Encyclopedia of Medical Devices and Instrumentation

Second Edition

John G. Webster

ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition

VOLUME 1

Alloys, Shape Memory – Brachytherapy, Intravascular

ENCYCLOPEDIA OF MEDICAL DEVICES AND INSTRUMENTATION, SECOND EDITION

ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition
Volume 1

Alloys, Shape Memory – Brachytherapy, Intravascular

Edited by

John G. Webster
University of Wisconsin–Madison

The *Encyclopedia of Medical Devices and Instrumentation* is available online at http://www.mrw.interscience.wiley.com/emdi

WILEY-INTERSCIENCE

A John Wiley & Sons, Inc., Publication

Published by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Encylopedia of medical devices & instrumentation/by John G. Webster,

editor in chief. – 2nd ed.
p. ; cm.
Rev. ed. of: Encyclopedia of medical devices and instrumentation. 1988.
Includes bibliographical references and index.
ISBN-13 978-0-471-26358-6 (set : cloth)
ISBN-10 0-471-26358-3 (set : cloth)
ISBN-13 978-0-470-04066-9 (v. 1 : cloth)
ISBN-10 0-470-04066-1 (v. 1 : cloth)

1. Medical instruments and apparatus–Encyclopedias. 2. Biomedical engineering–Encyclopedias. 3. Medical physics–Encyclopedias. 4. Medicine–Data processing–Encyclopedias. I. Webster, John G., 1932- . II. Title: Encyclopedia of medical devices and instrumentation.

[DNLM: 1. Equipment and Supplies–Encyclopedias–English. W 13

E555 2006]
R856.A3E53 2006
610.2803–dc22
2005028946

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTRIBUTOR LIST

ABDEL HADY, MAZEN, *McMaster University, Hamilton, Ontario Canada*, Bladder Dysfunction, Neurostimulation of

ABEL, L.A., *University of Melbourne, Melbourne, Australia*, Ocular Motility Recording and Nystagmus

ABREU, BEATRIZ C., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

ALEXANDER, A.L., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

ALI, ABBAS, *University of Illinois, at Urbana-Champaign*, Bioinformatics

ALI, MÜFTÜ, *School of Dental Medicine, Boston, Massachusetts*, Tooth and Jaw, Biomechanics of

ALPERIN, NOAM, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

ANSON, DENIS, *College Misericordia, Dallas, Pennsylvania*, Environmental Control

ARENA, JOHN C., *VA Medical Center and Medical College of Georgia*, Biofeedback

ARIEL, GIDEON, *Ariel Dynamics, Canyon, California*, Biomechanics of Exercise Fitness

ARMSTRONG, STEVE, *University of Iowa, Iowa City, Iowa*, Biomaterials for Dentistry

ASPDEN, R.M., *University of Aberdeen, Aberdeen, United Kingdom*, Ligament and Tendon, Properties of

AUBIN, C.E., *Polytechniquie Montreal, Montreal Quebec, Canada*, Scoliosis, Biomechanics of

AYRES, VIRGINIA M., *Michigan State University, East Lansing, Michigan*, Microscopy, Scanning Tunneling

AZANGWE, G., Ligament and Tendon, Properties of

BACK, LLOYD H., *California Institute of Technology, Pasadena, California*, Coronary Angioplasty and Guidewire Diagnostics

BADYLAK, STEPHEN F., *McGowan Institute for Regenerative Medicine, Pittsburgh, Pennsylvania*, Sterilization of Biologic Scaffold Materials

BANDYOPADHYAY, AMIT, *Washington State University, Pullman, Washington*, Orthopedic Devices, Materials and Design for

BANERJEE, RUPAK K., *University of Cincinnati, Cincinnati, Ohio*, Coronary Angioplasty and Guidewire Diagnostics

BARBOUR, RANDALL L., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

BARKER, STEVEN J., *University of Arizona, Tucson, Arizona*, Oxygen Monitoring

BARTH, ROLF F., *The Ohio State University, Columbus, Ohio*, Boron Neutron Capture Therapy

BECCHETTI, F.D., *University of Michigan, Ann Arbor, Michigan*, Radiotherapy, Heavy Ion

BELFORTE, GUIDO, *Politecnico di Torino – Department of Mechanics*, Laryngeal Prosthetic Devices

BENKESER, PAUL, *Georgia Institute of Technology, Atlanta, Georgia*, Biomedical Engineering Education

BENNETT, JAMES R., *University of Iowa, Iowa City, Iowa*, Digital Angiography

BERSANO-BEGEY, TOMMASO, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

BIGGS, PETER J., *Harvard Medical School, Boston, Massachusetts*, Radiotherapy, Intraoperative

BIYANI, ASHOK, *University of Toledo, and Medical College of Ohio, Toledo, Ohio*, Human Spine, Biomechanics of

BLOCK, W.F., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

BLUE, THOMAS E., *The Ohio State University, Columbus, Ohio*, Boron Neutron Capture Therapy

BLUMSACK, JUDITH T., *Disorders Auburn University, Auburn, Alabama*, Audiometry

BOGAN, RICHARD K., *University of South Carolina, Columbia, South Carolina*, Sleep Laboratory

BOKROS, JACK C., *Medical Carbon Research Institute, Austin, Texas*, Biomaterials, Carbon

BONGIOANNINI, GUIDO, *ENT Division Mauriziano Hospital, Torino, Italy*, Laryngeal Prosthetic Devices

BORAH, JOSHUA, *Applied Science Laboratories, Bedford, Massachusetts*, Eye Movement, Measurement Techniques for

BORDEN, MARK, *Director of Biomaterials Research, Irvine, California*, Biomaterials, Absorbable

BORTON, BETTIE B., *Auburn University Montgomery, Montgomery, Alabama*, Audiometry

BORTON, THOMAS E., *Auburn University Montgomery, Montgomery, Alabama*, Audiometry

BOSE SUSMITA,, *Washington State University, Pullman, Washington*, Orthopedic Devices, Materials and Design for

BOVA, FRANK J., *M. D. Anderson Cancer Center Orlando, Orlando, FL*, Radiosurgery, Stereotactic

BRENNER, DAVID J., *Columbia University Medical Center, New York, New York*, Computed Tomography Screening

BREWER, JOHN M., *University of Georgia*, Electrophoresis

BRIAN, L. DAVIS, *Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, Ohio*, Skin, Biomechanics of

BRITT, L.D., *Eastern Virginia Medical School, Norfolk, Virginia*, Gastrointestinal Hemorrhage

BRITT, R.C., *Eastern Virginia Medical School, Norfolk, Virginia*, Gastrointestinal Hemorrhage

BROZIK, SUSAN M., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

BRUNER, JOSEPH P., *Vanderbilt University Medical Center, Nashville, Tennessee*, Intrauterine Surgical Techniques

BRUNSWIG NEWRING, KIRK A., *University of Nevada, Reno, Nevada*, Sexual Instrumentatio n

BRUYANT, PHILIPPE P., *University of Massachusetts, North Worcester, Massachusetts*, Nuclear Medicine, Computers in

BUNNELL, BERT J., *Bunnell Inc., Salt Lake City, Utah*, High Frequency Ventilation

CALKINS, JERRY M., *Defense Research Technologies, Inc., Rockville, Maryland*, Medical Gas Analyzers

CANNON, MARK, *Northwestern University, Chicago, Illinois*, Resin-Based Composites

CAPPELLERI, JOSEPH C., *Pfizer Inc., Groton, Connecticut*, Quality-of-Life Measures, Clinical Significance of

CARDOSO, JORGE, *University of Madeira, Funchal, Portugal*, Office Automation Systems

CARELLO, MASSIMILIANA, *Politecnicodi Torino – Department of Mechanics*, Laryngeal Prosthetic Devices

CASKEY, THOMAS C., *Cogene Biotech Ventures, Houston, Texas*, Polymerase Chain Reaction

CECCIO, STEVEN, *University of Michigan, Ann Arbor, Michigan*, Heart Valve Prostheses, In Vitro Flow Dynamics of

CHAN, JACKIE K., *Columbia University, New York, New York*, Photography, Medical

CHANDRAN, K.B., *University of Iowa, Iowa City, Iowa*, Heart Valve Prostheses

CHATZANDROULIS, S., *NTUA, Athens, Attiki, Greece*, Capacitive Microsensors for Biomedical Applications

CHAVEZ, ELIANA, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CHEN, HENRY, *Stanford University, Palo Alto, California*, Exercise Stress Testing

CHEN, JIANDE, *University of Texas Medical Branch, Galveston, Texas*, Electrogastrogram

CHEN, YAN, *Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, Ohio*, Skin, Biomechanics of

CHEYNE, DOUGLAS, *Hospital for Sick Children Research Institute*, Biomagnetism

CHUI, CHEN-SHOU, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

CLAXTON, NATHAN S., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

CODERRE, JEFFREY A., *Massachus etts Institute of Technology, Cambridge, Massachusetts*, Boron Neutron Capture Therapy

COLLINS, BETH, *University of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

COLLINS, DIANE, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CONSTANTINOU, C., *Columbia University Radiation Oncology, New York, New York*, Phantom Materials in Radiology

COOK, ALBERT, *University of Alberta*, *Edmonton, Alberta, Canada*, Communication Devices

COOPER, RORY, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CORK, RANDALL C., *Louisiana State University, Shreveport, Louisiana*, Monitoring, Umbilical Artery and Vein, Blood Gas Measurements; Transcuta neous Electrical Nerve Stimulation (TENS); Ambulatory Monitoring

COX, JOSEPHINE H., *Walter Reed Army Institute of Research, Rockville, Maryland*, Blood Collection and Processing

CRAIG, LEONARD, *Feinberg School of Medicine of Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

CRESS, CYNTHIA J., *University of Nebraska, Lincoln, Nebraska*, Communicative Disorders, Computer Applications for

CUMMING, DAVID R.S., *University of Glasgow, Glasgow, United Kingdom*, Ion-Sensitive Field-Effect Transistors

CUNNINGHAM, JOHN R., *Camrose, Alberta, Canada*, Cobalt 60 Units for Radiotherapy

D'ALESSANDRO, DAVID, *Montefiore Medical Center, Bronx, New York*, Heart–Lung Machines

D'AMBRA, MICHAEL N., *Harvard Medical School, Cambridge, Massachusetts*, Cardiac Output, Thermodilution Measurement of

DADSETAN, MAHROKH, *Mayo Clinic, College of Medicine, Rochester, Minnesota*, Microscopy, Electron

DALEY, MICHAEL L., *The University of Memphis, Memphis, Tennessee*, Monitoring, Intracranial Pressure

DAN, LOYD, *Linköping University, Linköping, Sweden*, Thermocouples

DAS, RUPAK, *University of Wisconsin, Madison, Wisconsin*, Brachytherapy, High Dosage Rate

DATTAWADKAR, AMRUTA M., *University of Wisconsin, Madison, Madison, Wisconsin*, Ocular Fundus Reflectometry

DAVIDSON, MICHAEL W., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

DE LUCA, CARLO, *Boston University, Boston, Massachusetts*, Electromyography

DE SALLES, ANTONIO A.F., *UCLA Medical School, Los Angeles, California*, Stereotactic Surgery

DECAU, SABIN, *University of Maryland, School of Medicine*, Shock, Treatment of

DECHOW, PAUL C., *A & M University Health Science Center, Dallas, Texas*, Strain Gages

DELBEKE, JEAN, *Catholique University of Louvain, Brussels, Belgium*, Visual Prostheses

DELL'OSSO, LOUIS F., *Case Western Reserve University, Cleveland, Ohio*, Ocular Motility Recording and Nystagmus

DELORME, ARNAUD, *University of San Diego, La Jolla, California*, Statistical Methods

DEMENKOFF, JOHN, *Mayo Clinic, Scottsdale, Arizona*, Pulmonary Physiology

DEMIR, SEMAHAT S., *The University of Memphis and The University of Tennessee Health Science Center, Memphis, Tennessee*, Electrophysiology

DEMLING, ROBERT H., *Harvard Medical School*, Skin Substitute for Burns, Bioactive

DENNIS, MICHAEL J., *Medical University of Ohio, Toledo, Ohio*, Computed Tomography

DESANTI, LESLIE, *Harvard Medical School*, Skin Substitute for Burns, Bioactive

DEUTSCH, STEVEN, *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

DEVINENI, TRISHUL, *Conemaugh Health System*, Biofeedback

DI BELLA EDWARD, V.R., *University of Utah*, Tracer Kinetics

DIAKIDES, NICHOLAS A., *Advanced Concepts Analysis, Inc., Falls Church, Virginia*, Thermography

DOLAN, PATRICIA L., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

DONOVAN, F.M., *University of South Alabama*, Cardiac Output, Indicator Dilution Measurement of

DOUGLAS, WILSON R., *Children's Hospital of Philadelphia, Philadelphia, Pennsylvania*, Intrauterine Surgical Techniques

DRAPER, CRISSA, *University of Nevada, Reno, Nevada*, Sexual Instrumentation

DRZEWIECKI, TADEUSZ M., *Defense Research Technologies, Inc., Rockville, Maryland*, Medical Gas Analyzers

DURFEE, W.K., *University of Minnesota, Minneapolis, Minnesota*, Rehabilitation and Muscle Testing

DYRO, JOSEPH F., *Setauket, New York*, Safety Program, Hospital

DYSON, MARY, *Herts, United Kingdom*, Heat and Cold, Therapeutic

ECKERLE, JOSEPH S., *SRI International, Menlo Park, California*, Tonometry, Arterial

EDWARDS, BENJAMIN, *University of Wisconsin-Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

EDWARDS, THAYNE L., *University of Washington, Seattle, Washington*, Chromatography

EKLUND, ANDERS, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

EL SOLH, ALI A., *Erie County Medical Center, Buffalo, New York*, Sleep Studies, Computer Analysis of

ELMAYERGI, NADER, *McMaster University, Hamilton, Ontario, Canada*, Bladder Dysfunction, Neurostimulation of

ELSHARYDAH, AHMAD, *Louisiana State University, Baton Rouge, Louisiana*, Ambulatory Monitoring; Monitoring, Umbilical Artery and Vein, Blood Gas Measurements

FADDY, STEVEN C., *St. Vincents Hospital, Sydney, Darlinghurst, Australia*, Cardiac Output, Fick Technique for

FAHEY, FREDERIC H., *Childrens Hospital Boston*, Computed Tomography, Single Photon Emission

FAIN, S.B., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

FELDMAN, JEFFREY, *Childrens Hospital of Philadelphia, Philadelphia, Pennsylvania*, Anesthesia Machines

FELLERS, THOMAS J., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

FERRARA, LISA, *Cleveland Clinic Foundation, Cleveland, Ohio*, Human Spine, Biomechanics of

FERRARI, MAURO, *The Ohio State University, Columbus, Ohio*, Drug Delivery Systems

FONTAINE, ARNOLD A., *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

FOUST, MILTON J., JR, *Medical University of South Carolina Psychiatry and Behavioral Sciences, Charleston, South Carolina*, Electroconvulsive Therapy

FRASCO, PETER, *Mayo Clinic Scottsdale, Scottsdale, Arizona*, Temperature Monitoring

FRAZIER, JAMES, *Louisiana State University, Baton Rouge, Louisiana*, Ambulatory Monitoring

FREIESLEBEN DE BLASIO, BIRGITTE, *University of Oslo, Oslo, Norway*, Impedance Spectroscopy

FRESTA, MASSIMO, *University of Catanzaro Magna Græcia, Germaneto (CZ), Italy*, Drug Delivery Systems

FREYTES, DONALD O., *McGowan Institute for Regenerative Medicine, Pittsburgh Pennsylvania*, Sterilization of Biologic Scaffold Materials

FROEHLICHER, VICTOR, *VA Medical Center, Palo Alto, California*, Exercise Stress Testing

FUNG, EDWARD K., *Columbia University, New York, New York*, Photography, Medical

GAGE, ANDREW A., *State University of New York at Buffalo, Buffalo, New York*, Cryosurgery

GAGLIO, PAUL J., *Columbia University College of Physicians and Surgeons*, Liver Transplantation

GARDNER, REED M., *LDS Hospital and Utah University, Salt Lake City, Utah*, Monitoring, Hemodynamic

GEJERMAN, GLEN, *Hackensack University Medical, Hackensack, New Jersey*, Radiation Therapy, Quality Assurance in

GEORGE, MARK S., *Medical University of South Carolina Psychiatry and Behavioral Sciences, Charleston, South Carolina*, Electroconvulsive Therapy

GHARIEB, R.R., *Infinite Biomedical Technologies, Baltimore, Maryland*, Neurological Monitors

GLASGOW, GLENN P., *Loyola University of Chicago, Maywood, Illinois*, Radiation Protection Instrumentation

GLASGOW, GLENN, *University of Wisconsin-Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

GOEL, VIJAY K., *University of Toledo, and Medical College of Ohio, Toledo, Ohio*, Human Spine, Biomechanics of

GOETSCH, STEVEN J., *San Diego Gamma Knife Center, La Jolla, California*, Gamma Knife

GOLDBERG, JAY R., *Marquette University Milwaukee, Wisconsin*, Minimally Invasive Surgery

GOLDBERG, ZELENNA, *Department of Radiation Oncology, Davis, California*, Ionizing Radiation, Biological Effects of

GOPALAKRISHNAKONE, P., *National University of Singapore, Singapore*, Immunologically Sensitive Field-Effect Transistors

GOPAS, JACOB, *Ben Gurion University of the Negev, Beer Sheva, Israel*, Monoclonal Antibodies

GORGULHO, ALESSANDRA, *UCLA Medical School, Los Angeles, California*, Stereotactic Surgery

GOUGH, DAVID A., *University of California, La Jolla, California*, Glucose Sensors

GOUSTOURIDIS, D., *NTUA, Athens, Attiki, Greece*, Capacitive Microsensors for Biomedical Applications

GRABER, HARRY L., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

GRAÇA, M., *Louisiana State University, Baton Rouge, Louisiana*, Boron Neutron Capture Therapy

GRANT, WALTER III, *Baylor College of Medicine, Houston, Texas*, Radiation Therapy, Intensity Modulated

GRAYDEN, EDWARD, *Mayo Health Center, Albertlea, Minnesota*, Cardiopulmonary Resuscitation

GREEN, JORDAN R., *University of Nebraska, Lincoln, Nebraska*, Communicative Disorders, Computer Applications for

HAEMMERICH, DIETER, *Medical University of South Carolina, Charleston, South Carolina*, Tissue Ablation

HAMAM, HABIB, *Université de Moncton, Moncton New Brunswick, Canada*, Lenses, Intraocular

HAMMOND, PAUL A., *University of Glasgow, Glasgow, United Kingdom*, Ion-Sensitive Field-Effect Transistors

HANLEY, JOSEPH, *Hackensack University Medical, Hackensack, New Jersey*, Radiation Therapy, Quality Assurance in

HARLEY, BRENDAN A., *Massachusetts Institute of Technology*, Skin Tissue Engineering for Regeneration

HARPER, JASON C., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

HASMAN, ARIE, *Maastricht, The Netherlands*, Medical Education, Computers in

HASSOUNA, MAGDY, *Toronto Western Hospital, Toronto, Canada*, Bladder Dysfunction, Neurostimulation of

HAYASHI, KOZABURO, *Okayama University of Science, Okayama, Japan*, Arteries, Elastic Properties of

HENCH, LARRY L., *Imperial College London, London, United Kingdom*, Biomaterials: Bioceramics

HETTRICK, DOUGLAS A., *Sr. Principal Scientist Medtronic, Inc., Minneapolis, Minnesota*, Bioimpedance in Cardiovascular Medicine

HIRSCH-KUCHMA, MELISSA, *University of Central Florida NanoScience Technology Center, Orlando, Florida*, Biosurface Engineering

HOLDER, GRAHAM E., *Moorfields Eye Hospital, London, United Kingdom*, Electroretinography

HOLMES, TIMOTHY, *St. Agnes Cancer Center, Baltimore, Maryland*, Tomotherapy

HONEYMAN-BUCK, JANICE C., *University of Florida, Gainesville, Florida*, Radiology Information Systems

HOOPER, BRETT A., *Areté Associates, Arlington, Virginia*, Endoscopes

HORN, BRUCE, *Kaiser Permanente, Los Angeles, California*, X-Rays Production of

HORNER, PATRICIA I., *Biomedical Engineering Society Landover, Maryland*, Medical Engineering Societies and Organizations

HOROWITZ, PAUL M., *University of Texas, San Antonio, Texas*, Fluorescence Measurements

HOU, XIAOLIN, *Risø National Laboratory, Roskilde, Denmark*, Neutron Activation Analysis

HOVORKA, ROMAN, *University of Cambridge, Cambridge, United Kingdom*, Pancreas, Artificial

HUANG, H.K., *University of Southern California*, Teleradiology

HUNT, ALAN J., *University of Michigan, Ann Arbor, Michigan*, Optical Tweezers

HUTTEN, HELMUT, *University of Technology, Graz, Australia*, Impedance Plethysmography

IAIZZO, P.A., *University of Minnesota, Minneapolis, Minnesota*, Rehabilitation and Muscle Testing

IBBOTT, GEOFFREY S., *Anderson Cancer Center, Houston, Texas*, Radiation Dosimetry, Three-Dimensional

INGHAM, E., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

ISIK, CAN, *Syracuse University, Syracuse, New York*, Blood Pressure Measurement

JAMES, SUSAN P., *Colorado State University, Fort Collins, Colorado*, Biomaterials: Polymers

JENSEN, WINNIE, *Aalborg University, Aalborg, Denmark*, Electroneurography

JIN, CHUNMING, *North Carolina State University, Raleigh, North Carolina*, Biomaterials, Corrosion and Wear of

JIN, Z.M., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

JOHNSON, ARTHUR T., *University of Maryland College Park, Maryland*, Medical Engineering Societies and Organizations

JONES, JULIAN R., *Imperial College London, London, United Kingdom*, Biomaterials: Bioceramics

JOSHI, ABHIJEET, *Abbott Spine, Austin, Texas*, Spinal Implants

JUNG, RANU, *Arizona State University, Tempe, Arizona*, Functional Electrical Stimulation

JURISSON, SILVIA S., *University of Missouri Columbia, Missouri*, Radionuclide Production and Radioactive Decay

KAEDING, PATRICIA J., *Godfrey & Kahn S.C., Madison, Wisconsin*, Codes and Regulations: Medical Devices

KAMATH, CELIA C., *Mayo Clinic, Rochester, Minnesota*, Quality-of-Life Measures, Clinical Significance of

KANE, MOLLIE, *Madison, Wisconsin*, Contraceptive Devices

KATHERINE, ANDRIOLE P., *Harvard Medical School, Boston, Massachusetts*, Picture Archiving and Communication Systems

KATSAGGELOS, AGGELOS K., *Northwestern University, Evanston, Illinois*, DNA Sequencing

KATZ, J. LAWRENCE, *University of Missouri-Kansas City, Kansas City, Missouri*, Bone and Teeth, Properties of

KESAVAN, SUNIL, *Akebono Corporation, Farmington Hills, Michigan*, Linear Variable Differential Transformers

KHANG, GILSON, *Chonbuk National University*, Biomaterials: Tissue Engineering and Scaffolds

KHAODHIAR, LALITA, *Harvard Medical School, Boston, Massachusetts*, Cutaneous Blood Flow, Doppler Measurement of

KIM, MOON SUK, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

KIM, YOUNG KON, *Inje University, Kimhae City, Korea*, Alloys, Shape Memory

KINDWALL, ERIC P., *St. Luke's Medical Center, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

KING, MICHAEL A., *University of Massachusetts, North Worcester, Massachusetts*, Nuclear Medicine, Computers in

KLEBE, ROBERT J., *University of Texas, San Antonio, Texas*, Fluorescence Measurements

KLEIN, BURTON, *Burton Klein Associates, Newton, Massachusetts*, Gas and Vacuum Systems, Centrally Piped Medical

KNOPER, STEVEN R., *University of Arizona College of Medicine*, Ventilatory Monitoring

KONTAXAKIS, GEORGE, *Universidad Politécnica de Madrid, Madrid, Spain*, Positron Emission Tomography

KOTTKE-MARCHANT, KANDICE, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Vascular Graft Prosthesis

KRIPFGANS, OLIVER, *University of Michigan, Ann Arbor, Michigan*, Ultrasonic Imaging

KULKARNI, AMOL D., *University of Wisconsin–Madison, Madison, Wisconsin*, Ocular Fundus Reflectometry, Visual Field Testing

KUMARADAS, J. CARL, *Ryerson University, Toronto, Ontario, Canada*, Hyperthermia, Interstitial

KUNICKA, JOLANTA, *Bayer HealthCare LLC, Tarrytown, New York*, Differential Counts, Automated

KWAK, KWANJ JOO, *University of Miami Miller School of Medicine, Miami, Florida*, Microscopy, Scanning Force

LAKES, RODERIC, *University of Wisconsin-Madison*, Bone and Teeth, Properties of

LAKKIREDDY, DHANUNJAYA, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

LARSEN, COBY, *Case Western Reserve University, Cleveland, Ohio*, Vascular Graft Prosthesis

LASTER, BRENDA H., *Ben Gurion University of the Negev, Beer Sheva, Israel*, Monoclonal Antibodies

LATTA, LOREN, *University of Miami, Coral Gables, Florida*, Rehabilitation, Orthotics in

LEDER, RON S., *Universidad Nacional Autonoma de Mexico Mexico, Distrito Federal*, Continuous Positive Airway Pressure

LEE, CHIN, *Harvard Medical School, Boston, Massachusetts*, Radiotherapy Treatment Planning, Optimization of; Hyperthermia, Interstitial

LEE, HAI BANG, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

LEE, SANG JIN, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

LEI, LIU, *Department of General Engineering, Urbana, Illinois*, Bioinformatics

LEI, XING, *Stanford University, Stanford, California*, Radiation Dose Planning, Computer-Aided

LEWIS, MATTHEW C., *Medical College of Wisconsin, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

LI, CHAODI, *University of Notre Dame, Notre Dame, Indiana*, Bone Cement, Acrylic

LI, JONATHAN G., *University of Florida, Gainesville, Florida*, Radiation Dose Planning, Computer-Aided

LI, QIAO, *University of Michigan, Ann Arbor, Michigan*, Immunotherapy

LI, YANBIN, *University of Arkansas, Fayetteville, Arkansas*, Piezoelectric Sensors

LIBOFF, A.R., *Oakland University, Rochester, Michigan*, Bone Ununited Fracture and Spinal Fusion, Electrical Treatment of

LIGAS, JAMES, *University of Connecticut, Farmington, Connecticut*, Respiratory Mechanics and Gas Exchange

LIMOGE, AIME, *The René Descartes University of Paris, Paris, France*, Electroanalgesia, Systemic

LIN, PEI-JAN PAUL, *Beth Israel Deaconess Medical Center, Boston, Massachusets*, Mammography

LIN, ZHIYUE, *University of Kansas Medical Center, Kansas City, Kansas*, Electrogastrogram

LINEAWEAVER, WILLIAM C., *Unive rsity of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

LIPPING, TARMO, *Tampere University of Technology, Pori, Finland*, Monitoring in Anesthesia

LIU, XIAOHUA, *The University of Michigan, Ann Arbor, Michigan*, Polymeric Materials

LLOYD, J.J., *Regional Medical Physics Department, Newcastle-upon-Tyne, United Kingdom*, Ultraviolet Radiation in Medicine

LOEB, ROBERT, *University of Arizona, Tuscon, Arizona*, Anesthesia Machines

LOPES DE MELO, PEDRO, *State University of Rio de Janeiro, Térreo Salas, Maracanã, Thermistors*

LOUDON, ROBERT G., Lung Sounds

LOW, DANIEL A., *Washington University School of Medicine, St. Louis, Missouri*, Radiation Therapy Simulator

LU, LICHUN, *Mayo Clinic, College of Medicine, Rochester, Minnesota*, Microscopy, Electron

LU, ZHENG FENG, *Columbia University, New York, New York*, Screen-Film Systems

LYON, ANDREW W., *University of Calgary, Calgary, Canada*, Flame Atomic Emission Spectrometry and Atomic Absorption Spectrometry

LYON, MARTHA E., *University of Calgary, Calgary, Canada*, Flame Atomic Emission Spectrometry and Atomic Absorption Spectrometry

MA, C-M CHARLIE, *Fox Chase Cancer Center, Philadelphia, Pennsylvania*, X-Ray Therapy Equipment, Low and Medium Energy

MACIA, NARCISO F., *Arizona State University at the Polytechnic Campus, Mesa, Arizona*, Pneumotachometers

MACKENZIE, COLIN F., *University of Maryland, School of Medicine*, Shock, Treatment of

MACKIE, THOMAS R., *University of Wisconsin, Madison, Wisconsin*, Tomotherapy

MADNANI, ANJU, *LSU Medical Centre, Shreveport, Louisiana*, Transcutaneous Electrical Nerve Stimulation (TENS)

MADNANI, SANJAY, *LSU Medical Centre, Shreveport, Louisiana*, Transcutaneous Electrical Nerve Stimulation (TENS)

MADSEN, MARK T., *University of Iowa, Iowa City, Iowa*, Anger Camera

MAGNANO, MAURO, *ENT Division Mauriziano Hospital, Torino, Italy*, Drug Delivery Systems

MANDEL, RICHARD, *Boston University School of Medicine, Boston, Massachusetts*, Colorimetry

MANNING, KEEFE B., *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

MAO, JEREMY J., *University of Illinois at Chicago, Chicago, Illinois*, Cartilage and Meniscus, Properties of

MARCOLONGO, MICHELE, *Drexel University, Philadelphia, Pennsylvania*, Spinal Implants

MAREK, MIROSLAV, *Georgia Institute of Technology, Atlanta, Georgia*, Biomaterials, Corrosion and Wear of

MARION, NICHOLAS W., *University of Illinois at Chicago, Chicago, Illinois*, Cartilage and Meniscus, Properties of

MASTERS, KRISTYN S., *University of Wisconsin, Madison, Wisconsin*, Tissue Engineering

MAUGHAN, RICHARD L., *Hospital of the University of Pennsylvania*, Neutron Beam Therapy

MCADAMS, ERIC, *University of Ulster at Jordanstown, Newtownabbey, Ireland*, Bioelectrodes

MCARTHUR, SALLY L., *University of Sheffield, Sheffield, United Kingdom*, Biomaterials, Surface Properties of

MCEWEN, MALCOM, *National Research Council of Canada, Ontario, Canada*, Radiation Dosimetry for Oncology

MCGOWAN, EDWARD J., *E.J. McGowan & Associates*, Biofeedback

MCGRATH, SUSAN, *Dartmouth College, Hanover, New Hampshire*, Oxygen Analyzers

MEEKS, SANFORD L., *University of Florida, Gainesville, Florida*, Radiosurgery, Stereotactic

MELISSA, PETER, *University of Central Florida NanoScience Technology Center, Orlando, Florida*, Biosurface Engineering

MENDELSON, YITZHAK, *Worcester Polytechnic Institute*, Optical Sensors

METZKER, MICHAEL L., *Baylor College of Medicine, Houston, Texas*, Polymerase Chain Reaction

MEYEREND, M.E., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

MICHLER, ROBERT, *Montefiore Medical Center, Bronx, New York*, Heart–Lung Machines

MICIC, MIODRAG, *MP Biomedicals LLC, Irvine, California*, Microscopy and Spectroscopy, Near-Field

MILLER, WILLIAM, *University of Missouri Columbia, Missouri*, Radionuclide Production and Radioactive Decay

MITTRA, ERIK, *Stony Brook University, New York*, Bone Density Measurement

MODELL, MARK, *Harvard Medical School, Boston, Massachusetts*, Fiber Optics in Medicine

MORE, ROBERT B., *RBMore Associates, Austin, Texas* Biomaterials Carbon

MORE, ROBERT, *Austin, Texas*, Heart Valves, Prosthetic

MORROW, DARREN, *Royal Adelaide Hospital, Adelaide, Australia*, Intraaortic Balloon Pump

MOURTADA, FIRAS, *MD Anderson Cancer Center, Houston, Texas*, Brachytherapy, Intravascular

MOY, VINCENT T., *University of Miami, Miller School of Medicine, Miami, Florida*, Microscopy, Scanning Force

MÜFTÜ, SINAN, *Northeastern University, Boston, Massachusetts*, Tooth and Jaw, Biomechanics of

MURPHY, RAYMOND L.H., Lung Sounds

MURPHY, WILLIAM L., *University of Wisconsin, Madison, Wisconsin*, Tissue Engineering

MURRAY, ALAN, *Newcastle University Medical Physics, Newcastle upon Tyne, United Kingdom*, Pace makers

MUTIC, SASA, *Washington University School of Medicine, St. Louis, Missouri*, Radiation Therapy Simulator

NARAYAN, ROGER J., *University of North Carolina, Chapel Hill, North Carolina*, Biomaterials, Corrosion and Wear of

NATALE, ANDREA, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

NAZERAN, HOMER, *The University of Texas, El Paso, Texas*, Electrocardiography, Computers in

NEUMAN, MICHAEL R., *Michigan Technological University, Houghton, Houghton, Michigan*, Fetal Monitoring, Neonatal Monitoring

NEUZIL, PAVEL, *Institute of Bioengineering and Nanotechnology, Singapore*, Immunologically Sensitive Field-Effect Transistors

NICKOLOFF, EDWARD L., *Columbia University, New York, New York*, X-Ray Quality Control Program

NIEZGODA, JEFFREY A., *Medical College of Wisconsin, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

NISHIKAWA, ROBERT M., *The University of Chicago, Chicago, Illinois*, Computer-Assisted Detection and Diagnosis

NUTTER, BRIAN, *Texas Tech University, Lubbock, Texas*, Medical Records, Computers in

O'DONOHUE, WILLIAM, *University of Nevada, Reno, Nevada*, Sexual Instrumentation

ORTON, COLIN, *Harper Hospital and Wayne State University, Detroit, Michigan*, Medical Physics Literature

OZCELIK, SELAHATTIN, *Texas A&M University, Kingsville, Texas*, Drug Infusion Systems

PANITCH, ALYSSA, *Arizona State University, Tempe, Arizona*, Biomaterials: An Overview

PAOLINO, DONATELLA, *University of Catanzaro Magna Græcia, Germaneto (CZ), Italy*, Drug Delivery Systems

PAPAIOANNOU, GEORGE, *University of Wisconsin, Milwaukee, Wisconsin*, Joints, Biomechanics of

PARK, GRACE E., *Purdue University, West Lafayette, Indiana*, Porous Materials for Biological Applications

PARMENTER, BRETT A., *State University of New York at Buffalo, Buffalo, New York*, Sleep Studies, Computer Analysis of

PATEL, DIMPI, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

PEARCE, JOHN, *The University of Texas, Austin, Texas*, Electrosurgical Unit (ESU)

PELET, SERGE, *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

PERIASAMY, AMMASI, *University of Virginia, Charlottesville, Virginia*, Cellular Imaging

PERSONS, BARBARA L., *University of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

PIPER, IAN, *The University of Memphis, Memphis, Tennessee*, Monitoring, Intracranial Pressure

POLETTO, CHRISTOPHER J., *National Institutes of Health*, Tactile Stimulation

PREMINGER, GLENN M., *Duke University Medical Center, Durham, North Carolina*, Lithotripsy

PRENDERGAST, PATRICK J., *Trinity College, Dublin, Ireland*, Orthopedics, Prosthesis Fixation for

PREVITE, MICHAEL, *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

PURDY, JAMES A., *UC Davis Medical Center, Sacramento, California*, Radiotherapy Accessories

QI, HAIRONG, *Advanced Concepts Analysis, Inc., Falls Church, Virginia*, Thermography

QIN, YIXIAN, *Stony Brook University, New York*, Bone Density Measurement

QUAN, STUART F., *University of Arizona, Tucson, Arizona*, Ventilatory Monitoring

QUIROGA, RODRIGO QUIAN, *University of Leicester, Leicester, United Kingdom*, Evoked Potentials

RAHAGHI, FARBOD N., *University of California, La Jolla, California*, Glucose Sensors

RAHKO, PETER S., *University of Wisconsin Medical School*, Echocardiography and Doppler Echocardiography

RALPH, LIETO, *University of Wisconsin–Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

RAMANATHAN, LAKSHMI, *Mount Sinai Medical Center*, Analytical Methods, Automated

RAO, SATISH S.C., *University of Iowa College of Medicine, Iowa City, Iowa*, Anorectal Manometry

RAPOPORT, DAVID M., *NYU School of Medicine, New York, New York*, Continuous Positive Airway Pressure

REBELLO, KEITH J., *The Johns Hopkins University Applied Physics Lab, Laurel, Maryland*, Micro surgery

REDDY, NARENDER, *The University of Akron, Akron, Ohio*, Linear Variable Differential Transformers

REN-DIH, SHEU, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

RENGACHARY, SETTI S., *Detroit, Michigan*, Human Spine, Biomechanics of

REPPERGER, DANIEL W., *Wright-Patterson Air Force Base, Dayton, Ohio*, Human Factors in Medical Devices

RITCHEY, ERIC R., *The Ohio State University, Columbus, Ohio*, Contact Lenses

RIVARD, MARK J., *Tufts New England Medical Center, Boston, Massachusetts*, Imaging Devices

ROBERTSON, J. DAVID, *University of Missouri, Columbia, Missouri*, Radionuclide Production and Radioactive Decay

ROTH, BRADLEY J., *Oakland University, Rochester, Michigan*, Defibrillators

ROWE-HORWEGE, R. WANDA, *University of Texas Medical School, Houston, Texas*, Hyperthermia, Systemic

RUMSEY, JOHN W., *University of Central Florida, Orlando, Florida*, Biosurface Engineering

RUTKOWSKI, GREGORY E., *University of Minnesota, Duluth, Minnesota*, Engineered Tissue

SALATA, O.V., *University of Oxford, Oxford, United Kingdom*, Nanoparticles

SAMARAS, THEODOROS, *Aristotle University of Thessaloniki Department of Physics, Thessaloniki, Greece*, Thermometry

SANGOLE, ARCHANA P., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

SARKOZI, LASZLO, *Mount Sinai School of Medicine*, Analytical Methods, Automated

SCHEK, HENRY III, *University of Michigan, Ann Arbor, Michigan*, Optical Tweezers

SCHMITZ, CHRISTOPH H., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

SCHUCKERS, STEPHANIE A.C., *Clarkson University, Potsdam, New York*, Arrhythmia Analysis, Automated

SCOPE, KENNETH, *Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

SCOTT, ADZICK N., *University of Pennsylvania, Philadelphia, Pennsylvania*, Intrauterine Surgical Techniques

SEAL, BRANDON L., *Arizona State University, Tempe, Arizona*, Biomaterials: An Overview

SEALE, GARY, *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

SEGERS, PATRICK, *Ghent University, Belgium*, Hemodynamics

SELIM, MOSTAFA A., *Cleveland Metropolitan General Hospital, Palm Coast, Florida*, Colposcopy

SETHI, ANIL, *Loyola University Medical Center, Maywood, Illinois*, X-Rays: Interaction with Matter

SEVERINGHAUS, JOHN W., *University of California in San Francisco*, CO_2 Electrodes

SHALODI, ABDELWAHAB D., *Cleveland Metropolitan General Hospital, Palm Coast, Florida*, Colposcopy

SHANMUGASUNDARAM, SHOBANA, *New Jersey Institute of Technology, Newark, New Jersey*, Polymeric Materials

SHARD, ALEXANDER G., *University of Sheffield, Sheffield United Kingdom*, Biomaterials, Surface Properties of

SHEN, LI-JIUAN, *National Taiwan University School of Pharmacy, Taipei, Taiwan*, Colorimetry

SHEN, WEI-CHIANG, *University of Southern California School of Pharmacy, Los Angeles, California*, Colorimetry

SHERAR, MICHAEL D., *London Health Sciences Centre and University of Western Ontario, London, Ontario, Canada*, Hyperthermia, Interstitial

SHERMAN, DAVID, *The Johns Hopkins University, Baltimore, Maryland*, Electroencephalography

SHI, DONGLU, *University of Cincinnati, Cincinnati, Ohio*, Biomaterials, Testing and Structural Properties of

SHUCARD, DAVID W.M., *State University of New York at Buffalo, Buffalo, New York*, Sleep Studies, Computer Analysis of

SIEDBAND, MELVIN P., *University of Wisconsin, Madison, Wisconsin*, Image Intensifiers and Fluoroscopy

SILBERMAN, HOWARD, *University of Southern California, Los Angeles, California*, Nutrition, Parenteral

SILVERMAN, GORDON, *Manhattan College*, Computers in the Biomedical Laboratory

SILVERN, DAVID A., *Medical Physics Unit, Rabin Medical Center, Petah Tikva, Israel*, Prostate Seed Implants

SINHA, PIYUSH, *The Ohio State University, Columbus, Ohio*, Drug Delivery Systems

SINHA, ABHIJIT ROY, *University of Cincinnati, Cincinnati, Ohio*, Coronary Angioplasty and Guidewire Diagnostics

SINKJÆR, THOMAS, *Aalborg University, Aalborg, Denmark*, Electroneurography

SLOAN, JEFFREY A., *Mayo Clinic, Rochester, Minnesota*, Quality-of-Life Measures, Clinical Significance of

SO, PETER T.C., *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

SOBOL, WLAD T., *University of Alabama at Birmingham Health System, Birmingham, Alabama*, Nuclear Magnetic Resonance Spectroscopy

SOOD, SANDEEP, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

SPECTOR, MYRON, *Brigham and Women's Hospital, Boston, Massachusetts*, Biocompatibility of Materials

SPELMAN, FRANCIS A., *University of Washington*, Cochlear Prostheses

SRINIVASAN, YESHWANTH, *Texas Tech University, Lubbock, Texas*, Medical Records, Computers in

SRIRAM, NEELAMEGHAM, *University of Buffalo, Buffalo, New York*, Cell Counters, Blood

STARKO, KENTON R., *Point Roberts, Washington*, Physiological Systems Modeling

STARKSCHALL, GEORGE, *The University of Texas*, Radiotherapy, Three-Dimensional Conformal

STAVREV, PAVEL, *Cross Cancer Institute, Edmonton, Alberta, Canada*, Radiotherapy Treatment Planning, Optimization of

STENKEN, JULIE A., *Rensselaer Polytechnic Institute, Troy, New York*, Microdialysis Sampling

STIEFEL, ROBERT, *University of Maryland Medical Center, Baltimore, Maryland*, Equipment Acquisition

STOKES, I.A.F., *Polytechniquie Montreal, Montreal Quebec, Canada*, Scoliosis, Biomechanics of

STONE, M.H., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

SU, XIAO-LI, *BioDetection Instruments LLC, Fayetteville, Arkansas*, Piezoelectric Sensors

SUBHAN, ARIF, *Masterplan Technology Management, Chatsworth, California*, Equipment Maintenance, Biomedical

SWEENEY, JAMES D., *Arizona State University, Tempe, Arizona*, Functional Electrical Stimulation

SZETO, ANDREW Y.J., *San Diego State University, San Diego, California*, Blind and Visually Impaired, Assistive Technology for

TAKAYAMA, SHUICHI, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

TAMUL, PAUL C., *Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

TAMURA, TOSHIYO, *Chiba University School of Engineering, Chiba, Japan*, Home Health Care Devices

TANG, XIANGYANG, *GE Healthcare Technologies, Wankesha, Wisconsin*, Computed Tomography Simulators

TAYLOR, B.C., *The University of Akron, Akron, Ohio*, Cardiac Output, Indicator Dilution Measurement of

TEMPLE, RICHARD O., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

TEN, STANLEY, *Salt Lake City, Utah*, Electroanalgesia, Systemic

TERRY, TERESA M., *Walter Reed Army Institute of Research, Rockville, Maryland*, Blood Collection and Processing

THAKOR, N.V., *Johns Hopkins University, Baltimore, Maryland*, Neurological Monitors

THIERENS, HUBERT M.A., *University of Ghent, Ghent, Belgium*, Radiopharmaceutical Dosimetry

THOMADSEN, BRUCE, *University of Wisconsin–Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

TIPPER, J.L., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

TOGAWA, TATSUO, *Waseda University, Saitama, Japan*, Integrated Circuit Temperature Sensor

TORNAI, MARTIN, *Duke University, Durham, North Carolina*, X-Ray Equipment Design

TRAN-SON-TAY, ROGER, *University of Florida, Gainesville, Florida*, Blood Rheology

YENI, YENER N., *Henry Ford Hospital, Detroit, Michigan*, Joints, Biomechanics of

YLI-HANKALA, ARVI, *Tampere University of Technology, Pori, Finland*, Monitoring in Anesthesia

YOKO, KAMOTANI, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

YOON, KANG JI, *Korea Institute of Science and Technology, Seoul, Korea*, Micropower for Medical Applications

YORKE, ELLEN, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

YOSHIDA, KEN, *Aalborg University, Aalborg, Denmark*, Electroneurography

YOUNGSTEDT, SHAWN D., *University of South Carolina, Columbia, South Carolina*, Sleep Laboratory

YU, YIH-CHOUNG, *Lafayette College, Easton, Pennsylvania*, Blood Pressure, Automatic Control of

ZACHARIAH, EMMANUEL S., *University of Medicine and Dentistry of New Jersey, New Brunswick, New Jersey*, Immunologically Sensitive Field-Effect Transistors

ZAIDER, MARCO, *Memorial Sloan Kettering Cancer Center, New York, New York*, Prostate Seed Implants

ZAPANTA, CONRAD M., *Penn State College of Medicine, Hershey, Pennsylvania*, Heart, Artificial

ZARDENETA, GUSTAVO, *University of Texas, San Antonio, Texas*, Fluorescence Measurements

ZELMANOVIC, DAVID, *Bayer HealthCare LLC, Tarrytown, New York*, Differential Counts, Automated

ZHANG, MIN, *University of Washington, Seattle, Washington*, Biomaterials: Polymers

ZHANG, YI, *University of Buffalo, Buffalo, New York*, Cell Counters, Blood

ZHU, XIAOYUE, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

ZIAIE, BABAK, *Purdue University, W. Lafayette, Indiana*, Biotelemetry

ZIELINSKI, TODD M., *Medtronic, Inc., Minneapolis, Minnesota*, Bioimpedance in Cardiovascular Medicine

ZIESSMAN, HARVEY A., *Johns Hopkins University*, Computed Tomography, Single Photon Emission

PREFACE

This six-volume work is an alphabetically organized compilation of almost 300 articles that describe critical aspects of medical devices and instrumentation.

It is comprehensive. The articles emphasize the contributions of engineering, physics, and computers to each of the general areas of anesthesiology, biomaterials, burns, cardiology, clinical chemistry, clinical engineering, communicative disorders, computers in medicine, critical care medicine, dermatology, dentistry, ear, nose, and throat, emergency medicine, endocrinology, gastroenterology, genetics, geriatrics, gynecology, hematology, heptology, internal medicine, medical physics, microbiology, nephrology, neurology, nutrition, obstetrics, oncology, ophthalmology, orthopedics, pain, pediatrics, peripheral vascular disease, pharmacology, physical therapy, psychiatry, pulmonary medicine, radiology, rehabilitation, surgery, tissue engineering, transducers, and urology.

The discipline is defined through the synthesis of the core knowledge from all the fields encompassed by the application of engineering, physics, and computers to problems in medicine. The articles focus not only on what is now useful but also on what is likely to be useful in future medical applications.

These volumes answer the question, "What are the branches of medicine and how does technology assist each of them?" rather than "What are the branches of technology and how could each be used in medicine?" To keep this work to a manageable length, the practice of medicine that is unassisted by devices, such as the use of drugs to treat disease, has been excluded.

The articles are accessible to the user; each benefits from brevity of condensation instead of what could easily have been a book-length work. The articles are designed not for peers, but rather for workers from related fields who wish to take a first look at what is important in the subject.

The articles are readable. They do not presume a detailed background in the subject, but are designed for any person with a scientific background and an interest in technology. Rather than attempting to teach the basics of physiology or Ohm's law, the articles build on such basic concepts to show how the worlds of life science and physical science meld to produce improved systems. While the ideal reader might be a person with a Master's degree in biomedical engineering or medical physics or an M.D. with a physical science undergraduate degree, much of the material will be of value to others with an interest in this growing field. High school students and hospital patients can skip over more technical areas and still gain much from the descriptive presentations.

The *Encyclopedia of Medical Devices and Instrumentation* is excellent for browsing and searching for those new divergent associations that may advance work in a peripheral field. While it can be used as a reference for facts, the articles are long enough that they can serve as an educational instrument and provide genuine understanding of a subject.

One can use this work just as one would use a dictionary, since the articles are arranged alphabetically by topic. Cross references assist the reader looking for subjects listed under slightly different names. The index at the end leads the reader to all articles containing pertinent information on any subject. Listed on pages xxi to xxx are all the abbreviations and acronyms used in the *Encyclopedia.* Because of the increasing use of SI units in all branches of science, these units are provided throughout the *Encyclopedia* articles as well as on pages xxxi to xxxv in the section on conversion factors and unit symbols.

I owe a great debt to the many people who have contributed to the creation of this work. At John Wiley & Sons, Encyclopedia Editor George Telecki provided the idea and guiding influence to launch the project. Sean Pidgeon was Editorial Director of the project. Assistant Editors Roseann Zappia, Sarah Harrington, and Surlan Murrell handled the myriad details of communication between publisher, editor, authors, and reviewers and stimulated authors and reviewers to meet necessary deadlines.

My own background has been in the electrical aspects of biomedical engineering. I was delighted to have the assistance of the editorial board to develop a comprehensive encyclopedia. David J. Beebe suggested cellular topics such as microfluidics. Jerry M. Calkins assisted in defining the chemically related subjects, such as anesthesiology. Michael R. Neuman suggested subjects related to sensors, such as in his own work—neonatology. Joon B. Park has written extensively on biomaterials and suggested related subjects. Edward S. Sternick provided many suggestions from medical physics. The Editorial Board was instrumental both in defining the list of subjects and in suggesting authors.

This second edition brings the field up to date. It is available on the web at http://www.mrw.interscience.wiley.com/emdi, where articles can be searched simultaneously to provide rapid and comprehensive information on all aspects of medical devices and instrumentation.

JOHN G. WEBSTER
University of Wisconsin, Madison

LIST OF ARTICLES

ABBREVIATIONS AND ACRONYMS

AAMI	Association for the Advancement of Medical Instrumentation
AAPM	American Association of Physicists in Medicine
ABC	Automatic brightness control
ABET	Accreditation board for engineering training
ABG	Arterial blood gases
ABLB	Alternative binaural loudness balance
ABS	Acrylonitrile–butadiene–styrene
ac	Alternating current
AC	Abdominal circumference; Affinity chromatography
ACA	Automated clinical analyzer
ACES	Augmentative communication evaluation system
ACL	Anterior chamber lens
ACLS	Advanced cardiac life support
ACOG	American College of Obstetrics and Gynecology
ACR	American College of Radiology
ACS	American Cancer Society; American College of Surgeons
A/D	Analog-to-digital
ADC	Agar diffusion chambers; Analog-to-digital converter
ADCC	Antibody-dependent cellular cytotoxicity
ADCL	Accredited Dosimetry Calibration Laboratories
ADP	Adenosine diphosphate
A-D-T	Admission, discharge, and transfer
AE	Anion exchange; Auxiliary electrode
AEA	Articulation error analysis
AEB	Activation energy barrier
AEC	Automatic exposure control
AED	Automatic external defibrillator
AEMB	Alliance for Engineering in Medicine and Biology
AES	Auger electron spectroscopy
AESC	American Engineering Standards Committee
AET	Automatic exposure termination
AFO	Ankle-foot orthosis
AGC	Automatic gain control
AHA	American Heart Association
AI	Arterial insufficiency
AICD	Automatic implantable cardiac defibrillator
AID	Agency for International Development
AIDS	Acquired immune deficiency syndrome
AL	Anterior leaflet
ALG	Antilymphocyte globulin
ALS	Advanced life support; Amyotropic lateral sclerosis
ALT	Alanine aminotransferase
ALU	Arithmetic and logic unit
AM	Amplitude modulation
AMA	American Medical Association
amu	Atomic mass units
ANOVA	Analysis of variance
ANSI	American National Standards Institute
AP	Action potential; Alternative pathway; Anteroposterior
APD	Anterioposterior diameter
APL	Adjustable pressure limiting valve; Applied Physics Laboratory
APR	Anatomically programmed radiography
AR	Amplitude reduction; Aortic regurgitation; Autoregressive
Ara-C	Arabinosylcytosine
ARD	Absorption rate density
ARDS	Adult respiratory distress syndrome
ARGUS	Arrhythmia guard system
ARMA	Autoregressive-moving-average model
ARMAX	Autoregressive-moving-average model with external inputs
AS	Aortic stenosis
ASA	American Standards Association
ASCII	American standard code for information interchange
ASD	Antisiphon device
ASHE	American Society for Hospital Engineering
ASTM	American Society for Testing and Materials
AT	Adenosine-thiamide; Anaerobic threshold; Antithrombin
ATA	Atmosphere absolute
ATLS	Advanced trauma life support
ATN	Acute tubular necrosis
ATP	Adenosine triphosphate
ATPD	Ambient temperature pressure dry
ATPS	Ambient temperature pressure saturated
ATR	Attenuated total reflection
AUC	Area under curve
AUMC	Area under moment curve
AV	Atrioventricular
AZT	Azido thymidine
BA	Biliary atresia
BAEP	Brainstem auditory evoked potential
BAPN	Beta-amino-proprionitryl
BAS	Boston anesthesis system
BASO	Basophil
BB	Buffer base
BBT	Basal body temperature

BCC	Body-centered cubic
BCD	Binary-coded decimal
BCG	Ballistocardiogram
BCLS	Basic cardiac life support
BCRU	British Commitee on Radiation Units and Measurements
BDI	Beck depression inventory
BE	Base excess; Binding energy
BET	Brunauer, Emmett, and Teller methods
BH	His bundle
BI	Biological indicators
BIH	Beth Israel Hospital
BIPM	International Bureau of Weights and Measurements
BJT	Bipolar junction transistor
BMDP	Biomedical Programs
BME	Biomedical engineering
BMET	Biomedical equipment technician
BMO	Biomechanically optimized
BMR	Basal metabolic rate
BOL	Beginning of life
BP	Bereitschafts potential; Break point
BR	Polybutadiene
BRM	Biological response modifier
BRS	Bibliographic retrieval services
BSS	Balanced salt solution
BTG	Beta thromboglobulin
BTPS	Body temperature pressure saturated
BUN	Blood urea nitrogen
BW	Body weight
CA	Conductive adhesives
CABG	Coronary artery by-pass grafting
CAD/CAM	Computer-aided design/computer-aided manufacturing
CAD/D	Computer-aided drafting and design
CADD	Central axis depth dose
CAI	Computer assisted instruction; Computer-aided instruction
CAM	Computer-assisted management
cAMP	Cyclic AMP
CAPD	Continuous ambulatory peritoneal dialysis
CAPP	Child amputee prosthetic project
CAT	Computerized axial tomography
CATS	Computer-assisted teaching system; Computerized aphasia treatment system
CAVH	Continuous arteriovenous hemofiltration
CB	Conjugated bilirubin; Coulomb barrier
CBC	Complete blood count
CBF	Cerebral blood flow
CBM	Computer-based management
CBV	Cerebral blood volume
CC	Closing capacity
CCC	Computer Curriculum Company
CCD	Charge-coupled device
CCE	Capacitance contact electrode
CCF	Cross-correlation function
CCL	Cardiac catheterization laboratory
CCM	Critical care medical services
CCPD	Continuous cycling peritoneal dialysis
CCTV	Closed circuit television system
CCU	Coronary care unit; Critical care unit
CD	Current density
CDR	Complimentary determining region
CDRH	Center for Devices and Radiological Health
CEA	Carcinoembryonic antigen
CF	Conversion factor; Cystic fibrosis
CFC	Continuous flow cytometer
CFR	Code of Federal Regulations
CFU	Colony forming units
CGA	Compressed Gas Association
CGPM	General Conference on Weights and Measures
CHO	Carbohydrate
CHO	Chinese hamster ovary
CI	Combination index
CICU	Cardiac intensive care unit
CIF	Contrast improvement factor
CIN	Cervical intraepithelial neoplasia
CK	Creatine kinase
CLAV	Clavicle
CLSA	Computerized language sample analysis
CM	Cardiomyopathy; Code modulation
CMAD	Computer managed articulation diagnosis
CMI	Computer-managed instruction
CMRR	Common mode rejection ratio
CMV	Conventional mechanical ventilation; Cytomegalovirus
CNS	Central nervous system
CNV	Contingent negative variation
CO	Carbon monoxide; Cardiac output
COBAS	Comprehensive Bio-Analysis System
COPD	Chronic obstructive pulmonary disease
COR	Center of rotation
CP	Cerebral palsy; Closing pressure; Creatine phosphate
CPB	Cardiopulmonary bypass
CPET	Cardiac pacemaker electrode tips
CPM	Computerized probe measurements
CPP	Cerebral perfusion pressure; Cryoprecipitated plasma
CPR	Cardiopulmonary resuscitation
cps	Cycles per second
CPU	Central Processing unit
CR	Center of resistance; Conditioned response; Conductive rubber; Creatinine
CRBB	Complete right bundle branch block
CRD	Completely randomized design
CRL	Crown rump length
CRT	Cathode ray tube
CS	Conditioned stimulus; Contrast scale; Crown seat
CSA	Compressed spectral array
CSF	Cerebrospinal fluid
CSI	Chemical shift imaging
CSM	Chemically sensitive membrane
CT	Computed tomography; Computerized tomography
CTI	Cumulative toxicity response index
CV	Closing volume

C.V.	Coefficient of variation
CVA	Cerebral vascular accident
CVP	Central venous pressure
CVR	Cardiovascular resistance
CW	Continuous wave
CWE	Coated wire electrodes
CWRU	Case Western Reserve University
DAC	Digital-to-analog converter
DAS	Data acquisition system
dB	Decibel
DB	Direct body
DBMS	Data base management system
DBS	Deep brain stimulation
dc	Direct current
DCCT	Diabetes control and complications trial
DCP	Distal cavity pressure
DCS	Dorsal column stimulation
DDC	Deck decompression chamber
DDS	Deep diving system
DE	Dispersive electrode
DEN	Device experience network
DERS	Drug exception ordering system
DES	Diffuse esophageal spasm
d.f.	Distribution function
DHCP	Distributed Hospital Computer Program
DHE	Dihematoporphyrin ether
DHEW	Department of Health Education and Welfare
DHHS	Department of Health and Human Services
DHT	Duration of hypothermia
DI	Deionized water
DIC	Displacement current
DIS	Diagnostic interview schedule
DL	Double layer
DLI	Difference lumen for intensity
DM	Delta modulation
DME	Dropping mercury electrode
DN	Donation number
DNA	Deoxyribonucleic acid
DOF	Degree of freedom
DOS	Drug ordering system
DOT-NHTSA	Department of Transportation Highway Traffic Safety Administration
DPB	Differential pencil beam
DPG	Diphosphoglycerate
DQE	Detection quantum efficiency
DRESS	Depth-resolved surface coil spectroscopy
DRG	Diagnosis-related group
DSA	Digital subtraction angiography
DSAR	Differential scatter-air ratio
DSB	Double strand breaks
DSC	Differential scanning calorimetry
D-T	Deuterium-on-tritium
DTA	Differential thermal analysis
d.u.	Density unit
DUR	Duration
DVT	Deep venous thrombosis
EA	Esophageal accelerogram
EB	Electron beam
EBCDIC	Extended binary code decimal interchange code
EBS	Early burn scar
EBV	Epstein–Barr Virus
EC	Ethyl cellulose
ECC	Emergency cardiac care; Extracorporeal circulation
ECCE	Extracapsular cataract extinction
ECD	Electron capture detector
ECG	Electrocardiogram
ECM	Electrochemical machining
ECMO	Extracorporeal membrane oxygenation
ECOD	Extracranial cerebrovascular occlusive disease
ECRI	Emergency Care Research Institute
ECS	Exner's Comprehensive System
ECT	Electroconvulsive shock therapy; Electroconvulsive therapy; Emission computed tomography
EDD	Estimated date of delivery
EDP	Aortic end diastolic pressure
EDTA	Ethylenediaminetetraacetic acid
EDX	Energy dispersive X-ray analysis
EEG	Electroencephalogram
EEI	Electrode electrolyte interface
EELV	End-expiratory lung volume
EER	Electrically evoked response
EF	Ejection fraction
EF	Electric field; Evoked magnetic fields
EFA	Estimated fetal age
EGF	Epidermal growth factor
EGG	Electrogastrogram
EIA	Enzyme immunoassay
EIU	Electrode impedance unbalance
ELF	Extra low frequency
ELGON	Electrical goniometer
ELISA	Enzyme-linked immunosorbent assay
ELS	Energy loss spectroscopy
ELV	Equivalent lung volume
EM	Electromagnetic
EMBS	Engineering in Medicine and Biology Society
emf	Electromotive force
EMG	Electromyogram
EMGE	Integrated electromyogram
EMI	Electromagnetic interference
EMS	Emergency medical services
EMT	Emergency medical technician
ENT	Ear, nose, and throat
EO	Elbow orthosis
EOG	Electrooculography
EOL	End of life
EOS	Eosinophil
EP	Elastoplastic; Evoked potentiate
EPA	Environmental protection agency
ER	Evoked response
ERCP	Endoscopic retrograde cholangiopancreatography
ERG	Electron radiography; Electroretinogram
ERMF	Event-related magnetic field
ERP	Event-related potential
ERV	Expiratory reserve volume

ESCA	Electron spectroscopy for chemical analysis
ESI	Electrode skin impedance
ESRD	End-stage renal disease
esu	Electrostatic unit
ESU	Electrosurgical unit
ESWL	Extracorporeal shock wave lithotripsy
ETO, Eto	Ethylene oxide
ETT	Exercise tolerance testing
EVA	Ethylene vinyl acetate
EVR	Endocardial viability ratio
EW	Extended wear
FAD	Flavin adenine dinucleotide
FARA	Flexible automation random analysis
FBD	Fetal biparietal diameter
FBS	Fetal bovine serum
fcc	Face centered cubic
FCC	Federal Communications Commission
Fct	Fluorocrit
FDA	Food and Drug Administration
FDCA	Food, Drug, and Cosmetic Act
FE	Finite element
FECG	Fetal electrocardiogram
FEF	Forced expiratory flow
FEL	Free electron lasers
FEM	Finite element method
FEP	Fluorinated ethylene propylene
FES	Functional electrical stimulation
FET	Field-effect transistor
FEV	Forced expiratory volume
FFD	Focal spot to film distance
FFT	Fast Fourier transform
FGF	Fresh gas flow
FHR	Fetal heart rate
FIC	Forced inspiratory capacity
FID	Flame ionization detector; Free-induction decay
FIFO	First-in-first-out
FITC	Fluorescent indicator tagged polymer
FL	Femur length
FM	Frequency modulation
FNS	Functional neuromuscular stimulation
FO	Foramen ovale
FO-CRT	Fiber optics cathode ray tube
FP	Fluorescence polarization
FPA	Fibrinopeptide A
FR	Federal Register
FRC	Federal Radiation Council; Functional residual capacity
FSD	Focus-to-surface distance
FTD	Focal spot to tissue-plane distance
FTIR	Fourier transform infrared
FTMS	Fourier transform mass spectrometer
FU	Fluorouracil
FUDR	Floxuridine
FVC	Forced vital capacity
FWHM	Full width at half maximum
FWTM	Full width at tenth maximum
GABA	Gamma amino buteric acid
GAG	Glycosaminoglycan
GBE	Gas-bearing electrodynamometer
GC	Gas chromatography; Guanine-cytosine
GDT	Gas discharge tube
GFR	Glomerular filtration rate
GHb	Glycosylated hemoglobin
GI	Gastrointestinal
GLC	Gas–liquid chromatography
GMV	General minimum variance
GNP	Gross national product
GPC	Giant papillary conjunctivitis
GPH	Gas-permeable hard
GPH-EW	Gas-permeable hard lens extended wear
GPO	Government Printing Office
GSC	Gas-solid chromatography
GSR	Galvanic skin response
GSWD	Generalized spike-wave discharge
HA	Hydroxyapatite
HAM	Helical axis of motion
Hb	Hemoglobin
HBE	His bundle electrogram
HBO	Hyperbaric oxygenation
HC	Head circumference
HCA	Hypothermic circulatory arrest
HCFA	Health care financing administration
HCL	Harvard Cyclotron Laboratory
hcp	Hexagonal close-packed
HCP	Half cell potential
HDPE	High density polyethylene
HECS	Hospital Equipment Control System
HEMS	Hospital Engineering Management System
HEPA	High efficiency particulate air filter
HES	Hydroxyethylstarch
HETP	Height equivalent to a theoretical plate
HF	High-frequency; Heating factor
HFCWO	High-frequency chest wall oscillation
HFER	High-frequency electromagnetic radiation
HFJV	High-frequency jet ventilation
HFO	High-frequency oscillator
HFOV	High-frequency oscillatory ventilation
HFPPV	High-frequency positive pressure ventilation
HFV	High-frequency ventilation
HHS	Department of Health and Human Services
HIBC	Health industry bar code
HIMA	Health Industry Manufacturers Association
HIP	Hydrostatic indifference point
HIS	Hospital information system
HK	Hexokinase
HL	Hearing level
HMBA	Hexamethylene bisacetamide
HMO	Health maintenance organization
HMWPE	High-molecular-weight polyethylene
HOL	Higher-order languages
HP	Heating factor; His-Purkinje
HpD	Hematoporphyrin derivative
HPLC	High-performance liquid chromatography
HPNS	High-pressure neurological syndrome
HPS	His-Purkinje system
HPX	High peroxidase activity

HR	Heart rate; High-resolution
HRNB	Halstead-Reitan Neuropsychological Battery
H/S	Hard/soft
HSA	Human serum albumin
HSG	Hysterosalpingogram
HTCA	Human tumor cloning assay
HTLV	Human T cell lymphotrophic virus
HU	Heat unit; Houndsfield units; Hydroxyurea
HVL	Half value layer
HVR	Hypoxic ventilatory response
HVT	Half-value thickness
IA	Image intensifier assembly; Inominate artery
IABP	Intraaortic balloon pumping
IAEA	International Atomic Energy Agency
IAIMS	Integrated Academic Information Management System
IASP	International Association for the Study of Pain
IC	Inspiratory capacity; Integrated circuit
ICCE	Intracapsular cataract extraction
ICD	Intracervical device
ICDA	International classification of diagnoses
ICL	Ms-clip lens
ICP	Inductively coupled plasma; Intracranial pressure
ICPA	Intracranial pressure amplitude
ICRP	International Commission on Radiological Protection
ICRU	International Commission on Radiological Units and Measurements
ICU	Intensive care unit
ID	Inside diameter
IDDM	Insulin dependent diabetes mellitus
IDE	Investigational device exemption
IDI	Index of inspired gas distribution
I:E	Inspiratory: expiratory
IEC	International Electrotechnical Commission; Ion-exchange chromatography
IEEE	Institute of Electrical and Electronics Engineers
IEP	Individual educational program
BETS	Inelastic electron tunneling spectroscopy
IF	Immunofluorescent
IFIP	International Federation for Information Processing
IFMBE	International Federation for Medical and Biological Engineering
IGFET	Insulated-gate field-effect transistor
IgG	Immunoglobulin G
IgM	Immunoglobulin M
IHP	Inner Helmholtz plane
IHSS	Idiopathic hypertrophic subaortic stenosis
II	Image intensifier
IIIES	Image intensifier input-exposure sensitivity
IM	Intramuscular
IMFET	Immunologically sensitive field-effect transistor
IMIA	International Medical Informatics Association
IMS	Information management system
IMV	Intermittent mandatory ventilation
INF	Interferon
IOL	Intraocular lens
IPC	Ion-pair chromatography
IPD	Intermittent peritoneal dialysis
IPG	Impedance plethysmography
IPI	Interpulse interval
IPPB	Intermittent positive pressure breathing
IPTS	International practical temperature scale
IR	Polyisoprene rubber
IRB	Institutional Review Board
IRBBB	Incomplete right bundle branch block
IRPA	International Radiation Protection Association
IRRAS	Infrared reflection-absorption spectroscopy
IRRS	Infrared reflection spectroscopy
IRS	Internal reflection spectroscopy
IRV	Inspiratory reserve capacity
IS	Image size; Ion-selective
ISC	Infant skin servo control
ISDA	Instantaneous screw displacement axis
ISE	Ion-selective electrode
ISFET	Ion-sensitive field effect transistor
ISIT	Intensified silicon-intensified target tube
ISO	International Organization for Standardization
ISS	Ion scattering spectroscopy
IT	Intrathecal
ITEP	Institute of Theoretical and Experimental Physics
ITEPI	Instantaneous trailing edge pulse impedance
ITLC	Instant thin-layer chromatography
IUD	Intrauterine device
IV	Intravenous
IVC	Inferior vena cava
IVP	Intraventricular pressure
JCAH	Joint Commission on the Accreditation of Hospitals
JND	Just noticeable difference
JRP	Joint replacement prosthesis
KB	Kent bundle
Kerma	Kinetic energy released in unit mass
KO	Knee orthosis
KPM	Kilopond meter
KRPB	Krebs-Ringer physiological buffer
LA	Left arm; Left atrium
LAD	Left anterior descending; Left axis deviation
LAE	Left atrial enlargement
LAK	Lymphokine activated killer
LAL	Limulus amoebocyte lysate
LAN	Local area network
LAP	Left atrial pressure
LAT	Left anterior temporalis
LBBB	Left bundle branch block
LC	Left carotid; Liquid chromatography

LCC	Left coronary cusp
LCD	Liquid crystal display
LDA	Laser Doppler anemometry
LDF	Laser Doppler flowmetry
LDH	Lactate dehydrogenase
LDPE	Low density polyethylene
LEBS	Low-energy brief stimulus
LED	Light-emitting diode
LEED	Low energy electron diffraction
LES	Lower esophageal sphincter
LESP	Lower esophageal sphincter pressure
LET	Linear energy transfer
LF	Low frequency
LH	Luteinizing hormone
LHT	Local hyperthermia
LL	Left leg
LLDPE	Linear low density polyethylene
LLPC	Liquid-liquid partition chromatography
LLW	Low-level waste
LM	Left masseter
LNNB	Luria-Nebraska Neuropsychological Battery
LOS	Length of stay
LP	Late potential; Lumboperitoneal
LPA	Left pulmonary artery
LPC	Linear predictive coding
LPT	Left posterior temporalis
LPV	Left pulmonary veins
LRP	Late receptor potential
LS	Left subclavian
LSC	Liquid-solid adsorption chromatography
LSI	Large scale integrated
LSV	Low-amplitude shear-wave viscoelastometry
LTI	Low temperature isotropic
LUC	Large unstained cells
LV	Left ventricle
LVAD	Left ventricular assist device
LVDT	Linear variable differential transformer
LVEP	Left ventricular ejection period
LVET	Left ventricular ejection time
LVH	Left ventricular hypertrophy
LYMPH	Lymphocyte
MAA	Macroaggregated albumin
MAC	Minimal auditory capabilities
MAN	Manubrium
MAP	Mean airway pressure; Mean arterial pressure
MAST	Military assistance to safety and traffic
MBA	Monoclonal antibody
MBV	Maximum breathing ventilation
MBX	Monitoring branch exchange
MCA	Methyl cryanoacrylate
MCG	Magnetocardiogram
MCI	Motion Control Incorporated
MCMI	Millon Clinical Multiaxial Inventory
MCT	Microcatheter transducer
MCV	Mean corpuscular volume
MDC	Medical diagnostic categories
MDI	Diphenylmethane diisocyanate; Medical Database Informatics
MDP	Mean diastolic aortic pressure
MDR	Medical device reporting
MDS	Multidimensional scaling
ME	Myoelectric
MED	Minimum erythema dose
MEDPAR	Medicare provider analysis and review
MEFV	Maximal expiratory flow volume
MEG	Magnetoencephalography
MeSH	Medline subject heading
METS	Metabolic equivalents
MF	Melamine-formaldehyde
MFP	Magnetic field potential
MGH	Massachusetts General Hospital
MHV	Magnetic heart vector
MI	Myocardial infarction
MIC	Minimum inhibitory concentration
MIFR	Maximum inspiratory flow rate
MINET	Medical Information Network
MIR	Mercury-in-rubber
MIS	Medical information system; Metal-insulator-semiconductor
MIT	Massachusetts Institute of Technology
MIT/BIH	Massachusetts Institute of Technology/ Beth Israel Hospital
MMA	Manual metal arc welding
MMA	Methyl methacrylate
MMECT	Multiple-monitored ECT
MMFR	Maximum midexpiratory flow rate
mm Hg	Millimeters of mercury
MMPI	Minnesota Multiphasic Personality Inventory
MMSE	Minimum mean square error
MO	Membrane oxygenation
MONO	Monocyte
MOSFET	Metal oxide silicon field-effect transistor
MP	Mercaptopurine; Metacarpal-phalangeal
MPD	Maximal permissible dose
MR	Magnetic resonance
MRG	Magnetoretinogram
MRI	Magnetic resonance imaging
MRS	Magnetic resonance spectroscopy
MRT	Mean residence time
MS	Mild steel; Multiple sclerosis
MSR	Magnetically shielded room
MTBF	Mean time between failure
MTF	Modulation transfer function
MTTR	Mean time to repair
MTX	Methotroxate
MUA	Motor unit activity
MUAP	Motor unit action potential
MUAPT	Motor unit action potential train
MUMPI	Missouri University Multi-Plane Imager
MUMPS	Massachusetts General Hospital utility multiuser programming system
MV	Mitral valve
MVO_2	Maximal oxygen uptake
MVTR	Moisture vapor transmission rate
MVV	Maximum voluntary ventilation
MW	Molecular weight

NAA	Neutron activation analysis
NAD	Nicotinamide adenine dinucleotide
NADH	Nicotinamide adenine dinucleotide, reduced form
NADP	Nicotinamide adenine dinucleotide phosphate
NAF	Neutrophil activating factor
NARM	Naturally occurring and accelerator-produced radioactive materials
NBB	Normal buffer base
NBD	Neuromuscular blocking drugs
N-BPC	Normal bonded phase chromatography
NBS	National Bureau of Standards
NCC	Noncoronary cusp
NCCLS	National Committee for Clinical Laboratory Standards; National Committee on Clinical Laboratory Standards
NCRP	National Council on Radiation Protection
NCT	Neutron capture theory
NEEP	Negative end-expiratory pressure
NEMA	National Electrical Manufacturers Association
NEMR	Nonionizing electromagnetic radiation
NEQ	Noise equivalent quanta
NET	Norethisterone
NEUT	Neutrophil
NFPA	National Fire Protection Association
NH	Neonatal hepatitis
NHE	Normal hydrogen electrode
NHLBI	National Heart, Lung, and Blood Institute
NIR	Nonionizing radiation
NIRS	National Institute for Radiologic Science
NK	Natural killer
NMJ	Neuromuscular junction
NMOS	N-type metal oxide silicon
NMR	Nuclear magnetic resonance
NMS	Neuromuscular stimulation
NPH	Normal pressure hydrocephalus
NPL	National Physical Laboratory
NR	Natural rubber
NRC	Nuclear Regulatory Commission
NRZ	Non-return-to-zero
NTC	Negative temperature coefficient
NTIS	National Technical Information Service
NVT	Neutrons versus time
NYHA	New York Heart Association
ob/gyn	Obstetrics and gynecology
OCR	Off-center ratio; Optical character recognition
OCV	Open circuit voltage
OD	Optical density; Outside diameter
ODC	Oxyhemoglobin dissociation curve
ODT	Oxygen delivery truck
ODU	Optical density unit
OER	Oxygen enhancement ratio
OFD	Object to film distance; Occiputo-frontal diameter
OHL	Outer Helmholtz layer
OHP	Outer Helmholtz plane
OIH	Orthoiodohippurate
OPG	Ocular pneumoplethysmography
OR	Operating room
OS	Object of known size; Operating system
OTC	Over the counter
OV	Offset voltage
PA	Posterioanterior; Pulmonary artery; Pulse amplitude
PACS	Picture archiving and communications systems
PAD	Primary afferent depolarization
PAM	Pulse amplitude modulation
PAN	Polyacrylonitrile
PAP	Pulmonary artery pressure
PAR	Photoactivation ratio
PARFR	Program for Applied Research on Fertility Regulation
PARR	Poetanesthesia recovery room
PAS	Photoacoustic spectroscopy
PASG	Pneumatic antishock garment
PBI	Penile brachial index
PBL	Positive beam limitation
PBT	Polybutylene terephthalate
PC	Paper chromatography; Personal computer; Polycarbonate
PCA	Patient controlled analgesia; Principal components factor analysis
PCG	Phonocardiogram
PCI	Physiological cost index
PCL	Polycaprolactone; Posterior chamber lens
PCR	Percent regurgitation
PCRC	Perinatal Clinical Research Center
PCS	Patient care system
PCT	Porphyria cutanea tarda
PCWP	Pulmonary capillary wedge pressure
PD	Peritoneal dialysis; Poly-p-dioxanone; Potential difference; Proportional and derivative
PDD	Percent depth dose; Perinatal Data Directory
PDE	Pregelled disposable electrodes
p.d.f.	Probability density function
PDL	Periodontal ligament
PDM	Pulse duration modulation
PDMSX	Polydimethyl siloxane
PDS	Polydioxanone
PE	Polyethylene
PEEP	Positive end-expiratory pressure
PEFR	Peak expiratory now rate
PEN	Parenteral and enteral nutrition
PEP	Preejection period
PEPPER	Programs examine phonetic find phonological evaluation records
PET	Polyethylene terephthalate; Positron-emission tomography
PEU	Polyetherurethane
PF	Platelet factor
PFA	Phosphonoformic add
PFC	Petrofluorochemical
PFT	Pulmonary function testing
PG	Polyglycolide; Propylene glycol

PGA	Polyglycolic add
PHA	Phytohemagglutinin; Pulse-height analyzer
PHEMA	Poly-2-hydroxyethyl methacrylate
PI	Propidium iodide
PID	Pelvic inflammatory disease; Proportional/integral/derivative
PIP	Peak inspiratory pressure
PL	Posterior leaflet
PLA	Polylactic acid
PLATO	Program Logic for Automated Teaching Operations
PLD	Potentially lethal damage
PLED	Periodic latoralized epileptiform discharge
PLT	Platelet
PM	Papillary muscles; Preventive maintenance
PMA	Polymethyl acrylate
p.m.f.	Probability mass function
PMMA	Polymethyl methacrylate
PMOS	P-type metal oxide silicon
PMP	Patient management problem; Poly(4-methylpentane)
PMT	Photomultiplier tube
PO	Per os
Po_2	Partial pressure of oxygen
POBT	Polyoxybutylene terephthalate
POM	Polyoxymethylene
POMC	Patient order management and communication system
POPRAS	Problem Oriented Perinatal Risk Assessment System
PP	Perfusion pressure; Polyproplyene; Postprandial (after meals)
PPA	Phonemic process analysis
PPF	Plasma protein fraction
PPM	Pulse position modulation
PPSFH	Polymerized phyridoxalated stroma-free hemoglobin
PR	Pattern recognition; Pulse rate
PRBS	Pseudo-random binary signals
PRP	Pulse repetition frequency
PRO	Professional review organization
PROM	Programmable read only memory
PS	Polystyrene
PSA	Pressure-sensitive adhesive
PSF	Point spread function
PSI	Primary skin irritation
PSP	Postsynaptic potential
PSR	Proton spin resonance
PSS	Progressive systemic sclerosis
PT	Plasma thromboplastin
PTB	Patellar tendon bearing orthosis
PTC	Plasma thromboplastin component; Positive temperature coefficient; Pressurized personal transfer capsule
PTCA	Percutaneous transluminal coronary angioplasty
PTFE	Polytetrafluoroethylene
PTT	Partial thromboplastin time
PUL	Percutaneous ultrasonic lithotripsy
PURA	Prolonged ultraviolet-A radiation
PUVA	Psoralens and longwave ultraviolet light photochemotherapy
P/V	Pressure/volume
PVC	Polyvinyl chloride; Premature ventricular contraction
PVI	Pressure–volume index
PW	Pulse wave; Pulse width
PWM	Pulse width modulation
PXE	Pseudo-xanthoma elasticum
QA	Quality assurance
QC	Quality control
R-BPC	Reverse bonded phase chromatography
R/S	Radiopaque-spherical
RA	Respiratory amplitude; Right arm
RAD	Right axis deviation
RAE	Right atrial enlargement
RAM	Random access memory
RAP	Right atrial pressure
RAT	Right anterior temporalis
RB	Right bundle
RBBB	Right bundle branch block
RBC	Red blood cell
RBE	Relative biologic effectiveness
RBF	Rose bengal fecal excretion
RBI	Resting baseline impedance
RCBD	Randomized complete block diagram
rCBF	Regional cerebral blood flow
RCC	Right coronary cusp
RCE	Resistive contact electrode
R&D	Research and development
r.e.	Random experiment
RE	Reference electrode
REM	Rapid eye movement; Return electrode monitor
REMATE	Remote access and telecommunication system
RES	Reticuloendothelial system
RESNA	Rehabilitation Engineering Society of North America
RF	Radio frequency; Radiographic-nuoroscopic
RFI	Radio-frequency interference
RFP	Request for proposal
RFQ	Request for quotation
RH	Relative humidity
RHE	Reversible hydrogen electrode
RIA	Radioimmunoassay
RM	Repetition maximum; Right masseter
RMR	Resting metabolic rate
RMS	Root mean square
RN	Radionuclide
RNCA	Radionuclide cineagiogram
ROI	Regions of interest
ROM	Range of motion; Read only memory
RP	Retinitis pigmentosa
RPA	Right pulmonary artery
RPP	Rate pressure product
RPT	Rapid pull-through technique
RPV	Right pulmonary veins
RQ	Respiratory quotient

RR	Recovery room
RRT	Recovery room time; Right posterior temporalis
RT	Reaction time
RTD	Resistance temperature device
RTT	Revised token test
r.v.	Random variable
RV	Residual volume; Right ventricle
RVH	Right ventricular hypertrophy
RVOT	Right ventricular outflow tract
RZ	Return-to-zero
SA	Sinoatrial; Specific absorption
SACH	Solid-ankle-cushion-heel
SAD	Source-axis distance; Statistical Analysis System
SAINT	System analysis of integrated network of tasks
SAL	Sterility assurance level; Surface averaged lead
SALT	Systematic analysis of language transcripts
SAMI	Socially acceptable monitoring instrument
SAP	Systemic arterial pressure
SAR	Scatter-air ratio; Specific absorption rate
SARA	System for anesthetic and respiratory gas analysis
SBE	Subbacterial endocarditis
SBR	Styrene-butadiene rubbers
SC	Stratum corneum; Subcommittees
SCAP	Right scapula
SCE	Saturated calomel electrode; Sister chromatid exchange
SCI	Spinal cord injury
SCRAD	Sub-Committee on Radiation Dosimetry
SCS	Spinal cord stimulation
SCUBA	Self-contained underwater breathing apparatus
SD	Standard deviation
SDA	Stepwise discriminant analysis
SDS	Sodium dodecyl sulfate
S&E	Safety and effectiveness
SE	Standard error
SEC	Size exclusion chromatography
SEM	Scanning electron microscope; Standard error of the mean
SEP	Somatosensory evoked potential
SEXAFS	Surface extended X-ray absorption fine structure
SF	Surviving fraction
SFD	Source-film distance
SFH	Stroma-free hemoglobin
SFTR	Sagittal frontal transverse rotational
SG	Silica gel
SGF	Silica gel fraction
SGG	Spark gap generator
SGOT	Serum glutamic oxaloacetic transaminase
SGP	Strain gage plethysmography; Stress-generated potential
SHE	Standard hydrogen electrode
SI	Le Système International d'Unités
SEBS	Surgical isolation barrier system
SID	Source to image reception distance
SIMFU	Scanned intensity modulated focused ultrasound
SIMS	Secondary ion mass spectroscopy; System for isometric muscle strength
SISI	Short increment sensitivity index
SL	Surgical lithotomy
SLD	Sublethal damage
SLE	Systemic lupus erythemotodes
SMA	Sequential multiple analyzer
SMAC	Sequential multiple analyzer with computer
SMR	Sensorimotor
S/N	Signal-to-noise
S:N/D	Signal-to-noise ratio per unit dose
SNP	Sodium nitroprusside
SNR	Signal-to-noise ratio
SOA	Sources of artifact
SOAP	Subjective, objective, assessment, plan
SOBP	Spread-out Bragg peak
SP	Skin potential
SPECT	Single photon emission computed tomography
SPL	Sound pressure level
SPRINT	Single photon ring tomograph
SPRT	Standard platinum resistance thermometer
SPSS	Statistical Package for the Social Sciences
SQUID	Superconducting quantum interference device
SQV	Square wave voltammetry
SR	Polysulfide rubbers
SRT	Speech reception threshold
SS	Stainless steel
SSB	Single strand breaks
SSD	Source-to-skin distance; Source-to-surface distance
SSE	Stainless steel electrode
SSEP	Somatosensory evoked potential
SSG	Solid state generator
SSP	Skin stretch potential
SSS	Sick sinus syndrome
STD	Source-tray distance
STI	Systolic time intervals
STP	Standard temperature and pressure
STPD	Standard temperature pressure dry
SV	Stroke volume
SVC	Superior vena cava
SW	Standing wave
TAA	Tumor-associated antigens
TAC	Time-averaged concentration
TAD	Transverse abdominal diameter
TAG	Technical Advisory Group
TAH	Total artificial heart
TAR	Tissue-air ratio
TC	Technical Committees
TCA	Tricarboxylic acid cycle
TCD	Thermal conductivity detector
TCES	Transcutaneous cranial electrical stimulation

TCP	Tricalcium phosphate
TDD	Telecommunication devices for the deaf
TDM	Therapeutic drug monitoring
TE	Test electrode; Thermoplastic elastomers
TEAM	Technology evaluation and acquisition methods
TEM	Transmission electron microscope; Transverse electric and magnetic mode; Transverse electromagnetic mode
TENS	Transcutaneous electrical nerve stimulation
TEP	Tracheoesophageal puncture
TEPA	Triethylenepho-sphoramide
TF	Transmission factor
TFE	Tetrafluorethylene
TI	Totally implantable
TICCIT	Time-shared Interaction Computer-Controlled Information Television
TLC	Thin-layer chromatography; Total lung capacity
TLD	Thermoluminescent dosimetry
TMJ	Temporomandibular joint
TMR	Tissue maximum ratio; Topical magnetic resonance
TNF	Tumor necrosis factor
TOF	Train-of-four
TP	Thermal performance
TPC	Temperature pressure correction
TPD	Triphasic dissociation
TPG	Transvalvular pressure gradient
TPN	Total parenteral nutrition
TR	Temperature rise
tRNA	Transfer RNA
TSH	Thyroid stimulating hormone
TSS	Toxic shock syndrome
TTD	Telephone devices for the deaf
TTI	Tension time index
TTR	Transition temperature range
TTV	Trimming tip version
TTY	Teletypewriter
TUR	Transurethral resection
TURP	Transurethral resections of the prostrate
TV	Television; Tidal volume; Tricuspid valve
TVER	Transscleral visual evoked response
TW	Traveling wave
TxB_2	Thrombozame B^2
TZ	Transformation zone
UES	Upper esophageal sphincter
UP	Urea-formaldehyde
UfflS	University Hospital Information System
UHMW	Ultra high molecular weight
UHMWPE	Ultra high molecular weight polyethylene
UL	Underwriters Laboratory
ULF	Ultralow frequency
ULTI	Ultralow temperature isotropic
UMN	Upper motor neuron
UO	Urinary output
UPTD	Unit pulmonary oxygen toxicity doses
UR	Unconditioned response
US	Ultrasound; Unconditioned stimulus
USNC	United States National Committee
USP	United States Pharmacopeia
UTS	Ultimate tensile strength
UV	Ultraviolet; Umbilical vessel
UVR	Ultraviolet radiation
V/F	Voltage-to-frequency
VA	Veterans Administration
VAS	Visual analog scale
VBA	Vaginal blood volume in arousal
VC	Vital capacity
VCO	Voltage-controlled oscillator
VDT	Video display terminal
VECG	Vectorelectrocardiography
VEP	Visually evoked potential
VF	Ventricular fibrillation
VOP	Venous occlusion plethysmography
VP	Ventriculoperitoneal
VPA	Vaginal pressure pulse in arousal
VPB	Ventricular premature beat
VPR	Volume pressure response
VSD	Ventricular septal defect
VSWR	Voltage standing wave ratio
VT	Ventricular tachycardia
VTG	Vacuum tube generator
VTS	Viewscan text system
VV	Variable version
WAIS-R	Weschler Adult Intelligence Scale-Revised
WAK	Wearable artificial kidney
WAML	Wide-angle mobility light
WBAR	Whole-body autoradiography
WBC	White blood cell
WG	Working Groups
WHO	World Health Organization; Wrist hand orthosis
WLF	Williams-Landel-Ferry
WMR	Work metabolic rate
w/o	Weight percent
WORM	Write once, read many
WPW	Wolff-Parkinson-White
XPS	X-ray photon spectroscopy
XR	Xeroradiograph
YAG	Yttrium aluminum garnet
ZPL	Zero pressure level

CONVERSION FACTORS AND UNIT SYMBOLS

SI UNITS (ADOPTED 1960)

A new system of metric measurement, the International System of Units (abbreviated SI), is being implemented throughout the world. This system is a modernized version of the MKSA (meter, kilogram, second, ampere) system, and its details are published and controlled by an international treaty organization (The International Bureau of Weights and Measures).

SI units are divided into three classes:

Base Units

length	meter[†] (m)
mass[‡]	kilogram (kg)
time	second (s)
electric current	ampere (A)
thermodynamic temperature§	kelvin (K)
amount of substance	mole (mol)
luminous intensity	candela (cd)

Supplementary Units

plane angle	radian (rad)
solid angle	steradian (sr)

Derived Units and Other Acceptable Units

These units are formed by combining base units, supplementary units, and other derived units. Those derived units having special names and symbols are marked with an asterisk (*) in the list below:

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
*absorbed dose	gray	Gy	J/kg
acceleration	meter per second squared	m/s^2	
*activity (of ionizing radiation source)	becquerel	Bq	1/s
area	square kilometer	km^2	
	square hectometer	hm^2	ha (hectare)
	square meter	m^2	

[†]The spellings "metre" and "litre" are preferred by American Society for Testing and Materials (ASTM); however, "−er" will be used in the Encyclopedia.

[‡]"Weight" is the commonly used term for "mass."

§Wide use is made of "Celsius temperature" (t) defined $t = T - T_0$ where T is the thermodynamic temperature, expressed in kelvins, and $T_0 = 273.15\,K$ by definition. A temperature interval may be expressed in degrees Celsius as well as in kelvins.

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
*capacitance	farad	F	C/V
concentration (of amount of substance)	mole per cubic meter	mol/m^3	
*conductance	siemens	S	A/V
current density	ampere per square meter	A/m^2	
density, mass density	kilogram per cubic meter	kg/m^3	g/L; mg/cm^3
dipole moment (quantity)	coulomb meter	C·m	
*electric charge, quantity of electricity	coulomb	C	A·s
electric charge density	coulomb per cubic meter	C/m^3	
electric field strength	volt per meter	V/m	
electric flux density	coulomb per square meter	C/m^2	
*electric potential, potential difference, electromotive force	volt	V	W/A
*electric resistance	ohm	Ω	V/A
*energy, work, quantity of heat	megajoule	MJ	
	kilojoule	kJ	
	joule	J	N·m
	electron volt†	eV†	
	kilowatt hour†	kW·h†	
energy density	joule per cubic meter	J/m^3	
*force	kilonewton	kN	
	newton	N	$kg·m/s^2$
*frequency	megahertz	MHz	
	hertz	Hz	1/s
heat capacity, entropy	joule per kelvin	J/K	
heat capacity (specific), specific entropy	joule per kilogram kelvin	J/(kg·K)	
heat transfer coefficient	watt per square meter kelvin	$W/(m^2·K)$	
*illuminance	lux	lx	lm/m^2
*inductance	henry	H	Wb/A
linear density	kilogram per meter	kg/m	
luminance	candela per square meter	cd/m^2	
*luminous flux	lumen	lm	cd·sr
magnetic field strength	ampere per meter	A/m	
*magnetic flux	weber	Wb	V·s
*magnetic flux density	tesla	T	Wb/m^2
molar energy	joule per mole	J/mol	
molar entropy, molar heat capacity	joule per mole kelvin	J/(mol·K)	
moment of force, torque	newton meter	N·m	
momentum	kilogram meter per second	kg·m/s	
permeability	henry per meter	H/m	
permittivity	farad per meter	F/m	
*power, heat flow rate, radiant flux	kilowatt	kW	
	watt	W	J/s
power density, heat flux density, irradiance	watt per square meter	W/m^2	
*pressure, stress	megapascal	MPa	
	kilopascal	kPa	
	pascal	Pa	N/m^2
sound level	decibel	dB	
specific energy	joule per kilogram	J/kg	
specific volume	cubic meter per kilogram	m^3/kg	
surface tension	newton per meter	N/m	
thermal conductivity	watt per meter kelvin	W/(m·K)	
velocity	meter per second	m/s	
	kilometer per hour	km/h	
viscosity, dynamic	pascal second	Pa·s	
	millipascal second	mPa·s	

†This non-SI unit is recognized as having to be retained because of practical importance or use in specialized fields.

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
viscosity, kinematic	square meter per second	m^2/s	
	square millimeter per second	mm^2/s	
	cubic meter	m^3	
	cubic decimeter	dm^3	L(liter)
	cubic centimeter	cm^3	mL
wave number	1 per meter	m^{-1}	
	1 per centimeter	cm^{-1}	

In addition, there are 16 prefixes used to indicate order of magnitude, as follows:

Multiplication factor	*Prefix*	*Symbol*	*Note*
10^{18}	exa	E	
10^{15}	peta	P	
10^{12}	tera	T	
10^{9}	giga	G	
10^{8}	mega	M	
10^{3}	kilo	k	
10^{2}	hecto	h[a]	[a]Although hecto, deka, deci, and centi are
10	deka	da[a]	SI prefixes, their use should be avoided
10^{-1}	deci	d[a]	except for SI unit-multiples for area and
10^{-2}	centi	c[a]	volume and nontechnical use of
10^{-3}	milli	m	centimeter, as for body and clothing
10^{-6}	micro	μ	measurement.
10^{-9}	nano	n	
10^{-12}	pico	p	
10^{-15}	femto	f	
10^{-18}	atto	a	

For a complete description of SI and its use the reader is referred to ASTM E 380.

CONVERSION FACTORS TO SI UNITS

A representative list of conversion factors from non-SI to SI units is presented herewith. Factors are given to four significant figures. Exact relationships are followed by a dagger (†). A more complete list is given in ASTM E 380-76 and ANSI Z210. 1-1976.

To convert from	*To*	*Multiply by*
acre	square meter (m^2)	4.047×10^{3}
angstrom	meter (m)	$1.0 \times 10^{-10\dagger}$
are	square meter (m^2)	$1.0 \times 10^{2\dagger}$
astronomical unit	meter (m)	1.496×10^{11}
atmosphere	pascal (Pa)	1.013×10^{5}
bar	pascal (Pa)	$1.0 \times 10^{5\dagger}$
barrel (42 U.S. liquid gallons)	cubic meter (m^3)	0.1590
Btu (International Table)	joule (J)	1.055×10^{3}
Btu (mean)	joule (J)	1.056×10^{3}
Bt (thermochemical)	joule (J)	1.054×10^{3}
bushel	cubic meter (m^3)	3.524×10^{-2}
calorie (International Table)	joule (J)	4.187
calorie (mean)	joule (J)	4.190
calorie (thermochemical)	joule (J)	$4.184^{\dagger}$
centimeters of water (39.2 °F)	pascal (Pa)	98.07
centipoise	pascal second (Pa·s)	$1.0 \times 10^{-3\dagger}$
centistokes	square millimeter per second (mm^2/s)	$1.0^{\dagger}$

To convert from	*To*	*Multiply by*
cfm (cubic foot per minute)	cubic meter per second (m^3/s)	4.72×10^{-4}
cubic inch	cubic meter (m^3)	1.639×10^{-4}
cubic foot	cubic meter (m^3)	2.832×10^{-2}
cubic yard	cubic meter (m^3)	0.7646
curie	becquerel (Bq)	$3.70 \times 10^{10\dagger}$
debye	coulomb-meter (C·m)	3.336×10^{-30}
degree (angle)	radian (rad)	1.745×10^{-2}
denier (international)	kilogram per meter (kg/m)	1.111×10^{-7}
	tex	0.1111
dram (apothecaries')	kilogram (kg)	3.888×10^{-3}
dram (avoirdupois)	kilogram (kg)	1.772×10^{-3}
dram (U.S. fluid)	cubic meter (m^3)	3.697×10^{-6}
dyne	newton(N)	$1.0 \times 10^{-6\dagger}$
dyne/cm	newton per meter (N/m)	$1.00 \times 10^{-3\dagger}$
electron volt	joule (J)	1.602×10^{-19}
erg	joule (J)	$1.0 \times 10^{-7\dagger}$
fathom	meter (m)	1.829
fluid ounce (U.S.)	cubic meter (m^3)	2.957×10^{-5}
foot	meter (m)	$0.3048^{\dagger}$
foot-pound force	joule (J)	1.356
foot-pound force	newton meter (N·m)	1.356
foot-pound force per second	watt(W)	1.356
footcandle	lux (lx)	10.76
furlong	meter (m)	2.012×10^2
gal	meter per second squared (m/s^2)	$1.0 \times 10^{-2\dagger}$
gallon (U.S. dry)	cubic meter (m^3)	4.405×10^{-3}
gallon (U.S. liquid)	cubic meter (m^3)	3.785×10^{-3}
gilbert	ampere (A)	0.7958
gill (U.S.)	cubic meter (m^3)	1.183×10^{-4}
grad	radian	1.571×10^{-2}
grain	kilogram (kg)	6.480×10^{-5}
gram force per denier	newton per tex (N/tex)	8.826×10^{-2}
hectare	square meter (m^2)	$1.0 \times 10^{4\dagger}$
horsepower (550 ft·lbf/s)	watt(W)	7.457×10^2
horsepower (boiler)	watt(W)	9.810×10^3
horsepower (electric)	watt(W)	$7.46 \times 10^{2\dagger}$
hundredweight (long)	kilogram (kg)	50.80
hundredweight (short)	kilogram (kg)	45.36
inch	meter (m)	$2.54 \times 10^{-2\dagger}$
inch of mercury (32 °F)	pascal (Pa)	3.386×10^3
inch of water (39.2 °F)	pascal (Pa)	2.491×10^2
kilogram force	newton (N)	9.807
kilopond	newton (N)	9.807
kilopond-meter	newton-meter (N·m)	9.807
kilopond-meter per second	watt (W)	9.807
kilopond-meter per min	watt(W)	0.1635
kilowatt hour	megajoule (MJ)	$3.6^{\dagger}$
kip	newton (N)	4.448×10^2
knot international	meter per second (m/s)	0.5144
lambert	candela per square meter (cd/m^2)	3.183×10^3
league (British nautical)	meter (m)	5.559×10^2
league (statute)	meter (m)	4.828×10^3
light year	meter (m)	9.461×10^{15}
liter (for fluids only)	cubic meter (m^3)	$1.0 \times 10^{-3\dagger}$
maxwell	weber (Wb)	$1.0 \times 10^{-8\dagger}$
micron	meter (m)	$1.0 \times 10^{-6\dagger}$
mil	meter (m)	$2.54 \times 10^{-5\dagger}$
mile (U.S. nautical)	meter (m)	$1.852 \times 10^{3\dagger}$
mile (statute)	meter (m)	1.609×10^3
mile per hour	meter per second (m/s)	0.4470

To convert from	*To*	*Multiply by*
millibar	pascal (Pa)	1.0×10^{2}
millimeter of mercury (0 °C)	pascal (Pa)	1.333×10^{2}†
millimeter of water (39.2 °F)	pascal (Pa)	9.807
minute (angular)	radian	2.909×10^{-4}
myriagram	kilogram (kg)	10
myriameter	kilometer (km)	10
oersted	ampere per meter (A/m)	79.58
ounce (avoirdupois)	kilogram (kg)	2.835×10^{-2}
ounce (troy)	kilogram (kg)	3.110×10^{-2}
ounce (U.S. fluid)	cubic meter (m^3)	2.957×10^{-5}
ounce-force	newton (N)	0.2780
peck (U.S.)	cubic meter (m^3)	8.810×10^{-3}
pennyweight	kilogram (kg)	1.555×10^{-3}
pint (U.S. dry)	cubic meter (m^3)	5.506×10^{-4}
pint (U.S. liquid)	cubic meter (m^3)	4.732×10^{-4}
poise (absolute viscosity)	pascal second (Pa·s)	0.10†
pound (avoirdupois)	kilogram (kg)	0.4536
pound (troy)	kilogram (kg)	0.3732
poundal	newton (N)	0.1383
pound-force	newton (N)	4.448
pound per square inch (psi)	pascal (Pa)	6.895×10^{3}
quart (U.S. dry)	cubic meter (m^3)	1.101×10^{-3}
quart (U.S. liquid)	cubic meter (m^3)	9.464×10^{-4}
quintal	kilogram (kg)	1.0×10^{2}†
rad	gray (Gy)	1.0×10^{-2}†
rod	meter (m)	5.029
roentgen	coulomb per kilogram (C/kg)	2.58×10^{-4}
second (angle)	radian (rad)	4.848×10^{-6}
section	square meter (m^2)	2.590×10^{6}
slug	kilogram (kg)	14.59
spherical candle power	lumen (lm)	12.57
square inch	square meter (m^2)	6.452×10^{-4}
square foot	square meter (m^2)	9.290×10^{-2}
square mile	square meter (m^2)	2.590×10^{6}
square yard	square meter (m^2)	0.8361
store	cubic meter (m^3)	1.0†
stokes (kinematic viscosity)	square meter per second (m^2/s)	1.0×10^{-4}†
tex	kilogram per meter (kg/m)	1.0×10^{-6}†
ton (long, 2240 pounds)	kilogram (kg)	1.016×10^{3}
ton (metric)	kilogram (kg)	1.0×10^{3}†
ton (short, 2000 pounds)	kilogram (kg)	9.072×10^{2}
torr	pascal (Pa)	1.333×10^{2}
unit pole	weber (Wb)	1.257×10^{-7}
yard	meter (m)	0.9144†

A

ABLATION. See TISSUE ABLATION.

ABSORBABLE BIOMATERIALS. See BIOMATERIALS, ABSORBABLE.

ACRYLIC BONE CEMENT. See BONE CEMENT, ACRYLIC.

ACTINOTHERAPY. See ULTRAVIOLET RADIATION IN MEDICINE.

ADOPTIVE IMMUNOTHERAPY. See IMMUNOTHERAPY.

AFFINITY CHROMATOGRAPHY. See CHROMATOGRAPHY.

ALLOYS, SHAPE MEMORY

YOUNG KON KIM
Inje University
Kimhae City
Korea

INTRODUCTION

An alloy is defined as a substance with metallic properties that is composed of two or more chemical elements of which at least one is an elemental metal (1). The internal structure of most alloys starts to change only when it is no longer stable. When external influences, such as pressure and temperature, are varied, it will tend to transform spontaneously into a mixture of phases, the structures, compositions, and morphologies of which differ from the initial one. Such microstructural changes are known as phase transformation and may involve considerable atomic rearrangement and compositional change (2,3).

Shape memory alloys (SMAs) exhibit a unique mechanical "memory", or restoration force characteristic, when heated above a certain phase-transformation temperature range (TTR), after having been deformed below the TTR. This thermally activated shape recovering behavior is called the shape memory effect (SME) (3–5). This particular effect is closely related to a martensitic phase transformation accompanied by subatomic shear deformation resulting from the diffusionless, cooperative movement of atoms (6,7). The name martensite was originally used to describe the very fine, hard microstructure found in quenched steels (8). The meaning of this word has been extended gradually to describe the microstructure of nonferrous alloys that have similar characteristics.

SMAs have two stable phases: a high temperature stable phase, called the parent or austenite phase and a low temperature stable martensite phase. Martensite phases can be induced by cooling or stressing and are called thermally induced martensite (TIM) or stress induced martensite (SIM), respectively (8). The TIM forms and grows continuously as the temperature is lowered, and it shrinks and vanishes as the temperature is raised. The SIM is generated continuously with increasing applied stress on the alloy. On removing the applied stress, SIM disappears gradually at a constant temperature. If the temperature is sufficiently low when stressing, however, the SIM cannot return to its initial structure when the stress is removed. When the temperature is increased above the TTR, the residual SIM restores the original structure, resulting in shape recovery (9). Surprisingly, this process can be reliably repeated millions of times, provided that the strain limits are not breached. If dislocations or slips intervene in this process, the shape memory becomes imperfect. When the applied stress on a SMA is too great, irreversible slip occurs, and the SMA cannot recover its original shape even after heating above TTR (10). However, it can remember this hot parent pattern. In the next cooling cycle, the SMA changes slightly and remembers the cool-martensite pattern. A SMA trained with this repeated cyclic treatment is called a two-way SMA (9). A schematic explanation of the SME related to the two-dimensional (2D) crystal structure (11) is shown in Fig. 1. When a SMA is cooled below its TTR, the parent phase begins to form TIM without an external shape change. This TIM can be changed into SIM easily by mechanical deformation below the TTR. When the deformed SMA is heated above its TTR, however, it cannot hold the deformed shape anymore, and the SMA returns to its original shape, resulting in a reverse martensitic phase transformation.

A SMA also shows rubber-like behavior at temperatures above its TTR. When a SMA is deformed isothermally above its TTR, only SIM is produced, until plastic deformation occurs. Then, the SIM disappears immediately after removing the applied load, resulting in a much greater amount of recovering strain, in excess of the elastic limit, compared to the conventional elastic strain of a metal. This rubber-like behavior at a constant temperature above TTR is called superelasticity (12). A schematic explanation of superelasticity is shown in Fig. 2.

These contrasting behaviors of superelasticity and SME are a function of the testing temperature. If a SMA is tested below its TTR, it shows SME, while a SMA that is deformed above its TTR shows superelasticity.

It is convenient to subdivide the superelastic behavior into two categories, "superelasticity" and "rubber-like behavior", depending on the nature of the driving forces and mechanism involved. If it is triggered by SIM formation and subsequent reversion, the terminology superelasticity is used. By contrast, rubber-like behavior does not involve phase transformation, but involves deformation of the martensite itself. It is closely related to the reversible movement of deformed twin boundaries or martensite boundaries (10).

An example of SME in a shape-memory suture needle (13) is shown in Fig. 3. Figure 3a shows a curved needle with the shape preset by a heat-treatment process. When the shape-memory needle is cooled below its TTR, it is readily amenable to a change in shape with forceps (b). On heating it above TTR, thermal energy causes the needle to recover its original curved shape (c).

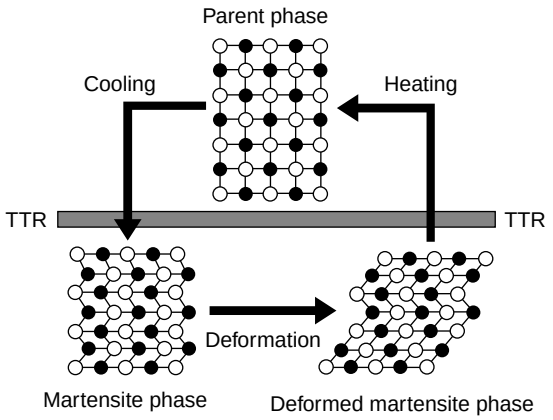


Figure 1. Schematic illustration of the shape memory effect. The parent phase is cooled below TTR to form a twinned (self-accommodated) martensite without an external shape change. Deformed martensite is produced with twin boundary movement and a change of shape by deformation below the TTR. Heating above the TTR results in reverse transformation and leads to shape recovery.

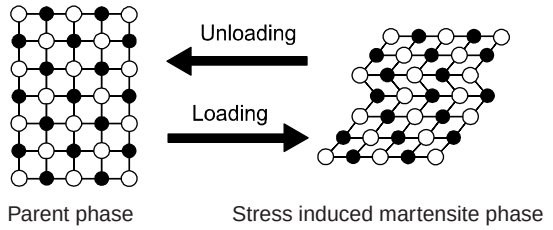


Figure 2. Schematic illustration of the superelasticity of a SMA above TTR. During the loading process, the applied load changes the parent phase into stress-induced martensite, which disappears instantly on unloading.

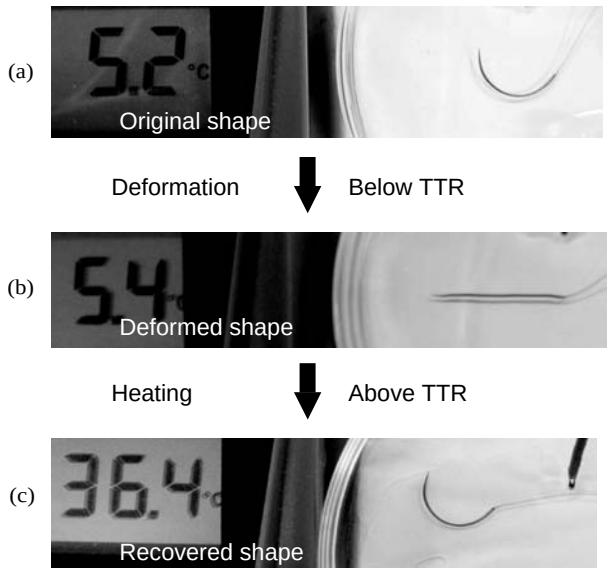


Figure 3. Shape-memory effect in a SMA suture needle. (a) Cooling the SMA suture needle below its TTR, (b) straightening the SMA suture needle below its TTR, (c) recovering the original shape of the SMA suture needle above its TTR.

HISTORY OF SHAPE MEMORY ALLOYS

The first observed shape memory phenomenon was pseudoelasticity. In 1932, Oelander observed it in a Au–Cd alloy and called it "rubber-like" behavior (14). Owing to the great amount of reversible strain, this effect is also called "superelasticity". The SME was discovered in 1938 by Greninger and Mooradian (15), while observing the formation and disappearance of martensite with falling and rising temperature in a brass (Cu–Zn alloy) sample. The maximum amount of reversible strain was observed in a Cu–Al–Ni single crystal with a recoverable elastic strain of 24% (16). In 1949, Kurdjumov and Khandros (17), provided a theoretical explanation of the basic mechanism of SME, the thermoelastic behavior of the martensite phase in Au–Cd alloy. Numerous alloy systems have been found to exhibit shape memory behavior. However, the great breakthrough came in 1963, when Buehler et al. (4) at the U.S. Naval Ordnance Laboratory discovered the SME in an equiatomic alloy of Ni–Ti, since then popularized under the name nitinol (Nickel–Titanium Naval Ordnance Laboratory). Partial listings of SMAs include the alloy systems: Ag–Cd, Au–Cd, Au–Cu, Cu–Zn, Cu–Zn–X (X = Si, Sn, Al, Ga), Cu–Al–Ni, Cu–Au–Zn, Cu–Sn, Ni–Al, Ni–Nb, Ni–Ti, Ni–Ti–Cu, Ti–Pd–Ni, In–Tl, In–Cd, Mn–Cd, Fe–Ni, Fe–Mn, Fe–Pt, Fe–Pd, and Fe–Ni–Co–Ti (9). It took several years to understand the microscopic, crystallographic, and thermodynamic properties of these extraordinary metals (18–20). The aeronautical, mechanical, electrical, biomedical, and biological engineering communities, as well as the health professions, are making use of shape memory alloys for a wide range of applications (9). Several commercial applications of Ni–Ti and Cu–Zn–Al SMAs have been developed, such as tube-fitting systems, self-erectable structures, clamps, thermostatic devices, and biomedical applications (5,21–23).

Andreasen suggested the first clinical application of Ni–Ti SMA in 1971. He suggested that nitinol wire was useful for orthodontics by reason of its superelasticity and good corrosion resistance (24). Since then, Ni–Ti alloys have been used in a broad and continually expanding array of biomedical applications, including various prostheses and disposables used in vascular and orthopedic surgery. Medical interventions have themselves been driven toward minimally invasive procedures by the creation of new medical devices, such as guide wires, cardiovascular stents, filters, embolic coils, and endoscopic surgery devices. The Ni–Ti SMA stent was first introduced in 1983 when Dotter (25) and Cragg (26) simultaneously published the results of their experimental studies. However, their studies were unsuccessful because of the unstable introduction system and the intimal hyperplasia in the stent-implanted region (27). In 1990, Rauber et al. renewed the effort to use a Ni–Ti alloy as a stent, significantly reducing intimal hyperplasia by using a transcatheter insertion method (28). In 1992, Josef Rabkin reported successful results in the treatment of obstructions in vascular and nonvascular systems in 268 patients (29). In 1989, Kikuchi reported that a guidewire constructed from kink-resistant titanium–nickel alloy was helpful for angiography and interventional

procedures (30). Guidewires are used for needles, endoscopes, or catheters, to gain access to a desired location within the human body. In 1989, the U.S. Food and Drug Administration approved the use of a Mitek anchor constructed of nitinol for shoulder surgery (31). Since then, many devices and items have been developed with nickel–titanium SMAs.

NICKEL–TITANIUM SHAPE MEMORY ALLOY

Physical Properties

Some of the physical properties of 55-Nitinol are listed in Table 1 (32,33). Nitinol has good impact properties, low density, high fatigue strength, and a nonmagnetic nature. The excellent malleability and ductility of nitinol enable it to be manufactured in the form of wires, ribbons, tubes, sheets, or bars. It is particularly useful for very small devices.

Phase Diagram and Crystal Structures

A Ti–Ni equilibrium phase diagram (34) is very useful for understanding phase transformation and alloy design; a modified one is shown in Fig. 4 (35). There is a triangular region designated "TiNi" near the point of equiatomic composition. The left slope (solubility limit) is nearly vertical with temperature. This means that a precipitation-hardening process cannot be used on the Ti-rich side in bulk alloys. By contrast, the right slope is less steep than the left. Therefore, the precipitation-controlling process can adjust transformation temperatures for practical application of SMAs on the Ni-rich side. The crystal structure of the upper part of this triangle, $> 1090\ ^\circ$C, is body centered cubic (bcc). The lower part is a CsCl-type ordered structure (B2) from 1090 °C to room temperature. A schematic atomic configuration of the B2 structure is shown in Fig. 5 (36). In 1965, Wang determined the lattice constant of the B2 crystal as $a_0 = 3.01$ Å (6). He proposed that the Ni–Ti crystal structure is not a simple CsCl-type structure, but has a disordered 9 Å superlattice and an ordered 3 Å CsCl-type sublattice. As the temperature is lowered, the ordered CsCl structure is slightly tilted instantaneously and cooperatively into a close-packed structure, called martensite, with a 2D dimensional close-packed plane (basal plane) (6,37). The martensite unit cell is described as a monoclinic (B19′) configuration, as shown in Fig. 6.

Table 1. Some of the Physical and Mechanical Properties of Nominal 55-Nitinol[a]

Density	6.45 g/cm^3
Melting point	1310 °C
Magnetic permeability coefficient	< 1.002
Electrical resistivity	
20 °C	80 μΩ · cm
900 °C	132 μΩ · cm
Thermal expansion	10.4×10^{-6}/ °C
Hardness,	
950 °C furnace cooled	89 R_B
950 °C quenched	89 R_B
Yield strength	103–138 MPa (15–20 $\times$ 10^3 psi)
U.T.S.	860 MPa (125 $\times$ 10^3 psi)
Elongation	60%
Young's modulus	70 GPa (10.2 $\times$ 10^6 psi)
Shear modulus	24.8 GPa (3.6 $\times$ 10^6 psi)
Poisson's ratio	0.33
Fatigue (Moore test) stress 10^7 counts	480 MPa (70 $\times$ 10^3 psi)
Charpy impact	
Unnotched (RT)[b]	155 ftlb
Unnotched (−80 °C)	160 ftlb
Notched (RT)	24 ftlb
Notched (−80 °C)	17 ftlb

[a]Reproduced with permission from Biocompatibility of Clinical Implant Materials volume I, Ed. By D. F. Williams, 1981, Table 2 on page 136, Castleman L. S. and Motzkin S. M., copyright CRC press, Boca Raton Florida. See Refs. (32) and (33).
[b]Room temperature = RT.

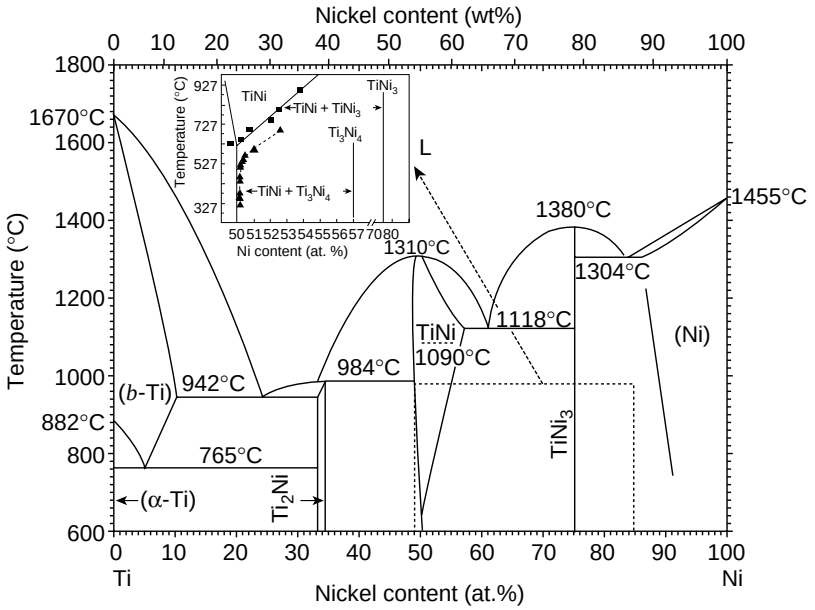


Figure 4. Phase diagram of a Ti–Ni alloy and details of the TiNi and $TiNi_3$ phases (35). (Reproduced with permission from Binary Alloy Phase Diagrams, 2nd ed., Vol. 3, 1990, Phase diagram of a Ti-Ni alloy on page 2874, T. B. Massalski, H. Okamoto, P. R. Subramanian, and L. Kacprzak, ASM International.)

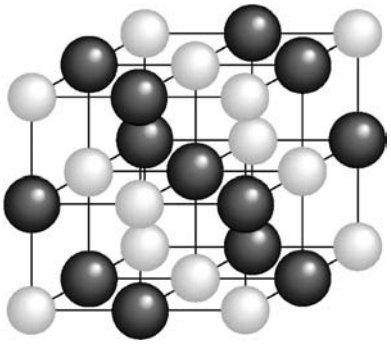


Figure 5. Schematic 3D diagram of the Ni–Ti atomic model in the stable high temperature phase (CsCl-type structure; lattice constant; $a_0 = 3.01$ Å).

The twin-type stacking of the thermally induced martensite structure shown on the left (a) has a readily deformable crystalline arrangement, from the twin structure to the detwinned structure shown on the right (b) (9,38). Diagram (b) of the detwinned structure shows relatively planar atomic stacking layer by layer alternately along the {111} basal plane of the deformed martensite crystal (39). Since martensitic transformation in Ni–Ti SMAs demonstrates an abnormal heat capacity change, it is regarded as a crystallographic distortion instead of a crystallographic transformation. The Ni–Ti martensite transformation is accompanied by a large latent heat of enthalpy ($\Delta H \sim 4{,}150$ J/mol). This extraordinarily latent heat of transformation was considered to be owing to a portion of the electrons undergoing a "covalent-to-metallic" electron-state transformation (11).

Thermomechanical Properties

The mechanical properties of Ni–Ti SMAs are closely dependent on the testing temperature. If a mechanical stress is applied to the SMA below the TTR, then the metastable parent structure of the Ni–Ti alloy is susceptible to transformation into the martensite. However, if the testing temperature exceeds the TTR, then, in the absence of stress, the reverse transformation happens. Figure 7 shows an example of a uniaxial compressive stress-strain curve of a Ni–Ti alloy above its TTR, which shows its superelasticity (40).

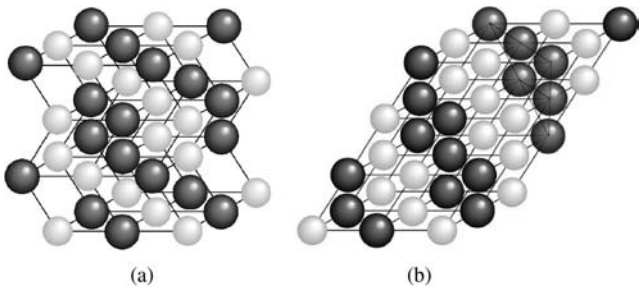


Figure 6. Schematic 3D diagram of the Ni–Ti atomic stacking model of low temperature stable monoclinic structured martensite (a) twin-type stacking of martensite, (b) detwinned-type stacking of martensite).

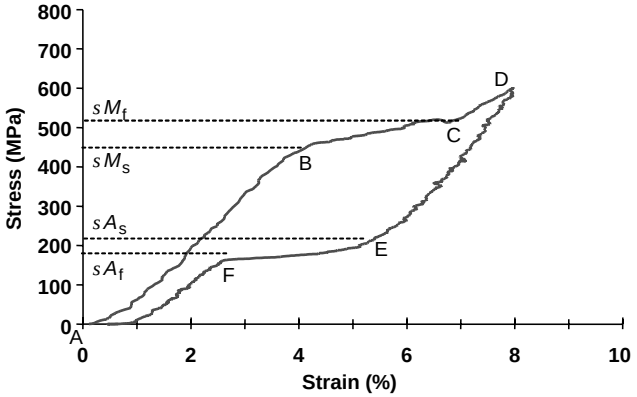


Figure 7. Compressive stress–strain curve of a heat-treated 6-mm-diameter Ni–Ti rod at 4 °C. Three distinct stages are observed on the stress–strain curve (σM_f: stress-induced martensite finishing stress, σM_s: stress-induced martensite starting stress, σA_s: parent phase starting stress, σA_f: parent phase finishing stress) (40).

With stress below the martensite starting stress (σM_s), the Ni–Ti alloy behaves in a purely elastic way, as shown in section AB. As soon as the critical stress is reached at point B, corresponding to stress level σM_s, forward transformation (parent phase-to-martensite) is initiated and SIM starts to form. The slope of section BC (upper plateau) reflects the ease with which the transformation proceeds to completion, generating large transformational strains. When the applied stress reaches the value of the martensite finishing stress (σM_f), the forward transformation is completed and the SMA is fully in the SIM phase. For further loading above σM_f, the elastic behavior of martensite is observed again until plastic deformation occurs, as represented in section CD. For stress beyond D, the material deforms plastically until fracture occurs. However, if the stress is released before reaching point D, the strain is recovered in several stages. The first stage is elastic unloading of the martensite, as shown in section DE. On arriving at stress σA_s, at E, the reverse martensite transformation starts and the fraction of martensite decreases until the parent phase is completely restored at F. Section FA represents the elastic unloading of the parent phase. If some irreversible deformation has taken place during either loading or unloading, the total strain may not be recovered completely. Owing to the stress differences between σM_f and σA_s and between σM_s and σA_f, a hysteresis loop is obtained in the loading–unloading stress–strain curve. Increasing the test temperature results in an increase in the values of the critical transformation stresses, while the general shape of the hysteresis loop remains the same. The area enclosed by the loading and unloading curves represents the energy dissipated during a stress cycle. As part of the hysteresis loop, both the loading and unloading curves show plateaus, at which point large strains are accommodated on loading, or recovered on unloading, with only a small change in stress (19). This behavior of Ni–Ti SMAs is much like that of natural tissues, such as hair and bone, and results in a "superelastic" ability to withstand and recover from large deforming stresses.

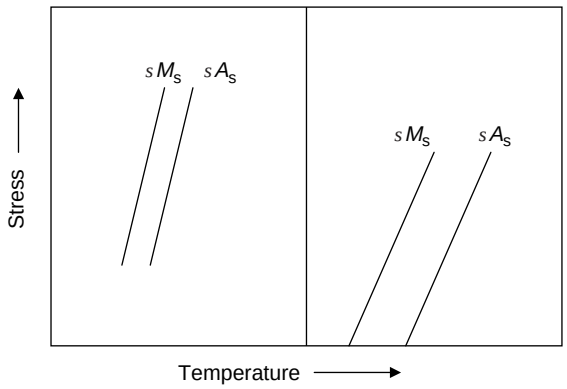


Figure 8. The effect of compressive (a) and tensile (b) loading on martensite formation and disappearance in 20.7% Tl–In alloy (19). (From J. Mat. Sci. Vol. 9, 1974, Figure 2 on page 1537, Krishnan R. V., Delaey L., Tas H. and Warlimont H., Kluwer Academic Publisher. Reproduced with kind permission of Springer Science and Business Media.)

In 1974, Krishnan argued that Burkart and Read had found the effects of compressive and tensile stress on martensite formation and disappearance in Tl–In SMAs (19). The transformation stresses σM_s and σA_s have a linear relation with testing temperature, as shown in Fig. 8 (41). They inferred that σM_s is a linear function of temperature, and the stresses σM_s and σA_s increase with temperature.

Another important thermomechanical property of SMAs is the relationship between the plateau stress of the martensite phase transformation and the enthalpy change of that reaction. As the stress-induced martensitic transformation is a second-order transformation, the amount of transformation depends on its temperature, so the high temperature state has a larger energy barrier of SIM and needs more energy to overcome this larger reverse martensitic transformation barrier. The enthalpy change of the parent phase to martensitic transformation ($\Delta H^{\text{p-m}}$) can be calculated theoretically using the modified Clausius–Clapeyron equation (20), shown in Eq. 1.

$$\frac{\delta\sigma^{\text{p}-\text{m}}}{\delta T} = \frac{\rho\Delta H^{\text{p-m}}}{\varepsilon^{\text{p}-\text{m}} T_{\text{o}}} \tag{1}$$

Where $\Delta H^{\text{p}-\text{m}}$ is the enthalpy change of the parent phase to the martensite phase at T_{o}; $\sigma^{\text{p}-\text{m}}$ is the stress at which stress-induced martensite is formed at the testing temperature, T; ρ is the density of the SMA; and $\varepsilon^{\text{p}-\text{m}}$ is the strain corresponding to complete transformation. $\delta\sigma^{\text{p}-\text{m}}$ and $\varepsilon^{\text{p}-\text{m}}$ can be taken from the stress–strain curves. Kim compared the theoretically calculated $\Delta H^{\text{p}-\text{m}}$ of a Ni–Ti alloy using stress–strain curves and Eq. 1 with an experimentally acquired value (9). He reported that the theoretical value of $\Delta H^{\text{p}-\text{m}}$ for a Ni–Ti alloy calculated from the stress–strain curves was 6.24 cal/g. The experimental value of the enthalpy change ($\Delta H^{\text{p}-\text{m}}$) of an 8% prestrained Ni–Ti wire sample from DSC measurement was 6.79 cal·g^{-1}. Based on this result, he inferred that Ni–Ti alloys undergo thermomechanical-phase transformation by exchanging thermal energy into mechanical energy and vice versa (9).

MANUFACTURING METHODS

Alloy Refining

The Ni–Ti SMAs can be refined using either the vacuum-induction melting method or the consumable arc melting technique. In vacuum-induction melting, a prerequisite in working with Ni–Ti is a high purity graphite crucible. To prevent impurities, the crucible should be connected to the pouring lip mechanically to keep the molten Ni–Ti compound from contacting anything, but the high density, low porosity graphite. Elemental carbon is very reactive with Ni or Ti alone and any contact with either will ruin the purity of the desired sample. However, there is very little reaction with the crucible in the consumable arc melting process. This method yields a product that is relatively free of impurities. Once the Ni–Ti alloy is cast using the melting technique, it is ready for hot or cold working into more practical forms and consecutive annealing treatment (42).

Mechanical Processing

When hot working a piece of Ni–Ti alloy, the temperature should be below that where incipient melting of the secondary phase can occur. This temperature should also be held constant for a period of time sufficient for certain nonequilibrium phases to return to solution, which makes the remaining alloy homogeneous. Andreasen suggested that the optimum hot working temperature is 700–800 °C for forging, extrusion, swaging, or rolling. If cold rolling is desired, then the alloy should be annealed before the oxide is removed (42). The most common form of Ni–Ti alloy is a wire. To make a wire, the Ni–Ti alloy ingot must be rolled into a bar at high temperature. Swaging the bar, followed by drawing, and a final annealing, reduce the alloy to wire form. To soften the wire, it should be annealed between 600 and 800 °C for a short period. When the Ni–Ti alloy is drawn down to 0.8 mm through a carbide die, the maximum reduction in area with each pass should be within 10%. Once this diameter is reached, a diamond die is used to draw the alloy with a 20% area reduction per die. The Ni–Ti alloy is annealed again at 700 °C and allowed to cool to room temperature between passes (42). By contrast, the extrusion method is used for the tube-making process, which enables a substantially greater reduction in cross-sectional area as compared to drawing wire. Laser cutting of Ni–Ti tubes has been used to make vascular stents (43). Most Ni–Ti alloys require a surface finishing procedure after the final machining process, such as chemical leaching, cleaning, rinsing, and surface modification.

Shape Memory Programming

There are two steps in the shape memory programming of a Ni–Ti alloy. First, the Ni–Ti alloy sample must be deformed to the desired shape and put into a constraining mold or fixture. The next step is shape memory heat treatment in a furnace at 400–600 °C. The shape recovery efficiency of a Ni–Ti alloy can be controlled by changing the heat treatment conditions or the degree of deformation. In general, there are three different ways to control the TTR of a Ni–Ti SMA: altering the chemical composition,

changing the heat treatment conditions, and varying the degree of deformation (13).

Chemical Composition Effect on TTR. The shape memory characteristic is limited to Ni–Ti alloys with near-equiatomic composition, as shown in Fig. 4. A pure stoichiometric (50 at%) Ni–Ti alloy will have a nickel content of $\sim$ 55 wt%. Increasing the nickel concentration lowers the characteristic transformation temperature of the alloy. The limit of the nickel concentration for a SMA is $\sim$ 56.5 wt%, owing to the formation of a detrimental second phase. In addition, the shape memory properties of a Ni–Ti alloy can be readily modified by adding ternary elements that are chemically similar to Ni or Ti. Adding a small amount of a transition metal such as Co, Fe, or Cr, instead of Ni, depresses the TTR, such that the SME occurs at well-below ambient temperature (44). When larger ions are substituted for smaller ions, the transformation temperature increases. Concerning ternary additions to alloys, Murakami et al. (45) proposed that the stability of the parent phase is controlled by ion–core repulsive interactions such that when larger ions are substituted for smaller ions, the transformation temperature increases. Based on this hypothesis, substitutions of Au and Zr should increase the recovery temperature of Ni–Ti alloys, Al and Mn should decrease it, and Co and Fe should cause little change. The effects of Au, Zr, Al, and Mn were predicted correctly, but those of Co and Fe were not. Similarly, Morberly suggested that if $>$ 7.5% copper is added to a Ni–Ti alloy, up to 30%, the addition of Cu increases and narrows the TTR (46).

Mechanical Deformation Effect on the TTR. Many investigators have reviewed the effect of mechanical deformation on the TTR (5,36,47). They found that the degree of deformation affects the TTR of a SMA, and the stress slope ($d\sigma/dT$) is a very important fundamental descriptor of SMAs. The residual stress from prior cold work can have a major effect on the transformation behavior. As a result, retention of the parent phase is a function of the stress and heat treatment history. Lee et al. reported that bending beyond the yielding point broadened the TTR and increased the stored internal energy (48). Figure 9 shows an example of transition temperature variation with respect to uniaxial prestrain of a Ni–Ti alloy wire (13). When the prestrain $>$ 8%, the shape recovery transition temperature (A_s) and the martensite starting temperature (M_s) are increased with increasing prestrain. However, the enthalpy change of the cooling cycle is almost the same because most stored internal energy in SIM is already liberated during the heating cycle (36).

Heat Treatment Effect on the TTR. The TTR of a Ni–Ti alloy can be controlled by the final annealing temperature and time. Kim insisted that a higher annealing temperature gives a lower transition temperature and a wider TTR (9). Moreover, he showed that a larger grain size has a lower transition temperature because the annealed large grains have much more transformable volume than smaller grains, so they need more energy for second-phase nucleation and growth inside the grain (49).

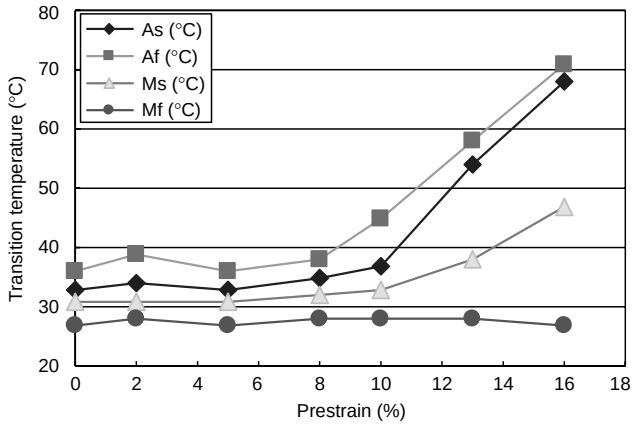


Figure 9. Transition temperatures of prestrained Ni–Ti alloy wire (13).

Figure 10 shows an example of the heat treatment temperature effect on SME (40). When a Ni–Ti rod is heat treated for 30 min at 600 °C, the rod shows superelasticity at room temperature. This indicates that the TTR is lower than the testing temperature. By contrast, when a Ni–Ti rod is heat treated for 30 min at $<$ 500 °C, the rod shows SME at room temperature, which suggests that the TTR is higher than the testing temperature. These results clearly show that the SME is closely related to the heat treatment temperature (9,50).

Methods of Measuring Transition Temperatures

There are many measurable parameters that accompany the shape memory transformation of a Ni–Ti alloy, for example, hardness, velocity of sound, damping characteristic, elastic modulus, thermal expansion, electrical resistivity, specific heat, latent heat of transformation, thermal conductivity, and lattice spacing. Of these, the electrical resistivity and latent heat of transformation are useful for measuring the TTR of a SMA.

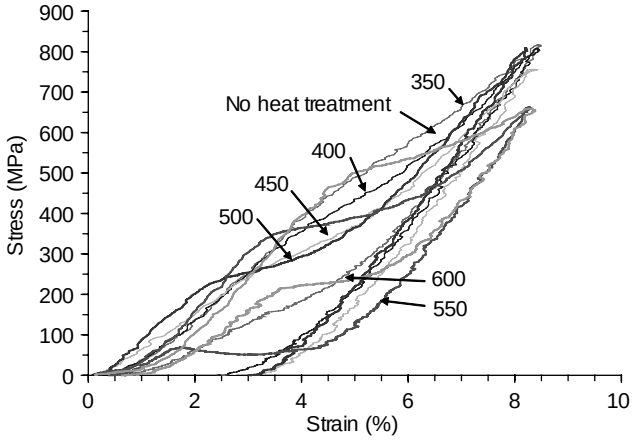


Figure 10. The room temperature compression stress–strain curves of heat-treated ϕ6-mm Ni–Ti rods for 30 min at 350–600 °C. The numbers pointing to the graphs are the annealing temperatures (40).

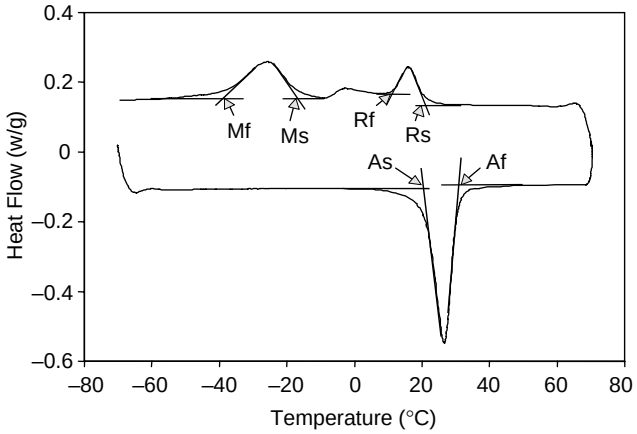


Figure 11. A cyclic DSC curve of the specific heat versus temperature for a Ni–Ti alloy wire from −70 to 70 °C. The lower and upper parts of the cyclic curve represent heating and cooling processes, respectively. (A_s: shape recovery starting temperature, A_f: shape recovery finishing temperature, M_s: martensitic transformation starting temperature, M_f: martformation starting temperature, R_f: R phase transformation finishing temperature) (40).

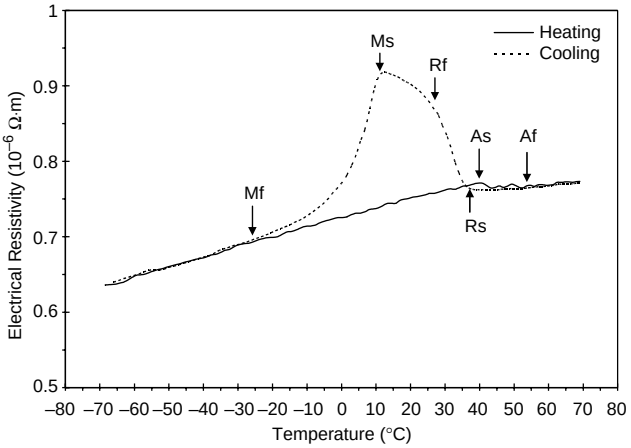


Figure 12. The electrical resistivity versus temperature curve of a ϕ1.89 mm Ni–Ti alloy wire that was heat treated at 550 °C (53).

DSC Measurement. Differential scanning calorimetry (DSC) is a thermal analysis technique that determines the specific heat, heat of fusion, heat of reaction, or heat of polymerization of materials. It is accomplished by heating or cooling a sample and reference under such conditions that they are always maintained at the same temperature. The additional heat required by the sample to maintain it at the same temperature is a function of the observed chemical or physical change (50). Figure 11 shows a typical DSC curve of the specific heat change of a Ni–Ti alloy (40). The lower curve is the heating curve and the upper one is the cooling curve. Each peak represents a phase transformation during the thermal cycle. The area under the curve represents the enthalpy change (ΔH) during the phase transformation. The arrows on Fig. 11 indicate the transition temperatures. The advantage of DSC measurement is that samples can be small and require minimal preparation. In addition, it can detect the residual strain energy, diffusing DSC peaks (51).

Electrical Resistivity Measurement. The shape memory transition temperature can also be determined from the curve of the electrical resistance versus temperature using a standard four-probe potentiometer within a thermal scanning chamber. In 1968, Wang reported the characteristic correlation between the shape memory phase transformations of a Ni–Ti alloy and the irreversible electrical resistivity curves (52). He proposed that the electrical resistivity curve in the same temperature range has a two-step process on cooling, that is, from the parent phase via R-phase to the final martensite phase, and a one-step process on heating, that is, from the martensite to the parent phase. Figure 12 plots the electrical resistivity versus temperature curves of a ϕ1.89 mm Ni–Ti alloy wire that was heat treated at 550 °C for 30 min. During the heating process, the electrical resistivity increases up to temperature A_s, and then it decreases until temperature A_f is reached. This suggests the restoration of the parent structure accompanying this resistivity change. During the cooling cycle, however, a triangular curve appears. The increasing part of this triangular curve from R_s to R_f represents the formation of an intermediate R phase resulting in a further increase in electrical resistivity. The decreasing part represents the thermal energy absorption of the martensitic phase transformation.

Corrosion Resistance

The Ni–Ti SMAs are an alloy of nickel, which is not corrosion resistant in saline solutions, such as seawater, and titanium, which has excellent corrosion resistance under the same conditions. The corrosion resistance of Ni–Ti alloys more closely resembles that of titanium than that of nickel. The corrosion resistance of Ni–Ti alloys is based mainly on the formation of a protective oxide layer, which is called passivation (9). If the alloy corrodes, then breakdown of the protective oxide film on the alloy's surface occurs locally or generally, and substantial quantities of metallic ions are released into the surrounding solution. Therefore, corrosion resistance is an important determinant of biocompatibility (54–56). The Pourbaix diagram is a useful means of measuring corrosion. It is a potential versus pH diagram of the redox and acid–base chemistry of an element in water. It is divided into regions where different forms of the metal predominate. The three regions of interest for conservation are corrosion, immunity, and passivity. The diagram may indicate the likelihood of passivation (or corrosion) behavior of a metallic implant *in vivo*, as the pH varies from 7.35 to 7.45 in normal extracellular fluid, but can reach as low as 3.5 around a wound site (57). An immersion test is also used for determining the concentration of released metallic ions, corrosion rates, corrosion types, and passive film thickness in saline, artificial saliva, Hank's solution, physiological fluids, and so on (58).

Some surface modifications have been introduced to improve the corrosion properties of Ni–Ti alloys, and prevent the dissolution of nickel. These include titanium

nitride coating of the Ni–Ti surface and chemical modification with coupling agents for improving corrosion resistance. However, when the coating on a Ni–Ti alloy is damaged, corrosion appears to increase in comparison with an uncoated alloy (56). Laser surface treatment of Ni–Ti leads to increases in the superficial titanium concentration and thickness of the oxide layer, improving its cytocompatibility up to the level of pure titanium (9). Electropolishing methods and nitric acid passivation techniques can improve the corrosion resistance of Ni–Ti alloys owing to the increased uniformity of the oxide layer (59).

Biocompatibility

Biocompatibility is the ability of a material or device to remain biologically inactive during the implantation period. The purpose of a biocompatibility test is to determine potential toxicity resulting from contact of the device with the body. The device materials should not produce adverse local or systemic effects, be carcinogenic, or produce adverse reproductive or developmental effects, neither directly nor through the release of their material constituents (60). Therefore, medical devices must be tested for cytotoxicity, toxicity, specific target-organ toxicity, irritation of the skin and mucosal surfaces, sensitization, hemocompatibility, short-term implantation effects, genotoxicity, carcinogenicity, and effects on reproduction.

The biocompatibility of a Ni–Ti alloy must include the biocompatibility of the alloy's constituents. As Ni–Ti alloys corrode, metallic ions are released into the adjacent tissues or fluids by some mechanisms other than corrosion (61). Although Ni–Ti alloys contain more nickel than 316L stainless steel, Ni–Ti alloys show good biocompatibility and high corrosion resistance because of the naturally formed homogeneous TiO_2 coating layer, which has a very low concentration of nickel. Although Ni–Ti alloys have the corrosion resistance of titanium, the passivated oxide film will dissolve at some rate; furthermore, the oxide layer does not provide a completely impervious barrier to the diffusion of nickel and titanium ions (62,63).

Many investigators have reported on the biocompatibility of Ni–Ti alloys. Comparing the corrosion resistance of common biomaterials, the biocompatibility of Ni–Ti ranks between that of 316L stainless steel and Ti6A14V, even after sterilization. Some of these findings are listed here. Thierry found that electropolished Ni–Ti and 316L stainless steel alloys released similar amounts of nickel after a few days of immersion in Hank's solution (64). Trepanier reported that electropolishing improved the corrosion resistance of Ni–Ti stents because of the formation of a new homogeneous oxide layer (59). In a short-term biological safety study, Wever found that a Ni–Ti alloy had no cytotoxic, allergic, or genotoxic activity and was similar to the clinical reference control material AISI 316 LVM stainless steel (65). Motzkin showed that the biocompatibility of nitinol is well within the limits of acceptability in tissue culture studies using human fibroblasts and buffered fetal rat calvaria tissue (66). Ryhanen reported that nitinol is nontoxic, nonirritating, and very similar to stainless steel and Ti–6Al–4V alloy in an *in vivo* soft tissue and inflammatory response study (67). Castleman found no significant histological compatibility differences between nitinol and Vitallium (Co–Cr alloy) (68). However, Shih reported that nitinol wire was toxic to primary cultured rat aortic smooth muscle cells in his cytotoxicity study using a supernatant and precipitate of the corrosion products (69). Moreover, he found that the corrosion products altered cell morphology, induced cell necrosis, and decreased cell numbers.

MEDICAL DEVICES

The Ni–Ti alloys have been used successfully for medical and dental devices because of their unique properties, such as SME, superelasticity, excellent mechanical flexibility, kink resistance, constancy of stress, good elastic deployment, thermal deployment, good corrosion resistance, and biocompatibility. Recently, Ni–Ti alloys have found use in specific devices that have complex and unusual functions, for example, self-locking, self-expanding, or compressing implants that are activated at body temperature (58). Some popular examples of Ni–Ti medical devices have been selected and are reviewed below.

Orthodontic Arch Wires

A commercially available medical application of nitinol is the orthodontic dental arch wire for straightening malpositioned teeth, marketed by Unitek Corporation under the name Nitinol Active-Arch (70). This type of arch wire, which is attached to bands on the teeth, is intended to replace the traditional stainless steel arch wire. Although efforts have been made to use the SME in orthodontic wires (71), the working principle of Nitinol Active-Arch wire is neither the SME nor pseudoelasticity, but the rubber-like behavior and relatively low Young's modulus (30 GPa) of nitinol in the martensitic condition. This modulus is very low in comparison with the modulus of stainless steel (200 GPa). Comparing the bending moment change of nitinol and stainless steel wire undergoing a constant change in deflection (72), stainless steel wire shows a much larger change in moment than the moment change of nitinol wire. Clinically, this means that for any given malocclusion nitinol wire will produce a lower, more constant force on the teeth than would a stainless steel wire of equivalent size. Figure 13 shows a clinical example of orthodontic treatment using a superelastic Ni–Ti arch wire (73). This wire showed faster movement of teeth and shorter chair time than conventional stainless steel wire.

Guidewires

One typical application of superelasticity is the guidewires that are used as guides for the safe introduction of various therapeutic and diagnostic devices. A guidewire is a long, thin metallic wire that is inserted into the body through a natural opening or a small incision. The advantages of using superelastic guide wire are the improvement in kink resistance and steerability. A kink in a guidewire creates a difficult situation when the time comes to remove it from a

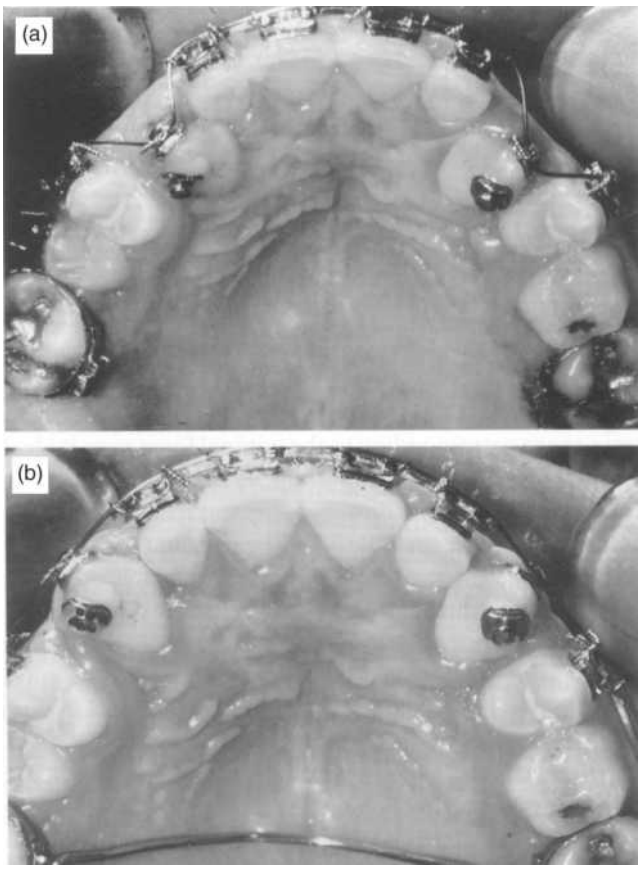


Figure 13. Orthodontic treatment using a Ni–Ti superelastic arch wire. (a) Malaligned teeth before treatment and (b) normally aligned teeth after the first stage of treatment (73). (Reprinted with permission from Shape memory materials, Ed. By K. Otsuka and C. M. Wayman, 1998, Figure 12.3 on page 270, S. Miyazaki, Cambridge University Press.)

complex vascular structure. The enhanced twist resistance and flexibility make it easier for the guidewire to pass to the desired location (74). Figure 14 shows the tip of a guidewire. The curved "J" tip of the guidewire makes it easy to select the desired blood vessel.

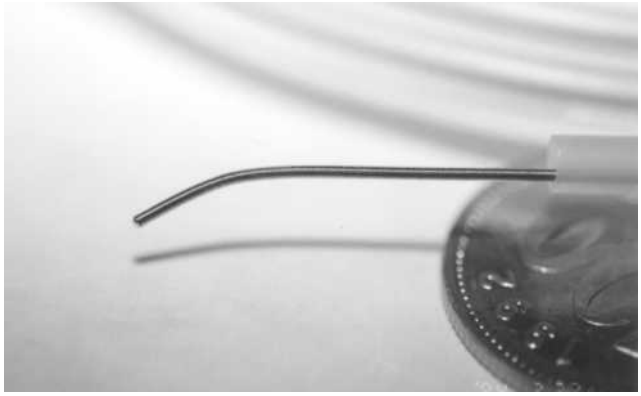


Figure 14. Photograph of the tip of a commercial Ni–Ti guidewire (FlexMedics, USA) (75).

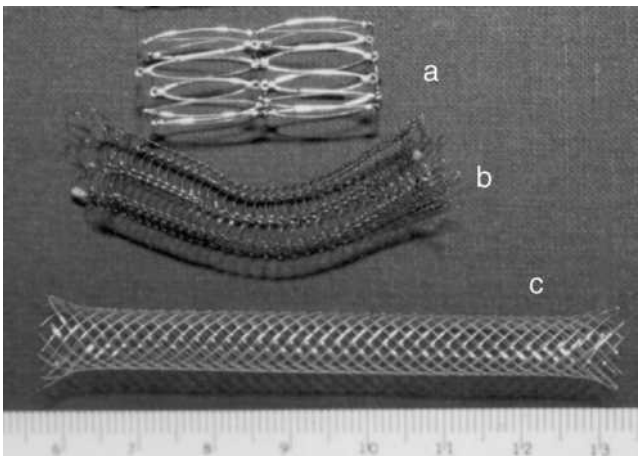


Figure 15. Commercial Ni–Ti stents (a) Gianturco stent, (b) self-expanding nitinol stent with the Strecker stent design, (c) Wall stent (76).

Stents

A stent is a slender metal mesh tube that is inserted inside a luminal cavity to hold it open during and after surgical anastomosis. Superelastic nitinol stents are very useful for providing sufficient crush resistance and restoring lumen shape after deployment through a small catheter (25–27). Figure 15 shows three examples of commercial self-expandable Ni–Ti superelastic stents: a Gianturco stent for the venous system, a Strecker stent for a dialysis shunt, and a Wall stent for a hepatic vein. Figure 16 shows the moment of expansion of a Ni–Ti self-expandable stent being deployed from the introducer. The driving force of the self-expanding stent is provided by the superelasticity of the Ni–Ti alloy. Some clinical limitations of Ni–Ti stents remain unresolved and require further development; these are the problems of intimal hyperplasic and restenosis (78).

Orthopedic Applications

Dynamic compression bone plates exhibiting the SME are one of the most popular orthopedic applications of nitinol, followed by intramedullary fixation nails. Fracture healing

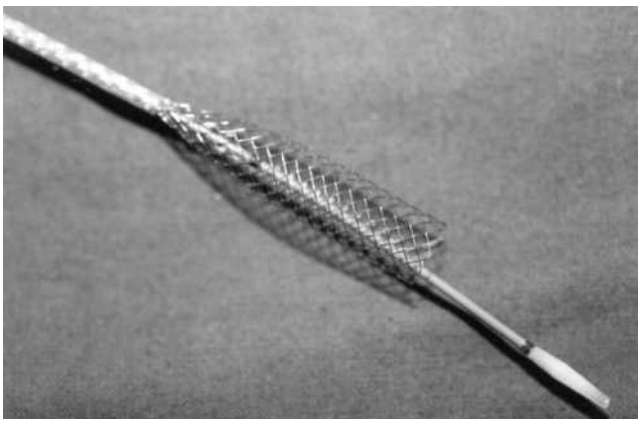


Figure 16. Deployment of a commercial Ni–Ti self-expandable stent (Taewoong Medical, Korea) (77).

in long bones can be accelerated when bone ends are held in position with compression between the bone fragments. Using this method, the undesirable surface damage and wear of the holes that occur in a conventional dynamic bone plate are avoided, while continuous compression is assured, even if bone resorption occurs at the fracture sites. The effect continues as long as the original shape is not reached (79).

Historically, the first orthopedic application of a SMA was a Nitinol Harrington instrument for scoliosis treatment that was introduced in 1976 by Schmerling (80), which enabled the surgeon to restore any relaxed corrective force postoperatively simply by the external application of heat. In addition, it could be used initially to apply a more appropriate set of corrective forces. Figure 17 shows an example of a Ni–Ti shape memory clamp in small bone surgery (81). Six months after surgery, a non-union was present, although the outcome in this patient was assessed as good.

CONCLUSIONS

A shape memory alloy is a metallic substance that has a memory for shape combined with superelasticity. The mechanisms of a nickel–titanium alloy's shape memory effect and superelasticity are described based on thermally induced or stress induced martensite phase transformations. Some of the physical properties of nickel–titanium alloys and a phase diagram are included for reference. The thermomechanical characteristics, corrosion properties, and biocompatibility of Ni–Ti shape memory alloys are reviewed for the design of shape memory devices. Manufacturing methods, including refining, processing, shape memory programming, and transformation temperature range measuring methods are summarized for practical applications. Finally, some applications in medical devices are reviewed as examples of current trends in the use of shape memory alloys. In conclusion, Ni–Ti shape memory alloys are a very useful biocompatible material because of their unique mechanical properties and good corrosion resistance. A better understanding of shape memory alloys should allow further developments in this area.

Figure 17. Failed arthrodesis of the carpometacarpal joint when only one titanium–nickel (TiNi) clamp was used (81). (From Arch. Orthop. Trauma Surg., Vol. 117, 1998. Figure 1 on page 342, Musialek J., Filip P. and Nieslanik J. Reproduced with kind permission of Springer Science and Business Media.)

BIBLIOGRAPHY

Cited References

1. Properties and selection. Metal handbook 8th edition volume 1. American Society for Metals; 1961. p 1.
2. Jena AK, Chaturvedi MC. Phase transformation in materials. Prentice Hall; 1992. p 1–3.
3. Park JB, Kim YK. Metallic biomaterials. In: Bronzino JD, editor. The biomedical engineering handbook. 2nd ed. Volume 1, CRC Press; 2000. p 37-1–37-20.
4. Buehler WJ, Gilfrich JV, Wiley RC. Effect of low-temperature phase changes on the mechanical properties of alloys near composition TiNi. J Appl Phys 1963;34:1475–1477.
5. Wayman CM, Duerig TW. An introduction to martensite and shape memory. In: Duerig TW, Melton KN, Stoeckel D, Wayman CM, editors. Engineering aspects of shape memory alloys. Butterworth-Heinemann; 1990. p 3–20.
6. Wang FE, Buehler WJ, Pickart SJ. Crystal structure and a unique martensite transition on TiNi. J Appl Phys 1965;36: 3232–3239.
7. Ling CH, Kaplow R. Stress-induced shape changes and shape memory in the R and Martensite transformations in equiatomic NiTi. Metal Trans A 1981;12A:2101–2111.
8. In: Nishiyama Z, Fine ME, Meshii M, Wayman CM, editors. Martensitic Transformation. London: Academic Press; 1978. p 1–13.
9. Kim YK. Thermo-mechanical study of annealed and laser heat treated nickel–titanium alloy dental arch wire. Ph.D. dissertation, University of Iowa, Iowa, Dec. 1989.
10. Wayman CM, Bhadeshia H. Phase transformations, Nondiffusive. In: Cahn RW, Haasen P, editor. Physical Metallurgy. 4th ed. Volume 2, North-Holland: 1996. p 1507–1554.
11. Wang FE, Pickart SJ, Alperin HA. Mechanism of the TiNi transformation and the crystal structures of TiNi-II and TiNi-III phases. J Appl Phys 1972;43:97–112.
12. Otsuka K, Wayman CM, Nakai K, Sakamoto H, Shimizu K. Superelasticity effects and stress-induced martensite transformations in Cu–Al–Ni alloys. Acta Metallurgica 1976;24: 207–226.
13. Kim YK, Doo JK, Park JP. The application of shape memory alloy to abdominoscopic suture needles, In: Shin KS, Yoon JK, Kim SJ, editors. Proceeding of 2nd Pacific RIM International conference on Advanced Materials and Processing. Korean Institue of Metals and Materials; 1995. p 1691–1696.
14. Oelander A. Z Kristallogr 1932;83A:145. as cited in Lieberman DS. Crystal geometry and mechanisms of phase transformations in crystalline solids. In: Aaronson HI, editor. Phase Transformations. American Society for Metals; 1970. p 1–58.
15. Greninger AB, Mooradian VG. Strain transformation in metastable beta copper-zinc and beta copper-tin alloys. Am Inst Mining Met Eng 1937;19:867.
16. Bush RE, Leudeman RT, Gross PM. Alloys of improved properties. AMRA CR 65-02/1, AD629726, U.S. Army Materials Research Agency, 1966.
17. Kurdjumov GV, Khandros LG. Dokl Akad Nauk SSSR 1949;66:211. (as cited in Delaey L, Krishnan RV, Tas H, Warlimont H. Review: thermoelasticity, pseudoelasticity and memory effects associated with martensitic transformations. J Mater Sci 1974;9:1521–1535.
18. Delaey L, Krishnan RV, Tas H, Warlimont H. Review Thermoelasticity, pesudoelasticity and the memory effects associated

with martensitic transformations Part 1 Structural and microstructural changes associated with the transformations. J Mat Sci 1974;9:1521–1535.
19. Krishnan RV, Delaey L, Tas H, Warlimont H. Review Thermoelasticity, pesudoelasticity and the memory effects associated with martensitic transformations Part 2 The macroscopic mechanical behaviour. J Mater Sci 1974;9:1536–1544.
20. Warlimont H, Delaey L, Krishnan RV, Tas H. Review Thermoelasticity, pesudoelasticity and the memory effects associated with martensitic transformations Part 3 Thermodynamics and kinetics. J Mater Sci 1974;9:1545–1555.
21. Melton KN. General applications of SMA's and smart materials. In: Duerig TW, Melton KN, Stoeckel D, Wayman CM, editors. Engineering aspects of shape memory alloys. Butterworth-Heinemann; 1990. p 220–239.
22. Miayazaki S. Medical and dental applications of shape memory alloys, In: Duerig TW, Melton KN, Stoeckel D, Wayman CM, editor. Engineering aspects of shape memory alloys. Butterworth-Heinemann; 1990. p 267–281.
23. Filip P. Titanium-Nickel shape memory alloys in medical applications. In: Brunette DM, Tengvall P, Textor M, Thomsen P, editor. Titanium in Medicine. Springer; 2001. p 53–86.
24. Andreasen GF, Hilleman TB. An evaluation of 55 cobalt substituted Nitinol wire for use in orthodontics. JADA 1971;82: 1373–1375.
25. Dotter CT, Bushmann RW, McKinney MK, Rosch J. Transluminal expandable nitinol coil stent grafting: preliminary report. Radiology 1983;147:259–260.
26. Cragg A, Lund G, Rysavy J, Castaneda F, Castaneda-Zuniga W, Amplatz K. Nonsurgical placement of arterial endoprostheses: a new technique using nitinol wire. Radiology 1983;147: 261–263.
27. Rösch J, Keller FS, Kaufman JA. The Birth, Early Years, and Future of Interventional Radiology. JVIR 2003;14(7):841–853.
28. Rauber K, Franke C, Rau WS, Syed Ali S, Bensmann G. Perorally insertable endotracheal stents made from NiTi memory alloy - an experimental animal study. Rofo Fortschr Geb Rontgenstr Neuen Bildgeb Verfahr 1990;152(6):698–701.
29. Rabkin JE, Germashev V. The Rabkin nitinol coil stent: a five-year experience. In: Castaneda-Zuniga WR, Tadavarthy SM, editors. Interventional Radiology, 2nd ed. Williams & Wilkins; 1992. p 576–581.
30. Kikuchi Y, Graves VB, Strother CM, McDermott JC, Babel SG, Crummy AB. A new guidewire with kink-resistant core and low-friction coating. Cardiovasc Intervent Radiol 1989;12(2): 107–109.
31. Kauffman GB, Mayo I. The story of Nitinol: the serendipitous discovery of the memory metal and its applications. Chem Educator 1997;2(2):S1430–4171; http://chemeducator.org/bibs/0002002/00020111.htm, Feb.2. 2005.
32. Castleman LS, Motzkin SM. The Biocompatibility of Nititnol. In: Williams DF, editor. Biocompatibility of Clinical Implant Materials volume I. CRC Press; 1981. p 129–154.
33. Cross WB, Karitos AH, Wasilewski RJ. Nitinol characterization study. NASA CR-1433, National Aeronautics and Space Administration, Houston. 1969.
34. Buehler WJ, Wiley RC. TiNi-ductile intermetallic compound. Trans ASM 1962;55:269–276.
35. Otsuka K, Kakeshita T. Science and Technology of Shape memory Alloys: New Developments. MRS Bull 2002;27(2): 91–100.
36. Kim YK. The study of the shape recovery temperature change of cold-worked nickel-titanium alloys. Inje J 1994;10(1): 341–352.
37. Aboelfotoh MO, Aboelfotoh HA, Washburn J. Observations of pretransformation lattice instability in near equiatomic NiTi alloy. J Appl Phys 49(10): 1978; 5230–5232.
38. Ling HC, Kaplow R. Phase transitions and shape memory in NiTi. Metal Trans A 1980;11A:77–83.
39. Chandra K, Purdy GR. Obseravation of thin crystals of TiNi in premartensite states. J Appl Phys 19(5): 1968; 2176–2181.
40. Shin SH. The study of heat-treatment temperature effect on hardness and compressional properties of nickel-titanium alloy. Master. dissertation, Inje University, Korea, Dec. 1998.
41. Burkart MW, Read TA. Trans Met Soc AIME 1953;197:1516. (as cited in Krishnan RV, Delaey L, Tas H, Warlimont H. Review Thermoelasticity, pesudoelasticity and the memory effects associated with martensitic transformations Part 2 The macroscopic mechanical behaviour. J Mat Sci 1974;9:1536–1544.
42. Andreasen GF, Fahl JL. Alloys, Shape Memory. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation. Volume 1, New York: Wiley-Interscience; 1988. p 15–20.
43. Thierry B, Merhi Y, Bilodeau L, Trepanier C, Tabrizian M. Nitinol versus stainless steel stents: acute thrombogenicity study in an ex vivo porcine model. Biomaterials 2002;23:2997–3005.
44. Moberly WJ, Melton KN. Ni–Ti–Cu shape memory alloys. Engineering Aspects of Shape Memory Alloys. London: Butterworth-Heinemann; 1990. p 46–57.
45. Murakami Y, Asano N, Nakanishi N, Kachi S. Phase relation and kinetics of the transformations in Au–Cu–Zn thernary alloys. Jpn J Appl Phys 1967;6:1265–1271.
46. Gil FJ, Planell JA. Effect of copper addition on the superelastic behavior of Ni–Ti shape memory alloys for orthodontic applications. J Biomed Mater Res Appl Biomat 1999;48:682–688.
47. Goldstein D, Kabacoff L, Tydings J. Stress effects on Nitinol phase transformations. J Metals 1987;39(3):19–26.
48. Lee JH, Park JB, Andreasen GF, Lakes RS. Thermo mechanical study of Ni–Ti alloys. J Biomed Mater Res 1988;22: 573–588.
49. Kim YK. The Grain size distribution study of heat treated Ni–Ti alloy. Inje J 1993;9(2):857–868.
50. Differential Scanning Calorimetry, Dept. of Polymer Science, University of Southern Mississippi. Available at http://www.psrc.usm.edu/macrog/dsc.htm. Accessed Feb. 8. 2005.
51. Harrison JD. Measurable changes concomitant with the shape memory effect transformation, In: Duerig TW, Melton KN, Stoeckel D, Wayman CM, editors. Engineering aspects of shape memory alloys. Butterworth-Heinemann; 1990. p 106–111.
52. Wang FE, DeSavage BF, Buehler WJ, Hosler WR. The irreversible critical range in the TiNi transition. J Appl Phys 1968;39(5):2166–2175.
53. Kim YK. unpublished experimental data (ykkimbme.inje.ac.kr).
54. Cutright DE, Bhaskar SN, Perez B, Johnson RM, Cowan GS, Jr. Tissue reaction to nitinol wire alloy. Oral Surg 1973;35(4): 578–584.
55. Castleman LS, Motzkin SM, Alicandri FP, Bonawit VL. Biocompatibility of Nitinol alloy as an implant material. J Biomed Mater Res 1976;10:695–731.
56. Castleman LS, Motzkin SM. The biocompatibility of Nitinol. In: Williams DF, editor. Biomcompatibility of Clinical Implant Materials. Volume 1, CRC Press; 1981. p 129–154.
57. Park JB. Metallic implant materials. Biomaterials Science and Engineering. Joon Bu Park: Plenum Press; 1984. p 193–233.
58. Ryhaenen J. Biocompatibility Evaluation of Nickel-Titanium Shape Memory Metal Alloy. Academic Dissertation, University hospital of Oulu, on May 7th, 1999.
59. Trepanier C, Tabrizian M, Yahia LH, Bilodeau L, Piron DL. Effect of modification of oxide layer on NiTi stent corrosion resistance. J Biomed Mater Res (Appl Biomater) 1998;43:433–440.
60. Kammula RG, Morris JM. Considerations for the Biocompatibility Evaluation of Medical Devices. Available at http://www.devicelink.com/mddi/archive/01/05/008.html. Medical Device Link. Accessed Feb. 2. 2005.

61. Austian J. Toxicological evaluation of biomaterials: primary acute toxicity screening program. Artif Organs 1977;1:53–60.
62. Mears DC. The use of dissimilar metals in surgery. J Biomed Mater Res 1975;9:133–148.
63. Williams DF. Future prospects for biomaterials. Biomed Eng 1975;10:206–218.
64. Thierry B, Tabrizian M, Trepanier C, Savadogo O, Yahia LH. Effect of surface treatment and sterilization processes on the corrosion behavior of NiTi shape memory alloy. J Biomed Mater Res 2000;51:685–693.
65. Wever DJ, Veldhuizen AG, Sanders MM, Schakenraad JM, Horn V, Jr. Cytotoxic, allergic and genotoxic activity of a nickel–titanium alloy. Biomaterials 1997;18(16):1115–1120.
66. Motzkin SM, Castleman LS, Szablowski W, Bonawit VL, Alicandri FP, Johnson AA. Evaluation of nitinol compatibility by cell culture. Proc. 4th New England Bioeng Conf. New Haven: Yale University; 1976. p 301.
67. Ryhanen J, Kallioinen M, Tuukkanen J, Junila J, Niemela E, Sandvik P, Serlo W. In vivo biocompatibility evaluation of nickel-titanium shape memory metal alloy: muscle and perineural tissue responses and encapsule membrane thickness. J Biomed Mater Res 1998;41(3):481–488.
68. Castleman LS. Biocompatibility of Nitinol alloy as an implant material. Proceeding of the 5th Annual International Biomaterials Symposium. Clemson (SC): Clemson University; 1973.
69. Shih CC, Lin SJ, Chen YL, Su YY, Lai ST, Wu GJ, Kwok CF, Chung KH. The cytotoxicity of corrosion products of nitinol stent wire on cultured smooth muscle cells. J Biomed Mater Res 2000;52:395–403.
70. Andreasen GF. Method and system for orthodontic moving of teeth. US pat. 4,037,324. 1977.
71. Andreasen GF. A clinical trial of alignment of teeth using a 0.019 inch thermal nitinol wire with a transition temperature range between 31 °C and 45 °C. Am J Orthod 1980;78(5):528–537.
72. Andreasen GF, Morrow RE. Laboratory and clinical analyses of nitinol wire. Am J Orthod 1978;73:142–151.
73. Miyazaki S. Medical and dental applications of shape memory alloys. In: Otsuka K, Wayman CM, editor. Shape memory materials. Cambridge University Press; 1998. p 267–281.
74. Mooney MR, Mooney JF, Pedersen WR, Goldenberg IF, Gobel FL. The Ultra-Select guidewire: a new nitinol guidewire for coronary angioplasty. J Invasive Cardiol 1991;3(5):242–245.
75. Kim YK. unpublished photograph (ykkimbme.inje.ac.kr).
76. Kim YK. unpublished photograph (ykkimbme.inje.ac.kr).
77. Kim YK. unpublished photograph (ykkimbme.inje.ac.kr).
78. Palmaz JC, Kopp DT, Hayashi H, Schatz RA, Hunter G, Tio FO, Garcia O, Alvarado R, Rees C, Thomas SC. Normal and stenotic renal arteries: experimental balloon-expandable intraluminal stenting. Radiology 1987;164(3):705–708.
79. Kousbroek R. Shape memory alloys, In: Ducheyne P, editor. Metal and Ceramic Biomaterials, Volume II: Strength and Surface. CRC Press; Chapt. 3. 1984. p 63–90.
80. Schemerling MA, Wilkov MA, Sanders AE, Woosley JE. Using the shape recovery of Nitinol in the Harrington rod treatment of scoliosis. J Biomed Mater Res 1976;10:879–892.
81. Musialek J, Filip P, Nieslanik J. Titanium-nickel shape memory clamps in small bone surgery. Arch Orthop Trauma Surg 1998;117:341–344.

Reading List

Castleman LS, Motzkin SM. The Biocompatibility of Nititnol. In: Williams DF, editor. Biocompatibility of Clinical Implant Materials volume I. CRC Press; 1981. p 129–154.
Kousbrock R. Shape Memory Alloys, In: Ducheyne P, Hastings GW, editors. Metal and Ceramic Biomaterials Volume II. CRC Press; 1984. p 63–90.
Otsuka K, Wayman CM, editors. Shape Memory Materials. Cambridge University Press; 1998.
Nishiyama Z, Fine ME, Meshii M, Wayman CM, editors. Martensitic Transformation. London: Academic Press; 1978.
Jena AK, Chaturvedi MC. Phase Transformation in Materials. New Jersey: Prentice Hall; 1992.
Duerig TW, Melton KN, Stockel D, Wayman CM. Engineering Aspects of Shape Memory Alloys. London: Butterworth-Heinemann; 1990.
Wayman CM, Bhadeshia HKDH. Phase Transformations, Nondiffusive. Cahn RW, Hassen P, editors. Physical Metallurgy. 4th ed. Amsterdam: North-Holland; 1996. p 1507–1554.
Filip P. Titanium–Nickel Shape Memory Alloys in Medical Applications. In: Brunette DM, Tengvall P, Textor M, Thomsen P, editors. Titanium in Medicine. Berlin: Springer; 2001. p 53–86.
Perkins J, editor. Shape Memory Effects in Alloys. New York: Plenum Press; 1975.

See also HIP JOINTS, ARTIFICIAL; SPINAL IMPLANTS.

AMBULATORY MONITORING

HAIBO WANG
AHMAD ELSHARYDAH
RANDALL CORK
JAMES FRAZIER
Louisiana State University

INTRODUCTION

Due to advances in technology, especially computer sciences, ambulatory monitoring with medical instruments has increasingly become an important tool in the diagnosis of some diseases and medical conditions. Some devices used or in development for current clinical practice are shown in Table 1.

The ideal device for ambulatory monitoring should be consistently sensitive, accurate, lightweight, noninvasive, and easy to use. The Holter monitor is a popular device for ambulatory monitoring. Therefore, this article will start with a discussion of the Holter monitor.

AMBULATORY MONITORING WITH A HOLTER DEVICE

A Holter monitor is a continuous recording of a patient's ecocardiogram (ECG) for 24 h as shown in Fig. 1. It was named in honor of Norman J. Holter for his contribution in creating the world's first ambulatory ECG monitor in 1963 (1). Since it can be worn during the patient's regular daily activities, it helps the physician correlate symptoms of dizziness, palpitations, and syncope with intermittent

Table 1. Current Devices for Ambulatory Monitoring

Devices	Uses
Holter monitoring	Cardiac arrhythmia and ischemia
Ambulatory BP monitoring	Hypertension and hypotension
Ambulatory glucose monitoring	Hyperglycemia and hypoglycemia

Figure 1. Holter monitor.

cardiac arrhythmias. When compared with the ECG, which lasts < 1 min. The Holter monitor is more likely to detect abnormal heart rhythm. It can also help evaluate the patient's ECG during episodes of chest pain due to cardiac ischemia. The common clinical applications for Holter monitor are summarized in Table 2 (2,3).

The basic components of a Holter monitor include at least a portable ECG recorder and a Holter analyzer (scanner). Functional characteristics of both components have improved dramatically since the first Holter monitor was developed > 40 years ago.

Portable ECG Recorder

The recorder is a compact, light-weight device used to record an ambulatory patient's three or more ECG leads, typically for 24 h for dysrhythmia or ischemia detection. There are two types of the recorder available on the market: the classical cassette type (tape) or the newer digital type (flash memory card). The cassette recorder uses magnetic tape to record ECG information. The tape needs to be sent to the physician's office for analysis with a scanner to produce a patient's report. The problems with this type of recorder are its limited memory for ECG recording, its inability to transmit ECG information to the service center digitally, and its difficulty in processing the information with computer software. Therefore, it normally takes days to produce a monitoring report for a patient.

The newer digital recorder has an increased memory compared to the classical cassette type, making it possible to extend the monitoring time beyond 24 h if indicated. More importantly, the recorded signs are digital, which can be transmitted to the service center or on-call physician by digital transmission system via a phone line, email, or wireless technology, and can be processed rapidly with computer software. Therefore, the patient's report will be available for the patient's physician much sooner. If indicated, treatment can be started without delay. Additionally, a patient event button has been incorporated in some of newer recorders to allow correlation of symptoms and activity with ECG changes to obtain more clinical data useful for making a correct clinical diagnosis.

Another new development is the Cardiac Event Monitoring (CEM), which is similar to Holter monitor for recording an ambulatory patient's ECG. The difference between them is that the CEM is an event-triggered device, only recording the patient's ECG when they experiences a detectable symptoms (4). As a result, the CEM makes prolonged monitoring possible even with a limited recorder memory.

Table 2. Clinical Application of the Holter Monitor

1. Evaluation of symptomatic events: dizziness, syncope, heart palpitations, fatigue, chest pain, shortness of breath, episodic diaphoresis.
2. Detection of asymptomatic dysrhythmia: asymptomatic atrial fibrillation.
3. Evaluation of rate, rhythm or ECG interval changes during drug therapy.
4. Evaluation for specific clinical situations: postmyocardial infarction, postcoronary bypass surgery, postpercutaneous transluminal coronary angioplasty, postpacemaker implant, first or second degree heart block, possible pacer malfunction, automatic implanted defibrillator functions.
5. Evaluation of ECG changes during specific activities.

Holter Analyzer (Scanner)

There are different types of Holter analyzer systems currently available. The original system for analyzing the Holter cassette was a scanner with manual observer detection. With manual observer detection, the trained technician watches for audiovisual clues of abnormal beats while playing the tape back at 60–120 times real time. The process is time consuming. It requires a skilled technician who can withstand high boredom and fatigue levels to minimize possible human error rate.

The modern Holter analyzer system has been revolutionized due to the application of computer technologies. It is available in a variety of options, such as auto analysis and complete editing capabilities. Some of the newer systems provide easy-to-use professional features that allow rapid review of recorded information, producing a fast, accurate report.

Future Development

Ambulatory Holter monitoring is a valuable tool in patient care and is becoming more and more popular. Integration of computer technology, digital technology, wireless technology, and nanotechnology may lead to an ideal Holter device, which is minimal in size and weight, user-friendly, noninvasive, sensitive and accurate, wirelessly connected to a physician on-call center, and with automatic data analysis capacity. Newer analysis techniques involving fuzzy logic, neural networks and genetic algorithms will also enhance automatic detection of abnormal ECG. Hopefully, such an ideal ambulatory Holter monitor will be available in the near future.

AMBULATORY BLOOD PRESSURE MONITORING

Introduction

An ambulatory blood pressure monitoring device is a noninvasive instrument used to measure a patient's 24 h ambulatory blood pressure as shown in Fig. 2. The first device was developed by Hinman in 1962 (5). He used a

Figure 2. Ambulatory blood pressure monitoring device.

microphone placed over the brachial artery distal to a compression cuff and a magnetic tape recorder for recording of onset and disappearance of Korotkoff sounds. It weighed ~ 2.5 kg and was obviously inconvenient for an ambulatory patient to use. The first fully automatic device was developed, using compressed carbon dioxide to inflate the cuff. An electronic pump was introduced later and automatic data recording systems have been used since 1979.

Since then, the techniques for ambulatory blood pressure monitoring have been improved significantly. The modern device is light-weighted, compact in size, accurate, and automated in nature. It can be belt-worn and battery powered. The newest generation available in the current market is fully automatic, microprocessor-controlled, digitalized in memory, and extremely light weight (< 500 g).

Basic Techniques

The techniques for ambulatory blood pressure monitoring include auscultation, cuff oscillometry, and volume oscillometry.

Auscultation is a technique based on detection of onset and disappearance of Korotkoff sounds via a piezoelectric microphone taped over an artery distal to a deflating compression cuff. The Korotkoff sound is produced by turbulent flow while arterial blood flows through a segment of artery narrowed by a blood pressure cuff. The pressure at the onset of sound corresponds to systolic blood pressure, and at the disappearance of the sound to diastolic pressure. The advantage of this technique is simplicity, but the device is sensitive to background noise. This technique may also underestimate systolic pressure due to his flow dependency.

Cuff oscillometry is a technique based on detection of cuff pressure oscillations or vibrations to calculate systolic and diastolic values using an algorithmic approach. The systolic pressure corresponds to the cuff pressure at which oscillations first increase, and the diastolic pressure corresponds to the cuff pressure at which oscillations cease to decrease. The endpoints are estimated by analysis of oscillation amplitudes and cuff pressures. Different algorithms are used by different manufacturers, which may result in variability among different devices. This technique is insensitive to background noise, but arm movement may cause an errant reading. It may overestimate systolic pressure because of transmitted cuff pressure oscillations.

Volumetric oscillometry is a technique based on detection of finger volume pulsations under a cuff. The pressures are estimated as the cuff pressures at which finger volume oscillations begin (systolic pressure) and become maximal (mean pressure). Diastolic pressure is then derived from the known systolic and mean pressures. One problem with this technique is that this finger pressure may have a variable relationship to the brachial pressure. Another problem is that the technique cannot directly assess diastolic pressure.

Despite some problems associated with the mentioned techniques, their accuracy has been confirmed by validation testing using mercury sphygmomanometry and intraarterial measurement. The discrepancy is generally < 5 mmHg (0.399 kPa) between ambulatory devices and readings taken by trained professionals.

Patients are advised to wear the monitor for a period of 24 h, preferably during a normal working day. The monitor is preprogrammed to measure and record blood pressure at certain time intervals, preferably every 15–20 min during daytime hours and every 20–30 min during nighttime hours. Patients are also advised to document their activity during the testing period for assessment of any stress-related blood pressure.

The monitoring device consists of a small central unit and an attached cuff. The central unit contains a pump for cuff inflation and deflation, and the memory device, such as tape or digital chip, for recording. The time intervals between the measurements, maximal and minimal inflation pressures, and deflation rate are programmable according to the physician's order. The recording pressures can be retrieved from the tape or memory chip for analysis. Due to recent applications of digital technology and advanced software programs, a large amount of data can be stored in a small chip, and analysis can also be done automatically to generate a patient's report for the physician's use. A complete patient's report normally contains all blood pressure readings over a 24 h period, heart rates, mean arterial pressures, and statistic summaries for daytime, nighttime, and 24 h periods.

New Clinical Concepts Related to Ambulatory Blood Pressure Monitoring

A few new considerations related to ambulatory blood pressure monitoring have emerged. These include blood pressure load, pressure dipping, pressure variability, and white-coat hypertension. Health professionals need to understand these concepts in order to properly interpret or use data collected from monitoring.

Blood Pressure Load. This is defined as the proportion of the 24 h pressure recordings above the thresholds for waking and sleep blood pressure. The threshold commonly used for estimating the pressure load during waking hours is 140/90 and 120/80 mmHg (15.99/10.66 kPa) during sleep. Blood pressure load is helpful in the diagnosis of hypertension and in the prediction of end-organ damage. It has been considered closely correlated with left ventricle

hypertrophy. It has been reported that the incidence of left ventricular hypertrophy is $\sim 90\%$ in untreated patients with systolic blood pressure loads $> 50\%$, and $\sim 70\%$ with diastolic blood pressure loads $< 40\%$ (6,7).

Dipping and Circadian Blood Pressure Variability. Dipping is a term used to describe the circadian blood pressure variation during 24 h ambulatory blood pressure monitoring. In normotensive patients there is circadian blood pressure variability. Typically, the peak blood pressures occur around 6 a.m., and then taper to lower levels during the evening hours and further at night with the lowest levels between 2 and 4 a.m.. A patient whose blood pressure drops by at least 10% during sleep is considered normal (a dipper), and by $< 10\%$ abnormal (nondipper). In comparison to dippers, nondippers have been reported associated with higher prevalence of left ventricular hypertrophy, albuminuria, peripheral arterial changes, and cerebral lacunae. Nondippers have also been reported to have increased cardiovascular mortality rates (8).

White-Coat Hypertension. This is a condition in which blood pressure is persistently elevated in the presence of a doctor, but falls to normal levels when the patient leaves the medical facilities. Measurement by a nurse or trained nonmedical staff may reduce this effect. Because decisions regarding treating hypertension are usually made on the basis of isolated office blood pressure reading, a doctor may incorrectly diagnose this group of patients as sustained hypertension and prematurely start the therapy. This phenomenon has been reported in 15–35% of patients currently diagnosed and treated as hypertensive. However, white-coat hypertension can be easily detected by either ambulatory blood pressure monitoring or self-monitoring at home. It may or may not be benign, requiring definitive outcome studies to rule out any end-organ damages. It also requires continued surveillance by self-monitoring at home and repeat ambulatory blood pressure monitoring every 1–2 years (9,10).

Interpretation of Ambulatory Blood Pressure Profile

Normal ambulatory blood pressure values for adults are currently defined to be $< 135/85$ mmHg (17.99/11.33 kPa) during the day, $< 120/75$ mmHg (15.99/9.99 kPa) during the night, and $< 130/80$ mmHg (17.33/10.66 kPa) over 24 h. Daytime and night time blood pressure loads should be less 20% above normal values. Mean day-time and nighttime (sleep) blood pressure measurements should differ by at least 10%. The ambulatory blood pressure profile should also be inspected in relation to diary data and time of drug therapy.

Indications of Ambulatory Blood Pressure Monitoring

Although ambulatory blood pressure monitoring was originally developed as a research tool, it has widely been applied in clinical practice to help diagnose and manage hypertensive patients. It is indicated to rule out white-coat hypertension, to evaluate drug-resistant hypertension, to assess symptomatic hypertension or hypotension, to diagnose hypertension in pregnancy, and to assess adequacy of blood pressure control in patients at high risk of cardiovascular diseases.

White-Coat Hypertension. Office-based blood pressure measurement cannot differentiate sustained hypertension from white-coat hypertension. Historical appraisal and review of self-recorded blood pressures may aid in identification of patients with white-coat hypertension. However, ambulatory blood pressure monitoring is more effective in this clinical scenario to rule out white-coat hypertension. Recognition and proper management of patients with white-coat hypertension may result in a reduction in medication use and eliminate related cost and side effects. Although white coat hypertension may be a prehypertensive state and can eventually evolve to sustained hypertension, data collected from ambulatory blood pressure monitoring suggest, patients with white coat hypertension who maintain low ambulatory blood pressures (< 130–$135/80$ mmHg) have a low cardiovascular risk status and no demonstrable end-organ damage (11).

Drug-Resistant Hypertension. Drug resistant hypertension is defined as a condition when adequate blood pressure control ($< 140/90$ mmHg) (18.66/11.99 kPa) cannot be achieved despite the use of appropriately combined antihypertensive therapies in proper dosages for a sufficient duration. Ambulatory blood pressure monitoring helps evaluate whether additional therapy is needed. The causes include true drug-resistant hypertension as well as other conditions such as superimposition of white-coat hypertension on existing hypertension, patient's noncompliance, pseudohypertension secondary to brachial artery calcification, and sleep apnea and other sleep disorders. Ambulatory blood pressure monitoring can help differentiate the true drug resistant hypertension from the above-mentioned conditions (12).

Episodic Hypertention. A single office-based measurement of blood pressure may or may not detect episodic hypertension as in pheochromocytoma. In this clinical scenario the 24 h ambulatory blood pressure monitoring is a useful diagnostic tool. It is indicated if a patient's symptoms or signs are suggestive of episodic hypertension (13).

Borderline or Labile Hypertension. Patients with borderline hypertension often demonstrate only some (but not all) elevated blood pressure readings in office-based measurement, 24 h ambulatory blood pressure monitoring can benefit these patients and provide a useful diagnostic information for physician's use (14).

Hypertension with End-Organ Damage. Patients who exhibit worsening of end-organ damage may suggest inadequate 24 h blood pressure control. Occasionally, those patients may demonstrate adequate blood pressure control based on the office-based measurements. In this condition, a 24 h blood pressure monitoring is needed to rule out inadequate blood pressure control, which is associated with worsening of end-organ damage (15).

Hypentensive Patients with High Risk of Cardiovascular Events. Some hypertensive patients are at particularly high

risk of cardiovascular events, such as those with diabetes and/or past stroke. Those patients require rigorous blood pressure control over 24 h. Ambulatory blood pressure monitoring can be applied to assess the 24 h control (15).

Suspected Syncope or Orthostatic Hypotension. Transient hypotensive episodes and syncope are difficult to assess with the office-based blood pressure measurements, but are readily recorded with ambulatory blood pressure monitoring. Therefore, if symptoms and signs are suggestive of syncope or orthostatic hypertension, patients can benefit from 24 h blood pressure monitoring, especially in conjuction with Holter monitoring (15).

Hypertension in Pregnancy. About 10% of pregnancies may be complicated by hypertension. At the same time, white-coat hypertension may affect up to 30% of patients. It is important to differentiate true hypertension in pregnancy from white-coat hypertension, to avoid unwarranted hospitalizations or medication use. In this clinical scenario, ambulatory blood pressure monitoring would help to rule out white-coat hypertension and identify pregnancy-induced hypertension (16).

Clinical Research. Since ambulatory blood pressure monitoring can provide more samples of blood pressure measurements, data from this device is therefore much more statistically significant than a single isolated office-based reading. Therefore, statistical significance of clinical studies can possibly be achieved with smaller numbers of patients. This is very important for the efficient study of new therapeutic agents (17).

Limitations of Ambulatory Blood Pressure Monitoring

Although ambulatory blood pressure monitoring has been proved useful in the diagnosis and management of hypertension, the technology remains underused secondary to lack of experience in interpretation of results, unfamiliarity with devices, and some economic issues. Adequate staff training, regular calibration of devices, and good quality control are required. The patient's diary of daily activities and time of drug treatment are also needed for proper data analysis and interpretation.

Future Development

Like any other ambulatory device, an ideal noninvasive ambulatory blood pressure monitoring device should be user-friendly, light-weight, compact in size, digitalized for automated data management, and low in cost. Application of newer technologies will make such devices available, hopefully, in the near future.

AMBULATORY BLOOD GLUCOSE MONITORING

Introduction

Diabetes is one of most common diseases suffered by millions of people around the world. It is essential to monitor blood glucose to ensure overall adequate blood glucose control. Traditional standard blood glucose monitoring devices require invasive blood samplings and are therefore unsuitable for ambulatory blood glucose monitoring. Development of minimally invasive or noninvasive ambulatory glucose monitoring devices that provide accurate, near-continuous measurements of blood glucose level have the potential to improve diabetes care significantly. Such devices will provide information on blood glucose levels, as well as rate and direction of change, which can be displayed to patients in real-time and be stored for later analysis by physicians. Guardian RT system recently developed by Medtronic MiniMed is an example (Fig. 3). It provides continuous real-time glucose readings around the clock. Due to the huge market potential, many biomedical and medical instrument companies are developing similar devices for ambulatory glucose monitoring. Several innovative devices have recently been unveiled; many more are still in development. It is expected that some of them will be eventually U. S. Food and Drug Administration (FDA) approved as a replacement for standard blood glucose monitors, providing patients with a new option for long-term, daily monitors in the near future. The FDA is concerned about the accuracy of ambulatory continuous glucose monitoring devices when compared to the accuracy of standard monitoring devices. This issue will be eventually eliminated as related technologies become more and more mature. Technically, a typical ambulatory glucose monitoring device consists of a glucose sensor to measure glucose levels and a memory chip to record data information.

Figure 3. Ambulatory glucose monitoring with guardian real time system (Medtronic MiniMed).

Glucose Sensors

The glucose sensors for ambulatory glucose monitoring devices are either minimally invasive or completely noninvasive. A variety of technologies have emerged over the past decade aiming at development of ideal glucose sensors suitable for ambulatory monitoring.

A typical minimally invasive ambulatory continuous glucose sensor is a subcutaneous device developed by Minimed, Inc. (18). The sensor is designed to be inserted into a patient's abdominal subcutaneous tissue. It measures glucose levels every 10 s and records means > 5 min intervals. The technology involves measurement of glucose levels of

interstitial fluid via the subcutaneous sensor. The blood glucose levels are then derived from the measured interstitial fluid glucose levels. The detection mechanism involves use of a low fluorescence molecule. Electrons are transferred from one part of the molecule to another when excited by light. This prevents bright fluorescence from occurring (19). When bound to glucose, the molecule prevents the electrons from interfering with fluorescence, and the molecule becomes a bright fluorescent emitter. Therefore, the glucose levels can be determined based on the brightness of fluorescence. The glucose information will be transmitted from the sensor to a watch-like device worn on the wrist. Using this type of sensor, two devices have been developed by the company. One is a device that can be worn by the patient for a few days to record the glucose levels for the physician's analysis. The other is a device that can alert patients of impending hyperglycemia or hypoglycemia if the glucose levels go beyond the physician's predetermined upper and lower limits. The sensor can also work in conjunction with an implanted insulin pump, creating a "biomechanical" or artificial pancreas in response to the change at the glucose levels (20). It is predictable that such a biomechanical pancreas will eventually benefit millions of diabetic patients whose glucose control is dependant on insulin.

Complete noninvasive sensors for ambulatory glucose monitoring are even more attractive since they do not need any blood or interstitial samples to determine glucose levels. Several such sensors have recently been developed based on different technologies. For example, a glucose sensor that can be worn like a wristwatch has been developed by Pendragon Medical AG (Zurich, Switzerland). This sensor can continuously monitor blood glucose level without the need for a blood sample. It is based on impedance spectroscopy technology (21). The principle of this technology relates to the fact that blood glucose changes produce significant conductivity changes, causing electric polarization of cell membranes. At the same time, the sensor generates an electronic field that fluctuates according to the electrical conductivity of the body. A micro antenna in the sensor then detects these changes and correlates them with changes in serum glucose. With this technology, blood glucose levels can be monitored noninvasively in real time. Another promising noninvasive sensor is based on the possibility of measuring glucose by detecting small changes in the retinal capillaries. By scanning the retinal microvasculature, the sensor can directly measure glucose levels in aqueous humor using a reflectometer. Recently, a plastic thin sensor, which can be worn like a contact lens, has been innovated (22,23). The sensor changes its color based on the concentration of glucose, from red, which indicates dangerously low glucose levels, to violet, which indicates dangerously high glucose concentrations. When glucose concentration is normal, the sensor is green. Integration of the sensor material into commercial contact lenses may also be possible with this technology.

Memory Chips

Memory chips are used to record glucose data information for later use by the physician. The digital chips have many advantages, such as compact size, large memory, easy data transmission via wire or wireless, and possible autoanalysis with computer software. Patients can also upload their glucose data from digital memory chips to web-based data management systems, allowing diabetic patients and their health care providers to analyze and communicate glucose information using the internet.

Significances of Ambulatory Glucose Monitoring

Ambulatory glucose monitoring can provide continuous data on blood glucose levels. Such data can improve diabetic care by enabling patients to adjust insulin delivery according to the rate and direction of blood glucose change, and by warning of impending hypoglycemia and hyperglycemia. Doctors can use ambulatory glucose monitoring to help diagnose problematic cases, fine-tune medications, and get tighter control of blood glucose levels for high risk patients. Obviously, the monitoring will improve overall blood glucose control, reducing short-term adverse complications and delaying onset of long-term serious complications, such as end-stage renal disease, heart attack, blindness, stroke, neuropathy, and lower extremity amputation.

In addition, continuous ambulatory glucose monitoring is a key step toward the development of artificial pancreas, which could deliver insulin automatically in response to blood glucose levels. It is expected that such an artificial pancreas would greatly benefit many diabetic patients and provide them new hope for better quality of life.

Future Development

Although many continuous ambulatory glucose monitoring devices are still in the stage of clinical trials, there is little doubt as to the value of the devices in management of diabetic patients. It is expected that millions of diabetic patients will be benefited once such devices are widely available. At the same time, introduction of more and more new devices highlights the need for careful evaluation to ensure accuracy and reliability. Cooperation between the manufacturers and physicians to fine-tune the technology will eventually lead to approval of the devices by the FDA to replace traditional invasive standard glucose monitoring. Technology for continuous ambulatory glucose monitoring is also required to make an artificial pancreas, which would offer great hope for millions of patients with diabetes.

CONCLUSION

Ambulatory monitoring has increasingly provided a powerful alternative tool to diagnose and manage some diseases. Continuous advancement in a variety of technologies provides more and more innovative ambulatory devices to serve the patients' need. Applications of information technology and specialized software tools make autotransmission and autoanalysis of ambulatory monitoring data possible. Clinicians will be able to monitor their ambulatory patients distantly without a hospital or office visits. In addition, integration of the technology of continuous ambulatory monitoring with an implantable automatic therapeutic pump may create a biomechanical system in response to specific abnormal changes. The artificial pancreas currently in development is a typical example for

such hybrid devices. Such devices will be available in the market in the near future.

BIBLIOGRAPHY

Cited References

1. Holter NJ. New method for heart studies: Continuous electrocardiography of active subjects over long periods is now practical. Science 1961;134:1214–1220.
2. Heilbron EL. Advances in modern electrocardiographic equipment for long-term ambulatory monitoring. Card Electrophysiol Rev 2002;6(3):185–189.
3. Kadish AH, et al. ACC/AHA clinical competence statement on electrocardiography and ambulatory electrocardiography: a report of the ACC/AHA/ACPASIM task force on clinical competence. Circulation 2001;104:3169–3178.
4. Kinlay S, et al. Event recorders yield more diagnoses and are more cost-effective than 48 hour Holter monitoring in patients with palpitations. Ann Intern Med 1996;124:16–20.
5. Hinman AT, Engel BT, Bickford AF. Portable blood pressure recorder accuracy and preliminary use in evaluation intradaily variations in pressure. Am Heart J 1962;63:663–668.
6. Zachariah PK, et al. Blood pressure load: A better determinant of hypertension. Mayo Clin Proc 1998;63:1085–1091.
7. White WB, Dey HM, Schulman P. Assessment of the daily blood pressure load as a determinant of cardiac function in patients with mild-to-moderate hypertension. Am Heart J 1989;118:782–795.
8. Pickering TG. The clinical significance of diurnal blood pressure variations: dippers and nondippers. Circulation 1990;81:700–702.
9. Verdecchia P, et al. White-coat hypertension: not guilty when correctly defined. Blood Press Monit 1998;3:147–152.
10. Pickering TG, et al. How common is white coat hypertension. Hypertension 1988;259:225–228.
11. Palatini P, et al. Target-organ damage in stage-1 hypertensive subjects with white coat and sustained hypertension: results from the HARVEST study. Hypertension 1998;31:57–63.
12. Brown MA, Buddle ML, Martin A. Is resistant hypertension really resistant ? Am J Hypertens 2001;14:1263–1269.
13. Myers MG, Haynes RB, Rabkin SW. Canadian hypertension society quidelines for ambulatory blood pressure monitoring. Am J Hypertens 1999;12:319–331.
14. Pickering T. for the American Society of Hypertension ad-hoc Panel. Recommendations for the use of home (self) and ambulatory blood pressure monitoring. Am J Hypertens 1996;9:1–11.
15. O'Brien E, et al. Use and interpretation of ambulatory blood pressure monitoring: recommendations of the British Hypertension Society. BMJ 2000;320:1128–1134.
16. Halligan A, et al. Twenty-four-hour ambulatory blood pressure measurement in a primigravid population. J Hypertens 1993;11:869–873.
17. Conway J, et al. The use of ambulatory blood pressure monitoring to improve the accuracy and reduce the numbers of subjects in the clinical trials of antihypertensive agents. J Clin Exper Hypertension 1986;8:1247–1249.
18. Cross TM, et al. Performance evaluation of the MinMed continuous glucose monitoring system during patient home use. Diab Technol Ther 2000;2:49–56.
19. Pickup JC, Shaw GS, Claremont DJ. *In vivo* molecular sensing in diabetes mellitus: an implantable glucose sensor with direct electron transfer. Diabetes 1989;32:213–217.
20. Jaremko J, Rorstad O. Advances toward the implantable artificial pancreas for treatment of diabetes. Diabs Care 1998;21:444–450.
21. Caduff A, et al. First human experiments with a novel noninvasive, non-optical continuous glucose monitoring system. Biosens Bioelecs 2003;19:209–217.
22. Badugu R, Lakowicz JR, Geddes CD. Ophthalmic glucose sensing: a novel monosaccharide sensing disposable and colorless contact lens. Analyst (England) 2004;129:516–521.
23. Badugu R, Lakowicz JR, Geddes CD. Ophthalmic glucose monitoring using disposable contact lenses—a review. J Fluoresc 2004;14:617–633.

See also ARRHYTHMIA ANALYSIS, AUTOMATED; BIOTELEMETRY; HOME HEALTH CARE DEVICES; PACEMAKERS.

ANALYTICAL METHODS, AUTOMATED

LAKSHMI RAMANATHAN
Mount Sinai Medical Center
LASZLO SARKOZI
Mount Sinai School of Medicine

INTRODUCTION

The chemical composition of blood, urine, spinal fluid, sweat, provides a wealth of information on the well being or illness of the individual. The presence, concentration, and activity of chemical constituents are indicators of various organ functions. Concentrations higher or lower than expected sometimes require immediate attention. Some of the reasons to analyze body fluids:

1. Screening of an apparently healthy population for unsuspected abnormalities.
2. Confirming or ruling out a diagnosis.
3. Monitoring changes during treatment, improvement of condition or lack of improvement.
4. Detecting or monitoring drug levels for diagnosis or maintenance of optimal therapeutic levels.

By the 1950s, demands of clinicians for laboratory tests increased rapidly. Classical methods of manual laboratory techniques could not keep up with these demands. The cost of performing large numbers of laboratory tests by manual methods became staggering and the response time was unacceptable.

The article in the first edition of this Encyclopedia published in 1988 describes the history of laboratory instrumentation during the previous three decades (1). Reviewing that long list of automated instruments, with the exception of a few, all became museum pieces. During the last 15 years the laboratory landscape changed drastically. In addition, new group of automated instruments were introduced during this period. They were developed to perform bedside or near patient testing, collectively called Point of Care Testing instruments. In this period in addition to new testing instruments, perianalytical instrumentation for specimen handling became available. Their combined result is increased productivity and reduction of manpower requirements, which became imperative due to increased cost of healthcare and dwindling resources.

This article will present some financial justification of these investments.

PATIENT PREPARATION, SPECIMEN COLLECTION, AND HANDLING

The prerequisites for accurate testing include proper patient preparation, specimen collection, and specimen handling. Blood specimens yield the most information about the clinical status of the patient though in many cases urine is the preferred sample. For specialized tests, other body fluids that include sweat and spinal fluid are used. When some tests, such as glucose and lipids, require fasting specimens, patients are prepared accordingly.

Common errors affecting all specimens include the following:

- Inaccurate and incomplete patient instructions prior to collection.
- Wrong container/tube used for the collection.
- Failure to label a specimen correctly.
- Insufficient amount of specimen to perform the test.
- Specimen leakage in transit due to failure to tighten specimen container lids.
- Interference by cellular elements of blood.

Phlebotomy techniques for blood collection have considerably improved with better gauge needles and vacuum tubes for collection. The collection tubes are color coded with different preservatives so that the proper container can be used for a particular analyte. The cells should be separated from the serum by centrifugation within 2 h of collection. Grossly or moderately hemolyzed specimens may be unsuitable for certain tests. If not separated from serum or plasma, blood cells metabolize glucose and produce a false decrease of ~5%/h in adults. The effect is much greater in neonates (2). If there is a delay in separating the cells from the serum, the blood should be collected in a gray top tube containing sodium fluoride as a preservative that inhibits glycolysis.

Urine collection is prone to errors as well, some of which include (3):

- Failure to obtain a clean catch specimen.
- Failure to obtain a complete 24 h collection/aliquot or other timed specimen.
- No preservative added if needed prior to the collection.

Once specimens are properly collected and received in the clinical laboratory, processing may include bar coding, centrifugation, aliquoting, testing and reporting of results.

AUTOMATED ANALYZERS

A large variety of instruments are available for the clinical chemistry laboratory. These may be classified in different ways based on the type of technology applied, the test menu, the manufacturer, and the intended application. Depending on the size of the laboratory, the level of automation varies. Clinical chemistry analyzers can be grouped according to throughput of tests and diversity of tests performed and by function, such as immunoassay analyzers, critical care blood gas analyzers, and urinalysis testing systems. Point of Care analyzers vary in terms of accuracy, diversity and menu selection.

Some of the features to consider while evaluating low or high volume analyzers are listed below:

- Test menu available on instrument:
  - Number of different measured assays onboard simultaneously.
  - Number of different assays programmed/calibrated at one time.
  - Number of user-defined (open) channels.
- Reagents:
  - Preparation of reagents if any.
  - Storage of reagents.
  - On board stability.
  - Bar-coding for inventory control.
- Specimen volume:
  - Minimum sample volume.
  - Dead volume.
- Instrument supplies:
  - Use of disposable cuvettes.
  - Washable/reusable cuvettes.
- Clot detection features along with quantitation of hemolysis and turbidity detection.
- Auto dilution capabilities of analyzer.
- Frequency of calibration.
- Quality control requirements.
- Stat capability.
- LIS interface.
- Maintenance procedures on instrument; anticipated downtime.
- Analyzer costs expressed in cost per reportable test.

Our goal is not to review every analyzer available on the market. We have chosen a few of the instruments–vendors. This is by no means endorsing any particular vendor, but merely discussing some of the most frequently utilized features or describing our personal experiences. The College of American Pathologists has provided excellent surveys of instruments and the reader is referred to those articles for more complete details (4).

CHEMISTRY ANALYZERS

Routine chemistry analyzers have broad menus capable of performing an average of 45 (20 to >70) different on board tests simultaneously, selected from an available menu of 26

Table 1. Automated Analyzers from Different Manufacturers

Instrument Type	Generic Menu	Vendor
Routine chemistry	Electrolytes, BUN, Glucose, Creatinine, Protein, Albumin, Lipids, Iron, Drugs of abuse, Therapeutic drug monitoring, etc.	Abbott, Bayer, Beckman Dade, J&J, Olympus, Roche
Immunoassays	Tumor markers, Cardiac markers, Anemia, B12, Folate and misc. Endocrine tests	Abbott, Bayer, Beckman, DPL, J&J, Olympus, Roche
Critical Care	Blood gases, Cooximetry, Electrolytes Ionized calcium, Lactate, Hematocrit	Abbott, Bayer, Instrumentation Lab, Nova, Radiometer, Roche

to >100 different analytes (5,6). Selection is based on test menu, analytic performance, cost (reagents, consumables and labor), instrument reliability (downtime etc.), throughput, and ease of use, customer support and robotic connectivity, if needed. Some automated analyzers from different manufacturers are listed in Table 1.

General Chemistry

Virtually all automated chemistry analyzers offer random access testing, multiple tests can be performed simultaneously and continuously. This is different from batch-mode instruments that perform a single test on a batch of samples loaded on the instrument (Abbott TDX and COBAS Bio). Many analyzers are so-called "open systems" that use reagents from either the instrument manufacturer or different vendors. The advantage of these systems being that the customer has a choice of reagent vendors and the reagent can be selected based on performance and cost.

An example of a closed system is a line of analyzers manufactured by Ortho Clinical Diagnostics. The Vitros 950 and the analyzers in this category use a unique, dry chemistry film-based technology developed by Kodak. The slide is a dry, multilayer, analytical element coated on a polyester support. A 10 μL drop of patient sample is deposited on the slide and is evenly distributed by the spreading layer to the underlying layers that contain the ingredients for a particular chemical reaction.

The reaction slide (Fig. 1.) for albumin shows the reactive ingredient is the dye (bromcresol green), which is in the reagent layer. The inactive ingredients that include polymeric beads, binders, buffer, and surfactants are in the spreading layer. When the specimen penetrates the reagent layer, the bromcresol green (BCG) diffuses to the spreading layer and binds to albumin from the sample. This binding results in a shift in wavelength of the reflectance maxima of the free dye. The color complex that forms is measured by reflectance spectrophotometry. The amount of albumin-bound dye is proportional to the concentration of albumin in the sample. Once the test is completed the slide is disposed into the waste container.

Some manufacturers close their system by labeling their individual reagent packs with unique barcodes, rejecting packs not distributed by them. Examples of "open systems" include analyzers manufactured by Olympus, Roche (Fig. 2.), Beckman, Dade and Abbott. Many instruments have both open and closed channels allowing greater flexibility in the use of reagents. In addition to diverse menus, open and closed channels, compatibility of analyzers

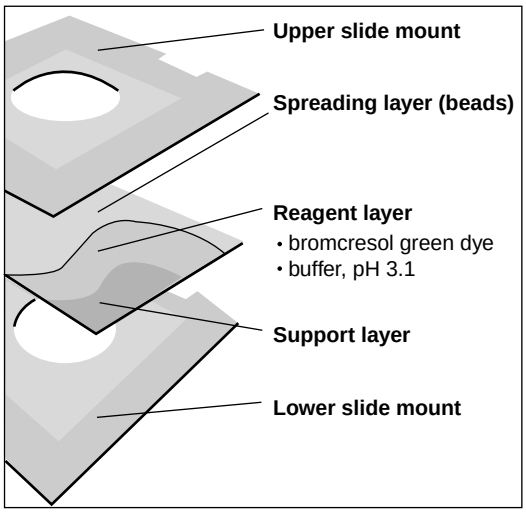


Figure 1. The Vitros 950 (J&J Diagnostics) slide is a dry, multilayer, analytical element coated on a polyester support. A drop of patient sample is deposited on the slide and is evenly distributed by the spreading layer to the underlying layers that contain the ingredients for a particular chemical reaction.

Figure 2. The Roche/Hitachi Modular™ analytic system has a theoretical throughput of 3500–5000 tests or 150–250 samples/h. They test 24 different analytes simultaneously with a total menu of > 140 available tests.

with perianalytical technology is becoming an important feature.

Perianalytical systems include front-end automation with specimen processing and aliquoting, track systems or other technologies to move specimens between instruments in the laboratory, and robots to place specimens on and remove them from the analyzers.

Immunoassay Analyzers

Immunoassay systems are presently the fastest growing areas of the clinical laboratory where advances in immunochemical methodology, signal detection systems, microcomputers and robotic processing are taking place at an accelerated pace (7). At present, manufacturers have high volume immunoassay analyzers that can be modularly integrated along with chemistry and hematology analyzers into fully automated laboratory systems. In addition, expanding menus of homogeneous immunoassays allow integration into many laboratories using "open reagent kits" designed for use on automated clinical chemistry analyzers.

One of the several analyzers in this category is the Bayer Advia Centaur (Fig. 3.)

Of the different enzyme immunoassays (EIA) available, only the two homogeneous methods, EMIT and CEDIA have been easily adapted to fully automated chemistry analyzers (8–11). The other EIAs require a separation step to remove excess reagent that will interfere with the quantitation of the analyte. Abbott uses a competitive assay involving a fluorescent-labeled antigen that competes for a limited number of sites on antigen specific antibody. The amount of analyte is inversely proportional to the amount of fluorescence polarization. Chemiluminescence technology is used in the Bayer ACS and Roche Elecsys systems combines very high sensitivity with low levels of background interference. Essentially, it involves a sandwich immunoassay direct chemiluminometric technology, which uses constant amounts of two antibodies. The first antibody in the Lite Reagent is a polyclonal goat anticompound antibody labelled with acridinium ester. The second antibody in the Solid Phase is a monoclonal mouse anticompound antibody, which is covalently coupled to paramagnetic particles. A direct relationship exists between the amount of compound present and the amount of relative light units (RlU) detected by the system (Table 2).

Critical Care Analyzers

Blood gas measurements performed on arterial, venous, and capillary whole blood includes electrolytes and other tests in addition to the gases. These tests are listed in Table 3.

The Nova CCX series combines blood gas measurements with co-oximetry, electrolytes, a metabolic panel and hematology on 50 μL of whole blood. Several blood gas analyzers are utilizing the concept of "Intelligent Quality Management" whereby the analyzers run controls automatically at specified time intervals set by the operator. If a particular analyte is not within the specified range, the analyzer will not report out any patient results on the questionable test. Selected blood gas and critical care analyzers are listed in Table 4.

The unique specimen and turnaround time requirements for blood gases have prevented the tests from

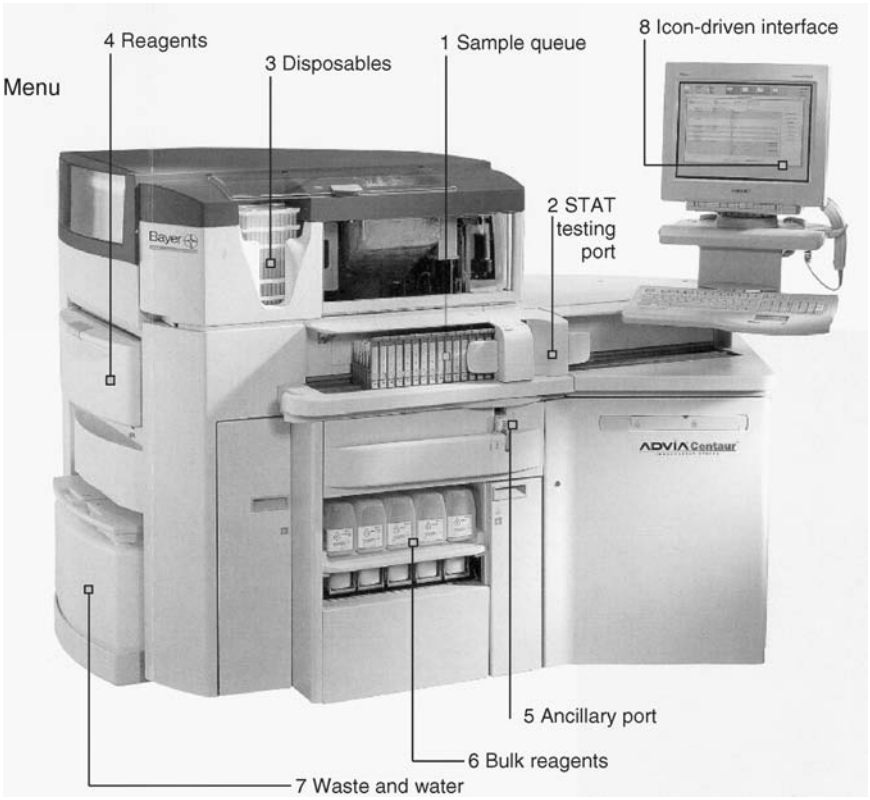


Figure 3. The Bayer Advia Centour system has large on-board capacity for reagents and supplies combined with automated maintenance and monitoring features streamline operations. Categories such as fertility, therapeutic drug monitoring, infectious disease, allergy, cardiovascular, anemia, and oncology, therapeutic drug monitoring and thyroid tests are available. Up to 30 different reagent packs can be placed on the instrument. It has a throughput of 240 tests/h.

Table 2. Immunoassay Analyzers

Manufacturer	Model	Methodology
Abbott diagnostics	Axsym	FPIA, MEIA
	TDX, IMX	FPIA
	ADX	FPIA
	Architect	Chemiluminescence
Bayer Diagnostics	ACS 180	Chemiluminescence
	Centaur	Chemiluminescence
	Immuno 1	EIA
Beckman Coulter	Access	EMIT
	LX-20	EMIT
	DCI	Chemiluminescence
Boehringer Manheim ES-300	EIA	
	Elecsys	Chemiluminescence
Dade Behring	Opus Magnum	EIA
	Stratus	FIA
	ACA	EIA, Petinia
	ACA	EIA, turbidimetric
Diagnostic Product Corp.	Immulite	Chemiluminescence, EIA
Nichols Diagnostics CLS ID	Chemiluminescence	
Ortho Clinical	Eci	Chemiluminescence

Table 3. Test Menus for Critical Care Analyzers

Category	Tests Included
Blood gases	pH, pCO_2, pO_2 and other calculated parameters
Electrolytes	Sodium, potassium, chloride, bicarbonate, ionized calcium
Co-oximetry	Carboxyhemoglobin, methemoglobin, total hemoglobin, O_2 saturation
Metabolic panel	Glucose, blood urea nitrogen, creatinine, lactate
Hematology	Hematocrit, hemoglobin, activated clotting time

being performed in combination with general chemistry tests.

Point of Care Testing

Point of care testing (POCT) is defined as laboratory diagnostic testing performed close to the location of the patient. Recent advances over the last decade have resulted in smaller, more accurate devices with a wide menu of tests (12,13). Today POCT can be found from competitive sports to the prison system, from psychiatric counseling to preemployment and shopping mall health screening. Use of POCT devices can be found in mobile transport vehicles such as ambulances, helicopters cruise ships and even the space shuttle.

The advantage of POCT is the ability to obtain extremely rapid laboratory results. However, it is necessary to be aware of the limitations of POCT devices in clinical practice. Venous blood samples often have to be drawn and sent to the main laboratory for confirmation if the results are not within a certain specified range. Another disadvantage of POCT is costs.

In compliance with the guidelines set by federal, state regulatory agencies and the College of American Pathologists (CAP), point of care testing programs are usually overseen by dedicated staff under the direction of the central laboratory. The responsibilities of the POCT staff include education and training of hospital staff, troubleshooting of equipment, maintaining quality control and

Table 4. Partial List of Critical Care Instruments

Vendor	Instrument
Abbott (iSTAT)	iSTAT
Bayer	200,300 800 series, Rapidpoint
Diametrics	IRMA
Instrumentation Lab	1600, 1700 series, Gem series
NOVA	Stat profile series, CCX
Radiometer	ABL series
Roche (AVL)	900 series, Omni and Opti series

Figure 4. The Roche Accu-Check is a small, easy to use blood glucose meter; it is widely used by our Point of Care Testing program. Test results are downloaded to the Laboratory Information System.

quality assurance standards. For a successful POCT program, the laboratory and clinical staff need to effectively work together.

The handheld Accu-Chek POCT device is shown on Fig. 4.

The most widely used point of care tests are bedside glucose testing, critical care analysis, urinalysis, coagulation, occult blood and urine pregnancy testing. Selected point of care devices are listed in Table 5. Other available POCT tests: cardiac markers, pregnancy, influenza A/B, Rapid Strep A, *Helicobacter pylori*, urine microalbumin and creatinine.

CLINICAL LABORATORY AUTOMATION

Historical Perspective

Along with innovations in instrumentation, automating perianalytical activities such as centrifuging, aliquoting, delivering specimens to the automated testing instruments, recapping and storing plays significant role in the modern clinical laboratory (14). Robotic systems that automate some or virtually all of the above functions are available. Automated laboratory and information systems offer benefits in terms of speed, operating efficiency, integrated information sharing and reduction of error.

Table 5. Selected Point of Care Devices

Test	Vendor
Bedside glucose test	Abbott (Medisense PCx)
	Bayer
	Ortho (Lifescan: One Touch)
	Roche
Critical care	Abbott (iSTAT)
	Bayer (Rapidpoint)
	IL (Gem series)
Coagulation	Abbott (iSTAT)
	Bayer (Rapid point)
	Hemosense
	ITC (Hemochron series)
	Medtronics (Hepcon)
	Roche (Coaguchek)
Fecal occult blood	Helena
	Smithkline Diagnostics
Urinalysis	Bayer (Multistix and Clintek)
	Roche (Chemstrip and CUA)

However, the individual needs of each laboratory have to be considered in order to select the optimum combination of instrumentation and perianalytical automation. For small laboratories, front-end work cell automation may be applied economically. For large commercial reference labs and hospital labs, total laboratory automation (TLA) is appropriate where samples move around the whole lab, or from place to place (15).

Clinical laboratory automation evolved with the development of the hematology "Coulter Counter" and the chemistry "AutoAnalyzer" in the 1950s. Automated cell counting by the Coulter involved placing a sample of whole blood in a hemocytometer and using a microscope to count the serial passage of individual cells through an aperture. Likewise, the automated analysis of patient samples for several chemistries dramatically changed the testing process in the chemistry laboratory.

In the 1980s in Japan, Dr. Sasaki's group developed a point-to-point laboratory system that was based on overhead conveyor transportation, delivering specimens placed in 10 position racks (16). These initial designs are the basis of several automation systems available today.

Automation Options and System Design

Available options for automation include the following:

- Interfaced instruments (some can be operated as stand alone analyzers and later linked to a modular system).
- Modular instruments (including, processing, and instrument work cells).
- Multidiscipline platforms (including multifunction instruments and multiwork cells).
- Total laboratory automation robotics system that automates virtually all routine functions in the laboratory.

Automation system design usually rests on the needs of the user. However, the following concepts should be considered:

- Modern information technology with hardware and operating systems that are vertically upgraded.
- Transportation system management at both the local level (device) and overall system level.
- Specimen tracking so that any specimen can be located in the automation system.
- Reflex testing where an additional test can be performed at the same instrument or the specimen can be retrieved to another instrument.
- Information systems agreement with the Laboratory Information System (LIS).

The ability to interface between the hospital LIS and the laboratory automation system (LAS) has been significantly

enhanced by the implementation of the HL7 system-to-system interface. The National Committee on Clinical Laboratory Standards (NCCLS) has issued a proposal level standard (Auto 3 –P) that specifies the HLA interface as the system-to-system communications methodology for connecting an LIS and an LAS (17–21).

NCCLS Guidelines

Components of an optimized laboratory automation system per NCCLS may include:

Preprocessing specimen sorting.
Automated specimen centrifugation.
Automated specimen aliquoting.
Specimen–aliquot recapping/capping.
Specimen integrity monitoring.
Specimen transportation.
Automated specimen sampling.
Automated specimen storage and retrieval.

It is also recommended that process control software should support:

Specimen routing.
Reflex testing.
Repeat testing.
Rules based processing.
Patient data integration.

Available Automation Systems

In the mid-1990s, several laboratory automation technologies implemented hardware-based automation solutions that were centered on defining a limited number of specimen containers compatible with the transportation system. By limiting the number of specimen containers, the hardware can be better defined and more efficient. The original Coulter IDS automation system and the original Hitachi CLAS were based on fixed, rigid or hard-coded hardware technologies.

In the Hitachi CLAS and modular systems, the automation transportation devices use the Hitachi 747 five-place specimen container rack. In order to move the specimen container rack from one analyzer to the next, the automation system must carry along four other patient specimens. The requirement to carry along additional specimens along with the target specimen creates significant mathematical complexity in routing and scheduling of tests. The use of a simple specimen container per specimen carrier model allows the routing of an individual specimen to a workstation without interrupting the flow of other individual specimens in the system.

Total laboratory automation is used to describe the Beckman Coulter IDS system (22). We have two parallel systems in our laboratory (Fig. 5). The basic components include the inlet module, where samples are placed, a centrifuge, serum level sensor, decapping unit, aliquoter-labeler units, outlet units, refrigerated storage unit and a disposal unit. A line PC that interacts with the LIS and all the individual components of the automation system controls the entire system. Each of the automated instruments has their own individual attachment for the handling of specimens being received from the robotic system. View of our automated (perianalytical and analytical) clinical laboratory is shown on Fig. 6.

Work Cell Technologies

The work cell model can be divided into two basic approaches. The first includes all instruments from the same discipline (Chemistry). The second approach is the development of a platform that includes multiple disciplines. An example of this is the Bayer Advia work cell in which chemistry, hematology, immunoassay, and urinalysis processing can take place on one platform. However, this work cell does not have front-end specimen processing and handling capability. Several automated work cells are available in the market at the present time. They include Abbott (Abbott hematology work cell), Beckman-Coulter (Acel-Net work cell), Bayer (Advia work cell), Johnson and Johnson (lab interlink labframe select), and Roche (modular system). The work cell technology varies from simple specimen transportation to complex specimen management.

LABORATORY AUTOMATION-A FINANCIAL PERSPECTIVE

Several studies are being reported on the financial aspects of automation. The most significant impact has been the reduction in FTEs and improvement in turnaround time. A retrospective analysis of 36 years of the effects of initially automation followed by total laboratory automation in the clinical chemistry laboratory at Mount Sinai Medical Center indicated that workload was significantly increased with a reduction of personnel (23). We present these productivity changes in Table 6.

Increased productivity resulted in significant reduction of performing laboratory tests (Table 7).

The effect of increased productivity is illustrated by the drastic reduction of cost/test (Fig. 7)

CALCULATIONS FOR NET PRESENT VALUE OF THE MOUNT SINAI CHEMISTRY AUTOMATION PROJECT (FIG. 8)

Net Present Value

The Net Present Value (NPV) is the value of the net cash flows generated by the project in 1998 $ (the year in which the project was initiated). The NVP is calculated by discounting the value of the annual cash flows [using values taken from the Present Value Interest Factor (PVIF) table for a given project length and cost of capital] to the purchasing value of the dollar at the date of inception of the project (1998). The length of the investment project is a conservative estimate of the useful economic lifetime of the investment project. In this case, we believe that after 8 years additional investments in upgrades

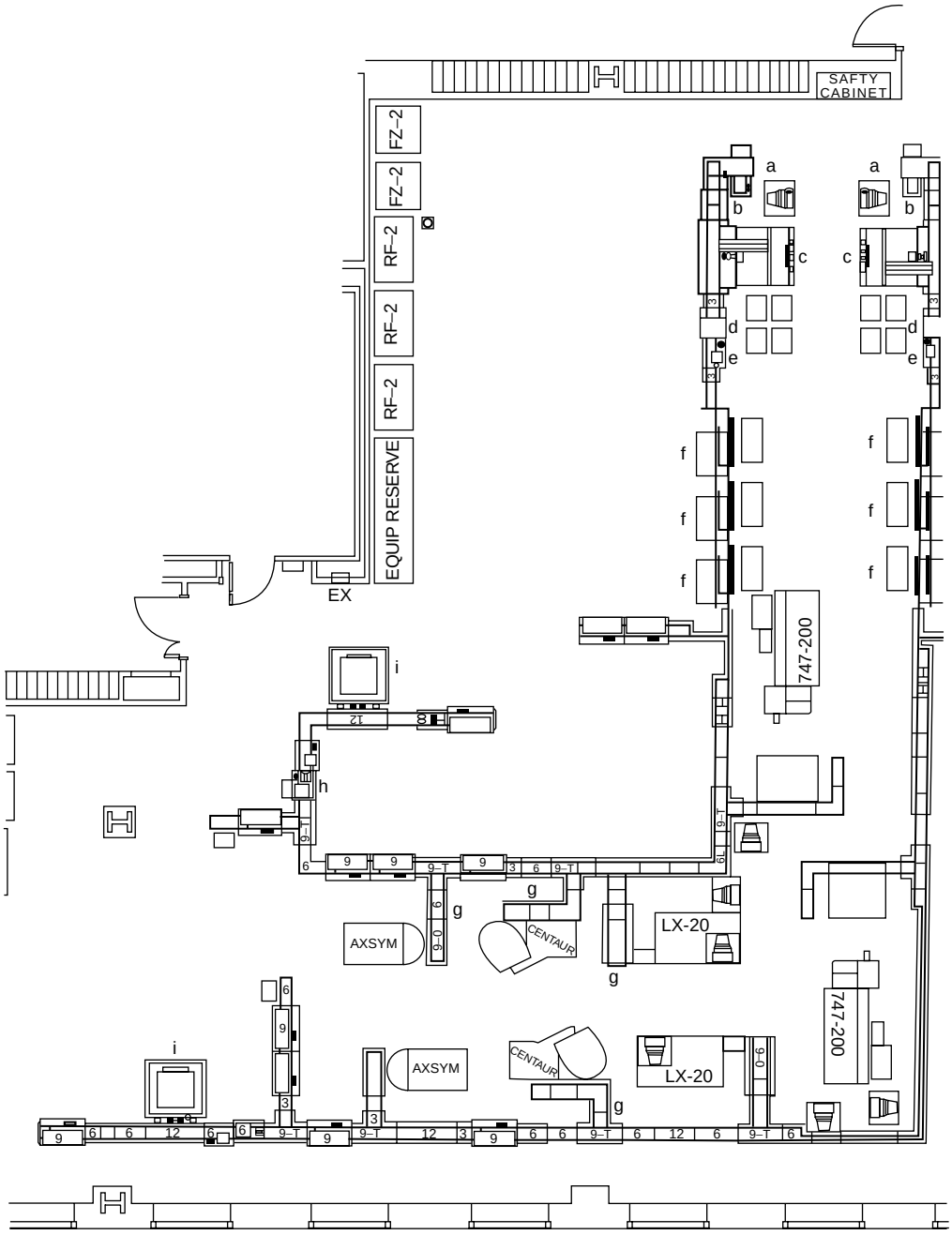


Figure 5. Floor plan of our Total Laboratory Automation. (a) Sample reception. Specimens are picked up, 5 at a time, from 50-position racks and loaded into individual tube holders. A bar-code verification unit determines the legibility of the labels, determines if the specimen is on the right processing location, rejects the suspect samples into a holding area and accepts the correct ones by a message to the Laboratory Information System:"Specimen Received". (b) Sample transport. The transport lanes are conveyor belts that move the samples about the system. (c) Centrifugation. Samples are loaded and unloaded automatically. The rotor has 50 positions. In our laboratory 350 specimens/h can be processed in these centrifuges. (d) Serum level detection. After centrifugation the samples are lowered into an optical well and based on the transmitted information the amount of the available serum is calculated. (e) Cap removal. A gentle rocking motion removes the cups without creating potentially hazardous aerosols. (f) Tube labeling and sample aliquoting. For each primary serum sample secondary aliquot tuber are prepared. The tube labeler prints a bar-code label and applies it to each aliquot tube. The number of aliquot tubes is defined by the system. Disposable pipette tips transfer the serum from the primary to the secondary (aliquot) tubes. The primary tubes are directed to a storage unit. (g) Instrument connections. Several instruments are connected to the transport system. Connection units load and unload samples. Samples not going to the analyzer can continue down the main line. (h) Cap replacement. When the testing of a secondary aliquot tube has been completed, the tube is directed toward an outlet unit, stockyard or storage locker. Before storage, the tube can receive a clean cap. (i) Refrigerated Sample storage. It holds up to 3000 tubes. Samples can be retrieved automatically through a request in the computer and sent to the location requested by the operator.

Figure 6. A portion of the Chemistry automated Core Laboratory at The Mount Sinai Hospital, New York.

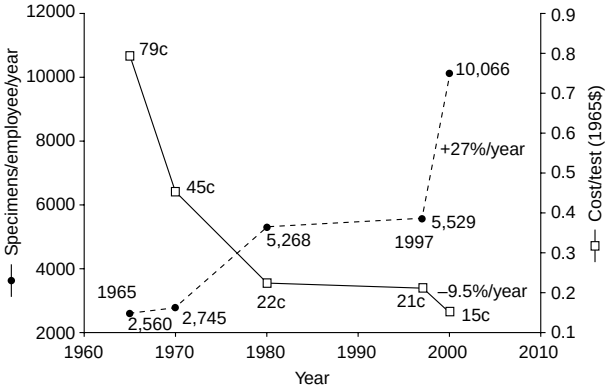


Figure 7. Automation increased Productivity and reduced cost. While the number of specimens processed increased from 2.500 to 10,000 specimens/year the cost/test was reduced from $0.79 to 0.15 (in 1965$).

beyond normal maintenance may be required. The cost of capital used was the interest rate of the lease taken out to finance the project. The relevant calculations are shown below:

1. Total cost of the lease (capital and interest):	$121,963/month×60 months =	$7,318,480
2. Total interest paid over the life of the lease:	$7,318,380 − $6,194,650 =	$1,123,730
3. Annual interest payments:	$1,123,730/5 years =	$ 224,730
4. Interest rate paid on lease:	($224,746/$6,194,650) × 100 =	**3,628%**

Negative Cash Flows

Negative cash flows represent money spent on the project. This includes capital outlays, lease payments ($3,140,000 or $741,921/year for 5 years, represented the portion on chemistry automation), project-related expenses (annual maintenance contract, years 1999 and 2000 = $74,240 annually, 2001–2005 = $39,000 annually.

Positive Cash Flows

Positive cash flows are those that represent money saved and/or costs avoided as the result of the chemistry automation project. There are recurring positive cash flows, resulting from savings that are essentially perpetual, such as salaries and benefits of workers replaced permanently by the chemistry automation project. Savings realized in a given year that are not expected to be repeated in subsequent years are nonrecurring positive cash flows. Staff pay raises during the years 1998, 1999, 2000, and 2002 were

Table 6. Increased Productivity

Year	Tech Staff	Other Staff	Total Staff	No. of Tests/Tech	No. of Tests/tot. Staff	No. of Tests/Specimen	Total No. of Specimens
1965	19	6.00	24.00	14,000	10,600	4.2	2,560
1970	34	17.00	51.00	36,205	24,150	8.8	2,745
1980	39	22.00	61.00	82,359	53,732	10.0	5,268
1997	38	17.00	55.00	94,058	66,099	11.8	5,529
2000	29	13.00	42.00	151,190	104,558	10.4	10,066
2002	29	39.00	35.00	169,582	128,530	10.5	12,190

Table 7. Cost/Test Reduction

Year	Tech Salary, $	Salary $/Test	Supplies $/Test	Total $/Test	Salary 1965 $/Test	Supplies 1965 $/Test	Total 1965 $/Test
1965	5,170	0.70	0.19	0.79	0.70	0.09	0.79
1970	9,114	0.38	0.17	0.55	0.31	0.14	0.45
1980	16,500	0.37	0.20	0.57	0.14	0.08	0.22
1997	38,000	0.66	0.41	1.07	0.13	0.08	0.21
2000	41,000	0.45	0.36	0.81	0.08	0.07	0.15
2002	44,000	0.38	0.34	0.72	0.07	0.06	0.13

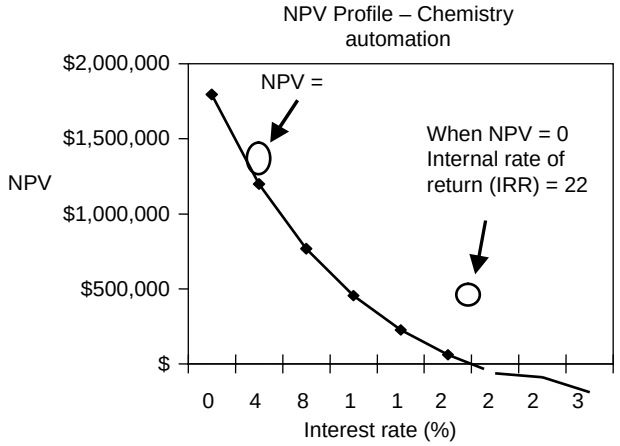


Figure 8. The Internal Rate of Return for the Chemistry Automation project was a remarkable 22%.

financed directly from chemistry automation project savings. These amounts are not reflected in the net salary and benefits savings. As such they are positive cash flow, since they represent costs covered by the automation savings that otherwise would have had to be financed through other sources.

The NPV Profile and Internal Rate of Return

The interest rate employed to discount the value of cash flows to the baseline year (1998) is the marginal cost of capital (the interest rate of the lease), which is 4%. The interest rate at which the NPV equals zero is especially interesting. This is called the internal rate of return (IRR) of the project. For all interest rates below the IRR, the NPV will generate positive values. In order to determine the IRR, we construct an NPV profile for different interest rates and locate the rate where the NPV crosses the X axis where the NPV $= 0$ (Fig. 8.). It shows that the chemistry automation project can tolerate interest rates up to 18.0% (the IRR) and still generate positive returns.

The Payback Period and Average Return on Investment

Although the NPV and IRR are vastly superior indicators of project profitability because of their use of discounted cash flows, the payback period and return on investment (ROI) are still key determinant of project viability by a majority of financial managers.

The Payback Period. After 8 years the raw dollar value of positive cash flows is \$5,371,743 versus negative cash flows of \$4,0053,085. The payback period therefore:

$$8\,\text{years} \times (\$4,053,085/\$5,371,743)$$

$$8\,\text{years} \times 0.755 = 6.04\,\text{years}$$

Average ROI

Average Annual Cash Outlay:	\$4,053,085/8 years = \$506,635/ year
Average Annual Net Return:	(5,371,743 − \$4,053,085)/8 years = \$164,822
Average ROI:	(\$164,832/\$506,635) × 100 = 32.5%

CONCLUSIONS

Productivity is a key issue for labs.

The major financial benefit of automation is increased productivity.

Perianalytical automation increased our chemistry productivity by 120% (from 5,530 to 12,190 specs/tech/year).

Perianalytical automation reduced our chemistry labor cost/test by 42% (from 66¢ to 38¢/test).

Automation is a key solution for staff shortages.

Speedy implementation, speedy labor reductions and speedy revenue generation improve financial performance.

To achieve financial success, laboratorians must understand key financial principles.

ACKNOWLEDGMENTS

We thank E. Simson for practical advice on the financial perspective and M. Gannon for teaching us the meaning and calculation of the Net Present Value.

BIBLIOGRAPHY

Cited References

1. Eggert AA. Analytical methods, automated. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation. Hoboken (NJ): John Wiley & Sons; 1988.
2. Narayanan S. The preanalytical phase: an important component of laboratory medicine. Am J Clin Pathol 2000;113:429–452.
3. Labcorp directory of services and interpretive guide 2003.
4. Ford A. Latest chemistry wish list in low volume labs. CAP today 2004; April 44–58.
5. Lee-Lewandrowski E, Lewandrowski K. Contemporary instruments in the clinical laboratory: A brief overview. In: Lewandrowski K, editor. Clinical chemistry. Philadelphia: Lippincott Williams & Wilkins; 2002.
6. Aller R. Chemistry analyzers. CAP today 1999; July 58–83.
7. Adesoji BB, Peterson JR. Immunoassay and immunoassay analyzers: A perspective from the clinical laboratory. In: Lewandrowski K, editor. Clinical chemistry. Philadelphia: Lippincott Williams & Wilkins; 2002.
8. Goslimg JP. A decade of development in immunoassay technology. Clin Chem 1990;36:1408–1427.
9. Ehrhardt V, Assmann G, Batz O, et al. Results of the multicentre evaluation of an electrochemiluminescence immunoassay for hcg on elecsys 2010. Wein Klin Wochenschr 1998;3 (Suppl): 61–67.
10. Aller RD, Smalley D. Of all analyzers, immunoassay the trickiest. CAP today 2000;April 30–64.
11. Ford A. Automated immunoassay analyzers: the latest lineup. CAP today 2003; June 72–96.
12. Jacobs E. Acute care and stat lab testing. In: Lewandrowski K, editor. Clinical chemistry. Philadelphia: Lippincott Williams & Wilkins; 2003.

13. Ford A. Choosing cost-efficiency in low-volume labs. CAP today 2003; June 32–52.
14. Markin RS. Clinical laboratory automation. In: Henry JB, editor. Clinical diagnosis and management by laboratory methods. Philadelphia: W.B. Saunders; 2001.
15. Ford A. Laboratory automation systems and work cells. CAP today 2003; May: 35–52.
16. Sasaki M. Completed automatic clinical laboratory using a sample transportation system: the belt-line system. Jpn J Clin Pathol 1984;32:119–126.
17. NCCLS laboratory automation: specimen container/specimen carrier; proposed standard. NCCLS document auto 1 P; December 1995.
18. NCCLS laboratory automation: bar codes for specimen container identification; proposed standard. NCCLS document 2 P; April 1999.
19. NCCLS laboratory automation: communications with automated clinical laboratory systems, instruments, devices and information systems; proposed standard. NCCLS document 3 P; December 1998.
20. NCCLS laboratory automation: systems operational requirements and information elements; proposed standard. NCCLS document auto 4 P; October 1999.
21. NCCLS laboratory automation; electromechanical interface; proposed standard. NCCLS document auto 5 P; April 1999.
22. Markin RS, Whalen SA. Laboratory automation; trajectory, technology and tasks. Clin Chem 2000;46:764–771.
23. Sarkozi L, Simson E, Ramanathan L. The effects of total laboratory automation on the management of a clinical chemistry laboratory. Retrospective analysis of 36 years. Clinica Chimica Acta 2003;329:89–94.

See also BLOOD COLLECTION AND PROCESSING; COMPUTERS IN THE BIOMEDICAL LABORATORY; CYTOLOGY, AUTOMATED; DIFFERENTIAL COUNTS, AUTOMATED.

ANALYZER, OXYGEN.

See OXYGEN ANALYZERS.

ANESTHESIA MACHINES

ROBERT LOEB
University of Arizona
Tuscon, Arizona

JEFFREY FELDMAN
Children's Hospital of Philadelphia
Philadelphia, Pennsylvania

INTRODUCTION

On October 16, 1846, W. T. G. Morton gave the first successful public demonstration of inhalational anesthesia. Using a hastily devised glass reservoir to deliver diethyl ether, he anesthetized a patient before an audience at the Massachusetts General Hospital (Fig. 1). This glass reservoir thus became the first, crude, anesthesia machine. The technology of anesthesia machines has advanced immeasurably in the ensuing 150 years. Modern anesthesia machines are used to administer inhalational anesthesia safely and precisely to patients of any age, in any state of health, for any duration of time, and in a wide range of operating environments.

Figure 1. A reproduction of the Morton Inhaler, ~1850. (Image © by the Wood Library-Museum of Anesthesiology, Park Ridge, Illinois.)

The term anesthesia machine colloquially refers to all of the medical equipment used to deliver inhalational anesthesia. Inhalational anesthetics are gases that, when inhaled, produce a state of general anesthesia, a drug-induced reversible loss of consciousness during which the patient is not arousable, even in response to painful stimulation. Inhalational anesthetics are supplied as either compressed gases (e.g., nitrous oxide), or volatile liquids (e.g., diethyl ether, sevoflurane, or desflurane). In recent years, the anesthesia machine has been renamed the anesthesia delivery system, or anesthesia workstation because modern devices do more than simply deliver inhalational anesthesia. Defined precisely, the term "anesthesia machine" specifically refers to that component of the anesthesia delivery system that precisely mixes the compressed and vaporized gases that are inhaled to produce anesthesia. Other components of the anesthesia delivery system include the ventilator, breathing circuit, and waste gas scavenger system. Anesthesia workstations are anesthesia delivery systems that also incorporate patient monitoring and information management functions (Fig. 2).

The most obvious goals of general anesthesia are to render a patient unaware and insensible to pain so that surgery or other medically necessary procedures can be performed. In the process of achieving these goals, potent medications are administered that interfere with normal body functions, most notably circulation of blood and the ability to breathe (see the text box Typical Process of Delivering General Anesthesia). The most important goal of anesthesia care is therefore to keep the patient safe and free from injury.

Patient safety is a major principle guiding the design of the anesthesia workstation. Precise control of the dose of anesthetic gases and vapors reduces the risk of administering an overdose. The ventilator and breathing circuit are fundamental components of the anesthesia delivery system designed to allow for continuous delivery of oxygen to the lungs and removal of exhaled gases. To fulfill national and international standards, anesthesia delivery systems must have essential safety features and meet specified minimum performance criteria (1–6)

Typical Process of Delivering General Anesthesia

Check the anesthesia delivery system for proper function: At the start of each day, the anesthesia provider places disposable components on the breathing circuit and performs an equipment check to ensure proper function of the anesthesia workstation (7).

Identify the patient and confirm the surgical site: Healthcare institutions are required to have formal procedures to identify patients and the site of surgery before the patient is anesthetized.

Establish venous access to administer medications and fluids: Using this catheter, drugs can be administered intravenously and fluids can be given to replace loss of blood or other body fluids.

Attach physiologic monitors: Monitoring the effects of anesthesia on the body is of paramount importance to guide the dose of anesthetic given and to keep the patient safe. Typical monitors include a blood pressure cuff, electrocardiogram, and pulse oximeter. Standards require that additional monitors be used during most anesthesia care (8).

Have the patient breathe 100% oxygen through a mask and circuit attached to the anesthesia machine: A tightly fitting mask is held over the patient's face while 100% oxygen is administered using the anesthesia machine. The goal is to eliminate the nitrogen in the lungs and provide a reservoir of oxygen to sustain the patient from the time anesthesia is induced until mechanical ventilation is established.

Inject a rapidly acting sedative–hypnotic medicine into the patient's vein: This injection induces general anesthesia and often causes the patient to stop breathing. Typical induction medications (e.g., thiopental, propofol) are quickly redistributed and metabolized, so additional anesthetics must be administered shortly thereafter to maintain anesthesia.

Breathe for the patient: This is typically accomplished by holding a mask attached to the breathing circuit tightly over the patient's face and squeezing the bag on the anesthesia machine to deliver oxygen to the lungs. This process is also known as manual ventilation.

Inject a neuromuscular blocking drug to paralyze the patient's muscles: Profound muscle relaxation makes it easier for the anesthesia provider to insert a tracheal tube into the patient's trachea. Neuromuscular blockers are also often used to make it easier for the surgeon to perform the procedure.

Insert a tube into the patient's trachea: This step is called endotracheal intubation and is used to establish a secure path for delivering oxygen and inhaled anesthetics to the patient's lungs as well as eliminating carbon dioxide.

Confirm correct placement of the endotracheal tube: This step is fundamental to patient safety. Numerous methods to confirm correct placement have been described. Identifying the presence of carbon dioxide in the exhaled gas is considered the best method for confirming tube placement. Continuous monitoring of carbon dioxide in the exhaled gases is considered a standard of care during general anesthesia.

Deliver anesthetic agents: General anesthesia is typically maintained with inhaled anesthetic gases. Dials are adjusted on the anesthesia machine to dispense a specified concentration of anesthetic vapor mixed with oxygen and air or nitrous oxide.

Begin mechanical ventilation: The anesthesia delivery system is switched from spontaneous to mechanical ventilation mode, and a ventilator, built into the anesthesia delivery system, is set to breathe for the patient. This frees the anesthesia provider's hands and ensures that the patient breathes adequately during deeper levels of anesthesia and while under the effect of neuromuscular blockers. The ability to deliver anesthetic gases while providing mechanical ventilation is a unique feature of the anesthesia machine.

Adjust ventilation and depth of anesthesia: During the case, the gas flows are reduced to minimize anesthetic usage. The inhaled anesthetic concentration is adjusted to optimize the depth of anesthesia in response to changing levels of surgical stimulus. The ventilator settings are tuned to optimize the patient's ventilation and oxygenation status. Information form the physiologic monitors helps to guide these adjustments.

Establish spontaneous ventilization: Toward the end the operation, the magnitude of ventilation is decreased. The patient responds by starting to breathe spontaneously, at which time the anesthesia delivery system is switched from mechanical to spontaneous ventilation mode and the patient continues to breath from the bag on the anesthesia machine.

Remove the endotracheal tube: At the end of the case, the anesthetic gases are turned off and the patient regains consciousness. The endotracheal tube is removed and the patient breathes oxygen from a cylinder while being transported to the recovery area.

System Overview

Anesthesia delivery systems allow anesthesia providers to achieve the following goals:

1. Precisely deliver a prescribed concentration of inhaled gases to the patient.
2. Support multiple modes of ventilation (i.e., spontaneous, manually assisted, and mechanically controlled).
3. Precisely deliver a wide variety of prescribed ventilator parameters.
4. Conserve the use of anesthetic vapors and gases.
5. Minimize contamination of the operating room atmosphere by anesthetic vapors and gases.
6. Minimize the chance of operator errors.
7. Minimize patient injury in the event of operator error or equipment malfunction.

Figure 2. Four contemporary anesthesia workstations. The top two are manufactured by GE Healthcare, and the bottom two by Draeger Medical.

These goals will be discussed further in the following section, which describes the major components of the anesthesia delivery system. The following overview of anesthesia delivery system function will refer to these goals.

The anesthesia delivery system consists of four components: a breathing circuit, an anesthesia machine, a waste gas scavenger system, and an anesthesia ventilator. The breathing circuit is the functional center of the system, since it is physically and functionally connected to each of the other components and to the patient's airway (Fig. 3). There is a one-way flow of gas from the anesthesia machine into the breathing circuit, and from the breathing circuit into the scavenger system. There is a bidirectional flow of gas between the breathing circuit and the patient's lungs, and between the breathing circuit and the anesthesia ventilator or reservoir bag. The ventilator and the reservoir bag are functionally interchangeable units, which are used during different modes of ventilation (Goal 2). During spontaneous and manually assisted modes of ventilation, the elastic reservoir bag is used as a source of inspired gas and a low impedance reservoir for exhaled gas. The anesthesia ventilator is used during mechanically controlled ventilation to automatically inflate the lungs using prescribed parameters (Goal 3).

During inhalation, gas flows from the anesthesia ventilator or reservoir bag through the breathing circuit to the patient's lungs. The patient's bloodstream takes up a small portion of gas (e.g., oxygen and anesthetic agent) from the lungs and releases carbon dioxide (CO_2) into the lungs.

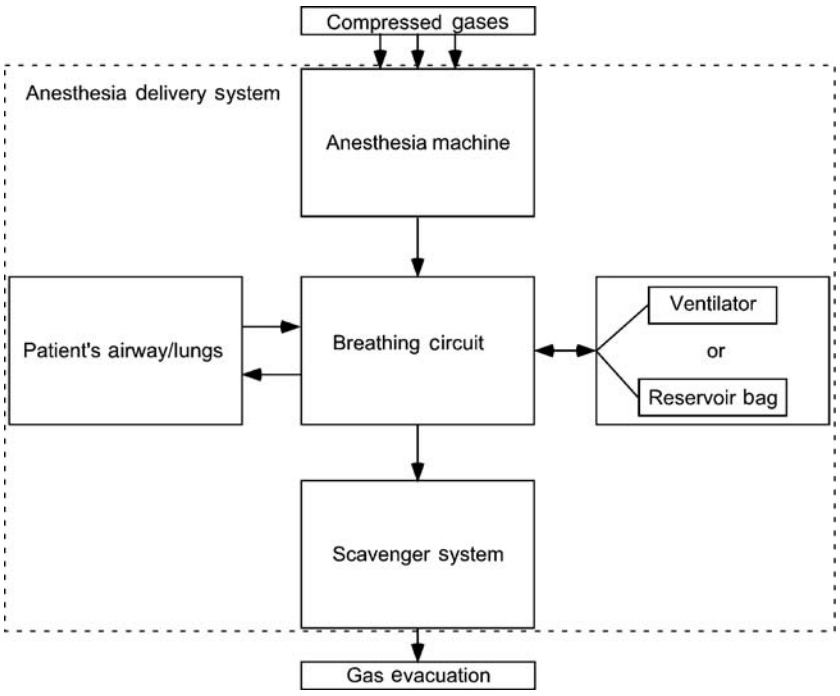


Figure 3. Block diagram of anesthesia delivery system components. The arrows show the direction of gas flow between components.

During exhalation, gas flows from the patient's lungs through the breathing circuit back to the anesthesia ventilator or reservoir bag. This bulk flow of gas, between the patient and the ventilator or reservoir bag, constitutes the patient's pulmonary ventilation; the volume of each breath is referred to as tidal volume, and the total volume exchanged during one minute is referred to as minute volume.

Over time, the patient absorbs oxygen and anesthetic agents from, and releases CO_2 to, the gas in the breathing circuit. Without intervention, the gas within the breathing circuit would progressively decrease in total volume, oxygen concentration, and anesthetic concentration. The anesthesia provider, therefore, dispenses fresh gas into the breathing circuit, replacing the gas absorbed by the patient. Using the anesthesia machine, the anesthesia provider precisely controls both the flow rate and the concentration of various gases in the fresh gas (Goal 1). The anesthesia machine is capable of delivering a total fresh gas flow that far exceeds the volume of gas absorbed by the patient. When higher fresh gas flows are used (for example, to rapidly change the concentration of gases in the breathing circuit), the excess gas is vented into the scavenger system to be evacuated from the operating room (Goal 5).

To conserve the use of anesthetic gases (Goal 4), the anesthesia provider will use a fresh gas flow rate that is significantly lower than the patient's minute volume. In this situation, the patient reinhales gas that they had previously exhaled into the breathing circuit (this is called rebreathing). Carbon dioxide absorbent contained within the breathing circuit prevents the patient from rebreathing CO_2, which would be deleterious. All other gases (oxygen, nitrous oxide, nitrogen, and anesthetic vapors) can be rebreathed safely.

During the course of a typical anesthetic, the anesthesia provider will use a relatively high fresh gas flow at the beginning and end of the anesthetic when a rapid change in anesthetic concentration is desired, and a lower fresh gas flow when little change in concentration is desired. The technique of closed circuit anesthesia refers to the process of adjusting the fresh gas flow to exactly match the amount of gas used by the patient so that no gas is vented to the scavenging system.

Because anesthesia delivery systems provide critical life support functions to unconscious patients, equipment malfunctions and user errors can have catastrophic consequences. In 1974, the American National Standards Institute published an anesthesia machine standard that specified minimum performance and safety requirements for anesthesia gas machines (Goals 6 and 7). That standard was a landmark one, in that it was the first systematic approach to standardize the safety requirements for a medical device. Similar standards have since been written for other medical equipment, and the anesthesia machine standards have been regularly updated.

Breathing Circuit (Semiclosed Circle System)

The semiclosed circle system is the most commonly used anesthesia breathing circuit, and the only type that will be discussed in this article. It is so named because expired gases can be returned to the patient in a circular fashion (Fig. 4). The components of the circle system include a carbon dioxide absorber canister, two one-way valves, a reservoir bag, an adjustable pressure-limiting valve, and tubes that connect to the patient, ventilator, anesthesia machine, and scavenger system.

During inspiration, the peak flow of gas exceeds 25 $L \cdot min^{-1}$, far in excess of the rate of fresh gas supply. As a result, the patient will inspire both fresh gas and gas stored in the reservoir bag or ventilator bellows. Inspired gas travels through the carbon dioxide absorber canister, past the one-way inspiratory valve, to the patient. During

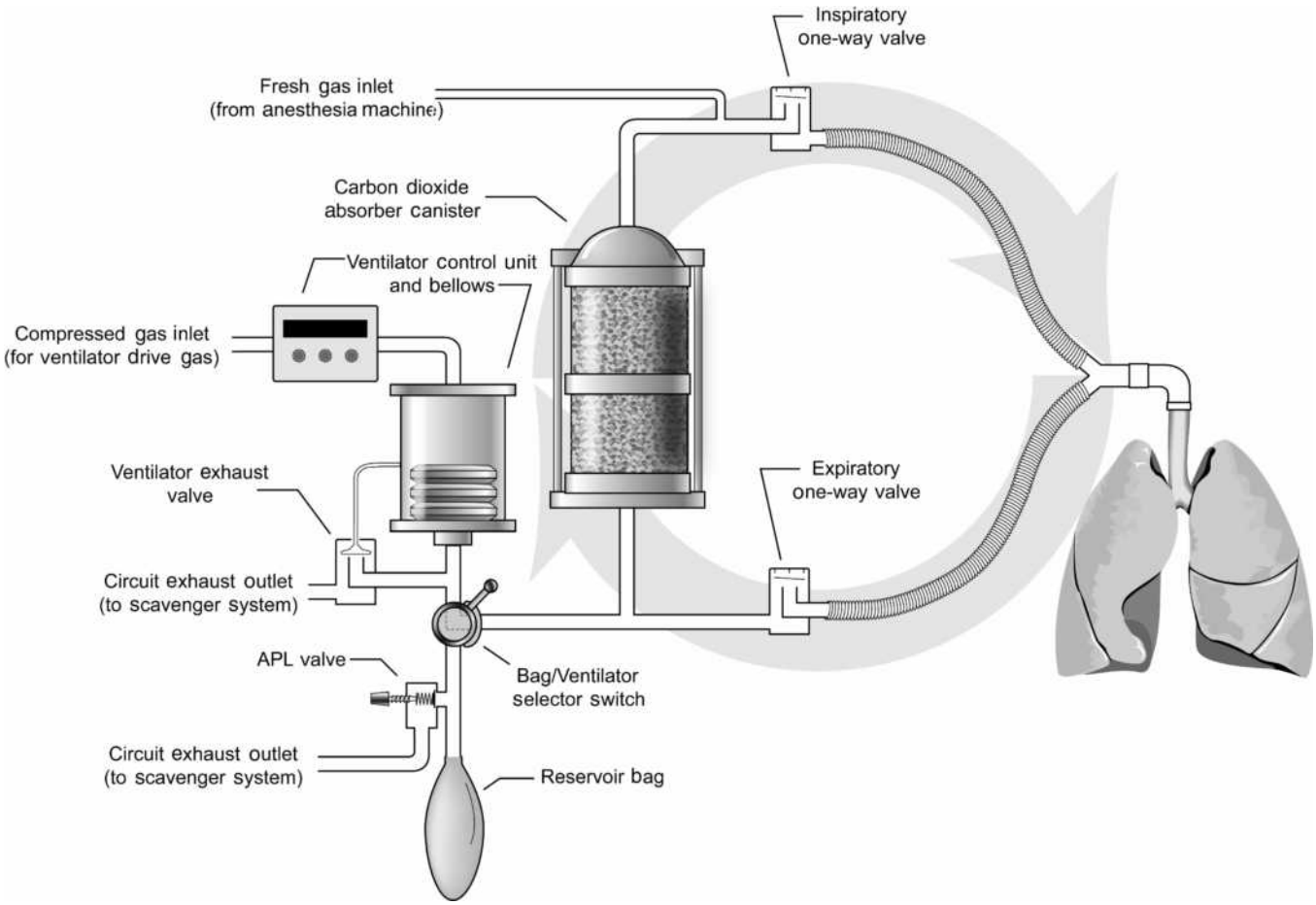


Figure 4. This schematic of the circle breathing circuit shows the circular arrangement of components. The one-way valves permit flow in only one direction.

exhalation, gas travels from the patient, past the one-way expiratory valve, to the reservoir bag (or ventilator bellows, depending upon the position of the bag–ventilator selector switch). The one-way valves establish the direction of gas flow in the breathing circuit. Carbon dioxide is not rebreathed because exhaled gas is directed through the carbon dioxide absorber canister prior to being reinhaled. Fresh gas from the anesthesia machine flows continuously into the breathing circuit. During inhalation, this gas joins with the inspiratory flow and is directed toward the patient. During exhalation, the fresh gas enters the breathing circuit and travels retrograde through the carbon dioxide absorber canister toward the reservoir bag (it does not travel toward the patient because the inspiratory one-way valve is closed during exhalation). Thus, during exhalation, gas enters the reservoir bag from the expiratory limb and from the carbon dioxide absorber canister. Once the reservoir bag is full, excess returning gas is vented out the adjustable pressure-limiting (APL) valve to the scavenger system (when the ventilator is used, the excess gas is vented out the ventilator exhaust valve). The total fresh gas flow will therefore control the amount of gas that is rebreathed. At high fresh gas flows, the exhaled gases are washed out through the scavenging system between each inspiration. At low fresh gas flow, very little exhaled gas is forced out to the scavenging system and most of the exhaled gas is reinhaled in subsequent breaths.

CIRCLE SYSTEM COMPONENTS

CO_2 Absorbents

Alkaline hydroxides of sodium, potassium, calcium, and barium in varying concentrations are most commonly used as carbon dioxide absorbents. These alkaline hydroxides irreversibly react with carbon dioxide to eventually form carbonates, releasing water and heat. Absorbent granules are 4- to 8-mesh in size (25–35 granules cm^{-3}) to maximize the surface area available for chemical reaction and minimize the resistance to gas flow through the absorber canister. Ethyl violet is incorporated into the granules as a pH indicator; fresh granules are white, while a purple color indicates that the absorbent needs to be replaced. Absorber canisters are constructed with transparent sides so that absorbent color can be easily monitored during use. Canisters have a typical capacity of 900–1200 cm^3 and the absorbent is good for ~10–30 h of use, depending on the operating conditions.

Many of the absorbent materials have the potential to interact with anesthetic agents to degrade the anesthetics

and produce small amounts of potentially toxic gases, such as carbon monoxide. This is especially true if the absorbents are allowed to dessicate by exposure to high flows of dry gas (e.g., leaving the fresh gas flowing on the anesthesia machine over a weekend). Periodic replacement of absorbent, especially at the end of a weekend is therefore desirable. Newer absorbent materials, which are more costly, are designed to reduce or eliminate the potential for producing toxic gases by eliminating the hydroxides of sodium, barium, and potassium.

Unidirectional Valves

The inspiratory and expiratory one-way valves are simple, passive devices. Each has an inlet tube that is capped by a valve disk. When the pressure in the inlet tube exceeds that in the outlet tube, the valve opens to allow gas to flow downstream. The valve disks are light in weight to minimize gas flow resistance. Each valve has a clear dome to allow visual monitoring of valve function. Rarely, valves malfunction by failing to open or close properly. Carbon dioxide rebreathing can occur if either valve becomes incompetent (i.e., fails to close properly). This can occur if a valve disk becomes warped, sticks open due to humidity, or fails to seat properly.

Reservoir Bag

The reservoir bag is an elastic bag that serves three functions in the breathing circuit. First, it is a compliant element of an otherwise rigid breathing circuit that allows changes in breathing circuit gas volume without changes in breathing circuit pressure. Second, it provides a means for manually pressurizing the circuit to control or assist ventilation. Third, it provides a safety limit on the peak pressure that can be achieved in the breathing circuit. It acts as a pressure-limiting device in the event that fresh gas inflow exceeds APL valve outflow. Reservoir bags are designed such that, at fresh gas flow rates below 15 $L \cdot min^{-1}$, the breathing circuit pressure will remain < 35 cm H_2O (3.4 kPa) until the bag reaches more than twice its full capacity. Yet, inspiratory pressures up to 70 cm H_2O (6.9 kPa) can be achieved by quickly compressing the reservoir bag.

APL Valve

The APL valve (euphemistically referred to as the pop-off valve) is a spring-loaded device that controls the flow of gas from the breathing circuit to the scavenger system. The valve opens when the pressure gradient from the circuit to the scavenger exceeds the force exerted by the spring (as discussed later, the pressure in the scavenger system is regulated to be equal to atmospheric pressure plus or minus a few cm H_2O). When the patient is breathing spontaneously, the anesthesia practitioner minimizes the spring tension allowing the valve to open with minimal end-expiratory pressure (typically < 3 cm H_2O, or 0.3 kPa). When the anesthesia practitioner squeezes the reservoir bag to manually control or assist ventilation, the APL valve opens during inhalation. Part of the gas exiting the reservoir bag escapes to the scavenger system and the remainder is directed toward the patient. By turning a knob, the anesthesia practitioner increases the pressure on the spring so that the APL valve remains closed until the pressure in the circuit achieves a level that is adequate to inflate the patients lungs; the APL valve thus opens toward the end of inhalation, once the lungs are adequately inflated. Continual adjustment of the APL valve is sometimes needed to adapt to changing fresh gas flow rate, circuit leaks, pulmonary mechanics, and ventilation parameters.

Bag–Ventilator Selector Switch

During mechanical ventilation, the reservoir bag and APL valve are disconnected from the breathing circuit and an anesthesia ventilator is connected to the same spot. Modern breathing circuits have a selector switch that quickly toggles the connection to either the ventilator or the reservoir bag and APL valve.

VIRTUES AND LIMITATIONS OF THE CIRCLE BREATHING CIRCUIT

Primary advantages of the circle breathing system over other breathing circuits include conservation of anesthetic gases and vapors, ease of use, and humidification and heating of inspired gases.

As stated previously, anesthetic agents are conserved when very low fresh gas flows are used with the circle breathing system. The minimum adequate flow is one that just replaces the gases taken up by the patient; for a normal adult, flows below 0.5 $L \cdot min^{-1}$ can be achieved during anesthesia maintenance. It is customary to use higher fresh gas flow rates in the range of 1–2 $L \cdot min^{-1}$, but this is still well below typical minute ventilation rates of 5–10 $L \cdot min^{-1}$ which is the fresh gas flow that would be required for a nonrebreathing ventilation system.

The circle breathing circuit is easy to use because the same fresh gas settings can be used with patients of various sizes. A 100 kg adult and a 1 kg infant can each be anesthetized with a circle breathing system and a fresh gas maintenance flow rate of 1–2 $L \cdot min^{-1}$. Since the larger patient would take up more anesthetic agent and more oxygen, and would give off more carbon dioxide, higher *minimal* flows would be required for the larger patient and the carbon dioxide absorbent would become exhausted quicker. Also, for convenience, a smaller reservoir bag and smaller bore breathing tubes would be selected for the smaller patient. But, otherwise, the system would function similarly for both patients.

Humidification and warming of inspired gases is another advantage of rebreathing. Fresh gas is mixed from compressed gases that contain zero water vapor, and breathing this dry gas can have detrimental effects on lung function. But, within the circle breathing system, inspired gas is humidified by the admixture of rebreathed gas, and by the water vapor that forms as a byproduct of carbon dioxide absorption. Both of these mechanisms also act to warm the inspired gas. By using low flows, enough heat and humidity is conserved to eliminate the need to actively heat and humidify inspired gas.

Most disadvantages of the circle breathing system are due to the large circuit volume. Internal volumes are primarily

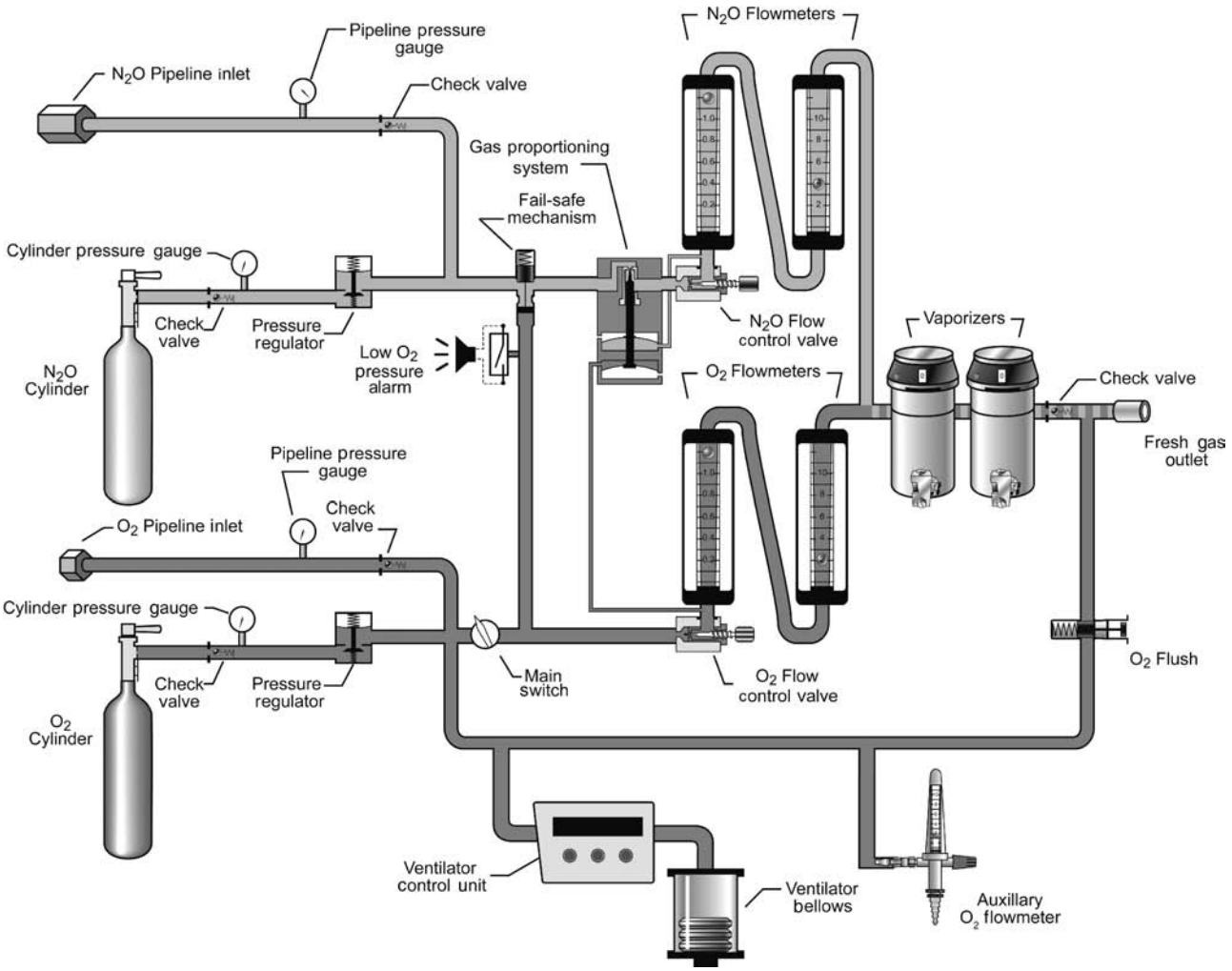


Figure 5. Schematic showing the internal piping and placement of components within the anesthesia machine. Dark gray indicates oxygen (O_2) and light gray indicates nitrous oxide (N_2O).

determined by the sizes of the absorbent canister, reservoir bag, and breathing hoses; 3–6 L are typical. Large circuits are physically bulky. They also increase the time required to change inspired gas concentrations because the large reservoir of previously exhaled gas is continually added to fresh gas. Finally, large circuits are more compliant, which degrades the efficiency and accuracy of ventilation. This effect will be discussed further in the section on ventilators.

Anesthesia Machine

The anesthesia machine is used to accurately deliver into the breathing circuit a precise flow and concentration of gases and vapors. Anesthesia machines are manufactured to deliver various compressed gases; all deliver oxygen, most deliver nitrous oxide or air, some deliver helium or carbon dioxide. They have one or more vaporizers that convert liquid anesthetic agents into anesthetic vapors; currently used inhaled vapors include halothane, enflurane, isoflurane, sevoflurane, and desflurane. Anesthesia machines include numerous safety features that alert the anesthesia provider to malfunctions and avert use errors.

The anesthesia machine is a precision gas mixer (Fig. 5). Compressed gases enter the machine from the hospital's centralized pipeline supply or from compressed gas cylinders. The compressed gases are regulated to specified pressures, and each passes through its own flow controller and flow meter assembly. The compressed gases then are mixed together and may flow through a single vaporizer where anesthetic vapor is added. The final gas mixture then exits the common gas outlet (also called the fresh gas outlet) to enter the breathing circuit.

ANESTHESIA MACHINE COMPONENTS

Compressed Gas Inlets

Compressed gases from the hospital pipeline system or from large compressed gas cylinders enter the anesthesia machine through flexible hoses. The inlet connector for each gas is unique in shape to prevent the connection of the wrong supply hose to a given inlet. The standardized design of each hose-inlet connector pair conforms to the Diameter Indexed Safety System (DISS) (9).

Anesthesia machines also have inlet yokes that hold small compressed gas cylinders; these cylinders provide compressed gas for emergency backup and for use in locations without piped gases. Each yoke is designed to prevent incorrect placement of a cylinder containing another gas. Two pins located in the yoke must insert into corresponding holes on the cylinder valve stem. The standardized placement of these pins and corresponding holes, referred to as the Pin Indexed Safety System (PISS), is unique for each gas (10).

Pressure Regulators And Gauges

Gauges on the front panel of the anesthesia machine display the cylinder and pipeline inlet pressures of each gas. Gases from the pipeline inlets enter the anesthesia machine at pressures of 45–55 psig (310–380 kPa), whereas gases from the compressed gas cylinders enter at pressures up to 2000 psig (1379 kPa). (Pressure conversion factors: 1 psig = 0.068 atm = 51.7 mmHg = 70.3 cm H_2O = 6.89 kPa.) Pressure regulators on each cylinder gas inlet line reduce the pressure from each cylinder to $\sim$ 45 psig (310 kPa). The pressure regulators provide a relatively constant outlet pressure in the presence of a variable inlet pressure, which is important since the pressure within a gas cylinder declines during use. Lines from the pipeline inlet and the cylinder inlet (downstream of the pressure regulator) join to form a common source line for each gas. Gases are preferentially used from the pipelines, since the pressure regulators are set to outlet pressures that are less than the usual pipeline pressures.

Flow Controllers And Meters

A separate needle-valve controls the flow rate of each compressed gas. Turning a knob on the front panel of the anesthesia machine counterclockwise opens the needle valve and increases the flow; turning it clockwise decreases or stops the flow. A flowmeter assembly, located above each flow-control knob, shows the resulting flow rate. The flowmeter consists of a tapered glass tube containing a movable float; the internal diameter of the tube is larger at the top than at the bottom. Gas flows up through the tube, which is vertically aligned, and in doing so blows the float higher in the tube. The float balances in midair partway up the tube when its weight equals the force of the gas traveling through the space between the float and the tube. Thus, the height to which the float rises within the tube is proportional to the flow rate of the gas. Flow rate is indicated by calibrated markings on the tube alongside the level of the float.

Each flowmeter assembly is calibrated for a specific gas. The density and viscosity of the gas significantly affects the force generated in traveling through the variable-sized annular orifice created by the outer edge of the float and the inner surface of the tube. Temperature and barometric pressure affect gas density, and major changes in either can alter flowmeter accuracy. Accuracy is also impaired by dirt or grease within the tube, static electricity between the float and the tube, and nonvertical alignment of the tube.

To increase precision and accuracy, some machines indicate gas flow rate past a single needle valve using two flowmeter assemblies, one for high flows and the other for low flows. These flowmeters are connected in series and the flow rate is indicated on one flowmeter or the other. A flow rate below the range of the high-flow meter shows an accurate flow rate on the low flow meter and an unreadable low flow rate on the high flow meter. While, a flow rate that exceeds the range of the low-flow meter shows an accurate flow rate on the high flow meter and an unreadable high flow rate on the low flow meter.

Each gas, having passed through its individual flow controller and meter assembly, passes into a common manifold before continuing on. Only the individual gas flow rates are indicated on the flowmeters; the user must calculate the total gas flow rate and the percent concentration of each gas in the mixture.

Vaporizers

Vaporizers are designed to add an accurate amount of volatilized anesthetic to the compressed gas mixture. Anesthetic vapors are pharmacologically potent, so low concentrations (generally $< 5\%$) are typically needed. The volatilized gases contribute to the total gas flow rate and dilute the concentration of the other compressed gases. The user can calculate these effects since they are not displayed on the machine front panel; luckily, these are generally negligible and can be ignored. Even though most anesthesia machines have multiple vaporizers, only one is used at a time; interlock mechanisms prevent a vaporizer from being turned on when another vaporizer is in use. Vaporizers are anesthetic agent specific and keyed filling systems prevent filling a vaporizer with the wrong liquid anesthetic.

All current anesthesia machines have direct-setting vaporizers that add a specified concentration of a single anesthetic vapor to the compressed gas mixture. Variable-bypass vaporizers are the most common (Fig. 6). In these, the inflowing compressed gas mixture is split into two streams. One stream is directed through a bypass channel and the other is directed into a chamber within the vaporizer that contains liquid anesthetic agent. The gas entering the vaporizing chamber becomes saturated with anesthetic vapor at a concentration that depends on the vapor pressure of the particular liquid anesthetic. For example, sevoflurane has a vapor pressure of 157 mmHg (20.9 kPa) at 20 °C, so the gas within the vaporizing chamber is about 20% sevoflurane (at sea level). This highly concentrated anesthetic mixture exits the chamber (now, at a flow rate greater than that entering the chamber, due to the addition of anesthetic vapor) to join, and be diluted by, gas that traversed the bypass channel. A dial on the vaporizer controls the delivered anesthetic concentration by regulating the resistance to flow along each path. For example, setting a sevoflurane vaporizer to a dialed concentration of 1% splits the inflowing compressed gas mixture so that one-twenty-fourth of the total is directed through the vaporizing chamber and the remainder is directed through the bypass. Direct-reading variable-bypass vaporizers are calibrated for a specific agent, since each anesthetic liquid has a different vapor pressure. Vapor pressure varies with

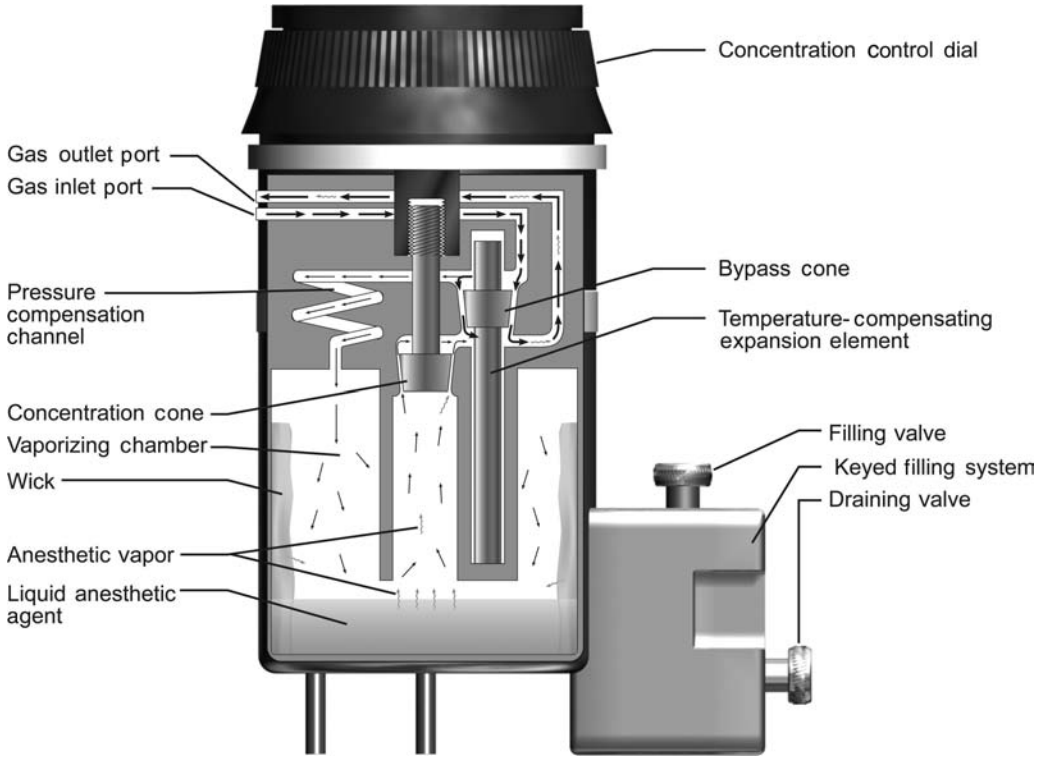


Figure 6. Schematic of a variable-bypass vaporizer. Arrows indicate direction of gas flow; heavier arrows indicate larger flow rates. Gas enters the Inlet Port and is split at the Bypass Cone into two streams. One stream is directed through a bypass channel and the rest enters the Vaporizing Chamber. Gas entering the Vaporizing Chamber equilibrates with Liquid Anesthetic Agent to become saturated with Anesthetic Vapor. This concentrated anesthetic mixture exits the chamber to join, and be diluted by, gas that traversed the bypass channel. The Concentration Control Dial is attached to the Concentration Cone, which regulates resistance to flow exiting the Vaporizing Chamber and thus controls the anesthetic concentration dispensed from the Outlet Port.

temperature, so vaporizers are temperature compensated; at higher temperatures, a temperature sensitive valve diverts more gas through the bypass channel. Vaporizers are designed to ensure that the gas within the liquid-containing chamber is saturated with anesthetic vapor. A cotton wick within the chamber promotes saturation by increasing the surface area of the liquid. Thermal energy is required for liquid vaporization (heat of vaporization). To minimize cooling of the anesthetic liquid, vaporizers are constructed of metals with high specific heat and high thermal conductivity so that heat is transferred easily from the surroundings. The output of variable-bypass vaporizers varies with barometric pressure; delivered concentration increases as barometric pressure decreases.

Desflurane vaporizers are designed differently because desflurane has such a high vapor pressure (664 mmHg, or 88.5 kPa, at 20 °C) and low boiling point (22.8 °C). Uncontrollably high output concentrations could easily occur if desflurane were administered at room temperature from a variable-bypass vaporizer. In a desflurane vaporizer, the liquid desflurane is electrically heated to a controlled temperature of 39 °C within a pressure-tight chamber. At this temperature, the vapor pressure of desflurane is 1500 mmHg (200 kPa) and the anesthetic vapor above the liquid is a compressed gas. The concentration dial on the vaporizer regulates a computer-assisted flow proportioning mechanism that meters pressurized desflurane into the incoming gas mixture to achieve a set output concentration of desflurane vapor. Room temperature does not affect the output concentration of the vaporizer, nor does barometric pressure. The vaporizer requires electrical power for the heater, the onboard computer, and two electronic valves.

Safety Systems

By written standard, the anesthesia machine has numerous safety systems designed to prevent use errors. Some of these, such as the DISS and PISS systems to prevent compressed gas misconnections, interlock mechanisms to prevent simultaneous use of multiple vaporizers, and keyed filler systems to prevent misfilling of vaporizers, have already been discussed. Others are presented, below.

Failsafe Mechanism And Oxygen Alarm. The anesthesia machine has a couple of safety systems that alert the user and stop the flow of other gases when the oxygen supply runs out (for example, when an oxygen tank becomes depleted). An auditory alarm sounds and a visual message appears to alert the user when the oxygen supply pressure falls below a predetermined threshold pressure of ~ 30 psig (207 kPa). A failsafe valve in the gas line supplying each flow controller-meter assembly, except oxygen, stops the

flow of other gases. The failsafe valve is either an on–off valve or a pressure-reducing valve that is controlled by the pressure within the oxygen line. When the oxygen supply pressure falls below the threshold level, the failsafe valves close to stop the flow, or proportionally reduce the supply pressure, of all the other gases. This prevents administration of hypoxic gases (e.g., nitrous oxide, helium, nitrogen, carbon dioxide) without oxygen, which could rapidly cause injury to the patient, but it also prevents administration of air without oxygen. The failsafe mechanisms do *not* prevent delivery of hypoxic gas mixtures in the presence of adequate oxygen supply pressure; the gas proportioning system, described below, prevents this.

Gas Proportioning System. Anesthesia machines are equipped with proportioning systems that prevent the delivery of high concentrations of nitrous oxide, the most commonly used non-oxygen containing gas. A mechanical or pneumatic link between the oxygen and nitrous oxide lines ensures that nitrous oxide does not flow without an adequate flow of oxygen. One such mechanism, the Datex-Ohmeda Link-25 system, is a chain linkage between sprockets on the nitrous oxide and oxygen flow needle valves. The linkage is engaged whenever the nitrous oxide is set to exceed three-times the oxygen flow, or when the oxygen flow is set to less than one-third of the nitrous oxide flow; this limits the nitrous oxide concentration to a maximum of 75% in oxygen. Another mechanism, the Draeger Oxygen Ratio Monitor Controller (ORMC), is a slave flow control valve on the nitrous oxide line that is pneumatically linked to the oxygen line. This system limits the flow of nitrous oxide to a maximum concentration of $72 \pm 3\%$ in oxygen. Both of the above systems control the ratios of nitrous oxide and oxygen, but do not compensate for other gases in the final mixture; a hypoxic mixture (oxygen concentration $< 21\%$) could be dispensed, therefore, if a third gas were added in significant concentrations.

Oxygen Flush. Each anesthesia machine has an oxygen flush system that can rapidly deliver 45–70 $L \cdot min^{-1}$ of oxygen to the common gas outlet. The user presses the oxygen flush valve in situations where high flow oxygen is needed to flush anesthetic agents out of the breathing circuit, rapidly increase the inhaled oxygen concentration, or compensate for a large breathing circuit leak (for example, during positive pressure ventilation of the patient with a poorly fitted face mask). The oxygen flush system also serves as a safety system because it bypasses most of the internal plumbing of the anesthesia machine (e.g., safety control valves, flow controller-meter assemblies, and vaporizers) and because it is always operational, even when the anesthesia machine's master power switch is off.

Monitors and User-Interface Features. Written standards specify that all anesthesia machines must be equipped with essential safety monitors and user-interface features. To protect against hypoxia, each has an integrated oxygen analyzer that monitors the oxygen concentration in the breathing circuit whenever the anesthesia machine is powered on. The oxygen monitor must have an audible alarm that sounds whenever the oxygen concentration falls below a preset threshold, which cannot be set $< 18\%$. To protect against dangerously high and low airway pressures, the breathing circuit pressure is continuously monitored by an integrated system that alarms in the event of sub-atmospheric airway pressure, sustained high airway pressure, or extremely high airway pressure. To protect against ventilator failure and breathing circuit disconnections, the breathing circuit pressure is monitored to ensure that adequate positive pressure is generated at least a few times a minute whenever the ventilator is powered on; a low airway pressure alarm (AKA disconnect alarm) is activated whenever the breathing circuit pressure does not reach a user-set threshold level over a 15 s interval. User-interface features protect against mistakes in gas flow settings. Oxygen controls are always positioned to the right of other gas flow controls. The oxygen flow control knob has a unique size and shape that is different from the other gas control knobs. The flow control knobs are protected against their being bumped to prevent accidental changes in gas flow rates. All gas flow knobs and vaporizer controls uniformly increase their settings when turned in a clockwise direction.

LIMITATIONS

Anesthesia machines are generally reliable and problem-free. Limitations include that they require a source of compressed gases, are heavy and bulky, are calibrated to be accurate at sea level, and are designed to function in an upright position within a gravitational field. Machine malfunctions are usually a result of misconnections or disconnections of internal components during servicing or transportation. Aside from interlock mechanisms that decrease the likelihood of wrong gas or wrong anesthetic agent problems, there are no integrated monitors to ensure that the vaporizers are filled with the correct agents and the flow meters are dispensing the correct gases. Likewise, except for oxygen, the gas supply pressures and anesthetic agent levels are not automatically monitored. Thus, problems can still result when the anesthesia provider fails to diagnose a problem with the compressed gas or liquid anesthetic supplies.

Ventilator

General anesthesia impairs breathing by two mechanisms, it decreases the impetus to breath (central respiratory depression), and it leads to upper airway obstruction. Additionally, neuromuscular blockers, which are often administered during general anesthesia, paralyze the muscles of respiration. For these reasons, breathing may be supported or controlled during anesthesia to ensure adequate minute ventilation. The anesthesia provider can create intermittent positive pressure in the breathing circuit by rhythmically squeezing the reservoir bag. Ventilatory support is often provided in this way for short periods of time, especially during the induction of anesthesia. During mechanical ventilation, a selector switch is toggled to disconnect the reservoir bag and APL valve from the breathing circuit and connect an anesthesia ventilator instead. Anesthesia

ventilators provide a means to mechanically control ventilation, delivering consistent respiratory support for extended periods of time and freeing the anesthesia provider's hands and attention for other tasks. Most surgical patients have normal pulmonary mechanics and can be adequately ventilated with an unsophisticated ventilator designed for ease of use. But, high performance anesthesia ventilators allow safe and effective ventilation of a wide variety of patients, including neonates and the critically ill.

Most anesthesia ventilators are pneumatically powered, electronically controlled, and time cycled. All can be set to deliver a constant tidal volume at a constant rate (volume control). Many can also be set to deliver a constant inspiratory pressure at a constant rate (pressure control). All anesthesia ventilators allow spontaneous patient breaths between ventilator breaths (intermittent mandatory ventilation, IMV), and all can provide PEEP during positive pressure ventilation (note that in some older systems PEEP is set using a PEEP-valve integrated into the expiratory limb of the breathing circuit, and is not actively controlled by the ventilator). In general, anesthesia ventilators do not sense patient effort, and thus do not provide synchronized modes of ventilation, pressure support, or continuous positive airway pressure (CPAP).

As explained above, the anesthesia delivery system conserves anesthetic gases by having the patient rebreathe previously exhaled gas. Unlike intensive care ventilators, which deliver new gas to the patient during every breath, anesthesia ventilators function as a component of the anesthesia delivery system and maintain rebreathing during mechanical ventilation. In most anesthesia ventilators, this is achieved by incorporating a bellows assembly (see Fig. 4). The bellows assembly consists of a distensible bellows that is housed in a clear rigid chamber. The bellows is functionally equivalent to the reservoir bag; it is attached to, and filled with gas from, the breathing circuit. During inspiration, the ventilator injects drive gas into the rigid chamber; this squeezes the bellows and directs gas from the bellows to the patient via the inspiratory limb of the breathing circuit. The drive gas, usually oxygen or air, remains outside of the bellows and never enters the breathing circuit. During exhalation, the drive gas within the rigid chamber is vented to the atmosphere, and the patient exhales into the bellows through the expiratory limb of the breathing circuit.

The bellows assembly also contains an exhaust valve that vents gas from the breathing circuit to the scavenger system. This ventilator exhaust valve serves the same function during mechanical ventilation that the APL valve serves during manual or spontaneous ventilation. However, unlike the APL valve, it is held closed during inspiration to ensure that the set tidal volume dispensed from the ventilator bellows is delivered to the patient. Excess gas then escapes from the breathing circuit through this valve during exhalation.

The tidal volume set on an anesthesia ventilator is not accurately delivered to the patient; it is augmented by fresh gas flow from the anesthesia machine, and reduced due to compression-loss within the breathing circuit. Fresh gas, flowing into the breathing circuit from the anesthesia machine, augments the tidal volume delivered from the ventilator because the ventilator exhaust valve, which is the only route for gas to escape from the breathing circuit, is held closed during inspiration. For example, at a fresh gas flow rate of 3 $L \cdot min^{-1}$ (50 $mL \cdot s^{-1}$), and ventilator settings of 10 breaths min^{-1} and an I/E ratio of 1:2 (inspiratory time = 2 s), the delivered tidal volume is augmented by 100 mL per breath (2 s per breath × 50 $mL \cdot s^{-1}$). Conversely, the delivered tidal volume is reduced due to compression loss within the breathing circuit. The magnitude of this loss depends on the compliance of the breathing circuit and the peak airway pressure. Circle breathing circuits typically have a compliance of 7–9 $mL \cdot cm^{-1}$ H_2O (70–90 $mL \cdot kPa^{-1}$), which is significantly higher than the typical 1–3 $mL \cdot cm^{-1}$ H_2O (10–30 $mL \cdot kPa^{-1}$) circuit compliance of intensive care ventilators, because of their large internal volume. For example, when ventilating a patient with a peak airway pressure of 20 cm H_2O (2 kPa) using an anesthesia ventilator with a breathing circuit compliance of 10 mL·cm H_2O, delivered tidal volume is reduced by 200 mL per breath.

LIMITATIONS

Until recently, anesthesia ventilators were simple devices designed to deliver breathing circuit gas in volume control mode. The few controls consisted of a power switch, and dials to set respiratory rate, inspiratory/expiratory (I/E) ratio, and tidal volume. While simple to operate, these ventilators had a number of limitations. As discussed above, delivered tidal volume was altered by peak airway pressure and fresh gas flow rate. Tidal volume augmentation was particularly hazardous with small patients, such as premature infants and neonates, since increasing the gas flow on the anesthesia machine could unintentionally generate dangerously high tidal volumes and airway pressures. Tidal volume reduction was particularly hazardous since dramatically lower than set tidal volumes could be delivered, unbeknown to the provider, to patients requiring high ventilating pressures (e.g., those with severe airway disease or respiratory distress syndrome). Worse yet, the pneumatic drive capabilities of these ventilators were sometimes insufficient to compensate for tidal volume losses due to compression within the breathing circuit; anesthesia ventilators were unable to adequately ventilate patients with high airway pressures (> 45 cm H_2O) requiring large minute volumes (> 10 $L \cdot min^{-1}$). Another imperfection of anesthesia ventilators is that they are pneumatically powered by compressed gases. The ventilator's rate of compressed gas consumption, which is approximately equal to the set minute volume (5–10 $L \cdot min^{-1}$ in a normal size adult), is not a concern when central compressed gas supplies are being used. But the ventilator can rapidly deplete oxygen supplies when compressed gas is being dispensed from the emergency backup cylinders attached to the anesthesia machine (e.g., a backup cylinder could provide over 10 h of oxygen to a breathing circuit at low flow, but would last only one-hour if also powering the ventilator). Lastly, anesthesia ventilators that do not sense patient effort are unable to provide synchronized or supportive modes of ventilation. This limitation is most significant during spontaneous ventilation, since CPAP and pressure support cannot be provided to compensate

for the additional work of breathing imposed by the breathing circuit and endotracheal tube, or to prevent the low lung volumes and atalectasis that result from general anesthesia. New anesthesia ventilators, introduced in the past 10 years, address many of these limitations as discussed later in the section on New Technologies.

Scavenger System

Waste anesthetic gases are vented from the operating room to prevent potentially adverse effects on health care workers. High volatile anesthetic concentrations in the operating room atmosphere can cause problems such as headaches, dysphoria, and impaired psychomotor functioning; chronic exposure to trace levels has been implicated as a causative factor for cancer, spontaneous abortions, neurologic disease, and genetic malformations, although many studies have not borne out these effects. The National Institute for Occupational Safety and Health (NIOSH) recommends that operating room levels of halogenated anesthetics be < 2 parts per million (ppm) and that nitrous oxide levels be < 25 ppm. Waste gases can be evacuated from the room actively via a central vacuum system, or passively via a hose to the outside; alternatively, the waste gas can pass through a canister containing activated charcoal, which absorbs halogenated anesthetics.

The scavenger system is the interface between the evacuation systems described in the preceding sentence and the exhaust valves on the breathing circuit and ventilator (i.e., APL valve and ventilator exhaust valve). It functions as a reservoir that holds waste gas until it can vent to the evacuation system. This is necessary because gas exits the exhaust valves at a non-constant rate that may, at times, exceed the flow rate of the evacuation system. The scavenger system also ensures that the downstream pressure on the exhaust valves does not become too high or too negative. Excessive pressure at the exhaust valve outlet could cause sustained high airway pressure leading to barotrauma and cardiovascular collapse; whereas, excessive vacuum at the exhaust valve outlet could cause sustained negative airway pressure leading to apnea and pulmonary edema.

There are two categories of scavenger systems, open and closed. Open scavenger systems can only be used with a vacuum evacuation system. In an open scavenger system, waste gas enters the bottom of a rigid reservoir that is open to the atmosphere at the top, and gas is constantly evacuated from the bottom of the reservoir into the vacuum. Room air is entrained into the reservoir whenever the vacuum flow rate exceeds the waste gas flow rate, and gas spills out to the room through the openings in the reservoir whenever the waste gas flow rate exceeds the vacuum flow rate. The arrangement of the components prevents spillage of waste gas out of the reservoir openings unless the *average* vacuum flow rate is less than the *average* flow out of the exhaust valves.

Closed scavenger systems consist of a compliant reservoir bag with an inflow of waste gas from the exhaust valves of the breathing system and an outflow to the active or passive evacuation system. Two or more valves regulate the internal pressure of the closed scavenger system. A negative pressure release valve opens to allow entry of room air whenever the pressure within the system becomes too negative, < -1.8 cm H_2O (-0.18 kPa) (i.e., in situations where the evacuation flow exceeds the exhaust flow and the reservoir bag is collapsed). A positive pressure release valve opens to allow venting of waste gas to the room whenever the pressure within the scavenger system becomes too high, > 5 cm H_2O (0.5 kPa) (i.e., in situations where the reservoir bag is full and the exhaust flow exceeds the evacuation flow). Thus, the pressure within the scavenger system is maintained between -1.8 and 5.0 cm H_2O.

Integrated Monitors

All anesthesia delivery systems have integrated electronic safety monitors intended to avert patient injuries. Included are (*1*) an oxygen analyzer, (*2*) an airway pressure monitor, and (*3*) a spirometer.

The oxygen analyzer measures oxygen concentration in the inspiratory limb of the breathing circuit to guard against the administration of dangerously low inhaled oxygen concentrations. Most analyzers use a polarographic or galvanic (fuel cell) probe that senses the rate of an oxygen-dependent electrochemical reaction. These analyzers are inexpensive and reliable, but are slow to equilibrate to changes in oxygen concentration (response times on the order of 30 s). They also require daily calibration. Standards stipulate that the oxygen analyzer be equipped with an alarm, and be powered-on whenever the anesthesia delivery system is in use.

The airway pressure monitor measures pressure within the breathing circuit, and warns of excessively high or negative pressures. It also guards against apnea during mechanical ventilation. Most anesthesia delivery systems have two pressure gauges: an analog Bourdon tube pressure gauge that displays instantaneous pressure on a mechanical dial, and an electronic strain-gauge monitor that displays a pressure waveform. Most electronic pressure monitors embody an alarm system with variable-threshold negative pressure, positive pressure, and sustained pressure alarms that can be adjusted by the user. An apnea alarm feature, which is enabled whenever the ventilator is powered-on, ensures that positive pressure is sensed within the breathing circuit at regular intervals. On some anesthesia delivery systems pressure is sensed within the circle system absorber canister; on other systems it is sensed on the patient side of the one-way valves; the latter gives a more accurate reflection of airway pressure.

The spirometer measures gas flow in the expiratory limb of the breathing circuit and guards against apnea and dangerously low or high respiratory volumes. A number of different techniques are commonly used to measure flow. These include spinning vanes, rotating sealed spirometers, ultrasonic, and variable orifice differential pressure. Respiratory rate, tidal volume, and minute volume are derived from the sensor signals and displayed to the user. Some machines also display a waveform of exhaled flow versus time. Most spirometers have an alarm system with variable-threshold alarms for low and high tidal volume, as well as an apnea alarm that is triggered if no flow is detected during a preset interval.

In addition to these standard monitors, some anesthesia workstations have integrated gas analyzers that measure inhaled and exhaled concentrations of oxygen, carbon dioxide, nitrous oxide, and volatile anesthetic agents. Although stand-alone gas analyzers are available, they are likely to be integrated into the anesthesia workstation because they monitor gas concentrations and respiratory parameters that are controlled by the anesthesia delivery system.

Other patient monitors, such as electrocardiography, pulse oximetry, invasive and noninvasive blood pressure, and thermometry may also be integrated into the anesthesia workstation; but often stand-alone monitors are placed on the shelves of the anesthesia delivery system. In either case, standard patient monitors must be used during the conduct of any anesthetic to evaluate the adequacy of the patient's oxygenation, ventilation, circulation, and body temperature. Monitoring standards, which have contributed to the dramatic increase in anesthesia safety, were initially published by the American Society of Anesthesiologists in 1986 and have been continually evaluated and updated (8).

New Technologies

The anesthesia delivery system as described thus far has evolved incrementally from a pneumatic device designed in 1917, by Henry Boyle for administration of anesthesia using oxygen, nitrous oxide and ether. The evolution of Boyle's machine has occurred in stages. In the 1950s and 1960s the failsafe devices and fluidic controlled ventilators were added. In the 1970s and early 1980s the focus was on improving safety with features, such as gas proportioning systems, safety alarms, electronically controlled ventilators, and standardization of the user interface to decrease errors. In the late 1980s and 1990s, monitors and electronic recordkeeping were integrated to create anesthesia workstations. Since 2000 the focus has been on improving ventilator performance, incorporating automated machine self-checks, and transitioning to electronically controlled and monitored flow meters and vaporizers. Some of the new technologies that have been introduced in the last few years are discussed, below.

BREATHING CIRCUIT

As discussed above, the tidal volume set on an anesthesia ventilator is not accurately delivered to the patient because of two breathing circuit effects. First, a portion of the volume delivered from the ventilator is compressed within the breathing circuit and does not reach the patient. Second, fresh gas flowing into the breathing circuit augments the delivered tidal volume. A number of techniques are used to minimize these effects in new anesthesia delivery systems.

Two techniques have been used to minimize the effect of gas compression. First, smaller, less compliant breathing circuits are being used. This has been achieved by minimizing the use of compliant hoses between the ventilator and breathing circuit and by decreasing the size of the absorber canister. A tradeoff is that the absorbent must be changed more frequently with a smaller canister, hence new breathing circuits are designed so that the carbon dioxide absorbent can be exchanged during use. Second, many new machines automatically measure breathing circuit compliance during an automated preuse checkout procedure and then compensate for breathing circuit compliance during positive pressure ventilation; the ventilator continually senses airway pressure and delivers additional volume to make up for that lost to compression.

A number of techniques have also been used to eliminate augmentation of tidal volume by fresh gas flowing into the circuit. In one approach, the ventilator automatically adjusts its delivered volume to compensate for the influx of fresh gas into the breathing circuit. The ventilator either adjusts to maintain a set exhaled tidal volume as measured by a spirometer in the expiratory limb of the breathing circuit, or it responds to maintain a set inhaled tidal volume sensed in the inspiratory limb, or it modifies its delivered volume based on the total fresh gas flow as measured by electronic flowmeters in the anesthesia machine. None of the above methods requires redesign of the breathing circuit, except for the addition of flow sensors that communicate with the ventilator.

In a radically different approach, called fresh gas decoupling, the breathing circuit is redesigned so that fresh gas flow is channeled away from ventilator-delivered gas during inspiration, which removes the augmenting effect of fresh gas flow on tidal volume. An example of such a breathing circuit is illustrated in Fig. 7. In this circuit, during inhalation, gas dispensed from a piston driven ventilator travels directly to the patient's lungs; retrograde flow is blocked by a passive fresh gas decoupling valve, and expiratory flow is blocked by the ventilator-controlled expiratory valve, which is actively closed during the inspiratory phase. Fresh gas does not contribute to the delivered tidal volume; instead it flows retrograde into a nonpressurized portion of the breathing circuit. During exhalation, the ventilator-controlled expiratory valve opens, and the ventilator piston withdraws to actively fill with a mixture of fresh gas and gas from the reservoir bag. This design causes a number of other functional changes. First, the breathing circuit compliance is lower during positive pressure ventilation, since only part of the breathing circuit is pressurized during inspiration (the volume between the fresh gas decoupling valve and the ventilator-controlled expiratory valve). Second, the reservoir bag remains in the circuit during mechanical ventilation. As a result, it fills and empties with gas throughout the ventilator cycle, which is an obvious contrast to the absence of bag movement during mechanical ventilation with a conventional circle breathing circuit.

ANESTHESIA MACHINE

Many new anesthesia machines have electronic gas flow sensors instead of tapered glass tubes with internal floats. Advantages include (*1*) improved reliability and reduced maintenance; (*2*) improved precision and accuracy at low-flows; and (*3*) ability to automatically record and use gas flows (for instance to adjust the ventilator). The electronic sensors operate on the principle of heat transfer, measuring the energy required to maintain the temperature of a heated element in the gas flow pathway. Each sensor is calibrated for a particular gas, since every gas has a different specific heat index. Gas flows are shown on dedicated light-emitting diode (LED) displays or on the

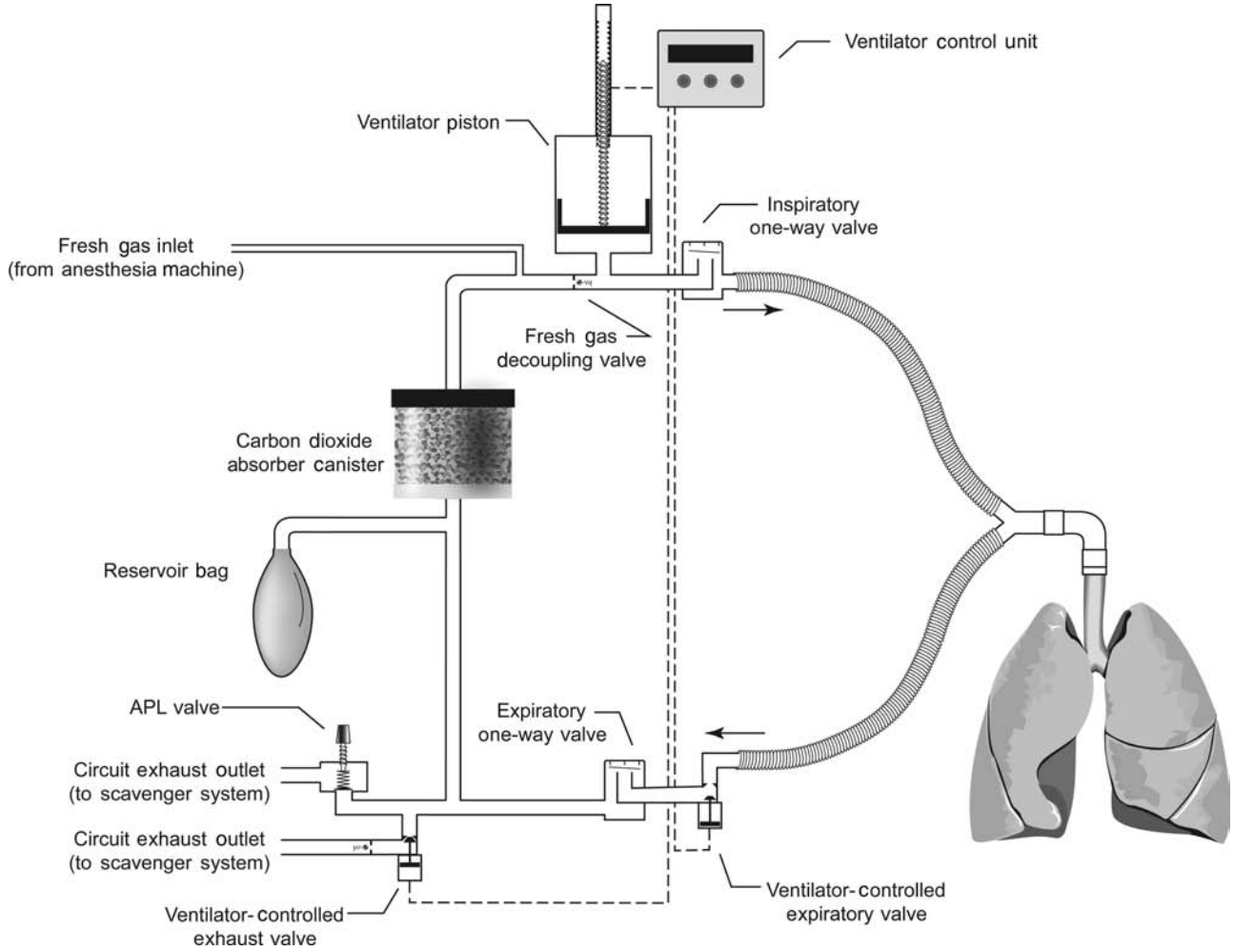


Figure 7. Example of a breathing circuit with fresh gas decoupling. This breathing circuit is used in the Draeger Fabius anesthesia machine. It contains three passive one-way valves and two active valves that are controlled by the ventilator during mechanical ventilation.

main anesthesia machine flat panel display. Most anesthesia machines still regulate the flow of each gas using mechanical needle valves, but in some these have been replaced with electronically control valves. Electronically controlled valves provide a mechanism for computerized gas proportioning systems that limit the ratios of multiple gases. Some machines with electronic flow control valves allow the user to select the balance gas (i.e., air or nitrous oxide) and set a desired oxygen concentration and total flow, leaving the calculation of individual gas flow rates to the machine.

Most new anesthesia machines continue to use mechanical vaporizers as described above, but a few incorporate electronic vaporizers. These operate on one of two principles: either computer-controlled variable bypass, or computer-controlled liquid injection. Computerized variable bypass vaporizers control an electronic valve that regulates the flow of gas exiting from the liquid anesthetic containing chamber to join the bypass stream. The valve is adjusted to reach a target flow that is based upon the: (*1*) dial setting, (*2*) temperature in the vaporizing chamber, (*3*) total pressure in the vaporizing chamber, (*4*) bypass flow, and (*5*) liquid anesthetic identity. Computerized injectors continuously add a measured amount liquid anesthetic directly into the mixed gas coming from the flowmeters based upon the: (*1*) dial setting, (*2*) mixed gas flow, and (*3*) liquid anesthetic identity. Electronic vaporizers offer a number of advantages. First, they provide a mechanism for vaporizer settings to be automatically recorded and controlled. Second, a number of different anesthetics can be dispensed (one at a time) using a single control unit, provided that the computer knows the identity of the anesthetic liquid.

VENTILATOR

Anesthesia ventilator technology has improved dramatically over the past 10 years and each new machine brings further advancements. As discussed above, most new ventilators compensate for the effects of circuit compliance and fresh gas flow, so that the set tidal volume is accurately delivered. Older style ventilators notoriously delivered low tidal volumes to patients requiring high airway pressures, but new ventilators overcome this problem with better flow generators, compliance compensation and feedback

control. Many new anesthesia ventilators offer multiple modes of ventilation (in addition to the traditional volume control), such as pressure control, pressure support, and synchronized intermittent mandatory ventilation. These modes assess patient effort using electronic flow and pressure sensors that are included in many new breathing circuits. Lastly, some anesthesia ventilators use an electronically controlled piston instead of the traditional pneumatically compressed bellows. Piston ventilators, which are electrically powered, dramatically decrease compressed gas consumption of the anesthesia delivery system. However, they actively draw gas out of the breathing circuit during the expiratory cycle (as opposed to bellows, which fill passively) so they cannot be used with a traditional circle system (see Fig. 7 for an example of a piston ventilator used with a fresh gas decoupled breathing circuit).

AUTOMATED CHECKOUT

Many new anesthesia delivery systems feature semiautomated preuse checkout procedures. These ensure that the machine is functioning properly prior to use by (*1*) testing electronic and computer performance, (*2*) calibrating flow sensors and oxygen monitors, (*3*) measuring breathing circuit compliance and leakage, and (*4*) testing the ventilator.

Future Challenges

The current trend is to design machines that provide advanced capabilities through the use of computerized electronic monitoring and controls. This provides the infrastructure for features such as closed-loop feedback, smart alarms, and information management that will be increasingly incorporated in the future. We can anticipate closed-loop controllers that automatically maintain a user-set exhaled anesthetic concentration (an indicator of anesthetic depth), or exhaled carbon dioxide concentration (an indicator of adequacy of ventilation). We can look forward to smart alarms that pinpoint the location of leaks or obstructions in the breathing circuit, alert the user and switch to a different anesthetic when a vaporizer becomes empty, or notify the user and switch to a backup cylinder if a pipeline failure or contamination event is detected. We can foresee information management systems that automatically incorporate anesthesia machine settings into a nationwide repository of anesthesia records that facilitate outcomes-guided medical practice, critical event investigations, and nationwide access to patient medical records. Anesthesia machine technology continues to evolve.

BIBLIOGRAPHY

Cited References

1. ASTM F1850. Standard Specification for Particular Requirements for Anesthesia Workstations and Their Components. ASTM International; 2000.
2. ISO 5358. Anaesthetic machines for use with humans. International Organization for Standardization; 1992.
3. ISO 8835-2. Inhalational anaesthesia systems—Part 2: Anaesthetic breathing systems for adults. International Organization for Standardization; 1999.
4. ISO 8835-3. Inhalational anaesthesia systems—Part 3: Anaesthetic gas scavenging systems—Transfer and receiving systems. International Organization for Standardization; 1997.
5. ISO 8835-4. Inhalational anaesthesia systems—Part 4: Anaesthetic vapour delivery devices. International Organization for Standardization; 2004.
6. ISO 8835-5. Inhalational anaesthesia systems—Part 5: Anaesthetic ventilators. International Organization for Standardization; 2004.
7. Anesthesia Apparatus Checkout Recommendations. United States Food and Drug Administration. Available at http://www.fda.gov//cdrh/humfac//anesckot.html. 1993.
8. Standards for Basic Anesthetic Monitoring, American Society of Anesthesiologists. Available at http://www.asahq.org/publicationsAndServices/standards/02.pdf. Accessed 2004.
9. CGA V-5. Diameter Index Safety System (Noninterchangeable Low Pressure Connections for Medical Gas Applications). Compressed Gas Association; 2000.
10. CGA V-1. Compressed Gas Association Standard for Compressed Gas Cylinder Valve Outlet and Inlet Connections. Compressed Gas Association; 2003.

Reading List

Dorsch J, Dorsch S. Understanding Anesthesia Equipment. 4th ed. Williams & Wilkins; 1999.
Brockwell RC, Andrews JJ. Inhaled Anesthetic Delivery Systems. In: Miller RD. et al. editors. Miller's Anesthesia. 6th ed. Philadelphia: Elsevier Churchill Livingstone; 2005.
Ehrenwerth J, Eisenkraft JB, editors. Anesthesia Equipment: Principles and Applications. St. Louis: Mosby; 1993
Lampotang S, Lizdas D, Liem EB, Dobbins W. The Virtual Anesthesia Machine. http://vam.anest.ufl.edu/.

See also CONTINUOUS POSITIVE AIRWAY PRESSURE; EQUIPMENT ACQUISITION; EQUIPMENT MAINTENANCE, BIOMEDICAL; GAS AND VACUUM SYSTEMS, CENTRALLY PIPED MEDICAL; VENTILATORY MONITORING.

ANESTHESIA MONITORING. See MONITORING IN ANESTHESIA.

ANESTHESIA, COMPUTERS IN

LE YI WANG
HONG WANG
Wayne State University
Detroit, Michigan

INTRODUCTION

Computer applications in anesthesia patient care have evolved with advancement of computer technology, information processing capability, and anesthesia devices and procedures.

Anesthesia is an integral part of most surgical operations. The objectives of anesthesia are to achieve hypnosis (consciousness control), analgesia (pain control), and immobility (body movement control) simultaneously throughout surgical operations, while maintaining the vital functions of the body. Vital functions, such as respiration and circulation of blood, are assessed by signs such as blood pressures, heart rate, end-tidal carbon dioxide (CO_2), oxygen saturation by pulse oximetry (SpO_2), and so on. These objectives are

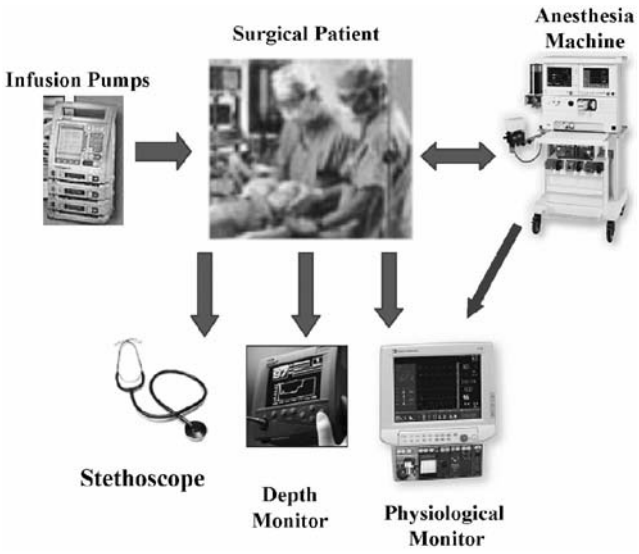


Figure 1. Some anesthesia equipment in an operating room.

carefully balanced and maintained by a dedicated anesthesia provider using a combination of sedative agents, hypnotic drugs, narcotic drugs, and, in many surgeries, muscle relaxants. Anesthesia decisions and management are complicated tasks, in which anesthetic requirements and agent dosages depend critically on the surgical procedures, the patient's medical conditions, drug interactions, and coordinated levels of anesthesia depth and physiological variables. Anesthesia decisions impact significantly on surgery and patient outcomes, drug consumptions, hospital stays, and therefore quality of patient care and healthcare cost (Fig. 1).

Computer technologies have played essential roles in assisting and improving patient care in anesthesia. Development of computer technology started from its early stages of bulky computing machines, progressed to minicomputers and microcomputers, exploded with its storage capability and computational speed, and evolved into multiprocessor systems, distributed systems, computer networks, and multimedia systems. Computer applications in anesthesia have taken advantage of this technology advancement. Early computer applications in medicine date back to the late 1950s when some hospitals began to develop computer data processing systems to assist administration, such as storage, management, and analysis of patient and procedural data and records. The main goals were to reduce manpower in managing ever-growing patient data, patient and room scheduling, anesthesia supply tracking, and billing. During the past four decades computer utility in anesthesia has significantly progressed to include computer-controlled fluid administration and drug dispersing, advanced anesthesia monitoring, anesthesia information systems, computer-assisted anesthesia control, computer-assisted diagnosis and decisions, and telemedicine in anesthesia.

COMPUTER UTILITY IN ADVANCED ANESTHESIA MONITORING

The quality of anesthesia patient care has been greatly influenced by monitoring technology development. A

(a) Paramagnetic Oxygen Analyzer in 1940 (b) Carbon Dioxide Monitor in 1960

Figure 2. Anesthesia monitoring devices without computer technologies. (Courtsey of Sheffield Museum of Anesthesia used with permission.)

patient's state during a surgery is assessed using vital signs. In earlier days of anesthesiology, vital signs were limited to manual measurements of blood pressures, stethoscope auscultation of heart–lung sounds, and heart rates. These values were measured intermittently and as needed during surgery. Thanks to advancement of materials, sensing methods, and signal processing techniques, many vital signs can now be directly, accurately, and continuously measured. For example, since the invention of pulse oximetry in the early 1980s, this noninvasive method of continuously monitoring the arterial oxygen saturation level in a patient's blood (SpO_2) has become a standard method in the clinical environment, resulting in a significant improvement of patient safety. Before this invention, blood must be drawn from patients and analyzed using laboratory equipment.

Integrating these vital signs into a comprehensive anesthesia monitoring system has been achieved by computer data interfacing, multisignal processing, and computer graphics. Advanced anesthesia monitors are capable of acquiring multiple signals from many vital sign sensors and anesthesia machine itself, displaying current readings and historic trends, and providing audio and visual warning signals. At present, heart rate, electrocardiogram (ECG), arterial blood pressures, temperature, ventilation parameters (inspired–expired gas concentration, peak airway pressure, plateau airway pressure, inspired and expired volumes, etc.), end-tidal CO_2 concentrations, blood oxygen saturation (SpO_2), and so on, are routinely and reliably monitored (Figs. 2 and 3).

However, there are still many other variables reflecting a patient's state that must be inferred by the physician, such as anesthesia depth and pain intensity. Pursuit of new

Figure 3. An anesthesia monitor from GE Healthcare in 2005. (Courtsey of GE Healthcare.)

physiological monitoring devices for direct and reliable measurements of some of these variables is of great value and imposes great challenges at the same time (1). Anesthesia depth has become a main focus of research in the anesthesia field. At present, most methods rely in part or in whole on processing of the electroencephalogram (EEG) and frontalis electromyogram (FEMG) signals. Proposed methods include the median frequency, spectral edge frequency, visual evoked potential, auditory evoked potential, entropy, and bispectral index (2,3). Some of these technologies have been commercialized, leading to several anesthesia depth monitors for use in general anesthesia and sedation. Rather than using indirect implications from blood pressures, heart rate, and involuntary muscle movements to derive consciousness levels, these monitors purport to give a direct index of a patient's anesthesia depth. Consequently, combined effects of anesthesia drugs on the patient anesthesia depth can potentially be understood clearly and unambiguously. Currently (the year 2005), the BIS Monitor by Aspect Medical Systems, Inc. (www.aspectmedical.com), Entropy Monitor by GE Healthcare (www.gehealthcare.com), and Patient State Analyzer (PSA) by Hospira, Inc. (www.hospira.com) are three FDA (U.S. Food and Drug Administration) approved commercial monitors for anesthesia depth.

Availability of commercialized anesthesia depth monitors has prompted a burst of research activity on computerized depth control. Improvement of their reliability remains a research frontier. Artifacts have fundamental impact on reliability of EEG signals. In particular, muscle movements, eye blinks, and other neural stimulation effects corrupt EEG signals, challenging all the methods that rely on EEG to derive anesthesia depth. As a result, reliability of these devices in intensive care units (ICU) and emergency medicine remains to be improved.

Another area of research is pain-intensity measurement and monitoring. Despite a long history of research and development, pain intensity is still evaluated by subjective assessment and patient self-scoring. The main thrust is to establish the relation between subjective pain scores, such as the visual analog scale (VAS) system, and objective measures of vital signs. Computer-generated objective and continuous monitoring of pain will be a significant advance in anesthesia pain control. This remains an open and active area of research and development (R&D). As an intermediate step, patient-controlled analgesia (PCA) devices have been developed (see, e.g., LifeCare PCA systems from Hospira, Inc., which is a 2003 spin-off of Abbott Laboratories) that allow a patient to assess his/her pain intensity and control analgesia as needed.

Currently, anesthesia monitors are limited to data recording and patient state display. Also, their basic functions do not offer substantial interaction with human and environment. Future monitors must enhance fundamentally human-factors design: Intelligent human–machine interface and integrated human–machine–environment systems (4). Ideally, a monitor will intelligently organize observation data into useful information, adapt its functions according to surgical and anesthesia events, select the most relevant information to display, modify its display layouts to reduce distraction and amplify essential information, tune safety boundaries for its warning systems on the basis of the individual medical conditions of the patient, analyze data to help diagnosis and treatment, and allow user-friendly interactive navigation of the monitor system. Such a monitor will eventually become an extension of a physician's senses and an assistant of decision-making processes.

COMPUTER INFORMATION TECHNOLOGY IN ANESTHESIA

Anesthesia Information Systems

Patient information processing systems have undergone a long history of evolution. Starting in the 1960s, some computer programming software and languages were introduced to construct patient information systems. One example is MUMPS (Massachusetts General Hospital Utility Multi-Programming System), which was developed in Massachusetts General Hospital and used by other hospitals, as well as the U.S. Department of Defense and the U.S. Veteran's Administration. During the same period, Duke University's GEMISCH, a multi-user database programming language, was created to streamline data sharing and retrieval capabilities.

Currently, a typical anesthesia information system (AIS) consists of a central computer station or a server that is interconnected via wired or wireless data communication networks to many subsystems. Subsystems include anesthesia monitors and record-keeping systems in operating rooms, preoperative areas, postanesthesia care units (PACU), ICUs; data entry and record systems of hospital testing labs; office computers of anesthesiologists. The system also communicates with hospital mainframe information systems to further exchange information with in- and out-patient care services, patient database, and billing systems.

Information from an operating room is first collected by medical devices and anesthesia monitors and locally sorted and recorded in the record-keeping system. Selected data are then transmitted to the mainframe server through the data network. Anesthesia events, procedures, physician observations and diagnosis, patient care plans, testing results, drug and fluid data can also be entered into the record-keeping system, and broadcast to the main server and/or other related subsystems.

The main server and observation station provide a center in which patient status in many operating rooms, preoperative area and PACUs can be simultaneously monitored in real time. More importantly, the central anesthesia information system keeps accurately patient data and makes them promptly accessible to many important functions, including patient care assessment, quality assurance, room scheduling, physician assignment, clinical studies, medical billing, regulation compliance, equipment and personnel utility, drug and blood inventory, postoperative in-patient and out-patient service, to name just a few examples (Fig. 4).

One example of AIS is the automation software system CareSuite of PICIS, Inc. (www.picis.com). The system delivers comprehensive perioperative automation. It provides surgical

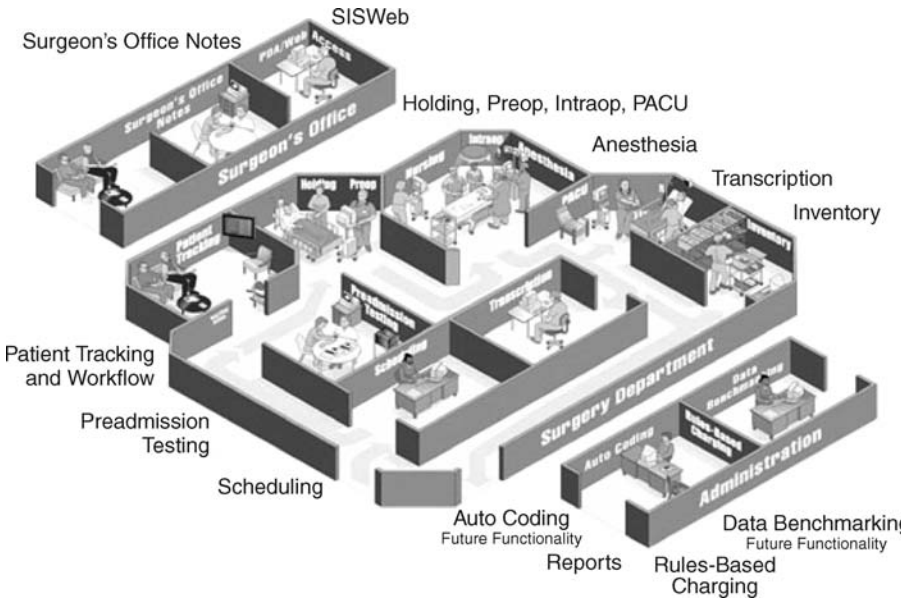


Figure 4. An illustration of a surgical/anesthesia information management system from surgical information systems, Inc. (www. orsoftware.com).

and anesthesia supply management, intraoperative nursing notes, surgical infection control monitoring, adverse event tracking and intervention, patient tracking, resource tracking, outlier alerts, anesthesia record, anesthesia times, case compliance, and so on. Similarly, the surgical and anesthesia management software by Surgical Information Systems (SIS), Inc. (www.orsoftware.com) streamlines patient care and facilitates analysis and performance improvement.

Anesthesia information systems are part of an emerging discipline called medical informatics, which studies clinical information and clinical decision support systems. Although technology maturity of computer hardware and software has made medical information systems highly feasible, creating seamless exchange of information among disparate systems remain a difficult task. This presents new opportunities and challenges for broader application of medical informatics in anesthesia practice.

Computer Simulation: Human Patient Simulators

Human patient simulators (HPS) are computerized mannequins whose integrated mechanical systems and computer hardware, and sophisticated computer software mimic authentically many related physiological, pathological, and pharmacological aspects of the human patient during a surgery or a medical procedure (Fig. 5). The mannequins are designed to simulate an adult or pediatric patient of either gender under a medical stress condition. They are accommodated in a clinical setting, such as an operating room, a trauma site, an emergence suite, or an ICU. Infusion pumps use clean liquid (usually water) through bar-coded syringes, whose barcodes are read by the HPS code recognition system to identify the drugs, to administer the simulated drugs, infusion liquids, or transfused blood during an anesthesia administration. The HPS allows invasive procedures such as intubation.

The HPS responds comprehensively to administered drugs, surgical events, patient conditions, and medical crisis; and displays on standard anesthesia monitors most related physiological vital signs, including blood pressures, heart rate, EKG, and oxygen saturations. They also generate normal and adventurous heart and lung sounds for auscultation. All these characteristics are internally generated by the computer software that utilizes mathematics models of typical human patients to simulate the human responses. The patient's physical parameters (age, weight, smoker, etc.), preexisting medical conditions (high blood pressure, asthma, diabetic, etc.), surgical procedures, clinical events, and critical conditions are easily programmed by a computer Scenario Editor with a user-friendly graphical interface. The Scenario Editor also allows interactive reprogramming of scenarios during an operation. This on-the-fly function of scenario generation is especially useful for training. It gives the instructor great flexibility to create new scenarios according to the trainee's reactions to previous scenarios.

The HPS is a great educational tool that has been used extensively in training medical students, nurse anesthetists, anesthesia residents, emergency, and battlefield

Figure 5. Human patient simulator complex at Wayne State University.

Figure 6. Anesthesia resident training on an HPS manufactured by METI, Inc. (Used with permission.)

medics. Its preliminary development can be traced back to the 1950s, with limited computer hardware or software. Its more comprehensive improvement occurred in pace with computer technology in the late 1960s when highly computerized models were incorporated into high fidelity HPS systems with interfaces to external computers.

Due to rareness of anesthesia crisis, student and resident training on frequent and repeated critical medical conditions and scenarios is not possible in operating rooms. The HPS permits the trainee to practice clinical skills and manage complex and critical clinical conditions by generating and repeating difficult medical scenarios. The instructor can design individualized programs to evaluate and improve trainees' crisis management skills. For invasive skills, such as intubation, practice on simulators is not harmful to human patients. Catastrophic or basic events are presented with many variations so that trainees can recognize their symptoms, diagnose their occurrences, treat them according to established guidelines, and avert disasters. For those students who have difficulties to transform classroom knowledge to clinical hands-on skills, the HPS training is a comfortable bridge for them to rehearse in simulated clinical environments (5) (Fig. 6).

There are several models of HPSs on market. For example, the MedSim-Eagle Patient Simulator (Fig. 7) is a realistic, hands-on simulator of the anesthetized or critically ill patient, developed at Stanford University and manufactured by Eagle Simulation, Inc. METI (Medical Education Technologies, Inc.) (www.meti.com) manufactures adult HPS (Stan), pediatric HPS (PediaSim), emergency care simulator (ECS), pediatric emergency simulator (PediaSim-ECS), and related simulation suites such as airway tools (AirSim), surgical training tools (SurgicalSim). Laerdal Medical AS (www.laerdal.com) has developed a comprehensive portable HPS that can be operated without the usual operating room settings for HPS operations.

Human simulations, however, are not a total reality. Regardless how comprehensive the HPS has become, real environments are far more complex. There are many complications that cannot be easily simulated. Consequences of overly aggressive handling of certain medical catastrophic events may not be fully represented. Issues like these have prompted further efforts in improving HPS technologies and enhancing their utilities in anesthesia education, training, and research.

Figure 7. A human patient simulator SimMan, by Laerdal Medical Corporation. (Used with permission.)

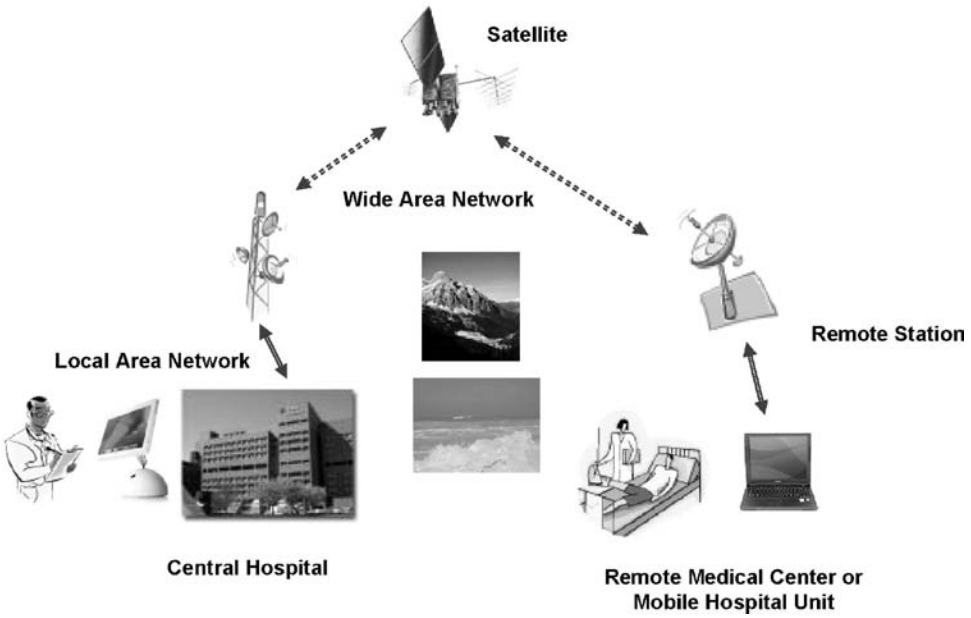


Figure 8. Telemedicine connects remote medical centers for patient care.

Large Area Computer Networks: Telemedicine in Anesthesia

Telemedicine can be used to deliver healthcare over geographically separated locations. High speed telecommunication systems allow interactive video-mediated clinical consultation, and the possibility in the future of remote anesthesia administration. Wide availability of high speed computer and wireless network systems have made telemedicine a viable area for computer applications. Telemedicine could enable the delivery of specialized anesthesia care to remote locations that may not be accessible to high quality anesthesia services and knowledge, may reduce significantly travel costs, and expand the supervision capability of highly trained anesthesiologists.

In a typical telemedicine anesthesia consultation, an anesthesiologist in the consultation center communicates by a high speed network with the patient and the local anesthesia care provider, such as a nurse, at the remote location (Fig. 8). Data, audio and video connections enable the parties to transfer data, conduct conversations on medical history and other consultation routines, share graphs, discuss diagnosis, and examine the patient by cameras. The anesthesiologist can evaluate the airway management, ventilation systems, anesthesia monitor, and cardiovascular systems. Heart and lung sound auscultation can be remotely performed. Airway can be visually examined. The anesthesiologist can then provide consultation and instructions to the remote anesthesia provider on anesthesia management.

Although telemedicine is a technology-ready field of computer applications and has been used in many medicine specialties, at present its usage for systematic anesthesia consultation remains at its infancy. One study reports a case of telemedicine anesthesia between the Amazonian rainforests of Ecuador and Virginia Commonwealth University, via a commercially developed telemedicine system (6). In another pilot study, the University Health Network in Toronto utilized Northern Ontario Remote Telecommunication Health (NORTH) Network to provide telemedicine clinical consultations to residents of central and northern Ontario in Canada (7).

COMPUTER-AIDED ANESTHESIA CONTROL

The heart of most medical decisions is a clear understanding of the outcome from drug administration or from specific procedures performed on the patient. To achieve a satisfactory decision, one needs to characterize outcomes (outputs), establish causal links between drugs and procedures (inputs) and the outcomes, define classes of decisions in consideration (classes of possible actions and controllers), and design actions (decisions and control). Anesthesia providers perform these cognitive tasks on the basis of their expertise, experience, knowledge of guidelines, and their own subjective judgments. It has long been perceived in the field of anesthesiology that computers may help in this decision and control process.

At a relatively low level of control and decision assistance, there has been routine use by anesthesia providers of computers to supply comprehensive and accurate information about anesthesia drugs, procedures, and guidelines in relation to individual patient care. Thanks to miniaturization and internets, there are now commonly and commercially available digital reference databases on anesthesia drugs, their detailed user manuals, and anesthesia procedures. With a palm-held device, all information becomes readily available to anesthesia providers in operating rooms, and other clinical settings. New data can be routinely downloaded to keep information up-to-date.

More challenging aspects of computer applications are those involving uncertainty, control, and intelligence that are the core of medical decision processes. These include individualized models of human patients, outcome prediction, computer-assisted control, diagnosis, and decision assistance. Such tools need to be further developed and commercialized for anesthesia use.

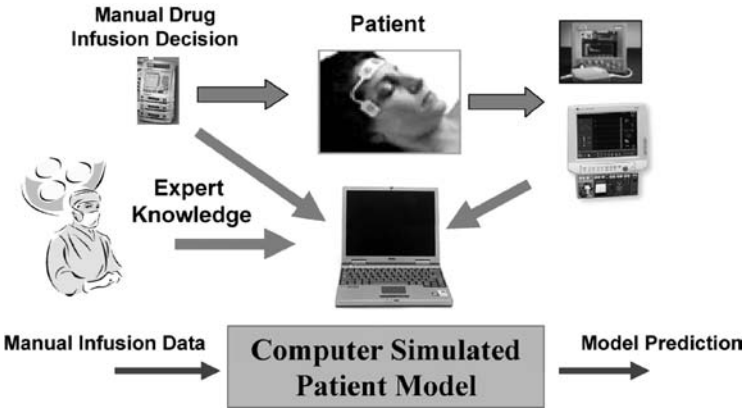


Figure 9. Utility of patient models to predict outcomes of drug infusion.

Patient Modeling and Outcome Prediction

Response of a patient's physiological and pathological state to drugs and procedures is the key information that an anesthesiologist uses in their management. The response, that is, the outcome, can be represented by either the values of the patient vital signs such as anesthesia depth and blood pressures, or consequence values such as length of ICU stay, hospital stay, complications. Usually, drug impact on patient outcomes is evaluated in clinical trials on a large and representative population and by subsequent statistical analysis. These population-based models (average responses of the selected population) link drug and procedural inputs to their effects on the patient state. These models can then be used to develop anesthesia management guidelines (Fig. 9).

For real-time anesthesia control in operating rooms, the patient model also must represent dynamic aspects of the patient response to drugs and procedures (8). This real-time dynamic outcome prediction requires a higher level of modeling accuracy, and is more challenging than off-line statistical analysis of drug impact. Real-time anesthesia control problems are broadly exemplified by anesthesia drug infusion, fluid resuscitation, pain management, sedation control, automated drug rates for diabetics, and so on.

There have been substantial modeling efforts to capture pharmacokinetic and pharmacodynamic aspects of drug impact as well as their control applications (9). These are mostly physiology-based and compartment-modeling approaches. By modeling each process of infusion pump dynamics, drug propagation, concentration of drugs on various target sites, effect of drug concentration on nerve systems, physiological response to nerve stimulations, and sensor dynamics, an overall patient response model can be established. Verification of such models has been performed by comparing model-predicted responses to measured drug concentration and physiological variables. These models have been used in evaluating drug impact, decision assistance and control designs.

Computer Automation: Anesthesia Control Systems

At present, an anesthesiologist decides on an initial drug control strategy by reviewing the patient's medical conditions, then adapts the strategy after observing the patient's actual response to the drug infusion. The strategy is further tuned under different surgical events, such as incision, operation, and closing. Difficulties in maintaining smooth and accurate anesthesia control can have dire consequences, from increased drug consumption, side effects, short- and long-term impairments, and even death. Real-time and computer-assisted information processing can play a pivotal role in extracting critical information, deriving accurate drug outcome predictions, and assisting anesthesia control.

Research efforts to develop computer-assisted anesthesia control systems have been ongoing since the early 1950s (10–14). The recent surge of interest in computer-assisted anesthesia diagnosis, prediction, and controls is partly driven by the advances in anesthesia monitoring technologies, such as depth measurements, computer-programmable infusion pumps, and multisignal real-time data acquisition capabilities. These signals provide fast and more accurate information on the patient state, making computer-aided control a viable possibility. Research findings from computer simulations, animal studies, and limited human trials, have demonstrated that many standard control techniques, such as proportional-integral-derivative (PID) controllers, nonlinear control techniques, fuzzy logic, model predictive control, can potentially provide better performance under routine anesthesia conditions in operating rooms (Fig. 10).

Target Concentration and Effect Control. Target concentration or drug effect control is an open-loop control strategy. It relies on computer models that relate drug infusion rates to drug concentrations on certain target sites or to drug effects on physiological or nerve systems. Since at present drug concentration or drug effects are not directly measured in real-time, feedback control is often not possible. Implementation of this control strategy can be briefly described as follows. For a prespecified time interval, the desired drug concentration profile is defined. This profile is usually determined *a priori* by expert knowledge, safety mandates, and smooth control requirements. A performance index is then devised that includes terms for control accuracy (to follow the desired profiles closely), drug consumption, constraints on physiological variables (safety constraints), and so on. Then, an optimal control is derived by optimizing the performance index under the given constraints and the dynamic

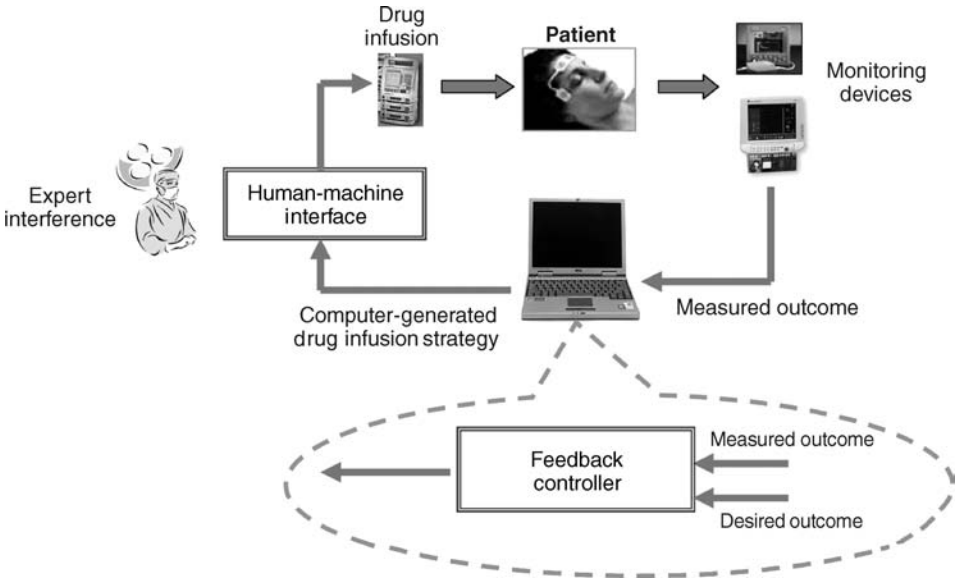


Figure 10. Computer-assisted drug infusion control (a) without expert interference: automated anesthesia feedback control. (b) With expert interference: anesthesia decision assistance systems.

models of the patient. One common method of designing optimal control strategies is dynamic programming, although many other optimal or suboptimal control design methodologies for nonlinear control systems are also available in the control field. Due to lack of feedback correction in target concentration control, optimality and accuracy of control actions may be compromised. However, even this open-loop control has seen many successful applications, such as glucose level control. Feedback control may become feasible in the future when new sensors become available to measure directly drug concentration.

Automatic Feedback Control. Computer-assisted anesthesia control has been frequently compared to autopilot systems in aviation. The autopilot system controls flying trajectories, altitude and position, airplane stability and smoothness, automatically with minimum human supervision. Success of such systems has resulted in their ubiquitous applications in most airplanes. It was speculated that an anesthesia provider's routine control tasks during a surgery may be taken over by a computer that adjusts automatically drug infusions to maintain desirable patient states. Potential and speculated advantages of such systems may include less reliance on experience, avoidance of fatigue-related medical mistakes, smoother control outcomes, reduced drug consumptions, and consequently faster recovery. So far, these aspects have been demonstrated only in a few selective cases of research subjects.

System Identification and Adaptive Control for Individualized Control Strategies. One possible remedy for compensating variations in surgical procedures and patient conditions in control design is to use real-time data to adjust patient models that are used in either target concentration control or feedback control. Successful implementation of this idea will generate individualized models that will capture the unique characteristics of the patient. This real-time patient model can then be used to tune the controllers that will deliver best performance for the patient. This control tuning is the core idea of adaptive control systems: Modifying control strategies on the basis of individually identified patient models. Adaptive control has been successfully applied to a vast array of industrial systems. The main methods of model reference adaptive control, gain scheduling, self-tuning regulators, and machine learning are potentially applicable in anesthesia control. This is especially appealing since variations and uncertainties in patient conditions and surgical events and procedures are far more complicated than industrial systems.

Most control algorithms that have been employed in anesthesia control are standard. The main difficulties in applying automated anesthesia control are not the main control methodologies, but rather an integrated system with high reliability and robustness, and well-designed human-machine interaction and navigation. Unlike an airplane in midair or industrial systems, anesthesia patients vary vastly in their responses to drugs and procedures. Control strategies devised for a patient population may not work well in individual patients. Real-time, on-site, and automatic calibration of control strategies are far more difficult than designing an initial control strategy for a patient population. Adaptation adds a layer of nonlinear feedback over the underlying control, leading to adaptive PID, tuned fuzzy, adaptive neural frameworks, and so on. Stability, accuracy, and robustness of such control structures are more difficult to establish. Furthermore, human interference must be integrated into anesthesia control systems to permit doctors to give guidelines and sometimes take control. Due to high standard in patient safety, at present automated anesthesia control remains largely in a phase of research, and in a very limited sense, toward technology transfer to medical devices. It will require a major commercialization effort and large clinical studies to transform research findings into product development of anesthesia controllers.

Moreover, medical complications occur routinely, which cannot be completely modeled or represented in control

strategies. Since such events usually are not automatically measured, it is strenuous to compensate their impact quickly. In addition, medical liability issues have raised the bar of applying automated systems. These concerns have curtailed a widespread realization of automated anesthesia control systems, despite a history of active research over four decades on anesthesia control systems.

COMPUTER INTELLIGENCE: DIAGNOSIS AND DECISION ASSISTANCE

In parallel to development of automatic anesthesia control systems, a broader application of computers in anesthesia management is computer-aided anesthesia diagnosis, decision assistance, and expert systems. Surveys of anesthesia providers have indicated that the field of anesthesiology favors system features that advise or guide rather than control (15). Direct interventions, closed-loop control, lockout systems, or any other coercive method draw more concerns. In this aspect, it seems that anesthesia expert decision support systems may be an important milestone to achieve before automated systems.

Anesthesia Diagnosis and Decision Assistance

Computer-aided diagnosis will extract useful information from patient data and vital-sign measurements, apply computerized logic and rigorous evaluations of the data, provide diagnosis on probable causes, and suggest guideline-driven remedy solutions. The outcome of the analysis and diagnosis can be presented to the anesthesia care provider with graphical displays, interactive user interfaces, and audio and visual warnings.

Decision assistance systems provide decision suggestions, rather direct and automatic decision implementations. Such systems provide a menu of possible actions for an event, or dosage suggestions for control purposes, and potential consequences of selected decisions. Diagnosis of possible causes can remind the anesthesiologist what might be overlooked in a crisis situation. The system can have interactive interfaces to allow the physician to discuss further actions and the corresponding outcomes with the computer. This idea of physician-assistant systems aims to provide concise, timely, and accurate references to the anesthesiologist for improved decisions. Since the physician remains as the ultimate decision maker, their management will be enhanced by the available information and diagnosis, but not taken over.

Suggested remedies of undesirable events are essentially recommendations from anesthesia management guidelines, brought out electronically to the anesthesiologist. Some computer simulators for anesthesia education are developed on the basis of this idea. For example, Anesthesia Simulator by Anesoft Corporation (www.anesoft.com) contains a software module of expert consultation that incorporates anesthesia emergency scenarios and suggests expert advices. Utility of expert systems in resident training has been widely accepted. However, decision support systems in the operating rooms are slow in development and acceptance. Generally speaking, a decision support system must interact with the complicated cognitive environment of the operating rooms. To make such systems a useful tool, they must be designed to accommodate the common practice in which the anesthesiologist thinks, sees, and reasons, rather than imposing a complicated new monitoring mode for the clinician to be retrained. This is again an issue of human-factors design.

Dosage recommendations for anesthesia drugs are internally derived from embedded modeling and control strategies. In principle, the control strategies discussed in the previous sections can support the decision assistance system. By including the physician in the decision loop, some issues associated with automated control systems can be alleviated. Reliability of such control strategies, user interfaces, and clinical evidence of cost-effectiveness of the decision support system will be the key steps toward successful clinical applications of such systems.

FUTURE UTILITY OF COMPUTER TECHNOLOGY IN ANESTHESIA

The discussions in the previous sections outline briefly critical roles that computers have played in improving anesthesia management. New development in computer-related technologies are of much larger potential.

Micro-Electro-Mechanical Systems (MEMS) is a technology that integrates electrical and mechanical elements on a common silicon material. This technology has been used in developing miniature sensors and actuators, such as micro infusion pumps and *in vivo* sensors. Integrated with computing and communication capabilities, these devices become smart sensors and smart actuators. The MEMS technology has reached its maturity. Further into the realms of fabrication technology at atom levels, emergence of nanotechnology holds even further potential of new generations of medical devices and technologies. There are many exciting possibilities for utility of these technologies in anesthesia: *In vitro* sensors based on nanodevices can potentially pinpoint drug concentrations at specific target sites, providing more accurate values for automated anesthesia drug control; Microactuators can directly deliver drugs to the target locations promptly and accurately, reducing drastically reliance on trial-and-error and sharpened experience in anesthesia drug infusion control; MEMS and nanosensors together with computer graphical tools will allow two-dimensional (2D) or three-dimensional (3D) visual displays of drug propagation, drug concentration, distributed blood pressures, heart and lung functions, brain functions, consequently assisting anesthesiologists in making better decisions about drug delivery for optimal patient care.

On another frontier of technology advancement, computer parallel computing (many computers working in symphony to solve complicated problems), computer imaging processing, data mining (extracting useful information from large amount of data), machine intelligence, wireless communication technologies, and human-factors science and design provide a vast opportunity and a promising horizon in advancing anesthesia management.

Advanced anesthesia control systems will manage routine drug infusion with their control actions tuned to

individual patients' conditions and surgical procedures, relieving anesthesiologists from stressful and strenuous routine tasks to concentrate on higher level decisions in patient care.

Patient physiological conditions can be more accurately and objectively measured by computer-processed sensor and imaging information.

Computer-added imaging processing will make it possible to consolidate information from CT-Scan (Computed Tomography), TEE (Transesophageal Echocardiography), MRI (Magnetic Resonance Imaging), and fMRI (Functional Magnetic Resonance Imaging) into regular anesthesia monitoring.

Anesthesia decisions will be assisted by computer database systems and diagnosis functions.

Anesthesia monitoring devices will become wireless, eliminating the typical spaghetti conditions of monitoring cables in operating rooms.

Anesthesia information systems will become highly connected and standard in anesthesia services, automating and streamlining the total patient care system: From patient admission to patient discharge, as well as follow-up services.

BIBLIOGRAPHY

Cited References

1. Penelope M, et al. Advanced patient monitoring displays: tools for continuous informing. Anesthes Analges 2005;101: 161–168.
2. Bonhomme V, et al. Auditory steady-state response and bispectral index for assessing level of consciousness during propofol sedation and hypnosis. Intravenous Anesthes 2000; 91:1398–1403.
3. Drummond JC. Monitoring depth of anesthesia. Anesthesiology 2000;93(3):876–882.
4. Blike GT. Human factors engineering: It's all about 'usability'. ASA Newslett Oct. 2004;68.
5. Peteani LA. Enhancing clinical practice and education with high-fidelity human patient simulators. Nurse Educ 2004; 29(1):25–30.
6. Stephen W, et al. Case report of remote anesthetic monitoring using telemedicine. Anesthes Analges 2004;98:386–388.
7. Wong DT, et al. Preadmission anesthesia consultation using telemedicine technology: A pilot study. Anesthesiology June 2004;100(6):1605–1607.
8. Wang LY, Yin G, Wang H. Identification of Wiener models with anesthesia applications. Int J Pure Appl Math Sci 2004; 35–61.
9. Shafer A, Doze VA, Shafer SL, White PF. Pharmacokinetics and pharmacodynamics of propofol infusions during general anesthesia. Anesthesiology 1988;69:348–356.
10. Eisenach JC. Reports of scientific meetings-workshop on safe feedback control of anesthetic drug delivery. Anesthesiology August 1999;91:600–601.
11. Linkens DA, Hacisalihzade SS. Computer control systems and pharmacological drug administration: A survey. J Med Eng Technol 1990;14(2):41–54.
12. Mortier EM, et al. Closed-loop controlled administration of Propofol using bispectral analysis. Anaesthesia 1998;53: 749–754.
13. Rao RR, et al. Automaded regulation of hemodynamic variables. IEEE Eng Med Biol Mag 2001;20:24–38.
14. Tackley RM, et al. Computer controlled infusion of propofol. Br J Anesthes 1989;62:46–53.
15. Beatty PT, et al. User attitudes to computer-based decision support In anesthesia and critical care: A preliminary survey. Internet J Anesthesiol 1999;3(1).

See also ANESTHESIA MACHINES; CARDIAC OUTPUT, THERMODILUTION MEASUREMENT OF; ELECTROCARDIOGRAPHY, COMPUTERS IN; ELECTROENCEPHALOGRAPHY; MEDICAL RECORDS, COMPUTERS IN; MONITORING IN ANESTHESIA.

ANGER CAMERA

MARK T. MADSEN
University of Iowa

INTRODUCTION

In nuclear medicine, radioactive tracers are used to provide diagnostic information for a wide range of medical indications. Gamma-ray emitting radionuclides are nearly ideal tracers, because they can be administered in small quantities and yet can still be externally detected. When the radionuclides are attached to diagnostically useful compounds (1), the internal distribution of these compounds provides crucial information about organ function and physiology that is not available from other imaging modalities. The Anger camera provides the means for generating images of the radiopharmaceuticals within the body. Example images of some common studies are shown in Figs. 1 and 2.

Initially, nonimaging detectors were used to monitor the presence or absence of the radiotracer. However, it was clear that mapping the internal distribution of the radiotracers would provide additional diagnostic information. In 1950, Benedict Cassen introduced the rectilinear scanner. The rectilinear scanner generated images of radionuclide distributions by moving a collimated sodium iodide detector over the patient in a rectilinear fashion. The detected count rate modulated the intensity of a masked light bulb that scanned a film in an associated rectilinear pattern. While this device did produce images, it was very slow and had no capability for imaging rapidly changing distributions. The rectilinear scanner was used into the 1970s, but was finally supplanted by the Anger camera (2–5).

The Anger camera, also referred to as the scintillation camera (or gamma camera), is a radionuclide imaging device that was invented by Hal O. Anger. It is the predominant imaging system in nuclear medicine and is responsible for the growth and wide applicability of nuclear medicine. Anger was born in 1920. He received his BS degree in electric engineering from the University of California at Berkeley in 1943 and in 1946 he began working at the Donner Laboratories, where he developed a large number of innovative detectors and imaging devices including the scintillation well counter and a whole body rectilinear scanner using 10 individual sodium iodide probes. In 1957, he completed his first gamma

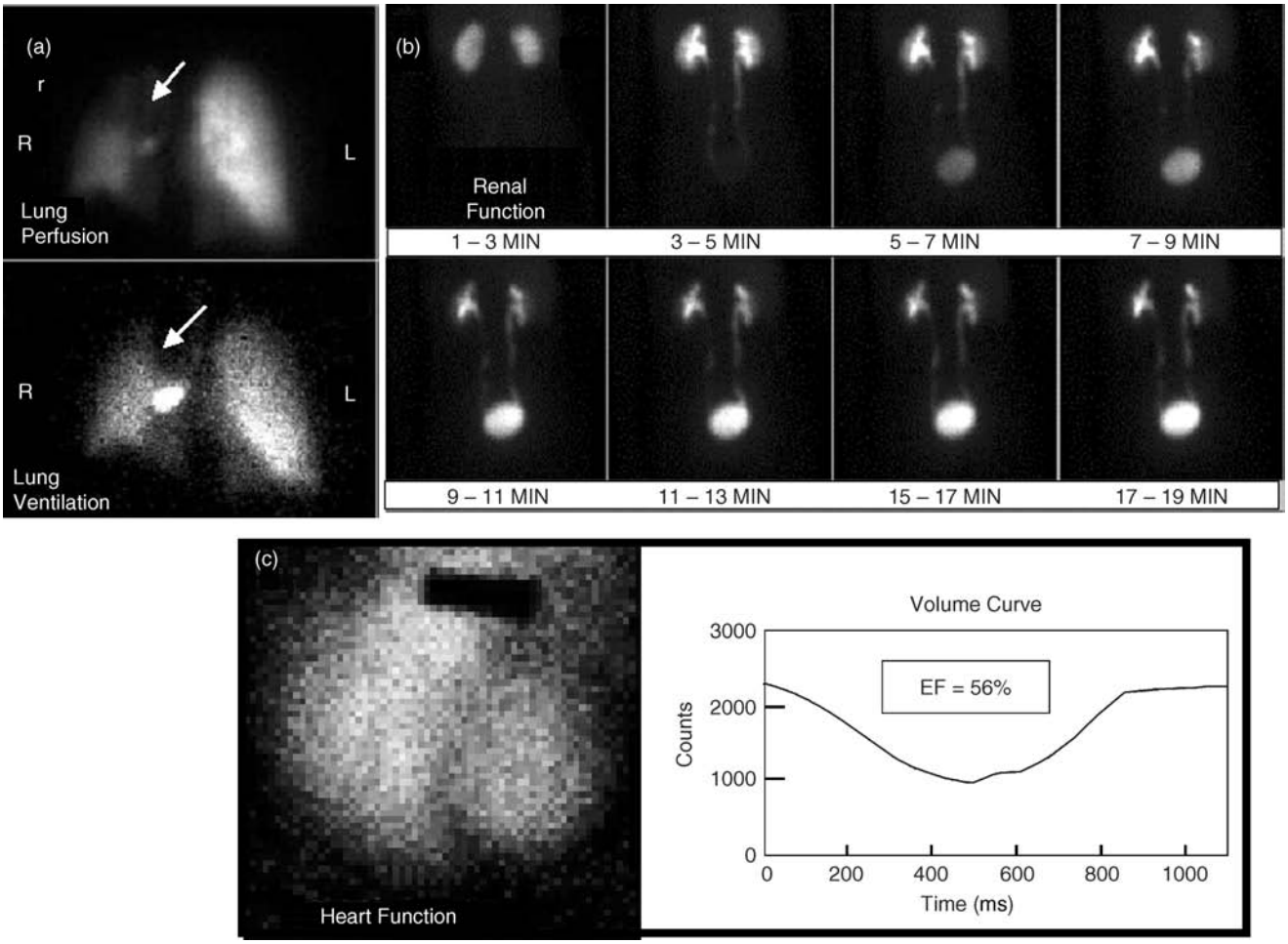


Figure 1. Anger camera clinical images. A. Lung ventilation and perfusion images are used to diagnose pulmonary emboli. B. Renal function images are used to diagnose a variety of problems including renal hypertension and obstruction. C. Gated blood pool studies permit evaluation of heart wall motion and ejection fraction.

imaging camera that he called a scintillation camera and is often referred to as the Anger camera (6). Anger's scintillation camera established the basic design that is still in use today. The first Anger camera had a 10 cm circular field of view, seven photomultiplier tubes, pinhole collimation, and could only be oriented in one direction. A picture of this initial scintillation camera is shown in Fig. 3 and a schematic drawing of the electronics is shown in Fig. 4. In 1959, Anger adapted the scintillation camera for imaging positron emitting radionuclides without collimation using coincidence between the camera and a sodium iodide detector. He also continued improving the scintillation camera for conventional gamma emitting radionuclides. By 1963, he had a system with a 28 cm field of view and 19 photomultiplier tubes (7). This device became commercialized as the nuclear Chicago scintillation camera. Throughout the 1960s, 1970s, and 1980s Anger remained active at Donner labs developing numerous other radionuclide imaging devices. He has received many prestigious awards including the John Scott Award (1964), a Guggenheim fellowship (1966), an honorary Doctor of Science degree from Ohio State University (1972), the Society of Nuclear Medicine Nuclear Pioneer Citation (1974), and the Society of Nuclear Medicine Benedict Cassen Award (1994) (8–10).

About the same time that the Anger camera was introduced, the molybdenum-99/technetium-99m radionuclide generator became available. This finding is mentioned because the advantages offered by this convenient source of ^{99m}Tc had a large influence on the development of the Anger camera. Technetium-99m emits a single gamma ray at 140 keV, has a 6 h half-life and can be attached to a large number of diagnostically useful compounds. Because it is available from a generator, it also has a long shelf life. The ^{99m}Tc is used in $> 80\%$ of nuclear medicine imaging studies. As a result, both the collimation and detector design of the Anger camera has been optimized to perform well at 140 keV (1).

SYSTEM DESCRIPTION

The Anger camera is a position sensitive gamma-ray imaging device with a large field of view. It uses one

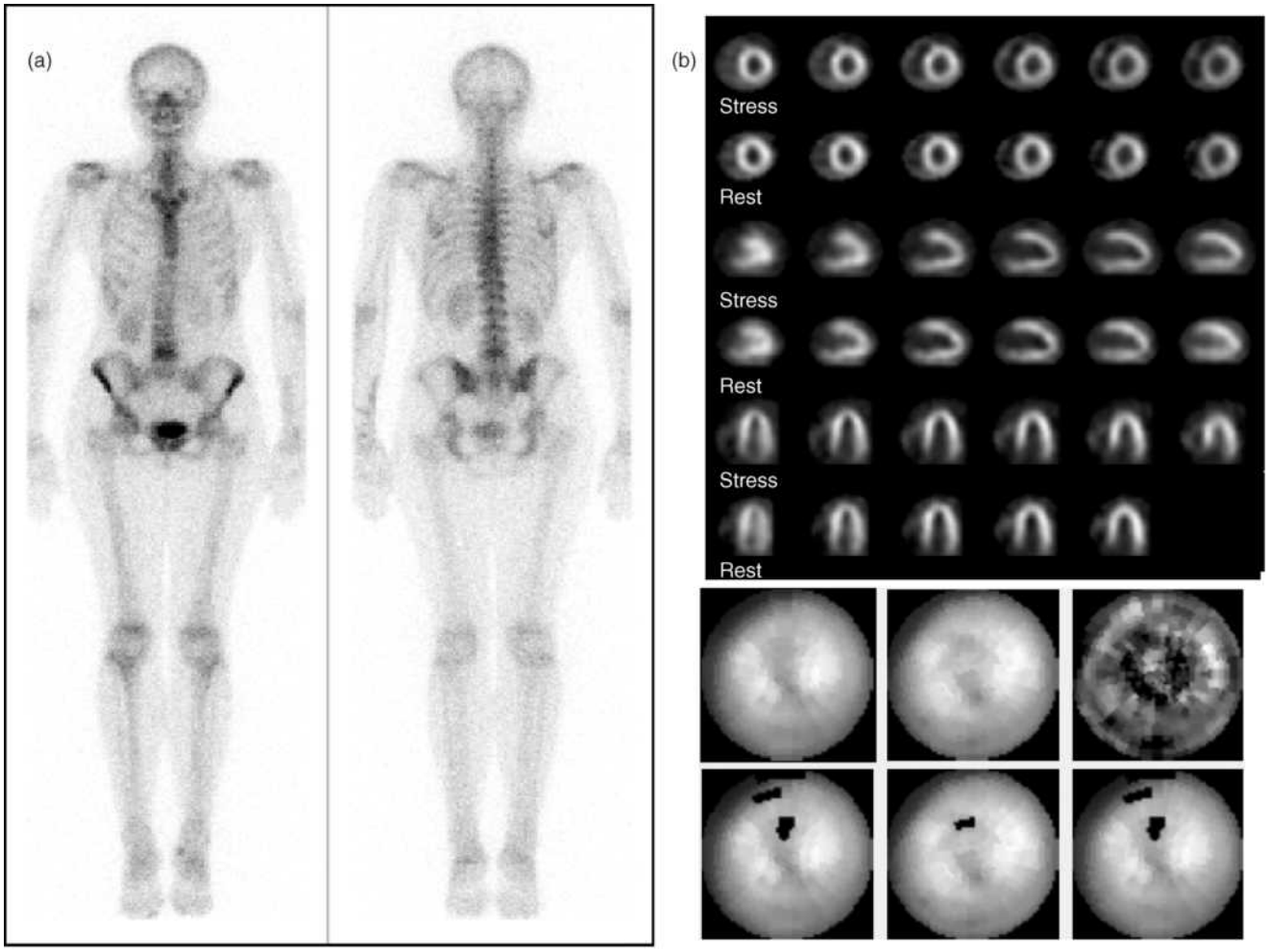


Figure 2. Anger camera clinical images. A. Bone scans are used to evaluate trauma and metastatic spread of cancer. B. Myocardial perfusion studies are used to evaluate coronary artery disease.

large, thin sodium iodide crystal for absorbing gamma-ray energy and converting that into visible light. The light signal is sampled by an array of photomultilier tubes that convert the light signal into an electronic pulse. The pulses from individual PMTs are combined in two ways. An energy pulse is derived from the simple summation of the PMT signals. The X and Y locations of the event are calculated from the sum of the PMT signals after position-dependent weighting factors have been applied. When a signal from a detected event falls within a preselected energy range, the X and Y locations are recorded in either list or frame modes. The components that make up the Anger camera are shown in Fig. 5 and are described in detail in the following section (11,12).

Sodium Iodide Crystal

Sodium iodide activated with thallium, NaI(Tl), is the detecting material used throughout nuclear medicine. Sodium iodide is a scintillator giving off visible light when it absorbs X- or gamma-ray energy. At room temperature, pure NaI has very low light emission, however, when small amounts (parts per million, ppm) of thallium are added, the

Figure 3. Initial Anger camera used to image a patient's thyroid with I-131. The field of view of this device was 10 cm. (Reprinted from Seminars in Nuclear Medicine, Vol 9, Powell MR, H.O. Anger and his work at Donner Laboratory, 164–168., 1979, with permission from Elsevier.)

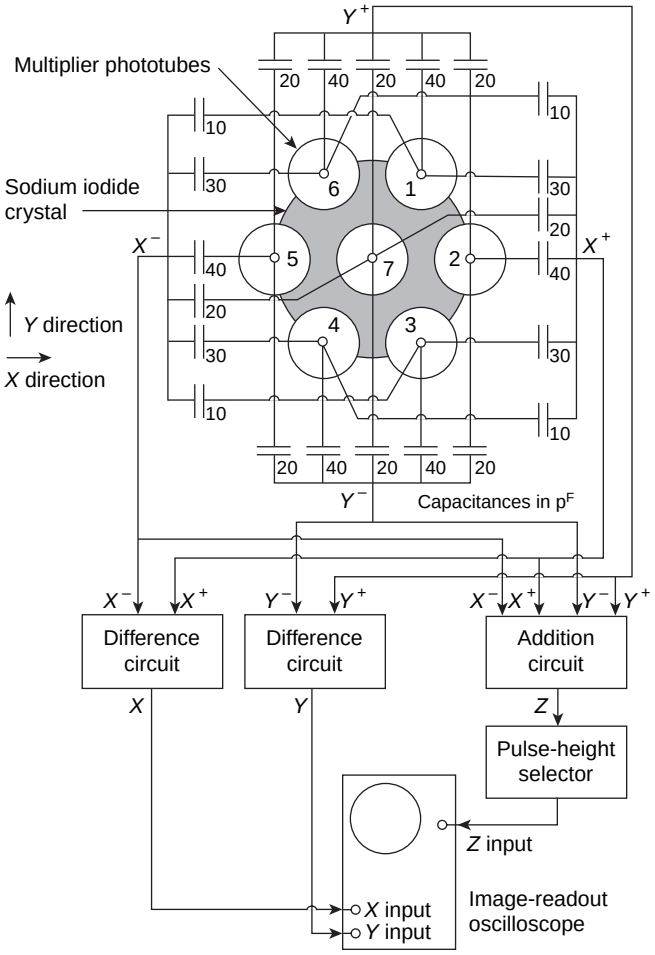


Figure 4. Electronic schematic for Anger's first scintillation camera. The photomultiplier tube position weighting was accomplished with a capacitor network. (From Instrumentation in Nuclear Medicine, Hine and Sorenson, Elsevier, 1967.)

efficiency for light emission is greatly enhanced. This is an especially important aspect for its application in the Anger camera since the light signal is used to determine both the energy and location of the gamma-ray interaction with the detector. In addition to its high light ouput, NaI(Tl) has several other desirable properties. It has a relatively high effective atomic number ($Z_{\text{eff}} = 50$) and the density is 3.67 g·cm^{-3}. This results in a high detection efficiency for gamma rays under 200 keV with relatively thin crystals (8,12–15).

Sodium iodide is grown as a crystal in large ingots at high temperatures (> 650 °C). The crystals are cut, polished, and trimmed to the required size. For Anger cameras, the crystals are typically 40 × 55 cm and 9.5 mm thick. Because NaI(Tl) absorbs moisture from the air (hygroscopic), it must be hermetically sealed. Any compromise of this seal often results in the yellowing of the crystal and its irreversible destruction.

In addition to the need to keep the crystal hermetically sealed, the temperature of the detector must be kept relatively constant. Temperature changes > 2 °C·h^{-1} will often shatter the detector.

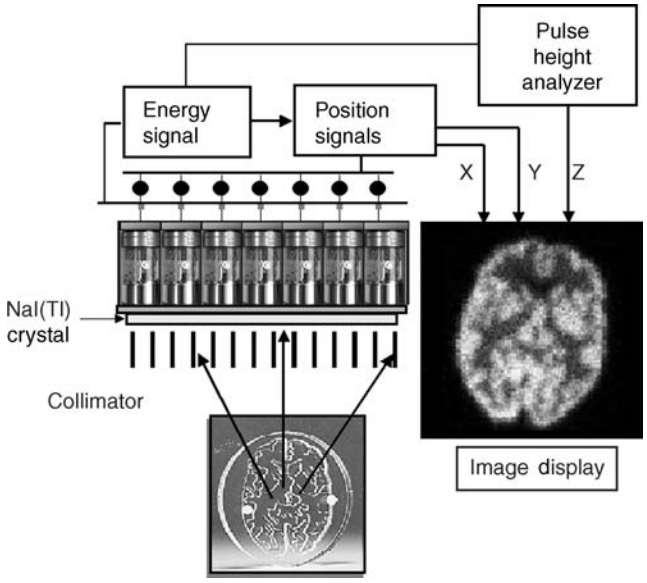


Figure 5. Anger camera components.

Light Pipe

The scintillation light generated in the crystal is turned into electronic signals by an array of photomultiplier tubes (PMTs). These signals provide both event energy and localization information. It is desirable that the magnitude of the signal from the photomultiplier tube be linearly related to the event location as shown in Fig. 6. However, when the PMTs are in close proximity to the crystal, the relationship between the signal magnitude and the event location is very nonlinear. In early designs of the Anger camera, a thick transparent material referred to as a light pipe was coupled to the crystals to improve spatial linearity and uniformity. Glass, lucite, and quartz have been used for this purpose. Design enhancement of the light pipe included sculptured contouring to improve light collection and scattering patterns at the PMT interface to reduce positional nonlinearities (Fig. 7). In the past decade, many of the spatial nonlinearities have been corrected algorithmically operating on digitized PMT signals. This has allowed manufacturers to either reduce the thickness of the light pipe or completely eliminate it (2,16,17).

PMT Array

The visible light generated by the absorption of a gamma ray in the NaI(Tl) crystal carries location and energy information. The intensity of the scintillation is directly proportional to the energy absorbed in the event. To use this information, the scintillation must be converted into an electronic signal. This is accomplished by photomultiplier tubes. In a PMT, the scintillation light liberates electrons at the photocathode and these electrons are amplified through a series of dynodes. The overall gain available from a photomultiplier tube is on the order 10^6.

Photomultiplier tubes are manufactured in a wide variety of shapes and sizes. Those with circular, hexagonal, and square photocathodes have all been used in Anger cameras. Hexagonal and square PMTs offer some advantages for

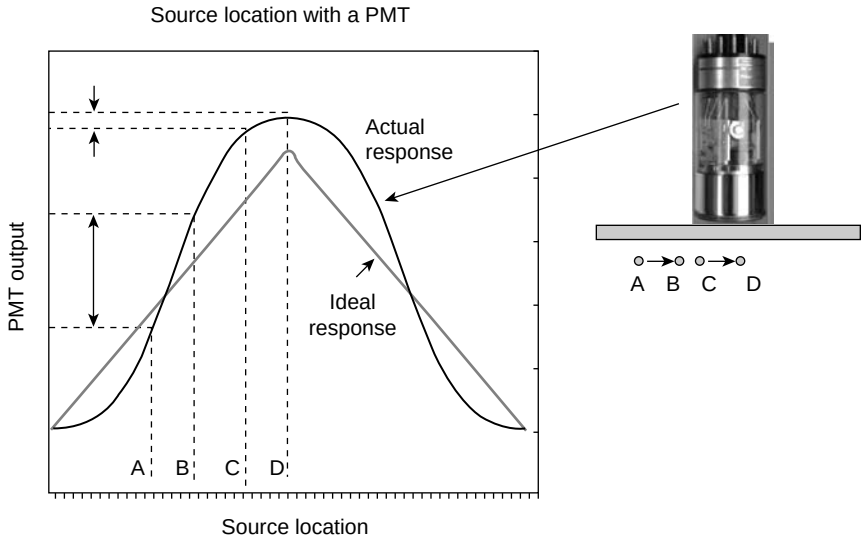


Figure 6. Photomultiplier tube response with source position. The ideal response can be approximated by interposing a light pipe between the crystal and photomultiplier tube.

close packing the PMTs over the surface of the detector. However, the sensitivity of all PMTs falls off near the edge of the field, so that "dead" space between the PMTs is unavoidable.

It is important to determine the energy of the detected event. Gamma rays that are totally absorbed produce a scintillation signal that is directly proportional to the gamma-ray energy. Thus, the signal resulting from the unweighted sum of the PMTs represents gamma-ray energy. This signal is sent to a pulse height analyzer. Scattered radiation from within the patient can be rejected by setting an energy window to select events that have resulted from the total absorption of the primary gamma ray. Gamma rays that have been scattered in the patient necessarily lose energy and are (largely) not included.

The position of the gamma-ray event on the detector is determined by summing weighted signals from the PMTs (2,6,7,12,16,18,19). Each PMT contributes to four signals: X^+, X^-, Y^+, Y^-. The magnitude of the contribution is determined both by the amount of light collected by the PMT and its weighting factor. For the tube located exactly at the center of the detector, the four weighting factors are equal. A tube located along the x axis on the left side (e.g., tube 5 in Fig. 4) contributes equally to Y^+ and Y^-, has a large contribution to X^-, and a small contribution to X^+. In Anger's original design, the weighting factors were provided by capacitors with different levels of capacitance (Fig. 4). In commercial units, the capacitor network was replaced by resistors (Fig. 8). In the past decades, the resistor weighting matrix has been largely supplanted by digital processing where nonlinear weighting factors can be assigned in software (Fig. 9).

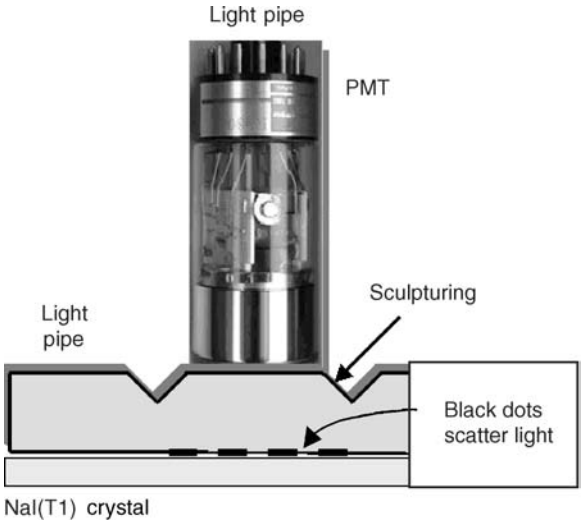


Figure 7. The light pipe is a transparent light conductor between the crystal and the photomultiplier tubes. The sculpturing grooves and black dot pattern spread the light to improve the positioning response of the photomultiplier tube.

It is clear that PMTs located near the event collect most of the scintillation light while those far away get relatively little. Because each PMT has an unavoidable noise component, the PMTs that receive the least light increase the error associated with the event localization. Initially, all the PMTs were included. Later, in order to eliminate PMTs that have little real signal, diodes were used to set current thresholds. In digital Anger cameras, the PMT thresholds are set in software (20–22).

The weighted signals from the PMTs are summed together to generate four position signals: X^+, X^-, Y^+, and Y^-. The X and Y locations are determined from: $(X^+ - X^-)/Z$ and $(Y^+ - Y^-)/Z$, where Z is the energy signal found from the unweighted sum of the PMT signals discussed above. This energy normalization is necessary to remove the size dependence associated with the intensity of the scintillation. This is not only important for imaging radionuclides with different gamma-ray energies, but it also improves the spatial resolution with a single gamma ray energy because of the finite energy resolution of the system. The energy signal is also sent to a pulse height analyzer where an energy window can be selected to include only those events associated with total absorption of the primary gamma.

Image Generation

Anger camera images are generated in the following way (see Fig. 10). A gamma ray is absorbed in the NaI(Tl)

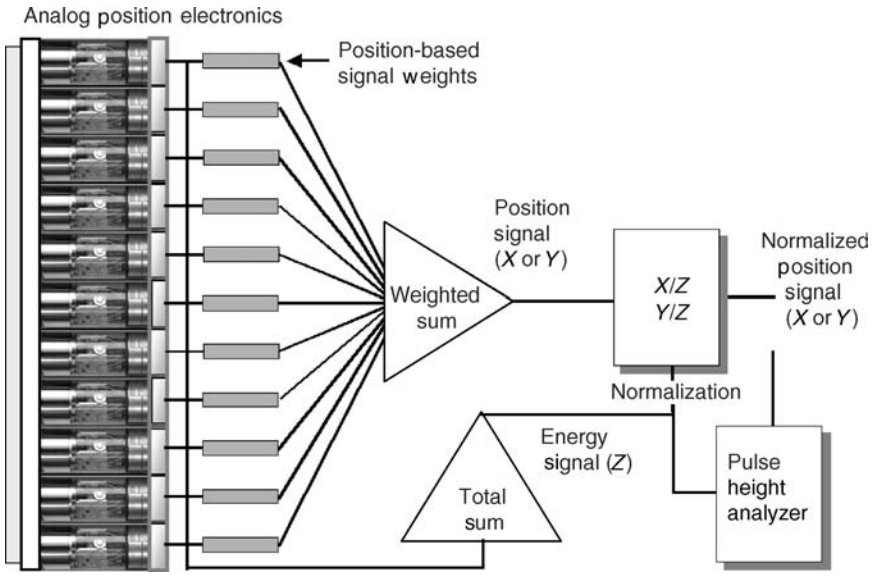


Figure 8. The Anger camera of the 1980s used resistors to provide position weighting factors. The energy signal was used both for normalization and scatter discrimination.

crystal and the resulting scintillation light is sampled by the PMT array to determine the event energy and location. The energy signal is sent to a pulse height analyzer and if the signal falls within the selected energy window, a logic pulse is generated. At the same time, the X and Y coordinates of the event are determined and the logic pulse from the PHA enables the processing of this information. For many years, Anger camera images were generated photographically with the enabled X and Y signals intensifying a dot on a cathode ray tube (CRT) viewed by a camera. In modern Anger cameras, the CRT has been replaced with computer memory and the location information is digital. The X and Y coordinate values, still enabled by the PHA, point to a memory element in a computer matrix. The contents of that memory element are incremented by 1. Information continues to accrue in the computer matrix until the count or time stopping criteria are met.

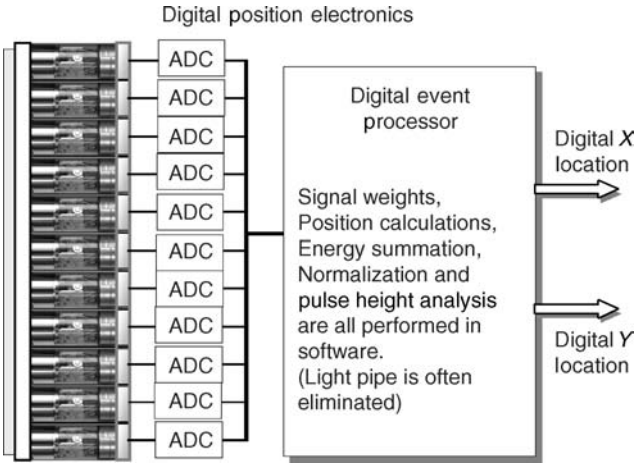


Figure 9. Digital Anger camera electronics. The photomultiplier tube signals are digitized so that signal weighting, energy and position determination are performed in software rather than with digital electronics.

The image generation described in the previous paragraph is referred to as frame or matrix mode. The information can also be stored in list mode where the X and Y coordinate of each event is stored sequentially along with a time marker. The list mode data can then be reconstructed at a later time to any desired time or spatial resolution.

Collimation

In order to produce an image of a radionuclide distribution, it has to be projected onto the detector. In a conventional camera, image projection is accomplished by the camera lens. However, gamma rays are too energetic to be focused with optics or other materials. The first solution to projecting gamma-ray images was the pinhole collimator. The pinhole collimator on an Anger camera is conceptually identical to a conventional pinhole camera. There is an inversion of the object and the image is magnified or minified depending on the ratio of the pinhole to detector distance and the object to pinhole distance. Pinhole collimators are typically constructed out of tungsten and require lead shielding around the "cone". Because the amount of magnification depends on the source to pinhole distance, pinhole images of large, three-dimensional (3D) distributions are often distorted. In addition, the count sensitivity falls off rapidly for off-axis activity. A better solution for most imaging situations is a multiholed parallel collimator (Fig. 11). As the name implies, the parallel collimator consists of a large number of holes with (typically) lead septae. Most parallel collimators have hexagonal holes that are $\sim$ 1.5 mm across and are 20–30 mm long. The septal walls are typically 0.2 mm thick. The parallel hole collimator produces projections with no magnification by brute force. Gamma rays whose trajectories go through the holes reach the detector while those with trajectories that intersect the septae are absorbed. Less than 1 out of 5000 gamma rays that hit the front surface of the collimator are transmitted through to the detector to form the image (2,7,11,12,20,23,24).

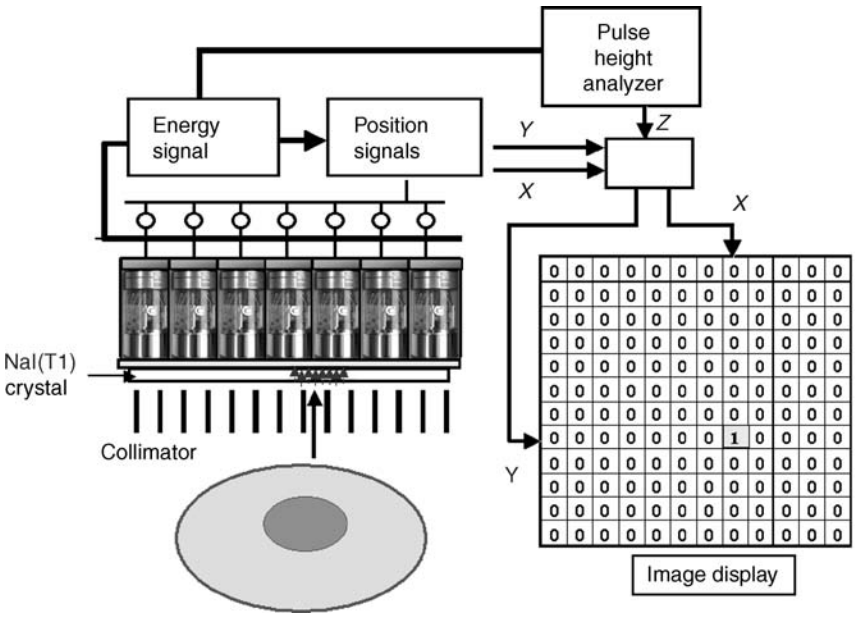


Figure 10. Schematic of Anger camera showing how image information is acquired.

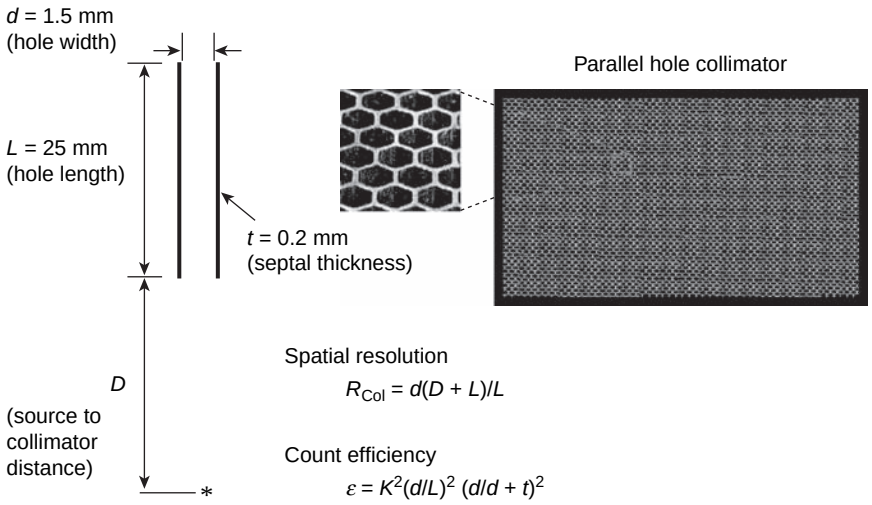


Figure 11. Collimation. Collimators are the image forming aperture of the Anger Camera, but are also the limiting component in spatial resolution and count sensitivity.

The spatial resolution of the collimator, R_{col}, is determined by the hole size (d), hole length (L), and the source to collimator distance (D): $R_{col} = d(L + D)/L$. The efficiency of a parallel hole collimator is expressed as $K^2(d/L)^2\ (d/d + t)^2$, where K is a shape factor constant equal to 0.26 and t is the septal wall thickness. Collimation design is an optimization problem since alterations in d and L to improve resolution will decrease count sensitivity. Collimator spatial resolution has a strong dependence on the source to collimator distance. As shown in Fig. 12, the spatial resolution rapidly falls with distance. However, the count sensitivity of a parallel hole collimator is *not* affected by the source distance because the parallel hole geometry removes the divergent rays that are associated with the inverse square loss. Another factor that influences collimator design is the energy of the gamma ray being imaged. Higher energy gamma rays require thicker septae and

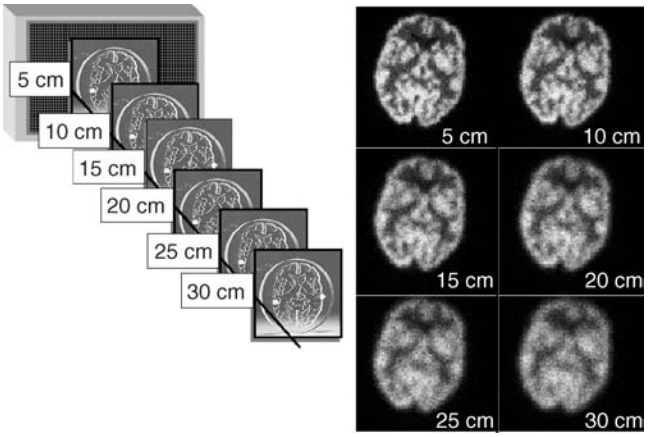


Figure 12. Spatial resolution dependence on source to collimator distance.

larger holes resulting in poorer resolution and count sensitivity (2,7,25).

CORRECTIONS

The analog Anger camera had a number of limitations because of nonlinearities in the position signals and because of the uneven light collection over the NaI(Tl) crystal. As digital approaches became viable over the last two decades, a number of corrections to improve energy, spatial, and temporal resolution have evolved. All of these corrections, and particular those involving spatial and energy resolution, require system stability. This challenge was significant because of the variation in PMT output associated with environmental conditions and aging. In order for corrections to be valid over an extended period of time, a method to insure PMT stability has to be implemented. Several different approaches have evolved. In one method, PMT gains are dynamically adjusted to maintain consistent output signals in response to stabilized light emitting diodes (LEDs). An LED is located beneath each PMT where its light is sampled 10–100 times·s^{-1}. Gains and offsets on the PMTs are adjusted so that the resulting signal is held close to its reference value. Another approach uses the ratio of photopeak/Compton plateau counts from a ^{99m}Tc or ^{57}Co source as the reference.

Flood Field Image

When the Anger camera is exposed to a uniform flux of gamma rays, the resulting image is called a flood field image. Flood field images are useful for identifying nonuniform count densities and may be acquired in two different ways. An intrinsic flood field is obtained by removing the collimation and placing a point source of activity 1.5–2 m from the detector. An extrinsic flood field is obtained with the collimator in place and with a large, distributed source (often called a flood source) placed directly on the collimator. Flood field sources using ^{57}Co are commercially available. Alternatively, water-filled flood phantoms are available into which ^{99m}Tc or other radionuclide can be injected and mixed.

Energy Correction

The energy signal represents the total energy absorbed in a gamma-ray interaction with the detector. This signal is determined by the intensity of the scintillation and by how much of the scintillation light is captured by the PMTs. Because the efficiency for sampling the scintillation is position dependent, there are fluctuations in the energy signals across the detector as shown in Fig. 13. These variations degrade energy resolution and have a significant effect on the performance of the scintillation camera that limit corrections for nonuniformity. The idea of using a reference flood field image to correct nonuniformities has been around for a long time. However, if the reference flood field image is acquired with little or no scattered radiation (as it often is), the correction factors are not appropriate during patient imaging. The reason is that scattered radiation and the amount of scatter entering the selected energy

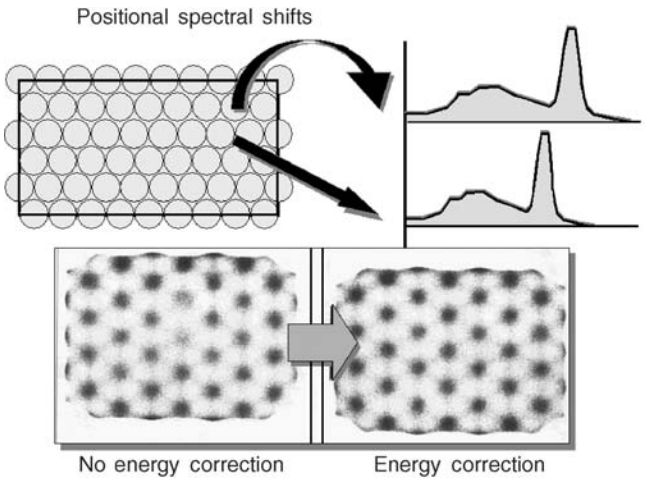


Figure 13. Energy correction. Local spectral gain shifts are evident across the crystal because of the variable sampling imposed by the photomultiplier tube array.

window will be position dependent. Energy correction electronics was introduced in the late 1970s that essentially generated an array of energy windows that are adjusted for the local energy spectra. Typically, the detector field of view is sampled in a 64 × 64 matrix and a unique energy window is determined for each matrix element. With the energy windows adjusted on the local photopeaks, the variations in the scatter component are greatly reduced. As shown in Fig. 13, energy correction does not significantly improve intrinsic field uniformity. Its role is to reduce the influence of source scatter on the other correction factors (11,20,26–28).

Spatial Linearity Correction

The nonlinearities in the PMT output with respect to source location causes a miss-positioning of events when Anger logic is used. This finding can be demonstrated by acquiring an image of a straight line distribution or a grid pattern. The line image will have a "wavy" appearance (Fig. 14). In the early 1980s, a method to improve the spatial linearity was developed. An imaging phantom array of precisely located holes in a sheet of lead is placed on the uncollimated detector and is exposed to a point source of ^{99m}Tc located 1–2 m away. The image of the hole pattern is used to calculate corrective x and y offsets for each point in the field of view. These correction factors are stored in a ROM. When an event is detected and the Anger logic produces x and y coordinates, the offsets associated with these coordinates are automatically added generating the new, accurate event location. Improving the spatial linearity has a profound affect on field uniformity as can be seen in Fig. 14 (17,28–31).

Uniformity Correction

After the energy and spatial linearity corrections have been made, there are still residual nonuniformities that are present in a flood field image. Typically, these will vary $< 10\%$ from the mean count value for the entire field. A high count reference flood field image can be acquired and

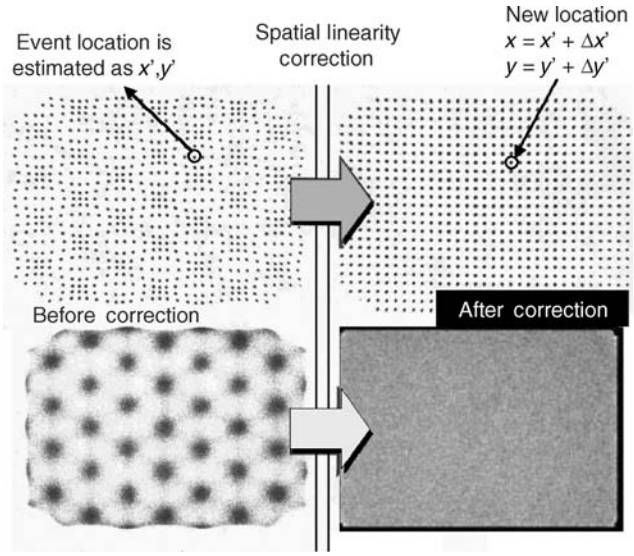


Figure 14. Spatial linearity correction. Accurate correction for inaccurate event localization has a profound effect on field uniformity.

this image is then used to generate regional flood correction factors that are then applied to subsequent acquisitions (Fig. 15) (17,29,32).

Pulse Pileup Correction

An Anger camera processes each detected event sequentially. Because the scintillation persists with a decay time of 230 ns, pulses from events occurring closely in time are distorted from summation of the light. This distortion is referred to as pulse pileup. As the count rate to the detector increases, the amount of pulse pileup also increases and becomes significant at count rates $> 30{,}000$ cps. For much of conventional nuclear medicine imaging, this is not a problem since the count rate is typically well below that level. There are certain applications such as coincidence positron emission tomography (PET) imaging where the detectors are exposed to event rates that can exceed 1,000,000 cps. Because the Anger logic used to establish the event location is essentially a centroid method, pulse pileup causes errors. An example of this is shown in Fig. 16, which shows an Anger camera image of four high-count rate sources. In addition to the actual sources, false images of source activity between the sources are also observed (33,34). The effects of pulse pileup can be minimized by electronic pulse clipping, where the pulse is forced to the baseline before all the light has been emitted and processed. While this increases the count rate capability, it compromises both spatial and energy resolution, which are optimal when the whole pulse can be sampled and integrated. One approach to reduce losses in spatial and energy resolution is to alter the integration time event-by-event, based on count rate demands. In addition, algorithms have been developed that can extrapolate the pulse to correct for

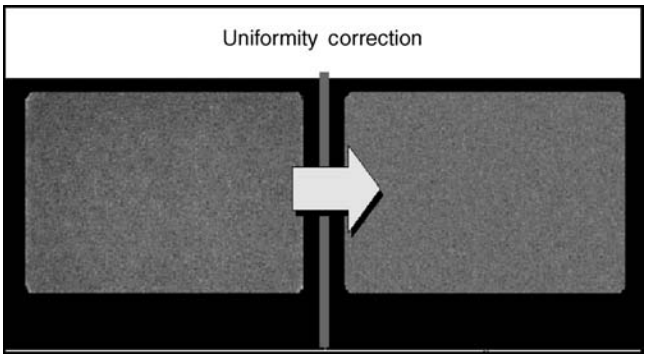


Figure 15. Uniformity correction. Residual non-uniformities can be reduced by skimming counts based on a reference flood field image.

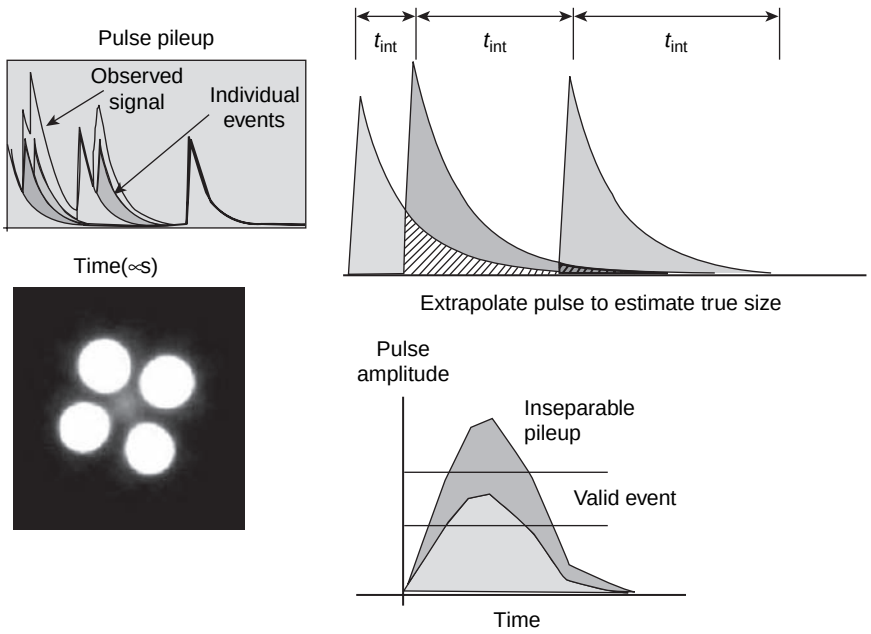


Figure 16. Pulse pileup correction. Pulse pileup correction improves the count rate capability and reduces the spurious placement of events.

its contribution to a second pileup pulse. This process can be repeated if a third pileup is also encountered. When this correction is performed at the PMT level, it reduces the "false" source images discussed above (35).

PERFORMANCE

Uniformity

When the Anger camera is exposed to a uniform flux of gamma rays, the image from that exposure should have a uniform count density. Anger cameras with energy, spatial linearity, and uniformity correction are uniform to within 2.5% of the mean counts.

Intrinsic Spatial Resolution

The intrinsic spatial resolution refers to the amount of blurring associated with the Anger camera independent of the collimation. It is quantified by measuring the full width at half-maximum (fwhm) of the line spread response function. The intrinsic spatial resolution for Anger cameras varies from 3 to 4.5 mm depending on the crystal thickness and the size of the PMTs. Another way of evaluating intrinsic spatial resolution for gamma-ray energies < 200 keV is with a quadrant bar phantom consisting of increasingly finer lead bar patterns where the bar width is equal to the bar spacing (Fig. 17). Typically an Anger camera can resolve a 2 mm bar pattern.

Extrinsic Spatial Resolution

The extrinsic spatial resolution, also referred to as the system spatial resolution, refers to the amount of blurring associated with Anger camera imaging. It depends on the collimation, gamma-ray energy, and the source to collimator distance. The standard method for determining the extrinsic resolution is from the fwhm of the line spread response function generated from the image of a line source positioned 10 cm from the collimator. Typical values for the extrinsic spatial resolution range from 8 to 12 mm.

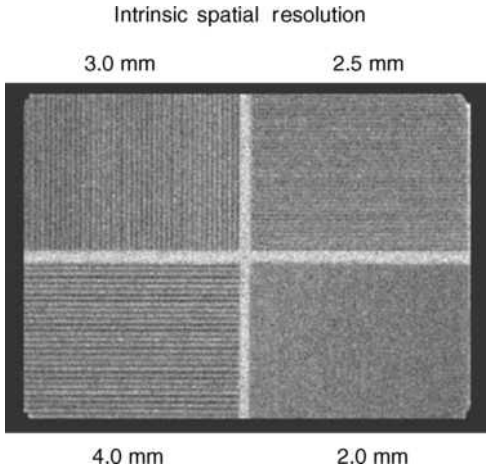


Figure 17. Intrinsic spatial resolution is routinely assessed with bar pattern images. An Anger camera can typically resolve the 2 mm bar pattern.

Energy Resolution

The energy signal generated from gamma-ray absorption in the detector has statistical fluctuations that broaden the apparent energy peaks of the gamma rays. Energy resolution is determined from $100\% \times \text{fwhm}/E\gamma$, where fwhm is of energy peak and $E\gamma$ is the gamma-ray energy. At 140 keV the energy resolution of an Anger camera is 10%. Good energy resolution is important because it permits the discrimination of scattered radiation from the patient. Gamma rays that are scattered in the patient necessarily loose energy and these scattered photons degrade image quality.

Spatial Linearity

Spatial linearity refers to the accurate positioning of detected events. On an Anger camera with spatial linearity correction, the misplacement of events is < 0.5 mm.

Multiwindow Spatial Registration

Because the Anger camera has energy resolution, it can acquire images from radionuclides that emit more than one gamma ray or from two radionuclides. However, the images from different energy gamma rays may have slightly different magnifications or have offsets because of imperfections in the energy normalization. The multiwindow spatial registration parameters quantifies the misalignment between different energy gamma rays (Fig. 18). For Anger cameras, the multiwindow spatial registration is < 2 mm, which is well below the system spatial resolution and therefore is not perceptible.

Count Rate Performance

The count rate capability of Anger camera ranges from 100,000 to 2,000,000 cps depending on the sophistication of the pulse handling technology as discussed above. Anger cameras that are used for conventional nuclear medicine imaging are designed to operate with maximum count rates of 200,000–400,000 cps range, whereas Anger cameras that are used for coincidence imaging require count rate capabilities that exceed 1,000,000 cps (25,28,30,32,36–41).

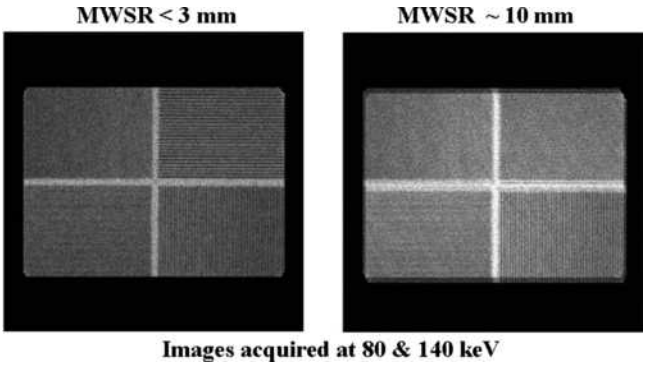


Figure 18. Multi-window spatial registration refers to ability to accurately image different gamma ray energies simultaneously. The figure on the right is an example of poor multi-window spatial registration.

SUMMARY

The Anger camera has been the primary imaging device in nuclear medicine for > 30 years and is likely to remain in that role for at least the next decade. Although it has evolved with the development of digital electronics, the basic design is essentially that promulgated by H.O. Anger. Special purpose imaging instruments based on semiconductor cadmium zinc telluride detectors are actively being pursued as imaging devices for ^{99m}Tc and other low energy gamma emitters. Their pixilated design removes the need for Anger logic position determination and the direct conversion of the absorbed energy into an electronic signal removes the need for photomultiplier tubes allowing compact packaging. However, over the range of gamma-ray energies encountered in nuclear medicine, NaI(Tl) still provides the best efficiency at a reasonable cost.

BIBLIOGRAPHY

Cited References

1. Banerjee S, Pillai MR, Ramamoorthy N. Evolution of Tc-99m in diagnostic radiopharmaceuticals. Semin Nucl Med 2001;31: 260–277.
2. Hine GJ, editor. Instrumentation in Nuclear Medicine. Volume 1, New York: Academic Press; 1967.
3. Pollycove M, Fish MB, Khentigan A. Clinical radioisotope organ imaging–diagnostic sensitivity and practical factors. Rectilinear scanner versus the anger-type scintillation camera. J Nucl Med 1967;8:321–322.
4. Blahd WH. Ben Cassen and the development of the rectilinear scanner. Semin Nucl Med 1996;26:165–170.
5. McCready VR. Milestones in nuclear medicine. Eur J Nucl Med 2000;27:S49–S79.
6. Anger H. A new instrument for mapping gamma ray emitters. Bio Med Quart Rep 1957; UCRL-3653 38.
7. Anger H. Scintillation camera with multichannel collimators. J Nucl Med 1964;5:515–531.
8. Hine GJ. The inception of photoelectric scintillation detection commemorated after three decades. J Nucl Med 1977;18:867–871.
9. Powell MR. H.O. Anger and his work at the Donner Laboratory. Semin Nucl Med 1979;9:164–168.
10. Tapscott E. Nuclear medicine pioneer: Hal O. Anger. First scintillation camera is foundation for modern imaging systems. J Nucl Med 1998;39:15N, 19N, 26N–27N.
11. Murphy PH, Burdine JA. Large-field-of-view (LFOV) scintillation cameras. Semin Nucl Med 1977;7:305–313.
12. Cherry S, Sorenson J, Phelps M. Physics in Nuclear Medicine. Philadelphia: W. B. Saunders; 2003.
13. Muehllehner G. Effect of crystal thickness on scintillation camera performance. J Nucl Med 1979;20:992–993.
14. Royal HD, Brown PH, Claunch BC. Effects of a reduction in crystal thickness on Anger-camera performance. J Nucl Med 1979;20:977–980.
15. Keszthelyi-Landori S. NaI(Tl) camera crystals: imaging capabilities of hydrated regions on the crystal surface. Radiology 1986;158:823–826.
16. Anger H. Scintillation Camera. Rev Sci Instrum 1958;29: 27–33.
17. Genna S, Pang SC, Smith A. Digital scintigraphy: principles, design, and performance. J Nucl Med 1981;22:365–371.
18. Anger H. Scintllation camera with 11 inch crystal. UCRL-11184 (1963).
19. Scrimger JW, Baker RG. Investigation of light distribution from scintillations in a gamma camera crystal. Phys Med Biol 1967;12:101–103.
20. White W. Resolution, sensitivity, and contrast in gamma-camera design: a critical review. Radiology 1979;132:179–187.
21. Zimmerman RE. Gamma cameras—state of the art. Med Instrum 1979;13:161–164.
22. Goodwin PN. Recent developments in instrumentation for emission computed tomography. Semin Nucl Med 1980;10:322–334.
23. Strand SE, Lamm IL. Theoretical studies of image artifacts and counting losses for different photon fluence rates and pulse-height distributions in single-crystal NaI(T1) scintillation cameras. J Nucl Med 1980;21:264–275.
24. Kereiakes JG. The history and development of medical physics instrumentation: nuclear medicine. Med Phys 1987;14:146–155.
25. Chang W, Li SQ, Williams JJ, Bruch PM, Wesolowski CA, Ehrhardt JC, Kirchner PT. New methods of examining gamma camera collimators. J Nucl Med 1988;29:676–683.
26. Budinger TF. Instrumentation trends in nuclear medicine. Semin Nucl Med 1977;7:285–297.
27. Myers WG. The Anger scintillation camera becomes of age. J Nucl Med 1979;20:565–567.
28. Heller SL, Goodwin PN. SPECT instrumentation: performance, lesion detection, and recent innovations. Semin Nucl Med 1987;17:184–199.
29. Muehllehner G, Colsher JG, Stoub EW. Correction for field nonuniformity in scintillation cameras through removal of spatial distortion. J Nucl Med 1980;21:771–776.
30. Muehllehner G, Wake RH, Sano R. Standards for performance measurements in scintillation cameras. J Nucl Med 1981;22: 72–77.
31. Johnson TK, Nelson C, Kirch DL. A new method for the correction of gamma camera nonuniformity due to spatial distortion. Phys Med Biol 1996;41:2179–2188.
32. Murphy PH. Acceptance testing and quality control of gamma cameras, including SPECT. J Nucl Med 1987;28:1221–1227.
33. Strand SE, Larsson I. Image artifacts at high photon fluence rates in single-crystal NaI(T1) scintillation cameras. J Nucl Med 1978;19:407–413.
34. Patton JA. Instrumentation for coincidence imaging with multihead scintillation cameras. Semin Nucl Med 2000;30: 239–254.
35. Wong WH, Li H, Uribe J, Baghaei H, Wang Y, Yokoyama S. Feasibility of a high-speed gamma-camera design using the high-yield-pileup-event-recovery method. J Nucl Med 2001;42: 624–632.
36. O'Connor MK, Oswald WM. The line resolution pattern: a new intrinsic resolution test pattern for nuclear medicine [see comments]. J Nucl Med 1988;29:1856–1859.
37. De Agostini A, Moretti R. Gamma-camera quality control procedures: an on-line routine. J Nucl Med Allied Sci 1989;33:389–395.
38. Lewellen TK, Bice AN, Pollard KR, Zhu JB, Plunkett ME. Evaluation of a clinical scintillation camera with pulse tail extrapolation electronics. J Nucl Med 1989;30:1554–1558.
39. Links JM. Toward a useful measure of flood-field uniformity: can the beauty in the eye of the beholder be quantified? [editorial] [see comments]. Eur J Nucl Med 1992;19:757–758.
40. Hander TA, Lancaster JL, Kopp DT, Lasher JC, Blumhardt R, Fox PT. Rapid objective measurement of gamma camera resolution using statistical moments. Med Phys 1997;24:327–334.
41. Smith EM. Scintillation camera quality control, Part I: Establishing the quality control program. J Nucl Med Technol 1998;26:9–13.

See also COMPUTED TOMOGRAPHY, SINGLE PHOTON EMISSION; IMAGING DEVICES; MICROPOWER FOR MEDICAL APPLICATIONS; NUCLEAR MEDICINE INSTRUMENTATION.

ANGIOPLASTY. See CORONARY ANGIOPLASTY AND GUIDEWIRE DIAGNOSTICS.

ANORECTAL MANOMETRY

ASHOK K. TUTEJA
University of Utah
Salt Lake City, Utah

GEORGE E. WAHLEN
Veterans Affairs Medical Center
and the University of Utah
Salt Lake City, Utah

SATISH S.C. RAO
University of Iowa College of
Medicine
Iowa City, Iowa

INTRODUCTION

The most commonly performed test is the evoluation of anorectal function. These tests can provide useful information regarding the pathophysiology of disorders that affect defecation, continence, or anorectal pain. Anorectal manometry quantifies anal sphincter tone and assesses anorectal sensory response, recto anal reflexes, and rectal compliance. Sensory testing is usually performed along with anorectal manometry and is generally considered a part of the manometry.

The functional anatomy of the anorectum, the equipment, and the technique used for performing anorectal manometry and the parameters for measuring and interpreting the test are discribed in this article. The indications for anorectal manometry are shown in Table 1.

FUNCTIONAL ANATOMY AND PHYSIOLOGY OF THE ANORECTUM

The neuromuscular integrity of the rectum, anus, and the pelvic floor musculature help to maintain normal fecal continence and evacuation. The rectum is an S-shaped muscular tube, which serves as a reservoir and as a pump for emptying stool. The anus is a 2–4 cm long muscular cylinder, which at rest forms an angle with the axis of the rectum of $\sim 90^\circ$. During voluntary squeeze the angle becomes more acute, $\sim 70^\circ$, and during defecation, the angle becomes more obtuse, ~ 110–130° (1,2). The puborectalis muscle, one of the pelvic floor muscles, is responsible for these changes. The anal canal is surrounded by specialized muscles that form the anal sphincters [internal anal sphincter (IAS) and the external anal sphincter (EAS)]. The IAS is 0.5 cm thick and is an expansion of circular smooth muscle layer of the colon. It is an involuntary muscle innervated by fibers of the autonomic nervous system. The EAS is composed of striated muscle, is 0.6–1 cm thick, and is under voluntary control (3). The anus is normally closed by the tonic activity of the IAS. This barrier is reinforced during voluntary squeeze by the EAS. The IAS contributes ~ 70–85% of the resting anal pressure. The IAS does not completely seal the anal canal and requires the endo-anal cushions to interlock and seal the canal. The anal mucosal folds, together with the expansive anal vascular cushions, provide a tight seal. These barriers are further augmented by the puborectalis muscle, which forms a flap-like valve that creates a forward pull and reinforces the anorectal angle to prevent fecal incontinence (3,4). The rectum and the IAS are innervated by the autonomic nervous system. The EAS and the anoderm are supplied by somatic nerves. The mucosa of the rectum and proximal anal canal is lack of somatic sensory innervation. The pudendal nerve, arising from second, third, and fourth sacral nerves is the principal somatic nerve and innervates the EAS, the puborectalis muscle, and the anal mucosa (5).

Table 1. Indications for Anorectal Manometry

Fecal Incontinence
Chronic idiopathic constipation
Diagnosis of Hirschsprung's disease and/or follow up
Megarectum
Pelvic floor dyssynergia
Rectocele
Solitary rectal ulcer
Rectal prolapse
Functional anorectal pain
Neurological diagnostic investigations
Biofeedback training
Pre- and Postsurgery (pouch)

EQUIPMENT FOR ANORECTAL MANOMETRY

The manometric system has two major components: the manometric probe and the pressure recording apparatus. Several types of probes and pressure recorders are available. Each system has distinct advantages and disadvantages. The most commonly used probes and recording devices are reviewed here (6).

Water-Perfused Catheter

This catheter has multiple canals through which water is perfused slowly using a pneumohydraulic pump (Arndorfer, Milwaukee, WI; MUI Scientific Ltd., Toronto, Canada). The infusion rate is 0.5 mL $\cdot$ canal^{-1} $\cdot$ min^{-1}. In the catheter with helicoidal configuration the side holes of the canals are arranged radially and spaced 1, 2, 3, 4, 5, and 8 cm from the "0" reference point. A compliant balloon is tied to one end of the probe, which has a 200 mL capacity. The catheter is placed inside the anorectum, but the pressure transducers are located outside the body and across the flow of water. Resistance generated to the flow of water by luminal contractile activity is quantified as intraluminal pressure. The transducers located on the perfusion pump and the perfusion ports must be at the same level during calibration and when performing the study. The maintenance of the water perfused system requires relatively skilled personnel and air bubbles in the water tubing can affect the pressure recordings. The probe and the recording system are inexpensive and versatile. The closely spaced pressure sensors along the probe can record rectal and anal canal pressures and discriminate between EAS and IAS activity (7).

Solid-State Probe

This system has pressure sensors (microtransducers) that are mounted on the probe. This allows more accurate measurement by placing the pressure sensors directly at the source of pressure activity. The transducers are true strain gauge, that is, they consist of a pressure sensitive diaphragm with semiconductor strain gauges that are mounted on its inner surface. Currently, this is the most accurate catheter system for performing manometery (8). It is user friendly, offers higher fidelity, and is free of limitations imposed by the perfused system. However, it is expensive.

Amplifier–Recorder

The pressure signals that are obtained from the transducer are amplified and recorded on computerized small size amplifiers and recorders (e.g., Polygraph-Medronics/Functional Diagnostics, Minneapolis, MN; Insight, Sandhill Scientific Ltd. Littleton, CO; 7-MPR, Gaeltec, Isle of Sky, UK, and others). They are small, compact, and not only serve as amplifiers and recorders, but also facilitate analysis of data and provide convenient storage for future retrieval of data or for generating a database. No one system is ideal, although each has its strengths and weakness.

STUDY PROTOCOL

General Instructions for Patients Undergoing Anorectal Manometry

In order to maximize uniformity, the manometry should be accomplished with the rectum emptied of feces. The preparation cannot be indispensable for incontinent patients. Constipated patients must be examined several hours after a 500 mL tap water enema or a single Fleets phospho-soda enema. Patients may continue with their routine medications, but the medications should be documented to facilitate interpretation of the data. Patients may eat or drink normally up to the time of the test. Upon arrival at the motility laboratory, the patient may be asked to change into a hospital gown.

The duration of the test is $\sim$ 1 h. The manometry catheter is inserted into the rectum while patients lie on their left side. Patients will feel movement of the catheter and distension of the balloon. After the test, patients can drive home and resume their usual work and diet. It is a safe procedure. There should be little, if any, discomfort during manometry. No anesthetic is used. Absolute contraindication to manometry is recent surgery of the rectum and anal canal, relative contraindication is a poorly compliant patient and rectum loaded with stool.

Patient Position and Digital Examination

It is recommended that the patient is placed in the left lateral position with knees and hips bent to 90°. After explaining the procedure, a digital rectal examination is performed using a lubricated gloved finger. The presence of tenderness, stool, or blood on the finger glove should be noted.

Probe Placement

Next, the lubricated manometry probe is gently inserted into the rectum and oriented such that the most distal sensor (1 cm level) is located posteriorly at 1 cm from anal verge. The markings on the shaft of the probe should aid this orientation.

Run-in Time

After probe placement, a rest (run-in) period should be allowed ($\sim$ 5 min) to give the subject time to relax and allow the sphincter tone to return to basal levels.

Resting Anal Pressure

Currently, two methods are available for assessing this function (9). *Station pull-through*: In this technique, the most distal sensor of a multiport catheter assembly is placed 5 cm above the anal margin. At every 30 s intervals, the catheter is withdrawn by 0.5 cm either manually or with a probe withdrawal device (10). As the sensors straddle the high pressure zone, there is a step up of pressure. The length and the highest pressure of the anal sphincter is then measured. Because pull-through excites anal contraction and the individual is conscious of these movements, the recorded pressure is high (10). For the same reason, a rapid pull through is not an accurate method and is not advisable for measuring anal sphincter function. *Stationary method*: Uses radially arranged multiport catheter, at least three sensors, 1 cm apart that is placed in the anal sphincter zone, that is, 0–3 cm from the anal verge (11). After allowing the tracings to stabilize, the highest sphincter pressure that is observed at any level in the anal canal is taken as the maximum resting sphincter pressure. Resting pressures can be expressed as the average obtained from each transducer or as a range to identify asymmetry of anal canal pressures (12).

Normal anal canal pressures vary according to sex, age, and techniques used (10). Normal values for anorectal manometry are shown in Table 2. There are normal variations in external sphincter pressures both radially and longitudinally (12,14). Anterior quadrant pressures are lower in the orad part of anal canal while posterior quadrant pressures are lower in the distal part of the anal canal. In the mid-anal canal, pressures are equal in all four quadrants. Manometry also enables routine calculation

Table 2. Suggested List of Tests/Maneuver Based on Indication (s)[a]

	Indications for maneuver	
Test	Incontinence	Constipation
Resting pressure	Yes	Yes
Squeeze pressure/duration	Yes	No
Cough reflex	Yes	No
Attempted defecation	No	Yes
RAIR	No	Yes
Rectal sensation	Yes	Yes
Rectal compliance	Optional	Optional

[a]From Ref. 13 with permission.

of anal canal length. Overall pressures are higher in men and younger persons and men have longer anal canals than women. But there is considerable overlap in values and disagreement among various studies about the effect of age and gender on anal canal pressures (10–12,15). Furthermore, subjects with values outside the normal range may not have clinical symptoms and patients with clinical symptoms may exhibit normal values (16).

Squeeze Sphincter Pressure

This pressure can be measured with either the station pull-through or the alternative technique. In the station pull-through technique; after placing the multiport assembly as describe above, at each level the subject is asked to squeeze and to maintain the squeeze for as long as possible (at least 30 s). Alternatively, with a multiport catheter is place, the subject is instructed to squeeze on three separate occasions, with a minutes' rest between each squeeze to allow for recovery from fatigue. The average of the three highest sphincter pressures recorded at any level in the anal canal is taken as the maximum anal squeeze pressure (13). The duration of maximum sustained squeeze should also be determined and is defined as the time interval in seconds during which the subject can maintain a squeeze pressure at or above 50% of the maximum pressure.

Weak squeeze pressures may be a sign of external sphincter damage, neurological damage of the motor pathways, or a poorly compliant patient. Squeeze pressures should be evaluated together with response to cough reflex (16).

Response to Increases in Intraabdominal Pressure

An increase in intraabdominal pressure brought about by asking the subject to blow up a party balloon or by coughing is associated with a reflex increase in the activity of the EAS (11); also called the cough reflex. This reflex response causes the anal sphincter pressure to rise above that of the rectal pressure so that continence is maintained. The response may be triggered by receptors in the pelvic floor and mediated through a spinal reflex arc. In patients with complete supra conal spinal cord lesions, this reflex response is present, but the voluntary squeeze may be absent whereas in patients with lesions of the cauda equina or of the sacral plexus, both the reflex response and the voluntary squeeze are absent.

Rectoanal Pressure Changes During Attempted Defecation

In this maneuver, the subject is asked to bear down, and simulate the act of defecation. The side holes of catheter are located within the anal canal and the rectal balloon is kept inflated. The normal response consists of an increase in rectal pressure coordinated with a relaxation of the intraanal pressure. Alternatively, there may be a paradoxical increase in anal canal pressures, or absent relaxation or incomplete relaxation of the anal sphincter (Fig. 1) (17). It must be appreciated that laboratory conditions may induce artifactual changes, which is a learned response and is under voluntary control.

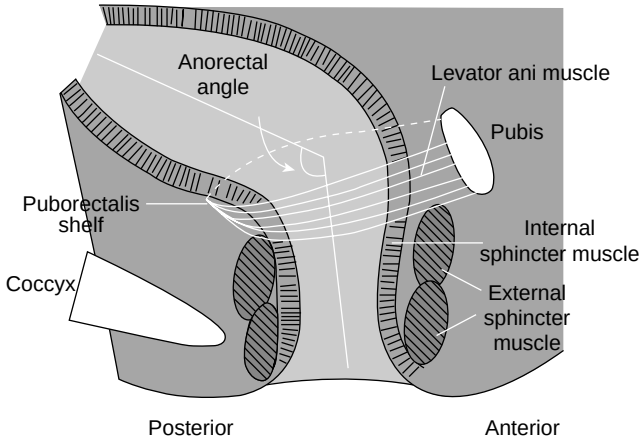


Figure 1. Structures of the anorectum: Reprinted from Ref. 17 with permission from American Gastroenterological Association.

Rectoanal Inhibitory Reflex

This consists of reflex relaxation of the IAS in response to rectal distension. The catheter is positioned with its side holes within the anal canal. Volumes of air are rapidly inflated in the rectal balloon and removed. The inflated time is 10 mL · s. The reflux is evoked with 10, 20, 40, 60, 80, 140, and 200 mL. As the volume of rectal distension is increased, the amplitude and duration of IAS relaxation increases (7). The absolute or relative amplitude of the IAS relaxation depends on the preexisting tone of the IAS and the magnitude of its contribution to the basal anal tone. This reflex may facilitate sampling of rectal contents by the sensory receptors in the upper anal canal and may also help to discriminate flatus, from liquid or solid stools. This reflex is regulated by the intrinsic myenteric plexus. In patients with Hirschsprung's disease and in those with a history of rectal resection and colo- or ileo-anal anastomosis, this reflex is absent. However, in patients with spinal cord injury and in patients with transaction of the hypogastric nerves or lesions of the sacral spinal cord, it is present (18).

Sensory Testing

Rectal Sensory Function. In this technique, the rectal sensory threshold for three common sensations (first detectable sensation, the sensation of urgency to defecate, and the sensation of pain or maximum tolerable volume) is assessed. This can be assessed either by the intermittent rectal distension or by the ramp inflation method.

Intermittent Rectal Distension. This technique is performed by inflating a balloon in the rectum using a handheld syringe. After each inflation, the balloon is deflated completely and after a rest period it is reinflated to the next volume (19).

Ramp Inflation. In this method, the rectum is progressively distended without deflation. This is performed by continuously inflating the balloon at a constant rate with a peristaltic pump or a syringe using increasing volumes of air or fluid or in a stepwise fashion, with a 1 min interval between each incremental inflation of 10–30 cm^3. It is known that the type of inflation (phasic vs. continuous) and the speed of continuous inflation affect the threshold

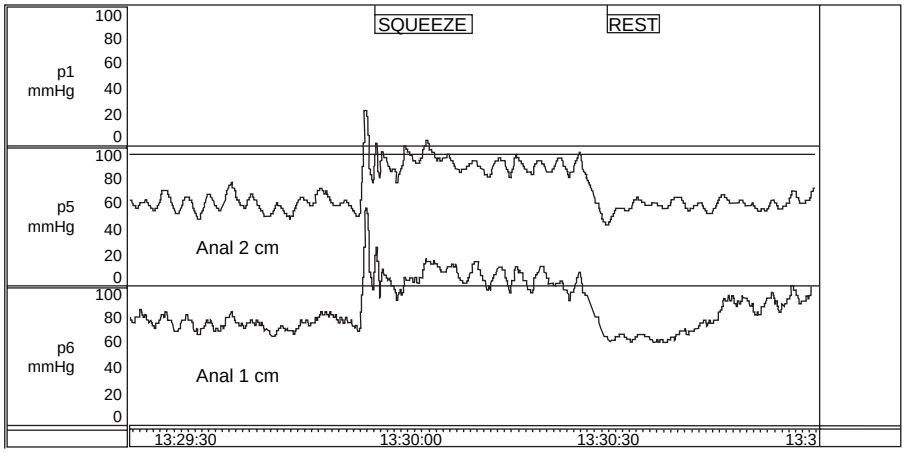


Figure 2. Normal squeeze profile.

volume required for healthy control subjects to perceive distension (20). Also the size and shape of the balloon will affect the threshold volume. Some of this variability can be reduced by using a high compliance balloon and a continuous-infusion pump or a barostat (21).

The maximum tolerable volume or pain threshold may be reduced in patients with a noncompliant rectum (e.g., proctitis) abdominoperineal pull-through, and rectal ischemia (9). Pain threshold also may be reduced in patients with irritable bowel syndrome (22). Higher sensory threshold is seen in autonomic neuropathy, congenital neurogenic anorectal malformations (spinal bifida, Hirschprung's disease, meningocele) and with somatic alteration in rectal reservoir (megarectum, descending perineum syndrome) (20,23). Rectal sensory threshold is altered by change in rectal wall compliance and sensory data should be interpreted along with measurement of rectal compliance (24).

Anal Sensation. At present, assessment of anal canal sensation is not of established value for the diagnosis and treatment of patients with constipation or fecal incontinence (9).

Rectal Compliance

The capacity and distensibility of the rectum are reflected by its compliance. It is a measure of the rectal reservoir function and is defined as the change in rectal volume per unit change in rectal pressure (11). The rectal compliance can be measured by the balloon distension method or more accurately by using a computerized barostat. The higher the compliance, the lower the resistance to distension and vice versa. Low rectal compliance is also seen in patients with acute ulcerative colitis, radiation proctitis, and low spinal cord lesions (20). High compliance is seen in patients with megarectum. Decreased rectal compliance can result in decreased rectal capacity, fecal urgency, and may contribute to fecal incontinence (25).

MANOMETRIC FEATURES OF FECAL INCONTINENCE AND CONSTIPATION

Fecal Incontinence

Anorectal manometry can provide useful information regarding the pathophysiology and management of fecal incontinence (26). Anal sphincter pressures may be decreased in patients with fecal incontinence; either circumferentially or in one quadrant of the anal canal (Fig. 2). Manometry can also determine if compensatory squeeze pressure can be activated. A reduced resting pressure correlate with predominant weakness of IAS and decreased squeeze pressures correlate with EAS defects (27). Two large studies have reported that maximum squeeze pressure has the greatest sensitivity and specificity in discriminating fecal incontinence from continent and healthy controls (28,29). The ability of the EAS to contract reflexly can also be assessed during abrupt increases of intra-abdominal pressure (e.g., when coughing). This reflex response causes the anal sphincter pressure to rise above that of the intrarectal pressure to preserve continence. This reflex response is absent in patients with lesions of the cauda equina or the sacral plexus (18,30). On sensory testing, both hyper- and hyposensitivity can be seen. Assessment of rectal sensation is useful in patients with fecal incontinence associated with neurogenic problems, such as diabetes mellitus (decrease in rectal sensations) or multiple sclerosis (increase in rectal sensation) (31). In some patients, rectal sensory thresholds may be altered because of changes in the compliance of the rectal wall. Patients with megarectum have decreased rectal sensation; and can present with fecal incontinence. Patients with incontinence often have lower rectal compliance (i.e., chronic rectal inschemia, proctitis).

Because of the wide range of normal values in anorectal physiologic testing, no single test can predict fecal incontinence. However, a combination of the tests with clinical evaluation is helpful in assessment of patients with fecal incontinence (32). Anorectal manometry is also useful in evaluating the responses to biofeedback training as well as assessing objective improvement following drug therapy or surgery.

Constipation

Anorectal manometry is useful in the diagnosis of dyssynergic defecation. Manometry helps to detect abnormalities during attempted defecation. Normally, when subjects bear down or attempt to defecate, there is a rise in rectal pressure, which is synchronized with a relaxation of the EAS (Fig. 3). The inability to perform this coordinated

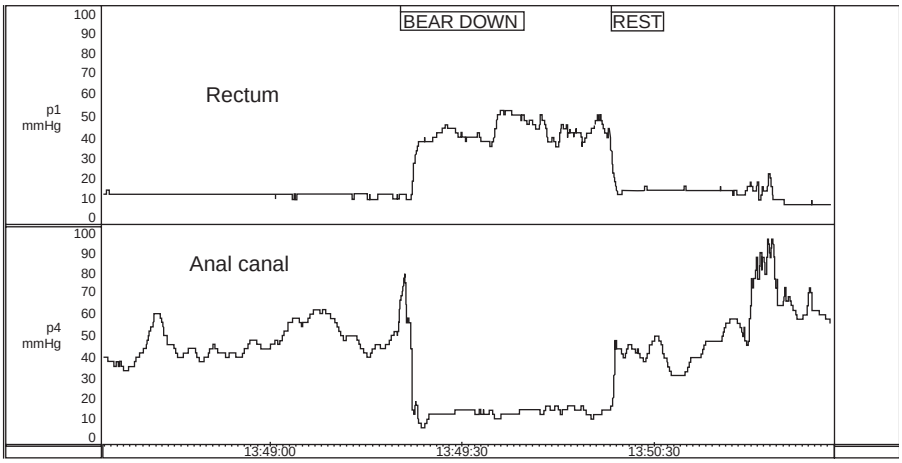


Figure 3. Strain maneuver: A normal coordinated response of the anorectum during attempted defecation shows a rise in rectal pressure associated with a decrease in anal sphincter pressure.

movement represents the chief pathophysiologic abnormality in patients with dyssynergic defecation (17). This inability may be due to impaired rectal contraction, paradoxical anal contraction, impaired anal relaxation, or a combination of these mechanisms (Fig. 4) (Fig. 5). Anorectal manometry also helps to exclude the possibility of Hirschsprung's disease. The absence of the rectoanal inhibitory reflex accompanied by a normal intrarectal pressure increase during distension of the intrarectal balloon is evidence of denervation of the intrinsic plexus at the recto-anal level. Megarectum can cause a falsely negative reflex. In this condition, there is hypotonia of the rectal wall due to a deficiency of visceroelastic properties of the rectum and high degrees of rectal distension are necessary to produce the reflex. In addition to the motor abnormalities, sensory dysfunction may be present. The rectal sensations are reduced in patients with megarectum. The first sensation and the threshold for a desire to defecate may be higher in $\sim$ 60% of patients with dyssynergic defecation (33). The threshold for urge to defecate may be absent on elevated in patients with chronic constipation. Maximum tolerable volume can also be elevated (34). But is not clear whether these findings are the cause or secondary to constipation. When rectal sensation is impaired, neuromuscular conditioning using biofeedback technique can be effective in improving the dysfunction.

Select Appropriate Test/Maneuver

Because anorectal manometry consists of several maneuvers, it is important to determine whether a patient needs all of the maneuvers or only a selection from the array of tests described below. The patient's symptoms and the reason for referral are helpful in choosing the appropriate list. A suggested list is given in Table 3.

Prolonged Anorectal Manometry. It is now feasible to perform anorectal manometry for prolonged periods of time outside the laboratory setting. With the use of this technique, it is possible to measure physiologic functions of the anal sphincter while the person is mobile and free (35). This technique shows promise as an investigational procedure, but its clinical applicability has not been established.

Clinical Utility and Problems with Anorectal Manometry. A systematic and careful appraisal of anorectal function can provide valuable information that can guide treatment of patients with anorectal disorders. Prospective studies have shown that manometric tests of anorectal function provide not only an objective diagnosis, but also a better understanding of the underlying pathophysiology. In

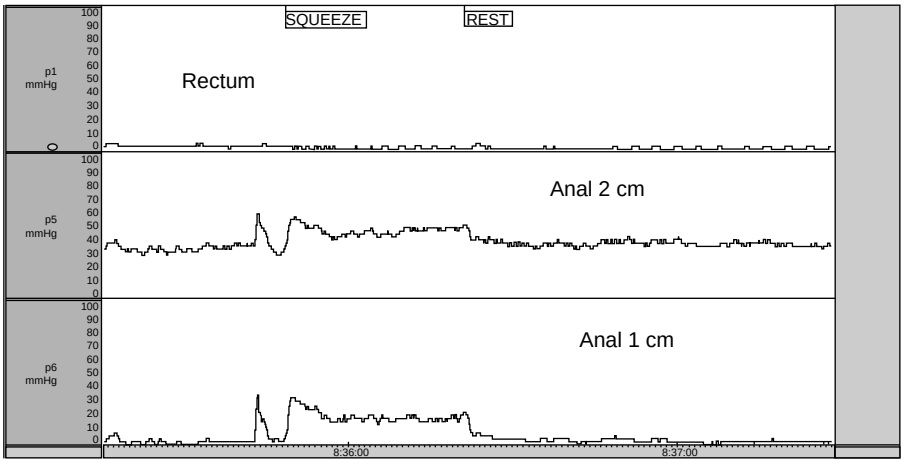


Figure 4. Weak resting and squeeze anal sphincter pressure in a patient with fecal incontinence.

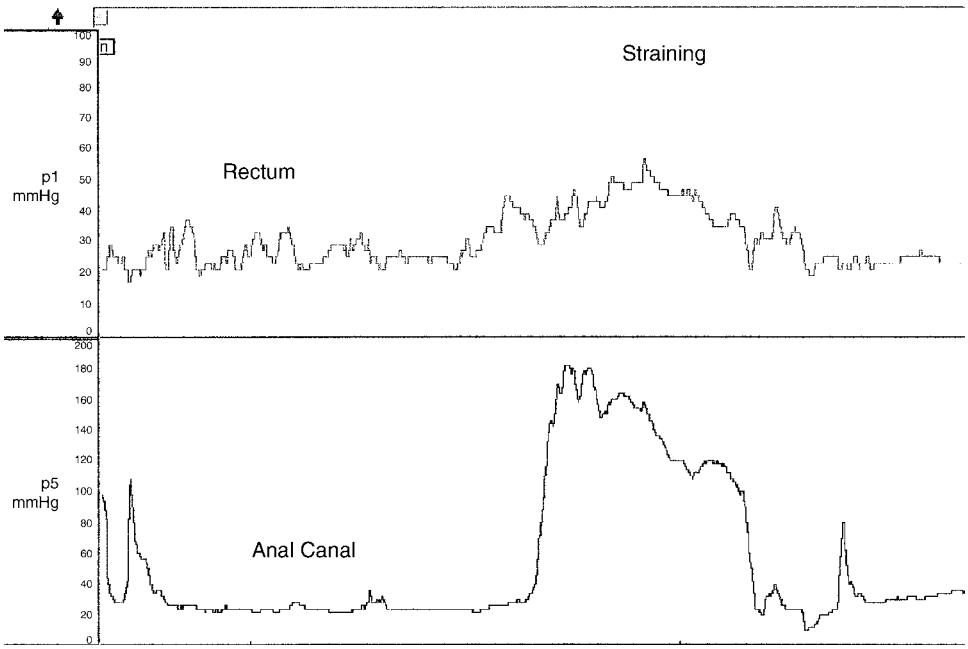


Figure 5. Dyssynergic defecation. During strain maneuver there is rise in intrarectal pressure together with a paradoxical rise in anal sphincter pressure.

addition, it provides new information that could influence the management and outcome of patients with disorder of defecation (36,37).

Anorectal manometry has gained wide acceptance as a useful method to objectively assess the physiology of defecation. However, there are some problems with anorectal physiologic testing. There is a lack of uniformity with regards to the anorectal manometry equipments, methods of performance, and interpretation of the tests. A multiplicity of catheter designs exists, including water-perfused catheters, microtransducers, and microballoons. The techniques of manometric measurement are variable. The catheter can be left at one position (stationary technique), it can be manually moved from one position to another (manual pull through technique), or it can be automatically delivered from one position to another (automatic pull-through technique). If the automated technique is selected, pressure can be recorded while the catheter is at rest or in motion. Pressure can be recorded in centimeters of H_2O, millimeters of mercury (mmHg), or kilopascals (kPa). There is also a relative lack of normative data stratified for age and gender. A more uniform method of performing these tests and interpreting the results is needed to facilitate a wider use of this technecology for the assessment of patients with anorectal disorders. Recently, experts from the American and European Motility Society have described a consensus document, where minimum standards for performing ARM have been described (13). By adopting such standards it is possible to standardize the technique globally that should help diagnosis and interpretation.

Medical Terms:

Compliance: It is defined as the capacity of the organ to stretch (expand) in response to an imposed force.

Defecation: The discharge of feces from the rectum.

Distal: Situated away from the center of the body, or from the point of origin, in contrast to proximal.

Table 3. Normal Manometric Data During Anorectal Manometry[a,b]

	All (n = 45)	Male (N = 18)	Female (N = 22)
Length of anal sphincter, cm	3.7 (3.6–3.8)	4.0 (3.8–4.2)	3.6 (3.4–3.8)
Maximum anal rest pressure, mmHg	67 (59–74)	71 (52–90)	64 (53–75)
Sustained squeeze pressure, mmHg	138 (124–152)	163 (126–200)	117 (100–134)
Squeeze duration, s	28 (25–31)	32 (26–38)	24 (20–28)
% increase in anal sphincter pressure during squeeze	126 (89–163)	158 (114–202)	103 (70–136)
Rectal pressure when squeezing, mmHg	19 (14–23)	24 (15–33)	16 (11–21)
Anal pressure during party balloon inflation, mmHg	127 (113–141)	154 (138–170)	106 (89–123)
Rectal pressure during party balloon inflation, mmHg	63 (54–72)	66 (51–81)	62 (51–73)

[a]Mean 95% cl.
[b]From Ref. 11 with permission.

Dyssynergia:	When an act is not performed smoothly or accurately because of lack of harmonious association of its various components; when there is lack of coordination or dyssynergia of the abdominal and pelvic floor muscles that are involved in defecation it is called dyssynergic defecation
ENS:	Abbreviation for enteric nervous system.
Endoanal Cushion:	Within the anus. Anal mucosal folds together with anal vascular cushion.
High pressure zone:	Intense compression area.
Intrinsic Plexus:	A network or inter-joining of nerves and blood vessels or of lymphatic vessels belonging entirely to a part.
Myenteric Plexus:	A plexus of unmyelinated fibers and postganglionic autonomic cell bodies lying in the muscular coat of the esophagus, stomach, and intestines; it communicates with the subserous and submucous plexuses, all subdivisions of the enteric plexus.
Orad:	In a direction toward the mouth.
Phasic:	In stages, in reference to rectal balloon distension for sensory testing.
Proctalgia:	Pain in the anus, or in the rectum.
Proximal:	Nearest the trunk or the point of origin, in contrast to distal.
Supraconal:	Above a condyle.
Tone:	Normal tension or resistance to stretch.

BIBLIOGRAPHY

Cited References

1. Strohbehn K. Normal pelvic floor anatomy. Obstet Gynecol Clin N Am 1998;25:683–705.
2. Whitehead WE, Schuster MM. Anorectal physiology and pathophysiology. Am J Gastroenterol 1987;82:487–497.
3. Matzel KE, Schmidt RA, Tanagho EA. Neuroanatomy of the striated muscular anal continence mechanism. Implications for the use of neurostimulation. Dis Colon Rectum 1990;33: 666–673.
4. Fernandez-Fraga X, Azpiroz F, Malagelada JR. Significance of pelvic floor muscles in anal incontinence. Gastroenterology 2002;123:1441–1450.
5. Gunterberg B, Kewenter J, Petersen I, Stener B. Anorectal function after major resections of the sacrum with bilateral or unilateral sacrifice of sacral nerves. Br J Surg 1976;63: 546–554.
6. Sun WM, Rao SS. Manometric assessment of anorectal function. Gastroenterol Clin N Am 2001;30:15–32.
7. Sun WM, Read NW. Anorectal function in normal human subjects: effect of gender. Int J Colorectal Disease 1989;4: 188–196.
8. Rao SSC. Book Chapter—Colon Transit and Anorectal Manometry. In: Rao SSC, editors. Gastrointestinal Motility: Tests and Problem-Orientated Approach. New York: Kluwer Academic/Plenum Publishers; 1999. pp 71–82.
9. Diamant NE, Kamm MA, Wald A, Whitehead WE. AGA technical review on anorectal testing techniques. Gastroenterology 1999;116:735–760.
10. McHugh SM, Diamant NE. Effect of age, gender, and parity on anal canal pressures. Contribution of impaired anal sphincter function to fecal incontinence. Dig Dis Sci 1987; 32:726–736.
11. Rao SS. Manometric tests of anorectal function in healthy adults. Am J Gastroenterol 1999;94:773–783.
12. Taylor BM, Beart RW, Jr., Phillips SF. Longitudinal and radial variations of pressure in the human anal sphincter. Gastroenterology 1984;86:693–697.
13. Rao SS. Minimum standards of anorectal manometry. Neurogastroenterol Motil 2002;14:553–559.
14. McHugh SM, Diamant NE. Anal canal pressure profile: a reappraisal as determined by rapid pullthrough technique. Gut 1987;28:1234–1241.
15. Pedersen IK, Christiansen J. A study of the physiological variation in anal manometry. Br J Surg 1989;76: 69–70.
16. Azpiroz F, Enck P, Whitehead WE. Anorectal functional testing: review of collective experience. Am J Gastroenterol 2002;97:232–240.
17. Rao SS. Dyssynergic defecation. Gastroenterol Clin N Am 2001;30:97–114.
18. MacDonagh R, et al. Anorectal function in patients with complete supraconal spinal cord lesions. Gut 1992;33:1532–1538.
19. Wald A. Colonic and anorectal motility testing in clinical practice. Am J Gastroenterol 1994;89:2109–2115.
20. Sun WM, et al. Sensory and motor responses to rectal distention vary according to rate and pattern of balloon inflation. Gastroenterology 1990;99:1008–1015.
21. Whitehead WE, Delvaux M. Standardization of barostat procedures for testing smooth muscle tone and sensory thresholds in the gastrointestinal tract. The Working Team of Glaxo-Wellcome Research, UK. Dig Dis Sci 1997;42: 223–241.
22. Mertz H, et al. Altered rectal perception is a biological marker of patients with irritable bowel syndrome. Gastroenterology 1995;109:40–52.
23. Sun WM, Read NW, Miner PB. Relation between rectal sensation and anal function in normal subjects and patients with faecal incontinence. Gut 1990;31:1056–1061.
24. Rao SS, et al. Anorectal sensitivity and responses to rectal distention in patients with ulcerative colitis. Gastroenterology 1987;93:1270–1275.
25. Salvioli B, et al. Rectal compliance, capacity, and rectoanal sensation in fecal incontinence. Am J Gastroenterol 2001;96: 2158–2168.
26. Tuteja AK, Rao SS. Review article: Recent trends in diagnosis and treatment of faecal incontinence. Aliment Pharmacol Ther 2004;19:829–840.

27. Engel AF, Kamm MA, Bartram CI, Nicholls RJ. Relationship of symptoms in faecal incontinence to specific sphincter abnormalities. Int J Colorectal Disease 1995;10:152–155.
28. Felt-Bersma RJ, Klinkenberg-Knol EC, Meuwissen SG. Anorectal function investigations in incontinent and continent patients. Differences and discriminatory value. Dis Colon Rectum 1990;33:479–485; discussion 485–486.
29. Sun WM, Donnelly TC, Read NW. Utility of a combined test of anorectal manometry, electromyography, and sensation in determining the mechanism of 'idiopathic' faecal incontinence. Gut 1992;33:807–813.
30. Sun WM, et al. Anorectal function in patients with complete spinal transection before and after sacral posterior rhizotomy. Gastroenterology 1995;108:990–998.
31. Caruana BJ, Wald A, Hinds JP, Eidelman BH. Anorectal sensory and motor function in neurogenic fecal incontinence. Comparison between multiple sclerosis and diabetes mellitus. Gastroenterology 1991;100:465–470.
32. Tjandra JJ, et al. Anorectal physiological testing in defecatory disorders: a prospective study. Aust N Z J Surg 1994;64: 322–326.
33. Rao SS, Welcher KD, Leistikow JS. Obstructive defecation: a failure of rectoanal coordination. Am J Gastroenterol 1998;93: 1042–1050.
34. Read NW, et al. Anorectal function in elderly patients with fecal impaction. Gastroenterology 1985;89:959–966.
35. Kumar D, et al. Prolonged anorectal manometry and external anal sphincter electromyography in ambulant human subjects. Dig Dis Sci 1990;35:641–648.
36. Rao SS, Patel RS. How useful are manometric tests of anorectal function in the management of defecation disorders? Am J Gastroenterol 1997;92:469–475.
37. Vaizey CJ, Kamm MA. Prospective assessment of the clinical value of anorectal investigations. Digestion 2000;61:207–214.

See also BIOFEEDBACK; ESOPHAGEAL MANOMETRY; GASTROINTESTINAL HEMORRHAGE.

ANTIBODIES, MONOCLONAL. See MONOCLONAL ANTIBODIES.

APNEA DETECTION. See VENTILATORY MONITORING.

ARRHYTHMIA, TREATMENT. See DEFIBRILLATORS; PACEMAKERS.

ARRHYTHMIA ANALYSIS, AUTOMATED

STEPHANIE A. C. SCHUCKERS
Clarkson University
Potsdam, New York

INTRODUCTION

Sudden cardiac death is estimated to affect $\sim 400{,}000$ people annually (1). Most of these cases are precipitated by ventricular fibrillation (VF), a chaotic abnormal electrical activation of the heart. Ventricular fibrillation disturbs systemic blood circulation and causes immediate death if therapy in the form of an electrical shock is not immediately applied. In fact, survival depends dramatically on the time it takes for therapy to arrive (2). Automated arrhythmia detection is a key component for speeding up defibrillation therapy through medical devices that detect arrhythmia and provide treatment automatically without human oversight. Examples of devices include implantable defibrillators, public access automated external defibrillators, and more.

Arrhythmias generally are an abnormal electrical activation of the heart. These abnormalities can occur in the atrial chambers, ventricular chambers, or both. Since the ventricles are the chambers responsible for providing blood to the body and lungs, disruptions in the electrical system that stimulates the mechanical contraction of the heart can be life threatening. Examples of ventricular arrhythmias include VF and ventricular tachycardia (VT), as seen in Fig. 1. Atrial arrhythmias including atrial fibrillation (AF), atrial flutter (AFl), and supraventricular tachycardia (SVT) are not immediately life threatening, but can cause uncomfortable symptoms and complications over the long term.

Automated arrhythmia analysis is the detection of arrhythmias through the use of a computer. This article focuses on arrhythmia detection performed *without human oversight*. The primary focus will be algorithms developed for the implantable cardioverter defibrillator (ICD). An implantable cardioverter defibrillator is a device that provides an electrical shock to ventricular fibrillation and tachycardia to terminate it and restart NSR. The implantable cardioverter defibrillator was developed in the late 1970s and FDA-approved in the mid-1980s (4–7). A

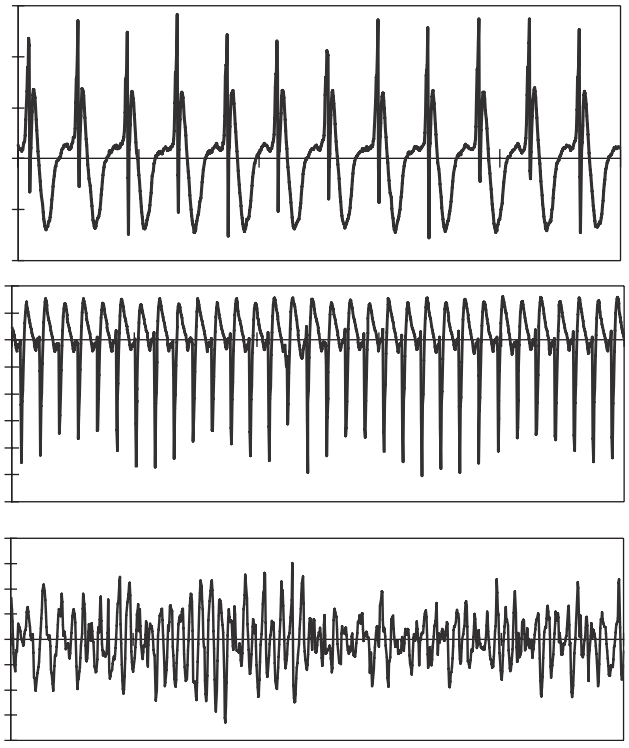


Figure 1. Unipolar electrograms (measurements of the electrical activity from inside the heart) for normal sinus rhythm (NSR), ventricular tachycardia (VT), and VF [10 s of the passage are shown (AAEL234) (3)].

Table 1. Truth Table Used to Determine Sensitivity and Specificity, Measurements of Automated Algorithm Performance

Device/Truth->	VT/VF	All Others
VT/VF	True positive	False positive
All others	False negative	True negative

catheter placed in the right ventricle is used for both sensing and therapy. This device has a long history of arrhythmia detection algorithms developed in research laboratories and brought to the marketplace. Other medical devices that use purely automated arrhythmia detection include the automatic external defibrillator and the implantable atrial defibrillator. Semiautomated arrhythmia detection is used in ambulatory and bedside monitoring. These topics will be touched on briefly.

It is important to consider the measurements used to assess the performance of automated arrhythmia analysis. Sensitivity is defined as the percent correct detection of disease, while specificity is the percent correct detection of not disease. Take the case of an implantable defibrillator that detects ventricular tachycardia and ventricular fibrillation. Consider the truth table in Table 1.

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

A false positive is one minus the specificity, while a false negative is one minus the sensitivity.

Early Work

The earliest examples of computer-based arrhythmia analysis are semiautomated approaches for bedside and ambulatory monitoring. Ambulatory monitoring typically uses a 24 or 48 h, three-lead, portable electrocardiogram (ECG) recorder that the patient wears to diagnosis arrhythmias. Arrhythmia analysis is done in an off-line fashion with technician oversight, such that it is not purely automated (8). Other ambulatory monitors include loop recorders or implanted monitors like Medtronic Reveal Insertable Loop Recorder that permanently records with patient interaction. Clinical bedside monitors typically are also not fully automated, but are used as initial alarms, which is then over read by clinical staff.

In ambulatory monitoring, in addition to detection of arrhythmias, it is typical to also detect premature ventricular contractions (PVCs). These PVCs are beats that form ectopically in the ventricle and result in an early, wide ECG beat and occur alone or in small groups of two or more. They are considered a potential sign of susceptibility to arrhythmias.

Use of correlation is a common tool in surface arrhythmia analysis (9–18). Correlation waveform analysis (CWA) uses the correlation coefficient between a previously stored template of sinus rhythm and the unknown cycle under analysis. The correlation coefficient, used by CWA, is computed as

$$\rho = \frac{\sum_{i=1}^{i=N}(t_i - \bar{t})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{i=N}(t_i - \bar{t})^2 \sum_{i=1}^{i=N}(s_i - \bar{s})^2}}$$

where ρ = the correlation coefficient, N = the number of template points, t_i = the template points, s_i = the signal points under analysis, $\bar{t}$ = the average of the template points, and $\bar{s}$ = the average of the signal points. The correlation coefficient falls within a range $-1 < \rho < +1$, where $+1$ indicates a perfectly matched signal and template.

To compute CWA, a beat detector (described in more detail in the section implantable cardioverter defibrillators) finds the location of each beat. From the location of each beat, the template is aligned with the beat under analysis, typically using the peak, and the correlation coefficient is calculated. Often, the template is shifted and the procedure is repeated to determine the best alignment indicated by the highest correlation coefficient. An example of CWA is shown in Fig. 2. Sustained high correlation indicates normal sinus rhythm and low indicates an arrhythmia.

Examples of other features used in PVC and arrhythmia detection include timing, width, height, area, offset, first spectral moment (5–25 Hz), T-wave slope, and others (9,13,14,17,20–24). Another early approach was to develop a database of electrocardiogram templates grouping similar shaped complexes based on shape, width, and prematurity (17,25–28).

Some of the earliest algorithms for *purely automated* arrhythmia detection involved algorithms for newly developing implantable devices for SVT termination and the developing implantable defibrillator (29,30). With problems of inducing ventricular arrhythmias in the devices for SVT termination, focus shifted to the implantable

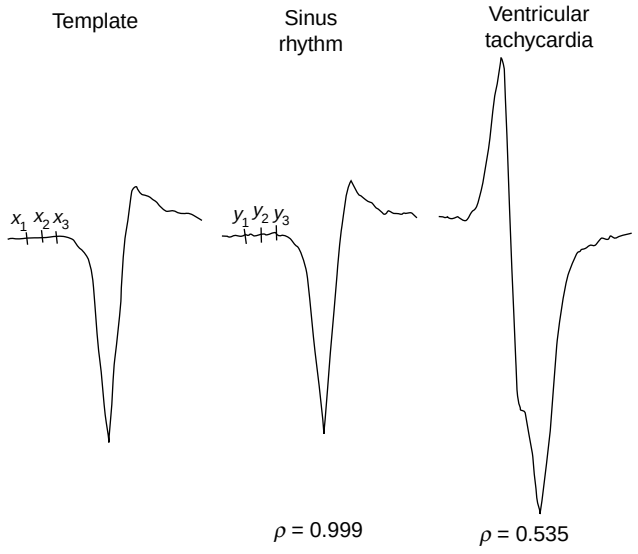


Figure 2. Example of use of correlation waveform analysis. The first electrogram is the stored normal sinus rhythm template, the second is a NSR beat with correlation equal to 0.999 and the third is a ventricular tachycardia with correlation equal to 0.535. (Used with permission from Ref. 19.)

defibrillator (31,32). In the early 1980s, Furman (29) proposed that two sensors (atrial and ventricular) be required for automatic diagnosis of tachycardia (with even a possible refinement of a third sensor for hestidine (His) bundle detection). He also suggested examining the QRS configuration for a match with sinus rhythm as a schema for diagnosing supraventricular tachycardia. Other early work included the development of algorithms for surface electrocardiography that used an esophogeal electrode for analysis of atrial information (33,34).

The original detection mechanism in the first implantable defibrillator, the AICD, was the probability density function (PDF). This algorithm utilized the derivative of the signal to define the duration of time that the signal departed from baseline (31,32) and was empirically based upon the observation that the ventricular fibrillation signal spends the majority of its time away from the electrocardiographic isoelectric baseline when compared to sinus rhythm or supraventricular rhythms (Fig. 1). The PDF was supplanted at a very early stage by intrinsic heart rate measures.

The need to identify and cardiovert ventricular tachycardia in addition to detecting and defibrillating ventricular fibrillation, and the recognition that sufficiently slow VT might have rates similar to those that may occur during sinus rhythm or supraventricular tachycardias resulted in several changes being incorporated into the second generation of devices. An alternative time-domain method called temporal electrogram analysis was incorporated into some second-generation devices (35). This algorithm employed positive and negative thresholds, or rails, placed upon electrograms sensed during sinus rhythm. A change in electrogram morphology was identified when the order of the excursion of future electrograms crossed the predetermined thresholds established during sinus rhythm. The combination of this morphologic method with ventricular rate was intended to differentiate ventricular tachycardia from other supraventricular tachycardias including sinus tachycardia.

Experience with probability density function and temporal electrogram analysis in first- and second-generation devices was disappointing. Probability density function was found to be unable to differentiate sinus tachycardia, supraventricular tachycardia, ventricular tachycardia, and ventricular fibrillation whose respective rates exceeded programmed device thresholds for tachycardia identification (36). A similar experience was encountered with temporal electrogram analysis. As a result, these criteria were utilized less and less frequently as increasing numbers of second-generation devices were implanted. By 1992, $< 15\%$ of all ICDs implanted worldwide utilized either algorithm for tachycardia discrimination (37).

IMPLANTABLE CARDIOVERTER DEFIBRILLATORS

Over time, the implantable cardioverter defibrillator added capabilities to pace terminate and cardiovert ventricular tachycardia and provide pacemaker functions, single and dual chamber. Early reviews of automated algorithms particularly for implantable defibrillators are given in Refs. 38–40. More recent reviews of automated arrhythmia detection algorithms include a thorough review by Jenkins and the author in 1996 (41) and reviews incorporating recent developments in dual chamber algorithms in 1998 and 2004 (42,43).

Rate-Based Analysis

The main method for detection of arrhythmias after initial use of PDA and TEA was the use of intrinsic heart rate for detection of ventricular tachycardia and ventricular fibrillation. To this day, all algorithms in ICDs have rate as a fundamental component for detection of arrhythmias. Since implantable defibrillators have a stable catheter screwed in the apex of right ventricle, the rate of the ventricles can be determined with little of the noise that is present at the surface of the body, like motion artifact and electromyogram noise. Ventricular tachycardia and ventricular fibrillation have rates of $\sim$ 120–240 beats per minute and $>$ 240 beats per minute, respectively.

Many approaches abound for arrhythmia detection using rate, but the general procedure is the same (Fig. 3). First, each ECG beat must be detected. Second, the time between beats (or the cycle length) is determined. Most algorithms rely on cycle length (CL) values over beats per minute. The value of the CL determines the zone it falls into: Normal, VT, or VF. In some cases, zones may be further divided depending on the device and therapy options. The thresholds that define the zones are programmable. Each zone has a programmable counter that will determine when therapy will need to be considered. In addition, each zone has a reset mechanism that may be different depending on the zone. For example, typical VT zones require X consecutive beats within the zone or the counter is reset. While VF often has an X of Y criteria, for example, 12 of 16 beats. This flexibility is due to the fact that VF is of varying amplitude, morphology, and rate, such that each beat may not be detected reliably and/or may not be in the VF zone. The CL thresholds, counters, and associated therapies are all programmable.

The fundamental basis of automated rate algorithms is the detection of each beat. Many approaches have been suggested and utilized including fixed thresholds, exponentially varying thresholds, amplitude gain control, and

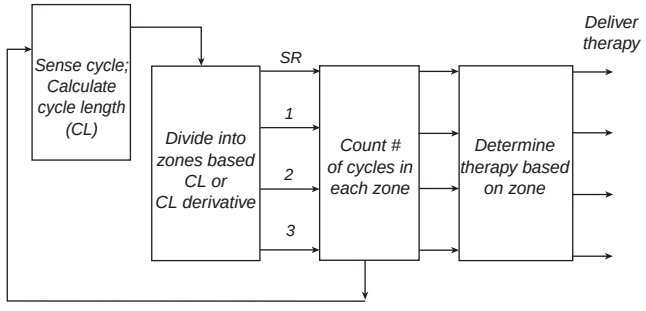


Figure 3. Typical rate-based arrhythmia detection scheme for implantable cardioverter defibrillators. First, each beat is detected and the cycle length between beats determined. A counter is incremented in the zone that the CL falls and therapy is delivered when the counter reaches a programmed threshold.

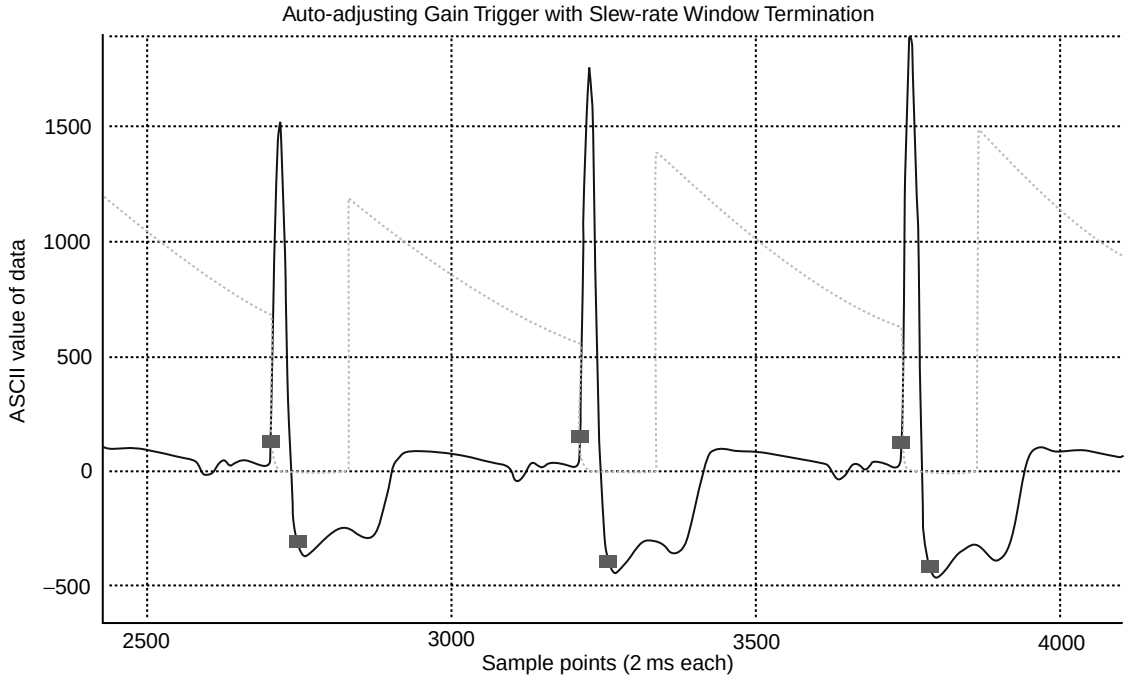


Figure 4. Example of beat detector that utilizes an exponentially decaying threshold. After a beat is detected, a blanking period prevents detection of the same beat twice. Then, the threshold (dotted line) for determination of the next beat is calculated as a percentage of the peak amplitude of the previous beat. This threshold exponentially decays, such that beats that are smaller than the currently detected beat will be not be missed (47).

others (9,44–46). In implantable devices, most are hardware-based and beyond the scope of this article. For example, one software method relies on an exponentially varying threshold (Fig. 4) (47). After a beat is detected, there is first a blanking period to prevent a beat from being detected more than once. After the blanking period, the threshold for detection of the next beat is set as a percentage of the previous beat. This threshold then decays exponentially such that subsequent beats that have a smaller peak amplitude will be detected. Most beat detectors also have a floor for the threshold, that is, the smallest amplitude by which a beat can be detected to prevent detection of noise as a beat.

An example of one rate-based algorithm is given in Fig. 5 (40). This algorithm uses three CL thresholds, fibrillation detection interval (FDI), fast tachycardia interval (FTI), and tachycardia detection interval (TDI), and two counters, VF counter (VFCNT) and VT counter (VTCNT). These are combined to result in three zones, VF, fast VT, and slow VT, which can have different therapeutic settings, utilizing defibrillation shock, cardioversion, and antitachycardia pacing. Therapy for ventricular fibrillation is given when 18 of 24 beats are shorter than the FDI. Therapy for slow ventricular tachycardia is delivered when 16 beats counted by the VTCNT are between the FDI and TDI thresholds. The VTCNT will be reset by one long CL greater than TDI. The fast VT zone is a combination of these techniques.

A thorough description of the rate-based algorithms is given in Ref. 40. While many additional features have been added to refine the decision, the main structure of automated arrhythmia detection algorithms still rely on this fundamental approach (42).

As can be see from Fig. 1, heart rate in VT and VF increase substantially over normal sinus rhythm. This is a reliable means of detecting VT and VF for implantable devices, resulting in high sensitivity. Unfortunately, while providing high sensitivity, heart rate also increases for normal reasons, exercise, stress, resulting in sinus tachycardia or for nonventricular-based arrhythmias like atrial fibrillation, supraventricular tachycardia, atrial flutter, and so on, which do not require therapy from the ICD. Thus, rate-based algorithms have low specificity. False therapies have been estimated in as much as 10–40% in the early devices (48–50). Morphology and other extended algorithmic approaches have long been suggested as a means to increase specificity.

Early rate-based algorithms to prevent false therapies, due to sinus tachycardia, atrial fibrillation, and so on, include onset and rate stability. Rate-based methods were chosen initially over morphology due to the simplicity of calculations in battery operated devices.

Onset is the difference between the rate changes during the onset of sinus tachycardia compared to those of VT, since the onset of VT is typically sudden compared to sinus tachycardia. False therapies due to sinus tachycardia are determined by onset. Figure 6a shows the sudden onset of ventricular tachycardia.

Rate stability is used to prevent false therapies due to AF. In AF, it is common for the ventricle to respond to the atrium at a fast rate. This response is typically irregular since atrial fibrillation, by definition has an irregular rate,

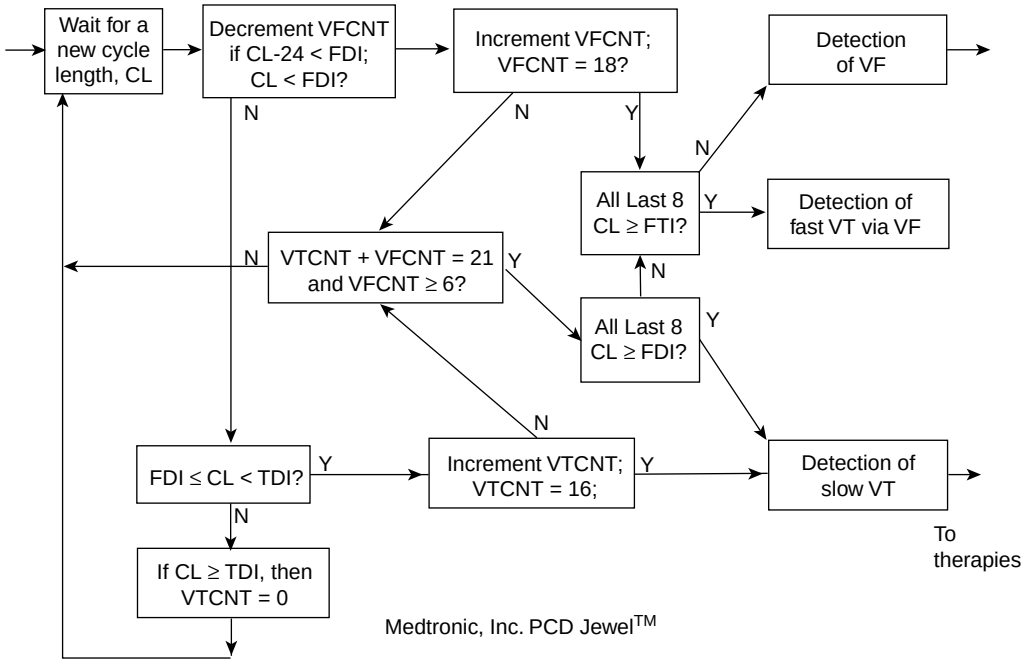


Figure 5. Example of rate-based algorithm for the Medtronic PCD Jewel. The VFCNT is the VF counter, FDI is the fibrillation detection interval, FTI is the fast tachycardia interval, VTCNT is the VT counter, and TDI is the tachycardia detection interval. This algorithm has three zones that has associated programmable therapies including defibrillation shock, cardioversion, and antitachycardia pacing (40).

and since not every beat is conducted from the atrium to the ventricle. Rate stability considers the stability of the ventricular rate, since VT typically has a stable rate compared to the ventricular response to atrial fibrillation. Figure 6b shows an irregular ventricular response to atrial flutter.

Rate and rate-derived measures that measure onset and stability (based on cycle-by-cycle interval measurements) include average or median cycle length, rapid deviation in cycle length (onset), minimal deviation of cycle length (stability), and relative timing measures in one or both chambers or from multiple electrodes within one or more chambers. Among the methods most widely used for detection of VT in commercially available single chamber antitachycardia devices have been combinations of rate, rate stability, and sudden onset (51–56). Pless and Sweeney published an algorithm for (1) sudden onset, (2) rate stability, and (3) sustained high rate (57). This schema among others (58,59) was a forerunner of many of the methods introduced into tachycardia detection by ICDs (60).

Morphological Pattern Recognition

Instead of relying purely on rate, it has been suggested that morphology may provide the means for automated arrhythmia detection to separate VT and VF from rhythms with fast rates that do not need therapy. Morphology in this context refers to characteristics of the electrogram waveform itself, which are easily identifiable and measurable. Such features might include peak-to-peak amplitude, slew rate (a measure of waveform), sequence of slope patterns, sequence of amplitude threshold crossings, statistical pattern recognition of total waveform shape by correlation coefficient measures, and others (61,62). Figure 1 shows an example of distinctly different waveforms recorded from the right ventricular apex during SR, VT, and VF (3). Furthermore, morphology in the ventricle appears normal

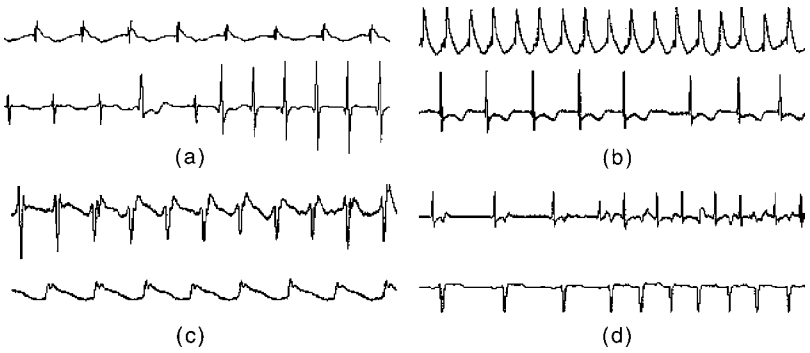


Figure 6. Atrial (top) and ventricular (bottom) electrograms. (a) Sudden onset of VT with normal atrial electrogram, (b) irregular response of ventricular to atrial flutter, (c) simultaneous atrial flutter and ventricular tachycardia, and (d) sudden onset of supraventricular tachycardia with ventricular response.

even during supraventricular arrhythmias since the rhythm typically is conducted normally in the ventricles.

The ICDs often have two channels of electrograms. The first channel is typically bipolar, which is associated with two electrodes on the lead, or an electrode–coil combination, both located within the ventricle. This channel provides a near-field electrical view of the ventricle and is usually used for beat detection because the electrogram typically has a narrow QRS (the main depolarization of the electrogram). The second electrode configuration is far-field which, for example, may use an electrode in the ventricle versus the implantable device casing. The far-field electrode combination is used primarily for giving the electrical shock. However, this far-field view typically has a more global perspective of the electrogram depolarization and is helpful in differentiating the morphology changes between normal beats and ventricular abnormalities.

Template-Based Algorithms

Correlation Waveform Analysis. Lin et al. (19,62,63) investigated three techniques for morphologic analysis of VT: correlation waveform analysis, amplitude distribution analysis, and spectral analysis. Correlation waveform analysis (CWA) is a classic method of pattern recognition applied to the surface electrocardiogram, described earlier, but was first applied to intracardiac signals in this study. Correlation waveform analysis was shown to be superior and has the advantage of being independent of amplitude and baseline fluctuations. However, it requires heavy computational demands. Less computationally demanding algorithms based on the same principle have been developed and are described in the next section.

Less Computationally Demanding Template-Based Algorithms. Another template matching algorithm based on raw signal analysis measured the area of difference between electrograms, that is, adding absolute values of the algebraic differences between each point on the electrogram and corresponding point on the SR template (64,65). The area of difference was expressed as a percentage of the total area of the template. The measurement of an area of difference is simple computationally, but has the disadvantage of producing erroneous results in the face of baseline and amplitude fluctuations, and this method fails to produce a bounded measure. An improvement on this technique by signal normalization and scaling to create a metric bounded by ± 1 was utilized by Throne et al. (66).

Steinhaus et al. (67) modified correlation analysis of electrograms to address computational demand by applying data compression to filtered data (1–11 Hz) by retaining only samples with maximum excursion from the last saved sample. The average squared correlation coefficient (ρ^2) was used for separation of SR and VT. Comparison with noncompressed correlations demonstrated that data compression had negligible effects on the results.

Throne et al. (66) designed four fast algorithms and compared discrimination results to CWA performance. These morphological methods were the bin area method (BAM); derivative area method (DAM); accumulated difference of slopes (ADIOS); and normalized area of difference (NAD). All four techniques are independent of amplitude fluctuations and three of the four are independent of baseline changes.

The bin area method is a template matching algorithm that compares corresponding area segments or bins of the template with the signal to be analyzed. Each bin (average of three consecutive points) is adjusted for baseline fluctuations by subtracting the average of the bins over one cycle and normalized to eliminate amplitude variations. This BAM equation is given in the following equation:

$$\rho = 1 - \sum_{i=1}^{i=M} \left| \frac{T_i - \overline{T}}{\sum_{k=1}^{k=M} \left| T_k - \overline{T} \right|} - \frac{S_i - \overline{S}}{\sum_{k=1}^{k=M} \left| S_k - \overline{S} \right|} \right|$$

where the bins are $S_1 = s_1 + s_2 + s_3$, $S_2 = s_4 + s_5 + s_6, \cdots$, $S_M = s_{N-2} + s_{N-1} + s_N$ and the average of M bins is $\overline{S} = (1/M)\Sigma_{k=1}^{k=M} S_k$. The bins and average of the bins is calculated similarly for the template. The BAM metric falls between -1 and $+1$, allowing a comparison to CWA.

Normalized area of difference is identical to BAM except that the average bin value is not removed. By not removing the average value the algorithm avoids one division that would otherwise increase computational demand each time the BAM algorithm is applied. The NAD is independent of amplitude changes.

The DAM uses the first derivative of the template and the signal under analysis. The method creates segments from zero crossings of the derivative of the template. It imposes the same segmentation for analysis of the derivative of the signal to be compared. The segments are normalized, but are not adjusted for baseline variations since derivatives are by their nature baseline independent. The DAM metric is calculated as follows:

$$\rho = 1 - \sum_{i=1}^{i=M} \left| \frac{\dot{T}_i}{\sum_{k=1}^{k=M} \left| \dot{T}_k \right|} - \frac{\dot{S}_i}{\sum_{k=1}^{k=M} \left| \dot{S}_k \right|} \right|$$

where $\dot{T}_k$ represents the kth bin of the first derivative of the template. The DAM metric falls between -1 and $+1$.

The ADIOS is similar to DAM in that it also employs the first derivative of the waveforms. A template is constructed of the sign of the derivative of the ventricular depolarization template. This template of signs is then compared to the signs of the derivative for subsequent depolarizations. The total number of sign differences between the template and the current ventricular depolarization is then computed as

$$\rho = \sum_{i=1}^{i=N} \text{sign}(\dot{t}_i) \oplus \text{sign}(\dot{s}_i)$$

where $\oplus$ is the exclusive or operator. The number of sign changes is bounded by 0 and the maximum number of points in the template (N), that is, $\rho \in \{0, \ldots, N\}$.

Evaluation of these four algorithms was performed on 19 patients with 31 distinct ventricular tachycardia morphologies. Three of the algorithms (BAM, DAM, and

NAD) performed as well or better than correlation waveform analysis, but with one-half to one-tenth the computational demands.

A morphological scheme for analysis of ventricular electrograms (SIG) was devised for minimal computation (68) and compared to NAD. The SIG is a template-based method that creates a boundary window enclosing all template points that form a signature of the waveform to be compared. Equivalent results of VT separation were seen in the two techniques at two thresholds, but at an increased safety margin of separation SIG outperformed NAD and yielded a fourfold reduction in computation.

Another simplified correlation-type algorithm has been designed using electrogram vector timing and correlation, developed for the Guidant ICD (69). In this algorithm, the rate (near-field) channel is used for determining the location of each beat. The peak of the near-field electrogram or fiducial point is used for alignment of the template with the beat under analysis. From this fiducial point, eight specific points are chosen on the shock (far-field) electrogram. The amplitude of the shock channel at the rate-channel fiducial point is one point. In addition, amplitudes at the turning point, intermediate, and baseline values on the shock channel are selected as shown in Fig. 7. This provides an eight-point template that is compared to subsequent beats using the square of the correlation coefficient, as follows:

$$\mathrm{FCC} = \frac{(8\sum t_i s_i - (\sum t_i)(\sum s_i))^2}{(8\sum t_i^2 - (\sum t_i)^2)(8\sum s_i^2 - (\sum s_i)^2)}$$

where each summation is $i = 1$–8.

When an unknown beat is analyzed, the exact same timing relative to the fiducial point as the template is used for selecting the amplitudes of the unknown beat. For beats that have a different morphology, those points will not be associated with the same amplitudes as the normal template beat and, thus, the correlation coefficient will be low. To incorporate this into an overall scheme to detect an arrhythmia, morphology was calculated for a sliding window of 10 beats. If 8 or more beats were detected as abnormal, a VT was detected.

Another algorithm that reduces computational complexity of the standard correlation algorithm uses the wavelet transform of the sinus beat for the template (70). Wavelets can reduce the number of coefficients needed to characterize a beat while still retaining the important morphologic information. The sinus electrogram is transformed using the Haar (square) wavelet, considering a family of 48 wavelets over 187.5 ms window aligned by the fiducial point of the QRS. The wavelet transform is simplified by removing the standard factor of square root of 2. In addition, wavelet coefficients that do not carry much information, defined by a threshold, are set to zero. The remaining coefficients are normalized. This gives a variable template size, depending on the electrogram, but typically between 8 and 20 coefficients. To analyze an unknown electrogram, the electrogram is aligned using the peak (negative or positive) point. The wavelet transform is computed for the unknown electrogram and each coefficient is compared using the absolute difference in wavelet coefficients (c_i) between the template and unknown beat. A match is determined by the following equation:

$$\mathrm{Match}\% = \left|1 - \frac{\sum |c_i^{\mathrm{template}} - c_i^{\mathrm{unknown}}|}{\sum |c_i^{\mathrm{template}}|}\right| * 100$$

The nominal threshold used in this study is 70%. This morphology algorithm is incorporated into an overall rate scheme by remaining inactive until a ventricular tachycardia has been detected by the rate algorithm. Then, the morphology is calculated for the preceding eight beats. A VT is detected if six or more beats are detected as abnormal.

A novel way of testing this algorithm was used. Instead of, as in most tests, using data prerecorded in laboratory conditions, this algorithm was downloaded to the Medtronic clinical ICDs and tested off-line, while the device functioned with its regular algorithm.

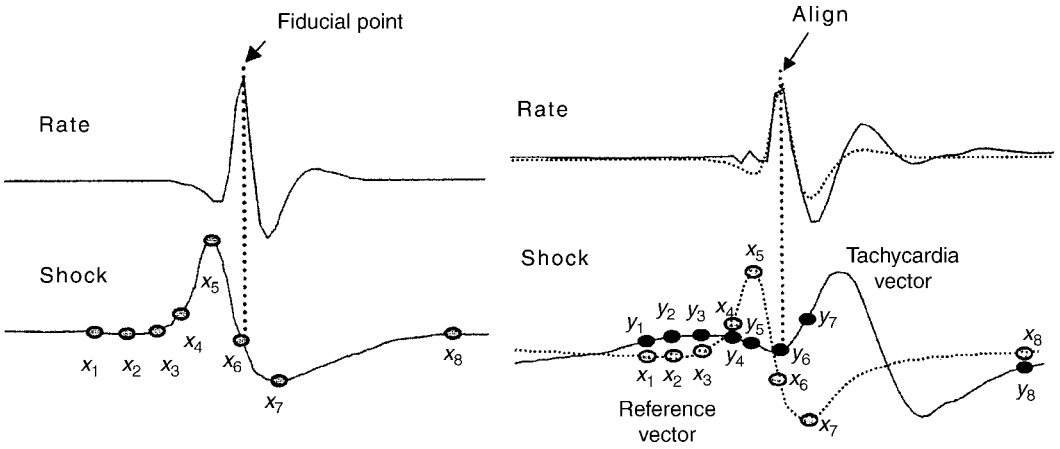


Figure 7. Example for vector timing and correlation algorithm. Alignment of the template is based on the peak of the rate electrogram channel. From the peak, eight specific points on the shock channel are automatically selected for the template (left). These exact points in time relative to the fiducial point selected from the template are applied to the "unknown" beat (right). (Used with permission from Ref. 69.)

Another algorithm that utilizes morphology is termed morphology discrimination (MD) in St. Jude implantable cardioverter defibrillators (71,72). This algorithm uses an area-based approach. First, the template is defined based on the three consecutive peaks with the largest area. The area under each peak is normalized by the maximum peak area. When analyzing an unknown beat, the beat is aligned with the template using the dominant peak of the unknown beat. If this peak does not have same polarity, the second largest peak is used. If this also does not have the same polarity, a nonmatch is declared. Once the unknown beat and template are aligned, the morphology score is determined by the following:

$$\text{Score} = (1 - ||N\text{Area}A - N\text{Area}A'| + |N\text{Area}B - N\text{Area}B'| + |N\text{Area}C - N\text{Area Peak}C'||) * 100$$

where NArea stands for the normalized area of the three corresponding peaks of the template (A, B, C) and test complexes (A', B', C'). For arrhythmia diagnosis, once the rate criteria is met, the algorithm determines the number of matching complexes in the morphology window. If the number of matching complexes equals or exceeds a programmed number of matching complexes, VT is not confirmed and therapy is not delivered. This is repeated for as long as the rate criteria has been met or VT is confirmed.

Template matching by CWA was further examined for distinction of multiple VTs of unique morphologies in the same patient (73,74). It was hypothesized that, in addition to a SR template, a second template acquired from the clinical VT could provide confirmation of a later recurrence of the same VT. The recognition of two or more different VTs within the same patient could play an important role in future devices in the selection of therapy to be delivered to hemodynamically stable versus unstable VTs.

Considerations for Template Analysis. While template-based algorithms appear the most promising, several issues need to be addressed. The first is that it is necessary that the normal sinus rhythm beat or template remain stable, that is, does not change over time or due to position or activity. Several studies using temporary electrodes saw changes in the morphology of normal rhythm due to increase in rate or positional changes (75–78) but further studies with fixed electrodes showed no changes in morphology due to heart rate or position, with some changes in amplitude (76,79). A second consideration is that paroxysmal (sudden) bundle branch block (BBB) may be misdiagnosed as ventricular tachycardia (80). While this may result in a false therapy, it does not result in withholding of therapy during life-threatening arrhythmias (the more critical mistake).

Feature-Based Algorithms

Depolarization Width for Detection of Ventricular Tachycardia. Depolarization width (i.e., duration) in ventricular electrograms has been used as a discriminant of supraventricular rhythm (SR) from VT (81,82). Electrogram width is available in the Medtronic single chamber ICDs. This criterion uses a slew threshold to find the beginning and end of the QRS. Analysis of electrogram width compared to a patient-specific width threshold is performed using the previous eight beats after a VT detected by the rate component of the algorithm. If a minimum of six complexes are greater than the width, then a VT is detected. Otherwise, the counter is reset. This algorithm is not appropriate in patients with BBB that have a wider width for normal beats. Exercise induced variation should be considered in programming (83). Electrogram width has been shown to be sensitive to body position and changes over longer periods of time (6 months in this study) (84).

Amplitude and Frequency Analysis. Amplitude and frequency are distinguishing characteristics of arrhythmia. Amplitude during ventricular tachycardia is typically higher and during ventricular fibrillation is lower than normal sinus rhythm (85,86). These differences have not been considered pronounced and consistent enough, such that a classifier could be based on them.

Frequency-domain analysis is often proposed for classification of rhythms (87) but little success has been solidly demonstrated for the recognition of VT (63). Distinctly different morphological waveforms (SR vs VT), which are easily classified in the time domain, can exhibit similar or identical frequency components if one focuses on the depolarization component alone. Examination of longer segments of 1000–15,000 ms yields the same phenomenon because the power present in small visually distinctive high frequency notches is insignificant compared to the remainder of the signal, and changes in polarity of the waveform, easily recognized in the time domain, are simply not revealed by frequency analysis (63). Frequency-domain recognition of AF (88) and VF (89,90) is perhaps more promising. However, frequency has not been applied in commercial applications given the success of rate and time-domain morphology approaches.

Other Morphologic Approaches. Other approaches that have been suggested in the literature include use of neural networks (91–97). Neural network approaches utilize either features, the time-series, or frequency components as inputs to the neural network. The network is trained on one dataset and tested on a second. Limitations with the approaches developed thus far are related to the fact that there is only limited data for development of the neural network. One problem is that in some studies the training set and test set both include samples from the same patient. Thus, these networks cannot be considered a general classifier for all patients, since it did not have a valid test set for assessing results on unseen patients. Ideally, three sets should be utilized: training, validation, and testing. The purpose of the validation set is to test the generalization of the network, such that it is not overtrained. Plus, it is typical practice to retrain neural networks until good results are achieved on the validation set. A separate testing set verifies that success on the validation set was not just by chance. Until large datasets are available for development of the algorithms, neural networks will not be considered for clinical use. Furthermore, neural networks generally have not achieved much

acceptance by the clinical community who prefer methods that are tied to underlying physiologic understanding.

Dual-Chamber Arrhythmia Detection

Since dual-chamber pacemakers have been combined into ICDs, the possibility of the use of information from the atrial electrogram for arrhythmia diagnosis has opened up. The most prevalent cause of delivery of false therapy is AF, which accounts for > 60% of all false shocks according to the literature. The simple addition of an atrial sensing lead can dramatically change the false detection statistics.

The first two-channel algorithm for intracardiac analysis incorporated timing of atrial activation as well as ventricular into the diagnostic logic of arrhythmia classification (98,99). This scheme was based on earlier work in which an esophageal pill electrode (33) provided P-wave identification as an adjunct to surface leads in coronary care and Holter monitoring (34). The early argument for adding atrial sensing for improvement of ICD tachycardia detection was advanced conceptually by Furman in 1982 (29), was demonstrated algorithmically by Arzbaecher et al. in 1984 (58), and was further confirmed by Schuger (100). This simple two-channel analysis offers a first-pass method for confirming a VT diagnosis when the ventricular rate exceeds the atrial (Fig. 8). Recognition of a run of short intervals was followed by a comparison of atrial versus ventricular rate. With both chambers (atrial and ventricular) under analysis, most supraventricular arrhythmias could be detected by an $N : 1$ ($A : V$) relationship, and most ventricular arrhythmias could be detected by a $1 : N$ ($A : V$) relationship. Ambiguity occurred in tachycardias characterized by a 1 : 1 relationship, where SVT with 1 : 1 ventricular conduction could be confounded with ventricular tachycardia with retrograde 1 : 1 atrial conduction. In addition, an $N : 1$ ($A : V$) relationship should not be an automatic detection of atrial arrhythmia, since a concurrent ventricular arrhythmia could be masked by a faster atrial arrhythmia, as seen in Fig. 6c. Thus the limitations of two-channel timing analysis, although powerful, needs to be addressed by more advanced logical relationships.

A system designed for two-channel analysis using rate in both chambers plus three supplemental time features (onset derived by median filtering, regularity, and multiplicity) was designed for real-time diagnosis (101) of spontaneous rhythms. This system was an integration of previously tested stand-alone timing schemes (102,103). The combined system is able to recognize competing atrial and ventricular tachycardias and produces joint diagnoses of the concurrent rhythms. Simultaneous VT and atrial flutter is classified via atrial rate, ventricular rate, and a lack of multiplicity. Fast ventricular response in AF is detected via the regularity criterion. Onset (employed in 1:1 tachycardias) utilizes a median filter technique (102).

Commercially, each manufacturer now has available algorithms that utilize information from both chambers for making the diagnosis, particularly for improving specificity. These algorithms are implemented in commercial devices and continually updated and improved. Examples of the algorithms are in the following paragraphs. Reviews are given in Refs. 42,43 along with a thorough comparison of clinical results of the various commercial dual-chamber algorithms (43). Other comparisons include Refs. 104,105.

The first actual realization of a two-channel ICD appeared with the introduction into clinical trials (1995) of the ELA *Defender,* a dual chamber sensing and pacing ICD that uses both atrial and ventricular signals for its tachycardia diagnoses (106) (Fig. 9). The first step after a fast rate is detected is to consider stability of the ventricular rate. If the rhythm is not stable, atrial fibrillation is detected and no therapy delivered. The next consideration is the association between the A and V. For $A : V$ association of $1 : N$ or no association, a VT is detected. For $N : 1$ association, atrial arrhythmia is detected and no therapy delivered. For 1:1 association, the last step is consideration of chamber of onset, ventricular acceleration will result in VT therapy, while no acceleration or atrial acceleration will result in no therapy. An example of a sudden onset in the atrium due to SVT is seen in Fig. 6d. The most recent algorithm, PARAD+ incorporates additional features after the association criteria (second step) (107). If there is no PR association, a second criteria is considered where a single

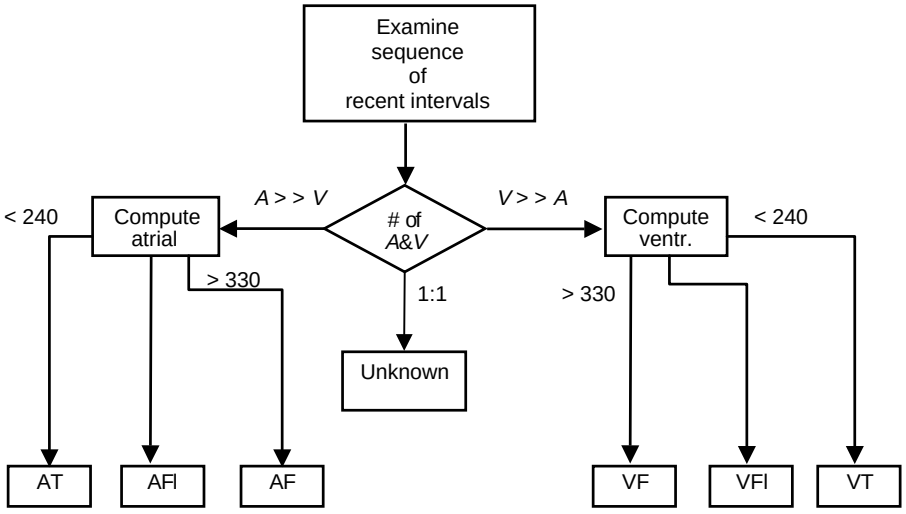


Figure 8. Basic dual chamber arrhythmia detection algorithm. For a sequence of intervals, the number of atrial (A) intervals is compared to the number of ventricular (V) intervals. If there are more V than A, a diagnosis of ventricular fibrillation (VF), ventricular flutter (VFl), or ventricular tachycardia (VT) is made based on the rate. If there are more V than A beats, a diagnosis of atrial fibrillation (AF), atrial flutter (AFl), or atrial tachycardia is made (AT) (58).

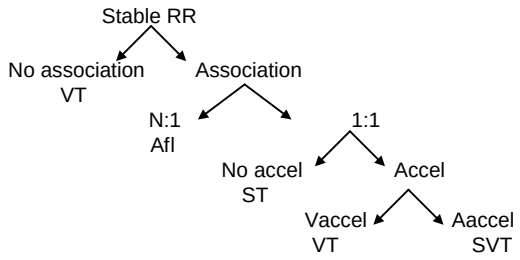


Figure 9. Flowchart of dual-chamber arrhythmia detection algorithm using simple rate-based features. For unstable RR interval (time between beats), atrial fibrillation is detected. For stable RR and no association between the atrium (A)and ventricle (V), ventricular tachycardia (VT) is detected. For $N:1$ ($A:V$) association, atrial flutter (AFl) is detected. For $1:1$ ($A:V$) association with no acceleration (Accel), sinus tachycardia (ST) is detected. Last, with a ventricular acceleration, VT is detected and with atrial acceleration (Aaccel) supraventricular tachycardia (SVT) is detected (106).

long ventricular cycle will result in the diagnosis of atrial fibrillation (for the next 24 consecutive cycles) while no long ventricular cycles will result in VT detection.

The Guidant Ventak AV III DR algorithm uses the following scheme, shown in Fig. 10 (104). First, it checks if the ventricular rate is greater than the atrial rate (by 10 bpm). If yes, then VT is detected. If no, then more analysis is performed. If the atrial rate is greater than the atrial fibrillation threshold and the RR intervals are not stable, then supraventricular rhythm is classified. If the RR intervals are stable, VT is detected. If the atrial rate is not greater than the atrial fibrillation threshold, then ventricular tachycardia is detected if the RR intervals are stable and there is a sudden onset of ventricular rate. An updated algorithm from Guidant is described in the next section.

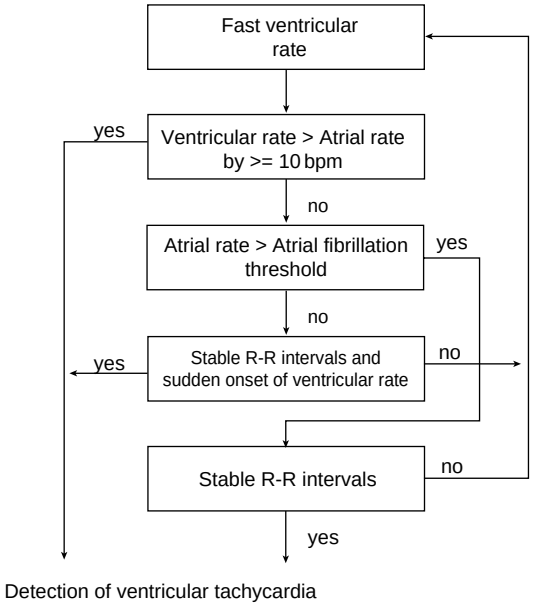


Figure 10. Algorithm for Guidant Ventak AV III DR, using comparison of the rate of A and V, stability and onset. (Used with permission from Ref. 104.)

Perhaps the most complex of the dual chamber algorithms rests with the PR Logic algorithm, by Medtronic (42). This algorithm uses a series of measurements from the timing of the atrial ventricular depolarizations to create a code (1 of 19 possible). These codes are then used for ultimate diagnosis. The main component of the algorithm is the timing between the atrial and ventricular beat to determine if the conduction was antegrade or retrograde. For a given ventricular RR interval, if the atrial beat falls 80 ms before or 50 ms after the ventricular beat, the rhythm is considered junctional. Outside of this, if the atrial beat (P-wave) falls within the first one-half of the RR interval, it is considered retrograde conduction. If the atrial beat falls in the second half of the RR interval it is considered antegrade. This is performed for the previous two beats and incorporated in the code. There are only a few programmable components in the algorithm. The first is the type of SVT for which rejection rules apply (AF/AFL, ST, SVT). The second is the SVT limit. The rest are not programmable.

Dual-Chamber with Ventricular Morphological Analysis

The Photon DR from St. Jude incorporates morphology in its dual chamber defibrillator algorithm. The MD in the ventricular chamber described earlier is incorporated in the full algorithm as follows (108). For $V > A$, ventricular tachycardia is detected. For $V < A$, a combination of morphology discrimination and interval stability is used to inhibit therapy for atrial fibrillation/flutter and SVT. For the branch $V = A$, morphology discrimination and sudden onset is used to inhibit therapy for ST and SVT. This algorithm has an automatic template feature update (ATU) for real-time calibration of the sinus template.

A new dual chamber algorithm from Guidant uses the vector timing and correlation (VTC) algorithm, described earlier (69). If the V rate exceeds A rate by $>$ 10 bpm, a VT is detected. Otherwise, VTC algorithm is implemented. If the atrial rate does not exceed the AF threshold, then VTC will be used for diagnosis. Otherwise, stability will be used for diagnosis. Therapy would be inhibited for an unstable ventricular rhythm.

Two-Channel Morphological Analysis. An early algorithm that uses morphological analysis of both the intraatrial signal and the intraventricular signal (109,110) is based on strategy developed previously for surface and esophageal signals (111). A five-feature vector was derived for each cycle containing an atrial and a ventricular waveform metric (ρ_a, ρ_v), where ρ is the correlation coefficient for each depolarization, and AA, AV, and VV interval classifiers (short, normal, and long). Single-cycle codes were mapped to 122 diagnostic statements. The eight most current cycles were employed for a contextual interpretation of the underlying rhythm. This addition of morphological analysis of both atrial and ventricular channels combined with rate determination in each channel on a cycle-by-cycle basis, dramatically demonstrated the power of modern signal processing in the interpretation of arrhythmias.

One aspect in which analysis of the atrial morphology would be very useful in ICDs is the separation of antegrade versus retrograde atrial conduction. During a 1:1 tachycardia, it is difficult to separate an SVT with 1:1

anterograde conduction (forward conduction from the sinus node through the atrium and AV node to the ventricle) versus a ventricular arrhythmia with retrograde conduction (retrograde conduction from the ventricle through the AV node to the atrium). To differentiate these cases, morphology differences in the atrial electrogram could be utilized, where abnormal morphology would indicate retrograde conduction. Various methods have been described in the literature which use similar approaches as ventricular morphology (112–118).

Distinction of Ventricular Tachycardia and Ventricular Fibrillation

Discriminating between VT and VF might be useful to allow unique zone settings for choice of therapy. Antitachycardia pacing is a lower energy therapy used to treat VT, which is not painful to the patient. Currently, there is difficulty in detecting each VF cycle, leading to electrogram dropout, which leads physicians to expand the VF detection zone to eliminate the possibility of misdiagnosing VF (119,120). Therefore, many VTs are detected as VF and given shock therapy directly. While these are typically fast VTs, there is a possibility that fast VTs can be terminated using anti-tachycardia pacing protocols, with only limited delay of shock therapy, if fast VTs and VF could be differentiated (121). In one study, 76% of fast VTs would have received shock therapy if programmed traditionally (121). However, by expanding the fast VT zone, 81% diagnosed as fast VT were effectively pace-terminated. More sophisticated digital signal processing techniques could be applied to ensure proper separation of VT and VF by methods more intelligent than counting alone.

For separation of VT and VF, CWA using a sinus rhythm template was tested on a passage of monomorphic ventricular tachycardia and a subsequent passage of ventricular fibrillation in each patient (122–124). The standard deviation of the correlation coefficient (ρ) of each class (VT and VF) was used as a discriminant function. This scheme was based upon the empiric knowledge that correlation values are more tightly clustered in the cycle-by-cycle analysis of monomorphic VT and more broadly distributed in the dissimilar waveforms in VF. Results showed easy separation of sinus rhythm from VT and VF; however in the VT/VF separation, standard deviation only achieved limited success Standard deviation requires patient-specific thresholds, may not hold for all template-based algorithms, and adds further computational requirements to the algorithm; therefore, it is not a promising algorithm in its present form for discrimination of VT from VF.

Throne et al. (125) addressed the problem of separating monomorphic and polymorphic VT/VF by using scatter diagram analysis. A moving average filter was applied to rate and morphology channels and plotted as corresponding pairs of points on a scatter diagram with a 15×15 grid. The percentage of grid blocks occupied by at least one sample point was determined. Investigators found that monomorphic VTs trace nearly the same path in two-dimensional space and occupy a smaller percentage of the graph than nonregular rhythms such as polymorphic VT or VF.

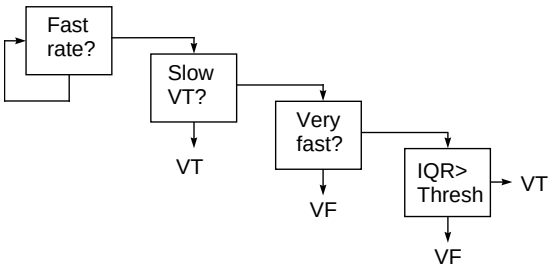


Figure 11. Basic algorithm for separation of VT and VF using PSC. Once a fast rate is detected, VT and VF are detected for slow fast rates and very fast rates, respectively. Only in the overlap between fast VT and VF rates, is the morphology algorithm implemented. Interquartile range (IQR) of the paired signal concordance over a passage is used.

A magnitude-squared coherence function was developed by Ropella et al. (126), which utilizes filtering and Fourier transformation of intraventricular electrograms derived from two leads with a sliding window to distinguish monomorphic ventricular tachycardia from polymorphic ventricular tachycardia and ventricular fibrillation. This method, while elegant, requires multiple electrode sites and is at present too computationally demanding for consideration in battery operated devices. As technology advances, the possibility of hardware implementation of frequency-based methods such as magnitude-squared coherence and time-domain CWA may become feasible.

A similar algorithm uses two "unipolar" ventricular electrograms, 1 cm apart, to compare the paired signal concordance (PSC) between the electrograms using correlation analysis (127). During normal rhythms and VT, the two closely spaced electrograms will exhibit high correlation, while during VT, the two electrogram will experience low correlation. Considering only rhythms that have a fast rate in the overlap between fast VT and VF rates, the variability of the correlation, measured by interquartile range, over a passage distinguishes VT from VF (Fig. 11).

Complexity measurements have also been utilized for distinction of ventricular tachycardia and fibrillation, including approximate entropy (128), Lempil–Ziv complexity (129), least-squares prony modeling algorithm (130).

OTHER DEVICES THAT USE AUTOMATED ARRHYTHMIA DETECTION

Other commercially available devices that use automated arrhythmia detection algorithms include the automatic external defibrillator and the implantable atrial defibrillator.

Automatic External Defibrillators

Recently, automatic external defibrillators (AEDs) have become widespread and available. The AED is able to determine if the rhythm for an unresponsive, pulseless patients is shockable or unshockable and is able to apply therapy automatically or to inform the user to deliver the

therapy (131–133). The AEDs are available on location for large organizations, such as airports, airplanes, businesses, sporting events, schools, and malls (134). This expanded availability dramatically increases the possibility that victims of ventricular fibrillation could receive defibrillation in a timely manner, thus, improving survival rates (135).

The AEDs, operating in a truly automated mode, must be exquisitely accurate in their interpretation of the ECG signal (136,137). In an AED, shockable rhythms are rhythms that will result in death if not treated immediately and include coarse ventricular fibrillation and ventricular tachycardia. Nonshockable rhythms are rhythms where no benefit and even possible harm may result from therapy and include supraventricular tachycardia, atrial fibrillation normal sinus rhythm, and asystole. Asystole is not considered shockable for these devices since the leads may be misplaced and no signal captured. Intermediate rhythms are rhythms that may or may not receive benefit from therapy and include fine VF and VT. Ventricular tachycardia is an intermediate rhythm because often it is hemodynamically tolerated in the patient. A rate threshold is usually programmed in the device (even though there is still no universally accepted or obvious delineation in rate for hemodynamic tolerance in the literature).

Contrary to the ICD, AEDs have an extremely necessary requirement for accurate specificity (shocks when not needed) since these devices are expected to be used by untrained personnel. Algorithms must consider large variations in cardiac rhythms and artifact from CPR or patient movement. The risk to the patient and the technician is too great to allow public use for devices with decreased specificity. Given that a bystander is on the scene and that trained help may soon be available, specificity is more important for the first, or immediate response. However, sensitivity must be considered to ensure the AEDs potential to save lives is maximized. In the AED, more battery power can be utilized (since the device is not implanted), and therefore more sophisticated schemes borrowed from ICD technology have been considered.

Current devices use numerous schemes for determining if the patient is in ventricular fibrillation. Common components include isoelectric content (like PDF algorithm of the ICD), zero crossings, rate, variability, slope, amplitude and frequency (138), all similar techniques to those described in the ICD literature. A review of AED algorithms is given in Ref. 138.

One example of a recent algorithm described in the literature is for a programmable automatic external defibrillator designed to be used in the hospital setting for monitoring and automatic defibrillation, if needed (139). This device uses a programmable rate criterion to detect shockable rhythms. In addition, the device has an algorithm which distinguishes VT and SVT rhythms below a SVT threshold. The algorithm uses three features to discriminate between SVT and VT. The first is the modulation domain function that uses amplitude and frequency characteristics. The second, called waveform factor (WF), provides a running average of the electrocardiogram signal amplitude normalized by the R-wave amplitude. The WF for one beat is as follows:

$$\mathrm{WF}_i = \frac{100 \times \sum_{1}^{N} \mathrm{abs}(A_n)}{N \times \mathrm{abs}(A_r)}$$

where N is the total number of samples between the previous and current beat, A_n is the nth amplitude of the signal, and A_r is the peak. The algorithm uses an eight beat running average. The SVT rhythms would have a small WF value, while VT (which has a wide QRS complex on the surface of the body) would have a high WF value. This algorithm should not be used with patients who have bundle branch block, chronic or paroxysmal. The third feature is called amplitude variability analysis factor, which uses distribution of the average derivative. The measurement is found by calculating the number of derivatives which fall into the baseline bin as a percentage of the total number of sample points. Amplitude variability analysis (AVA) is calculated as follows:

$$\mathrm{AVA} = \frac{100 \times \sum n(i)}{N}$$

where the summation is performed across the baseline bins and $n(i)$ is the number of samples for the ith bin. The baseline bins are defined as 25% of the total bins centered at $n(i)_{\max}$. Exact use of these features in the AED algorithm are not described.

Implantable Atrial Defibrillators

Implantable atrial defibrillators are used in patients with paroxysmal or persistent atrial fibrillation, particularly those which are symptomatic and drug refractory (140,141). Goals in atrial defibrillators are different than ICDs since atrial tachyarrhythmias are typically hemodynamically tolerable, therefore, more time and care can be used to make the decision. The challenge is that the device must sense low, variable amplitude atrial signals, while not sensing far-field ventricular waves. Furthermore, there are also multiple therapies available, antitachycardia pacing for atrial tachycardia, and cardioversion for atrial fibrillation. Lastly, some atrial defibrillators have been combined with ICDs such that back-up ventricular defibrillation therapy is available in this susceptible population (140,142).

A review of algorithms used in atrial defibrillators in given in Ref. 143. For example, Medtronic has a dual-chamber defibrillator that has both atrial and ventricular therapies (140,142). The algorithm for detection uses the same algorithm as the dual-chamber ventricular (only) defibrillator, PR Logic. In addition to this algorithm, there are two zones used for detection of atrial tachycardia and of atrial fibrillation. If the zones overlap, AT is detected if it is regular and AF if it is irregular. The purpose of multiple zones is similar to ventricular devices, in that a variety of therapies can be selected and utilized for each zone. For this device, this includes pacing algorithms for prevention, pacing therapies for termination, and high voltage shocks.

CONCLUSION

This article focuses on the overall approaches used for automated arrhythmia detection. However, this review did not delve into the specifics of comparisons of sensitivity and specificity results for the various algorithms. While, each paper referenced gives performance for a specific test database, it is difficult to compare the results from one study to another. There have been some attempts to develop standardized datasets, including surface electrocardiograms from Physionet (including the MIT-BIH databases) (144) and American Heart Association (145), and intracardiac electrograms, in addition to surface, from Ann Arbor Electrogram Libraries (3). Use of these datasets allows for comparisons, but does not address the differences between performance at the system level that incorporates the hardware components of the specific device. A comprehensive description of the pitfalls in comparing results from one study to another is given in (43). These include limitations of (*1*) benchtesting that does not incorporate specific ICD-system differences and spontaneous arrhythmias, (*2*) limited storage in the ICD making gold standard clinical diagnosis difficult, (*3*) great variations in settings of rate based thresholds and zones, (*4*) variability of types of rhythms included in the study, among others.

In conclusion, examples from the long history of automated arrhythmia detection for implantable cardioverter defibrillators is given with a brief mention of automated external defibrillators and implantable atrial defibrillators. The ICDs are beginning to reach maturity in terms of addressing both sensitivity and specificity in performance of the algorithms to achieve close to perfect detection of life-threatening arrhythmias, with greatly reduced false therapies. In the meantime, automated external defibrillators and implantable atrial defibrillators have learned many lessons from the ICD experience to provide accurate arrhythmia diagnosis. Devices on the horizon incorporating automated arrhythmia detection may include wearable external defibrillators (146,147), wearable wireless monitors, and beyond. This rich area of devices that detect and treat life-threatening arrhythmias shall reduce the risk of sudden cardiac death.

BIBLIOGRAPHY

Cited References

1. American Heart Association. Heart Disease and Stroke Statistics—2005 Update. American Heart Association. Available at http://www.americanheart.org. Accessed 2005.
2. Vilke GM, et al. The three-phase model of cardiac arrest as applied to ventricular fibrillation in a large, urban emergency medical services system. Resuscitation 2005;64:341–346.
3. Jenkins JM, Jenkins RE. Arrhythmia database for algorithm testing: surface leads plus intracardiac leads for validation. J Electrocardiol 2003;36(Suppl):157–1610. Available at http://www.electrogram.com.
4. Nisam S, Barold S. Historical evolution of the automatic implantable cardioverter defibrillator in the treatment of malignant ventricular tachyarrhythmias. In: Alt E, Klein H, Griffin JC, editors. The implantable cardioverter/defibrillator. Berlin: Springer-Verlag; 1992. pt. 1, p 3–23.
5. Mower MM, Reid PR. Historical development of automatic implantable cardioverter-defibrillator. In: Naccarelli GV, Veltri EP, editors. Implantable Cardioverter-Defibrillators, Boston: Scientific Publications; 1993. Chapt. 2. p 15–25.
6. Mower MJ. Clinical and historical perspective. In: Singer I, editor. Implantable Cardioverter Defibrillator. Armonk (NY): Futura; 1994, Chapt. 1. p 3–12.
7. Bach SM, Shapland JE. Engineering aspects of implantable defibrillators. In: Saksena S, Goldschlager N, editors. Electrical Therapy for Cardiac Arrhythmias. Philadelphia: Saunders; 1990. Chapt. 18. p 371–383.
8. Kennedy HL. Ambulatory (Holter) Electrocardiography Technology. Cardiol Clin 1992;10:341–359.
9. Feldman CL, Amazeen PG, Klein MD, Lown B. Computer detection of ventricular ectopic beats. Comput Biomed Res 1970 Dec; 3(6):666–674.
10. Thomas LJ, et al. Automated Cardiac Dysrhythmia Analysis. Proc IEEE Sept 1979;67(9):1322–1337.
11. Collins SM, Arzbaecher RC. An efficient algorithm for waveform analysis using the correlation coefficient. Comput Biomed Res 1981 Aug; 14(4):381–389.
12. Hulting J, Nygards ME. Evaluation of a computer-based system for detecting ventricular arrhythmias. Acta Med Scand 1976;199(12):53–60.
13. Thakor NV. From Holter monitors to automatic defibrillators: developments in ambulatory arrhythmia monitoring. IEEE Trans Biomed Eng 1984 Dec; 31(12):770–778.
14. Feldman CL, Hubelbank M. Cardiovascular Monitoring in the Coronary Care Unit. Med Instrum 1977;11:288–292.
15. Hubelbank M, Feldman CL. A 60x computer-based Holter tape processing system. Med Instrum 1978;Nov–Dec; 12(6):324–326.
16. Govrin O, Sadeh D, Akselrod S, Abboud S. Cross-correlation technique for arrhythmia detection using PR and PP intervals. Comput Biomed Res 1985 Feb; 18(1):37–45.
17. Shah PM, et al. Automatic real time arrhythmia monitoring in the intensive coronary care unit. Am J Cardiol 1977 May 4; 39(5):701–708.
18. Lipschultz A. Computerized arrhythmia monitoring systems: a review. J Clin Eng 1982 Jul–Sep; 7(3):229–234.
19. Lin D, et al. Analysis of time and frequency domain patterns of endocardial electrograms to distinguish ventricular tachycardia from sinus rhythm. Comp Cardiol 1987; 171–174.
20. Knoebel SB, Lovelace DE, Rasmussen S, Wash SE. Computer detection of premature ventricular complexes: a modified approach. Am J Cardiol 1976 Oct; 38(4):440–447.
21. Yanowitz F, Kinias P, Rawling D, Fozzard HA. Accuracy of a continuous real-time ECG dysrhythmia monitoring system. Circulation. 1974 July; 50(1):65–72.
22. Mead CN, et al. A detection algorithm for multiform premature ventricular contractions. Med Instrum 1978;12:337–339.
23. Knoebel SB, Lovelace DE. A two-dimensional clustering technique for identification of multiform ventricular complexes. Med Instrum 1978;12:332–333.
24. Cheng QL, Lee HS, Thakor NV. ECG waveform analysis by significant point extraction. II. Pattern matching. Comput Biomed Res 1987 Oct; 20(5):428–442.
25. Spitz AL, Harrison DC. Automated family classification in ambulatory arrhythmia monitoring. Med Instrum 1978 Nov–Dec; 12(6):322–323.
26. Oliver GC, et al. Detection of premature ventricular contractions with a clinical system for monitoring electrocardiographic rhythms. Comput Biomed Res 1971 Oct; 4(5): 523–541.
27. Yanowitz F, Kinias P, Rawling D, Fozzard HA. Accuracy of a continuous real-time ECG dysrhythmia monitoring system. Circulation 1974 July; 50(1):65–72.

28. Cooper DH, Kennedy HL, Lyyski DS, Sprague MK. Holter triage ambulatory ECG analysis. Accuracy and time efficiency. J Electrocardiol. 1996 Jan; 29(1):33–38.
29. Furman S, Fisher JK, Panizzo F. Necessity of signal processing in tachycardia detection. In: Barold SS, Mugica J, editors. The Third Decade of Cardiac Pacing: Advances in Technology and Clinical Applications. Mt Kisco (NY): Futura; 1982. Pt. 3, Chapt. 1. p 265–274.
30. Jenkins J, et al. Present state of industrial development of devices. PACE May–June 1984;7(II):557–568.
31. Mirowski M, Mower MM, Reid PR. The automatic implantable defibrillator. Am Heart J 1980;100:1089–1092.
32. Langer A, Heilman MS, Mower MM, Mirowski M. Considerations in the development of the automatic implantable defibrillator. Med Instrum May–June 1976;10:163–167.
33. Arzbaecher R. A pill electrode for the study of cardiac dysrhythmia. Med Instrum 1978;12:277–281.
34. Jenkins JM, Wu D, Arzbaecher R. Computer diagnosis of supraventricular and ventricular arrhythmias. Circulation 1979;60:977–987.
35. Paul VE, O'Nunain S, Malik M. Temporal electrogram analysis: algorithm development. PACE Dec. 1990;13:1943–1947.
36. Toivonen L, Viitasalo M, Jarvinen A. The performance of the probability density function in differentiating supraventricular from ventricular rhythms. PACE May 1992;15: 726–730.
37. DiCarlo L, et al. Tachycardia detection by antitachycardia devices: present limitations and future strategies. J Interven Cardiol 1994;7:459–472.
38. Pannizzo F, Mercando AD, Fisher JD, Furman S. Automatic methods for detection of tachyarrhythmias by antitachycardia devices. PACE Feb. 1988;11:308–316.
39. Lang DJ, Bach SM. Algorithms for fibrillation and tachyarrhythmia detection. J Electrocardiol 1990;23(Suppl):46–50.
40. Olson WH. Tachyarrhythmia sensing and detection. In: Singer I, editor. Implantable Cardioverter Defibrillator. Armonk (NY): Futura; 1994. Chapt. 4. p 71–107.
41. Jenkins JM, Caswell SA. Detection algorithms in implantable defibrillators. Proc IEEE 1996;84:428–445.
42. Olson WH. Dual chamber sensing and detection for implantable cardioverter-defibrillators. In: Singer I, Barold SS, Camm AJ, editors. Nonpharmacological Therapy of Arrhythmias for the 21st century. Armonk (NY): Futura; 1998. p 385–421.
43. Aliot E, Mitzsche R, Ribart A. Arrhythmia detection by dual-chamber implantable cardioverter defibrillators: a review of current algorithms. Europace 2004;6:273–286.
44. Thakor NV, Webster JG. Design and evaluation of QRS and noise detectors for ambulatory ECG monitors. Med Biol Eng Comput 1982;20:709–714.
45. Jalaleddine S, Hutchens C. Ambulatory ECG wave detection for automated analysis: a review. ISA Trans 1987;26(4): 33–43.
46. Warren JA, et al. Implantable cardioverter defibrillators. Proc IEEE 1996;84:468–479.
47. MacDonald R, Jenkins J, Arzbaecher R, Throne R. A software trigger for intracardiac waveform detection with automatic threshold adjustment. Proc Computers Cardiol IEEE 30276-6574 1990; 167–170.
48. Winkle RA, et al. Long-term outcome with the automatic implantable cardioverter-defibrillator. J Am Coll Cardiol May 1989;13:1353–1361.
49. Grimm W, Flores BF, Marchlinski FE. Electrocardiographically documented unnecessary, spontaneous shocks in 241 patients with implantable cardioverter defibrillators. PACE Nov. 1992;15:670–669.
50. Nunain SO, et al. Limitations and late complications of third-generation automatic cardioverter-defibrillators. Circulation April 15 1995;91:2204–2213.
51. Warren J, Martin RO. Clinical evaluation of automatic tachycardia diagnosis by an implanted device. PACE 1986; 9:16.
52. Nathan AW, Creamer JE, Davies DW. Clinical experience with a new versatile, software based, tachycardia reversion pacemaker. J Am Coll Cardiol 1987;7:184A.
53. Olson W, Bardy G, Mehra R. Onset and stability for ventricular tachycardia detection in an implantable pacer-cardioverter-defibrillator. Comp Cardiol 1987;34:167–170.
54. Tomaselli G, Scheinman M, Griffin J. The utility of timing algorithms for distinguishing ventricular from supraventricular tachycardias. PACE March–April 1987;10:415.
55. Geibel A, Zehender M, Brugada P. Changes in cycle length at the onset of sustained tachycardias-importance for antitachycardia pacing. Am Heart J March 1988;115:588–592.
56. Ripley KL, Bump TE, Arzbaecher RC. Evaluation of techniques for recognition of ventricular arrhythmias by implanted devices. IEEE Trans Biomed Eng June 1989;36:618–624.
57. Pless BD, Sweeney MB. Discrimination of supraventricular tachycardia from sinus tachycardia of overlapping cycle length. PACE Nov–Dec 1984;7:1318–1324.
58. Arzbaecher R, et al. Automatic tachycardia recognition. PACE May–June 1984;7:541–547.
59. Jenkins JM, et al. Tachycardia detection in implantable antitachycardia devices. PACE Nov–Dec 1984;7:1273–1277.
60. Swerdlow CD, et al. Discrimination of ventricular tachycardia from sinus tachycardia and atrial fibrillation in a tiered-therapy cardioverter. J Am Coll Cardiol 1994;23:1342–1355.
61. Pannizzo F, Furman S. Pattern recognition for tachycardia detection: a comparison of methods. PACE July 1987;10:999.
62. Santel D, Mehra R, Olson W. Integrative algorithm for detection of ventricular tachyarrhythmias from the intracardiac electrogram. Comp Cardiol 1987; 175–177.
63. Lin D, DiCarlo LA, Jenkins JM. Identification of ventricular tachycardia using intracavity ventricular electrograms: analysis of time and frequency domain patterns. PACE Nov. 1988;1592–1606.
64. Tomaselli GF, et al. Morphologic differences of the endocardial electrogram in beats of sinus and ventricular origin. PACE Mar. 1988;11:254–262.
65. Langberg JL, Gibb WJ, Auslander DM, Griffin JC. Identification of ventricular tachycardia with use of the morphology of the endocardial electrogram. Circulation June 1988;77: 1363–1369.
66. Throne RD, Jenkins JM, Winston SA, DiCarlo LA. A comparison of four new time domain methods for discriminating monomorphic ventricular tachycardia from sinus rhythm using ventricular waveform morphology. IEEE Trans Biomed Eng June 1991;38:561–570. (U. S. Pat. No. 5,000,189 Mar. 19, 1991).
67. Steinhaus BM, et al. Detection of ventricular tachycardia using scanning correlation analysis. PACE Dec. 1990;13: 1930–1936.
68. Greenhut SE, et al. Separation of ventricular tachycardia from sinus rhythm using a practical, real-time template matching computer system. PACE Nov. 1992;15:2146–2153.
69. Gold MR, et al. Advanced rhythm discrimination for implantable cardioverter defibrillators using electrogram vector timing and correlation. J Cardiovasc Electrophysiol 2002; 13:1092–1097.
70. Swerdlow CD, et al. Discrimination of ventricular tachycardia from supraventricular tachycardia by a downloaded wavelet transform morphology algorithm. J Cardiovasc Electrophysiol 2002;13:432–441.

71. Duru F, et al. Morphology discriminator feature for enhanced ventricular tachycardia discrimination in implantable cardioverter defibrillators. PACE 2000;23:1365–1374.
72. Boriani G, et al. Clinical evaluation of morphology discrimination: an algorithm for rhythm discrimination in cardioverter defibrillators. PACE 2001;24:994–1001.
73. Throne RD, Jenkins JM, Winston SA, DiCarlo LA. Use of tachycardia templates for recognition of recurrent monomorphic ventricular tachycardia. Comp Cardiol 1989;171–174.
74. Stevenson SA, Jenkins JM, DiCarlo LA. Analysis of the intraventricular electrogram for differentiation of distinct monomorphic ventricular arrhythmias. J Am Coll Cardiol (submitted June 1995;).
75. Paul VE, et al. Variability of the intracardiac electrogram: effect on specificity of tachycardia detection. PACE Dec. 1990;13:1925–1829.
76. Finelli CJ, et al. Intraventraventricular electrogram morphology: effect of increased heart rate with and without accompanying changes in sympathetic tone. Comp Cardiol 1990; 115–118.
77. Rosenheck S, Schmaltz S, Kadish AH, Morady F. Effect of rate augmentation and isoproterenol on the amplitude of atrial and ventricular electrograms. Am J Cardiol July 1 1990;66:101–102.
78. Belz MK, et al. The effect of left ventricular intracavitary volume on the unipolar electrogram. PACE Sept. 1993;16: 1842–1852.
79. Caswell SA, et al. Chronic bipolar electrograms are stable during changes in body position and activity: implications for antitachycardia devices. PACE April 1995;18:871.
80. Throne RD, Jenkins JM, Winston SA, DiCarlo LA. Paroxysmal bundle branch block of supraventricular origin: a possible source of misdiagnosis in detecting ventricular tachycardia using ventricular electrogram morphology. PACE April 1990;13:453–458.
81. Gilberg JM, Olson WH, Bardy GH, Mader SJ. Electrogram width algorithms for discrimination of supraventricular rhythm from ventricular tachycardia. PACE April 1994;17:866.
82. Unterberg C, et al. Long-term clinical experience with the EGM width detection criteria for differentiation of supraventricular and ventricular tachycardia in patients with implantable cardioverter defibrillators. PACE 2000;23: 1611–1617.
83. Kingenheben T, Sticherling C, Skupin M, Hohnloser SH. Intracardiac QRS electrogram width—an arrhythmia detection feature for implantable cardioverter defibrillators: exercise induced variation as a base for device programming. PACE 1998;21:1609–1617.
84. Favale S, et al. Electrogram width parameter analysis in implantable cardioverter defibrillators: influence of body position and electrode configuration. PACE 2001;24:1732–1738.
85. Leitch JW, et al. Correlation between the ventricular electrogram amplitude in sinus rhythm and in ventricular fibrillation. PACE Sept. 1990;13:1105–1109.
86. Ellenbogen KA, et al. Measurement of ventricular electrogram amplitude during intraoperative induction of ventricular tachyarrhythmias. Am J Cardiol Oct. 15 1992;70:1017–1022.
87. Pannizzo F, Furman S. Frequency spectra of ventricular tachycardia and sinus rhythm in human intracardiac electrograms: application to tachycardia detection for cardiac pacemakers. IEEE Trans Biomed Eng June 1988; 421–425.
88. Slocum J, Sahakian A, Swiryn S. Computer discrimination of atrial fibrillation and regular atrial rhythms from intra-atrial electrograms. PACE May 1988;11:610–621.
89. Aubert AE, et al. Frequency analysis of VF episodes during AICD implantation. PACE June 1988;11(Suppl):891.
90. Lovett EG, Ropella KM. Autoregressive spread-spectral analysis of intracardiac electrograms: comparison of Fourier analysis. Comp Cardiol 1992; 503–506.
91. Minami K, Nakajima H, Toyoshima T. Real-time discrimination of ventricular tachyarrhythmia with Fourier-transform neural network. IEEE Trans Biomed Eng 1999;46:179–185.
92. Yan MC, Jenkins JM, DiCarlo LA. Feasibility of arrhythmia recognition by antitachycardia devices using time-frequency analysis with neural network classification. PACE 1995;18: 871.
93. Farrugia S, Yee H, Nickolls P. Implantable cardioverter defibrillator electrogram recognition with a multilayer perceptron. PACE Jan. 1993;16:228–234.
94. Leong PH, Jabri MA. MATIC-An intracardiac tachycardia classification system. PACE Sept. 1982;15:1317–1331.
95. Rojo-Alvarez JL, et al. Automatic discrimination between supraventricular and ventricular tachycardia using a multilayer perceptron in implantable cardioverter defibrillators. PACE 2002;25:1599–1604.
96. Wang Y, et al. A short-time multifractal approach for arrhythmia detection based on fuzzy neural network. IEEE Trans Biomed Eng 2001;48:989–995.
97. Al-Fahoum AS, Howitt I. Combined wavelet transformation and radial basis neural networks for classifying life-threatening cardiac arrhythmias. Med Biol Eng Comp 1999;27: 566–573.
98. Arzbaecher R, et al. Automatic tachycardia recognition. PACE May–June 1984;7:541–547.
99. Jenkins JM, et al. Tachycardia detection in implantable antitachycardia devices. PACE Nov–Dec 1984;7:1273–1277.
100. Schuger CD, Jackson K, Steinman RT, Lehmann MH. Atrial sensing to augment ventricular tachycardia detection by the automatic implantable cardioverter defibrillator: a utility study. PACE Oct. 1988;11:1456–1463.
101. Caswell SA, DiCarlo LA, Chiang CJ, Jenkins JM. Automated analysis of spontaneously occurring arrhythmias by implantable devices: limitation of using rate and timing features alone. J Electrocardiol 1994;27(Suppl):151–156.
102. Chiang CJ, Jenkins JM, DiCarlo LA. Discrimination of ventricular tachycardia from sinus tachycardia by antitachycardia devices: value of median filtering. Med Engr Phys Nov. 1994;16:513–517.
103. Chiang CJ, Jenkins JM, DiCarlo LA. The value of rate regularity and multiplicity measures to detect ventricular tachycardia in atrial fibrillation of flutter with a fast ventricular response. PACE Sept. 1994;17:1503–1508.
104. Hintringer F, Schwarzacher S, Eibl G, Pachinger O. Inappropriate detection of supraventricular arrhythmias by implantable dual chamber defibrillators: a comparison of four different algorithms. PACE 2001;24:835–841.
105. Hintringer F, et al. Comparison of the specificity of implantable dual chamber defibrillator detection algorithms. PACE 2004;27:976–982.
106. Lavergne T, et al. Preliminary clinical experience with the first dual chamber pacemaker defibrillator. PACE 1997;20: 182–188.
107. Mletzko R, et al. Enhanced specificity of a dual chamber ICD arrhythmia detection algorithm by rate stability criteria. PACE 2004;27:1113–1119.
108. Bailin SJ, et al. Clinical investigation of a new dual-chamber implantable cardioverter defibrillator with improved rhythm discrimination capabilities. J Cardiovasc Electrophysiol 2003;14:144–149.
109. Chiang CJ, et al. Real-time arrhythmia identification from automated analysis of intraatrial and intraventricular electrograms. PACE Jan. 1993;16:223–227.

110. Caswell SA, et al. Pattern recognition of cardiac arrhythmias using two intracardiac channels. Comp Cardiol 1993; 181–184.
111. DiCarlo LA, Lin D, Jenkins JM. Automated interpretation of cardiac arrhythmias. J Electrocardiol Jan. 1993;26:53–67.
112. Amikan S, Furman S. A comparison of antegrade and retrograde atrial depolarization in the electrogram. PACE May 1983;6:A111.
113. Wainwright R, Davies W, Tooley M. Ideal atrial lead positioning to detect retrograde atrial depolarization by digitization and slope analysis of the atrial electrogram. PACE Nov–Dec. 1984;7:1152–1157.
114. Davies DW, Wainwright RJ, Tooley MA. Detection of pathological tachycardia by analysis of electrogram morphology. PACE March–April 1986;9:200–208.
115. McAlister HF, et al. Atrial electrogram analysis: antegrade versus retrograde. PACE Nov. 1988;11:1703–1707.
116. Throne RD, et al. Discrimination of retrograde from anterograde atrial activation using intracardiac electrogram waveform analysis. PACE Oct. 1989;12:1622–1630.
117. Saba S, et al. Use of correlation waveform analysis in discrmination between anterograde and retrograde atrial electrograms during ventricular tachycardia. J Cardiovasc Electrophysiol 2001;12:145–149.
118. Strauss D, Jung J, Rieder A, Manoli Y. Classification of endocardial electrograms using adapted wavelet packets and neural networks. Ann Biomed Eng 2001;29:483–492.
119. DiCarlo LA, et al. Impact of varying electrogram amplitude sensing threshold upon the performance of rate algorithms for ventricular fibrillation detection. Circulation Oct. 1994;90:I–176.
120. Caswell SA, et al. Ventricular tachycardia versus ventricular fibrillation: Discrimination by antitachycardia devices. J Electrocardiol 1996;28:29.
121. Wathen MS, et al. PainFREE Rx II Investigators. Prospective randomized multicenter trial of empirical antitachycardia pacing versus shocks for spontaneous rapid ventricular tachycardia in patients with implantable cardioverter-defibrillators: Pacing Fast Ventricular Tachycardia Reduces Shock Therapies (PainFREE Rx II) trial results. Circulation 2004;110:2591–2596.
122. Jenkins JM, Kriegler C, DiCarlo LA. Discrimination of ventricular tachycardia from ventricular fibrillation using intracardiac electrogram analysis. PACE April 1991;14:718.
123. DiCarlo LA, Jenkins JM, Winston SA, Kriegler C. Differentiaion of ventricular tachycardia from ventricular fibrillation using intraventricular electrogram morphology. Am J Cardiol Sept. 15 1992;70:820–822.
124. Jenkins JM, Caswell SA, Yan MC, DiCarlo LA. Is waveform analysis a viable consideration for implantable devices given its computational demand? Comp Cardiol 1993; 839–842.
125. Throne RD, et al. Scatter diagram analysis: a new technique for discriminating ventricular tachyarrhythmias. PACE July 1994;17:1267–1275.
126. Ropella KM, Baerman JM, Sahakian AV, Swiryn S. Differentiation of ventricular tachyarrhythmias. Circulation Dec. 1990;82:2035–2043.
127. Caswell SA, Jenkins JM, DiCarlo LA. Comprehensive scheme for detection of ventricular fibrillation for implantable cardioverter defibrillators. J Electrocardiol 1998;30: 131–136.
128. Schuckers SA. Use of approximate entropy measurements to classify ventricular tachycardia and fibrillation. J Electrocardiol 1998;31(Suppl):101–105.
129. Zhang HX, Zhu YX, Wang ZM. Complexity measure and complexity rate information based detection of ventricular tachycardia and fibrillation. Med Biol Eng Comp 2000;38: 553–557.
130. Chen SW. A two-stage discrmination of cardiac arrhythmias using a total least squares-based prony modeling algorithm. IEEE Trans Biomed Eng 2000;47:1317–1327.
131. American Heart Assocation. Emergency Cardiac Care Committee and Subcommittees. Guidelines for cardiopulmonary resuscitation and emergency cardiac care. JAMA 1992;268: 2171–2302.
132. Aronson AL, Haggar B. The automatic external defibrillator-pacemaker: clinical rationale and engineering design. Med Instrum 1986;20:27–35.
133. Charbonnier FM. External defibrillators and emergency external pacemekers. Proc IEEE 1996;84:487–499.
134. Weisfeldt ML, et al. American Heart Association Report on the Public Access Defibrillation Conference December 8–10, 1994. Automatic External Defibrillation Task Force. Circulation 1995;92:2740–2747.
135. Dimmit MA, Griffiths SE. What's new in prehospital care? Nursing 1992;22:58–61.
136. Association for the Advancement of Medical Instrumentation. Automatic external defibrillators and remote-control defibrillators [American National Standard]. AAMI 1993; ANSI/AAMI DF39-1993.
137. American Heart Association. AED Task Force, Subcommittee on Safety and Efficacy. Automatic External Defibrillators for Public Access Use: Recommendations for Specifying and Reporting Arrhythmia Analysis Algorithm Performance, Incorporating new Waveforms, and Enhancing Safety. AHA 1996.
138. Charbonnier FM. Algorithms for arrhythmia analysis in AEDs. In: Tacker WA Jr, editor. Defibrillation of the Heart: ICDs, AEDs and Manual. St Louis (MO): Mosby/Yearbook; 1994.
139. Mattioni T, et al. Performance of an automatic external cardioverter-defibrillator algorithm in discrimination of supraventricular from ventricular tachycardia. Am J Cardiol 2003;91:1323–1326.
140. Sopher SM, Camm AJ. Atrial defibrillators. In: Singer I, Barold SS, Camm AJ, editors. Nonpharmacological Therapy of Arrhythmias for the 21st century. Armonk (NY): Futura; 1998. p 473–489.
141. Gold MR, et al. Clinical experience with a dual-chamber implantable cardioverter defibrillator to treat atrial tachyarrhythmias. J Cardiovasc Electrophysiol 2001;12: 1247–1253.
142. Swerdlow CD, et al. Detection of atrial fibrillation and flutter by a dual-chamber implantable cardioverter-defibrillator. Circulation 2000;101:878–885.
143. KenKnight BH, Lang DJ, Scheiner A, Cooper RAS. Atrial defibrillation for implantable cardioverter-defibrillators: lead systems, waveforms, detection algorithms, and results. In: Singer I, Barold SS, Camm AJ, editors. Nonpharmacological Therapy of Arrhythmias for the 21st century. Armonk (NY): Futura; 1998. p 457–471.
144. Costa M, Moody GB, Henry I, Goldberger AL. PhysioNet: an NIH research resource for complex signals. J Electrocardiol 2003;36(Suppl) 139–144. Available at http://www.physionet.org.
145. American Heart Association ECG Database, Available from ECRI, 5200 Butler Pike, Plymouth Meeting, PA 19462 USA, http://www.ecri.org/.
146. Reek S, et al. Clinical efficacy of a wearable defibrillator in acutely terminating episodes of ventricular fibrillation using biphasic shocks. PACE 2003;26:2016–2022.
147. Feldman AM, et al. Use of a wearable defibrillator in terminating tachyarrhythmias in patients at high risk for sudden death: results of WEARIT/BIROAD. PACE 2004;27:4–9.

See also AMBULATORY MONITORING; DEFIBRILLATORS; ELECTROCARDIOGRAPHY, COMPUTERS IN; EXERCISE STRESS TESTING.

ARTERIAL TONOMETRY. See TONOMETRY, ARTERIAL.

ARTIFICIAL BLOOD. See BLOOD, ARTIFICIAL.

ARTIFICIAL HEART. See HEART, ARTIFICIAL.

ARTIFICIAL HEART VALVE. See HEART VALVE PROSTHESES.

ARTIFICIAL HIP JOINTS. See HIP JOINTS, ARTIFICIAL.

ARTIFICIAL LARYNX. See LARYNGEAL PROSTHETIC DEVICES.

ARTIFICIAL PANCREAS. See PANCREAS, ARTIFICIAL.

ARTERIES, ELASTIC PROPERTIES OF

KOZABURO HAYASHI
Okayama University of Science
Okayama, Japan

INTRODUCTION

The elastic properties of the arterial wall are very important because they are closely related to arterial physiology and pathology, especially via effects on blood flow and arterial mass transport. Furthermore, stresses and strains in the arterial wall are prerequisite for the understanding of the pathophysiology and mechanics of the cardiovascular system. Stresses and strains cannot be analyzed without exact knowledge of the arterial elasticity.

STRUCTURE OF ARETRIAL WALL AND BASIC CHARACTERISTICS

Arteries become smaller in diameter with increasing distance from the heart, depending on functional demands (1). In concert with this reduction in size, their structure, chemical composition, and wall thickness-inner diameter ratio gradually change in a way that leads to a progressive increase both in stiffness and in their ability to change their inner diameter in response to a variety of chemical and neurological control signals.

Arterial wall is inhomogeneous not only structurally, but also histologically. It is composed of three layers (intima, media, and adventitia), which are separated by elastic membranes. Because the media is much thicker than the other two layers and supports load induced by blood pressure, its mechanical properties represent the properties of arterial wall. The media is mainly composed of elastin, collagen, and cells (smooth muscle cell and fibroblast). Roughly speaking, elastin gives an artery its elasticity, while collagen resists tensile forces and gives the artery its burst strength. Smooth muscle cells contract or relax in response to mechanical, chemical, and the other stimuli, which alters the deformed configuration of arteries. The wall compositions vary at different locations depending on required functions. For example, collagen and smooth muscle increase and elastin decreases at more distal sites in conduit arteries; the ratio of collagen to elastin increases in more distally located arteries. Collagen and elastin are essentially similar proteins, but collagen is very much stronger and stiffer than elastin. Therefore, the change of arterial diameter developed by blood pressure pulsation depends on the arterial site; it is larger in more proximal arteries.

Like most biological soft tissues, arteries undergo large deformation when they are subjected to physiological loading, and their force-deformation and stress–strain relations are nonlinear partly because of the above-mentioned inhomogeneous structure and partly because of the nonlinear characteristics of each component itself. Since collagen is a long-chained high polymer, it is intrinsically anisotropic. Moreover, not only collagen and elastin fibers, but also cells, are oriented in tissues and organs in order so that their functions be most effective. Inevitably, the arterial wall is mechanically anisotropic like many other biological tissues. Biological soft tissues including arterial wall demonstrate opened hysteresis loops in their force–deformation and stress–strain curves, which means that those tissues are viscoelastic. In such materials, the stress state is not uniquely determined by current strain, but depends also on the history of deformation. When a viscoelastic tissue is elongated and maintained at some length, load does not stay at a specific level, but decreases rather rapidly at first and then gradually (relaxation). If some constant load is applied to the tissue, it is elongated with time rather rapidly at first and then gradually (creep). Viscoelastic materials generally show different stress–strain properties under different strain rates. It is true, and higher strain rates give higher stresses. However, such a strain rate effect is not so much in biological soft tissues like arteries, namely, their elastic properties are not more sensitive to strain rate. Therefore, it is not always necessary to consider viscoelasticity for arterial mechanics; it is very often enough to assume wall material to be elastic. Many biological soft tissues contain water of $> 70\%$. Therefore, they hardly change their volume even if load is applied, and they are almost incompressible. The incompressibility assumption is very important in the formulation of constitutive laws of soft tissues, because it imposes a constraint on the strains and they are not independent.

MEASUREMENT OF ARTERIAL ELASTICITY

In Vitro Tests

It is widely recognized that the mechanical properties of blood vessels do not change for up to 48 h if tissues are stored at $\sim 4\ ^{\circ}C$ (1). One of the basic methods for the determination of the mechanical properties of biological tissues is uniaxial tensile testing on excised specimens. In this test, an increasing force is steadily applied to the longitudinal direction of a specimen, and the resulting specimen deformation is measured, which gives relations between stress (force divided by specimen cross-sectional area) and strain (specimen elongation divided by reference specimen length). This *in vitro* test is simple but, nevertheless, provides us with basic and useful information on

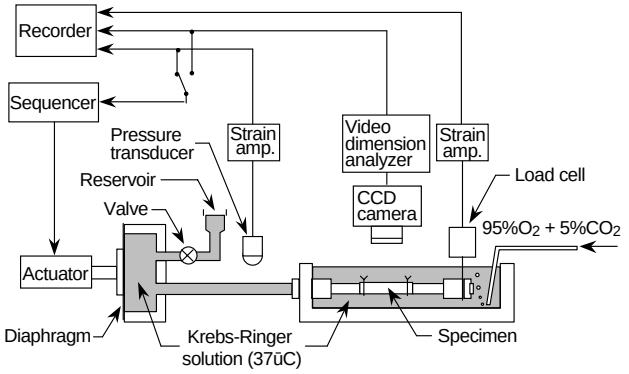


Figure 1. An *in vitro* experimental setup for the pressure–diameter–axial force test of a tubular arterial specimen. Internal pressure or outer diameter can be controlled with a feedback system (2).

the mechanical properties of tissues. Dumbbell-shape specimens, helically stripped specimens, and ring specimens are commonly used for arterial walls.

Under *in vivo* conditions, arteries are tethered to or constrained by perivascular connective tissues and side branches, and pressurized by blood from inside. These forces develop multiaxial stresses in the wall. For the determination of the mechanical characteristics of arteries under multiaxial conditions, biaxial tensile tests on flat specimens are utilized to simultaneously apply forces in the circumferential and longitudinal directions; however, the effect of wall radial stress is ignored in this case.

Although stress–strain data obtained from the above-mentioned uniaxial and biaxial tests on flat, strip, and ring specimens are often used to represent the elastic properties of arterial walls, the data obtained from pressure–diameter tests on the tubular segments of blood vessels are more important and realistic. An example of the test devices is shown in Fig. 1 (2). A tubular specimen is mounted in the bath filled with Krebs–Ringer solution, which is kept at 37 °C and aerated with 95% O_2 and 5% CO_2 gas mixture. Then, it is extend to the *in vivo* length to mimic the *in vivo* condition, because arteries inside the body are tethered to the surrounding tissues as mentioned above and, therefore, they are extended in the axial direction. A diaphragm-type actuator, which is controlled with a sequencer, is incorporated in the device for the application of internal pressure to the specimen. The internal pressure or specimen diameter can be controlled with the sequencer during pressure–diameter tests. The pressure is measured with a fluid-filled pressure transducer, while the outer diameter of the specimen is determined with a video dimension analyzer combined with a CCD camera. If the measurements of axial force are required in order to obtain pressure–diameter–axial force relations for the purpose of determining multiaxial constitutive laws, a load cell attached to one end of the vessel can be used.

In Vivo Measurements

It may be more realistic to obtain data from *in vivo* experiments under *in situ* conditions rather than to get data from *in vitro* biomechanical tests. As a result of recent progress in ultrasonic techniques, arterial diameter and even arterial wall thickness can be measured noninvasively with fairly good precision. These methods are being used not only for *in vivo* animal experiments, but also for clinical diagnosis of vascular diseases. It is true that the data obtained from these experiments and clinical cases are very useful, and provide important information concerning arterial mechanics. On the other hand, it is also true that many factors considerably affect the results obtained. These include physiological reactions to momentary changes in body and ambient conditions as well as the effects of anesthesia and respiration. In addition, since there has been some difficulty in applying the methods to small-diameter blood vessels, accurate measurements of vascular diameter and wall thickness with current techniques have been mostly limited to aortas and large arteries.

Before noninvasive ultrasonic techniques were developed, *in vivo* measurements of vascular diameter were invasively performed following surgical exposure of blood vessels, using strain gauge-mounted cantilevers, strain gauge-pasted calipers, and sonomicrometers. For example, a pair of miniature ultrasonic sensors may be used for the measurement of the outer diameter of a blood vessel (3). They are attached to the adventitial surface of a blood vessel so as to face each other across the vascular diameter. The diameter is determined from the transit time of the pulses between the two sensors. Similar sonomicrometers have been used for the measurement of arterial diameter not only in anesthetized, but also in conscious animals.

The noninvasive measurement of the elastic properties of arteries offers several significant advantages over invasive techniques. First, the nontraumatic character of the measurement guarantees a physiological state of the arterial wall, whereas such key functional elements of the wall as endothelium and smooth muscle might be affected in certain invasive measurement techniques. Second, it is of great clinical interest because it allows the monitoring of many outpatients and, therefore, it is well adapted for epidemiological or cross-sectional studies.

Noninvasive measurement of the arterial diameter can be done with ultrasonic echo-tracking techniques; recent improvements of the original technique have been proposed, which include digital tracking, prior inverse filtering, and coupling with B-mode imaging (1).

There exist no direct ways to measure pressure noninvasively in large central arteries, such as the aorta. Thus, regardless of the progress of ultrasonic and magnetic resonance imaging techniques which allow for the noninvasive measurement of vascular diameter, mechanical properties, such as compliance and elastic modulus cannot be derived from first principles. Therefore, primarily for clinical use, as an indirect, but noninvasive way of estimating the mechanical properties, the pulse wave velocity, c (see the next section), is often obtained from the measurements of pulsation at two distinct points along the vessel. One of the major drawbacks of this technique is low accuracy. The other one is that it yields a single value for the wave velocity. Because of the nonlinear elastic properties of the arterial wall, the pulse wave velocity sensitively changes depending on blood pressure. Therefore, the determination of a single value or a typical value of the arterial stiffness

estimated from the pulsation does not provide a full description of the mechanical properties of the arterial wall.

MATHEMATICAL EXPRESSION OF ARTERIAL ELASTICITY

Uniaxial Tensile Behavior

There are many tensile test data from arterial walls in humans and animals (4). Arterial walls exhibit nonlinear force-deformation or stress–strain behavior, having higher distensibility in the low force or stress range and losing it at higher force or stress. To represent strain in such biological soft tissues that deform largely and nonlinearly, we commonly use extension ratio, λ, which is defined by the ratio of the current length of a specimen (L) to its initial length (L_0). If we plot a stress/extension ratio curve as the slope of a stress/extension ratio curve versus stress, we can see that the relation is composed of one or two straight lines (1). Each line is described by

$$dT/d\lambda = BT + C \tag{1}$$

where T is Lagrangian stress defined by F/A_0 (F, force; A_0, cross-sectional area of an undeformed specimen), and B and C are constants. This is also expressed by

$$T = A[\exp B(\lambda - 1) - 1] \tag{2}$$

where A is equal to C/B. This type of exponential formulation is applicable to the description of the elastic behavior of many other biological soft tissues (5).

Pressure–Diameter Relations

For practical purposes, it is convenient to use a single parameter that expresses the arterial elasticity under living conditions. In particular, for noninvasive diagnosis in clinical medicine, material characterization should be simple, yet quantitative. For this purpose, several parameters have been proposed and commonly utilized (1). These include pressure–strain elastic modulus, E_p and vascular compliance, C_v. Pulse wave velocity, c, which was mentioned above, is also used to express elastic properties of the arterial wall. These parameters are described by

$$E_p = \Delta P/(\Delta D_o/D_o) \tag{3}$$

$$C_v = (\Delta V/V)/\Delta P \tag{4}$$

and

$$c^2 = (S/\rho)(\Delta P/\Delta S) = (V/\rho)(\Delta P/\Delta V) \tag{5}$$

where D_o, V, and S are the outer diameter, volume, and luminal area of a blood vessel at pressure P, respectively, and ΔD_o, ΔV, and ΔS are their increments for the pressure increment, ΔP. The parameter ρ is the density of the blood.

To calculate these parameters, we do not need to measure the wall thickness; for E_p and C_v, we need to know only pressure–diameter and pressure–volume data, respectively, at a specific pressure level. However, we should remember that these parameters express the stiffness or distensibility of a blood vessel. Therefore, they are structural parameters, and do not rigorously represent the inherent elastic properties of the wall material; in this sense, they are different from the elastic modulus which is explained below. In addition, these parameters are defined at specific pressures, and give different values at different pressure levels because the pressure–diameter relations of arteries are nonlinear.

To overcome this shortcoming, several functions have been proposed to mathematically describe pressure–diameter, pressure–volume, and pressure-luminal area data, and one or several parameters included in these equations have been used for the expression of the elastic characteristics of arteries. Among these functions, the following equation is one of the simplest and most reliable for the description of pressure-diameter relations of arteries in the physiological pressure range (6):

$$\ln(P/P_s) = \beta(D_o/D_s - 1) \tag{6}$$

where P_s is a standard pressure and D_s is the wall diameter at pressure P_s. A physiologically normal blood pressure like 100 mmHg (13.3 kPa) is recommended for the standard pressure, P_s. As an example, Fig. 2 shows the pressure–diameter relationships of a human femoral artery under normal and active conditions of vascular smooth muscle and the relations between the logarithm of pressure ratio, P/P_s, and distension ratio, D_o/D_s. Figure 2a demonstrates nonlinear behavior of the artery under both conditions, while Fig. 2b shows the close fit of the data to Eq. 6 over a rather wide pressure range. The coefficient, β, called the stiffness parameter, represents the structural stiffness of a vascular wall; it does not depend upon pressure. This parameter has been used for the evaluation of the stiffness of arteries not only in basic investigations, but also in clinical studies (1).

As can be seen from Fig. 2a, under the normal condition, arteries greatly increase the diameter with pressure under a low pressure range, say < 60 mmHg (8 kPa), and then gradually lose the distensibility at higher pressures. When vascular smooth muscle cells are activated by stimuli, arteries are contracted and their diameter decreases in a

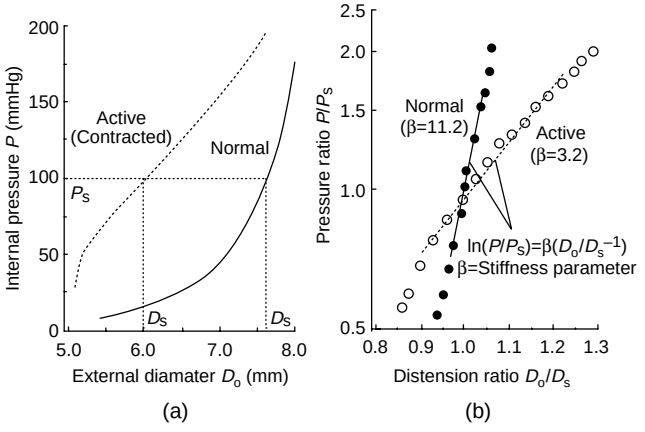


Figure 2. Pressure–diameter (a) and pressure ratio–distension ratio (b) relations of a human femoral artery under normal and active conditions (*in vitro* study) (1,7).

physiological pressure range and below the range [< 200 mmHg (26.6 kPa) in Fig. 2], and their pressure–diameter curves become greatly different from those observed under the normal condition.

To express the elastic properties of wall material, it is necessary to use a material parameter such as elastic modulus or Young's modulus, which is the slope of a linear stress–strain relation. For arterial walls that have nonlinear stress–strain relations, the following incremental elastic modulus has been often used for this purpose (8):

$$H_{\theta\theta} = 2D_i^2 D_o(\Delta P/\Delta D_o)/(D_o^2 - D_i^2) + 2PD_o^2/(D_o^2 - D_i^2) \quad (7)$$

where D_i is the internal diameter of a vessel. This equation was derived using the theory of small elastic deformation superposed on finite deformation in the case of a pressurized orthotropic cylindrical tube.

To calculate this modulus, it is necessary to know the thickness or internal diameter of a vessel. In *in vitro* experiments, we can calculate them from D_o, the internal and external diameters under no-load conditions measured after pressure–diameter testing, the *in vivo* axial extension ratio, and assuming the incompressibility of wall material. Noninvasive measurement of wall thickness or internal diameter on intact vessels has been rather difficult compared with the measurement of external diameter; however, it is now possible with high accuracy ultrasonic echo systems as mentioned above.

Constitutive Laws

Mathematical description of the mechanical behavior of a material in a general form is called a constitutive law or constitutive equation. We cannot perform any mechanical analyses without knowledge of constitutive laws of materials. Strain energy functions are commonly utilized for formulating constitutive laws of biological soft tissues that undergo large deformation (5). Let W be the strain energy per unit mass of a tissue, and ρ_0 be the density in the zero-stress state. Then, $\rho_0 W$ is the strain energy per unit volume of the tissue in the zero-stress state, and this is called the strain energy density function. Because arterial tissue is considered as an elastic solid, a strain energy function exists, and the strain energy W is a function solely of the Green strains:

$$W = W(E_{ij}) \quad (8)$$

where E_{ij} are the components of the Green strain tensor with respect to a local rectangular Cartesian coordinate system.

Under physiological conditions, arteries are subjected to axisymmetrical loads, and the axes of the principal stresses and strains coincide with the axes of mechanical orthotropy. Moreover, the condition of incompressibility is used to eliminate the radial strain E_{rr}, and therefore the strain energy function becomes a function of the circumferential and axial strains $E_{\theta\theta}$ and E_{zz}. Then, the constitutive equations for arteries are

$$\sigma_{\theta\theta} - \sigma_{rr} = (1 + 2E_{\theta\theta})\partial(\rho_0 W)/\partial E_{\theta\theta} \quad (9)$$

and

$$\sigma_{zz} - \sigma_{rr} = (1 + 2E_{zz})\partial(\rho_0 W)/\partial E_{zz} \quad (10)$$

where $\sigma_{\theta\theta}$, σ_{zz}, and σ_{rr} are Cauchy stresses in the circumferential, axial, and radial directions, respectively. Thus, we need to know the details of the strain energy function to describe stress–strain relations.

Three major equations have so far been proposed for the strain energy function of arterial wall. Vaishnav et al. (9) advocated the following equation:

$$\rho_0 W = (c/2)\exp(b_1 E_{rr}^2 + b_2 E_{\theta\theta}^2 + b_3 E_{zz}^2 + 2b_4 E_{rr}E_{\theta\theta} + 2b_5 E_{\theta\theta}E_{zz} + 2b_6 E_{zz}E_{rr}) \quad (11)$$

where $E_{\theta\theta}$ and E_{zz} are Green strains in the circumferential and axial directions, respectively, and A, B, and so on, are constants.

Chuong and Fung (10) proposed another form with an exponential function:

$$\rho_0 W = (c/2)\exp(b_1 E_{rr}^2 + b_2 E_{\theta\theta}^2 + b_3 E_{zz}^2 + 2b_4 E_{rr}E_{\theta\theta} + 2b_5 E_{\theta\theta}E_{zz} + 2b_6 E_{zz}E_{rr}) \quad (12)$$

where c, b_1, b_2, and so on, are material constants.

Later, Takamizawa and Hayashi (11) proposed a logarithmic form of the function described by

$$\rho_0 W = -C\ln(1 - a_{\theta\theta}E_{\theta\theta}^2/2 - a_{zz}E_{zz}^2/2 - a_{\theta z}E_{\theta\theta}E_{zz}) \quad (13)$$

where C, $a_{\theta\theta}$, a_{zz}, and $a_{\theta z}$ characterize the elastic properties of a material.

By using one of these strain energy equations or another type of equation for W in Eqs. 9 and 10, and applying the equations of equilibrium and boundary conditions, we determine the values of material constants. Although all of the proposed formulations describe quite well the elastic behavior of arterial walls, we prefer to reduce the number of constants included in the equations in order to handle them more easily, as well as to give physical meanings to the constants. For this reason, the logarithmic expression (Eq. 13) may be advantageous.

ELASTIC PROPERTIES OF NORMAL ARTERIES

Figure 3 shows β values of common carotid arteries, intracranial vertebral arteries, and coronary arteries obtained from autopsied human subjects of different ages (7). Note that arterial stiffness is much larger in the coronary arteries than in the other arteries, and also that intracranial vertebral arteries are considerably stiffer than extracranially located common carotid arteries. As can be seen from this figure, almost all the data obtained from normal human aortas and conduit arteries show that the structural stiffness of wall (e.g., E_p and β) increases with age rather gradually until the age of 40 years, and rapidly thereafter; on the other hand, the wall compliance (C_v) decreases with age. The stiffness of intracranial arteries like the intracranial vertebral artery progressively increases until 20 years, and then more slowly (6). There seems to be almost no age-related change in the human coronary artery (12).

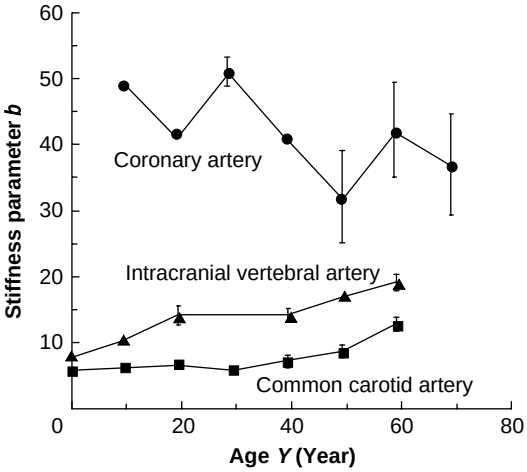


Figure 3. Age-related changes in the wall stiffness of human arteries. (*In vitro* studies, reproduced from Ref. 7.)

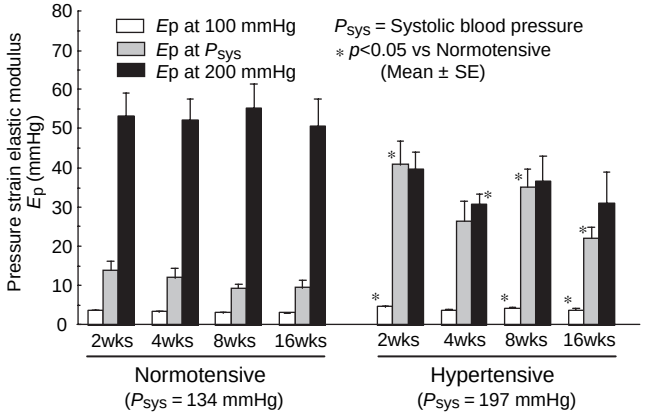


Figure 4. Structural stiffness of thoracic arotas in normotensive and hypertensive rats (7).

Most of the data obtained from arteries in animals indicated that the incremental elastic modulus or the slope of stress–strain curve increases with age, although there are several opposite data (13).

ELASTIC PROPERTIES OF DISEASED ARTERIES

Hypertension

Hypertension is recognized as one of the important risk factors for many cardiovascular diseases, including atherosclerosis and stroke. Elevated blood pressure exerts influences on the synthetic activity of vascular smooth muscle cells, and is believed to induce changes in structure and morphology of the arterial wall, its mechanical properties, and vascular contractility. It is therefore very important to understand arterial mechanics in hypertension. However, results from the extensive literature concerning the elastic properties of hypertensive arteries are contradictory and inconclusive (1,7,13,14). As mentioned above, when we analyze the reported data, we should remember that the values of such parameters as E_p (Eq. 3) and C_v (Eq. 4) are dependent on pressure. Without this consideration, comparisons between the results from different studies have little meaning.

Several studies have been performed to determine the pressure–diameter or pressure–volume relationships of aortas and arteries in hypertensive animals and humans. For example, comparison of hypertensive rats to normotensive controls has shown that at 100 mmHg (13.3 kPa) and also at the working pressure (systolic blood pressure before sacrifice, P_{sys}) of each group, the pressure strain elastic modulus, E_p of the thoracic aorta is greater in hypertensives than in normals; whereas at 200 mmHg (26.6 kPa) the E_p values in the hypertensive animals are slightly lower than those of the normals (Fig. 4) (7). These results do not depend on the duration of hypertension for 2–16 weeks.

With regard to the inherent elastic modulus of wall material calculated from pressure–diameter data, it has been shown that the incremental elastic moduli of the rat thoracic aorta ($H_{\theta\theta}$ in Eq. 7) at systolic blood pressure levels have significant correlations with blood pressure until 8 weeks after the induction of hypertension; at 16 weeks, however, the correlation disappears and the elastic modulus tends to be at the same level as that in control, normotensive rats (Fig. 5) (7). There are no significant differences in the incremental elastic modulus at

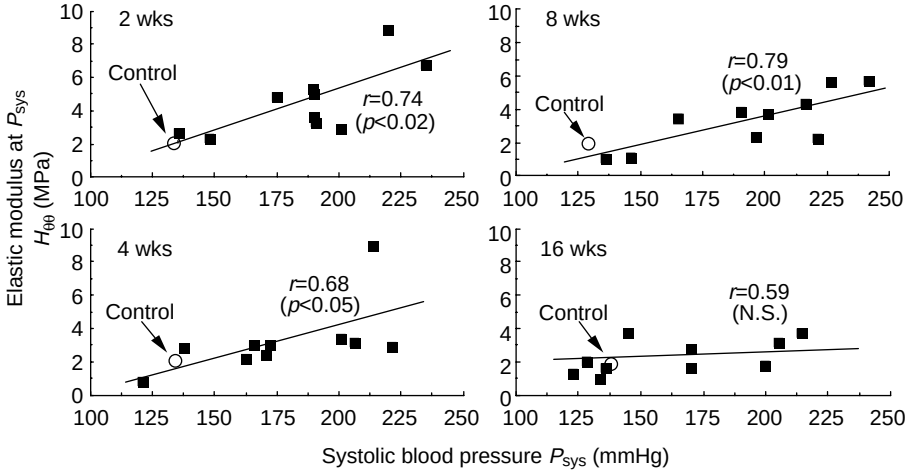


Figure 5. Incremental elastic modulus versus blood pressure in the rat aorta at 2, 4, 8, and 16 weeks after the treatment for hypertension (7).

100 mmHg (13.3 kPa) regardless of the period of hypertension. The aortic wall in hypertensive rats seems to restore the *in vivo* elastic properties to a normal level in 16 weeks due to the functional adaptation and remodeling of the wall.

In connection with the elastic properties of wall, many data have shown that the arterial wall is thickened by hypertension and hypertrophy occurs (14). Wall thickness critically depends on pressure level and, therefore, we have to pay attention to the pressure for the thickness measurement when we interpret the results. If the thickness of wall at the *in vivo* blood pressure level is used to calculate *in vivo* wall stress in the circumferential direction (hoop stress), the stress is independent of the degree of hypertension and is always maintained at a control, normal level even at 2 weeks after the induction of hypertension. This phenomenon is attributable to a functional adaptation and remodeling of the arterial wall (15).

Atherosclerosis

Effects of flow dynamics and wall shear stress on the initiation and development of atherosclerosis have been studied extensively. However, less attention has been paid to the mechanical properties of atherosclerotic wall tissue. Does atherosclerosis stiffen the arterial wall or increase the elastic modulus of wall? The results obtained have been conflicting and inconclusive, as shown in Table 1 (7). One of the reasons for this is that the structural stiffness of arterial wall and the elasticity of wall material have been confusingly used for the expression of the elastic properties of atherosclerotic wall. In this table, the elastic modulus represents the elasticity of wall material, which corresponds to the slope of stress–strain curve and is given by, for example, the incremental elastic modulus, $H_{\theta\theta}$; the stiffness is the structural stiffness expressed by, for example, the stiffness parameter, β.

Several studies have shown that the arterial wall is stiffened by the development of atherosclerosis. However, others have presented different results. Thus it has not been clear whether atherosclerosis increases the elastic modulus of arterial wall. We can see from Table 1 that atherosclerosis is mostly accompanied by wall thickening. This might be a reason why there are no data indicating a decrease in the structural stiffness associated with atherosclerosis. The structural stiffness is determined not only by the elastic modulus of wall material, but also by wall dimensions such as wall thickness.

A detailed and systematic study on the mechanical properties and morphology of atherosclerotic aortas in the rabbit has shown that the changes in the wall stiffness (β) and the elastic modulus ($H_{\theta\theta}$) are not always correlated

Table 1. Distributions of Reported Data of the Elasticity, Stiffness, and Thickness of Atherosclerotic Wall[a]

	No. of Data	Increase, %	No Change, %	Decrease, %
Elastic modulus	17	29	47	24
Stiffness	17	71	29	0
Wall thickness	12	67	25	8

[a]See Ref. 7

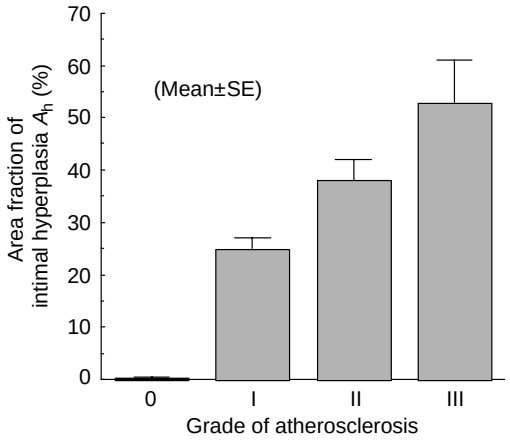


Figure 6. Area fraction of intimal hyperplasia in atherosclerotic thoracic aortas in the rabbit (7). Atherosclerosis was induced by the combination of denudation of endothelial cells and cholesterol diet.

with the time of cholesterol diet feeding (16). Thus, the grade of atherosclerosis was defined from the percent fraction of the luminal surface area stained with Sudan IV as well as from wall stiffening. The area fraction of intimal hyperplasia increases with the grade (Fig. 6). Likewise, wall thickness steadily increases with the progression

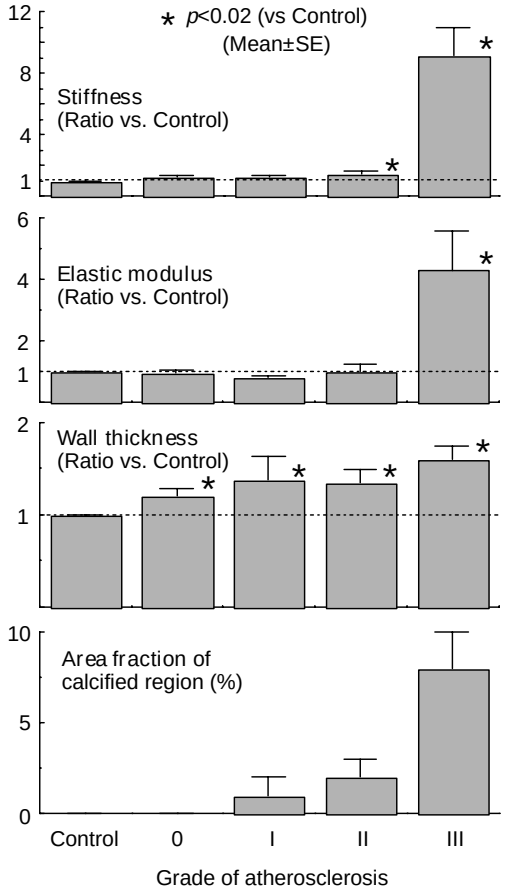


Figure 7. Elasticity, wall thickness, and calcification of athorosclerotic thoracic aortas in the rabbit (7).

of atherosclerosis (Fig. 7). However, the elastic modulus is not significantly different from the control artery until the highest grade of atherosclerosis. On the other hand, there appears a significant increase in the arterial stiffness in the grade II atherosclerosis, which is attributable to the wall thickening. Significantly increased calcification and intimal hyperplasia are observed in the wall of the grade III atherosclerosis. From these results, it is concluded that the progression of atherosclerosis induces wall thickening, followed by wall stiffening. However, even if atherosclerosis is advanced, there is essentially no change in the elastic modulus of wall material unless considerable calcification occurs in the wall. Calcified aortas have high elastic moduli. At the most advanced stage of atherosclerosis, the arterial wall has high structural and material stiffness due to calcification and wall hypertrophy (16).

BIBLIOGRAPHY

Cited References

1. Hayashi K, et al. Techniques in the Determination of the Mechanical Properties and Constitutive Laws of Arterial Walls. In: Leondes C, editor. Cardiovascular Techniques – Biomechanical Systems: Techniques and Applications Vol. II. Boca Raton (FL): CRC Press; 2001. p 6-1–6-61.
2. Hayashi K, Mori K, Miyazaki H. Biomechanical response of femoral vein to chronic elevation of blood pressure in rabbits. Am J Physiol 2003;284:H511–H518.
3. Hayashi K, Nakamura T. Implantable miniature ultrasonic sensors for the measurement of blood flow. Automedica 1989;12:53–62.
4. Abe H, Hayashi K, Sato M, editors. Data Book on Mechanical Properties of Living Cells, Tissues, and Organs. Tokyo: Springer-Verlag; 1996. p 25–125.
5. Fung YC. Biomechanics: Mechanical Properties of Living Tissues. New York: Springer-Verlag; 1993. p 242–320.
6. Hayashi K, et al. Stiffness and elastic behavior of human intracranial and extracranial arteries. J Biomech 1980;13: 175–184.
7. Hayashi K. Mechanical Properties of Soft Tissues and Arterial Walls. In: Holzapfel G, Ogden RW, editors. Biomechanics of Soft Tissue in Cardiovascular Systems. (CISM Courses and Lectures No. 441), Wien: Springer-Verlag; 2003. p 15–64.
8. Hudetz AG. Incremental elastic modulus for orthotropic incompressible arteries. J Biomech 1979;12:651–655.
9. Vaishnav RN, Young JT, Patel DJ. Distribution of stresses an strain energy density through the wall thickness in a canine aortic segment. Circ Res 1973;32:577–583.
10. Chuong CJ, Fung YC. Three-dimensional stress distribution in arteries. Trans ASME J Biomech Eng 1983;105:268–274.
11. Takamizawa K, Hayashi K. Strain energy density function and uniform strain hypothesis for arterial mechanics. J Biomech 1987;20:7–17.
12. Hayashi K, Igarashi Y, Takamizawa K. Mechanical properties and hemodynamics in coronary arteries. In: Kitamura K, Abe H, Sagawa K, editors. New Approaches in Cardiac Mechanics. New York: Gordon and Breach; 1986. p 285–294.
13. Hayashi K. Experimental approaches on measuring the mechanical properties and constitutive laws of arterial walls. Trans ASME J Biomech Eng 1993;115:481–488.
14. Humphrey JD. Cardiovascular Solid Mechanics: Cells, Tissues, and Organs. New York: Springer-Verlag; 2002. p 365–386.
15. Matsumoto T, Hayashi K. Response of arterial wall to hypertension and residual stress. In: Hayashi K, Kamiya A, Ono K, editors. Biomechanics: Functional Adaptation and Remodeling. Tokyo: Springer-Verlag; 1996. p 93–119.
16. Hayashi K, Ide K, Matsumoto T. Aortic walls in atherosclerotic rabbits -Mechanical study. Trans ASME J Biomech Eng 1994;116:284–293.

See also BLOOD PRESSURE, AUTOMATIC CONTROL OF; BLOOD RHEOLOGY; CUTANEOUS BLOOD FLOW, DOPPLER MEASUREMENT OF; HEMODYNAMICS; INTRAAORTIC BALLOON PUMP; TONOMETRY, ARTERIAL.

ASSISTIVE DEVICES FOR THE DISABLED.

See ENVIRONMENTAL CONTROL.

ATOMIC ABSORPTION SPECTROMETRY.

See FLAME ATOMIC EMISSION SPECTROMETRY AND ATOMIC ABSORPTION SPECTROMETRY.

AUDIOMETRY

THOMAS E. BORTON
BETTIE B. BORTON
Auburn University Montgomery
Montgomery, Alabama

JUDITH T. BLUMSACK
Disorders Auburn University
Auburn, Alabama

INTRODUCTION

Audiology is, literally, the science of hearing. In many countries around the world, audiology is a scientific discipline practiced by audiologists. According to the American Academy of Audiology, "an audiologist is a person who, by virtue of academic degree, clinical training, and license to practice and/or professional credential, is uniquely qualified to provide a comprehensive array of professional services related to the prevention of hearing loss and the audiologic identification, assessment, diagnosis, and treatment of persons with impairment of auditory and vestibular function, as well as the prevention of impairments associated with them. Audiologists serve in a number of roles including clinician, therapist, teacher, consultant, researcher and administrator" (1).

An important tool in the practice of audiology is audiometry, which is the measurement of hearing. In general, audiometry entails one or more procedures wherein precisely defined auditory stimuli are presented to the listener in order to elicit a measurable behavioral or physiologic response. Frequently, the term audiometry refers to procedures used in the assessment of an individual's threshold of hearing for sinusoidal (pure tones) or speech stimuli (2). So-called conventional audiometry is conducted with a calibrated piece of electronic instrumentation called an audiometer to deliver controlled auditory signals to a listener. Currently, however, an expanded definition of audiometry also includes procedures for measuring various physiological and behavioral responses to the presentation

of auditory stimuli, whether or not the response involves cognition. More sophisticated procedures and equipment are increasingly used to look beyond peripheral auditory structures in order to assess sound processing activity in the neuroauditory system.

Today, audiologists employ audiometric procedures and equipment to assess the function of the auditory system from external ear to brain cortex and serve as consultants to medical practitioners, education systems, the corporate and legal sectors, and government institutions such as the Department of Veterans Affairs. Audiologists also use audiometric procedures to identify and define auditory system function as a basis for nonmedical intervention with newborns, young children with auditory processing disorders, and adults who may require sophisticated amplification systems to develop or maintain their communication abilities and quality of life.

The purpose of this chapter is to acquaint the reader with the basic anatomy of the auditory system, describe some of the instrumentation and procedures currently used for audiometry, and briefly discuss the application of audiometric procedures for the assessment of hearing.

AUDIOMETRY AND ITS ORIGINS

Audiometry refers broadly to qualitative and quantitative measures of auditory function/dysfunction, often with an emphasis on the assessment of hearing loss. It is an important tool in the practice of audiology, a healthcare specialty concerned with the study of hearing, and the functional assessment, diagnosis, and (re)habilitation of hearing impairment. The profession sprang from otology and speech pathology in the 1920s, about the same time that instrumentation for audiometry was being developed. Audiometry grew rapidly in the 1940s when World War II veterans returned home with hearing impairment related to military service (3). Hearing evaluation, the provision of hearing aids, and auditory rehabilitation were pioneered in the Department of Veterans Affairs and, subsequently, universities began programs to educate and train audiologists for service to children and adults in clinics, hospitals, research laboratories and academic settings, and private practice. Audiometry is now a fundamental component of assessing and treating persons with hearing impairment.

Audiometers

Audiometers are electroacoustic instruments designed to meet internationally accepted audiometric performance standards for valid and reliable assessment of hearing sensitivity and auditory processing capability under controlled acoustic conditions. The audiometer was first described around the turn of the twentieth century (4) and was used mainly in laboratory research at the University of Iowa. A laboratory assistant at the university, C. C. Bunch, would later publish a classic textbook describing audiometric test results in patients with a variety of hearing disorders (5). The first commercial audiometer, called the Western Electric 1A, was developed in the early 1920s by the Bell Telephone Laboratories in the United States, and was described by Fowler and Wegel in 1922 (6). More than 20 years elapsed before the use of audiometers for hearing assessment was widely recognized (7), and it was not until the early 1950s (8) that audiometry became an accepted clinical practice. Since that time, electroacoustic instrumentation for audiometry has been described in standards written by scientists and experts designed to introduce uniformity and facilitate the international exchange of data and test results. The American National Standards Institute (ANSI), the International Standards Organization (ISO), and the International Electrotechnical Commission (IEC) are recognized bodies that have developed accepted standards for equipment used in audiometry and in psychophysical measurement, acoustics, and research. In the present day, audiometric procedures are routinely used throughout the world for identifying auditory dysfunction in newborns, assessing hearing disorders associated with ear disease, and monitoring the hearing of patients at risk for damage to the auditory system (e.g., because of exposure to hazardous noise, toxic substances).

Types of Audiometers

In the United States, ANSI (9) classifies audiometers according to several criteria, including the use for which they are designed, how they are operated, the signals they produce, their portability, and other factors.

In general, Type IV screening audiometers (designed to differentiate those with normal hearing sensitivity from those with hearing impairment) are of simpler design than those instruments used for in-depth diagnostic evaluation for medical purposes (Type I audiometers). There are automatic and computer-processor audiometers (Bekesy or self-recording types), extended high-frequency (Type HF), free-field equivalent audiometers (Type E), speech audiometers (Types A, B, and C depending on available features), and others for specific purposes.

Audiometers possess one or more oscillation networks to generate pure tones of differing frequencies, switching networks to interrupt and direct stimuli, and attenuating systems calibrated in decibels (dB) relative to audiometric zero. The intensity range for most attenuation networks usually approximates 100 dB, typically graduated in steps of 5 dB. The "zero dB" level represents normal hearing sensitivity across the test frequency range for young adults under favorable, noise-free laboratory conditions. Collection of these hearing level reference data from different countries began in the 1950s and 1960s. These reference levels have been accepted by internationally recognized standards organizations. Audiometers also include various types of output transducers for presentation of signals to the listener, including earphones, bone-conduction vibrators, and loudspeakers. As the audiometrically generated signals are affected substantially by the electroacoustic characteristics of these devices (e.g., frequency response), versatile audiometers have multiple calibrated output networks to facilitate switching between transducers depending on the clinical application of interest.

Figure 1 shows the general layout of a Type I diagnostic audiometer. Such instruments are required by standards governing them to produce a stable output at a range of

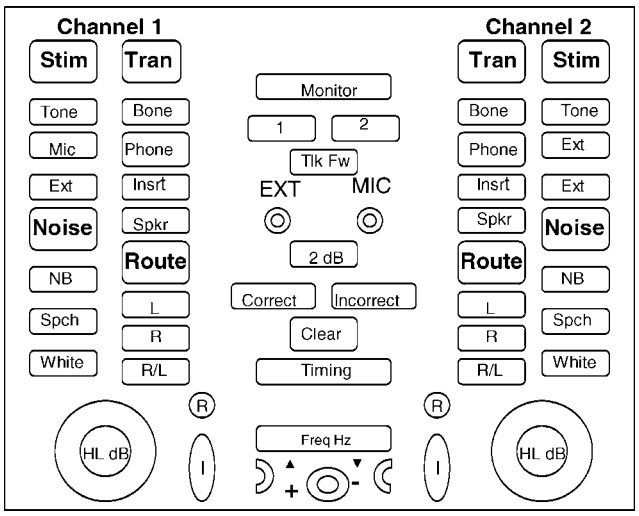


Figure 1. Typical diagnostic audiometer.

operating temperatures and humidity and meet a wide variety of electrical and other safety standards, in addition to precise electroacoustic standards for frequency, intensity, spectral purity, maximum output sound pressure level (SPL), and harmonic distortion. Type II audiometers have fewer required features and less flexibility, and Type IV audiometers have even more limitations.

Audiometric Calibration

To ensure that an audiometer is performing in accordance with the relevant standard, the instrument's electroacoustic characteristics are checked and adjusted as necessary, usually following a routine procedure. These calibration activities may be conducted at the manufacturing facility or an outside laboratory, but are most often accomplished on site at least annually. Calibration of speakers in a sound field is typically conducted on site because the unique acoustic characteristics of a specific field cannot easily be reproduced in a remote calibration facility.

Calibration of audiometers is routinely checked using instruments such as oscilloscopes, multimeters, spectrometers, and sound-level meters to verify frequency, intensity, and temporal characteristics of the equipment. Output transducers such as earphones and bone stimulators can be calibrated in two ways: (*1*) using "real ear" methods, involving individuals or groups of persons free from ear pathology and who meet other criteria, or (*2*) using hard-walled couplers (artificial ears) and pressure transducers specified by the relevant standard.

Audiometric Standards

Electroacoustic instrumentation for audiometry has been described in national and international standards written by scientists and experts designed to introduce uniformity and facilitate the international exchange of data and test results. ANSI, ISO, and IEC are recognized bodies that have developed accepted standards for equipment used in audiometry and acoustics. Some standards relate to equipment, others to audiometric procedures, and still others to the environment and conditions in which audiometry should be conducted (10).

The aim of standards for audiometric equipment and procedures is to assure precision of equipment functions to help ensure that test results can be interpreted meaningfully and reliably within and between clinics and laboratories using different equipment and personnel in various geographical locations. The results of audiometry often help to provide a basis for decisions regarding intervention strategies, such as medical or surgical intervention, referral for cochlear implantation, hearing aid selection and fitting, application of assistive listening devices, or selection of appropriate educational or vocational placement. As in any measurement scheme, audiometric test results can be no more precise than the function of the equipment and the procedures with which those measurements are made.

PURE TONE AUDIOMETRY

Psychophysical Methods

Audiometry may be conducted with a variety of methodologies depending on the goal of the procedure and subject variables such as age, mental status, and motivation. For example, the hearing sensitivity of very young children may be estimated by assessing the effects of auditory stimuli presented in a sound field on startle-type reflexes, level of arousal, and localization. Patients who are developmentally delayed may be taught with reinforcement to push a button upon presentation of a test stimulus. Children of preschool age may be taught to make a motor response to auditory test stimuli using play audiometric techniques.

In conventional audiometry, auditory stimuli are presented through special insert or supra-aural earphones, or a bone oscillator worn by the patient. When indicated, a sound field around the listener may be created by presenting stimuli through strategically placed loudspeakers. Most threshold audiometric tests in school-aged children and adults can be conducted using one of two psychophysical methods originally developed by Gustav Fechner: (*1*) the method of adjustment, or (*2*) the method of limits (11). In the method of adjustment, the intensity of an auditory stimulus is adjusted by the listener according to some criterion (just audible or just inaudible), usually across a range of continuously or discretely adjusted frequencies. Nobel Prize Laureate Georg von Bekesy initially introduced this methodology into the practice of audiometry in 1947 (12). With this approach, listeners heard sinusoidal stimuli that changed from lower to higher test frequencies, and adjusted the intensity of continuous and interrupted tones from "just inaudible" to "just audible". As shown in Fig. 2, this methodology yielded information about the listener's auditory threshold throughout the test frequency range. The relationship of threshold tracings for the pulsed and continuous stimuli added additional information about the site of lesion causing the hearing loss (13,14).

Later, it was found that the tracing patterns tracked by hearing-impaired patients at their most comfortable loudness levels, instead of their threshold levels, yielded additional useful diagnostic information (15). A myriad of

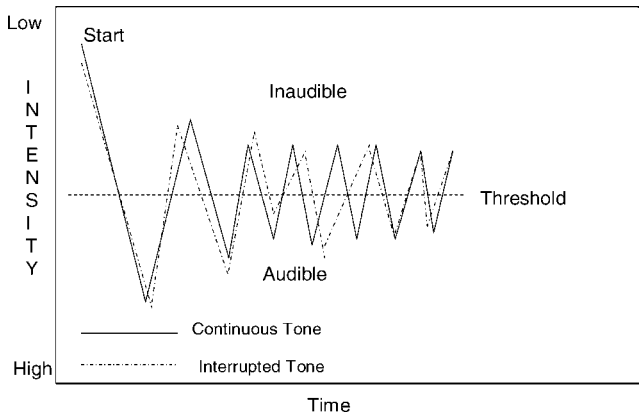


Figure 2. Bekesy audiometric tracing of continuous and interrupted tones around auditory threshold.

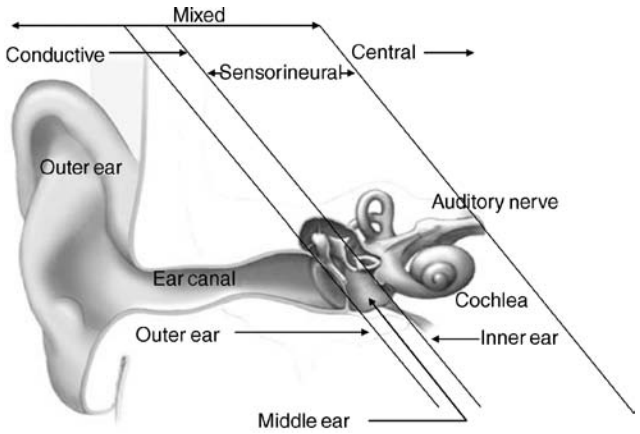


Figure 3. Anatomy of the peripheral auditory mechanism. Adapted from medical illustrations by NIH, Medical Arts & Photography Branch.

factors associated with this psychophysical measurement method, including the age of the listener, learning effects, and the length of time required to obtain stable test results, make these methods unsuitable for routine diagnostic purposes, especially in young children. Nevertheless, audiometers incorporating this methodology are manufactured to precise specifications (9) and are routinely used in hearing conservation programs to record the auditory sensitivity of large numbers of employees working in industrial or military settings.

In most clinical situations today, routine threshold audiometry is conducted using the method of limits. In this approach, the examiner adjusts the intensity of the auditory stimuli of various frequencies according to a predetermined schema, and the listener responds with a gross motor act (such as pushing a button or raising a hand) when the stimulus is detected. Although auditory threshold may be estimated using a variety of procedural variants (ascending, descending, mixed, adaptive), research has established (16) that an ascending approach in which tonal stimuli are presented to the listener from inaudible intensity to a just audible level is a valid and reliable approach for cooperative and motivated listeners, and the technique most parsimonious with clinical time and effort. In this approach, tonal stimuli are presented at intensity levels below the listener's threshold of audibility and raised in increments until a response is obtained. At this point, the intensity is lowered below the response level and increased incrementally until a response is obtained. When the method of limits is used, auditory threshold is typically defined as the lowest intensity level that elicits a reliable response from the patient on approximately 50% or more of these "ascending" trials.

Sound Pathways of the Auditory System

The fundamental anatomy of the ear is depicted in Fig. 3. Sound enters the auditory mechanism by two main routes, air conduction and bone conduction. Most speech and other sounds in the ambient environment enter the ear by air conduction. The outer ear collects and funnels sound waves into the ear canal, provides a small amount of amplification to auditory signals, and conducts sound to the tympanic membrane (eardrum). Acoustic energy strikes the tympanic membrane, where it is converted to mechanical energy in the form of vibrations to be conducted by small bones across the middle ear space to the inner ear. These mechanical vibrations are then converted to hydraulic energy in the fluid-filled inner ear (cochlea). This hydrodynamic form of energy results in traveling waves on cochlear membranous tissues. Small sensory hair cells are triggered by these waves to release neurotransmitters, resulting in the production of neural action potentials that are conducted through the auditory nerve (N. VIII) via central auditory structures in the brainstem to the auditory cortex of the brain, where sound is experienced.

Disorders affecting different sections of the ear depicted in Fig. 3 produce different types of hearing impairment. The outer ear, external auditory canal, and ossicles of the middle ear are collectively considered as the conductive system of the ear, and disorders affecting these structures produce a conductive loss of hearing. For example, perforation of the tympanic membrane, presence of fluid (effusion) in the middle ear due to infection, and the disarticulation of one or more bones in the middle ear all produce conductive hearing loss. This type of hearing loss is characterized by attenuation of sounds transmitted to the inner ear, and medical/surgical treatment often fully restores hearing. In a small percentage of cases, the conductive disorder may be permanent, but the use of a hearing aid or other amplification device can deliver an adequate signal to the inner ear that usually permits excellent auditory communication.

The inner ear and auditory nerve comprise the sensorineural mechanism of the ear, and a disorder of this apparatus often results in a permanent sensorineural hearing loss. Sensorineural disorders impair both perceived sound audibility and sound quality typically because of impaired frequency selectivity and other effects. Thus, in sensorineural-type impairments, sounds become difficult to detect, and they are also unclear, leading to poor understanding of speech. In some cases, conductive and sensorineural disorders simultaneously co-exist to produce

a mixed-type hearing impairment. Listeners with this disorder experience the effects of conductive and sensorineural deficits in combination.

Finally, the central auditory system begins at the point the auditory nerve enters the brainstem, and comprises the central nerve tracts and nuclear centers from the lower brainstem to the auditory cortex of the brain. Disorders of the central auditory nervous system produce deficits in the ability to adequately process auditory signals transmitted from the outer, middle, and inner ears. The resulting hearing impairment is characterized not by a loss of sensitivity to sound, but rather difficulties in identifying, decoding, and analyzing auditory signals, especially in difficult listening environments with background noise present. Auditory processing disorders require sophisticated test paradigms to identify and diagnose.

The Audiogram

The results of basic audiometry may be displayed in numeric form or on a graph called an audiogram, as shown in Fig. 4. As can be seen, frequency in hertz (Hz) is depicted on the abscissa, and hearing level (HL) in dB is displayed on the ordinate. Although the normal human ear can detect frequencies below 100 Hz and as high as 20,000 Hz, the audible frequency range most important for human communication lies between 125 and 8000 Hz, and the audiogram usually depicts this more restricted range. For special diagnostic purposes, extended high frequency audiometers produce stimuli between 8000 and 20,000 Hz, but specialty audiometers and earphones must be used to obtain thresholds at these frequencies. A few commercially available audiometers produce sound pressure levels as high as 120 dB HL, but such levels are potentially hazardous to the human ear and hearing thresholds poorer than 110 dB do not represent "useful" hearing for purposes of communication.

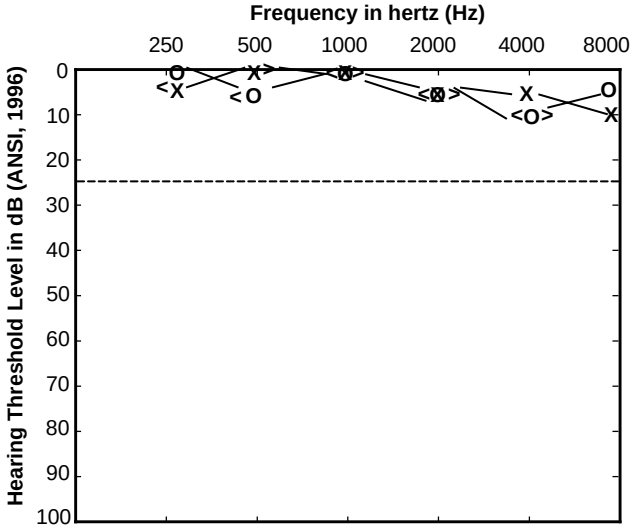


Figure 4. Graphic audiogram for a normal hearing listener. Bone conduction, right ear = <; Bone conduction, left ear = >; Air conduction, right ear = O; Air conduction, left ear = X.

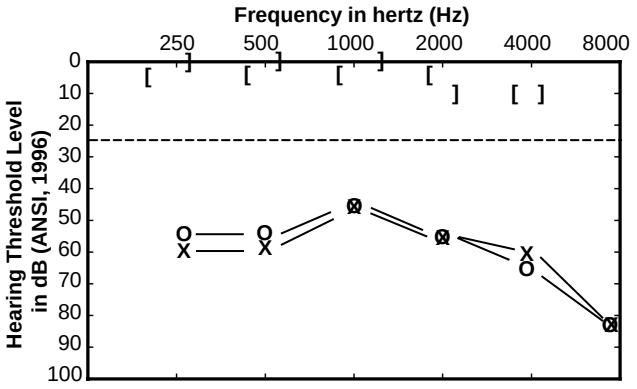


Figure 5. Graphic audiogram for a listener with conductive hearing loss. Bone conduction, right ear = [; Bone conduction, left ear =]; Air conduction, right ear = O; Air conduction, left ear = X.

The dashed line across the audiogram in Fig. 4 at 25 dB HL represents a common depiction of the boundary between normal hearing levels and the region of hearing loss (below the line) in adults. The recorded findings on this audiogram represent normal test results from an individual with no measurable loss of hearing sensitivity.

Figure 5 displays test results for a listener with a middle ear disorder in both ears and a bilateral conductive loss of hearing, which is moderate in degree, and similar in each ear. Bone conduction responses for the two ears are within normal limits (between 0 and 25 dB HL), suggesting normal sensitivity of the inner ear and auditory nerve, while air conduction thresholds are depressed below normal, suggesting obstruction of the air conduction pathway to the inner ear. Thus, conductive hearing losses are characterized on the audiogram by normal bone conduction responses and depressed air conduction responses. In

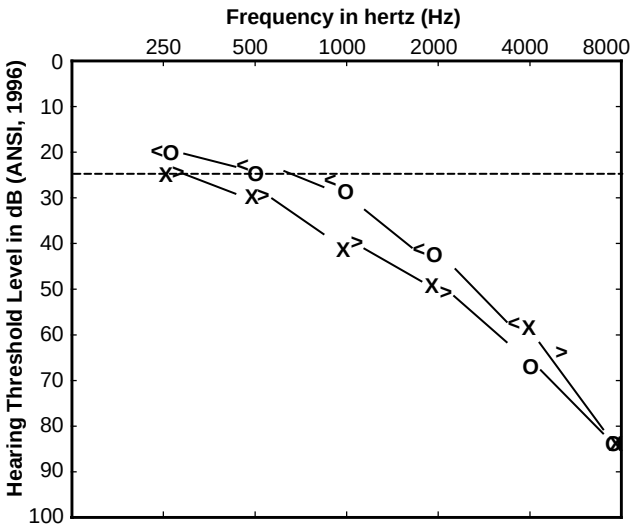


Figure 6. Graphic audiogram. Bone conduction, right ear = <; Bone conduction, left ear = >; Air conduction, right ear = O; Air conduction, left ear = X.

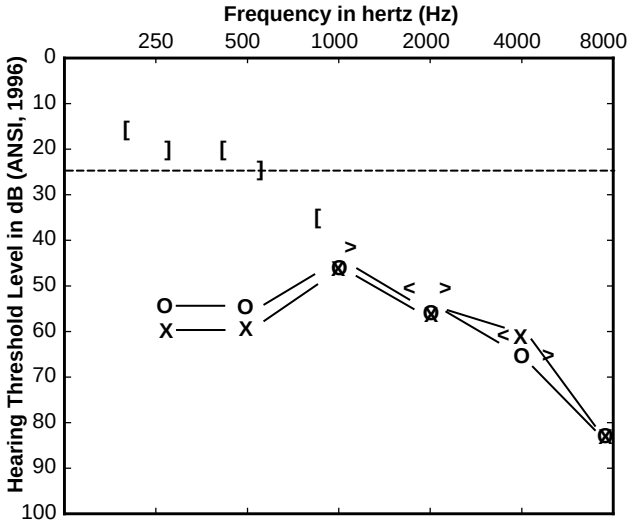


Figure 7. Graphic audiogram for a listener with mixed hearing loss. Masked bone conduction, right ear = [; Masked bone conduction, left ear =]; Unmasked bone conduction, right ear = <; Unmaksked bone conduction, left ear = >; Air conduction, right ear = O; Air conduction, left ear = X.

sensorineural-type hearing losses, air conduction and bone conduction responses in each ear are equally depressed on the audiogram. Figure 6 shows a high frequency loss of hearing in both ears, falling in pattern and sensorineural in type. Air and bone conduction hearing sensitivity is similar in both ears, suggesting that the cause of the hearing loss is not in the conductive mechanism (outer and middle ears).

The audiogram shown in Fig. 7 depicts a mixed-type loss of hearing in both ears. The gap between air and bone conduction thresholds in the two ears at the lower frequencies suggests a conductive disorder affecting the outer or middle ears. However, at frequencies above 500 Hz, hearing sensitivity via both air and bone conduction pathways in the two ears is nearly identical, which points to a disorder affecting the inner ear or auditory nerve.

In summary, an audiogram displays the results of basic audiometry in a stylized "shorthand", so that the hearing impairment can be readily characterized according to type of loss, degree of deficit, configuration (shape) of loss, and the degree of symmetry between the two ears. Such findings constitute the basis for first-order description of a listener's hearing sensitivity across the audible frequency range and provide important clues about the cause of hearing loss, the effects of the impairment on auditory communication ability, and the prognosis for treatment and rehabilitation.

SPEECH AUDIOMETRY

The first attempts to categorize hearing impairment on the basis of tests using speech stimuli were made in the early 1800s, when sounds were ranked according to their intensity and used to estimate the degree of hearing loss (17). Throughout the 1800s, refinements were introduced in methodologies for using speech stimuli to evaluate hearing. These improvements included the control of word intensities by varying distance between speaker and listener, the introduction of whispered speech to reduce differences in audibility between words, recording speech stimuli on the phonograph devised by Edison in 1877, and standardizing words lists in English and other languages (17). Most of the early research on speech perception focused on the sensitivity of the auditory system to speech, but progress in this area accelerated in the early 1900s because of work at the Bell Telephone Laboratories centered on the discrimination of speech sounds from each other. This emphasis led to the development of modern materials for assessing speech recognition at the Harvard Psychoacoustic Laboratories (18), which have been refined and augmented since that time.

Although pure tone audiometry provides important information about hearing sensitivity, as well as the degree, configuration, and type of hearing loss in each ear, it provides little information about a listener's auditory communication status and the ability to hear and understand speech in quiet as well as difficult listening situations. Attempts to predict speech recognition ability from the pure tone audiogram, even with normal hearing listeners, have met with only partial success, and the task is particularly complicated when listeners have a hearing impairment.

Instrumentation

Speech audiometry is conducted in the "speech mode" setting of a clinical audiometer. Speech stimuli are presented through the same types of transducers as those used for pure tone audiometry. Live speech stimuli via microphone and monitored with a VU meter can be used for speech audiometry, or recorded speech materials can be presented by CD or tape and routed through the audiometer to either one ear or both ears simultaneously by earphone or loudspeaker. Recorded speech materials typically include a calibration tone, and the input level is adjusted for individual recordings to a specified intensity level. Many different speech audiometric tests have been developed, and most currently in use are available in recorded form. Monitored live-voice presentation enables greater flexibility, but recorded speech materials enhance consistency across test conditions and avoid performance differences related to talker speech and vocal eccentricities.

In general, speech audiometry is conducted with the examiner in one room and the listener in another. With this arrangement, the examiner is able to observe the listener and maintain easy communication through microphones in both rooms, but the speech stimuli can be presented under carefully controlled conditions.

Speech Recognition Threshold

Speech recognition threshold (SRT) testing typically entails presentation of spondees (two-syllable, compound words), spoken with equal stress on each syllable (e.g., baseball, toothbrush, airplane). The use of these words for audiometric purposes has been investigated extensively, especially with respect to similarity in audibility (19).

Audiologists now generally select stimulus words from a list of commonly accepted spondees, and the words are presented at varying intensities using protocols similar to those used for pure tone audiometry. The speech recognition threshold (SRT) is the lowest intensity level at which the patient correctly responds to (repeats, writes down, points to) approximately 50% of the words, with the goal of determining the threshold of hearing for speech. The relationship between thresholds for speech and pure tone was identified in the early part of the twentieth century (20) and later described in detail (21,22). For purposes of clinical speech audiometry, speech recognition thresholds are expected to be within ± 6 dB of the average of the patient's pure tone air conduction thresholds at 500, 1000, and 2000. However, if the pure tone air conduction thresholds slope steeply, the speech recognition threshold is expected to agree with the average of the two best pure tone thresholds in the range of 500–2000 Hz.

The expected agreement between pure tone thresholds and speech recognition thresholds enables audiologists to use the SRT as a cross-check of pure tone air conduction threshold values. Disagreement between SRTs and pure tone threshold averages occurs for a variety of reasons. For example, poor agreement may exist between pure tone thresholds and SRTs in each ear if the patient misunderstands instructions regarding the test procedure for pure tone audiometry, or if the patient attempts to deceive the audiologist regarding actual hearing sensitivity.

SRTs can also be used to estimate/predict pure tone air conduction thresholds in the so-called speech frequency range of 500–2000 Hz in patients who are difficult to test with pure tones. Young children, for example, may reliably repeat or point to pictures of spondees (baseball, toothbrush) while exhibiting inconsistent responses to more abstract pure tones. Speech recognition thresholds have also been used as a basis for predetermining the presentation level for suprathreshold speech stimuli.

Speech Detection Threshold

Whereas the SRT represents the least intensity at which 50% of the speech stimuli presented to the listener can be recognized, the Speech Detection Threshold (SDT), sometimes called the Speech Awareness Threshold (SAT), represents the lowest intensity at which the patient exhibits an awareness of the presence of speech stimuli. If thresholds for spondaic words cannot be established, because of language impairment or other limitations such as young age or inability to speak because of injury, the SDT may represent a useful estimate of the level at which the patient indicates awareness of the presence of speech. In this type of speech threshold testing, the patient is not required to repeat the speech stimulus, which may be just a simple word or nonsense sound, but, instead, the patient simply responds with a hand movement or other gesture to indicate that a sound was detected. The SDT is obtained with protocols similar to those used for speech recognition measurement, and it is expected to be approximately 7–9 dB less intense than the value that would be obtained for the SRT (23,24).

Suprathreshold Speech Audiometry

In suprathreshold speech audiometry, speech stimuli (live-voice or recorded) are presented at levels well above the speech threshold in order to assess the listener's ability to understand speech. Depending on the purpose of the evaluation, the stimuli may be presented in quiet or in the presence of noise (e.g., speech babble, speech-spectrum noise), and the stimuli may be single nonsense syllables, monosyllabic words, nonsense sentences, or sentences. For some purposes, the stimuli are intentionally degraded by filtering or mixing them with noise, and depending on the purpose of suprathreshold evaluation, stimuli may be presented to one ear only (monaurally) or to both ears (binaurally). When stimuli are presented binaurally (both ears simultaneously), they may be identical (diotic) or different (dichotic). Stimuli may be presented at a specified level greater than speech recognition threshold or at varying intensity levels to establish a performance-intensity function. In suprathreshold testing, patient performance is often characterized in terms of percent correct, and standardized norms are used to interpret results.

Purposes for assessment of speech understanding include assessing auditory communication impairment, evaluating effectiveness of a hearing aid fitting, facilitating a comparison between hearing aids, and detecting possible VIIIth nerve or central auditory processing disorder. Suprathreshold stimuli may also be used to determine most comfortable and uncomfortable listening levels for purposes related to hearing aid fitting.

ELECTROPHYSIOLOGIC AUDIOMETRY

Auditory Evoked Potentials—Introduction

The electrophysiological response of the auditory system is often used by audiologists to evaluate auditory function. The techniques are derived from electroencephalography (EEG), which is the measurement of ongoing neural activity and has long been used to monitor brain function. The EEG can be recorded with surface electrodes attached to the scalp and connected to instrumentation that amplifies and records neural activity. Embedded in ongoing EEG activity is the brain's specific response to sensory stimulation. Auditory nervous system responses can be intentionally evoked with an auditory stimulus (such as an acoustic click) presented via an earphone (or other transducer) coupled to the ear. Neural responses that are time-linked to the stimulus can be recorded and differentiated from background EEG activity and other electrical noise sources (e.g., muscle artifact, 60 Hz electrical line noise).

Auditory Evoked Potentials—Clinical Applications

Auditory evoked response recording is an important tool for estimation of auditory sensitivity, particularly when conventional audiometry cannot be used. Evoked auditory potentials are also used routinely to assess the integrity of the auditory system (e.g., in tumor detection, auditory processing assessment, intra-operative monitoring),

but the following discussion will focus on threshold estimation/prediction.

Auditory evoked responses are used in place of conventional audiometry primarily in (*1*) infant hearing screening and assessment, (*2*) auditory evaluation of noncooperative children and adults, and (3) threshold estimation for people whose neurological status precludes use of conventional techniques. Although evoked potentials are not true measures of hearing, evoked potentials can be used in conjunction with other tests and information to estimate or predict hearing sensitivity. The capacity to make such estimates has important implications for early identification and rehabilitation of hearing impairment in newborns and young children, provision of auditory rehabilitation to people who have neurological problems, and even evaluation of nonorganic hearing impairment.

Historical Perspective

Early work indicated that ongoing EEG activity can be modified by sensory input (25). In order for a response specific to sensory stimulation to be observed, however, it was necessary to develop techniques to extract the sensory response from the ongoing EEG voltages. One important extraction technique that was developed involved algebraic summation (often called averaging) of responses that are linked in time to the sensory stimulus (26). When a bioelectric potential that is time-locked to a stimulus is recorded repeatedly and added to itself, the amplitude of the observed response will gradually increase with each stimulus repetition. In contrast, as EEG voltages during the same recording period are random, EEG voltages, when repeatedly summed, will gradually diminish or average out. Signal averaging was a critical advancement toward the clinical use of auditory evoked potentials. Other developments followed, and clinical applications of auditory potentials have now been investigated extensively. Measurement and assessment of evoked potentials are currently standard components of audiological practice.

Instrumentation

Many systems for recording auditory evoked potentials are now commercially available and are used widely. Components of the recording equipment typically include a stimulus generator capable of generating a variety of stimuli (e.g., clicks, tone bursts, tones), an attenuator, transducers for stimulus presentation (e.g., insert earphones, standard earphones, bone oscillator), surface electrodes, a differential amplifier, amplifier, filters, analog-to-digital converter, a signal averaging computer, data storage, display monitor, and printer. A simplified schematic diagram of a typical instrument is shown in Fig. 8.

In preparation for a typical single-channel recording, three electrodes are placed on the scalp. The electrodes often are called noninverting, inverting, and ground, but other terminology may be used (e.g., positive/negative or active/passive). A typical electrode montage is shown in Fig. 8, but electrode placements may vary depending on the potentials being recorded and the judgment of the clinician. Unwanted electrical or physiologic noise that may distort or obscure features of the response is reduced

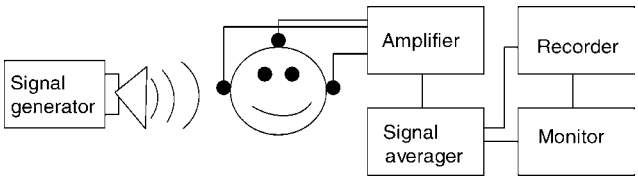


Figure 8. Simple block diagram of an auditory evoked response audiometry system.

by the use of differential amplifiers with high common mode rejection ratios and filters. It is important to note that electrode placement, stimulus polarity, stimulus presentation rate, number of signal presentations, signal repetition rate, filter characteristics, stimulus characteristics, and sampling rate during analog-to-digital conversion all affect the recording, and so must be controlled by the clinician.

Classification of Evoked Auditory Potentials

After the onset of an auditory stimulus, neural activity in the form of a sequence of waveforms can be recorded. The amount of time between the onset of the stimulus and the occurrence of a designated peak or trough in the waveform is called the latency. The latency of some auditory evoked potentials can be as short as a few thousandths of a second, and other latencies can be 400 ms or longer. Auditory evoked potentials are often classified on the basis of their latencies. For example, a system of classification can divide the auditory evoked potentials into "early" [$<$ 15 ms (e.g., electrocochleogram and auditory brainstem response)], "middle" [15–80 ms (e.g., Pa, Nb, and Pb)], and "late" [$>$ 80 ms (e.g., P300)] categories. Various classification systems based on latencies have been described, and other forms of classification systems based on the neural sites presumed to be generating the potentials (e.g., brainstem, cortex) are also sometimes used.

It is important to note that recording most bioelectric potentials requires only passive cooperation from the patient, but for some electrical potentials originating in the cortex of the brain, patients must provide active, cognitive participation. In addition, certain potentials are affected by level of consciousness. These factors, combined with the purpose of the evaluation, are important in the selection of the waveforms to be recorded.

Auditory Threshold Estimation/Prediction with AEPs

Auditory threshold estimations/predictions have been made on the basis of early, middle, and late potentials, but the evoked potentials most widely used for this purpose are those recorded within the first 10–15 ms after stimulus onset, particularly the so-called auditory brainstem response (ABR). An ABR evoked by a click consists of 5–7 peaks that normally appear in this time frame (27–29). Typical responses are shown in Fig. 9. The figure depicts three complete ABRs, and each represents the algebraic average of 2048 responses to a train of acoustic transient stimuli. The ABR is said to be time-locked such that each of the prominent peaks occurs in the normal listener at

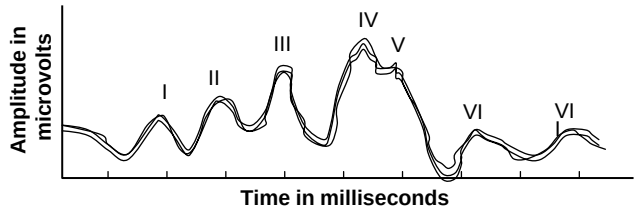


Figure 9. Normal auditory brainstem response; three complete responses.

predictable time periods after stimulation. Reliability is a hallmark of the ABR, and helps assure the audiologist that a valid estimation of conduction time through auditory brainstem structures has been made. A brief, gated, square-wave signal (click) stimulus is often used to generate the response, and stimulus intensity is reduced until the amplitude of the most robust peak (Wave V) is indistinguishable from the baseline voltage.

The response amplitude and latency (which lengthens as stimulus intensity decreases) are used to estimate behavioral auditory thresholds. In some equipment arrangements, computer software is used to statistically analyze the potentials for threshold determination purposes. ABR thresholds obtained with click stimuli correlate highly with behavioral thresholds at 2000–4000 Hz when hearing sensitivity ranges from normal to the severe range hearing impairment. Click stimuli are commonly used in clinical situations because their transient characteristics can excite many neurons synchronously, and thus a large amplitude response is evoked. However, variability limits the usefulness of click-evoked thresholds for prediction/estimation of auditory sensitivity of any particular patient (30), and the frequency specificity desired for audiometric purposes may not be obtained. As a result, gated tone bursts of differing frequencies are often used to estimate hearing sensitivity across the frequency range. These tonal stimuli can be embedded in bursts of noise to sharpen the frequency sensitivity and specificity of the test procedure.

In recent years, another evoked potential technique similar to the ABR has been developed to improve frequency specificity in threshold estimation while maintaining good neural synchronization. This technique, the auditory steady-state response (ASSR), uses rapidly (amplitude or frequency) modulated pure tone carrier stimuli (see Fig. 10). Evidence suggests that the ASSR is particularly useful when hearing sensitivity is severely impaired because high intensity stimuli can be used

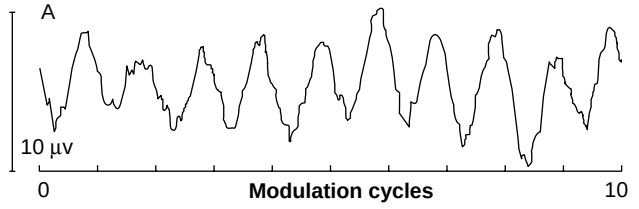


Figure 10. Auditory steady-state response.

(31). Research on the ASSR is ongoing, and this technique currently is considered to be a complement to click and tonal ABR in threshold estimation/prediction.

ACOUSTIC IMMITTANCE MEASUREMENT

Introduction

One procedure that helps audiologists interpret the results of conventional audiometry and other audiological tests is a measure of the ease with which energy can flow through the ear. Heaviside (1850–1925) coined the term impedance, as applied to electrical circuitry, and these principles were later applied in the United States during 1920s to acoustical systems (32). Mechanical impedance-measuring devices were initially designed for laboratory use, but electroacoustic measuring instruments were introduced for clinical use in the late 1950s (33). As acoustic impedance is difficult and expensive to measure accurately, measuring instruments using units of acoustic admittance are now widely used. The term used to describe measures incorporating the principles of both acoustic impedance and its reciprocal (acoustic admittance) is acoustic immittance. Modern instrumentation permits an estimate of ear canal and middle ear acoustic immittance (including resistive and reactive components).

Instrumentation

Commonly available immittance measuring devices (see Fig. 11) employ a probe-tone delivered to the tympanic membrane through the external ear canal. Sinusoids of differing frequencies are presented through a tube encased in a soft probe fitted snugly in the ear canal. The probe also contains a microphone and tubing connected to an air pump so that air pressure in the external ear canal can be varied from -400 to $+400$ mmH_2O.

Immittance devices also typically have a signal generator and transducers that can be used to deliver high intensity tones at various frequencies for the purpose of acoustic reflex testing. The American National Standards Institute (ANSI) has published a standard (S3.39-1987) for immittance instruments (34).

Immittance Measurement Procedures

As mentioned earlier in this chapter, the middle ear transduces acoustic energy into mechanical form. The transfer function of the middle ear can be estimated by measuring acoustic immittance at the plane of the tympanic membrane. These measures are often considered (in conjunction with the results of other audiological tests) in determining the site of lesion of an auditory disorder.

Static Acoustic Immittance

The acoustic immittance of the middle ear system is usually estimated by subtracting the acoustic immittance of the ear canal. This value is termed the compensated static acoustic immittance and is measured in acoustic mhos (reciprocal of acoustic ohms). The peak compensated

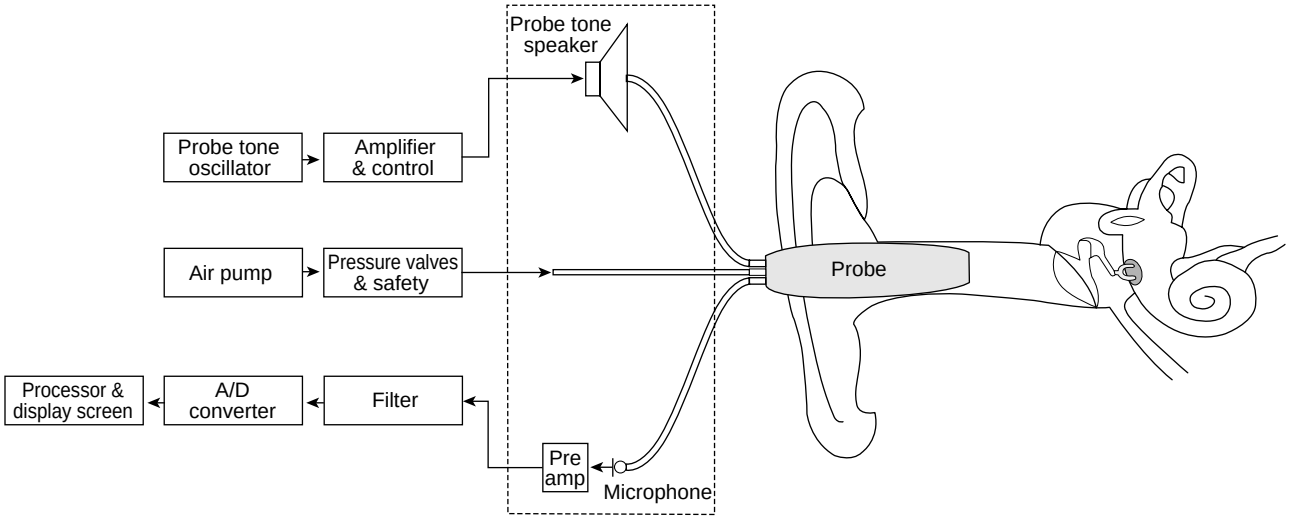


Figure 11. Block diagram of an acoustic immittance measuring device.

static acoustic immittance is obtained by adjusting the air pressure in the external ear canal so that a peak in the tympanogram exists. The magnitude of this peak, relative to the uncompensated immittance value, is clinically useful, because it can be compared with norms (e.g., 0.3 to 1.6 mmho) to determine the presence of middle ear pathology. It is important to note that at ear canal pressures of + 200 daPa or more, the sound pressure level (SPL) in the ear canal is directly related to the volume of air in the external ear canal, because the contribution of the middle ear system is insignificant at that pressure. A measure of the external ear canal volume is a valuable measure that can be used to detect tympanic membrane perforations otherwise difficult to detect visually. That is, a large ear canal volume (i.e., a value considerably greater than 1.5 mL) indicates a measurement of both the external ear canal and the middle ear as a result of a perforation in the tympanic membrane.

Dynamic Acoustic Immittance (Tympanometry)

The sound pressure of the probe-tone directed at the eardrum is maintained at a constant level and the volume velocity is measured by the instrumentation while positive and negative air pressure changes are induced in the external ear canal.

The procedure is called tympanometry and the resulting changes in immittance are recorded graphically as a tympanogram. A typical tracing is seen in Fig. 12.

Admittance is at its maximum when the pressures on both sides of the tympanic membrane are equal. Sound transmission decreases when pressure in the ear canal is greater or less than the pressure at which maximum admittance occurs. As a result, in a normal ear, the shape of the tympanogram has a characteristic peaked shape (see Fig. 13) with the peak of admittance occurring at an air pressure of 0 decapascals (daPa).

Tympanograms are sometimes classified according to shape (Fig. 13) (35).

The Type A tympanogram shown in Fig. 13, so-called because of it's resemblance to the letter "A", is seen in normal ears. When middle ear effusion is present, the fluid contributes to a decrease in admittance, regardless of the changes of pressure in the external ear canal. As a result, a characteristically flat or slightly rounded Type B tympanogram is typical. When the Eustachian tube malfunctions, the pressure in the middle ear can become negative relative to the air pressure in the external auditory canal. As energy flow through the ear is maximal when the pressure differential across the tympanic membrane is zero, tympanometry reveals maximum admittance when the pressure being

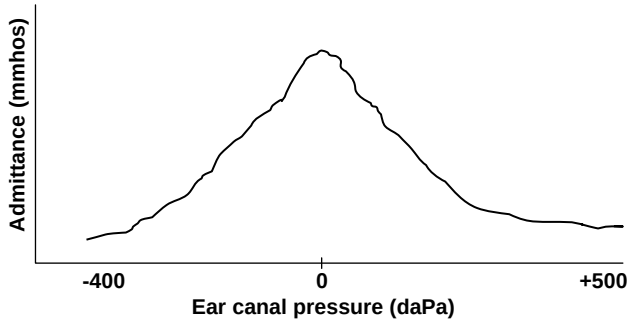


Figure 12. Tympanogram of a normal ear.

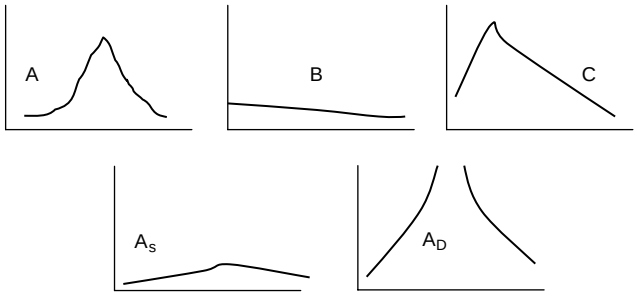


Figure 13. Tympanogram types (see text for descriptions).

varied in the external ear canal matches the negative pressure in the middle ear. At that pressure, the peak admittance will be normal but will occur at an abnormal negative pressure value. This tympanogram type is termed a Type C. It is important to note that variations exist of the type A tympanogram associated with specific pathophysiology affecting the middle ear. For example, if the middle ear is unusually stiffened by ear disease, the height of the peak may be reduced (Type A_s). Similarly, if middle ear pathology such as a break in the ossicular chain occurs, the energy flow may be enhanced, which is reflected in the Type A_d tympanogram depicted in Fig. 13.

Multifrequency Tympanometry

Under certain circumstances, particularly during middle ear testing of newborns and in certain stages of effusion in the middle ear, responses to the 226 Hz probe-tone typically used in tympanometry may fail to reveal immittance changes caused by disorders of the middle ear. In these circumstances, tympanometry with probe-tone frequencies above 226 Hz may be very useful in the detection of middle ear dysfunction. With probe-tone frequencies above 226 Hz, tympanometric shapes are more complex. More specifically, multifrequency tympanometric tracings normally progress through an expected sequence of shapes as probe frequency increases (36), and deviations from the expected progression are associated with certain pathologies.

Acoustic Reflex Measurement

In humans, a sufficiently intense sound causes a reflexive contraction of the middle ear muscles in both ears, acoustically stiffening the middle ear systems in each ear, called the acoustic reflex and is a useful tool in the audiometric test battery. When the reflex occurs, energy flow through both middle ears is reduced, and the resulting change in immittance can be detected in the probe ear by an immittance measuring device. Intense tones can be introduced to the probe ear (ipsilateral stimulation) or by earphone to the ear opposite the probe ear (contralateral stimulation).

One acoustic reflex measure is the minimum sound intensity necessary to elicit the reflex. The minimum sound pressure level necessary to elicit the reflex is called the acoustic reflex threshold. Acoustic reflex thresholds that are from 70 to 100 dBHL are generally considered to be in the normal range when pure tone stimuli are used. In general, the acoustic reflex thresholds in response to broadband noise stimuli tend to be lower than those for pure tones. Reduced or elevated thresholds, as well as unusual acoustic reflex patterns, are used by audiologists to localize the site of lesion and as one method of predicting auditory sensitivity.

OTOACOUSTIC EMISSIONS

When sound is introduced to the ear, the ear not only is stimulated by sound, it can also generate sounds that can be detected in the ear canal. The generated sounds, so-called otoacoustic emissions, have become the basis for the development of another tool that audiologists can use to assess the auditory system. In the following section, otoacoustic emissions will be described, and their relationship to conventional audiometry will be discussed.

Otoacoustic Emissions—Historical Perspective

Until relatively recently, the cochlea was viewed as a structure that converted mechanical energy from the middle ear into neural impulses that could be transmitted to and used by the auditory nervous system. This conceptual role of the cochlea was supported by the work on human cadavers of Georg von Bekesy during the early and middle 1900s, and summarized in 1960 (37). In Nobel Prize-winning research, von Bekesy developed theories to account for the auditory system's remarkable frequency sensitivity, and his views were widely accepted. However, a different view of the cochlea was proposed by one of von Bekesy's contemporaries, Thomas Gold, who suggested that processing in the cochlea includes an active process, a mechanical resonator (38). This view, although useful in explaining cochlear frequency selectivity, was not widely embraced at the time it was proposed.

In later years, evidence in support of Gold's idea of active processing in the cochlea accumulated. Particularly significant were direct observations of outer hair cell motility (39). In addition, observed differences in inner hair cell and outer hair cell innervation such as direct efferent innervation of outer but not inner hair cells (40) suggested functional differences in the two cell types. Most relevant to the present discussion were reports of the sounds that were recorded in the ear canal (41) and attributed to a mechanical process occurring in the cochlea, which are now known as otoacoustic emissions.

Otoacoustic Emissions—Description

Initially, otoacoustic emissions (OAEs) were thought to originate from a single mechanism, and emissions were classified on the basis of the stimulus conditions under which they were observed. For example, spontaneous otoacoustic emissions (SOAEs) are sounds that occur spontaneously without stimulation of the hearing mechanism. Two categories of otoacoustic emissions that are most widely used clinically by audiologists are (*1*) transient otoacoustic emissions (TOAEs), which are elicited by a brief stimulus such as an acoustic click or a tone burst, and (*2*) distortion product otoacoustic emissions (DPOAEs), which are elicited by two tones (called primaries) that are similar, but not identical, in frequency. A third category of otoacoustic emissions that may prove helpful to audiologists in the future is the stimulus frequency otoacoustic emission (SFOAE), which is elicited with a pure tone. Currently, SFOAEs are used by researchers studying cochlear function, but they are not used widely in clinical settings.

Recent research indicates that, contrary to initial thinking, otoacoustic emissions are generated by at least two mechanisms, and a separate classification system has been proposed to reflect improved understanding of the physical basis of the emissions. Specifically, it is believed that the

mechanisms that give rise to evoked otoacoustic emissions include (*1*) a nonlinear distortion source mechanism and (*2*) a reflection source that involves energy reflected from irregularities within the cochlea such as variations in the number of outer hair cell motor proteins or spatial variations in the number and geometry of hair cell distribution (42). Emissions currently recorded in the ear canal for clinical purposes are thought to be mixtures of sounds generated by these two mechanisms.

Instrumentation

Improved understanding of the mechanisms that generate otoacoustic emissions may lead to new instrumentation that can "unmix" evoked emissions. Currently, commercially available clinical equipment records "mixed" emissions and includes a probe placed in the external ear canal that both delivers stimuli (i.e., pairs of primary tonal stimuli across a broad range of frequencies, clicks or tone-bursts) and records resulting acoustic signals in the ear canal. The microphone in the probe equipment is used in (*1*) the verification of probe fit, (*2*) monitoring probe status (e.g., for cerumen occlusion), (*3*) measuring noise levels, (*4*) verifying stimulus characteristics, and (*5*) detecting emissions. Otoacoustic measurement recording entails use of probe tips of various sizes to seal the probe in the external ear canal and hardware/software that control stimulus parameters and protocols for stimulus presentation. The computer equipment performs averaging of responses time-locked to stimulus presentation, noise measurement, artifact rejection, data storage, and so on, and can provide stored normative data and generate printable reports. An example of a typical DPOAE data display is shown in Fig. 14.

It is important to note that outer or middle ear pathology can interfere with transmission of emissions from the cochlea to the ear canal, and thus the external ear canal and middle ear status are important factors in data interpretation. Also, although otoacoustic emissions ordinarily are not difficult to record and interpret, uncooperative patient behavior and high noise levels can hamper or even preclude measurement of accurate responses.

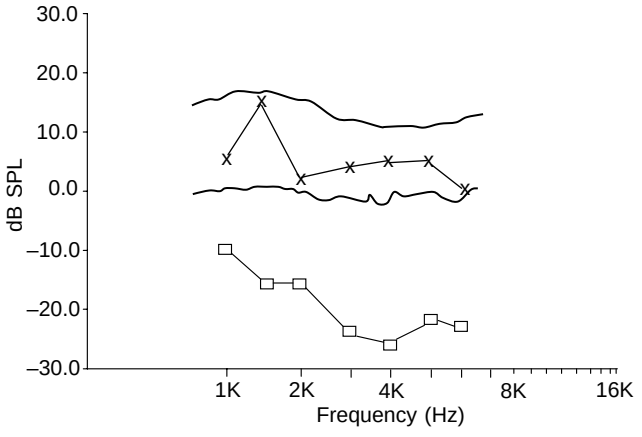


Figure 14. DPAOE responses from 1000 to 6,000 Hz from the left ear of a normal listener. X = DPOAE response amplitudes; squares = physiologic noise floor; bold curves = 95% confidence limits for normal ears.

Otoacoustic Emissions—Clinical Applications

Measurement of otoacoustic emissions is used routinely as a test battery component of audiometric evaluations in children and adults, and it is particularly useful for monitoring cochlear function (e.g., in cases of noise exposure and during exposure to ototoxic medication) as well as differentiating cochlear from neural pathology. Currently, otoacoustic emission evaluation is also useful, either alone or in combination with evoked potential recording, in newborn hearing screening. In addition, OAE assessment is sometimes used in preschool and school-age hearing screening, as well as with patients who may be unwilling to cooperate during audiometry.

Otoacoustic Emissions and Audiometric Threshold Prediction/Estimation

Audiologists do not use otoacoustic emissions as a measure of "hearing" because OAEs constitute an index of cell activity in the inner ear, not "hearing." Research suggests, however, that otoacoustic emissions may become an important indicator for predicting/estimating auditory thresholds when conventional audiometry cannot be conducted (42).

Many sources of variability exist that affect DPOAE use in audiometric threshold prediction/estimation, including variability with respect to etiology of the hearing loss, age, gender, and uncertainty regarding locations of DPOAE generation and their relationship to audiometric test frequencies. Individual DPOAE amplitude variation, intra- and inter-subject variations occurring at different frequencies and at different stimulus levels, and the mixing of emissions from at least two different regions of the cochlea (as described above) can reduce frequency selectivity and specificity in DPOAE measurement.

It has been suggested that developing methods to "unmix" the emissions associated with different generators (e.g., through the use of suppressor tones to reduce or eliminate one source component or with the use of Fourier analysis to analyze the emissions) may reduce variability and improve specificity in threshold prediction/estimation and determination of etiology (41). It is likely that future commercial otoacoustic measurement instruments will enable the user to differentiate distortion source emissions from reflection source emissions and that this improvement will lead to more widespread use of otoacoustic emissions in audiometric threshold estimation and prediction.

BIBLIOGRAPHY

1. American Academy of Audiology. Scope of practice. Audiol Today 2004;15(3):44–45.
2. American National Standards Institute. Methods of manual pure tone threshold audiometry. (ANSI S3.21-2004). New York: ANSI; 2004.
3. Bergman M. On the origins of audiology: American wartime military audiology. Audiol Today Monogr 2002;1:1–28.

4. Seashore CE. An Audiometer. University of Iowa Studies in Psychology (No. 2). Iowa City, NA: University of Iowa Press; 1899.
5. Bunch CC. Clinical Audiometry. St. Louis, MO: C. V. Mosby; 1943.
6. Fowler EP, Wegel RL. Presentation of a new instrument for determining the amount and character of auditory sensation. Trans Am Otol Soc 1922;16:105–123.
7. Bunch CC. The development of the audiometer. Laryngoscope 1941;51:1100–1118.
8. Carhart R. Clinical application of bone conduction audiometry. Archi Otolaryngol 1950;51:798–808.
9. American National Standards Institute. Specification of audiometers. (ANSI-S3.6-1996). New York: ANSI; 1996.
10. American National Standards Institute. Maximum permissible ambient noise levels for audiometric test rooms. (ANSI-S3.1-1999 [R2003]). New York: ANSI; 1999.
11. Breakwell GM, Hammond S, Fife-Shaw C. Research Methods in Psychology. 2nd ed. London: Sage; 2000. p 200–205.
12. von Bekesy G. A new audiometer. Acta Oto-laryngologica Stockholm 1947;35:411–422.
13. Jerger J. Bekesy audiometry in analysis of auditory disorders. J Speech Hear Res 1960;3:275–287.
14. Jerger J, Herer G. Unexpected dividend in Bekesy audiometry. J Speech Hear Disord 1961;26:390–391.
15. Jerger S, Jerger J. Diagnostic value of Bekesy comfort loudness tracings. Arch Otolaryngol 1974;99:351–360.
16. Carhart R, Jerger J. Preferred method for clinical determination of pure-tone thresholds. J Speech Hear Disord 1959;24: 330–345.
17. Feldmann H. A history of audiology: A comprehensive report and bibliography from the earliest beginnings to the present. Translations Beltone Institute Hear Res 1960;22:1–111. [Translated by Tonndorf J. from Die Geschichtliche Entwicklung der Horprufungsmethoden, Kuze Darstellung and Bibliographie von der Anfrongen bis zur Gegenwart. In: Leifher H, Mittermaiser R, Theissing G, editors. Zwanglose Abhandunger aus dem Gebeit der Hals-Nasen-Ohren-Heilkunde. Stuttgart: Georg Thieme Verlag; 1960.]
18. Hudgins CV, Hawkins, JE Jr, Karlin JE, Stevens SS. The development of recorded auditory tests for measuring hearing loss for speech. Laryngoscope 1947;57:57–89.
19. Olsen WO, Matkin ND. Speech audiometry. In: Rintelmann WF, editor. Hearing Assessment. 2nd ed. Austin, TX: Pro-Ed; 1991. 39–140.
20. Fletcher H. Speech and Hearing. Princeton NJ: Von Nostrand; 1929.
21. Fletcher H. A method of calculating hearing loss for speech from an audiogram. Acta Otolaryngologica 1950; (Suppl 90): 26–37.
22. Carhart R. Observations on the relations between thresholds for pure tones and for speech. J Speech Hear Disord 1971; 36:476–483.
23. Beattie RC, Edgerton BJ, Svihovec DV. An investigation of Auditec of St. Louis recordings of Central Institute for the Deaf spondees. J the Am Audiol Soc 1975;1:97–101.
24. Cambron NK, Wilson RH, Shanks JE. Sondaic word detection and recognition functions for female and male speakers. Ear Hear 1991;12:64–70.
25. Loomis A, Harvey E, Hobart G. Disturbances of patterns in sleep. J Neurophysiol 1938;1:413–430.
26. Clark WA Jr, Goldstein MH Jr, Brown RM, Molnar CE, O'Brien DF, Zieman HE. The average response computer (ARC): A digital device for computing agerages and amplitudes and time histograms of physiological responses. Trans of IRE 1961;8:46–51.
27. Jewett DL, Romano MN, Williston JS. Human auditory evoked potentials: Possible brainstem components detected on the scalp. Science 1970;167:1517–1518.
28. Jewett DL, Williston JS. Auditory evoked far fields averaged from the scalp of humans. Brain 1971;94:681–696.
29. Stapells DR, Oates P. Estimation of the pure-tone audiogram by the auditory brainstem response: A review. Audiol Neurootol 1997;3(5):257–280.
30. Rance G, Briggs RJS. Assessment of hearing in infants with moderate to profound impairment: The Melbourne experience with auditory stead-state evoked potential testing. Ann Otol Rhinol Laryngol 2002;111(5) (Part 2, Suppl. 189):22–28.
31. Swanepoel D, Roode R. Auditory steady-state responses for children with severe to profound hearing loss. Arch of Otolaryngol Head Neck Surg 2004;130(5):531–535.
32. Margolis RH, Hunter LH. Acoustic Immittance Measurements. In: Roeser RJ, Valente M, Hosford-Dunn H, editors. Audiology Diagnosis. New York: Thieme; 2002. pp 381–423.
33. Terkildsen K, Nielsen S. An electroacoustic impedance measuring bridge for clinical use. Arch Otolaryngol 1960;72:339–346.
34. American National Standards Institute. National Standard Specifications for instruments to measure aural acoustic impedance and admittance. (ANSI S3.39-1987). New York: ANSI; 1987.
35. Jerger J. Clinical experience with impedance audiometry. Arch Otolaryngol 1970;92:311–324.
36. Vanhuyse VJ, Creten WL, Van Camp KJ. On the W-notching of tympanograms. Scand Audiol 1975;4:45–50.
37. von Bekesy G. Experiments in Hearing. New York: McGraw Hill; 1960.
38. Gold T. Hearing. II. The physical basis of the acton of the cochlea. Proc Roy Soc Brit 1948;135:492–498.
39. Brownell WE, Bader CR, Bertrand D, Ribaupierre Y. Evoked mechanical responses of isolated cochlear outer hair cells. Science 1985;227:194–196.
40. Smith CA. Innervation pattern of the cochlea. Ann Otol Rhinol Otolaryngol 1961;70:504–527.
41. Kemp DT. Stimulated acoustic emissions from within the human auditory system. J Acoust Soc Am 1978;64:1386–1391.
42. Shera CA. Mechanisms of mammalian otoacoustic emission and their implications for the clinical utility of otoacoustic emissions. Ear Hear 2004;25:86–97.

See also COCHLEAR PROSTHESES; COMMUNICATION DEVICES.

AUDITORY IMPLANTS.

See COCHLEAR PROSTHESES.

AUGMENTATIVE COMMUNICATION SYSTEM.

See COMMUNICATION DEVICES.

B

BACTERIAL DETECTION SYSTEMS. See MICROBIAL DETECTION SYSTEMS.

BALLOON PUMP. See INTRAAORTIC BALLOON PUMP.

BANKED BLOOD. See BLOOD COLLECTION AND PROCESSING.

BAROTRAUMA. See HYPERBARIC MEDICINE.

BARRIER CONTRACEPTIVE DEVICES. See CONTRACEPTIVE DEVICES.

BIOCERAMICS. See BIOMATERIALS: BIOCERAMICS.

BIOCOMPATIBILITY OF MATERIALS

ZHOU XIANG
MYRON SPECTOR
Brigham and Women's Hospital
Boston, Massachusetts

INTRODUCTION

Most surgical specialties have been revolutionized by the introduction of implantable devices. These advances have been founded, in large part, on biomaterials science and engineering. One of the critical determinants of the performance of the device relates to its compatibility with the structure and function of the tissue or organ (a structure comprised of two or more tissues) in which it is implanted. Moreover, the tissue response to the implant should not impede the required function of the device. This article will deal with the response to biomaterials implanted into solid tissues. Biocompatibility issues related to blood-contacting applications will be outside the scope of this discussion.

Implantation of the device requires the production of a surgical wound, and in this respect the tissue response to the implant may be looked upon as the modification of the wound healing response by the very presence of the implant. In vascularized tissues, the creation of a surgical wound elicits an inflammatory process that can be considered part of the natural course of healing. The end result of the healing process is tissue similar to that naturally occurring at the site (the process of "regeneration") or scar tissue (the process of "repair"), which in many tissues and organs comprises fibrous tissue. Infectious organisms (viz., bacteria) when present serve as a persistent injurious agent that prolongs and can further incite the inflammatory process, not only jeopardizing the performance of the implant, but threatening the life of the individual. The principles of biocompatibility including the mechanisms underlying inflammatory and infectious processes apply regardless of the type of material of fabrication of the implant. There are, however, features of the biomaterials that can affect certain aspects of these processes. It can therefore be instructive to also consider issues of biocompatibility in the context of the various classes of materials: metals, ceramics, and polymers.

CLASSES OF MATERIALS

The term "biomaterials" generally refers to synthetic materials and treated natural materials that are employed for the fabrication of implantable devices that are to replace or augment tissue or organ function. An understanding of the chemical makeup of biomaterials can provide insights into their biocompatibility for selected applications. Generally, biomaterials may be considered "inert" or "reactive" with the biological milieu. In the latter case, the reactivity could relate to the release of moieties or the adsorption of biological molecules. Inert materials may also release small amounts of ions and molecules or nonspecifically bind biomolecules. The feature that distinguishes inert from reactive biomaterials is the degree to which the interactions of the implant with the biological environment affects the tissue response and device performance. Those materials designed to effect specific tissue responses through their reactivity may also be referred to as "bioactive".

This article provides a brief description of materials used for the fabrication of implantable devices relative to issues related to biocompatibility. A more comprehensive discussion of biomaterials can be found elsewhere in this encyclopedia.

Metals

In metals, closely packed arrays of positively charged atoms are held together in a loosely associated "cloud" of free electrons. The essential features of the metallic bond are that it is nondirectional and the electrons are freely mobile. The metals most often used for the fabrication of implantable devices are stainless steel, cobalt–chromium alloys, and titanium and titanium alloy. The specific members of these families used as biomaterials are usually identified by a designation provided by the American Society for Testing and Materials (ASTM). Metallic materials have certain properties that make them ideal for load-bearing applications; in particular, they can maintain very high strength under the aggressive aqueous environment in the body.

The biocompatibility of the implantable metallic materials is related to their corrosion resistance, in that they can generally be considered as inert. While they release detectable levels of metal ions (1,2), these ions have not yet been found to significantly affect tissue or organ function or cause pathological changes. One conundrum related to biomaterials is that while the ions released from certain metallic devices are known to be carcinogenic when administered to certain animals models (3,4) and when encountered by humans in certain circumstances (1), there have not yet been definitive studies relating the incidence of

various cancers to the ions released by metallic implants (see later section).

The following sections provide a brief description of three metal systems most frequently used for the fabrication of implantable devices.

Stainless Steels. The stainless steels, like all steels, are iron-based alloys. Chromium is added to improve the corrosion resistance through the formation of a chromium-oxide surface layer; at least 17% chromium is required for the term "stainless" to be used. Carbon and nickel are employed as alloying elements to increase strength. The most common type of stainless steel used for implants is 316L (American Iron and Steel Institute designation; ASTM F-138), containing 17–19% chromium, 13–15.5% nickel, and <0.03% carbon. Despite their very good corrosion resistance, stainless steels are subject to several types of corrosion processes, including crevice, pitting, intergranular, and stress corrosion. These processes can profoundly degrade the mechanical strength of the alloy and can lead to the elevated release of metallic ions into the surrounding tissue with undesirable biological consequences.

Alloying with chromium generates a protective, self-regenerating oxide film that resists perforation, has a high degree of electrical resistivity, and, thus, provides a major protection against corrosion; formation of the chromium oxide "passivation" layer is facilitated by immersion of the alloy in a strong nitric acid solution. The nickel imparts more corrosion resistance and ease of fabrication. The molybdenum addition provides resistance to pitting corrosion. Other alloying elements facilitate manufacturing processes.

Cobalt–Chromium Alloys. Surgical cobalt–chromium alloy is a cobalt-based system with chromium added for increased corrosion resistance. Its composition contains 27–30% chromium and 5–7% molybdenum. Tungsten is added to the wrought alloy to enhance ductility. As with stainless steels, the chromium content of this alloy generates a highly resistive passive film that contributes substantially to corrosion resistance. The Co–Cr Mo (F-75) alloy has superior corrosion resistance to the F-138 stainless steel, particularly in crevice corrosion. There is an extensive, decades-long history of biocompatibility in human implantation. The Co–Cr Mo devices are currently produced by a process referred to as hot isostatic pressing that results in parts with more favorable strength characteristics than results from casting processes. The Co–Cr Tn–Ni Alloy (ASTM F90) is very different from the F-75 alloy with which it is often confused. Parts can be fabricated from this alloy by hot forging and cold drawing; it is not used in the cast form. In clinical practice, F90 is used to make wire and internal fixation devices (e.g., plates, intramedullary rods, and screws).

Unalloyed Titanium (ASTM F-67) and Titanium Alloy (ASTM F-136). Titanium and its alloy with 6% aluminum and 4% vanadium, Ti-6A1-4V, are used for their excellent corrosion resistance and their modulus of elasticity that is approximately one-half that of stainless steel and cobalt–chromium alloys. This lower modulus results in devices with lower stiffness that may be advantageous in certain applications such as implants in bone as they will result in less stress-shielding of bone. The alloy of titanium has much better material properties than the pure titanium. Problems with titanium are its severe notch sensitivity and poor wear resistance.

Titanium and its alloys are of particular interest for biomedical applications due to their outstanding biocompatibility. In general, their corrosion resistance significantly exceeds that of the stainless steels and the cobalt–chromium alloys. In saline solutions near neutral pH, the corrosion rate is extremely small, and there is no evidence of pitting, intergranular, or crevice corrosion. Unalloyed titanium is used less frequently than the alloy for the fabrication of implants. It is, however, available in various configurations, such as plain wire for manufacturing purposes. In addition, it is used to produce porous coatings for certain designs of total joint replacement prostheses.

ASTM F-136 specifies a titanium alloy with a content of 5.5–6.5% Al, 3.5–4.5% V, 0.25% Fe (maximum), 0.05% N (maximum) and 0.08% C (maximum), 0.0125% H (maximum), and other 0.1% (maximum 0.4% total). Developed by the aircraft industry as one of a number of high strength titanium alloys, this particular formulation has a yield strength reaching 1110 MPa. The ASTM F-136 specification limits the oxygen to an especially low level of 0.13% maximum. This is also known in the industry as the ELI (extra low interstitial) grade. Limiting the level of oxygen improves the mechanical properties of the material, particularly increasing its fatigue life. One interesting feature of titanium and its alloys is the low modulus of elasticity of 100 GPa as compared to 200 GPa for the cobalt–chromium alloys. This feature leads to their use in plates for internal fixation of fractures. Some have found that the lower stiffness of these plates may decrease the severity of bone stress-shielding that results in osteopenia under these devices.

One of the weaknesses of titanium is its poor wear resistance. It appears that this problem relates to the mechanical stability of the passive film covering the surface of the alloy. On a carefully polished surface, the film is highly passive but mechanically weak. The poor wear resistance of titanium can result in the release of particulate wear debris if the material is not judiciously employed in the fabrication of implants (5). While the metal in bulk form may be biocompatible adverse cell and tissue responses may be elicited by titanium particles (5–8).

Permanent and Absorbable Synthetic and Natural Polymers

Polymers consist of long chains of covalently bonded molecules characterized by the repeated appearance of a monomeric molecular unit. They can be produced de novo by the polymerization of synthetic monomers or prepared from natural polymers isolated from tissues. Most synthetic and natural polymers have a carbon backbone. Bonding among polymer chains results from much weaker secondary forces—hydrogen bonds or van der Waal's forces. Covalent bonding among chains, referred to as cross-linking, can be produced in certain polymer systems. Physical

entanglements of the long polymer chains, the degree of crystallinity, and chemical cross-linking among chains play important roles in determining polymer properties. The molecular bonding of the backbone of the polymer can be designed to undergo hydrolysis or enzymatic breakdown thus allowing for the synthesis of absorbable, as well as permanent, devices.

Polymeric materials are generally employed for the fabrication of implants for soft tissue applications that require a greater degree of compliance than can be achieved with metals. However, they have also been shown to be of value as implants in bone for indications that would also benefit from their lower modulus of elasticity, and the ability of some to be polymerized *in vivo* so as to adapt to defects of complex shape. For some indications, the radiolucency of polymeric materials may be an important benefit. Because of the limited strength and wear resistance of polymers, care must be given to the load-bearing requirements of the applications in which they are used.

The following sections provide a summary of a few of the most frequently used polymers.

Polymethyl Methacrylate (PMMA). Polymethyl methacrylate (9) is used in a self-curing form as a filling material for defects in bone and as a grouting agent for joint replacement prostheses. It can be shaped *in vivo* while in a dough stage prior to complete polymerization and thus makes a custom implant for each use. Its purpose is the redistribution of stress in a more even pattern to the surrounding bone. Often referred to as "bone cement", when PMMA is employed for joint arthroplasty it acts as a grout to support the prosthesis rather than a glue; it has minimal adhesive properties. The time-dependent properties of PMMA during curing require an understanding of its handling characteristics. Immediately after mixing, the low viscosity permits interdigitation with cancellous bone. Viscosity rises quickly once the chemical setting reaction begins requiring that the prosthesis be accurately positioned and stationary to achieve maximum fixation.

The chemical toxicity of the methyl methacrylate monomer and the heat generated during the polymerization exothermic reaction need be considered when using PMMA. While the toxicity of the monomer has not prevented the material from being employed successfully for a wide array of applications, there are efforts to reduce the monomer content (10) and to employ alternative activating agents that would be less toxic (11).

Its brittle nature after curing and low fatigue strength make PMMA vulnerable to fracture under high mechanical loading and production of wear debris in situations when other harder materials rub against it. The cellular response to PMMA particles can promote an inflammatory response (12,13) that can result in osteolysis. This process has been referred to as "cement disease" (14). This underscores the importance of reducing the circumstances that can result in the production of PMMA debris.

Silicone. Silicones (15) are polymers having a backbone comprising alternating silicon and oxygen atoms with organic side groups bonded to the silicon through covalent bonding with the carbon atom. One form of silicone commonly used for the fabrication of implants is polydimethylsiloxane (PDMS). In PDMS, methyl (CH_3) side groups are covalently bonded to the silicon atom and it can be used in three forms: (a) a fluid comprising linear polymers of varying molecular weight (i.e., chain length); (b) a cross-linked network referred to as a gel; and (c) a solid elastomer comprising a highly cross-linked gel filled with small particles of silica. In considering the performance of silicone implants, the role of each form of PDMS in the device need be considered. Attributing specific biological responses to individual components of a silicone device is complicated by the many molecular forms of silicone that the implant comprises.

The PDMS elastomers contain a noncystalline silica particles 7–22 nm in diameter that has been surface-treated to facilitate chemical bonding of the particle to the PDMS gel. Addition of the silica particle to a highly cross-linked PDMS gel is done to modify the mechanical properties of the elastomer. One of the challenges in investigating tissue responses that may have been elicited by the release of the silica particles is their small size that requires transmission electron microscopy (TEM) for analysis.

Polyethylene. Ultrahigh molecular weight polyethylene (UHMWPE) (16,17) has a very low frictional coefficient against metal and ceramics, and is therefore used as a bearing surface for joint replacement prostheses. Moreover, the wear resistance of UHMWPE is higher than other polymers investigated for this application. Low strength and creep, however, present potential problems.

The term polyethylene refers to plastics formed from polymerization of ethylene gas. The possibilities for structural variation on molecules formed by this simple repeating unit for different molecular weight, crystallinity, branching, cross-linking, and so on, are so numerous and dramatic with such a wide range of attainable properties that the term polyethylene truly refers to a subclass of materials. The earliest type of polyethylene was made by reacting ethylene at high (20,000–30,000 lb/in.^{2}) pressure and temperatures of 200–400 °C with oxygen as catalyst. Such material is referred to as conventional or low density polyethylene. Much polyethylene is produced now by newer, low pressure techniques using aluminum–titanium (Ziegler) catalysts. This is called linear polyethylene due to the linearity of its molecules, in contrast to the branched molecules produced by high pressure processes. The linear polymers can be used to make high density polyethylene by means of the higher degree of crystallinity attained with the regularly shaped molecules. Typically, there is no great difference in molecular weight between the low density and high density varieties, (e.g., 100,000–500,000). However, if the low pressure process is used to make extremely long molecules, (i.e., UHMWPE), the result is different and quite remarkable. This material, with a molecular weight between 1 and 10 million, is less crystalline and less dense than high density polyethylene and has exceptional mechanical properties. The material is used in very demanding applications and is by far the most successful polymer used in total joint replacements. It far outperforms the various acrylics, fluorocarbons,

polyacetals, polyamides, and polyesters that were tried for such purposes. In recent years, cross-linking methods including chemical agents and ionizing radiation have been implemented in an attempt to further improve the wear performance of UHMWPE.

The principal biocompatibility issue with polyethylene relates to the inflammatory response provoked by particles of the polymer, as can be generated by the wear of total joint replacement prostheses prostheses (18,19).

Absorbable Polymers. Absorbable polymers have been used in the fabrication of surgical implants for decades in the form of absorbable sutures. More recently, this class of materials has been investigated for the application of resorbable devices including fracture fixation implants and scaffolds for tissue engineering. The principal issues associated with the implementation of absorbable polymers as implants include: mechanical properties (i.e., strength), degradation rate, and biological response to the degradation products.

Synthetic Absorbable Polymers. One of the classes of polymers used frequently for the fabrication of absorbable implants is the alpha-hydroxy acids including L-lactic acid, glycolic acid, and dioxanone. These molecules normally are used in their polymeric forms: poly L-lactic acid (PLLA), polyglycolic acid (PGA), and polydioxanone. Copolymers of lactic and glycolic acids are also frequently employed.

This particular class of polyester undergoes breakdown as a result of the hydrolytic scission of the ester bond. The access of water to this bond in PGA is much greater resulting in a more rapid degradation rate compared to that which occurs with PLLA, which has a bulkier CH_3 side group instead of the H atom in PGA. The copolymer of polylactic and polyglycolic acid can be designed to have an intermediate degradation rates. While the majority of the breakdown of these polymers is due to hydrolytic scission there is some lesser extent of nonspecific enzymatic action.

Several factors affect the rate of breakdown of these polymers: the relative amount of monomers comprising the copolymers, the degree of crystallinity, and the surface area. These polymers are normally broken down to natural body components excreted in the urine or exhaled. The process of degradation involves the gradual decrease in the average molecular weight of the polymer as hydrolysis proceeds. At some point, the molecular weight decreases to the extent that the polymer becomes soluble in the aqueous environment and there is a bolus release of the molecules. Depending on the mass of the implanted device, the concentration of the molecules may elicit an inflammatory response (20).

Natural Absorbable Polymers. Myriad devices are fabricated from collagen (21), the principal structural protein of the body. The collagen molecule comprises three tightly coiled helical polypeptide chains. *In vivo* the collagen molecule, tropocollagen, is assembled to form fibrils that in turn assume various orientations and configurations to form the architecture of various tissues. The wide array of properties of tissues comprising collagen, from dermis to musculoskeletal tissues including articular cartilage, meniscus, and ligament, is due to differences in the chemistry, density, and orientation of the fibrils formed from the collagen molecule.

Collagen is soluble in specific solutions in which the chains can become disentangled to produce gelatin. It can be isolated from tissue and purified through the use of several agents: acids, alkalis, enzymes, and salt. Treatment in acid results in the elimination of acidic proteins and glycosaminoglycans that result in the dissociation of the collagen fibrils. A similar effect can be achieved using alkaline extraction with the removal of basic proteins. Proteolytic enzymes that cleave the telopeptides, that serve as natural cross-linking agents for collagen, allow for the dissolution of collagen molecules and aggregates in aqueous solutions. Salt extraction leads to the removal of newly synthesized collagen molecules and certain noncollagenous molecules thus facilitating the disaggregation of collagen fibrils.

That collagen is soluble in acidic medium facilitates its extraction from tissues and reprocessing into biomaterials. Several factors are critical determinants of the properties of reconstituted collagen biomaterials. The degree to which denaturation or degradation of the collagen structures isolated from tissue occurs will affect the mechanical properties. These properties will also be affected by the degree to which the material is subsequently cross-linked.

An important biological property related to the molecular structure of collagen is the collagen-induced blood platelet aggregation. The quarternary structure of collagen resulting from the periodic aggregation of the collagen molecules has been well documented. Methods for isolating and purifying collagen fibrils, that result in the preservation or destruction of this quarternary structure are employed to produce either hemostatic or thromboresistant biomaterials. Another factor relates to the removal of soluble components that might serve as antigens. The immunogenicity can be reduced, to clinically nonconsequential levels, by chemically modifying the antigen molecules.

A wide variety of methods have been employed for the fabrication of collagen sutures, fleeces for hemostasis, and sponge-like materials for scaffolds for tissue engineering.

Ceramics

Ceramics are typically three-dimensional (3D) arrays of positively charged metal ions and negatively charged nonmetal ions, often oxygen. The ionic bond localizes all the available electrons in the formation of a bond. Network organization ranges from highly organized, crystalline, 3D arrays to amorphous, random arrangements in glassy materials.

Ceramics, for reasons described above, may be the most chemically inert implant materials currently in use. However, their relatively low tensile strength, high modulus, and brittleness limit the applications in which they may be used. Current techniques allowing the formation of ceramic coatings on metallic substrates have revitalized interest in ceramics for hard tissue applications.

Aluminum Oxide. Aluminum oxide has been found of value for the articulating components of total joint arthroplasties because of its high wear resistance and its low coefficient of friction when prepared in congruent, polished geometries. The brittle nature of alumina remains a detriment.

Calcium Phosphates/Hydroxyapatite. Calcium-based ceramics, closely related to the naturally occurring hydroxyapatite in bone, have generated a large amount of interest in recent years. The ability to bond directly to bone as well as their osteoconductive capability promise to enhance biological fixation of implant devices. Hydroxyapatite is only slightly resorbable and is used in both dense and porous forms as a permanent implant. Tricalcium phosphate is bioabsorbable to varying degrees, depending on formulation and structure. There are currently a wide array of calcium phosphate materials undergoing investigation as bone graft substitute materials (see the Section on Calcium Phosphate Materials as Matrices for Bone Regeneration: Bone Graft Substitute Materials).

Composite Materials

Composite materials are combinations of two or more materials, and usually more than one material class (i.e., metals polymers, ceramics). They are used to achieve a combination of mechanical properties for specific applications. Composite technology, much of it developed for the aerospace industry, is beginning to make it way into biomedical materials. Carbon fiber reinforced polymers are being investigated as substitutes for metals. The advantage is that devices with comparable strength, but with significantly lower stiffness can be produced. Moreover, these types of composite devices are radiolucent.

BIOLOGICAL RESPONSE TO BIOMATERIALS

The biological processes comprising the tissue response are affected by implant-related factors including (22):

1. The "dead space" created by the presence of the implant.
2. Soluble agents released by the implant (e.g., metal ions or polymer fragments).
3. Insoluble particulate material released from the implant (e.g., wear debris).
4. Chemical interactions of biological molecules with the implant surface.
5. Alterations in the strain distribution in tissue caused by the mismatch in modulus of elasticity between the implant and surrounding tissue, and the movement of the implant relative to adjacent tissue as a result of the absence of mechanical continuity.

Study of the tissue response to implants requires methodology capable of measurements at the molecular, cellular, and tissue levels. Moreover, time is an important variable owing to the criticality of the temporal relationship between the molecular and cellular protagonists of the biological reactions, and because implant-related factors act with different time constants on the biological responses. The dynamic nature of implant-tissue interactions requires that the final assessment of tissue compatibility be qualified by the time frame in which it has been evaluated.

The tissue response to an implant is the cumulative physiological effect of (*1*) modulation of the acute wound healing response to the surgical trauma of implantation and the presence of the implant, (*2*) the subsequent chronic inflammatory reaction associated with the presence of the device, and (*3*) remodeling of surrounding tissue as it adapts to the presence of the implant (23). Moreover, the healing and stress-induced adaptive remodeling responses of different tissues vary considerably. In this regard, the response of various tissues to the same implant can vary greatly.

In considering the biological response that might be elicited by an implant, the healing–remodeling characteristics of the four basic types of tissue: connective tissue, muscle, epithelia, and nerve, should be recalled. The characteristics of the parenchymal cells in each type of tissue can provide a basis for understanding the tissue response to an implant. The following characteristics of an implant site are determinants of the biological response:

1. Vascularity.
2. The nature of the parenchymal cell with respect to its capability for mitosis and migration, because these processes determine the regenerative capability of the tissue.
3. The presence of regulatory cells such as macrophages/histiocytes.
4. The effect of mechanical strain, associated with deformation of the extracellular matrix, on the behavior of the parenchymal cell.

Surgical wounds in avascular tissue (e.g., the cornea and inner third of the meniscus) will not heal because of the limited potential for the proliferation and migration of surrounding parenchymal cells and the absence of a fibrin clot in the wound site into which the cells can migrate. Gaps between an implant and surrounding avascular tissue can remain indefinitely. Implant sites in vascular tissues in which the parenchymal cell does not have the capability for mitosis heal by the formation of scar in the gap between the implant and surrounding tissue. Moreover, adjacent cells that have died as a result of the implant surgery will be replaced by fibroblasts and scar tissue.

Normal Local Tissue Response

Wound Healing. Implantation of a medical device initiates a sequence of cellular and biochemical processes that lead to "healing by second intention" (i.e., healing by the formation of granulation tissue within a defect; as opposed to the healing of an incision, i.e., healing by first intention). The first phase of healing in vascularized tissues is inflammation, which is followed by a reparative phase, the replacement of the dead or damaged cells by healthy cells. The pathway that the reparative process takes depends

on the regenerative capability of the cells comprising the injured tissue (i.e., the tissue or organ into which the implant has been placed). Cells can be distinguished as labile, stable, or permanent based on their capacity to regenerate. Labile cells continue to proliferate throughout life, replacing cells that are continually being destroyed. Epithelia and blood cells are examples of labile cells. Cells of splenic, lymphoid, and hematopoietic tissues are also labile cells. Stable cells retain the capacity for proliferation, although they do not normally replicate. These cells can undergo rapid division in response to a variety of stimuli and are capable of reconstitution of the tissue of origin. Stable cells include the parenchymal cells of all of the glandular organs of the body (e.g., liver, kidney, and pancreas), mesenchymal derivatives such as fibroblasts, smooth muscle cells, osteoblasts and chondrocytes, and vascular endothelial cells. Permanent cells are those which cannot reproduce themselves after birth. Examples are nerve cells.

Tissues comprised of labile and stable cells have the capability for regeneration after surgical trauma. The injured tissue is replaced by parenchymal cells of the same type, often leaving no residual trace of injury. However, tissues comprised of permanent cells are repaired by the production of fibrocollagenous scar. Despite the capability of many tissues to undergo regeneration, destruction of the tissue stroma, remaining after injury or constructed during the healing process, will lead to formation of scar. The biological response to materials thereby depends on the influence of the material on the inflammatory and reparative stages of wound healing. Does the material yield leachables or corrosion products that interfere with the resolution of inflammation initiated by the surgical trauma? Does the presence of the material interfere with the stroma required for the regeneration of tissue at the implant site? These are the types of questions that need to be addressed when assessing the "biocompatibility" of materials.

A number of systemic and local factors influence the inflammatory-reparative response. Systemic influences include age, nutrition, hematologic derangements, metabolic derangements, hormones, and steroids. While there is a prevailing "conventional wisdom" that the elderly heal more slowly than the young, there are few control data and animal experiments to support this notion. Nutrition can have a profound effect on the healing of wounds. Prolonged protein starvation can inhibit collagen formation, while high protein diets can enhance the rate of tensile strength gained during wound healing. Local influences that can affect wound healing include infection, inadequate blood supply, and the presence of a foreign body.

Fibrous Tissue Interface. The very presence of the implant provides a dead space in tissue that attracts macrophages to the implant–tissue interface (24). These cells are attracted to the prosthesis as they are to any dead space (e.g., bursa or joint space), presumably because of certain microenvironmental conditions (e.g., low O_2 and high lactate). In this regard it is not clear why macrophages are absent from the surface of osseointegrated implants (see below).

Macrophages along with fibroblasts of the scar comprise synovial tissue (25) that can be considered the chronic inflammatory response to implants (unless the device is apposed by osseous tissue, i.e., osseointegrated). This process is often termed "fibrous encapsulation" (26,27). The presence of regulatory cells such as macrophages at the implant-tissue interface can profoundly influence the host response to a device because these cells can release proinflammatory mediators if irritated by the movement of the device or substances released from the biomaterial (28). The inflammatory response of the synovial tissue around implants is comparable to the inflammation that can occur in the synovium lining any bursa (e.g., bursitis); hence, the response to implants has been termed "implant bursitis" (25).

Response to Implants in Bone: Osseointegration, Bone Ingrowth, Chemical Bonding of Bone to a Biomaterial. Wound healing governs the makeup of the tissue that forms around implants. Because of its capability for regeneration bone should be expected to appose implants in osseous tissue, and form within the pore spaces of porous coatings. Is this bone bonded in any way to the implant? Bonding of a prosthesis to bone would enhance its stability, limiting the relative motion between the implant and bone. In addition, bonding might provide a more favorable distribution of stress to surrounding osseous tissue.

Bonding of bone to an implant can be achieved by mechanical or chemical means., Interdigitation of bone with PMMA bone cement or with irregularities in implant topography, and bone ingrowth into porous surfaces, can yield interfaces capable of supporting shear and tensile as well as compressive forces. These types of mechanical bonding have been extensively investigated and are reasonably well understood. Chemical bonding of bone to materials could result from molecular (e.g., protein) adsorption–bonding to surfaces with subsequent bone cell attachment. This phenomenon has undergone intensive investigation in recent years but is not yet as well understood as mechanical bonding.

The term "osseointegration" has been used to describe the presence of bone on the surface of an implant with no histologically (light microscopy) demonstrable intervening nonosseous (e.g., fibrous) tissue. All implants in bone should become osseointegrated unless the bone regeneration process is inhibited.

The bone ingrowth into a porous-surface coating on an implant leads to an interlocking bond that can serve to stabilize the device. In order for the porous material to accommodate the cellular and extracellular elements of bone, the average pore diameter should be above $\sim$100 μm.

The bone ingrowth process proceeds in two stages. The surgical trauma of implantation initially leads to the regeneration of bone throughout the pores of the coating. Then mechanical stress-induced remodeling leads to resorption of bone from certain regions of the implant and continued formation and remodeling of bone in other regions.

Previous investigations have provided evidence of bone bonding to many different types of calcium phosphate

materials, calcium carbonate substances, and calcium-containing "bioactive" glasses. Chemical bonding was evidenced by the high strength of the implant-bone interface that could not be explained by a mechanical interlocking bond alone. In addition, TEM has shown that there is no identifiable border between these calcium-containing implants and adjacent bone.

Many recent studies have investigated the bonding of bone to one particular calcium phosphate mineral, hydroxyapatite, chosen because its relationship to the primary mineral constituent of bone; natural bone mineral is a calcium-deficient carbonate apatite. Experiments have been performed on both hydroxyapatite coated metallic implants and on particulate and block forms of the mineral employed as bone substitute materials. Histology of specimens from animals and retrieved from human subjects show that a layer of new bone $\sim$100 μm in thickness covers most of the hydroxyapatite surface within a few weeks of implantation and remains indefinitely. This layer of bone is attached to the surrounding osseous tissue by trabecular bridges.

In studying the mechanism of bone bonding, researchers have found that within days of implantation, biological apatites precipitate (from body fluid) onto the surface of the calcium-containing implants. These biological apatites are comparable to the carbonate apatite that is bone mineral. Proteins probably adsorb to this biological mineral layer thereby facilitating bone cell attachment and the production of osteoid directly onto the implant. This osteoid subsequently undergoes mineralization as it does normally in osteogenesis, thus forming a continuum of mineral from the implant to the bone. In this light, the bone cell responds to the biological apatite layer that has formed on the implant and not directly to the implant itself. Recent studies have shown that this biological apatite layer forms on many different calcium phosphate substances, explaining why bone bonding behavior has been reported for many different types of calcium phosphate materials. Of course, the clinical value of this phenomenon will depend (*1*) for coatings, on how well these substances can be bonded to implants; and (*2*) for bone graft substitute materials, their strength, modulus of elasticity, and ability to be resorbed. However, the finding that bone can become chemically bonded to certain biomaterials is a significant advance in our understanding of the implant-bone interface.

Effects of Implant-Induced Alterations of the Mechanical Environment. The presence of the implant can alter the stress distribution in the extracellular matrix (ECM), and thereby reduce or increase the strains experienced by the constituent cells. Many studies have demonstrated immobilization-induced atrophy of certain tissues resulting from the decrease in mechanical strains. Loss of bone mass around stiff femoral stems and femoral condylar prostheses of total hip and knee replacement devices has been associated with the reduced strains due to "stress shielding". Hyperplasia and hypertrophy of tissue in which mechanical strains have increased due to the presence of an implant have also been evidenced.

Criteria for Assessing Acceptability of the Tissue Response

The *in vivo* assessment of tissue compatibility of biomaterials requires that certain criteria be implemented for determining the acceptability of the tissue response relative to the intended application of the material–device. The biomaterial–device should be considered biocompatible only in the context of the criteria used to assess the acceptability of the tissue response. In this regard, every study involving the *in vivo* assessment of tissue compatibility should provide a working definition of biocompatibility. Biomaterials–devices implanted into bone can become apposed by the regenerating osseous tissue, and thus be considered compatible with bone regeneration. However, altered bone remodeling around the device due to stress shielding, with a net loss of bone mass (i.e., osteopenia), could lead to the assessment that the material–device is not compatible with normal bone remodeling. In situations in which the implant is surrounded by fibrous tissue the macrophages on the surface of the material are the expected response to the dead space produced by the very presence of the implant. The synovial tissue thus produced might be considered an acceptable response relative to the chemical compatibility of the material. Utilization of the thickness of the scar capsule around implants alone as a measure of biocompatibility is problematic because it can be influenced by movement of the tissue at the site relative to the implant.

The cellular and molecular make-up of tissue and the interactions among these components are complex. Therefore, criteria for assessing certain features of the biocompatibility of biomaterials–devices should focus on specific aspects of the biological response. The tissue compatibility of materials should be assessed specifically in the context of the effects of the material–device on certain aspects of the response. Moreover, it is important to note that materials yielding acceptable tissue compatibility in one site of implantation might yield unfavorable results in another site.

Degeneration of the Biomaterial–Tissue Interface

As noted earlier it is the wound healing response that initially establishes the tissue characteristics of the implant–tissue interface. Several agents have the potential for initiating degenerative changes in the interface tissue. Others probably act as promoters to stimulate the production of proinflammatory mediators that stimulate tissue degradation, and potentiate the failure process. Of the many factors affecting the implant–tissue interface, motion of the prosthetic component and particulate debris are two of the most important. However, it is difficult to determine the causal relationships between these factors and implant failure from only studying the end-stage tissue. Other histopathological findings and laboratory studies indicate that metal ions and immune reactions might play roles in the degenerative processes leading to prosthesis loosening in certain patients. Systemic diseases and drugs employed for the treatment of the disorders could also serve as factors contributing to the breakdown of the implant–bone interface. Finally, there might be interindividual differences in genetically determined cellular responses that could explain why prostheses fail in

some patients in whom there is a low mechanical risk factor for failure.

Effects of Implant Movement. Movement of the implant relative to the surrounding tissue can interfere with the wound healing response by disrupting the granulation tissue. In the case of implants in bone this relative movement, if excessive, can destroy the stroma required for osseous regeneration, and a fibrous scar results. Another important effect of implant motion is the formation of a bursa within connective tissue in which shearing and tensile movement has led to disruption of tissue continuity and led to the formation of a void space or sack (lined by synovial-like cells). It is to be expected, then, that tissue around prosthetic components removed due to loosening might display features of synovial-like tissue. The presence of synovial cells (macrophage and fibroblast-like cells) is important because they could be activated by other agents, such as particulate debris, to produce proinflammatory molecules. The process of activation of this tissue might be similar to that which occurs in inflammatory joint synovium or bursitis.

An explanation of how prosthetic motion leads to the formation of the synovial-like tissue can be found in previous studies (29) that have shown that "synovial lining is simply an accretion of macrophages and fibroblasts stimulated by mechanical cavitation of connective tissue". These findings are based on experiments in which the mechanical disruption of connective tissue was produced by injection of air and/or fluid into the subcutaneous space of animals (30). The resulting sack was initially described as a "granuloma pouch". Later studies (29) demonstrated that the membrane lining the pouch displayed the characteristics of synovium, and referred to this tissue as "facsimile synovium".

Prosthetic motion can also contribute to wear of the prosthetic component abrading against the bone cement sheath or surrounding bone, thereby generating increased amounts of particulate debris that might contribute to activation of the macrophages and synovial-like cells at the implant–tissue interface.

Effects of Implant-Derived Particles. Particulate debris can be generated from the abrasion of the implant against surrounding tissue. Understandably, the potential for wear is greater with materials–devices rubbing against a hard surface such as bone and with the articulating components of joint replacement prostheses. This particulate debris can induce changes in the tissue around the implants. Adverse responses have been found to both metallic and polymeric particles. The biological reactions to particles are related to (*1*) particle size, (*2*) quantity, (*3*) chemistry, (*4*) topography, and (*5*) shape. While it is not clear what role each of these factors play in the biological response, particle size appears to be particularly important. Particles small enough to be phagocytosed (<10 μm) elicit more of an adverse cellular response than larger particles.

Particulate metallic particles (viz., cobalt–chromium alloy particles) can induce rapid proliferation of macrophages and focal degeneration of synovial tissues (31). Because previous animal investigations and histopathological studies of tissues retrieved from human subjects have suggested that titanium alloy is more "biocompatible" than cobalt–chromium alloys it has been assumed that titanium particulate debris would be less problematic than particles of cobalt–chromium alloy. Histology of pigmented tissue surrounding titanium implants has generally revealed considerably fewer macrophages and multinucleated foreign body giant cells than seen around cobalt–chromium alloy particles and polymeric particulate debris. However, titanium alloy particles generated by the abrasion of femoral stems against bone cement in human subjects can cause histiocytic and lymphoplasmacytic reactions to the metallic particles (32). Titanium particles have also been found to cause fibroblasts in culture to produce elevated levels of PGE_2. These findings show that there may be adverse aspects of the biological response to titanium particles as well as to cobalt–chromium alloy particulate debris.

Many investigations evaluating the histological response to polyethylene and polymethylmethacrylate particles in animals and in tissue recovered from revision surgery have revealed the histiocytic response to these polymer particles. Moreover, it has been shown that this macrophage response can lead to bone resorption.

Synovial cells also respond to calcium-containing ceramic particles (33). Local leukocyte influx, proteinase, PGE_2, and tumor necrosis factor (TNF) levels have been measured after injection of calcium containing ceramic materials into the "air pouch model" described above. The TNF was detected in significant amounts after injection of the ceramics. These substances also provoked elevated leukocyte counts and levels of proteinase and PGE_2, showing that substances with surface chemistries that elicit a beneficial tissue response (e.g., bone bonding) when implanted in bulk form can cause destructive cellular reactions when present in particulate form.

Investigations indicate that most biomaterials, when present in particulate form in a size range small enough to be phagocytosed (<10 μm), can elicit a biological response that could cause the bone resorption that initiates and promotes the loosening process. This degenerative process has been referred to as "small particle disease".

Metallic Ions. Animal and human investigations have revealed elevated levels of metal ions in subjects with certain types of implants (viz., total joint replacement prostheses). Our knowledge is still incomplete with respect to the mechanisms of metal ion release. Results are often variable with respect to the concentration of specific metal ions in certain tissues and fluids. The fact that metal in ionic form is often not distinguished from that present as particles serves to confound interpretation of results.

Rises in serum and urinary chromium levels in patients who have undergo conventional cemented cobalt–chromium alloy hip replacement have previously been reported (34). However, an attempt to determine the valency of chromium as either $+3$ (III) or $+6$ (VI) from the concentration of metal ion in blood clot was not successful. This experiment was based on the fact that erythrocytes display

a unidirectional uptake of Cr(VI) while effectively excluding Cr(III). The distinction of the valency of chromium is important because Cr(VI) is much more biologically active than Cr(III).

Unfortunately, our knowledge of the local and systemic biological and clinical sequelae of metal ion release has not significantly advanced over the past several years. Addition of cobalt ions in the form of cobalt fluoride solutions to the media of synovial cells can stimulate their production of neutral proteinases and collagenase (35). These findings may be relevant to findings of tissue degradation (e.g., osteolysis) around implants in that metal ions could activate synovial cells in the surrounding synovial tissue to produce agents that promote tissue degeneration.

Diseases and Drugs. There has been little work correlating the failure of implants with disease states and drugs employed to treat the disorders. Some observations indicate that antiinflammatory agents, as well as certain anticancer drugs, can reduce the amount of bone formation around devices in the early stages of wound healing after implantation. Little is known, however, about the role of these and other agents on tissue remodeling and degeneration at the biomaterial-tissue interface.

Immune Reactions

It is not infrequent that two patients matched for sex, age, weight, activity level, and other factors, that might be expected to affect the performance of the prosthesis (implanted with the same device by same surgeon), have very different outcomes. This suggests that immune reactions, or genetically determined responses, might play a role in the failure of prostheses in some patients.

Immune responses include antibody and cell-mediated reactions and activation of the complement system. Certain small molecules released from implants (e.g., metal ions), while not antigenic themselves, can bind to existing antibodies and then to larger antigenic molecules or carrier proteins and subsequently elicit antibody production by activation of B lymphocytes by the small molecule (the "hapten") and by activation of helper T lymphocytes by the carrier protein to which it is bound.

The cell types that might be expected to occur at sites of antibody and cell-mediated reactions are not often found in tissue retrieved with revised devices. These cells include lymphocytes and plasma cells. The finding of occasional lymphocytic infiltrates in peri-implant tissue does not provide enough information for the role of immune reactions to implant. Immune reactions to polymeric materials (viz., silicone) have also been suggested as the cause of certain systemic diseases. However, mechanisms for such a response, and its prevalence, remain in question. Much more additional work is necessary to determine the role of immune reactions in the response to implantable devices.

The complement system, comprising circulating proteins and cell-surface receptors, plays an important role in immune processes engaged in the host defense against infectious agents. Te complement system consists of 20–30 proteins circulating in blood plasma. Most of these are inactive until they are cleaved by the chemical action of an enzyme of the interaction with a biomaterial surface. Once activated, the proteins can initiate a cascade of reactions resulting in the mobilization of immune cells resulting in inflammatory processes. Previous studies have demonstrated that many biomaterials can activate (cleave) certain molecules (C_3 and C_5) in the complement system and thereby stimulate the alternative pathway of the immune response. It has been suggested that complement activation by biomaterials could play a role in adverse reactions to certain devices. However, additional studies are required.

One form of cell-mediated immune reaction associated with implants, that has been studied, is the delayed hypersensitivity response. "Metal allergy" has been incriminated as the cause of failure in certain patients (36). However, results obtained to date are not definitive. "The incidence of metal sensitivity in the normal population is high, with up to 15% of the population sensitive to nickel and perhaps up to 25% sensitive to at least one of the common sensitizers Ni, Co, and Cr. The incidence of metal sensitivity reactions requiring premature removal of an orthopedic device is probably small (less than the incidence of infection). Clearly, there are factors not yet understood that caused one patient, but not another, to react" (37).

A similar situation exists with respect to sensitivity reactions to polymeric materials including bone cement (PMMA). The monomer of PMMA is a strong skin sensitizer (38). However, failure of cemented devices has not yet been correlated with a hypersensitivity response in patients.

The fact that there is no clear etiology of the prosthesis loosening in some patients while in other individuals with multiple risk factors for failure the prosthesis functions well has suggested that there may be genetic determinants for loosening.

Carcinogenicity

Chromium and nickel are known carcinogens and cobalt is a suspected carcinogen. Therefore, it is understandable that some concern might be raised about the release of these metal ions into the human body from implants. Fortunately, there have been few reports of neoplasms around implanted devices (e.g., total joint replacement prostheses). While no causal relationship has been evidenced, there is a high enough index of suspicion to warrant serious investigation of this matter through epidemiological and other studies. The use of porous coated metallic devices (with large surface area) in younger patients (e.g., noncemented total joint replacement prostheses) has added to concern about the long-term clinical consequences of metal ion release because of the significant increase in exposure of patients to metal ions.

Prior publications have reviewed the relationship of metallic ion release to oncogenesis (1), and reports of neoplasms found around orthopedic implants have been reviewed (39). The difference in the tumor types, time to appearance, and type of prosthesis confounds attempts to conclude an association of the neoplasm to the implant materials and released moieties.

In an epidemiological investigation conducted in New Zealand (40), >1300 total joint replacement patients were followed to determine the incidence of remote site tumors. The incidences of tumors of the lymphatic and hemapoietic systems were found to be significantly greater than expected in the decade following arthroplasty. It is important to note that the incidences of cancer of the breast, colon, and rectum were significantly less than expected. The investigators acknowledged that while the association might be due, in part, to an effect of the prosthetic implants, other mechanisms, particularly drug therapy, require consideration. Somewhat similar results were obtained from another recent study (41) of the cancer incidence in 443 total hip replacement (McKee–Farrar) patients operated on between 1967 and 1973 (followed to the end of 1981). The risk of leukemias and lymphomas increased while the risk of breast cancer decreased. The authors concluded that the local occurrence of cancer associated with prostheses made of cobalt–chromium–molybdenum as reported in the literature as well as animal experiments indicate that "chrome–cobalt–alloy plays some role in cancerogenesis (sic)".

In a recent publication (42), a nationwide cohort study performed in Sweden to evaluate cancer incidence among > 100,000 hip replacement patients found no overall cancer excess relative to the general population. The standardized incidence ratios (SIRs) were, however, elevated for prostate cancer and melanoma and reduced for stomach cancer risk. Long-term follow-up (> or = 15 years) revealed an excess of multiple myeloma. The study found no material increase in risk for bone or connective tissue cancer. The investigators noted that, while hip implant patients had similar rates of most types of cancer to those in the general population, excesses in certain types of cancers warranted further investigation, particularly because of the ever-increasing use of hip implants at younger ages.

BACTERIAL INFECTION

Biomaterial surfaces can provide favorable substrates for the colonization by bacteria. The adherence of bacteria to solid surfaces is facilitated by the their production of a "biofilm". The biofilm is a complex structure comprising bacterial cells encapsulated in a polymeric matrix. The detailed composition of the matrix has yet to be fully determined. Little is still known about how certain biomaterials may favor the production and adherence of a biofilm. Studies are still seeking to understand how certain material characteristics might favor bacterial colonization.

Infection following material implantation may be defined as multiplication of pathogenic microorganisms in the tissue of the host after a material implantation, causing disease by local cellular injury, secretion of a toxin, or antigen–antibody reaction in the host. The pathogenic microorganisms could be bacteria, fungi or viruses; the mostly common pathogenic microorganisms are bacteria.

Implant-related infection is one of the most serious and difficult complications to treat, often requiring reoperation including the surgical removal of the implant, and it may result in osteomyelitis, amputation, or even death. About 25–50% of infected vascular prostheses for cardiac, abdominal, and extremity vessel replacement cases result in amputation or death (43–45). Infectious complications are the principal concerns in the use and development of implanted materials for several indications.

The Biological Response to Bacterial Infection

We have noted above that the surgical trauma associated with the implantation of medical device is an injury that elicits an inflammatory response. Bacteria are another form of injurious agent that similarly elicits inflammation. While there are similarities in the cellular reactions to these two forms of injury there are important differences. That cells of the immune system are involves in the inflammatory reaction to bacterial infection is cause for use of the term "immune response" to describe the biological process elicited by a bacterial infection.

The immune response (also called the "immune reaction) is a defense function of the body that protects it against invading pathogens, foreign proteins, and malignancies. It consists of the "humoral immune response" and the "cell-mediated immune response". In the humoral immune response, B lymphocytes produce antibodies that react with specific antigens brought by invading pathogens, foreign proteins, and malignancies. The antigen–antibody reactions activate the complement cascade, which causes the lysis of pathogens or cells bearing those antigens. The humoral response may begin immediately on invasion by an antigen in acute type or up to days later in chronic type. In the cell-mediated immune response, T lymphocytes mobilize tissue macrophages in the presence of foreign antigen, which causes the pathogens or cells bearing those antigens been taken by phagocyte.

An implant-related infection occurs when an adequate number of a sufficiently virulent organism overcomes the host's immune response and establishes a focus of infection at the implant site. Implant-related infections remain a formidable challenge to the surgeon as well as material scientist. The high success rate obtained with antibiotic therapy in most bacterial diseases has not been obtained with implant-related infections for several known and as yet unidentified reasons. One important factor is that the sites on and around implants colonized by bacteria have little or no blood supply and thus do not allow blood-borne antimicrobial agents to reach the bacteria. The biofilm in which the bacteria grow may also shield the pathogens from the antimicrobial agents. Illness, malnutrition, and inadequacy of the immune system may be other factors that allow for the development implant-related infections.

Classification of Pathogens in Environment

Pathogens in the environment can be divided into three categories: primary pathogen, opportunistic pathogen, and nonpathogen. Primary pathogens are organisms that can cause infection in normal host when it has attached to the host's tissue and has gained sufficient numbers. It is also called a professional pathogen. Only a very small proportion of microbial species may be considered to be primary or

professional pathogens, and even among these species only a relatively small number of clones have been shown to cause infection. Pathogenic organisms are highly adapted to the pathogenic state and have developed characteristics that enable them to be transmitted, to attach to surfaces, to invade tissue, to escape host defenses, to multiply, and thus to cause infection.

Opportunistic pathogens are those organisms that can only cause infection in impaired hosts. For opportunistic pathogens, the state is the main determinant of whether infection will be the outcome of their interaction with the host's local tissue. This group of organisms may lack effective means to overcome an unimpaired host's defense mechanisms. They have limited growth opportunities outside their restricted niche in an unimpaired individual. As a result, infection may be only a rare consequence of the host–microbe encounter.

Nonpathogens are harmless members of the normal flora in healthy individuals. They may, however, in some rare situations, act as virulent invaders in an individual with severe deficiencies in host defense mechanisms.

The Most Common Bacteria that Cause Implant-Related Infection and Routes of Transmitting Pathogens

Studies of infected implants that have been retrieved for analysis indicate that a few species seem to dominate implant-related infections. Coagulase negative staphylococci are most frequently involved in the implant-related infections. Aerobic Gram-negative bacteria and anaerobic bacteria, which are usually present in deep infections (46), can also cause implant-related infection. *Staphylococcus aureus* (*S. aureus*) and *Staphylococcus epidermidis* (*S. epidermidis*) have been most frequently isolated from infected implant material surfaces. However, *Escherichia coli* (*E. coli*), *Pseudomonas aeruginosa*, β-hemolytic streptococci, and enterococci have also been isolated (43,47). These bacteria more often act as a component of mixed infections.

The different physical and chemical properties of implant material surfaces appear to be responsible for favoring infection with certain bacteria. *Staphylococcus aureus* is mostly involved in infection of metallic implants, such as metallic artificial joints, whereas *S. epidermidis* is a primary cause of infection of polymeric biomaterial implants, such as vascular grafts, catheters, and shunts (47).

The most pathogenic species is *S. aureus* because the infection caused by *S. aureus* often results in a much higher rate of mortality, and it is rarely cleared without removal of the implant. The recent emergence of *S. aureus* that are more resistant to all approved antibiotics raises more serious concerns for the future (48).

Implant-related infections may be the result of bacterial contamination of the implant material prior to its implantation. Pathogenic microorganisms may obtain access into the body by, direct contact, airborne spreading, contaminated water transmission, and blood stream transmission. If microorganisms exit in the host's tissue or on the skin or mucous membrane away from the implant site and break through blood barrier (e.g., associated dental treatment) they can gain access to the blood circulation. This can bring the microorganisms to the implant site (i.e., hematogenous infection).

Risk Factors for Implant-Related Infections

Implant-related infections occur when an adequate number of a sufficiently virulent organism overcomes the host's defense systems. During this process, many factors may be involved or even cooperate in establishing a focus of infection at the local implant area. The risk factors may involve the implant material, the process of implantation surgery, and the host.

Material-Related Factors. A large surface increases the possibility of microorganism attachment and thus can lay a role in implant-related infection. The avascular zone surrounding a device also favors infection as it contains tissue fluid and is often free of microorganism-monitoring agents because of the absence of blood circulation. A small initial number of microorganisms can grow to significant numbers and cause infection without interruption.

Sterile implant materials are commonly packaged in sterile paper, cloth, and plastic bags. However, all of these may be accidentally broken without notice and allow bacterial contamination of the implant.

Implantation-Related Factors. Implantation-related factors include the operating environment, skin and wound care, and surgical technique. In the operating room, airborne microorganisms (usually Gram-positive bacteria) are a source of wound contamination, originating with operating room personnel. Each person in the operating room sheds as many as 5000–55,0000 particles/min. Conventional operating room air may contain 10–15 bacteria/ft^3(49).

The microorganisms present on the host's skin are another source of wound contamination. Although the skin and hair can be sterilized with disinfectant agents, it is almost impossible to sterilize the hair follicles and sebaceous glands because the disinfectants now used in surgery do not penetrate an oily environment. Many disinfectants that do penetrate the oily environment, such as hexachlorophene, are absorbed by the body and have potentially toxic side effects. For this reason, skin preparations now used in surgery have a limited effect on sebaceous glands and hair follicles where microorganisms normally reside and reproduce (49). Because the skin can never be disinfected completely, the number of residual microorganisms present on the skin after disinfection builds the possibility of infection.

Any factor or event that delays wound healing increases the risk of implant-related infection. Ischemic necrosis, seroma, hematoma, wound infection, and suture abscesses are common preceding events for implant-related infection. Surgical technique and operating time also contribute to infection rates.

Host-Related Factors. Systematic factors that can contribute to implant-related infections include: immunological status, nutrition, chronic disease, and infection at a remote site or bacteremia caused by other reasons. A

deficiency in the host's defense mechanisms predisposes the host to infection by specific groups of opportunistic pathogens (49). Deficiencies in the immune system may be acquired (such as acquired immune deficiency syndrome, AIDS) or may result from congenital abnormalities. Malnutrition and chronic disease decreases both the immune and inflammatory response to microorganism invasion. Although the contaminating microorganisms may be few in number, the altered host's defense mechanisms implies that even small bacterial counts have to be regarded as highly virulent species.

If there is infection at a remote site in the host, the microorganisms can be brought to the implant site by blood stream and cause implant-related infection. Under several other conditions which the blood barrier are broken, such as a dental treatment, microorganisms are transported by the blood stream to find their way to the implant site, causing hematogenous infections (50,51). Infection of total hip arthroplasties after dental treatment is not rare (52).

The Most Common Feature of Implant-Related Infection: Biofilm

At the implant site, the surface of the material is immersed in the tissue fluid of the host's local tissue. If microorganisms appear, they have a strong tendency to colonize surfaces to form a microecosystem in which various microbial strains and species grow in a complex community-like structure, which is called biofilm.

Biofilms are defined as bacterial populations reside and produce in matrix that adheres to a surface, interface, or each other. During most implant-related infections, microbial products may assist the development and persistence of the infection in association with adsorbed macromolecules from the biological environment in which the implant material is placed. In the presence of implant material, bacteria elaborate a fibrous exopolysaccharide material called the "glycocalyx". The glycocalyx modifies the local environment in favor of the pathogen by hiding and protecting the organism from surfactants, antibodies, phagocytes, and antimicrobial agents. This increases the population of microorganisms on the surface of implant materials *in vivo* (53,54). These protective biofilms may act as bases in predisposing to tissue invasion and also result in the persistence of infection. Biofilms are implant-associated and troublesome. They have been reported to be 500 times more resistant to antibiotics than planktonic cells (55).

Growth of the organisms is the main mechanism of multiplication in a biofilm and eventually leads to the formation of a thick film. The biofilm is formed in three phases. The first phase is the formation of "conditioning film". As soon as the implantation of material performs, the material surfaces adsorbed macromolecules from the surrounding fluid, forming a conditioning film. The macromolecules are a number of extracellular proteins that interact with host intracellular matrix and blood proteins. For example, joint materials adsorb macromolecules from synovial fluid, bone materials adsorb macromolecules from plasma, while dental materials adsorb macromolecules from salivary fluid. The conditioning film forms within seconds of exposure of the implant to a body fluid (56). This conditioning film provides a suitable substrate for microorganism's adhesion.

The second phase is an initial, reversible adherence of microorganisms to the conditioning film. This adhesion depends on the physicochemical characteristics of the microbial cell surface, the material surface and the conditioning film. Microorganisms can reach the surface via various transport mechanisms, such as diffusion, convection or sedimentation (57). Implants can become contaminated before or during surgery, and more likely, by hematogenous seeding (58). Several factors are reported to contribute to this initial adherence, including surface hydrophobicity, proteinaceous adhesins, and capsular polysaccharides, such as fibrinogen, fibronectin, thrombospondin, von Willebrand factor, collagen, bone sialoprotein, and elastin (59). It seems that different bacteria is helped by different group of factors. *Staphylococcus aureus* appears to be enhanced by mostly fibronectin and plasma glycoprotein in adhering to polymethylmethacrylate *in vivo*, and this may contribute to the establishment of infection (60).

The third phase of biofilm development is microcolony formation and exopolymer production, which results in the firm anchoring of the biofilm and complex biofilm architecture. The adhered organisms multiply and form microcolonies and higher ordered structure glycocalyx. As soon as glycocalyx have been formed, the organisms gain some resistance. In favor of the protection, the microorganisms keep multiplying within this matrix. New layers of film are added and allowing the microorganisms room to multiply, forming thick biofilms (56). Biofilm protects the resident microorganisms against environmental attacks and antibiotics. However, the mechanism for resistance is not well understood.

As biofilms grow thicker and thicker, microorganisms on the periphery of the expanding biofilm may detach, which plays a large part in the pathogenesis of septic processes (61,62).

Latent Infections

As the biofilm protects microorganisms against environmental attacks and antibiotics, the microorganisms can survive in the biofilm for a long period of time when the host's defense system is strong or sufficient concentration of antibiotics is exit. There is no clinical symptom or sign of infection, but the microorganisms exit in the implant site. However, if the host's defense system becomes weakened or a sufficient concentration of antibiotics is no longer administered, the microorganisms may become more active, and cause a latent infection.

Latent infection also can be caused in other situations by a remote wound. The remote wound can give microorganisms access to the blood circulation. The blood stream can bring the microorganisms to the implant, causing a latent infection. Infection of total hip arthroplasty after dental treatment is not rare (52).

The Outcomes of Infection

Implant-related infection produces all the symptoms of infectious inflammation with a wide spectrum of severity.

The clinical presentation is determined largely by the virulence of the infecting pathogen, the extent of the area involved, the location of the infection and the nature of the infected host tissue. The infection may cause large changes to the host's internal environment as well as the implant material. As we mentioned in former section, *S. aureus* is a common pathogen in implant-related infection. In the mean while, it is particularly a very virulent pathogen in this setting and usually produces a fulminant infection.

The early stage of implant-related infection may be obvious or obscure. Signs and symptoms vary with the location and extent of tissue or organ involvement. Common characteristics of infection, such as fever, chills, nausea, vomiting, malaise, erythema, swelling, and tenderness may or may not be present. The classic triad is fever, swelling, and tenderness or pain. Tenderness or pain probably is the most common and earliest symptom. Swelling may be mild. Fever is not always a consistent finding.

During the mid-late stage of infection, the severity of the infection, its specific microorganism, and the particular tissue, site, and material involved all introduce morphologic variations in the basic patterns of acute and chronic infection. The implant-related infection can appear as serous inflammation, fibrinous inflammation, suppurative or purulent infection, abscesses, or more seriously lead septicemia, septic shock, or patient death.

Serous inflammation is marked by the outpouring of a thin fluid derived from the blood serum. Fibrinous inflammation is a fibrinous exudate develops when the vascular leaks are large enough. These two are the clinical appearance of mild infection. Fibrinous exudate may convert to scar tissue if the infection is controlled in this stage (63).

Suppurative or purulent infection is commonly seen in implant-related infection. It is characterized by the production of large amounts of pus or purulent exudate consisting of neutrophils, necrotic cells, and tissue fluid. Certain pyogenic (pus-producing) microorganisms (e.g., staphylococci) produce this localized suppuration.

Abscesses are focal localized collections of purulent inflammatory tissue caused by suppuration buried in a tissue. They are produced by pyogenic bacteria. Abscesses have a central region that appears as a cavity full of pus that consists of necrotic tissue, died white blood cells, bacteria, and material. There is usually a zone of neutrophils around this necrotic focus. Vascular dilation, parenchymal and fibroblastic proliferation occurs outside this region, indicating the beginning of repair. Sometimes, the abscess may become walled off by connective tissue that limits it from further spread (64).

Microorganisms on the periphery of the expanding biofilm may detach or separate. These microorganisms may present in blood and can be confirmed by blood culture, which is called bacteremia. If the bacteria are strong enough to survive in blood and produce toxin, it is called septicemia, which can be life-threatening.

Diagnosis of Implant-Related Infection

The specific diagnosis of implant-related infection is dependent, in large part, upon isolation of the pathogen by aspiration of secreted fluid or by culture of tissue obtained at debridement. However, there are other assessments that can indicate an infection, such as blood count and different morphologic examinations.

Roentgenographic studies are helpful in implant-related infections. Plain roentgenograms show soft tissue swelling, joint space narrowing or widening, bone destruction and non-X-ray transparent implant materials. These roentgenograms can reveal (*1*) abnormal lucencies at the material–host tissue interface, (*2*) bone or periosteal reaction, (*3*) motion of components on stress views, or (*4*) changes in the position of implant materials. If initial roentgenograms are normal in the evaluation of a suspected implant-related infection, other imaging modalities that show soft tissue swelling and loss of normal fat planes about the involved site should be used.

Computed tomography (CT) scanning can help determine the extent of surrounding tissue involvement. Pus within the cavity can cause an increased density on the CT scan. Adjacent soft tissue abscesses also are easily seen. However, the use of CT scan is limited if the implanted material is made of metal.

Magnetic resonance imaging (MRI) can also be implemented for evaluating implant-related infections. The images can reflect the increase in water content resulting from edema in the implant or surrounding tissue due to infection. The MRI detects changes much earlier in the course of disease than roentgenograms, because it shows the condition of the surrounding soft tissue.

Fate of Material During Infection

The biofilm and bacterial modification of the microenvironment around implants may affect the biomaterial in several ways. Biodegradable materials, such as collagen, may degrade faster than expected due the elevated levels of enzymes and the changes in pH. Such implants may collapse before being replaced by host tissue and thus become a component of the local abscess. The low pH often found at sites of infection may also accelerate the corrosion of metallic materials. Implant-related infections indirectly affect implant materials by causing the destruction of surrounding tissue thus contributing to loosening of the implant.

Treatments of Implant-Related Infection

Successful treatment of an implant-related infection often depends on both extensive and meticulous surgical debridement and effective antimicrobial therapy. Debridement should be emphasized because infection often persists despite treatment with systemic antibiotic therapy in the absence of extensive and meticulous debridement of the implant-tissue interface.

It is of paramount importance to confirm the microorganism causing the infection. Distinguishing infection from pure inflammation is also very important because they have some similarities. The timing and selection of bacteria culture are critical. Many implant-related infections are deep seated, and adequate culture specimens are difficult to obtain. In spite of this, every effort should be made to obtain a culture specimen. The preferred specimen in implant-related infections is aspirated fluid. A deep

wound biopsy or a curetted specimen after cleaning the wound is acceptable.

Antibiotic therapy should begin as early as implant-related infection is diagnosed. Treatment with systemic antibiotic therapy can at least prevent the infection from transmission. In the meanwhile, local symptoms and signs of infection should be observed carefully. If there are signs show the infection is not under control, a debridement may be indicated.

Treatment of an implant-related infection at mid–late stage may require both systemic antimicrobial treatment and local surgical treatment. Antibiotic treatment alone sometimes may still be sufficient at this stage, however, it should be performed under careful observation and cannot last long. If there are signs that the infection is not under control, surgery is needed.

Surgery may go hand in hand with antibiotic treatment. The purpose of surgery is clearance of the necrotic tissue with the bacteria and augmentation of the host response. Debridement and irrigation removes necrotic and avascular tissue, bacteria, and harmful bacterial products. It is essential when pus is found on aspiration, signifying an abscess, or when roentgenographic changes indicating pus, necrotic material, and chronic inflammation. If an abscess has formed, removal of the implant is indicated.

Frequently, the only way to treat an infected large implant is to remove it in associate with antibiotic treatment.

If the infection is very severe, septic shock may exist and threaten the patient's life. At this situation, the most important work is antishock and save life. Supported by antishock and antibiotic treatment, removal of implant and open debridement, or even amputation must be performed.

How to Prevent Implant-Related Infection

Recognizing the unique characteristics and outcomes of implant-related infections, the best course is prevention. The close relationship and cooperation of implant designer, manufacturer, and surgeon are necessary for prevention of implant-related infection. The implant must be kept free of bacteria, while the surgeon should evaluate the risk of infection in each patient by considering both host- and surgeon-dependent factors. Simply stated, it is much easier to prevent an implant-related infection than to treat it.

Sterilization Methods. There are several methods for sterilizing implants prior to implantation. The first concern when choosing a sterilization method is the physical and chemical properties of the implant material itself as well as the packaging material required to maintain the implant sterile prior to delivery to the operative site. Autoclaving is the method of choice for the sterilization of metallic or heat-resistant implants. The advantages of autoclaving are efficacy, speed, process simplicity, and no toxic residues. The disadvantages are the relatively high temperature of the process (121 °C) may damage some non-heat resistant implant materials as well as the packaging materials. Thus, most nonmetallic implants and packaging materials cannot be sterilized by this method.

Ethylene oxide (EtO) gas sterilization is a low temperature sterilizing process. It is compatible with a wide range of implant and packaging materials. It is commonly used to sterilize a wide range of medical implants, including surgical sutures, absorbable and nonabsorbable meshes, absorbable bone repair devices, heart valves, and vascular grafts. The advantages of EtO are its efficacy, high penetration ability, and compatibility with a wide range of materials. The main disadvantage is EtO residuals in the sterilized materials.

Exposure to ^{60}Co gamma rays is another widespread sterilizing method. Gamma rays have a high penetrating ability. This method of sterilization is widely used for medical products, such as surgical sutures and drapes, syringes, metallic bone implants, knee and hip prostheses. The advantages of ^{60}Co gamma-ray sterilization are efficacy, speed, process simplicity, no toxic residues. The main disadvantages are the very high costs and incompatibility of some radiation sensitive materials such as the fluoropolymer and polytetrafluoroethylene (PTFE).

Medical implant may also be sterilized with machine-generated accelerated electrons called electron beam sterilization. It has a similar range of applications and material compatibility characteristics as the ^{60}Co process. However, the main disadvantages are short penetration distance. This limits its usage. A unique application for this method is the on-line sterilization of small, thin materials immediately following primary packaging.

Several new technologies are emerging that have potential utility for implant material sterilization, such as gaseous chlorine dioxide, low temperature gas plasma, gaseous ozone, vapor-phase hydrogen peroxide, and machine-generated X rays. Machine-generated X rays have the advantage of a nonisotopic source and penetrating power similar to gamma rays.

Prevention of Operative Contamination, Wound Sepsis Contiguous to the Implant, and Hematogenous Infection. The importance of irrigation during and at the end of surgery has been well documented. The principles of no dead space, no avascular tissue, evacuation of hematomas, and soft tissue coverage should be practiced strictly during implantation surgery. Good surgical technique and minimal operating times also contribute to lowering of infection rates. Prophylactic antibiotics are definitely indicated when implants are involved. New methods are on the way of developing. Direct local delivery of polyclonal human antibodies to abdominal implant sites reduced infection severity and mortality in an animal model of implant-related peritoneal infection (65).

Any bacteremia can induce an implant-related infection by the hematogenous route (51,66,67). Dentogingival infections and manipulations are known causes of streptococcal and anaerobic infections in prostheses (51). Pyogenic skin processes can cause staphylococcal and streptococcal infections of joint replacement. Genitourinary and gastrointestinal tract procedures or infections are associated with Gram-negative bacillary, enterococcal, and anaerobic infections of prostheses (66,68). Twenty-to-forty percent of prosthetic joint infections are caused by the hematogenous route (68).

To prevent hematogenous implant-related infection, any factor that might predispose to infection should be avoid before insertion of the implant material. For elective implant surgery, the patient should be evaluated for the presence of pyogenic dentogingival pathology, skin, and other even very small local infection. Perioperative antibiotic prophylaxis is also very important. It has been reported to reduce infections in total joint replacement surgery (69).

For patients with indwelling implant materials, it is of paramount important to diagnose and treat infections in any location as early as possible. This reduce the risk of seeding bacteria to the implant materials hematogenously in large extent. Any situation likely to cause bacteremia, even a dental care should be avoided or seriously evaluated.

Antimicrobial Biomaterials

Two factors necessary for an implant-related infection are attachment of bacteria to the implant surface and multiplication of bacteria to a significant number. Antibacteria materials are designed to decrease the surface attachment of microorganism and/or inhibit the multiplying of microorganism.

A variety of antibiotics incorporating implant materials were designed to inhibit the microorganism against multiplying. Antibiotics for incorporation in materials should have a broad antibacterial spectrum, sufficient bactericidal activity, high specific antibacterial potency, low rate of primary resistant pathogens, minimal development of resistance during therapy, low protein binding, low sensitizing potential, marked water solubility, and stable (70). During these years, various antibiotics have been evaluated both *in vitro* and *in vivo*. The studies are most regarding their suitability for incorporation in materials.

Cefazolin loaded bone matrix gelatin (C-BMG) was made from putting cefazolin into BMG by vacuum adsorption and freeze-drying techniques (71). It was tested for repair of long segmental bone defects and preventing infection in animal experiment. The effective inhibition time to staphylococcus aureus of C-BMG was 22 days in vitro, while 14 days in vivo. The drug concentration in local tissues (bone and muscle) were higher than that of plasma, and the drug concentration in local tissues was higher in early stage, later it kept stable low drug release.

Rifampin-bonded gelatin-sealed polyester was tested in another animal experiment (72). Their results indicated that rifampin-bonded gelatin-sealed polyester grafts were significantly more resistant to bacteremic infection than were silver/collagen-coated polyester grafts.

For developing antibiotics delivery biodegradable materials, polylactide–polyglycolide copolymers were mixed with vancomycin (73). The mixture was compressed and sintered at 55 °C to form beads of different sizes. The biodegradable material released high concentrations of antibiotic *in vitro* for the period of time needed to treat infection. The diameter of the sample inhibition zone ranged from 6.5 to 10 mm, which is equivalent to 12.5–100% of relative activity. By changing the processing parameters, the release rate of the beads was able to be controlled. This provides advantages of meeting the specific requirement for prevention of implant-related infection.

Besides incorporating antibiotics in implant materials, antimicrobial materials have been made in different ways. Bovine serum albumin was used to coat material surfaces by using carbodiimide, a cross-linking agent (74). The inhibition rate of the albumin coating on bacterial adherence remained high throughout the experiment. This suggests the potential use of this cross-linked albumin coating to reduce bacterial adherence and thus the subsequent possibility of prosthetic or implant infection *in vivo*.

In the future, the use of implant materials will surely increase with growing demands for a higher quality of life. In 1997, operating expenses allocated to tissue engineering exceed $450 million and fund the activities of nearly 2500 scientists and support personnel. Growth rate is 22.5% / annum (75). At the beginning of 2001, operating expenses allocated to tissue engineering exceed $600 million and fund the activities of nearly 3300 scientists and support personnel. Spending by tissue engineering firms has been growing at a compound annual rateof 16% (76). However, implant-related infection remains a significant problem in this field. Research on the development of biomaterial surfaces with antimicrobial properties has increased to an annual expenditure of ~$430 million (75).

BIBLIOGRAPHY

1. Black J. Metallic ion release and its relationship to oncogenesis. In: Fitzgerald RHJ, editor. The Hip, Proceedings of the Thirteenth Open Scientific Meeting of the Hip Society. St. Louis: C.V. Mosby; 1985. p 119–213.
2. Bartolozzi A, Black J. Chromium concentrations in serum, blood clot and urine from patients following total hip arthroplasty. Biomaterials 1985;6:2–8.
3. Takamura K, Hayashi K, Ishinishi N, Yamada T, Sugioka Y. Evaluation of carcinogenicity and chronic toxicity associated with orthopedic implants in mice. J Biomed Mater Res 1994;28:583–589.
4. Bouchard PR et al. Carcinogenicity of CoCrMo (F-75) implants in the rat. J Biomed Mater Res 1996;32:37–44.
5. Lombardi AV, Mallory TH, Vaughn BK, Drouillard P. Aseptic loosening in total hip arthroplasty secondary to osteolysis induced by wear debris from titanium-alloy modular femoral heads. J Bone Jt Surg 1989;71A:1337.
6. Blaine TA et al. Increased levels of tumor necrosis factor-a and interleukin-6 protein and messenger RNA in human peripheral blood monocytes due to titanium particles. J Bone Jt Surg 1996;78-A:1181–1192.
7. Gonzales JB, Purdon MA, Horowitz SM. In vitro studies on the role of titanium in aseptic loosening. Clin Orthop 1996;330:244–250.
8. Goodman SB, Fornasier VL, Lee J, Kei J. The effects of bulk versus particulate titanium and cobalt chrome alloy implanted into the rabbit tibia. JBMR 1990;24:1539–1549.
9. Donkerwolcke M, Burny F, Muster D. Tissues and bone adhesives—historical aspects. Biomaterials 1998;19:1461–1466.
10. Nivbrant B, Karrholm J, Rohrl S, Hassander H, Wesslen B. Bone cement with reduced proportion of monomer in total hip arthroplasty: preclinical evaluation and randomized study of 47 cases with 5 years' follow-up. Acta Orthop Scand 2001;72:572–584.

11. de la Torre B et al. Biocompatibility and other properties of acrylic bone cements prepared with antiseptic activators. J Biomed Mater Res 2003;66B:502–513.
12. Thomson LA, Law FC, James KH, Matthew CA, Rushton N. Biocompatibility of particulate polymethylmethacrylate bone cements: a comparative study *in vitro* and *in vivo*. Biomaterials 1992;13:811–818.
13. Davis RG, Goodman SB, Smith RL, Lerman JA, Williams RJ., III The effects of bone cement powder on human adherent monocytes/macrophages *in vitro*. J Biomed Mater Res 1993;27: 1039–1046.
14. Jones LC, Hungerford DS. Cement Disease. Clin Orthop Rel Res 1987;225:192–206.
15. LeVier RR, Harrison MC, Cook RR, Lane TH. What is silicone? Plas Reconstr Surg 1992;92:163–167.
16. Bobyn JD, Spector M. Polyethylene. In: Encycolpedia of Materials Science and Engineering. New York: Pergamon Press; 1987. p 3649.
17. Li S, Burstein AH. Ultra-high molecular weight polyethylene. J Bone Jt Surg 1994;76-A:1080–1090.
18. Schmalzried TP, Jasty M, Harris WH. Periprosthetic bone loss in total hip arthroplasty: polyethylene wear debris and the concept of the effective joint space. J Bone Jt Surg 1992;74-A:849–863.
19. Green TR, Fisher J, Stone M, Wroblewski BM, Ingham E. Polyethylene particles of a 'critical size' are necessary for the induction of cytokines by macrophages in vitro. Biomaterials 1998;19:2297–2302.
20. Vert M, Pascal C, Chabot F, Leray J. Bioresorbable plastic materials for bone surgery. In: Hastings GW, Ducheyne P, editors. Macromolecular Biomaterials. Vol. Chap. 5 Boca Raton: CRC Press, Inc.; 1984. p 119–142.
21. Yannas IV. Natural materials. In: Ratner BD, Hoffman AS, Schoen FJ, Lemons JE, editors. Biomaterials Science: An Introduction to Materials in medicine. San Diego: Academic Press; 1992. p 84–94.
22. Spector M, Lalor PA. *In vivo* assessment of tissue compatibility. In: Ratner BD, Hoffman AS, Schoen FJ, Lemons JE, editors. Biomaterials Science: An Introductory Text. San Diego; CA: Academic Press; 1996. p 220–227.
23. Spector M, Cease C, Xia T-L. The local tissue response to biomaterials. CRC Crit Rev Biocompat 1989;5:269–295.
24. Silver IA. The physiology of wound healing. In: Hunt TK, editor. Wound Healing and Wound Infection. New York: Appleton-Century-Crofts; 1984. p 11–28.
25. Spector M. et al. Synovium-like tissue from loose joint replacement prostheses: comparision of human material with a canine model. Sem Arthr Rheum 1992;21:335–344.
26. Coleman DL, King RN, Andrade JD. The foreign body reaction: a chronic inflammatory response. J Biomed Mater Res 1974;8:199–211.
27. Laing PG, Ferguson AB, Hodge ES. Tissue reaction in rabbit muscle exposed to metallic implants. J Biomed Mater Res 1967;1:135–149.
28. Anderson JM, Miller K. M. Biomaterial biocompatibility and the macrophage. Biomaterials 1984;5:5–10.
29. Edwards JCW, Sedgwick AD, Willoughby DA. The formation of a structure with the features of synovial lining by subcutaneous injection of air: an in vivo tissue culture system. J Pathol 1981;134:147–156.
30. Selye H. Use of "granuloma pouch" technic in the study of antiphlogistic corticoids*. Proc Soc Exp Biol Med 1953;82: 328–333.
31. Howie DW, V-Roberts B. The synovial response to intraarticular cobalt-chrome wear particles. Clin Orthop 1988;232: 244–254.
32. Agins HJ et al., Metallic wear in failed titanium-alloy total hip replacements. J Bone Jt Surg 1988;70A:347–356.
33. Nagase M, Baker DG, Schumacher, Jr., HR. Prolonged inflammatory reactions induced by artificial ceramics in the rat air pouch model. J Rheum 1988;15:1334–1338.
34. Bartolozzi A, Black J. Chromium concentrations in serum, blood clot and urine from patients following total hip arthoplasty. Biomaterials 1985;6:2–8.
35. Ferguson GM, Watanabe S, Georgescu HI, Evans CH. The synovial production of collagenase and chondrocyte activating factors in response to cobalt. J Orth Res 1988;6:525–530.
36. Merritt K, Rodrigo JJ. Immune response to synthetic materials. Clin Orthop 1996;326:71–79.
37. Merritt K, Brown SA. Biological effects of corrosion products from metals. In: Fraker A, editor. Vol. STP 859 Corrosion and Degradation of Implant Material. Philadelphia: American Society for Testing and Materials; 1985. p 195–207.
38. Merritt K. Role of medical materials, both in implant and surface applications, in immune response and in resistance to infection. Biomaterials 1984;5:47–53.
39. Martin A, Bauer TW, Manley MT, Marks KE. Osteosarcoma at the site of total hip replacement. JB J S 1988;70A:1561–1567.
40. Gillespie WJ, Frampton CMA, Henderson RJ, Ryan PM. The incidence of cancer following total hip replacement. JB J S 1988;70B:539–542.
41. Visuri T, Koskenvuo M. Cancer risk after McKee-Farrar total hip replacement. Acta Orthop Scand 1989;60:25.
42. Signorello LB et al. Nationwide study of cancer risk among hip replacement patients in Sweden. J Natl Cancer Inst 2001;93:1405–1410.
43. Gristina AG. Implant-associated infection. In: Ratner BD, Schoen FJ, Lemons JE, editors. Biomaterials science: an introduction to materials in medicine. San Diego: Academic Press; 1996.
44. Gristina AG, Oga M, Webb LX, Hobgood CD. Adherent bacterial colonization in the pathogenesis of osteomyelitis. Science 1985;228:990–993.
45. Gristina AG, Costerton JW. Bacterial adherence to biomaterials and tissue. The significance of its role in clinical sepsis. J Bone Joint Surg Am 1985;67:264–273.
46. Fitzgerald RH, Jr., Infected Total Hip Arthroplasty: Diagnosis and Treatment. J Am Acad Orthop Surg 1995;3:249–262.
47. Gristina AG. Biomaterial-centered infection: microbial adhesion versus tissue integration. Science 1987;237: 1588–1595.
48. Proctor RA. Toward an understanding of biomaterial infections: a complex interplay between the host and bacteria. J Lab Clin Med 2000;135:14–15.
49. William C. General Principles of Infection. Campbell's Operative Orthopaedics. Mosby, Inc.; 1998.
50. Stinchfield FE, Bigliani LU, Neu HC, Goss TP, Foster CR. Late hematogenous infection of total joint replacement. J Bone Joint Surg Am 1980;62:1345–1350.
51. Lindqvist C, Slatis P. Dental bacteremia–a neglected cause of arthroplasty infections? Three hip cases. Acta Orthop Scand 1985;56:506–508.
52. LaPorte DM, Waldman BJ, Mont MA, Hungerford DS. Infections associated with dental procedures in total hip arthroplasty. J Bone Joint Surg Br 1999;81:56–59.
53. Costerton JW, Irvin RT, Cheng KJ. The bacterial glycocalyx in nature and disease. Annu Rev Microbiol 1981;35:299–324.
54. Gristina AG, Kolkin J. Current concepts review. Total joint replacement and sepsis. J Bone Joint Surg Am 1983;65:128–134.
55. Costerton JW, Lewandowski Z, Caldwell DE, Korber DR, Lappin-Scott HM. Microbial biofilms. Annu Rev Microbiol 1995;49:711–745.

56. Barry D. Infected orthopedic prostheses. Infections associated with indwelling medical devices. Washington, DC: ASM Press; 1994.
57. van Loosdrecht MC, Lyklema J, Norde W, Zehnder AJ. Influence of interfaces on microbial activity. Microbiol Rev 1990;54:75–87.
58. Schmalzried TP, Amstutz HC, Au MK, Dorey FJ. Etiology of deep sepsis in total hip arthroplasty. The significance of hematogenous and recurrent infections. Clin Orthop 1992; 200–207.
59. Patti JM, Allen BL, McGavin MJ, Hook M. MSCRAMM-mediated adherence of microorganisms to host tissues. Annu Rev Microbiol 1994;48:585–617.
60. Vaudaux P, Suzuki R, Waldvogel FA, Morgenthaler JJ, Nydegger UE. Foreign body infection: role of fibronectin as a ligand for the adherence of Staphylococcus aureus. J Infect Dis 1984;150:546–553.
61. Neu TR, Marshall KC. Bacterial polymers: physicochemical aspects of their interactions at interfaces. J Biomater Appl 1990;5:107–133.
62. Busscher HJ, Van der Mei HC. Relative importance of surface-free energy as a measure of hydrophobicity in bacterial adhesion to solid surfaces. In: Doyle RJ, Rosenberg M, editors. Microbial cell surface hydrophobicity. Washington DC: American Society for Microbiology; 1990. p 335–359.
63. Mandell B, Dolin. Principles and Practice of Infectious Disease. Churchill Livingstone; 2000.
64. Kuma C. Robbins Patholog. Basis Disease, 1999.
65. Poelstra KA et al. Prophylactic treatment of gram-positive and gram-negative abdominal implant infections using locally delivered polyclonal antibodies. J Biomed Mater Res 2002;60:206–215.
66. Ahlberg A, Carlsson AS, Lindberg L. Hematogenous infection in total joint replacement. Clin Orthop 1978; 69–75.
67. Lattimer GL, Keblish PA, Dickson TB, Jr., Vernick CG, Finnegan WJ. Hematogenous infection in total joint replacement. Recommendations for prophylactic antibiotics. JAMA 1979;242:2213–2214.
68. Inman RD, Gallegos KV, Brause BD, Redecha PB, Christian CL. Clinical and microbial features of prosthetic joint infection. Am J Med 1984;77:47–53.
69. Norden CW. A critical review of antibiotic prophylaxis in orthopedic surgery. Rev Infect Dis 1983;5:928–932.
70. Wahlig H, Dingeldein E. Antibiotics and bone cements. Experimental and clinical long-term observations. Acta Orthop Scand 1980;51:49–56.
71. You HB, Chen AM. The effect of cefazolin loaded bone matrix gelatin on repairing large segmental bone defects and preventing infection after operation, Zhongguo Xiu Fu Chong Jian Wai Ke Za Zhi 2000;14:162–165.
72. Goeau-Brissonniere O.A. et al. Comparison of the resistance to infection of rifampinbonded gelatin-sealed and silver/collagen-coated polyester prostheses. J Vasc Surg 2002;35: 1260–1263.
73. Liu SJ et al. In vitro elution of vancomycin from biodegradable beads. J Biomed Mater Res 1999;48:613–620.
74. An YH et al. Prevention of bacterial adherence to implant surfaces with a crosslinked albumin coating in vitro. J Orthop Res 1996;14:846–849.
75. Lysaght MJ, Nguy NA, Sullivan K. An economic survey of the emerging tissue engineering industry. Tissue Eng 1998;4: 231–238.
76. Lysaght MJ, Reyes J. The growth of tissue engineering. Tissue Eng 2001;7:485–493.

See also ALLOYS, SHAPE MEMORY; POLYMERIC MATERIALS; POROUS MATERIALS FOR BIOLOGICAL APPLICATIONS; RESIN-BASED COMPOSITES.

BIOELECTRODES

ERIC MCADAMS
University of Ulster at Jordanstown
Newtownabbey,
Ireland

INTRODUCTION

Biomedical electrodes are used in various forms in a wide range of biomedical applications, including:

1. The detection of bioelectric events such as the electrocardiogram (ECG).
2. The application of therapeutic impulses to the body [e.g., cardiac pacing and defibrillation and transcutaneous electrical nerve stimulation (TENS)].
3. The application of electrical potentials in order to facilitate the transdermal delivery of ionized molecules for local and systemic therapeutic effect (iontophoresis).
4. The alternating current (ac) impedance characterization of body tissues.

Good electrode design is not as simple and straightforward a matter as is often assumed, and all electrode designs are not equal in performance (1). One must, therefore, not simply choose an electrode with as conductive a metal plate as possible, which unfortunately, was and appears to still be the case in many designs. Probably due to this mistaken view, it would appear that the associated electronic systems are often first developed and the electrode design is left to the end, almost as an afterthought. If the clinician is to properly diagnose the patient's cardiac problem, for example, it is imperative that the measured biosignal is clear, undistorted, and artefict-free. Unfortunately, monitoring bioelectrodes, if they are not chosen correctly, give rise to significant problems that make biosignal analysis difficult, if not impossible. Similarly, stimulation electrodes must be well-chosen if they are to optimally supply the therapeutic waveforms without causing trauma to the patient.

Current or charge is carried by ions inside the patient's body and by electrons in the electronic device itself and in its leads. The "charge-transfer" mechanism between current/charge carriers takes place at the electrode-patient interface and is of major importance in the design of an optimal electrode. Both the electrode–electrolyte interface and the skin under the electrode (collectively known as the contact) give rise to potentials and impedances that can distort the measured biosignal or adversely affect the electrotherapeutic procedure.

Implanted electrodes are generally made from inert or noble materials that do not react with surrounding tissues. Unfortunately, as a consequence, they tend to give rise to large interface impedances and unstable potentials. Implanted biosignal monitoring electrodes, in particular, require stable potentials and low interface impedances to minimize biosignal distortion and artifact problems. External biosignal-monitoring electrodes can generally use

high electrical performance nonnoble materials such as silver–silver chloride without fear of biocompatibility problems (2). They do, however, have to address the additional and very significant problem of the skin with its sizeable impedance and unstable potential. Along with the desired biosignal, one amplifies the difference between the two contact potentials. If the contact potentials were identical (highly improbable), they would cancel each other out due to the use of a differential amplifier. If the potential mismatch were very large (several hundred mV), the amplifier would not be able to cope and would saturate. If the mismatch in contact potentials is small and stable, this mismatch will be amplified along with the biosignal, and the biosignal will appear shifted up or down on the oscilloscope screen or printout paper, which would generally not be a major problem as the additional voltage offset can be easily removed. What is a significant problem, however, is when the contact potentials fluctuate with time. Their mismatch, therefore, varies and the baseline of the biosignal is no longer constant, which leads to the problem termed baseline wonder or baseline drift, which makes analysis of some of the key features of the biosignal difficult. Filtering out of the drift is often not an option, as the filtering often also removes key components of the desired biosignal.

Large mismatched contact impedances can cause signal attenuation, filtering, distortion, and interference in biosignal monitoring. If contact impedances are significant compared with the input impedance of the amplifier, they can give rise to signal attenuation as a result of the voltage divider effect. Attenuation of the signal is not a major problem, after all, the amplifier is going to be used to amplify the signal by a factor of around 1000 (in the case of an ECG). A significant problem develops, however, because the contact impedance varies with frequency. The frequency-dependence of the contact impedance is a consequence of the presence of parallel capacitances at the electrode-electrolyte interface or at the skin under the electrode. At very high frequencies, the contact impedances are very small and, therefore, no attenuation of the high frequency parts of the biosignal exists. At low frequencies, the contact impedances can be very large and, hence, significant attenuation of low frequency components of the biosignal can exist. The overall signal is not only attenuated, it is also distorted with its low frequency components selectively reduced. The measurement system in effect acts as a high pass filter and the signal is differentiated. In the case of the ECG, the P, S, and T waves are deformed, leading in particular to a modification of the S–T segment. The S–T segment is of vital importance to the electrocardiologist, hence the importance of avoiding such biosignal distortions.

50/60 Hz interference can be amplified along with any monitored biosignal due to the mismatch of the contact impedances. Displacement currents flow from power lines through the air to the monitor cables and then through the electrodes and the patient to ground. If the contact impedances are not identical, the displacement currents flowing through the two contact impedances connected to a differential amplifier will give rise to different voltages at the amplifier's inputs. This 50 Hz offset voltage will be amplified along with the desired biosignal and its amplitude is proportional to an electrode–skin impedance mismatch (3).

Other applications, such as electrical impedance plethysmography and electrical impedance tomography (4), do not monitor intrinsic biosignals emanating from the body, but inject small currents or voltages into the body and record the resultant voltages or currents. The electrical properties of the body or a body segment can then be calculated. In many of these applications, the magnitude and mismatch of contact impedances can give rise to significant errors or artifacts (5). As relatively high frequencies are often involved in these techniques, even the series resistance of the gel pad (which is generally ignored) may become significant.

Although interface impedance and potential are generally less critical for implanted stimulation electrodes, many such electrodes (e.g., implanted pacing electrodes) are used to monitor biosignals as well as to deliver the required stimulation impulses. Even in the case of a purely stimulating electrode, a low interface impedance is required to minimize energy waste and to prolong the life of the power source. Various techniques are therefore used to effectively decrease the otherwise large interface impedances of the noble or inert materials used for their biocompatible properties. Electrode material and high electrical performance is generally less critical for external stimulation electrodes such as TENS and external defibrillation electrodes. Current density distribution is of major importance in these applications in order to avoid electrical hotspots and resultant burns to the skin. In some applications, such as TENS and external pacing, it is even sometimes advantageous to use a relatively resistive electrode material or gel, as this has been found to optimize current density distribution under the stimulation electrode.

As in the above applications, the avoidance of current density hotspots is one of several key factors in iontophoretic, transdermal delivery (6). An additional important constraint that is generally not relevant in other electrotherapies is the maintenance of the delicate electrochemical balance at the electrode/reservoir/skin interface. The electrode potential and impedance, as well as the composition of the drug reservoir, must generally remain within certain narrow ranges in order to avoid the deterioration of the electrode, the contamination of the drug reservoir, and the irritation of the patient's skin.

The electrical properties of the electrode contacts are, therefore, of great importance in most applications. Ideally, the contact with the patient should give rise to the following:

- Zero potential. Unfortunately, zero potential is not possible and a more realistic goal is to achieve a low, stable potential at each of the contacts.
- Zero Impedance. Unfortunately, zero impedance too is not possible and a more realistic goal is to achieve impedances at the two contacts that are low and as similar as possible.

The potentials and impedances of the electrode–electrolyte interface and the skin will therefore be studied in more depth in the following sections.

ELECTRICAL PROPERTIES OF ELECTRODE–SKIN INTERFACE

As briefly outlined above, the electrode–electrolyte interface and the skin under the electrode both give rise to potentials and impedances that can either distort any measured biosignal or give rise to problems during electrical stimulation.

The Electrode–Electrolyte Interface

The Electrode–Electrolyte Potential. When a metallic electrode comes in contact with an electrolyte (in body tissues or in an electrode gel), an ion–electron exchange occurs as a result of an electrochemical reaction. A tendency exists for metal atoms M to lose n electrons and pass into the electrolyte as metal ions, M^{+n}, causing the electrode to become negatively charged with respect to the electrolyte (Fig. 1). Reaction (**1**) is termed oxidation.

$$M \rightarrow M^{+n} + ne^- \qquad (\mathbf{1})$$

Similarly, under equilibrium conditions, some of the ions in solution M^{+n} take n electrons from the metal and deposit onto the electrode as metal atoms M. The electrode becomes positively charged with respect to the electrolyte. Reaction (**2**) is termed reduction.

$$M^{+n} + ne^- \rightarrow M \qquad (\mathbf{2})$$

The overall chemical reaction taking place at the interface is therefore

$$M \rightleftharpoons M^{n+} + ne^- \qquad (\mathbf{3})$$

Under equilibrium conditions, the rate at which metal atoms lose electrons and pass into solution is exactly balanced by the rate at which metal ions in solution deposit onto the electrode as metal atoms. The current flowing in one direction, i_0, is equal to and cancels out the current flowing in the opposite direction. The electrode is said to be behaving reversibly and the common value of currents, io, is termed the exchange current (density). Although the net current flowing through the electrode interface is zero, a potential difference is found to exist between the electrode and the electrolyte and depends on the position of the equilibrium between the two processes (**1**) and (**2**). Generally, the metal is negative relative to the electrolyte. The potential difference depends on the relative activities (or concentrations) of the ions present and on the electrode metal (7). This potential has been termed the equilibrium, reversible, or half-cell (i.e., one interface only) potential in the literature.

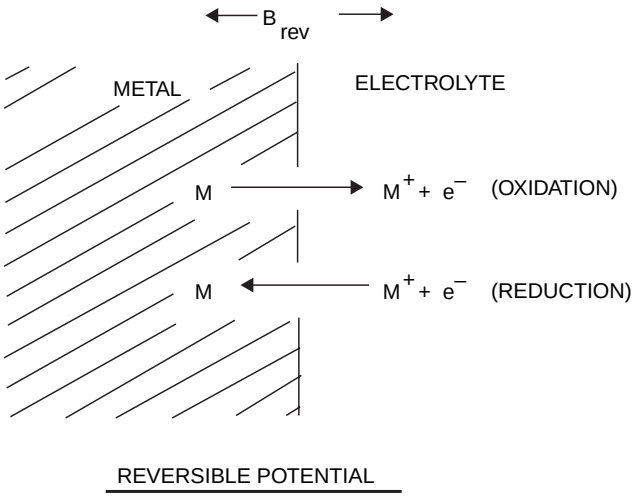


Figure 1. The electrode–electrolyte interface and reactions involved in generating its reversible or equilibrium potential.

When trying to measure the potential of a half-cell (i.e., one interface only), one is immediately faced with a problem, as one requires two electrodes to make a potential measurement, thus effectively creating an electrochemical cell with two electrode–electrolyte interfaces. One, therefore, measures not only the potential of the electrode–electrolyte interface under study, but also that of the second electrode used to complete the circuit. If one uses the same metal for the second electrode to that used in the first, the potentials will be identical (in theory at least) and will cancel each other out. The measured potential will be (theoretically) equal to zero. (In practice, however, slight differences in the composition of the metal used, the electrode surfaces, and in the gel will result in differences in the two half-cell potentials.) If, on the other hand, one uses a different metal for the second electrode, the measured potential of the cell will be due to the combination of the potentials of the two half-cells. It will be impossible to separate the potential of the half-cell under investigation.

In order to resolve this problem, early electrochemists decided to measure all electrode interface potentials with respect to a standardized electrode or reference electrode. The standard hydrogen electrode (SHE) was chosen to be the universal reference electrode and its half-cell potential was specified as zero. Other metal-to-ion interface potentials were then measured with reference to SHE and the entire measured offset voltage was attributed to electrode system being tested.

Hydrogen electrode consists of a platinized plate submerged to one-half its height HCl over which hydrogen gas at atm is bubbled. The half-cell potential of SHE depends on concentration of hydrogen ions in the solution, hence it is quite stable and reproducible. At the time that this decision was reached, the necessary glass blowing and silver soldering were common skills and the SHE was thus easy and inexpensive to make. Although, however, it is no longer convenient for modern routine measurements as a reference electrode (the flowing hydrogen gas is potentially explosive), electrode potentials are standardized with respect to the SHE (7).

The reversible, equilibrium, or half-cell potential of a given electrode–electrolyte interface depends on the activity (almost synonymous with concentration) of the ions taking part in the reactions (Table 1). This potential, E_{rev}, is given by the Nernst equation,

$$E_{rev} = E_0 + [RT/nF]\text{In}[\text{activity of oxidized form/activity of reduced form}] \qquad (1)$$

Table 1. Reversible Potentials for Common Electrode Materials at 25 °C[a]

Metal and Reaction	Potential E^V,V
$Al \rightarrow Al^{3+} + 3e^-$	−1.706
$Zn \rightarrow Zn^{2+} + 2e^-$	−0.763
$Cr \rightarrow Cr^{3+} + 3e^-$	−0.744
$Fe \rightarrow Fe^{2+} + 2e^-$	−0.409
$Cd \rightarrow Cd^{2+} + 2e^-$	−0.401
$Ni \rightarrow Ni^{2+} + 2e^-$	−0.230
$Pb \rightarrow Pb^{2+} + 2e^-$	−0.126
$H_2 \rightarrow 2H^+ + 2e^-$	0.000 by definition
$Ag + Cl^- \rightarrow AgCl + e^-$	+0.223
$2\,Hg + 2Cl^- \rightarrow Hg_2Cl_2 + 2e^-$	+0.268
$Cu \rightarrow Cu^{2+} + 2e^-$	+0.340
$Cu \rightarrow Cu^+ + e^-$	+0.522
$Ag \rightarrow Ag^+ + e^-$	+0.799
$Au \rightarrow Au^{2+} + 3e^-$	+1.420
$Au \rightarrow Au^+ + e^-$	+1.680

[a]The metal undergoing the reaction shown has the magnitude and polarity of standard half-cell potential, E_0. Listed when the metal is referenced to the standard hydrogen electrode (3).

E_{rev} is the reversible, equilibrium, or half-cell potential
E_0 is the standard half-cell potential (measured relative to the standard hydrogen electrode)
R the universal gas constant,
n the number of electrons involved in reaction,
T the absolute temperature (K).

Activity, $a = \gamma C$, where C is concentration and γ, the activity coefficient, is a measure of the interaction between ions. When solution is infinitely dilute, $\gamma = 1$ and activity is equal to concentration.

Note the two components of E_{rev}. One is constant, E_0, wheras the other will vary due to slight variations in concentration, from one electrode to another. If two chemically identical electrodes make contact with the same electrolyte/body, the two interfaces should, in theory, develop identical half-cell potentials. When connected to a differential amplifier, the half-cell potentials of such electrodes would cancel each other out and the offset voltage would be zero. The electrode potentials would, therefore, make zero contribution to a biosignal they were being used to detect. Unfortunately, slight differences in electrode metal or gel result in the creation of offset voltages, which can greatly exceed the physiological variable to be measured. Generally, a more significant problem is that the electrode offset voltage can fluctuate with time, thus distorting the monitored biosignal (8).

The Electrode–Electrolyte Impedance. It has already been stated that in the electrode and the connecting lead, electrical charge is carried by electrons, whereas in the gel and in the human body, charge is carried by ions. A transition exists at the interface between the electrode and the electrolyte where charge is transferred from one kind of carrier to the other. In order for some of the ions in the electrode gel or in the body fluids to transfer their charge across the interface, many must first diffuse to

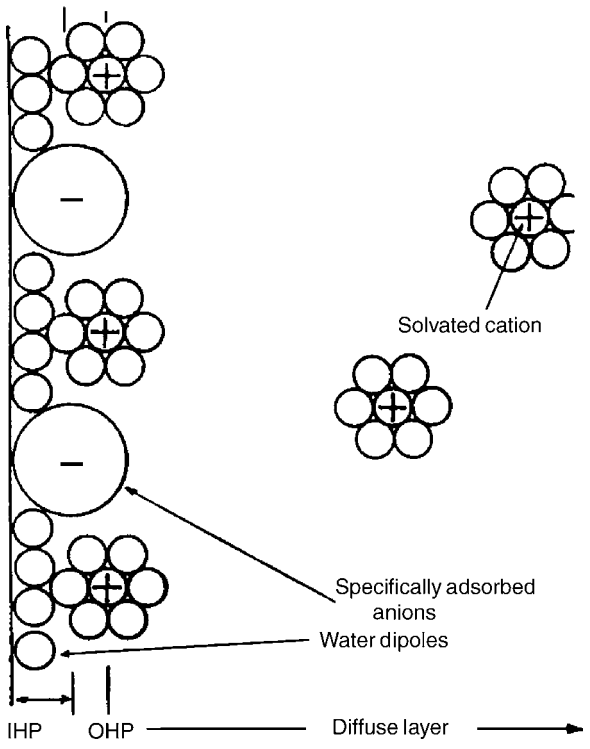


Figure 2. The electrode–electrolyte interface. A metal in an electrolyte forms a double layer of charge. (Redrawn from Ref. 7).

the electrode–electrolyte interface under the influence of electrostatic attraction. Here they stick (or adsorb, as it is termed in electrochemistry) to the electrode surface and form the outer Helmholtz plane (OHP) (7). If the electrode has a negative charge relative to the electrolyte, positive ions will be attracted to the interface region and adsorb onto the electrode surface. As a consequence, there is a layer of negative charge on the metal surface and a layer of equal but opposite charge on the electrolyte side of the interface, both separated by a small distance across the OHP (see Fig. 2). A double layer of charge therefore exists at the interface, and such a system behaves like a parallel-plate capacitor. Not altogether surprising, the interface's capacitance is often termed the double-layer capacitance, C_{dl}, and is connected in parallel to the charge-transfer resistance in our simple equivalent circuit model.

Just in case one believed that the electrode interface was that simple, one must point out, for example, that as well as the cations electrostatically attracted to the negatively charged electrode surface (coulombic adsorption) anions may exits that are adsorbed on the electrode surface and form the inner Helmholtz layer or plane (IHP). These anions have tended to lose their hydration sphere and, consequently, are in close contact with the electrode. As they are negative ions adsorbed onto a negative electrode surface, electrostatic forces cannot be responsible. Some force *specific* to the ion (rather than its electric charge) must be responsible, hence the use of the term specific

adsorption to describe this phenomenon. The van der Waals or chemical forces is thought to be responsible (7).

In order to understand some aspects of the double-layer capacitance, it is good to consider the basic equation for a parallel-plate capacitor. If two identical conductive plates, each of area A cm^2, are separated by a distance d cm, which is filled with a material of dielectric constant ε_0, then the capacitance of this parallel-plate capacitor, C_{pp}, is given by:

$$C_{pp} = \varepsilon_0 A/d \tag{2}$$

and the magnitude of the capacitive impedance, Z_{pp}, is given by

$$Z_{pp} = 1/2\pi f C_{pp} \tag{3}$$

where f is the frequency of the applied ac signal and π is a constant.

Some dc (or faradaic) current does, however, manage to leak across the double layer due to electrochemical reactions (**1**) and (**2**) taking place at the interface. These reactions experience a charge transfer resistance, R_{CT}, which can be thought of as shunting the nonfaradaic, double-layer capacitance and whose expression can be derived from the Butler–Volmer equation.

For small applied signal amplitudes (7),

$$R_{CT} = \frac{RT}{nF}\frac{1}{i_0} \tag{4}$$

A good electrode, from an electrical point of view, will have a very low value of R_{CT}. Charge will be transferred across the interface almost unimpeded and little voltage will be dropped across the interface. One should note that R_{CT} is inversely proportional to i_0. i_0 is the exchange current [i.e., the current flowing across the interface (in both directions) under equilibrium conditions (no net current flow)]. Simplistically, if an interface can cope with large currents under equilibrium conditions, it will be able to cope well with currents under nonequilibrium conditions. A good electrode system will therefore be characterized by a large value of exchange current or a low value of R_{CT}.

The interface impedance should theoretically be well represented by an equivalent circuit model comprising the double-layer capacitance in parallel with the charge transfer resistance, R_{CT}. Both are in series with R_{TOTAL}, the relatively small resistance due to the sum of the lead and electrolyte resistances.

Complex Impedance Plot. If, for each frequency of ac signal used to measure the impedance, the real part of the measured impedance (Z' or R_S) is plotted on the x axis and the imaginary part (Z'' or X_S) on the y axis of a graph, one obtains a Nyquist or complex impedance plot. The impedance locus for the above simple equivalent circuit model (Fig. 3.) of the interface impedance is plotted on a complex impedance plot in Fig. 4. *Note*: Electrochemists plot–X_S versus R_S and not X_S versus R_S as electrode (and tissue) impedances tend to be capacitive and thus negative. It is generally found easier to look at the plots with the Z'' axis inverted. Low frequency data are on the right side of the plot and higher frequencies are on the left, which is generally the case for electrode interface data.

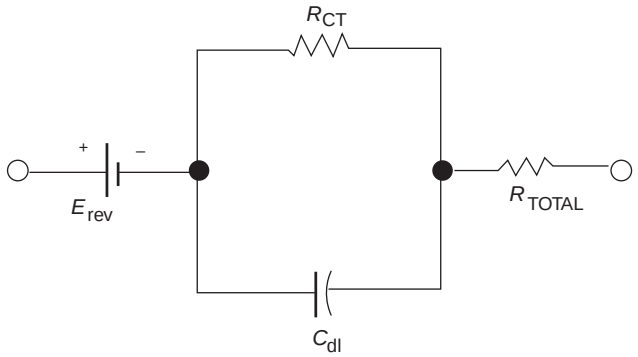


Figure 3. Simple equivalent circuit model of the electrode-electrolyte interface. C_{dl} represents the double-layer capacitance, R_{CT} the charge transfer resistance, R_{TOTAL} the sum of the lead and electrolyte resistances, and E_{rev} represents the reversible or equilibrium potential.

The impedance locus has the form of a semi-circle with high and low frequency intercepts with the real axis at 90° (due to the presence of C_{dl} in parallel with R_{CT}). At very low frequencies, the impedance is equal to $R_{TOTAL} + R_{CT}$, the diameter of the semicircle being equal to R_{CT}. At higher frequencies, the impedance is influenced by the value of the parallel capacitance C_{dl}. As the capacitive impedance decreases with increasing frequency, current therefore flows through it and the total impedance of the parallel combination decreases. The reactive component and the phase angle increases from zero, reaches a maximum value (which depends on the relative sizes of R_{TOTAL} and R_{CT}), and then decreases again toward zero (see Figs. 4 and 5). The frequency at which the reactive component reaches its maximum value (ω_0) is given by $\omega_0 = 1/R_{CT}C_{dl}$ (Fig. 4).

At high frequencies, the impedance is determined by the series resistance R_{TOTAL}.

Bode Plot. Another popular method of presenting impedance data is the Bode plot. The impedance is plotted

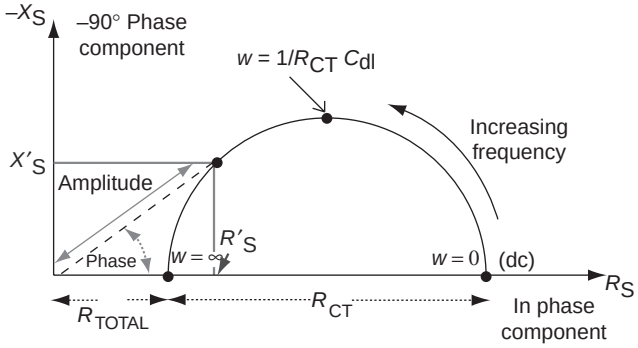


Figure 4. Impedance plot for simple equivalent circuit model of the electrode–electrolyte interface. The impedance locus is semicircular as a result of the parallel combination of C_{dl} (the double layer capacitance) and R_{CT} (the charge transfer resistance), both of which are in series with R_{TOTAL}, the sum of the lead and electrolyte resistances.

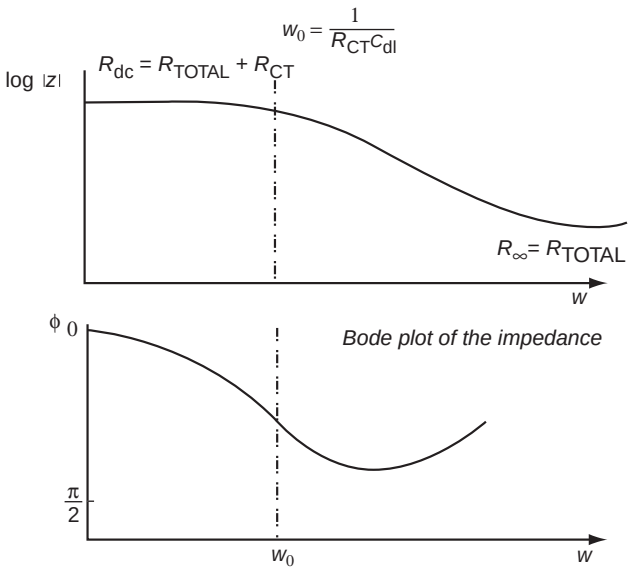


Figure 5. Bode plot for the simple '3-component' equivalent circuit model. The magnitude of the impedance decreases from its low frequency value ($R_{TOTAL} + R_{CT}$) to R_{TOTAL} at very high frequencies. The phase angle of the overall interface impedance increases from 0° at low frequencies, reaches a maximum, which depends on the relative sizes of R_{TOTAL} and R_{CT} and then decreases again.

with log frequency on the x axis and both the absolute value of the impedance ($|Z| = [Z'^2 + Z''^2]^{1/2}$) and the phase-shift ($\phi = \tan^{-1}[Z''/Z']$) on the y axis. Unlike the complex impedance plot, the Bode plot explicitly shows frequency information.

The Bode plot for the electric circuit of Fig. 3 is shown in Fig. 5.

As in the complex impedance plot, the magnitude of the impedance is equal to $R_{TOTAL} + R_{CT}$ at very low frequencies (R_{dc}). The phase angle is zero at this point as the impedance is purely resistive. At very high frequencies, the magnitude of the impedance (R_∞) is equal to that of the series resistance R_{TOTAL}. At frequencies in between these two limits, the interface impedance is influenced by the value of the parallel capacitance C_{DL}. As the capacitive impedance decreases with increasing frequency, current therefore flows through it and the total impedance of the parallel combination decreases. The phase angle increases from zero, reaches a maximum value (generally less than 90° or π/2 rad), and then decreases again toward zero (see Fig. 5).

The above model can be used to explain most key aspects of the electrode–electrolyte interface. It must be pointed out, however, that the equivalent circuit is a gross approximation.

For example, diffusion of ions to the interface from the bulk of the electrolyte (gel or patient) takes place at a finite rate and thus gives rise to impedance to current flow, especially at low frequencies. The diffusion (often termed Warburg) impedance is generally located in series with the charge transfer resistance, both of these being in parallel with the double-layer capacitance. The diffusion impedance has been ignored in the above model as it tends not to be observed for many biomedical electrode systems over the range of frequencies typically used.

A further simplification is the use of a simple capacitance in the above model. Such ideal capacitive behavior is rarely observed with solid metal electrodes. Instead, an empirical pseudo capacitance or constant phase angle impedance, Z_{CPA}, is often used that has a constant phase angle, much like a capacitor.

$$Z_{CPA} = K(i\omega)^{-\beta} \tag{5}$$

where K is a measure of the magnitude of Z_{CPA} and has units of $\Omega s^{-\beta}$, and β is constant such that $0 < \beta < 1$. The phase angle of this empirical circuit element ($\phi = \beta\pi/2$ radians or $90\beta°$) generally lies between 45° and 90° (9). Typically, β has a value of 0.8 for many biomedical electrode systems.

Fricke (10) used the term polarization to describe the constant phase angle impedance and postulated that it was due to spontaneous depolarization of the electrode. Although he did not enlarge on the hypothesis, many authors have used Fricke's terminology over the intervening years. The present author must concur with Cole and Curtis' observation that "the use of the term polarization for describing the unexplained effects occurring at the metal–electrolyte interface is only an admission of our ignorance" (11).

The two most likely causes of the observed constant phase angle impedance are specific adsorption and surface roughness effects (12). With solid biomedical electrodes, the nonideal behavior is probably due to the surface roughness of the electrodes (13), which is supported by reports that roughing an electrode surface decreases the measured value of phase angle.

It is also naive to think that surface effects will only distort the nonfaradaic impedance and will have no effect on R_{CT} as assumed in the above model. It is more realistic that surface effects will affect the parallel combination of C_{dl} and R_{CT} giving rise to skewed (14,15) or distorted (16) arcs. The simple equivalent circuit used in this presentation is, however, a useful approximation that enables qualitative interpretation of much of the published data.

Polarization. Since the work of Fricke (10), the term polarization has been used to describe just about anything associated with the electrode–electolyte interface—frequency-dependence, nonlinearity, noise, and so on. Polarization has been defined as "the departure of the electrode potential from the reversible value upon the passage of faradaic current" (7).

Under equilibrium conditions, the electrode potential E is equal to its reversible potential E_{rev}. When a dc or faradaic current, i_{dc}, is applied to the electrode interface, it must flow through the resistance R_{CT}, which is in parallel with C_{dl}. From Ohm's law, the voltage dropped across this charge transfer resistance will be equal to i_{dc} multiplied by R_{CT} (Fig. 6). The electrode potential E is now given by:

$$E = E_{rev} + i_{dc}R_{CT} \tag{6}$$

The electrode, therefore, is no longer operating at its equilibrium or reversible value E_{rev}. This change in the

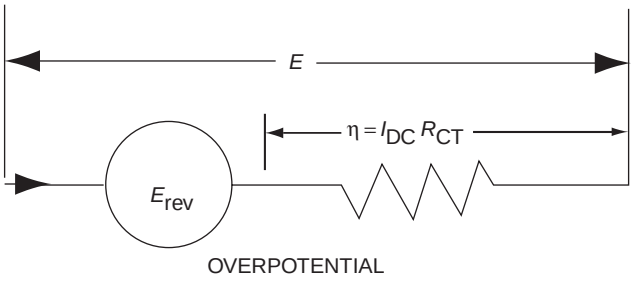


Figure 6. Polarization, the departure of the electrode potential from its reversible value upon the passage of faradaic current.

electrode interface's potential from its equilibrium value is termed polarization. The degree of polarization is measured by the additional voltage dropped across R_{CT}, (or overpotential, η, as it is termed in electrochemistry) where:

$$\eta = /E - E_{rev}/ \tag{7}$$

An ideal nonpolarizable electrode would have a value of R_{CT} equal to zero and, hence, would exist no resistance to faradaic current. The nonfaradaic impedance would effectively be shorted out and the total interface impedance would be zero. In this case, current would pass freely across the interface unimpeded. Measured biosignals, for example, would be unattenuated and undistorted. A perfect electrode system! The electrode potential would always remain constant at its reversible value.

A perfectly polarizable electrode would not permit the flow of any dc or faradaic current as the charge-transfer resistance in this case is infinite. Such an electrode is sometimes termed a blocking electrode. No faradaic charge would cross the interface, even for large overpotentials, and the electrode couples capacitvely with the tissues/electrolyte in this extreme case (Fig. 7).

Real electrodes are, however, neither perfectly polarizable nor perfectly nonpolarizable. Any net current flow across an electrode–electrolyte interface will experience a finite faradaic impedance across which an overpotential will develop.

An electrode system that has a very low value of R_{CT} lets current traverse the interface almost unimpeded, wastes little energy at the interface, has a relatively small overpotential, and has a relatively nonpolarizable electrode system. Such electrode systems are highly sought after, especially when recording small biosignals from the body surface.

Electrodes made of noble metal come closest to behaving as perfectly polarizable electrodes. As these metals are inert, they tend not to react chemically with the surrounding electrolyte or tissue. Noble metals are, therefore, generally used in the construction of implant electrodes where chemical reaction with surrounding tissues must be avoided in order to minimize tissue toxicity problems. Little steady current can pass in such cases as the charge transfer resistance for these electrodes is therefore very large (Fig. 7). The small current that does pass represents the charging and discharging of the double-layer capacitance. A problem, therefore, exists when designing implant electrodes. For a biocompatibility point of view, one requires a noble, hence polarizable, electrode system, whereas from an electrical performance point of view, one requires a nonpolarizable system. A compromise is achieved by using a polarizable electrode and roughening the surface of the electrode, thus decreasing the large interface impedance.

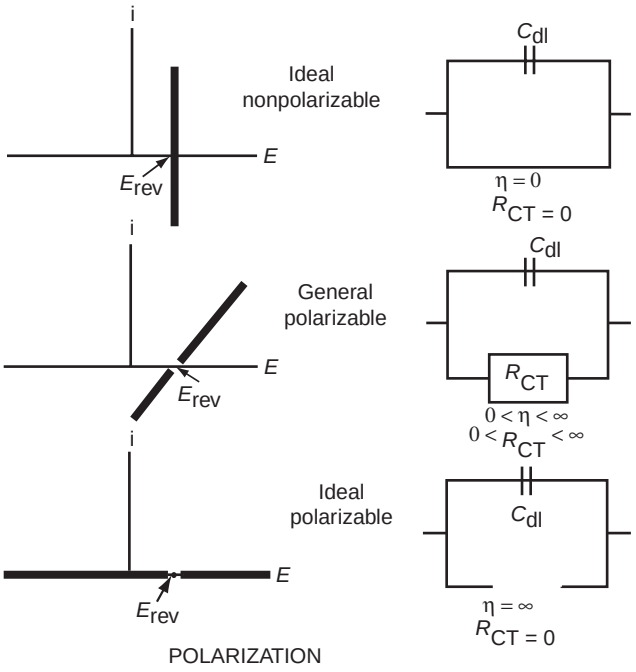


Figure 7. The dc current–voltage plots for Ideal Nonpolarizable, General Polarizable, and Ideal Polarizable electrode interfaces.

Transient Response and Tissue Damage. The response of the electrode system to sine waves of varying frequencies has been considered above (Complex Impedance and Bode plots) as this is a very useful tool in analyzing circuits or, in this case, electrode systems. Equally relevant is the response of an electrode system to voltage and current steps or pulses, as these will approximate therapeutic stimulation applications.

It must be borne in mind that the conversion from electrical to ionic current takes place at the electrode-tissue interface. Based on the simple equivalent circuit model, current can flow either through the parallel resistance or through the double-layer capacitance.

Current flowing through the parallel resistance involves faradaic charge transfer reactions. At the anode, the electrolysis of water and the oxidation of organic compounds can occur. The oxidation of the electrode itself can also occur, which results in the dissolution of metal. At the cathode, hydrogen ions are reduced to form hydrogen gas, which results in a change in pH near the electrode. The new chemical by products in all of these reactions may lead to tissue damage and, hence, faradaic charge transfer reactions must be avoided (17,18). Current must not, therefore, be allowed to flow through R_{CT}.

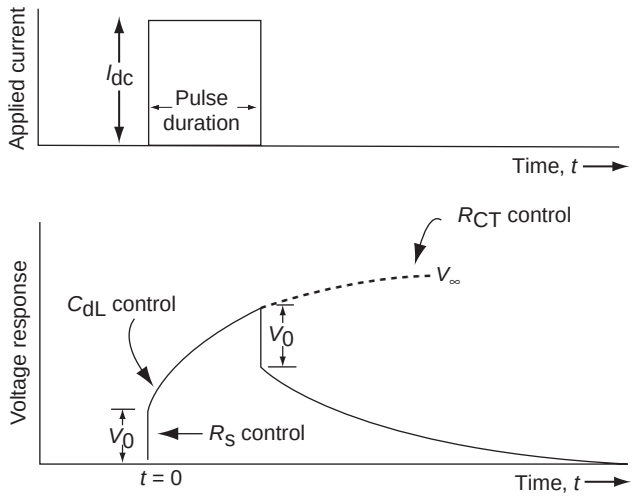


Figure 8. Voltage response to a pulse or step in current.

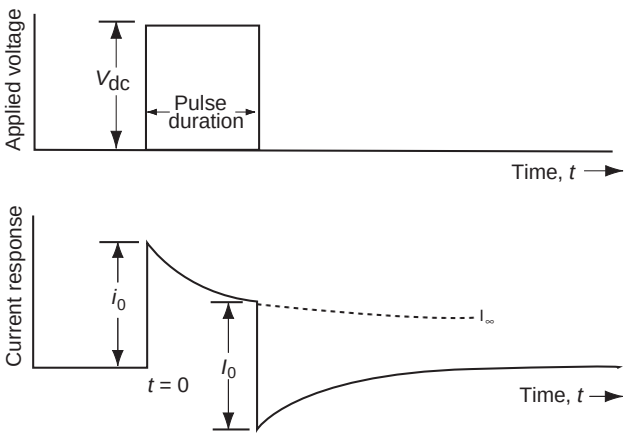


Figure 9. Current response to a pulse or step in voltage.

For current flowing through the double-layer capacitance, no charge actually crosses the electrode-tissue interface. Instead, ions in the tissue are attracted or repelled by charges on the electrode, resulting in transient pulses of ionic current. As no net current flows through the interface and electrochemical reactions are not involved, capacitive current is relatively safe. One must therefore seek, as far as possible, to couple capacitively with tissue when seeking to stimulate tissue without causing trauma.

If one applies a pulse of current of amplitude I_{dc} at time $t = 0$, the voltage response of the electrode-interface equivalent circuit model and, it is believed, the electrode-patient system is as shown in Fig. 8.

$$V(t) = I_{dc}R_{TOTAL} + I_{dc}R_{CT}(1 - \exp[-t/R_{CT}C_{dl}]) \quad (8)$$

At $t = 0$, the applied current flows unopposed through the capacitor and, hence, only sees the series resistance R_{TOTAL}. The initial voltage response is, therefore,

$$V_0 = I_{dc}R_{TOTAL} \quad (9)$$

The voltage response is then observed to gradually increase from V_0. The initial increase in voltage with time is inversely proportional to the magnitude of the capacitance.

For long pulse durations, all of the current will flow through the resistances R_{CT} and R_{TOTAL}. The total resistance seen by the current is, therefore,

$$Z_{(t=\infty)} = R_{TOTAL+}R_{CT} \quad (10)$$

and the limit voltage V_∞ is given by

$$V_\infty = I_{dc}(R_{TOTAL} + R_{CT}) \quad (11)$$

The voltage response will reach this limit value V_∞ in a time period of approximately five time constants, T, where $T = C_{dl}R_{CT}$.

If a perfect step in voltage, V_{dc}, is applied to the electrode system or the three-component model, the current response is as shown in Fig. 9 and given by the equation

$$I(t) = V_{dc}\left\{\left(\frac{1}{R_{TOTAL} + R_{CT}}\right)\right\} + \left(\frac{R_{CL}/R_{TOTAL}}{R_{CT} + R_{TOTAL}}\right)\exp\left[-\frac{R_{CT} + R_{TOTAL}}{R_{CT}R_{TOTAL}C_{dl}}t\right] \quad (12)$$

At the beginning of the voltage step, the resultant current jumps to a relatively large value I_0, where

$$I_0 = V_{dc}/R_{TOTAL} \quad (13)$$

As time passes, the resultant current decreases exponentially with an initial slope inversely proportional to the capacitance C_{dl}. Eventually, the current will reach a limiting value of I_∞, where

$$I_\infty = \frac{V_{dc}}{R_{TOTAL} + R_{CT}} \quad (14)$$

Tissue Damage. The rise in voltage response or the decrease in current response is often attributed in the literature to a mysterious phenomena called polarization. It is, in fact, nothing more than the transient response of a simple three-component circuit model.

When one applies pulses to an electrode-tissue interface (either for implanted or surface stimulation), the applied or resultant current initially flows into the patient via the double-layer capacitance. No reactions are associated with the capacitive flow of current and, hence, few undesirable effects exist when using short duration pulses (i.e. with a duration less than one time constant). As the pulse duration is increased, however, progressively more current flows through the parallel charge transfer resistance and the patient is more at risk due to the reaction byproducts, which is especially true if the pulse is applied for a length of time longer than five time constants giving the voltage or current response time to level off and reach its steady-state value of V_∞ (or I_∞). At this point, most of the current is flowing through the charge-transfer resistance and charge is therefore injected into the tissues via a faradaic process. The byproducts of the electrochemical reactions involved in

the charge-transfer process will diffuse into the skin or tissue, causing chemical injury (19). As the leveling off of the voltage response is indicative of pure charge transfer control, this feature must be avoided at all costs by either avoiding long duration pulses or by using electrode systems with very long time constants, $T = C_{dl}R_{CT}$. In both cases, charge will be largely applied to the body via relatively safe capacitive processes.

It may not always be possible to use sufficiently short pulse durations to avoid tissue damage and still achieve therapeutic effect. It must also be noted that even when using short pulses where the current or voltage response has not leveled off, some current still flows through R_{CT} and a faradaic charge-transfer reaction takes place, with the associated, albeit reduced, problems (18). Over extended periods of stimulation, the byproducts will accumulate in the tissues. In the past, however, the applied waveforms found experimentally to minimize electrode and tissue damage are consistent with the basic goal of minimizing the flow of faradaic current across the electrode interface (20) and thus ensuring little, if any, net transfer of charge. Decreasing the duration, amplitude, and rate of current pulses have all been suggested as each ensures C_{dl} control of the transients and thus avoids, to some extent at least, the undesirable faradaic processes.

An alternative to using short pulse durations is to use electrode systems with long time constants [i.e., with a large value of R_{CT} (faradaic charge-transfer resistance) or C_{dl} (double-layer capacitance)]. As we have seen, noble metals react with difficulty with surrounding tissues and thus have large values of R_{CT} (19). For a given pulse duration, they tend to couple capacitively with tissue. The reactions on both anode and cathode are reversible, involving surface oxygen, and this reversibility explains the observed identical nature of anodic and cathodic waveforms (19). As a result, noble electrodes are widely used for physiological stimulation. Unfortunately, noble metals give rise to large interface impedances. From a biocompatibility point of view, one requires a noble (hence, large impedance) electrode system, whereas from an electrical performance point of view, one requires a low impedance system. A compromise is achieved by using such polarizable electrode materials and roughening the surface of the electrode, thus decreasing the large interface impedance. Roughening the electrode surface gives rise to a significant increase in the interface capacitance, thus increasing further the time constant ($T\uparrow = C_{dl}\uparrow R_{CT}$) and ensuring that the voltage or current response is even more dominated by the highly desirable capacitive processes while decreasing the interface impedance.

It would be a gross oversimplification to attempt to demystify biomedical electrode design by stating that all high performance biomedical electrodes simply have rough surfaces. However, a significant element of truth exists in the statement. For example, terms like activated, sintered, and porous have been used to describe implant electrodes for cardiac pacing and indicate that the electrode fabrication process results, deliberately or otherwise, in a rough-surfaced electrode. It could be argued that it is often the surface finish rather than the electrode metal that gives rise to the favorable electrical properties reported, especially the low interface impedances.

Most electrical stimulation electrodes rely, to some extent, on faradaic mechanisms at the interface between the metal and the tissue. Even in the case of noble metal electrodes with short pulse durations, byproducts of the electrochemical reactions involved will accumulate over time in the tissues when a signal is applied in one direction (monophasic) and will eventually give rise to irritation. With surface stimulation, for example, early designs of electrodes incorporated thick pads of electrolyte-impregnated lint in order to distance the patient's skin from the electrode–electrolyte interface and the undesirable byproducts. Obviously, with implanted electrodes, that short term solution is not possible.

In particular, monophasic anodic pulses must be avoided as they will cause corrosion problems. Additionally, for most applications, cathodic stimulation has a lower threshold than anodic stimulation. Even in the case of monophasic cathodic pulses, however, current flows in only one direction and the chemical reactions at the interface are not reversed.

Biphasic waveforms are preferred in most electrotherapies as the byproducts of the forward reaction are thought to be recaptured by the reverse reaction (21). In using charge balanced waveforms, it is often believed that because no net charge transfer exists across the electrode–skin interface, no net flow of potentially harmful byproducts into the skin. Unfortunately, the electrochemical reaction that occurs to enable the flow of current during the first phase is not necessarily that involved in the second. Byproducts of the first reaction are therefore not always recaptured and may escape from the interface into the patient (22). However, it must be pointed out that the use of charge balanced biphasic waveforms does indeed greatly minimize the problem and, hence, its widespread use in a range of surface and implant applications. Additionally, surface biphasic stimulation is found to be more comfortable than monophasic.

Limit Voltages and Currents of Linearity. Although the non-linearity of the skin's electrical properties has been investigated under a range of conditions, the phenomenon is still far from well understood. "There appears to be, so far, no model available which accounts for both the linear and nonlinear behavior of the electrodes in the frequency and time domains" (23). It is an important feature of an electrode as 'it appears . . . that electrodes often introduce nonlinear characteristics that are erroneously ascribed to the biological system under study (24).

Schwan proposed empirical relationships for the limit current of linearity and the limit voltage of linearity.

Limit Current of Linearity. If has been observed that electrode–tissue interface impedance nonlinear behavior is first evidenced at low frequencies. As the applied current amplitude is increased, progressively higher frequency points are affected. Schwan proposed a limit current of

linearity i_L. He observed that the relationship between the angular velocity of a given impedance point and the current amplitude required to drive it into nonlinearity (deviate by more than 10% from its linear, small-signal value) was well expressed by the empirical relationship

$$i_L = B\,\omega^{\beta} \tag{15}$$

where B is a constant particular to the electrode system and β is the fractional power that appears in equation 5.

Schwan and others (25–28) have observed that this empirical relationship is valid for many electrode systems over wide frequency ranges.

The presence of β (a parameter describing the frequency dependence of the linear interface impedance) in a relationship describing the nonlinearity of the system was found most intriguing.

The solution to this mystery is quite simple when it is approached from the right direction. Generally, researchers have assumed that the observed nonlinear behavior is attributable to the high frequency Z_{CPA} impedance (Eq. 5), which they observe under linear, small-signal conditions and over the limited frequency ranges they use. However, in parallel with Z_{CPA} is the charge transfer resistance R_{CT}, which, in the linear range, has a very large value R_{CT}^{0}, where

$$R^{0}_{CT(0)} = \frac{RT}{nF}\,\frac{1}{i_0} \tag{16}$$

As a result, its contribution is either not observed or ignored.

The value of the charge-transfer resistance can be derived from the Butler–Volmer equation and is very nonlinear, decreasing rapidly with applied signal (ac or dc) amplitude. Compared with R_{CT}, Z_{CPA} is relatively linear. R_{CT} is therefore the source of the observed nonlinear behavior.

As the applied current amplitude is increased, the charge transfer resistance decreases rapidly, causing the diameter of the impedance locus to decrease. As the low frequency end of the arc is dominated by the charge transfer resistance, the effects of such nonlinearity will be first evidenced at these frequencies. Low frequency points are therefore the first to deviate significantly (by more than 10%) from their small-signal, linear values, as observed by Schwan and others. As the applied signal amplitude is further increased, the diameter of the impedance locus decreases further, and progressively higher frequencies are affected (29,30).

Simplistically, it can be shown that the following approximations can be made over limited ranges of frequency or applied signal amplitude:

- Approximate relationship between applied current and R_{CT}

$$i \propto 1/R_{CT} \tag{17}$$

- Approximate relationship between R_{CT} and the frequency at which nonlinearity occurs

$$R_{CT} \propto \omega^{-\beta} \tag{18}$$

Then, by cancelling R_{CT}, in the above two equations,

- Approximate relationship between applied current and the frequency at which nonlinearity occurs

$$i_L \propto \omega^{\beta} \tag{19}$$

as found by Schwan. The presence of β in the expression of an electrode system's nonlinear behavior is therefore simply due to the presence of a very nonlinear resistance in parallel with a relatively linear, frequency-dependent Z_{CPA}. A more accurate calculation based on the equivalent circuit model outlined above and the Butler–Volmer equation was published (29).

Limit Voltage of Linearity. Schwan and others (25,28) also postulated that the electrode–electrolyte interface impedance becomes nonlinear at a certain limit voltage, V_L, which they found to be independent of the frequency of the applied signal.

Using the Butler–Volmer equation and the equation for the impedance of the equivalent circuit model, the voltage limit of linearity can be calculated for a range of frequencies (31). It can be shown that the charge-transfer resistance decreases pseudo exponentially with applied voltage amplitude, initially causing low frequency impedance points on the locus to deviate from their small-signal values (32,33).

At very low frequencies, such that $\omega \rightarrow 0$, the voltage limit of linearity, V_L, approximates to the voltage at which the charge-transfer resistance decreases by 10% from its small-signal value, which occurs at $V_L = 40/n$ mV, where n is the number of electrons per molecule oxidized or reduced (31).

As the applied voltage is increased above this low frequency limiting value, the charge transfer resistance further decreases and affects progressively higher frequency points (i.e., become nonlinear). The derived log (f) versus V_L plot is found to be a straight line over a wide range of frequencies. V_L is observed to increase only very slightly with frequency, which would agree qualitatively with Onaral and Schwan's results (28), where V_L increased from 106 to only 129 mV over the frequency range of 10 mHz–100 Hz for platinum electrodes in saline, which would also explain why, in the past, V_L has been assumed constant and independent of the applied frequency.

Electrode Metals. As biocompatibility is of great importance in implants, implant electrode materials are generally confined to those that are essentially inert and do not react with the surrounding tissues. As cardiac pacing electrodes were among the first implanted and have had a long, generally successfully and well-researched history, most conclusions drawn on the suitability of materials for implant electrodes are based on pacing electrodes.

Implant electrodes are and have been generally made from noble metals such as gold, platinum, iridium, rhodium, and palladium. Platinum has been the most widely used as it has excellent corrosion resistance and produces relatively low polarization (34). Platinum, however, is mechanically relatively soft and for many applications is

alloyed with much harder iridium, producing platinum-iridium. Other noble metal alloys that have been used include gold–platinum–rhodium, platinum–rhodium, and gold–palladium–rhodium.

Passive metals, such as titanium, tantalum, zirconium, tungsten, and chromium, have been successfully implanted. Titanium has been widely used because it forms a nonconducting oxide layer at the surface. This coating prevents charge transfer at the electrode interface. Titanium, therefore, exhibits a high resistance to corrosion. Stainless steel is similar in that it acquires a protective oxide layer that renders it inert. Although stainless steels were used in early pacing electrodes, they do not appear to have the required corrosion resistance for long-term use. Stainless-steel pacing electrodes were discontinued after the 1960s because of unreliable corrosion resistance (34,35).

Some early pacing electrodes were made of Elgiloy (an alloy of Fe, Ni, CO, Cr, and MO from Elgin Watch Co.) However, Elgiloy has marginal corrosion resistance and produces a relatively high polarization overvoltage. It was discontinued in the 1980s. Carbon is an inert, nonmetallic element that has similar electrochemical characteristics to noble metals and continues to be used successfully as an implant electrode. Materials such as zinc, copper, mercury, nickel, lead, copper, silver, silver chloride, iron, and mild steel have been found toxic to body tissues and are normally not used.

Biocompatibility has been defined as the ability of a material to perform with an appropriate host response in a specific application (36). Strictly speaking, no such thing as a biocompatible material exists as an implant's biocompatibility will also depend on a range of variables including its shape and surface finish.

Stimulation threshold is a key parameter in implant stimulation electrode design. When activitated vitreous carbon electrodes were first introduced in pacing electrodes, they were found to have relatively low chronic thresholds. These thresholds were thought to be the result of the superior biocompatibility of the carbon electrode. Other researchers similarly interpreted the low thresholds observed for their new exotic materials such as indium oxide, titanium nitride, and semimetal ceramics. Stokes (34), however, concluded that "material selection appears to have little or nothing to do with threshold evolution—as long as the material is biocompatible and reasonably corrosion resistant. Thus our experiments with biocompatible materials such as carbon, titanium, platinum, iridium oxide, and many more have all produced about the same results when tested as polished electrodes, all other factors held equal". Stokes went on to point out "while the bulk properties of an electrode material are important, it is the electrode-tissue interface that determines the electrode's performance. In fact, the surface microstructure of the electrode is critical" (34). It would appear that the microstructure of an electrode surface may affect cellular adhesion and activation, thus reducing the foreign body response. It is, therefore, the surface structure of many of the new materials (resulting from their fabrication process) that gives rise to the observed positive effect on threshold evolution over time, rather than the biocompatibility of the bulk material.

Another advantage of porous and microporous implant surfaces is their reduced interface impedance. Although interface impedance is generally less critical for implanted stimulation electrodes, many such electrodes (e.g., implanted pacing electrodes) are used to monitor biosignals as well as to deliver the required stimulation impulses. Decreased interface impedance helps in this regard.

Implanted biosignal monitoring electrodes require stable potentials as well as low interface impedances to minimize biosignal recording problems. These metals have high positive standard electrode potentials (E^0 in Eq. 1) and are the lowest ones on the electromotive series. As noble metal electrodes do not tend to react chemically with the electrolyte, the Nernst equation is not defined and the measured potential is often influenced more by any traces of impurities on the surface than by the intrinsic properties of the metal itself. The electrode potential can drift randomly, especially immediately following implantation. It may fluctuate widely under apparently identical circumstances, which is an inherent disadvantage of noble materials.

External biosignal monitoring electrodes can generally use high electrical performance nonnoble materials such as silver–silver chloride without fear of biocompatibility problems (2). Silver–silver chloride has been found to be an excellent electrode sensor material as, when it is in contact with a chloride gel, it has the following characteristics:

1. A low, stable electrode potential.
2. A low level of intrinsic noise.
3. A small value of charge transfer resistance (i.e., it is relatively nonpolarizable).
4. A small interface impedance.

A silver–silver chloride electrode is generally made by the deposition of a layer of silver chloride onto a silver electrode. Silver chloride is a sparingly soluble salt and, thus, effectively provides the silver electrode with a saturated silver–chloride buffer, which facilitates exchanges of charge between the silver electrode and the sodium chloride environment of the gel and human body. The system behaves as a reversible chloride ion electrode, and the Nernst potential, in this case, depends on the activity (which is closely related to concentration) of the environment chloride ions and not on that of the silver ions. The potential of this electrode is, therefore, quite stable (as well as small) when the electrode is placed in an electrolyte containing Cl as the principal anion—as is the case in the human body and electrode gels (2).

Electrical noise (potential fluctuations) can occur spontaneously at the electrode interface without any physiological input. Ag/AgCl electrodes have been shown to be particularly stable and resistant to noise (37).

A silver–silver chloride electrode has a relatively large value of exchange current density (2) (Eq. 4) and, hence, a very low value of charge transfer resistance, R_{CT}. Charge is transferred across the interface with relative ease and little voltage is dropped across the interface. The electrode therefore operates close to its equilibrium or reversible

potential. Ag-AgCl electrodes are, therefore, relatively nonpolarizable.

When a smooth-surfaced electrode is chlorided, the AgCl deposit can give rise to a very rough surface and thus to relatively very low interface impedances (37,38). K, the magnitude of the interface pseudocapacitance (Eq. 5), is observed to decrease following the deposition of an AgCl layer (39). However, although AgCl facilitates the interfacial electrochemistry, it is very resistive having a resistivity of around 10^5–$10^6\,\Omega\cdot$cm (2). As the layer thickness increases, the series resistance, R_{TOTAL} will therefore increase. This series resistance dominates the very high frequency interface impedance, and the latter will also increase with chloride deposit. Therefore, an optimal layer thickness exists, for a given frequency, that decreases the interface impedance and yet does not significantly increase the series resistance, R_{TOTAL} (29). The optimal silver chloride layer thickness consequently depends on the frequency range of interest (40).

Tin-stannous chloride, a material somewhat similar to silver–silver chloride, was used in some biosignal electrodes (41).

Electrode material and high electrical performance is generally less critical for external stimulation electrodes such as TENS and external pacing electrodes (current density distribution is the key concern). The majority of commercially available TENS electrodes are molded from an elastomer such as silicone rubber or a plastic such as ethylene vinyl acetate and loaded with electrically conductive carbon black. Mannheimer and Lampe (42) pointed out that the only tangible disadvantage with having a large electrode interface impedance is that more power will be required from the stimulator to drive the stimulating current through the electrodes into the patient.

Graphite-loaded polyesters and similar materials are used in external pacing electrodes, for example. Some are constructed using tin as the metal layer. In early electrodes, the combination of tin and the chloride-based gel gave rise to pitting of the metal. Improvements made to the gels and the use of high purity tin have effectively removed this problem.

Although silver-silver chloride has been and still is used in some external electrostimulation electrodes, it should be used with care. Silver chloride is deposited electrolytically and can therefore be either removed by the passage of current or a thicker, high resistance layer deposited, depending on the polarity of the electrode, which can be a significant problem in iontophoretic transdermal drug delivery and may cause problems in multifunction pads, which include a silver–silver chloride layer to enable distortion-free monitoring of the ECG through electrodes designed to deliver the pacing or defibrillation impulses.

The Skin

Structure of the Skin. The skin is a multi layered organ that covers and protects the body. It is made up of three principal layers—the epidermis, the dermis, and the subcutaneous layer. (*Note:* In the literature, variations exist in the terminology used to denote these layers.)

The epidermis, the outermost layer, is around 100 μm thick, depending on body site. It is the strongest layer, providing a protective barrier against the outside hostile environment. Unlike any other organ of the body, the epidermis renews itself continually. It can be subdivided into several layers, with the basal layer forming the innermost layer and the stratum corneum the outermost layer. Cells in the basal layer constantly multiply and, as they are pushed up toward the skin's external surface, the cells undergo changes. Eventually, layers of compacted, flattened, nonnucleated, dehydrated cells (called corneocytes) form the stratum corneum. These dead cells are continuously being shed and are replaced from the underlying epidermal layers. The intercellular spaces between corneocytes are occupied by arrays of bilaminar membranes with the morphological features of polar lipids (43). This matrix appears to serve to bind the cells and the stratum corneum has been described in terms of corneocyte bricks surrounded by lipid mortar (44). On average, the stratum corneum comprises around 20 cell layers thick and has a thickness of around 10–15 μm. Thickness will, however, vary with the number of cell layers making up the stratum corneum and the state of hydration. On some body areas, it can be several hundred micrometers thick. The epidermal layer is traversed by numerous skin appendages such as hair follicles, sebaceous glands, and sweat glands.

The underlying layers of the epidermis are, in contrast, a relatively aqueous environment. The transition from an essentially nonconductive, lipophilic membrane (the stratum corneum) to an aqueous tissue (viable epidermis and dermis) gives rise to the skin's barrier properties.

The dermis is the second layer of the skin and, with an approximate thickness of 2 mm, is considerably thicker than the epidermis. It is formed from a dense network of connective tissue made of collagen fibers, giving the skin much of its elasticity and strength. Embedded in the dermis are blood vessels, hair follicles, sebaceous and sweat glands, and several types of sensory nerve endings.

The final layer of the skin (the subcutaneus layer) is found beneath the dermis layer. It contains structures of connective tissues and enables the skin on most parts of the body to move freely across the underlying bone structures. It is one of the body's areas for fat storage and acts as a cushion to protect delicate organs lying beneath the skin.

Skin Impedance

Electrical Properties of the Skin. As the stratum corneum is relatively nonconductive, it presents a high impedance to the transmission of electric currents. As a result, the impedance of the skin is the largest component of the overall interelectrode impedance (Fig. 10). Nonetheless, due to the stratum corneum's dielectric properties and its thinness, it permits capacitive coupling between a conductive metal electrode placed on the skin surface and the underlying conductive tissues. One can imagine the relatively nonconductive stratum corneum sandwiched between the conductive electrode and the conductive tissues underlying the stratum corneum forming a parallel-plate capacitor. The stratum corneum's electrical impedance is, therefore, often represented by a simple capacitor, C_{SP}. (The subscript SP refers to Skin and Parallel.)

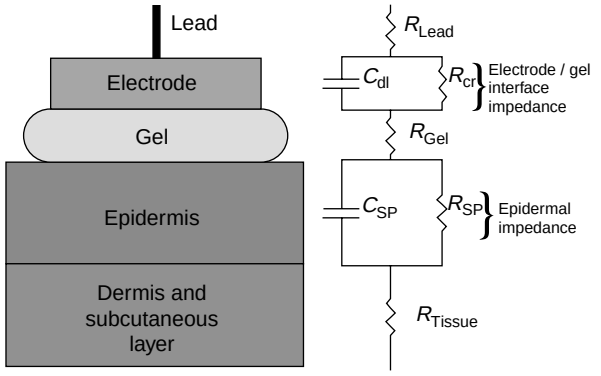


Figure 10. Schematic representation of the skin and its equivalent circuit model.

Some ions do, however, manage to cross the stratum corneum via paracellular pathways and through the skin's appendages (hair follicles, sweat ducts, sebaceous glands, imperfections in the integrity of the skin). As skin appendages extend through the stratum corneum barrier they can act as shunts to the interior. The flow of ionic current can be represented electrically by a large resistance, R_{SP}, in parallel with C_{SP}.

The underlying layers of the epidermis are, in contrast, relatively conductive and can be collectively represented by a tissue resistance, R_{Tissue}.

A simple equivalent circuit model of the overall electrode-gel-skin system is therefore shown in Fig. 10 and includes the electrode lead resistance, R_{Lead}; the electrode–gel interface impedance (the double-layer capacitance, C_{dl}, in parallel with the charge transfer resistance, R_{CT}); the gel resistance, R_{Gel}; the skin impedance (the parallel combination of a capacitance, C_{SP} and a resistance, R_{SP}), and the underlying tissue resistance, R_{Tissue}.

It must be borne in mind that this equivalent circuit model is a simplification of the rather complex electrical properties of the skin. For example, it has been found experimentally (45,46) that the capacitance of the skin is better described by an empirical, constant phase angle impedance, Z_{CPA}, where

$$Z_{CPA(S)} = K_S(j\omega)^{-\alpha} \qquad (20)$$

[which is similar to the empirical expression for the pseudo capacitance often used to represent the nonideal capacitive properties of the electrode interface's double-layer capacitance (Eq. 5)].

K has units of $\Omega \cdot s^{-\alpha}$. The parameter α is constant such that $0 \leq \alpha \leq 1$. The fractional power, α, of the capacitive impedance has been found to be related to the degree of hydration of the stratum corneum (47). If the epidermal layer behaved as a simple capacitance, α would equal unity. The actual value of α, normally around 0.8–0.9, is a measure of the deviation from this ideal behavior.

The use of C_{SP} will, however, be sufficient for this present review. The lead resistance, the gel resistance, the tissue resistance, and the electrode–gel interface impedance are all relatively small in comparison with the large skin impedance. Skin impedance, therefore, generally dominates and will be studied in more depth.

The Skin's Parallel Capacitance, C_{SP}. It was suggested above that the electrode–skin interface can be approximated by a capacitor with the stratum corneum forming the dielectric layer sandwiched between the electrode and the underlying tissues that form the conductive plates. It can be seen from Eq. 2 that the skin's capacitance will increase as the thickness of the stratum corneum decreases, its dielectric constant increases, or the area of the electrode increases.

The number of cell layers in the stratum corneum can range from 12 to 30 (48). Epidermal thickness can, therefore vary greatly for different body sites within the range of about 10 to well over 100 μm (49). The stratum corneum can be, for example, as thick as 400–600 μm in the palm and plantar areas and as little as 10–20 μm on the back, legs, and abdomen (50). The value of the capacitance of the skin is related to the thickness and composition of the stratum corneum and has a typical value in the range $0.02–0.06\ \mu F \cdot cm^{-2}$ when measured using electrodes with "wet" electrolyte gels several minutes following electrode application (51,52). As the stratum corneum is typically at least 10 times as thick on the palms of the hands and soles of the feet as compared with other body areas, the skin capacitance at these points is considerably smaller than at other sites on the body. The stratum corneum on the face and scalp is not as thick as on other body parts and is characterized by large capacitance values.

Dark-skinned subjects have stratum corneum layers that are more dense and contain more layers of cells than fair-skinned subjects (48). Not surprisingly, they are characterized by skin capacitances that are much lower (skin impedances, Eq. 3, much higher) than those for fair-skinned subjects. One should therefore take care when assessing a new electrode system or associated device that they are tested on a range of subjects and skin sites. What may work well on a subject with low skin impedance in a warm and humid environment may be found later to fail on a high impedance subject, especially in a cold or dry environment.

The Skin's Parallel Resistance, R_{SP}. Although the stratum corneum does not easily allow foreign substances to traverse it, some current, carried by ions, manages to flow through it. The difficulty or resistance, this current experiences in passing through the skin is represented in the equivalent circuit (Fig. 10) by the parallel resistance, R_{SP}.

The skin's resistance is highly dependent on the presence and activity of sweat glands and on the presence of other appendageal pathways. An average human skin surface is believed to contain between 200 and 250 sweat ducts on every square centimeter (53). The density of sweat glands varies greatly over the body surface with a value of approximately 370 per cm^2 on the palms of the hands and the soles of the feet and a value of approximately 160 per cm^2 on the forearm (49). The diameter of the ducts can range from 5 to 20 μm. It is, therefore, not surprising that R_{SP} is reported to vary greatly from patient to patient, from body site-to-body site, and with time. The measured values of R_{SP} are much smaller on areas with high densities of sweat glands, such as the palms of the hands (in spite of the thicker stratum corneum layer), especially when the

glands are active in response to thermal or psychophysiological stimuli.

An average human skin surface is reported to contain between 40 and 70 hair follicles per square centimeter (53). The presence of a high density of hair follicles (which act as low resistance shunts) gives rise to a very low value of skin parallel resistance, R_{SP}. However, this observation is counterbalanced by the difficulty in making firm mechanical and electrical contact to hirsute body sites or patients. In such cases, the skin impedance is very large at best. Generally, the electrodes fall off and, hence, require the shaving of the skin site prior to electrode application.

Observed intersite and interpatient variations in skin impedance tend to be due to large variations in R_{SP}. In the low frequency range, dominated by R_{SP}, regional differences in skin impedance were observed by Rothman (54), Lawler et al. (55), and Rosell et al. (56). Low frequency skin impedance was observed to decrease in the following order: thumb, forearm, abdomen and, smallest of all, forehead. Similarily, Almasi and Schmitt (57) observed the low frequency skin (10 Hz) impedance to decrease in the order of outer forearm, leg, inner forearm, back, chest, earlobes, and forehead. The forehead appears to have a very low skin impedance value (58), presumably as a result of the stratum corneum on the face and scalp being thinner than that on other body parts (48) and the presence of a high density of sweat glands. Almasi and Schmitt (57) plotted their average impedance values for the body sites on a complex impedance plot and found that most of the points lay along a "smooth common locus of monotonically increasing phase angle and impedance magnitude." This behavior was successfully interpreted by McAdams and Jossinet (32), who showed that such frequency loci were formed when the skin's parallel resistance varied greatly from site to site while the skin's capacitance remained relatively constant. Two body sites did not fit the locus and, hence, the physical explanation; these sites were the palm and fingertips. These body sites have much larger epidermal thicknesses and, hence, have skin capacitance values much smaller than other body sites.

One must be very careful when assessing different electrode designs or gels. Testing different electrodes on different patients is certain to give misleading results due to the intersubject variations, unless, of course, large numbers of subjects are used and statistically significant differences are observed.

R_{SP} varies greatly over time due to a number of parameters including room temperature and psychophysiological stimuli. The latter effect is exploited in so-called lie detectors. Schmitt and Almasi (59) reported that a considerable daily variation exists in a given subject, and seasonal changes have also been reported (60). Testing a range of electrodes on the same subject but on different days is, therefore, not optimal either, as day-to-day variations in skin impedance, especially fluctuations in R_{SP}, will nullify the validity of this approach. For example, Searle and Kirkup (61) found that the diurnal variations on a given subject for a given electrode was much larger than any difference between the range of electrode designs they tested in any one recording session. It should be further noted that the electrode test sites should be allowed to recover for several days between experiments to enable the skin to recover. For example, pealing off an adhesive electrode will remove some of the underlying stratum corneum. Any electrode subsequently tested on the site will benefit from this prior skin stripping (see below).

Electrode designs must, therefore, be compared *in vivo* by testing them at the same time on the same subject. One must still bear in mind the significant differences in skin impedance that exist over the subject's body, as outlined above. Even testing the electrodes at the same time on a limb of a given subject remains problematic. The different skin sites involved, even if located close together, will give rise to significant differences in the measured electrode–skin impedances, which may be wrongly attributed to the electrode designs or gels under test. Searle and Kirkup (61), for example, showed that testing a range of dry electrode metals on the inner forearm gave rise to potentially very misleading results. Electrodes placed closer to the wrist gave rise to lower impedances due to the presence of a higher concentration of sweat glands.

Electrodes must therefore be repeatedly tested at the same time, under the same conditions, varying their relative positions in order to clearly establish their relative performances. McAdams et al. developed a four-channel impedance monitoring system to enable the simultaneous comparison of electrode designs/gels (62).

Skin Potential Motion Artifact. A potential difference E_S, given by the Nernst equation, exists across the epidermis as a result of ionic concentration differences. This potential varies from patient to patient, from site to site, and depends on gel composition (if used) and skin condition.

The skin surface is normally negative with respect to the inside of the body. Skin potential becomes more negative when sweat glands are active, and palmer and plantar surfaces, with their higher sweat gland concentrations, are the most negative. Increasing gel concentrations of NaCl or KCl also render the site more negative. The parameter E_S has a typical value of 15–30 mV (63).

The dependence of the skin potential on the thickness of the epidermal layer is important to many ECG recording applications. If the thickness of the layer is changed by stretching or pressing down on the skin, the skin potential can vary by 5–10 mV compared with, for example, the 1 or 2 mV ECG signal. As these fluctuations generally result from patient movement, they are termed motion artifact. Motion or skin-deformation artifact is a serious problem during exercise cardiac stress testing of patients on treadmills or exercise bicycles, during ambulatory monitoring, and while monitoring patients lying in bed (64,65) (Fig. 11).

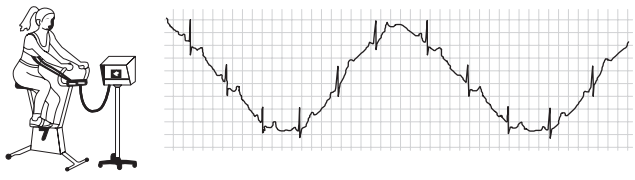


Figure 11. Disturbance of biosignal due to patient movement. (Redrawn from Ref. 65.)

Abrading or puncturing the skin is often used in stress testing to remove or bypass the problem source. Although skin potential increases with gel concentration, artifact gradually decreases with time as the conductive electrode gel soaks into the skin and renders the stratum corneum more conductive. High concentration gels are often used for short-term diagnostic applications where the risk of skin irritation is outweighed by the need for clear traces.

In general, hydrogel-based electrodes (see below) should not be used for stress testing or in other monitoring applications that are likely to suffer motion artifact problems. Hydrogels tend not to hydrate the skin and, hence, do not actively attack the source of the problem. The same comment applies for dry (gel-less) electrodes.

In stress testing and ambulatory event monitoring, modified electrode locations are used to avoid muscular or flabby areas of the body and thus minimize skin-deformation (and EMG) artifact. Stress loops are formed in the connecting leads, which are taped to the patient and used to avoid direct pull on electrodes and the underlying skin. The use of foam-backed electrodes tends to absorb any pull on the electrode and minimizes artifact.

Electrode Gels and Their Effects. Dry electrodes are successfully used in some monitoring applications. Suitably designed gel-less electrodes have advantages when used in the home environment where the patient may not remember or have the time to apply gel electrodes prior to use (66).

For many home-based monitoring applications, electrodes are manufactured from noncorroding materials such as stainless steel, which can be repeatedly washed and reused. Unfortunately, such polarizable materials give rise to poor electrical performances. In order to ensure good, stable electrode potentials, silver–silver chloride electrodes should be used (see below).

Jossinet and McAdams (67) demonstrated that the impedance of a dry electrode decreases pseudoexponentially due to the gradual buildup of sweat under an occlusive, gel-less electrode and the resultant progressive hydration of the underlying skin. Searle and Kirkup (61) reported that the decrease in skin impedance of dry electrodes is polyexponential and requires two time constants, one very short (~45 s) and the other almost 10 times longer (~450 s), possibly indicating two different processes at work.

Given that the surface of the skin is irregular, a flat dry electrode will initially only make contact with a few 'peaks' on the skin surface. Therefore, a smaller effective contact area exists than one would otherwise expect. However, as sweat builds up under the occlusive, dry electrode, a better contact with the skin will result in a relatively rapid increase in the measured value of C_{SP}. Human sweat contains a small amount of sodium chloride [~0.1–0.4% NaCl (49)], and hence serves as a weak electrolyte. It is suggested by the author that this accounts for the shorter time constant. (The longer time constant is probably indicative of the progressive hydration on the underlying skin resulting in a gradual decrease in R_{SP}.) As will be outlined below, R_{SP} is observed to decrease with a time constant of around 10 min in the presence of an electrolyte gel, which agrees quite well with the 7.5 min observed by Searle and Kirkup (61).

Before leaving gel-less electrodes, it should be pointed out that in certain applications that employ very high frequency signals, such as electrical impedance tomography (EIT), the use of a gel pad may not be needed as it will contribute a small but significant contact resistance to the desired measurement (5). In such instances, the use of a very thin spray of moisture onto the electrode surface prior to its firm application to the patient's skin may be all that is required. Profiled dry electrodes firmly pressed onto the skin may also be adequate for certain home-based biosignal applications. If skin impedance is a problem with standard button electrode designs, this can be addressed by increasing the electrode area in the noncritical axis. For example, long, narrow, dry electrodes are used for precordial ECG recording, which enable a large contact area while ensuring sufficient interelectrode distances on the chest (66).

Electrode gels serve (*1*) to ensure a good electrical contact between the electrode and the patient's skin, (*2*) to facilitate the transfer of charge at the electrode-electrolyte interface between the two kinds of charge carrier (electrons in the electrode and ions in the gel), and (*3*) to decrease the large impedance of the stratum corneum.

Two main types of electrode gel exist, viz. wet gels (often described as pastes, creams, or jellies) and hydrogels.

Wet gels are generally composed of water, a thickening agent, a bactericide/fungicide, an ionic salt, and a surfactant (68). The ionic salt is present to achieve the appropriate electrical conductivity of the gel, which will depend on the specific application. As the major portion of ions present in tissue fluids and sweat are sodium, potassium, and chloride (Cl^-), in order to ensure biocompatibility, the ionic salts most commonly used in electrode gels are NaCl (sodium chloride) and KCl (potassium chloride). High concentrations of these salts tend to be better tolerated by the body than other salts. The ions in the gel serve not only to ensure electrical conductivity of the gel but to decrease the skin impedance by diffusing into the skin due to the existing concentration gradient. A relatively high concentration of electrolyte will also decrease the value of the charge transfer resistance (thus rendering the electrode more nonpolarizable).

When a standard pregelled wet electrode is applied to the skin, the gel rapidly fills up the troughs on the electrode and skin surfaces, thus ensuring maximum effective contact area. The skin capacitance, C_{SP}, is therefore observed to initially increase rapidly in value following electrode application and then to remain relatively constant (32). [A similar effect was probably noticed by Searle and Kirkup (61) as a result of sweat accumulation under a dry occlusive electrode.] Although C_{SP} does not exhibit a strong time dependence, it does vary with the electrolyte composition and concentration (49), increasing with increasing concentration (69).

Following electrode application, the skin's parallel resistance, R_P, generally decreases with time in a pseudo exponential manner as the ions in the gel diffuse through the skin rendering it more conductive (32,70) (see Fig. 12).

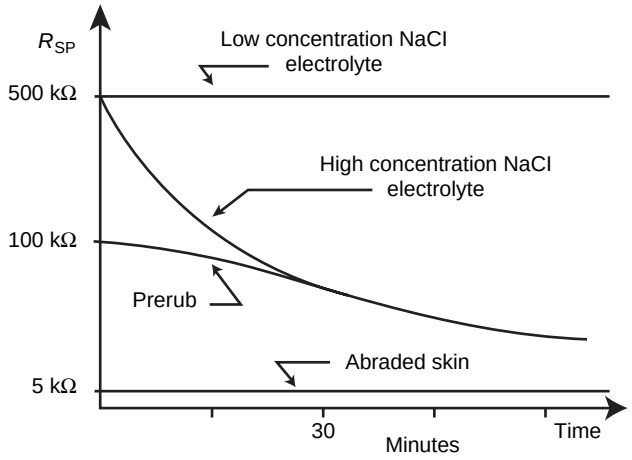


Figure 12. Variation of skin resistance with time for a range of gel concentrations and skin preparation techniques (1).

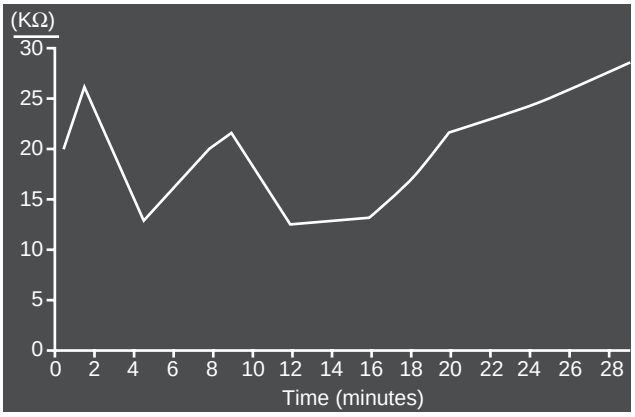


Figure 13. Variation of skin resistance with time for a hydrogel-based electrode. Fluctuations are due to variations in subjects state of relaxation/arousal (32).

It has been observed, however, that when a cold gel is applied to a skin site, the measured value of R_{SP} is observed to initially increase (32,56), which is attributed to the cold gel causing the sweat pores to contract. Once the gel and skin has warmed up, the value of R_{SP} is observed to decrease as the electrolyte ions diffuse through the epidermal layer.

The skin temperature effect should be borne in mind when assessing a range of electrode designs. If, for example, the patient/subject removes his/her shirt just before the tests, the electrodes tested at the start of the experiment will have an advantage (i.e., smaller skin impedance) over those tested later as the uncovered skin sites will gradually cool down with time following removal of the shirt. Meaningful *in vivo* assessment of electrodes is not straightforward, and wrong conclusions can very easily be made by the unaware or the unscrupulous.

The time constant for the skin's parallel resistance, R_{SP}, appears to be inversely proportional to the concentration of the gel. The decay has a time constant of around 10 min (1), thus indicating that it takes almost 1 h for the electrode-skin impedance to decrease to its lowest value. For example, 50/60 Hz interference, linked to mismatch of electrode-skin impedances, is often observed experimentally to decrease with time. One should, therefore, where possible, apply the electrodes to the patient first, for example, before setting up the rest of the measurement system, to enable the skin impedance to decrease as much as possible.

High salt concentrations give rise to a more rapid diffusion of ions into the skin and a more rapid decrease in the skin's parallel resistance, R_{SP} (1,32) (Fig. 13). Such aggressive gels tend to be used in short-term biosignal monitoring applications such as stress testing, where instant, high quality traces are required (71). Biological tissues cannot tolerate long-term exposure to salt concentrations, which depart significantly from physiological levels [~0.9% NaCl for body fluids and around 0.1–0.4% NaCl for human sweat (49)]. Aggressive gels (5% NaCl) should not be used, for example, for the long-term monitoring of bed-ridden patients or for the monitoring of neonates. In the latter case, the incompletely formed skin is very susceptible to skin irritation problems. In any monitoring application, aggressive gels should not be used in combination with skin abrasion (see below). It is especially to be avoided in longer term monitoring applications where the removal of the body's defensive barrier coupled with the long-term exposure to an aggressive gel will lead to sever discomfort to the patient.

The second kind of electrode gel commonly used in electrode systems are hydrogels. Hydrogel-based electrodes have recently become popular for numerous biomedical applications including resting ECG. Hydrogels are solid gels, which originally incorporated natural hydrocolloids (e.g., Karaya gum a polysaccharide obtained from a tree found in India) (68). The use of natural hydrocolloids give rise to variable performances and, in some cases, an unattractive color. Synthetic (e.g., polyvinyl pyrrolidone) hydrocolloids are now widely used.

The use of such solid gels entails numerous advantages when they are used in conjunction with screen-printing or similar technologies. The use of an adhesive hydrogel pad dispenses with the need of the standard gel-impregnated sponge, gel- retaining ring, and surrounding disk of adhesive foam that are used in wet-gel electrode designs. It is possible to construct thin, lightweight, highly flexible electrode arrays with accurately defined electrode/gel areas, shapes, and interelectrode distances (72,73).

Hydrogels also tend to cause less skin irritation compared with wet gels. A simplistic explanation of the advantageous/disadvantageous features of hydrogels is that hydrogel serves principally to ensure a good electrical contact between the skin and the electrode and that they do not significantly affect (compared with wet gels) the properties of the stratum corneum.

The impedance of the gel layer can be represented by a simple resi stance in series with the impedances of the skin and the electrode plate–electrolyte interface. The magnitude of the gel resistance will depend on the composition and concentration of the gel and on the dimensions of the gel layer. Hydrogels are generally more resistive than wet gels. Typical resistivities for wet gels are of the order of 5–500 Ω·cm (the higher the salt concentration, the lower the

resistivity) compared with 800–8000 Ω·cm for hydrogels (the higher resistivity hydrogels tend to be used in cardiac pacing electrodes). Wet ECG electrodes, for example, have a gel layer thickness of around 0.3 cm and typical areas of 3 cm^2. The resistance of a wet gel layer is, therefore, generally in the range 0.5–50 Ω. Although hydrogels have higher resistivities, this disadvantage is generally compensated for by the use of larger gel areas, which need not necessarily entail the use of a larger overall electrode area as the adhesive hydrogel may not require the use of a large surrounding disk of adhesive foam. Another way to compensate for hydrogel's inherent disadvantage is to decrease the gel layer thickness, a variable generally ignored in electrode design even though it can have a significant effect on electrical performances. Many commercial hydrogels used in biosignal monitoring electrodes have layer thicknesses of around 1 mm (compared with around 3 mm for pregelled wet electrodes) and, coupled with larger areas of around 7 cm^2, can lead to hydrogel pad resistances in the range 10–100 Ω (5).

It is suggested that further improvements can be made to the performances of hydrogel electrodes (and wet electrodes) by the use of even thinner gel layers. It must be borne in mind that gel-layer resistance is not solely determined by the dimensions and properties of the gel pad. When a large area gel pad is used in conjunction with a small area sensor, the dimensions of the smaller sensor will largely determine the magnitude of the gel-layer resistance, the overlapping section of gel pad carrying relatively little current, which is important in both biosignal monitoring and electrostimulation applications.

Hydrogels, being hydrophilic, are used for wound dressings in order to absorb exudate. They are, therefore, poor at hydrating the skin and will even absorb surface moisture. With hydrogel electrodes, R_{SP} is observed to fluctuate with sweat gland activity and the subject's state of mental arousal, decreasing during increased activity and gradually increasing again as the hydrogel absorbs the excess surface moisture (74,75). In contrast, C_{SP} remains relatively constant after a slight initial increase (32).

Hydrogels are therefore not only more resistive than wet gels, but they hydrate the skin less effectively and give rise to higher skin impedances (i.e., higher values of R_{SP} and lower values of C_{SP}). Typical values of R_{SP} for hydrogels can be as high as 15 MΩ·cm^2 compared with a high of 5 MΩ·cm^2 for wet gels (75). Once again, this disadvantage can be overcome, at least partially, by the use of larger hydrogel pad areas. An additional way of increasing the value of C_{SP} is the use of thinner hydrogel pads (32).

Skin Preparation Techniques. In the clinical environment, the skin site is often degreased using an alcohol wipe prior to electrode application, which probably removes some of the loose, outermost cells of the stratum corneum and the poorly conducting lipid substances from the surface of the skin (55). However, the use of alcohol wipes may initially increase the impedance of the skin by dehydrating the outer layers of the skin (76). Motion artifact also may increase initially following application of alcohol to the skin (64). When wet gel electrodes are applied to alcohol-wiped skin, the gel will eventually penetrate the degreased skin more readily once the electrode has been on the skin for several minutes, leading to a more rapid decrease in skin impedance and possibly to a decrease in motion artifact, which may not be the case, however, in the case of hydrogel electrodes, which do not actively hydrate the skin. The use of an alcohol wipe accompanied by vigorous rubbing should result in low initial impedances due to the additional mild abrasion.

A related method of rapidly decreasing skin impedance is to prerub the skin site with a high concentration electrolyte, thus forcing the gel into the outer layers of the skin, resulting as in a significant decrease in R_{SP} (Fig. 12) and an increase in C_{SP}, especially when accompanied by vigorous rubbing. Arbo-prep cream is supplied for this purpose and it is claimed to reduce skin resistance by up to 90% (from 40 or 50 to 4 or 5 kΩ, according to an advertisement). Some commercial gels such as Hewlett-Packard's Redox paste contain abrasives such as crushed quartz, which, when rubbed into the skin prior to electrode application, greatly reduce skin impedance. Such aggressive gels should only be used in short-term biosignal monitoring applications such as stress testing where high quality traces are required.

The outer layers of the stratum corneum can also be removed by rubbing the skin with abrasive pads especially designed for this purpose, which can give rise to a major decrease in R_{SP} (Fig. 12) and an increase in C_{SP}.

Unomedical, for example, markets a small disposable skin preparation abrasive pad that, when adhered to the finger tip, can be used to dramatically reduce skin impedance. A Skin Rasp, which resembles a strip of Velcro, is marketed by Medicotest for this purpose. The Quinton Quick-Prep Applicator, rotates the abrasive center of the Quick-Prep electrodes, causing a marked decrease in skin impedance. ECG electrodes are often supplied with abrasive pads built into the electrode release backing.

In skin stripping, the stratum corneum is progressively removed by repeatedly applying and removing adhesive tape to and from the skin (55,77). Skin stripping can greatly decrease skin impedance as a consequence of a dramatic decrease in the value of R_{SP} and an increase in C_{SP}. As the outermost layers of the stratum corneum are the most resistive, the most significant decrease in skin impedance is achieved with the first few strippings (77). Therefore, no need exists for the complete removal of the stratum corneum, which would obviously be clinically unacceptable due to the discomfort (pain, bleeding, or irritation) caused to the patient during and following the recording. The more the skin is abraded for a given gel composition, the sooner discomfort develops and the more severe the irritation. The level of irritation also varies with the salt concentration and the additives present in the gel.

As pointed out above, abrading or stripping the skin is often used in stress testing to decrease motion artifact (63) as well as the 50/60 Hz noise induced by any mismatch of the contact impedances. High concentration gels are also often used for such demanding applications, rapidly soaking the skin and, thus, effectively removing the source of the problem.

The use of both skin abrasion/stripping and an aggressive gel will, however, maximize the potential for severe

skin irritation problems. These approaches should not be used together. Even with the use of mild gels, it is probably unwise to abrade the skin for long-term monitoring applications. The increased length of exposure of the abraded skin to the gel will be conducive to skin irritation. Somewhat surprisingly, long-term monitoring electrodes are sometimes commercially supplied with integral abrasive pads, which is not only risky but probably unnecessary as the use of a suitable mild gel would eventually decrease the skin impedance without the need for skin abrasion.

ELECTRODE DESIGN

External Biosignal Monitoring Electrodes

Historical Background. In 1887, Augustus Waller, using Etienne Jules Marey's modification of the capillary electrometer, obtained surface ECGs (as opposed to recording directly from the exposed heart of an animal) of one of his patients, 'Jimmy'. The patient turned out to be his pet dog. Waller used two buckets of saline to measure the canine ECG, one for the front paws and one for the hind paws (78–80).

Waller eventually succeeded in recording the first human ECG in 1887 using the capillary electrometer (Fig. 14) (81). However, he initially concluded "I do not imagine that electrocardiography is likely to find any very extensive use in the hospital. It can at most be of rare and occasional use to afford a record of some rare anomaly of cardiac action" (82).

It was, therefore, left to a more visionary and tenacious Dutchman, Willem Einthoven, to establish the clinical relevance of this strange new trace and to develop and commercialize a clinically acceptable system based on the string galvanometer. Einthoven's achievement was truly awesome. However, it must be pointed out that he did build (very significantly, it is conceded) on the work of earlier pioneers. The electrode system used, for example, was

Figure 14. Human subject connected to capillary electrometer via large area bucket electrodes (81).

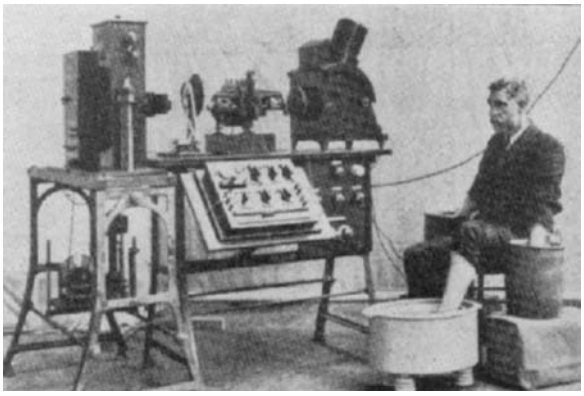


Figure 15. Early commercial ECG machine and electrodes (84).

Waller's bucket electrode, whereas the moving photographic plate recording technique was originated by Marey (83).

The input impedance of Einthoven's galvanometer was such that very low contact impedances were necessary, hence, the very large bucket electrodes (Fig. 15) (84). Obviously, the range of applications was somewhat limited.

Realistically, only one's limbs could be conveniently placed into the buckets. Hence, the use of limb leads exists in electrocardiograpghy, even to the present day. It is, therefore, important to note several points. Present state-of-the-art is often based on historical quirks rather than on a profound scientific basis. The monitoring device and amplifier determined the electrode size, design, and location of the electrodes, which in turn determined the clinical application and the presentation of the physiological data.

Einthoven's device in its early form could not be used for the monitoring of bed-ridden patients or for ambulatory monitoring. These applications had to wait for improvements to be made to the amplifiers, which then enabled the use of smaller electrodes that could be more conveniently attached in other anatomical locations. However, the early monitoring locations and the form of the signals observed became accepted as standard and there is often considerable resistance to novel monitoring scenarios (e.g., smart clothing), which require or are based on different lead systems and present physiological data in a different format to that familiar to the clinician.

In the 1920s vacuum tubes were used to amplify the electrocardiogram instead of the mechanical amplification of the string galvanometer, which lead to smaller, more rugged systems that were transportable (Fig. 16) (84). The input impedances of the new ECG monitors were larger, and the large metal buckets could be replaced by smaller metal-plate electrodes (still large compared with present-day electrodes) (83). These advances enabled bedside monitoring, and, by the 1930s, some ECG devices could be carried to the patient's home. Not unsurprisingly, the new plate electrodes were attached to the limbs, both for historical and practical reasons. The metals used were chosen for their availability and ease of machining (Fig. 17). They included German silver, nickel-silver,

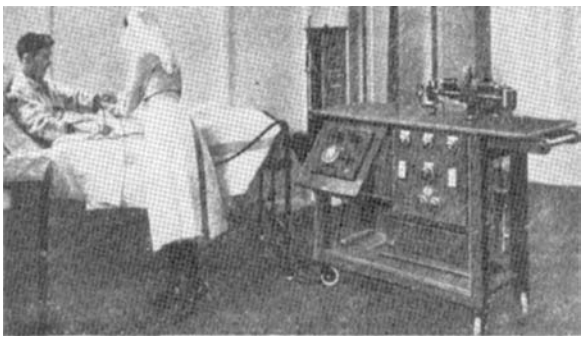


Figure 16. Mobile ECG used to monitor bed-ridden patient in hospital ward circ. 1920 (84).

and nickel-plated steel. The foil plates were used in conjunction with moistened pads of paper toweling, lint, cotton gauze, or sponge and were generally held in place with rubber straps. Around 1935, conductive gels were developed to replace the soaked pads. A wide range of gel ingredients were assessed, and it was noticed that the presence of an abrasive in the gel greatly reduced the skin impedance (5).

They also noted that slightly abrading the skin before applying the electrolyte helped achieve very low skin impedances.

A dry version of the plate electrode was reported by Lewes (85). The multipoint stainless-plate electrode resembled a large nutmeg grater that penetrated the skin when firmly strapped onto the limb or applied to the skin with a slight rotary movement, thus resulting in a very significant reduction in skin impedance.

Modern versions of the limb plate electrode still exist. Some have a convenient spring clip mechanism, which dispenses with the need for the rubber strap.

In the 1930s, clinicians, some using electrodes held on the chest by the patient himself or by another member of clinical staff, experimented with precordial leads and established their clinical value (86). In 1938, the American Heart Association and the Cardiac Society of Great Britain defined the standard positions and wiring of chest leads V1–V6 (87).

Research then focused on the development of electrodes that could be conveniently attached to the chest to enable convenient routine clinical measurements. Several designs

Figure 17. Metal plate limb electrode.

involved a rubber bulb, which was used to create suction sufficient to hold the metal electrode on the chest. One of the first suction electrodes was developed by Rudolph Burger in 1932 for the precordial leads (88). The suction electrode shown in Fig. 18a (85,89) is one developed by Ungerleider (89). Another more recent system incorporated the multipoint electrode of Lewes (85) into the suction head (Fig. 18b). The most popular suction electrode design, widely used around the world and still in use today, was developed by Welch (90) and often called the Welch or Suction cup/bulb electrode (Fig. 19) (3). It consists of a hollow, metallic, cylindrical electrode that makes contact with the skin at its base. A rubber suction bulb fits over the other end of the cylinder. The suction bulb was squeezed while the electrode was held against the skin. Upon releasing the bulb, the electrode is held in place. The suction electrode can be used anywhere on the chest and can even be used on hairy subjects. A single electrode can, if necessary, be used to take a measurement at a given location and then moved to another site.

Although the Welch cup electrode became widely used as a precordial electrode, it could only be realistically used for resting (supine) diagnostic ECG recording. The weight and bulk of the electrode generally rules out its use on upright, ambulatory, or clothed subjects. Since then, more suitable, lightweight, low profile suction electrodes have been developed that are pneumatically connected to remote vacuum pumps (37). Some arrays of suction electrodes are commercially available, for example, for use with exercise bicycles for cardiac stress testing (72).

A method had to be invented to attach small disks of suitable metal and their conductive gel coating to a patient's chest (in the case of ECG) or to other body parts in the case of other biosignal applications, such as EEG and EMG. Simply taping a metal disk to the skin site with a sandwiched gel layer was a method often used (91).

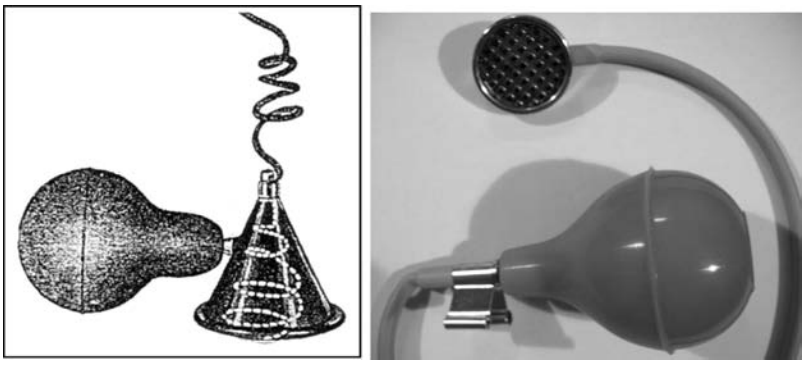


Figure 18. Early designs of suction precordial electrode (a) Ungerleider (89). (b) Lewes (85).

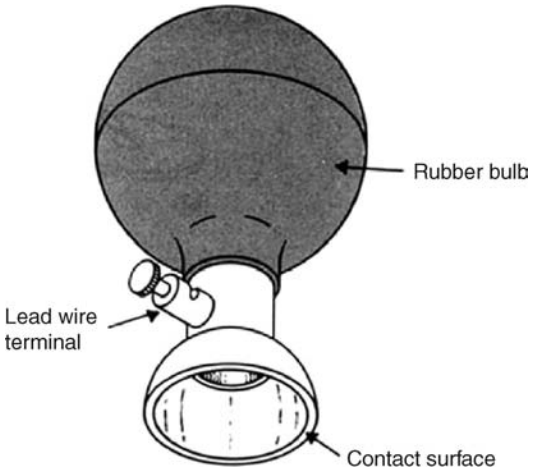


Figure 19. A metallic suction electrode is often used as a precordial electrode on clinical electrocardiographs (3).

Conically formed metal disk electrodes were and often still are used for EEG recordings (Fig. 20). The base of the metal cone is attached to the patient's scalp using elastic bandages, wire mesh, or more recently, using a strong adhesive such as colloidon. Aperture exists in the apex of the cone to enable the introduction into the recessed electrode of electrolyte gel or to enable the abrasion of the underlying skin by means of a blunt hypodermic needle. The cone electrodes were often made of gold as it has high conductivity and inertness, desirable in reusable electrodes. More recently, Ag/AgCl has been used.

Early plate electrode designs were presumably very messy and gave rise to considerable artifact problems. The observed artifacts were attributed to disturbance of the double-layer region at the electrode/skin (or, more precisely, electrode/electrolyte) interface [termed the electrokinetic effect by Khan and Greatbatch (94)]. When the electrode moves with respect to the electrolyte, the distribution of the double layer of charge on electrode interface was thought to change and cause transient fluctuations in the half cell potential or give rise to a streaming potential.

Recessed or floating electrodes were introduced in an effort to protect the electrode–gel interface from such mechanical disturbance and resultant movement artifact.

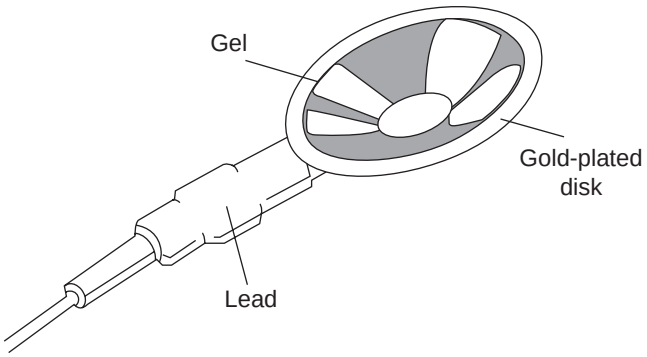


Figure 20. Conically formed metal disk EEG electrodes.

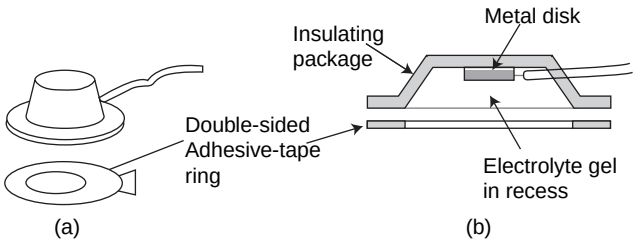


Figure 21. Examples of a floating or recessed biosignal electrode. (a) Recessed electrode with top-hat structure. (b) Cross-sectional view of the electrode in (a)(3).

A metal disk was recessed in a plastic housing that was filled with electrolyte gel prior to application to the patient. The top hat-shaped container was adhered to the skin by means of an annulus of double-sided adhesive tape, (Fig. 21) (3). Later a gel-impregnated sponge was used to ensure good electrical contact between the electrode disk and the skin surface. The electrode disk was, therefore, not in direct contact with the skin, which was found to reduce motion artifact. At first, various metal plates were used as the electrode conductor, then a sintered Ag/AgCl disk with preattached wire ensured better performances for more demanding applications.

Modern Disposable Electrodes. The top hat housing was eventually replaced with a smaller retaining ring or plastic cup and the electrode was held in place by means of a surrounding disk of adhesive foam. The plastic cup holds the gel-impregnated sponge in place and stops the gel from spreading beyond the set boundary, either during storage or use on the patient. Low cost Ag/AgCl-plated plastic eyelets (part of a snap fastener) are used in these disposable electrodes and the leads are connected to the electrodes via the electrodes' snap fastener studs. The rigid retaining ring was, however, uncomfortable as it did not allow the electrode to conform optimally to body contours. It was eventually removed in many modern disposable electrodes and the recess is now often formed by a hole in the adhesive foam layer. The backing label serves to hold the snap and eyelet in place as well as to present the company's logo (Fig. 22). The resultant electrode structure is much more flexible and more comfortable to wear.

The use of a snap fastener-style connection in disposable electrodes has one significant drawback for certain applications. The male stud protruding from the back of the electrode and the female connector required on the connecting lead results in a relatively heavy, large-profile electrode/connector interface, which is less than optimal for applications such as neonatal and pediatric monitoring. The use of such electrodes in long-term monitoring of bedridden patents could lead to considerable discomfort and the heavy connection could also give rise to significant motion artifact problems. The integrated lead design seeks to overcome these disadvantages. A thin, highly flexible lead wire is bonded directly to the back of a specially designed Ag/AgCl-coated eyelet, which results in a very low profile, lightweight electrode-connector system much used in neonatal and pediatric monitoring and attractive for long-term monitoring applications (Fig. 23).

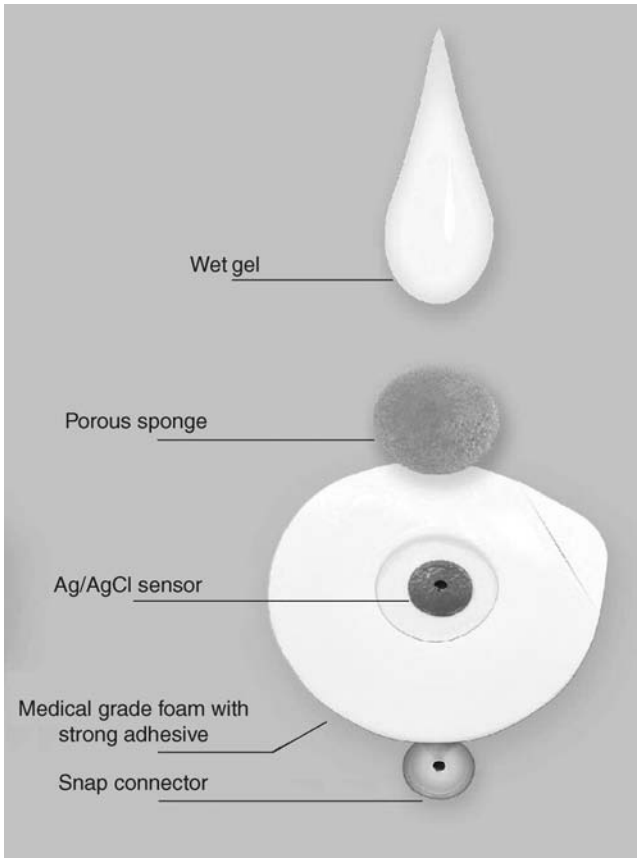


Figure 22. Modern wet gel disposable electrode. (Courtesy of Unomedical A/S.)

In an effort to decrease motion artifact, many electrode designs feature an offset center. The connector, often in the form of a snap fastener, is separated from the gelled sensor by a strip of metal or similar conductive layer. The connector is thus 1 or 2 cm away from the metal–gel–skin interface, and it is possible to connect the lead to the electrode or to pull on the connector without pulling directly on the gelled, skin site, thus causing artifact problems. This design appears well suited for stress testing applications although arguably less so for long-term monitoring of bed-ridden patients due to the bulky connector. The invention was patented by Manley (93) and the concept

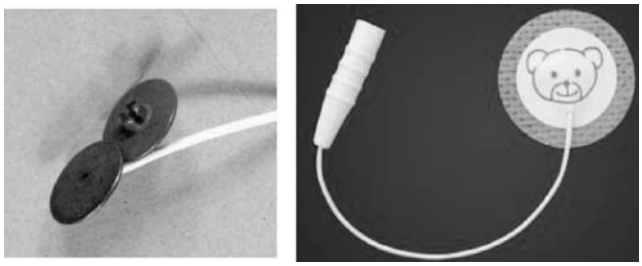


Figure 23. (a) Specially designed Ag/AgCl-coated eyelet with bonded lead. (b) Low profile, lightweight pediatric electrode with bonded lead. (Courtesy of Unomedical A/S.)

Figure 24. An example of an off-set connector electrode. (Courtesy of Unomedical A/S.)

has been commercially exploited very successfully by Ambu A/S.

More recently, many other manufactures supply electrodes with offset connectors (Fig. 24). Leadlock have developed an electrode that incorporates a slit in the foam backing. Once the electrode has been applied to the patient and the lead connected to it, part of the foam backing is used tape down the lead, locking it in place and minimizing direct pull on the connector and underlying electrode/skin interface.

A wide range of backing materials now exists, and some are better suited for specific monitoring applications, types of patients, skin types, and so on. Given the great variety of materials, adhesives, and designs used, the following comments are generalizations.

Open-cell foam layers, made from a plastic such as polyethylene, are much used and have thicknesses typically in the range 1–2 mm. They have a 10–100 μm coating of a pressure-sensitive adhesive, generally a polymeric hydrophobic substance (68,94). Adhesive foams generally give rise to firm adhesion, are resistant to liquids, and tend to cushion lead pull, thus giving rise to less artifact. They are, therefore, generally well-suited to cardiac stress testing and similar applications. However, as they are occlusive and generally have a relatively aggressive adhesive, they can give rise to more skin irritation and they must be used with caution for neonatal and long-term monitoring.

Porous, breathable layers, such as nonwoven clothes or tapes, have the advantages of being soft, stretchable, and conformable with the skin. (*Note:* The term micropore, although sometimes used loosely to describe any breathable backing material, is strictly a 3M product.) Porous layers tend to cause less mechanically induced trauma to the skin, which can occur with more rigid materials and is due to shearing of underlying skin layers. As they are highly air permeable and use milder adhesives, porous tapes cause less skin irritation and are well suited for long-term monitoring. Larger backing areas tend to be required. The gelled center can, however, pull away from the skin as a result of the stretchable backing. Ambu A/S

use a central ring of adhesive around the gelled eyelet to minimize this problem.

As we have already seen, wet (as opposed to solid gels) vary in composition and concentration depending on the application. Aggressive gels with higher concentrations of electrolyte or including abrasive particles are used for short-term, demanding monitoring applications such as cardiac stress testing. Mild gels are used in pediatric and neonatal applications due to the increased vulnerability of the patient's skin. It should be noted that no matter how hypoallergenic a gel or an adhesive is claimed to be, some patients will experience some form of skin reaction to one of the components.

Solid Conductive Adhensive Electrodes. The growing monitoring market has led to the development of even lower-cost disposable electrodes. Solid conductive adhesive or hydrogel electrodes were first introduced by LecTec Corp around 1980 (95). Hydrogels are composed of a hydrocolloid, alcohol, a conductive salt, water, and a preservative. The hydrocolloid use can be either natural (e.g., Karaya gum) or synthetic (e.g., polyvinyl pyrrolidone) (68). Early hydrogel electrodes were based on the natural hydrocolloid, Karaya, which comes from the bark of a tree. The rather unaesthetic appearance of these early gels and the variations in their electrical and mechanical properties limited their widespread acceptance. The use of synthetic hydocolloids, with their more attractive appearance and performances, has led to the recent revolution in electrode design.

Solid adhesive gels reduce the number of electrode parts required, dispensing with the need of a gel-impregnated sponge or a surrounding disk of adhesive tape, which gives rise to small-area, low profile electrodes suitable for neonatal monitoring, especially when coupled with integrated leads as discussed above (Fig. 23b).

Tab solid adhesive electrodes are now widely used for many biosignal monitoring (and stimulation) applications. Thin, highly flexible metallic/conductive foils or printed conductive ink layers are laminated with solid, adhesive hydrogels. A section of the foil or printed layer is left uncovered. Once the electrodes are cut out, the exposed conductive tab acts as a means of connection, the leads being connected via alligator clips (Fig. 25). Electrode design is therefore very simple and manufacturing costs are low. These flexible, low profile electrodes are best used for short-term, resting diagnostic monitoring. Tab electrodes are not suitable for ambulatory or long-term monitoring as the tab connection will cause the electrode to peal off quite easily when pulled from any angle other than directly downward. Also, hydrogels are hydrophilic and tend to absorb moisture, lose their adhesive properties over time, and fall off the patient if an additional adhesive backing is not used. Hydrogels, being solid, do not leave a messy residue on the skin requiring cleaning. Tab electrodes are also repositionable and are reuseable (on the same patient!) in certain home monitoring applications.

When used with an adhesive backing layer, the hydrophilic hydrogels tend to be relatively nondrying (a significant problem with pregelled wet electrodes) and their electrical properties may even improve as they absorb moisture. As they do not actively hydrate or otherwise affect the skin, they tend to be relatively nonirritating compared with wet gels.

Figure 25. Hydrogel-based tab electrode with connector. (Courtesy of Unomedical A/S.)

Some disadvantages exist, however, associated with hydogels. Hydrogels are more resistive than wet gels and, hence, the gel pad resistance will be higher, which can be compensated for by using larger hydrogel pad areas and thinner layer thicknesses as compared with those used with wet gels. Although the area of the solid adhesive gel in a tab electrode, for example, is considerably larger for this reason than that in a standard disposable wet gelled electrode, the absence of a surrounding adhesive layer results in the tab electrode having a smaller overall area.

Hydrogels, being hydrophilic, are poor at hydrating the skin and may even absorb surface moisture. They, therefore, give rise to larger skin impedances. This disadvantage can also be overcome, at least partially, by the use of larger hydrogel pad areas. Hydrogels are also more expensive than wet gels but generally lead to less expensive electrodes due to the simpler designs involved.

Hydrogels are more sensitive to motion artifact as they do not actively hydrate the skin. They are, therefore, not well-suited for stress testing.

The use of such solid gels entails numerous advantages when they are used in conjunction with screen printing technology (73,96), especially for body surface mapping and similar applications. It is possible to construct thin, lightweight, highly flexible electrode arrays with accurately defined electrode/gel areas, shapes, and interelectrode distances for a wide range of novel stimulation and biosignal recording applications. As the solid gel will not spread between electrodes, it is possible to position electrodes very close together without electrical shorting (Fig. 26).

Wearable Electrodes for Personalized Health. The recent and continuous trend toward home-based and ambulatory monitoring for personalized healthcare, although exciting and potentially leading to a revolution in healthcare provision, necessitates even more demanding performance

Figure 26. Cardiac mapping electrode harness.

criteria for the monitoring sensors (97,98). Many groups around the world are seeking to incorporate electrodes into clothing in order to monitor military personnel, firefighters and eventually the average citizen who wishes to monitor his or her health. Systems already exist on the market (e.g., Life Shirt) that resemble waistcoats into which one plugs-in standard ECG electrodes and other sensors. These sensors are removed and replaced periodically by the subject and, hence, require the knowledgable involvement of the motivated wearer, presently military personnel, athletes, rescue workers, and so on.

For the more widespread use of wearable monitoring systems, especially by the average citizen, the system must be very easy and comfortable to use and require no preparation—literally as simple as putting on their shirt. Electrodes must, therefore, (*1*) require no prepping, (*2*) be located in the correct location once the smart garment is put on, (*3*) make good electrical contact with the skin, (*4*) not give rise to motion artifact problems, (*5*) not cause discomfort or skin irritation problems, and (*6*) be reusable and machine-washable. Although much work has been carried out in this novel area, it is not surprising given the above list of required performance criteria that the electrodes/sensors tend to form the bottleneck in the success of the overall monitoring systems. One must, therefore, not simply choose an electrode with as conductive a metal element as possible. Unfortunately, it would often appear that the associated electronic systems are first developed and the electrode design is left to the end, almost as an afterthought. The author would, therefore, suggest that researchers start with the desired biosignal and establish the optimal body site(s) and electrode design for the given application before developing the rest of the monitoring system. This process may involve the use of novel lead or montage electrode positions in order to conveniently pick up artifact-free signals. Although this method will necessitate clinicians interpreting nontraditional waveforms, it will at least enable feasible monitoring and, as it involves novel body sites and electrode designs, it may well be patentable. After all, if it is not patented and commercialized, it will not benefit the patient.

One of the most promising smart garments is that developed under a European Fifth Framework programme called WEALTHY (Wearable Health Care System) (Fig. 27). WEALTHY is a wearable, fully integrated system, able to monitor a range of physiological parameters including electrocardiogram, respiration, posture, temperature, and a movement index. Fabric electrodes are made using conductive fibers woven into the stretchable yarn of the body-contour hugging garment and connections are integrated into the fabric structure (Fig. 27b). Various membranes are being assessed to ensure optimal electrode-skin contact and minimize skin irritation. The garment is comfortable and can be worn during everyday activities. It is washable and easy to put on.

External Electrostimulation Electrodes

Historical Background. The evolution of external stimulation electrode design shares some of the key landmarks as the development of biosignal monitoring electrode and, hence, this section will be somewhat shorter.

From the mid-1700s, when electrostatic generators were used to deliver arguably therapeutic impulses to various parts of the body, handheld (by the practitioner) electrodes had to be designed capable of delivering the impulses to the patient without shocking the practitioner who was holding them against the body part in question. The electrodes used tended to be simply long metal rods insulated with wooden handles (Fig. 28) (99). Although the electrodes were initially terminated in a simple metallic sphere, more exotic terminations were soon invented as these were observed to lead to different therapeutic effects on the body by means of the variations in the streams of the electric fluid. A modern parallel would be the use of different pencil electrodes (ball, loop, and needle) in electrosurgery for different effect.

Figure 27. (a) The WEALTHY physiological monitoring vest with integrated sensors. (b) An early version of the fabric electrodes.

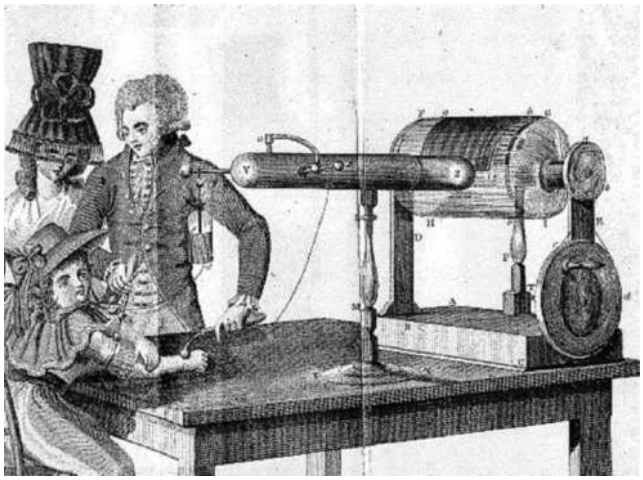


Figure 28. Early electrostatic generator and handheld electrodes (99).

Around 1800 came the discovery of Galvanism (dc current) and Voltaic Piles (early batteries). Numerous examples exist of practitioners using their handheld probe for localized effect, and the second contact to the patient was made by means of a container of water into which the patient put a hand or foot (Fig. 29) (100).

Following the discoveries of self-and mutual induction (~1830), Guillaume Duchenne made great contributions to the clinical application of the new Faradic current. At that time, much interest existed in the localization of what became known as motor points. It was common to combine the prevailing interest in acupuncture and use needles to stimulate muscles and nerves under the skin, termed electropuncture (83). Duchenne was not happy with this approach and developed his own electrodes for localized electrization. His electrodes were in various shapes (disks spheroids, and cones) covered with leather

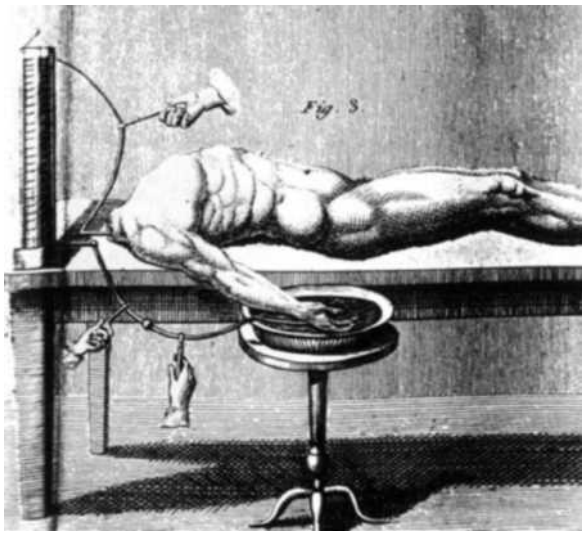


Figure 29. Container electrode. Used on the dead and the living (100).

Figure 30. Duchenne's moistened conical electrodes for localized electrization (101).

moistened with salt water prior to application (101). (Fig. 30).

During the 1900s, as with biosignal monitoring electrodes, electrostimulation electrodes involved the use of simple metal buckets or receptacles, filled with water or another electrolyte, into which the subject introduced their foot or hand, especially in early iontophoretic applications. Obviously, the range of applications was somewhat limited. Metal probes were still manually pressed against skin for short-term applications. The electrodes were either gelled before application or had moistened chamois coverings similar to those used by Duchenne. In the 1950s, early external pacing and defibrillator electrodes, termed paddles because of their shape, consisted of bare metal disks made of noncorrosive material and were simply pressed against the patient's chest (102) see Fig. 31.

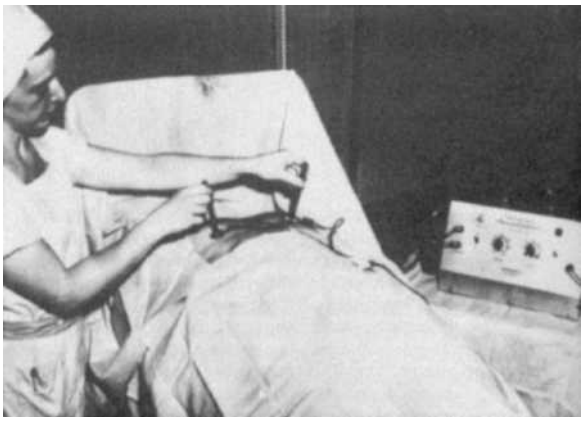


Figure 31. Early pacing equipment and handheld electrodes (102).

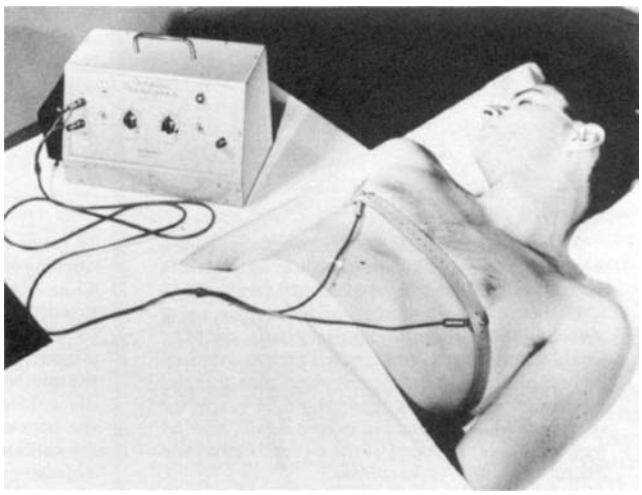


Figure 32. Early pacing equipment and metal-plate electrodes (103).

Rigid metal plate electrodes were eventually held in place with rubber straps on the limbs and even the thorax (Fig. 32) (103). Some of these electrodes were and still are made with rigid stainless-steel plates (104). The foil plates were generally used in conjunction with moistened pads of paper toweling, lint, cotton gauze, or sponge. The pads were moistened by the therapist prior to electrode application with water or electrolyte. Such electrodes could be easily reused by simply washing and regelling the electrode. Being rigid, however, these plate electrodes did not always make optimal contact with the body surface and gave rise to current density hot spots. External cardiac pacing at this time, for example, was very painful (83).

Malleable metal foil electrodes were the next evolutionary step in electrode design. Malleable electrodes have been made using a range of metals including tinplate lead and aluminium foils (105). Such electrodes had the advantage of being able to conform, to some extent, with body contours, thus ensuring a better, more comfortable contact between the electrode and the patient than was the case with rigid plates. Wrinkles in malleable metal foil could, however, encourage preferential current flow through small areas of the gel and into the patient.

More convenient, disposable pregelled foil electrodes were then developed for a range of external electrostimulation applications. The metal foil was laminated onto an adhesive foam backing. A gel-impregnated sponge layer was located on top of the metal layer and the complete electrode is attached to the patient by means of the surrounding layer of adhesive backing foam.

Unfortunately, the wet gel in these disposable pregelled electrodes tended to pool to one side, depending on how they were stored, giving rise once again to current density hotspots. More recently, the gel-impregnated sponge layer has been replaced by a conductive adhesive gel layer, as it does not have the potential for pooling to one area during storage and it does not squeeze out under pressure (68) (Fig. 33).

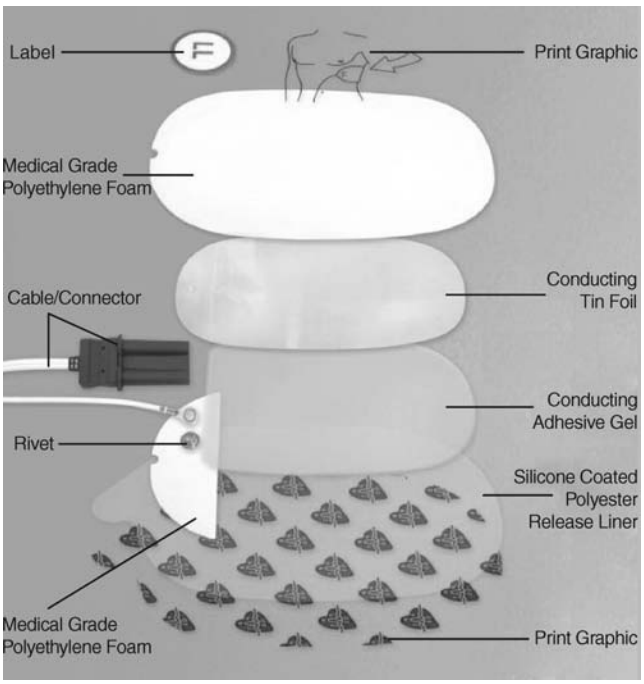


Figure 33. Construction of a modern external pacing or defibrillation electrode (courtesy Unomedical A/S).

Current Density Considerations. The distribution of current density under an electrode is an important parameter when designing and using electrostimulation electrodes. In the simplest case, current density (the amount of current per unit of conduction area) is inversely proportional to the electrode/skin contact area. For a given current, the current density under a small-area electrode will be higher and more localized than that under a large-area electrode. Generally, an optimal electrode area exists for a given therapeutic application, based on a range of criteria including the anatomical position and size of the nerve/muscle/organ and the relative positions and sizes of the electrodes. Small electrodes are, therefore, well existd to target precisely known points such as motor points. If a small electrode is used in conjunction with a large-area electrode, the effect is more pronounced under the smaller of the two. In such monopolar stimulation, the small electrode is often used as the active electrode to target the therapeutic effect. The larger electrode is simply used to complete the electrical circuit and is termed the indifferent or dispersive electrode. The use of two equally sized electrodes is termed biopolar stimulation. In TENS, for example, biopolar stimulation is often used to stimulate large muscle groups sandwiched between the (large) electrodes. Too large an area of electrode, however, may cause the current to spread to neighboring tissues.

High current densities can cause tissue injury due to, among other things, heating effects. The passage of electricity through any conductor will cause the dissipation of heat within that conductor. The amount of heat generated in a tissue depends on spatial and temporal patterns of current density and tissue resistivity (49).

The total energy dissipated at an electrode–skin interface is given by the formula:

$$E = I^2Rt \tag{21}$$

where

E is energy dissipated (J)
I is root-mean-squared (rms) electrode current (A)
t is the duration of current flow (s)
R is the real part of the impedance at the electrode site (Ω).

The change in temperature at the skin, ΔT, site is proportional to the energy dissipated and, hence, ΔT is proportional to I^2Rt. When skin or muscle tissue is heated to about 45 °C for prolonged periods, thermal damage can result. For short durations (i.e., <5 s), a temperature rise approaching 70 °C would be needed to cause heat damage.

As the electrode–skin resistance, R, is not generally known for a given site, it is often found convenient to use a heating factor (HF), where

$$\mathrm{HF} = I^2t \quad (\mathrm{A}^2 \cdot \mathrm{s}) \tag{22}$$

Assuming uniform current density distribution under an electrode, it is possible to calculate the minimum area of electrode necessary to achieve therapeutic effect and avoid tissue trauma (42). In theory, the applied currents flowing through standard dispersive electrodes used for electrosurgery, for example, will generally not give rise to sufficiently high overall current densities to cause thermal damage. However, analysis shows that current density distribution is not uniform under a stimulation electrode and that localized hotspots can occur and cause considerable pain and trauma to the patient when applying apparently safe' therapeutic impulses (49). At best, in cases such as TENS, the applied current may have to be limited to less than therapeutic values due to the patients discomfort (68).

Many potential sources exist of accidentally high current densities. Wrinkles or breaks in the metal electrodes, gel squeezing out from under the electrode or drying out, electrodes partially peeling off the skin, poor electrode application, and so on can encourage preferential current flow through small areas of the gel and into the patient. However, current density hotspots can also occur due to poor electrode design, and a considerable amount of research has been and is being spent investigating this important problem.

In this presentation, stimulation electrodes have been divided into conductive electrodes and resistive electrodes in order to facilitate the review of the various design features.

With highly conductive metal electrodes, such as those used for external cardiac pacing, defibrillation, or electrosurgery, current density hotspots are observed to occur under the perimeter of the electrode, often evidenced in the past by annular-shaped burns to the patient (49,106).

Current density hotspot problems are now often studied using thermal imaging cameras. Thermograms of the patient's skin (or a substitute such as pig skin) are taken immediately following the application of a given series of pulses and the removal of the electrode under test (Fig. 34). Increases in skin temperature reflect the magnitude of the current density at a particular point (107).

Wiley and Webster (108) showed that current flow through a circular electrode placed on a semi-infinite medium could be solved analytically. They found that for an electrode of radius, a, and total current, I_0, into the electrode, the current density into the body as a function of radial distance from the center, r, was given by:

$$J(r,0) = \frac{J_0}{2[1-(r/a)^2]^{1/2}} \quad (\mathrm{A} \cdot \mathrm{cm}^{-2}) \tag{23}$$

where $J_0 = I_0/\pi a^2$, (i.e., a hypothetical uniform current density).

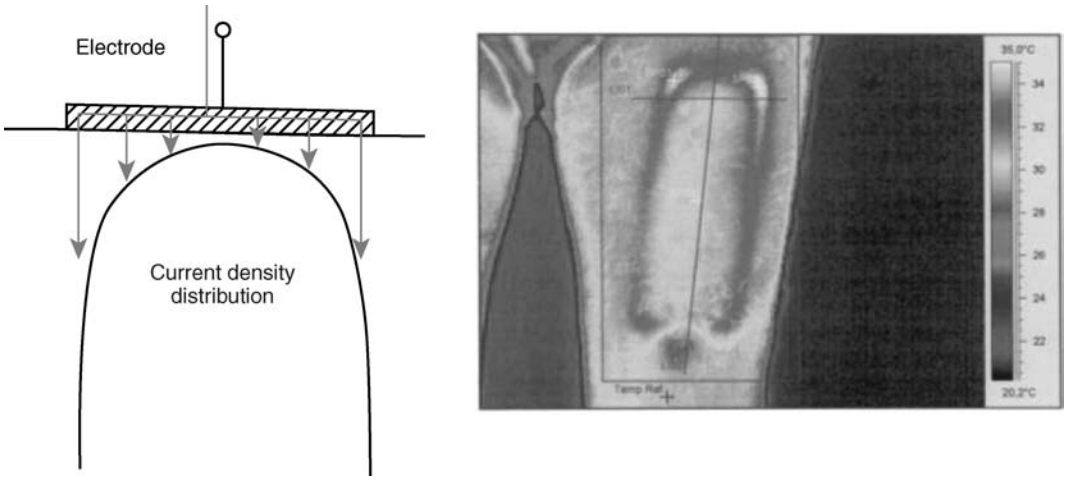


Figure 34. (a) Schematic representation of the current density distribution under a conductive electrode plate. (b) Thermal image of the skin under an electrosurgical electrode following testing.

As a result of the approximations made in deriving this simple equation, the value of current density at the edge (when $r = a$) would theoretically approach infinity. More realistically, the current density at the perimeter can be around three times higher than that at the center of the electrode (109) (Fig. 34). The above equation shows that the middle portion of the electrode is relatively ineffective in carrying the current as half the total current flows through an outermost annulus 0.14 a wide or one-seventh of the radius.

Efforts in this area concentrate on encouraging more of the current to flow through the central portion of the electrode.

A main concern in the design of the conductive stimulation electrodes used for external cardiac pacing, defibrillation, or electrosurgery is the decrease in the high current densities observed at the edges.

In TENS, relatively resistive conductive rubber is often used and the opposite problem develops. When current is introduced into the conductive rubber (via a small metallic connector), it tends to flow into the skin immediately under the connector rather than laterally through the resistive electrode. Efforts in this area concentrate on encouraging the current to flow laterally through more of the electrode surface.

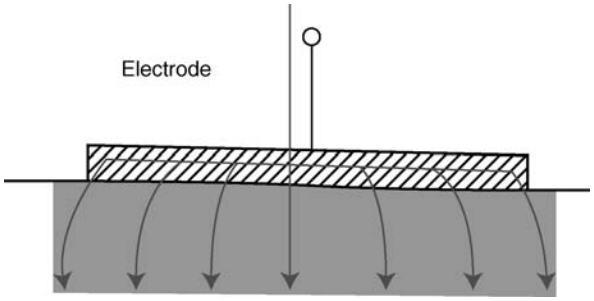


Figure 35. Schematic representation of the current density distribution under an electrode plate coupled with a larger gel pad. Current is allowed to spread out beyond the boundaries of the metal plate, thus minimizing the edge effect.

Modern Electrode Designs

Conductive Electrodes. Electrosurgery, external cardiac pacing, and defibrillation share a common problem: Electrodes tend to deliver or sink a substantial portion of the outgoing or incoming current through their peripheral area as opposed to providing a uniform current density along their surface. This problem is referred to in the literature as the fringe, edge, or perimeter effect.

Many suggestions have been made to reduce this edge effect observed with metal electrodes, including:

1. Increasing overall area of the electrode. Obviously, an increase in electrode area will lead to a decrease in current density (110,111). However, it is generally not practical to use very large electrodes as the applied electrical field must be sufficiently focused to stimulate the targeted tissues and them alone. Also, a strong commercial interest exists in decreasing the size of the electrodes to save money and to facilitate packaging and storage of the electrodes.
2. Avoiding sharp edges in the metal plate (110,111). It has long been observed that square or rectangular electrodes with angular edges concentrate the electrical field at their corners, giving rise to current density hotspots in these locations. Using round electrodes or rectangular ones with rounded edges have been found advantageous in this regard.
3. Making the gel pad slightly larger than the electrode to enable the electric field lines to spread out before reaching the skin (111) (Fig. 35). Using a gel pad much larger than the size of the metal plate has less effect than would be expected as the perimeter of such a large gel pad will carry little current and the additional gel is electrically redundant, which arguably applies to some of the snap connection electrodes used in TENS that do not have an additional current dispersing element.
4. Increasing the overall resistance or thickness of the gel layer in order to give the current more time to spread out evenly through the gel (110,111). It is well-known that, in applications such as external cardiac pacing, the use of relatively resistive gels decreases the pain and burning to the patient's chest. Krasteva and Papazov (110) suggest that the use of a layer of intermediate resistivity, comparable with that of the underlying tissues, optimally improves the distribution. However, in other applications such as external defibrillation, a high resistance gel pad would lead to energy wastage and a decrease in the desired therapeutic effect. Taken to its logical conclusion, this approach results in the coating of the electrode metal plate surface with a dielectric film. Such capacitive electrodes have been shown to give rise to nearly uniform current densities (107).
5. Increasing the resistance or thickness of the gel at the edges. Kim et al. (112) proposed covering the electrode metal with resistive gel of increasing resistivity as one moved out from the center toward the periphery, according to a specific relation with respect to the electrode radius. Although an intriguing concept, the commercial manufacture of such an electrode system is not yet feasible.
6. Making the electrode conductive plate progressively more resistive toward the peripheral edge of the electrode. Wiley and Webster (108) suggested subdividing the electrode plate into concentric segments and connecting external resistors to the individual segments. The connected resistors had progressively higher resistances toward the periphery in order to equalize the currents in the separate segments. A simpler system that has been successfully commercialized was patented by Netherly and Carim (113). A resistive layer is deposited on the outer edge of the electrode conductive plate, thus forcing more current to flow through the central portion of the electrode (Fig. 36). Krasteva and Papazov (110) demonstrated theoretically that a high resistivity perimeter ring

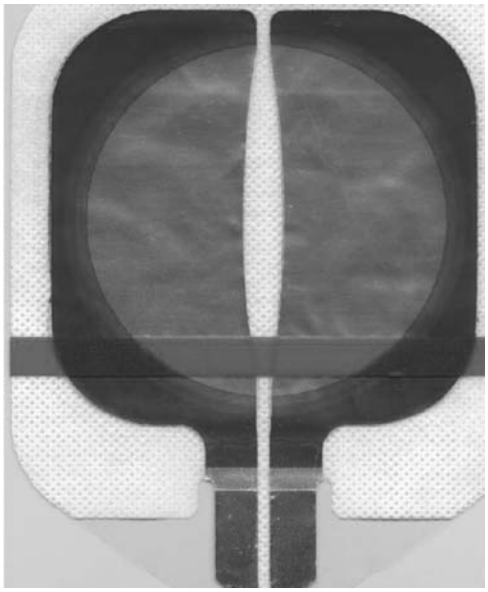


Figure 36. 3M's electrosurgical dispersive electrode. Note the green lossy dielectric material deposited around the peripheral edge of the electrode.

decreased the maximum periphery current by 12% without increasing the total interface resistance, hence the resistance to the defibrillation current.

Another related approach to this problem starts with a resistive graphite-based conductive layer and progressively builds up multilayered (at least two) coatings of more conductive silver/silver chloride toward the center of the electrode (114). The perimeter resistance is approximately 200 times that of the center, this is the technique exploited in Medtronic EDGE system electrodes for defibrillation, noninvasive pacing, and ECG monitoring (Fig. 37). It is claimed that the design distributes the current density evenly over the entire surface area of the electrode, rather than concentrating it at the edges.

7. Scalloping or otherwise shaping the edges of the metallic plate so that the length of the perimeter is increased and, hence, the peripheral current density is decreased. Over the years, various designs have incorporated this concept. For example, it has been shown that using a figure-eight design rather than a rectangular metal plate reduced the maximum temperature (reflective of current density) by 30–50% (107). An alternative design is shown in Fig. 38. Caution is advised with this approach as the formation of fingers in the metal layer may serve only to concentrate the current at the tips of the fingers, and one could be effectively left with a reduced peripheral area.
8. Making holes in the central portion of the metal plate in order to provide internal peripheral edges to block the lateral flow of current. Some early claims were made that holes in the metal layer improved current density under the electrode. Presumably, it was believed that the holes blocked the current from flowing from the connector to the edge of the metal plate, forcing it to flow into the patient at the edges. It is the authors belief that such holes in the metal plate achieve little apart from further decreasing the area of the electrode and, if anything, increasing the current density at the edges. This impression appears to be confirmed by the work of Krasteva and Papazov (110), who investigated electrode structures with openings in the metal plate for skin breathing.

Figure 37. Medtronic's EDGE system electrode.

Figure 38. An electrostimulation electrode with cut-out metal plate in an effort to increase peripheral edge. The green sponge impregnated with gel has largely been removed to facilitate inspection of the underlying plate.

The author has suggested that the use of concave slits in the metal layer rather than circular holes may well have a favorable effect on current density distribution with the concave internal peripheral edges effectively blocking the lateral flow of current, forcing the trapped current to flow into the gel and, thus achieving a more uniform current density distribution over the surface of the conductive layer (115). Early work on the project with an industrial

partner appeared promising, but the work was never completed.

Resistive Electrodes. A TENS electrode system appears relatively simple and generally comprises a conductive plate, an ion-containing gel, a means of attachment to the skin, and a means of connection to the stimulators lead. Mannheimer and Lampe (42) pointed out, however, that of all the component parts of the overall TENS system, the electrode–skin interface has probably been the least understood and the most problematic. In addition to influencing the effectiveness of the treatment, poor electrode design can give rise to electrically, chemically, and mechanically induced skin irritation and trauma to the patient.

Initially, electrodes originally designed for ECG and other biosignal monitoring applications were used with TENS units, and some still are. Larger, more suitable electrode designs were eventually developed in order to reduce the current densities under the electrodes, to reduce skin irritation problems, and to increase stimulation comfort (116).

A large percentage of commercially available TENS electrodes are now molded from an elastomer (e.g., silicone rubber) or a plastic (e.g., ethylene vinyl acetate) and loaded with electrically conductive carbon black (Fig. 39) (3). Very few irritation or allergic reactions have been reported for conductive rubber electrodes as they do not generate the corrosion products often observed with metal electrodes (42). The great advantage of such electrodes is that they can be molded into almost any size or shape and a wide range of choice exists in the market. They can be made sufficiently thin to have high flexibility and, thus, are able to conform with body contours, making them suitable for a wide range of TENS applications.

Conductive rubber electrodes are often used in conjunction with an electrolyte gel and attached to the patient using elastic straps or custom-cut disks or patches of adhesive tape. Expanded polyester foam tends to give the most secure adhesion. However, as this backing is occlusive, the use of foam can give rise to skin irritation problems. Breathable, cloth-like fabrics allow the transmission of air and moisture and generally cause less skin irritation problems. Cloth-like materials tend to stretch, however (an advantage when it accommodates skin stretching due to movement), which can lead to the electrode working loose and making poor contact with the skin, possibly resulting in current density hot spots.

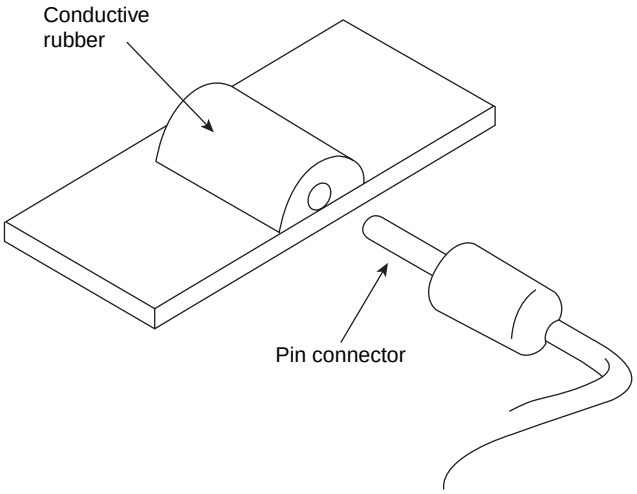


Figure 39. Carbon-filled silicone rubber electrode (3).

Wet gels can squeeze out from under parts of the electrode and give rise to increased current densities in other areas. The use of hydrogel minimizes this problem source (68). Conductive, adhesive pads of solid hydrogel help ensure firm electrical contact between the electrode and the skin, reduce the incidence of current density hotspots, and often simplify the design of the electrode. As a large surrounding disk of adhesive tape is not required, the electrode size can be reduced to the active electrode area. These solid gel pads can be, depending on the application, replaced, refreshed, or simply reused in various semi- or totally-reusable electrode systems. In some applications, the gel pads can be removed and the conductive rubber electrode cleaned and regelled with a fresh gel pad for further use. In other cases, the electrode can be intermittently reused, on the same patient, by rehydrating the gel pad. Such reusable electrodes are ideal for home-based patient use.

One disadvantage with such conductive rubber electrodes is that they are relatively resistive. More power is required to drive the stimulating current through the resistive electrodes into the body and achieve the desired stimulation. Therefore, some reduction in battery life may occur which is generally not a significant problem, however.

A more serious problem involves current density distribution under the resistive electrode. When current is applied through the conductive rubber (via a small metallic connector), it tends to flow into the skin immediately under the connector rather than laterally through the resistive electrode, thus giving rise to a current density hotspot under the connector, which effectively, is the opposite problem to that encountered when using highly conductive electrodes.

Efforts to overcome this problem include incorporating conductive elements in the rubber to more evenly to help spread the current over the entire interface surface. Some electrodes have a thin metallic layer coated onto the back of the conductive rubber, which appear to give rise to the most uniform current density profiles (68).

The growing home-based market has led to the great variety of low cost disposable and reusable electrodes that are generally based on solid adhesive gels. Some electrodes are made using conductive cloth-like materials, thin metallic foils, aluminized carbon-filled mylar, or wire strands. Electrical connection is generally made to these electrodes via alligator clips, snap fasteners, or pin connectors. Many of these hydrogel-based electrodes can be trimmed to the desired size or shape by simply cutting with a pair of scissors. The current density profiles under these electrodes will very much depend on the relative resistivities of the metal and gel layers as well as on the actual design.

Snap fastener designs resembling standard ECG electrodes and are available with hydrodel pads or sponge

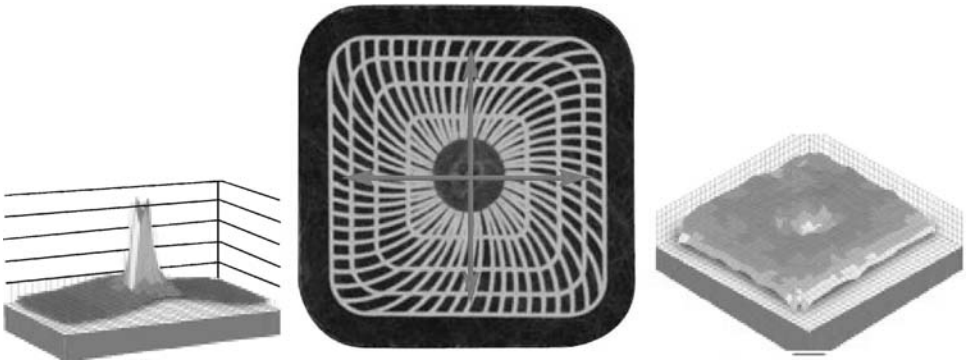


Figure 40. (a) Current density distribution under a conventional snap electrode. (b) Axelguaard's UltraStim snap electrode with current controlling grid. (c) Current density distribution under Axelguaard's UltraStim snap electrode. (Courtesy of Axelgaard Ltd.)

disks containing low chloride wet gels (to minimize skin irritation). These desists may require an additional current-dispersing element to ensure that the current spreads out beyond the immediate confines of the eyelet electrode (68). Axelgaard Ltds UltraStim Snap Electrodes feature a highly conductive grid pattern printed on to a conductive flexible layer and coated with a moderately conductive adhesive gel layer. The conductivity of the conductive pattern is controlled through the use of various grid designs with preselected line widths and spacing as well as thickness and ink compositions (117). The pattern is thus used to control and optimize the spread of electric current over the surface of the electrode with an intentional current drop off toward the edge of the electrode (Fig. 40).

It is interesting to note that considering current density under conductive and resistive electrodes leads to a similar optimal design. To improve conductive electrodes, one places a resistive layer between it and the patient. To improve a resistive electrode, one puts a conducting layer either behind or in front of the resistive layer. Such sandwich electrode designs appear promising for a range of electrostimulation applications.

Garment Electrodes. A range of researchers in the TENS, FES, and body-toning areas of electrostimulation are endeavoring to incorporate electrodes in to body hugging garments to enable the convenient and accurate application of a (large) number of electrodes to the body part to be stimulated. The use of a large number of electrodes can enable, for example, several muscle groups to be stimulated together or sequentially in a coordinated manner to achieve a more natural movement of a limb. Garments are already on the market that resemble tight-fitting cycling shorts and have integrated wires and connectors for the attachment of standard TENS (or similar) snap electrodes prior to application. Other, more challenging designs include the integration of reusable electrodes into a stretchable garment.

Implant Electrodes

Implantable monitors/stimulators and their electrodes are used, or are being developed, for a wide range of applications, including cardiac pacing and defibrillation, cochlear implants; urinary control, phrenic nerve stimulation for respiration control; functional electrical stimulation of limbs; vagal stimulation for control of epilepsy, spinal stimulation for chronic pain relief, deep-brain stimulation for Parkinsons disease or depression, bonehealing, and several visual neuroprostheses.

Implanted monitoring electrodes are used to more accurately pick up the desired signal while minimizing the contributions of extraneous signals. Implanted stimulation electrodes deliver the applied waveform more selectively to the targeted tissue, making the therapy more effective and, as the stimulation electrode is generally implanted away from cutaneous pain receptors and afferent nerve fibers, more comfortable for the patient. One significant drawback, however, is the greater potential for damage from improper electrode design, installation, and use.

The design of an implant electrode will depend greatly on the anatomical structure it is to be implanted against, into, or around. Electrodes can be or have been implanted in, on, or near a given muscle; in, on, or around a given nerve; in, on, or around a given bone; in, on, or around the spinal cord; and in or on the surface of the cerebellar cortex.

A review of all of these designs is beyond the scope of this chapter. The reader is referred to the appropriate chapters in this Encyclopedia.

To facilitate this overview of some of the key design possibilities, two main application areas will be concentrated on: muscular and neural electrodes. Muscular (especially cardiac) electrodes, using more traditional electrode fabrication, are presented in a separate section. Neural electrodes will be largely covered in the section on newer microelectrodes constructed using thin-film and similar techniques. These categories are very loose and a considerable degree of overlap obviously exists between applications and the various electrode fabrication techniques. Once again, the reader is referred to the appropriate chapters in this Encyclopedia for more detailed descriptions of electrodes and their fabrication for specific applications.

Cardiac electrodes are the most important example of muscular electrodes. As cardiac pacemakers and defibrillators have the longest and most successful track records as implantable devices, much of the science underpinning the newer (and future) implantable devices (muscular, neural, and other) has been developed by the cardiac implant pioneers. Key contributions were made in the areas of implant electrodes, biomaterials, and powersources, to

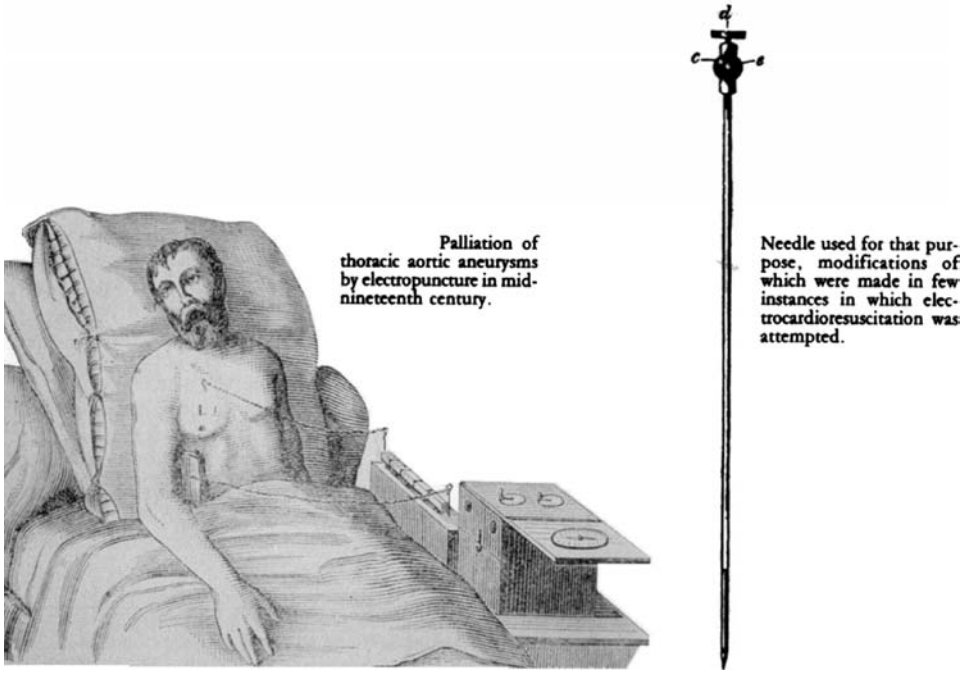


Figure 41. Kimer's electropuncture of the heart (103).

name but a few. The present review of electrodes is, therefore, largely based on cardiac electrodes. However, ideas can be gleaned from this area and, once suitably customized, applied to others with success. The advantageous aspects of biphasic impulses (discovered in the 1950s, arguably earlier) is a good example of something being rediscovered in a range of stimulation applications over the past few decades

Historical Background. Implant stimulation electrodes, or at least percutaneous stimulation electrodes, predate the earliest biosignal monitoring electrodes. In the early 1800s, there appeared a renewal of interest in acupuncture in Europe that had been introduced into Europe in the second half of the eighteenth century by Jesuit missionaries. In 1825, Sarlandière was the first to apply an electric (galvanic) current to thin metal needle electrodes (derived from acupuncture needles) thus creating electropuncture for the application of current to specific points on or in the body (98,118).

Electropuncture soon became the accepted method of stimulating muscles, nerves, or organs beneath the skin (83). Electropuncture of the heart was first attempted by Krimer in 1828 without recorded success (Fig. 41) (103). This technique was then abandoned for several decades. Meanwhile, W. Morton successfully introduced the use of ether as an anesthetic in 1846. Eventually, chloroform was found to be more suitable although cardiac arrest was a frequent complication of chloroform anesthesia in those early days. In 1871, Steiner overanesthetized horses, dogs, cats, and rabbits to produce cardiac arrest. He reported successfully applying an intermittent galvanic current to a percutaneous needle in the heart to evoke rhythmic contractions. Terms such as galvano and farado puncture soon started to appear in the literature (103).

In the early 1900s, cardio-stimulating drugs such as epinephrine were injected directly into the heart of sudden death victims by means of a large needle inserted through the chest wall to restore automatic activity. It was eventually established that one of the key factors in the occasional success of these intracardiac injection procedures was the actual puncture of the heart wall rather than the medication administered. Based on this observation, Hymen went on to build the first hand-cranked, spring-driven artificial pacemaker (119). He used transthoracic needle electrodes plunged into the atrium and even introduced the concept of using a biopolar needle arrangement as in having the two electrodes so close together that only a small pathway is concerned in the electric arc established by the heart muscle, an irritable point is produced (103).

In the 1950s, Lillehei, Weirich, and others pioneered the use of cardiac pacing for the management of heart block accidentally resulting from cardiac surgery and for other emergency cardiac treatment. Slender wire electrodes were implanted into the myocardium before closing the chest with the connecting leads thus exiting through the chest wall. Pacing impulses could then be delivered through these wires for a week or so until the heart healed. Once the heart had recovered, the electrodes were pulled out. Early versions of these electrodes consisted of silver-plated braided copper wires insulated with polyethylene or Teflon (103,120).

In 1958, Furman and Schwedel reported the first instance of transvenous pacing of the heart. They inserted a unipolar catheter electrode into the right ventricle of the patient through a superficial vein and paced the heart via the endocardial surface. The electrode used was a solid copper wire with a bare terminal tip (120). The electrode was withdrawn once the patients heart resumed its own idioventrcular rhythm. Although the cardiac pacing employed by Lillehei et al. and by Furman and Schwedel

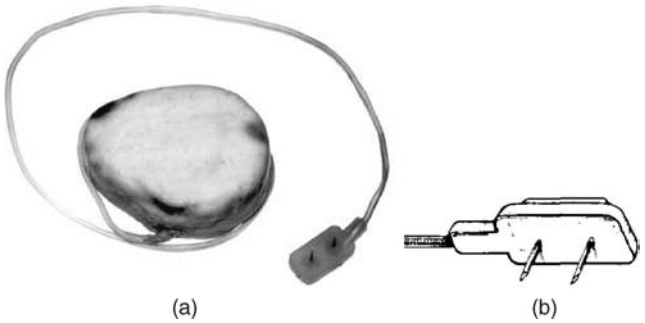


Figure 42. (a) Chardack, Gage, and Greatbatch's wholly implantable pacemaker and the Hunter–Roth bipolar intramyocardial electrode (83) (b) Diagram of an early Hunter–Roth lead with two bipolar myocardial pin electrodes (35).

was much better tolerated than the external stimulation of Zoll, perielectrode infection was a major drawback as was the transport of the external pacemaker. The development of an implantable pacemaker, therefore, became the goal of several groups around the world.

In 1958, Senning and Elmqvist successfully implanted the first Pacemaker without leads emerging from the patient's chest to invite infection. The implanted unit was powered by cells that were recharged from outside the body using a line-connected, vacuum-tube radio-frequency generator. The electrodes/leads used were stainless-steel wires. The second version of the unit failed due to a lead fracture one week following implantation. It was then decided to abandon pacemaker therapy for this patient until better leads were developed (120).

At that time, many of the electrodes used were unipolar. The active electrode tended to consist of the bared tip of an insulated wire implanted in the myocardium whereas the indifferent electrode was a similar wire implanted subcutaneously in the chest wall. Unfortunately, the stimulation threshold was observed to rise following implantation during longer-term pacing. This increase in threshold was thought to be due to the development of scar tissue around the active electrode. Hunter and Roth developed a bipolar electrode system in 1959. This electrode consisted of two rigid, 0.5 cm long, stainless-steel pins attached to a silicone rubber patch. The cathode–anode pins were positioned in to myocardial stab wounds surgically created for the purpose and the pad was then sutured to the epicardial surface (35). The lead wire was a Teflon-coated, multistrand stainless-steel wire with an outer sleeve made from silicone rubber tubing (121).

In 1960, Chardack, Gage, and Greatbatch successfully produced a wholly-implantable battery-powered pacemaker (Fig. 42a). Initially, they used a pair of multistrand stainless steel wires in a Teflon sleeve with the bare ends sutured to the myocardium (35). Other metallic formulations were tried, such as solid wire, silver wire, stainless steel, orthodontic gold, and platinum and its alloys (122).

They eventually adopted the Hunter–Roth intramyocardial electrode (Fig. 42a and b). Considerable surgery was required as the pacemaker had to be implanted into the abdomen and the electrodes were sutured to the heart wall. The bipolar electrode did, however, dispense with the need of a dispersive chest electrode and the associated pain it caused (103). Stimulation thresholds tended to stabilize at much lower levels with this electrode (120), which enabled successful pacing for many months.

Breakage of lead wires, due to metal fatigue, was a major concern. One of the main problem areas occurred at the point were the two metal components were welded together (121). Corrosion also occurred at the small-area stainless-steel anode, causing cessation of pacing within a few months.

Chardack et al. (123) devised a replacement for the Hunter–Roth electrode based on a continuous helical coil of platinum-iridium (Fig. 43). The electrode was simply a few turns of the coiled lead wire, exposed and extended to enable fibrous tissue to grow between the spirals and firmly anchor the electrode in place. The use of a helical coil greatly increased flexibility and decreased the number of fatigue failures, as did the use of one continuous wire (without a join) for both lead and electrode. The use of the same metal for lead and electrode also had the advantage of preventing corrosion from galvanic action. Additionally, platinum-iridium is more corrosion resistant than the metals used in many electrodes prior to Chardack's electrode.

The sutureless screw-in lead was later introduced by Hunter in 1973 (35). The screw-in electrode was simply rotated into the myocardium and did not require a stab wound or sutures for insertion. The electrode was effectively the means of attachment. As this corkscrew electrode tended not to dislodge, it dominated pacing for a long time and is still used today for many epimyocardial implants (Fig. 44).

A thoracotomy was required to attach many of the above electrodes to the heart, which complicated surgical procedure and resulted in a 10% early mortality (122). The first so-called modern pacemaker, which combined an implanted generator and a transvenous lead, was developed simultaneously in 1962 by Parsonnet and Lagergren (124,125). The endocardial catheter electrodes could be

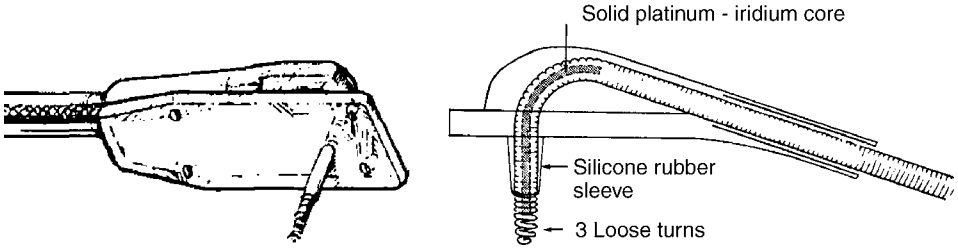


Figure 43. The "Chardack" electrode (a) (35) (b) Chardack et al. (123).

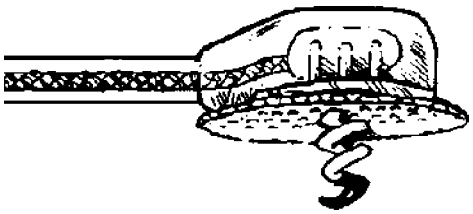


Figure 44. The corkscrew myocardial electrode (35).

installed under local anesthesia, and this approach virtually eliminated early mortality. As they did not require the opening of the chest cavity, the use of catheter leads opened the field of pacemaker implantation to non-surgeons in later decades.

To minimize the risk of venous perforation, the electrode leads were made flexible by winding bands of stainless steel around a core of textile fibers (120). The electrode was a small stainless-steel cylinder at the end of the catheter.

As time passed, the transvenous route progressively evolved over the myocardial approach, so much so that, at present, the transverse route is almost exclusively used for pacemaker implantation.

Cardiac pacing has been the earliest and most successful example of implanted electrodes and associated hardware. Many present and future developments in other implanted electrostimulation (and biosignal recording) areas are and will be based, to a large extent, on the pioneering work carried out in the pacing area.

Some Modern Electrode Designs. With the early transvenous leads, the stimulation threshold was observed to greatly increase if the electrode pulled away even slightly from the myocardium. A wide variety of active fixation devices was therefore invented. These devices included springs, deployable radiating needles, barbs, hooks, claws and screws designed to anchor the electrodes by actively penetrating the myocardium (35,126). The "Bisping" transvenous screw-in electrode is the most popular, as it allows the screw helix to be extended from the tip once the lead has been successfully threaded through the vein and located against the desired part of the heart (Fig. 45). It can be used as a combined anchor and electrode. The screw can be retracted allowing for an easier extraction of the lead, when necessary. (*Note:* A similar design of electrode is used for detecting the fetal electrocardiogram during labor. The intracutaneous needles are screwed in to the fetus' presenting scalp. Similar designs are also used in EEG monitoring.)

A wide variety of passive fixation devices were also invented. Various tines, flanges, and other soft, pliant projections were formed at the distal end of the lead, generally as an extension of the silicone or polyurethane lead insulation, and designed to passively and atraumatically wedge the electrode between endocardial structures such as trabeculae (Fig. 46). In some designs, the electrode has the form of a closed-loop helical coil that, when twisted clockwise, becomes lodged in the trabeculae (126).

Early electrodes had smooth metal surfaces. Techniques were then developed to roughen the surface in the hope of encouraging tissue in-growth, thus locking the electrode in place, minimizing mechanical irritation and excessive fibrous encapsulation, and ensuring low chronic stimulation thresholds. Studies found that porous electrodes did indeed achieve better fixation, thinner fibrous capsules, and stable thresholds.

A variety of porous electrode tips have been developed including totally porous structures such as CPIs meshed screen electrode and electrodes whose surfaces had been textured using a range of techniques (Fig. 47). Porous surfaces have been generated by coating metal surfaces with metallic granules, by sintering metal spheres to form a network of cavities, and by laser-drilling the surface of electrodes (126).

Not only does roughening improve electrode fixation and threshold stability, it has been found to have a very advantageous effect on interface impedance. From a stimulation point of view, one is keen to use a small-area electrode to increase current density at the small tip and

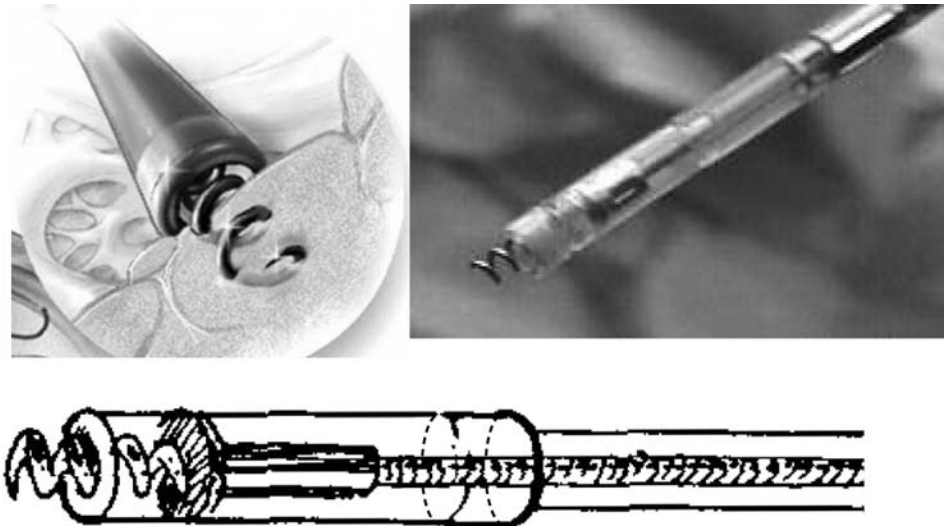


Figure 45. The "Bisping" transvenous screw-in lead with the helical screw electrode extended. [From S.S. Barold's The Third Decade of Cardiac Pacing (35).]

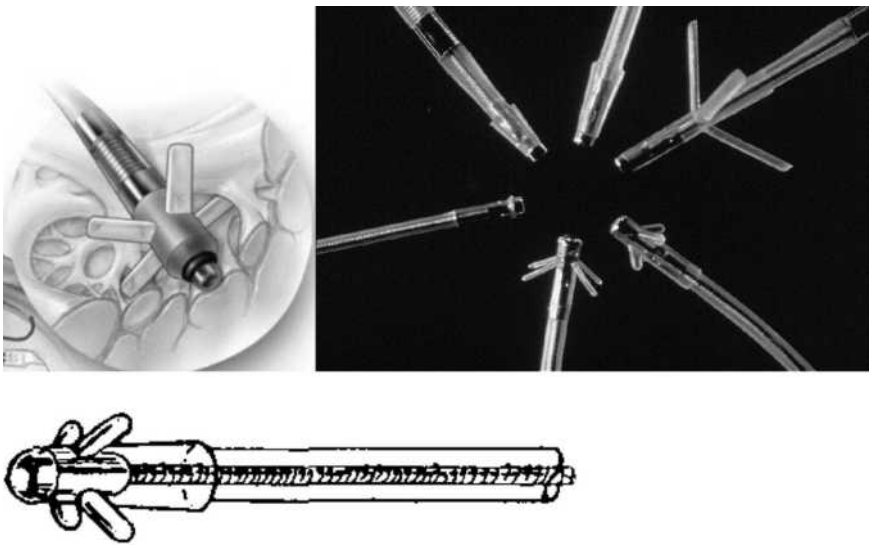


Figure 46. Transvenous lead with silicone rubber tines (35).

thus decrease stimulation threshold. A small-area electrode also has a high pacing impedance that can decrease current drain on the generator and thus prolong implant life (35). However, one would also like to have a low sensing impedance in order to avoid excessive attenuation of the cardiac signal. Two ways that exit these conflicting criteria can be optimized, modifying the electrode surface or modifying its design.

Roughening the surface of a small-area electrode increases its effective area without changing its geometric or outer envelope surface area (35). Many electrode systems have been developed that incorporate this concept—electrodes with terms such as activated, porous, and sintered in their company's description. These electrodes have been found to be effective in lowering the electrode interface impedance under small-signal sensing conditions. [Note: the electrode–electrolyte interface is very nonlinear and, hence, smaller under stimulation.] Unfortunately, the reduction in interface impedance has been erroneously interpreted as rendering the electrode nonpolarizable. As stated previously, the word polarization appears to be used in a rather vague manner and has been used as the explanation of, among other things, the nonlinearity of the interface impedance as well as its frequency- and time-dependence. The fact that the current or voltage response to a step in voltage or current is not a simple step has been attributed to polarization. The observed transient responses are merely due to the presence of the double layer capacitance (see Figs. 8, and 9). Roughening the surface of an electrode effectively increases the area of the interface and the value of C_{dl}, which in turn results in an increase in the response's time constant ($T = R_{CT}C_{dl}$). The observed response thus looks stretched out along the time axis. This flattened response has been mistaken for that of a purely resistive, nonpolarizable electrode. At any rate, roughening the surface of the electrode almost gives us the best of both worlds, a noble or inert electrode with a low interface impedance.

Another way of achieving a small simulation surface area (high current density) while ensuring a large-sensing surface area (low interface impedance) is to modify the design of the electrode.

The porous electrode of Amundson involved a hemispherical platinum screen that enclosed a ball of compacted 20 μm diameter platinum–iridium fibers (127). As electrolyte could penetrate this three-dimensional (3D) or multilayered electrode, the design resulted in a major increase in

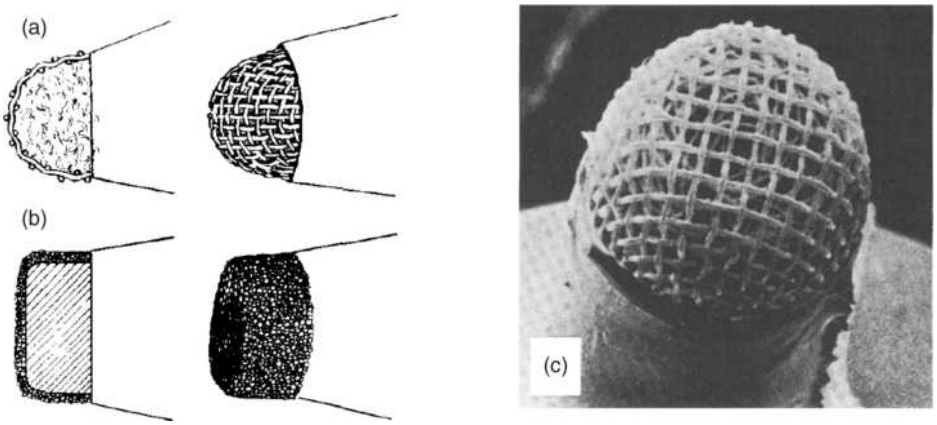


Figure 47. (a) Cross-section of a totally porous electrode (35). (b) Cross-section of a porous surface electrode (35). (c) Photo of totally porous electrode (126).

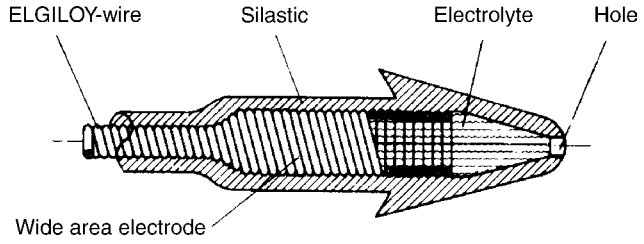


Figure 48. The differential current density (DCD) electrode (129).

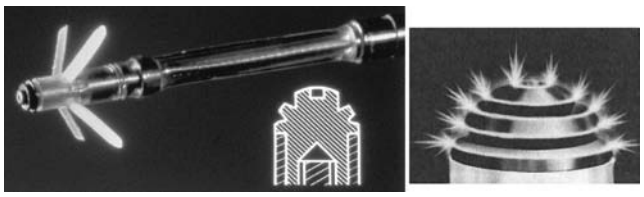


Figure 49. The target tip electrode. Microporous, plantinized plantinum electrode. The target appearance is due to shallow grooves separated by peaks. (Courtesy Medtronic, Inc.)

effective surface area as well as promoting tissue in-growth and long-term stability of thresholds. Lagergren et al. (128) introduced the birdcage design, which also exploited some of these features (126).

One interesting example of a modified design by Parsonnet et al. involved the use of an electrolyte-filled hollow electrode, called a differential current density (DCD) electrode (129). The actual stimulating electrode is the mouth of the electrolyte filled pore, which can be small to provide high current density at the point of contact with tissue (Fig. 48). The inside of the hollow electrode chamber has a large metallic surface (a helical coil forming a cylinder) and thus gives rise to a low electrode-electrolyte interface impedance.

Figure 48 appears to be an electrode design that could readily be customized and used in a wide range of monitoring or stimulation applications. The electrode–electrolyte interface is effectively recessed and protected from any disturbance, a further advantage to those already listed above.

Several other designs exist that aim to achieve a similar effect by manipulating the current density distribution around an electrode tip. Electrodes with complex shapes have irregular patterns of current density with localized hotspots at points of greatest curvature (126). It is possible to exploit these areas of high current density for stimulation purposes while the larger overall surface area gives rise to a low interface impedance (130). A hollow, ring-tipped electrode (effectively similar to the DCD electrode) has a large current density at its annular mouth while having a large electrode–electrolyte interface area. Such electrodes are reported to have better stimulation thresholds and sensing characteristics than hemispherical designs and have proved popular. Several manufactures have combined this ring-tip design with increased surface porosity (126). Other related designs include a dish-shaped electrode for edge-focusing of current (with laser-drilled pores for interface impedance reduction) and a grooved hemispherical platinum electrode coated with platinum black particles (target-tip electrode, Fig. 49) (126,130).

Steroid-eluting electrodes were introduced in 1983 in an effort to minimize the growth of connective tissue. The first-generation electrode was made of titanium, with a platinum-coated porous titanium surface (Fig. 50). The electrodes incorporated a silicone core that was impregnated with a small quantity an anti-inflammatory corticosteroid (34). Upon implant, the steroid is gradually eluted into the interface between the lead electrode and the endocardium, reducing the inflammation and fibrosis that would normally occur. Steroid-eluting leads are characterized by a lower long-term capture threshold. Similar improvements in capture thresholds have been achieved (131,132).

Most cardiac electrodes now involve the combination of drugs and complex surface structures at the macro and micro scales.

For newer applications, such as Cochlear implant electrodes, an array of electrodes is involved. In some such multielectrode applications, one may be interested not only in the current density profiles under the surfaces of the individual electrodes, but in the interplay between the electrical fields produced by the electrodes in the array in the hope of achieving more effective stimulation or more effectively imitate physiological stimulation. For example, the Clarion hi-focus electrode system of Advanced Bionics Corp. incorporates 16 electrodes in a flexible array that are designed to deliver improved focused stimulation to the auditory nerve (133).

Microelectrodes. Over the past few years, exciting developments have taken place in areas of biomedical

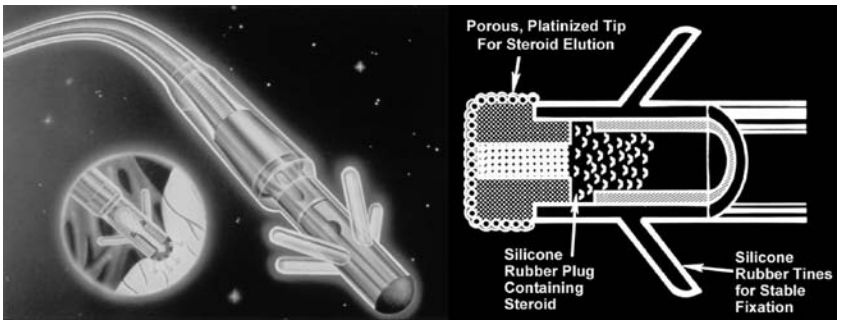


Figure 50. (a) Steroid-eluting electrode. (b) Cross-sectional diagram of an early design of steroid-eluting electrode. Behind the electrode is the silicone rubber plug compounded with steroid. (Courtesy Medtronic, Inc.)

engineering that involve implantable devices for the recording or stimulation of the nervous system.

In the previous section, we saw the success in commercializing pacemakers. Other implant devices that have also reached the patient in clinical routine practice or research settings include Cochlear implants to restore hearing; deep-brain stimulators to alleviate symptoms of Parkinson's Disease and depression; vagal nerve stimulators to minimize the effects of epilepsy; as well as FES systems to restore or improve function in the upper extremity, lower extremity, bladder and bowel, and respiratory system (134,135). Other areas of research that are likely to come to fruition within the next few years include various visual prostheses to restore functional vision in the profoundly blind and the exploration of the brain–computer interface (134,136,137).

Much of the early research in these areas started around the 1960s (135). Where possible and appropriate, surface and percutaneous electrodes were first used to establish the feasibility of the given recording/therapy. Early implant electrodes involved fine metallic wires or small disks placed near, in, on, or around the targeted muscle or nerve. The fabrication of these electrodes was time-consuming and the electrode properties were not very reproducible given the variations in areas, surfaces, inter-electrode distances, and so on, which was particularly a problem when several electrodes were to be used in an array. As the demands on human implantable diagnostic/stimulation devices increases, an increased need for a larger number of smaller-area electrodes with well-defined and reproducible surfaces and dimensions generally occors. Although, due to their high level of specificity, muscle-based electrodes will continue to be used, new electrode designs tend to concentrate more on direct nerve stimulation as this may provide more complete muscle recruitment and the same electrode may successfully recruit several muscles, thus reducing the number of electrode leads required (135). Electrodes are, therefore, needed that can interface electrically with the neural system at the micrometer scale (136).

For example, the goal for a high resolution retinal prosthesis is a 1000-electrode stimulating array in a 5×5 mm package (137). If this area of research is to be clinically successful and if the other areas are to continue to improve, microelectrodes must be (and are being) manufactured using the thin-film technologies associated with the IC circuit industry. Microfabrication involves either material deposition or removal. Either rigid silicon wafers or flexible polyimide substrates act as platforms for the microelectrodes and associated circuitry. The deposited films (for connectors, leads, electrodes, or insulation) are produced by electroplating, evaporation, and sputtering. The layers can be photo-patterned and etched to submicrometer resolutions and finally encapsulated in biomaterials such as diamond-like carbon, bioceramic, or a biocompatible polymer. Processes such as photolithography, reactive ion etching (RIE), CMOS processing, MEMS processing, focused ion beam patterning, and AFM lithography can be used to achieve the desired microelectrode design.

The benefits of a microfabrication approach include a high degree of reproducibility in physical, chemical, and electrical characteristics. Microfabrication is a high yield, low cost process once the design and processing sequence have been developed. Additionally, precise control of the spatial distribution of electrode sites exists, which may be of interest when seeking to optimally stimulate or record from a target site. A high packing density of electrode sites for a given implant volume is also readily achievable using photolithographic techniques. The possibility exits of incorporating the interface circuitry directly on the microsensor platform thus reducing the need for complex interconnections.

The widespread availability of silicon micromanufacturing techniques has enabled the fabrication of a range of silicon-based wedge- or needle-shaped electrodes to allow penetration of the nervous tissue. 3D arrays of such structures have been developed for insertion into, for example, the cortex to detect local potentials (134).

1D arrays of electrodes are fabricated using lithographic patterning and deposition of thin-film metal leads and electrodes onto not only silicon, but also glass and even flexible polyimide substrates (136). Much of the work on silicon-based microprobe fabrication has been pioneered at the Center for Integrated Sensors and Circuits at the University of Michigan.

A 3D electrode array can be fabricated by assembling a range of 1D probes (such as those shown in Fig. 51). As each probe has multiple recording sites along its length, the complete volume of the tissue under study can be assessed, giving rise to very dense sampling. The Michigan Probe has evolved a large number of single-shaft, multishaft, and 3-D-stacked microelectrode arrays (136).

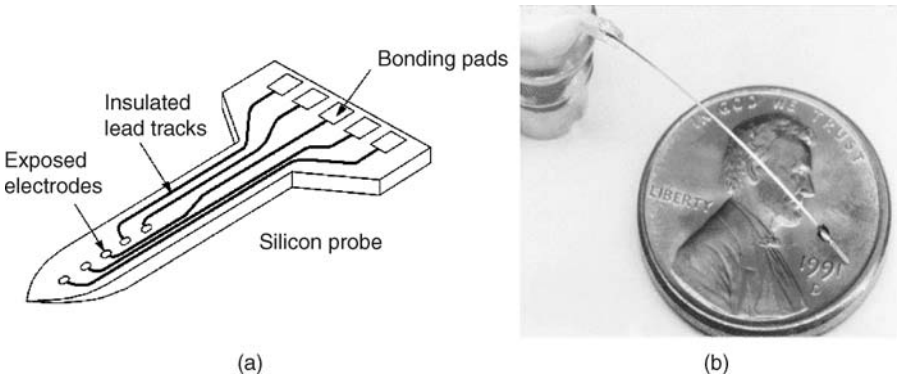


Figure 51. (a) Multielectrode silicon probe. [After Drake et al. (138).] (b) Michigan micromachined multielectrode probe for recording and stimulation of central nervous system.

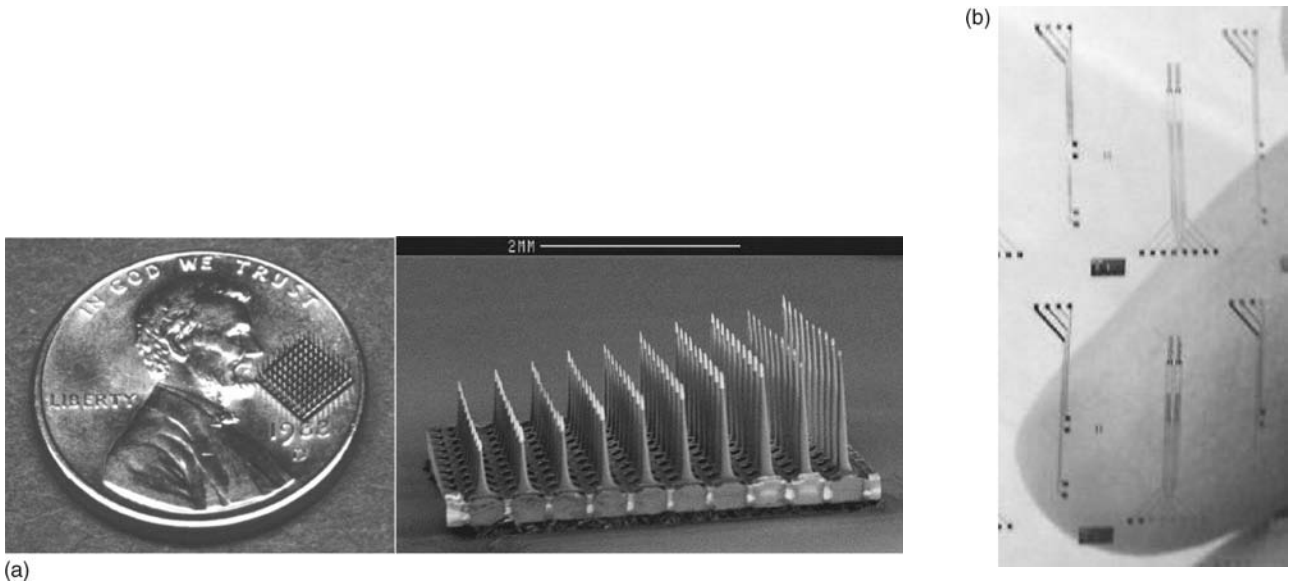


Figure 52. (a) Utah electrode array shown on a U.S. penny to convey size. (b) Modified Utah electrode array in which the length of the needles is uniformly graded (134).

Recent improvements in silicon microtechnology have made it possible to create not only planar microelectrodes but also penetrating brush electrode structures for *in vivo* measurements. In contrast to the University of Michigan's planar devices, the 2D and 3D cortical multimicroelectrode arrays developed at the University of Utah are fabricated out of a single solid block of silicon. Etching of the block results in a 10×10 array of needles, each 1.0–1.5 mm long, arranged on a 4.2×4.2 mm base. The metal and insulation layers are then applied, creating 35–75 μm long platinum recording tips (134,136,139) (Fig. 52a).

This design has the advantage of placing a relatively large number of recording sites in a compact volume of the cortex. However, with a single recording site on the end of each needle set at a fixed depth into the cortex, this version of the Utah array is classified as a 2D array as all the electrodes are in the same plane (140). The Utah probe can achieve high-density sampling by spacing many needles close together but does not have multiple sites along each shaft. When the length of the needles in such an array is graded (the array is said to be slanted, Fig. 52b) or the needles have some other distribution of lengths, these arrays are termed 3D as the electrode tips are no longer in the same plane. These designs are thought to give the better spatial selectivity (134,136).

Implanting such needle or brush electrode systems is obviously associated with damage of the tissue. Moreover, the stiffness of many systems may lead to damage of nervous tissue, especially if relative movement exists between the sharp needles and the delicate tissues. Breakage of the brittle needle is also a concern. Considerable efforts are therefore being directed at miniaturizing the width of the needles or at introducing more flexible materials.

For example, some versions of the Michigan Probe consist of four parallel, dagger-like probes connected to a micro-silicon ribbon cable. The ribbon cable is semiflexible and allows the probes to move up and down with the cortex as it pulses (139).

In the development of subretinal stimulating arrays using current silicon micromanufacturing techniques, it has been pointed out that a planar, rigid implant is likely to mechanically damage the compliant, spherical retina (137). Concerns have also been expressed regarding the use of penetrating microelectrodes, the relative micromotion between the array and the retina potentially provoking mechanical damage and a significant encapsulation response (134). The ideal retinal-stimulating electrode would therefore have the flexibility to match the curvature of the retina and the next generation of electrode arrays are likely to be constructed on flexible substrates.

Microelectrode arrays on flexible substrate have been demonstrated in a range of applications including the European project "Microcard", Si-Based Multifunctional Microsystem needle for Myocardial Ischemia Monitoring. Initially, work centered on silicon-based microprobes to monitor the electrical impedance of tissue, tissue temperature, pH, and local ionic concentrations of potassium, sodium, and calcium. These parameters were found to vary considerably when, for example, a heart undergoes an ischaemic phase, thus establishing the clinical value of the technique and device, (140).

In the course of the silicon probe development, it was foreseen that the brittle nature of silicon could make intact probe removal difficult. Additionally, the rigid needle could cause damage to the delicate tissues. The thrust of the project thus changed to the development of flexible, polymer-based probes.

Thin-film devices for the measurement of tissue impedance and ion concentrations were manufactured on flexible polyimide substrates (Fig. 53a) (141). Gold thin-film electrodes were deposited using an improved photolithography process for 1 μm resolution. Polyimide insulation

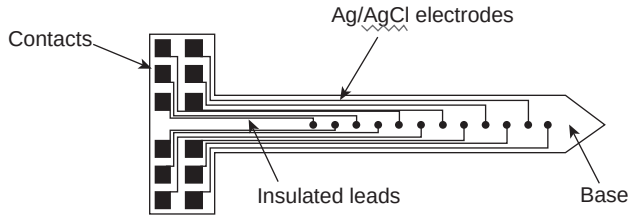


Figure 53. Microfabrication of sensors onto flexible substrates. 1D probe electrode array. [After Mastrototaro et al. (141).]

layers were spin-coated onto the PTFE surface after suitable conditioning, and they proved to be insulating and continuous.

Electrochemical characterization of the gold thin-film impedance electrodes showed them to possess high interface impedance. Pt and IrO oxide coatings were electrochemically applied to the gold thin-film surface and resulted in a drastic reduction in interface impedance for monitoring or stimulation applications (142).

Encircling neural electrodes may be of a cuff or spiral design. The term cuff electrodes applies to those devices that engulf the entire circumference of a nerve First model, which rather stiff, carried only one or two electrodes and they were made using a platinum foil electrodes that were located on the inside of a cylinder of silicone rubber, which was wrapped around a nerve (Fig. 54b). (136). It is generally recommended that the diameter of the cuff be 50% larger than the nerve diameter to avoid nerve compression and necrosis due to swelling and fibrous tissue ingrowth. Cuff electrodes do however have a long and successful track record in a range of FES applications (143).

The spiral electrode is a loose, open helix that is wound around the nerve (143). The open design can accommodate swelling and is very flexible. A version of this electrode is marketed by Cyberonics for use with their vagus nerve stimulator (Fig. 54b). New designs of nerve cuff electrodes seek to reshape the geometry of the nerve to more selectively stimulate or record from particular nerve fascicles. Efforts are also directed at controlling the electrical fields generated by the electrode arrays to better focus the stimulation (135).

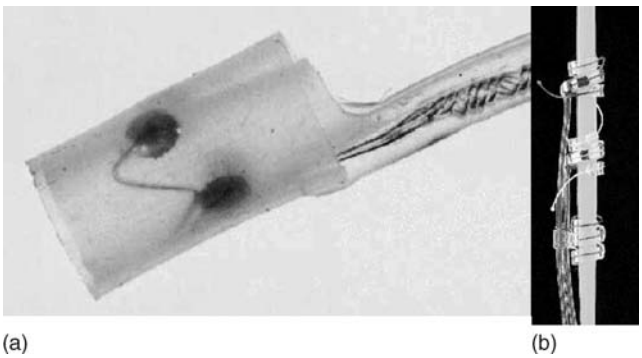


Figure 54. (a) Neural cuff electrode. (b) Spiral electrode (Cyberonics).

As part of a European project NEUROS, NIBEC developed a flexible thin-film-based stimulation and sensing cuff electrode for FES-related application. IrO, Pt, and Au electrodes were deposited onto a polyimide substrate. In order to facilitate implantation and ensure good contact between nerves, fascicles and electrode surface was self-curling. Polyimide resin with a thermal expansion coefficient differing from that of the polyimide substrate was chosen so that the curing process gives rise to a residual stress and curl in the device. The diameter of the electrode cylinder could be made less than 1 mm.

Diamond-like carbon (DLC) encapsulation was deposited onto the device using a plasma-enhanced chemical vapor deposition (PECVD) process. Adhesion to the polyimide substrate was found to be satisfactory following the addition of a silane adhesion layer at the interface (144,145).

With the aid of micofabrication techniques, one can control the area and properties of the electrodes and greatly decrease them in size. However, as electrode area decreases, the interface impedances increase with resultant difficulties in making accurate measurement. The key to success in this case is in the choice of electrode design, material, and electrode surface topography.

A similar concept to Chardack's differential current density pacemaker electrode was suggested for use in thin-film electrodes. The metal electrode is housed within a hollow chamber (Fig. 55). The chamber is filled with electrolyte and has a small aperture to enable electrical contact with tissues (146). As the metal–electrolyte interface is relatively large, the interface impedance is relatively small. The interface is also protected from mechanical disturbance (similar to the floating electrode) and, hence, should suffer from less artifact. As the small aperture determines the area of contact with the tissue, the effective stimulation or recording area is very small.

Other 3D designs with etched meshes should be assessed for their potentially larger interfacial areas.

Once again, surface roughness is an important factor in decreasing interface impedance and possibly in helping anchor the electrode in position. Rough-surfaced electrodes must be used with caution, depending on the application, in case the surface causes damage to the surrounding tissues. Certain materials and the electrode fabrication processes involved may well result in favorable macro-, micro-, and nano- surface features. Presently, investigators are studying modifications to the electrode surface using such things as nanotubes. Nanotechnology offers much promise for

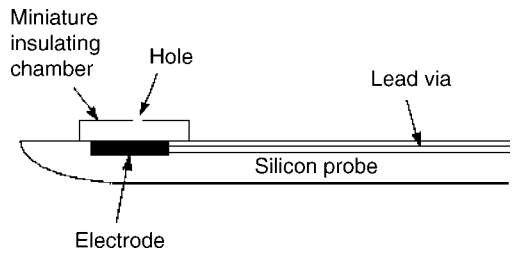


Figure 55. Thin-film differential current density electrode. [After Prohaska et al. (146).]

new sensor devices, particularly in the biomedical sector. Not only do individual nanotubes offer the possibility of using them as ultrafine needles for *in vivo* probing at the cellular level, but surfaces can be created with optimal distributions of clusters of nanotubes to maximize performance.

ELECTRODE STANDARDS

The Association for the Advancement of Medical Instrumentation (AAMI) produce a range of labeling, electrical, and other performance requirements for manufacturers and users to help ensure acceptable levels or product safety and efficacy. Some of the key electrode-related standards are briefly reviewed below.

Standards For Biosignal Monitoring Electrodes

Standards for Disposable ECG Electrodes. ANSI/AAMI EC 12 (2000)

Introduction. In an effort to minimize ECG recording problems associated with the performance of electrodes coupled to a standard ECG monitor or electrocardiograph, AAMI has proposed a series of simple bench tests designed to assess pregelled, disposable ECG electrodes.

Although originally conceived to assess disposable ECG electrodes, these standards are widely used to assess other biosignal monitoring electrodes, which is a consequence of the lack of other widely accepted standards for these monitoring applications and to the general applicability of the ECG standards to the other applications.

Although AAMI also lays down stipulations for electrode labeling, adhesion testing, and soon, only the electrical performance requirements are reviewed here.

AC Impedance. The average value of 10 Hz impedance for at least 12 electrode pairs connected gel-to-gel, at a level of impressed current not exceeding 100 μA peak-to-peak, shall not exceed 2 kΩ. None of the individual pair impedances shall exceed 3 kΩ.

Low impedance electrodes are desirable to avoid signal attenuation and distortion and to minimize 50/60 Hz interference pickup. High electrode impedances can also give rise to serious burns when the ECG electrodes are used in the presence of electrosurgery or defibrillator discharges. (147).

The impedance of the skin's outer layer, the stratum corneum, is many times larger than that of the metal/electrolyte interface, and hence, the former is of key concern when endeavoring to ensure good electrode performance. The skin preparation technique, the extent of diaphoresis, and the ability of the electrode gel in penetrating and reducing the skin impedance are generally more important than the electrode–electrolyte interface impedance.

The Standards Committee decided that the electrode gel-to-gel impedances should be significantly less than the expected impedance of clean, dry skin to ensure a minimal contribution by the electrode itself to the overall impedance (147). In the UBTL tests carried out on behalf of the Standards Committee, it was found that the mean 10 Hz impedance of a standard pair of ECG electrodes on unabraded skin was of the order of 100 kΩ. The AAMI committee chose 2000 kΩ as a reasonable limit for 10 Hz gel-to-gel impedance to ensure that the electrodes did not contribute significantly to the overall impedance nor to power dissipation in the presence of defibrillation overload and electrosurgery currents.

As the electrode-gel interface impedance is nonlinear and decreases with applied signal amplitude, the standard stipulates that the level of impressed current must not exceed 0.1 mA peak-to-peak when carrying out the test.

In the UBTL tests, it was found that the impedance as measured on abraded skin correlated well (99%) with the impedance measured with the electrodes connected gel-to-gel, whereas the impedance measured with electrodes applied to clean, dry skin correlates very poorly (47%) with the gel-to-gel measurements. Obviously, the bench test simply evaluates the ac impedance performance of the electrode-gel interface and will, therefore, not accurately predict or represent the clinical performance of an electrode on intact skin. For example, cases of electrodes that performed poorly as per the AAMI bench test exist, yet which proved very satisfactory *in vivo*. Conversely, some of the best electrodes according to the bench tests performed relatively badly in vivo (148).

DC Offset Voltage. After a 1 min stabilization period, a pair of electrodes connected gel-to-gel, shall not exhibit an offset voltage greater than 100 mV.

Ideally, the potentials of both electrodes used to monitor a biosignal should be identical and, thus, cancel each other out. Slight differences in the gels and metals used, however, result in an offset voltage. The potentials of the skin sites further complicate the recording, especially as these latter potentials (and their amplified difference) tend to be much larger. If the overall electrode-skin potential difference is larger than 300 mV, the amplifier may saturate and the biosignal will not be observed.

The UBTL report studied the correlation between gel-to-gel and electrode-skin offset voltages and found that gel-to-gel offsets were in the order of 2.5 times smaller than those recorded *in vivo* for the same electrodes on a patient's skin. As the maximum allowable *in vivo* dc offset should be less than 300 mV, the Committee decided that the limit for gel-to-gel dc offset should therefore be less than 300/2.5 mV (i.e., 100 mV).

Some reviewers of the standard argued that the limit should be reduced to 10 mV as this would help minimize motion artifact problems. The Committee rejected this suggestion, pointing out that no clear evidence exists that links high gel-to-gel offset voltages with motion artifact (largely caused by skin deformation).

Offset Instability and Internal Noise. After a 1 min. stabilization period, a pair of electrodes connected gel-to-gel, shall not generate a voltage greater than 150 $\mu V_{p\text{-}p}$ in the passband (first-order frequency response) of the 0.15–100 Hz for a period of 5 min following the stabilization period.

This standard is concerned with the problem of baseline wander, which introduces a low frequency component into the monitored biosignal making accurate diagnosis difficult. The American College of Cardiology's Task Force on the Quality of Electrocardiographic Records judged that drift rates less than 400 $\mu V \cdot s^{-1}$, although not highly rated, were not considered unacceptable.

The UBTL report detailed several experimental limitations that prohibited their detailed study of *in vivo* dc offset drift. Consequently, no correlational analysis was carried out between dc offset drift measurements made with electrodes applied to human skin and those joined gel-to-gel. They, however, decided to use the factor of 2.5 they had observed between clinical and bench test result for dc offsets, given that the measurement techniques are fundamentally similar. A limit of 150 $\mu V \cdot s^{-1}$ was therefore arrived at by dividing the 400 $\mu V \cdot s^{-1}$ baseline drift rating by a factor of 2.5. As the test circuit used in the bench test differentiates the offset voltage, the offset instability requirement is specified in μV rather than $\mu V \cdot s^{-1}$.

The Committee was contacted and asked to decrease the limit from 150 μV to 40 μV p-p in order to be in line with the AAMI standard "Cardiac monitors, heart rate meters and alarms (EC13)". The working group agreed that this requirement could be made more stringent but refused to decrease the limit to 40 $\mu V_{p\text{-}p}$. This requirement is under study and may well be altered.

This calculation involved in reaching the 150 $\mu V \cdot s^{-1}$ limit implies that skin potential fluctuations are only 2.5 times larger than those of the electrode–gel interface, which is most unlikely, and problems developing from drifting electrolyte/skin potentials will depend on skin preparation, electrode design, and electrode gel rather on than the electrode–gel interface characteristics per se (64).

Defibrillation Overload Recovery. Five seconds after each of four capacitor discharges, the absolute value of polarization potential of a pair of electrodes connected gel-to-gel shall not exceed 100 mV. Also during the 30 s interval following each polarization potential measurement, the rate of change of the residual polarization potential shall be no greater than $\pm 1\,mV \cdot s^{-1}$.

It is important that a clinician, having defibrillated a patient, be able to see a meaningful ECG within 5–10 s in order to judge the efficacy of the delivered impulse and to decide if another is required. The offset voltage across the electrode-skin interfaces, which drastically increased as a result of the defibrillation impulse, must therefore return to below 300 mV within 5 s following the discharge. Once again, using the 2.5 factor between bench test and *in vivo* potentials, this requirement translates to a gel-to-gel bench test offset voltage under 100 mV within 5 s of applying an overload of 2 mC (representing the worst possible situation encountered *in vivo* where the defibrillator paddles are placed in immediate contact with the ECG electrodes). Electrodes made of stainless steel, for example, tend to acquire offset voltages of several hundred mV for minutes and, consequently, no ECG trace is observable on the monitor (68).

Following the initial 5 s the EGC must not only be visible on the monitor but must also be recognizable and clinically useful. Hence, the stipulation that the offset voltage should not drift with time by more than $\pm 1\,mV \cdot s^{-1}$.

The UBTL results indicate good correlation exists between the results of this bench tests and animal tests, particularly at the higher recovery voltages encountered with non-Ag/AgCl electrodes.

Although only a very low percentage of ECG electrodes are, in fact, subjected to defibrillation impulses *in vivo*, the AAMI committee decided after some deliberation to insist that all ECG electrodes meet the proposed standard as it is impossible to guarantee that a given electrode would not be used in an emergency defibrillation situation.

Bias Current Tolerance. The observed dc voltage offset change across a pair of electrodes connected gel-to-gel shall not exceed 100 mV when the electrode pair is subjected to a continuous 200 nA dc current over the period recommended by the manufacturer for the clinical use of the electrodes. In no case shall this period be less than 8 h.

When a dc current passes through the metal-gel interface of an electrode, the electrode potential deviates from its equilibrium value and the electrode is said to be polarized. If the current is maintained indefinitely, the reactants become depleted causing the electrode potential to deviate further, possibly exceeding the limit allowable at the input of the ECG recording device.

Although most modern ECG recorders pass less than 10 nA of bias current through the electrodes, some older models can have bias currents as high as 1000 nA. A number of cardiac monitor manufacturers use dc bias currents to sense high electrode impedances to warn of disconnected leads or poorly affixed electrodes. The standard for cardiac monitors permits input bias currents of up to 200 nA. UBTL, therefore, adopted the 200 nA limit on the dc input bias current suggested for cardiac monitors for the tests. The ability of an electrode to cope with this value of bias current must therefore be demonstrated by not exceeding the AAMI dc offset requirement of 100 mV over the time period recommended by the manufacturer for the clinical use of the electrodes.

The 200 nA current level is generally well-tolerated by Ag/AgCl electrodes. Stainless-steel electrodes rapidly fail this test even at 10 nA with major increases in electrode potential.

Discussion. The AAMI standards bench tests are currently the only widely accepted electrode standard tests in use. The tests are simple and inexpensive to set up and have been widely embraced by manufacturers and users for production quality control purposes. One must bear in mind, however, that these tests evaluate only the electrode-gel interface and that they do not include the more important properties of the gel–skin interface. Assessment of the clinical performance of electrode impedance using the proposed bench tests is only relevant if the skin has been suitably abraded. Skin abrasion is not widely used by the clinical community and, hence, the relevance of at least some of the standard tests to the clinical situation is open to question.

Especially several decades ago, fulfillment of the AAMI requirements was commonly quoted as a guarantee

of the high *in vivo* electrical performance of an electrode. An electrode with, for example, a dc offset of 1 mV was widely believed by customers to be a much better electrode than one with an offset of 5 mV. This naivety appears to be on the wane, however, and manufacturers and customers are shifting toward low cost electrodes that score less highly in the AAMI tests but are good enough for a given application.

The author once supplied a leading company with dry metal-loaded polymer electrodes. The company connected the electrodes together and tested them as per the AAMI standards (for pregelled electrodes). Perfect electrical performances were measured given that what was effectively being assessed was metal-to-metal contact. Direct current offsets of 0 mV were obtain. Once the dry electrodes were applied to a patient's skin, a less than favorable result was obtained.

The attitude to adopt, therefore, when interpreting AAMI standard bench tests results for pregelled, disposable ECG electrodes is that electrodes that meet the AAMI standards have a tendency rather than a certainty to perform well *in vivo*. Electrodes that perform better as per the bench tests do not necessarily perform better *in vivo*. They are a useful set of tests nonetheless.

The ANSI/AAMI standard tests were conceived such that the test apparatus needed can be readily assembled by an electrode manufacturer. However, one can can buy a convenient-to-use, custom-built electrode tester (as per AAMI standards) called the Xtratek electrode tester ET65A (Direct Design Corporation, Lenexa, Kansas.) (Fig. 56).

Electrocardiograph surface electrode testers also exist for the *in vivo* testing of the quality of (*1*) the design ECG electrodes, (*2*) the application of the electrodes, and (*3*) the skin preparation technique used.

The electrode tester generally measures the ac impedance and dc offset of the electrode-patient system. These measurements can be used, for example, in stress testing to decide if the skin sites have been sufficiently well prepared (i.e., contact impedances are low enough) to proceed with the clinical procedure. They can also be used to detect the presence of loose cables or bad contacts.

Figure 56. Early version of the Xtratek electrode tester ET65A. (Direct Design Corporation; Lenexa, Kansas.)

Standards for Stimulation Electrodes

Although not covered in this article, the following standards exist that stipulate minimum labeling, safety, and performance requirements for the given stimulators. The rationale for the standards is also presented.

- Transcutaneous electrical nerve stimulators ANSI/AAMI NS4.
- Implantable spinal cord stimulators ANSI/AAMI NS14.
- Implantable peripheral nerve stimulators ANSI/AAMI NS15.

Standards for Automatic External Defibrillators and Remote-Control Defibrillators. ANSI/AAMI DF 80 (2003)

AC Small Signal Impedance. The 10 Hz impedance for any of at least 12 electrode pairs connected gel-to-gel, at a level of impressed current not exceeding 100 μA peak-to-peak, shall not exceed 3 kΩ. The impedance at 30 kHz shall be less than 5 Ω. The rationale for this requirement is based on the performance criteria in ANSI/AAMI EC 12 for disposable ECG electrodes. Interestingly, the permissible gel-to-gel 10 Hz impedance for large-area defibrillation pads is higher than that allowed for small-area ECG electrodes. The gel-to-gel impedance measured at 30 kHz will be largely that of the gel pads as the interface impedances at this frequency will be almost zero.

AC Large Signal Impedance. The impedance of an electrode pair connected gel-to-gel, in series with a 50 Ω load and measured at the maximum rated energy of the defibrillator shall not exceed 3 Ω. A value of 50 Ω is thought to represent the typical (rather low) *in vivo* transthoracic impedance between the electrodes. One wants the delivered energy to be dissipated in the patient's chest and not in the electrodes where the wasted energy may give rise to skin burns. The above requirement is therefore thought to provide a reasonable limit on the impedance contributed to the overall impedance by the electrode pair during defibrillation (<6%).

Combined Offset Instability and Internal Noise. A pair of electrodes connected gel-to-gel shall generate, after a 1 min stabilization period, a voltage no greater than 100 μV peak-to-peak in the pass band of 0.5–40 Hz, for a period of 5 min following the stabilization period. The rationale for this requirement is based on the performance criteria in ANSI/AAMI EC 12 for Disposable ECG electrodes. The frequency range used is more limited in recognition that the cardiac monitor bandwidth is more appropriate in this application.

Defibrillation Recovery. The potential of a pair of gel-to-gel electrodes in series with a 50 Ω resistor and subjected to three shocks at 360 J or maximum energy at 1 min intervals shall not exceed 400 mV at 4 s and 300 mV at 60 s after the last shock delivery. The rationale for this requirement is largely based on the performance criteria in ANSI/AAMI EC 12 for Disposable ECG electrodes. An actual

defibrillation impulse is applied instead of that from a simulation circuit. The offset voltage across the simulated electrode-patient load must return to below 400 mV within 4 s following the discharge (slightly different values, 300 mV and 5 s, are used in ANSI/AAMI EC 12). As the patient's chest is represented by the 50 Ω resistor, no need exists for the 2.5 factor used in ANSI/AAMI EC 12 to correlate bench test and *in vivo* results.

DC Offset Voltage. A pair of electrodes connected gel-to-gel shall, after a 1 min stabilization period, exhibit an offset voltage no greater than 100 mV. The rationale for this requirement is based on the performance criteria in ANSI/AAMI EC 12 for disposable ECG electrodes.

Universal-Function Electrodes. With conventional defibrillators, it has been customary to use separate pregelled ECG electrodes for monitoring and defibrillator paddle electrodes for defibrillation. The monitoring electrodes are not capable of effectively delivering a defibrillation shock, and the paddle electrodes have only limited monitoring capability. For recent applications, particularly automatic external defibrillation, it is very desirable to use self-adhesive pregelled disposable combination electrodes that perform well in the dual monitoring and defibrillation functions. These electrodes may also be used for delivery of transcutaneous pacing. Hence, combination electrodes may become preferred for defibrillation, and it is appropriate in a standard for defibrillators to consider their use and to outline a few requirements for them.

If the electrodes are designed and intended for use in multiple modes (i.e., monitoring, defibrillation, and pacing) the electrode shall meet all (of the above) requirements after 60 min of pacing at the maximum current output and maximum pacing rate through a pair of gel-to-gel electrodes in series with a 50 Ω resistor.

No general performance standards exist for combination pacing/defibrillation/monitoring electrodes, the (above) requirements define the basic minimum controls necessary to ensure safe and reliable operation.

Standards for Electrosurgical Devices. ANSI/AAMI HF 18 (2001)

Introduction. Although AAMI lays down stipulations for the testing of a range of parameters, only the key electrical performance requirements for the dispersive electrodes are reviewed below.

Maximum Safe Temperature Rise. The maximum patient tissue temperature rise shall not exceed 6 °C when the dispersive electrode carries a current of 700 mA under the test conditions below, unless the device is labeled in accordance with 4.1.4.2 (i.e., for use on infants). For devices labeled for use on infants, the maximum patient tissue temperature rise shall not exceed 6 °C when the dispersive electrode carries a current of 500 mA under the test conditions stipulated in the standard. In monopolar electrosurgical procedures, the dispersive electrode must be able to reliably conduct the required surgical current without generating a significant rise in skin temperature. It is widely accepted that the maximum safe skin temperature for short-term and long-term exposure is 45 °C, as normal resting skin temperature varies between 29° and 33 °C. Electrodes must not generate skin temperature increases approaching 12 °C. A 6 °C increase in temperature is therefore thought to represent an acceptable upper limit.

The temperature measurement method must have an overall accuracy of better than 0.5 °C and a spatial resolution of at least one sample per square centimeter of the electrode thermal pattern. The thermal pattern must include the area extending 1 cm beyond the geometry of the electrode under test. This degree of special resolution is stipulated as electrosurgical burns may be confined to very small areas and these must be detected. As current tends to flow to the edge of the electrode and spread out further in the skin, the test requires that the surround area of skin is also scanned.

The electrode under test is to carry a current from an electrosurgical generator of 700 mA_{rms} for 60 s, unless the device is labeled in accordance with 4.1.4.2, in which case the test current may be 500 mA. A current of 700 mA applied for 60 s yields a heating factor of 30 A^2 s. [Heating Factor $= I^2t$ (A^2s).] This value is far in excess of the maximum likely current and duration for a TUR (transurethral resection) procedure. A more realistic heating factor is less than 10 and, hence, the stipulated testing procedure is very conservative.

These tests must be conducted on human volunteers or on a suitably structured surrogate medium. When human volunteers are used, the tester must include a variety of body types in the sample group rather than concentrate on a single body type (thin, average, or thick layers of subcutaneous body fat). If surrogate media are used, the tester must demonstrate that the media are electrically and thermally similar to human volunteers. Human volunteer subjects are the reference standard. Current density distribution under an electrode depends on a wide range of factors, including the electrical properties of the skin and underlying tissues, hence, the need to test a given electrode on a wide range of individuals. The use of a surrogate material, even pig skin, which is commonly used, will not necessarily replicate with sufficient accuracy the clinical performance of the electrode. If a surrogate medium is used, the tester must demonstrate the equivalence of the test medium to human tissue. It is the Committee's view that no adequate surrogate medium has yet been suggested or used that has all of the properties of human tissue for the purpose of determining electrode performance.

Nessler et al. (149) point out that the above experiments are laborious, time-consuming, and expensive to perform. They have developed a new test device, swaroTEST, which includes a surrogate electronic skin, which, they claim, simulates the relevant electrical features of human skin and thus can replace the required volunteer experiments (Fig. 57). The device consists of a 3D resistor network representing the electric features of the skin and muscle tissue, and a temperature-sensing array (one transistor for each cm^2) to measure the resultant temperature increase after a standardized current load (700 mA hf current during 60 s, proposed in the relevant AAMI HF-18 standard). The authors claim that a comparison of results obtained with their device and those with thermo camera images of

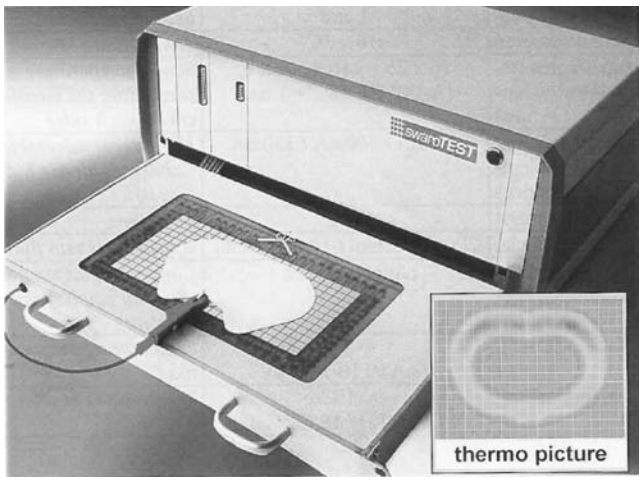


Figure 57. Swaro Test device with a measuring board electronic skin to simulate the electrical properties of human skin. It is hoped that this device will replace the human volunteer experiments required for ANSI.AAMI HF 18.

volunteer experiments correspond sufficiently well to justify the acceptance of their test device as a surrogate medium.

Electrode Contact Impedance. The electrode contact impedance must be low enough that the dispersive electrode represents the preferred current pathway, thus avoiding skin burns at alternative pathways. For conductive electrodes, the maximum electrode contact impedance shall not exceed 75 Ω over the frequency range of 200 kHz–5 MHz when measured as described on a human subject. The frequency range of 200kHz–5 MHz encompasses the frequency ranges of existing generators. As electrode-tissue impedance increases as applied current decreases, the committee decided on an impedance measuring current of 200 mA as it represents the lower limit of average currents reported for TUR procedures. Under these conditions, a maximum contact impedance value of 75 Ω was judged an acceptable for the conductive electrodes.

For capacitive electrodes, the minimum capacitance shall be no less than 4 nF (0.004 μF) when measured as described. In this case, electrode contact impedance is measured by placing the capacitively coupled dispersive electrode under test on a rigid metal plate larger than the electrode contact area. The test current and frequencies are the same as those specified for conductive electrodes. Their impedance characteristics are described in terms of capacitance as their impedances vary as the inverse of the frequency. The majority of capacitive electrodes that have been found to be clinically acceptable typically have a capacitance value of 4 nF, hence the minimum acceptable capacitance value specified by the Committee.

SUMMARY

With external biosignal monitoring electrodes, difficult challenges exist in the exciting new area of personalized health. Such electrodes must form part of the patient's (or health-conscious citizen's) clothing and must continue to work, day after day, wash after wash, without gelling or preparation of any kind, without suffering from motion artifacts, and without causing skin irritation, which is no mean achievement.

An old monitoring problem still remains to be adequately conquered. A convenient and rapid method of applying many high performance electrodes to the head of a patient for EEG measurement awaits invention. The problem (and that of ECG ambulatory monitoring above) can be side-stepped to some extent by finding new electrode positions (montages or leads) that avoid the most problematic skin sites, hairy head in EEG and muscle and flabby areas in ECG.

For external stimulation, exciting new areas include public access defibrillation. The electrodes and their application to the victim must be almost literally fool-proof, given the seriousness of the possible consequences for all concerned. The electrodes must work after having been stored in the most inhospitable locations and possibly under extreme temperature fluctuations, for example, in the trunk of a car in the desert. In the more mainstream areas of cardiac pacing and defibrillation and electrosurgery, the optimal distribution of current density under the electrodes remains a goal still to be achieved. The solution to this problem offers the hope of decreased electrode areas and the design of truly multifunction pads.

The integration of electrodes into garments for FES and body toning is a relatively new area with considerable possibilities.

At present, implant electrodes and associated technologies already offer amazing potential for the deaf, lame, and even the blind. The development of multimicroelectrode arrays and waveforms that can help optimally shape the electrical fields to facilitate more effective and natural stimulation is a thrilling prospect. The interface properties of the microelectrodes will require further research so as to offset the potentially high interface impedances. Ideas already exploited in cardiac pacing, for example, may prove rewarding when adapted for these areas.

It is hard to overstate the potential of research being undertaken in the area of brain–machine interface. We live in exciting times.

BIBLIOGRAPHY

Cited References

1. Gatzke RD. In: Miller HA, Harrison DC, editors. The electrode: A measurement systems viewpoint. Biomedical Electrode Technology. New York: Academic Press; 1974.
2. Janz GJ, Ives DJG. Silver-silver chloride electrodes. Ann NY Acad Sci 1968;148:210–221.
3. Webster JG. Medical Instrumentation: Application and Design. 3rd ed. New York: John Wiley & Sons; 1998.
4. Cheney M, Isaacson D, Newell JC. Electrical impedance tomography. SIAM Rev 1999;41:85–101.
5. McAdams ET, Jossinet J, Lackermeier A, Risacher F. Factors affecting the electrode-gel-skin interface impedance in electrical impedance tomography. Med Biol Eng Comput 1996;34(6):397–408.

6. Singh S, Singh J. Transdermal drug delivery by passive diffusion and iontophoresis: A review. Med Res Rev 1993;13(5):569–621.
7. Bard AJ, Faulkner LR. Electrochemical Methods. New York: John Wiley & Sons; 1980.
8. Almasi JJ, Hart MW, Schmitt OH, Watanabe Y.Bioelectrode voltage offset time profiles and their impact on ECG measurement standards. Can Med Biol Eng Conf 4th, Winnipeg, Manitoba, Canada, 1972.
9. McAdams ET. Effect of surface topography on the electrode-electrolyte interface impedance, Part 1: The high frequency, small signal interface impedance. Surface Topogr 1989;2: 107–122.
10. Fricke H. The theory of electrolytic polarization. Philos Mag 1932;7:310–318.
11. Cole KS, Curtis HJ. Transverse electric impedance of squid giant axon. J Gen Physiol 1938;22:3764.
12. Sluyters-Rechbach M, Sluyters JH. Sine wave methods in the study of electrode processes. In: Bard AJ, editor. Electroanalytical Chemistry, vol. 4. New York: Marcel Dekker; 1970. pp 1–128.
13. De Levie R. The influence of surface roughness of solid electrodes on electrochemical measurements. Electrochim Acta 1965;10:113–130.
14. de Levie R. On the impedance of electrodes with rough interfaces. J Electroanal Chem 1989;261:1–9.
15. Maritan A, Toigo F. On skewed arc plots of impedance of electrodes with an irreversible electrode process. Electrochim. Acta 1990;35:141–145.
16. McAdams ET. Effect of surface topography on the electrode-electrolyte interface impedance, Part 2: The low frequency ($F<1$ Hz), small signal interface impedance. Surface Topogr 1989;2:223–232.
17. Bergveld P. Med Biol Eng Comput 1976;14:479–482.
18. Brummer SB, Robblee LS, Hambrecht FT. Criteria for selecting electrodes for electrical stimulation: Theoretical and practical considerations. Ann NY Acad Sci USA 1983;405: 159–171.
19. Brummer SB, Turner MJ. Electrical stimulation of the nervous system: The principle of safe charge injection with noble metal electrodes. Bioelectrochem Bioenerget 1975; 2:13–25.
20. Dymond AM. IEEE Trans BME 1976;23:274–280.
21. Lilly JC, Hughes JR, Alvord EC, Galkin TW. Brief, noninjurious electric waveform for stimulation of the brain. Science 1955;121:468–469.
22. Weinman J. Biphasic stimulation and electrical properties of metal electrodes. J Appl Physiol 1965;20:787–790.
23. Fischler H, Schwan HP. Polarisation impedance of pacemaker electrodes: *In vitro* simulating practical operation. Med Biol Eng Comput 1981;19:579–588.
24. Schwan HP. Electrode polarization impedance and measurements in biological materials. Ann NY Acad Sci USA 1968;148:191–209.
25. Schwan HP. Alternating current electrode polarisation. Biophysik 1966;3:181–201.
26. Simpson RW, Berberian JG, Schwan HP. Nonlinear AC and DC polarization of platinum electrodes. IEEE Trans Biomed Eng 1980;27:166–171.
27. Jaron D, Briller SA, Schwan HP, Geselowitz DB. Nonlinearity of cardiac pacemaker electrodes. IEEE Trans Biomed Eng 1969;16:132–138.
28. Onaral B, Schwan HP. Linear and non-linear properties of platinum electrode polarization: Part 1.Frequency dependence at very low frequencies. Med Biol Eng Comput 1982;20:299–306.
29. McAdams ET, Henry P, Anderson JMcC, Jossinet J. Optimal electrolytic chloriding of silver ink electrodes for use in electrical impedance tomography. Clin Phys Physiol Meas 1992;13(Suppl 1):19–23.
30. McAdams ET, Jossinet J. The importance of electrode-skin impedance in high resolution electrocardiography. Automedica 1991;13:187–208.
31. McAdams ET, Jossinet J. A Physical Interpretation of Schwan's limit voltage of linearity. Med Biol Eng Comput 1994; March: 126–130.
32. McAdams ET. Jossinet J. DC nonlinearity of the solid electrode-electrolyte interface impedance. Inn Tech Biol Med 1991;12:329–343.
33. McAdams ET, Jossinet J. A physical interpretation of Schwan's limit current of linearity. Ann Biomed Eng 1992; 20:307–319.
34. K Stokes. Cardiac pacing electrodes. Proc IEEE 1996;84(3): 457–467.
35. Stokes K. Implantable pacing lead technology. IEEE Eng Med Biol 1990;9(2):43–49.
36. Williams DF. The Williams Dictionary of Biomaterials. Liverpool University Press; 1999.
37. Geddes LA. Electrodes and the Measurement of Bioelectric Events. New York: John Wiley & Sons; 1972.
38. Crenner F, Angel F, Ringwald C. Ag/AgCl electrode assembly for thin smooth muscle electromyography. Med Biol Eng Comput 1989;27:346–356.
39. Kingma YJ, Lenhart J, Bowes KL, Chambers MM, Durdle NG. Improved Ag/AgCl pressure electrodes. Med Biol Eng Comput 1983;21:351–357.
40. Geddes LA, Baker LE, Moore AG. Optimum electrolytic chloriding of silver electrodes. Med Biol Eng 1969;7:49–56.
41. Heath R. Tin-stannous chloride electrode element. U.S. Patent 4,852,585, 1989.
42. Mannheimer JS. Lampe GN. Clinical Transcutaneous Electrical Nerve Stimulation. Philadelphia, F.A. Davis, PA; 1987.
43. Prausnitz MR, Bose VG, Langer R, Weaver JC. Electroporation of mammalian skin: A mechanism to enhance transdermal drug delivery. Proc Natl Acad Sci USA 1993;90:10504–20508.
44. Brown L, Langer R. Transdermal derlivery of drugs. Ann Rev Med 1988;39:221–229.
45. Rosendal T. Further studies on the conducting properties of human skin to direct and alternating current. Acta Physiol Scand 1945;8:183–202.
46. Rosendal T. Concluding studies on the conducting properties of human skin to alternating current. Acta Physiol Scan 1945;9:39–49.
47. Salter DC. A study of some electrical properties of normal and pathological skin in vivo. Ph.D. dissertation, University of Oxford. Oxford (UK): 1980.
48. Klingman AM. Skin permeability: Dermatologic aspects of transdermal drug delivery. Am Heart J 1984;108(1):200–207.
49. Reilly JP. Electrical Stimulation and Electropathology. Cambridge, UK: Cambridge University Press; 1992.
50. Chien YW. Transdermal controlled-release drug administration.In: Swarbrick J. editor Novel Drug Delivery Systems. New York: Marcel Dekker Inc.; 1982. p 149.
51. Edelberg R. Electrical properties of the skin. In: Elden HR, editor. A Treatise of the Skin. New York: John Wiley & Sons; 1971.
52. Yamamoto Y, Yamamoto T. Dispersion and correlation of the parameters for skin impedance. Med Biol Eng Comput 1978;16:592–594.

53. Chien YW. Development of transdermal drug delivery systems. Drug Develop Industr Pharm 1987;13(4&5):589–651.
54. Rothman S. Electrical behavior. In:Physiology and Biochemistry of the Skin. Chicago (IL): The University of Chicago Press; 1956. p 9–25.
55. Lawler JC, Davis MJ, Griffith EC. Electrical characteristics of the skin. J Invest Dermatol 1960; 301–308.
56. Rosell J, Colominas J, Riu P, Pallas-Areny R, Webster JG. Skin impedance from 1 Hz to 1 MHz. IEEE Trans Biomed Eng 1980;35:649–651.
57. Almasi JJ, Schmitt OH. Systemic and random variations of ECG electrode system impedance. Ann N Y Acad Sci 1970;170:509–519.
58. Grimnes S. Dielectric breakdown of human skin in vivo. Med Biol Eng Comput 1983;21:379–381.
59. Schmitt OH, Almasi JJ. Electrode impedance and voltage offset as they affect efficacy and accuracy of VCG and ECG measurements. Proc. XIth International Vectorcardiography Symposium, New York, 1970; 245–253.
60. Yamamoto T, Yamamoto Y. Analysis for the change of skin impedance. Med Biol Eng Comput 1977;15:219–227.
61. Searle A, Kirkup L. A direct comparison of wet, dry and insulating bioelectric recording electrodes. Physiol Meas 2000;21:271–283.
62. McAdams ET, Lackermeier A, Woolfson ET, Moss GP, McCafferty DF, In vivo ac impedance monitoring of percutaneous drug delivery. Proc. 9th Int. Conf. on BioImpedance, Heidelberg, Germany, 1995: 344–347.
63. De Talhouet H, Webster JG. The origin of skin-stretch-caused motion artefacts under electrodes. Physiol Meas 17:81–93.
64. Tam HW, Webster JG. Minimizing motion artifact by skin abrasion. IEEE Trans Biomed Eng 1977; BME 24:134–140.
65. Zinc R, Distortion and interference in the measurement of electrical signals from the skin (ECG, EMG, EEG). Innovation and Technology in Biology and Medicine, 12, special issue. 1991; 1: 46–59.
66. McLaughlin J, McAdams ET, Anderson JMcC. Novel dry electrode ECG sensor system. 16th Annual Int Conf IEEE Eng Med Biol Soc Baltimore (MD), Nov. 1994:804.
67. Jossinet J, McAdams ET. Skin Impedance. Innovation and technology in biology and medicine, 12, special issue. 1991;1:21–31.
68. Carim HM. Bioelectrodes. In: Webster JG. editor Encyclopaedia of Medical Devices and Instrumentation. New York: Wiley & Sons; 1988. p 195–226.
69. Oh SY, Leung L, Bommannan D, Guy RH, Potts RO. Effect of current, ionic strength and temperature on the electrical properties of skin. J Controlled Release 1993;27:115–125.
70. Olson WH, Schmincke DR, Henley BL. Time and frequency dependence of disposable ECG electrode-skin impedance. Med Instrum 1979;13:269–272.
71. McAdams ET. Surface biomedical electrode technology. Int Med Device Diagnost Ind 1990; 44–48.
72. McAdams ET, McLaughlin JA, Anderson J McC. Multi-electrode systems for electrical impedance tomography. Physiol Meas 1994;15:A101–A106.
73. McAdams ET, McLaughlin J, Brown BN, McArdle F. In: London HD, editor. The NIBEC EIT harness, Clinical and Physiological Applications of Electrical Impedance Tomography. Chapt 8, UCL Press; 1993. p 85–92.
74. McAdams ET, Jossinet J. Hydrogel electrodes in bio-signal recording. Proceedings of the 12th Annual International Conference of the IEEE, Philadelphia, PA: Engineering in Medicine and Biology Society; 1990: 1490–1491.
75. McAdams ET, Lackermeier A, Jossinet J. AC impedance of the hydrogel-skin interface. 16th Annual Int. Conf IEEE Eng in Med and Biol Soc Baltimore (MD), 1994: 870–871.
76. Carim HM, Hawkinson RW. EKG electrode electrolyte-skin AC impedance studies. Proc. 4th Ann Conf IEEE Eng Med Biol Soc 1982:503–504.
77. Yamamoto T, Yamamoto Y. Electrical properties of the epidermal stratum corneum. Med Biol Eng 1976;14:151–158.
78. Geddes LA. A. Historical perspectives 2: The electrocardiograph. In: Bronzino JD, editor. The Biomedical Engineering Handbook. Boca Raton FL: CRC Press; 1995; p 788–798.
79. Waller AD. A demonstration on man of electromotive changes accompanying the heart's beat. J Physiol 1887; 8:229–234.
80. Waller AD. On the electromotive changes connected with the beat of the mammalian heart, and of the human heart in particular. Phil Trans R Soc London Ser B 1989;180:169–194.
81. Waller AD. Introductory address on the electromotive properties of the human heart. Brit Med J 1888;2:751–754.
82. Barker LF:. Electrocardiography and phonocardiography: A collective review. Bull Johns Hopkins Hosp 1910;21:358–359.
83. Rowbottom ME, Susskind C. In: Electricity and Medicine: History of their Interaction. San Francisco (CA): San Francisco Press; 1984.
84. Barron SL. The development of the electrocardiograph in Great Britain. Br Med J 1950;1:720–725.
85. Lewes D. Multipoint electrocardiography without skin preparation. Lancet 1965;2:17–18.
86. Wolferth CC, Wood FC. The electrocardiographic diagnosis of coronary occlusion by the use of chest leads. Am J Med Sci 1932;183:30–35.
87. Barnes AR, et al. Standardization of precordial leads. Am Heart J 1938;15:235–239.
88. Burch GE, DePasquale NP: A History of Electrocardiography with a New Introduction by Joel D Howell, 2nd ed. San Francisco, CA: Jeremy Norman; 1990.
89. Ungerleider HE. A new precordial electrode. Am Heart J 1939;18:94.
90. Welch W. Self-retaining electrocardiographic electrode. JAMA 1951;147:1042.
91. Jasper HH, Carmichael L. Electrical potentials from the intact human. Science 1935;81:51–53.
92. Khan A, Greatbatch W. Physiologic electrodes. In: Ray CD. editor. Medical Engineering. Chicago, IL: Year Book Medical Publishers; 1974.
93. Manley AG, Medical electrode. US patent 3,977,392, 1976.
94. K Krug, Marecki NM. Porous and other medical and pressure sensitive adhesives. Adhes Age 1983;26(12):19–23.
95. Hymes AC. Monitoring and stimulating electrode. U.S. Patent 4,274,420, June 23, 1981.
96. Dempsey GJ, McAdams ET, McLaughlin J, Anderson JMcC. NIBEC cardiac mapping harness. 14th Annual Int. Conf. IEEE Eng. In Med and Biol Soc Paris, France, Nov 1992: 2702–2703.
97. Lymberis A. Research and Development of Smart Wearable Health Applications: The Challenge Ahead, Wearable eHealth Systems for Personalised Health Management, Studies in Health Technology and Informatics 108. Lymberis A, de Rossi D, editors, IOS Press; 2004.
98. Axisa F, Schmitt PM, Gehin C, Delhomme G, McAdams E, Dittmar A. Flexible technologies and smart clothing for citizen medicine, home healthcare and disease prevention. IEEE Trans Inform Technol Biomed 2005;9(3): 325–336.
99. Adams G. An Essay on Electricity. London; 1785.

100. Aldini G. Account of Late Improvements in Galvanism. London; 1803.
101. Duchenne GBA. In: De l'Électrisation Localisée et de son Application á la Physiologie, á la Pathologie et à la Thérapeutique. 1855.
102. Duchenne GBA. In: Baillière JB et al., editors. Mécanisme de la Physionomie Humaine. 1876.
103. Schechter DC. In: Exploring the Origins of Electrical Cardiac Stimulation. Medtronic; 1983.
104. Robinson AJ, Snyder-Mackler L. Clinical Electrophysiology: Electrotherapy, Electrophysiologic Testing. Baltimore (MD): Williams and Wilkins; 1995.
105. Low J, Reed A, Electrotherapy Explained: Principles and Practice. Oxford: Butterworth–Heinmann Ltd; 1994.
106. Stankevich BA. 4% of professional liability claims involve electromedicine equipment. Mod Health Care 1980;10(12): 74–76.
107. Pearce JA. The thermal performance of electrosurgical dispersive electrodes. Ph.D. dissertation. Purdue University, West Lafayette (IN); 1980.
108. Wiley JD, Webster JG. Analysis and control of the current distribution under circular dispersive electrodes. IEEE Trans Biomed Eng 1982;29:381–385.
109. Caruso PM, Pearce JA, DeWitt DP. Temperature and current density distributions at electrosurgical dispersive electrode sites. Proc 7th N Engl Bioeng Conf., Troy, New York, March 22–23, 1979: 373–376.
110. V Krasteva, Papazov S. Estimation of current density distribution under electrodes for external defibrillation. BioMedical Engineering Online, 2002; 1:7. Available http://www.biomedical-engineering-online.com/content/1/1/7.
111. Y Kim, Schimpf PH. Electrical behavior of defibrillation and pacing electrodes. Proc IEEE 1996;84(3):446–456.
112. Kim Y, Fahy JB, Tupper B. Optimal electrode designs for electrosurgery, defibrillation, and external cardiac pacing IEEE Trans Biomed Eng 1986;33:845–853.
113. Netherly SG, Carim HM. Biomedical electrode with lossy dielectric properties. US pat 5,836,942, 1998.
114. Ferrari RK. X-ray transmissive transcutaneous stimulating electrode. US pat 5,571,165, 1996.
115. McAdams ET, Andrews P. Biomedical electrodes and biomedical electrodes for electrostimulation. US pat 2003, 134,545, 2003.
116. Szeto AYJ. Pain relief from transcutaneous electrical nerve stimulation (TENS). In: Webster JG. ed. Encyclopedia of Medical Devices and Instrumentation. New York: John Wiley & Sons; 1988. p 2203–2220.
117. AXELGAARD J. Reverse current controlling electrode. US pat 2004,158,305, 2004.
118. Sarlandière. "Mémoires sur l'électropuncture considérée comme moyen nouveau de traiter efficacement la goutte, les rhumatismes et les affections nerveuses. Paris, 1825.
119. Hyman AS. Resuscitation of the stopped heart by intracardial therapy. Arch Intern Med 1932;50:283.
120. Mittal T. Pacemakers – A journey through the years. Ind J Thorac Cardiovasc Surg 2005;21:236–249.
121. Myers GH, Parsonnet V. Pacemaker electrodes In: Myers GH, Engineering in the Heart and Blood. New York: Wiley-Interscience; 1969.
122. Greatbatch W, Holmes CF. History of implantable devices. IEEE Eng Med Biol 1991; Sept: 36–49.
123. Chardack WM, Gage AA, Greatbatch W. Correction of complete heart block by a self-contained and subcutaneously implanted pacemaker. J Thorac Cardiovasc Surg 1961;42: 418.
124. Lagergren H, Johansson L. Intracardiac stimulation for complete heart block. Acta Chir Scand 1963;125:562–566.
125. Parsonnet V, Zucker IR, Asa MM.Preliminary investigation of the development of a permanent implantable pacemaker utilizing an intracardiac dipolar electrode. Clin Res 1962; 10:391.
126. Timmis G. The electrobiology and engineering of pacemaker leads. In: Saksena S, Goldschlager N. eds. Electrical Therapy for Cardiac Arrhythmias. New York: Saunders W.B. Co.; 1990.
127. Admundson DC, McArthur W, Mosharrafa M. The porous endocardial electrode. PACE 1979;2:40–50.
128. Lagergren H, Edhag O, Wahlberg I. A low threshold nondislocating endocardial electrode. J Thorac Cardiovasc Surg 1976;72:259.
129. Lewin G, Myers GH, Parsonnet V, Zucker IR. A non-polarizing electrode for physiological stimulation. Trans Am Soc Artific Intern Organs 1967;13:345.
130. Ellenbogen KA, Wood MA. Cardiac Pacing and ICDs. New York: Blackwell Science Inc; 2002.
131. Mond H, Stokes KB. The electrode-tissue interface: The revolutionary role of steroid elution. PACE 1991;15:95–107.
132. Mond HG, Stokes KB. The steroid-eluting electrode: A 10-year experience. Pacing Clin Electrophysiol 1996 Jul; 19(7):1016–1020.
133. Lenarz T, Battmer R-D, Goldring JE, Neuburger J, Kuzma J, Reuter G. New electrode concepts (Modiolus–Hugging Electrodes). Adv Otorhinolaryngol Basel Karger 2000;57:347–353.
134. Maynard EM. Visual prostheses. Annu Rev Biomed Eng 2001;3:145–168.
135. Peckham PH, Knutson JS. Functional electrical stimulation for neuromuscular applications. Annu Rev Biomed Eng 2005;7:327–360.
136. Rutten WLC. Selective electrical interfaces with the nervous system. Annu Rev Biomed Eng 2002;4:407–452.
137. Weiland JD, Liu W, Humayun MS. Retinal prosthesis. Annu Rev Biomed Eng 2005;7:361–401.
138. Drake KL, Wise KD, Farraye J, Anderson DJ, BeMent SL. Performance of planar multisite microprodes in recording extracellular single-unit intracortical activity. IEEE Trans Biomed Eng 1988;35:719–732.
139. Schwartz AB. Cortical neural prosthetics. Annu Rev Neurosci 2004;27:487–507.
140. Aguiló J. Microprobe multisensor for graft viability monitoring during organ preservation and transplantation. 2nd Annual International IEEE-EMB Special Topic Conference on Microtechnologies in Medicine & Biology, Madison, WI. February 2002; 15–20.
141. Mastrototaro JJ, Massoud HZ, Pilkington TC, Ideker RE. Rigid and flexible thin-film multielectrode assays for transmural cardiac recording. IEEE Trans Biomed Eng 1992; 39:271–279.
142. Linquette-Mailley SC, Hyland M, Mailley P, McLaughlin J, McAdams ET. Electrochemical and structural characterisation of electrodeposited iridium oxide thin film electrodes applied to neurostimulating electrical signal. Mater Sci Eng 2002;21:167–175.
143. Mortimer JT, Bhadra N. Peripheral nerve and muscle stimulation. In: Horch KW, Dhillon GS, editors. Neuroprosthetics: Theory and Practice (Series on Bioengineering & Biomedical Engineering), vol. 2. New York: World Scientific; 2004. p 638–744.
144. Hyland M, McLaughlin J, Zhou DM, McAdams E. Surface modification of thin film gold electrodes for improved in vivo performance. Analyst 1996;121:705–709.
145. Rieger R, Taylor J, Comi E, Donaldson N, Russold M, Mahony CMO, McLaughlin JA, McAdams E, Demosthenous A, Jarvis JC. Experimental determination of compound A-P

direction and propagation velocity from multi-electrode nerve cuffs. Med Biol Eng Comput Phys 2004;26:531–534.
146. Prohaska OJ, Olcaytug P, Pfundner P, Dragaun H. Thin film multiple electrode probes: Possibilities and limitations. IEEE Trans Biomed Eng 1986;33:223–229.
147. Schoenberg AG, Klingler DR, Baker CD, Worth NP, Booth HE, Lyon PC. Final report:Development of test methods for disposable ECG Electrodes. UBTL Technical Report No. 1605–005, Salt Lake City (UT); 1979.
148. Hollander JI, ECG-Electrodes.Report No. 83.336, MFI-TNO, Utrecht, The Netherlands, 1983.
149. Nessler N, Reischer W, Salchner M. Electronic skin replaces volunteer experiments. Measure Sci Rev 2003;3(2):71–74.

Reading List

Bell GH, Knox AC, Small AJ. Electrocardiography electrolytes. Br Heart J 1939;1:229–236.
Geddes LA, Baker LE. Principles of Applied Biomedical Instrumentation, 3rd edition. New York: John Wiley & Sons; 1989.
Licht S. History of electrotherapy. In: Licht S. ed. Therapeutic Electricity and Ultraviolet Radiation. New Haven, CT: Elizabeth Licht Pub; 1959. p 1–69.

See also DEFIBRILLATORS; ELECTROCARDIOGRAPHY, COMPUTERS IN; ELECTROENCEPHALOGRAPHY; ELECTROSURGICAL UNIT (ESU); FUNCTIONAL ELECTRICAL STIMULATION; TRANSCUTANEOUS ELECTRICAL NERVE STIMULATION (TENS).

BIOFEEDBACK

JOHN C. ARENA
VA Medical Center and Medical Collage of Georgia
TRISHUL DEVINENI
Conemaugh Health System
EDWARD J. MCGOWAN
E.J. McGowan & Associates

INTRODUCTION

Biofeedback is a term that first arose in the 1960s for a methodology that uses instrumentation to record the physiological responses of organisms and then in real time give information about those physiological responses back to the organism. It is presumed that by getting such timely feedback about physiological responding, the organism will learn, through a trial and error basis, how to control the desired physiological response.

The most concise definition of biofeedback is probably that of Olton and Noonberg (1), who characterized it as, "any technique that increases the ability of a person to control voluntarily physiological activities by providing information about those activities" (p. 4). In practice, the process of clinical biofeedback training involves the use of a machine (usually a computer-based system in contemporary applications), which allows a therapist to monitor the patient's bodily responses (most commonly surface muscle tension or surface skin temperature). Information concerning the patient's physiological responses are then relayed back to the patient, generally either through an auditory modality (a tone that goes higher or lower depending on, say, electrical activity of the target muscles increasing or decreasing) and/or a visual modality (now usually a computer screen where, e.g., surface skin temperature is sampled and then graphed on a second by second basis in real time). Through this physiological feedback, it is anticipated that the patient will learn how to control his/her bodily responses through mental means.

Biofeedback arose as an application of the learning theories of B.F. Skinner, Hull, Thorndike, Dollard and Miller, and John Watson. In particular, Neal Miller postulated that the established principles of learning that had so far been applied to overt behaviors could validly be applied to behaviors that were covert and presumed not under voluntary control.

Classical conditioning is also referred to as Pavlovian Conditioning after the seminal work of Russian scientists Pavlov and Sechenov in the early twentieth century. Classical conditioning is a laboratory learning paradigm by which a neutral stimulus (conditioned stimulus; CS) comes to elicit a new response (conditioned response; CR) by repeated pairing in close temporal proximity with another stimulus (unconditioned stimulus; UCS) that already elicits that response (unconditioned response; UCR). In subsequent presentations of the CS, the organism will then emit the UCR without pairing of the UCS. For example, the UCS might be food and the UCR is salivation, the CS, the ringing of a bell, is presented immediately prior in temporal pairing with the UCS, food. After repeated pairing of the ringing bell with food, the organism will come to salivate in response to the bell's ringing. The behaviors conditioned in this paradigm are typically unlearned, such as most physiological responses utilized in biofeedback practice. However, the learning paradigm most often appealed to as the theoretical underpinning of the field of biofeedback is not classical conditioning. Rather, biofeedback is generally considered a form of operant conditioning. This learning theory postulates that the consequence of a response changes the likelihood that the organism will produce that response again. The essential assumption of operant conditioning is that behavior is lawful and follows the rules of cause and effect and probability.

A basic supposition of operant conditioning is, if you wish a behavior to continue, you reinforce or reward that behavior. If you wish a behavior to decrease, or to stop completely, you do not reinforce that behavior. Thus, within the theoretical framework of operant conditioning, the main way that you strengthen a behavior is to follow it in close temporal proximity with a reward. The definition of a reinforcer is any stimulus change that occurs after a response and tends to increase the likelihood that a response will be repeated. It is important that the reinforcer follow the desired behavior quickly such that the delay in the presentation of the reward is kept to an optimally short delay. As the delay in the reward increases, the effectiveness of the reinforcer is generally decreased. There are many examples of positive reinforcement in our everyday life—receiving a bonus for outstanding work, receiving an A in a course for intensive studying, scoring a touchdown in a football game and the crowd's adulation.

In addition to positive reinforcement, there are three other possible consequences to behavior in operant

conditioning: (a) Negative reinforcement involves the removal of a consequence to a response that results in reduced likelihood that the behavior will be repeated in the future. (b) Positive punishment involves adding a consequence when a response is performed that serves to decrease the likelihood of the response occurring in the future. Examples are plentiful: child misbehaves, parent scolds child, child less likely to misbehave; drive over the speed limit, get ticket, less likely to speed; do poorly at work, get demoted, less likely to perform poorly on the job. Negative punishment involves the removal of a consequence to a behavior that serves to reduce the likelihood of that behavior occurring in the future. An examples is a parent playing with their child who is clearly enjoying the playtime; the child starts to yell loudly, the parent stops playing with the child, parent less likely to engage in positive play with the child. In clinical biofeedback applications, these other types of reinforcement contingencies are rarely used, with biofeedback clinicians preferring to use positive rewards to influence behavior.

Perhaps the most important principle in operant conditioning that directly involves clinical biofeedback training is that of shaping. Shaping is the learning process by which the predefined target response is achieved through gradual and systematic reinforcement. The training begins with a simple, existing response and basic criteria for reinforcement, with gradually more stringent criteria applied for reinforcement in order to achieve more complex and reliable responses. After the initial behavior is reliably performed, reinforcement is given contingent on the performance of more complex or difficult responses. This pattern of increasingly stringent contingent reinforcement continues until the final target behavior is achieved. Shaping can be assisted by modeling the skill to be learned before shifting the reinforcement schedule. In clinical biofeedback, the target behavior in shaping is often tailored to the individual learning style and abilities of the patient. For example, a patient undergoing EMG biofeedback training may be initially reinforced for detecting gross changes in electrical activity in a particular muscle group. Following the initial success, the reinforcement is tapered and made contingent on the patient being able to detect ever more subtle changes in muscular tension in the target region. This may continue until the patient is unable to further demonstrate more refined skill in detecting muscle tension.

There are other principles within operant conditioning that also apply to biofeedback practice. One of these involves discrimination training, in which the organism demonstrates the ability to differentiate between at least two stimulus conditions by emitting a different response to each stimuli. In clinical biofeedback, the concept of discrimination applies to the patients' ability to distinguish relatively subtle differences in physiological states. Another important principle, indeed the mortar that lays the foundation of clinical biofeedback training, is the concept of generalization: If a response is conditioned to one stimulus, the organism may also respond to a similar stimulus (generalization), but not to a dissimilar stimulus (discrimination). Discrimination learning is a goal in the early stages of biofeedback training, while generalization is a longer range goal. Clearly, the overarching aim of clinical biofeedback training is to take the learned process applied in the office setting and have that learning process come to apply to the everyday ''real world'' setting.

In addition to basic operant conditioning principles, there are general principles in basic psychophysiology that are important in clinical biofeedback training. The law of initial values states that the autonomic nervous system response to stimulation is a function of the prestimulus level (2). The higher the level of the response measure prior to a stressful stimulus being presented, the smaller the increase in response to the stressor, which is often referred to as a ceiling effect. Conversely, the higher the level of the measure prior to a relaxing stimulus being presented, the larger the decrease in response to the relaxing stimulus. When prestimulus response values are low prior to the presentation of a relaxing stimulus, this will lower the magnitude of the response and is often referred to as a floor effect. While the law has been shown to generally hold for measures of respiration and cardiovascular activity (such as heart rate and the vasomotor response), measures such as salivation and electrodermal response have not been found to be influenced by prestimulus values.

Homeostasis refers to the tendency of any organism to strive to maintain a state of equilibrium or rest. Homeostasis is believed to be maintained by a negative feedback loop, which is a theorized set of bodily mechanisms that provide information. This information directs the organism's physiological systems to decrease activity if levels of functioning are higher than normal, or to increase activity if levels are diminished relative to normal. Thus, all organisms strive to return to prestimulus levels of physiological arousal when presented with any stimulus.

Theories Underlying Clinical Biofeedback Training

There are two general theories underlying the use of biofeedback for most chronic benign medical disorders, such as anxiety, headache, musculoskeletal pain, and incontinence (3). The first is a direct psychophysiological theory, which attributes the etiology and/or maintenance of the disorder to specific physiological pathology. This biofeedback training modulates in a therapeutic direction. For example, it has traditionally been assumed that tension headache is caused by sustained contraction of skeletal muscles in the forehead, neck, and shoulder regions. Through the use of biofeedback, the patient learns to decrease muscle tension levels, leading to a decrease in headache activity. The second theory is predominantly psychological and postulates that there is a relationship between situational stress and the disorder in question. Through the use of biofeedback, the patient learns to regulate physiological responses such as muscle tension levels or sympathetic nervous system activity. This regulation leads to a decrease in overall stress levels, which brings about symptomatic relief. This amelioration of symptoms brought about by the learning of voluntary control of specific peripheral responses is postulated to be underpinned by central changes that occur along pain-relay and sympathetic pathways. For the case of headache treatment, this means that both muscle

relaxation and hand warming may indirectly dampen central brain mechanisms involved in the onset of headache. It is not necessary to view these theories as competing; they may be more appropriately viewed as complementary. Most clinicians subscribe to both theories, depending on the patient's presenting problem, clinical findings, and medical history.

Biofeedback Modalities and Instrumentation

Biofeedback instruments are first and foremost psychophysiological measuring instruments, to which has been added the capability to display the value of the measured parameter(s) in a form understandable to the subject. Feedback can be visual, auditory, or tactile.

The physiological measures generally employed in clinical biofeedback training are surface electromyographic activity (EMG) and skin surface temperature. Less often used physiological responses include measures of neuronal activity using electroencephalography (EEG), and measures of electrodermal response (skin resistance, skin conductance), cardiovascular activity (simple heart rate, heart rate variability, blood pressure, and vasomotor activity), and respiration (generally, respiration rate and depth).

Biofeedback instruments can be used to gain insight into a subject even if biofeedback therapy is not the goal. When assessment is the goal, care must be taken that the results are not contaminated by unintended biofeedback, such as the subject viewing the display. Subject spatial position with respect to windows, doors, and the professional must also be considered to avoid influencing the results. Environmental control of ambient temperature, humidity, and drafts is strongly recommended to minimize effects of these stressors on the subject.

Biofeedback instruments fall into three categories: research, clinical, and trainers. Research instruments are often configured from very flexible laboratory modules and their use is beyond the scope of this article. Many modern clinical instruments, however, are precise enough for basic or clinical research.

Clinical instruments are generally accurate, calibrated, and reliable. They are available as either stand-alone discrete units measuring a single parameter, or as computer-based multimodality systems. Some are comprehensive and accurate enough to be used for research. Trainers are single modality instruments intended to be purchased by or loaned/rented to the subject. They are less accurate and expensive than clinical instruments. Modern technology has made most trainers accurate and reliable despite their relatively low cost.

To protect both the subject and the professional, it is recommended that only FDA listed equipment be used. Even third-party software should meet this recommendation unless it is to be used only for educational purposes.

The biofeedback professional should have the proper academic credentials for the applications offered. Certification in use of the specific instrument modalities and applications is also recommended. The Biofeedback Certification Institute of America (BCIA), an affiliate of the Association for Applied Psychophysiology and Biofeedback (AAPB), offers certification in many areas. The American Physical Therapy Association (APTA) offers certification in EMG treatment of urinary incontinence. It is further recommended that the biofeedback professional receive training for the specific instruments used as there are some differences among instruments of the same type.

Most modern instruments are manufactured using standards traceable to the National Bureau of Standards (NBS). Modern electronic technology permits instruments to remain calibrated throughout their life, with the exception of sensors, which must be replaced occasionally. Instrument performance while attached to a subject, even the same subject at different times, can vary widely from improperly placed sensors, electromagnetic interference (EMI) and the subjects' condition. Even the professional is not a psychophysiological constant. Therefore it is strongly recommended that the using professional invest in test instruments and fixtures for each modality to be used so they can independently establish that the instrument(s) are operating correctly. The test device could be as simple as a laboratory thermometer used in a stirred water bath to verify temperature or a precision resistor to verify electrodermal accuracy. Biopotentials (EMG, EEG, ECG) require an electrode meter, and electrical safety, a volt Ω-ampere meter. Some commercial ECG signal simulators also provide sine and square-wave outputs, useful for testing EMG, EEG, and ECG instruments.

Circuitry and software of commercial instruments varies as to amplifier bandwidth, filter cutoffs, signal rectifiers, and integrators, which make it difficult to measure accurately direct comparisons between instruments of the same type from different manufacturers. Comparing results from different instruments of the same model from one manufacturer depends on the technology used. Older instruments with discrete component filters vary more than modern instruments due to component tolerances. Modern computer-based instruments employing software filters are more similar.

Professionals making comparisons with other professionals should use relative values such as percent decrease in finger temperature. A rough rule of thumb follows: Instruments vary about $\pm$ 10% across models and manufacturers, but subjects can vary as much as an order of magnitude in some measures. Lesson: Rely on known standard inputs to evaluate instrument performance.

Safety falls into three major areas: environmental, biological, and electrical. Environmental safety concerns trip hazards, sharp edges–corners, heaters, lamps and machinery in the subject–professional area. Biological safety concerns disinfecting re-useable sensors, subject chair, area, and so on. Disposable biopotential (EMG,EEG,ECG) electrodes are recommended. Until disposable EEG electrodes become available, they and all reusables should be cleaned and disinfected between uses. Electrical safety concerns both subject and professional. Battery powered instruments are intrinsically safe but if connected to a alternating current (ac), line operated device (i.e., computer, recorder, oscilloscope), through a nonisolated interface, they become a potential hazard.

Most computer-based instruments are isolated to stringent standards and provide complete electrical safety for

the connected subject. By isolating the computer system with a medical grade power transformer, the professional is also protected. Use the (recommended) ac current meter to verify isolation.

EMG Biofeedback Instrumentation

Biofeedback applications of EMG range from simple relaxation, using electrodes placed on the forehead, to complex neuromuscular retraining of stroke (cardiovascular accident, or CVA) victims utilizing four EMG channels on each of the affected and unaffected sides to retrain functional movements using both inhibition and reinforcement learning techniques.

The EMG signals for clinical biofeedback are the summation of muscle cell action potentials generated by the underlying muscles seen at the skin. They are acquired from surface electrodes applied over (or in the vicinity of) muscle(s) to be monitored. Signals from the electrodes are amplified and conditioned by high performance differential amplifiers. Signal characteristics of interest are in the range of 0.1–2000 μV amplitude, over a bandwidth of ~25–500 Hz. The amplified signal is then processed to a form suitable for display to the subject.

EMG Electrodes. Muscle signals are acquired using two (active) electrodes located along the muscle fiber axis. A third electrode (common, often erroneously called ground) establishes the instrument common at the subject common potential. Well-placed surface electrodes provide an adequate EMG signal to enable subjects to learn control and change in the desired direction (i.e., relaxation, EMG lower; reeducation, EMG higher).

Needle or wire subcutaneous electrodes can collect a more comprehensive representation of the EMG (motor and nerve cellular action potentials), but their use is limited to Neurology or research, due to the complexity and invasive nature of the procedures.

A surface electrode is a complex electrochemical network. The manufacturing process reduces and stabilizes all internal parameters, leaving only the electrolyte–subject skin interface for the clinician to cope with. Fortunately, EMG amplifier technology has considerably reduced requirements for electrode preparation (for most applications) to a good skin cleaning using alcohol or a commercial prep solution. A surface electrode basically consists of a contact resistance (i.e., 10–100 kΩ), paralleled by a capacitance (i.e., 1 nF), in series with a half-cell potential (i.e., 300 mV "battery").

The care required in electrode–lead–amplifier placement and mechanical stabilization depends on procedure dynamics and electrical environment considerations. For relatively static applications (i.e., relaxing in a chair), in a relatively benign electrical environment, much less preparation and virtually no stabilization are required. For dynamic (i.e., treadmill) applications in a hostile electrical environment (i.e., central urban), considerable care and expertise are required to obtain reliable signals. Proper electrode choice and stabilization means (i.e., taping) are required. Electrode mechanical stabilization is required in dynamic applications to minimize disturbing the half-cell potentials that form at the electrolyte–skin interface.

Other considerations for electrode selection and placement include surface curvature, movement, and sweating.

Today, most EMG biofeedback clinicians use disposable electrodes in consideration of disease and litigation problems. One newer type of surface electrode (hydrogel) can be moved between (properly prepared) sites on one patient during one session, but using any type of electrodes between subjects is not recommended.

EMG Amplifiers. The EMG amplifiers have benefited considerably from integrated circuit technology and today achieve performance inconceivable two decades ago. Many products still connect to electrodes with a cable or leads, but some amplifiers are small enough to mount directly on the electrode structure. While the amplifier characteristics are exceptional, the performance of the EMG channel can be compromised by asymmetries in electrodes and connecting leads, resulting in unequal electrical induction, causing artifact to appear as a differential error signal. The following are considered to be minimal specifications for a modern EMG amplifier:

- Low internal noise (<0.5 μV, p–p).
- High input impedance (Zin >100 MΩ).
- Flat bandwidth and sharp high and low frequency cutoffs (>18 dB/octave).
- High common mode rejection ratio (CMRR $> 10^7$).
- Common mode input range (CMR $> \pm 200$ mV).
- Static electricity shock protection (>2000 V).
- Gain stability (all causes) $> \pm 1\%$.

Bandwidths (3 dB) in common use for clinical biofeedback are 25–500 Hz, primarily used for neuromuscular reeducation training, and 100–200 Hz used for relaxation training and where bandwidth must be limited because of EMI considerations. A hardware or software switch provides bandwidth selection. The narrow bandwidth loses some of the EMG frequency spectrum, but is adequate for many applications, having the advantages of lower in-band noise and less artifact susceptibility. Some manufacturers believe that a single bandwidth of ~30–300 Hz captures most of the EMG signal and provides simplicity.

EMG Biofeedback Application Example. Consider the problem of acquiring a 1 μV signal from a muscle using two (active) surface electrodes and a reference (common) electrode. Characteristics of the reference electrode are not as critical as the actives because the high CMRR of the amplifier minimizes disturbances in its impedance and half-cell potential.

The two active signal acquiring electrodes are effectively back to back, in series with the (muscle) signal source impedance and generator. Impedance of the active electrodes is probably close, since they were applied in a like manner, and the very high input impedance of the amplifier makes differences negligible.

Active electrode half-cell potentials are a different matter. While they are similar, they are of such a great magnitude (i.e., 300 mV) compared to the signal (i.e., 1 μV) that small abrupt changes in either half-cell potential will be coupled through the amplifier as a large artifact, which is the reason why it is so important to mechanically stabilize surface electrodes. Small disturbances of electrodes and leads outside of the frequency band of interest are not seen.

Leads connecting electrodes to the amplifier must also be stabilized to reduce artifact. They should be twisted and taped down for dynamic applications. Triode (equilateral triangle group) electrodes allow placing the miniaturized amplifier directly on the electrode(s), reducing lead length to zero. Amplifier and lead mass must also be considered in dynamic applications.

In some applications, such as treadmill and (internal) pelvic floor applications, some artifact must be tolerated and either average or quiescent levels used for training.

Following successful acquisition and amplification of the EMG signal it is usually rectified in a precision operational circuit or subroutine to produce the time-variant average that is needed for quantification and display. Older instruments provide integral average and modern instruments provide root-mean-square (rms) average that is loosely held as an analog of power by some authors. A related instrument response parameter is the time constant (TC) of integration that follows rectification.

Neuromuscular reeducation applications require a short (i.e., 50 ms) TC to sense the slightest voluntary (phasic) response and relaxation training as long as one second, to provide a more slowly changing display promoting relaxation, when tonic levels are of more interest.

Therapists having detailed knowledge of the frequency spectral characteristics of the specific muscle(s) being trained view (on an oscilloscope) the raw (filtered but not rectified ac) signal for assessment and training effectiveness. Raw EMG is seldom used as feedback due to its' visual complexity.

Stand-alone instruments use either analogue operational or microprocessor implemented filtering, rectification, averaging, and display. Most modern instruments are microcomputer based. The biofeedback instrument manufacturer provides signal acquisition and conditioning hardware to be interfaced, and software to be installed in a PC. Most recent biofeedback software has been for use on Windows operating system PCs.

Some stand-alone instruments have data storage, statistical reporting, and downloading capabilities. Microcomputer-based systems provide very sophisticated data reporting capabilities, with raw data exporting for advanced analysis. Some systems feature general purpose software for more advanced users and application specific software for less sophisticated users performing repetitive procedures, such as pelvic floor muscle strengthening and synchrony training for urinary incontinence.

Temperature Biofeedback Instrumentation

Temperature biofeedback is primarily used to improve poor peripheral blood flow caused by chronic sympathetic arousal. Other medical conditions causing low peripheral blood flow must be ruled out prior to attempting biofeedback. Temperature is then taken as a partial integration of blood flow and thus of peripheral vasoconstriction caused by sympathetic arousal.

Training criteria are determined by the therapist for the particular subject, but digit temperature training goals range from 31.1 to 35 °C for most subjects.

There are several types of low cost "instruments" primarily used for group education. These include glass-alcohol thermometers, liquid-crystal (i.e., mood-dots and biotic bands), and digital thermometers. This section concerns itself with clinical grade instruments employing sensors, amplifiers, and displays.

Temperature Sensors. Small sensor(s) are affixed to finger(s) or toe(s) with porous surgical tape at nearly zero tension so as not to occlude blood flow, or provide a heat reservoir. The sensor(s) are in good thermal contact with, but electrically isolated from, the subject. Accuracy and linearity of modern commercial sensors are excellent, but it is essential to use sensors having a short thermal time-constant and minimal hysteresis so minute changes in temperature (i.e., blood flow) are immediately communicated to the subject.

Modern sensors are thermistors, compensated negative coefficient resistors that provide a linear negative change in resistance over the temperature range of 21–37.8 °C. A small direct current (dc) voltage is impressed across the thermistor resulting in increased current flow for increases in heat (due to increased blood flow) from the subject.

Temperature Amplifiers. Whether the temperature instrument is stand-alone or one modality of a multimodality system, the circuit consists of a compensated thermistor sensor, powered by a constant voltage, and a current-to-voltage amplifier that produces a temperature proportional voltage that is then displayed and/or digitized for further processing.

Amplifier characteristics depend on the thermistor resistance range, the excitation voltage, and the required output voltage. Bandwidths of 0–5 Hz provides adequate response and minimizes noise, improving resolution. Since signals are high level, a single-ended current amplifier can be used.

Absolute accuracy is important because both subjects and therapists tend to discuss readings in their respective groups for comparative purposes. Subjects particularly are very sensitive to numbers and an instrument differences could impede training progress.

Most instruments specify ± 0.56 °C absolute accuracy. It is often better than claimed and can be checked in a stirred waterbath against a standard lab mercury thermometer.

Resolution of most instruments is selectable for either 0.056 or 0.0056 °C resolution. Subjects tend to make large improvements in initial states of training and the coarser (0.056 °C) resolution is adequate. There is some question by these authors about whether the 0.0056 °C resolution has clinical utility.

Temperature Displays. A time-based line graph is the most common display, as it shows past history and current trend. Bar graphs, digital meters, and computer animation displays are also used. These should be selected by the therapist–subject team to be relevant to the subject's beliefs; and sometimes varied to keep motivation high.

Temperature Biofeedback Application. Sensor attachment is often the back of the finger or toe of interest using a breathable tape to prevent heat build-up. Tape tension should be just enough to affix the sensor without constricting blood flow. Lead length should be sufficient to allow the subject to achieve a comfortable position without lead tension.

Room temperature should be in the comfort range for the subject (i.e., 21.1–23.9 °C) and without any breeze. Relative humidity should be between 30 and 50% to avoid humidity stress or evaporation cooling.

A location free of transient noises (i.e., office activity, elevator, emergency vehicles) will facilitate subject focus and relaxation.

EEG Biofeedback Instrumentation

The EEG biofeedback instrumentation is similar to EMG biofeedback instrumentation in acquiring the biopotential signal. The reader is encouraged to review the section on EMG biofeedback instrumentation for a background on the following section.

Whereas EMG instrumentation acquires surface muscle signals at the skin, EEG instruments acquire signals on the scalp generated by brain cellular activity below. The EMG activity in this region is considered artifact and care must be taken in EEG signal processing to minimize EMG contamination. When large EMG artifacts cannot be removed from the EEG signal, feedback must be blocked or held constant until the EMG artifact has passed.

The EEG signal is acquired on the scalp using surface electrodes, generally different from EMG electrodes, in either a differential (bipolar) or referential (monopolar) mode. High performance differential amplifiers increase and condition the signals. Signals of interest are in the range of 0.5–100 V (p–p), over the frequency range of 1–30 Hz. Newer clinical research extends the frequency range to 1–50 Hz.

Modern EEG biofeedback instruments perform digital filtering on the amplified and conditioned EEG signal to obtain the frequency band(s) or full spectrum of interest (i.e., fast fourier transform, FFT).

Viewing the full frequency spectrum EEG signal is of interest in assessing brain state(s) and is sometimes used as feedback, displayed as bilateral (left and right hemisphere) displays back to back about as vertical centerline.

More often the feedback display is less complicated, consisting of vertical bars or lights representing several spectral bands logically combined to reward increases in the band of interest (i.e., beta, 16–20 Hz) and decreases in the band to be reduced (i.e., theta, 4–9 Hz). The EMG activity is also monitored and acts to inhibit or freeze the feedback if present.

This logic is also used to drive a computer generated graphic animation sequence (i.e., games) and/or audio feedback comprising several types of files and songs (i.e., wav, midi). Auditory feedback is useful for closed eyes training and can also be used to provide simultaneous reward-inhibit band information when the visual presentation does not provide it, as in the case of animation sequence visual feedback.

EEG Electrodes. The EEG electrodes are placed on the scalp in a standardized grid called the 10–20 system. Coordinates are determined with respect to the midline, between the nasion and the inion, and a line between the right and left ears, in percent of the distance from the reference axis.

Bipolar recordings are made between (2) active scalp electrodes, with a neutral site serving as amplifier common (i.e., center forehead). An active pair of electrodes is used for each additional site to be monitored. Only one common is required unless the amplifiers have isolated commons, in which case they must be tied to the subject, usually at the same neutral site.

Monopolar recordings are made with a single scalp electrode (+) with respect to a chosen reference site (−). Amplifier common is placed at a third (neutral) site. For bilateral monopolar recording, the left channel has a scalp electrode (+) somewhere on the left side of mid-line with its reference (−) on the left ear. Amplifier common could be either the left mastoid bone or forehead.

The right channel placement somewhat mirrors the left with the active scalp electrode (+) on the right side of midline, the active reference (−) on the right ear, and the common on the right mastoid bone or forehead. The left and right active electrodes are placed according to training objectives, not necessarily symmetrical.

Multichannel systems are monopolar having a scalp electrode for each of the (20) sites in the 10–20 system, all having the same (forehead) reference and instrument common.

An Electro-Cap having 20 scalp and one reference–common electrode, sized to the subject, is worn. Electrode sites are prepared through holes in each electrode. A cable harness connects the cap to a junction-box for connecting amplifiers to desired sites, or directly to the instrument, for the 20 channel system. Since all sites are measured referentially (i.e., monopolar), differential comparison between any sites can then be accomplished in software or circuitry.

It is important that both active electrodes (i.e., +, −) have the same electrode material and electrolyte, since the initial stage of the amplifier is dc coupled, and a large difference in half-cell potentials could saturate (block) the amplifier. The (2) active electrodes must be prepared similarly even if one is a scalp (cup or disk) placement and the other an ear-clip.

Gold, silver–silver chloride, and tin are common active electrode materials. The common electrode can be a different type, usually a disposable silver–silver chloride EMG electrode.

Electrode sites are prepared by mild abrading and infusing a conductive "prep" gel to stabilize and improve conductance of the scalp. This is followed by application of

an thixo-tropic electrolyte (i.e., 10–20 paste). The cup or disk electrode is then pressed on the paste until it is firmly sealed. Viscous force holds the scalp electrode in place during recording. Hair is managed with tape or cottonballs, which also helps retain and stabilize the placement. Ear clip electrodes are spring retained.

Reusable cup scalp and ear clip electrodes are in common use despite potential health and litigation problems. Several disposable systems have been or are being developed, but are not in wide use at this writing.

Careful electrode preparation will result in an impedance of <10 kΩ, usually measured using a 20 Hz ac impedance meter. The actual value required depends on the electrical environment of the treatment area. The EEG electrodes are generally unshielded, so electric induction is reduced with low impedance placements.

Electrode half-cell potentials must also be checked differentially with an electrode meter. Differences of more than ± 25 mV between the active electrodes usually indicates unstable placements or different materials used. If a different type of common electrode was used, a large half-cell potential difference between each active and the common is not problematical, as long as they are similar and stable.

EEG Amplifiers. Like EMG amplifiers, EEG amplifiers make good use of integrated circuit and surface-mount technologies to achieve performance never before seen. While amplifier characteristics are exceptional, the effective performance of the EEG channel can be compromised by asymmetries in electrodes, cables, and leads. The resulting imbalance between the plus (+) and minus (−) inputs to the differential amplifier can cause unequal electric induction producing an error that shows as increased noise or offset. Fortunately, most unwanted electric fields are outside the amplifier bandwidth. Still, good practice in electrode preparation, lead routing, and electrical environment control are necessary to avoid a setup that is electrostaticly "hot".

The following parameters are considered minimum for the modern EEG differential amplifier:

- Low internal voltage and current noise(<1 μV, 100 pA, p–p)
- High input impedance (Zin $>10^8$).
- Bandwidth (1–50 Hz).
- Frequency cutoffs (>18 dB/octave).
- High common mode rejection ratio ($>10^7$).
- Common mode input range (greater than ± 200 mV).
- Static electricity shock protection (>2000 V).
- Gain stability (all causes) greater than $\pm 1\%$.

The amplified EEG signal is then digitally filtered to produce the desired bands of frequencies to be used for assessment, training, or mapping.

EEG Displays and Feedback. The modern EEG instrument utilizes a computer to perform the signal processing, auditory and visual display generation, data collection–reduction, and reporting necessary for effective assessment and training. Today's sophisticated programs require a fairly high performance computer to perform all these tasks in apparent real time. Most manufacturers design around readily available Windows based personal computers having the following minimum characteristics:

Processor:	Pentium 3
Speed:	600 MHz
Memory	256 MB
Storage	10 GB
Display resolution	800×600
Video memory	32 MB ($4 \times$ AGP)
Operating system	Windows '98, or later
Auditory system	Sound Blaster Live
Speakers	Good quality with sub-woofer

Recent advances in clinical biofeedback have shown subjects can comprehend complex visual and multiparameter auditory feedback simultaneously. A higher performance computer with more working and video memory, and the very best auditory system is suggested for these applications.

A very useful feature of Windows'98 and later operating systems is the support of multiple display monitors. This makes it possible to stretch the display onto a second monitor. The therapist can construct displays having highly technical items on their monitor while the subject sees nontechnical items, such as animation.

A common use of the digitally filtered EEG signal is for treatment of Attention Deficit Disorder. Other applications using other bands of frequencies include hyperactivity and performance enhancement.

Full frequency spectrum displays are used to "quiet" a "noisy" brain under the guidance of a skilled therapist. Animation sequence displays accompanied by contingent auditory feedback provide effective training tools for the therapist.

Some 20 channel systems are capable of mapping the entire EEG frequency spectrum at every electrode location in the 10–20 coordinate system in (nearly) real time. The displays are updated in 50 ms providing a useful assessment mode.

EEG Biofeedback Application. The most important factor is the therapist's training. The EEG assessment and biofeedback treatment has progressed considerably beyond that required to use the less complex modalities of temperature, electrodermal, and EMG. Fortunately, few schools are keeping up with advances. The International Society for Neuronal Regulation (ISNR) is a good source to inquiry. The Neurofeedback division of the Association for Applied Psychophysiology and Biofeedback (AAPB) and ISNR have together addressed several critical issues in EEG biofeedback methodology and clinical application.

As with most psychophysiological assessment and biofeedback applications environmental temperature (21.1–23.9 °C), humidity (30–50%), illumination (as required), auditory noise (transient free, white noise is usually acceptable), and therapist's physical position with respect to the subject must be considered and controlled.

An electrically quiet environment is also desirable. Power lines, TV and radio antennas, and mobile communications are frequent sources of interference. Basement locations usually offer the electrically quietist locations and well-placed electrodes along with carefully routed leads provide the best immunity under the therapist's control.

Electrodermal Biofeedback Instrumentation

Electrodermal activity (EDA) is essentially a measure of palmar sweat gland activity of the fingers or hands. Bulk tissue impedance is largely ignored, being clinically less significant than the eccrine sweat gland activity, modulated by the sympathetic nervous system that modern instruments measure.

Electrodermal activity can be measured either as resistance, called skin resistance activity (SRA = SRL + SRR) or it's reciprocal skin conductance activity (SCA = SCL + SCR). The SCA has the advantage of being a linear function of the number of sweat glands conducting, while SRA has a hyperbolic relationship to the number of sweat glands conducting.

Electrodermal activity includes both the tonic (average) level (SRL or SCL) and short-term (phasic) response (SRR or SCR). Training goals for SCA are to lower the tonic level, indicative of chronic sympathetic arousal and reduce the phasic changes percent, indicative of over reactivity, either spontaneously or in response to stimuli.

The magnitude of phasic changes tend to occur as a percentage of the tonic level. For conductance, a 1 μS SCR change from a tonic level of 5 μS SCL is approximately equivalent to a 2 μS SCR change from a 10 μS tonic SCL level (i.e., 20%).

Most modern instruments measure conductance although one manufacturer recently reverted to resistance so that the same programmable amplifier could be used to measure any variable resistance sensor (i.e., temperature). Since conductance and resistance are reciprocals, either can be displayed with the conversion made in circuitry or software.

Subject phasic responses are delayed by 1–3 s neurophysiologically. This latency limits the usefulness of EDA as an early in-training feedback modality. It is an excellent assessment modality where the subject receives no feedback.

The EDA is a passive electrical parameter and must be elicited by impressing a small voltage on (conductance), or passing a small current through (resistance) the two electrodes, usually placed on the fingers. Either method is referred to as "excitation".

The normal range of EDA for human subjects is

Conductance:	0.5–50 μS
Resistance:	2 MΩ–20 kΩ

Most untrained subjects range between 2 and 10 μS. A value of 50 μS would be indicative of a soaking wet hand.

Electrodermal Electrodes. Normal placement of electrodes is on the palmar surface of the index and middle fingers of either hand, selection dictated by other modality sensors on a particular hand. Electrodes are held on the fingers by Velcro straps set for just enough tension to hold them on without causing blood "pulsing" or pounding in the fingers. Silver–silver chloride is the most common electrode material, but gold, stainless steel, and nickel-plated brass are also used.

Early instruments used monopolar (dc) excitation. At the values then used, electrode polarization and subsequent amplifier blocking tended to occur. Modern instruments use 10 μA current (resistance) or 200 mV (conductance) and reverse the excitation several times per second resulting in no net charge, thus avoiding electrode polarization. These values also limit current to 10 μA as required by the FDA.

Electrolyte between electrode and skin is not normally used clinically for EDA, but should be used, along with finger cleaning for clinical research or published studies.

Electrodermal Biofeedback Amplifier. Modern EDA amplifiers employ a switching technique to reverse electrode excitation several times per second to avoid electrode polarization. This requires a reversible voltage (conductance) or current (resistance) source, circuits easily implemented by operational amplifier techniques. Values are sampled after transient effects caused by the switching. The EDA is a relatively slowly changing modality and an effective amplifier bandwidth of 1–2 Hz is adequate.

The entire EDA (EDL + EDR) should be measured and later separated. Phasic changes (EDR) can be ac coupled to remove the EDL and amplified to produce the desired display sensitivity. This produces bipolar components to each phasic response (EDR), which can be difficult for subjects to understand.

Another method is to introduce an offset in the amplifier equal to the EDL resulting in only the phasic (EDR) component that can then be amplified separately. This method requires a relatively stable EDL or a circuit that samples the EDL and adjusts the offset continuously.

Electrodermal Feedback. Use of EDA as an assessment (without feedback) modality is highly recommended since it provides insight into the chronic sympathetic arousal level and reactivity of the subject.

Using EDA as a feedback modality in the beginning subject is discouraged as it is so nonspecifically reactive, subject relaxation and learning may be impeded. Advanced subjects, however, may find it challenging to control EDR while performing tasks.

For assessment, the time line graph of the entire EDA provides the most information on current and recent past history to the therapist.

The line graph, bar graph or contingent animation are all useful visual displays for the advanced subject.

Pitch-proportional (SCA) or pitch inversely proportional (SRA) auditory tones are useful for eyes-closed or task performance training. Care must be taken in choosing the pitch range to avoid alarming the subject with a "siren" effect.

Electrodermal Biofeedback Application. Silver–silver chloride electrodes are recommended. Low cost velcro-strap electrodes are affordable for each subject's course of training. Using these between subjects is not recommended from disease and therapist liability considerations. Disposable Ag/Ag CL EMG electrodes provide an alternative, but affixing them to digits is not as convenient.

Hand washing prior to the session is advisable for both uniformity and sanitation.

Use of electrolyte is not required for most clinical training, but should be used for clinical research or published studies. Saline gel or cream is suitable for most subjects. Non-saline gel is useful for allergic subjects. Hand washing following the session is good practice.

A small percentage (i.e., 7%) of subjects show little or no EDA. Always check the instrument with a known conductance–resistance standard (or the therapist) before concluding that the instrument is mal-functioning.

Since EDA is an elicited or exosomatic measure, care must be taken to place sensors from all used instrument modalities so that the current from EDA does not interfere with or impede other simultaneous measurements. Manufacture of multimodality biofeedback systems provide guidelines to minimize these situations. This is of particular concern if two EDA channels are used on the same subject.

As EDA is a relatively high level signal and excitation switching is slow, the modality is relatively impervious to electromagnetic interference.

Cardiopulmonary Biofeedback

Respiration (RSP) training has been shown to greatly improve the subjects' ability to relax and maintain self-regulation in the face of psychological and performance stressors. It tends to reduce performance anxiety and increases oxygen uptake and waste expulsion. It also increases peripheral circulation by reducing sympathetically activated vasoconstriction. Heart rate increase during inhalation is normal, implemented by the sympathetic nervous system. Para-sympathetic action slows heart rate during exhalation. Heart rate variability (HRV) training has also provided benefits in reducing rapid heart rate and tachycardia.

The HRV (measured as HR max - HR min) is frequently as high as 20 BPM in healthy 20 year-old adults, and decreases by age 50 often to ~10 BPM. Athletically active and physically well-conditioned individuals have higher variation in heart rate. Heart rate variability training aims at increasing the variability, and frequently increases it significantly higher than 20 BPM. Some authors have claimed that it goes as high as 50 BPM in peak training. Higher HRV is considered desirable in disease prevention and health promotion applications, and lower HRV correlates with cardiovascular morbidity and mortality. For example, lower HRV is a strong independent predictor of post-MI death (4).

By combining respiratory (RSP) and heart rate (HR) instrumentation, the interplay between respiration and heart rate, referred to as respiratory sinus arrhythmia (RSA), can be assessed and used as biofeedback to train subjects to optimize their natural cardiopulmonary rhythms under the influence of stressors.

Norms for HRV training have been published (5). One of the persistent problems in the field is the failure of researchers and practitioners alike to adhere to standard nomenclature and normative values for training. The following HRV frequency ranges have been established as standard for cardiopulmonary training:

Cardiac Rhythms. High frequency: 0.15–0.4 Hz; low frequency: 0.04–0.15 Hz; very low frequency: 0.0033–0.04 Hz; and ultra low frequency: <0.0033 (beyond clinical biofeedback measurement technology) (5).

Recently, a more careful examination of HRV has shown the very slow rhythms to be of interest in assessing dysregulation, other than that caused by the ANS, to be discussed in a separate HRV section, below. The very low frequency (VLF) range is to some extent correlated with dysregulation. Rhythms in the low frequency range are to some extent correlated with optimal homeostasis. Higher HR oscillations are found in rhythms in this range.

The RSP rates vary from 2 to 30 bpm requiring channel bandwidths of 0–5 Hz to faithfully reproduce the RSP waveform. Trained subject's RSP rates at rest are individually optimum and range from 6 to 9 pm.

The HR rates vary from ~40–180 bPM. Trained subject's HRs rest at ranges between 60 and 80 BPM The HRV (RSA) of 8–16 BPM as a function of RSP is normal and desirable.

Some instruments are capable of RSA assessment and biofeedback only. Instruments designed for HRV generally can also perform RSA procedures.

Respiratory Sinus Arrhythmia Instruments. Instruments for RSA have one or two respiration channels and one heart rate channel. Two respiration channels are desirable to train the subject to breathe abdominally, not thoracically.

The RSA instruments are computer implemented. Time line graphs of abdominal and thoracic RSP along with a beat-by-beat line graph of HR comprise an excellent assessment and training display. The line or (better) filled line graph is the most common display as it shows recent past history as well as current performance. Digital meters showing RSP and HR rates can be added, but the display tends to be too complicated, particularly for beginning subjects.

Raw data is saved, so the session can be replayed as desired. Statistics can be generated, but care must be taken to consider the effect of artifact. Both RSP and PPG HR channels are susceptible to movement artifacts. Some instrument software permits editing the raw data to minimize artifact contamination of statistics.

Respiration Sensors and Amplifiers. Respiration sensors for biofeedback comprise a stretchable segment and a belt or chain that is wrapped around the circumference of the abdominal and the thoracic regions of the subject. A slight prestretch (set at the point of maximum exhalation) allows the sensor to operate over the full range of circumference change caused by breathing. Care must be taken that restrictive clothing does not impede breathing.

Circumference change caused by breathing is a relative measure affected by subject breathing, posture, type of sensor, even temperature depending on the type of sensor used. Three types of sensors are in common use.

1. Rubber Bellows (Air Filled): Following placement the system is sealed, and sensor internal pressure changes with stretch. A transducer (half or full bridge strain gage) measures pressure changes. A dc coupled bridge amplifier, with offset control, provides amplification and the ability to "position" the output at the desired level for display and/or quantification. Typical pressure variations are on the order of ±15 mmHg (1.99 kPa) gage.
2. Tubular liquid-filled strain gage: An elastometric tube filled with a conductive thixo-tropic liquid. Changes in length and diameter of the tube, caused by breathing, varies the resistance of the gauge. Since the liquid is ionic, excitation (current or voltage) is reversed several times per second to prevent polarization. The amplifier comprises a reversible voltage or current source, gain, and offset capability to provide the desired output signal positively proportional to inhalation.
3. Magneto-position transducer: A magnetic armature is moved within an excited coil to produce a current proportional to movement. An elastomeric tubing provides a restorative force to track breathing movements. The amplifier converts the magnetically induced signal, provides gain and offset to provide a voltage positively proportional to inhalation.

Heart Rate Sensors and Amplifiers. Heart rate for clinical biofeedback is measured using either a finger photoplethysmograph (PPG) sensor or by acquiring the electrocardiogram (ECG) biopotential signal. The PPG sensor is easy to use, but its' output is subject to artifact from any disturbance of the placement. Acquiring the ECG requires placement of surface electrodes, but provides the virtually artifact-free signal that is required for reliable HRV analysis, and will be discussed in that section.

The PPG sensors employ an (invisible) infrared (IR) source of ~0.9 nm wavelength to illuminate the vascular bed of a finger (or toe). A photocell detects the backscatter caused by the opacity of blood at that wavelength. It is held in place with a Velcro strap or elastic band. The amplifier provides gain producing a pulse signal having information on HR, peripheral pulse amplitude (PPA), pulse volume (PPV) and rise and decay characteristics of the pulse.

Heart Rate Variability Instruments. Heart rate variability is useful for more than RSA training. Other factors, such as chemoreceptors, baroreceptors, the renin-angiotension system and various disease states affect HRV. Many of these variations occur at frequencies too slow to be perceived by or used as feedback for subjects.

Research has shown the frequency spectrum of HRV from 0.003 to 0.4 Hz to be useful in assessing disregulation of various systems. When HRV spectral components are distributed throughout the range of 0.003–0.4 Hz, HRV is said to be incoherent. In the coherent state, virtually all the energy occurs at one frequency in the range of 0.08–0.12 Hz and is individual specific.

The most comprehensive analysis of the slowest rhythms requires a 24 h data string of HR, such as is obtained in a Holter portable monitor of ECG, so various segments of Circadian rhythms can be analyzed. Data must be edited for artifact before spectral analysis by fast fourier transform (FFT) is performed. The comprehensive analysis can give insight into undiagnosed medical conditions.

In clinical psychophysiology, 5 min HRV data resolves frequencies from 0.003 to 0.4 Hz, adequate for observing ANS activity. Some instruments have added a 60 s HRV data gathering to provide 0.06–0.4 Hz spectral data, more easily obtained in the clinical situation, and adequate for determining coherence.

The ECG amplitude ranges between 0.3 and 2 mV for the QRS complex that is used to determine the interbeat interval from which the frequency spectrum is derived. The exact characteristics of the ECG signal are not as important in HRV applications as in clinical cardiology.

ECG Sensors and Amplifiers. The ECG is most reliably obtained by placement of chest electrodes using pregelled disposable Ag/AgCl sensors. For 24 h HRV studies, chest placement is mandatory.

In clinical psychophysiology, it is preferable not to remove clothing, so the wrist–wrist or wrist–ankle placement of sensors is preferred. Polarity of the ECG signal is important so the amplifier leads must be connected according to the manufacturer's instructions.

The EMG disposable sensors can be used to acquire the ECG in the clinic, but are not recommended for longer term studies as their adhesive surface area is considerably smaller than disposable ECG sensors, promoting half-cell disturbance artifact.

Typical ECG Amplifier Specifications:

Low internal noise (<2 μV p–p).
High Input Impedance [Zin> 10 MΩ].
Bandwidth (0.16–250 Hz).
Bandwidth cutoffs (>18 dB/octave).
Notch filter (60 Hz, in the United States).
Common mode rejection ratio [CMRR $> 10^7$].
Common mode input range (CMR ± 200 MV).
Static electricity shock protection (>2000 V).

Heart Rate Variability Instruments. The HRV instruments are computer implemented. They are also capable of performing RSA procedures using either ECG or PPG sensors to acquire HR.

Most HRV/RSA software is written for the Windows operating system. It is recommended that a fairly high performance computer be used to reduce HRV analysis computational time. A computer similar to that recommended for EEG is suitable.

Some instruments also support TMP and EDA in addition to the ECG, RSP, and PPG modalities.

One manufacturer offers a dual instrument interface making it possible for two computers to access one multimodality instrument, to perform simultaneous RSA/HRV, EEG, and other psychophysiological modality assessment and biofeedback procedures, with synchronized data collection. This instrument capability makes heart–mind interaction training and clinical research possible.

Heart Rate Variability Application. As with all biofeedback procedures establishing comfortable levels of temperature and humidity, with absence of transient auditory noise, is essential for focused, efficient, and reputable performance.

The Electromagnetic Interference (EMI) environment should meet the requirements of the most sensitive modality used (i.e., EEG).

Digitization of the ECG signal should be at least 256 $s \cdot s^{-1}$, with 512 $s \cdot s^{-1}$ recommended.

APPLIED CLINICAL EXAMPLES: TENSION AND MIGRAINE HEADACHE

We will now describe the typical EMG and hand surface temperature biofeedback procedures for tension and migraine headache, which we have used both clinically and in our research (3,6–8). As noted above, it is important to remember that when referring to EMG, the authors are alluding to surface electromyography, which uses noninvasive electrodes, is painless, and involves measuring the pattern of many motor action potentials; this is in contrast to EMG used as a diagnostic procedure in neurology, which uses invasive needle electrodes, measure the activity of a single motor unit, and is often quite painful.

The standard EMG biofeedback procedure for tension headache involves measuring the muscle tension in the frontalis muscle region by placing electrodes ~2.5 cm above each eyebrow and a ground electrode in the center of the forehead (9). The frontalis region has traditionally been assumed in clinical practice to be the best overall indicator of general muscular tension throughout the body. The standard thermal biofeedback training procedure involves attaching a sensitive temperature sensor, called a thermister, to a fingertip (usually the ventral surface of the index finger of the nondominant hand) with care taken not to create a tourniquet or inhibit circulation to this phalange.

EMG biofeedback is the modality most commonly used for tension headache, with the psychophysiological rationale being that muscle tension levels in the forehead, neck, and facial areas are directly causing or maintaining/exacerbating the headaches. It is also believed that individuals suffering from tension headache have high levels of stress and using EMG biofeedback as a general relaxation technique reduces their levels of stress, enabling tension headache sufferers to better cope with their headache activity.

Hand surface temperature biofeedback for migraine headache also has two possible mechanisms of action. The psychophysiological theory states that temperature biofeedback prevents the first of the two stages of migraine (vasoconstriction of the temporal artery and arterioles; the second stage is vasodilation, which causes the actual pain) from occurring by decreasing sympathetic arousal and increasing vasodilation to the temporal artery and arterioles. An alternative mechanism of action is the use of temperature as a general relaxation technique.

The EMG or temperature signal is then electronically processed using transducers to provide the patient with information on changes in the electrical activity of the muscles or surface skin temperature on a moment by moment basis. Generally, the signals are sampled every one-tenth of a second and integrated over the entire second. Both are quite sensitive, with the EMG sensor generally detecting changes of magnitude as low as a hundredth of a microvolt. The temperature sensor typically detects changes as low as one-tenth of a degree Fahrenheit. Through this feedback, the patient undergoing EMG biofeedback training learns how to relax the musculature of the face and scalp, and also learns how to detect early symptoms of increased muscle tension. In temperature biofeedback, the patient is taught how to detect minute changes in peripheral skin temperature, with the training goal being to increase hand temperature rapidly upon detection of low hand temperature. For EMG biofeedback, the feedback signal is usually auditory, and may consist of a tone that varies in pitch, a series of clicks that vary in frequency, and so on. Given the choice, >80% of patients receiving thermal biofeedback choose the visual display. The feedback display can be the pen on a voltmeter, a numeric output of the actual surface skin temperature, or a changing graph on a video screen. The format of the visual feedback display does not seem to affect learning or treatment outcome (10).

Common Type of Feedback Schedules in Clinical Applications. One of the challenges faced by both the biofeedback clinician and patient is selecting what type of feedback is most appropriate to facilitate learning to achieve rapid therapeutic benefit. There has been little research in this area; however, there are abundant anecdotal reports among the biofeedback community. There are three types of feedback schedules in clinical practice. By far, the most widely used method for delivering feedback is an analog display, which provides continuous information to the patient. For example, a tone that varies in relative pitch and frequency depending on an increase or decrease in the response being measured. However, in many applications, this may provide too much information to the patient, leading to information processing overload, retarding the learning process. A second type of feedback schedule employed in clinical practice is a binary display, where the patient receives information that is discrete, depending on achievement of a predefined training threshold. In threshold training, the feedback is turned on or off depending on whether the patient falls above or below the threshold. Threshold training is a clinical application of an operant shaping procedure, where the patient is reinforced for achieving successively closer approximations to the training goal. The third type of feedback schedule is an aggregate display of the training progress. In this type of feedback, the patient is given summary information at the conclusion of the treatment session (e.g., data averaged over each min interval in a 20 min training session). In

clinical practice, the integrated display of aggregate feedback is the most commonly applied training schedule.

Training to Criterion. Training to criterion is a term used by clinicians that involves continuing biofeedback training until the patient achieves a specified criterion of a learning end state. For example, biofeedback training will persist until the tension headache patient has demonstrated reduced muscle tension levels in the frontalis region to a stable 1-μV level. Although there is compelling logic behind this notion, there is little empirical data to support the practice of training to criterion. Exceptions to this are a report by Libo and Arnold (11) who found that every patient who achieved training criteria on both EMG and finger temperature also reported long-term improvement, and 73% of patients who did not improve failed to achieved training criterion in either modality. In another study, Blanchard et al. (12) presented data supporting the concept of training to criterion. They observed a discontinuity in outcome for migraine headache patients who achieve 35.6 °C or higher at any point during temperature biofeedback training. Those who reached this level had a significantly higher likelihood of experiencing a clinically meaningful reduction in headache activity (at least 50%) than those who reached lower maximum levels. This apparent threshold was replicated in a subsequent study (13). More representative of the research is a recent study by Tsai et al. (14), where they found no evidence to support the concept of training to criterion in a study of hypertensives. Fifty-four stage I or stage II hypertensives were taught thermal, EMG, and respiratory sinus arrhythmia biofeedback. Most participants (76%) achieved the thermal criterion; only 33 and 41% achieved the EMG and respiratory sinus arrhythmia criterion, respectively. Achievement of the criterion level in any of the three modalities was not associated with a higher improvement rate. These results contradict the notion that training to criterion is associated with clinical improvement.

Electrode Placement. An important consideration to be made by the clinician utilizing EMG biofeedback is at what sites to place the electrodes. This decision depends in large part on which of the two general theories underlying the use of biofeedback the clinician adheres to. In most instances, electrode placements appear not to matter. However, for tension headache this may not be the case.

Although the vast majority of published reports on tension headache utilize the frontalis region electrode placement, their is some controversy about this practice. This is perhaps because the Task Force Report of the Biofeedback Society of America, in their influential position paper on tension headache (15), strongly implied that frontal placement was the "gold standard" for biofeedback with tension headache sufferers, making no mention of other site placements. In the standard placement, muscle activity is detected not only in the forehead, but probably also from the rest of the fact, scalp, and neck, down to the clavicles (16).

Some writers (17,18) advocated attaching electrodes to other sites, such as the back of the neck or temporalis area, especially if the patient localizes his/her pain there. However, three of the four studies that compared biofeedback training from different sites between subjects found no advantage of one cite over the other (19–21). Arena et al. (22) published the only systematic comparison of a trapezius (neck and shoulder region) versus frontal EMG biofeedback training regimen with tension headache sufferers. They found clinically significant (50% or greater) decreases in overall headache activity in 50% of subjects in the frontal biofeedback group versus 100% in the trapezius biofeedback group. The trapezius biofeedback group was more effective in obtaining significant clinical improvement than the frontal biofeedback group. Thus, there is some limited support for the use of an upper trapezius electrode placement with tension headache sufferers. More research needs to be done in this area.

Discrimination Training. A concept in clinical biofeedback applications that is quite often discussed, particularly among those practitioners of EMG biofeedback training, is that of discrimination training. In this procedure, patients are taught to discriminate high levels of muscle tension from moderate and low levels. Feedback is given contingent upon successful differentiation among these varying levels of muscle tension. For example, a patient is asked to produce maximal muscle tension in a particular region, and given feedback reflective of this high level of muscle activity, followed by instruction to halve this level and consequent feedback. Then, finally, they are asked to halve this again, that is produce one-quarter of the initial level of muscle activity, followed by appropriate feedback reflecting success at this level. To our knowledge, there is little reliable data demonstrating that individuals specifically taught a muscle discrimination training procedure have clinical outcomes superior to those taught a standard tension-reduction method.

Sensitivity–Gain. The gain or sensitivity of the feedback signal is important to facilitate the training process in clinical biofeedback. Too high a gain may interfere with learning by providing indiscriminate feed back for extraneous responding, leading to frustration on the part of the learner. In addition, in many response measures, too high a sensitivity leads to increased artifact. Conversely, setting the gain too low leads to lack of feedback for responses that may be clinically meaningful, thereby interfering with the learning process. In clinical practice, there are established ranges in various applications, depending on the response measure employed, individual differences in patient responsivity, and the nature of the disease state. Sensitivity may be adjusted as needed using a shaping procedure. Some response measures involve more frequent changes in gain settings than others. For example, gain is frequently adjusted in EMG biofeedback applications, because the goal often is detection of quite subtle muscular activity changes, but infrequently changed in hand surface temperature training, where gross changes in skin temperature are usually necessary for clinical improvement.

Session Length and Outline. Treatment sessions usually last 30–50 min; 15–40 min of each session is devoted to the actual feedback training. In our research (and in our

clinical work), we have typically used the following format for biofeedback training sessions:

1. Attachment of electrodes and initial adaptation: 10 min.
2. In-session baseline, during which patients are asked to sit quietly with their eyes closed: 5 min.
3. Self-control 1, during which patients are asked to attempt to decrease their forehead muscle tension levels in the absence of the feedback signal: 3 min.
4. Feedback training, with the feedback signal available: 20 min.
5. Self-control 2, during which patients are asked to continue to decrease their forehead muscle tension levels in the absence of the feedback signal: 3 min.

The two self-control conditions are included to determine whether generalization of the biofeedback response has occurred. Generalization involves preparing the patient to, or determining whether or not the patient can, carry the learning that may have occurred during the biofeedback session into the "real world". If the patient can decrease muscle tension without any feedback prior to the biofeedback condition (Self-control 1 condition), then the clinician can assume that between-session generalization has occurred. If the patient can decrease their muscle tension without any feedback following the biofeedback condition (Self-control 2 condition), then the clinician can assume that within-session generalization has occurred.

There are other methods clinicians use to train for generalization of the biofeedback response. For example, in an attempt to make the office biofeedback training simulate real world situations, many clinicians initially train patients on a recliner; then, once they have mastered the rudiments of biofeedback in this extremely comfortable chair, they progress to, respectively, a less comfortable office chair (with arms), an uncomfortable office chair (without arms), and, lastly, the standing position. Finally, giving the patient homework assignments to practice the biofeedback response in the real world is a routine way of preparing them for generalization.

BRIEF REVIEW OF CLINICAL OUTCOME LITERATURE FOR BIOFEEDBACK

Anxiety Disorders–Stress Reduction

Biofeedback as a general relaxation technique has been in existence since the late 1960s. Indeed, it is common practice to call any form of biofeedback "biofeedback assisted relaxation", stressing the stress-reduction quality of the procedure. Where diagnoses are given, it is usually generalized anxiety disorder, although mostly the research defines anxiety or stress by global self-report measures or a simple paper and pencil instrument such as the Spielberger State-Trait Anxiety Inventory (i.e., scoring in the ninetieth percentile or above), rather than standard criteria such as the American Psychiatric Association's Diagnostic and Statistical Manual IV: revised (23). The primary modalities used for anxiety and stress reduction are EMG, hand surface temperature, and EEG. Nearly all the research has demonstrated that biofeedback is superior to placebo and wait-list controls for the treatment of stress and anxiety. There is some data to suggest (24) that EEG biofeedback to increase alpha waves may be superior to forehead EMG biofeedback and EEG biofeedback to decrease alpha waves in terms of decreasing heart rate activity, but not in terms of decreasing self-reported anxiety levels, where their were no differences between the three groups. When biofeedback has been compared to relaxation therapy, there is no difference between the two treatments in terms of their clinical efficacy (25).

One typical study was that of Spence (26). He took 55 anxious subjects, and gave them either electrodermal response, hand surface temperature, or forehead EMG biofeedback based on a pretreatment psychophysiological assessment (subjects were given feedback corresponding to that physiological parameter that changed the most during stress). All groups reported significant reductions in their anxiety symptoms, and 15 months later 76% of subjects were still symptom-free for anxiety, regardless of the type of feedback they received.

Moore (27) reviewed the EEG biofeedback treatment of anxiety disorders and pointed out that there are many limitations in the research to date. Unfortunately, many of his concerns hold for the EMG and temperature biofeedback literature as well, such as comparisons to relevant placebos, examination of such factors as duration of treatment, type and severity of anxiety, and so on.

Tension and Vascular (Migraine And Combined Migraine–Tension) Headache

By a large margin, the greatest number of articles supporting the efficacy of biofeedback for any disorder in the clinical treatment literature pertains to its use with vascular and tension headache. For both types of headache, biofeedback has been shown to be superior to both pharmacological and psychological placebo, as well as wait list control, in numerous controlled outcome studies. Biofeedback for headache is usually compared to relaxation therapy or cognitive therapy (a form of psychotherapy focusing on changing an individual's pain-and stress-related self-statements and behaviors). Arena and Blanchard have recently reviewed the biofeedback treatment outcome literature on tension and vascular headache (3,7,8).

With tension headache, the biofeedback approach used is EMG (muscle tension) feedback from the forehead, neck, and/or shoulders. For relaxation therapy alone, successful tension headache treatment outcomes generally range from 40 to 55%, for EMG biofeedback alone, this value ranges from 50 to 60%, and for cognitive therapy, from 60 to 80%; when EMG biofeedback and relaxation are combined, the average number of treatment successes improves from $\sim$50 to $\sim$75%; when relaxation and cognitive therapy are combined, success increases from 40 to 65%. When compared to relaxation therapy, there is usually comparable efficacy.

For patients with pure migraine headache, hand surface temperature (or thermal) is the biofeedback modality of choice, and it leads to clinically significant improvement in

40–60% of patients. Cognitive therapy by itself gets ~50% success. A systematic course of relaxation training seems to help when added to thermal biofeedback (increasing success from ~40 to 55%), but cognitive therapy added to the thermal biofeedback and relaxation does not improve outcome on a group basis. Relaxation training alone achieves success in from 30 to 50% of patients, and adding thermal biofeedback boosts that success (from ~30 to 55%). There appears to be a trend in the literature for thermal biofeedback to be superior to relaxation therapy.

For patients with both kinds of primary benign headache disorders (migraine and tension type), the results with thermal biofeedback alone are a bit lower, averaging 30–45% success; relaxation training alone leads to 20–25% success. Thermal biofeedback consistently appears to be superior to relaxation therapy with combined headache. The best results come when thermal biofeedback and relaxation training are combined. With this combination treatment, results show 50–55% success rates (adding thermal biofeedback to relaxation raises success from 20 to 55%; adding relaxation therapy to thermal biofeedback increases success from 25 to 55%). Most experts strongly recommend a combination of the two treatments for these headache sufferers.

Lower Back Pain

Arena and Blanchard (7) recently reviewed the biofeedback literature for low back pain and concluded that biofeedback appears to hold promise as a clinically useful technique in the treatment of patients with back pain. While the evidence indicates that optimal clinical improvement is clearly obtained when biofeedback is used within the context of a comprehensive, multidisciplinary pain management program, the cumulative weight of the evidence suggest that EMG biofeedback is likely to be helpful, as a single therapy, in the treatment of musculoskeletal low back pain, obtaining success rates of from 35 to 68% improvement on follow-up.

However, there were many concerns about the literature. Only two studies have directly compared biofeedback to relaxation therapy, and both of these studies were significantly flawed so as to limit definitive conclusions. Direct comparisons of biofeedback to relaxation therapy are clearly needed. Longer (at least 1 year) and larger scale (at least 50/group) follow-up studies are required. Evaluations of treatments based on diagnosis (i.e., the cause of the pain) should be conducted. Comparisons of various biofeedback treatment procedures, such as paraspinal versus frontal electrode placement, or training while supine versus training while standing, are necessary. Finally, further evaluations of patient characteristics predictive of outcome, such as gender, race, chronicity, psychopathology, and psychophysiological reactivity, are needed.

Myofascial Pain Dysfunction

Myofascial pain dysfunction (MPD) syndrome, also known as temporomandibular joint (TMJ) syndrome, is considered a subtype of craniomandibular dysfunction that is caused by hyperactivity of the masticatory muscles. It is characterized by diffuse pain in the muscles of mastication, mastication muscle tenderness, and joint sounds and limitations. Although disagreement exists as to the cause of the hyperactivity (e.g., occlusal problems vs. psychological stress), several researchers have examined the use of EMG biofeedback as a treatment, which can provide relief by teaching patients to relax the muscles of the jaw. Consistent with the logic of this approach, the most common electrode placement is on the masseter muscle, although frontal muscle placements have also been used. Excellent overviews of the treatment of MPD syndrome can be found (28,29).

Arena and Blanchard (7) recently reviewed the MPD biofeedback literature and noted that, although the majority of the studies had significant limitations, when taken as a whole they appeared to be quite impressive in support of the efficacy of EMG biofeedback for MPD syndrome. EMG biofeedback is at least as effective as traditional dental treatments such as occlusal splint therapy. Curiously, it was noted that no MPD syndrome study of biofeedback as a treatment in and of itself had been published since 1989. Given the extremely positive results, this observation is somewhat perplexing.

Deficiencies in the research on biofeedback treatment for MPD syndrome are similar to those discussed in the lower back pain section, above. Large scale outcome studies are needed, comparing (a) masseter versus frontal versus temporalis placement sites; (b) biofeedback versus relaxation; (c) biofeedback versus traditional dental strategies, and, (d) biofeedback in conjunction with other treatments versus traditional dental strategies. The latter approach has been used by Turk and co-workers (30–32), in a number of recent, methodologically elegant studies. In these studies various combinations of biofeedback, stress management training, and intraoral appliances were used, with results showing strong support for combined treatments. Finally, lack of long-term follow-ups, or for that matter, any follow-up at all, is a serious limitation that needs to be corrected.

Fibromyalgia

There have been a number of studies examining the efficacy of EMG biofeedback in the treatment of fibromyalgia (see Arena and Blanchard (5), for a review of the studies before 2000). The majority of the studies concluded that EMG biofeedback is useful in reducing fibromyalgic pain. Fibromyalgia is a type of nonarticular, noninflammatory rheumatism that is characterized by diffuse pain, sleep disturbance, tenderness, and functional impairment. Three studies have been published since 2000. Mueller et al. (33) gave 38 fibromyalgia patients EEG biofeedback, noted statistically significant decreases in pain, mental clarity, mood, and sleep. Van Santen et al. (34) compared physical fitness traing to EMG biofeedback and usual treatment on 143 female patients with fibromyalgia. They found no difference between the three groups on any measure. Recently, Drexler et al. (35) broke 24 female fibromyalgia patients down into those with abnormal psychological test (MMPI) results and those with normal psychological test profiles. Psychologically abnormal

individuals were helped more by the biofeedback training than were psychologically normal individuals. Given the relatively promising results [all five of the pre-2000 studies (36–40) obtained positive results], it appears that large scale, controlled EMG biofeedback studies looking at factors such as psychological profiles and gender would now be appropriate.

Biofeedback for Gastrointestinal Disorders: Constipation Pain, Irritable Bowel Syndrome, Urinary, and Fecal Incontinence

The biofeedback literature on treatment of constipation pain, especially in children, is both impressive and growing. In adults, Jorge et al. (41) recently reviewed the literature and noted that, overall, mean percentage of success is 68.5% for studies that examine constipation attributable to paradoxical puborectalis syndrome. Mason et al. (42) examined 31 consecutive patients who received biofeedback training for idiopathic constipation. Twenty-two of the patients felt subjectively symptomatically improved. They noted that the symptomatic improvement produced by biofeedback in constipated patients was associated with improved psychological state and quality of life factors.

In the constipation pain literature regarding children, three studies particularly stand out. Benninga et al. (43) gave 29 children who suffered from constipation and encopresis an average of five sessions of EMG biofeedback of the external anal sphincter. At 6 weeks, 55% were symptom free. Another group of investigators (44) placed 13 children who suffered from constipation into a standard medical care group, while another group of 13 children were placed in a EMG biofeedback (of the external anal sphincter–from 1 to 6 sessions) plus standard medical care group. At 16 month follow-up, all children were significantly improved, with the biofeedback plus standard medical care group significantly more improved than the standard medical care only group.

One large scale study, however, does not support the efficacy of EMG biofeedback for constipation pain. In a procedure similar to Cox et al. (44), van der Plas et al. (45) placed 94 children who suffered from constipation into a standard medical care group, while another group of 98 children were placed in a five-session EMG biofeedback (of the external anal sphincter) plus standard medical care group. At 18 month follow-up, over one-half of the children in both groups were significantly improved, with no significant difference between the two groups. In spite of this large scale study suggesting no advantage to the inclusion of EMG biofeedback to conventional medical care, we believe that there is sufficient evidence to conclude that EMG biofeedback is a useful technique in treating the pain of both adult and childhood constipation, especially when the patient has proven refractory to standard medical care.

Biofeedback for irritable bowel syndrome has been in existence since 1972, but nearly all of the studies are small and uncontrolled. The type of feedback is generally thermal biofeedback, however, two groups have used novel feedback approaches with some success. Leahy et al. (46) have developed an electrodermal response biofeedback device that uses a computer biofeedback game based on animated gut imagery. This significantly reduced symptoms in 50% of 40 irritable bowel syndrome patients. Radnitz and Blanchard (47), using an electronic stethoscope placed on subjects' abdomens, gave bowel sound biofeedback to five individuals with irritable bowel syndrome. Three of the five patients had reductions in their chronic diarrhea by over 50% (54, 94, and 100%). Results were maintained at 1- and 2-year follow-up (48). Large scale controlled outcome studies comparing biofeedback to pharmacological and dietary interventions for irritable bowel syndrome symptoms need to be conducted.

Biofeedback For Cancer Chemotherapy Effects

Biofeedback has been used to decrease the negative side effects of cancer chemotherapy, especially the anticipatory nausea. While biofeedback assisted relaxation does seem to help these patients, biofeedback by itself (i.e., not using a relaxation emphasis), while reducing physiological arousal, does not reduce the anticipatory nausea. This is an area where relaxation therapy seems to have a clear advantage over biofeedback. For example, Burish and Jenkins (49) randomly assigned 81 cancer chemotherapy patients to one of six groups in a 3 (EMG biofeedback/skin temperature biofeedback/no biofeedback) $\times$ 2 (relaxation/no relaxation) factorial design. They concluded, "The findings suggest that relaxation training can be effective in reducing the adverse consequences of chemotherapy and that the positive effects found for biofeedback in prior research were due to the relaxation training that was given with the biofeedback, not the biofeedback alone" (p. 17).

Biofeedback for Cardiovascular Reactivity: Hypertension, Raynaud's Disease, and Cardiac Arrhythmia

Biofeedback has been used as a treatment for essential hypertension since the late 1960s. The type of feedback used is either direct blood pressure feedback or temperature biofeedback. There appears to be no difference in terms of clinical outcomes between the two biofeedback modalities. In a recent influential meta-analysis of 22 randomized controlled outcome studies, Nakao et al. (50) found that biofeedback resulted in averaged blood pressure decreases of 7.3/5.8 mmHg (0.97/0.77 kPa) compared to clinical visits or nonintervention controls. It resulted in averaged blood pressure decreases of 4.9/3.5 mmHg (0.65/0.46 kPa) compared to sham or nonspecific behavioral interventions. Statistical analysis indicated that, after controlling for the effects of initial blood pressures, biofeedback decreased blood pressure more than nonintervention controls, but not more than sham or nonspecific behavioral interventions. Further analyses revealed that when the treatments were broken down into two types, biofeedback assisted relaxation, as opposed to simple biofeedback that did not offer other relaxation procedures, was superior to sham or nonspecific behavioral intervention. Nakao et al. (50) concluded that, "Further studies will be needed to determine whether biofeedback itself has an antihypertensive effect beyond the general relaxation response" (p. 37).

It has long been believed that temperature biofeedback is more efficacious than medication in the treatment of Raynaud's disease (51,52). Raynaud's disease is a disease of the peripheral circulatory system that is caused by insufficient blood supply to the hands and feet. It can result in cyanosis, numbness, pain, and, in extreme cases, gangrene and subsequent amputation of the affected finger or toe. The vasospastic attacks are triggered by cold and, to a lesser extent, anxiety and stress. Recent data, however, has failed to transparently support the belief that temperature biofeedback is a more effective treatment than medication for Raynaud's disease.

The Raynaud's treatment study (53) was a large, multicenter randomized controlled trial comparing sustained relief nifedipine, pill placebo, temperature biofeedback, and EMG biofeedback (a behavioral control) on 313 individuals diagnosed with primary Raynaud's disease. Results indicated that while nifedipine was significantly different from medication placebo in reducing vasospastic attacks, temperature biofeedback was not significantly different from psychological placebo (EMG biofeedback) in reducing vasospastic attacks. Comparison of nifedipine and temperature biofeedback indicated a nonsignificant ($p = 0.08$) trend for the nifedipine to result in greater reductions in vasospastic attacks. However, 15% of the nifedipine group had to discontinue the treatment due to adverse reactions to the medication. The interpretation of the biofeedback results of the Raynaud's treatment study, however, have been criticized by the behavioral investigators of the project (54). They note that a substantial proportion of subjects in the temperature group (65%) did not achieve learning, compared to only 33% in a normal comparison group who achieved successful learning by the end of the 10 biofeedback sessions in the protocol.

EEG Biofeedback (Neurofeedback)

Ramirez et al. (55) exhaustively reviewed the scientific literature on EEG biofeedback treatment of Attention Deficit Disorder (ADD). These authors conclude that, as in many other areas of clinical biofeedback practice, the positive evidence from anecdotal sources and case reports is plentiful, but a dearth of rigorous studies does not allow firm inferences to be drawn about the therapeutic efficacy of enhanced alpha wave activity and hemispheric lateralized biofeedback training. The EEG biofeedback training with a combined training goal of modifying the pattern of theta and beta wave activity has shown promising implications for management of ADD in adults. Studies using the theta/beta training paradigm have reported significant improvement in academic, intellectual, and adaptive behavioral functioning following EEG treatment. Other studies using sensorimotor rhythm training (recording from the "Rolandic" cortex) have produced behavioral and cognitive improvements in ADD patients. Unfortunately, these studies like those in other therapeutic areas of biofeedback are plagued with methodological problems including small sample sizes, absent or inadequate placebo controls, no randomization to treatment conditions, and insufficient follow-up of patient status. However, some authors of recent nonrandomized studies contend that EEG biofeedback shows promising evidence of therapeutic efficacy on the core symptoms of childhood ADHD in comparison to or in combination with standard stimulant medication therapy, family counseling, and an individualized educational intervention (56).

Another clinical problem in which EEG biofeedback was tested in the early 1970s was for control of frequent and disabling seizures. These studies have been reviewed by Lubar (57). The most common types of EEG recording and feedback training successfully studied in human subjects are EEG alpha rhythm (8–13 Hz) recorded from the occipital region of the brain, theta activity (4–7 Hz), and beta activity (>14 Hz). With the introduction over the recent decades of effective and relatively safe antiepileptic drugs, interest in systematic research and clinical application of EEG biofeedback as a nonmedication method of seizure control has waned. However, intractable seizures are still encountered in routine clinical practice despite all pharmacotherapeutic efforts. Implantable stimulatory devices and surgical interventions are reserved for highly selected patients and carry significant risks. For those patients with uncontrolled seizure disorder who have been unresponsive to standard anticonvulsant medication regimens and/or are not candidates for surgical treatment, Lubar has advocated that they be considered for a trial of sensorimotor rhythm EEG biofeedback training for the most common types of psychomotor seizures (57). Note that the equipment is expensive and the training procedures are complex and time consuming, and thus practitioners familiar with EEG biofeedback treatment of epilepsy may be difficult to find.

Quantitative EEG recording and specialized biofeedback training protocols have been developed and tested in the treatment of addictive disorders such as alcoholism. Peniston et al. (58) studied a protocol for enhancing EEG alpha and theta wave activity, and improving "synchrony" among the brain wave rhythms along the power spectrum. Peniston et al. (58) propose that alcoholics have a predisposition to "brain wave desynchrony" and deficient alpha activity and show a vulnerability to alcohol's capacity to produce reinforcing (pleasant and relaxing) levels of slow brain wave activity. These investigators have evaluated the treatment in a series of studies suggesting that their neurotherapy protocol reduces subjective craving among severe alcohol abusers, improves psychological functioning on personality measures, increases alpha and theta activity levels, increases beta-endorphin levels, and increases time to relapse. However, a large, independent randomized controlled trial of Neurotherapy did not show the incremental benefit to relapse prevention of adding electronic neurotherapy to a traditional residential treatment program for severe, chronic alcoholics (59). Although widely practiced, the clinical utility and theoretical rationale of EEG biofeedback in treating alcohol abusers remains controversial among the scientific biofeedback community. While promising data exists to suggest the potential role of EEG biofeedback in substance abuse treatment, further research is needed to illuminate the conceptual basis of such treatment and the reliability of clinical improvements for alcoholism and other addictive disorders.

There are very limited data from controlled studies on the use of EEG biofeedback for control of symptoms associated with Tourette syndrome, a behavioral impulse control disorder characterized by a constellation of motor and vocal tics (involuntary behaviors). A few scattered case reports describe positive results using a course of EEG sensorimotor rhythm biofeedback training to treat complex motor tics and associated attention deficit symptoms (60). There may be overlap in this treatment approach with the observation that epilepsy cases with motor involvement show some remediation following sensorimotor EEG biofeedback training. There is speculation and anectodal reports to suggest that anxiety, depression, and attentional symptoms associated with complex tics in Tourette's may be the most responsive targets for psychophysiological treatment. Because of the multiple symptom clusters of Tourette's, and the likelihood that different treatment protocols are needed to address the range of affected behaviors, focusing treatment on the whole condition is difficult, and often a prioritization of the most severe problems must occur to serve as the focus of clinical attention. Most patients are managed on a medication regimen by their physicians and EEG biofeedback is seen a useful adjunct in selected cases that have not responded adequately to pharmacologic management alone or where medication usage is to be reduced for some reason.

Cardiovascular Reactivity

Heart rate variability (HRV) biofeedback is being studied as a psychophysiological means of managing heart problems such as cardiac arrhythmia. Earlier isolated attempts at biofeedback interventions with cardiovascular ailments using simpler unitary heart rate (beats/min) or blood pressure (mmHg) measures have not been remarkably successful on modifying disease states. Heart rate variability is derived from the standard deviation of the beat-to-beat time interval (in ms) recorded in the laboratory with an ECG machine or with a Holter monitor using 24 h ambulatory monitoring methods. Heart rate variability has been proposed as a more robust metric of overall cardiac health in that it provides an indirect marker of the heart's ability to respond to normal regulatory impulses that affect its rhythm. With higher levels of HRV, it is proposed that there is a better balance between the combined sympathetic and parasympathetic inputs to the heart. Generally, greater HRV is associated with relaxed states and slow or regular breathing pattern. The HRV biofeedback training is claimed to offer a more precise method for helping clients to moderate the heightened sympathetic activity that is associated with elevated stress, anxiety, and dysphoric mood. Relatively greater levels of heart rate variability have been associated with better heart health. Biofeedback of breathing rate and depth is also used to increase respiratory sinus arrhythmia, which may be connected to therapeutic increases in HRV. A few small-scale studies have been conducted that show the clinical potential of HRV biofeedback in cardiovascular diseases (61); however, the results are inconsistent, and methodological problems abound with the existing studies. As yet, there is little evidence from larger scale randomized controlled trials conducted at independent laboratories demonstrating the therapeutic efficacy of HRV biofeedback in specific cardiovascular disease states such as cardiac arrhythmia.

Incontinence: Urinary and Fecal

EMG biofeedback training of the bladder-urinary sphincter and pelvic floor musculature has been found to be an efficacious intervention for urge urinary incontinence, especially among female geriatric populations, and is usually related to destrusor muscle contraction instability or reduced bladder volume (62). Some form of Kegel exercises are often used to train the muscles of the pelvic floor that are in continuity with the external urethral sphincter. Biofeedback with behavior modification training of the anorectal-anal sphincter musculature along the pelvic floor has been reported to be successful in treatment of fecal incontinence of various etiologies (63). Small, insertable EMG sensors are usually used in current treatment protocols for urinary incontinence in female patients and for fecal incontinence. A second EMG channel with abdominal placement is often recommended to better isolate contraction of the pelvic muscles from activity of accessory muscle of the legs, buttocks, and abdomen during the training exercises. Some degree of voluntary contractibility of sphincter muscles and rectal sensitivity are necessary for successful biofeedback treatment. While biofeedback training for urinary incontinence has a longer history of usage, and thus a larger empirical base (64), there is considerable evidence to suggest the efficacy of EMG biofeedback in a majority of adult patients with fecal incontinence (65). Unfortunately, there was a great deal of variability in biofeedback instrumentation used among these studies, treatment protocols followed, and outcome measures with uncertain validity.

Stroke and Mild Traumatic Brain Injury

There is very limited research in the area of EEG biofeedback in treatment and rehabilitation of the neurological impairments resulting from stroke or closed head injury. There are a number of anecdotal reports and small case series that suggest a place for quantitative EEG analysis in the functional assessment of neurological symptoms secondary to stroke and head injury. A highly individualized QEEG protocol used in these studies is sometimes called EEG entrainment feedback recording from the surface of brain regions suspected to be pathologically involved in the functional impairments (66,67). Neuromuscular reeducation is a general term used to describe assessment and treatment methods that may include EMG biofeedback and are applied to helping neurologically impaired patients (such as poststroke patients) with regaining gross motor functions necessary for carrying out activities of daily living and ambulation. In a meta-analysis of controlled trials using EMG biofeedback for neuromuscular reeducation in hemiplegic stroke patients, the authors concluded that EMG biofeedback resulted

in significant functional gains (68). While these results are promising, the specific effects of EMG biofeedback in stroke rehabilitation remain unclear as some of the studies reviewed included other interventions such as physical therapy or gait training as part of the rehabilitation program.

Sexual Dysfunction

Surface EMG biofeedback of targeted abnormalities in pelvic floor musculature are implicated in the pathogenesis of vulvovaginal pain (vulvodynia) syndromes such as dsypareunia resulting from vulvar vestibulitis. These have been successfully used in the stabilization of pelvic floor muscles leading to 83% reduction of pain symptoms, improved sexual function, and psychological adjustment at 6-month follow-up (69). As in other EMG biofeedback protocols for assessment and modification of abnormal pelvic floor musculature, an individualized assessment is performed to identify the patient's specific neuromuscular abnormality, with subsequent biofeedback training designed to modify the muscle tension and contractile weakness of the target muscles. However, gynecological surgery (vestibulectomy), on average, appears to produce superior outcomes (70).

FUTURE DIRECTIONS OF BIOFEEDBACK THERAPY

There are five areas that biofeedback research and clinical work are heading or should focus on. They are (*1*) expanding upon and refining current research; (*2*) applications of biofeedback and psychophysiological assessment to the "real world" environment (i.e., ambulatory monitoring); (*3*) applications of biofeedback training to new populations; (*4*) applications of biofeedback to applicatuons of biofeedback to the prinany the primary care setting; (*5*) alternative methods of treatment delivery.

1. Expanding Upon and Refining Current Research. Although biofeedback is considered a mature treatment, there are surprisingly many basic areas that need to be explored further. Such basic questions as (a) whether massed versus distributed practice produces greater physiological learning, (b) whether the presence or absence of the therapist in the room retards or enhances the acquisition of the biofeedback response, (c) the usefulness of coaching, (d) is their any value in training to criterion, (e) whether group biofeedback enhances or retards psychophysiological learning or clinical affects clinical outcome, have not been satisfactorily answered. Moreover, with the notable exception of headache, nearly every area is missing large scale treatment outcome studies (i.e., 25 or greater subjects per condition), in which biofeedback treatment is compared to placebo, another psychological treatment, conventional medical treatment, and so on. Many studies fail to describe the instrumentation and biofeedback procedures sufficiently to allow replication of the research. Often diagnostic criteria are not given, or diagnoses are commingled (e.g., conduct disorder children with attention deficit disorder children, or generalized anxiety disorder with simple phobias). Such failures to answer basic research questions or to conduct research in a scientifically acceptable manner are troubling and need to be corrected.
2. Applications of Biofeedback and Psychophysiological Assessment To the "Real World" Environment (i.e., ambulatory monitoring). Biofeedback clinicians have attempted to use their psychophysiological monitoring equipment to assist in setting treatment goals and to further explore the relationship between the mechanism of action believed to be involved in the underlying pathology of the disorder in question. For example, many clinicians use ambulatory blood pressure monitors to determine when their hypertensive patients are most reactive (work, driving, etc.) and tailor exercises to be maximized around those situations of elevated blood pressures. Use of such ambulatory equipment for other responses such as EMG, hand temperature, and respiration would be quite useful and such studies need to be performed.

   There have been only a few studies examining the relationship between ambulatory monitoring in the naturalistic environment and the presumed pathological response underlying the disease, with the exception of bruxism and temporomandibular joint dysfunction, where measurement in the natural environment by telemetry, portable tape recording, and digital EMG integration have been reported (71). Unfortunately, with those exceptions, when such studies have been conducted, they have arrived at negative findings, quite possibly due to the difficulty in reducing the data and inability to control all the relevant variables (sleep is a relatively controlled environment). For example, Hatch et al. (72) had 12 tension headache subjects and nine nonheadache controls wear a computer-controlled EMG activity recorder in their natural environment for 48–96 consecutive hours. The EMG activity of the posterior neck or frontal muscles was recorded 24 h/day. During waking hours, subjects rated their perceived levels of stress, pain, and negative affect at 30-min intervals. The EMG activity of headache and control subjects did not differ significantly, and EMG activity did not covary with stress, pain, or negative affect. Cross-correlations among EMG activity, pain, and stress revealed little evidence of leading, contemporaneous, or lagging relationships. Interrupted time series analysis showed no consistent muscle hyperactivity during a headache attack compared to a headache-free baseline period.

   Arena and co-workers designed a portable activity monitor for simultaneously recording and quantifying surface EMG signals and body movements in the natural environment (73). Two independent channels record EMG activity from symmetric muscle groups to determine contraction patterns. The EMG signals are amplified, filtered, integrated for 1 s, and converted to a digital value. Full scale was

jumper selected to accommodate a wide range of muscle activity. Electrode resistance >20 kΩ generates an alarm to signal poor contact or lead-off condition. The EMG voltages less than a preset threshold are not integrated.

The movement sensors are electrolytic potentiometers whose output are proportional to angular position and linear acceleration. The outputs are differentiated and summed to obtain angular acceleration with minimal response to linear movement. The peak value and 1 s integral are converted to digital values.

Subjective evaluation of pain and activity may be annotated by a 16-button keypad. An hourly auditory alarm reminds the user to enter subjective evaluations.

Data is saved in static random-access memory in binary coded 3-byte words. Power is supplied by a 9 V alkaline battery and converted to ±5 V by switching regulators. At the end of 18 h of recording, all power is turned off except for standby power to memory. A low-battery condition will also switch power to the standby mode. Data retained in the standby mode is valid for at least 7 days.

Arena et al. demonstrated that the device is highly reliable in 26 healthy controls (74). They then had 18 tension-type headache subjects and 26 control subjects wear the device attached to the bilateral upper trapezius muscles for 5 consecutive days for up to 18 h a day (75). During waking hours, subjects rated their perceived levels of stress, pain, and physical activity at 60-min intervals. Similar to Hatch, the EMG activity of headache and control subjects did not differ significantly, and EMG activity did not covary with stress, pain, or physical activity levels. Examination of cross-correlations among EMG activity, pain, physical activity, and stress revealed little evidence of an isomorphic, precursor or consequence relationship. There were no consistent differences between a headache and nonheadache state on muscle activity levels.

Arena et al. (75) concluded that there were so many variables entering into the natural environment, that use of such devices required a sophistication not available to the average clinician or researcher, and that treating headache patients nomothetically as an aggregate, rather than idiographically, as individuals, may also present difficulties. For example, some individuals may lie down as soon as a headache begins, while others may continue with their daily routine. Other individuals may have consequence, precursor, or isomorphic relationships between their head pain, except the changes occur on an every 5 min basis rather than a 1 h basis. With still others, to identify a relationship 5 consecutive days is not enough. Given the fact that the technology has increased exponentially to allow much more sophisticated data reduction and statistical analysis, we feel that the time is now ripe for a renaissance in such an area of research, which has been dormant for nearly a decade.

3. Applications of Biofeedback Training to Other Populations. As biofeedback is considered to be an established field, investigators have begun to take the treatments and expand it to other, similar clinical problems. For example, in the field of headache, biofeedback has been shown to be effective with the elderly, children, and pregnant women (6). Areas that need to be explored further to determine whether biofeedback treatment effects can be generalized are headaches in depressed individuals, headaches in individuals following cerebral aneurysm, headaches due to eyestrain, posttraumatic headache, and headache in multiple sclerosis patients. Similarly, the anxiety disorders literature needs to be expanded to include children, the elderly, anxiety due to a medical condition, and so on.
4. Application in Primary Care Medical Settings. Because of the growing recognition of the high prevalence of psychosomatic and psychophysiological disorders that present in primary care settings, the increased availability and implementation of psychophysiological assessment and biofeedback interventions in these healthcare settings appears to be timely (76). Many behavioral medicine interventions including biofeedback may be more efficiently and effectively delivered in these primary practice settings as the focus of these interventions is often toward the goals of preventing or slowing disease progression rather than treating severe or complicated problems that are well established. This approach is in contrast to conventional practice where patients with complicated medical problems or cooccurring psychological symptoms are referred out to specialty behavioral medicine clinics or other specialists (e.g., physical therapists) for psychophysiological treatment. As many chronic health problems are progressive in nature, by the time referral is made, the patient's condition is likely to have worsened to the point where behavioral intervention including biofeedback training may have far less impact than had it been instituted earlier in the disease course.

However, for biofeedback to be successfully integrated into the busy primary care office practice setting, certain modifications will have to be made in the context of assessment protocol and treatment delivery. First, behavioral assessment will have to be brief, but informative and practical, yielding results that are helpful to the primary care team in managing the patient's medical problems. The assessment results will have to be readily incorporated into the medical record of the patient rather than assigned privileged status, as mental health records frequently are, and therefore accessible to few if any providers for reference in primary care delivery. Second, the psychophysiology assessment and biofeedback treatment program will have to be carefully standardized and mid-level providers such as nurses, physician assistants, psychology technicians, or other mental health therapists trained in the competent and efficient delivery of these services. A doctoral level psychologist on staff or

consulting from another facility should be available to supervise these services to monitor quality and assess outcomes. Third, the rapid advancement of biofeedback equipment in terms of measurement accuracy, increased reliability with precision electronics, lowered cost, much improved portability through miniaturization, and enhanced patient convenience with alternative sensor technology has enabled the possibility of almost entirely home-based, self-administered treatment. Instruction and support can be provided by nursing staff, consultant psychologist, or other health professional through less frequent office visits and telephone consultation as needed. Arena and Blanchard have recently outlined in greater detail steps one should take to apply behavioral medicine techniques such as biofeedback to the trement of headaches in the primary care setting (3).

5. Alternative Methods of Treatment Delivery. The availability of relatively low cost, high precision biofeedback training devices lends itself to the possibility of a limited-therapist-contact, largely home-based treatment regimen. Blanchard and co-workers published three separate studies (77–79) evaluating a treatment regimen of three sessions (>2 months) combining thermal biofeedback and progressive relaxation training. In all three instances, very positive results were found for this attenuated form of treatment. Similar results were reported by Tobin et al. (80).

   We believe that some limited therapist contact is often necessary, so that patients understand the rationale for the treatment and that problems (trying too hard, thermistor misplacement, etc.) can be caught and corrected early. We also believe that detailed manuals to guide the home training, and telephone consultation to troubleshoot problems, are crucial in this approach. Given the national push for improving the efficiency of treatments, this approach has much to recommend it. We should also note that this home-based approach was not as successful as office-based treatment of essential hypertension with thermal biofeedback (81).

   This limited therapist contact does not have to be face-to-face with the therapist, however. It can be conducted via the Internet or using a videoconferencing telemedicine application. Devineni and Blanchard (under review 82) conducted a randomized controlled study of an Internet-delivered behavioral regimen composed of progressive relaxation, limited biofeedback with autogenic training, and stress management in comparison to a symptom monitoring waitlist control. Thirty-nine percent of treated individuals showed clinically significant improvement on self-report measures of headache symptoms at post-treatment. At 2-month follow-up, 47% of participants maintained improvement. There was a 35% within-group reduction of medication usage among the treated subjects. The Internet program was noticeably more time cost-efficient than traditional clinical treatment. Treatment and follow-up dropout rates, 38.1 and 64.8%, respectively, were typical of behavioral self-help studies.

   Arena et al. (83) recently reported a small ($n = 4$) uncontrolled study investigating the feasibility of an Internet and/or telemedicine delivery modality for relaxation therapy and thermal biofeedback for vascular headache. Each subject was over the age of 50 and had suffered from headaches for >20 years. Subjects came into the clinic for treatment but never saw the therapist in person. Instead, all treatment was conducted through the use of computer terminals and monitors. The only difference between this treatment and office-based treatment was the physical presence of the therapist. Results indicated one of the subjects was a treatment success (>50% headache improvement, and two others had between 25 and 50% improvements. Thus, it seems that further exploration into the potential of telemedicine and internet delivery of psychophysiological interventions is warranted.

In summary, an attempt has been made to review the basic theories underlying the application of biofeedback training to the amelioration of a broad range of general medical and psychiatric disorders. Also covered are the main types of biofeedback systems, technical specifications of instrumentation, and engineering design considerations for major functional components of biofeedback apparatus. A sampling of the many clinical problem areas in which psychophysiological assessment technology and biofeedback instrumentation have been utilized with varying degrees of success is discussed. This coverage of instrumentation is not exhaustive. Given the rich basic science underpinnings of biofeedback and the wide appeal among both health professionals and the general public, this coverage was by necessity selective and opportunistic. Throughout this article are found references to key resources of primary literature and authoritative reviews of the biofeedback field. This technological area is one of the most promising of both professional psychology practice and consumer oriented general healthcare. The field of biofeedback is far from matured, with medical application of basic research finding beginning only ~30 years ago. The field is probably in its second generation with a more rigorous examination of its scientific underpinnings, casting aside unproven or implausible theories related to disease process and treatment efficacy, and development of a sound empirical basis for its assessment methods and interventions. We are witnessing the continued rapid advancement of microcomputer technology and digital electronics coupled with the accumulation of knowledge of the basic mechanisms involved in human health and disease. There is a great potential for an evolution of biofeedback from its early origins with focus on nonpathological states and development of body awareness, human potential, and wellness to a more refined and sophisticated understanding and application of techniques toward the maintenance of health and prevention of disease states that is seamlessly integrated into the individual's lifestyle.

BIBLIOGRAPHY

Cited References

1. Olton DS, Noonberg AR. Biofeedback: Clinical applications in behavioral medicine. Englewood Cliffs, NJ.: Prentice Hall; 1980.
2. Wilder J. The law of initial values. Psychosom Med 1950;12: 392–400.
3. Arena JG, Blanchard EB. Assessment and treatment of chronic benign headache in the primary care setting. In: O'Donohue W, Cummings N, Henderson D, Byrd M, editors. Behavioral integrative care: Treatments that work in the primary care setting. New York: Allyn & Bacon; 2005. p 293–313.
4. Carney RM, Blumenthal JA, Stein PK, Watkins L, Catellier D, Berkman LF, Czajkowski SM, O'Connor C, Stone PH, Freedland KE. Depression, heart rate variability, and acute myocardial infarction. Circulation 2001; 104:2024–2028.
5. Task Force of the European Society of Cardiology and The North American Society of Pacing and Electrophysiology. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. Eur Heart J 1996;17:354–381.
6. Arena JG, Blanchard EB. Biofeedback training for chronic pain disorders: A primer. In: Gatchel RJ, Turk DC, editors. Chronic pain: Psychological perspectives on treatment. 2nd ed. New York: Guilford Publications; 2002. p 159–186.
7. Arena JG, Blanchard EB. Biofeedback Therapy for Chronic Pain Disorders. In: Loeser JD, Turk D, Chapman RC, Butler S, editors. Bonica's Management of Pain. 3rd ed. Baltimore: Williams & Wilkins; 2001. p 1755–1763.
8. Blanchard EB, Arena JG. Biofeedback, relaxation training and other psychological treatments for chronic benign headache. In: Diamond ML, Solomon GD, editors. Diamond's and Dalessio's The Practicing Physician's Approach to Headache. 6th ed. New York: W. B. Saunders; 1999. p 209–224.
9. Lippold DCJ. Electromyography. In: Venables PH, Martin I, editors. Manual of Psychophysiological Methods. New York: John Wiley & Sons; 1967.
10. Evans DD. A comparison of two computerized thermal biofeedback displays in migraine headache patients and controls. [Unpublished dissertation]. State University of New York at Albany; 1988.
11. Libo LM, Arnold GE. Does training to criterion influence improvement? A follow-up study of EMG and thermal biofeedback. J Behav Med 1983;6:397–404.
12. Blanchard EB, Andrasik F, Neff DF, Saunders NL, Arena JG, Pallmeyer TP, Teders SJ, Jurish SE, Rodichok LD. Four process studies in the behavioral treatment of chronic headache. Behav Res Ther 1983;21:209–220.
13. Morrill B, Blanchard EB. Two studies of the potential mechanisms of action in the thermal biofeedback treatment of vascular headache. Headache 1989;29:169–176.
14. Tsai P, Calderon KS, Yucha CB, Tian L. Biofeedback training to criteria and blood pressure reduction. Proceedings of the 34th Annual Meeting of the Association for Applied Psychophysiology and Biofeedback. Wheat Ridge, Colorado: AAPB; 2003.
15. Budzynski T. Biofeedback in the treatment of muscle-contraction (tension) headache. Biofeedback Self Regul 1978; 3:409–434.
16. Basmajian JV. Facts vs. myths in EMG biofeedback. Biofeedback Self Regul 1976;1:369–378.
17. Belar CD. A comment on Silver and Blanchard's (1978) review of the treatment of tension headaches by EMG biofeedback and relaxation training. J Behav Med 1979;2:215–218.
18. Hudzinski LG. Neck musculature and EMG biofeedback in treatment of muscle contraction headache. Headache 1983;23:86–90.
19. Hart JD, Cirhanski KA. A comparison of frontal EMG biofeedback and neck EMG biofeedback in the treatment of muscle-contraction headache. Biofeedback Self Regul 1981;6:63–74.
20. Philips C. The modification of tension headache pain using EMG biofeedback. Behav Res Ther 1977;15:119–129.
21. Philips C, Hunter M. The treatment of tension headache. II. Muscular abnormality and biofeedback. Behav Res Ther 1981;19:859–489.
22. Arena JG, Bruno GM, Hannah SL, Meador KJ. A comparison of frontal electromyographic biofeedback training, trapezius electromyographic biofeedback training and progressive muscle relaxation therapy in the treatment of tension headache. Headache 1995;35:411–419.
23. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders IV-TR. Washington, DC: APA Press; 2000.
24. Rice KM, Blanchard EB, Purcell M. Biofedback treatments of generalized anxiety disorder: preliminary results. Biofeedback Self Regul 1993;18:93–105.
25. Eppley KR, Abrams AJ, Shear J. Differential effects of relaxation techniques on trait anxiety: a meta analysis. J Clin Psychol 1989;45:957–974.
26. Spence J. Maximization of biofeedback following cognitive stress pre-selection in generalized anxiety. Percept Mot Skills 1986;63:239–242.
27. Moore NC. A review of EEG biofeedback treatment of anxiety disorders. Clin Electroencephalog 2000;31:1–6.
28. Crider AB, Glaros AG. A meta-analysis of EMG biofeedback treatment of temporomandibular disorders. J Orofac Pain 1999;13:29–37.
29. Gevirtz RN, Glaros AG, Hooper D, Schwartz MS. Temporomandibular disorders. In: Schwartz MS, editor. Biofeedback: A practitioner's guide. 2nd ed. New York: Guilford; 1995. p 411–428.
30. Turk DC, Rudy TE, Kubinski JA, Zaki HS, Greco CM. Dysfunctional patients with temporomandibular disorders: Evaluating the efficacy of a tailored treatment protocol. J Consult Clin Psychol 1996;64:139–146.
31. Turk DC, Zaki HS, Rudy TE. Effects of intraoral appliance and biofeedback/stress management alone and in combination in treating pain and depression in patients with temporomandibular disorders. J Prosthet Dent 1993;70:158–164.
32. Greco CM, Rudy TE, Turk DC, Herlich A, Zaki HH. Traumatic onset of temporomandibular disorders: Positive effects of a standardized conservative treatment program. Clin J Pain 1997;13:337–347.
33. Mueller HH, Donaldson CC, Nelson DV, Layman M. Treatment of fibromyalgia incorporating EEG-Driven stimulation: a clinical outcomes study. J Clin Psychol 2001;57: 933–952.
34. van Santen M, Bolwijn P, Verstappen F, Bakker C, Hidding A, Houben H, van der Heijde D, Landewe R, van der Linden S. A randomized clinical trial comparing fitness and biofeedback training versus basic treatment in patients with fibromyalgia. J Rheumatol 2002;29:575–581.
35. Drexler AR, Mur EJ, Gunther VC. Efficacy of an EMG-biofeedback therapy in fibromyalgia patients. A comparative study of patients with and without abnormality in (MMPI) psychological scales. Clin Exp Rheumatol 2002; 20:677–682.
36. Nolli M et al. Evaluation of chronic fibromyalgic pain before and after EMG-BFB training. Algos 1986;3:249–253.

37. Ferraccioli G et al. EMG-biofeedback training in fibromyalgia syndrome. J Rheumatol 1987;14:820–825.
38. Waylonis GW, Perkins RH. Post-traumatic fibromyalgia: A long-term follow-up. Am J Phys Med Rehabil 1994;73:403–412.
39. Buckelew SP et al. Self-efficacy predicting outcome among fibromyalgia patients. Arthritis Care Res 1996;9:97–104.
40. Sarnoch H, Adler F, Scholz OB. Relevance of muscular sensitivity, muscular activity, and cognitive variables for pain reduction associated with EMG biofeedback for fibromyalgia. Percept Mot Skills 1997;84:1043–1050.
41. Jorge JM, Habr-Gama A, Wexner SD. Biofeedback therapy in the colon and rectal practice. Appl Psychophysiol Biofeedback 2003;28:47–61.
42. Mason HJ, Serrano-Ikkos E, Kamm MA. Psychological state and quality of life in patients having behavioral treatment (biofeedback) for intractable constipation. Am J Gastroenterol 2002;97:3154–3159.
43. Benninga MA, Buller HA, Taminiau JA. Biofeedback training in chronic constipation. Arch Dis Child 1993;68:126–129.
44. Cox DJ, Sutphen J, Borowitz S, Dickens MN, Singles, Whitehead WE. Simple electromyographic biofeedback treatment for chronic pediatric constipation/encopresis: Preliminary report. Biofeedback Self Regul 1994;19:41–50.
45. Van der Plas RN et al. Biofeedback training in treatment of childhood constipation: A randomized controlled trial. Lancet North Am Ed 1996;348:776–780.
46. Leahy A, Clayman C, Mason I, Lloyd G, Epstein O. Computerized biofeedback games: a new method for teaching stress management and its use in irritable bowel syndrome. J R Coll Phys London 1998;32:552–556.
47. Radnitz CL, Blanchard EB. Bowel sound biofeedback as a treatment for irritable bowel syndrome. Biofeedback Self Regul 1988;13:169–179.
48. Radnitz CL, Blanchard EB. A 1- and 2-year follow-up study of bowel sound biofeedback as a treatment for irritable bowel syndrome. Biofeedback Self Regul 1989;14:333–338.
49. Burish TG, Jenkins RA. Effectiveness of biofeedback and relaxation training in reducing side effects of cancer chemotherapy. Health Psychol 1992;11:17–23.
50. Nakao M, Yano E, Nomura S, Kuboki T. Blood pressure-lowering effects of biofeedback treatment in hypertension: A meta-analysis of randomized controlled trials. Hypertens Res 2003;26:37–46.
51. Freedman RR. Long-term effectiveness of behavioral treatments for Raynaud's Disease. Behav Ther 1987;18:387–399.
52. Sedlacek K, Taub E. Biofeedback treatment of Raynaud's Disease. Prof Psychol Res Pr 1996;27:548–553.
53. Raynaud's Treatment Study Investigators. Comparison of sustained-release nifedipine and temperature biofeedback for treatment of primary Raynaud phenomenon. Results from a randomized clinical trial with 1-year follow-up. Arch Intern Med 2000;24:1101–1108.
54. Middaugh SJ, Haythornthwaite JA, Thompson B, Hill R, Brown KM, Freedman RR, Attanasio V, Jacob RG, Scheier M, Smith EA. The Raynaud's Treatment Study: biofeedback protocols and acquisition of temperature biofeedback skills. Appl Psychophysiol Biofeedback 2001;26:251–278.
55. Ramirez PM, DeSantis D, Opler LA. EEG biofeedback treatment of ADD: A viable alternative to traditional medical intervention? Adult Attention Deficit Disorder: Brain Mechanisms and Life Outcomes. In: Wasserstein J, et al., editors. Ann. N. Y. Acad. Sci. New York: New York Academy of Sciences; 2001.
56. Monastra VJ, Monastra DM, George S. The effects of stimulant therapy, EEG biofeedback, and parenting style on the primary symptoms of attention-deficit/hyperactivity disorder. Appl Psychophysiol Biofeedback 2002;27:231–249.
57. Luber JF. Electroencephalographic biofeedback methodology and the management of epilepsy. Integr Physiol Behav Sci 1998;33:1053–1088.
58. Peniston EG, Kulkosky PJ. Neurofeedback in the treatment of addictive disorders. In: Evans JR, Abarbanel A, editors. Introduction to Quantitative EEG and Neurofeedback. San Diego: Academic Press; 1999. p 157–179.
59. Taub E, Steiner SS, Weingarten E, Walton KG. Effectiveness of broad spectrum approaches to relapse prevention in severe alcoholism: A long-term, randomized, controlled trial of Transcendental Meditation, EMG biofeedback, and electronic neurotherapy. Alcohol Treat Q 1994;11:187–220.
60. Tansey MA. A simple and a complex tic (Gilles de la Tourette's syndrome): Their response to EEG sensorimotor rhythm biofeedback training. Int J Psychophysiol 1986;4:91–97.
61. Brody C, Davison ET, Brody J. Self-regulation of a complex ventricular arrhythmia. Psychosom: J Consult-Liaison Psychiat 1985;26:754–756.
62. Burgio KL, Locher JL, Goode PS, Hardin JM, McDowell BJ, Dombrowski M, Candib D. Behavioral vs. drug treatment for urge urinary incontinence: A randomized controlled trial. J Am Med Assoc 1998;280:1995–2000.
63. Jorge JMN, Habr-Gama A, Wexner SD. Biofeedback therapy in the colon and rectal practice. Appl Psychophysiol Biofeedback 2003;28:47–61.
64. Tries J, Brubaker L. Application of biofeedback in the treatment of urinary incontinence. Prof Psychol Res Pr 1996;27:554–560.
65. Norton C, Kamm MA. Anal sphincter biofeedback and pelvic floor exercises for faecal incontinence in adults—A systematic review. Aliment Pharmacol Ther 2001;15:1147–1154.
66. Rozelle GR, Budzynski TH. Neurotherapy for stroke rehabilitation: A single case study. Biofeedback Self Regul 1995;20:211–228.
67. Byers AP. Neurofeedback therapy for a mild head injury. J Neurother 1995;1:22–37.
68. Schleenbaker RE, Mainous AG. Electromyographic biofeedback for neuromuscular reeducation in the hemiplegic stroke patient—A meta-analysis. Arch Phys Med Rehabil 1993;74:1301–1304.
69. Glazer HI, Rodke G, Swencionis C, Hertz R, Young AW. Treatment of Vulvar Vestibulitis Syndrome with Electromyographic Biofeedback of Pelvic Floor Musculature. J Reprod Med 1995;40:283–290.
70. Bergeron S, Binik YM, Khalife S, Pagidas K, Glazer HI, Meana M, Amsel R. A randomized comparison of group cognitive-behavioral therapy, surface electromyographic biofeedback, and vestibulectomy in the treatment of dyspareunia resulting from vulvar vestibulitis. Pain 2001;91:297–306.
71. Burger C, Rough J. An EMG integrator for muscle activity studies in ambulatory subjects. IEEE Trans Biomed Eng 1983; 66–69.
72. Hatch JP, Prihoda TJ, Moore PJ, Cyr-Provost M, Borcherding S, Boutros NN, Seleshi E. A naturalistic study of the relationships among electromyographic activity, psychological stress, and pain in ambulatory tension-type headache patients and headache-free controls. Psychosom Med 1991;53:576–584.
73. Searle JR, Arena JG, Sherman RA. A portable activity monitor for musculoskeletal pain disorders. Proc Annu Int Conf IEEE Eng Med Biol Soc. 1989.

74. Arena JG, Bruno GM, Brucks AG, Searle JD, Sherman RA, Meador KJ. Reliability of an ambulatory electromyographic activity device for musculoskeletal pain disorders. Int J Psychophysiol 1994;17:153–157.
75. Arena JG, Bruno GM, Brucks AG, Searle JD, Sherman RA, Meador KJ (unpublished manuscript). The measurement of surface EMG in tension-headache subjects in the natural environment: Ambulatory recordings of data from five consecutive days.
76. Gatchel RJ, Oordt MS. Future trends and opportunities. In: Gatchel RJ, Oordt MS, editors. Clinical Health Psychology and Primary Care: Practical Advice and Clinical Guidance for Successful Collaboration. Washington, DC: American Psychological Association; 2003.
77. Blanchard EB, Andrasik F, Appelbaum KA, Evans DD, Jurish SE, Teders SJ, Rodichok LD, Barron KD. The efficacy and cost-effectiveness of minimal-therapist contact, nondrug treatments of chronic migraine and tension headache. Headache 1985a;25:214–220.
78. Blanchard EB, Appelbaum KA, Nicholson NL, Radnitz CL, Morrill B, Michultka D, Kirsch C, Hillhouse J, Dentinger MP. A controlled evaluation of the addition of cognitive therapy to a home-based biofeedback and relaxation treatment of vascular headache. Headache 1990a;30:371–376.
79. Jurish SE, Blanchard EB, Andrasik F, Teders SJ, Neff DF, Arena JG. Home versus clinic-based treatment of vascular headache. J Consult Clin Psychol 1983;51:743–751.
80. Tobin DL, Holroyd KA, Baker A, Reynolds RVC, Holms JE. Development and clinical trial of a minimal contact, cognitive-behavioral treatment for tension headache. Cognit Ther Res 1988;12:325–339.
81. Blanchard EB, McCoy GC, Musso A, Gerardi RJ, Cotch PA, Siracusa K, Andrasik F. A controlled comparison of thermal biofeedback and relaxation training in the treatment of essential hypertension: I. Short-term and long-term outcome. Behav Ther 1986;17:563–579.
82. Devineni T, Blanchard EB. A Randomized Controlled Trial of an Internet-based Treatment for Chronic Headache. Behav Res Ther 2005;43:277–292.
83. Arena JG, Dennis N, Devineni T, McClean R, Meador KJ. A pilot study of the feasibility of a telemedicine delivery system for psychophysiological treatments for vascular headache. Tele Meo J E-Health 2005;10:449–454.

See also BIOELECTRODES; ELECTROENCEPHALOGRAPHY; ELECTROGASTROGRAM; ELECTROMYOGRAPHY.

BIOHEAT TRANSFER

JONATHAN W. VALVANO
The University of Texas
Austin, Texas

INTRODUCTION

Bioheat transfer is the study of the transport of thermal energy in living systems. Because biochemical processes are temperature dependent, heat transfer plays a major role in living systems. Also, because the mass transport of blood through tissue causes a consequent thermal energy transfer, bioheat transfer methods are applicable for diagnostic and therapeutic applications involving either mass or heat transfer. This article presents the characteristics of bioheat transfer that distinguish it from nonliving systems, including the effects of blood perfusion on temperature distribution, coupling with biochemical processes, therapeutic and injury processes, and thermoregulation.

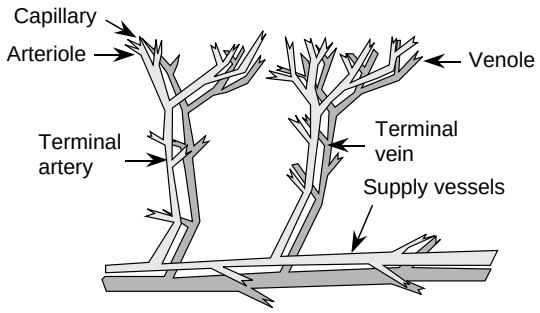


Figure 1. Countercurrent blood vessels have arterial blood flowing in the opposite direction as venous blood.

The study of bioheat transfer involves phenomena that are not found in systems that are not alive. For example, blood perfusion is considered a three-dimensional (3D) process as fluid traverses in a volumetric manner through tissues and organs via a complex network of branching vessels. Heat transfer is affected by vessel geometry, local blood flow rates, and thermal capacity of the blood (1). One factor that makes modeling blood perfusion difficult is the complex network of pairs of arteries and veins with countercurrent flow (2), as shown in Fig. 1. Arterial and venous blood temperatures may be different, and it is possible that neither is equal to the local tissue temperature. These temperatures may vary as a function of many transient physiological and physical parameters. The regulation of temperature and blood flow is quite nonlinear and has presented a major challenge to understand and model. Nevertheless, these critical processes must be accounted for in the design of many types of systems that interface with humans and animals.

Many scientists view life from either the macroscopic (systems) or the microscopic (cellular) level, but in reality one must be aware that life processes exist continuously throughout the spectrum. In order to better understand life processes at the molecular level, a significant research effort is underway associated with molecular biology. Because temperature and blood flow are critical factors, bioengineers are collaborating with molecular biologists to understand and manipulate the molecular and biochemical processes that constitute the basis of life. Research has found that the rates of nearly all physiological functions are altered 6–10%/°C (3). Similarly, heat can be added or removed during therapeutic or diagnostic procedures to produce or measure a targeted effect, based on the fact that a change in local temperature will have a large effect on rates of biochemical process rates. Thus, the measurement and control of temperature in living tissues is of great value in both the assessment of normal physiological function and the treatment of pathological states.

The study of the effects of temperature alterations on biochemical rate processes has been divided into three broad categories: hyperthermia (increased temperature), hypothermia (decreased temperature), and cryobiology

(subfreezing temperature). An extensive review of these domains has been published (4), to which the reader is referred for further details and bibliography.

Effects of Blood Perfusion on Heat Transfer

Blood perfusion through the vascular network and the local temperature distribution are interdependent. Many environmental (e.g., heat stress and hypothermia), pathophysiologic (e.g., inflammation and cancer), therapeutic (e.g., heating–cooling pads) situations create a significant temperature difference between the blood and the tissue through which it flows. The temperature difference causes convective heat transport to occur, altering the temperatures of both the blood and the tissue. Perfusion-based heat transfer interaction is critical to a number of physiological processes, such as thermoregulation and inflammation.

The convective heat transfer depends on the rate of perfusion and the vascular anatomy, which vary widely among the different tissues, organs of the body, and pathology. Diller et al. published an extensive compilation of perfusion data for many tissues and organs and for many species (5). Charney reviewed the literature on mathematical modeling of the influence of blood perfusion on bioheat transfer phenomena (6).

The rate of perfusion of blood through different tissues and organs varies over the time course of a normal day's activities, depending on factors, such as physical activity, physiological stimulus, and environmental conditions. Further, many disease processes are characterized by alterations in blood perfusion, and some therapeutic interventions result in either an increase or decrease in blood flow in a target tissue. For these reasons, it is very useful in a clinical context to know what the absolute level of blood perfusion is within a given tissue. Many thermal techniques have been developed that directly measure heat flux to predict blood perfusion by exploiting the coupling between vascular perfusion and local tissue temperature using inverse mathematical solutions.

In 1948, Pennes (7) published the seminal work describing the mathematical coupling between the mass transfer of blood perfusion and the thermal heat transfer. His work consisted of a series of experiments to measure temperature distribution as a function of radial position in the forearms of nine human subjects. A butt-junction thermocouple was passed completely through the arm via a needle inserted as a temporary track, with the two leads exiting on opposite sides of the arm. The subjects were unanesthetized so as to avoid the effects of anesthesia on blood perfusion. Following a period of normalization, the thermocouple was scanned transversely across the mediolateral axis to measure the temperature as a function of radial position within the interior of the arm. The environment in the experimental suite was kept thermally neutral during experiments. Pennes' data showed a temperature difference of 3–4° between the skin and the interior of the arm, which he attributed to the effects of metabolic heat generation and heat transfer with arterial blood perfused through the microvasculature.

Pennes proposed a model to describe the effects of metabolism and blood perfusion on the energy balance within tissue. These two effects were incorporated into the standard thermal diffusion equation, which is written in its simplified form as

$$\rho c \frac{\partial T}{\partial t} = \nabla \cdot k \nabla T + \rho_{\mathrm{bl}} c_{\mathrm{bl}} w (T_a - T) + Q_{\mathrm{met}} \qquad (1)$$

Metabolic heat generation, Q_{met}, is assumed to be homogeneously distributed throughout the tissue of interest as rate of energy deposition per unit volume. It is assumed that the blood perfusion effect is homogeneous and isotropic, and that thermal equilibration occurs in the microcirculatory capillary bed. In this scenario, blood enters capillaries at the temperature of arterial blood, T_a, where heat exchange occurs to bring the temperature to that of the surrounding tissue, T. There is assumed to be no energy transfer either before or after the blood passes through the capillaries, so that the temperature at which it enters the venous circulation is that of the local tissue. The total energy exchange between blood and tissue is directly proportional to the density, ρ_{bl}, specific heat, c_{bl}, and perfusion rate, w, of blood through the tissue. The basic principle that couples mass transfer to heat transfer is the change in sensible energy caused by the moving blood. The units of perfusion in equation 1 are volume of blood per volume of tissue per time (s^{-1}). This thermal transport model is analogous to the process of mass transport between blood and tissue, which is confined primarily to the capillary bed.

A major advantage of the Pennes model is that the added term to account for perfusion heat transfer is linear in temperature, which facilitates the solution of Eq. 1. Since the publication of this work, the Pennes model has been adapted by many researchers for the analysis of a variety of bioheat transfer phenomena. These applications vary in physiological complexity from a simple homogeneous volume of tissue to thermal regulation of the entire human body (8,9). As more scientists have evaluated the Pennes model for application in specific physiological systems, it has become increasingly clear that many of the assumptions to the model are not valid. For example, it is now well established that the significant heat transfer due to blood flow occurs in the terminal arterioles (vessels 60–300 μm in diameter) (10–17). Thermal equilibration is essentially complete for vessels <60 μm (precapillaries and capillaries). Therefore, no significant heat transfer occurs in the capillary bed; the exchange of heat occurs in the larger components of the vascular tree. The vascular morphology varies considerably among the various organs of the body, which contributes to the need for specific models for the thermal effects of blood flow (as compared to the Pennes model that incorporates no information concerning vascular geometry). It would appear as a consequence of these physiological realities that the validity of the Pennes model is questionable.

Many investigators have developed alternative models for the exchange of heat between blood and tissue. These models have accounted for the effects of vessel size, countercurrent heat exchange, as well as a combination of partial countercurrent exchange and bleed-off perfusion. All of these models provided a larger degree of rigor in the analysis, but at the compromise of greater complexity and

reduced generality. These studies also led to an increased appreciation of the necessity for a more explicit understanding of the local vascular morphology as it governs bioheat transfer, which has given rise to experimental studies to measure and characterize the 3D architecture of the vasculature in tissues and organs of interest (18). The quantitative analysis of the effects of blood perfusion on the internal temperature distribution in living tissue remains a topic of active research after one-half of a century of study (19).

THERAPEUTIC APPLICATIONS OF BIOHEAT TRANSFER

The elevation of tissue temperature into the 40–42 °C range provides some relief from pain (analgesia). In addition, wound healing can be enhanced by this modest increase in temperature. Increased temperature does not cause healing by itself, but rather it creates the improved conditions for the natural processes to heal wounds.

Hot or cold packs can be used to create therapeutic heating for injuries near the skin surface. Heating packs are effective for injuries such as sprains, muscle strains, and postoperative swelling. Elevated temperatures cause an increase in blood perfusion, supplying nutrients to the injured tissue. During the first 12–24 h after injury, cold packs will reduce perfusion, thereby reducing vascular pressure and tissue swelling. Afterward, hot packs (up to 45 °C) are applied to increase perfusion and promote healing (20).

When the injury is deeper, the therapy requires a volumetric heater, such as radio frequency (rf) electromagnetic heating, microwave frequency electromagnetic heating, and ultrasonic heating. The Industrial-Medical-Scientific (ISM) frequencies are 6.78, 13.56, 27.12, and 40.68 MHz. The ISM frequencies typically used in medical applications of microwaves are 915 MHz and 2.45 GHz. Medical ultrasonic devices operate in the 500 kHz to 10 MHz range. There are three engineering parameters to consider when designing a therapeutic device. The first parameter is the amount of local volumetric heat generation, the second parameter is the shape of the heating field, and the third parameter is the depth of penetration. Higher frequencies have shorter wavelengths, causing higher absorption and less penetration. The electrical properties of the tissue, which depend on structure and composition, strongly affect the effectiveness of volumetric heating using rf and microwave EM fields. Similarly, acoustic properties of the tissue are important in ultrasonic systems. Often the design of an effective therapeutic device hinges on proper control of the boundary conditions where energy is transferred across the transducer–tissue interface.

There is a systemic effect of local heating, controlled by the hypothalamus, involving both neuronal and hormonal signals. Local heating of organs or peripheral muscles can also cause a spinal cord mediated response. A local release of bradykinins can affect the vascular tone of the terminal arterioles (see Fig. 1), which in turn affect the vascular resistance to blood flow. In general, an increase in local temperature causes an increase in local blood flow, whereas a decrease in local temperature creates a decrease in local blood flow, but the behavior is highly complex.

Smooth muscles surrounding the 40–200 μm diameter arterioles play a dominate role in controlling local blood flow. Normal capillary pressure is ~3.3 kPa (25 Torr). When local tissues are heated, these arterioles dilate causing an increase in capillary pressure and capillary blood flow. Edema occurs when this pressure widens gaps in the capillary wall causing excess fluid to leak from the vascular to intravascular space. High capillary flow promotes healing by removing wastes, delivering nutrients, and supplying oxygen. Leukocytes (white blood cells) control the healing process by first breaking down, then removing damaged and dead tissue.

The volumetric heating created by electromagnetic fields in governed by the electrical conductivity, σ ($S \cdot m^{-1}$), the imaginary part of the electrical permittivity, ε'' ($F \cdot m^{-1}$), and the magnitude of the local electric field, $|\boldsymbol{E}|$ ($V \cdot m^{-1}$):

$$q''' = (\sigma + \omega\varepsilon'')|\boldsymbol{E}|^2 \tag{2}$$

where ω is the angular frequency of the field in rad s^{-1}. Direct heating from the magnetic fields in medical applications is usually neglected. Magnetic fields can be used to heat tissue, but because of Faraday's law of induction, the time-varying magnetic field will induce an electric field, and it is this electric field that heats the tissue. A comprehensive review of electromagnetic heating can be found in Roussy and Pearce (21). Tables of electrical and acoustic properties can be found in Diller (5).

THERMOREGULATION

Thermoregulation is an elaborate control system, used by mammals, to maintain internal body temperatures near a physiological set point under a large spectrum of environmental conditions and metabolic rate activities. Even though there have been many years of research, much remains unknown about the human thermoregulatory system. Therefore, active investigation continues. Heat transfer due to conduction, convective heat transfer via the blood flow, local generation of thermal energy, and thermal boundary conditions comprise the major components of thermoregulation. Once these individual mechanisms are understood, they can be combined to create mathematical models to simulate and predict thermoregulatory behavior. The mathematical models are used to design systems to interact thermally with the human body (such as a space suit) without compromising the health and safety of the subject.

The prevailing theory is that the main objective of a human thermoregulation system is to maintain the body core temperature at a constant value consistent with that required for normal physiological functions, regardless of the environmental conditions. An alternative theory, suggested by Chappuis et al. (22) and Webb (23), is that the goal of the human regulation system is to maintain the body's energy balance. In this theory, tissue temperatures are a result, rather than a cause, of the regulation process.

Nunneley (24,25) showed that temperature and internal energy storage of the human body vary with time of day, metabolic activity, and individuality of the human. To maintain body core temperature, the thermoregulatory system incorporates a number of energy production and dissipation mechanisms, many of them controlled by feedback from other body parameters. Examples of such feedback control are that for sweating, shivering, and varying blood flow.

Ganong (26) showed that the main control center for feedback mechanisms is located in the hypothalamus of the brain, where reflex responses operate to maintain the body temperature within its narrow range. The signals that activate the hypothalamic temperature regulating centers come largely from two sources: the temperature-sensitive cells in the anterior hypothalamus and cutaneous temperature receptors. The cells in the anterior hypothalamus sense the temperature of the body core or, specifically, the temperature of the arterial blood that passes through the head.

The theory of energy regulation is based on the demonstration of the existence of temperature sensors at several levels in the skin enabling the sensing of heat flow within and from the body. Evidence has also shown neurological sensing of thermal gradients, of which changes relate to the thermal regulating responses. Therefore, the hypothesis behind the theory of energy content regulation based on Webb's experimental observations is "Heat (energy) regulation achieves heat (energy) balance over a wide range of heat (energy) loads. Heat flow to or from the body is sensed, and physiological responses defend the body heat (energy) content. Heat (energy) content varies over a range that is related to heat (energy) load. Changes in body heat (energy) content drive deep body temperatures" (23). Energy regulation involves constantly changing metabolic energy production and the adjustment of heat losses to maintain a system in equilibrium. This mechanism is opposed to temperature regulation where adjustments are required to maintain body temperature.

The thermal energy balance over time within the human body combines the heat added by internal production minus the heat lost by various heat-transfer processes.

$$\Delta E = M - (W + Q_{\text{conv}} + Q_{\text{cond}} + Q_{\text{rad}} + Q_{\text{evap}} + Q_{\text{resp}}) \qquad (3)$$

where ΔE is the rate of energy storage in the body (W), M is the metabolic energy production (W), W is the external work (W), Q_{conv} is the surface heat loss by convection (W), Q_{cond} is the surface heat loss by conduction (W), Q_{rad} is the surface heat loss by radiation (W), Q_{evap} is the surface heat loss by evaporation (W), and Q_{resp} is the respiratory heat loss (W).

The human body produces energy, exchanges heat with the environment, and loses heat by evaporation of body fluids. Thermal energy is produced by metabolism, a biochemical process occurring in cells has adenosine triphosphate (ATP) is combined with oxygen to produce the various life functions. Fulcher (27) defined the basal metabolic rate as "the minimal metabolism measured at a temperature of thermal neutrality in a resting homeotherm with normal body temperature several hours after a meal and not immediately after hypothermia". Energy is also produced at an increased rate due to muscle activity, including physical exercise and shivering, and by food intake. Therefore, the total energy production in the body is determined by the energy needed for basic body processes plus any external work. Since the body operates with $<100\%$ efficiency only a fraction of the metabolic rate is applied to work. The remainder shows up as heat. The mechanical efficiency associated with metabolic energy utilization is zero for most activities except when the person is performing external mechanical work, such as in walking upstairs, lifting something to a higher level, or cycling on an exercise machine. When external work is dissipated into heat in the human body, the mechanical efficiency is negative. An example of negative mechanical efficiency is walking downstairs.

Convection, radiation, conduction, and evaporation of sweat at the skin surface allow heat to be dissipated from the body. There is also heat transfer, especially when the environmental air temperature is extremely high or low, through the respiratory tract and lungs. Storage of energy takes place whenever there is an imbalance of production and dissipation mechanisms. In many instances, such as astronauts in space suits or military personnel in chemical defense garments, energy storage is forced due to the lack of appropriate heat exchange with the environment.

The human thermoregulatory system is quite complicated and behaves mathematically in a highly nonlinear manner. It contains multiple sensors, multiple feedback loops, and multiple outputs. The control mechanisms to release excess energy include the production of sweat, and vasodilatation of the blood vessels in the skin. Conversely, to conserve energy there can be shivering of the muscles, and vasoconstriction of blood vessels, which engage in the transportation of heat to the surface of the body.

Heat transfer within the body is due to the internal conductance that governs the flow of heat from the core, through the tissue, to the surface. This component of heat transfer is governed by peripheral blood flow, the core-skin temperature gradient, and the conductivity of the body tissue. Blood flow provides the majority of the peripheral conductance where there is convection between blood and tissue and countercurrent heat exchange between the arteries and the veins. Blood flow is controlled according to metabolic needs of the body as well as the need to maintain the appropriate core temperature. When the core becomes too hot, the blood vessels in the skin dilate to allow increased blood flow to the surface of the skin. Then, the environment cools the blood and the cooler blood returns to the core. Increased blood flow to the skin surface increases extravascular pressure enabling greater sweat production, again adding to the cooling process. In contrast, when the core becomes too cold, blood flow to the skin is constricted to conserve the body's internal energy.

Sweating is centrally controlled by the hypothalamus. When the body senses an increase temperature the hypothalamus increases nerve impulses to the sweat glands. Shivering, on the other hand, is an involuntary response of the skeletal muscles when passive body cooling exceeds metabolic energy production.

This section briefly introduced the concepts of thermoregulation. For a quantitative analysis of this topic, see Wissler (8,9).

THERMAL INJURY

Thermal injury is defined as irreversible changes to living tissue caused by temperature. Injury can occur when the tissue temperature exceeds the range between which normal life processes exist. Both high and low temperature can cause irreversible changes to biomolecules, resulting in injury. Common examples are burns and frostbite. Recently, it has been discovered that under some kinds of moderate thermal stress that is subthreshold to injury, cells produce molecules that render temporary protection against levels of many types of stress (thermal, mechanical, chemical, etc.) that would normally cause injury. These protective molecules are called heat shock proteins, and they are the subject of widespread investigation to identify the kinetics of their expression and function. It is an effort to develop applications in which they may be induced either before or even after a traumatic event.

The most commonly encountered type of thermal injury is the burn. Accidental burns are encountered most frequently in domestic and industrial settings as well as many other venues of activity. Most burns result from the propagation of heat inward into tissues as a result of contact at the surface (skin) with a hot solid, liquid, or vapor. One exception is electrical burns in which the tissue temperature is elevated owing to I^2R dissipation of electric energy when a voltage is applied. In this case, the primary source of heating is internal since the impedance of muscle is higher than of skin and fat.

It is generally assumed that thermal burns can be modeled as a simple Arrhenius rate process such that

$$\Omega(t) = \int_0^t A e^{\Delta E/RT(\tau)} d\tau \tag{4}$$

where Ω is a dimensionless damage parameter (e.g., $\Omega = 1$ means first degree burn), A is a frequency factor (s^{-1}), ΔE is the activation energy in $J \cdot mol^{-1}$, R is the universal gas constant ($8.314\,J \cdot mol^{-1} \cdot K^{-1}$), and T is the tissue temperature (K). The constants A and ΔE are tissue parameters, and $T(\tau)$ is the time history of the tissue temperature (28).

This model was first posed for predicting the severity of a burn as a function of the temperature and time of exposure at the skin surface by Moritz and Henriques (29) shortly after World War II. They also performed experiments to determine threshold conditions for eliciting first and second degree burns in humans and applied this data to determine values for the scaling constant and activation energy in their Arrhenius model. Their experiments were conducted at temperatures between 44 and 70°C and exposure times between 1 and 25,000 s. The model parameter values are $A = 3.1 \times 10^{98}\,s^{-1}$ and $\Delta E = 6.28 \times 10^5\,kJ \cdot mol^{-1}$. Over the ensuing 50 years many subsequent investigators have studied this process with mathematical models and experimental investigations (30–40). Although a considerable body of literature has been accrued, there is by no means a consensus on how to accurately predict the occurrence of thermal injury over the wide range of conditions that cause burns.

SUBZERO EFFECTS

One application of subzero temperatures is the long-term preservation of biologic tissue. Therapeutic devices based on subzero temperatures can be used to destroy cancerous cells or remove necrotic tissue. The rate of biochemical processes is governed by local temperature. Lowering the temperature has the effect of reducing reaction rates, and at sufficiently low temperatures, a state of suspended animation can be achieved. Because of the water component of physiological fluids, temperatures low enough to affect suspended animation normally result in freezing. The freezing of native biomaterials is nearly always lethal to the affected tissue upon thawing. The formation of ice has two damaging effects. The first effect is mechanical as intracellular ice crystals physically damage cell structures. The second and more lethal effect is osmotic. The local concentration of ions, such as Na^+, K^+, and Cl^-, are critical for sustaining life, and a high concentration of these ions is produced as liquid water freezes into ice. The effected injury can be used to benefit in cryosurgery for the purpose of destroying a target tissue, such as cancer. Alternatively, the tissue can be modified prior to freezing by the introduction of a chemical cryoprotective agent (CPA) to afford protection from freeze–thaw injury. The CPA either protects against the injurious effects of ice formation or blocks the formation of ice so that a glassy state results, which is called vitrification. Organ transplantation, blood banks, and animal husbandry are three applications that require the successful long-term cryopreservation of biologic tissue. The response of living biomaterials to freezing and thawing is intimately tied to the thermal history during processing, especially at subzero temperatures. Thus, bioheat transfer analysis has played a key role in the design and development of effective cryopreservation techniques. Polge was first to report the successful use of glycerol to freeze fowl sperm >55 years ago (41). Successes were reported in succession for other types of tissues having rather simple cell structures, such as erythrocytes, gametes and various cells obtained from primary cultures (42–44). Most of these cryopreservation techniques were derived via largely empirical methods, and starting in the 1970s it came to be realized that the cryopreservation of more complex systems, such as multicellular tissues and whole organs require a more rigorous scientific understanding of the mechanisms of the governing biophysical processes and cellular response to freezing and thawing. Since that time engineers have made significant contributions to the developing science of cryobiology, not the least of which has been to identify some of the key biophysical problems to be solved (45,46).

MEASUREMENT OF THERMAL CONDUCTIVITY AND THERMAL DIFFUSIVITY

While the other sections in this article presented brief overviews of the various disciplines within the field of

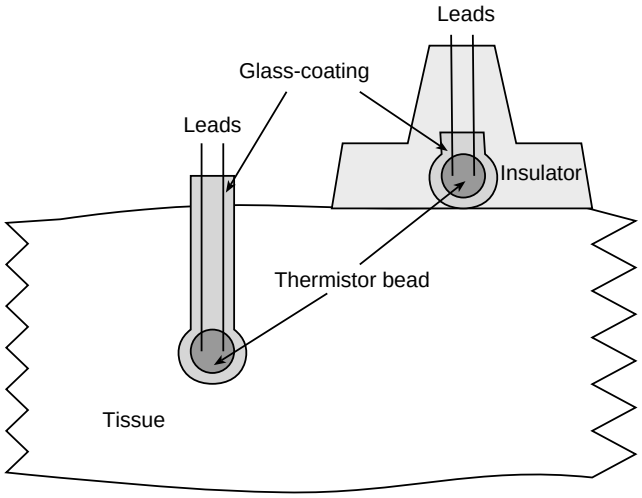


Figure 2. A glass-coated thermistor is placed into or on the surface of the tissue of interest.

bioheat transfer, this section will present in detail a specific measurement technique. In particular, this section presents an instrument used to measure thermal properties in tissue. The section begins with definitions of thermal properties, overviews of the technique, then develops the heat-transfer equations that form the basis of the instrument. Finally, calibration methods and error analyses are presented.

Definitions of Thermal Properties

Thermal conductivity (k) is the ability of a material to transport heat in the steady state. In one dimension, the total heat (Q) transported across a flat surface of area A and thickness Δx is related to the temperature gradient across the surface (ΔT) and the thermal conductivity of the material.

$$Q = -kA\frac{\Delta T}{\Delta x} \tag{5}$$

Thermal diffusivity (α) is the ability of a material to conduct heat in the transient state. Thermal properties of conductivity and diffusivity are related. The quotient of conductivity divided by diffusivity equals density times specific heat.

$$\frac{k}{\alpha} = \rho c \tag{6}$$

Diffusivity is often defined in the partial differential equation used to describe transient heat transfer. Assuming homogeneous thermal properties, the Fourier conduction equation in one dimension is

$$\frac{\partial^2 T}{\partial x^2} = \frac{1}{\alpha}\frac{\partial T}{\partial t} \tag{7}$$

Measurement Technique

The technique involves inserting a thermistor into the tissue of interest or placing it on the tissue surface, as shown in Fig. 2. Thermometrics P60DA102M and Fenwal 121-102EAJ-Q01 are glass probe thermistors that make excellent transducers (shown on the left of Fig. 2). The diameter of these thermistors is ~0.15 cm. The glass-coated spherical probes provide a large bead size and a rugged, stable transducer. The Thermometrics BR55KA102M and Fenwal 112-102EAJ-B01 bead thermistors also provide excellent results (shown on the right of Fig. 2).

If the tissue is living, the properties measured are called effective thermal conductivity, k_{eff}, and effective thermal diffusivity, α_{eff}. Effective thermal properties include the contribution to heat transfer due to intrinsic conduction added to the contribution caused by the transport of blood through the tissue.

In the constant temperature heating technique (47–52), the instrument first measures the baseline tissue temperature, T_s. Then, an electronic feedback circuit applies a variable voltage, $V_0(t)$, in order to maintain the average thermistor temperature at a predefined constant, T_h. The electrical circuit used to implement the constant temperature heating technique is shown in Fig. 3. Three high quality, gold-plated, electromagnetic relays are used to switch the thermistor (R_s) between "heat" and "sense" mode. Figure 3 shows the position of the three relays in "heat" mode. Initially, the instrument places the circuit in "sense" mode with the three relays in

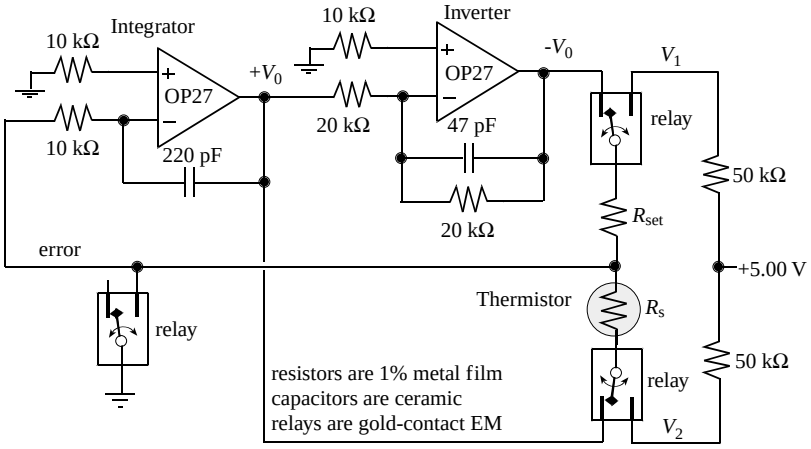


Figure 3. Instrumentation used for the constant temperature heating technique.

the opposite position as shown in Fig. 3. A precision +5.00 V reference (PMI REF02) supplies voltage to the four-resistor bridge, formed by the two 50 kΩ, R_{set}, and R_s resistors. The voltage difference $V_2 - V_1$ is fed to a differential amplifier, passed through a low pass filter, then fed to a 12-bit ADC.

Fundamental Equations

Resistance calibration is performed to determine the relationship between the ADC sample and the unknown R_s. Next, temperature calibration is performed by placing the thermistor adjacent to an accurate temperature monitor and placing the combination in a temperature- controlled waterbath. The thermistor resistance varies nonlinearly with its temperature. For small temperature ranges equation 8 can be used for temperature calibration.

$$R_s = R_{\text{O}}\, \text{e}^{\beta/(T_s+273.15)} \tag{8}$$

where T_s is the temperature in degrees Celsius, and R_s is the thermistor resistance in ohms.

In heat mode, the integrator–inverter circuit varies the voltage across the thermistor until the thermistor resistance, R_s, matches the fixed resistor, R_{set}. It takes just a few milliseconds for the electrical control circuit to stabilize. Once stable, R_s is equal to R_{set}, meaning the volume average thermistor temperature is equal to a constant. The instrument uses a calibration temperature versus resistance curve to determine the heated temperature T_h from the fixed resistor R_{set}. The power applied to the thermistor, P, is calculated from $(V_0)^2/R_{\text{set}}$. The applied thermistor power includes a steady state and a transient term:

$$P(t) = A + Bt^{-1/2} \tag{9}$$

In order to measure thermal conductivity, thermal diffusivity, and tissue perfusion the relationship between applied thermistor power, P, and resulting thermistor temperature rise, $\Delta T(t) = T_h - T_s$, must be known. In the constant temperature method, ΔT is constant. The thermistor bead is treated as a sphere of radius a embedded in a homogeneous medium. Since all media are considered to have constant parameters with respect to time and space, the initial temperature will be uniform when no power is supplied to the probe.

$$T_b = T_m = T_s = T_a + \frac{Q_{\text{met}}}{w\rho_{\text{bl}}c_{\text{bl}}} \qquad \text{at} \quad t = 0 \tag{10}$$

Let V be the temperature rise above baseline, $V = T - T_s$. Both the thermistor bead temperature rise (V_b) and the tissue temperature rise (V_m) are initially zero.

$$V_b = V_m = 0 \qquad \text{at} \quad t = 0 \tag{11}$$

To solve this coupled thermistor–tissue system, equation 7 is written in spherical coordinates and the applied power is deposited into the thermistor, while the perfusion heat sink is added to the tissue, equation 1. Assuming the venous blood temperature equilibrates with the tissue temperature and that the metabolic heat is uniform in time and space, the Pennes' bioheat transfer equation in spherical coordinates is given by

$$\rho_b c_b \frac{\partial V_b}{\partial t} = k_b \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial V_b}{\partial r}\right) + \frac{A + Bt^{-1/2}}{4/3\pi a^3} \qquad r < a \tag{12}$$

$$\rho_m c_m \frac{\partial V_m}{\partial t} = k_m \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial V_m}{\partial r}\right) - w\rho_{bl}c_{bl}V_m \qquad r > a \tag{13}$$

where w is the tissue perfusion (s^{-1}). Perfect thermal contact is assumed between the finite-sized spherical thermistor and the infinite homogeneous perfused tissue. At the interface between the bead and the tissue, continuity of thermal flux and temperature leads to the following boundary conditions:

$$V_b = V_m \qquad \text{at } r = a \tag{14}$$

$$k_b \frac{\partial V_b}{\partial r} = k_m \frac{\partial V_m}{\partial r} \qquad \text{at } r = a \tag{15}$$

The other boundary conditions are necessary at positions $r \to 0$ and $r \to$ infinity. Since no heat is gained or lost at the center of the thermistor:

$$k_b \frac{\partial V_b}{\partial r} = 0 \qquad \text{as } r \to 0 \tag{16}$$

Because the thermistor power is finite and the tissue is infinite, the tissue temperature rise at infinity goes to zero:

$$V_m \to 0 \qquad \text{as } r \to \text{infinity} \tag{17}$$

It is this last initial condition that allows the Laplace transform to be used to solve the coupled partial differential equations. The Laplace transform converts the partial differential equations into ordinary differential equations that are independent of time t. The steady-state solution allows for the determination of thermal conductivity and perfusion (49).

$$V_b(r) = \frac{A}{4\pi a\, k_b}\left(\frac{k_b}{k_m(1+\sqrt{z})} + \frac{1}{2}\left[1 - \left(\frac{r}{a}\right)^2\right]\right) \tag{18}$$

$$V_m(r) = \frac{A}{4\pi r\, k_m}\left(\frac{e^{(1-r/a)\sqrt{z}}}{1+\sqrt{z}}\right) \tag{19}$$

where z is a dimensionless Pennes' model perfusion term ($w\rho_{\text{bl}}\, c_{\text{bl}}\, a^2/k_m$). The measured thermistor response, ΔT, is assumed be the simple volume average of the thermistor temperature:

$$\Delta T = \frac{\int_0^a V_b(r) 4\pi r^2 dr}{4/3\pi a^3} \tag{20}$$

Inserting equation 18 into Eq. 20 yields the relationship used to measure thermal conductivity assuming no perfusion (49).

$$k_m = \frac{1}{\frac{4\pi a\, \Delta T}{A} - \frac{0.2}{k_b}} \tag{21}$$

A similar equation allows the measurement of thermal diffusivity from the transient response, again assuming no

perfusion (49).

$$\alpha_m = \left(\frac{a}{\sqrt{\pi}\,\frac{B}{A}\left(1 + 0.2\,\frac{k_m}{k_b}\right)} \right)^2 \tag{22}$$

Calibration Equations

The first calibration determines relationship between the ADC sample and the thermistor resistance when in sense mode. For this calibration, precision resistors are connected in place of the thermistor, and the computer-based instrument is used to sample the ADC in sense mode. A simple linear equation works well for converting ADC samples to measured resistance. In this procedure, the device acts like a standard ohmmeter.

The second calibration determines the relationship between thermistor temperature and its resistance. The instrument measures resistance, and a precision thermometer determines true temperature. Equation 23 yields an accurate fit over a wide range of temperature:

$$T = \frac{1}{H_0 + H_1 \ln(R) + H_3 [\ln(R)]^3} - 273.15 \tag{23}$$

where T is in degrees Celsius. Temperature resistance data are fit to Eq. 23 using nonlinear regression to determine the calibration coefficients H_0, H_1, and H_3.

The applied power, $P(t)$, is measured during a 30 s transient while in heat mode. Nonlinear regression is used to calculate the steady-state and transient terms in equation 9. Figure 4 shows some typical responses. The steady-state response (time equals infinity) is a measure of the thermal conductivity. The transient response (slope) indicated the thermal diffusivity.

The third calibration maps measured power to thermal properties while operating in heat mode. Rather than using the actual probe radius (a) and probe thermal conductivity (k_b), as shown in Eqs. 21 and 22, the following empirical equations are used to calculate thermal properties.

$$k_m = \frac{1}{(c_1\,\Delta T/A) + c_2} \tag{24}$$

$$\alpha_m = \left(\frac{c_3}{B/A(1 + k_m/c_4)} \right)^2 \tag{25}$$

The coefficients c_1, c_2, c_3, and c_4 are determined by operating the probe in two materials of known thermal properties. Typically, agar-gelled water and glycerol are used as thermal standards. This empirical calibration is performed at the same temperatures at which the thermal property measurements will be performed.

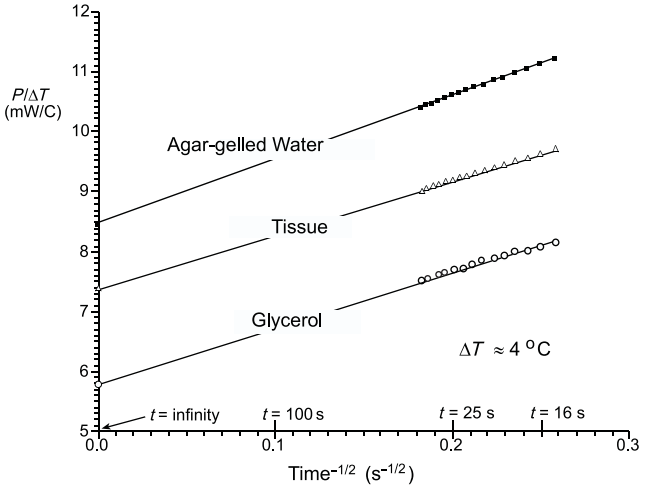


Figure 4. Typical $P/\Delta T$ versus $t^{-1/2}$ data for the constant temperature heating technique. The agar-gelled water and glycerol curves are used for empirical calibration.

Error Analysis

It is assumed that the baseline tissue temperature, T_0, is constant during the 30 s transient. Patel has shown that if the temperature drift, dT_0/dt, is $>0.002\,^{\circ}\mathrm{C}\cdot\mathrm{s}^{-1}$, then significant errors will occur (52). The electronic feedback circuit forces T_h to a constant. Thus, if T_0 is constant then ΔT does not vary during the 30 s transient.

The time of heating can vary from 10 to 60 s. Shorter heating times are better for small tissue samples and for situations where there is baseline tissue temperature drift. Another advantage of shorter heating times is the reduction in the total time required to make one measurement. Longer heating times increase the measurement volume and reduce the effect of imperfect thermistor–tissue coupling. Typically, shorter heating times are used *in vivo* because it allows more measurements to be taken over the same time period. On the other hand, longer heating times are used *in vitro* because accuracy is more important than measurement speed.

Thermal probes must be constructed in order to measure thermal properties. The two important factors for the thermal probe are thermal contact and transducer sensitivity. The shape of the probe should be chosen in order to minimize trauma during insertion. Any boundary layer between the thermistor and the tissue of interest will cause a significant measurement error. The second factor is transducer sensitivity that is the slope of the thermistor voltage versus tissue thermal conductivity. Equation 21 shows for a fixed ΔT k_m and k_b the thermistor power (A) increases linearly with probe size (a). Therefore larger probes are more sensitive to thermal conductivity. For large tissue samples, multiple thermistors can be wired in parallel, so they act electrically and thermally as one large device. There are two advantages to using multiple thermistors. The effective radius, $a = c_1/4\pi$, is increased from $\sim$0.08 cm for a typical single P60DA102M probe to $\sim$0.5 cm for a configuration of three P60DA102M thermistors. The second advantage is that the three thermistors are close enough to each other that the tissue between the probes will be heated by all three thermistors. This cooperative heating tends to increase the effective measurement volume and reduce the probe/tissue contact error. Good mechanical–thermal contact is critical. The probes are calibrated after they are constructed, so that the thermistor geometry is incorporated into the coefficients

c_1, c_2, c_3, and c_4. The same waterbath, and probe configuration should be used during the calibration and during the tissue measurements.

Calibration is a critical factor when using an empirical technique. For temperatures $<0\,^{\circ}$C, ice and ethylene glycol are used as thermal standards. For temperatures between 0 and 15 °C, agar-gelled water and ethylene glycol can be used as thermal standards. For temperatures between 15 and 75 °C, agar-gelled water and glycerol were used. To prevent convection, 1 g of agar/100 mL of water should be added. A mixture of water and glycerol can be used to estimate the accuracy of the technique. The mass fraction, m, can be used to determine the true thermal properties of the mixture (53,54). The ability to determine measurement accuracy is critical for the acceptance of new technology. These two equations provide for the capability to create reference materials of known thermal properties, which can be used to experimentally determine measurement accuracy.

$$k_m = m\,k_g + (1-m)k_w + 1.4\,m(m-1)(k_w - k_g - 2) - 0.014\,m(m-1)(T - 20\,^{\circ}\mathrm{C}) \quad (26)$$

$$\alpha_m = m\,\alpha_g + (1-m)\alpha_w \quad (27)$$

where T is in degrees Celsius. Self-heat thermistors have also been successfully used to measure the convective heat transfer coefficient on the endocardial surface of the heart (55,56).

ADDITIONAL STUDIES

In this article, the general concepts of bioheat transfer were introduced, and a detailed design and analysis of an instrument that measures thermal properties was presented. Although out of print, the 1985 book *Heat Transfer in Medicine and Biology*, is a wonderful collection of detailed works that address a wide spectrum of topics in bioheat transfer. The book *Optical-Thermal Response of Laser Irradiated Tissue* covers the issues involved in high temperature effect such as tissue damage and thermal ablation. Valvano's chapter titled Temperature Measurements, in *Advances In Heat Transfer: Bioengineering Heat Transfer*, covers many practical issues involved in measuring temperature in the biomedical setting. An in depth treatment of bioheat transfer topics can be found in the 2005 edition of *CRC Handbook of Heat Transfer*. This reference has excellent treatments of thermoregulation and low temperature effects.

BIBLIOGRAPHY

Cited References

1. Baish JW, Ayyaswamy PS, Foster KR. Heat transport mechanisms in vascular tissues: a model comparison. J Biomech Eng 1986;108:324–331.
2. Baish JW. Heat transport by countercurrent blood vessels in the presence of an arbitrary temperature gradient. J Biomech Eng 1990;112:207–211.
3. Johnston KA, Bennett AF, editors. Animals and Temperature: Phenotypic and Evolutionary Adaptation. Cambridge: Cambridge University Press; 1996.
4. Diller KR. Modeling of bioheat transfer processes at high and low temperatures. Adv Heat Trans 1992;22:157–357.
5. Diller KR, Valvano JW, Pearce JA. Bioheat Transfer. In: Kneith F, editor. CRC Handbook of Heat Transfer, 2nd ed. 2005.
6. Charney CK. Mathematical models of bioheat transfer. Adv Heat Trans 1992;22:19–155.
7. Pennes HH. Analysis of Tissue and Arterial Blood Temperature in the Resting Human Forearm. J Appl Phys 1948;1:93–102.
8. Wissler E. A review of human thermal models. In: Morrison MB, editor. Environmental Ergonomics. New York: Taylor and Francis; 1988. p 267–285.
9. Wissler EH. Mathematical simulation of human thermal behavior using whole-body models. In: Shitzer A, Eberhart RC, editors. Heat Transfer in Medicine and Biology. Vol. 1. New York: Plenum Press; 1985. p 325–373.
10. Chato JC. Heat transfer to blood vessels. J Biomech Eng 1980;102:110–118.
11. Chen MM, Holmes KR. Microvascular contributions in tissue heat transfer. Annals NY Acad Sci 1980;335:137–150.
12. Weinbaum S, Jiji L, Lemons DE. Theory and Experiment for the Effect of Vascular Temperature on Surface Tissue Heat Transfer—Part 1: Anatomical Foundation and Model Conceptualization. ASME J Biomech Eng 1984;106:246–251.
13. Weinbaum S, Jiji L, Lemons DE. Theory and Experiment for the Effect of Vascular Temperature on Surface Tissue Heat Transfer—Part 2: Model Formulation and Solution. ASME J Biomech Eng 1984;106:331–341.
14. Weinbaum S, Jiji L. A New Simplified Bioheat Equation for the Effect of Blood Flow on Average Tissue Temperature. J of Biomech Eng 1985;107:131–139.
15. Charny CK, Weinbaum S, Levin RL. An Evaluation of the Weinbaum-Jiji Bioheat Equation for Normal and Hyperthermic Conditions. ASME J Biomech Eng 1990;112:80–87.
16. Xu LX, Chen MM, Holmes KR, Arkin H. The Evaluation of the Pennes, the Chen-Holmes, the Weinbaum-Jiji Bioheat Transfer Models in the Pig Kidney Cortex. ASME WAM Proc HDT 1991;189:15–21.
17. Arkin H, Xu LX, Holmes KR. Recent Developments in Modeling Heat Transfer in Blood Perfused Tissues. IEEE Trans Biomed Eng 1994;41(2):97–107.
18. Wissler EH. Pennes' 1948 paper revisited. J Appl Physiol 1998;85:35–41.
19. Pennes HH. Analysis of tissue and arterial blood temperatures in the resting forearm. J Appl Physiol 1948;1:93–122 (republished for fiftieth anniversary issue of J Appl Physiol 1998;85:5–34).
20. Scully RM, Barnes MR. Physical Therapy. Philadelphia: J.B. Lippincott Co.; 1989.
21. Roussy G, Pearce JA. Foundations And Industrial Applications Of Microwaves Physical And Chemical Processes. New York: John Wiley & Sons, Inc.; 1995.
22. Chappuis P et al. Heat storage regulation in exercise during thermal transients. J Appl Physiol 1976;40:384–392.
23. Webb P. The physiology of heat regulation. Am J Physiol 1995;268:R838–R850.
24. Nunneley SA. Water cooled garments: a review. Space Life Sci 1970;2:335–360.
25. Nunneley SA. Physiological response of women to thermal stress: A review. Med Sci Sports 1978;10:250–255.
26. Ganong WF. Review of Medical Physiology. 16th ed. Norwalk (CT): Appleton and Lange; 1993.
27. Fulcher CWG. Control of a liquid cooling garment for extravehicular astronauts by cutaneous and external auditory meatus temperatures, Ph.D. dissertation, Houston (TX): University of Houston; 1970.

28. Thomsen S. Mapping of thermal injury in biologic tissues using quantitative pathologic techniques. Proc SPIE 1999;3594-09:822–897.
29. Moritz AR, Henriques FC. Studies of Thermal Injury. II. The Relative Importance of Time and Surface Temperature in the Causation of Cutaneous Burns. Am J Path 1947;23:695–720.
30. Büttner K. Effects of extreme heat and cold on human skin. I. analysis of temperature changes caused by different kinds of heat application. J Appl Physiol 1951;3:691–702.
31. Büttner K. Effects of extreme heat and cold on human skin. II. surface temperature, pain and heat conductivity in experiments with radiant heat. J Appl Physiol 1951;3: 691–702.
32. Stoll AM. A computer solution for determination of thermal tissue damage integrals from experimental data. I R E Trans Med Electron 1960;7:355–358.
33. Stoll AM, Chianta MA. Burn production and prevention in convective and radiant heat transfer. Aerospace Med 1968;39:1232–1238.
34. Stoll AM, Green LC. Relationship between pain and tissue damage due to thermal radiation. J Appl Physiol 1959;14:373–382.
35. Ross DC, Diller KR. An experimental investigation of burn injury in living tissue. J Heat Trans 1976;98:292–296.
36. Lawrence JC, Bull JP. Thermal conditions which cause skin burns. J Inst Mech Eng Eng Med 1976;5:61–63.
37. Takata AN. Development of criterion for skin burns. Aerospace Med 1974;45:634–637.
38. Welch AJ, Polhamus GD. Measurement and prediction of thermal injury in the retina of Rhesus monkey. IEEE Trans Biomed Eng 1984;BME-31:633–644.
39. Thomsen S, Pearce JA, Cheong WF. Changes in birefringence as makers of thermal damage in tissues. IEEE Trans Biomed Eng 1989;BME-36:1174–1179.
40. Pearce JA, Thomsen S. Kinetic models of tissue fusion processes. Proc SPIE Laser Tissue Int III 1992;1643.
41. Polge C, Smith AU, Parkes AS. Revival of spermatozoa after vitrificaiton and dehydration at low temperatures. Nature 1949;164:666.
42. Lovelock JE The mechanism of the protective action of glycerol against haemolysis by freezing and thawing. Biochim Biophys Acta 1953;11:28–36.
43. Strumia MM, Clawell LS, Strumia PV. The preservation of blood for transfusion. J Lab Clin Med 1960;56:576–593.
44. Whittingham DG, Leibo SP, Mazur P. Survival of mouse embryos frozen to −196°C and −296°C. Science 1972;178: 411–414.
45. McGrath JJ, Diller KR, editors. Low Temperature Biotechnology: Emerging Applications and Engineering Contributions. New York: ASME; 1988. p 1–380.
46. Diller KR, Ryan TP. Heat transfer in living systems: current opportunities. J Heat Trans 1998;120:810–829.
47. Bowman HF. Estimation of Tissue Blood Flow. In: Shitzer, Eberhart, editors. Heat Transfer in Medicine and Biology. New York: Plenum; 1985. p 193–230.
48. Chato JC. Measurement of Thermal Properties of Biological Materials. In: Shitzer, Eberhart, editors. Heat Transfer in Medicine and Biology. New York: Plenum; 1985. p 167–192.
49. Valvano JW, et al. The simultaneous measurement of thermal conductivity, thermal diffusivity and perfusion in small volumes of tissue. J Biomech Eng 1984;106:192–197.
50. Valvano JW, et al. Thermal conductivity and diffusivity of biomaterials measured with self-heated thermistors. Intern J Thermophys 1985;6:301–311.
51. Valvano JW, Chitsabesan B. Thermal conductivity and diffusivity of arterial wall and atherosclerotic plaque. Lasers Life Sci 1987;1:219–229.
52. Patel PA, et al. A self-heated thermistor technique to measure effective thermal properties from the tissue surface. J Biomech Eng 1987;109:330–335.
53. Rastorguev YL, Ganiev YA. Thermal conductivity of aqueous solutions or organic materials. Russ J Phys Chem 1966;40: 869–871.
54. Touloukian YS, et al. Thermophysical Properties of Matter: Thermal Conductivity. Vol. 3. New York: IFI/Plenum; 1970. p 120, 209.
55. dos Santos I, et al. An instrument to measure the heat convection coefficient on the endocardial surface. Physiol Measur 2003;24:321–335.
56. dos Santos I, et al. In vivo measurements of heat transfer on the endocardial surface. Physiol Measur 2003;24:793–804.

Reading List

Welch AJ, van Gemert M, editors. Optical-Thermal Response of Laser Irradiated Tissue. New York: Plenum Press; 1995.
Valvano JW. Temperature Measurements. In: Advances In Heat Transfer: Bioengineering Heat Transfer. Vol. 22. New York: Academic Press; 1992. p 359–436.
Roussy G, Pearce JA. Foundations And Industrial Applications of Microwaves Physical And Chemical Processes. New York: John Wiley & Sons, Inc.; 1995.
Shitzer, Eberhart, editors. Heat Transfer in Medicine and Biology. New York: Plenum; 1985.
Kreith F, editor. CRC Handbook of Heat Transfer. 2nd ed. 2005.

See also CYSTIC FIBROSIS SWEAT TEST; HYPERTHERMIA, SYSTEMIC; TEMPERATURE MONITORING; THERMOMETRY.

BIOIMPEDANCE IN CARDIOVASCULAR MEDICINE

DOUGLAS A. HETTRICK
TODD M. ZIELINSKI
Medtronic, Inc.
Minneapolis, Minnesota

INTRODUCTION

Historical Context

Electrical impedance measurements have been applied to the study of biologic systems for nearly 200 years. Indeed, the history of continuously flowing electricity began with Luigi Galvani's famous experiments on bioelectricity at the University of Bologna (1,2). It was not until the 1870s however, that Hermann Müller in Konigsberg/Zürich discovered the capacitive properties of tissue and the anisotropy of muscle conductance based on alternating current measurements. In 1864, James C. Maxwell contrived his now famous equations by specifically calculating the resistance of a homogeneous suspension of uniform spheres as a function of their volume concentration (1). In 1928, Kenneth S. Cole expanded on Maxwell's model by determining

the impedance of a suspension of capacitively coated spheres over a range of frequencies.

Several additional and important developments occurred before the start of World War II. Rudolph Hoebers studied the conductivity of blood and found it to be dependent on the stimulation frequency. Simultaneously, the electrical properties of proteins and amino acids were discovered and extensively studied by Oncley, Fricke, and Wyman (3). These contributions lead to further developments in the science of biophysics and electrophysiology.

Bioimpedance research accelerated after World War II. In 1950, Nyboer et al. launched an investigation into thoracic electrical bioimpedance (TEB) as an alternative to invasive methods of measuring cardiac function and published a novel method termed "Impedance Plethysmography" (4,5). However, Kubicek and Patterson were credited with the development of the original TEB system in conjunction with the National Aeronautics and Space Administration in the mid-1960s (6). This device was designed to monitor stroke volume (SV) and cardiac output (CO) noninvasively during space flight. In addition, Djordjevich and Sadove coined the term "electrohemodynamics" in 1981 to describe a science that relates the theories of fluid mechanics and elasticity to the continuous impedance signal and to the time variations of arterial blood pressure (7). Jan Baan et al. introduced the impedance or conductance catheter technique to measure real time chamber volume in the mid-1980s (8). This technique revolutionized the study of cardiovascular mechanics in both the laboratory and clinical settings by making the study of ventricular-pressure volume relationships practical.

More recently, bioimpedance applications have continued to expand, especially in the area of implantable devices. Modern pacemakers and defibrillators routinely use biompedance measurements to verify pacing lead performance and position, monitor minute ventilation and thoracic fluid content, and optimize programmable device features such as pacing rate and AV delay in a closed-loop fashion (9–11).

The terms bioimpedance or tissue impedance describe both the resistive and reactive components of tissue at the applied stimulus frequency. The capacitive reactive components of the measured tissue impedance change at higher frequencies due to the relative conductive properties of tissue fluids and cellular membranes. Bioimpedance methods can be categorized into two areas: impedance plethysmography and impedance cardiography. Impedance plethysmography, by definition, refers to the measurement of a volume change in a heterogeneous tissue segment using electrical impedance in which the changing impedance waveform (ΔZ) is used to determine cardiac, respiration, and peripheral volume change as a function of time. In contrast, impedance cardiography is a subdivison of impedance plethysmography that focuses on the measurement of cardiac stroke volume and widely uses the first derivative (dZ/dt) of the changing impedance waveform (ΔZ) to monitor fiducial time element points such as cardiac valve opening and closing. Both methods primarily use a single low frequency stimulus current (<100 kHz) where most of the elements in the current paths are primarily resistive. Techniques such as impedance plethysmography and impedance cardiography primarily depend on resistive rather than reactive components of the blood impedance. Thus, applications using low frequency stimulus current to primarily measure the resistive component of bioimpedance will be categorized in this article as resistive applications of bioimpedance.

The second general category of bioimpedance measurement involves estimation of fluid volume distributions such as intracellular and extracellular volume, percent body fat vs. percent muscle mass, and cell and tissue viability. This area primarily employs a multifrequency stimulus current bandwidth (>1000 Hz) where most of the elements in the current path contain significant resistive and reactive components. Thus, applications using high frequency stimulus current to measure the resistive and reactive components of bioimpedance will be categorized as reactive applications of bioimpedance.

Bioimpedance Theory

When constant electric current is applied between two electrodes through a biological medium and the corresponding voltage is measured between the two source poles, the resultant impedance or bioimpedance is determined by Ohm's law. The recorded voltage is the sum of the potential difference contributions due to the electrical conductivity properties of the tissue medium. The exchange of electrons from source to sink occurs from electrons of the metal electrode (such as platinum or silver-silver chloride) to ions of the tissue medium. The electrode is the site of charge carrier exchange between electrons and ions and thus serves as a transducer of electrical energy. Impedance measurements most commonly use a two-electrode (bipolar) or four-electrode (tetrapolar) arrangement (Fig. 1). In the bipolar arrangement, the electrodes serve as both the current source (anode and cathode, respectively) and as the measurement electrodes. In the tetrapolar arrangement, one electrode serves as current source anode, one as the current source cathode, and the remaining two electrodes serve as the respective measurement electrodes. A disadvantage of the bipolar electrode system is electrode polarization due to a frequency-dependent polarization impedance. Therefore,

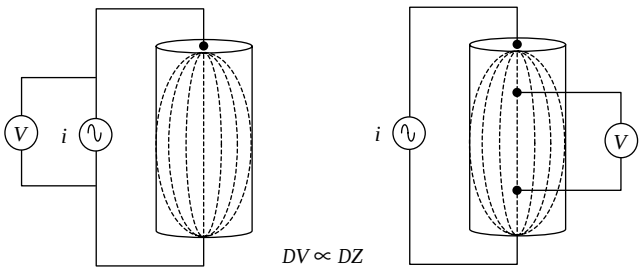


Figure 1. Bipolar (left) and tetrapolar (right) electrode configurations. With a constant current stimulation source, the change in measured voltage (ΔV) is proportional to the change in calculated impedance (ΔZ). Tetrapolar systems require additional electrodes but avoid electrode polarization effects.

the measured voltage in a bipolar impedance system reflects the combined impedance of both the tissue segment and the electrode tissue interface. When the voltage is measured with an isolated high input impedance electrode system, such as with the tetrapolar lead configuration, minimal current flows in the isolated sensing electrodes, thus problems with electrode polarization can be effectively reduced (12).

Endogenic ionic current movement between and within cellular structures encompasses the electrical properties of tissues and the term bioelectricity. In tissue and the living cell, an inseparable alliance exists between electricity and chemistry (1). The perception of current through human tissue is dependent on frequency, current density, effective electrode area, and current duration. The maximum sensitivity of the nervous system is approximately in the range of 10 to 1000 Hz for sine waves. At frequencies greater than 1 kHz, the sensitivity is strongly reduced.

Measured bioimpedance is a function of the real and reactive components of the tissue medium at the applied frequency of the stimulus current. Tissue characteristics at low frequencies are almost independent of cellular membrane reactance and internal intracellular resistivity. Thus, most of the applied current is conducted via the extracellular fluid. The cellular membrane behavior at intermediate or high frequencies is primarily a characteristic of membrane reactance and internal resistivity. The membranes are an impure reactance and, therefore, show a dielectric loss and a phase angle, which is independent of frequency (13). At higher frequencies, the cell's membrane reactance and resistance become negligible and the applied current is conducted through both the intracellular and extracellular fluid.

RESISTIVE APPLICATIONS OF BIOIMPEDANCE

Transthoracic Bioimpedance

The Cylindrical Model. Many applications of bioimpedance measurement focus on primarily resistive changes. These techniques are all based on the cylindrical model (Fig. 2) represented by a tissue volume with uniform cross-sectional area (A), length (L), and resistivity (ρ) (14).

$$R = \rho \frac{L}{A} \tag{1}$$

The resistance of the vessel segment is directly proportional to the resistivity of the conductive medium and length of the vessel segment and inversely proportional to the cross-sectional area of the vessel segment (Eq. 1). As shown in Fig. 2, as the cross-sectional area of the vessel segment increases from A_1 to A_3, the measured resistance decreases. Resistivity (ρ) is a tissue property that varies substantially between tissues. Typical tissue resistivities are shown in Table 1 (15).

Equation 1 can be modified to determine the volume of the tissue by multiplying both sides of the equation by L, substituting impedance (Z) for resistance, and solving for volume (V) as a function of time (Eq. 2):

$$V(t) = \rho \frac{L^2}{Z(t)} \tag{2}$$

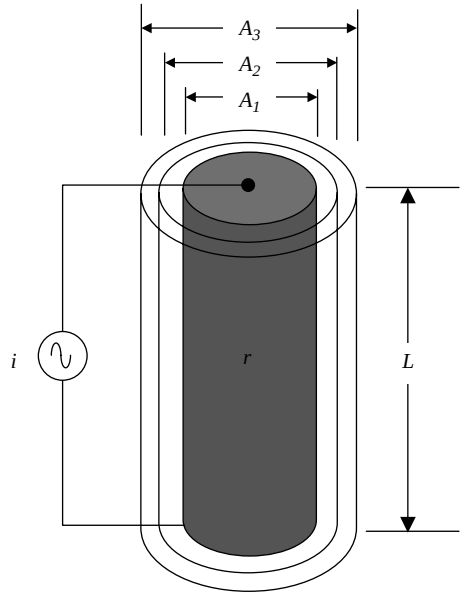


Figure 2. Cylindrical model of a vessel segment. A_1–A_3 represent cross-sectional area changes of the tissue of the interest (e.g., blood). Tissue length (L) is often determined by measurement electrode spacing. Blood resistivity (ρ) also determines the resistance (R) of the tissue volume. Resistance measured is directly proportional to the measured voltage and indirectly proportional to the constant alternating current (i) applied to the vessel segment.

The cylindrical model is based on several important assumptions: The electrical field, and hence current density, is homogeneous within the tissue of interest, the current is completely confined to the tissue of interest, and the values shown in Table 1 do not account for tissue anisotropy. Most biological tissues have lower resistivity in the longitudinal direction of cell or fiber orientation (12,16–18). For example, the ratio of resistivity in the transverse to parallel direction can be > 3 in cardiac tissue (18). These assumptions may not be valid for some applications of bioimpedance such as the conductance catheter technique for chamber volume estimation (see below). Therefore, the

Table 1. Various Tissue Resistivities in ohms-meter ($\Omega \cdot$ m) (15)

Tissue	ρ ($\Omega \cdot$ m)
Blood (Hematocrit = 45)	1.6
Plasma	0.7
Heart Muscle (Longitudinal)	2.5
Heart Muscle (Transverse)	5.6
Skeletal Muscle (Longitudinal)	1.9
Skeletal Muscle (Transverse)	13.2
Lung	21.7
Fat	25

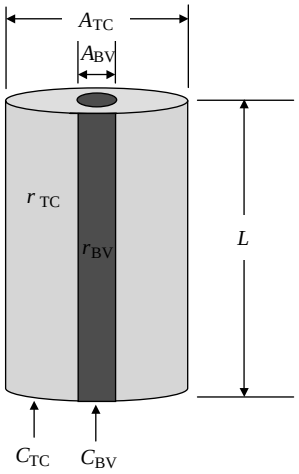


Figure 3. Parallel-column model of the thoracic cavity. This two-column model (C_{TC} and C_{BV}) represents a thoracic cavity segment of length (L), cross-sectional areas of the great blood vessels (A_{BV}), and thoracic cavity segment (A_{TC}), and resistivities of the thoracic cavity segment tissues (ρ_{TC}) and blood volume (ρ_{BV}).

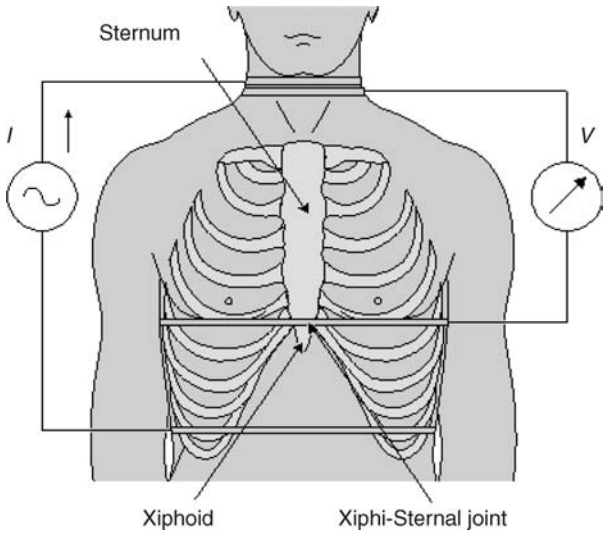


Figure 4. Transthoracic band electrode placement for stroke volume estimates. The two outer-band electrodes supply the stimulus current (I); the two inner electrodes measure the corresponding voltage (V). Impedance is calculated from the ratio of V/I (15).

cylindrical model must be adjusted for particular applications.

The Parallel-Column Model. The parallel-column model, first described by Nyboer (5) (Fig. 3) is closely related to the cylindrical model, but accounts for current leakage into surrounding tissues. The model consists of a smaller cylindrical conductor (C_{BV}) of length L representing the large blood vessels of the thoracic cavity (i.e., aortic and pulmonary arteries) embedded in a larger cylindrical conductor (C_{TC}) of the same length (L) representing the tissues of the thoracic cavity. C_{BV} consists of blood with specific resistivity (ρ_{BV}) and time-varying cross-sectional area (A_{BV}). C_{TC} is assumed to be heterogeneous (i.e., bone, fat, muscle) with specific resistivity (ρ_{TC}) and constant cross-sectional area (A_{TC}). Thus, the cylindrical model of the time-varying volume can be modified:

$$V_T(t) = \rho_{RV}\frac{L^2}{Z_{BV}(t)} + \rho_{TC}\frac{L^2}{Z_{TC}} \tag{3}$$

where $V_T(t)$ represents the total volume change. As the distribution of the measured resistance and the net resistivity of the parallel tissues are unknown, calculation of absolute volume can be problematic. However, the constant volume term drops out when the change in volume is calculated from Eq. 3:

$$\Delta V_{BV}(t) \cong \rho_{BV}\left(\frac{L^2}{Z_0^2}\right)\bullet\Delta Z(t) \tag{4}$$

where Z_0 is the basal impedance measured and $\Delta Z(t)$ is the pulsatile thoracic impedance change. Thus, Eq. 4 links the parallel cylindrical model to TEB estimates of stroke volume (ΔV_{BV}).

Noninvasive Measurement of Cardiac Output. Transthoracic electrical bioimpedance (TEB) was first introduced by Patterson et al. in 1964 (19). As shown in Fig. 4 (13), this system employs two pairs of band electrodes positioned at the superior and inferior ends of the thorax in the cervical and substernal regions, respectively. The outer electrode pair drives a constant current (I) and the inner electrode pair is used to measure the corresponding voltage (V), which is a function of the varying impedance changes during respiration and the cardiac cycle.

Noninvasive measurements of stroke volume can be determined with the configuration in Fig. 4 by applying Eq. 5.

$$SV = \left[\rho_{BV}\left(\frac{L^2}{Z_0^2}\right)\bullet\Delta Z(t)\right]\bullet LV_{ET} \tag{5}$$

Cardiac output may then be determined by multiplying stroke volume (SV) by heart rate (HR). $\Delta Z(t)$ is the measured time-varying impedance signal, and Z_0 represents the nonpulsatile basal impedance.

Sramek et al. (20) modified Patterson et al.'s (19) parallel cylinder model into a truncated cone in order to improve stroke volume predictions (Eq. 6). The physical volume of the truncated cone was determined to be one-third the volume of the larger thoracic cylinder model.

$$SV = \left(\frac{(L)^3}{4.2}\right)\bullet LV_{ET}\bullet\left(\frac{(dZ/dt)_{max}}{Z_0}\right) \tag{6}$$

Sramek et al. (20) also found that in a large normal adult population, the measured linear distance (L) is equal to 17% of body height (cm). Cardiac output is directly proportional to body weight (21). As ideal body weight is a linear function of overall height (22), the proportionality of height (H) to cardiac output can be represented in the first term of Eq. 6 by $(0.17H)^3/4.2$.

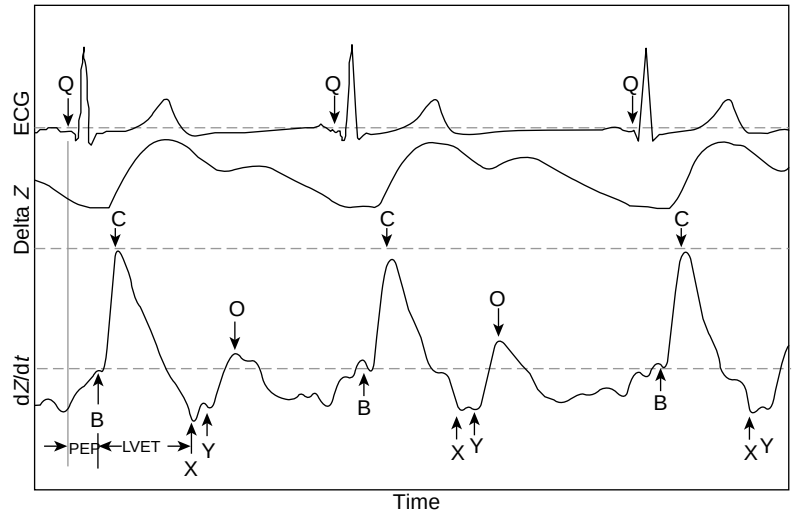


Figure 5. Impedance cardiography waveforms. Three waveforms depict the electrocardiogram (ECG), transthoracic impedance change as a function of time (ΔZ), and first-time derivative of the impedance change dZ/dt. Impedance waveforms are intentionally inverted to show a positive deflection during cardiac contraction. Fiducial points on the dZ/dt waveform are represented by the opening of the aortic and pulmonic valves (B), closure of the aortic (X) and pulmonic (Y) valves, mitral valve opening (O), ventricular pre-ejection period (PEP), and left ventricular ejection time (LVET). Q represents the end of atrial contraction (46).

Other empirical modifications to the original model have also been proposed in order to improve CO and SV estimates (20,22–28). In addition, various modifications to the external lead configuration have also been proposed to improve SV estimates, including the application of tranesophageal electrodes (29,30).

Several commercial bioimpedance systems are available for clinical noninvasive estimation of cardiac output and other hemodynamic parameters. The advantages of such systems include noninvasive application, relatively low cost, and lack of noninvasive alternatives. However, these techniques have gained somewhat limited clinical acceptance due to suspect reliability over a wide range of clinical conditions.

A myriad of validation studies of TEB estimates of cardiac output have been published with equivocal results (6,7,19,24,28,31–40). For example, Engoren et al. (35) recently compared cardiac output as determined by bioimpedance, thermodilution, and the Fick method and showed that the three methods were not interchangeable in a heterogenous population of critically ill patients. Their data showed that measurements of cardiac output by thermodilution were significantly greater than by bioimpedance. However, the bioimpedance estimates varied less than the thermodilution estimates for each subject. In contrast, a meta-analysis of impedance cardiography validation trials by Raaijmakers et al. (38) showed an overall correlation between cardiac output measurements using transthoracic electrical bioimpedance cardiography and a reference method of 0.82 (95% CI: 0.80–0.84). The performance of impedance measurement of cardiac output was similar in various groups of patients with different diseases with the exception of cardiac patients, in which group the correlation was decreased. Additional investigations by Kim et al. (41) and Wang et al. (42) used a detailed 3D finite element model of the human thorax to determine the origin of the transthoracic bioimpedance signal. Contrary to the theory that lead to the parallel column model formulae, these investigators determined that the measured impedance signal was determined by multiple tissues and other factors that make reliable estimates of cardiac output over a wide variety of physiological conditions difficult. Nevertheless, commercially available impedance plethysmographs provide estimates of cardiac output that may be useful for assessing relative changes in cardiac function during acute interventions, such as optimization of implantable pacemaker programmable options such as AV delay (11).

Cardiac Cycle Event Detection. TEB also focuses on measurements of the change in impedance (ΔZ) and the impedance first time derivative (dZ/dt) measured simultaneously with the electrocardiogram (ECG). Figure 5 depicts a typical waveform of the aforementioned parameters. Note that the impedance change (ΔZ) and the impedance first time derivative (dZ/dt) are inverted by convention (43). The value of dZ/dt is measured from zero to the most negative point on the waveform. The ejection time (LV_{ET}) is an important parameter in determining stroke volume (Eqs. 5 and 6). This systolic time interval allows an estimation of cardiac contractility. The Heather Index (HI) is another proposed index of contractility from systolic time intervals determined by the dZ/dt waveform (33,44) (Eq. 7):

$$HI = \frac{dZ/dt_{\max}}{QZ_1}$$

where $dZ/dt_{\max}$ (point C) is the maximum deflection of the initial waveform derived from the ΔZ waveform and QZ_1 is the time from the beginning of the Q wave to peak $dZ/dt_{\max}$ (Fig. 5). In this figure, point Q represents the time between the end of the ECG p-wave (atrial contraction) and the beginning of the QRS wave (ventricular depolarization) (45). Point B depicts the opening of the aortic and pulmonic valves. After the ventricles depolarize and eject the blood volume into the aortic and pulmonary arteries, points X and Y represent the end systolic component of the cardiac cycle as closure of the aortic and pulmonic valves, respectively.

Passive mitral valve opening and passive ventricular filling begins at point O. Although the timing of the various

fiducial notches of the dZ/dt waveform is well known, controversy remains with the origins of the main deflections and are not well understood (15).

Signal Noise. In TEB measurements, several filtering techniques have been proposed to attenuate undesired noise sources depending on which component of the impedance waveform is desired (i.e., respiratory, cardiac, or mean impedance) (47). Most of the signal processing techniques for impedance waves use ensemble averaging for the elimination of motion artifacts (48). A recent signal processing technique described by Wang et al. (48) uses the time-frequency distribution to identify fiducial points on the dZ/dt signal for the computation of left ventricular ejection time and dZ/dt_{max}. As shown in Fig. 5, many of the fiducial points on the dZ/dt waveform are clearly identifiable, but may be somewhat more difficult to observe under severe interference conditions.

Filtering techniques have also been proposed to eliminate noise caused by respiration such as narrow band-pass filtering around the cardiogenic frequency. However, such filtering techniques often eliminate the high frequency components of the cardiac signal and introduce phase distortion (49). To help alleviate this problem, various techniques to identify breathing artifacts with forward and backward filtering have been employed (50,51). Despite these techniques, motion artifact remains with unknown frequency spectra that may overlap the desired impedance frequency spectra during data acquisition. Adaptive filters represent another approach and may eliminate the motion artifact by tracking the dynamic variations and reduce noise uncorrelated to the desired impedance signal (49). Raza et al. (51) developed a method to filter respiration and low frequency movement artifacts from the cardiogenic electrical impedance signal. Based on this technique, the best range for the cutoff frequency appears to be from 30–50% of the heart rate under supine, sitting, and moderate exercise conditions (51).

Applications of Transthoracic Bioimpedance. ***Hypertension.*** TEB has emerged as a noninvasive tool to assess hemodynamic parameters, especially within the framework of hypertension monitoring (24–28,52–54). Measurement of the various hemodynamic components such as stroke volume, ejection time, systemic vascular resistance, aortic blood velocity, thoracic fluid content, and contractility (i.e., Heather Index) using impedance cardiography in patients with hypertension allows more complete characterization of the condition, a greater ability to identify those at highest risk, and allows more effectively targeted drug management (25,53). Several studies have used TEB to evaluate hemodynamic parameters and demonstrated that TEB-guided therapy improves blood pressure control (53,55,56). For example, in a three month clinical study by Taler et al. (55) 104 hypertensive patients were randomized to either TEB-guided therapy or standard therapy. The results showed improved blood pressure control in the TEB-guided group. The investigators concluded that measurement of hemodynamic parameters with TEB methods was more effective than clinical judgment alone in guiding selection of antihypertensive therapies in patients resistant to empiric therapy (53).

Pacemaker Programming. TEB estimates of cardiac index technique has been investigated as a noninvasive method to optimize AV delay intervals in pacemaker patients in an open-loop fashion (57,58). Ovsyshcher et al. (58) measured stroke volume changes at various programmed AV delays via impedance cardiography in dual-chamber pacemaker patients. The optimal and worst programmed AV delays were identified as the settings that produced the highest and lowest cardiac index, respectively. As shown in Fig. 6 (58), the highest cardiac index values resulted with mean AV delays <200 ms and the lowest cardiac index values resulted with mean AV delays >200 ms.

More recently, a study by Tse et al. (59) evaluated AV delay interval optimization during permanent left ventricular pacing using transthoracic impedance cardiography in conjuction with Doppler echocardiography over a range of AV intervals. This study revealed no significant difference between the optimal mean AV delay interval determined by transthoracic impedance cardiography and that determined by Doppler echocardiography. However, as shown in Fig. 7, the mean cardiac output at different AV

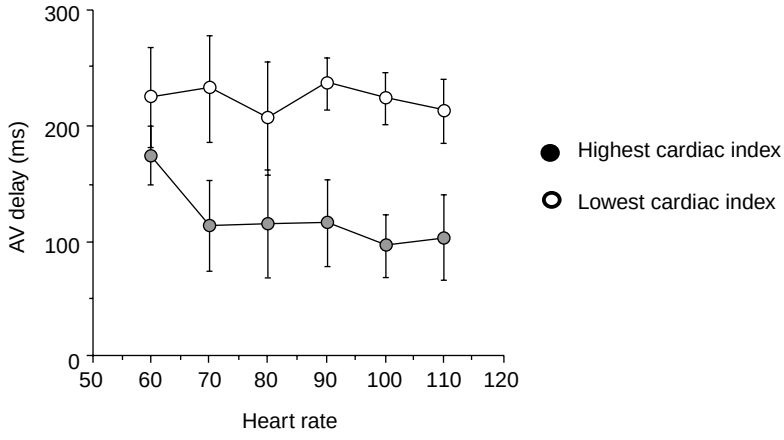


Figure 6. Highest and lowest cardiac indices at varied AV delays and pacing rates. Highest cardiac index values (closed circles) resulted with mean AV delays <200 ms and the lowest cardiac index values (open circles) resulted with mean AV delays >200 ms (58).

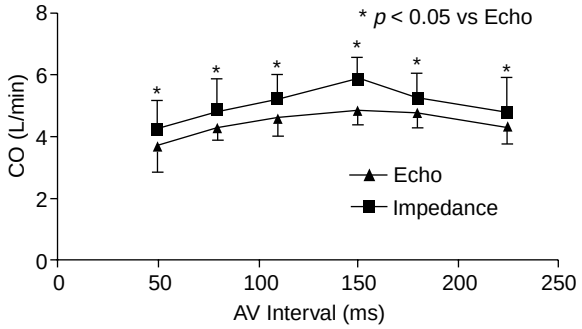


Figure 7. Cardiac output measured by impedance cardiography and Doppler echocardiography at various AV delay intervals (59).

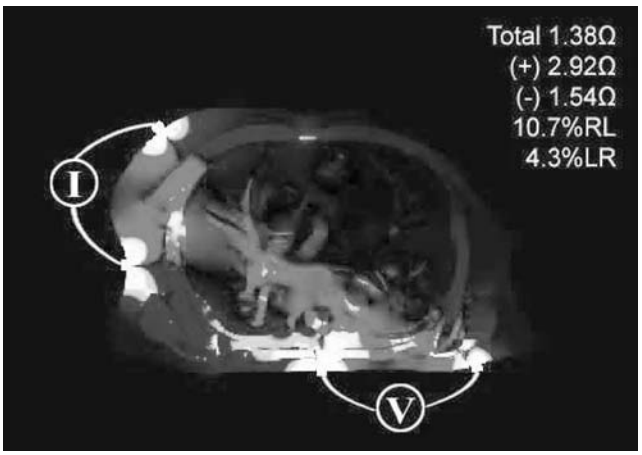


Figure 9. High resolution computer simulated model of the thorax. Regions of both positive and negative sensitivity contribute to the total impedance measured. Negative impedance sensitivity regions, (1.54 Ω) and positive impedance sensitivity regions (red, 2.92 Ω) both contribute to the measure total impedance of 1.38 Ω. RL = contribution of right lung to measured impedance. LR = concontribution of left lung to measured impedance (62).

delay intervals was significantly higher when measured by transthoracic impedance cardiography than when measured by Doppler echocardiography.

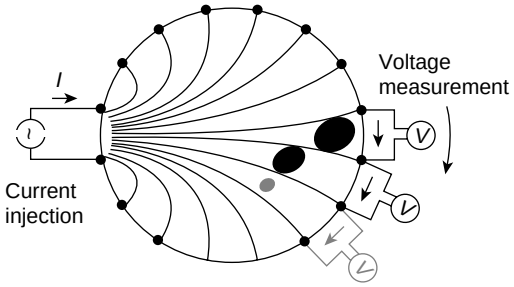
Electrical Impedance Tomography. Electrical impedance tomography (EIT) is a technique to reconstruct low resolution cross-sectional images of the body based on differential tissue resistivity (60). The image is created using an array of 16–32 electrodes, usually positioned around the thorax (Fig. 8). Impedance is computed from all electrodes as the drive electrodes rotate sequentially about the tissue surface. The "image" is then reconstructed using standard tomographic techniques. The advantages of EIT include low cost and the potential for ambulatory applications. The disadvantages include the low resolution of the image, the contribution of "out-of-plane" tissues to the "in-plane" image, and the limited clinical applications.

Recent improvements in hardware and software systems that increase the accuracy and speed of regional lung volume change have maintained interest in this technology (60–62). Besides pulmomary monitoring, other potential applications of EIT include neurophysiology, stroke detection, breast cancer detection, gastric emptying, and cryosurgery (63–66).

Lead Field Theory. An analysis of sensitivity is crucial to interpretation and application of EIT images as well as other bioimpedance applications. The sensitivity distribution of an impedance measurement provides the relation between the measured impedance resulting from the conductivity distribution of the measured region. It describes the relative contribution of each region to the measured impedance signal. The contribution of any region to the measurement is not always intuitively obvious and the magnitude of the sensitivity may be less than zero (Fig. 9). Therefore, the relative contribution of various tissues to the reconstructed "image" can be difficult to interpret.

The applicability of lead field theory in impedance measurements has been shown theoretically by Geselowitz (67). According to that theory, appropriate selection of the electrode configuration enables increased measurement sensitivity and selectivity to particular regions (68). Also, the measured impedance change (ΔZ) can be evaluated from the change in conductivity within a volume conductor $\Delta\sigma$ and the sensitivity distribution S by (67):

$$\Delta Z = \int_v \frac{1}{(\Delta)\sigma} \bullet S_{dv} \tag{8}$$



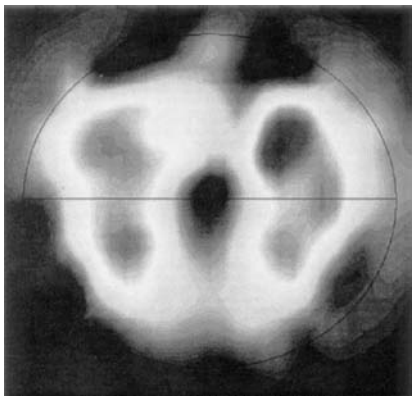


Figure 8. Left: Cartoon representation of a system to generate an electrical impedance tomographic image: 16 electrodes around the chest inject currents and record the resultant voltage in a sequential manner. Right: Electrical impedance tomographic image of the thoracic cavity. Heart and lung tissue are distinguishable (60).

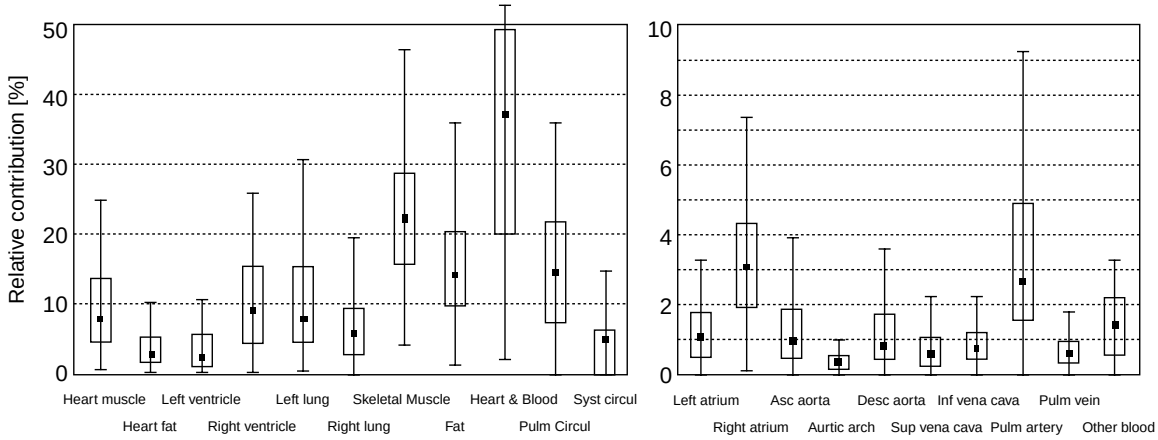


Figure 10. Simulated measurement sensitivities of tissues. Values are indicated for each tissue type in addition to three tissue groups consisting of pulmonary circulation, systemic circulation, and all the blood masses and heart muscle (68).

The measurement sensitivity S is obtained by first determining the current fields generated by a unit current applied to the current injection electrodes and the voltage measurement electrodes. These two lead fields form the combined sensitivity field of the impedance measurement associated with the electrode configuration by:

$$S = J_{\mathrm{LE}} \bullet J_{\mathrm{LI}} \tag{9}$$

where:

S = the scalar field giving the sensitivity to conductivity changes at each location,

J_{LI} = the lead field produced by current excitation electrodes,

J_{LE} = the lead field produced by voltage measurement leads.

Therefore, sensitivity at each location depends on the angle and magnitude of the two fields and can be positive, negative, or null. The relative magnitude of the sensitivity field in a tissue segment provides a measure of how conductivity variation in that tissue segment will affect the detected ΔZ (69).

Lead field theory suggests that the relative contribution of a tissue to the measured impedance depends on the properties of the tissue, the symmetric arrangement of the tissues, and the geometry of the applied current and voltage electrodes. The precise relative contribution of various tissues to measure impedance is therefore difficult to predict (Fig. 10) (68).

Intrathoracic Bioimpedance

Minute Ventilation. As described earlier, respiratory rate can be estimated with TEB. However, intrathoracic impedance sensing has also been applied to measure respiratory rate and minute ventilation in implantable devices such as pacemakers and implantable cardiac defibrillators (ICDs). Intrathoracic impedance vector configurations typically consist of a tripolar arrangement with bipolar pacing or ICD leads placed in the right ventricle (RV) and the device "can" (metal case or housing) placed subcutaneously in the left or right pectoral region. A variety of anode/cathode electrode arrangements are possible with current source electrodes such as the proximal electrode (RV-ring) to can or the distal electrode (RV-tip) to can and voltage sense electrodes between RV-coil to can. Typically, a low energy pulse of low current amplitude (1 mA with pulse duration of 15 μs) is delivered every 50 milliseconds (10). Figure 11 depicts a typical lead arrangement used for intrathoracic impedance measurements.

The electric fields generated with this electrode configuration must be arranged to intersect in parallel in order to provide the greatest sensitivity. The sensitivity of an electrode is proportional to the current density of the applied stimulus. Moreover, the sensitivity is highest close to the current-injecting electrodes and lowest toward the center of the tissue medium within the lead field vector,

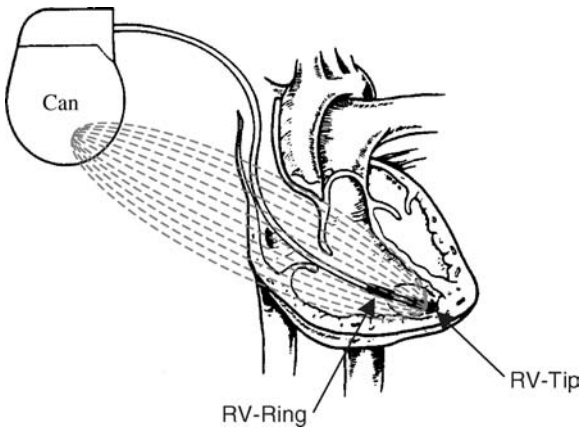


Figure 11. Lead configuration for intracardiac impedance measurments. Stimulus current injected from RV-tip to can. Voltage is measured from RV-Ring to can. The large size of the can reduces electrode polarization effects of the tripolar lead configuration.

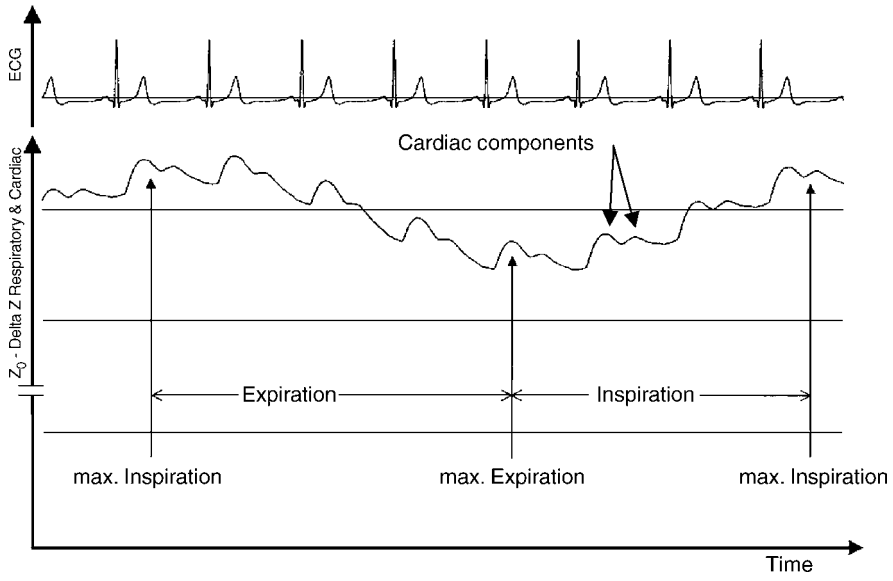


Figure 12. Respiratory variation in impedance waveform. Electrocardiogram (ECG) shown with ΔZ waveform. ΔZ waveform is comprised of higher frequency cardiac components superimposed on the lower frequency respiratory variation component (46).

because the applied field current density is lowest in this region.

Experimental evidence indicates that the frequency and amplitude of the respiratory component of the bioimpedance signal are related to changes in both the respiratory rate and the tidal volume and, hence, the minute ventilation (MV). MV sensing in rate-adaptive pacing systems has also been shown to closely correlate with carbon dioxide production ($V\mathrm{CO}_2$) (10). This relationship has been applied in some commercially available pacemakers with automatic rate-adaptive pacing features (9,10).

As shown in Fig. 12 (46), the amplitude of impedance changes during respiration are significantly larger than the higher frequency cardiac components. By magnitude, the change in the cardiac component of the impedance waveform is in the range of 0.1–0.2 Ω, which correlates to approximately 0.3–0.5% of the thoracic impedance (ΔZ) (46). Moreover, each component has a different frequency, typically 1.0–3.0 Hz for cardiac activity and 0.1–1.0 Hz for respiratory activity (9). This differentiation allows extraction of each signal by specific filtering techniques.

In general, the minute-ventilation sensor is characterized by a highly proportional relationship to metabolic demand over a wide variety of exercise types (10). However, optimal performance of impedance-based MV sensors to control pacing rate during exercise often requires careful patient-specific programming.

Fluid Status. Fluid congestion in the pulmonary circulation due to volume overload results in preferential transport of fluid primarily into the extracelluar fluid space and not into the intracellular compartments. Clinical symptoms to assess fluid overload include hypertension, increased weight, pulmonary or peripheral edema, dyspnea, and left ventricular dysfunction. Recently, implantable device-based bioimpedance measurements have been applied to detect thoracic fluid accumulation in patients with congestive heart failure (CHF) and to provide early warning of decompensation caused by factors such as volume overload and pulmonary congestion (32,70,71). This application is the result of a substantial body of new and historical experimental evidence (32,70–73).

Externally measured transthoracic impedance techniques have been shown to reflect alterations in intrathoracic fluid and pulmonary edema in acute animal and human studies (72). The electrical conductivity and the value for transthoracic impedance are determined at any point in time by relative amounts of air and fluid within the thoracic cavity (73). Additional studies have suggested that transthoracic impedance techniques provide an index of the fluid volume in the thorax (32,71). Wang et al. (70) employed a pacing-induced heart failure model to demonstrate that measurement of chronic impedance using an implantable device effectively revealed changes in left ventricular end-diastolic pressure in dogs with pacing-induced cardiomyopathy (Fig. 13) (70). Several factors were identified that may influence intrathoracic impedance with an implantable system, including (*1*) fluid accumulation in the lungs due to pulmonary vascular congestion, pulmonary interstitial congestion, and pulmonary edema; (*2*) as heart failure worsens, heart chamber dilation and venous congestion occur and pleural effusion may develop; and (*3*) after implant, the tissues near the pacemaker pocket swell and surgical trauma can cause fluid buildup (70).

Yu et al. (74) also showed that sudden changes in thoracic impedance predicted eminent hospitalization in 33 patients with severe congestive heart failure (NYHA Class III–IV). During a mean follow-up of 20.7 ± 8.4 months, 10 patients had a total of 25 hospitalizations for worsening heart failure. Measured impedance gradually decreased before admission by an average of $12.3 \pm 5.3\%$ ($p < 0.001$) over a mean duration of 18.3 ± 10.1 days. The decline in impedance also preceded the symptom onset by a mean lead time of 15.3 ± 10.6 days ($p < 0.001$). During hospitalization, impedance was inversely correlated with

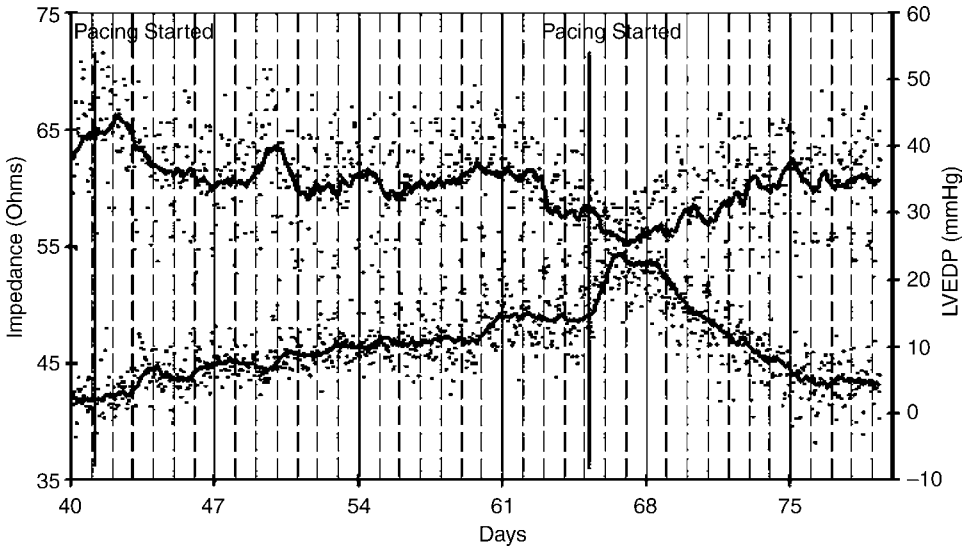


Figure 13. Impedance vs. LVEDP during pacing-induced heart failure. Intrathorcic impedance via an implantable device-lead configuration and LVEDP are inversely correlated in a canine model of pacing-induced cardiomyopathy. A general trend for impedance to decrease as heart failure developed is shown. Once pacing induced heart failure was terminated, LVEDP and impedance returned to basal levels (70).

pulmonary wedge pressure (PWP) and volume status with $r = -0.61$ ($p < 0.001$) and $r = -0.70$ ($p < 0.001$), respectively. Automated detection of impedance decreases was 76.9% sensitive in detecting hospitalization for fluid overload with 1.5 false-positive (threshold crossing without hospitalization) detections per patient-year of followup. Thus, intrathoracic impedance from the implanted device correlated well with PWP and fluid status, and may predict eminent hospitalization with a high sensitivity and low false-alarm rate in patients with severe heart failure (Fig. 14) (74). Some commercially available implantable devices for the treatment of CHF or ventricular tachyarrhythmias now continually monitor intrathoracic impedance and display fluid status trends. This information is then provided to the clinician via direct-device interrogation or by remote telemetry.

Volume Conductance Catheter. The conductance catheter technique, first described by Baan et al. (8), enables continuous measurements of chamber volume, particularly left ventricular (LV) volume. This method has been used extensively to assess global systolic and diastolic ventricular function (75). In many respects, the conductance catheter revolutionized the study of cardiovascular mechanics in both the laboratory and clinical settings by making the study of ventricular pressure-volume relationships practical. The technique led to a renaissance of cardiac physiology over the past 25 years (76) by increasing the understanding of the effect of pharmacologic agents, disease states, pacing therapies, and other interventions on cardiovascular function. Conductance catheter systems are available for clinical and laboratory monitoring applications, including a miniature system capable of measuring LV volume and pressure in mice (77).

The conductance methodology is based on the parallel cylinder model (Fig. 3). However, the cylindrical model assumes that the volume of interest has a uniform cross-sectional area across its length. Therefore, the ventricular volume is subdivided into multiple segments determined by equipotential surfaces bounded by multiple sensing electrodes along the axis of the conductance catheter (Fig. 15).

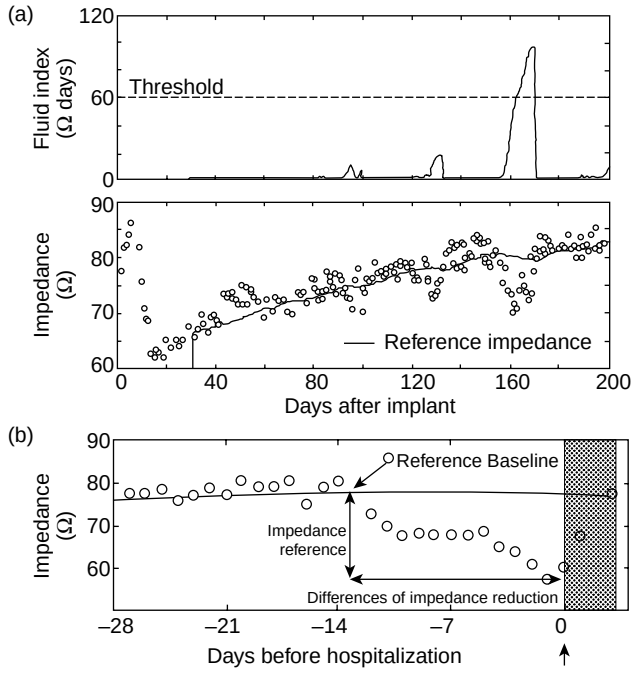


Figure 14. Fluid status monitoring with an implanted device. A: Operation of algorithm for detecting decreases in impedance over time. Differences between measured impedance (bottom; °) and reference impedance (solid line) are accumulated over time to produce fluid index (top). Threshold values are applied to fluid index to detect sustained decreases in impedance, which may be indicative of acutely worsening thoracic congestion. B: Example of impedance reduction before heart failure hospitalization (arrow) for fluid overload and impedance increase during intensive diuresis during hospitalization. Label indicates reference baseline (initial reference impedance value when daily impedance value consistently falls below reference impedance line before hospital admission). Magnitude and duration of impedance reduction are also shown. Days in hospital are shaded (74).

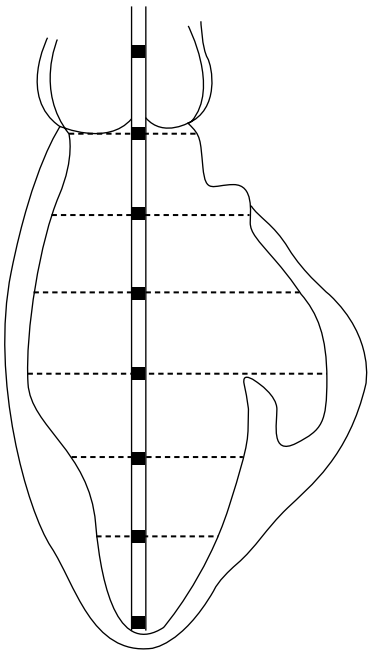


Figure 15. Conductance catheter modeled in the left ventricle (LV). Stimulus current is injected from the proximal and distal catheter electrodes. Voltage is measured between the remaining adjacent electrode pairs. Total conductance is calculated by the summation of all segmental conductances measured in the individual segments.

The two most distal electrodes are used to generate an electric field, typically 0.4 mA p-p, at 20 kHz. The remaining electrodes are used in pairs to measure the conductance of several segments ($n =$ number of segments), which represent the instantaneous volumes of the corresponding segment. The conductance is then converted to volume by modifying Eq. 2:

$$V(t) = \rho L^2 \sum_{i=1}^{n} G_i(t) \tag{10}$$

where G is the time-varying conductance of segment i. However, the conductance technique also violates two other key assumptions of the cylindrical model. First, the electrical field generated by the drive current electrodes is not homogenous and, second, the electric field is not confined to the chamber of interest (i.e., the LV). Thus, the multiple segment cylindrical model has been modified in order to allow conductance catheter estimates of volume to agree with gold standard estimates such as echocardiography (Eq. 11):

$$V(t) = \left(\frac{\rho L^2 \Sigma G(t)}{\alpha}\right) - V_P \tag{11}$$

where correction factor α accounts for nonhomogeneity of the electric field and the correction factor V_P accounts for the current leakage into the surrounding tissues. The terms α and V_p are related and may vary somewhat during the cardiac cycle (16,78). Various methods have been applied to determine the values of α and V_P, including the method of hypertonic saline injection. Recently, the concept of dual-frequency excitation has been applied to estimate V_P for conductance volume measurements in mice (79). This method takes advantage of the relative reactive components of impedance between blood and tissue (80). Despite some theoretical limitations regarding the basic assumptions of field heterogeneity and current leakage, the conductance catheter technique has also been applied to the study of biomechanics in other chambers besides the left ventricle, including the right ventricle (81), right and left atria (82), and aorta (83,84).

Other Pacing Applications. Intracardiac impedance, or transvalvular impedance (TVI), can be used in the assessment of cardiac hemodynamics. This method involves determining the impedance between pacemaker leads in the right atrium and ventricle using a typical dual-chamber pacing configuration. The TVI waveform can be categorized into atrial, valvular, and ventricular components (Fig. 16) (85). Information derived from the atrial component may be useful to identify the loss of electrical capture in the atrium, or the impairment in atrial hemodynamic function associated with supraventricular tachyarrhythmias. The valvular and ventricular components may provide information on the presence, timing, and strength of ventricular mechanical activity (86).

In a study performed by Gasparini et al. (85), the representative TVI tracings (Fig. 16) were recorded from atrial ring to ventricular tip. TVI was measured by application of 64 Hz subthreshold current pulses of 125 μs duration and the amplitude ranging from 15 to 45 μA. The TVI

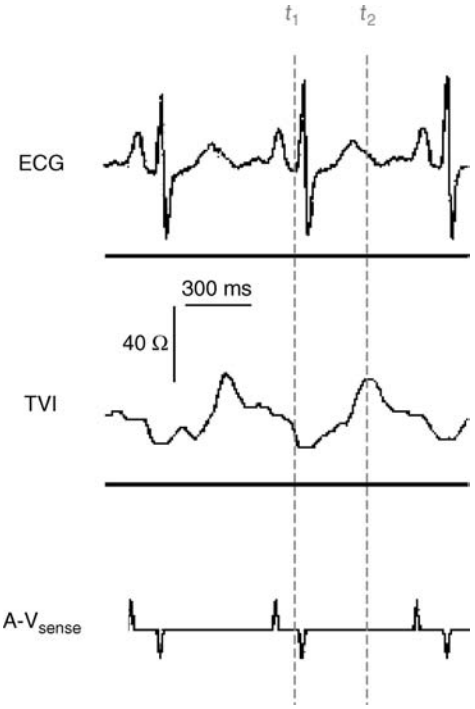


Figure 16. TVI during spontaneous A-V sequential activity. Fiducial points on the TVI waveform used to optimize A-V delay. t_1 corresponds to the end of atrial systole; t_2 corresponds to the end of ventricular ejection (85).

signal was recorded without high-pass filtering to determine the absolute minimum and maximum impedance in each cardiac cycle, which were assumed to reflect the end-diastolic volume and the end-systolic volume, respectively. TVI may represent a useful approach to determine hemodynamic parameters such as stroke volume, ejection fraction, pre-ejection interval, and atrio-ventricular delay. One significant advantage with this technique is that the source and sense leads are of those typically used in pacing systems and offer the advantage of a high signal-to-noise ratio (86). Moreover, the use of this technique has the potential to differentiate atrial from ventricular function that would be paramount if this technique is used for atrioventricular delay optimization (87).

Hematocrit Measurement. Measurement of the resistivity of whole blood has been investigated by a number of researchers, particularly in the area of transthoracic impedance techniques (88–91). A number of investigators have found blood resistivity to be an exponential function of hematocrit (Fig. 17) (15,89–95). These studies have demonstrated a strong correlation between the electrical resistivity of blood at frequencies between 20 to 50 kHz, as the red blood cell is the major resistive component in blood, compared with the relatively conductive plasma. Pop et al. (93) employed a four ring catheter electrode system with narrow electrodes spacing (2 mm center-to-center) to estimate hematocrit in the right atrium of anesthetized pigs. As shown in Fig. 17, good correlation existed between the hematocrit of blood and its electrical resistivity ($r^2 = 0.95$–0.99). Moreover, this study also showed a strong correlation between whole blood viscosity and electrical resistivity.

This interesting observation implies that intracardiac impedance has potential to monitor thrombosis risk in patients with hyperviscosity.

Blood Flow Conductivity Based on Erythrocyte Orientation. The electrical properties of blood are of practical interest in medicine because blood has the highest conductivity of all living tissues (89,96,97). Blood is a heterogeneous suspension of erthrocyctes that have a higher resistivity than the suspending fluid (plasma). The resistivity of blood is a function of the resistivities of plasma, the (fractional) packed-cell volume or hematocrit, and the orientation of the erythrocytes, due to their biconcave shape (98). The orientation of the erythrocytes can be influenced by the viscous forces in flowing blood, resulting in a shear rate-dependent resistivity. In stationary blood, the erythrocytes assume a random distribution while in flowing blood, the plane of the erythrocytes becomes oriented parallel to the axis of flow (99). Thus, minimum resistance occurs when the erythrocytes are oriented in an axial direction, parallel to the stream line. Conversely, maximum resistance occurs when the erythrocytes are oriented in a transverse direction to the stream line (100). The electrical properties of pulsatile blood flow are important when applying transthoracic bioimpedance to estimate cardiac output. In an experiment performed by Katsuyuki et al. (101), erythrocyte orientation, deformation, and axial accumulation caused differences in resistance between flowing and resting blood. Frequency characteristics of blood resistance under pulsatile flow showed that at low pulse rates, the resistance change was minimal, whereas at higher pulse rates, the resistance change increased because the orientation of the erythrocytes cannot follow the rapid changes of pulsatile blood flow. These results suggest that one mechanism of the varying resistance of blood in the aorta during pulsatile blood flow occurs because the orientation of the erythrocyte changes due to shear as a function of heart rate. Therefore, hemodynamic parameters such as cardiac output measured by impedance plethysmography must take into account the anisotropic electrical properties of oriented erythrocytes in blood. Moreover, the resulting resistance of flowing blood depends on the direction of the electrical field applied for impedance measurement and may be affected by the orientation of the erythrocytes during pulsatile flow (101).

REACTIVE APPLICATIONS OF BIOIMPEDANCE

Tissue Impedance

The reactive component of tissue impedance does not contribute significantly to measured impedance when the driving frequency range is less than 1 kHz (8,15). However,

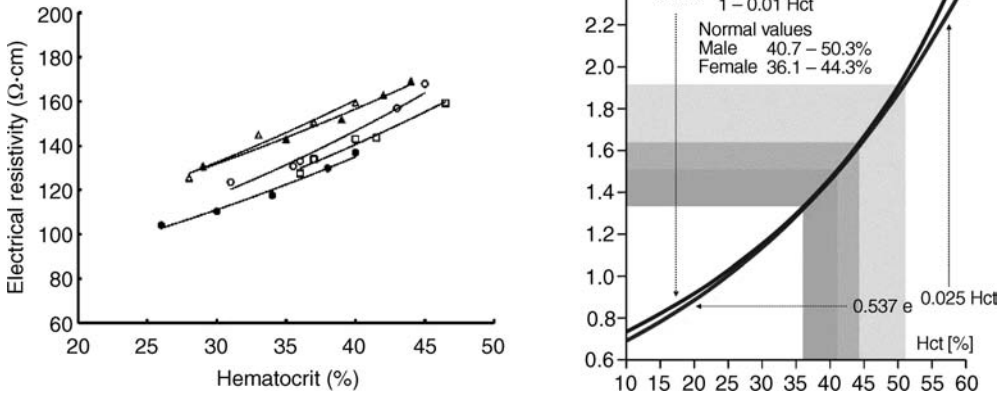


Figure 17. Left figure depicts the correlation between hematocrit of blood and electrical resistivity in five subjects. Right figure depicts a similar correlation between hematocrit of blood and electrical resistivity based on equations by Maxwell–Fricke (upper curve) and Geddes and Sadler (lower curve) (15,93).

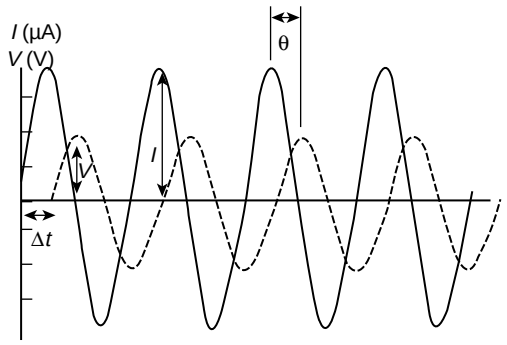


Figure 18. Relationship in phase angle and amplitude for tissue electrical properties. This example shows a capacitive tissue segment since the current waveform (I) leads the voltage waveform (V) by phase angle (θ) (103).

at higher driving current frequencies, the reactive component may contribute more substantially. As different tissues have different reactance, different frequencies may be selected for impedance measurement in order to discriminate various tissues (15,102).

Tissue impedance is characterized by four components: the in-phase component of voltage (V) with respect to the current intensity (I), the tissue resistance (R), and the phase angle (θ). The phase angle represents the time delay between the voltage and current intensity waves due to the capacitance of cell membranes (Fig. 18) (103).

Figure 19 shows cellular tissue structure representing alternating current distribution between a bipolar electrode pair at high and low frequencies. The change in polarity that occurs with AC current causes the cell membrane to charge and discharge at the rate of the applied frequency, and the impedance decreases as a function of increased frequency, because the amount of conducting volume increases through intracellular space. At higher frequencies, the rate of cell membrane charge and discharge becomes such that the effect of the cellular membrane on measured impedance becomes insignificant and the current flows through the intracellular and extracellular space (104).

Capacitance causes the voltage to lag behind the current (Fig. 18), creating a phase shift that is quantified as the angular transformation (θ) of the ratio of reactance to resistance (105). Note that the uniform orientation of cells in a tissue (Fig. 19) can result in anisotropy of electrical properties. That is, impedance will be lower in the longitudinal versus transverse direction of the tissue segment cellular structure (12,18).

The parallel-column model (Fig. 3) must be modified to describe higher frequency applications of bioimpedance in which the capacitive properties of the cell membranes become important. The Cole–Cole plot (Fig. 20) is a useful characterization of the three element RC model that describes the behavior of tissue impedance as a function of frequency (f), impedance (Z), resistance (R), reactance (X_C), and phase angle (ϕ) (103). The real components (R_1 and R_2) can be plotted versus the negative imaginary component of the capacitor (C) with reactance (X_C) in the complex series impedance ($R + jX_C$), with the frequency as a parameter where $j = (\sqrt{-1})$ (15). As the frequency is

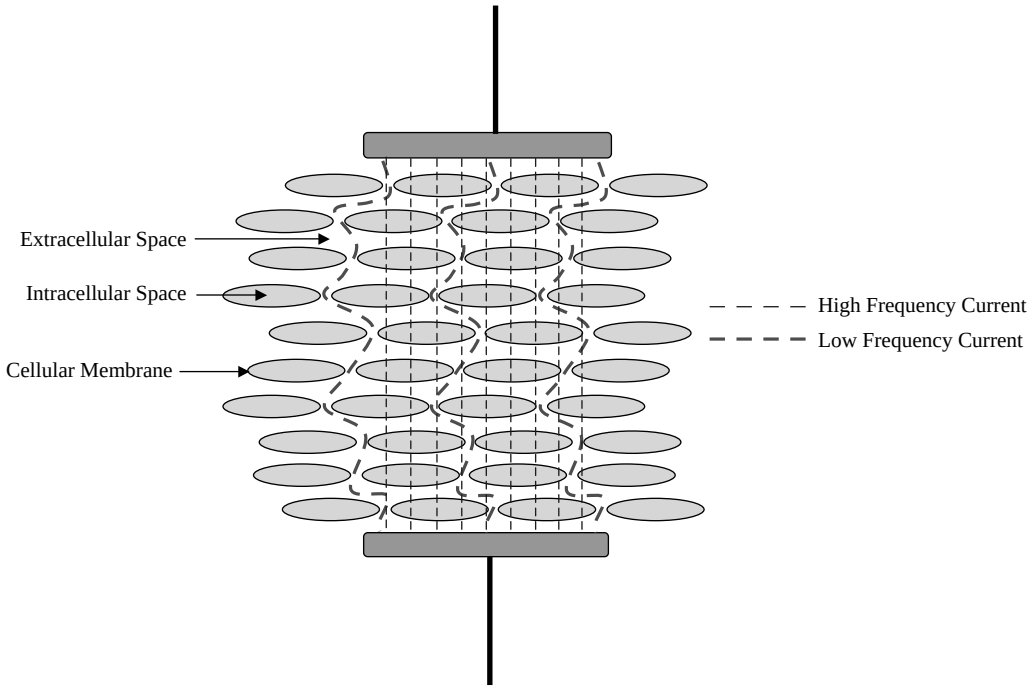


Figure 19. Low and high frequency current distribution is a cellular structure. The low frequency stimulus current flows through the highly conductive extracellular space, whereas the high frequency stimulus current flows through both the extracellular and intracellular space once the reactance of the capacitive cellular membrane is reduced.

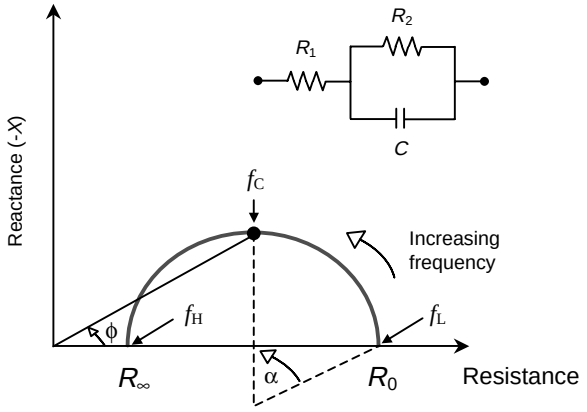


Figure 20. Cole–Cole plot and equivalent tissue impedance circuit. Resistance (abscissa) and reactance (ordinate) plotted as a function of frequency. A three-element electrical equivalent tissue impedance model is shown. At low frequency (f_L), the equivalent circuit is resistive and $R_0 = R_1 + R_2$. As the frequency increases, the phase angle (ϕ) increases until the resistance and reactance are equal at the characteristic frequency of the tissue (f_C). As the frequency increases beyond the characteristic frequency, the reactive element C is reduced to a low impedance and the tissue displays purely resistive properties where $R_\infty = R_1$. The depressed locus at angle α is presumed to represent electrode polarization.

changed between R_0 and R_∞, the impedance will change continually along a curve in the R-X plane. At very low frequencies (f_L), the capacitive component of the system is effectively an open circuit so the reactance is equal to zero and the measured impedance (Z) is purely resistive (R_0). As the frequency increases, reactance (X_C) increases in proportion to resistance causing the phase angle (ϕ) to increase until a maximum angle is reached at the critical (characteristic) frequency (f_C). As shown in Fig. 20, phase angle is positively associated with reactance and negatively associated with resistance (106). Beyond the critical frequency, the reactance begins to decrease in proportion to resistance with increasing frequency and, at very high frequencies (f_H), the capacitive component is essentially short-circuited so the measured impedance is purely resistive at R_∞ (105).

If impedance of a tissue is measure over a broad spectrum, then the resultant impedance Cole–Cole plot can be fit to the three element model or other similar lumped-parameter models. Changes in the model elements can reflect changes in tissue properties due to pathological conditions such as ischemia (see below). In many biologic systems, the center of loci of the plot lies below the real axis and is represented by the angle α, a fixed number between 0 and 1 (2). This behavior can only be modeled by adding an inductive element to the electrical parameter model shown in Fig. 20. However, the physiological interpretation of the inductance is uncertain. Fricke et al. hypothesized that a possible source of this observed inductance might be electrode polarization (107). These investigations demonstrated behavior similar to constant depression angle of electrode polarization. They demonstrated that a frequency-dependent resistance and reactance could mathematically assume a constant depression angle (2). However, the physiologic explanation for $\alpha >$ zero remains controversial. An additional theory related to the origin of the depressed loci is the distribution of time constants in a heterogeneous tissue segment. This distribution could result from variability in cell size or variability in properties of the individual cells (2).

Ischemia Detection. Tissue degradation due to ischemia can alter both the real and reactive components of bioimpedance (40). The dielectric polarization of matter (e.g., myocardial tissue) is given by the dimensionless parameter ε', which is called dielectric permittivity. ε' describes the capacitance increase of a capacitor filled with matter:

$$\varepsilon' = \frac{C}{C_0} \tag{12}$$

where:

C = a capacitor with matter (i.e., cellular structure),
C_0 = vacuum capacitor.

As the dielectric polarization processes are frequency-dependent, they show relaxation phenomena with increasing frequency (108). The relaxation process is defined by the complex dielectric permittivity ε, thus:

$$\varepsilon(\omega) = \varepsilon'(\omega) - i\varepsilon''(\omega) \tag{13}$$

where:

ε' = dielectric permittivity,
ε'' = dielectric loss factor,
$\omega = 2\pi f$,
f = frequency of stimulus current,
i = imaginary unit ($\sqrt{-1}$).

The method of dielectric spectroscopy has been proposed to investigate heart tissue during global ischemia, because the dielectric polarization of matter can be measured by the application of weak electric fields. An electrical circuit model to describe myocardial ischemia, initially developed by Gersing (109) and modified by Schaefer et al. (108), is depicted in Fig. 21. This model can be considered as a variation of the simplified three element physiologic model as shown in Fig. 20. The resistance R_{ext} describes the properties of the extracellular electrolyte, and the resistance R_{int} describes the intracellular cytosol. This model assumes that the transcellular current has to pass the membrane with capacitance C_m and the resistance R_m, through the cytosol, and from cell to cell through the interstitial membranes described by C_{is} or, alternatively, through gap junctions with resistance R_g (108,109). Application of this model enables quantification of the variation of intracellular coupling via gap junctions due to myocardial ischemia (108–111).

The measurement of alterations in impedance spectra with ischemia is often referred to as impedance spectroscopy. Myocardial electrical impedance (MEI), a specific application of impedance spectroscopy, has been shown to

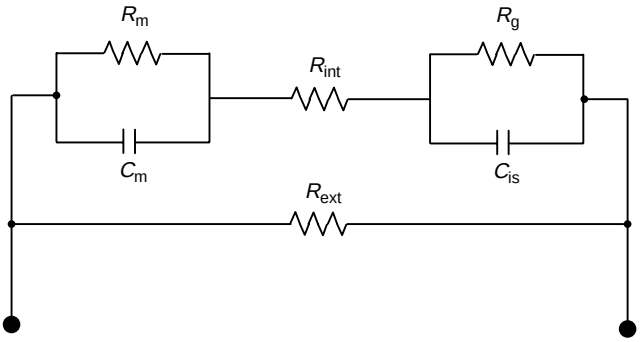


Figure 21. Electronic equivalent model of heart tissue. Electrical elements consist of the intracellular (R_{int}) and extracellular resistance (R_{ext}), cellular membrance capacitance (C_m) and resistance (R_m), cell-to-cell interstitial membrane capacitance (C_{is}), and gap junction resistance (R_g).

identify localized and global myocardial tissue in various disease states in *in vitro* and *in vivo* experimental models (112).

Recently, MEI has been used in conjunction with electrocardiogram (ECG) ST-segment deviations to assess the magnitude of the ischemic region of the myocardium (103,112–116).

Injury currents, secondary to myocardial ischemia result in ST-segment displacements in the ECG of patients with myocardial ischemia (117). Injury currents deriving from resting depolarization in ischemic myocardial cells are associated with slow conduction through the myocardium. The mechanisms by which these injury currents correlate with the impedance spectroscopy alterations in the ischemic myocardial tissue are well described (103,108,109,112–117). Figure 22 depicts a segment of the myocardium with subendocardial and transmural ischemic tissue. Blood flow through the heart is interrupted during ischemia, and the tissue undergoes progressive changes leading to irreversible loss of its viability (108). Transmural ischemia causes ST-segment elevation and subendocardial ischemia causes ST-segment depression (117). The electrocardiographic differences between transmural and subendocardial ischemia are clinically important. Supply ischemia, as occurs following total interruption of flow through a coronary artery supplying a large area of the left ventricle, typically causes ST-segment elevation (63). In contrast, demand ischemia, as occurs during a stress test, begins in the subendocardial regions of the left ventricle and causes ST-segment depression (117).

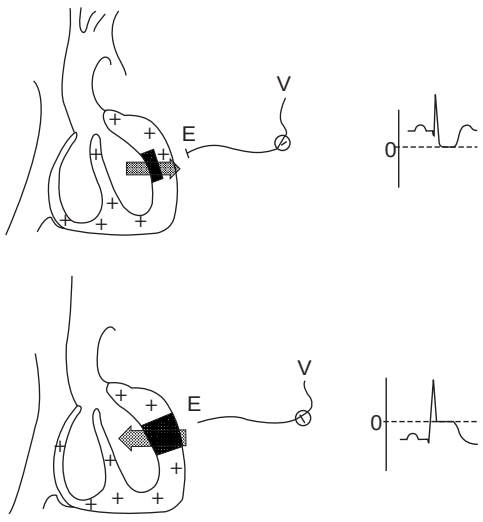


Figure 22. Ischemic regions of myocardial tissue and corresponding ST-segment. Subendocardial ischemia (top) with depressed ST-segment. Transmural ischemia (bottom) with elevated ST-segment (117).

The mechanism by which MEI changes with ischemia is not certain, but may well be associated with ultrastructural changes or cellular biochemical changes that occur in the myocardial tissue similar to those viewed by ST-segment deviations (113). The increase in MEI may result from reductions in the conductive fluid volume in the affected region of the myocardium (113). Gap junctions play a critical role in the propagation of electrical impulse in the heart, and its conductivity has been shown to be reduced and eventually abolished during ischemia and rapidly restored during reperfusion (103). Thus, gap junction closure is a reasonable hypothesis to explain observed impedance changes with ischemia. The intraischemic variation of intracellular and extracellular coupling is one possible explanation for the observed impedance changes of the dielectric frequency spectrum (108).

As MEI correlates with myocardial tissue viability (118,119), the measure has several important potential monitoring applications. Intraoperatively, MEI could be used to detect ischemia in aortic or myocardial tissue during cardiopulmonary bypass surgery as an early indication of damage. Following cardiopulmonary bypass, MEI could be used to assess reperfusion afforded by the new grafts. MEI could also aid in drug titration after cardiac surgery as well as to chronically monitor tissue perfusion with implantable devices such as pacemakers or cardioverter-defibrillators, or with patients whom have received a heart transplant (12,113,120).

In a study performed by Howie et al. (113), acute ischemia was induced in anesthetized dogs via left anterior descending (LAD) coronary artery occlusion for randomly assigned periods of 15, 30, 45, 60, or 120 min. MEI was simultaneously recorded using ventricular pacing leads sutured into the exposed heart tissue. As shown in Fig. 23, MEI increased immediately after LAD coronary artery occlusion and returned to baseline following reperfusion. A statistically significant increase occurred from baseline impedance when compared at 64, 68, 72, 76, and 80 min (113). This intracardiac technique used by Howie et al. suggested other possible applications for MEI with implantable devices and intracardiac pacing/monitoring leads. However, further development in the direction of optimal electrode placement to isolate the targeted tissue region and obtain the highest quality data for diagnosis of tissue alteration is warranted (12).

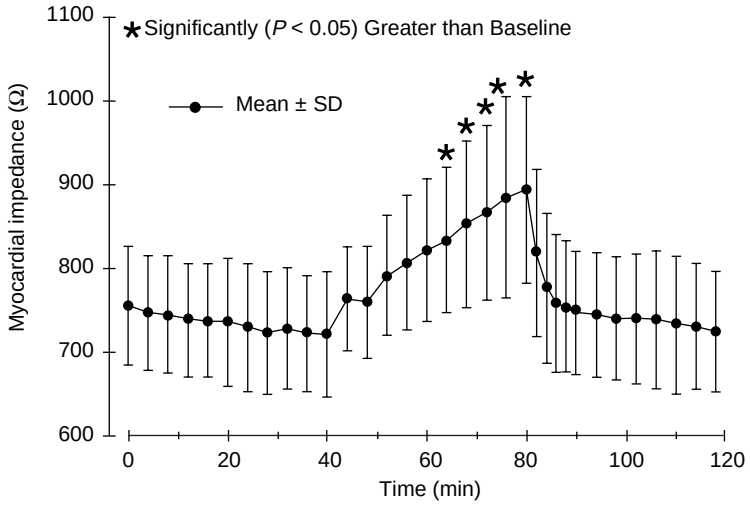


Figure 23. Change in myocardial impedance during LAD coronary artery occlusion (113).

Dialysis. Whole-body bioimpedance spectroscopy has been proposed by several investigators for measuring extracellular (ECW) and intracellular (ICW) water volumes in dialysis patients in order to assess nutritional status and to monitor hydration (121–125). An adequate assessment of body water compartments is crucial in dialysis patients because overhydration and underhydration are often difficult to detect and may result in severe morbidity in this population (126). Despite the continuous progress in the delivery of renal replacement therapy, mortality in patients on maintenance dialysis remains higher than in the general population (127).

During acute volume overload, most of the extra fluid collects in the ECW not the ICW. At very low frequencies, current only penetrates the ECW because the cell membrane acts as a capacitor and the impedance becomes equal to the ECW resistance (see Fig. 19). At very high frequencies, the injected current penetrates both the ECW and the ICW, and the impedance represents the total body water (TBW) resistance (125). Several investigators (128–131) have used single- and multiple-frequency impedance to monitor fluid shifts during hemodialysis. However, when attempting to determine precise fluid volumes from the measured impedance, difficulties occur due to the complex geometry of the human body and electrical inhomogeneity of nonconducting elements such as bone and fat (125). Signal processing methods to account for these aforementioned difficulties are described in the literature (104,124,125,132).

Whole-body bioelectrical impedance measurements typically apply single (e.g., 50 kHz) (133) or multifrequency (e.g., 5 to 1000 kHz) alternating currents applied via cutaneous electrodes placed on the hands and feet with more proximal electrodes uses for voltage measurements (126). The precise method for calculation of body fluid volumes depends on whether the single-frequency or multiple-frequency method is applied. The single-frequency method often uses an empirically derived regression formula to assess TBW, whereas the multiple-frequency method predicts the volume of TBW and ECW from a general mixture theory, assuming specific resistance values for ECW and ICW (104,126,134). Moreover, the contribution of body weight, which is strongly related to ECW and TBW, is greater in the regression approach compared with the mixture approach (126).

Although reliable measurements of fluid content in dialysis patients have been reported (121–131), uncertainty remains regarding the agreement of whole-body bioimpedance in dialysis patients with tracer dilution techniques, which are considered the gold standard methods (126). One explanation for the lack of satisfactory agreement between techniques is that whole-body bioimpedance techniques consider the body as a multiple conductive cylinder model (e.g., arms, legs, trunk) connected in series (Fig. 24). With conductors connected in series, conductors with the smallest cross-sectional area (e.g., extremities) will determine most of the resistance, whereas the component with the largest cross-sectional area (e.g., trunk) will have minimal contribution to the resistance

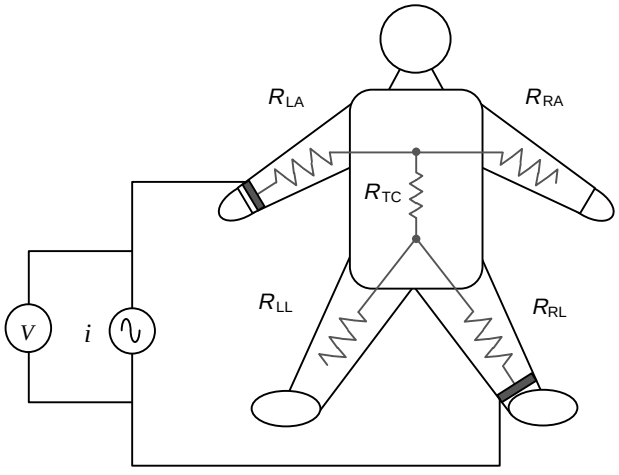


Figure 24. Whole-body impedance measurement technique. Total Conductance (C_T) = Left Arm Conductance ($1/R_{LA}$) + Thor-Thoracic Cavity Conductance ($1/R_{TC}$) + Right Leg Conductance ($1/R_{RL}$).

although it contains a significant amount of body water (126). However, assessment of the sum of segmental bioelectrical impedance analysis measurements, which take into account resistance of the extremities and the trunk independently, have been shown to detect changes in trunk water more accurately (135). A seminal study performed by Patterson et al. (136) used multiple linear regression analysis combining data measured independently from the arms, legs, and trunk correlated with weight change on patients undergoing hemodialysis, gave a correlation coefficient of 0.87, whereas the correlation coefficient from measurements between just the wrist and ankle was 0.64.

Pulmonary Edema Detection. Patients developing pulmonary edema initially accumulate fluid in the interstitial spaces of the lung. As the condition progresses, fluid ultimately accumulates in the alveoli. To accurately measure pulmonary fluid status, the different bioelectric properties of blood, lung tissue, and extravascular fluid must be considered, and an impedance parameter not influenced by the patient's geometry should be used (137). Thus, using a dual-frequency measurement of thoracic impedance, an impedance ratio can be calculated that represents the ratio between intracellular and extracellular water. This ratio, therefore, changes as a result of the fluid shift caused by edema formation. As the low frequency current only passes through the extracellular resistance, the measured low frequency impedance (Z_{LF}) over a specified thoracic length equals the total extracellular resistance. As the frequency is increased, current is divided over the intracellular and extracellular compartments. Therefore, the measured high frequency impedance (Z_{HF}) over a specified thoracic length equals the parallel equivalent of intracellular and extracelluar impedance. Thus, a dual-frequency impedance ratio that represents the intracellular/extracellular impedance fraction can be defined by Z_{HF}/Z_{LF}. As pulmonary fluid accumulates in the extracellular space, the impedance ratio increases (137).

BIBLIOGRAPHY

Cited References

1. Grimnes S, Martinsen O. Bioimpedance and Bioelectricity Basics. London: Academic Press; 2000. p 313–320.
2. Ackmann J, Seitz M. Methods of complex impedance measurements in biologic tissue. CRC Crit Rev Biomed Eng 11(4):281–311.
3. Schwan HP. The bioimpedance field: Some historical observations. Proceedings of the 4th International Conference on Electrical Bioimpedance. 1995; 1–4.
4. Nyboer J, et al. Radiocardiograms: Electrical impedance changes of the heart in relation to electrocardiograms and heart sounds. J Clin Invest 1940; 19:963.
5. Nyboer J. Electrical impedance plethysmography: A physical and physiologic approach to peripheral vascular study. Circulation 1950;2:811–821.
6. Kubicek W, et al. Development and evaluation of an impedance cardiac output system. Aerospace Med 1966;37:1208–1212.
7. Djordjevich L, Sadove M. Basic principles of electrohemodynamics. J Biomed Eng 1981; 3:25–33.
8. Baan J, et al. Continuous measurement of left ventricular volume in animals and humans by conductance catheter. Circulation 1984;70(5):812–823.
9. Min M, Parve T, Kink A. Thoracic impedance as a basis for pacing control. Ann NY Acad Sci 1999;873: 155–166.
10. Ellenbogen K, Wood M. Cardiac Pacing and ICD's. 3rd ed. Malden (MA): Blackwell Science; 2002. p 106–109.
11. Kindermann M, et al. Optimizing the AV delay in DDD pacemaker patients with high degree AV block: Mitral valve doppler versus impedance cardiography. PACE 1997;20(10 pt 1):2453–2462.
12. Steendijk P, et al. The four-electrode resistivity technique in anisotropic media: Theoretical analysis and application on myocardial tissue in vivo. IEEE Trans Biomed Eng 1993;40(11):1138–1147.
13. Nyboer J. Electrorheometric properties of tissues and fluids. Ann NY Acad Sci 1970;170(2):410–420.
14. Geddes LA, Hoff HE. The measurement of physiological events by electrical impedance, a review. Am J Med Electron 1964; 14:16–27.
15. Malmivuo J, Plonsey R. Bioelectromagnetism. In: Principles and Application of Bioelectric and Biomagnetic Fields. United Kingdom: Oxford Univeristy Press; 1995. p 141, 405–419.
16. Hettrick DA, Battocletti JH, Ackmann JJ, Linehan JH, Warltier DC. Effects of physical parameters on the cylindrical model for conductance volume measurement. Ann Biomed Eng 1997;24:126–134.
17. Hettrick D, et al. Finite element model determination of correction factors used for measurement of aortic diameter via conductance. Ann Biomed Eng 1999;27(2): 151–159.
18. Roberts DE, Hersh LT, Scher AM. Influence of cardiac fiber orientation on wavefront voltage, conduction velocity, and tissue resistivity in the dog. Circ Res 1979; 44:701–712.
19. Patterson R, et al. Development of an electrical impedance plethysmography system to monitor cardiac output. Proceedings of the First Annual Rocky Mountain Bioengineering Symposium, 1964: 56–71.
20. Sramek B, Rose D, Miyamoto A. A stroke volume equation with a linear base impedance model and its accuracy, as compared to thermodilution and magnetic flowmeter techniques in humans and animals. Proceedings of the Sixth International Conference on Electrical Bioimpedance, Zadar, Yugoslavia, 1983: 38.
21. Milnor WR. Hemodyanmics. Baltimore: Williams & Wilkins Co.; 1982. p 155.
22. Statistical Bulletin. Metropolitan Life Foundation, January-June 1983. p 64.
23. Quail A, et al. Thoracic resistivity for stroke volume calculation by impedance cardiography. J Appl Physiol 1981;50:191.
24. Albert N, et al. Equivalence of bioimpedance and thermodilution in measuring cardiac output in hospitalized patients with advanced, decompensated chronic heart failure. Am J Crit Care 2004;13:469–479.
25. Van DeWater J, et al. Impedance cardiography: The next vital sign technology? Chest 2003;123: 2028–2033.
26. Drazner M, et al. Comparison of impedance cardiography with invasive hemodynamic measurments in patients with heart failure secondary to ischemic or nonischemic cardiomyopathy. Am J Cardiol 2002;89:993–995.
27. Sageman W, Riffenburgh R, Spiess B. Equivalence of bioimpedance and thermodilution in measuring cardiac index

after cardiac surgery. J Cardiothroracic Vasc Anesth 2002; 16:8–14.
28. Yung G, et al. Comparison of impedance cardiography to direct Fick and thermodilution cardiac output determination in pulmonary arterial hypertension. Congest Heart Fail 2004;10(Suppl 2):7–10.
29. Patterson RP. Possible technique to measure ventricular volume using electrical impedance measurements with an oesophageal electrode. Med Biol Eng Comput 1987;25:677–679.
30. Hettrick DA, et al. Correlation of esophageal conductance measurements with aortic and left ventricular diameters and stroke volume. IEEE Trans Biomed Eng 2000;47:559–564.
31. Patterson R. Handbook of Biomedical Engineering. Bioelectric Impedance Measurements. Boca Raton, FL: CRC Press; 1995. p 1223–1230.
32. Ebert T, et al. The use of thoracic impedance for determining thoracic blood volume changes in man. Aviation Space Environ Med 1986 57:49–53.
33. Mancini R, et al. Cardiac output and contractility indices: Establishing a standard in response to low-to-moderate level exercise in healthy men. Arch Phys Med Rehabil 1979;60:567–573.
34. Levett J, Replogle R. Thermodilution cardiac output: A critical analysis and review of the literature. J Surg Res 1979;27:392–404.
35. Engoren M, Barbee D. Comparison of cardiac output determined by bioimpedance, thermodilution, and the Fick method. Am J Crit Care 2005;14(1):40–45.
36. Imhoff M, Lehner J, Lohlein D. Noninvasive whole-body electrical bioimpedance cardiac output and invasive thermodilution cardiac output in high-risk surgical patients. Crit Care Med 2000;28:2812–2818.
37. Cotter G, et al. Accurate, Noninvasive continuous monitoring of cardiac output by whole body electrical bioimpedance. Chest 2004;125:1431–1440.
38. Raaijmakers E, et al. A meta-analysis of published studies concerning the validity of thoracic impedance cardiography. Ann NY Acad Sci 1999;873: 121–134.
39. Patterson R, Witsoe D. Impedance stroke volume compared with dye and electromagnetic flowmeter values during drug induced inotropic and vascular changes in dogs. Ann NY Acad Sci 1999;873:143–148.
40. Min M, Ollmar S, Gersing E. Electrical impedance and cardiac monitoring: Technology, potential and applications. Int J Bioelectromag 2003;5(1):53–56.
41. Kim D, et al. Origins of the impedance change in impedance cardiography by a three-dimensional finite element model. IEEE Trans Biomed Eng 1988 ;35(12): 993–1000.
42. Wang L, Patterson R. Multiple sources of the impedance cardiogram based on 3-D finite difference human thorax models. IEEE Trans Biomed Eng 1995;42(4): 141–148.
43. Wang X, et al. An impedance cardiography system: A new design. Ann Biomed Eng 1989;17: 535–556.
44. Summers R, Kolb J, Woodward L. Differentiating systolic from diastolic heart failure using impedance cardiography. Academ Emerg Med 1999;6(7):693–699.
45. Lababidi Z, et al. The first derivative thoracic impedance cardiogram. Circulation 1970;41(4): 651–658.
46. Osypka M, Berstein D. Electrophysiologic principles and theory of stroke volume determination by thoracic electrical bioimpedance. Non-invasive monitoring using thoracic bioimpedance. AACN Clin Issues 1999;10(3): 385–399.
47. Christov I. Dynamic powerline interference subtraction from biosignals. J Med Eng Technol 2000;24(4):169–172.
48. Wang X, Sun H, Van DeWater J. An advanced signal processing technique for impedance cardiography. IEEE Trans Biomed Eng 1995;42(2):224–230.
49. Barros A, Yoshizawa M, Yasuda Y. Filtering noncorrelated noise in impedance cardiography. IEEE Trans Biomed Eng 1995;42(3):324–327.
50. Eiken O, Segerhammer P. Elimination of breathing artifacts from impedance cardiograms at rest and during exercise. Med Biol Eng Comput 1988;13–16.
51. Raza S, Patterson R, Wang L. Filtering respiration and low-frequency movement artifacts from the cardiogenic electrical impedance signal. Med Biol Eng Comput 1992; 556–561.
52. Abdelhammed A, et al. Noninvasive hemodynamic profiles in hypertensive subjects. Am J Hypertens 2005;18:51S–59S.
53. Ventura H, Taler S, Strobeck J. Hypertension as a hemodynamic disease: The role of impedance cardiography in diagnostic, prognostic, and therapeutic decision making. Am J Hypertens 2005;18:26S–43S.
54. Alfie J, Galarza C, Waisman G. Noninvasive hemodynamic assessment of the effect of mean arterial pressure on the amplitude of pulse pressure. Am J Hypertens 2005;18:60S–64S.
55. Taler S, Textor S, Augustine J. Resistant hypertension: Comparing hemodynamic management to specialist care. Hypertension 2002;39:982–988.
56. Sharman D, Gomes C, Rutheford J. Improvement in blood pressure control with impedance cardiograph-guided pharmacologic decisions making. Congest Heart Fail 2004;10: 54–58.
57. Eugene M, et al. Assessment of the optimal atrio-ventricular delay in DDD paced patients by impedance plethysmography. Eur Heart J 1989;10:250–255.
58. Ovsyshcher I, et al. Measurements of cardiac output by impedance cardiography in pacemaker patients at rest: Effects of various atrioventricular delays. JACC 1993; 21(3):761–767.
59. Tse H, et al. Impedance cardiography for atrioventricular interval optimization during permanent left ventricular pacing. PACE 2003;26(Pt II):189–191.
60. Wolf GK, Arnold JH. Noninvasive assessment of lung volume: Respiratory inductance plethysmography and electrical impedance tomography. Crit Care Med 2005;33(3): S163–S169.
61. Coulombe N, et al. A parametric model of the relationship between EIT and total lung volume. Physiol Meas 2005;26(4):401–411.
62. Zhang J, Patterson RP. EIT images of ventilation: What contributes to the resistivity changes? Physiol Meas 2005; 26(2):S81–S92.
63. Edd JF, Horowitz L, Rubinsky B. Temperature dependence of tissue impedivity in electrical impedance tomography of cryosurgery. IEEE Trans Biomed Eng 2005;52(4): 695–701.
64. Xiao C, Lei Y. Analytical solutions of electric potential and impedance for a multilayered spherical volume conductor excited by time-harmonic electric current source: Application in brain EIT. Phys Med Biol 2005;750(11): 2663–2674.
65. Clay MT, Ferree TC. Weighted regularization in electrical impedance tomography with applications to acute cerebral stroke. IEEE Trans Med Imag 2002; 21(6):629–637.

66. Zlochiver S, Rosenfeld M, Shimon A. Contactless bio-impedance monitoring technique for brain cryosurgery in a 3D head model. Ann Biomed Eng 2005;33(5):616–625.
67. Geselowitz D. An application of electrocardiographic lead theory to impedance plethysmography. IEEE Trans Biomed Eng 1971;18(1):38–41.
68. Kauppinen P, et al. Application of computer modelling and lead field theory in developing multiple aimed impedance cardiography measurments. J Med Eng Technol 1999;23(5):169–177.
69. Kauppinen P, et al. Lead field theoretical approach in bioimpedance measurements: Towards more controlled measurment sensitivity. Ann NY Acad Sci 1999; 135–142.
70. Wang L, Lahtinen S, Lentz L, et al. Feasibility of using and implantable system to measure thoracic congestion in an ambulatory chronic heart failure canine model. PACE 2005; 28:404–411.
71. Pomerantz M, et al. Transthoracic electrical impedance for the early detection of pulmonary edema. 1969;66:260–268.
72. Fein A, et al. Evaluation of transthoracic electrical impedance in the diagnosis of pulmonary edema. Circulation 1979;60:1156–1160.
73. Gotshall R, Davrath L. Bioelectric impedance as an index of thoracic fluid. Aviation Space Environ Med 1999;70(1):58–61.
74. Yu C, et al. Intrathoracic impedance monitoring in patients with heart failure. Correlation with fluid status and feasibility of early warning preceding hospitalization. Circulation 2005;112:841–848.
75. Steendijk P, et al. Pressure-volume measurements by conductance catheter during cardiac resynchronization therapy. Eur Heart J Suppl 2004;6(Suppl D): D35–D42.
76. Baan J, Van der Velde E, Steendijk P. Ventricular pressure-volume relations in vivo. Eur Heart J 1992;13(Suppl E):2–6.
77. Segers P, et al. Conductance catheter based assessment of arterial input impedance, arterial function, and ventricular-vascular interaction in mice. Am J Physiol Heart Circu Physiol 2005;288:H1157–H1164.
78. Szwarc RS, Laurent D, Allegrini PR, Ball HA. Conductance catheter measurement of left ventricular volume: evidence for nonlinearity within cardiac cycle. Am J Physiol 1995;268: H1490–H1498.
79. Georgakopoulus D, Kass DA. Estimation of parallel conductance by dual-frequency conductance catheter in mice. Am J Physiol Heart Circu Physiol 2000;279:H443–H450.
80. Gawne TJ, Gray KS, Goldstein RE. Estimating left ventricular offset volume using dual frequency conductance catheters. J Appl Physiol 1987;63:872–876.
81. Nicolosi AC, Hettrick DA, Warltier DC. Assessment of right ventricular function in swine using sonomicrometry and conductance. Ann Thorac Surg 1996;61:1281–1387.
82. Schwartzman D, et al. Atrial pacing lead location alters left atrial-ventricular mechanical coupling relationships independent of AV delay in humans: A dual-chamber pressure-volume analysis. Heart Rhythm 2005;2:S85.
83. Hettrick DA, et al. In vivo measurement of real time aortic segmental volume using the conductance catheter. Ann Biomed Eng 1998;26:431–440.
84. Kornet L, et al. Conductance method for the measurement of cross-sectional areas of the aorta. Ann Biomed Eng 1999;27:141–150.
85. Gasparini G, et al. Rate-responsive pacing regulated by cardiac hemodynamics. Europace 2005; 7:234–241.
86. Di Gregorio F, et al. Transvalvular impedance (TVI) recording under electrical and pharmacological cardiac stimulation. PACE 1996;19(Pt.II):1689–1693.
87. Salo R. Application of impedance volume measurement to implantable devices. Int J Bioelectromagn 2003;5(1): 57–60.
88. Fricke H, Morse S. The electrical resistance and capacity of blood for frequencies between 800 and 4.5 MHz. J Gen Physiol 1926;9:153–167.
89. Geddes L, Sadler C. The specific resistance of blood at body temperature. IEEE Trans Biomed Eng 1973;20: 336–339.
90. Hill D, Thompson F. The effect of hematocrit on the resistivity of human blood at 37 degrees celsius and 100 kHz. Med Biol Eng 1975;March:182–186.
91. Mohapatra S, Costeloe K, Hill D. Blood resistivity and its implications for the calculation of cardiac output by the thoracic electrical impedance technique. Intens Care Med 1977;3:63–67.
92. Fuller H. The electrical impedance of plasma: A laboratory simulation of the effect of changes in chemistry. Ann Biomed Eng 1991;19:123–129.
93. Pop G, et al. Catheter based impedance measurements in the right atrium for continuously monitoring hematocrit and estimating blood viscosity changes. An in vivo feasibility study in swine. Biosens Bioelectron 2004;19:1685–1693.
94. Geddes L, Sadler C. The specific resistance of blood at body temperature. Med Biol Eng 1973;11(5):336–339.
95. Fricke H. A mathematical treatment of the electric conductivity and capacity of disperse systems. Physiol Rev 1924;4:575–587.
96. Sigman E, Kolin A, Katz L. Effect of motion on the electrical conductivity of blood. Am J Physiol 1937;118:708.
97. Gollan F, Namon R. Electrical impedance of pulsatile blood flow in rigid tubes and in isolated organs. Ann NY Acad Sci 1970;170(2):568–576.
98. Visser K. Electric properties of flowing blood and impedance cardiography. Ann Biomed Eng 1989;17: 463–473.
99. Peura R, et al. Influence of erythrocyte velocity on impedance plethysmographic measurements. Med Biol Eng Comput 1978;16:147–154.
100. Tanaka K, et al. The impedance of blood: The effects of red cell orientation and its application. Japan J Med Electron Biol Eng 1970;8:14–21.
101. Katsuyuki S, Hiroshi K. Electrical characteristics of flowing blood. IEEE Trans Biomed Eng 1979;26(12):686–695.
102. Lozano A, Rossell J, Pallas-Areny R. Two frequency impedance plethysmograph: Real and imaginary parts. Med Biol Eng Comput 1990;28(1):38–42.
103. Padilla F, et al. Protection afforded by ischemic preconditioning is not mediated by effects on cell-to-cell electrical coupling during myocardial ischemia reperfusion. Am J Heart Circ Physiol 2003;285:H1909–H1916.
104. De Lorenzo A, et al. Predicting body cell mass with bioimpedance by using theoretical methods: A technological review. J Appl Physiol 1997;82(5):1542–1558.
105. Baumgartner R, Chumlea W, Roche A. Bioelectric impedance phase angle and body composition. Am J Clin Nutr 1988; 48:16–23.
106. Barnett A, Bango S. The physiological mechanisms involved in the clinical measure of phase angle. Am J Physiol 1936; 114:366–382.
107. Fricke H. The theory of electrolyte polarization. Phil Mag 1932;14:310.
108. Schaefer M, et al. The complex dielectric spectrum of heart tissue during ischemia. Bioelectrochemistry 2002;58:171–180.

109. Gersing E. Impedance spectroscopy on living tissue for determination of the state of organs. Bioelectrochem Bioenerget 1998;45:145–149.
110. Owens L, et al. Correlation of ischemia-induced extracellular and intracellular ion changes to cell-to-cell electrical uncoupling in isolated blood-perfused rabbit hearts. Circulation 1996;94:10–13.
111. Schafer M, Gebhard Gersing E. Characterization of organ tissue during the transition between life and death: Cardiac and skeletal muscle. Med Biol Eng Comput 1999;37:100–101.
112. Dzwonczyk R, et al. Myocardial electrical impedance responds to ischemia and reperfusion in humans. IEEE Trans Biomed Eng 2004;51(12):2206–2209.
113. Howie M, Dzwonczyk R, McSweeney T. An evaluation of a new two-electrode myocardial electrical impedance monitor for detecting myocardial ischemia. Anesthesia Analgesia 2001;92:12–18.
114. Sezer M, et al. New support for clarifying the relation between ST segment resolution and microvascular function: Degree of ST segment resolution correlates with the pressure derived collateral flow index. Heart 2004;90: 146–150.
115. Leung J, et al. Automated electrocardiograph ST segment trending monitors: Accuracy in detecting myocardial ischemia. Anesthesia Analgesia 1998;87:4–10.
116. Leung J, et al. Electrocardiographic ST-segment changes during acute, severe isovolemic hemodilution in humans. Anesthesiology 2000;93:1004–1010.
117. Katz A. Physiology of the Heart. 2nd ed. New York: Raven Press; 1992. p 609–637.
118. Garrido H, et al. Bioelectrical tissue resistance during various methods of myocardial preservation. Ann Thorac Surg 1983;36:143–151.
119. Gebhard M, et al. Impedance spectroscopy: A method for surveillance of ischemia tolerance of the heart. Thorac Cardiovasc Surg 1987;35:26–32.
120. Mueller J, et al. Electric impedance recording: A noninvasive method of rejection diagnosis. J Extra Corpor Technol 1992;23:49–55.
121. Matthie J, et al. Development of commercial complex bioimpedance spectroscopic system for determining intracellular and extracellular water volumes. Proceedings of the 8th International Conference on Electrical Bioimpedance, Kupio, Finland, 1992. p 203–205.
122. Van Loan M, et al. Use of bioimpednace spectroscopy to determine extracellular fluid, intracellular fluid, total body water, and fat-free mass. In: Human body composition: In vivo methods, models and assessment. New York: Plenum; 1993. p 67–70.
123. Van Marken Lichtenbelt W, et al. Validation of bioelectric impedance measurements as a method to estimate body water compartments. Am J Clin Nutr 1994;60: 159–166.
124. Van Loan M, et al. Fluid changes during pregnancy: Use of bioimpedance spectroscopy. J Appl Physiol 1995;27:152–158.
125. Jaffrin M, et al. Continuous monitoring of plasma, interstitial, and intracellular fluid volumes in dialyzed patients by bioimpedance and hematocrit measurements. ASAIO J 2002;48:326–333.
126. Cox-Reijven P, et al. Role of bioimpedance spectroscopy in assessment of body water compartments in hemodialysis patients. Am J Kidney Dis 2001; 38(4):832–838.
127. Mancini A, et al. Nutritional status in hemodialysis patients and bioimpedance vector analysis. J Renal Nutrition 2003;13(3):199–204.
128. DeVries P, et al. Measurement of transcellular fluid shifts during hemodialysis. Med Biol Eng Comput 1989;27:152–158.
129. Sinning W, et al. Monitoring hemodialysis with bioimpedance: What do we really measure? ASAIO J 1993;39:M584–M589.
130. Scanferla F, et al. On-line bioelectric impedance during hemodialysis: Monitoring of body fluids and cell membrane status. Nephrol Dial Transplant 1990; 5(Suppl 1):167–170.
131. Jaffrin M, et al. Extracellular and intracellular fluid volume during dialysis by multifrequency impedancemetry. ASAIO J 1996;42:M533–M537.
132. Hanai T. Electrical properties of emulsions. In: Emulsions Science. London: Academic Press; 1968. p 354–477.
133. Foley K, et al. Use of single-frequency bioimpedance at 50 kHz to estimate total body water in patients with multiple organ failure and fluid overload. Crit Care Med 1999;27(8):1472–1477.
134. Ward L, Elia M, Cornish B. Potential errors in the application of mixture theory to multifrequency bioelectrical impedance analysis. Physiol Meas 1998;19:53–60.
135. Zhu F, et al. Estimation of body fluid changes during peritoneal dialysis by segmental bioimpedance analysis. Kidney Int 2000;57:299–306.
136. Patterson R, et al. Measurement of body fluid volume change using multisite impedance measurements. Med Biol Eng Comput 1988;26:33–37.
137. Raaijmakers E, et al. Estimation of non-cardiogenic pulmonary edema using dual-frequency electrical impedance. Med Biol Eng Comput 1998;36:461–466.

Further Reading

Cole KS, Cole RH. Dispersion and absorption in dielectrics. J Chem Phys 1941;9:341–351.

See also ELECTROCARDIOGRAPHY, COMPUTERS IN; EXERCISE STRESS TESTING; FLOWMETERS, ELECTROMAGNETIC; IMPEDANCE PLETHYSMOGRAPHY; NEONATAL MONITORING; PHONOCARDIOGRAPHY.

BIOINFORMATICS

ALI ABBAS
LEI LIU
University of Illinois
Urbana, Illinois

INTRODUCTION

The past two decades have witnessed revolutionary changes in biomedical research and biotechnology and an explosive growth of biomedical data. High throughput technologies developed in automated DNA sequencing, functional genomics, proteomics, and metabolomics enable production of such high volume and complex data that the data analysis becomes a big challenge. Consequently, a promising new field, bioinformatics has emerged and is growing rapidly. Combining biological studies with computer science, mathematics, and statistics, bioinformatics develops methods, solutions, and software to discover patterns, generate models, and gain insight knowledge of complex biological systems.

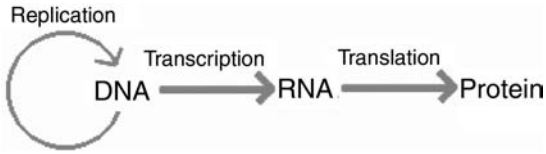


Figure 1. Central dogma of molecular biology.

Before bioinformatics is discussed further, a brief review of the basic concepts in molecular biology, which are the foundations for bioinformatics studies, is provided. The genetic information is coded in DNA sequences. The physical form of a gene is a fragment of DNA. A genome is the complete set of DNA sequences that encode all the genetic information for an organism, which is often organized into one or more chromosomes. The genetic information is decoded through complex molecular machinery inside a cell composed of two major parts, transcription and translation, to produce functional protein and RNA products. These molecular genetic processes can be summarized precisely by the central dogma shown in Fig. 1. The proteins and active RNA molecules combined with other large and small biochemical molecules, organic compounds, and inorganic compounds form the complex dynamic network systems that maintain the living status of a cell. Proteins form complex 3D structures that carry out functions. The 3D structure of a protein is determined by the primary protein sequence and the local environment. The protein sequence is decoded from the DNA sequence of a gene through the genetic codes as shown in Table 1. These codes have been shown to be universal among all living forms on earth.

The high throughput data can be generated at many different levels in the biological system. The genomics data are generated from the genome sequencing that deciphers the complete DNA sequences of all the genetic information in an organism. We can measure the mRNA levels using microarray technology to monitor the gene expression of all the genes in a genome known as transcriptome. Proteome is the complete set of proteins in a cell at a certain stage, which can be measured by high throughput 2D gel electrophoresis and mass spectrometry. We also can monitor all the metabolic compounds in a cell known as metabolome in a high throughput fashion. Many new terms ending with "ome" can be viewed as the complete set of entities in a cell. For example, the "interactome" refers to the complete set of protein-protein interactions in a cell.

Bioinformatics is needed at all levels of high throughput systematic studies to facilitate the data analysis, mining, management, and visualization. But more importantly, the major task is to integrate data from different levels and prior biological knowledge to achieve system-level understanding of biological phenomena. As bioinformatics touches on many areas of biological studies, it is impossible to cover every aspect in a short chapter. In this chapter, the authors will provide a general overview of the field and focus on several key areas, including sequence analysis, phylogenetic analysis, protein structure, genome analysis, microarray analysis, and network analysis.

Sequence analysis often refers to sequence alignment and pattern searching in DNA and protein sequences. This area can be considered classic bioinformatics, which can be dated back to 1960s, long before the word bioinformatics appeared. It deals with the problems such as how to make an optimal alignment between two sequences and how to

Table 1. The Genetic Code

	Second Position				
First Position	T	C	A	G	Third Position
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA *Stop*[end]	TGA *Stop*[end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Stop*[end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

(a)	Sequence 1	C	O	U	N	T	I	N	G	
	Sequence 2	N	T	I	G					
(b)	Sequence 1	C	O	U	N	T	I	N	G	Possible Alignment
	Sequence 2	-	-	-	N	T	I	G	-	(Shifting Sequence 2)
(c)	Sequence 1	C	O	U	N	T	I	N	G	Possible Alignment
	Sequence 2	-	-	-	N	T	-	I	G	(Shifting Sequence 2 and inserting a gap)
(d)	Sequence 1	C	O	U	N	T	I	N	G	Possible Alignment
	Sequence 2	-	-	-	N	T	I	-	G	(Shifting Sequence 2 and inserting a gap)

Figure 2. Possible alignments of two sequences.

search sequence databases quickly with an unknown sequence. Phylogenetic analysis is closely related to sequence alignment. The idea is to use DNA or protein sequence comparison to infer evolution history. The first step in this analysis is to perform multiple sequence alignment. Then, a phylogenetic tree is built based on the multiple alignments. The protein structure analysis involves the prediction of protein secondary and tertiary structures from the primary sequences. So far, the analyses focus on individual sequences or a handful of sequences. The next three areas are involved in system-wide analysis. Genome analysis mainly deals with the sequencing of a complete or partial genome. The problems include genome assembly, gene structure prediction, gene function annotation, and so on. Many techniques of sequence analysis are used in genome analysis, but many new methods were developed for the unique problems. Microarray technologies provide an opportunity for biologists to study the gene expression at a system level. The problems faced in the analysis are completely different from sequence analysis. Many statistical and data mining techniques are applied in the field. Network analysis is another system level study of the biological system. Biological networks can be divided into three categories: metabolic network, protein-protein interaction network, and genetic network. The questions in this area include network modeling, network inference from high throughput data, such as microarray, and network properties study. In the following several sections, the authors will provide a more in-depth discussion of each area.

SEQUENCE ALIGNMENT

Pair-Wise Sequence Alignment

Sequence alignment can be described by the following problem. Given two strings of text, X and Y (which may be DNA or amino acid sequences), find the optimal way of inserting dashes into the two sequences so as to maximize a given scoring function between them. The scoring function depends on both the length of the regions of consecutive dashes and the pairs of characters that are in the same position when gaps have been inserted. The following example from Abbas and Holmes (1) illustrates the idea of sequence alignment for two strings of text. Consider the two sequences, COUNTING and NTIG, shown in Fig. 2a. Figures 2b, 2c, and 2d show possible alignments obtained by inserting gaps (dashes) at different positions in one of the sequences. Figure 2d shows the alignment with the highest number of matching elements. The optimal alignment between two sequences depends on the scoring function that is used. As shall be shown, an optimal sequence alignment for a given scoring function may not be.

Now that what is meant by an optimal sequence alignment has been discussed, the motivation for doing so must be explained. Sequence alignment algorithms can detect mutations in the genome that lead to genetic disease and also provide a similarity score, which can be used to determine the probability that the sequences are evolutionarily related. Knowledge of evolutionary relation between a newly identified protein sequence and a family of protein sequences in a database may provide the first clues about its 3D structure and chemical function. Furthermore, by aligning families of proteins that have the same function (and may have very different sequences), a common subsequence of amino acids can be observed that is key to its particular function. These subsequences are termed protein motifs. Sequence alignment is also a first step in constructing phylogenetic trees that relate biological families of species.

A dynamic programming approach to sequence alignment was proposed by Needleman and Wunsch (2). The idea behind the dynamic programming approach can be explained using the two sequences, CCGAT and CA-AT, of Fig. 3a. If this alignment is broken into two parts (Fig. 3b),

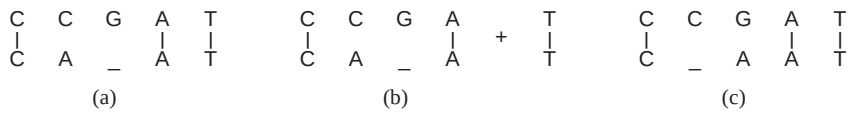


Figure 3. Overview of the dynamic programming approach.

two alignments exist: the left is the alignment of the two sequences CCGA and CA-A, and the right is the alignment of the last elements T-T. If the scoring system is additive, then the score of the alignment of Fig. 3b is the sum of the scores of the four base-alignment on the left plus the score of the alignment of the pair T-T on the right. If the alignment in Fig. 3a is optimal, then the four-base alignment in the left-hand side of Fig. 3b must also be optimal. If this were not the case (e.g., if a better alignment would be obtained by aligning A with G), then the optimal alignment of Fig. 3c would lead to a higher score than the alignment shown in Fig. 3a. The optimal alignment ending at any stage is therefore equal to the total (cumulative) score of the optimal alignment at the previous stage plus the score assigned to the aligned elements at that current stage.

The optimal alignment of two sequences ends with either the last two symbols aligned, the last symbol of one sequence aligned to a gap, or the last symbol of the other sequence aligned to a gap. In the author's analysis, x_i refers to the ith symbol in sequence 1 and y_i refers to the jth symbol in sequence 2 before any alignment has been made. The authors will use the symbol $S(i,j)$ to refer to the cumulative score of the alignment up until symbols x_i and y_j, and the symbol $s(x_i,y_j)$ to refer to the score assigned to matching elements x_i and y_j. The authors will use d to refer to the cost associated with introducing a gap.

1. If the current stage of the alignment matches two symbols, x_i and y_j, then the score, $S(i,j)$, is equal to the previous score, $S(i–1,j–1)$, plus the score assigned to aligning the two symbols, $s(x_i,y_j)$.
2. If the current match is between symbol x_i in sequence 1 and a gap in sequence 2, then the new score is equal to the score up until symbol x_{i-1} and the same symbol y_j, $S(i–1,j)$, plus the penalty associated with introducing a gap, $–d$
3. If the current match is between symbol y_j in sequence 2 and a gap in sequence 1, then the new score is equal to the previous score up until symbol y_{j-1} and the same symbol x_i, $S(i,j–1)$, plus the gap penalty $–d$

The optimal cumulative score at symbols x_i and y_j is:

$$S(i,j) = \max \begin{cases} S(i-1,j-1) + s(x_i,y_j) \\ S(i-1,j) - d \\ S(i,j-1) - d \end{cases}$$

The previous equation determines the new elements at each stage in the alignment by successive iterations from the previous stages. The maximum at any stage may not be unique. The optimal sequence alignment (s) is the one that provides the highest score, which is usually performed using a matrix representation, where the cells in the matrix are assigned an optimal score, and the optimal alignment is determined by a process called trace back (3,4).

The optimal alignment between two sequences depends on the scoring function that is used, which brings the need for a score that is biologically significant and relevant to the phenomenon being analyzed. Substitution matrices present one method of achieving this alignment using a "log-odds" scoring system. One of the first substitution matrices used to score amino acid sequences was developed by Dayhoff et al. (5). Other matrices such as the BLOSUM50 matrix (6) were also developed and use databases of more distantly related proteins.

The Needleman–Wunsch (N–W) algorithm and its variation (3) provide the best *global* alignment for two given sequences. Smith and Waterman (7) presented another dynamic programming algorithm that deals with finding the best *local* alignment for smaller subsequences of two given sequences rather than the best global alignment of the two sequences. The local alignment algorithm identifies a pair of subsegments, one from each of the given sequences, such that no other pair of subsegments exist with greater similarity.

Heuristic Alignment Methods

Heuristic search methods for sequence alignment have gained popularity and extensive use in practice because of the complexity and large number of calculations in the dynamic programming approach. Heuristic approaches search for local alignments of subsegments and use these alignments as "seeds" in which to extend out to longer sequences. The most widely used heuristic search method available today is BLAST (Basic Local Alignment Search Tool) by Altschul et al. (8). BLAST alignments define a measure of similarity called MSP (Maximal Segment Pair) as the highest scoring pair of identical length subsegments from two sequences. The lengths of the subsegments are chosen to maximize the MSP score.

Multiple Sequence Alignments

Multiple sequence alignments are alignments of more than two sequences. The inclusion of additional sequences can improve the accuracy of the alignment, find protein motifs, identify related protein sequences in a database, and predict protein secondary structure. Multiple sequence alignments are also the first step in constructing phylogenetic trees.

The most common approach for multiple alignments is progressive alignment, which involves choosing two sequences and performing a pairwise alignment of the first to the second. The third sequence is then aligned to the first and the process is repeated until all the sequences are aligned. The score of the multiple alignment is the sum of scores of the pairwise alignments. Pairwise dynamic programming can be generalized to perform multiple alignments using the progressive alignment approach; however, it is computationally impractical even when only a few sequences are involved (9). The sensitivity of progressive alignment was improved for divergent protein sequences using CLUSTAL-W (10) (available at http://clustalw.genome.ad.jp/).

Many other approaches to sequence alignment have been proposed in the literature. For example, a Bayesian

approach was suggested for adaptive sequence alignments (11,12). The data that is now available from the human genome project has suggested the need for aligning whole genome sequences where large-scale changes can be studied as opposed to single-gene insertions, deletions, and nucleotide substitutions. MuMMer (12) follows this direction and performs alignments and comparisons of very large sequences.

PHYLOGENETIC TREES

Biologists have long built trees to classify species based on morphological data. The main objectives of phylogenetic tree studies are (*1*) to reconstruct the genealogical ties between organisms and (*2*) to estimate the time of divergence between organisms since they last shared a common ancestor. With the explosion of genetic data in the last few years, tree building has become more popular, where molecular-based phylogenetic studies have been used in many applications, such as the study of gene evolution, population subdivisions, analysis of mating systems, paternity testing, environmental surveillance, and the origins of diseases that have transferred species.

From a mathematical point of view, a phylogenetic tree is a rooted binary tree with labeled leaves. A tree is binary if each vertex has either one or three neighbors. A tree is rooted if a node, R, has been selected and termed the root. A root represents an ancestral sequence from which all other nodes descend. Two important aspects of a phylogenetic tree are its topology and branch length. The topology refers to the branching pattern of the tree, and the branch length is used to represent the time between the splitting events (mutations). Figure 4a shows a rooted binary tree with six leaves. Figure 4b shows all possible distinct rooted topologies for a tree with three leaves.

The data that is used to construct trees is usually in the form of contemporary sequences and is located at the leaves. For this reason, trees are represented with all their leaves "on the ground level" rather than at different levels.

The tree-building analysis consists of two main steps. The first step, estimation, uses the data matrix to produce a tree, $\tilde{T}$, that estimates the unknown tree, T. The second step provides a confidence statement about the estimator $\tilde{T}$, which is often performed by bootstrapping methods.

Tree-building techniques can generally be classified into one of four types: distance-based methods, parsimony methods, maximum likelihood methods, and Bayesian methods. For a detailed discussion of each of these methods, see Li (13).

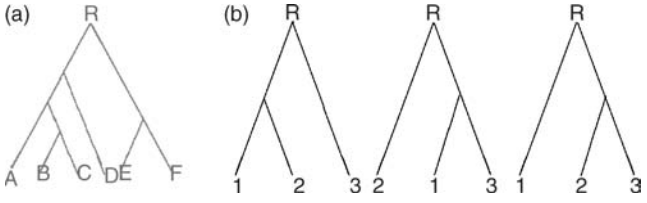


Figure 4. (a) Rooted tree with six leaves. (b) All possible topologies for three leaves.

Tree-building methods can be compared using several criteria such as accuracy (which method gives the true tree, T, when we know the answer?), consistency (when the number of characters increases to infinity, do the trees provided by the estimator converge to the true tree?), efficiency (how quickly does a method converge to the correct solution as the data size increases?), and robustness (is the method stable when the data does not fulfill the necessary assumptions?). To clarify some of these issues, read Holmes (14), where a geometric analysis of the problem is provided and these issues are further discussed.

The second part of the tree-building analysis is concerned with how close we believe the estimated tree is to the true tree. This analysis builds on a probability distribution on the space of all trees. The difficult part of this problem is that, exponentially, many possible trees exist. A nonparametric approach using a multinomial probability model on the whole set of trees would not be feasible as the number of trees is (2N-3)!!. The Bayesian approach defines parametric priors on the space of trees, and then computes the posterior distribution on the same subset of the set of all trees. This analysis enables confidence statements in a Bayesian sense (15).

PROTEIN FOLDING, SIMULATION, AND STRUCTURE PREDICTION

The main motivation for this study is that the structure of a protein greatly influences its function. Knowledge of protein structure and function can help determine the chemical structure of drugs needed to reverse the symptoms that develop due to its malfunction.

The structure of a molecule consists of atoms connected together by bonds. The bonds in a molecular structure contribute to its overall potential energy. The authors shall neglect all quantum mechanical effects in the following discussion and consider only the elements that contribute largely to the potential energy of a structure [as suggested by Levitt and Lifson (16)].

1. **Pair Bonds:** A bond that exists between atoms physically connected by a bond and separated by a distance b. It is like a spring action where energy is stored above and below an equilibrium distance, b_0. The energy associated with this bond is $U(b) = \frac{1}{2}K_b(b - b_0)^2$, where b_0 can be determined from X rays and K_b can be determined from spectroscopy.
2. **Bond Angles:** This bond exists when an angular deviation from an equilibrium angle, θ_0, occurs between three atoms. The bond angle energy associated with the triplet is $U(\theta) = \frac{1}{2}K_\theta(\theta - \theta_0)^2$.
3. **Torsion Angles:** This bond exists when a torsion angle, ϕ, exists between the first and fourth atoms on the axis of the second and third atoms. The energy associated with this bond is $U(\phi) = K_\phi(1 - \cos(n\phi + \delta))$, where θ is an initial torsion angle.
4. **Nonbonded pairs:** Bonds also exist between atoms that are not physically connected in the structure. These bonds include:

a. Van der Waal forces, which exist between nonbonded pairs and contribute to energy, $U(r) = \varepsilon[(\frac{r_0}{r})^{12} - 2(\frac{r_0}{r})^6]$, r_0 is an equilibrium distance and ε a constant.
b. Electrostatic interactions, which contribute to an energy of $U(r) = \alpha\frac{q_i q_j}{r}$; and
c. Hydrogen bonds, which result from van Der Waals forces and the geometry of the system, and contribute to the potential energy of the structure.

The total potential energy function of a given structure can thus be determined by the knowledge of the precise position of each atom. The three main techniques that are used for protein structure prediction are homology (comparative modeling), fold recognition and threading, and *ab initio* folding.

Homology or Comparative Modeling. Comparative modeling techniques predict the structure of a given protein sequence based on its alignment to one or more protein sequences of known structure in a protein database. The approach uses sequence alignment techniques to establish a correspondence between the known structure "template" and the unknown structure. Protein structures are archived for public use in an Internet-accessible database known as the Protein Data Bank (http://www.rcsb.org/pdb/) (17).

Fold Recognition and Threading. When the two sequences exhibit less similarity, the process of recognizing which folding template to use is more difficult. The first step, in this case, is to choose a structure from a library of templates in the protein databank, called fold recognition. The second step "threads" the given protein sequence into the chosen template. Several computer software programs are available for protein structure prediction using the fold recognition and threading technique such as PROSPECT (18).

Ab Initio (New Fold) Prediction. If no similarities exist with any of the sequences in the database, the *ab initio* prediction method is used. This method is one of the earliest structure prediction methods, and uses energy interaction principles to predict the protein structure (16,19,20). Some of these methods include optimization where the objective is to find a minimum energy structure (a local minimum in the energy landscape has zero forces acting on the atoms and is therefore an equilibrium state).

Monte Carlo sampling is one of the most common techniques for simulating molecular motion. The algorithm starts by choosing an initial structure, A, with potential energy, $U(A)$. A new structure, B, is then randomly generated. If the energy of the new structure is less than that of the old structure, the new structure is accepted. If the energy of the new structure is higher than the old structure, then we generate a random number, RAND, from a uniform distribution U(0,1). The new structure is accepted if $e - \frac{\Delta E}{\mathrm{KT}} > RAND$, where $\Delta E = E_B - E_A$ is the difference in energy levels, K is Boltzman's constant, and T is the temperature in kelvins. Otherwise, the new structure is rejected. Another random structure is then generated (either from the new accepted structure or from the old structure if the first one was rejected) and the process is repeated until some termination condition is satisfied (e.g., the maximal number of steps has been achieved).

Another type of analysis uses molecular dynamics uses equations of motion to trace the position of each atom during folding of the protein (21). A single structure is used as a starting point for these calculations. The force acting on each atom is the negative of the gradient of the potential energy at that position. Accelerations, a_i, are related through masses, m_i, to forces, F_i, via Netwon's second law ($F_i = m_i a_i$). At each time step, new positions and velocities of each of the atoms are determined by solving equations of motion using the old positions, old velocities, and old accelerations. Beeman (22) showed that new atomic positions and velocities could be determined by the following equations of motion

$$x(t+\Delta t) = x(t) + v(t)\Delta t + [4a(t) - a(t+\Delta t)]\frac{(\Delta t)^2}{6}$$

$$v(t+\Delta t) = v(t) + [2a(t+\Delta t) + 5a(t) - a(t-\Delta t)]\frac{\Delta t}{6}$$

where $x(t)$ = position of the atom at time t, $v(t)$ = velocity of the atom at time t, $a(t)$ = acceleration at time t, and Δt = time step in the order of 10^{-15} s for the simulation to be accurate.

In 1994, the first large-scale experiment to assess protein structure prediction methods was conducted. This experiment is known as CASP (Critical Assessment of techniques for protein Structure Prediction). The results of this experiment were published in a special issue of *Proteins* in 1995. Further experiments were developed to evaluate the fully automatic web servers for fold recognition. These experiments are known as CAFASP (Critical Assessment of Fully Automated Structure Prediction). For a discussion on the limitations, challenges, and likely future developments on the evaluation of the field of protein folding and structure prediction, the reader is referred to Bourne (23).

GENOME ANALYSIS

Analysis of completely sequenced genomes has been one of the major driving forces for the development of the bioinformatics field. The major challenges in this area include genome assembly, gene prediction, function annotation, promoter region prediction, identification of single nucleotide polymorphism (SNP), and comparative genomics of conserved regions. For a genome project, one must ask several fundamental questions: How can we put the whole genome together from many small pieces of sequences? where are the genes located on a chromosome? and what are other features we can extract from the completed genomes?

Genome Assembly

The first problem is pertaining to the genome mapping and sequence assembly. During the sequencing process, large DNA molecules with millions of base pairs, such as a

human chromosome, are broken into smaller fragments ($\sim$ 100 kb) and cloned into vector such as bacterial artificial chromosome (BAC). These BAC clones can be tiled together by physical mapping techniques. Individual BACs can be further broken down into smaller random fragments of 1–2 kb. These fragments are sequenced and assembled based on overlapping fragments. With more fragments sequenced, enough overlaps will exist to cover most of the sequence. This method is often referred as "shotgun sequencing". Computer tools were developed to assemble the small random fragments into large contigs based on the overlapping ends among the fragments using similar algorithms as the ones used in the basic sequence alignment. The widely used ones include PHRAP/Consed (24,25) and CAP3 (26). Most of the prokaryotic genomes can be sequenced directly by the shotgun sequencing strategy with special techniques for gap closure. For large genomes, such as the human genome, two strategies exist. One is to assemble large contigs first and then tile together the contigs based on the physical map to form the complete chromosome (27). Another strategy is called Whole Genome Shotgun Sequencing (WGS) strategy, which assemble the genome directly from the shotgun sequencing data in combination with mapping information (28). WGS is a faster strategy to finish a large genome, but the challenge of WGS is how to deal with the large number of repetitive sequences in a genome. Nevertheless, WGS has been successfully used in completing the *Drosophila* and human genomes (29,30).

Genome Annotation

The second problem is related to deciphering the information coded in a genome, which is often called genome annotation. The process includes the prediction of gene structures and other features on a chromosome and the function annotation of the genes. Two basic types of genes exist in a genome: RNA genes and protein encoding genes. RNA genes produce active RNA molecules such as ribosomal RNA, tRNA, and small RNA. The majority of genes in a genome are protein encoding genes. Therefore, the big challenge is how to find the protein encoding region in a genome. The simplest way to search for a protein encoding region is to search for open reading frames (ORF), which is a contiguous set of codons between two stop codons. Six possible reading frames for a given DNA sequence exist, three of which start at the first, second, and third base. The other three reading frames are at the complementary strand. The longest ORFs between the start codon and the stop codon in the same reading frame provide good, but not sufficient, evidence of a protein encoding region. Gene prediction is generally easier and more accurate in prokaryotic than eukaryotic organisms due to the intron/exon structure in eukaryote genes. Computational methods of gene prediction based on the Hidden Markov Model (HMM) have been quite successful, especially in prokaryote genome. These methods involve training a gene model to recognize genes in a particular organism. As a result of the variations in codon usage, a model must be trained for each new genome. In a prokaryote genome, genes are packed densely with relatively short intergenic sequences. The model reads through a sequence with unknown gene composition and find the regions flanked by start and stop codons. The codon composition of a gene is different from that of an intergenic region and can be used as a discriminator for gene prediction. Several software tools, such as GeneMark (31) and Glimmer (32) are widely used HMM methods in prokaryotic genome annotation. Similar ideas are also applied to eucaryote gene prediction. As a result of the intron/exon structure, the model is much more complex with more attention on the boundary of intron and exon. Programs such as GeneScan (33) and GenomeScan (34) are HMM methods for eukaryote gene prediction. Neural network-based methods have also been applied in eukaryote gene prediction, such as Grial (35). Additional information for gene prediction can be found using expressed sequence tags (ESTs), which are the sequences from cDNA libraries. As cDNA is derived from mRNA, a match to an EST is a good indication that the genomic region encodes a gene. Functional annotation of the predicted genes is another major task in genome annotation. This process can be also viewed as gene classification with different functional classification systems such as protein families, metabolic pathways, and gene ontology. The simplest way is to infer annotation from the sequence similarity to a known gene (e.g., BLAST search against a well-annotated protein database such as SWISS-PROT). A better way can be a search against protein family databases [e.g., Pfam (36)], which are built based on profile HMMs. The widely used HMM alignment tools include HMMER (37) and SAM (38). All automated annotation methods can produce mistakes. More accurate and precise annotation requires manual checking and a combination of information from different sources.

Besides the gene structures, other features such as promoters can be better analyzed with a finished genome. In prokaryotic organisms, genes involved in the same pathway are often organized in an operon structure. Finding operons in a finished genome provides information on the gene regulation. For eukaryotic organisms, the completed genomes provide upstream sequences for promoter search and prediction. Promoter prediction and detection has been a very challenging bioinformatics problem. The promoter regions are the binding sites for transcription factors (TF). Promoter prediction is to discover the sequence patterns that are specific for TF binding. Different motif finding algorithms have been applied including scoring matrix method (39), Gibbs sampling (40), and Multiple EM for Motif Elicitation (MEME) (41). The results are not quite satisfactory. Recent studies using comparative genomics methods on the problem have produced some promising results and demonstrated that the promoters are conserved among closely related species (42). In addition, microarray studies can provide additional information for promoter discoveries (see the section on microarray analysis).

Comparative Genomics

With more and more genomes being completely sequenced, comparative analysis becomes increasingly valuable and provides more insights of genome organization and

evolution. One comparative analysis is based on the orthologous genes, called clusters of orthologous groups (COG) (43). Two genes from two different organisms are considered orthologous genes if they are believed to come from a common ancestor gene. Another term, paralogous genes, refers to genes in one organism and are related to each other by gene duplication events. In COG, proteins from all completed genomes are compared. All matching proteins in all the organisms are identified and grouped into orthologous groups by speciation and gene duplication events. Related orthologous groups are then clustered to form a COG that includes both orthologs and paralogs. These clusters correspond to classes of functions. Another type of comparative analysis is based on the alignment of the genomes and studies the gene orders and chromosomal rearrangements. A set of orthologous genes that show the same gene order along the chromosomes in two closely related species is called a synteny group. The corresponding region of the chromosomes is called synteny blocks (44). In closely related species, such as mammalian species, the gene orders are highly conserved. The gene orders are changed by chromosomal rearrangements during evolution including the inversion, translocation, fusion, and fission. By comparing completely sequenced genomes, for example, human and mouse genomes, we can reveal the rearrangement events. One challenging problem is to reconstruct the ancestral genome from the multiple genome comparisons and estimate the number and types of the rearrangements (45).

MICROARRAY ANALYSIS

Microarray technologies allow biologists to monitor genome-wide patterns of gene expression in a high throughput fashion. Gene expression refers to the process of transcription. Gene expression for a particular gene can be measured as the fluctuation of the amount of messenger RNA produced from the transcription process of that gene in different conditions or samples.

DNA microarrays are typically composed of thousands of DNA sequences, called probes, fixed to a glass or silicon substrate. The DNA sequences can be long (500–1500 bp) cDNA sequences or shorter (25–70 mer) oligonucleotide sequences. The probes can be deposited with a pin or piezoelectric spray on a glass slide, known as spotted array technology. Oligonucleotide sequences can also be synthesized *in situ* on a silicon chip by photolithographic technology (i.e., Affymetrix GeneChip). Relative quantitative detection of gene expression can be carried out between two samples on one array (spotted array) or by single samples comparing multiple arrays (Affymetrix GeneChip). In spotted array experiments, samples from two sources are labeled with different fluorescent molecules (Cy3 and Cy5) and hybridized together on the same array. The relative fluorescence between each dye on each spot is then recorded and a composite image may be produced. The relative intensities of each channel represent the relative abundance of the RNA or DNA product in each of the two samples. In Affymetrix GeneChip experiments, each sample is labeled with the same dye and hybridized to different arrays. The absolute fluorescent values of each spot may then be scaled and compared with the same spot across arrays. Figure 5 gives an example of a composite image from one spotted array.

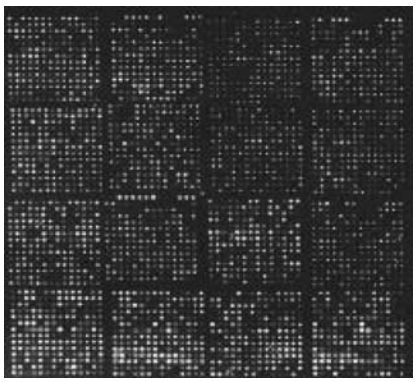


Figure 5. An image from a spotted array after laser scanning. Each spot on the image represents a gene and the intensity of a spot reflects the gene expression.

Microarray analyses usually include several steps including: image analysis and data extraction, data quantification and normalization, identification of differentially expressed genes, and knowledge discovery by data mining techniques such as clustering and classification. Image analysis and data extraction is fully automated and mainly carried out using a commercial software package or a freeware depending on the technology platforms. For example, Affymetrix developed a standard data processing procedure and software for its GeneChips (for detailed information, see http://www.affymetrix.com); GenePix is widely used image analysis software for spotted arrays. For the rest of the steps, the detailed procedures may vary depending on the experiment design and goals. We will discuss some of the procedures below.

Statistical Analysis

The purpose of normalization is to adjust for systematic variations, primarily for labeling and hybridization efficiency, so that the true biological variations can be discover as defined by the microarray experiment (46,47). For example, as shown in the self-hybridization scatter plot (Fig. 6) for a two-dye spotted array, variations (dye bias) between dyes is obvious and related to spot intensities. To correct the dye bias, one can apply the following model:

$$\log_2(R/G) \rightarrow \log_2(R/G) - c(A)$$

where R and G are the intensities of the dyes; A is the signal strength ($\log_2(R^*G)/2$); M is the logarithm ratio ($\log_2(R/G)$); $c(A)$ is the locally weighted polynomial regression (LOWESS) fit to the MA plot (48,49).

After correction of systematic variations, we want to determine which genes are significantly changed during the experiment and to assign appropriately adjusted p values to the genes. For each gene, we wish to test the null hypothesis that the gene is not differentially expressed. The P value is the probability of finding a result

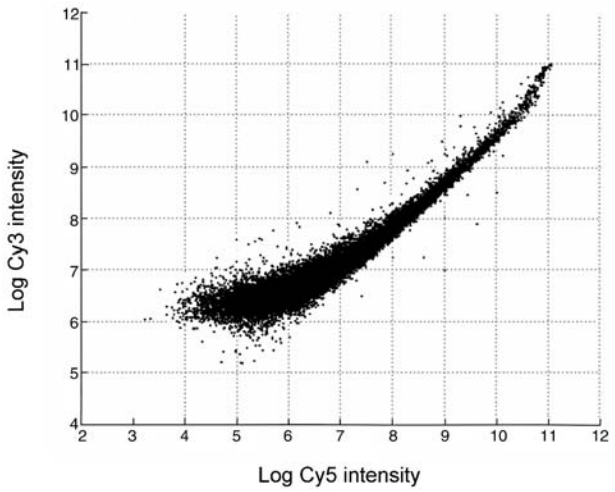


Figure 6. Self-hybridization scatter plot. The y axis is the intensity from one dye; the x axis is the intensity from the other dye. Each spot is a gene.

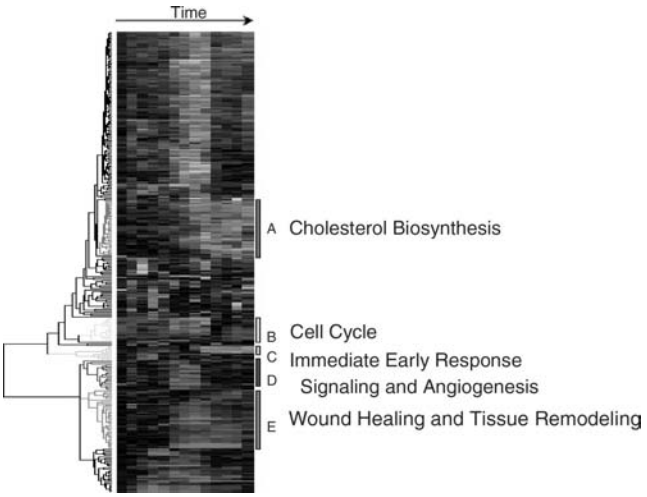


Figure 7. Hierarchical clustering of microarry data. Rows are genes. Columns are RNA samples at different time points. Values are the signals (expression levels) that are represented by the color spectrum. Green represents down-regulation whereas red represents up-regulation. The color bars beside the dendrogram show the clusters of genes that exhibit similar expression profiles (patterns). The bars are labeled with letters and description of possible biological processes involving the genes in the clusters. [Reprinted from Eisen et al. (58).]

by chance. If P value is less than a cut-off (e.g., 0.05), one would reject the null hypothesis and state that the gene is differentially expressed (50). Analysis of variance (ANOVA) is usually used to model the factors for a particular experiment. For example,

$$\log(m_{ijk}) = \mu + A_i + D_j + V_k + \varepsilon_{ijk}$$

where m_{ijk} is the ratio of intensities from the two dye-labeled samples for a gene; μ is the mean of ratios from all replicates; A is the effect of different arrays; D is the dye effects; and V is the treatment effects (51). Through F test, it will be determined if the gene exhibits differential expression between any V_k. For a typical microarray, thousands of genes exist. We need to perform thousands of tests in an experiment at the same time, which introduce the statistical problem of multiple testing and adjustment of p value. False discovery rate (FDR) (52) has been commonly adopted for this purpose.

For Affymetrix GeneChips analysis, even though the basic steps are the same as spotted microarrays, because of the difference in technology, different statistical methods were developed. Besides the statistical methods provided by Affymetrix, several popular methods are packaged into software such as dChip (53) and RMA (54) in Bioconductor (http://www.bioconductor.org). With rapid accumulation of microarray data, one challenging problem is how to compare microarray data across different technology platforms. Some recent studies on data agreements have provided some guidance (55–57).

Clustering and Classification

Once a list of significant genes is obtained from the statistical test, different data mining techniques would be applied to find interesting patterns. At this step, the microarray dataset is organized as a matrix. Each column represents a condition; each row represents a gene. An entry is the expression level of the gene under the corresponding condition. If a set of genes exhibit the similar fluctuation under all of the conditions, it may indicate that these genes are co-regulated. One way to discover the co-regulated genes is to cluster genes with similar fluctuation patterns using various clustering algorithm. Hierarchical clustering was the first clustering method applied to the problem (58). The result of hierarchical clustering forms a 2D dendrogram as shown in Fig. 7. The measurement used in the clustering process can be either a similarity, such as Pearson's correlation coefficient, or a distance, such as Euclidian distance.

Many different clustering methods have been applied later on, such as k means (59), self-organizing map (60), and support vector machine (61). Another type of microarray study involves classification techniques. For example, we can use the gene expression profile to classify cancer types. Golub et al. (62) first reported using classification techniques to classify two different types of leukemia as shown in Fig. 8. Many commercial software packages (e.g., GeneSpring and Spotfire) offer the use of these algorithms for microarray analyses.

COMPUTATIONAL MODELING AND ANALYSIS OF BIOLOGICAL NETWORKS

The biological system is a complex system involving hundreds of thousands of elements. The interaction among the elements forms an extremely complex network. With the development of high throughput technologies in functional genomics, proteomics, and metabolomics, one can start looking into the system-level mechanisms governing the interactions and properties of biological networks. Network modeling has been used extensively

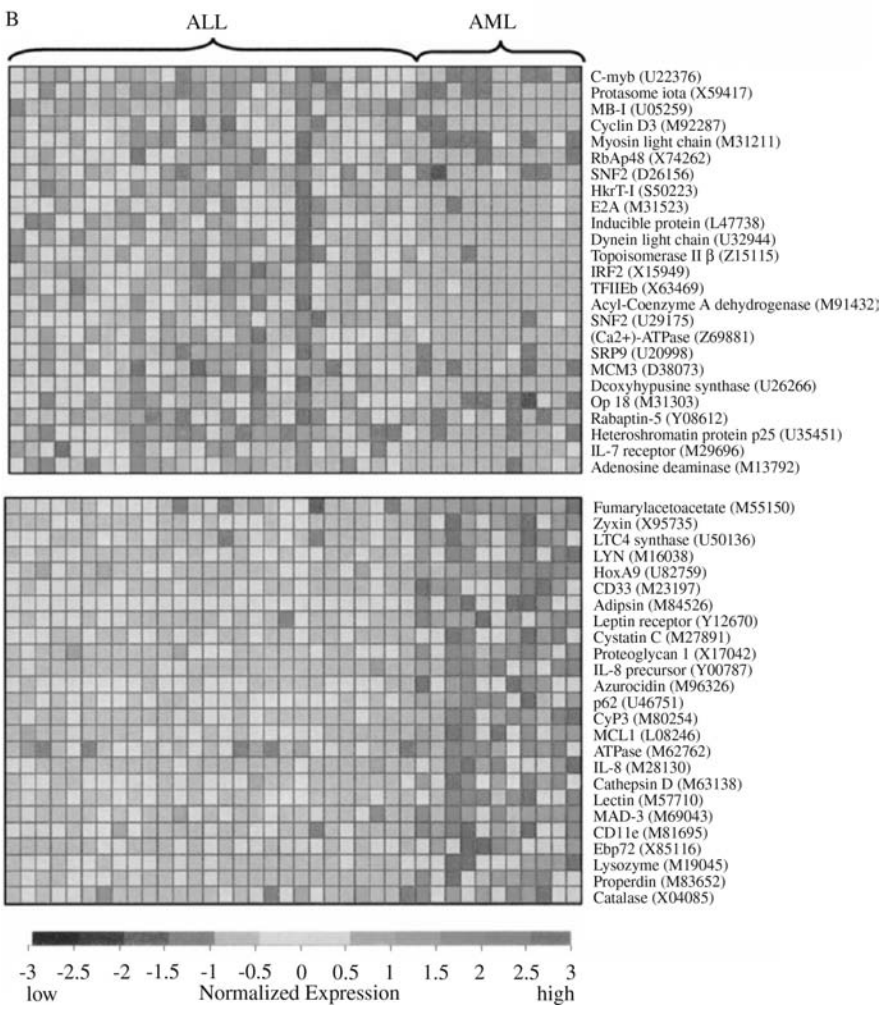


Figure 8. An example of microarray classification. Genes distinguishing acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). The 50 genes most highly correlated with the ALL-AML class distinction are shown. Each row corresponds to a gene, with the columns corresponding to expression levels in different samples. Expression levels for each gene are normalized across the samples such that the mean is 0 and the SD is 1. The scale indicates SDs above or below the mean. The top panel shows genes highly expressed in ALL, the bottom panel shows genes more highly expressed in AML. [Reprinted from Golub et al. (62).]

in social and economical fields for many years (63). Many methods can be applied to biological network studies.

The cellular system involves complex interactions between proteins, DNA, RNA, and smaller molecules and can be categorized in three broad subsystem, metabolic network or pathway, protein network, and genetic or gene regulatory network. *Metabolic network* represents the enzymatic processes within the cell, which provide energy and building blocks for cells. It is formed by the combination of a substrate with an enzyme in a biosynthesis or degradation reaction. Considerable information about metabolic reactions has been accumulated through many years and organized into large databases, such as KEGG (64), EcoCyc (65), and WIT (66). *Protein network* refers to the signaling networks where the basic reaction is between two proteins. Protein-protein interactions can be determined systematically using techniques such as yeast two-hybrid system (67) or derived from the text mining of literatures (68). *Genetic network or regulatory network* refers to the functional inference of direct causal gene interactions (69). One can conceptualize gene expression as a genetic feedback network. The network can be inferred from the gene expression data generated from microarray or proteomics studies in combination with computation modeling.

Metabolic network is typically represented as a graph with the vertex being all the compounds (substrates) and the edges being reactions linking the substrates. With such representation, one can study the general properties of the metabolic network. It has been shown that metabolic network exhibits typical property of small world or scale-free network (70,71). The distribution of compound connectivity follows a power law as shown in Fig. 9. Nodes serving as hubs exist in the network. Such property makes the network quite robust to random deletion of nodes, but vulnerable to selected deletion of nodes. For example, deletion of hub nodes will cause the network collapse very quickly. A recent study also shows that the metabolic network can be organized in modules based on the connectivity. The connectivity is high within modules, but low between modules (72).

Flux analysis is another important aspect in metabolic network study. Building on the stoichiometric network analysis, which only uses the well-characterized network topology, the concept of elementary flux modes was introduced (73,74). An elementary mode is a minimal set of enzymes that could operate at steady state, with the

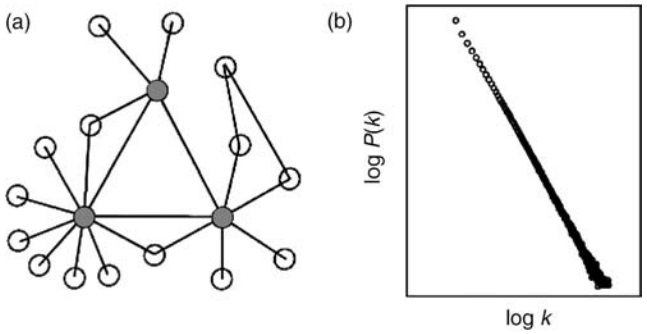


Figure 9. a. In the scale-free network, most nodes have only a few links, but a few nodes, called hubs (filled circle), have a very large number of links. b. The network connectivity can be characterized by the probability, $P(k)$, that a node has k links. $P(k)$ for a scale-free network has no well-defined peak, and for large k, it decays as a power-law, $P(k) \approx k^{-\gamma}$, appearing as a straight line with slope -γ on a log–log plot. [Reprinted from Jeong et al. (70).]

enzymes weighted by the relative flux they need to carry out the mode to function. The total number of elementary modes for given conditions has been used as a quantitative measure of network flexibility and as an estimate of fault-tolerance (75,76).

A system approach to model regulatory networks is essential to understand their dynamics. Recently, several high-level models have been proposed for the regulatory network including Boolean models, continuous systems of coupled differential equations, and probabilistic models. *Boolean networks* assume that a protein or a gene can be in one of two states, active or inactive, represented by 1 or 0. This binary state varies in time and depends on the state of the other genes and proteins in the network through a discrete equation:

$$X_i(t+1) = F_i[X_1(t), \ldots, X_N(t)] \qquad (4)$$

Thus, the function F_i is a Boolean function for the update of the ith element as a function of the state of the network at time t (69). Figure 10 gives a simple example.

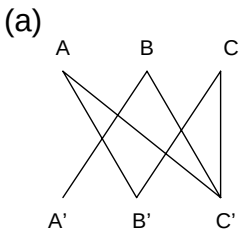
Gene expression patterns contain much of the state information of the genetic network and can be measured experimentally. We are facing the challenge of inferring or reverse engineering the internal structure of this genetic network from measurements of its output. Genes with similar temporal expression patterns may share common genetic control processes and may, therefore, be related functionally. Clustering gene expression patterns according to a similarity or distance measure is the first step toward constructing a wiring diagram for a genetic network (78).

Differential equations can be an alternative model to the Boolean network and applied when the state variables X are continuous and satisfy a system of differential equations of the form

$$\frac{dX_i}{dt} = F_i[X_1(t), \ldots, X_N(t), I(t)]$$

where the vector *I(t)* represents some external input into the system. The variable X_i can be interpreted as representing concentrations of proteins or mRNAs. Such a model has been used to model biochemical reactions in the metabolic pathways and gene regulation (69).

Bayesian networks are provided by the theory of graphical models in statistics. The basic idea is to approximate a complex multidimensional probability distribution using a product of simpler local probability distributions. Generally, a Bayesian network model is based on a directed acyclic graph (DAG) with N nodes. In genetic network, the nodes may represent genes or proteins and the random variables X_i levels of activity. The parameters of the model are the local conditional distributions of each random variable given the random variables associated with the



(b)

A'= B
B'= A OR C
C'= (A AND B) OR (B AND C) OR (A AND C)

(c)

Input			Output		
A	B	C	A'	B'	C'
0	0	0	0	0	0
0	0	1	0	1	0
0	1	0	1	0	0
0	1	1	1	1	1
1	0	0	0	1	0
1	0	1	0	1	1

Figure 10. Target Boolean network for reverse engineering. (a) The network wiring and (b) logical rules determine (c) the dynamic output. The challenge lies in inferring (a) and (b) from (c). [Reprinted from Liang et al. (77).]

parent nodes

$$P(X_1, \ldots, X_N) = \prod_i P(X_i | X_j : j \in N^-(i)) \quad (4)$$

where $N^-(i)$ denotes all the parents of vertex i. Given a dataset D representing expression levels derived using DNA microarray experiments; it is possible to use learning techniques with heuristic approximation methods to infer the network architecture and parameters. As data from microarray experiments are still limited and insufficient to completely determine a single model, people have developed heuristics for learning classes of models rather than single models, for instance, for a set of co-regulated genes (69). Bayesian networks have recently been shown to combine heterogeneous datasets, for instance, microarray data with functional annotation and mutation data to produce an expert system (79).

In this chapter, some major development in the field of bioinformatics were reviewed and some basic concepts in the field were introduced covering six areas: sequence analysis, phylogenetic analysis, protein structure analysis, genome analysis, microarray analysis, and network analysis. Due to the limited space, some topics have been left out. One such topics is text mining, which uses Natural Language Processing (NLP) techniques to extract information from the vast amount of literature in biological research. Text mining has become an integral part in bioinformatics. With the continuing development and maturing of new technologies in many system-level studies, the way that biological research is conducted is undergoing revolutionary change. Systems biology is becoming a major theme and driving force. The challenges for bioinformatics in the post-genomics era lie on the integration of data and knowledge from heterogeneous sources and system-level modeling and simulation providing molecular mechanism for physiological phenomena.

BIBLIOGRAPHY

Cited References

1. Abbas A, Holmes S. Bioinformatics and management science. Some common tools and techniques. Operations Res 2004; 52(2):165–190.
2. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
3. Gotoh O. An improved algorithm for matching biological sequences. J Mol Biol 1982;162:705–708.
4. Durbin S, Eddy S, Krogh A, Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge, (UK): Cambridge University Press; 1998.
5. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure. vol. 5, supplement 3. National Biomedical Research Foundation. Washington, (DC): 1978. p 345–352.
6. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
7. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
8. Altschul SF, Gish W, Miller W, Myers E, Lipman J. Basic local alignment search tool. J Molec Biol 1990;215:403–410.
9. Lipman JD, Altschul SF, Kececioglu JD. A tool for multiple sequence alignment. Proc Natl Acad Sci 1989;86:4412–4415.
10. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22: 4673–4680.
11. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AN, Wootton J. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. Science 1993;262:208–214.
12. (a) Delcher et al. 2002. (b) Zhu J, Liu JS, Lawrence CE. Bayesian adaptive sequence alignment algorithms. Bioinformatics 1998; 14:25–39.
13. Li WH. Molecular Evolution. Boston, MA: Sinauer Associates; 1997.
14. Holmes S. Bootstrapping phylogenetic trees. To appear in Statistical Science. Submitted in (2002).
15. Li S, Pearl DK, Doss H. Phylogenetic tree construction using MCMC. J Am Statist Assoc 2000;95:493–503.
16. Levitt M, Lifson S. Refinement of protein confirmations using a macromolecular energy minimization procedure. J Mol Biol 1969;46:269–279.
17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
18. Xu Y, Xu D. Protein threading using PROSPECT: Design and evaluation. Proteins Structure, Function, and Genetics 2000;40:343–354.
19. Levitt M, Warshel A. Computer simulation of protein folding. Nature 1975;253:694–698.
20. Nemethy G, Scheraga HA. Theoretical determination of sterically allowed conformations of a polypeptide chain by a computer method. Biopolymers 1965;3:155–184.
21. Levitt M. Molecular dynamics of native protein: Computer simulation of the trajectories. J Mol Biol 1983;168:595–620.
22. Beeman D. Some multi-step methods for use in molecular dynamics calculations. J Comput Phys 1976;20:130–139.
23. Bourne PE. CASP and CAFASP experiments and their findings. Methods Biochem Anal 2003;44:501–507.
24. Gordon D, Abajian C, Green P. Consed: A graphical tool for sequence finishing. Genome Res 1998;8(3):195–202.
25. Gordon D, Desmarais C, Green P. Automated finishing with autofinish. Genome Res 2001;11(4):614–625.
26. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res 1999;9(9):868–877.
27. Waterston RH, Lander ES, Sulston JE. On the sequencing of the human genome. Proc Natl Acad Sci USA 2002; 99(6):3712–3716.
28. Myers EW, et al. A whole-genome assembly of Drosophila 2000;287(5461):2196–2204.
29. Adams MD, et al. The genome sequence of Drosophila melanogaster Science 2000;287(5461):2185–2195.
30. Venter JC, et al. The sequence of the human genome. Science 2001;29:1304–1351.
31. Lukashin AV, Borodovsky M. GeneMark.hmm: New solutions for gene finding. Nucleic Acids Res 1998;26(4):1107–1115.
32. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic Acids Res 1999;27(23):4636–4641.
33. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997;268:78–94.
34. Yeh R-F, Lim LP, Burge CB. Computational inference of homologous gene structures in the human genome. Genome Res 2001;11:803–816.

35. Xu Y, Uberbacher CE. Automated gene identification in large-scale genomic sequences. J Comp Biol 1997;4:325–338.
36. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, Eddy SR. The Pfam Protein Families Database. Nucleic Acids Res 2004;32:D138–D141.
37. Eddy S. Profile hidden Markov models. Bioinformatics 1998;14:755–763.
38. Krogh A, Brown M, Mian IS, Juolander K, Haussler D. Hidden Markov models in computational biology applications to protein modeling. J Mol Biol 1994;235: 1501–1531.
39. Stomo GD, Hartzell GW. Identifying protein-binding sites from unaligned DNA fragments. Proc Natl Acad Sci 1989;86:1183–1187.
40. Lawrence CE, Reilly AA. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins Struct Funct Genet 1990;7:41–51.
41. Bailey LT, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology 1994; 28–36.
42. Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 2004;428:617–624.
43. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science 1997; 631–637.
44. O'Brien SJ, Menotti-Raymond M, Murphy WJ, Nash WG, Wienberg J, Stanyon R, Copeland NG, Jenkins NA, Womack JE, Graves JAM. The promise of comparative genomics in mammals. Science 1999;286:458–481.
45. Bourque G, Pevzner AP. Genome-scale evolution: Reconstructing gene orders in the ancestral species. Genome Res 2002;12:26–36.
46. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. Bioinformatics 2003;19(2):185–193.
47. Bajesy et al. 2005.
48. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002;30(4):e15.
49. Yang YH, Thorne N. Normalization for two-color cDNA microarray data. Science and statistics: A festschrift for terry speed. In: Goldstein D, ed. IMS Lecture Notes, Monograph Series. Vol. 40; 2003. p 403–418.
50. Smyth GK, Yang YH, Speed TP. Statistical issues in microarray data analysis. In: Brownstein MJ, Khodursky AB, eds. Functional Genomics: Methods and Protocols. Methods in Molecular Biology. vol. 224. Totowa, (NJ): Humana Press; 2003. p 111–136.
51. Kerr M, Churchill G. Analysis of variance for gene expression microarray data. J Comp Biol 2000;7:819–837.
52. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Royal Statist Soc B 1995;57(1):289–300.
53. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proc Natl Acad Sci 2001;98:31–36.
54. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. Bioinformatics 2003;19(2):185–193.
55. Wang H, He X, Band M, Wilson C, Liu L. A study of inter-lab and inter-platform agreement of DNA microarray data. BMC Genomics 2005;6(1):71.
56. Jarvinen A, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi O, Monni O. Are data from different gene expression microarray platforms comparable? Genomics 2004;83: 1164–1168.
57. Culhane AC, Perriere G, Higgins DG. Cross-platform comparison and visualisation of gene expression data using coinertia analysis. BMC Bioinformatics 2003;4:59.
58. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95(25):14863–14868.
59. Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. J Comp Biol 1999;6(3/4):281–297.
60. Tamayo P, Solni D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 1999;96(6):2907–2912.
61. Alter O, Brown PO, Bostein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 2000;97(18):10101–10106.
62. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: Class discovery and cass prediction by gene expression monitoring. Science 1999;286:531–537.
63. Sole RV, Ferrer-Cancho R, Montoya JM, Valverde S. Selection, tinkering, and emergence in complex networks. Complexity 2003;8:20–33.
64. Kanehisa M. A database for post-genome analysis. Trends Genet 1997;13:375–376.
65. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD. EcoCyc: A comprehensive database resource for Escherichia coli. Nucleic Acids Res 2005;33:D334–D337.
66. Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E, Kyrpides N, Fonstein M, Maltsev N, Selkov E. WIT: Integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res 2000;28(1): 123–125.
67. Fields S, Song OK. A novel genetic system to detect protein-protein interactions. Nature 1989;340:245–246.
68. Daraselia N, Yuryev A, Egorov S, Novichkova S, Nikitin A, Mazo I. Extraction of human protein interactions from MEDLINE using full-sentence parser. Bioinformatics 2003;19:1–8.
69. Baldi P, Hatfield GW. Microarrays and Gene Expression. Cambridge, (UK): Cambridge University Press; 2001.
70. Jeong H, Tombor B, Albert1 R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. Nature 2000; 407:651–654.
71. Wagner A, Fell DA. The small world inside large metabolic networks. Proc Royal Soc Lond B 2001;268:1803–1810.
72. Guimera R, Nunes Ameral AL. Functional cartography of complex metabolic networks. Nature 2005;433:895–900.
73. Schuster S, Hilgetag C, Woods JH, Fell DA. Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. J Math Biol 2002;45(2):153–181.
74. Schuster S, Fell DA, Dandekar T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nature 2000;18: 326–332.
75. Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED. Metabolic network structure determines key aspects of functionality and regulation. Nature 2002;420:190–193.

76. Cakir T, Kirdar B, Ulgen KO. Metabolic pathway analysis of yeast strengthens the bridge between transcriptomics and metabolic networks. Biotechnol Bioeng 2004;86:251–260.
77. Liang S, Fuhrman S, Somogyi R. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. Pacific Symp Biocomput 1998;3:18–29.
78. Somogyi R, Fuhrman S, Wen X. Genetic network inference in computational models and applications to large-scale gene expression data. Cambridge, (MA): MIT Press; 2001.
79. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc Natl Acad Sci USA 2003;100: 8348–8353.

Further Reading

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.

Baldi P, Chauvin Y, Hunkapillar T, McClure M. Hidden Markov models of biological primary sequence information. Proc Natl Acad Sci USA 1994;91:1059–1063.

Baldi P, Brunak S. Bioinformatics: The Machine Learning Approach. 2nd ed. Cambridge, (MA): MIT Press; 2001.

Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y. Predicting function: From genes to genomes and back. J Mol Biol 1998;283:707–725.

Bower J, Bolouri H. Computational Modeling of Genetic and Biochemical Networks. Cambridge, (MA): MIT Press; 2001.

Bray N, Dubchak I, Pachter L. AVID: A global alignment program. Genome Res 2003;13(1):97–102.

Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. Nature Genetics 1999;21:33–37.

Brudno M, CB Do, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou A. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 2003;13:(4):721–731.

Brudno M, Malde S, Poiakov A, Do C, Couronne O, Dubchak I, Batzoglou A. Glocal alignment: Finding rearrangements during alignment. Bioinformatics Special Issue on the Proceedings of the ISMB 2003;19:54i–62i.

Bryant SH, Altschul SF. Statistics of sequence-structure threading. Curr Opin Structur Biol 1995;5:236–244.

Cohen FE. Protein misfolding and prion diseases. J Mol Biol 1999;293:313–320.

Diaconis P, Holmes S. Random walks on trees and matchings. Electron J Probabil 2002;7:1–17.

Doyle JC. Robustness and dynamics in biological networks. In: The First International Conference on Systems Biology. New York: Japan Science and Technology Corporation, MIT Press; 2000.

Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Statistic Assoc 2002;97:77–87.

Eddy S, Mitchison G, Durbin R. Maximum discrimination hidden Markov models of sequence consensus. J Comput Biol 1995; 2:9–23.

Eddy SR. Non-coding RNA genes and the modern RNA world. Nature Rev Genet 2001;2:919–929.

Efron B, Halloran EE, Holmes S. Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci 1996;93:13429–13434.

Farris JS. The logical basis of phylogenetic analysis. In: Platnick N, Funk V, eds. Advances in Cladistics. vol. 2. 1983. p 7–36.

Fedorov AN, Baldwin TO. Contranslational protein folding. Biol Chem 1997;272(52):32715–32718.

Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 1981;17(6):368–376.

Felsenstein J. 1993. (Phylogeny Inference Package) version 3.5c. Department of Genetics, University of Washington, Seattle, WA. Available http://evolution.genetics.washington.edu/phylip.html.

Fischer D, Barret C, Bryson K, Elofsson A, Godzik A, Jones D, Karplus KJ, Kelley LA, MacCallum RM, Pawowski K, Rost B, Rychlewski L, Sternberg M. CAFASP-1: Critical assessment of fully automated structure prediction methods. Proteins 1999;3:209–217.

Fitch WM, Margoliash E. Construction of phylogenetic trees. Science 1967;155:279–284.

Foulds LR, Graham RL. The Steiner problem in Phylogeny is NP-complete. Adv Appl Math 1982;3:43–49.

Friedman N, Linial M, Nachman I, Peter D. Using Bayesian networks to analyze expression data. J Comp Bio 2000;7: 601–620.

Gardner M. The Last Recreations. New York: Copernicus-Springer Verlag; 1997.

Geman S, Geman D. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. IEEE Trans Pattern Anal Machine Intell 1984;6:721–741.

Gibson KD, Scheraga HA. Revised algorithms for the build-up procedure for predicting protein conformations by energy minimization. J Comp Chem 1987;9:327–355.

Goloboff PA. SPA. 1995. (S)ankoff (P)arsimony (A)nalysis, version 1.1. Computer program distributed by J. M. Carpenter, Department of Entomology, American Museum of Natural History, New York.

Gribaldo S, Cammarano P. The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. J Mol Evol 1998;47(5):508–516.

Haeckel E. Morphologie der Organismen: Allgemeine Grundzuge der Organischen FormenWissenschaft, Mechanisch Begrundet durch die von Charles Darwin Reformirte Descendenz-Theorie. Berlin: Georg Riemer; 1866.

Hannenhalli S, Pevzner PA. Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. STOC 1995; 178–189.

Helden JV, Andre B, Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Bio 1998;281: 827–842.

Hooper E. The River. Boston, (MA): Little, Brown; 1999.

Huelsenbeck J, Ronquist F. 2002. Mr. Bayes. Bayesian inference of phylogeny. Available at http://morphbank.ebc.uu.se/mrbayes/links.php.

Jukes T, Cantor C. Evolution of protein molecules. In: eds. Munro HN, Mammalian Protein Metabolism. New York: Academic Press; 1969. p 21–132.

Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequences features by using general scoring schemes. Proc Natl Acad Sci USA 1990;87(6):2264–2268.

Keith JM, Adams P, Bryant D, Kroese DP, Mitchelson KR, Cochran DAE, Lala GH. A simulated annealing algorithm for finding consensus sequences. Bioinformatics 2002;18: 1494–1499.

Kent WJ. BLAT–the BLAST-like alignment tool. Genome Res 2002;12(4):656–664.

Kirkpatrick S, Gelatt CD Jr, Vecchi MP. Optimization by simulated annealing. Science 1983;220:671–680.

Korf I, Flicek P, Duan D, Brent MR. Integrating genomic homology into gene structure prediction. Bioinformatics 2001;17:S140–S148.

Levitt M. Protein folding by restrained energy minimization and molecular dynamics. J Mol Biol 1983;170:723–764.
Ly DH, Lockhart DJ, Lerner RA, Schultz PG. Mitotic misregulation and human aging. Science 2000;287:1241–1248.
Ma B, Tromp J, Li M. PatternHunter: Faster and more sensitive homology search. Bioinformatics 2002;18:440–445.
Ma B, Wang Z, Zhang K. Alignment between two multiple alignments. In: Combinatorial Pattern Matching: 14th Annual Symposium, CPM 2003, Morelia, Michoacán, Mexico, June 25–27. Lecture Notes in Computer Science, vol. 2676. Heidelberg, Germany: Springer-Verlag; 2003.
Maddison D, Maddison W. 2002. Sinauer. Available at http://phylogeny.arizona.edu/macclade.
McAdams H, Shapiro L. Circuit simulation of genetic networks. Science 1995;269:650–656.
Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E. Simulated Annealing. J Chem Phys 1953;21:1087–1092.
Mjolsness E, Sharp DH, Rinetz J. A connectionsit model of development. J Theor Biol 1991;152:429–453.
Morales LB, Garduno-Juarez R, Romero D. Applications of simulated annealing to the multiple-minima problem in small peptides. J Biomol Struc Dyn 1991;8:721–735.
Morgenstern B. Dialign2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 1999;15:211–218.
Mountain JL, Cavalli-Sforza LL. Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. Proc Natl Acad Sci USA 1994;91:6515–6519.
Muckstein U, Hofacker IL, Stadler PF. Stochastic pairwise alignments. Bioinformatics 2002;18(sup. 2):S153–S160.
Notredame C, Higgins D, Heringa J. T-Coffee: A novel method for multiple sequence alignments. J Mol Biol 2000;302:205–217.
Peitsch MC. ProMod and Swiss-Model: Internet-based tools for automated comparative protein modeling. Biochem Soc Trans 1996;24:274–279.
Pevzner PA. Computational Molecular Biology, an Algorithmic Approach. Cambridge, (MA): MIT Press; 2000.
Pieper U, Eswar N, Ilyin VA, Stuart A, Sali A. ModBase, a database of annotated comparative protein structure models. Nucleic Acids Res 2002;30:255–259.
Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. Adv Protein Chem 1968;23:283–438.
Rannala B, Yang Z. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J Mol Evol 1996;43:304–311.
Richards FM. The protein folding problem. Sci. Am 1991; January: 54–63.
Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4(4): 406–425.
Schlick T. Optimization methods in computational chemistry. In: Reviews in Computational Chemistry, III. New York: VCH Publishers; 1992. p 1–71.
Schmulevich I, Dougherty E, Kim S, Zhang W. Probabilistic Boolean networks: A rule-based uncertainty model for gene regulatory networks. Bioinformatics 2002;18:261–274.
Schröder E. Vier combinatorische probleme. Z Math Phys 1870;15:361–376.
Shannon CE. A mathematical theory of communication. Bell Sys Tech J 1948;27:379–423, 623–656.
Snow ME. Powerful simulated annealing algorithm locates global minima of protein folding potentials from multiple starting conformations. J Comput Chem 1992;13:579–584.
Stanley R. Enumerative Combinatorics. vol. I. 2nd ed. Cambridge (MA): Cambridge University Press; 1996.
Swofford DL. PAUP. Phylogenetic analysis using parsimony. V4.0. Boston, (MA): Sinauer Associates; 2001.
Tozeren A, Byers SW. New Biology for Engineers and Computer Scientists. Englewood Cliffs, (NJ): Prentice Hall; 2003.
Wang LS, Jansen R, Moret B, Raubeson L, Warnow T. Fast phylogenetic methods for the analysis of genome rearrangement data: An empirical study. Proc of 7th Pacific Symposium on Biocomputing, 2002.
Watson JD, Crick FH. A structure for deoxyribose nucleic acid. Nature 1953; April.
White KP, Rifkin SA, Hurban P, Hogness DD. Microanalysis of drosphila development during metamorphosis. Science 1999; 286:2179–2184.
Winkler H. Verbeitung und Ursache der Parthenogenesis im Pflanzen und Tierreiche. Jena: Verlag Fischer; 1920.
Xu J, Hagler A. Review: Chemoinformatics and drug discovery. Molecules 2002;7:566–600.
Yang Z, Rannala B. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. Mol Biol Evol 1997;14:717–724.

See also COMPUTERS IN THE BIOMEDICAL LABORATORY; DNA SEQUENCE; MEDICAL EDUCATION, COMPUTERS IN; POLYMERASE CHAIN REACTION; STATISTICAL METHODS.

BIOLOGIC THERAPY.

See IMMUNOTHERAPY.

BIOMAGNETISM

DOUGLAS CHEYNE
Hospital for Sick Children
Research Institute

JINI VRBA
VSM MedTech Ltd.

INTRODUCTION

The science of *biomagnetism* refers to the measurement of magnetic fields produced by living organisms. These tiny magnetic fields are produced by naturally occurring electric currents resulting from muscle contraction, or signal transmission in the nervous system, or by the magnetization of biological tissue. The first observation of biomagnetic activity in humans was the recording of the magnetic field produced by the electrical activity of the heart, or *magnetocardiogram*, by Baule and McFee in 1963 (1). In 1968, David Cohen (2) at the Massachusetts Institute of Technology reported the first measurement of the alpha rhythm of the human brain, demonstrating that it was possible to measure magnetic fields of biological origin that are only several hundred femtotesla in magnitude (1 femtotesla $= 10^{-15}$ T)—more than 1 million times smaller than the earth's magnetic field ($\sim 5 \times 10^{-5}$ T). These early measurements were achieved using crude instruments consisting of inductance coils of 1–2 million windings in magnetically shielded enclosures and using extensive signal averaging. Instruments with increased sensitivity and performance based on the *superconducting quantum interference device*, or SQUID became available shortly after these pioneering measurements. The SQUID is a highly sensitive magnetic flux detector based on the

properties of electrical currents flowing in superconducting circuits, as predicted by Nobel laureate Brian Josephson in 1962 (3). The SQUID was soon adapted for use in biomagnetic measurements (4) and by the early 1970s, measurements of the spontaneous activity of the human heart (5) and brain (6) had been achieved without the need for signal averaging using superconducting sensing coils coupled to SQUIDs immersed in cryogenic vessels containing liquid helium. Thereafter, the field of biomagnetism continued to expand with the further development of SQUID based instrumentation during the 1970s and 1980s. The introduction in 1992 of multichannel biomagnetometers capable of simultaneous measurement of neuromagnetic activity from the entire the human brain (7,8) has resulted in widespread interest in the field of *magnetoencephalography* or *MEG* as a new method of studying human brain function.

Biomagnetic measurements are considered to have a number of advantages over more traditional electrophysiological measurements of heart and brain activity, such as the electrocardiogram or electroencephalogram. One significant advantage is that propagation of magnetic fields through the body is less distorted by the varying conductivities of the overlying tissues in comparison to electrical potentials measured from the surface of the scalp or torso, and can therefore provide a more precise localization of the underlying generators of these signals. In applications such as MEG and magnetocardiography (MCG), these measurements are completely passive and can be made repeatedly without posing any risk or harm to the patient. Also, biomagnetic signals are a more direct measure of the underlying currents in comparison to surface electrical recordings that measure volume conducted activity that must be subtracted from a reference potential at another location complicating the interpretation of the signal. In addition, magnetic measurements from multiple sites can be less time consuming since there is no need to affix electrodes to the surface of the body. As a result, biomagnetic measurements provide an accurate and noninvasive method for locating sources of electrical activity in the human body. The development of multichannel MEG systems has dramatically increased the usefulness of this technology in clinical assessment and treatment of various brain disorders. This has resulted in the recognition of routine clinical procedures by health agencies in the United States for the use of MEG to map sensory areas of the brain or localize the origins of seizure activity prior to surgery. Clinical applications of MCG have also been developed although to a lesser extent than MEG. This includes the assessment of coronary artery disease and other disorders affecting the propagation of electrical signals in the human heart. Another biomagnetic technique, known as *biosusceptometry*, involves measuring magnetized materials in the human body by measuring their moment as they are moved within a strong magnetic field. These measures can provide useful information regarding the concentration of ferromagnetic or strongly paramagnetic materials in various organs of the body, such as iron particles in the lung or iron-containing proteins in the liver. In addition, novel biomagnetometer systems are now available for the assessment of fetal brain and heart function in utero, and may provide a new clinical tool for the assessment of fetal health. Currently, there are >100 multichannel MEG systems worldwide and advanced magnetometer systems specialized for the measurement of magnetic signals from the heart, liver, lung, peripheral nervous system, as well as the fetal heart and fetal brain are currently being commercially developed. Although biomagnetism is still regarded as a relatively new field of science, new applications of biomagnetic measurements in basic research and clinical medicine are rapidly being developed, and may provide novel methods for the assessment and treatment of a variety of biological disorders. The following section reviews the current state of biomagnetic instrumentation and signal processing and its application to the measurement of human biological function.

BIOMAGNETIC INSTRUMENTATION

SQUID Sensors and Electronics

The SQUID sensor is the heart of a biomagnetometer system and provides high sensitivity detection of very small magnetic signals. The most popular types of SQUIDs are direct current (dc) and radio frequency (rf) SQUIDs, deriving their names from the method of their biasing. The modern commercial biomagnetometer instrumentation uses dc SQUIDs implemented in low temperature superconducting materials (usually Nb). In recent years, there has been significant progress in the development of high Tc SQUIDs, both dc and rf. These devices are usually constructed from $YBa_2Cu_3O_{7-x}$ ceramics. However, due to their poorer low frequency performance and difficulties with reproducible large volume manufacturing they are not yet suitable for large-scale applications. An excellent review of SQUID operation can be found in (9).

The rf SQUID was popular in the early days of superconducting magnetometry because they required only one Josephson junction. However, in majority of low Tc commercial applications, the rf SQUIDs have been displaced by dc SQUIDs due to their greater sensitivity, although in recent years, interest in rf SQUIDs has been renewed in connection with high Tc superconductivity. The operation of SQUIDs is illustrated in Fig. 1a. The dc SQUID can be modeled as a superconducting ring interrupted by two resistively shunted Josephson junctions as in Fig. 1a (11). The Josephson junctions are superconducting quantum mechanical devices that allow passage of currents with zero voltage, and when voltage is applied to them, they exhibit oscillations with a frequency to voltage constant of $\sim$484 MHz $\cdot$ μV. The resistive shunting causes the Josephson junctions to work in a nonhysteretic mode, which is necessary for low noise operation (9). An example of a thin-film dc SQUID, consisting of a square washer and Josephson junctions near the outside edge is shown in Fig. 1b (12,13). The usual symbol used to represent a dc SQUID is shown in Fig. 1c.

The SQUID ring (or washer) must be coupled to the external world and to the electronics that operates it (see Fig. 2a). When the dc SQUID is current biased, its *I–V* characteristics is similar to that of a nonhysteretic Josephson junction and the critical current I_0 is modulated

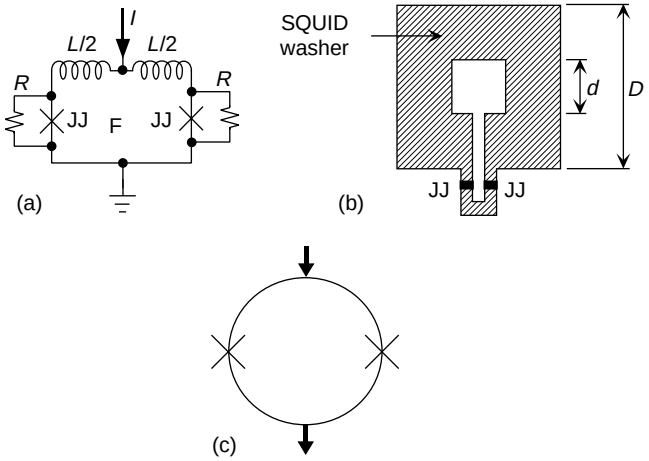


Figure 1. Thin-film dc SQUID. (a) Schematic diagram indicating inductances of the SQUID ring and shunting resistors to produce nonhysteretic Josephson junctions. (b) Diagram of a simple SQUID washer with Josephson junctions near the outer edge. (c) Symbolic representation of a dc SQUID, where the Josephson junctions are indicated by 'x'. (Reproduced with permission from Ref. 10).

by magnetic flux externally applied to the SQUID ring. The modulation amplitude is roughly equal to Φ_0/L (9), where Φ_0 is the flux quantum with magnitude $\sim 2.07 \times 10^{-15}$ Wb and L is inductance of the SQUID ring. The critical current is maximum for applied flux $\Phi = n\Phi_0$ and minimum for $\Phi = (n + 1/2)\Phi_0$. For monotonically increasing flux the average SQUID voltage oscillates as in Fig. 2d with period equal to 1 Φ_0. The SQUID transfer function is periodic (Fig. 2d) and to linearize it, the SQUID is operated in a feedback loop as a null detector of magnetic flux (14). Most SQUID applications use analogue feedback loop whereby a modulating flux with $\pm 1/4$ Φ_0 amplitude is applied to the SQUID sensor through the feedback circuitry (Fig. 2a,b). The modulation, feedback signal, and the flux transformer output are superposed in the SQUID, amplified, and demodulated in a lock-in detector fashion. The demodulated output is integrated, amplified, and fed back as a flux to the SQUID sensor to maintain its total input close to zero. The modulation flux superposed on the dc SQUID transfer function is shown in Fig. 2d and the modulation frequencies are typically several hundreds of kilohertz.

For satisfactory MEG operation, the SQUID system must exhibit large dynamic range, excellent interchannel matching, good linearity, and satisfactory slew rates. The analogue feedback loop is not always adequate and the dynamic range can be extended by implementing digital integrator as shown in Fig. 2c, and by utilizing the flux periodicity of the SQUID transfer function (15). The dynamic range extension works in the following manner: The loop is locked at a certain point on the SQUID transfer function and remains locked for the applied flux in the range of ± 1 Φ_0, Fig. 2d. When this range is exceeded, the loop lock is released and the locking point is shifted by 1 Φ_0 along the transfer function. The flux transitions along the transfer function are counted and are merged with the signal from the digital integrator to yield 32 bit dynamic range. This "flux slipping" concept can also be implemented using four-phase modulation (16), where the feedback loop jumps by $\Phi_0/2$ and can also provide compensation for the variation of SQUID inductance with the flux changes.

Flux Transformers

The purpose of flux transformers is to couple the SQUID sensors to the measured signals and to increase the overall magnetic field sensitivity. The flux transformers are superconducting and consist of one or more pickup coil(s) that are exposed to the measured fields. The pickup coil(s) are connected by twisted leads to a coupling coil that inductively couples the measured flux to the SQUID ring (as

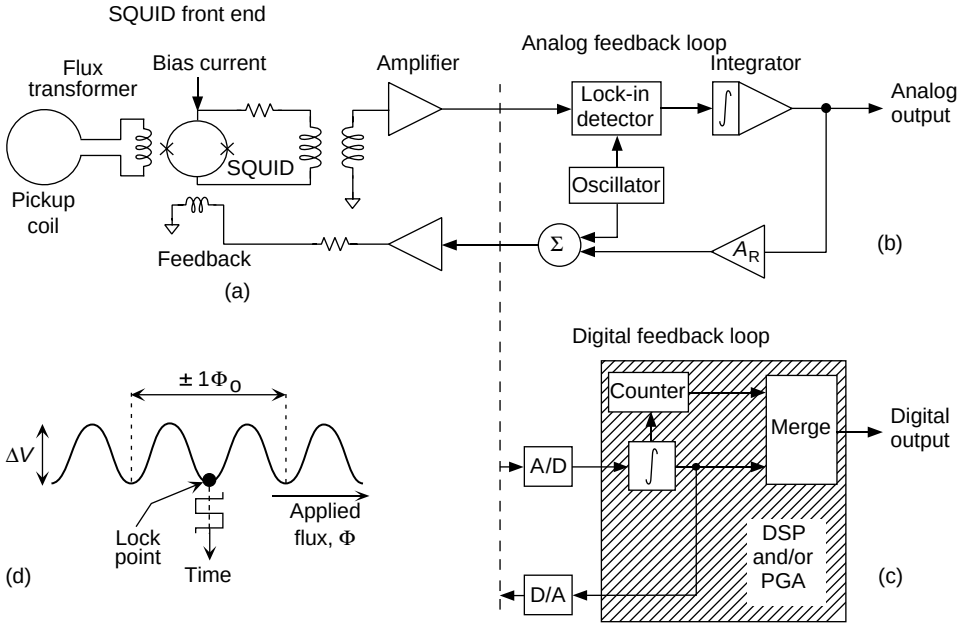


Figure 2. Examples of SQUID electronics, where the SQUID is operated as a null detector. (a) SQUID sensor is coupled to an amplifier. (b) Analogue feedback loop. (c) Digital feedback loop using digital signal processor (DSP) or a programmable logic array (PGA). (d) Feedback loop modulation. (Adapted with permission from Ref. 10).

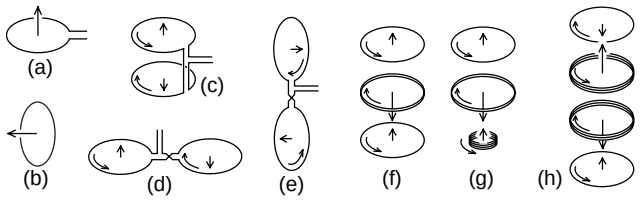


Figure 3. Examples of hardware flux transformers for biomagnetic applications. It is assumed that the scalp surface is at the bottom of the figure, (a) Radial magnetometer; (b) tangential magnetometer; (c) radial first-order gradiometer; (d) planar first-order gradiometer; (e) radial gradiometer for tangential fields; (f) second-order symmetric gradiometer; (g) second-order asymmetric gradiometer; (h) third-order gradiometer. (Reproduced with permission from Ref. 10).

shown in Fig. 2a). Because the flux transformers are superconducting, their gain is noiseless and their response is independent of frequency. The flux transformer pickup coil can have diverse configurations as shown in Fig. 3. A single loop of wire acts as a magnetometer and is sensitive to the magnetic field component perpendicular to its area, Fig. 3a and b. Two magnetometer loops can be combined with opposite orientation and connected by the same wire to the SQUID sensor. The loops are separated by a distance b and such a device is called a first-order gradiometer Fig. 3c–e, and the distance b is referred to as gradiometer baseline. The magnetic fields detected at the two coils are subtracted and the gradiometer acts as a spatial differential detector (this differential action is comparable to differential detection of electric signals (e.g., in electroencephalography, EEG). Fields induced by distant sources will be almost completely canceled by a gradiometer because both its coils will detect similar signals. On the other hand, near sources will produce markedly different fields at the two gradiometer coils and will be detected. Thus the gradiometers diminish the effect of the environmental noise that is typically generated by distant sources while remaining sensitive to near sources (e.g., neural sources). Similarly, first-order gradiometers can be combined with opposing polarity to form second-order gradiometers (Fig. 3f,g) and second-order gradiometers can be combined to form third-order gradiometers, (Fig. 3h). The flux transformers in Fig. 3 are called hardware flux transformers, because they are directly constructed in hardware by interconnecting various coils.

The main types of flux transformers used in commercial practice as the primary sensors are magnetometers (Fig. 3a), radial gradiometers (Fig. 3c), and planar gradiometers (Fig. 3d). These different sensor types will measure different spatial pattern of magnetic flux when placed over a current dipole as shown in Fig. 4. The radial magnetometer produces a field map with one maximum and one minimum, symmetrically located over the dipole with zero field measured directly above the dipole (Fig. 4a). The radial gradiometer in Fig. 4b produces similar field pattern as the magnetometer, except that the pattern is spatially tighter since it subtracts two field patterns measured at different distances from the dipole. The planar gradiometer field patterns are quite different from that of the

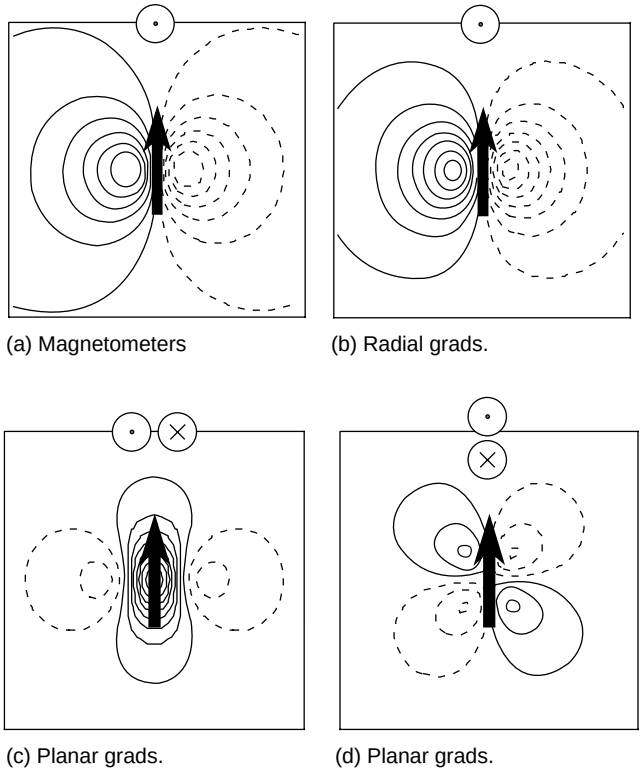


Figure 4. Response to a point dipole of several flux transformer types. A tangential dipole is positioned 2 cm deep in a semi infinite conducting space bounded by $x_3 = 0$ plane and its field is scanned by a flux transformer with its sensing coil positioned at $x_3 = 0$. Dipole position is indicated by a black arrow. Dimensions of each map are 14×14 cm. Schematic top view of the flux transformers is shown in the upper part of each figure. Solid and dashed lines indicate different field polarities. (a) Radial magnetometer; (b) radial gradiometer with 4 cm baseline; (c) planar gradiometer with 1.5 cm baseline aligned for maximum response; (d) planar gradiometer with 1.5 cm baseline aligned for minimum response. (Reproduced with permission from Ref. 10).

radial devices. If the two coils of the planar gradiometer are aligned perpendicular to the dipole, as in Fig. 4c, the planar gradiometer exhibits a peak directly above the dipole; if the two coils were aligned parallel to the dipole, the planar gradiometer exhibits a weak, clover-leaf pattern. When two orthogonal planar gradiometers are positioned at the same location, their two independent components can determine orientation of the current dipole located directly under the gradiometers (17).

In the absence of noise, there are no practical differences between these types of flux transformers. However, in the presence of noise, the signal-to-noise ratios (SNR) can differ greatly, resulting in significant performance differences between devices. For MEG applications, the magnitude of both the detected brain signal and environmental noise increases with increasing gradiometer baseline (distance between coils). Since the signal and noise functional dependencies on baseline are different, SNR exhibits a peak corresponding to an optimum baseline of ~3–8 cm for first-order radial gradiometers (10). Magnetometers can be thought of as gradiometers with very long baseline

and are not optimal because they can be overly sensitive to environmental noise. Planar gradiometers have good SNR for shallow brain sources but are suboptimal for deeper sources due to their short baselines resulting in poor depth sensitivity. Too long a baseline can also result in greater sensitivity to noise sources arising from the body itself, such as the magnetic field of the heart that may then contaminate the MEG signal. A detailed comparison of gradiometer design and performance can be found in (10).

Noise Cancelation

Introduction. Since biomagnetic measurements must be made in real world settings, the influence of noise on the measurements is a major concern in the design of biomagnetic instrumentation. Environmental noise affects biomagnetometer systems even when they are operated within shielded rooms. Environmental noise results from moving magnetic objects and currents (cars, trains, elevators, power lines, etc.). These noise sources are many orders of magnitude larger than signals of biomagnetic origin as shown in Fig. 5a. Note also, that only SQUID magnetometers have sufficient sensitivity for measuring biomagnetic signals of interest [atomic magnetometers are not yet suitable for biomagnetic applications (19)]. For MEG applications, the resolution or white noise level of the sensors should be much less than the "noise" level of brain activity ($\sim$30 fT $\cdot$ Hz$^{1/2}$). An example of background brain activity is shown in Fig. 5b. Also, certain MEG signal interpretation methods require the white noise to be as low as possible, however, the noise level cannot be made lower than the contribution of noise from the cryogenic vessel (dewar) itself. As a compromise, the majority of the existing MEG systems exhibit intrinsic noise levels of $<$10 fT $\cdot$ Hz$^{1/2}$ (typically $\sim$5 fT $\cdot$ Hz$^{1/2}$), yet are able to tolerate unwanted environmental noise many orders of magnitude greater.

Magnetic Shielding. Magnetic shielding is the most straightforward, though most costly method for reduction of environmental noise. A variety of shielded rooms have been used for biomagnetic applications and their relative shielding performance is shown in Fig. 6. The simplest shielding is accomplished through eddy currents by using a thick layer of high conductivity metal (20). Eddy current shielding is not effective at low frequencies, and therefore shielded rooms utilize high permeability μ-metal, which depending on the number of layers, can provide attenuation in the range from $\sim$30 to $\sim 10^5$ (21–24). Low frequency attenuation of nearly 10^8 was demonstrated with a whole-body, high Tc superconducting shield (25).

Environmental noise can also be reduced by active shielding, which can be employed either in unshielded environments (26), or in combination with shielded rooms (24,27,28). Active shielding system consists of a reference magnetometer, feedback electronics, and a set of compensating coils. The references measure the environmental noise and provide a signal that is amplified and fed into the

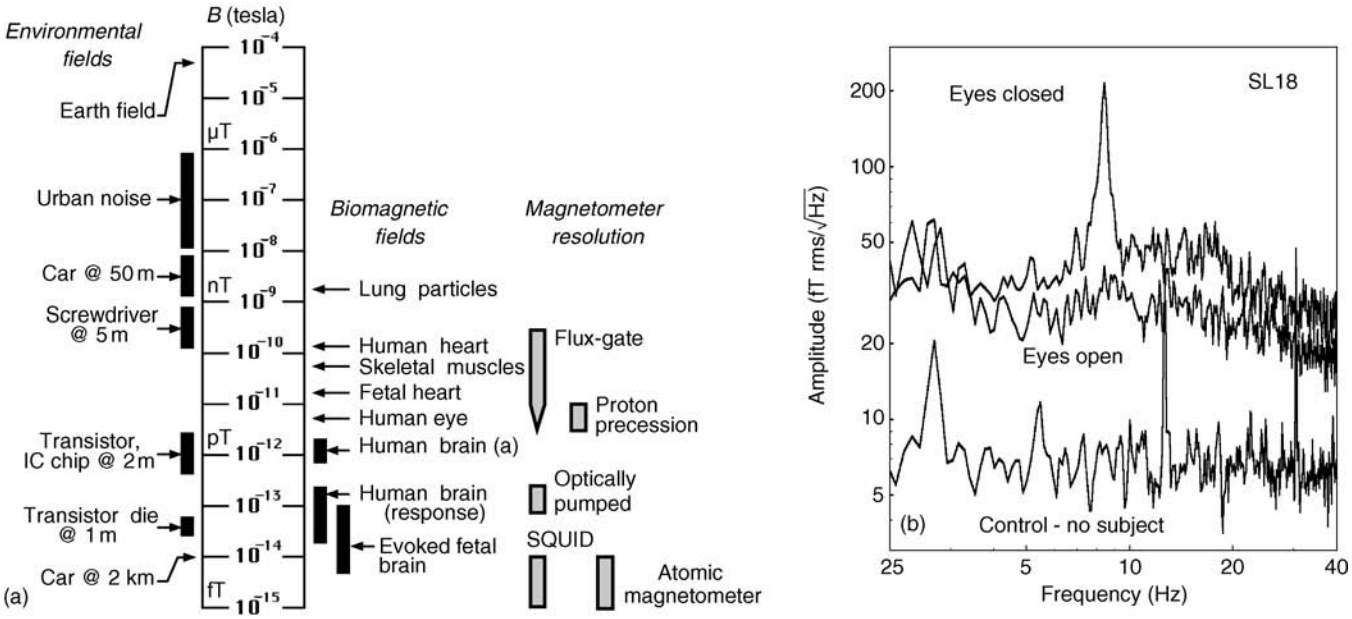


Figure 5. Environmental and brain generated noise. (a) Comparison of biomagnetic fields, environmental noise, and sensitivity in 1 Hz bandwidth of various types of magnetometers. (b) Spontaneous brain activity and the system noise measured in an unshielded environment, noise cancelation by synthetic third-order gradiometer, primary sensors are radial first-order gradiometers with 5 cm baseline. Control trace was collected with no subject in the helmet, large lines correspond to signals due to nearby rotating machinery. Eyes closed and open were collected with the subject in the MEG helmet. The presence of alpha activity (peak at 8 Hz) is visible in the eyes closed condition. (Reproduced with permission from Ref. 18).

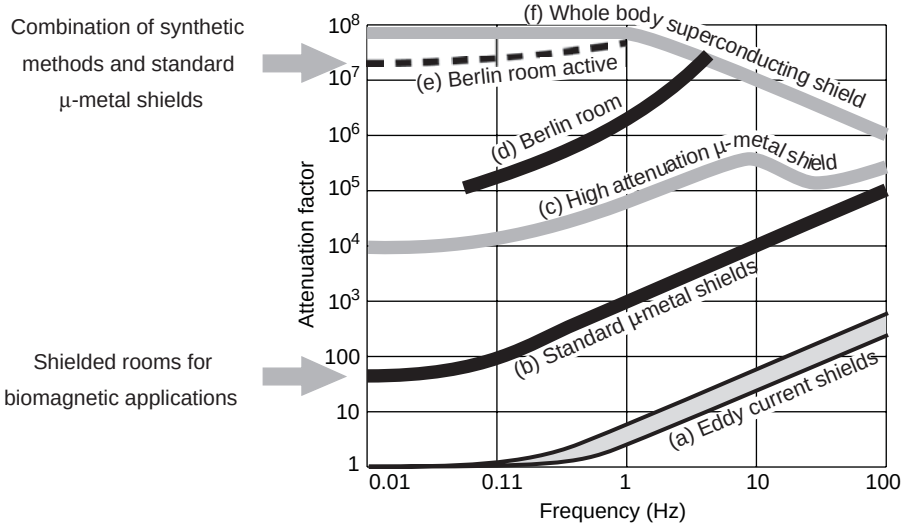


Figure 6. Noise attenuation of various shielded rooms as a function of frequency. (a) Eddy current Al rooms. (b) Standard μ-metal rooms used for MEG applications. (c,d) High attenuation μ-metal rooms. (e) Combination of high attenuation m-metal room in "d" and active shielding. (f) Whole-body high temperature superconducting shield. (Adapted with permission from Ref. 18).

compensating coils to reduce the noise. In general, the active shielding reduces the magnetic field noise due to far field sources and is effective only for magnetometers with no noise cancelation, while it has only a small effect on first-order gradiometers or magnetometers with noise cancelation. For higher order gradiometers, active shielding actually degrades system performance since the active coils can produce higher order gradients that are larger than that of the environmental noise.

Noise Reduction Using Higher Order Gradients. Since hardware noise cancelation (shielding or active noise cancelation) is usually not sufficient, additional methods, implemented in software or firmware, are employed. These methods either incorporate information from additional reference sensors or operate directly on the primary sensors. The reference sensors are typically a combination of SQUID magnetometers and gradiometers and the noise is canceled by synthesizing either higher order gradiometers or adaptively minimizing noise. The principle of synthetic gradiometer operation is similar for all gradiometer orders, and the method is illustrated for first-order gradiometer synthesis in Fig. 7a (29). The primary magnetometer detects the magnetic field component parallel to its coil normal, **p** (unit vector). The three reference magnetometers are orthogonal and their vector output, **r**, corresponds to the environmental field at the reference location, $\mathbf{r} \approx \mathbf{B}$. Then, if α_p is the primary magnetometer gain and α_r the reference gain (identical for all three references), the synthetic first-order gradiometer, $g^{(1)}$, can be derived as

$$g^{(1)} = m_p - \frac{\alpha_p}{\alpha_r}(\mathbf{p} \cdot \mathbf{r}) \approx \alpha_p \mathbf{p} \cdot \mathbf{G} \cdot \mathbf{b} \qquad (1)$$

where **b** is the gradiometer baseline (a vector connecting the primary sensor and the reference centers), and **G** is the first gradient tensor at the coordinate origin. Equation 1 states that the synthetic first-order gradiometer is a projection of the first-gradient tensor to the primary magnetometer orientation, **p**, and the baseline, **b**. To synthesize a second-order gradiometer, a primary hardware or synthetic first-order gradiometer, and a tensor first-gradient reference are used (Fig. 7b). Similar to Eq. 1, it can be shown that the synthetic second-order gradiometer output is a projection of the second gradient tensor to the coil orientation **p** and the first- and second-order gradiometer baselines $\mathbf{b}_1$ and $\mathbf{b}_2$. Synthesis of third- and higher order gradiometers is similar (29).

Adaptive methods can also be applied in addition to the synthetic gradiometers and can incorporate the same references as the gradiometers, but their coefficients are explicitly computed to minimize correlated noise (29). The advantage of synthesizing higher order gradiometers is that their coefficients are universal, independent of the noise character or sensor orientation (18). In contrast, the coefficients determined to adaptively minimize background noise are not universal because they depend on the noise character and sensor orientations (18) and assume that the noise environment is unchanging.

The noise cancelation achieved by various methods is illustrated in Fig. 8. The upper trace (a) shows the magnetic field noise outside a shielded room; and trace (b) shows the field noise after attenuation by the shielded room. The difference of the two slopes is due to the

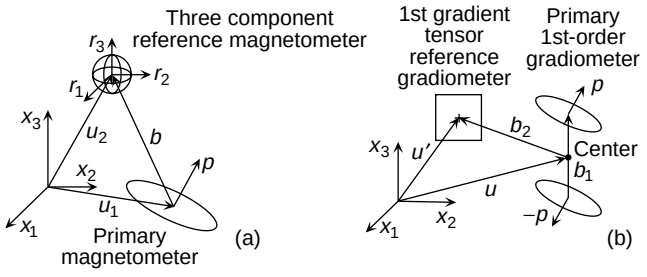


Figure 7. An illustration of gradiometer synthesis. (a) Synthesis of a first-order gradiometer from a primary magnetometer sensor and a vector magnetometer reference. (b) Synthesis of a second-order gradiometer from hardware first-order gradiometer and a first-gradient tensor reference. (Adapted from Ref. 30).

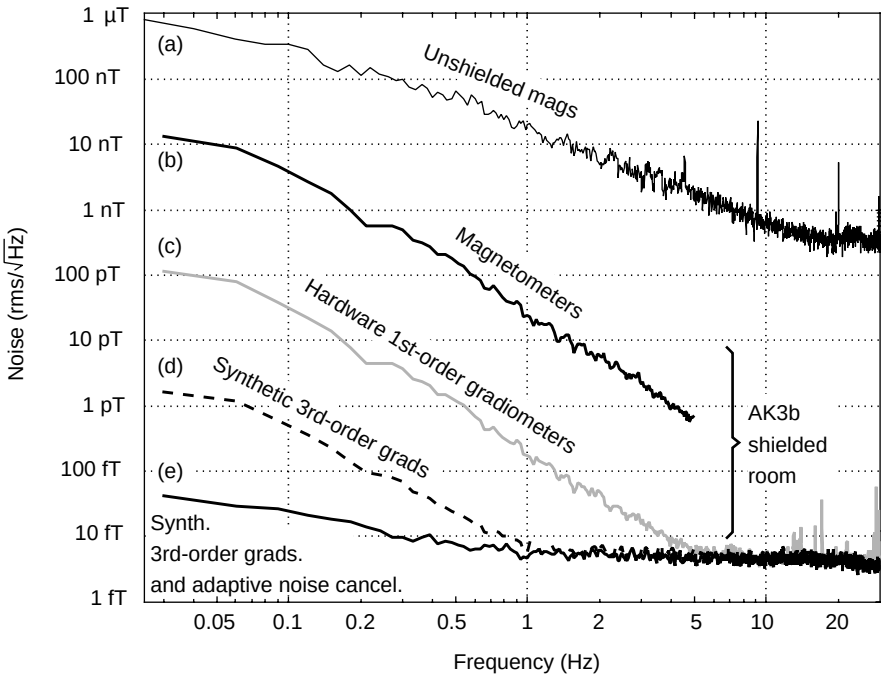


Figure 8. Reduction of environmental noise by a moderately shielded room, synthetic gradiometers, and adaptive methods. (a) Magnetic field noise outside a shielded room. (b) Field noise after attenuation by the shielded room. (c) Noise reduction by hardware first-order gradiometer with 5 cm baseline. (d) Noise reduction by synthetic third-order gradiometer (nearly four orders of magnitude lower noise than that of a shielded magnetometer in "b"). (e) Noise reduction by addition of adaptive methods to synthetic third-order gradiometer. (Adapted from Ref. 31).

frequency dependent eddy current shield that is part of the room. Hardware first-order radial gradiometers with 5 cm baseline reduce noise by nearly a factor of 100; and (c) a synthetic third-order gradiometer; (d) reduces the noise by almost another factor of 100. The low frequency environmental noise can further be reduced by adaptive method (e). The combination of all methods in Fig. 8 achieves attenuation of $>10^7$ at low frequencies.

Additional noise reduction methods can be employed in systems with a large number of channels. The simplest method is spatial filtering using Signal Space Projection (SSP) (32–34), which projects out from the measurement the noise components oriented along specific spatial vectors in signal space. The method works best when the signal and noise subspaces are nearly orthogonal. Related to SSP is noise elimination by rotation in signal space (35), which avoids loss of degrees of freedom encountered in SSP. These methods are discussed further in the Signal Interpretation section. More recently Signal Space Separation (SSS) has been proposed as a noise cancelation method in MEG (36). This approach was first proposed by Ioannides et al. (37) and reduces environmental noise by retaining only the "internal" component of the spherical expansion of the measured signal. This method can be applied to a number of problems inherent in biomagnetic measurements, including environmental noise reduction and motion compensation.

Cryogenics

The sensing elements of a biomagnetometer system (SQUIDs, flux transformers, and their interconnections) are superconducting and must be maintained at low temperatures. Since all commercial systems use low temperature superconductors, they must be operated at liquid He temperatures of 4.2 K. These temperatures can be achieved either with cryocoolers or by a cryogenic bath in contact with the superconducting components. The cryocoolers are attractive because they eliminate the need for periodic refilling of the cryogenic container. However, because they contribute magnetic and electric interference, vibrational noise, thermal fluctuations, and Johnson noise from metallic parts (38), they are not yet commonly used in MEG instrumentation. Present commercial biomagnetometer systems rely on cooling by liquid He bath in a nonmagnetic vessel with an outer vacuum space also referred to as a Dewar. An example of how the components may be organized within the Dewar for an MEG system is shown in Fig. 9a (39). The primary sensing flux transformers are positioned in the Dewar helmet area. The reference system for the noise cancelation is positioned close to the primary sensors and the SQUIDs with their shields are located at some distance from the references, all immersed in liquid He or cold He gas. The Dewar is a complex dynamic device that incorporates various forms of thermal insulation, heat conduction, and radiation shielding, as shown Fig. 9b. Most commercial MEG and MCG systems have reservoirs holding up to 100 L of liquid He and can be operated for periods of several days before refilling. An excellent review of the issues associated with the Dewar construction is presented in (38).

Biomagnetometer Systems: Overview

Even though magnetic fields have been detected from many organs, so far the most important application of biomagnetism has been the detection of neuromagnetic activity of the human brain. This interest led to the development of sophisticated commercial MEG systems. The current generation of these systems consists of helmet shaped multisensor arrays capable of measuring activity

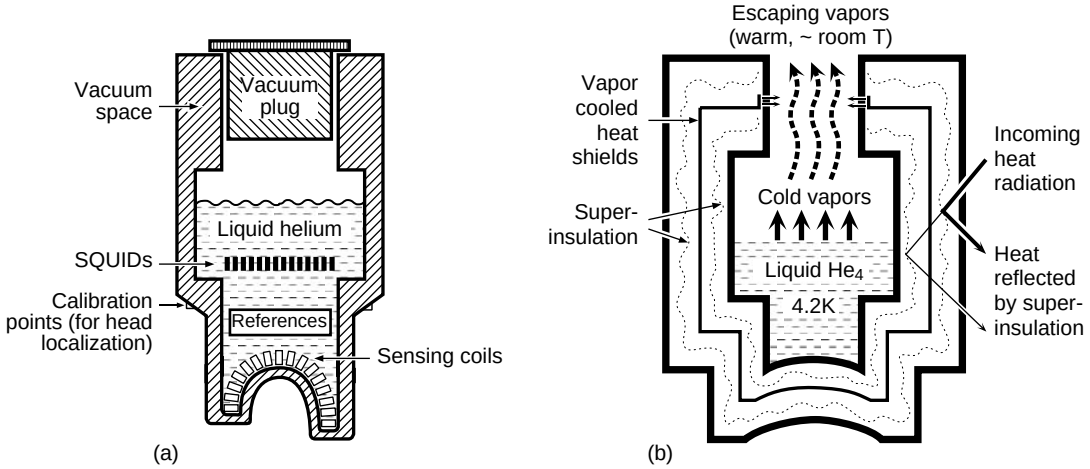


Figure 9. Schematic diagram of cryogenic containers used for whole-cortex MEG. (a) Placement of various MEG components relative to the cryogenic Dewar. (b) Principles of the Dewar operation. Reproduced with permission from (10).

simultaneously from the entire cerebrum. In contrast, multichannel magnetocardiogram (MCG) systems consist of a flat array of radial or vector devices (40–45) or systems with a smaller number of channels operating at liquid N_2 temperatures (46–51) for better placement over the chest directly above the heart. These flat array systems can also be placed over other areas of the body to measure peripheral nerve, gastrointestinal, or muscle activity. These systems can even be placed over the maternal abdomen to measure heart and brain activity of the fetus and a custom shaped multichannel array specifically designed for fetal measurements has recently been introduced (39,52).

MEG Systems. A diagram of a generic MEG system is shown in Fig. 10. The SQUID sensors and their associated flux transformers are mounted within a liquid He dewar suspended in a movable gantry to allow for supine or seated patient position. The patient rests on an adjustable chair or a bed. All signals are preamplified and transmitted from the shielded room to a central workstation for real-time acquisition and monitoring of the magnetic signals. At present, the majority of MEG installations use magnetically shielded rooms, however, progress is being made toward unshielded operation (18,40). The MEG measurements are often complemented by simultaneous EEG measurements or peripheral measures of muscle activity or eye movement. Most MEG installations have provisions for stimulus delivery in order to study brain responses to sensory stimulation and video and intercom systems in order to interact with the patient from outside the shielded room. Multichannel MEG systems are commercially available from a number of manufacturers (39,53–56).

For MEG localization of brain activity to be useful, particularly in clinical applications, it must be accurately known relative to brain anatomy. The anatomical information is usually obtained by magnetic resonance imaging

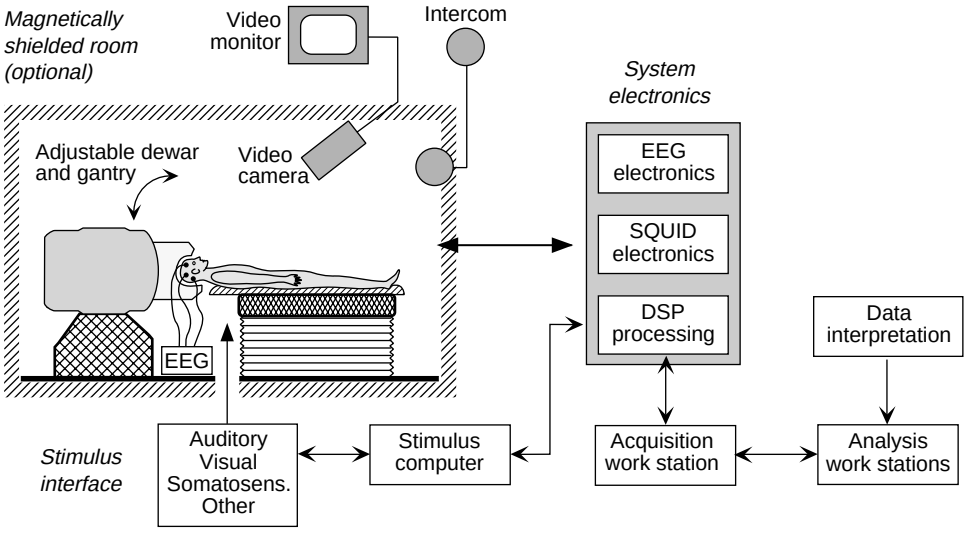


Figure 10. Schematic diagram of a typical MEG installation in a magnetically shielded room. (Reproduced with permission from Ref. 10).

(MRI), and the MRI images are required during the MEG interpretation phase. The registration of the MEG sensors to the brain anatomy is performed in two steps. First, the head position relative to the MEG sensor array is determined in order to accurately position MEG sources within a head-based coordinate system. Second, the head position relative to the MRI anatomical image is determined to allow transfer of MEG sources to the anatomical images. There are different methods for such registration. The simplest one uses a small number of anatomical markers positioned on identical locations on the head surface that can be measured both by MEG and MRI (e.g., small coils for MEG and lipid contrast markers for MRI) usually placed at anatomical landmarks near the noise and ears (18). To improve localization accuracy, the head shape can be digitized in the MEG coordinate system by a device mounted on the dewar (57) or by the MEG sensors (10). The surface of the head can also be constructed from segmented MRI and the transformation between the two systems can be determined by alignment of the two surfaces (58–60).

Biosusceptometers. A somewhat different system design is encountered in biomagnetometer systems used for the measurement of magnetic materials in the human body, such as iron content in the liver or magnetic contaminants in the lung. These instruments contain both SQUID sensing coils and a superconducting magnet operated in persistent mode. The system is suspended over the patient's body on a bed with a waterbag placed between the patient and dewar to provide continuity of the diamagnetic properties of body tissue. Figure 11 illustrates the layout of a biosusceptometer system for liver measurements with a patient in a supine position on a moveable bed. The patient is moved vertically relative to the SQUID gradiometer-magnet system and flux changes due to the susceptibility of the liver are monitored. These measures of magnetic moment can then be used to estimate the concentration of the paramagnetic compounds within the liver (62–64).

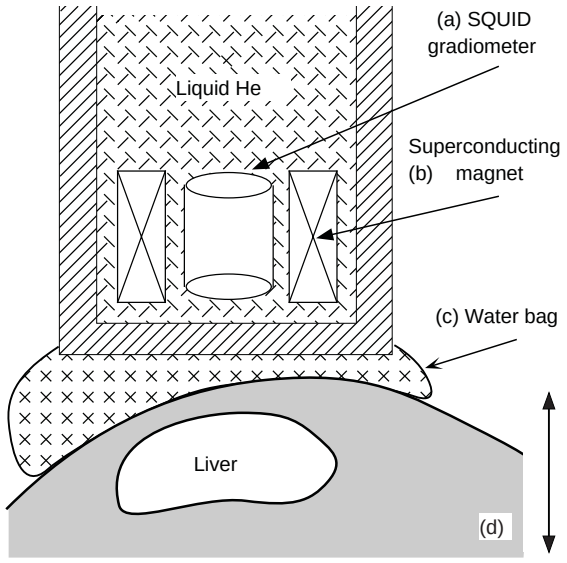


Figure 11. Schematic diagram of a liver susceptometer. (a) SQUID gradiometer. (b) Superconducting magnet. (c) Bag filled with water to simulate the diamagnetism of human body tissue. (d) Patient on a bed that is vertically movable. (Reproduced with permission from Ref. 61).

Signal Interpretation

Biomagnetometers measure the distribution of magnetic field outside of the body. Although the observed field patterns provide some information about the underlying physiological activity, ideally one would like to invert the magnetic field and provide a detailed image of the current distribution within the body. Such inversion problems are nonunique and ill defined. The nonuniqueness is either physical (65) or mathematical due to being highly underdetermined (i.e., there are many more sources than sensors). In order to determine the current distribution, it is necessary to provide additional information, constraints, or simplified mathematical models of the sources. The field of source modeling in both MEG and MCG has been an intensive area of study over the last 20 years. In the following section we shall review briefly various methods of source analysis as it is applied to MEG, although these methods apply to other biomagnetic measurements such as MCG, with the main difference being the physical geometry of the conductor volumes containing the sources. For detailed reviews of mathematical approaches used in biomagnetism (see 66–69).

Neural Origin of Neuromagnetic Fields. Magnetic fields of the brain measured by MEG are thought to be the primarily due to activation of neurons in the gray matter of the neocortex, whereas action potentials in the underlying fiber tracts (white matter) have been shown to produce only poorly synchronized quadrupolar sources associated with weak fields (70,71). Some subcortical structures have also been shown to produce weak yet measurable magnetic fields, but are difficult to detect without extensive signal processing (72,73). The generation of magnetic fields in the human brain is illustrated in Fig. 12. The neocortex of the brain (shown in Fig. 12a) contains a large number of pyramidal cells arranged in parallel (Fig. 12b) that in their resting state maintain an intracellular potential of ca. -70 mV. Excitatory (or inhibitory) synaptic input near the cell body or at the superficial apical dendrites results in the flow of charged ions across the cell membrane producing a graded depolarization (or hyperpolarization) of the cell. This change in polarization results in current flow inside the cell, called *impressed* current and corresponding return or *volume* currents that flow through the extracellular space in the opposite direction. Studies carried out in the early 1960s (74,75) demonstrated that these extracellular or volume currents are main generators of electrical activity measured in the electroencephalogram or EEG. The combination of excitatory and inhibitory synaptic inputs to different cortical layers can produce a variety of sink and source patterns through the depth of the cortex, each associated with current flow along the axes of elongated pyramidal cells toward or away from the cortical surface.

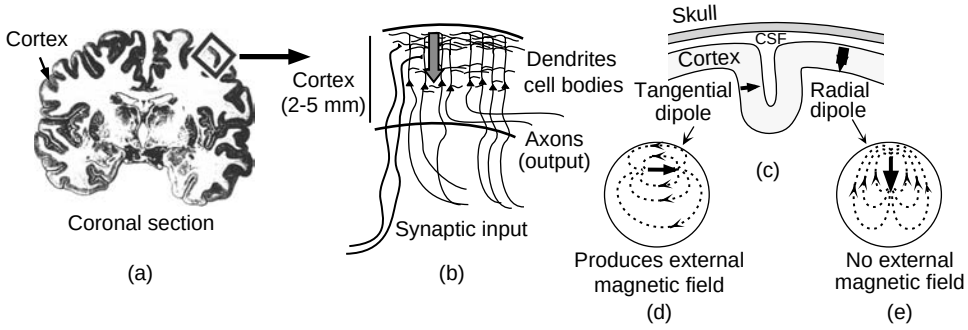


Figure 12. Origin of the MEG signal. (a) Coronal section of the human brain. The neocortex is indicated by dark outer surface. (b) Pyramidal cells in the cortex have vertically oriented receptive areas (dendrites). Depolarization of the dendrites at the cortical surface due to excitatory synaptic input results in Na^+ ions entering the cell producing a local current source and a current sink at the cell body, resulting in intracellular current flowing toward the cell body (arrow). (c) The cortex has numerous sulci and gyri resulting in currents flowing either tangentially or radially relative to the head surface. (d) Tangential currents will produce magnetic fields that are observable outside the head if modeled as a sphere. (e) Radial currents will not produce magnetic fields outside of the head if modeled as a sphere. (Adapted from Ref. 10).

Synchronous activity in large populations of these cells summate to produce the positive and negative time-varying voltages measured at the scalp surface in the EEG (76).

Okada et al. (77) carried out extensive studies over the last 20 years on the neural origin of evoked magnetic fields using small array "microSQUID" systems to measure directly magnetic fields from *in vitro* preparations in the turtle cerebellum and mammalian hippocampus. These studies have shown that although both extracellular and intracellular currents may contribute to externally measured magnetic fields, it is primarily intracellular or impressed currents flowing along the longitudinal axis of pyramidal cells that are the generators of evoked magnetic fields. A recent review of this work is presented in (78). Note that, since MEG measures mainly intracellular currents and EEG the return volume currents, the pattern of electrical potential over the scalp due to an underlying current source will reflect current flow in opposite direction to that of the magnetic field, as has been demonstrated in physical models (79) and human brain activity (80). In addition, activation of various regions of the enfolded cortical surface (the gyri and sulci) will result in current flow that is either radial or tangential to the scalp surface, respectively (Fig. 12c). If the brain is modeled as a spherical conducting volume, then due to axial symmetry it can be shown that only the tangential currents will produce fields outside the sphere (81) (Fig. 12d and e). Using *in vivo* preparations in the porcine brain, it has been experimentally demonstrated that, in contrast to the EEG, magnetic fields are relatively undistorted by the presence of the skull, and are generated primarily in tangentially oriented tissue (78). It has been recently shown, however, that MEG is insensitive only to a relatively small percentage of the total cortical surface in humans due to this tangential constraint (82). There is some uncertainty as to the extent of cortical activation typically measured by MEG. Current densities in the cortex have been estimated to be on the order of 50 $\text{pA} \cdot \text{m} \cdot \text{mm}^2$(83) suggesting that cortical areas of at least 20 mm^2 must be activated in order to produce a sufficiently large external field to be observed outside the head (66,68). However, current densities as high as 1000 $\text{pA} \cdot \text{m} \cdot \text{mm}^2$ have been recorded *in vitro* (77) indicating that much smaller areas of activation may be observed magnetically.

Equivalent Current Dipoles. The equivalent current dipole or ECD (81,84) is the oldest and most frequently used model for brain source activity. It is based on the assumption that activation of a specific cortical region involves populations of functionally interconnected neurons (macrocolumns) within a relatively small area. When measured from a distance, this local population activity can be modeled by a vector sum or "equivalent" current dipole that represents the aggregate activity of these neurons. The ECD analysis proceeds by estimating a priori the number of equivalent dipoles and their approximate locations, and then adjusting the dipole parameters (location and orientation) by a nonlinear search that minimizes differences between the field computed from the dipole model and the measured field (Fig. 13). This can be done at one time sample, or it can be extended to a time segment, where several dipoles are assumed to have fixed positions in space, but variable amplitude. Such models are referred to as "spatiotemporal" dipole models (85). The dipole fit procedures require the calculation of the magnetic field produced by a current dipole at each sensor: also termed the *forward solution*. Since the frequency range of interest for biomagnetic fields is <1 kHz, the quasistatic approximations of Maxwell's equations apply. If the head is assumed to be approximately spherical in shape it can be represented by a uniformly conducting sphere and the radial magnetic field of an ECD with magnitude $\mathbf{q}$, is given by the radial component of the well-known Biot–Savart law, $B_{\text{rad}}(\mathbf{r}) = \mathbf{B}(\mathbf{r}) \cdot \mathbf{r}/|\mathbf{r}|$, where the Biot–Savart

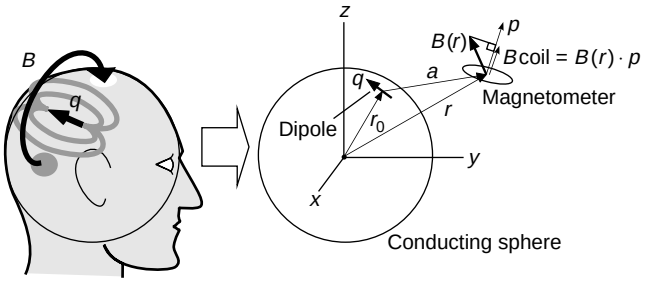


Figure 13. Magnetic fields due to an equivalent current dipole source will exit and reenter the head that can be modeled as a spherical shaped conducting medium. Calculation of the field magnitude ($\mathbf{B}_{coil}$) measured by a magnetometer coil due to a current dipole $\mathbf{q}$ at location $\mathbf{r_0}$ inside a sphere is given by the projection of the calculated vector field $\boldsymbol{B(r)}$ onto the direction normal to the surface area of the coil indicated by the unit vector $\mathbf{p}$, such that $\boldsymbol{B}_{coil} = \boldsymbol{B}(\boldsymbol{r}) \cdot \boldsymbol{p}$. The orientation of $\mathbf{q}$ is assumed to be tangential to the sphere surface. For gradiometer devices, the measured output of the gradiometer can be calculated as the difference between the field magnitudes calculated separately at each of the coils.

vector field, $\mathbf{B}(\mathbf{r})$, is given by

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \frac{\mathbf{q} \times (\mathbf{r} - \mathbf{r}_o)}{|\mathbf{r} - \mathbf{r}_o|^3} \tag{2}$$

where $\mathbf{r}_o$ is the ECD position and $\mathbf{r}$ is the position where the field is measured. For multiple ECDs or continuously distributed sources, Eq. 2 will also include the sum over all sources or the integral over the volume of the conducting sphere.

Generally, the vector of the external magnetic field is produced by both the primary current density reflecting the impressed (intracellular) currents, and volume currents that produce "secondary sources" on the surface of the volume conductor. For complex shapes, the calculation of the external field also requires knowledge of the conductivity profile of the conducting volume. The assumption of spherical symmetry, however, simplifies the calculation, and the vector field $\mathbf{B(r)}$ due to a current dipole $\mathbf{q}$ in a sphere at location $\mathbf{r}_o$ (Fig. 13b) is given by Sarvas (81) as

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi F^2} \{F\mathbf{q} \times \mathbf{r}_o - [(\mathbf{q} \times \mathbf{r}_o) \cdot \mathbf{r}]\nabla F\} \tag{3a}$$

where

$$F = a(ra + r^2 - \mathbf{r}_o \cdot \mathbf{r}) \tag{3b}$$

and

$$\nabla F = (r^{-1}a^2 + a^{-1}\mathbf{a} \cdot \mathbf{r} + 2a + 2r)\mathbf{r} - (a + 2r + a^{-1}\mathbf{a} \cdot \mathbf{r})\mathbf{r}_o \tag{3c}$$

and $\mathbf{a} = \mathbf{r} - \mathbf{r}_o$, $a = |\mathbf{a}|$, $r = |\mathbf{r}|$ and the permeability of free space $\mu_0 = 4\pi \times 10^{-7}$ H · m. The sensing coil measures the component of the vector field $\mathbf{B(r)}$ perpendicular to its surface area as shown in Fig. 13b. If the field is measured only in the radial direction, Eq. 3 simplifies to the radial component of the Biot–Savart law (Eq. 2), and the volume currents do not contribute any field. It can be seen from Fig. 13 that the definition of the origin of the theoretical sphere relative to the head will influence the calculation of the external magnetic field and thus plays a significant role in the accuracy of the single sphere approach. Since the head is not perfectly spherical, improved accuracy of the forward solution can be achieved by using more realistic models of the conducting surfaces and boundary element methods for the calculation of the magnetic field (69), but these methods are more computationally demanding. A simple improvement over the single sphere model can be achieved by using a multiple-sphere model, where independent spheres are determined for each sensor by evaluating local head curvature in the sensor vicinity (86).

The ECD procedure is very sensitive to the SNR and dc offsets, and therefore works best when applied to averaged brain responses that are well time locked to a sensory or motor event and requires an accurate estimate of signal baseline (e.g., prestimulus activity). This approach has proven useful for modeling simple patterns of focal brain activity, yet is compromised by interaction or "cross-talk" between simultaneously active sources, requiring that the number of dipoles be correctly specified. Also, ECD models do not correctly describe spatially "extended" sources: areas of cortical activity that may extend over an area of several square centimeters.

Minimum Norm. The dipole model assumes that the brain activity is localized in one or several small areas of the brain. Sometimes it is required to obtain a more general solution without an a priori assumption about the source distribution. This can be obtained by minimum norm methods, first proposed for MEG by Hämäläinen and Ilmoniemi (84). This inverse problem is underdetermined, solutions are diffuse, and the unweighted minimum norm favors solutions close to the sensors. The minimum norm method has subsequently been adapted to produce more localized solutions. The algorithm, FOCal Undetermined System Solution (FOCUSS) utilizes a recursive linear estimation based on weighted pseudoinverse solution (87) and the Minimum Current Estimate (MCE) utilizes the L1-norm approach (88). A related method, Magnetic Field Tomography (MFT) (89) utilizes weights and a regularization parameter that are optimized according to the given experimental geometry and noise. Another minimum norm-based method is the algorithm LORETA (LOw Resolution Electromagnetic TomogrAphy) (90). This algorithm introduces a spatial second derivative operator (Laplacian) into the weighting function and seeks the minimum norm solution subject to the maximum smoothness condition. This requirement is justified on a physiological assumption that neighboring points in the brain are likely to be synchronized. The method produces low spatial resolution that is a consequence of the smoothness constraint. Methods based on simulated (surrogate) data have also been proposed for producing distributed, unbiased solutions based on the minimum norm (91).

Bayesian Inference. Bayesian inference has also been applied to the biomagnetic inverse problem, using probability distributions of many possible source solutions. This approach can easily incorporate a priori information that may influence the likelihood of features of the current

distribution based on anatomy, maximum current strength, smoothness, and so on (92,93). This method determines expectation and variance of the a posteriori source current probability distribution given source prior probability distribution and data set (94,95). The model can include probability weightings determined from other imaging techniques such as functional MRI (fMRI) or positron emission tomography (PET) to influence the MEG current images.

Signal Space Projection. Signal space projection. (33,34) and beamformers are spatial filters that can separate signal from noise on the basis of their relationship in signal space (a M-dimensional space, where M is the number of MEG channels). The application of spatial filtering to MEG was first proposed by Robinson and Rose (96). This original article sparked growing interest in spatial filtering by the MEG community that still continues. The spatial filtering depends on the assumption that component vectors corresponding to different neuronal sources have distinct and stable (fixed) directions in signal space, and only their magnitudes are functions of time. If the vectors are defined by modeling the field produced by known dipole sources, SSP can be used as a spatial filter that passes only signals corresponding to these known sources. Thus, we can define the output of a spatial filter as $y_\theta(t) = \mathbf{P}_\parallel \mathbf{m}(t)$, where $\mathbf{P}_\parallel$ is the parallel projection operator (95) constructed from the forward solutions of the dipole source(s) of interest, and $\mathbf{m}(t)$ represents a vector of instantaneous MEG measurement at time t. The output of the spatial filter then provides a time series that is the estimate of changing strength of the dipole source(s) over time. Alternatively, if the vectors associated with artifact patterns are known, SSP can be used to remove these artifacts from the signal using orthogonal projection operators (32). If the signal vectors are determined from patterns in the data, the source model need not even be known. Note that restricting all sources to current dipoles in a known volume conductor model reduces SSP to a multiple dipole approximation (34).

Beamformers. The SSP method does not separate well sources that are not in orthogonal subspaces. To overcome this limitation, source analysis can be done by beamforming (borrowed from radio-communication and radar work). Beamformers utilize spatial and temporal correlations to obtain information about uncorrelated dipolar sources. The Linearly Constrained Minimum Variance (LCMV) beamformer in the form now used in MEG analysis was first described in 1972 (97) and can be used without specific information about source orientation. An introduction to the beamformers may be found in (98) and a relatively recent review of various beamforming techniques in (99). As in the case of SSP, if vector $\mathbf{m}(t)$ represent an instantaneous MEG measurement in M-dimensional space, we can define a spatial filter centered on the location "θ" as $y_\theta(t) = \mathbf{W}_\theta^T m(t)$, where $\mathbf{W}_\theta$ is a weight matrix. Only tangential sources contribute to the MEG signal. They can be decomposed into two orthogonal tangential directions and the corresponding forward solutions, $\mathbf{B}_{\theta 1}$ and $\mathbf{B}_{\theta 2}$, can be arranged in a forward solution matrix as $\mathbf{H}_\theta = [\mathbf{B}_{\theta 1}, \mathbf{B}_{\theta 2}]$. The beamformer weights are determined by minimizing the power projected from the location , $P_\theta = \mathbf{W}_\theta^T \mathbf{C} \mathbf{W}_\theta$, subject to the unity gain condition, $\mathbf{W}_\theta^T \mathbf{H}_\theta = \mathbf{I}$, where $\mathbf{C}$ is the covariance matrix of the measurement and $\mathbf{I}$ is the identity matrix. The weights are given as (100)

$$\mathbf{W}_\theta = \mathbf{C}^{-1}\mathbf{H}_\theta(\mathbf{H}_\theta^T \mathbf{C}^{-1}\mathbf{H}_\theta)^{-1} \tag{4}$$

An alternative approach known as synthetic aperture magnetometry (SAM) defines an optimal dipole orientation for each spatial filter location (101). Only one vector is retained, $\mathbf{H}_\theta = \mathbf{B}_\theta$ simplifying Eq. 6 to $\mathbf{W}_\theta = \mathbf{C}^{-1}\mathbf{B}_\theta (\mathbf{B}_\theta^T \mathbf{C}^{-1} \mathbf{B}_\theta)^{-1}$. This approach produces higher spatial resolution due to less projected sensor noise by the spatial filter (102). The beamformer weights can be used to compute the time course of the dipole magnitude variation or power at a single location in the brain independently of other active sources, provided sources are not highly correlated. An especially useful quantity is the normalized power $Z_\theta^2 = P_\theta / N_\theta$, where $N_\theta^2 = \mathbf{W}_\theta^T \Sigma \mathbf{W}_\theta$ is the sensor noise projected by the beamformer from location 'θ', and Σ is the sensor noise covariance matrix (100). In contrast to P_θ and N_θ, the parameter Z_θ^2 behaves gracefully through the center of the model sphere and does not exhibit a singularity. A spatial image of brain activity can be obtained by computing the normalized power at individual brain voxels, θ, one at a time over a region of interest.

Multiple Signal Classification. MUltiple SIgnal Classification (MUSIC) is a signal space scanning method and is related to beamforming (103,104). MUSIC requires an initial nonlinear step of partitioning the data covariance matrix into signal and noise subspaces using standard eigendecomposition methods. This partitioning can be more readily determined from the averaged data and as a result the method is more difficult to apply to spontaneous brain activity. Sources are located by scanning of the brain volume and at each location requiring that the dipole forward solution be orthogonal to the noise subspace (or parallel to the signal subspace). A more recent implementation known as recursively applied and projected MUSIC (RAP-MUSIC) projects out each located source and then repeats the scanning procedure (105). Similar to beamforming, MUSIC also assumes there are fewer sources than sensors, the sources are uncorrelated and the noise is white. In the limit of high SNR (e.g., averaged data), a small number of sources, and white noise, the MUSIC localizer function and beamformer based source power estimates differ only by a scaling factor.

Principal Component Analysis. Principal Component Analysis (PCA), for example (106,107), also determines the signal and noise subspaces. The method is based on second order statistics and attempts to fit dipoles into the orthogonal principal spatial vectors of the singular value decomposition of the data. For mixtures of components corresponding to nonorthogonal spatial vectors, the PCA cannot account for the structure of the data (108). The PCA has been shown to be potentially inaccurate, as it can mislocalize dipoles even in noiseless simulations.

Independent Component Analysis. Independent Component Analysis (ICA) is a relatively new technique that

allows separation of sources that are linearly mixed at the sensors. The method is also called blind source separation, because the source signals are not directly observed and nothing is known about their mixture (109,110). The method uses higher order statistics and in realistic situations is often more successful than PCA (108). The mixing model used for the separation is usually stated as $\mathbf{m}(t) = \mathbf{A}\mathbf{s}(t)$, where $\mathbf{m}(t)$ is the instantaneous vector of the measurement, $\mathbf{s}(t)$ is the instantaneous source activity vector, and $\mathbf{A}$ is the mixing matrix. The procedure provides solution for an unmixing matrix $\mathbf{B}$, such that the estimated source activity is given as $\hat{\mathbf{s}}(t) = \mathbf{B}m(t)$, where $\hat{\mathbf{s}}$ is the estimate of the source vector $\mathbf{s}$. The sources are assumed to be statistically independent and the separation is obtained by optimizing a contrast function, that is, a scalar measure of some distributional property of the output $\hat{\mathbf{s}}$. The contrast functions are based on entropy, mutual independence, high order decorrelations, and so on. The ICA has been applied to MEG and EEG to either remove artifacts or extract desired signals (111,112).

APPLICATIONS OF BIOMAGNETIC MEASUREMENTS

Magnetoencephalography: Basic Studies

The most prevalent and rapidly growing application of biomagnetism is the field of magnetoencephalography (MEG): the measurement of human brain activity. This field of basic and clinical research is also referred to as *neuromagnetism* or *magnetic source imaging* (113–117). The latter term is often used to refer to the localization of neural sources with respect to individual brain anatomy by the combination of MEG source modeling with structural imaging techniques such as MRI (see Magnetic Resonance Imaging). As noted in the introduction, the first magnetic fields recorded from the human brain involved the observation of spontaneous alpha rhythm activity. This was soon followed by the application of MEG measurements to the study of *evoked responses* of the human brain: Time-averaged responses to discrete sensory or motor events that provide sufficient SNR to allow for the localization of brain regions contributing to the externally measured field patterns. Due to its excellent time resolution and ability to measure neuronal function directly, MEG has continued to generate interest within the field of human neuroscience as a complement to other methods of functional brain imaging based on metabolic or hemodynamic changes in brain function, such as fMRI (see Magnetic Resonance Imaging) or PET (see Positron Emission Tomography). Present MEG practice includes measurement of both evoked and spontaneous signals and is used for clinical purposes and for investigation of a wide range of brain processes.

Somatosensory Evoked Fields. Evoked responses to stimulation of the human somatosensory system (somatosensory evoked fields or SEFs) were first reported by Brenner and colleagues in 1987 (118). The observed magnetic brain response to electrical stimulation of the digit was of great interest since it demonstrated a well-characterized dipolar field pattern over the scalp indicative of a single neural generator located in the underlying somatosensory cortex. Subsequent studies have shown that early components of the SEF occurring at latencies of 20–50 ms reflect early activation of the primary somatosensory cortex contralateral to the side of stimulation, and are generally well modeled as single ECD source in these brain regions [see (119) for a recent review]. The earliest component at a poststimulus latency of 20 ms (sometimes referred to as the "M20" or "N20m" since it is considered the magnetic equivalent of the negative N20 potential measured in the EEG) arises from the posterior bank of the central sulcus: a primary somatosensory projection area. By stimulating different body parts, it can be shown that the N20m source reflects the somatotopic or "homuncular" organization of the ascending neural pathways of the somatosensory system to this brain region (120). Figure 14 shows a typical SEF response to stimulation of the median nerve at the wrist and the localization of an ECD model fit to the N20m source in the corresponding somatosensory cortex.

When using extensive signal averaging, MEG recordings also reveal low amplitude high frequency oscillations in the 300–900 Hz range during the period of the N20m

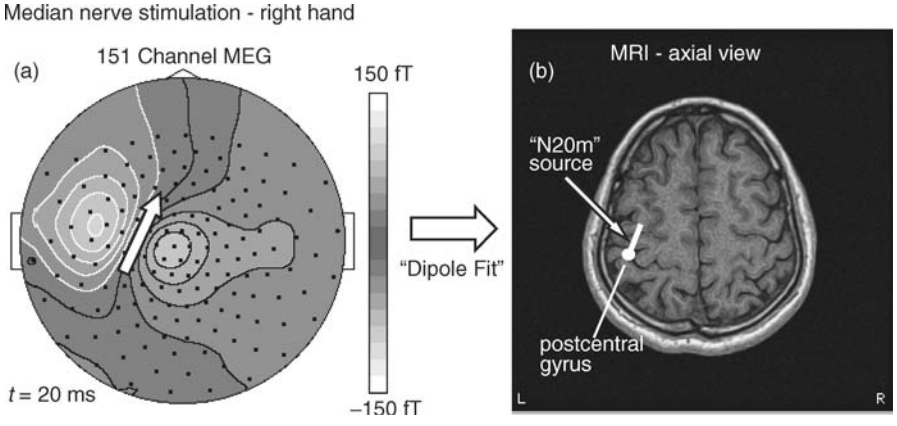


Figure 14. (a) Topographic map (polar projection with nose upwards) of the magnetic field pattern recorded from 151 MEG channels over the scalp at a latency of 20 ms following stimulation of the right median nerve (average of 600 stimuli). White contours indicate outgoing fields and solid contours, ingoing fields. Arrow indicates direction of current flow below the scalp corresponding to the dipolar field pattern over the left hemisphere. (b) Location of a single ECD source corresponding to the magnetic field pattern shown in (a) indicated by white dot with tail indicating direction of current flow superimposed on an axial slice of the individual's MRI. Location is in the hand region of the primary somatosensory cortex.

response (121). These oscillations have been proposed to reflect the activity of inhibitory interneurons in the somatosensory cortex (122) although cortico-thalamic pathways have also been shown to play a possible role (123). The N20m is followed by reversals of the same pattern at latencies of 30 and 40 ms that appear to reflect additional activation of somatosensory areas. These are followed by more complex and widespread activity from ~80 to 150 ms after stimulation that reflects bilateral activation of secondary somatosensory areas in the parietal operculum and is most likely related to higher order processing of somatosensory input (124). The MEG responses at latencies of 50–70 ms are elicited by mechanical stimulation of the digits (125) and reflect somatotopically organized sources in the primary somatosensory cortex (80). A number of MEG studies have used mechanical SEFs to demonstrate functional reorganization or "plasticity" of the somatosensory cortex resulting from anesthetic block or damage to the peripheral nerves or amputation (126), or even as the result of musical training (127).

Movement Related Fields. The first recordings of magnetic fields accompanying simple finger movements were reported in the early 1980s. Deecke et al. (128) observed slow magnetic field changes over sensorimotor areas of the brain preceding voluntary movements of the digits. These "readiness fields" begin approximately a half a second prior to the onset of a voluntary movement and are thought to represent activation of brain areas involved in motor preparation (129). Dipole source analysis suggests that premovement fields arise primarily from bilateral activation of the primary motor cortex (even for unilateral movements) with larger amplitude fields and dipole magnitudes the contralateral to the side of movement (130).

Movement-evoked fields (MEFs) accompany the onset and execution phase of simple movements. The first component (MEFI) is the largest in amplitude and begins ~100 ms after onset of EMG activity in the involved muscles. These responses appear to arise from sources in the postcentral gyrus, most likely reflecting sensory feedback to cortex from proprioceptors in the muscles (131) and are correlated with movement velocity (132). Movements made in response to a sensory cue show a very similar pattern of activity, but with a shorter latency of onset of premovement activity (133). Passive movements also elicit magnetic responses thought to reflect activation of the proprioceptive inputs to areas of the postcentral gyrus (134). MEG mapping studies have demonstrated activity in motor cortex during motor imagery providing evidence of the involvement of these brain areas in the simple imagination of movement (135).

MEG–EMG Coherence. By using a single channel magnetometer, Conway et al. (136) made the interesting observation of increased coherence (correlation in the frequency domain) between the surface electromyogram (EMG) in a contracting muscle and MEG recordings made over the contralateral motor areas. Subsequently, there has been a great deal of interest in the relationship between MEG–EMG coherence and the functional relationship between spontaneous cortical rhythms and EMG activity during movement (137). Interestingly, changes in the frequency of coherence varies with the strength of muscular contraction and recent studies have shown that MEG–EMG coherence may reflect the underlying physiology of tremor in patients with Parkinson's disease (138) or essential tremor (139).

Sensorimotor Rhythms. The MEG studies have also provided evidence for the functional significance of specific oscillatory brain activity in humans associated with both somatosensory stimulation and motor output. These centrally distributed rhythms were first observed in the EEG, and are predominant at frequencies ~10 Hz (also referred to as the mu rhythm) and in the range of 20–30 Hz. The MEG studies have been able to show that these are functionally independent cortical oscillations that originate from postcentral and precentral regions, respectively (140,141). These sensorimotor rhythms are suppressed during median nerve stimulation, followed by a transient increase or "rebound" of 20–30 Hz rhythms within 500 ms after stimulus onset. A similar pattern of suppression followed by rebound is observed during voluntary movements (142). These rhythmic changes are modulated by sensorimotor tasks such as movement or passive tactile stimulation and motor imagery or even observation of another individual's movements (140). Rhythmic activity is not amenable to the same signal averaging technique used for evoked fields and therefore the ECD source modeling approach is more difficult to apply. Spatial filtering methods, however, provide a new approach to the localization of frequency dependent power changes in cortical areas using MEG and have been applied to the localization of rhythmic changes induced by somatosensory stimulation (141,143) and voluntary movements of the digits (144).

Visual Evoked Fields. One of the first magnetic evoked responses recorded from the human brain was the visual evoked field or VEF reported by Brenner et al. in 1975 (145). Robust responses can be elicited at latencies of 100–150 ms following visual stimulation using light flashes or visual pattern contrast changes (e.g., reversing checkerboard stimuli). However, early VEF responses pose a challenge in terms of modeling their sources due to the complex enfolded cross-like shape of the primary (striate) visual cortex: also referred to as the "cruciform" model. More recently, investigators have successfully modeled early VEF components in primary visual cortex by stimulating restricted portions of the total visual field using both monochrome (146) and color (147) pattern stimuli and have produced source configurations that reflect the retinotopic organization of the primary visual cortex. Due to the difficulty in applying ECD models to the VEF response, spatial filtering methods such as beamforming have been found to be useful for imaging visual cortex function (148). Figure 15 shows the activation of primary visual cortex by a steady state visual pattern (reversing checkerboard) using the SAM beamforming algorithm. Visual stimuli also activate several nonprimary (extrastriate) visual areas depending on the attributes of the stimulus. A number of MEG studies have shown activation of brain areas related to higher order visual processes such as detection

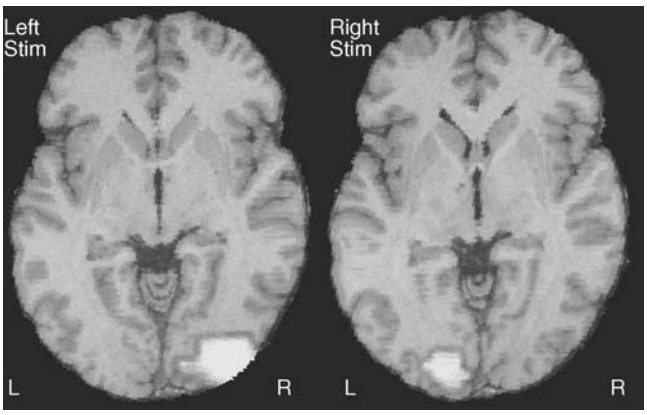


Figure 15. Images of following response to flickering checkerboard stimulus at $f = 17$ Hz presented to the left or right visual field, using a whole-head MEG recordings and the synthetic aperture magnetometry (SAM) beamformer algorithm. The image shows increased source power as yellow (lighter) colored areas at the posterior portion of the brain, corresponding to increased power at 34 Hz ($2f$) in the primary visual cortex of the contralateral hemisphere. Unpublished data. (Courtesy of K. Singh.)

of coherent motion (149,150). Clinical applications of VEFs have been limited, although abnormal VEFs have been reported in cases of strabismic amblyopia (151).

Auditory Evoked Fields. Auditory evoked fields were first reported by Reite et al. (152) and subsequent MEG mapping studies have demonstrated that responses at latencies of 50 and 100 ms reflect activity of primary auditory cortex in the temporal lobes (153). The largest response occurring ~100 ms following stimulus onset, termed the M100, has been the most extensively studied auditory evoked field response. The M100 is bilateral for both binaural and monaural stimuli and has been shown to reflect the frequency specific (tonotopic) organization of the primary auditory cortex (154). These magnetic evoked responses are of interest in the study of the functional organization of the auditory system as they reflect perceptual attributes of auditory stimulation such as perceived pitch or the frequency profiles of complex speech sounds (155,156). Auditory responses to repetitive (steady-state) auditory stimuli show enhanced amplitude in the EEG at presentation rates of ~40 Hz and were initially thought to represent volume conducted thalamic responses. However, MEG steady-state auditory responses were shown by Weinberg et al. (157) to reflect oscillations at the stimulus frequency in the auditory cortex. Subsequent studies have suggested that 40 Hz auditory responses reflect thalamocortical networks in the brain responsible for integration of sensory input (158). The 40 Hz MEG response has recently been used to measure temporal integration times in the primary auditory cortex (159).

MEG Studies of Higher Brain Function. Although early MEG studies have provided useful information regarding the early processing of sensory input and motor output, one of the more intriguing potential uses of MEG is the noninvasive study of higher brain function. The EEG studies of cognitive function have been carried out using event-related potentials (ERPs) for many decades. The MEG measurements using similar paradigms have helped gain a better understanding of the neural basis of many ERP components. Early MEG studies have had some success in measuring brain responses related to short-term memory (160), target detection tasks (161), or selective attention (162). More recently, the use of whole head MEG systems have enabled the study of more complex aspects of cognitive processing in humans, such as face recognition (163) and object naming (164). Basic research on brain mechanisms related to speech and language is also a promising area of application for MEG. Early studies have shown that speech processing is affected by incongruous visual feedback at the level of the auditory cortex (165). More recent studies have attempted to localize magnetic brain responses related to syntactic (166) and semantic (167) language processes as well as the processing of speech sounds (168). The MEG responses have also been used to study abnormal processing of sensory input during reading in dyslexic children (169) and during speech in stutterers (170).

A great deal of progress has been made in studying higher brain function with MEG by applying traditional source analysis methods to ERP components. However, these complex brain processes often involve activation of multiple brain regions complicating the interpretation of the data in terms of simple ECD models. Moreover, many of these processes may not be highly time-locked to specific sensory or motor events. More recent approaches have focused on oscillatory brain activity and synchronization or "phase-locking" between different cortical areas. Accordingly, this has produced increased interest in brain imaging methods with fine temporal resolution such as MEG. Rhythmic activity in the so-called gamma frequency band (30–90 Hz) is of particular interest since it is associated with cognitive processes such as feature binding within a sensory modality that may underlie perception (171). Recent studies have also described changes in neuromagnetic rhythms associated with observation and imitation of other individual's actions that appear to originate in brain areas associated with learning through imitation (172). Since changes in spontaneous brain rhythms are not necessarily time-locked to a stimulus onset, alternative signal processing techniques are required (173). The combination of spatial filtering source analysis methods and time frequency and phase analysis may be particularly well suited to measure these aspects of higher order brain function (141) and constitute a new and interesting avenue of research in human cognition.

Magnetoencephalography: Clinical Applications

Presurgical Functional Mapping. One of the more prevalent clinical applications of MEG is the localization of so-called "eloquent cortex" (those areas that subserve sensory, motor, speech, and memory function) prior to neurosurgery in order to prevent loss of these functions as a result of the surgical procedure. Due to displacement cortical tissue by space occupying lesions such as tumors, or natural variability in cortical morphology, identification of these brain

areas may not be possible by visual inspection alone and can be aided by functional localization of these areas using MEG. This is achieved by activating primary sensory areas associated with visual, somatosensory and auditory stimulation, and applying ECD models to the early evoked response—a method generally referred to as *presurgical functional mapping* (114,116). For example, the N20m source of the somatosensory evoked field can be consistently and reliably localized in most individuals and used as an estimate of the location of the central sulcus prior to surgical removal of brain tissue in the region of the primary motor or somatosensory cortex (174).

Determination of the language dominant hemisphere is also necessary prior to surgical resection of cortical tissue near language areas of the temporal lobe. This is routinely done through highly invasive procedures such as selective anesthesia of the left and right hemispheres (*Wada test*) or direct cortical stimulation intraoperatively. The use of MEG for the localization of brain areas that are specific to the processing of speech, as distinct from areas associated with the simple processing of auditory input, constitutes a challenging area of research, however, some recent progress has been made in this area (175–177). In addition to using MEG source imaging to identify functional and pathological brain areas in surgical planning, these functional data can also be incorporated into frameless stereotaxic *neuronavigation* systems. These systems allow surgeons to identify the corresponding brain areas in the functional and structural images during the surgical procedure. The MEG-based neuronavigation is rapidly becoming a useful clinical tool for the surgical treatment of epilepsy, tumors and other brain disorders (114,178,179).

Epilepsy. Due to its high temporal resolution and ability to localize focal brain activity, there has been a long interest in the application of MEG to the study of epilepsy. In many case, intractable seizures can be controlled by the removal of the epileptogenic zone: brain tissue from which seizure activity originates. The identification of this area may be aided by the measurement of abnormal electrical activity between seizures. These interictal (between seizure) spikes arise from an *irritative zone* that may be correlated with the epileptogenic zone in cases of focal epilepsy (180). The localization of ECD sources based on MEG recordings of interictal spiking activity has been shown to be highly correlated with the localization of this zone as identified by other methods such as direct intracranial monitoring from depth or subdural electrode grids [for recent reviews see (181–184)]. Interictal spikes are generally of much larger amplitude than sensory evoked responses and ECD models can be used to localize individual spike events without averaging. However, the area activated may be quite large and exhibit a high degree of spatial variability, and as a result the aggregate locations or clusters of many spike sources are often used to estimate the putative location of the irritative zone. Other methods, such as spatial filtering by beamformers (SAM), are currently being investigated and may help overcome some of the limitations of the single ECD approach to the localization of epileptogenic areas. Even in cases where the precise location of the epileptogenic zone is not clearly identified, MEG may help to guide the placement of subdural grids, and in some cases may be used to evaluate the propagation of abnormal electrical activity between multiple brain regions. The diagnostic yield of MEG measurements of interictal activity varies with different forms of epilepsy and appears to be highest for neocortical epilepsy (185) and can also aid in the differentiation of different types of epilepsy (186).

Since the site of brain pathology may not be known in advance, particularly in nonlesional epilepsy, the introduction of whole-head MEG systems has drastically improved the feasibility of using MEG as a routine clinical procedure for presurgical epilepsy evaluation allowing the data to be acquired in a more rapid and standardized manner. The main drawbacks to the application of MEG in epilepsy is the inability to measure brain activity during or just prior to seizure onset due to head movement, and the difficulty in performing long-term monitoring of interictal activity, although this is somewhat ameliorated by the introduction of MEG systems that allow recording from patients in the supine position and while asleep. Although there is some debate on the overall usefulness of MEG in the presurgical evaluation of epilepsy (183) comparisons with other modalities such as EEG, functional MRI, and intracranial electrical recordings (114,182) indicate that MEG provides useful complementary, and in some cases unique information for the surgical treatment of epilepsy.

Other Clinical Applications. Although presurgical functional mapping and epilepsy have been the main areas of clinical application of MEG, other brain disorders have been studied. This includes the use of MEG to study changes in electrical brain activity associated with tumors that often manifests as abnormal low frequency activity (187) or other disturbances in brain rhythmic activity (188) and may help identify the functional integrity of surrounding brain tissue (189). Abnormal low frequency activity has been associated with other brain lesions such as those due to stroke and in epilepsy (190). The MEG has also been used to study recovery after stroke due to functional reorganization of the cortex (191) and its relationship to rehabilitation and outcome (192) and in the evaluation of patients with mild head injury (193). Low frequency neuromagnetic activity has been hypothesized to be an index of spreading cortical depression associated with migraine (194).

The MEG studies have focused on pain related brain responses by selectively stimulating the $A\delta$ and C fiber systems painful CO_2 laser stimulation of the skin (195) or direct electrical stimulation of nerve fibers (196). This type of somatosensory stimulation produces long latency responses in secondary somatosensory areas located in the parietal operculum, and insula: brain regions known to be involved in the perception of pain. Such studies are promising for the clinical treatment of chronic pain, although are challenging due to the difficulty in discriminating activation of brain areas due to painful versus nonpainful somatosensory input, and the invasiveness of the procedure.

More recently, MEG measures have also been combined with neurochemical imaging methods such as magnetic resonance spectroscopy (MRS). In these studies, correlations have been found between MEG activity and changes in levels of neurotransmitter and other brain metabolites in ischemia or in brain areas harboring tumors (197). Although still a new area of study, there is a great deal of interest in the application of MEG to psychiatric disorders. For example, MEG studies have reported abnormal auditory evoked magnetic fields in schizophrenic patients (198) and patients with Alzheimer's disease (199) or in individuals with developmental disorders such as autism (200).

Magnetocardiography

The first biomagnetic measurements in humans were measurements of the magnetic field of the heart. The field of magnetocardiography or MCG has not expanded as rapidly as that of MEG, although a number of research centers have continued to develop the MCG method for the noninvasive evaluation of cardiac disease. As described in the Instrumentation section, MCG requires instrumentation designed for the adequate sampling of the heart's magnetic field over the chest and a number of instruments have been developed and installed at research centers around the world, including systems based on high temperature SQUIDs. Source modeling based on magnetic field measurements is somewhat simplified in the case of MEG due the ability to model the head as a spherically shaped conductor, whereas, modeling of the electrical activity of the heart requires realistic models of the conducting properties of the thorax and its influence on the distribution of magnetic fields arising from the heart. As a result, source localization methods in MCG often employ boundary element methods for forward calculations (201). Source localization in MCG is further complicated by the continuous movement of the heart itself. Nevertheless, MCG has been successfully used in the diagnosis of cardiac disease. For recent reviews see (202–204).

Since the 1980s a number of studies have focused on the use of MCG for the 3D localization of the origins of abnormal electrical activity of the heart. This includes abnormal activity underlying cardiac arrhythmias, such as Wolff–Parkinson–White syndrome, which involves abnormal electrical activity (preexcitation) in the accessory pathway. Recent studies have shown that MCG studies provide more accurate localization of the site of pathology than standard multichannel electrocardiogram techniques (205). The identification of the generators of heart arrhythmias is useful in presurgical evaluation for interventional procedures such as catheter ablation therapy, or in the screening of patients at risk for ventricular tachycardia (202) or coronary artery disease (206). Another application of MCG is in the assessment of ischemic areas of the heart after infarction by the detection of regions of low current density (207).

Fetal Studies

One of the more intriguing new applications of biomagnetism is the noninvasive measurement of activity of the fetal heart and brain. Since the first report of the detection of an evoked response from the fetal brain in 1985 (208) there has been a great of interest in developing instrumentation for the measurement of biomagnetic fields from the human fetus. The main challenges for the measurement of fetal MCG or MEG is the detection of biomagnetic sources that are distant from the detector array, and the difficulty in establishing the position of the fetal heart and brain during measurement. The latter has been partially resolved by the combination of fetal biomagnetic measurements with 3D ultasound imaging and new instrumentation has been recently designed for optimum placement and sensitivity of the sensory array to detect fetal heart and brain responses.

Fetal MCG. The largest biomagnetic signal arising from the fetus is the fetal magnetocardiogram or fMCG. The first recording of fMCG was demonstrated in 1984 (209). The fMCG signal magnitude is quite large, but due to proximity of the fetus to the mother's heart, signal processing methods are required to first remove the large maternal heart signal, after which the P, QRS, and T segments of the fMCG can be discerned with high reliability in fetuses beyond the twentieth week of gestation (210). Fetal heart rate variability has been shown to be a good indicator of fetal well being (211). This method has been applied to the detection of fetal arrhythmias (212) and may provide a useful diagnostic or screening tool for fetal congenital heart defects (213), or for the assessment of fetal health in high risk pregnancies. An overview of fMCG can be found in (214).

Fetal MEG. Due to the distance of the fetal brain from the surface of the maternal abdomen, the fetal MEG (fMEG) signal is difficult to detect without high sensitivity biomagnetometer with large coverage, large number of channels, and optimal placement of the sensor array. In addition, the fetal brain signals are small in comparison with an adult and their measurement is performed in the presence of strong interference from the maternal and fetal heart signals and various abdominal signals (intestinal electrical activity, uterine contractions, etc.). Measurements of fetal brain responses to sensory stimulation are also hampered by the difficulty in delivering the sensory stimulus to the fetus. However, fetal auditory evoked responses have been successfully recorded by presenting high amplitude auditory stimuli directly to the mother's abdomen (215,216). In order to successfully eliminate the interference due to cardiac signals, which can be >100 times larger than the fMEG, the latter efforts employed various signal extraction methods (spatial filtering, PCA, etc.) in addition to averaging. Magnitudes of fMEG responses to transient tone bursts are in the range of $\sim$8–180 fT and the latencies range from $\sim$125 to nearly 300 ms, decreasing with the increasing gestation age (217). The response is typically observed in not more than about 50% of examined subjects. Fetal responses to steady-state auditory clicks have also been reported (218) as well as spontaneous fetal brain activity in the form of burst suppression (219). The strength of these signals can be

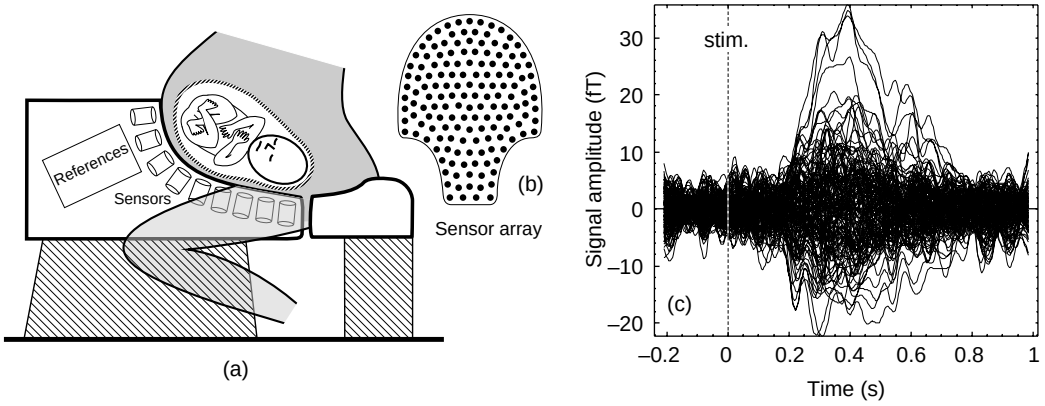


Figure 16. Dedicated system for fetal MEG measurement. (a) Schematic diagram of SQUID Array for Reproductive Assessment (SARA) (52). (b) Layout of 151 sensing channels. (c) Example of flash evoked fMEG response, overlay of 151 SARA channels. The fetus with gestation age of 28 weeks was stimulated by 33 ms duration flashes of 625 nm wavelength light (220). Vertical dashed line corresponds to the flash stimulus onset. (Adapted from Ref. 221).

relatively large, up to 500 fT, and can represent interference during evoked fMEG responses.

Early fMEG experiments used single or multiple-channel probes with relatively small area of coverage, requiring a search for the region with the largest signals. Recently, a dedicated fetal MEG system, the SQUID Array for Reproductive Assessment (SARA), was constructed and operated (52). The SARA system consists of an array of 151 SQUID sensors covering the mother's anterior abdominal surface in late gestation, from the perineum to the top of the uterus as shown in Fig. 16a and b. The primary sensor flux transformers are axial first-order gradiometers, with 8 cm baseline with a nominal SQUID sensor noise density of 4 $fT/Hz^{1/2}$. The SARA system is now being used routinely and was recently used to measure the first fetal visual evoked field to high intensity light stimuli presented to the maternal abdomen (220). An example of a flash evoked response from the fetal brain is shown in Fig. 16c.

Other Applications

Biosusceptometry. The greatest interest in biosusceptometry has stemmed from its potential to assess noninvasively iron overload in the human liver. This potentially fatal condition arises in individuals with hemoglobinopathies that require frequent blood transfusions (e.g., sickle cell anemia) or involve abnormal production of hemoglobin (hemachromatosis and beta-thalassemia). Standard methods for assessing iron overload can be highly invasive (e.g., liver biopsy) and biosusceptometry offers a safer and potentially more accurate diagnostic tool. Iron, which is normally strongly ferromagnetic, is stored in the liver bound by the proteins ferritin and hemosiderin and exhibits a strong paramagnetic response. As a result, measurement of the magnetic moment produced by placing the liver in a uniform magnetic field will be proportional to the total amount of iron in the liver: a method known as *biomagnetic liver susceptometry* (BLS). The basis for this technique was first proposed and the first measurements carried out in the late 1970s (222).

Most approaches to the measurement of hepatic iron concentration involve placing the patient's abdomen directly under a magnetic sensor that also contains a field coil that produces a magnetic field, and lowering the patient by a fixed distance to measure the change in field amplitude due to the magnetized liver (Fig. 11). In order to eliminate the effect of the surrounding air, a water-filled bellows is placed between the abdomen and the device to simulate the diamagnetic properties of the other tissues in the body. The main challenge to accurate estimates of hepatic iron content using BLS is the remaining effect of the varying susceptibility of the lungs and air filled compartments in the abdomen. Since this technique requires the application of a dc magnetic field to the body on the order of about 0.1 T, it is a much more invasive technology in comparison to MEG and MCG, and may be contraindicated in patients with implanted medical devices such as pacemakers. A detailed review of the clinical applications of BLS can be found in (223).

Peripheral Nerve Studies. It is known from the pioneering studies of Wikswo and colleagues (70) that the propagation of action potentials in nerve fibers produces quadrupolar like sources that have a rapidly diminishing magnetic field with distance. This is due to the fact that action potentials consist of a traveling wave of depolarization in the axon, followed closely by a wave of repolarization. In addition, due to varying conduction velocities in the peripheral nerves, action potentials in different axons will not necessarily summate to produce coherent synchronous activity. As a result, activation of compound nerve bundles does not produce coherent dipole-like sources as in the case of the neocortex. However, with sufficient signal averaging it is possible to record the magnetic signature of the sensory nerve action potentials noninvasively in the human: a technique

referred to as *magnetoneurography*. These measures have been achieved by placing single channel magnetometers or flat arrays of magnetic sensors over the peripheral nerve pathways and electrically stimulating the nerve. The predicted quadrapolar pattern of traveling actions potentials resulting from electrical stimulation of the finger was reported by Hoshiyama et al. (224) using a 12 channel "micro-SQUID" device placed over the wrist. Mackert et al. (225), using a 49 channel flat triangular array of first-order radial gradiometers were able to measure compound action potentials elicited by tibial nerve stimulation in sensory nerves entering the spinal cord at the lower lumbar region, and have recently using this method clinically to demonstrate impaired nerve conduction in the patients with S1 root compression.

Magnetopneumography. Magnetopneumography refers to the measurement of the remanent magnetism of ferromagnetic particles in the lungs. This technique may be used to assess lung contamination encountered in occupations that may involve the inhalation of ferromagnetic dust particles such as arc-welders, coalminers, asbestos, and foundry and steel workers. Similar to liver biosusceptometry, magnetopneumography involves the application of a weak dc magnetic field to the thorax. However, the field is applied for only a short interval in order to produce a remanent magnetization of ferromagnetic material, usually iron oxides such as magnetite. This remanent magnetic field is then measured to assess to total load of ferromagnetic particles in the lung. These measures can be used to evaluate the quantity and clearance rates of these substances (226,227). A related measure is *relaxation*: the decay of the remanent field due to the reorientation of the magnetic particles away from their aligned state after application of the dc field. Relaxation times are thought to reflect cellular processes in the lung associated with clearance or macrophage activity on the foreign particles. Recent studies have used magnetopneumography to study the effect of smoking on clearance times of inhaled magnetic particles (228).

Gastrointestinal System. Biomagnetic measurements have also been applied to other areas of the human body. The human gastrointestinal system produces electrical activity associated with the processes of peristalsis and digestion of food. For example, slow electrical activity at frequencies of $\sim$3 cycles/min (0.05 Hz) can be recorded from the human stomach using cutaneous surface electrodes or magnetically: a technique referred to as *magnetogastrography* (MGG). This activity arises from the smooth muscle of the stomach and the detection of changes in frequency with time has been proposed as a method of characterizing gastric disorders (229). Another novel application of biomagnetic instrumentation to gastrointestinal function, is the 3D tracking of the transport of magnetic materials through the gut. This technique has been termed *magnetic marker monitoring* (MMM) and can be used to monitor the passage and disintegration (by measuring decrease in magnetic moment) of magnetically labeled pharmaceutical substances through the gastrointestinal system (230).

FUTURE DIRECTIONS

Since its inception 40 years ago with the first recording of the magnetic field of the heart, the field of biomagnetism has expanded immensely to become a major field of basic and applied research. The field of magnetoencephalography, or MEG, has in recent years become a recognized neuroimaging technique, with the development of advanced instruments for the measurement of the electrical activity of the brain with exquisite temporal and spatial resolution. Biomagnetic instrumentation is now at a mature state, with commercially developed measurement systems available for a variety of biomagnetic applications. For example, whole head MEG systems are installed worldwide in $>$100 research laboratories and clinical centres and are now being used in routine clinical diagnostic procedures. Nevertheless, there remain many areas for further improvement of both instrumentation and data analysis approaches and techniques. In terms of instrumentation, biomagnetometer systems with increased number of sensing channels and capable of unshielded operation will likely be developed, and present systems that require frequent refilling with liquid Helium may be replaced by systems with longer hold times and less frequent cryogen replenishment. The latter may be accomplished either by incorporation of cryocoolers, or the use of sensors that do not require liquid He. The last two technical innovations, combined with production of larger numbers of MEG systems will also help reduce the cost of these instruments.

The analysis and interpretation of biomagnetic measurements is possibly the most significant area for continued research and development, and much progress has been made in the implementation of new signal processing algorithms for the extraction of biomagnetic signals, or improving the spatial resolution of source localization methods. There has been recent interest in combining MEG with its high temporal resolution and other functional imaging techniques such as functional MRI. In addition, advanced image processing techniques, such as the automated extraction of the cortical surface of the brain from structural MRI, will allow the use of more precise physical models of biomagnetic sources. Combination of MEG with its counterpart EEG may also help to develop more accurate models of brain activity. These advancements will aid the development of new clinical applications of biomagnetism such as the use of MEG to study psychiatric disorders, or to study the effects of drug treatments on brain processes related to cognitive deficits, or gain insight into the physiological mechanisms underlying various brain disorders in children, for example, learning disabilities, dyslexia, and autism. Finally, novel applications of biomagnetic measurements, for example, the measurement of heart and brain activity in the fetus, will lead to new applications of biomagnetism in clinical medicine and will further drive the development of improved technology. In sum, biomagnetism will continue

to grow as a novel and powerful noninvasive technique for the study of physiological processes in humans in both health and disease.

BIBLIOGRAPHY

Cited References

1. Baule GM, McFee R. Detection of the magnetic field of the heart. Am Heart J 1963;66:95–96.
2. Cohen D. Magnetoencephalography: evidence of magnetic fields produced by alpha-rhythm currents. Science 1968; 161:784–786.
3. Josephson BD. Possible new effect in superconductive tunneling. Phys Lett 1962;1:251–253.
4. Zimmerman JE. Recent developments in superconducting devices. J Appl Phys 1971;42:30–37.
5. Cohen D, Edelsack EA, Zimmerman JE. Magnetocardiograms taken inside a shielded room with a superconducting point-contact magnetometer. Appl Phys Lett 1970;16:278–280.
6. Cohen D. Magnetoencephalography: detection of the brain's electrical activity with a superconducting magnetometer. Science 1972;175:664–666.
7. Cheyne D, Vrba J, Crisp D, Betts K, Burbank M, Cheung T, Fife A, Haid G, Kubik P, Lee S, McCubbin J, McKay J, McKenzie D, Spear P, Taylor B, Tillotson M, Weinberg H, Basar E. Use of an unshielded, 64 channel, whole cortex MEG system in the study of normal and pathological brain function. Proceedings of the Satellite Symposium on Neuroscience and Technology, 14th Annu Conf IEEE Eng Med Biol Soc Lyon. France: 1992; 46–50.
8. Ahonen AI, Hämäläinen MS, Kajola MJ, Knuutila JET, Laine PP, Lounasmaa OV, Simola JT, Tesche CD, Vilkman VA. A 122-channel magnetometer covering the whole head. Proceedings of the Satellite Symposium on Neuroscience and Technology, 14th Annu Conf IEEE Eng Med Biol Soc Lyon. France: 1992; 16–20.
9. Clarke J. SQUIDs: theory and practice. In: Weinstock H, Ralston RW, editors. The New Superconducting Electronics. Dordrecht, Boston: Kluwer Academic; 1993. p 123–180.
10. Vrba J, Robinson SE. Signal processing in magnetoencephalography. Methods 2001;25:249–271.
11. Jaklevic R, Lambe RC, Silver AH, Mercereau JE. Quantum interference effects in Josephson tunneling. Phys Rev Lett 1964;12:159–160.
12. Jaycox JM, Ketchen MB. Planar coupling scheme for ultra low noise dc SQUIDs. IEEE Trans Magn 1981;MAG-17:400–403.
13. Ketchen MB, Jaycox JM. Ultra low noise tunnel junction dc SQUID with a tightly coupled planar input coil. Appl Phys Lett 1982;40:736–738.
14. Clarke J, Goubau WM, Ketchen MB. Tunnel junction DC SQUID: fabrication, operation, and performance. J Low Temp Phys 1976;25:99–144.
15. McKay J, Vrba J, Betts K, Burbank MB, Lee S, Mori K, Nonis D, Spear P, Uriel Y. Implementation of multichannel biomagnetic measurement system using DSP technology. Proc Can Conf Elect Comp Eng 1993; 1090–1093.
16. Robinson SE. A digital SQUID controller for unshielded biomagnetic measurement. In: Aine C, Okada Y, Stroink G, Swithenby S, Wood C, editors. Biomag96: Proc 10th Int Conf Biomag; 2000; New York: Springer; p 103–106.
17. Knuutial JET, Ahonen AI, Hamalainen MS, Kajola MJ, Petteri Laine P, Lounasmaa OV, Parkkonen LT, Simola JTA, Tesche CD. A 122-channel whole-cortex SQUID system for measureing the brain's magnetic fields. IEEE Trans Mag 1993;29:3315–3321.
18. Vrba J. Multichannel SQUID biomagnetic systems. In: Weinstock H, editor. Applications of Superconductivity. Dordrecht: Kluwer Academic Publishers; 2000. p 61–138.
19. Kominis IK, Kornack TW, Allred JC, Romalis MV. A subfemtotesla multichannel atomic magnetometer. Nature (London) 2003;422:596–599.
20. Zimmerman JE. SQUID instruments and shielding for low level magnetic measurements. J Appl Phys 1977;48:702–710.
21. H. G. Vacuumschmelze GmbH, Shielded room model AK-3.
22. Sullivan GW, Flynn ER. Performance of the Los Alamos Shielded Room. In: Atsumi K, Kotani M, Ueno S, Katila T, Williamson SJ, editors. Biomagnetism '87. Tokyo: Tokyo Denki University Press; 1987. p 486–489.
23. Erné SN, Hahlbohm HD, Scheer G, Trontelj Z. The Berlin magnetically shielded room (BMSR). In: Erné SN, editor. Biomagnetism. Berlin: Walter de Gruyter; 1981. p 79–87.
24. Bork J, Hahlbohm HD, Klein R, Schnabel A. The 8-layered magnetically shielded room of the PTB: Design and construction. In: Nenonen J, Ilmoniemi R, Katila T, editors. Biomag 2000. Proc 12th Int Conf Biomag. Espoo, Finland: Helsinki University of Technology; 2001. p 970–973.
25. Matsuba H, Shintomi K, Yahara A, Irisawa D, Imai K, Yoshida H, Seike S. Superconducting shielding enclosing a human body for biomagnetic measurement. In: Baumgartner C, Deecke L, Stroink G, Williamson SJ, editors. Biomagnetism: Fundamental research and clinical applications. Amsterdam, The Netherlands: Elsevier Science, IOS Press; 1995. p 483–489.
26. Matsumoto K, Yamagishi Y, Wakusawa A, Noda T, Fujioka K, Kuraoka Y. SQUID based active shield for biomagnetic measurements. In: Hoke M, Erné SN, Okada Y, Romani Gl, editors. Biomagnetism: clinical aspects. Proc 8th Int Conf Biomag. Amsterdam: Excerpta Medica; 1992. p 857–861.
27. Malmivuo J, Lekkala J, Kontro P, Suomaa I, Vihinin H. Improvement of the properties of an eddy current magnetic shield with active compensation. J Phys E: Sci Instr 1987;20:151–164.
28. ter Brake HJM, Wieringa HJ, Rogalla H. Improvement of the performance of a μ-metal magnetically shielded room by means of active compensation. Meas Sci Technol 1991;2: 596–601.
29. Vrba J. SQUID gradiometers in real environments. In: Weinstock H, editor. SQUID sensors: Fundamentals, Fabrication, and Applications. Dordrecht, Boston: Kluwer Academic Publishers; 1996. p 117–178.
30. Vrba J, Robinson SE. SQUID sensor array configurations for magnetoencephalography applications. Supercond Sci Technol 2002;15:R51–R89.
31. Vrba J. Magnetoencephalography: the art of finding a needle in a haystack. Phys C 2002;368:1–9.
32. Huotilainen M, Ilmoniemi RJ, Tiitinen H, Lavikainen J, Alho K, Kajola M, Naatanen R. The projection method in removing eye blink artefacts from multichannel MEG measurements. In: Baumgartner C, Deecke L, Stroink G, Williamson SJ, editors. Biomagnetism: Fundamental research and clinical applications. Elsevier Science, IOS Press; 1995. p 363–367.
33. Tesche CD, Uusitalo MA, Ilmoniemi RJ, Huotilainen M, Kajola M, Salonen O. Signal-space projections of MEG data characterize both distributed and well-localized neuronal sources. Electroencephalogr Clin Neurophys 1995; 95:189–200.

34. Uusitalo MA, Ilmoniemi RJ. Signal-space projection method for separating MEG or EEG into components. Med Biol Eng Comput 1997;35:135–140.
35. Parkkonen LT, Simola JT, Tuoriniemi JT, Ahonen AI. An interference suppression system for multichannel magnetic field detector arrays. In: Yoshimoto T, Kotani M, Kuriki S, Karibe H, Nakasato N, editors. Recent Advances in Biomagnetism. Sendai: Tohoku University Press; 1999. p 13–16.
36. Taula S, Kajola MJ, Simola JT. The Signal Space Separation method. 14th Conf Int Soc Brain Electromagnetic Topography. Santa Fe, NM; 2003.
37. Ioannides AA, Mütter J, Barna-Popescu EA. Irreducible tensor representation of MEG signals: Theory and applications. In: Nenonen J, Ilmoniemi R, Katila T, editors. Biomag 2000. Proc 12th Int Conf Biomag. Espoo, Finland: Helsinki University of Technology; 2001. p 883–886.
38. ter Brake HJM. Cryogenic systems for superconducting devices. In: Weinstock H, editor. Applications of superconductivity. Dordrecht: Kluwer Academic Publishers; 2000. p 561–639.
39. VSM MedTech Ltd. (CTF) 9 Burbidge St. Coquitlam, B.C., Canada (www.vsmmedtech.com).
40. Nowak H. Biomagnetic Instrumentation. In: Andrä W, Nowak H, editors. Magnetism in Medicine. Berlin: Wiley VCH; 1998. p 88–135.
41. Hoenig HE, Daalmans GM, Bär L, Bömmel F, Paulus A, Uhl D, Weisse HJ, Schneider S, Seifert H, Reichenberger H, Abraham-Fuchs K, Multichannel DC. SQUID sensor array for biomagnetic applications. IEEE Trans Magn 1991; 27:2777–2785.
42. Tsukada K, Kandori A, Miyashita T, Sasabuchi H, Suzuki H, Kondo S, Komiyama Y, Teshogawara K. A simplified superconducting quantum interference device system to analyze vector components of a cardiac magnetic field. Proc 20th Annu Int Conf IEEE/EMBS. Hong Kong; 1998. p 524–527.
43. Van Leeuwen P, Haupt C, Hoormann C, Hailer B, Mackert BM, Stroink G. A 67 channel biomagnetometer designed for cardiology and other applications. In: Yoshimoto T, Kotani M, Kuriki S, Karibe H, Nakasato N, editors. Recent Advances in Biomagnetism. Sendai: Tohoku University Press; 1999. p 89–92.
44. Erné SN, Pasquarelli A, Kammrath H, Della Penna S, Torquati K, Pizzella V, Rossi R, Granata C, Russo M. Argos 55 - the new MCG system in Ulm. In: Yoshimoto T, Kotani M, Kuriki S, Karibe H, Nakasato N, editors. Recent Advances in Biomagnetism. Sendai: Tohoku University Press; 1999. p 27–30.
45. Montonen J, Ahonen A, Hämäläinen M, Ilmoniemi R, Laine P, Nenonen J, Paavola M, Simelius K, Simola J, Katila T. Magnetocardiographic functional imaging studies in BioMag Laboratory. In: Aine CJ, Okada Y, Stroink G, Swithenby S, Wood C, editors. Biomag 96: Proc Tenth Int Conf Biomag. New York: Springer; 2000. p 494–497.
46. ter Brake HJM, Janssen N, Flokstra J, Veldhuis D, Rogalla H. Multichannel heart scanner based on high-Tc SQUIDs. IEEE Trans Appl Supercond 1997;7:2545–2548.
47. Seidel P, Schmidl F, Wunderlich S, Dörrer L, Vogt T, Schneidewind H, Weidl R, Lösche R, Leder U, Solbig O, Nowak H. High-Tc SQUID systems for practical use. IEEE Trans Appl Supercond 1999;9:4077–4080.
48. Kouznesov KA, Borgmann J, Clarke CJS. High-Tc second-order gradiometer for magnetocardiography in an unshielded environment. Appl Phys Lett 1999;75:1979–1981.
49. Zhang Y, Panaitov G, Wang SG, Wolters N, Otto R, Schubert J, Zander W, Krause HJ, Soltner H, Bousack H, Braginski A. Second-order, high-temperature superconducting gradiometer for magnetocardiography in an unshielded environment. Appl Phys Lett 2000;76:906–908.
50. Ludwig F, Jansman ABM, Drung D, Lindström M, Bechstien S, Beyer J, Flokstra J, Schurig T. Optimization of direct-coupled high-Tc SQUID magnetometers for operation in a magnetically shielded environment. IEEE Trans Appl Supercond 2001;11:1824–1827.
51. Barthelmess H-J, Halverscheid M, Schiefenhövel B, Heim E, Schilling M, Zimmerman R. Low-noise biomagnetic measurements with a multichannel dc-SQUID system at 77 K. IEEE Trans Appl Supercond 2001;11:657–660.
52. Robinson SE, BM, FA, HG, KP, I Sekachev, Taylor B, Tillotson M, Wong VJG, Lowery C, Eswaran H, Wilson D, Murphy P, Preissl H. A biomagnetic instrument for human reproductive assessment. In: Nenonen J, Ilmoniemi R, Katila T, editors. Biomag2000, Proc 12th Int Conf Biomag. Espoo, Finland: Helsinki University of Technology; 2001. p 919–922.
53. 4-D Neuroimaging Inc. 9727 Pacific Heights Blvd., San Diego, CA 92121-3719 (www.4dneuroimaging.com).
54. Neuromag Oy, P.O. Box 68, FIN-00511 Helsinki, Finland (www.neuromag.com).
55. Eagle Technology, Inc. 1-2-23 Hirosaka, Kanazawa Ishikawa 920-0962, Japan (www.eagle-tek.com).
56. Advanced Technologies Biomagnetics S.r.l, Via Martiri di Pietransieri 2, 65129 Pescara, Italy (www.atb-it.com).
57. Polhemus Inc., Hercules Drive, P.O. Box 560, Colchester, VT 05446.
58. Bramidis PD, Ioannides AA. Combination of point and surface matching techniques for accurate registration of MEG and MRI. In: Aine CJ, Okada Y, Stroink G, Swithenby S, Wood C, editors. Biomag 96: Proc 10th Int Conf Biomag. New York: Springer; 1997. p 1126–1129.
59. Abraham-Fuchs K, Lindner L, Weganer P, Nestel F, Schneider S. Fusion of biomagnetism with MRI or CT images by contour fitting. Biomed Eng 1991;36(Suppl.): 88–89.
60. Kober H, Nimsky C, Vieth J, Fahlbusch R, Ganslandt O. Coregistration of function and anatomy in frameless stereotaxy by contour fitting, Stereotact Funct. Neurosurg 2002;79:272–283.
61. Braginski A, Krause HJ, Vrba J. SQUID magnetometers. In: Francombe MH, editor. Handbook of Thin Film Devices, Volume 3: Superconducting Film Devices. San Diego: Academic Press; 2000. p 149–225.
62. Farrell DE, Tripp JH, Zanzucchi PE. Magnetic measurements of human iron stores. IEEE Transactions on Magnetics May-1980;16:818–823.
63. Paulson DN, Fagaly RL, Toussaint RM, Fischer R. Biomagnetic susceptometer with SQUID instrumentation. IEEE Trans Magn 1991;27:3249–3252.
64. Fischer F. Liver iron susceptometry. In: Andrä W, Nowak H, editors. Magnetism in medicine: a handbook. Berlin; New York: Wiley-VCH; 1998. p 286–301.
65. Helmholtz H. Über einige Gesetze der Vertheilung elektrischer Ströme in köperlichen Leitern mit Anwendung auf die thierisch-elektrischen Versuche. Ann Phys Chem 1853;89:211–233, 353–377.
66. Hämäläinen M, Hari R, Ilmoniemi RJ, Knuutila JET, Lounasmaa OV. Magnetoencephalography-theory, instrumentation, and applications to noninvasive studies of the working human brain. Rev Mod Phys 1993;65:413–497.
67. Jeffs B, Leahy R, Singh M. An evaluation of methods for neuromagnetic image reconstruction. IEEE Trans Biomed Eng 1987;34:713–723.

68. Baillet S, Mosher JC, Leahy R. Electromagnetic brain mapping. IEEE Signal Proc Mag 2001;18:14–30.
69. Mosher JC, Leahy RM, Lewis PS. EEG and MEG: forward solutions for inverse methods. IEEE Trans Biomed Eng 1999;46:245–259.
70. Wikswo JP, Barach JP, Freeman JA. Magnetic field of a nerve impulse: first measurements. Science 1980;208:53–55.
71. Swinney KR, Wikswo, Jr. JP. A calculation of the magnetic field of a nerve action potential. Biophys J 1980;32:719–731.
72. Tesche CD. Non-invasive imaging of neuronal population dynamics in human thalamus. Brain Res 1996;729:253–258.
73. Tesche CD, Karhu J, Tissari SO. Non-invasive detection of neuronal population activity in human hippocampus. Brain Res Cogn Brain Res 1996;4:39–47.
74. Humphrey DR. Re-analysis of the antidromic cortical response. II. On the contribution of cell discharge and PSPs to the evoked potentials. Electroencephalogr Clin Neurophysiol 1968;25:421–442.
75. Creutzfeldt OD, Watanabe S, Lux HD. Relations between EEG phenomena and potentials of single cortical cells. I. Evoked responses after thalamic and epicortical stimulation. Electroencephalogr Clin Neurophysiol 1966;20:1–18.
76. Mitzdorf U. Current source-density method and application in cat cerebral cortex: investigation of evoked potentials and EEG phenomena. Physiol Rev 1985;65:37–100.
77. Okada YC, Wu J, Kyuhou S. Genesis of MEG signals in a mammalian CNS structure. Electroencephalogr Clin Neurophysiol 1997;103:474–485.
78. Okada Y. Toward understanding the physiological origins of neuromagnetic signals. In: Lu Z-L, Kaufman L, editors. Magnetic Source Imaging of the Brain. Mahwah. New Jersey: Lawrence Erlbaum Associates; 2003. p 43–76.
79. Weinberg H, Brickett P, Coolsma F, Baff M. Magnetic localisation of intracranial dipoles: simulation with a physical model. Electroencephalogr Clin Neurophysiol 1986;64:159–170.
80. Cheyne D, Roberts LE, Gaetz W, Bosnyak D, Nahmias C, Christoforou N, Weinberg H. Somatotopic organization of human somatosensory cortex: a comparison of EEG, MEG and fMRI methods. In: Koga Y, Nagata K, Hirata K, editors. Brain Topography Today. Amsterdam: Elsevier Science; 1998. p 76–81.
81. Sarvas J. Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. Phys Med Biol 1987;32:11–22.
82. Hillebrand A, Barnes GR. A quantitative assessment of the sensitivity of whole-head MEG to activity in the adult human cortex. Neuroimage 2002;16:638–650.
83. Lu ZL, Williamson SJ. Spatial extent of coherent sensory-evoked cortical activity. Exp Brain Res 1991;84:411–416.
84. Hämäläinen MS, Ilmoniemi RJ. Interpreting measured magnetic fields of the brain: Estimates of current distribution. Report TKK-F-A559. Helsinki University of Technology: Espoo, Finland, 1984.
85. Scherg M, Von Cramon D. Two bilateral sources of the late AEP as identified by a spatio-temporal dipole model. Electroencephalogr Clin Neurophysiol 1985;62:32–44.
86. Huang MX, Mosher JC, Leahy RM. A sensor-weighted overlapping-sphere head model and exhaustive head model comparison for MEG. Phys Med Biol 1999;44:423–440.
87. Gorodnitsky IF, George JS, Rao BD. Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm. Electroencephalogr Clin Neurophysiol 1995;95:231–251.
88. Wagner M, Wischmann HA, Fuchs M, Kohler T, Drenckhahn R. Current density reconstruction using the L1 norm. In: Aine CJ, Okada Y, Stroink G, Swithenby S, Wood C, editors. Biomag96. New York: Springer-Verlag; 2000. p 393–396.
89. Ioannides AA, Bolton JPR, Clarke CJS. Continuous probabilistic solutions to the biomagnetic inverse problem. Inverse Problems 1990;6:523–542.
90. Pascual-Marqui RD, Michel CM, Lehmann D. Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. Int J Psychophysiol 1994;18:49–65.
91. David O, Garnero L, Cosmelli D, Varela FJ. Estimation of neural dynamics from MEG/EEG cortical current density maps: application to the reconstruction of large-scale cortical synchrony. IEEE Trans Biomed Eng 2002;49: 975–987.
92. Baillet S, Garnero L. A Bayesian approach to introducing anatomo-functional priors in the EEG/MEG inverse problem. IEEE Trans Biomed Eng 1997;44:374–385.
93. Schmidt DM, George JS, Wood CC. Bayesian inference applied to the electromagnetic inverse problem. Hum Brain Mapp 1999;7:195–212.
94. Hämäläinen MS, Haario H, Lehtinen MS. Inferences about sources of neuromagnetic fields using Bayesian parameter estimation. Espoo, Finland: Helsinki University of Technolgy; 1987.
95. Sorenson HW. Parameter estimation. New York: Marcel Dekker; 1980.
96. Robinson SE, Rose DF. Current source estimation by spatially filtered MEG. In: HokeM, Erné SN, Okada Y, Romani GL, editors. Biomagnetism: clinical aspects. Proc 8th Int Conf Biomag. Amsterdam: Excerpta Medica; 1992. p 761–765.
97. Frost OL. An algorithm for linearly constrained adaptive array processing. Proc IEEE 1972;60:926–935.
98. Van Veen B, Buckley K. Beamforming: A versatile approach to spatial filtering, in IEEE ASSP Mag.; 1988. p 4–24.
99. Godara LC. Application of antenna array to mobile communications, Part II: Beam-Forming and direction-of-arrival considerations. Proc IEEE 1997;85:1195–1245.
100. Van Veen BD, Van Drongelen W, Yuchtman M, Suzuki A. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Trans Biomed Eng 1997;44:867–880.
101. Robinson SE, Vrba J. Functional neuroimaging by synthetic aperture magnetometry (SAM). In: Yoshimoto T, Kotani M, Kuriki S, Karibe H, Nakasato N, editors. Recent Advances in Biomagnetism. Sendai: Tohoku University Press; 1999. p 302–305.
102. Vrba J, Robinson SE. Linearly constrained minimum variance beamformers, synthetic aperture magnetometry and MUSIC in MEG applications. IEEE, Proc 34th Asilomar Conf. Signals, Systems, Comput. Pacific Grove, CA: Omnipress; 2000. p 313–317.
103. Schmidt RO. Multiple emitter location and signal parameter estimation. IEEE Trans Anten Propagat 1986;AP-34:276–280.
104. Mosher JC, Lewis PS, Leahy RM. Multiple dipole modeling and localization from spatio-temporal MEG data. IEEE Trans Biomed Eng 1992;39:541–557.
105. Mosher JC, Leahy R. Source localization using recursively applied and projected (RAP) MUSIC. IEEE Trans Signal Proc 1999;47:332–340.
106. Maier J, Dagnelie G, Spekreijse H, van Dijk BW. Principal components analysis for source localization of VEPs in man. Vision Res 1987;27:165–177.
107. Achim A, Richer F, Saint-Hilaire JM. Methods for separating temporally overlapping sources of neuroelectric data. Brain Topogr 1988;1:22–28.

108. Jung TP, Makeig S, Mckeown MJ, Bell AJ, L. T, Sejnowski TJ. Imaging brain dynamics using independent component analysis. Proc IEEE 2001;89:1107–1122.
109. Cardoso J-F. Blind signal separation: Statistical principles. Proc IEEE 1998;86:2009–2025.
110. Hyvarinen A. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Networks 1999;10:626–634.
111. Makeig S, Jung TP, Bell AJ, Ghahremani D, Sejnowski TJ. Blind separation of auditory event-related brain responses into independent components. Proc Natl Acad Sci USA 1997;94:10979–10984.
112. Ziehe A, Muller KR, Nolte G, Mackert BM, Curio G. Artifact reduction in magnetoneurography based on time-delayed second-order correlations. IEEE Trans Biomed Eng 2000; 47:75–87.
113. Lewine JD, Orrison, Jr. WW. Magnetic source imaging: basic principles and applications in neuroradiology. Acad Radiol 1995;2:436–440.
114. Wheless JW, Castillo E, Maggio V, Kim HL, Breier JI, Simos PG, Papanicolaou AC. Magnetoencephalography (MEG) and magnetic source imaging (MSI). Neurologist 2004; 10:138–153.
115. Lu Z-L, Kaufman L, editors. Magnetic Source Imaging of the Brain. Mahwah, NJ: Lawrence Erlbaum Associates; 2003.
116. Gallen CC, Schwartz BJ, Bucholz RD, Malik G, Barkley GL, Smith J, Tung H, Copeland B, Bruno L, Assam S. Presurgical localization of functional cortex using magnetic source imaging. J Neurosurg 1995;82:988–994.
117. Roberts TP, Poeppel D, Rowley HA. Magnetoencephalography and magnetic source imaging. Neuropsychiat Neuropsychol Behav Neurol 1998;11:49–64.
118. Brenner D, Lipton J, Kaufman L, Williamson SJ. Somatically evoked magnetic fields of the human brain. Science 1978;199:81–83.
119. Kakigi R, Hoshiyama M, Shimojo M, Naka D, Yamasaki H, Watanabe S, Xiang J, Maeda K, Lam K, Itomi K, Nakamura A. The somatosensory evoked magnetic fields. Prog Neurobiol 2000;61:495–523.
120. Nakamura A, Yamada T, Goto A, Kato T, Ito K, Abe Y, Kachi T, Kakigi R. Somatosensory homunculus as drawn by MEG. Neuroimage 1998;7:377–386.
121. Curio G, Mackert BM, Burghoff M, Koetitz R, Abraham-Fuchs K, Harer W. Localization of evoked neuromagnetic 600 Hz activity in the cerebral somatosensory system. Electroencephalogr Clin Neurophysiol 1994;91:483–487.
122. Hashimoto I, Mashiko T, Imada T. Somatic evoked high-frequency magnetic oscillations reflect activity of inhibitory interneurons in the human somatosensory cortex. Electroencephalogr Clin Neurophysiol 1996;100:189–203.
123. Ikeda H, Leyba L, Bartolo A, Wang Y, Okada YC. Synchronized spikes of thalamocortical axonal terminals and cortical neurons are detectable outside the pig brain with MEG. J Neurophysiol 2002;87:626–630.
124. Hari R, Karhu J, Hämäläinen MS, Knuutila J, Salonen O, Sams M, Vilkman V. Functional organization of the human first and second somatosensory cortices: a neuromagnetic study. Eur J Neurosci 1993;5:724–734.
125. Suk J, Ribary U, Cappell J, Yamamoto T, Llinas R. Anatomical localization revealed by MEG recordings of the human somatosensory system. Electroencephalogr Clin Neurophysiol 1991;78:185–196.
126. Flor H, Elbert T, Knecht S, Wienbruch C, Pantev C, Birbaumer N, Larbig W, Taub E. Phantom-limb pain as a perceptual correlate of cortical reorganization following arm amputation. Nature (London) 1995;375:482–484.
127. Elbert T, Pantev C, Wienbruch C, Rockstroh B, Taub E. Increased cortical representation of the fingers of the left hand in string players. Science 1995;270:305–307.
128. Deecke L, Weinberg H, Brickett P. Magnetic fields of the human brain accompanying voluntary movement: Bereitschaftsmagnetfeld. Exp Brain Res 1982;48:144–148.
129. Cheyne D, Weinberg H. Neuromagnetic fields accompanying unilateral finger movements: pre-movement and movement-evoked fields. Exp Brain Res 1989;78:604–612.
130. Kristeva R, Cheyne D, Lang W, Lindinger G, Deecke L. Movement-related potentials accompanying unilateral and bilateral finger movements with different inertial loads. Electroencephalogr Clin Neurophysiol 1990;75:410–418.
131. Cheyne D, Endo H, Takeda T, Weinberg H. Sensory feedback contributes to early movement-evoked fields during voluntary finger movements in humans. Brain Res 1997;771:196–202.
132. Kelso JA, Fuchs A, Lancaster R, Holroyd T, Cheyne D, Weinberg H. Dynamic cortical activity in the human brain reveals motor equivalence. Nature (London) 1998;392:814–818.
133. Endo H, Kizuka T, Masuda T, Takeda T. Automatic activation in the human primary motor cortex synchronized with movement preparation. Cogn Brain Res 1999;8:229–239.
134. Xiang J, Hoshiyama M, Koyama S, Kaneoke Y, Suzuki H, Watanabe S, Naka D, Kakigi R. Somatosensory evoked magnetic fields following passive finger movement. Brain Res Cogn Brain Res 1997;6:73–82.
135. Lang W, Cheyne D, Hollinger P, Gerschlager W, Lindinger G. Electric and magnetic fields of the brain accompanying internal simulation of movement. Cogn Brain Res 1996;3:125–129.
136. Conway BA, Halliday DM, Farmer SF, Shahani U, Maas P, Weir AI, Rosenberg JR. Synchronization between motor cortex and spinal motoneuronal pool during the performance of a maintained motor task in man. J Physiol 1995;489(Pt. 3): 917–924.
137. Brown P. Cortical drives to human muscle: the Piper and related rhythms. Prog Neurobiol 2000;60:97–108.
138. Salenius S, Avikainen S, Kaakkola S, Hari R, Brown P. Defective cortical drive to muscle in Parkinson's disease and its improvement with levodopa. Brain 2002;125: 491–500.
139. Timmermann L, Gross J, Dirks M, Volkmann J, Freund HJ, Schnitzler A. The cerebral oscillatory network of parkinsonian resting tremor. Brain 2003;126:199–212.
140. Hari R, Salmelin R. Human cortical oscillations: a neuromagnetic view through the skull. Trends Neurosci 1997;20: 44–49.
141. Cheyne D, Gaetz W, Garnero L, Lachaux JP, Ducorps A, Schwartz D, Varela FJ. Neuromagnetic imaging of cortical oscillations accompanying tactile stimulation. Brain Res Cogn Brain Res 2003;17:599–611.
142. Feige B, Kristeva-Feige R, Rossi S, Pizzella V, Rossini PM. Neuromagnetic study of movement-related changes in rhythmic brain activity. Brain Res 1996;734:252–260.
143. Gaetz WC, Cheyne DO. Localization of human somatosensory cortex using spatially filtered magnetoencephalography. Neurosci Lett 2003;340:161–164.
144. Taniguchi M, Kato A, Fujita N, Hirata M, Tanaka H, Kihara T, Ninomiya H, Hirabuki N, Nakamura H, Robinson SE, Cheyne D, Yoshimine T. Movement-related desynchronization of the cerebral cortex studied with spatially filtered magnetoencephalography. Neuroimage 2000;12:298–306.
145. Brenner D, Williamson SJ, Kaufman L. Visually evoked magnetic fields of the human brain. Science 1975;190:480–482.
146. Supek S, Aine CJ, Ranken D, Best E, Flynn ER, Wood CC. Single vs. paired visual stimulation: superposition of early

neuromagnetic responses and retinotopy in extrastriate cortex in humans. Brain Res 1999;830:43–55.
147. Fylan F, Holliday IE, Singh KD, Anderson SJ, Harding GF. Magnetoencephalographic investigation of human cortical area V1 using color stimuli. Neuroimage 1997;6:47–57.
148. Singh KD, Barnes GR, Hillebrand A, Forde EM, Williams AL. Task-related changes in cortical synchronization are spatially coincident with the hemodynamic response. Neuroimage 2002;16:103–114.
149. Anderson SJ, Holliday IE, Singh KD, Harding GF. Localization and functional analysis of human cortical area V5 using magneto-encephalography. Proc R Soc London Sev B Biol Sci 1996;263:423–431.
150. Maruyama K, Kaneoke Y, Watanabe K, Kakigi R. Human cortical responses to coherent and incoherent motion as measured by magnetoencephalography. Neurosci Res 2002;44:195–205.
151. Anderson SJ, Holliday IE, Harding GF. Assessment of cortical dysfunction in human strabismic amblyopia using magnetoencephalography (MEG). Vision Res 1999;39:1723–1738.
152. Reite M, Edrich J, Zimmerman JT, Zimmerman JE. Human magnetic auditory evoked fields. Electroencephalogr Clin Neurophysiol 1978;45:114–117.
153. Romani GL, Williamson SJ, Kaufman L, Brenner D. Characterization of the human auditory cortex by the neuromagnetic method. Exp Brain Res 1982;47:381–393.
154. Pantev C, Hoke M, Lehnertz K, Lutkenhoner B, Anogianakis G, Wittkowski W. Tonotopic organization of the human auditory cortex revealed by transient auditory evoked magnetic fields. Electroencephalogr Clin Neurophysiol 1988;69: 160–170.
155. Pantev C, Lutkenhoner B. Magnetoencephalographic studies of functional organization and plasticity of the human auditory cortex. J Clin Neurophysiol 2000;17:130–142.
156. Roberts TP, Ferrari P, Stufflebeam SM, Poeppel D. Latency of the auditory evoked neuromagnetic field components: stimulus dependence and insights toward perception. J Clin Neurophysiol 2000;17:114–129.
157. Weinberg H, Cheyne D, Brickett P, Harrop R, Gordon R. An interaction of cortical sources associated with simultaneous auditory and somatosensory stimulation. In: Pfurtscheller G, Lopes da Silva FH, editors. Functional Brain Imaging. Lewiston, N.Y.: Hans Huber Publishers; 1988. p 83–88.
158. Ribary U, Ioannides AA, Singh KD, Hasson R, Bolton JP, Lado F, Mogilner A, Llinas R. Magnetic field tomography of coherent thalamocortical 40-Hz oscillations in humans. Proc Natl Acad Sci USA 1991;88:11037–11041.
159. Ross B, Picton TW, Pantev C. Temporal integration in the human auditory cortex as represented by the development of the steady-state magnetic field. Hear Res 2002;165:68–84.
160. Starr A, Kristeva R, Cheyne D, Lindinger G, Deecke L. Localization of brain activity during auditory verbal short-term memory derived from magnetic recordings. Brain Res 1991;558:181–190.
161. Mecklinger A, Maess B, Opitz B, Pfeifer E, Cheyne D, Weinberg H. A MEG analysis of the P300 in visual discrimination tasks. Electroencephalogr Clin Neurophysiol 1998;108:45–56.
162. Alho K. Cerebral generators of mismatch negativity (MMN) and its magnetic counterpart (MMNm) elicited by sound changes. Ear Hear 1995;16:38–51.
163. Halgren E, Raij T, Marinkovic K, Jousmaki V, Hari R. Cognitive response profile of the human fusiform face area as determined by MEG. Cereb Cortex 2000;10:69–81.
164. Salmelin R, Hari R, Lounasmaa OV, Sams M. Dynamics of brain activation during picture naming. Nature (London) 1994;368:463–465.
165. Sams M, Aulanko R, Hamalainen M, Hari R, Lounasmaa OV, Lu ST, Simola J. Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neurosci Lett 1991;127:141–145.
166. Pulvermuller F, Shtyrov Y, Ilmoniemi R. Spatiotemporal dynamics of neural language processing: an MEG study using minimum-norm current estimates. Neuroimage 2003;20:1020–1025.
167. Pylkkanen L, Marantz A. Tracking the time course of word recognition with MEG. Trends Cogn Sci 2003;7:187–189.
168. Roberts TP, Ferrari P, Poeppel D. Latency of evoked neuromagnetic M100 reflects perceptual and acoustic stimulus attributes, Neuroreport 1998;9:3265–3269.
169. Helenius P, Salmelin R, Service E, Connolly JF. Semantic cortical activation in dyslexic readers. J Cogn Neurosci 1999;11:535–550.
170. Salmelin R, Schnitzler A, Schmitz F, Jancke L, Witte OW, Freund HJ. Functional organization of the auditory cortex is different in stutterers and fluent speakers. Neuroreport 1998;9:2225–2229.
171. Tallon-Baudry C, Bertrand O. Oscillatory gamma activity in humans and its role in object representation. Trends Cogn Sci 1999;3:151–162.
172. Nishitani N, Hari R. Temporal dynamics of cortical representation for action. Proc Natl Acad Sci USA 2000;97:913–918.
173. Varela F, Lachaux JP, Rodriguez E, Martinerie J. The brainweb: phase synchronization and large-scale integration. Nat Rev Neurosci 2001;2:229–239.
174. Gallen CC, Sobel DF, Waltz T, Aung M, Copeland B, Schwartz BJ, Hirschkoff EC, Bloom FE. Noninvasive presurgical neuromagnetic mapping of somatosensory cortex. Neurosurgery 1993;33:260–268; discussion 268.
175. Papanicolaou AC, Simos PG, Castillo EM, Breier JI, Sarkari S, Pataraia E, Billingsley RL, Buchanan S, Wheless J, Maggio V, Maggio WW. Magnetocephalography: a noninvasive alternative to the Wada procedure. J Neurosurg 2004;100:867–876.
176. Naatanen R, Lehtokoski A, Lennes M, Cheour M, Huotilainen M, Iivonen A, Vainio M, Alku P, Ilmoniemi RJ, Luuk A, Allik J, Sinkkonen J, Alho K. Language-specific phoneme representations revealed by electric and magnetic brain responses. Nature (London) 1997;385:432–434.
177. Bowyer SM, Moran JE, Mason KM, Constantinou JE, Smith BJ, Barkley GL, Tepley N. MEG localization of language-specific cortex utilizing MR-FOCUSS. Neurology 2004;62: 2247–2255.
178. Firsching R, Bondar I, Heinze HJ, Hinrichs H, Hagner T, Heinrich J, Belau A. Practicability of magnetoencephalography-guided neuronavigation. Neurosurg Rev 2002; 25:73–78.
179. Ganslandt O, Fahlbusch R, Nimsky C, Kober H, Moller M, Steinmeier R, Romstock J, Vieth J. Functional neuronavigation with magnetoencephalography: outcome in 50 patients with lesions around the motor cortex. J Neurosurg 1999;91:73–79.
180. Rosenow F, Luders H. Presurgical evaluation of epilepsy. Brain 2001;124:1683–1700.
181. Stefan H, Hummel C, Scheler G, Genow A, Druschky K, Tilz C, Kaltenhauser M, Hopfengartner R, Buchfelder M, Romstock J. Magnetic brain source imaging of focal epileptic activity: a synopsis of 455 cases. Brain 2003.
182. Barkley GL. Controversies in neurophysiology. MEG is superior to EEG in localization of interictal epileptiform activity. Pro Clin Neurophysiol 2004;115:1001–1009.
183. Baumgartner C. Controversies in clinical neurophysiology. MEG is superior to EEG in the localization of interictal epileptiform activity: Con. Clin Neurophysiol 2004;115: 1010–1020.

184. Otsubo H, Snead III OC. Magnetoencephalography and magnetic source imaging in children. J Child Neurol 2001;16:227–235.
185. Stefan H, Hummel C, Hopfengartner R, Pauli E, Tilz C, Ganslandt O, Kober H, Moler A, Buchfelder M. Magnetoencephalography in extratemporal epilepsy. J Clin Neurophysiol 2000;17:190–200.
186. Baumgartner C, Pataraia E, Lindinger G, Deecke L. Magnetoencephalography in focal epilepsy. Epilepsia 2000;41 (Suppl 3): S39–47.
187. de Jongh A, Baayen JC, de Munck JC, Heethaar RM, Vandertop WP, Stam CJ. The influence of brain tumor treatment on pathological delta activity in MEG. Neuroimage 2003;20:2291–2301.
188. Taniguchi M, Kato A, Ninomiya H, Hirata M, Cheyne D, Robinson SE, Maruno M, Saitoh Y, Kishima H, Yoshimine T. Cerebral motor control in patients with gliomas around the central sulcus studied with spatially filtered magnetoencephalography. J Neurol Neurosurg Psychiat 2004;75: 466–471.
189. Schiffbauer H, Ferrari P, Rowley HA, Berger MS, Roberts TP. Functional activity within brain tumors: a magnetic source imaging study. Neurosurgery 2001;49:1313–1320; discussion 1320–1311.
190. Gallen CC, Tecoma E, Iragui V, Sobel DF, Schwartz BJ, Bloom FE. Magnetic source imaging of abnormal low-frequency magnetic activity in presurgical evaluations of epilepsy. Epilepsia 1997;38:452–460.
191. Rossini PM, Tecchio F, Pizzella V, Lupoi D, Cassetta E, Pasqualetti P, Paqualetti P. Interhemispheric differences of sensory hand areas after monohemispheric stroke: MEG/MRI integrative study. Neuroimage 2001;14:474–485.
192. Gallien P, Aghulon C, Durufle A, Petrilli S, De Crouy AC, Carsin M, Toulouse P. Magnetoencephalography in stroke: a 1-year follow-up study. Eur J Neurol 2003;10: 373–382.
193. Lewine JD, Davis JT, Sloan JH, Kodituwakku PW, Orrison Jr. WW. Neuromagnetic assessment of pathophysiologic brain activity induced by minor head trauma. Am J Neuroradiol 1999;20:857–866.
194. Bowyer SM, Aurora KS, Moran JE, Tepley N, Welch KM. Magnetoencephalographic fields from patients with spontaneous and induced migraine aura. Ann Neurol 2001;50:582–587.
195. Tran TD, Inui K, Hoshiyama M, Lam K, Qiu Y, Kakigi R. Cerebral activation by the signals ascending through unmyelinated C-fibers in humans: a magnetoencephalographic study. Neuroscience 2002;113:375–386.
196. Inui K, Tran TD, Qiu Y, Wang X, Hoshiyama M, Kakigi R. Pain-related magnetic fields evoked by intra-epidermal electrical stimulation in humans. Clin Neurophysiol 2002;113: 298–304.
197. Kamada K, Moller M, Saguer M, Ganslandt O, Kaltenhauser M, Kober H, Vieth J. A combined study of tumor-related brain lesions using MEG and proton MR spectroscopic imaging. J Neurol Sci 2001;186:13–21.
198. Reite M, Teale P, Rojas DC. Magnetoencephalography: applications in psychiatry. Biol Psychiat 1999;45:1553–1563.
199. Pekkonen E, Hirvonen J, Jaaskelainen IP, Kaakkola S, Huttunen J. Auditory sensory memory and the cholinergic system: implications for Alzheimer's disease. Neuroimage 2001;14:376–382.
200. Nishitani N, Avikainen S, Hari R. Abnormal imitation-related cortical activation sequences in Asperger's syndrome. Ann Neurol 2004;55:558–562.
201. Hren R, Zhang X, Stroink G. Comparison between electrocardiographic and magnetocardiographic inverse solutions using the boundary element method. Med Biol Eng Comput 1996;34:110–114.
202. Fenici R, Melillo G. Magnetocardiography: ventricular arrhythmias. Eur Heart J 1993;14(Suppl E): 53–60.
203. Stroink G, Moshage W, Achenbach S. Cardiomagnetism. In: Andrä W, Nowak H, editors. Magnetism in Medicine: A Handbook. Berlin: Wiley; 1998. p 136–189.
204. Tavarozzi I, Comani S, Del Gratta C, Di Luzio S, Romani GL, Gallina S, Zimarino M, Brisinda D, Fenici R, De Caterina R. Magnetocardiography: current status and perspectives. Part II: Clinical applications. Ital Heart J 2002;3:151–165.
205. Fenici R, Brisinda D, Nenonen J, Fenici P. Noninvasive study of ventricular preexcitation using multichannel magnetocardiography. Pacing Clin Electrophysiol 2003;26: 431–435.
206. Kanzaki H, Nakatani S, Kandori A, Tsukada K, Miyatake K. A new screening method to diagnose coronary artery disease using multichannel magnetocardiogram and simple exercise. Basic Res Cardiol 2003;98:124–132.
207. Leder U, Pohl HP, Michaelsen S, Fritschi T, Huck M, Eichhorn J, Muller S, Nowak H. Noninvasive biomagnetic imaging in coronary artery disease based on individual current density maps of the heart. Int J Cardiol 1998; 64:83–92.
208. Blum T, Saling E, Bauer R. First magnetoencephalographic recordings of the brain activity of a human fetus. Br J Obstet Gynaecol 1985;92:1224–1229.
209. Kariniemi V, Ahopelto J, Karp PJ, Katila TE. The fetal magnetocardiogram. J Perinat Med 1974;2:214–216.
210. Lowery CL, Campbell JQ, Wilson JD, Murphy P, Preissl H, Malak SF, Eswaran H. Noninvasive antepartum recording of fetal S-T segment with a newly developed 151-channel magnetic sensor system. Am J Obstet Gynecol 2003;188: 1491–1496; discussion 1496–1497.
211. Wakai RT, Leuthold AC, Martin CB. Atrial and ventricular fetal heart rate patterns in isolated congenital complete heart block detected by magnetocardiography. Am J Obstet Gynecol 1998;179:258–260.
212. Quartero HW, Stinstra JG, Golbach EG, Meijboom EJ, Peters MJ. Clinical implications of fetal magnetocardiography. Ultrasound Obstet Gynecol 2002;20:142–153.
213. Kahler C, Schleussner E, Grimm B, Schneider U, Haueisen J, Vogt L, Seewald HJ. Fetal magnetocardiography in the investigation of congenital heart defects. Early Hum Dev 2002;69:65–75.
214. Van Leeuwen P. Future topics in fetal magnetocardiography. In: Nenonen J, Ilmoniemi R, Katila T, editors. Biomag 2000: Proc 12th Int Conf Biomag. Espoo, Finland: Helsinki University of Technology; 2001. p 587–590.
215. Lengle JM, Chen M, Wakai RT. Improved neuromagnetic detection of fetal and neonatal auditory evoked responses. Clin Neurophysiol 2001;112:785–792.
216. Schneider U, Schleussner E, Haueisen J, Nowak H, Seewald HJ. Signal analysis of auditory evoked cortical fields in fetal magnetoencephalography. Brain Topogr 2001;14:69–80.
217. Schleussner E, Schneider U, Olbertz D, Kahler R, Huonker R, Michels W, Nowak H, Seewald HJ. Assessment of the fetal neuronal maturation using auditory evoked fields in fetal magnetoencephalography. In: Yoshimoto T, Kotani M, Kuriki S, Karibe H, Nakasato N, editors. Recent Advances in Biomagnetism. Sendai: Tohoku University Press; 1999. p 975–977.

218. Lowery C, Robinson S, Eswaran H, V. J, H. G, Cheung T. Detection of the transient and steady-state auditory evoked responses in the human fetus. In: Yoshimoto T, Kotani M, Kuriki S, Karibe H, Nakasato N, editors. Recent Advances in Biomagnetism. Sendai: Tohoku University Press; 1999. p 963–966.
219. Rose D, Eswaran H. Spontaneous neuronal activity in fetuses and newborns. Exp Neurol, in press.
220. Eswaran H, Wilson J, Preissl H, Robinson S, Vrba J, Murphy P, Rose D, Lowery C. Magnetoencephalographic recordings of visual evoked brain activity in the human fetus. Lancet 2002;360:779–780.
221. Vrba J, Robinson SE, McCubbin J, Murphy P, Eswaran H, Wilson JD, Preissl H, Lowery CL. Human fetal brain imaging by magnetoencephalography: verification of fetal brain signals by comparison with fetal brain models. Neuroimage 2004;21:1009–1020.
222. Paulson DN, Fagaly RL, Toussaint RM, Fischer F. Biomagnetic susceptometry with SQUID instrumentation. IEEE Trans Mag 1991;27:3249–3252.
223. Brittenham GM, Sheth S, Allen CJ, Farrell DE. Noninvasive methods for quantitative assessment of transfusional iron overload in sickle cell disease. Semin Hematol 2001; 38: 37–56.
224. Hoshiyama M, Kakigi R, Nagata O. Peripheral nerve conduction recorded by a micro gradiometer system (micro-SQUID) in humans. Neurosci Lett 1999;272:199–202.
225. Mackert BM, Curio G, Burghoff M, Trahms L, Marx P. Magnetoneurographic 3D localization of conduction blocks in patients with unilateral S1 root compression. Electroencephalogr Clin Neurophysiol 1998;109:315–320.
226. Le Gros V, Lemaigre D, Suon C, Pozzi JP, Liot F. Magnetopneumography: a general review. Eur Respir J 1989;2: 149–159.
227. Huvinen M, Oksanen L, Kalliomaki K, Kalliomaki PL, Moilanen M. Estimation of individual dust exposure by magnetopneumography in stainless steel production. Sci Total Environ 1997;199:133–139.
228. Moller W, Barth W, Kohlhaufl M, Haussinger K, Stahlhofen W, Heyder J. Human alveolar long-term clearance of ferromagnetic iron oxide microparticles in healthy and diseased subjects. Exp Lung Res 2001;27:547–568.
229. Moraes R, Toncon LE, Baffa O, Oba-Kunyioshi AS, Wakai RT, Leuthold AC. Adaptive, autoregressive spectral estimation for analysis of electrical signals of gastric origin. Physiol Meas 2003;24:91–106.
230. Kosch O, Osmanoglou E, Hartman V, Strenzke A, Weitschies W, Wiedenmann B, Monnikes H, Trahms L. Investigation of gastrointestinal transport by magnetic marker localization. Biomed Tech (Berl) 2002;47(Suppl 1, Pt 2): 506–509.

See also ELECTROCARDIOGRAPHY, COMPUTERS IN; ELECTROENCEPHALOGRAPHY; EVOKED POTENTIALS; PULMONARY PHYSIOLOGY.

BIOMATERIALS, ABSORBABLE

MARK BORDEN
Director of Biomaterials Research
Irvine, California

INTRODUCTION

Historically, the use of implants in orthopedic surgery has originated from fracture repair and joint replacement applications. During the late 1920s, stainless-steel bone implants such as Kirshner nails and Steinman pins were popularized for the surgical treatment of fractures (1). With the introduction of new surgical materials such as cobalt alloys, polyethylene and poly(tetrafluoroethylene) [Teflon], surgeons and engineers began working toward the design and fabrication of artificial joints. The advent of new high strength implant materials allowed researchers such Dr. John Charnley to begin pioneering work in total hip replacement surgery in the late 1930s (1,2). As advances in chemistry, metallurgy, and ceramics progressed throughout the years, a large variety of implants have entered the orthopedic market. Today, orthopedic implants are composed of specialized metals, ceramics, polymers, and composites that possess a large range in properties. Although these materials have been successfully fabricated into a variety of implants, one common issue has remained. Once the device has performed its required function and is no longer needed, it remains as a bystander in the now healthy tissue. The issue is that the long-term presence of an implant in the body can result in implant-related complications such as loosening, migration, mechanical breakdown and fatigue, generation of wear particles, and other negative effects (3–6). With prolonged patient life spans and higher activity levels, more and more people are now outliving the lifetime of their implants.

The potential for long-term implant problems has driven researchers to look to a unique category of materials that are capable of being completely resorbed by the body. These bioresorbable or biodegradable materials are characterized by the ability to be chemically broken down into harmless byproducts that are metabolized or excreted by the body. Materials of this type offer a great advantage over conventional nonresorbable implant materials. Bioresorbable implants provide their required function until the tissue is healed, and once their role is complete, the implant is completely resorbed by the body. The end result is healthy tissue with no signs that an implant was ever present. As the implant is completely gone from the site, long-term complications associated with nonresorbable devices do not exist.

ORTHOPEDIC APPLICATIONS OF RESORBABLE IMPLANTS

The ability of a resorbable implant to provide temporary fixation followed by complete resorption is a desirable property for a large variety of surgical applications. In relation to orthopedic surgery, this behavior is particularly useful based on the goal of restoring physiological function to the tissues and joints of the skeleton. In general, orthopedic surgery is often compared with carpentry in that the surgeon's instruments often consist of hammers, drills, and saws. Similar to carpentry, specialized screws, plates, pins, and nails are used to fix one material to another. In orthopedics, this fixation can be categorized into two main areas: bone-to-bone fixation and soft tissue-to-bone fixation. Bone fixation is used in the treatment of complex fractures and in reconstructive procedures of the skeleton. The implants used in these surgeries are designed to

maintain the position of the bone fragments, to stabilize the site, and to allow for eventual fusion of the fracture. As a result of the fracture healing process, the bone is remodeled so effectively that it is often difficult to locate the initial injury. With nonresorbable implants, the long-term presence of the device only serves as a source for potential complications. Resorbable implants, on the other hand, alleviate this concern by fully resorbing and allowing the bone to completely remodel into its normal physiological state.

In addition to bone fixation, soft tissue fixation is also an excellent application of resorbable implants. This type of reconstruction is often the result of trauma to joints such as the knee and shoulder. Typically developing from sports injuries or accidents, the goal is to restore stability to the joint by replacing or reconstructing the ligament or tendon interface to bone. In the knee, for example, the reconstruction of a torn anterior cruciate ligament (ACL) is a common sports medicine procedure. This type of surgical reconstruction consists of replacing the torn ACL with a bone-tendon-bone graft taken from the patient's patella and fixing the graft across the joint. During the procedure, the bony portion of the ACL graft is fixed in bone tunnels drilled into the tibia and femur. In order to stabilize the graft and aid in the formation of a stable bone-to-ligament interface, interference screws are used to fix the graft to the site. Once bone has been incorporated into the graft, the device is no longer needed.

Another example of soft tissue reconstruction is the repair of a tear in the rotator cuff tendon of the shoulder. This type of injury requires reestablishing the tendon-to-bone interface. To facilitate this process and restore stability to the shoulder, implants called suture anchors are used to provide a means to affix the torn tendon to the bone of the humerous. Just as the name describes, these implants function by providing an anchor in bone that allows the attached suture to tighten down on the tendon and pull it in contact with bone. As healing progresses, a stable interface develops and joint function is restored. Similar to other fixation applications, once the interface has fully healed, the implant is no longer needed.

FUNCTION OF A RESORBABLE IMPLANT

As seen from the various types of tissue fixation procedures within orthopedic surgery, resorbable implants are exposed to a variety of healing environments. Out of the currently used materials in orthopedic surgery, only the polymer and ceramic groups contain resorbable biomaterials. It is the specific properties of these materials that allow them to be used as resorbable devices. In evaluating a material for potential use as an implant, the key properties include implant biocompatibility, resorbability, and mechanical properties. The first criteria, biocompatibility, refers to the ability of the material to be implanted into the body without negatively affecting the surrounding tissue, which includes the absence of inflammation, toxicity (materials that kill surrounding cells), carcinogenicity (materials that can cause cancer), genotoxicicty (materials that damage the DNA of surrounding cells), and mutagenicity (materials that cause genetic mutations within the cell). More specifically related to bone, the implant must also be osteocompatible, which means the material does not interfere with the normal bone healing process (7).

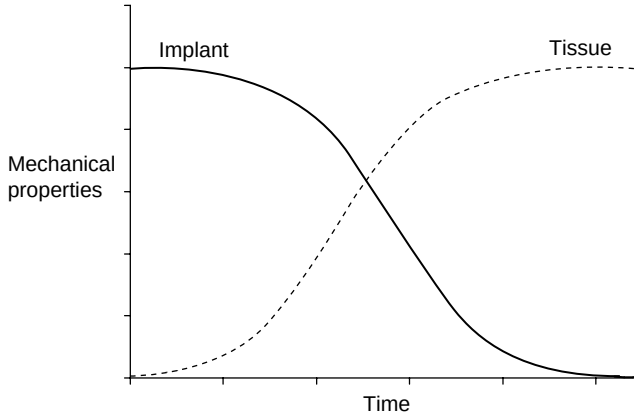


Figure 1. Optimal stress transfer profile for resorbable implants demonstrating the load-sharing properties of the implant.

Although biocompatibility has a direct effect on how the tissue surrounding the device heals and is an important property of the implant, the main criteria related to implant function are the resorbability and mechanical properties. Once the device is implanted, it provides immediate mechanical stabilization to the site while the tissue heals. As the regenerating bone, ligaments, or tendons become stronger over time, the implant site becomes less dependent on the device and more dependent on the healing tissue. This concept is shown in Fig. 1. In this situation, the implant provides all of the mechanical support immediately following placement. As the device begins to degrade, the mechanical properties decrease over time and are gradually transferred to the new tissue. During this period, the regenerating tissue responds to the gradual loads and begins to remodel and become stronger. In the healing of musculoskeletal tissue, the sharing of the load between the implant and the tissue results in further regeneration. Once healing is complete, the load is fully transitioned to the tissue, which is now mechanically independent from the implant. Upon final resorption of the device, the site is left fully functional and entirely free of any implant material.

The ability to gradually transfer load to regenerating tissue is an important part of the musculoskeletal healing process. This characteristic is only found in resorbable materials. Although metallic implants offer effective load-bearing properties in applications such as joint replacement and certain spinal surgeries, these high strength materials do not resorb and do not effectively transfer loads to the implant site. Due to the high strength of metals, these implants bear the majority of the force at the site and can shield the surrounding tissues from any load. This phenomenon is called stress shielding and can actually cause bone to resorb in certain areas around the implant (8,9). The stress-shielding effect is based on a concept called Wolff's Law, which describes the ability of bone to

dynamically respond to the presence or lack of stress by changing its density and strength.

When bone is subjected to new loads, the additional stress stimulates bone formation and the tissue increases in strength and density. When the remodeling process is complete, the stronger tissue can now fully support the added load. However, when a high strength material such as metal is placed in bone, the bone surrounding the implant is shielded from the normal stresses, which results in a decrease in the strength and density of this tissue and possible bone resorption. This phenomenon can cause complications such as implant loosening or fracture of the implant site. Polymer and ceramic materials, on the other hand, have mechanical properties that are similar to bone, which allows them to share the stresses with newly regenerating tissue thereby preventing resorption and other stress-shielding complications (10–12).

Although load transfer and strength retention are common properties of all resorbable implants, not all surgical sites heal at the same rate. In fracture fixation applications where bone-to-bone contact is maintained, healing can be as short as 6–8 weeks. However, in applications such as spinal fusion where significant amounts of tissue need to be formed in the intervertebral space, the healing process can take up to 6–12 months. Based on the dependence of implant function on the surgical site, the material choice becomes an important part of implant development. The challenge in designing an implant lies in choosing a material that correctly matches the function and strength requirements of the surgical application, which can be accomplished through a thorough understanding of the function of the implant, the load requirements of the implant site, and the properties of the material.

RESORBABLE POLYMERS

One of the most versatile materials used in orthopedic surgery are polymers. Polymers are a group of materials that are produced through a chemical reaction that results in a long chain of repeating molecules called monomers. In addition to polymers composed of a single monomer repeating unit, there are other materials, called copolymers, that have two or more monomer repeating units. By combining different monomers, the properties of the resultant copolymer can be specifically modified to serve a certain purpose. This versatility can also be achieved by modifying the polymerization reaction and the postprocessing techniques used to create polymer implants. Table 1 shows a few examples of the many properties that characterize polymers. These characteristics can be altered by changing the molecular weight, chemical structure, and morphology of the polymer or copolymer.

Table 1. Range of Common Properties Found in Orthopedic Polymers

Property	Range	
Resorbability	Fully Resorbable	Nonresorbable
Strength	Low Strength	High Strength
Moldability	Flexible	Rigid
Physical State	Gel/Liquid	Solid
Temperature Sensitivity	Flexible at higher Temperature	Rigid at all Temperatures
Radiation Resistance	Low	High

The molecular weight of a polymer is a measurement of the number of repeating units found in the entire molecule. During the formation of polymers and copolymers, the length of the molecule can be controlled to give a variety of molecular weights. The length of the polymer chain can be as small as a few thousand repeating units or as large as a million, which can have a significant effect on the degradation properties of the polymer. When a polymer breaks down, it occurs through random cleavage of the chemical bonds along the polymer chain. It is not until the polymer finally fragments into its monomer form that the material is absorbed by the surrounding tissue. Therefore, longer polymers chains with higher molecular weights will take a longer time to degrade because more bonds exist to the cleaved.

Additionally, the chemical structure can also affect degradation. As described previously, the backbone of a polymer consists of a long, continuous chain of monomer units linked together. In all resorbable polymers, it is the backbone of the polymer where degradation occurs. The typical linkage that allows polymers to break down is a carbon–oxygen–carbon (C-O-C) bond. This bond is found in ester, carbonate, carboxylic acid, and amide-based polymers. The degradation process occurs at this bond when the material is exposed to water. In a process called hydrolytic degradation, water molecules chemically react with the C-O-C bonds causing them to break apart at random areas throughout the polymer chain. The chemical structure of the polymer dictates the ability of the water molecules to access these bonds and start the degradation reaction. If the polymer is characterized by large bulky side chains or strong C-O-C bonds, it becomes difficult for the water molecule to penetrate the polymer chains to react with the backbone, which results in a prolonged degradation period. The opposite is true for polymers that tend to absorb water and do not have any large side chains. In these polymers, the water molecules can easily access the backbone and the degradation process proceeds at a relatively fast rate.

The final characteristic that can affect the degradation and strength of a polymer is the morphology. The morphology of the polymer refers to the orientation of the long polymer chains throughout the material. Polymer morphology can be classified into three groups: crystalline polymers, semicrystalline polymers, and amorphous polymers. The crystallinity of a polymer develops from areas within the material where the polymer chains are aligned and tightly packed together. This type of orientation forms dense crystalline regions within the random arrangement of the polymer chains. A highly organized polymer is considered crystalline, whereas a completely random orientation is considered amorphous. Semicrystalline polymers fall between these two extremes and exist with varying degrees of crystallinity (Fig. 2).

The effect of crystallinity on the degradation of the polymer is due to the tight orientation between the polymer chains in the crystalline regions. With highly crystalline

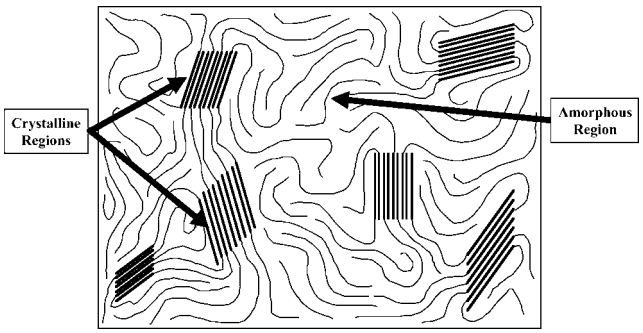


Figure 2. Semicrystalline polymer showing orientation of amorphous regions and crystalline regions.

polymers, the degradation rate is very slow due to the difficulty of water gaining access to the C-O-C bonds. These polymers degrade at a rate much slower than polymers that are completely amorphous with no crystalline regions (13). The crystallinity also affects the mechanical properties of the polymer. The dense, organized areas within crystalline polymer make these regions stronger than the unorganized, amorphous regions. As a result, an increase in crystallinity translates into an increase in mechanical properties.

The ability to alter the properties of a polymer has resulted in thousands of different materials used in a wide range of applications. However, only a few of these polymers can be effectively used as medical implants due to the strict requirements of surgical implants. The following sections describe some of the polymers currently used in orthopedic surgery.

Poly(hydroxy acids)

Poly(hydroxy acids) were the first group of resorbable materials to be used in surgery (14). The main polymers in this family are poly(lactic acid) (PLA), poly(glycolic acid) (PGA), and the copolymer poly(lactide-*co*-glycolide) (PLG). The basic chemical structure of these materials in shown in Fig. 3. Originally, PLA and PGA were initially used as a degradable sutures (15–18). However, since their initial success in the wound closure field, both of these polymers have been fabricated into several orthopedic implants including screws (19,20), plates (19,21), pins (22–25), suture anchors (26), and bone grafting scaffolds (27–30). In addition, several new devices composed of the PLG copolymer have been developed over the past 10 years (31–35).

Although the chemical structure of PLA and PGA is somewhat similar, the presence of a methyl group ($-CH_3$) in PLA significantly changes its physical properties compared with PGA. Comparatively, PGA has a lower strength and degrades in approximately 3–6 months, whereas certain forms of PLA can take 3–5 years to fully degrade. Although only a single methyl group differentiating PLA from PGA exists, the location of this side group close to the C-O-C bond makes it difficult for the water molecules to gain access to cleavage site, thereby prolonging degradation.

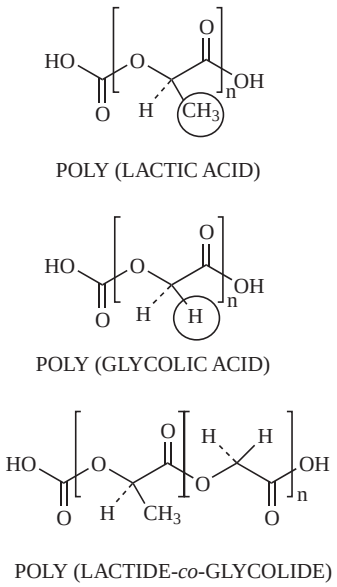


Figure 3. Chemical structure of poly(lactic acid), poly(glycolic acid), and the copolymer poly(lactide-*co*-glycolide).

In addition, the methyl group in PLA also gives the polymer a unique chemical orientation. As a monomer, lactic acid is a molecule that can have two different molecular orientations: L-lactic acid and D-lactic acid. These isomers are based on the orientation of the methyl and hydrogen groups on the molecule. Figure 4 shows the chiral nature of the lactic acid molecule and the resulting stereoregular polymers: poly(L-lactic acid) (PLLA), poly (D-lactic acid (PDLA), and poly (D,L-lactic acid) (PDLLA). Although three forms of PLA exist, in the medical field, poly(L-lactic acid) is used more often than poly(D-lactic acid) because the degradation product is the same as naturally occurring L-lactic acid (13).

Using the various forms of PLA, polymers with significantly different properties can be synthesized. The effect of the starting isomer on the physical properties of the material is dramatically seen in the properties of PLLA and PDLLA. In Fig. 4, the chemical structure of poly (L-lactic acid) is represented by a long chain with all of the $-CH_3$ groups on one side. This uniformity allows the chains to pack tightly together resulting in a highly crystalline material that has a high strength and long degradation period (3–5 years). Poly(D,L lactic acid), on the other hand, is characterized by either a random or alternating arrangement of the $-CH_3$ groups and –H groups. This molecule orientation prevents the polymer chains from packing together, resulting in a completely amorphous polymer with a lower strength and shorter degradation profile (9–12 months). In addition, the polymerization of L-lactic acid and D,L-lactic acid together results in a copolymer with properties in between PLLA and PDLLA. In recent years, the 70:30 combination of poly(L/D,L lactic acid) has gained popularity in orthopedic applications due to its ability to retain its strength for 9–12 months while being completely resorbed within 1.5–2 years (36–38). This copolymer appears to provide the best

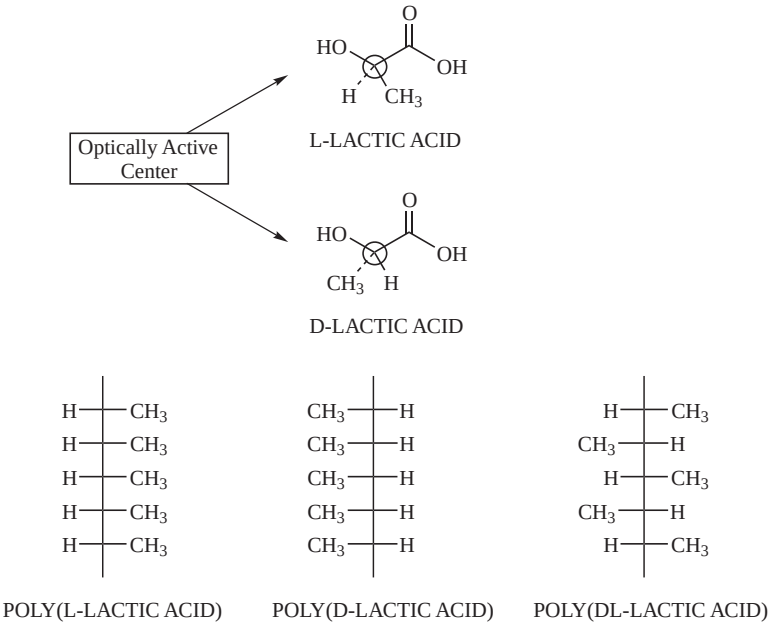


Figure 4. The optically active center in lactic acid is allows it to have two different molecular orientations. These orientations result in three types of stereoregular polymers.

of both worlds in that it has the strength retention of PLLA but has a degradation period only slightly longer than PDLLA.

In addition to the lactic acid-based copolymers, a combination of PLA and PGA has also been shown to be an effective implant material (31–35). Due to the large differences in the degradation properties of PLA and PGA, the poly(lactide-*co*-glycolide) (PLG) copolymer can be modified based on the PLA to PGA ratio to provide varying degradation periods. Common PLG copolymers used in orthopedic surgery have PLA/PGA ratios of 50:50, 75:25, and 85:15. This combination not only provides both slow and fast resorbing monomer units but also eliminates any crystallinity, making the copolymer completely amorphous. These materials have been commonly used as fracture implants due to the shorter 6–12 month degradation period.

Although PLA and PGA polymers have been successfully used in patients for several years, there have been certain cases where the abundance of acidic monomers at the site has caused inflammation and bone resorption (39–48). When PLA and PGA polymers near the end of degradation, they release lactic acid and glycolic acid, respectively. Although these degradation products can be metabolized by the body, if the surrounding tissue cannot absorb the acid in a timely manner, the build up of acid and resultant drop in pH at the implant site can cause bone to resorb. Historically, this effect has mainly been seen in the fast-resorbing PGA implants; however, a few cases have been reported with PLA (43,46,49,50). Although the bone resorption complication is detrimental to the healing of the implant site, the complication rate has been relatively low. In a review of over 2000 patients by Bostman, only 5% of the patients have shown implant-associated reactions (44).

Additionally, the copolymers PLG and PLDLLA have been shown to possess a more osteocompatible degradation profile due to a gradual release of the acidic byproducts (36,51–56), which has minimized acid dumping and the associated bone resorption complications. In a study by Eppley et al. (35), 1883 patients treated with PLG plates and screws for bone fixation in craniofacial procedures showed an implant-related complication rate of only 0.5%, which was well below the 5% rate reported by Bostman for PGA and PLA implants. Overall, the PLG and PLDLLA copolymers have been shown to be effective devices for fracture fixation, bone graft containment, and soft tissue fixation, and have begun to replace the outdated PLA and PGA devices (37,38,57,58).

Polycarbonates

Another group of resorbable polymers are the polycarbonates. Although the majority of the polymers and copolymers within the polycarbonate family are nonresorbable plastics used for industrial applications, a select few exist that are resorbable and can be used as orthopedic implants. One group of medical-grade polycarbonates are the copolymers based off of poly(trimethylene-carbonate) (PTMC) and poly(glycolic acid) or poly(lactic acid). These combinations offer the combined advantage of the processing versatility of PTMC and the resorbability and strength of PLA and PGA. The PTMC copolymers have been used for soft tissue fixation in shoulder surgery as suture anchors and soft tissue tacks (59–61).

Although the PTMC copolymers with PGA and PLA offer improved implant properties compared with PTMC alone, the degradation of the material still produces acidic monomers. In order to avoid the issues with glycolic acid- and lactic acid-based polymers and copolymers, an amino acid-based polycarbonate was developed by Joachim Kohn at Rutgers University. Designed specifically for orthopedic applications, the amino acid poly(carbonates) combine the biocompatibility of individual amino acids with the strength and processability of standard industrial poly(carbonates) (62–64). One such promising polymer, poly(DTE carbonate), is derived from the amino acid

tyrosine and has been shown to have excellent strength-retention properties, an optimal degradation profile, and biocompatible degradation products (65–68). Based on large amount of characterization data, a material safety file has been recently established at the U.S. Food and Drug Administration (FDA) that allows manufacturers to begin development of poly(DTE carbonate) implants. Due to the advantages of poly(DTE carbonate) over conventional resorbable polymers, amino acid-based poly(carbonate) implants may soon be a common sight in orthopedic operating rooms.

Other Resorbable Polymers

In addition to the widely used PLA and PGA polymers and the up-and-coming amino acid-based poly(carbonates), several other polymers have applications as medical devices. Although not specifically used in orthopedics, the poly(anhydride) family of polymers developed by Robert Langer at MIT has been effectively used as drug-delivery vehicles (69–73). The function of these resorbable implants is to provide a sustained and controlled release of drugs to a specific implant site. The device functions by releasing molecules entrapped within the implant as it degrades. Another polymer, poly(dioxanone), has been used as a resorbable suture material for several years (74–80). The flexibility of this polymer enables it to be used as a monofilament suture instead of the typical braided fiber of PGA, which provides the suture with an improved ability to move through tissue with less friction, thereby minimizing the tearing and pulling of the surrounding areas (81,82). Looking specifically at orthopedic applications, additional polymers currently in development include poly(caprolactone) (83–86), poly(hydroxybutyrate) (87–89), polyurethanes (90–93), and poly(phosphazenes) (94–96).

RESORBABLE CALCIUM CERAMICS

Aside from the polymers, the other group of resorbable implant materials are the calcium-based ceramics. Due to the similarity of these materials with the mineral content of bone, hydroxyapatite [$Ca_{10}(PO_4)_6(OH)_2$], calcium ceramics are highly biocompatible and osteocompatible materials that have a long history of clinical use. These materials are typically used in orthopedic surgery to fill voids in bone as self-setting cements or as porous blocks and granules.

Calcium Sulfate

One the first materials to ever be used as a filler for bone defects was calcium sulfate (Plaster of Paris) (97). In its dehydrated form (calcium sulfate hemihydrate), this material undergoes a chemical reaction when mixed with water that allows it to function as a resorbable cement. As the cement reacts, it transforms from a slurry, to a paste, to a dough, and then fully sets into its final hardened form (calcium sulfate dihydrate). This reaction is exothermic in that it produces heat; however, the increase in temperature is only slightly above body temp (37 °C). Figure 5 shows a typical timeline of the calcium sulfate setting reaction. In the slurry and paste form, the calcium sulfate is able to be added to a syringe and injected to the bone graft site. Near the end of the reaction, the cement becomes much thicker and has a putty-like consistency. During this phase, the doughy cement can be molded into a variety of shapes and provides a custom fit when placed directly at the implant site. Once the cement has fully hardened, it can be shaped by using powered surgical instruments such as osteotomes, burrs, and drills.

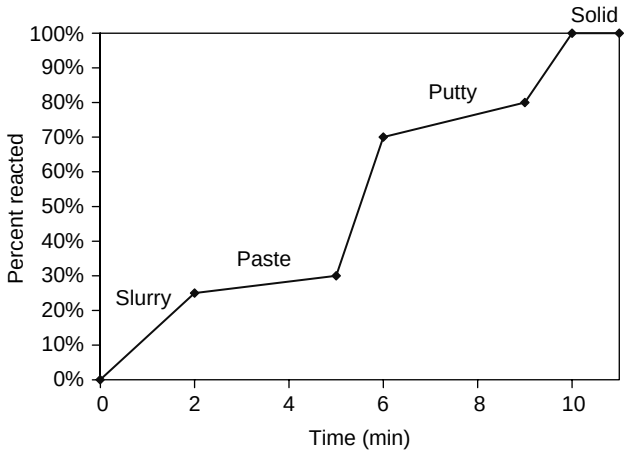


Figure 5. Typical setting reaction and phase changes for a calcium sulfate cement.

The resorption of calcium sulfate graft materials is based on the microstructure of the fully hardened cement. Figure 6 shows electron micrographs of the surface of fully reacted calcium sulfate dihydrate. These high magnification images show small calcium sulfate crystals packed together in a microporous structure. Upon implantation, the presence of these small pores allows the calcium sulfate to absorb water throughout the cement. Unlike polymers, which undergo active breakdown of the polymer chains, calcium sulfate materials are slowly dissolved by the water. As the material dissolves, Ca^{2+} and SO_4^{-3} ions are released over a 6–8 week period. During healing, bone formation initially begins on the outer area of the calcium sulfate and progresses inward as the cement slowly breaks apart. During the resorption process, the dissolution of the calcium sulfate material aids bone formation by providing a direct source Ca^{2+} ions to the surrounding osteoblasts. These cells absorb the calcium and use it during the mineralization phase of bone regeneration. From a mechanical standpoint, the hardened cement can provide initial stabilization to the site, but quickly loses it strength as the calcium sulfate begins to fragment. Although the strength of the calcium sulfate quickly decreases within the first few weeks, additional bone regeneration takes place within the cement and the implant site becomes mechanically stable. At the 6–8 week period, the majority of the calcium sulfate is resorbed by the body and has been replaced by bone.

In general, calcium sulfate cements and implants offer an effective means to fill small voids in bone resulting from cysts, tumors, or fractures (98–101). The initial strength

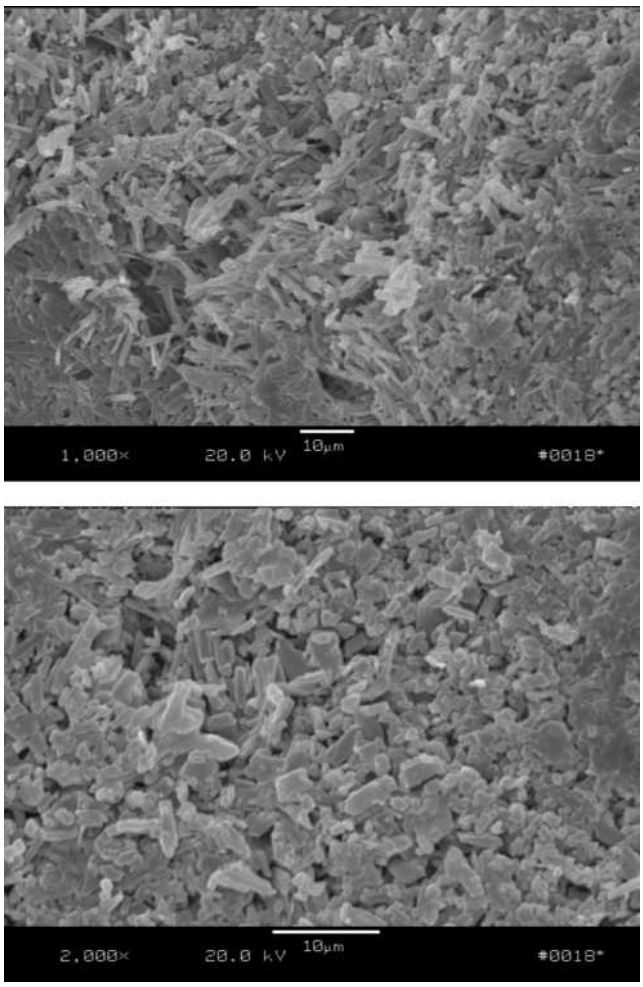


Figure 6. High magnification scanning electron micrographs of fully reacted calcium sulfate dihydrate showing crystalline structure and microporosity (1000X and 2000X magnification).

can also help maintain the spacing of fracture fragments and aid in placement of additional hardware. The moldability of the cement allows a custom fit to the defect site and makes the material easy to use. However, due to the quick resorption time and quick loss in strength, this material can not be effectively used in large defects or in areas under high mechanical loads. In these applications, supplemental hardware and grafting materials are needed to ensure complete bone regeneration (102,103). From a commercial standpoint, calcium sulfate graft materials are available in a cement form (requires mixing at the time of surgery) or in a preformed pellet form (fully reacted calcium sulfate dihydrate).

Calcium Phosphate. Calcium phosphates are another class of calcium containing bone graft materials that offer different properties than the calcium sulfates. As the name describes, these material are composed of varying amounts of calcium (Ca^{2+}) and phosphate (PO_4^{-3}). One of the first calcium phosphate materials to be used as a bone graft was hydroxyapatite, which was chosen because it is the main inorganic component of bone accounting for 40% of its weight. Most calcium phosphate graft materials are produced synthetically and can be chemically altered to create materials with different properties. By slightly varying the calcium-to-phosphate ratio, the resorption times and mechanical properties of these materials can be significantly altered. Hydroxyapatite [$Ca_{10}(PO_4)_6(OH)_2$] with a Ca/P ratio of 1.67 has slow resorption rate, which, depending on crystallinity, can be as little are 2–5% resorption per year. Tricalcium phosphate $Ca_3(PO_4)_2$ has a ratio of 1.5, which results in a much faster resorption time of 9–12 months.

Due to the chemical composition of calcium phosphates, the mechanism of resorption is different than the dissolution mechanism seen with calcium sulfates. The chemical similarity of calcium phosphates to bone results in a cell-mediated resorption profile. During healing, bone-resorbing cells called osteoclasts migrate to the surface of the calcium phosphate ceramics. Once activated, the osteoclasts release specific enzymes that dissolve the calcium phosphate into its base ions. As the osteoclasts tunnel through the calcium phosphate, bone-forming cells called osteoblasts trail behind filling in the region with new tissue. Similar to calcium sulfate, the calcium ions resulting from the resorption process are transported to the osteoblasts, which create new mineralized bone. Over time, the entire structure is slowly dissolved by the osteoclasts and replaced with new bone.

To facilitate this type of resorption process, many of the calcium phosphate bone graft materials exist as porous scaffolds (104–109). A typical example of an osteoconductive calcium phosphate bone graft scaffold is shown in Fig. 7. This material, called Pro Osteon (developed and manufactured by Interpore Cross), was one of the first porous calcium phosphates used in orthopedics (110–113). Derived from sea coral, it is fabricated by chemically converting the calcium carbonate skeleton of the coral into hydroxyapatite. This reaction can be run to completion to give a implant composed entirely of hydroxyapatite or intentionally stopped to result in an implant with a thin (4–10 μm) surface of hydroxyapatite over the calcium carbonate skeleton. The conversion of coral to Pro Osteon

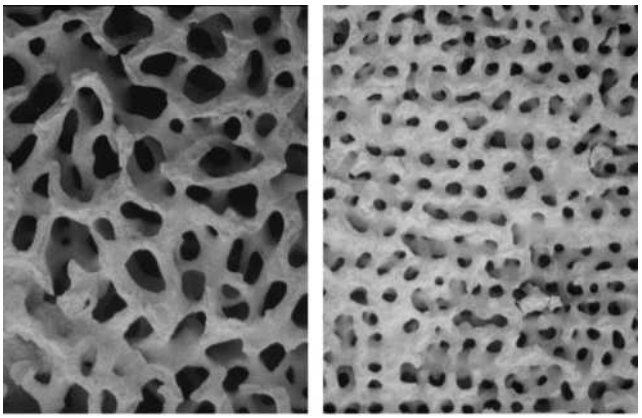


Figure 7. Photographs of a two commercially available calcium phosphate scaffolds derived from coral (Interpore Cross, Irvine, CA).

allows the relatively short degradation time of calcium carbonate (6–8 weeks) to be prolonged to 8–18 months for Pro Osteon R (HA layer on the calcium carbonate skeleton) and to 3–5 years for Pro Osteon HA (fully converted hydroxyapatite). With a natural pore structure similar to cancellous bone, the Pro Osteon graft materials offer an effective scaffold for new bone growth. Since development of the Pro Osteon bone graft materials, several other porous calcium phosphates have entered the market. These materials are synthetically made to mimic the porosity of cancellous bone, which is done through various foaming and void creation techniques.

In contrast to calcium sulfate graft materials, the slower resorption profile of porous calcium phosphate ceramics allow these material to be used in larger defects. In this scenario, the graft serves as a cellular "bridge" for continued bone growth. In bone grafting surgery, once a defect reaches a size when it can no longer completely heal itself, it is called a critical-sized defect. Typical bone regeneration can bridge empty gaps of up to 4 mm, but anything larger will not fill in with bone. A porous ceramic scaffold alleviates this problem by providing the means for bone to grow across the entire defect.

This effect was demonstrated in a study by Holmes who implanted a block of Pro Osteon 500R (calcium carbonate scaffold with an HA coating) into a rabbit tibial defect (114). The healing sequence of the this scaffold is shown in Fig. 8. As seen from cross sectional image of the implant before implantation (Fig. 8a), the structure is characterized by an open pore structure (black regions) within areas of calcium carbonate/HA ceramic (light-gray regions). After initial placement of the porous ceramic, cells migrated to the graft site and began to infiltrate the pore system. At the same time, proteins were released from surrounding bone and blood cells to stimulate the bone regeneration process, which was seen in the 6 week histology of the Pro Osteon 500R implant (Fig. 8b). In this image, bone formation was evident within the porosity of the scaffold, and osteoclasts were seen resorbing the scaffold (arrows). By 12 weeks, further bone growth was seen within the porosity, and significant portions of the scaffold were replaced by bone. At the 24 week time point, the scaffold was fully replaced by bone with the exception of the thin HA layer that once covered the calcium carbonate. As seen from this study, porous ceramics are capable of functioning as a scaffold for bone growth. The pore system allowed for immediate bone regeneration and the resorbability allowed the implant to be completely replaced by bone.

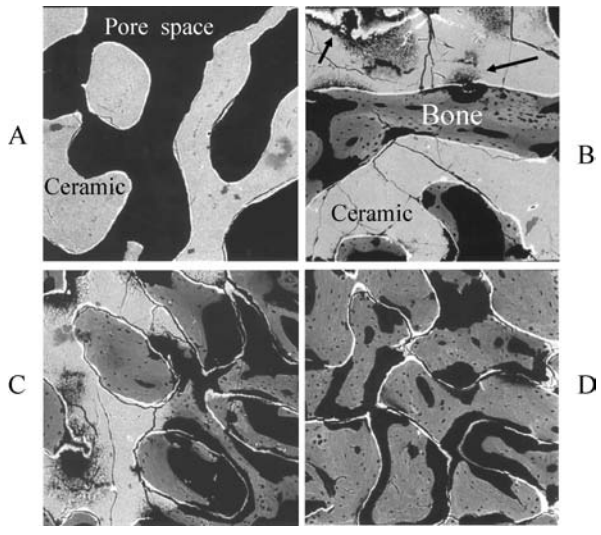


Figure 8. Typical healing mechanism of a porous ceramic implanted into a rabbit tibial defect. (Figure A – 0 weeks; Figure B – 6 weeks; Figure C – 12 weeks; Figure D – 24 weeks).

In addition to porous blocks and granules, calcium phosphates are also used in cement form (115–119). In this application, the base components that create calcium phosphates are provided in an unreacted form. With the addition of water, dilute acid, or other initiators, a chemical reaction takes place, and the components are converted to calcium phosphate. The result is a moldable paste or putty that can be shaped to the graft site and hardens into a solid mass. Although these cements have longer resorption times than calcium sulfate cements and can be used in broader applications, the resulting hardened cement does not possess the porosity to function as a scaffold for bone repair, which has limited the use of calcium phosphate cerments because surgeons prefer the porous blocks and granules over the self-setting cements.

RESORBABLE COMPOSITES

As discussed, both polymers and ceramics have properties suitable for fabricating orthopedic implants. However, certain drawbacks exist with these materials that can not be avoided no matter how the material is fabricated or chemically altered. One technique for combining the desirable properties of two or more materials is the fabrication of a composite. Composites used in the medical device area are fabricated by physically mixing two or more resorbable materials. One of the most common composite combinations is the creation of a polymer-ceramic composite. On their own, ceramics are excellent substrates for new bone growth due to the chemical similarity with bone mineral. However, their brittleness limits their use in load-bearing applications. Polymers, on the other hand, are elastomeric materials that can flex under deformation without major structural collapse. The combination of these two materials results in a high strength, yet ductile composite that allows for direct bone attachment on its surface. In this combination, the polymer adds to the overall mechanical properties of the composite, whereas the ceramic allows for bone formation directly on the ceramic phase.

The fabrication of a composite is a relatively straightforward process. Typically, ceramic particles in the shape of spheres or fibers are added to the polymer during processing. The various orientations of the particles within a polymer are shown in Fig. 9. As seen from these illustrations, each particle is surrounded by the polymer and serves to reinforce the polymer phase and improve it mechanical properties. Once fabricated in a block or rod

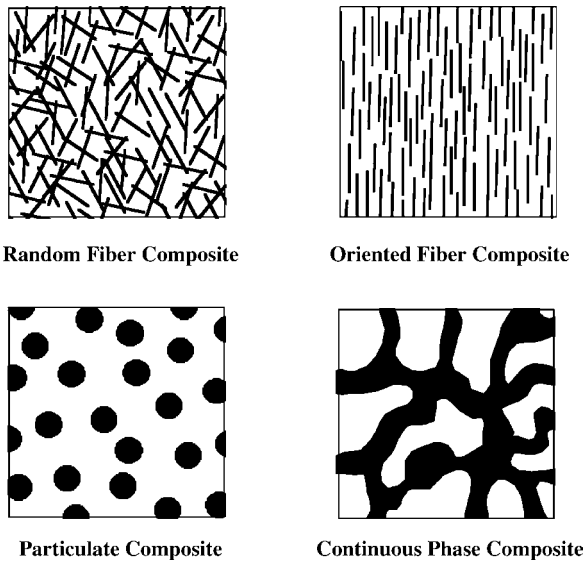


Figure 9. Orientation of various types of polymer-ceramic composites (ceramic is depicted as the black particles).

form, composites of these different types can be machined into a variety of implants such as fracture screws, pins, and plates. During the machining process, the ceramic on the outer surfaces of the implant are exposed. From a bone implant standpoint, the presence of the exposed calcium ceramic particles on the surface of the polymer aids in creating a solid bone-to-implant interface. In comparison, pure polymer implants typically heal with limited bone contact or a continuous layer of fibrous tissue usually covering the surface. Although the implant can still provide stabilization, it is not directly bonded to the surrounding bone. A composite implant improves on the stabilizing effect of the device through this bone-bonding ability.

In addition to the particulate ceramic composites, a new type of composite has recently been developed by Interpore Cross (Irvine, CA). This novel material consists of two intact, continuous phases of polymer and ceramic. Shown in Fig. 10, a continuous phase composite (CPC) is the result of infiltrating a porous ceramic block with polymer. The

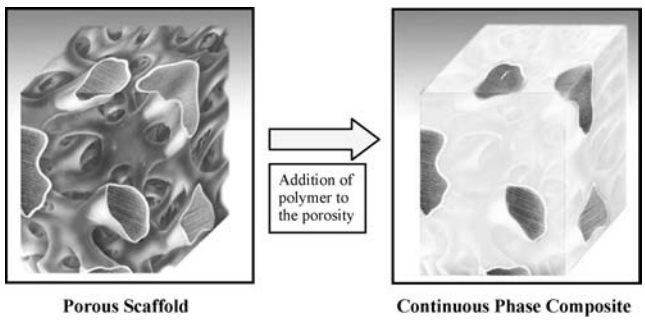


Figure 10. A continuous phase composite is formed when a polymer is infiltrated into the porosity a porous, ceramic scaffold. The result is a solid block with an intact polymer and ceramic phase.

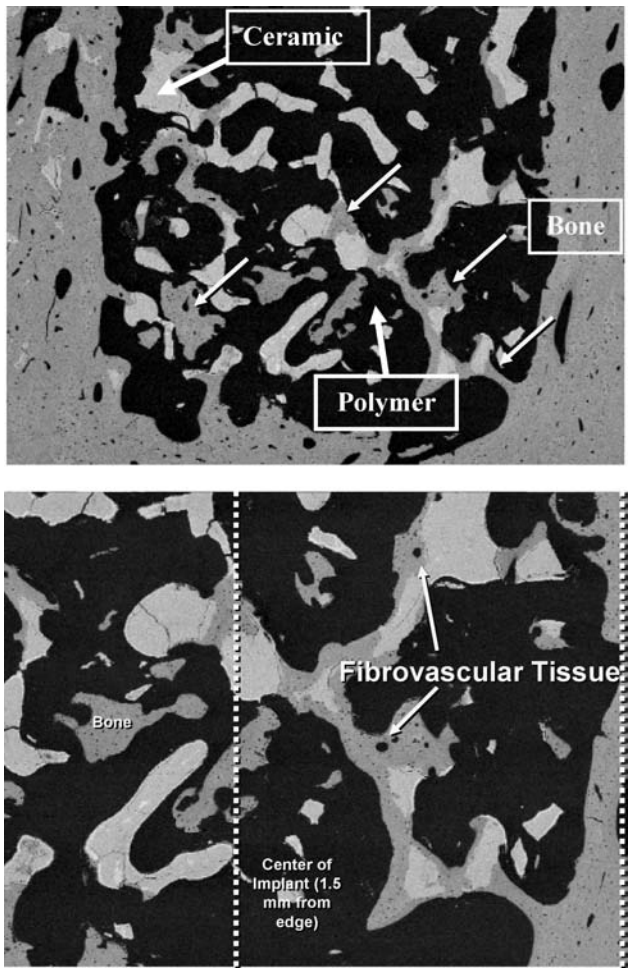


Figure 11. Backscattered electron microscope images demonstrating bone and blood vessel in-growth into a CPC implant (bone is shown in gray, ceramic is white, and polymer is black).

end result is a composite material with continuous seams of ceramic running through the polymer. Similar to the particulate composites, the CPC material will allow for bone growth on the surface and into the ceramic regions. However, the continuity of the ceramic phase throughout the composite gives the material a unique ability to allow for bone to penetrate into the center of a CPC implant. Figure 11 shows the histology a CPC implant composed of the Pro Osteon porous ceramic infiltrated with poly(L/D,L lactic acid) implanted in a sheep femur at 9 months. This backscattered electron microscope image shows the ability of a CPC implant to support bone and blood vessel in-growth into the center of the implant wall. In addition to acting as a structural implant, a CPC device also functions as an eventual scaffold for bone in-growth. From a healing standpoint, this type of composite will result in more bone formation at the site. Additionally, the presence of bone and blood vessels within the implant wall significantly improves the ability of the tissue to absorb the degradation products. Typically occurring at the surface of polymer implants, the presence of bone within the CPC material allows resorption throughout the entire device. This new

composite is currently being investigated for use as spinal fusion implants, fracture screws and plates, interference screws, and suture anchors.

CONCLUSION

As seen from this article, resorbable polymers and ceramics possess the desired properties needed for orthopedic implants. They have been shown to be versatile materials with a range in degradation rates and mechanical properties. The resorbable nature of these devices allows them to provide temporary stabilization and mechanical support. Combined with the ability to be completely resorbed by the body and replaced by natural tissue, these implants are highly desirable alternatives to their nonresorbing counterparts. The elimination of long-term implant complications and the ability to share load with regenerating tissues are large driving forces behind the use of these implants in orthopedics. With further advancements in biomaterial research, resorbable implants may soon become the standard of care.

BIBLIOGRAPHY

Cited References

1. History of Total Joint Replacement. Utah Hip and Knee Center. Available: http://www.utahhipandknee.com/history.htm. Accessed September 24, 2004.
2. History of Innovation. Zimmer Corporate Website. Available:http://www.zimmer.com/ctl?op=global&action=1&id=1420&template= CP. Accessed September 24, 2004.
3. Parker DA, Dunbar MJ, Rorabeck CH. Extensor mechanism failure associated with total knee arthroplasty: Prevention and management. J Am Acad Orthop Surg 2003;11:238–247.
4. Ries MD. Complications in primary total hip arthroplasty: Avoidance and management: wear. Instrum Course Lect 2003;52:257–265.
5. Sochart DH. Relationship of acetabular wear to osteolysis and loosening in total hip arthroplasty. Clin Orthop 1999; 135–150.
6. Moreland JR. Mechanisms of failure in total knee arthroplasty. Clin Orthop 1988; 49–64.
7. Borden MD. The development of bone graft materials using various formulations of demineralized bone matrix. In Laurencin CT, editor, Bone Graft Substitutes. Conshohocken, (PA): ASTM International; 2003.
8. Uhthoff HK, Finnegan M. The effects of metal plates on post-traumatic remodelling and bone mass. J Bone Joint Surg Br 1983;65:66–71.
9. van der EM, Dijkema AR, Klein CP, Patka P, Haarman HJ. Tissue reaction on PLLA versus stainless steel interlocking nails for fracture fixation: An animal study. Biomaterials 1995;16:103–106.
10. Bos RR, Rozema FR, Boering G, Nijenhuis AJ, Pennings AJ, Verwey AB. Bio-absorbable plates and screws for internal fixation of mandibular fractures. A study in six dogs. Int J Oral Maxillofac Surg 1989;18:365–369.
11. Viljanen J, Kinnunen J, Bondestam S, Majola A, Rokkanen P, Tormala P. Bone changes after experimental osteotomies fixed with absorbable self-reinforced poly-L-lactide screws or metallic screws studied by plain radiographs, quantitative computed tomography and magnetic resonance imaging. Biomaterials 1995;16:1353–1358.
12. van der EM, Klein CP, Blieck-Hogervorst JM, Patka P, Haarman HJ. Bone tissue response to biodegradable polymers used for intra medullary fracture fixation: A long-term in vivo study in sheep femora. Biomaterials 1999; 20:121–128.
13. Engelberg I, Kohn J. Physico-mechanical properties of degradable polymers used in medical applications: A comparative study. Biomaterials 1991;12:292–304.
14. Kulkarni RK, Pani KC, Neuman C, Leonard F. Polylactic acid for surgical implants. Arch Surg 1966;93:839–843.
15. Chu CC. A comparison of the effect of pH on the biodegradation of two synthetic absorbable sutures. Ann Surg 1982;195:55–59.
16. Chu CC. Mechanical properties of suture materials: An important characterization. Ann Surg 1981;193:365–371.
17. Chu CC, Campbell ND. Scanning electron microscopic study of the hydrolytic degradation of poly(glycolic acid) suture. J Biomed Mater Res 1982;16:417–430.
18. Chu CC. The in-vitro degradation of poly(glycolic acid) sutures–effect of pH. J Biomed Mater Res 1981;15:795–804.
19. Leenslag JW, Pennings AJ, Bos RR, Rozema FR, Boering G. Resorbable materials of poly(L-lactide). VI. Plates and screws for internal fracture fixation. Biomaterials 1987;8:70–73.
20. Tuompo P, Partio EK, Jukkala-Partio K, Pohjonen T, Helevirta P, Rokkanen P. Comparison of polylactide screw and expansion bolt in bioabsorbable fixation with patellar tendon bone graft for anterior cruciate ligament rupture of the knee. A preliminary study. Knee Surg Sports Traumatol Arthrosc 1999;7:296–302.
21. Koskikare K, Hirvensalo E, Patiala H, Rokkanen P, Pohjonen T, Tormala P, Lob G. Fixation of osteotomies of the distal femur with absorbable, self-reinforced, poly-L-lactide plates. An experimental study on rabbits. Arch Orthop Trauma Surg 1997;116:352–356.
22. Makela EA, Vainionpaa S, Vihtonen K, Mero M, Laiho J, Tormala P, Rokkanen P. Healing of physeal fracture after fixation with biodegradable self-reinforced polyglycolic acid pins. An experimental study on growing rabbits. Clin Mater 1990;5:1–12.
23. Pihlajamaki H, Bostman O, Hirvensalo E, Tormala P, Rokkanen P. Absorbable pins of self-reinforced poly-L-lactic acid for fixation of fractures and osteotomies. J Bone Joint Surg Br 1992;74:853–857.
24. Tormala P, Vasenius J, Vainionpaa S, Laiho J, Pohjonen T, Rokkanen P. Ultra-high-strength absorbable self-reinforced polyglycolide (SR-PGA) composite rods for internal fixation of bone fractures: In vitro and in vivo study. J Biomed Mater Res 1991;25:1–22.
25. Vainionpaa S, Kilpikari J, Laiho J, Helevirta P, Rokkanen P, Tormala P. Strength and strength retention in vitro, of absorbable, self-reinforced polyglycolide (PGA) rods for fracture fixation. Biomaterials 1987;8:46–48.
26. Barber FA, Deck MA. The in vivo histology of an absorbable suture anchor: A preliminary report. Arthroscopy 1995;11:77–81.
27. Borden M, El Amin SF, Attawia M, Laurencin CT. Structural and human cellular assessment of a novel microsphere-based tissue engineered scaffold for bone repair. Biomaterials 2003;24:597–609.
28. Borden M, Attawia M, Laurencin CT. The sintered microsphere matrix for bone tissue engineering: in vitro osteoconductivity studies. J Biomed Mater Res 2002;61:421–429.
29. Borden M, Attawia M, Khan Y, Laurencin CT. Tissue engineered microsphere-based matrices for bone repair: design and evaluation. Biomaterials 2002;23:551–559.

30. Thomson RC, Yaszemski MJ, Powers JM, Mikos AG. Fabrication of biodegradable polymer scaffolds to engineer trabecular bone. J Biomater Sci Polym Ed 1995;7:23–38.
31. Hollier LH, Rogers N, Berzin E, Stal S. Resorbable mesh in the treatment of orbital floor fractures. J Craniofac Surg 2001;12:242–246.
32. Edwards RC, Kiely KD, Eppley BL. The fate of resorbable poly-L-lactic/polyglycolic acid (LactoSorb) bone fixation devices in orthognathic surgery. J Oral Maxillofac Surg 2001;59:19–25.
33. Edwards RC, Kiely KD, Eppley BL. Resorbable PLLA-PGA screw fixation of mandibular sagittal split osteotomies. J Craniofac Surg 1999;10:230–236.
34. Westermark A. LactoSorb resorbable osteosynthesis after sagittal split osteotomy of the mandible: A 2-year follow-up. J Craniofac Surg 1999;10:519–522.
35. Eppley BL, Morales L, Wood R, Pensler J, Goldstein J, Havlik RJ, Habal M, Losken A, Williams JK, Burstein F, Rozzelle AA, Sadove AM. Resorbable PLLA-PGA plate and screw fixation in pediatric craniofacial surgery: Clinical experience in 1883 patients. Plast Reconstr Surg 2004;114:850–856.
36. Toth JM, Wang M, Scifert JL, Cornwall GB, Estes BT, Seim HB, III, Turner AS. Evaluation of 70/30 D,L-PLa for use as a resorbable interbody fusion cage. Orthopedics 2002;25: s1131–s1140.
37. Vaccaro AR, Madigan L. Spinal applications of bioabsorbable implants. Orthopedics 2002;25:s1115–s1120.
38. Lowe TG, Coe JD. Bioresorbable polymer implants in the unilateral transforaminal lumbar interbody fusion procedure. Orthopedics 2002;25:s1179–s1183.
39. Bostman OM. Osteolytic changes accompanying degradation of absorbable fracture fixation implants. J Bone Joint Surg Br 1991;73:679–682.
40. Pelto-Vasenius K, Hirvensalo E, Vasenius J, Rokkanen P. Osteolytic changes after polyglycolide pin fixation in chevron osteotomy. Foot Ankle Int 1997;18:21–25.
41. Bostman OM. Reaction to biodegradable implants. J Bone Joint Surg Br 1993;75:336–337.
42. Bostman OM. Intense granulomatous inflammatory lesions associated with absorbable internal fixation devices made of polyglycolide in ankle fractures. Clin Orthop 1992; 193–199.
43. Bostman OM. Osteoarthritis of the ankle after foreign-body reaction to absorbable pins and screws: A three- to nine-year follow-up study. J Bone Joint Surg Br 1998;80:333–338.
44. Bostman OM, Pihlajamaki HK. Adverse tissue reactions to bioabsorbable fixation devices. Clin Orthop 2000; 216–227.
45. Bostman O, Pihlajamaki H. Clinical biocompatibility of biodegradable orthopedic implants for internal fixation: A review. Biomaterials 2000;21:2615–2621.
46. Rovinsky D, Nissen TP, Otsuka NY. Osteolytic reaction to polylevolactic acid fracture fixation. Orthopedics 2001;24: 177–179.
47. Bostman O, Hirvensalo E, Makinen J, Rokkanen P. Foreign-body reactions to fracture fixation implants of biodegradable synthetic polymers. J Bone Joint Surg Br 1990;72:592–596.
48. Taylor MS, Daniels AU, Andriano KP, Heller J. Six bioabsorbable polymers: in vitro acute toxicity of accumulated degradation products. J Appl Biomater 1994;5:151–157.
49. Bergsma JE, de Bruijn WC, Rozema FR, Bos RR, Boering G. Late degradation tissue response to poly(L-lactide) bone plates and screws. Biomaterials 1995;16:25–31.
50. Bergsma EJ, Rozema FR, Bos RR, de Bruijn WC. Foreign body reactions to resorbable poly(L-lactide) bone plates and screws used for the fixation of unstable zygomatic fractures. J Oral Maxillofac Surg 1993;51:666–670.
51. Lanman TH, Hopkins TJ. Lumbar interbody fusion after treatment with recombinant human bone morphogenetic protein-2 added to poly(L-lactide-*co*-D,L-lactide) bioresorbable implants. Neurosurg Focus 2004;16:E9.
52. Couture DE, Branch CL, Jr. Posterior lumbar interbody fusion with bioabsorbable spacers and local autograft in a series of 27 patients. Neurosurg Focus 2004;16:E8.
53. Vaccaro AR, Robbins MM, Madigan L, Albert TJ, Smith W, Hilibrand AS. Early findings in a pilot study of anterior cervical fusion in which bioabsorbable interbody spacers were used in the treatment of cervical degenerative disease. Neurosurg Focus 2004;16:E7.
54. Cornwall GB, Ames CP, Crawford NR, Chamberlain RH, Rubino AM, Seim HB, III, Turner AS. In vivo evaluation of bioresorbable polylactide implants for cervical graft containment in an ovine spinal fusion model. Neurosurg Focus 2004;16:E5.
55. Lippman CR, Hajjar M, Abshire B, Martin G, Engelman RW, Cahill DW. Cervical spine fusion with bioabsorbable cages. Neurosurg Focus 2004;16:E4.
56. Robbins MM, Vaccaro AR, Madigan L. The use of bioabsorbable implants in spine surgery. Neurosurg Focus 2004;16:E1.
57. Ames CP, Crawford NR, Chamberlain RH, Cornwall GB, Nottmeier E, Sonntag VK. Feasibility of a resorbable anterior cervical graft containment plate. Orthopedics 2002;25: s1149–s1155.
58. DiAngelo DJ, Kitchel S, McVay BJ, Scifert JL, Cornwall GB. Bioabsorbable anterior lumbar plate fixation in conjunction with anterior interbody fusion cages. Orthopedics 2002;25: s1157–s1165.
59. Speer KP, Warren RF, Pagnani M, Warner JJ. An arthroscopic technique for anterior stabilization of the shoulder with a bioabsorbable tack. J Bone Joint Surg Am 1996;78:1801–1807.
60. Warner JJ, Miller MD, Marks P, Fu FH. Arthroscopic Bankart repair with the Suretac device. Part I: Clinical observations. Arthroscopy 1995;11:2–13.
61. Warner JJ, Miller MD, Marks P. Arthroscopic Bankart repair with the Suretac device. Part II: Experimental observations. Arthroscopy 1995;11:14–20.
62. Bourke SL, Kohn J. Polymers derived from the amino acid L-tyrosine: polycarbonates, polyarylates and copolymers with poly(ethylene glycol). Adv Drug Deliv Rev 2003;55:447–466.
63. Ertel SI, Kohn J. Evaluation of a series of tyrosine-derived polycarbonates as degradable biomaterials. J Biomed Mater Res 1994;28:919–930.
64. Ertel SI, Kohn J, Zimmerman MC, Parsons JR. Evaluation of poly(DTH carbonate), a tyrosine-derived degradable polymer, for orthopedic applications. J Biomed Mater Res 1995;29:1337–1348.
65. Choueka J, Charvet JL, Koval KJ, Alexander H, James KS, Hooper KA, Kohn J. Canine bone response to tyrosine-derived polycarbonates and poly(L-lactic acid). J Biomed Mater Res 1996;31:35–41.
66. Tangpasuthadol V, Pendharkar SM, Peterson RC, Kohn J. Hydrolytic degradation of tyrosine-derived polycarbonates, a class of new biomaterials. Part II: 3-yr study of polymeric devices. Biomaterials 2000;21:2379–2387.
67. Chaikof EL, Matthew H, Kohn J, Mikos AG, Prestwich GD, Yip CM. Biomaterials and scaffolds in reparative medicine. Ann N Y Acad Sci 2002;961:96–105.
68. Kohn J, Langer R. Poly(iminocarbonates) as potential biomaterials. Biomaterials 1986;7:176–182.

69. Ibim SM, Uhrich KE, Bronson R, El Amin SF, Langer RS, Laurencin CT. Poly(anhydride-*co*-imides): in vivo biocompatibility in a rat model. Biomaterials 1998;19:941–951.
70. Ibim SE, Uhrich KE, Attawia M, Shastri VR, El Amin SF, Bronson R, Langer R, Laurencin CT. Preliminary in vivo report on the osteocompatibility of poly(anhydride-co-imides) evaluated in a tibial model. J Biomed Mater Res 1998;43:374–379.
71. Uhrich KE, Ibim SE, Larrier DR, Langer R, Laurencin CT. Chemical changes during in vivo degradation of poly(anhydride-imide) matrices. Biomaterials 1998;19: 2045–2050.
72. Katti DS, Lakshmi S, Langer R, Laurencin CT. Toxicity, biodegradation and elimination of polyanhydrides. Adv Drug Deliv Rev 2002;54:933–961.
73. Attawia MA, Uhrich KE, Botchwey E, Langer R, Laurencin CT. In vitro bone biocompatibility of poly (anhydride-co-imides) containing pyromellitylimidoalanine. J Orthop Res 1996;14:445–454.
74. Ray JA, Doddi N, Regula D, Williams JA, Melveger A. Polydioxanone (PDS), a novel monofilament synthetic absorbable suture. Surg Gynecol Obstet 1981;153:497–507.
75. Ping OC, Cameron RE. The hydrolytic degradation of polydioxanone (PDSII) sutures. Part I: Morphological aspects. J Biomed Mater Res 2002;63:280–290.
76. Ping OC, Cameron RE. The hydrolytic degradation of polydioxanone (PDSII) sutures. Part II: Micromechanisms of deformation. J Biomed Mater Res 2002;63:291–298.
77. Ray JA, Doddi N, Regula D, Williams JA, Melveger A. Polydioxanone (PDS), a novel monofilament synthetic absorbable suture. Surg Gynecol Obstet 1981;153:497–507.
78. Bartholomew RS. PDS (polydioxanone suture): A new synthetic absorbable suture in cataract surgery. A preliminary study. Ophthalmologica 1981;183:81–85.
79. Lerwick E. Studies on the efficacy and safety of polydioxanone monofilament absorbable suture. Surg Gynecol Obstet 1983;156:51–55.
80. Cohen EL, Kirschenbaum A, Glenn JF. Preclinical evaluation of PDS (polydioxanone) synthetic absorbable suture vs chromic surgical gut in urologic surgery. Urology 1987;30:369–372.
81. Apt L, Henrick A. "Tissue-drag" with polyglycolic acid (Dexon) and polyglactin 910 (Vicryl) sutures in strabismus surgery. J Pediatr Ophthalmol 1976;13:360–364.
82. Homsy CA, McDonald KE, Akers WW, Short C, Freeman BS. Surgical suture-canine tissue interaction for six common suture types. J Biomed Mater Res 1968;2:215–230.
83. Rhee SH, Lee YK, Lim BS, Yoo JJ, Kim HJ. Evaluation of a novel poly(epsilon-caprolactone)-organosiloxane hybrid material for the potential application as a bioactive and degradable bone substitute. Biomacromolecules 2004;5: 1575–1579.
84. Rhee SH. Bone-like apatite-forming ability and mechanical properties of poly(epsilon-caprolactone)/silica hybrid as a function of poly(epsilon-caprolactone) content. Biomaterials 2004;25:1167–1175.
85. Ciapetti G, Ambrosio L, Savarino L, Granchi D, Cenni E, Baldini N, Pagani S, Guizzardi S, Causa F, Giunti A. Osteoblast growth and function in porous poly epsilon-caprolactone matrices for bone repair: a preliminary study. Biomaterials 2003;24:3815–3824.
86. Im SY, Cho SH, Hwang JH, Lee SJ. Growth factor releasing porous poly (epsilon-caprolactone)-chitosan matrices for enhanced bone regenerative therapy. Arch Pharm Res 2003;26:76–82.
87. Wang YW, Wu Q, Chen J, Chen GQ. Evaluation of three-dimensional scaffolds made of blends of hydroxyapatite and poly(3-hydroxybutyrate-*co*-3-hydroxyhexanoate) for bone reconstruction. Biomaterials 2005;26:899–904.
88. Yang M, Zhu S, Chen Y, Chang Z, Chen G, Gong Y, Zhao N, Zhang X. Studies on bone marrow stromal cells affinity of poly (3-hydroxybutyrate-*co*-3-hydroxyhexanoate). Biomaterials 2004;25:1365–1373.
89. Kose GT, Kenar H, Hasirci N, Hasirci V. Macroporous poly(3-hydroxybutyrate-*co*-3-hydroxyvalerate) matrices for bone tissue engineering. Biomaterials 2003;24:1949–1958.
90. Farso NF, Karring T, Gogolewski S. Biodegradable guide for bone regeneration. Polyurethane membranes tested in rabbit radius defects. Acta Orthop Scand 1992;63:66–69.
91. Gorna K, Gogolewski S. Preparation, degradation, and calcification of biodegradable polyurethane foams for bone graft substitutes. J Biomed Mater Res 2003;67A:813–827.
92. Grad S, Kupcsik L, Gorna K, Gogolewski S, Alini M. The use of biodegradable polyurethane scaffolds for cartilage tissue engineering: potential and limitations. Biomaterials 2003; 24:5163–5171.
93. Warrer K, Karring T, Nyman S, Gogolewski S. Guided tissue regeneration using biodegradable membranes of polylactic acid or polyurethane. J Clin Periodontol 1992;19:633–640.
94. Ambrosio AM, Sahota JS, Runge C, Kurtz SM, Lakshmi S, Allcock HR, Laurencin CT. Novel polyphosphazene-hydroxyapatite composites as biomaterials. IEEE Eng Med Biol Mag 2003;22:18–26.
95. Laurencin CT, El Amin SF, Ibim SE, Willoughby DA, Attawia M, Allcock HR, Ambrosio AA. A highly porous 3-dimensional polyphosphazene polymer matrix for skeletal tissue regeneration. J Biomed Mater Res 1996;30:133–138.
96. Laurencin CT, Norman ME, Elgendy HM, El Amin SF, Allcock HR, Pucher SR, Ambrosio AA. Use of polyphosphazenes for skeletal tissue regeneration. J Biomed Mater Res 1993;27:963–973.
97. Coetzee AS. Regeneration of bone in the presence of calcium sulfate. Arch Otolaryngol 1980;106:405–409.
98. Gitelis S, Piasecki P, Turner T, Haggard W, Charters J, Urban R. Use of a calcium sulfate-based bone graft substitute for benign bone lesions. Orthopedics 2001;24:162–166.
99. Ladd AL, Pliam NB. Use of bone-graft substitutes in distal radius fractures. J Am Acad Orthop Surg 1999;7:279–290.
100. Guarnieri R, Pecora G, Fini M, Aldini NN, Giardino R, Orsini G, Piattelli A. Medical grade calcium sulfate hemihydrate in healing of human extraction sockets: Clinical and histological observations at 3 months. J Periodontol 2004;75:902–908.
101. Kelly CM, Wilkins RM. Treatment of benign bone lesions with an injectable calcium sulfate-based bone graft substitute. Orthopedics 2004;27:s131–s135.
102. Urban RM, Turner TM, Hall DJ, Infanger S, Cheema N, Lim TH. Healing of large defects treated with calcium sulfate pellets containing demineralized bone matrix particles. Orthopedics 2003;26:s581–s585.
103. Turner TM, Urban RM, Hall DJ, Infanger S, Gitelis S, Petersen DW, Haggard WO. Osseous healing using injectable calcium sulfate-based putty for the delivery of demineralized bone matrix and cancellous bone chips. Orthopedics 2003;26:s571–s575.
104. Thalgott JS, Giuffre JM, Fritts K, Timlin M, Klezl Z. Instrumented posterolateral lumbar fusion using coralline hydroxyapatite with or without demineralized bone matrix, as an adjunct to autologous bone. Spine J 2001;1:131–137.
105. McConnell JR, Freeman BJ, Debnath UK, Grevitt MP, Prince HG, Webb JK. A prospective randomized comparison of coralline hydroxyapatite with autograft in cervical interbody fusion. Spine 2003;28:317–323.
106. Thalgott JS, Klezl Z, Timlin M, Giuffre JM. Anterior lumbar interbody fusion with processed sea coral (coralline

hydroxyapatite) as part of a circumferential fusion. Spine 2002;27:E518–E525.
107. Delecrin J, Takahashi S, Gouin F, Passuti N. A synthetic porous ceramic as a bone graft substitute in the surgical management of scoliosis: A prospective, randomized study. Spine 2000;25:563–569.
108. McAndrew MP, Gorman PW, Lange TA. Tricalcium phosphate as a bone graft substitute in trauma: Preliminary report. J Orthop Trauma 1988;2:333–339.
109. Bucholz RW, Carlton A, Holmes RE. Hydroxyapatite and tricalcium phosphate bone graft substitutes. Orthop Clin North Am 1987;18:323–334.
110. Holmes R, Mooney V, Bucholz R, Tencer A. A coralline hydroxyapatite bone graft substitute. Preliminary report. Clin Orthop 1984; 252–262.
111. Finn RA, Bell WH, Brammer JA. Interpositional "grafting" with autogenous bone and coralline hydroxyapatite. J Maxillofac Surg 1980;8:217–227.
112. Holmes RE. Bone regeneration within a coralline hydroxyapatite implant. Plast Reconstr Surg 1979;63:626–633.
113. Holmes RE, Salyer KE. Bone regeneration in a coralline hydroxyapatite implant. Surg Forum 1978;29:611–612.
114. Jamali A, Hilpert A, Debes J, Afshar P, Rahban S, Holmes R. Hydroxyapatite/calcium carbonate (HA/CC) vs. plaster of Paris: A histomorphometric and radiographic study in a rabbit tibial defect model. Calcif Tissue Int 2002;71:172–178.
115. Kenny SM, Buggy M. Bone cements and fillers: A review. J Mater Sci Mater Med 2003;14:923–938.
116. Horstmann WG, Verheyen CC, Leemans R. An injectable calcium phosphate cement as a bone-graft substitute in the treatment of displaced lateral tibial plateau fractures. Injury 2003;34:141–144.
117. Kamano M, Honda Y, Kazuki K, Yasudab M. Palmar plating with calcium phosphate bone cement for unstable Colles' fractures. Clin Orthop 2003; 285–290.
118. Zimmermann R, Gabl M, Lutz M, Angermann P, Gschwentner M, Pechlaner S. Injectable calcium phosphate bone cement Norian SRS for the treatment of intra-articular compression fractures of the distal radius in osteoporotic women. Arch Orthop Trauma Surg 2003;123:22–27.
119. Schildhauer TA, Bauer TW, Josten C, Muhr G. Open reduction and augmentation of internal fixation with an injectable skeletal cement for the treatment of complex calcaneal fractures. J Orthop Trauma 2000;14:309–317.

See also DRUG DELIVERY SYSTEMS; MATERIALS AND DESIGN FOR ORTHOPEDIC DEVICES; POROUS MATERIALS FOR BIOLOGICAL APPLICATIONS.

BIOMATERIALS: AN OVERVIEW

BRANDON L. SEAL
ALYSSA PANITCH
Arizona State University
Tempe, Arizona

INTRODUCTION

Biomaterials are materials that are used or that have been designed for use in medical devices or in contact with the body. Traditionally, they consist of metallic, ceramic, or synthetic polymeric materials, but more recent developments in biomaterials design have attempted to incorporate materials derived from or inspired by biological materials (e.g., silk and collagen). Often, the use of biomaterials focuses on the augmentation, replacement, or restoration of diseased or damaged tissues and organs. The prevalence of biomaterials within society is most evident within medical and dental offices, pharmacies, and hospitals. However, the influence of biomaterials has reached into many households with examples ranging from increasingly common news media coverage of medical breakthroughs to the availability of custom color non-corrective contact lenses.

The evolving character of the discipline of biomaterials is evidenced by how the term biomaterial has been defined. In 1974, the Clemson Advisory Board, in response to a request by the World Health Organization (WHO), stated that a biomaterial is a "systemically pharmacologically inert substance designed for implantation within or incorporation with living tissue" (1). Dr. Jonathan Black further modified this definition to state that a biomaterial is "any pharmacologically inert material, viable or nonviable, natural product or manmade, that is part of or is capable of interacting in a beneficial way with a living organism" (1). An National Institute of Health (NIH) consensus definition appeared in 1983 and defined biomaterials as "any substance (other than a drug) or combination of substances, synthetic or natural in origin, which can be used for any period of time, as a whole or as a part of a system that treats, augments, or replaces any tissue, organ, or function of the body" (2). Thus, relatively newer definitions of the term biomaterial recognize that more modern medical and diagnostic devices will rely increasingly upon direct biological interaction between biological molecules, cells, and tissues and the materials from which these devices are manufactured.

HISTORY OF BIOMATERIALS

Compared with the much larger field of materials science, the field of biomaterials is relatively new. Although there exist recorded cases of glass eyes and metallic or wooden dental implants (some of which can be dated back to ancient Egypt), the modern age of biomaterials could not have existed without the adoption of aseptic surgical techniques pioneered by Sir Joseph Lister in the mid-nineteenth century and indeed, did not fully emerge as an industry or discipline until after the development of synthetic polymers just prior to, during, and following World War II. Prior to World War II, implanted biomaterials consisted primarily of metals (e.g., steel, used in pins and plates for bone fixation, joint replacements, and the covering of bone defects). In the late 1940s, Harold Ridley observed that shards of poly(methyl methacrylate) (PMMA), from airplane cockpit windshields, embedded within the eyes of World War II aviators did not provoke much of an inflammatory response (3). This observation led not only to the development of PMMA intraocular lenses, but also to greater experimentation of available materials, especially polymers, as biomaterials that could be placed in direct contact with living tissue.

As the fields of cellular, molecular, and developmental biology began to grow during the 1970s and 1980s, new insights into the organization, function, and properties of biological systems, tissues, and interactions led to a greater understanding of how cells respond to their environment. This wealth of biological information allowed the field of biomaterials to undergo a paradigm shift. Instead of focusing primarily on replacing an organ or a tissue with a synthetic, usually nondegradable biomaterial, a new branch of biomaterials would attempt to combine biologically active molecules, therapeutics, and motifs into existing and novel biomaterial systems derived from both synthetic and natural sources (4–6). Although there exist many examples of successful, commercially available biomaterials consisting of metallic and ceramic bases, the focus of biomaterials research has shifted to the development of polymeric or composite materials with biologically sensitive or environmentally controlled properties. This change has resulted largely due to the reactivity and variety of chemical moieties that are found in or that can be engineered into natural and synthetic polymers. Indeed, by viewing biomaterials as materials designed to interact with biology rather than being inert substances, the field of biomaterials has exploded with innovative designs that promote cell attachment, encapsulation, proliferation, differentiation, migration, and apoptosis, and that allow the biomaterial to polymerize, swell, and degrade under a variety of environmental conditions and biological stimuli. Evidence of this polymer and composite revolution is the dramatic increase in the number of publications relating to biomaterials research. Figure 1 shows a plot of the number of journal articles with biomaterial or biomaterials in their title, abstract, or keyword as a function of publication year as searched in the Web of Science database. As seen in Fig. 1, publications matching the search criteria have increased exponentially starting around the early 1990s and continuing until the present. The number of scientific journals, shown in Table 1, related to research in the field of biomaterials has also grown. Although the number of journal articles related to biomaterials research may have resulted primarily from the large increase in the number of biomaterials-related scientific journals, the exponential growth of biomaterials is evidenced further by the growth of the number of bioengineering or biomedical engineering departments (the department in which most biomaterials programs reside) established at universities throughout the United States (Fig. 1).

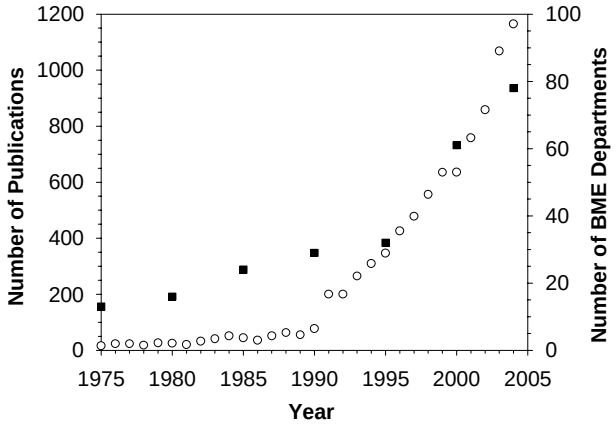


Figure 1. A plot of the number of publications (○) containing the word biomaterial or biomaterials in the title, abstract, or keywords, as searched in the Web of Science database, as a function of the publication year as well as a plot of the number of bioengineering or biomedical engineering departments (BME departments) within the United States as a function of time (■).

MARKET SIZE AND TYPES OF APPLICATIONS

The field of biomaterials, by nature, is interdisciplinary. Successful biomaterial designs have involved talents, knowledge, and expertise provided by physicians and clinicians, materials scientists, engineers, chemists, biologists, and physicists. As a result, it is not surprising that the biomaterials industry is both relatively young and very diversified. The diversity of this industry has resulted from the types of products created and marketed, the size and location of involved companies, and the types of regulatory policies imposed by government agencies and third party reimbursement organizations. Specifically, the biomaterials industry is part of the Medical Device and Diagnostic Industry, a multibillion dollar industry comprised of organizations that design, fabricate, and/or manufacture materials that are used in the health and life science fields. The end use applications are medical and dental devices, prostheses, personal hygiene products, diagnostic devices, drug delivery vehicles, and biotechnology systems. Some examples of these applications include full and hybrid artificial organs, biosensors, vascular grafts, pacemakers, catheters, insulin pumps, cochlear implants, contact lenses, intraocular lenses, artificial joints and bones, burn dressings, and sutures. Table 2 shows a list of some common medical devices that require various biomaterials, and Table 3 displays a list of the prevalence and market potential of a few of these applications (7).

GOVERNMENT REGULATION

Within the United States, in a research only environment, biomaterials by themselves do not necessarily require government regulation. However, if any biomaterial is used within a medical or diagnostic device designed and destined for commercialization, the biomaterials used within the medical device (as well as the device itself) are subject to the jurisdiction of the U.S. Food and Drug Administration (FDA) as set forth in the Federal Food, Drug, and Cosmetic Act of 1938, the Medical Device Amendments of 1976, and the Food and Drug Administration Modernization Act of 1997. These laws have empowered the FDA to regulate conditions involving premarket controls, postmarket reporting, production involving Good Manufacturing Practices, and the registration and listing of medical devices. Any biomaterial within a marketed medical device prior to the Medical Device Amendments of 1976 were grandfathered and are considered approved materials. Modifications to these materials or

Table 1. A List of Journals with Publications Related to the Field of Biomaterials[a]

Name of Journal	Name of Journal	Name of Journal
Advanced Drug Delivery Reviews (1987)	Biosensors and Bioelectronics (1985)	Journal of Biomaterials Science: Polymer Edition (1990)
American Journal of Drug Delivery (2003)	Cells and Materials (1991)	Journal of Biomedical Materials Research (1967)
American Society of Artificial Internal Organs Journal (1955)	Cell Transplantation (1992)	Journal of Controlled Release (1984)
Annals of Biomedical Engineering (1973)	Clinical Biomechanics (1986)	Journal of Drug Targeting (1993)
Annual Review of Biomedical Engineering (1999)	Colloids and Surfaces B: Biointerfaces (1993)	Journal of Long Term Effects of Medical Implants (1991)
Artificial Organs (1977)	Dental Materials (1985)	Journal of Nanobiotechnology (2003)
Artificial Organs Today (1991)	Drug Delivery (1993)	Macromolecules (1968)
Biomacromolecules (2000)	Drug Delivery Systems and Sciences (2001)	Materials in Medicine (1990)
Biofouling (1985)	Drug Delivery Technology (2001)	Medical Device and Diagnostics Industry (1996)
Biomedical Engineering OnLine (2002)	e-biomed: the Journal of Regenerative Medicine (2000)	Medical Device Research Report (1995)
Bio-medical Materials and Engineering (1991)	European Cells and Materials (2001)	Medical Device Technology (1990)
Biomaterial-Living System Interactions (1993)	Federation of American Societies for Experimental Biology Journal (1987)	Medical Plastics and Biomaterials (1994)
Biomaterials (1980)	Frontiers of Medical and Biological Engineering (1991)	Nanobiology
Biomaterials, Artificial Cells and Artificial Organs (1973)	IEEE Transactions on Biomedical Engineering (1954)	Nanotechnology (1990)
Artificial Cells, Blood Substitutes, and Immobilization Biotechnology (1973)	International Journal of Artificial Organs (1976)	Nature: Materials (2002)
Biomaterials Forum (1979)	Journal of Bioactive and Compatible	Tissue Engineering (1995)
Biomedical Microdevices (1998)	Polymers (2002) Journal of Biomaterials Applications (2001)	Trends in Biomaterials and Artificial Organs (1986)

[a]The date of first publication is listed in parentheses following the name of each journal.

new materials are subject to controls established by the FDA. These controls consist of obtaining an Investigational Device Exemption for the medical device, including the biomaterials used within the device, prior to conducting clinical trials.

In addition, biomaterials can be considered part of a Class I, II, or III device depending on FDA classifications and depending on whether or not the biomaterial is considered to be part of a biologic, drug, or medical device. Class I devices are generally considered those devices needing the least amount of regulatory control since they do not present a great risk for a patient. Examples include tongue depressors and surgical drills. Class II devices represent a moderate risk to patients and require additional regulation (e.g., mandatory performance standards, additional labeling requirements, and postmarket surveillance). Some examples include X-ray systems and cardiac mapping catheters. Class III devices (e.g., cardiovascular stents and heart valves), represent those devices with the highest risk to patients and require extensive regulatory control. Usually, for biomaterials in direct contact with tissue within the body, devices are considered Class III devices and are subject to a Premarket Approval process before they can be sold within the United States. In general, for most biomaterials, some of the tests the FDA reviews to evaluate biomaterial safety includes tests

Table 2. Some Common Uses for Biomaterials

Organ/Procedure	Associated Medical Devices
Bladder	Catheters
Bone	Bone plates, joint replacements (metallic and ceramic)
Brain	Deep brain stimulator, hydrocephalus shunt, drug eluting polymers
Cardiovascular	Polymer grafts, metallic stents, drug eluting grafts
Cosmetic enhancement	Breast implants, injectable collagen
Eye	Intraocular lenses, contact lenses
Ear	Artificial cochlea, artificial stapes
Heart	Artificial heart, ventricular assist devices, heart valves, pacemakers
Kidney	Hemodialysis instrumentation
Knee	Metallic knee replacements
Lung	Blood oxygenator
Reproductive system	Hormone replacement patches, contraceptives
Skin	Artificial skin, living skin equivalents
Surgical	Scalpels, retractors, drills
Tissue repair	Sutures, bandages

Table 3. A Summary of the Prevalence and Economic Cost of Some of the Healthcare Treatments Requiring Biomaterials for the Year 2000

Medical Application[a]	Incident Patient Population[a]	Prevalent Patient Population[a]	Total Therapy Cost (Billions of US Dollars)[a]
Dialysis	188,000	1,030,000	$67
Cardiovascular			
Bypass grafts	733,000	6,000,000	$65
Valves	245,000	2,400,000	$27
Pacemakers	670,000	5,500,000	$44
Stents	1,750,000	2,500,000	$48
Joint replacement	1,285,000	7,000,000	$41
Hips	610,000		
Knees	675,000		

[a]All data taken from that reported by Lysaght and O'Loughlin (7).

involving cellular toxicity (both direct and indirect), acute and chronic inflammation, debris and degradation byproducts and associated clearance events, carcinogenicity, mutagenicity, fatigue, creep, tribology, and corrosion. Further information regarding FDA approval for medical devices can be found on the FDA webpage, www.fda.gov.

Many FDA approved biomaterials continue to be monitored for efficacy and safety in an effort not only to protect patients, but also to improve biocompatibility and reduce material failure. Perhaps the best known example of an FDA regulated biomaterial is silicone. Silicone had been used since the early 1960s in breast implants. As a result, silicone breast implants were grandfathered into the Medical Device Amendments of 1976. During the 1980s, some concerns regarding the safety of silicone breast implants arose and prompted the FDA to request, in 1990, additional safety data from manufacturers of breast implants. Due to fears of connective tissue disease, multiple sclerosis, and other ailments resulting from ruptured silicone implants, the FDA banned silicone breast implants in 1992. Recently, however, manufacturers (e.g., the Mentor Corporation) have applied for and received premarket approval for the sale of silicone breast implants contingent on the compliance of various conditions (8). Thus, silicone is a good example of the complexity surrounding the testing of both efficacy and safety for biomaterials.

TYPES OF BIOMATERIALS

Similar to the field of materials science, the field of biomaterials focuses on four major types of materials: metals, ceramics, polymers, and composites. Examples of a few selected medical devices made from these materials are shown in Fig. 2. The materials selected for any particular application depend on the properties desired for a particular function or set of functions. In all materials applications, the structure, properties, and processing of the selected material will affect performance. As a result, physicians, scientists and engineers who design biomaterials need to understand not only mechanical and physical properties of materials, but also biological properties of materials. Mechanical and physical properties include strength, fatigue, creep resistance, flexibility, permeability

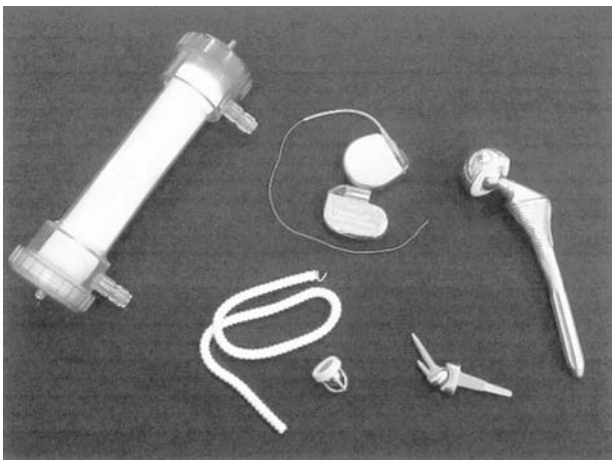


Figure 2. A representation of a few medical devices made from various biomaterials. From the upper left corner and moving clockwise, this picture shows a kidney hemodialyzer, two pacemakers, a hip replacement, an articulating wrist joint, a heart valve, and a vascular graft.

to gases and liquids, thermal and electrical properties, chemical reactivity, and degradation. Biological properties of materials largely focus on biocompatibility issues related to toxicity, immune system reactivity, thrombus formation, tribiology, inflammation, carcinogenic and teratogenic potential, integration with tissues and cells, and the ability to be sterilized. Regardless of the material, recent approaches to biomaterials research has focused on directing specific tissue interaction by using materials to introduce chemical bonds with the surrounding tissue, to act as scaffolds for tissue ingrowth, to introduce an inductive signal that will influence the behavior of surrounding cells or matrix, or to form new tissue when incubated or presented to transplanted cells.

Metals

Metals have been used as biomaterials for centuries. Although some fields (e.g., dentistry) continue to use amalgams, gold, and silver, most modern metallic biomaterials consist of iron, cobalt, titanium, or platinum bases. Since they are strong, metals are most often employed as biomaterials in orthopedic or fracture fixation medical devices; however, metals are also excellent conductors, and are therefore used for electrical stimulation of the heart, brain, nerves, muscle, and spinal cord. The most common alloys for orthopedic applications include stainless steel, cobalt, and titanium alloys. These alloys have enjoyed frequent use in medical procedures related to the function of joints and load-bearing. For example, metal alloys are commonly found in medical devices for knee replacement as well as in the femoral stem used in total hip replacements. Since all metals are subject to corrosion, especially in the salty, aqueous environment within the body, metals used as biomaterials often require an external oxide layer to protect against pitting and corrosion. These electrochemically inert oxide layers consist of Cr_2O_3 for stainless steel, Cr_2O_3 for cobalt alloys, and TiO_2 for

Figure 3. Photographs of stainless steel (a), cobalt–chromium (b), and titanium alloy (Ti6Al4V) (c) hip implants. (All three photographs are used with permission from the Department of Materials at Queen Mary University of London.)

titanium alloys. Figure 3 displays examples of three types of metallic hip replacements.

Stainless Steel Alloys. The stainless steel most commonly used as orthopedic biomaterials is classified 316L by the American Iron and Steel Institute. This particular austenitic alloy contains a very low carbon content (a maximum of 0.03%) and chromium content of 17–20%. The added chromium will react with oxygen to produce a corrosion-resistant chromium oxide layer. The 316L grade of stainless steel is a casting alloy, and its relatively high ductility makes this alloy amenable to extensive postcasting mechanical processing. Compared to cobalt and titanium alloys, stainless steel has a moderate yield and ultimate strength, but high ductility. Furthermore, it may be fabricated by virtually all machining and finishing processes and is generally the least expensive of the three major metallic alloys (4,5,9).

Cobalt Alloys. Cobalt alloys have been used since the early twentieth century as dental alloys and in heavily loaded joint applications. For use as a biomaterial, cobalt alloys are either cast (i.e., primarily formed within a mold) or wrought (i.e., worked into a final form from a large ingot). Two examples of cobalt alloys include Vitallium (designated F 75 by ASTM International), a cast alloy that consists of 27–30% chromium and >34% cobalt, and the wrought cobalt alloy MP35N (designated F 563 by ASTM International), which consists of 18–22% chromium, 15–25% nickel, and >34% cobalt. Compared to Vitallium, the MP35N alloy has demonstrated superior fatigue resistance, larger ultimate tensile strength, and a higher degree of corrosion resistance to chlorine. Consequently, this particular alloy is good for applications requiring long service life without fracture or stress fatigue. Compared to stainless steel alloys, cobalt-based alloys have slightly higher tensile moduli, but lower ductility. In addition, they are more expensive to manufacture and more difficult to machine. However, relative to stainless steel and titanium, cobalt-based alloys can offer the most useful balance of corrosion resistance, fatigue resistance, and strength (4,5,9).

Titanium Alloys. The most recent of the major orthopedic metallic alloys to be employed as biomaterials are titanium alloys. Although pure titanium is relatively weak and ductile, titanium can be stabilized by adding elements (e.g., aluminum and vanadium) to the alloy. Often, pure titanium (designated F 67 by ASTM International) is primarily used as a surface coating for orthopedic medical devices. For load-bearing applications, the alloy Ti6Al4V (designated F 136 by ASTM International) is much more widely used in implant manufacturing. As in the case of stainless steel and cobalt alloys, titanium contains an outer oxide layer, composed of TiO_2, that protects the implant from corrosion. In fact, of the three major orthopedic alloys, titanium shows the lowest rate of corrosion. Moreover, the density of titanium is almost half that of stainless steel and cobalt alloys. As a result, implants made from titanium are lighter and reduce patient awareness of the implant; however, titanium alloys are among the most expensive metallic biomaterials. Relative to stainless steel and cobalt alloys, titanium has a lower Young's modulus, which can aid in reducing the stresses around the implant by flexing with the bone. Titanium has a lower ductility than the other alloys, but does demonstrate high strength. These properties allow titanium alloys to play a diverse role as a biomaterial. Titanium alloys are used in parts for total joint replacements, screws, nails, pacemaker cases, and leads for implantable electrical stimulators (4,5,9).

Despite the reduced weight and improved mechanical match of titanium alloy implants to bone relative to stainless steel and chromium alloy implants, titanium alloy implants still exhibit issues with regard to mechanical mismatch. This problem stems from the large differences in properties (e.g., elastic moduli) between bone, metals, and polymers used as acetabular cups. For example, metals have elastic moduli ranging from ~100 to 200 GPa, ultrahigh molecular weight polyethylene has an elastic modulus of 1–2 GPa, and the elastic modulus of cortical bone is ~ 12 GPa (10). In addition, it is difficult to produce a titanium implant surface that is conducive to bone ingrowth or attachment. Novel titanium foams have been investigated as a method for reducing implant weight, better matching tissue mechanics, and improving bone ingrowth. The process involves mixing titanium powder with ammonium hydrogen carbonate powder and compressing and heating the mixture to form foams with densities varying from 0.2 to 0.65 times the density of solid titanium. These densities are close to those of cancellous bone (0.2–0.3 times the density of solid titanium) and cortical bone (0.5–0.65 times the density of solid titanium) (11). While they are preliminary, studies with novel materials such as these titanium foams illustrate a trend toward the development of materials that better mimic the properties of the native tissue they are designed to replace.

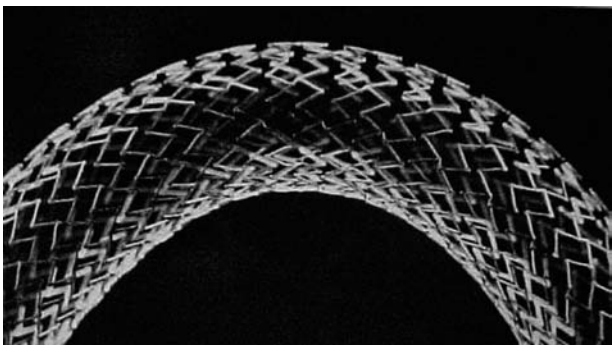


Figure 4. A photograph of the SMARTeR nitinol stent developed by the Cordis Corporation. (Reprinted from Ref. 12, with permission from Royal College of Radiologists.)

Other Metals. Besides stainless steel, cobalt alloys, and titanium alloys, there exist other examples of metals used as biomaterials. Some examples include nitinol, a single-phase nickel/titanium shape memory alloy, tantalum, a very dense, chemically inert, weak, but fatigue-resistant metal, and platinum, a very expensive metal used by itself or with iridum as a corrosion-resistant electrical conductor for electrode applications. Nitinol stents (e.g., that seen in Fig. 4) (12) and drug-eluting nitinol stents used for cardiovascular applications recently have seen enormous medical and commercial success. Indeed, metallic stents have significantly changed the way coronary blockages are treated (4,5,9).

Ceramics

Of the major types of materials used as biomaterials, ceramics have not been used as frequently as metals, polymers, or composites. However, ceramics continue to enjoy widespread use in certain bone-related applications (e.g., dentistry and joint replacement surgeries), due to their high compressive strength, high degree of hardness, excellent biocompatibility, superior tribological properties, and chemical inertness. Although they are very strong in compression, ceramics are susceptible to mechanical and thermal loading, have lower tensile strengths relative to other materials, and are very brittle in tension; this brittleness limits potential biomaterials applications.

Ceramics consist of a network of metal and nonmetal ions, with the general structure X_mY_n, arranged in a repeating structure. This structure depends on the relative size of the ions as well as the number of counterions needed to balance total charge. For example, if $m = n = 1$, and both ions are approximately the same size, then the structure would be of a simple cubic nature (e.g., CsCl or CsI); if the anion is much larger than the cation, then typically, a face centered cubic (fcc) structure would emerge (e.g., ZnS or CdS). If $m = 2$ and $n = 3$, as is the case with oxide ceramics (e.g., Al_2O_3), then a hexagonal closed pack structure would often result (13).

Ceramics used as biomaterials can be classified by processing–manufacturing methods, by chemical reactivity, or by ionic composition. Regarding chemical reactivity, ceramics can be bioinert, bioactive, or bioresorbable. Bioinert or nonresorbable ceramics are either porous or nonporous and are essentially not affected by the environment at the implant site. Bioactive or reactive ceramics are designed with specific surface properties that are intended to react with the local host environment and to elicit a desired tissue response. Bioresorbable ceramics dissolve over some prescribed period of time *in vivo* mediated by physiochemical processes. If one considers the application of bone replacement, then there would be about four ways for ceramics to interact with and attach to bone. First, a nonporous, inert ceramic material could be attached via glues, surface irregularities, or press-filling methods. Second, a porous, inert ceramic could be designed to have an optimal pore size, which promotes direct mechanical attachment of bone through bone ingrowth. Third, a nonporous, inert ceramic with a reactive surface could direct bone attachment via chemical bonding. Fourth, a nonporous or porous, resorbable ceramic could eventually be replaced by bone. When describing real examples of ceramics used as biomaterials, it is more useful to classify the ceramics based on ionic composition. This type of classification reveals a few major bioceramic groups: oxide ceramics, multiple oxides of calcium and phosphorus, glasses and glass ceramics, and carbon.

Oxide Ceramics. As their name implies, oxide ceramics consist of oxygen bound to a metallic species. Oxide ceramics are chemically inert, but can be nonporous or porous. One example of a nonporous oxide ceramic used as a biomaterial is aluminum oxide, Al_2O_3. Highly pure aluminum oxide (F 603 as designated by ASTM International), or alumina, has high corrosion resistance, good biocompatibility, high wear resistance, and good mechanical properties due to high density and small grain size. Aluminum oxide has been manufactured as an acetabular cup for total hip replacement. In comparison with metal or ultrahigh molecular weight polyethylene, Al_2O_3 provides better tribological properties by greatly decreasing friction within the joint and substantially increasing wear resistance. Recently, the FDA approved ceramic on ceramic hip replacements made from alumina and marketed by companies such as Wright Medical Technology and Stryker Osteonics. This ceramic on ceramic design is very resistant to wear and results in a much smaller amount of wear debris than traditional metal–polymer joints. With better wear properties and longer useful lifespan, ceramic on ceramic hip replacements likely will provide an attractive alternative to other biomaterial options, especially for younger patients that need better long-term solutions for joint replacements (4,5,9).

Ceramic oxides can also be porous. In bone formation, these pores are useful for allowing bone ingrowth, which will stabilize the mechanical properties of the implant without sacrificing the chemical inertness of the ceramic material. In general, there are three ways to make a porous ceramic oxide. First, a soluble metal or salt can be mixed with the ceramic and etched away. Second, a foaming agent that evolves gases during heating (e.g., calcium carbonate) can be mixed with the ceramic powder prior to firing. Third, the microstructure of corals can be used as a template to create a ceramic with a high degree

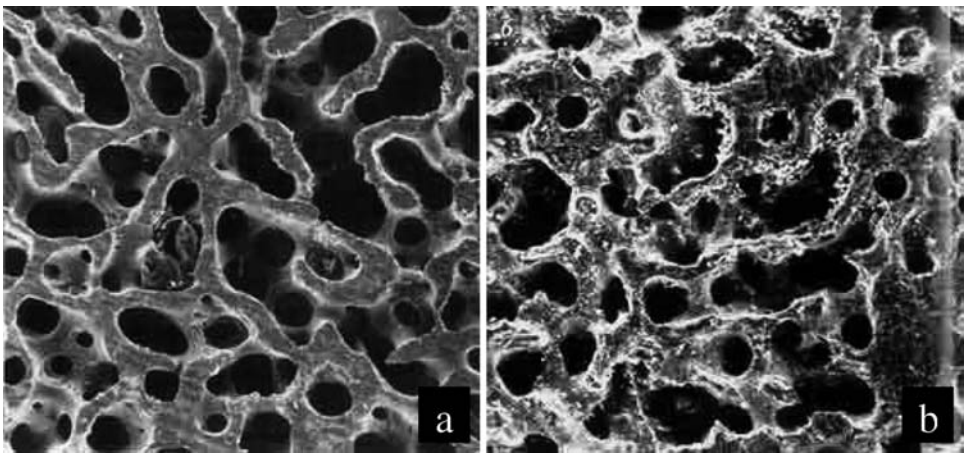


Figure 5. Photographs of (a) a cross-section of human cancellous bone and (b) coral of the genus Porites. These images illustrate how biologically derived materials (e.g., coral) can be used as scaffolds to create ceramic biomaterials that mimic the structure and porosity of natural bone. (Both photographs are used with permission from Biocoral, Inc.)

of interconnectivity and uniform pore size. In this third approach, coral is machined into the desired shape. Then, the coral is heated up to drive off carbon dioxide. The remaining calcium oxide provides a scaffold around which the ceramic material is deposited. After firing, the calcium oxide can be dissolved using hydrochloric acid. This dissolved calcium oxide will leave behind a very uniform and highly interconnected porous structure. Interestingly, the type of coral used will affect the pore size of the resulting ceramic. For example, if the genus *Porites* is used, then the pore size will range from 140 to 160 μm; the genus *Goniopora* will result in a pore size of 200–1000 μm (5). Porous ceramics do have many advantages for bone ingrowth, especially since the porous structure more closely mimics that of cancellous bone (see Fig. 5). However, the porous structure does result in a loss of strength and a tremendous increase in surface area that interacts with an *in vivo* saline environment.

Multiple Oxides of Calcium and Phosphorus. Aside from many types of proteins, the extracellular environment of bone contains a large concentration of organic mineral deposits known as hydroxyapatite. Chemically, hydroxyapatite generally has the following composition: $Ca_{10}(PO_4)_6(OH)_2$. Since hydroxyapatite is a naturally occurring ceramic produced by osteoblasts, it seemed reasonable to apply hydroxyapatite as filler or as a coating to allow better integration with existing bone. Coatings of hydroxyapatite have been applied (usually by plasma spraying) to metallic implants used in applications requiring bone ingrowth to provide a tight fit between bone and the implanted device, to minimize loosening over time, and to provide some measure of isolation from the foreign body response. Although hydroxyapatite is the most commonly used bioceramic containing calcium and phosphorus, there do exist other forms of calcium and phosphorus oxides including tricalcium phosphate, $Ca_3(PO_4)_2$, and octacalcium phosphate, $Ca_8H_2PO_4{\cdot}5H_2O$ (4,5,9,14).

Glasses and Glass Ceramics. Just as in the case of traditional glass, glass ceramics used as biomaterials contain large amounts of silica, SiO_2. Glass ceramics are formed using controlled crystallization techniques during which silica is cooled down at rate slow enough to allow the formation of a hexagonal crystal structure with small, crystalline grains (~1 μm) surrounded by an amorphous phase. Bioactive glass ceramics have been studied as biomaterials because they can attach directly to tissue via chemical bonds, they have a low thermal coefficient of expansion, they have good compressive mechanical strength, the mechanical strength of the glass–tissue interface is close to that of tissue, and they resist scratching and abrasion. Unfortunately, as with all ceramics, bioactive glasses are very brittle. Two well-known examples of commercially available glass ceramics include Bioglass, which consists of SiO_2, Na_2O, CaO, and P_2O_5, and Ceravital, which contains SiO_2, Na_2O, CaO, P_2O_5, K_2O, and MgO. Relative to traditional soda lime glass, bioactive glass ceramics contain lower amounts of SiO_2 and higher amounts of Na_2O and CaO. The high ratio of CaO to P_2O_5 in bioactive ceramics allows the rapid formation of a hydroxycarbonate apatite (HCA) layer at alkaline pH. For example, a 50 nm layer of HCA can form from Bioglass 45S5 after 1 h. The release of calcium, phosphorus, and sodium ions from bioactive ceramics also allows the formation of a water-rich gel near the ceramic surface. This cationic-rich environment creates a locally alkaline pH that helps to form HCA layers and provide areas of adhesion for biological molecules and cells (4,5,9).

Carbon. Processed carbon has been used in biomaterials applications as a bioceramic coating. Although carbon can exist in several forms (e.g., graphite, diamond), bioceramic carbons consist primarily of low temperature isotropic (LTI) and ultralow temperature isotropic (ULTI) carbon. This form of carbon is synthesized through the pyrolysis of hydrocarbon gases resulting in the deposition of isotropic carbon in a layer ~ 4 mm thick. Advantages to LTI and ULTI carbon include high strength, an elastic modulus close to that of bone, resistance to fatigue compared with other materials, excellent resistance to thrombosis, superior tribological properties, and excellent bond strength with metallic substances. The LTI carbon has been used as a coating for heart valves; however, applications remain limited primarily to coatings due to processing methods (4,5,9).

Polymers

Since the early to mid-twentieth century, the discovery of organic polymerization schemes and the advent of new polymeric species have fueled an incredible interest in the research of biomaterials. The popularity of polymers as potential biomaterials likely stems from the fact that polymers exist in a seemingly endless variety, can be easily fabricated into many forms, can be chemically modified or synthesized with chemically reactive moieties that interact with biological molecules or living tissues and cells, and can have physical properties that resemble that of natural tissues. Some disadvantages to polymeric biomaterials include relatively low moduli, instability following certain forms of sterilization, lot-to-lot variability, a lack of well-defined standards related to manufacturing, processing, and evaluating, and, for some polymers, hydrolytic instability, the need to add potentially toxic polymerization catalysts, and tissue biocompatibility of both the polymer and potential degradation byproducts. There also exist some characteristics of polymers that can be advantageous or disadvantageous depending on the application and type of polymer. Some of these characteristics include polymer degradation, chemical reactivity, polymer crystallinity, and viscoelastic behavior. Early examples of polymeric biomaterials included nylon for sutures and cellulose for kidney dialysis membranes, but more recent developments in the design of polymeric biomaterials are leading the field of biomaterials to embrace cellular and tissue interactions in order to directly induce tissue repair or regeneration.

Polymers consist of an organic backbone from which other pendant molecules extend. As their name implies, polymers consist of repeating units of one or more "mers". For example, polyethylene consists of repeating units of ethylene; nylon is comprised of repeating units of a diamine and a diacid. In general, polymers used as biomaterials are made in one of two ways: condensation or addition reactions. In condensation reactions, two precursors are combined to form larger molecules by eliminating a small molecule (e.g., water). Examples of condensation polymeric biomaterials include nylon, poly(ethylene terephthalate) (Dacron), poly(lactic acid), poly(glycolic acid), and polyurethane. In addition to synthetic polymers, biological polymers (e.g., cellulose and proteins) are formed through condensation-like polymerization mechanisms. The other major polymerization mechanism used to synthesize polymers is addition polymerization. In addition polymerization, an initiator or catalyst (e.g., free radical, heat, light, or certain ions) is used to promote a rapid polymerization reaction involving unsaturated bonds. Unlike condensation reactions, addition polymerization does not result in small molecular byproducts. Furthermore, polymers can be formed using only one type of monomer or a combination of several monomers susceptible to free radical initiation and propagation. Some examples of addition reaction polymeric biomaterials include polyethylene, poly(ethylene glycol) (PEG), poly(*N*-isopropylacrlyamide), and poly(hydroxyethyl methacrylate) (PHEMA). The chemical structure of various synthetic and natural polymers used as biomaterials are shown in Figs. 6a and b (15).

The properties of polymers are affected greatly by chemical composition and molecular weight. In general, as polymer chains become longer, their mobility decreases, but their strength and thermal stability increases. The tacticity and size of pendant chains off the backbone will affect temperature-dependent physical properties. For example, small side groups that are regularly oriented in an isotactic or syndiotactic arrangement will allow the polymer to crystallize much more readily than a polymer containing an atactic arrangement of bulky side groups. The crystalline and glass transition temperatures of polymers will affect properties (e.g., stiffness, mechanical moduli, and thermal stability) *in vivo* and will consequently influence the potential application and utility of the polymer system as a biomaterial. When the functionality of a monomer exceeds two, then the polymer will become branched upon polymerization. If a sufficient number of these high functionality monomers exist within the material, then the main chains of the polymer will become chemically cross-linked. Cross-linked polymers can be much stronger and more rigid than noncross-linked polymers. However, like linear and branched polymers, cross-linked polymers can be designed such that they degrade through hydrolytic or enzymatic mechanisms.

Due to their weaker moduli compared with that of metals or ceramics, polymers are not often used in load-bearing biomaterial applications. One exception to this observation is the example of ultrahigh molecular weight polyethylene (UHMWPE), which has a molecular weight $\sim 2{,}000{,}000\ \mathrm{g \cdot mol^{-1}}$ and has a higher modulus of elasticity than high or low density polyethylene. Additionally, UHMWPE is tough and ductile and demonstrates good wear properties and low friction. As a result, UHMWPE has been used extensively in the manufacturing of acetabular cups for total hip replacements. As an acetabular cup, UHMWPE is used in conjunction with metallic femoral stems to act as a load-bearing, low wear and friction interface. Some drawbacks to using UHMWPE include water absorption, cyclic fatigue, and a somewhat significant creep rate (4,5,9). Part of the problems surrounding UHMWPE involves its lower elastic modulus ($\sim$1–2 GPa) relative to bone ($\sim$12 GPa) and metallic implants ($\sim$100–200 GPa)

Polymers in Sutures. One of the first widespread uses of polymers as biomaterials involved sutures. In particular, polyamides and polyesters are among the most common suture materials. Nylons, an example of a polyamide, have an increased fiber strength due to a high degree of crystallinity resulting from interchain hydrogen bonding between atoms of the amide group. Nylon can be attacked by proteolytic enzymes *in vivo* and can absorb water. As a result, nylon has been used more as a short-term biomaterial. Polyester sutures, such as poly(glycolic acid), poly(lactic acid), and poly(lactic-*co*-glycolic acid) are readily degraded through hydrolytic mechanisms *in vivo*. Since one side chain of lactic acid contains a bulky hydrophobic methyl group (relative to the hydrogen side group of glycolic acid), polyesters comprised principally of lactic acid degrade at a rate slower than that of polyesters consisting mostly of glycolic acid. The degradation rate of copolymers

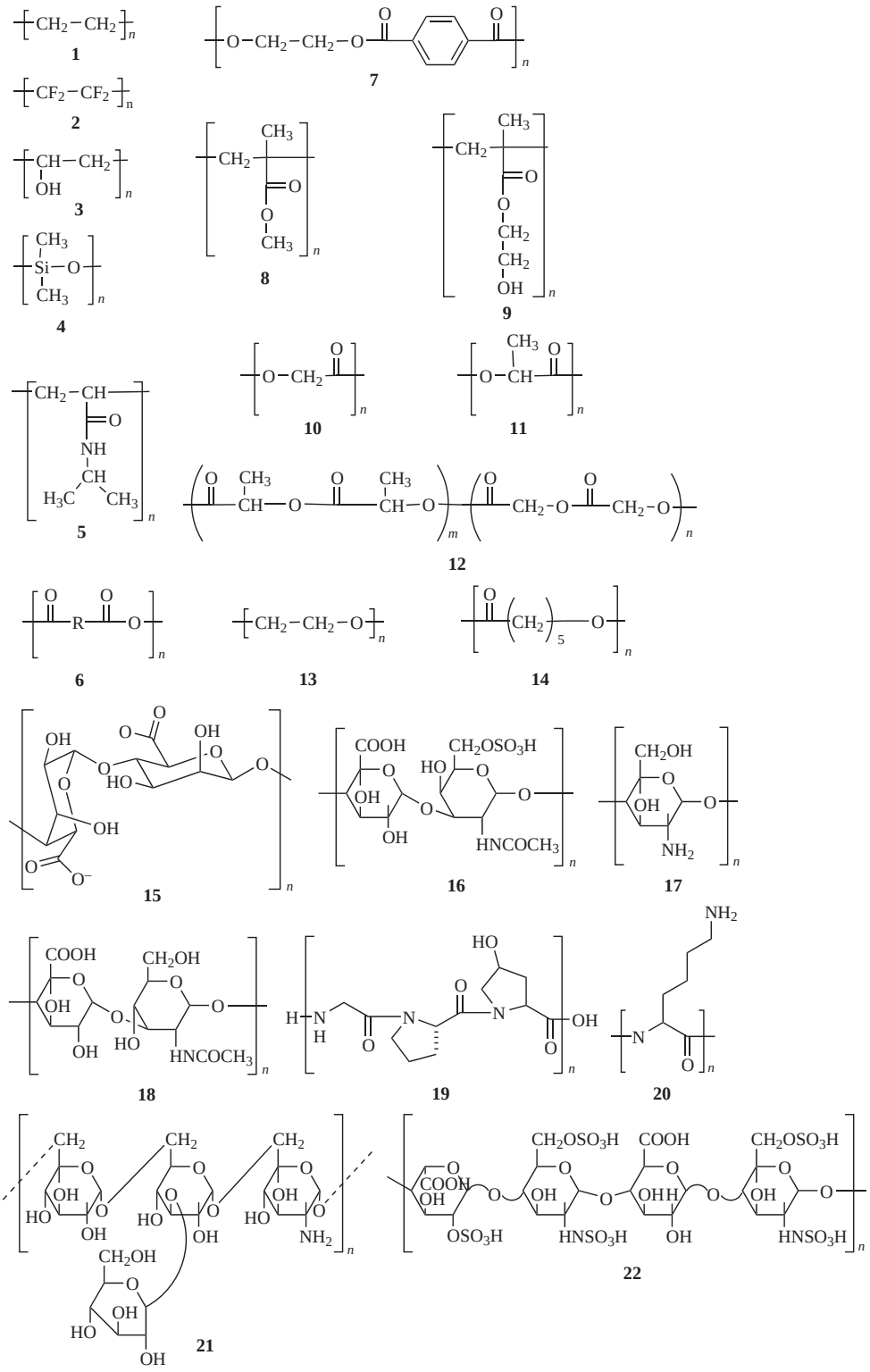


Figure 6. (a) Chemical schematics representing synthetic polymers used as biomaterials. The structures represent polyethylene (**1**), polytetrafluoroethylene (**2**), poly(vinyl alcohol) (**3**), poly(dimethyl siloxane) (**4**), poly(N-isopropylacrylamide) (**5**), polyanhydride (**6**), poly(ethylene terephthalate) (**7**), poly(methyl methacrylate) (**8**), poly(hydroxyethyl methacrylate) (**9**), poly(glycolic acid) (**10**), poly(lactic acid) (**11**), poly(lactic-co-glycolic acid) (**12**), poly(ethylene oxide) (**13**), and poly(e-caprolactone) (**14**). (Adopted from Ref. 15 with permission from Elsevier.) (b) Chemical schematics representing naturally derived polymers used as biomaterials. The structures represent alginate (**15**), chondroitin-6-sulfate (**16**), chitosan (**17**), hyaluronan (**18**), collagen (**19**), polylysine (**20**), dextran (**21**), and heparin (**22**). (Reprinted from Ref. 15 with permission from Elsevier.)

Figure 7. A photograph of a Carboflo vascular graft made out of expanded polytetrafluoroethylene (ePTFE) impregnated with carbon and marketed by Bard Peripheral Vascular, Inc.

of glycolic and lactic acid can be tailored based on the relative molar ratios of each monomer. Although the local pH of degrading polyesters can cause local inflammation concerns, the degradation byproducts of glycolic and lactic acid can be readily cleared through existing biochemical pathways. As a result, polyester sutures are commonly used within the body in applications where removal of sutures would warrant an invasive procedure (4,5,16).

Polymers in Cardiovascular Applications. Poly(ethylene terephalate) (Dacron) and expanded polytetrafluoroethylene (Teflon) have been used for decades as vascular grafts. An example of a Teflon vascular graft is shown in Fig. 7. Both of these polymers have excellent burst strengths and can be sutured directly to existing vasculature. For applications involving large diameter vascular grafts (> 6 mm), these two materials have worked well. However, neointimal hyperplasia and thrombus formation severely limit the patency of all known polymeric materials used for small diameter vascular grafts (17). Most current strategies to improve vascular graft patency involves chemically modifying the polymers used as vascular grafts to include the anticoagulant heparin, endothelial binding peptide analogues, and growth factors to stimulate endothelialization and minimize proliferation of smooth muscle into the lumen of the graft (15,18).

Polymers for Tissue Engineering. For many *in vivo* applications, researchers continue to evaluate a variety of polymeric biomaterials. Some more recent additions to the repertoire of biomaterials include naturally derived or recombinantly produced biological polymers. As an example, in the case of articular cartilage repair, it is evident that many types of polymers can be designed, modified, or combined with other materials to create new generations of biomaterials that promote healing and/or restore biological function. For example, synthetic polymers, such as poly(vinyl alcohol) (PVA), PMMA, poly(hydroxyethyl methacrylate), poly(*N*-isopropylacrylamide), polyethylene, poly(lactic acid), poly(glycolic acid), poly(lactic-*co*-glycolic acid), and poly(ethylene glycol) and naturally derived polymers (e.g., alginate, agarose, chitosan, hyaluornic acid, collagen, and fibrin) have been studied extensively with and without biochemical modifications to replace cartilage function or to promote neocartilage formation (15,19,20). These and other polymeric biomaterials have been used in studies related to liver, nerve, cardiovascular, bone, ophthalmic, skin, and pancreatic repair or restoration (15,21).

Hydrogels. As the name implies, hydrogels are polymer networks that contain large amounts of water (up to or $> 90\%$ water). As a result, hydrogels generally are hydrophilic materials, although, the presence of hydrophobic domains within the hydrogel backbone can enhance mechanical properties. To avoid dissolution into the aqueous phase, the polymeric component of the hydrogel must contain cross-links. The majority of hydrogel systems use chemical cross-links, such as covalent bonds to create a three-dimensional (3D) network; however, some hydrogels exist that rely on physical interactions to maintain gel integrity.

The high water content of hydrogels provides many benefits. First, of all the materials within materials science, the physical and mechanical properties of hydrogels most closely resemble those of biological tissue. Due to their polymeric content, hydrogels exhibit viscoelastic behavior. The elastic modulus, G', of many gel compositions reaches 1 MPa, but some hydrogels can be as strong as 20 MPa. These mechanical properties match well with those reported for many tissues. Second, the large presence of water within hydrogels can limit nonspecific interactions within the body, can shield the polymer from leukocytes and can decrease frictional effects at the site of implantation. Third, the relatively low concentration of polymer within the hydrogel can result in materials with higher porosities. Consequently, it is possible not only for cells to migrate within the hydrogel structure, but also for nutrients and waste products to diffuse into and out of the gel structure (15,22,23).

In addition to high water content, hydrogels possess other characteristics that are beneficial for biomedical applications. For example, chemical composition of polymers used in hydrogel formulations is amenable to chemical modification of the backbone and/or side group structures. These polymer derivatives allow the incorporation of various gelation chemistries, degradation rates and biologically active molecules. Although not a complete list, some of the polymers used as biomaterial hydrogels include poly(ethylene glycol), PVA, poly(hydroxyethyl methacrylate) PHEMA, poly(*N*-isopropylacrylamide), poly(vinyl pyrrolidone), dextran, alginate, chitosan, and collagen. These hydrogels, in addition to many others, are currently being explored as materials for use in cartilage, skin, liver, nerve, muscle, cardiovascular, and bone tissue engineering applications.

Poly(ethylene glycol). One of the most widely studied hydrogel materials is PEG, which contains repeats of the monomer CH_2CH_2O and exhibits a large radius of hydration due to its high hydrophilicity. As a result, PEG

can avoid detection by the body, and often is coupled to pharmaceuticals or other molecules to extend circulation half-life within the body. Of all the materials used in biomedical research, few polymers have better biocompatibility properties than PEG. Also, the chemical structure of PEG is fairly stable within aqueous environments, although hydrolytic degradation can occur. Furthermore, removal of PEG from the body is not a major concern since PEG, with a molecular weight $< 20{,}000\,g{\cdot}mol^{-1}$, can be cleared readily by the kidneys. Traditionally, PEG hydrogels have been cross-linked through chemical initiators, however, other work has shown that photoinitiators can be used to gel PEG *in situ*. Recently, more attention has focused on the use of star PEG, which contain a central core out of which proceeds several linear PEG arms. Consequently, these materials offer improved control over mechanical properties and biological interactions since each molecule of polymer contains many more potential sites for cross-linking or for incorporating biologically active molecules. Cell adhesion peptides, polysaccharides, and polysaccharide ligands have all been coupled to various PEG molecules and studied as biomaterials (15,23,24).

Acrylics. One of the greatest success stories involving polymeric biomaterials involves PMMA and PHEMA. Many polymers have not yet been approved by the FDA. However, many polymers of the acrylic family (e.g., PMMA used for bone cement and intraocular lenses) were grandfathered into the Medical Device Amendments of 1976 as approved materials. The PHEMA polymer allows for sufficient gas exchange, and both PHEMA and PMMA have excellent optical properties and a good degree of hydration. As a result, intraocular lenses, hard, and soft contact lenses (see Fig. 8) made in whole or in part from these polymers are commercially available (3,5). Even though contact lenses only touch the eye on one side, the polymers that comprise the contact lenses are still bathed in tears and are therefore subject to protein deposition. This protein deposition can cause eye irritation and lead to contact lens failure if the contacts are not properly cleaned. With the development of disposable contact lenses, however, the problem of protein buildup can be minimized since the useful lifespan of each contact lens does not have to extend very long.

Biomimetic Materials. Recently, the field of biomaterials has started to incorporate features found within mechanisms involved in biomolecular assembly and interaction (24). These biomimetic materials show great promise since assembly is directed through biological affinity, recognition and/or interactions. As a result, these materials often have properties more similar to those of natural materials. Biomimetic materials exist as polymer scaffolds and hydrogels, but can also consist of ceramic and metallic materials machined or chemically modified to mimic porous structure of tissue (e.g., bone). Further elucidation of mechanisms responsible for biological self-assembly most likely will lead to improved biomaterials that are capable of interacting very specifically with an environment containing cells, tissue, or ECM molecules. In addition, many researchers are borrowing biological concepts to provide appropriate signals for cellular proliferation or differentiation and to deliver pharmaceuticals in a much more controlled manner. The scanning electron micrograph (SEM) shown in Fig. 9 illustrates a biologically oriented approach of using a biomaterial like chitosan–collagen as a scaffold on which cells can adhere (25).

Drug Delivery. Applications involving biomaterials have evolved from those focused on mostly structural requirements to those combining multiple design considerations including structure, mechanics, degradation, and drug delivery. The latest trend in biomaterials design is to promote healing, repair, or regeneration via the delivery of pharmaceutical agents, drugs, or growth factors. There exist many examples of biomaterials used as delivery vehicles or as drugs (22); however, many of these examples

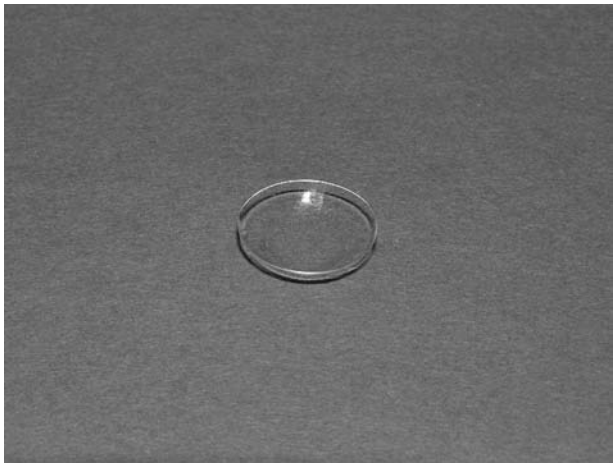


Figure 8. A photograph of a disposable contact lens made from PHEMA.

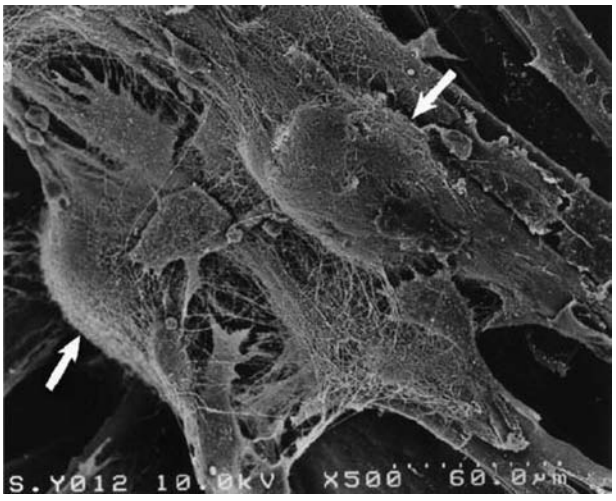


Figure 9. A SEM showing chondrocytes (denoted by white arrows) attached to a biomaterial scaffold comprised of chitosan-based hyaluoronic acid hybrid polymer fibers. (Reprinted from Ref. 25 with permission from Elsevier.)

are just beginning to transition from research materials into commercially available products. One example of a commercially available drug delivery biomaterial is known as Gliadel, made by Guilford Pharmaceuticals. Gliadel wafers consist of a polyanhydride polymer loaded with carmustine, a chemotherapeutic drug. This system is intended as a treatment for malignant gliomas. Following removal of the tumor, the Gliadel wafers are added to the cavity and allowed to degrade and release the carmustine in order to kill remaining tumor cells. In addition to cancer treatments, biomaterials as drug delivery vehicles have been extensively employed in cardiovascular applications. Recently, FDA approval was granted to several types of drug eluting metallic stents. Among these include the Sirolimus-eluting CYPHER stent manufactured by the Cordis Corporation, the TAXUS Express^2 Paclitaxel-eluting stent manufactured by the Boston Scientific Corporation. The purpose behind releasing the drugs from the stents is to decrease the occurrence of restenosis, or the renarrowing of vessels treated by the stent. As a result of the drug delivery aspect of the system, the stents are expected to have better long-term viability. Several more examples of drug delivery and biomaterial hybrid systems exist; however, a comprehensive review of biomaterials as drug delivery systems is beyond the scope of this article. It is important to note that more interest and attention have been given to modify biomaterials so that the material is more integrally involved in interacting with and manipulating organ and tissue biology.

FACTORS CONTRIBUTING TO BIOMATERIAL FAILURE

Although there exists a multitude of commercially available and successful metallic, ceramic, and polymeric biomaterials, biomaterials have and will continue to fail. The human body is a very hostile environment for synthetic and natural materials. In some instances, like orthopedic applications, it is much easier to understand why materials can fail since no material can survive cyclical loading indefinitely without showing signs of fatigue or wear. However, for most biomaterial failures, the exact reason for failure is still not well understood. Some factors contributing to the failure of a biomaterial include corrosion, wear, degradation, and biological interactions.

Corrosion

By weight, more than one-half of the human body consists of water. As a result, all implanted biomaterials will encounter an aqueous environment. Moreover, this aqueous environment is also very saline due to the presence of a relatively large concentration of extracellular salts. The aqueous and saline conditions of physiological solutions create favorable conditions for metallic corrosion. Corrosion involves oxidation and reduction reactions between a metal, ions, and species (e.g., dissolved oxygen). In fact, the lowest free energy state of many metals in and oxygenated and hydrated environment is an oxide. Most corrosion reactions are electrochemical. For example, if zinc metal is placed in an acidic environment (e.g., hydrochloric acid), hydrogen gas will evolve as the zinc become cationic and binds to chloride ions. The actual reaction consists of two half reactions. In the first reaction, zinc metal is oxidized to a Zn^{2+} state; the second reaction involves the reduction of hydrogen ions to hydrogen gas. During this process, the newly formed metal ions diffuse into solution. Both the oxidation and reduction reactions must occur at the same time to avoid charge buildup within the material. This process occurs at the surface and exposed pore of metals, and, in an attempt to passivate the surface to avoid this process, corrosion resistant oxides have been incorporated into an implant surface (13). Care must be taken, however, to ensure that the protective oxide coating is not damaged during processing, packaging, or surgical procedure.

In addition to oxidative corrosion, bimetallic or galvanic corrosion is a concern with implants composed of more than one type of metal, such as alloys with mixing defects and implants containing parts made from distinct metals. Galvanic corrosion can occur because all metals have a different tendency to corrode. If two distinct metals are in contact with one another through a conductive medium, oxidation of one metal will occur while reduction of the other occurs. In both oxidative corrosion and bimetallic corrosion, bits of metal, metal ions, and oxidative debris can enter the surrounding tissue and even travel to distant body parts. This can result in inflammation and even in metal toxicity.

Wear

In addition to corrosion, metal, as well as other materials can wear as a result of friction. For example, in hip implants, the acetabular cup is in contact with the ball of the metal or ceramic stem. Every time a movement occurs within the joint, rubbing between the ball and cup occurs and small wear particles of metal and polymer are left behind (see Fig. 10). More often than not, the particles are shed from the softer surface (e.g., ultrahigh molecular weight polyethylene); however, metal particles are also produced. The particles range in size from

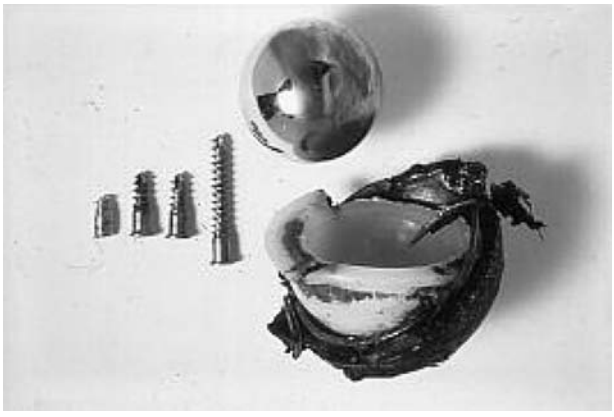


Figure 10. A photograph of some worn biomaterials. Examples in this photograph include screws, a femoral head replacement, and a polyethylene acetabular cup. (Used with permission from the Department of Materials at Queen Mary University of London.)

nanometers to microns with the smaller particles able to enter the lymph fluid and travel to distant parts of the body. The small particles increase the surface area of the material, and this increased surface area can result in increased corrosion (5,13). Thus, wear can lead to deleterious effects (e.g., corrosion), described above, and inflammation, as will be discussed below.

Degradation

Although not affected by corrosion, certain bioactive ceramics and polymers are susceptible to degradation. In the case of bioactive ceramics, however, this process is relatively slow compared with the potential rate of bone regeneration. For polymers, degradation can occur via hydrolytic or, in some cases, enzymatic mechanisms. The chemical structures of both polyamides and polyesters lend themselves toward enzymatic degradation. For polyesters, acidic or alkaline conditions will lead to a deesterification reaction that will eventually destroy the backbone of the polymer. The degradation rate varies greatly depending on the composition of the polymer. For example, within the body, poly(lactic acid) will degrade over many months to years, but poly(glycolic acid) can degrade over a few days or weeks. The degradation rate of polyamides is slower than that of polyesters, but is still an important design consideration when choosing a polymeric biomaterial for a specific application. For applications (e.g., sutures), degradation of the material is a beneficial property since the sutures only need to remain in place for a few days to weeks until the native tissue heals. For applications needing a material with a longer lifespan, degradation poses a larger problem.

Increasingly, degradable polymers or polymers with degradable cross-links are being studied as biomaterials. This interest in degradable systems stems largely from more current research involving tissue engineering and drug delivery (15,16,22,24,26,27). The philosophy of tissue engineering holds that the polymeric biomaterial acts as a scaffold with or without viable cells or biological molecules to promote tissue ingrowth. As cells proliferate and migrate within these scaffolds and begin to create new tissue, the material can and should degrade to leave, ultimately, regenerated or repaired tissue in its place. One of the engineering design constraints, therefore, is to balance the rate of degradation with that of tissue ingrowth. If the biomaterial degrades too rapidly and the newly formed tissue cannot provide the necessary mechanical support, then the biomaterial will have failed. At the same time, if the biomaterial degrades too slowly, then the process of tissue ingrowth may become inhibited or may not occur at all. To this end, more recent research has attempted to include enzymatically sensitive cross-links, usually made from synthetic peptide analogues of enzyme substrates, within polymer networks. Instead of relying upon relatively uncontrolled hydrolytic degradation, the polymeric biomaterial would degrade at a rate controlled by migrating cells. Thus, the cells themselves could degrade the material and produce new tissue in a much more controlled and physiologically relevant manner.

Biological Interactions

Most modern biomaterials are intended to come into direct contact with living tissue and biological fluids. This interaction often makes the biomaterial a target for the protective mechanisms within the body. These protective mechanisms include protein adsorption, hemostasis, inflammation and the foreign body response, and the immune response. Although it has been well established that all types of tissue-contacting biomaterials invoke some degree of biological response, it has only been during the past decade or so when investigations have revealed that all implanted tissue-containing biomaterials invoke an almost identical inflammatory and foreign body response regardless of whether the biomaterial is of metallic, ceramic, polymeric, or composite origin. Although future research in the field of biomaterials aims to better understand and to eventually mitigate the biological interactions that currently result in the failure of many biomaterials, the following biological responses remain of great importance when considering the design and potential applications of any biomaterial. In fact, most current obstacles related to the design of biomaterials involve the interaction of biomaterials with the body and the reaction of the body to biomaterials. As a result, current biomaterial research trends aim to provide an environment that allows the body to invade, remodel, and degrade the implanted material (23,27,28).

Protein Adsorption. As soon as a biomaterial comes into contact with biological fluid (e.g., blood) the material becomes coated with adsorbed proteins. This adsorption is very rapid and is based primarily on noncovalent interactions between various hydrophilic and hydrophobic domains within the adsorbed proteins and the surface of the implanted biomaterial. Initially, the composition of the protein layer depends on the relative concentration of various proteins within the biological fluid. Certain proteins (e.g., albumin) are very abundant in serum and will initially be found abundantly in the adsorbed protein layer. However, over time the adsorbed protein layer will change its composition as proteins with higher affinities for the surface of the material, but lower serum concentrations will displace proteins with lower affinities and higher serum concentrations. This rearrangement and equilibration of the protein layer is known as the Vroman effect. When biomaterials become coated with proteins, surrounding cells no longer see the surface of the material. Instead, they see a layer of serum-soluble proteins. Increasingly, biomaterials design has focused on optimizing surface chemistries and incorporating selective reactive domains that will promote a specific biological response. In reality, these engineered surfaces become masked by a nonspecific protein layer, and it is this protein layer that drives the biological response to an implanted biomaterial. Some successful examples of surface modifications aimed at reducing nonspecific protein adsorption involve the use of nonfouling hydrophilic polymers (e.g., PEG and dextran), the pretreatment of the biomaterial with a specific protein, and the replacement of certain chemically reactive functional groups with others. Time, however, remains the

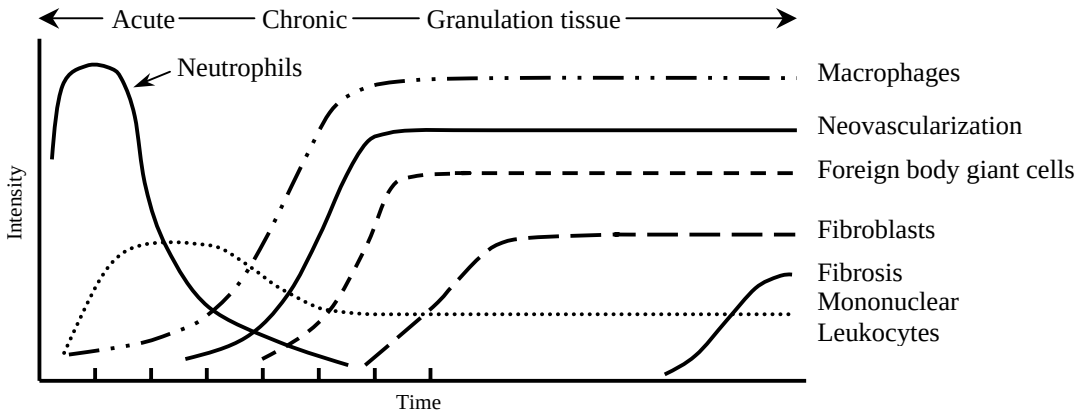


Figure 11. A schematic representing the temporal events involved in acute and chronic inflammation as well as the foreign body response. (Adopted from Anderson et al. as found in Ref. 5.)

largest obstacle with any of these surface treatments. Often, surface treatments will only function for a limited time before serum and other extracellular proteins becomes adsorbed. Once adsorbed, proteins can undergo conformational changes that expose cryptic sites, allow autoactivation, or that influence the behavior of other proteins or cells (5,6,18).

Blood Contact. Direct contact with blood is a major concern of all biomaterials regardless of whether or not they were designed for cardiovascular applications. During surgical implantation, blood vessels are broken, which results in an increased probability that the biomaterial will contact blood. Although exact mechanisms remain unclear, serum proteins (e.g., Factor XII, Factor XI, plasma prekallikrein, and high molecular weight kininogen) interact to initiate contact activation of the coagulation cascade through the intrinsic pathway. This calcium- and platelet-dependent cyclic network involves the activation of thrombin, which ultimately cleaves specific protein domains within fibrinogen and Factor XIII. Activated Factor XIIIa and fibrinogen then react to form a cross-linked fibrin clot. The formation of blood clots as well as the activation of various serum proteins and platelets can lead to local inflammation. Recent approaches have attempted to passivate the blood contact response of implanted biomaterials by incorporating heparin or other antithrombotic agents on the biomaterial surface (5,6,18).

Inflammation and the Foreign Body Response. The human body is well equipped to handle injuries that affect hemostasis. During trauma, proteins within blood can initiate a relatively large biological response lasting days, weeks, and even months. Initially, the area around a trauma site, including the implantation of a biomaterial, becomes inflamed. Inflammation is a normal process involved with healing that is characterized by four major events: swelling, pain, redness, and heat. The vasculature around an injury will become leaky to allow extravasation of various leukocytes (e.g., neutrophils and macrophages). With the presence of cytokines and other growth factors, leukocytes, primarily macrophages, are stimulated to remove bacteria and foreign material. Macrophages also recruit fibroblasts and other cells to the injury site to aid in healing by forming granulation tissue. Over the course of several days or weeks, this initial granulation tissue is remodeled and replaced with restored, functional tissue or, more commonly, scar tissue (Fig. 11).

In the case of implanted biomaterials, the implantation site is the injury site and will become inflamed. As a result, macrophages will be recruited to the site and attempt to remove the "foreign" biomaterial. Unlike smaller injuries, macrophages are unable to remove biomaterials through phagocytosis. When they become frustrated, macrophages will fuse together to form foreign body giant cells. These foreign body giant cells can secrete superoxides and free radicals, which can damage biomaterials, but these cells usually cannot completely remove the foreign biomaterial. In the event that the body cannot eliminate a foreign object through phagocytosis, activated macrophages and foreign body giant cells remain around the implant and can promote a chronic localized area of inflammation. Remaining fibroblasts and other cells around the biomaterial then will begin to secrete a layer of avascular collagen around the biomaterial to effectively encapsulate it and wall it off from the rest of the body (5,6,18). Although the function of some biomaterials is not affected by this foreign body response, biomaterials ranging from sensors to orthopedic implants to soft tissue replacements are adversely affected by this biological reaction. To date, it is not known how to minimize or eliminate an inflammation or foreign body reaction. However, a great deal of research is attempting to create biomaterials that do not evoke a tremendous inflammatory response or that degrade in a way that allows the restoration or repair of native tissue without the adverse affects of chronic inflammation.

Immune Response. The innate and adaptive immune responses of the body also pose a challenge for biomaterials designed for long-term applications. Increasingly, new biomaterials have attempted to incorporate cellular components in an attempt to create new tissues *in vitro* or to seed materials with autologous, allogeneic, or xenogenic cells, including stem cells, to promote tissue repair. Unfortunately, the adaptive immune response will actively eliminate allogeneic or xenogenic cell types. As a result,

biomaterials have been designed to act as barriers that limit lymphocyte activation. Often, cells are encapsulated in microspheres made from various polymers or layers of polymers. For example, pancreatic Islets of Langerhans from animal and human donors have been encapsulated within polymers [e.g., polysulfones, poly(*N*-isopropylacrylamide)] and alginates, to provide an immunoisolated environment that still retains enough permeability to allow for the diffusion of insulin. One of the major complications of this type of biomaterials design is to balance the creation of volume within the microsphere to accommodate enough Islets to allow for sufficient insulin production with the need to provide appropriate diffusion rates so that the cells within the center of the microsphere remain viable. As more polymeric biomaterials incorporate or consist of peptide and protein motifs, there remains a concern as to whether or not these motifs might elicit an adaptive immune response. Even if protein domains derived from human proteins are incorporated into biomaterials, these domains might not be presented the same way to lymphocytes. As a result, the body may start producing antibodies against these domains, which might also lead to certain forms of autoimmune diseases (29).

Although the adaptive immune system is playing an increasingly important role in the rejection of new types of biomaterials, the innate immune system remains a very large threat to the success of a biomaterial. As mentioned above, proteins bind to biomaterials upon implantation. One of the most abundant proteins within the blood is the complement protein C3. Within the blood, C3 can spontaneously hydrolyze to form an active convertase complex, which can cleave C3 into C3a and C3b. Although C3b is rapidly inactivated within the blood, it can remain active if it binds to a surface (e.g., a biomaterial). As a result, the alternative pathway of the complement system can be activated very rapidly leading to formation of membrane-attack complexes but more importantly, the formation of the soluble anaphylotoxins C3a, C4a, and C5a. These anaphylotoxins induce smooth muscle contraction, increase vascular permeability, recruit phagocytic cells, and promote opsonization by phagocytic cells. These phagocytic cells (e.g., macrophages) have receptors recognizing C3b. As a result, macrophages will attempt to engulf the C3b-coated biomaterial. When this fails, the macrophages will form foreign body giant cells, and the body will attempt to encapsulate the biomaterial in a manner similar to that described above for the inflammation and foreign body response (29). Overall, all of the above mentioned biological responses can affect the performance of any biomaterial, and active biomaterials research is striving not only to better understand the mechanisms of inflammation, protein adsorption, hemostasis, and innate and adaptive immune responses, but also to develop strategies to minimize, eliminate, evade, or alter adverse biological responses to materials.

BIOCOMPATIBILITY

Since biomaterials are intended for direct contact with biologically viable tissue, all biomaterials need to possess some degree of biocompatibility. In a manner similar to that of the term biomaterials, the term biocompatibility has experienced many changing definitions over the past several decades. Initially, biocompatibility implied that the biomaterial remained inert to its surroundings in order to refrain from being toxic, carcinogenic, or allergenic. As the definition of biomaterials evolved to include biologically derived materials and molecules, the term biocompatibility needed to encompass these changes. In 1987, David Williams suggested that biocompatibility is "the ability of a material to perform with an appropriate host response in a specific application" (30). Although there does not yet exist a universal consensus with regard to the definition of the term biocompatibility, the definition proposed by Williams provides enough generality to serve as an adequate and accurate description of biocompatibility.

Instead of remaining inert, biomaterials are becoming increasingly reliant on biochemical reactions and physiological processes in order to serve a useful function. In some cases (e.g., in the case of bone plates and artificial joints), biomaterials can remain inert and still provide satisfactory performance. In other instances (e.g., drug delivery vehicles), tissue engineering applications, and *in vivo* organ replacement therapies, biomaterials not only need to actively minimize or adapt to the surrounding biological responses (e.g., inflammation and foreign body responses), but also need to depend on interactions with surrounding tissues and cells in order to provide a useful function (15,22,24–27). In addition, the performance of traditionally inert biomaterials is being enhanced by incorporating chemical or mechanical modifications that interact with biology at the cellular level. For example, the bone-contacting surfaces of metallic femoral stems, for hip replacement, have been modified to contain bioactive ceramic porous networks or hydroxyapatite crystal networks. These ceramic networks allow better osteointegration of the implant with the host tissue and, in some cases, eliminate the need to use bone sealants (e.g., PMMA).

Obviously, if a successful biomaterial needs to show some level of biocompatibility, then there must exist various testing conditions and manufacturing standards to establish safety controls. Organizations [e.g., ASTM International and the International Organization for Standardization (ISO)] do have guidelines and standards for the testing and evaluation of biomaterial biocompatibility. These regulations include tests include the measuring of cytotoxicity, sensitization, skin irritation, intracutaneous reactivity, acute systemic toxicity, genotoxicity, macroscopic and microscopic evaluation of implanted materials and devices, hemocompatibility, subchronic and chronic toxicity, carcinogenicity, the effect of degradation byproducts, and the effect of sterilization (31). For many of these paramenters, the associated standards dictate the size and shape of the material to be tested, appropriate *in vitro* testing procedures and analysis schemes, and relevant testing and evaluation protocols for *in vivo* experimentation. Although standards related to the manufacturing and performance of some biomaterials exist, there remains a lack of uniform biocompatibility testing standards for new classes of biomaterials that rely heavily upon cellular and tissue interactions or that contain biologically active

molecules. New developments in biologically active biomaterials have resulted in not only nonuniform approaches to biocompatibility testing, but also confusions related to the regulatory classification of new types of biomaterials.

FUTURE DIRECTIONS

As more information becomes available regarding biological responses to materials, mechanisms that control embryonic development and early wound healing, and matrix biology, materials will be designed to more adequately address, promote or inhibit biological responses as needed. As a result, the field of biomaterials will not only incorporate principles from materials science and engineering, but also rely increasingly upon design constraints governed by biology (see Fig. 12). Recent trends in biomaterial research show an increased emphasis in designing materials that better match the biological environment with respect to mechanics and biological signals. Materials promote cell attachment using biologically derived signals, degrade through relevant enzymatic degradation and release and store bioactive factors using methods derived from biology. Continued adaptation of materials to more appropriately interact with the living system will result in devices that work with the body to promote tissue regeneration and healing.

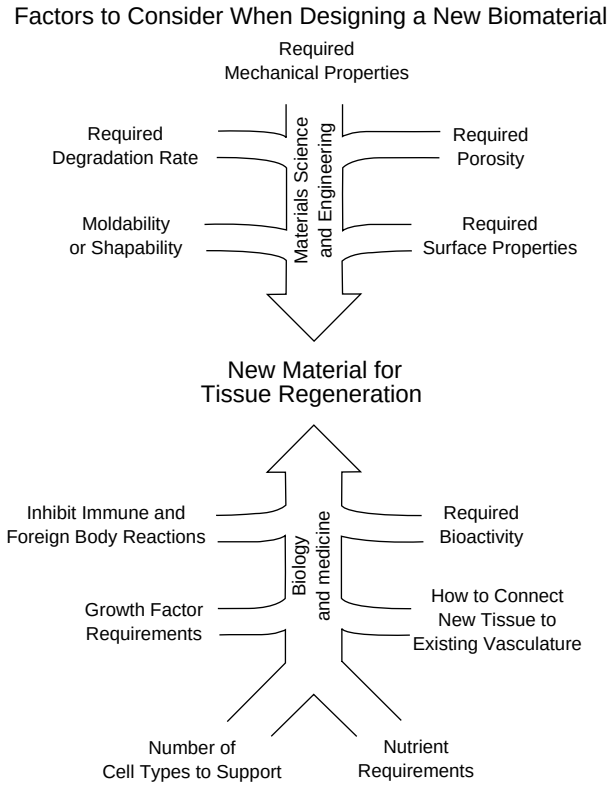


Figure 12. An illustration depicting the various engineering and biological factors that need to be considered in the design of modern biomaterials. Successful, new biomaterials will require the optimization of a variety of parameters and the cooperation of interdisciplinary scientists, engineers, and clinicians. (Reprinted from Ref. 15 with permission from Elsevier.)

BIBLIOGRAPHY

Cited References

1. Black J. The education of the bio-materialist—report of a survey. 1980–1981. J Biomed Mater Res 1982;16(2):159–167.
2. Galletti PM, Boretos JW. Report on the consensus development conference on clinical-applications of biomaterials, 1–3 november 1983. J Biomed Mater Res 1983;17(3):539–555.
3. Lindstrom RL. The polymethylmetacrylate (pmma) intraocular lenses. In: Steinert RF, et al. editors. Cataract surgery: Technique, Complications, and Management. Philadelphia: Saunders; 2003.
4. Barbucci R, editor. Integrated Biomaterials Science. New York: Kluwer Academic; 2002.
5. Ratner BD, Hoffman AS, Schoen FJ, Lemons JE, editors. Biomaterials Science: An Introduction to Materials in Medicine. San Diego: Academic Press; 1996.
6. Greco RS, editor. Implantation Biology: The Host Response and Biomedical Devices. Boca Raton (FL): CRC Press; 1994.
7. Lysaght MJ, O'Loughlin JA. Demographic scope and economic magnitude of contemporary organ replacement therapies. Asaio J 2000;46(5):515–521.
8. Harris G. Step forward for implants with silicone. New York: The New York Times; 2005.
9. Park JB, Lakes RS. Biomaterials: An introduction. 2nd ed. New York: Plenum Press; 1992.
10. Black J, Hastings G, editors. Handbood of Biomaterial Properties. New York: Chapman & Hall; 1998.
11. Wen CE, et al. Novel titanium foam for bone tissue engineering. J Mater Res 2002;17(10):2633–2639.
12. Graham, et al. The use of SMARTeR stents in patients with binary obstrevition. Cliro Radiol 2004;59:288–291.
13. Smith WF, Foundations of Materials Science and Engineering. 2nd ed. New York: McGraw-Hill, Inc.; 1993.
14. Fernandez E, et al. Calcium phosphate bone cements for clinical applications—part ii: Precipitate formation during setting reactions. J Mater Sci-Mater M 1999;10(3):177–183.
15. Seal BL, Otero TC, Panitch A. Polymeric biomaterials for tissue and organ regeneration. Mater Sci Eng R 2001; 34(4–5):147–230.
16. Langer R. 1994 Whitaker lecture—polymers for drug-delivery and tissue engineering. Ann Biomed Eng 1995;23(2): 101–111.
17. Greenwald SE, Berry CL. Improving vascular grafts: The importance of mechanical and haemodynamic properties. J Pathol 2000;190(3):292–299.
18. Dee KC, Puleo DA, Bizios R. An Introduction to Tissue-Biomaterial Interactions. Hoboken (NJ): John Wiley & Sons, Inc.; 2002.
19. Temenoff JS, Mikos AG. Review: Tissue engineering for regeneration of articular cartilage. Biomaterials 2000;21 (5):431–440.
20. Temenoff JS, Mikos AG. Injectable biodegradable materials for orthopedic tissue engineering. Biomaterials 2000;21(23): 2405–2412.
21. Palsson BØ, Bhatia SN. Tissue Engineering. Upper Saddle River (NJ): Pearson Prentice Hall; 2004.
22. Peppas NA, Bures P, Leobandung W, Ichikawa H. Hydrogels in pharmaceutical formulations. Eur J Pharm Biopharm 2000;50(1):27–46.
23. Hubbell JA. Biomaterials in tissue engineering. Bio-Technol 1995;13(6):565–576.
24. Sakiyama-Elbert SE, Hubbell JA. Functional biomaterials: Design of novel biomaterials. Ann Rev Mater Res 2001;31: 183–201.

25. Yamane S, et al. Feasebelety of chitosan-based hyaluoronic acid hybrid biomaterial for a novel scaffold in cartilage tissue engineering. Biomaterials 2005;26:611–619.
26. Gupta P, Vermani K, Garg S. Hydrogels: From controlled release to ph-responsive drug delivery. Drug Discov Today. 2002;7(10):569–579.
27. Ratner BD, Bryant SJ. Biomaterials: Where we have been and where we are going. Annu Rev Biomed Eng 2004;6:41–75.
28. Ratner BD. New ideas in biomaterials science—a path to engineered biomaterials. J Biomed Mater Res 1993;27(7): 837–850.
29. Janeway CA, Travers P, Walport M, Shlomchik MJ. Immunobiology: The Immune System in Health and Disease. 6th ed. New York: Garland Science; 2005.
30. Williams D. Revisiting the definition of biocompatibility. Med Device Technol 2003;14(8):10–13.
31. Bollen LS, Svensen O. Regulatory guidelines for biocompatibility safety testing. Med Plastics Biomater 1997;(May): 16.

See also ALLOYS, SHAPE MEMORY; BIOMATERIALS: BIOCERAMICS; BIOMATERIALS, CORROSION AND WEAR OF; BIOMATERIALS, TESTING AND STRUCTURAL PROPERTIES OF; POLYMERIC MATERIALS.

BIOMATERIALS: BIOCERAMICS

JULIAN R. JONES
LARRY L. HENCH
Imperial College London

INTRODUCTION

During the last century, there has been a revolution in orthopedics with a shift in emphasis from palliative treatment of infection in bone to interventional treatment of chronic age-related ailments. The evolution of stable metallic fixation devices, and the systematic development of reliable total joint prostheses were critical to this revolution in health care. Two alternative pathways of treatment for patients with chronic bone and joint defects are now possible: (*1*) transplantation or (*2*) implantation. Figure 1 shows how approaches to tissue repair have changed and how we think they need to develop.

At present the "gold standard" for the clinical repair of large bone defects is the harvesting of the patient's tissue from a donor site and transplanting it to a host site, often maintaining blood supply. This type of tissue graft (an *autograft*) has limitations; limited availability, morbidity at the donor site, tendency toward resorption, and a compromise in biomechanical properties compared to the host tissue.

A partial solution to some of these limitations is use of transplant tissue from a human donor, a *homograft*, either as a living transplant (heart, heart-lung, kidney, liver, retina) or from cadavers (freeze-dried bone). Availability, the requirement for lifetime use of immunosuppressant drugs, the concern for viral or prion contamination, ethical, and religious concerns all limit the use of homografts.

The first organ transplant (homograft) was carried out in Harvard in 1954. In the United States alone, there are now $>$80,000 organs needed for transplantation at one time, only a quarter of which will be found. The shortage of donors increases every year.

A third option for tissue replacement is provided by transplants (living or nonliving) from other species called *heterografts* or *xenografts*. Nonliving, chemically treated xenografts are routinely used as heart valve replacements (porcine) with $\sim$50% survivability after 10 years. Bovine bone grafts are still in use, but concern of transmission of prions (disease transmission) is growing.

The second line of attack in the revolution to replace tissues was the development of manmade materials to interface with living, host tissues (e.g., implants or prostheses made from biomaterials). There are important advantages of implants over transplants, including

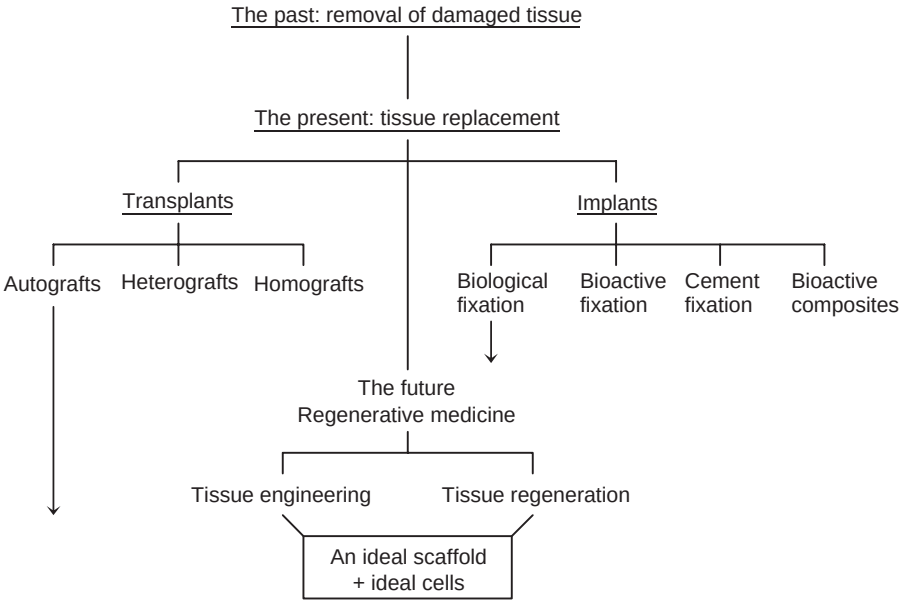


Figure 1. Schematic of the past, present and future of the treatment for diseased and damaged tissue.

availability, reproducibility, and reliability. Failure rate of the materials used in most prostheses are very low, at $< 0.01\%$ (1). As a result, survivability of orthopedic implants such as the Charnley low friction metal–polyethylene total hip replacement is very high up to 15 years (2).

Many implants in use today continue to suffer from problems of interfacial stability with host tissues, biomechanical mismatch of elastic moduli, production of wear debris, and maintenance of a stable blood supply. These problems lead to accelerated wear rates, loosening and fracture of the bone, interface or device that become worse as the patient ages (3). Repair of failed devices, called revision surgery, also becomes more difficult as the patient ages due to decreased quality of bone, reduced mobility, and poorer circulation of blood. In addition, all present day orthopedic implants lack two of the most critical characteristics of living tissues: (*1*) ability to self-repair; and (*2*) ability to modify their structure and properties in response to environmental factors such as mechanical load. The consequences of these limitations are profound. All implants have limited lifetimes. Many years of research and development have led to only marginal improvements in the survivability of orthopedic implants for >15years.

Ideally, artifical implants or devices should be designed to stimulate and guide cells in the body to regenerate tissues and organs to a healthy and natural state. We need to shift our thinking toward regenerative medicine (Fig. 1) (4).

BIOCERAMICS AS MEDICAL DEVICES

A bioceramic is a ceramic that can be implanted into a patient without causing a toxic response. Bioceramics can be classified into three categories; resorbable (e.g., tricalcium phosphate), bioactive (e.g., bioactive glass, hydroxyapatite), and nearly inert materials (e.g., alumina and zirconia) (5). A bioactive material is defined as a material that elicits a specific biological response at the interface of the material, which results in a formation of a bond between the tissue and that material (6). Bioceramics can be polycrystalline (alumina or hydroxyapatite), bioactive glass, bioactive glass–ceramic (apatite/ wollastonite, A/W), or used in bioactive composites such as polyethylene–hydroxyapatite.

This article begins with examples of successful bioceramics used clinically that improve the length and quality of life for patients. Developments of porous bioceramic and composite scaffolds for tissue engineering applications are then discussed. The article ends by discussing how bioactive ceramics may be the future of regenerative medicine due to their potential for guiding tissue regeneration by stimulating cells at the genetic level (Fig. 1).

NEARLY INERT BIOCERAMICS

High density, high purity α-alumina (Al_2O_3) was the first bioceramic widely used clinically, as the articulating surfaces of the ball and socket joints of total hip replacements because of its combination of low friction, high wear resistance, excellent corrosion resistance, good biocompatibility, and high strength (7). The physical properties of alumina depend on the grain size. Medical grade alumina exhibits an average grain size <4 μm, a compressive strength of 4.5 GPa and a Young's modulus of 400 GPa (8). Other clinical applications of Al_2O_3 include bone screws, jaw and maxillofacial reconstructions, middle ear bone substitutions, and dental implants.

Zirconia is a bioinert ceramic that has higher flexural strength and fracture toughness and a lower Young's modulus than alumina. Zirconia may therefore be suitable for bearing surfaces in total hip prostheses, however, there are concerns over the wear rate and radioactivity of the material in the body (9).

When an almost inert implant is implanted into soft or hard tissue, a nonadherent fibrous capsule surrounds the implant. If the implant is loaded such that interfacial movement can occur, the fibrous capsule can become several hundred micrometers thick and cause loosening of the implant, which will eventually lead to clinical failure (8).

An improved interface between nearly inert implants and tissue can be achieved by using an implant containing pores in excess of 100 μm in diameter. The fibrous connective (scar) tissue grows into these pores and anchors the implant in place. This technique is termed "biological fixation" (10). Viable bone requires pores of >200 μm. However, connective tissue still allows some movement of the prothesis, which will increase with age and cause bone resorption.

THE CHALLENGE FOR BIOCERAMICS

Bone is a natural composite of collagen (type I) fibers, noncollagenous proteins, and mineralized bone. It is a rigid material that exhibits a hierarchical structure with an outer layer of dense cortical bone and an internal structure of porous cancellous and trabecular (spongy) bone. Trabecular bone is orientated spongy bone that is found at the end of long bones and in vertebras (11). This structure provides excellent mechanical properties: cortical bone exhibits a compressive strength of 100–230 MPa and a Young's modulus of 7–30 GPa; cancellous bone exhibits a compressive strength of 2–12 MPa and a Young's modulus of 0.05–0.5 GPa (8). Bone is generated by cells called osteoblasts and resorbed by cells called osteoclasts, which remodel the bone in response to external stimuli such as mechanical load (11). In order to regenerate bone, the implant should exhibit a Young's modulus similar to that of the bone. If the modulus of the implant is higher than the bone then stress shielding can occur, where the stem supports the total load. If this occurs, osteoclasts resorb bone from the implant interface (12). An example of this is the use of alumina as in total hip replacements. Alumina exhibits a modulus 10–50 times higher than cortical bone. If the modulus of the implant is substantially lower than the bone, the implant is unlikely to be able to withstand the loading environment and will fracture.

Ceramics have the potential to prevent stress-shielding and have many properties that can aid bone regeneration (8). Therefore, we will concentrate on the use of bioceramics

in orthopedics, but will also describe adaption for soft tissue applications.

Osteoporosis is a condition where the density and strength of the trabecular bone decreases (13), due to osteoblasts becoming progressively less active and the pore walls (trabeculae) in the internal spongy bone are reduced in thickness and number causing spinal problems, hip fracture, and subsequent hip replacement operations.

The challenge for bioceramics is to replace old, deteriorating bone with a material that can function for as long as is needed, which may be > 20 years. There are two options to satisfy increasing needs for orthopedic repair in the new millennium: (*1*) improve implant survivability by 10–20 years; or (*2*) develop alternative devices that can regenerate tissues to their natural form and function. Decades of research have not been able to achieve the first, discussion of the second, the application of bioactive bioceramics, and their role in regenerative medicine, particularly in bone regeneration follows.

RESORBABLE BIOCERAMICS

Tricalcium phosphate (TCP) resorbs on contact with body fluid. Resorbable materials are designed to dissolve at the same rate that a tissue grows, so that they eventually are totally replaced by the natural host tissue. However, matching the resorption rate of TCP with bone growth is difficult and since TCP ceramics exhibit low mechanical strength, so they cannot be used in load bearing applications (14).

THE BIOACTIVE ALTERNATIVE

During the last decade, considerable attention has been directed toward the use of bioceramic implants with bioactive fixation, where bioactive fixation is defined as interfacial bonding of an implant to tissue by means of formation of a biologically active hydroxyapatite layer on the surface of the implant. This layer bonds to the biological apatite in bone (8).

An important advantage of bioactive fixation is that a bioactive bond forms at the implant–bone interface with a strength equal to, or greater than, bone after 3–6 months. The level of bioactivity of a specific material can be related to the time taken for $> 50\%$ of the interface to bond to bone ($t_{0.5bb}$) (15);

$$\text{Bioactivity index}, I_B = 100/t_{0.5bb} \quad (1)$$

Materials exhibiting an I_B value > 8 (class A), will bond to both soft and hard tissue. Materials with an I_B value < 8 (class B), but > 0, will bond only to bone. Biological fixation is capable of withstanding more-complex stress states than implants that only achieve morphological fixation, that is surface fixation to roughness (15). There are a number of bioactive bioceramics.

SYNTHETIC HYDROXYAPATITE

Synthetic hydroxyapatite (HA) $Ca_{10}(PO_4)_6(OH)_2$ has been developed to match the biological apatite in bone. Biological apatite, although similar, exhibits different stoichiometry, composition, and crystallinity to pure HA. Biological apatites are usually calcium deficient and carbonate substituted (primarily for phosphate groups) (16). Hydroxyapatite is a class B bioactive material, that is, it bonds only to bone and promotes bone growth along its surface (osteoconduction). The mechanism for bone bonding involves the development of a cellular bone matrix on the surface of the HA, producing an electron dense band 3–5 μm wide. Collagen bundles appear between this area and the cells. On contact with body fluid, a dissolution–precipitation process occurs at the HA surface resulting in the formation of carbonated apatite microcrystals, which are similar to biological HA and are incorporated into the collagen. As the site matures, collagen fibrils mineralize and the interfacial layer decreases in thickness as crystallites of the growing bone align with those of the implant. Commercial production methods for synthetic HA usually involve a dropwise addition of phosphoric acid to a stirring suspension of calcium hydroxide in water, which causes an apatite precipitate to form. Ammonia is added to keep the pH very alkaline and to ensure formation of HA when the precipitate is sintered at 1250 °C. Commercial dense HA exhibits a compressive strength in excess of 400 MPa and a Young's modulus of 12 GPa. There are many clinical applications for HA implants including the repair of bony defects and tooth root replacement (16).

Hydroxyapatite has also been used as a plasma-sprayed coating on porous metallic implants in total hip replacements, allowing a bond to form between the bone and the implant (17). Initial bone ingrowth is more rapid than uncoated porous metallic implants, but the long-term survivability of the implants will not be known until after 10-year follow-up clinical trails have been completed.

BIOACTIVE GLASSES

Bioactive glasses are class A bioactive materials (I_B value > 8) that bond to soft and hard tissue and are osteoconductive, which means that bioactive glass implants stimulate bone formation on the implant away from the host bone–implant interface (15). Bioactive glasses undergo surface dissolution in a physiological environment in order to form a hydroxycarbonate apatite (HCA) layer. This is very similar to the carbonate-substituted apatite in bone. The higher the solubility of a bioactive glass, the more pronounced is the effect of bone tissue growth. The structures of bioactive glasses are based on a cross-linked silica network modified by cations. The original bioactive glasses were developed by Hench and colleagues in the early 1970s (18) and were produced using conventional glass melt-processing techniques with a composition of 45S5, 46.1% SiO_2, 24.4% NaO, 26.9% CaO, and 2.6% P_2O_5, in mol percent. This composition was given the name Bioglass.

MECHANISM OF BIOACTIVITY OF BIOACTIVE GLASSES

When a glass reacts with an aqueous solution, both chemical and structural kinetic changes occur as a function of time within the glass surface (8). Accumulation of

dissolution products causes both the chemical composition and pH of solution to change. The formation of HCA on bioactive glasses and the release of soluble silica to the surrounding tissue are key factors in the rapid bonding of these glasses to tissue and the stimulation of tissue growth.

There are 11 stages in process of complete bonding of bioactive glass to bone. Stages 1–5 are chemical; stages 6–11 are biological (4,15);

1. Rapid exchange of Na^+ and Ca^{2+} with H^+ or H_3O^+ from solution (diffusion controlled with a $t^{1/2}$ dependence, causing hydrolysis of the silica groups, which creates silanols;

$$Si-O-Na^+ + H^+ + OH^- \rightarrow Si-OH^+ + Na^+_{(aq)} + OH^-$$

   The pH of the solution increases as a result of H^+ ions in the solution being replaced by cations.
2. The cation exchange increases the hydoxyl concentration of the solution, which leads to attack of the silica glass network. Soluble silica is lost in the form of $Si(OH)_4$ to the solution, resulting from the breaking of Si–O–Si bonds and the continued formation of Si–OH (silanols) at the glass solution interface:

$$Si-O-Si + H_2O \rightarrow Si-OH + OH-Si$$

   This stage is an interface-controlled reaction with a $t^{1.0}$ dependance.
3. Condensation and repolymerization of a SiO_2-rich layer on the surface, depleted in alkalis and alkali earth cations:
4. Migration of Ca^{2+} and $PO_4{}^{3-}$ groups to the surface through the SiO_2-rich layer, forming a $CaO–P_2O_5$-rich film on top of the SiO_2-rich layer, followed by growth of the amorphous $CaO–P_2O_5$-rich film by incorporation of soluble calcium and phosphates from solution.
5. Crystallization of the amorphous $CaO–P_2O_5$ film by incorporation of OH^- and $CO_3{}^{2-}$ anions from solution to form a mixed-HCA layer.
6. Adsorption and desorption of biological growth factors, in the HCA layer (continues throughout the process) to activate differentiation of stem cells.
7. Action of macrophages to remove debris from the site.
8. Attachment of stem cells on the bioactive surface.
9. Differentiation of stem cells to form bone growing cells, such as osteoblasts.
10. Generation of extra cellular matrix by the osteoblasts to form bone.
11. Crystallization of inorganic calcium phosphate matrix to enclose bone cells in a living composite structure.

Interfacial bonding occurs with bone because of the biological equivalence of the inorganic portion of bone and the growing HCA layer on the bioactive implant. For soft tissues, the collagen fibrils are chemisorbed on the porous SiO_2-rich layer via electrostatic, ionic and/or hydrogen bonding, and HCA is precipitated and crystallized on the collagen fiber and glass surfaces.

Reaction stages one and two are responsible for the dissolution of a bioactive glass, and therefore greatly influence the rate of HCA formation. Studies have shown that the leaching of silicon and sodium to solution, from melt-derived bioactive glasses, is initially rapid, following a parabolic relationship with time for the first 6 h of reaction, then stabilizes, following a linear dependence on time, which agree with the dissolution kinetics of soda lime–silica glasses:

$$Q = Kt^{\gamma} \quad \text{for total diffusion, or more generally} \tag{2}$$

$$Q = at^{\gamma} + bt \quad \text{for total diffusion and selective leaching} \tag{3}$$

where Q is the quantity of alkali ions from the glass, t is the duration of experiment, a,b are empirically determined constants, K is the reaction rate constant, assuming constant glass area and temperature, and $\gamma = 1/2$ (for stage 1) or 1 (for stage 2); as $t \rightarrow 0$ $\gamma = 1/2$, as $t \rightarrow \infty$ $\gamma = 1$ (19).

Phosphorous and calcium contents of the solution follow a similar parabolic trend over the first few hours, after which they decrease, corresponding with the formation of the Ca–P-rich film (stage 4). The pH change of the solution mirrors dissolution rates. An initial rapid increase of pH is a result of ion exchange of cations such as Na^+ from the glass with H^+ from solution during the first minutes of reaction at the bioactive glass surface. As release rate of cations decreases, the solution pH value tends toward a constant value.

For bioactive implants, it is necessary to control the solubility (dissolution rate) of the material. A low solubility material is needed if the implant is designed to have a long life, for example, a coating on orthopedic metals, such as synthetic hydroxyapatite on a titanium alloy femoral stem. A high solubility implant is required if it is designed to aid bone formation, such as 45S5 Bioglass powders for bone graft augmentation. A fundamental understanding of factors influencing solubility and bioreactivity is required when developing new materials for *in situ* tissue regeneration and tissue engineering.

FACTORS AFFECTING THE DISSOLUTION AND BIOACTIVITY OF GLASSES

Many factors affect the dissolution rate, and therefore bioactivity of bioactive glasses. The composition, initial pH, ionic concentration, and temperature of the aqueous environment have a large effect on the dissolution of the glass. The presence of proteins in the solution has been found to reduce dissolution rates due to the adsorption of serum proteins onto the surface of the bioactive glass, which form a barrier to nucleation of the HCA layer (19).

A change in geometry and surface texture of an implant will generally mean a change in the surface area/solution volume ratio (SA/V). An increase in the SA/V generally causes an increase in the dissolution rate, as the amount of surface exposed to solution for ion exchange increases. An increase in SA/V can be caused by a decrease in particle size or by an increase in surface roughness or porosity (19). A

similar effect occurs if the volume of surrounding solution increases (20).

If silicate glasses are considered to be inorganic polymers of silicon cross-linked by oxygen, the network connectivity is defined as the average number of cross-linking bonds for elements other than oxygen that form the backbone of a silicate glass network. The network connectivity can be used to predict solubility (21). Calculation of network connectivity is based on the relative numbers of network forming oxide species (those that contribute to cross-linking or bridging) and network-modifiers (nonbridging) present. Silicate structural units in a glass of low network connectivity are probably of low molecular mass and capable of going into solution. Consequently, glass solubility increases as network connectivity is reduced. The network connectivity can be used to predict bioactivity. Crystallization of a glass inhibits its bioactive properties, because ion exchange is inhibited by crystalline phases, and interferes with network connectivity.

Slight deviations in glass composition can radically alter the dissolution kinetics and the basic mechanisms of bonding. It is widely accepted that increasing silica content of melt-derived glass decreases dissolution rates by reducing the availability of modifier ions such as Ca^{2+} and HPO^{4-} to the solution and the inhibiting development of a silica-gel layer on the surface. The result is the reduction and eventual elimination of the bioactivity of the melt-derived bioactive glasses as the silica content approaches 60%. The addition of multivalent cations, such as alumina, stabilizes the glass structure by eliminating nonbridging oxygen bonds reducing the rate of break-up of the silica network and reducing the rate of HCA formation. Melt-derived glasses with >60 mol% silica are not bioactive. In order to obtain bioactivity at silica levels >60 mol%, the sol–gel process is used, which is a novel processing technique for the synthesis of tertiary bioactive glasses.

CLINICAL APPLICATIONS OF MELT-DERIVED BIOACTIVE GLASSES

Bioactive glasses have been used for >15 years to replace the small bones of the middle ear (ossicles) damaged by chronic infection (22). The glass bonds to the soft tissue of the eardrum and to the remaining bone of the stapes footplate, anchoring both ends of the implant without the formation of fibrous (scar) tissue.

In 1993, particulate bioactive glass, 45S5 Bioglass was cleared in the United States for clinical use as a bone graft material for the repair of periodontal osseous defects (Perioglas, USBiomaterials Alachua, Florida). The glass powder is inserted into the cavities in the bone between the tooth and the periodontal membrane and the tooth, which have eroded due to periodontitis. New bone is rapidly formed around the particles restoring the anchorage of the tooth in place (23). Since 1993, numerous oral and maxillofacial clinical studies have been conducted to expand the use of this material. More than 2,000,000 reconstructive surgeries in the jaw have been done using Perioglas. The same material has been used by several orthopedic surgeons to fill a variety of osseous defects and for clinical use in orthopedics, now termed NovaBone, it is now approved for clinical use worldwide.

GLASS-CERAMICS

Bioactive glasses and sintered HA do not have mechanical properties as high as that of cortical bone. Kokubo et al. (24) developed dense apatite/ wollastonite (A/W) glass-ceramics by heating crushed quenched melt-derived glass (MgO 4.6, CaO 44.7, SiO_2 34.0, P_2O_5 6.2, CaF_2 0.5 wt%) to 1050 °C at a rate of 5 °C · min. Oxyapatite (38 wt%) and β-wollastonite (34 wt%) precipitated and were homogeneously dispersed in a glassy matrix (28 wt%). A/W glass-ceramic (Cerabone) has a compressive strength of 1080 MPa and a Young's modulus of 118 GPa, an order of magnitude higher than cortical bone. On contact with body fluid, the A/W glass-ceramic forms a surface layer of carbonated apatite (HCA) similar to biological apatite. The release of calcium to solution causes a hydrated silica layer to form on the glass phase, providing nucleation sites for the HCA layer. Figure 2 shows how A/W glass-ceramics bond to bone more rapidly than sintered HA, but less rapidly than Bioglass. The A/W glass-ceramics are not resorbable, but due to the high compressive strengths A/W glass-ceramics are used as replacement vertebrae, iliac prostheses and in a granular form as bone defect fillers.

SOL–GEL-DERIVED BIOACTIVE GLASSES

Until the late 1980s, bioactive glasses were generally melt-derived, with the majority of research aimed at the 45S5 Bioglass composition (46.1% SiO_2, 24.4% NaO, 26.9% CaO,

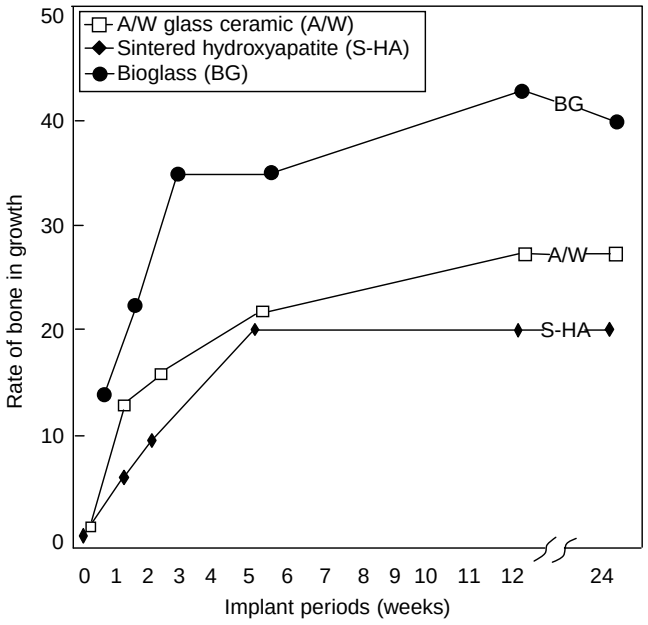


Figure 2. Graph of rate of bone ingrowth into the spaces between bioactive ceramic particles (diameters 300–500 μm) as a function of implantation time for Bioglass (BG), A/ W glass–ceramics (A/ W) and sintered hydroxyapatite (s-HA).

and 2.6% P_2O_5, in mol%) and apatite-wollastanite (A/W) glass-ceramics. The recognition that the silica gel layer plays a role in HCA nucleation and crystallization led to the development of the bioactive three component $CaO–P_2O_5–SiO_2$ sol–gel-derived glasses by Li et al. (25).

A sol is a dispersion of colloidal particles (solid particles with diameters 1–100 nm) in a liquid. A gel is an interconnected, rigid network of polymeric chains with average lengths > 1 μm in a continuous fluid phase and pores of submicrometer dimensions.

There are three methods that are used to produce sol–gel monoliths (26):

1. Network growth of discrete colloidal powders in solution.
2. Simultaneous hydrolysis and polycondensation of alkoxide or nitrate precursors, followed by hypercritical point drying of gels.
3. Simultaneous hydrolysis and polycondensation of alkoxide precursors, followed by ageing and drying under ambient conditions.

The pore liquid is removed from the three-dimensional (3D) network as a gas phase. A gel is defined as dried when the physically absorbed water is completely evacuated. This occurs between 100 and 180 °C. Homogeneous gel–glasses are only obtained when a sol–gel method using alkoxide precursors is employed (i.e., methods 2 or 3).

Liquid alkoxide precursors, such as $Si(OR)_4$, where R is CH_3, C_2H_5, and so on, are mixed with a solvent (usually water) and a catalyst. Tetraethylorthosilicate (TEOS) and tetramethylorthosilicate (TMOS) are the alkoxide precursors most commonly used for sol–gel derived silica. Hydrolysis and condensation (aqueous and alcoholic) reactions follow, forming a 3D SiO_2 network of continuous Si–O–Si links that span throughout the solvent medium. When sufficient interconnected Si–O–Si bonds are formed in a region, they respond cooperatively as a sol. A silica network is formed by the condensation reactions and the sol becomes a gel when it becomes rigid and it can support a stress elastically. The gel point is characterized by a steep increase in elastic and viscous moduli. A highly interconnected 3D gel network is obtained, composed of $(SiO_4)_4$ tetrahedra bonded either to neighboring silica tetrahedra via bridging oxygen (BO) bonds or by Si–O–Ca or Si–O–P nonbridging (NBO) bonds. The gel consists of interpenetrating solid and liquid phases: the liquid (a byproduct of the polycondensation reactions) prevents the solid network from collapsing and the solid network prevents the liquid from escaping. The gel is aged at ~60 °C to allow cross-linking of silica species and further network formation. The liquid is then removed from the interconnected pore network by evaporation of the solvent at elevated temperature to form a "xerogel". During this stage, the gel network undergoes considerable shrinkage and weight loss. This stage is critical in obtaining crack-free bodies as large capillary stresses can develop due to solvent evaporation through the pore network. To prevent cracking, the drying process must be controlled by decreasing liquid surface energy, by controlling the rates of hydrolysis and condensation using the precursors, and controlling the thermal drying conditions carefully. Extending the ageing time can help prevent cracking during drying (27). However, even under optimum conditions it is difficult to produce multicomponent crack-free silica-based glasses with diameters in excess of 10 mm.

Dried xerogels have a very large concentration of silanols on the surface of the pores, which renders them chemically unstable at room temperature. The gel is stabilized by sintering at > 500 °C, which removes chemically active surface groups (such as silanols or trisiloxane rings) from the pore network, so that the surface does not rehydroxylate in use. Thermal methods are most common, but chemical methods, involving replacing silanols with more hydrophobic and less reactive species are also possible.

In multicomponent systems, the stabilization process also decomposes other species in the gel such a nitrates or organics. In this thesis nitrates are present after drying. Such species are a source of inhomogeneity and are biologically toxic. Pure $Ca(NO_3)_2$ decomposes at 561 °C, therefore thermal stabilization must be carried out above this temperature.

Sol–gel derived bioactive glasses exhibit a mesoporous texture, that is, pores with diameters in the range 2–50 nm that are inherent to the sol–gel process. The textural properties of the glass are affected by each stage of the sol–gel process, that is, temperature, sol composition, ageing, drying rate, and stabilization temperatures and rates.

Advantages of Sol–Gel-Derived Glasses

There are several advantages of a sol–gel-derived glass over a melt-derived glass, which are important for biomedical applications. Sol–gel-derived glasses have (26):

1. Lower processing temperatures (600–700 °C for gel–glasses compared to 1100–1300 °C for melt-derived glasses).
2. The potential of improved purity, required for optimal bioactivity due to low processing temperatures and high silica and low alkali content.
3. Improved homogeneity.
4. Wider compositions can be used (up to 90 mol% SiO_2) while maintaining bioactivity.
5. Better control of bioactivity by changing composition or microstructure.
6. Structural variation can be produced without compositional changes by control of hydrolysis and polycondensation reactions during synthesis.
7. A greater ease of powder production.
8. Interconnected nanometer scale porosity that can be varied to control dissolution kinetics or be impregnated with biologically active phases such as growth factors.
9. A higher bioactivity due to the textural porosity (SA/V ratio two orders of magnitude higher than melt-derived glasses).
10. Gel–glasses are resorbable and the resorption rate can be controlled by controlling the mesoporosity.

11. Can be foamed to provide interconnected pores of 10–200 μm, mimicking the architecture of trabecular bone.

The mechanism for HCA formation on bioactive gel–glasses follows most of the same 11 stages as those for melt-derived glasses except that dissolution rates are much higher due to the mesoporous texture which creates a higher SA/ V ratio, increasing the area of surface exposed for cation exchange (stage 1) and silica network break-up (stage 2). There are also more sites available for HCA layer formation (19).

FEATURES OF CLASS A BIOACTIVE MATERIALS

An important feature of Class A bioactive materials is that they are osteoproductive as well as osteoconductive. In contrast, Class B bioactive materials exhibit only *osteoconductivity*, defined as the characteristic of bone growth and bonding along a surface. Dense synthetic HA ceramic implants exhibit Class B bioactivity. *Osteoproduction* occurs when bone proliferates on the surfaces of a material due to enhanced osteoblast activity. Enhanced proliferation *and* differentiation of osteoprogenitor cells, stimulated by slow resorption of the Class A bioactive particles, are responsible for osteoproduction.

Is Bioactive Fixation the Solution?

During the last decade, it has been assumed that improved interfacial stability achieved with bioactive fixation would improve implant survivability. Clinical trials have shown this to often not be the case. Replacement of the roots of extracted teeth with dense HA ceramic cones to preserve the edentulous alveolar ridge of denture wearers resulted in generally $<50\%$ survived at only 5 years. Early use of HA-coated orthopedic implants seldom survived 10 years $>85\%$ figure for cemented total hip prostheses (1). However, long-term success rates of bioactive HA coatings have improved during the last decade due to greater control of the coating process. The survivability of HA coated femoral stems is now equivalent at 10 years to cemented prostheses. It will take another 5 years to know if survivability is superior when HA coatings are used.

Why is bioactive fixation not a panacea to hip implant survivability? There are three primary reasons: (*1*) metallic prostheses with a bioactive coating still have a mismatch in mechanical properties with host bone, and therefore less than optimal biomechanical and bioelectric stimuli, at the bonded interface; (*2*) the bioactive bonded interface is unable to remodel in response to applied load; and (*3*) use of bioactive materials does not solve the problem of osteolysis due to wear debris generated from the polyethylene cups. Use of alumina–alumina bearing surfaces eliminates most wear debris from total hip prostheses, but increases the cost of the prosthesis by 200–300%. For younger patients the cost is acceptable, but for the general population it often is considered to be too expensive.

Most biomaterials in use today and the prostheses made from the materials have evolved from trial and error experiments. Optimal biochemical and biomechanical features that match living tissues have not been achieved, so it also is not surprising that long-term implant survivability has not been improved very much during the last 15 years.

THE BIOCOMPOSITES ALTERNATIVE

Bone is a natural composite of collagen fibers (polymer) and mineral (ceramic). Therefore to create an implant that mimics the mechanical properties of bone, a composite should provide high toughness, tensile strength, fatigue resistance, and flexibility while maintaining modulus similar to bone. Biocomposites are being developed to eliminate elastic modulus mismatch and stress shielding of bone. Two approaches have been tried. Bioinert composites, such as carbon–carbon fiber composite materials, are routinely used in aerospace and automotive applications. These lightweight, strong, and low modulus materials would seem to offer great potential for load-bearing orthopedic devices. However, delamination can occur under cyclic loading that releases carbon fibers into the interfacial tissues. The carbon fibers can give rise to a chronic inflammatory response. Thus, bioinert composites are not widely used and are unlikely to be a fruitful direction for development in the next decade.

BIOACTIVE COMPOSITES

The second approach is to make a bioactive composite that does not degrade, such as pioneered by at the IRC in Biomedical Materials, University of London. Bonfield and co-workers (28) increased the stiffness of a biocompatible polymer (polyethylene) from 1 to 8 GPa by adding a secondary phase with higher modulus (HA). The compressive strength of the composite, now called HAPEX, was 26 MPa. Addition of HA also meant that the composite would also bond to bone. Applications for HAPEX have included ossicular replacement prostheses and the repair of orbital floors in the eye socket. Ideally, it is possible to match the properties of both cancellous and cortical bone, although this is seldom achieved by the biocomposites available today. A challenge for the next decade is to use advanced materials processing technology to improve the interfacial bonding between the phases and reduce the size of the second-phase particles, thereby increasing the strength and fracture toughness of these new materials.

Another option is to use a resorbable polymer matrix for a biocomposite that will be replaced with mineralizing bone as the load on the device is increased. Work in this area is in progress, but it is difficult to maintain structural integrity as resorption occurs. The tissue engineering alternative is based upon this concept (29). Further details on biomedical composites can be found in a review by Thompson and Hench (30).

A NEW REVOLUTION IN ORTHOPEDICS?

We suggest that the orthopedics revolution of the last 30 years, the revolution of replacement of tissues by transplants and implants, has run its course. It has led to a

remarkable increase in the quality of life for millions of patients; total joint prostheses provide excellent performance and survivability for 15–20 years. Prostheses will still be the treatment of choice for many years to come for patients of 70 years or older. However, continuing the same approach of the last century; that is, modification of implant materials and designs is not likely to reach a goal of 25–30 years implant survivability, an increasing need of our ageing population. We need a change in emphasis in orthopedic materials research; in fact, we need a new revolution.

BIOCERAMICS IN REGENERATIVE MEDICINE

The challenge for the next millennium in bioceramics and biomedical materials in general is to shift the emphasis of research toward assisting or enhancing the body's own reparative capacity. We must recognize that within our cells lies the genetic information needed to replicate or repair any tissue. We need to learn how to activate the genes to initiate repair at the right site.

Our goal of regeneration of tissues should involve the restoration of metabolic and biochemical behavior at the defect site, which would lead to restoration of biomechanical performance, by means of restoration of the tissue structure leading to restoration of physiological function.

The concept requires that we develop biomaterials that behave in a manner equivalent to an autograft, that is, what we seek is a *regenerative allograft* or *scaffold*. This is a great challenge. However, the time is ripe for such a revolution in thinking and priorities. Regenerative medicine encompasses many fields. We concentrate here on the use of bioceramics in tissue engineering and regeneration applications that require scaffolds to promote tissue repair. Tissue regeneration techniques involve the use of a scaffold that can be implanted into a defect to guide and stimulate tissue regrowth *in situ*. The scaffold should resorb as the tissue grows, leaving no trace. In tissue engineering applications, the scaffolds are seeded with cells *in vitro* to produce the basis of a tissue before implantation; cells extracted from a patient, seeded on a scaffold of the desired architecture and the replacement tissue grown in the laboratory, ready for implantation. The use of the patient's own cells from the same patient would eliminate any chance of immunorejection (31).

GENETIC CONTROL BY BIOACTIVE MATERIALS

We have now discovered the genes involved in phenotype expression and bone and joint morphogenesis, and thus are on the way toward learning the correct combination of extracellular and intracellular chemical concentration gradients, cellular attachment complexes, and other stimuli required to activate tissue regeneration *in situ*. Professor Julia Polak's group at the Imperial College London Centre for Tissue Engineering and Regenerative Medicine has recently shown that seven families of genes are up- and down-regulated by bioactive glass extracts during proliferation and differentiation of primary human osteoblasts *in vitro* (32). These findings should make it possible to design a new generation of bioactive materials for regeneration of bone. The significant new finding is that low levels of dissolution of the bioactive glass particles in the physiological environment exert a genetic control over osteoblast cell cycle and rapid expression of genes that regulate osteogenesis and the production of growth factors.

Xynos et al. (33) showed that within 48 h a group of genes was activated including genes encoding nuclear transcription factors and potent growth factors. These results were obtained using cultures of human osteoblasts, obtained from excised femoral heads of patients (50–70 years) undergoing total hip arthroplasy.

In particular, insulin-like growth factor (IGF) II, IGF-binding proteins, and proteases that cleave IGF-II from their binding proteins were identified (34). The activation of numerous early response genes and synthesis of growth factors was shown to modulate the cell cycle response of osteoblasts to the bioactive glasses and their ionic dissolution products. These results indicate that bioactive glasses enchance osteogenesis through a direct control over genes that regulate cell cycle induction and progression. However, these molecular biological results also confirm that the osteoprogenitor cells must be in a chemical environment suitable for passing checkpoints in the cell cycle toward the synthesis and mitosis phases. Only a select number of cells from a population are capable of dividing and becoming mature osteoblasts. The others are switched into apoptosis and cell death. The number of progenitor cells capable of being stimulated by a bioactive medium decreases as a patient ages, which may account for the time delay in formation of new bone in augmented sites.

Enormous advances have been made in developmental biology, genetic engineering, cellular and tissue engineering, imaging and diagnosis, and in microoptical and micromechanical surgery and repair. Few of these advances have, as yet, been incorporated with the molecular design of new biomaterials. This must be a high priority for the next two decades of research. However, for large defects a scaffold is required to guide tissue regeneration in 3D. Ideally, the scaffold should also release active agents that can also stimulate the cells within the tissue.

AN IDEAL SCAFFOLD

An ideal scaffold is one that mimics the extracellular matrix of the tissue that is to be replaced so that it can act as a 3D template on which cells attach, multiply, migrate, and function. The criteria for an ideal scaffold for bone regeneration are that it (35,36):

1. Is made from a material that is biocompatible (i.e., not cytotoxic).
2. Acts as template for tissue growth in 3D.
3. Has an interconnected macroporous network containing pores with diameters in excess of 100 μm for cell penetration, tissue ingrowth and vascularization, and nutrient delivery to the center of the regenerating tissue on implantation.
4. Bonds to the host tissue without the formation of scar tissue (i.e., is made from an bioactive and osteoconductive–osteoproductive material).

5. Exhibits a surface texture that promotes cell adhesion, adsorption of biological metabolites.
6. Influences the genes in the bone generating cells to enable efficient cell differentiation and proliferation.
7. Resorbs at the same rate as the tissue is regenerated, with degradation products that are nontoxic and that can be easily be excreted by the body, for example, via the respiratory or urinary systems. Is made from a processing technique that can produce irregular shapes to match that of the defect in the patient. Has the potential to be commercially producible to the required ISO (International Standards Organization) or FDA (Food and Drug Administration) standards.
8. Can be sterilized and maintained as a sterile product to the patient.
9. Can be produced economically to be covered by national and/ or private healthcare insurances.

For *in situ* bone regeneration applications, the mechanical properties of the scaffold are also critical and the modulus and elastic strength the scaffold should be similar to that of the natural bone. However, for tissue engineering applications only the mechanical properties of the final tissue engineered construct are critical (36).

TYPES OF BIOCERAMIC SCAFFOLD

Many types of porous bioceramics have been developed and are reviewed in Ref. (37). The simplest way to generate porous scaffolds from ceramics such as HA or TCP is to sinter particles. Particles are usually mixed with a wetting solution, such as poly(vinyl alcohol), and compacted by cold isostatic pressing to form a "green" body, which is sintered (heated to $\sim$1200 °C) to improve mechanical properties. Porosity can be increased by adding fillers such as sucrose to the powder and the wetting solution, which burnout on sintering. Komlev et al. (38) produced porous HA scaffolds with interconnected interparticle pore diameters of $\sim$100 μm, and a tensile strength of $\sim$0.9 MPa by sintering HA spheres 500 μm in diameter.

Other techniques include adding a combustible organic material to a ceramic powder burned away during sintering leaving closed pores; freeze drying where ice crystals are formed in ceramic slurries and then sublimation of the ice leaves pores; polymer foam replication where the ceramic slurry is poured into a polymer foam, which is then burnt out on sintering leaving a pore network. Most of these techniques produced porous ceramics that were not suitable for tissue engineering applications. Typical problems were either that the pore diameters were too low, the pores were closed, the pore distributions were very heterogeneous or mechanical strengths were very low.

Recently, rapid prototyping has been adapted for producing scaffolds with controlled and homogeneous interconnected porosity (39). Rapid prototyping is a generic term for a processing technique that produces materials in a shape determined by CAD (computer aided design) software on a computer. Such materials are usually built up layer-by-layer using a liquid phase or slurry of the material that cures or sets on contact with a substrate. Specific techniques include stereolithography, selective laser sintering, fused deposition modeling and ink-jet printing. It is a challenge to apply these techniques to direct processing of bioactive ceramic scaffolds.

Perhaps the most successful technique for synthesis of porous HA that could be produced in any size of shape, with interconnected macropore diameters in excess of 100 μm is the gel-casting process.

GEL-CASTING OF HA

In the gel-casting of HA, aqueous suspensions of HA particles, dispersing agents, and organic monomers (6 wt% acrylate/diene) are foamed. The organic monomers must be water soluble and retain a high reactivity. Foaming is the incorporation of air into a ceramic to produce a porous material. Once the slurry has foamed, *in situ* polymerization of the monomers is initiated and cross-linking occurs, forming a 3D polymeric network (gel), which produces strong green bodies. Foaming is achieved by vigorous agitation at 900 rpm with the addition of a surfactant (Tergitol TMN10; polyethylene glycol trimethylnonyl ether) under a nitrogen atmosphere (40). Surfactants are macromolecules composed of two parts, one hydrophobic and one hydrophilic. Owing to this configuration, surfactants tend to adsorb onto gas–liquid interfaces with the hydrophobic part being expelled from the solvent and a hydrophilic part remaining in contact with the liquid. This behavior lowers the surface tension of the gas–liquid interfaces, making the foam films thermodynamically stable, which would otherwise collapse in the absence of surfactant (41). Once stable bubble formation is achieved, the polymerization process is initiated using ammonium persulphate and a catalyst (TEMED, *N*,*N*,*N*′,*N*′-tetramethylethylenediamine) and the viscous foam is cast into moulds immediately prior to gelation. The surfactant stabilises the air bubbles until gelation provides permanent stability (40).

The porous green bodies are then sintered to provide mechanical strength and to burnout the organic solvents. Foam volume (and hence porosity) can be controlled by the surfactant concentration in the slurry. The materials produced exhibited interconnected pores of maximum diameter of 100–200 μm, which is ideal for tissue engineering applications.

The gel-cast HA scaffolds satisfy many of the criteria of the ideal scaffold, however, the criteria of controlled resorbability and genetic stimulation are not fulfilled. A bioactive glass scaffold would fulfil these criteria and also be able to bond to soft tissue. However, producing a 3D macroporous scaffold from a glass is difficult.

POROUS MELT-DERIVED BIOACTIVE GLASSES

Theoretically, the gel-casting process could be applied to melt-derived bioactive glass powders. However, such glasses undergo surface reactions on contact with solutions to produce an HCA surface layer and it is desirable to control the reaction before a scaffold is ready for clinical use.

Livingston et al. (42) produced a simple sintered scaffold by mixing 45S5 melt-derived bioactive glass (Bioglass) powders, with a particle size range of 38–75 μm, with 20.2 wt% camphor ($C_{10}H_{16}O$) particles, with particle size range of 210–350 μm. The mixture was dry pressed at 350 MPa and heat treated at 640 °C for 30 min. The camphor decomposed to leave porous Bioglass blocks. Macropores were in the region of 200–300 μm in diameter, however, the total porosity was just 21% as there were large distances between pores.

Yuan et al. (43) produced similar scaffolds by foaming Bioglass 45S5 powder with a dilute H_2O_2 solution and sintered at 1000 °C for 2 h to produce a porous glass–ceramic. The pores were irregular in shape and relatively few in number, implying that interconnectivity was poor, but pore diameters were in the range 100–600 μm. The pores appeared to be more like orientated channels running through the glass, rather than an interconnected network. The samples were implanted into the muscle of dogs and were found for the first time to be osteoinductive. Bone was formed directly on the solid surface and on the surface of crystal layers that formed in the inner pores. Osteogenic cells were observed to aggregate near the material surface and secrete bone matrix, which then calcified to form bone. However, although the implants had a porosity of ~30% only 3% bone was formed. It seems that creating interconnected pore networks in bioactive glasses by sintering is not practical at the present time, although sol–gel derived bioactive glasses may do so.

SOL–GEL DERIVED BIOACTIVE GLASS FOAMS: AN IDEAL SCAFFOLD?

The foaming process has also been applied to sol–gel derived bioactive glasses (44). The resulting scaffolds exhibit the majority of the criteria for an ideal scaffold.

Figure 3 shows an scanning electron microscopy (SEM) micrograph of a typical foam of the 70S30C composition (70 mol% SiO_2, 30 mol% CaO). The scaffolds have a hierarchical pore structure similar to that of trabecular bone, with interconnected macropores with diameters in excess of 100 μm and a textural porosity with diameters in range 10–20 nm (mesopores), which are inherent to the sol–gel process. The scaffolds have the potential to guide tissue growth, with the interconnected macropores providing channels for cell migration, tissue ingrowth, nutrient delivery, and eventually vascularisation (blood vessel ingrowth throughout the regenerated tissue). The mesoporous texture enhances the resorbability and biaoctivity of the scaffolds and provides nucleation points for the HCA layer and sites for cell attachment for anchorage dependant cells such as osteoblasts. The bioactive glass composition contributes high bioactivity, controlled resorbability, and the potential for the ionic dissolution products (Si and Ca) to stimulate the genes in bone cells to enhance bone regeneration.

Figure 4 shows a flow chart of the sol–gel foaming process. Sol–gel precursors [e.g., tetraethoxyl orthosilicate (TEOS, $Si(OC_2H_5)_4$] are mixed in deionized water in the presence of an acidic hydrolysis catalyst. Simultaneous hydrolysis and polycondensation reactions occur beginning with the formation of a silica network. Viscosity of the sol increases as the condensation reaction continues and the network grows. Other alkoxides–salts can be added to introduce network modifiers (e.g., CaO species). On completion of hydrolysis, the sol is foamed by vigorous agitation with the addition of a surfactant. A gelling agent [hydrofluoric acid (HF), a catalyst for polycondensation] is added to induce a rapid increase in viscosity and reduce the gelling time.

The surfactant stabilized the bubbles that were formed by air entrapment during the early stages of foaming by lowering the surface tension of the solution. As viscosity rapidly increased and the gelling point was approached, the solution was cast into airtight moulds. The gelling point is the point at which the meniscus of the foamed sol does not move, even if the mold is tilted. Casting must

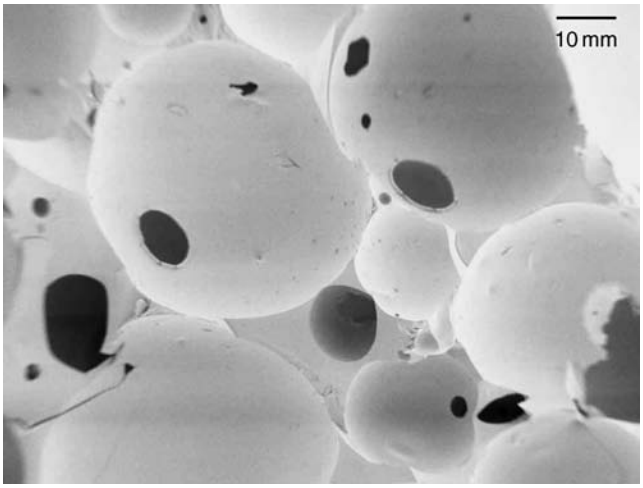


Figure 3. An SEM micrograph of a sol–gel derived bioactive glass foam scaffold.

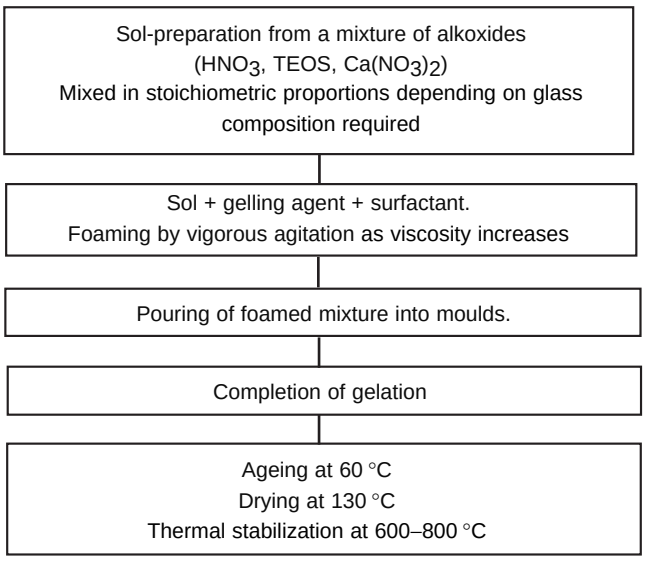


Figure 4. Flow chart of the sol–gel foaming process.

take place immediately prior to the gelation point. The gelation process provided permanent stabilization for the bubbles.

A foam scaffold is produced that sits in the liquid mixture of water and alcohol (pore liquor) produced as a byproduct of the polycondensation reaction. The foams are then subjected to the thermal treatments of ageing, drying, and stabilization. Ageing is done at 60 °C, leaving the foam immersed in its pore liquor. Ageing allows further cross-linking of the silica network and a thickening of the pore walls. Pore coarsening also occurs when larger pores grow at the expense of smaller ones.

Drying involves the evaporation of the pore liquor, which is critical and must be carried out under very carefully controlled conditions to prevent cracking under capillary pressure. Silica-based glasses that only contain the textural mesopores cannot be produced as monoliths with diameters in excess of 10 mm due to the high capillary stresses during drying. The formation of interconnected pore channels with large diameters allows efficient evaporation of the pore liquor; therefore very large crack-free scaffolds (in excess of 100 mm diameter) can be made. Thermal stabilization is carried out (again under carefully controlled heating regimes) at a minimum of 600 °C to ensure removal of silanol and nitrate groups from the glass.

The variables in each stage of the foaming process affect the final structure and properties of the foams (45,46). The percentage and pore volume of the textural mesopores can be controlled by the glass composition and the alkoxide: water ratio in initial sol preparation. Therefore the resorbability and bioactivity of the scaffolds can be easily controlled. The macropore diameters are little affected until the sintering temperature increases >800 °C. However, the glass composition, the foaming temperature, the surfactant concentration and type, the gelling agent concentration heavily affect the macropore diameters, and interconnectivity, which are vital for tissue engineering applications.

Three compositions have been successfully foamed; the tertiary 58S (60 mol% SiO_2, 36 mol% CaO, 4 mol% P_2O_5), the binary 70S30C (70 mol% SiO_2, 30 mol% CaO) composition, and 100S silica. The binary composition 70S30C (70 mol% SiO_2, 30 mol% CaO) has been found to be the most suitable to the foaming process, producing crack-free foams scaffolds with porosities in the range 60–95% (depending on the other variables in the process). Macropores were homogeneously distributed with diameters up of up to 600 μm and modal interconnected pore diameters of up to 150 μm.

Due to the nature of the sol–gel process the scaffolds can be produced in many shapes, which are determined simply by the shape of the casting mould. The scaffolds can be produced from various compositions of gel-derived glasses. All foam compositions can be easily cut to a required shape. Figure 5 shows foams produced in various shapes.

The only criterion not addressed is the matching of mechanical properties of the scaffolds to bone for *in situ* bone regeneration applications. The compressive strength of the foams (~2.5 MPa for 70S30C foams sintered at 800 °C) is less than that of trabecular bone (~10 MPa). However, the mechanical properties of these foams should be sufficient for tissue engineering applications, where bone would be grown on a scaffold in the laboratory before implantation. Work on improving the mechanical properties is ongoing.

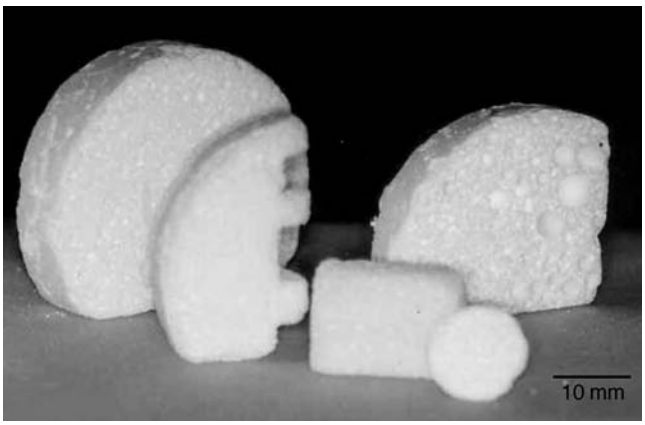


Figure 5. Sol–gel derived bioactive glass scaffolds. (Courtesy of Dr. P. Sepulveda.)

BIOLOGICAL RESPONSES TO SOL–GEL DERIVED BIOACTIVE GLASSES

The biological response to bioactive gel–glasses made from the $CaO–P_2O_5–SiO_2$ system provides evidence that bone regeneration is feasible. An important factor for future research is that the structure and chemistry of bioactive gel–glasses can be tailored at a molecular level by varying the composition (such as SiO_2 content) or the thermal or environmental processing history. The compositional range for Class A bioactive behavior is considerably extended for the bioactive gel–glasses over Class A bioactive glasses and glass–ceramics made by standard high temperature melting or hot pressing. Thus, gel–glasses offer several new degrees of freedom over the influence of cellular differentiation and tissue proliferation. This enhanced biomolecular control will be vital in developing the matrices and scaffolds for engineering of tissues and for the *in vivo* regenerative allograft stimulation of tissue repair.

Evidence of the regenerative capacity of bioactive gel–glasses powders is based on a comparison of the rates of proliferation of trabecular bone in a rabbit femoral defect (47). Melt-derived Class A 45S5 bioactive glass particles exhibit substantially greater rates of trabecular bone growth and a greater final quantity of bone than Class B synthetic HA ceramic or bioactive glass–ceramic particles. The restored trabecular bone has a morphological structure equivalent to the normal host bone after 6 weeks; however, the regenerated bone still contains some of the larger (>90 μm) bioactive glass particles. Wheeler et al. (48) showed that the use of bioactive gel–glass particles in the same animal model produces an even faster rate of trabecular bone regeneration with no residual gel–glass particles of either the 58S or 77S composition. The gel–glass particles resorb more rapidly during proliferation of

trabecular bone. Thus, the criteria of a regenerative allograft cited above appear to have been met. Recent results of *in vivo* subperiosteum implantation of 58S foams on the calvaria of New Zealand rabbits (49) showed that bone regeneration occurred more rapidly for 58S foams compared to 58S powder and that the regeneration was in line with that produced by compacted melt-derived Bioglass powders that are available commercially as Perioglas and Novabone.

SOFT TISSUE ENGINEERING

The interactions between cells and surfaces play a major biological role in cellular behavior. Cellular interactions with artificial surfaces are mediated through adsorbed proteins. A common strategy in tissue engineering is to modify the biomaterial surface selectively to interact with a cell through biomolecular recognition events. Adsorbed bioactive peptides can allow cell attachment on biomaterials, and allow 3D structures modified with these peptides to preferentially induce tissue formation consistent with the cell-type seeded, either on or within the device (50). The surface of the gel-derived foams has been modified with organic groups and proteins to create scaffolds that have potential for lung tissue engineering (50,51). If cells can recognize the proteins adsorbed on the surface of a biomaterial they can attach to it and start to differentiate, inducing tissue regeneration. However, if cells do not recognize the proteins, an immunogenic response may result, initiating a chronic inflammation that can lead to failure of the device. Besides promoting cell-surface recognition, bioactive peptides can be used to control or promote many aspects of cell physiology, such as adhesion, spreading, activation, migration, proliferation, and differentiation.

Three-dimensional scaffolds have been produced that allow the incorporation and release of biologically active proteins to stimulate cell function. Laminin was adsorbed on the textured surfaces of binary 70S30C (70 mol% SiO_2–30 mol% CaO) and ternary 58S (60 mol% SiO_2–36 mol% CaO–4 mol% P_2O_5) sol–gel derived bioactive foams. The covalent bonds between the binding sites of the protein and the ligands on the scaffolds surface do not denaturate the protein. *In vitro* studies show that the foams modified with chemical groups and coated with laminin maintained bioactivity, as demonstrated by the formation of the (HCA) layer formed on the surface of the foams on exposure to simulated body fluid (SBF). Sustained and controlled release from the scaffolds over a 30-day period was achieved. The laminin release from the bioactive foams followed the dissolution rate of the material network. These findings suggest that bioactive foams have the potential to act as scaffolds for soft tissue engineering with a controlled release of proteins that can induce tissue formation or regeneration.

The way that proteins or other bioactive peptides interact with surfaces can alter their biological functionality. In order to achieve full functionality, peptides have to adsorb specifically. They also must maintain conformation in order to remain functional biologically. Chemical groups, such as amine and mercaptan groups, are known to control the ability of surfaces to interact with proteins (51). In addition, these chemical groups can allow protein–surface interactions to occur such that the active domains of the protein can be oriented outward, where they can be maximally effective in triggering biospecific processes. Cell cultures of mouse lung epithelial cells (MLE-12) on modified 58S foam scaffolds showed that cells attached and proliferated best on 58S foam modified with amine groups (using amine propyltriethoxysilane, APTS) and coated with laminin (52).

SUMMARY

During the last century, a revolution in orthopedics occurred that has led to a remarkably improved quality of life for millions of aged patients. Specially developed bioceramics were a critical component of this revolution. However, survival of prostheses appears to be limited to ~20 years. We conclude that a shift in emphasis from replacement of tissues to regeneration of tissues should be the challenge for orthopedic materials in the new millennium. The emphasis should be on the use of materials to activate the body's own repair mechanisms, that is, regenerative allografts. This concept will combine the understanding of tissue growth at a molecular biological level with the molecular design of a new generation of bioactive scaffolds that stimulate genes to activate the proliferation and differentiation of osteoprogenitor cells and enhance rapid formation of extracellular matrix and growth of new bone *in situ*. The economic and personal benefits of *in situ* regenerative repair of the skeleton on younger patients will be profound.

BIBLIOGRAPHY

Cited References

1. Jones JR, Hench LL. Biomedical materials for the new millennium: A perspective on the future. J Mat Sci T 2001;17: 891–900.
2. Ratner BD, Hoffman AS, Schoen FJ, Lemmons JE. Biomaterials Science: An Introduction to Materials in Medicine. London: Academic Press; 1996.
3. Berry DJ, Harmsen WD, Cabanela ME, Morrey MF. Twenty-five-year survivorship of two thousand consecutive primary Charnley total hip replacements. J Bone Jt Surg 2002;84A(2): 171–177.
4. Hench LL, Polak JM. Third generation biomedical materials. Science 2002;295(5557):1014–1018.
5. Hench LL, Wilson J. An Introduction to Bioceramics. Singapore: World Scientific; 1993.
6. Hench LL. Biomaterials: A forecast for the future. Biomaterials 1998;19:1419–1423.
7. Black J, Hastings G. Handbook of Biomaterial Properties. London: Chapman and Hall; 1998.
8. Hench LL. Bioceramics. J Am Ceram 1998;81(7):1705–1728.
9. Hulbert SF. The use of alumina and zirconia in surgical implants. In: Hench LL, Wilson J, editors. An Introduction to Bioceramics. Singapore: World Scientific; 1993.
10. Hulbert SF, Bokros JC, Hench LL, Heimke G. Ceramics in Clinical Applications: Past, Present, and Future. In: Vincenzini P, editor. High tech Ceramics. Amsterdam: Elsevier; 1987.
11. Bilezikian JP, Raisz LG, Rodan GA. Principles of bone biology. London: Academic Press; 1996.

12. Sumner DR, Galante JO. Determinants of stress shielding—design versus materials versus interface. Clin Orthop Relat R 1992;274:202–212.
13. Marcus R, Feldman D, Kelsey JL. Osteroporosis. London: Academic Press; 1996.
14. de Groot K. Bioceramics of Calcium Phosphate. Boca Raton, FL: CRC Press; 1983.
15. Hench LL, West JK. Biological applications of bioactive glasses. Life Chem Rep 199;13:187–241.
16. LeGeros RZ, LeGeros JP. Dense Hydroxyapatite. In: Hench LL, Wilson J, editors. An Introduction to Bioceramics. Singapore: World Scientific; 1993.
17. Ducheyne P, Hench LL, Kagan A, Martens M, Burssens A, Mulier JC. The effect of hydroxyapatite impregnation of skeletal bonding of porous coated implants. J Biomed Mater Res 1980;14:225–237.
18. Hench LL, Splinter RJ, Allen WC, Greenlee TK. Bonding mechanism at the interface of ceramic prosthetic implants. J Biomed Mater Res 1971;74:1478–1570.
19. Sepulveda P, Jones JR, Hench LL. *In vitro* dissolution of melt-derived 45S5 and sol–gel derived 58S bioactive glasses. J Biomed Mater Res 2002;61(2):301–311.
20. Jones JR, Sepulveda P, Hench LL. Dose-dependent behaviour of bioactive glass dissolution. J Biomed Mater Res 2001;58:720–726.
21. Wallace KE, Hill RG, Pembroke JT, Brown CJ, Hatton PV. Influence of sodium oxide content on bioactive glass properties. J Mater Sci Mater Med 1999;10(12):697–701.
22. Wilson J, Douek E, Rust K. Bioglass® Middle Ear Devices: 10 Year Clinical Results. In: Hench LL, Wilson J, Greenspan DC, editors. Bioceramics 8. Oxford: Pergamon; 1995. p 239–245.
23. Fetner AE, Hartigan MS, Low SB. Periodontal repair using Perioglas® in non-human primates: Clinical and histologic observations. Comp Cont E Dent 1994;15(7):932–939.
24. Kokubo T, Ito S, Shigematsu M, Sakka S, Yamamuro T, Higashi S. Mechanical properties of a new type of apatite containing glass-ceramic for prosthetic application. J Mater Sci 1985;20:2001–2004.
25. Li R, Clark AE, Hench LL. Effect of structure and surface area on bioactive powders made by sol–gel process. In: Hench LL, West JK, editors. Chemical Processing of Advanced Materials. Vol. 56, New York: John Wiley & Sons; 1992. 627–633.
26. Hench LL, West JK. The Sol–Gel Process. Chem Rev 1990;90: 33–72.
27. Ishizaki K, Komarneni S, Nanko M. Sol–Gel Processing: Designing Porosity, Pore Size and Polarity and Shaping Processes. In: Ishizaki K, Komarneni S, Nanko M, editors. Porous Materials: Process Technology and Applications. London: Kluwer Academic Publishers; 1998. p 67–180.
28. Huang J, DiSilvo L, Wang M, Tanner KE, Bonfield W. *In vitro* mechanical and biological assessment of hydroxyapatite-reinforced polyethylene composite. J Mat S-M M 1997;8: 775–779.
29. Day R, Boccaccini AR, Roether JA, Surey S, Forbes A, Hench LL, Gabe S. The effect of Bioglass® on epithelial cell and fibroblast proliferation and incorporation into a PGA matrix. Gastroenterology 2002;122(4) T875 Suppl 1.
30. Thompson ID, Hench LL. Medical Applications of Composites. Comprehensive Composite Mater 2000; (6.39) 727–753.
31. Ohgushi H, Caplan AI. Stem Cell Technology and Bioceramics: From cell to Gene Engineering. J Biomed Mater Res B 1999;48:913–927.
32. Hench LL, Polak JM, Xynos ID, Buttery LDK. Bioactive Materials to Control Cell Cycle. Mat Res Innovat 2000;3: 313–323.
33. Xynos ID, Hukkanen MVJ, Batten JJ, Buttery LD, Hench LL, Polak JM. Bioglass® 45S5 Stimulates Osteoblast Turnover and Enhances Bone Formation In Vitro: Implications and Applications for Bone Tissue Engineering. Calciff Tiss 2000;67:321–329.
34. Xynos ID, Edgar AJ, Buttery LD, Hench LL, Polak JM. Ionic Dissolution Products of Bioactive Glass Increase Proliferation of Human Osteoblasts and Induce Insulin-like Growth Factor II mRNA Expression and Protein Synthesis, Biochem Biophys Res 2000;276:461–465.
35. Freyman TM, Yannas IV, Gibson LJ. Cellular materials as porous scaffolds for tissue engineering. Prog Mat Sci 2001;46: 273–282.
36. Holy CE, Fialkov JA, Davies JE, Shoichet MS. Use of a biomimetic strategy to engineer bone. J Biomed Mater Res 2003;65A:447–553.
37. Jones JR. Bioactive Glass 3D Scaffolds for Tissue Engineering, [dissertation]. London (UK): Imperial College London; 2002.
38. Komlev VS, Barimov SM. Porous hydroxyapaptite ceramics of bi-modal pore size distribution. J Mater Sci Mater Med 2002;13:295–299.
39. Chu GTM, Orton DG, Hollister SJ, Feinberg SE, Halloran JW. Mechanical and *in vivo* performance of hydroxyapatite implants with controlled architectures. Biomaterials 2002; 23:1283–1293.
40. Sepulveda P, Binner JGP, Rogero SO, Higa OZ, Bressiani JC. Production of porous hydroxyapatite by the gel-casting of foams and cytotoxic evaluation. J Biomed Mater Res 2000;50: 27–34.
41. Rosen MJ. Surfactants and Interfacial Phenomena. 2nd ed, New York: Wiley; 1989. p 277–303.
42. Livingston T, Ducheyne P, Garino J. *In vivo* evaluation of a bioactive scaffold for bone tissue engineering. J Biomed Mater Res 2002;62:1–13.
43. Yuan H, de Bruijn JD, Zhang X, van Blitterswijk CA, de Groot K. Bone Induction by porous glass ceramic made from Bioglass® (45S5). J Biomed Mater Res 2001;58(3): 270–276.
44. Sepulveda P, Jones JR, Hench LL, Bioactive sol–gel foams for tissue repair. J Biomed Mater Res 2002;59(2):340–348.
45. Jones JR, Hench LL. The effect of processing variables on the properties of bioactive glass foams. J Biomed Mater Res In press.
46. Jones JR, Hench LL. The effect of surfactant concentration and glass composition on the structure and properties of bioactive foam scaffolds. J Mat Sci In press.
47. Oonishi H, Hench LL, Wilson J, Sugihara F, Tsuji E, Kushitani S, Iwaki H. Comparative bone growth behaviour in granules of bioceramic materials of various sizes. J Biomed Mater Res 1999;44(1):31–43.
48. Wheeler DL, Hoellrich RG, McLoughlin SW, Chamerland DL, Stokes KE. In Vivo Evaluation of Sol–Gel Bioglass®–Biomechanical Findings. In: Sedel L, Rey C, editors. Bioceramics. Volume 10, 1997. p 349–350.
49. Cook R 58S sol–gel Bioglass: a study of osteoproductive, interfacial and handling properties using new microscopic techniques. [dissertation] London (UK). University of London; 2003.
50. Lenza RFS, Jones JR, Vasconcelos WL, Hench LL. *In vitro* release kinetics of proteins from bioactive foams. J Biomed Mater Res In press.
51. Lenza RFS, Jones JR, Vasconcelos WL, Hench LL. *In vitro* release kinetics of proteins from bioactive foams. J Biomed Mater Res In press.
52. Mansur HS, Vasconcelos WL, Lenza RFS, Oréfice RL, Reis EF, Lobato ZP. Sol–gel silica based networks with controlled chemical properties. J Non-Cryst 2000;273:109–115.
53. Tan A, Romanska HM, Lenza R, Jones J, Hench LL, Polak JM, Bishop AE. The effect of 58S bioactive glass sol–gel

derived foams on the growth of murine lung epithelial cells. Key Eng Mat 2003;240–242: 719–724.

References List

Clifford A, Hill R, Rafferty A, Mooney P, Wood D, Samuneva B, Matsuya S. The influence of calcium to phosphate ratio on the nucleation and crystallization of apatite glass-ceramics. J Mater Sci Mater Med 2001;12(5): 461–469.

Healy KE. Molecular engineering of materials for bioreactivity. Curr Op Sol 1999;4: 381–387.

See also BIOMATERIALS FOR DENTISTRY; BONE AND TEETH, PROPERTIES OF; HEART VALVE PROSTHESES; HIP JOINTS, ARTIFICIAL.

BIOMATERIALS: CARBON

ROBERT B MORE
RBMore Associates,
Austin, Texas

JACK C BOKROS
Medical Carbon Research Institute
Austin, Texas

INTRODUCTION

Inorganic, elemental carbon is one of the oldest, and yet newest, biomaterials. Carbon utilization began with prehistoric human's use of charcoal and continues today with a variety of applications exploiting the physicochemical, adsorptive, structural, and biocompatible properties of different forms of carbon. To date, the most important carbon biomaterials have been the isotropic pyrolytic carbons (PyC), produced in a fluidized bed, for use as structural blood contacting components for heart valve prostheses and for small joint orthopedic prostheses. Adsorptive properties of activated carbons also find widespread use for the removal of toxins from the body either by direct ingestion, dialysis, or by plasmapherisis.

Other carbons, such as carbon fibers and glassy carbons have been proposed for use in a variety of structural implants, but because of limited strength and durability, have not been generally accepted. However, carbon fibers and glassy carbons are used as electrodes and electronic components in biomedical analytical devices. Diamond-like coatings have been proposed to provide enhanced wear resistance for large orthopedic components, but this technology is still under development. For the future, carbon holds a central focus in nanotechnology with investigations into the use of fullerenes and carbon nanotubes as means of imaging and manipulating nanoscale bioactive molecules, as selective markers, and perhaps as inhibitors to virulent organisms such as the human immunodeficiency virus (HIV).

Elemental carbon is allotropic, meaning that it can exist in two or more forms (1). There are at least two perfectly crystalline allotropic forms: graphite and diamond, and a myriad of intermediate, imperfectly crystalline, amorphous structures (2). This diversity in structure leads to considerable variability in physical and mechanical properties ranging from graphite, one of the softest materials, to diamond, the hardest material known to human. Thus, carbon rather than being a single material is actually a spectrum of materials (3). For this reason, it is necessary to qualify the use of the term *carbon* as designating a generic material with a carbon elemental composition. A specific carbon material must then be qualified with a description of it's structure.

In general, most of the pure carbons are biocompatible in that they are bioinert, do not provoke thrombosis, hemolysis, inflammatory response, nor activate the complement system (4). Furthermore pure carbons are biostable: toxic products are not generated and the materials retain their properties. However, just because a candidate material is a carbon does not mean that its particular microstructure and properties are appropriate for the desired application. For example, structural applications such as cardiovascular and orthopedic prostheses require strength, fatigue resistance, wear resistance, low friction and durability, in addition to tissue compatibility (3). Not all carbons have the appropriate properties needed for structural use.

In order to appreciate the medically important carbons, some of the various forms of elemental carbon, their synthesis, structure, and properties will be presented and briefly discussed. We will then return to the important carbon biomaterials, discuss their utilization, and conclude with speculations as future directions.

BACKGROUND

Structure of Carbons

Diversity in carbon arises from the electronic configuration: $1s^2 2s^2 2p^2; {}^3P$, which allows the formation of a number of hybridized atomic orbitals that share four valence electrons to form covalent bonds with directional properties. On the basis of bond structures that arise from the hybridized orbital bonds, carbon compounds are classed as aliphatic and as aromatic (5). Originally, aliphatic meant "fatty" and aromatic meant "fragrant", but these descriptions no longer have any real significance. Aliphatic compounds are further subdivided into alkane, alkenes, alkynes, and cyclic aliphatic. Aromatic compounds are benzenes and compounds that resemble benzene in chemical behavior. With a few exceptions, organic compounds of medical importance tend to be aromatic or benzene-like. Details of electronic structure beyond that given here may be found in standard chemistry and organic chemistry textbooks (1,5).

Naturally Occurring Carbons

Diamond. Diamond is the ultimate polycyclic aliphatic system, but is not a hydrocarbon; rather, it is one of the allotropic forms of elemental carbon. In the diamond allotropic structure, one s and three p orbitals undergo hybridization to form the sp^3 orbital that has tetrahedral symmetry. This symmetry allows covalent bonds to connect each carbon atom with four others. Bond angles are 109.5° and the carbon–carbon bond length is 0.154 nm. Each carbon is bonded to the original plus three others and this structure propagates throughout the entire crystal forming one giant isotropic molecule of covalently bonded carbons (1,2), as shown in Fig. 1. The diamond crystallographic

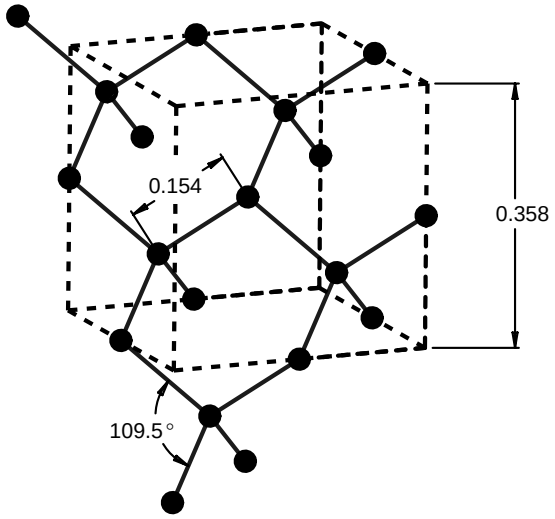


Figure 1. Crystallographic structure of diamond with tetrahedral bond angles of 109.5° and bond lengths of 0.154 nm. The unit cell with a length of 0.358 nm is shown by the dashed lines. The spheres represent the location of the atoms and not size.

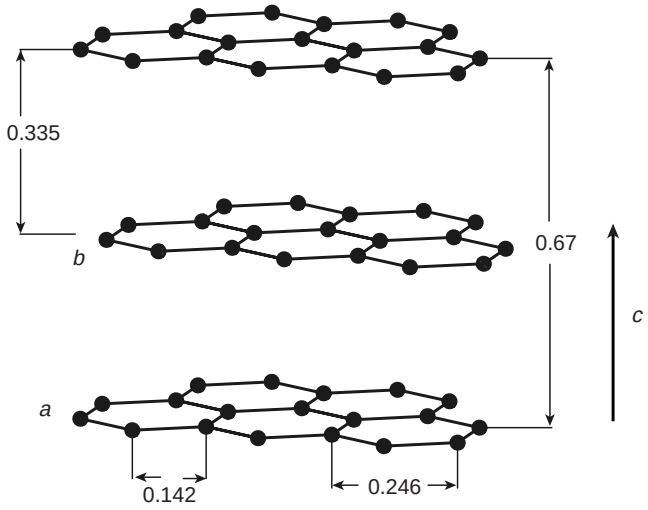


Figure 2. Crystallographic structure of graphite. Basal planes *a* and *b* contain the hexagonal covalently bonded carbons with bond angles of 120° and bond lengths of 0.142 nm. Because of sp^2 coordination, each basal plane is shifted one atomic position relative to one another. The successive basal planes are separated by 0.34 nm in the *c* direction. The distances 0.246 and 0.67 nm are the dimensions of the graphite hexagonal close-packed unit cell.

structure can be visualized as a repetition of the six-carbon cyclohexane "chair" configuration. Because of the large number covalent bonds with an interlocking isotropic orientation, the structure is very rigid. A large amount of energy is required to deform the crystal, hence diamond is the hardest material known.

Graphite. Where diamond is the ultimate polycyclic aliphatic system, graphite is the ultimate polycyclic aromatic system. Graphite has a layered structure consisting of planar arrays in which each carbon atom is bonded by two single bonds and one double bond with its three nearest neighbors. Where diamond can be visualized as a repeated cyclohexane chair, graphite is visualized as a repeated six-carbon benzene ring structure. Within a single plane, each carbon is bonded with a single atomic distance of 0.142 nm to its three nearest neighbors by sp^2 orbitals with hexagonal symmetry and 120° bond angles (1). Three of the four valence electrons are used to form these regular covalent σ (sigma) bonds, which forms the basal planes of hexagonal covalently bonded atoms as shown in Fig. 2. A single basal layer of the hexagonal carbons is known as a *graphene* structure.

The fourth π (pi) electron resonates between valence structures in overlapping *p* orbitals forming π bond donut-shaped electron clouds with one lying above and one below and perpendicular to the plane of the σ bonded carbons (2). Successive layers of the hexagonal carbons are held together at a distance of 0.34 nm by weak van der Waals forces or by interactions between the π orbitals of the adjacent layers (6,7). Thus the graphite structure is highly anisotropic, consisting of stacked parallel planes of strong covalent in-plane bonded carbons with the planes held together by much weaker van der Waals type forces. Because of weak interlayer forces, the layers are easily separated, which accounts for softness and lubricity of graphite. These weak interlayer forces also account for (a) the tendency of graphitic materials to fracture along planes, (b) the formation of interstitial compounds; and (c) the lubricating, compressive, and many other properties of graphite (2,6).

Amorphous Carbons. There are many cystallographically disordered forms of carbon with structures that are intermediate between those of graphite and diamond. The majority tends to be imperfectly layered graphene, turbostratic, and randomly oriented structures (2). X-ray diffraction patterns for amorphous carbons are broad and diffuse because of the small crystallite size, imperfections, and a turbostratic structure (2). In turbostratic structures, there is order within the graphene planes (denoted as *a* and *b*), but no order between planes (denoted as *c* direction) as shown in Fig. 3. Crystallographic defects such as lattice vacancies, kinked or warped layer planes, and possible aliphatic bonds tend to increase the turbostratic layer spacing relative to graphite and inhibit the ability of the layer planes to slip easily as occurs in graphite (2). Like graphite, there is strong in plane covalent bonding, but, because the ability of the planes to slip past one another is inhibited, the materials are much harder and stronger than graphite. Turbostratic carbons occur in a spectrum of amorphous ranging through mixed-amorphous structures and include materials such as soot, pitch, and coals.

Fullerenes. The recently discovered fullerenes (2,8,9) can occur naturally as a constituent of soot. Fullerenes are hollow cage-like structures that can be imagined as graphene sheets that have been folded or rolled into a ball or cylindrical tube. However, the structures are actually formed by the reassociation of individual carbon atoms rather than a folding or rolling of a graphene structure. The most famous fullerene is the ~1 nm diameter C_{60}

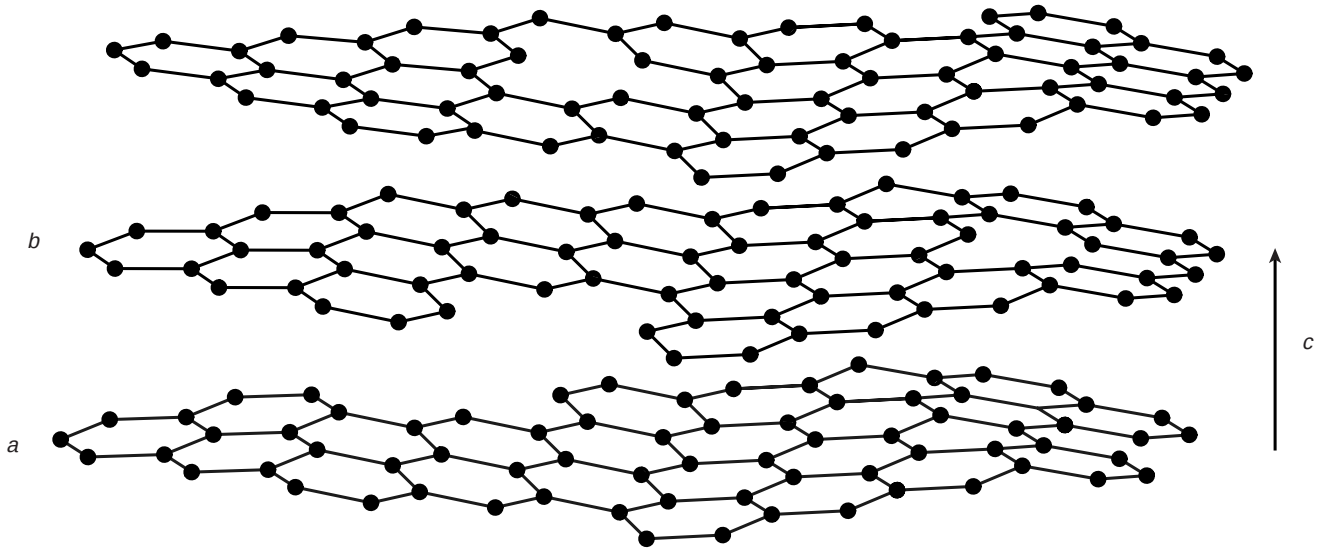


Figure 3. Turbostratic amorphous structure.

(60 carbon) buckministerfullerene (bucky ball) with a truncated icosahedron structure that resembles a European football. Because the structure is reminiscent of the geodesic dome designed by the architect Buckminister Fuller, the proposed structure was named after him (8).

Geometrically, the bucky ball has a repeating structure that consists of a pentagon surrounded by five hexagons (see Fig. 4). In order to wrap into a nonplanar ball, the graphitic p orbitals must assume an angle of 101.6° relative to the plane of the C bonds rather than 90° for graphite (9). There are a number of other possible carbon number ball structures, but the smallest sizes are thought to be limited to C_{60} and C_{70} by the molecular strain induced at the edge-sharing pentagons. Although remarkably stable, C_{60} can photodisassociate when pulsed with laser light and loose carbon C_2 pairs down to $\sim C_{32}$, where it explodes into open fragments because of strain energy (10).

Metals can also be inserted into the buckyball cage simply by conducting fullerene synthesis in the presence of metals (11). Such internally substituted endohedral fullerenes are fancifully called "dopyballs" for doped fullerenes (12) and denoted as M_aC_n, where M_a is the metal and C_n the carbon complex. "Fullerite" refers to a solid-state association of individual C_{60} molecules, named by analogy to graphite, in which the bucky balls assume a face-centered cubic (fcc) crystallographic structure with lattice constant $a = 1.417$ nm (13). Treatment of fullerite with 3 equiv of alkali metal, A_3C_{60}, makes it a super conductor at room temperature (14), whereas treatment with 6 equiv of alkali metal, A_6C_{60}, makes it an insulator.

An excellent introduction to fullerenes by Bleeke and Frey, Department of Chemistry, Washington University, St. Louis, MO, is available on the Internet at http://www.chemistry.wustl.edu/~edudev/Fullerene/fullerene.html (15).

Nanotubes. Although most likely synthetic, because of the basic fullerene structure, nanotubes will be discussed here. A nanotube consists of a single graphene sheet SWNT (single-wall nanotube) or multiple concentric graphene sheets MWNT (multiwall nanotube) rolled into a cylindrical tube (16). In MWNT, the nested concentric cylinders are separated by the ~ 0.34–0.36 nm graphite layer separation distance.

There are several different wrapping symmetries to give chiral, zigzag or arm chair nanotubes and the tubes may be end capped by a bucky ball half-sphere. Lengths range from well > 1 μm and diameters range from 1 nm for SWNT to 50 nm for MWNT. A zigzag SWNT is shown in Fig. 5. Additional information regarding nanotubes can be found at Tomanek's laboratory, at the University of Michigan. A very informative web page dedicated to nanotubes (17) is at http://www.pa.msu.edu/cmp/csc/nanotube.html.

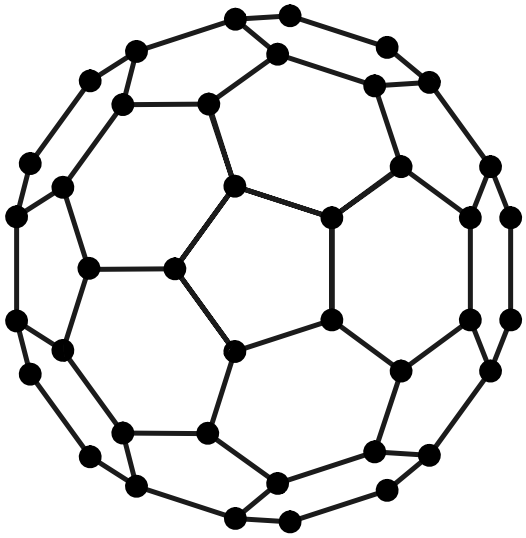


Figure 4. A surface view of a C_{60} structure, buckministerfullerene (buckyball), with an ~1 nm diameter, is shown.

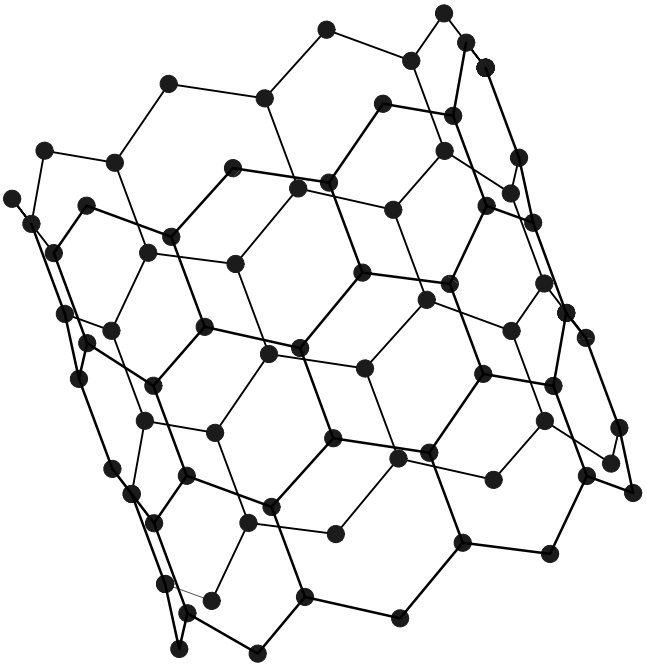


Figure 5. A short section of SWNT with a zigzag chiral symmetry is shown. The arrow indicates the long axis of the tube and bonds on the forward wall are more heavily drawn. Like the buckyball, this SWNT has a diameter of ~1 nm.

Synthetic Carbons

Carbon structures can be synthesized through a variety of processes. Because these processes define the resulting materials, both will be presented together. The most important synthesis processes include carbonization or pyrolysis and graphitization (2). Carbonization is a thermal process in which an organic precursor is converted into an all carbon residue with the diffusion of non-carbon volatile compounds to the atmosphere (2,6). The resulting all-carbon residue is known as a coke or a char.

Coke is a graphitizable carbon and chars are nongraphitizing (2). Cokes and chars are amorphous, lacking long-range crystallographic structure (turbostratic), with the degree of structure dependant on the precursor and the particular carbonization process. A coke may then be graphitized. In graphitization, residual non-carbon impurities are removed and the turbostratic structure is converted into a well-ordered graphite crystallographic structure by heating to high temperatures (6). A char, when graphitized retains its disordered turbostratic structure (2).

Synthetic Graphites. These carbons are prepared by grinding or milling a solid precursor material (coke) into fine particles, binding with a material such as coal tar pitch, and then molding into shape (2,6). The resulting material is then carbonized baked and graphitized. Typical milled grain sizes may range from 1 μm up to ~1 cm. The mixture of filler and binding may be doped or impregnated with the addition of non-carbon elements such as tungsten. The final properties of the molded graphite depend on the degree of graphitization and the grain size (6). Other important parameters are porosity, anisotropy, and density (6). One synthetic graphite used in medical devices, POCO AXF 5Q, has a grain size of 5 μm, a pore size of 0.6 μm, and a 23% volume total porosity. This particular graphite grade is often mixed with 10% by weight fine powdered tungsten before molding and baking to confer radio opacity (6).

Viterous, Glassy Carbons. Carbonization of certain polymer chars produces glassy carbons. These materials are amorphous, turbostratic, and thought to contain some sp^3 character in addition to sp^2(2). Precursors are polymers such as phenolformaldehyde, poly(furfuryl alcohol) and poly(vinyl alcohol). Shapes are attained by carbonization in molds and are limited to ~7 mm thickness because of volumetric shrinkage (~50%) and the need for gases generated during carbonization to diffuse out and not nucleate bubbles (18). The resulting material is hard, brittle, and difficult to machine.

Carbon Fibers. Thomas Edison produced the first carbon fiber in 1879 as a filament of an incandescent lamp and the first patent was issued in 1892 (2,3). Hiram Maxim received a process patent for the production of carbon fibers in 1899 (3). However, prior to the 1950s carbon fibers were of marginal strength and used primarily for their electrical properties.

Carbon fibers are highly oriented, small (with diameters on the order of 7 μm), crystalline filaments that are prepared by carbonization of polymeric filament precursors and sequential heat treatment. There are three classes of fibers based on the precursor material: PAN (polyacrylonitrile), rayon, and pitch (2). Other precursors and processes exist, but have not been as successful commercially (3). In general, fibers are classified according to structure and degree of crystallite orientation (2): high modulus (345 GPa and above), intermediate modulus (275 GPa), and low modulus (205 GPa and below). Structures are turbostratic and can contain mixed sp^2 and sp^3 bonding (2). Because of their small volume, tensile strengths can be quite high, on the order of 1000–7000 MPa.

Chemical Vapor Deposited (CVD) Carbons. Carbonization of a gaseous or liquid precursor such as gaseous hydrocarbons produces a material known as pyrolytic carbon or pyrolytic graphite (2). Thermal decomposition of hydrocarbons produces carbon free radicals in the vapor phase, which can then polymerize to form coatings on exposed surfaces. Common precursor hydrocarbons are methane, propane, and acetylene. The resulting turbostratic pyrolytic carbons can be isotropic or anisotropic depending on the pyrolysis reaction conditions (19). The coating process can be prolonged so as to produce structural components for heart valve and orthopedic prostheses with coating thickness on the order of 1 mm.

The pyrolytic carbons for medical applications are formed by CVD processes in fluidized-bed reactors (20). Propane is the precursor gas and an inert gas such as nitrogen or helium provides the levitation needed to fluidize a bed of small refractory particles. Graphite preformed substrates (e.g., POCO AXF 5Q) suspended in the fluidized

bed are coated with the pyrolytic carbon (20). The resulting coating structures are turbostratic and isotropic with very small randomly oriented crystallites: These crystallites will henceforth be designated as *isotropic fluidized-bed pyrolytic carbons*. Nonfluidized-bed CVD reactors tend to produce a highly anisotropic coating with column-like, (columnar) crystallites or laminar structures with the basal planes oriented parallel to the deposition surface (2,19).

Highly Oriented Pyrolytic Graphite (HOPG). Columnar and laminar pyrolytic carbons when annealed $>2700\,^{\circ}C$ are reordered, the turbostratic imperfections disappear and the resulting structure closely approaches the ideal graphite structure with an angular spread of the crystallite *c* axes of $<1^{\circ}$ (2).

Vapor-Phase Carbons. Carbon CVD coatings formed from solid precursors carbonized by vaporization are considered vapor-deposited coatings (VPC). Precursors can be graphite or vitreous carbon vaporized by heating to high temperature at low pressure to generate the carbon free radicals. This technique produces line-of-sight coatings of nanometer and micrometer level thickness. The VPC coatings tend to be turbostratic and amorphous (3).

Diamond-Like Carbon. Diamond-like carbon coatings containing mixed sp^3 and sp^2 bonds can be prepared by physical vapor deposition (PVD). These PVD methods produce carbon free radicals by ion beam sputtering, or laser or glow discharge of solid carbon targets. There are also mixed PVD/CVD methods such as plasma or ion beam deposition from hydrocarbon gas precursors (2).

Activation. Activated carbons have large surface areas and large pore volumes that lend to a unique adsorption capability (21). Activation is a thermal or chemical treatment that increases adsorption capability. The mechanisms for adsorption are complex and include physical and chemical interactions between the carbon surface and the sorbed substances. Activity includes (a) adsorption, (b) mechanical filtration, (c) ion exchange, and (d) surface oxidation (22). Of these, adsorption and surface oxidation are the most important for medical applications. Incompletely bonded basal plane carbons as occur at crystal edges exposed at the surface, as well as defects, are chemically active and can chemisorb substances, particularly oxidizing gases such as carbon monoxide and carbon dioxide (23). Surface oxidation involves the chemisorbance of atmospheric oxygen and further reaction of the sorbed oxygen with other substances (24). Physical adsorbance occurs because of charge interactions, and chemical adsorbance occurs because of reactions between the adsorbant and adsorbate (24).

Any high carbon material can be "activated" by various oxidizing thermal and chemical processes that increase porosity and active surface area, which increases the ability for chemisorption (25). A char is formed and then treated chemically or physically to generate pores and the surface oxidized (21). Surface oxide complexes such as phenols, lactones, carbonyls, carboxylic acids, and quinones form that have a strong affinity for adsorbing many substances such as toxins or impurities (26). Carbon fibers may be activated in order to enhance the ability to bind with a matrix material when used as a filler.

PROPERTIES

Representative physical and mechanical properties of the carbon allotropes are summarized in Table 1 (2,3,27). Materials included span the spectrum from natural diamond to natural graphite. There is considerable variability

Table 1. Representative Mechanical and Physical Properties for Carbon Allotrophs

Property	Natural Diamond	Amorphous Carbons	HOPG	Natural Graphite
Density, $g \cdot cm^{-3}$	3.5–3.53	1.45–2.1	2.25–2.65	2.25
Young's modulus, GPa	700–1200	17–31	20	4.8
Hardness, mohs	10	2–5		1–2
Hardness, DPH 500 g		150–(>230)		
Flexural strength, MPa		175–520	80 (*c*) 120 (*ab*)	
Compressive yield strength, MPa	8680–16530	700–900	100	
Fracture toughness, $MPa \cdot m^{1/2}$	3.4	0.5–1.67		
Poisson's ratio	0.1–0.29	0.2–0.28		
Wear resistance	Excellent	Poor to excellent	Poor	Poor
Electrical resistivity, Ω·cm	2700		0.15–0.25 (*c*) 3.5×10^{-5}–4.5×10^{-5}(*ab*)	0.006
Magnetic susceptibility, $\times 10^6$ emu/mol	−5.9			−6
Melting point, °C	3550		3650	3652–3697 (sublimes)
Boiling point, °C	4827			4220
CTE linear, $(20^{\circ}C)\mu m \cdot (m \cdot {}^{\circ}C)^{-1}$	1.18	2.6–6.5	−0.1 (*ab*) 20 (*c*)	0.6 (*ab*) 4.3 (*c*)
Heat capacity, J/g·°C	0.4715			0.69
Thermal conductivity, $W \cdot (m \cdot K)^{-1}$	2000	4.6–6.3	16–20 (*ab*) 0.8 (*c*)	19.6 (*ab*) 0.0573 (*c*)

[a]Values from Matweb.com and from Refs. (2,3).

in properties depending on the structure, anisotropy, and crystallinity, particularly in the amorphous carbons. Physical properties such as resistivity, coefficient of thermal expansion, thermal conductivity, and tensile strength (28) show profound sensitivity to direction in the graphitic materials. This anisotropy is most easily seen in HOPG by comparing the *ab* direction, parallel to the σ-bonded basal plane, to the perpendicular *c* direction. For example, the resistivity drops for HOPG because of the mobility of the π-electron clouds in the *ab* direction relative to the *c* direction (2). Diamond, with full covalent bonding, is an insulator.

Thermal conductivity, which occurs by lattice vibration, is related to a mean-free-path length for wave scattering. Little scattering occurs in the near-perfect graphite crystal basal plane, so the scattering path length and thermal conductivity are high in the *ab* direction. In the *c* direction, thermal conductivity is much lower because the amplitude of lattice vibration is considerably lower than for the *ab* direction (2). Thermal expansion is related to the interatomic spacing of the carbon atoms, bond strength, and vibration. As temperature increases, the atoms vibrate and the mean interatomic spacing increases. For weak bonding in the *c* direction, the interatomic vibrational amplitude and dimensional changes are larger than for the strongly bonded *ab* direction (2). The CTE values are stated for room temperature to $\sim 200\,^{\circ}C$; the negative values shown in Table 1 are possibly due to internal stresses and become positive at higher temperatures. Large anisotropic differences in CTE can result in large internal stresses and possible structural problems with heating and cooling over large temperature differences.

BIOCOMPATIBILITY

Pyrolytic carbons used in heart valve and orthopedic prostheses have a successful clinical experience as long-term implant materials for blood and skeletal tissue contact (3,29–31). These isotropic, fluidized-bed, pyrolytic carbons that were originated at General Atomics in the 1960s demonstrate negligible reactions in the standard Tripartite and ISO 10993-1 type biocompatibility tests. Results from such tests are given below in Table 2 (20). This material is so nonreactive that it has been proposed as a negative control for these tests. However, the surface is not totally inert and is capable of adsorption and desorption of a variety of substances including protein (32–39). The mechanism for biocompatibility is yet poorly understood, but probably consists of a complex, interdependent, and time-dependent series of interactions between the proteins and the carbon surface (32).

Because of the similarity in surface sp^2 and sp^3 character among the various pure carbons, most can be expected to have the tissue compatibility and biostability to perform well in these biocompatibility tests also. Vitreous carbons (40), activated carbons, and diamond-like coatings (41) are known to exhibit tissue compatibility, likewise the fullerenes will probably be found tissue compatible. As an extreme example, in testing the safety of an activated charcoal for hemoperfusion, Hill (42) introduced finely ground charcoal suspensions into the blood stream of rats in varying concentrations up to 20 mg/kg charcoal and observed no differences in survival or growth relative to controls over a 2-year observation period.

A reasonable working definition for biocompatibility has been given by Williams (43) as, "*The ability of a material to perform with an appropriate response in a specific application*". The important point here is that while many carbons provoke a minimal biological reaction, "*the specific application*" demands a complete set of mechanical and physical properties, in addition to basic cell compatibility. Because there are a number of possible microstructures, each with different properties, a given carbon will probably not have the entire set of properties needed for a specific application. Historically, the clinically successful isotropic, fluidized-bed, pyrolytic carbons required extensive development and tailoring to achieve the set of mechanical and physical properties needed for long-term cardiovascular and orthopedic applications (20,30–32).

Blood compatible glassy carbons, for example, are often proposed for use in heart valves. However, glassy carbons were evaluated in the early 1970s as a replacement for the polymer Delrin in Bjork–Shiley valve occluders and actually found to have inferior wear resistance and durability relative to the polymer (44). Thus, the fact that a material is carbon, a turbostratic carbon, or a pyrolytic carbon and is cell compatible, does not justify its use in a long-term implant devices (3,32,33). The entire range of physical and mechanical properties as dictated by the intended application are required.

MEDICAL APPLICATIONS

Activated Charcoal–Activated Carbons

Charcoal, the residue from burnt organic matter, was probably one of the first materials used for medical and

Table 2. Biological Testing of Pure PyC

Test description	Protocol	Results
Klingman maximization	ISO/CD 10993-10	Grade 1; not significant
Rabit pyrogen	ISO/ DIS 10993-11	Nonpyrogenic
Intracutaneous injection	ISO 10993-10	Negligible irritant
Systemic injection	ANSI/AAMI/ISO 10993-11	Negative—same as controls
Salmonella typhimurium Reverse mutation assay	ISO 10993-3	Nonmutagenic
Physiochemical	USP XXIII, 1995	Exceeds standards
Hemolysis–rabbit blood	ISO 10993-4/NIH 77-1294	Nonhemolytic
Elution test, L929 mammalian cell culture	ISO 10993-5, USP XXIII, 1995	Noncytotoxic

biocompatible applications. Prehistoric humans knew that pulverized charcoal could be placed under the skin indefinitely without ill effects, thus allowing decorative tattoos (45). Because granulated charcoal has an active surface area, it can adsorb toxins when ingested. Likewise, charcoal has long been used to clear water and other foods. The ancient Egyptians first recorded the medical use of charcoal $\sim$1500 BC (21). During the 1800s, the first formal scientific studies of charcoal as an antidote to treat human poisoning appeared in Europe and The United States. In some of these studies, the researchers demonstrated charcoals effectiveness by personally ingesting charcoal along with an otherwise fatal dose of strychnine or arsenic (21). Activation was discovered $\sim$1900 and activated charcoals were used as the sorbant in World War I gas masks (21).

Today's activated carbons or activated charcoals are derived from a number of precursor organic materials ranging from coal, wood, coconut shells, and bone. Chars are prepared by pyrolyzing the starting organic material using heat in the absence of oxygen. The char is then activated by treatment with chemicals or steam. Activated carbon has remarkable adsorptive properties that vary with the starting material and activation process. Common active surface areas are on the order of 1000–2000 m^2/g. Prior to the discovery of activation processes, charcoals were naturally oxidized by exposure to the atmosphere and moisture, as in charcoal, or oxidized in a more controlled activating process (46).

Orally administered activated carbon applications include use as an antidote to poisoning and to drug overdoses, where it acts at the primary site of drug adsorption in the small intestine. There are no contraindications for patients with intact GI tracts. There are numerous advantages and few disadvantages. One of the main disadvantages is that it is unpleasant for the health care professional to use because it can be messy, staining walls, floors, clothing, and so on. It may also be unpleasant to swallow because of a gritty texture (46).

There are extracorporal, parenteral, methods such as hemoperfusion, hemofilteration, and plasmapheresis where activated carbon is used to remove toxins from a patient's blood. The patient's heparinized blood is passed via an arterial outflow catheter into an extracorporal filter cartridge containing the activated carbon and then returned to the patient via a venous catheter. These techniques are effective when there is laboratory confirmation of lethal toxin concentrations in the blood and for poorly dialyzable and nondialyzable substances (47).

Pyrolytic Carbons

Isotropic, fluidized-bed PyCs, appropriate for cardiovascular applications originated at General Atomics in the late 1960s as a cooperative effort between an engineer, Jack Bokros, working with pyrolytic carbons as coatings for nuclear fuel particles and a surgeon, Vincent Gott, who was searching for thromboresistant materials for cardiovascular applications (48). Together, they tailored a specific fluidized-bed, isotropic pyrolytic carbon alloy with the biocompatibility, strength and durability needed for long-term structural applications in the cardiovascular system. The original material was a patented silicon-alloyed pyrolytic carbon given the tradename "Pyrolite" (20).

In the early 1960s, heart valve prostheses constructed from polymers and metal were prone to early failure from wear, thrombosis, and reactions with the biological environment. Prosthesis lifetimes were limited to several years because of wear in one or more of the valve components. Incorporation of PyC as a replacement for the polymeric valve components successfully eliminated wear as an early failure mechanism. Subsequently, in most valve designs, metallic materials were replaced with PyC also (20,29–33,49).

During the 1970s and 1980s Pyrolite was only available from a single source until the original patents expired. Since that time, several other sources have appeared with copies of the original silicon-alloyed General Atomics material. In the early 1990s, the Bokros group revisited the synthesis methods and found that with the then available technology for process control, that a pure carbon pyrolytic carbon could be made with better mechanical properties and potentially better biocompatibility than the original silicon-alloyed Pyrolite (20). This new pure isotropic, fluidized-bed, pyrolytic carbon material was patented and named On-X carbon. On-X carbon is currently utilized in mechanical heart valves and in small joint orthopedic applications.

These PyC materials are turbostratic in structure and isotropic with fine randomly oriented crystallite sizes on the order 2.5–4.0 nm and c layer spacing of $\sim$0.348 nm (50–52). Implants are prepared by depositing the hydrocarbon gas precursor coating in a fluidized bed on to a preformed graphite substrate to a thickness of $\sim$0.5 mm (29–32,53). The coatings then may be ground and polished if desired and subjected to a proprietary process that minimizes the degree of surface chemisorbed oxygen.

Some of the mechanical and physical properties of the pure and silicon-alloyed PyC materials appropriate for use in long-term implants are given Table 3 (3,20). A typical glassy carbon and a fine-grained synthetic graphite are also included for comparison. The PyC flexural strength, fatigue, and wear resistance provide adequate structural integrity for a variety of implant applications. The density is low enough to allow components to be actuated by flowing blood. Relative to orthopedic applications, Young's modulus is in the range reported for bone (54,55), which allows for compliance matching and minimizes stress shielding at the prosthesis bone interface. The PyC strain-to-failure is low relative to ductile metals and polymers; but it is high relative to ceramics. Because PyC is a nearly ideal linear elastic material, component design requires the consideration of brittle material design principals. Certain properties such as strength vary with the effective stressed volume, or stressed area as predicted by Weibull theory (56). Table 3 strength levels were measured for specimens tested in four-point bending, third-point loading (57) with an effective stressed volume of 1.93 mm^3. The Weibull modulus for PyC is $\sim$10 (57).

Fluidized-bed isotropic PyCs are remarkably fatigue resistant. There is strong evidence for the existence of a fatigue threshold that is very nearly the single cycle

Table 3. Biomedical Fluidized-Bed Pyrolytic Carbon Properties

Property	Pure PyC	Typical Si-Alloyed PyC	Typical Glassy Carbon	POCO Graphite AXF-5Q
Flexural strength, MPa	493.7 ± 12	407.7 ± 14.1	175	90
Young's modulus, GPa	29.4 ± 0.4	30.5 ± 0.65	21	11
Strain-to-failure, %	1.58 ± 0.03	1.28 ± 0.03		0.95
Fracture toughness, $MPa \cdot \sqrt{m}$	1.68 ± 0.05	1.17 ± 0.17	0.5–0.7	1.5
Hardness, DPH, 500 g load	235.9 ± 3.3	287 ± 10	150	120
Density, $g \cdot cm^{-3}$	1.93 ± 0.01	2.12 ± 0.01	< 1.54	1.78
CTE, $\mu m \cdot cm^{-1}$ EC	6.5	6.1		7.9
Silicon content, %	0	6.58 ± 0.32	0	0
Wear resistance	Excellent	Excellent	Poor	Poor

fracture strength (58–60). Paris-law fatigue crack propagation rate exponents are high; on the order of 80 and da/dN fatigue crack propagation testing displays clear evidence of a fatigue crack propagation threshold (58–63).

Crystallographic mechanisms for fatigue crack initiation and damage accumulation are not significant in the PyC at ambient temperatures (59,61). There have been no clear instances of fatigue failure in a clinical implant during the accumulated 30-year experience (64). Less than 60 out of >4 million implanted PyC components have fractured (65), and these were caused by damage from handling or cavitation (66–68).

The PyC wear resistance is excellent. Wear testing performed in the 1970s identified titanium alloy, cobalt chromium alloy, and PyC as low wear contact materials for use in contact with PyC (69,70). This study determined that wear in PyC occurred due to an abrasive mechanism and interpreted wear resistance as approximately proportional to the ratio $H^2/2E$, where H is the Brinell hardness number and E is Young's modulus. This criteria is related to the amount of elastic energy that can be stored in the wearing surface (70). The greater the amount of stored energy, the greater the wear resistance. Successful low wearing contact couples used for mechanical heart valves include PyC against itself, cobalt chromium alloy, and ELI titanium alloy.

Observed wear in retrieved PyC mechanical heart valve prosthesis implant components utilizing PyC coupled with cobalt chromium alloy is extremely low with PyC wear mark depths of <2 μm at durations of 17 years (71–73). Wear in the cobalt chromium components was higher, 19 μm at 12 years (71–73). But, wear in the cobalt chromium components was concentrated at fixed contact points instead of being distributed over a large area as for the PyC rotating disk. Wear depths in all PyC prostheses, with fixed contact points are also low, <3.5 μm at 13 years (74,75). In contrast, the wear depths in valves utilizing polymeric components such as the polyacetyl Delrin in contact with cobalt chromium and titanium alloys are much higher at 267 μm at 17 years (76). Incorporation of PyC in heart valve prostheses has eliminated wear as a failure mode (29,77).

The PyC is often used in contact with metals and behaves as a noble metal in the galvanic series. Testing using mixed potential corrosion theory and potentiostatic polarization has determined that no detrimental effects occur for PyC coupled with titanium or cobalt–chrome alloys (78,79). Use of PyC with stainless steel alloys is not recommended.

To date, PyC has been used in ~ 25 mechanical heart valve prosthesis designs. One such design, the On-X valve, by Medical Carbon Research Institute, http://www.mcritx.com, is shown in Fig. 6.

Pyrolytic carbon has a good potential for orthopedic applications because of advantages over metallic alloys and polymers (3,30,31):

1. A modulus of elasticity similar to bone to minimize stress shielding.
2. Excellent wear characteristics.
3. Excellent fatigue endurance.
4. Low coefficient of friction.
5. Excellent biocompatibility with bone and hard tissue.
6. Excellent biocompatibility with cartilage.
7. Fixation by direct bone apposition.

A brief comparison of PyC properties to those of conventional/orthopedic implant materials is given in Table 4. Pyrolytic carbon coatings for orthopedic implants can reduce wear, wear particle generation, osteolysis aseptic loosening, and thus extend implant useful lifetimes. Furthermore, good PyC compatibility with bone and the native joint capsule enables conservative hemiarthroplasty replacements as an alternative to total joint replacement.

Cook et al. (80) studied hemijoint implants with a PyC femoral head in the canine hip and observed a greater potential for acetabular cartilage survival in PyC than for cobalt–chromium–molybdenum alloy and titanium alloy femoral heads. There were significantly lower levels of gross acetabular wear, fibrillation, eburnation, glycosaminoglycan loss, and subchondral bone change for PyC than the metallic alloys.

Tian et al. (81) surveyed *in vitro* and clinical *in vivo* PyC orthopedic implant studies conducted during the 1970s through the early 1990s and concluded that PyC demonstrated good biocompatibility and good function in clinical applications.

A 10-year follow-up of PyC metacarpophalangeal (MCP) finger joint replacements implanted in patients at the Mayo Clinic, Rochester Minnesota (82) between 1979 and 1987, demonstrated excellent performance. Ascension Orthopedics was able to use these results in part to justify a FDA premarket approval application (PMA) for the semiconstrained, uncemented MCP finger joint replacement, PMA P000057, Nov. 2001.

Figure 6. On-X prosthetic heart valves manufactured by Medical Carbon Research Institute from the elementally pure, fluidized-bed isotropic pyrolytic carbon, On-X carbon. The valves consist of a central flow circular orifice with two semicircular occluder disks. A polymeric sewing cuff is used to attach the valve to the annulus tissue. Aortic and mitral valves with two different sewing cuff designs each are shown.

Currently, Ascension Orthopedics, http://www.ascensionortho.com, manufactures PyC prostheses for finger joints: metacarpophalengeal (MCP) and proximal interphalangeal (PIP) in addition to carpometacarpal (CMC) thumb and an elbow radial head (RH) prostheses (see Fig. 7). Because of the excellent PyC compatibility with bone and cartilage, the CMC and radial head are used in hemiarthroplasty directly contacting the native joint capsule and bone. Fixation is by direct bone opposition for all of the prostheses. To date, $\sim$6500 Ascension Orthopedics prostheses have been implanted. Another company, Bioprofile, http://www.bio-profile.com, manufactures hemiarthroplasty PyC prostheses for the wrist: scapoid, scapho-trapezo-trapezoid, trapezium bone, capitate head, and an elbow radial head.

Glassy carbons have been proposed as an attractive low cost alternative for a variety of orthopedic and cardiovascular devices (3). However, because of relatively low strength and poor wear resistance it has not been generally accepted as a suitable material for long-term critical implants. An example of poor glassy carbon durability when used for heart valve components was cited earlier in the text (44).

Carbon fibers are popular as high strength fillers for polymers and other material composites and have been proposed for use in tendon and ligament replacements in addition to orthopedic and dental implants (83–86). Spinal interbody fusion cages using PEEK and carbon fibers (86) are an example of an orthopedic application. However, the ultimate properties of the implant depend largely upon the

Table 4. Material Properties of Orthopedic Materials

Property	Unit	PyC	Al_2O_3	TZP	CoCrMo	UHMWPE
Density	$g \cdot cm^{-3}$	1.93	3.98	6.05	8.52	0.95
Bend strength	MPa	494	595	1000	690, uts	20
Young's modulus, E	GPa	29.4	400	150	226	1.17
Hardness, H	HV	236[a]	2400	1200	496	NA
Fracture toughness, K_{1c}	$MN \cdot m^{-3/2}$	1.68	5	7		
Elongation at failure	%	2	0.15		1	>300
Poisson's ratio		0.28	0.2	0.2	0.3	
$H^2/2E^b$		7.6	12.2		1.8	

[a]The hardness value for PyC is a hybrid definition that represents the indentation length at a 500 g load with a diamond penetrant indentor. Because PyC elastically completely recovers the microhardness indentation a replica material such as a cellulose acetate coating, or a thin copper tape is used to "record" the fully recovered indentation length. Although unusual, this operational definition for hardness is a common practice used throughout the PyC heart valve industry.

[b]Approximate values, there are no exact conversions.

Figure 7. Ascension Orthopedics small joint PyC prostheses for finger joints.

matrix in which the carbon fibers are included and the geometry and orientation of the fiber inclusions (3).

Diamond-like carbon (DLC) coatings may find use as low friction, wear resistant surfaces for joint articulating surfaces in orthopedic implants (87,88). However, the coating thickness is limited to the micrometer level; the technology is still in development and ultimately may not be competitive with the newer ceramic joint replacement materials.

Buckyballs (fullerenes) and carbon nanotubes are cage-like structures that suggest use as a means to encapsulate and selectively deliver molecules to tissues. Because of their nanometer dimensions, fullerenes can potentially travel throughout the body. Some current biomedical applications under study involve functionalizing fullerenes with a number of substances including DNA and peptides that can specifically target, mark, or interfere with active sites on enzymes and perhaps inhibit virulent organisms such as the human immunodeficiency virus (89–93). They may also be used to selectively block ion channels on membranes (94). Fullerenes are synthesized by CVD and PVD techniques and can have a variety of novel properties depending on preparation. Currently, there are production difficulties with separation and isolation of fullerenes from the rest of soot-like materials that can occur during synthesis. However, bulk separation methods have been developed and some commercial sources have appeared. See http://www. chemistry.wustl.edu/~edudev/Fullerene/fullerene.html#index and http://www.pa.msu.edu/cmp/csc/nanotube. html. There is a wealth of information available on

the Internet that is readily accessed. Medical applications of fullerenes are currently a topic of intense interest and activity and hold much promise for future developments.

CONCLUSION

Uses of carbon as a biomaterial range from burnt toast, as mother's first aid remedy for suspected poisoning, to the newly discovered fullerene nanomaterials as a possible means to treat disease on a molecular level. The most successful and widespread medical applications have been the use of activated carbons for detoxification and the use of the General Atomics family of isotropic, fluidized-bed, pyrolytic carbons for structural components of long-term critical implants. However, the successful biomedical application of carbon requires an understanding that carbon is a spectrum of materials with wide variations in structure and properties. While a given carbon may be biocompatible, it may not have the mechanical and physical properties needed for the intended application.

As for the future, additional applications of the biomedical PyC materials to orthopedic applications in larger joints and in the spine can be expected, especially if successful long-term hemiarthroplasty devices can be demonstrated. New cardiovascular devices can be expected, such as components for venous shunts and venous valves. The most exciting new developments will probably occur in nanotechnology with the creation of functional, fullerene type, materials, devices, and systems through control of matter at the scale of 1–100 nm, and the exploitation of novel properties and phenomena at the same scale.

BIBLIOGRAPHY

Cited References

1. Pauling L. College Chemistry. San Francisco: W.H. Freeman; 1964.
2. Pierson HO. Handbook of Carbon, Graphite, Diamond and Fullerenes. Park Ridge, New Jersey: Noyes Publications; 1993.
3. Haubold AD, More RB, Bokros JC. Carbons. In: Black J, Hastings G, editors. Handbook of Biomaterial Properties. London: Chapman & Hall; 1998. p 464–477.
4. Janvier G, Baquey C, Roth C, Benillan N, Belisle S, Hardy J. Extracorporeal circulation, hemocompatibility, and biomaterials. Ann Thorac Surg 1996;62:1926–1934.
5. Morrison RT, Boyd RN. Organic Chemistry. Boston: Allyn and Bacon; 1974.
6. Properties and Characteristics of Graphite for the Semiconductor Industry. In: Sheppard RG, Mathes DM, Bray DJ, editors. Decatur, TX: POCO Graphite, Inc.; November 2001. Can be downloaded from http://www.poco.com.
7. Spain IL. Electronic Transport Properties of Graphite, Carbons, and Related Materials. Chem Phys Carbon 1981; 16:119
8. Kroto HW, Heath JR, O'Brien SC, Curl RF, Smalley RE. C_{60}: Buckminsterfullerene. Nature (London) 1985;318(6042): 162–163.
9. Haddon RC, Brus LE, Raghavachari K. Electronic Structure and Bonding in Icosahedral C_{60}. Chem Phys Lett 1986; 125:459.
10. O'Brien SC, Heath JR, Curl RF, Smalley RE. Photophysics of Buckminsterfullerene and Other Carbon Cluster Ions. J Chem Phys 1988;88:220.
11. Heath JR, O'Brien SC, Zhang Q, Liu Y, Curl RF, Kroto HW, Tittel FK, Smalley RE. Lanthanum Complexes of Spheroidal Carbon Shells. J Am Chem Soc 1985;107:7779–7780.
12. Chai Y, Guo T, Jin C, Haufler RE, Chibante LPF, Fure J, Wang L, Alford JM, Smalley RE. Fullerenes with Metals Inside. J Phys Chem 1991;95:7564
13. Heiney PA, Fischer JE, McGhie AR, Romanow WJ, Denenstein AM, McCauley JP, Jr., Smith AB, III, Cox DE. Orientational Ordering Transition in Solid C_{60}. Phys Rev Lett 1991;66:2911.
14. Haddon RC, Hebard AF, Rosseinsky MJ, Murphy DW, Duclos SJ, Lyons KB, Miller B, Rosamilia JM, Fleming RM, Kortan AR, Glarum SH, Makhija AV, Muller AJ, Eick RH, Zahurak SM, Tycko R, Dabbagh G, Thiel FA. Conducting Films of C_{60} and C_{70} by Alkali-Metal Doping. Nature (London) 1991; 350:320.
15. Bleeke JR, Frey RF. Fullerene Science Module. St. Louis, MO: Department of Chemistry, Washington University; Available at http://www.chemistry.wustl.edu~/edudev/Fullerene/fullerene.html.
16. Iijima S. Helical microtubules of graphitic carbon. Nature (London) 1991;354:56.
17. Tomanek D, of the University of Michigan, nanotube web page http://www.pa.msu.edu/cmp/csc/nanotube.html.
18. Jenkins GM, Kawamura K. Polymeric Carbons–Carbon Fibers, Glass and Char. Cambridge: Cambridge University Press; 1976.
19. Bokros JC. Deposition, Structure and Properties of Pyrolytic Carbon. In: Walker PL, editor. Chemistry and Physics of Carbon. Volume 5, New York: Marcel Dekker, Inc.; 1969. p 1–118.
20. Ely JL, Emken MR, Accuntius JA, Wilde DS, Haubold AD, More RB, Bokros JC. Pure Pyrolytic Carbon: Preparation and Properties of a New Material, On-X Carbon for Mechanical Heart Valve Prostheses. J Heart Valve Dis 1998;7:626–632.
21. Cooney DO. Activated Charcoal: Antidotal and Other Medical Uses. New York: Marcel Dekker; 1980.
22. Baker FS, Miller CE, Repik AJ, Tolles ED. Activated Carbon, in Kirk–Othmer. Encyc Chem Technol 1992;4:1015–1037.
23. Puri Balwant Rai. Chemsiorbed oxygen evolved as carbon dioxide and its influence on surface reactivity of carbons. Carbon 1966;4:391–400.
24. Cheremishoff NP, Moressi AC. Carbon adsorption applications. In: Cheremisinoff NP, Ellerbusch F, editors. Carbon Adsorption Handbook. Ann Arbor: Ann Arbor Science; 1978.
25. Pradhan BK, Sandle NK. Effect of different oxidizing agent treatments on the surface properties of activated carbons. Carbon 1999;37:1323–1332.
26. McQreey RL. Carbon electrodes: structural effects on electron transport kinetics. In: Bard AJ, editor. Electroanalytical Chemistry. New York: Dekker; 1991.
27. See Matweb.com for a variety of properties for engineering materials.
28. Diefendorf RJ, Stover ER. Pyrolytic Graphites. . .How structure affects properties. Metals Prog 1962;8 (May): 103–108.
29. Schoen FJ. Carbons in Heart Valve Prostheses: Foundations and clinical Performance. In: Zycher M, editor. Biocompatible Polymers, Metals and Composites. Lancaster PA: Technomic; 1983. p 240–261.
30. Bokros J. Carbon biomedical devices. Carbon 1977;15:355–371.
31. Haubold AD, Shim HS, Bokros JC. Carbon in Medical Devices. In: Williams DF, editor. Biocompatibility of Clinical Implant Materials. Volume 2, Boca Raton, FL: CRC Press; 1981. p 3–42.

32. More RB, Haubold AD, Bokros JC. Pyrolytic Carbon for Long-Term Medical Implants. In: Ratner B, Hoffman A, Schoen F, Lemons J, editors. Biomaterials Science: An Introduction to Materials in Medicine. 2nd ed. Academic Press; 2004.
33. More RB, Sines G, Ma L, Bokros JC. Pyrolytic Carbon. Encyclopedia of Biomaterials and Biomedical Engineering. Marcel Dekker; 2004.
34. Baier RE, Gott VL, Feruse A. Surface Chemical Evaluation of Thromboresistant Materials Before and After Venous Implantation. Trans Am Soc Artif Intern Organs 1970; 16:50–57.
35. Lee RG, Kim SW. Adsorption of Proteins onto Hydrophobic Polymer Surfaces: Adsorption Isotherms and Kinetics. J Biomed Mater Res 1974;8:251.
36. Nyilas E, Chiu TH. Artificial Surface/Sorbed Protein Structure/Hemocompatibility Correlations. Artif Organs 1978;2 (Suppl): 56–62.
37. Salzman EW, Lindon J, Baier D, Merril EW. Surface-Induced Platelet Adhesion, Aggregation and Release. Ann NY Acad Sci 1977;283:114.
38. Feng L, Andrade JD. Protein Adsorption on Low-Temperature Isotropic Carbon: I Protein Conformational Change Probed by Differential Sacnning Calorimetry. J Biomed Mater Res 1994;28:735–743.
39. Chinn JA, Phillips RE, Lew KR, Horbett Fibrinogen and Albumin Adsorption to Pyrolite Carbon. Trans Soc Biomater 1994;17:250.
40. Guglielmotti MB, Renou S, Cabrini RL. A histomorphometric study of tissue interface by laminar implant test in rats. Int J Oral Maxillofac Implants 1999;14:565–570.
41. Santavirta S, Takagi M, Gomez-Barrena E, Nevalainen J, Lassus J, Salo J, Konttinen YT. Studies of host response to orthopedic implants and biomaterials. J Long Term Eff Med Implants 1999;9:67–76.
42. Hill JB, Horres CR. The BD Hemodetoxifier: Particulate release and its significance. In: Chang TMS, editor. Artificial Kidney, Artificial Liver and Artificial Cells. New York: Plenum Press; 1978. p 199–207.
43. Williams DF. The Williams' Dictionary of Biomaterials. United Kingdom: Liverpool University Press; 1999.
44. Fettel BE, Johnston DR, Morris PE. Accelerated life testing of prosthetic heart valves. Med Inst 1980;14(3): 161–164.
45. Bensen J. Pre-Survey on the Biomedical Applications of Carbon. 1969. North American Rockwell Corporation Report R-7855.
46. Ford X. Clinical Toxicology. 1st ed., W. B. Saunders Company; 2001.
47. Roberts X. Clinical Procedures in Emergency Medicine. 3rd ed., W. B. Saunders Company; 1998.
48. LaGrange LD, Gott VL, Bokros JC, Ramos MD. Compatibility of Carbon and Blood. In: Hegyeli RJ, editor. Artificial Heart Program Conference Proceedings. Washington, DC: US Government Printing Office; 1969. Chapt. 5. p 47–58.
49. Sadeghi H. Dysfonctions des prostheses valvulaires cardaques et leur traitment chirgical. Schwiez Med Wschr 1987; 117:1665–1670.
50. Kaae JL. The mechanism of deposition of pyrolytic carbon. Carbon 1985;23(6): 665–667.
51. Kaae JL, Wall DR. Microstructural Characterization of Pyrolytic Carbon for Heart Valves. Cells Mater 1996;6(4): 281–290.
52. Ma L, Sines G. High resolution structural studies of a pyrolytic carbon used in medical applications. Carbon 2002;40:451–454.
53. Akins RJ, Bokros JC. The Deposition of Pure and Alloyed Isotropic Carbons and Steady State Fluidized Beds. Carbon 1974;12:439–452.
54. Reilly DT, Burstein AH, Frankel VH. The Elastic Modulus for Bone. J Biomech 1974;7:271.
55. Reilly DT, Burstein AH. The Mechanical Properties of Bone. J Bone Jt Surg Am 1974;56:1001.
56. De Salvo G. Theory and Structural Design Applications of Weibull Statistics. 1970. WANL-TME-2688, Westinghouse Electric Corporation.
57. More RB, Kepner JL, Strzepa P. Hertzian Fracture in Pyrolite Carbon. In: Ducheyne P, Christiansen D, editors. Bioceramics. Volume 6, Oxford: Butterworth-Heinemann Ltd; 1993. p 225–228.
58. Gilpin CB, Haubold AD, Ely JL. Fatigue Crack Growth and Fracture of Pyrolytic Carbon Composites. In: Ducheyne P, Christiansen D, editors. Bioceramics. Volume 6, Oxford: Butterworth-Heinemann Ltd; 1993. p 217–223.
59. Ma L, Sines G. Fatigue of Isotropic Pyrolytic Carbon Used in Mechanical Heart Valves. J Heart Valve Dis 1996;5(Suppl.I): S59–S64.
60. Ma L, Sines G. Unalloyed Pyrolytic Carbon for Implanted Heart Valves. J Heart Valve Dis 1999;8(5): 578–585.
61. Ma L, Sines G. Fatigue Behavior of Pyrolytic Carbon. J Biomed Mat Res 2000;51:61–68.
62. Ritchie RO, Dauskardt RH, Yu W, Brendzel AM. Cyclic Fatigue-crack Propagation, Stress Corrosion and Fracture Toughness Behavior in Pyrolite Carbon Coated Graphite for Prosthetic Heart Valve Applications. J Biomed Mat Res 1990;24:189–206.
63. Beavan LA, James DW, Kepner JL. Evaluation of Fatigue in Pyrolite Carbon. In: Ducheyne P, Christiansen D, editors. Bioceramics. Volume 6, Oxford: Butterworth-Heinemann Ltd; 1993. p 205–210.
64. Bokros JC, Haubold AD, Akins RJ, Campbell LA, Griffin CD, Lane E. The durability of mechanical heart valves replacements: past experience and current trends. In: Bodnar E, Frater RWM, editors. Replacement Cardiac Valves. New York: Pergamon Press; 1991. p 21–47.
65. Haubold AD. On the Durability of Pyrolytic Carbon In Vivo. Med Prog Through Technol 1994;20:201–208.
66. Kelpetko V, Moritz A, Mlczoch J, Schurawitzki H, Domanig E, Wolner E. Leaflet Fracture in Edwards-Duromedics Bileaflet Valves. J Thorac Cardiovasc Surg 1989;97: 90–94.
67. Kafesjian R, Howanec M, Ward GD, Diep L, Wagstaff L, Rhee R. Cavitation Damage of Pyrolytic Carbon in Mechanical Heart Valves. J Heart Valve Dis 1994;3(Suppl I): S2–S7.
68. Richard G, Cao H. Structural failure of Pyrolytic Carbon Heart Valves. J Heart Valve Dis 1996;5(Suppl I): S79–S85.
69. Shim HS, Schoen FJ. The wear resistance of pure and silicon-alloyed isotropic carbons. Biomater Med Dev Art Org 1974;2(2): 103–118.
70. Shim HS. The wear of titanium alloy, and UHMW polyethylene caused by LTI carbon and Stellite 21. J Bioengr 1977;1:223–229.
71. More RB, Silver MD. Pyrolytic Carbon Prosthetic Heart Valve Occluder Wear: In Vivo vs. In Vitro Results for the Björk-Shiley Prosthesis. J Appl Biomater 1990;1:267–278.
72. More RB. An Examination of Two Retrieved Long-Term Human Implant Bjork-Shiley Valves. Med Prog Technol 1994;20:195–200.
73. More RB, Haubold AD, Silver MD. Pyrolytic Carbon Wear in Retrieved Mechanical Heart Valve Prosthesis Implants. 25th Annual Meeting of the Society for Biomaterials, 1999. p 553.
74. More RB, Chang BC, Hong YS, Cao BK, Butany J, Wear Analysis of Retrieved Mitral Bileaflet Mechanical Heart Valve Prostheses, Presented to the Society for Heart Valve Disease, 1st Biennial Symposium, London; June 2001.
75. More RB, Haubold AD, Silver MD. Pyrolytic Carbon Wear in Retrieved Mechanical Heart Valve Prosthesis Implants. 25th Annual Meeting of the Society for Biomaterials, 1999. p 553.

76. Wieting DW. The Björk-Shiley Delrin Tilting Disc Heart Valve: Historical Perspective, Design and Need for Scientific Analyses After 25 Years. J Heart Valve Dis 1996;5(Suppl I): S157–S168.
77. Schoen FJ, Titus JL, Lawrie GM. Durability of Pyrolytic Carbon-Containing Heart Valve Prostheses. J Biomed Mater Res 1982;16:559–570.
78. Griffin CD, Buchanan RA, Lemons JE. In Vitro Electrochemical Corrosion Study of Coupled Surgical Implant Materials. J Biomed Mater Res 1983;17:489–500.
79. Thompson NG, Buchanan RA, Lemons JE. In Vitro Corrosion of Ti-6Al-4V and Type 316L Stainless steel When Galvanically Coupled with Carbon. J Biomed Mater Res 1979;13:35–44.
80. Cook SD, Thomas KA, Kester MA. Wear characteristics of the canine acetabulum against different femoral prostheses. J Bone Joint Surg 1989;71B:189–197.
81. Tian CL, Hetherington VJ, Reed S. A Review of Pyrolytic carbon: Application in Bone and Joint Surgery. J Foot Ankle Surg 1993;32(5):490–498.
82. Cook SD, Beckenbaugh RD, Redondo J, Popich LS, Klawitter JJ, Linscheid RL. Long term follow-up of pyrolytic carbon metacarpophalangeal implants. J Bone Joint Surg 1999; 81A(5): 635–648.
83. Ferrari M. Clinical evaluation of fiber-reinforced epoxy resin posts and cast post and cores. Am J Dent 2000; 01-May-13 (Spec No): 15B–18B.
84. Pamula E. Studies on development of composite biomaterials for reconstruction of the larynx. Polim Med 2001;31(1–2):39–44.
85. Katoozian H. Material optimization of femoral component of total hip prosthesis using fiber reinforced polymeric composites. Med Eng Phys 2001;23(7):503–509.
86. Früh HJ. Fusion implants of carbon fiber reinforced plastic. Orthopade 2002;31(5):454–458.
87. Dearnaley G. Diamond-like carbon: a potential means of reducing wear in total joint replacements. Clin Mater 1993;12:237–244.
88. Lappalainen R, Anttila A, Heinonen H. Diamond coated total hip replacements. Clin Orthop 352 (July 1998): 118–127.
89. Pantarotto D, Partidos CD, Graff R, Hoebeke J, Briand JP, Prato M. Bianco A. Synthesis, structural characterization, and immunological properties of carbon nanotubes functionalized with peptides. J Am Chem Soc 2003 May 21; 125(20): 6160–6164.
90. Qingnuan L, Yan X, Xiaodong Z, Ruili L, Qieqie D, Xiaoguang S, Shaoliang C, Wenxin L. Preparation of (99m)Tc-C(60)(OH)(x) and its biodistribution studies. Nucl Med Biol 2002 Aug; 29(6): 707–710.
91. Gonzalez KA, Wilson LJ, Wu W, Nancollas GH. Synthesis and in vitro characterization of a tissue-selective fullerene: vectoring C(60)(OH)(16)AMBP to mineralized bone. Bioorg Med Chem 2002 Jun; 10(6): 1991–1997.
92. Wolff DJ, Barbieri CM, Richardson CF, Schuster DI, Wilson SR. Trisamine C(60)-fullerene adducts inhibit neuronal nitric oxide synthase by acting as highly potent calmodulin antagonists. Arch Biochem Biophys 2002 Mar 15; 399(2): 130–141.
93. Schinazi RF, Sijbesma R, Srdanov G, Hill CL, Wudl F. Synthesis and virucidal activity of a water-soluble, configurationally stable, derivatized C60 fullerene. Antimicrob Agents Chemother 1993 Aug.; 37(8): 1707–1710.
94. Park KH, Chhowalla M, Iqbal Z, Sesti F. Single-walled carbon nanotubes are a new class of ion channel blockers. J Biol Chem 2003 Dec. 12; 278(50): 50212–50216, Epub 2003 Sep. 30.

See also BIOMATERIALS: TISSUE ENGINEERING AND SCAFFOLDS; BIOMATERIALS, TESTING AND STRUCTURAL PROPERTIES OF; BIOSURFACE ENGINEERING; MATERIALS AND DESIGN FOR ORTHOPEDIC DEVICES; HEART VALVE PROSTHESES.

BIOMATERIALS CORROSION AND WEAR OF

ROGER J. NARAYAN
University of North Carolina
Chapel Hill, North Carolina

MIROSLAV MAREK
Georgia Institute of Technology
Atlanta, Georgia

CHUNMING JIN
North Carolina State University
Raleigh, North Carolina

INTRODUCTION

Many materials suffer degradation with time when exposed to aggressive chemical environments within the human body. In metallic biomaterials, degradation results from electrochemical corrosion. Ceramic and polymeric biomaterials may undergo physical or chemical deterioration processes. In addition, mechanical forces may act to increase damage by wear, abrasion, or environment-induced cracking processes.

Corrosion of implants, dental restorations, and other objects placed in the human body may result in degradation of function as a result of loss of mass, decrease in mechanical integrity, or deterioration of aesthetic qualities. The associated release of corrosion products and the flow of the corrosion currents also may cause inflammation, allergic reactions, local necrosis, and many other health problems.

For electronic conductors (e.g., metals), corrosive interaction with ionically conducting liquids (e.g, body fluids) is almost always electrochemical. The degradation of metals is due to an oxidation process that involves the loss of electrons. This process involves a change from a metallic state to an ionic state, in which the ions dissolve or form nonmetallic solid products. For the process to continue, the released electrons must be consumed in a complementary reduction, which usually involves species present in the biological environment (e.g., hydrogen ions or dissolved oxygen). The reaction resulting in oxidation is usually called an anodic process and reaction resulting in reduction is usually called a cathodic process. The metal is referred to as an electrode, and the liquid environment is referred to as an electrolyte.

For many metals, the most important environmental variables are the concentrations of chloride ions, hydrogen ions, and dissolved oxygen. In many human body fluids, the chloride ion concentration varies in a relatively narrow range near $0.1\,\mathrm{mol \cdot L^{-1}}$; however, it may be variable (e.g., urine) or lower (e.g., saliva) in certain body fluids. The hydrogen ion concentration is expressed as a pH value and is near neutral ($\mathrm{pH} = 7$) for plasma, interstitial fluid, bile, and saliva; however, it is more variable ($\mathrm{pH} = 4$–8) in urine and very low ($\mathrm{pH} = 1$–3) in gastric juice (1). The chloride concentration and pH are most important factors determining the rate of oxidation because of their effect on protective oxide passivating films on metals. The dissolved oxygen concentration affects mainly the cathodic process. The usual range of partial pressure of oxygen in body fluids is ~40–100 mmHg (5.33–13.33 kPa) (1–3).

For electrochemical oxidation to cause clinically relevant degradation of a material, the electrochemical reaction must be energetically possible (thermodynamics) and the reaction rate must be appreciable (kinetics). Oxidation of nickel, for example,

$$Ni \rightarrow Ni^{2+} + 2e^{-} \quad (1)$$

will proceed in the indicated direction if the potential of the electrode on which the reaction occurs is higher (more positive or less negative) than the equilibrium potential for a given electrolyte, and is a function of the energy change involved. The equilibrium potential also depends on temperature, pressure, and activity ($\approx$ concentration) of ions. The values of potentials for reactions between metals and their ions in water are given under standard conditions (temperature 25 °C, pressure 1 atm, and activity of ions equal to 1) and are written in the form of materials reduction. These values, known as standard single electrode potentials, are listed in the so-called electrochemical or electromotive (EM) series (1). Noble metals, which have no tendency to dissolve in water have positive standard single-electrode potentials. On the other hand, active metals with a high tendency to react with water exhibit negative potentials.

While the potential of the anodic process must be above the equilibrium potential for the reaction to proceed as oxidation, for the cathodic process the electrode potential must be below (more negative or less positive) the equilibrium potential for net reduction to occur. For reduction of hydrogen ions,

$$2H^{+} + 2e^{-} = H_2 \quad (2)$$

the equilibrium potential at normal body temperature (37 °C) and pH 7.4 (blood or interstitial fluids) is −0.455 V (SHE) (−0.697 V, SCE), while the equilibrium potential of the other likely cathodic reaction,

$$O_2 + 4\,H^{+} + 4e^{-} = 2\,H_2O \quad (3)$$

is −0.753 V (SHE) (0.511 V, SCE) at 40 mmHg (5.33 kPa) of oxygen partial pressure. Although reaction 3 is more sluggish than reaction 2, for most metals in the human body the electrode potential of reaction 3 is above the equilibrium potential of the hydrogen reaction 2, and reduction of oxygen is the dominant cathodic process.

In spontaneous electrochemical corrosion, at least two reactions occur simultaneously. At least one reaction occurs in the direction of oxidation, and at least one reaction occurs in the direction of reduction. Each reaction has its own equilibrium potential, and this potential difference results in a current flow, as the electrons released in oxidation flow to the sites of reduction and are consumed there. In the absence of a significant electrical resistance in the current path between the reaction sites, a common potential is established, which is known as mixed potential or corrosion potential (E_{corr}). At this potential, both reactions produce the same current in opposite directions in order to preserve electrical neutrality. The value of the oxidation current, which is equal to the absolute value of the reduction current, per unit area at this potential is known as the corrosion current density (i_{corr}). The oxidation and reduction reactions may be distributed uniformly on the same metal surface; however, there are often some regions of the biomaterial surface that are more favorable for oxidation and other regions that are more favorable for reduction. As a result, either local anodic and cathodic areas or completely separate anodes and cathodes are formed.

The corrosion rate (mass of metal oxidized per unit area and time) is proportional to the corrosion current density. The conversion is given by the Faraday's law, which states that an electric charge of 96,485 C is required to convert 1 equiv weight of the metal into ions, or vice versa. The shift of the potential of a reaction from the equilibrium value to the corrosion potential is called polarization by the flow of the current. The resulting current flowing at corrosion potential depends on the way the current of each reaction varies with the potential. If the current is controlled by the activation energy barrier for the reaction at the electrode surface, then the reaction rate increases exponentially with increasing potential for oxidation reactions and decreases exponentially with increasing potential for reduction reactions. The activation energy controlled current typically increases or decreases ten times for a potential change of ~50–150 mV. At high reaction rates, the current may become limited by the transport of reaction species to or from the electrodes; eventually, the corrosion process may become completely controlled by diffusion and independent of potential.

The vast majority of uses for metallic biomaterials in the human body are successful due to the phenomenon of passivity. In a passive state, these metals become covered with thin, protective films of stable, poorly soluble oxides or hydroxides when exposed to an aqueous electrolyte. Once this passivating film forms, the current density drops to a very low value and becomes much less dependent on the potential. The variation of the reaction current density with the potential can be illustrated in a polarization diagram. A schematic diagram in Fig. 1 shows some of the main reactions in corrosion and relevant parameters. A straight-line relationship in a semilogarithmic (E vs. log I) diagram indicates that an activation energy-controlled reaction is occurring. This electrochemical activity is known as Tafel behavior, and the slopes of the lines (~50–150 mV per 10-fold change in the current or current density) are equal to the values of the Tafel constants. When the oxidation reaction of the metal shows this relationship at the corrosion potential, it indicates that the metal is actively corroding. If a metal forms a passivating film when the potential exceeds a critical value in the active corrosion region, then the current density drops from a value called the critical current density for passivation (i_{crp}) at a primary passivation potential (E_{pp}) to a low current density in the passive state (i_p). This behavior is illustrated schematically in Fig. 2. For an electrode to maintain a stable passive state, the intersection of the oxidation (anodic) and reduction (cathodic) lines must occur in the region of passivity.

The polarization characteristics of a biomaterial can be experimentally determined using a device called a potentiostat, which maintains the sample potential at a set value

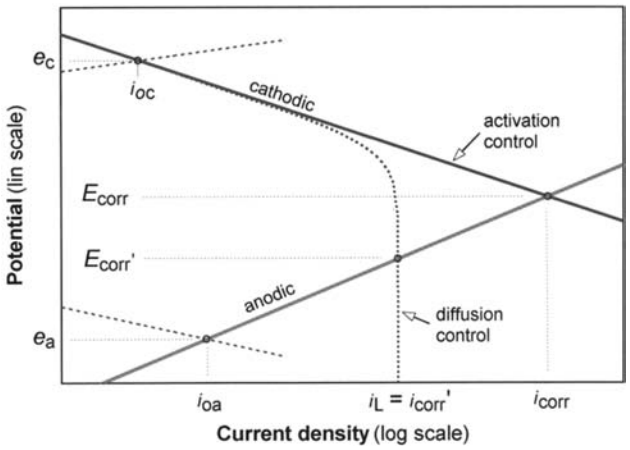


Figure 1. Schematic polarization diagram showing oxidation (anodic) and reduction (cathodic) reactions of a corrosion process, for reactions controlled by activation energy and by mass transport (diffusion). In this figure, e_a and e_c refer to equilibrium potentials of the anodic and cathodic process, respectively, i_{oa} and i_{oc} refer to exchange current densities, E_{corr} refers to mixed corrosion potential, and i_L refers to limiting current density.

versus a reference electrode by passing current between the sample and an auxiliary counterelectrode. A scan generator can be used to vary the controlled potential over a range of interest, and the E–i relationship can be determined. The relationship of main interest is usually the oxidation rate as a function of the potential, which can be depicted in an anodic polarization diagram. Since only a net current (difference between the absolute values of the oxidation and reduction currents) can be measured, the experimental polarization curve shows a value approaching zero at the intersection of the anodic and cathodic polarization curves.

Experimental anodic polarization curves for passivating metals and alloys often do not exhibit the passivation peak shown in Fig. 2, either because the metal forms an oxide in the electrolyte without undergoing active dissolution or because an oxide film already has formed as a result of exposure to air. More importantly for human body fluids and other chloride-containing electrolytes, the region of passivity is often limited by a localized passivation breakdown above a critical breakdown potential (E_b). When a breakdown occurs, intensive oxidation takes place within localized regions on the biomaterial surface, resulting in sometimes significant pit formation. In an experimental anodic polarization diagram, breakdown appears as a sharp increase in the measured current above the critical breakdown potential. Because of the destructive nature of surface pitting, the determination of critical breakdown potential is one of the most important ways of assessing the suitability of novel metallic biomaterials for use in medical devices.

The high current density in active pits is due to the absence of a passivating film, which results from local chemical and electrochemical reactions that change the electrolyte to become highly acidic and depleted in dissolved oxygen. A similar corrosion mechanism may occur in interstices known as crevices, where the transport of species to and from the localized corrosion cell is difficult. This process, known as crevice corrosion, does not require a potential exceeding the critical breakdown potential for the initiation of corrosion. Both pit and crevice corrosion cells may repassivate if the potential is lowered below a value needed for maintenance of a high oxidation rate on the bare (nonpassivated) metal surface. The potential below which active pits repassivate is called the repassivation or protection potential (E_p). The concept of a protection potential also applies to crevice corrosion. Experimentally, repassivation can be studied by reversing the anodic polarization scan and recording the potential at which the current returns to a passive state value (Fig. 3). Repassivation can also be examined by initiating pitting or crevice corrosion and lowering the potential in steps until the current

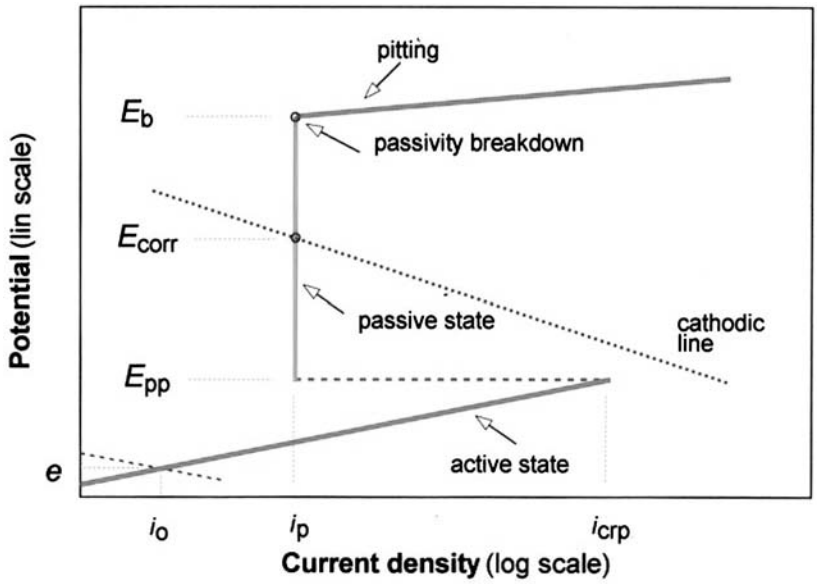


Figure 2. Schematic polarization diagram, showing a transition from active to passive state and a breakdown of passivity. In this figure, i_{crp} refers to critical current density for passivation, i_p refers to current density in the passive state, E_{pp} refers to primary passivation potential, and E_b refers to breakdown potential.

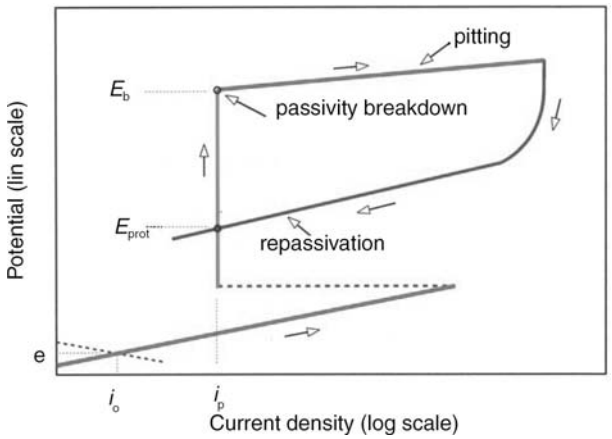


Figure 3. Schematic experimental cyclic polarization diagram for a passivating electrode, showing passivity breakdown and repassivation after potential scan reversal. In this figure, E_{prot} refers to protection (repassivation) potential.

shows low values that decrease with time (standard test methods F2129 and F746, respectively, ASTM 2005) (4). The difficulty in finding a reliable protection potential value is due to the fact that the ease of repassivation depends on the extent of pitting or crevice corrosion damage that has occurred before the potential drop.

For some polyvalent metals (e.g., chromium), soluble species (e.g., CrO_4^{2-}) become thermodynamically stable as the valence changes (e.g., from 3 to 6 in chromium) at potentials above those for a stable oxide. This process may result in another region of active dissolution at high potentials in a phenomenon known as transpassivity.

When corrosion is relatively uniform throughout the biomaterial, the most important corrosion parameter is the average corrosion rate. For biomaterials with very low corrosion rates, the average corrosion rate is mostly determined by sensitive electrochemical techniques. The average corrosion current density (i_{corr}) is usually determined either from the results of the polarization scan by extrapolating the anodic and cathodic lines to the corrosion potential or from calculating the value of the polarization resistance. The polarization resistance (R_p) is defined as the slope of the polarization curve at zero current density $[R_p = (dE/di_n)_{i_n=0}]$, in which i_n is the net (measured) current density.

When two or more dissimilar metals are placed in contact within an electrolyte, their interaction may cause galvanic corrosion. The oxidation degradation is enhanced for the metal with the lower individual corrosion potential, which becomes the anode of the cell. It is polarized towards a higher potential at the other electrode, which becomes the cathode. Since the oxidation current increase on the anode must be balanced by an identical reduction current increase on the cathode, a combination of a small anode with a large cathode is more detrimental than the reverse situation, since a larger increase in the oxidation current density is produced. In practical situations, resistance in the current path between the electrodes often reduces the galvanic effect. Differences in the concentrations of reaction species at different regions on the metal surface may result in a potential difference, which leads to additional polarization. An increase in the oxidation current density, a difference in the equilibrium potentials, and a flow of current may result. Differences in concentrations of hydrogen ions, dissolved metal ions, or dissolved oxygen may result in concentration cell corrosion.

Metal parts subjected to mechanical loading in a corrosive environment may fail by environment-induced cracking (EIC). Stress corrosion cracking (SCC) may occur in some biomaterials when they are subjected to static loading under certain environmental conditions. Corrosion fatigue (CF) may result from variable loading in reactive environments. When the failure can be attributed to the entry of hydrogen atoms into the metal, the phenomenon is referred to as hydrogen induced cracking (HIC). Environment-induced cracking may be caused by complex combinations of mechanical, chemical, and electrochemical forces; however, the exact mechanisms of this behavior are subject to significant controversy. In these cases, mechanical factors may play important roles in crack propogation.

Intergranular corrosion occurs when dissolution is confined to a narrow region along the grain boundaries. This process is either due to precipitation of corrosion susceptible phases or due to depletion in elements that provide corrosion protection along the boundaries, which is caused by precipitation of phases rich in those elements. Some stainless steel and nickel-chromium alloys may be sensitized due to precipitation of chromium-rich carbides along grain boundaries when heated to a specific temperature range. Sensitization is normally prevented from occurring in stainless steels currently used in medical devices, which contain very low amounts of carbon.

Passivating films may also be mechanically destroyed in wear- , abrasion- , erosion- , and fretting-corrosion processes. Wear-corrosion involves materials in a friction contact that exhibit substantial relative movement. Fretting occurs in situations in which there are only small relative movements between materials that are essentially fixed with respect to one another. The resulting wear debris may cause abrasion–corrosion behavior. Wear-corrosion may occur in artificial joints, including the metal ball of a hip joint in contact with the polyethylene cup. Fretting may take place between the ball and the stem of multicomponent hip implants. In both forms of corrosion, the narrow gap between contacting surfaces creates crevice conditions. In addition, the destructive effect of friction and abrasion on the protective surface film is superimposed on the corrosion mechanism in the crevice cell. Erosion corrosion may occur on devices exposed to rapidly flowing fluids, including the surfaces of artificial heart valves.

A wide variety of metals and alloys have been used in medical devices. The three most commonly used alloys are stainless steel, cobalt alloys, and titanium alloys (5). Type 316 LVM (low carbon, vacuum-melted) stainless steel is less corrosion resistant than cobalt or titanium alloys, and it is most often used for temporary implants (5–8). This material is referred to as an austenitic steel, because it contains an iron carbide phase called austenite (γ-iron). Implant-grade steel has a nominal composition of 18% chromium, 14% nickel, and 2.5% molybdenum; the

compositional limits and properties are specified by ASTM standards F 138 and F 139 for wrought steel and F 745 for cast steel (ASTM, 2005) (4). Chromium serves to improve corrosion resistance through the formation of a highly protective surface film rich in chromium oxide. Implant-grade steel has a low carbon content in order to prevent sensitization and intergranular corrosion. Alloying with - molybdenum further improves the resistance, especially to crevice corrosion and pitting. Nickel serves to stabilize the face-centered cubic (fcc) structure. On the other hand, manganese sulfide inclusions, which contribute to initiation of pitting, are minimized.

The corrosion resistance of stainless steel greatly depends on the surface conditions, and stainless steel implants are almost always electropolished and prepassivated by exposure to nitric acid (standard practice F86, ASTM 2005) (4). The breakdown potential is usually around 0.4 V (SCE), with a large hysteresis loop and a low protection potential (9). Considering that the potential in the human body is not likely to exceed about 0.5 V (SCE) (see Eq. 3 and its equilibrium potential), a well polished and passivated 361 LVM stainless steel is not very susceptible to pitting in the human body, especially for unshielded and undisturbed implant surfaces. Once localized attack is initiated, however, repassivation is difficult. As a result, stainless steel implants are very susceptible to crevice corrosion, especially when the crevice situation is combined with destruction of the surface film (e.g., fretting of bone plates under the screw heads). Small single component stainless steel implants, such as balloon-expandable vascular stents, that are made of high purity precursor materials and are subjected to a high quality surface treatment and inspection can achieve a breakdown potential in excess of 0.8 V (SCE); these materials are considered very resistant to localized corrosion (10). Stainless steel bars [containing 22% chromium 12.5% nickel , 5% manganese, and 2.5% molybdenum (ASTM F 1586)] and wires [containing 22% chromium, 12.5% nickel, 5% manganese, and 2.5% molybdenum (ASTM F 1314)] strengthened with nitrogen have shown a higher breakdown potential than ASTM F 138 steel (4).

Vitallium and other cobalt–chromium alloys were developed as a corrosion resistant, high strength alternative to stainless steel alloys. These materials were first used in dentistry, and were later introduced to orthopedics and other surgical specialties. The cast cobalt–chromium alloy most commonly used in medical devices (ASTM F 75) contains 28% chromium and 6% molybdenum (4). This alloy was found to be suitable for investment casting into intricate shapes. In addition, it exhibited very good corrosion and excellent wear resistance; however, it possessed low ductility. Alloys with slightly modified compositions were later developed for forgings (ASTM F 799) and wrought bars, rods, and wire (ASTM F 1537) (4). Alloy F75 has shown corrosion resistance superior to stainless steel in the human body. Laboratory studies reported a breakdown potential of 0.5 V (SCE) and protection potential of 0.4 V (SCE) (6,7,9,11). These properties have made it possible to use cobalt–chromium alloys for permanent implants. Cobalt–chromium alloys with porous surfaces have been used for bone ingrowth, although they have been superseded by even more crevice corrosion resistant and biocompatible titanium alloys. The excellent corrosion resistance of cobalt–chromium alloys can be attributed to a high chromium content and a protective surface film of chromium oxide. Concerns have been raised, however, regarding the release of biologically active hexavalent chromium ions (12). Other cobalt-based wrought surgical alloys include F90 (Co-Cr-W-Ni), F563 (Co-Ni-Cr-Mo-W-Fe), F563 (Co-Ni-Cr-Mo-W-Fe), F1058 (Co-Cr-Ni-Mo), and F688 (Co-Ni-Cr-Mo) (4). These alloys provide good to excellent corrosion behavior and a variety of mechanical properties, which depend on thermomechanical treatment. However, there is some concern regarding metal ion release in these alloys, which contain high nickel concentrations.

Titanium and titanium alloys have been used in orthopedic implants and other medical devices since the 1960s. Their popularity has rapidly increased because they possess high corrosion resistance, adequate mechanical properties, and relatively benign degradation products. Although titanium is thermodynamically one of the least stable structural metals in air and water, it acquires high resistance to corrosion due to a very protective titanium oxide film. Unalloyed titanium (ASTM F67 and F1341) and titanium-6 % aluminum, 4 % vanadium alloy (ASTM F136 and F1472 for wrought alloy and F1108 for castings) are commonly used in orthopedic prostheses (4). These materials exhibit a breakdown potential in body fluid substitutes well above the physiological range of potentials (several volts vs. SHE). In addition, they readily repassivate in biological fluids, which makes them highly resistant to pitting and crevice corrosion. The high crevice corrosion resistance and biocompatibility of titanium alloys have made it possible to create porous titanium surfaces that allow for bone ingrowth and cementless fixation of implants.

One shortcoming of titanium and titanium alloys is their relatively poor wear resistance (5). Since resistance to corrosion depends on the integrity of the protective oxide film, wear-corrosion remains a problem for titanium alloy prostheses. Surface treatments (including nitrogen diffusion hardening, nitrogen ion implantation, and thin-film deposition) may be used to provide more wear-resistant articulating surfaces. Another solution to titanium wear involves the use of multicomponent implants (e.g., implants that contain smooth surfaces made of cobalt–chromium alloy for articulating components and porous surfaces made out of titanium alloy for bone ingrowth and biological fixation). However, fretting corrosion may occur as a result of micromovement at the taper joints between the components, which may destroy the surface passivating films and increase overall corrosion rates (13–15). In spite of the very successful use of the Ti-6Al-4V alloy orthopedic implants, some concern remains regarding the possible toxicity of the aluminum and vanadium components within this alloy. A variety of vanadium-free or aluminum- , and vanadium-free alloys have been developed, including Ti-15Sn-4Nb-2Ta-0.2Pd, Ti-12Mo-6Zr-2Fe (TMZF), Ti-15Mo, and Ti-13Nb-13Zr (5). Ti-12Mo-6Zr-2Fe (TMZF) and Ti-13Nb-13Zr alloys exhibit lower elastic moduli and higher tensile properties. The alloying

elements also form highly protective oxides, which contribute to the excellent corrosion resistance of these materials (16).

An equiatomic nickel–titanium alloy (Nitinol) has received considerable interest as an implant material because of its shape memory and pseudoelasticity properties, the latter resulting in a very low apparent elastic modulus. This superelastic behavior has allowed the development of self-expandable vascular stents, bendable eyeglass frames, orthodontic dental archwire, and intracranial aneurysm clips. Several studies have shown good biocompatibility of Nitinol; however, clinical failures have also been reported (17–19). Laboratory studies have shown a wide variety of performance, with resistance to the breakdown of passivity ranging from poor to excellent (20–22). Resistance to the initiation of pitting critically depends on the surface conditions. A surface film that consists mostly of titanium oxide results in a high resistance to pitting; however, the presence of elemental nickel or nickel oxide reduces the breakdown potential. In addition, recent studies have shown that strained nickel–titanium alloy exhibits significant improved corrosion resistance over as-prepared materials. Other conditions that may affect corrosion resistance include surface roughness, the presence of inclusions, and the concentration of intermetallic species (23).

Another group of biomaterials is used in restorative dentistry and orthodontics. Materials for restorative dentistry must not only meet corrosion, wear, and compatibility considerations described earlier, but also satisfy aesthetic requirements and must have the capacity to be either precisely cast into intricate shapes or used to directly fill a prepared cavity in a tooth. Dental cast alloys can be roughly divided into three major groups of high noble alloys, seminoble alloys, and base alloys. The high noble alloys include those with a high percentage of gold or other noble metals (e.g., platinum), and derive their corrosion resistance mainly from a low thermodynamic tendency to react with the environment. Seminoble alloys often have complex compositions, and either possess a relatively low noble metal content or contain a significant concentration of silver. These materials possess a higher thermodynamic tendency to react than high noble alloys; however, their kinetics of aqueous corrosion in saliva is sufficiently slow, and allows these materials to provide adequate corrosion resistance under biological conditions. The main corrosion concern for seminoble alloys is their tendency to react with sulfur in food and drinks and form dark metallic sulfide film, resulting in the loss of aesthetic quality. Base dental cast alloys include cast titanium, titanium alloys, and nickel–chromium alloys. These materials lack the aesthetic qualities of noble alloys; however, they are resistant to sulfide tarnishing. Nickel–chromium alloys exhibit passivation behavior and some susceptibility to pitting and crevice corrosion. Cast titanium and titanium alloys exhibit highly protective passive films and high resistance to chloride corrosion; however, they demonstrate some susceptibility to fluoride attack, which is of some concern due to the prophylactic use of fluoride rinses and gels. Direct-filling metallic materials include unalloyed gold and dental amalgams, which are alloys of mercury, silver, tin, copper, and some other minor elements. Dental amalgams have a higher thermodynamic tendency for reaction with the oral environment than noble and seminoble cast dental alloys. In addition, these materials receive weaker protection by passivating surface films than implant alloys. However, these materials have shown adequate long-term clinical corrosion resistance. This property has been greatly improved by the transition from low copper amalgams, which contain a corrosion susceptible Sn–Hg structural phase, to high copper amalgams, which contain a more corrosion resistant Sn–Cu phase. Low copper amalgams exhibit breakdown of passivity and suffer from selective corrosion of the tin–mercury phase, which penetrates and weakens the structure. On the other hand, high copper amalgams do not show breakdown in laboratory testing and have demonstrated better clinical performance. The use of dental amalgam in dentistry has been on the decline as a result of concerns regarding the release of small amounts of toxic mercury and due to improvements in the performance of nonmetallic dental composites. Recent reviews on dental alloys and their corrosion behavior can be found in Refs. (24) and (25). Materials for orthodontic applications include cobalt–chromium alloys, titanium alloys, nickel–titanium alloys, which exhibit similar corrosion behavior in dental applications and medical applications.

Ceramic materials were first used in medical devices in the early 1970s. These materials are either crystalline or amorphous, and contain atoms linked by highly directional ionic bonds. Alumina (Al_2O_3) and zirconia (ZrO_2) exhibit high passivation tendencies and resistance to breakdown properties. These materials exhibit better corrosion resistance, hardness, stiffness, wear resistance, and biocompatibility properties than metal alloys. Zirconia and alumina used in medical devices exhibit full-densities and uniformly controlled small grain sizes ($<5\,\mu m$) (26). Full-density ceramics are used in medical devices, because voids may increase stresses and degrade mechanical properties. Ceramics containing uniform small grains are used in order to minimize internal stresses from thermal contraction. In addition, ceramics with small grain sizes exhibit enhanced wear, hardness, and strength properties (27–31). Typical material combinations for ceramic hip prostheses include ceramic-on-ceramic; ceramic-on-metal; and ceramic-on-polymer wear couples.

A ceramic coating material that may provide corrosion resistance to an orthopedic prosthesis is diamond-like carbon (DLC). Diamond-like carbon refers to amorphous carbon materials that contain some component of sp^3-hybridized atoms. Nano- or microcrystalline graphite regions may also be present within the amorphous matrix. Hydrogen-free diamond-like carbon exhibits atomic number densities $>3.19\,g\ atom{\cdot}cm^{-3}$. Hydrogenated diamond-like carbon (HDLC) contains up to 30 atomic percent hydrogen and up to 10 atomic percent oxygen within CH_3 and OCH_3 inclusions, which are surrounded by an amorphous carbon matrix. The density of hydrogenated coatings rarely exceeds $2.2\,g{\cdot}cm^{-3}$. Hydogenated or hydrogen-free diamond-like carbon coatings may provide a medical device with an atomically smooth, low friction, wear resistant, corrosion resistant hermetic seal

between the bulk biomaterial, and the surrounding tissues. Tiainen demonstrated extremely low corrosion rates for diamondlike carbon-coated metals (32). The hydrogen-free diamond-like carbon coated-cobalt–chromium–molybdenum alloy and cobalt–chromium–molybdenum alloy were placed in saline solution equivalent to placement in body fluid for 2 years at a temperature of 37 °C. The DLC-coated cobalt–chromium–molybdenum alloy exhibited 10^5 lower corrosion rate than cobalt–chromium–molybdenum alloy. Similarly, the corrosion rate of DLC-coated titanium–aluminum-vanadium alloy in saline solution has been shown to be extremely low.

Bioactive ceramic materials, which develop a highly adherent interface with bony tissue, have been developed for several medical and dental applications, including coatings for promoting bone ingrowth, grouting agents for hip arthroplasty, and replacements for autologous bone grafts. The most commonly used bioactive ceramics include hydroxyapatite, $Ca_{10}(PO_4)_6(OH)_2$, tricalcium phosphate, $Ca_3(PO_4)_2$, and $Na_2OCaOP_2O_5SiO_2$ glasses (e.g., Bioglass). These materials undergo chemical–biochemical processes, which are dependent on several material properties. For example, 45S5 Bioglass, which contains 45 wt% SiO_2 and 5:1 $CaO:P_2O_5$ ratio, forms SiOH bonds, hydrated silica gel, hydroxyl carbonate apatite layer, matrix, and bone at the material/tissue interface. Materials with high (>60 mol%) SiO_2, low CaO/P_2O_5 ratios, and additions of Al_2O_3, ZrO_2, or TiO_2 are not highly reactive in aqueous media, and do not demonstrate bonding to bone. For example, Bioglass degradation is highly dependent on composition. The dissolution behavior of calcium phosphate ceramics depends on their composition, crystallinity, and processing parameters. For example, materials with larger surface areas (e.g., powders) and smaller grain sizes resorb more rapidly due to preferential degradation at grain boundaries. Phase is another important factor, with alpha-tricalcium phosphate and beta-tricalcium phosphate degrading more slowly than hydroxyapatite. Hydrated forms of calcium phosphate are more soluble than nonhydrated forms. In addition, ionic substitutions affect resorption rate; CO_3^{2-}, Mg^{2+}, and Sr^{2+} increase and F^- decreases biodegradation. Finally, low pH conditions seen in infection and inflammation can result in locally active dissolution processes.

Polymers used in medicine include polyethylene, poly(methyl methacrylate), poly(dimethylsiloxane), poly(tetrafluoroethylene), and poly(ethyleneterephylate). These structures contain primarily covalent atomic bonds, and many undergo several *in vivo* degradation processes. Water, oxygen, and lipids may be absorbed by the polymer, which may result in local swelling. Polyamides avidly absorb lipids and undergo a stress-cracking process known as crazing; these materials may swell up to five volume percent, and can serve as locking inserts for screws. Desorption (leaching) of low molecular weight species can occur due to release of species remaining from fabrication or from chain scission processes, including free radical depolymerization and hydrolysis. Hydrolytic- and enzymatic-based degradation processes are also possible. Wettability also has a prominent effect on the degradation rate of polymers. Degradation of hydrophilic polymers occurs by surface recession, and may resemble uniform corrosion of metals. Hydrophobic polymers may absorb water and other polar species. As a result, the amorphous regions may dissolve preferentially to crystalline ones, increasing the surface area and the effective dissolution rate. A process similar to inter-granular corrosion may result, with abrupt loss of integrity and small particle release.

WEAR

Wear is the loss of material as debris when two materials slide against one another, which may result in abrasion, burnishing, delamination, pitting, scratching, or embedding of debris. The study of wear, friction, and lubrication was integrated in a 1966 British Department of Education and Science report into a new branch of science known as tribology. The term biotribology was coined in 1973 by Dowson to describe wear, friction, and lubrication in biological systems (33). Over the past 30 years, biotribologists have considered the wear properties of orthopedic, dental, cardiovascular, ophthalmic, and urologic devices, including artificial joints, dental restorations, artificial vessels, prosthetic heart valves, and urinary catheters.

Much of biotribology research has focused on orthopedic prostheses, including devices that replace the function of the hip, knee, shoulder, and finger joints. Hip prostheses have provided control of pain and restoration of function for patients with hip disease or trauma, including osteoarthritis, rheumatoid arthritis, osteonecrosis, posttraumatic arthritis, ankylosing spondylitis, bone tumors, and hip fractures. Polymers, metals, ceramics, and composites have been used on the bearing surfaces of orthopedic prostheses. At present, there are three material combinations used in hip prostheses: a metallic head articulating with a polymeric acetabular ceramic cup; a metallic head articulating with a metallic acetabular metallic cup; a ceramic head articulating with a ceramic acetabular polymeric cup.

Osteolysis and aseptic loosening (loosening in the absence of infection) are the major causes of hip prosthesis failure. In 1994, the National Institutes of Health concluded that the major issues limiting hip prosthesis lifetime include the long-term fixation of the acetabular component, biological response due to wear debris, and problems related to revision surgery (34). Although problems with acetabular fixation have been significantly reduced in the intervening years, wear and the biological response to wear debris remain major problems that reduce the longevity of hip prostheses.

Wear may affect the longevity and the function of hip and other orthopedic prostheses. Clinical practices, patient-specific factors, design considerations, materials parameters, and tissue-biomaterial interaction all play significant roles in determining implant wear rates (35). The complex interaction between these parameters makes it difficult to determine a relationship between the in vitro properties of biomaterials and the *in vivo* wear performance for joint prostheses. For example, particles produced by wear may excite both local and systemic inflammatory responses. In addition, the function of prostheses may be affected by the shape changes that are

caused by uneven wear of surfaces. More effective collaboration among clinicians, material scientists, and biologists is necessary to understand the underlying biological, chemical, mechanical, and patient related parameters associated with wear of prostheses.

Wear may occur via adhesive, abrasive, fatigue, or corrosive mechanisms (30–33). The wear process for a given medical device is usually a combination of these mechanisms; however, one mechanism often plays a dominant role. The most important wear mechanism in orthopedic prostheses is adhesive wear. Adhesive wear is caused by adhesive forces that occur at the junction between rough surfaces. Adhesive wear may occur at asperities, or regions of unevenness, on opposing surfaces. Extremely large local stresses and cold welding processes may occur at the junctions between materials. Material may be transferred from one surface to the other as a result of relative motion at the junction. The transferred fragments may be either temporarily or permanently attached to the counterface surface. During this process, the volume of wear material produced is proportional to both the sliding distance acting on the device and the load. The volume of wear materials produced is also inversely proportional to the hardness of the material. For acetabular hip and tibial knee prostheses, adhesive wear is dependent on the large-strain deformation of polyethylene. For acetabular components under multiaxial loading conditions, plastic strain is locally accumulated until a critical strain is reached. Adhesive wear and submicron wear particle release occurs if this critical value is exceeded (30). Although adhesive wear is the most commonly occurring wear mechanism, it is also the most difficult one to prevent.

Abrasive wear takes place when a harder material ploughs into the surface of a softer material, resulting in the removal of material and the formation of depressions on the surface of the softer material. In general, materials that possess higher hardness values exhibit greater resistance to abrasive wear; however, the relationship between resistance to abrasive wear and hardness is not directly proportional. Abrasive wear is called two-body wear when asperities on one surface plough into and cause abrasion on the counterface surface (36). For example, hip prosthesis simulator testing has shown a positive correlation between the surface roughness of the metallic femoral head and the amount of wear damage to the polyethylene acetabular cup. Isolated scratches on a metallic counterface may also participate in abrasive wear. Three-body wear can also occur if hard, loose particles grind between two opposing surfaces that possess similar hardness values. These loose particles may arise from the material surfaces or from the immediate environment, and may become either trapped between the sliding surfaces or embedded within one of the surfaces. For example, metal, polymer, or tissue (e.g., bone) particles embedded in a polyethylene-bearing surface may act to produce third-body wear in orthopedic prostheses. The overall rate of abrasive wear in polyethylene, metal, and ceramic orthopedic prosthesis components depends both on the surface roughness of the materials and the presence of hard third-body particles.

Fatigue wear is caused by the fracture of materials that results from cyclical loading (fatigue) processes. Surface cracks created by fatigue may lead to the generation of wear particles. Cracks deeper within the biomaterial may generate larger particles, in a process known as microcracking. This process typically occurs in metal components; however, has been observed in other materials (e.g., polyethylene) as well. Corrosive wear results from chemical or electrochemical reactions at a wear surface. For example, metals may react with oxygen at a wear surface (oxidation). The resulting oxide may have a lower shear strength than the underlying metal, and may exhibit a more rapid wear rate than the surrounding material. The rate of corrosive wear is governed by the reactivity of the biomaterial, the chemical properties of the implant site, and the mechanical activity of the medical device.

A film or layer of lubricant between the two bearing surfaces can serve to reduce frictional forces and wear. Lubrication can be divided into three regimes: full film (hydrodynamic) lubrication, boundary lubrication, and mixed lubrication. In full film lubrication, the sliding surfaces are entirely separated by a lubricant film that is greater in thickness than the roughness of the surfaces. In boundary lubrication, the surfaces are separated by an incomplete lubricant film, which does not prevent contact by asperities on the surfaces. A mixed lubrication is the one that encompasses aspects of full film and boundary lubrication, in which a region of the two surfaces exhibits boundary lubrication, and the remainder exhibits full film lubrication. The healthy synovial joint provides a low wear and low friction environment, which may exhibit combination of these lubrication modes. Under normal conditions, the hip, knee, and shoulder joints exhibit full film lubrication, in which the two opposing surfaces are entirely separated by a lubricant film of synovial fluid, which carries the load of the joint.

Wear testing is an important consideration during the development of novel biomaterials and medical devices. Any changes in biomaterial or implant design parameters, including composition, processing, and finishing, should be accompanied by studies that confirm that these changes provide either equivalent or improved wear performance to the implant under clinical conditions. As mentioned earlier, asperities on the contact surfaces generally have a significant effect on overall wear performance. In addition, wear has been described as an accumulative process, because overall wear behavior is highly dependent on the material and testing history. An isolated event during a wear test (e.g., the presence of a third-body wear particle) may have a significant impact on the behavior that is observed.

Wear can be assessed in several ways, including which involve changes in shape (dimensions), size, weight of the implant, weight of the debris, or location of radioactive tracers (37). A standard parameter, known as a wear factor, can be used to estimate the wear effects obtained from different wear tests. The wear factor (K) is defined as

$$K = V/LX \tag{4}$$

in which V is the volume of wear (mm^3), L is the applied load (N), and X is the sliding distance (m). Many parameters can influence the results of wear testing,

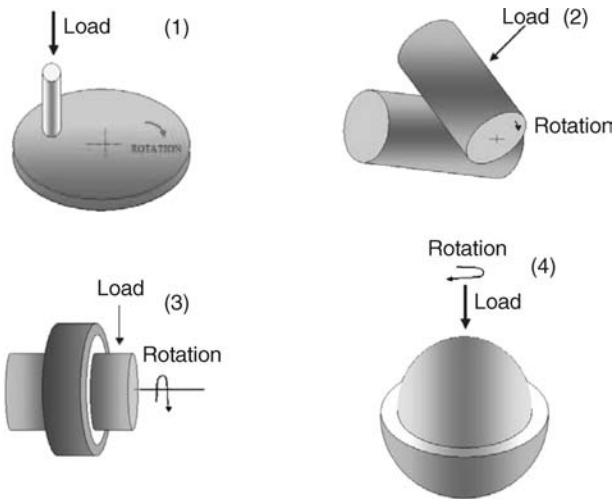


Figure 4. Schematic illustration of geometries in which wear phenomena are likely to occur: (1) pin-on-disk; (2) crossed cylinder; (3) journal bearing; and (4) ball-and-socket bearing (After Ref. 31.)

including test lubricants, test duration, sliding velocity, contact area, alignment, and vibration.

Wear studies fall under three broad categories: (a) screening studies that involve testing of materials with simple geometries under well-controlled conditions; (b) simulator studies that involve testing of partial or complete prostheses; and (c) *in vivo* and retrieval studies of complete implanted devices. Screening studies may provide a basis for comparing novel materials against established materials; however, they can only provide estimations for wear of medical device components. Screening studies involve four general types of geometries: (1) pin-on-disk; (2) crossed cylinder; (3) journal bearing; and (4) spherical or ball and socket bearing. Geometries (3) and (4) are most similar to those encountered in orthopedic protheses (Fig. 4) (35). Simulator studies can be used to assess biomaterials and compare the wear characteristics of materials within a medical device. Various design and material combinations can be examined prior to animal studies and the clinical trials. Clinical assessments and implant retrieval studies also provide useful information for improving biomaterials, medical device design, and manufacturing protocols.

Much of current biotribology research focuses on the relationship between wear performance and biomaterial properties, including composition, processing, and finishing. However, other parameters have significant impact on the wear performance of the prostheses, including surgical and patient factors. The wear performance of polymer, metal, and ceramic biomaterials is discussed below.

Ultrahigh molecular weight polyethylene is commonly used in load-bearing components of total joint prostheses (38–45). The use of a polyethylene/cobalt–chromium wear couple in orthopedic prosthesis was first advocated by Charnley. In many contemporary total hip prosthesis designs, an ultrahigh molecular weight polyethylene acetabular cup slides against a cobalt–chromium alloy femoral ball. Significant numbers of submicron-sized ultrahigh molecular weight polyethylene wear particles are commonly released from these prostheses with each movement of the joint. These particles may remain in the synovial fluid that serves to lubricate the joint (and contribute to third-body wear), embed in prosthesis surfaces, or enter lymphatic circulation. Immune cells (e.g., macrophages) may identify these particles as foreign materials and initiate an inflammatory response, which can lead to rapid bone loss (osteolysis), prosthesis loosening, or bone fracture (39). The volume and size of wear particles are critical factors that affect macrophage activation (40). These biological and physical effects of ultra-high molecular weight polyethylene wear particles are presently the leading cause of long term failure for metal-on-polyethylene hip prostheses (41,42).

Several mechanisms have been proposed to describe wear of ultrahigh molecular weight polyethylene prosthesis components. The wear mechanism of ultrahigh molecular weight polyethylene in hip prostheses has been described by Jasty et al. (43). They found that ultrahigh molecular weight polyethylene surfaces of retrieved implants contained numerous elongated fibrils, which were indicative of large strain deformation. This plastic deformation resulted from strain hardening of the material in the sliding direction and weakening of the material in the transverse direction. Once strain deformation of the surface has occurred, the surface will fragment during the relative motion, and micron- and submicron-sized wear particles will be released. Subsurface cracking, pitting, and delamination caused by oxidative embrittlement and subsurface stresses are responsible for wear of ultra-high molecular weight polyethylene tibial knee inserts.

The wear resistance of ultrahigh molecular weight polyethylene can be improved by reducing the plastic-strain deformation and increasing the oxidization stability (30). The large-strain plastic deformation of ultrahigh molecular weight polyethylene can be diminished by increasing the number of covalent bonds between the long molecular chains of the polymer, which reduces the mobility of the polymer chain and minimizes the creep of the polymer. This process can be achieved by chemical methods (e.g., silane reactions) or, more commonly, by exposing polyethylene to ionizing radiation (46–50). Gamma-ray, e-beam, or X-ray radiation is used to cleave C=C and C=H bonds in polyethylene, which leads to the formation of species with unpaired electrons (free radicals). If the carboncarbon bond is cleaved (chain scission), the polymer molecular weight is reduced. Cross-linking can occur if free radicals from separate chains react with one another, and form an inter-chain covalent bond. If cross-linking occurs as a result of recombination by two radicals cleaved from C=H bonds, it is referred to as an H-type cross-link. If one of the free radicals comes from the cleavage of the C=C bond, it is referred to as a Y-type cross-link. The Y-type cross-linking process can increase the extent of polymer side chain branching (51). The yield of cross-linking processes has been estimated to be three times greater than the yield of chain scission processes for radiation/ultrahigh molecular weight polyethylene interaction. Cross-linking is most significant in amorphous regions of ultrahigh molecular weight polyethylene. An 83% reduction in wear rate has

been reported for ultrahigh molecular weight polyethylene surfaces treated with 5 Mrad radiation (38).

However, not all of the free radicals recombine with other free radicals. In crystalline regions, where the spatial separation between free radicals is large, the residual free radicals become trapped. These species are often confined to the crystalline-amorphous interfaces (52,53). Residual free radicals can cause long-term embrittlement through a series of complex cascade reactions. The residual free radicals first react with oxygen, leading to the formation of oxygen-centered radicals. The oxygen-centered radicals can take a hydrogen atom from a nearby chain to form a hydroperoxide species and another free radical on a chain. This additional free radical can repeat the process by generating another hydroperoxide and forming another free radical on a chain. These unstable species may decay into carbonyl species after exposure to high temperatures or after long periods of time, resulting in lower molecular weights and recrystallization. These processes result in increased stiffness, which is highly undesirable for biotribological applications.

Significant research has been done on reducing the concentration of residual free radicals and limiting the brittleness of irradiated ultrahigh molecular weight polyethylene. One cross-linking postprocessing treatment involved annealing the polymer above its melting transition, which allowed the residual free radicals to be removed through recombination reactions. The polymer recrystallized on cooling; however, the covalent bonds obtained during cross-linking were maintained. Unfortunately, the ultrahigh molecular weight polyethylene exhibited slightly lower crystallinity after this treatment. Another treatment involved annealing the cross-linked polymer at a temperature below the peak melting transition. One advantage of this technique is that a greater degree of crystallinity is retained; however, only a partial reduction in the number of residual free radicals is achieved. Other treatments for residual free radicals include irradiation at room temperature followed by annealing at temperatures below the melting transition; irradiation at room temperature with gamma or electron beams followed by melting; or irradiation at high temperatures followed by melting (34).

The physical properties of the ultrahigh molecular weight polyethylene can be significantly altered by cross-linking and annealing treatments. The effect of these treatments is dependent on the cross-linking parameters (e.g., technique, radiation source, dose, temperature during irradiation) and the annealing parameters (e.g., annealing temperature, annealing time). For example, the ultimate elongation (<45%) and the work to failure for ultrahigh molecular weight polyethylene are reduced as the radiation dose level is increased. Large radiation doses also reduce the yield strength (<30%) and the modulus (<27%) of ultrahigh molecular weight polyethylene (34). In addition, toughness decreases as the radiation dose level is increased, since the energy absorption before failure decreases as the chain mobility is reduced (54).

One alternative to the use of ultrahigh molecular weight polyethylene involves the use of so-called metal-on-metal prostheses, which contain two metallic load-bearing components. The primary motivation for use of these implants is friction; metal-on-metal bearings generate less frictional torque during simulated gait than metal-on-polyethylene bearings (55–59). A stainless steel metal-on-metal hip prosthesis design was attempted by Wiles in 1938. Cobalt–chromium alloy/cobalt–chromium alloy prostheses designs were later developed by McKee and Watson-Ferrar (55,56). Although many of these early cobalt–chromium alloy metal-on-metal hip prostheses failed relatively soon after implantation, others have remained in place for >20 years (35). These first-generation metal-on-metal prostheses were displaced by ultrahigh molecular weight polyethylene/cobalt-chromium alloy prostheses in the 1970s for several reasons, including seizure of the cast metal surfaces (56). In the 1980s, second generation cobalt–chromium alloy/cobalt–chromium alloy prostheses were developed, which have again attracted interest from biomaterials researchers and prosthesis manufacturers (58). Earlier problems with seizing have been minimized through the use of wrought alloys, which are prepared using a thermal-mechanical forming process. Scholes et al. recently demonstrated using a hip simulator system that the mode of lubrication in metal-on-metal hip prostheses is strongly influenced by the diameter of the femoral head and diameter clearance (42). In small diameter joints, the wear rate increased as the diameter of the femoral head was increased. These results were attributed to the development of mixed lubrication in this system.

Alumina and zirconia have also been considered for use in orthopedic prostheses. Alumina exhibits very high hardness and elastic modulus values of 1900 $kgf \cdot mm^{-2}$ (Vickers hardness) and 380 GPa, respectively (60). This material is polished to provide an extremely smooth finish; surface roughness values $<0.005\,\mu m$ are routinely obtained. In addition, alumina surfaces are hydrophilic and may provide prostheses with full film lubrication (35). Fracture toughness and wear resistance can be improved by lowering grain size, increasing grain uniformity, increasing purity, and lowering porosity.

Alumina prostheses have demonstrated wear rates $<1\,mm{\cdot}million^{-1}$ cycles during simulator testing (35). In addition, the *in vivo* wear rate for early alumina-on-alumina hip prostheses was shown to be as low as 1 mm·year (61). However, retrieval studies involving early alumina-on-alumina hip prostheses found high rates of wear on some prostheses. Microseparation of the head and cup was shown to be responsible for this *in vivo* wear behavior (62). Insley et al. examined alumina-on-alumina prostheses with a laboratory simulator, and found that many very small (~40 nm) and some large (100–3000 nm) particles were generated under microseparation conditions (35). Zirconia is harder than alumina, and is used to fabricate smaller components that can withstand higher stresses. Deformation-induced phase transformation has a significant effect on the mechanical properties of zirconia (63). The crystalline phase of a pure zirconium changes from monoclinic to tetragonal during deformation, which is accompanied by volume expansion of ~3–4%. The addition of either yttrium oxide (Y_2O_3) or magnesium oxide (MgO) stabilizes the tetragonal phase at room temperature. However, aging can cause zirconia to return the more stable monoclinic phase, and can limit the lifespan of zirconia

prostheses (64,65). In addition, Sato et al. showed that the tetragonal-monoclinic transformation on the surface of zirconia prostheses can be promoted by the presence of water molecules in the environment (66). The resulting volume change can lead to the generation of surface microcracks and an increase in surface roughness. The failure of some zirconia prostheses has been attributed to this microcracking process. Finally, *in vivo* fracture of some zirconia prostheses has been attributed to variations in sintering.

Thermal oxidation of zirconium alloys has also been used to create biocompatible, corrosion- and wear-resistant surfaces for orthopedic prostheses (67). Wrought zirconium-2.65 weight % niobium alloy contains a two-phase microstructure, which consists of elongated hexagonal alpha-zirconium grains that are bordered by cubic beta-zirconium grains. This material is oxidized for up to 8 h in air at temperatures near 620 °C (the eutectoid temperature). The resulting ~5 μm thick monoclinic ZrO_2 surface contains 40 nm wide × 200 nm long grains that are arranged in a brickwork pattern, which is resilient to grain pull-out and lateral fracture. At the interface between the alloy and the surface oxide, regions of unoxidized niobium in beta-zirconium second-phase grains continue from the alloy into the oxide and serve to anchor the oxide to the alloy. The outermost portion of the oxide surface is burnished to create a smooth bearing surface. The oxide surface provides excellent wear behavior against polyethylene components, with reduced wear particle generation and inflammation.

Diamond-like carbon coatings on orthopedic prostheses can exhibit a wide range of elastic modulus and hardness values, which can be correlated with the fraction of sp^3-hybridized atoms within the coating (68–71). Collins et al. developed a relation between sp^3 fraction and Vickers hardness values for hydrogen-free diamond-like carbon coatings. They found that an sp^3 fraction of 10% corresponded to a hardness value of 2000–3000 Hv, an sp^3 fraction of 50% corresponded to a hardness value of 7000-8000 Hv, and an sp^3 fraction of 100% corresponded to a hardness value of 10,000 Hv (72). Schneider et al. found that hydrogen-free and diamond-like carbon films with sp^3 fractions between 0 and 90% provided elastic modulus values between 300 and 800 GPa. In contrast, typical hardness and elastic modulus values for hydrogenated diamond-like carbon films are <17 and <200 GPa, respectively (73).

The coefficients of friction values for diamond-like carbon coatings depend on ambient humidity, topology, and sliding partner (74). The most important parameter determining the coefficient of friction for hydrogenated diamond-like carbon thin films is relative humidity. Friction values for hydrogenated diamond-like carbon films can be as low as 0.01–0.3 in vacuum conditions, but greatly increase under humid conditions due to incomplete formation of the graphitic transfer surface. This variation in coefficient of friction values can be correlated with hydrogen/carbon ratio in the precursor material. As the hydrogen content in the precursor material increases, the friction coefficient demonstrates a greater positive correlation with ambient humidity. For example, hydrogenated diamond-like carbon films produced from hydrogen-diluted methane demonstrate lower friction coefficients under high humidity conditions than other hydrogenated diamond-like carbon films. On the other hand, hydrogen-free diamond-like carbon thin films maintain low friction coefficients (<0.1) under low and high humidity conditions (75,76).

The combination of high hardness and low coefficient of friction values allows diamond-like carbon coatings to provide significant wear protection to a bulk implant material (71). Hirvonen et al. found that the wear resistance of diamond-like carbon coatings is superior to that of silicon carbide, tungsten carbide–cobalt, silicon nitride, or alumina by factors of 40, 60, 230, and 290, respectively (77). Hydrogen-free diamond-like carbon thin films exhibit wear rates of 10^{-9} $mm^3 \cdot N^{-1} \cdot m^{-1}$, these values are ~100 times lower than those for hydrogenated diamond-like carbon thin films (10^{-7} $mm^3 \cdot N^{-1} \cdot m^{-1}$) under similar testing conditions (78). Many diamond-like carbon substrate materials are significantly softer than the diamond-like carbon coatings; high contact pressures can initiate substrate deformation and coating failure. Nitriding processes can harden the substrate surface, reduce subsurface deformation, and extend diamond-like carbon coating lifetimes (79).

The friction and wear properties of diamond-like carbon-coated metal hip prostheses against diamond-like carbon-coated polyethylene cups have been determined using both screening and simulator techniques (80). For example, Tiainen et al. demonstrated extremely low coefficients of friction for prostheses coated with hydrogen-free diamond-like carbon using a pulsed arc discharge method. They demonstrated coefficients of friction for diamond-like carbon/diamond-like carbon and metal–metal pairs of 0.05 and 0.14, respectively. In addition, they found that wear rate in the diamond-like carbon-coated metal/diamond-like carbon-coated metal wear couple was 10^5–10^6 times lower than that observed in conventional metal–polyethylene and metal–metal wear couples. They also observed that wear of polyethylene in the diamond-like carbon-coated metal/ultrahigh molecular weight polyethylene wear couple was 10–600 times lower than that observed in conventional metal/ultrahigh molecular weight polyethylene wear couples. On the other hand, other investigators have found little difference in wear rates between diamond-like carbon-coated prosthesis materials and conventional prosthesis materials. For example, Sheeja et al. found little difference in wear rates between cobalt-chromium–molybdenum alloy/ultrahigh molecular weight polyethylene and multilayered diamond-like carbon-coated cobalt–chromium–molybdenum alloy/ultrahigh molecular weight polyethylene wear couples (81,82). The seemingly contradictory results suggest other factors, such as use of lubricant, may play a significant role in determining wear rates (83–85). For example, physiologic lubricants may not allow graphitic layers to form on the surfaces of the test materials. In addition, diamond-like carbon coatings that contain particulates and pits may increase adhesive wear (86).

Adhesion of diamond-like carbon coatings is dependent on several factors, including film stress, film–substrate chemical bonding, and substrate topology (87,88). Large internal compressive stresses (as high as 10 GPa) are typically observed in

diamond-like carbon coatings. These stresses limit maximum diamond-like carbon coating thickness to 0.1–0.2 mm and prevent widespread medical use. Lifshitz et al. attributed these stresses to subplantation (low energy subsurface implantation) of carbon ions during coating formation (89). They suggested that carbon ions with energies between 10 and 1000 eV undergo shallow implantation to depths of 1–10 nm during growth of diamond-like carbon coatings. Carbon species are trapped in subsurface sites due to restricted mobility, leading to the development of very large internal compressive stresses.

Diamond-like carbon can be alloyed with metals in order to reduce internal compressive stresses and promote specific biological responses (90,91). Diamond-like carbon–metal composite coatings retain hardness and wear properties similar to those of unalloyed diamond-like carbon films, and exhibit excellent adhesion to metal alloy substrates (Fig. 5). The metal component can provide additional biological functionality to the implant surface; for example, silver has been shown to possess a wide antimicrobial spectrum against a broad range of Gram-negative bacteria (including *Pseudomonas aeruginosa*),

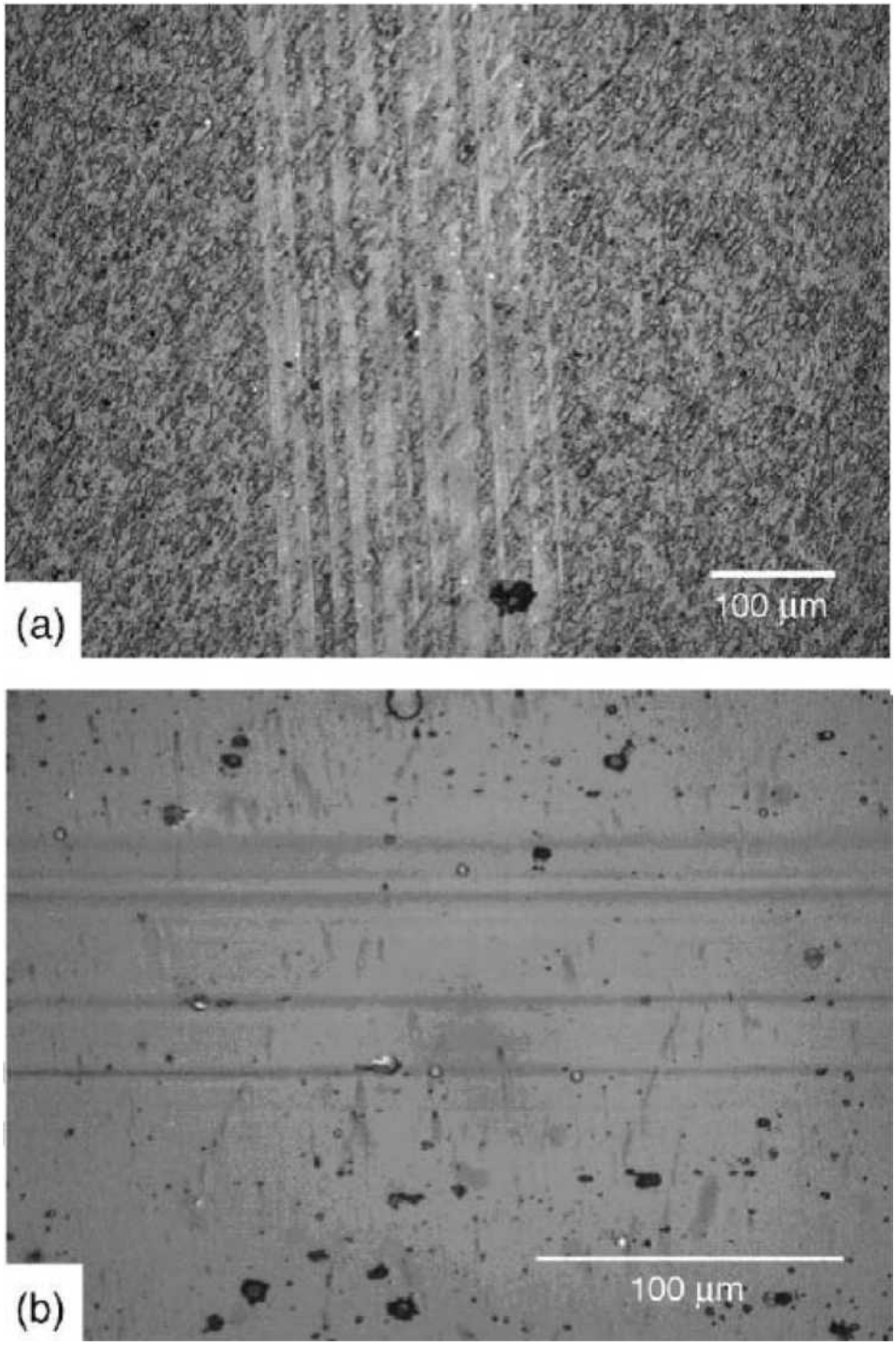


Figure 5. Wear testing of diamond-like carbon–metal composite coatings using a linear tribometer (screening wear test). (a) Wear track of functionally gradient diamond-like carbon–silver composite film after 10,000 cycles, (b) Wear track of functionally gradient diamond-like carbon–titanium composite film after 10,000 cycles.

Gram-positive bacteria (including methicillin-resistant *Staphylococcus aureus*), fungi, viruses, and yeasts. Films containing both silver and platinum may demonstrate enhanced antimicrobial activity due to formation of a galvanic couple that accelerates silver ion release. Narayan et al. showed that diamond-like carbon–silver–platinum nanocomposite films reduce bacterial colonization rates by 90% compared to uncoated silicon substrates (92).

Corrosion and wear are critical parameters that affect overall success of a biomaterial or a medical device design. The complicated interaction between material-, device-, surgical-, patient-specific parameters has made it difficult to predict clinical behavior of implanted medical devices. In addition, the large variation in measurement techniques for corrosion and wear has led problems in interpreting and comparing the work performed in the biomaterials community. If these issues can be successfully resolved, significant advances in these areas may be achieved in the coming years.

BIBLIOGRAPHY

Cited References

1. Burke DR. The Composition and Function of Body Fluids, 3rd ed. St. Louis: C. V. Mosby; 1980.
2. Lentner C, editor. Geigy Scientific Tables, Vol. 1: Units of Measurement, Body Fluids, Composition of the Body, Nutrition. Basel: CIBA-Geigy; 1981.
3. Lentner C, editor. Geigy Scientific Tables, Vol. 3: Physical Chemistry, Composition of Blood, Hematology, Somatometric Data. Basel: CIBA-Geigy; 1984.
4. ASTM, Annual Book of ASTM Standards. West Conshohocken (PA): ASTM. International; 1977.
5. Pilliar RM. Metals and irthopaedic implants — past successes, present limitations, future challenges. Shrivastava S, editor. Proceedings of the Materials & Processes for Medical Devices Conference. Materials Park (OH): ASM International; 2004.
6. Hoar TP, Mears DC. Corrosion-resistant alloys in chloride solutions: materials for surgical implants. Proc R Soc London 1966;294:486–510.
7. Sury P, Semlitsch M. Corrosion behavior of cast and forged cobalt-based alloys for double-alloy joint endoprostheses. J Biomed Mater Res 1978;12:723–741.
8. Steinemann S. Corrosion of surgical implants – *in vivo* and *in vitro* tests. In: Winder GD, editor. Evaluation of Biomaterials. Chichester: John Wiley & Sons Inc.; 1980.
9. Cahoon JR, Bandyopadhya R, Tennese L. The concept of protection potential applied to the corrosion of metallic orthopedic implants. J Biomed Mater Res 1975;9:259–264.
10. Bandy CR. Effects of composition on the electrochemical behavior of austenitic stainless steel in Ringer's solution. Corrosion (Houston) 1977;33:204–208.
11. Scales JT, Winter GD, Shirley HT. Corrosion of orthopaedic implants. J Bone Joint Surg 1959;41B:810–820.
12. Merritt K, Brown SA. Release of hexavalent chromium from corrosion of stainless steel and cobalt-chromium alloys. J Biomed Mater Res 1995;29:627–633.
13. McKellop HA, Sarmiento A, Brien W, Park SH. Interface corrosion of a modular head total hip prosthesis. J Arthroplasty 1992;7:291–294.
14. Brown SA, et al. Fretting corrosion accelerates crevice corrosion of modular hip tapers. J Appl Biomater 1995;6:19–26.
15. Kawale JS, Brown SA, Payer JH, Merritt K. Mixed-metal fretting corrosion of Ti6Al4V and wrought cobalt alloy. J Biomed Mater Res 1995;29:867–873.
16. Metikos-Hukovic M, Kwokal A, Piljac J. The influence of niobium and vanadium on passivity of titanium-based implants in physiological solution. Biomaterials 2003;24: 3765–3775.
17. Cutright DE, et al. Tissue reaction to nitinol wire alloy. Oral Surg Oral Med Oral Pathol 1973;35:578–584.
18. Kapanen A, Ryhänen J, Danilov A, Tuukkanen J. Effect of nickel-titanium shape memory metal alloy on bone formation. Biomaterials 2001;22:2475–2480.
19. Ryhänen J, et al. Bone healing and mineralization, implant corrosion, and trace metals after nickel-titanium shape memory metal intramedullary fixation. J Biomed Mater Res 1999;47:472–480.
20. Villermaux F, et al. Corrosion resistance improvement of NiTi osteosynthesis staples by plasma polymerized tetrafluorethylene coating. Biomed Mater Eng 1996;6:241–254.
21. Rondelli G, Vincentini B. Localized corrosion behavior in human body fluids of commercial NiTi orthodontic wires. Biomaterials 1999;20:785–792.
22. Carroll W, Kelly M, Brien B. Corrosion behavior of Nitinol wires in body fluid environment. Int Conf Shape Memory Superelastic Technol 1999; 240–249.
23. Montero-Ocampo C, Lopez H, Salinas RA. Effect of compressive straining on corrosion resistance of a shape memory Ni-Ti alloy in Ringer's solution. J Biomed Mater Res 1996;32: 583–591.
24. Shabalovskaya SA. Surface, corrosion and biocompatibility aspects of Nitinol as an implant material. Bio-Med Mater Eng 2002;12:69–109.
25. Niinomi M. Recent research and development in titanium alloys for biomedical applications and healthcare goods. Sci Technol Adv Mat 2003;4:445–454.
26. Senda T, Yasuda E, Kaji M, Bradt RC. Effect of Grain Size on the Sliding Wear and Friction of Alumina at Elevated Temperatures. J Am Ceram Soc 1999;82:1505–1511.
27. Dogan CP, Hawk JA. Role of composition and microstructure in the abrasive wear of high-alumina ceramics. Wear 1999; 225:1050–1058.
28. Rodríguez J, et al. Sliding wear of alumina/silicon carbide nanocomposites. J Am Ceram Soc 1999;55:2252–2306.
29. Webster TJ, Siegel RW, Bizios R. Osteoblast adhesion on nanophase ceramics. Biomaterials 1999;20:1221–1227.
30. Morsi K, Keshavan H, Bal S. Hot pressing of graded ultrafine-grained alumina bioceramics. Mater Sci Eng A 2004;386:384–389.
31. Kingery WD. Introduction to Ceramics. New York: John Wiley & Sons Inc.; 1976.
32. Tiainen VM. Amorphous Carbon as a Bio-mechanical coating-mechanical properties and biological applications. Diamond Related Mater 2001;10:153–160.
33. Hutchings IM. Biotribology-A Personal View. In: Hutchings IM, editor. Friction, Lubrication and Wear of Artificial Joints. Bury St. Edmunds (UK): Professional Engineering Publishing Ltd.; 2003.
34. Wright TM, Goodman SB, editors. Implant Wear in Total Joint Replacement: Clinical and Biologic Issues, Material and Design Considerations. Rosemont (IL): American Academy of Orthopaedic Surgeons; 2001.
35. Hutchings IM, editor. Friction, Lubrication and Wear of Artificial Joints. Bury St. Edmunds (UK): Professional Engineering Publishing Ltd; 2003.
36. Schmalzreid TP, Callaghn JJ. Current concepts review: Wear in total hip and knee replacement. J Bone Joint Surg Am 1999;81:115–136.

37. Clarke IC, McKellop HA. Wear Testing. In: von Recum AF, editor. Handbook of biomaterials evaluation: scientific, technical, and clinical testing of implant materials. New York: Macmillan; 1986.
38. McKellop H, Shen FW, DiMaio W, Lancaster JG. Wear of gamma-crosslinked polyethylene acetabular cups against roughened femoral balls. Clin Orthop 1999;369:73–82.
39. Unsworth A, Dowson D, Wright V. Some new evidence on human joint lubrication. Ann Rheum Dis 1975;34:277–285.
40. Green TR, et al. Effect of size and dose on bone resorption activity of macrophages by *in vitro* clinically relevant ultra high molecular weight polyethylene particles. J Biomed Mater Res 2000;B53:490–497.
41. Ingham E, Fisher J. Biological reactions to wear debris in total joint replacement. Proc Inst Mech Eng H 2000;214:21–37.
42. Scholes C, Unsworth A. Comparison of friction and lubrication of different hip prostheses. Proc Inst Mech Eng H 2000;214:49–57.
43. Jasty MJ, et al. Wear of polyethyene acetabular components in total hip arthroplasty: an analysis of 128 components retrieved at autopsy or revision operation. J Bone Joint Surg Am 1977;79:349–358.
44. Charnley JC. Arthroplasty of the hip: a new operation. Lancet 1961;280:1129–1132.
45. Charnley JC. Tissue reaction to polytetrafluorethylene. Lancet 1963;285:1379.
46. Collier JP, et al. Overview of polyethylene as a bearing material: comparison of sterilization methods. Clin Orthop Rel Res 1996;333:76–86.
47. Wang A, Sun DC, Stark C, Dumbleton JH. Wear Mechanisms of UHMWPE in total joint replacements. Wear 1998;181–183:241–249.
48. Collier JP, et al. Results of implant retrieval from postmortem specimens in patients with well-function, long term total hip replacement. Clin Orthop 1992;274:97–112.
49. Cameron HU. Tibial component wear in total knee replacement. Clin Orthop 1994;309:29–32.
50. Atkinson JR, Cicek RZ. Silane crosslinked polyethylene for prosthetic applications: Part II. Creep and wear behavior and a preliminary molding test. Biomaterials 1984; 5:326–335.
51. Muratoglu OK, et al. Larger diameter femoral heads used in conjunciton with a highly cross-linked ultra high molecular weight polyethylene: A New Concept. J Arthroplasty 2001;16: 24–30.
52. Pearson RW. Mechanism of the radiation crosslinking of polyethylene. J Polym Sci 1957;25:189–200.
53. Zhu QR, Horii F, Kitamaru R, Yamaoka H. C-13-NMR study of cross-linking and long-chain branching in linear polyethylene induced by Co-60 gamma-ray irradiation at different temperatures. J Polym Sci, Part A: Polym Chem 1990;28: 2741–2751.
54. Premnath V, Harris WH, Jasty M, Merrill EW. Gamma sterilization of UHMWPE articular implants: an analysis of the oxidation problem. Biomaterials 1996;17:1741–1753.
55. Steinberg DR, Steinberg ME. The early history of arthroplasty in the United States. Clin Orthop Relat Res 2000;374:55–89.
56. McKee GK, Watson-Farrar, J. Replacement of arthritic hip by the McKee-Farrar prostheses. J Bone Joint Surg 1966;48B: 245–259.
57. Wimmer MA, et al. The acting wear mechanisms on metal-on-metal hip joint bearings: in vitro results. Wear 2001; 250:129–139.
58. Dorr LD, et al. Total hip arthroplasty with use of the metasul metal-on-metal articulation: 4–7-year results. J Bone Joint Surg Am 2000;82:789–798.
59. Poggie RA, Afflitto R, St John K. The wear performance of precision Co–Cr–Mo alloy metal-on-metal hip bearings. Proc Conf Trans 12th Ann Int Symp Technol Arthroplasty 1999;11: 1–2.
60. Boutin P, et al. The use of dense alumina- alumina ceramic combination in total hip replacement. M J Biomed Mater Res 1988;22:1203–1232.
61. Willmann G. Development in medical-grade alumina during the past two decades. J Mater Process Technol 1996;56:168–176.
62. Nevelos J, et al. Microseparation of the centers of alumina-alumna artificial hip joints during simulator testing produces clinically relevant wear and patterns. J Arthroplasty 2000;15: 793–795.
63. Christel P, et al. Mechanical properties and short-term In vivo evaluation of yttrium-oxide-partially-stabilized zirconia. J Biomed Mater Res 1989;23:45–61.
64. Chevalier J, et al. Critical effect of cubic phase on aging in 3 mol% yttria-stabilized zirconia ceramics for hip replacement prosthesis. Biomaterials 2004;25:5539–5545.
65. Gremillard L, et al. Modeling the aging kinetics of zirconia ceramics. J Eur Ceram Soc 2004;24:3483–3489.
66. Sato T, Ohtaki S, Endo T, Shimada M. Science and technology of zirconia. Advances in Ceramics. Westerville (OH): American Ceramic Society; 1988.
67. Hobbs LW, et al. Oxidation microstructures and interfaces in the oxidized zirconium knee. Int J Appl Ceram Technol 2005;2:221–246.
68. Enke K, Dimigen H, Hubsch H. Frictional properties of diamondlike carbon layers. Appl Phys Lett 1980;36:291–292.
69. Pharr GM, et al. Hardness, elastic modulus, and structure of very hard carbon films produced by cathodic-arc deposition with substrate pulse-biasing. Appl Phys Lett 1996;68:779–781.
70. Ronkainen H, Varjus S, Koskinen J, Holmberg K. Differentiating the tribological performance of hydrogenated and hydrogen-free DLC coatings. Wear 2001;249:260–266.
71. Holmberg K, Mathews A. Coatings tribology: a concept, critical aspects, and future directions. Thin Solid Films 1994;253:173–178.
72. Collins CB, et al. Noncrystalline films with the chemistry, bonding, and properties of diamond. J Vac Sci Technol B 1993;11:1936–1941.
73. Schneider D, Schwarz T, Scheibe HJ, Panzner M. Nondestructive evaluation of diamond and diamond-like carbon films by laser induced surface acoustic waves. Thin Solid Films 1997;295:107–116.
74. Erdemir A, Bindal C, Pagan J, Wilbur P. Characterization of Transfer layers on Surfaces Sliding Against Diamond-like Hydrocarbon Films in Dry Nitrogen. Surf Coatings Technol 1995;76–77:559–563.
75. Liu Y, Erdemir A, Meletis EI. An investigation of the relationship between graphitization and frictional behavior of DLC coatings. Surface Coatings Technol 1996;86:564–568.
76. Oguri K, Arai T. Friction mechanisms of Diamond-like carbon with silicon coatings formed by plasma-assisted chemical vapor-deposition. J Mater Res 1992;7:1313–1316.
77. Hirvonen JP, Koskinen J, Lappalainen R, Anttila A. Preparation and properties of high density hydrogen free hard carbon films with direct ion beam or arc discharge deposition. Mater Sci Forum 1990;52:197.
78. Voevodin AA, Donley MS, Zabinski JS, Bultman JE. Mechanical and tribological properties of diamond-like carbon coatings prepared by pulsed laser deposition. Surface Coatings Technol 1995;77:534–539.
79. Voevodin AA, Phelps AW, Zabinski JS, Donley MS. Friction induced phase transformation of pulsed laser deposited diamondlike carbon. Diamond Related Mater 1996;5:1264–1269.

80. Santavirta S, Lappalainen R, Heinonen H, Anttila A. Some relevant issues related to the use of amorphous diamond coatings for medical applications. Diamond Related Mater 1998;7:482–485.
81. Sheeja D, et al. Mechanical and tribological characterization of diamond-like carbon coatings on orthopedic materials. Diamond Related Mater 2001;10:1043–1048.
82. Sheeja D, Tay BK, Lau SP, Nung LN. Tribological characterization of diamond-like carbon coatings on Co-Cr-Mo alloy for orthopaedic applications. Surface Coatings Technol 2001;146: 410–416.
83. Ahlroos T, Saikko V. Wear of prosthetic joint materials in various lubricants. Wear 1997;211:113–119.
84. Saikko V, Ahlroos T. Phospholipids as boundary lubricants in wear tests of prosthetic joint materials. Wear 1997;207:86–91.
85. Affatato S, Frigo M, Toni A. An *in vitro* investigation of diamond-like carbon as a femoral head coating. J Biomed Mater Res 2000;53:221–226.
86. Dong H, Shi W, Bell T. Potential of improving tribological performance of UHMWPE by engineering the Ti6Al4V counterfaces. Wear 1999;229:146–153.
87. Morshed MM, McNamara BP, Cameron DC, Hashmi MSJ. Effect of surface treatment on the adhesion of DLC film on 316L stainless steel. Surface Coatings Technol 2003;163:541–545.
88. Schwan J, et al. Stress-induced formation of high-density amorphous carbon thin films. J Appl Phys 1997;82:6024–6030.
89. Lifshitz Y, et al. Growth mechanisms of DLC films from C+ ions- experimental studies. Diamond Related Mater 1995;4: 318–323.
90. Narayan RJ, Scholvin D. Nanostructured carbon-metal composite films. J Vac Sci Technol B 2005;23:1041–1046.
91. Narayan RJ. Pulsed laser deposition of functionally gradient diamondlike carbon-metal nanocomposites. Diamond Related Mater 2005;14(8):1319–1330.
92. Narayan RJ, et al. Antimicrobial properties of diamond-like carbon-silver-platinum nanocomposite thin films. J Mater Eng Perform 2005;14:435–440.

See also BIOMATERIALS, TESTING AND STRUCTURAL PROPERTIES OF; HEART VALVE PROSTHESES; HIP JOINTS, ARTIFICIAL.

BIOMATERIALS FOR DENTISTRY

STEVE ARMSTRONG
University of Iowa

INTRODUCTION

Gold has been used for dental purposes for at least 2500 years; the fabrication of gold crowns and bridgework flourished in Etruria and Rome as early as 700–500 BC. Gold leaf came into use during the fifteenth century for the restoration of carious teeth. Restorative materials and techniques continued to develop through the nineteenth century including the use of waxes, fused porcelain, "silver paste" amalgam, cements, vulcanite, the angle handpiece, and a gold inlay casting machine. A rapid development in materials and instrumentation has occurred since the 1950s, to include the high speed handpiece, steel and diamond cutting instruments, adhesive techniques to metal, ceramics, enamel and dentin, resin-based composites, glass ionomers, base-metal alloys for partial dentures, metal-ceramic systems, high-strength all ceramic structures, and titanium alloys for dental implants. This increasing complexity and body of knowledge has lead to the establishment of uniform material standards and the recognition of the science of dental materials as a distinct and essential branch of dentistry.

Biomaterials are used in the oral cavity either to restore function, comfort, or aesthetics caused by developmental disorders, disease, or trauma. More elective procedures are being requested and performed purely for aesthetic purposes as the incidence of caries has dropped in certain population groups and as patients have become more aware of various restorative or cosmetic options. However, the replacement of diseased tooth structure or missing teeth accounts for the bulk of work in restorative dentistry. The instruments and materials used in the surgical aspects of oral, maxillofacial and periodontal surgery have much in common with medicine. This article will focus on those commonly used materials in the restoration of individual teeth or the replacement of missing teeth.

Restorative materials include noble and base metal alloys, resin-based composites (RBCs), glass ionomers, ceramics, acrylics, and amalgam alloys. Techniques to apply these materials include both direct and indirect approaches. Materials or "fillings" can be directly placed in a prepared cavity by the use of adhesives and/or retentive-type preparations. Full or partial coverage crowns, bridges, and dentures are fabricated indirectly by dental laboratories or computer aided milling machines and then attached or cemented into the mouth for the coverage of missing or weakened tooth structure or the replacement of missing teeth. Various forms of ceramic or metallic implants can be placed into the upper or lower jaw bones to serve as tooth root substitutes upon which a prosthesis is attached to replace missing teeth. Auxiliary materials, such as waxes, gypsum, and impression materials are also utilized during clinical and laboratory steps but will not be covered in this article.

DIRECTLY PLACED RESTORATIVE MATERIALS: 'FILLINGS'

Amalgam

Dental amalgam has been very successfully used in dentistry for 150 years and is one of the most technique insensitive dental restorative materials available. Dental amalgams are inexpensive and have demonstrated a relatively long service life. The disadvantages of amalgam are the silver color and the presence of mercury. The presence of mercury requires regulatory control of wastewater effluent and has raised unsubstantiated health concerns regarding mercury toxicity to the individual patient.

Dental amalgam is a mixture of mercury and a solid metal alloy of silver, tin, copper, and sometimes zinc, palladium, indium, and selenium. Once the mercury and alloy is mixed, the plastic mass is condensed into the prepared cavity and carved to required form before hardening. The alloy particles are microspheres of various sizes, irregular lathe-cut particles or mixtures of the two. "High

copper" alloys (13–30%) have essentially replaced the low copper alloys of the past. These high copper alloys, along with the addition of zinc for manufacturing procedures, have improved early clinical strength, lowered creep, and improved corrosion resistance. The mixing of mercury and the alloy is referred to as trituration or amalgamation. A surface reaction occurs between the alloy and liquid mercury that binds the unreacted particles together by a surrounding matrix of reaction products. Increasing the copper content eliminates most of the weakest and most corrosive phase ($Sn_{7-8}Hg$) from the setting reaction.

After the carious lesion is removed, the plastic dental amalgam is condensed into the prepared cavity before hardening and subsequently retained by the mechanical resistance and retention form of the surgically prepared cavity. An adhesive liner is not required. The material slowly corrodes and the corrosion products "self-seal" the margin between tooth and amalgam, thereby protecting the tooth from leakage of oral fluids and bacteria and their byproducts. Dental amalgam is brittle and undergoes creep at mouth temperature, which can lead to marginal or bulk fracture and clinical failure. However, if the cavity is well designed and the amalgam placed with technical competence, many years of service should be expected. A vast number of studies have shown the safety and efficacy of dental amalgam. When a dentist is faced with a patient's request to remove amalgam fillings due to a claimed medical malady, the dentist is professionally obligated to explain that the possibility of their medical condition(s) being related to the presence of dental amalgam fillings is extremely remote. These patients typically face a complex problem with biological, psychological, and social components unrelated to mercury intake.

Resin Composites

Modern day resin-based composites placed with dental adhesives have replaced the silicates and acrylic resins of the past and are now widely used throughout dentistry. In addition to the treatment of decay and trauma, RBCs are used in aesthetic or cosmetic dentistry procedures due to their versatility and conservative nature. Discolored, misshapen, or misaligned dentition can be aesthetically treated with cosmetic "bonding". The RBCs are composed of four main components: (*1*) a continuous organic polymer matrix, (*2*) a dispersion of inorganic filler particles, (*3*) silane coupling agents to bind the filler particles with the polymer matrix, and (*4*) an initiator–accelerator system. They also contain various pigments for matching tooth shades and ultraviolet (UV) absorbers to minimize oxidative color changes. The two most common oligomers are the dimethacrylates 2,2- bis[4(2-hydroxy-3-metharcyloyloxy-propyloxy)-phenyl] (bis-GMA) and urethane dimethacrylate (UDMA). Diluents such as triethylene glycol dimethacrylate (TEGDMA) are added to reduce the viscosity for the addition of filler and to obtain clinically acceptable handling properties. The inorganic filler particles can be of borosilicate, lithium aluminum silicate, barium aluminum silicate, strontium or zinc glasses, quartz, or colloidal silica. The combination of relatively larger glass or quartz particles and a significant addition of colloidal silica are referred to as a "hybrid". A useful classification method is by filler particle size (Table 1) with minifill hybrids and microfills being the most popular types. Recently, using a proprietary process, manufacturers have been able to produce a smaller average silica particle size (0.02 μm) as compared with traditional microfills (0.04 μm). Marketed as "nanofills", these smaller silica particles are produced in a nondrying method thereby avoiding agglomeration due to physical forces and thusly enabling a higher degree of filler loading. The microfill RBCs polish to the most enamel-like surface, but lack the strength of the hybrids due to the lower filler volume. Newer formulations of minifill hybrids can be used as "universal" RBCs possessing both the strength for posterior chewing forces and acceptable surface finish for use in aesthetic anterior regions. Clinical studies showed that the long-term wear resistance of RBC restorations placed on posterior teeth is still inferior to the dental amalgam restorations.

Table 1. Classification of Resin-Based Composites by Filler Particle Size

Class	Particle Size, μm	Filler Loading, vol%
Macrofill	>10	50–70
Midifill	1–10	50–70
Minifill	0.1–1	50–70
Microfill	0.01–0.1	20–50

Polymerization occurs through either a self-cured free radical initiation when a peroxide–amine system is mixed or through light-activated free radical initiation when a diketone–amine system is exposed to blue light. The photoactivator, most commonly camphoroquinone, is added in small amounts, which forms a free radical when exposed to 467-nm blue light. Dual-cure varieties of RBCs are available as well. Halogen light-curing units are most commonly used, but several other light curing units are currently being marketed to include: plasma-arc, laser and light-emitting diodes. To insure proper polymerization care must be taken to match the unit's spectral emission and the RBCs spectral requirements. One aspect that all RBCs currently share is 2–4% volumetric shrinkage of the continuous polymeric network upon polymerization. Shrinkage induces residual stress that can disrupt the adhesive bond between the RBC and the tooth structure, damage enamel or the RBC. Incremental placement of light-cured RBCs can help to reduce the net effect of this shrinkage stress. Manufacturer's are currently working to develop no- or low shrinkage RBCs; several approaches are noted: (*1*) addition of ring-opening monomers that expand upon polymerization (spiroorthocarbonates, oxiranes, vinylcyclopropanes), (*2*) low shrinkage cyclopolymerizable di- and multifunctional oligomers synthesized through the reaction of acrylates and formaldehyde, and (*3*) the addition of strain-absorbing polybutadiene rubber polymer adsorbed onto the fumed silica.

Variations in filler loading, viscosity, and polymerization initiating systems allow RBCs to be used in a wide variety of clinical situations, to include: sealants, cements, crown core buildups, so-called flowables and packables,

provisional restorations, and in a variety of laboratory processed RBCs for adhesive cementation as inlays, onlays, crowns and veneers. Several manufacturer's have promoted fiber reinforced RBCs for use as bridges, as well, but these lack convincing clinical data for their recommended usage.

Glass Ionomers

Smith (1968) developed glass ionomer cements (GIC) by combining the polyacrylic acid from polycarboxylate cement, which is strongly adhesive to tooth structure, and the aluminosilicate glass from the fluoride-containing silicate cements. The GICs form a true chemical bond to the tooth mineral (hydroxyapatite) by ionic bonding between calcium and carboxylic ions and act as chemotherapeutic agents in the treatment and prevention of dental caries through the release of fluoride.

The GIC are composed of a basic ion-leachable aluminosilicate glass powders that, upon exposure to water-soluble homopolymer or copolymers of alkenoic acids, form a matrix of continuous polysalts (polyalkenoates) surrounded by partially solubilized glass filler particles. The clinical placement technique must account for the relatively slow-setting reaction and moisture sensitivity. Fluoride easily passes in and out of the matrix without any degradative effects due to the substitution of carboxyl ions for fluoride within the salt matrix. Fluoride "recharging" has been clearly demonstrated in vitro to prevent the demineralization of tooth structure at the margins adjacent to the GIC under an artificial caries challenge. In vivo evidence is not as clearly demonstrated, but when evaluating clinical data from high caries risk cohort populations the anticariogenic effect of GIC is elucidated.

Since their development GIC has been modified in a number of different ways to improve their clinical handling properties and durability. A significant development was the addition of water soluble monomers, for example, HEMA, and the grafting of methacrylate side groups on the polyacid polymer. By the addition of visible light initiator–accelerator systems these resin-modified glass ionomers (RMGI) can be command set with a light curing unit while also self-curing through the acid–base reaction. These improvements to the conventional GIC and RMGI have made these materials widely used as restorative materials. The self-adhesive and self-cure properties of GIC, along with improvements in strength have allowed these materials to be used in nontraditional field situations without the luxuries of electricity or modern equipment. Auxiliary personnel have been trained to remove caries with sharp hand instruments with the cavity then filled with a heavier filled GIC. This treatment has aided thousands through a technique termed Atraumatic Restorative Treatment or "ART". These materials are also used as cements and liners.

Adhesives, Cements, and Liners

Adhesive dentistry has become increasingly important as the use of dental amalgam and direct compacted gold foil and cemented gold restorations has declined. Unlike dental amalgam, RBCs require an adhesive liner for placement and retention. A durable bond of RBCs to enamel can be accomplished by first cleaning and demineralizing the surface with a 30–40% phosphoric acid, followed by a polymerizable methacrylate monomer [bis(GMA), UDMA, TEGDMA], which diffuses into the porosities created by the acid etching. However, bonding to dentin is a much greater challenge due to the compositional differences in dentin relative to enamel and its extremely variable clinical presentation. Dentin contains less inorganic components and more organic components and water than enamel. Dentin is made permeable by tubules that travel from the dental pulp through the dentin to the coronal enamel.

Similar to enamel, the dentin is treated with an acid that removes any smear layer and exposes the collagen fibers by demineralizing the surface. An adhesive primer containing a hydrophilic solvent (water, acetone, ethanol, or HEMA) and an amphipathic monomer (hydrophilic–hydrophobic functionalities) then penetrates the exposed collagen network. After the solvent is evaporated from the primed surface, an adhesive monomer is applied that attaches to the hydrophobic functionality of the primer to create a wetted surface for subsequent copolymerization with the RBC. This bonding process is also approached with so-called "self-etch" adhesives. The initial acid application is eliminated and a water-soluble acidic polymer is included in the primer to simultaneously demineralize the tooth surface while penetrating with the adhesive monomers and oligomers. Regardless of the approach, the adhesive liner penetrates into the exposed collagen network and also partially into the dentinal tubules. The interdiffusion of the synthetic adhesive polymer within the collagen network forms a micromechanical bond and is commonly referred to as a hybrid layer (Fig. 1). In the last 20 years,

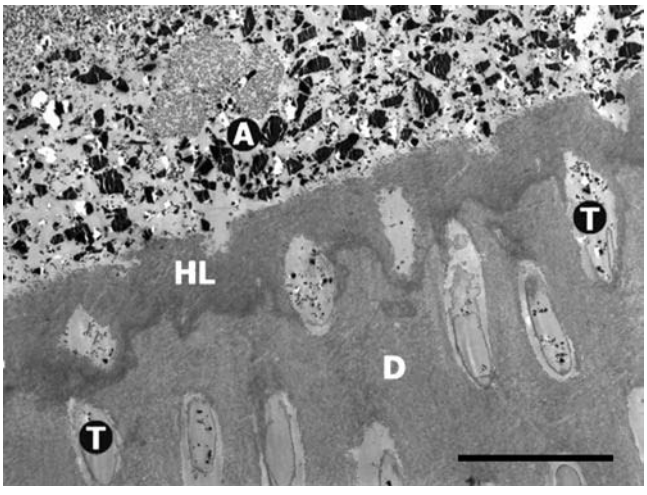


Figure 1. Scanning electron micrograph (SEM) image of a total-etch three-step dental adhesive system applied to dentin. A = filled adhesive resin (ceramic filler particles visible within resin matrix), HL = hybrid layer (interdiffusion zone of adhesive polymers and collagen network), T = dentinal tubules, D = laboratory demineralized dentin. Original magnification = 2000X, black bar=10 μm. (Photomicrograph courtesy of Marcos Vargas, University of Iowa College of Dentistry.)

dentin bonding has improved the clinical success of adhesively placed restorations, but difficulties remain. Completely penetrating the exposed collagen matrix with the adhesive monomers can be hindered by the presence of excess solvent, dentinal fluids, or by physical blockage of the interpenetrating microchannels between collagen fibrils. Even if the adhesive fully wets the dentin surface, suboptimal polymerization may reduce bonding effectiveness. This union of restorative material to enamel and dentin is critical not only for retaining restorations in place, but also for sealing the margin from the passage of bacterial fluids, molecules, or ions. Leakage between the interface of the dental restoration and the wall of the cavity preparation has been associated with marginal discoloration, secondary caries, and pulpal pathology.

Adhesive liners or bonding agents are also used in conjunction with resin-based composite cements when adhesively cementing crowns or other fixed appliances into or onto prepared teeth. The ceramic elements or oxides of the internal surface of dental porcelain–ceramic, metals, and resin-based composite restorative materials are mechanically or chemically roughened before applying a silane coupling agent. The silane bonds to the ceramic surface with both covalent and hydrogen-bonding protecting from hydrolytic degradation while making the surface hydrophobic and organophilic for resin cement wettability and copolymerization. Self-adhesive cements include the glass ionomer and polycarboxylate cements. Zinc phosphate and zinc oxide eugenol are nonadhesive cements that act as mortar or a luting agent.

Additional uses of the above mentioned cements include (*1*) cavity liners to achieve a physical barrier to bacteria and their products and/or to provide a therapeutic effect, such as fluoride release from glass ionomer or pulpal obtundent with zinc oxide eugenol, (*2*) cavity bases to block out undercuts in cavity preparations for indirect restorations or for insulating the pulp from thermal changes, and (*3*) temporary or provisional restorations. The beneficial effects of adhesion to tooth structure and fluoride release obtained from adhesive liners and glass ionomer materials are rapidly replacing the traditional liner, base, and cement materials, that is, zinc phosphate, polycarboxylate, and zinc oxide eugenol.

PROSTHETIC RESTORATIVE MATERIALS: CROWNS, BRIDGES, DENTURES

Metals and Alloys

Noble and base metal alloys are used for (*1*) crowns and bridges with fused porcelain in esthetic areas, (*2*) inlays, onlays, crowns, and bridges without porcelain veneering in the posterior or nonaesthetic regions of the mouth, and (*3*) partial and complete removable denture bases. Base metals commonly used in dental alloys include, nickel, chromium, copper, zinc, gallium, silver, indium, and tin. Silver, a "precious" metal, is not considered a noble metal in dentistry due to its corrosion in the oral cavity. The noble metals utilized in dentistry are gold, platinum, palladium, iridium, rhodium, and ruthenium. Cold-worked or wrought noble and base alloys can be cast with or "soldered" (brazed) to cast structures as attachments or clasps to removable partial dentures. High purity gold, being soft and malleable, can be cohesively adapted in the form of gold foil into small cavity preparations by careful condensation techniques. This process develops adequate hardness and physical properties through work hardening, resulting in a clinically successful, long-lasting restoration. However, the compacted gold foil restoration is becoming increasingly uncommon due to the success of adhesively bonded tooth colored restorations and the skill and time required to properly place gold foil. Almost all fixed-dental prostheses contained a minimum of 75% gold before the dramatic increase in the price of gold after the United States separated gold from monetary standards in 1969. The increase in gold prices, and rise in palladium prices three decades later, has lead to an increased use of alternative alloys containing base metals.

The lost wax technique became common in dentistry after W.H. Taggart introduced his casting machine in 1907. A wax pattern of the desired restoration is fabricated and then invested in a ceramic material (casting investment), which is subsequently heated to burn out the wax pattern, that is "lost wax". A molten metal alloy is then cast into the resultant space previously occupied by the wax. The restoration is recovered, finished, and polished before cementing or delivering to the patient. The investment must be able to expand enough upon setting and heating to compensate for the wax and metal shrinkage if a precise fit is to be obtained.

Alloys should (*1*) produce no toxic, carcinogenic, or allergic reactions; (*2*) resist corrosion and physical changes in the oral environment; (*3*) possess physical properties, that is, strength, fusing temperature, thermal conductivity, and coefficient of thermal expansion, appropriate for the desired application; (*4*) be able to be fabricated in a technically feasible manner; and (*5*) be available and relatively inexpensive. The alloys used for metal-ceramic or commonly termed porcelain-fused-to-metal (PFM) restorations must possess a fusion temperature range that is substantially higher ($>100\,^{\circ}C$) than the ceramic firing temperature, have sufficient creep resistance at that temperature, and have the ability to form a good bond between its oxide surface and the ceramic veneer.

Wrought stainless steel alloys are used in orthodontic brackets and wires, endodontic instruments, prefabricated temporary crowns and space maintainers. In addition, wrought cobalt–chromium–nickel, nickel–titanium, copper–nickel–titanium, and beta-titanium alloys are also used as orthodontic wires. Nickel–titanium and copper–nickel–titanium orthodontic wires have a unique superelastic (pseuedoelastic) property that delivers a constant low-level force over an extended range of deformation. Nickel–titanium and copper–nickel–titanium alloys are also the shape memory alloy (SMA), that is, they can be deformed plastically below its transition temperature range (TTR), then after heating through and above the TTR, they will return to their original desired shape due to a crystallographic transformation from martensitic phase into austenitic phase. Titanium and titanium alloys, especially due to their thin stable oxide layers, are very important endosseous dental implant materials and, with the

recent refinement in casting techniques, can be used for crowns, partial dentures and complete denture bases.

Ceramics

The first porcelain was developed in the T'ang Dynasty from 618 to 906 AD and the first suggested use of porcelain for dentistry was by Pierre Fauchard in France after the porcelain formula was brought from China by a Jesuit priest. Several developments followed but the current approaches to ceramic-metal crowns and bridges occurred from the 1950s–1960s. High-fusing alloys combined with the development of low-fusing thermally compatible leucitic porcelains permitted the fabrication of ceramic-metal restorations. High expansion leucite was mixed with feldspar glass during manufacturing to refine the coefficient of thermal expansion (α) creating a successful junction between dental porcelain and metal. The combination of which must be thermally compatible for fabrication and so that the veneering materials surface is left in a residual state of compression. The α of the veneering material is generally $\sim 0.5–1.0 \times 10^{-6}\,^{\circ}C^{-1}$ lower than the core material so that upon cooling the inner core will contract more resulting in a residual compressive state resisting crack formation and propagation of the relatively brittle veneering material.

Minor refinements have continued with metal–ceramic systems, but more recently significant advances have occurred in the area of "all-ceramic" systems, in which the metal is replaced with a ceramic core upon which veneering porcelain is fused. This eliminates the masking of the metal with opaquing agents and greatly simplifies the aesthetic technical procedures by the dental technician. However, more tooth structure may need to be removed in preparation for the construction of these all-ceramic restorations and their brittle nature does not yet permit their function in long-span bridges.

Ceramic systems can be classified by their fusion temperature, clinical usage, processing methods, or crystalline phase. Table 2 lists some crystalline types used for the three major applications of ceramics in dentistry: (*1*) metal–ceramic crowns and fixed-partial dentures (bridges); (*2*) all-ceramic crowns, inlays, onlays, veneers, and shortspan bridges; and (*3*) denture teeth.

Table 2. Classification of Dental Ceramic Materials

	Fabrication	Crystalline (dispersed) Phase
Ceramic–Metal (PFM)	Sintered	Leucite ($KAlSi_2O_6$)
All-Ceramic	Machined	Alumina (Al_2O_3)
		Feldspar ($KAlSi_3O_8$)
		Mica ($KMg_{2.5}Si_4O_{10}F_2$)
	Slip-cast	Alumina (Al_2O_3)
		Spinel ($MgAl_2O_4$)
		Zirconia (ZrO_2)
	Heat-pressed	Leucite ($KAlSi_2O_6$)
		Lithium disilicate ($Li_2Si_2O_5$)
	Sintered	Alumina (Al_2O_3)
		Leucite ($KAlSi_2O_6$)
Denture Teeth		Feldspar ($KAlSi_3O_8$)

[a]Adapted from Ref. (1) p. 553.

"Porcelains" are composed of kaolin (clay), feldspar, and quartz (flint), while the dental porcelains being quite similar, are fabricated from silica (SiO_2), soda (NaO_2), potash (K_2O), alumina (Al_2O_3), with the addition of pigments, opacifiers and fluxes. Naturally occurring minerals such as feldspar ($K_2O\ Al_2O_3\ 6SiO_2$), quartz, and nepheline syenite have been utilized to provide these constituents. The use of feldspar has lead to the term feldspathic porcelain, however, feldspar is not necessarily present in the final processed porcelain, nor is it essential to form leucite, the major crystalline phase of the porcelain. Like all dental ceramics, dental porcelain is composed of a glassy (vitreous) matrix phase surrounding a dispersed crystalline phase. The glassy matrix, composing 75–85% of the porcelain, is formed by heating the raw materials into a glassy state then quenching. This pyrochemical reaction produces a supercooled liquid of metastable equilibrium that is quenched then ground into a fine powder. These fine powders or frit can be reheated and will fuse at a lower temperature with little pyroplastic flow giving increased homogeneity, translucency, smoother texture, a lower fusion temperature, and less shrinkage. The temperature at which the surface glassy phase softens allowing the fritted particles to coalesce without further pyrochemical change is called sintering. These sintered dental porcelains, in general, will have little change in the physical, chemical, or optical properties of the glassy matrix upon repeated firings during the necessary steps of the restoration fabrication. However, if improperly fired or over fired, the dispersed leucite crystals can be altered leading to reduced strength or porcelain–metal (core) thermal incompatibilities. During the fritting process the silica matrix is disordered due to the rapid cooling from the molten state and also by the addition of fluxes that break up the silica tetrahedral network by occupying oxygen. These alkali ions reduce the number of cross-linkages between the silicon–oxygen tetrahedra by randomly occupying space in the open network. The net effect of flux (LiO_2, Na_2O, K_2O, BaO, CaO, MgO, ZnO) addition is lower softening or fusion temperature, decreased viscosity, production of glassy phase, increased α, decreased strength, lowered chemical resistance, increased risk of devitrification during repeated firing cycles. The lower fusing temperatures and increased α made possible the modern day metal–ceramic systems. Three components of the porcelain to metal bond are classically described: (*1*) mechanical interlocking through good wetting of the porcelain on the roughened metal surface, (*2*) chemical physical bonding between the oxides of the porcelain and the oxides on the metal surface, and (*3*) a controlled mismatch in α leading to residual compressive forces in the porcelain (described earlier). Any of these may predominate depending on the ceramic system.

Achieving superior esthetics, in general, is simplified by having a ceramic core. However, strength, wear, fit, and longevity must be proven in controlled clinical trials. Increasing the strength of the ceramic core to perform comparably to metal substructures is approached by manipulating the crystalline phase for reinforcement.

Techniques for fabricating all-ceramic systems include (*1*) sintering with alumina-based, magnesia- based, and leucite-reinforced ceramics; (*2*) heat-pressed techniques with leucite-reinforced and lithium-disilicate-based ceramics; (*3*) slip-casting with alumina-, spinel-, and zirconia-based ceramics; and (*4*) the machining of manufactured ceramic blocks available in several types of ceramic. One method uses computer-aided designing/computer aided machining (CAD/CAM) technology to fabricate inlays, onlays, veneers, and crowns. An "optical" impression is obtained from the prepared tooth with an optical scanner and the computerized image of the restoration is designed by the computer software. Subsequently, a ceramic block is machined in the computer-controlled milling machine according to the design and later cemented into or on the tooth by the dentist.

Slip-cast all-ceramics are fabricated by a process very similar to that used for the production of common objects such as plumbing fixtures and beer steins. Successive layers of ceramic slurry are applied to porous refractory gypsum that draws in the water depositing a solid layer of alumina on its surface. The ceramic buildup is dried then sintered for 4 h at 1100 °C, the porous alumina coping is then carved into the desired shape before infiltrating with a slurry of lanthanum aluminosilicate glass by firing at 1120–1150 °C for 3–5 h. The resultant ceramic is a three-dimensional (3D) interpenetrating network of alumina and glass of high strength due to the presence of densely packed alumina and low porosity. The excess glass is removed and the core is subsequently veneered with a thermally compatible veneering ceramic. Improved translucency (esthetics) can be obtained by glass infiltrating a core composed of magnesium spinel and alumina. The strongest slip-cast material currently available contains tetragonal zirconia along with alumina and glass. When a load is induced on the tetragonal zirconia it absorbs energy by transforming into a monoclinic crystal form accompanied by a volume increase of 3% in a crack arresting manner. Flexural strengths (380–700 MPa) and fracture toughness (2–7 $MPa \cdot m^{1/2}$) for these core materials are in the following rank order: spinel < alumina < zirconia.

Prosthetic Resin Materials

Poly(methylmethacrylate) was introduced as a denture base material in 1937 and in roughly a decade had virtually replaced the use of vulcanite. Acrylic polymers also enjoy a wide variety of uses in additional prosthetic applications, such as, artificial denture teeth, provisional restorations and temporary crowns, denture base repair, relining and rebasing materials, and obturators for maxillofacial defects.

Denture base materials are typically fabricated from heat-cured poly(methylmethacrylate) and rubber-reinforced poly(methylmethacrylate) and perform surprisingly well. These plastics are supplied in a powder liquid or gel form. The 10 poly(methylmethacrylate) powder is modified with ethyl, butyl, or other alkyl methacrylates for impact resistance and contains benzoyl peroxide or diisobutylazonitrile to initiate polymerization when mixed with the liquid monomer. Pigments are added to obtain natural tissue appearance, for example, mercuric sulfide, cadmium sulfide, cadmium selenide, ferric oxide, or carbon black. Various glasses, ceramics and polymer fibers have been added as dispersed phases to various products in an attempt to reinforce the acrylic polymers. The liquid component is methyl methacrylate, modified with various other monomers while including an inhibitor such as hydroquinone to prevent premature polymerization for adequate shelf life. The liquid of cold-, self-, or autocuring resins contain tertiary amine or sulfinic acid chemical accelerators to allow the polymerization of the monomer at room temperature. Plasticizers for resilience and cross-linkers for hardness and decreased solubility may also be included. Denture base resins can also be fabricated through pressure, heat and light-activated techniques with compositional modifications for the various initiation reactions and physical handling properties during fabrication. A number of general requirements for denture base resins are outlined in ANSI/ADA Specification No. 12 (ISO 1567) providing guidance to dentists and dental manufacturers.

Denture teeth are also fabricated from acrylic and modified acrylic materials and are generally preferred over porcelain denture teeth due to wear characteristics, phonetics and technical considerations during fabrication and repair. Temporary or provisional restorations are also fabricated from acrylic-based resins, placed during an interim period in or over the coronal aspects of the tooth while a crown or bridge is fabricated in the dental laboratory. Due to ease of fabrication and tooth-like appearance these are much more popular than aluminum shells or polycarbonate crowns that typically must be relined before temporary cementation to the tooth.

Defects of the head and neck resulting from cancer surgery, accidents and congenital deformities have been corrected with a wide variety of maxillofacial resin materials, including poly(methylmethacrylate), plasticized polyvinylchloride, polyurethane, heat- vulcanized and room temperature-vulcanized (RTV) silicone and a whole host of various other elastomers. It is important to use prosthetic resin materials with color stability, ease of fabrication, dimensional stability, edge strength, flexibility, low thermal conductivity, biocompatibility, and surface texture to achieve clinical success and patient acceptance. Silicones are the most widely used materials for facial restorations in the United States, with RTV Silicone MDX-4-4210, possessing surface texture and hardness within the range of human skin. The prosthesis can be held in place by tissue undercuts, the patient's glasses or dentures, medical grade adhesives, magnetic attachment to endosseous implant-retained metallic attachments or bars or through a combination of methods. A mold is made from an impression of the defect upon which a prosthesis is fabricated and color matched by mixing small amounts of pigments into the elastomer. Surface coloration and texturing is completed and the patient returns periodically for esthetic touchups to achieve a lifelike match to the skin.

Implants

The surgical placement of endosseous dental implants to support dental restorations has become a routine aspect of

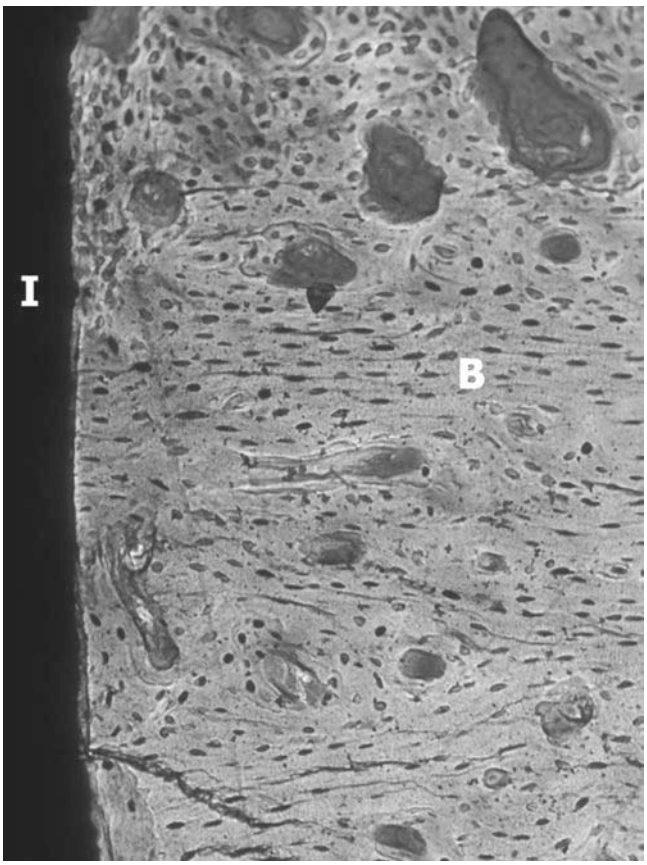


Figure 2. Dental implant demonstrating osseointegration. I = implant, B = bone. (Courtesy of John Keller, University of Iowa College of Dentistry.)

dental care enjoying a high success rate. Commercially pure titanium (CpTi) and Ti-6Al-4V are the materials most commonly used for endosseous dental implants. The stable oxides surfaces formed on CpTi and Ti- 6Al-4V have proven to successfully biointegrate with bone. The terms osseointegration and functional ankylosis are used to describe the direct bone apposition on the implant surface giving evidence to support a direct biochemical bonding (Fig. 2). The clinical success of these implants depends not only on osseointegration, but also the quantity, quality and distribution of bone at the implant site, the technical skill during surgical placement, and the timing and degree of mechanical loading under function. Many other factors play a role but six biological and technical factors are recognized as key to implant success: (*1*) implant surface texture, (*2*) biocompatibility, (*3*) implant design, (*4*) host tissue condition, (*5*) surgical technique, and (*6*) loading conditions.

Lower success is observed in areas of the mouth that may have less cancellous bone or thin cortical plates such as the posterior regions of the maxilla. Therefore, attempts are made to manipulate the osseous response to the implant so that the bone quantity and quality at the implant interface is optimized for the clinical requirements. This will not only help the routine implant but is especially important for those patients with: (*1*) poor bone quality, (*2*) heavy masticatory loading, and (*3*) the need for multiple tooth replacement. Surgical procedures are also utilized that enhance the osseous tissues at the intended implant site by auto- or allo-grafting.

Implant surface modifications have been heavily investigated with every major manufacturer offering various implant designs and surface textures. Two of the most thoroughly investigated and successful surfaces are machined titanium and titanium plasma-sprayed (TPS) surface, with the latter significantly increasing the surface area for bone contact. These rougher surfaces have been shown to require higher forces to be removed from the bone than do smoother surface implants and may allow: shorter healing periods, the use of less invasive shorter implants and may not require bicortical implant engagement. Improving the bone adaptation through microretentive mechanisms can be divided into those that attempt to enhance the immigration of new bone through surface topography, that is, osteoconduction, and those that attempt to manipulate the type of cell response and growth for new bone formation, that is, osteoinduction. Osteinductive methods also include the use of the implant as a delivery device for biomolecules for the induction of the desired response. A complex cascade of molecular and cellular processes occur after the placement of the implant into a surgical site, many of which are just now beginning to be understood, leading to the possibility of implant-mediated tissue engineering. Calcium phosphates or "hydroxyapatite" can also be coated on titanium implants and have been documented to create a very intimate bone-to-implant contact with a reduced healing period; however, the long-term results are less favorable than that achieved with TPS due to surface degradation and coating separation problems.

Macroretentive features are also part of the implant design including: screw threads, solid body press-fit designs, and sintered bead technology. These macroretentive features are intended to improve initial implant stability and enhanced bone ingrowth. Without the aid of a periodontal ligament (present between the natural tooth root and bone) the bone responds most favorable to compressive loading, which must accounted for in the implant design.

The original guidelines for implant success have changed over time with an increased use of nonsubmerged (not covered by the gum tissue) and single-staged surgical techniques (immediate abutment placement). These time- and cost-saving changes have come about after studies revealed similar end-results, in terms of "biological width" (composed of junctional epithelial and connective tissue attachment) and clinical survival rates. Considerations such as these are blurring the distinction between the clinical healing period (Phase I) and the functional period (Phase II). Additionally, studies have shown similar clinical predictability using both solely implant supported and mixed tooth-implant-supported fixed partial dentures (bridges), while the use of cemented rather than screw-retained prostheses have reduced technique complications.

Abutments of either titanium or alumina, as compared with those abutments of more "esthetic" quality, such as gold alloys or dental porcelains, are most likely to have favorable soft-tissue healing with formation of a

physiological epithelial and connective tissue attachment without subsequent bone resorption. Therefore, surgical techniques to help hide the prosthesis-abutment in the casual viewing region of a patient's mouth by careful peri-implant soft tissue manipulation with proven implant materials is currently recommended.

As work continues to optimize the osteoconductive (passive) response to implant surfaces, research will also progress toward predictable osteoinductive (active) responses. As hard and soft tissue responses are optimized through surgical protocols and biomaterial influences, healing phases will be shortened, retention rates will be increased, and loading capabilities will be improved, allowing the placement of fewer, less invasive and less expensive implants for predictable long-term, implant supported prostheses.

SUMMARY

This article briefly reviewed those commonly used metal, ceramic, polymer, and composite materials used in dentistry for the restoration of individual teeth or the replacement of missing teeth. Restorative materials include noble and base metals, resin based composites, glass ionomers, ceramics, acrylics, and amalgam alloys. These materials are either directly placed into the prepared tooth cavity or cemented in place after laboratory fabrication. An increasing number of "fillings" are retained by the use of dental adhesives. The dentist, in consult with the patient, must take several factors into consideration in the selection of restorative materials, to include (1) chewing forces, (2) esthetic demands' (3) strength of remaining tooth structure, (4) diet, (5) hygiene, and (6) cost. No one material type possesses all the desired physical properties; therefore, several materials are required for successful dental restoration.

Our population is living longer while retaining more of their teeth. With the increased emphasis on preventive dental care, increased awareness and the desire for health, our population will require more partial and single tooth restorations or replacements, especially in the area of root caries and less of a need for removable partial dentures, complete dentures and fixed bridges. Improvements in adhesive dental procedures and direct placed tooth colored resin-based composites will allow more conservative dental care. The interplay of biomaterials and biomolecules may also lead to the predictable regeneration of hard and soft tissues, while tissue engineering may someday lead to the induction of whole tooth regeneration.

BIBLIOGRAPHY

Cited References

1. Craig RG, Powers JM. Restorative Dental Materials. 11th ed. St. Louis: Mosby; 2002.

Reading List

Anusavice KJ. Phillip's Science of Dental Materials. 10th ed. Philadelphia: W. B. Saunders; 1996.

Denry IL. Recent advances in ceramics for dentistry. Crit Rev Oral Biol Med 1996;7(2):134–143.

Ferracane JL. New polymer resins for dental restoratives. Oper Dent 2001; Suppl 6: 199–209.

Keller JC. Dental Implants: The relationship of materials characterization to biologic properties. In: Bronzino JD, editor. The Biomedical Engineering Handbook. 2nd ed. Boca Raton, FL: CRC Press LLC; 2000.

Osborne JW. Mercury, its impact on the environment and its biocompatibility. Oper Dent 2001; Suppl 6:87–104.

Salvi GE, Lang NP. Changing paradigms in implant dentistry. Crit Rev Oral Biol Med 2001;12(3):262–272.

Smith DC. A new dental cement. Br Dent J 1968;125:381–384.

Stanford CM. Surface modifications of implants. Oral Maxillofacial Surg Clin N Am 2002;14:39–51.

See also BIOCOMPATIBILITY OF MATERIALS; BONE AND TEETH, PROPERTIES OF; RESIN-BASED COMPOSITES; TOOTH AND JAW, BIOMECHANICS OF.

BIOMATERIALS, POLYMERS

MIN ZHANG
University of Washington
Seattle, Washington

SUSAN P. JAMES
Colorado State University
Fort Collins, Colorado

INTRODUCTION

Biomaterials are materials of synthetic as well as of natural origin in contact with tissue or biological fluids, including metals, ceramics, polymers, and composites. The main advantages of polymeric biomaterials over ceramics and metals are the variety of composition, properties and available forms (solid, hydrogel, and solution), and ease of fabrication into complex shapes (films, sheets, fibers, powders, etc.) and structures because synthetic polymers are easily tailored to specific applications. In addition, polymeric materials are much lighter than metals and ceramics. Since most natural biomaterials are polymeric, mimicking the function and/or structure of natural materials (e.g., skin) is more easily achieved with polymers or polymeric composites than metals and ceramics. As with other biomaterials, there are some basic requirements for polymeric biomaterials. They must be (*1*) nontoxic, for example, not causing carcinogenesis, pyrogenicity, hemolysis, sustained inflammation, and allergy; (*2*) biocompatible, that is, not causing foreign body reactions, such as complement activation, thrombus formation, collagenous tissue encapsulation, calcification, and compatibility with the contact tissue in physical and mechanical properties; (*3*) sterilizable with autoclave, dry heating, ethylene oxide, gas plasma or γ irradiation, or be produced in a sterile fashion so no postmanufacture sterilization is required (1a).

Synthetic polymers have been widely used in biomedical devices, for example, hard and soft tissue implants, extracorporeal devices, drug delivery systems, and medical disposable supplies. They exhibit diverse properties, ranging from hydrophobic, non-water-absorbing materials (e.g., polyethylene, polypropylene, and polytetrafluoroethylene), to hydrophilic, water-swelling hydrogels [e.g., poly(hydroxyethyl methacrylate)], and to water-soluble materials [e.g.,

poly(vinyl alcohol)] (2a). Biological polymers are obtained from animals, plants, bacteria, or other living creatures. Their remarkable advantages over synthetic polymers include their excellent physiological activities. These activities include cellular activity regulation (e.g., hyaluronic acid), selective cell adhesion (e.g., collagen), and similar properties to natural tissues (3). Most biopolymers are biodegradable, so they are suitable for use in temporary medical devices, drug-delivery systems, and tissue engineering scaffolds.

This section introduces the synthesis methods of generic polymers, and the effect of composition and structure on their properties. Following this, 13 groups of polymers that have found wide biomedical application are reviewed and their properties and uses are discussed.

POLYMERIZATION–SYNTHESIS

Polymers are long molecules made up of a large number of simple repeating units. They are prepared from monomers through a process called polymerization. Small monomer molecules react chemically to form either linear chains or three-dimensional (3D) networks. Conventionally polymerization mechanisms are divided into two main categories: condensation polymerization (also called step-growth polymerization) and addition polymerization (also called chain polymerization).

Condensation or Step-Growth Polymerization

Condensation polymerization occurs between an organic base (e.g., an alcohol and amine) and an organic acid (e.g., carboxylic acid and acid chloride), and a small molecule (e.g., water) is condensed out during the reaction. In a condensation reaction to combine two monomers together to form a dimer, each monomer molecule loses an atom or a group of atoms at the reactive end, leading to the formation of a covalent bond between the two molecules, while the eliminated atoms bond with others to form small molecules. The dimers can react with each other or with unreacted monomers until finally a long molecule is generated. An equilibrium exists between the reactants and products during condensation polymerization. The condensate (e.g., water) should be removed to drive the reaction toward the product direction. A high molecular weight product can be obtained only after a sufficiently long reaction time.

The reaction between a diacid and a dialcohol to produce an ester is a typical condensation polymerization example.

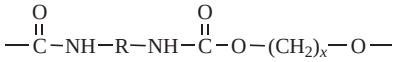
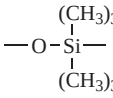
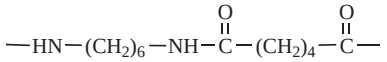
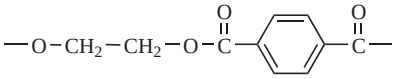
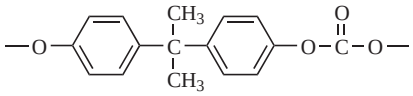
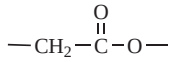
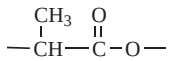
$$n\,\mathrm{HOOC{-}R_1{-}COOH} + n\,\mathrm{HO{-}R_2{-}OH} \rightarrow$$

$$\mathrm{HO}\left[\mathrm{{-}OCR_1COOR_2{-}O}\right]_n\mathrm{H} + (2n-1)\mathrm{H_2O}$$

Some condensation polymers used as biomaterials are given in Table 1. They are typically synthesized from reactions of acids and alcohols to produce polyesters, reactions of acids with amines to produce polyamides, or reactions of alcohols or amines with isocyanates to produce polyurethanes or polyurea, respectively.

Synthesis of biopolymers is very complicated, but typically involves enzyme-catalyzed condensation polymerization occurring in animal or plant cells, or in microorganisms via their metabolic pathways (4,5a).

Table 1. Typical Condensation Polymers

Polymer	Repeat Unit
Polyurethane	
Silicone rubber	
Polyamide (Nylon 66)	
Poly(ethylene terephthalate)	
Polycarbonate	
Polyacetal	
Polyglycolic acid (PGA)	
Polylactic acid (PLA)	

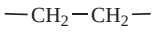
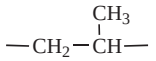
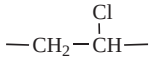
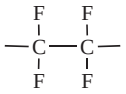
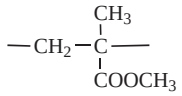
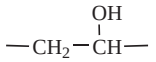
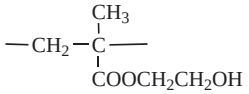
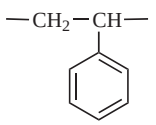
Addition or Chain Polymerization

Addition polymerization occurs among small molecules with double bonds. Polymer chains are formed by opening-up double bonds of unsaturated monomer units and successive addition to a growing chain with an active center. No small molecule byproducts are formed during addition polymerization. Consequently, the composition of the repeating unit of the polymer is identical to that of its monomer. Addition polymerization takes place in three distinct steps: initiation, propagation, and termination. Initiation occurs by an attack on the monomer molecule by a free radical, a cation, an anion, or Ziegler–Natta catalysts; accordingly, addition polymerization can be divided into four types: free-radical polymerization, cationic polymerization, anionic polymerization, and coordination polymerization. No matter how the reaction is initiated, once a reactive center is created, many monomers are added onto it and the molecule chain grows very large within a few seconds or less, so the addition polymer size (i.e., molecular weight) is independent of reaction time. Unlike condensation polymerization, no dimer, trimer, or other intermediates can be found in addition polymerization. Polymerization of ethylene is a typical example of addition polymerization.

$$n\,\mathrm{H_2CCH_2} \rightarrow \left(\mathrm{H_2C{-}CH_2}\right)_n$$

Typical addition polymers are listed in Table 2.

Table 2. Typical Addition Polymers

Polymer	Repeat Unit
Polyethylene	
Polypropylene	
Polyvinyl chloride	
Poly(terafluoroethylene) (Teflon)	
Poly(methyl methacrylate)	
Poly(vinyl alcohol)	
Poly(hydroxyethyl methacrylate)	
Polystyrene	

The stereoregularity and branching of addition polymers can be controlled through varying the type of initiator and the reaction conditions (6). Inonic addition polymerization can lead to some control of tacticity and a stereoregular structure. Polymers produced through coordination polymerization have a high degree of stereoregularity. The polyethylene produced with peroxide initiator is highly branched. By using a Ziegler–Natta catalyst, linear, high density polyethylene or ultrahigh molecular weight polyethylene (UHMWPE) can be obtained. Living free-radical polymerization is a newer form of free-radical polymerization. By using rapid initiation, slow propagation, and inhibition of termination and transfer reactions, the molecular structure of polymers can be precisely controlled. This method can be applied to vinyl monomers to produce block, graft, star polymers, polymer brushes, and many other architectures (5a,7).

Molecular Weight and Its Distribution

Almost all polymers consist of molecules (a.k.a., chains) with a variety of lengths, so it is only possible to quote an average value of molecular weight. The length of a polymer molecule is represented by the degree of polymerization (DP), which is equal to the number of repeat units in the chain. The relationship between polymer molecular weight (MW) and degree of polymerization can be expressed as

$$\text{MW of polmer} = \text{DP} \times \text{MW of repeating units}$$

The number-average molecular weight (M_n) and weight-average molecular weight (M_w) are the most commonly used average values of molecular weight. The number average molecular weight is defined as the sum of the products of the molecular weight of each fraction (M_i) multiplied by its mole fraction (x_i) (Eq. 1), which can be obtained using gel filtration chromatography, light scattering, or ultracentrifugation. Whereas M_w is the sum of the products of the MW of each fraction (M_i) multiplied by its weight fraction (w_i) (Eq. 2), which can be measured with osmometry.

$$M_n = \sum x_i M_i \tag{1}$$

$$M_w = \sum w_i M_i \tag{2}$$

The ratio of M_w/M_n is defined as the polydispersity index (PDI), representing the breadth of the molecular weight distribution. When all the polymer chains have the same length, the ratio is 1. A low polydispersity index is necessary to control physical and mechanical properties of polymers because the short chains usually present when PDI is high degrade properties.

COMPOSITION, STRUCTURE, AND PROPERTIES

The structure and behavior of polymers is strongly temperature dependant. There are two major transition temperatures for polymers: T_g and T_m. The glass-transition temperature is a second-order transition temperature, associated with the amorphous regions of polymers. It marks the onset of significant molecular motion. Above T_g, polymers soften, and become rubber-like and more easily deformed. Below T_g, polymers become hard and brittle, and glass-like. Applications of polymeric biomaterials are related with their T_g values. For example, silicone rubber with a T_g of $-127\,^{\circ}C$ is soft and acts as an elastomer at $37\,^{\circ}C$ (human body temperature), while poly(methyl methacrylate) (PMM) used as bone cement with a T_g of $105\,^{\circ}C$ retains high strength, stiffness, and creep resistance at $37\,^{\circ}C$. The melting temperature is associated with crystalline regions of polymers. It is a first-order transition temperature. Above T_m, the polymer is in a melt liquid state. Polymeric materials with a low T_m value can be melt processed.

Polymer molecules can be linear, branched, or a cross-linked network. Schematic representations are given in Fig. 1. Based on their molecular structure and their mechanical and thermal behavior, polymers are classified into three major categories: thermoplastics, thermosets and elastomers. They are discussed in detail respectively.

Thermoplastics

These polymers soften and harden reversibly with changes in temperature. Both linear and branched polymers are thermoplastic. The properties of thermoplastics can be changed by controlling the following factors.

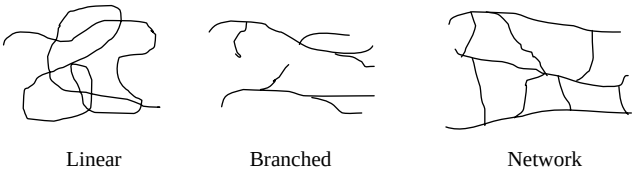
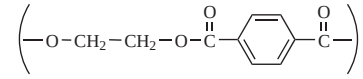


Figure 1. Schematic representation of different types of polymer molecules.

Molecular Weight. The molecular weight and its distribution have a great effect on the properties of thermoplastics. By increasing molecular weight, the polymer chains become longer and more entangled, resulting in a higher melting temperature and improved strength, including resistance to creep (8a). The strength properties increase with the molecular weight rapidly at first, but level off after reaching a certain point. Uniform molecular length is also important, because short molecules act as plasticizers, which decrease the mechanical properties of polymers. For example, ultrahigh molecular weight polyethylene must have a high molecular weight (MW, $2–4 \times 10^6$ g·mol^{-1}) and narrow MW distribution (i.e., low PDI) to obtain excellent mechanical properties for orthopedic applications.

Chemical Composition. The changes in the composition of the backbone or side chains also affect the properties of polymers. When atoms that can increase the flexibility of polymer chains (e.g., O and S) are incorporated into the carbon backbone, for example, polyethylene oxide (CH_2CH_2O) (T_g: $-41\,^\circ$C; T_m: $69\,^\circ$C), the glass transition and melting temperatures will decrease. On the other hand, the insertion of groups that stiffen the polymer chains markedly raises the T_g and T_m and leads to higher strength and stiffness, as seen in the case of polyethylene terephthalate  (T_g: 60/85 °C; T_m: 280 °C).

The replacement of pendant hydrogen atoms in polyethylene by other atoms also changes the polymer properties. Large atoms or groups (e.g., methyl groups in polypropylene), hinder the rotation about the backbone, resulting in higher T_g and T_m, strength, and stiffness. Similar results are observed for the polymers with polar pendant atoms or groups [e.g. poly(vinyl chloride), PVC]. van der Waals forces are enforced and even hydrogen bonds may be formed among the chains of these polymers.

Branching. Branching prevents dense packing and crystallization of the polymer chain, and thus reduces the density, melting temperature, strength, and stiffness of polymers. For example, branched low density polyethylene (LDPE) is much weaker than linear high density polyethylene (HDPE).

Tacticity. When the repeat units of a polymer are nonsymmetrical, the location of the side atoms or groups also plays an important role in the structure and properties of the polymer. If all the side atoms or groups are on one side of the main chain, the polymer is termed isotactic. In a syndiotactic polymer, the side groups alternatively appear on both sides of the chain. In both cases, the polymer is stereoregular or stereotactic. Polymer chains with stereoregularity are better able to crystallize, resulting in high T_m, stiffness and less solubility. In an atactic polymer, side groups are randomly distributed along the molecular chain. Atactic polymers give poor packing, low density, low stiffness, and strength. A typical example of the importance of tacticity is polypropylene. Isotactic and syndiotactic polypropylene have a high T_m (176 °C for isotactic, 150 °C for syndiotactic) and good mechanical properties. They are widely used as biomaterials, whereas atactic propylene is an amorphous waxlike polymer without any application in biomedical field (8b,9).

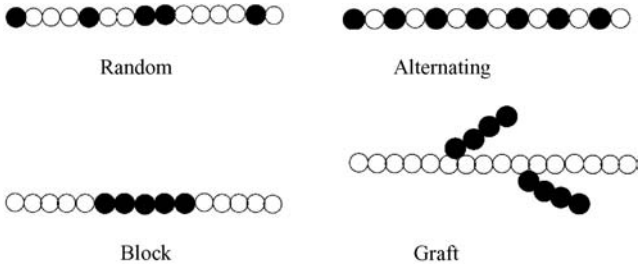


Figure 2. Four types of copolymers.

Copolymer. Copolymers contain two or more different types of repeat units. According to the distribution of the repeat units, they are divided into four types (Fig. 2). Copolymers are synthesized to obtain the desirable combination of properties of simple homopolymers.

Temperature and Time. Amorphous thermoplastics exhibit viscoelastic behavior, meaning that their properties are time or temperature dependent. At low temperatures or high rates of loading, the polymers behave in a brittle manner. However, at high temperature or low rates, the materials behave as a viscous liquid with chains easily passing one another. The application temperature of most biomaterials is 37 °C, which cannot be changed, but the properties of the materials can be controlled by selecting appropriate T_g.

Thermosets

Thermosets are highly cross-linked 3D molecular networks. High density cross-links between molecules restrict the motion of the chains of thermosets, leading to a high T_g, good strength, stiffness, and hardness, but poor ductility. The hard and stiff thermosetting polymers find uses in hard tissues (e.g., bone and teeth). Poly(carboxylic acid) cross-linked with zinc is a hard and rigid cement used for dental restorations (2a,10a). Epoxy resin is sometime used to fill the cavities of the teeth and provide hardness.

Elastomers

Elastomers are lightly cross-linked macromolecular networks. The cross-linking can be covalent or physical. In

thermoplastic elastomers, the hard and tightly packed domains with high T_g (e.g., the hard isocyanate segments in polyurethane elastomer) act as physical cross-links. The loose cross-links prevent viscous plastic deformation while retaining large elastic deformation, so elastomers can be easily stretched to high extension and will go back to their original position on removal of the stress. To act as a biomedical elastomer, a T_g much lower than 37 °C is required, and the polymer should not easily crystallize.

POLYMERS USED AS BIOMATERIALS

Many polymers have been synthesized for biomedical applications. This section just focuses on 13 groups of polymers most commonly used in clinical practice. Each part of the following deals with one polymer or one group of polymers with similar structures. The synthesis, structure, properties, and applications of these polymers will be discussed. Their trade names and related ASTM standards are listed in Tables 3 and 4, respectively, while the thermal and mechanical properties are summarized in Table 5.

Polyolefins

Polyolefins are a group of thermoplastics polymers derived from simple olefins. The most important polyolefins are polyethylene, polypropylene, and their copolymers.

Polyethylene. Commercially available polyethylene has four major grades: LDPE, linear low density (LLDPE), high density (HDPE), and UHMWPE. These materials have good toughness and excellent chemical resistance, and can be easily processed into products at low cost.

Low density polyethylene is produced through the free-radical polymerization of ethylene gas at high pressure (100–300 MPa) in the presence of peroxide initiator (5c). The synthesis conditions lead to the highly branched structure, low density (0.915–0.935 $g{\cdot}cm^{-3}$) and crystallinity of LDPE (11a). By using Ziegler–Natta catalyst, HDPE can be synthesized at a low temperature (60–80 °C) and pressure (~10 MPa). Unlike LDPE, HDPE is linear. The linearity leads to good packing of the molecular chains, high crystallinity, and density (0.94–0.965 $g{\cdot}cm^{-3}$) of HDPE. The LLDPE is produced by a low pressure process in the presence of metal catalysts. Up to 10% of a 1-alkene (e.g., butene-1, hexene-1, and octane) is used as the comonomer. Unlike LDPE, the side chains of LLDPE are very short, resulting in better properties than LDPE (12). All three types of polyethylene can be melt-processed through extrusion or molding. The LDPE cannot withstand sterilization temperatures, so only HDPE and LLDPE are used for biomedical applications (2a). The HDPE is used in tubing for catheters and drains, and in pharmaceutical bottles and nonwoven fabrics. The LLDPE is frequently used for pouches and bags due to its excellent puncture resistance. Biocompatibility tests for PE used as human tissue contact devices, short-term implantation of 30 days or less and fluid transfer devices are given in ASTM F 639 (not applicable for UHMWPE).

When molecular weights of the linear polyethylene obtained through Ziegler–Natta catalyst are $>1\times10^6 g{\cdot}mol^{-1}$, there is a sudden jump in the properties (13). The melt viscosity becomes extremely high so that the polyethylene cannot be processed with conventional extrusion and injection molding. Also it is practically insoluble in all solvents, so only sintering at high temperature and pressure may be used to fabricate the desired products. The polyethylene has excellent mechanical properties and a very low coefficient of friction and wear. It is termed UHMWPE. The UHMWPE, currently

Table 3. Structures and Trade Names of Polymeric Biomaterials

Polymeric Biomaterials	Structure	Trade Names
Cross-Linked UHMWPE	UHMWPE Cross-Linked through Radiation or Chemical Reactions	Crossfire (Stryker Howmedica Osteonics), Marathon (DePuy Orthopaedics), and Durasul (Sulzer Medica)
Polypropylene	Linear macromolecules	Prolene (Ethicon), Surgipro (Syneture)
PTFE	Linear macromolecules, highly crystalline	Teflon (DuPont)
Expanded PTFE	PTFE with microporous structure	GoreTex (Gore)
PMMA	Atactic, linear macromolecules	Plexiglas (Rohm & Haas), Lucite (DuPont)
Polyurethane	Thermoplastic segmented polyurethane elastomers	Biomer (Ethicon), Pellethane (Dow Chemical), Tecoflex (Thermedics)
Polyamide	Linear macromolecules, strong intermolecular hydrogen bonding	Nylon (DuPont)
PET	Linear, stiff macromolecules	Dacron (DuPont)
Polyacetal	Closely packed, linear molecules	Delrin (DuPont)
Poly(hydroxyethyl methacrylate)	Hydrogel	Hydron (Hydron Technologies)
PGA	Biodegradable polymer	Dexon (American Cyanamid)
PLGA	Biodegradable polymer	Vicryl (Ethicon)
Poly(ethylene oxide/propylene oxide) copolymers	PEO-PPO-PEO triblock copolymer	Pluronic F127 (BASF)
Hyaluronan	Crosslinked hyaluronan Hyaluronan with carboxyl groups esterified by alcohol	Hylan (Biomatrix)Hyaff (Fidia Advanced Biopolymer)

Table 4. Polymeric Biomaterials and Related ASTM Standards

Polymer	ASTM Standards	Scope
Polyethylene	F 639	Specifies requirements and physical/biological test methods for PE plastics used in medical devices (not applicable to UHMWPE).
UHMWPE	F 648	Specifies property requirements for UHMWPE powder and fabricated forms used for surgical implants, such as joint implants.
PVC	F 665	Classifies formulations of PVC plastics used for short-term biomedical application.
PTFE	F 754	Specifies the performance of PTFE in sheet, tube and rod shapes used for surgical implants.
PMMA bone cement	F 451	Specifies composition, physical performance, packaging requirements, and biocompatibility of acrylic bone cement.
	F 2118	Provides test methods to evaluate the fatigue properties of acrylic bone cement.
Polyurethane	F 624	Provides guide to evaluate thermoplastic polyurethane in solid and solution forms for biomedical applications.
Silicone rubbers	F 2038	Provides information about formulation and use of silicone elastomers, gels, and foams used in medical applications.
	F 2042	Provides information about fabrication and processing of silicone elastomers, gels and foams used in medical applications.
Polycarbonates	F 997	Specifies requirements and test methods for polycarbonates used for medical devices.
Polyacetal	F 1855	Specifies requirements and test methods for polyacetal used for medical devices.
L-PLA	F 1925	Specifies requirements for virgin poly(L-lactic-acid) resin used for surgical implants.
	F 1635	Defines testing methods to assess biodegradation rates and changes in material and properties of poly(L-lactic-acid) resin and devices.

used for fabrications of acetabular cups in hip replacements (Fig. 3), tibial plateau and patellar surfaces in knee replacements, sliding core in spinal disk replacements, and glenoid components in shoulder replacements, has a MW $\sim 2\text{–}4 \times 10^6$ g·mol^{-1}. The property requirements of UHMWPE for surgical implants are defined by ASTM F648. Before 1995, γ radiation in air was a standard method to sterilize UHMWPE orthopedic implants. However, free radicals within UHMWPE from gamma radiation caused oxidation and property degradation on shelf aging and *in vivo*. By 1998, all of the major orthopedic manufactures in the United States changed to use gamma radiation in an inert or a reduced oxygen environment, ethylene oxide, or gas plasma to sterilize UHMWPE (14). Despite the recognized success of UHMWPE as loading bearing surfaces in joint arthroplasties, UHMWPE wear debris and associated osteolysis and loosening of implants remains a major obstacle limiting the longevity of current joint replacements. To further improve the wear resistance, highly cross-linked UHMWPE is developed by

Table 5. Properties of Polymeric Biomaterials[a]

Polymer	T_g, °C	T_m, °C	Tensile Strength, MPa	Tensile Modulus, GPa	Elongation, %
Polyethylene					
LDPE	−120	115	7.6	0.096–0.26	150
HDPE	−120	137	23–40	0.41–1.24	400–500
UHMWPE	−120	130–145	30	1.1–2.0	300
Polypropylene	−20	175	28–36	1.1–1.55	400–900
Polyvinyl chloride	80	180	40–50	2.4–4.1	2–80
Teflon	117	327	15–35	0.40	2–5
Poly(methyl methacrylate)	105		50–75	2.0–6.0	2–10
Polyurethane (elastomer)			23–58		400–600
Silicone rubbers	−123				
Soft			6.0		600
Hard			7.0		350
Polyamide					
Nylon 6	50/100	270	70	0.7	300
Nylon 66	50	280	75		300
Poly(ethylene terephthalate)	60–85	280	50–70	3.0–4.0	30–300
Polycarbonate	150	230	65	2.4	110
Polyacetal	−85	181	65–80	5.0–13.0	9.5–12
Poly(lactic acid)					
L-PLA	54–59	159–178	28–50	1.2–3.0	2.0–6.0
DL-PLA	51		29	1.9	5.0
Poly(glycolic acid)	35	210			
Collagen fibers			50–1000	1.0	10
Cellulose acetate	230		13–60	0.45–2.8	1.9–9.0

[a]Refs. 5b 9, 10c, and 17e.

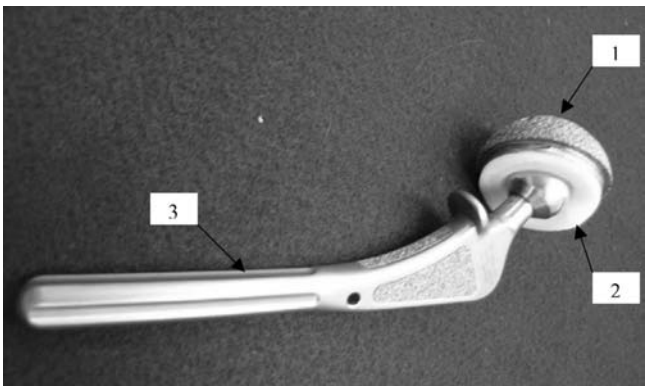


Figure 3. A total joint replacement: (1) metal backing of acetabular cup, (2) UHMWPE acetabular cup, (3) metallic femoral prosthesis.

cross-linking conventional UHMWPE through chemical reactions, or through gamma or electron beam radiation. Cross-linked UHMWPEs available on the market include Crossfire (Stryker Howmedica Osteonics), Marathon (DePuy Orthopaedics), and Durasul (Sulzer Medica). Both research and clinical applications have demonstrated that cross-linking dramatically reduces the wear rate of UHMWPE (14).

Polypropylene. Like linear polyethylene, syndiotactic and isotactic polypropylene are also polymerized through Ziegler–Natta catalyst. Although polypropylene is similar to polyethylene in structure, polypropylene has a lower density ~ 0.90 g·cm^{-3} and a higher T_g ($-12\,^\circ$C) and T_m (125–167 °C). The higher melting temperature makes polypropylene suitable for autoclave sterilization (15). The chemical resistance of polypropylene is similar to high density polyethylene, while its stress-cracking resistance and creep resistance is superior to that of polyethylene. It has an exceptionally long flex life, and thus is used to make integrally molded hinges in finger joint prostheses (11a). Also as a suture material, polypropylene yarn (e.g., Prolene from Ethicon, Surgipro from Syneture) has been used clinically (2a,10b). It causes least fibroblastic response compared with other nondegradeable suture materials and does not lose strength after it is implanted.

Poly(vinyl chloride)

Poly(vinyl chloride) is a linear, atactic polymer synthesized through free-radical polymerization. Due to the large volume and high polarity of the chlorine atoms, it is difficult for the molecular chains to rotate and disentangle and hydrogen bonds are formed between adjacent chains, resulting in high strength and stiffness, and T_g (80 °C) and T_m (180 °C) (10c). Pure PVC is hard and brittle, but with the addition of plasticizers, it becomes soft and flexible. In the medical formulations of PVC, di-2-ethylhexylphthalate (DEHP or DOP) is used as a plasticizer. Plasticized PVC is used in temporary blood storage bags, catheters, cannulae, and dialysis devices. The PVC may pose problems for long-term applications because of possible extraction of the plasticizer by body fluid. Standard classification for vinyl chloride plastics used in biomedical applications is provided by ASTM F665.

Poly(tetrafluoroethylene)

Poly(tetrafluoroethylene) (PTFE), commonly known as Teflon (DuPont), is made from tetrafluoroethylene through free-radical polymerization in the presence of excess of water for removal of heat. The polymer is highly crystalline ($>94\%$ crystallinity, T_m 327 °C), dense (2.2 g·cm^{-2}) and insoluble in all common solvents. It is very stable both thermally and chemically, and as a result it is very difficult to process. The PTFE can only be sintered into products at a temperature $>327\,^\circ$C under pressure.

The PTFE has excellent lubricity, its coefficient of friction is very low (0.1), but it is not wear resistant, its modulus of elasticity and tensile strength are very low, and even more importantly it can not maintain shape very well due to the cold-flow (5c). The use of Teflon as the acetabular component material by Charnley (2b) in his total hip replacement design 40 years ago caused a catastrophe. All Teflon cups failed *in vivo*, requiring revision surgery.

Although not suitable for load bearing surfaces, PTFE can be used for other biomedical applications because of its excellent biocompatibility and stability. Standard specifications for the implantable PTFE are given in ASTM F754. Expanded PTFE (ePTFE) vascular grafts, made by stretching paste-extruded PTFE tubes at a temperature $<T_m$ and then sintering, are soft microporous tubes (GoreTex). They show good clinical results as medium diameter (5–11 mm) vessel grafts, (e.g., femoral and popliteal artery replacements) (10d,16a). However, intimal hyperplasia of smooth muscle cells at the anastomosis frequently leads to their failure. The ePTFE grafts are also popular in hemodialysis as an interposition between radial artery and cubital vein. However, thrombosis occurring at the graft venous ends may be a concern (16b). The PTFE fabrics find applications in heart valve prosthesis as suture ring (10b). Sheets or films of PTFE or its composite with graphite are widely used by plastic surgeons in reconstruction of the maxillofacial areas (10b). The PTFE tubes are used for middle ear drain, while PTFE shunts are used to carry cerebral spinal fluid from brain to venous in the treatment of hydrocephalus.

Poly(methyl methacrylate)

The commercially important poly(methyl methacrylate) (PMMA) is atactic, which is produced by free-radical polymerization of methyl methacrylate monomer (liquid) using initiator, or thermal, or photochemical initiation. Because of the bulky side chains, atactic PMMA is completely amorphous (T_g: 105 °C) and has an excellent light transparency (92% transmission) and a high index of refraction (1.49). The transparent material is familiar as Plexiglas or Lucite. For the same side group argument presented above, the strength and stiffness of PMMA are also relatively high. As a hard thermoplastic polymer, PMMA can be easily formed into any desirable shapes by regular cast, molding, or machining.

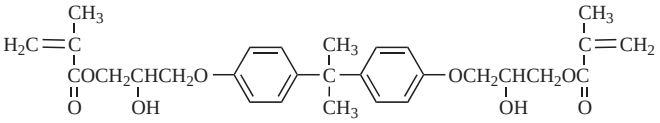


Figure 4. Chemical formula of bisphenol-A-glycidyl dimethacrylate (a new bone cement).

Poly(methyl methacrylate) is highly biocompatible and chemical resistant. It has been used in a variety of medical applications including hard contact lenses and intraocular lenses, membranes for blood dialysis, cranioplasty, bone cement for joint prostheses fixation, dentures, and maxillofacial implants.

The PMMA resin used for bone cement and dentures are usually formulated from two components: prepolymerized PMMA solid particles and the methyl methacrylate monomer liquid. When the two components are mixed during a clinical procedure, an easily moldable dough is obtained that cures in ~ 10 min. The liquid monomer polymerizes by free-radical reaction and binds the solid particles together. The composition of a commercial bone cement product (Surgical Simplex) is given in Table 6. Dibenzoyl peroxide is included in the solid component to initiate polymerization of the methyl methacrylate monomer. *N*, *N*-Dimethyl-*p*-toluidine is an activator to promote self-curing of the monomer at room temperature. Hydroquinone is added as an inhibitor to prevent premature polymerization of the monomer during storage. Barium sulfate ($BaSO_4$) is a radioopacifier. The methyl methacrylate–styrene copolymer is added to adjust the mixing and handling characteristics (e.g., viscosity, exotherm) of the cement. The physical and mechanical properties of the cement can be controlled through changing the composition and relative proportions of the components, solid particle size and its molecular weight. The requirements for PMMA bone cement are given in ASTM F451. The test methods for its fatigue performance are provided in ASTM F2118.

The PMMA bone cement is inert (no bioactivity), and the fatigue failure occurring at the cement-prosthesis and the cement-bone is a main cause of implant loosening (16c). New bioactive bone cements (BABCs) have been developed to improve the bonding strength and biochemical properties of PMMA bone cement (17a). These BABCs consist of bioactive glass ceramic powder (e.g., $CaOMgOSiO_2P_2O_5CaF_2$) and a bisphenol-A-glycidyl dimethacrylate-based resin shown in Fig. 4.

Table 6. Composition of PMMA Bone Cement (Surgical Simplex)

Components	Composition	Amount, %
Powder (40 g in a packet)	Polymethyl methacrylate	15.0 (w/o)[a]
	Methyl methacrylate –styrene copolymer	75.0 (w/o)
	$BaSO_4$	10.0 (w/o)
	Dibenzoyl peroxide	Trace
Liquid (20 mL in an ampoule)	Methyl methacrylate	97.4 (v/o)[b]
	N, *N*-Dimethyl-*p*-toluidine	2.6 (v/o)
	Hydroquinone	Trace

[a]w/o: weight percentage.
[b]v/o: volume percentage.

Polyurethanes

Polyurethanes are a family of heterogeneous polymers containing the urethane linkage (NHCOO) and frequently the urethane groups are not the predominant functional groups (5d,18). The most common method to synthesize polyurethanes consists of two steps (Fig. 5). The first step involves formation of an isocyanate terminated prepolymer from polyester or polyether polyols and di- or higher isocyanate. Subsequent reaction of the prepolymer with a chain extender, usually a diol or diamine, produces a multiblock copolymer. These two reactions are a step-growth polymerization, but no condensed byproduct is eliminated, so this type of polymerization is often referred to as a polyaddition or rearrangement polymerization. Due to the multiple choices in the chemistry and molecular weight of the various components, polyurethanes exhibit a broad range of physical properties: from hard and brittle thermoset polymers, through thermoplastic elastomers, to viscous materials. Thermoplastic segmented polyurethanes usually are valuable in producing medical devices (e.g., extruded blood tubing), while the cross-linked ones have received more attention for long-term devices and implants. The ASTM F624 provides test methods to evaluate properties of thermoplastic polyurethanes.

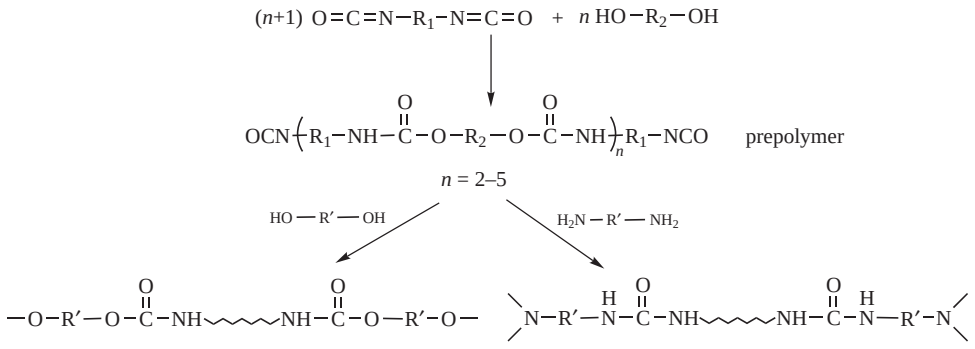


Figure 5. Two-step synthesis of polyurethane.

The isocyanates used for linear polyurethane elastomers are aromatic diisocyanates [e.g., toluene diisocyanate (TDI) and 4,4′-diphenylmethane diisocyanate (MDI)], which comprise the hard segments of the thermoplastic elastomers with the chain extenders. The soft segments of the elastomers are blocks of polyether polyols (usually polyethylene glycol PEG) or polyester polyols. These two types of segments tend to aggregate into different domains, resulting in microphase separation with the hard blocks acting as physical crosslinks in the thermoplastic elastomer.

Polyurethane elastomers are good materials for use in medical devices due to their good mechanical properties and blood compatibility. The very high flexural endurance of polyurethanes makes them major candidates for cardiovascular implants. Biomer (Ethicon, NJ), Pellethane (Dow Chemical, TX), and Tecoflex (Thermedics, MA) are all polyurethanes under different trade names. They have been widely used for cardiac guiding catheters, pacemaker lead insulation, vascular prostheses, artificial heart assist devices, blood tubing, and hollow fiber dialysers. Polyurethanes are also used extensively for wound healing. Bioclusive (Johnson and Johnson Medical, Inc.) and Opsite (Smith and Nephew) are two types of nonabsorbent wound dressings made from polyurethane films.

Although these ether-based polyurethanes are stable *in vitro*, environmental stress cracking after implanted makes their long-term biostability questionable. The microcracks caused by biological peroxidation of the ether linkage not only weaken the materials, but also serve as nucleation sites for thrombus formation (19a). To solve the problem, polycarbonate-based polyurethanes have been developed and investigated to provide an unsurpassed combination of biostability, strength, flexibility, and ease of manufacture.

Silicone Rubbers

The basic repeat unit of silicone rubbers is dimethyl siloxane. They are made by vulcanization of silicone prepolymers or by ring-opening polymerization of octamethylcyclotetrasiloxane. Silicone prepolymers are obtained by the hydrolysis of dimethyldichlorosilane with water. The hydrolysis product is dimethylsilanol, which is unstable and condenses to low molecular weight silicone prepolymers in the presence of hydrochloric acid (5d,12,17b) (Fig. 6). Silicone prepolymers are also useful as silicone oil. The vulcanization of the prepolymers can be performed at room or at high temperature. Room temperature vulcanization silicone rubbers are available in two formats: one-component or two-component. The one-component silicone rubbers result from a reaction between atmospheric moisture and a mixture of silicone prepolymers and catalyst, whereas the two-component systems result from a reaction between prepolymers and a cross-linker added to initiate the reaction. In heat vulcanization, peroxides are used to produce free radicals on heating that react with the side groups of prepolymers to form cross-links. Several copolymer silicone rubbers have been developed to meet diverse biomedical applications. Copolymers made from dimethyl siloxane and a small amount of methylvinyl siloxane result in medium and hard grades of silicone rubbers. Soft grades of silicone rubber are copolymers with phenyl-methyl-siloxane. Requirements for medical grade silicone gels and elastomers are provided in ASTM F2038 and F2042.

Silicone rubbers have many excellent properties, (e.g., extreme inertness, nonadhesion, high oxygen permeability, thermal and oxidative stability, and high flexibility at low temperature). A major disadvantage of silicone rubbers is poor resistance to tearing. Silicone rubbers are one of the widely used polymeric biomaterials in modern medicine. Since Alfred Swanson introduced silicone rubber for small flexible joints in 1960s, $>600{,}000$ silicone rubber hinge or end-bearing prostheses have been implanted to treat arthritic conditions of finger and wrist joints. These silicone implants are successful in relieving pain and restoring motion of joints at initial stage, but their long-term durability and biocompatibility have been questioned. Silicone microparticles from fragmentation or wear may cause immune reactions (20a). The largest use of silicone rubber and gel has been in breast augmentation and reconstruction (19b). Gel bleed, calcified deposit, and autoimmune diseases are concerns related to these gel-filled silicone rubber bag implants. In the earliest prosthetic cage-and-ball heart valves, silicone rubber was used for the ball, but uptake of blood lipids by the silicone led to the swelling and fracturing of the ball after several years of service. Presently, silicone rubber is only used for the suture ring in bioprosthetic heart valves (10d). Other medical applications of silicone rubber include soft contact lenses, catheters and drainage tubing, oxygenator membranes, wound dressings, and facial implants.

Polyamides

Polyamides are known as nylons. They are divided into two types: dyadic nylons and monadic nylons. The dyadic nylons, (e.g., nylon 66 and nylon 610) are made through condensation polymerization of diamine and dicarboxylic acid or its derivatives (Fig. 7a). There are two numbers following the name, the first for the number of carbon atoms in the diamine and the second for that in the diacid. The monadic nylons (e.g., nylon 6 and nylon 11) are made through self-amidation of an amino acid or through ring-opening polymerization of a cyclic lactam (Fig. 7b). The single number following the name represents the number of monomer carbon atoms. These polymers have high crystallinity and very strong intermolecular hydrogen bonding between amide groups. Thus, they have excellent fiber forming properties, and the strength along the fiber

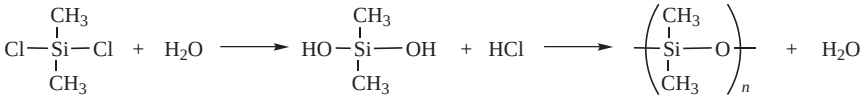


Figure 6. Formation of silicone prepolymer.

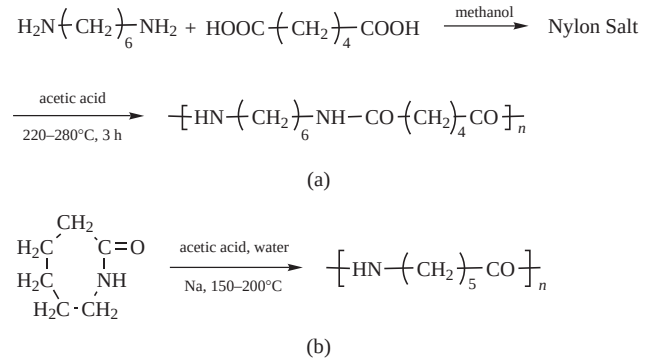


Figure 7. Synthesis of nylons: (a) nylon 66, (b) nylon 6.

direction is very high. The number and distribution of amide groups play an important role in determining the properties of nylons. Generally, the T_g and tensile properties increase with increasing the number of amide groups. For example, nylon 66 is stronger than nylon 610 and nylon 6 is stronger than nylon 11 (11a).

Nylon plastics are stiff and tough, and have a high abrasion resistance; but they are sensitive to water. Water adsorption reduces their strength and lowers their T_g. It is the amorphous region of polyamide chains that is sensitive to the attack of water. The greater the degree of crystallinity, the less the water adsorption. Nylons are used as surgical sutures, components of dialysis devices, hypodermic syringes, and intracardiac catheters.

Polyesters

Polyethylene Terephthalate. Polyethylene Terephthalate (PET) is known as Dacron. Commercially, it is manufactured by ester interchange polymerization between dimethyl terephthalate and excess glycol (1:1.7). The reaction has two stages (5d,12) (Fig. 8). In the first stage, methanol is displaced from dimethyl terephthalate by glycol. In the second stage, the excess glycol is driven off under vacuum. The polymer is semicrystalline with a T_m value of 265 °C and a T_g value of 80–120 °C depending on crystallinity. The PET fibers made from melt spinning have high strength and good crease resistance, so they are used as nondegradeable surgical sutures. Like PTFE, PET is also hydrophobic and hydrolysis resistant. The knitted or woven PET tubes are widely used for large diameter (12–30 mm) and medium diameter (5–11 mm) vascular grafts. However, Dacron graft devices are not fully satisfactory. Thrombosis is a major problem. Another drawback of Dacron grafts is the need for preclotting the grafts with autologous blood before implantation to prevent bleeding from their micropores (16a).

Polycarbonates. Bisphenol A polycarbonate is the only commercially significant polycarbonate product, so this material is often referred to as polycarbonate. It is prepared either by the reaction of phosgene with bisphenol A, or by ester interchange of a diphenyl carbonate with bisphenol A (Fig. 9). Polycarbonate is a clear, tough material. It has excellent mechanical and thermal properties (T_g 150 °C). The high transparency and impact strength make polycarbonate useful as lenses for eyeglasses and safety glasses, and housings for oxygenators and heart–lung bypass machines (2a). Requirements for medical grade polycarbonates are given in ASTM F 997.

Polyacetal

Polyacetal, also called poly(oxymethylene) is known as Delrin (DuPont). It is prepared from formaldehyde in an inert hydrocarbon solvent along with an initiator (ring-scission polymerization) (5d). The polymer has a high melting temperature (184 °C) and low glass transition temperature (−82 °C) (5b). It is lubricious, strong, and has good dimensional stability, resistance to creep and fatigue, high abrasion, and chemical resistance. A cementless polyacetal isoelastic femoral stem was introduced in the early 1970s to solve two important problems in total hip replacement (THR): stress shielding and cement disease. The modulus of elasticity of polyacetal (∼ 5–13 GPa) is close to that of bone, thus providing the condition of isoelasticity. However, the prostheses were unsuccessful in clinical practice due to high rate of loosening (20). The successful use of polyacetal in medical devices includes use in the valve disk in Penn State circulatory-assitant devices, which is one of only two approved for heart replacement by the U.S. Food and Drus admistration (FDA) (16d). Specifications for polyacetal are given in ASTM F 1855.

Hydrogels

Hydrogels are 3D networks of hydroplilic polymers held together by chemical or physical crosslinks. Typical methods to prepare hydrogels include irradiation, chemical reactions, and physical association. Hydrogels have inherently weak mechanical properties, so hydrophobic

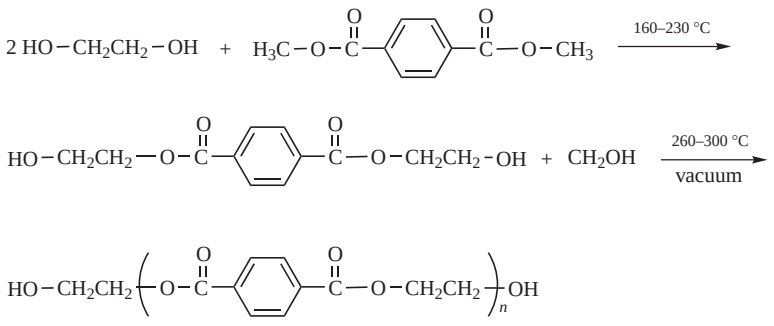


Figure 8. Synthesis of poly(ethylene terephthalate).

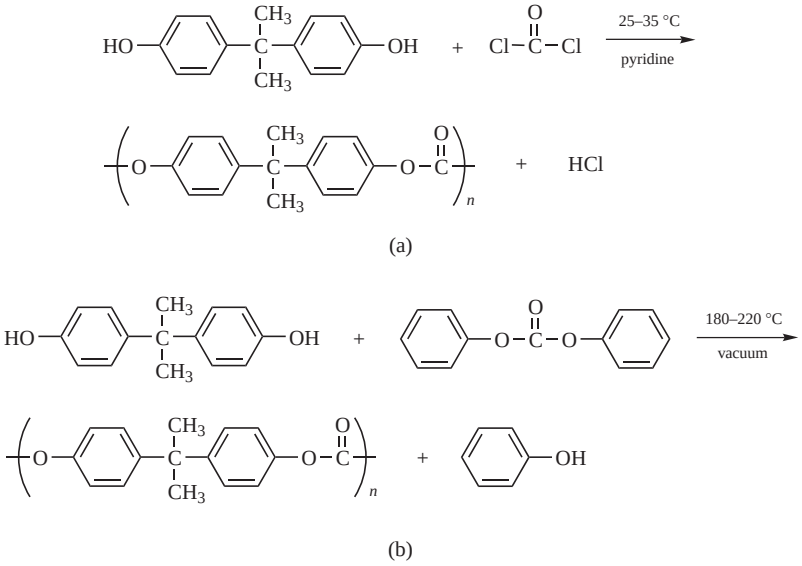


Figure 9. Synthesis of polycarbonate through (a) phosgene, (b) ester interchange.

constituents are often incorporated to improve the mechanical strength. The low interfacial tension with surrounding biological fluids makes hydrogels desirable for nonfouling surfaces (10c). Hydrogels tend to calcify under physiological conditions, limiting the use of hydrogels as implantable biomaterials (17c).

Poly(hydroxyethyl methacrylate) (PHEMA) is the most frequently used hydrogel. It is a rigid acrylic polymer when dry, but it takes up ~40% water when wet and changes into an elastic gel. PHEMA hydrogel is transparent, and thus is used for soft contact lens (17c). Other more hydrophilic hydrogel monomers, such as methacrylic acid and 1-vinyl-2-pyrrolidone, are often copolymerized with hydroxyethyl methacrylate to improve the oxygen permeability coefficient and water adsorption of PHEMA hydrogel. This hydrogel was also the first successfully used for wound dressings under the trade name Hydron (17c).

Poly(vinyl alcohol) is a water-soluble polymer. Solutions of PVA are used as ophthalmic lubricants and viscosity-increasing agents. The solution thickens the natural film of tears in eyes. Poly(vinyl alcohol) can crystallize even in its highly hydrolyzed state, and thus it has a relatively high tensile strength for a hydrogel. The PVA hydrogel is a candidate material for artificial articular cartilage in reconstructive joint surgery, and has been used for releasing bovine serum albumin. Physically cross-linked PVA hydrogels prepared by free-thaw processes have been investigated for protein-releasing matrix (17c).

Other synthetic hydrogels of biomedical interest include polyacrylamides, poly(*N*-vinyl-2-pyrrolidone), poly(methacrylic acid), and poly(ethylene oxide).

Biodegradable Synthetic Polymers

The interest in biodegradable polymers has dramatically increased in recent years, because these biomaterials do not permanently leave residuals in the implantation site, do not elicit permanent foreign-body reactions, and avoid second surgeries in the case of temporary implants like fracture fixation devices (10c,11b,17d). The major biomedical applications of biodegradable polymers include: temporary scaffolding and barrier, drug delivery matrix, tissue engineering scaffolds, and adhesives. Biodegradable polymers are generally hydrophilic. They degrade either by simple hydrolysis without enzyme catalysis or by enzymatic mechanisms. The degradation products should be inert or a natural metabolite of the body (11b).

The most commercially successful biodegradable polymers are polyglycolide (PGA), polylactide (PLA) and their copolymer poly(glycolide-*co*-lactide) (PLGA). The polymers PGA, PLA, and PLGA are a group of poly α-hydroxy acids belonging to absorbable polyesters. They degrade by bulk hydrolysis of their ester bonds. The hydrolysis byproduct of PLA is lactic acid, which is a normal byproduct of anaerobic metabolism in the human body (17e). The degradation of PGA involves both hydrolytic scission and enzymatic degradation, and the product is glycolic acid that can be eliminated by the metabolic pathway as carbon dioxide and water (1b).

Polylactide is prepared in the solid state through ring-opening polymerization. It has four stereoisomers: D-PLA, L-PLA, DL-PLA, and *meso*-PLA. The most frequently used forms in biomedical practice are L-PLA, and DL-PLA (2c). The L-PLA is a semicrystalline polymer with a T_m of 159–178 °C and T_g of 54–59 °C (17e). Compared with other biodegradable polymers, L-PLA exhibits high strength and modulus, and a very slow biodegradation rate. It is suitable for light load-bearing applications (e.g., orthopedic fixation devices, vascular grafts, and surgical meshes). Self-reinforced L-PLA bolts, screws, pins and anchors have been used for bone fracture fixation (1b). The DL-PLA is an amorphous polymer with a T_g of 51 °C. It degrades very fast, and thus usually is used for drug delivery (11b). The L-PLA polymer and its *in vitro* degradation tests are specified in ASTM F 1925 and F 1635, respectively.

Polyglycolide can be synthesized either by direct polycondensation of glycolic acid or by ring-opening polymerization of the cyclic dimers of glycolic acid. Due to tight molecular packing, PGA has a high melting point (225–230 °C). Its T_g ranges from 35 to 40 °C. The PGA degrades

faster than LPLA because it is more hydrophilic (11b) and PGA can be melt spun into fibers and fabricated into sutures, meshes, and surgical products. Dexon (American Cyanamid) is a trade name of polyglycolide products. It has been successfully used for wound closure and sutures (10c,11b). The Properties and degradation rate of PLGA can be controlled by varying the ratio of monomers. Vicryl from Ethicon is a poly(glycolide-L-lactide) random copolymer with 90:10 ratio of glycolide to lactide. It completely degrades *in vivo* after 90 days, and has been successfully used as surgical meshes and sutures, and for wound closure and drug delivery (10c,21b). Several other biodegradable polymers have been developed and investigated for drug delivery, tissue engineering, and medical devices [e.g., polyorthoesters, poly(ϵ-caprolactone), polydioxanone, and polyanhydride] (11b).

The concern with PLA and PGA is that their degradation products (lactic acid and glycolic acid) may significantly lower the local pH in a closed and less body-fluid-buffered region, leading to irritation at the site of polymer implant (11b).

Smart Polymers

Smart polymers are "polymers that respond with large property changes to small physical or chemical stimuli" (22). The most common stimuli are pH values and temperatures. Some polymers [e.g., poly(hydroxypropyl acrylate), poly(*N*-isopropylacrylamide), and poly(ethylene oxide/propylene oxide) copolymers] exhibit thermally induced precipitation and have a lower critical solution temperature (LCST). The polymers are soluble in water below LCST, but precipitate sharply as temperature is raised above LCST. Pluronic F127 (BASF, NJ) is a polyethylene oxide–polypropylene oxide–polyethylene oxide triblock copolymer with a LCST around the physiological temperature. It is used for controlled release of drugs including proteins and liposomes (23).

In general, pH-responsive hydrogels can be prepared from polymers with ionizable groups (e.g., carboxyl, sulfonic, amino, and phosphate groups). The pH change influences the ionization degree of the polymer to govern its solubility in water. Lysozyme (a cationic protein) immobilized within a hydrogel with phosphate groups is released at pH 7.4 (enteric conditions), but is not released at pH 1.4 (gastric conditions), ensuring the drug is delivered to the small intestines, and is not released in the stomach (17f).

Besides drug delivery, smart polymers can be used for sensors, chemical valves, mechanochemical actuators, specialized separation systems, and artificial muscles.

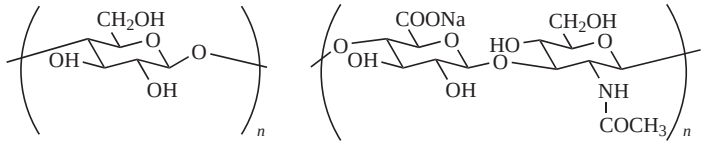
Biopolymers

Biopolymers are polymers of natural origin that can be obtained from animals, plants, and microorganisms. They are produced during the growth cycles of all organisms by enzyme catalyzed stepwise polymerization rather than a chain polymerization (5a). Biopolymers most frequently used for medical applications are polysaccharides and proteins.

Cellulose is a polysaccharide of plant origin. It is the primary structural component of plant cell walls. The molecular chain of cellulose is linear, consisting of D-glucose residues linked by β(1-4) glycoside bonds (Fig. 10a). The strongly hydrogen-bonded structure make cellulose highly crystalline and exceptional in strength, but insoluble in water and most organic solvents, and infusible. Ether and ester derivatives of cellulose, such as cellulose acetate and hydroxyethyl cellulose, have been developed to improve its processability (24). Cellophane (regenerated cellulose) semipermeable membrane was first used for hemodialysis to remove blood waste in the 1960s, and it is still in use today, but mostly in hollow fiber forms. Cellulose acetate (di- and triacetate) membranes and hollow fibers also find in hemodialysis. Hydroxypropylmethylcellulose is the most widely used hydrophilic drug delivery matrix (17f). The matrix tablets can be formed by compression, slugging, or wet granulation. Cellulose and its derivatives are also used as wound dressings.

Hyaluronan (HA) is a natural mucopolysaccharide, present in connective tissues of all vertebrates, which consists of repeating disaccharides of *N*-acetylglucosamine and glucuronic acid (Fig. 10b). Besides inherent biocompatibility, hyaluronan has some other unique properties: viscoelasticity, hydrophilicity, lubricity, and biological activity (regulator of cellular activity) (25). Unlike cellulose, hyaluronan is soluble in water forming a viscous solution, and it is biodegradable. Many derivatives have been developed to improve its residence time in water. Hylan (crosslinked hyaluronan, Biomatrix) and Hyaff (hyaluronan with carboxyl groups esterified by alcohol, Fidia Advanced Biopolymer) are two groups of widely used and commercialized hyaluronan derivatives. Native hyaluronan and Hylan are used for viscosupplementation to treat arthritis, viscoseparation to prevent adhesion and facilitate wound healing after surgeries, and viscoaugmentation to correct scars and facial wrinkles. Hyaff is widely used for wound dressings as sheets, meshes, or nonwoven fleeces. Hyaluronan and its derivatives have also been investigated for lubricious coatings of biomedical devices, control released drug delivery and tissue engineering scaffolds (26).

Collagens are a family of structural fiber proteins present in all animals (27). They are the most abundant proteins in



(a) (b)

Figure 10. Structure of polysaccharides: (a) cellulose, (b) hyaluronan.

vertebrates, widely distributed in pliant connective tissues and tensile structures (e.g., tendon and ligament) as scaffolds to keep their shape, maintain integrity, and support tensile stress. A collagen molecule is composed of three polypeptide chains wound into a triple helix, stabilized by interchain hydrogen bonds. They are subject to degradation by lysozomal enzymes and collagenase. Glutaraldehyde and carbodiimides can be used to crosslink collagens to improve their mechanical properties (10c). Collagens are probably the most common proteins used as biomaterials. They are used in artificial skins, in soft tissue, and plastic surgery to fill up tissue defects, in surgical sutures, vascular grafts, and corneal replacements (11c).

POLYMERIC BIOMATERIALS EVALUATION

Bulk Characterization

Initial characterizations give information about the composition, structure, and physical and mechanical properties of the investigated polymers to examine if the properties match the specific application requirements, and the reproducibility of batch-to-batch properties. Infrared (IR) spectroscopy and nuclear magnetic resonance (NMR) are often used to analyze the chemical composition and structures of polymers. Crystalline and multiphase structures are usually determined using wide- (WAXS) or small-angle (SAXS) X-ray scattering, or transmission electron microscopy (TEM). Differential scanning calorimetry (DSC), and dynamic mechanical analysis (DMA) can provide thermal transition information of polymers (e.g., T_m and T_g). Furthermore, DMA can provide significant insight into the viscoelastic nature of the polymer. The mechanical properties are determined using the standard ASTM methods. A second level of characterizations needs to be made on the candidate materials to investigate if sterilization and physiological environments' exposure significantly change their properties.

Surface Characterization

Biomaterials contact the body through their surface, and thus the surface chemistry and topography determine the host's response to the materials. The surface chemistry of polymers can be characterized with X-ray photoelectron spectroscopy (XPS), attenuated total reflectance Fourier transform infrared (ATR–FTIR) and secondary ion mass spectroscopy (SIMS). Scanning electron microscopy (SEM) and atomic force microscopy (AFM) are typical techniques to determine the surface topography. A contact angle goniometer is used to measure the wettability (i.e., hydrophilicity) of surfaces.

Biocompatibility Assessment

Biocompatibility tests generally include two levels. The first level tests are biosafety testing, while the second level involves biofunctional testing (28). In biosafety tests, the materials or their extracts are tested to see if they are toxic to cultured cells, cause hemolysis or allergic responses, induce heritable genetic alterations, or tissue necrosis after animal implantation. Those materials passing the first level tests need to be further inspected with the second level tests. This level of testing focuses on the specific functions of a medical device, in which the responses of all the cell and tissue types contacting the device are investigated with both *in vitro* and *in vivo* methods. The functionality of the medical device is also tested during this phase of testing, and changes in material can have significant effects on functionality just as changes in medical device design or function can have significant effects on the biocompatibility of the material.

BIBLIOGRAPHY

Cited References

1. (a) Ikada Y. Polymeric biomaterials in medical systems. (b) Mauliagrawal C. Biodegradable polymers for orthopaedic applications. In: Reis RL, Cohn D, editors. Polymer Based Systems on Tissue Engineering, Replacements and Regeneration. Dordrecht: Kluwer Academic Publishers; 2002.
2. (a) Cooper SL, Visser SA, Hergenrother RW, Lamba NMK. Polymers. (b) Hallab NJ, Jacobs JJ, Katz JL. Orthopedic applications. (c) Kohn J, Abramson S, Langer R. Bioresorbable and bioerodible materials. In: Ratner BD, Hoffman AS, Schoen FJ, Lemons JE, editors. Biomaterials Science: An Introduction to Materials in Medicine. 2nd ed., San Diego: Academic Press; 2004.
3. Ikada Y. Biological Materials. In: Barbucci R, editor. Integrated Biomaterials Science. New York: Kluwer Academic/Plenum Publishers; 2002.
4. Cheng HN, Gross RA. Polymer Biocatalysis and Biomaterials. In: Cheng HN, Gross RA, editors. Polymer Biocatalysis and Biomaterials. Washington, (DC): American Chemical Society; 2005.
5. Rodriguez F, Cohen C, Ober CK, Archer LA. Principles of Polymer Systems, 5th ed., New York: Taylor & Francis; 2003. Chapt 4(a); Chapt 3(b); Chapt 15(c); Chapt 16(d).
6. Young RJ. Introduction to Polymers. London: Chapman and Hall; 1981. Chapt 2.
7. Matyjaszewski K. Comparison and classification of controlled/living radical polymerizations. In: Matyjaszewski K, editor. Controlled/Living Radical Polymerization: Progress in ATRP, NMP, and RAFT. Washington, (DC): American Chemical Society; 2000.
8. Chanda M. Advanced Polymer Chemistry: a Problem Solving Guide. New York: Marcel Dekker; 2000; Chapt 1(a); Chapt 2(b).
9. Askeland DR, Phule PP. The Science and Engineering of Materials. 4th ed., Boston: PWS Publishing Company; 2003; Chapt 15.
10. Bhat SV. Biomaterials. Boston: Kluwer Academic Publishers; 2002; Chapt 12(a); Chapt 8(b); Chapt 5(c); Chapt 9(d); Chapt 6(e).
11. (a) Lee HB, Khang G, Lee JH.. Polymeric biomaterials. (c) Chu CC. Biodegradable polymeric biomaterials: an updated overview. (d) Li ST. Biologic biomaterials: tissue-derived biomaterials (collagen). In: Park JB, Bronzino JD, editors. Biomaterials: Principles and Applications. Boca Raton (FL): CRC Press; 2003.
12. Alger MSM. Polymer Science Dictionary. London: Elsevier Science Publishers; 1989.
13. Birnkraut HW. Synthesis of UHMW-PE. In: Willert H-G, Buchhorn GH, Eyerer P, editors. Ultra-High Molecular Weight Polyethylene as Biomaterial in Orthopedic Surgery. Toronto: Hogrefe & Huber Publishers; 1991.

14. Kurtz SM, Muratoglu OK, Evans M, Edidin AA. Advances in the processing, sterilization of ultra-high molecular weight polyethylene for total joint arthroplasty. Biomaterials 1999;20:1659–1688.
15. Teoh SH, Tang ZG, Hastings GW. Thermoplastic polymers in biomedical applications: Structures, properties and and processing. In: Black J, Hastings G, editors. Handbook of Biomaterial Properties. London: Chapman & Hall; 1998.
16. (a) Ramshaw JAM, Werkmeister JA, Edwards GA. Tissue-polymer composite vascular prostheses. (b) Kowligi RR, Edwin TJ, Banas C, Calcote RW. Vascular grafts: materials, methods, and clinical applications. (c) Planell JA, Vila MM, Gil FJ, Driessens FCM. Acrylic bone cements. (d) Felder G III, Donachy JH, Sr. Fabrication techniques and polymer considerations for the blood contacting components of the Penn State circulatory–assist devices. In: Wise DL, Trantolo DJ, Altobelli DE, Yaszemski MJ, Gresser JD, Schwartz ER, editors. Encyclopedic Handbook of Biomaterials and Bioengineering, Part B: Applications Vol. 2. New York: Marcel Dekker; 1995.
17. (a) Tomita N, Fujita H, Nagata K. Polymers for artificial joints. (b) El-Zaim HS, Heggers JP. Silicones for pharmaceutical and biomedical applications. (c) Kishida A, Ikada Y. Hydrogels for biomedical and pharmaceutical applications. (d) Rokkanen PU. Bioabsorbable polymers for medical applications with an emphasis on orthopedic surgery. (e) Domb AJ, Kumar N, Sheskin T, Bentolila A, Slager J, Teomim D. Biodegradable polymers as drug carrier systems. (f) Miyata T, Uragami T. Biological stimulus-responsive hydrogels. (g) Dumitriu S. Polysaccharides as biomaterials. In: Dumitriu S, editor. Polymeric Biomaterials, 2nd ed. New York: Marcel Dekker; 2002.
18. Lamba NMK, Woodhouse KA, Cooper SL. Polyurethanes in Biomedical Applications, Boca Raton (FL): CRC Press; 1998. Ch. 2.
19. (a) Szycher M, Reed AM. Biodurable polyurethane elastomers. (b) Tiffany JS, Petraitis DJ. Silicone biomaterials. In: Wise DL, Trantolo DJ, Altobelli DE, Yaszemski MJ, Gresser JD, Schwartz ER, editors. Encyclopedic Handbook of Biomaterials and Bioengineering, Part A: Materials (Vol. 2). New York: Marcel Dekker; 1995.
20. (a) Pappas MA, Schmidt CC, Shanbhag AS, Whiteside TA, Rubash HE, Herndon JH. Biological response to particulate debris from nonmetallic orthopedic implants. (b) Gresser JD, Trantolo DJ, Lyons CH, Nagaoka H, Shuster L, Swift RM, Wise DL. In vitro and in vivo release of naltrexone from two types of poly(lactide-co-glycolide) matrices. In: Wise DL, Trantolo DJ, Altobelli DE, Yaszemski MJ, Gresser JD, editors. Human Biomaterials Applications. Totowa (NJ): Humana Press; 1996.
21. Minovic A, Milosev I, Pisot V, Cor A, Antolic V. Isolation of polyacetal wear particles from periprosthetic tissue of isoelastic femoral stems. J Bone Joint Surg 2001; 83B:1182–1190.
22. Hoffman AS, Stayton PS, Bulmus V, Chen G. Really smart bioconjugates of smart polymers and receptor proteins. J Biomed Mater Res 2000;52:577–586.
23. Chandaroy P, Sen A, Hui SW. Temperature-controlled release from liposomes encapsulating Pluronic F127. J Controlled Release 2001;76:27–37.
24. Klemm D, Philipp B, Heinze T, Heinze U, Wagenknecht W. Comprehensive Cellulose Chemistry Vol. 2: Functionalization of Cellulose. Weinheim, Germany: Wiley-VCH; 1998.
25. Laurent TC. The Chemistry, Biology and Medical Applications of Hyaluronan and its Derivative. London: Portland Press; 1998.
26. Kennedy JF, Philips GO, Williams PA. Hyaluronan 2000. Cambridge (England): Woodhead Publishing Limited; 2002.
27. Wainwright SA, Biggs WD, Currey JD, Gosline JM. Mechanical Design in Organism. Princeton (NJ): Princeton University Press; 1982. Chapt 3.
28. Zhang M. Biocompatibility of materials. In: Shi D, editor. Biomaterials and Tissue Engineering, Heidelberg: Springers; 2003. p 83–137.

See also BONE CEMENT, ACRYLIC; CONTACT LENSES; LENSES, INTRAOCULAR; POLYMERIC MATERIALS.

BIOMATERIALS, SURFACE PROPERTIES OF

SALLY L. MCARTHUR
ALEXANDER G. SHARD
University of Sheffield

INTRODUCTION

In the broadest of definitions, biomaterials are nonliving materials that come into contact with biological systems. The point of contact between the two different phases is at the interface, or surface, of the material. It is quite common for the surface of a material to have properties that are not trivially related to the bulk of the material. These differences can arise because of a number of processes, such as surface segregation, surface reactions, contamination, scratching, and phase separation. It should therefore be recognized that the interactions between a biomaterial and the biological medium and in turn, the physical and chemical activity or stability of a medical device, can depend critically upon the properties of the surface.

In general, materials selection for biomedical devices and applications is based on a combination of physical properties, manufacturability, and availability. In many cases, materials are chosen because they have been used previously in medical devices and as such, detailed records of *in vivo* behavior and performance already exist. Due to the costs and time involved with the testing of new materials to meet regulatory standards for safety and efficacy, relatively few materials are currently used in the manufacture of biomedical devices. The most common of these are titanium-based alloys, 316L stainless steels, ultrahigh molecular weight polyethylene (UHMWPE), expanded poly(tetrafluoroethylene) (e-PFTE), poly(ethylene terephthalate) (PET), poly(hydroxyethyl methacrylate) (pHEMA), polyglycolic and lactic acids (PGA and PLA), polystyrene, polyurethanes, hydroxyapatite, alumina, and zirconia.

Of course, mechanical properties are only one of a number of materials characteristics that may be required for biomedical applications. Each biomedical application may desire a range of properties that are directly influenced by the nature of the surface. Specific characteristics and modifications made to biomaterial surfaces include:

1. Orthopedic Devices

Improved wear resistance and frictional properties for joints and bearing surfaces via cross-linking of UHMWPE and the introduction of carbide, nitride and crystalline structures on metallic components.

Bone conductive coatings for improved osseointegration via implantation of specific chemical species (e.g., Ca, P) into metals and deposition of hydroxyapatite coatings

2. Cardiovascular Devices

Improved hemocompatibility via diamond-like carbon (DLC) coatings (e.g., mechanical heart valve leaflets) and via the immobilization of biomolecules to promote epithelialization

Short- and long-term drug delivery via degradable polymeric coatings (e.g., stents).

Polymeric barrier coatings to prevent transmission of electrical signals and improve corrosion resistance (e.g., pacemaker cases and leads).

3. Diagnostics, Sensors and *in vitro* Applications,

Reflective coatings for optical sensors.

Nonfouling coatings to prevent protein and cell attachment and reduce background signal in biological assays and sensors.

Oriented biomolecule immobilization for DNA, protein, and antibody arrays.

Topographical and chemical patterning of microfluidic devices and sensors for the control of fluid flow and chemical mixing.

4. Tissue Engineering

Improved cell proliferation and growth in culture via oxidation of polystyrene to produce a hydrophilic substrate (tissue culture polystyrene, TCPS).

Immobilization of biological ligands for controlled cell adhesion (e.g., RGD and other cell receptor binding domains).

In this article, we intend to provide a broad overview of the basic properties of surfaces, their interactions with biological systems, and how surfaces can be changed to suit particular biomaterial applications. Of particular importance is the requirement for surface characterization. As stated earlier, differences between surface and bulk properties can arise via a number of different processes. However they arise, it is important to ensure that the surface properties of the material are verified before ascribing any biological effect to the material. To complete this article, we provide an outline of the most commonly employed surface characterization techniques and include references to more detailed texts to aid the interested reader.

PROPERTIES OF SURFACES

One of the most important properties of a surface or interface is that it exhibits free energy. This means that if the surface was extended in some way so that it had a larger area then work would have to be done. If this was not the case, then for fluid interfaces at least, the surface could grow without limit, eventually resulting in a homogenous mixture. The existence of surface energy leads to a tendency for surfaces to contract resulting in a higher pressure on the inside of curved surfaces. Measuring the interfacial energy between liquids and air is relatively trivial, as the surface may be extended without producing a bulk strain. Thus the energy required to extend an area of surface or, more usually, the force of contraction normal to a length of surface can be directly obtained. Liquid surface tensions scale with the strength of intermolecular interactions in the bulk of the liquid, so hydrocarbons typically have surface energies of $\sim$25 $mN \cdot m^{-1}$, water has 72 $mN \cdot m^{-1}$ due to hydrogen bonding and the metallic bonding in mercury results in a surface energy of over 470 $mN \cdot m^{-1}$ (1).

In contrast, the surface energy of solids cannot be obtained directly. There are, however, a variety of methods of estimating it from a series of contact angle experiments and it is found that the surface energies between solids and air are very similar to those of analogous liquids. However, for most biomaterial applications it is the solid–water interfacial energy that is important. One should note that "low energy" hydrophobic surfaces typically have interfacial energies with both air and water of $\sim$30 $mN \cdot m^{-1}$. Hydrophilic surfaces, such as clean glass or aluminium, can have high surface energies in air, higher than 80 $mN \cdot m^{-1}$, but have negligibly small interfacial energies with water. The difference between the air and water interfacial energies for glass and aluminium is greater than the surface energy of water, and hence water does not form drops, but completely wets these materials. Protein adsorption, described in detail later, can be thought of as being driven by the minimization of surface energy. A comparison of the surface energies for hydrophilic and hydrophobic materials gives an appreciation of the strength and importance of the hydrophobic interaction, described later, during this process.

Other properties of surfaces that are important are chemistry, mechanical properties, and topology. In the context of a biomaterial, the chemistry of a surface will determine the initial interactions with proteins through ionic, hydrogen bonding, and hydrophobic interactions as well as the promotion of specific interactions by the presence of surface bound ligands. It is important to realize that the surface chemistry of a material may bear little or no resemblance to the bulk chemistry. In many cases, this is due to the presence of thin layers of contaminants that naturally accumulate on the surfaces of all materials. The deliberate alteration of biomaterial surface chemistry is carried out to enhance or inhibit certain properties, usually the alteration of protein adsorption and cell attachment. Whether the surface chemistry is a result of contamination or modification, it is important to specifically characterize the surface to ensure that correlations between biomaterial chemistry and performance are correctly obtained. The mechanical properties of a biomaterial surface are also of some importance, particularly for cellular attachment. It is generally found that cells attach more strongly to rigid substrates and will migrate from soft-to-hard materials. Note that the mechanical properties of some materials, in

particular polymers, may be somewhat different at the surface compared to the bulk. It has been observed, for example, that the glass transition temperature of polymers is reduced close to an interface. The topology of a surface is also important as it defines the surface area of the interface, and has been shown to influence cell behavior (2).

The Adsorption of Proteins at Surfaces

One of the most important events that occur in biomaterial applications is the sequestration of proteins from solution to the surface of the material. Proteins are polyamino acids in which, for each protein, there is a predetermined and specific sequence of amino acids. This sequence is termed the primary structure of the protein. The secondary structure consists of a variety of common folding motifs, such as α-helices and β-sheets. The tertiary structure of the protein comprises the folding and packing of the secondary structure into a particular three-dimensional (3D) shape. For most proteins, the tertiary structure creates unique, and often rather small sites of activity that allow the protein to function (e.g., cell-binding domains). In contrast, synthetic macromolecules form random coils because they lack the well-defined structure that allows the strong bonding that occurs between different parts of the protein chain.

When one considers protein adsorption at interfaces, it is common to draw analogies to the adsorption of synthetic macromolecules. While these comparisons are extremely useful, it is important to remember that proteins are capable of site specific and highly selective binding, whereas synthetic macromolecules in general are not. Examples of such selective binding include the much utilized affinity of avidin for biotin and the binding of antigens to antibodies. Protein adsorption occurs primarily due to a number of intermolecular forces. These include ionic and hydrogen bonding and the hydrophobic interaction (3). Although the ionic interaction is rather strong in solid materials, in aqueous media it is diminished due to strong ion–dipole interactions with water, the high dielectric constant of water and the presence of other solvated ions that cause a decrease in the effective range of ionic interactions. Nevertheless, ionic interactions are important at short ranges and can have a strong effect on the rate of adsorption of proteins at surfaces. It is commonly observed, for example, that a protein adsorbs most rapidly to an uncharged surface when it is at its isoelectric point, that is, when it is itself uncharged. The presence of a charged interface can decrease or increase the rate of adsorption depending on whether the protein has a like or an unlike total charge. Furthermore, if the protein has a dipole moment, then it may be possible to influence the orientation of the protein upon adsorption at a charged surface.

Hydrogen bonding is a particularly strong example of a dipole–dipole interaction. A hydrogen atom bound to an electronegative element such as oxygen or nitrogen forms a strong association with a lone pair of electrons on another electronegative atom, which may be part of another molecule. There is no great driving force for the formation of hydrogen bonds in the presence of water, since water very effectively makes such bonds. Without the generation of highly specific geometries of complementary hydrogen donors and acceptors, hydrogen bonding is almost certainly not a major driving force for adsorption of proteins at interfaces.

The "hydrophobic interaction" is something of a misnomer, since the driving force is in fact the formation of hydrogen bonds in water and not the attraction between two hydrophobic species. Water cannot form hydrogen bonds with regions of predominantly hydrocarbon species, whether these are part of a protein or on a surface. The result is that at such an interface water is in a state of higher free energy than if the interface was not present. Hydrocarbons thus tend to aggregate together to minimize the area of contact between themselves and water and lower the free energy of the system as a whole. These interactions are critical to the folding of proteins, with the interior of the protein generally consisting of hydrophobic amino acids and the exterior of hydrophilic amino acids. It is undoubtedly also an important interaction in the adsorption of proteins at interfaces. While the exterior surface of most globular proteins contain few hydrophobic sites, if the protein can unfold upon the surface (denature) then many more such sites become available.

When a surface is exposed to a solution of a single protein it is generally found that adsorption occurs rapidly and in many cases is diffusion limited. It is usual for adsorption to reach a maximum at a single layer with close contact between adsorbed proteins. Following adsorption, the rate of desorption from the surface is extremely slow. Proteins cannot commonly be removed from surfaces simply by changing the protein solution for pure solvent. However, if other proteins are present there may be exchange between adsorbed and solvated proteins. This includes self-exchange, as has been demonstrated by the exchange of unlabelled proteins with their radiolabeled analogues (4). Different proteins can have different affinities for surfaces, so that one protein may adsorb initially because it is in a high solution concentration, but at later times be displaced by other proteins that have higher affinity, but are in low concentrations. This effect is named after Leo Vroman and the classic example is the adsorption of proteins from serum that occurs in the order albumin, fibrinogen, and high molecular weight kininogen. It is also thought that immunoglobulin G adsorbs transiently between albumin and fibrinogen (5). This exchange can take place in a matter of seconds in pure serum, but may take minutes or hours in diluted serum. It is also noted that the amount of protein that can be exchanged in this manner diminishes the longer the protein is in contact with the surface. This indicates that the initial state of adsorption is metastable and that some activation energy barrier needs to be overcome for an adsorbed protein to reach a free energy minimum. It is possible that this energy barrier relates to the unfolding of tertiary or secondary structure and represents a denaturation of the protein.

Although the precise details of protein adsorption are unclear, it is generally agreed that the stability of adsorbed protein layers derives from the large number of contact points possible between a single protein molecule and a surface. Although each individual contact may be weak and temporarily displaced by smaller molecules the probability of breaking enough bonds for the protein to actually desorb is extremely small. The stability of an adsorbed layer is

therefore related to both the strength of individual interactions with the surface and the number of interactions. One should expect on this basis that, neglecting the detailed interactions and protein conformation, a high molecular weight protein should displace a low molecular weight protein because it is able to form more bonds to the surface. It is instructive to note that this trend is at least partially followed in the Vroman effect, the exception being high molecular weight kininogen that has a slightly lower molecular weight than fibrinogen.

Cell Behavior at Surfaces

In comparison with protein adsorption, the adhesion of cells to a biomaterial surface is a rather slow process. In standard cell culture, the adsorption and equilibration of proteins at the surface will occur much more rapidly than cellular attachment. The behavior of cells at a surface is thought to be governed by the initial layer of protein. Cells with surfaces via interactions of their transmembrane proteins (e.g., integrins) with proteins in the extracellular matrix. One approach to encourage cell adhesion is to incorporate such specific sequences at the surface of the biomaterial. A variety of suitable peptide sequences have been reported. From fibronectin, the RGD sequence mentioned above and also REDV, which targets integrins found in endothelial cells, but not blood platelets. Laminin contains sequences such as YIGSR and SIKVAV, which may be employed to encourage nerve cell growth (6).

If cell attachment and growth is to be discouraged, then the biomaterial surface should adsorb as few proteins as possible or only adsorb proteins that are not implicated in cellular adhesion. In the first alternative, this is typically achieved by using a hydrogel-like polymer layer, such as grafted chains of polyethylene glycol. These highly hydrated films provide few sites for protein attachment and cell attachment is also strongly discouraged. It is commonly observed that cells attach poorly to hydrophobic surfaces; this may indicate that there is a selective adsorption of proteins that do not contain binding domains for cells. The modification of surfaces to promote and inhibit cell attachment is discussed later in this article.

Once a cell has formed attachment points at a surface it will strengthen these by accumulating integrin receptors in the vicinity of each site. These eventually form a focal adhesion that acts as a connection between the actin cytoskeleton of the cell and the surface. As these adhesive contacts are made the cell spreads upon the surface and will then enter the normal cell cycle. The formation of focal adhesions is critical to the survival of the cell, without sufficient spreading a cell will normally die. There are proteins that trigger signals from the focal adhesion to the cellular interior such as focal adhesion kinase, which may be implicated in this decision making process.

The movement of mammalian cells is achieved by crawling. This involves the myosin driven contraction of actin filaments in the cell to supply the mechanical power, the detachment of focal adhesions at the trailing edge of the cell and the formation of new adhesions at the leading edge (7). Cells will generally move in the direction in which they can make the largest number of focal adhesions. The surface of a biomaterial can thus be tailored to concentrate cells in particular locations.

SURFACE MODIFICATION

In many cases, surface characteristics can be modified by designing the chemical constituents of the materials, for example, surface segregating components in polymer blends to alter frictional properties; or induced during the manufacturing process, for example, the introduction of topography via die and mould design. However, it is not always possible or practical to use these approaches and secondary processing capable of inducing specific surface properties without detrimentally affecting the bulk characteristics is often required.

In broad terms, surface modification techniques can be divided into two categories: those that treat the existing surface and those that result in the addition of a surface coating. As shown in Table 1, there are a number of different surface modification techniques that are currently used in industry or applied to bioengineering research. In this section, we give a brief overview of a number of these techniques, discuss their advantages, and limitations and give some specific examples of their application.

Plasma Treatment and Polymerization

Plasma-based modifications have been applied, with varying degrees of success, to biomaterials and biomedical devices since the early 1960s. Also termed radio frequency glow discharge (rfgd), the process involves the volatilization of a liquid or gaseous "monomer" into an evacuated process chamber. An electric field at rf is applied across the vapor, ionizing a fraction of the molecules and generating electrons, ions, free radicals, photons, and molecules in both ground and excited states, within the gas plasma. When the resultant reactive species impinge on a surface within the plasma zone, they create reactive sites resulting in alteration of the surface chemistry and properties.

There are two classes of glow discharge plasma modification, treatment and Polymerization. Plasma treatment results in the introduction of chemical species or physical changes to the surface of the material. Plasma treatments are often used to etch polymeric, metallic and ceramic surfaces, remove contaminants, and improve adhesion and hydrophilicity (8). Chemical modifications resulting from plasma treatments can also be used as an activation step for graft polymerization. Plasma generated radicals can be used to initiate polymerization of monomers in the liquid or gas phase, resulting in surface -grafted polymer layers. Typically "monomers" used for plasma treatment include oxygen, argon, ammonia, air, and water.

Plasma Polymerization occurs when a plasma is struck in an organic vapor and results in the deposition of a polymeric film from the vapor phase. Excitation of the monomer results in reactive species impinging on a surface within the plasma zone creating reactive sites that are then used for the covalent attachment of other species and subsequent growth of a coating of controllable thickness (typically tens of nanometers). A wide range of monomers can be used to produce plasma polymer coatings suitable

Table 1. Methods and Applications of Surface Modification Commonly Used in Biomedical Devices

Method	Application	References
Plasma polymerization	Organic and inorganic coatings for use as barrier coatings (thermal and chemical). Improved abrasion resistance, electrical and optical properties. Control of chemical functionality, cell and protein adhesion	8,9
Plasma treatment	Introduction of chemical functionality, crosslinking of polymers for improved wear and frictional properties	9,10
Plasma immersion or source ion implantation (PIII)	Wear resistance and improved friction properties for metals ceramics and polymers. Improved biocompatibility	11–13
Radiation techniques [ultraviolet (UV), gamma and laser irradiation]	Polymer grafting, introduction of topographical features and chemical functionality	14–17
Ion implantation	Improved wear and friction properties. Implantation of specific elements can improve cellular integration on polymers and metals	18–20
IBAD	Enhanced cell and tissue compatibility, antimicrobial properties, friction, wear, and chemical stability.	20–22
Polymer grafting	Nonfouling and biomimetic surfaces. Control of hydrophilicity, introduction of chemical functionality. Chemical, thermal, and biologically responsive coatings	23–25
Biomolecule immobilization	Biomimetic surfaces, introduction of specific biological function and activity.	26–29

for biomaterials applications. Table 2 lists some of the most common monomers and their applications. In general, plasma polymers tend to be highly cross-linked and do not reproduce the chemistry of the monomer. In the last 10 years, there has been increasing interest in the production of plasma polymers with the functionality and specific characteristics of their parent monomer. This can range from simple systems for retaining more amine or acid functionality in coatings (30) to more complex cases such as optimizing the protein resistance of poly(ethylene oxide)-like plasma polymers (31) or the production of thermally responsive *N*-isopropylacrylamide (NIPPAm) surfaces (32).

A range of deposition parameters can be used to manipulate the characteristics of a plasma polymer and encourage coating properties that are commensurate with those of a traditionally synthesized polymer. Lower deposition powers, pulsing of the power supply, and copolymerization have all been used to modify the coating properties (30,33,34). The resulting materials have been shown to retain higher monomer functionality and in some cases specific physicochemical properties normally associated with multi-step polymer grafted surfaces (32).

Ion Implantation and Ion-Beam Assisted Deposition

As is the case with plasma techniques discussed previously, the key difference between these two ion-based surface modification techniques is that ion implantation is a surface treatment while ion-beam assisted deposition (IBAD), as the name suggests, results in a surface coating. In both cases ionized species are produced via an ion source and accelerated in an electric field to reach the surface with kiloelectronvolt energies. Parameters affecting the process include beam energy, dose, and current density as well as the nature of the ion species (20).

In polymers, ion impacts and interactions induce modifications of the macromolecular structure through gas evolution, formation of double bonds, chain scissions, and cross-linking over a thickness corresponding to the penetration depth of the ions. Factors such as chain scission and cross-linking obviously have diametrically opposing effects on the properties of the polymeric surface. Generally, manufacturers utilize ion-beam implantation to increase cross-link density, a factor that can improve wear properties at load bearing interfaces and create polymers with improved chemical resistance. In metals and ceramics, ion implantation can be used to induce the formation of new surface phases, surface disorder. The formation of hard-phase nitride, carbide, and oxide precipitates via ion implantation has been used to harden the surfaces of Ti and Ti alloy orthopedic implants, improving their wear resistance (35). The application of specific ion species such as nitrogen and calcium enables the generation of specific chemical

Table 2. Common Plasma Polymerization Monomers and the Coating Properties They Produce

Monomer	Coating Properties and Applications
Organosilanes (silanes and disiloxanes)	Thermal and chemical resistance Specific electrical and optical properties
Fluorine (e.g.) and hydrocarbon (octadiene) containing	Hydrophobic coatings Chemical barrier coatings, non cell adhesive.
Acid containing (acrylic acid)	Hydrophilic coatings
Amine containing (heptylamine, allylamine)	Acid and amine functionality used for polymer and biomolecule immobilization Controlled cell attachment and growth
Ethylene oxide containing (glymes, diethylene glycol vinyl ether)	Nonfouling coatings Controlled protein and cell adhesion

changes at the surface. The bone conductivity, corrosion and wear resistance of Ti alloys have all been shown to improve after calcium and phosphorous ion implantation (35). On polymeric materials, nitrogen ion implantation has been used to induce complex crosslinked surfaces with increased solvent and wear resistance (36), while the incorporation of silver (Ag) ions has been used to impart antimicrobial properties on indwelling catheters (21). Ion-beam assisted deposition, combines ion-beam implantation with physical vapor deposition (PVD), producing a low stress, uniform and adherent coating via interactions of the ions from the beam with the coating atoms (20). Ion-beam assisted deposition (IBAD) has been used to produce a variety of metallic and inorganic coatings on Co–Cr and Ti alloys, alumina, and UHMWPE (37). Titanium alloy coatings have been produced on Co–Cr components to improved cellular integration in orthopedic applications (38) and bone conductivity has been improved on a variety of metallic substrates with the deposition of adherent hydroxyapatite coatings. Commercially, IBAD is used to produce DLC coatings that are chemically inert, optically transparent, have a low friction coefficient, and are extremely hard. These DLC coatings are used to treat the bearing surfaces of orthopedic implants to improve wear and friction properties and reduce the incidence of wear debris (39). On polymeric substrates, IBAD is used to produce adhesive silver coatings for antimicrobial applications (40).

Plasma Immersion Ion Implantation

Plasma immersion ion implantation (PIII) has a critical advantage over the standard ion implantation methods discussed in the previous section: it is not a line of sight technique, a factor that enables the modification of complex shapes commonly found in biomedical applications. Unlike ion implantation, PIII samples are pulse-biased to a high negative potential relative to the chamber wall and surrounded by high density plasma. Ions generated in the plasma are accelerated across the sheath formed around the samples and are implanted into the surface. Gaseous plasmas can be induced using a variety of sources including rf and microwave, and combining these gas plasmas with a metallic plasma allows interface mixing that result in metallic coatings with low intrinsic stress, significantly reducing the risk of coating delamination (12). Plasma immersion ion implantation has been used to surface modify skeletal prosthetic implants with Ti alloy coatings (for cell recruitment) while maintaining the mechanical properties of the Co–Cr substrate (41) and to deposit carbon and Ti–N coatings on both metals and polymers for improved wear and scratch resistance (12).

Wet Chemical Techniques

There is a vast array of wet chemical techniques that can be used to modify the surfaces of biomaterials. In their simplest incarnation, wet chemical routes for surface modification can involve the immersion of a device in a chemical bath to adsorb polymer to the surface. The complexity of the modification increases incrementally through the grafting of polymers to form nonfouling and bioresponsive coatings toward the construction of biomimetic surfaces that attempt to imitate the outer surface of a cell and can contain a range of lipids, proteins, and sugars.

Grafted polymer layers can be used to manipulate both the physical and the chemical characteristics of a surface. Grafting can be achieved via a number of routes including covalent coupling, surface graft polymerization; surface segregation and interpenetration of a substrate. One of the most popular current applications is in the generation of nonfouling surfaces via the immobilization of water-soluble polymers like polyethylene oxide (PEO). While there is considerable debate on the efficacy of these coating, recent reviews on surface modification for nonfouling behavior have detailed the critical roles of polymer molecular weight, graft density, and residual charge on the performance of these types of grafted polymer layers (23,42).

An alternative approach to polymer grafting is the self-assembly of molecules to form monolayers. Self-assembled monolayers (SAMs) can be formed spontaneously via a range of specific molecule–surface interactions. Common systems include alkane thiol on gold or silver and chlorosilanes on hydroxyl-terminated surfaces. By tailoring the headgroup chemistry of the immobilized molecules, surface can be designed with a range of properties. Commercially, these systems are currently used as platforms for biosensors and bioarray technologies (43). The well-defined nature of these systems has resulted in their extensive application in research as model systems for protein and cell–surface interaction studies (42).

Increasing focus on the development of coatings capable of eliciting specific biological responses has seen a significant research focus on the incorporation of peptide sequences, particularly from the receptor-binding domains of adhesion proteins, in order to promote cell adhesion (6,27). In more general terms, there is significant interest in the immobilization of a range of biomolecules. Array and sensor technologies require antibody, protein, and DNA immobilization, while the immobilized proteoglycans such as heparin have been used to the modulate hemocompatibility of biomedical devices (44). As illustrated in Fig. 1, immobilization strategies for biomolecules can be as simple as nonspecific adsorption or as complex as molecular imprinting. The adsorption of biomolecules tends to result in coatings with limited functionality as the molecules are randomly oriented and tend to be desorbed from the surface over time. Covalent immobilization can eliminate problems associated with desorption, but there is often little control over conformation or orientation and thus activity of the biomolecule can be limited. The use of spacer polymer chains or amino acid sequences between the surface and the protein can reduce the denaturation of the molecule. Examples of this type of approach are the site specific modification of proteins with cysteine that enable immobilization of the proteins in specific orientations on gold (45). The success of strategies designed to present biological ligands can also be maximized if the immobilized molecule is coupled to a surface capable of preventing nonspecific adsorption. In general terms, biological response is influenced by the presentation, average density and the spatial distribution of the immobilized molecule

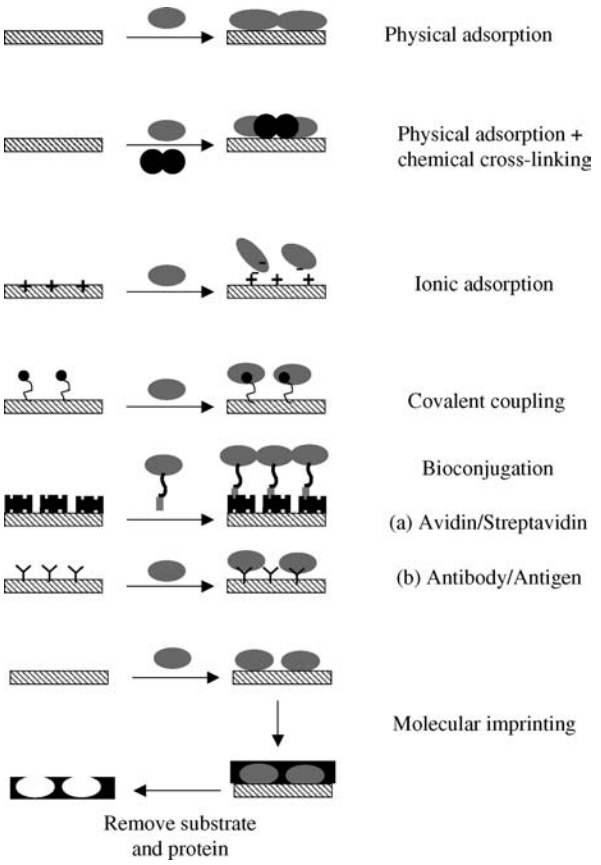


Figure 1. A range of immobilization strategies for biomolecules at interfaces.

(1). One of the most challenging applications for wet chemical surface modification lies in the development of surfaces that borrow from structures observed in Nature. Surfaces that mimic the structure of a cell wall are increasingly sought for use in biosensors and as model systems to further investigate cell, protein, and pharmaceutical interactions. These systems generally consist of a lipid bilayer that may containing a range of different lipids, transmembrane and membrane proteins, and in some cases oligosaccharides. Critically, these structures need to retain their fluidity; molecules need to be able to move within and through the structure in order to maintain their activity and function. The most simple cell mimetic surfaces discussed in the literature have been based on transferring lipid bilayers onto glass substrates. Under these conditions, a thin film of water lubricates the interface between the glass and bilayer and allows free lateral diffusion. These types of bilayers generally have poor long-term stability, show limited transmembrane protein activity, and cannot be transferred through the air–water interface without disrupting the structure (46). An alternative approach lies in either the formation of hybrid bilayers, where the inner leaflet is formed from alkane thiol on gold (47), or the deposition of the lipids on a polymer support (46). Hydrogel layers can also be used and act as a hydrated cushion that is both a self-lubricating and a spacer, creating an area for protein insertion without affecting protein function. At present, there are a number of commercial biosensors that utilize this type of technology for disease diagnosis (48).

SURFACE ANALYSIS

The study of surfaces and coatings is an advanced field and ranges from the investigation of elementary chemical processes on single crystals in ultrahigh vacuum (UHV) to the analysis of rather more "dirty" and real surfaces in engineering applications. The development of techniques suitable for surface science has a long history, the driving force for which has only recently included attempting to understand and control biomaterial surface interactions. Table 3 details some of the more common surface analysis techniques used today in the characterization of biomaterials. The physical principles of all the techniques commonly employed today are well understood and have been for many decades. Many of the approaches described here have their origin in the study of elementary chemical and physical processes, the semiconductor industry, and from engineering disciplines.

When considering the choice of surface analytical tools, it is important to appreciate the questions that require answering. A single technique cannot generally provide a complete picture of the surface characteristics. In this section, a description of the some of the most commonly used techniques is provided with reference to more detailed and extensive reviews. It is important to note that the techniques fall into two classes, those that operate inside a vacuum and those that can directly probe the biomaterial–water interface. While it is obviously preferable to use those techniques that can perform under ambient conditions, in general these techniques are either not as informative or not as surface sensitive as the vacuum techniques. For this reason, the vacuum techniques are commonly utilized to provide a detailed characterization, but in doing so it must always be under the assumption that the surface is the same in vacuum as it is under water. This is a rather large assumption, particularly if the material is able to reorganize itself relatively easily. The surface energy change following immersion in water can be rather large, as indicated above, and in mixed biomaterial phases components that are absent at the surface in vacuum may dominate when the material is immersed in an aqueous environment. There is evidence for this kind of surface reorientation in the contact angle hysteresis of water on some polymers.

Hysteresis is the difference in contact angle between a water contact line advancing or receding across a surface. For some polymers, the advancing angle is high and the receding angle is low, indicating that at the polymer–air interface the polymer is hydrophobic and at the polymer–water interface it is hydrophilic. As long as the surface is flat and homogenous, this is evidence of surface reorganization. For some materials it is possible to reduce the rate of reorganization by cooling. This is typically achieved by hydrating the sample in air, and then freezing the sample in liquid nitrogen prior to entry into the vacuum chamber. The sample needs to be held at low temperature while the ice on the surface sublimes and then vacuum techniques

Table 3. Surface Analysis Techniques Used in the Characterization of Biomaterials

Technique	Sampling Depth/Height (Spatial Resolution)	Information Obtained	References
Ultrahigh Vacuum			
Static secondary ion mass spectrometry	<5 nm (500 Å)	Chemical Spectroscopy and Imaging	49,50
X-ray photoelectron spectroscopy	2–10 nm (5 μm)	Elemental, chemical Spectroscopy and Imaging	51
Ambient Techniques			
Attenuated total reflectance Fourier transform Infrared (ATR/FTIR) spectrometry	> 100 nm (1 μm)	Chemical	52
Contact angle measurement	<1 nm (1 mm)	Surface free energy, wettability	53
Atomic force microscopy (AFM) Imaging	Atomic = 20 μm	Topography, coverage, atomic structure	54
Surface force measurement	Å = μm	Chemical, conformational, structural	
Ellipsometry	Å = 300 nm	Layer thickness, adsorption kinetics	55,56
Streaming potential measurements/electroosmosis	Not applicable	Electrokinetics	57
Surface plasmon resonance	~300 nm	Adsorbed mass and adsorption kinetics	58

can be applied to the surface, which should not have reorganized from the hydrated state (59).

Vacuum Techniques

Traditional surface analysis techniques are usually based on ultrahigh vacuum instrumentation. One of the reasons for this was that much of the initial interest in the field was concentrated upon extremely clean and often highly reactive surfaces. To maintain the surface in this state during the analysis it is important to prevent undesired gas or vapor molecules sticking to the surface and changing its characteristics. This is only possible in ultrahigh vacuum ($<10^{-9}$ mbar or so). A second important reason is that to study just the surface and eliminate contributions from the bulk of the material it is necessary to use probe species that strongly interact with matter, such as ions and electrons. These cannot penetrate through more than a few atomic layers, and hence provide highly surface sensitive information. However, the detection of such species normally requires that they travel a considerable distance from the surface. At atmospheric pressure the average distance traveled prior to interaction with gaseous species is too short for the detection systems to work. A vacuum of, typically, 10^{-7} mbar or better is required for the techniques described here to operate. In addition, many of the components necessary for the production and detection of probe species can only be operated in vacuum; at atmospheric pressure, they would be irreparably damaged. We will now briefly describe some of the key surface analysis techniques used in the characterization of biomaterials. More detailed information and discussion on the interpretation of these techniques can be found in books by Vickerman (54) and Watts (51).

X-Ray Photoelectron Spectroscopy

During X-ray photoelectron spectroscopy (XPS) analysis, the sample is illuminated with X rays of a particular energy that causes electrons to be emitted from the sample. This phenomenon is called the photoelectric effect and it is generally found that an X-ray photon imparts all of its energy to a single electron during the process. Since the electrons are bound in orbitals of well-defined energy (binding energy) that are characteristic of the material, the outgoing electrons have a kinetic energy that is essentially the difference between the photon and binding energies. For core level electrons, this binding energy is characteristic of the nucleus to which the electron is closely bound. Only those electrons that are generated close to the surface can escape from the sample without loss of energy due to inelastic collisions with other atoms. Thus, by analyzing the number of electrons emitted from the surface as a function of electron kinetic energy it becomes possible to identify the elements present on, or near to, the material surface. As long as the photon energy is significantly larger than the binding energy of the electron, the probability of generating a photoelectron from a core level is independent of the chemical situation of the element. This means that it is possible not only to identify the elements present but, with appropriate sensitivity factors, quantify the relative amounts of each element.

The chemical situation of the element may, however, have an influence on the binding energy of the core level electron. Chemical bonds to different elements may cause some charge to accumulate on the element of interest, this will directly affect the binding energy of the core electrons. This change in the binding energy is termed the "chemical shift" by analogy to nuclear magnetic resonance (NMR) spectroscopy. The appearance of a chemical shift is extremely useful in surface analysis as it can provide information on how the various elements in the surface are bonded to each other. Where the same element is in a range of chemical environments it is often possible to deduce the fraction of atoms in each environment by careful curve fitting of the spectrum. An example of these types of chemical shifts is illustrated in Fig. 2. In this case, there are chemical shifts evident in a high resolution carbon 1*s*

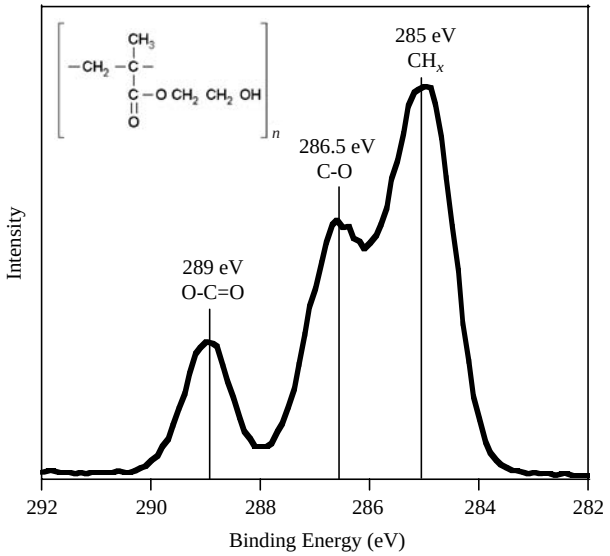


Figure 2. The XPS high resolution C1*s* spectrum of a HEMA contact lens. The figure illustrates the various chemical shifts associated with the chemical bonds in the polymer chain.

(C1*s*) spectrum due to the various chemical bonds present in poly(hydroxyethylmethacrylate), pHEMA, a common soft contact lens material.

The surface sensitivity of XPS is dependent on the angle at which electrons are emitted from the sample. For smooth, flat samples it is possible to enhance the surface sensitivity by analyzing electrons which are emitted at a grazing incidence from the sample. By collecting at a number of different angles it is possible to obtain information on the depth distribution of components close to the material surface. This depth profiling capability is particularly important when thin films and coatings of < 10 nm thickness are being studied.

The XPS has been utilized to chemically characterize biomaterials in four principal areas: identification and characterization of the surface chemistry of bulk polymers; characterization of surface specific modifications; characterization of coatings; detection of biomolecules. Factors that are commonly investigated using XPS include: surface oxidation and reorientation of polymer segments; surface segregation (blooming) of plasticizers, additives, and low molecular weight fragments; adventitious contaminants such as silicones and protein adsorption.

Secondary Ion Mass Spectrometry

The impact of a high kinetic energy (typically 1–100 keV) ion, atom, or molecule causes material to be sputtered from a surface. The origin of the vast majority of ejected species is from the topmost layer of the sample. Therefore analysis of the sputtered fragments can provide information on the composition of the material surface. A small proportion of the ejected atomic and molecular fragments are ionized, these are called the secondary ions. Secondary ion mass spectrometry (SIMS) is the application of mass spectrometry to the secondary ions. Note that SIMS is an ablative, destructive technique and this can be used to advantage in generating a depth profile of layered surfaces. The use of SIMS for depth profiling is termed "dynamic" SIMS and is most often employed in the study of layered materials in which the elemental composition of the layers is of interest, for example, doping levels in semiconductors. It is not possible to obtain more detailed chemical information using dynamic SIMS because of the damage induced by the high energy primary ions. In contrast, "static" SIMS employs a low density, low dose ion bombardment such that the probability of two ion impacts occurring at the same place on the sample is negligibly small ($< 10^{13}$ ions $\cdot$ cm^{-2}). The mass spectrum then contains information that is characteristic of the undamaged surface. This information is particularly useful in the analysis of organic materials, when the normal rules of organic mass spectrometry can be applied to the interpretation of SIMS data. Most modern static SIMS instruments are based on time-of-flight mass analyzers (TOF–SIMS), which have a far greater combined sensitivity and mass resolution than quadrupole or magnetic sector detectors. The probability of ion generation is influenced by a daunting range of factors and thus SIMS is regarded as a nonquantitative technique. However, it is commonly found that in a range of similar materials the characteristic ion intensities are approximately proportional to the concentration of species from which they are generated. With a suitable set of calibration data it is then possible to use SIMS in a quantitative manner. The application of TOF–SIMS in the analysis of biomaterials and biological interfaces has historically revolved around the characterization of polymeric interfaces. This has included the study of degradation pathways for biodegradable polymers, the monitoring of coating chemistries, detection of surface contamination and surface chemical characterization of copolymer systems. The surface sensitivity of TOF–SIMS has lead to its application in the detection and identification of biomolecules adsorbed at interfaces. The process is not without its problems though as the largest ions detected from any protein are the immonium ions ($^{+}NH_2{=}CHR$) from each amino acid (MW<200). As a result of this fragmentation, the identification of proteins is often more like a jigsaw puzzle, where the amino acid fragments have to be pieced together using pattern recognition or multivariate analysis techniques, to identify and quantify the parent molecules (60). These types of statistical analysis are being increasingly used to analyze, compare and reconstruct data collected in TOF–SIMS.

In addition to spectroscopy, TOF–SIMS can be used in an imaging mode to chemically map the surface of a material. There is always a trade off between high special resolution and high mass resolution, but with the advent of liquid metal ion sources (e.g., Ga^+ and In^+), systems are typically capable of spatial resolution of < 10 μm, while retaining atomic mass resolution. As a result there is increasing application of TOF–SIMS for the chemical imaging of a range of biomaterial surfaces. Significantly, developments in ion sources have shown that polyatomic (e.g., Au_3) and cluster ion (C_{60}) sources can significantly improve the molecular ion yield of both biological and polymeric materials. With the development of integrated freeze hydration stages for sample preparation, this has

lead to increased activity in the application of TOF–SIMS in the analysis of cell membranes and other hydrated biological systems (49).

AMBIENT TECHNIQUES

Atomic Force Microscopy

Atomic force microscopy (AFM) can be utilized to characterize surfaces via either an imaging or a spectroscopic mode. There are two common methods of imaging utilized in AFM, contact, and tapping mode. Both can be performed in either air or liquid, a factor that makes AFM particularly attractive in biomaterial research. In contact mode, the tip is scraped across the surface, while in tapping mode, the tip is in intermittent contact with the surface and as such, there is limited substrate disturbance. As a result, tapping mode AFM is more common for the characterization of biomaterials and biological surfaces. Common applications of AFM to biomaterial surfaces include: surface topography and coating continuity assessment, measurement and monitoring of coating thickness (see Fig. 3) and phase imaging. The last application is an extension of tapping mode imaging that gives nanometer-scale two-dimensional (2D) information about surface structure. It can be used to locate and characterize the distribution of discrete phases within polymer blends such as polyurethane–urea, and a range of other polymers of biological importance. There is also considerable interest in using AFM to image the surfaces of cells and biomolecules on surfaces (61,62).

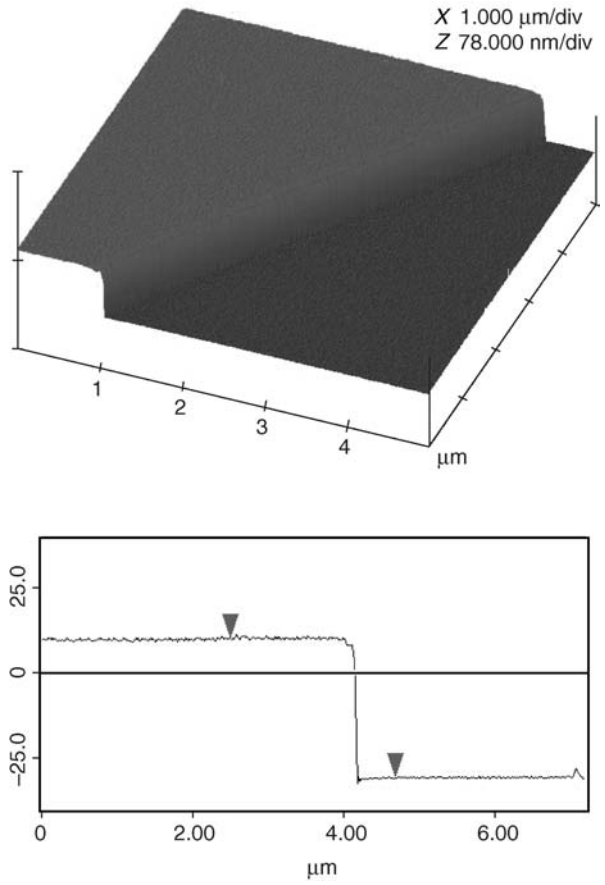


Figure 3. The AFM tapping mode image of a 40 nm step between coated and uncoated regions on a silicon wafer. The step is produced by masking a section of the sample surface prior to plasma polymerization. Once the coating is deposited, the mask is removed and surface imaged. This enables the plasma polymer thickness to be measured with nm resolution. (Image courtesy of Dr. P.G. Hartley, CSIRO Molecular Science, Australia.)

Surface Force Measurements

The forces that act between particles and surfaces determine a wide range of interactions. They control the stability of dispersions and emulsions; determine the adhesion of colloids onto surfaces and the adsorption properties of proteins and cells at surfaces. Surface force measurements are used in variety of industries and increasingly there is interest in the application of surface force analysis to biomaterials. with the aim of characterizing protein–protein, cell–protein, protein–surface and cell–surface interactions.

There are a number of techniques that can be used to measure force interactions between surfaces. They are divided into two classes based on the method used to determine surface separation. Absolute surface separation can only be determined using an interferometric technique such as the surface force apparatus (SFA). These techniques are limited by the need for a transparent substrate and specific geometric configuration. In response, noninterferometric techniques have been developed that can employ a wider variety of substrates, and that rely on indirect determination of the surface separation rather than interferometry. One of these is AFM surface force measurement.

Force measurements can be made with an AFM using both a bare tip and a tip modified with a probe particle. If the results from these measurements are to be quantitative, knowledge of the radius of curvature for the probe or tip is critical. While it is possible to measure the nominal radius of a bare AFM tip, the indirect nature of the measurement adds to the error associated with the resulting force calculations (63). The Derjaguin approximation is only valid when the radius is much larger than the surface separation, which is not necessarily the case if an unmodified tip is used. If a colloid probe is utilized, the radius of curvature can be measured easily and accurately using scanning electron microscope (SEM) or optical microscopy.

The zero surface separation for noninterferometric techniques are set at the hard wall encountered when the two surfaces are forced together. Termed the constant compliance regime, this is the region of the force curve in which the displacement of the colloid probe is linear with respect to the surface motion. This assumption of hard wall contact is an inherent limitation of these techniques, particularly if the surface deforms or compresses under pressure, as is often the case with polymer surfaces. The compression of a polymer layer can have a number of effects on the force curve. In the first instance, compression of a dense polymer layer can result in a compliance line mimicking hard wall contact, with a layer of compressed polymer between the probe and the substrate. If the polymer is less densely

packed, the probe may displace the material, squeezing the polymer out of the gap between the probe and the surface. This results in discontinuities in the compliance region as the force increases and the probe pushes through the polymer layers (63). As a result, conclusions about absolute layer thickness cannot be inferred from this data.

While XPS is able to characterize the chemistry of a surface in the dry state, surface force measurements are well suited to characterizing the intermolecular forces and stability in a variety of environments. A number of studies have investigated the surface force characteristics of rfgd films, grafted polymer layers, and adsorbed protein films in a variety of media (64). Other studies have used surface force measurements to characterize the interactions of polymer modified surfaces in an attempt to elucidate parameters that control the structure of the polymer layer (64). Surface force techniques have been used to investigate the effects of molecular weight, ionic strength, charge density, and polymer concentration on the interaction forces of adsorbed and grafted polymers layers. Increasingly, strategies to eliminate protein adsorption are based on the characterization and modification of the surfaces and thus the interaction forces that govern protein adsorption (65). A number of theoretical studies have also used surface force interactions as design parameters when modeling polymer coatings capable of resisting protein adsorption (66).

In addition to these standard modes of operation, by chemically modifying the AFM cantilever it is possible to map specific interactions between the tip and the surface. Depending on the type of modification made to the cantilever, a range of interactions can be investigated. With a cantilever modified with a receptor specific integrin, dynamic force spectroscopy can be used to identify and map receptor sites on a cell surface (67). By modifying the cantilever with specific chemical functional groups, differences in frictional properties and the distribution of different phases can be probed where there is no topographical variation (68).

Optical Techniques

The refractive index close to an interface can be measured by a number of optical techniques. The two most commonly employed for this purpose are surface plasmon resonance (SPR) (58) and ellipsometry (55,56). Since proteins have a higher refractive index ($\sim$1.55) than water ($\sim$1.33) it is possible to monitor the amount of protein adsorption at an interface through a measurement of the refractive index and thickness of the adsorbed layer. These techniques have found utility in a wide range of areas relevant to biomaterials research, such as the study of protein adsorption to biomaterials. ligand–receptor interactions and the dissolution and swelling of polymers. The advantage over traditional approaches such as enzyme linked immunosorbent assay (ELISA), fluorescent labeling or radiolabeling is that the proteins under study do not have to be chemically altered in any way.

The disadavantages of these techniques are that the substrate must be flat and conform to a number of optical requirements for the techniques. Additionally, these techniques cannot directly distinguish between different proteins, since all proteins have roughly equivalent optical densities. To determine the identity of proteins adsorbed from a mixed solution it is necessary to subsequently expose the surface to antibodies specific to each protein of interest. Binding of the antibody can be monitored as an increase in the adsorbed layer thickness, however, this approach is difficult to employ quantitatively as there may be nonspecific and competitive adsorption of the antibody as well as a limited availability of binding sites on the adsorbed target protein.

Surface Plasmon Resonance

A surface plasmon is a collective oscillation of electrons that can be excited in certain metals such as silver and gold. The frequency of this oscillation depends on the refractive index of the dielectric material close to the metal interface. If the metal is a thin film it is possible to excite surface plasmons by reflecting light of a wavelength greater than the thickness of the metal from the reverse side of the film. The ability to cause this excitation depends on the wavelength of light, the refractive index of the material through which the light travels (which remains constant in this geometry) and the angle of reflection. When the conditions are correct, light is absorbed. In the usual set up for surface plasmon resonance (SPR) instruments the light undergoes total internal reflection at the interface and the angle at which light is absorbed is monitored. If there is a change in the refractive index close to the interface then a corresponding shift in the angle at which light is absorbed can be followed. The sensitivity of SPR decreases exponentially in distance from the surface with a decay length of the order of the wavelength of the light. If the layer to be analysed is significantly smaller than the decay length, which is usual for protein adsorption, then it can be assumed that any change in refractive index is proportional to the mass of adsorbed protein.

Ellipsometry

When light is reflected at an oblique angle from a planar surface it commonly undergoes a change in polarization. By analyzing these changes, it is possible to infer both the optical properties and thickness of thin films on the surface. To obtain the most complete characterization, a large number of different wavelengths of light or different angles of incidence must be studied. The measured data is then compared to the expected polarization changes calculated from a model, and parameters in the model changed (such as thickness or refractive index) to find a fit between the two. The sensitivity of ellipsometry is comparable to SPR ($\sim 0.01/\mathrm{g{\cdot}cm^2}$), however, it is able to analyze comparatively thick layers of material.

CONCLUSION

While materials selection for most biomedical devices needs to be based upon bulk properties, in this article we have provided a broad overview of the basic properties of surfaces, and introduced some of the reasons why the surface properties may significantly 3 influence the efficacy of biomaterials

and biomedical devices. Surface modification aims to tailor the surface characteristics of a material for a specific application without detrimentally affecting the bulk properties. Throughout this article we have shown how a range of physical, chemical, and biological modifications can be made to surfaces and used to manipulate surface characteristics. Finally, we discussed a range of highly sensitive surface analytical methods that can be utilized to investigate both the nature of an interface and its interactions with biological environments. As is always the case with review articles of this type, it is impossible to give detailed accounts of all of the material being discussed. We have included a range of references (*Reading List*) to aid the reader in further developing their understanding of each of the specific concepts and techniques. Additionally, we have included a list of more general references that cover many of the fundamental concepts discussed within this article.

BIBLIOGRAPHY

Cited References

1. Mittal KL, editor. Contact Angle, Wettability and Adhesion. Utrecht: VSP; 1993.
2. Berry CC, Campbell G, Spadiccino A, Robertson M, Curtis ASG. The influence of microscale topography on fibroblast attachment and motility. Biomaterials 2004;25(26):5781–5788.
3. Andrade JD. Principles of Protein Adsorption. In: Andrade JD, editor. Surfaces and Interfacial Aspects of Biomedical Polymers. Vol. 2: Protein Adsorption. New York: Plenum Press; 1985.
4. Underwood PA, Steele JG. Practical limitations of estimation of protein adsorption to polymer surfaces. J Immunol Methods 1991;142(1):83–94.
5. Leduc CA, Vroman L, Leonard EF. A Mathematical-Model for the Vroman Effect. Ind Eng Chem Res 1995;34(10):3488–3495.
6. Hubbell JA. Bioactive Biomaterials. Curr Opin Biotechnol 1999;10:123–129.
7. Bray D. Cell Movements:From Molecules to Motility. 2nd ed. New York: Garland; 2001. p 372.
8. Chu PK, Chen JY, Wang LP, Huang N. Plasma-surface modification of biomaterials. Mat Sci Eng R 2002;36(5-6): 143–206.
9. Favia P, d'Agostino R. Plasma Treatments and Plasma Depositions of Polymers for Biomedical Applications. Surf Coat Tech 1998;98:1102–1106.
10. Aronsson BO, Lausmaa J, Kasemo B. Glow discharge plasma treatment for surface cleaning and modification of metallic biomaterials. J Biomed Mater Res 1997;35(1):49–73.
11. Mandl S, Rauschenbach B. Plasma immersion ion implantation. New technology for homogeneous modification of the surface of medical implants of complex shapes. Biomed Tech 2000;45(7–8):193–198.
12. Bilek MMM, McKenzie DR, Tarrant RN, Lim SHM, McCulloch DG. Plasma-based ion implantation utilising a cathodic arc plasma. Surf Coat Tech 2002;156(1–3):136–142.
13. Shin GH, Lee YH, Lee JS, Kim YS, Choi WS, Park HJ. Preparation of plastic and biopolymer multilayer films by plasma source ion implantation. J Agric Food Chem 2002;50(16):4608–4614.
14. McPherson TB, Shim HS, Park K. Grafting of PEO to glass, nitinol, and pyrolytic carbon surfaces by gamma irradiation. J Biomed Mater Res 1997;38(4):289–302.
15. Benson RS. Use of radiation in biomaterials science. Nucl Instrum Meth B 2002;191:752–757.
16. Welle A, Gottwald E. UV-based patterning of polymeric substrates for cell culture applications. Biomed Microdevices 2002;4(1):33–41.
17. Zhang F, Kang ET, Neoh KG, Wang P, Tan KL. Surface modification of stainless steel by grafting of poly(ethylene glycol) for reduction in protein adsorption. Biomaterials 2001;22(12):1541–1548.
18. Krupa D, Baszkiewicz J, Kozubowski J, Barcz A, Sobczak J, Bilinski A, Rajchel B. The influence of calcium and/orphosphorus ion implantation on the structure and corrosion resistance of titanium. Vacuum 2001;63(4): 715–719.
19. Braceras I, Alava JI, Onate JI, Brizuela M, Garcia-Luis A, Garagorri N, Viviente JL, de Maeztu MA. Improved osseointegration in ion implantationtreated dental implants. Surf Coat Tech 2002;158:28–32.
20. Cui FZ, Luo ZS. Biomaterials modification by ion-beam processing. Surf Coat Tech 1999;112(1–3):278–285.
21. Bambauer R, Mestres P, Schiel R, Schneidewind JM, Latza R, Bambauer S, Sioshansi P. Surface treated catheters with ion beam based process for blood access. Ther Apher 2000;4(5):342–347.
22. Li DJ, Zhao J, Gu HQ. Hemocompatibility of DLC coatings synthesized by ion beam assisted deposition. Sci China Ser E-Technol Sci 2001;44(4):427–431.
23. Kingshott P, Griesser HJ. Surfaces that resist bioadhesion. Curr Opin Solid St M 1999;4:403–412.
24. Bures P, Huang YB, Oral E, Peppas NA. Surface modifications and molecular imprinting of polymers in medical and pharmaceutical applications. J Control Release 2001;72(1–3):25–33.
25. Kato K, Uchida E, Kang ET, Uyama Y,Ikada Y. Polymer surface with graft chains. Prog Polym Sci 2003;28(2):209–259.
26. Cai KY, Lin SB, Yao KD. Advances in research on surface engineering of biomaterials for tissue engineering. Prog Chem 2001;13(1):56–64.
27. Sakiyama-Elbert SE, Hubbell JA. Functional biomaterials: Design of novel biomaterials. Ann Rev Mater Res 2001; 31:183–201.
28. Massia SP, Stark J. Immobilized RGD peptides on surface-grafted dextran promote biospecific cell attachment. J Biomed Mater Res 2001;56(3):390–399.
29. Whitesides GM, Ostuni E, Takayama S, Jiang X, Ingber DE. Soft Lithography in Biology and Biochemistry. Annu Rev Biomed Eng 2001;3:335–373.
30. Beck AJ, Jones FR, Short RD. Plasma copolymerization as a specific route to the fabrication of new surfaces with controlled amounts of specific chemical functionality. Polymer 1996;37:5537–5539.
31. Shen MC, Martinson L, Wagner MS, Castner DG, Ratner BD, Horbett TA. PEO-like plasma polymerized tetraglyme surface interactions with leukocytes and proteins: in vitro and in vivo studies. J Biomater Sci Polym Ed 2002;13(4):367–390.
32. Pan YV, Wesley RA, Luginbuhl R, Denton DD, Ratner BD. Plasma polymerized *N*-isopropylacrylamide: synthesis and characterization of a smart thermally responsive coating. Biomacromolecules 2001;2(1):32–36.
33. Fraser S, Short RD, Barton D, Bradley JW. A multi-technique investigation of the pulsed plasma and plasma polymers of acrylic acid: Millisecond pulse regime. J Phys Chem B 2002;106(22):5596–5603.
34. Han LCM, Timmons RB. Pulsed-plasma polymerization of 1-vinyl-2- pyrrolidone: Synthesis of a linear polymer. J Polym Sci Pol Chem 1998;36(17):3121–3129.

35. Hanawa T. In vivo metallic biomaterials and surface modification. Mat Sci Eng A-Struct 1999;267(2):260–266.
36. Guzman L, Celva R, Miotello A, Voltolini E, Ferrari F, Adami M. Polymer surface modification by ion implantation and reactive deposition of transparent films. Surf Coat Tech 1998;104:375–379.
37. Cui FZ, Luo QL, Feng J. Highly adhesive hydroxyapatite coatings on titanium alloy formed by ion beam assisted deposition. J Mater Sci Mater M 1997;8:403–405.
38. Howlett CR, Zreiqat H, Wu Y, McFall DW, McKenzie DR. Effect of ion modification of commonly used orthopedic materials on the attachment of human bone-derived cells. J Biomed Mater Res 1999;45(4):345–354.
39. Sioshansi P, Tobin EJ. Surface treatment of biomaterials by ion beam processes. Surf Coat Tech 1996;83(1–3):175–182.
40. Davenas J, Thevenard P, Philippe F, Arnaud MN. Surface implantation treatments to prevent infection complications in short term devices. Biomol Eng 2002;19(2–6):263–268.
41. Leng YX, Chen JY, Zeng ZM, Tian XB, Yang P, Huang N, Zhou ZR, Chu PK. Properties of titanium oxide biomaterials synthesized by titanium plasma immersion ion implantation and reactive ion oxidation. Thin Solid Films 2000;377:573–577.
42. Ostuni E, Chapman RG, Holmlin RE, Takayama S, Whitesides GM. A survey of structure-property relationships of surfaces that resist the adsorption of protein. Langmuir 2001;17:5605–5620.
43. Textor M, Ruiz L, Hofer R, Rossi A, Feldman K, Hahner G, Spencer ND. Structural chemistry of self-assembled monolayers of octadecylphosphoric acid on tantalum oxide surfaces. Langmuir 2000;16(7):3257–3271.
44. Chandy T, Das GS, Wilson RF, Rao GHR. Use of plasma glow for surface-engineering biomolecules to enhance bloodcompatibility of Dacron and PTFE vascular prosthesis. Biomaterials 2000;21(7):699–712.
45. Peluso P, Wilson DS, Do D, Tran H, Venkatasubbaiah M, Quincy D, Heidecker B, Poindexter K, Tolani N, Phelan M, Witte K, Jung LS, Wagner P, Nock S. Optimising antibody Immobilization strategies for the construction of protein microarrays. Anal Biochem 2003;312:113–124.
46. Sackmann E, Tanaka M. Supported membranes on soft polymer cushions: fabrication, characterization and applications. Trends Biotechnol 2000;18:58–64.
47. Plant AL. Supported hybrid bilayer membranes as rugged cell membrane mimics. Langmuir 1999;15(15):5128–5135.
48. Krishna G, Schulte J, Cornell BA, Pace RJ, Osman PD. Tethered bilayer membranes containing ionic reservoirs: Selectivity and conductance. Langmuir 2003;19(6):2294–2305.
49. Winograd N. Prospects or imaging TOF-SIMS: from fundamentals to biotechnology. Appl Surf Sci 2003;203:13–19.
50. Castner DG, Ratner BD. Biomedical surface science: Foundations to frontiers. Surface Sci 2002;500(1–3):28–60.
51. Watts JF, Wolstenholme J. An Introduction to Surface Analysis by XPS and AES. Chichester: John Wiley & Sons; 2003.
52. Chittur K. FTIR/ATR for protein adsorption to biomaterials surfaces. Biomaterials 1998;19:357–369.
53. Adamson AP, Gast AW. Physical Chemistry of Surfaces. New York: John Wiley & Sons; 1997.
54. Vickerman JC, editor. Surface analysis: The Principal Techniques. Chichester: John Wiley & Sons; 1997, p. 457.
55. Elwing H. Protein adsorption and ellipsometry in biomaterials research. Biomaterials 1998;19:397–406.
56. Arwin H. Ellipsometry on thin organic layers of biological interest: characterization and applications. Thin Solid Films 2000;377:48–56.
57. Hunter RJ. Zeta Potential in Colloid Science. London: Academic Press; 1988. p 67.
58. Green RJ, Frazier RA, Shakesheff KM, Davies MC, Roberts CJ, Tendler SJB. Surface plasmon resonance analysis of dynamic biological interactions with biomaterials. Biomaterials 2000;21(18):1823–1835.
59. Lewis KB, Ratner BD. Observation of surface restructuring of polymers using ESCA. J Colloid Interface Sci 1993;159:77–85.
60. Wagner MS, Castner DG. Characterization of adsorbed protein films by time of flight secondary ion mass spectrometry (ToF-SIMS) in conjunction with prinicpal component analysis (PCA). Langmuir 2001;17:4649–4660.
61. Boonaert C, Rouxhet P, Dufrene Y. Surface propaerties of microbial cells probed at the nanometre scale with atomic force microscopy. Surf Interface Anal 2000;30:32–35.
62. Fritz M, Radmacher M, Cleveland JP, Allersma MW, Stewart RJ, Gieselmann R, Janmey P, Schmidt CF, Hansma PK. Imaging globular and filamentous proteins in physiological buffer solutions with tapping mode atomic force microscopy. Langmuir 1995;11:3529–3535.
63. Hartley PG. Farinato R, Dubin P, editors. Measurement of Colloidal Interactions Using the Atomic Force Microscope, in Colloid-Polymer Interactions: From Fundamentals to Practice. New York: John Wiley & Sons; 1999.
64. Hartle PG, McArthur SL, McLean KM, Griesser HJ. Physicochemical properties of polysaccharide coatings based on grafted multilayer assemblies. Langmuir 2002;18(7):2483–2494.
65. Leckband D, Sheth S, Halperin A. Grafted poly(ethylene oxide) brushes as nonfouling surface coatings. J Biomater Sci Polym Ed 1999;10(10):1125–47.
66. Halperin A. Polymer brushes that resist adsorption of model proteins: Design parameters. Langmuir 1999;15:2525–2533.
67. Evans E. Energy landscapes of biomolecular adhesion and receptor anchoring at interfaces explored with dynamic force microscopy. Faraday Discuss 1998;111:1–16.
68. Sun S, Chong KS, Leggett GJ. Nanoscale molecular patterns fabricated by using scanning near-field optical lithography. J Am Chem Soc 2002;124(11):2414–2415.

Reading List

Andrade JD, editor. Surfaces and Interfacial Aspects of Biomedical Polymers. New York: Plenum Press; 1985.
Malmsten M, editor. Biopolymers at Interfaces. New York: Marcel Dekker; 1998/2004.
Castner DG, Ratner BD. Biomedical surface science: Foundations to frontiers. Surface Sci 2002;500(1–3):28–60.
Adamson AW, Gast AP. Physical Chemistry of Surfaces. New York: John Wiley & Sons; 1997.

See also BIOCOMPATIBILITY OF MATERIALS; BIOSURFACE ENGINEERING; MICROSCOPY, SCANNING TUNNELING.

BIOMATERIALS, TESTING AND STRUCTURAL PROPERTIES OF

DONGLU SHI
University of Cincinnati
XUEJUN WEN
Clemson University
Clemson, South Carolina

INTRODUCTION

Tissue transplantation and synthetic devices have been utilized in order to substitute the function of lost

or damaged hard tissue, such as bone and tooth. Tissue transplants can be autologous, allogeneic, or xenogeneic. However, the use of autologous tissue involves additional surgery and donor site morbidity while the use of allogeneic or xenogeneic tissue involves the risks of immune rejection and disease transmission. Therefore, synthetic hard tissue implants are very necessary. Metals, ceramics, composites, and even polymers are investigated as candidates for the hard tissue replacements. For heavy loaded applications, such as hip prostheses, metals (e.g., Ti–alloys, Co–Cr), and strong inert ceramics (e.g., alumina, zirconia) are extensively studied. Unfortunately, various problems related to both the metallic materials and the bioinert ceramics, for example, corrosion, elastic modulus mismatch (stress concentration and shielding), and bioinertness (only physical connection with host) with metals, and brittle, elastic modulus mismatch, and bioinertness with bioinert ceramics. For these reasons, bioceramics are showing very promising results in the high bioactivity and the formation of interfacial chemical bond with host tissue, which was called osseointegration (1). So far, several bioactive ceramics have been proposed for hard tissue replacements, hydroxyapatide (HA) and bioactive glasses are the most acceptable materials for hard tissue applications (2). The advantages of bioceramics over inert ceramics and metals allow for developing better hard tissue replacements with the characteristics of bioactive and elastic modulus more close to that of bone (2). On the other hand, the mechanical properties of bioceramics are fairly poor when compared with their replaced natural hard tissues. The poor mechanical properties, especially inside the body aqueous environments, limit their applications to only small, unloaded, and low loaded implants, powders, coatings, composites, porous scaffolds for tissue engineering, and so on. Bioceramic coatings and porous scaffold are showing the most promising results for the future hard tissue replacements (3–5). There are various methods developed to produce HA coatings (3–5). Among these techniques, plasma spraying has widely been used. However, this method is not applicable for deposition HA films onto a porous substrate.

In order to obtain a bone substitute possessing both desirable mechanical properties and bioactivity, two major deposition routes in coating the bioactive HA on a highly porous alumina substrate with the similar range of tensile and compressive strength as natural bone were developed. Coated reticulated bioactive substrates can provide the needed mechanical strength for the replacement of the bear-loading functions. The first one is a suspension method in which the ceramic substrates are first coated with a suspension containing the HA powder followed by a sintering with an appropriate time–temperature cycle to densify the HA coating. The second one is a synthesis route or called thermal deposition.

The techniques of coating uniformly thin layers of bioactive HA onto highly porous alumina substrate, the structural properties, especially the interfaces between the coating and the substrate, and the bioactive behavior of the coated substrate in the simulated body fluid (SBF) will be presented in this article.

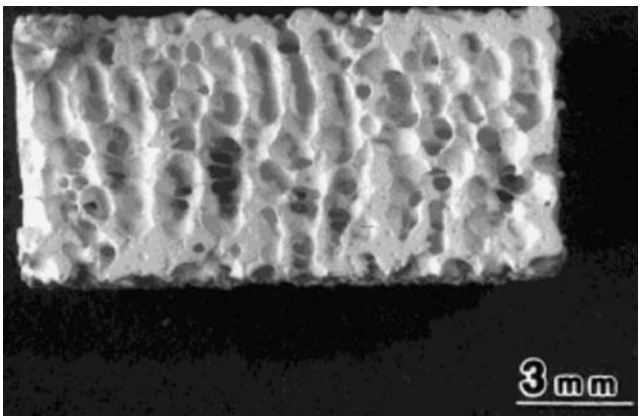


Figure 1. Cross-section of reticulated alumina substrate, showing the interconnected porosity.

MATERIALS AND METHODS

Reticulated alumina Al_2O_3 was used as the substrate materials. Figure 1 is the gross morphology of the substrate; and Fig. 2 is the cross-section showing the interconnected pores. The average size of the pores is 500 μm, which are large enough to allow the ingrowth of bone tissue. Substrates were cleaned using ultrasonic cleaner in acetone and dried at 100 °C before applying the coating.

Suspension Method

The coating suspension was made up of finely milled ceramic powders, an organic solvent, and a binder. The binder was used to prevent the precipitation of particles and to provide bonding strength to the coating after drying. One important property of the suspension is its viscosity. Specifically, when the porous substrate was immersed in the coating suspension, the suspension must be fluid enough to enter, fill, and uniformly coat the substrate skeleton. Low viscosity could result in undesirable thin films while highly viscous slurry would block the pore, thus impairing the interconnectivity of the pores. The viscosity

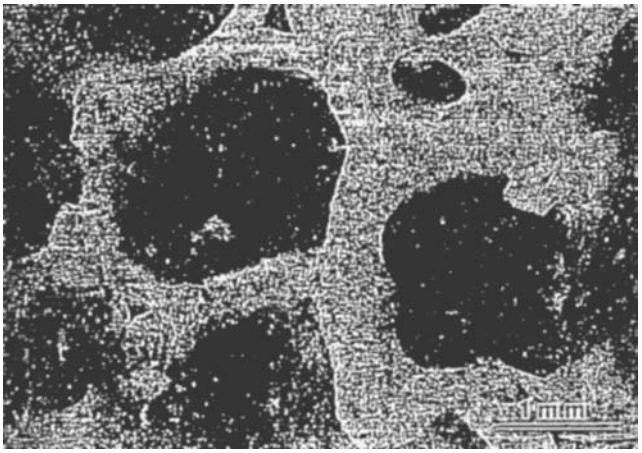


Figure 2. Cross-section of reticulated alumina substrate, showing the interconnected porosity.

Table 1. Sample Group Assignment

Sample	Source and Treatment	SSA, $m^2 \cdot g^{-1}$
HA	Commercial	63.02
HA600	HA heated at 600 °C, 30 min	40.92
HA900	HA heated at 900 °C, 30 min	17.45
SHA700	Synthesized HA heated at 700 °C, 4 h	15.19
SHA800	Synthesized HA heated at 800 °C, 4 h	1.49
SHA900	Synthesized HA heated at 900 °C, 4 h	1.23
No. 1	HA heated at 700 °C, 3.5 h, particle size: 40–100 mesh	24.38
No. 2	HA heated at 700 °C, 3.5 h, particle size: 100–200 mesh	25.70
No. 3	HA heated at 700 °C, 3.5 h, particle size: $>$ 200 mesh	27.02

was controlled by the relative amount of solvent, binder, and particles. The porous substrates were subsequently immersed into the mixture. After the porous substrates were completely infiltrated, they were spun briefly in a high speed centrifuge for removal of excess solution. Coated substrates were dried in an oven at 100 °C. The dried specimens were heated in air to 400 °C for 1 h to burn out the organic binder from suspension. During the burnout process, a slow and controlled heating rate was necessary to avoid bubbling in the coating. Then the samples were subject to sintering at different temperatures. Two types of suspension were developed for coating. One was prepared by suspending HA particles (300 mesh) (Chemat Technology, Inc.) in an organic binder–solvent system without glass sintering aid, glass frits. The other was prepared by partially substitution of HA by sintering aid, glass frits (65%), which have good adhesion to the Al_2O_3 substrate and a weak reaction with HA during firing. The glass frits used in this work were borosilicate glasses containing $\sim$75% of a mixture of SiO_2 and B_2O_3, and 20 wt% alkali metal oxides. After melting, the glass was quenched in water and ground in a ball mill into a glass frit of the desired particle size (325 mesh).

Thermal Deposition

Mixing calcium 2-ethyl hexanoate with bis(2-ethyhexyl) phosphite stoichiometrically in ethanol. The viscosity of the solution was controlled by the quantity of ethanol added. The mixture was stirred for 2 days at room temperature. Then mixture was used to coat porous substrates. The coating method used is the same as used in suspension method described earlier. Coated substrates were air-dried and calcinated up to 1000 °C at a heating rate of 2 °C/min. Then the samples were subject to sintering at different temperatures. For phase analysis purpose, the HA was prepared in the powder form as well through same procedures. Briefly, the solution was open to the air and stirred to vaporize the solvent in chemical hood. Finally, a highly viscous, translucent mixture was obtained and then subject to calcinations at desirable temperatures.

Mechanical strength measurements were carried out on an Instron testing unit. Bars of the porous substrate ($5\times5\times60$ mm^3) were cut using a diamond saw. Tension tests were performed in three-point bending. Compression tests were made on cylinder-shaped sample of 10-mm height and 23 mm diameter. X-ray diffraction (XRD), Scanning electron microscopy (SEM) with an energy dispersive spectrometer (EDS) was utilized to study the coating structure, surface morphology, and the interface structure. The coating bonding strength was measured through tape test (ASTM D 3359), which was originally designed for organic coatings on metallic substrates. This method was used to find the relative bonding strength. All the tests were performed on dense alumina substrates with one or multilayer HA coatings. Permacel 670 tape (Permacel) was used in the test. After removal from the coating, the tape was examined under a light microscope. Sintering process and chemical bonding of the sintering products were examined using differential thermal analysis (DTA) and Fourier transform infrared spectroscopy (FTIR).

The *in vitro* tests were conducted to evaluate the bioactivity of the synthetic HA produced by thermal deposition method (commercial HA as control). All the samples were tested in the powder form (Table 1). Table 1 summarizes the different treatments used to obtain a variety of crystalline structures in the materials; and also names the samples according to the treatment conditions, such as HA, HA600, HA900, SHA700, SHA800, SHA900, No. 1, No. 2, and No. 3. The HA sample group refers to commercial hydroxyapatite samples. HA600 and HA 900 groups refer to commercial HA samples treated for 30 min under 600 and 900 °C, respectively. The SHA700, SHA800, SHA900 conditions refer to synthesized HA using the thermal deposition method described earlier and heated for 4 h at 700, 800, and 900 °C, respectively. Sample No. 1, No. 2, and No. 3 refer to commercial HA and are heat-treated at 700 °C for 3.5 h and with different specific surface area (SSA). Sample No. 3 has the highest SSA; Sample No. 1 has the lowest SSA; and Sample No. 2 is in the middle. The simulated body fluid (SBF) solution that had ionic concentrations close to human blood plasma, as shown in Table 2, was prepared by dissolving reagent grade NaCl, $NaHCO_3$, KCl, $K_2HPO_4{\cdot}3H_2O$, $MgCl_2{\cdot}3H_2O$, $CaCl_2$, and Na_2SO_4 in ion-exchanged distilled water. The solution was buffered at pH 7.4 with 1 *M* HCl and tris(hydroxymethyl) aminomethane, $(CH_2OH)_3CNH_2$ at 37 °C. Powders were immersed into solution at 1 mg/mL ratio and maintained at 37 °C at periods ranging from 15 min to 9 weeks. The calcium concentrations in the solutions were measured by inductively coupled plasma (ICP). Subsequent to immersion, the solutions were vacuum filtered. The powders were gently rinsed with alcohol, ion-exchanged distilled water and then dried at room temperature. The surface microstructures before and after immersion of SBF solution were analyzed via scanning electron microscope (SEM). Both

Table 2. Ionic Concentration of SBF in Comparison With Those of Human Blood Plasma

	Concentration, m*M*							
	Na^+	K^+	Ca^{2+}	Mg^{2+}	HCO_3^-	Cl^-	HPO_4^{2-}	SO_4^-
Blood plasma	142.0	5.0	2.5	1.5	27.0	103.0	1.0	0.5
SBF	142.0	5.0	2.5	1.5	4.2	147.8	1.0	0.5

XRD and FTIR determined the contents of the phases that were present in the coatings. Measurements were obtained on a Philips X-ray diffractometer with CuK-radiation at 35 kV and 23 mA.

RESULTS

Suspension Method

The phase diagram for anhydrous calcium phosphates (Fig. 3) shows that the liquid phase appears at a temperature $> 1500\,^{\circ}C$, and presumably the induced liquid could improve the bonding between HA and the Al_2O_3 substrate during sintering. However, such as liquid-enhanced bonding process was not experimentally observed. Meanwhile, XRD analysis of the coating made from HA solution (without glass additive) showed that HA was decomposed, and in turn, the bioactivity of the coating was changed. A high density of cracks was found to exist in the coating. The adherence of the coating to the Al_2O_3 substrate was low and the coating layer could be peeled off by scraping. These results indicated that HA solution without sintering aid, such as glass frits, was not suitable for this particular application.

Using the sintering aid, glass frits, a well-bonded HA coating was produced. Figures 4–6 are SEM photographs of the surfaces and interfaces of the coatings made from solution with glass frits. As is apparent from Figs. 5 and 6, the glass–HA ceramic layer is firmly attached to the alumina substrate. An average coating thickness is 15 μm. The interfacial strength between the coating and the

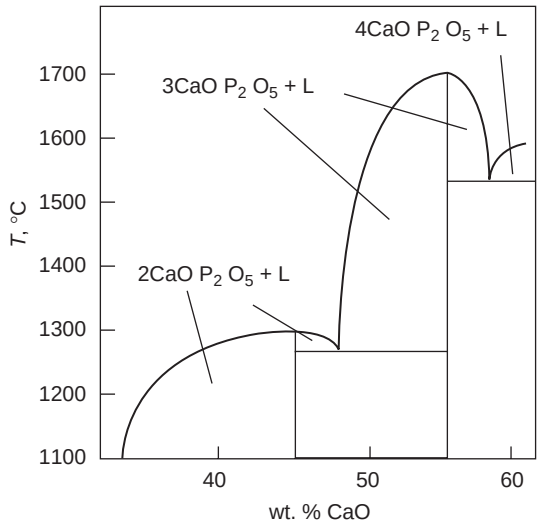


Figure 3. Phase diagram of the system $CaO–P_2O_5$, indicating the appearance of the liquid phase at high temperatures.

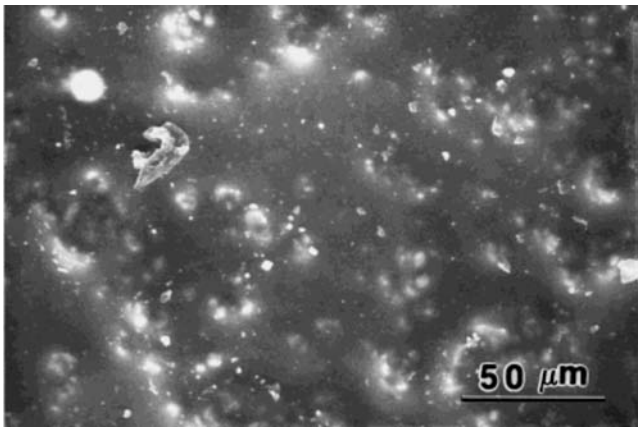


Figure 4. The SEM photograph of the coated surface.

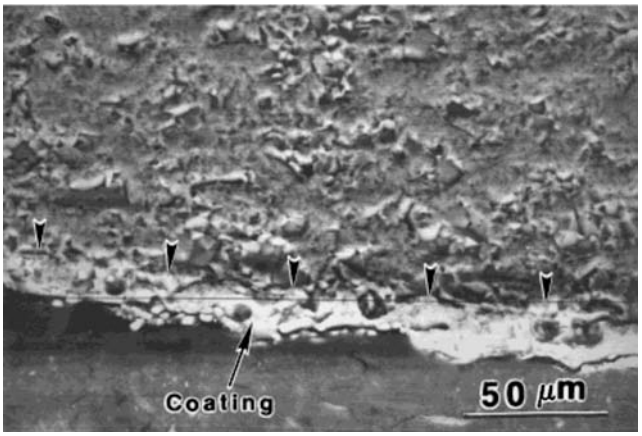


Figure 5. The SEM photograph of the coating interface for dense alumina. The interface is indicated by arrows.

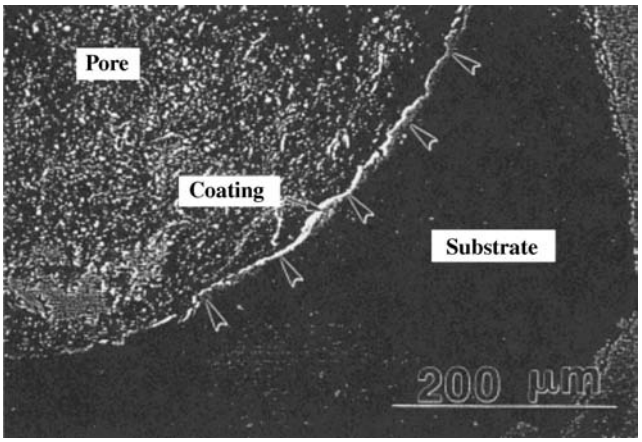


Figure 6. The SEM photograph of the coating interface for porous alumina. The interface is indicated by arrows.

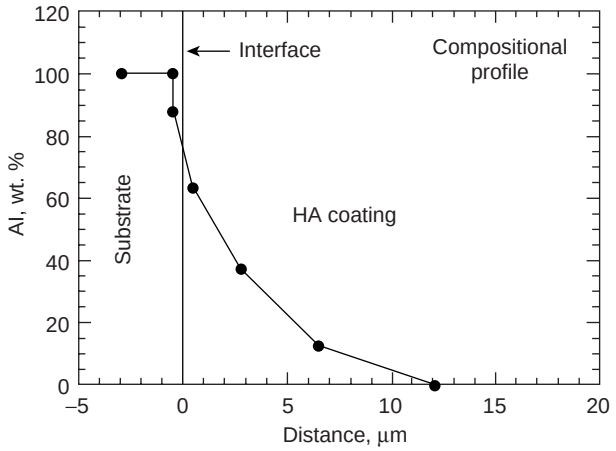


Figure 7. Aluminum content profile along the interface. The solid line is a guide to the eye.

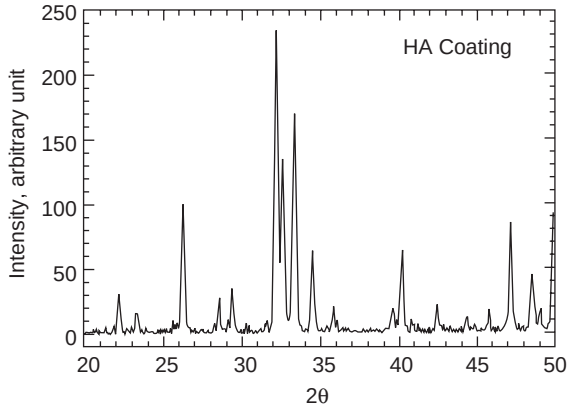


Figure 8. The XRD spectra showing a typical HA pattern of the coating surface.

substrate depends on the adherence of the glass to the Al_2O_3. Figure 7 is the aluminum compositional profile across the interface. The data showed that the aluminum concentration decreased when scanning from the substrate to the outer surface of the coating. It was concluded that aluminum ions diffused during sintering, and consequently bonded the glass to the alumina substrate by ion diffusion. This diffusion bonding is attributed to the formation of a eutectic compound at the interface during the sintering, and thus ensures the strong bonding between the coating and the substrate.

The above results indicate that the development of glass frits is essential for having an excellent adhesion of HA coating to the alumina substrate. Great efforts were therefore made in the preparation of the glass frits. As a sintering aid, the glass must wet the substrate and HA, and its melting point should be lower than the decomposing temperature of HA (1300 °C). Furthermore, for successful coating, optimizing the coefficient of thermal expansion of the glass to match the substrate is critical. It has been known that the mismatch in expansion coefficients between the coating and the substrate materials will give rise to interfacial stress that weakens the bonding strength or leads to the cracking and spalling of the coating. The magnitude of this stress is proportional to the difference between the thermal expansion coefficients of the coating and the substrate. The expansion coefficient of HA (13.3×10^{-6}/°C) is relatively higher than that of alumina (8.0×10^{-6}/°C). The expansion coefficient of the selected glass should then be an intermediate one to reduce this difference. Another important aspect of the glass is chemical durability. For biological applications, it is essential for glass to be nontoxic and stable in the body fluid. The dissolution of the glass will lead to the degradation of the coating. The HA particles will escape from the coating, and this will have an extremely negative effect, such as interfacial loosening and tissue inflammation, on the bone regeneration. The optimal properties of the glass can be achieved by adjusting the glass compositions. The glass selected in this work was borosilicate glass. Its expansion coefficient is compatible with that of Al_2O_3 substrate. No crack was found in the coating (Fig. 4). The XRD pattern of the coating (Fig. 8) shows that there is no negative reaction between the glass and HA. Figure 8 is a typical HA diffraction pattern. Mechanical properties of HA-coated reticulated alumina are ~7–10.35 MPa for compressive strength and 5–8 MPa for tensile strength. Compared to previously reported porous materials, such as porous HA (1.3 MPa for compressive strength and 2.5 MPa for tensile strength) and coralline ceramic (5.8 MPa for compressive strength and 1.3 MPa for tensile strength) (6), a substantial increase in strength was obtained. These values are comparable to those of cancellous bone (2–12 MPa for compressive strength and 10–20 for tensile strength) (2). Some much stronger substrate materials, such as fiber reinforced composites, are excellent candidates for the HA coating using the developed approaches discussed in this article. The mechanical properties of the substrate can also be significantly improved by other ceramics routes. In addition, after bone ingrowth, the strength of the implant (bone composites) will be expected to further increase by a factor of 2–7 as previously demonstrated (7).

Thermal Deposition Method

Figure 9 shows the FTIR spectra of an unfired sample and samples fired at 500, 600, 700, and 900 °C. According to standard IR transmission spectra, peaks observed at 3573 and 631 cm^{-1} are assigned to OH stretching and librational modes. Peaks ~600 and 1100 cm^{-1} are due to the bending and stretching modes of PO bonds in the phosphate groups (8). These are characteristic peaks of HA. At a sintering temperature of 500 °C, PO bonds formed, but hydroxyl groups were not detected. Compared with the spectra of the unfired sample, most organic groups were burned out by this temperature. At 600 °C, all the characteristic lines of HA were recorded, but some organic residual could still be seen. At 700 °C and higher the peak positions match all those of standard HA, and the organic groups were not detectable.

Figure 10 shows the XRD spectra of HA sintered at different temperatures in the range of 600–900 °C. The results of the XRD are quite consistent with that of the FTIR. The crystalline phase started to form at 600 °C, and

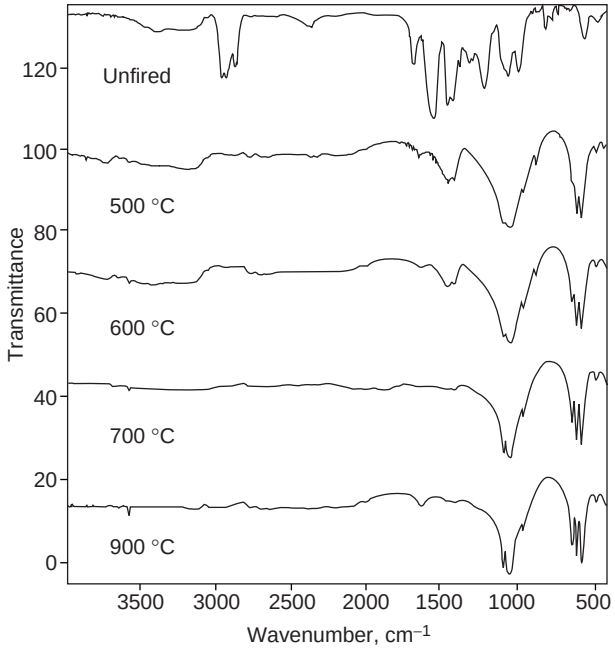


Figure 9. The FTIR spectra of HA-coated samples sintered for 4 h at the temperature indicated (thermal deposition).

all peaks were attributed to the HA phase. According to the Scherrer equation,

$$\Delta(2\theta) = K\lambda/D \tag{1}$$

where D is the crystallite dimension; K is the Scherer constant (here $K = 0.9$); is the X-ray wavelength in angstroms; $\Delta(2)$ is the true broadening of the diffraction peak at half-maximum intensity. The crystallite size is inversely proportional to the peak width. The broadening of peaks was evident at lower sintering temperatures, indicating the initial state of crystal formation. At higher sintering temperatures, the sharpening of peaks evidenced the growth of crystals. The peak shift could also be noted by comparing it with the standard XRD spectra of HA. At lower temperatures the shift was considerable, suggesting great lattice distortion.

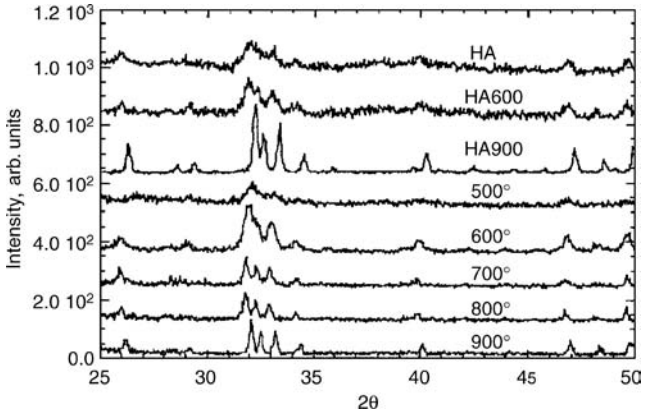


Figure 10. The XRD spectra of HA samples sintered at the temperatures indicated.

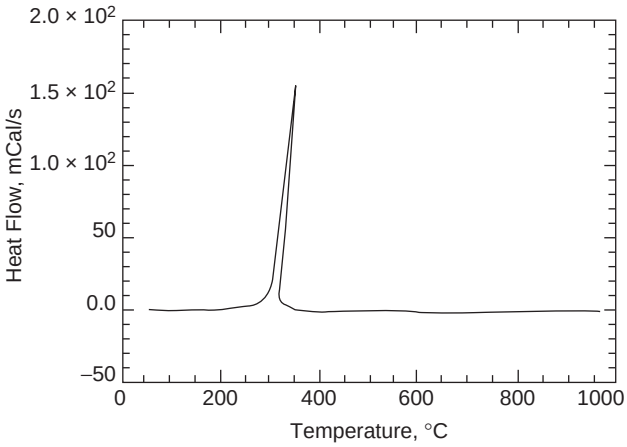


Figure 11. The DTA profile of the HA-coated sample showing a reaction near 300 °C.

Monitoring the sintering process and the evolution of chemical bonds is important in determining the bioactivity of the sintered products. The material with more lattice defects would be expected to be more reactive (2). This assumption will be experimentally verified later in *in vitro* tests. The high temperature and long sintering time will result in well-crystallized products. Therefore, to enhance the bioactivity, low temperature and short sintering time is preferred. It is critical to find an optimal sintering procedure so that the sintered HA is poorly crystallized but with no organic residuals. The DTA profile (Fig. 11) shows that burn-out of organic residuals occurs over the temperature range of 300–350 °C. In the current work, the samples were baked at 500 °C to burn out the organic groups. At this temperature, the structure of the sample is amorphous and most of the organic groups can be easily removed. This procedure will be helpful in eliminating residual carbon in the coating. Without this treatment, some organics could be incorporated in the final crystal lattice. It was found that the carbon disappeared at much lower temperatures than those samples treated in a rapid sintering process because most organic groups were not burned out at low temperatures. Therefore, a higher temperature is needed to remove them. However, the reactivity of HA is considerably reduced. It should be noted that the removal of the organic residue is not only related to the microstructure, but also to the macrostate of the samples. For example, for a thick and dense coating, a high temperature is needed to remove the residual carbon.

The bonding strength between the HA coating and the substrate was determined using the tape test. No peeling of the coating film was observed for all samples, indicating a strong bonding between the HA coating and the substrate. Figure 12 is the SEM micrographs showing the surfaces of the HA coating on dense alumina. As can be seen in this figure, the coating is fairly porous, which contributes to the bioactivity when immersed in SBF. Figure 13 is an SEM image of HA-coated reticulated alumina with significant open pores in the matrix. Figure 13b is the X-ray map recorded with Ca K_α lines for coated porous alumina. As can be seen in Fig. 13b, the distribution of calcium demonstrated that HA is uniformly coated on the skeleton of the

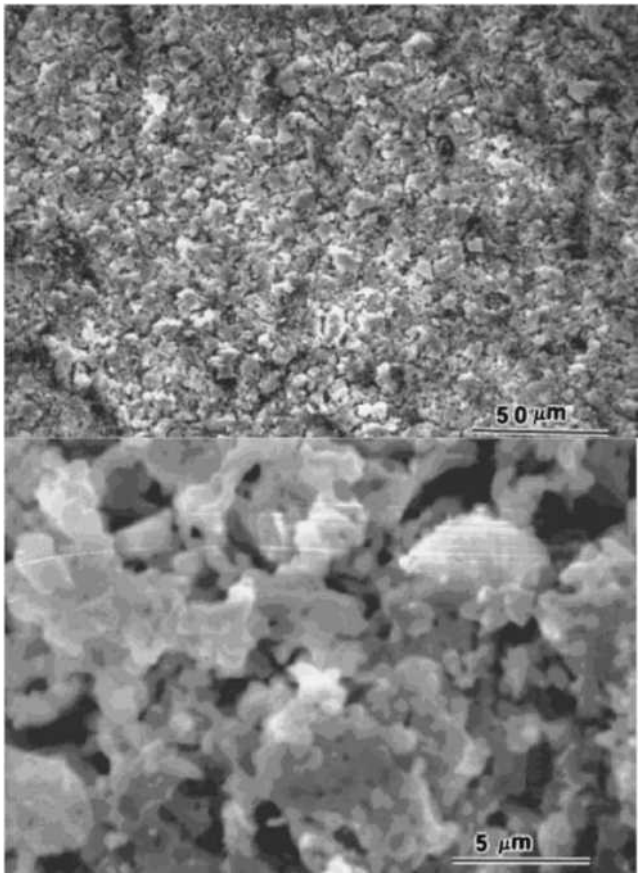


Figure 12. The SEM photographs showing the coated surface for a dense alumina substrate.

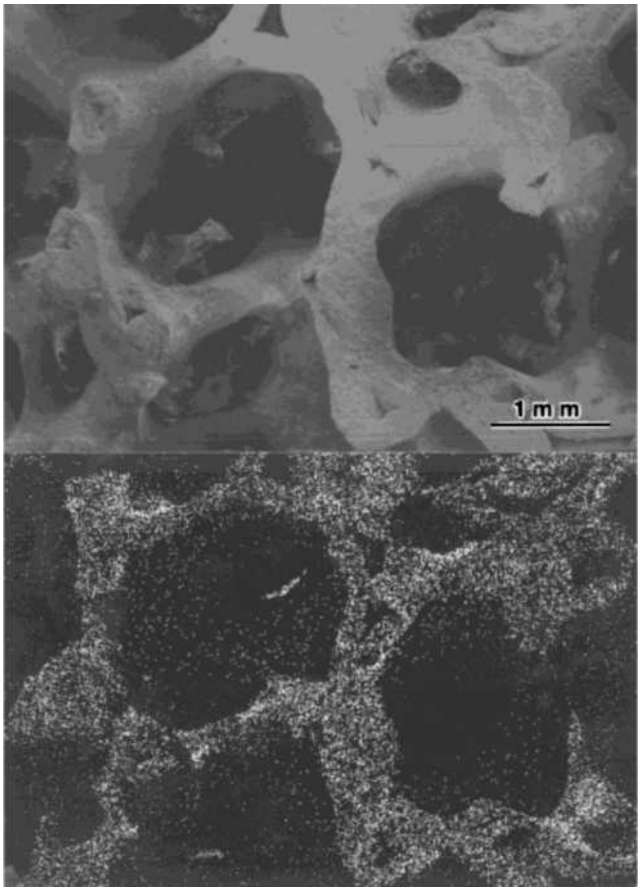


Figure 13. The SEM photographs showing (a) porous alumina substrate and (b) an X-ray map recorded with Ca K_α lines.

substrate. Figure 14 is the SEM micrograph showing the interface between the HA coating and the substrate. The coating thickness was ~ 1 μm, which can also be altered by a second or third coating.

Bioactivity Test

Due to bioactivity of HA, dissolution occurs after the sample is immersed in SBF. Consequently, some of the elements such as calcium in the solution are expected to change. The elemental-concentration changes of calcium in the SBF solution as a function of time are given in Figs. 15,16. As can be seen in Fig. 15, both HA and HA600 exhibit an immediate uptake of the Ca concentration. Initially, there is a high rate of ion uptake, suggesting the formation of a new phase on the HA surface in supersaturated solution. After 24 h, with the depletion of supersaturation, the reaction proceeds at a lower rate of uptake. For HA900, there is an induction time of 60 min prior to a detectable decrease in Ca concentration, and the initial rate of Ca uptake is much lower than those of HA and HA600. The SHA700 behaves similarly to HA and HA600 with a slow reaction rate, as can be seen in Fig. 16. However, the reaction behaviors of SHA800 and SHA900 significantly differ from the HA samples. During the first hour, an increase of Ca concentration was measured, indicating that dissolution of SHA800 and SHA900 has surpassed the new phase formation. Note that the rise in supersaturation for SHA900 is greater than that for SHA800. The ion uptake takes place after this initial dissolution. Another difference between HA and SHA series is that the latter took longer to reach the solid–solution equilibrium stage, clearly indicating a slower reaction rate in the HA series. These results suggest that the dissolution and precipitation rates are

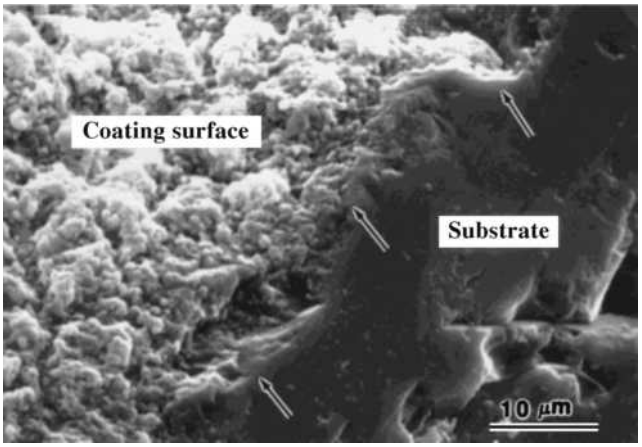


Figure 14. The SEM photograph showing the interface between the HA coating and the dense alumina substrate.

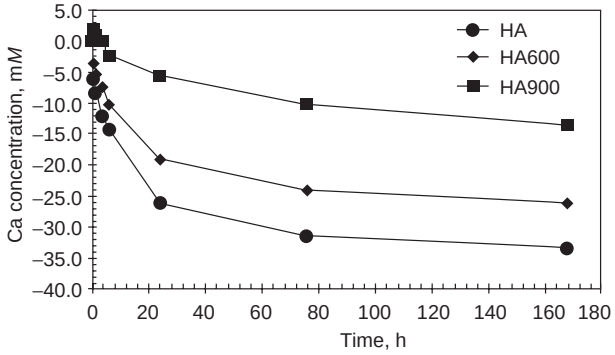


Figure 15. The Ca concentration in SBF versus immersion time for the HA group sintered at temperature indicated.

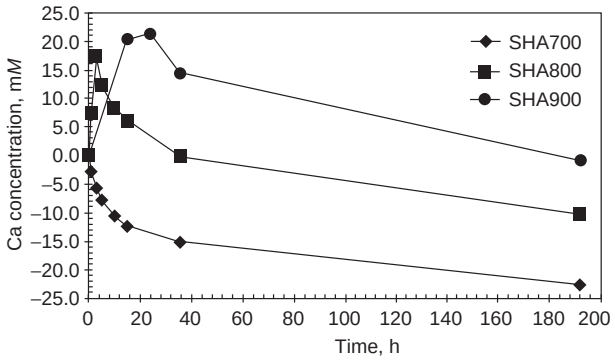


Figure 16. The Ca concentration in SBF versus immersion time for the SHA group sintered at temperature indicated.

critically dependent on the crystal structures developed in the HA samples.

Figure 17 represents the Ca concentration in the solution as a function of immersion time for samples Nos. 1, 2, and 3. All these samples are commercial HA heat treated at 700 °C for 30 min. Therefore, these samples are of the same structural crystallinity, but with different specific surface areas. Sample No. 3 has the highest SSA; Sample No. 1 has the lowest SSA; and Sample No. 2 is in the middle. They were tested at a ratio of 1 mg · mL^{-1} SBF. It is apparent in Fig. 17 that the rates of precipitation are highly dependent on the surface area. From these kinetic curves, the initial

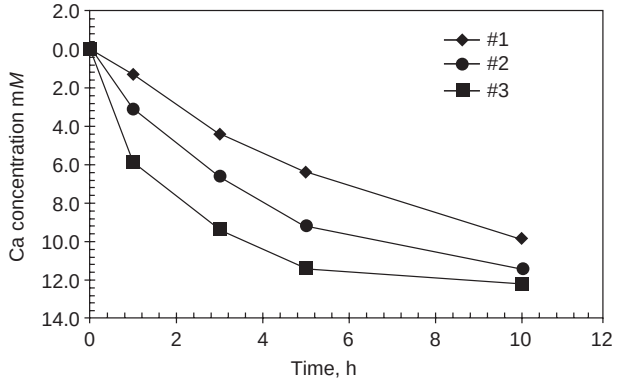


Figure 17. The Ca concentration in the solution as a function of immersion time for samples No. 1, No. 2, and No. 3.

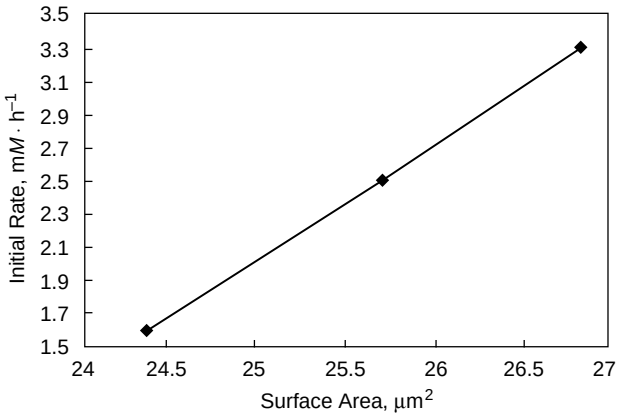


Figure 18. Initial reaction rate versus specific area for the samples showing in Fig. 17.

rate of precipitation, R_0, was determined by the slope of the first two data points. As shown in Fig. 18, there is a linear relationship between the initial precipitation rate and the surface area.

Figure 19 is a plot of Ca concentration versus immersion time for HA, HA600, SHA800, and SHA900. Samples of each group have been selected to have the same surface area. As can be seen, the initial rates of HAs and SHAs separate into two branches. The HA group exhibits an initial gradual decrease, while that of SHA group increases quite rapidly. However, as can also be seen in this figure, calcium concentration of SHA800 initially increases, but reaches a peak at 3 h, and thereafter decreases. In SHA900, although with a different rate, the calcium concentration always increases up to 9 h. Therefore, it can be concluded that the specific surface area is not the only factor that affects the reaction behavior of various HA samples. As discussed later, the degree of crystillinity in fact plays an even more important role in the reaction rates. The SHA700 sample behaves similarly to HA and HA600 with a slow reaction rate as can be seen in (Fig. 19). However, the reaction behaviors of SHA800 and SHA900 significantly differ from the HA samples. During the first hour, an increase of Ca concentration was measured indicating that dissolution of SHA800 and SHA900 has surpassed the new phase formation. It is noted that the rise in super-

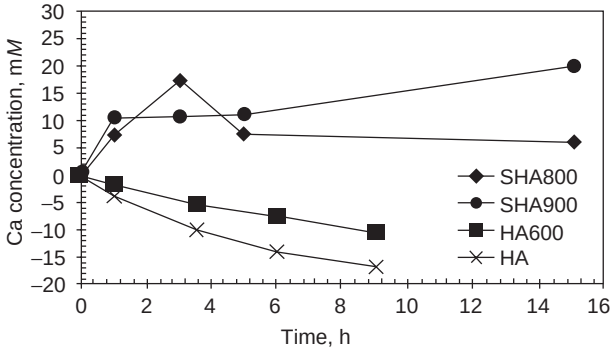


Figure 19. The Ca concentration versus immersion time for some of the typical HA and SHA samples.

Figure 20. The SEM photograph showing the surface morphology of the SHA700 immersed in SBF for 1 week.

saturation for SHA900 is greater than that for SHA800. The ion uptake takes place after this initial dissolution. Another difference between HA and SHA series is that the latter took longer time to reach the solid–solution equilibrium stage, clearly indicating a slower reaction rate in the HA series. These results suggest that the dissolution and precipitation rates are critically dependent on the crystal structures developed in the HA samples.

Figures 20 and 21 are surfaces of SHA700 and SHA900 coatings after immersion in SBF for 1 week and 9 weeks, respectively. As can be seen, the morphology of these two surfaces is quite different. The SHA700 coating is fully converted to flake shape with an average diameter of 1–10 μm. The surface of the flakes exhibits fine, needle-like structures within 1 week, which have been identified to be HCA by IR analysis. For SHA900 coating, after 9 weeks of immersion in SBF, a layer of precipitation has been observed under high magnification, which is shown to be amorphous or poorly crystallized new phase instead of HCA.

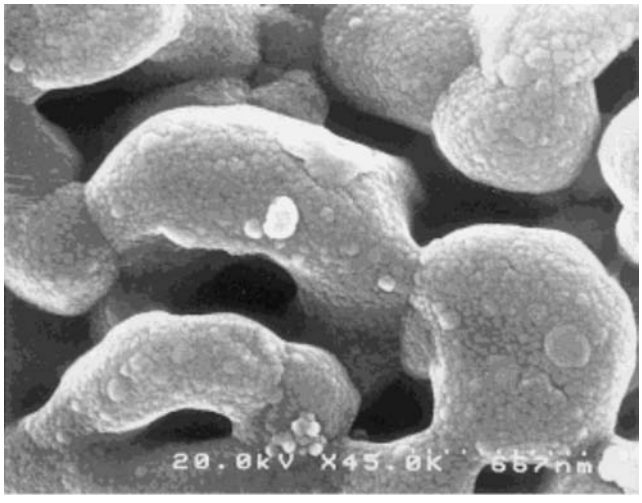


Figure 21. The SEM photograph showing the surface morphology of the SHA900 immersed in SBF for 9 weeks.

Figure 22 shows the FTIR spectra of HA and SHA samples after immersion in SBF at time periods up to 1 week. The absorption bands at 1460 cm^{-1} (high C=O region) and 872 cm^{-1} (low CO region) are characteristic features of HCA (8). As can be seen from spectrum of HA (Fig. 22a), these bands became significantly greater after 76 h of immersion indicating an increase in carbonate content. A gradual reduction of the splitting of the major PO_4^{3-} absorption bands (1100–1000 and 600–550 cm^{-1}) with immersion time is also observed, suggesting the formation of amorphous or fine, poorly crystallized new phases. For HA900, a broad band appears in the high energy C=O region (Fig. 22b). However, the low energy C=O band at 872 cm^{-1} is not recorded. At the same time, a gradual reduction of the splitting of the major PO_4^{3-} bands is observed, indicating again the formation of amorphous or fine, poorly crystallized new phases. The HCA phase cannot be identified from these weak bands, and it is likely that some intermediate phases other than HCA have formed. The HCA peaks have appeared in the spectra of SHA700 within 7 days (Fig. 22c). A time-dependent increase in the carbonate band intensities accompanied by a reduction of splitting of the major PO_4^{3-} bands is again recorded. Similar changes have occurred in the spectra of immersed SHA800 (Fig. 22d). However, no characteristic HCA peaks are recorded for SHA900 up to 3 weeks, only a broad band has appeared in the high energy C=O region (Fig. 22c). This trend seems to indicate that the reactivity of HA is considerably reduced at higher temperatures.

DISCUSSION

Structural Effects

The results indicate that *in vitro* behavior of the HA coatings is strongly affected by the structural characteristics induced by heat treatment. The SBF used in this work represents physiological ion concentration in human body, and it is supersaturated with respect to HA. In this chemical environment, HA is the most stable phase among all the calcium phosphate phases, thus the HCA formation is thermodynamically possible. However, only HA, HA600, and SHA700 have led to immediate Ca ions uptake. The HA900, SHA800, and SHA900 samples show a partial dissolution prior to precipitation. The difference in the dissolution ability of the HA samples is not the only factor in bioactivity. Figure 10 shows XRD spectra of HA sintered at different temperatures in the range of 600–900 °C. The structural evolution begins from an amorphous state in the commercial HA. Crystalline phase started to form at 600 °C, and all peaks were attributed to the HA phase. In addition, relative peak intensities are in agreement with the expected values for HA. Therefore, it can be decided that the structure consists primarily of crystalline HA, no additional peaks were observed to appear at any firing temperatures. However, the peak shift could be noted by comparing with the standard XRD spectra of HA. At lower temperatures, the shift was considerable suggesting great lattice distortion. The breadth of the peaks was used as an indicator of crystal dimension in the direction perpendicular to the diffracting plane hkl. The crystal size D is

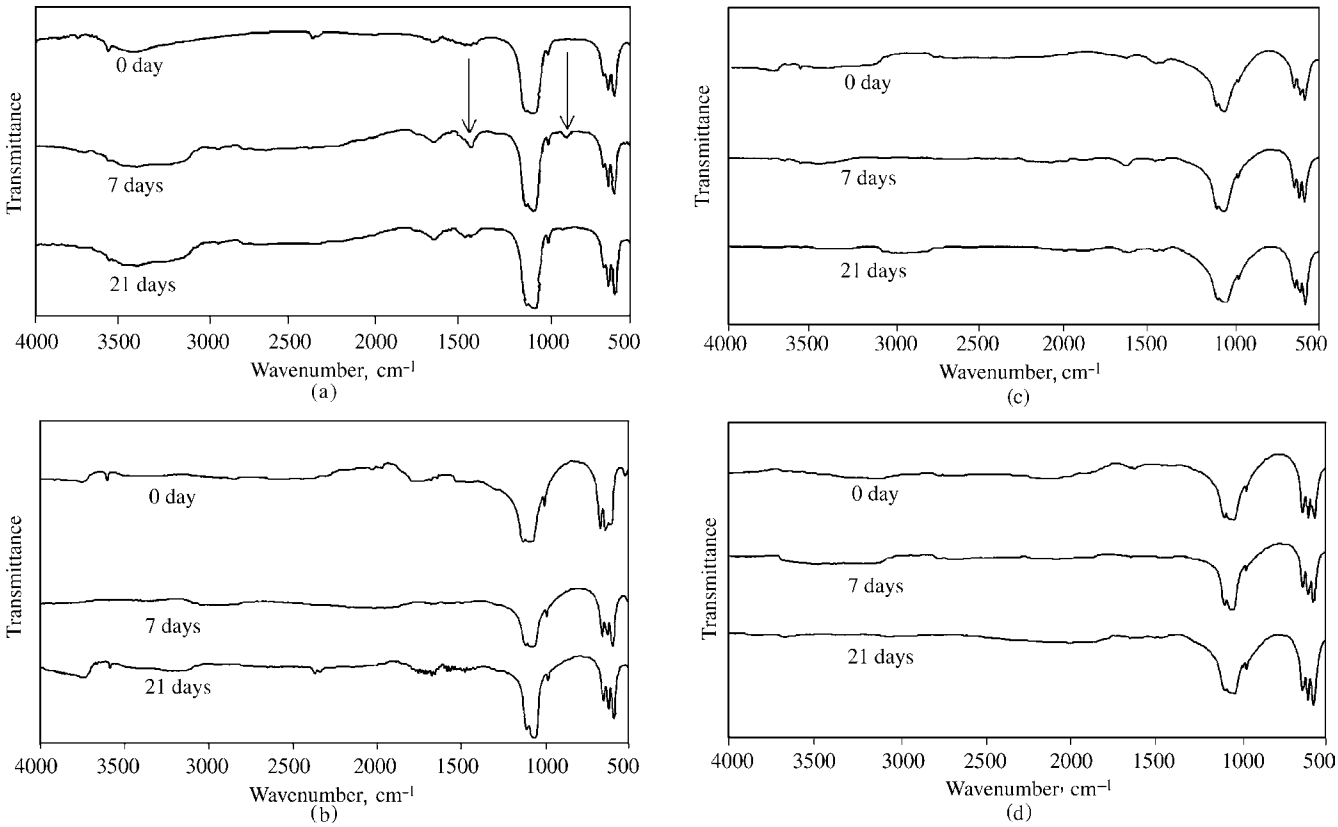


Figure 22. The FTIR spectra of HA and SHA samples after immersion in SBF for the time indicated.

inversely proportional to the peak breadth according to the Scherer equation. The contribution to the peak breadth from instrumental broadening was determined to be $\sim 0.12\,^{\circ}C$ (0.002 rad), independent of 2θ. This amount subtracted from the total experimental width is the value of true broadening, assuming the two contributions add linearly. The peak breadth (D002) is given as a function of temperature in (Fig. 23). It can be seen that the peak breadth decreases with sintering temperature, indicating that the crystal size increases with increasing sintering temperature, from 600 to 900 °C. On the basis of above analysis, the important difference with annealing temperature was the size of the individual crystals and the amount of crystal defects.

It is possible that the crystal growth rate is controlled by more than one of the elementary rate controlling mechanisms. The rate controlling process can change depending on particle size, solution concentration, and surface properties of the crystallites. The mechanisms of crystal growth are usually interpreted from measured reaction rates at different driving forces or from the activation energies of reactions. It is common practice to fit the data to an empirical rate law, which is represented by simple empirical kinetics (9):

$$R_g = k_g s \sigma^n \qquad (2)$$

where k_g is the rate constant for crystal growth, s is a function of the total number of available growth sites, is the degree of supersaturation, and n is the effective order of reaction. A broad empirical test for growth mechanism can be achieved from a logarithmic plot of Eq. 2. From the n value, the probable mechanism can be deduced. It is possible that the crystal growth rate is controlled by more than one of the elementary rate controlling mechanisms listed above. Under these circumstances, the rate-limiting steps are dependent on the jump frequency of lattice ions: (*1*) through the solution for mass transport control; (*2*) to the crystal surface for adsorption control, or (*3*) along the crystal surface or into a crystal lattice kink site for spiral and polynuclear control. The rate controlling process can change depending on particle size, solution concentration, and surface properties of the crystallites. A broad empirical

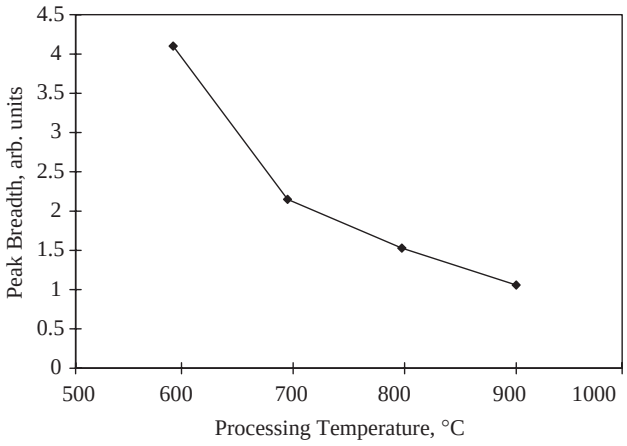


Figure 23. The X-ray diffraction peak breadth versus processing temperature for the HA synthesized in this study.

test for growth mechanism can be achieved by plotting the data according to Eq. 2. An effective order reaction in the range $0 < n < 1.2$, $n \sim 2$, or $n > 2.5$ indicates that the rate controlling process is one of adsorption and /or mass transport, surface spiral, or polynucleation, respectively (9). Experimentally, it is found that the growth rates of the calcium phosphates are insensitive to changes in fluid dynamics indicating surface controlling mechanisms rather than mass transport of ions to the crystal surfaces. A dynamic fluid may effect the growth rate, but not in a pronounced fashion. However, in our experiment, we found the growth to be insensitive. But we have no comparison of growth rates in both static and dynamic fluids.

Temperature Dependence of Activation Energy

Activation energies, obtained from experiments at different temperatures, can be used to differentiate between volume diffusion and surface controlled processes. The activation energy for volume diffusion, reflecting the temperature dependence of the diffusion coefficient, usually lies between 16 and 20 $kJ \cdot mol^{-1}$, while for a surface reaction the value may be in excess of 35 $kJ \cdot mol^{-1}$. If a reaction has activation energy of < 20 $kJ \cdot mol^{-1}$, it is safe to assume that it is overwhelmingly controlled by volume diffusion. However, if the activation energy is higher than 35 $kJ \cdot mol^{-1}$, it is quite certain that an adsorption process predominates. In all other cases, both adsorption and volume diffusion mechanisms may participate for a first-order reaction (10).

Figure 17 represents the Ca concentration in solutions as a function of immersion time at different surface areas. These samples were the same kind of powders to ensure that they have the same crystal structure and surface morphology; while the ratio of surface area to volume of SBF was different. It is apparent in Fig. 17 that the rates of precipitation were highly dependent on the surface area. Based on the empirical kinetics in Eq. 2, to build a relationship between the reaction rate R_g and surface area s, the degree of supersaturation σ should be kept at a constant value. The corresponding reaction rates were calculated by a simple fitting procedure from the above kinetic plots. As shown in Fig. 18, there was a linear relationship between the precipitation rates and the total surface area, which is in agreement with the above empirical kinetics equation. This result also showed that crystallization occurred only on the added seed materials without any secondary nucleation or spontaneous precipitation. Furthermore, the advantage of porous bioceramics over dense bioceramics was proved by this relationship.

The initial precipitation rate was not used here because of the following considerations. First, the empirical fitting procedure used to calculate R_0 is greatly affected by the slower rates occurring after the initial fast stage of the precipitation process. Thus, the fitting data could not represent the true initial rate. Second, initial rate was a complicated factor. Rapid adjustment of surface composition usually happened when the solids were introduced into the growth or dissolution media. In the case of HA, initial uptaking surges were observed, which might be attributed to calcium ion adsorption. Therefore, considerable uncertainties can arise if too much emphasis is placed upon initial rates of reaction.

Another point needed to be noted was that in this test, the different surface areas were not originated from the distribution of particle sizes, considering that different particle sizes might bring in the factor of surface morphology, which has great influence on the reaction rate. The effect of particle size would be demonstrated later. In the current method, the same powders were used, so that the factor of morphology was eliminated and a linear relationship was obtained.

The particles of different sizes behaved differently under the same surface area to volume (SA/V) test conditions. When comparing the 40–100-mesh and < 200-mesh particles at SA/V of 0.02 $m^2 \cdot mL^{-1}$, it is apparent that the Ca adsorption rate is slower for the smaller particles. This may be attributed to physical differences such as the radius of curvature and surface roughness.

Figure 19 is a plot of Ca concentration verses immersion time for HA, HA600, and SHA800, SHA900. Samples of each group were tested under the same SA/V ratio. As can be seen, the initial rate of HA was greater than that of HA600; the behavior of SHA800 differed from the one of SHA900. Therefore, it is concluded that the specific surface area is not the only reason that affects the reaction behavior of various HA powders, the degree of crystallinity also plays an important role in their reaction rates.

Chemical reactions, specifically in this case, the process of nucleation and crystal growth from solution, is described as an activated process with temperature, which is represented by the following relationship:

$$\text{rate} \propto \exp\left(-\frac{E_a}{kT}\right) \tag{3}$$

where E_a is the activation energy, so that reaction rate increases exponentially with temperature increase. The reaction rate constant K is related to temperature by an Arrhenius equation:

$$K = K_0 \exp\left(-\frac{E_a}{kT}\right) \tag{4}$$

By keeping σ at a constant value, plot $\ln R$ versus $1/T$, the slope of the curve will be E_a/k, and consequently E_a can be calculated.

According to the procedures described above, the activation energy for HA, HA600, HA900, and SHA700 was calculated. The parameter σ was selected at $\Delta Ca = -8mM$ for all the reaction temperatures. The computed activation energy was listed in Table 3. The above results showed that the activation energy increased with the sintering temperature for HA powders. The activation energy of synthesized HA700 was much higher than those of HA and HA600.

In Vitro Biochemistry Behavior of Hydroxyapatite

The formation of biological apatite on the surface of implanted synthetic calcium phosphate ceramics goes

Table 3. Activation Energies for the Samples Indicated

Samples	Activation energy, $kJ \cdot mol^{-1}$
HA	66.3
HA600	80.3
HA900	172.7
SHA700	130.4

through a sequence of chemical reactions. It has been shown that the reaction rate *in vitro* appears to correlate with the rate of apatite mineral formation *in vivo*.

Therefore, the laboratory observations can be projected to the *in vivo* situation. The *in vitro* behavior of bioceramics is determined by its stability at ambient and body temperatures. Many factors have significant influence on their stability, including the pH and supersaturation of the solution, crystallinity, structure defects, and porosity of the material (11,12). Driessens (13) showed that, among the phases composed of calcium and phosphate, hydroxyapatite is the most stable at room temperature when in contact with SBF, which was used to represent the ionic concentrations of plasma. Generally, SBF will initiate a partial dissolution of the HA material causing the release of Ca^{2+}, HPO_4^{2-}, and PO_4^{3-}, and increasing the supersaturation of the microenvironment with respect to HA phase that is stable in this environment. Following this initial dissolution is the reprecipitation. Carbonate ions, together with other electrolytes, which are from the biological fluids, become incorporated in the new apatite microcrystals forming on the surfaces of the HA.

Since any clinical use of calcium phosphate bioceramics involves contact with water, it is important to understand the stability of HA in the presence of water at ambient temperatures. As Driessens showed (13), there were only two classes of calcium phosphate materials that were stable at room temperature when in contact with aqueous solution. Temperature, ionic strength, and pH are major parameters influencing the stability of calcium phosphate. In the body, temperature and ion strength are constant, therefore, the pH value at the local tissue determines which form of calcium phosphate is the most stable. At a pH < 4.2, the component $CaHPO_4{\cdot}2H_2O$ was the most stable, while at higher pH (> 4.2), HA was the stable phase. However, HA does not form at the first place. Other mineral phases such as dicalcium phosphate dihydrate (DCPD), octacalcium phosphate (OCP), and amorphous tricalcium phosphate (TCP) form as precursor phases that transform to HA.

Therefore, in this *in vitro* test, at biological pH value, only HA or its precursor phase can be found in contact with SBF. It is believed that synthetic HA ceramic surfaces can be transformed to biological apatite through a set of reactions including dissolution, precipitation, and ion exchange. Following the introduction of HA to SBF, a partial dissolution of the surface is initiated causing the release of Ca^{2+}, HPO_4^{2-}, and PO_4^{3-}, which increases the supersaturation of the microenvironment with respect to the stable (HA) phase. Carbonated apatite can form using the calcium and phosphate ions released from partially dissolving ceramic HA and from the biological fluids that contain other electrolytes, such as CO_3^{2-} and Mg^{2+}. These become incorporated in the new CO_3-apatite microcrystals forming on the surfaces of ceramic HA crystals. The *in vitro* reactivity of HA is governed by a number of factors, which can be considered from the two aspects: *in vitro* environment and properties of HA material.

CONCLUSIONS

In order to produce highly strengthened porous bioactive materials for bone substitutes, suspension method and thermal deposition method, were employed to coat the inner-pore surfaces of a porous ceramic substrate. A thin layer of HA has been uniformly coated onto inner-pore surfaces of reticulated alumina substrates. The *in vitro* bioactivity of HA coatings was found to be strongly affected by structure characteristics, which are a combination of crystallinity and specific surface area. The bioactivity is reduced at a higher degree of crystallinity, which is likely related to the higher driving force for the formation of a new phase, and the reaction rate was proportional to the surface area. The surface morphology and HA treating temperature also have a direct affect on the reaction rates of the HA coatings. The calcium absorption rate is slower for smaller particles; this could be attributed to physical differences including radius of curvature and surface roughness. The activation energy increased with the heat-treatment temperature for HA powders.

BIBLIOGRAPHY

Cited References

1. Shinzato S, et al. Bioactive bone cement: Effect of silane treatment on mechanical properties and osteoconductivity. J Biomed Mater Res 2001;55(3):277–284.
2. Hench LL. Introduction to Bioceramics. Singapore: World Scientific; 1993. p 139–180.
3. Barth E, Hero H. Bioactive glass ceramic on titanium substrate: the effect of molybdenum as an intermediate bond coating. Biomaterials 1986;7(4):273–276.
4. Kasuga T, et al. Bioactive calcium phosphate invert glass-ceramic coating on beta-type Ti-29Nb-13Ta-4.6Zr alloy. Biomaterials 2003;24(2):283–290.
5. Livingston T, Ducheyne P, Garino J. In vivo evaluation of a bioactive scaffold for bone tissue engineering. J Biomed Mater Res 2002;62(1):1–13.
6. Roy DM, Linnehan SK. Hydroxyapatite formed from coral skeletal carbonate by hydrothermal exchange. Nature (London) 1974;247(438):220–222.
7. Holmes R, et al. A coralline hydroxyapatite bone graft substitute. Preliminary report. Clin Orthop 1984;188:252–262.
8. Radin SR, Ducheyne P. The effect of calcium phosphate ceramic composition and structure on in vitro behavior. II. Precipitation. J Biomed Mater Res 1993;27(1):35–45.
9. Nielsen AE. Electrolyte Crystal Growth Mechanisms. J Crystal Growth 1984;67: 289–310.
10. Gengwei J. Development of Bioactive Materials using Reticulated Ceramics for Bone Substitue. Ph. D. dissertation, University of Cincinnati; 2000. p 118.
11. Margolis HC, Moreno EC. Kinetics of hydroxyapatite dissolution in acetic, lactic, and phosphoric acid solutions. Calcif Tissue Int 1992;50(2):137–143.
12. Christoffersen J, Christoffersen MR. Kinetics of Dissolution of Calcium Hydroxyapatite. 5. The Acidity Constant for the Hydrogen Phosphate Surface Complex. J Crystal Growth 1982;57: 21–26.
13. Driessens FCM. Formation and Stability of Calcium Phosphates in Relation to the Phase Composition of the Mineral in Calcified Tissues. In: de Groot K, editor. Bioceramics of Calcium Phosphate. Boca Raton, (FL): CRC Press; 1983; p 1–31.

See also BIOMATERIALS, CORROSION AND WEAR OF; NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY; POROUS MATERIALS FOR BIOLOGICAL APPLICATIONS.

BIOMATERIALS: TISSUE-ENGINEERING AND SCAFFOLDS

GILSON KHANG
Chonbuk National University
SANG JIN LEE
MOON SUK KIM
HAI BANG LEE
Korea Research Institutes of Chemical Technology

INTRODUCTION

Tissue engineering offers an alternative to whole organ and tissue transplantation for diseased, failed, or abnormally functioning organs. Millions suffer from end-stage organ failure or tissue loss annually. In the United State alone, at least 8 million surgical operations are carried out each year at a total national healthcare cost exceeding $400 billion annually (1–4). Approximately 500,000 coronary artery bypass surgeries are conducted in the United States annually (5). Autologous and allogenic natural tissue, such as the saphenous vein or the internal mammary artery, is generally used for coronary artery replacement. The results have been favorable for these procedures with patency rates generally ranging from 50–70%. Failures are caused by intimal thickening due largely to adaptation of the vessel in response to increased pressure and wall shear stress, compression, inadequate graft diameter, and disjunction at the anastomosis. Also, successful treatment has been limited by the poor performance of the synthetic materials used, such as polyethyleneterephthalate (PET, Dacron) and expanded polytetrafluoroethylene (ePTFE, Gore-Tex), which are used for tissue replacement due to plaguing problems (6). For example, in cases of tumor resection in the head, neck, and upper and lower extremities, as well as in cases of trauma and congenital abnormalities, there are often outline defects due to the loss of soft tissue, this tissue is largely composed of subcutaneous adipose tissue (7). The defects lead to abnormal cosmesis, affect the emotional comfort of patients, and may impair function. A surgeon would prefer to use an autologous adipose tissue to sculpt contour deformities. Because mature adipose tissue does not transplant effectively, numerous natural, synthetic, and hybrid materials have been used to act as adipose surrogates. Despite improved patient outcomes, the use of many of these materials results in severe problems, such as unpredictable outcomes, fibrous capsule contraction, allergic reactions, suboptimum mechanical properties, distortion, migration, and long-term resorption.

To offset the short supply of donor organs as well as the problems caused by the poor biocompatibility of the biomaterials used, a new hybridized method of "tissue engineering", which combines both cells and biomaterials has been introduced (8). To reconstruct new tissue by tissue engineering, a triad of components are requried: (*1*) harvested and dissociated *cells* from the donor tissue including nerve, liver, pancreas, cartilage, and bone as well as embryonic stem, adult stem, or precursor cell; (*2*) *scaffolds* made of biomaterials on which cells are attached and cultured, then implanted at the desired site of the functioning tissue; (*3*) *growth factors* that promote and/or prevent cell adhesion, proliferation, migration, and differentiation by up-regulating or down-regulating the synthesis of protein, growth factors, and receptors (see Fig. 1). In a typical application for cartilage regeneration, donor cartilage is harvested from the patient and dissociated into individual chondrocyte cells using enzymes as collagenase, and then mass cultured *in vitro*. The chondrocyte cells are then seeded onto a porous and synthetic biodegradable scaffold. This cell–polymer structure is massively cultured in a bioreactor. The abnormal tissue is removed and the cell–polymer structure is then implanted in the patient. Finally, the synthetic biodegradable scaffold resorbs into the body and the chondrocyte cell produces collagen and glycosaminoglycan as its own natural extracellular matrix (ECM), which results in regenerated cartilage. This approach can theoretically be applied to the manufacture of almost all organs and tissues except for organs such as the brain (3).

In this section, a review is given of the biomaterials and procedures used in the development of tissue-engineered scaffolds, including: (*1*) natural and synthetic biomaterials, (*2*) natural–synthetic hybrid scaffolds, (*3*) the fabrication methods and techniques for scaffolds, (*4*) the required physicochemical properties for scaffolds, and (5) cytokine-released scaffolds.

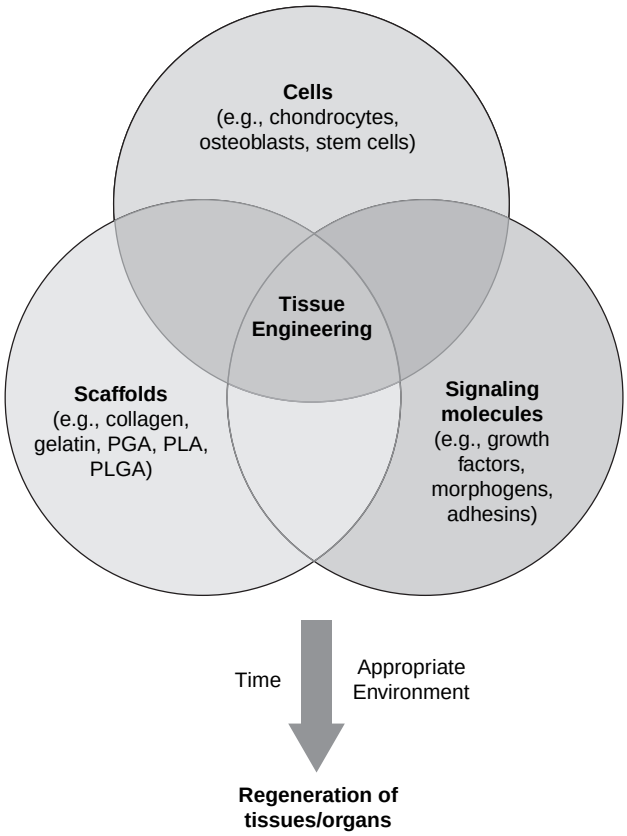


Figure 1. Tissue engineering triad. The combination of three key elements, cells, biomaterials, and signaling molecules, results in regenerated tissue-engineered neo-organs.

BIOMATERIALS FOR TISSUE ENGINEERING

The Importance of Scaffold Matrices in Tissue Engineering

Scaffolds play a very critical role in tissue engineering. Scaffolds direct the growth (*1*) of cells seeded within the porous structure of the scaffold, or (*2*) of cells migrating from surrounding tissue. Most mammalian cell types are anchorage dependent; the cells die if an adhesion substrate is not provided. Scaffold matrices can be used to achieve cell delivery with high loading and efficiency to specific sites. Therefore, the scaffold must provide a suitable substrate for cell attachment, cell proliferation, differentiated function, and cell migration. The prerequisite physicochemical properties of scaffolds are (*1*) to support and deliver the cells; (*2*) to induce, differentiate, and promote conduit tissue growth; (*3*) to target the cell-adhesion substrate, (*4*) to stimulate cellular response; (*5*) to create a wound healing barrier; (*6*) to be biocompatible and biodegradable; (*7*) to have relatively easy processability and malleability into the desired shapes; (*8*) to be highly porous with large surface–volume; (*9*) to have mechanical strength and dimensional stability; and (*10*) to have sterilizability (9–16). Generally, three-dimensional (3D) porous scaffolds can be fabricated from natural and synthetic polymers (Fig. 2 shows these chemical structures), ceramics, metal, and in a very few cases, composite biomaterials and cytokine-releasing materials.

Natural Polymers

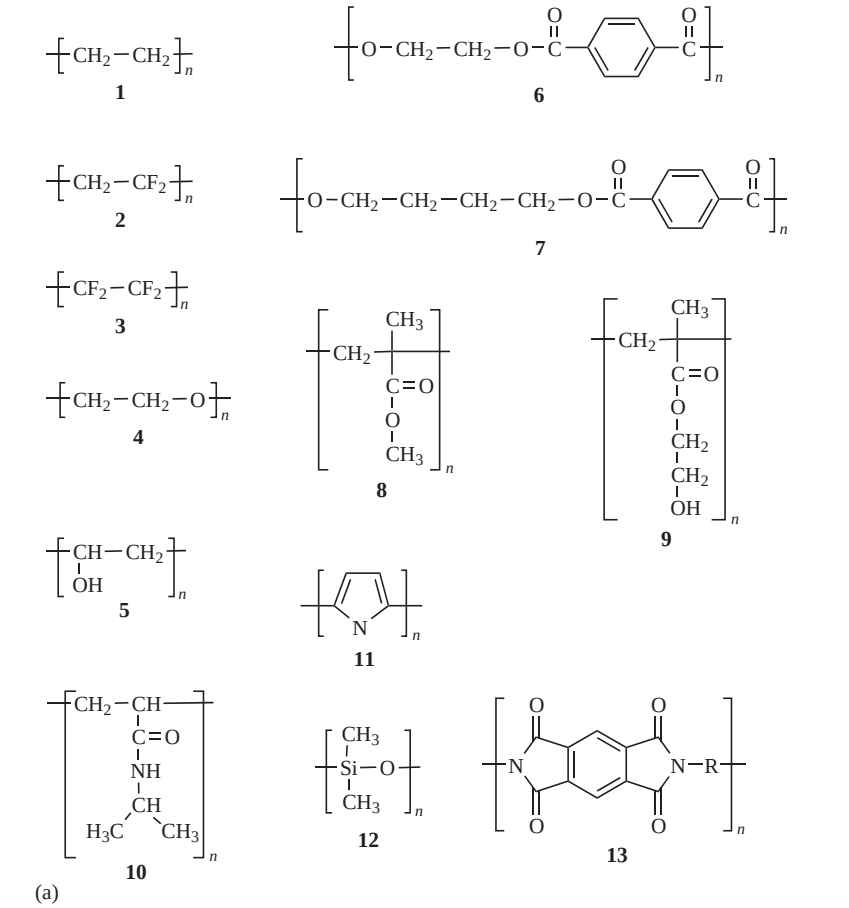
Many naturally occurring scaffolds can be used for tissue engineering purposes. One such example is the ECM, which is composed of very complex biomaterials and controls cell function. For the ECM used in tissue engineering, natural and synthetic scaffolds are designed to mimic specific function. The natural polymers used are alginate, proteins, collagens (gelatin), fibrins, albumin, gluten, elastin, fibroin, hyarulonic acid, cellulose, starch, chitosan (chitin), sclerolucan, elsinan, pectin (pectinic acid), galactan, curdlan, gellan, levan, emulsan, dextran, pullulan, heparin, silk, chondroitin 6-sulfate, polyhydroxyalkanoates, and others. Much of the interest in these natural polymers comes from their biocompatibility, relatively abundance and commercial availability, and ease of processing (17).

Alginate. Alginate (from seaweed) is composed of two repeating monosaccharides: L-guluronic acid and D-mannuronic acid. Repeating strands of these monomers form linear, water-soluble polysaccharides. Gelation occurs by interaction of divalent cations (e.g., Ca^{2+}, Mg^{2+}) with blocks of guluronic acid from different polysaccharide chains (as shown in Fig. 3). From this gelation property, the encapsulation of calcium alginate beads impregnated with various pharmaceutics, cytokines, or cultured cells, has been extensively investigated. Varying the preparation conditions of the gelation can control structure and physicochemical properties. Calcium alginate scaffolds do not degrade by hydrolytic reaction, whereas they can be degraded by a chelating agent such as ethyleneaminetetraaceticacid (EDTA) or by an enzyme. Also, the diffusion of calcium ions from an alginate gel can cause dissociation between alginate chains, which results in a decrease of mechanical strength over time. One of the disadvantages of an alginate matrix is a potential immune response and the lack of complete degradation, since alginate is produced in the human body (10). For these reasons, the chemical modification and incorporation of biological peptides, such as Arg-Gly-Asp cell adhesion peptides, have been used to improve the functionality and flexibility of natural scaffolds and their potential application (18).

Many researchers have studied the encapsulation of chondrocytes. Growth plate chondrocytes, fetal chondrocytes, and mesenchymal stem cells derived from bone marrow have been encapsulated in alginate (19). In each system, the chondrocytes demonstrated a differentiated phenotype, producing an ECM and retaining the cell morphology of typical chondrocytes. In addition, novel hybrid composites, such as alginate/agarose (a thermosensitive polysaccharide), alginate/fibrin, alginate/collagen and alginate/hyaruronic acid, and different gelling agents (water, sucrose, sodium chloride, and calcium sulfate) were investigated to optimize the advantages of each component material for tissue engineered cartilage (20–22). It was found that this hybrid material provides a reason why the microenvironments of composite materials affect chondrogenesis.

Collagen. At least 22 types of collagen exist in the human body. Among these, collagen types I, II, and III are the most abundant and ubiquitous. Conformation of the collagen chain consists of triple helices that are packed or processed into microfibrils. Molecularly, the three repeating amino acid sequences, such as glycine, proline, and hydroxyproline, form protein chains resulting in the formation of a triple helix arrangement. Type I collagen is the most abundant and is the major constituent of bone, skin, ligament, and tendon, whereas type II collagen is the collagen in cartilage. Collagen can promote cell adhesion as demonstrated by the Asp-Gly-Glu-Ala peptide in type I collagen, which functions as a cell-binding domain. Due to the abundance and ready accessibility of these tissues, they have been used frequently in the preparation of collagen (23).

The purified collagen materials obtained from either molecular or fibrillar technology are subjected to additional processing to fabricate the materials into useful scaffold types for specific tissue-engineered organs. Collagen can be processed into several types such as membrane (film and sheet), porous (sponge, felt, and fiber), gel, solution, filamentous, tubular (membrane and sponge), and composite matrix for the application of tissue repair, patches, bone and cartilage repair, nerve regeneration, and vascular and skin repair with/without cells (24). The Physicochemical properties of collagen can be improved by the addition of a variety of homogeneous and heterogeneous composites. Homogeneous composites can be formed between ions, peptides, proteins, and polysaccharides in a collagen matrix by means of ionic and covalent bonding, entrapment, entanglement, and coprecipitation. Heterogeneous composites, such as collagen–synthetic polymers, collagen–biological polymers, and collagen–ceramic hybrid



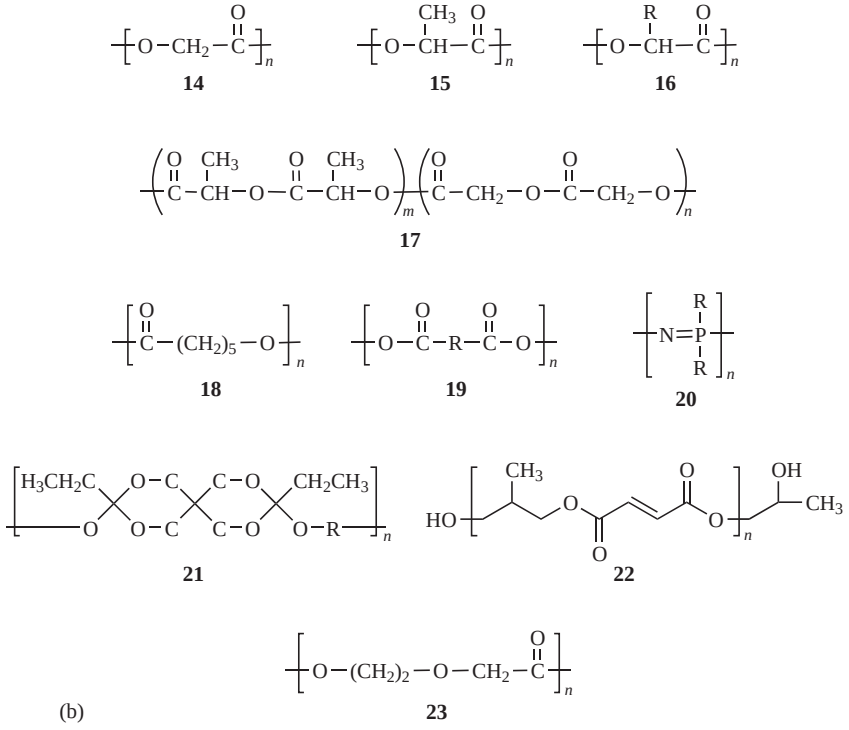
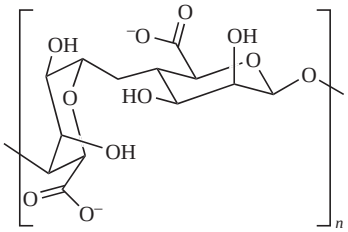
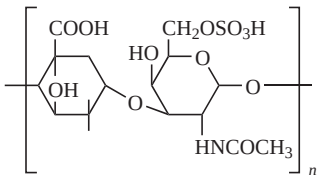


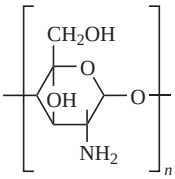
Figure 2. Chemical structures of some commonly used biodegradable and nondegradable polymers in tissue engineering. (a) Synthetic nondegradable polymers: (**1**). polyethylene, (**2**). poly(vinylidene fluoride), (**3**). polytetrafluoroethylene, (**4**). poly-(ethylene oxide), (**5**). poly(vinyl alcohol), (**6**). poly-(ethyleneterephthalate), (**7**). poly(butyleneterethphalate), (**8**). poly(methylmethacrylate), (**9**). poly- (hydroxymethylmetacrylate), (**10**). poly(*N*-isopropylacrylamide), (**11**). polypyrrole, (**12**). poly(dimethyl siloxane), and (**13**). polyimides. (b) Synthetic biodegradable polymers: (**14**). poly(glycolic acid), (**15**). poly(lactic acid), (**16**). poly(hydroxyalkanoate), (**17**). poly(lactide-*co*-glycolide), (**18**). poly(ε-caprolactone), (**19**). polyanhydride, (**20**). polyphsphazene, (**21**). poly(orthoester), (**22**). poly-(propylene fumarate), and (**23**). poly(dioxanone). (c) Natural polymers: (**24**). alginate, (**25**). chondroitin-6-sulfate, (**26**). chitosan, (**27**). hyarunonan, (**28**). collagen, (**29**). polylysine, (**30**). dextran, and (**31**). heparin. (d) PEO-based hydrogels: (**32**). Pluronic, (**33**). Pluronic R, (**34**). Tetronic, and (**35**). Tetronic R.



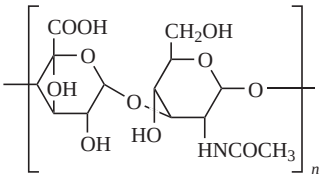
24



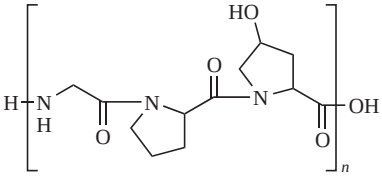
25



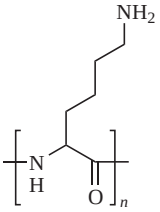
26



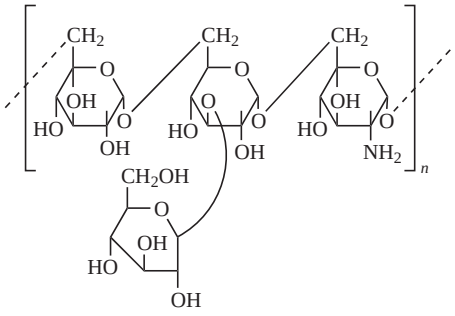
27



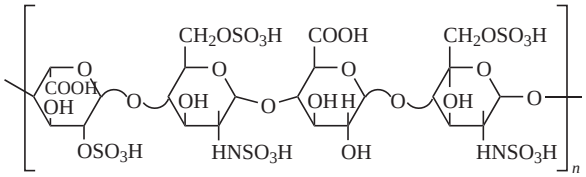
28



29

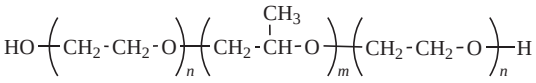


30

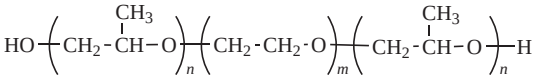


31

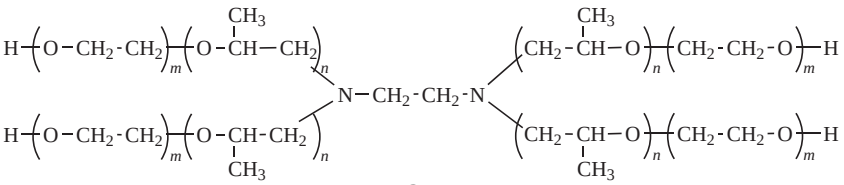
(c)



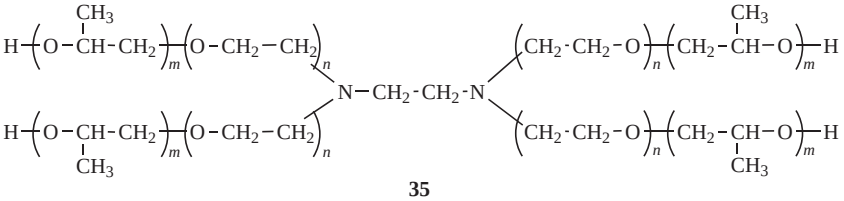
32



33



34



35

(d)

Figure 2. (*continued*)

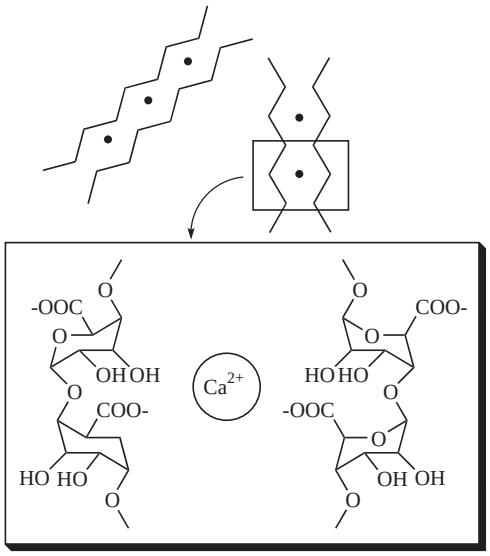


Figure 3. Schematic representation of the guluronate junction zone in alginate; eggbox model. The circles represent calcium ions.

polymers (collagen–nano-hydroxyapatite and collagen–calcium phosphate) have been investigated for use in tissue-engineered products (10).

Fibrin. Fibrin plays a major role during wound healing as a hemostatic barrier to prevent bleeding and to support a natural scaffold for fibroblasts. Actual polymerization is triggered by the conversion of fibrinogen to fibrin monomer by thrombin, and gelation occurs within 30–60 s. One advantage of using fibrin in this manner is its ability to completely fill the defect by gelling in situ. Fibrin sealant composed of fibrinogen and thrombin in addition to antifibronolytic agents has been used already in such surgical applications as sealing lung tears, cerebral spinal fluid leaks, and bleeding ulcers, because of its natural role in wound healing. Fibrin sealant might be made from autologous blood or from recombinant proteins (22). Fibrin gels can degrade either through hydrolytic or proteolytic means. Fibrinogen is commercially available from several manufacturers, so the cost of the fabrication of fibrin gels is relatively low. Recently, much work has been done to develop fibrin as a potential tissue-engineered scaffold matrix, especially for cartilage, which is formed from a fibrin/chondrocyte construct. Biochemical and mechanical analysis has demonstrated its cartilage-like properties. In neural tissue engineering, fibrin modified the incorporation of bioactive peptide in fibrin gels (25). Also, fibrin/hydroxyapatite hybrid composites have been investigated to optimize the mechanical strength of tissue-engineered subchondral bone substitutes.

Hyaluronan. Hyaluronic acid, a natural glycosaminoglycans polymer, can be found in abundance within cartilaginous ECM. It has some disadvantages in its natural form, such as high water solubility, fast resorption, and fast tissue clearance times, which are not conducive to biomaterials. To overcome these undesirable characteristics, chemical modifications were made to increase biocompatibility, tailor the degradation rate, control water solubility, and to fit the mechanical property. To increase hydrophobicity, esterification was carried out to increase the hydrocarbon content of the added alcohol, which resulted in tailored degradation rates since hydrophobicity directly influences hydration and the deesterification reaction (10). Another approach, the condensation reaction between the carboxylic group of unmodified hyaluronan molecules with the hydroxyl group of other hyalunonic acid molecules, was used to fabricate the sponge form. Then, bone marrow-derived mesenchymal progenitor cells were seeded to induce chondrogenesis and osteogenesis on this scaffold. Results from animal studies indicate that modified hyaluronic acid can successfully support mesenchymal stem cell proliferation and differentiation for osteochondral application (15). Also, a sulfate reaction on a hyaluronan gel created a variety of sulfate derivatives, ranging from one-to-four sulfate groups per disaccharide subunit. A crosslinking network hydrogel can be formed by using diamines from individual hyaluronic acid chains. Chondrocytes seeded on sulfated hyaluronic acid hydrogels appear to have good cell compatibility with the tissue-engineered cartilage. The benzyl ester hyaluronan products HYAFF-11 and LaserSkin (Fidia Advanced Biopolymers, FAB, Abano Terme, Italy) have been introduced to engineer skin bilayers *in vitro* (26).

Chitosan. Chitosan, a polysaccharide derived from chitin, is composed of a simple glucosamine monomer and has physicochemical properties similar to many glycosaminoglycans. Chitosan is relatively biocompatible and biodegradable; it does not evoke a strong immune response. It is relatively cheap due to its abundance and good reactivity with diverse methods of chemical processing. Chitin is typically extracted from arthropod shells by means of acid–alkali treatment to hydrolyze acetamido groups from the *N*-acetylglucosamine resulting in the production chitosan. It has a molecular weight of 800,000–1,500,000 $g \cdot mol^{-1}$ and dissolves easier than the native chitin polymer (27). For its use in the tissue-engineered cartilage, a 3D composite, such as a chondroitin sulfate A/chitosan hydrogel scaffold, was prepared. This hydrogel supported a differentiated phenotype of seeded articular chondrocytes and type II collagen and proteoglycan production (28). Also, the organic–inorganic hybrid scaffold, used as a chitosan/tricalcium phosphate scaffold, was fabricated for tissue-engineered bone. When osteoblast cells collected from rat fetal calvary were seeded onto a chitosan/tricalcium phosphate scaffold, the cells proliferated in a multiplayer manner and deposited a mineralized matrix (29).

Agarose. Agarose is another type of marine source polysaccharide purified extract from sea creatures, such as agar or agar-bearing algae. One of the unique properties of agarose is the formation of a thermally reversible gel, which starts to set at a concentration in excess of 0.1% at a temperature $\sim 40\,^{\circ}C$ and a gel melting temperature of $90\,^{\circ}C$. Agarose gel is widely used in the electrophoresis of proteins and nucleic acid. Its good gelling behavior could make it a suitable injectable bone substitute and cell

carrier matrix (17). Allogenic chondrocyte-seeded agarose gels have been used as a model to repair osteochondral defects in vivo. The repaired tissues were scored histologically based on the intensity and extent of the proteoglycan and the type II collagen immunoassay, the structural features of the various cartilaginous zones, integration with host cartilage, and the morphological features and arrangement of chondrocytic cells. The allogenic chondrocyte–agarose-grafted repairs had a higher semiquantitative score than control grafts. These results showed a good potential for use in tissue engineering (30). More detailed studies, such as the in vivo mechanical properties, biocompatibility and toxicity, and the balance degradation and synthesis kinetics of agarose-based tissue-engineered products, must be undertaken to further successful agarose applications (31).

Small Intestine Submucosa. Porcine small intestine submucosa (SIS) is an important material for natural ECM scaffolds (15). Many experiments have shown systematically that an acellular resorbable scaffold material, derived from SIS, is rapidly resorbed, supports early and abundant new blood vessel growth, and serves as a template for the constructive remodeling of several body tissues including musculoskeletal structures, skin, body wall, dura mater, urinary bladder, and blood vessels (32). The SIS material consists of a naturally occurring ECM, rich in components that support angiogenesis, including fibronectin, glycosaminoglycans including heparin, several collagens (including types I, III, IV, V, and VI), and angiogenic growth factors such as basic fibroblast growth factor and vascular endothelial cell growth factor (33). For these reasons, SIS scaffolds have been successfully used to reconstruct the urinary bladder, for vascular grafts, to reconstruct cartilage and bone alone or as a composite with synthetic polymers and inorganic biomaterials (34).

Acellular Dermis. Acellular human skin, that is skin removed of all cellular components, may be one of the most significant ECMs. An acellular dermis can be seeded with fibroblasts and keratinocytes to fabricate a dermal–epidermal composite for the regeneration of skin. AlloDerm (LifeCell, Branchburgh, NJ) is a typical commercialized product, a split-thickness acellular allograft prepared from human cadaver skin and cryopreserved for off-shelf use (35). Alloderm has been successful in the treatment of burn patients because of its nonantigenic dermal scaffold that includes elastin, proteoglycan, and basement membrane.

Poly(hydroxyalkanoates). Poly(hydroxyalkanoates) are entirely natural and are obtained from the microorganism *Alcaligen eutrophus* as Gram-negative bacteria. The physical properties of polyhydroxybutyrate (PHB) are similar to nondegradable polypropylene. Its copolymers with hydroxyvalerate [poly(hydroxybutylate-*co*-hydrovalerate); PHBV] have a modest range of mechanical properties and a correspondingly modest range of chemical compositions for monomers and processing conditions. Due to their good processability, these polymers can be manufactured into many forms, such as fibers, meshes, sponges, films, tubes, and matrices through standard processing techniques.

The family of poly(hydroxyalkanoates) does not appear to cause any acute inflammation, abscess formation, or tissue necrosis in whethers in the form of nonporous disks or cylinders, adjacent tissues (36). To optimize the mechanical property of PHBV, organic–inorganic hybrid composites such as PHBV–hydroxyapatite were developed for the tissue engineering of bone; hydroxyapatite promotes osteoconductive activity (13). Also, Schwann cell-seeded PHB was applied to regenerate a nerve in the shape of a conduit to guide and induce neonerve tissue at the nerve ends. Good nerve regeneration in PHB conduits as compared to nerve grafts was observed. The shape, mechanical strength, porosity, thickness, and degradation rate of PHB and its copolymers can be engineered.

Other Natural Polymers. Excluding those polymers discussed in the Natural Polymers section above, other natural polymers, are proteins, albumin, gluten, elastin, fibroin, cellulose, starch, sclerolucan, elsinan, pectin (pectinic acid), galactan, curdlan, gellan, levan, emulsan, dextran, pullulan, heparin, silk, and chondroitin 6-sulfate. Although they are not discussed here, these biopolymers are of interest because of their unusual and useful functional properties as well as their abundance. This group of natural polymers are (*1*) biocompatible and nontoxic, (*2*) easily processed as film and gel, (*3*) heat stable and thermal processable over a broad temperature range, and (*4*) water soluble (17). *In vivo* and *in vitro* experiments, and physicochemical modifications should be performed in the near future to promote the use of these natural polymers in tissue-engineered scaffolds.

Synthetic Polymers

Natural polymers are not used more extensively because they are expensive, differ from batch to batch, and there is a possibility of cross-contamination from unknown viruses or unwanted diseases due to their isolation from plant, animal, and human tissue. Alternatively, synthetic polymeric biomaterials have easily controlled physicochemical properties and quality, and no immunogenecity. Also, they can be processed by various techniques and supplied consistently in large quantities. To adjust the physical and mechanical properties of a tissue-engineered scaffold at a desired place in the human body, the molecular structure, and molecular weight are adjusted during the synthetic process. Synthetic polymers are largely divided two categories: biodegradable, and nonbiodegradable. Some nondegradable polymers include poly(vinylalcohol) (PVA), poly(hydroxylethylmethacryalte), and poly(*N*-isopropylacryamide). Some synthetic degradable polymers are in the family of poly(α-hydroxy ester)s, such as polyglycolide (PGA), polylactide (PLA) and its copolymer poly(lactide-*co*-glycolide) (PLGA), polyphosphazene, polyanhydride, poly(-propylene fumarate), polycyanoacrylate, polycaprolactone, polydioxanone and biodegradable polyurethanes.

Between these two polymers, synthetic biodegradable polymers are preferred for use in tissue-engineered scaffolds because they have minimal chronic foreign body reactions and they promote the formation of completely natural tissue. That is, they can form a temporary scaffold

for mechanical and biochemical support. More detailed polymer fabrication methods are discussed in the section, Scaffold Fabrication and Characterization.

Poly(α-Hydroxy Ester)s. The family of poly(α-hydroxy acid)s, such as PGA, PLA, and its copolymer PLGA, are among the few synthetic polymers approved for human clinical use by the U.S. Food and Drug Administration (FDA). These polymers are extensively used or tested as scaffold materials, because they are as bioerodible with good biocompatibility, have controllable biodegradability, and relatively good processability (37). This family of poly(α-hydroxy éster)s has been used for three decades: PGA as a suture; PLA in bone plate, screw and reinforced materials; and PLGA in surgical and drug delivery devices. The safety of these materials has been prove for many medical applications (38–47).

These polymers degrade by nonspecific hydrolytic scission of their ester bonds. Polyglycolide biodegrades by a combination of hydrolytic scission and enzymatic (esterase) action producing glycolic acid, which can either enter the tricarboxylic acid (TCA) cycle or be excreted in urine and eliminated as carbon dioxide and water. The hydrolysis of PLA yields lactic acid, which is a normal byproduct of anaerobic metabolism in the human body and is incorporated in the TCA cycle to be excreted finally by the body as carbon dioxide and water. With the addition of a methyl group to glycolide, PLA is much more hydrophobic than the highly crystalline PGA. As a result, PLA has a slower degradation rate over a year's time. The degradation time of PLGA as a copolymer can be controlled from weeks to over a year by varying the ratio of monomers, its molecular weight, and the processing conditions. The synthetic methods and physicochemical properties, such as melting temperature, glass transition temperature, tensile strength, Young's modulus, and elongation, were reviewed elsewhere (48).

The mechanism of biodegradation of poly(α-hydroxy acid)s is bulk degradation, which is characterized by a loss in polymer molecular weight, while its mass is maintained. Mass maintenance is useful for tissue-engineering applications that require a specific shape. However, a loss in molecular weight causes a significant decrease in mechanical properties. Degradation depends on its chemical history, porosity, crystallinity, steric hindrance, molecular weight, water uptake, and pH. Degradable products, such as lactic acid and glycolic acid, decrease the pH in the surrounding tissue resulting in inflammation and potentially poor tissue development. The PGA, PLA, and PLGA scaffolds are applied for the regeneration of all tissue, including skin, cartilage, blood vessel, nerve, liver, dura mater, bone, and other tissue (10,12,17). For the application of these polymers as scaffolds, the development of fabrication methods for porous structures is also important.

The hybrid structure of chondrocytes and fibroblast/PGA fiber felts was successfully tested in the regeneration of cartilage and skin, respectively (49). Also, porous PLGA scaffolds with an average pore sizes of 150–300 or 500–710 μm were seeded with osteoblast cells, which resulted in good bone generation. Composites of PLA/tricalcium phosphate and PLA/hydroxyapatite were attempted to induce bone formation both *in vitro* and *in vivo* (13,50). Porous PLA tubes with an inside diameter of 1.6 mm, are outside diameter of 3.2 mm, and lengths of 12 mm, were implanted into 12 mm gaps in the rat sciatic nerve model. Compared to control grafts, both the number and density of axons were significantly less for the tabulated implants. The PGA tube was also tested for the regeneration of vascular grafts, and showed good *in vivo* results.

To improve the physicochemical properties of poly(α-hydroxy acid)s for use as scaffold materials, the chemical modification of both end groups of PLA and PGA was undertaken; the additional reaction of the moieties helps to control the biological and/or physical properties of biomaterials (17). For example, poly(lactic acid-*co*-lysine-*co*-aspartic acid) (PLAL-ASP) was synthesized to add a cell adhesion property. Similarly, a copolymer of lactide and ε-caprolactone was synthesized to improve the elastic property of PLA. The PLA-poly(ethylene oxide) (PEO) copolymers were synthesized to have the degradative and mechanical properties of PLA and the biological control offered by PEO and its functionalization (51). One of the unique characteristics of PLA-PEO block copolymers is its temperature sensitivity. Because of the hydrophobicity of PLA and hydrophilicity of PEO, the sol–gel property can be applied to injectable cell carriers. Also, a nano-hybrid composite with other materials has been developed for application to all organs in the body.

Poly(vinyl Alcohol). Poly(vinyl alcohol) is synthesized from poly(vinyl acetate) by saponification. The result is a hydrogel that contains some water, which is similar to cartilage. It is relatively biocompatible, swells with a large amount of water, easily sterilized, and easily fabricated and molded into desired shapes. It has a reactive pendant alcohol group that can be modified by chemical cross-linking, physical cross-linking, or by incorporating an acrylate group, which results in improvement of its mechanical properties. A typical commercialized PVA gel is Salubria (Salumedica, Atlanta, GA), which was created by completing a series of freeze–thaw cycles with PVA polymers and 0.9% saline solution. By changing the ratio of PVA and H_2O, the molecular weight of PVA, and the quantity and duration of the freeze–thaw cycles, the physical properties of the PVA hydrogel can be controlled. Poly(vinyl alcohol) has been used in cartilage regeneration; it has similar mechanical properties needed in breast augmentation, diaphragm replacement, and bone replacement (10). One significant drawback is that it is not fully biodegradable because of the lack of labile bonds within the polymer backbone. So, it is recommended that low molecular weight PVA, $\sim 15{,}000\ \mathrm{g \cdot mol^{-1}}$, which can be absorbed through the kidney, might be applied to tissue-engineered scaffolds.

Polyanhydride. Polyanhydride is synthesized by the reaction of diacids with anhydride to form acetyl anhydride prepolymers. High molecular weight anhydrides are synthesized from the anhydride prepolymer in a melt condensation. Polyanhydrides are modified to increase their physical properties by a reaction with imides (17). A typical example of this is copolymerization with an

aromatic imide monomer that results in the polyanhydride-*co*-imide used in hard tissue engineering. To control degradability and to enhance mechanical properties, photo-crosslinkable functional groups were introduced by the substituted methacrylate groups on polyanhydrides for orthopedic tissue engineering (48,50). The degradation mechanism of polyanhydrides is a highly predictable and controlled, surface erosion whereas that of poly(α-hydroxy ester) is bulk erosion. To optimize the degradation behavior of anhydride-based copolymers, the polymer backbone chemistry needs to be controlled to achieve a ratio of monomer and molecular weight.

Poly(Propylene Fumarate). Poly(propylene fumarate) and its copolymer, a biodegradable and unsaturated linear polyester, were synthesized as potential scaffold biomaterials. The degradation mechanism is a hydrolytic chain scission similar to poly(α-hydroxy ester). The mechanical strength and degradable behaviors were controlled by crosslinking with a vinyl monomer at the unsaturated double bonds. The physical properties are enhanced by a composite with degradable bioceramic β-tricalcium phosphate, which is used as injectable bone (52). Copolymerization of propylene fumarate with ethylene glycol can be made elastic with poly(propylene fumarate) and used as a cardiovascular stent. New materials for propylene fumarate polymers are continually being investigated through copolymer synthesis, hybrid composites, and blends.

PEO and Its Derivatives. Poly(ethylene oxide) is one of the most important and widely used polymers in biomedical applications because of its excellent biocompatibility (51,53,54). It can be produced by anionic or cationic polymerization from ethylene oxide by initiators. Poly(ethylene oxide) is used to coat materials used in medical devices to prevent tissue and cell adhesion, as well as in the preparation of biologically relevant conjugates, and in induction cell membrane fusion. These PEO hydrogels can be fabricated by crosslinking reactions which gamma rays, electron beam irradiation, or chemical reactions. This hydrogel can be used for drug delivery and tissue engineering. Vigilon (C.R. Bard, Inc., Murray Hill, NJ) is a radiated crosslinked, high molecular weight PEO, which swells with water and is used as a wound-covering material. The hydroxyl in the glycol end group is very active, making it appropriate for chemical modification. The attachment of bioactive molecules, such as cytokines and peptides to PEO or poly(ethylene glycol) (PEG) promotes the efficient delivery of bioactive molecules. See the section, Cytokine Release System for Tissue Engineering, for a more detailed explanation.

To synthesize biodegradable PEO, block copolymerization with PGA or PLA degradable units has been carried out. The hydrogel can be polymerized into two- or three-block copolymers such as PEO-PLA, PEO-PLA-PEO, and PLA-PEO-PLA. For the biodegradable block, ε-caprolactone, δ-valerolactone, and PLGA can be used (50). A characteristic of this series of hydrogels is a temperature-sensitive phenomena. A solid state at room temperature changes to a gel state at body temperature. Hence, biodegradable hydrogels are very useful in injectable cell loading scaffolds (55). After injection of the chondrocyte cell hybrid structure and biodegradable hydrogels, the hydrogels degrade in vivo and neocartilage tissue remains.

Also, the copolymers of PEO and poly(propylene oxide) (PPO), including PPO-PEO-PPO or PEO-PPO-PEO block copolymers, are the basis for the commercially available Pluronics and Tetronics. Pluronics form a thermosensitive gel by shrinking hydrophobic segments of the copolymer PPO (54). The physicochemical property of the hydrogel can be varied with the composition and structure of the ratio of PPO and PEO. Some have been approved by the FDA and EPA for use as food additives, pharmaceutical ingredients, and agricultural products. Although the polymer is not degraded by the body, the gels dissolve slowly and the polymer is eventually cleared. Chondrocytes-loaded Pluronics, when directly injected at the injured site containing tissue-engineered cartilage, maintained its original shape in the developing neocartilage (56). Also, these polymers are used in the treatment of burn patients and for protein delivery. The advantages of these injectable hydrogels include: (*1*) no need for surgical intervention, (*2*) easy pore-size manipulation, and (*3*) no need for complex shape fobrication.

Polyphosphazene. Polyphossphazene consists of an inorganic backbone of alternating single and double bonds between phosphorous and nitrogen atoms, while most of the polymer is made up of a carbon–carbon organic backbone (10,12,17). It has side groups that can react with other functional groups which result in block or star polymers. Biological and physical properties can be controlled by the substitution of functional side groups. For example, the rate of degradation can be varied by controlling the proportion of hydrolytically labile side groups. The wettability such as hydrophilicity, hydrophobicity, and amphiphilicity, of polyphosphazene might be dependent on the properties of the side group. It can be made into films, membranes, and hydrogels for scaffold applications by cross-linking or grafting modifications (48). The cytocompatibility of highly porous polyphosphazene scaffolds offers possibilities for skeletal tissue engineering. Also, the blend of polyphosphazene with PLGA may be modified and its miscibility and degradability determined (57).

Biodegradable Polyurethane. Polyurethane is one of the most widely used polymeric biomaterials in biomedical fields due to its unique physical properties, such as durability, elasticity, elastomer-like character, fatigue resistance, compliance, and tolerance. Moreover, the reactivity of the functional group of the polyurethane backbone can be achieved by the attachment of biologically active biomolecules and the adjustment of their hydrophilicity–hydrophobicity (58). Recently, the synthesis of a new generation of nontoxic biodegradable peptide-based polyurethanes was achieved. Typical biodegradable polyurethane is composed of an amino acid-based hard segment (such as lysine diisocyanate), a polyol soft segment (such as a hydroxyl doner-like polyester), and sugar (59). Hence, the degradation products of these nontoxic lysine diisocyanate-based urethane polymers are nontoxic lysine and the polyol. If the covalent bonding of various proteins, such

as cytokines, growth factors, and peptides, are introduced in the polymer backbone, the controlled release of the bioactive molecules can be achieved in a degradable manner using polyurethane scaffolds. The mechanisms of degradation are hydrolysis, oxidation—both thermal, and enzymatic. Both the chemistry and the composition of soft and hard segments play an important role in the degradability of polyurethane. Poly(urethane-urea) matrices with lysine diisocyanate as the hard segment and glucose, glycerol, or PEG as the soft segments have been studied. In the application of biodegradable polyurethane as a scaffold various types of cells, such as chondrocytes, bone marrow stromal cells, endothelial cells, and osteoblast cells, were successfully adhered and proliferated. Also, toxicity, induction of a foreign body reaction, and antibody formation were not observed in in vivo experiments. The long-term safety and biocompatibility of biodegradable polyurethane must be continuously monitored for use in tissue-engineered scaffold substrates.

Other Synthetic Polymers. Besides the synthetic polymers already introduced in the above sections, many other synthetic polymers, either degradable or nondegradable, are being developed and tested to mimic the natural tissue and wound-healing environment. Examples are poly(2-hydroxyethylmethacrylate) hydrogel, injectable poly(*N*-isopropylacrylamide) hydrogel, and polyethylene for neocartilage; poly(iminocarbonates) and tyrosine-based poly(iminocarbonates) for bone and cornea; crosslinked collagen–PVA films and an injectable biphasic calcium phosphate–methylhydroxypropylcellulose composite for bone regeneration materials; a polyethylene oxide-*co*-polybutylene terephthalate for bone bonding; poly(orthoester) and its composites with ceramics for tissue-engineered bone; synthesized conducting polymer polypyrrole–hyaruronic acid composite films for the stimulation of nerve regeneration; and peptide-modified synthetic polymers for the stimulation of cell and tissue.

It is very important for the design and synthesis of more biodegradable and biocompatible scaffold biomaterials to mimic the natural ECM in terms of bioactivity, mechanical properties, and structures. The more biocompatible biomaterials tend to elicit less of an immune response and reduce an inflammatory response at the implantation site.

Bioceramic Scaffolds

Bioceramic is a term used for biomaterials that are produced by sintering or melting inorganic raw materials to create an amorphous or a crystalline solid body that can be used as an implant. Porous final products have been used mainly as scaffolds. The components of ceramics are calcium, silica, phosphorous, magnesium, potassium, and sodium. Bioceramics used in tissue engineering might be classified as nonresorbable (relatively inert), bioactive, or surface active (semi-inert), and biodegradable or resorbable (non-inert). Alumina, zirconia, silicone nitride, and carbons are inert bioceramics. Certain glass ceramics are dense hydroxyapatites [$9CaO \cdot Ca(OH)_2 \cdot 3P_2O_5$] and semi-inert (bioactive). Calcium phosphates, aluminum–calcium–phosphates, coralline, tricalcium phosphates ($3CaO \cdot P_2O_5$), zinc-calcium-phosphorous oxides, zinc-sulfate-calcium-phosphates, ferric–calcium–phosphorous–oxides, and calcium aluminates are resorbable ceramics (60). Among these bioceramics, synthetic apatite and calcium phosphate minerals, coral-derived apatite, bioactive glass, and demineralized bone particles are widely used in the hard tissue engineering area, hence, they will be discussed in this section.

Synthetic crystalline calcium phosphate can be crystallized into salts such as hydroxyapatite and β-whitlockite, depending on the Ca/P ratio. These salts are very tissue compatible and are used as bone substitutes in a granular, sponge form or as a solid block. The apatite formed with calcium phosphate is considered to be closely related to the mineral phase of bone and teeth. The chemical composition of crystalline calcium phosphate is a mixture of $3CaO \cdot P_2O_5$, $9CaO \cdot Ca(OH)_2 \cdot 3P_2O_5$ and calcium pyrophosphate ($4CaO \cdot P_2O_5$). The active exchange of ions occurs on the surface and leads to the exchange composition of minerals (9,61). When porous ceramic scaffolds were implanted in the body, both with or without cells for tissue-engineered bone, the delivery of some elements to the new bone was at the interface between the materials and the osteogenic cells.

Tricalcium phosphate is the rapidly resorbable calcium phosphate ceramic resulting in resorption 10–20 times faster than hydroxyapatite (13). Porous tricalcium phosphate may stimulate local osteoblasts for new bone formation. Injectable calcium phosphate cement containing β-tricalcium phosphate, dibasic dicalcium phosphate, and tricalcium phosphate monoxide, was investigated for the treatment of distal radius fractures. Calcium sulfate hemihydrate (plaster of Paris), as a synthetic graft material, was also tested for tissue-engineered bone.

Coral-derived apatite (Interpore; Interpore international, Irvine, CA) is a natural substance made by marine vertebrate (62). The porous structure of coral, which is structurally similar to bone, is a unique physicochemical property that promotes its use as a scaffold matrix for bone. The main component of natural coral is calcium carbonate or aragonite, the metastable form of calcium carbonate. This compound can be converted to hydroxyapatite by a hydrothermal exchange process, which results in a mixture of hydroxyapatites, $9CaO \cdot Ca(OH)_2 \cdot 3P_2O_5$, and fluoroapatite, $Ca_5(PO_4)_3F$. For tissue-engineered bone, the hybrid structure of porous coral-derived scaffolds and mesenchymal stem cells were demonstrated in vitro. The results showed the differentiation of bone marrow derived from stem cells to osteoblasts; successive mineralizations were successfully accomplished (63).

Glass ceramics are polycrystalline materials manufactured by controlled crystallization of glasses using nucleating agents, such as small amounts of metallic agent Pt groups, TiO_2, ZrO_2, and P_2O_5, which result in a fine-grained ceramic that possesses excellent mechanical and thermal properties (60,61). Typical bioglass ceramics developed for implantations are SiO_2-CaO-Na_2O-P_2O_5 and Li_2O-ZnO-SiO_2 systems. These bioglass scaffolds are suitable for inducing direct bonding with bone. Bonding to bone is related to the composition of each component.

One significant natural bioactive material is the demineralized bone particle, which is a powerful inducer of new

bone growth (38,41). Demineralized bone particles contain many kinds of osteogenic and chondrogenic cytokines such as bone morphogenetic protein, and are widely used as filling agent for bony defects. Because of their improved availability through the tissue bank industry, demineralized bone particles are widely used in clinical settings. To achieve more optimal results in the application of demineralized bone particles to tissue engineering, nanohybridization with synthetic (PLGA/demineralized bone particle hybrid scaffolds) and with natural organic compounds (collagen/demineralized bone particle hybrid scaffolds), has been carried out.

Porosity—the size of the mean diameter and the surface area—is a critical factor for the growth and migration of tissue into bioceramic scaffolds (60). Several methods have been introduced to optimize the fabrication of porous ceramics, such as dip casting, starch consolidation, the polymeric sponge method, the foaming method, organic additives, gel casting, slip casting, direct coagulation consolidation, hydrolysis-assisted solidification, and freezing methods. Therefore, it is very important to choose an appropriate method of preparation based on the physical properties of the desired organs.

A CYTOKINE-RELEASE SYSTEM FOR TISSUE ENGINEERING

Growth factors, a type of cytokine, are polypeptides that transmit signals to modulate cellular activity and tissue development including cell patterning, motility, proliferation, aggregation, and gene expression. As in the development of tissue-engineered organs, regeneration of functional tissue requires maintenance of cell viability and differentiated function, encouragement of cell proliferation, modulation of the direction and speed of cell migration, and regulation of cellular adhesion. For example, transforming growth factor-β_1 (TGF-β_1) might be required to induce osteogenesis and chondrogenesis from bone marrow derived mesenchymal stem cells. Also, brain-derived neurotrophic factor (BDNF) can be enhanced to regenerate the spinal cord after injury. The easiest method for the delivery of growth factor is injection near the site of cell differentiation and proliferation (4). The most significant problems associated with the direct injection method are that the growth factors have a relatively short half-life, have a relatively high molecular weight and size, display very low tissue penetration, and have potential toxicity at systemic levels (4,10,11,16).

A promising technique for the improvement of their efficacy is to locally control the release of bioactive molecules for a specified release period to promote impregnation into a biomaterial scaffold. Through impregnation into the scaffold carrier, protein structure and biological activity can be stabilized to a certain extent, resulting in prolonging the release time at the local site. The duration of cytokine release from a scaffold can be controlled by the types of biomaterials used, the loading amount of cytokine, the formulation factors, and the fabrication process. The release mechanisms are largely divided into three categories: (*1*) diffusion controlled, (*2*) degradation controlled, and (*3*) solvent controlled. The mechanism of biodegradable scaffold materials was regulated by degradation control, whereas that of the nondegradable material was regulated by diffusion and/or solvent control. The desired release pattern, such as a constant, pulsatile, and time programed behavior over the specific site and injury can be achieved by the appropriate combination of these mechanisms. Also, the cytokine-release system's geometries and configurations can be altered to produce the necessary scaffold, tube, microsphere, injectable form or fiber (46,51,54).

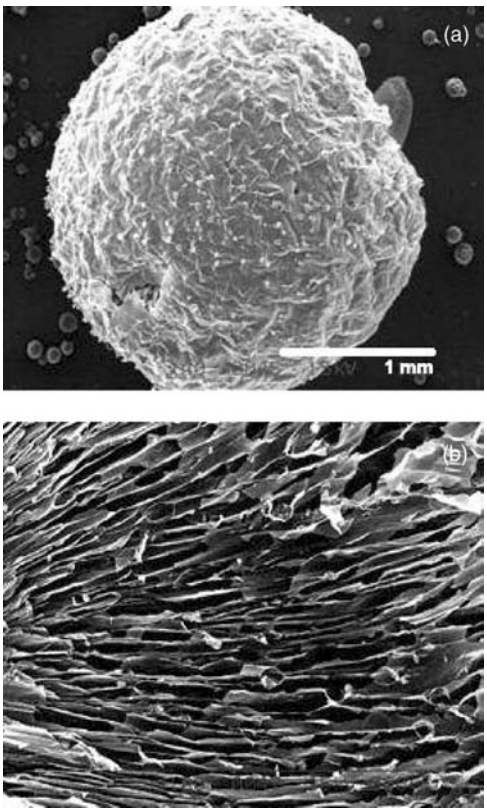


Figure 4. (a) Bone marrow-derived mesenchymal stem cells impregnated TGF-β_1 loaded alginate beads (original magnifications 40×), and (b) inner structure of alginate beads (original magnifications 100×).

Figures 4–6 show the TGF-β_1 loaded alginate bead and the release pattern of TGF-β_1 from alginate beads for the chondrogenesis from bone marrow-derived mesenchymal stem cells (64). The pore structure of 10 μm width and 100 μm length, was well suited to promote cell proliferation (Fig. 4); TGF-β_1 released at a near zero-order rate for 35 days (Fig. 5). By using the alginate bead with TGF-β_1 delivery system, chondrogenesis was successfully attained, as shown in Fig. 6.

To fabricate a new sustained delivery device for nerve growth factor (NGF), we developed NGF-loaded biodegradable PLGA films by a novel and simple sandwich solvent casting method for possible applications in the central nervous system (45). The release of NGF from the NGF-loaded PLGA films was prolonged $>$ 35 days with a zero-order rate, without initial burst, and controlled by variation of different molecular weights and different NGF loading amounts as shown in Fig. 7. After 7 days, NGF

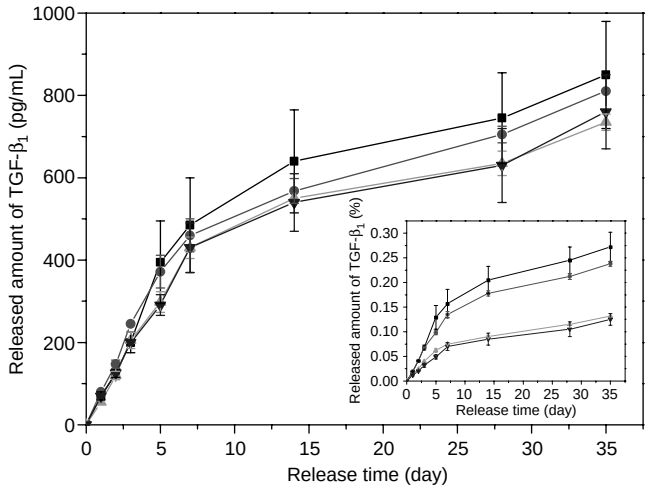


Figure 5. Release pattern of TGF-β_1 from TGF-β_1 loaded alginate beads; (■) 0.5 μg TGF-β_1, (●) 0.5 μg TGF-β_1 with heparin, (▲) 1.0 μg TGF-β_1, and (▼) 1.0 μg TGF-β_1 with heparin.

was released in a phosphate buffered saline solution (PBS; pH 7.0) and rat pheochromocytoma (PC-12) cells were cultured on the NGF-loaded PLGA film for 3 days. The released NGF stimulated neurite sprouting in the cultured

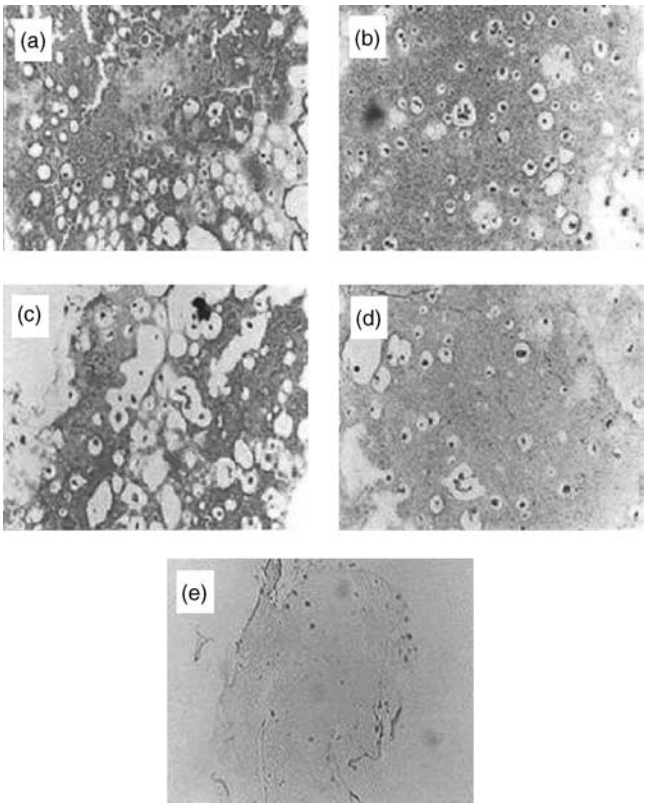


Figure 6. Safranin-O staining of chondrogenesis cells from bone marrow-derived mesenchymal stemcells in alginate beads. We can observe typical chondrocyte cells in alginate beads; (a) 0.5 μg·mL^{-1} TGF-β_1, (b) 1.0 μg·mL^{-1} TGF-β_1, (c) 0.5 μg·mL^{-1} TGF-β_1 with heparin (d) 1.0 μg·mL^{-1} TGF-β_1 with heparin, and (e) control (without TGF-β_1) (Original magnification 100×).

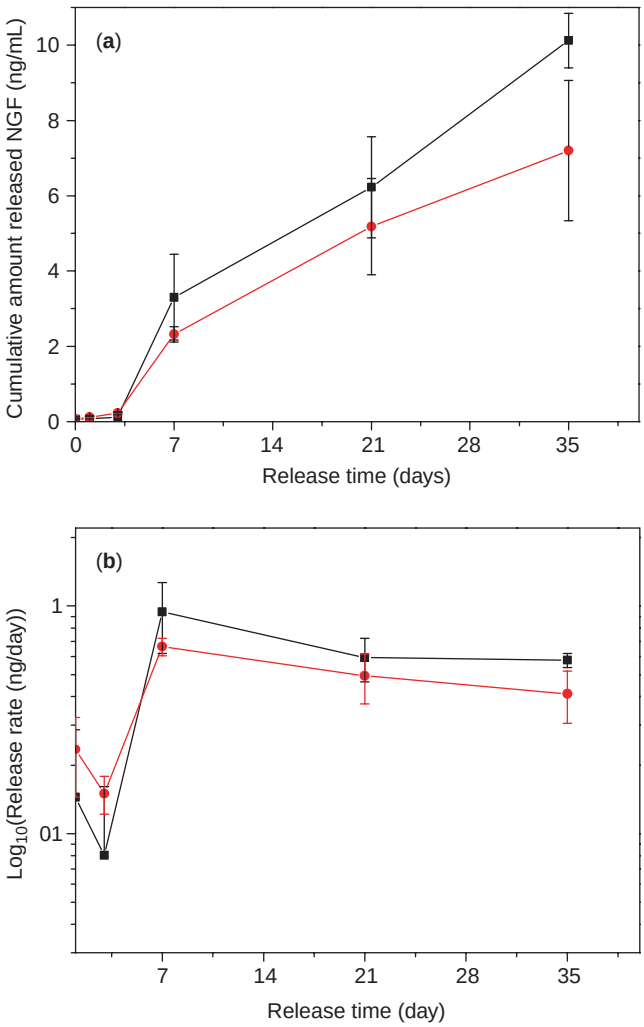


Figure 7. (a) Release profiles and (b) logarithmic plot of release rate for NGF from NGF-loaded PLGA films of 43,000 g/mol. (●) 25.4 ng, and (■) 50.9 ng NGF/cm^2 PLGA.

PC-12 cells; the remaining NGF in the NGF/PLGA film at 37°C for 7 days was still bioactive, as shown in Fig. 8. These studies suggest that NGF-loaded PLGA sandwich film can be released in the delivery system over the desired time period, thus, it can be a useful neuronal growth culture serving as a nerve contact guidance tube for applications in neural tissue engineering.

One serious problem during the fabrication of cytokine-loaded scaffolds is the denaturation and deactivation of cytokines, which result in loss of biological activity (65,66). Hence, the optimized method must be developed for stabilized cytokine-release scaffolds. For example, the release of NGF from a PLGA matrix was investigated using codispersants, such as polysaccharides (dextran) and proteins (albumin and β-lactoglobulin), with different molecular weights and charges. Negatively charged codispersants stabilized NGF in the PLGA system. Similarly, albumin stabilized epidermal growth factor (EGF) and heparin stabilized other growth factors.

Another available emerging technology is the "tethering of protein", that is, immobilization of protein on the surface

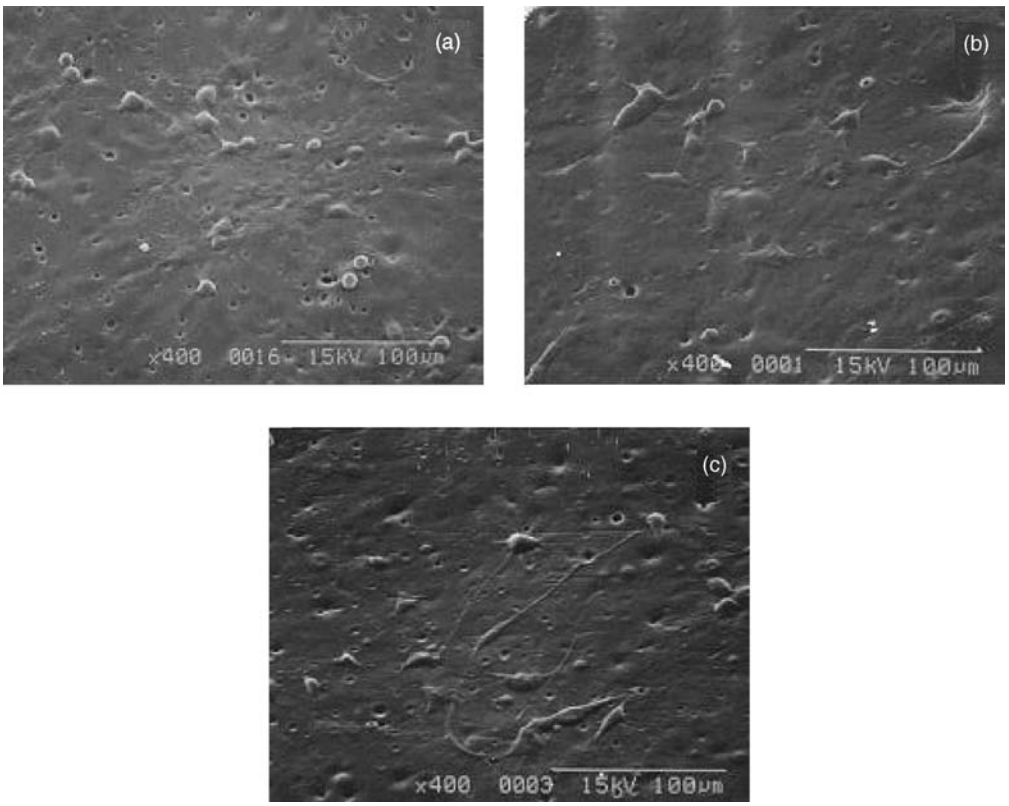


Figure 8. Effect of NGF released on neurites formation of PC-12 cells for 3-day cultivation on control (a) PLGA, (b) 25.4 ng, and (c) 50.9 ng NGF/cm^2 PLGA just after 7 days. There were total medium changes (Molecular weight of PLGA; 83,000 g/mol, original magnification; 400×).

of a scaffold matrix. Immobilization of insulin and transferrin to the poly(methylmetacrylate) films stimulates the growth of fibroblast cells compared to the same concentrations of soluble or physically adsorbed proteins (67). For the enhancement of cytokine activity, the PEO chain was applied as a short spacer between the surface of the scaffold and the cytokine. Tethered EGF, immobilized to the scaffold through the PEO chain, showed better DNA synthesis or cell rounding compared to the physically adsorbed EGF surface (68).

Conjugation of cytokine with an inert carrier prolongs the short half-life of protein molecules. Inert carriers are albumin, gelatin, dextran, and PEG. In PEGylation, PEG conjugated cytokine is most widely used for the release. This carrier appears to decrease the rate of cytokine degradation, attenuate the immunological response, and reduce clearance by the kidneys (69). Also, this PEGylated cytokine can be impregnated into scaffold materials by physical entrapment for sustained release. For example, the NGF-conjugated dextran (70,000 $g \cdot mol^{-1}$) impregnated polymeric device was implanted directly into the brain of adult rats. Conjugated NGF could penetrate into the brain tissue 8 times faster than the unconjugated NGF. This conjugation method can be applied to the delivery of proteins and peptides. Immobilized RGD (arginin-glycine-aspartic acid) and YIGSR (tyrosin-leucineglycine-serine-arginine), which are typical ECM proteins, can enhance cell viability, function, and recombinant products in the cell (70).

Gene-activating scaffolds are being designed to deliver the targeted gene that results in the stimulation of specific cellular responses at the molecular level (4,3,11). Modification of bioactive molecules with resorbable biomaterial systems obtain specific interactions with cell integrins resulting in cell activation. These bioactive bioglasses and macroporous scaffolds also can be designed to activate genes that stimulate regeneration of living tissue (9). Gene delivery would be accomplished by complexation with positively charged polymers, encapsulation, and gel by means of the scaffold structure (51). Methods of gene delivery for gene-activating scaffolds are almost the same methods as for those with protein, drug, and peptides.

SCAFFOLD FABRICATION AND CHARACTERIZATION

Scaffold Fabrication Methods

Engineered scaffolds may enhance the functionalities of cells and tissues to support the adhesion and growth of a large number of cells because they provide a large surface area and pore structure within a 3D structure. The pore structure needs to provide enough space, permit cell suspension, and allow penetration of the 3D structure. Also, these porous structures help to promote ECM production, transport nutrients from nutrient media, and excrete waste products (10,12,15). Therefore, an adequate pore size and a uniformly distributed, and an interconnected pore structure, which allow for easy distribution of cells

throughout the scaffold structure, are very important. Scaffold structures are directly related to their fabrication methods; over 20 methods have been proposed (10,71).

The most common and commercialized scaffold is the PGA nonwoven sheet (Albany International Research Co., Mansfield, MA; porosity $\sim 97\%$, ~ 1–5 mm thick); it is one of the most tested scaffolds for tissue-engineered organs. To stabilize dimensionally and provide mechanical integrity, fiber-bonding technology was developed using heat and PLGA or PLA solution spray coating methods (72).

Porogen leaching methods have been combined with polymerization, solvent casting, gas foaming, or compression molding of natural and synthetic scaffolds biomaterials. The leaching of pore-generating particles such as sodium chloride crystal, sodium tartrate, and sodium citrate were sieved using a molecular sieve (10,71). PLGA, PLA, collagen, poly(orthoester), or SIS-impregnated PLGA scaffolds were successfully fabricated into a biodegradable sponge structure by this method with $> 93\%$ porosity and a desired pore size of 1000 μm. By using the solvent casting/particulate leaching method, complex geometries, such as tube, nose, and specific organ types (e.g., nano-composite hybrid scaffolds), could be fabricated by means of conventional polymer-processing techniques, such as calendaring, extrusion, and injection. Complex geometry can be fabricated from porous film lamination (33,39,42,47). The advantage of this method is its easy control of porosity and geometry. However, the disadvantages include: (*1*) the loss of water-soluble biomolecules or cytokines during the leaching porogen process, (*2*) the possibility that the remaining porogen as a salt can be harmful to the cell culture, and (*3*) the different geometry surface and cross-section that results.

The gas-foaming method consists of a solid scaffold matrix exposured to a sudden expansion of CO_2 gas under high pressure, which results in the formation of a sponge structure due to nucleation and expansion in a dissolved CO_2 scaffold matrix. The PLGA scaffolds with $> 93\%$ porosity and ~ 100 μm median pore size were developed by this method (71). A significant advantage is that there is no loss of bioactive molecules in the scaffold matrix, since there is no more need for the leaching process and there is no residual organic solvent. The disadvantage is the presence of a skimming film layer on the scaffold surface, which results in a need for an additional process to remove the skin layer.

The phase-separation method is divided into the freeze–drying, freeze–thaw, freeze–immersion precipitation, and emulsion freeze–drying techniques (37,72,73). Phase separation by freeze–drying can be induced by the appropriate concentration of polymer solution obtained by rapid freezing. Then, the used solvent is removed by freeze–drying, leaving in porous structure made up of a portion of the solvent. These can be collagen scaffolds with pores ~ 50–150 μm; collagen–glycosaminoglycan blend scaffolds with an average pore size ~ 90–120 μm; or chitosan scaffolds with a pore size ~ 1–250 μm, dependent on the freezing conditions (71). Also, scaffold structures of synthetic polymers, such as PLA or PLGA, have been successfully made much $> 90\%$ porosity and ~ 15–250 μm size by this method. The freeze–thaw technique induces phase separation between a solvent and a hydrophilic monomer upon freezing, followed by the polymerization of the hydrophilic monomer by means of ultraviolet (UV) irradiation and removal of the solvent by thawing. This technique leads to the formation of a macroporous hydrogel. A similar method is the freeze–immersion precipitation technique. The polymer solution is cooled, immersed in a nonsolvent, and then the vaporized solvent leads to a porous scaffold structure. Also, the emulsion freeze–drying method is used to fabricate a porous structure. Mixtures of polymer solution and nonsolvent are thoroughly sonicated, freezed quickly in liquid nitrogen at $-198\,^{\circ}C$, and then freeze–dried, resulting in a sponge structure. The advantage of these techniques is that they result in the loading of hydrophilic or hydrophobic bioactive molecules, whereas the disadvantages are relatively small pore sized scaffolds with precise pore structures that are hard to control (73).

Nano-electrospinning of PGA, PLA, PLGA, caprolactone copolymers, collagen, and elastin, has been extensively developed (74). For example, electrostatic processing can consistently produce PGA fiber diameters $\leqslant 1$ μm. By controlling the pick-up of these fibers, the orientation and mechanical properties can be tailored to the specific needs of the injured site. Also, collagen electrospinning was performed utilizing type I collagen dissolved in 1,1,1,3,3,3-hexafluoro-2-propanol with a concentration of $0.083\ g \cdot mL^{-1}$. The optimally electrospun type I collagen nonwoven fabric appeared with an average diameter of 100 ± 40 nm, which resulted in biomimicking fibrous scaffolds.

Injectable gel scaffolds have also been reported (10,16,51,54). An injectable, gelforming scaffold offers several advantages: (*1*) it can fill any space based on its ability to flow; (*2*) it can load various types of bioactive molecules and cells by simple mixing; (*3*) it does not contain residual solvents that may be present in a preformed scaffold; and (*4*) it does not require a surgical procedure for placement. Typical examples are thermosensitive gels such as Pluronics and PEG-PLGA-PEG triblock copolymer, pH sensitive gels such as chitosan and its derivates, an ionically cross-linked gel such as alginate, and fibrin and hyaluronan gels, as well as others previously introduced in the Natural Polymers section. In the near future, multifunctional gels which are tissue-specific, have a very fast sol–gel transition, are fully degradable over the necessary time period will be available.

Newly hybridized fabrication techniques such as organic–inorganic and synthetic–natural techniques at the nanosize level that biomimic, are also being developed for use in engineered scaffolds.

Physicochemical Characterization of Scaffolds

For the successful achievement of 3D scaffolds, several characterization methods are needed. These methods can be divided into four categories. (*1*) Morphology—porosity, pore size, and surface area; (*2*) mechanical properties—compressive and tensile strength; (*3*) bulk properties—degradation and its relevant mechanical properties; and (*4*) surface properties—surface energy, chemistry, and charge.

Porosity is defined as the fraction of the total volume occupied by voids that appear as percentages. The most widely used methods for the measurement of porosity are mercury porosimetry, scanning electron microscopy (SEM), and confocal laser microscopy.

Mechanical properties are extremely important when designing tissue-engineered products. Conventional testing instruments can be used to determine the mechanical properties of a porous structure. Mechanical tests can be divided into (*1*) creep tests, (*2*) stress–relaxation tests, (*3*) stress–strain tests, and (*4*) dynamic mechanical tests. These test methods are similar to those used for conventional biomaterials.

The rate of degradation of manufactured scaffolds is a very important factor in the design of tissue-engineered products. Ideally, the scaffold constructs provide mechanical and biochemical supports until the entire tissue regenerates, then the scaffold completely biodegrades at a rate consistent with tissue generation. Immersion studies are commonly conducted to track the degradation of the biodegradable matrix. Changes in weight loss and molecular weight can be evaluated by the chemical balance of the matrix, by SEM, and by gel permeation chromatography. These results produce the mechanism of biodegradation.

It is generally recognized that the adhesion and proliferation of different types of cells on polymeric materials depend largely on the materials' surface characteristics, such as wettability (hydrophilicity/hydrophobicity of surface free energy), chemistry, charge, roughness, and rigidity (37,40,41,44,45). The 3D aspects of tissue engineering are more important for cell migration, proliferation, DNA/RNA synthesis, and phenotype presentation on the scaffold materials. Surface chemistry and charge can be analyzed by electron scanning chemical analysis and streaming potential, respectively. Also, wettability of the scaffold surface can be measured by the contact angle using static and dynamic methods.

SURFACE MODIFICATION OF SCAFFOLDS FOR THE IMPROVEMENT OF BIOCOMPATIBILITY

As explained above, the surface properties of scaffold materials are very important. For example, the hydrophobic surfaces of PLA, PGA, and PLGA possess high interfacial free energy in aqueous solutions, which tends to unfavorably influence their cell, tissue, and blood compatibility in the initial stage of contact. Moreover, it does not allow the nutrient media to permeate into the center of the scaffolds. For these reasons, a surface treatment is applied by several methods: (*1*) chemical treatment using oxidants, (*2*) physical treatment using glow discharge, and (*3*) a blend with hydrophilic biomaterials or bioactive molecules.

The physicochemical treatment has been demonstrated to improve the wetting property and hydrophilicity of PLGA porous scaffolds fabricated by the emulsion freeze–drying method (37,45). The chemical treatments were 70% perchloric acid, 50% sulfuric acid, and 0.5 *N* sodium hydroxide solution. The physical methods included corona and plasma treatments generated by a radiofrequency glow discharge. After treatment, water contact angles decreased (Fig. 9). The wetting property of chemically treated PLGA scaffolds also ranked in the order of perchloric acid, sulfuric acid, and sodium hydroxide solution by blue dye intrusion experiment, whereas physical methods had no effect, as shown in Fig. 10. Thus, the chemical treatment method may be useful in uniform cell seeding into porous biodegradable PLGA scaffolds. Wettability plays an important role in cell adhesion, spreading, and growth on the PLGA surface, and the intrusion of nutrient media into the PLGA scaffold.

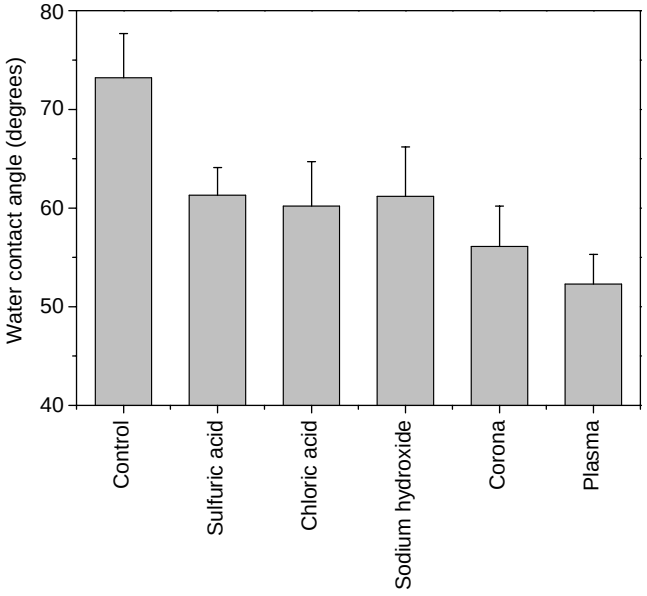


Figure 9. Changes of water contact angles after physicochemical treatment. The significant decreasing of water contact angle, that is, increased hydrophilicity, was observed.

Scaffolds impregnated with bioactive and hydrophilic material might be better for cell proliferation, differentiation, and migration due to cell stimulation. To give scaffolds new bioactive functionality from SIS powder as a natural source, scaffolds consisting of porous SIS/PLA and SIS/PLGA as a natural–synthetic composite, were prepared by the solvent casting–salt leaching method for use in tissue-engineered bone. A uniform distribution of good interconnected pores from the surface-to-core region was observed (pore size 40–500 μm), independent of the SIS amount, by using the solvent casting–salt leaching method. Porosities, specific pore areas as well as pore size distribution were also similar. After the fabrication of SIS/PLGA hybrid scaffolds, the wetting properties were greatly improved resulting in more uniform cell seeding and distribution, as shown in Fig. 11. Five different scaffolds, a PGA nonwoven mesh scaffold without glutaraldehyde (GA) treatment, PLA scaffolds without and with GA treatment, PLA/SIS scaffolds without and with GA treatment, were implanted into the back of nude mouse to observe the effect of SIS on the induction of cell proliferation by hematoxylin and eosin using von Kossa staining, for 8 weeks. It was observed that the effect of PLA/SIS scaffolds with GA treatment on bone induction is stronger than PLA scaffolds, that is the effects of PLA/SIS scaffolds with GA treatment $>$ PLA/SIS scaffolds without GA treatment $>$ PGA nonwoven $>$ PLA scaffolds only with GA treatment $=$ PLA scaffolds only without GA treatment for osteoinduction activity (Fig. 12).

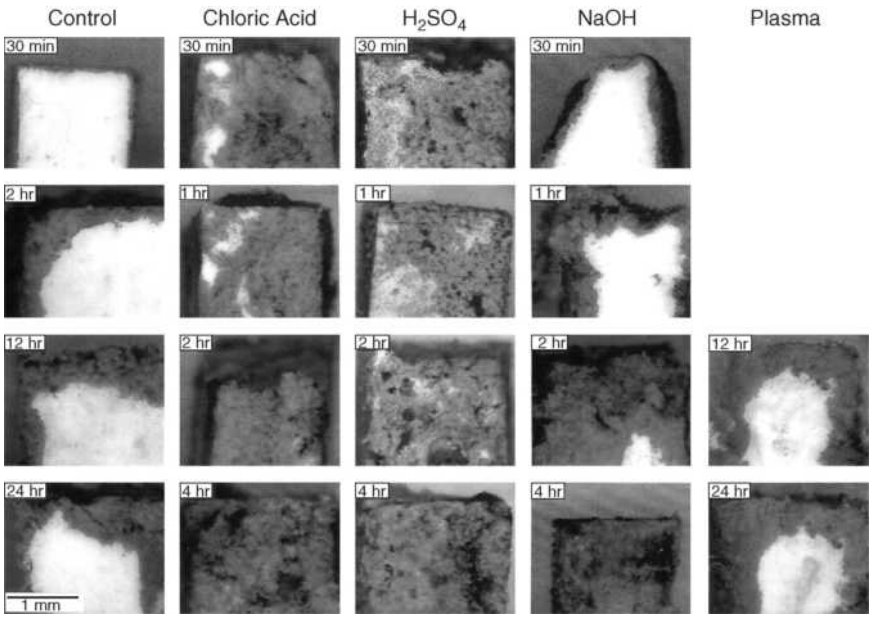


Figure 10. Wetting properties of physicochemically treated porous PLGA scaffolds by blue dye intrusion methods for 0.5, 1, 2, 4, 12, and 24 h.

STERILIZATION METHODS FOR SCAFFOLDS

The sterilizability of polymeric scaffold biomaterials is an important property, since polymers have lower thermal and chemical stability than other materials, such as ceramics and metals. Consequently, polymers are more difficult to sterilize using conventional techniques. Commonly used sterilization techniques are dry heat, autoclaving, radiation, and ethylene oxide gas (EOG). In addition, plasma glow discharge and electron beam sterilization recently were proposed due to their convenience (6,75).

In dry heat sterilization, the temperature varies between 160 and 190 °C. This temperature is above the melting and softening temperatures of many linear polymers, such as PLGA, resulting in the shrinking of the scaffold dimension. The PLA scaffolds were sterilized at 129 °C for 60 s, resulting in a minimal change in tensile properties. One of the significant problems was a decrease in molecular weight, which might have an affect on the degradation kinetics of the polymers. In the case of polyamide (Nylon) used as a nonbiodegradable polymer, oxidation occurs at the dry sterilization temperature, even though this is below its melting temperature. The only polymers that can safely be dry sterilized are polytetrafluoroethylene (PTFE) and silicone rubber. However, ceramic and metallic scaffolds were safe in this temperature range.

Steam sterilization (autoclaving) is performed under high steam pressure at a relatively low temperature (125–130 °C). However, if the polymer is subjected to attack by water vapor, this method cannot be employed. The PVC, polyacetals, PE (low density variety), and polyamides belong to this category. In the poly(α-hydroxy ester) family, a trace of water can deteriorate the PLGA backbone.

Chemical agents such as EOG and propylene oxide gases, and phenolic and hypochloride solutions are used widely for sterilizing all biomaterials, since they can be used at relatively low temperatures. Chemical agents

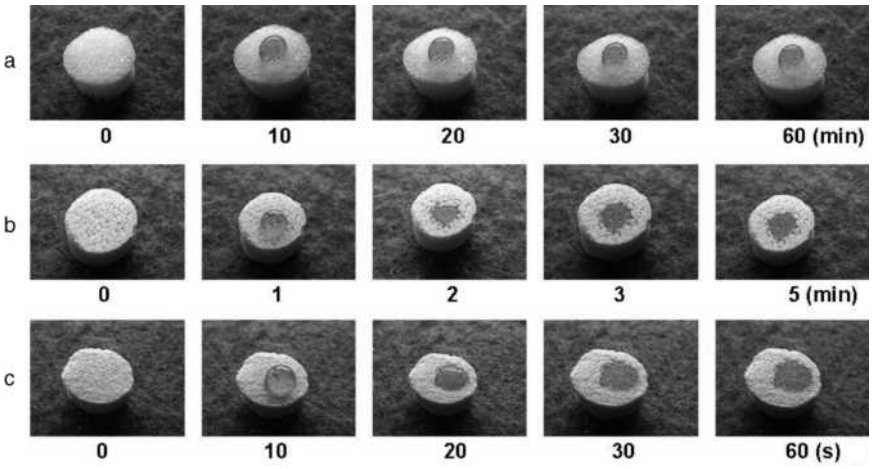


Figure 11. Wetting properties of SIS impregnated PLGA scaffolds by red dye intrusion methods. We observed the rapid penetration of water into SIS/PLGA scaffolds compared to the control PLGA scaffolds; (a) control PLGA, (b) 40% SIS/PLGA, and (c) 160% SIS/PLGA scaffolds.

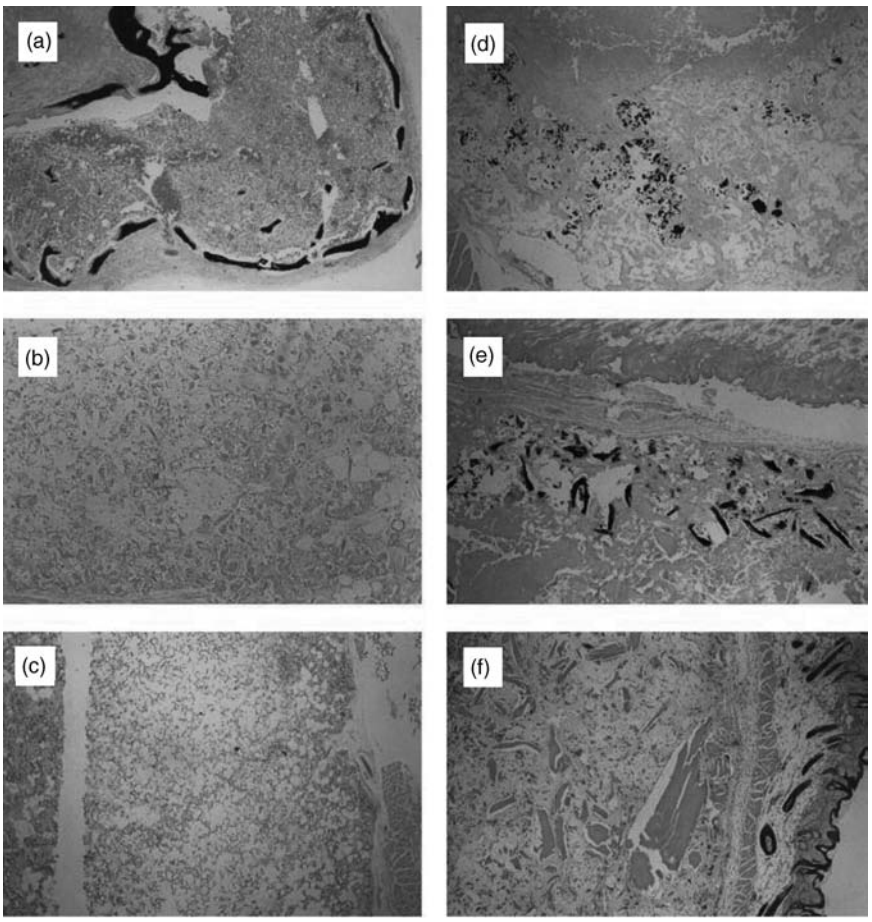


Figure 12. Photomicrographs of von Kossa and H&E histological sections of implanted (a) PGA nonwoven, (b) PLA scaffold only without GA treatment, (c) PLA scaffold only with GA treatment, (d) SIS/PLA scaffold without GA treatment, (e) SIS/PLA scaffold with GA treatment, and (f) SIS/PLA scaffold with GA treatment (H&E) (Original magnification 100×).

sometimes cause polymer deterioration even when sterilization takes place at room temperature. However, the time of exposure is relatively short (overnight), and most scaffolds can be sterilized with this method. The cold EOG sterilization method is the most widely used, with conditions of 35°C and 95% humidity. While the hot EOG method, which uses 60°C and 95% humidity, can cause shrinkage of the PLGA scaffold. One significant problem is residual EOG, which is harmful on the surface and within the polymer. Therefore, it is important that the scaffolds are subjected to adequate degassing or aeration subsequent to EOG sterilization, so that the concentration of residual EOG can be reduced to acceptable levels.

Radiation sterilization using isotopic ^{60}Co can also deteriorate polymers, since at high dosages the polymer chains can be dissociated or crosslinked according to the characteristics of the chemical structures. At a 2.5 Mrad dose, the tensile strength and molecular weight of PLGA decreases. Also, there is a rapid decrease in the molecular weight of the PGA nonwoven felt with increasing doses of radiation. It is important to remember that the properties and useful lifetime of the PLGA implant can be significantly affected by irradiation. In the case of polyethylene, it becomes a brittle and hard material at doses as high as 25 Mrad; This is due to a combination of random chain scission crosslinking. Polypropylene will often discolor during irradiation giving the product an undesirable tint, but a more severe problem is the embrittlement resulting in flange breakage, luer cracking, and tip breakage. The physical properties continued to deteriorate with time following irradiation.

Sterilization methods might significantly affect the physicochemical properties of the scaffold matrix. The specific effects with various methods are determined by the kinds of scaffold materials themselves, the scaffold preparation methods, and the sterilization factors. It is essential that a new standard for sterilizing scaffold devices be designed and established.

CONCLUSIONS

Tissue engineering, including regenerative medicine in recognition of its tremendous potential, has received a revolutionary "research push." As a result, there have been many reports on the successful regeneration of tissues and organs including skin, bone, cartilage, the peripheral and central nerves, tendon, muscle, cornea, bladder and urethra, and liver as well as composite systems like the human phalanx and joint, using scaffold biomaterials from polymers, ceramic, metal, composites and its hybrids. As previously emphasized, scaffold materials must contain a site of cellular and molecular induction and adhesion, and must allow for the migration and proliferation of cells through porosity. They must also maintain strength, flexibility, biostability, and biocompatibility to mimic a more natural, 3D environment. From this standpoint, control over a precise biochemical signal must be fostered by the combination

of a scaffold matrix and bioactive molecules including genes, peptide molecules, and cytokines. Moreover, the combination of cells and redesigned bioactive scaffolds should expand to a tissue level of hierarchy. To achieve this goal, novel scaffold biomaterials, scaffold fabrication methods, and characterization methods must be developed.

ACKNOWLEDGMENTS

This work was supported by grants from the Korea Ministry of Wealth and Health (0405-BO01-0204-0006) and stem cell Research Center (SC3100).

BIBLIOGRAPHY

Cited References

1. Langer R, Vacanti J. Tissue engineering. Science 1993;260: 920–926.
2. Nerem RM , Sambanis A. Tissue engineering: from biology to biological substitutes. Tissue Eng 1995;1:3–13.
3. Griffith LG, Naughton G. Tissue engineering—Current challenges and expanding opportunity. Science 2002;295:1009–1014.
4. Baldwin SP, Saltzman WM. Materials for protein delivery in tissue engineering. Adv Drug Deliv Rev 1998;33:71–86.
5. Mann BK, West JL. Tissue engineering in the cardiovascular system: Progress toward a tissue engineered heart. Anat Record 2001;263:367–371.
6. Lee HB, Khang G, Lee JH. Chapter 3, Polymeric biomaterials. In: Park JB, Bronzino JD, editors. Biomaterials: Principles and Applications. Boca Raton: CRC Press; 2003.
7. Patrick CW, Jr. Tissue engineering strategies for adipose tissue repair. Anat Record 2001;263:361–376.
8. Petit-Zeman S. Regenerative medicine. Nature Biotech 2001;19:201–206.
9. Hench LL, Polak JM. Third-generation biomedical materials. Science 2002;295:1014–1017.
10. Seal BL, Otero TC, Panitch A. Polymeric biomaterials for tissue and organ generation. Mater Sci Eng 2001;R34:147–230.
11. Babensee JE, McIntire LV, Mikos AG. Growth factor delivery for tissue engineering. Pharm Res 2000;17:497–504.
12. Chaignaud BE, Langer R , Vacanti JP. Chapter 1, The history of tissue engineering using synthetic biodegradable polymer scaffolds and cells. In: Atala A, Mooney DJ, editors. Synthetic Biodegradable Polymer Scaffolds. Boston: Birkhauser; 1996.
13. Rose FRA, Oreffo ROC. Bone tissue engineering: Hope vs Hype. Biochem Biophys Res Commun 2002;292:1–7.
14. Freyman TM, Yannas IV, Gibson LJ. Cellular materials as porous scaffolds for tissue engineering. Prog Mater Sci 2001;46:273–282.
15. Woolverton CJ, Fulton JA, Lopina ST, Landis WJ. Chapter 3, Mimicking the natural tissue environment. In: Lewandrowski K-U, Wise DL, Trantolo DJ, Gresser JD, Yasemski MJ, Altobeli DE, editors. Tissue Engineering and Biodegradable Equivalents: Scientific and Clinical Applications. New York: Marcel Dekker; 2002.
16. Tabata Y. The importance of drug delivery systems in tissue engineering. PSTT 2000;3:80–89.
17. Wong WH, Mooney DJ. Chapter 4, Synthesis of properties of biodegradable polymers used as synthetic matrices for tissue engineering. In: Atala A, Mooney DJ, editors. Synthetic Biodegradable Polymer Scaffolds. Boston: Birkhauser; 1996.
18. Rwoley JA, Madlambayan G, Mooney DJ. Alginate hydrogels as synthetic extracellular matrix. Biomaterials 1999;20:45–53.
19. Shakibaei M, De Souza P. Differentiation of mesenchymal limb bud cells to chondrocytes in alginate bead. Cell Biol Int 1997;21:75–86.
20. Madihally SV, Matthew HW. Porous chitosan scaffolds for tissue engineering. Biomaterials 1999;20:1133–1142.
21. Kang HW, Tabata Y, Ikada Y. Fabrication of porous gelatin scaffolds for tissue engineering. Biomaterials 1999;20:1339–1344.
22. Dunn CJ, Goa KL. Fibrin sealant: A review of its use in surgery and endoscopy. Drugs 1999;58:863–886.
23. Mayne R, Burgeson RE. Structure and function of collagen types. In: Mecham RP, editor. Biology of extracellular matrix: A Series. Orlando: Academic Press; 1987.
24. Li S-T. Chapter 6, Biologic biomaterials: Tissue-derived biomaterials (Collagen). In: Park JB, Bronzino JD, editors. Biomaterials: Principles and Applications. Boca Raton (FL): CRC Press; 2003.
25. Schense JC, Bloch J, Aebischer P, Hubbell JA. Enzymatic incorporation of bioactive peptides into fibrin matrices enhances neurite extension. Nature Biotech 2000;18:415–419.
26. Zacchi V, Soranzo C, Cortivo R, Radice M, Brun P, Abatangelo G. In vitro engineering of human skin-like tissue. J Biomed Mater Res 1998;40:187–194.
27. Malette WG, Quigley HJ, Gaines RD, Johnson ND, Rainer WG. Chitosan: a new hemostatic. Ann Thorac Surg 1983;36: 55–58.
28. Sechriest VF, Miao YJ, Niyibizi C, Westerhauzen-Larson A, Matthew HW, Evans CF, Fu FH, Suh J-K. GAG-augmented polysaccharide hydrogel: a novel biocompatible and biodegradable material to support chondrogenesis. J Biomed Mater Res 2000;49:534–541.
29. Lee YM, Park YJ, Lee SJ, Ku Y, Han SB, Choi SM, Klokkevoid PR, Chung CP. Tissue engineered bone formation using chitosan/tricalcium phosphate sponges. J Periodontol 2000; 71:410–417.
30. Lee DA, Noguchi T, Knight MM, O'Donnell L, Bently G, Bader DL. Response of chondrocyte subpopulations cultured within unloaded and loaded agarose. J Orthop Res 1998;16: 726–733.
31. Lee DA, Frean SP, Lee P, Bader DL. Dynamic mechanical compression influences nitric oxide production by articular chondrocytes seeded in agarose. Biochem Biophys Res Commun 1998;251:580–585.
32. Badylak SF, Record R, Lindberg K, Hodde J, Park K. Small intestine submucosa: a substrate for in vitro cell growth. J Biomater Sci, Polymer Ed 1998;9:863–878.
33. Khang G, Shin P, Kim I, Lee B, Lee SJ, Lee YM, Lee HB, Lee I. Preparation and characterization of small intestine submucosa particle impregnated PLA scaffolds: The application of tissue engineered bone and cartilage. Macromol Res 2002;10:158–167.
34. Badylak SF. The extracellular matrix as a scaffolds for tissue reconstruction. Cell Develop Biol 2002;13:377–383.
35. Gustafson C-J, Katz G. Cultured autologous keratinocytes on a cell-free dermis in the treatment of full-thickness wounds. Burns 1999;25:331–335.
36. Williams SF, Martin DP, Horowitz DM, Peoples OP. PHA applications: Addressing the price performance issue, I. Tissue Engineering. Int J Biolog Macromol 1999;25:111–121.
37. Khang G, Lee HB. Chapter 67. Cell-synthetic surface interaction: Physicochemical surface modifications. In: Atala A, Lanza R, editors. Orlando: Academic Press; 2001.
38. Khang G, Seong H, Lee HB. Sustained delivery of drugs with biodegradable. In: Hsuie GH, Okano T, Kim YU, Sung W-W, Yui N, Park KD, editors. Taipei, Taiwan: Princeton International Publishing Co.; 2002.
39. Khang G, Lee SJ, Han CW, Rhee JM, Lee HB. Preparation and characterization of natural/synthetic hybrid scaffolds.

In: Elcin M, editor. London, England: Kluwer-Plenum Press; 2003.
40. Khang G, Lee JH, Lee I, Rhee JM, Lee HB. Interaction of different types of cells on PLGA surfaces with wettability chemogradient. Macromol Res 2000;8:276–284.
41. Khang G, Choi MK, Rhee JM, Rhee SJ, Lee HB, Iwasaki Y, Nakabayashi N, Ishihara K. Biocompatibility of poly(MPC-*co*-EHMA)/PLGA blends. Macromol Res 2001;9:107–115.
42. Khang G, Park CS, Rhee JM, Lee SJ, Lee YM, Lee I, Choi MK, Lee HB. Preparation and characterization of demineralized bone particle impregnated PLA scaffolds. Macromol Res 2001;9:267–276.
43. Choi HS, Khang G, Shin H-C, Rhee JM, Lee HB. Preparation and characterization of fentanyl-loaded PLGA microspheres; *In vitro* release profiles. Int J Pharm 2002;234:195–203.
44. Lee SJ, Khang G, Lee YM, Lee HB. Interaction of human chondrocyte and fibroblast cell onto chloric acid treated poly(α-hydroxy acid) surface. J Biomater Sci, Polym Ed 2002;13:197–212.
45. Khang G, Choi CW, Rhee JM, Lee HB. Interaction of different types of cells on physicochemically treated PLGA surfaces. J Appl Polym Sci 2002;85:1253–1262.
46. Khang G, Jeon EK, Rhee JM, Lee I, Lee SJ, Lee HB. Controlled release of NGF from sandwiched PLGA films for the application of neural tissue engineering. Macromol Res 2003; 11:334–340.
47. Jang JW, Lee B, Han CW, Lee I, Lee HB, Khang G. Preparation and characterization of ipriflavone-loaded PLGA scaffolds for tissue engineered bone. Polymer(Korea) 2003;27:226–234.
48. Khon J, Langer R. Chapter 2.5, Bioresorbable and bioerodible materials. In: Ratner BD, Hoffman AS, Scheon FJ, Lemons JE, editors. Biomaterials Science: An Introduction to Materials in Medicine, San Diego: Academic Press; 1996.
49. Vacanti CA, Langer R, Schloo B, Vacanti JP. Synthetic polymers seeded with chondrocytes provide a template for new cartilage formation. Plast Reconstr Surg 1991;88:753–759.
50. Burg KJL, Porter S, Kellam JF. Biomaterials developments for tissue engineering. Biomaterials 2000;21:2347–2359.
51. Gutowska A, Jeong B, Jasionowski M. Injectable gel for tissue engineering. Anat Record 2001;263:342–349.
52. Suggs LJ, Krishna RS, Garcia CA, Peter SJ, Anderson JM, Mikos AG. In vitro and in vivo degradation of poly(propylene fumarate-*co*-ethylene glycol) hydrogel. J Biomed Mater Res 1998;42:312–320.
53. Harris JM, editor. Poly(ethylene glycol) Chemistry: Biotechnical and Biomedical Applications. New York: Plenum Publish. Co.; 1997.
54. Qui Y, Park K. Environment-sensitive hydrogels for drug delivery. Adv Drug Deliv Rev 2001;53:321–339.
55. Webb D, An YH, Gutowska A, Mironov VA, Friedman RJ. Propagation of chondrocytes using thermosensitive polymer gel culture. Orthoped J Musc Orthoped Surg 2000;3:18–22.
56. Sims CD, Butler P, Casanova R, Lee BT, Randolph MA, Lee A, Vacanti CA, Yaremchuk MJ. Injectable cartilage using polyethylene oxide polymer substrate. Plast Reconstruct Surg 1996;95:843–850.
57. Laurencin CT, El-Amin SF, Ibim SE, Willoughby DA, Attawia M, Allcock HR, Ambrosio AA. A highly porous 3-dimensional polyphosphazene polymer matrix for skeletal tissue engineering. J Biomed Mater Res 1996;30:133–138.
58. Agarwal S, Gassner R, Piesco NP, Ganta SR. Chapter 7, Biodegradable urethanes for biomedical applications. In: Lewandrowski K-U, Wise DL, Trantolo DJ, Gresser JD, Yasemski MJ, Altobeli DE, editors. Tissue Engineering and Biodegradable Equivalents: Scientific and Clinical Applications. New York: Marcel Dekker; 2002.
59. Zhang JY, Beckman EJ, Piesco NP, Agarwal S. A new peptide based urethane polymer: synthesis, degradation, and potential to support cell growth in vitro. Biomaterials 2000;21:1247–1258.
60. Billotte WG. Chapt. 2, Ceramic biomaterials. In: Park JB, Bronzino JD, editors. Biomaterials: Principles and Applications. Boca Raton (FL): CRC Press; 2003.
61. Hench LL. Bioactive ceramics. Ann NY Acad Sci 1988; 523:54–71.
62. Frician JC, Bareille R, Rouais F. In vitro dissolution of coral in periodontal or fibroblast cell culture. J Dent Res 1998;77:406–411.
63. Yoshikawa T, Oghushi H, Uemura T. Human marrow cells derived cultured bone in porous ceramics. Bio-Med Mater Eng 1998;8:311–320.
64. Unpublished data.
65. Krewson C, Dause R, Mak M, Saltzman WM. Stabilization of nerve growth factor in controlled release polymers and in tissue. J Biomater Sci, Polym Ed 1996;8:103–117.
66. Haller MF, Saltzman WM. Localized delivery of proteins in the brain. Pharm Res 1998;15:377–385.
67. Ito Y, Lui SQ, Imanishi Y. Enhancement of cell growth on growth factor-immobilized polymer films. Biomaterials 1991; 12:449–453.
68. Khul PR, Grriffith-Cima LG. Tethered epidermal growth factor as a paradigm for growth factor-induced stimulation from the solid phase. Nature Med 1996;2:1022–1027.
69. Duncan R, Spreafico F. Polymer conjugates. Pharmacokinetic considerations for design and development. Clin Pharmacokinet 1994;27:290–306.
70. Massia SP, Hubbell JA. Covalent surface immobilization of Arg-Gly-Asp- and Tyr-Ile-Gly-Ser-Arg-containing peptides to obtain well-defined cell-adhesive substrate. Anal Biochem 1990;187:292–301.
71. Leibmann-Vinson A, Hemperly JJ, Guarino RD, Spargo CA, Heidaran MA. Chapter 36, Bioactive extracellular matrices: Biological and biochemical evaluation. In: Lewandrowski KU, Wise DL, Trantolo DJ, Gresser JD, Yasemski MJ, Altobeli DE, editors. Tissue Engineering and Biodegradable Equivalents: Scientific and Clinical Applications. New York: Marcel Dekker; 2002.
72. Thompson RC, Wake MC, Yasemski MJ, Mikos AG. Biodegradable polymer scaffolds to regenerate organs. Adv Polym Sci 1995;122:245–274.
73. Khang G, Jeon JH, Cho JC, Lee HB. Fabrication of tubular porous PLGA scaffolds by emulsion freeze drying methods. Polymer(Korea) 1999;23:471–177.
74. Bowlin GL, Pawlowski KJ, Boland ED, Simpson DG, Fenn JB, Wnek GE, Stitzel JD. Chapter 9, Electrospinning of polymer scaffolds for tissue engineering. In: Lewandrowski K-U, Wise DL, Trantolo DJ, Gresser JD, Yasemski MJ, Altobeli DE, editors. Tissue Engineering and Biodegradable Equivalents: Scientific and Clinical Applications. New York: Marcel Dekker; 2002.
75. Athanasious KA, Neiderauer GG, Agrawal CM. Sterilization, toxicity, biocompatibility and clinical applications of polylactic acid/polyglycolic acid copolymers. Biomaterials 1996;17:93–102.

Reading List

Jeon EK, Khang G, Lee I, Rhee JM, Lee HB. Preparation and release profile of NGF-loaded PLA scaffolds for tissue engineered nerve regeneration. Polymer(Korea) 2001;25:893–901.

See also ENGINEERED TISSUE; STERILIZATION OF BIOLOGIC SCAFFOLD MATERIALS.

BIOMECHANICS OF EXERCISE FITNESS

GIDEON ARIEL
Ariel Dynamics
Canyon, California

INTRODUCTION

Normal human development spans a lifetime from infancy to old age. Modern civilization is confronted with the lengthening of that time and its effect on the individual and society. Housing improvements, employment alterations, labor saving devices, and modern medicine are but a few of the factors protecting humanity from those instances which previously shortened life. While many of the difficult, threatening experiences have been eliminated or reduced in severity, problems remain to be solved. Concerns for the quality of life as people become older include maintaining self-sufficiency. Many solutions conflict with beliefs generally termed "current wisdom" in areas, such as training, dieting, exercising, and aging. While society ages, the challenge for each individual is to strive to retain the lowest "biological" age while their "chronological" birthdays increase. The dilemma concerns the best way to accomplish this task.

The main purpose of this article is to focus on the biomechanical principles of movement, the scientific bases of training and fitness, and the optimization of human performance at any age. These are not just nonsense concepts added to the quantities of known theories, but are objectively quantifiable procedures that encompass our understandings and can produce precise conclusions. Mathematical principles and gravitational formulations provide the cornerstones for optimizing human performance. Biological, anatomical, physiological, and medical discoveries are always under investigation, challenge, and improvement and these findings will be incorporated into many of the current theories. Figure 1 illustrates just part of the anatomy and its complicated structure. The struggle will continue among scientists to establish new principles for revolutionizing the world of gerontology, diet, physical fitness and training, and amplifying those factors necessary for extending life not only in length, but also in quality. Scientists with expertise in many different areas will be addressing the problems associated with aging from their specialized perspective.

In order to address the optimization of human movement and performance, the underlying philosophical premise metaphorically compares life with sport. The goal is that everyone should be a gold medalist in their own body regardless of age. Most people, however, do not achieve their Gold Medal because their goals, potential, and/or timing are uncoordinated or nonexistent. For example, an individual may envision themselves as a tennis champion, yet lack the requisite physical and physiological traits of the greatest players. Given this situation, can a person's potential be maximized? Achieving one's maximum potential necessitates tools applicable to everyone for improving their performance, whether in tennis, fitness, overcoming physical handicaps, or fighting disease. Useful tools must be based, however, on correct, substantive scientific principles.

SCIENTIFIC PRINCIPLES FOR QUANTIFYING MOTION

Human movement has fascinated humans for centuries including some of the world's greatest thinkers, such as Leonardo da Vinci, Giovanni Borelli, Wilhelm Braune, and others. Many questions posed by these stellar geniuses have been or can be addressed by the relatively new area of Biomechanics. Biomechanics is the study of the motion of living things, primarily, and it has evolved from a fusion of the classic disciplines of anatomy, physiology, physics, and engineering. Bio refers to the biological portion, incorporating muscles, tendons, nerves, and so on, while mechanics is associated with the engineering concepts based upon the laws described by Sir Isaac Newton. Human bodies consist of a set of levers that are powered by muscles. Quantification of movements, whether human, animal, or inanimate objects, can be treated within biomechanics according to Newtonian equations. It may seem obvious, with the perfect vision of hind sight, that humans and their activities, such as the wielding of tools (e.g., hammer, axe) or implements (e.g., baseball bat, golf club, discus), must obey the constraints of gravitational bodies, just as bridges, buildings, and cars do. For some inexplicable reason, humans and their activities had not been subjected to the appropriate engineering concepts that architects would use when determining the weight of books to be housed in a new library or engineers would apply to designing a bridge to span a wide, yawning abyss. It was not until Newton's

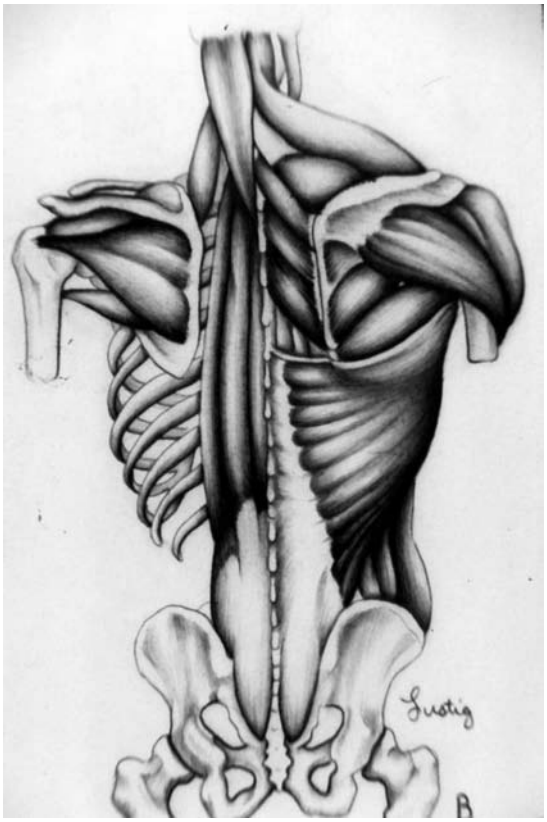


Figure 1. The human structure.

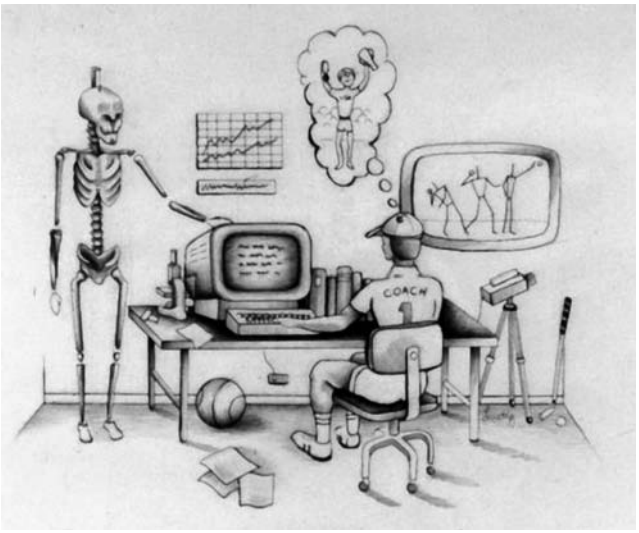


Figure 2. The modern coach and his tools.

apple fell again during the twentieth century that biomechanics was born.

Biomechanics, then, is built on a foundation of knowledge and the application of the basic physical laws of gravitational effects as well as those of anatomy, chemistry, physiology, and other human sciences. Early quantification efforts of human movement organized the body as a system of mechanical links. Activities were recorded on movie film that normally consisted of hundreds of frames for each of the desired movement segment. Since each frame of the activity had to be processed individually, the task was excessively lengthy, tedious, and time intensive. Figure 2 illustrates an abstract of todays sophisticated coaching tools in athletics. The hand calculations of a typical 16 segment biomechanical human required many hours for each frame, necessitating either numerous assistants or an individual investigator's labor-of-love and, frequently, both. Unfortunately, these calculations were susceptible to numerical errors.

The introduction of large, main-frame computers improved reliability and reasonableness of the results, replacing much of the skepticism or distrust associated with the manually computed findings. Computerization accelerated the calculations of a total movement much more rapidly than had been previously possible, but presented new difficulties to overcome. Many of the early biomechanical programs were cumbersome, time intensive main-frame endeavors with little appeal except to the obsessed, devotee of computers, and movement assessment. However, even these obstacles were conquered in the ever expanding computerization era. The computerized hardware/software system provides a means to objectively quantify the dynamic components of movement in humans regardless of the nature of the event. Athletic events, gait analyses, job-related actions as well as motion by inanimate objects, including machine parts, air bags, and auto crash dummies are all reasonable analytic candidates. Objectivity replaces mere observation and supposition.

One of the most important aspects included in the Bio portion of biomechanics is the musculoskeletal system. Voluntary human movement is caused by muscular contractions that move bones connected at joints. The neuromuscular system functions as a hierarchical system with autonomic and basic, life sustaining operations, such as heart rate and digestion, controlled at the lowest, noncognitive levels and with increasing complexities and regulatory operations, such as combing the hair or kicking a ball, controlled by centers that are further up the nervous system. Interaction of the various control centers is regulated through two fundamental techniques each governed like a servosystem.

The first technique equips each level of decision making with subprocessors that accept the commands from higher levels as well as accounting for the inputs from local feedback and environmental information sensors. Thus, a descending pyramid of processors is defined that can accept general directives and execute them in the presence of varying loads, stresses, and other perturbations. This type of input–output control is used for multimodal processes, such as maintaining balance while walking on an uneven terrain, but would be inappropriate for executing deliberate, volitional, complex tasks like the conductor using the baton to coordinate the music of the performing musicians.

The second technique utilized by the brain to control muscular contractions applies to the operation of higher level systems that generate output strategies in relation to behavioral goals. These tasks use information from certain sensory inputs, including joint angle, muscle loading, and muscular extension or flexion that are assessed, transmitted to higher centers for computation, which then executes the set of modified neural transmissions received. Cognitive tasks requiring the type of informational input that influences actions are the ones with which humans are most familiar since job execution requires more thought than breathing or standing upright. A frequently misunderstood concept is that limb movement is possible only through contractions of individual muscle fibers. For most cases of voluntary activity, muscles work in opposing pairs with one set of muscles opening or extending the joint (extensors) while the opposite muscle group closes or flexes the joint. The degree of contraction is proportional to the frequency of signals from the nerve as signaled from the higher centers. Movement control is provided by a programmable mechanism so that when flexors contract, the extensors relax, and vice versa. The motor integration programmed generated in the higher, cognitive levels regulates not only the control of the muscle groups around a joint, but also those necessary actions by other muscles and limbs to redistribute weight, to counteract shifts in the center of gravity.

One of the most important, but frequently misunderstood, concepts of the nervous system is the control and regulation of coordinated movement. When a decision is made to move a body segment, the prime muscles or agonists receive a signal to contract. The electrical burst stimulates the agonist muscular activity causing an acceleration of the segment in the desired direction. At the same time, a smaller signal is transmitted to the opposite muscle group, or antagonist, which causes it to function as a joint stabilizer. With extremely rapid movements, the antagonist is frequently stimulated to slow the limb in time to

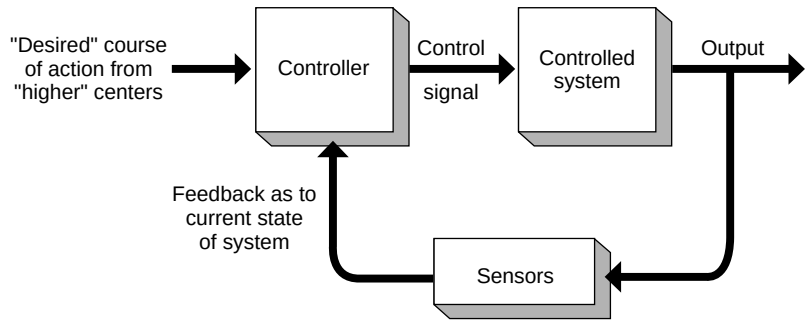


Figure 3. Feedback control mechanism of movement.

protect the joint from injury. It is the strength and duration of the electrical signal to both the agonist and antagonist that govern the desired action. The movement of agonists and antagonists, whether a cognitive process, such as throwing a ball, or an acquired activity, such as postural control, is controlled by the nervous system. Figure 3 illustrates a flowchart for the control system for movement. Many ordinary voluntary human activities resulting from agonist–antagonist muscular contraction are classified by different terms, isotonic, variable resistance, and ballistic. Slower movements, demonstrating smaller, more frequent, electrical signal alterations, are intricately controlled by both agonist and antagonist. These types of motion are tracking movements.

One control mechanism available involves the process of information channeled between the environment and the musculature. Closed-loop control involves the use of feedback whereby differences between actual and desired posture are detected and subsequently corrected, whereas open-loop control utilizes feed forward strategies that involve the generation of a command based on prior experience rather than on feedback. Braitenberg and Onesto (ARBIB) proposed a network for converting space into time by providing that the position of an input would determine the time of the output. This open loop system would trigger a preset signal from the nervous system to the muscle generating a known activity. Kicking a ball, walking, throwing a baseball, swinging a golf club, and hand writing are considered ballistic movements.

When a limb moves, a sophisticated chain of events occurs before, during, and after the movement is completed. The fineness of control depends on the number of muscle fibers innervated by each motor neuron. A motor unit is generally defined as a single motor neuron and the number of muscle fibers it innervates. Fine control is achieved when a single motor neuron innervates just a few fibers. Less fine control, as in many large muscle groups, is attained when individual motor units innervate hundreds or even thousands of fibers. The more neurons there are, the finer the ability to maneuver, as with eye movements or delicate hand manipulations. In contrast to the high innervations ratio of the eye, the biceps of the arm has a very low rate of nerve-to-muscle fiber resulting in correspondingly more coarse movements.

While the amount of nervous innervations is important when anticipating the precision of control, the manner of interaction and timing between muscles, nerves, and desired outcome is probably more important when evaluating performance. Recognizable actions elicit execution of patterned, synchronous nervous activity. Frequently repeated movements are usually performed crudely in the beginning stages of learning, but become increasingly more skilled with use and/or practice. Consider the common activity of handwriting and the execution of one's own signature. The evolution from a child's irregular, crude printing to an adult's recognizable, consistently repeatable signature is normal. Eventually, the individual's signature begins to appear essentially the same every time and is uniquely different from any other person. Not only can the person execute handwritten signatures consistency, but can use chalk to sign the name in large letters on a blackboard producing a recognizably similar appearance. The individuality of the signature remains whether using the fine control of the hand or recruiting the large shoulder and arm muscles not normally required for the task. Reproduction of recognizable movements occurs from preprogrammed control patterns stored in the brain and recruited as necessary. Practicing a golf swing until it results in a 300 yard drive down the middle of the fairway, getting the food-laden fork from the plate into the mouth, and remembering how to ride a bike after a 30 year hiatus illustrate learned behavior that has become "automatic" with practice and can be recalled from the brain's storage for execution.

Volitional tasks require an integration of neurological, physiological, biochemical, and mechanical components. There are many options available when performing a task, such as walking, but eventually, each person will develop a pattern that will be recognizable as that skill, repeatable, and with a certain uniqueness associated with that particular individual. Although any person's movement could be quantified with biomechanical applications and compared to other performers in a similar group, for example, the gold, silver, and bronze medalists in an Olympic event, perhaps it will be the ability to compare one person to themselves that will provide the most meaningful assistance in the assault on aging.

There are many areas of daily living in which biomechanical analyses could be useful. Biomechanics could be utilized to design a house or chair to suit the body or to lift bigger, heavier objects with less strain. This science could be useful in selecting the most appropriate athletic event for children or for improving an adult's performance. With increasing international interest in competitive athletics, it was inevitable that computers would be used for the analysis of sports techniques. Computer calculations can provide information that surpasses the limits of what the

human eye can see and intuition can deduce. Human judgment, however, is still critically important. As in business and industry, where decisions are based ultimately upon an executive's experience and interpretive ability, the coach or trainer is, and will remain, the ultimate decision maker in athletic training. Rehabilitation and orthopedic specialists can assess impaired movement relative to normal performance and/or apply computerized biomechanical techniques to the possibilities of achieving the restoration of normal activities. With the increase in the population of older citizens, erotological applications will increase. The computer should be regarded as one more tool, however, complex, which can be skillfully used by humans in order to achieve a desired end.

One factor that humans have lived with is change. The environment in which we live is changing during every one of the $\sim$ 35 million min of our lives. The human body itself changes from birth to maturity and from maturity to death. The moment humans first picked up a stone to use as a tool, the balance between humans and the environment was altered. After that adaptation, the ways in which the surrounding world changed resulted in different effects and these were no longer regular or predictable. New objects were created from things that otherwise would have been discounted. These changes were made possible by humans due to the invention of tools. The more tools humans created, the faster was the rate of environmental change. The rate of change due to tools has reached such a magnitude that there is danger to the whole environment and frequently to the people who use the tools, such as occurred during the Industrial Revolution, as well as in our own times with such problems as carpal tunnel syndrome. Human beings seem to have become so infatuated with their ability to invent things that they have concentrated almost exclusively upon improving the efficiency, safety, durability, cost, or aesthetic appeal of the device. It is ironic that with all of the innovative development, little consideration has been given to the most complex system with the most sophisticated computer in the world: the human body.

When they talk about their physical goals in work or in sports, people usually say they would like to do their best, meaning, reach their maximum output. It is a matter of achieving their absolute limit in speed, strength, endurance or skill and combining the elements with accuracy. This is no different than an athlete training for maximum performance in the Olympic games. The difficulty with focusing everything on maximum performance is that only a single goal, getting the highest results—fastest, biggest, quickest, longest, or most graceful—is considered a superlative or acceptable achievement. Maximums do not take into consideration other aspects of body performance that often prove to be just as important to the individual. Emphasis upon the demands for maximum performance is frequently portrayed with the thought that Winning isn't everything, it's the only thing. Figure 4 illustrates todays sophisticated biomechanical system to quantify human performance.

Imagine for a moment a maximum performance in the car industry—the perfect automobile. It is incredibly graceful and the aerodynamic, functional lines make it a thing of beauty. It accelerates from 0 to 60 miles $\cdot$ h^{-1} within a few seconds. It brakes, corners, and steers with a fineness that would permit a shortsighted 75-year old to compete at Le Mans. The suspension is so smooth that a passenger can pour liquids without spilling a drop. The car requires only minimal maintenance while averaging 50 miles $\cdot$ gal^{-1} in city driving. Best of all, it is the vehicle of the common man at a price of $5000. If all that sounds impossible—it is. Incorporating all of these maximums into a single automobile exceeds the ability of any designer or manufacturer. Instead, the individual shopping for a car must choose the attributes he or she feels are most important.

Figure 4. The modern biomechanical system.

Therein lies the problem, some goals are partly, if not wholly, incompatible with others. An automatic transmission uses more gas than a standard shift, but it does make driving easier. Sleek aerodynamic lines add grace and reduce drag, but they can also lessen head room. High performance engines provide power, but require constant care. The solution is a compromise, a willingness to make tradeoffs.

This same spirit of compromise, of accepting something less than a single maximum, should govern the operation of the most important machine in our lives—our body. Reality must be applied when comparing ourselves to Olympic athletes or, with the progression of age, mimicking various youthful physical activities. For example, there is no need to have an endurance capacity equal to the current gold medalist or the strength level equivalent to the World heavyweight record holder. Likewise, senior citizens may resist relinquishing their drivers' licenses despite their slower reaction times, poorer eyesight, and/or hearing, as well as frequently suffering from some type of chronic disease that may further reduce their strength, joint mobility, or even cognitive processes, such as memory or decision making.

Instead of a maximum, what most people really want from their bodies is to optimize their performances and lives. They seek the most efficient use of energy, of bodily action consonant with productive output, health, and enjoyment. Many people are beginning to appreciate that certain types of exercise add to the vitality of the

cardiovascular system, lessen the risk of heart attack, and make it possible to live longer and more active lives. In other words, the willingness to sacrifice 20 yards on a drive off the golf tee may mean that the golfer's feet will be able to walk the entire course without being tortured during every step. The desire is to play a couple of hours of winning tennis, stroking the ball with pace and purpose, but not if the extra zing means a tennis elbow that will be sore for several weeks. Sensible joggers prefer to run 6 rather than 10 miles a day in 40 min, if the latter leads to tender knees and shin splints. In other words, human beings must compromise between anatomy (the structural components) and physiology (the bodily processes). A correct balance between the two, at all ages, will assist in optimizing bodily efficiency.

In addition to the desire for our internal environment to be physical fit, pertinent questions should be posed about our external environment. For example, is it really necessary for that designer chair to cause a bone ache deep in the buttocks after sitting for 5 min? Can a person not spend a day laboring over a desk or piece of machinery without feeling as if a rope had been tightly tied around the shoulders at the end of the project? Why must a weekend with shovel or rake inevitably produce lower back pain on Monday? Why is it that some individuals who are 50 years old seem able to work and play as if 10–20 years younger, while some 30 year olds act as if infected with a malignant decrepitude? The answer is that, as with the anatomy and physiology achieving optimal coordination, so should the whole human organism coordinate better with its environment.

Perhaps these examples could be dismissed as the minor aches of a hypochondriac society overly concerned with its comfort. But the overall health facts for the United States and many other modern civilizations appall even those jaded by constant warnings of disaster. The American Heart Association, in urging the 2005 Congress to fund prevention programs, contends that the Number One killer of Americans is heart disease, stroke, and other cardiovascular diseases. In addition, a total of 75 million Americans are afflicted with chronic disease. On any given day, >1 million workers do not show up for their jobs because of illness, and sickness prevents a million of these from returning in <1 week. Twenty-eight million Americans have some degree of disability. Perhaps not coincidentally, a quarter of the population is classified as overweight. At least 3 million citizens have diabetes, and one-half are unaware of the problem, and the United States accounts for most of the deaths due to cardiovascular disease. The health profile of the future, the condition of the youth of today, offers no comfort. About 1 in 5 youngsters still cannot pass even a simple test of physical performance. More than 9 million American children under the age of 15 have a chronic ailment. From one-third to one-half of U.S. children are overweight and one-third of America's young men fail to meet military physical fitness requirements.

In pursuit of technological achievement, Americans have almost ignored the one major element besides food and rest needed to sustain the human body: physical activity. This has lent impetus to a subtle yet deadly disease that has reached epidemic proportions in this country and others. Cardiovascular disease is often referred to as hypokinetic disease or lack of-motion disease. Unfortunately, degeneration with Americans begins earlier rather than later. One study indicates that middle age characteristics start to show at approximately age 26. The peak age for heart disease among American men is 42 years. In Europe, it is 10 years later. A corporate wide employee health survey conducted by a large computer manufacturer indicated that smokers have 25% higher healthcare costs and 114% longer hospital stays than nonsmokers. People who did not exercise have 36% higher healthcare costs and 54% longer hospital stays than people who did exercise. Overweight people have 7% higher healthcare costs and 85% longer hospital stays than people who are not. In general, people with poor health habits have higher healthcare costs, longer hospital stays, lower productivity, more absenteeism, and more chronic health problems than those who do not. Some questions both workers and their companies should ask are (*1*) How many heart attacks, strokes, cancers, or coronary by-pass operations did your company pay for last year? (*2*) How much better would profits have been if heart diseases had been reduced 10, 20, or 30%? (*3*) How much would corporate profits increase if employee healthcare costs were reduced by 10%?

One large U.S. corporation developed a comprehensive wellness program at numerous sites. During the first year, grievances decreased by 50%, on-the-job accidents by 50%, lost time by 40%, and sickness and accident payments by 60%. The corporation estimated at least a 3:1 return per dollar invested.

The requirement for such an optimum way of life is a scientific analysis of the way people live and use their bodies. Only after such a quantitative examination can a concept of cost be determined or a better way of doing something that is more efficient and less damaging to the body, discovered. For example, rapid weight loss may result from running long distances, such as 15 miles a day, fasting drastically, or performing aerobics for 5 h a day. However, such excessive training regimens may be as detrimental to the body as sitting all day in an easy chair and simply ignoring one's obesity.

Evolution, culture, and the changing demands of existence have tended to develop forces and stresses upon the body that are not necessarily in harmony with the basic design and structure of the human equipment. Standing upright, humans employ one pair of extremities for support and the other pair capable of tremendous versatility. It would seem that of all animals, humans, fortuitously assisted by the evolution of their brain and other organs, optimized the use of their body. Unfortunately, the human body has had to pay a stiff price for its upright posture. Human vertical posture is inherently unstable; therefore, humans must devote more neuromuscular effort and control to maintain balance, than four-legged animals. There is a tendency to lean forward, which adds to the ability to move in that direction, but increases the risk of falling.

A complex neuromuscular process is constantly at work to prevent humans from toppling. Many things may interfere with this balancing act, such as consuming too much whiskey or walking on an icy sidewalk. These interruptions

of the flow of information to and from the brain centre which coordinates the balancing process can result in staggering or falling. This postural condition creates a constant strain on all the muscles employed to retain balance and upon the set of bones forming the spine. The spine is basically a tower of I-beams supports the skeletal frame and, in order to remain in good health, proper mechanical alignment is essential. Any deviation from this mechanical alignment will result in pain relating to non-alignment, such as low back or neck pain. The vulnerability of the back is threatened frequently by work, recreation situations, and furnishings, since their uses subject an already tenuous upright position to undergo increased stresses. As the body compensates for alignment problems by creating excess bone tissue and neural pain, certain arthritic conditions may be the result.

Correction or prevention in tools or activities may assist in the optimization of performance and in more closely aligning the biological with the chronological age. Clearly, optimization and compensation may conflict within the human mechanism since a logical idea may violate physical principles. Based on this introduction of merely a few of the internal and external challenges to the human organism, the need for adequate and accurate assessments, improved tools, and human behavioral modifications becomes more apparent.

With each passing year, the composition of the population in America and probably many other modern societies is becoming older. This population increase of older citizens appears to be due, in part, to the large number of individuals of all ages who are experiencing modifications of lifestyle in a variety of ways, including better working conditions, improved health–medical opportunities, and changing activity levels. Pollock et al. (1) noted that the activity levels of elderly people have increased during the previous 20 years. However, it was estimated that only 10% of elderly individuals participate in regular vigorous physical activity and that 50% of the population who are 60 or more years of age described their lifestyles as sedentary.

Scientific studies and personal experiences continue to link many of the health problems and physical limitations found in the aged to lifestyle. Sedentary living appears to be a major contributor to the significantly adverse effect on health and physical well being. Certainly, there is increasing evidence indicating the vital need for improved national and international policies for better fitness, health, and sports for older individuals. In order to address some of these indicators, new attitudes and policies must emphasize activities and resources to meet the minimal requirements for keeping older people in good health, preventing their deterioration with age, and meeting the special interests of individuals with various disorders. In addition to the difficulties that hospitals, insurance companies, children of the elderly, and legislators face, the medical and scientific communities require time to determine the most appropriate solutions for improving the quality of these lengthening lives.

Many of the myths about aging are being disproved while the true nature of age-related changes appears to be less bleak than previously thought. Disuse and disease, not age alone, are increasingly, revealed as culprits. There is an increasing awareness of the need for more emphasis on fitness to maintain wellness and prevent degenerative illness, for more research to understand the aging body of the healthy older person, and to determine the exercise needs of the ill and/or the handicapped. Pollock et al. (1) noted that physical capacity decrements are normally associated with the aging process. This loss has been attributed to the influence of disease, medication, age, and/or sedentary lifestyle. Additionally, it was noted that the majority of the elderly do not exercise and that it is unclear whether the reduced state of physical conditioning associated with aging results from the deconditioning due to sedentary living, age, or both.

It is a fact of life that muscle tissue suffers some diminution from age. Age-associated changes in organ and tissue function, such as a decline in fat-free mass, total body and intracellular water, and an increase in fat mass (2) may alter the physiological responses to exercise or influence the effect(s) of medication. However, any discussion about age realistically utilizes arbitrary time periods apportioned eons ago by men who evaluated time relative to the number of revolutions of the earth around the sun and the rotation of the earth on its own axis. These predetermined periods may or may not have any relationship with the aging of the cells in the body. The linkage between the chronological age and the biological age of people is imprecise. Perhaps a more accurate consideration of the relationship between chronological and biological age would be one that is nonlinear, may differ with gender, or be dependent on other factors.

It is an inevitable evolutionary consequence that individuals within a species differ in many ways. The characterization of an individual on the basis of a chronological age scale may be practical, but biologically inappropriate. It may be that use or functional activities may have a greater influence on determining biological age rather than the number of times the earth has revolved around the sun. It appears that biological age can be affected by genetic code, nutrition and, most physical activity. Astrand (3) suggested that as an individual ages, the genetic code may have more of an effect on the function of systems with key importance in physical performance. He also noted that a change in lifestyle, at almost any chronological age, can definitely modify the biological age, either upward or downward. It has been suggested that the disparity of older persons is a hallmark of aging itself (4). It is important to determine how much age variance is due to the passage of time and how much is caused by the accumulation of other, nontime dependent, alterations. Previous attitudes towards physical adversities observed in the elderly were that they were attributable to disease. More recently, a third dimension associated with poor health in older persons has been described by Bortz and Bortz (4) as The Disuse Syndrome. For example, one of the most common markers of aging was thought to be a decreased lean body mass. However, analysis of 70 year old weight lifters revealed no such decline. The components of the Disuse Syndrome have been similarly grouped by Kraus and Raab (5) in their book, Hypokinetic Disease, and are (*1*) cardiovascular vulnerability; (*2*) musculoskeletal fragility; (*3*) obesity; (*4*) depression; (*5*) premature aging.

Use is a universal characteristic of life. When any part of the body has little or no use, it declines structurally and

functionally. The effects of disuse can be observed on any body part, such as atrophied intestinal mucosa, when a loop is excluded from digestive functions or the lung becomes atelectatic when not aerated. A lack of adequate conditioning and physical activity causes alterations in the heart and circulatory system, as well as the lungs, blood volume, and skeletal muscle (6–9). During prolonged bed rest, blood volume is reduced, heart size decreases, myocardial mass falls, blood pressure response to exercise increases, and physical performance capability is markedly reduced. On the other hand, although acute changes within the cardiovascular system result in response to increased skeletal muscle demands during exercise, there is evidence that chronic endurance exercise produces changes in the heart and circulation that are organic adaptations to the demands of chronic exercise (10–15).

Cardiac performance undergoes direct and indirect age-associated changes. There is a reduction in contractility of the myocardium (16) and this increased stiffness impairs ventricular diastolic relaxation and increases end diastolic pressure (17). This suggests that exercise-induced increases in heart rate would be less well tolerated in older individuals than in younger populations. The decline in maximal heart rate is known and the cause is multifactorial, but is mostly related to a decrement in sympathetic nervous system response. Fifty percent of Americans who are >65 years of age have a diagnostically abnormal resting electrocardiogram (18). Another factor associated with aging is a progressive increase in rigidity of the aorta and peripheral arteries due to a loss of elastic fibers, increase in collagenous materials, and calcium deposits (19). When aortic rigidity increases, the pulse generated during systole is transmitted to the arterial tree relatively unchanged. Therefore, systolic hypertension predominates in elderly hypertensive patients.

Other bodily systems demonstrate age-related alterations. Baroreceptor sensitivity decreases with age and hypertension (20,21) such that rapid adjustment of the cerebral circulation to changes in posture may be impaired. Kidney function reveals a defect in renal concentrating ability and sluggish renal conservation of sodium intake causes elderly patients to be more susceptible to dehydration (22). Hyaline cartilage on the articulator surface of various joints shows degenerative changes and clinically represents the fundamental alteration in degenerative osteoarthritis (23). A decrease in bone mineral density (osteoporosis) can reduce body stature as well as predispose the individual to spontaneous fractures. Older women are more prone to osteoporosis than older men and this may reflect hormonal differences (23). Older persons are less tolerant of high ambient temperatures than younger people (24) due to a decrease in cardiovascular and hypothalamic function which compromises the heat dissipating mechanisms. Heat dissipation is further compromised by the decrease in fat-free mass, intracellular and total body water, and an increase in body fat.

Unfortunately, the effects of disuse on the body manifest themselves slowly since humans normally have redundant organs that can compensate for ineffectiveness or disease. In addition, humans are opaque so that disease or deterioration are externally unobservable and, thus, go unheeded (e.g., the early changes in bones due to osteoporosis are subclinical and are normally detected only after becoming so pronounced that fractures ensue). Cummings et al. (25) mentioned the difficulty of distinguishing manifestations in musculoskeletal changes due to disease related to aging. Muscle mass relative to total body mass begins decreasing in the fifth decade and becomes markedly reduced during the seventh decade of life. This change results in reduced muscular strength, endurance, size, as well as a reduction in the number of muscle fibers. Basmajian and De Luca (26) reported numerous alterations in the electrical signals associated with voluntary muscular contractions with advancing age. As yet, there are no findings published that have definitively located age-related musculoskeletal changes in either the nervous or the muscular system. The diaphragm and cardiac muscle do not seem to incur age changes. Perhaps this is due to constant use, from exercise, or possibly a genetic survival mechanism.

There is growing consensus that many illnesses are preventable by good health practices including physical exercise. Milliman and Robertson (27) reported that, of the 15,000 employees of a major computer company, the nonexercisers accounted for 30% more hospital stays than the exercisers. Lane et al. (28) reported that regular runners had only two-thirds as many physician visits as community matched controls. The beneficial effect of exercise on diabetes has long been recognized and is generally recommended as an important component in the treatment of diabetes (29). Regular endurance exercise favorably alters coronary artery disease risk factors, including hypertension, triglyceride and high density lipoprotein cholesterol concentrations, glucose tolerance, and obesity. In addition, regular exercise raises the angina threshold (30).

Jokl (31) suggested three axioms of gerontology that are affected by exercise. He contents that sustained training results in the following: (*1*) decline of physique with age; (*2*) decline of physical fitness with age; (*3*) decline of mental functions with age.

Health in older people is best measured in terms of function, mental status, mobility, continence, and a range of activities of daily living. Preventive strategies appear to be able to forestall the onset of disease. Whether exercise can prevent the development of atherosclerosis, delay the occurrence of coronary artery disease, or prevent the evolution of hypertension is at present debatable. But moderate endurance exercise significantly decreases cardiovascular mortality (32). Endurance exercise can alter the contributions of stress, sedentary lifestyle, obesity, and diabetes to the development of coronary artery disease (33).

For example, the four-time Olympic discus champion, Al Oerter, at the age of 43, focused his training to qualify for the 1980 Olympic Games that would have been his fifth consecutive Olympiad Oerter threw his longest throw [220 f (67.05 m)] but, since the United States boycotted the 1980 Moscow Olympic Games, his chance was denied. By the time of the 1984 Los Angeles Games, Oerter was 47 years old. Even at an age well beyond most Olympic competitors, he again threw his best, exceeding 240 f (73.15 m) in practice sessions. Oerter's physique and strength suggested that his biological age was less than his chronological age.

Biologically, he was probably between 25 and 30, although chronologically he was 15–20 years older. Unfortunately, in the competition that determined which athletes would represent the United States, Oerter suffered an injury that precluded him from trying to achieve an unprecedented fifth consecutive Olympic Gold medal.

PRINCIPLES FOR EXERCISE AND TRAINING

Physical fitness and exercise have become, as previously discussed, an increasing concern at nearly all levels of American society. The goal of attaining peak fitness has existed for centuries, yet two problems continue to obfuscate understanding. The ability to assess strength and/or to exercise has occupied centuries of thought and effort. For examples, Milo the Greek lifted a calf each day until the baby grew into a bull. Since this particular procedure is not commonly available, humans have attempted to provide more suitable means to determine strength levels and ways to develop and maintain conditioning. Technology for assessing human performance in exercise and fitness evaluations, in both theory and practice, exhibits two problems. First, a lack of clearly defined and commonly accepted standards results in conflicting claims and approaches to both attaining and maintaining fitness. Second, a lack of accurate tools and techniques for measuring and evaluating the effectiveness of a given device designed to diagnose present capabilities for exercising or even to determine which exercises are appropriate to provide "fitness", regardless of age or gender. Vendors and consumers of fitness technology have lacked sound scientific answers to simple questions regarding the appropriateness of exercise protocols.

Reviewing studies conducted to determine the effects of strength training on human skeletal muscle suggests many benefits with appropriate exercise. In general, strength training that uses large muscle groups in high resistance, low repetition efforts increases the maximum work output of the muscle group stressed (34). Since resistance training does not change the capacity of the specific types of skeletal muscle fibers to develop different tensions, strength is generally seen to increase with the cross-sectional area of the fiber (35). The human body can exercise by utilizing its own mass (e.g., running, climbing, sit ups). These and other forms of nonequipment based exercises can be quite useful. In addition, there are various types of exercise equipment that allow selection of a weight or resistance and then the exercise against that machine resistance is performed.

The relationship between resistance exercises and muscle strength has been known for centuries. Milo the Greek's method of lifting a calf each day until it reached its full growth probably provides the first example of progressive resistance exercises. It has been well-documented in the scientific literature that the size of skeletal muscle is affected by the amount of muscular activity performed. Increased work by a muscle can cause that muscle to undergo compensatory growth (hypertrophy), whereas disuse leads to wasting of the muscle (atrophy).

The goal of developing hypertrophy has stimulated the medical and sports professions, especially coaches and athletes, to try many combinations and techniques of muscle overload. Attempts to produce a better means of rehabilitation, an edge in sporting activities, as a countermeasure for the adverse effects of space flight, or as a means to improve or enhance bodily performances throughout a lifetime have only scratched the surface of the cellular mechanisms and physiological consequences of muscular overload.

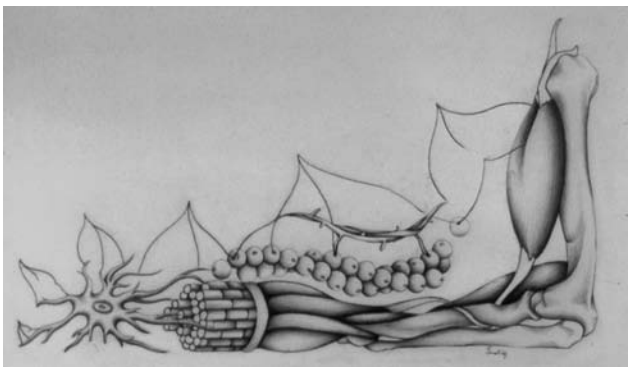


Figure 5. Integration of our muscular system.

Muscular strength can be defined as the force that a muscle group can exert against a resistance in a maximal effort. In 1948, Delorme and Watkins (36) adopted the name "progressive resistance exercise" for his method of developing muscular strength through the utilization of counter balances and weight of the extremity with a cable and pulley arrangement. This technique gave load-assisting exercises to muscle groups that did not perform antigravity motions. McQueen (37) distinguished between exercise regimes for producing muscle hypertrophy and those for producing muscle power. He concluded that the number of repetitions for each set of exercise determines the different characteristics of the various training procedures. Figure 5 illustrates the complexity of the skeletal–muscular structure.

When muscles contract, the limbs may appear to move in unanticipated directions. One type of motion is a static contraction, known as an isometric type of contraction. Another type of contraction is a shortening or dynamic contraction that is called an isotonic contraction. Dynamic contractions are accompanied by muscle shortening and by limb movement. Dynamic contractions can exhibit two types of motion. One activity is a concentric contraction in which the joint angle between the two bones become smaller as the muscular tension is developed. The other action is an eccentric contraction in which, as the muscles contract, the joint angle between the bones increases. Owing to ambiguity in the literature concerning certain physiologic terms and differences in laboratory procedures, the following terms are defined below.

1. *Muscular strength*: the contractile power of muscles as a result of a single maximum effort.
2. *Muscular endurance*: ability of the muscles to perform work by holding a maximum contraction for a given length of time or by continuing to move submaximal load to a certain level of fatigue.

3. *Isometric*: a muscular contraction of total effort but with little or no visible limb movement (sometimes referred to as static or anaerobic).
4. *Isotonic*: a muscular contraction of less than total effort with visible limb movement (sometimes called dynamic or aerobic).
5. *Isokinetic training (accommodating resistance)*: muscular contraction at a constant velocity. In other words, as the muscle length changes, the resistance alters in a manner that is directly proportional to the force exerted by the muscle.
6. *Concentric contraction*: an isotonic contraction in which the muscle length decreases (that is, the muscle primarily responsible for movement becomes shorter).
7. *Eccentric contraction*: an isotonic contraction in which the muscle length of the primary mover and the angle between the two limbs increases during the movement.
8. *Muscle overload*: the workload for a muscle or muscle group that is greater than that to which the muscle is accustomed.
9. *Variable resistance exercise*: as the muscle contracts, the resistance changes in a predetermined manner (linear, exponentially, or as defined by the user).
10. *Variable velocity exercise*: as the muscle contracts with maximal or submaximal tension, the speed of movement changes in a predetermined manner (linear, exponentially, or as defined by the user).
11. *Repetitions*: the number of consecutive times a particular movement or exercise is performed.
12. *Repetition maximum (1 RM)*: the maximum resistance a muscle or muscle group can overcome in a maximal effort.
13. *Sets*: the number of groups of repetitions of a particular movement or exercise.

Based on evidence presented in these early studies (36–38), hundreds of investigations have been published relative to techniques for muscular development, including isotonic exercises, isometric exercises, eccentric contractions, and many others. The effectiveness of each exercise type has been supported and refuted by numerous investigations, but no definitive, irrefutable conclusions have been established.

Hellebrandt and Houtz (38) shed some light on the mechanism of muscle training in an experimental demonstration of the overload principle. They found that the repetition of contractions that place minimal stress on the neuromuscular system had little effect on the functional capacity of the skeletal muscles. They also found that the amount of work done per unit of time is the critical variable upon which extension of the limits of performance depends. The speed with which functional capacity increases suggests that the central nervous system, as well as the contractile tissue, is an important contributing component of training.

Results from the work of Hellebrandt and Houtz (38) suggest that an important consideration in both the design of equipment for resistive exercise and the performance of an athlete or a busy executive is that the human body relies on preprogrammed activity by the central nervous system. Since most human movements are ballistic and the neural control of these patterns differs from slow controlled movements, it is essential that training routines employ programmable motions to suit specific movements. This control necessitates exact precision in the timing and coordination of both the system of muscle contraction and the segmental sequence of muscular activity. Research has shown that a characteristic pattern of motion is present during any intentional movement of body segments against resistance. This pattern consists of reciprocally organized activity between the agonist and antagonist. These reciprocal activities occur in consistent temporal relationships with the variables of motion, such as velocity, acceleration, and forces.

In addition to the control by the nervous system, the human body is composed of linked segments, and rotation of these segments about their anatomic axes is caused by force. Both muscle and gravitational forces are important in producing these turning effects, which are fundamental in body movements in all sports and daily living. Pushing, pulling, lifting, kicking, running, walking, and all human activities result from the rotational motion of the links which, in humans, are the bones. Since force has been considered the most important component of athletic performance, many exercise equipment manufacturers have developed various types of devices employing isometrics and isokinetics. When considered as a separate entity, force is only one factor influencing successful athletic performance. Unfortunately, these isometric and isokinetic devices inhibit the natural movement patterns of acceleration and deceleration.

The three factors underlying all athletic performances and the majority of routine human motions are force, displacement, and the duration of movement. In all motor skills, muscular forces interact to move the body parts through the activity. The displacement of the body parts and their speed of motion are important in the coordination of the activity and are also directly related to the forces produced. However, it is only because of the control provided by the brain that the muscular forces follow any particular displacement pattern and, without these brain centre controls, there would be no skilled athletic performances. In every planned human motion, the intricate timing of the varying forces is a critical factor in successful performances. In any human movement, the accurate coordination of the body parts and their velocities is essential for maximizing performances. This means that the generated muscular forces must occur at the right time for optimum results. For this reason, the strongest weightlifter cannot put the shot as far as the experienced shotputter, although the weightlifter possesses greater muscular force, he has not trained his brain centers to produce the correct forces at the appropriate time. Older individuals may be unable to walk up and down stairs or perform many of the daily, routine functions that had been virtually automatic before the deterioration produced by weakness, disease, or merely age.

There are significant differences in the manner of execution of the various resistive training methods. In isotonic exercises, the inertia, which is the initial resistance, must be overcome before the execution of the movement progresses. The weight of the resistance cannot be heavier than the maximum strength of the weakest muscle acting in a particular movement or the movement cannot be completed. Consequently, the amount of force generated by the muscles during an isotonic contraction does not maintain maximum tension throughout the entire range of motion. In an isokinetically loaded muscle, the desired speed of movement occurs almost immediately and the muscle is able to generate a maximal force under a controlled and specifically selected speed of contraction.

The use of the isokinetic principle for overloading muscles to attain their maximal power output has direct applications in the fields of sport medicine and athletic training. Many rehabilitation programmed utilize isokinetic training to recondition injured limbs of athletes to their full range of motion. The unfortunate drawback to this type of training is that the speed is constant and there are no athletic activities that are performed at a constant velocity. The same disadvantage applies to normal human activities.

In isotonic resistive training, if more than one repetition is to be used, a submaximal load must be selected for the initial contractions in order to complete the required repetitions. Otherwise, the entire regimen would not be completed, owing to fatigue or, the inability to perform. A modality that can adjust the resistance so that it parallels fatigue to allow a maximum effort for each repetition would be a superior type of equipment. This function could be accomplished by manually removing weight from the bar while the subject trained. This is neither convenient nor practical. With the aid of the computer, the function can be performed automatically.

Another drawback with many isotonic types of resistive exercises is that the inertia resulting from the motion changes the resistance depending on the acceleration of the weight and of the body segments. In addition, since overload on the muscle changes due to both biomechanical levers and the length–tension curve, the muscle is able to achieve maximal overload only in a small portion of the range of motion. To overcome this shortcoming in resistive training, some strength training devices have been introduced that have "variable resistance" mechanisms, such as a cam, in them. However, these variable resistance systems increase the resistance in a linear fashion and this linearity may not truly accommodate the individual. When including inertial forces to the variable resistance mechanism, the accommodating resistance can be canceled by the velocity of the movement.

There seem to be unlimited training methods and each is supported and refuted by as many "experts". In the past, the problem of accurately evaluating the different modes of exercise was rendered impossible because of the lack of adequate diagnostic tools. For example, when trying to evaluate isotonic exercises, the investigator does not know exactly the muscular effort nor the speed of movement, but knows only the weight that has been lifted. When a static weight is lifted, the force of inertia provides a significant contribution to the load and cannot be quantified by feel or observation alone. In the isokinetic mode, the calibration of the velocity is assumed, but has been poorly verified since the mere rotation of a dial to a specific speed setting does not guarantee the accuracy of subsequently generated velocity. In fact, discrepancies as great as 40% have been observed when verifying the bar velocity.

Most exercise equipment currently available lack intelligence. In other words, the equipment is not aware that a subject is performing an exercise or how it is being conducted. Verification of the speed is impossible since a closed-loop feedback and sensors are absent. However, with the advent of miniaturized electronics in computers, it became possible to unite exercise equipment with the computer's artificial intelligence. In other words, it became possible for exercise equipment to adapt to the user rather than forcing the user to adapt to the equipment.

HIGH TECHNOLOGY TOOLS

High technology refers to the use of advanced, sophisticated, space age mathematical and electronic methods and devices for creating tools that can enhance human activities as well as expanding the horizons for future inventions. NASA put a man on the moon, sent exploratory spacecraft to Mars and beyond, and is sending shuttle missions to the Space Station. Polymer science invented plastics, mechanical science produced the automobile, and aeronautical engineering developed the airplane. Despite all of the knowledge and explosive developments since the rock became a tool, few advances have considered first the most important component in a complicated system, the human body.

The usual developmental cycle creates something and humans must adapt to it rather than the reverse. Computers can provide precise computations rapidly for complex problems that would otherwise require enormous quantities of time, talent, and energy to complete. The strength of these electronic wizards to follow instructions exactly, remember everything, and perform calculations within thousandths of a second has made them indispensable in finance, industry, and government. Application of the computer was a perfect enhancement for the human mind in order to quantify and evaluate movement performances. Used in conjunction with the human mind's ability to deduce, interpret, and judge, the computer provides the necessary enhancement to surpass the limits of what the eye can see or what intuition can surmise. Technological advances, such as these, can assist humans irrespective of their age.

For good health, it is necessary to follow a training method that incorporates all of the various bodily systems. In other words, the body should be treated as a complex, but whole, entity rather than as isolated parts. While it is not wrong to evaluate one's diet, an assessment of health would be incomplete without consideration of physical training, stress reduction, and other components that constitute the integrated organism of the human body. For a person to be able to jog 5 miles it is not important only to

run, but to develop the cardiovascular system in a systematic way to achieve a healthy status. Strength exercise, flexibility routines, proper nutrition and skill are necessary to achieve this goal.

Two sophisticated systems have been developed to analyze human performance and both are appropriate for the assault on aging. These systems include tools to (*1*) assess movements of the human body and (*2*) assist in exercising human beings. The first one is the biomechanical system that was developed to analyze movement performance. Currently, biomechanical analyses are routinely performed on a wide range of human motions in homes, work settings, recreation, hospitals, and rehabilitation centers. The second system, which incorporates space age technology, allows diagnoses and training of the musculoskeletal system. Each of these systems will be discussed subsequently in detail. Both of these technologies and the scientific principles and techniques discussed may help achieve physical and mental goals. The technological advances provide tools for quantification of the results and to analyze the potential of a person. With this information and these tools, it should be possible to train the various body systems for optimal results at any age.

The first commercially available computerized biomechanical system was described in 1973 (39) and that system can serve to illustrate the general concepts and procedures associated with biomechanical quantification of movement. Figure 6 illustrates device system. The computerized hardware–software system provides a means to objectively quantify the dynamic components of movement in humans, such as athletic events, gait analyses, work actions, as well as motion by inanimate objects, including such items as machinery actions, air bag activation, and auto crash dummies. This objective technique replaces mere observation and supposition. This system provides a means to quantity motion utilizing input information from any or all of the following mediums: visual (video), electromyography (EMG), force platforms, or other signal processing diagnostic equipment.

Figure 6. Analyses of vertical jump.

The Ariel Performance Analysis System provides a means of measuring human motion based on a proprietary technique for the processing of multiple high speed video recordings of a subject's performance (40–42). This technique demonstrates significant advantages over other common approaches to the measurement of human performance. First, except in those specific applications requiring EMG or kinetic (force platform) data, it is noninvasive. No wires, sensors, or markers need be attached to the subject. Second, it is portable and does not require modification of the performing environment. Cameras can be taken to the location of the activity and positioned in any convenient manner so as not to interfere with the subject. Activities in the workplace, home, hospital, therapist's office, health club, or athletic field can be studied with equal ease. Third, the scale and accuracy of measurement can be set to whatever levels are required for the activity being performed. Camera placement, lens selection, shutter and film speed may be varied within wide limits to collect data on motion of only a few centimeters or of many meters, with a duration from a few milliseconds to a number of seconds. Video equipment technology currently available is sufficiently adequate for most applications requiring accurate motion analysis. Determination of the problem, error level, degree of quantification, and price affect the input device selection.

A typical kinematic analysis consists of four distinct phases: data collection (filming); digitizing; computation; and presentation of the results. Data collection is the only phase that is not computerized. In this phase, video recordings of an activity are made using two or more cameras with only a few restrictions: (*1*) all cameras must record the action simultaneously. (*2*) If a fixed camera is used, it must not move between the recording of the activity and the recording of the calibration points. These limiting factors are not necessary when a panning camera and associated mechanism are used. A specialized device accompanied by specialized software was developed to accommodate camera movement particularly for use with gait analysis and some longer distance sporting events, such as skiing or long jumping. (*3*) The activity must be clearly seen throughout its duration from at least two camera views. (*4*) The location of at least six fixed noncoplanar points visible from each camera view (calibration points) must be known. These points need not be present during the activity as long as they can be seen before or after the activity. Usually they are provided by some object or apparatus of known dimensions that is placed in the general area of the activity, filmed and then removed. (*5*) The speed of each of the cameras (frames/second) must be accurately known, although the speeds do not have to be identical. (*6*) Some

Figure 7. Digitizing system.

event or time signal must be recorded simultaneously by all cameras during the activity in order to provide synchronization.

These rules for data collection allow great flexibility in the recording of an activity. Figure 7 illustrates a modern digitizing system to quantify human movement. Information about the camera location and orientation, the distance from camera to subject, and the focal length of the lens is not needed. The image space is self-calibrating through the use of calibration points that do not need to be present during the actual performance of the activity. Different types of cameras and different film speeds can be used and the cameras do not need to be mechanically or electronically synchronized. The best results are obtained when camera viewing axes are orthogonal (90° apart), but variations of 20–30° can be accommodated with negligible error. Initially, the video image is captured by the computer and stored in memory. This phase constitutes the "Grabbing" mode. Brightness, contrast, saturation, and color can be adjusted so that the grabbed picture may, in fact, be better than the original. Grabbing the image and storing it on computer memory eliminates any further need for the video apparatus.

Digitizing is the third step in biomechanical quantification. The image sequence is retrieved from computer memory and displayed, one frame at a time, on the digitizing monitor. Using a video cursor, the location of each of the subject's body joints (e.g., ankle, knee, hip, shoulder, elbow) is selected and stored in computer memory. In addition, a fixed point, which is a point in the field of view that does not move, is digitized for each frame as an absolute reference. The fixed point allows for the simple correction of any registration or vibration errors introduced during recording or playback. At some point during the digitizing of each view, a synchronizing event must be identified and, additionally, the location of the calibration points as seen from that camera must be digitized. This sequence of events is repeated for each camera view. This type of digitizing is primarily a manual process.

An alternative digitizing option permits the procedure to proceed automatically using any number of marker sets. This requires that the subject have the markers placed on the body prior to the filming phase. The types of markers and their placements have a substantial number of adherents particularly in the rehabilitation, gait measurement, and computer game communities. This type of digitizing combines manual and automatic, so that the activity progresses under manual control with computer-assisted selection of the joint segments or points. User participation in the digitizing process provides an opportunity for error checking and visual feedback which rarely slows the digitizing process adversely. A trained operator, with reasonable knowledge about digitizing and anatomy, can rapidly produce high quality digitized images. It is essential that the points are selected precisely because all subsequent information is based on the data provided in this phase.

The computation phase of analysis is performed after all camera views have been digitized. At this point in the procedures, the three-dimensional (3D) coordinates of the joints centers of a body are calculated. The transformation methods for transforming the data to two-dimensional (2D) or 3D coordinates are Direct Linear Transformation, Multiplier, and Physical Parameters Transformation. This phase computes the true 3D image space coordinates of the subject's body joints from the 2D digitized coordinates obtained from each camera's view. The Direct Linear Transformation Computation is determined by first relating the known image space locations of the calibration points to the digitized coordinate locations of those points. The transformation is then applied to the digitized body joint locations to yield true image space locations. This process is performed under computer control with some timing information provided by the user. The information needed includes, for example, starting and ending points if all the data are not to be used, as well as a frame rate for any image sequence that differs from the frame rate of the cameras used to record the sequence. The Multiplier technique for transformation is less rigorous mathematically and is utilized for those situations when no calibration device was used and only a few objects in the background are available to calibrate the area. This situation usually occurs when a nonscientific, third-party recorded the pictures such as a home video or even a televised sporting event. The third type of transformation, the Physical Parameters Transformation, is primarily applied with panning camera views or when greater accuracy is required on known image sources.

Following data transformation, a smoothing or filtering operation is performed on the image coordinates to remove small random digitizing errors and to compute body joint

velocities and accelerations. Smoothing options include polynomial, cubic and quintic splines, a Butterworth second-order digital and fast Fourier filters (43–45). Smoothing may be performed automatically by the computer or interactively with the user controlling the amount of smoothing applied to each joint. Error measurements from the digitizing phase may be used to optimize the amount of smoothing selected. Another unique feature is the ability to display the Power Spectrum for each of the x, y, and z coordinates. This enhancement permits the investigator to evaluate the effect of the smoothing technique and the chosen value selected for that curve by examining the Power Spectrum. Thus, the investigator can determine the method and level of smoothing that best meets the requirements of the specific research. After smoothing, the true 3D body joint displacements, velocities, and accelerations will have been computed on a continuous basis throughout the duration of the sequence.

Analogue data can be obtained from as many as 256 channels for input into the analogue-to-digital (A/D) system. Processing of the analogue signals, such as those obtained from transducers, thermistors, accelerometers, force platforms, EMG, ECG, EEG, or others, can be recorded for analysis and, if needed, synchronized with the video system. The displayed video picture and the vectors from the force plate can be synchronized so that the force vectors appear to be "inside the body". At this point, optional kinetic calculations can be performed to provide for measurement and analysis of the external forces that are applied to the body during movement. Inverse Dynamics are used to compute joint forces and torques as well as energy and momentum parameters of single or combined segments. External forces include anything external to the body that is applying force or resistance such as a golf club held in the hand. The calculations that are performed are made against the force distribution of the body.

The presentation phase of analysis allows computed results to be viewed and recorded in a number of different formats. Body position and motion can be presented in both still frame and animated stick figure format in 3D. Multiple stick figures may be displayed simultaneously for comparison purposes. Joint velocity and acceleration vectors may be added to the stick figures to show the magnitude and direction of body motion parameters. Copies of these displays can be printed for reporting and publication. Results can also be reported graphically. Plots of body joints and segments, linear and angular displacements, velocities, accelerations, forces, and moments can be produced in a number of format options. An interactive graphically oriented user interface allows the selection and plotting of such results to be simple and straightforward. In addition, body motion parameter results may also be reported in numerical form and printed as tables.

Utilizing this computerized system for biomechanical quantification of various movements performed by the elderly may assist in developing strategies of exercise, alterations in lifestyle, modifications in environmental conditions, and inventions to ease and/or extend independence. For example, rising from a chair is a challenging task for many elderly persons and getting up quickly is associated with a particularly high risk for falling. Hoy and Marcus (46) observed that older women moved more slowly and altered their posture to a greater extent than younger women. The strength levels were greater for the younger subjects, but it could not be concluded that strength was the causal mechanism for the slower speed. Following an exercise program affecting a number of muscle groups, younger and older women significantly increased in strength. Results of this study suggest that age-associated changes in muscle strength have an important effect on movement strategies used during chair rising. Following participation in a strength-training program, biomechanical assessment revealed changes in movement strategies that increased both static and dynamic stability. Other areas appropriate for biomechanical assessment would be on the well-known phenomenon of increased postural sway (47) and problems with balance (48–50) in the aged.

It is also important to study the motor patterns used by older persons while performing locomotor tasks associated with daily life such as walking on level ground and climbing or descending stairs. Craik (51) demonstrated that older subjects walking at the same speed as younger ones exhibited similar movement characteristics. Perhaps the older subjects selected slower movement speeds that produced apparent rather than real reductions in performance. These types of locomotors studies are easily assessed by biomechanical procedures. A biomechanical inquiry by Williams (52) examined the age-related differences of intralimb coordination by young and old individuals. Williams observed a similarity of general intralimb coordination for both old and young participants for level ground motions. One age-related change was suggested with regard to the additional balance constraints required for going up stairs because of adjustments not required on level ground. More profound differences were observed by Light et al. (53) with complex, multilimb coordinated movements performed in a standing position which necessitated dynamic balance control. These types of tasks showed significant age-dependent changes. Compared with younger subjects, the older participants were slower in all timing components, had less predominance in their movement patterns, less coupling of their limbs for movement end-points, and were more susceptible to environmental uncertainties. The alterations in movement performance reflected age-related loss in the ability to coordinate fast, multilimb movements performed from an upright stance suggesting that older individuals may have uncoordinated and unpredictable movement patterns when required to move quickly. Additionally, it was suggested that the more uncertain the environment, the greater the disturbance on the movement, thus, increasing the risk of falling. These studies provide realistic examples of one role biomechanics can perform by not only specifically identifying the locus of change but also providing objective quantification.

Another interesting application of the biomechanical system involves a multidimensional study of Alzheimer's disease currently in progress at a leading medical school. The study's strength is similar to the blind men who must integrate all of the information each has gathered in order to accurately describe the elephant. Examination of the

brain's response to specific drugs and at varying dosages, magnetic resonance imaging (MRI), thermographic, endocrine, and hormonal changes, vascular chemistry, as well as other aspects are being evaluated for each patient and their specific motor performances are being quantified biomechanically with the Ariel Performance Analysis system. Preliminary evidence indicates that performance on a simple bean-bag tossing skill improves daily although there is no cognitive recognition of the task. The activity of tossing a bean bag into a target circle from a standing position employs postural adjustments as well as coordinated arm and hand directed skills. Skill acquisition, or motor learning, involves both muscular capability and neural control mechanisms. Both activities involve closed- and open-loop mechanisms. The goal-directed movements needed to perform the bean-bag toss require the anticipatory postural adjustments that are inherent in an open-loop control. Because these findings suggest that muscular control and skill acquisition remain viable, this enables investigators to narrow the direction of the research and continue the study while continuously honing the focus. With each scientific finding, the research can be directed toward identification of the underlying cause.

The preceding discussion has described a computerized biomechanical system that can be utilized for the quantification of activities and performance levels particularly where appropriate for gerontological issues. Following the identification and definition of an activity, a second and equally necessary component follows. This is the ability to evaluate, test, and/or train the musculoskeletal components of the body in a manner appropriate to the specifically identified task(s) and according to the capabilities of the age and health of the individual. The integration of both technological assessment tools should assist the individual and others involved in their daily life to identify and measure those portions of an exercise program that can enhance performance, fitness status, or exercise capabilities for each gender and at different ages. In other words, one of the principles should be remembered is the goal of optimizing performance at every age.

For centuries, many devices have been created specifically for strength development. These devices include treadmills, bicycle ergometers, rowing machines, skiing simulators, as well as many of the more traditional resistive exercises with dumbbells, bar bells, and commercially available weight equipment. Figure 8 illustrates one of these equipment. Each type of exercise has some advantages, but none are designed to cope with the difficulties inherent with the gravitational effects that affect the multilinked human body performing on various exercise equipment.

All systems that employ weights as the mechanism for resistance have major drawbacks in four or more areas, as follows: (*1*) biomechanical considerations; (*2*) inertia; (*3*) risk of injury; (*4*) unidirectional resistance.

The biomechanical parameters are extremely important for human performance and should be incorporated into exercise equipment. The biomechanical factors were discussed previously. Inertia is the resistance to changes in motion. In other words, a greater force is required to begin moving weights than is necessary to keep them moving.

Figure 8. The computerized exercise equipment.

Similarly, when the exercising person slows at the end of a movement, the weights tend to keep moving until slowed by gravity. This phenomenon reduces the force needed at the end of a motion sequence. Inertia becomes especially pronounced as acceleration and deceleration increase, effectively reducing the useful range of motion of weight-based exercise equipment.

The risk of injury is obvious in most weight-based exercise equipment. When weights are raised during the performance of an exercise, they must be lowered to their original resting position before the person using the equipment can release the equipment and stop exercising. If the person exercising loses their grip, or is unable to hold the weights owing to exhaustion or imbalance, the weights fall back to their resting position; serious injuries can, and have, occurred. Finally, while being raised or lowered, weights, whether on exercise equipment or free standing, offer resistance only in the direction opposite to that of gravity. This resistance can be redirected by pulleys and gears but still remains unidirectional.

In almost every exercise performed, the muscle or muscles being trained by resistance in one direction are balanced by a corresponding muscle or muscles that could be trained by resistance in the opposite direction. With weight-based systems, a different exercise, and often a different mechanism, is necessary to train these opposing muscles. Exercise mechanisms that employ springs, torsion bars, and the like are able to overcome the inertia problem of weight-based mechanisms and, partially, to compensate the unidirectional force restriction by both expanding and compressing the springs. However, the serious problem of safety remains. An additional problem is the fixed, nonlinear resistance that is characteristic of springs and is usually unacceptable to most exercise equipment users.

The third resistive mechanism commonly employed in existing exercise equipment is a hydraulic mechanism. Hydraulic devices are able to overcome the inertial problem of weights, the safety problem of both weights and springs, and, with the appropriate selection or configuration, the unidirectional problem. However, previous applications of the hydraulic principle have demonstrated a serious

deficiency that has limited their popularity in resistive training. This deficiency is that of a fixed or a preselected flow rate through the hydraulic system. With a fixed-flow rate, it is a well established fact that resistance is a function of the velocity of the piston and, in fact, varies quite rapidly with changes in velocity. It becomes difficult for a person exercising to select a given resistance for training due to the constraint of moving either slower or faster than desired in order to maintain the resistance. Additionally, at any given moment, the user is unsure of just what the performing force or velocity actually is.

In the field of rehabilitation (54) especially, isokinetic or constant velocity training equipment is a technology that has enjoyed wide acceptance. These mechanisms typically utilize active or passive hydraulics or electric motors and velocity-controlling circuitry. The user or practitioner selects a constant level of velocity for exercise and the mechanism maintains this velocity while measuring the force exerted by the subject. Although demonstrating significant advantages over weight-based systems, isokinetic systems possess a serious limitation. There are virtually no human activities that are performed at a constant velocity. Normal human movement consists of patterns of acceleration and deceleration. When a person learns to run, ride a bike, or write, an acceleration–deceleration sequence is established that may be repeated at different rates and with different levels of force, but always with the pattern unique to that activity. To train, rehabilitate, or diagnose at a constant velocity is to change the very nature of the activity being performed and to violate most biomechanical performance principles.

FEEDBACK CONTROL OF EXERCISE

A newer form of exercise equipment can determine the level of effort by the person, compare it to the desired effort, and then adjust accordingly. The primary advantage of this resistive mechanism is that the pattern of resistance or the pattern of motion is fully programmable. The concept of applying a pattern of resistance or motion to training and rehabilitation was virtually impossible until the invention of computerized feedback control. Prior to the introduction of computerized feedback control, fitness technology could provide only limited modes of resistance and motion. Bar bells or weights of any type provide an isotonic or constant resistance type of training only when moved at a constant velocity. Typically, users are instructed to move the weights slowly to avoid the problem of inertia resulting from the acceleration or deceleration of mass. Weights used with cams or linkages that alter the mechanical advantage can provide a form of variable resistance. However, the pattern is always fixed and the varying mechanical advantage causes a variation in velocity that increases inertial effects. Users must move the weights slowly to preserve the resistance pattern. Another deficiency with these types of equipment is that they do not approximate the body or limb movement pattern of a normal human activity.

An exercise machine controlled by a computer possesses several unique advantages over other resistive exercise mechanisms, both fixed and feedback controlled. The most significant of these advances is the introduction of software to the human/computer feedback loop. The computer and its associated collection of unique programs can regulate the resistance to vary with the measured variables of force and displacement as well as modify the resistance according to data obtained from the feedback loop while the exercise progresses. This modification can, therefore, reflect changes in the pattern of exercise over time. The unique programmed selection can effect such changes in order to achieve a sequential or patterned progression of resistance for optimal training effect. The advantage of this capability over previous systems is that the user can select the overall pattern of exercise and the machine assumes responsibility for changing the precise force level, the speed of movement, and the temporal sequence to achieve that pattern.

There are a wide range of treadmills, bikes, and exercise devices currently available that employ electrical control features. These include such options as fat burn, up hill training, or cardiac modes. These types of equipment change the speed or elevations with preprogrammed actions that are determined at the manufacturing center when the machines are made rather than by the person exercising. The exerciser can select the programs presented on the control panel, but the response by the machine to the user is not at all related to the performance but rather to the preset events stored in the memory. Therefore, the person may be running "uphill" on the treadmill as determined by the imbedded system, but not with responsive interaction between the equipment and the individual moment by moment. This is a limitation of most of the exercise equipment available in the marketplace of the twenty-first century.

In the early 1980s, the first resistive training and rehabilitation device to employ computerized feedback control of both resistance and motion during exercise was introduced to overcome the lack of machine–human interactivity (55). For the first time, a machine dynamically adapted to the activity being performed rather than the traditional approach of modifying the activity to conform to the limitations of the machine. Biomechanical results previously calculated could be used to program the actual patterns of motion for training or rehabilitation. The equipment utilizes a passive hydraulic resistance mechanism operating in a feedback-controlled mode under control of the system's computer.

A simplified functional description of this mechanism, the Ariel Computerized Exercise System, and its operation is described. A hydraulic cylinder is attached to an exercise bar through a mechanical linkage. As the bar is moved, the piston in the hydraulic cylinder moves which pushes oil from one side of the cylinder, through a valve, and into the other side of the cylinder. When the valve is fully open there is no resistance to the movement of oil and, thus, no resistance in the movement of the bar. As the valve is closed, it becomes harder to push the oil from one side of the cylinder to the other and, thus, harder to move the bar. When the valve is fully closed, oil cannot flow and the bar will not move. In addition to the cylinder, the resistance mechanism contains sensors to measure the applied force on the bar and the motion of the bar. To describe

the operation of the computerized feedback loop, assume the valve is at some intermediate position and the bar is being moved at some velocity with some level of resistance. If the computer senses that the bar velocity is too high or that bar resistance is too low, it will close the valve by a small amount and then check the velocity and resistance values again. If the values are incorrect, it will continue to regulate the opening of the valve and continually check the results until the desired velocity or resistance is achieved. Similar computer assessments and valve adjustments are made for every exercise. Thus, an interactive feedback loop between the computer and the valve enable the user to exercise at the desired velocity or resistance. The feedback cycle occurs hundreds of times a second so that the user experiences no perceptible variations from the desired parameters of exercise.

There are a number of advantages in a computerized feedback controlled resistance mechanism over devices that employ weights, springs, motors, or pumps. One significant advantage is safety. The passive hydraulic mechanism provides resistance only when the user pushes or pulls against it. The user may stop exercising at any time and the exercise bar will remain motionless. Another advantage is that of bidirectional exercise. The hydraulic mechanism can provide resistance with the bar moving in each direction, whereas weights and springs provide resistance in only one direction. Opposing muscle groups can be trained in a single exercise. Two additional problems associated with weight training, noise and inertia, are also eliminated because the hydraulic mechanism is virtually silent and full resistance can be maintained at all speeds. Figure 9 illustrates an olympic training system utilized by the olympic athletes.

The Ariel Computerized Exercise System allows the user to set a pattern of continuously varying velocity or resistance. The pattern can be based on direct measurements of that individual's motion derived from the biomechanical analysis or can be designed or created by the user with a goal of training or rehabilitation. During exercise, the computer uses the pattern to adjust bar velocity or bar resistance as the subject moves through the full range of motion. In this manner, the motion parameters of almost any activity can be closely duplicated by the exercise system allowing training or rehabilitation using the same pattern as the activity itself.

The software consists of two levels. One level of software is invisible to the individual using the equipment since it controls the hardware components. The second level of software allows interaction between the user and the computer. The computer programs necessary to provide the real-time feedback control, the data program and storage, and the additional performance manipulations are extensive. The software provides computer interaction with the individual operator by automatically presenting a menu of options when the system is activated. Selection of the diagnostics option allows several parameters about that person to be evaluated and stored if desired. Some of the diagnostic parameters available include range of motion, maximum force, and maximum speed that the individual can move the bar for the specific activity selected. The maximum force and maximum speed data

Figure 9. Olympic training on the computerized exercise system.

can be determined at each discrete point in the range of movement as well as the average across the entire range. The diagnostic data can be used solely as isolated pre- and post-test measurements. However, the data can also be stored within the person's profile so that subsequent actions and tests performed on the equipment can be customized to adjust to that specific individual's characteristics.

The controlled velocity option permits the individual to control the speed of bar movement. The pattern of the velocity can be determined by the person using the equipment and these choices of velocity patterns include: (*1*) isokinetic, which provides a constant speed throughout the range of motion; (*2*) variable speed, in which the speed at the beginning of the motion and the speed at the end of the stroke are different with the computer regulating a smooth transition between the two values; and (*3*) programmed speed, which allows the user to specify a unique velocity pattern throughout the range of movement. For each of the choices, determination of the initial and final velocities is at the discretion of the individual through an interactive menu. The number of repetitions to be performed can be indicated by the person. Also, it is possible to designate different patterns of velocity for each direction of bar movement.

The controlled resistance option enables the person to control the resistance or amount of force required to move the bar. The alternatives include (*1*) isotonic, which provides a constant amount of force for the individual to overcome in order to move the bar; (*2*) variable resistance, in which the force at the beginning of the motion and the force at the end of the movement are different with the computer regulating a smooth transition between the two values; (*3*) programmed resistance, which permits the individual to specify a unique force pattern through-out the range of movement. An interactive menu enables the person to indicate the precise initial and final values, the number of repetitions to be used, and each direction of bar motion for the three choices. The controlled work option allows the individual to determine the amount of work, in Newton/meters or joules, to be performed rather than the number of repetitions. In addition, the person can choose either velocity or resistance as the method for controlling the bar movement. As with the previous options, bidirectional control is possible. The data storage capability is useful in the design of research protocols. The software allows an investigator to program a specific series of exercises and the precise manner in which they are to be performed, for example, number of repetitions and amount of work, so that the user need only select their name from the graphic menu and the computer will then guide the procedures. Data gathered can be stored for subsequent analysis. The equipment is fully operational for all options irrespective of whether the data storage option is activated.

Numerous features further enhance the application of this advanced fitness technology. Individual exercise programs can be created and saved on the computer, a CD, an internet file, or a USB disk. Users can perform their individual program at any time merely by loading it from any of the memory options used. Measurements of exercise results can be automatically saved and progress monitored by comparing current performance levels to previous ones. Performance can be measured in terms of strength, speed, power, repetitions, quantity of work, endurance and fatigue. Comparison of these quantities can be made for flexors versus extensors, right limb versus left limb, as well as between different dates and different individuals. Visual and audio feedback are provided during exercise to ensure that the subject is training in the proper manner and to provide motivation for optimal performance. Accuracy of measurement is essential and it is deemed as one of the most important considerations in the software. Calibration of the equipment is performed dynamically and is a unique feature that the computerization and the feedback system allow. Calibration is performed using weights with known values and the procedure can be performed for both up and down directions. This type of calibration is unique since the accuracy of the device can be ascertained throughout the range of motion.

FUTURE DEVELOPMENTS

As discussed previously, a large diagnostic and/or exercise system exists, but sheer bulk precludes its convenient use at home or in small spaces. One future goal is to develop a computerized, feedback-controlled, portable, battery-powered, hydraulic musculoskeletal exercise assessment and training equipment based on the currently available full-sized system. The device will be portable, compact, and operate at low voltage. Although physical fitness and good health have become increasingly more important to the American public, no compact, affordable, accurate device either for measurement or conditioning human strength or performance exists. This deficit hinders both America's ability to provide convenient, affordable, and accurate diagnostic and exercise capabilities for hospital or home-bound patients, children or elderly, to adequately perform within small-spaced military areas, as would be found in submarines, or in NASA shuttle projects to explore the frontiers of space. Figure 10 illustrates an astronaut running on a computerized treadmill in a zero gravity environment.

Figure 10. Motion analysis in space.

The frame will be compact and light-weight with a target weight of < 10 kg. This is an ambitious design goal that will require frame materials to have maximum strength/weight ratios and the structure must be engineered with attention directed toward compactness, storage size, and both ease and versatility of operation. The design of a smaller and lighter hydraulic valve, pack, and cylinder assembly is envisioned. Software can be tailored to specific applications such as for the very young or the aged, specific orthopedic and/or disease training, or other applications.

Another future development will be the ability to download programs through the Internet. For example, each patient could have one of the small exercise devices at home. His/her doctor can prescribe certain diagnostic activities and exercise regiments and transmit them via the Internet. The individual can perform the exercises at home and then submit the results to the doctor electronically. Biomechanical quantification of performances will become available electronically by downloading the software and executing the procedures on the individual's personal computer. Parents will be able to assist their child's athletic and

growth performances, doctors or physical therapists can compare normal gait with their patient's, and many other uses which may not be apparent at this time. The Internet can also function as a conduit between a research site and a remote location. Consider a hypothetical example of the National Institute of Health conducting a study on the effects of exercise on various medical, chemical, neural, and biomechanical factors for a large number of subjects around the world. The exercise equipment could be linked directly with Internet sources; the other data could be collected, and sent to the appropriate participating institutes. Findings from each location could then be transmitted to the main data collection site for integration.

CONCLUSION

National and international attitudes and policies focused on improving the health of children, workers, and the elderly must be directed towards good nutrition and improving lifestyles. It is made abundantly clean in print and televised media, that obesity has become a severe threat to the health and well being of Americans. That this problem is or will become an international epidemic may depend on the manner in which it is addressed. Exercise is no substitute for poor lifestyle practices, such as excessive alcohol consumption, smoking, overeating, and poor dietary practices. Attention must be directed to the importance of creative movements, posture, perceptual motor stimulation, body awareness, body image, and coordination. However, the importance of physical activity is too valuable to be limited to the young and healthy. Exercise, sports, and other physical activities must include all ages without regard to their frailty or disabilities.

The laws of nature rule the human body. Chemical and biological laws affect food metabolism, neurological transmissions within the nervous system and the target organs, hormonal influences, and all other growth, maintenance, and performance activities. Mechanical influences occur at the joints according to the same laws that return the pole vaulter to earth. Food, water, air, and environmental factors interact with work and societal demands. Human life is an interplay of external and internal processes and energy and, according to the second law of thermodynamics, the system will move toward increased disorder over time (56).

In terms of the universe, the first law of thermodynamics states that the total energy of the universe is constant. The second law states that the total entropy of the universe is increasing. The measure of a system's disorder is referred to as entropy and Eddington said, Whenever you conceive of a new theory of unusually attractiveness, but it does not in some way conform to the second law, then that theory is most certainly wrong (57). Everyone inevitably grows older. Delaying the process of disorder by keeping the subsystems of the organism at a low level of entropy does not flaunt the second law, but rather exploits it.

Science and technology have afforded us the ability to quantify movement so that humans can use their bodies more efficiently. Normal movement of small children can be reflected in improved diapers that do not alter their gait.

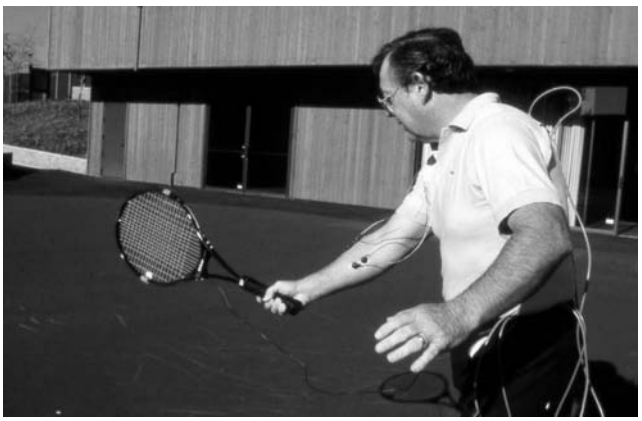


Figure 11. The EMG analysis of a tennis stroke.

Assessment of workplace activities can identify movements that are biomechanically inappropriate for healthy workers. Changing the design of the work bench, providing variable height stools for the conveyor belt operators, and evaluating the job requirements to assist in matching the employee to the work, improved wheelchair design, and adaptations in housing for the elderly are just a few examples of how biomechanical analysis can be applied. Figure 11 shows how athletic performance and equipment are assessed scientifically.

Not only has scientific and technological means provided quantitative assessment abilities, but has also allowed the development of improved means for exercising. Exercise equipment has become so sophisticated that it is appropriate for all ages. The youngest and the oldest can benefit from improved muscular health; the weakest and the strongest can always improve or, at the very least, sustain, healthy muscles; and those with compromised health or bodily functions should enjoy the opportunities to improve their musculature.

Logically, consumption of proper food, sleeping or resting sufficiently, and engaging in an appropriately amount of intense physical activity should keep the tissues and organs functioning maximally. To extend and improve the length and the quality of life depends on an increased awareness of human anatomy, biology, and physiology with continuous research efforts in these and other areas which impact human life. The aging process cannot be overcome, but it should be possible to negate many of the debilitating aspects of it. The Declaration of the United States of America is the only document of any country in history which includes the statement of "pursuit of happiness" and this concept should apply to the health and quality of life for all peoples, regardless of location, and at every age: from infancy to the twilight years.

BIBLIOGRAPHY

Cited References

1. Pollack L, Lowenthal DT, Graves JE, Carroll JF. The elderly and endurance training. In: Shephard RJ, Astrand P-O, editors. Endurance in Sport. London: Blackwell Scientific Publications; 1992. p 390–406.

2. Sidney K, Shephard R, Harrison J. Endurance training and body composition in the elderly. Am J Clin Nutr 1977;30:326–333.
3. Astrand P-O. Influences of biological age and selection. In: Shephard RJ, Astrand P-O, editors. Endurance in Sport. London: Blackwell Scientific Publications; 1992. p 285–289.
4. Bortz WM IV, Bortz WM II. Aging and the disuse syndrome - effect of lifetime exercise. In: Harris S, Harris R, Harris WS, editors. Optimization of human performance. Physical Activity, Aging and Sports, Vol. H.- P, Program and Policy. Albany (NY): Center for the Study of Aging; p 44–50.
5. Kraus H, Raab W. Hypokinetic diseases—diseases produced by the lack of exercise. Philadelphia: Thomas; 1961.
6. Bove AA. Heart and circulatory function in exercise. In: Lowenthal DT, Bharadwaja, Oaks WW, editors. Therapeutics Through Exercise. New York: Grune & Stratton; 1979. p 21–31.
7. Saltin B. et al. Response to exercise after bed rest and after training. Circulation 38(7): 1.
8. Erick H, Knottinggen A, Sarajas SH. Effects of physical training on circulation at and during exercise. Am J Cardiol 1963;12:142.
9. Saltin B, et al. Responses to exercise after bed rest and after training. Circulation 38(8): 1.
10. Clausen JP. Effect of physical training on cardiovascular adjustments to exercise. Ph Rev 1977;37:779.
11. Scheuer J, Tipton CM. Cardiovascular adaptations to physical training. Ann Rei Physiol 1977;39:221.
12. Ritzer TF, Bove AA, Lynch PR. Left ventricular size and performance followir term endurance exercise in dogs. Fed Proc 1977;36:447.
13. Miller PB, Johnson RL, Lamb LE. Effects of moderate exercise during four w(bed rest on circulatory function in man. Aerosp Med 1965;38:1077.
14. Oscai LB, Williams BT, Hertig BA. Effect of exercise on blood volume. J A Physiol 1968;24:622.
15. Hanson JS, Tabakin BS, Levy AM, Nedde W. Long term physical training and cardiovascular dynamics in middle aged men. Circulation 1968;38:783.
16. Becklake B, et al. Age changes in myocardial function and exercise response. Prog Cardiovasc Dis 1965;19:1–21.
17. Templeton G, Platt M, Willerson J, Weisfeldt M. Influence of aging on left ventricular hemodynamics and stiffness in beagles. Circ Res 1979;44:189–194.
18. Gottlieb SO, et al. Silent ischemia on Holter monitoring predicts mortality in high risk infarction patients. JAMA 1988;259:1030–1035.
19. Dustan H. Atherosclerosis complicating chronic hypertension. Circulation 1974;50:871.
20. Gribbin B, Pickering T, Sleight P, Peto R. Effect of age and high blood presst baroflex sensitivity in man. Circ Res 1971;29:424.
21. Bristow J, et al. Diminished baroflex sen; in high blood pressure. Circulation 1969;39:48.
22. Papper S. The effects of age in reducing renal function. Geriatrics 1973;28:83–87.
23. Lane C, et al. Long distance running, bone density, and osteoarthritis. JAMA 1986;255:1147–1151.
24. Shock N. Systems integration. In: Finch C, Hayflick L, editors. Handbook of the Bio1 Aging. New York: Van Nostrand Reinhold; 1977. p 639–665.
25. Cummings S, et al. Epidemiology of osteop and osteoporotic fractures. Epidemiol Rev 1985;7:178–208.
26. Basmajian JV, De Luca CJ. Muscles Alive. Baltimore: Williams & Wilkins; 1985.
27. Milliman and Robertson, Inc. Health risks and behavior: The impact on medical costs. C Data Corporation. 1987.
28. Lane N, et al. Long distance rut bone density and osteoarthritis. JAMA 1986;255:1147–1151.
29. Felig P, Koivisto V. The metabolic response to exercise: Implications for diabetes. In: Lowenthal DT, Bharadwaja K, Oaks WW, editors. Therapeutics Through Exercise. New York: Grune & Stratton; 1979. p 3–20.
30. Pollock M, Wilmore J, editors. Exercise in Health and Disease: Evaluation and Prescription for Prevention and Rehabilitation. Philadelphia: Saunders; 1990.
31. Jokl E. Physical activity and aging. In: Harris S, Harris R, Harris WS, editors. Physical Activity, Aging and Sports. Vol. II, Albany: Center for the Study of Aging; 1992. p 12–20.
32. Paffenbarger R, Hyde R, Wing A, Hsieh C. Physical activity and all-cause mortality and longevity of college alumni. N Engl J Med 1986;314:605–613.
33. Kannel W, et al. Prevention of cardiovascular disease in the elderly. J Am Coll Cardiol 1987;10:25A–8A.
34. Dudley G, Fleck S. Strength and endurance training: Are they mutually exclusive? Sports Med 1987;4:79–85.
35. McDonaugh M, Davies C. Adaptive response of mammalian skeletal muscle to exercise with high loads. Eur J Appl Phys 1984;52:139–155.
36. Delorme TL, Watkins AL. Techniques of progressive resistance exercise. Arch Phys Med 1948;29:645–667.
37. McQueen I. Recent advances in the technique of progressive resistance exercise. Br Med 1954;2:328–338.
38. Hellebrandt F, Houtz S. Mechanism of muscle training in man: Experimental demonstration of overload principle. Physiol Ther Rev 1956;36:371–376.
39. Ariel GB. Computerized biomechanical analysis of human performance. Mechanics and Sport, Vol. 4, New York: The American Society of Mechanical Engineers; 1973. p 267–275.
40. Wainwright RW, Squires RR, Mustich RA. Clinical significance of ground reaction forces in rehabilitation and sports medicine. Presented at the Canadian Society for Biomechanics, 5th Biannial Conference on Biomechanics and Symposium on Human Locomotion; 1988.
41. Llacera I, Squires RR. An analysis of the shoulder musculature during the forehand racquetball serve. Las Vegas: Presented at the American Physical Therapy Association meeting; 1988.
42. Susanka P. Biomechanical analyses of men's handball. Presented at International Handball World Federation 12th Men's Handball World Championship, Prague, Czechoslovakia, Charles University; 1990.
43. Reinsch C. Smoothing by spline functions. Numer Math 1967;10:177–183.
44. Wood GA, Jennings LS. On the use of spline functions for data smoothing. J Biomech 1975;12(6): 477–479.
45. Kaiser JF. Digital Filters. In: Liu D, editor. Digital Filters and the Fast Fourier Transform, 5-79. Stroudsburg (PA): Dowden, Hutchinson & Ross; 1975.
46. Hoy MG, Marcus R. Effects of age and muscle strength on coordination of rising from a chair. In: Woollacott M, Horak F, editors. Posture and Gait: Control Mechanisms. Vol. II. Portland (OR): University of Oregon Books; 1992. p 187–190.
47. Teasdale N, Stelmach GE, Bard C, Fleury M. Posture and elderly persons: Deficit; the central integrative mechanisms. In: Woollacott M, Horak F, editors. Posture and Gait: Control Mechanisms. Vol. II. Portland, (OR): University of Oregon Books; 1992. p 203–207.
48. Vamos L, Riach CL. Postural stability limits and vision in the older adult. In: Woollac M, Horak F, editors. Posture and Gait: Control Mechanisms. Vol. II. Portland (OR): University of Oregon Books; 1992. p 212–215.
49. Frank J, et al. Control of upright stand active, healthy elderly. In: Woollacott M, Horak F, editors. Posture and Gait:

Control Mechanisms. Vol. II. Portland (OR): University of Oregon Books; 1992. p 216–219.
50. Panzer V, Kaye J, Edner A, Holme L. Standing postural control in the elderly and v elderly. In: Woollacott M, Horak F, editors. Posture and Gait: Control Mechanisms. Vol. II. Portland (OR): University of Oregon Books; 1992. p 220–223.
51. Craik R. Changes in locomotion in the aging adult. In: Woollacott MH, Shumway-Co A, editors. Development of Posture and Gait across the Life Span. Columbia (SC): University of South Carolina Press; 1989. p 150–153.
52. Williams K. Intralimb coordination of older adults during locomotion: Stair climbing. In: Woollacott M, Horak F, editors. Posture and Gait: Control Mechanisms. Vol. II. Portland (OR): University of Oregon Books; 1992. p 208–211.
53. Light KE, Tang PF, Krugh CR. Performance differences between young and elderly females in a step-reach task. In: Woollacott M, Horak F, editors. Posture and Gait: Control Mechanisms. Vol. II. Portland (OR): University of Oregon Books; 1992. p 287–290.
54. Jacobs I, Bell DG, Pope J. Comparison of isokinetic and isoinertial lifting tests as predictors of maximal lifting capacity. Eur J Appl Physiol 1988;57:146–153.
55. Ariel GB. Computerized dynamic resistive exercise. In: Landry F, Orban WAR, editors. Mechanics of Sports and Kinanthropometry. Book 6. Miami (FL): Symposia Specialists, Inc.; 1978. p 45–51.
56. Benson H. University Physics. New York: John Wiley & Sons; 1991.
57. Eddington A. The Nature of the Physical World. Cambridge: New Press; 1928.

See also EXERCISE STRESS TESTING; HUMAN SPINE, BIOMECHANICS OF; JOINTS, BIOMECHANICS OF; LOCOMOTION MEASUREMENT, HUMAN; REHABILITATION AND MUSCLE TESTING.

BIOMECHANICS OF JOINTS. See JOINTS, BIOMECHANICS OF.

BIOMECHANICS OF SCOLIOSIS. See SCOLIOSIS, BIOMECHANICS OF.

BIOMECHANICS OF SKIN. See SKIN, BIOMECHANICS OF.

BIOMECHANICS OF THE HUMAN SPINE. See HUMAN SPINE, BIOMECHANICS OF.

BIOMECHANICS OF TOOTH AND JAW. See TOOTH AND JAW, BIOMECHANICS OF.

BIOMEDICAL ENGINEERING EDUCATION

PAUL BENKESER
Georgia Institute of Technology
Atlanta, Georgia

INTRODUCTION

Biomedical engineering is that interdisciplinary field of study combining engineering with life sciences and medicine. It is a relatively new field of study that has only recently experienced sufficient maturity to enable it to establish its own identity. Often, this field will be described using the term bioengineering. In 1997, the Bioengineering Definition Committee of the National Institutes of Health released the following definition of the field (1): "Bioengineering integrates physical, chemical, mathematical, and computational sciences and engineering principles to study biology, medicine, behavior, and health. It advances fundamental concepts; creates knowledge from the molecular to the organ systems level; and develops innovative biologics, materials, processes, implants, devices and informatics approaches for the prevention, diagnosis, and treatment of disease, for patient rehabilitation, and for improving health."

While many use biomedical engineering and bioengineering interchangeably, it is generally accepted today that bioengineering is a broader field that combines engineering with life sciences, but is not necessarily restricted to just medical applications.

The Biomedical Engineering Society further elaborated on the definition of biomedical engineering as part of a guide on careers in the field. In it is stated (2): "A Biomedical Engineer uses traditional engineering expertise to analyze and solve problems in biology and medicine, providing an overall enhancement of health care. Students choose the biomedical engineering field to be of service to people, to partake of the excitement of working with living systems, and to apply advanced technology to the complex problems of medical care. The biomedical engineer works with other health care professionals including physicians, nurses, therapists and technicians. Biomedical engineers may be called upon in a wide range of capacities: to design instruments, devices, and software, to bring together knowledge from many technical sources to develop new procedures, or to conduct research needed to solve clinical problems."

Educational programs in the field of biomedical engineering had their origins in a handful of specialized graduate training programs in the 1950s focusing primarily on diagnostic and therapeutic devices and instrumentation. By 2004, there were undergraduate and graduate programs in biomedical engineering at ~100 universities in the United States. The diversity in the content of undergraduate educational programs that was commonplace in its early years is gradually diminishing as the field has matured. While the current undergraduate programs still maintain their own unique identity, there has been a steady movement toward the definition of a core curriculum in the field.

The purpose of this article is to give the reader some historical perspective on the origins of educational programs in the field, the challenges associated with preparing bachelor-level graduates for careers in the field, and the current state-of-the-art in undergraduate biomedical engineering curriculums.

HISTORY

The first steps toward establishing biomedical engineering as a discipline occurred in the 1950s as several formalized training programs were created. Their establishment was significantly aided by the National Institutes of Health

creation of training grants for doctoral studies in biomedical engineering. The Johns Hopkins University, the University of Pennsylvania, the University of Rochester, and Drexel University were among the first to be awarded these grants.

During the late 1960s and early 1970s, growing opportunities in the field helped prompt the development of a second generation of biomedical engineering programs and departments. These included Boston University in 1966; Case Western Reserve University in 1968; Northwestern University in 1969; Carnegie Mellon University, Duke University, Rensselaer Polytechnic Institute and a joint program between Harvard and the Massachusetts Institute of Technology in 1970; Ohio State University and University of Texas, Austin, in 1971; Louisiana Tech, Texas A&M and the Milwaukee School of Engineering in 1972; and the University of Illinois, Chicago in 1973 (3). Many of these first and second generation of programs were concentrating the training of their students in areas defined either using quasiclassical engineering terminology, such as bioinstrumentation, biomaterials and biomechanics, or by application area, such as rehabilitation engineering or clinical engineering.

The late 1990s witnessed a substantial increase in the growth of the number of departments and programs in biomedical engineering, especially at the undergraduate level. The growth of this third generation of programs was fueled in part by grants from The Whitaker Foundation to help institutions establish or develop biomedical engineering departments or programs. In 2004, ~100 universities have programs or departments in biomedical engineering, including 33 offering undergraduate degree programs accredited by the Engineering Accreditation Commission of the Accreditation Board for Engineering and Technology (ABET) (4). The growth in the numbers of ABET accredited degree programs is illustrated in Fig. 1.

The arrival of the third generations of programs coincided with the development of several new areas of training in biomedical engineering, such as systems biology–physiology, and tissue, cellular, and biomolecular engineering. These areas typically require significantly more training in life sciences than was present in the first and second generation biomedical engineering training programs. This presented significant challenges for undergraduate programs trying to add this life science content to their curricula without increasing the number of credit hours required for the programs. Many of these programs accomplished this by creating a new generation of courses in which the engineering and life science concepts are integrated together within courses. The integration of such courses into the curriculum is discussed in more detail in the Curriculum section.

CAREER PREPARATION

The design of a high quality educational program should always start with its educational objectives. By using the definition established by ABET, these program educational objectives are statements that describe the expected accomplishments of graduates during the first several years following graduation (5). This requires programs to be cognizant of the needs of prospective employers of its graduates and design learning environments and curricula to meet those needs. This is particularly challenging task for a relatively new and evolving field like biomedical engineering.

Biomedical engineers are employed in industry, in research facilities of educational and medical institutions, in teaching, in government regulatory agencies, and in hospitals. They often serve as integrators or facilitators, using their skills in both the engineering and life science fields. They may work in teams in industry to help design devices, systems, and processes that require an in-depth understanding of both living systems and engineering. Frequently, biomedical engineers will be found in technical sales and marketing positions in companies seeking to provide their customers with technically trained individuals who are capable of better understanding their needs and communicating those needs back to product development teams. Government regulatory positions, such as those with the Food and Drug Administration, often involve testing medical devices for performance and safety. In research institutions, biomedical engineers participate in or direct research activities in collaboration with other

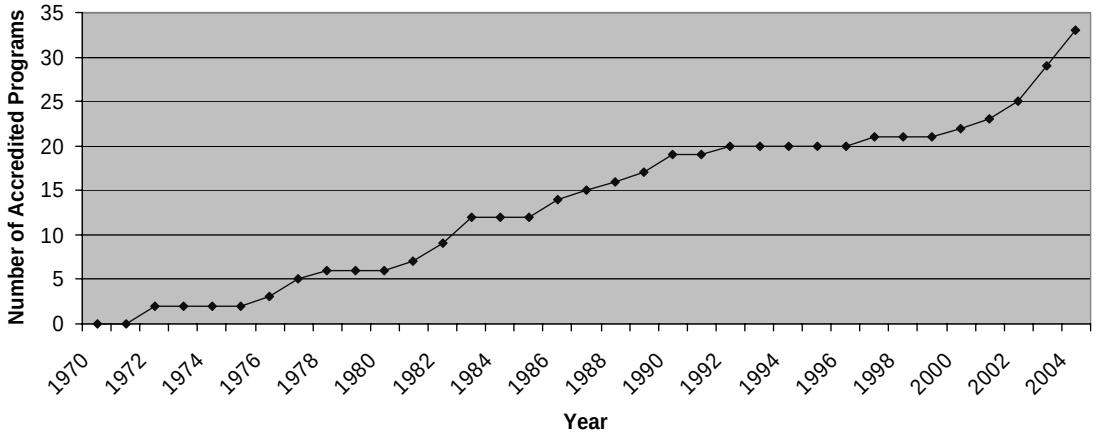


Figure 1. Number of ABET accredited programs in biomedical enginering.

researchers with such backgrounds as medicine, biology, chemistry, and a variety of engineering disciplines.

According to the U.S. Department of Labor's Bureau of Labor Statistics, manufacturing industries employed 38% of all biomedical engineers, primarily in the pharmaceutical and medicine manufacturing and medical instruments and supplies industries (6). Others worked in academia, hospitals, government agencies, or as independent consultants. Employment of biomedical engineers is expected to increase faster than the average for all occupations through 2012 (6). The demand for better and more cost-effective medical devices and equipment designed by biomedical engineers is expected to increase with the aging of the population and the associated increased focus on health issues. Most of these employment opportunities will be filled with graduates from B.S. and M.S. degree programs. However, for research-oriented jobs, like faculty positions in academia and research and development positions in industry, employers typically require their employees to possess a Ph.D. degree.

The needs of the employers that hire biomedical engineers undoubtedly vary by industry and job title. However, there are some skills that appear to be in universal demand by all employers of biomedical engineers. They include proficient oral and written communication skills, the ability to speak the languages of engineering and medicine, a familiarity with physiology and pathophysiology, and teamwork skills (6).

In spite of this seemingly impressive list of career paths and options, one of the most significant challenges facing entry-level biomedical engineers are prospective employers who complain that they do not understand what skill sets biomedical engineers possess (7). The perception is that the skills possessed by engineers from other disciplines, like electrical or mechanical engineering, are more predictable and in large part are independent of the university where the engineer was educated. It is likely that this perception is a result of a combination of two factors. First, often the individuals responsible for making the hiring decisions in companies did not receive their degrees in biomedical engineering, and thus do not have first-hand experience with the training received by biomedical engineers. Second, until relatively recently, many undergraduate biomedical engineering programs lacked a substantive core curriculum and were structured in such a way that students had to select from one of several "tracks" offered by the program. These tracks were typically pattered along traditional engineering lines, such as bioelectronics and biomechanics, in an attempt to address another concern expressed by prospective employers—that graduates of bachelor degree programs in biomedical engineering were too broadly trained and thus lacked sufficient depth of engineering skills. Due to this perceived lack of depth, it is not uncommon to find employers for which the entry-level degree for biomedical engineering positions is the masters degree. Undoubtedly the presence of these tracks, and their variability from program to program, contributed to the confusion in industry over the what skill sets they should expect from a biomedical engineer.

As a result of these concerns over depth, breadth, and uniformity of curriculum, the biomedical engineering education community has recognized the need to reach consensus on what constitutes a core undergraduate curriculum in biomedical engineering. This has become one of the major initiatives of the National Science Foundation (NSF) sponsored VaNTH Engineering Research Center (ERC) for Biomedical Engineering Educational Technologies. This ERC, a collaboration of teams from Vanderbilt University, Northwestern University, The University of Texas at Austin, Harvard University, and the Massachusetts Institute of Technology, was created in 1999. Its vision is to transform bioengineering education to produce adaptive experts by developing, implementing and assessing educational processes, materials, and technologies that are readily accessible and widely disseminated (8). The Whitaker Foundation has also sponsored workshops at its 2000 and 2005 Biomedical Engineering Summit meetings with the goal of delineating the core topics in biomedical engineering that all biomedical engineering students should understand. White papers from these meetings can be found at the Foundation's web site (9).

In spite of the movement toward the creation of a common core curriculum in undergraduate programs of study in biomedical engineering, there will undoubtedly continue to be some differences in curricula between programs. This is not only permitted in the current accreditation review process of ABET, but in some sense encouraged. Within the past decade this process has changed from a prescriptive evaluation to an outcomes-based assessment centered on program-defined missions and objectives (10). Thus, it will be incumbent on programs to work closely with the prospective employers of their graduates to ensure that the programs provide the graduates with the skills the employers desire. For example, Marquette University's biomedical engineering program has an established industrial partners program with >30 companies participating (11).

THE UNDERGRADUATE CURRICULUM

Contained within this section is a description of a core undergraduate curriculum that the author believes the biomedical engineering educational community is converging upon. The contents are based upon reviews of curriculums from biomedical engineering programs (12) and information disseminated by the Curriculum Project of the VaNTH ERC (13). It is important to note that the core described herein is not being presented as the prescription for what a biomedical engineering curriculum should look like, but rather a reflection of the current trends in the field.

Course Requirements

To be accredited by ABET, the curriculum must include the following:

- One year of an appropriate combination of mathematics and basic sciences.
- One-half year of humanities and social sciences.
- One and one-half years of engineering topics and the requirements listed in ABETs Program Criteria for bioengineering.

A year is defined as 32 semester or 48 quarter hours.

The typical math and science content of biomedical engineering curriculums, as described in Table 1, are similar to those of other engineering disciplines. The notable exceptions are courses in biology and organic chemistry. In general, most programs rely on other departments within their university to provide the instruction for these courses. Some programs with large enrollments have been successful in collaborating with faculty in their university's science departments to create biology and chemistry courses for biomedical engineering students. For example, the School of Chemistry and Biochemistry at the Georgia Institute of Technology has created organic and biochemistry courses specifically designed for biomedical engineering students.

Table 1. Mathematics and Science Core Curriculum Subjects

Subject	Sampling of Topical Coverage
Math	Linear algebra
	Differential, integral, multivariable calculus
	Differential equations
	Statistics
Physics	Classical mechanics, oscillations and waves
	Electromagnetism, light and modern physics
Computer Science	Algorithms, data structures, program design and flow control
	Graphics and data visualization
	Higher level programming language (e.g., Matlab, Java, C++)
Chemistry	General and inorganic chemistry
	Organic chemistry
	Biochemistry
Biology	Modern biological principles
	Genetics
	Cell biology

The core biomedical engineering content is described in Table 2. Depending on the size of the program, some of the content may be delivered in courses outside of biomedical engineering (e.g., thermodynamics from mechanical engineering). It is not uncommon to find some variability between programs in the content of their core curriculums. This will likely always be the case as each program must provide the curriculum that best enables its graduates to achieve the program's unique educational objectives.

ABET Criterion 3 stipulates that engineering programs must demonstrate achievement of a minimum set of program outcomes. These outcomes are statements that describe skills that "students are expected to know or be able to do by the time of graduation from the program" (5). A closer examination of these skills suggests that they can be divided into two sets as illustrated in Table 3. The first set, "domain" skills, is one that engineering educators are typically adept in both teaching and quantitatively measuring achievement. Programs generally use courses, like those listed in Tables 1 and 2, to develop these domain

Table 2. Biomedical Engineering Core Curriculum Subjects

Subject	Sampling of Topical Coverage
Biomechanics	Principles of statics
	Mechanics of biomaterials
	Dynamics
Biotransport	Mass transfer
	Heat transfer
	Momentum transfer
Biothermodynamics	Thermodynamic principles
	Mass and energy balances
Biomaterials	Metals, polymers and composite materials
	Biocompatibility
Bioinstrumentation	Instrumentation concepts
	Amplifiers and filters
	Sensors and transducers
Biofluids	Blood vessel mechanics
	Hydrostatics and steady flow models
	Unsteady Flow and non-uniform geometric models
Systems Physiology	Cellular metabolism
	Membrane dynamics
	Homeostatis
	Endocrine, cardiovascular and nervous systems
	Muscles
Biosignal Analysis	Digital signal processing theory
	Filtering
	Frequency-domain characterization of signals

skills in their students. The second set, "professional" skills, is more difficult to teach and assess. However, these professional skills are often the ones most frequently cited by employers of engineers as the most important skills they value in their employees.

Humanities and social science courses are integral to the achievement of these professional skills. However, programs must avoid employing the "inoculation" model for teaching these skills to the students. In this model, it is assumed that students can learn these skills by simply taking isolated courses in ethics, technical communications, and so on. There are several problems with this model. It can decontextualize these skills, treating them as add-ons and not an integral part of everyday engineering practice. This is a false and even dangerous message to give the students—that written and oral communication and ethical behavior are peripheral to the real world of engineering. This message is further driven home because the faculty responsible for teaching these skills is humanities or social science faculty not engineering faculty. In addition, the complexity of these skills to be learned is too great for students to master within the framework of isolated courses. Research suggests, however, that students need quasirepetitive activity cycles and practice in multiple settings to develop proficiency in these professional skills (14–16).

Professional Skills

Before describing methods of developing these professional skills in students, it is necessary to establish operational

Table 3. Program Outcomes Specified in ABET Criterion 3

Domain Skills	Professional Skills
An ability to apply knowledge of mathematics, science, and engineering	An ability to function on multi-disciplinary teams
An ability to design and conduct experiments, as well as to analyze and interpret data	An understanding of professional and ethical responsibility
An ability to design a system, component, or process to meet desired needs	An ability to communicate effectively
An ability to identify, formulate, and solve engineering problems	The broad education necessary to understand the impact of engineering solutions in a global and societal context
A knowledge of contemporary issues	A recognition of the need for, and an ability to engage in lifelong learning
An ability to use the techniques, skills, and modern engineering tools necessary for engineering practice	

descriptions for these constructs. Such descriptions serve two functions. They reveal the complexity of the particular skills in terms of the subskills required to demonstrate the higher level skills specified in the ABET lists. Descriptions can also serve as articulations of learning outcomes, which can be designed toward and assessed. The following represents one interpretation of the variables that are indicators of these constructs (16).

Ability to communicate effectively.

Oral + written communication skills

- Convey information and ideas accurately and efficiently.
- Articulate relationships among ideas.
- Inform and persuade.
- Assemble and Organize evidence in support of an argument.
- Make communicative purpose clear.
- Provide sufficient background to anchor ideas–information.
- Be aware of and address multiple interlocutors.
- Clarify conclusions to be drawn from information.

Ability to function on multidisciplinary teams.

Team − collaboration skills + communication skills 3

- Help group develop and achieve team goals.
- Avoid contributing excessive or irrelevant information.
- Confront others directly when necessary.
- Demonstrate enthusiasm and involvement.
- Monitor group progress and complete tasks on time.
- Facilitate interaction with other members.

Understanding professional and ethical responsibilities.

- Recognize moral problems and issues in engineering.
- Comprehend, clarify, and critically assess opposing arguments.
- Form consistent and comprehensive viewpoints based on facts.
- Develop imaginative responses to problematic conflicts.
- Think clearly in the midst of uncertainty and ambiguity.
- Appreciate the role of rationale dialogue in resolving moral conflicts.
- Ability to maintain moral integrity in face of pressures to separate professional and personal convictions.

Broad education necessary to understand the impact of engineering solutions in a global and societal context.

- Identify human needs or goals technology will serve.
- Analyze and evaluate the impact of new technologies on economy, environment, physical and mental health of manufacturers, uses of power, equality, democracy, access to information and participation, civil liberties, privacy, crime and justice.
- Identify unintended consequences of technology development.
- Create safeguards to minimize problems.
- Apply lessons from earlier technologies and experiences of other countries.

Recognition of the need for, and an ability to engage in life-long learning.

- Identify learning needs and set specific learning objectives.
- Make a plan to address these objectives.
- Evaluate inquiry.
- Assess the reliability of sources.
- Evaluate how the sources contribute to knowledge.
- Question the adequacy and appropriateness of forms of evidence used to report back on learning needs.
- Apply knowledge discovered to the problem.

Table 4. Repetitive Activities in the Problem-Solving Cycle

	Professional Skill				
Activity	Communicate	Teams	Responsibilities	Impact	Learning
Identifying learning–knowledge needs as a team–individual					X
Acquiring knowledge needed to solve problem			X	X	X
Reporting back to team	X	X			X
Digging deeper and solving	X	X	X	X	X
Presenting solution to audience of experts	X				
Writing a report on problem solution	X				

There are a variety of methods programs can employ to foster the development of these professional skills in students. These include the use of team-based capstone design experiences, facilitating student participation in coop and internship experiences, encouraging involvement in undergraduate research projects, and incorporating oral and written communication exercises throughout the curriculum.

The creation of new undergraduate biomedical engineering programs has led to the development of some new approaches to professional skills development. For example, the Wallace H. Coulter Department of Biomedical Engineering at Georgia Tech and Emory University has implemented an integrative approach to the development of professional skills and adaptive expertise in its program. This approach anchors professional skills development in the context of team-based problem solving and design experiences over the four-year curriculum. This approach provides multiple opportunities for the students to work on and develop skills and knowledge in a variety of "real-world" engineering settings in which these professional skills are practiced. Within each experience, activities are identified within the problem-solving cycle that help students develop these professional skills. These activities are described in Table 4.

The need for real-world problems cannot be overstated. Authentic, open-ended problems are needed as contexts and catalysts for the development of these professional skills. Not only do they help prepare students for the professional practice of engineering, but they are also a significant motivator for the students to delve more deeply into the problem space. Moreover, the use of these skills in the context of a large problem makes them central not peripheral to biomedical engineering problem solving. They begin to understand the value of clear, thoughtful communication, and collaboration when confronting complex problems. They see how ethical issues can arise when seeking design solutions. If the problems are authentic, then the information needed to solve them must be found in multiple places, which helps students to develop inquiry and research skills for lifelong learning.

SUMMARY

The field of biomedical engineering had its foundations laid roughly 50 years ago. Undergraduate degree programs in the field followed shortly thereafter. Fueled in part by generous support from the Whitaker Foundation, there has been a significant increase in the number of new undergraduate degree programs in the field. This has led to a significant increase in student interest in the field. This growth has increased the need for the biomedical engineering education community to work with industry to better define the skills that graduates need to obtain to lead productive careers in the field. There exists a movement, led by the NSF VaNTH ERC, to define a core undergraduate curriculum within the constraints imposed by ABET accreditation criteria. The VaNTHs vision to transform bioengineering education to produce adaptive experts has been adopted by new undergraduate degree programs and has produced demonstrated examples of pedagogical advances in the field of engineering education.

BIBLIOGRAPHY

Cited References

1. NIH working definition of bioengineering. 1997, July 24. National Institutes of Health Bioengineering Consortium. Available at http://www.becon.nih.gov/bioengineering_definition.htm. Accessed 2004 Nov. 18.
2. Planning a career in biomedical engineering. 1999. Biomedical Engineering Society. Available at http://www.bmes.org/careers.asp. Accessed 2004 Nov. 18.
3. A history of biomedical engineering. 2002, May. The Whitaker Foundation. Available at http://www.whitaker.org/glance/definition.html. Accessed 2004 Nov. 19.
4. Accredited engineering programs. 2004. Accreditation Board for Engineering and Technology. Available at http://www.abet.org/ accredited_programs/engineering/EACWebsite.asp. Accessed 2004 Nov. 19.
5. Criteria for accrediting engineering programs. 2004. Accreditation Board for Engineering and Technology. Available at http://www.abet.org/criteria.html. Accessed 2004 Nov. 19.
6. Bureau of Labor Statistics, U.S. Department of Labor, Occupational Outlook Handbook. 2004–2005 edition, Biomedical Engineers. Available at http://www.bls.gov/oco/ocos262.htm. Accessed 2005 Feb. 10.
7. RA Linsenmeier, What makes a biomedical engineer, IEEE Eng Med Biol Mag 2003;22(4):32–38.
8. Cordray DS, Pion GM, Harris A, Norris P. The value of the VaNTH Engineering Research Center. IEEE Eng Med Biol Mag 2003;22(4):47–54.
9. Biomedical Engineering Educational Summit, The Whitaker Foundation. Available at http://summit.whitaker.org/. Accessed 2005 Feb 10.
10. Enderle J, Gassert J, Blanchard S, King P, Beasley D, Hale P, Aldridge D. The ABCs of preparing for ABET. IEEE Eng Med Biol Mag 2003;22(4):122–132.

11. Waples LM, Ropella KM. University partnerships in biomedical engineering. IEEE Eng Med Biol Mag 2003;22(4): 118–121.
12. The biomedical engineering curriculum database 2004. The Whitaker Foundation. Available at http://www.whitaker.org/academic/database/index.html. Accessed 2004 Nov. 18.
13. VaNTH ERC curriculum project (2004, June 10). VaNTH ERC [Online]. Available at http://www.vanth.org/curriculum/. Accessed [2004 Nov. 18].
14. Bransford JD, Brown AL, Cocking RR, editors. How People Learn: Brain, Mind, Experience, and School. Washington: National Academy Press; 1999.
15. Harris TR, Bransford JD, Brophy SP. Roles for learning sciences and learning technologies in biomedical engineering education: A review of recent advances. Annu Rev Biomed Eng 2002;4:29–48.
16. Benkeser PJ, Newstetter WC. Integrating soft skills in a BME curriculum, Proc. 2004 ASEE Annu Conf. 2004, June. American Society for Engineering Education. Available at http://www.asee.org/about/events/conferences/search.cfm. Accessed 2004 Nov. 19.

See also BIOINFORMATICS; MEDICAL EDUCATION, COMPUTERS IN; MEDICAL ENGINEERING SOCIETIES AND ORGANIZATIONS.

BIOSURFACE ENGINEERING

PETER MOLNAR
MELISSA HIRSCH-KUCHMA
JOHN W. RUMSEY
KERRY WILSON
JAMES J. HICKMAN
University of Central Florida
Orlando, Florida

INTRODUCTION

One primary reason there is a tremendous amount of interest in cellular patterning techniques is that numerous examples in nature use these techniques to segregate cells into tissues, vessels, and organs. The idea of templates in nature abounds for the creation of organized biological systems using both inorganic (1), as well as organic template systems (2). There is also a certain allure to being able to integrate electrically active cells directly to electronic devices using standard electronic fabrication techniques. Researchers have attempted to use surface cues to pattern cells since as early as 1917, in which spider webs were used to pattern cells (3). Most of the early work on cellular patterning used topographical cues until 1988, when a landmark publication by Kleinfeld et al. (4) used lithographic templating to fabricate simple patterns of cortical neurons. This was an adaptation of standard technology developed by the electronics industry to create computer chips that was then applied to the creation of patterns to guide neuronal cell attachment. It was at this point that interest in this field mushroomed, as the idea of creating neuronal networks from living neurons has potential applications in understanding biological information processing, creating hybrid computer systems, as well as a whole host of biomedical applications. Some prominent efforts to use cell patterning have been for spinal cord repair, creation of *in vitro* test bed systems to study diseases, as well as blood vessel formation from patterned endothelial cells (5). The initial lithography-based technique used by Kleinfeld et al. has been extended to include many other methods for the patterning of cells, including self-assembled monolayer (SAM) patterning, laser ablation, microcontact printing or stamping, ink jet printing, AFM printing, as well as patterning using microfluidic networks. Methods have also been extended in the area of topographical cues for 3D patterning, which has evolved from early work that used scratched grooves in glass surfaces. At this point, depending on the facilities that one has available, some form of mask making and pattern templating is available to just about any laboratory in the world. However, the biological interactions with these patterns that have been created are still not well developed, and this limits applications at this point.

There are many reasons for the lack of long-term applications of this technique, even after the large amount of work that has been done in this area. The first is that, in many instances, the patterns direct the initial attachment of cells, but as the extracellular matrix is deposited by the cells, long-term adherence to the patterns is not maintained. Another issue is that longer term cell survival is also dependent on factors besides the surface, such as media composition, cell–cell contact interactions, and the lack of growth factors that are normally present from other support cells and tissues. Defined systems are being developed in an attempt to control these other variables in addition to the surface (6,7), but these efforts have been limited to date. However, as progress is made in these areas, it will open up possible applications in tissue engineering, tissue repair, biosensors, and functional *in vitro* test beds.

PATTERNING METHODS

Many methods have been developed for creating templates to be used for cell patterning. These can be divided roughly into two categories: those that are derived from photolithography techniques and those that depend on physical segregation, although there is some crossover in the methods between the two. The photolithography-based systems typically use some sort of organic layer, from polymers to monolayers, which is illuminated, either directly with a pattern or through the template pattern to be created. This can involve many or a few steps depending on the particular method used. Generally, a second layer is deposited in the area where material was removed. Specific variations of this technique are discussed in this section.

The second major category, physical segregation, involves the actual placement of the molecules or cells in a pattern on a surface. Stamping is the most well-known method of creating a molecular-based template and of all the techniques is probably the most economical, but other techniques have been investigated for physical placement of cells in a desired pattern on the surface. Finally, both of these methods, which are 2D in nature, are now being extended into 3D patterning using many of the same

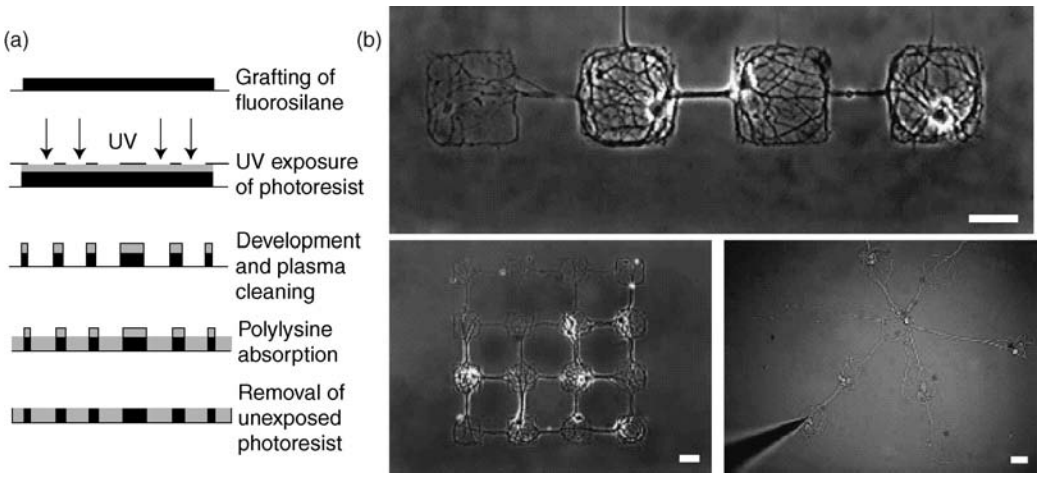


Figure 1. Protocol and images of the pattern. (a) Protocol of photolithography. A glass coverslip (continuous line) is coated with domains of fluorosilane molecule (dark) and with regions of polylysine (light gray) according to the pattern designed on the mask (dashed line), using ultraviolet (UV) exposure of a spin-coated photoresist (gray). (b) Images of neural networks of controlled architecture. Top: linear network. Bottom left: matrix 4×4. Bottom right: star. Cell bodies of neurons are restricted to squares or disks of 80 μm and neurites to line (80 μm length, 2–4 μm wide). Square and disk diameters are 80 μm for each figure. Scale bar is 50 μm (Ref. 9).

techniques, or combinations thereof, described therein. Below there is a brief description, along with the appropriate references, of the techniques that can be used to create these cellular templates.

Photolithography

A variety of photolithographic techniques has been employed to pattern proteins and cells on surfaces from the micrometer to the nanometer scale (8). The basic tools needed are a radiation source and a photomask. The photomask can be created using standard photolithography processes developed for the electronics industry. Irradiation of the surface through the photomask is used to create the patterns by ablation or by using a photosensitive material such as a photoresist as shown in Fig. 1.

Kleinfeld et al. (4) first demonstrated that dissociated neurons could be grown on 2D substrates consisting of lithographically defined patterned monolayers of diamines and triamines with alkylsilanes. The method used by Kleinfeld et al. started with a clean silicon or quartz surface that was spin coated with a layer of photoresist (an organic photosensitive polymer used in the electronics industry). The resist was exposed to UV light with a patterned photomask and then developed. The surface was refluxed in the presence of an alkylsilane, and the photoresist was then stripped off so that areas the photomask covered were reduced to bare silicon or quartz. These areas were then reacted with an aminosilane to form the patterned surface. The patterned cells developed electrical excitability and immunoreactivity for neuron-specific proteins. A further modification of this technique eliminated the photoresist from the pattern formation by direct ablation of the SAM layer (10).

Patterns of self-assembled monolayers formed from organosilanes on glass or silicon substrates and on gold surfaces can be made by using a photoresist mask and deep UV radiation (10–18). Monolayers can also be directly ablated with various forms of radiation such as UV, X ray, ions, and electrons (8) depending on the resolution needed for the patterns. Organosilanes self-assemble and condense onto substrates that have surface –OH functionalities (19). The –SH functionality of the alkanethiol (20) is also highly reactive to ozone and other irradiation sources and has been used in patterning (21). Methods using X-ray or extreme UV (EUV) radiation give better resolution than the traditional photolithography using deep UV and photoresist masks. The ablated regions of the SAM can then be reacted with an organosilane or alkanethiol with different characteristics from the original layer to enable cell growth (22). Azides (23) and aromatic hydrocarbon silanes (24) have also been shown to be reactive for creating patterns.

A typical method to prepare a patterned glass silanated surface is illustrated next. The glass must first be acid cleaned or oxygen plasma cleaned to maximize the surface –OH functional density. Next, the glass is reacted with a silane that contains -chloro, -methoxy, or -ethoxy bonds in the presence of a small amount of water that acts as a catalyst. A mask is then used to protect certain areas of the surface while allowing the radiation source to ablate others in the desired pattern. The ablated regions of the surface can then be coated with a different silane with different properties than the original, thus forming the patterned surface; an example of this is shown in Fig. 2.

The photoreactivity of polymers, such as poly(ethylene) glycol and polystyrene, has also been used to pattern surfaces (25). Biologically based polymers, such as poly-L-lysine and extracellular matrix proteins, have been ablated tso create patterns (26). Polymer photolithography followed by protein adsorption has been combined to create patterned cytophobic and cytophilic areas. Patterning of perfluoropolymers followed by adsorption of poly-L-lysine

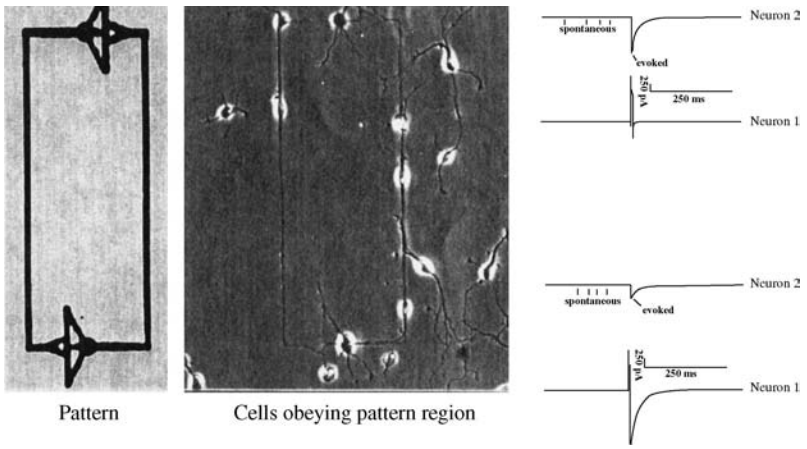


Figure 2. Micrograph of circuit-patterned day 2 *in vitro* hippocampal neurons plated onto DETA/15F modified glass coverslips. Electrophysiology of day 12 *in vitro* hippocampal neurons displaying both spontaneous and evoked activity on a DETA/15F line-space patterned surface. Top trace: post-synaptic neuron. Bottom trace: stimulated presynaptic neuron.

and albumin has been one method used (27). Photosensitive polymers have been treated with UV radiation to create an –COOH functional unit on the surface and then patterned via the linkage of proteins to form cytophilic and cytophobic regions (28). A bioactive photoresist (bioresist) has been developed that does not require the use of solvents that can denature the biomolecule that is patterned (29).

Photolithographic protein patterning requires the attachment of photosensitive groups to proteins on a substrate. Patterns can be made using a patterned mask and selective ablation. This method has been shown to be useful to produce micropatterned cultures (30). Various methods are used to covalently attach proteins in patterns to surfaces (31), to pattern using biomolecule photoimmobilization (32), and to create density gradients of photoreactive biomolecules (33). Heterobifunctional crosslinker molecules have been used to attach proteins to silanated surfaces both before and after the photolithographic patterning step (34). Protein patterning has been achieved using a micromirror array (MMA), which can transfer a pattern from the mirrors that are switched on, ablating a photolabile protecting group (35). Photolithography was also used to pattern thermosensitive copolymers through polymer grafting (36). The surface micropattern appeared and disappeared interchangeably, as observed under a phase-contrast microscope, by varying the temperature between 10 and 37 °C. The copolymer-grafted polystyrene surface was hydrophobic at 37 °C and hydrophilic at 10 °C.

Photolithography provides high resolution patterning and the ability to make complex patterns on surfaces. Unlike stamping techniques, the patterns are more permanent; however, the process can be relatively expensive as it generally requires the use of a laser and clean room facilities for the mask production. To attach proteins, such as ECM proteins, the use of a covalently attached crosslinker is necessary, and stamping techniques are generally preferred for this application.

Microcontact Printing (Stamping)

Microcontact printing was introduced by George Whiteside's group at Harvard in 1994 (37) to pattern self-assembled monolayers on gold substrates to control surface properties, cell adhesion, proliferation, and protein secretion by patterned cells. The basic method to create surface patterns by microcontact printing has not changed much since then. Usually, a poly(dimethylsiloxane) (PDMS) stamp is created using a molding technique from a master pattern relief mold and then used to transfer chemical patterns to flat or curved surfaces. The master is usually prepared from silicon by standard photolithography and/or etching, but other substrates can also be used. The transferred chemical patterns can be created using a compound that binds covalently to the substrate (e.g., self-assembled monolayers or proteins immobilized by crosslinkers) (38) or a compound that binds noncovalently, such as absorbed extracellular matrix proteins (39). This methodology is illustrated in Fig. 3. Oliva et al. presented a novel method to couple proteins to patterned surfaces based on the strong interaction of protein A and the Fc fragment of immunoglobulins. This method involved the creation of a covalently coupled Fc fragment and the target protein (41). Methods have also been developed to transfer proteins from a fluid phase to a surface using hydrogels as the stamp (42). Moreover, recently introduced techniques are allowing the creation of protein gradients with microcontact printing (43). Although alignment of the stamp/patterns with surface features such as microelectrodes is more difficult than in the case of photolithographic patterning, several groups are beginning to address this issue (44). Microcontact printing is usually a favored method among biologists compared with photolithography, because (*1*) the equipment and controlled environment facilities required for photolithography are not routinely available to cell biologists and (*2*) the steps are simpler to pattern proteins, the molecules of greatest interest to biologists, using microcontact printing than with photolithography and crosslinkers. The refinement of the PDMS molding technique has directly led to the development of another important patterning method, microfluidics, which gained wider applications with the introduction of microelectromechanical systems (MEMS) and "lab-on-a-chip" systems. However, initial results using PDMS indicated some transfer of the PDMS to the surface during the stamping process. This can be troublesome in cell patterning applications as PDMS can be toxic to cells or mask the chemical functionality of interest. Methods of "curing" the stamps or presoaking to enable better release of the compound has been reported (45).

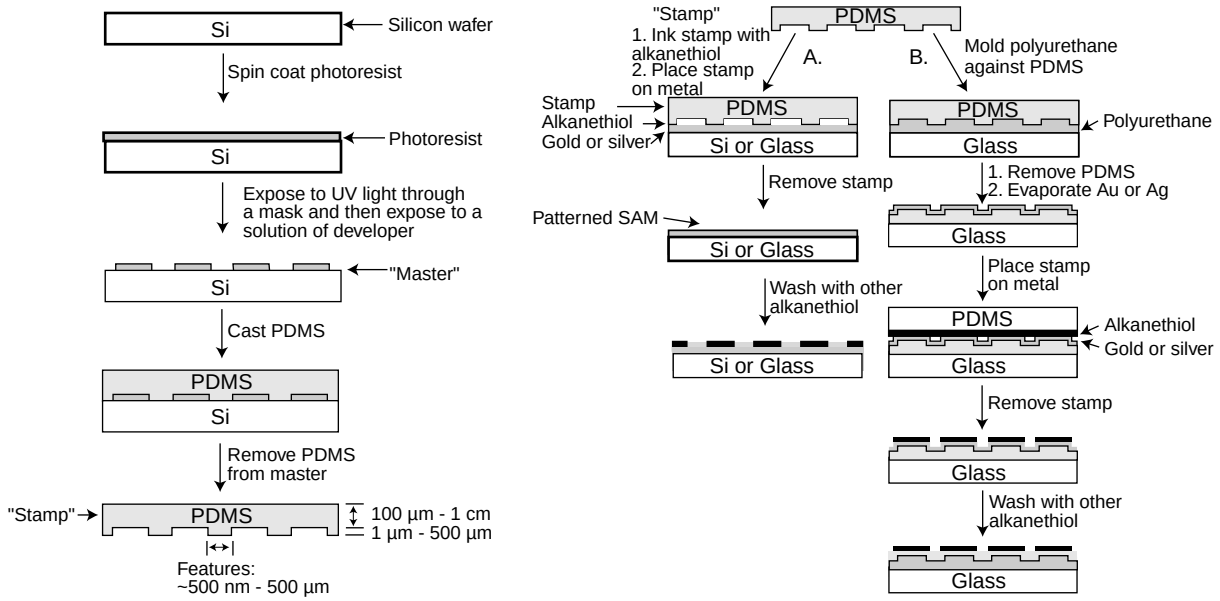


Figure 3. The creation of the master stamp is indicated on the left side of the figure, and its use to make patterns is indicated on the flow chart to the right (40).

Inkjet Printing

Inkjet technology used in desktop printers can be applied to the creation of viable cell patterns by printing proteins on to a surface (46). It is a fast and inexpensive method that does not require any contact (such as with stamping) to the surface. This method is desirable for high throughput printing of surfaces, and there is good control of the drop volume and of the alignment of the pattern. Printing occurs when small volumes of a protein solution or a solution containing cells is pumped through a nozzle in a continuous jet or small droplets of the solution are formed either by an acoustic or thermal pulse. The drop size is in the range of 10–20 pL (8). Inkjet printing and the use of computer-aided design (CAD) have impacted the biomaterials field greatly in the areas of biosensor development (47), immobilization of bacteria on biochips (48), DNA arrays and synthesis (49), microdeposition of proteins on cellulose (50), and free-form fabrication techniques to create acellular polymeric scaffolds. A drawback of this method is that the resolution, in the 20–50 μm range, is limited by the statistical variations in the drop direction and spreading on the surface (8). Other high throughput printing methods have also been adapted to pattern proteins on surfaces for biological application such as the already developed DNA spotter used to create DNA microchips (51).

Patterning via Microfluidic Networks

Synthetic surfaces may also be patterned using microfluidic networks (μFNs) to selectively generate regions with greater cytophilicity. This method involves the use of a microfluidic network fabricated in an elastomeric polymer, usually polydimethylsiloxane (PDMS), to direct a protein solution to the regions where cell-adhesion is desired. Gravity and pressure-driven flows are the most common methods for circulating the solution.

The most basic application of this method involves allowing a solution of the material to be patterned noncovalently to the substrate surface using the microchannels as guides. Some of the earliest work done with this method by Folch and Toner (52) involved patterning of various polymer surfaces with human plasma fibronectin (Fn) and collagen to create adhesion promoting domains for hepatocyte/fibroblast cocultures. Similar work by Chiu et al. (53) involved the patterning of glass coverslips with fibrinogen (Fb) and bovine serum albumin (BSA) for patterned cocultures with bovine adrenal capillary endothelial cells (BCEs) and human bladder cancer cells (ECVs). Further work by Takayama et al. (54) demonstrated an added degree of sophistication of this technique by using the laminar flow characteristics of microchannels to generate patterns using a single microchannel.

In addition to simple noncovalent binding of proteins to the substrate, it is possible to use a crosslinking agent to covalently link a molecule of interest. Delamarche et al. (55) used a hydroxylsuccinimidyl ester to chemically couple immunoglobulin G (IgG) to various substrates. Another more commonly used method involves functionalized silane SAMs on substrates, such as 3-aminopropyltriethoxysilane (APTES), and a crosslinking reagent, such glutaraldehyde (GA), to achieve crosslinking of protein molecules to a substrate. Romanova et al. (25) demonstrated the applicability of this method using microfluidic patterning to study controlled growth of *Aplysia* neurons on geometric patterns of poly-L-lysing and collagen IV. Yet another variation on this method was developed by Itoga et al. (56), who generated patterns by photopolymerization of acrylamide on 3-methacryloxypropyltrimethoxysilane modified glass coverslips. In this method, the acrylamide monomer was flowed through microchannels adhered to the derivatized coverslip and cross-linked via photopolymerization to generate cytopholic poly-acrylamide regions.

Perhaps the most sophisticated application of microfluidic patterning of substrates for cell adhesion was demonstrated by Tan and Desai (57,58), who demonstrated the fabrication of complex multilayer cocultures for biomimetic blood vessels. In this work, the 3D structure of blood vessels was recreated by differential deposition of protein and cell layers on a glass coverslip. By alternately layering proteins (Collagen I, collagen/chitosan, and matrigel) and cell types found in blood vessels (fibroblasts, smooth muscle, and endothelial cells), it was possible to recreate layers mimicking the adventitial, medial, and intimal layers observed in blood vessels.

Topographical (3D) Patterning Methods

Control of Cell Placement, Movement, and Process Growth Based on Topographical Clues. It has been known for many years that cells react to topographical clues in their environment (59). Originally, natural fibers were used to create topological clues. Later, fabrication methods developed for the microchip industry were adapted to micromachine silicon surfaces for cell culture applications (45). For the creation of micro- or nanotopography, sophisticated methods and equipment are necessary, which are available in most electronics laboratories, but are usually not available to cell biologists. In recent years, as a response to the increased need for high- throughput screening methods (planar patch clamp, lab-on-a-chip) and in response to the challenge of biological applications of nanoscience, several multidisciplinary team/centers have been established with microfabrication capabilities. The most commonly used fabrication methods are (*1*) silicon etching (60), (*2*) photoresist-based methods (61), (*3*) PDMS molding (62), and (*4*) polymer/hydrogel molding (63). Methods have also been developed for the creation of complex 3D structures by the rapid prototyping/layer-by-layer technique (58). Tan et al. (62) used 3D PDMS structures not only to control attachment and morphology of cells but also to measure attachment force through flexible microneedles as the culture substrate. Xi et al. used AFM cantilevers to demonstrate and measure the contracting force of cardiac muscle cells (64).

3D Patterning of Living Cells. Several hydrogel scaffold-based methods have also been developed to create 3D patterns from cells. For example, photo-polymerization of hydrogels can be used to create patterns of entrapped cells (65). These scaffolding methods offer possible intervention in spinal cord injury (66). Layer-by-layer methods have also been used to create complex "tissue analog" cellular structures, such as blood vessels (57).

Other Patterning Methods

Several other methods, based on microcontact printing and PDMS stamping, have been developed to create cellular patterns. Folch et al. used an inexpensive method to create microwells for coculture experiments based on a reusable elastomeric stencil (i.e., a membrane containing thru holes), which seals spontaneously against the surface (67,68). Gole and Sastry developed a novel method to pattern surfaces with lipids followed by selective protein incorporation into the lipid patterns that result in complex protein patterns on the surface (69). A scanning electrochemical microscope has also been adapted to pattern self-assembled monolayers on surfaces with high spatial resolution by either chemical removal of SAMs (70) or by gold deposition (71). Amro et al. used an AFM tip to directly print nanoscale patterns using the so-called dip-pen technique (72–74). However, none of these methods has been proven beyond the demonstration step.

Applications

Many *potential* applications for cell patterning remain in the biomedical and biotechnology fields. However, there has been limited success to date, besides demonstrations in the literature, for any of the applications envisioned by the host of researchers in this area. However, the promise of the use of cellular patterning for real applications is bolstered by the success that has been achieved for patterning of DNA, RNA, and protein arrays (42,51) as well as for enzymatic biosensors, such as simple pregnancy tests. Much like with cellular patterning, there was a period of time during the development of molecular patterning techniques before the applications became relevant, and the authors believe this is the situation that exists for cell patterning at this time. One reason for this long development stage is that viable and reproducible cellular patterns have many other variables that are not a major issue with biomolecule patterning applications. The cellular media are very important for cell survival, especially long term, as well as cell preparation, which has a significant affect on an extracellular attachment. In addition, no universal combination will be good for every cell, as each cell type has a unique environment that it needs to survive and function, and some aspect of these factors needs to be reproduced for long-term applications. That said, many examples of tissues exhibit some segregation, including blood vessels, lung tissue, the lining of the stomach and intestines, as well as a host of other tissues, which could benefit from this methodology. However, one of the most studied cell tissues is that of the central nervous system (CNS), which exhibits a complicated network of structures that will be difficult to reproduce for reconstruction or repair of neuronal tissue or for other *in vivo* as well as *in vitro* applications. To date, there has been some success in manipulating cells in patterns and controlling certain variables that would be necessary for the creation of functional tissues, but a complete system, using this methodology, has not been reported in the literature. However, there has been some success in demonstrating the intermediate steps that will be necessary for the realization of applications of this method for biomedical and biotechnological applications. Cell attachment has been demonstrated by several researchers, and generally, pattern adherence is maintained for approximately 1 week although longer times have been demonstrated (75). Control of cell morphology and differentiation are two important factors that are necessary in creating functional systems, and these have also been demonstrated. For neuronal-based systems, the primary variable, that of axonal polarity, has also been demonstrated. Brief descriptions and progress in these areas are described below.

Controlling Cell Attachment, Morphology, and Differentiation. *In vivo*, cells are arranged in distinct patterns (76). This patterning effect is dictated during development, with cues provided by both physical contact with other cells and chemicals present in the extracellular matrix (77). Because the random arrangement of cells cultured *in vitro* does not represent the complex architecture seen in tissues, studies of many cell types lack a clear *in vivo* relationship. Consequently, techniques to create defined and reproducible functional patterns of cells on surfaces have been created.

Controlling Attachment and Morphology. Several methods have been developed to control the attachment of cells on surfaces creating patterns that more accurately mimic conditions found *in vivo*. Using photolithography, microcontact printing, and microstamping, groups have been able create 2D patterns that guide cell attachment and alignment (37,78–81). Three-dimensional patterning techniques have also been employed to influence cell orientation and polarity of neurons and osteoblasts (82–84). Cells have also been attached to capillaries and microfluidic devices using SAMs and protein adsorption or microcontact printing (85,86). Additionally, cell attachment and proliferation has been enhanced using biomolecules attached covalently, by stamping, or microcontact printing to surfaces (37,78–82,84–87).

Controlling Morphology and Differentiation of Cell Types Other Than Neurons. Microfabrication and photolithography used to create microtextured membranes for cardiac myocyte culture showed greater levels of attachment and cell height relative to 2D culture techniques (88). Similar techniques applied to vascular smooth muscle also showed the ability to control shape and size of the cells (89). Furthermore, microtextured surfaces were shown to influence gene expression and protein localization in neonatal cardiomyocytes (90). Cues provided to cells by the topography of their extracellular environment are thought to play a role in differentiation. The generation of microtopographical surfaces in titanium has been used to regulate the differentiation of osteoblasts *in vitro* (91).

Study of Axon Guidance in Neurons. Using photolithographic techniques and SAMs, the 2D patterns created were shown to influence neuronal polarity (22). Photolithographically fabricated 3D surfaces demonstrated that topology also influenced the orientation of neurons and the polarity of axonal outgrowth (83). The ability to guide neurite outgrowth and axonal elongation has significant applications in the areas of spinal cord repair, synapse, formation, and neural network formation. Initial studies using striped patterns on glass coverslips showed that neurons would adhere and preferentially extend axons along the length of the pattern (4,92). Growth of neurons on micropatterned 2D surfaces showed preferential axon extension along the length of the pattern as well as increased axon extension (92–96).

The use of 3D microchannels and microstructured surfaces has also been shown to increase the complexity of neuronal architecture, increase neurite growth, and enhance cell activity (61). Photolithography has also been used to pattern neurons and control axon elongation for the formation of neuronal networks (25,97). The synapses formed by these hippocampal neurons showed strong electrophysiological activity up to 17 days in culture (10). These network formations show promise for use in screening pharmacological agents as well as for electronic connection.

BIBLIOGRAPHY

Cited References

1. Fritz M, Belcher AM, Radmacher M, Walters DA, Hansma PK, Stucky GD, Morse DE, Mann S. Flat pearls from biofabrication of organized composites on inorganic substrates. Nature Biotechnol 1994;371:49–51.
2. Noctor SC, Flint AC, Weissman TA, Dammerman RS, - Kriegstein AR. Neurons derived from radial glial cells establish radial units in neocortex. Nature 2001: 409(6821): 714–720.
3. Harrison RG. The reaction of embryonic cells to solid structures. J Exp Zool 1914;17:521–544.
4. Kleinfeld D, Kahler KH, Hockberger PE. Controlled outgrowth of dissociated neurons on patterned substrates. J Neurosci 1988;8(11):4098–4120.
5. Spargo BJ, Testoff MA, Nielsen TB, Stenger DA, Hickman JJ, Rudolph AS. Spatially controlled adhesion, spreading, and differentiation of endothelial-cells on self-assembled molecular monolayers. Proc Nat Acad Sci 1994;91(23):11070–11074.
6. Das M, Molnar P, Gregory C, Riedel L, Jamshidi A, Hickman JJ. Long-term culture of embryonic rat cardiomyocytes on an organosilane surface in a serum-free medium. Biomaterials 2004;25(25):5643–5647.
7. Das M, Bhargava N, Gregory C, Riedel L, Molnar P, Hickman JJ. Adult rat spinal cord culture on an organosilane surface in a novel serum-free medium. In Vitro Animal Cell Develop Bio 2005. In press.
8. Geissler M, Xia Y. Patterning: Principles and some new developments. Adv Mater 2004;16(15):1249–1269.
9. Wyart C, Ybert C, Bourdieu L, Herr C, Prinz C, Chatenay D. Constrained synaptic connectivity in functional mammalian neuronal networks grown on patterned surfaces. J Neurosci Methods 2002;117(2):123–131.
10. Dulcey CS, Georger JHJr. Krauthamer V, Stenger DA, Fare TL, Calvert JM. Deep UV photochemistry of chemisorbed monolayers: Patterned coplanar molecular assemblies. Science 1991;252(5005):551–554.
11. Dressick WJ, Calvert JM. Patterning of self-assembled films using lithographic exposure tools. Appl Phys Part 1 1993; 32(12B):5829–5839.
12. Bhatia SK, Teixeira JL, Anderson M, Shriver-Lake LC, Calvert JM, Georger JH, Hickman JJ, Dulcey CS, Schoen PE, Ligler FS. Fabrication of surfaces resistant to protein adsorption and application to two-dimensional protein patterning. Anal Biochem 1993;208(1):197–205.
13. Liu J, Hlady V. Chemical pattern on silica surface prepared by UV irradiation of 3-mercaptopropyltriethoxy silane layer: Surface characterization and fibrinogen adsorption. Colloids Surfaces B-Biointerfaces 1996;8(1–2):25–37.
14. Dressick WJ, Dulcey CS, Chen MS, Calvert JM. Photochemical studies of (aminoethylaminomethyl)phenethyltrimethoxysilane self-assembled monolayer films. Thin Solid Films 1996;285:568–572.
15. Georger JH, Stenger DA, Rudolph AS, Hickman JJ, Dulcey CS, Fare TL. Coplanar patterns of self-assembled monolayers for selective cell-adhesion and outgrowth. Thin Solid Films 1992;210(1–2):716–719.

16. Stenger DA, Georger JH, Dulcey CS, Hickman JJ, Rudolph AS, Nielsen TB, McCort SM, Calvert JM. Coplanar molecular assemblies of aminoalkylsilane and perfluorinated alkylsilane—characterization and geometric definition of mammalian-cell adhesion and growth. J Am Chem Soc 1992;114 (22):8435–8442.
17. Calvert JM. Lithographic patterning of self-assembled films. J Vacuum Sci Technol B 1993;11(6):2155–2163.
18. Ravenscroft MS, Bateman KE, Shaffer KM, Schessler HM, Jung DR, Schneider TW, Montgomery CB, Custer TL, Schaffner AE, Liu QY, Li YX, Barker JL, Hickman JJ. Developmental neurobiology implications from fabrication and analysis of hippocampal neuronal networks on patterned silane- modified surfaces. J Am Chem Soc 1998;120(47): 12169–12177.
19. Plueddemann EP. Silane adhesion promoters in coatings. Progr Organic Coatings 1983;11(3):297–308.
20. Whitesides GM, Laibinis P, Folkers J, Prime K, Seto C, Zerkowski J. Self-assembly—alkanethiolates on gold and hydrogen-bonded networks. Abstr Papers Am Chem Soc 1991;201:103-INOR.
21. Gillen G, Wight S, Bennett J, Tarlov MJ. Patterning of self-assembled alkanethiol monolayers on silver by microfocus ion and electron-beam bombardment. Appl Phys Lett 1994;65(5): 534–536.
22. Stenger DA, Hickman JJ, Bateman KE, Ravenscroft MS, Ma W, Pancrazio JJ, Shaffer K, Schaffner AE, Cribbs DH, Cotman CW. Microlithographic determination of axonal/dendritic polarity in cultured hippocampal neurons. J Neurosci Methods 1998;82(2):167–173.
23. Matsuda T, Sugawara T. Development of surface photochemical modification method for micropatterning of cultured-cells. J Biomed Mater Res 1995;29(6):749–756.
24. Dulcey CS, Georger JH, Chen MS, McElvany SW, Oferrall CE, Benezra VI, Calvert JM. Photochemistry and patterning of self-assembled monolayer films containing aromatic hydrocarbon functional groups. Langmuir 1996;12(6):1638–1650.
25. Romanova EV, Fosser KA, Stanislav SR, Nuzzo RG, Sweedler JV. Engineering the morphology and electrophysiological parameters of cultured neurons by microfluidic surface patterning. FASEB J 2004.
26. Corey JM, Wheeler BC, Brewer GJ. Compliance of hippocampal neurons to patterned substrate networks. J Neurosci Res 1991;30(2):300–307.
27. Griscom L, Degenaar P, LePioufle B, Tamiya E, Fujita H. Techniques for patterning and guidance of primary culture neurons on micro-electrode arrays. Sen Actuators B 2002;83 (1–3):15–21.
28. Nicolau DV, Taguchi T, Taniguchi H, Tanigawa H, Yoshikawa S. Patterning neuronal and glia cells on light-assisted functionalised photoresists. Biosens Bioelectron 1999;14(3):317–325.
29. He W, Halberstadt CR, Gonsalves KE. Lithography application of a novel photoresist for patterning of cells. Biomaterials 2004;11:2055–8063.
30. Liu GY, Amro NA. Positioning protein molecules on surfaces: A nanoengineering approach to supramolecular chemistry. Proc Nat Acad Sci 2002;99(8):5165–5170.
31. Pirrung MC, Huang CY. A general method for the spatially defined immobilization of biomolecules on glass surfaces using "caged" biotin. Bioconjugate Chem 1996;7(3):317–321.
32. Sigrist H, Collioud A, Clemence JF, Gao H, Luginbuhl R, Sanger M, Sundarababu G. Surface immobilization of biomolecules by light. Opt Eng 1995;34(8):2339–2348.
33. Herbert CB, McLernon TL, Hypolite CL, Adams DN, Pikus L, Huang CC, Fields GB, Letourneau PC, Distefano MD, Hu WS. Micropatterning gradients and controlling surface densities of photoactivatable biomolecules on self-assembled monolayers of oligo(ethylene glycol) alkanethiolates. Chem Bio 1997;4(10): 731–737.
34. Sorribas H, Padeste C, Tiefenauer L. Photolithographic generation of protein micropatterns for neuron culture applications. Biomaterials 2002;23(3):893–900.
35. Lee K-N, Shin D-S, Lee Y-S, Kim Y-K. Protein patterning by virtual mask photolithography using a micromirror array. J Micromech Microeng 2003;13(1):18–25.
36. Chen GP, Imanishi Y, Ito Y. Effect of protein and cell behavior on pattern-grafted thermoresponsive polymer. J Biomed Mater Res 1998;42(1):38–44.
37. Singhvi R, Kumar A, Lopez GP, Stephanopoulos GN, Wang DI, Whitesides GM, Ingber DE. Engineering cell shape and function. Science 1994;264(5159):696–698.
38. Lahiri J, Ostuni E, Whitesides GM. Patterning ligands on reactive SAMs by microcontact printing. Langmuir 1999;15 (6):2055–2060.
39. Cornish T, Branch DW, Wheeler BC, Campanelli JT. Microcontact printing: A versatile technique for the study of synaptogenic molecules. Mol Cell Neurosci 2002;20(1):140–153.
40. Kane RS, Takayama S, Ostuni E, Ingber DE, Whitesides GM. Patterning proteins and cells using soft lithography. Biomaterials 1999;20(23–24):2363–2376.
41. Oliva AA, James CD, Kingman CE, Craighead HG, Banker GA. Patterning axonal guidance molecules using a novel strategy for microcontact printing. Neurochem Res 2003;28 (11):1639–1648.
42. Martin BD, Gaber BP, Patterson CH, Turner DC. Direct protein microarray fabrication using a hydrogel "stamper". Langmuir 1998;14(15):3971–3975.
43. Mayer M, Yang J, Gitlin I, Gracias DH, Whitesides GM. Micropatterned agarose gels for stamping arrays of proteins and gradients of proteins. Proteomics 2004;4(8):2366–2376.
44. Lauer L, Ingebrandt S, Scholl M, Offenhausser A. Aligned microcontact printing of biomolecules on microelectronic device surfaces. IEEE Trans Biomed Eng 2001;48(7):838–842.
45. Craighead HG, James CD, Turner AMP. Chemical and topographical patterning for directed cell attachment. Curr Opin Solid State Mater Sci 2001;5(2–3):177–184.
46. Roth EA, Xu T, Das M, Gregory C, Hickman JJ, Boland T. Inkjet printing for high-throughput cell patterning. Biomaterials 2004;25(17):3707.
47. Newman JD, Turner APF, Marrazza G. Ink-jet printing for the fabrication of amperometric glucose biosensors. Anal Chim Acta 1992;262(1):13–17.
48. Xu T, Jin J, Gregory C, Hickman JJ, Boland T. Inkjet printing of viable mammalian cells. Biomaterials 2005;26(1):93–99.
49. Schena M, Heller RA, Thieriault TP, Konrad K, Lachenmeier E, Davis RW. Microarrays: Biotechnology's discovery platform for functional genomics. Trends Biotechnol 1998;16(7):301–306.
50. Roda A, Guardigli M, Russo C, Pasini P, Baraldini M. Protein microdeposition using a conventional ink-jet printer. Biotechniques 2000;28(3):492–496.
51. Flaim CJ, Chien S, Bhatia SN. An extracellular matrix microarray for probing cellular differentiation. Nature Methods 2005;2(2):119–125.
52. Folch A, Toner M. Cellular micropatterns on biocompatible materials. Biotechnol Progr 1998;14(3):388–392.
53. Chiu DT, Jeon NL, Huang S, Kane RS, Wargo CJ, Choi IS, Ingber DE. Whitesides GM. Patterned deposition of cells and proteins onto surfaces by using three-dimensional microfluidic systems. Proc Natl Acad Sci USA 2000;97(6):2408–2413.
54. Takayama S, McDonald JC, Ostuni E, Liang MN, Kenis PJA, Ismagilov RF, Whitesides GM. Patterning cells and their

environments using multiple laminar fluid flows in capillary networks. Proc Natl Acad Sci USA 1999;96(10):5545–5548.
55. Delamarche E, Bernard A, Schmid H, Michel B, Biebuyck H. Patterned delivery of immunoglobulins to surfaces using microfluidic networks. Science 1997;276(5313):779–781.
56. Itoga K, Yamamoto JK, Kikuchi A, Okano T. Micropatterned surfaces prepared using a liquid crystal projector-modified photopolymerization device and microfluidics. J Biomed Mater Res 2004;69A:391–397.
57. Tan W, Desai TA. Microscale multilayer cocultures for biomimetic blood vessels. J Biomed Mater Res 2005;72A(2):146–160.
58. Tan W, Desai TA. Layer-by-layer microfluidics for biomimetic three-dimensional structures. Biomaterials 2004;25(7–8): 1355–1364.
59. Curtis A, Wilkinson C. Topographical control of cells. Biomaterials 1997;18(24):1573–1583.
60. Turner S, Kam L, Isaacson M, Craighead HG, Shain W, Turner J. Cell attachment on silicon nanostructures. J Vacuum Sci Technol B 1997;15(6):2848–2854.
61. Mahoney MJ, Chen RR, Tan J, Saltzman WM. The influence of microchannels on neurite growth and architecture. Biomaterials 2005;26(7):771–778.
62. Tan JL, Tien J, Pirone DM, Gray DS, Bhadriraju K, Chen CS. Cells lying on a bed of microneedles: An approach to isolate mechanical force. Proc Natl Acad Sci USA 2003;100(4):1484–1489.
63. Recknor JB, Recknor JC, Sakaguchi DS, Mallapragadaa SK. Oriented astroglial cell growth on micropatterned polystyrene substrates. Biomaterials 2004;25(14):2753–2767.
64. Xi JZ, Schmidt J, Montemagno C. Development of self-assembled muscle-MEMS microdevices. Biophys J 2004;86 (1):481A.
65. Albrecht DR, Tsang VL, Sah RL, Bhatia SN. Photo- and electropatterning of hydrogel-encapsulated living cell arrays. Lab Chip 2005;5(1):111–118.
66. Bloch J, Fine EG, Bouche N, Zurn AD, Aebischer P. Nerve growth factor- and neurotrophin-3-releasing guidance channels promote regeneration of the transected rat dorsal root. Exper Neurol 2001;172(2):425–432.
67. Folch A, Jo BH, Hurtado O, Beebe DJ, Toner M. Microfabricated elastomeric stencils for micropatterning cell cultures. J Biomed Mater Res 2000;52(2):346–353.
68. Ostuni E, Kane R, Chen CS, Ingber DE, Whitesides GM. Patterning mammalian cells using elastomeric membranes. Langmuir 2000;16(20):7811–7819.
69. Gole A, Sastry M. A new method for the generation of patterned protein films by encapsulation in arrays of thermally evaporated lipids. Biotechnol Bioeng 2001;74(2):172–178.
70. Shiku H, Uchida I, Matsue T. Microfabrication of alkylsilanized glass substrate by electrogenerated hydroxyl radical using scanning electrochemical microscopy. Langmuir 1997;13(26):7239–7244.
71. Turyan I, Matsue T, Mandler D. Patterning and characterization of surfaces with organic and biological molecules by the scanning electrochemical microscope. Anal Chem 2000;72 (15):3431–3435.
72. Amro NA, Xu S, Liu GY. Patterning surfaces using tip-directed displacement and self-assembly. Langmuir 2000; 16(7): 3006–3009.
73. Schwartz PV. Molecular transport from an atomic force microscope tip: A comparative study of dip-pen nanolithography. Langmuir 2002;18(10):4041–4046.
74. Agarwal G, Sowards LA, Naik RR, Stone MO. Dip-pen nanolithography in tapping mode. J Am Chem Soc 2003; 125(2):580–583.
75. Chang JC, Brewer GJ, Wheeler BC. A modified microstamping technique enhances polylysine transfer and neuronal cell patterning. Biomaterials 2003;24:2862–2870.
76. Curtis A, Riehle M. Tissue engineering: the biophysical background. Phys Med Biol 2001; 46(4):R47–R65.
77. Curtis A, Wilkinson C. Reactions of cells to topography. J Biomater Sci Pol Educ 1998;9:1313–1329.
78. Li B, Ma Y, Wang S, Moran PM. A technique for preparing protein gradients on polymeric surfaces: effects on PC12 pheochromocytoma cells. Biomaterials 2005;26:1487–1495.
79. McFarland CD, Thomas CH, DeFilippis C, Stelle JG, Healy KE. Protein adsorption and cell attachment to patterned surfaces. J Biomed Mater Res 2000;49(2):200–210.
80. Veiseh M, Wickes BT, Castner DG, Zhang MQ. Guided cell patterning on gold-silicon dioxide substrates by surface molecular engineering. Biomaterials 2004;25(16):3315–3324.
81. Zheng H, Berg MC, Rubner MF, Hammond PT. Controlling cell attachment selectively onto biological polymer-colloid templates using polymer-on-polymer stamping. Langmuir 2004;20(17):7215–7222.
82. Ber S, Kose GT, Hasirci V. Bone tissue engineering on patterned collagen films: An *in vitro* study. Biomaterials 2005; 16:1977–1986.
83. Dowell-Mesfin NM, Abdul-Karim MA, Turner AM, Schanz S, Craighead HG, Roysam B, Turner JN, Shain W. Topographically modified surfaces affect orientation and growth of hippocampal neurons. J Neural Eng 2004;1(2): 78–90.
84. Mohammed JS, DeCoster MA, McShane MJ. Micropatterning of nanoengineered surfaces to study neuronal cell attachment *in vitro*. Biomacromolecules 2004;5(5):1745–1755.
85. Mrksich M, Chen CS, Xia YN, Dike LE, Ingber DE, Whitesides GM. Controlling cell attachment on contoured surfaces with self- assembled monolayers of alkanethiolates on gold. Proc Natl Acad Sci USA 1996;93(20):10775–10778.
86. Theibaud P, Lauer L, Knoll W, Offenhausser A. PDMS device for patterned application of microfluids to neuronal cells arranged by microcontact printing. Biosen Bioelectro 2002; 17:87–93.
87. Scholl M, Sprossler C, Denyer M, Krause M, Nakajima K, Maelicke A, Knoll W, Offenhausser A. Ordered networks of rat hippocampal neurons attached to silicon oxide surfaces. J Neurosci Methods 2000;104(1):65–75.
88. Deutsch J, Motiagh D, Russell B, Desai TA. Fabrication of microtextured membranes for cardiac myocyte attachment and orientation. J Biomed Mater Res 2000;53(3):267–275.
89. Goessl A, Bowen-Pope DF, Hoffman AS. Control of shape and size of vascular smooth muscle cells *in vitro* by plasma lithography. J Biomed Mater Res 2001;57(1):15–24.
90. Motlagh D, Senyo SE, Desai TA, Russell B. Microtextured substrata alter gene expression, protein localization and the shape of cardiac myocytes. Biomaterials 2003;24(14):2463–2476.
91. Zinger O, Zhao G, Schwartz Z, Simpson J, Wieland M, Landolt D, Boyan B. Differential regulation of osteoblasts by substrate microstructural features. Biomaterials 2005;26 (14):1837–1847.
92. Matsuzawa M, Liesi P, Knoll W. Chemically modifying glass surfaces to study substratum-guided neurite outgrowth in culture. J Neurosci Methods 1996;69(2):189–196.
93. Clark P, Britland S, Connolly P. Growth cone guidance and neuron morphology on micropatterned laminin surfaces. J Cell Sci 1993;105:203–212.

94. Saneinejad S, Shoichet MS. Patterned poly(chlorotrifluoroethylene) guides primary nerve cell adhesion and neurite outgrowth. J Biomed Mater Res 2000;50(4):465–474.
95. Tai H, Buettner HM. Neurite outgrowth and growth cone morphology on micropatterned surfaces. Biotechnol Prog 1998;14:364–370.
96. Zhang ZP, Yoo R, Wells M, Beebe TP, Biran R, Tresco P. Neurite outgrowth on well-characterized surfaces: Preparation and characterization of chemically and spatially controlled fibronectin and RGD substrates with good bioactivity. Biomaterials 2005;26(1):47–61.
97. Heller DA, Garga V, Kelleher KJ, Lee TC, Mahbubani S, Sigworth LA, Lee TR, Rea MA. Patterned networks of mouse hippocampal neurons on peptide-coated gold surfaces. Biomaterials 2005;26(8):883–889.

See also BIOCOMPATIBLITY OF MATERIALS; BIOMATERIALS, SURFACE PROPERTIES OF; BIOMATERIALS: TISSUE ENGINEERING AND SCAFFOLDS.

BIOMEDICAL EQUIPMENT MAINTENANCE.

See EQUIPMENT MAINTENANCE, BIOMEDICAL.

BIOSENSORS.

See IMMUNOLOGICALLY SENSITIVE FIELD-EFFECT TRANSISTORS.

BIOTELEMETRY

BABAK ZIAIE
Purdue University
W. Lafayette, Indiana

INTRODUCTION

The ability to use wireless techniques for measurement and control of various physiological parameters inside human and animal bodies has been a long-term goal of physicians and biologists going back to the early days of wireless communication. From early on, it was recognized that this capability could provide effective diagnostic, therapeutic, and prosthetic tools in physiological research and pathological intervention. However, this goal eluded scientists prior to the invention of transistor in 1947. Vacuum tubes were too bulky and power hungry to be of any use in many wireless biomedical applications. During the late 1950s, MacKay performed his early pioneering work on what he called Endoradiosonde (1). This was a single-transistor blocking oscillator designed to be swallowed by a subject and was able to measure pressure and temperature in the digestive track. Following this early work, came a number of other simple discrete systems each designed to measure a specific parameter (temperature, pressure, force, flow, etc.) (2). By the late 1960s, progress in the design and fabrication of integrated circuits provided an opportunity to expand the functionality of these early systems. Various hybrid single and multichannel telemetry systems were developed during the 1970s and the 1980s (3). In addition, implantable therapeutic and prosthetic devices started to appear in the market. Cardiac pacemakers and cochlear prosthetics proved effective and reliable enough to be implanted in thousands of patients. We direct the interested readers to several excellent reviews published over the past several decades summarizing these advances in their perspective time periods. These include a review article by W. H. Ko and M. R. Neuman in the *Science* covering the technologies available in the 1960s (4) and another similar paper by Topich covering the 1970s period (5). Three subsequent reviews detailed the efforts in the 1980s (6–8) followed by the most recent article published in 1999 (9). An outdated, but classic reference book in biotelemetry, is by MacKay, which still can be used as a good starting point for some simple single channel systems and includes some ingenious techniques used by early investigators to gain remote physiological information (10).

The latter part of the 1990s witnessed impressive advances in microelectromechanical (MEMS) based transducer and packaging technology, new and compact power sources (high efficiency inductive powering and miniature batteries), and CMOS low power wireless integrated circuits that provided another major impetus to the development of biotelemetry systems (11–18). These advances have created new opportunities for increased reliability and functionality, which had been hard to achieve with pervious technologies. The term biotelemetry itself has been for most part superseded by Microbiotelemetry or Wireless Microsystems to denote these recent changes in technology. Furthermore, the burgeoning area of nanotechnology is poised to further enhance these capabilities beyond what have been achievable using current miniaturization techniques. This is particularly true in the biochemical sensing and chemical delivery areas and will undoubtedly have a major impact on the future generations of implantable biotelemetry microsystems.

This review article is intended to complement and expand the earlier reviews by emphasizing newer developments in the area of biomedical telemetry in particular attention is paid to the opportunities created by recent advances in the area of microbiotelemetry (i.e., systems having volumes $\sim$1 cm^3 or less) by low power CMOS wireless integrated circuits, micromachined–MEMS transducers, biocompatible coatings, and advanced batch-scale packaging. We have both expanded and narrowed the traditional definition of biotelemetry by including therapeutic–rehabilitative microsystems and excluding wired devices that although fit under the strict definition of biotelemetry; do not constitute an emerging technology. In the following sections, after discussing several major components of such biotelemetry microsystems, such as transducers, interface electronics, wireless communication, power sources, and packaging, we will present some selected examples to demonstrate the state of the art. These include implantable systems for biochemical and physiological measurements, drug delivery microsystems, and neuromuscular and visual prosthetic devices. Although our primary definition of biotelemetry encompass devices with active electronics and signal processing capabilities, we will also discuss passive MEMS-based transponders that do not require on-board signal processing and can be interrogated using simple radio-frequency (rf) techniques. Finally, we should mention that although in a strict sense biotelemetry encompasses systems targeted for physiological measurements, this

narrow definition is no longer valid or desirable. A broader scope including neuromuscular stimulation and chemical delivery is currently understood to be more indicative of the term biotelemetry.

BIOTELEMETRY SYSTEMS

For the purpose of current discussion biotelemetry systems can be defined as a group of medical devices that (*1*) incorporate one or several miniature transducers (i.e., sensors and actuators), (*2*) have an on-board power supply (i.e., battery) or are powered from outside using inductive coupling, (*3*) can communicate with outside (bidirectional or unidirectional) through an rf interface, (*4*) have on-board signal processing capability, (*5*) are constructed using biocompatible materials, and (*6*) use advanced batch-scale packaging techniques. Although one microsystem might incorporate all of the above components, the demarcation line is rather fluid and can be more broadly interpreted. For example, passive MEMS-based microtransponders do not contain on-board signal processing capability, but use advanced MEMS packaging and transducer technology and are usually considered to be a telemetry device. We should also emphasize that the above components are interrelated and a good system designer must pay considerable attention from the onset to this fact. For example, one might have to choose a certain power source or packaging scheme to accommodate the desired transducer, interface electronics, and wireless communication. In the following sections, we will discuss various components of a typical biotelemetry system with more attention being paid to the wireless communication block. For other components, we provide a brief discussion highlighting major recent developments and refer the reader to some recent literature is these areas.

Transducers

Transducers are interfaces between biological tissue and readout electronics–signal processing. Their performance is critical to the success of the overall microsystem (19–24). Current trend in miniaturization of transducers and their integration with signal processing circuitry have considerably enhanced their performance. This is particularly true with respect to MEMS-based sensors and actuators, where the advantages of miniaturization have been prominent. Development in the area of microactuators has been lagging behind the microsensors due to the inherent difficulty in designing microdevices that efficiently and reliably generate motion. Although some transducing schemes, such as electrostatic force generation, has advantageous scaling properties in the microdomain, problems associated with packaging and reliability has prevented their successful application. The MEMS-based microsensors have been more successful and offer several advantages compared to the macrodomain counterparts. These include lower power consumption, increased sensitivity, higher reliability, and lower cost due to batch fabrication. However, they suffer from a poor signal/noise ratio, hence requiring a close by interface circuit. Among the many microsensors designed and fabricated over the past two decades, physical sensors have been by and large more successful. This is due to their inherent robustness and isolation from any direct contact with biological tissue in sensors, such as accelerometers and gyroscopes. Issues related to packaging and long-term stability have plagued the implantable chemical sensors. Long-term baseline and sensitivity stability are major problems associated with implantable sensors. Depending on the type of the sensor, several different factors contribute to the drift. For example, in implantable pressure sensors, packaging generated stresses due to thermal mismatch and long-term material creep are the main sources of baseline drift. In chemical sensors, biofouling and fibrous capsule formation is the main culprit. Some of these can be mitigated through clever mechanical design and appropriate choice of material, however, some are more difficult to prevent (e.g., biofouling and fibrous capsule formation). Recent developments in the area of antifouling material and controlled release have provided new opportunities to solve some of these long standing problems (25–27).

Interface Electronics

As mentioned previously, most miniature and MEMS-based transducers suffer from poor signal/noise ratio and require on-board interface electronics. This, of course, is also more essential for implantable microsystems. The choice of integrating the signal processing with the MEMS transducer on the same substrate or having a separate signal processing chip in close proximity depends on many factors, such as process complexity, yield, fabrication costs, packaging, and general design philosophy. Except for post-CMOS MEMS processing methods, which rely on undercutting micromechanical structures subsequent to the fabrication of the circuitry (28), other integrated approaches require extensive modifications to the standard CMOS processes and have not been able to attract much attention. Post-CMOS processing is an attractive approach although packaging issues still can pose roadblocks to successful implementation. Hybrid approach has been typically more popular with the implantable biotelemetry microsystem designers providing flexibility at a lower cost. Power consumption is a major design consideration in implantable wireless microsystems that rely on batteries for an energy source. Low power and subthreshold CMOS design can reduce the power consumption to nanowatt levels (29–33). Important analogue and mixed-signal building blocks for implantable wireless microsystems include amplifiers, oscillators, multiplexers, A/D and D/A converters, and voltage references. In addition, many such systems require some digital signal processing and logic function in the form of finite-state machines. In order to reduce the power consumption, it is preferable to perform the DSP functions outside the body although small finite-state machines can be implemented at low power consumptions.

Wireless Communication

The choice of appropriate communication scheme for a biotelemetry system depends on several factors, such as (*1*) number of channels, (*2*) device lifetime, and (*3*) transmission range. For single (or two) channel systems, one can choose a variety of modulation schemes and techniques.

These systems are the oldest type of biotelemetry devices (1) and can range from simple blocking oscillators to single channel frequency modulation (FM) transmitters. They are attractive since one can design a prototype rather quickly using off-the-shelf components. Figure 1 shows a schematic of the famous blocking oscillator first used by MacKay to transmit pressure and temperature (10). It consists of a single bipolar transistor oscillator configured to periodically turn itself on and off. The oscillation frequency depends on the resonant frequency of the tank circuit that can be made to vary with parameters, such as pressure, by including a capacitive or inductive pressure sensor. The on–off repetition frequency can be made to depend on the temperature by incorporating a thermistor in the circuit. This is an interesting example of an ingenious design that can be accomplished with a minimum amount of effort and hardware. An example of a more recent attempt at single channel telemetry is a two-channel system designed by Mohseni et al. to transmit moth electromyograms (34). The circuit schematic and a picture of the fully assembled device are shown in Fig. 2. As can be seen, each channel

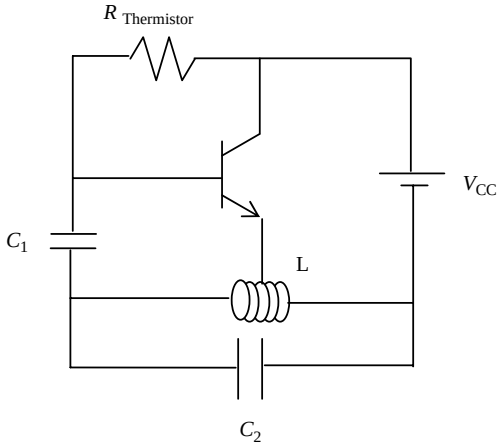
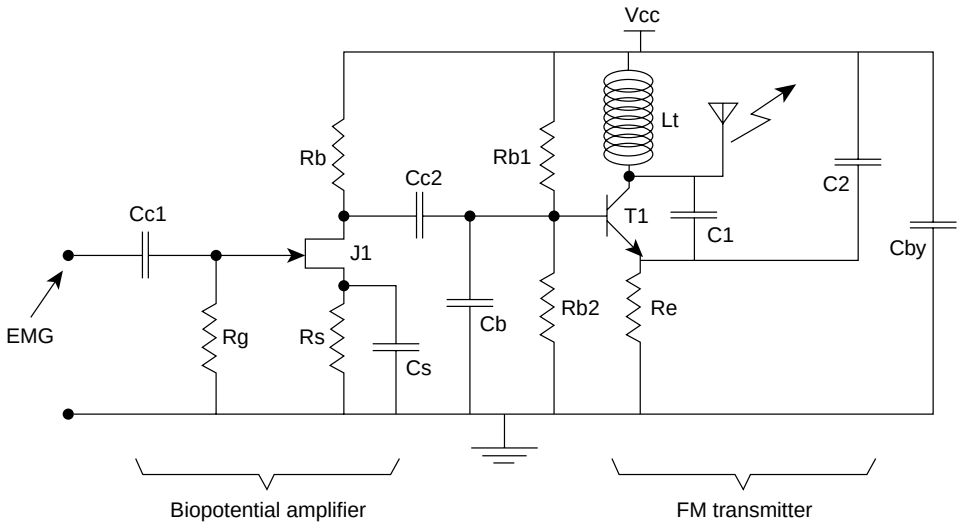


Figure 1. Schematic circuit of a blocking oscillator used to transmit pressure and temperature.



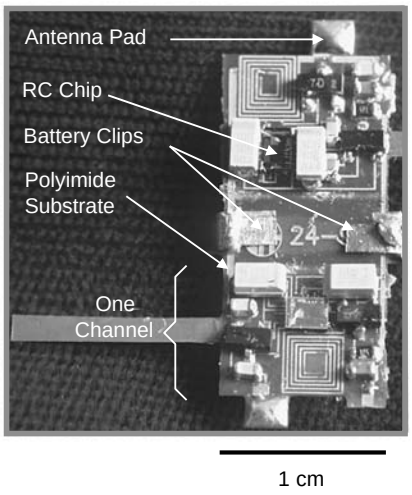


Figure 2. Schematic diagram and photograph of a biotelemetry system used to transmit flight muscle electromyograms in moths showing the polyimide flex circuit and various components (the Colpitts Oscillator inductor is used as the transmitting antenna).

consists of a biopotential amplifier followed by a Colpitts oscillator with operating frequency tunable in the 88–108 MHz commercial FM band. The substrate for the biotelemetry module was a polyimide flex circuit in order to reduce the weight such that the Moth can carry the system during flight. The overall system measures $10 \times 10 \times 3$ mm, weighs 0.74 g, uses two 1.5 V batteries, dissipates $\sim$ 2 mW, and has a transmission range of 2 m.

Multichannel systems are of more scientific and clinical interest. These systems rely on different and more elaborate communication schemes. For the purpose of current discussion, we will divide these systems into the ones that operate with a battery and the ones that are powered from outside using an inductive link. Battery-operated biotelemetry microsystems rely on different communication schemes than the inductively powered ones. Figure 3 shows a schematic block diagram of a time-division multiplexed multichannel system. It consists of several transducers with their associated signal conditioning circuits. These might include operations, such as simple buffering, low level amplification, filtering, or all three. Subsequent to signal conditioning, different channels are multiplexed using an analogue MUX. Although recent advances in AD technology might allow each channel to be digitized prior to multiplexing, this is not an attractive option for biotelemetry systems (unless there are only a few channels), since it requires an increase in power consumption that most biotelemetry systems cannot afford. All the timing and framing information is also added to the outgoing multiplexed signal at this stage. After multiplexing, an AD converter is used to digitize the signal. This is followed by a rf transmitter and a miniature antenna. The transmitted signal is picked up by a remote receiver and the signal is demodulated and separated accordingly. The described architecture is the one used currently by most investigators. Although over the years many different modulation scheme (pulse-width-modulation, pulse-position-modulation, pulse-amplitude-modulation, etc.) and system architectures have been tried; due to the proliferation of inexpensive integrated low power AD converters, the pulse-code-modulation (PCM) using an integrated AD is the dominant method these days.

The transmission of the digitized signal can be accomplished using any of the several digital modulation schemes (PAM, PFM, QPSK, etc.), which offer standard trade offs between transmitter and receiver circuit complexity, power consumption, and signal/noise ratio (35). Typical frequencies used in such systems are in the lower UHF range (100–500 MHz). Higher frequencies result in smaller transmitter antenna at the expense of increased tissue loss. Although tissue loss is a major concern in transmitting power to implantable microsystems, it is less of an issue in data transmission, since a sensitive receiver outside the body can easily demodulate the signal. Recent advances in low power CMOS rf circuit design has resulted in an explosive growth of custom made Application Specific Integrated Circuits (ASIC), and off-the-shelf rf circuits suitable for a variety of biotelemetry applications (36–38). In addition, explosive proliferation of wireless communication systems (cell phones, wireless PDAs, Wi-Fi systems, etc.) have provided a unique opportunity to piggyback major WLAN manufacturers and simplify the design of biotelemetry microdevices (39,40). This cannot only increase the performance of the system, but also creates a standard platform for many diverse applications. Although the commercially available wireless chips have large bandwidths and some superb functionality, their power consumption is higher than what is acceptable for many of the implantable microsystems. This, however, is going to change in the future by the aggressive move

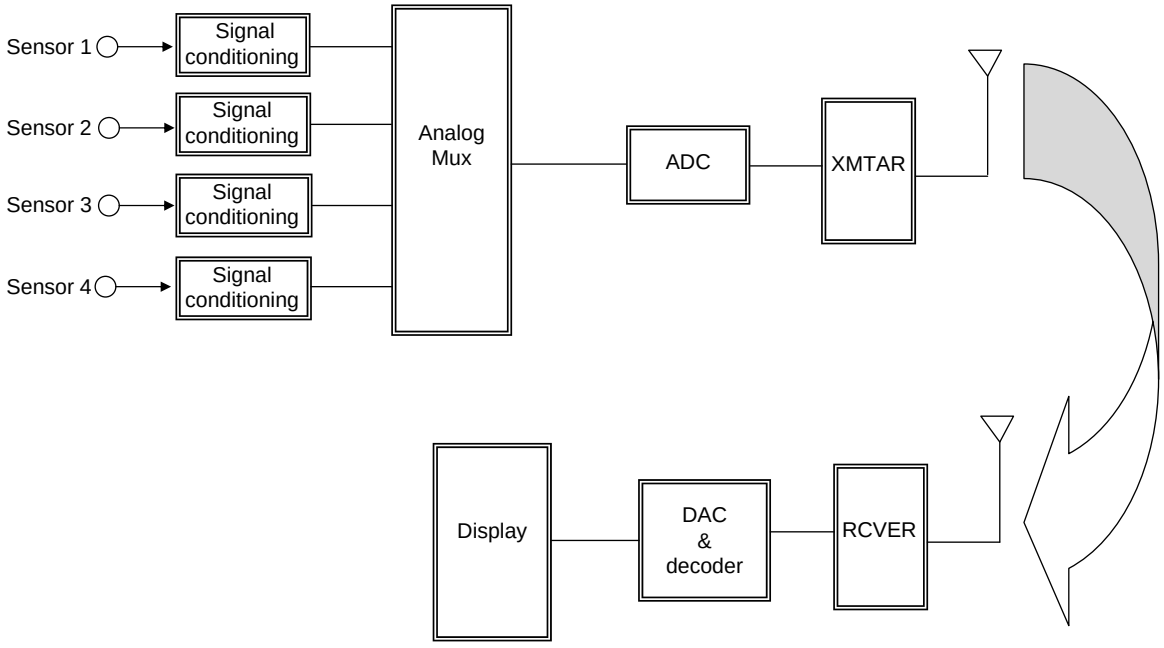


Figure 3. Block diagram of a multichannel biotelemetry system.

toward lower power handheld consumer electronics. A particularly attractive WLAN system suitable for biotelemetry is the Bluetooth system (41). This system, which was initially designed for wireless connection of multiple systems (computer, fax, printer, PDA, etc.) located in close proximity, has been adopted by many medical device manufacturer for their various biotelemetry applications. The advantage of Bluetooth compared to other Wi-Fi system, such as 902.11-b, is its lower power consumption at the expense of a smaller data rate (2.4 GHz carrier frequency, 1 Mbps data rate, and 10 m transmission range). This is not critical in most biotelemetry applications since the frequency bandwidth of most physiologically important signals are low (< 1 kHz). However, note that since the Bluetooth carrier frequency is rather high (2.4 GHz), the systems using Bluetooth or similar WLAN devices can not operate from inside the body and has to be worn by the subject on the outside.

Inductively powered telemetry systems differ from the battery operated ones in several important ways (42). First and foremost, the system has to be powered by an rf signal from outside; this puts several restrictions on frequency and physical range of operation. For implantable systems, the incoming signal frequency has to remain low in order for it to allow enough power to be coupled to the device (this means a frequency range of 1–10 MHz, see next section). In addition, if the device is small, due to a low coupling coefficient between the transmitter and receiver coil, the transmission range is usually limited to distances < 10 cm. Finally, in inductively powered systems, one has to devise a method to transmit the measured signal back to the outside unit. This can be done in several different ways with the load-modulation being the most popular method (43). In "load modulation", the outgoing digital stream of data is used to load the receiver antenna by switching a resistor in parallel with the tank circuit. This can be picked up through the transmitter coil located outside the body. A second technique that is more complex requires an on-chip transmitter and a second coil to transmit the recorded data at a different frequency. The inward link can be easily implemented using amplitude modulation, that is, the incoming rf signal that powers the microsystem is modulated by digitally varying the amplitude. It is evident that the modulation index cannot be 100% since that would cut off the power supply to the device (unless a storage capacitor is used). The coding scheme is based on the pulse time duration, that is,"1" and "0" have the same amplitude, but different durations (42). This modulation technique requires a simple detection circuitry (envelope detector) and is immune to amplitude variations, which are inevitable in such systems.

In addition to the above mentioned differences between the battery operated and inductively powered biotelemetry systems, the implanted circuit in the latter case also includes several modules that are unique and require special attention. These have mostly to do with power reception (rectifier and voltage regulator), clock extraction, and data demodulation. Figure 4 shows a block diagram of the receiver circuit for an inductively powered microsystem currently being developed in the author's laboratory for the measurement of intraocular pressure in glaucoma patients. It consists of a full-bridge rectifier, a voltage regulator, a piezoresistive pressure sensor, and voltage to frequency converter. The incoming rf signal is first rectified and used to generate a stable voltage reference being used by the rest of the circuit (amplifiers, filters, etc.). The clock is extracted from the incoming rf signal and is used wherever it is needed in the receiver circuit. The pressure sensor bridge voltage is first amplified and converted to a stream of pulses having a frequency proportional to the pressure. This signal is then used to load-modulate the tank circuit. The receiver circuitry for most of the reported inductively powered biotelemetry systems were fabricated through CMOS foundries, such as MOSIS. This is due to the fact that one can simply design a single chip performing all of the mentioned functions in a CMOS technology, and hence save valuable space. In the sections dealing with various applications, we will describe several other inductively powered telemetry systems.

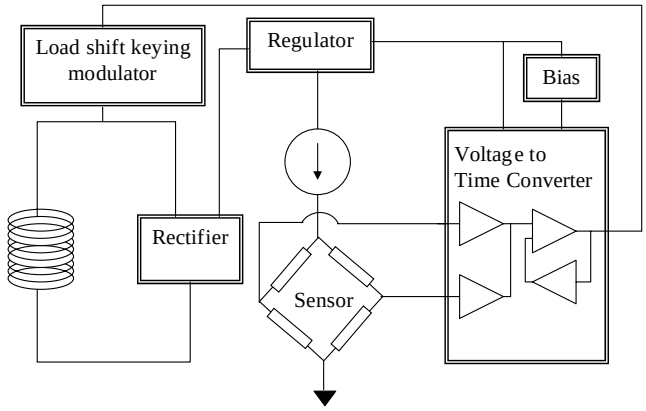


Figure 4. Block diagram of an implantable biotelemetry system used in the measurement of intraocular pressure in glaucoma patients.

There has not been much effort in the area of antenna design for biotelemetry applications. This is due to the basic fact that these systems are small and operate at low frequencies, hence, most antennas employed in such systems belong to the"small antenna" category, that is, the antenna size is much smaller than the wavelength. In such cases it is difficult to optimize the design and most investigators simply use a short electrical or magnetic dipole. For example, in many situations the inductor in the output stage can be used to transmit the information. Or alternatively, a short wire can be used in the transmitter as an electrical dipole. These antennas are usually low gain and have an omnidirectional pattern (44). Systems operating at higher frequencies, such as externally worn Wi-Fi modules, however, can benefit from an optimized design.

In addition to using an rf signal to transmit information that constitutes the majority of the work in the biotelemetry area, the use of ultrasound and infrared (IR) have also been explored by some investigators (45,46). The use of ultrasound is attractive in telemetering physiological information from aquatic animals and divers. This is due to the fact that rf signals are strongly absorbed by seawater while ultrasound is not affected to the same extent. The use of IR is also limited to some specific areas, such as systems that can be worn by the animal on the outside and are not

impeded by solid obstructions. This is due to the inability of IR to negotiate solid opaque objects (line of sight propagation) and its severe absorption by tissue. The advantage of free space IR transmission lies in its very wide bandwidth making it useful for transmitting neural signals. The rf, ultrasonic, and IR systems share many of the system components discussed so far, with the major difference between them having to do with the design and implementation of the output stage. The output transmitter for the ultrasonic biotelemetry systems is usually an ultrasonic transducer, such as PZT or PVDF, whereas for the IR systems it is usually a simple light-emitting diode (LED). The driver circuitry has to be able to accommodate the transducers, that is, a high voltage source for driving the ultrasonic element and a current.

Power Source

The choice of power source for implantable wireless microsystems depends on several factors, such as implant lifetime, system power consumption, temporal mode of operation (continuous or intermittent), and size. Progress in battery technology is incremental and usually several generations behind other electronic components (47). Although lithium batteries have been used in pacemakers for several years, they are usually large for microsystem applications. Other batteries used in hearing aids and calculators are smaller, but have limited capacity and can only be used for low power systems requiring limited lifespan or intermittent operation. Inductive powering is an attractive alternative for systems with large power requirements (e.g., neuromuscular stimulators) or long lifetime (e.g., prosthetic systems with >5 years lifetime) (14,15). In such systems, a transmitter coil is used to power a microchip using magnetic coupling. The choice of the transmission frequency is a trade-off between adequate miniaturization and tissue loss. For implantable microsystems, the frequency range of 1–10 MHz is usually considered optimum for providing adequate miniaturization while still staying below the high tissue absorption region ($>$10 MHz) (48). Although the link analysis and optimization methods have been around for many years (49), recent integration techniques that allow the fabrication of microcoils on top of CMOS receiver chip has allowed a new level of miniaturization (50). For applications that require the patient to carry the transmitter around, a high efficiency transmitter is needed in order to increase the battery lifetime. This is particularly critical in implantable microsystem, where the magnetic coupling between the transmitter and the receiver is low ($<1\%$). Class-E power amplifier/transmitters are popular among microsystem designers due to their high efficiency ($>80\%$) and relatively easy design and construction (51,52). They can also be easily amplitude modulated through supply switching.

Although ideally one would like to be able to tap into the chemical reservoir (i.e., glucose) available in the body to generate enough power for implantable microsystems (glucose-based fuel cell), difficulty in packaging and low efficiencies associated with such fuel cells have prevented their practical application (53). Thin-film batteries are also attractive, however, there still remain numerous material and integration difficulties that need to be resolved (54). Another alternative is nuclear batteries. Although they have been around for several decades and were used in some early pacemakers, safety and regulatory concerns forced medical device companies to abandon their efforts in this area. There has been a recent surge of interest in microsystem nuclear batteries for military applications (55). It is not hard to envision that due to the continuous decrease in chip power consumption and improve in batch scale MEMS packaging technology, one might be able to hermetically seal a small amount of radioactive source in order to power an implantable microsystem for a long period of time. Another possible power source is the mechanical movements associated with various organs. Several proposals dealing with parasitic power generation through tapping into this energy source have been suggested in the past few years (56). Although one can generate adequate power from activities, such as walking, to power an external electronic device, difficulty in efficient mechanical coupling to internal organ movements make an implantable device hard to design and utilize.

Packaging and Encapsulation

Proper packaging and encapsulation of biotelemetry microsystems is a challenging design aspect particularly if the device has to be implanted for a considerable period. The package must accomplish two tasks simultaneously: (*1*) protect the electronics from the harsh body environment while providing access windows for transducers to interact with the desired measurand, and (*2*) protect the body from possible hazardous material in the microsystem. The second task is easier to fulfill since there is a cornucopia of various biocompatible materials available to the implant designer (57). For example, silicon and glass, which are the material of choice in many MEMS applications, are both biocompatible. In addition, polydimethylsiloxane (PDMS) and several other polymers (e.g., polyimide, polycarbonate, parylene) commonly used in microsystem design are also accepted by the body. The first requirement is, however, more challenging. The degree of protection required for implantable microsystems depends on the required lifetime of the device. For short durations (several months), polymeric encapsulants might be adequate if one can conformally deposit them over the substrates (e.g., plasma deposited parylene) (58). These techniques are considered non-hermetic and have a limited lifetime. For long-term operation, hermetic sealing techniques are required (59). Although pacemaker and defibrillator industries have been very successful in sealing their systems in tight titanium enclosures; these techniques are not suitable for microsystem applications. For example a metallic enclosure prevents the transmission of power and data to the microsystem. In addition, these sealing methods are serial in nature (e.g., laser or electron beam welding) and are not compatible with integrated batch fabrication methods used in microsystem design. Silicon–glass electrostatic and silicon–silicon fusion bonding are attractive methods for packaging implantable microsystems (60). Both of these bonding methods are hermetic and can be performed at the wafer level. These are particularly attractive for

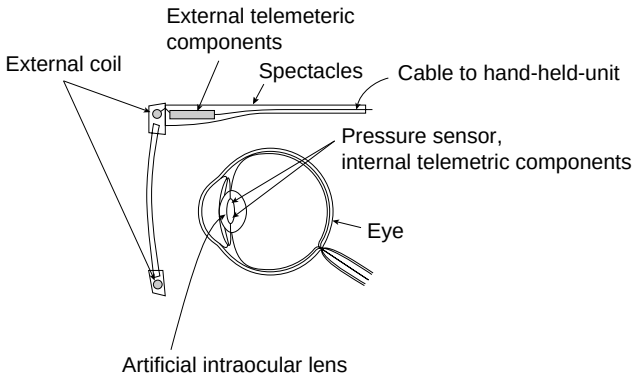


Figure 5. Schematic of the IOP measurement microsystem (61).

inductively powered wireless microsystems since most batteries cannot tolerate the high temperatures required in such substrate bondings. Other methods, such as metal electroplating, have also been used to seal integrated MEMS microsystems. However, their long-term performance is usually inferior to the anodic and fusion bondings. In addition to providing a hermetic seal, the package must allow feeedthrough for transducers located outside the package (18). In macrodevices, such as pacemakers, where the feedthrough lines are large and not too many, traditional methods, such as glass–metal or ceramic–metal has been employed for many years. In microsystems, such methods are not applicable and batch scale techniques must be adopted.

DIAGNOSTIC APPLICATIONS

Diagnostic biotelemetry microsystems are used to gather physiological or histological information from within the body in order to identify pathology. Two recent examples are discussed in this category. The first is a microsystem designed to be implanted in the eye and to measure the intraocular pressure in order to diagnose low tension glaucoma. The second system, although not strictly implanted, is an endoscopic wireless camera-pill designed to be swallowed in order to capture images from the digestive track.

Figure 5 shows the schematic diagram of the intraocular pressure (IOP) measurement microsystem (61,62). This device is used to monitor the IOP in patients suffering from low tension glaucoma, that is, the pressure measured in the doctor's office is not elevated (normal IOP is ~10–20 mmHg, 1.33–2.66 kPa) while the patient is showing optic nerve degeneration associated with glaucoma. There is great interest in measuring the IOP in such patients during their normal course of daily activity (exercising, sleeping, etc). This can only be achieved using a wireless microsystem. The system shown in Fig. 5 consists of an external transmitter mounted on a spectacle, which is used to power a microchip implanted in the eye. A surface micromachined capacitive pressure sensor integrated with CMOS interface circuit is connected to the receiving antenna. The receiver chip implemented in an n-well 1.2 μm CMOS technology has overall dimensions of 2.5×2.5 mm^2 and consumes 210 μW (Fig. 6). The receiver polyimide-based antenna is, however, much

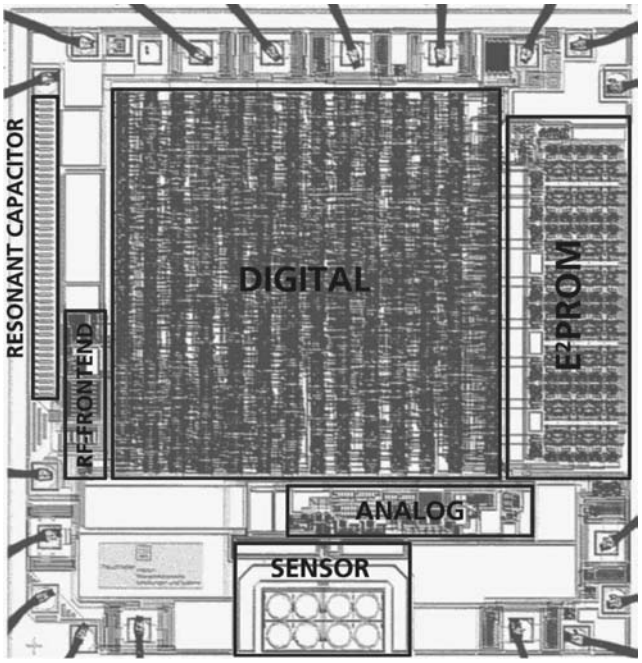


Figure 6. Micrograph of the IOP measurement microsystem receiver chip showing surface micromachined capacitive pressure sensors and other parts of the receiver circuitry (62).

larger (1 cm in diameter and connected to the receiver using flip chip bonding) requiring the device to be implanted along with an artificial lens. The incoming signal frequency is 6.78 MHz, while the IOP is transmitted at 13.56 MHz using load-modulation scheme. This example illustrates the levels of integration that can be achieved using low power CMOS technology, surface micromachining, and flip chip bonding.

The second example in the category of diagnostic microsystems is an endoscopic wireless pill shown in Fig. 7 (63,64). This pill is used to image small intestine, which is a particularly hard area to reach using current fiber optic technology. Although these days colonoscopy and gastroscopy are routinely performed, they cannot reach the small intestine and many disorders (e.g., frequent bleeding) in this organ have eluded direct examination. A wireless endoscopic pill cannot only image the small intestine, but also will reduce the pain and discomfort associated with regular gastrointestinal endoscopies. The endoscopic pill is a perfect example of what can be called *Reemerging Technology*, that is, the rebirth of an older technology based on new capabilities offered by advances in modern technology. Although the idea of a video pill is not new, before the development of low power microelectronics, white LED, CMOS image sensor, and wide-band wireless communication, fabrication of such a device was not feasible. The video pill currently marketed by Given Imaging Inc. is 11 mm in diameter and 30 mm in length (size of a large vitamin tablet) and incorporates: (*1*) a short focal length lens, (*2*) a CMOS image sensor (90,000 pixel), (*3*) four white LEDs, (*4*) a low power ASIC transmitter, and (*5*) two batteries (enough to allows the pill to go though the entire digestive track). The pill can capture and transmit

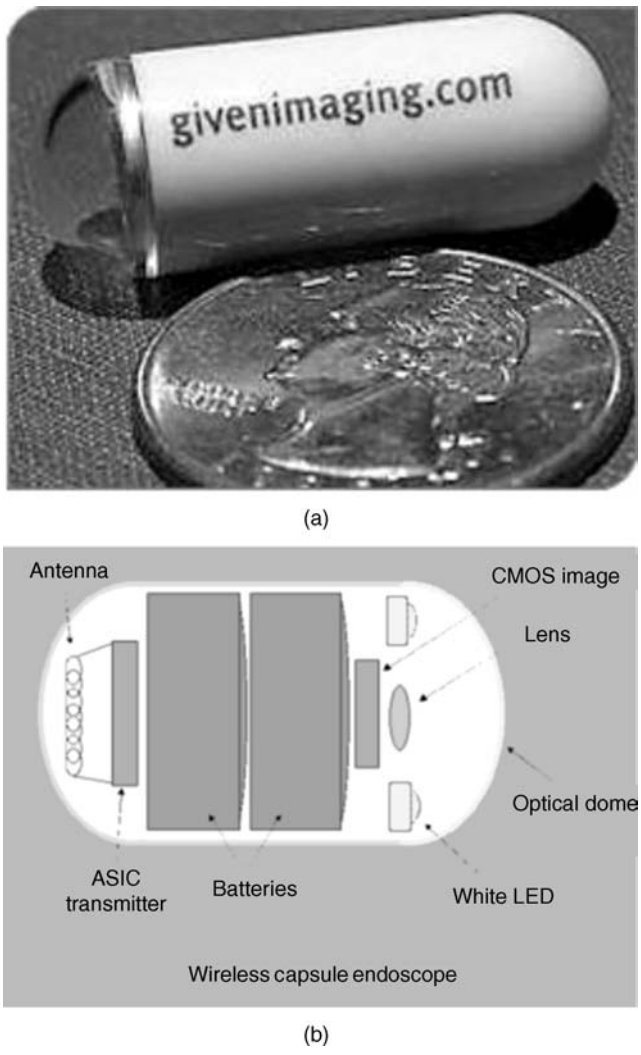


Figure 7. A photograph (a) and internal block diagram (b) of Given Imaging wireless endoscopic pill. (Courtesy Given Imaging.)

two images per second to an outside receiver capable of storing up to 5 h of data.

THERAPEUTIC APPLICATIONS

Therapeutic biotelemetry microsystems are designed to alleviate certain symptoms and help in the treatment of a disease. In this category, two such biotelemetry microsystems unit be described. The first is a drug delivery microchip designed to administer small quantities of potent drugs upon receiving a command signal from the outside. The second device is a passive micromachined glucose transponder, which can be used to remotely monitor glucose fluctuations allowing a tighter blood glucose control through frequent measurements and on-demand insulin delivery (pump therapy or multiple injections).

Figure 8 shows the central component of the drug delivery microchip (65,66). It consists of several microreservoirs (25 nL in volume) etched in a silicon substrate. Each microreservoir contains the targeted drug and is covered by a thin gold membrane (0.3 μm), which can be dissolved through the application of a small voltage (1 V vs. Saturated Calomel Electrode). The company marketing this technology (MicroCHIPS Inc.) is in the process of designing a wireless transceiver that can be used to address individual wells and release the drug upon the reception of the appropriate signal (67). Another company (ChipRx Inc.) is also aiming to develop a similar microsystem (Smart Pill) (68). Their release approach, however, is different and is based on conductive polymer actuators acting similar to a sphincter, opening and closing a tiny reservoir. Due to the potency of many drugs, safety and regulatory issues are more stringent in implantable drug delivery microsystems and will undoubtedly delay their appearance in the clinical settings.

Figure 9 shows the basic concept behind the glucose-sensitive microtransponder (69). A miniature MEMS-based microdevice is implanted in the subcutaneous tissue and an interrogating unit remotely measures the glucose levels without any hardwire connection. The microtrasponder is a passive LC resonator, which is coupled to a glucose-sensitive hydrogel. The glucose-dependent swelling and deswelling of the hydrogel is coupled to the resonator causing a change the capacitor value. This change translates into variations of the resonant frequency, which can be detected by the interrogating unit. Figure 10 shows the schematic drawing of the microtransponder with a capacitive sensing mechanism. The glucose sensitive hydrogel is mechanically coupled to a glass membrane and is separated from body fluids (in this case interstitial fluid) by a porous stiff plate. The porous plate allows the unhindered flow of water and glucose while blocking the hydrogel from escaping the cavity. A change in the glucose concentration of the external environment will cause a swelling or deswelling of the hydrogel, which will deflect the glass membrane and change the capacitance. The coil is totally embedded inside the silicon and can achieve a high quality factor and hence increased sensitivity by utilizing the whole wafer thickness (reducing the series resistance). The coil-embedded silicon and the glass substrate are hermetically sealed using glass–silicon anodic bonding.

REHABILITATIVE MICROSYSTEMS

Rehabilitative biotelemetry microsystems are used to substitute a lost function, such as vision, hearing, or motor activity. In this category, two microsystems are described. The first one is a single-channel neuromuscular microstimulator used to stimulate paralyzed muscle groups in paraplegic and quadriplegic patients. The second microsystem is a visual prosthetic device designed to stimulate ganglion cells in retina in order to restore vision to people afflicted with macular degeneration or retinitis pigmentosa.

Figure 11 shows a schematic of the single channel microstimulator (13). This device is $10 \times 2 \times 2$ mm^3 in dimensions and receives power and data through an inductively coupled link. It can be used to stimulate paralyzed muscle groups using thin-film microfabricated electrodes located at the ends of a silicon substrate. A hybrid capacitor

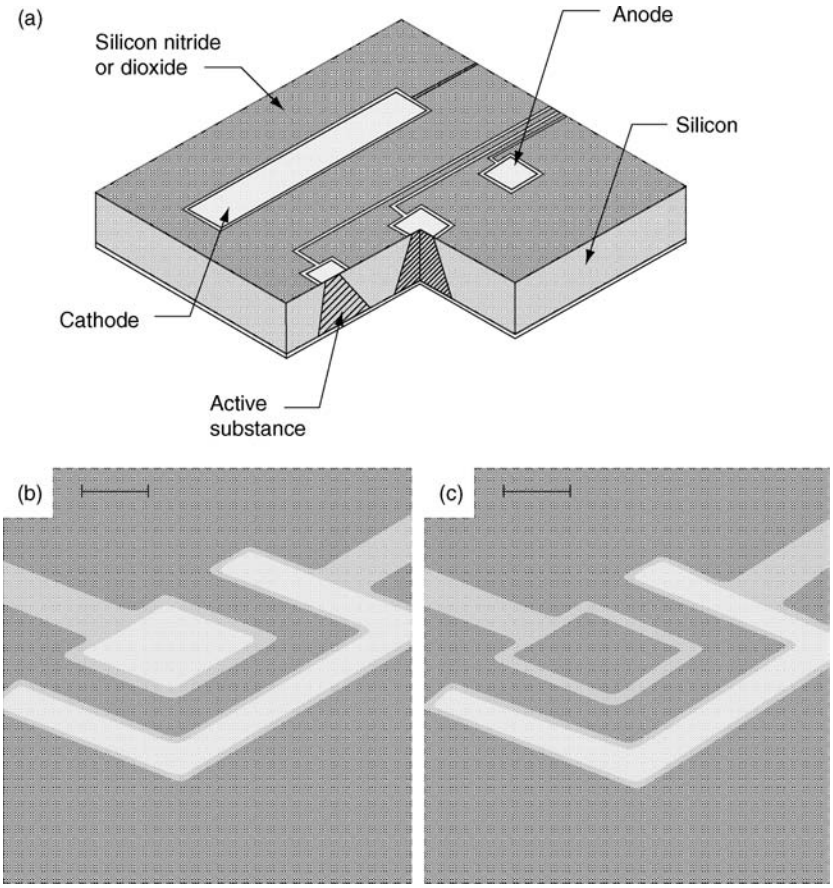


Figure 8. MicroCHIP drug delivery chip (a), a reservoir before and after dissolution of the gold membrane (b,c), the bar is 50 μm (65).

is used to store the charge in between the stimulation pulses and to deliver 10 mA of current to the muscle every 25 ms. A glass capsule hermetically seals a BiCMOS receiver circuitry along with various other passive components (receiver coil and charge storage capacitor) located on top of the silicon substrate. Figure 12 shows a photograph of the microstimulator in the bore of a gauge 10 hypodermic needle. As can be seen, the device requires a complicated hybrid assembly process in order to attach a wire-wound coil and a charge storage capacitor to the

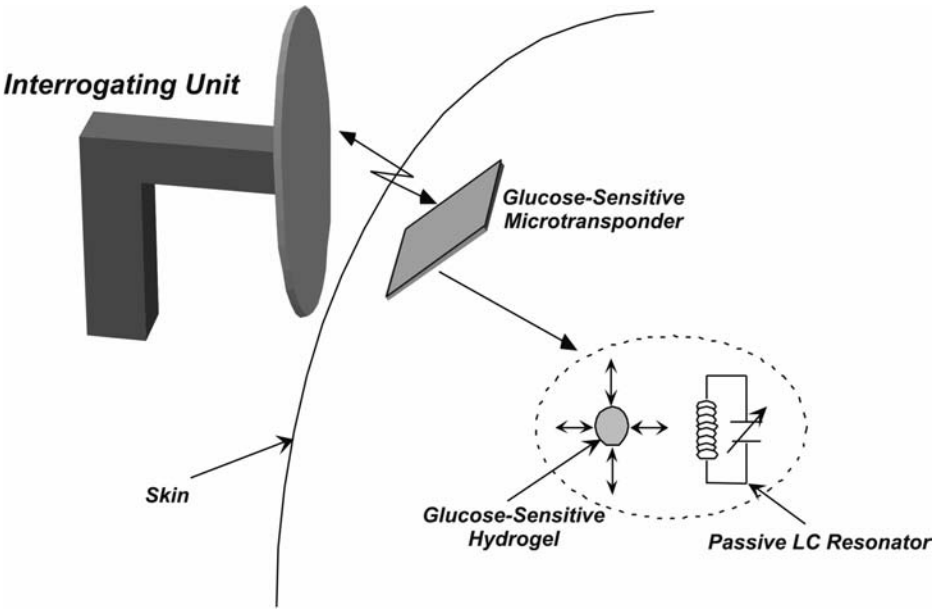


Figure 9. Basic concept behind the glucose-sensitive microtransponder.

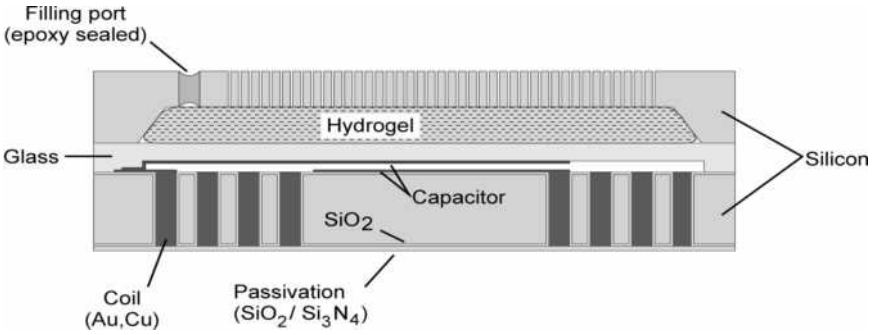


Figure 10. Cross-section of glucose microtransponder.

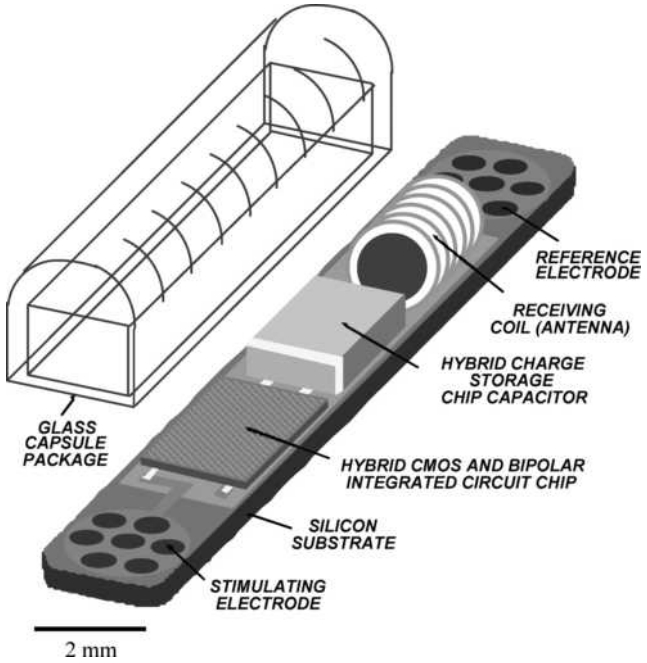


Figure 11. Schematic of a single-channel implantable neuromuscular microstimulator.

Figure 12. Photograph of the microstimulator in the bore of a gage 10 hypodermic needle.

receiver chip. In a subsequent design targeted for direct peripheral nerve stimulation (requiring smaller stimulation current), the coil was integrated on top of the BiCMOS electronics and on-chip charge storage capacitors were used thus considerably simplifying the packaging process. Figure 13 shows a micrograph of the chip with the electroplated copper inductor (70). A similar microdevice (i.e., a single channel microstimulator) was also developed by another group of investigators with the differences mainly related to the packaging technique (laser welding of a glass capsule instead of silicon–glass anodic bonding), chip technology (CMOS instead of BiCMOS), and electrode material (tantalum and iridium instead of iridium oxide) (42). Figure 14 shows a photograph of the microstimulator developed by Troyk, Loeb, and their colleagues.

Figure 15 shows the schematic of the visual prosthetic microsystem (71,72). A spectacle mounted camera is used to capture the visual information followed by digital conversion and transmission of data to a receiver chip

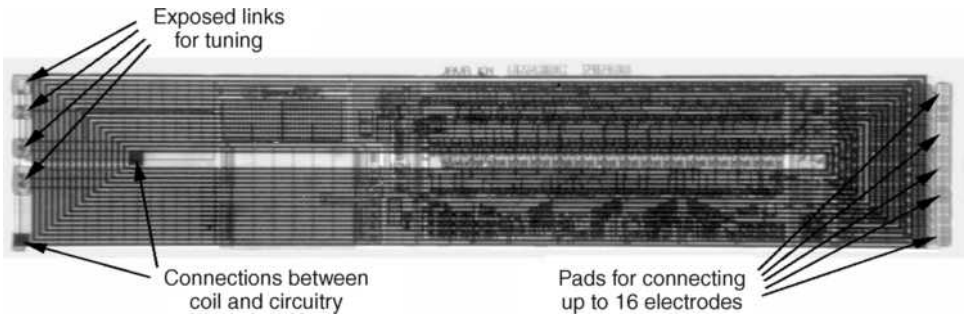


Figure 13. Microstimulator chip with integrated receiver coil and on-chip storage capacitor (70).

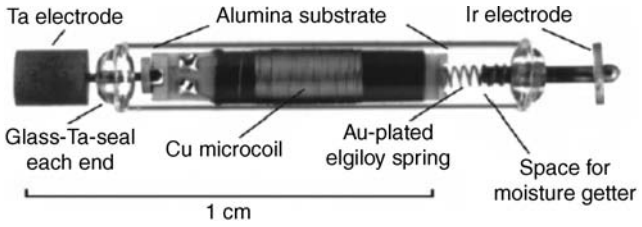


Figure 14. Photograph of a single channel microstimulator developed by Troyk (42).

implanted in the eye. The receiver uses this information to stimulate the ganglion cells in the retina through a microelectrode array in sub or epi-retinal location. This microsystem is designed for patients suffering from macular degeneration or retinitis pigmentosa. In both diseases, the light sensitive retinal cells (cones and rods) are destroyed while the more superficial retinal cells, that is, ganglion cells, are still viable and can be stimulated. Considering that macular degeneration is an age related pathology and will be afflicting more and more people as the average age of the population increases, such a microsystem will be of immense value in the coming decades. There are several groups pursuing such a device with different approaches to electrode placement (epi-or subretinal), chip design, and packaging. A German consortium that has also designed the IOP measurement microsystem is using a similar approach in antenna placement (receiver antenna in the lens), chip design, and packaging technology to implement a retinal prosthesis (61). Figure 16 shows photographs of the retinal stimulator receiver chip, stimulating electrodes, and polyimide antenna. The effort in the United States is moving along a similar approach (72,72).

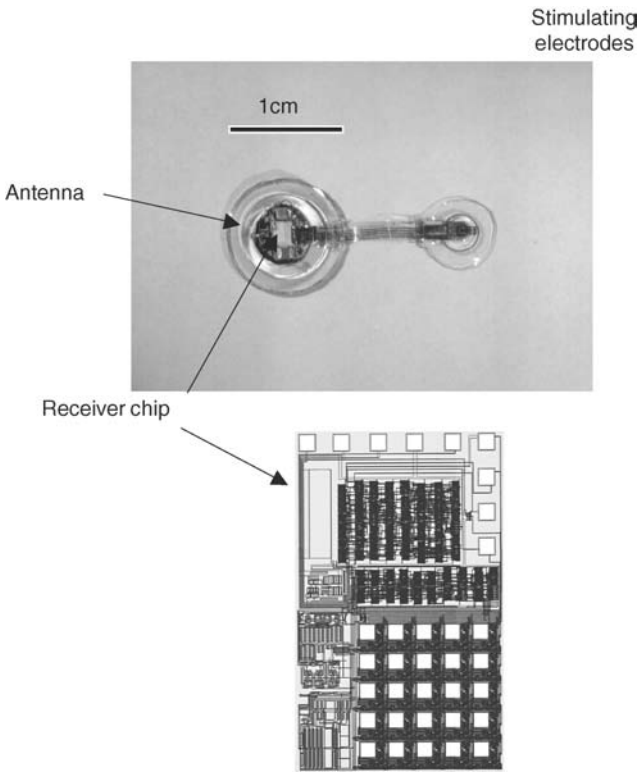


Figure 16. Retinal stimulator receiver chip, stimulating electrodes, and polyimide antenna (61). Chip size.

CONCLUSIONS

In this article, several biotelemetry microsystems currently being developed in the academia and industry were reviewed. Recent advances in MEMS-based transducers, low power CMOS integrated circuit, wireless communication transceivers, and advanced batch scale packaging have provided a unique opportunity to develop implantable biotelemetry microsystems with advanced functionalities not achievable previously. These systems will be indispensable to the twenty-first century physician by providing assistance in diagnosis and treatment. Future research and development will probably be focused on three areas: (*1*) nanotransducers, (*2*) self-assembly, and (*3*) advanced biomaterials. Although MEMS-based sensors and actuators have been successful in certain areas (particularly physical sensors), their performance could be further improved by utilizing nanoscale fabrication technology. This is particularly true in the area of chemical sensors where future diagnostic depends on detecting very small amounts of chemicals (usually biomarkers) well in advance of any

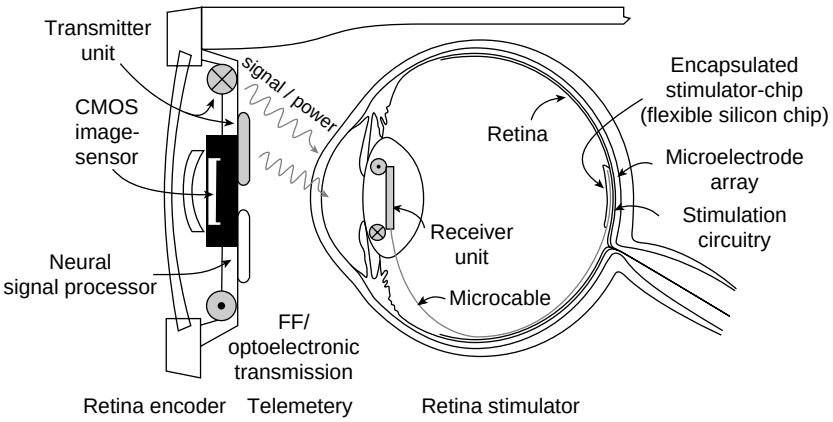


Figure 15. Schematic of a visual prosthetic microsystem (61).

physical sign. Nanosensors capable of high sensitivity chemical detection will be part of the future biotelemetry systems. In the actuator–delivery area, drug delivery via nanoparticles is a burgeoning area that will undoubtedly be incorporated into future therapeutic microsystems. Future packaging technology will probably incorporate self-assembly techniques currently being pursued by many micro–nanoresearch groups. This will be particularly important in microsystems incorporating multitude of nanosensors. Finally, advanced nanobased biomaterials will be used in implantable microsystems in order to enhance biocompatibility and prevent biofouling. These will include biocompatible surface engineering and interactive interface design (e.g., surfaces that release anti-inflammatory drugs in order to reduce postimplant fibrous capsule formation).

BIBLIOGRAPHY

Cited References

1. MacKay RS, Jacobson B. Endoradiosonde. Nature (London) 1957;179:1239–1240.
2. MacKay RS. Biomedical telemetry: The Formative Years. IEEE Eng Med Biol Mag 1983;2:11–17.
3. Knutti JW, Allen HV, Meindl JD. Integrated Circuit Implantable telemetry Systems. IEEE Eng Med Biol Mag 1983;2:47–50.
4. Ko WH, Neuman MR. Implant Biotelemetry and Microelectronics. Science 1867;156:351–360.
5. Topich JA. Medical Telemetry. CRC Handbook of Engineering in Medicine and Biology; 1976, p 41–75.
6. Jeutter DC. Biomedical Telemetry Techniques. CRC Crit Rev Biomed Eng 1982;11:121–174.
7. Meindl JD, et al. Implantable Telemetry. Methods Animal Exper 1986;3:37–111.
8. Kimmich HP. Biotelemetry. Encyclopedia of Medical Devices In: Webster JG, editor. 1988, p 409–425.
9. Santic A. Biomedical Telemetry. Wiley Encyclopedia of Electrical and Electronics Engineering, In: Webster JG, editor. 1999. p 438–454.
10. MacKay RS. Biomedical Telemetry, Sensing and Transmitting Biological Information from Animals and Man, 2nd ed. Piscataway (NJ): IEEE Press; 1993.
11. Wise KD. Special Issue on Sensors, Actuators, and Microsystems Proc. IEEE, 1998;86.
12. Cameron T, et al. Micromodular Implants to Provide Electrical Stimulation of Paralyzed Muscles and Limbs. IEEE Trans. Biomed. Eng. 1997;44:781–790.
13. Ziaie B, Nardin M, Coghlan AR, Najafi K. A Single Channel Microstimulator for Functional Neuromuscular Stimulation. IEEE Trans Biomed Eng 1997;44:909–920.
14. Hamici Z, Itti R, Champier J. A High-Efficiency Power and Data Transmission System for Biomedical Implanted Electronic Devices. Measurement Sci Technol. 1996;7:192–201.
15. Heetderks WJ. RF Powering of Millimeter- and Submillimeter-Sized Neural Prosthetic Implants. IEEE Trans Biomed Eng 1988;35:323–327.
16. Gray PR, Meyer RG. Future Directions in Silicon ICs for RF Personal Communications. Proceedings of the Custom Integrated Circuits Conference. 1995, p 83–89.
17. Abidi AA. RF CMOS Come of Age. IEEE Microwave Mag 2003;4:47–60.
18. Ziaie B, Von Arx JA, Dokmeci MR, Najafi K. A Hermetic Glass-Silicon Micropackage with High-Density on-chip Feedthroughs for Sensors and Actuators. IEEE J Microelectromech Systems, 1996;5:166–179.
19. Gopel W, Hesse J, Zemel JN. Sensors: A Comprehensive Survey, Vols. 1–8, New York: VCH Publishers; 1989.
20. Hohler JM, Sautz HP, editor. Microsystem Technology: A Powerful Tool for Biomolecular Studies. Boston: Birkhauser; 1999.
21. Taylor RF, Schultz JS. Handbook of Chemical and Biological Sensors, Boston: IOP Press; 1996.
22. Rogers EK. editor, Handbook of Biosensors and Electronic Nose, Boca Raton, (FL): CRC Press; 1997.
23. Hak GA. editor, The MEMS Handbook. Boca Raton (FL): CRC Press; 2001.
24. Webster JG. editor, The Measurement Instrumentation and Sensors Handbook, Boca Raton (FL): CRC Press; 1998.
25. Zhang M, Desai T, Ferrari M. Proteins and Cells on PEG Immobilized Silicon Surfaces. Biomaterials 1998;19:953–960.
26. Branch DW, Wheeler BC, Brewer GJ, Leckband DE. Long-Term Stability of Grafted Polyethylene Glycol Surfaces for use with Microstamped Substrates in Neuronal Cell Culture. Biomaterials, 2001;22:1035–1047.
27. Alcantar NA, Aydil ES, Israelachvili JN. Polyethylene Glycol-Coated Biocompatible Surfaces. J Biomed Mat Res 2000;51:343–351.
28. Baltes H, Paul O, Brand O. Micromachined Thermally Based CMOS Microsensors. Proc IEEE 1998;86:1660–1678.
29. Stotts LJ. Introduction to Implantable Biomedical IC Design. IEEE Circuits Devices Mag 1989;5:12–18.
30. Stouraitis T, Paliouras V. Considering the Alternatives in Low-Power Design. IEEE Circuits Devices Mag 2001;17:22–29.
31. Tsividis Y, Krishnapura N, Palakas Y, Toth L. Internally Varying Analog Circuits Minimize Power Dissipation. IEEE Circuits Devices Mag 2003;19:63–72.
32. Benini L, De Micheli G, Macii E. Designing Low-power Circuits: Practical Recipes. IEEE Circuits Systems Mag 2001;1:6–25.
33. Rajput SS, Jamuar SS. Low Voltage Analog Circuit Design Techniques. IEEE Circuits Systems Mag 2002;2:24–42.
34. Mohseni P, et al. An Ultra-Light Biotelemetry Backpack for Recording EMG Signals in Moths. IEEE Trans Biomed Eng. June 2001;48:734–737.
35. Proakis JG, Salehi M. Communication System Engineering. Pearson Education; 2001.
36. Lee TH. Design of CMOS Radiofrequency Integrated Circuits. Cambridge: Cambridge University Press, 1998.
37. Razavi B. Challenges in Portable RF Transceiver Design. IEEE Circuits Devices Mag 1996;12:12–25.
38. Larson LE. Integrated Circuit Technology Options for RFICs-Present Status and Future Directions. IEEE J Solid-State Circuits 1998;33:387–399.
39. Crow BP, Wudiaja I, Kim LG, Saki PT. IEEE 802.11 Wireless Local Area Networks. IEEE Commun Mag 1997;35:116–126.
40. Chatschik B. An Overview of the Bluetooth Wireless technology. IEEE Commun Mag 2001;39:86–94.
41. Saltzstein WE. Bluetooth and Beyond: Wireless Options for Medical Devices. Med Device Diagnostic Ind June 2004.
42. Troyk P. Injectable Electronic Identification, Monitoring, and Stimulation Systems. Ann Rev Biomed Eng 1999;1:177–209.
43. Finkenzeller K. RFID Handbook, New York: John Wiley & Sons, Inc; 2003.
44. Kraus JD. Antenna. New York: McGraw-Hill; 2001.
45. Woodward B, Istepanian RSH.Acoustic Biotelemetry of Data from Divers, Proc 15th Annu Int IEEE Eng Med Biol Soc Conf Paris 1992;1000–1001.
46. Kawahito S, et al. A CMOS Integrated Circuit for Multichannel Multiple-Subject Biotelemetry using Bidirectional Optical Transmissions. IEEE Trans Biomed Eng 1994;41: 400–406.

47. Linden D, Reddy T. Handbook of Batteries. New York: McGraw-Hill; 2001.
48. Foster KR, Schwan HP. Handbook of Biological Effects of Electromagnetic Fields, In: Polk C, Postow E, editor. Boca Raton (FL): CRC Press; 1996.
49. Ko WH, Liang SP, Fung CDF. Design of Radio-Frequency Powered Coils for Implant Instruments. Med Biol Eng Computing 1977;15:634–640.
50. Ashby KB, et al. High Q Inductors for Wireless Applications in a Complementary Silicon Bipolar Process. IEEE J Solid-State Circuits 1996;31:4–9.
51. Sokal NO, Sokal AD. Class E-A New Class of High-Efficiency Tuned Single-Ended Switching Power Amplifiers. IEEE J. Solid-State Circuits 1975;10:168–176.
52. Ziaie B, Rose SC, Nardin MD, Najafi K. A Self-Oscillating Detuning-Insensitive Class-E Transmitter for Implantable Microsystems. IEEE Trans Biomed Eng 2001;48:397–400.
53. Mehta V, Cooper JS. Review and Analysis of PEM Fuel Cell Design and Manufacturing. J Power Sources 2003;114:32–53.
54. Singh D, et al. Challenges in Making of Thin Films for $Li_xM_nyO_4$ Rechargeable Lithium Batteries for MEMS. J Power Sources 2001;97–98:826–831.
55. Lal A, Blanchard J. Daintiest Dynamos: Nuclear Microbatteries. IEEE Spectrum 2004;42:36–41.
56. Starner T. Human Powered Wearable Computing. IBM J Systems 1996;35:618–629.
57. Ratner BD, Schoen FJ, Hoffman AS, Lemons JE. Biomaterials Science: An Introduction to Materials in Medicine. New York: Elsevier Books; 1997.
58. Loeb GE, Bak MJ, Salcman M, Schmidt EM. Parylene C as a Chronically Stable reproducible Microelectrode material. IEEE Trans Biomed Eng 1977;24:121–128.
59. Nichols MF. The Challenges for Hermetic Encapsulation of Implanted Devices. Critical Rev Biomed Eng 1994;22:39–67.
60. Schmidt MA. Wafer-to-Wafer Bonding for Microstructure Formation. Proc IEEE 1998;86:1575–1585.
61. Mokwa W, Schenakenberg U. Micro-Transponder Systems for Medical Applications. IEEE Trans Instr Meas 2001;50:1551–1555.
62. Stangel K, et al., A Programmable Intraocular CMOS Pressure Sensor System Implant. IEEE J Solid-State Circuits 2001;36:1094–1100.
63. Iddan G, Meron G, Glukhovsky A, Swain P. Wireless Capsule Endoscopy. Nature (London) 2000;405:417.
64. http://www.givenimaging.com.
65. Santini JT, Cima MJ, Langer R. A Controlled-Release Microchip. Nature (London) 1999;397:335–338.
66. Santini JT, et al. Microchips as Controlled Drug-Delivery Devices. Angew Chem 2000;39:2396–2407.
67. Available at http://www.mchips.com.
68. Available at http://www.chiprx.com.
69. Lei M, et al. A Hydrogel-Based Wireless Chemical Sensor. Proc IEEE MEMS 2004;391–394.
70. Von Arx JA, Najafi K. A Wireless Single-Chip Telemetry-Powered Neural Stimulation System. IEEE Solid-State Circuits Conf 1999;15–17.
71. Liu W, et al. Retinal Prosthesis to Aid the Visually Impaired. IEEE Systems, Man, and Cybernetics, Conf 1999;364–369.
72. Humayun MS, et al. Towards a Completely Implantable, Light-Sensitive Intraocular Retinal Prosthesis. Proc 23rd Ann IEEE EMBS Conf 2001;3422–3425.

See also BIOFEEDBACK; BLADDER DYSFUNCTION, NEUROSTIMULATION OF; MONITORING, INTRACRANIAL PRESSURE; NEONATAL MONITORING; PACEMAKERS.

BIRTH CONTROL. See CONTRACEPTIVE DEVICES.

BLEEDING, GASTROINTESTINAL. See GASTROINTESTINAL HEMORRHAGE.

BLADDER DYSFUNCTION, NEUROSTIMULATION OF

MAGDY HASSOUNA
Toronto Western Hospital
NADER ELMAYERGI
MAZEN ABDELHADY
McMaster University

INTRODUCTION

The discovery of electricity introduced enormous changes to human society: Electricity not only improved daily life, but also opened up new opportunities in scientific research. The effects of electrical stimulation on muscular and nervous tissue have been known for several centuries, but the underlying electrophysiological theory to explain these effects was first derived after the development of classical electrodynamics and the development of nerve cell models (1).

Luigi Galvani first suggested that electricity could produce muscular contraction in his animal experiments (2). He found that a device constructed from dissimilar metals, when applied to the nerve or muscle of a frog's leg, would induce muscular contraction. His work formed the foundation for later discoveries of transmembrane potential and electrically mediated nerve impulses. Alessandro Volta, the inventor of the electrical battery (or voltaic pile) (3), was later able to induce a muscle contraction by producing a potential with his battery and conducting it to a muscle strip. The use of Volta's battery for stimulating nerves or muscles became known as galvanic stimulation.

Another basis for modern neural stimulators was the discovery of the connection between electricity and magnetism, demonstrated by Oersted in 1820; he described the effect of current passing through a wire on a magnetized needle. One year later, Faraday showed the converse—that a magnet could exert a force on a current-carrying wire. He continued to investigate magnetic induction by inducing current in a metal wire rotating in a magnetic field. This device was a forerunner of the electric motor and made it possible to build the magneto-electric and the induction coil stimulator. The latter, the first electric generator, was called the Faraday stimulator. Faradic stimulation could produce sustained titanic contractions of muscles, instead of a single muscle twitch as galvanic stimulation had done.

Duchenne used an induction coil stimulator to study the anatomy, physiology, and pathology of human muscles. Finally, he was able to study the functional anatomy of individual muscles (4,5). This work is still valid for the investigation of functional neuromuscular stimulation.

Another basis for modern stimulator devices lay in the work of Chaffee and Light (6). They examined the problem

of stimulating neural structures deep in the body, while avoiding the risk of infection from percutaneous leads: They implanted a secondary coil underneath the skin and placed a primary coil outside the body, using magnetic induction for energy transfer and modulation. Further improvement was achieved by radio frequency (rf) induction (7,8). The Glenn group developed a totally implanted heart pacemaker—one of the first commercially available stimulators. In the ensuing years, stimulators for different organ systems were developed, among them the abovementioned heart pacemaker, a diaphragmatic pacemaker (7,8), and the cochlear implant (9).

BLADDER STIMULATION

Electrical stimulation of the bladder dates back to 1878. The Danish surgeon M.H. Saxtorph treated patients with urinary retention by inserting a special catheter with a metal electrode into the urinary bladder transurethrally and placing a neutral electrode suprapubically (10). Also, Katona et al. (11) described their technique of intraluminal electrotherapy, a method that was initially designed to treat a paralytic gastrointestinal tract, but was later used for neurogenic bladder dysfunction in patients with incomplete central or peripheral nerve lesions (11,12).

Further interest in the electrical control of bladder function began in the 1950s and 1960s. The most pressing question at that time was the appropriate location for stimulation. Several groups attempted to initiate or prevent voiding (in urinary retention and incontinence, respectively) by stimulation of the pelvic floor, the detrusor directly, the spinal cord, or the pelvic and sacral nerves or sacral roots. Even other parts of the body, such as the skin, were stimulated in an attempt to influence bladder function (13).

In 1954, McGuire performed extensive experiments of direct bladder stimulations in dogs (14) with a variety of electrodes, both single and multiple, in a variety of positions. Boyce and associates continued this research (15).

It was realized that with a single pair of electrodes, the maximal response was obtained when the electrodes were placed on both lateral bladder walls so that the points of stimulation encompassed a maximal amount of detrusor muscle. When this was performed in human studies, an induction coil for direct bladder stimulation was implanted in three paraplegic men with complete paralysis of the detrusor muscle. The secondary coil was implanted in the subcutaneous tissue of the lower abdominal wall. Of the three, only one was a success, with the other a failure and the third only partially successful (15).

In 1963, Bradley and associates published their experience with an implantable stimulator (16). They were able to achieve complete bladder evacuation in the chronic dog model over 14 months. However, when the stimulator was implanted in seven patients, detrusor contraction was produced, but bladder evacuation resulted in only two. Further experiments were performed in the sheep, calf, and monkey in an attempt to resolve species discrepancies. These animals were chosen because, in the sheep and calf, the bladder is approximately the same size as in the human, and this similarity could determine whether more power is needed for a bladder larger than that of the dog. In addition, the pelvis of monkeys and humans is similarly deep; thus, the influence (if any) of pelvic structure could be investigated. The results showed that a larger bladder needs more power and wider contact between the electrodes and that differences in structure do not necessitate different stimulation techniques (13,16).

PELVIC FLOOR STIMULATION

In 1963, Caldwell described his clinical experience with the first implantable pelvic floor stimulator (17). The electrodes were placed into the sphincter, with the secondary coil placed subcutaneously near the iliac spine. Though this device was primarily designed for the treatment of fecal incontinence; Caldwell also treated urinary incontinence successfully.

Another approach to pelvic floor stimulation for females is intravaginal electrical stimulation, reported initially by Magnus Fall's group (1977) (18). They published numerous studies dealing with this subject in the ensuing years and found that intravaginal electrical stimulation also induces bladder inhibition in patients with detrusor instability. Lindstram, a member of the same group, demonstrated that bladder inhibition is accomplished by reflexogenic activation of sympathetic hypogastric inhibitory neurons and by central inhibition of pelvic parasympathetic excitatory neurons to the bladder (13,19). The afferent pathways for these effects could be shown to originate from the pudendal nerves.

POSTERIOR TIBIAL OR COMMON PERONEAL

Another interesting application of electrical stimulation for inhibition of detrusor activity is the transcutaneous stimulation of the posterior tibial or common peroneal nerve. This technique, drawn from traditional Chinese medicine, is based on the acupuncture points for inhibition of bladder activity and was reported by McGuire et al. in 1983 (20).

A percutaneous tibial nerve stimulation (PTNS) (Urgent PC, CystoMedix, Anoka, MN) was approved by the Food and Drug Administration in 2000. A needle is inserted $\sim$5 cm cephalad from the medial malleolus and just posterior to the margin of the tibia. Stimulation is done using a self-adhesive surface stimulation electrode without an implanted needle electrode (21). Current data describe results after an initial treatment period of 10–12 weeks. If patients get a good response, they are offered tapered chronic treatment. As in sacral root neuromodulation, PTNS seems less effective for treating chronic pelvic pain (22).

More substantial data, in particular on objective parameters and long-term follow up, are needed, as are studies looking into the underlying neurophysiological mechanisms of this treatment modality. Although minimally invasive, easily applicable, and well tolerated, the main disadvantage of PTNS seems to be the necessity of chronic treatment. The development of an implantable subcutaneous stimulation device might ameliorate this problem (23). It has never found widespread acceptance.

PELVIC NERVE STIMULATION

Pelvic nerves do not tolerate chronic stimulation and the pudendal nerves are activated, increasing outflow resistance. Also, in humans the fibers of the parasympathetic nervous system innervating the bladder split early in the pelvis, forming a broad plexus unsuitable for electrode application (24).

DETRUSOR STIMULATION

Direct detrusor stimulation offers high specificity to the target organ (25), but its disadvantages are electrode displacement and malfunction due to bladder movement during voiding, and fibrosis (even erosion) of the bladder wall. In 1967, Hald et al. (26) reported their experience of direct detrusor stimulation with a radio-linked stimulator in four patients, three with upper motor-neuron lesions and one with a lower motor-neuron lesion. The receiver was placed in a paraumbilical subcutaneous pocket. Two wires from the receiver were passed subcutaneously to the ventral bladder wall, where they were implanted. A small portable external transmitter generated the necessary energy. The procedure worked in three patients; in one it failed because of technical problems (13).

SPINAL CORD STIMULATION

The first attempt to achieve micturition via spinal cord stimulation was through the exploration of the possibility of direct electrical activation of the micturition center in the sacral segments of the conus medullaris. This was conducted by Nashold, Friedman, and associates, and had reported that the region for optimal stimulation was S1–S3.

Effectiveness was determined not only by location, but also by frequency. In two subsequent experiments, the same group compared the stimulation of the dorsal surface of the spinal cord at LS, S1, and S2 with depth stimulation (2–3 mm) at S1 and S2 in acute and chronic settings (27). It was only through the latter, the depth stimulation, that voiding was produced: High bladder pressures were achieved by surface stimulation, but external sphincter relaxation did not occur, and was noted only after direct application of the stimulus to the micturition center in the spinal cord. Stimulation between L5 and S1 produced pressure without voiding, even with depth stimulation (13).

Jonas et al. continued the investigation of direct spinal cord stimulation to achieve voiding (28–30). They compared 12 different types of electrodes: three surface (bipolar surface electrode, dorsal column electrode, and wrap-around electrode) and nine depth electrodes. These differed in many parameters (e.g., bipolar–tripolar, horizontal–vertical–transverse). Regardless of the type of electrode, the detrusor response to stimulation was similar. Interestingly, the wrap-around surface electrode with the most extended current spread provoked the same results as the coaxial depth electrode with the least current spread, prompting those authors to theorize that current does not cross the midline of the spinal cord. Unfortunately, no real voiding was achieved. It was found that the stimulation of the spinal cord motor centers stimulates the urethral smooth and striated sphincteric elements simultaneously: The expected detrusor contraction resulted, but sphincteric contraction was associated. The sphincteric resistance was too high to allow voiding: It allowed only minimal voiding at the end of the stimulation, so-called poststimulus voiding (13). These results contrasted with the earlier work of Nashold and Friedman (27,31).

Thurhoff et al. (32) determined the existence of two nuclei, a parasympathetic and a pudendal nucleus. The parasympathetic nucleus could be shown within the pudendal nucleus; thus, at the level of the spinal cord, stimulation of the bladder separate to that of the sphincter is difficult.

SACRAL ROOT STIMULATION

Based on the hypothesis that different roots would carry different neuronal axons to different locations. The culmination of these studies led to the feasibility of sacral rootlet stimulation.

It appears that sacral nerve-root stimulation is the most attractive method since the space within the spinal column facilitates mechanically stable electrode positioning and the application of electrodes is relatively simple due to the long intraspinal course of the sacral roots.

The University of California, San Francisco (UCSF) group performed numerous experiments on a canine model (33), as the anatomy of bladder innervation of the dog is similar to that of the human. After laminectomy, the spinal roots were explored and stimulated, either intradurally or extradurally, but within the spinal canal, in the following modes:

1. Unilateral stimulation of the intact sacral root at various levels.
2. Simultaneous bilateral stimulation of the intact sacral root at various levels.
3. Stimulation of the intact ventral and dorsal root separately.
4. Stimulation of the proximal and distal ends of the divided sacral root.
5. Stimulation of the proximal and distal ends of the divided dorsal and ventral roots (13).

From these studies, it became clear that stimulating the intact root is least effective and stimulating the ventral component is most effective and that no difference exists between right- and left-root stimulation (33).

However, this stimulation also causes some sphincteric contraction, owing to the presence of both autonomic and somatic fibers in the ventral root, and the studies were continued with the addition of neurotomy to eliminate the afferent fibers. These experiments showed that, to achieve maximally specific detrusor stimulation, the dorsal component must be separated from the ventral component and the somatic fibers of the root must be isolated and selectively cut (34).

The experiments also demonstrated that stimulation with low frequency and low voltage can maintain adequate sphincteric activity, but that stimulation with high frequency and low voltage will fatigue the external sphincter and block its activity. When high frequency/low voltage stimulation is followed by high voltage stimulation, bladder contraction will be induced and voiding achieved.

The finding that detrusor contraction can be activated separately from sphincteric activity and that adequate sphincteric contraction can be sustained without exciting a detrusor reaction made it seem possible that a true bladder pacemaker could be achieved. In addition, in histological and electron microscopic examination of the stimulated sacral roots, no damage was found when they were compared with the contralateral nonstimulated roots. Neither the operation nor the chronic stimulation damaged the ventral root, and the responses remained reliable and stable (13).

Tanagho's group later performed detailed anatomical studies on human cadavers. The aim was to establish the exact anatomical distribution of the entire sacral plexus, following it from the sacral roots in the spinal cord through the sacral foramen inside the pelvic cavity. Emphasis was placed on the autonomic pelvic plexus as well as the somatic fibers. With this anatomical knowledge, the stimulation of human sacral roots in neurogenic bladder dysfunction was developed and made clinically applicable as a long-term treatment (35). Direct electrical stimulation was performed through a permanently implanted electrode, placed mostly in contact with S3 nerve roots in the sacral foramen, after deafferentation.

The stimulation of sacral rootlet bundles isolated from the rest of the sacral root gave the same increase of bladder pressure when stimulated close to the exit from the dura, in the mid-segment, or close to the origin in the spinal cord. This could make the stimulation more selective, eliminating detrusor-sphincter dyssynergia.

In additional work, taking advantage of the knowledge that high frequency current can block large somatic fibers, electrical blockade of undesired responses was tested to replace selective somatic neurotomies. High frequency sinusoidal stimulation was effective in blocking external sphincter activity. However, the sinusoidal waveform is not efficient. Alternate-phase, rectangular wave is more efficient and induces the same blockade: alternating pulses of high frequency and low amplitude followed by pulses of low frequency and high amplitude were effective in inducing low pressure voiding without the need for somatic neurotomies. This approach has not yet been tried clinically, but it might prove to be the answer to the problem of detrusor-sphincter dyssynergia in electrically stimulated voiding (13).

The three main devices used for sacral neuromodulation is the Medtronic InterStim, the Finetech–Brindley (VOCARE) bladder system, and the rf BION systems. Each is explained in detail below.

MEDTRONIC INTERSTIM

Indications for use: urge incontinence, retention and urgency frequency, male and female dysfunctional voiding syndromes and postprostatectomy incontinence. There are also benefits beyond voiding disorders, including re-establishment of pelvic floor awareness, resolution of pelvic floor muscle tension and pain, reduction in bladder pain (interstitial cystitis) and normalization of bowel function.

The basic concept behind the implantable pulse generator (IPG) that provides stimulation to the sacral nerve is not far removed from the concepts behind cardiac pacing. A long-lived battery encased in biocompatible material is programmed to deliver pulses of electricity to a specific region of the body through an electrode at the end of an encapsulated wire.

Medtronic is the manufacturer of the InterStim neurostimulator. Earl Bakken, the founder of the company, first created a wearable, battery-operated pacemaker at the request of Dr. C. Walton Lillehei, a pioneer in open-heart surgery at the University of Minnesota Medical School Hospital, who was treating young patients for heart block.

The Itrel I, the first-generation neurostimulator, was introduced in 1983. Current versions are used for the treatment of incontinence, pain, and movement disorders.

System Overview

There are two established methods for sacral root neuromodulation using the Medtronic InterStim system.

1. An initial test phase, then the more permanent hardware is implanted.
2. An alternative method uses a staged testing–implant procedure, where a chronic lead is implanted and connected to a percutaneous extension and test stimulator.

Testing Phase (See Fig. 1). The testing hardware consists of a needle, test lead, test stimulator, interconnect cabling and a ground pad (Fig. 1).

- Needle (see Figs. 2 and 3).

A 20-gauge foramen needle with a bevelled tip is used to gain access to the sacral nerve for placing the test stimulation lead. The stainless steel needle is depth-marked along its length and electrically insulated along its center length. The portion near the hub is exposed to allow connection to

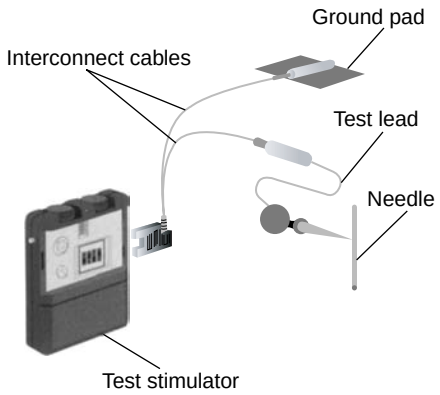


Figure 1. Test stimulation system.

Figure 2. Model 041828 (20 gauge) 3.5 in. (88.9 mm) foramen needles.

Figure 3. Model 041829 (20 gauge) 5 in. (127 mm) foramen needles.

the test stimulator. By stimulating through the uninsulated tip of the needle, the physician can determine the correct SNS site for the test stimulation lead.

- Test lead (see Fig. 4).

The initial test lead is a peripheral nerve evaluation (PNE) test lead with a coiled, seven-stranded stainless steel wire coated with fluoropolymer. Its electrode is extended to 10 mm (0.4 in.) to increase the length of coverage and reduce the effects of minor migration. Depth indicators help to align the lead electrode with the needle tip. The lead contains its own stylet, which is removed once the correct position has been found, leaving the lead flexible and stretchable, to mitigate migration.

- Test stimulator (see Fig. 5).

The most current version of test stimulators is the model 3625. The model 3625 test stimulator can be used both for patient screening, where the patient is sent home with the device, and for intraoperative usage in determining lead placement thresholds. It provides output characteristics that are similar to those of the implantable neurostimulator and can be operated in either monopolar or bipolar modes. It is battery operated by a regular, 9 V battery. The physician sets the maximum and minimum amplitude settings, allowing the patient to control the amplitude (within those maximum and minimum settings) to whatever level is comfortable.

The safety features of the stimulator include; an automatic output shut-off occurs when the amplitude is turned up too rapidly (as when the control is inadvertently bumped), a loose device battery will cause output shut-off also to prevent intermittent stimulation and shock to the patient, and sensors, which detect when electrocautery is being used, shut the output off. Turning the test stimulator off for a minimum of 3 s can reset the protection circuitry.

- Interconnect cables (see Fig. 6).

Single-use electrical cables are used to hook the test stimulation lead to the model 3625 test stimulator during the test stimulation procedure in the physician's office and when the patient goes home for the evaluation period.

Figure 4. Model 3057 test stimulation lead.

Figure 5. Model 3625 sacral nerve test stimulator.

The patient cable is used to deliver acute stimulation during the test procedure. The insulated tin-plated copper cable has a 2 mm socket at one end and a spring-activated minihook at the other end. The minihook makes a sterile connection to the foramen needle, test stimulation lead, or implant lead. The socket end is connected to the test stimulator by a long screener cable, the latter being a two-wire cable with a single connector to the model 3625 test stimulator at one end; one of the wires is connected to the patient cable and the other to the ground pad. After the test stimulation, the patient cable is removed and a short screener cable is substituted for at-home use. This cable is connected to the ground pad and directly to the test lead. It is designed to withstand the rigours of home use and can be disconnected, to facilitate changing clothes (13).

- Ground pad.

The ground pad provides the positive polarity in the electrical circuit during the test stimulation and the at-home trial. It is made of silicone rubber and is adhered to the patient's skin. As described above, for the at-home trial a short screener cable is substituted for the long screener cable and connected directly to the lead.

Surgical Technique Used for Acute Testing Phase:. The aims of percutaneous neurostimulation testing (PNE) are to check the neural and functional integrity of the sacral nerves, to determine whether neurostimulation is beneficial for each particular patient, and to clarify which sacral spinal nerves must be stimulated to achieve the optimum therapeutic effect in each individual case.

Local anesthetic is injected into the subcutaneous fatty tissue and the muscles, but not into the sacral foramen itself. The S3 foramen is localized on one side with a 20-gauge foramen needle. By stimulating through the uninsulated tip of the needle, the physician can find the correct sacral nerve stimulation site for placement of the test stimulation lead. Once the location of the S3 foramen is established, tracing of the other foramina is done. The

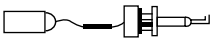


Figure 6. Model 041831 patient cable.

portion near the hub is exposed to allow connection to the test stimulator.

Keeping the needle at a 60° angle to the skin surface with a rostrocaudal and slightly lateral pointing tip of the needle will ensure that the needle is inserted into the targeted foramen. The puncture should progress parallel to the course of the sacral nerve, which normally enters at the upper medial margin of the foramen. This method achieves optimal positioning of the needle for stimulation and avoids injuring the spinal nerve. The insulated needle (cathode) is then connected to an external, portable pulse generator (Medtronic model 3625 test stimulator) via a connection cable. The pulse generator itself is connected to a neutral electrode (anode) attached to the shoulder.

Because patient sensitivity varies, the voltage used is between 1–6 V, which starts at 1 and is increased in 20 Hz increments. Stimulation of the S3 evokes the "bellows" effect (contraction of the levator ani and the sphincter urethrea). Also, there is plantar flexion of the foot on the ipsilateral side. If plantar flexion of the entire foot is observed, the gastrocnemius muscle should be palpated, because a strong contraction usually indicates stimulation of S2 fibers and should be avoided.

Stimulation of S3 generally produces the most beneficial effect. Furthermore, most patients will not tolerate the permanent external rotation of the leg caused by stimulation of S2. Occasionally, stimulating S4 also causes clinical improvement. Stimulation of S4 provokes a strong contraction of the levator ani muscle, accompanied by a dragging sensation in the rectal region. If stimulating one side produces an inadequate response, the contralateral side should be tested; the aim is to obtain a typical painless stimulatory response.

Once the optimal stimulation site has been identified, the obturator is removed from the foramen needle, and a temporary wire test lead (Medtronic model 3057 test lead) is inserted through the lumen of the needle. Once the test lead has been inserted into the needle, the latter must not be advanced any further in order to avoid severing the lead. The needle is then carefully removed from the sacral foramen, leaving the test lead in place. The stimulation is then repeated to check the correct position of the test electrode. To mitigate migration the lead contains its own stylet, which is removed once the correct position has been found, leaving the lead flexible and stretchable.

A repetition of the test stimulation, confirming the correct position of the test lead, is therefore mandatory at this stage; otherwise the test lead cannot be reinserted.

After correct positioning, the test lead is coiled on the skin and fixed with adhesive transparent film. Finally, the correct position of the wire is radiologically confirmed and the portable external impulse generator is connected.

Percutaneous Extension Hardware (see Fig. 7). If acute testing is inconclusive, or when there is a need for positive fixation of the test lead, percutaneous extension hardware is the best method used. Also called the staged implant, it is an alternative method for patient screening.

The chronic lead is implanted in the normal manner and is connected to a percutaneous extension (model 3550-05). The extension is designed to provide a connection between the chronic lead and the external test stimulator. Positive contact is made using four set screws; the connection is sealed with a silicone boot that covers the set screws. The percutaneous extension, which is intended for temporary use, features four insulated wires, wound together and sized for a small incision, so that they can be brought through the skin. The percutaneous extension is then connected to the screener cable, as described above (13).

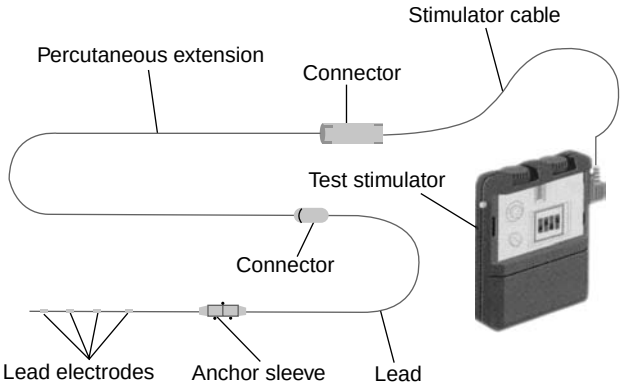


Figure 7. Percutaneous extension system.

Chronic System. The chronic system consists of an implantable neurostimulator, a lead, an extension, a physician programmer and a patient programmer.

- Neurostimulator (see Fig. 8).

The implantable neurostimulator (Medtronic model 3023) weighs ~42 g and has a volume of 22 cm^3. It comprises ~70% battery and 30% electronics. The physician has unlimited access to programmable parameters such as amplitude, frequency, and pulse width. Each parameter can be changed by means of an external, physician programmer that establishes a rf link with the implanted device. A patient programmer provides limited access to allow the patient to turn the neurostimulator on and off, or to change amplitude within a range established by the physician (via the physician programmer) (13).

The external titanium container of the neurostimulator may be used in either a monopolar configuration (lead negative, can positive) or a bipolar configuration, which will result in marginally better longevity. The life of the neurostimulators is usually ~7–10 years. Factors that affect this are the mode, programming of the amplitude, pulse width and frequency, and the use of more than one active electrode.

- Implantable lead system (see Fig. 9–14).

Figure 8. Model 3023 implantable neurostimulator.

Figure 9. Model 3080 lead.

Figure 10. Model 3092 lead.

Figure 11. Model 3093 lead.

Figure 12. Model 3886 lead.

Figure 13. Model 3889 lead.

Figure 14. Model 3966 lead.

The lead is a quadripolar design, with four separate electrodes that can be individually programmed to plus, minus, or off. This allows the physician to optimize the electrode configuration for each patient and to change programming, without additional surgery, at a later date, to adapt to minor lead migration or changing disease states. The electrode sizes, spacing, and configurations have been designed specifically for SNS.

The lead is supplied with multiple stylets and anchors, to accommodate physician preferences. A stylet (straight or bent) is inserted into the lumen of the lead to provide extra stiffness during implant. Two different degrees of stiffness provide the physician with options to tailor the handling and steering properties of the lead, as preferred. The stylet must be removed before connection with the mating component.

The physician also has a choice of anchors, which allow fixation of the lead to stable tissue to prevent dislodging of the lead after implantation. Three anchor configurations are available: a silicone rubber anchor fixed in place on the lead has wings, holes and grooves to facilitate suturing; a second type, also made of silicone, slides into place anywhere along the lead body, and must be sutured to the lead to hold it in place; a new plastic anchor is also available, which can be locked in place anywhere along the lead body without a suture to the lead.

- Quadripolar extension (see Fig. 15).

Figure 15. Series 3095 extension.

Figure 16. Physician programmer.

The quadripolar extension, which is available in varying lengths to facilitate flexibility in IPG placement, is designed to provide a sealed connection to the lead. This extension provides the interface with the neurostimulator. Positive contact is made with four set screws, and the connection is sealed with a silicone boot covering the screws.

- Physician programmer (Fig. 16).

The console programmer (Medtronic model 8840 N'Vision) is a microprocessor-based system that the physician uses to program the implanted neurostimulator noninvasively. The programmer uses an application-specific memory module, installed by means of a plug-in software module.

- Patient programmer (Fig. 17).

The patient programmer also communicates with the implanted neurostimulator by an rf link. The patient can adjust stimulation parameters within the range set by the physician. This range is intended to allow the patient to turn the device on or off, and to change amplitude for comfort (as during postural changes), without returning to the physician's office.

Surgical Technique Used for Chronic Implantable System. The sacral foramen electrode and impulse generator are implanted under general anesthesia. Long-acting muscle relaxants must not be used, as these would impair the intraoperative electrostimulation.

The patient is placed in the prone position with a 45° flexion of the hip and knee joints. An 8 cm long midline

Figure 17. Patient programmer.

incision is made above the sacrum, reaching one-third caudal and two-thirds cranial from the S3 foramen. After transection of the subcutaneous fat, the muscle fascia (thoracolumbar fascia) is incised approximately 1 cm lateral of the midline in a longitudinal direction.

Usually, the Gluteus maximus has to be incised over a length of 1–2 cm for good exposure of the S3 foramen and a little further caudal if implantation of the S4 foramen is intended. The paraspinal muscles are then divided longitudinally and the dorsal aspect of the sacrum is exposed.

Intraoperative test stimulation, using the same equipment as for the acute testing phase, will confirm the precise location of the foramen selected. The foramen needle is left in place to avoid relocalisation of the foramen while preparing the permanent electrode for implantation. Proximal to the four contact points of the permanent electrode, a silicon rubber cuff is glued to the electrode body. The cuff is fitted with three eyelets to accommodate nonabsorbable atraumatic needle-armed sutures.

After removal of the foramen needle, the permanent electrode (Medtronic quadripolar lead, model 3080) is gently inserted into the foramen. Renewed test stimulation will determine the most effective contact point between the electrode and spinal nerve; the most distal contact point is termed "0", with the subsequent three being numbered 1–3 sequentially. An identical motor response at all four contact points is ideal. If only one contact gives a satisfactory response, the electrode should be repositioned at a different angle to the foramen and the test stimulation repeated. The preattached sutures are then used to secure the electrode to the ligaments overlying the periosteum of the sacral bone. Test stimulation should be repeated at this stage to confirm an appropriate position of the electrode after fixation.

A small skin incision is now made in the flank between the iliac crest and the 12th rib on the side where the electrode has been placed. A subcutaneous tunnel is formed between the two wounds, starting from the flank incision and running toward the sacral incision.

The obturator of the tunneling device is removed and the silicone sheath, which is open at both ends, left in place. The free end of the electrode is guided through the sheath to the flank incision, after the stylet has been removed from the electrode.

The silicone sheath is now removed from the flank incision, the proximal end of the electrode is marked with a suture, and the electrode is buried in a subcutaneous pocket that has been created at the site of the flank incision. The flank incision is temporarily closed, leaving the marking suture exposed between the skin sutures. The sacral incision is then closed in layers and covered with a sterile dressing.

The patient is now positioned on the contralateral flank. The flank and abdomen on the side chosen previously for placement of the Medtronic InterStim model 3023 implantable pulse generator are disinfected and the surgical field is draped with a sterile cover. The flank incision is now reopened, and a subcutaneous tunnel is again created between the flank incision and the subcutaneous pocket in the lower abdomen through which a connecting extension cable (Medtronic quadripolar extension, model 3095) between electrode and impulse generator is guided.

Once the electrode has been connected to the extension cable in the area of the flank incision, the contact point is sealed with a silicone cover, fixed with two sutures and placed subcutaneously. The flank incision is closed in layers and covered with a sterile dressing.

Finally, the other end of the connecting cable is attached to the impulse generator. The generator is attached to the rectus fascia using two nonabsorbable sutures. The abdominal incision is closed in two layers and covered with sterile dressings.

On the first postoperative day, anterior–posterior and lateral radiographs of the implant are obtained to verify that all components are correctly positioned and will act as a control for comparison in case of subsequent problems.

Modifications of the surgical procedure include placement of the pulse generator in the gluteal area thus avoiding repositioning of the patient during the procedure and implantation of bilateral electrodes, which should be powered by a two-channel pulse generator (Medtronic Synergy, model 7427) for adequate synchronous independent stimulation of each side. The implant remains deactivated at least until the day following surgery and will be activated by a telemetric programming unit (Medtronic Console Programmer, model 7432) allowing programming of all features of the implant by the physician during the initial activation and follow-up stages.

NEW MEDTRONIC TINED LEAD PERCUTANEOUS IMPLANT (SEE FIGS. 18 AND 19)

Tined leads offer sacral nerve stimulation through a minimally invasive implant procedure. The use of local anesthesia allows for patient sensory response during the implant procedure. This response helps ensure optimal lead placement and may result in better patient outcomes. With previous lead designs, many physicians used general anesthesia, which did not allow for patient sensory response.

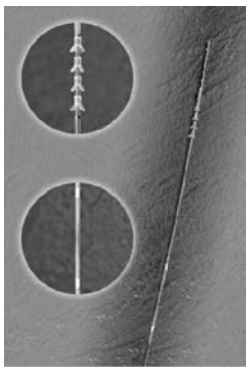


Figure 18. Tined lead percutaneous implant.

Among the advantages of the minimally invasive implant procedure are the radiopaque markers to identify where tines are deployed, helping physicians in identifying the exact lead location relative to the sacrum and nerves. Tactile markers indicate lead deployment, and a white marker bands on the lead and tactile markers aid in proper lead placement and to notify the physician when the tines are ready to be deployed.

Percutaneous lead placement allows use of local anesthesia. This reduces the risks of general anesthesia and surgical incision and may facilitate faster patient recovery time as a result of less muscle trauma and a minimized surgical incision. Also, it may reduce surgical time as a result of a sutureless anchoring procedure and reduced number of surgical steps.

To date, a positive response to the PNE test has been the only predictive factor for the long-term efficacy of sacral nerve stimulation therapy. Current studies show that up to 40% of patients who experience improvement in symptoms during PNE test stimulation with a temporary lead do not have this improvement carried through after neurostimulator implantation (36). A study by Spinelli et al. looked at patients who underwent tined lead implant without PNE testing, and reported a positive outcome of 80% during the screening phase, which was maintained at an average follow up of 11 months, resulting in a higher success rate than that currently reported in the literature (37).

The development of the new tined lead allows fully percutaneous implantation of the permanent lead and offers the possibility of a longer and more reliable screening period than that possible with the PNE test. The advantage for patient screening are that the permanent tined lead is less prone to migration, hence if the results of screening are positive, the lead is already in the precise place where positive results were obtained, and there is a decrease in false-positive and false-negative results after screening (37). However, use (or lack thereof) of PNE testing in conjunction with the tined lead differs from center to center, depending on fiscal and/or other reasons.

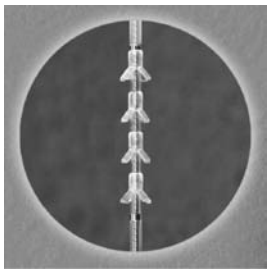


Figure 19. Tined lead percutaneous implant.

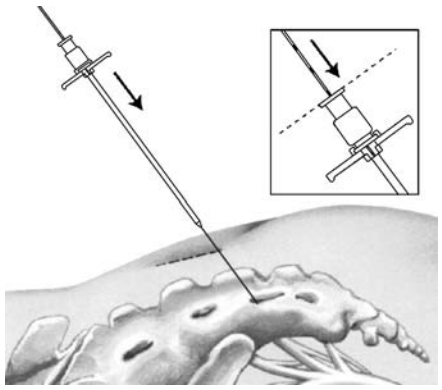


Figure 20. The foramen needle stylet and directional guide.

The tined lead models 3093 and 3889 are designed to work with the current lead introducer model 355018 or 042294.

Surgical Technique for Tined Lead Implant. The foramen needle is inserted and tested for nerve response. The foramen needle stylet is then removed and replaced with the directional guide (see Fig. 20). The foramen needle itself is then removed.

A small incision is made on either side of the directional guide, which is followed by fitting the dilator and the introducer sheath over the directional guide and advanced into the foramen (see Fig. 21). The guide and the dilator are then removed, leaving the introducer sheath in place.

The lead is then inserted into the introducer sheath and advanced until visual marker band C on the lead lines up with the top of the introducer sheath handle. Using fluoroscopy, electrode 0 of the lead is confirmed to be proximal to the radiopaque marker band at the distal tip of the sheath (see Fig. 22).

While holding the lead in place, the introducer sheath is retracted until visual marker band D on the lead lines up with the introducer sheath handle. Using fluoroscopy, radiopaque marker band at the tip of the sheath is

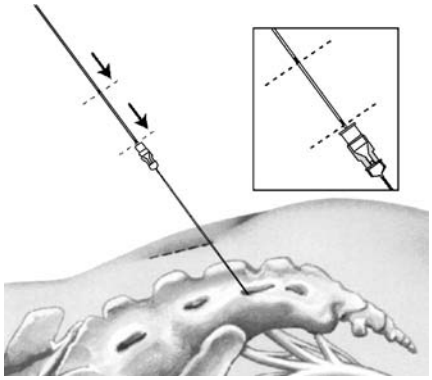


Figure 21. Fitting the dilator and introducer sheath.

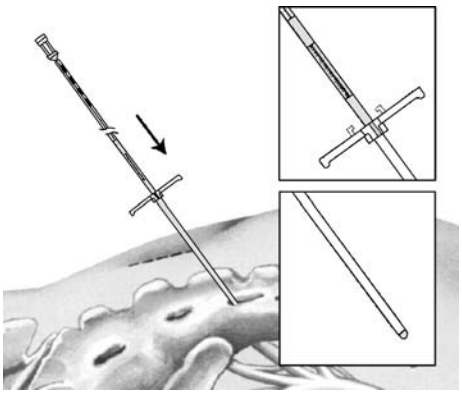


Figure 22. Confirming lead proximal to radiopaque marker band.

confirmed to be proximal to electrode 3 and adjacent to radiopaque marker band A on the lead (see Fig. 23).

Test stimulation of the various electrodes (0, 1, 2, 3) is done and the responses are observed. If necessary, the lead is repositioned within the foramen. When the lead is in the proper position, the lead is held in place and the introducer sheath and lead stylet are carefully withdrawn, thereby deploying the tines and anchoring the lead.

FINETECH–BRINDLEY (VOCARE) BLADDER SYSTEM

Introduction (see Fig. 24)

Indications for use: The VOCARE bladder system is indicated for the treatment of patients who have clinically complete spinal cord lesions with intact parasympathetic innervation of the bladder and are skeletally mature and neurologically stable. However, patients with other neurological disorders, including multiple scleroses, spinal cord tumours, transverse myelitis, cerebral palsy and meningomyelocele, have also benefited from the implant(38). *A secondary use of the device is to aid in bowel evacuation and promote penile erection*.

The sacral anterior root stimulation (SARS) system was developed by Brindley with the support of the Medical Research Council (Welwyn Garden City, Herts, UK), is manufactured by Finetech Medical Ltd. in England, and is marketed as the Vocare system by NeuroControl Corporation (Cleveland, OH) (1).

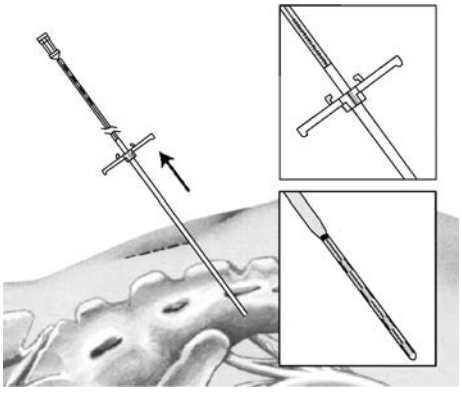


Figure 23. Confirming marker band proximal to electrode 3.

Beginning in 1969, Brindley developed a new device to stimulate sacral roots at the level of the cauda equina. This technique, first tested in baboons, led to the development of a stimulator that was first successfully implanted in a patient in 1978 (39).

Hardware

The Finetech–Brindley bladder controller is composed of external and internal equipment.

1. External (see Fig. 25):

   One analog and three digital versions of the external controller are available in different countries (1). This device has no batteries but is powered and controlled by rf transmission from a portable external controller operated by the user and programmed by the clinician. It consists of a transmitter block connected to the control box via a transmitter lead. The patient holds the transmitter over the implanted receiver to apply stimulation. A new, smaller control box that is more powerful will be available in the coming months (39).
2. Internal (see Fig. 26):

   The internal equipment consists of three main parts: (1) the electrodes, (2) the cables, (3) and the receiver block.

   Two types of electrodes are used, depending on the approach (intra- or extradural).

   For intradural implantation the electrode mounts in which the anterior sacral roots are trapped are called "books" because of their shape.

   The two-channel implant has an upper book with only 1 slot. Trapping of S3 and S4 roots is often sufficient to obtain bladder contractions. In males, S2 roots were trapped in the upper book and S3 and S4 roots, in the lower book.

   The three-channel implant is composed of two electrode books. The upper book contains three parallel slots for S3 and S2 roots and the lower contains one slot for S4 roots. There are three electrodes in each slot (one cathode in the center and two anodes at the two ends) to avoid stimulation of unwanted structures.

   The four-channel implant has two books like those of the three- channel implant, and the four slots allow independent stimulation of four sets of nerve fibers. It is used in patients who retained sacral-segment pain sensitivity.

   The special eight-channel implant allowed the stimulation of four anterior roots and the destruction of any of the four posterior roots, if necessary, after implantation. It is no longer used.

   For extradural implantation the cables end with three helical electrodes (a cathode between two anodes) and are attached to the roots with a strip of Dacron-reinforced silicone rubber. The cables used

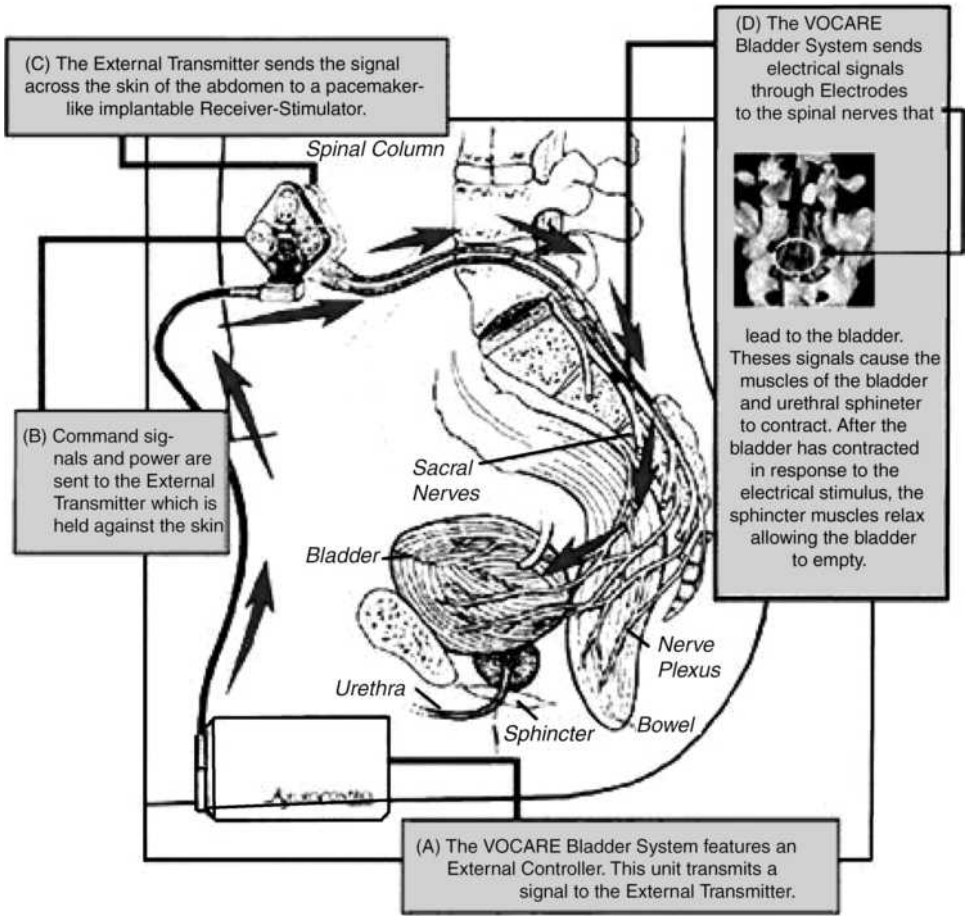


Figure 24. VOCARE bladder system.

Figure 25. External equipment. (a) New control box. (b) Original control box. (c) Transmitter lead. (d) Transmitter block.

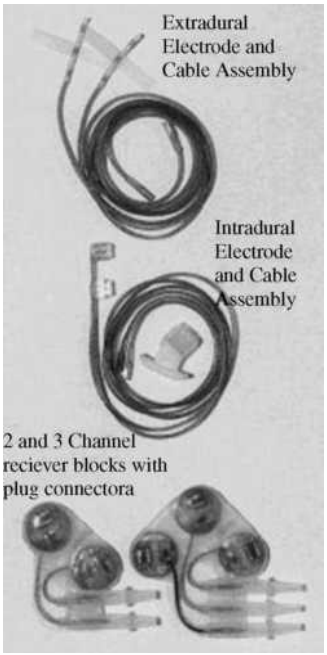


Figure 26. Internal equipment.

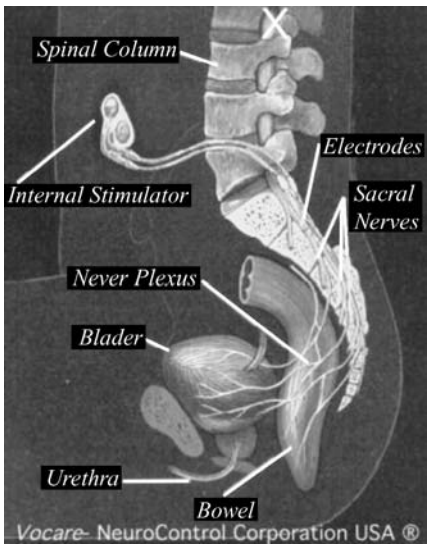


Figure 27. Finetech–Brindley system.

are encapsulated in silicone rubber, and the wires are made of 90% platinum and 10% iridium and connect the electrodes to the radio receiver block. The radio receiver block, which contains two, three, or four radioreceivers imbedded in silicone rubber, is activated by pulse-modulated rf waves (39).

Surgical Technique for Finetech–Brindley System (see Fig. 27):. The surgical technique for ***intrathecal*** implantation developed by Brindley et al. (40) involves laminectomy of the fourth and fifth lumbar vertebrae and the first two pieces of the sacrum, exposing 10–12 cm of dura. The dura and arachnoid are opened at the midline to expose the roots. The roots are identified by their size and situation and by perioperative stimulation during the recording of bladder pressure and observation of skeletal muscle responses with the naked eye.

The S2 anterior roots contract the triceps surae, the glutei, and the biceps femoris. The S3 anterior roots innervate the pelvic floor and the toe flexors. The S4 anterior roots innervate the pelvic floor. The sphincters (anorectal and urethral) are innervated predominantly by S4 and also by S3 and S2. The detrusor response is always obtainable by stimulation of S3 and S4 and sometimes achievable by stimulation of S2.

The roots are split into the anterior and posterior components. The identity of the posterior root is confirmed by electrical stimulation and then a segment measuring ~20–40 mm in length is removed. When the S5 root has been identified, it is resected if no bladder response is obtained (39).

If a posterior rhizotomy is performed, stimulation can be applied to mixed sacral nerves in the sacral spinal canal extradurally, since the action potentials generated on the afferent axons do not reach the spinal cord. This has the advantage that the electrodes can be placed extradurally, reducing the risk of leakage of cerebrospinal fluid along the cables, and reducing the risk of breakage of the cables where they cross the dura. In addition, the extradural nerves are more robust than the intradural roots, being covered with epineurium derived from the dura, and require less dissection than the intradural roots; therefore, there is less risk of neuropraxia of the axons, which could otherwise lead to a delay in usage of the stimulator but not usually in permanent loss of function (1,41).

The benefits of a posterior rhizotomy include abolition of the neurogenic detrusor over activity, resulting in increased bladder capacity and compliance, reduced incontinence, and protection of the kidneys from ureteric reflux and hydronephrosis. The rhizotomy also reduces detrusor-sphincter dyssynergia, which improves urine flow, and prevents autonomic dysreflexia arising from distension or contraction of the bladder or bowel. In addition, a posterior rhizotomy improves implant-driven micturition. However, there are also drawbacks with a rhizotomy. They include abolition of reflex erection, reflex ejaculation, reflex defecation and sacral sensation, if present. Still, in many subjects with spinal lesions, these reflexes are not adequately functional, and function can be restored by other techniques (42).

The surgical technique for *extradural* implantation involves laminectomy of the first three pieces of the sacrum. It may also involve laminectomy of the L5 vertebra, depending on whether it is decided to implant electrodes on S2 roots (39). Extradural electrodes are used for patients in whom arachnoiditis makes separation of the sacral roots impossible. In some centers, however, extradural electrodes are used for all or nearly all patients.

After electrode implantation, the operation proceeded with closure of the dura, tunneling of the leads to a subcutaneous pocket in the flank, and closure of the skin. The patient is turned over and the leads are prepared for connection to the implantable stimulator.

At this time the leads are connected via an aseptic cable to an experimental stimulator. Prior to stimulation the bladder is filled with 200 mL saline using a transurethral filling catheter. The experimental stimulator consisted of two synchronized current sources with a common cathode. Pressure responses are elicited using pulse trains of 3–5 s duration; containing identical monophasic rectangular pulses delivered at a rate of 25 pulses $\cdot s^{-1}$. Stimulation is usually limited to the S3 and S4 ventral roots since they contain most of the motoneurons innervating the lower urinary tract.

After 15–20 min of experimental stimulation the leads are disconnected from the stimulator and the normal procedure is resumed with implantation stimulator.

A two-channel transurethral pressure catheter is used to measure intravesical and intraurethral pressure. The urethral pressure sensor is positioned at the level of the external sphincter such that in response to suprathreshold stimulation a maximal pressure response is measured. Pressures are sampled at 8 Hz, displayed on a monitor, and stored in a portable data logger (43).

All patients are followed up according to a fixed protocol. Urodynamic measurements are taken at 2 days, 15 days, 4 months, and 1 year after surgery and every 2–3 years thereafter. Renal ultrasound examination is performed every year. Stimulation is performed for the first time

Figure 28. BION microstimulator.

between days 8 and 14, depending on the level of the spinal cord lesion (33).

PUDENDAL NERVE STIMULATION FOR THE TREATMENT OF THE OVERACTIVE BLADDER (rf BION) (SEE FIGS. 28 AND 29)

Indications for use: The rf Bion system is still relatively new, and though no clear, established indications have been set so far, its activity on the pudendal nerve and inhibition of the detruser muscle makes it ideal for overactive bladder disorders.

Electrical stimulation of the pudendal nerve has been demonstrated to inhibit detrusor activity and chronic electrical stimulation may provide effective treatment for overactive bladder disorders (44). The hurdle to date has been the technical challenge of placing and maintaining an electrode near the pudendal nerve in humans; however, recent development of the BION has made chronic implantation feasible.

The BION is a small, self-contained microstimulator that can be injected directly adjacent to the pudendal nerve (see Fig. 28). The ischial spine is an excellent marker for the pudendal nerve as it re-enters the pelvis through Alcock's canal. This is a very consistent anatomical landmark in both men and women. Also, the implanted electrode is protected in this area by both the sacral tuberous and sacrospinous ligaments. Stimulation in this area activates afferent innervation over up to three sacral segments. Efferent stimulation also provides direct activation of the external urethral sphincter, the external anal sphincter, and the levator ani muscles, which may be of some benefit in bladder control. The external components of this neural prosthesis include a coil that is worn around the subject's hips and a controller that is worn around the shoulder or waist.

The technique chosen to implant the device is that of the transperineal pudendal block. This approach is minimally invasive and is well established. A special implant tool was

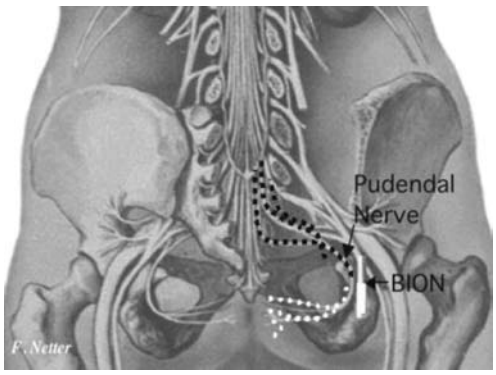


Figure 29. Placement of BION system near pudendal nerve.

devised to facilitate placement. The BION implantation technique was developed in cadavers. The optimum insertion location is 1.5 cm medial to the ischeal tuberosity using a vaginal finger to guide the implant toward the ischial spine where electrical stimulation of the pudendal nerve may be confirmed (see Fig. 29).

A percutaneous stimulation test (PST) was developed and proved to be a very effective way to assess acute changes in bladder volumes while stimulating the pudendal nerve. A baseline cystometrogram (CMG) was obtained followed by percutaneous pudendal nerve stimulation for 10 min with a repeat CMG.

The first implant was done on August 29, 2000. The BION was implanted under local anesthesia with intravenous sedation. Proper placement was verified by palpation and EMG activity. An intermittent stimulation mode of 5 s on 5 s off was used. Subjects returned 5–7 days later for activation, to distinguish between postoperative pain and potential stimulation pain. Subjects were followed up at 15, 30, and 45 days after activation. At each follow-up visit they underwent another cystometrogram and brought in a 72 h voiding diary. The results indicated a favorable response to maximum cystometric capacity throughout the study period. Diary entries verified improvement—incontinent episodes decreased by 65%, and both daytime and nighttime voids were decreased, as was pad use per day.

FUTURE DIRECTION OF THE THERAPY HARDWARE

The ongoing development of tools and hardware is driven by the desire to reduce the invasiveness of the implant and the likelihood of adverse events. Development efforts are concentrated on system components and tools that will allow implantation of the lead system through small incisions or percutaneous approaches. It is inevitable that the size of the neurostimulator will be reduced as future generations of the device are developed; more efficient power batteries and packaging will drive this aspect of development.

A rechargeable power battery may allow a smaller device. Although a smaller device would be welcomed, attaining this goal with a rechargeable battery is not seen as the best approach. A rechargeable neurostimulator would

require the patient frequently to recharge the unit; this would inconvenience the patient and could reduce patient compliance. Additionally, a rechargeable battery would be more expensive than a nonrechargeable one owing to the technology required and the additional equipment necessary for recharging. Furthermore, this would not eliminate the need for periodic replacement of the neurostimulator every 5–10 years. System components will be optimized for the therapy, to reduce the time needed for management of both implant and patient. The incorporation of microprocessors and implementation of features such as a battery gauge will provide additional operational information while decreasing the time needed to manage the patient.

Physicians will be able to analyze system use, lead status, and other parameters. The addition of sensing technology may provide an opportunity to create a closed-loop system that captures data to optimize both diagnosis and functioning.

Bilateral stimulation may provide more efficacious therapy. There is considerable interest in this approach, and it seems to be a probable avenue of research in the near future. However, any use of bilateral stimulation would have to justify the larger neurostimulator, the extra lead system, and the additional costs associated with this approach; at present, there is no scientific experience to support this approach.

Apart from a reduction in the size of the implanted device, enhanced physician control is the most likely development to occur in the foreseeable future. Graphics-based programming and control will simplify device programming; it will allow more complex features to be incorporated in the neurostimulator without adding undue complexity to the physician programmer. Management of patient data files will become easier as additional data-management features are added to the programmer; the physician will be able to obtain a patient-programming history and other patient-management data. It is conceivable that, in the not-so-distant future, the physician may be able to access patient-device data over the Internet, thus making unnecessary some clinic visits and allowing for remote follow-up of patients who are on holiday or have moved house.

Future devices may allow software loading in a noninvasive manner, to upgrade the device long after implantation. Such capability could be used to provide new therapy algorithms as well as new therapy waveforms.

The future will also bring enhanced test stimulation devices, which will provide improved fixation during the test stimulation period. The development of new leads is one such focus with the aim of allowing a longer test stimulation period without lead migration.

The future application of SNS is dependent on new clinical research. Pelvic disorders, such as pelvic pain and sexual dysfunction, appear likely to be the First areas of investigation; sacral anterior root stimulation for spinal cord injury may also provide a worthwhile avenue of enquiry. The development of these applications—or of any other, for that matter—will potentially require new waveforms and the development of new therapy algorithms. The future is as open as the availability of resources and the application of science allow (45).

BIBLIOGRAPHY

Cited References

1. Jezernik S, Craggs M, Grill WM, et al.Electrical stimulation for the treatment of bladder dysfunction: current status and future possibilities. Neurol Res 2002;24(5):413–430.
2. Galvani L. De virfbus electricitatis in motu musculari, commentaffus. De Bononiensi Scientarium et Artium Instituto Atque Academia. 1791;7:363–418.
3. Volta A. Letter to Sir Joseph Banks, March 20, 1800. On electricity excited by the mere contact of conducting substances of different kinds. Philos Trans R Soc London (Biol) 1800;90:403–431.
4. Duchenne GBA. De I'dlectrisation localisde et de son application & la physiologie, & la pathologie et b la therapeutique. Paris; 1855.
5. Duchenne GBA. Physiologie des moumments demonstrde b l'aide de l'experimentation electrique et de l'observation clinique, et applicable b I'6tude des paralysies et des deformations. Paris; 1867.
6. Chaffee EL, Light RE. A method for remote control of electrical stimulation of the nervous system. Yale J Biol Med 1934;7:83.
7. Glenn WWL, Phelps ML. Diaphragm pacing by electrical stimulation of the phrenic nerve. Neutosurgery 1985; 17:974–1044.
8. Glenn WWL, Mauro A, Longo E, et al.Remote stimulation of the heart by radiofrequency transmission. N Engl J Med 1959;261:948.
9. House WF. Cochlear implants. Ann Otol Rhinol Laryngol 1976;85(27):1–93.
10. Saxtorph MH. Strictura urethrae—Fistula petinei—Retentio urinae. Clinisk Chirurgi. Copenhagen: Gyldendalske Fodag; 1878.
11. Katona F, Benyo L, Lang J. Uber intraluminare elektrotherapie vor verschiedenen paralytischen zustlinden des gastrointestinalen tractes mit quadrangularstrom. Zentralbl Chir 1959;84:929.
12. Matona F. Stages otvegetative afferentation in reorganization of bladder control during electrotherapy. Urol Int 1975;30:192–203.
13. Schlote N, Tanagho EA. Electrical Stimulation of the lower urinary tract: historical overview. In: Jonas U, Grunewald V, editors. New Perspectives in sacral nerve stimulation. Dunitz; 2002. p 1–8.
14. McGuire WE. Response of the neurogenic bladder to vadous electrical stimuli [dissertation]. Department of Surgery, Bowman Gray School of Medicine; 1955.
15. Boyce WH, Lathem JE, Hunt LD. Research related to the development ofan artificial electrical stimulator for the paralyzed human bladder: a review. J Urology 1964;91:41–51.
16. Bradley WE, Chou SN, French LA. Further experience with the radio transmitter receiver unit for the neurogenic bladder. J. Neurosurg 1963;20:953–960.
17. Caldwell KPS. The electrical control of sphincter incompetence. Lancet 1963;2:174.
18. Fall M, Erlandson BE, Carlsson CA, Lindstr6m S. The effect of intravagnal electrical stimulation on the feline urethra and urinary bladder. Scand) Urol Nephrol (Supp) 1977;44: 19–30.
19. Lindstrim S, Fall M, Carlsson CA, Edandson BE. The neurophysiologcal basisaf bladder inhlbition in response to intravaginal electrical stimulation. Urology 1983;129:405–410.
20. McGuire EL, ZIang SC, Horwinski ER, Lytton B. Treatment of motor and sensory detrusor instability by electical stimulation. J Urol 1983;129:78–79.
21. Govier FE, Litwiller S, Nitti V, Kreder KJ, Jr., Rosenblatt P. Percutaneous afferent neuromodulation for the refractory

overactive bladder: results of a multicenter study. J Urol 165:1193, 2001.
22. van Balken MR, Vandoninck V, Messelink BJ, Vergunst H, Heesakkers JP, Debruyne FM, et al. Percutaneious tibial nerve stimulation as neuromodulative treatment of chronic pelvic pain. Eur Urol; 43:158, 2003.
23. van Balken, Michael R, Vergunst Henk, Bemelmans Bart LH. The use of Electrical Devices for the Treatment of Bladder Dysfunction: A Review of Methods. Urol September 2004;172(3):846–851.
24. Ingersoll EH, Jones LL, Hegre ES. Effect on urinary bladder of unilateral stimulation of pelvic nerves in the dog. Am Physiol 1957;189:167.
25. Hald T, Agrawal O, Mantrowitz A. Studies in stimulation of the bladder and its motor nerves. Surgery 1966;60:848–856.
26. Hald T, Meier W, Khalili A, et al. Clinical experience with a radio-linked bladder stimulator. J Urol 1967;97:73–78.
27. Friedman H, Nashold BS, Senechat R. Spinal cord stimulation and bladder function in normal and paraplegic animals. J Neurosurg 1972;36:430–437.
28. Jonas U, Heine JR, Tanagho EA. Studies on the feasibility of urinary bladder evacuation by direct spinal cord stimulation. 1. Parameters of most effective stimulation. Invest Urol 1975;13:142–150.
29. Jonas U, James LW, Tanagho EA. Spinal cord stimulation versus detrusor stimulation. A comparative study in six acute dogs. Invest Urol 1975;13:171–174.
30. Jonas U, Tanagho EA. Studies on the feasibility of urinary bladder evacuation by direct spinal cord stimulation. II. Poststimulus voiding: a way to overcome outflow resistance. Invest Urol 1975;13:151–153.
31. Nashold BS, Friedman H, Boyarsky S. Electrical activation of micturition by spinal cord stimulation. J Surg Res 1971;11:144–147.
32. Thirhoff JW, Bazeed MA, Schmidt RA, et al. Regional topography of spinal cord neurons innervating pelvic floor muscles and bladder neck in the dog: a study by combined horseradish peroxidase histochemistry and autoradiography. Utol Int 1982;37:110–120.
33. Tanagho EA, Schmidt RA. Bladder pacemaker: scientific basis and clinical future. Urology 1982;20:614–619.
34. Schmidt RA, Bruschini H, Tanagho EA. Sacral root stimulation in controlled micturition: peripheral somatic neurotomy and stimulated voiding. Invest Urol 1979;17:130–134.
35. Probst M, Piechota HA, Hohenfeliner M, et al.Neurostimulation for bladder evacuation: is sacral root stimulation a substitute for microstimulation? Br J Urol 1997;79:554–566.
36. Bosch JLHR, Groen J. Sacral nerve neuromodulation in the treatment of patients with refractory motor urge incontinence: long-term results of a prospective longitudinal study. J Urol 2000;163:1219.
37. Spinelli M, Giardiello G, Gerber M, Arduini A, Van Den Hombergh U, Malaguti S. New Sacral Neuromodulation Lead For Percutaneous Implantation Using Local Anesthesia: Description And First Experience. J Urol 2003;170(5): 1905–1907.
38. Brindley GS. The first 500 patients with sacral anterior root stimulator implants: general description. Paraplegia 1994; 32:795–805.
39. Egon G, Barat M, Colombel P, et al.Implantation of anterior sacral root stimulators combined with posterior sacral rhizotomy in spinal injury patients. World J Urol 1998;16:342–349.
40. Brindley GS, Polkey CE, Ruston DN. Sacral anterior root stimulators of bladder control in paraplegia. Paraplegia 1982;28:365–381.
41. Brindley GS, Polkey CE, Rushton DN, Cardozo L. Sacral anterior root stimulators for bladder control in paraplegia: The first 50 cases. J Neurol Neurosurg Psychiat 1986;49: 1104–1114.
42. Rijkhoff N. Neuroprostheses to treat neurogenic bladder dysfunction: current status and future perspectives. Childs Nerv Syst 2004 Feb; 20(2): 75–86.
43. Rijkhoff N, Wijkstra H, Kerrebroeck P, et al.Selective detrusor activation by sacral ventral nerve-root stimulation: results of intraoperative testing in humans during implantation of a Finetech-Brindley system. World J Urol 1998;16: 337–341.
44. Vodusçek DB, Light KJ, Libby JM. Detrusor inhibition induced by stimulation of pudendal nerve afferents. Neurourol Urodyn 1986;5:381.
45. Gerber M, Swoyer J, Tronnes C. Hardware: development and function. New Perspectives in sacral nerve stimulation. In: Jonas U, Grunewald V, editors. Dunitz: 2002. p 81–88.

See also BIOTELEMETRY; FUNCTIONAL ELECTRICAL STIMULATION; TRANSCUTANEOUS ELECTRICAL NERVE STIMULATION (TENS).

BLIND AND VISUALLY IMPAIRED, ASSISTIVE TECHNOLOGY FOR

ANDREW Y. J. SZETO
San Diego State University
San Diego, California

INTRODUCTION

Severe visual impairment represents one of the most serious sensory deficits that a human being can have. When this sensory input channel is so impaired that little useful information can pass through it, assistive devices that utilize alternative sensory input channels are often necessary. Familiar examples include the use of Braille and the white cane, respectively, for reading and obstacle avoidance by persons who are blind. Both of these assistive devices provide environmental information to the user via the sense of touch. Other assistive devices provide environmental feedback via the sense of hearing.

In the material that follows, examples of available assistive technology and promising new assistive technology under development for persons who are blind or severely visually impaired are presented. This article begins with an overview of the prevalence and impairments associated with blindness impairments and follows with an examination of reading aids, independent living aids, and mobility aids. The article concludes with a brief look at kinds of assistive technology likely to be available in the near future for persons with severe visual impairments.

The term blindness has many connotations and is difficult to define precisely. To many people, blindness refers to the complete loss of vision with no perception of light. The U.S. government, however, defines blindness as the best corrected visual acuity of 20/200 or worse in the better seeing eye. The acuity designation 20/200 means that a vision impaired person is able to see at a distance of 20 ft (6.09 m) what a person with normal visual acuity is able see at 200 ft (60.96 m). Low vision is defined as the

Table 1. Prevalence of Blindness and Low Vision Among Adults 40 Years and Older in the United States[a]

	Blindness		Low Vision		All Vision Impaired	
Age, Years	Persons	%	Persons	%	Persons	%
40–49	51,000	0.1	80,000	0.2	131,000	0.3
50–59	45,000	0.1	102,000	0.3	147,000	0.4
60–69	59,000	0.3	176,000	0.9	235,000	1.2
70–79	134,000	0.8	471,000	3.0	605,000	3.8
>80	648,000	7.0	1,532,000	16.7	2,180,000	23.7
Total	*937,000*	*0.8*	*2,361,000*	*2.0*	*3,298,000*	*2.7*

[a]Abstracted from Ref. 3 Arch. Ophthalmol. Vol. 122, April 2004.

best corrected visual acuity that is worse than 20/40 in the better seeing eye. People with extreme tunnel vision (a visual field that subtends an angle $>20^\circ$ regardless of the acuity within that visual angle) also are classified as being legally blind and thus qualify for certain disability benefits.

It is important to realize that a great majority ($\sim$70–80%) of people with severe impairments has some degree of usable vision (1,2). The severity of vision loss can vary widely and result in equally varying degrees of functional impairment. Although the degree of impairment may differ from one person to another, people who are blind or have low vision experience the common frustration of not being to see well enough to perform common everyday tasks.

The prevalence of blindness and low vision among adults 40 years and older is given in Table 1. According to the National Eye Institute (2), a component of the National Institutes of Health in the United States Department of Health and Human Services, the leading causes of vision impairment and blindness are primarily age-related eye diseases. These include age-related macular degeneration, cataract, diabetic retinopathy, and glaucoma. The 2000 census data revealed >5 million people of all ages in America have visual impairments severe enough to significantly interfere with their daily activities.

CONSEQUENCES OF SEVERE VISUAL IMPAIRMENTS

The two major difficulties faced by persons who are blind or severely visually impaired are access to reading material and independent travel or mobility. Simple-to-sophisticated technology has been used in a variety of assistive devices to help overcome these problems. The term reading is used in this context to include access to all material printed on paper or electronically. Reading material can include text, pictures, drawing, tables, maps, food labels, signs, mathematical equations, and graphical symbols. Safe and independent mobility is used to encompass both obstacle avoidance and navigation. For safe and independent mobility, the first concern is avoiding obstacles, such as curbs, chairs, low hanging branches, and platform drop-offs. After the sight impaired traveler has gained an awareness of the basic spatial relationships between objects within the travel environment, their needs wayfinding or navigational assistance, which involves knowing one's position, one's heading with respect to the intended destination, and a suitable path to reach it.

LOW VISION READING AIDS

People with low vision significantly out number those who are totally without sight (Table 1). Hence, the consumer market for low vision aids is much larger than the one dedicated to people with zero vision. The technology used in low vision aids is rather straightforward and the technologically used is relatively mature. Hence, only a brief overview of such assistive devices will be presented before discussing the more challenging issues faced by persons with zero useful vision. For readers desiring detailed product information about low vision aids, a search of the Internet using the term low vision aids will yield a bounty of pictures, product specifications, and purchasing information.

All low vision aids aim to maximize an individual's residual vision to its fullest. Low vision aids can be categorized as optical, nonoptical, and electronic. Optical aids include handheld magnifying glasses, telescopes mounted on eyeglass frames, and even microscope lenses. Nonoptical aids include enlarged high contrast print and high intensity lamps.

Electronic low vision aids represent the highest level in terms of cost, complexity, and performance. They include electronic video magnifiers that project printed material on a closed circuit monitor, regular television, or computer screen. Electronic video magnifiers can maximize readability of the written material by providing a wide range of magnification, brightness, contrast, type of fonts, and foreground and background colors. A good example of a modern closed circuit TV type of electronic low vision aid is the Optelec Traveller (Fig. 1). This portable video

Figure 1. This portable video magnifier has a built-in 6 in. (15.24 cm) color screen and can magnify text and pictures up to 16 times. (Courtesy of Optelec International, New York.)

Figure 2. Closed-circuit television with computer based text-to-speech output, a talking computer.

magnifier has a built-in 6 in. (15.24 cm) color screen and can magnify text and pictures up to 16 times and more if its video signal is sent to a television set.

People with tunnel vision or central blind spots due to macular degeneration often find it difficult and tiring to read an entire computer screen. For such individuals, the advent of the talking computer (Fig. 2) represented a major technological breakthrough. The capability and flexibility of such a computer or reading machine addressed many of their needs as well as the needs of persons without any useful vision.

READINGS AIDS FOR THE BLIND

For persons with essentially zero useful vision, the tactile sense has been utilized as an alternative sensory input channel for reading. One of the oldest reading substitutes for the blind is Braille, a six dot matrix code that Louise Braille adapted in 1824 for use by blind persons to read written text. The standard Braille cell consists of two columns and three rows of dots separated by 2.3 mm with 4.1 mm separating adjacent cells. Each Braille cell occupies a rectangular area of 4.3×8.6 mm and can represent 2^6-1 (or 63) possible symbols within that areas. Grade I Braille maps each cell a one-to-one basis to each letter of the alphabet, basic punctuation marks, and simple abbreviations so that Grade I Braille has an informational density of approximately 1 bit per 6 mm^2 of surface area. For greater informational compactness and faster reading rates, Grade II Braille uses combinations of dots to represent contractions, frequently used words, prefixes, and suffices. Grade III Braille is even more compact and affords the highest reading rates, but very few people ever master it. The largest proportion of Braille literature is produced at the Grade II level, which can be read at up to 200 words per minute (4) by those proficient in Braille. Braille is a unique reading aid that not only gives blind persons access to printed material but also provides them with a writing medium.

Despite Braille's unique place as a complete writing system that is spatially distributed and retains many advantages of a printed page, Braille is a specialized code that only a small percentage of blind individuals learn to use. This is especially true for persons who become blind after the age of 15 years. Given the difficulty of mastering Braille, the lack of up-to-date Braille printed material, and advances in alternate technologies such as electronic reading machines, many blind individuals choose to not bother with Braille.

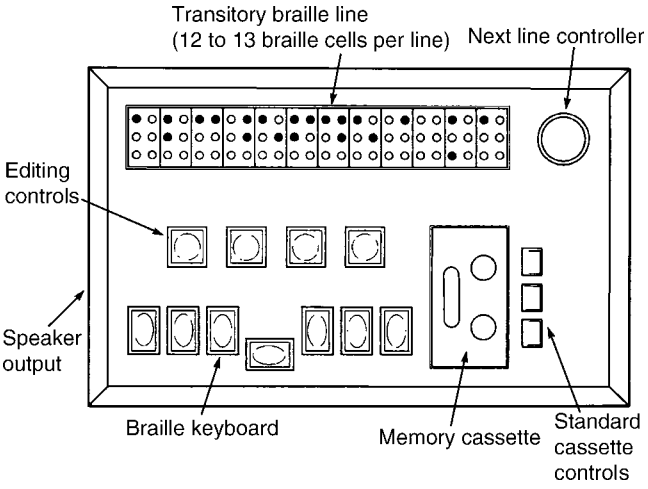


Figure 3. Portable refreshable Braille reader that can playback or store messages using a cassette recorder. The reader has a single-line tactile display, a Braille keyboard, and a tape cassette for data storage and recall. (Picture taken from Fig. 2.8 of Ref. 5.)

Other disadvantages of Braille printed material include the cost to produce it, store it, and maintain it. Embossable Braille paper is not only bulky, heavy, and expensive, the pattern of raised dots (laboriously and noisily impressed into the paper) is fragile and short lived. Assistive technologies such as portable Braille readers (Fig. 3) have mitigated some of the inconveniences associated with Braille (5), but these electronic Braille readers–recorders often do not display the two-dimensional (2D) information embedded in graphs, tables, and mathematical formulas. The single and dual line tactile displays found in most portable readers also makes the rapid search for content via headings very difficult.

Refreshable Braille readers can be used as a computer interface for accessing information on the computer screen. Some full-sized electronic Braille displays are 80 cells long and cost upward of $10,000. The dots in these transient Braille displays are produced by pins raised and lowered (refreshed) to form Braille characters. Refreshable Braille readers allow users to access any portion of the screen information via specialized control buttons and status Braille cells. Tactually distinguishable arrow keys offer screen cursor control while extra status cells provide additional information about text-attributes or line and colon positions.

Refreshable Braille displays are especially useful for deaf blind individuals and users working with computer programming languages. For example, the Braille Voyager 44 (Fig. 4), made by F.J. Tieman BV, has a 44 cell Braille display, and 5 thumb keys for screen navigation. Using its built-in macro program, USB connection, and any screen reader, the Voyager enables a user to access many features of the Windows operating system.

Figure 4. The Braille Voyager 44 made by Manufacturer: F.J. Tieman BV. It has a 44 cell Braille display and 5 navigation keys.

Despite Braille's many drawbacks and limited popularity, its long history, status as the only complete writing and reading system for the blind, and tenacity of advocates like the American Federation for the Blind combine to keep Braille viable as an informational medium. Nonprofit groups like the Braille Institute produce millions of pages of Braille each year for business, schools, government agencies and individuals across the nation. They sell recreational reading material in Braille to both children and adults and provide low cost transcription, embossing, and tactile graphic services.

For the majority of blind persons who do not know Braille, reading material converted into the audio format (aka talking books) and played back on variable speed tape recorders have proven to be popular and convenient to use. To overcome spoken speech's inherently slower reading rate, variable speed tape recorders with special electronic circuits that compensate for the pitch change during high speed playback (1.5–3 times normal speed) can be used. Obtaining reading material in audio form for playback on such recorders also has become more convenient as vendors make downloading of electronic text and audio files available to their subscribers (6).

Although audio books are popular for persons with severe visual impairments, this approach does not work for reading the newspaper, daily mail, memoranda, cookbooks, technical reports, handwritten notes, and common everyday correspondence, such as utility bills and bank statements. Before the advent of a reading machine, which has now become part of a general purpose talking computer, persons with no useful vision relied on human readers with its attendant inconvenience, loss of independence, and lack of privacy.

For severely sight impaired individuals and even those who know Braille, the power, convenience, and versatility of a reading machine, also known as a talking computer, have made it the preferred method of accessing most reading material. First marketed in the early 1980s, reading machines of today are affordable, compact, and can reliably and rapidly convert alphanumeric text into synthetic speech. In addition to a synthetic voice that reads aloud the actual text, the talking computer or reading machine also provides auditory feedback of cursor location and navigational commands.

A talking computer or dedicated reading machine contains artificial intelligence that converts alphanumeric text into spoken speech. The multistep process begins with an optical device that scans the text of a printed document or web page and, using optical character recognition, converts that alphanumeric text string into prefixes, suffixes, and root words (Fig. 5).

The process through which the text string is converted into speech output is somewhat complex and undergoing refinement. The clarity and naturalness of the voice output depend on the text-to-speech technique employed. In general, clearer and more natural costlier sounding speech requires more memory and greater processing power and is thus more expensive.

After the written material has been converted into a text string by optical character recognition software, one of

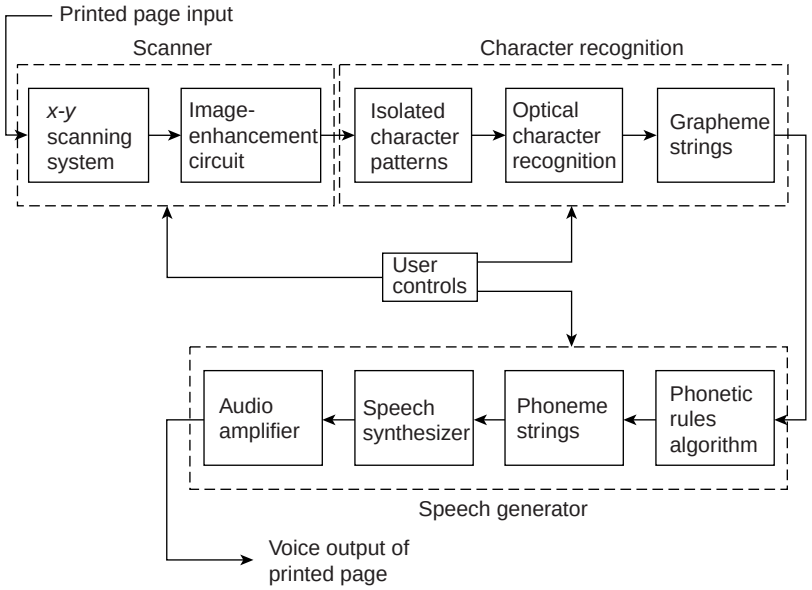


Figure 5. Functional components of a reading machine. (Taken from Fig. 2.18 of Ref. 5.)

three basic methods can be employed to convert the string into speech sounds. The first method is called whole word look-up. It produces the most intelligible, life-like speech, but it is the most memory intensive even for modest sized vocabularies. Despite steady advances in low cost, high density memory chips, whole-word-look-up tends to be prohibitively expensive or the vocabulary is limited (7).

A less memory intensive approach is the letter-to-sound conversion in which a synthetic sound processor divides the text string into basic letter groups and then follows certain pronunciation rules for the creation of speech. Many languages (especially American English) are replete with numerous exceptions to the usual rules of pronunciation. Hence, the quality of the speech output using letter-to-sound conversion depends on the sophistication of the rules and the number of exceptions employed (7,8).

The third method of converting text into speech is called morphonemic text-to-speech conversion. This approach relies on prestored combination of morphemes (basic units of language such as prefixes, suffixes, and roots) and their corresponding speech sounds. Some 8000 morphemes can generate $\sim 95\%$ of the English words (8,9) so this approach avoids the memory demands of the whole word look-up approach. Morphonemic based speech generation generally yields synthetic speech output that is more intelligible than the letter to speech approach, but is more demanding computationally. Continuing advances in technology have now made single chip text to speech converters powerful, capable, and affordable in consumer electronics (10).

A blind individual using a computer running a text-to-speech program can now hear what is on the screen and use cursor keys to select a specific part of the screen to read. Equipped with such a computer, high speed connection to the Internet, and a modern reading machine, sight impaired individuals now have wide access to news, e-mail, voice messaging, and Internet's vast repository of information. These powerful information technologies have reduced the social isolation formerly felt by blind persons while also broadening their employment opportunities.

One example of how recent technological advances are improving access to reading materials is the Spoken Interface that Apple Computer unveiled at the 2005 Annual Technology & Persons with Disabilities Conference held in Los Angeles. Because Spoken Interface is a screen reader that is fully integrated into Apple's operating system, assistive technology developers should be able to set up easy inter-operability between their software and the operating platform with little additional modifications.

Another example of a low cost, user friendly, and powerful text-to-speech software is the TextAloud MP3 by Nextup Technologies (http://www.nextuptech.com/about.html). This software converts any text into natural sounding speech or into MP3 files for downloading and later playback on portable electronic devices (e.g., MP3 players, pocket PCs, and portable data assistants).

MANDATED WEB ACCESSIBILITY

With so much information available on the Internet and the blind people's increasing dependence on it, the United States government included web accessibility in its 1998 amendment of the Rehabilitation Act (11). Section 508 of this law requires that when Federal agencies develop, procure, maintain, or use electronic information technology, they must ensure that this technology offers comparable access to Federal employees who have disabilities. Although the scope of Section 508 is limited to the Federal sector, these requirements have gradually spread to the private sector, especially to large corporations that deal frequently with the Federal government.

The accessibility requirements of Section 508 are reflected in several guidelines, including as the Web Content Accessibility Guidelines (WCAG) from the World Wide Web Consortium (W3C). The WCAG recommendations, which are updated periodically, include implementing standardized style sheets instead of custom HTML tags and offering closed-captioning and transcripts of multimedia presentations. Other recommendations for making a web site compliant (12) include the following: provide text alternates to images; make meaning independent of color; identify language changes; make pages style sheet independent; update equivalents for dynamic content; include redundant text links for server-side image maps; use client-side image maps when possible; put row and column headers in data tables; associate all data cells with header cells; title all frames; make the site script independent.

An assortment of adaptive hardware and software can be effectively utilized once a web site satisfies the WCAG recommendations (13). Persons with low vision can change their browser settings or use screen magnifiers. Internet users who are blind or have very limited vision can use text-based Web browsers with voice-synthesized screen readers, audio browsers, or refreshable Braille displays to read and interact with the Web.

Recent efforts to increase internet's compatibility with assistive technologies used by sight impaired persons include the development and implementation of search engines that read aloud their results using male and female voices. Some websites offer speech-synthesized renditions of articles from news organizations like BBC, Reuters, and the New York Times (14).

While internet accessibility by persons with severe visual impairments is improving, a number of problems and challenges remain. Screen readers or Braille keyboards that blind people use to navigate the Internet cannot scan or render graphical elements into a readable format. Spam, security checks, popup ads, and other things that slow down a sighted person's Web searches are even worse impediments for those with severe visual impairments using assistive technology.

INDEPENDENT LIVING AIDS

Because blindness and severe visual impairments are so pervasive in their impact, numerous and relatively low cost assistive devices have been developed to make non-reading activities of daily living (ADL) easier. In general, these ADL devices rely on the users' auditory or tactile sense for their operation.

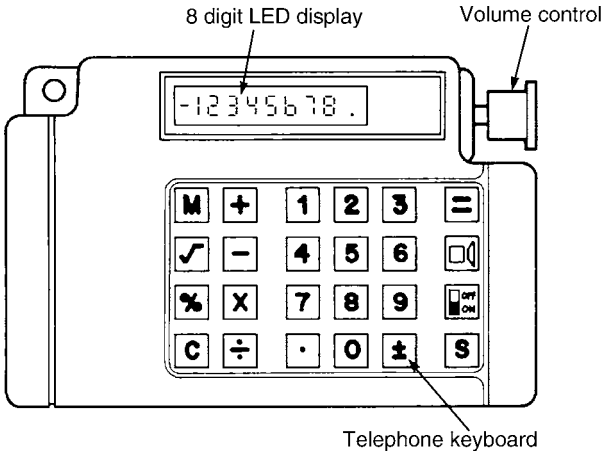


Figure 6. A talking calculator with a female voice speaks the individual digits or whole integers. Its large 8 digit LCD readout is ~0.6 in. (1.52 cm) high. The calculator can add, subtract, divide, multiply and calculate percentages. (Reproduced from Ref. 5.)

A quick check of electronic catalogs on the Internet shows that many types of independent living aids are available. For example, special clocks and timers that give both voice and vibratory alarms are available in various sizes and features. Other assistive devices for ADL include talking wrist watches, push-button padlocks, special money holders, Braille embossed large push button phones, and jumbo sized playing cards embossed with Braille. Personal care items for the blind include talking bathroom scales, thermometers, glucose monitors, blood pressure gauges, and prescription medicine organizers. Educational aids that facilitate note taking, calculating, searching, printing, and organizing information include talking calculators (Fig. 6), pen-like handheld scanner for storing text, letter writing guides with raised lines, Braille metal guides and styluses, and signature guides.

MOBILITY AIDS

For persons with severe visual impairments, the advent of powerful and affordable reading machines and the vast amount information (already in electronic form) on the internet, the problem of access to reading materials has been significantly ameliorated. In contrast, their other major problem (the ability to travel safely, comfortably, gracefully, and independently through the environment) has only been partially solved.

Despite years of effort and some major advances in technology, there is no widely accepted electronic travel aid (ETA). Most blind individuals rely on the sighted human guide, a guide dog, and the familiar white cane. The human sighted guide offers companionship, intelligence, wayfinding capability, route recall, and adaptability. Unfortunately, human guides are not always available, and their very presence constitutes a lack of independence. A guide dog or animal guide has been popular, but not every blind person can independently care for a living animal nor afford the cost of its care. In some social situations, a guide dog can be awkward or unacceptable. The white cane, which is both a tool and a symbol for the blind, can alert sight-impaired travelers to obstacles in their path, but only those at ground level and < 5 ft. (1.5 m) away. Above ground obstacles and especially those at head height remain a source of apprehension and danger for travelers depending on just the white cane.

To understand why decades of research and development efforts have not yielded an efficacious and widely accepted electronic travel aid, one needs to realize that mobility aids must deal with a very different set of constraints and inputs than do reading aids. An identification error made by reading aids results only in misinformation, mispronunciation, or inconvenience. In contrast, a failed detection of an obstacle or step-down or a missed landmark can lead to confusion, frustration, apprehension, and physical injury.

Another major difference between a mobility aid and a reading aid lies in their operating milieu. Mobility aids must detect and analyze unconstrained, long range, and highly variable environmental inputs, that is, obstacles of differing sizes, textures, and shapes distributed over a 180° wide area. In contrast reading machines must identify and convert into intelligible speech inputs that are often well defined and short ranged, for example, high contrast printed alphanumeric symbols and punctuation marks (15).

To further complicate matters, users of reading aids often have the luxury of focusing all or most of their attention on the task at hand: interpreting the output of the reading aid. Users of mobility aids, however, must divide their attention among several demanding tasks associated with traveling, such as avoiding obstacles, listening to environmental cues, monitoring their physical location, recalling the memorized route, and interpreting the auditory or tactile cues from their mobility aid. Given these challenges, today's mobility aids represent a much less satisfactory solution (in comparison to available reading aids) to the problem of independent and safe mobility for persons with severe visual impairments.

THE IDEAL MOBILITY AID

Before examining the capabilities of currently available mobility aids, it is desirable to enumerate the fundamental features of an ideal electronic travel or mobility aid (Table 2) (16–18). The first three items of an ideal mobility aid can be categorized as nearby obstacle avoidance; features 4–7 fall under the category of navigational guidance or wayfinding; and features 8–10 represent good ergonomic design or user friendliness.

CONVENTIONAL ELECTRONIC TRAVEL AIDS

Standard or conventional electronic travel aids detect nearby obstacles, but provide no wayfinding assistance. Obstacle detection entails the transmission of some sort of energy into the surrounding space and the detection of the reflections. After analyzing the reflected signals, the ETA warns the traveler of possible obstacles using either auditory feedback or tactile feedback.

Table 2. The Ideal Mobility Aid

	Capabilities and Features	Description
Feature No. 1	Obstacle detection	Detect nearby obstacles that are ahead, at head level, and at ground level and indicate their approximate locations and distances without causing sensory overload.
Feature No. 2	Warn of impending Obstacles	Reliably locate and warn of impending potholes, low obstacles, step-downs and step-ups.
Feature No. 3	Guidance around obstacles	Guide the traveler around impending obstacles.
Feature No. 4	Ergonomically designed	Offer voice and/or tactile feedback of traveler's present location. Capable of voice input operation and/or have tactually distinct push buttons
Feature No. 5	Wayfinding	Able to monitor the traveler's present location and indicate the direction toward the destination
Feature No. 6	Route recall	Be able to remember a previous route and warn of changes in the environment due to construction or other blockages
Feature No. 7	Operational flexibility	Reliably function in a variety of settings, that is, outdoors, indoors, stairways, elevators, and cluttered open spaces
Feature No. 8	User friendliness	Be portable, rugged, fail-safe, and affordable for a blind user
Feature No. 9	Cosmesis	Be perceived by potential users as cosmetically acceptable and comfortable to use in terms of size, styling, obtrusiveness, and attractiveness
Feature No. 10	Good battery life	Have rechargeable batteries that can last for at least 6 h per charge

The LASER CANE (Fig. 7) is one of the few conventional ETAs that can serve as a stand-alone, primary travel aid because it has obstacle detection (features 1–3 of Table 2) and is reasonably user friendly and cosmetic (features 8–10). The laser cane's shaft houses three narrow-beam lasers; the lasers scan upward, forward, and downward. Reflections from objects in these zones are detected by three optical receivers also housed in the shaft. The UP channel monitors head level obstacles and causes high pitched beeps to be emitted. The FORWARD channel monitors objects located 4–10 ft. (1.21–3.01 m) ahead of the cane's tip and produces warning signals in the form of either vibrations in the handle of the cane or a medium (1600 Hz) audio tone. Obstacles encountered by the DOWN channel produce a low frequency (200 Hz) warning tone (19). Because the laser cane is swept through an arc $\sim$3–4 ft. (0.91–1.21 m) wide in the direction of the intended path (in a manner similar to standard long cane usage), the laser cane augments the auditory and tactile feedback of an ordinary white cane by detecting objects at greater distances and, most importantly, head level obstructions.

The laser cane's main drawbacks include it being somewhat costly and fragile. It also cannot monitor the traveler's geographic location nor guide the traveler toward the intended destination (features 5 and 6). Field tests and consumer feedback revealed that laser obstacle detection can be highly variable because certain surfaces and objects reflect laser light better than others. For example, the laser beam mostly passes through glass so that the laser cane may miss glass doors or large glass windows ahead.

Although the laser cane is imperfect, it has one major advantage as an ETA; It is failsafe. Should its batteries run down or its electronics malfunction, the laser cane can still serve as a standard long cane (20) and thus still be useful to the traveler.

Another commercially available electronic travel aid is the Sonic Guide, an eyeglass frame equipped with one ultrasonic transmitter and two receivers embedded in

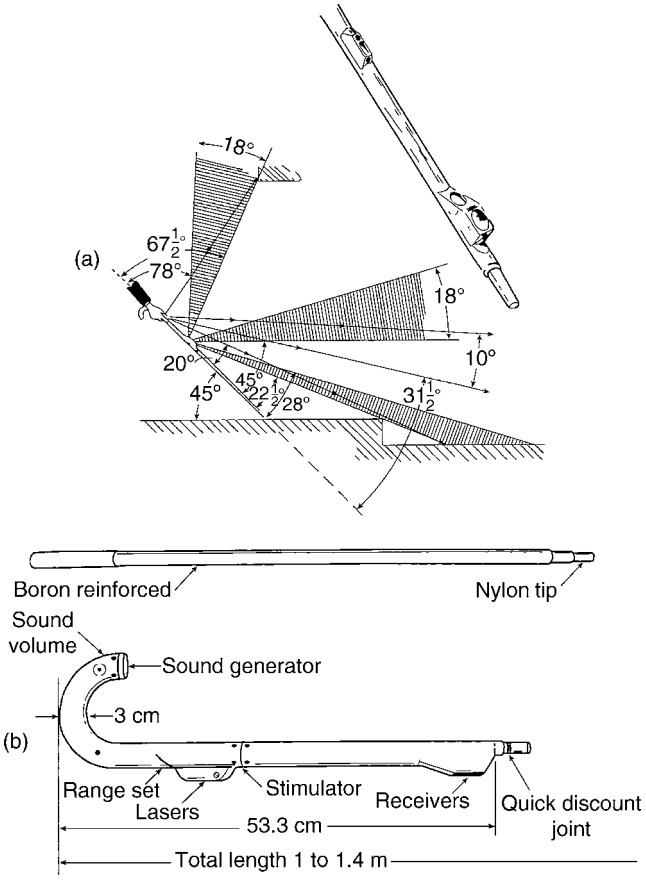


Figure 7. The laser cane projects three narrow beams of laser light. If any of the beams (up, forward, and downward) encounter an object and is reflected back to the receivers in the cane's shaft, a tactile or auditory warning is generated. (Reproduced from Ref. 19.)

the nose piece (21). The pulsed ultrasonic beam radiates through a forward solid angle of $\sim 100^\circ$. Objects in the environment reflect ultrasound back to the two receivers with time delays proportional to their distance and angle with respect to the wearer's head. The wearer is given awareness of his surroundings via binaural auditory feedback of the reflected signals, recreating the experience of echolocation as found in bats or dolphins. An object's distance is displayed in terms of frequencies proportional to the object's distance from the user. The azimuth of an object relative to the user's head is displayed via the relative intensity of tones sent to the ears (stereoscopic aural imaging). As a result, the binaural sounds heard by the user changes as he moves or turns his head.

To circumvent Sonic Guide's tendency to interfere with normal hearing, Kuc (22) investigated the utility of using vibrotactile feedback via a pair of sonar transceivers and vibrators worn on the wrists. Being on opposite sides of the body, the dual sonar transceivers offered better left–right obstacle discrimination than could a single sonar unit embedded in the nose piece of the eyeglasses. The wrist mounted pager-like device vibrated at a frequency inversely related to the reflecting object's distance from that side of the body.

Unfortunately, neither the original eyeglass frame based Sonic Guide nor the wrist worn sonar guide can serve as a stand-alone travel aid because neither can detect impending step-ups, step-downs, or other tripping hazards in the pathway. Other user comments about the Sonic Guide include interference with normal hearing, sensory overload, and difficulty in combining the aid's feedback with other important environmental cues such as the sound of traffic at street intersections, tactile feedback from a white cane, or the subtle pull of a guide dog.

In contrast to Sonic Guide's rich auditory feedback, the Mowat Sensor implements the design philosophy that simpler is better. The Mowat Sensor is a handheld ultrasonic flash light that acts like a clear path detector. It measures $6 \times 2 \times 1$ in. ($15 \times 5 \times 2.5$ cm), weighs 6.5 oz (184.2 g), can be easily carried in a pocket or purse, and is manufactured by Pulse Data International Ltd. of New Zealand and Australia.

The Mowat device emits a pulsed elliptical ultrasonic beam $\sim 15^\circ$ wide by 30° high, a beam pattern that should detect doorway sized openings located some 6 ft. (1.8 m) away. Reflections from objects in the beam pattern cause the Mowat to produce vibrations that are inversely proportional to the object's distance from the detector. As the traveler points at and gets closer to the object, the Mowat vibrates faster and faster. As the traveler aims moves away from that object, the vibrations slow and then cease. Objects outside of Mowat's beam pattern produce no vibrations.

The Sonic Guide, Mowat Sensor, and their various derivatives share similarities while representing two divergent design philosophies. They all employ ultrasound instead of laser light to detect nearby obstacles. None of them can detect tripping hazards, such as impending step-ups, step-downs, uneven concrete walkways, or small low obstacles in the path of travel so they cannot serve as a stand-alone travel aid. The Mowat sensor scans a small portion of the environment, displays limited data from that region, and offers easily interpreted vibratory information to the user. Alternatively, the Binaural Sonic Guide sends a broad sonic beam into much of the traveler's forward environment, displays large amounts of environmental information, and leaves it up to the user to select which portion of the auditory feedback to monitor and which to ignore.

While similar in concept, obstacle detection via ultrasound and obstacle detection via laser light interact with the environment differently. For example, hard vertical surfaces and glossy painted surfaces reflect sound and light very well so they tend to be detected by both methods at greater distances than oblique surfaces or dark cloth covered soft furnishings. Transparent glass, however, reflects sound very well, but laser light very poorly. Hence an ultrasonic beam would readily note the presence of a glass door whereas laser light could miss it entirely. Sonar based ETAs, however, are susceptible to spurious sources of ultrasound such as squealing air breaks on buses. Such sources and even heavy precipitation can cause the sonar sensor to signal the presence of a phantom obstacle or produce unreliable feedback. Furthermore, because all ETAs display environmental information via the sense of touch or hearing, severe environmental noise and wearing gloves or ear muffs can reduce a user's ability to monitor an ETAs feedback signals.

Other drawbacks of conventional electronic travel aids include the lack of navigational guidance (features 5 and 6 of Table 2), thus limiting the blind traveler to familiar places or necessitating directional guidance from a sighted guide until they have memorized the route. Furthermore, conventional ETAs often require the user to actively scan the environment and interpret the auditory and tactile feedback from the aids. These somewhat burdensome tasks require conscious effort and can slow walking speed.

INTELLIGENT ELECTRONIC TRAVEL AIDS

Recent advances in technology have sparked renewed efforts to develop mobility aids that address some of the aforementioned drawbacks. One promising intelligent electronic travel aid, under development at the University of Michigan Mobile Robotics Laboratory, is the GuideCane (23). The GuideCane (Fig. 8) is a semiautonomous robotic guide that improves user friendliness by obviating the burden of constant scanning while also guiding the traveler around obstacles, not merely detecting them. It consists of a self-propelled and servocontrolled mobile platform connected to a cane. An array of 10 ultrasonic sensors is mounted on the small platform. The sensors emit slightly overlapping signals to detect ground-level obstacles over a 120° arc ahead of the platform. The sonar units, made by Polaroid Corporation, emit short bursts of ultrasound and then uses the time of flight of the reflections to gauge the distance to the object. The sonar has a maximum range of 30 ft. (10 m) and an accuracy of $\sim 0.5\%$ (24).

When walking with the GuideCane, the user indicates his intended direction of travel via a thumb-operated minijoystick mounted at the end of a cane attached to the

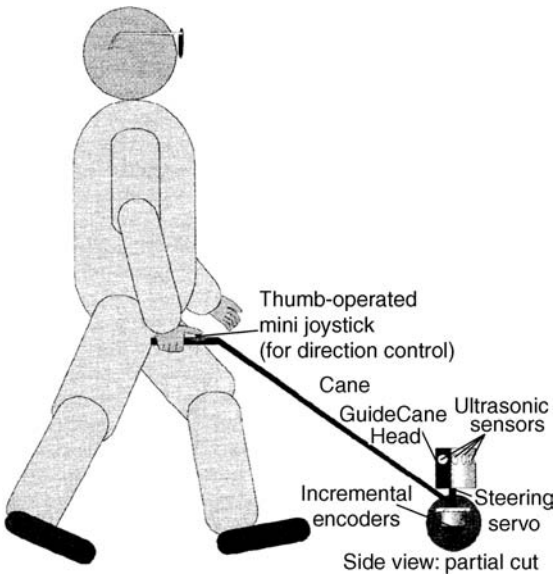


Figure 8. The GuideCane functions somewhat like a robotic guide dog. It is able to scan the environment and steer around obstacles by using its ultrasonic sensors, steering servomotors, and on-board computer to keeps track of nearby obstacles and the intended path of travel. (Reprinted from figure on p. 435 of Ref. 23.)

platform. The mobile platform maintains a map of its immediate surroundings and self-propels along the indicated direction of travel until it detects an obstacle at which time the robotic guide steers itself around it. The blind traveler senses the GuideCane's change of direction and follows it accordingly.

Like the Laser cane, the GuideCane can function as a stand-alone travel aid because it gives advance warning of impending step-downs and tripping hazards. Its bank of 10 ultrasonic detectors and ability to navigate around detected obstacles make the GuideCane easier and less mentally taxing to use than the Laser Cane. To address the wayfinding needs of the blind traveler, efforts are underway to add GPS capability, routing software and area maps to the GuideCane. The drawbacks of the wheel mounted GuideCane, however, include its size and weight and its inability to detect head height objects.

NAVIGATIONAL NEEDS

Electronic travel aids like those described above are becoming proficient at detecting and enabling the traveler to avoid obstacles and other potential hazards. Avoiding obstacles, however, represents only a partial solution to a blind person's mobility problem. Many visually impaired or blind travelers hesitate to visit unfamiliar places because they fear encountering an emergency or possibly getting lost. Their freedom of travel is hampered by having to pre-plan their initial trip to a new place or needing to enlist the help of a sighted person.

Furthermore, blind pedestrians, even those with training in orientation and mobility, often experience difficulty in unfamiliar areas and areas with free flowing traffic, such as parking lots, open spaces, shopping malls, bus terminals, school campuses, and roadways or sidewalks under construction. They also have difficulty crossing nonorthogonal, multiway traffic intersections (25). Conventional traffic signals combined with audible pedestrian traffic signals have proven somewhat helpful in reducing the pedestrian accident rates at intersections (26–28), but audible traffic signals offer guidance only at traffic intersections and not other important landmarks.

One proposed solution for meeting the wayfinding needs of blind travelers is the Talking Sign, a remote infrared signage technology that has been under development and testing at The Smith-Kettlewell Eye Research Institute in San Francisco, CA (29,30). The Talking Signs system consists of strategically located modules that transmit environmental speech messages to small, hand held receivers carried by blind travelers (Fig. 9). The repeating and directionally selective voice messages are transmitted to the receiver by infrared (IR) light (940 nm, 25 kHz). Guided by these orientation aids, blind travelers can know their present location and move in the direction from which the desired message, for example, Corner of Front Street and Main Street, is being broadcasted, thus finding their way without having to remember the precise route.

The Talking Sign and other permanently mounted voice output devices, however, require standardization, costly retrofitting of existing buildings, and the possession of a suitable transceiver to detect or activate the installed devices. Retrofitting buildings with such devices is not cost effective due to their inherent inflexibility and the need for many users to justify the implementation costs. What's especially frustrating for persons with severe visual impairments is that talking signs may not reflect their travel patterns or be available at unfamiliar locations and wide open spaces. To be truly useful, talking signs would have to be almost ubiquitous and universally adopted.

GPS NAVIGATIONAL AIDS

In addition to obstacle avoidance, the ideal navigational aid also must address two other key aspects of independent travel: orientation (the ability to monitor one's position in relationship to the environment) and route guidance (the ability to determine a safe and appropriate route for reaching one's destination). As an orientation aid, the Global Positioning System (GPS) seems promising. For route guidance, a notebook computer or personal data assistant (PDA) equipped with speech input/output software, route planning software (artificial intelligence), and digital maps have been proposed (18,31,32). A voice operable, handheld GPS unit used in combination with obstacle detecting ETAs like the Laser Cane might constitute the ideal navigational aid for blind persons.

Several GPS equipped PDAs have recently become available. For example, the iQue 3600 ($600 from Garmin International Inc., Olathe, Kansas) is a handheld device that combines a PDA and mapping software with a built-in GPS receiver. The iQue 3600 uses the Palm operating system and offers a color screen and voice output turn-by-turn navigational guidance. For someone who already possesses a PDA (e.g., Palm Pilot or Microsoft's Pocket

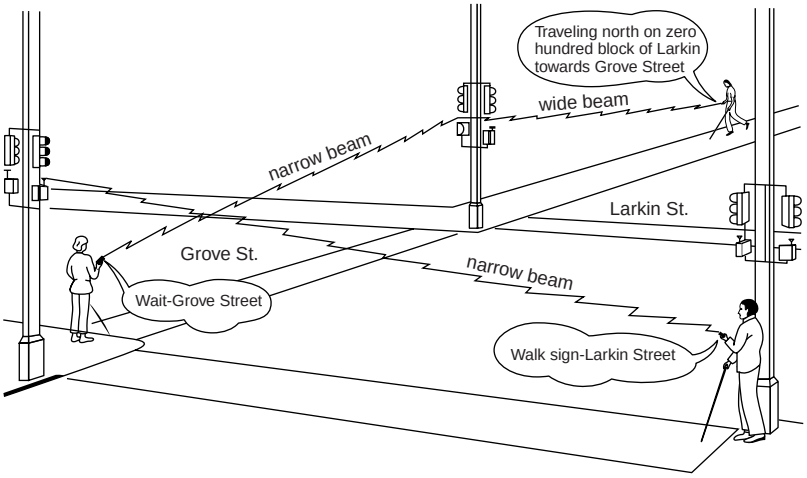


Figure 9. Talking Signs not only gives location information, but also tells the pedestrian the current status of the pedestrian cycle, aids in finding the cross-walk, and indicates the direction of the destination corner. (Reproduced from Ref. 29.)

PC), various third party software and GPS add-on units can be used.

While promising as a navigational aid for persons with severe visual impairments, GPS equipped portable PDAs (or notebook computers) have significant limitations. To fully appreciate these limitations, a brief review of how the Global Positioning System works (Fig. 10) would be apropos.

Global Positioning System (GPS) began some 30 years ago when Aerospace Corporation in Southern California studied ways to improve radio navigation systems for the military (33). Although GPS was not fully operational at the outbreak of the Persian Gulf War in January 1991, its exceptional performance in accurately locating fighting units evoked a strong demand from the military for its immediate completion.

Currently, 24 satellites of the GPS circle the earth every 12 h at a height of 20,200 km. Each satellite continuously transmits pseudorandom codes at 1575.42 and 1227.6 MHz. The orbital paths of the satellites and their altitude enable an unobstructed observer to see between five and eight satellites from any point on the earth. Signals from different visible satellites arrive at the GPS receiver with different time delays. The time delay needed to achieve coherence between the satellites' pseudorandom codes and the receiver's internally generated code equals the time-of-flight delay from a given satellite. GPS signals from at least four satellites are analyzed to determine the receiver's longitude, latitude, altitude (as measured from earth's center) and the user's clock error with respect to system time (33).

For civilian applications, position accuracy of a single channel receiver is about 100 m and its time accuracy is ~ 340 ns. Greater accuracy, usually within 1 m, can be achieved using differential GPS wherein signals from additional satellites are analyzed and/or the satellite signals are compared with and corrected by a GPS transceiver at a known fixed location (33).

At first glance, GPS signals seem fully able to meet the orientation needs of persons with severe visual impairments. The GPS signals are sufficiently accurate if combined with differential GPS and signals are immune to weather and are available at any time of the day, anywhere there's a line of sight to at least four GPS satellites. Lastly, a GPS receiver is relatively inexpensive, $< \$200$.

Unfortunately, just equipping blind persons with a voice-output GPS receiver for wayfinding outdoors is insufficient. The GPS signals are often unavailable or highly attenuated under bridges, inside natural canyons, and between tall buildings in urban areas. The altitude GPS information is generally not useful, and its longitudinal and latitudinal coordinates are useless when unaccompanied by local area maps (17). For college campuses or even major metropolitan areas, the location of major buildings and their entrances in terms of longitude and latitude coordinates are rarely available. Without these key pieces

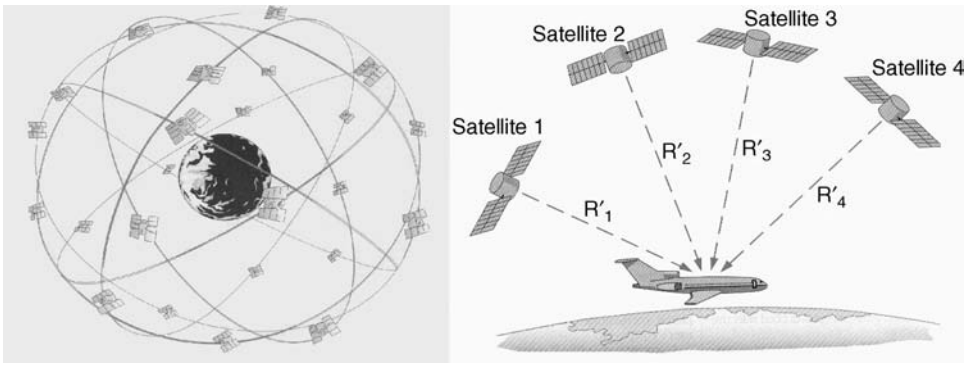


Figure 10. Synchronized signals from four satellites are analyzed by the mobile receiver to determine its precise position in three dimensions. The distances for the four satellites include an unknown error due to the inaccuracy of the receiver's clock and Doppler effects. (Reprinted from Ref. 33.)

of information, the GPS navigational aid is unable to offer directional guidance to the blind traveler.

INDOOR NAVIGATIONAL AIDS

One of the key characteristics of an ideal mobility aid is that the device reliably function indoors, outdoors, and within changing environments (Table 2). When used in combination with detailed local area maps, the GPS receiver and voice output could form the basis for a navigational aid. However, GPS signals may not be available at all times and are totally absent indoors. To function indoors, an electronic navigational aid will need to rely on some other set of electronic beacons as signposts.

For wayfinding within a large building, several investigators have borrowed the idea found in the Hansel and Gretel fairy tale about two children leaving a trail of bread crumbs to find their way home again. Instead of bread crumbs, Szeto (18) proposed placing small, low cost electronic beacons along corridors or at strategic locations (e.g., elevators, bathrooms, stairs) of buildings visited. Acting like personal pathmarkers, these radio frequency (RF) emitting electronic beacons would be detected by the associated navigational aid and guide the traveler back to a previous location or exit. To avoid confusion with other users, the electronic beacons could be keyed to work with one navigational aid.

Kulyukin et al. (34) recently studied the efficacy of using Radio Frequency Identification (RFID) tags in robot-assisted indoor navigational for the visually impaired. They described how strategically located, passive (nonpowered) RFID tags could be detected and identified by a RFID reader employing a square 200×200 mm antenna and linked to a laptop computer. In field tests, wall-mounted RFID tags responded to the spherical electromagnetic field from an RFID antenna at a distance of ~ 1.5 m. Since each tag is given a unique identifier, its location inside a building can be easily recalled and used to locate one's position inside a building.

In comparison to wall-mounted Talking Signs, the approach of Szeto (18) and Kulukin et al. (34) seems to be less costly and more flexible. Placing disposable electronic beacons in the hands of individual travelers does not require permanent retrofits of buildings, can be cost effective even for single users, and easily changes with the travel patterns of the user.

The electronic beacons and handheld electronic transceivers also should be economically feasible because they utilize a technology that's being developed for the mass market. World's largest retailer, Wal-Mart, has mandated 2008 as the year when all its suppliers must implement an RFID tracking system for their deliveries. It is likely that RFID tags, antenna, and handheld interrogators developed for inventory tracking can be adapted for use in an indoor navigational aid.

Although not yet a reality, a low cost, portable, handheld, indoor–outdoor mobility aid that embodies many of the features listed in Table 2 is clearly feasible. The needed technological infrastructure will soon be in place. For obstacle avoidance, the Laser Cane, Guide Cane, or their variants can be used. For indoor wayfinding and route guidance, the blind traveler could augment the cane or guide dog with a handheld voice output electronic navigational aid linked to strategically placed electronic beacons. For outdoor wayfinding, the blind traveler could augment the Laser Cane with a handheld mobility aid equipped with a GPS receiver, compass, local area maps, and wireless internet link.

The intelligent navigational aid just described would address the mobility needs of the blind by responding to voice commands; automatically detecting GPS signals or searching for the presence of electronic beacons; wirelessly linking to the local area network to obtain directory information; converting the GPS coordinates or the signals from electronic beacons into a specific location on a digital map; and, with the help of routing finding software, generating step-by-step directions to the desired destination.

FUTURE POSSIBILITIES

Of course, the ultimate assistive technology for overcoming the many problems of severe visual impairment would be an artificial eye. Since the mid-1990s, research by engineers, ophthalmologists, and biologists to develop a bionic eye have grown and artificial retina prototypes are nearing animal testing. An artificial eye would incorporate a small video camera to capture light from objects and transmit the image to a wallet-sized computer processor that in turn sends the image to an implant that would stimulate either the retina (35) or visual cortex (36).

Researchers at Stanford University recently announced progress toward an artificial vision system that can stimulate a retina with enough resolution to enable a visually impaired person to orient themselves toward objects, identify faces, watch television, read large fonts, and live independently (37). Their optoelectronic retinal prosthesis system is expected to stimulate the retina with resolution corresponding to a visual acuity of 20/80 by employing 2500 pixels per square millimeter. The researchers see the device as being particularly helpful for people left blind by retinal degeneration. Although such developments are exciting, tests with human subjects on practical but experimental prototypes won't likely occur for another 6–8 years (38).

What else the does the future hold in terms of assistive technology in general and mobility aids in particular? In an address to the CSUN 18th Annual Conference on Technology and Persons with Disabilities in 2003, futurist and U.S. National Medal of Technology recipient, Ray Kurzweil, presented his vision of the sweeping technological changes that he expected to take place over the next few decades (39). His comments are worthy of reflection and give cause for optimism.

With scientific and technological progress doubling every decade, Kurzweil envisions ubiquitous computers with always-on Internet connections, systems that would allow people to fully immerse themselves in virtual environments, and artificial intelligence embedded into Web sites by 2010. Kurzweil (39) expects the human brain to be fully reverse-engineered by 2020, which would result in computers with enough power to equal human intelligence. He forecasted the emergence of systems that provide subtitles for deaf people around the world, as well as listening systems geared

toward hearing-impaired users. Blind people would be able to take advantage of pocket-sized reading devices within a decade or have retinal implants that restore useful vision in 10–20 years. Kurzweil believed that people with spinal cord injuries would be able to resume fully functional lives by 2020, either through the development of exoskeletal robotic systems or techniques that bridged severed neural pathways, possibly by wirelessly transmitting nerve impulses to muscles. Even if one-half of what Kurzweil predicted became reality, the future of assistive technology for the blind is bright and an efficacious intelligent mobility aid for such persons will soon be commercially available.

BIBLIOGRAPHY

Cited References

1. Beck AF, Stern A, Uslan MM, Wiener WR, editors. Access to Mass Transit for Blind and Visually Impaired Travelers. New York: American Foundation for the Blind; 1990.
2. National Eye Institute and Prevent Blindness America®, Vision Problems in the U.S., 4th ed., 2002.
3. Arch Ophtha Imol. April 2004; 122.
4. Allen J. Electronic aids for the severely visually handicapped. CRC Crit Rev Bioeng 1971;1:137–167.
5. Servais SB. Visual Aids. In: Webster JG, Cook AM, Tompkins WJ, Vanderheiden GC, editors. Electronic Devices for Rehabilitation. New York: John Wiley & Sons, Medical; 1985. p 31–78.
6. Independent Living Aids, Inc. (No date) [Online] product catalog. Available at http://www.independentliving.com/home.asp. Accesse May 2005
7. Allen J. Linguistic-based algorithms offer practical text-to-speech systems. Speech Technol 1981;1(1):12–16.
8. Breen A. Speech synthesis models: a review. Elect Commun Eng J 1992;4(1):19–31.
9. O'Shaughnessy D. Interacting with computers by voice: Automatic speech recognition and synthesis. Proc IEEE 2003;91(9): 1272–1305.
10. Jackson G, et al. A single-chip text-to-speech synthesis device utilizing analog nonvolatile multi-level flash storage. IEEE J. Solid State Cir Nov 2002;37(11):1582–1592.
11. Thatcher J, et al. Constructing Accessible Websites, ISBN: 1904151000, New York: Glasshaus; 2002.
12. Matthews W. 13 rules for accessible web pages, August 07, 2000 of the Federal Computer Week. (No date). [Online]. Available at http:// www.fcw.com/fcw/articles/2000/0807/cov-access2-08-07-00.asp. Accessed March 2005.
13. Lazzaro JJ. Adaptive Technologies for Learning and Work Environments. 2nd ed., New York: The American Library Association; 2000.
14. Tucker A. Net surfing for those unable to see, Baltimore Sun, p. 1C. [Online] Available at http://www.baltimoresun.com/features/lifestyle/bal-to.blind16mar16,1,1345515.story?ctrack=1&cset=true. Accessed March 16, 2005.
15. Shao S. Mobility Aids For The Blind. In: Webster JG, Cook AM, Tompkins WJ, Vanderheiden JC, editors. Electronic Devices for Rehabilitation. New York: John Wiley & Sons, Medical; 1985. p. 79–100.
16. Farmer LW. Mobility Devices. In: Welsh RL, Blasch BB, editors. Foundation of Orientation and Mobility. New York: American Foundation for the Blind; 1980. p 206–209.
17. Bentzen BL. Orientation aids. In: Blasch B, Weiner W, Welsh W, editors. Foundations of Orientation and Mobility. 2nd ed. New York: American Foundation for the Blind; 1997. p 284–316.
18. Szeto AYJ. A navigational aid for persons with severe visual impairments: a project in progress. Proceeding of the 25th Annual International Conference IEEE Engineering and Medicine & Biology Society; Vol 25(2), Cancun, Mexico, Sep. 2003 p 1637–1639.
19. Nye PW, Bliss JC. Sensory aids for the blind: a challenging problem with lessons for the future. Proc IEEE 1970;58: 1878–1879.
20. Cook AM, Hssey SM. Assistive Technologies: Principles and Practice. 2nd ed., St. Louis, (MO): Mosby; 2002. p. 423–426.
21. Kay L. A sonar aid to enhance spatial perception of the blind: Engineering design and evaluation. Radio Elect Eng 1974;44(11):605–627.
22. Kuc R. Binaural sonar electronic travel aid provides vibrotactile cues for landmark, reflector motion and surface texture classification. IEEE Trans Biomed Eng Oct 2002; 49(10):1173–1180.
23. Shovel S, Ulrich I, Borenstein J. Computerized Obstacle Avoidance Systems for the Blind and Visually Impaired. In: Teodorescu HNL, Jain LC, editors. Intelligent Systems and Technologies in Rehabilitation Engineering. Boca Raton(FL): CRC Press; 2001.
24. Polaroid Corp, Ultrasonic Ranging System—Description, operation and use information for conducting tests and experiments with Polaroid's Ultrasonic Ranging System, Ultrasonic Components Group, 119 Windsor Street, Cambridge (MA).
25. National Safety Council, Pedestrian accidents, National Safety Council Accident Facts (Injury Statistics), 1998.
26. Szeto AYJ, Valerio N, Novak R. Audible pedestrian traffic signals: Part 1. Prevalence and impact. J Rehabilit R & D 1991;28(2):57–64.
27. Szeto AYJ, Valerio N, Novak R. Audible pedestrian traffic signals: Part 2. Analysis of sounds emitted. J Rehabilit R & D 1991;28(2):65–70.
28. Szeto AYJ, Valerio N, Novak R. Audible pedestrian traffic signals: Part 3. Detectability. J Rehabilit R & D 1991;28(2):71–78.
29. Farmer LW, Smith DL. Adaptive technology. In: Blasch B, Weiner W, Welsh R. Foundations of Orientation and Mobility. 2nd ed., New York: American Foundation for the Blind; 1997. p. 231–259.
30. Brabyn J, Crandall W, Gerrey W. Talking Signs®: A Remote Signage Solution for the Blind, visually Impaired and Reading Disabled, Proceeding of the 15th Annual International Conference in IEEE Engineering in Medicine & Biology Society; 1993; Vol. 15: p. 1309–1311.
31. Vogel S. A PDA-based navigational system for the blind. [Online], Available at http://www.cs.unc.edu/~vogel/IP/IP/IP_versions/IPfinal_SusanneVogel. Accessed Spring 2003.pdf.
32. Helal A, Moore SE, Ramachandran B. Drishti: An integrated navigation system for visually impaired and disabled. Procedings of the 5th International Symposiun on Wearable Computers, Zurich, Switzerland; October 2001; p. 149–155.
33. Getting IA. The Global Positioning System. IEEE Spectrum Dec. 1993;30(12):36–47.
34. Kulyukin V, Gharpure C, Nicholson J, Pavithran S. RFID in Robot-Assisted indoor Navigation for the Visually Impaired. Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems; Sept. 28–Oct. 2, 2004; Sendai, Japan, p 1979–1984.
35. Wyatt J, Rizzo J. Ocular implants for the blind. IEEE Spectrum May 1996;33(5):47–53.
36. Normann RA, Maynard EM, Guillory KS, Warren DJ. Cortical implants for the blind. IEEE Spectrum May 1996;33 (5):54–59.
37. Palanker D, Vankov A, Huie P, Baccus S. Design of a high-resolution optoelectronic retinal prosthesis. J Neural Eng 2005;2:105–120.

38. Braham R. Toward an artificial eye. IEEE Spectrum May 1996;33(5):20–21.
39. Kurzweil R. The future of intelligent technology and its impact on disabilities. J Visual Impairment Blindness Oct 2003;97(10):582–585.

Reading List

Cook AM, Hussey SM. Assistive Technologies: Principles and Practice. 2nd ed., St. Louis (MO): Mosby, Inc.; 2002. A thorough text on assistive technologies that is especially suited for the rehabilitation practitioner or those in allied health.

Smith RV, Leslie JH Jr., editors. Rehabilitation Engineering. Boca Rotan (FL): CRC Press; 1990. Contains diverse articles that should be of particular interest to practitioners in the rehabilitation field although several of the articles present definitive state-of-the-art information on rehabilitation engineering.

Webster JG, Cook AM, Tompkins WJ, Vanderheiden GC, editors. Electronic Devices for Rehabilitation. New York: John Wiley & Sons Inc. Medical; 1985. Though somewhat dated, this book offers a comprehensive overview of rehabilitation engineering and describes many of the design issues that underlie various types of assistive devices. A useful introductory text for undergraduate engineering students interested in rehabilitation.

Golledge RG. Wayfinding Behavior: Cognitive Mapping and Other Spatial Processes. Baltimore: John Hopkins University Press; 1999. A good reference that covers the cognitive issues of wayfinding behavior in blind and sighted humans.

Teodorescu HNL, Jain LC, editors. Intelligent Systems and Technologies in Rehabilitation Engineering. Boca Rotan (FL): CRC Press; 2000. A compendium of technical review articles covering intelligent technologies applied to retinal prosthesis, auditory & cochlear prostheses, upper and lower limb orthoses/prostheses, neural prostheses, pacemakers, and robotics for rehabilitation.

Yonaitis RB. Understanding Accessibility-A Guide to Achieving Compliance on Websites and Intranets, ISBN: 1-930616-03-1, HiSoftware, 2002. This is a free booklet in electronic form for complying with the Federal government's "Section 508" of the Workplace Rehabilitation Act (amendments of 1998). The book gives a brief and clear discussion of accessibility testing and how to integrate this activity into web design and related tasks.

Blasch B, Weiner W, Welsh R. Foundations of Orientation and Mobility. 2nd ed., New York: American Foundation for the Blind; 1997. A useful book for general background regarding the issues of orientation and mobility.

IEEE Spectrum, Vol. 33(5), May 1996, carries six special reports on the development of an artificial eye. Articles in this issue examine physiology of the retina, neural network signal processing, electrode array design, and sensor technology.

Journal of Visual Impairment and Blindness, Vol. 97(10), Oct. 2003, is a special issue that focused on the impact of technology on blindness.

Speech Technology, a magazine published bimonthly by AmCom Publications, 2628 Wilhite Court, Suite 100, Lexington, KY 450503. This magazine regularly covers the development and implementation of technologies that underlie speech recognition and speech generation. For example, its March/April 2005 issue contained articles on the following topics: guide to speech standards; applications of transcription; role of speech in healthcare, embedding speech into mobile devices, technology trends, new products, and speech recognition software.

See also COMMUNICATION DEVICES; ENVIRONMENTAL CONTROL; MOBILITY AIDS; VISUAL PROSTHESES.

BLOOD BANKING. See BLOOD COLLECTION AND PROCESSING.

BLOOD CELL COUNTERS. See CELL COUNTERS, BLOOD.

BLOOD COLLECTION AND PROCESSING

TERESA M. TERRY
JOSEPHINE H. COX
Walter Reed Army Institute of Research
Rockville, Maryland

INTRODUCTION

Phlebotomy may date back to the Stone Age when crude tools were used to puncture vessels to allow excess blood to drain out of the body (1). This purging of blood, subsequently known as blood letting, was used for therapeutic rather than diagnostic purposes and was practiced through to modern times. Phlebotomy started to be practiced in a more regulated and dependable fashion after the Keidel vacuum tube for the collection of blood was manufactured by Hynson, Wescott, and Dunning. The system consisted of a sealed ampoule with or without culture medium connected to a short rubber tube with a needle at the end. After insertion onto the vein, the stem of the ampoule was crushed and the blood entered the ampoule by vacuum. Although effective, the system did not become popular until evacuated blood collection systems started to be used in the mid-twentieth century. With evacuated blood collection systems came a new interest in phlebotomy and blood drawing techniques and systems. A lot of technical improvements have been made, not only are needles smaller, sharper, and sterile, they are also less painful. The improved techniques of obtaining blood samples assure more accurate diagnostic results and less permanent damage to the patient. Today, the main purpose of phlebotomy synonymous with venipuncture is to obtain blood for diagnostic testing.

Venipuncture Standards and Recent Standard Changes

The Clinical and Laboratory Standards Institute (CLSI, formerly the National Committee for Clinical Laboratory Standards, NCCLS) develops guidelines and sets standards for all areas of the laboratory (www.CLSI.org). Phlebotomy program approval as well as certification examination questions are based on these important national standards. Another agency that affects the standards of phlebotomy is the College of American Pathologists (CAP; www.CAP.org). This national organization is an outgrowth of the American Society of Clinical Pathologists (ASCP). The membership in this specialty organization is made up of board-certified pathologists only and offers, among other services, a continuous form of laboratory inspection by pathologists. The CAP Inspection and Accreditation Program do not compete with the Joint Commission on Accreditation of Health Care Organizations

(JCAHO) accreditation for health care facilities, because it was designed for pathology services only.

The CLSI have published the most current research and industry regulations on standards and guidelines for clinical laboratory procedures (2,3). The most significant changes to specimen collection are (*1*) collectors are now advised to discard the collection device without disassembling it, this reflects the Occupational Safety and Health Administration's (OSHA) mandate against removing needles from tube holders; (*2*) the standard now permits gloves to be applied just prior to site preparation instead of prior to surveying the veins; (*3*) collectors are advised to inquire if the patient has a latex sensitivity; (*4*) sharp containers should be easily accessible and positioned at the point of use; (*5*) there is a caution recommended against the use of ammonia inhalants on fainting patients in case the patient is asthmatic; (*6*) collectors must attempt to locate the median cubital vein on either arm before considering alternative veins due to the proximity of the basilica vein to the brachial artery and the median nerve; (*7*) forbids lateral needle relocation in an effort to access the basilica vein to avoid perforating or lacerating the brachial artery; (*8*) immediate release of tourniquet "if possible" upon venous access to prevent the effects of hemoconcentration from altering test results.

The Role of the Phlebotomist Today

Professionalism. Phlebotomists are healthcare workers and must practice professionalism and abide by state and federal requirements. A number of agencies have evolved offering the phlebotomist options for professional recognition (1). Certification is a process that indicates the completion of defined academic and training requirements and the attainment of a satisfactory score on a national examination. Agencies that certify phlebotomists and the title each awards include the following: American Society of Clinical Pathologists (ASCP): Phlebotomy Technician, PBT (ASCP); American Society for Phlebotomy Technology (ASPT): Certified Phlebotomy Technician, CPT (ASPT); National Certification Agency for Medical Laboratory Personnel (NCA): Clinical Laboratory Phlebotomist (CLP) (NCA); National Phlebotomy Association (NPA); Certified Phlebotomy Technician, CPT (NPA). Licensure is defined as a process similar to certification, but at the state or local level. A license to practice a specific trade is granted through examination to a person who can meet the requirements for education and experience in that field. Accreditation and approval of healthcare training programs provides an individual with an indication of the quality of the program or institution. The accreditation process involves external peer review of the educational program, including an on-site survey to determine if the program meets certain established qualification or educational standards referred to as 'essentials'. The approval process is similar to accreditation; however, programs must meet educational—standards and competencies—rather than essentials, and an on-site survey is not required.

Public Relations and Legal Considerations. The Patient's Bill of Rights was originally published in 1975 by the American Hospital Association. The document, while not legally binding, is an accepted statement of principles that guides healthcare workers in their dealings with patients. It states that all healthcare professionals, including phlebotomists, have a primary responsibility for quality patient care, while at the same time maintaining the patient's personal rights and dignity. Two rights especially pertinent to the phlebotomist are the right of the patient to refuse to have blood drawn and the right to have results of lab work remain confidential. Right of Privacy: "An individual's right to be let alone, recognized in all United States jurisdictions, includes the right to be free of intrusion upon physical and mental solitude or seclusion and the right to be free of public disclosure of private facts. Every healthcare institution and worker has a duty to respect a patient's or client's right of privacy, which includes the privacy and confidentiality of information obtained from the patient–client for purposes of diagnosis, medical records, and public health reporting requirements. If a healthcare worker conducts tests on or publishes information about a patient–client without that person's consent, the healthcare worker could be sued for wrongful invasion of privacy, defamation, or a variety of other actionable torts." In 1996, the Health Insurance Portability and Accountability Act (HIPAA) law was signed. It is a set of rules to be followed by health plans, doctors, hospitals, and other healthcare providers. Patients must be able to access their record and correct errors and must be informed of how their personal information will be used. Other provisions involve confidentiality of patient information and documentation of privacy procedures.

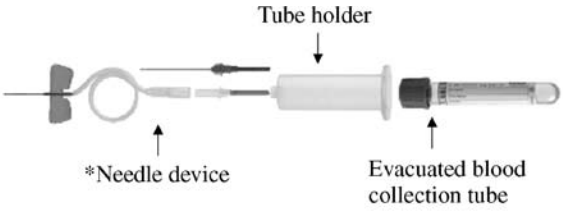
SAFETY

Universal Precautions

An approach to infection control that is mandated by federal and state laws is the so-called Universal Precautions. The guidelines for Universal Precautions are outlined by OSHA (www.OSHA.gov). According to the concept of Universal Precautions, all human blood and certain human body fluids are treated as if known to be infectious for human immunodeficiency virus (HIV), hepatitis B virus (HBV), and other blood borne pathogens. For blood collections, the use of needles with a safety device or a needle integrated into a safety device and the use of gloves is now mandatory in most institutions. Biohazard material should be disposed of in an appropriately labeled biohazard container. Needles and other sharp instruments should be disposed of in rigid puncture-resistant biohazard containers.

First Aid Procedures

Most phlebotomy programs require cardio pulmonary resuscitation (CPR) certification as a prerequisite or include it as part of the course and in the event of an emergency situation: basic First Aid Procedures should be performed by the phlebotomist. These procedures are not in the scope of this article and training needs to be performed by qualified experts.



*A butterfly with luer adapter is shown as well as a standard needle

Figure 1. Basic components of the Evacuated Blood Collection System.

BLOOD COLLECTION SYSTEM AND EQUIPMENT

Blood Collection System

The components of the Evacuated Blood Collection System are shown in Fig. 1. The system consists of the following;

Plastic evacuated collection tube: The tubes are designed to fill with a predetermined volume of blood by vacuum. The rubber stoppers are color coded according to the additive that the tube contains (see Table 1). Evacuated collection devices are supplied by many vendors worldwide. These evacuated collection devices use similar color coding systems, proprietary additives, and recommended uses. Various sizes are available.

Tube holder (single use): For use with the evacuated collection system.

Needles (also available with safety device): The gauge number indicates the bore size: the larger the gauge number, the smaller the needle bore. Needles are available for evacuated systems and for use with a syringe, single draw, or butterfly system.

Additional Materials

Tourniquet: Wipe off with alcohol and replace frequently. Nonlatex tourniquets are recommended.

Table 1. Tube Guide[a]

Tube Top Color	Additive	Inversions at Blood Collection[a]	Laboratory Use
Gold or Red/Black	Clot activator Gel for serum separation	5	Tube for serum determinations in chemistry. Blood clotting time: 30 min
Light Green or Green/Gray	Lithium heparin Gel for plasma separation	8	Tube for plasma determinations in chemistry
Red	Clot activator	5	Tube for serum determination in chemistry, serology, and immunohematology testing
Orange or Gray/Yellow	Thrombin	8	Tube for stat serum determinations in chemistry. Blood clotting occurs in < 5 min
Royal Blue	Clot activator K_2EDTA, where EDTA = ethylenediaminetetraacetic acid	5 8	Tube for trace-element, toxicology and nutritional chemistry determinations.
Green	Sodium heparin Lithium heparin	8 8	Tube for plasma determination in chemistry
Gray	Potassium oxalate/sodium fluoride Sodium fluoride/Na_2EDTA Sodium fluoride (serum tube)	8 8 8	Tube for glucose determination. Oxalate and EDTA anticoagulants will give plasma samples. Sodium fluoride is the antiglycolytic agent
Tan	K_2EDTA	8	Tube for lead determination. This tube is certified to contain $< 0.01\ \mu g \cdot mL^{-1}$ lead
Lavender	Spray-coated K_2EDTA	8	Tube for whole blood hematology determination and immunohematology testing
White	K_2EDTA with gel	8	Tube for molecular diagnostic test methods such as polymerase chain reaction (PCR) and/or DNA amplification techniques.
Pink	Spray-coated K_2EDTA	8	Tube for whole blood hematology determination and immunohematology test. Designed with special cross-match label for required patient information by the AABB[b]
Light Blue	Buffered sodium citrate (3.2%) Citrate, theophylline, adenosine, dipyridamole (CTAD)	3	Tube for coagulation determinations. The CTAD for selected platelet function assays and routine coagulation determination

[a]Reproduced from Becton Dickinson www.bd.com/vacutainer. Evacuated collection devices made by other manufacturers use similar color coding systems and additives. Recommended inversion times and directions for use are provided by each supplier.

[b]AABB = American Association of Blood Banks.

Gloves: Worn to protect the patient and the phlebotomist. Nonlatex gloves are recommended.

Antiseptics–Disinfectants: 70% isopropyl alcohol or iodine wipes (used if blood culture is to be drawn).

Sterile gauze pads: For application on the site from which the needle is withdrawn.

Bandages: Protects the venipuncture site after collection.

Disposal containers: Needles should never be broken, bent, or recapped. Needles should be placed in a proper disposal unit immediately after use.

Syringe: May be used in place of evacuated collection system in special circumstances.

Permanent marker or pen: To put phlebotomist initials, time, and date of collection on tube as well as any patient identification information not provided by test order label.

BEST SITES FOR VENIPUNCTURE

The most common sites for venipuncture are located in the antecubital (inside elbow) area of the arm (see Fig. 2). The primary vein used is the median cubical vein. The basilica and cephalic veins can be used as a second choice. Although the larger and fuller median cubital and cephalic veins of the arm are used most frequently, wrist and hand veins are also acceptable for venipuncture. Certain areas are to be avoided when choosing site such as; (*1*) Skin areas with extensive scars from burns and surgery (it is difficult to puncture the scar tissue and obtain a specimen); (*2*) the upper extremity on the side of a previous mastectomy (test results may be affected because of lymphedema); (*3*) site of a hematoma (may cause erroneous test results). If another site is not available, collect the specimen distal to the hematoma; (*4*) Intravenous therapy (IV)/blood transfusions (fluid may dilute the specimen, so collect from the opposite arm if possible); (*5*) cannula/fistula/heparin lock (hospitals have special safety and handling policies regarding these devices). In general, blood should not be drawn from an arm with a fistula or cannula without consulting the attending physician; (*6*) edematous extremities (tissue fluid accumulation alters test results).

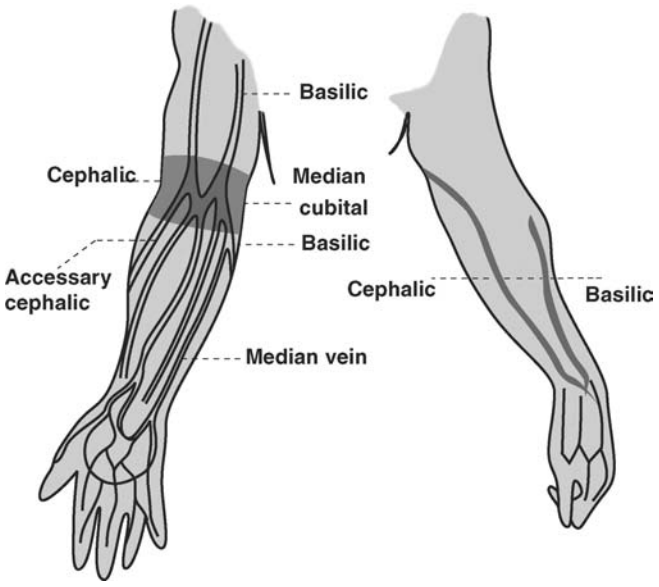


Figure 2. Venipuncture sites.

ROUTINE PHLEBOTOMY PROCEDURE

Venipuncture is often referred to as "drawing blood". Most tests require collection of a blood specimen by venipuncture and a routine venipuncture involves the following steps *Note*: The following steps were written using guidelines established by the CLSI/NCCLS (3).

1. **Prepare Order.** The test collection process begins when the physician orders or requests a test to be performed on a patient. All laboratory testing must be requested by a physician and results reported to a physician. The form on which the test is ordered and sent to the lab is called the test requisition. The requisition may be a computer-generated form or a manual form.
2. **Greet and Identify Patient.** Approach the patient in a friendly, calm manner. Identify yourself to the patient by stating you name. Provide for their comfort as much as possible. The most important step in specimen collection is patient identification. When identifying a patient, ask the patient to state their name and date of birth. Outpatients can use an identification card as verification of identity. Even if the patient has been properly identified by the receptionist, the phlebotomist must verify the patient's ID once the patient is actually called into the blood drawing area. The phlebotomist should ask for two identifiers that match the test requisition form (e.g., name and social security or name and date of birth).
3. **Verify Diet Restrictions and Latex Sensitivity.** Once a patient has been identified, the phlebotomist should verify that the patient has followed any special diet instructions or restrictions. The phlebotomist should also inquire about the patients' sensitivity to latex.

   Assemble Supplies: See the section on Blood Collection System and Equipment

   Position Patient

   A patient should be either seated or lying down while having blood drawn. The patient's arm that will be used for the venipuncture should be supported firmly and extended downward in a straight line.

4. **Apply Tourniquet.** A tourniquet is applied to increase pressure in the veins and aid in vein selection. The tourniquet is applied 3–4 in. (7.62–10.18 cm) above the intended venipuncture site. Never leave the tourniquet in place longer than 1 min.

5. **Select a Vein.** Palpate and trace the path of veins in the antecubital (inside elbow) area of the arm with the index finger. Having a patient make a fist will help make the veins more prominent. Palpating will help to determine the size, depth, and direction of the vein. The main veins in the antecubital area are the median cubical, basilica, and cephalic (see the section; Best Sites for Venipucture). Select a vein that is large and well anchored.
6. **Put on Gloves.** Properly wash hands followed by glove application.
7. **Cleanse Venipuncture Site.** Clean the site using a circular motion, starting at the center of the site and moving outward in widening concentric circles. Allow the area to air dry.
8. **Perform Venipuncture.** Grasp patients arm firmly to anchor the vein. Line the needle up with the vein. The needle should be inserted at a 15–30° angle BEVEL UP. When the needle enters the vein, a slight "give" or decrease in resistance should be felt. At this point, using a vacuum tube, slightly, with firm pressure, push the tube into the needle holder. Allow tube to fill until the vacuum is exhausted and blood ceases to flow to assure proper ratio of additive to blood. Remove the tube, using a twisting and pulling motion while bracing the holder with the thumb. If the tube contains an additive, mix it immediately by inverting it 5–10 times before putting it down.
9. **Order of Draw.** Blood tubes are drawn in a particular order to ensure integrity of each sample by lessening the chances of anticoagulants interference and mixing. The order of draw also provide a standardized method for all laboratories (3,4).

   Blood Cultures: With sodium polyanethol sulfonate anticoagulant and other supplements for bacterial growth.
   Light Blue: Citrate Tube (*Note*: When a citrate tube is the first specimen tube to be drawn, a discard tube should be drawn first). The discard tube should be a nonadditive or coagulation tube.
   Gold or Red/Black: Gel Serum Separator Tube, no additive.
   Red: Serum Tube, no additive.
   Green: Heparin Tube.
   Light Green or Green/Gray: Gel Plasma Separator Tube with Heparin.
   Lavender: EDTA Tube.
   Gray: Fluoride (glucose) Tube.

10. **Release the Tourniquet.** Once blood begins to flow the tourniquet may be released to prevent hemoconcentration.
11. **Place the Gauze Pad.** Fold clean gauze square in half or in fourths and place it directly over the needle without pressing down. Withdraw the needle in one smooth motion, and immediately apply pressure to the site with a gauze pad for 3–5 min, or until the bleeding has stopped. Failure to apply pressure will result in leakage of blood and hematoma formation. Do not bend the arm up, keep it extended or raised.
12. **Remove and Dispose of the Needle.** Needle should be disposed of immediately by placing it and the tube holder in the proper biohazard sharps container. Dispose of all other contaminated materials in proper biohazard containers.
13. **Bandage the Arm.** Examine the patients arm to assure that bleeding has stopped. If bleeding has stopped, apply an adhesive bandage over the site.
14. **Label Blood Collection Tubes.** Specimen tube labels should contain the following information: patient's full name, patient's ID number, date, time, and initials of the phlebotomist must be on each label of each tube.
15. **Send Blood Collection Tube to be Processed.** Specimens should be transported to the laboratory processing department in a timely fashion. Some tests may be compromised if blood cells are not separated from serum or plasma within a limited time.

SPECIMEN PROCESSING

Processing of blood is required in order to separate out the components for screening, diagnostic testing, or for therapeutic use. This section will concentrate primarily on processing of blood for screening purposes and diagnostic testing. An overview of the main blood processing procedures, specimen storage, and common uses for each of the components is provided in Table 2. Because there are many different blood components and many different end uses for these components, the list is not comprehensive and the reader should refer to other specialized literature for further details. The OSHA regulations require laboratory technicians to wear protective equipment (e.g., gloves, labcoat, and protective face gear) when processing specimens. Many laboratories mandate that such procedures are carried out in biosafety cabinets.

Whole Blood Processing

Because whole blood contains all but the active clotting components, it has the ability to rapidly deteriorate and all blood components are subject to chemical, biological, and physical changes. For this reason, whole blood has to be carefully handled and any testing using whole blood has to be performed as soon as possible after collection to ensure maximum stability. Whole blood is typically used for the complete blood count (CBC). The test is used as a broad screening test to check for such disorders as anemia, infection, and many other diseases (www.labtestsonline.org). The CBC panel typically includes measurement of the following: white blood, platelet and red blood cell count, white blood cell differential and evaluation of the red cell compartment by analysis of hemoglobin and hematocrit, red cell distribution width and mean corpuscular volume, and mean corpuscular hemoglobin. The CBC assays are now routinely performed with automated

Table 2. Blood Processing Procedures and Specimen Storage

Component	Processing	Short-Term Storage	Long-Term Storage	Uses
Red blood cells	Gravity and/or centrifigation	~1 month at 4 °C	Frozen up to 10 years	Transfusion
Plasma	Gravity and/or centrifigation	Use immediately	Frozen up to 7 years	Serology, diagnostics, immune monitoring source of biologics
Serum	Clotting and centrifugation	Use immediately	Frozen up to 7 years	Serology, diagnostics, immune monitoring source of biologics
Platelets	Plasma is centrifuged to enrich for platelet fraction	Five days at room temperature	Cannot be cryopreserved	Transfusion
Granulocytes	Centrifigation and separation from red blood cells	Use within 24 h	Cryopreserved in liquid nitrogen[a]	Transfusion
Peripheral blood mononuclear cells	Ficoll–hypaque separation	Use immediately	Cryopreserved in liquid nitrogen[a]	Immune monitoring, specialized expansion and reinfuison
Albumin, immune globulin, specific immune globulins, and clotting factor concentrates	Specialized processing, fractionation and separation	Not applicable	Variable	Multiple therapeutic uses

[a]Although it has been shown that cells can be stored indefinitely in liquid nitrogen, the functionality of the cells would have to be assessed and storage lengths determined for each type of use proposed.

analyzers in which capped evacuated collection devices are mixed and pierced through the rubber cap. Whole blood drawn in EDTA (lavender) tubes are usually used, although citrate (blue top) vacutainers will also work (although the result must be corrected because of dilution). Blood is sampled and diluted, and moves through a tube thin enough that cells pass by one at a time. Characteristics about the cell are measured using lasers or electrical impedance. The blood is separated into a number of different channels for different tests.

The CBC technology has expanded in scope to encompass a whole new field of diagnostics, namely, analytical cytometry. Analytical cytometry is a laser-based technology that permits rapid and precise multiparameter analysis of individual cells and particles from within a heterogeneous population of blood or tissues. Analytical cytometry is now routinely used for diagnosis of different pathological states. This technique can be used to examine cell deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) content, cell-cycle distribution, cellular apoptosis, tumor ploidy, cell function measurements (i.e., oxidative metabolism, phagocytosis), cellular biochemistry (i.e., intracellular pH, calcium mobilization, membrane potential, microviscosity, glutathione content), and fluorescence image analysis of individual blood cells. Since the blood is truly a window on what happens in the body, it is possible to use blood samples for a wide array of diagnostic and research purposes.

A second important use for whole blood is in the setting of HIV infection and treatment as a way to monitor CD4 T cell counts and percentages. These single or multiplatform tests use fresh whole blood with EDTA as anticoagulant (< 18 h after collection) samples are run with or without lysis of red cells and fixation of the lymphocytes. There are several different companies that make specialized equipment for enumeration of CD4 cell counts, the basic principle of which is to use a fluorescently tagged anti CD4 cell surface marker. The results for the CD4 or other subset are expressed as a percentage of total gated lymphocytes. In order to determine the absolute CD4 cell counts, the percent CD4 must be multiplied by an absolute lymphocyte count derived from a hematology analyzer or by an integrated volumetric analysis method (5–7).

Another common use for whole blood is for the detection of secretion of cytokines from antigen or mitogen stimulated lymphocytes. This assay can provide information on a patients T cell response to pathogens [e.g., cytomegalo virus (CMV) and Epstein Ban virus (EBV), HIV, and tuberculosis (TB)]. The technique can also be used to monitor vaccine induced responses or responses to immunotherapy. Whole blood is drawn into heparin, 0.5–1 mL of blood is stimulated with antigen of interest and costimulatory antibodies in the presence of Brefeldin-A. The latter inhibits transport of proteins from the Golgi so that secreted cytokines accumulate inside the cell. The samples are incubated at 37 °C for 6 h, after which they can be placed at 4 °C overnight or processed immediately. The samples are treated with EDTA to reduce clumping, red cells are lysed, and the sample is fixed by addition of paraformaldehyde. At this stage, the samples can be stored frozen for up to 4 months prior to detection of cell surface markers and intracellular cytokines by flow cytometry (8).

Serum Processing

Because of the ease of performing serum separation and the fact that so many tests rely on the use of serum, the technique has become routine in clinical and diagnostic laboratories. Specimens are drawn into tubes that contain no additives or anticoagulants (Table 1). Two commonly used tubes are the red serum collection tubes or commercially available serum separation tubes. Serum is obtained

by drawing the blood into a red top or the serum separator tube, allowing it to clot, and centrifuging to separate the serum. The time allowed for clotting depends on the ambient temperature and the patient sample. The typical recommendation is to allow the tube to clot for 20–30 min in a vertical position. A maximum of 1 h should suffice for all samples except those from patients with clotting disorders. Once the clot has formed, the sample is centrifuged for a recommended time of 10 min at 3000 revolutions per minute (rpm). The serum is transferred into a plastic transport tube or for storage purposes into a cryovial. Many tests collected in the serum separator tubes do not require transferring the supernatant serum unless the serum is to be stored frozen. Specimens transported by mail or stored >4 h should be separated from the clot and placed into a transport tube. Polypropylene plastic test tubes or cryovials are more resistant to breakage than most glass or plastic containers, especially when specimens are frozen. Caution needs to be observed with serum separator tubes for some tests since the analyte of interest may absorb to the gel barrier. Erroneous results may be obtained if the serum or plasma is hemolyzed, lipemic, or icteric. As eloquently described by Terry Kotrla, phlebotomist at Austin Community College these conditions cause specimen problems. (www.austin.cc.tx.us/kotrla/ PHBLab15SpecimenProcessingSum03.pdf).

1. **Hemolysis** is a red or reddish color in the serum or plasma that will appear as a result of red blood cells rupturing and releasing the hemoglobin molecules. Hemolysis is usually due to a traumatic venipuncture (i.e., vein collapses due to excessive pressure exerted with a syringe, " digging" for veins, or negative pressure damages innately fragile cells. Gross hemolysis (serum or plasma is bright red) affects most lab tests performed and the specimen should be recollected. Slight hemolysis (serum or plasma is lightly red) affects some tests, especially serum potassium and LDH (lactate dehydrogenase). Red blood cells contain large amounts of both of these substances and hemolysis will falsely elevate their measurements to a great extent. In addition to hemolysis caused during blood draw procedures, blood collection tubes (for serum and or whole blood) that are not transported correctly or in a timely fashion to the processing laboratory may be subject to hemolysis. Extremes of heat and cold in particular can cause red blood cells to lyze and sheering stresses caused by shaking of the specimens during transport may cause lysis. Finally, incorrect centrifugation temperatures and speeds may cause hemolysis of red blood cells.
2. **Icterus**. Serum or plasma can be bright yellow or even brownish due to either liver disease or damage or excessive red cell breakdown inside the body. Icterus can, like hemolysis, affect many lab tests, but unfortunately, recollection is not an option since the coloration of the serum or plasma is due to the patient's disease state.
3. **Lipemia**. Occasionally, serum or plasma may appear milky. Slight milkiness may be caused when the specimen is drawn from a nonfasting patient who has eaten a heavy meal. A thick milky appearance occurs in rare cases of hereditary lipemia.

Both for serum and plasma there are documented guidelines for specimen handling dependent on which analyte, is being examined. The kinds of tests that can be done on blood samples is ever expanding and includes allergy evaluations, cytogenetics, cytopathology, histopathology, molecular diagnostics, tests for analytes, viruses, bacteria, parasites, and fungi. Incorrect preparation, shipment, and storage of specimens may lead to erroneous results. The guidelines for preparing samples can be obtained from the CLSI (9). Diagnostic testing laboratories (e.g., Quest diagnostics) provide comprehensive lists of the preferred specimen type, transport temperature, and rejection criteria (www.questest.com).

Plasma Processing

Specimens are drawn into tubes that contain anticoagulant (Table 1.). The plasma is obtained by drawing a whole blood specimen with subsequent centrifugation to separate the plasma. Plasma can be obtained from standard blood tubes containing the appropriate anticoagulant or from commercially available plasma separation tubes. The plasma separation tubes combine spray-dried anticoagulants and a polyester material that separates most of the erythrocytes and granulocytes, and some of the lymphocytes and monocytes away from the supernatant. The result is a convenient, safe, single-tube system for the collection of whole blood and the separation of plasma. Samples can be collected, processed, and transported *in situ* thereby reducing the possibility of exposure to bloodborne pathogens at the collection and sample processing sites. One drawback is that plasma prepared in a plasma separation tube may contain a higher concentration of platelets than that found in whole blood. For plasma processing, after drawing the blood, the tube for plasma separation must be inverted five to six times to ensure adequate mixing and prevent coagulation. The recommended centrifugation time is at least 10 min at 3000 rpm. Depending on the tests required, plasma specimens may be used immediately, shipped at ambient or cooled temperatures, or may require freezing. The plasma is transferred into a plastic transport tube or for storage purposes into a cryovial. Some tests require platelet poor plasma, in which case the plasma is centrifuged at least two times.

Processing and Collection of Peripheral Blood Mononuclear Cells (PBMC) from Whole Blood

Peripheral blood mononuclear cells are a convenient source of white blood cells, T cells comprise $\sim 70\%$ of the white cell compartment and are the work-horses of the immune system. These T cells play a crucial role in protection from or amelioration of many human diseases and can keep tumors in check. The most readily accessible source of T

cells is the peripheral blood. Thus collection, processing, cryopreservation, storage, and manipulation of human PBMC are all key steps for assessment of vaccine and disease induced immune responses. The assessment of T cell function in assays may be affected by procedures beginning with the blood draw through cell separation, cryopreservation, storage, and thawing of the cells prior to the assays. Additionally, the time of blood collection to actual processing for lymphocyte separation is critical. Procedures for PBMC collection and separation are shown in Table 3 along with potential advantages and disadvantages.

When conducting cellular immunology assays, the integrity of the PBMC, especially the cellular membranes, is critical for success. A correct cellular separation process yields a pure, highly viable population of mononuclear cells consisting of lymphocytes and monocytes, minimal red blood cell and platelet contamination, and optimum functional capacity. The standard method for separation of PBMC is the use of Ficoll-hypaque gradients as originally described by Boyum in 1968 (10). A high degree of technical expertise is required to execute the procedure from accurate centrifuge rpm and careful removal of the cellular interface to avoid red cell contamination. Within the last 10 years, simplified separation ficoll procedures have largely replaced the standard ficoll method, two such procedures are outlined below (Adapted from Ref. 11) and in Fig. 3. The simplicity of these methods, superior technical reliability, reduced interperson variability, faster turnaround, and higher cell yields makes these the methods of choice.

The Cell Preparation Tube (CPT) method is described below and in greater detail in literature provided by Becton Dickinson (http://www.bd.com/vacutainer/products/molecular/citrate/procedures.asp). Vacutainer cell preparation tubes (VACUTAINER CPT tubes, Becton Dickinson) provide a convenient, single-tube system for the collection and separation of mononuclear cells from whole blood. The CPT tube is convenient to use and results in high viability of the cells after transportation. The blood specimens in the tubes can be transported at ambient temperature, as the gel forms a stable barrier between the anticoagulated blood and ficoll after a single centrifuge step. Cell separation is performed at the processing–storage laboratory using a single centrifugation step. This reduces the risk of sample contamination and eliminates the need for additional tubes and processing reagents. In many instances, and in particular when biosafety level 2 (BL2) cabinets are not available on site, the CPT method is useful because the centrifugation step can be done on site and the remaining processing steps can be performed after shipment to a central laboratory within the shortest time possible, optimally within 8 h. The central laboratory can complete cell processing in a BL2 cabinet and set up functional assays or cryopreserve the samples as needed.

Centrifuge speed is critical for PBMC processing. The centrifugal force is dependent on the radius of the rotation of the rotor, the speed at which it rotates, and the design of the rotor itself. Centrifugation procedures are given as xg measures, since rpm and other parameters will vary with the particular instrument and rotor used. The rpms may be calculated using the following formula where $r =$ radius of rotor $g =$ gravity; $g = 1.12\,r\,(\text{rpm}/1000)^2$. This conversion can be read-off a nomogram chart available readily online or in centrifuge maintenance manuals. Typically laboratory centrifuges can be programmed to provide the correct rpm.

Protocol 1. Separation of PBMC Using CPT Tubes

1. **Materials and Reagents:** Vacutainer CPT tubes (Becton Dickinson); Sterile Phosphate Buffered Saline (PBS) without Ca^+ and Mg^+, supplemented with antibiotics (Penicillin and Streptomycin); Sterile RPMI media containing 2% fetal bovine serum (FBS) and supplemented with antibiotics.

   The CPT tubes are sensitive to excessive temperature fluctuations, resulting in deterioration of the gel and impacting successful cellular separation. This problem is particularly serious in tropical countries where ambient storage temperatures may be $> 25\,^{\circ}\text{C}$. Following PBMC separation, one

Table 3. Stages and Variables in the Separation of PBMC from Whole Blood

Procedure/Technology	Alternatives	Advantages	Disadvantages
PBMC collection	Heparin	Greater cellular stability than EDTA	Impacts DNA isolation. Plasma from whole blood cannot be used for PCR based assays[a]
	EDTA		Time dependent negative impact on T cell responses
	Sodium Citrate	Greater cellular stability than EDTA	
PBMC separation	Standard Ficoll		Technically challenging Time consuming
	CPT	Rapid Technically easy and less inter-person variability Blood is drawn into same tube that is used for separation	Subject to temperature fluctuations manifested by gel deterioration and contamination in PBMC fraction.
	Accuspin/Leucosep	Rapid Technically easy and less inter-person variability	

[a]The inhibitory effects of heparin on DNA isolation can be removed by incubation of plasma or other specimens with silicon glass beads or by heparinase treatment prior to DNA extraction.

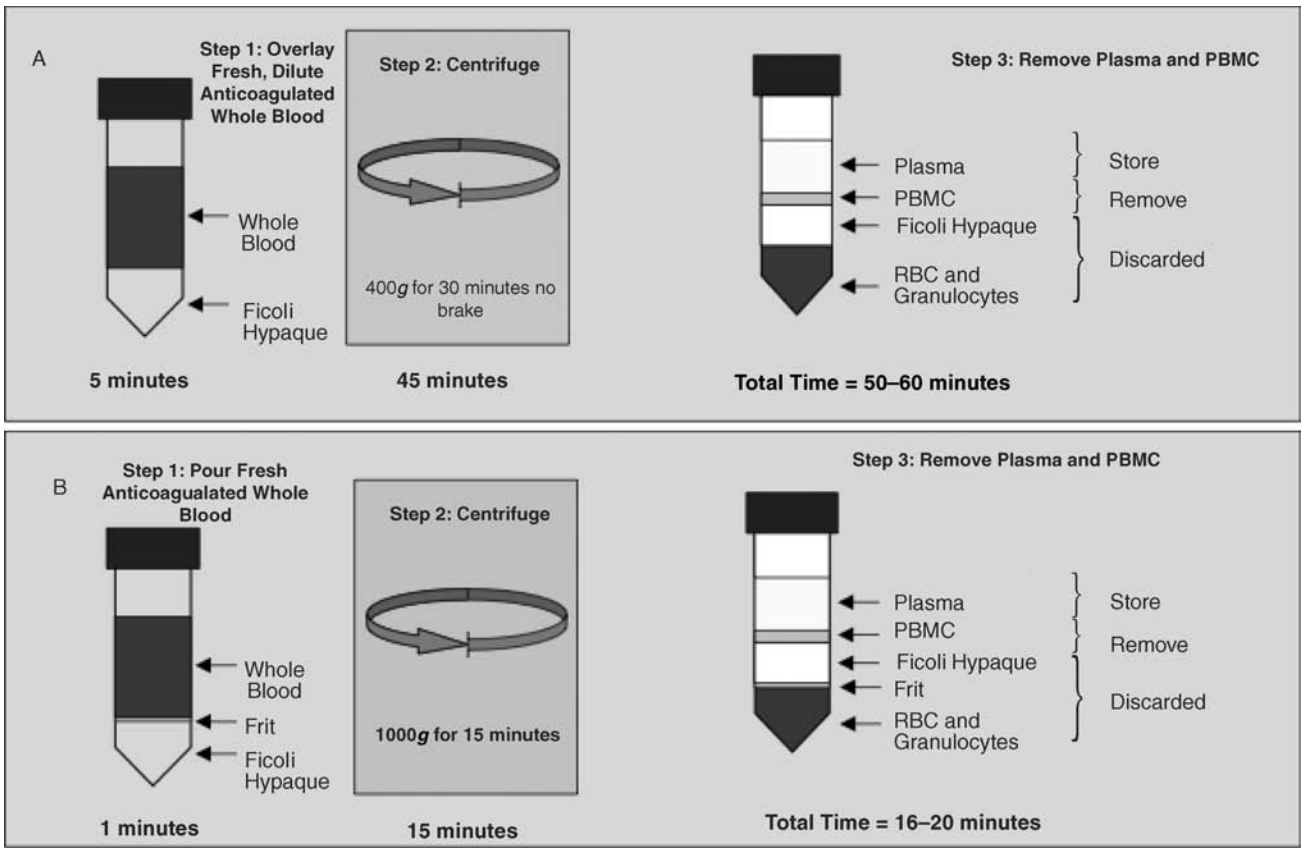


Figure 3. Gradient separation of peripheral blood mononuclear lymphocytes (PBMC). (a) The standard ficoll gradient method. (b) The Accuspin or Leucosep method. RBC = red blood cells, g = gravity. (Graphic courtesy of Greg Khoury and Clive Gray, National Institute for Communicable Diseases, Johannesburg, South Africa.)

may macroscopically observe the presence of gel spheres in the cellular layer, which are very difficult to distinguish from the actual PBMC. This has been observed after storage at temperatures $> 25\,^{\circ}$C. Where possible, the tubes should be stored at no $> 25\,^{\circ}$C. Once the tubes have blood drawn into them, an attempt should be made to keep them at temperatures of between 18 and 25 °C. Blood filled CPT should under no circumstances be stored on ice or next to an ice pack. It is recommended that they are separated from any ice-packs by bubble wrap or other type of insulation within a cooler so that the temperature fluctuations are kept to a minimum.

2. **Method:** (a) Specimens should be transported to the laboratory as soon as possible after collection. The manufacturer recommends the initial centrifugation to separate the lymphocytes be within 2 h. The samples may then be mixed by inversion and the processing completed preferably within 8 h after centrifugation. If there is a significant time delay, the specimens should be put into a cooler box and transported at room temperature (18–25 °C). (b) Spin tubes at room temperature (18–25 °C) in a horizontal rotor (swinging bucket head) for a minimum of 20 min and maximum of 30 min at $\sim 400\,g$. The brake is left off to assure that the PBMC layer is not jarred or disturbed while the centrifuge rotors are being mechanically halted. (c) Remove the tubes from the centrifuge and pipete the entire contents of the tube above the gel into a 50 mL tube. This tube will now contain both PBMC and undiluted plasma. An additional centrifugation step will allow removal of undiluted plasma if desired. Wash each CPT tube with 5 mL of PBS/1% Penicillin/Streptomycin (Pen/Strep). This wash step will remove cells from the top of the gel plug. Combine with cells removed from tube. This wash increases yield of cells by as much as 30–40%. (d) Spin down this tube at $300\,g$ for 15–20 min at room temperature with the brake on. (e) The PBMC pellet is resuspended in RPMI, 2% FBS and washed one more time to remove contaminating platelets. The PBMC are counted and cryopreserved or used as required.

3. **Separation of PBMC Using Accuspin or Leucosep Tubes.** More recently, the Leucosep and Accuspin tube have become available. Further information on the Leucosep is available at www.gbo.com and for the Accuspin at www.sigmaaldrich.com The principle of these tubes is the

same. The tube is separated into two chambers by a porous barrier made of highly transparent polypropylene (the frit). This biologically inert barrier allows elimination of the laborious overlaying of the sample material over Ficoll. The barrier allows separation of the sample material added to the top from the separation medium (ficoll added to the bottom). Figure 3 shows a comparison of the standard ficoll method and the Accuspin or Leucosep method. The tubes are available in two sizes and may be purchased with or without Ficoll. There is an advantage of buying the tubes without Ficoll because they can be stored at room temperature rather than refrigerated. This may be an important problem if cold space is limiting or cold chain is difficult. The expiration date of the Ficoll will not affect the tube expiration. The following procedure describes the separation procedure for Leucosep tubes that are not prefilled with Ficoll-hypaque. The Accuspin procedure is virtually identical. Note that whole blood can be diluted 1:2 with balanced salt solution. While this dilution step is not necessary, it can improve the separation of PBMC and enhance PBMC yield. The procedure is carried out using aseptic technique.

Protocol 2: Separation of PBMC Using Accuspin or Leucosep Tubes

1. Warm-up the separation medium (Ficoll-hypaque) to room temperature protected from light.
2. Fill the Leucosep tube with separation medium: 3 mL for the 14 mL tube and 15 mL for the 40 mL tube.
3. Close the tubes and centrifuge at 1000 *g* for 30 s at room temperature.
4. Pour the whole blood or diluted blood into the tube: 3–8 mL for the 14 mL tube and 15–30 mL for the 50 mL tube.
5. Centrifuge for 10 min at 1000 *g* or 15 min at 800 *g* in a swinging bucket rotor, with the centrifuge brake off. The brake is left off to assure that the PBMC layer is not jarred or disturbed while the centrifuge rotors are being mechanically halted.
6. After centrifugation, the sequence of layers from top to bottom should be plasma and platelets; enriched PBMC fraction; Separation medium; porous barrier; Separation medium; Pellet (erythrocytes and granulocytes).
7. Plasma can be collected to within 5–10 mm of the enriched PBMC fraction and further processed or stored for additional assays.
8. Harvest the enriched PBMC and wash with 10 mL of PBS containing 1% Pen/Strep and centrifuge at 250 *g* for 10 min.
9. The PBMC pellet is resuspended in RPMI, 2% FBS and washed one more time to remove contaminating platelets. The PBMC are counted and cryopreserved or used as required.

1. **Specimen Rejection Criteria**

Incomplete or inaccurate specimen identification.

Inadequate volume of blood in additive tubes (i.e., partially filed coagulation tube) can lead to inappropriate dilution of addition and blood.

Hemolysis (i.e., potassium determinations)

Specimen collected in the wrong tube (i.e., end product is serum and test requires plasma).

Improper handling (i.e., specimen was centrifuged and test requires whole blood).

Insufficient specimen or quantity not sufficient (QNS).

For PBMC, the rejection criteria are not usually evaluated at the time of draw due to the complexity of the tests performed. However, a minimum of 95% viability would be expected after PBMC separation unless the specimens have been subjected to heat or other adverse conditions (see note below).

The optimal time frame between collection of blood sample to processing, separation and cryopreservation of PBMC should be < 8 h or on the same day as collection. It is not always feasible to process, separate and cryopreserve PBMC within 8 h when samples are being shipped to distant processing centers. Under these conditions, PBMC left too long in the presence of anticoagulants or at noncompatible temperatures, adversely affect PBMC function and causes changes which affect the PBMC separation process (11).

There have been significant revisions to the procedures for the handling and processing of blood specimens; specimens for potassium analysis should not be recentrifuged because centrifugation may cause results to be falsely increased; the guidelines recommend that serum or plasma exposed to cells in a blood-collection tube prior to centrifugation should not exceed 2 h; storage recommendations for serum–plasma may be kept at room temperature up to 8 h, but for assays not completed within 8 h, refrigeration is recommended (2–8 °C), if the assay is not completed within 48 h serum–plasma should be frozen at or below −20 °C.

2. **Disclaimer.** The opinions or assertions contained herein are the private views of the author, and are not to be construed as official, or as reflecting true views of the Department of the Army or the Department of Defense.

BIBLIOGRAPHY

Cited References

1. McCall RE, Tankersley CM. Phlebotomy Essentials. Philadelphia: J.B. Lippincott; 1993.
2. Ernst DJ, Szamosi DI. 2005. Medical Laboratory Observer Clinical Laboratory, Specimen-collection standards complete major revisions, Available at www.mlo-online.com, Accessed 2005 Feb.
3. Arkin CF, et al. Procedures for the Collection of Diagnostic Blood Specimens by Venipuncture; CLSI (NCCLS) Approved

Standard – 5th ed. H3-A5, Vol 23, Number 32. NCCLS, Wayne (PA).
4. Becton, Dickinson and Company. 2004, BD Vacutainer Order of Draw for Multiple Tube Collections. Available at www.bd.com/vacutainer.
5. Centers for Disease Control and Prevention (CDC). Revised Guidelines for performing CD4+ T-cell determinations in persons infected with human immunodeficiency virus (HIV). MMWR. 1997;46(No.RR-2):1–29.
6. Deems D, et al. 1994, FACSCount White paper. Becton Dickenson. Available at www.bdbiosciences.com/immunocytometry systems/ whitepapers/pdf/FcountWP.pdf.
7. Dieye TN, et al. Absolute CD4 T-Cell Counting in Resource-Poor Settings: Direct Volumetric Measurements Versus Bead-Based Clinical Flow Cytometry Instruments. J Acquir Immune Defic Syndr 2005;39:32–37.
8. Maino VC, Maecker HT. Cytokine flow cytometry: a multiparametric approach for assessing cellular immune responses to viral antigens. Clin Immunol. 2004;110:222–231.
9. Wiseman JD et al. Procedures for the Handling and Processing of Blood Specimens; CLSI (NCCLS) Approved Guideline. 2nd ed. H18-A2. Vol. 19 Number 21. 1999. NCCLS, Wayne, PA
10. Boyum A. Isolation of mononuclear cells and granulocytes from human blood. Scand J Clin Lab Invest 1968;21:77–89.
11. Cox J et al. Accomplishing cellular immune assays for evaluation of vaccine efficacy. In: Hamilton RG, Detrich B, Rose NR; Manual Clinical Laboratory Immunology 6th ed. Washington (DC): ASM Press; 2002. Chapt. 33. pp 301–315.

See also ANALYTICAL METHODS, AUTOMATED; CELL COUNTERS, BLOOD; DIFFERENTIAL COUNTS, AUTOMATED.

BLOOD FLOW. See BLOOD RHEOLOGY; HEMODYNAMICS.

BLOOD GAS MEASUREMENTS

AHMAD ELSHARYDAH
RANDALL C. CORK
Louisiana State University
Shreveport, Louisiana

INTRODUCTION

Blood gas measurement–monitoring is essential to monitor gas exchange in critically ill patients in the intensive care units (1,2), and "standard of care" monitoring to deliver general anesthesia (3). It is a cornerstone in the diagnosis and management of the patient's oxygenation and acid–base disorders (4). Moreover, it may indicate the onset or culmination of cardiopulmonary problems, and may help in evaluating the effectiveness of the applied therapy. Numerous studies and reports have shown the significance of utilizing blood gas analyses in preventing serious oxygenation and acid–base problems. This article gives a summarized explanation of the common methods and instruments used nowadays in blood gas measurements in clinical medicine. This explanation includes a brief history of the development of these methods and instruments, the principles of their operation, a general description of their designs, and some of their clinical uses, hazards, risks, limitations, and finally the direction in the future to improve these instruments or to invent new ones. Blood gas measurement in clinical medicine can be classified into two major groups: (1) Noninvasive blood gas measurement, which includes blood oxygen–carbon dioxide measurement–monitoring by using different types of pulse oximeters (including portable pulse oximeters), transcutaneous oxygen partial pressure–carbon dioxide partial pressure (PO_2/PCO_2) monitors, intrapartum fetal pulse oximetry, cerebral oximetry, capnometry, capnography, sublingual capnometry, and so on; (2) invasive blood gas measurement, which involves obtaining a blood sample to measure blood gases by utilizing blood gas analyzers (in a laboratory or by using a bedside instrument), or access to the vascular system to measure/monitor blood gases. Examples include, but not limited to, mixed venous oximetry (SvO_2) monitoring by utilizing pulmonary artery catheter or jugular vein (SvO_2) measurement (5); continuous fibroptic arterial blood gas monitoring, and so on. In this article, we will talk about some of these methods; others have been mentioned in other parts of this encyclopedia.

BASIC CONCEPTS IN INVASIVE AND NONINVASIVE BLOOD GAS MEASUREMENTS

The Gas Partial Pressure

Gases consist of multiple molecules in rapid, continuous, random motion. The kinetic energy of these molecules generate a force as the molecules collide with each other and bounce from one surface to another. The force per unit area of a gas is called pressure, and can be measured by a device called a manometer. In a mixture of gases (e.g., a mixture of O_2, CO_2, and water vapor), several types of gas molecules are present within this mixture, and each individual gas (e.g., O_2 or CO_2) in the mixture is responsible for a portion of the total pressure. This portion of pressure is called partial pressure (P). According to Dalton's law, the total pressure is equal to the sum of partial pressures in a mixture of gases. Gases dissolve freely in liquids, and may or may not react with the liquid, depending on the nature of the gas and the liquid. However, all gases remain in a free gaseous phase to some extent within the liquid. Gas dissolution in liquids is a physical, not chemical, process. Therefore, gases (e.g., CO_2, O_2) dissolved in liquid (blood) exist in two phases: liquid and gaseous phase. Henry's law states that the partial pressure of a gas in the liquid phase equilibrates with the partial pressure of that gas in the gaseous phase (6,7).

BLOOD GAS ELECTRODES

Basic Electricity Terms

Electricity is a form of energy resulting from the flow of electrons through a substance (conductor). Those electrons flow from a negatively charged pole called Cathode, which has an excess of stored electrons, to a positively charged pole called Anode, which has a relative shortage

of electrons. The potential is the force responsible for pumping these electrons between the two poles. The greater the difference in electron concentration between these two poles, the greater is the potential. Volt is the potential measurement unit. The electrical current is the actual flow of electrons through a conductor. Ampere (amp) is the unit of measurement for the electrical current. Conductors display different degree of electrical resistance to the flow of the electrical current. The unit of the electrical resistance is ohm (Ω). Ohm's law states: voltage = current × resistance.

The Principles of Blood Gas Electrodes

Blood gas electrodes are electrochemical devices used to measure directly pH and blood gases. These blood gas electrodes use electrochemical cells. The electrochemical cell is an apparatus that consists of two electrodes placed in an electrolyte solution. These cells usually incorporated together (one or more cells) to form an electrochemical cell system. These systems are used to measure specific chemical materials (e.g., PO_2, PCO_2 and pH). The basic generic blood gas electrode consists of two electrode terminals, which are also called half-cells: one is called the working half-cell where the actual chemical analysis occurs, or electrochemical change is taken place; and the other one is called the reference half-cell. The electrochemical change occurring on the working terminal is compared to the reference terminal, and the difference is proportional to the amount of blood gas in the blood sample (6,7).

PO_2 Electrode

The PO_2 electrode basically consists of two terminals (1). The cathode, which usually made of platinum (negatively charged) and (2) the anode, which usually made of silver–sliver chloride (positively charged). How does this unit measure PO_2 in the blood sample? As shown in Fig. 1, the electricity source (battery or wall electricity) supplies the platinum cathode with energy (voltage of ~700 mV). This voltage attracts oxygen molecules to the cathode surface, where they react with water. This reaction consumes four electrons for every oxygen molecule reacts with water and produces four hydroxyl ions. The consumed four electrons, in turn, are replaced rapidly in the electrolyte solution as silver and chloride react at the anode. This continuous reaction leads to continuous flow of electrons from the anode to the cathode (electrical current). This electrical current is measured by using an ammeter (electrical current flow meter). The current generated is in direct proportion to the amount of dissolved oxygen in the blood sample, which in direct proportion to PO_2 in that sample.

Oxygen Polarography

The electrical current and PO_2 have a direct (linear) relationship when a specific voltage is applied to the cathode. Therefore, a specific voltage must be identified, to be used in PO_2 analysis. The polarogram is a graph that shows the relationship between voltage and current at a constant PO_2. As shown in Fig. 2, when the negative voltage applied to the cathode is increased, the current increases initially, but soon it becomes saturated. In this plateau region of the polarogram, the reaction of oxygen at the cathode is so fast that the rate of reaction is limited by the diffusion of oxygen to the cathode surface. When the negative voltage is further increased, the current output of the electrode increases rapidly due to other reactions, mainly, the reduction of water to hydrogen. If a fixed voltage in the plateau region (e.g., −0.7 V) is applied to the cathode, the current output of the electrode can be linearly calibrated to the dissolved oxygen. Note that the current is proportional not to the actual concentration, but to the activity or equivalent partial pressure of dissolved oxygen. A fixed voltage between −0.6 and −0.8 V is usually selected as the

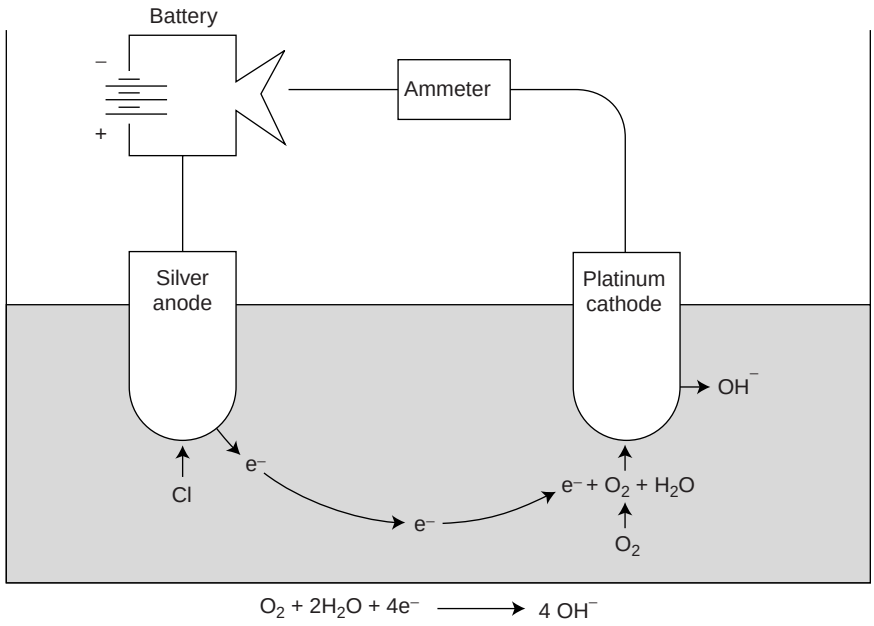


Figure 1. PO_2 electrode.

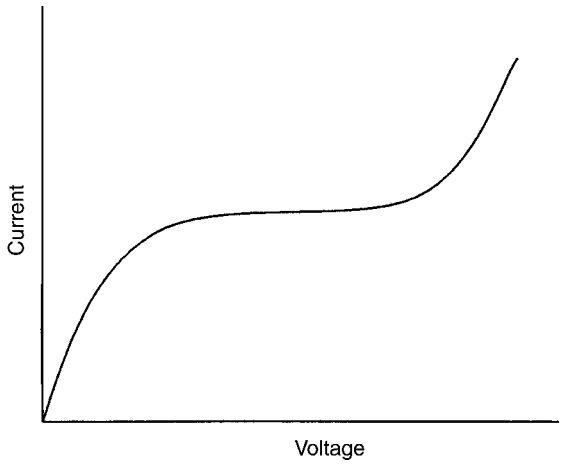


Figure 2. Polarogram.

polarization voltage when using Ag/AgCl as the reference electrode.

pH Electrode

The pH electrode uses voltage to measure pH, rather than actual current as in PO_2 electrode. It compares a voltage created through the blood sample (with unknown pH) to known reference voltage (in a solution with known pH). To make this possible, the pH electrode basically needs four electrode terminals (Fig. 3), rather than two terminals (as in the PO_2 electrode). Practically, one common pH-sensitive glass electrode terminal between the two solutions is adequate. This glass terminal allows the hydrogen ions to diffuse into it from each side. The difference in the hydrogen ions concentration across this glass terminal creates a net electrical potential (voltage). A specific equation is used to calculate the blood sample pH, using the reference fluid pH, the created voltage, and the fluid temperature.

PCO_2 Electrode

The PCO_2 electrode is a modified pH electrode. There are two major differences between this electrode and the pH electrode. The first difference is that in this electrode, the blood sample comes in contact with a CO_2 permeable membrane (such as Teflon, Silicone rubber), rather than a pH-sensitive glass (in the pH electrode), as shown in Fig. 4. The CO_2 from the blood sample diffuses via the CO_2 permeable (silicone) membrane into a bicarbonate solution. The amount of the hydrogen ions produced by the hydrolysis process in the bicarbonate solution is proportional to the amount of the CO_2 diffused through the silicone membrane. The difference in the hydrogen ions concentration across the pH-sensitive glass terminal creates a voltage. The measured voltage (by voltmeter) can be converted to PCO_2 units. The other difference is that the CO_2 electrode has two similar electrode terminals (silver–silver chloride). However, the pH electrode has two different electrode terminals (silver–silver chloride and mercury–mercurous chloride).

BLOOD GAS PHYSIOLOGY (8,9)

Oxygen Transport

Oxygen is carried in the blood in two forms: A dissolved small amount and a much bigger, more important component combined with hemoglobin. Dissolved oxygen plays a small role in oxygen transport because its solubility is so low, 0.003 mL O_2/100 mL blood per mmHg (133.32 Pa). Thus, normal arterial blood with a PO_2 of $\sim$100 mmHg (13332.2 Pa) contains only 0.3 mL of dissolved oxygen per 100 mL of blood, whereas $\sim$20 mL is combined with hemoglobin. Hemoglobin consists of heme, an iron–porphyrin compound, and globin, a protein that has four polypeptide chains. There are two types of chains, alpha and beta, and differences in their amino acid sequences give rise to different types of normal and abnormal human hemoglobin, such as, hemoglobin F (fetal) in the newborn, and hemoglobin S in the sickle cell anemia patient. The combination of oxygen (O_2) with hemoglobin (Hb) (to form oxyhemoglobin–HbO_2) is an easily reversible. Therefore, blood is able to transport large amounts of oxygen.

The relationship between the partial pressure of oxygen and the number of binding sites of the hemoglobin that have oxygen attached to it, is known as the oxygen dis-

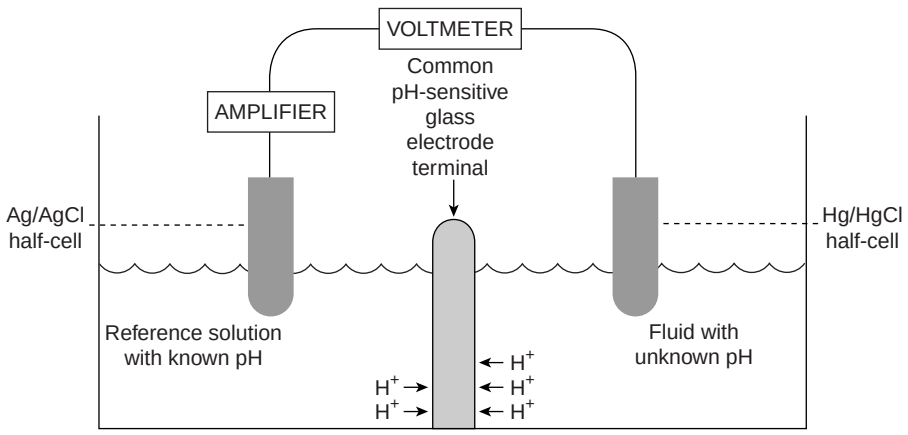


Figure 3. pH electrode.

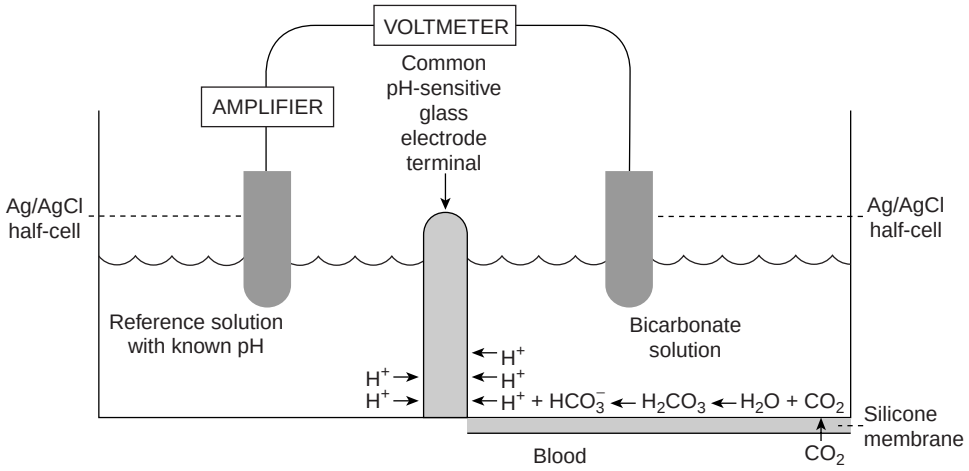


Figure 4. PCO_2 electrode.

sociation curve (Fig. 5). Each gram of pure hemoglobin can combine with 1.39 mL of oxygen, and because normal blood has ~15 g Hb/100 mL, the oxygen capacity (when all the binding sites are full) is ~20.8 mL O_2/100 mL blood. The total oxygen concentration of a sample of blood, which includes the oxygen combined with Hb and the dissolved oxygen, is given by $(Hb \times 1.36 \times SaO_2) + (0.003 \times PaO_2)$ Hb is the hemoglobin concentration.

The characteristic shape of the oxygen dissociation curve has several advantages. The fact that the upper portion is almost flat means that a fall of 20–30 mmHg in arterial PO_2 in a healthy subject with an initially normal value (e.g., ~100 mmHg or 13332.2 Pa) causes only a minor reduction in arterial oxygen saturation. Another consequence of the flat upper part of the curve is that loading of oxygen in the pulmonary capillary is hastened. This results from the large partial pressure difference between alveolar gas and capillary blood that continues to exist even when most of the oxygen has been loaded. The steep lower part of the oxygen dissociation curve means that considerable amounts of oxygen can be unloaded to the peripheral tissues with only a relatively small drop in capillary PO_2. This maintains a large partial pressure difference between the blood and the tissues, which assists in the diffusion process. Various factors affect the position of the oxygen dissociation curve, as shown in Fig. 5. It is shifted to the right by an increase of temperature, hydrogen ion concentration, PCO_2, and concentration of 2,3-diphosphoglycerate in the red cell. A rightward shift indicates that the affinity of oxygen for hemoglobin is reduced. Most of the effect of the increased PCO_2 in reducing the oxygen affinity is due to the increased hydrogen concentration. This is called the Bohr effect, and it means that as peripheral blood loads carbon dioxide, the unloading of oxygen is assisted. A useful measure of the position of the dissociation curve is the PO_2 for 50% oxygen saturation; this is known as the $P50$. The normal value for human blood is ~27 mmHg (3599.6 Pa).

Carbon Dioxide Transport

Carbon dioxide is transported in the blood in three forms: dissolved, as bicarbonate, and in combination with proteins such as carbamino compounds (Fig. 6). Dissolved carbon dioxide obeys Henry's law (as mentioned above). Because carbon dioxide is some 24 times more soluble than oxygen in blood, dissolved carbon dioxide plays a much more significant role in its carriage compared to oxygen. For example, ~10% of the carbon dioxide that evolves into the alveolar gas from the mixed venous blood comes from the dissolved form. Bicarbonate is formed in blood by the following hydration reaction:

$$CO_2 + H_2O \leftrightarrow H_2CO_3 \leftrightarrow H^+ + HCO_3^-$$

The hydration of carbon dioxide to carbonic acid (and vice versa) is catalyzed by the enzyme carbonic anhydrase (CA), which is present in high concentrations in the red cells, but is absent from the plasma. However, some carbonic anhydrase is apparently located on the surface of the endothelial cells of the pulmonary capillaries. Because of the presence of carbonic anhydrase in the red cell, most of the hydration of

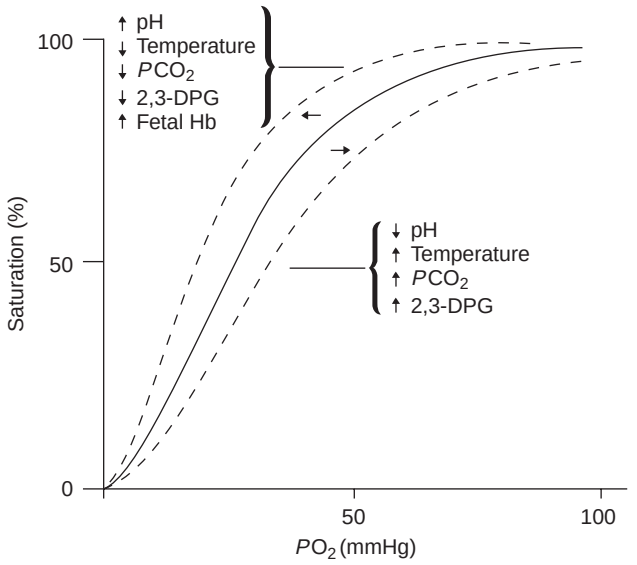


Figure 5. Oxygen dissociation curve and the effects of different factors on it.

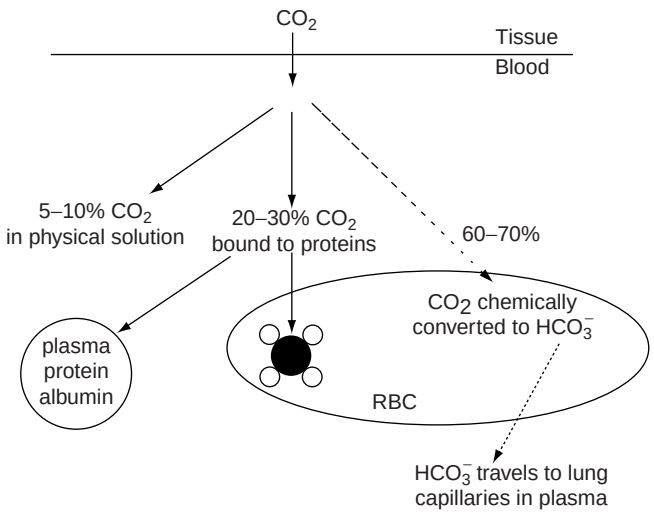


Figure 6. Carbon dioxide transport in blood.

carbon dioxide occurs there, and bicarbonate ion moves out of the red cell to be replaced by chloride ions to maintain electrical neutrality (chloride shift). Some of the hydrogen ions formed in the red cell are bound to Hb, and because reduced Hb is a better proton acceptor than the oxygenated form, deoxygenated blood can carry more carbon dioxide for a given PCO_2 than oxygenated blood can. This is known as the Haldane effect. Carbamino compounds are formed when carbon dioxide combines with the terminal amine groups of blood proteins. The most important protein is the globin of hemoglobin. Again, reduced hemoglobin can bind more carbon dioxide than oxygenated hemoglobin, so the unloading of oxygen in peripheral capillaries facilitates the loading of carbon dioxide, whereas oxygenation has the opposite effect. The carbon dioxide dissociation curve, as shown in Fig. 7, is the relationship between PCO_2 and total carbon dioxide concentration. Note that the curve is much more linear in its working range than the oxygen dissociation curve, and also that, as we have seen, the lower the saturation of hemoglobin with oxygen, the larger the carbon dioxide concentration for a given PCO_2.

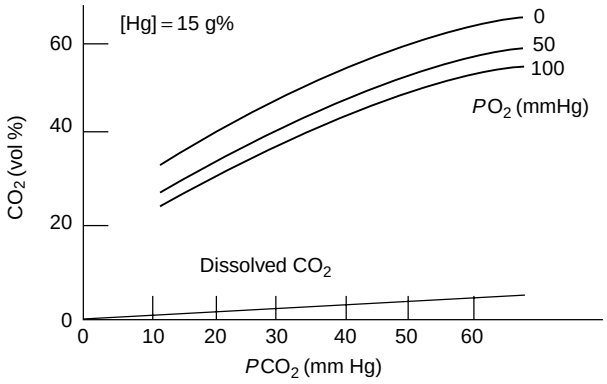


Figure 7. The carbon dioxide dissociation curve showing the effect of PO_2 variations.

OXIMETRY

Historical Development

Oximetry has its origins in the early 1860s (10), when Felix Hoppe-Seyler described the hemoglobin absorption of light using the spectroscope. He demonstrated that the light absorption was changed when blood was mixed with oxygen, and that hemoglobin and oxygen formed a compound called oxyhemoglobin. Soon after, George Gabriel Stokes reported that hemoglobin was in fact the carrier of oxygen in the blood. In 1929, Glen Allan Millikan (11), an American physiologist, began construction of a photoelectric blood oxygen saturation meter, which, used to measure color changes over time when desaturated hemoglobin solutions were mixed with oxygen solutions in an experimental setting. The use of photoelectric cells later proved to be crucial to the development of oximeters. In 1935, Kurt Kramer demonstrated, for the first time, *in vivo* measurement of blood oxygen saturation in animals. The same year, Karl Matthes introduced the ear oxygen saturation meter. This was the first instrument able to continuously monitor blood oxygen saturation in humans. In 1940, J.R. Squire introduced a two-channel oximeter that transmitted red and infrared (IR) light through the web of the hand. In 1940, Millikan and colleagues developed a functioning oximeter, and introduced the term "oximeter" to describe it. The instrument used an incandescent, battery-operated light and red and green filter. In 1948, Earl Wood of the Mayo Clinic made several improvements to Millikan's oximeter, including the addition of a pressure capsule. Then, in the 1950s, Brinkman and Zijlstra of the Netherlands developed the reflectance oximetry. However, oximetry did not fully achieve clinical applicability until the 1970s.

Principles of Operation

It is important to understand some of the basic physics principles that led to the development of oximetry and pulse oximetry. This is a summary of these different physics principles and methods (7,12).

Spectrophotometry. The spectroscope is a device which was used initially to measure the exact wavelengths of light emitted from a light generator (bunsen burner) (10). Each substance studied with the spectroscope has its unique light emission spectrum, in other words, each substance absorbed and then emitted light of different wavelengths. The graph of the particular pattern of light absorption–emission of sequential light wavelengths called the absorption spectrum. Figure 8 reveals the absorption spectra of common forms of hemoglobin.

Colorimetry. Colorimetry is another method of qualitative analysis (10). In this method, the color of known substance is compared of that of unknown one. This method is not highly exact, because it depends on visual acuity and perception.

Photoelectric Effect. The photoelectric effect is the principle behind spectrophotometry. It is defined as the ability of light to release electrons from metals in proportion to the

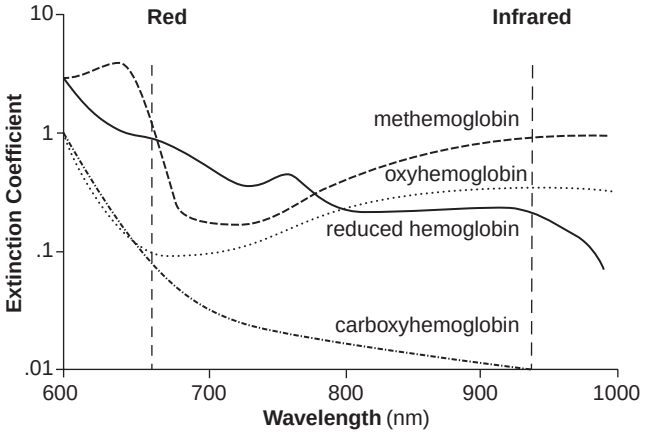


Figure 8. Absorption spectra of common forms of hemoglobin. Absorption spectra of oxyhemoglobin, deoxyhemoglobin, methemoglobin, carboxyhemoglobin.

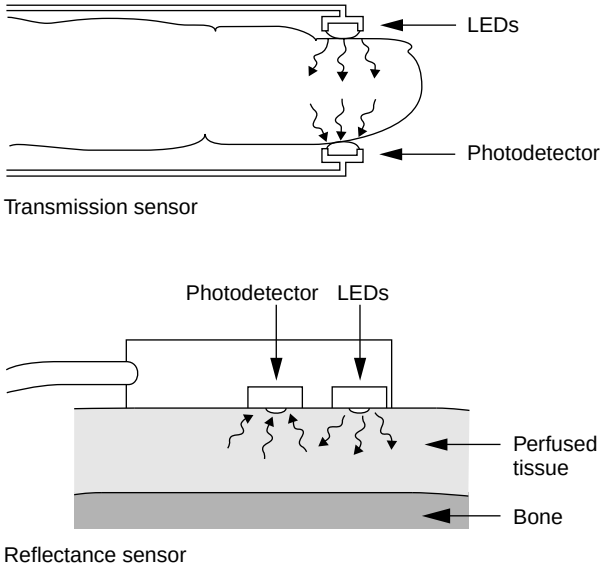


Figure 9. Major components of transmission and reflection oximeters.

intensity of the light. In spectrophotometry, light passes via a filter that converts the light into a specific wavelength. This light then passes through a container that contains the substance being analyzed. This substance absorbs part of this light and emits the remaining part, which goes through a special cell. This cell is connected to a photodetector, which detects and measures the emitting light (spectrophotometry). This method can be used for quantitive as well as qualitative analyses.

Lambert–Beer Law. This law combines the different factors that affect the light absorption of a substance:

$$\log 10 I_o/I_x = kcd$$

I_o = intensity of light incident on the specimen

I_x = intensity of the transmitted light

I_o/I_x = optical density

As shown in the above formula, the concentration of absorbing substance, the path length of the absorbing medium (d) and the characteristics of the substance and the light wavelength (k = constant) all affect light absorption (12).

Transmission Versus Reflection Oximetry. When the light at a particular wavelength passes through a blood sample, which contains Hb, this light would be absorbed, transmitted, or reflected. The amount of the absorbed, transmitted, or reflected light at those particular wavelengths is determined by various factors, including the concentration (Lambert–Beer law) and the type of the Hb present in the blood sample. The amount of light transmitted through the blood sample at a given wavelength is related inversely to the amount of light absorbed or reflected. The transmission oximetry is a method to determine the arterial oxygen saturation (S_aO_2) value by measuring the amount of light transmitted at certain wavelengths. On the other side, in the reflection oximetry, measuring the amount of light reflected is used to determine the S_aO_2 value. The significant difference between these two methods is the location of the photodetecor (Fig. 9). In the reflection method, the photodetector is on the same side of the light source. However, in the transmission oximetry, it is on the opposite site side of the light source (7,12).

Oximetry Versus Cooximetry. Each form of hemoglobin (e.g., oxyhemoglobin, deoxygenated hemoglobin, carboxyhemoglobin, methemoglobin) has its own unique absorption–transmission–reflection spectrum. By plotting the relative absorbance to different light wavelengths for both oxyhemoglobin and deoxygenated Hb as shown in Fig. 10. It is clear that these two hemoglobins absorb light differently at different light wavelengths. This difference is big in some light wavelengths (e.g., 650 nm in the red region), and small or not existing in other light wavelengths. The isosbestic point (13) is the light wavelength at which there is no difference between these two hemoglobins in absorbing light ($\sim$805 nm near the IR region). The difference in these two wavelengths can be used to calculate the S_aO_2.

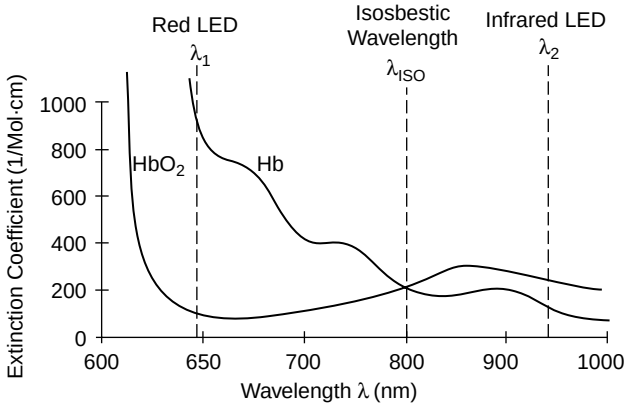


Figure 10. Light absorption spectra of oxygenated and deoxygenated hemoglobin.

However, these two hemoglobins are not only the hemoglobins exist in the patient's blood. There are other abnormal hemoglobins (dyshemoglobins) that can join these two hemoglobins in some abnormal conditions (such as carboxyhemoglobin and methemoglobin). Each one of these dyshemoglobins has its unique transmission–reflection–absorption spectrum. Some of these spectra are very close to the oxyhemoglobin spectrum at the routinely used two light wavelengths (see above). This makes these two wavelengths are incapable in detecting those dyshemoglobins. Therefore, the use of regular oximeters in these conditions may lead to erroneous and false readings, which may lead to detrimental effects on the patient's care. To overcome this significant problem a special oximeter (cooximeter, i.e., cuvette oximeter) is needed when there is a suspicion of presence of high level of dyshemoglobins in the patient's blood. Functional S_aO_2 is the percentage of oxyhemoglobin compared to sum of oxy- and deoxyhemoglobins. Therefore, the abnormal hemoglobins are not directly considered in the measurement of functional S_aO_2 by using regular oximetry. Cooximetry uses four or more light wavelengths, and has the ability to measure carboxyhemoglobin and methemoglobin as well as normal hemoglobins. The fractional S_aO_2 measures the percentage of oxyhemoglobin to all hemoglobins (normal and abnormal) present in the blood sample (14,15).

EAR OXIMETRY

Historical Development

In 1935, Matthes (16,17) showed that transmission oximetry could be applied to the external ear. However, a major problem with noninvasive oximetry applied to the ear was the inability to differentiate light absorption due to arterial blood from that due to other ear tissue and blood. In the following years, two methods were tried to solve this problem. The first was increasing local perfusion by heating the ear, applying vasodilator, or rubbing the ear. The second was comparing the optical properties of a "bloodless" earlobe (by compressing it using a special device) to the optical properties of the perfused ear lobe. Arterial S_aO_2 was then determined from the difference in these different measurements. This step was a significant step toward an accurate noninvasive measurement of S_aO_2. In 1976, Hewlett-Packard (18) used the collected knowledge about ear oximetry to that date to develop the model 47201A ear oximeter, Fig. 11.

HEWLETT-PACKARD EAR OXIMETER

This oximeter (18) is based on the measured light transmission at eight different wavelengths, which made this sensor less accurate and more complex than pulse oximeters. It used a high intensity tungsten lamp that generated a broad spectrum of light wave lengths. This light passes through light filters, then enters a fiberoptic cable, which carrys the filtered light to the ear. A second fibroptic cable carries the light pulses transmitted through the ear to the device for detection and analysis. The ear probe is relatively bulky ($\sim 10 \times 10$ cm) equipped with a temperature-controlled heater (to keep temperature of 41°C). It is attached to the antihelix after the ear has been rubbed briskly. This monitor is no longer manufactured because of its bulkiness and cost, and because of the development widely of a more accurate, smaller, and cost-effective monitor, the pulse oximeter.

Figure 11. The Hewlett-Packard Model 47201A ear oximeter.

PULSE OXIMETRY

Historical Development

In the early 1970s, Takuo Aoyagi (16,19,20), a Japanese physiological bioengineer, introduced pulse oximetry, the underlying concept of which had occurred to him while trying to cancel out the pulsatile signal of an earpiece densitometer with IR light. In early 1973, Dr. Susumu Nakajima, a Japanese surgeon, learned of the idea and ordered oximeter instruments from Nihon Kohden. After several prototypes were tested, Aoyagi and others delivered the first commercial pulse oximeter in 1974. This instrument was the OLV-5100 ear pulse oximeter, (Fig. 12). In 1977, the Minolta Camera Company

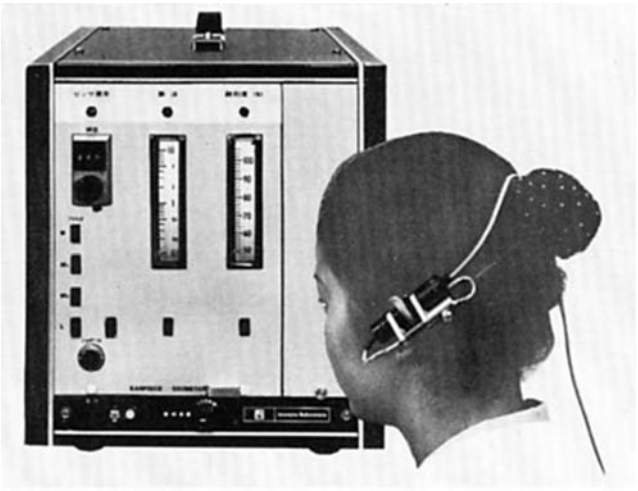


Figure 12. The OLV-5100 ear pulse oximeter, the first commercial pulse oximeter, it was introduced by Nihon Kohden in 1974.

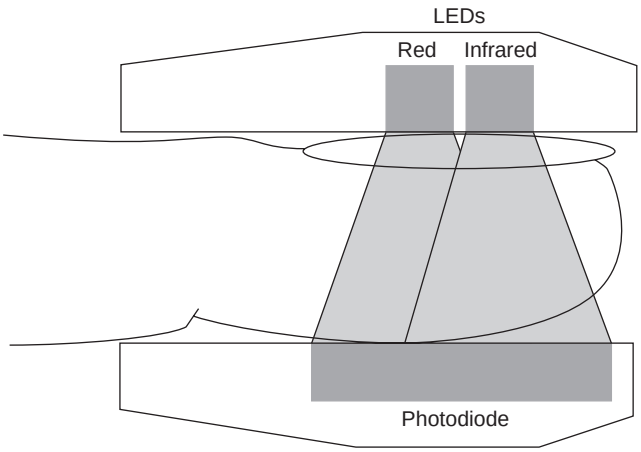


Figure 13. The basic components of a pulse oximeter sensor. Two LEDs with different wavelengths as light sources and a photodiode as receiver.

introduced the Oximet MET-1471 pulse oximeter with a fingertip probe and fiberoptic cables. Nakajima and others tested the Oximet MET-1471 and reported on it in 1979. In the years since, pulse oximetry has become widely used in a number of fields, including Anesthesia, intensive care, and neonatal care.

Principles of Operation

Pulse oximetry differs from the previously described oximetry in that it does not rely on absolute measurements, but rather on the pulsations of arterial blood. Oxygen saturation is determined by monitoring pulsations at two wavelengths and then comparing the absorption spectra of oxyhemoglobin and deoxygenated hemoglobin (20,21). Pulse oximetry uses a light emitter with red and infrared LEDs (light-emitting diodes) that shine through a reasonably translucent site with good blood flow (Fig. 13). Typical adult–pediatric sites are the finger, toe, pinna (top), or lobe of the ear. Infant sites are the foot or palm of the hand and the big toe or thumb. On the opposite side of the emitter is a photodetector that receives the light that passes through the measuring site. There are two methods of sending light through the measuring site (see above) (Fig. 9). The transmission method is the most common type used, and for this discussion the transmission method will be implied. After the transmitted red (R) and IR signals pass through the measuring site and are received at the photodetector, the R/IR ratio is calculated. The R/IR is compared to a "look-up" table (made up of empirical formulas) that converts the ratio to pulse oxygen saturation (S_pO_2) value. Most manufacturers have their own tables based on calibration curves derived from healthy subjects at various S_pO_2 levels. Typically, an R/IR ratio of 0.5 equates to approximately 100% S_pO_2, a ratio of 1.0 to $\sim$82% S_pO_2, while a ratio of 2.0 equates to 0% S_pO_2. The major change that occurred from the eight-wavelength Hewlett-Packard oximeters (see above) of the 1970s to the oximeters of today was the inclusion of arterial pulsation to differentiate the light absorption in the measuring site due

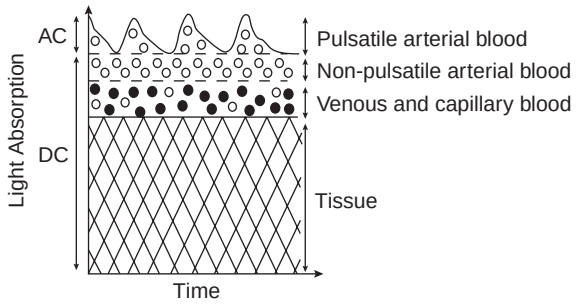


Figure 14. Schematic Representation of light absorption in adequately perfused tissue.

to skin, tissue, and venous blood from that of arterial blood. At the measuring site there are several light absorbers (some of them are constant) such as skin, tissue, venous blood, and the arterial blood (Fig. 14). However, with each heart beat the heart contracts and there is a surge of arterial blood, which momentarily increases arterial blood volume across the measuring site. This results in more light absorption during the surge. Light signals received at the photodetector are looked at as a waveform (peaks with each heartbeat and troughs between heartbeats). If the light absorption at the trough, which should include all the constant absorbers, is subtracted from the light absorption at the peak, then the resultants are the absorption characteristics due to added volume of blood only, which is arterial blood. Since peaks occur with each heartbeat or pulse, the term "pulse oximetry" was applied.

New Technologies

Conventional pulse oximetry accuracy degrades during motion and low perfusion. This makes it difficult to depend on these measurements when making medical decisions. Arterial blood gas tests have been and continue to be commonly used to supplement or validate pulse oximeter readings. Pulse oximetry has gone through many advances and developments since the Hewlett-Packard Model 47201A ear oximeter invention in 1976. There are several types of pulse oximeters manufactured by different companies available in the market nowadays. Different technologies have been used to improve pulse oximetry quality and decrease its limitations, which would lead eventually to better patient care. Figure 15 shows a modern pulse oximeter (Masimo Rad-9) designed by Masimo using the Signal Extraction Technology (Masimo SET) (22,23), is a software system composed of five parallel algorithms designed to eliminate nonarterial "noise" in a patient's blood flow. This monitor display includes: S_pO_2, pulse rate, alarm, trend, perfusion index (PI) (24), signal IQ, and plethysmographic waveform. Moreover, Masimo manufactures a handheld pulse oximeter by utilizing the same technology (Masimo SET) as shown in Fig. 16. Its small size ($\sim 15.7 \times 7.6 \times 3.5$ cm) and broad catalog of features make it suited for hospital, transport, and home use. Nellcor (25) uses the OxiMax technology to produce a list of pulse oximetry monitors and sensors. These sensors have a small digital memory chip that transmits

Figure 15. Masimo Rad-9 pulse oximeter.

Figure 17. Nellcor N-595 pulse oximeter.

sensor-specific data to the monitor. These chips contain all the calibration and operating characteristics for that sensor design. This gives the monitor the flexibility to operate accurately with a diverse range of sensor designs without the need for calibrating each sensor to the specific monitors. This opens a new area of pulse oximetry innovations. Figure 17 reveals some of the Nellcor monitors and sensors available. Furthermore, Nellcor has designed a handheld pulse oximeter, as shown in Fig. 18, compatible with its line of OxiMax pulse oximetry sensors. Nellcor combines two advanced technologies in measuring blood gases: the OxiMax technology and Microstream CO_2 technology (see the section Capnography) to produce a S_pO_2 and end-tidal CO_2 partial pressure (PetCO_2) handheld capnograph–pulse oximeter to monitor both S_pO_2 and PetCO_2. Several additional parameters are now available on the modern oximeter, and they add additional functionality for these monitors and decrease their limitations. One of these parameters is called the "perfusion index" (PI) (24,26). This is a simple measure of the change that has occurred in the tissue-under-test (e.g., the finger) over the cardiac cycle. When this parameter was first recognized as being something that a pulse oximeter could measure, it was difficult to imagine a value to the measurement because it is affected by so many different physiological and environmental variables, including systemic vascular resistance, volume status, blood pressure, and ambient temperature. But as time continues to pass since its introduction, more applications for PI have been found. The most obvious use for perfusion index is as an aid in sensor placement. It provides a means to quantify the validity of a given sensor site and, where desired, to maximize measurement accuracy. Perfusion index has also provided a simple and easy way test for sufficient collateral blood flow in the ulnar artery to allow for harvest of the radial artery for coronary artery bypass graft (CABG) surgery and for monitoring peripheral perfusion in critically ill patients.

Figure 16. Masimo Rad-5 handheld pulse oximeter.

Clinical Uses

Pulse oximeters are widely used in clinical practice (27–30). They are used extensively in the intensive care units to monitor oxygen saturation, and to detect and prevent hypoxemia. Monitoring oxygen saturation during anesthesia is a standard of care, which is almost always done by pulse oximeters. Pulse oximeters are very helpful in monitoring patients during procedures like bronchoscopy, endoscopy, cardiac catheterization, exercise testing,

Figure 18. Nellcor N-45 handheld pulse oximeter.

Figure 19. Portable Nonin Onyx 9500 pulse oximeter.

and sleep studies. Also, they are commonly used during labor and delivery for both the mother and infant. These sensors have no significant complications related to their use. There are several types of portable pulse oximeters on the market. These oximeters are small in size, useful for patients transport, and can be used at home. Figure 19 shows one of these pulse oximeters.

Accuracy and Limitations

The accuracy of pulse oximeters in measuring exact saturation has been shown to be $\sim \pm 4\%$ as compared to blood oximetry measurements. Several studies have shown that with low numbers of S_aO_2, there is a decreased correlation between S_pO_2 and S_aO_2, especially when $S_aO_2 < 70\%$ and in unsteady conditions (31). However, newer technologies have improved accuracy during these conditions substantially. Another factor that influences the accuracy of pulse-oximetry is the response time. There is a delay between a change in S_pO_2 and the display of this change. This delay ranges from 10 to 35 s. Pulse-oximeters have several limitations that may lead to inaccurate readings. One of its most significant limitations is that it estimates the S_aO_2, not the arterial oxygen tension (P_aO_2). Another limitation is the difficulty these sensors have in detecting arterial pulsation in low perfusion states (low cardiac output, hypothermia etc.) (32). Furthermore, the presence of dyshemoglobins (e.g., methemoglobin, carboxyhemoglobin) (15) and diagnostic dyes (e.g., methylene blue, indocyanine green, and indigo carmine) (33) affects the accuracy of these monitors, leading to false readings. High carboxyhemoglobin levels will falsely elevate S_pO_2 readings, which may lead to a false sense of security regarding the patient's oxygenation, and possible disastrous outcome. CO-oximetry should be used to measure S_aO_2 in every patient who is suspect for elevated carboxyhemoglobin (such as fire victims). Methemoglobinemia may lead to false $\sim 85\%$ saturation reading. The clinician should be alert to the potential causes and possibility of methemoglobinemia (e.g., nitrites, dapsone, and benzocaine). The CO-oximetry is also indicated in these patients. Vascular dyes may also affect the S_pO_2 readings significantly, especially methylene blue, which is also used in the treatment of methemoglobinemia. Brown, blue, and green nail polish may affect S_pO_2 too. Therefore, routine removal of this polish is recommended. The issue of skin pigmentations effect on S_pO_2 reading is still controversial. Motion artifacts are a common problem in using pulse oximeters, especially in the intensive care units.

Future Directions for Pulse Oximeters

As mentioned above, there are several limitations with the recent commercially available pulse oximeters. Pulse oximeters technology is working on decreasing those limitations and improving pulse oximeters function (34). In the future, techniques to filter out the noise component common to both R and IR signals, such as Masimo signal extraction, will significantly decrease false alarm frequency. Pulse oximeters employing more than two wavelengths of light and more sophisticated algorithms will be able to detect dyshemoglobins. Improvements in reflection oximetry, which detects backscatter of light from light-emitting diodes placed adjacent to detectors, will allow the probes to be placed on any body site. Scanning of the retinal blood using reflection oximetry can be used as an index of cerebral oxygenation. Combinations of reflectance oximetry and laser Doppler flowmetry may be used to measure microcirculatory oxygenation and flow.

Continuous Intravascular Blood Gas Monitoring (CIBM)

The current standard for blood gas analysis is intermittent blood gas sampling, with measurements performed *in vitro* in the laboratory or by using bedside blood gas analyzer. Recently, miniaturized fiberoptic devices have been developed that can be placed intravascularly to continuously measure changes in PO_2, PCO_2, and pH. These devices utilize two different technologies: Electrochemical sensors technology, based on a modified Clark electrode, and optode (photochemical/optical) technology (35,36).

Optode (Photochemical–Optical) Technology. An optode unit consists of optical fibers with fluorescent dyes encased in a semipermeable membrane. Each analyte, such as hydrogen ion, oxygen, or carbon dioxide, crosses the membrane and equilibrates with a specific chemical fluorescent dye to form a complex. As the degree of fluorescence changes with the concentration of the analyte, the absorbance of a light signal sent through the fiberoptic bundles changes, and a different intensity light signal is returned to the microprocessor. Optode technology has accuracy comparable to that of a standard laboratory blood gas analyzer. However, several reasons and problems, including the cost (see below) still limit the use of this monitor routinely.

At present, the Paratrend 7+ (PT7+; Diametric Medical Inc., High Wycombe, U.K.; distributed by Philips Medical Systems), and Neotrend (NT) are the only commercially available multiparameter CIBM systems. The original probe of Paratrend 7 (PT7) was introduced in 1992. It consists of a hybrid probe incorporating four different sensors: miniaturized Clark electrode to measure PO_2, optode to determine PCO_2, and pH (absorbance sensors, phenol red in bicarbonate solution), and a thermocouple (copper, constantan) to measure temperature and allow temperature correction of the blood gas values. All these sensors were encased in a heparin-coated microporous polyethylene tube that was permeable to the analytes to be measured. This sensor was modified in 1999. In the new sensor (PT7+) (Fig. 20), the Clark electrode was replaced by an optical PO_2 sensor. According to the manufacturer, this new PO_2 sensor is more accurate and has a faster response time.

Clinical Uses

Continuous intravascular blood gas monitoring has been applied in various clinical settings (36,37) in the operating room and the intensive care unit. In the operating room, especially in adults undergoing one lung ventilation for major surgery (e.g., one lung ventilation for thoracoscopic surgery or lung transplantation, major cardiac or vascular surgery). The most common site for CIBM measurement is the radial artery in adults and the femoral artery in children. The umbilical artery is used for probe insertion in neonates. Reports and studies showed that performance and accuracy of CIBM devices appear to be sufficient for clinical use.

Figure 20. The Paratrend 7+ (PT7+; Diametric Medical Inc.) sensor.

Limitations and Complications

Reliable intravascular blood gas measurement depends on a number of mechanical, electrical, and physicochemical properties of the CIBM probe as well as the conditions of the vessel into which the probe is inserted (36,37). Therefore, several factors can affect the performance of CIBM, including mechanical factors related to the intraarterial probe (e.g., not advanced adequately in the artery, the sensor becomes attached to the wall of the vessel), factors related to the artery itself (e.g., vasospasm), interference from electrocautery and ambient or endoscopic light, or related to the "flush" solution used to flush the intraarterial catheter, which may lead to false measurements. Complications may include thrombosis, ischemia, vasospasm, and failure. Although CIBM appears to be advantageous, there are no prospective, randomized, double-blind studies of its impact on morbidity and mortality. Future outcome studies should focus on well-defined groups of selected patients who might benefit from CIBM (e.g., critically ill patients with potentially rapid and unexpected changes in blood gas values). Furthermore, no data is available on the cost/benefit ratio of CIBM, and more studies are still needed to know if this monitor is cost-effective.

Intrapartum Fetal Pulse Oximetry

Intrapartum fetal pulse oximetry is a direct continuous noninvasive method of monitoring fetal oxygenation (38). Persistent fetal hypoxemia may lead to acidosis and neurological injury, and current methods to confirm fetal compromise are indirect and nonspecific. Therefore, intrapartum fetal pulse oximetry may improve intrapartum fetal assessment and, most important, improve the specificity of detecting fetal compromise (39,40). Intrapartum fetal pulse oximetry may monitor, not only the fetal heart rate (FHR), but also the arterial oxygen saturation and peripheral perfusion may be assessed.

Principle of Operation and Placement

The fetus *in utero* does not have an exposed area that would allow placement of a transmission sensor (38). Thus, reflectance sensors have been designed where the light-emitting diodes are located adjacent to the photodetector (Fig. 9). During labor, the sensor is placed transvaginally between the uterine wall and the fetus, with contact on the fetal presenting part, usually the soft tissue of the fetal cheek. Monitoring of fetal oxygen saturation has been encumbered by multiple technical obstacles (38). For example, reflectance sensors not directly attached to the fetus, work only when in contact with fetal skin and may not produce an adequate S_pO_2 signal when contact is suboptimal during intense uterine contractions or during episodes of fetal movement. In this situation, sensor position may require adjustment. Improved reflectance sensor contact has been attempted via a variety of sensor modifications, including suction devices, application with glue, and direct attachment to the fetal skin with a special clip. The Nellcor (Fig. 21) sensors have been developed with a "fulcrum" modification, which mechanically places the sensor surface into better contact with the fetal skin. Other technical

Figure 21. Nellcor OxiFirst fetal pulse oximeter.

advances, such as modification of the red light-emitting diode from a 660 to a 735 nm wavelength, have resulted in improved registration times.

Future Direction of Intrapartum Fetal Pulse Oximetry. Ideally, calibration of these monitors in human fetuses should be done by simultaneous measurement of S_pO_2 and preductal S_aO_2. Because the access to fetal circulation during labor is not feasible, calibration of these monitors is still a major problem (38). It appears that well-designed animal laboratory studies and human infant and neonatal studies will have to suffice for calibration and validation of these monitors. To make this monitor more valuable and accurate as a guide for obstetric and neonatal management during labor, prospective studies with a larger number of abnormal fetuses will be necessary to determine duration and level of hypoxia leading to metabolic acidosis in humans. Also, more studies are needed to answer questions about its safety and efficacy. Finally, further refinements in equipment design should improve the accuracy of S_pO_2 determination and the ability to obtain an adequate signal. Decreased signal-to-noise ratios, motion artifacts (e.g., contractions, fetal movement, maternal movement), impediments to light transmission (e.g., vernix, fetal hair, meconium), and calibration difficulties are unique obstacles in accurately assessing the fetus by this monitor. Technical development goals of fetal pulse oximetry should include improvement of sensor optical design, hardware, and software modification to obtain high signal quality and precise calibration. Major advantages of fetal oxygen saturation monitoring include its ease of interpretation for clinicians of varying skills, being noninvasive method, and the ability to monitor fetal oxygenation continuously during labor. However, more studies are needed to evaluate its safety, efficacy, and cost issues (41). When these issues are resolved, intrapartum fetal oxygen saturation monitoring could perhaps be one of the major advances in obstetrics during the twenty-first century.

TRANSCUTANEOUS BLOOD GAS MONITORING (TCM)

Historical Development

The possibility of continuously monitoring arterial blood oxygen and carbon dioxide using a heated surface electrode on human skin was discovered in the early 1970s and made commercially available by 1976 (42). In 1951, Baumberger and Goodfriend published an article showing a method to determine the arterial oxygen tension in man by equilibration through intact skin. By immersing a finger in a phosphate buffer solution heated to 45 °C, they found that the PO_2 of the buffer approached that of the alveolar air. They showed that if skin blood flow increased by the highest tolerable heat (45°C), the surface PO_2 rises to arterial blood PO_2. A few years later (in 1956), Clark invented the membrane covered platinum polarographic electrode to measure O_2 tissue tensions. By 1977, at least three commercial transcutaneous PO_2 ($tcPO_2$) electrodes were available (Hellige, Roche, RADIOMETER). These devices were applied initially to premature infants in an effort to reduce the incidence of blindness due to excessive oxygen administration. Throughout more than three decades, the TCM technology has been closely linked to the care of neonates; however, recent studies suggest that TCM technology may work just as well for older children and adults (29). The TCM offers continuous noninvasive measurement of blood gases, which is especially advantageous in critically ill patients in whom rapid and frequently life-threatening cardiopulmonary changes can occur during short periods of time. However, with the widespread use of pulse oximetry, the use of transcutanenous blood gas monitors has decreased.

Blood Gas Diffusion Through the Skin

The human skin consists of three main layers: the stratum corneum, epidermis, and dermis (Fig. 22). The thickness of the human skin varies with age, sex, and region of the body. The thickness of the stratum corneum varies from 0.1 to 0.2 mm depending on the part of the body. This is nonliving layer composed mainly of dehydrated cells (dead layer), which do not consume oxygen or produce carbon dioxide. The next layer is the epidermis layer, which consists of proteins, lipids, and melanin-forming cells. The epidermis is living, but is blood-free. The thickness of this layer $\sim$0.05–1 mm. Underneath the epidermis is the dermis, which consists of dense connective tissue, hair follicles, sweat glands, fat cells, and capillaries. These capillaries receive blood from arterioles and drain in venules. Arteriovenous anastomoses innervated by nerve fibers are commonly found in the dermis of the palms, face, and ears. These shunting blood vessels regulate blood flow

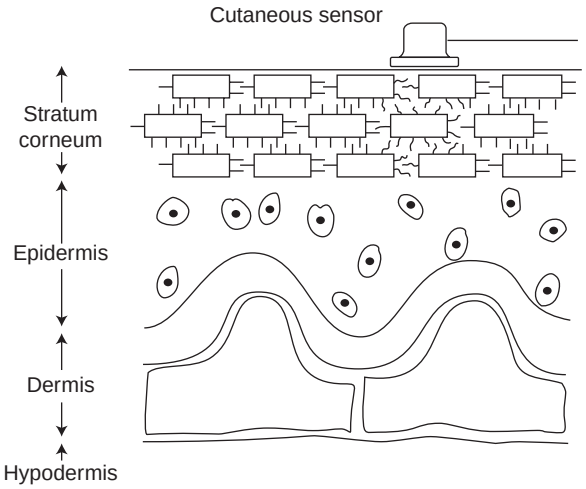


Figure 22. Human skin.

through the skin. Heat increases blood flow through these channels almost 30-fold. Gas diffusion through the skin occurs due to a partial pressure difference between the blood and the outermost surface of the skin. Diffusion of blood gases through the skin normally is very low, however, the heated skin ($\sim 43\,^{\circ}C$) becomes considerably more permeable to these gases.

Principle of Transcutaneous PO_2 Measurement (tcPO_2)

The probe used to measure the tcPO_2 is based on the idea of oxygen polarography (see above). This probe (7,12) consists of a platinum cathode and a silver reference anode encased in an electrolyte solution and separated from the skin by a membrane permeable to oxygen (usually made of Teflon, polypropylene, or polyethylene). The electrode is heated, thereby melting the crystalline structure of the stratum corneum, which otherwise makes this skin layer an effective barrier to oxygen diffusion. The heating of the skin also increases the blood flow in the capillaries underneath the electrodes. Oxygen diffuses from the capillary bed through the epidermis and the membrane into the probe, where it is reduced at the cathode, thereby generating an electric current that is converted into partial pressure measurements and displayed by the monitor. Because of an *in vitro* drift inevitably occurring inside the probe, where several chemical reactions are going on, the tcPO_2 sensor must be calibrated before using, and be repeated every 4–8 h. Since the O_2-dependent current flow exhibits a linear relationship at a fixed voltage, only two known gas mixtures are required for the calibration. Two *in vitro* calibration techniques can be employed: by using two precision gas mixtures (e.g., nitrogen and oxygen), and by using a "zero O_2 solution" (e.g., sodium sulfite) and room air.

The Transcutaneous PO_2 Sensor. The modern transcutaneous PO_2 sensors still use the principles used by Clark decades ago (7,12). Figure 23 illustrates a cross-sectional diagram of a typical Clark-type sensor. This particular sensor consists of three glass-sealed platinum cathodes that are separately connected via current amplifiers to a surrounding Ag–AgCl cylindrical ring. A buffered KCl electrode, which has a low water content to reduce drying of the sensor, is used. The following basic reactions happen between the two electrodes:

At the anode (+ electrode):

$$4\,Ag + 4\,Cl^- \leftrightarrow 4\,AgCl + 4\,e^-$$

(the electrons complete the circuit)

At the cathode (− electrode):

$$4\,H^+ + 4\,e^- + O_2 \leftrightarrow 2\,H_2O$$

(the electrons are boiled off of the platinum electrode)

Overall:

$$4\,Ag + 4\,Cl^- + 4\,H^+ + O_2 \leftrightarrow 4\,AgCl + 2\,H_2O$$

The two electrodes are covered with a thin layer of electrolytic solution that is maintained in place by a membrane that allows slow diffusion of O_2 from the skin into the sensor. The diffusion of O_2 through the skin is normally very low. Under normal physiological conditions, the PO_2 measured at the surface of the skin using a nonheated transcutaneous PO_2 electrode is near zero, regardless of the underlying blood PO_2. In order to facilitate O_2 diffusion through the skin, abrasion of the skin and drug-induced hyperemia through the application of nicotinic acid cream were initially used. However, since direct skin heating gives a more prolonged and consistent effect, a heating element is now used in all commercial transcutaneous PO_2 sensors. Generally, temperatures between 43 and 44° yield adequate vasodilatation of the cutaneous blood vessels with minimal skin damage. Heating the skin speeds up O_2 diffusion through the stratum corneum. In addition, it also causes vasodilatation of the dermal capillaries, which increases blood flow to the region of skin in contact with the sensor. With increased blood flow, more O_2 is available to the tissues surrounding the capillaries in the skin, and consequently the PO_2 of the blood in these capillary loops approximate more closely that of the arterial blood. Heating the blood also shifts the oxygen dissociation curve to the right. Therefore, the binding of hemoglobin with O_2 is reduced and the release of O_2 to the cells is increased. Simultaneously, skin heating also increases local tissue O_2 consumption. Fortunately, however, these two factors tend to cancel each other.

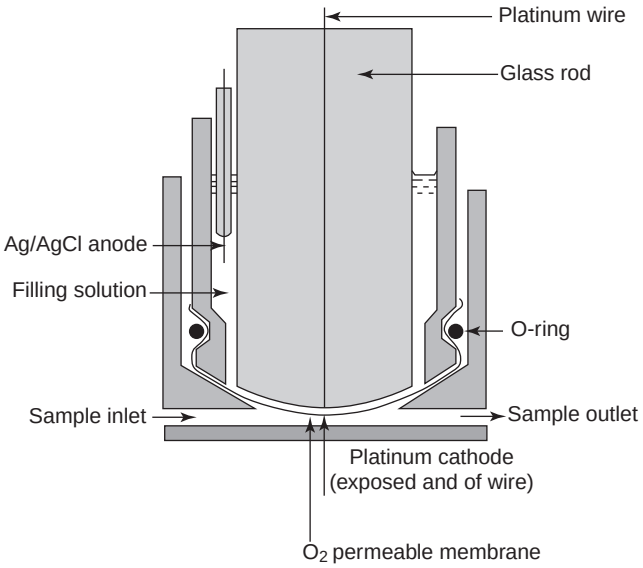


Figure 23. A cross-sectional diagram of a typical Clark-type sensor.

Transcutaneous PCO_2 Monitoring

Continuous PCO_2 monitoring is helpful in monitoring lung ventilation during spontaneous breathing or artificial ventilation. It makes it easier to adjust the parameters of the ventilator and prevent respiratory acidosis or alkalosis.

The Transcutaneous PCO_2 Sensor

The typical sensor is similar the O_2 sensor that was described above, as shown in Fig. 24. This sensor (7,12) consists of glass pH electrode with a concentric Ag–AgCl reference electrode that also serves as a temperature-

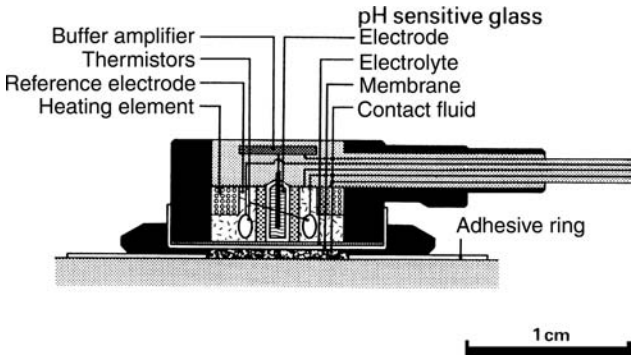


Figure 24. The transcutaneous PCO_2 sensor.

controlled heater. A buffer electrolyte (e.g., HCO_3^-) is placed on the surface of the electrode and a thin CO_2 permeable membrane (e.g., Teflon) stretched over the electrode separates the sensor from its surroundings. As CO_2 molecules diffuse via the CO_2-permeable membrane into the HCO_3^- containing solution, the following chemical reaction occurs:

$$CO_2 + H_2O \leftrightarrow H_2CO_3 \leftrightarrow H^+ + HCO_3^-$$

A potential between the pH and the reference electrodes is generated as a result of this reaction. This potential is proportional to the CO_2 concentration. Measurement of pH with a pH electrode can lead to estimation of the skin PCO_2, which correlates with the $PaCO_2$. According to the Henderson–Haselbach relationship, pH is proportional to the negative logarithm of PCO_2.

Skin temperature must be considered when analyzing PCO_2 measurements, because the skin heating can affect the transcutaneous PCO_2 sensor reading. This effect is due to the high temperature coefficient of the PCO_2 sensor. Heating the sensor results in an increase in PCO_2, since CO_2 solubility decreases, increase in local tissue metabolism, and increase in the rate of CO_2 diffusion through the stratum corneum. Therefore, the transcutaneous PCO_2 values are usually higher than the corresponding arterial PCO_2. Calibration of the PCO_2 sensor is different from the PO_2 sensor calibration. In the CO_2 sensor case, the voltage signal generated in the PCO_2 sensor is proportional to the logarithm of the CO_2 concentration (not to CO_2 concentration directly, as the case in PO_2). Therefore, there is no "zero point" calibration in transcutaneous PCO_2 sensor as there is with a transcutaneous PO_2 sensor. For this reason, one needs two different precisely analyzed gas mixtures for calibration. Usually, gas mixtures containing 5 and 10% CO_2 are used for calibrating the PCO_2 sensor. On the other side, PCO_2 sensor calibration must be done at the temperature at which it will be operated.

Clinical Applications of Transcutaneous PO_2 and PCO_2 Monitoring

Transcutaneous PO_2 and PCO_2 monitoring have found numerous applications in clinical medicine and research (42,43) during the past two decades: (1) neonatology: tcPO_2 monitoring remains the most commonly used technique to guide oxygen therapy in premature infants. In low birth weight infants, tcPO_2 is one of the best available monitor of ventilation. (2) Fetal monitoring: specially designed electrodes attached to the fetal scalp have been used. Changes in tcPO_2 rapidly reflected changing maternal and fetal conditions. Some studies showed that fetal tcPO_2 is considerably affected by local scalp blood flow, therefore repeated episodes of asphyxia, which may lead to increase in catecholamines, can reduce fetal scalp blood flow and lead to misleading reduction in tcPO_2. (3) Sleep studies: pulse oximetry and combined tcPO_2–tcPCO_2 electrode are used in sleep studies. This combination made it possible to study the ventilator response of hypoxia in sleeping infants. (4) Peripheral circulation: tcPO_2 electrodes are extensively used in evaluation peripheral vascular disease (44). Furthermore, transcutaneous oximetry has been used in several clinical situations such as prediction of healing potential for skin ulcers or amputation sites, assessment of microvascular disease (45), and determination of cutaneous vasomotor status. Figure 25 shows one of the commercially available transcutaneous blood gas monitor.

Figure 25. Radiometer TCM 4 transcutaneous blood gas monitor.

CAPNOMETRY AND CAPNOGRAPHY

Introduction

Capnometry is the measurement of carbon dioxide (CO_2) in the exhaled gas. Capnography is the method of displaying CO_2 measurements as waveforms (capnograms) during the respiratory cycle. The end-tidal PCO_2($PetCO_2$) is the maximum partial pressure of the exhaled CO_2 during tidal breathing (just before the beginning of inspiration). The measurement of CO_2 in respiratory gases was first accomplished in 1865, using the principle of Infrared(IR) absorption. Capnography was developed in 1943 and introduced to clinical practice in the 1950s (27). Since then, capnometry–capnography has gone through significant advances. Now capnography is a "standard of care" for general anesthesia (3), as described by the American Society of Anesthesiologists (ASA).

Measurement Techniques

Capnometry most commonly utilizes IR light absorption or mass spectrometry. Other technologies include Raman spectra analysis and a photoacoustic spectra technology (46,47)

Infrared Light Absorption Technique. This is the most common technique used to measure CO_2 in capnometers.

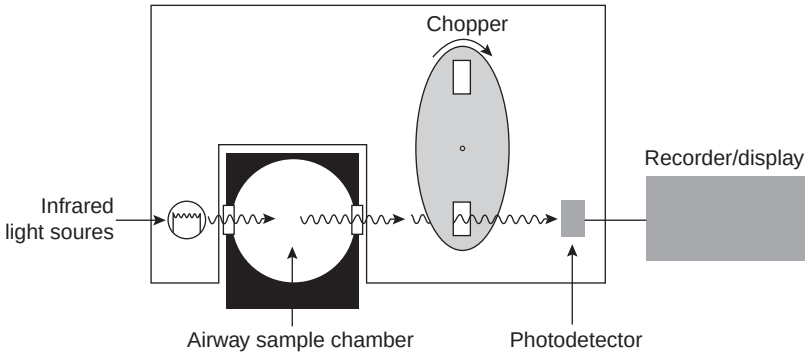


Figure 26. Single-beam infrared CO_2 analyzer, used in some mainstream sampling systems.

This method is cheaper and simpler than mass spectrometry. However, it is less accurate and has a slower response time ($\sim$0.25 vs. 0.1 s for mass spectrometry). There are two types of IR analyzers, a double and a single beam. The double-beam positive-filter model consists of an IR radiation source, which radiates to two mirrors. The two beams pass via a filter to two different chambers (sample chamber and reference chamber), and then to a photodetector. Consequently, it is possible to process the detector output electronically to indicate the concentration of CO_2 present. The single-beam negative-filter (Fig. 26), utilizes only one beam without using a reference. The principle behind this technique is that gases generally absorb electromagnetic IR radiation, and gas molecules with two or more atoms, provided these atoms are dissimilar (e.g., CO_2, but not O_2) absorb IR radiation in the range 1000–15000 nm. By filtering particular wavelengths, carbon dioxide and other gases can be measured. Carbon dioxide absorbs IR radiation strongly between 4200 and 4400 nm. Nitrous oxide and water have absorption peaks close to this area. Thus, there is a potential for the introduction of error with these substances in this method.

Raman Spectrography. Raman spectrography uses the principle of "Raman Scattering" for CO_2 measurement. The gas sample is aspirated into an analyzing chamber, where the sample is illuminated by a high intensity monochromatic argon laser beam. The light is absorbed by molecules, which are then excited to unstable vibrational or rotational energy states (Raman scattering). The Raman scattering signals (Raman light) are of low intensity and are measured at right angles to the laser beam. The spectrum of Raman scattering lines can be used to identify all types of molecules in the gas phase. Raman scattering technology has been incorporated into many newer anesthetic monitors (RASCAL monitors) to identify and quantify instantly CO_2 and inhalational agents used in anesthesia practice (48).

Mass Spectrography. The mass spectrograph separates molecules on the basis of mass to charge ratios. A gas sample is aspirated into a high vacuum chamber, where an electron beam ionizes and fragments the components of the sample. The ions are accelerated by an electric field into a final chamber, which has a magnetic field, perpendicular to the path of the ionized gas stream. In the magnetic field, the particles follow a path wherein the radius of curvature is proportional to the charge: mass ratio. A detector plate allows for determination of the components of the gas and for the concentration of each component. Mass spectrometers are quite expensive and too bulky to use at the bedside and are rarely used presently. They are either "stand alone", to monitor a single patient continuously, or "shared", to monitor gas samples sequentially from several patients in different locations (multiplexed). Up to 31 patients may be connected to a multiplexed system, and the gas is simultaneously sampled from all locations by a large vacuum pump. A rotary valve (multiplexer) is used to direct the gas samples sequentially to the mass spectrometer. In a typical 16-station system, with an average breathing rate of 10 breaths $\cdot$ min^{-1}, each patient will be monitored about every 3.2 min. The user can interrupt the normal sequence of the multiplexer and call the mass spectrometer to his patient for a brief period of time (46–48).

Photoacoustic Spectrography. Photoacoustic gas measurement is based on the same principles as conventional IR-based gas analyzers: the ability of CO_2, N_2O and anesthetic agents to absorb IR light (46,49). However, they differ in measurement techniques. While IR spectrography uses optical methods, photoacoustic spectrography (PAS) uses an acoustic technique. When an IR energy is applied to a gas, the gas will expand and lead to an increase in pressure. If the applied energy is delivered in pulses, the gas expansion would be also pulsatile, resulting in pressure fluctuations. If the pulsation frequency lies within the audible range, an acoustic signal is produced and is detected by a microphone. Potential advantages of PAS over IR spectrometry are higher accuracy, better reliability, less need of preventive maintenance, and less frequent need for calibration. Furthermore, as PAS directly measures the amount of IR light absorbed, no reference cell is needed and zero drift is nonexistent in PAS. The zero is reached when there is no gas present in the chamber. If no gas is present there can be no acoustic signal (49).

CO_2 Sampling Techniques

Sidestream versus Mainstream. Capnometers that are used in clinical practice use two different sampling techniques (50) (Fig. 27): sidestream or mainstream. A mainstream (flow-through) capnometer has an airway adaptor cuvette attached in-line and close to the endotracheal tube. The cuvette incorporates an IR light source and sensor that

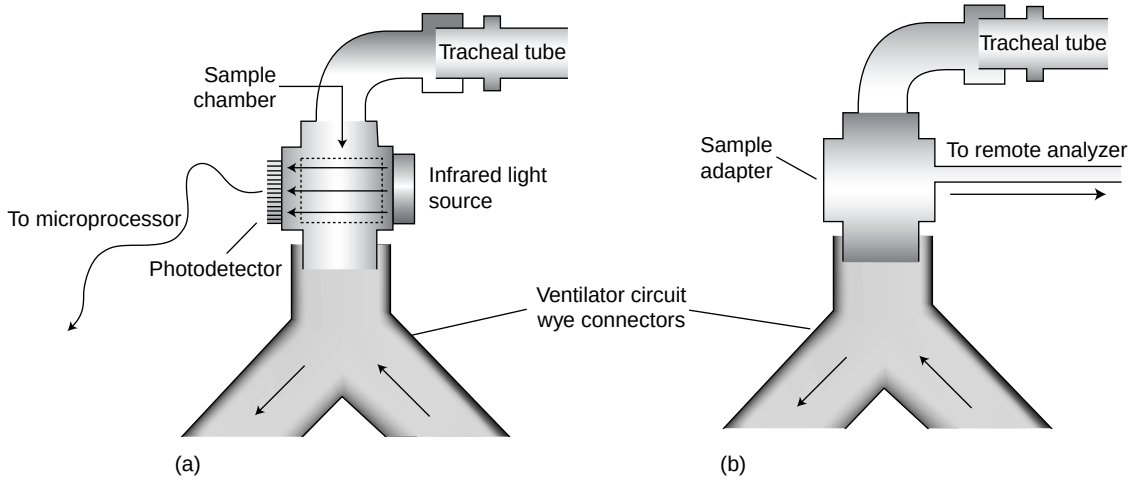


Figure 27. Sidestream vs. mainstream CO_2 sampling techniques. (a) Mainstream CO_2 sampling. (b) Sidestream sampling.

senses carbon dioxide absorption to measure PetCO_2. A sidestream capnometer uses a sampling line that attaches to a T-piece adapter at the airway opening, through which the instrument continually aspirates tidal airway gas for analysis of carbon dioxide. The main advantage of the mainstream analyzer is its rapid response, because the measurement chamber is part of the breathing circuit. The sample cuvette lumen, through which inspired and expired gases pass, is large in order to minimize the work of breathing, and pulmonary secretions generally do not interfere with carbon dioxide analysis. Compared with sidestream (aspiration) sampling, the airway cuvette is relatively bulky and can add dead space. However, within the past few years lighter and smaller airway cuvettes have been developed to allow its use in neonates. The sidestream PCO_2 analyzer adds only a light T-adapter to the breathing circuit, and can be easily adapted to non-intubation forms of airway control. Because the sampling tubing is small bore, it can be blocked by secretions. During sidestream capnography, the dynamic response, the steepness of the expiratory upstroke and aspiratory downslope, tends to be blunted because of the dispersive mixing of gases through the sampling line, where gas of high PCO_2 mixes with gas of low PCO_2. In addition, a washout time is required for the incoming sampled gas to flush out the volume of the measuring chamber. The overall effect is an averaging of the capnogram, resulting in a lowering of the alveolar plateau and an elevation of the inspiratory baseline. Thus, PetCO_2 may be underestimated and rebreathing can be simulated. These problems are exacerbated by high ventilatory rates and by the use of long sampling catheters. In addition, the capnogram is delayed in time by transport delay, the time required to aspirate gas from the airway opening adapter through the sampling tubing to the sampling chamber.

Micro-Stream Technology. Micro-stream technology (51) is a new CO_2 sampling techneque that uses a low aspiration rate (as low as 50 $mL \cdot min^{-1}$), such as NBP-75, Nellcor Puritan Bennett, as shown in Fig. 28. In addition, this technology uses a highly CO_2-specific IR source, where the IR emission exactly matches the absorption spectrum of the CO_2 molecules. The advantages of this technology, compared to the traditional high flow side-stream capnometer (150 $mL \cdot min^{-1}$), is that it gives more accurate PetCO_2 measurements and better waveforms in neonates and infants with small tidal volumes and high respiratory rates. Furthermore, these low flow capnometers are less likely to aspirate water and secretions into the sampling tubes, resulting in either erroneous PetCO_2 values or in total occlusion of sampling tube.

Phases of Capnography

A normal single breath capnogram (time capnogram) is shown in Fig. 29. Time capnogram is the partial pressure of expired CO_2 plotted against time on the horizontal axis. This capnogram can be divided into inspiratory (phase 0) and expiratory segments. The expiratory segment, similar to a single breath nitrogen curve or single breath CO_2 curve, is divided into phases I, II, and III, and occasionally, phase IV, which represents the terminal rise in CO_2 concentration. The angle between phase II and III is the alpha angle. The nearly 90° angle between phase III and the descending limb is the beta angle. Changes in time

Figure 28. Nellcor Microstream $ETCO_2$ breath sampling unit.

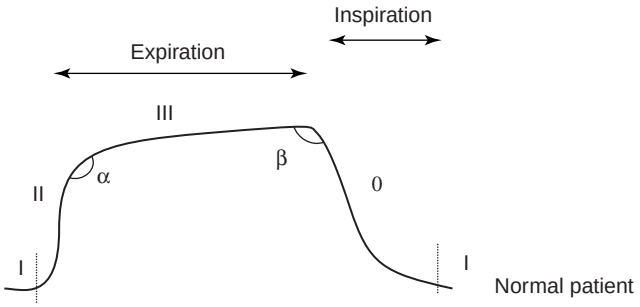


Figure 29. Normal single breath capnogram.

capnogram help to diagnose some of the breathing and ventilation problems, especially during anesthetic management of patients undergoing surgery (e.g., bronchospasm, esophageal intubation, CO_2 rebreathing, and even cardiac arrest).

Clinical Uses of Capnography

Capnography and capnometry (52) are safe, noninvasive test, and have few hazards. They are widely used in clinical medicine. Their uses include, but are not limited to (*1*) evaluating the exhaled CO_2, especially end-tidal CO_2 in mechanically ventilated patients during anesthesia. (*2*) Monitoring the severity of pulmonary disease and evaluating response to therapy, especially therapy intended to improve the ratio of dead space to tidal volume (V_D/V_T) and the matching of ventilation to perfusion (V/Q). (*3*) Determining that tracheal rather than esophageal intubation has taken place (low or absent cardiac output may negate its use for this indication) (53). Colorimetric CO_2 detectors are adequate devices for this purpose. (*4*) Evaluating the efficiency of mechanical ventilatory support by determination of the difference between the arterial partial pressure for CO_2 (PCO_2) and the $PetCO_2$. Figure 30 shows a combined handheld capnograph/pulse oximeter.

Limitations

Note that although the capnograph provides valuable information (52) about the efficiency of ventilation, it is not a replacement or substitute for assessing the PCO_2. The difference between PCO_2 and $PetCO_2$ increases as dead-space volume increases. In fact, the difference between the PCO_2 and $PetCO_2$ has been shown to vary within the same patient over time. Alterations in breathing pattern and tidal volume may introduce error into measurements designed to be made during stable, steady-state conditions. Interpretation of results must take into account the stability of physiologic parameters, such as minute ventilation, tidal volume, cardiac output, ventilation/perfusion ratios, and CO_2 body stores. Certain situations may affect the reliability of the capnogram. The extent to which the reliability is affected varies somewhat among types of devices (IR, photoacoustic, mass spectrometry, and Raman spectrometry). Furthermore, the composition of the respiratory gas mixture may affect the capnogram (depending on

Figure 30. Nellcor OxiMax NBP-75 handheld capnograph/ pulse oximeter.

the measurement technology incorporated). The IR spectrum of CO_2 has some similarities to the spectra for both oxygen and nitrous oxide. High concentrations of either or both oxygen or nitrous oxide may affect the capnogram, and, therefore, a correction factor should be incorporated into the calibration of any capnograph used in such a setting. The reporting algorithm of some devices (primarily mass spectrometers) assumes that the only gases present in the sample are those that the device is capable of measuring. When a gas that the mass spectrometer cannot detect (such as helium) is present, the reported values of CO_2 are incorrectly elevated in proportion to the concentration of helium in the gas mixture. Moreover, the breathing frequency may affect the capnograph. High breathing frequencies may exceed the response capabilities of the capnograph. In addition, the breathing frequency, > 10 breaths $\cdot$ min^{-1}, has been shown to affect devices differently. Contamination of the monitor or sampling system by secretions or condensate, a sample tube of excessive length, a sampling rate that is too high, or obstruction of the sampling chamber, can lead to unreliable results. Use of filters between the patient airway and the sampling line of the capnograph may lead to lowered $PetCO_2$ readings. Inaccurate measurement of expired CO_2 may be caused by leaks of gas from the patient–ventilator system preventing collection of expired gases, including, leaks in the ventilator circuit, leaks around tracheal tube cuffs, or uncuffed tracheal tubes.

Sublingual Capnometry

Sublingual capnometry is a method to measure the partial pressure of carbon dioxide under the tongue ($PSLCO_2$). This method is being used mainly in the critical care units to evaluate patients with poor tissue perfusion and multiple organ dysfunction syndrome.

Pathophysiologic Basis. Significant increases in the partial pressure of carbon dioxide (PCO_2) in tissue have

Figure 31. Nellcor CapnoProbe Sublingual capnometer.

been associated with hypoperfusion, tissue hypoxia, and multiple organ dysfunction syndrome (54). When perfusion of the intestinal mucosa is compromised, CO_2 accumulates in the gut. The high diffusability of CO_2 allows for rapid equilibration of PCO_2 throughout the entire gastrointestinal (GI) tract. The vasculature of the tongue and the GI tract are controlled by similar neuronal pathways. Thus, the vasculatures of both respond similarly during vasoconstriction (55). Because the tongue is the most proximal part of the GI tract, measurement of PCO_2 can be conveniently and noninvasively obtained by placing a sensor under the tongue. Clinical studies have demonstrated that $PSLCO_2$ can be used in the assessment of systemic tissue hypoperfusion and hypercapnia (56).

Capnometer Components and Principle of Operation. Figure 31 shows a commercially available sublingual capnometer (Nellcor CapnoProbe Sublingual System). This system (25) consists of two components: (1) SLS-l Sublingual Sensor: This sensor contains an optrode (a sensitive analyte detector) consists of an optical fiber capped with a small silicone membrane containing a pH-sensitive solution (Fig. 32). When the optrode is brought into contact with sublingual tissue, CO_2 present in the tissue freely diffuses across the silicone membrane into the fluorescent dye solution. No other commonly encountered gases or liquids can pass across the membrane. The CO_2 dissolves and forms carbonic acid, which in turn lowers the pH of the solution. The fluorescence intensity of the dye in the solution is directly proportional to pH.

This single use sensor is packaged in a sealed metal canister. Inside the canister, the sensor tip is enclosed in a gas permeable reservoir that contains a buffer solution. The solution prevents the optrode from drying out. The solution also allows calibration just prior to use, as it is in equilibrium with a known concentration of CO_2 within the canister. To begin use, the clinician opens the canister and inserts the cable handle into the SLS-l Sublingual Sensor. This action initiates a calibration cycle that allows the

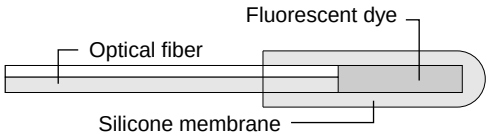


Figure 32. Diagram of Nellcor CapnoProbe sublingual capnometer basic components.

instrument to observe the sensor signal at the known PCO_2 of the calibrant (2). The N-80 instrument contains a precision optical component that emits light at two wavelengths in the violet and blue portions of the visible spectrum. The fluorescence intensity generated by the violet wavelength is insensitive to pH, whereas that generated by the blue wavelength is strongly sensitive to pH. The light is launched into an optical fiber and delivered to the tip of the disposable sensor. The green fluorescent light generated in the optrode is directed back to the N-80 instrument through an optical fiber. The light is then ratio metrically quantitated and directly correlated to $PSLCO_2$. When the fluorescence intensity of the optrode has stabilized (within 60–90 s), the N-80 instrument reports the measured value of $PSLCO_2$ on its LCD screen.

Chemical Colorimetric Airway Detector

This is a device (57,58) that uses a pH-sensitive indicator to detect breath-by-breath exhaled carbon dioxide (Fig. 33).

Figure 33. Nellcor Easy Cap II Pedi-Cap chemical colorimetric CO_2 detector.

The colorimetric airway detector is interposed between the endotracheal tube (ETT) and the ventilation device. Both adult and pediatric adaptors exist, but they cannot be used in infants who weigh < 1 kg. Because of excessive flow resistance, they are not suited for patients who are able to breathe spontaneously. Excessive humidity will render them inoperative in 15–20 min. The devices can be damaged by mucous, edematous, or gastric contents, and by administration of intratracheal epinephrine. Despite these drawbacks, colorimetric sensors have been found to be useful in guiding prehospital CPR (cardiopulmonary resuscitation) both in intubated patients and those with a laryngeal mask airway.

Role of Capnometry–Capnography in CPR. The relationship between cardiac output and $P\text{et}CO_2$ is logarithmic (59). Decreased presentation of CO_2 to the lungs is the major rate-limiting determinant of the $P\text{et}CO_2$ during low pulmonary blood flow. Capnography can detect the presence of pulmonary blood flow even in the absence of major pulses (pseudoelectromechanical dissociation, EMD) and also can rapidly indicate changes in pulmonary blood flow (cardiac output) caused be alterations in cardiac rhythm. Data suggests that $P\text{et}CO_2$ correlates with coronary perfusion pressure, cerebral perfusion pressure, and blood flow during CPR. This correlation between perfusion pressure and $P\text{et}CO_2$ is likely to be secondary to the relationship of $P\text{et}CO_2$ and cardiac output (60).

SUMMARY

Blood gas measurement methods and instruments have gone through significant improvements and advances in the last few decades. Invasive techniques have moved steadily toward using smaller instruments and closer to the patient's bed (bedside), which requires smaller blood sample. These improvements have made these devices more convenient, need less personnel to operate them, and more cost-effective. These bedside devices have comparable accuracy and reliability to the traditional central laboratory instruments. Continuous intravascular blood gas monitoring is a new invasive technique that uses miniaturized fiberoptic devices. This method has been used in different clinical settings with good results. However, several limitations and complications still exist. More studies and improvements are needed to know its cost-effectiveness in clinical medicine and to minimize its complications (such as ischemia and thrombosis). Other invasive instruments and methods were discussed in other parts of this encyclopedia. On the other hand, noninvasive blood gas measurement methods and devices improved greatly since its instruction to clinical medicine. These devices have been used extensively during anesthesia administration and in critical care units. Using pulse oximetry is "standard of care" in anesthesia practice. The use of pulse oximeter decreased the risk of hypoxia and its deleterious effect significantly. However, several limitations to its use still exist, especially in low perfusion states, during the presence of dyshemoglobins and motion artifacts. New technologies, such as the Oxi-Max and the Signal Extraction technologies, have been developed to overcome some of these limitations, and to add more features for these instruments (such as scanning the retina as an index of cerebral oxygenation and using Laser Doppler flowmetry–reflection oximetry to measure microcirculatory oxygenation and flow). Other types of pulse oximetry have been introduced to clinical medicine. Intrapartum fetal pulse oximetry is an example of these new pulse oximeters. This device provides a continuous, noninvasive method of monitoring fetal oxygenation, which may help in detecting persistent fetal hypoxemia and improve intrapartum fetal assessment. However, more studies are needed to evaluate its safety, efficacy, and cost issues. Transcutaneous blood gas monitoring is another noninvasive method to measure blood gases. This method is losing ground and popularity against the newer pulse oximeters, which have replaced this method in several situations. Capnometry–capnography has been used extensively in anesthesia practice in the last two decades. New CO_2 sampling technique such as microstream technology has been introduced. This method uses a low aspiration rate, making it more accurate than the previous techniques. Moreover, this technique can detect very small amount of CO_2. Photoacoustic sepectrography is a new technique has been developed to measure $P\text{et}CO_2$. This method is more reliable, accurate, and needs less calibration than the traditional methods. Studies showed an encouraging result and an important role for capnometry–capnography in cardiopulmonary resuscitation, which may lead to widespread use of these devices in CPR, in prehospital and inhospital settings.

BIBLIOGRAPHY

Cited References

1. Cadden K, Norman E, Booth J. Use of ABG in trauma for early recognition of acidosis and hypoxemia (Abstract). Respir Care 2001;46:1106.
2. Levin KP, Hanusa BH, Rotondi A, Singer DE, et al. Arterial blood gas and pulse oximetry in initial management of patients with community-acquired pneumonia. J Gen Intern Med 2001;9:590.
3. Morgan Jr GE, Mikhail MS, Murray MJ. Clinical Anesthesiology. 3rd ed. New York: Lange Medical Books/McGraw-Hill; 2002. p 124–125.
4. Severinghaus JW, Astrup P, Murray JF. Blood gas analysis and critical care medicine. Am J Resp Crit Care Med 1998;157:S114–S122.
5. Schell RM, Cole DJ. Cerebral monitoring: Jugular venous oximetry. Anesth Analg 2000;90:559.
6. Goodwin ALP. Physics of Gases, Anaesthesia and Intensive Care Medicine. (UK): The Medicine Publishing Company, Ltd.; 2003.
7. Malley WJ. Clinical Blood Gases: Assessment and Intervention. 2nd ed. New York: Elsevier; 2005.
8. West JB. Respiratory Physiology-Essentials. 6th ed. Baltimore: Williams & Wilkins; 2000.
9. Murray JF, Nadel JA. Textbook of Respiratory Medicine. 3rd ed. New York: W. B. Saunders.
10. Severinghaus JW, Astrup PB. History of blood gas analysis. VI: Oximetry. J Clin Monit 1986;2:270–288.
11. Millikan GA, Pappenheimer JR, Rawson AJ, et al. Continuous measurement of oxygen saturation in man. Am J Physiol 1941;133:390.

12. Adams AP, Hahn CEW. Principles and Practice of Blood Gas Analysis. London: Franklin Scientific Products; 1979.
13. Payne JB, Severinghaus JW. Pulse Oximetry. New York: Springer-Verlag; 1986.
14. Baker SJ, Tremper KK. The effect of carbon monoxide inhalation on pulse oximetry and transcutaneous PO_2. Anesthesiology 1987;66:677–679.
15. Baker SJ, Tremper KK, Hyatt J. Effects of methemoglobinemia on pulse oximetry and mixed venous oximetry. Anesthesiology 1989;70:112–117.
16. Severinghaus JW, Honda Y. History of blood gas analysis. VII. Pulse oximetry. J Clin Monit 1987;3:135–138.
17. Severinghaus JW. History and recent developments in pulse oximetry. Scand J Clin Lab Invest 1993;214 (1 Suppl): 105–111.
18. Merrick EB, Hayes TJ. Continuous noninvasive measurements of arterial blood oxygen levels. Hewlett-Packard J 1976;28920:2–9.
19. Aoyagi T, Miyasaka K. Pulse oximetry: its invention, contribution to medicine, and future tasks. Anesth Analg 2002;94(1 Suppl): S1–3.
20. Tremper KK, Barker SJ. Pulse oximetry. Anesthesiology 1989;70:98–108.
21. Welch JP, DeCesare MS, Hess D. Pulse oximetry: Instrumentation and clinical applications. Respir Care 1990;35: 584–601.
22. Robertson F, Hoffman G. Clinical evaluation of Masimo SET and Nellcor N395 oximeters during signal conditions in difficult-to-monitor neonates. Anesthesiology 2002;96:A556.
23. Barker SJ. The performance of six "motion-resistant" pulse oximeters during motion, hypoxemia, and low perfusion in volunteers. Anesthesiology 2001;95:A587.
24. Pologe JA, Tobin RM. Method and apparatus for improved photoplethysmographic perfusion-index monitoring. US patent 5,766,127. 1998.
25. http://www.nellcor.com.
26. Lima AP, Beelen P, Bakker J. Use of peripheral perfusion index derived from the pulse oximetry signal as a noninvasive indicator of perfusion. Crit Care Med 2002;30(6):1210–1213.
27. Soubani AO. Noninvasive monitoring of oxygen and carbon dioxide. Am J Emerg Med 2001;19(2).
28. Shapiro BA, Harrison RA, Cane RD. Clinical Application of Blood Gases. 4th ed. Chicago: Year Book Medical Publishers, Inc.; 1989.
29. Williams AJ. ABC of oxygen: Assessing and interpreting arterial blood gases and acid–base balance. Br Med J 1998;317(7167): 1213–1216.
30. Poets FC, Southall DP. Noninvasive monitoring of oxygenation in infants and children: Practical considerations and areas of concern. Pediatrics 1994;93(5): 737–746.
31. Severinghaus JW, Naifeh KH, Koh SO. Errors in 14 pulse oximeters during profound hypoxia. J Clin Monit 1989;5: 72–81.
32. Clayton DG, Webb RK, Ralston AC, et al. A comparison of the performance of 20 pulse oximeters under conditions of poor perfusion. Anaesthesia 1991;46:3–10.
33. Sidi A, Paulus DA, Rush W, et al. Methylene blue and indocyanine green artifactually low pulse oximetry readings of oxygen saturation. Studies in dogs. J Clin Monit 1987; 3:249–256.
34. Lynn LA. Interpretive Oximetry: Future Directions for Diagnostic Applications of SpO_2 Time-Series. Anesthesia Analgesia 2002;94:S84–S88.
35. Zimmerman JL, Dellinger RP. Initial evaluation of a new intra-arterial blood gas system in humans. Crit Care Med 1993;21:495–500.
36. Ganter M, Zollinger A. Continuous intravascular blood gas monitoring: development, current techniques, and clinical use of a commercial device. Br J Anaesth 2003;91:397–407.
37. Coule LW, Truemper EJ, Steinhart CM, Lutin WA. Accuracy and utility of a continuous intra-arterial blood gas monitoring system in pediatric patients. Crit Care Med 2001;29(2): 420–426.
38. Dildy GA. Intrapartum fetal pulse oximetry: past, present, and future. Am J Obstet Gynecol 1996;175(1): 1–9.
39. Dildy GA, Van den Berg PP, Katz M, et al. Intrapartum fetal pulse oximetry: fetal oxygen saturation trends during labor and relation to delivery outcome. Am J Obstet Gynecol 1994; 171:679–684.
40. Papiernik E. Fetal pulse oximetry: correlation between changes in oxygen saturation and neonatal outcome. Eur J Obstet Gynecol Reprod Biol 1994;57:73–77.
41. McNamara H, Chung DC, Lilford R, Johnson N. Do fetal pulse oximetry readings at delivery correlate with cord blood oxygenation and acidaemia. Br J Obstet Gynaecol 1992;99: 735–738.
42. Severinghaus JW. The current status of transcutaneous blood gas analysis and monitoring. Blood Gas News 1998; 9(2).
43. Franklin ML. Transcutaneous measurement of partial pressure of oxygen and carbon dioxide. Respir Care Clin North Am 1995;1:119–131.
44. Padberg FT, Back TL, Thompson PN, et al. Transcutaneous oxygen ($TcPO_2$) estimates probability of healing in the ischemic extremity. J Surg Res 1996;60:365–369.
45. Rooke TW. The use of transcutaneous oximetry in the noninvasive vascular laboratory. Int Angiol 1992;11(1): 46–40.
46. Tremper KK, Barker SJ. Fundamental principles of monitoring instrumentation. In: Miller RD, editor. Anesthesia. Vol I, 3rd ed. New York: Churchill Livingstone; 1990. p 957–999.
47. Raemer BD, Philip JH. Monitoring anesthetic and respiratory gases. In: Blitt CE, editor. Monitoring in Anesthesia and Critical Care Medicine. 2nd ed. New York: Churchill Livingstone; 1990. p 373–386.
48. Graybeal JM, Russell GB. Relative agreement between Raman and mass spectrometry for measuring end-tidal carbon dioxide. Respir Care 1994;39:190–194.
49. Mollgaard K. Acoustic gas measurement. Biomed Instr Technol 1989;23:495–497.
50. Block FE, McDonald JS. Sidestream versus mainstream carbon dioxide analyzers. J Clin Monit 1992;8:139–141.
51. Casti A, Gallioli G, Scandroglio G, Passaretta R, Borghi B, Torri G. Accuracy of end-tidal carbon dioxide monitoring using the NBP-75 microstream capnometer. A study in intubated ventilated and spontaneously breathing nonintubated patients. Euro J Anesthesiol 2000;17:622–626.
52. AARC Clinical Practice Guidelines. Capnography/capnometry during mechanical ventilation (2003 update). Respir Care 2003;48:534–539.
53. Shibutani K, Muraoka M, Shirasaki S, Kubal K, Sanchala VT, Gupte P. Do changes in end-tidal PCO_2 quantitatively reflect changes in cardiac output? Anesth Analg 1994;79(5): 829–833.
54. Rackow EC, et al. Sublingual capnometry and indexes of tissue perfusion in patients with circulatory failure. Chest 2001;120:1633–1638.
55. Weil MB, et al. Sublingual capnometry: a new noninvasive measurement for diagnosis and quantitation of severity of circulatory shock. Crit Care Med 1999;27:1225–1229.
56. Marik PE. Sublingual capnography: a clinical validation study. Chest 2001;120:923–927.

57. Kelly JS, Wilhoit RD, Brown RE, James R. Efficacy of the FEF colorimetric end-tidal carbon dioxide detector in children. Anesth Analg 1992;75:45–50.
58. Nakatani K, Yukioka H, Fujimori M, et al. Utility of colorimetric end-tidal carbon dioxide detector for monitoring during prehospital cardiopulmonary resuscitation. Am J Emerg Med 1999;17:203–206.
59. Ornato JP, Garnett AR, Glauser FL, Virginia R. Relationship between cardiac output and the end-tidal carbondioxide tension. Ann Emerg Med 1990;19:1104–1106.
60. White RD, Asplin BR. Out of hospital quantitative monitoring of end-tidal carbondioxide pressure during CPR. Ann Emerg Med 1994;23:25–30.

See also CHROMATOGRAPHY; FIBER OPTICS IN MEDICINE; PERIPHERAL VASCULAR NONINVASIVE MEASUREMENTS.

BLOOD PRESSURE MEASUREMENT

CAN ISIK
Electrical Engineering and Computer Science Department, Syracuse University Syracuse, New York

INTRODUCTION

Blood pressure is an important signal in determining the functional integrity of the cardiovascular system. Scientists and physicians have been interested in blood pressure measurement for a long time. The first blood pressure measurement is attributed to Reverend Stephen Hales, who in the early eighteenth century connected water-filled glass tubes in the arteries of animals and correlated their blood pressures to the height of the column of fluid in the tubes. It was not until the early twentieth century that the blood pressure measurement was introduced into clinical medicine, albeit with many limitations.

Blood pressure measurement techniques are generally put into two broad classes: direct and indirect. Direct techniques of blood pressure measurement, which are also known as invasive techniques, involve a catheter to be inserted into the vascular system. The indirect techniques are noninvasive, with improved patient comfort and safety, but at the expense of accuracy. The accuracy gap between the invasive and the noninvasive methods, however, has been narrowing with the increasing computational power available in portable units, which can crunch elaborate signal processing algorithms in a fraction of a second.

During a cardiac cycle, blood pressure goes through changes, which correspond to the contraction and relaxation of the cardiac muscle, with terminology that identifies different aspects of the cycle. The maximum and minimum pressures over a cardiac cycle are called the systolic and diastolic pressures, respectively. The time average of the cardiac pressure over a cycle is called the mean pressure, and the difference between the systolic and diastolic pressures is called the pulse pressure.

Normal blood pressure varies with age, state of health, and other individual conditions. An infant's typical blood pressure is 80/50 mmHg (10.66/6.66 kPa) (systolic/diastolic). The normal blood pressure increases gradually and reaches 120/80 (15.99/10.66 kPa) for a young adult. Blood pressure is lower during sleep and during pregnancy. Many people experience higher blood pressures in the medical clinic, a phenomenon called the "white coat effect." Therefore, the ranges given in Table 1 are used as guidelines rather than as diagnostic facts.

Table 1. Classification of Blood Pressure for Adults

Category	Systolic—mmHg		Diastolic—mmHg
Normal	<120	and	<80
Prehypertension	120–139	or	80–89
Stage 1 Hypertension	140–159	or	90–99
Stage 2 Hypertension	160 or higher	or	100 or higher

DIRECT TECHNIQUES

The operation of direct measurement techniques can be summarized in very simple terms: They all use a pressure transducer that is coupled to the vascular system through a catheter or cannula that is inserted to a blood vessel, followed by a microcontroller unit with electronics and algorithms for signal conditioning, signal processing, and decision making. There are many advantages of this set of techniques, including:

- The pressure is measured very rapidly, usually within one cardiac cycle.
- The measurement is done to a very high level of accuracy and repeatability.
- The measurement is continuous, resulting in a graph of pressure against time.
- The measurement is motion tolerant.

Therefore, the direct techniques are used when it is necessary to accurately monitor patients' vital signs, for example, during critical care and in the operating room. Although direct techniques have a lot in common, there are differences in the details of various approaches.

Extravascular Transducers

The catheter in this type of device is filled with a saline solution, which transmits the pressure to a chamber that houses the transducer assembly. As a minor disadvantage, this structure affects the measured pressure through the dynamic behavior of the catheter. As the catheter has a known behavior, this effect can be minimized to insignificant levels through computational compensation (1).

Intravascular Transducers

The transducer is at the tip of the catheter in this type of device. Then the measured signal is not affected by the hydraulics of the fluid in the catheter. The catheter diameter is larger in this class of transducers.

Transducer Technology

A wide spectrum of transducer technologies is available to build either kind of transducer. They include metallic or semiconductor strain is gauges, piezoelectric, variable capacitance, variable inductance, and optical fibers. Appropriate driver and interface circuitry accompanies each technology (2).

Other Applications of Direct Pressure Measurement

Another advantage of direct measurement techniques is that they are not limited to measuring the simple arterial pressure. They can be used to obtain central venous, pulmonary arterial, left atrial, right atrial, femoral arterial, umbilical venous, umbilical arterial, and intracranial pressures by inserting the catheter in the desired site (3).

Sources of Errors

Direct blood pressure measurement systems have the flexibility of working with a variety of transducers/probes. It is important that the probes are matched with the appropriate compensation algorithm. Most modern equipment does this matching automatically, eliminating the possibility of operator error. An additional source of error occurs when air bubbles get trapped in the catheter. This changes the fluid dynamics of the catheter, causing an unintended mismatch between the catheter and its signal processing algorithm. This may cause distortions in the waveforms and errors in the numeric pressure values extracted from them. It is difficult to recognize this artifact from the waveforms, so it is best to avoid air bubbles in the catheter.

NONINVASIVE (INDIRECT) TECHNIQUES

An overwhelming majority of blood pressure measurements do not require continuous monitoring or extreme accuracy. Therefore, noninvasive techniques are used in most cases, maximizing patient comfort and safety. Currently available devices for noninvasive measurement are

- Manual devices: These devices use the auscultatory technique.
- Semiautomatic devices: These devices use oscillatory techniques.
- Automatic devices: Although most of these devices use oscillatory techniques, some use pulse-wave velocity or plethysmographic methods.

The Auscultatory Technique

In the traditional, manual, indirect measurement system, an occluding cuff is inflated and a stethoscope is used to listen to the sounds made by the blood flow in the arteries, called Korotkov sounds. When the cuff pressure is above the systolic pressure, blood cannot flow, and no sound is heard. When the cuff pressure is below the diastolic pressure, again, no sound is heard. A manometer connected to the cuff is used to identify the pressures where the transitions from silence to sound to silence are made. This combination of a cuff, an inflating bulb with a release valve, and a manometer is called a sphygmomanometer and the method an auscultatory technique. Usually, the cuff is placed right above the elbow, elevated to the approximate height of the heart, and the stethoscope is placed over the brachial artery. It is possible to palpate the presence of pulse under the cuff, rather than to use a stethoscope to listen to the sounds. The latter approach works especially well in noisy places where it is hard to hear the heart sounds.

Figure 1. Blood pressure waveform, and systolic, diastolic, and mean pressures, from an invasive monitor screen (4).

This method has various sources of potential error. Most of these sources are due to misplacement of the cuff, problems with hearing soft sounds, and using the wrong cuff size. Using a small cuff on a large size arm would result in overestimating the blood pressure, and vice versa. Nevertheless, an auscultatory measurement performed by an expert healthcare professional using a clinical grade sphygmomanometer is considered to be the gold standard in noninvasive measurements.

Oscillatory Techniques

Most automatic devices base their blood pressure estimations on the variations in the pressure of the occluding cuff, as the cuff is inflated or deflated. These variations are due to the combination of two effects: the controlled inflation or deflation of the cuff and the effect of the arterial pressure changes under the cuff. The Korotkov sounds are not used in the oscillatory techniques.

The cuff pressure variation data may be collected while the cuff is being inflated or deflated. Furthermore, the inflation or deflation during the data collection may be controlled in a continuous fashion or in a step-wise fashion. This variability gives four different strategies in data collection. Their differences may seem insignificant at first, but they have significant effects on the way a variety of algorithms are designed.

Data in Fig. 2 were collected using an experimental system. The cuff is first rapidly inflated to a value higher than the anticipated systolic pressure, an approximate pressure of 170 mmHg (22.66 kPa) in this case. Then it is deflated in small steps until the cuff pressure is below the anticipated diastolic pressure, $\sim$50 mmHg (6.66 kPa). Please note that when the cuff pressure is very high or very low, the arterial blood pressure variations contribute very little to the cuff pressure trajectory. As a matter of fact, the height of those pulses above the cuff pressure baseline is at their maximum when the baseline pressure is equal to the mean arterial pressure (MAP). We demonstrate this in Fig. 3, with a plot of pulses relative to their baseline pressure (pulse-wave amplitude), against their respective baseline cuff pressures. Please note that only

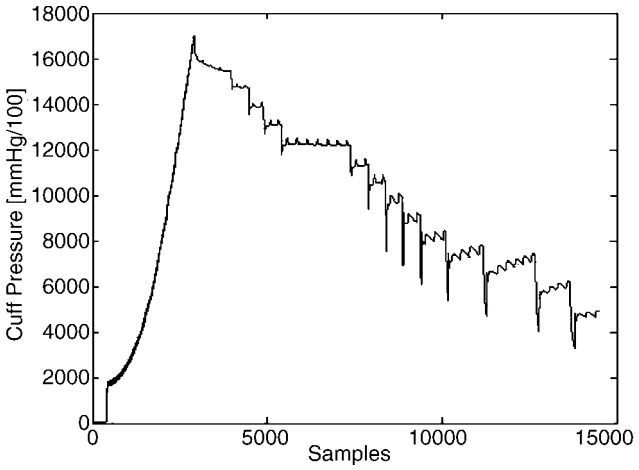


Figure 2. Cuff pressure trajectory when data are collected during step-wise deflation of the cuff (5).

a few of the pulses observed in Fig. 2 are transferred to Fig. 3 to maintain clarity.

Figure 4 shows a cycle of data collected during the continuous inflation of the cuff as well as the pulse-wave amplitude. The pulse-wave amplitude is obtained by subtracting the baseline cuff pressure from the raw pressure data. Next, we will return to the example developed in Figs. 2 and 3 and continue with the estimation of blood pressure values.

It seems trivial to pick the pulse with the tallest height above baseline and to select its baseline pressure to be the MAP. So, for the example at hand, MAP would be just under 100 mmHg (13.33 kPa), as shown in Fig. 5. The systolic and diastolic pressures are then estimated from the MAP using a variety of heuristic rules. A common class of these heuristic rules works as follows. First, the peak values (heights) of the pulse-wave amplitudes are connected to form an envelope. Again, the baseline pressure at the peak of this envelope is the MAP value. Then, the height of the MAP pulse is reduced by a predetermined systolic ratio, and the intersection of this "systolic height" with the envelope to the right of the MAP pulse is selected as the systolic location. The baseline pressure at this location is assigned as the estimate of the systolic pressure, as depicted in Fig. 5. The diastolic pressure is estimated in a similar fashion by using a ratio of its own to arrive at the

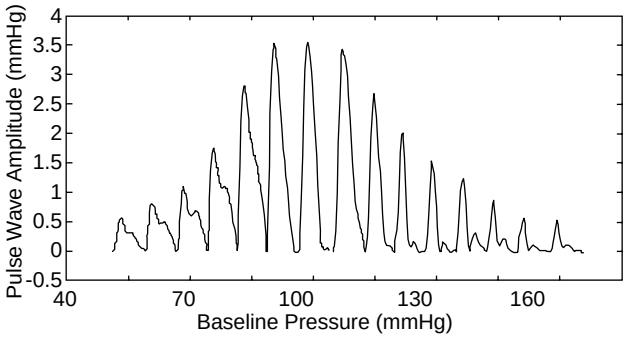


Figure 3. Pulse-wave amplitude profile (6).

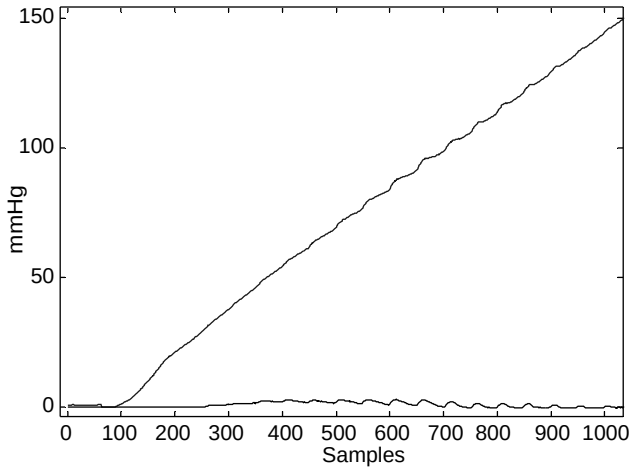


Figure 4. Cuff pressure trajectory and pulse-wave amplitude when data are collected during continuous inflation of the cuff.

diastolic height and then by finding the corresponding intersection with the envelope to the left of the MAP pulse.

In the example shown in Fig. 5, the systolic ratio and diastolic ratio were arbitrarily selected as 0.5 and 0.7, respectively. In a realistic system, those ratios would be found statistically (using methods such as regression, fuzzy rule-based systems, neural networks, or evolutionary algorithms) to minimize deviations between estimated and actual blood pressure values.

Algorithmic Components of Blood Pressure Measurement

In the earlier measurement units, it was a combination of hardware and software that controlled the various aspects of the automated measurement (or estimation) of blood pressure. With the ever increasing computational power of microcontrollers, all decision making and control are now implemented in software and with more elaborate algorithms. Here are some functions that are included in a measurement system. Please refer to Fig. 6 for a typical organization of such algorithms in a blood pressure measurement system.

- **Inflation/deflation control**: Whether data collection is done during inflation or deflation, continuously

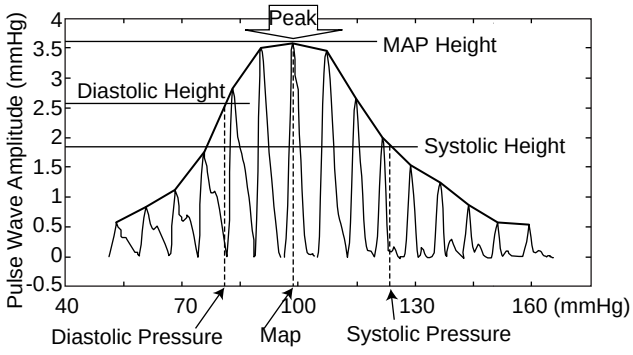


Figure 5. Blood pressure estimation from pulse-wave amplitude profile.

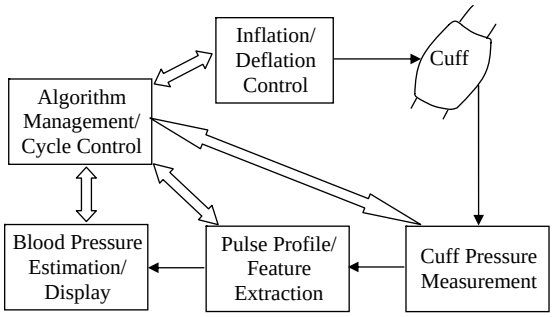


Figure 6. A typical organization of algorithmic components of (oscillatory) blood pressure measurement.

or in steps, there are many challenges to appropriately controlling the air pump. They include maintaining a smooth baseline cuff pressure without filtering out the arterial variations; adjusting the pump speed to variations arising from different cuff sizes, arm sizes, and cuff tightness; and selecting the range of cuff pressures for which data will be collected.

- **Pulse detection**: This is a fundamental part of extracting features from raw cuff pressure data. It becomes especially challenging when conditions such as arrhythmia, or tremors, affect the regularity of pulses. Pattern recognition techniques with features found in time, frequency, or wavelet domains are used to deal with difficult situations.
- **Blood pressure estimation**: The indirect method of measurement is a process of estimating pressures with the use of features extracted from cuff-pressures or other transducer data. This algorithm used to be limited to linear interpolation, as described in the example of Fig. 5. Recently, more elaborate decision-making and modeling tools such as nonlinear regression, neural networks, and fuzzy logic also are being used for this purpose.

Sources of Inaccuracy

Many factors contribute to the inaccuracies in the automated measurement of blood pressure. The following are some of the more significant sources of error:

- **Sparseness of data**: An important design criterion of a blood pressure monitor is to go through a cycle as quickly as possible. However, the faster a device functions, the fewer pulses it will have in a cycle. A cycle time of 1 min would yield about 60–70 pulses, whereas a 20-min cycle would have only 20–23 pulses. The oscillatory techniques are based on collecting cuff pressure due to pulses at baseline pressures that change from above systolic to below diastolic. If we divide a cuff pressure range of about 150–180 mmHg (10.99–23.99 kPa) by the number of pulses in a cycle, we can see that the baseline increment between successive pulses varies from 2 to 3 mmHg (0.26 to 0.39 kPa) in a 1 min cycle to 6 to 9 mmHg (0.79 to 1.19 kPa) in a 20 s cycle. This quantization error affects the accuracy in the estimate of the mean arterial pressure as well as the shape of the pulse envelope, hence, the accuracy of the systolic and diastolic values. Various curve-fitting and interpolation techniques are used to remedy this problem.
- **Pulse extraction uncertainty**: Whether the baseline cuff pressure is varied continuously or in steps, figuring out where one pulse ends and another one starts is not a trivial matter. An inspection of Fig. 2 will show that many artifacts in the data stream may confuse a pulse extraction algorithm and cause errors in the pulse-wave amplitude profile in Fig. 3. In addition, common factors such as an irregularity in the pulses as in arrhythmia, small wrinkles, or folds in the cuff changing its volume suddenly during data collection, or small movements of the patient may amplify those artifacts. A variety of pattern recognition techniques are employed to improve the accuracy of pulse detection (7).
- **Motion artifacts**: The performance of the oscillatory techniques depends on all measurements during a cycle. Therefore, any error caused by a motion of the patient may affect the accuracy of the blood pressure estimations. A comparative study of six noninvasive devices has found that average percent errors due to motion artifacts may be as high as 39% (8). Remedies to this source of error may be a combination of three strategies: (*1*) to identify and compensate for minor artifacts, (*2*) to identify and discard data that include significant artifacts or to repeat the entire cycle if the estimates are deemed unreliable, and (*3*) to incorporate features from additional sensors or monitors such as electrocardiogram (EKG) to help identify motion artifacts (8,9).

Other Blood Pressure Measurement Techniques

Oscillometry is by far the most common technique in automatic noninvasive blood pressure measurement. However, other methods are found in commercial units or in units that are being developed. In this section, a few of these methods are summarized and references are given for further information. It should be noted that algorithmic components and sources of inaccuracy presented within the context of oscillatory technique may apply to other automated measurement methods.

Arterial Tonometry. This relatively new technique in blood pressure measurement is inspired by the tonometry devices that were made in the mid-1950s to measure intraocular pressure. The arterial tonometry device is based on a pressure sensor and pneumatic actuator combination, which is placed on the wrist, above the radial artery. When the pressure applied on the artery is adjusted to the appropriate level (called the hold-down pressure), the portion of the artery wall that is facing the actuator is partially flattened. This configuration maximizes the energy transfer between the artery and the sensor, yielding pulses with the highest amplitude. The relative amplitudes of the tonometry pulses are calibrated to the systolic and the diastolic pressures. Tonometry is suitable for continuous monitoring applications. Sensor placement sensiti-

vity, calibration difficulties, and motion sensitivity are problems that need improvement (10,11).

Pulse-Wave Velocity. A pulse wave is generated by the heart as it pumps blood, and it travels ahead of the pumped blood. By solving analytical equations of fluid dynamics, it has been shown that changes in blood pressure heavily depend on changes in pulse-wave velocity. Blood pressure can be continuously calculated from pulse wave velocity, which in turn is calculated from EKG parameters and peripheral pulse wave measured by an SpO_2 probe on the finger or toe. This method is suitable for continuous monitoring as well as for detecting sudden changes in blood pressure to trigger an oscillometric cycle (12).

Plethysmographic Methods. In this method, changes in the blood volume during a cardiac cycle are sensed using a light emitter and receiver at the finger. Tissue and blood have different infrared light absorbance characteristics. That is, the tissue is practically transparent to the infrared light, whereas blood is opaque to it. A prototype of a ring-like sensor/signal processor/transmitter combination has been reported (13,14)

DIFFERENT FORMS OF BLOOD PRESSURE MEASUREMENT DEVICES

The techniques, algorithms, and transducers discussed in the previous sections have led to a variety of forms of devices, differentiated by where in the body the measurements are taken, or for what purpose the device is used.

Ambulatory Blood Pressure Monitoring

These portable and wearable devices monitor the patient's blood pressure over a long period, say for 24 h. While the patient is following her daily routine, the device periodically takes measurements and saves the results. These measurements are later downloaded for analysis by a physician. The first ambulatory devices, introduced in the early 1960s, were rudimentary and used tape recorders to capture the Korotkoff sounds with an occluding cuff. Most current ambulatory devices use the oscillatory technique. As the patient is subjected to repeated blood pressure measurements with an ambulatory device, it is essential to improve motion tolerance, patient comfort, measurement time, and of course overall accuracy of measurement algorithms that are employed in ambulatory monitors (15).

Ambulatory devices have been instrumental in clinical research and practice. Through their use, there have been significant improvements in our understanding of blood pressure dynamics in a variety of physiological and psychological conditions, and concepts such as "white-coat hypertension," "episodic hypertension," and "circadian rhythm of blood pressure" (e.g., daytime/nighttime variations of blood pressure) have been investigated and added to the medical lexicon (16).

Wrist Blood Pressure Monitoring

These monitors have smaller cuffs than their upperarm-attached counterparts. Hence, they are more compact and more conducive to self-measurement. It is important that the monitors are held at the heart level for correct measurement. They are popular with the home users but typically less accurate than the full-size arm monitors.

Finger Blood Pressure Monitoring

Finger monitors are not nearly as common as the arm or wrist monitors. The approaches used are auscultatory and plethysmographic.

Semiautomatic Blood Pressure Monitoring

The semiautomatic devices have cuffs that are inflated manually by an attached bulb, like a sphygmomanometer. Once the cuff is inflated, the monitor functions in the same manner as an automatic device, taking cuff-pressure measurements while releasing the pressure in a controlled way. These devices are more economical and have longer battery lives than their fully automated counterparts.

ACCURACY OF BLOOD PRESSURE MEASUREMENT DEVICES

Blood pressure measurement devices play an important role in medicine, as they measure one fundamental vital sign. In addition to this traditional use, noninvasive blood pressure devices, especially the automated ones, have become ubiquitous in the home, regularly used by lay people. Two widely used protocols for testing the accuracy of these devices are those set by the Association for the Advancement of Medical Instrumentation (AAMI), a pass/fail system published in 1987 and revised in 1993, and the protocols of the British Hypertension Society (BHS), an A–D graded system, established in 1990 and revised in 1993. These protocols describe in detail the process manufacturers should follow in validating the accuracy of their devices. Their numeric accuracy thresholds can be summarized as follows. A device would pass the AAMI protocols if its measurement error has a mean of no >5 mmHg (0.66 kPa) and a standard deviation of no >8 mmHg (1.06 kPa). The BHS protocol would grant a grade of A to a device if in its measurements 60% of the errors are within 5 mmHg, 85% of the errors are within 10 mmHg (1.33 kPa), and 95% within 15 mmHg (1.99 kPa). BHS has progressively less stringent criteria for the grades of B and C, and it assigns a grade D if a device performs worse than C.

The European Society of Hypertension introduced in 2002 the International Protocol for validation of blood pressure measuring devices in adults (17). The working group that developed this protocol had the benefit of analyzing many studies performed according to the AAMI and BHS standards. One of their motivations was to make the validation process simpler, without compromising its ability to assess the quality of a device. They achieved it by simplifying the rules for selecting subjects for the study. Another change was to devise a multistage process that recognized devices with poor accuracy early on. This is a pass/fail process, using performance requirements with multiple error bands.

Whether blood pressure measurement devices are used by professionals or lay people, their accuracy is important.

Table 2. Summary of Accuracy of Blood Pressure Measurement Devices

Device Type	Number Surveyed	Recommended?		
		Yes	Questionable	No
Manual, clinical	4	1	1	2
Auto, clinical	6	3	2	1
Auto, home, arm	20	4	4	12
Auto, home, wrist	4	0	2	2
Ambulatory	50	26	5	19
Total	84	34	14	36

Yet, most devices in the market have not been evaluated for accuracy independently, using the established protocols (18). In their study, O'Brien et al. surveyed published independent evaluations of manual sphygmomanometers, automated devices for clinical use, and automated devices for personal use. If a device was found acceptable by AAMI standards, and received a grade of A or B by BHS standards, for both systolic and diastolic measurements, then it was "recommended". Otherwise it was not recommended. Few studies they surveyed had issues such as specificity, so devices reported in those studies were "questionably recommended."

Table 2 summarizes the result of their survey. It is interesting to note that of the four clinical grade sphygmomanometers, a kind that is highly regarded by health-care providers, only one was "recommended". Overall, the number of devices "not recommended" is more than the number of "recommended" devices. What one should take away from this analysis is that at every level of quality, price, and target market, it is essential to research the accuracy of a device before investing in it and relying on it.

BIBLIOGRAPHY

Cited References

1. Gibbs NC, Gardner RM. Dynamics of invasive pressure monitoring systems: Clinical and laboratory evaluation. Heart Lung 1988;17:43–51.
2. Webster JG, editor. Medical Instrumentation: Application and Design, 3rd ed. New York: Wiley; 1998.
3. Hambly P. Measuring the blood pressure. Update Anaesthesia 2000;11(6).
4. Philips Invasive Monitoring literature. Available at http://www.medical.philips.com/main/products/patientmonitoring/products/invasivepressure/.
5. Colak S, Isik C. Blood pressure estimation using neural networks. IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, Boston, July 2004.
6. Colak S, Isik C. Fuzzy pulse qualifier. 23rd International Conference of the North American Fuzzy Information Processing Society (NAFIPS 2004) Proceedings, Banff, June 2004.
7. Osowski S, Linh TH. ECG beat recognition using fuzzy hybrid neural network. IEEE Trans Biomed Eng 2001;48:1265–1271.
8. Revision Labs, Beaverton, OR, Noninvasive Blood Pressure Measurement and Motion Artifact: A Comparative Study, December 3, 1998. Available at http://www.monitoring.welchallyn.com/pdfs/smartcufwhitepaper.pdf.
9. Dowling Jr NB. Measuring blood pressure in noisy environments. US patent No. 6,258,037 B1, July 10, 2001.
10. Sato T, Nishinaga M, Kawamoto A, Ozawa T, Takatsuji H. Accuracy of a continuous blood pressure monitor based on arterial tonometry. Hypertension 1993;21:866–874.
11. Matthys K, Verdonck P. Development and modelling of arterial applanation tonometry: A review. Technol Health Care 2002;10:65–76.
12. Williams B. Pulse wave analysis and hypertension: Evangelism versus skepticism. J Hypertension 2004;22:447–449.
13. Yang BH, Asada HH, Zhang Y. Cuff-less continuous monitoring of blood pressure, d'Arbeloff Laboratory of Information Systems and Technology, MIT, Progress Report No. 2–5, March 31, 2000. Available at http://darbelofflab.mit.edu/ProgressReports/HomeAutomation/Report2-5/Chapter01.pdf.
14. Rhee S, Yang BH, Asada HH. Artifact-resistant power-efficient design of finger-ring plethysmographic sensors. IEEE Trans Biomed Eng 2001;48:795–805.
15. McGrath BP. Ambulatory blood pressure monitoring. Med J Australia 2002;176:588–592.
16. National High Blood Pressure Education Program (NHBPEP) Working Group Report On Ambulatory Blood Pressure Monitoring. NIH Publication 92-3028. Reprinted February 1992. Available at http://www.nhlbi.nih.gov/health/prof/heart/hbp/abpm.txt.
17. O'Brien E, Pickering T, Asmar R, Myers M, Parati G, Staessen J, Mengden T, Imai Y, Waeber B, Palatini P. Working Group on Blood Pressure Monitoring of the European Society of Hypertension International Protocol for validation of blood pressure measuring devices in adults. Blood Pressure Monitoring 2002;7:3–17. Available at http://www.eshonline.org/documents/InternationalPS2002.04.29.pdf.
18. O'Brien E, Waeber B, Parati G, Staessen J, Myers MG. Blood pressure measuring devices: Recommendations of the European Society of Hypertension. Br Med J 2001;398. Available at http://bmj.bmjjournals.com/cgi/content/full/322/7285/531.

Further Reading

O'Brien E, Atkins N, Staessen J. State of the market: A review of ambulatory blood pressure monitoring devices. *Hypertension* 1995;26:835–842.

U.S. Food And Drug Administration. Non-Invasive Blood Pressure (NIBP) Monitor Guidance. March 10, 1997. Available at http://www.fda.gov/cdrh/ode/noninvas.html.

See also ARTERIES, ELASTIC PROPERTIES OF; BLOOD PRESSURE, AUTOMATIC CONTROL OF; CAPACITIVE MICROSENSORS FOR BIOMEDICAL APPLICATIONS; LINEAR VARIABLE DIFFERENTIAL TRANSFORMERS.

BLOOD PRESSURE, AUTOMATIC CONTROL OF

YIH-CHOUNG YU
Lafayette College
Easton, Pennsylvania

INTRODUCTION

Arterial pressure is one of the vital indexes of organ perfusion in human bodies. Generally speaking, blood pressure is determined by the amount of blood the heart pumps and the diameter of the arteries receiving blood from the heart. Several factors influence blood pressure. The

nervous system helps to maintain blood pressure by adjusting the size of the blood vessels, and by influencing the heart's pumping action. The heart pumps blood to make sure a sufficient amount of blood circulates to all the body tissues for organ perfusion. The more blood the heart pumps and the smaller the arteries, the higher the blood pressure is. The kidneys also play a major role in the regulation of blood pressure. Kidneys secrete the hormone rennin, which causes arteries to contract, thereby raising blood pressure. The kidneys also control the fluid volume of blood, either by retaining salt or excreting salt into urine. When kidneys retain salt in the bloodstream, the salt attracts water, increasing the fluid volume of blood. As a higher volume of blood passes through arteries, it increases blood pressure.

Hypertension is defined as abnormal high systemic arterial blood pressure, systolic and diastolic arterial pressures > 140 and 95 mmHg (18.662 and 12.664 kPa). The causes of hypertension might be due to acute myocardial infarction, congestive heart failure, and malignant hypertension. Postoperative cardiac patients may experience hypertension because of pain, hypothermia, reflex vasoconstriction from cardiopulmonary bypass, derangement of the rennin-angiotension system, and ventilation difficulties. A prolonged postoperative hypertension could lead to complications, including myocardial ischemia, myocardial infarction, suture line rupture, excessive bleeding, and arrhythmia. As a result, clinical treatment to postoperative hypertension is needed to reduce the potential risk of complications.

Postoperative hypertension is usually treated pharmacologically in the intensive care unit (ICU). Sodium nitroprusside (SNP) is one of the most frequently used pharmaceutical agents to treat hypertensive patients and is a vasodilating drug that can reduce the peripheral resistance of the blood vessel, and thus causes the reduction of arterial blood pressure. A desired mean arterial pressure (MAP) can be achieved by monitoring MAP and regulating the rate of SNP infusion. The mean arterial pressure can be measured from a patient by using an arterial pressure transducer with appropriate signal amplification. Low pass filtering is used to remove high frequency noise in the pressure signal and provide MAP for monitoring purpose. Administration of SNP infusion could be performed by manual operation. The drug infusion rate should be adjusted frequently in response to the spontaneous pressure variation and patient's condition changes. In addition, blood pressure response to the drug infusion changes over time and varies from patient to patient. Therefore, this manual approach is extremely difficult and time consuming for the ICU personnel. As the result, the use of control techniques to regulate the infusion of the pharmaceutical agents and maintain MAP within a desired level automatically has been developed in the last 30 years.

IVAC Corporation developed an automatic device, TITRATOR, to infuse SNP and regulate MAP in postoperative cardiac patients in early 1990s. Clinical evaluation for the clinical impact of this device in multiple centers was reported by Chitwood et al. (1). Patients who participate in this trial were treated by either automatic or manual control. The automated group showed a significant reduction in the number of hypertensive episodes per patient. Chest tube drainage, percentage of patients receiving transfusion, and total amount transfused were all reduced significantly by the use of an automated titration system. Although TITRATOR was not commercialized successfully due to economic reasons, the promising clinical experiences encouraged future development of automatic blood pressure regulation devices.

An automatic blood pressure control system usually includes three components: sensors, a controller, and a drug delivery pump. This article provides an overview of automatic control schemes, including proportional-integral-derivative (PID) controllers, adaptive-controllers, rule-based controllers, and artificial neural network controllers that regulate mean arterial blood pressure using SNP. A brief description of each control strategy is provided, followed by examples from literature. Testing of the control performance in computer simulations, animal studies, and clinical trials, is also discussed.

CONTROL SCHEMES

PID Controller

The PID control of MAP determines the SNP infusion rate, $u(t)$, based on the difference between the desired output and the actual output,

$$u(t) = K_{\mathrm{P}} e(t) + K_{\mathrm{I}} \int_{t_0}^{t_1} e(t)dt + K_{\mathrm{D}} \frac{d}{dt} e(t) \qquad (1)$$

where $e(t) = P_{\mathrm{d}}(t) - P_{\mathrm{m}}(t)$, $P_{\mathrm{d}}(t)$ is the desired MAP, and $P_{\mathrm{m}}(t)$ is the actual mean arterial pressure. The parameters K_{P}, K_{I}, and K_{D} are the proportional, integral, and differentiation gain respectively. The design of this type of controller involves the selection of appropriate control gains, K_{P}, K_{I}, and K_{D}, such that the actual blood pressure, $P_{\mathrm{m}}(t)$, can be stabilized and maintained close to the desired level, $P_{\mathrm{d}}(t)$. Typical components of the automatic blood pressure control system, including the PID controller, the infusion pump, the patient, as well as the patient monitor along with physiologic sensors are illustrated in Fig. 1.

Sheppard and co-worker (2–4) developed a PI-type controller, by setting $K_{\mathrm{D}} = 0$ in (1), to regulate SNP, which has been tested over thousands of postcardiac-surgery patients in the ICU. The control gains were tuned to satisfy an acceptable settling time with minimal overshoot. The discrete-time PI controller updates the infusion rate as

$$u(k) = u(k-1) + \Delta u(k) \qquad (2)$$

where $u(k-1)$ is the previous infusion rate a minute ago and $\Delta u(k)$ is the infusion rate increment defined by,

$$\Delta u(k) = K\{0.4512\, e(k) + 0.4512\, [e(k) - e(k-1)]\} \qquad (3)$$

where $e(k)$ and $e(k-1)$ are the current and previous error, respectively. The gain K in Eq. 3 as well as the further correction of $\Delta u(k)$ were determined by the region of current MAP $P_m(k)$ as described in the following:

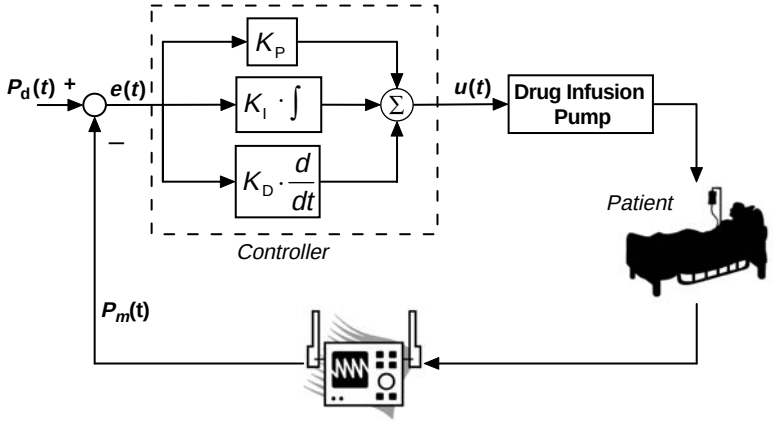


Figure 1. Proportional-integral-derivative control scheme of blood pressure control.

Rule 1 If $P_m(k) \geq P_d + 5$, then $K = -1$ and $\Delta u(k) = \Delta u(k)$ from Eqs. 3–2

Rule 2 If $P_d \leq P_m(k) < P_d + 5$, then $K = -0.5$ and $\Delta u(k) = \Delta u(k)$ from Eq. 3

Rule 3 If $P_d - 5 \leq P_m(k) < P_d$, then $K = -1$ and $\Delta u(k) = \Delta u(k)$ from Eq. 3

Rule 4 If $P_m(k) \leq P_d - 5$, then $K = -2$ and $\Delta u(k) = \Delta u(k)$ from Eq. 3

Rule 5 If $P_m(k) < P_d - 5$ and $\Delta u(k) > 0$, then $\Delta u(k) = 0$

Rule 6 If $P_m(k) \geq P_d$ and $\Delta u(k) > 7$, then $\Delta u(k) = 7$

These rules were designed to provide a boundary for the controller and achieve the optimal performance with the minimal pharmacological intervention. As a result, the controller is a nonlinear PI-type controller.

The automatic blood pressure controller described herein performed better than human operation in a comparison study (5). Automatic blood pressure regulation exhibits approximately one-half of the variation observed during manual control; MAP are more tightly distributed about the set-point, as shown in Fig. 2 (2). Forty-nine postcardiac surgery patients in ICU were managed by the automatic controller. The patients' MAPs were maintained within ± 5 mmHg (± 0.667 kPa) of the desired MAP 94% of the total operation time (103 out of the 110 operation hours). A group of 37 patients were managed with manual operation provided by experienced personals, with which only 52% of the time the patients' MAPs were within the prescribed range.

Adaptive Controller

The PID controller considered previously was with the control gains determined prior to their implementation. The control gains were usually tuned to satisfy the performance criterion in simulation or animal studies where the parameters characterizing the system dynamics were fixed variables. In clinical applications, the cardiovascular vascular dynamics change over time as well as from patient to patient. In addition, the sensitivity to drugs varies from one patient to another and even with the same patient at different instant. Therefore, it would be beneficial if the control gains can be adjusted automatically during operation to adapt the differences between patients as well as physiologic condition changes in a patient over time. This type of controllers is called adaptive controller.

An adaptive control system usually requires a model, representing plant (the patient and the drug infusion system) dynamics. Linear black box models, expressed by

$$\begin{aligned} y(k) &= \frac{B(q^{-1})}{A(q^{-1})}u(k) + \frac{C(q^{-1})}{A(q^{-1})}n(k) \\ A(q^{-1}) &= 1 + a_1 q^{-1} + a_2 q^{-2} + \cdots + a_n q^{-n} \\ B(q^{-1}) &= 1 + b_1 q^{-1} + b_2 q^{-2} + \cdots + b_l q^{-l} \\ C(q^{-1}) &= 1 + c_1 q^{-1} + c_2 q^{-2} + \cdots + c_m q^{-m} \end{aligned} \tag{4}$$

are typically used to represent the plant dynamics. A, B, and C are polynomials in the discrete shift operator q, where a_i, b_i, and c_i are coefficients in the polynomials; $y(k)$, $u(k)$, and $n(k)$ are the model input, output, and noise, respectively. Depending on the polynomials B and C, the model in Eq. 4 can be classified as autoregressive [AR, $B(q^{-1}) = 0$, $C(q^{-1}) = 1$], autoregressive with inputs [ARX, $C(q^{-1}) = 1$], autoregressive moving average [ARMA, $B(q^{-1}) = 0$], and autoregressive moving average with inputs (ARMAX). The coefficients of the polynomials are time-varying, much slower than the plant dynamic changes. The controller updates the control input, $u(k)$, by taking the model parameter changes into consideration. General reviews and descriptions on adaptive control theory can be found in literature (6–8). Three types of adaptive control schemes are frequently used in blood pressure controller design: self-tuning regulator, model reference adaptive control, and multiple model adaptive control.

Self-Tuning Regulator. The self-tuning regulator (STR) is based on the idea of separating the estimation of unknown parameters from the design of the controller. It is assumed that a priori knowledge of the model structure, that is, l, m, and n in Eq. 4. In choosing l, m, and n, one must compromise between obtaining an accurate representation of the system dynamics while keeping the system representation simple. The parameters of the regulator are adjusted by using a recursive parameter estimator and a

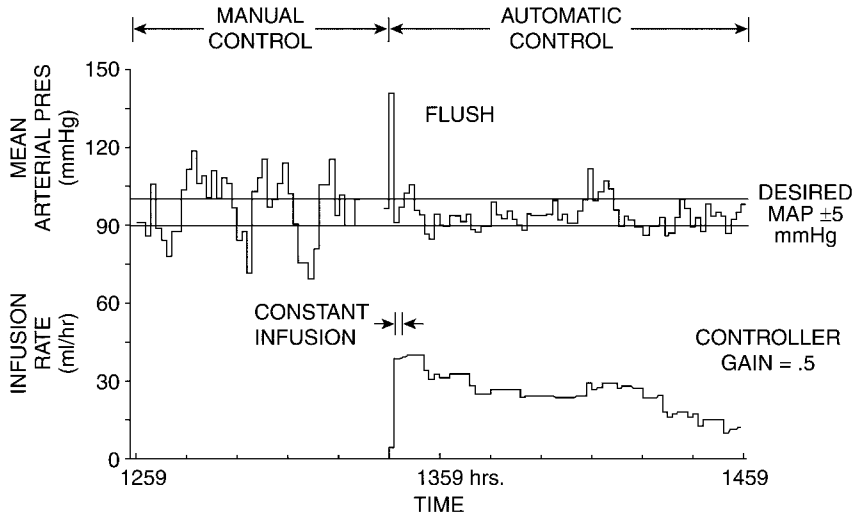


Figure 2. Comparison of manually controlled SNP infusion with computer control in the same patient. (Redrawn with permission from L.C. Sheppard, Computer control of the infusion of vasoactive drugs, *Ann. of Biomed. Eng.*, Vol. 8: 431–444, 1980. Pergamon Press, Ltd.)

regulator design calculation as shown in Fig. 3. The parameter estimates are treated as if they are true or at least asymptotically the true parameters. Several algorithms are available for parameter estimation, including recursive least-squares, generalized least-squares, stochastic approximation, maximum likelihood, instrumental variables, and Kalman filter. Every technique has its advantages and disadvantages. Descriptions of parameter estimation algorithms can be found in (9). Various approaches are available for regulator design calculation, such as minimum variance, gain and phase margin analysis, pole placement, and linear quadratic Gaussian (LQG). More detailed information of STR can be found in literature (6–8).

Various STR-type blood pressure controllers have been developed and tested in computer simulations, animal experiments, as well as clinical studies. Arnsparger et al. (10) used a second-order ARMA model to design the STR. A recursive least-mean-squares estimator was used to estimate the model parameters. The parameter estimates were then used to calculate the control signal, the drug infusion rate, based upon a minimum variance or a one-step-ahead control law. Both algorithms were implemented in microprocessor and tested in dog experiments for comparison. Both controllers were able to maintain the MAP at the desired level. However, the one-step-ahead controller performed better in the test with less variation in the infusion rate.

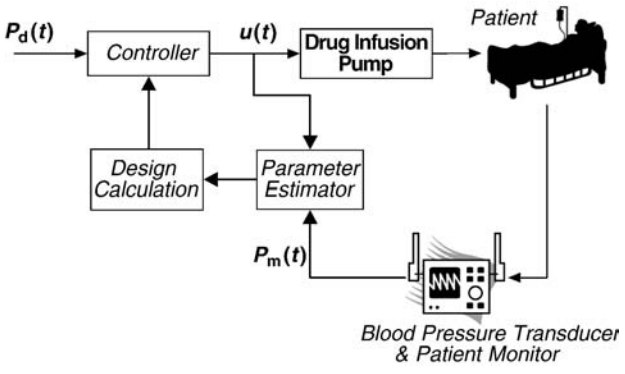


Figure 3. Configuration of self-tuning regulator for blood pressure control.

A combination of proportional derivative with minimum variance adaptive controller was designed by Meline et al. (11) to regulate MAP using SNP. The plant dynamics was represented by a fifth-order ARMAX model, while the model parameters were estimated through a recursive least-squares algorithm. The controller was tested on ten dog experiments as well as human subjects (12). Twenty patients with postsurgical hypertension were randomly assigned to either the manual group, where SNP was administrated by experience nurse, or the automatic group. Statistical analysis showed that MAP was maintained within $\pm$ 10% from the desired MAP for 83.3% of the total operation time in the "automatic" group versus 66.1% of the total operation time in the "manual" group. This implies the automatic control performed better than the manual operation.

A pole-assignment STR was designed by Mansour and Linkens (13) to regulate blood pressure using a fifth-order ARMAX model. The model parameters were identified through a recursive weighted least-squares estimator. These parameters were then used to determine appropriate feedback gains for the controller. Pole-placement algorithm was used because of its robustness to a system with nonminimum phase behavior or unknown time delay. Effectiveness of the controller was evaluated extensively in computer simulation, using a clinically validated model developed by Slate (3) as shown in Fig. 4(2). The controller demonstrated a robust performance even with the inclusion of the recirculation term or a variable time delay.

Voss et al. (14) developed a control advance moving average controller (CAMAC) to simultaneously regulate arterial pressure and cardiac output (CO) using SNP and dobutamine. CAMAC is a multivariable STR, which has the advantage of controlling nonminimum phase plants with unknown or varying dead times. The controller determines the drug infusion rates based on the desired MAP and CO, past inputs, past outputs, and a on-line recursive least-squares estimator with an exponential forgetting factor identifying the subject's response to the drugs.

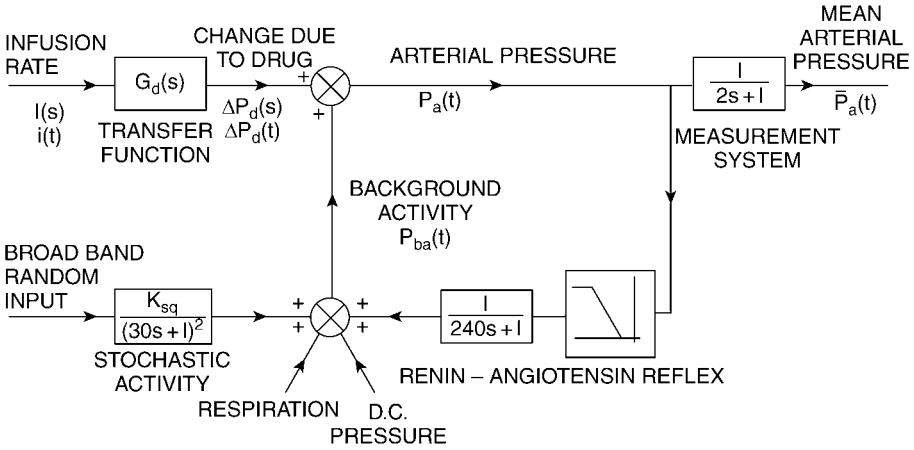


Figure 4. Model of MAP in response to SNP infusion. (Redrawn with permission from L.C. Sheppard, Computer control of the infusion of vasoactive drugs, *Ann. Biomed. Eng.* 1980; 8: 431–444. Pergamon Press, Ltd.)

The plant model for designing the controller and estimator was a second-order ARMAX model. The control algorithm was designed and tested in simulations prior to dog experiments. Although animal studies demonstrated that the controller was capable to maintain MAP and CO at their desired level, changing vasomotor tone and the lack of high frequency excitation signals could lead to inaccuracy in the parameter estimation, causing poor performance in transient response.

Model Reference Adaptive Control. The basic principle of the model reference adaptive control is illustrated in Fig. 5). The desired input–output response is specified by the reference model. The parameters of the regulator are adjusted by the error signal, the difference between the reference model output and the system output, such that the system output follows the reference output. More detailed information about MRAC can be found in Ref. 7.

The use of MRAC to regulate blood pressure was introduced by Kaufmann et al. (15). The format of the reference model was adopted form that developed by Slate (3). Controller design and evaluation were carried out in computer simulation. The controller with adaptation gains showed lower steady-state error than that with nonadaptive gains in simulations, particularly when a process disturbance was introduced. Animal studies were conducted to compare the performance of the MRAC with that of a well-tuned PI controller. Neosynephrine was introduced to change the transfer function characteristics of the subjects during experiments. The MRAC was superior to the PI controller and maintained MAP closed to the reference with an error within ± 5 mmHg (± 0.667 kPa) regardless of the plant characteristic changes due to drug intervention.

Pajunen et al. (16) designed a MRAC to regulate blood pressure using SNP with the ability to adjust the reference model by learning the patient's characteristics, represented by the model parameters, coefficients and time delays, of the transfer function. These model parameters were assumed to be unknown and exponentially time-varying. The time-varying reference model was automatically tuned to achieve the optimal performance while meeting the physical and clinical constraints imposed on the drug infusion rate and MAP. Extensive computer simulation was used to evaluate the robustness of the controller. The MAP was maintained within ± 15 mmHg (± 2 kPa) around the set-point regardless of changes in patient's characteristics and the presence of high level noises.

Polycarpou and Conway (17) designed a MRAC to regulate MAP by adjusting SNP infusion rate. The plant model was a second-order model discretized from the Slate's model (3). Time delay terms in the model were assumed to be known while the model parameters were constant with nonlinear terms. The constant terms were assumed to be known and the nonlinear terms were estimated by a radial basis function (RBF) neural network. The resulting parameter estimates were then used to update the control law such that the system output follows the reference model. Although the RBF was able to model the unknown nonlinearity and thus improve the closed-loop characteristics in computer simulation, the

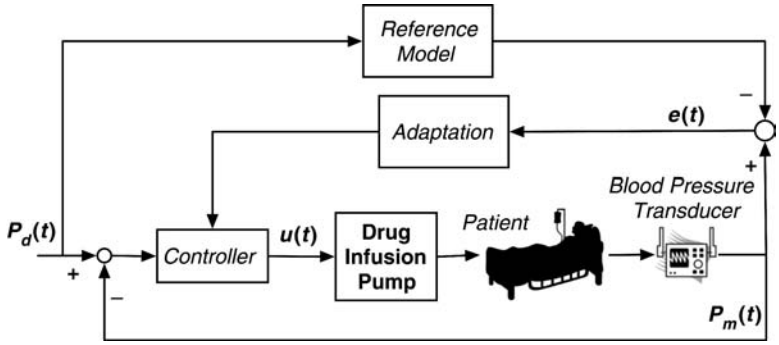


Figure 5. Configuration of MRAC for blood pressure regulation.

assumption that the model parameters and time delays were known would need further justification in practical applications.

Multiple Model Adaptive Control. The concept of multiple model adaptive control (MMAC) was first introduced by Lainiotis (18). This technique assumes that the plant response to the input can be represented by a bank of models. A controller is designed *a priori* to give a specified performance for each particular model. A probability, $P(q_i|t)$, describing the accuracy of each model, q_i, to represent the actual system, is calculated and used as the weighting factor to update the control input,

$$u = \sum_{i=1}^{N} u_i \cdot P(q_i|t) \tag{5}$$

where u_i is the control input based on the model q_i. As the response of the system changes, the probability, $P(q_i|t)$, will also be adjusted accordingly such that the model closest represents current dynamics gets the greatest probability. As a result, the contribution of the control input, obtained from the model with the greatest probability, to the updated control input in equation 5 is more significant than the inputs from other models with lower probabilities. Configuration of the MMAC is illustrated in Fig. 6.

He et al. (19) introduced the first blood pressure controller using the MMAC technique. There were eight plant models derived from Slate's model (3) for controller design. Each plant model contains a constant model gain between 0.32 and 6.8, representing the plant gain of 0.25–9 in Slate's model (3), along with the same time constants and delays at their nominal values. A proportional-plus-integral (PI) type controller was designed for each plant model. These controllers were with the same time constant but different gains. Computer simulation was used to test the controller performance in response to the variations of model parameters and the presence of background noise. The controller was able to settle MAP within 10 min with the error within ± 10 mmHg (± 1.333 kPa) from the set-point. The control algorithm was further tested in animal experiments. The controller stabilized MAP in < 10 min with ± 5 mmHg (± 0.667 kPa) error from its set-point, regardless of the plant characteristic changes due to neosynephrine injection, the sensitivity of the subject to the SNP infusion, and the background noise. The mean error was < 3 mmHg (0.4 kPa) over the entire studies.

Martin et al. (20) developed a MMAC blood pressure controller with seven models modified from Slate's model (3). The model gains in the seven models were from 0.33 to 9.03 to cover the variation of the plant gain between 0.25 and 10.86. The other model parameters were held constant at their nominal values. A pole-placement compensator was designed for each model. A Smith predictor was used to remove the effects of infusion delay, and thus simplify the control analysis and design. A PI unit was included to achieve zero steady-state error. Two constrains were used to limit the infusion rate when the patient's blood pressure is too low or the resulting SNP infusion rate from the controller is beyond the preset threshold. The controller was able to maintain MAP with the settling time < 10 min, the maximum overshoot < 10 mmHg (1.333 kPa), and the steady-state error within ± 5 mmHg (± 0.667 kPa) around the pressure set-point in computer simulations. The controller was also tested on 5 dogs as well as 19 patients during cardiac surgery with the aid of a supervisor module, which oversees the overall environment and thus improves the safety (21,22).

Yu et al. (23) designed a MMAC to control MAP and CO by adjusting the infusion rates of SNP and dopamine for congested heart failure subjects. There were 36 linear multiinput and multioutput (MIMO) models, represented by first-order transfer functions with time delays, to cover the entire range of possible dynamics. A model predictive controller [MPC, (24)] was designed for each individual model to find a sequence of control signals such that a quadratic cost function can be minimized. In order to save computation time, only the control signals corresponding to the six models with the highest probability weights were used to determine the drug infusion rates. The control

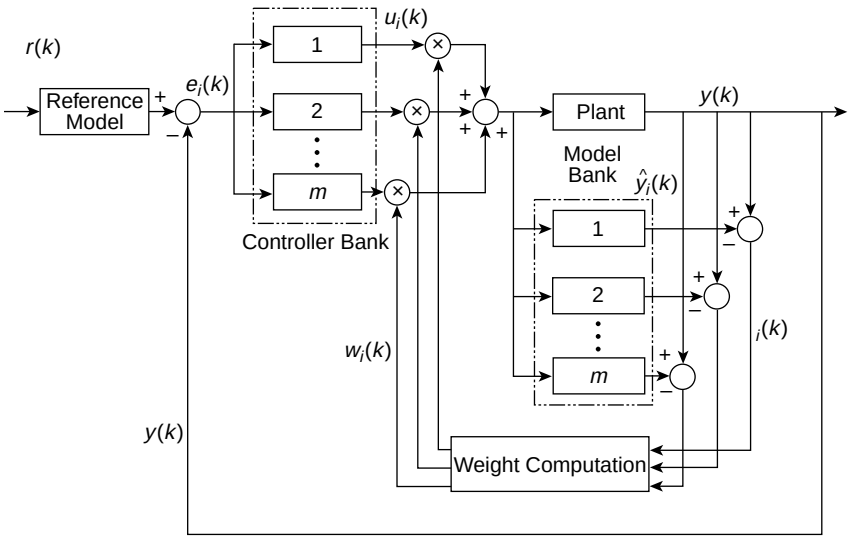


Figure 6. Block diagram of multiple model predictive control. [Redrawn with permission from Rao et al., Automated regulation of hemodynamic variables, *IEEE Eng. Med. Biol.* 2001; 20 (1): 24–38. (© Copyright 2004 IEEE).]

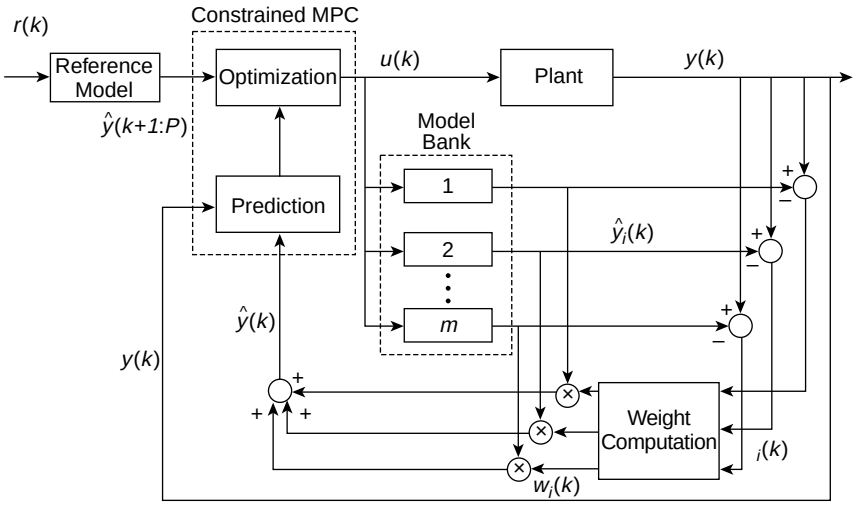


Figure 7. Modified multiple model predictive control strategy. [Redrawn with permission from Rao et al., Automated regulation of hemodynamic variables, *IEEE Eng. Med. Biol*. 2001; 20 (1): 24–38. (© Copyright 2004 IEEE).]

algorithm was tested in six dogs, including some cases with induced heart failure. It took 3–10.5 min to settle MAP within ± 5 mmHg (± 0.667 kPa) of the steady-state set-point with the mean of 5.8 min in all cases. The overshoot was between 0 and 12 mmHg (1.6 kPa) with the average of 5.92 mmHg (0.79 kPa). The standard deviation of MAP about its set-point was 4 mmHg (0.533 kPa).

The major challenge of implementing MPC in MMAC is the computation time, especially for a large model bank. Rao et al. (25) designed a MMAC with a single constrained MPC as shown in Fig. 7. The model bank, constituted first-order-plus-time-delay MIMO models spanning sufficient spectrum of model gains, time constants, and time delays, was run in parallel to obtain the possible input–output characteristics of a patient's response to drug dosages. A Bayesian weight was generated for each model based on the patient's response to drugs. The MPC used the combination of model weights to determine the optimal drug infusion rates. This control scheme combines the advantages of model adaptation according to patient variations, as well as the ability to handle explicit input and output constraint specifications. The controller effectively maintained MAP and cardiac output in seven canine experiments (26). Figure 8 illustrates the results of control MAP and CO using SNP and dopamine in on study. High levels of fluothane were introduced to reduce CO, mimicking congestive heart failure. The controller achieved both set-points of MAP = 60 mmHg (8 kPa) and CO = 2.3 $\text{L}\cdot\text{min}^{-1}$ in $\sim$ 12 min. In average over the entire studies, MAP was maintained within ± 5 mmHg (± 0.667 kPa) of its set-point 89% of the time with a standard deviation of 3.9 mmHg (0.52 kPa). Cardiac output was held within ± 1 $\text{L}\cdot\text{min}^{-1}$ of the set-point 96% of the time with a standard deviation of 0.5 $\text{L}\cdot\text{min}^{-1}$. Manual regulation was performed in the experiments for comparison. The MAP was kept within ± 5 mmHg (± 0.667 kPa) of its set-point 82% of the time with a standard deviation of 5.0 mmHg (0.667 kPa) while CO stayed in the ± 1 $\text{L}\cdot\text{min}^{-1}$ band of the set-point 92% of the time with a standard deviation of 0.6 $\text{L}\cdot\text{min}^{-1}$. Clearly, the automatic control performed better than the manual approach.

Rule-Based Controller

The blood pressure controllers discussed previously rely on mathematical models that can characterize plant dynamics, including the drug infusion system, human cardiovascular dynamics, and pharmacological agents. Identifying such mathematical forms could be a challenge due to the complexity of human body. Despite this, there exist experienced personnel, whose ability to interpret linguistic statements about the process and to reason in a qualitative fashion prompts the question: "can we make comparable use of this information in automatic controllers?"

In rule-based or intelligent control, the control law is generated from linguistic rules. This model-free controller usually consists of an inference engine and a set of rules for reasoning and decision making. A typical control rules are represented by *if* <*condition*> *then* <*action*> statements. Rule-based approaches have been proposed as a way of dealing with the complex natural of drug delivery systems and, more importantly, as a way of incorporating the extensive knowledge of clinical personnel into the automatic controller design.

One of the most popular rule-based control approaches is fuzzy control. Fuzzy control approach is based on fuzzy set theory and is a rule-based control scheme where scaling functions of physical variables are used to cope with uncertainty in the plant dynamics. A typical fuzzy controller, shown in Fig. 9, usually includes three components: (*1*) membership functions to fuzzify the physical input, (*2*) an inference engine with a decision rule base, and (*3*) a defuzzifier that converts fuzzy control decisions into physical control signals. More details on fuzzy set theory and its control applications are available in (27–29).

Isaka et al. (30) applied an optimization algorithm to determine the membership functions of a fuzzy blood pressure controller using SNP. This method reduced the time and efforts to determine appropriate values for a large number of membership functions. In addition, it also provided the knowledge of the effect of membership functions to the fuzzy controller performance, as well as the effect of

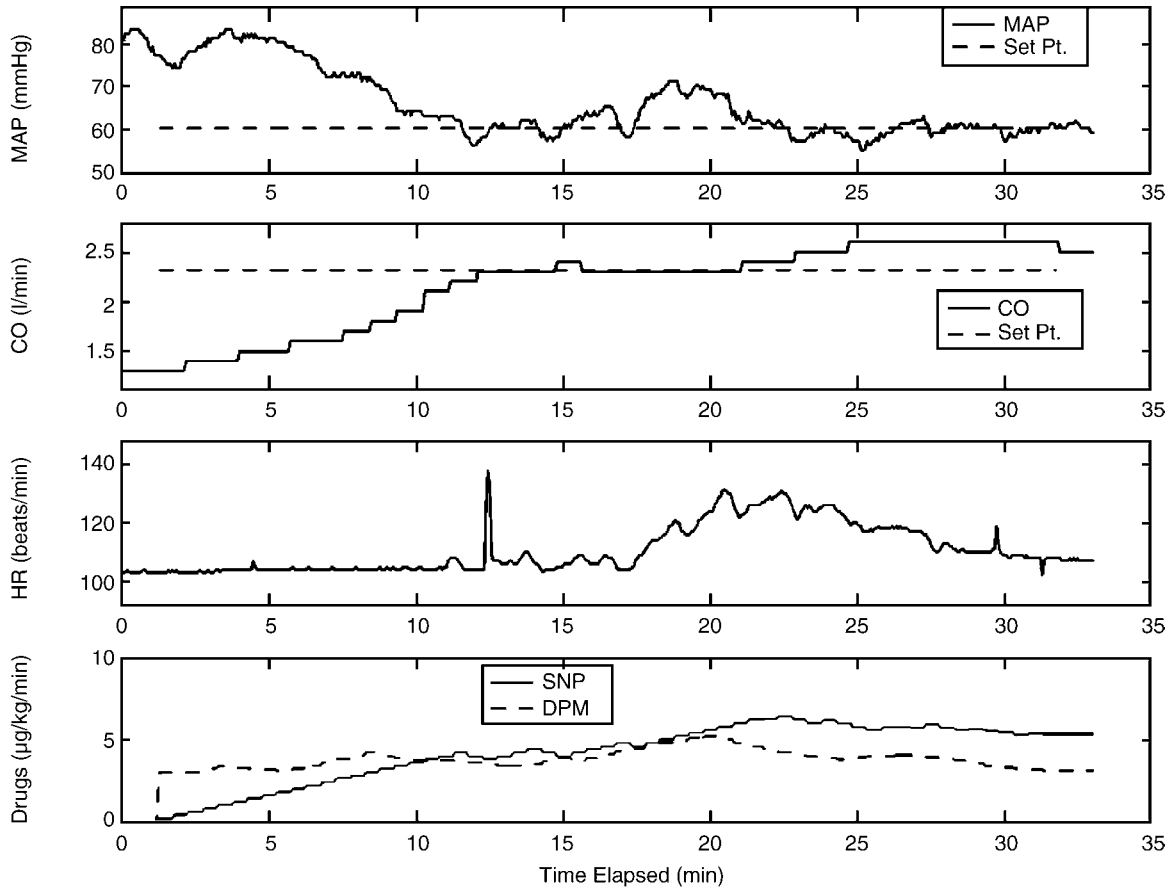


Figure 8. Multiple model adaptive control of MAP and CO using SNP and dopamine in canine experiment. (© Copyright 2004 IEEE).

plant parameter variations to the changes in membership functions. Efficacy of using this controller to regulate MAP by infusing SNP was evaluated in computer simulation model proposed by Slate (3). The MAP was initialized at 120 mmHg (16 kPa) at the beginning of simulation. The target MAP value was first set at 80 mmHg (10.665 kPa) and then changed to 110 mmHg (14.665 kPa). The target MAP values were achieved in < 3 min with overshoots < 10 mmHg (1.333 kPa).

Ying et al. (31) designed an expert-system-shell-based fuzzy controller to regulate MAP using SNP. The controller was a nonlinear PI-type control while the control gains were predetermined by analytically converting the fuzzy control algorithm. This converting process provided the advantage of execution time reduction. The controller was further fine-tuned to be more responsive to the rapid and large changes of MAP. It was successfully tested in 12 postsurgical patients for the total of 95 hs and 13 min. MAP was maintained within $\pm 10\%$ of its target value, 80 mmHg (10.665 kPa), 89.3% of the time over the entire test.

Neural-Network Based Controller

Artificial neural networks (ANN) are computation models that have learning and adaptation capabilities. An ANN-based controller is usually more robust than the traditional controllers in the presence of plant nonlinearity and uncertainty if the controller is trained properly. A survey article about the use of ANN in control by Hunt et al. (32) provides more detailed information.

The use of ANN-type controller in arterial blood pressure regulation was investigated in feasibility studies in either computer simulation or animal experiments. Chen et al. (33) designed an ANN-type adaptive controller to control MAP using SNP. The controller was tested in computer simulation with various gains and different levels of noise. The controller was able to maintain MAP close to the set point, 100 mmHg (13.33 kPa) with error within ± 15 mmHg (± 2 kPa) in an acceptable tolerance settling time < 20 min.

Kashihara et al. (34) compared various controllers, including PID, adaptive predictive control using ANN (APP_{NN}), a combined control of PID with APP_{NN}, a fuzzy controller, and a model predictive controller, to maintain MAP for acute hypotension using norepinephrine. The controllers were tested in computer simulation and animal studies. The controllers based on neural network approach were more robust in the presence of unexpected hypotension and unknown drug sensitivity. Adding an ANN or a fuzzy logic scheme to the PID or adaptive controller improved the ability of the controller to handle unexpected conditions more effectively.

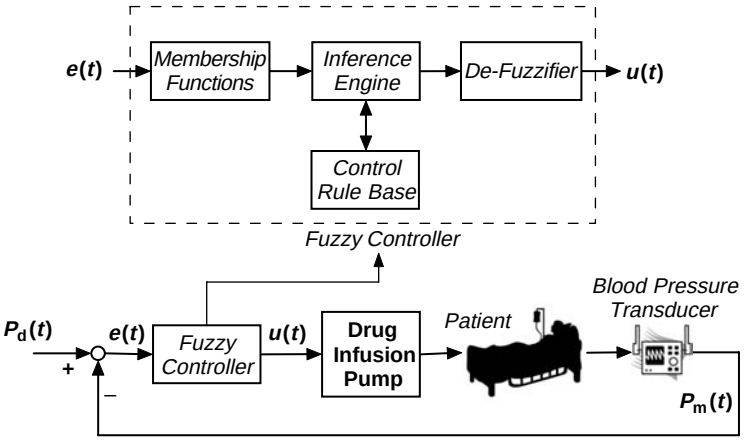


Figure 9. Block diagram of a fuzzy controller. [Redrawn with permission from Isaka et al., An optimization approach for fuzzy controller design, 1992; SMC 22: 1469–1473. (© Copyright 2004 IEEE).]

DISCUSSION

Numerous controllers have been developed since 1970s to regulate SNP and control MAP for hypertension patients. The control strategies can generally be classified as PID control, adaptive control (including STR, MRAC, and MMAC), rule-based control, as well as neural network control. Most controllers were developed and tested in computer simulation and animal experiments successfully. A few controllers were tested clinically with satisfactory results. Table 1 summarizes the control algorithms reviewed in this article.

Controller performance is influenced by several factors, including the fit of the process model to the plant, signal conditioning of the sensors under various clinical environments, as well as the diagnosis ability of the devices. Model selection is crucial for the stability and robustness of a controller. In blood pressure regulation, variable time delay, patient's sensitivity to SNP, and rennin regulatory mechanism are important factors. These factors could cause parameter variations in the plant model that might reduce the performance of a fixed-gain controller. Adaptive controllers that can adjust the control signal based on the estimation of model parameters or the probability of model errors could overcome the limit of the fixed-gain control. Some controllers have been tested in the laboratory with promising results. However, clinical applications of this type of controllers were very few. Rule-based and ANN controllers do not need a specific plant model for control design. The training signals or information must provide a broad coverage of possible events in the clinical environment to assure the reliability of the control algorithm.

Patient care practices and other aspects of the clinical environment must be considered in the design of a clinical useful system. A supervisory algorithm that can detect potential risks, determine appropriate control signals to

Table 1. Summary of Blood Pressure Controllers Reviewed in this Article

		Controller Performance			Controller Test		
Articles	Control Scheme	Settling Time (min)	Overshoot, mmHg	Steady-State about the Set-Point, mmHg	Simulation	Animal Studies	Clinical Studies
Slate et al.(2–4)	Nonlinear PI	<10		± 10	x	x	x
Arnsparger et al.(10)	STR	2	30	10		x	
Mansour et al.(13)	STR	5–20	<10	± 5	x		
Voss et al.(14)	CAMAC	1.3–7.3	0–22	−4–9.8	x	x	
Kaufmann et al.(15)	MRAC (w/known time-constant and delay)	<5		± 5	x	x	
Pajunen et al.(16)	MRAC (w/time-varying parameters)	<5	<15	± 15	x		
Polycarpou et al.(17)	MRAC	5		± 10	x		
He et al.(19)	MMAC	<8	<5	± 5	x	x	
Martin et al.(20–22)	MMAC	<10	<10	± 5	x	x	x
Yu et al.(23)	MMAC	3–10.5	0–12	± 5	x	x	
Rao et al.(25,26)	MMPC	12		± 5	x	x	
Isaka et al.(30)	Fuzzy controller	<3	<10	± 5	x		
Ying et al.(31)	Fuzzy controller			± 8	x		x
Chen et al.(33)	ANN	5 to 20		± 15	x		
Kashihara et al.(34)	ANN	2		± 5	x	x	

stably maintain a patient's blood pressure near the set point, and identify excessive noise or artifact in sensor measurements would be beneficial (21,22,31). The supervisor oversees the entire conditions of the control environment and directs the controller to take control actions efficiently and safely. Control decision is based upon sensor measurements. It is very important that the supervisor is able to process measurements and detect the nonphysiological signals, such as the noisy signals duo to suction the airway and flushing the arterial catheter, and thus avoid acting on unreliable information. In addition, the supervisor must have the ability to assure the proper operation of the infusion system for drug delivery. This monitoring system should be able to detect the potential faults that could prevent abnormal operation of the device (e.g., blood clotting, infusion kinking, leakage, and infusion pump stoppage).

CONCLUSION

Because of the quick action of SNP in blood pressure reduction, frequent monitoring of MAP followed by infusion rate adjustment is necessary. The use of manual control to achieve desired MAP would be burdensome to ICU personnel, who are already loaded with many duties. Successful development of a blood pressure controller that could automatically maintain patient's MAP within a preset range with self-monitoring capability would reduce the workload of the patient care providers and improve the patient's quality of life in the clinical environment.

Blood pressure control systems designed previously provide valuable experiences for further development. The future controller should be able to adapt the characteristic changes (represented by gains, time delays, and time constants) from patient to patient as well as the variations within a patient over time. In order to improve the reliability and safety of the controller, incorporating a supervisory scheme that can monitor system operation as well as identify and manage unexpected mechanical errors and clinical environment changes with the control system would be essential.

BIBLIOGRAPHY

Cited References

1. Chitwood Jr WR, Cosgrove 3rd DM, Lust RM. Multicenter trial of automated nitroprusside infusion for postoperative hypertension. Titrator Multicenter Study Group. Ann Thorac Surg 1992;54:517–522.
2. Sheppard LC. Computer control of the infusion of vasoactive drugs. Ann Biomed Eng 1980;8:431–444.
3. Slate JB. Model-based design of a controller for infusing sodium nitroprusside during postsurgical hypertension. Ph.D. dissertation. University of Wisconsin-Madison, 1980.
4. Slate JB, Sheppard LC. Automatic control of blood pressure by drug infusion. Proc Inst Electr Eng 1982;129, (Pt. A):639–645.
5. de Asla RA, Benis AM, Jurado RA, Litwak RS. Management of postcardiotomy hypertension by microcomputer-controlled administration of sodium nitroprusside. J Thrac Cardiovas Surg 1985;89:115–120.
6. Astrom KJ. Theory and application of adaptive control— a survey. Automatica 1983;19:471–486.
7. Astrom KJ, Wittenmark B. Adaptive Control 2nd ed. New York: Addison-Wesley; 1994.
8. Goodwin GC, Sin KS. Adaptive Filtering, Prediction, and Control. Englewood Cliffs (NJ): Prentice Hall; 1984.
9. Ljung L. System Identification: Theory for the User. 2nd ed. Englewood Cliffs (NJ): Prentice Hall; 1998.
10. Arnsparger JM, McInnis BC, Glover Jr JR, Norman NA. Adaptive control of blood pressure. IEEE Trans Biomed Eng 1983;BME-30:168–176.
11. Meline LJ, Westenskow DR, Pace NL, Bodily MN. Computer controlled regulation of sodium nitroprusside infusion. Anesth Analog 1985;64:38–42.
12. Waller JL, Roth JV. Computer-controlled regulation of sodium nitroprusside infusion in human subjects. Anesthesiology 1985;63:A192.
13. Mansour NE, Linkens DA. Pole-assignment self-tuning control of blood pressure in postoperative patients: a simulation study. Proc Inst Electr Eng 1989;136, (Pt. D):1–11.
14. Voss GI, Katona PG, Chizeck HJ. Adaptive multivariable drug delivery: control of arterial pressure and cardiac output in anesthetized dogs. IEEE Trans Biomed Eng 1987;BME-34:617–623.
15. Kaufman H, Roy R, Xu X. Model reference control of drug infusion rate. Automatica 1984;20:205–209.
16. Pajunen GA, Steinmetz M, Shankar R. Model reference adaptive control with constraints for postoperative blood pressure management. IEEE Trans Biomed Eng 1990;BME-37:679–687.
17. Polycarpou MM, Conway JY. Indirect adaptive nonlinear control of drug delivery systems. IEEE Trans Auto Control 1998;AC-43:849–856.
18. Lainiotis DG. Partition: a unifying framework for adaptive systems II: control. Proc IEEE 1976;64:1182–1198.
19. He WG, Kaufman H, Roy R. Multiple model adaptive control procedure for blood pressure control. IEEE Trans Biomed Eng 1986;BME-33:10–19.
20. Martin JF, Schneider AM, Smith NT. Multiple-model adaptive control of blood pressure using sodium nitroprusside. IEEE Trans Biomed Eng 1987;BME-34:603–611.
21. Martin JF, Schneider AM, Quinn ML, Smith NT. Improved safety and efficacy in adaptive control of arterial blood pressure through the use of a supervisor. IEEE Trans Biomed Eng 1992;BME-39:381–388.
22. Martin JF, Smith NT, Quinn ML, Schneider AM. Supervisory adaptive control of arterial blood pressure during cardiac surgery. IEEE Trans Biomed Eng 1992;BME-39:389–393.
23. Yu C, Roy RJ, Kaufman H, Bequette BW. Multiple-model adaptive predictive control of mean arterial pressure and cardiac output. IEEE Trans Biomed Eng 1992;BME-39:765–778.
24. Garcia CE, Prett DM, Morari M. Model predictive control: theory and practices—a survey. Automatica 1989;25:335–348.
25. Rao RR, Palerm CC, Aufderheide B, Bequette BW. Automated regulation of hemodynamic variables. IEEE Eng Med Biol 2001;20 (1):24–38.
26. Rao RR, Aufderheide B, Bequette BW. Experimental studies on multiple-model predictive control for automated regulation of hemodynamic variables. Trans Biomed Eng 2003;50 (3):277–288.
27. Zadeh LA. Fuzzy sets. Inform Contr 1965;8:338–353.
28. Tong RM. A control engineering review of fuzzy systems. Automatica 1977;13:559–569.
29. Sugeno M. An introductory survey of fuzzy control, Inform. Science 1985;36:59–83.
30. Isaka S, Sebald AV. An optimization approach for fuzzy controller design. IEEE Trans Sys, Man, Cyber 1992;SMC 22:1469–1473.

31. Ying H, McEachern M, Eddleman DW, Sheppard LC. Fuzzy control of mean arterial pressure in postsurgical patients with sodium nitroprusside infusion. IEEE Trans Biomed Eng 1992;BME-39:1060–1070.
32. Hunt KJ, Sbarbaro D, Zbikowski R, Gawthrop PJ. Neural networks for control systems—a survey. Automatica 1992;28: 1083–1112.
33. Chen CT, Lin WL, Kuo TS, Wang CY. Adaptive control of arterial blood pressure with a learning controller based on multilayer neural networks. IEEE Trans Biomed Eng 1997;BME-44:601–609.
34. Kashihara K, et al. Adaptive predictive control of arterial blood pressure based on a neural network during acute hypotension. Ann Biomed Eng 2004;32:1368–1383.

See also BIOFEEDBACK; BLOOD PRESSURE MEASUREMENT; DRUG INFUSION SYSTEMS; HEMODYNAMICS; PHYSIOLOGICAL SYSTEMS MODELING.

BLOOD RHEOLOGY

ROGER TRAN-SON-TAY
University of Florida
Gainesville, Florida

WEI SHYY
University of Michigan
Ann Arbor, Michigan

INTRODUCTION

Blood rheology has had broad impact in our understanding of diseases and in the development of medical technology. Rheology is the science dealing with the flow and deformation of matter. Therefore, it encompasses work in mechanical, chemical, and biomedical engineering. It plays a vital role not only in the design, manufacture, and testing of materials, but also in the health of the human body. Biorheology is therefore concerned with the description of the flow and deformation of biological substances. More specifically, hemorheology, or blood rheology, deals with the rheological behavior of blood, including plasma and cellular constituents.

Blood flow is known to be responsible for the delivery of oxygen to tissue and the removal of carbon dioxide. However, it also plays a pivotal role in the transport of substances (nutrients, metabolites, hormones, cells, etc.) involved not only in the maintenance of the body and its immune response, but also in diseases. For example, cancer cells are transported through blood as they spread from one tissue to another in a process known as metastasis.

The rheology of blood is altered in a number of pathological conditions. Sickle cell disease is a genetic disease producing abnormal hemoglobin causing red blood cells (RBCs) to become crescent shaped when they unload oxygen molecules or when the oxygen content of the blood is lower than normal. Under these conditions, the sickle hemoglobin aggregates and the RBCs become rigid, and consequently obstruct and/or damage the capillaries. Sickle cell disease is also known as sickle cell anemia because of the abnormally low oxygen-carrying capacity of the blood due to an insufficient number of RBCs and an abnormal hemoglobin. During a heart attack or stroke, there is a partial or complete occlusion of blood vessels due to the formation of a blood clot that alters blood flow. It is clear that many diseases and factors (artherosclerosis, hypertension, vasodilator agents, etc.) can compromise blood flow by occluding vessels or modifying their rheological properties. However, the study presented here will focus mainly on the rheology of blood.

It is important to recognize that the rheological properties of blood and its components, that is, blood cells, are important in the aptitude of blood to perform its functions correctly. The ability of a blood cell to flow into capillaries or migrate through tissues is governed, but its rheological properties. In addition, flow is expected to affects cells in two ways: (*1*) the fluid moving over or around the cell will exert mechanical stress on the cell, and (*2*) the motion of fluid will alter the concentration of chemical species in the immediate surrounding of the cell, leading to the mass transport of nutrients, waste products, drugs, hormones, and so an, to and from the cell. Finally, blood rheology can also have an indirect but critical role in our immune system and in diseases since a given applied stress, such as fluid shear, can generate a signal that can induce or modify cellular response.

Blood rheology is an extremely broad subject that cannot be covered in a single review. Therefore, the scope of the present article is to provide an understanding of blood rheology, and an appreciation of its contributions to the improvement of our understanding and assessment of diseases. The article also provides a review of the most common methods used to measure the rheological properties of blood and blood cells.

RHEOLOGICAL PROPERTIES OF BLOOD

Is blood a Newtonian fluid? A Newtonian fluid is a fluid that has a viscosity that is constant and independent of the properties of the flow. This simple question is not easily answered because blood is a complex fluid. It is a non-Newtonian fluid that behaves as a Newtonian fluid under certain conditions. For example, for a shear rate $> 100\,\mathrm{s}^{-1}$ and a vessel/tube diameter $> 500\,\mu\mathrm{m}$, blood behaves as a Newtonian fluid. Blood is composed of formed elements (red cells, white cells, platelets, etc.) suspended in plasma. Blood cells are viscoelastic particles (possess both viscous and elastic properties), whereas plasma is Newtonian. Therefore, depending on the characteristics of the flow and size of the vessel (extent of deformation), the size and properties of the blood cells may not play a major role in the flow characteristics of blood. The behavior of blood needs to be described as a function of the size of the vessel and the rate of flow. Because of that behavior, the concept of apparent and relative viscosities is introduced. The viscosity value of a non-Newtonian fluid depends on the experimental conditions and instrument used to perform the measurement. Therefore, that measured viscosity is called the apparent viscosity, μ_{app}. The relative apparent viscosity, μ_{rel}, is defined as

$$\mu_{\mathrm{rel}} = \frac{\mu_{\mathrm{app}}}{\mu_{\mathrm{p}}} \tag{1}$$

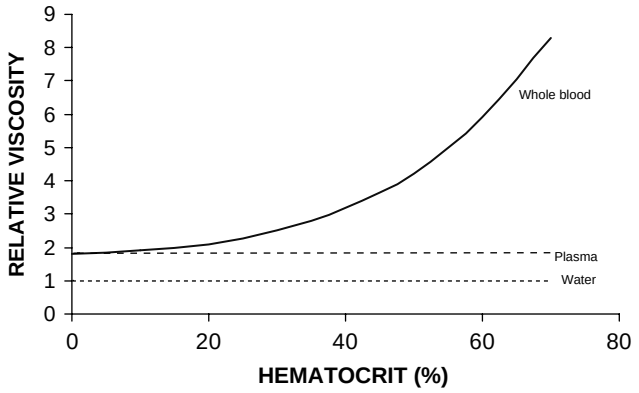


Figure 1. Effect of hematocrit on blood viscosity. Plasma has a relative viscosity of ~ 1.8 at 37 °C, that is a viscosity of ~ 1.2 mPa · s. Human blood at 40% hematocrit has a viscosity of ~3–4 mPa · s. However, it is a function of both hematocrit and shear rate.

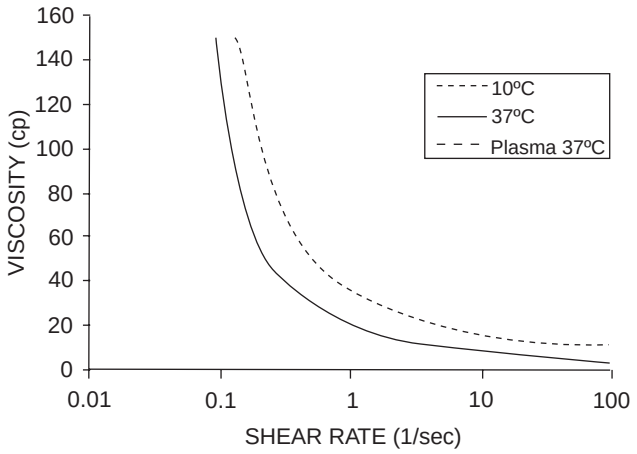


Figure 2. Effect of shear rate on blood viscosity. The Newtonian behavior of plasma and non Newtonian behavior of blood are clearly demonstrated. Plasma has a constant viscosity of ~ 1.2 mPa·s at 37 °C. The effect of temperature on blood viscosity is also shown.

where μ_p is the viscosity of plasma or other suspending medium.

In general, the viscosity of plasma is ~ 1.8 times the viscosity of water (termed relative viscosity) at 37 °C and is related to the protein composition of the plasma. Whole blood has a relative viscosity of ~ 4 depending on hematocrit (RBC concentration), temperature, and flow rate. The average hematocrit for a man and a woman is 42 and 38%, respectively. Hematocrit is an important determinant of the viscosity of blood. As hematocrit increases, there is a disproportionate (exponential) increase in viscosity (Fig. 1). For example, at a hematocrit of 40%, the relative viscosity is 4. At a hematocrit of 60%, the relative viscosity is ~ 8. Therefore, a 50% increase in hematocrit from a normal value increases blood viscosity by $\sim 100\%$. Such changes in hematocrit and blood viscosity occur in patients with polycythemia.

Because blood is non-Newtonian, the effect of shear rate is important. Figure 2 illustrates the shear thinning characteristic (decrease in viscosity as the shear rate increases) of blood at two different temperatures. On the other hand, it is clearly seen that plasma is Newtonian. It has a viscosity of ~ 1.2 cP or 1.2 mPa·s at 37 °C. The poise, P, is a unit of viscosity. The different viscosity units are related as follows: $1\,\mathrm{P} = 1\,\mathrm{dyn{\cdot}s{\cdot}cm^{-2}} = 0.1\,\mathrm{N{\cdot}s{\cdot}m^{-2}} = 0.1\,\mathrm{Pa{\cdot}s}$; therefore, $1\,\mathrm{cP} = 1\,\mathrm{mPa{\cdot}s}$. The fact that blood viscosity increases at low shear is one of the key factors for the initiation of atherosclerosis at specific sites in the arterial system. Increases in the viscosity of blood and plasma reflect clinical manifestations of atherothrombotic (formation of fibrinous clot) vascular disease. High blood viscosity invariably accompanies degenerative diseases. It is therefore not surprising that many treatments involve lowering blood viscosity to treat or prevent heart attacks, strokes, atherosclerosis, and so on.

Temperature also has a significant effect on viscosity. This can be seen in Figs. 3 and 4 where the effect of temperature on the viscosity of plasma and human blood is shown. Temperature has a similar effect on plasma and water. As temperature decreases, viscosity increases. Viscosity increases $\sim 2\%$ for each degree celcius decrease in temperature. This effect has several implications. For example, when whole-body hypothermia is used during certain surgical procedures, it increases blood viscosity and therefore augments resistance to blood flow.

The viscoelastic profile of normal human blood can be divided into three regions depending on the shear rate levels. In the low shear rate region ($\dot{\gamma} \leq 20\,\mathrm{s^{-1}}$), red cells are in large aggregates and as the shear rate increases, the size of the aggregates diminishes. Blood viscoelasticity is dominated by the aggregation properties of the red blood cells. In this region, human blood behaves like a Casson fluid with a small but finite yield stress (i.e., blood will not flow or deform unless the applied stress exceeds that critical stress),

$$\sqrt{\tau} = a + b\sqrt{\dot{\gamma}} \tag{2}$$

where a and b are constant (Fig. 5). The magnitude of the rheological parameters like yield stress and viscosity depends on various factors such as plasma protein concentration,

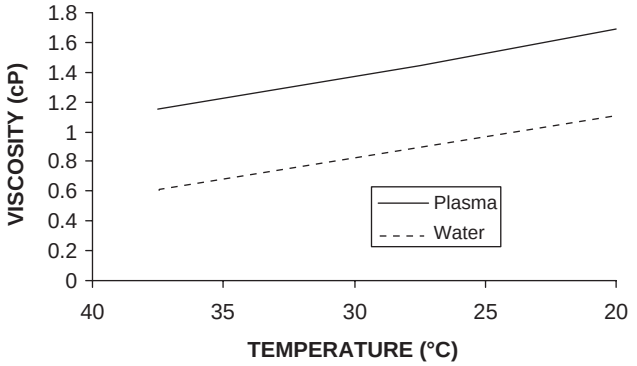


Figure 3. Effect of temperature on plasma viscosity. Temperature has similar effects on the viscosity of plasma and water. Viscosity increases $\sim 2\%$ for each degree celcius decrease in temperature.

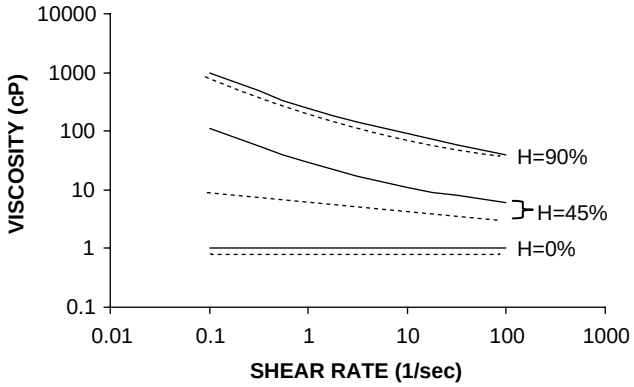


Figure 4. Effects of shear rate and hematocrit on blood viscosity. The shear thinning characteristics of blood is well illustrated in this figure. The solid and dashed lines represent, respectively, whole blood and washed red blood cells in a saline solution at 45 and 90% red cell volume concentrations. (Adapted from Ref. 1.)

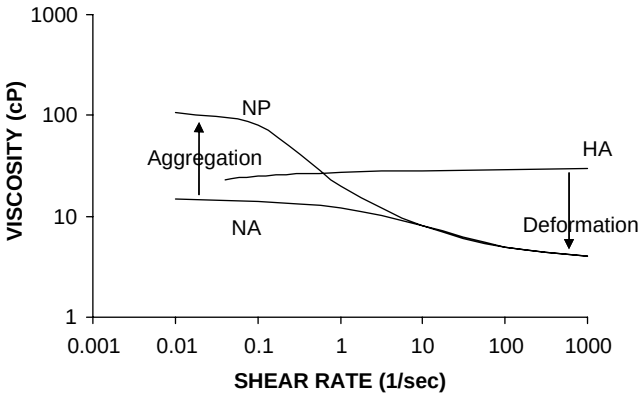


Figure 6. Effects of cell aggregation and deformability on blood viscosity. The logarithmic relation between the apparent viscosity and shear rate in three types of suspensions, each containing 45% human RBCs by volume is shown. Suspending plasma viscosity = 1.2 mPa.s; NP = normal RBCs in plasma; NA = normal RBCs in 11% albumin; HA = hardened RBCs in 11% albumin solution. (Adapted from Ref. 3.)

hematocrit, and properties of the blood cells. At low flow rates, there are increased cell-to-cell and protein-to-cell adhesive interactions that can cause erythrocytes (RBC) to adhere to one another and increase the blood viscosity. However, at shear rates $> 100\,s^{-1}$, cell aggregation and rouleaux formation break up and blood behaves as a Newtonian fluid with a viscosity of $\sim$3–4 mPa·s depending on the hematocrit and other factors (Fig. 6). In the mid-shear rate range ($20 \leq \dot{\gamma} \leq 100\,s^{-1}$), the cells are progressively disaggregated with increasing shear rate. Increasing shear rate causes the cells to deform and orient in the direction of flow, and the viscoelasticity of the blood is dominated by the deformability of the RBC. Figure 6 also demonstrates the effect of cell deformabilty on blood viscosity. It is seen that deformable cells lower the blood viscosity as compared to rigid ones.

However, when the dimensions of the cells are not negligible in comparison with the diameter of the vessel through which flow is occurring, the two phase nature of blood has to be recognized. Thus blood flow through vessels narrower than 500 μm in diameter is accompanied by several anomalous effects which can be directly traced to the two phase nature of blood. The important artifacts are the Fahraeus (a decrease in the hematocrit of the vessel–tube as compared to the larger feeding vessel–reservoir hematocrit) and Fahraeus–Lindqvist (a decrease in the vessel–tube apparent viscosity as compared to the larger feeding vessel–reservoir viscosity) effects. The effects are more pronounced as the vessel–tube diameter decreases. For vessel diameters < 10 μm (capillaries), blood cells must travel in single file and the flow must be analyzed as creeping (low Reynold number) flow of a Newtonian fluid with particles embedded in it. The transition from a single file to suspension flow occurs in the diameter range of 10–25 μm and this domain is difficult to analyze.

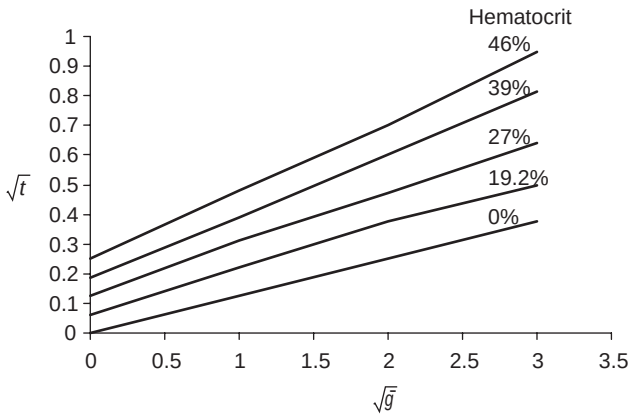


Figure 5. Blood behavior at low shear—casson behavior. These plots were generated from blood data obtained at 25 °C. They show that blood has a yield stress that depends on hematocrit. (Adapted from Ref. 2.)

To add to the complexity of blood behavior, vessels also affect blood flow characteristics. For example, the most noticeable feature of blood flow in the arteries is the pulsatile nature of the flow. However, this feature is lost in the microcirculation because its effect has been dampened by the viscoelastic blood vessels. The flow in the microcirculation occurs at very low Reynolds number (i.e., inertial forces due to transient and convective accelerations are negligible) and is determined by a balance of viscous stress and pressure gradient. Individual cells must be recognized. In the capillaries, 3–10 μm diameter vessels, cells flow in single line and their flow–deformation must be analyzed. Finally, in the veins, where $\sim$80% of the total volume of blood is located, the most noticeable feature is that vessels can collapse and that their mechanical properties cannot be neglected.

The above descriptions represent just some of the rheological characteristics of blood, but many more factors can affect its behavior. However, they prove the point that blood is an extremely complex fluid with many facets and that only a few have been unveiled so far.

CORRELATION BETWEEN BLOOD RHEOLOGICAL PROPERTIES AND CLINICAL CONDITIONS

Blood is a complex fluid whose flow (rheological) properties are significantly affected by the arrangement, orientation, and deformability of red blood cells. Variations in blood rheology among healthy individuals are very small. Thus, changes due to disease or surgical intervention can be readily identified, making blood rheology a useful clinical marker. Variations in blood rheology are observed in such conditions as cardiovascular disease, peripheral vascular disease, sickle cell anemia, diabetes, and stroke.

Although studies of blood rheology date from at least the early studies of Poiseuille (4), the discipline of clinical hemorheology is relatively new. It underwent a rapid growth in 1970–1980, in large part due to support by pharmaceutical companies and equipment manufacturers. Various instruments and devices were developed specifically for studying blood rheology.

Some of the early clinical tests dealing with blood rheology were on blood coagulation and the formation of blood clots. Most people think of blood in its liquid state, but its ability to thicken into a blood clot is a vital part of the body's natural defense. This process of forming a clot is referred to as coagulation. Blood coagulation, or blood clotting, is a complex process involving platelets, coagulation factors present in the blood and blood vessels. If blood becomes too thin, it loses the ability to form the blood clots that stop bleeding. When blood becomes too thick, the risk of blood clots developing within the blood vessels rises creating a potentially life-threatening condition. Blood disorders occur when hemostasis falls out of balance. Hemostasis is achieved when blood chemicals, hormones and proteins are correctly balanced. Hemostasis refers to the complicated chemical interplay that maintains blood fluidity (e.g., viscosity, elasticity, and other rheological properties).

Coagulation, or the lack thereof, is a key factor in various diseases. Sometimes thrombi (large clots) can completely occlude vessels. This can lead to ischemia, and ultimately death in any part of the body. Myocardial infarction and stroke are among the major life-threatening conditions caused by vessel occlusion due to clots. Conversely, there are various coagulatory disorders in which thrombus formation does not occur when it should. These bleeding disorders include various forms of hemophilia [e.g., (5,6)].

A great deal of research has focused on the effects of rheology on thrombus formation. Various *ex vivo* and *in vitro* systems have been designed to mimic *in vivo* blood flow in order to study thrombus formation within the circulatory system (7), and on various devices. For example, these systems have been used to model blood flow in order to study thrombus formation on stents (8) and mechanical heart valve prostheses (9). In addition, some research has focused on the effect of shear on thrombus dissolution. These studies suggest that thrombi lysis is accelerated with increasing shear rates (10,11).

It is impossible to cite all the contributions of blood rheology to our understanding of diseases in this review, but it is clear that viscosity was and still is clinically the most commonly used rheological property. The principal factors determining blood viscosity are hematocrit, plasma viscosity, cell aggregation, and cell deformability. Earlier rheological work was mainly performed on whole blood and on RBCs because the latter are by far the most numerous cells in our body (99% of the blood cells are RBCs). However, in the last two decades, the focus has shifted toward understanding the rheology of leukocytes or white blood cells (WBC) because they have been found to be bigger and more rigid that RBC. The major motivation behind all these blood cell studies is that the ability of a cell to deform and flow through the capillaries and/or to migrate in the tissue is determined by its rheological properties, and this ability is vital in its response to disease/infection. These properties, in turn, are a manifestation of the underlying structure of the cell and the organization of the structural components (microfilaments (F-actin), microtubules, intermediate filaments, lipid bilayer) in the cellular cytoplasm and cortex.

Because blood rheology is a very broad subject, this article focuses on the role of RBC deformability in clinical studies. The role of other blood cell types, cytoskeleton, proteins, adhesion molecules, and so on., although important and of interest, is beyond the scope of this article and will not be addressed.

RBC Deformability

Deformability is a term used to describe the ability of a body (cell in the present context) to change its shape in response to an applied force. A very important characteristics of a normal RBC is that it has a surface area $\sim 30\%$–40% greater than that of a sphere of equal volume. Other major determinants of RBC deformability include rheological properties of the cell membrane, and intracellular fluid.

Cell deformability can be determined by direct microscopic measurement (micropipette) or indirect estimation (filtration). By using micropipettes with diameters $\geq 3\,\mu m$, the entire RBC can be aspirated. The deformability of the cell can be estimated from the pressure required for its total aspiration.

The importance of cell deformability is well established in the studies of the rheological behavior of RBCs in the capillary network. It was clearly demonstrated that reductions in RBC deformability may adversely affect capillary perfusion (12) and that many diseases manifest reductions in RBC deformability (13–15). Of the many determinants of capillary perfusion, the size of the undeformed RBC relative to the capillary diameter may play the greatest role in affecting capillary perfusion. For example, studies of the passage of RBCs through capillary-sized pores of polycarbonate sieves (16) reveal that the flow resistance may increase 30–40 times as the ratio of pore to cell diameter is reduced from 1 to 0.1. Furthermore, after entry into a capillary, the ability of RBCs to deform may play an equally important role as RBCs negotiate irregularities in the capillary lumen, as manifested by encroachment of endothelial cell nuclei on the capillary lumen (17). It was also demonstrated that the microvascular network may passively compensate for increased RBC stiffness by

shunting RBCs within the capillary network through pathways of lesser resistance (18).

There are many methods for assessing the erythrocyte deformability but only two (filtration of RBCs through pores of 3–5 mm diameter and the measurement of RBC elongation using laser diffractometry) have been widely applied clinically. A brief description of these two methods is provided below.

Erythrocyte Filtration. Filtration method has been commonly used to study the deformability of RBC. The basic idea is to force the RBC suspension to flow through 3–5 μm pores (by using a negative or positive pressure or gravity), and obtain the relationship between pressure and flow rate to estimate the deformability of the cells. Either the flow rate is measured under a constant pressure or the pressure is measured under a constant flow rate. Contaminants, such as WBC, which is poorly deformable affect the experiment by plugging the pores.

The techniques for whole-blood filtration are almost all derived from that described by Reid et al. in 1976 (19). The results of these methods are expressed as volume of blood cells (VBCs) in the time unit. However, this technique is susceptible to aggregation of RBCs and contamination with leukocytes. A modified version of the apparatus was developed to reduce these problems (20). Nevertheless, WBC contamination remains an issue with whole-blood filtration techniques.

A common draw back among all these filtrometry-based instruments is the lack of any measure of individual cell volume, thereby making it difficult to distinguish changes in RBC filtration due to the volume distribution (or aggregates) within the RBC sample from those due to intrinsically less deformable cells.

Another filtration technique that is commonly used is the Bowden assay (21,22). However, this assay involves the migration of cells (WBCs) through a filter membrane with pores of defined diameter and is beyond the scope of this article.

Erythrocyte Elongation. The Ektacytometer (23) combines viscosity with laser diffractometry. It consists of a transparent cylindrical Couette or a cone-plate viscometer, which allows a helium–neon laser beam to pass through the erythrocyte test suspension during rotational shear. The laser diffracted image becomes elliptical as the RBCs are sheared, and the ratio of the major over minor axes of the image is called the elongation index. This dynamic measurement of RBC elongation has been used for the rheological studies of congenital defects of RBC membrane protein (24), and many blood disorders (25–27).

It is important to remember that the two methods described above provide information on the bulk deformability of the RBC only, and are not suited for characterizing the deformability of subpopulations. Alternative methods need to be used for these studies. Some of the methods that have been developed for specifically characterizing the rheological properties of individual cells are provided in the next section.

RHEOLOGICAL PROPERTIES MEASUREMENTS

The goal of this section is to provide an overview of the most commonly used techniques for characterizing the rheological properties of blood. Therefore, devices will be divided into two groups: one for characterizing fluids, the other for individual cells.

Techniques for Measuring the Rheological Properties of a Fluid

Cylindrical Tube. The first studies of blood rheology have been done in cylindrical tubes. In his quest toward developing a better method for measuring blood pressure, French physician and physiologist Jean Louis (or sometimes called Leonard) Marie Poiseuille (1799–1869) studied the flow of liquid through tubes. (There is some confusion about Poiseuille's precise name and year of birth, sometimes quoted as 1797.) In 1838, he established a series of meticulously executed experiments: At a given temperature the rate of water flow through tubes of very fine bore is inversely proportional to the length of the tube and directly proportional to the pressure gradient and to the fourth power of the tube diameter. In 1840 and 1846, he formulated and published an equation known as Poiseuille's law (or Hagen-Poiseuille law, named also after the German hydraulic engineer Gotthilf Heinrich Ludwig Hagen who independently carried out friction experiments in low speed pipe flow in 1840) based on his experimental pipe flow observation. Little is known of the life of Jean Leonard Marie Poiseuille. However, he made important contributions to the experimental study of circulatory dynamics. His law can be successfully applied to blood flow in capillaries and veins, and to air flow in lung alveoli, as well as for the flow through hypodermic needle or tubes, in general (28,29).

The derivation of Poiseuille's law for a Newtonian fluid, that is, the viscosity of the fluid is constant and independent of the properties of the flow. (Viscosity is a property of fluid related to the internal friction of adjacent fluid layers sliding past one another, as well as the friction generated between the fluid and the wall of the vessel. This internal friction contributes to the resistance to flow.) For example, water and plasma are Newtonian fluids. For a Newtonian, laminar (nonturbulent) case, the flow through a cylindrical tube is one-dimensional (1D) (Fig. 7) reaching the fully

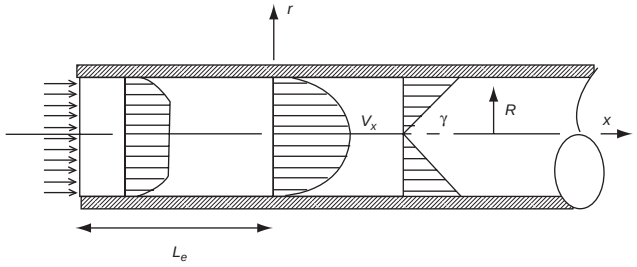


Figure 7. Flow in a cylindrical tube. Fully developed, laminar, viscous flow in tubes produces a parabolic velocity profile. The shear stress varies linearly with the radial distance r. Parameters are as follows: velocity field, $V_x = V_x(r)$, $V_r = 0$, $V_\theta = 0$; shear rate, $\dot{\gamma} = -(\partial V_z/\partial r)$; lines of shear, straight lines parallel to the tube axis; shearing surfaces, concentric cylinders.

developed state, that is, exhibiting no variation in velocity profiles in the streamwise, x direction, and the streamwise velocity, V_x, has the following distribution on any cross-section

$$V_x = \frac{R^2 \Delta P}{4 \mu L}\left[1 - \left(\frac{r}{R}\right)^2\right] \tag{3}$$

where ΔP is the pressure drop between two points located at a distance L apart, R is the tube radius, r is the radial distance from the tube axis, and μ is the fluid viscosity.

For 1D laminar flows in pipe, the pressure gradient, ΔP, necessary to produce a given flow rate, Q, is proportional to the viscosity and, as shown from the above equation, inversely proportional to the fourth power of the tube radius,

$$\frac{\Delta P}{L} = \frac{8 \mu Q}{\pi R^4} \tag{4}$$

This equation has important clinical implications. It tells us that for a given pressure drop, a 10% change in vessel radius will cause $\sim$50% change in blood flow. Conversely, for a fixed flow, a 10% decrease in vessel radius will cause $\sim$50% increase in the required pressure difference. Poiseuille law tells us the consequences of having a reduced vessel lumen like in arteriosclerosis.

Thus, the average fluid velocity can be expressed in terms of the volumetric flow rate Q or pressure ΔP

$$\overline{V} = \frac{Q}{\pi R^2} = \frac{R^2 \Delta P}{8 \mu L} \tag{5}$$

For a Newtonian fluid, the shear stress distribution in the tube is linear with r,

$$\tau = \mu \dot{\gamma} = -\mu \frac{dV_z}{dr} = \frac{r \Delta P}{2L} \tag{6}$$

The shear stress is the frictional force per unit area as one layer of fluid slides past an adjacent layer. Therefore, the maximum shear stress occurs along the tube wall and is equal to

$$\tau_z = \frac{R \Delta P}{2L} = \frac{4 \mu Q}{\pi R^3} = \frac{4 \mu \overline{V}}{R} \tag{7}$$

The viscosity of blood and other fluids have been characterized with cylindrical tube devices. For example, Cannon–Fenske viscometers are cylindrical tubes used to measure the viscosity of fluids. However, they are not commonly used for measuring non–Newtonian fluids because the shear rate generated in these devices is not constant so its effect on viscosity cannot be easily characterized.

Viscometers. Viscometers are designed to measure the viscosity of fluids. They come in many forms (e.g., concentric cylinders, parallel disks) (30), but the review will focus only on instruments that have been used to characterize biological fluids (31,32). The most common is the cone-plate arrangement (Fig. 8) because it produces a linear velocity profile and, consequently, a constant shear rate throughout the gap for small cone angles.

For a cone-plate viscometer of radius R and cone angle α_o, the relevant parameters, such as viscosity and shear rate, can be found by setting the angular speed of rotation

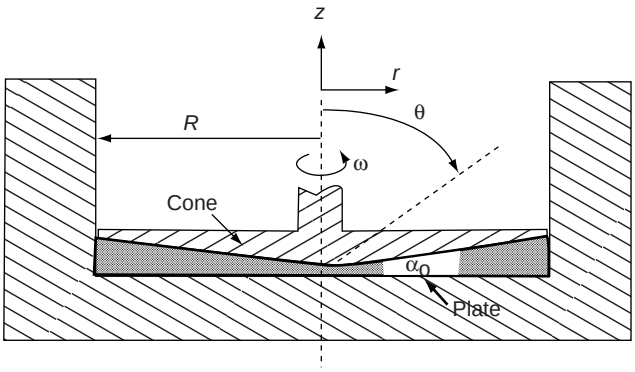


Figure 8. Cone-plate viscometer. For small cone angles, the flow between the cone and the plate is linear. Parameters are as follows: velocity field, $V_\phi = V_\phi(r,\theta)$, $V_\theta = 0$, $V_r = 0$; shear rate, $\dot{\gamma} = -(\frac{1}{r})(\frac{\partial V_\phi}{\partial \theta})$; lines of shear, circles of constant r and z; shearing surfaces, cones of constant θ.

of the cone, ω, and observing the resultant torque, T. These relationships are provided through the expressions of the shear stress, τ, and shear rate, $\dot{\gamma}$.

$$\tau = \frac{3T}{2\pi R^3} = \frac{\mu \omega}{\alpha_o} \tag{8}$$

$$\dot{\gamma} = -\frac{\sin\theta}{r}\frac{d}{d\theta}\left(\frac{V_\phi}{\sin\theta}\right) \cong -\frac{1}{r}\frac{dV_\phi}{d\theta} = \frac{r\omega}{d} = \frac{\omega}{\alpha_o} \tag{9}$$

In Eq. 9, d represents the gap width at a radial distance, r.

In order to measure the elastic properties, dynamic testing needs to be performed. These viscometers have to be run in an oscillatory mode so that elastic effects can be detected. In general, the rheological properties of the fluid can be described in terms of the complex viscosity η^*, or complex modulus G^*. These complex parameters are composed of a viscous and an elastic components, and are related to each other by the angular frequency of the oscillation ω ($G^* = \omega\eta^*$). It is common to mix the notation and use the viscous component, η', of η^*, and the elastic component or storage modulus, G', of G^*, to characterize the viscoelastic properties of the fluid. The viscous and elastic components represent, respectively, energy lost irreversibly and stored reversibly by the sample during an oscillatory cycle.

Microrheometers. As opposed to viscometers, rheometers are devices that measure not only viscosity, but also other rheological properties, like elasticity and yield stress. However, that characterization is very casual since many viscometers have been modified, as described above, to measure the viscoelastic properties of fluids.

As their names indicate, microrheometers have been developed to measure the rheological properties of small volume biological fluids. Typically, these machines require one drop of fluid or less. The design of the magneto-acoustic ball microrheometer for measuring the rheological properties of a liquid is shown in Fig. 2 (33). This instrument requires a much smaller sample size (20 μL, i.e., about a drop) than traditional rheometers, and opaque suspensions

can be studied with it. The small-volume rheometer permits accurate temperature control and rapid temperature changes for kinetics studies, if needed (34).

The instrument itself consists of a magnetically driven 0.8 or 1.3 mm stainless steel ball that is tracked by ultrasonic echo location as it moves within the sample fluid. Using a system consisting of a time-to-voltage Converter (TVC), a pulse generator, a differential amplifier, and an oscilloscope (150 MHz), ball displacements as small as 3 μm are measured. The improved microrheometer (34) is capable of accurately measuring the viscosity of water in the very short chamber. Two measurements can be made (*1*) a falling-ball viscosity and (*2*) an oscillating-ball frequency dependent viscoelastic measurement. Parameters measured are η, the falling ball or steady-state viscosity; η', the viscous or loss modulus; and G', the elastic or storage modulus.

An experiment with the falling ball consists of dropping the ball along the centerline of a 10 mm long tube with a radius of 1.6 mm. The tube is surrounded by a large flow-through chamber for accurate temperature control (Fig. 9). The terminal velocity of the ball, V, is inversely proportional to the viscosity, η. In rheology, it is common to denote the viscosity as η for a viscoelastic fluid. This velocity–viscosity relationship is readily derived from the Stokes drag equation:

$$\eta = 2[(\rho_s - \rho)R^2 g]/9\,VK \tag{10}$$

where ρ_s and ρ are the ball and fluid density, respectively, R is the ball radius, g is the acceleration due to gravity, and K is the wall correction factor to account for the tube wall effect.

Oscillating ball experiments are operated over a frequency range of 1–20 Hz, whereby the sinusoidal driving force and the resulting sinusoidal sphere displacement are recorded. The magnitude of the displacement sinusoid and its phase shift relative to the driving force provide a measure for η' and G'. The system is calibrated with a series of Newtonian silicone oils from 0.1 to 100 P (1 P = 0.1 Pa·s). For a ball oscillating in a viscoelastic medium, the viscous and elastic moduli, when inertia is negligible, are defined as

$$\eta' = \frac{F_o \sin\phi}{6\pi K R \omega X_o} \tag{11}$$

and

$$G' = \frac{F_o \cos\phi}{6\pi K R X_o} \tag{12}$$

where F_o is the magnitude of the oscillating magnetic force, ω is the angular frequency of oscillation ($\omega = 2\pi f$), and X_o is the amplitude of the ball displacement.

The viscoelastic properties of blood have been characterized using the instruments described above in an oscillatory mode. However, these viscoelastic data, although useful as a tool for comparing different blood types and diseases, are not widely used because they are difficult to relate to the mechanical properties of the blood cells. For a non Newtonian fluid, the apparent viscosity (i.e., the slope of the curve of shear stress vs. shear rate at a particular value of shear rate) is used instead of viscosity since the latter is no longer a constant value and will depend on the rate and extent of deformation.

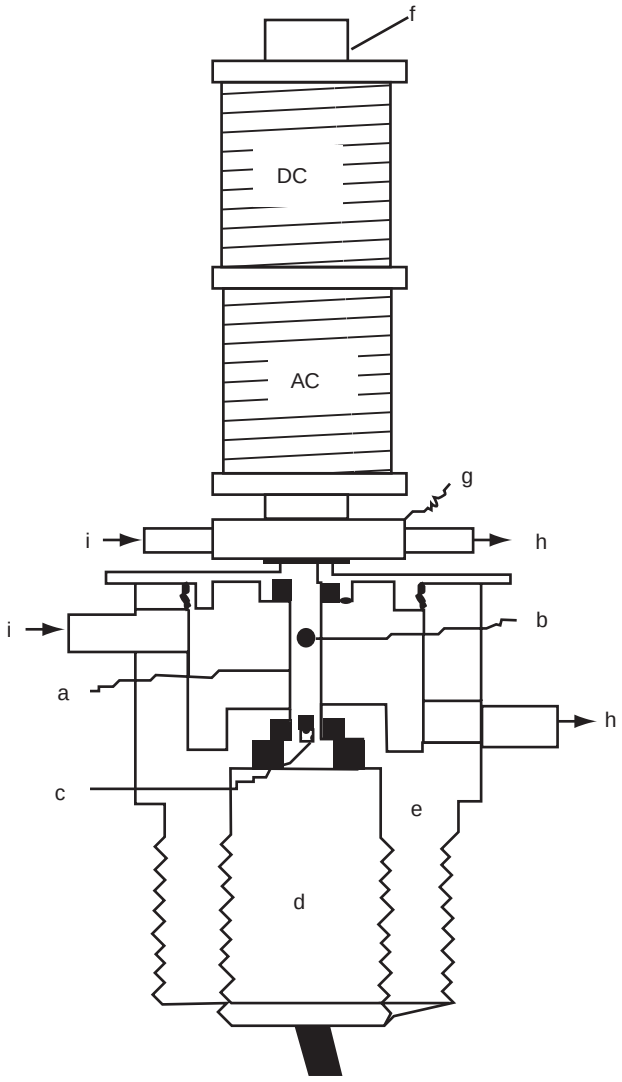


Figure 9. The magneto-acoustic ball microrheometer. (a) sample chamber, (b) stainless steel ball, (c) ultrasound crystal, (d) ultrasound transducer, (e) water jacket, (f) electromagnet, (g) electromagnet bath cap, (h) water flow outlet, (i) water flow inlet.

Techniques for Measuring the Rheological Properties of Blood Cells

Micropipette. The most popular technique for measuring the mechanical properties of blood cells is the micropipette technique (35). The micropipette manipulation technique has been used for studying liquid drops, cells, and aggregates. It has been used to investigate the effects of diseases (36,37) as well as treatments (38,39).

Micropipettes are made from 1mm capillary-glass tubing pulled to a fine point by quick fracture to give an orifice of desired diameter with a square end. The micropipette technique has been extensively used to characterize the mechanical properties of the RBC, but its use is limited in

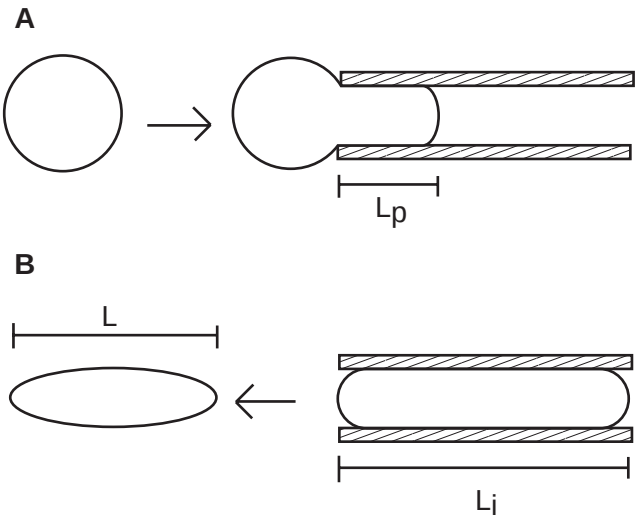


Figure 10. Schematic of the micropipette technique. Two micropipette experiments for determining cell viscosity and surface tension are depicted. (a) Aspiration experiment; for a given aspiration pressure, the length of the aspirated cell, Lp, is tracked as a function of time. (b) Recovery experiment; a cell fully aspirated inside a pipette is expelled from it. The length of the cell, L, is recorded as a function of time.

practice by the complexity of the theory associated with the biconcave shape of the cell. As an example of the micropipette technique, work on the characterization of WBCs will be presented below. The theoretical work is simplified because the shape of a WBC can be treated as a sphere. An equivalent simulation for a RBC will require extensive numerical work. Two typical types of experiment, aspiration and recovery (Fig. 10), are usually performed to determine the mechanical properties of individual WBCs (40–43).

Aspiration. Passive leukocytes (WBCs) are aspirated at a constant pressure into a micropipette. The length of the aspirated cell, L_p, is measured over time to generate an aspiration curve (Fig. 10a). Viscosity values, μ, can be derived from the slope, dL_p/dt, of the aspiration curves (42):

$$\mu = \frac{(\Delta P)R_p}{(dL_p/dt)m(1-\frac{1}{\bar{R}})} \tag{13}$$

where ΔP is the aspiration pressure, R_p is the pipet radius, $\bar{R} = R/R_p$, R is the radius of the cell outside the pipette, and $m = 6$. Figure 11 shows the aspiration of a white blood cell (lymphocyte) into a 4 μm diameter pipette. Fluorescence is used to better see the deformation of the cell nucleus.

Recovery. White blood cells are drawn by a small suction pressure into a micropipette, held there for ~15 s, and quickly expelled out. The changing length of the cell, L, as it recovers its spherical shape (Fig. 10b), is recorded as a function of time, t, and is described by a polynomial (43):

$$\frac{L}{D_o} = \frac{L_i}{D_o} + A\bar{t} + B(\bar{t})^2 + C(\bar{t})^3 \tag{14}$$

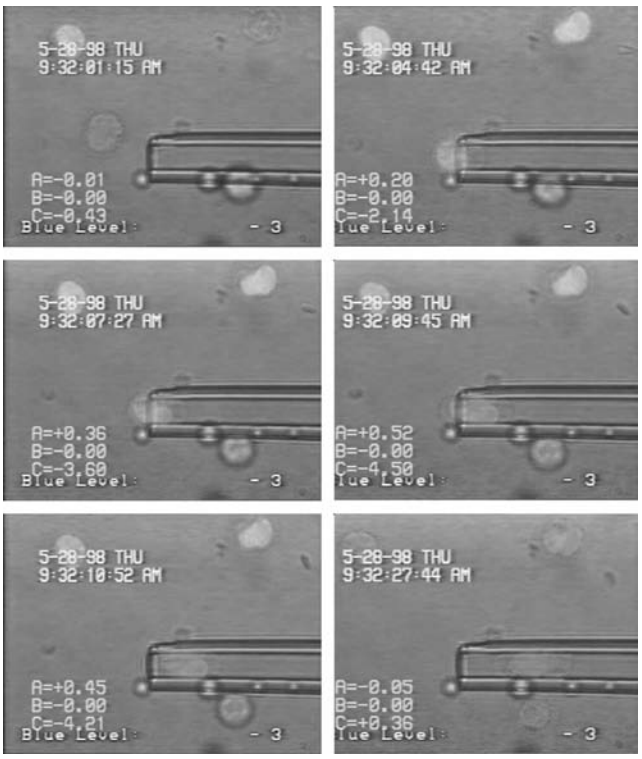


Figure 11. Flow of a lymphocyte inside a 4 μm pipette. The flow of an 8 μm lymphocyte (a WBC) inside a 4 μm diameter micropipette as a function of time is shown. Fluorescent technique was used to track the cell nucleus as well as the cellular membrane.

where A, B, C are known functions of (L_i/D_o). The parameters L_i and D_o are the initial deformed length and resting diameter of the cell, respectively. The variable $\bar{t} = 2t/[(\mu/T_o)D_o]$ represents a dimensionless time, where μ is the cell viscosity, and T_o is the surface tension of the membrane.

Figure 12 shows a lymphocyte (about 8 μm in diameter) aspirated inside a 4 μm diameter pipette (top picture). The cell is then expelled from the pipette and recovers its initial shape (bottom, left-hand side pictures). Pictures generated on the right hand side are from numerical simulation (44). In addition to the experimental techniques, significant progress has been made in the computational capabilities to simulate the dynamic behavior of blood at both large vessel and cellular scales (45).

Rheoscope

To allow direct observation of suspended cells during shear stress application, a modified cone-plate viscometer, called a rheoscope (Fig. 13), has been developed (46). In which the cone and plate counterrotate. This gives the advantage that a particle midway between the cone and plate is subjected to a well-defined shear stress field and remains nearly stationary in the laboratory frame of reference so that it can be studied without the help of high speed cinematography. It is important to note that, for an identical speed of rotation, that the shear rate generated in the rheoscope is twice that in the

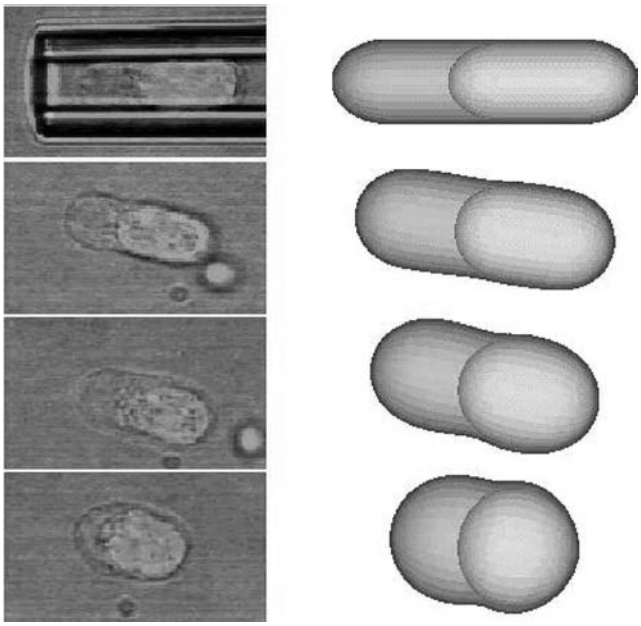


Figure 12. Pictures of a lymphocyte inside a pipette and its recovery. The top left frame is a picture of a lymphocyte (a type of leukocyte) aspirated inside a micropipette, the frames below it show the cell recovering its initial shape. Pictures on the right hand side are those generated by a theoretical compound drop model (44).

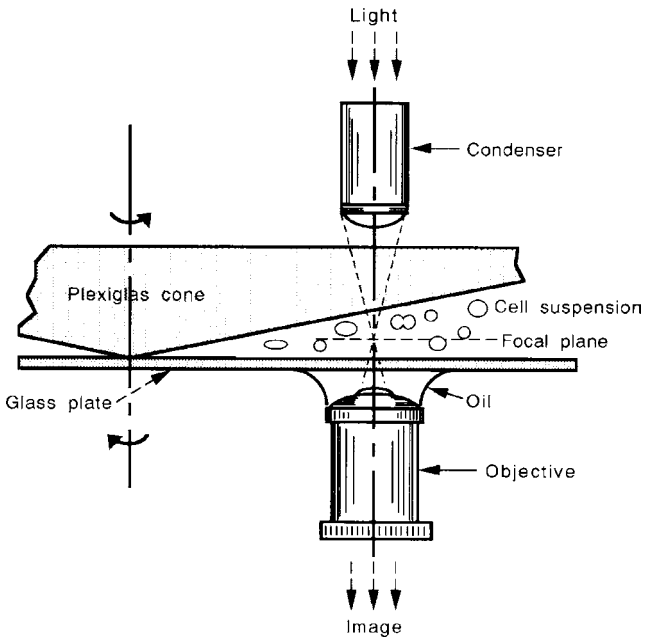


Figure 13. The Rheoscope. The main feature of this instrument is the counterrotation of the cone and plate. This gives the advantage that a particle midway between the cone and plate, subjected to a well-defined shear stress can be studied without the help of high speed cinematography. At a distance r from the axis of rotation, the shear rate is equal to $\dot{\gamma} = 2r\omega/d$, where d is the local gap width and ω is the angular velocity of the cone and plate

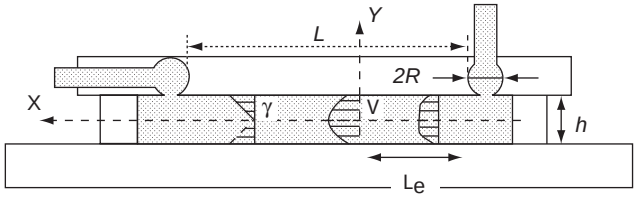


Figure 14. Parallel-Plate Flow Channel. The velocity profile is parabolic, and the shear rate is linear with y. Parameters are as follows: velocity field, $V_x = V_x\ (y)$, $V_y = 0$, $V_z = 0$; shear rate, $\dot{\gamma} = (\partial V_x/\partial y)$; lines of shear, straight lines parallel to the channel axis; shearing surfaces, plane surfaces parallel to the channel axis.

conventional cone-plate viscometer because of the counter rotation of the cone and plate. Effects of shear and mechanical properties of individual cells and anchorage dependent cells have been determined with the rheoscope (e.g., 1,36,47–49). The rheoscope is popular for studying RBC under shear because the analysis of the behavior of RBC is simplified since its shape is an ellipsoid.

Parallel Plate Flow Channel. Another popular technique for characterizing the rheological properties of blood cells under flow or for studying the effects of shear stress on anchorage dependent cells is the parallel plate flow system (Fig. 14).

The flow between infinite parallel plates is often referred to as plane Poiseuille flow. The velocity field reduces to one component in the direction of flow, V_x,

$$V_x = \frac{h^2 \Delta P}{8\mu L}\left[1 - \left(\frac{2y}{h}\right)^2\right] \tag{15}$$

where h is the distance between the plates, μ is the fluid viscosity, ΔP is the pressure drop between the inlet and outlet located at a distance L apart, and y is the vertical distance from the origin taken at the centerline of the channel.

The relationship between the pressure drop and volumetric flow rate Q is

$$\Delta P = \frac{12\mu QL}{wh^3} \tag{16}$$

where w is the channel width.

From the velocity field, Eq. 5, the shear rate across the channel gap is readily derived:

$$\dot{\gamma} = -\frac{dV_x}{dy} = \frac{12Q}{wh^3}y \tag{17}$$

For a Newtonian fluid, the relationship between shear rate and shear stress, τ, is linear, and the shear stress across the flow channel gap is

$$\tau = \mu\dot{\gamma} = \frac{12\mu Q}{wh^3}y \tag{18}$$

From the above equation, the shear stress is zero along the channel centerline, and the maximum shear stress, τ_s, is at the surface of the plate

$$\tau_s = \frac{h\Delta P}{2L} = \frac{6\mu Q}{wh^2} \tag{19}$$

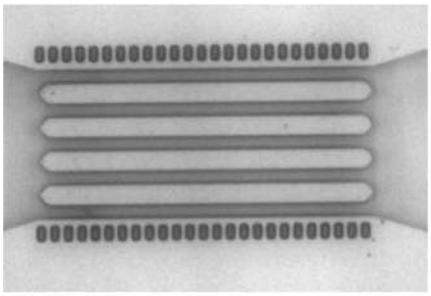
Vote stressed that, although they allow direct observation of individual particles, these devices do not provide a direct measurement of the particle viscosity. It is necessary to know the material constitutive properties of the particles, that is, expressions that relate stresses to strains, or to develop a mathematical model in order to determine the mechanical properties of the particles. Existing models can describe fairly accurately the rheological behavior of blood cells, but the exact rheological property values of these cells, with the exemption of red blood cells, are not known. That topic will not be covered here. For a review on the mechanical properties of blood cells, the reader is referred to Waugh and Hochmuth (35), and for a discussion on the discrepancy between the rheological data on white blood cells reported in the literature to Kan et al. (44).

CONCLUSION

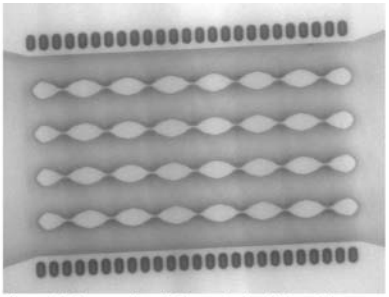
Some of the major advances in blood rheology are now being linked to the development and application of microfabrication to medicine. As reviewed by Voldman et al. (50) and Shyy et al. (51), these new tools will be used to better characterize the rheological properties of blood and blood cells, as well as to detect and diagnose cardiovascular and blood related diseases. Microfabrication is a process used to construct objects with dimensions in the micrometer to millimeter range. These objects are composed of miniature structures that can include moving parts such as cantilevers.

Soft lithography is a tool for micro/nanofabrication. It provides a convenient and effective method for the formation and manufacturing of micro- and nanostructures. Soft lithography is the collective name for a set of techniques that include replica molding, microcontact printing, microtransfer molding (52). The major advantages of soft lithography are that it is very fast as compared to conventional methods, relatively inexpensive, and applicable to almost all polymers. It is possible to go from design to production of replicated structures in < 24 h. In soft lithography, a master mold is first made by a lithographic technique, and an elastomeric stamp is then cast using the master mold. The elastomeric stamp with patterned relief structures on its surface is used to generate patterns and structures with feature sizes as small as 30 nm (53). Polydimethylsiloxane (PDMS) is the polymer of choice for many biological applications because it is optically transparent, isotropic, homogeneous, durable, and has interfacial properties that are easy to modify (53,54). As opposed to photolithography, soft lithography provides a mean for producing nonplanar surfaces.

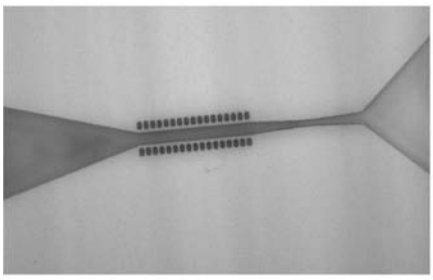
Microfabricated devices, also known as microelectromechanical systems (MEMS), are ideal tools for studying specific biological phenomena, for example, cell adhesion, as well as biological systems such as the microcirculation. For example, the fabrication of in vitro blood vessels can help to (*1*) determine blood cell distributions during blood flows through both arterial and venous type bifurcations, with successive bifurcations arranged as in microcirculation; (*2*) validate computer simulations and experimental methods used for *in vivo* measurements of parameters such as blood average velocity and hematocrits; and (*3*) to separate vessel or wall effects from hemodynamics effects. Figure 15 shows channels that were created using soft lithography. The different configurations shown are a series of 10 mm diameter channel in parallel, a series of



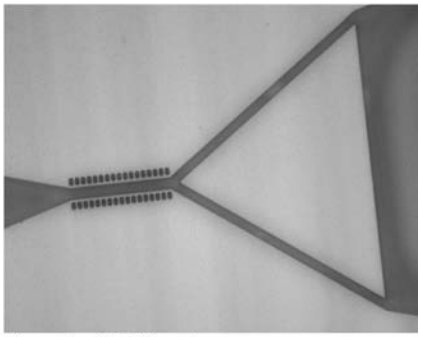
Series of 5 channels (150 μm long × 10 μm width) in parallel with smooth surface



Series of channels with endothelial cell pattern surface



Channel with constriction, smooth surface



Channel with bifurcation
Smooth surface

Figure 15. Photographs of microfabricated channels. The dashed lines on the top and bottom of the channels are length scales and each line represents 100 μm. Channels were created with inner diameters ranging from 5 to 100 μm.

channels with an endothelial cell like pattern surface, a channel with a constriction, and a channel with a bifurcation.

The application of MEMS in the area of microfluidics is a new and emerging field. It includes the development of miniature devices in fluid control, fluid measurement, and medical testing. Because of our need and interest in understanding, preventing, and treating diseases, most of the emerging MEMS applications are expected to be in biomicrofluidics. One of these applications is the development of gene chips and related deoxyribonucleic acid (DNA) tools. Others applications include microscale chemical analysis, known as microelectrophoresis, blood chemistry measurements using micromachined thin-film sensor, micropumps, drug delivery systems, glucose sensors, and chip-based microflow cytometry devices.

The major advantages of MEMS-based devices are that they are smaller and have the potential to be less expensive, more durable, and more reliable than conventional techniques. In addition, they can perform chemical tests much faster.

Finally, another area that will benefit from the advancement of technology is the development of blood substitutes. This is a needed area since the demand for blood continues to outpace the supply, especially in developing countries. With new technologies, the life span of blood substitutes based on cell-free hemoglobin, which is presently too short, could be lengthened. New methods could also be developed in order to produce human red cells, and other cell types, in culture.

Needless to say that, in order to be successful, all these new technical advances will need to be combined with an improved understanding of cell biology and rheology. This advancement will also depend on the successful incorporation of more sophisticated mathematical and computational techniques. In general, the field of biorheology is moving towards better understanding phenomena at the molecular level. For example, the knowledge of the signal transduction pathways associated with the responses of cells to deformation is essential in the field of tissue engineering, where mechanical environment during growth can affect cell response and the material properties of the tissue construct or living implant.

It is clear that blood rheology plays a major role in the maintenance of the human body and in the development of artificial organs and other medical devices, but our understanding, of that role and of the clinical implications of having an altered blood rheology, is far from being complete.

BIBLIOGRAPHY

Cited References

1. Chien S. Red Cell Deformability and its Relevance to Blood Flow. Ann Rev Physiol 1987;49:177–192.
2. Cokelet GR, et al. The rheology of human human blood measurement near and at zero shear rate. Trans Soc Rheol 1963;7:303–317.
3. Chien S, Usami S, Dellenbeck RJ, Gregersen M. Shear dependent deformation of erythrocytes in rheology of human blood. Am J Physiol 1970; 219:136–142.
4. Sutera SP. The history of Poiseuille's law. Ann Rev Fluid Mech 1993;25:1–19.
5. White GC, Montgomery RR. Clinical aspects of and therapy for von Willebrand disease. In: Hoffman R, et al., editors. Hematology: Basic principles and practice. 3rd ed. New York: Churchill Livingstone Inc.; 2000. 1946–1958.
6. Curry H. Bleeding Disorder Basics. Ped Nursing 2004;30(5): 402–404 and 428–429.
7. Hafezi-Moghadam A, Thomas KL, Cornelssen C. A novel mouse-driven *ex vivo* flow chamber for the study of leukocyte and platelet function. Am J Phys Cell Physiol 2004;286: C876–C892.
8. Sakakibara M, et al. Applicaion of *ex vivo* flow chamber system for assessment of stent thrombosis. Arterioscler Thromb Vasc Biol 2002;22(8):1360–1364.
9. Zimmer R, Steegers A, Paul R, Affeld K, Reul H. Velocities, shear stresses and blood damage potential of the leakage jets of the Medtronic Parallel bileaflet valve. Int J Artif Organs 2000;23(1):41–48.
10. Komorowicz E, Kolev K, Lerant I, Machovich R. Flow rate-modulated dissolution of fibrin with clot-embedded and circulating proteases. Circ Res 1998;82:1102–1108.
11. Blinc A, et al. Flow through clots determine rate and pattern of fibrinolysis. Thromb Haemost 1994;71:230–235.
12. Driessen GK, et al. Effect of reduced red cell "deformability" on flow velocity in capillaries of rat mesentery. Pflügers Arch 1980;388:75–78.
13. Schmalzer EM, Manning RS, Chien S. Filtration of sickle cells: recruitment into a rigid fraction as a function of density and oxygen tension. J Lab Clin Med 1989. 113:727–734.
14. Schmid-Schonbein H, Volger E. Red-cell aggregation and red-cell deformability in diabetes. Diabetes 1976;25(Suppl. 2): 897–902.
15. Schrier SL, Rachmilewitz E, Mohandas N. Cellular and membrane properties of alpha and beta thalassemic erythrocytes are different: implications for differences in clinical manifestations. Blood 1989;74:2194–2202.
16. Reinhart WH, Chien S. Roles of cell geometry and cellular viscosity in red cell passage through narrow pores. Am J Physiol 1985;248(Cell Physiol. 17):C473–C479.
17. Secomb TW, Hsu R. Motion of red blood cells in capillaries with variable cross-sections. J Biomech Eng 1996;118:538–544.
18. Lipowsky HH, Cram LE, Justice W, Eppihimer M. Effect of erythrocyte deformability on *in vivo* red cell transit time and hematocrit and their correlation with *in vitro* filterability. Microvasc Res 1993;46:43–64.
19. Reid HL, Barnes AJ, Lock PJ, Dormandy JA, Dormandy TL. A simple method for measuring erythrocyte deformability. J Clin Pathol 1976;29(9):855–858.
20. Dodds AJ. et al. Haemorheological response to plasma exchange in Raynaud's syndrome. Br Med J 1979. 1186–1187,
21. Boyden S. The chemotactic effect of mixtures of antibody and antigen on polymorphonuclear leucocytes. J Exp Med 1962; 115:453–466.
22. Zigmond SH, Hirsch JG. Leukocyte locomotion and chemotaxis. New methods for evaluation, and demonstration of a cell-derived chemotactic factor. J Exp Med 1973;137:387–410.
23. Bessis M, Mohandas N, A diffractometric method for the measurement of cellular deformability. Blood Cells 1975;1: 307–313.
24. Bull B, Feo C, Bessis M, Behavior of elliptocytes under shear stress in the rheoscope and the ektacytometer. Cytometry 1983;3:300–304.
25. Bareford D, et al. Erythrocyte deformability in peripheral occlusive arterial disease. J Clin Pathol 1985;38:135–139.
26. Erythrocyte deformability in peripheral occlusive arterial disease. J Clin Pathol 1985;38:135–139.
27. Yip R, et al. Red cell membrane stiffness in iron deficiency. Blood 1983;62:99–106.

28. Fung YC. Biomechanics—Circulation 2nd ed. New York: Springer-Verlag; 1997.
29. Pedley TJ. Pulmonary Fluid Dynamics. Ann Rev Fluid Mech 1977;9:229–274.
30. Ferry JD. Viscoelastic Properties of Polymers. 3rd ed. New York: John Wiley; 1980.
31. Whitemore RL. Rheology of the Circulation. New York: Pergamon Press; 1968.
32. Tran-Son-Tay R. Techniques for Studying the Effects of Physical Forces on Mammalian Cells and for Measuring Cell Mechanical Properties. In: Frangos JA, editor. Physical Forces and the Mammalian Cell. New York: Academic Press; 1993;p. 1–59.
33. Tran-Son-Tay R, Beaty BB, Acker DN, Hochmuth RM. Magnetically Driven, Acoustically Tracked Translating Ball Rheometer for Small, Opaque Samples. Rev Sci Instru 1988; 59:1399–1404.
34. Tran-Son-Tay R. A Microrheometer for Studying the Rheological Properties of Sickle Cell Suspensions and Hemoglobin. Proceedings of The Fourth China-Japan-USA-Singapore Conference on Biomechanics. In: Yang G, Hayashi K, Woo SL-Y, Goh JCH, editors. International Academic Publishers; 1995; p. 429–432.
35. Waugh RE, Hochmuth RM. Mechanics and Deformability of Hematocytes. In: Schneck DJ, Bronzino JD, editors. Biomechanics—Principles and Applications. New York: CRC Press; 2202; 227–239.
36. Linderkamp O, Ruef P, Zilow EP, Hoffmann GF. Impaired deformability of erythrocytes and neutrophils in children with newly diagnosed insulin-dependent diabetes mellitus. Diabetologia 1999;42:865–869.
37. Perrault CM, et al. Altered Rheology of Lymphocytes in the Diabetic Mouse. Diabetologia 2004;47:1722–1726.
38. Tsai MA, Frank RS, Waugh RE. Paasive Mechanical Behavior of Human Neutrophils: Effect of Cytochalasin B. Biophys J 1994;66:2166.
39. Thomas SJ, et al. Effects of X-Ray Radiation on the Rheological Properties of Platelets and Lymphocytes. Transfusion 2003;43:502–508.
40. Evans EA,, Yeung A. Apparent viscosity and Cortical Tension of Blood Granulocytes Determined by Micropipet Aspiration. Biophys J 1989;56:151.
41. Hochmuth RM, et al. Viscosity of Passive Neutrophils Undergoing Small Deformations. Biophy J 1993;64:1596–1601.
42. Needham D, Hochmuth RM. Rapid Flow of Passive Neutrophils into a 4 mm Pipet and Measurement of Cytoplasmic Viscosity. J Biomech Eng 1990;112:269.
43. Tran-Son-Tay R, Needham D, Hochmuth RM. Recovery of Passive Neutrophils after Large Deformation: Liquid Drop Model. Proceedings of the ASME, Adv Bioeng 1991;20:421–424.
44. Kan H-C, et al. Effects of Nucleus on Leukocyte Recovery. Ann Biomed Eng 1999;27(5):648–655.
45. Shyy W, et al. Moving Boundaries in Micro-Scale Biofluid Dynamics. Appl Mecha Rev 2001;54:405–453.
46. Schmid-Schoenbein H, et al. A Counter-Rotating "Rheoscope Chamber for the Study of the Microrheology of Blood Cell Aggregation by Microscopic Observation and Microphotometry. Microvasc Res 1973;6:366–376.
47. Tran-Son-Tay R, Sutera SP, Rao PR. Determination of RBC Membrane Viscosity from Rheoscopic Observations of Tank-Treading Motion. Biophy J 1984;46:65–72.
48. Tran-Son-Tay R, Sutera SP, Zahalak GI, Rao PR. Membrane Stress and Internal Pressure in Red Blood Cells Freely Suspended in Shear Flow. Biophys J 1987;51:915–924.
49. Glover S, et al. Phosphorylation of Tyrosine 397 Critically Mediates Gastrin-Releasing Peptide's Morphogenic Properties. J Cellular Physiol 2004;199:77–88.
50. Voldman J, Gray ML, Schmidt MA. Microfabrication in biology and medicine. Annu Rev Biomed Eng 1999;1:401–425.
51. Shyy W, Tran-Son-Tay R, N'Dri N. Micro-Nano Coupling in Biological Systems. In: Harik VM, Luo LS, Salas M, editors. Nano-Scale Mechanics of Solid and Liquid Materials Systems. The Netherlands: Kluwer Academic; 2003.
52. Maddou M. Fundamentals of Microfabrication: The Science of Miniaturization. 2nd ed. Washington (DC): CRC Press; 2001.
53. Xia Y, Whitesides GM. Soft lithography. Ann Rev Mater Sci 1998;28:153–184.
54. Branham ML, et al. Rapid Prototyping of Micropatterned Substrates Using Conventional Laser Printers. J Materials Res 2002;17(7):1559–1562.

Reading List

Adjizian JC, et al. Clinical applications to the Ektacytometer. Clin Hemorheol 1984;4:245–254.
Chien S, et al. Effects of hematocrit and plasma proteins of human blood rheology at low shear rates. J Appl Physiol 1966;21:81–87.

See also CELL COUNTERS, BLOOD; HEMODYNAMICS.

BLOOD, ARTIFICIAL

BRIAN WOODCOCK
University of Michigan
Ann Arbor, Michigan

INTRODUCTION

Blood has such a multitude of physiological functions; it provides a circulating volume to transport substrate and metabolites, and it transports the most valuable of substrates, oxygen. It is an organ intimately involved in the immune system, delivering antibodies and cellular elements to sites of infection. It carries the instruments for coagulation. It is the communication highway for the endocrine system. It is a metabolic organ containing enzyme systems to convert molecules to active and inactive forms. Blood is intimately involved in temperature regulation. The manufacture of an artificial substitute to fulfill all those purposes is beyond the capability of current science. However, several of the functional capabilities of blood have been incorporated into various blood substitutes.

The most basic function of blood is to provide a circulating volume for transportation of substrate and metabolites. Supplementation of intravascular volume with crystalloid and colloid fluids has been a part of medical practice for a century. Recent progress has concentrated on the development of substitute solutions that can transport oxygen. These solutions are known as oxygen therapeutics or red cell substitutes.

Currently, the only available oxygen therapeutic is typed and cross-matched allogeneic human blood. This is made available in the civilian setting by the Red Cross, the American Blood Centers, and the blood banking system. Blood shortages, due to increasing blood usage and declining blood donations (1), are one of the factors driving the

search for alternatives. In the past two decades there has also been an increasing concern regarding the infectious risks of blood borne pathogens (2,3). Awareness of the potential significance of the problem occurred with the increasing risk of human immunodeficiency virus (HIV) transmission from blood transfusion during the 1980s. Improved screening for HIV in the 1990s saw a dramatic improvement in blood safety so that now the risk of contracting HIV from a unit of blood is approaching 1/1,000,000. But there are still substantial concerns regarding not only the risk of contracting acquired immune deficiency syndrome (AIDS) or hepatitis, and also of other infectious diseases newly recognized as possibly being transmissible by transfusion, such as bovine spongiform encephalopathy (mad cow disease) and West Nile virus.

Another concern leading to the need for blood substitutes is that some groups Jehovah's Witnesses, have religious beliefs that cause them to refuse all blood products.

Banked blood is often wasted while active bleeding continues in the surgical setting. While hemorrhage continues, blood products administered are rapidly lost through the site of bleeding. Blood substitutes could be used as a resuscitation bridge until bleeding is controlled (4).

Initial research on artificial blood was lead by the U.S. military, which needed a ready supply of a substitute that could be stored easily and indefinitely in the field and not require typing or cross-matching. These considerations would also make a blood substitute valuable to the emergency medical services in ambulance and helicopter transfers.

All these concerns have stimulated efforts to develop red cell substitutes for use in the routine clinical setting. In an initial approach, prior to World War II, the defense department sought a hemoglobin solution that could be stored indefinitely at room temperature, preferably in a powdered form to be dissolved in normal saline, and that could be transfused without a need for cross-matching. Although the development of a reconstitutable powder has not been feasible, there are several hemoglobin-based products that are in various stages of clinical testing (5).

Other molecules apart from hemoglobin have been assessed for the function of oxygen transportation. Most success has been achieved with emulsions of perfluorochemicals.

Perfluorochemicals are inert liquids, which have a solubility for oxygen and carbon dioxide 20 times that of water. These liquids are immiscible in water and an emulsion form is required to allow them to mix with the recipient's blood after administration. A comparison of the advantages and disadvantages of perfluorocarbon and hemoglobin solutions is shown in Table 1.

Emulsions of perfluorochemicals have completed animal testing and have been investigated in the clinical setting. However, difficulties in their use have been observed, to date they have not been made available for routine use.

Table 1. Comparison of Perfluorocarbon and Hemoglobin Solutions

	Advantages	Disadvantages
Perfluorocarbon Emulsions	High O_2 Solubility Inert Ample supply	Requires high PO_2 Long tissue life Short vascular life Toxicities
Hemoglobin Solutions	Carry O_2 at normal P_aO_2 Unloads like RBCs May be stored dry?	Vasoconstriction Supply Short vascular life Toxicities

CURRENT RISKS OF BANKED BLOOD

Risks of Transfusion

Blood transfusion is safer today than it has ever been, with a death rate for each blood transfusion of ~ 1 in 300,000 (6). Two-thirds of these deaths are due to clerical errors (i.e., the wrong blood given to the wrong patient). Other risks associated with blood transfusion include infection and immune reactions (7,8). However, 20 million blood transfusions are administered each year in the United States, with an impressive safety record (9).

Infection

The risk of transmission of infection includes viral agents, other exotic infectious agents, and the risk of bacterial contamination of blood products. Blood donors with a history of risk factors are excluded from donation. Screening donated units and elimination of units that contain known infectious agents eliminates most of the remaining risk of infection.

HIV

The risk of HIV transmission from blood transfusion has caused the most public concern, though it is difficult to accurately assess the true risk of transmission because it is small and cases can be determined only after a significant period. Initial testing for HIV consisted of antibody testing alone, but this was felt to leave a risk of HIV from individuals who were infected, but had not yet sero-converted. In March 1996, HIV antigen p24 testing was instituted in the United States and only 3 out of 18 million units were identified as being antibody negative and antigen positive over the next 18 months (10). Blood donations are now tested for HIV-1 and HIV-2 (11). The apparent risk of HIV transmission is currently ~ 1 in 1,000,000.

Hepatitis

Hepatitis has a higher prevalence; 1:60,000 for hepatitis B, and 1:103,000 for hepatitis C, but is much less feared by the general public. Antigen screening tests have reduced the risk of post-transfusion hepatitis (B or C) to < 1 in 34,000 (12). Hepatitis G has more recently been recognized, and has a high incidence worldwide of 1 (13)–7% (14). Approximately 2% of blood donors and 15–20% of intravenous drug abusers in the United States have detectable hepatitis G (15). It may be identified by the polymerase chain reaction test, but this has not been implemented as a routine screening test. Fortunately, it appears that the hepatitis G virus is not responsible for non-A, non-B, non-C post-

transfusion hepatitis and the results of infection appear to be minimal (16,17), although there is a weak link between hepatitis G and fulminant hepatitis in rare cases (18).

Other Viruses

Creutzfeldt–Jakob disease (vCJD) can be transmitted by transfer of central nervous system tissue (or extract). However, no cases have been definitively linked to blood transfusion (19).

In the United Kingdom an outbreak of bovine spongiform encephalopathy (BSE) or "Mad Cow Disease" has led to concern that a new variant vCJD could be transferred to the human population through consumption of contaminated beef products. Transmission of BSE by blood transfusion can occur in sheep (20). In the United Kingdom, blood products for transfusion are leucodepleted, which is thought to reduce the risk of transmission of vCJD. The possibility that infection might occur has led the U.S. Food and Drug Administration (FDA) to institute a policy "deferring", that is, declining, blood donations from anyone who has lived in the United Kingdom for a cumulative period of more than 6 months during the years 1980–1996 (21).

West Nile virus transmission has occurred in four patients who received solid organ donations from an infected donor. The organ donor had received blood transfusions from 63 donors, and follow up of those donors showed that one of them was viremic at the time of donation (22).

Bacterial contamination of stored blood is rare (1:500,000), but has a mortality rate of 25–80%. The most common infectious contaminants in red cells are gram-negative species such as *Pseudomonas* or *Yersinia* (23). Platelets are stored at room temperature, allowing rapid bacterial proliferation, and there is a risk of 1:3000–7000 of bacterial infection with these units. Bacterial contamination of platelets is typically with Gram-positive staphylococci.

Immune Reactions

Minor immune reactions, such as febrile reactions, are common and may be discomforting to the patient, but are not associated with significant morbidity, though the transfusion may have to be stopped and the product discarded. The risk of serious immune reactions is small, but present. Most acute hemolytic reactions are due to clerical errors, because cross-matching should predict and prevent these events. However, immune reactions remain the most common cause of fatality associated with transfusions.

Graft versus host disease may occur rarely after transfusion, most commonly after transfusion of nonirradiated blood components to patients with immunodeficiency. Transfusion-associated graft versus host disease has a high mortality and is rapidly fatal. Immunodeficienct patients should receive irradiated units. Immunocompetent individuals may develop graft versus host disease if common histocompatibility leukocyte antigen haplotypes between the donor and recipient prevent destruction of stem cells transfused. This can occur between first-degree family members and therefore, relative-to-patient-directed donations, which are often preferred by patients because of a perceived reduction in risk of infection, may in fact carry an increased risk of initiating transfusion-associated graft versus host disease.

Transfusion-related immunomodulation has been recognized since the mid-1970s, but is not well quantified. Exposure to allogeneic blood can cause both allosensitization and immunosuppression. Studies have demonstrated a beneficial effect of allogeneic blood transfusion on transplant organ survival, but increases in the rates of cancer recurrence and postoperative infection have also been noted (23,24). Leukocyte depletion and removal of plasma may ameliorate the effects of TNF- suppression and interleukin induction (25).

The risk of infection and concerns regarding immune reactions and immunomodulation have been a large incentive to the development of oxygen-carrying colloids.

PERFLUOROCHEMICAL EMULSIONS

Perfluorochemicals (PFCs) are chemically inert liquids with a high solubility for gases. The PFCs that have been used as blood substitutes are 8–10-carbon atom structures that are completely fluorinated. The PFCs are chemically inert, clear, odorless liquids with a density nearly twice that of water. The solubility of PFCs for oxygen is nearly 20 times that of water.

In 1965, Clark and Gollan (26) performed an experiment to see whether an animal could survive if it breathed liquid PFC equilibrated with 1 atm of oxygen. A rat could be submerged beneath this liquid for 30 min and be retrieved in good condition. Respiration of liquid PFC allowed oxygen absorption from the lungs together with CO_2 excretion.

An intravenous injection of PFC is immediately lethal because the injectate is immiscible with water and forms a liquid embolus. An emulsion of an immiscible liquid can, however, mix with water or blood. In 1968, Gehes (27) produced a microemulsion (particle size, 0.1 μm) of a PFC in normal saline. An exchange transfusion could be done, eliminating all normal blood elements, and a rat with a hemoglobin of 0 could survive breathing 100% oxygen.

Because of the inert nature of these compounds, they are not metabolized, but are cleared from the vascular space by the reticulo-endothelial system (RES), and ultimately collected in the liver and spleen. Eventually, the PFC slowly leaves the body as vapor in the respiratory gas.

Perfluorochemical Oxygen Content

Because PFCs transport oxygen by simple solubility, the amount of oxygen they carry is directly proportional to the percentage of PFC in the bloodstream and to the P_aO_2.

Hemoglobin carries most of the oxygen in whole blood and does so in a nonlinear fashion. The plot of oxygen content against P_aO_2 for hemoglobin, known as the oxygen dissociation curve, is seen in Fig. 1. Perfluorochemicals (PFCs) carry oxygen by direct solubility, as does plasma. Because of the shape of the oxygen dissociation curve, hemoglobin is fully saturated and carries little or no more oxygen above a PO_2 of 90 mmHg (11.99 kPa). Because oxygen content dissolved in plasma, or carried by PFCs,

is linearly related to PO_2, additional oxygen is dissolved in the plasma phase or by PFC as oxygen tension increases.

Arterial content of oxygen (C_aO_2) is defined as the volume of oxygen in milliliters (mL) carried by each 100 mL of blood and is defined as follows:

$$C_aO_2 = (Hb \times 1.34 \times SaO_2) + (0.003 \times P_aO_2) \approx 20\,mL/100\,mL \quad (1)$$

(Hb = hemoglobin: S_aO_2, arterial oxygen saturation: P_aO_2, arterial oxygen tension)

With a normal 15 g of hemoglobin and normal P_aO_2 and S_aO_2 values of 90 mmHg (11.99 kpa) and 97%, respectively, an arterial oxygen content of 20 $mL \cdot dL^{-1}$ is obtained.

When blood contains an oxygen carrying PFC, the oxygen content equation requires a third term to represent the contribution from perfluorocarbon.

$$C_aO_2 = (Hb \times 1.34 \times S_aO_2) + (0.003 \times P_aO_2) + (0.057 \times Fct/100 \times P_aO_2) \quad (2)$$

where Fct = fluorocrit, which is the fraction of the blood volume that is PFC (analogous to the Hct).

Note that the solubility factor of PFC in the third term should be 0.06, that is, 20 times that of the solubility factor for oxygen in plasma (0.003), however, it is reduced by the amount of the plasma solubility factor to account for the plasma displaced by the presence of PFC.

The PFCs carry much more oxygen than plasma, but hemoglobin itself is able to carry much more than any PFC. Figure 1 shows that blood with a Hct of 45% will have a C_aO_2 of 20 mL/100 mL at a PO_2 of 100 mmHg (13.33 kPa), but a solution with a Fct of 45% would have an oxygen content of 2.7 mL/100 mL. Because the PFC carries oxygen by direct solubility, it also releases it in direct proportion to the PO_2, unlike the cooperative binding effect of Hg, with which the PO_2 has to fall below the elbow of the curve for oxygen release to occur. The potential contribution of PFCs to oxygen transport can be assessed by looking at the oxygen consumption required by tissues, the Fct and the PO_2 required to allow this quantity of oxygen to be released in the tissues (28).

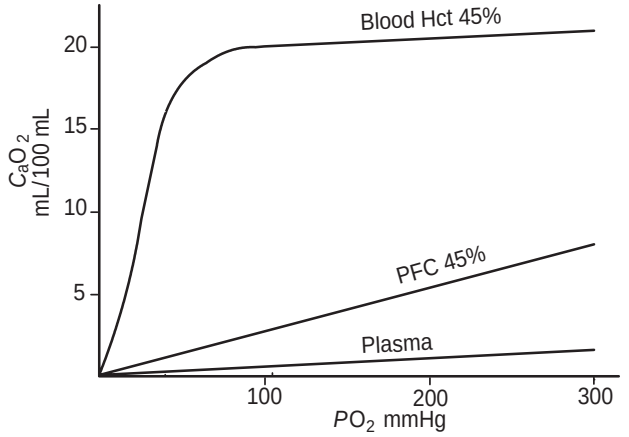


Figure 1. Oxygen content plotted against PO_2 for whole blood, plasma, and a perfluorochemical emulsion, with a 45% content of PFC. Hemoglobin saturates at a PO_2 of 100 mmHg (13.33 kPa), but the curve continues to rise because of the dissolved oxygen in plasma, so the whole blood and plasma lines are parallel above a PO_2 of 100. The line for perflubron is similar to plasma, but has a higher slope because of the greater affinity for oxygen (Hct = hemocrit).

Mixed venous blood has an oxygen content of 15 mL/100 mL; therefore 5 mL/100 mL of oxygen is consumed in the periphery.

If we assume a mixed venous oxygen tension (P_vO_2) level of 40 mmHg (5.33 kPa) and an oxygen extraction of 5 mL/dL, a bloodless animal could survive with a Fct of 45% with a P_aO_2 of 235 mmHg (31.33 kPa). Any increase in P_aO_2 above this value would raise the P_vO_2 by the same amount (Fig. 2). The elevated P_vO_2 in these circumstances could have the beneficial effect of increasing the pressure gradient for oxygen diffusion from the vascular space into the tissues and cells, theoretically increasing tissue oxygenation.

It is, however, difficult to manufacture a 45% emulsion of PFC. In the late 1970s, the Green Cross Corporation in Japan developed a product called Fluosol DA 20%. This solution contains only 10% PFC (Fluosol DA 20% is 20% by weight, 10% by volume). To supply 5 mL/100 mL of oxygen consumption and a P_vO_2 of 40 mmHg (5.33 kPa), a bloodless animal with a fluocrit of 10% would require a P_aO_2 of 920 mmHg (122.65 kPa).

In practice, this would make it difficult to completely replace the blood with PFC emulsion. Although the emulsion is cleared from the vascular space within 24 h, the long tissue half-life of a PFC in the body (months to years), make it unfeasible to continuously redose the patient.

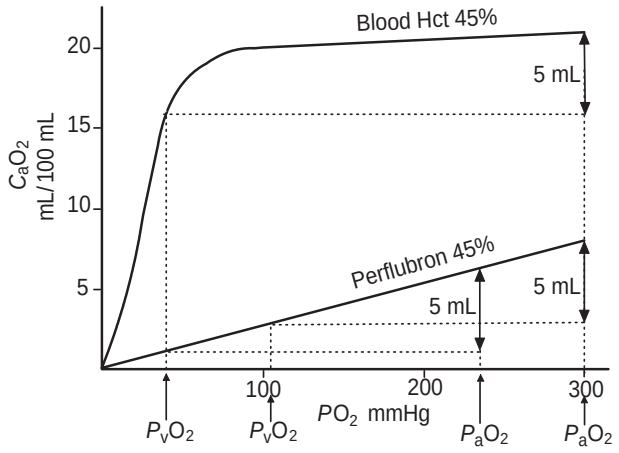


Figure 2. At an arterial PO_2 (P_aO_2) of 300 mmHg (13.33 kPa) blood has an oxygen content of 21 mL/100 mL. If 5 mL/100 mL are extracted the venous PO_2 (P_vO_2) will be < 50 mmHg. A PFC with a Fct of 45% could carry enough oxygen at a P_aO_2 of 235 to deliver 5 mL and give a similar P_vO_2. Increasing the P_aO_2 to 300 mmHg (39.99 kPa), increases the oxygen carried by the PFC so the P_vO_2 will be > 100 mmHg (13.33 kPa) after delivery of 5 mL O_2. (Adapted, with permission, from Woodcock BJ, Tremper KK. Red Blood Cell Substitutes. In: Evers AS, Maze M, editors. Anesthetic Pharmacology, Physiological Principles and Clinical Practice: A Companion to Miller's Anesthesia. Philidelphia: Churchill Livingstone; 2004.)

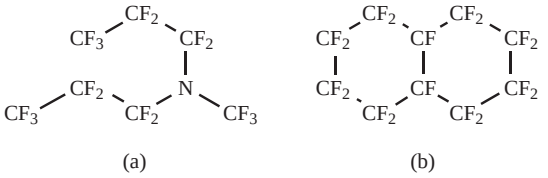


Figure 3. Structures of the constituents of Fluosol DA; perfluorodecalin (a) and perfluortripropylamine (b).

CLINICAL STUDIES WITH PFCS

Fluosol DA

Fluosol DA 20% was developed in the 1970s and was an emulsion composed of two perfluorochemicals; perfluorodecalin, which has emulsion stability and perfluorotripropylamin, which has a shorter half-life (29) (Fig. 3). The surfactant used to maintain this emulsion of Fluosol DA was Pluronic F68, an emulsifier used in other medical products. The product needed to be frozen to maintain stability until used.

Initial clinical studies were in patients who were actively bleeding, required surgery, and refused blood transfusion. A 20-$mL \cdot kg^{-1}$ body weight of Fluosol DA 20% PFC emulsion was transfused preoperatively with a maximum of one additional dose postoperatively (29,30). The patients achieved a maximal fluorocrit of $< 3\%$ with an intravascular half-life of ~ 19 h. Subsequent studies confirmed the oxygen-carrying contribution of the PFC, but could not show a beneficial effect on patient outcome (31).

Fluosol DA 20% gained FDA approval, not for use as a red cell substitute but for intracoronary infusion in patients undergoing angioplasty. Fluosol is no longer being manufactured because of low demand.

Perflubron

Second generation PFC emulsions have been developed. One of these, Oxygent, employs perfluoro-octylbromide or perflubron (32) (Fig. 4). This PFC has one bromine replacing fluorine, making it radiopaque. The emulsion (Oxygent: Alliance Pharmaceutical, San Diego, CA) contains 60% PFC by weight, or 30% by volume. It is emulsified with lecithin (egg yolk phospholipids) and is stable at room temperature for > 6 years.

Because of a short intravascular half-life and the need for a high F_iO_2, it has been suggested that these PFC solutions should be used in conjunction with acute normovolemic hemodilution. In this setting, patients who are expected to have significant surgical blood loss should have two to four units of blood removed immediately before surgery. Initial volume expansion is with crystalloid or colloid. As surgical bleeding continues, the PFC is administered at the first transfusion trigger, allowing for adequate oxygen delivery at low hemoglobin levels. When surgical bleeding stops, the patient would receive the previously harvested autologous blood.

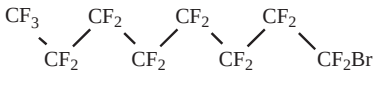


Figure 4. Structure of perfluoro-octylbromide (Perflubron, Oxygent).

The European Perflubron Emulsion Study Group looked at this technique of normovolemic hemodilution in orthopedic surgery. Perflubron was administered in response to transfusion triggers of tachycardia, hypotension, increased cardiac, and decreased mixed venous PO_2. The triggers for transfusion were at least as effectively reversed by perflubron, with 100% oxygen ventilation, as by autologous transfusion or colloid administration (33). Patients who received perflubron had higher mixed venous PO_2 levels, which continued for longer than the other treatment groups. Allogeneic blood transfusion requirements were not reduced by perflubron administration.

A phase III cardiac surgery study of perflubron was voluntarily suspended in 2001 because of an increased incidence of stroke, but it has not been determined if this was due to PFC administration.

A phase III study of perflubron in noncardiac surgery, again from the European Perflubron Emulsion Study Group (34), in 492 patients showed an intraoperative reduction in blood transfusion requirement, but there was no significant reduction by time of the postoperative period. Of the patients in the intent-to-treat population, 33% experienced low (< 20 $mL \cdot kg^{-1}$) blood loss. When these patients were excluded to leave a protocol defined target population, the patients treated by acute normovolemic hemodilution (ANH) and PFC administration experienced a reduction in transfusion that remained significant throughout the study (3.4 vs. 4.9 units). The PFC group also had a significantly greater avoidance of transfusion (26 vs. 16%). The clinical relevance of these findings is hard to elucidate. The control group did not undergo ANH and had higher transfusion triggers during the operative period, it is difficult to tell if the benefit accrued was due to the PFC or the hemodilution technique (35).

Microcirculation and Oxygen-Carrying Colloids

Oxygen-carrying colloids differ from red cells in terms of size. Free hemoglobin molecules in solution range in size from 68,000 to 500,000 Da, and the fluorochemical emulsion particles are $\sim 0.2\,\mu m$. These sizes are within the range of many of the plasma proteins, and like plasma proteins, PFCs are able to cross the capillary basement membrane and participate in extravascular circulation. Animal data have shown that these oxygen-carrying colloids leave the intravascular space and rejoin it through lymphatic circulation (36). Thus, these colloids may be able to provide increased oxygen delivery to the extravascular space (37,38). In addition, studies on microcirculation have shown that plasma flow continues through capillaries even when the capillary is closed to red blood cells. Oxygen delivery to tissues can be provided, even though red cells are not present in the capillaries, through the circulation of oxygen-carrying colloids.

This potential application of blood substitutes, as "therapeutic oxygenating agents" for ischemic tissue, has been

investigated for myocardial ischemia (39–41) and for enhancing the effectiveness of radiation therapy and chemotherapy of ischemic tumors by rendering the tumor hyperoxic (42,43).

Perfluourochemicals have been used in other circumstances as oxygen-carrying molecules. The PFCs have been used for cardiplegia (44) and preservation of transplanted organs (45–47). The PFCs have also been studied in models of myocardial and cerebral infarcts to minimize infarct size (45,48,49).

Liquid Ventilation with PFC

Perfluorochemicals have been used for liquid ventilation in the treatment of acute respiratory distress syndrome (ARDS). The PFCs bind oxygen and carbon dioxide avidly. Perflubron (LiquiVent, Alliance Pharmaceutical Corp., San Diego, CA) can be instilled into the endotracheal tube of a ventilated patient with ARDS until a fluid level is seen outside the patient. The ET is then connected to the normal ICU ventilator and sufficient gas transfer occurs across the PFC–gas interface to allow oxygenation of the patient and CO_2 clearance (50,51). This PFC has excellent surfactant properties and is possibly able to stent open alveoli (leading it to be termed "liquid PEEP" or "PEEP in a bottle") (52). There are also benefits of increased secretion clearance and possible antiinflammatory effects (53,54). Studies to date have not shown any benefit over conventional ventilation (55).

Future of PFCs

The PFCs have inherent limitations of a short endovascular half-life and the requirement for high inspired oxygen. These problems limit the use of PFC emulsions to acute settings in which supplemental oxygen is readily available. It is yet to be seen whether PFCs can have a useful role as blood substitutes.

HEMOGLOBIN SOLUTIONS

Hemoglobin Solutions: Oxygen Content

When a solution of free hemoglobin (FHb) has the same $P50$ as blood (27 mmHg 3.59 kPa), there is no difference between the hemoglobin solution oxygen-content curve and the curve for normal whole blood.

$$C_aO_2 = (\mathrm{FHb} \times 1.34 \times \mathrm{F}S_aO_2) + (\mathrm{Hb} \times 1.34 \times S_aO_2) + (0.003 \times P_aO_2) \quad (3)$$

[FHb is the concentration of free hemoglobin solution in blood ($g \cdot dL^{-1}$), and FS_aO_2, is the saturation of FHb.]

In clinical practice, arterial content of oxygen can be calculated using a single term for hemoglobin and saturation. A spectrophotometrically measured total hemoglobin should be used, which will measure total hemoglobin present, whereas the hematocrit only measures red blood cell hemoglobin. The saturation measured by an oximeter gives a mean saturation of both forms of hemoglobin because the device measures the amount of oxyhemoglobin and divides it by total hemoglobin to achieve the calculated saturation.

Formulation of Hemoglobin Solutions

A solution of hemoglobin from lyzed human red cells is unusable as a blood substitute for a variety of reasons. Hemoglobin outside the red cell membrane loses its tetrameric form and breaks down into dimers. The abundance of dimers in plasma has a pronounced oncotic effect creating an excessively high colloid oncotic pressure (41). This would draw fluid from the extracellular space and would cause an increase in circulating blood volume. The dimers have a molecular weight of 32,000 Da and are able to cross the renal glomerular basement membrane leading to a potent osmotic diuretic effect. The loss of oxygenated hemoglobin in the urine gave the early solutions the reputation of being "red mannitol".

The loss of 2,3-DPG, which is normally maintained inside the RBC, reduces the P50 of hemoglobin to 12–14 mmHg (1.59–1.86 kPa) from a normal level of 26. This shifts the oxygen dissociation curve markedly to the left, meaning that the free hemoglobin will avidly bind oxygen during passage through the lungs but will not release it in the peripheral tissues unless the PO_2 is extremely low.

The early attempts at making a hemoglobin solution used resuspended hemoglobin filtered from lyzed, outdated, human blood. This solution caused a high incidence of renal failure, which proved not to be due to the free hemoglobin, but due to the "stroma" of residual red blood cell elements left after cell lysis (56).

Several types of hemoglobin solution have been developed, and have taken different approaches to these problems. One product is produced from modified polymerized bovine hemoglobin (Hemopure, HBOC-201, Biopure, Cambridge, USA) (57,58). The dimer form is polymerized to form a larger roughly octomeric molecules (Fig. 5). It is

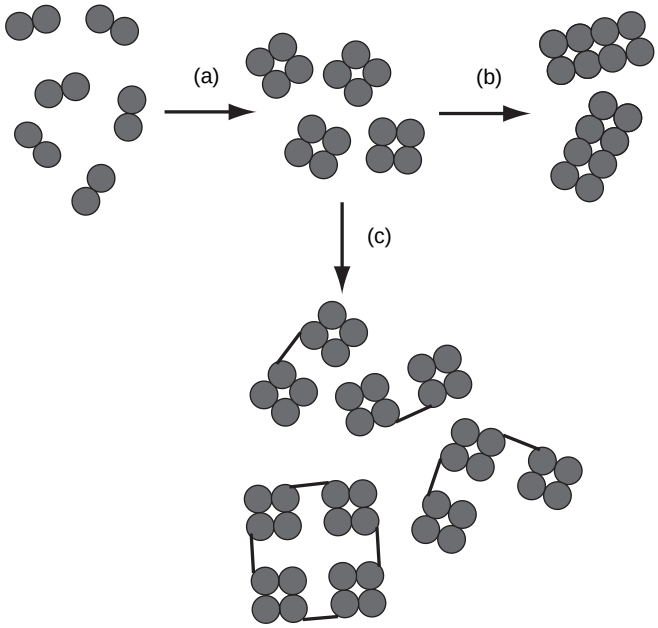


Figure 5. Hemoglobin dimer subunits can be polymerized (a) to form tetramers. These can be further polymerized (b) to form octamers or larger units, or can be cross-linked (c) to increase molecular size.

Table 2. Hemoglobin Solutions in Clinical Trials

Product (Manufacturer)	Configuration	*P*50	Status
	Bovine Hemoglobins		
HBOC201 Hemopure (Biopure)	Polymerized with glutaraldehyde	34 mmHg	Licensed in South Africa and for veterinary use. FDA review; further studies required
PEG-hemoglobin (Enzon)	PEG (polyethylene glycol) conjugated noncross-linked, encapsulated	20 mmHg	Phase 1 melanoma studies completed Now discontinued
	HumanHemoglobins		
Polyheme (Northfield)	Polymerized, tetramer-free	30 mmHg	Phase 3 study, before FDA review
Hemolink (Hemosol)	Polymerized with *o*-raffinose	32 mmHg	Under review United Kingdom and Canada
DCLHb (Baxter)	Cross-linked with diaspirin	32 mmHg	Discontinued
PHP (Curacyte)	Pyridoxylated		Studied in SIRS
	Recombinant Hemoglobins		
Optro (Somatogen)	Genetically fused	17 mmHg	Discontinued

then filtered to remove the smaller molecular weight tetrameric hemoglobin molecules. This reduces the oncotic pressure and prevents passage of the molecule though the glomerulus into the urine. Polymerization also increases the *P*50 into the normal range and allows the hemoglobin solution to release oxygen in the tissues (41).

Other products have been developed from outdated human blood: Cross-linking hemoglobin molecules with pyridoxal-5-phosphate (Fig. 5), which acts as an artificial 2, 3-DPG, can increase the *P*50. This technique is used in PolyHeme polymerized hemoglobin solution (Northfield Pharmaceutical, Chicago, IL) (59–61), and pyridoxylated hemoglobin polyoxyethylene conjugate or PHP hemoglobin (62). Diaspirin Cross-Linked Hemoglobin, DCLHb (Baxter Healthcare, Chicago, IL) (63,64) and Hemolink (Hemosol, Toronto, Ontario, Canada) also have a *P*50 within the normal range and have an increased size due to cross-linking or polymerization.

The size of the hemoglobin molecule can also be increased by attaching the dimer to a large non-hemoglobin molecule, for example, polyethylene glycol used in PEG–hemoglobin (Enzon, Piscataway, NJ). A final method of increasing molecular size is to encapsulate hemoglobin in phospholipid vesicles or liposomes (65).

Hemoglobin cannot only be obtained from outdated human blood, but also from bovine blood. There is a tremendous supply of bovine blood, since nearly 1 million units/day are produced as a byproduct of meat production. Bovine hemoglobin does not normally require 2, 3-DPG, and maintains a *P*50 in the range of 32 mmHg (4.26 kPa), even as a dimer (66). Concern over the spread of BSE or variant vCJD may impact the development of bovine products (67). The FDA has restricted the import of bovine products from Europe because of the outbreak of Mad Cow disease there (68).

Recombinant hemoglobin has been used as an alternative source of hemoglobin. Optro (Somatogen), was engineered to have a *P*50 in the normal range. However, manufacture has been discontinued (67).

The hemoglobin solutions that have undergone clinical trial are listed in Table 2. Since these solutions are produced from different hemoglobin sources, and have different sizes and different *P*50 values, each needs to be evaluated as a separate drug with its own effectiveness and toxicity profile.

All of these products have a relatively short intravascular half-life compared to hemoglobin contained in red blood cells, and are cleared from the vascular space by the reticuloendothelial system with a half-life of $\sim$24 h.

Hemodynamic Effects of Hemoglobin Solutions

Pulmonary and systemic hypertension have been observed in studies in animals (63,69–72) and human clinical studies (36,38,57,64,73–78). The cause of this appears to be related to the nitric oxide (NO) scavenging properties of hemoglobin. Endothelium derived relaxant factor (EDRF) was identified as NO in the 1990s and its role in controlling vascular resistance was elucidated. Nitric oxide is produced in the endothelial cells of blood vessel walls, and produces smooth muscle relaxation, thereby causing vasodilatation. As blood flow increases, nitric oxide is carried away, reducing its concentration and causing vasoconstriction (79). Binding of nitric oxide to hemoglobin plays an important role in its removal, and thereby the control of vascular tone. Free hemoglobin binds nitric oxide more avidly than hemoglobin within the red cell and nitric oxide clearance is increased (80). This leads to hypertension in the pulmonary and systemic vascular beds. Smaller hemo-

globin molecule size appears to increase NO clearance, and pulmonary hypertension becomes more problematic. The severity of this side effect varies with the different forms and preparations of hemoglobin.

The observed vasoconstriction with hemoglobin solutions led to the evaluation of DCLHb solution as an "all-in-one" therapeutic strategy in vasodilated shock states (38) as a vasopressor in critically ill patients with septic shock. In septic shock or systemic inflammatory response syndrome, the vasodilated state may be due to excessive synthesis of NO. The scavenging effect of the hemoglobin solution may be beneficial in restoring vascular tone in this setting (37). Tissue perfusion may be improved because of the small size of the hemoglobin molecules compared to red cells, improving oxygen delivery through the microcirculation. In a mouse model of gram-negative sepsis, hemoglobin solution was associated with increased circulating tumor necrosis factor and increased lethality (81). In pigs with septic shock, DCLHb administration restored blood pressure and allowed a reduction in dopamine infusion being used for resuscitation (82). Further work is required to determine the utility of hemoglobin solutions in critical illness (83).

Although NO scavenging may play a major role in the vasoconstriction seen with hemoglobin solutions it may not explain the whole picture. Hemoglobin molecules with similar rates of NO binding may have strikingly different degrees of vasoconstriction. The fall in PO_2 in terminal arterioles may play a large role in the autoregulation of microvascular circulation by causing vasoldilation. Hemoglobin solutions may deliver excess oxygen to the terminal capillaries causing vasoconstriction and paradoxically reducing oxygen delivery to the tissue (84).

This proposed effect on microcirculation has led to the adoption of a "counterintuitive" approach to hemoglobin solution development. An anemic but hyperoncotic solution with a very low $P50$ could delay oxygen release and prevent vasoconstriction (85). Such a solution has been developed in MalPEG-Hb (maleimide-activated polyethylene glycol conjugated hemoglobin). This MalPEG-Hb solution has a $P50$ of 5.5 mmHg and has been shown to improve microvascular circulation in hypovolemic shock in hamsters (86).

Duration of Action of Hemoglobin Solutions

Just as free hemoglobin is scavenged following intravascular hemolysis, the RE system scavenges hemoglobin solution from the blood stream. The effective intravascular life of the hemoglobin solution is $\sim$ 12–48 h, depending on the dose (66). Therefore, hemoglobin solutions, like the PFCs, will not have a role in maintaining oxygen-carrying capacity in chronically anemic patients. They may be used in acute, limited blood-loss situations as resuscitative fluids, especially in combination with perioperative autologous hemodilution to minimize loss of the patients own blood (87,88).

Hematopoietic Effect of Hemoglobin Solutions

Hemoglobin solutions may have a profound effect in stimulating erythropoiesis. Studies in which animals are hemodiluted to a low Hct, and given a transfusion of hemoglobin solution, demonstrate a pronounced level of red blood cell production exceeding that expected even with the administration of exogenous erythropoietin (36).

Free iron liberated when hemoglobin solutions are broken may stimulate erythropoiesis, reducing the requirement for subsequent red cell transfusion (89,90). Transfusion of diaspirin cross-linked hemoglobin (DCLHb) to transfuse cardiac surgery patients in the postbypass period resulted in 19% of patients not requiring packed red blood cell transfusion, although the DCLHb was rapidly cleared from the circulation (77). After acute normovolemic hemodilution (ANH) with HBOC-201 in human volunteers there were increases in serum iron, ferritin, and erythropoietin that did not occur following ANH with Ringer's lactate solution (91).

Other Uses of Hemoglobin Solutions

Hemoglobin solutions have an oxygen delivery curve similar to that of hemoglobin in red blood cells. At low temperatures the curve is shifted to the left, preventing release of the bound oxygen, they therefore may not be of any benefit as a constituent of cardioplegia solutions or organ preservatives used for transplanted organs.

Artifical oxygen carriers have a potential for abuse by elite athletes, and international sporting federations have already added the class of agent to their banned substance lists (92).

Clinical Trials: Hemoglobin Solutions

HBOC-201 (Hemopure). The bovine product HBOC-201 (Hemopure, Biopure, Cambridge, MA) is produced from modified polymerized bovine hemoglobin and has been studied in patients undergoing abdominal aortic aneurysm surgery. Treatment with the hemoglobin solution produced an increase in pulmonary and systemic vascular resistance and an associated decrease in cardiac output (57). In another study in aortic surgery HBOC-201, 27% of patients did not need blood transfusion, but the median transfusion requirement was not decreased. Mean arterial blood pressure increased by 15% in this study (75). Vasoconstriction and hypertension have been noted in many studies with HBOC-201. A study looking at preoperative hemodilution with bovine HBOC-201 before liver resection showed an increase in systemic vascular resistance and a fall in cardiac output (73). More recently, a study by Wahr et al. (36) found this product to be useful in reducing the amount of allogeneic blood in operative patients. Increases in pulmonary or systemic resistance were not seen, but a mild increase in blood pressure occurred.

In postoperative cardiac surgery patients, administration of up to 1000 mL HBOC-201 prevented packed red blood cell administration in 34% of patients who would otherwise have received it (93). Although the HBOC-201 had a short duration in the circulation, Hct was restored rapidly, perhaps due to a hematinic effect. The HBOC-201 patients had a slightly greater increase in blood pressure after transfusion.

HBOC-201 has been administered to patients with sickle cell anemia and may have a role in the treatment of vasoocclusive or aplastic crises in this disease (94,95).

The HBOC-201 can be used instead of PRBCs to transfuse patients with severe autoimmune hemolytic anemia, the hemoglobin solution does not have the surface antigens associated with red blood cells (96). In the presence of a Hct of 4.4% the hemoglobin solution reversed both lactic acidosis and myocardial ischemia.

In South Africa, HBOC-201 (Hemopure) has been licensed to treat anemia in surgical patients; this is the first time a hemoglobin solution has reached the market in humans. It is also licensed, as Oxyglobin, for veterinary use in the United States and Europe. The FDA approval for human use in the United States has been delayed pending a request for further information and animal testing.

Diaspirin Cross-Linked Hemoglobin, DCLHb

The DCLHb (Baxter Healthcare, Chicago, IL) has been developed from outdated human blood. Clinical studies to date have had varying results with DCLHb.

The DCLHb effectively scavenges NO and has been noted to increase pulmonary and systemic vascular resistance in the hemodilution model in animals to the point of reducing cardiac index and oxygen delivery relative controls (63). Pulmonary and systemic hypertension with rises in SVR and PVR also occurred in human studies of DCLHb given after cardiac surgery, and this led to decreased cardiac output compared to patients given red cell transfusion (77). Another study (76) showed that DCLHb could be used to reduce the number of patients requiring perioperative packed red blood cell (PRBC) transfusions following vascular, orthopedic, and abdominal surgery compared with patients randomized to PRBC transfusions. However, the total PRBC and other blood product requirements of the two groups were similar over the subsequent week. Side effects of hypertension, jaundice and hemoglobinuria were noted. One death from respiratory distress syndrome was attributed to DCLHb, and the study was terminated early.

This solution was also investigated in a randomized study of patients with acute ischemic stroke (64). Patients receiving the hemoglobin solution had more deaths, serious adverse outcomes, and worse outcome scale scores.

Studies in patients with hemorrhagic shock in the United States and Europe were discontinued because of increased mortality in the U.S. study, and a lack of benefit with DCLHb administration in the European study (97).

Because of these results Baxter has discontinued further work with DCLHb. He subsequently acquired Somatogen, manufacturer of Optro, a first generation recombinant hemoglobin, development of which has since been discontinued.

PolyHeme

PolyHeme (Northfield Pharmaceutical, Chicago, IL) has been used in trials treating acute trauma patients. Gould et al. administered 1–20 units of PolyHeme to 171 patients with urgent blood loss; 81 patients received 5 or more units and 34 received 10–20 units (98). Overall there was a mortality of 10.5 compared to 16% of historical controls who declined blood transfusion because of religious reasons during surgery, and had Hb levels < 8 g/dL. However, there was no comparison with controls receiving transfusion of allogeneic packed red blood cells. An earlier study, also by Gould, randomly assigned 44 trauma patients to either receive blood (23 patients) or up to 6 units of PolyHeme (21 patients) (60). There were no adverse effects on pulmonary and systemic vascular resistance or cardiac output from the administration of PolyHeme. There were no differences in patient outcomes, although there was a reduced need for red blood cells in the PolyHeme group at day 1, which was no longer seen by day 3. The lack of effect of PolyHeme on systemic and peripheral vascular resistance is attributed to its manufacturing process, which filters out the smaller tetrameric hemoglobin (59,60). It is speculated that the smaller size hemoglobin elements can defuse through the vessel wall, increasing NO scavenging and producing vasoconstriction.

Studies have found an intravascular half-life of 24 h of PolyHeme, which is longer than that found in previous studies that found a halflife in the range of 9–12 h. It may be that the half-life of these products is dose dependent; studies showing a larger dose produces a longer half-life.

PolyHeme is currently being assessed in a large multicenter study compared to saline for trauma patients. An application to the FDA for approval may follow this study and PolyHeme could have a role in future clinical practice.

Hemolink

Hemolink (Hemosol, Toronto, Ontario, Canada) is an O-raffinose cross-linked human hemoglobin, which has been shown to cause vasoconstriction in animals (70). As with other cross-linked hemoglobin solutions, the hemodynamic effects are less pronounced than with unmodified hemoglobin solutions (99) but exceed that of polymerized hemoglobin solutions.

A Phase I study in healthy human volunteers showed the solution was well tolerated apart from some moderate to severe abdominal pain, which occurred in all subjects at higher doses. Blood pressure rose by 14% following administration, and with higher doses this elevation lasted 24 h (78). The findings of a Phase II study in coronary artery bypass (CAB) surgery (74) reported a 7–10% increase in blood pressure, which was not statistically significant, and a reduction in the number of patients requiring red cell transfusion from 57 to 10%. A further report by the same group using intraoperative autologous donation and volume replacement with Hemolink or pentastarch in CAB surgery abolished the need for intraoperative transfusion (0 vs. 17% in the pentastarch group) (100). The reduction in transfusion requirement continued at 1 day (7 vs. 37%) and 5 days (10 vs. 47 %) after surgery. Adverse effects included hypertension (43 vs. 17%) and atrial fibrillation (37 vs. 17%).

A similar study in CAB patients using intraoperative autologous blood donation (IAD) and volume replacement with Hemolink or pentastarch showed a reduction in transfusion from 76% in the pentastarch group to 56% with Hemolink (101). This was compared to an historical group of patients in whom IAD was not used, who required transfusion in 95% of operations. Hypertension was again noted in the Hemolink treated group.

Hemolink is under review for approval by the drug agencies of the United States, Canada, and the United Kingdom.

PEG Hemoglobin

The PEG Hemoglobin (Enzon, Piscataway, NJ) is a bovine hemoglobin conjugated with polyethylene glycol. This increases the molecular size without using cross-linking between molecules. Retention in the circulation is increased as the conjugated molecule does not cross the glomerular basement membrane. Administration in dogs resulted in no elevation of blood pressure and it was well tolerated (102).

The PEG hemoglobin has been used to sensitize tumors to chemotherapy (103) and radiotherapy in rodents (104). The small molecular size, compared to red cells, improves microvascular oxygenation, and tumors that are hypoxic become more responsive to chemotherapy and radiotherapy.

The PEG hemoglobin has also been used in rabbits, in the preservative perfusate, and for protection of transplanted hearts during the ischemic period with improvement in cardiac function post-transplant (105).

PHP

Pyridoxilated hemoglobin polyoxyethylene conjugate or PHP (Curacyte, Chapel Hill, NC) is a human hemoglobin solution, cross-linked by pyridoxal-5-phosphate, which is being developed as a NO scavenger for use in septic shock and systemic inflammatory response syndrome (62). A Phase II study has been completed and a Phase III study is in progress looking at PHP for the treatment of NO induced shock.

Encapsulated Hemoglobins

The problems of small hemoglobin molecule size, leading to NO scavenging, high oncotic pressures, and osmotic diuresis can be countered by encapsulating the hemoglobin. This mimics the natural presentation of hemoglobin in whole blood and is sometimes referred to as "neo red cells" (65). Most work has used liposome encapsulated hemoglobin (LEH), but biodegradable polymer microcapsules have also been used (106). Circulation time of the hemoglobin is increased by encapsulation and can be increased, from 18 to 65 h, by polyethylene glycol (PEG) modification (107), these long-lasting derivatives have been named "stealth" liposomes. However, there is significant accumulation of liposomes in the liver and spleen, when LEH is given, causing vacuolization seen on liver biopsy. Liver transaminases may also be elevated (41).

BIBLIOGRAPHY

Cited References

1. Epstein JS. The US blood supply. Am Fam Phys 2000;61: 549–550.
2. American Society of Anesthesiologists Task Force on Blood Component Therapy, Practice Guidelines for blood component therapy. Anesthesiology 1996;84:732–747.
3. Spahn DR, Casutt M. Eliminating blood transfusions: new aspects and perspectives. Anesthesiology 2000;93:242–255.
4. Cohn SM. Blood substitutes in surgery. Surgery 2000;127: 599–602.
5. Ketcham EM, Cairns CB. Hemoglobin-based oxygen carriers: development and clinical potential. Ann Emerg Med 1999; 33:326–337.
6. Myhre BA, Bove JR, Schmidt PJ. Wrong blood—a needless cause of surgical deaths. Anest Analg 1981;60:777–778.
7. Nichollis MD. Transfusions: morbidity and mortality. Anaesth Intensive Care 1993; 15–19.
8. Sazama K. Reports of 355 transfusion-associated deaths: 1976 through 1985. Transfusion (Paris) 1990;30:583–590.
9. Newman RJ, Podolsky D, Loeb P. Bad blood. US News World Rep 1994;116:68–70.
10. Lackritz EM. Prevention of hiv transmission by blood transfusion in the developing world: achievements and continuing challenges. AIDS 1998;12:81–86.
11. Chamberland M, Khabbaz RF. Emerging issues in blood safety. Inf Dis Clin North Am 1998;12:217–229.
12. Dodd RY. The risk of transfusion-transmitted infection. N Engl J Med 1992;327:419–421.
13. Yoshikawa A, Fukuda S, Itoh K, Kosaki N, Suzuki T, Hirakawa K, Nakao H, Inoue T, Fukuda M, Okamoto H. Infection with hepatitis g virus and its strain variant, the gb agent (gbv-c), among blood donors in japan. Transfusion (Paris) 1997;37:657–663.
14. Tacke M, Kiyosawa K, Stark K, Schlueter V, Ofenloch-Haehnle B, Hess G, Engel AM. Detection of antibodies to a putative hepatitis g virus envelope protein. Lancet 1997;349:318–320.
15. Fiebig EW, Busch MP. Emerging infections in transfusion medicine. Clin Lab Med 2004;24:797.
16. Alter MJ, Gallagher M, Morris TT, Moyer LA, Meeks EL, Krawczynski K, Kim JP, Margolis HS. Acute non-a-e hepatitis in the united states and the role of hepatitis g virus infection. Sentinel counties viral hepatitis study team. N Engl J Med 1997;336:741–746.
17. Alter HJ, Nakatsuji Y, Melpolder J, Wages J, Wesley R, Shih JW, Kim JP. The incidence of transfusion-associated hepatitis g virus infection and its relation to liver disease. N Engl J Med 1997;336:747–754.
18. Karayiannis P, Thomas HC. Current status of hepatitis g virus (gbv-c) in transfusion: is it relevant? Vox Sang 1997;73:63–69.
19. Ricketts MN, Cashman NR, Stratton EE, ElSaadany S. Is creutzfeldt-jakob disease transmitted in blood? Emerg Infect Dis 1997;3:155–163.
20. Houston F, Foster JD, Chong A, Hunter N, Bostock CJ. Transmission of bse by blood transfusion in sheep. Lancet 2000;356:999–1000.
21. Mitka M. Blood groups differ on donor deferral. JAMA 2001;285:1694–1695.
22. Iwamoto M, Jernigan DB, Guasch A, Trepka MJ, Blackmore CG, Hellinger WC, Pham SM, Zaki S, Lanciotti RS, Lance-Parker SE, DiazGranados CA, Winquist AG, Perlino CA, Wiersma S, Hillyer KL, Goodman JL, Marfin AA, Chamberland ME, Petersen LR. West Nile Virus in Transplant Recipients Investigation Team, Transmission of west nile virus from an organ donor to four transplant recipients. N Engl J Med 2003;348:2196–203.
23. Blumberg N, Triulzi DJ, Heal JM. Transfusion-induced immunomodulation and its clinical consequences. Transfus Med Rev 1990;4:24–35.
24. Bordin JO, Blajchman MA. Immunosuppressive effects of allogeneic blood transfusions: implications for the patient with a malignancy. Hematol Oncol Clin North Am 1995;9: 205–218.
25. Biedler AE, Schneider SO, Seyfert U, Rensing H, Grenner S, Girndt M, Bauer I, Bauer M. Impact of alloantigens and

storage-associated factors on stimulated cytokine response in an in vitro model of blood transfusion. Anesthesiology 2002;97:1102–1109.
26. Clark Jr LC, Gollan F. Survival of mammals breathing organic liquids equilibrated with oxygen at atmospheric pressure. Science 1966;152:1755–1766.
27. Gehes RP, Monroe RG, Taylor K. Survival of rats having red cells totally replaced with emulsified fluorocarbon. Fed Pro 1968;27:384.
28. Woodcock BJ, Tremper KK. Red Blood Cell Substitutes. In: Evers AS, Maze M, editors. Anesthetic Pharmacology: Physiological Principles and Clinical Practice: A Companion to Miller's Anesthesia. Philidelphia: Churchill Livingstone; 2004.
29. Tremper KK, Friedman AE, Levine EM, Lapin R, Camarillo D. The preoperative treatment of severely anemic patients with a perfluorochemical oxygen-transport fluid, Fluosol-DA. N Engl J Med 1982;307:277–283.
30. Tremper KK, Levine EM, Waxman K. Clinical experience with Fluosol-DA (20%) in the United States. Int Anesthesiol Clinic 1985;23:185–197.
31. Gould SA, Rosen AL, Sehgal LR, Sehgal HL, Langdale LA, Krause LM, Rice CL, Chamberlin WH, Moss GS. Fluosol-DA as a red-cell substitute in acute anemia. N Engl J Med 1986;314:1653–1656.
32. Keipert PE, Faithfull NS, Bradley JD, Hazard DY, Hogan J, Levisetti MS, Peters RM. Oxygen delivery augmentation by low-dose perfluorochemical emulsion during profound normovolemic hemodilution. Adv Exp Med Biol 1994;345: 197–204.
33. Spahn DR, van Brempt R, Theilmeier G, Reibold JP, Welte M, Heinzerling H, Birck KM, Keipert PE, Messmer K, Heinzerling H, Birck KM, Keipert PE, Messmer K. Perflubron emulsion delays blood transfusions in orthopedic surgery. European perflubron emulsion study group. Anesthesiology 1999;91:1195–208.
34. Spahn DR, Waschke KF, Standl T, Motsch J, Van Huynegem L, Welte M, Gombotz H, Coriat P, Verkh L, Faithfull S, Keipert P. Use of perflubron emulsion to decrease allogeneic blood transfusion in high-blood-loss non-cardiac surgery: results of a European phase 3 study. Anesthesiology 2002;97:1338–1349.
35. Tremper KK. Perfluorochemical "red blood cell substitutes": the continued search for an indication. Anesthesiology 2002;97:1333–1334.
36. Wahr JA, Levy JH, Kindscher J. Hemodynamic effects of a bovine based oxygen carrying solution in surgical patients. Anesthesiology 1996;85:A347.
37. Creteur J, Vincent JL. Hemoglobin solutions: an "all-in-one" therapeutic strategy in sepsis? Crit Care Med 2000;28: 894–896.
38. Reah G, Bodenham AR, Mallick A, Daily EK, Przybelski RJ. Initial evaluation of diaspirin cross-linked hemoglobin (DCLHB) as a vasopressor in critically ill patients. Crit Care Med 1997;25:1480–1488.
39. Robalino BD, Marwick T, Lafont A, Vaska K, Whitlow PL. Protection against ischemia during prolonged balloon inflation by distal coronary perfusion with use of an autoperfusion catheter or Fluosol. J Am Coll Cardiol 1992;20:1378–1384.
40. Kent KM, Cleman MW, Cowley MJ, Forman MB, Jaffe CC, Kaplan M, King SB 3rd, Krucoff MW, Lassar T, McAuley B. et al. Reduction of myocardial ischemia during percutaneous transluminal coronary angioplasty with oxygenated Fluosol. Am J Cardiol 1990;66:279–284.
41. Creteur J, Sibbald W, Vincent JL. Hemoglobin solutions–not just red blood cell substitutes. Crit Care Med 2000;28: 3025–3034.
42. Teicher BA, Schwartz GN, Dupuis NP, Kusomoto T, Liu M, Liu F, Northey D. Oxygenation of human tumor xenografts in nude mice by a perfluorochemical emulsion and carbogen breathing. Artif Cells, Blood Sub Immobil Biotechnol 1994; 22:1369–1375.
43. Teicher BA. An overview on oxygen carriers in cancer therapy. Artif Cells, Blood Sub Immobil Biotechnol 1995; 23:395–405.
44. Martin SM, Laks H, Drinkwater DC, Stein DG, Capouya ER, Pearl JM, Barthel SW, Chang P, Kaczer E, Bhuta S. Perfluorochemical reperfusion yields improved myocardial recovery after global ischemia. Ann Thorac Surg 1993;55: 954–960.
45. Kloner RA, Hale S. Cardiovascular applications of fluorocarbons in regional ischemia/reperfusion. Artif Cells Blood Substit Immobil Biotechnol 1992;22:1069–1081.
46. Segel LD, Follette DM, Iguidbashian JP, Contino JP, Castellanos LM, Berkoff HA, Kaufman RJ, Schweighardt FK. Posttransplantation function of hearts preserved with fluorochemical emulsion. J Heart Lung Trans 1994;13: 669–680.
47. Grunert A, Qiu H, Muller I, Schuh S, Steinbach G, Wennauer R, Wolf C, Von Schenck H. A new extracorporeal perfusion system: prolongation of liver organ vitality beyond 24 hours. Ann N Y Acad Sci 1994;723:488–490.
48. Premaratne S, Harada RN, Chun P, Suehiro A, McNamara JJ. Effects of perfluorocarbon exchange transfusion on reducing myocardial infarct size in a primate model of ischemia-reperfusion injury: a prospective, randomized study. Surgery 1995;117:670–676.
49. Cole DJ, Schell RM, Drummond JC, Pryzbelski RJ, Marcantonio S. Focal cerebral ischemia in rats: effect of hemodilution with alpha-alpha cross-linked hemoglobin on brain injury and edema. Can J Neurol Sci 1993;20:30–36.
50. Gauger PG, Overbeck MC, Chambers SD, Cailipan CI, Hirschl RB. Partial liquid ventilation improves gas exchange and increases EELV in acute lung injury. J Appl Physiol 1998;84:1566–1572.
51. Hirschl RB, Pranikoff T, Wise C, Overbeck MC, Gauger P, Schreiner RJ, Dechert R, Bartlett RH. Initial experience with partial liquid ventilation in adult patients with the acute respiratory distress syndrome. JAMA 1996;275:383–389.
52. Wong DH. Liquid ventilation: more than "PEEP in a bottle"? Crit Care Med 1999;27:1052–1053.
53. Colton DM, Till GO, Johnson KJ, Dean SB, Bartlett RH, Hirschl RB. Neutrophil accumulation is reduced during partial liquid ventilation. Crit Care Med 1998;26:1716–1724.
54. Mrozek JD, Smith KM, Bing DR, Meyers PA, Simonton SC, Connett JE, Mammel MC. Exogenous surfactant and partial liquid ventilation: physiologic and pathologic effects. Am J Resp Crit Care Med 1997;156:1058–1065.
55. Hirschl RB, Conrad S, Kaiser R, Zwischenberger JB, Bartlett RH, Booth F, Cardenas V. Partial liquid ventilation in adult patients with ARDS: a multicenter phase I-II trial. Adult PLV Study Group. Ann Surg 1998;228:692–700.
56. Rabiner SF, Friedman LH. The role of intravascular haemolysis and the reticulo-endothelial system in the production of a hypercoagulable state. Br J Haematol 1968;14: 105–118.
57. Kasper SM, Grune F, Walter M, Amr N, Erasmi H, Buzello W. The effects of increased doses of bovine hemoglobin on hemodynamics and oxygen transport in patients undergoing preoperative hemodilution for elective abdominal aortic surgery. Anesth Analg 1998;87:284–291.
58. Standl T, Burmeister MA, Horn EP, Wilhelm S, Knoefel WT, Schulte am Esch J. Bovine haemoglobin-based oxygen

carrier for patients undergoing haemodilution before liver resection. Br J Anaesth 1998;80:189–194.
59. Johnson JL, Moore EE, Offner PJ, Haenel JB, Hides GA, Tamura DY. Resuscitation of the injured patient with polymerized stroma-free hemoglobin does not produce systemic or pulmonary hypertension. Am J Surg 1998;176:612–617.
60. Gould SA, Moore EE, Hoyt DB, Burch JM, Haenel JB, Garcia J, DeWoskin R, Moss GS. The first randomized trial of human polymerized hemoglobin as a blood substitute in acute trauma and emergent surgery. J Am Coll Surg 187:113–20; discussion 1998; 120–122.
61. Sehgal LR, Rosen AL, Gould SA, Sehgal HL, Moss GS. Preparation and in vitro characteristics of polymerized pyridoxylated hemoglobin. Transfusion (Paris) 1983;23: 158–162.
62. Privalle C, Talarico T, Keng T, DeAngelo J. Pyridoxalated hemoglobin polyoxyethylene: a nitric oxide scavenger with antioxidant activity for the treatment of nitric oxide-induced shock. Free Rad Biol Med 2000;28:1507–1517.
63. DeAngeles DA, Scott AM, McGrath AM, Korent VA, Rodenkirch LA, Conhaim RL, Harms BA. Resuscitation from hemorrhagic shock with diaspirin cross-linked hemoglobin, blood, or hetastarch. J Trauma-Injury Inf Crit Care 42:406–412; discussion 1997; 412–414.
64. Saxena R, Wijnhoud AD, Carton H, Hacke W, Kaste M, Przybelski RJ, Stern KN, Koudstaal PJ. Controlled safety study of a hemoglobin-based oxygen carrier, DCLHB, in acute ischemic stroke. Stroke 1999;30:993–996.
65. Rudolph AS. Encapsulated hemoglobin: current issues and future goals. Artif Cells, Blood Sub Immobiliz Biotechnol 1994;22:347–360.
66. Hughes GS Jr., Antal EJ, Locker PK, Francom SF, Adams WJ, Jacobs EE Jr Physiology and pharmacokinetics of a novel hemoglobin-based oxygen carrier in humans. Crit Care Med 1996;24:756–764.
67. Winslow RM. Blood substitutes. Adv Drug Del Rev 2000;40: 131–142.
68. USDA Interim Rule on Import Restrictions of Ruminant Material from Europe. Fed Proc 1998;63:406–408.
69. Ulatowski JA, Nishikawa T, Matheson-Urbaitis B, Bucci E, Traystman RJ, Koehler RC. Regional blood flow alterations after bovine fumaryl beta beta-crosslinked hemoglobin transfusion and nitric oxide synthase inhibition. Crit Care Med 1996;24:558–565.
70. Ning J, Wong LT, Christoff B, Carmichael FJ, Biro GP. Haemodynamic response following a 10% topload infusion of Hemolink™ in conscious, anaesthetized and treated spontaneously hypertensive rats. Transfus Med 2000;10:13–22.
71. Krieter H, Hagen G, Waschke KF, Kohler A, Wenneis B, Bruckner UB, van Ackern K. Isovolemic hemodilution with a bovine hemoglobin-based oxygen carrier: effects on hemodynamics and oxygen transport in comparison with a non-oxygen-carrying volume substitute.[See comment]. J Cardiothor Vas Anesthes 1997;11:3–9.
72. Maxwell RA, Gibson JB, Fabian TC, Proctor KG. Resuscitation of severe chest trauma with four different hemoglobin-based oxygen-carrying solutions. J Trauma-Injury Inf Crit Care 49:200–209; discussion 2000; 209–211.
73. Standl T, Wilhelm S, Horn EP, Burmeister M, Gundlach M, Schulte am Esch J. Preoperative hemodilution with bovine hemoglobin. Acute hemodynamic effects in liver surgery patients. Anaesthesist 1997;46:763–770.
74. Cheng DC, Ralph-Edwards A, Mazer CD, Carmichael FJL, Biro GP. The hemodynamic effects of the red cell substitute Hemolink™ (o-rafinose cross-linked human hemoglobin) on vital signs in patients undergoing CABG surgery. Anesthesiology 2000;93:A-180.
75. LaMuraglia GM, O'Hara PJ, Baker WH, Naslund TC, Norris EJ, Li J, Vandermeersch E. The reduction of the allogenic transfusion requirement in aortic surgery with a hemoglobin-based solution. J Vasc Surg 2000;31:299–308.
76. Schubert A, Mascha E, O'Hara JF. Synthetic hemoglobin reduces perioperative blood transfusions in vascular, orthopedic and abdominal surgery. Anesthesiology 2000;93:180.
77. Lamy ML, Daily EK, Brichant JF, Larbuisson RP, Demeyere RH, Vandermeersch EA, Lehot JJ, Parsloe MR, Berridge JC, Sinclair CJ, Baron JF, Przybelski RJ. Randomized trial of diaspirin cross-linked hemoglobin solution as an alternative to blood transfusion after cardiac surgery. The DCLHB cardiac surgery trial collaborative group. Anesthesiology 2000;92:646–656.
78. Carmichael FJ, Ali AC, Campbell JA, Langlois SF, Biro GP, Willan AR, Pierce CH, Greenburg AG. A phase I study of oxidized raffinose cross-linked human hemoglobin. Crit Care Med 2000;28:2283–2292.
79. Patel RP. Biochemical aspects of the reaction of hemoglobin and no: implications for hb-based blood substitutes. Free Rad Biol Med 2000;28:1518–1525.
80. Kim HW, Greenberg AG. Ferrous sulphate scavenging of endothelium derived nitric oxide is a principal mechanism for hemoglobin mediated vasoactivities in isolated rat thoracic aorta. Artf Cells, Blood Sub Immobil Biotechnol 1997;25:121–133.
81. Su D, Roth RI, Levin J. Hemoglobin infusion augments the tumor necrosis factor response to bacterial endotoxin (lipopolysaccharide) in mice. Crit Care Med 1999;27:771–778.
82. Freilich E, Freilich D, Hacker M, Leach L, Patel S, Hebert J. The hemodynamic effects of diaspirin cross-linked hemoglobin in dopamine-resistant endotoxic shock in swine. Art Cells, Blood Sub Immobiliz Biotechnol 2002;30:83–98.
83. Zimmerman JJ. Deciphering the dark side of free hemoglobin in sepsis. Crit Care Med 1999;27:685–686.
84. Winslow RM. Current status of blood substitute research: towards a new paradigm. J Intern Med 2003;253:508–517.
85. Kramer GC. Counterintuitive red blood cell substitute–polyethylene glycol-modified human hemoglobin. Crit Care Med 2003;31:1882–1884.
86. Wettstein R, Tsai AG, Erni D, Winslow RM, Intaglietta M. Resuscitation with polyethylene glycol-modified human hemoglobin improves microcirculatory blood flow and tissue oxygenation after hemorrhagic shock in awake hamsters. Crit Care Med 2003;31:1824–1830.
87. Slanetz PJ, Lee R, Page R, Jacobs EE Jr, LaRaia PJ, Vlahakes GJ. Hemoglobin blood substitutes in extended preoperative autologous blood donation: an experimental study. Surgery 1994;115:246–254.
88. Lee R, Neya K, Svizzero TA, Vlahakes GJ. Limitations of the efficacy of hemoglobin-based oxygen-carrying solutions. J Appl Physiol 1995;79:236–242.
89. Vlahakes GJ. Hemoglobin solutions come of age. Anesthesiology 2000;92:637–638.
90. Levy JH. Hemoglobin-based oxygen-carrying solutions: close but still so far. Anesthesiology 2000;92:639–641.
91. Hughes GS Jr, Francome SF, Antal EJ, Adams WJ, Locker PK, Yancey EP, Jacobs EE Jr. Hematologic effects of a novel hemoglobin-based oxygen carrier in normal male and female subjects. J Lab Clin Med 1995;126:444–451.
92. Schumacher YO, Ashenden M. Doping with artificial oxygen carriers: an update. Sports Med 2004;34:141–150.
93. Levy JH, Goodnough LT, Greilich PE, Parr GV, Stewart RW, Gratz I, Wahr J, Williams J, Comunale ME, Doblar D, Silvay G, Cohen M, Jahr JS, Vlahakes GJ. Polymerized bovine hemoglobin solution as a replacement for allogeneic red blood cell transfusion after cardiac surgery:

results of a randomized, double-blind trial. J Thor Cardiovas Sur 2002;124:35–42.
94. Gonzalez P, Hackney AC, Jones S, Strayhorn D, Hoffman EB, Hughes G, Jacobs EE, Orringer EP. A phase I/II study of polymerized bovine hemoglobin in adult patients with sickle cell disease not in crisis at the time of study. J Investig Med 1997;45:258–264.
95. Feola M, Simoni J, Angelillo R, Luhruma Z, Kabakele M, Manzombi M, Kaluila M. Clinical trial of a hemoglobin based blood substitute in patients with sickle cell anemia. Surg Gynecol Obs 1992;174:379–386.
96. Mullon J, Giacoppe G, Clagett C, McCune D, Dillard T. Transfusions of polymerized bovine hemoglobin in a patient with severe autoimmune hemolytic anemia. N Engl J Med 2000;342:1638–1643.
97. Sloan EP. The clinical trials of diaspirin cross-linked hemoglobin (DCLHB) in severe traumatic hemorrhagic shock: the tale of two continents. Int Care Med 2003;29: 347–349.
98. Gould SA, Moore EE, Hoyt DB, Ness PM, Norris EJ, Carson JL, Hides GA, Freeman IH, DeWoskin R, Moss GS. The life-sustaining capacity of human polymerized hemoglobin when red cells might be unavailable. J Am Coll Surg 195: 445–452; discussion 2002; 452–455.
99. Lieberthal W, Fuhro R, Freedman JE, Toolan G, Loscalzo J, Valeri CR. O-raffinose cross-linking markedly reduces systemic and renal vasoconstrictor effects of unmodified human hemoglobin. J Pharmacol Exper Therap 1999;288:1278–1287.
100. Cheng DC, Mazer CD, Martineau R, Ralph-Edwards A, Karski J, Robblee J, Finegan B, Hall RI, Latimer R, Vuylsteke A. A phase ii dose-response study of hemoglobin raffimer (Hemolink) in elective coronary artery bypass surgery. J Thor Cardiov Sur 2004;127:79–86.
101. Greenburg AG, Kim HW, Hemolink Study Group. Use of an oxygen therapeutic as an adjunct to intraoperative autologous donation to reduce transfusion requirements in patients undergoing coronary artery bypass graft surgery. J Am Coll Surg 198:373–383; discussion 2004; 384–385.
102. Conover CD, Lejeune L, Shum K, Gilbert C, Shorr RG. Physiological effect of polyethylene glycol conjugation on stroma-free bovine hemoglobin in the conscious dog after partial exchange transfusion. Artif Organs 1997;21:369–378.
103. Teicher BA, Ara G, Herbst R, Takeuchi H, Keyes S, Northey D. Peg-hemoglobin: effects on tumor oxygenation and response to chemotherapy. In Vivo 1997;11:301–311.
104. Linberg R, Conover CD, Shum KL, Shorr RG. Increased tissue oxygenation and enhanced radiation sensitivity of solid tumors in rodents following polyethylene glycol conjugated bovine hemoglobin administration. In Vivo 1998;12: 167–173.
105. Serna DL, Powell LL, Kahwaji C, Wallace WC, West J, Cogert G, Smulowitz P, Steward E, Purdy RE, Milliken JC. Cardiac function after eight hour storage by using polyethylene glycol hemoglobin versus crystalloid perfusion. ASAIO J 2000;46:547–552.
106. Meng FT, Zhang WZ, Ma GH, Su ZG. The preparation and characterization of monomethoxypoly(ethylene glycol)-*b*-poly-*dl*-lactide microcapsules containing bovine hemoglobin. Artf Cells, Blood Sub Immobil Biotechnol 2003;31:279–292.
107. Phillips WT, Klipper RW, Awasthi VD, Rudolph AS, Cliff R, Kwasiborski V, Goins BA. Polyethylene glycol-modified liposome-encapsulated hemoglobin: a long circulating red cell substitute. J Pharmacol Exp Ther 1999;288:665–670.

See also BIOCOMPATIBILITY OF MATERIALS; BLOOD COLLECTION AND PROCESSING; BLOOD GAS MEASUREMENTS.

BONDING, ENAMEL.

See RESIN-BASED COMPOSITES.

BONE AND TEETH, PROPERTIES OF

RODERIC LAKES
University of Wisconsin
Madison, Wisconsin

J. LAWRENCE KATZ
University of Missouri
Kansas City, Missouri

INTRODUCTION

Bone has a variety of functions in the body of which some of the most important are structural in nature: protection of vulnerable body parts, support of the body, and to provide muscle attachments. A knowledge of the mechanical and adaptive properties of bone is useful in the design and use of prostheses that replace a bone or a portion of a bone. Mechanical properties of bone are also of interest in trauma biomechanics and in efforts to prevent injury to the body. As for teeth, they also are replaced by artificial materials which are called upon to perform the mechanical functions of the original tooth. This article contains a survey of known properties of bone and teeth and their components collagen and apatite, with an emphasis on bone and its mechanical properties.

A voluminous literature is available dealing with the properties of bone and to a lesser extent of teeth's major constituents: collagen and apatite. Reported properties are often found to differ. Some of the differences arise from the fact that bone and tooth structures are of biological origin and consequently vary depending on the individual and on the part of the body from which the specimen is taken. Other differences are due to experimental technique and variations in environmental conditions during experiments. Of necessity, results presented in this article are selected from a large mass of published reports. The authors have endeavored to select results obtained by good techniques and representative of accepted values. Nevertheless, other results obtained by equally good techniques may be expected to differ somewhat as a result of biological variability. Therefore, it is suggested that additional measurements of the properties of bone and teeth can be found in several of the source books listed in the Bibliography [see (1–10)].

MECHANICAL PROPERTIES OF COMPACT BONE

Compact Bone Structure

The mechanical properties of bone are inseparably related to its structure. Bone tissue is a complex composite material that at different levels of scale exhibits fibrous, porous, and particulate microstructural features (1–5). The following constituents are present in bone: mineral, protein, other organic materials, and fluids such as water. The mineral, principally a carbonated apatite, where the

carbonate group substitutes in part for the phosphate group [$Ca_{10}(PO_4,CO_3)_6(OH)_2$] (5,6). In mature bovine cortical bone (similar to human cortical bone) it occurs as microcrystalline inclusions of plate-like shape (mineralites) of dimensions $\sim 0.7 \times 11 \times 17$ nm (11). However, in young postnatal bovine bone, the mineralites are thicker, shorter, and narrower, [i.e. $2 \times 6 \times 9$ nm (12)]. These mineralite sizes measured by AFM are closest to the actual values as, "AFM yields the full three-dimensional structure of mineralites rather than a projection $\cdots$" thus providing the full shape of each mineralite measured. The other major techniques for measuring mineralite sizes, transmission electron microscopy (TEM) and X-ray diffraction line broadening, each suffer from artifacts that result in increased sizes of some of the dimensions (3).

On the ultrastructural level (nanoscale), the mineral crystallites are in intimate apposition with fibrils of the protein collagen. A, "(d)iagrammatic depiction of the supramolecular packing of collagen molecules in a fibril $\cdots$" plus the possible arrangement of the mineralites within a fibril is given as Fig. 7 in Ref. 12. These fibrils are from 20 to 200 nm in diameter and are organized into fibers that are in turn arranged in bone into lamellae or layers. The fibers in each lamella run longitudinally, spirally, or nearly circumferentially. Moreover, the orientation of these layers are different in alternate lamellae. A micromechanical model for the Young's modulus of bone has been proposed, based on these histological features, however, it has not yet been tested experimentally. The organization of collagen fibrils differs in woven bone and lamellar bone. Mineralized collagen fibrils and isolated crystals from the mid-diaphyses of human fetal femurs were observed with scanning, TEM, and high resolution electron microscopy. The apatite crystals in woven bone are also platelet shaped, similar to mature crystals from lamellar bone. Average crystal dimensions are considerably smaller in woven bone than those of mature crystals in lamellar bone. In diseased bone such as that affected by osteogenesis imperfecta, the apatitic crystals occur in various sizes and shapes; they are oriented and aligned with respect to collagen in a manner that differs from that found in normal calcified tissues.

In compact cortical bone, the lamellae are layers arranged circumferentially around a central canal and form the Haversian system (secondary osteon). Haversian canals typically contain small blood vessels. Haversian bone occurs in the cortices of bones in adult humans (1–3,5) and in the bones of various large animals (1,2). Osteons are roughly cylindrical structures up to $\sim$ 200 μm in diameter. In a long bone, they tend to run approximately parallel to each other and to the bone axis. Figure 1 displays the typical structure of human cortical bone, whereas Fig. 2 displays the typical structure of human cancellous (also known as trabecular or spongy) bone. The large circular or elliptical features in the former figure are the cross-sections of osteons, the concentric layers are lamellae, and the central dark circles are the cross-sections of Haversian canals. The numerous spots among the lamellae are the lacunae, in which the osteocytes, or bone cells live. The lacunae are roughly ellipsoidal and have dimensions of $\sim 10 \times 15 \times 25$ μm. The osteocytes have many thin processes that occupy channels in the bone matrix known as canaliculi; they are too small to be visible in Fig. 1. The mechanical properties of bone depend on its degree of bulk and surface hydration and the details of its structure that depend on variables such as the age, state of health, and level of physical activity of the individual, the location of the bone in the body, as well as the rate and direction at which load is applied to the bone.

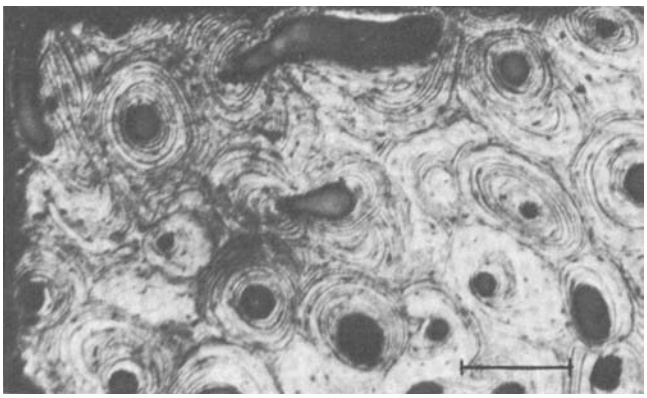


Figure 1. Human Haversian bone. Reflected light micrograph of a specimen cut perpendicular to the bone axis. Scale mark: 200 μm.

The emphasis in this article is on wet skeletal tissues, bone and teeth, from healthy adults, with occasional references to dry tissues for comparison. However, there are many studies of mammalian skeletal tissues as they provide a useful counterpart to the human studies; studies of the properties of bovine bone and teeth are the most numerous in this respect. The structures of both mature bovine cortical and cancellous bone are very similar to those corresponding human tissues Figs. 1 and 2, respectively. Young bovine cortical bone structure is quite different as seen in Fig. 3. This plexiform (or lamellar) bone resorbs and remodels to Haversian bone as the animal matures. Comparison of the young and mature bovine bone properties provides added insight into the importance of structure in determining the mechanical properties of bone. Thus, where appropriate, properties of bovine bone and teeth are included in Tables 1–7.

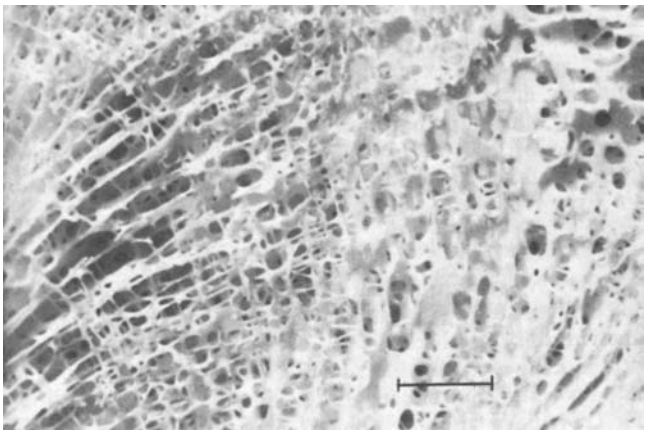


Figure 2. Human cancellous bone from the proximal femur. The marrow has been removed. Scale mark: 5 mm.

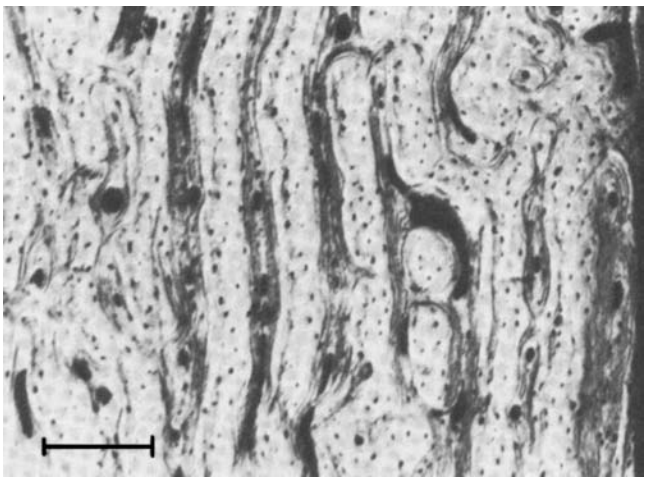


Figure 3. Bovine plexiform bone. Reflected light micrograph of a specimen cut perpendicular to the bone axis. Scale mark: 200 μm.

Elasticity of Compact Bone: Comparison with Teeth and Other Materials

Elastic materials deform under load; elasticity entails reversible behavior in which the material returns to its original configuration after the load is removed. If the load is sufficiently small, there is a linear relationship between stress and strain. For simple tension or compression, the constant of proportionality is referred to as Young's modulus E and for shear or torsion it is the shear modulus G(4,29). The stiffness of human compact bone as quantified by Young's modulus typically lies between 12 and 25 GPa for tension or compression depending on sample orientation and experimental technique (17), (1 GPa = 145,000 lb·in.$^{-2}$). It has been suggested that the tensile and compressive elastic moduli are not equal, but at small strain, the best evidence indicates that tensile and compressive properties are identical (13).

In Table 1, some examples of the elastic moduli of wet human bone and teeth are compared with those of various other materials. The mechanical testing strain rates for bone are faster than those encountered by bones in walking, but are perhaps comparable to those in vigorous activities. They are slower than those that occur during fracture. Compact bone is ∼ 10 times stiffer than most "rigid" polymers, but is about one-tenth as stiff as common metals such as steel. Table 1 also shows the relationship between stiffness and density. This relationship governs the structural efficiency of whole bones. In view of the fact that the skeleton represents ∼ 17% of the weight of the human body, structural efficiency of bones is relevant to their performance in the body. For example, the overall rigidity per unit weight of a bone acting as a short column or as a tensile member is proportional to the modulus to density ratio, E/ρ, of the bone *tissue* of which it is made. By contrast, for a given weight of material, the Euler buckling load of a bone acting as a slender column or the bending stiffness of a bone acting as a beam of constant shape is proportional to E/ρ^2(2,6). Density enters these relations in a different way since the rigidity of a short column depends on its cross-sectional area while the bending rigidity of a beam is governed by its moment of inertia.

A more complete listing of the properties of both human and bovine bone and teeth is given in Table 2, the latter data are included because of the similarity in hierarchical structure and organization of mature bovine compact bone with that of human compact bone. These data were obtained using various techniques, mechanical testing, ultrasonic wave propagation (UWP). In the low megahertz (MHz) region (2–5 MHz), scanning acoustic microscopy (SAM) in the high megahertz region (400–600 MHz), and nanoindentation. The first three techniques also were used to obtain the teeth data. Here too, the properties are strongly dependent on the sample location and orientation, for example, the range in dentin properties over the sample surface obtained by SAM (21), Table 2. Corresponding properties of human trabecular bone volumes and individual trabeculae are given in Table 7.

Because SAM is not as well documented compared to either UWP or mechanical testing (MT) in obtaining elastic properties, a description of the technique follows below.

Scanning Acoustic Microscopy

The development of SAM (30,31); see also (9) enabled the analysis of the biomechanical properties of materials at much higher resolution than was previously achieved using traditional ultrasonic wave propagation techniques.

Table 1. Bone, Teeth, and Other Materials: Stiffness

Material	Young's Modulus E, GPa	Density (ρ), g·cm^{-3}	E/ρ	E/ρ^2
Human compact bone				
longitudinal direction (13)	17	1.8	9.4	5.2
transverse direction	12.5			
Tooth dentin (7,14)	18	2.1	8.6	4.1
Tooth enamel (7,15)	50	2.9	17	6.0
Polyethylene (high density) (4)	0.5	0.95	0.53	0.55
Polymethyl methacrylate (4)	3.0	1.2	2.5	2.2
Steel(structural)	200	7.9	25	3.2
Aluminum	70	2.7	26	9.5
Granite	70	2.8	25	9.1
Concrete	25	2.3	11	4.6
Wood(pine)	11	0.6	18	30

Table 2. Elastic Properties of Wet Human and Bovine Bone and Teeth

Material	Young's Modulus, GPa	References
Human compact bone		
axial direction (femur)	27.7 (U)[a]	16
	17.6 (M)[a]	17
	23.4 (S)[a]	18
(osteons)	22.4 (N)[a]	19
(interstitial lamellae)	25.7 (N)	19
radial direction (femur)	18.9 (U)	16
	12.5 (M)	17
	13.0 (N)	19
Human teeth		
dentin	13.0 (S)	20
	16.4–38.6 (S)	21
enamel	62.7 (S)	20
Bovine compact bone		
axial direction (tibia)	36.0 (M)	2
axial direction (femur)	22.7 (M)	17
axial direction (femur)	21.9 (U)	22
radial direction (tibia)	22.8 (M)	2
radial direction (femur)	10.3 (M)	17
radial direction (femur, average)	13.1 (U)	22
Bovine teeth		
dentin	26.3 (U)	23,24
dentin	25.2 (U)[b]	25
enamel	105.1 (U)	23,24
enamel	97.8 (U)[b]	25

[a] U = Ultrasonics; M = Mechanical Testing; S = Scanning Acoustic Microscopy; N = Nanoindentation.

[b] Average of measurements of several samples.

Table 5. Comparison of Materials: Tensile Strength

Material	Strength σ_{ult}, MPa	Density ρ, g · cm^{-3}	σ_{ult}/ρ
Human femoral compact bone (17),			
longitudinal direction	148	2.0	74
transverse direction	49	2.0	25
Bovine femoral plexiform bone, (13)			
longitudinal direction	167	2.0	83
Tooth dentin (8), average	271	2.1	130
Tooth enamel (8), average	275	2.9	95
Polyethylene (high density)	20–40	0.95	21–42
Poly(methyl methacrylate) (PMMA)	70	1.2	59
Steel(structural)	400	7.8	51
Aluminum(1100-H14)	110	2.7	41
Granite	20	2.8	7.2
Concrete(compression)	28	2.3	12

A significant advantage of SAM is the ability to investigate the properties of internal and subsurface structures in addition to the surface properties of most materials, including those that are optically opaque. Another advantage of special interest for studying biological materials is that a liquid couplant must be used to transmit the acoustic waves from the acoustic lens to the specimen being studied, thus the specimen is kept wet during all measurements. Therefore, fresh tissue specimens can be used as well as embedded specimens. In addition, the use of high quality,

Table 3. Elastic Anisotropy of Bovine and Human Bone

	Tensorial Elastic Moduli (GPa), Determined Ultrasonically Wet Bovine Femur (26)		Dry Human Femur (27)
Elastic constant	Haversian (transverse isotropic)	Plexiform (orthotropic)	Haversian (transverse isotropic)
C_{11}	21.2	22.4	23.4
C_{22}	21.0	25.0	
C_{33}	29.0	35.0	32.5
C_{44}	6.30	8.20	8.71
C_{55}	6.30	7.10	
C_{66}	5.40	6.10	
C_{12}	11.7	14.0	9.06
C_{13}	12.7	15.8	
C_{23}	11.1	13.6	9.11

Table 4. Elastic Anisotropy of Bone[a]

Young's Moduli, GPa		Shear Moduli, GPa		Poisson's Ratios, Dimensionless	
Human	Bovine	Human	Bovine	Human	Bovine
$E = 17.0$[b]	22	$G = 3.6$	5.3	$v = 0.58$	0.30
$E = 11.5$[c]	15	$G = 3.3$	6.3	$v = 0.31$	0.11
$E = 11.5$[d]	12	$G = 3.3$	7.0	$v = 0.31$	0.21

[a] Technical elastic moduli. Wet human femoral bone by mechanical testing (13) and bovine femoral bone by ultrasound (23).

[b] Radial direction.

[c] Circumferential direction.

[d] Longitudinal direction.

Table 6. Elastic Properties of Apatites

Material	Young's modulus, GPa	Bulk modulus, GPa	Shear modulus, GPa	Reference
Hydroxyapatite				
(polycrystalline)	117	88.0	45.5	27
(single crystal, modeling[a])	120	111	45.3	8
Fluoroapatite				
(polycrystalline)	120	94.0	46.4	27
(single crystal, ultrasonics[a])	130	117	49.4	28
Chloroapatite				
(polycrystalline)	94.3	68.5	37.1	27

[a]Calculated from single-crystal elastic constants.

high frequency focusing acoustic lenses permits examination of the elastic properties of biological materials on a microscope scale comparable to the optical histology studies. The heart of the SAM is a spherical lens formed at the interface between a high acoustic velocity solid (e.g., sapphire) and the low acoustic velocity couplant liquid. Due to the high acoustic "refractive index", spherical aberration is negligible and an acoustic beam displaying significant convergence can be obtained. A radio frequency (RF) signal is generated from the transmitter. The acoustic lens is equipped with a piezoelectric transducer that converts the RF signal into an acoustic wave. This signal is then made to converge by the lens and propagates to the specimen through the coupling liquid. When reaching the surface of the specimen, a part of the acoustic wave is reflected back through the lens to the transducer, that, acting now as a receiver, transforms the acoustic signal into an RF signal. The amplitude of the echo reflected back to the lens is a measure of the acoustic reflectivity of the surface of the investigated material at the point in focus. It is proportional to the reflection coefficient, r, which is related to the acoustic impedances of the liquid couplant, Z_1, and the investigated material, Z_2, by the equation r given in Fig. 4. Acoustic impedance, Z, is measured in Rayls and is defined as $Z = \rho v$, where ρ is the material density and v is the velocity of the longitudinal (dilatational) acoustic wave propagating in the direction perpendicular to the surface. The images present the variations in acoustic signals that originate either from the intrinsic acoustic reflectivity of the material surface or through interference occurring between different surface and subsurface reflected signals. In the former case, surface imaging, the acoustic reflectivity variations are a result of the local changes in acoustic impedance (32).

Table 7. Elastic Modulus of Trabecular Bone Volumes and Trabeculae

Trabeculae	Elastic Modulus, GPa	Reference
Human trabecular bone (Trabeculae)	17.4 (S)	18
(longitudinal)	19.4 (N)	19
(transverse)	15.0 (N)	19
Human trabecular bone volumes		
(proximal tibia)	0.445 (M)	3

Figure 5 is a plot of elastic stiffness (GPa) versus reflection coefficient, r, for a wide range of materials. Bone has an acoustic impedance in the neighborhood of $Z = 7.5$ Mrayls, yielding a reflection coefficient in the neighborhood of $r = 0.67$. Of course, Z for bone can vary over a considerable range depending on a number of factors that would affect both the density and structural cohesivity of the specific specimen being measured. Likewise, Z for water will vary depending on the temperature at which the couplant fluid is maintained during the experiment, for example, at 0 °C, $Z(H_2O) = 1.40$ Mrayl; at body temperature, 37 °C, $Z(H_2O) = 1.51$ Mrayl; and at 60 °C, $Z(H_2O) = 1.53$, due mainly to the variations in water's acoustic properties with temperature.

As described above, the high frequency mode at 400 and 600 MHz also has been used to study the *in vitro* micromechanical elastic properties of human trabecular and

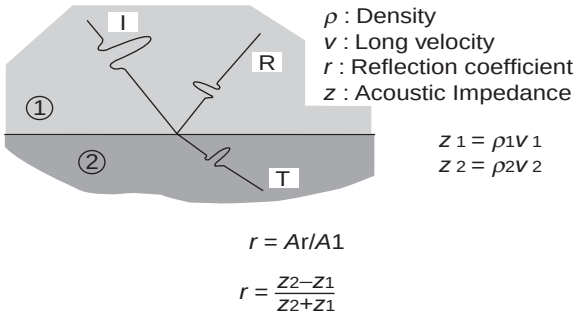


Figure 4. Schematic diagram of the incident, reflected, and transmitted signals at the interface between two materials.

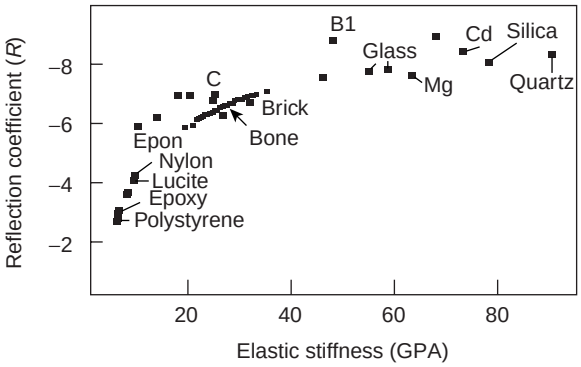


Figure 5. Reflection coefficient, r versus Modulus, E.

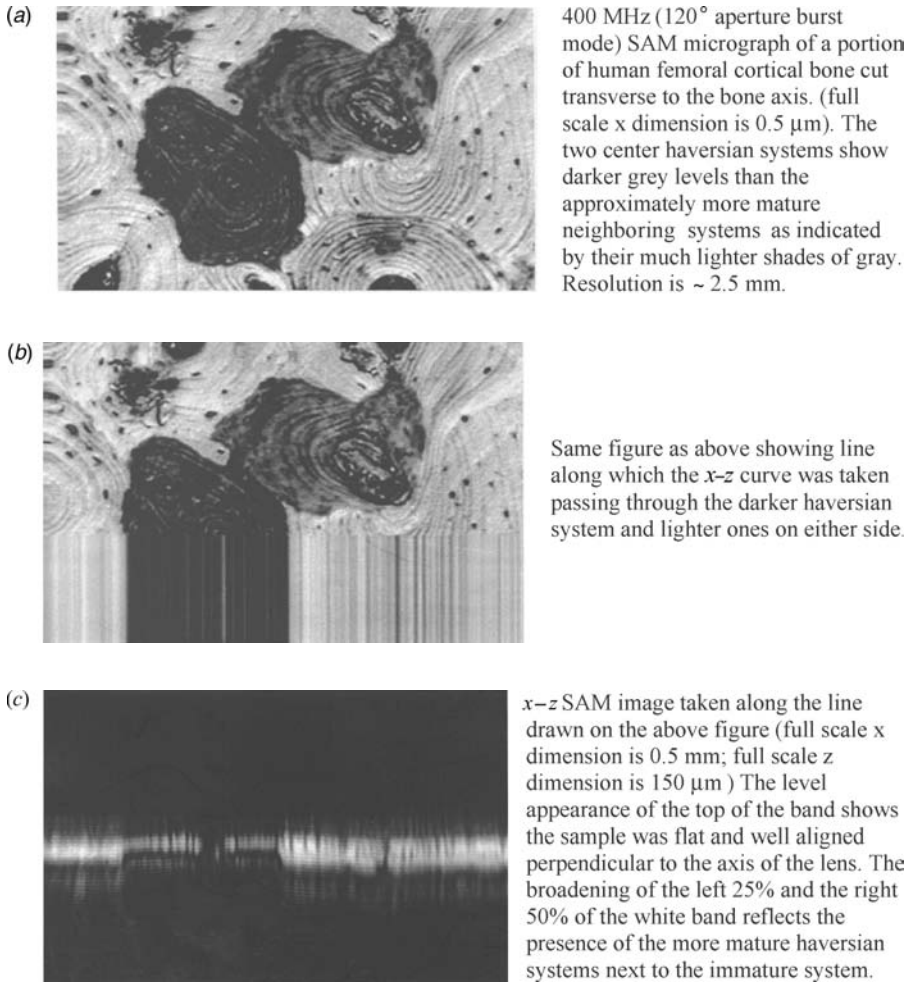


Figure 6. (a) The 400 MHz SAM image of two hypomineralized osteons from human femoral cortical bone. (b) Same area as in part a showing the line along which an x–z interference image was taken. (c) The x–z Interference image taken along the line shown in part b.

compact cortical femoral bone (15,31). In this study, they were able to image the properties of the individual osteonic lamellae at high enough resolution so that three new micromechanical observations were made: (*1*) the outermost lamellae, always appear to be more compliant; (*2*) the outermost lamellae of adjacent abutting osteons appear to have the same acoustic impedance (and thus Young's modulus), even though structurally distinct; and (*3*) adjacent lamellae within an osteon alternate in their acoustic impedance (and thus Young's modulus).

Scanning acoustic microscopy is particularly powerful in providing physical evidence of the possible differences in sound velocity in regions of significant differences in gray level, that is, the darker the gray level, the lower the acoustic impedance, the brighter the gray level, the higher the acoustic impedance, Fig. 6a–c. Figure 6a is a SAM micrograph (400 MHz Burst mode, 120° aperture lens) of human femoral cortical bone (18). The two much darker Haversian systems in the middle of the image are surrounded by the types of Haversian systems (secondary osteons) usually observed in normal bone. The startling difference in gray levels could arise due to out-of-focus artifacts developed in cutting or polishing the specimen. However, performing a x–z curve along the line depicted in Fig. 6b, which goes through the lower, darker Haversian system and the surrounding tissues, yields the image, Fig. 6c; note that the upper level of the broad band is essentially level, indicating that the specimen surface is everywhere level and normal to the acoustic beam. More important, the narrow band and secondary reflection corresponding to the dark Haversian reflects the lower Z and r for the Haversian. A possible explanation for the Haversians is that there was an arrested state of development leading possibly to a hypomineralized area. Unfortunately, we do not have information concerning the possibility of a disease state or a drug modality responsible for these two underdeveloped Haversians that possibly formed just prior to the death of the individual. They are illustrated here to show the ability of the SAM to permit highly sensitive measurements, at high resolution, of variations in Z and r due to remodeling.

In order to convert from reflection coefficient to Young's modulus on the SAM scans of materials of unknown properties, taken with the Olympus UH3 SAM, it is necessary to develop three calibration curves—voltage (V) versus reflection coefficient (r); reflection coefficient (r) versus acoustic impedance (Z); and acoustic impedance (Z) versus Young's modulus (E)—based on using the values of known materials ranging from polymers at the low end of r to metals and ceramics at the high end (18).

The stiffness of compact bone tissue depends on the bone from which it is taken. Fibular bone has a Young's modulus $\sim 18\%$ greater, and tibial bone $\sim 7\%$ greater than that of femoral bone (33). The differences are associated with differences in the histology of the bone tissue. Femoral bone, for example, has a high proportion of osteons that appear light in the polarizing microscope and that have a preponderantly circumferential collagen fiber orientation (34). The histology of a bone is unquestionably related to its function in the body. The relationship has been explored in some detail in the context of bones of very different function in various species of animals (1,2).

Bone is elastically anisotropic, that is, its properties depend on direction. Such behavior is unlike that of steel, aluminum, and most plastics, but is similar to that of wood. Anisotropic properties of bone are shown in Tables 3 and 4. As can be seen from the data in Table 4, human compact femoral bone is ~ 1.5 times as stiff in the longitudinal direction as it is in the transverse directions. The shear modulus G, determined from a torsion test upon a specimen aligned with the bone axis, is ~ 3.3 GPa (13), so that $E/G = 5.15$. Such a small value of G in comparison with Young's modulus E is a further manifestation of the anisotropy of cortical bone. For normal isotropic materials (positive Poisson's ratio), E/G, must lie between 2 and 3 and is typically ~ 2.6. Detailed studies of the anisotropy of bone have been conducted using ultrasonic methods as well as by mechanical testing. The elastic constants given in Table 3 are components of the elastic modulus tensor C_{ijkl} in the following relation between stress σ_{ij} and strain e_{kl} in an anisotropic material. The 3 axis is here assumed to be the bone's longitudinal axis, the 2 direction is circumferential, and the 1 direction is radial.

$$\sigma_{ij} = \sum_{k=1}^{3}\sum_{l=1}^{3} C_{ijkl} e_{kl}$$

In this equation, indexes i and j can have values 1, 2, or 3, so that there are nine components of stress of which six are independent. Since there are six independent components of strain, the number of independent C values is reduced from 81 to 36. Consideration of conservation of mechanical energy further reduces the number of elastic constants to 21, for the least symmetric anisotropic material (7). Materials that have some structural symmetry are described by fewer anisotropic elastic constants. For example, an orthotropic material, (e.g., wood), with three perpendicular planes of symmetry is described by nine constants. A material with transverse isotropic symmetry, (i.e., one that appears the same under an arbirary rotation about an axis), is described by five constants. In crystal physics, the former symmetry is referred to as orthotropic, while the latter is referred to as hexagonal. An isotropic material appears the same under any rotation and has the same properties in any direction. Two independent elastic constants are needed for the description of the elastic behavior of such a material. Equation 1 can be reduced to a more tractable form by a compaction of the counting indices, the so-called Einstein notation, that is, $\sigma_i = C_{ij} e_j$, where $11 \rightarrow 1$, $22 \rightarrow 2$, $33 \rightarrow 3$, $23 \rightarrow 4$, $13 \rightarrow 5$, $12 \rightarrow 6$ (e.g., $C_{1123} \rightarrow C_{14}$). This reduced notation is used in Table 3.

Both structural considerations and experiments indicate that human compact bone has five independent elastic constants and therefore exhibits transverse isotropic symmetry (27,35,36). Results of a different ultrasonic experiment suggest different stiffnesses in the radial and circumferential directions (16). Based on these results, it has been proposed that human compact bone is orthotropic. The difference between the stiffnesses in the radial and circumferential direction is, however, small and has been attributed to the gradient in porosity going from the periostium to the endostium. Thus, for modeling purposes transverse isotropy is the appropriate choice of symmetry (37,38).

Note that the elastic moduli found at high frequencies by ultrasonic methods are greater than those obtained statically or at low frequencies via mechanical testing machines. This is a result of the rate dependence (viscoelasticity) of bone. In the mechanics of whole bones, the principal macroscopic manifestation of cortical bone tissue anisotropy is that the bending rigidity of a bone (the femur) is much greater than its torsion rigidity (39). The degree of anisotropy and the symmetry of bone tissue depends on the species and on the location of the bone in the body. Bovine plexiform bone is stiffer than human Haversian bone, but dry bone is stiffer than the same type of bone when wet. Bovine plexiform bone is orthotropic and has significantly different elastic moduli in the longitudinal, radial, and circumferential directions (40). These properties reflect the laminar architecture of this type of bone, as shown in cross-section in Fig. 2. The scale mark is in the radial direction and is perpendicular to the laminae. Bovine plexiform bone is a type of primary bone that occurs in relativley young cattle. It is often used for experiments as a result of its availability. Canine femoral bone is also orthotropic (16), however, canine mandibular bone and possibly also human mandibular bone, is transversely isotropic (41).

Anisotropic properties of bone may also be expressed in terms of the technical elastic constants, Young's modulus E, shear modulus G, and Poisson's ratio, v, for different directions. Poisson's ratio is minus the transverse strain divided by the axial strain in the direction of stretching or of compressing force. Representative values for dry human femoral bone are shown in Table 3. The 3 direction is the long axis of the bone, the 2 direction is circumferential, and the 1 direction is radial. We note that Poisson's ratio in isotropic materials must be less than one-half (7). In anisotropic materials, larger values are permissible, so that the values reported for bone do not violate any physical law.

Bone mineral density plays an important role in determining all of the elastic properties described above. Under normal physiological and physical conditions, as bone density increases so do the respective elastic properties. However, bone density alone does not always determine fracture risk in pathologies such as osteporosis. A general term in vogue now, 'bone quality', is being used to qualify what is information is necessary to determine when bone is at risk of failure due to reduced density. Structure–property relationships are the key here, especially in understanding when trabecular bone will fail. Even under the conditions of reduced density, the appropropriate structural organization

of the bone may still provide adequate support during normal function. Similarly, good density alone, if associated with a genetic pathology, such as found in osteopetrosis, will not provide adequate prrotection against fracture. The hardness and elastic moduli of osteopetrotic cortical bone are even well below that of osteoporotic bone even though the density of the former is in the normal range (26,42). Indeed, osteopetrotic bone tends to fracture quite readily.

It is clear that for a detailed modeling of the elastic properties of bone, its complex hierarchical structure must be taken into account (37,38,43).

Strength

The ultimate strength of bone tissue refers to the maximum stress the material can withstand before breaking. The tensile strength of human compact bone (17) in a direction parallel to the osteons is about 150 MPa or 21,000 $\mathrm{lb \cdot in.^{-2}}$ (1 MPa = 145 $\mathrm{lb \cdot in.^{-2}}$). As indicated in Table 5, bone is stronger than various plastics, concrete, brick, some metals, (e.g., aluminum), and most woods. Although bone is stronger than aluminum, commonly used aluminum alloys are stronger than bone. Even so, bone has a lower density than aluminum and a much lower density than steel. The criterion for structural strength in beam bending is material strength divided by the 1.5 power of density [3.×]. For bending of plate-shaped structural elements the criterion is material strength divided by the square of the density. The strength to density ratio for bone is greater than that for structural steel. Therefore bone has very favorable properties in comparison with steel and is competitive with aluminum alloys. Bone is anisotropic in its strength as well as in its elastic behavior. In particular, bone is considerably weaker when loaded transversely than when loaded along the osteon direction. Fortunately, bone in the body does not normally experience significant transverse loads. Several investigators have explored the dependence of bone strength upon age. Tensile strength of adult compact bone decreases 4% per decade of age (44) as does its shear strength (45). In other studies, no significant age dependence of strength was found (46,47). More recently, it was found that the tensile strength of femoral bone decreases 2.1% per decade of age while that of tibial bone decreases 1.2% per decade of age (48). The decrease in strength was statistically significant in the case of femoral bone, but not in the case of tibial bone. Both kinds of bone exhibit significant decreases, 6.8% per decade for femur and 8.4% per decade for tibia, in energy absorption to fracture, a measure of toughness. No significant difference between age-matched males and females was found for any mechanical property (48). Tibial bone is stronger in tension than femoral bone by $\sim$ 19% (49,50) and is $\sim$ 4% stronger than fibular bone.

Currey observes that the average stiffness of human and sheep bone increases monotonically with age (in humans from age 2–50), while energy absorption to fracture decreases. The lower mineralization and stiffness and higher toughness seen in young bones is considered adaptive in view of the many falls and bumps experienced by children and young animals (1).

The fracture behavior of a material is not described fully by its yield and ultimate strengths alone; toughness is also important. Toughness is seen as an important characteristic of bone in view of the microcracks that occur in living bone. Fracture of laboratory specimens containing controlled notches provides information concerning the toughness of materials (49,50). For example, the nominal fracture strength of a tensile specimen with an edge notch of length a is $\sigma_{\mathrm{ult}} = K_{1c} a^{-1/2}/Y$ in which Y depends on the specimen geometry and K_{1c} is called the critical stress intensity factor. A large value of K_{1c} results in high strength even in the presence of a notch; K_{1c} is a measure of material toughness.

The critical stress intensity factor K_{1c} for fracture of compact tension specimens of bovine femur bone is from 2.2 to 4.5 $\mathrm{MN \cdot m^{-3/2}}$, and the specific surface energy for fracture is from 390 to 560 $\mathrm{J \cdot m^{-2}}$ (49). The toughness does not, however depend on the sharpness of the controlled notch in the expected way; the significance of this observation is discussed in the section Compact Bone as a Composite Material. As for the age dependence of bone fracture toughness, the energy absorbed to fracture is observed to decrease during childhood and early middle age, and the elastic modulus increases (51). Recently, fracture surfaces from accident victims have been examined microscopically (52). Such surfaces are observed to be much rougher and more intricate than fracture surfaces produced in laboratory specimens of dead bone, which suggests a greater toughness in living bone. The influence of bone viability on its mechanical properties is not well understood. The dependence of bone elastic modulus on viability has been explored in several studies, but the evidence for a difference in elasticity is not compelling.

Currey (1), compares density, modulus, strength and toughness of bones from deer antler, cow thigh, and whale tympanic bulla. As one might expect, modulus increases and toughness decreases with density. Strength is the highest for the femoral bone. Properties vary considerably between these types of bone; the difference is attributed largely to mineralization but partly to histology. The high mineralization of the whale ear bone is responsible for its stiffness, density and brittleness. The stiffness and density are useful given the acoustic function of the ear bone. The brittleness is not a problem since the ear bone is deep within the skull. Conversely, antler is subjected to repeated severe impacts during use, so toughness is beneficial.

Yielding and Plastic Deformation

Wet bone when loaded sufficiently exhibits a yield point, $\sigma_y = 114$ MPa (13). Beyond this critical stress level, the material does not recover upon the release of the load; permanent or plastic deformation has occurred. In bone as in most materials that exhibit yielding, the yield point is approximated by the proportional limit. The proportional limit is the boundary between the linear and nonlinear portions of the stress–strain curve. The mechanism for yield in bone differs from that in metals. In bone, yield occurs as a result of microcracking and other microdamage while in metals, yield results from motion of dislocations. Published reports have not been in agreement as to the amount of plastic deformation in bone. Much of the disagreement has been attributed to the fact that the observed

plastic deformation is highly sensitive to the hydration of the bone. Dry specimens or specimens with dried surfaces or dried and rewetted surfaces behave in a much more brittle manner than those which have been kept fully hydrated during preparation and testing. In the latter case (9), the maximum strain at fracture of bone under tension in the longitudinal direction is 0.031. Yielding of bone has considerable relevance to the function of bone in the body since very much mechanical energy can be absorbed in plastic deformation without fracture. In certain injuries, plasticity in bone can result in residual deformation, which is obvious on a clinical radiograph (53).

Bone Strain *In Vivo*

To ascertain the significance of bone elasticity and strength data in connection with the function of bone in the body, it is desirable to know what levels of stress and strain occur in bones during normal activities and under traumatic conditions. It is possible to make inferences from macroscopic measurements of forces acting on the extremities. The validity of such inferences is rendered uncertain by the fact that muscle forces cannot generally be uniquely determined. It has become possible to determine bone strains explicitly in various animals (54) and in humans (55) by directly cementing foil strain gages to bone surfaces (56). In a human volunteer, maximum strain along the tibia axis was $\sim 3.5 \times 10^{-4}$ during normal walking at $1.4\ \mathrm{m \cdot s^{-1}}$ and 8×10^{-4} during running at $2.2\ \mathrm{m \cdot s^{-1}}$. Strains of similar magnitudes have been observed in animals such as sheep (53). The largest strain magnitude observed in the normal activity of an animal was 3.2×10^{-3} in the tibia of a galloping horse (57). In comparison (5), in tension in the longitudinal direction human bone yields at a strain of 6.7×10^{-3} and fractures at a strain of 0.03. The strain levels observed *in vivo* are significant in view of the fatigue properties of bone. Race horses can experience bone strain exceeding 4.8×10^{-3} at maximum effort (58).

Fatigue

Bone, like other materials, accumulates damage when loaded repeatedly and this damage can lead to fracture at a lower stress than the ultimate strength measured using a single load cycle. The results of *in vitro* mechanical fatigue studies on dead bone suggest that bone may accumulate significant fatigue damage during normal daily activity. Biological bone remodelling is concluded to be essential to the long-term structural integrity of the skeletal system (59). It is notable that dead bone, unlike steel, does not exhibit an endurance limit. The endurance limit is defined as a stress below which the material can withstand an unlimited number of cycles of repetitive load without breaking in fatigue. In the absence of an endurance limit, dead bone subjected to the prolonged cyclic load of daily activity must eventually break. Living bone, by contrast, is able to repair microdamage generated during fatigue.

The resistance of human cortical bone to fatigue fracture is more strongly controlled by strain range than stress range. Fatigue strength shows a weak positive correlation with bone density and with bone modulus and a weak negative correlation with porosity (59). Fatigue strength in uniaxial tests is lower than in bending tests. In immature bone (60), bone fatigue resistance for a given strain range decreases with maturation, while the elastic modulus, density, and ash content increase with maturation. Cracking or fracture of bone due to fatigue manifests itself clinically as 'stress fractures' that can occur in individuals who suddenly increase their level of physical activity (61). In the case of a person in poor physical condition who sustains a fatigue fracture during military training, it is called a "march fracture". Laboratory results for fatigue in bone are not inconsistent with clinical experience with fatigue fractures in athletes and military recruits (62). In particular, a recruit may accumulate in 6 weeks of training a load history equivalent to 100–1000 miles of very rigorous exercise, associated with peak bone strains of 0.002 and a maximum strain range of 0.004, which can be sufficient to precipitate a fatigue fracture (62).

Stress Concentrations

It is common practice in orthopedic surgery to drill holes in bones for the placement of screws. Such holes cause the bone to be weaker than an intact bone, so that it may fracture as a result of less trauma than would ordinarily be required (63,64). This phenomenon may be understood in view of the fact that holes in elastic solids are known to cause a concentration of stress in the vicinity of the hole. According to the theory of elasticity, the stress concentration is not dependent on a decrease in the amount of material available to bear the load; it can be severe even for small holes. Laboratory studies of whole bone fracture have disclosed that a 3 mm diameter hole weakens a tibia by 40% in bending and 12% in torsion (65). A 2.8-or a 3.6-mm hole reduces the strength of dog bones (63) under rapidly applied torsion by a factor of 1.6. It is of interest to compare the observed stress concentration with predicted values based on the theory of elasticity. The predicted stress concentration factor for a hole in a field of shearing stress is 4.0 provided that the hole is small compared with the structure as a whole (66). The discrepancy has been attributed to the fact that intact bone already has stress raisers by virtue of its heterogeneous structure and porosity (63). However, in specimens loaded in the linear elastic range, well below yield or fracture, distributions and concentrations of strain differ from predicted values (67). Similar discrepancies have been reported in manmade fibrous composites without preexisting porosity (68,69). The role of the fibrous architecture of bone in relation to stress concentrations is discussed in the section on composite properties of bone.

Viscoelasticity

Bone exhibits viscoelastic behavior, that is, the stress depends not only on the strain, but also on the time history of the strain. Such behavior can manifest itself as creep, which is a gradual increase in strain under constant stress; stress relaxation, which is a gradual decrease in stress in a specimen held at constant strain; load-rate dependence of the stiffness; attenuation of sonic or ultrasonic waves; or energy dissipation in bone loaded dynamically (10).

Experimental modalities based on each of the above phenomena have been used in the study of bone (70–74). The results have been converted to a common representation via the interrelationships inherent in the linear theory of viscoelasticity, to permit a direct comparison of results (75,76). In the case of tension–compression, there is very significant disagreement among the published results. This disagreement may result from nonlinear viscoelastic behavior not accounted for in the transformation process, or from experimental artifacts. In the case of shear deformation, however, there is good agreement between results obtained in different kinds of experiments. The loss tangent, which is proportional to the ratio of energy dissipated to energy stored in a cycle of deformation, achieves a minimum value of ~ 0.01 at frequencies from 1 to 100 Hz. At lower and higher frequencies, the loss tangent, hence the magnitude of viscoelastic effects is greater (e.g., 0.08 at 1 MHz and at 1 μHz). To compare, the loss tangent of quartz may be $< 10^{-6}$, in metals, from 10^{-4} to 0.01, in hard plastics from 0.01 to 0.1, and in soft polymers, it may attain values > 1. It is notable that the minimum energy dissipation in bone occurs in a frequency range characteristic of load histories during normal activities.

A synopsis of wet bone viscoelastic behavior in shear is presented in Fig. 7. The tan δ attains a broad minimum over the frequency range associated with most bodily activities. Some authors have suggested that the viscoelastic behavior in bone confers a shock- absorbing role. The observed minimum in damping at frequencies of normal activities is not supportive of such an interpretation.

Some authors refer to three element spring dashpot models used in tutorials on viscoelasticity. The behavior of such a model corresponds to a Debye peak in tan δ, also shown in the figure. This corresponds, by Fourier transformation, to a single exponential in the creep or relaxation behavior. The tan δ of bone occupies a much larger region of the frequency domain than a Debye peak, so the spring dashpot model is not appropriate.

Compact Bone as a Composite Material

Early composite models (83–87) for bone were two-phase models involving the mineral and protein phases only. At the ultrastructural level one may imagine the mineral crystals as a particulate reinforcing phase and the surrounding collagen as a matrix phase. Strong arguments were presented that bone cannot be described simply as a compound bar or as a material similar to prestressed concrete. Based on measured Young's moduli of 114 GPa for hydroxyapatite (24) and 1.2 GPa for collagen derived from tendon (83), rigorous upper and lower bounds on the elastic modulus of compact bone were calculated (85,86). Young's modulus of bone lay between these bounds and was less than the value obtained by a simple rule of mixtures approach. Similar comparisons were made for other natural collagen-apatite composites such as dentin and enamel from human teeth.

The wide disparity in the upper and lower bounds in this approach limited its applicability. Later, a composite model of the elastic properties of cortical bone was attempted in order to explain the angular dependence of the elastic properties of both wet and dried bovine cortical bone as determined in an ultrasonic wave propagation experiment (87). However, this modeling failed due to the much stronger decay in the angular dependence of the elastic modulus calculated than was found experimentally. Both this

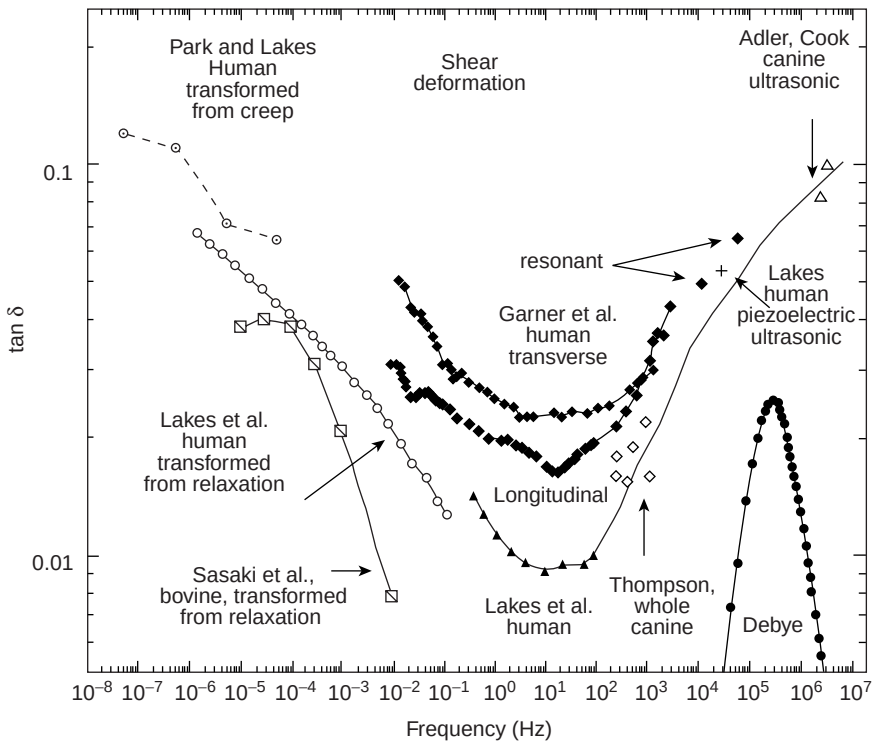


Figure 7. Viscoelastic behavior of wet compact bone in shear. Comparison of results of various authors, adapted from (3). Low frequency tan δ inferred from slope of long-term creep by Park and Lakes, 1986 (77), center circles, ⊙. Results calculated from integration of constitutive equation of Sasaki et al. 1993 (78) for torsion relaxation in bovine bone (slant squares, □). Damping adapted from data of Lakes et al., 1979 (75) for wet human tibial bone at 37°C (Calculated from relaxation, circles, ○; directly measured, ▲). Direct tan δ measurements by Garner et al., 2000 (79) for wet human bone in torsion, diamonds, →. Damping data of Thompson, 1971 (80) for whole dog radius at acoustic frequencies (diamonds, □). Damping of wet human femoral bone by Lakes, 1982 (81) via a piezoelectric ultrasonic oscillator (cross, +). Damping at ultrasonic frequency for canine bone by Adler and Cook (1975) (82) at room temperature (open triangles, △). Theoretical debye peak corresponding to an exponential in the time domain, solid circles •.

attempt (88) and the earlier calculations (84–87) failed to model the properties of compact bone because they did not incorporate the dependence of its properties on its complex hierarchical structure. Inclusion of the hierarchical structural organization of bone was accomplished by adaptation of the hollow fiber composite model (89) so that it resembled the structure of Haversian bone with the osteons viewed as hollow fibers embedded in a matrix (36,37). Subsequently, the same model (35,36,89) was adapted to calculate the viscoelastic properties of bone (90). In recent years, homogenization techniques have been applied to obtain even further improvements In the composite modeling (91,92).

The matrix is the ground substance that comprises the cement lines between osteons and is thought to be principally composed of mucopolysaccharides. The stiffness of the ground substance has not been determined experimentally, but it has been inferred from the composite model in conjunction with experimental data from the ultrasound studies on whole bone (93,94). Based on this calculation, the ground substance was computed to be about one-quarter as stiff as the osteon itself. The view of the ground substance as a compliant interface is also supported by the results of several experimental studies. Localized slippage occurs at the cement lines in specimens of bovine plexiform bone (95) of human Haversian bone (77) subjected to prolonged stress. Under such conditions, the ground substance at the cement lines appears to behave in a viscous manner. Considering the full range of load rates and frequencies, one may view the ground substance as a compliant and highly viscoelastic material. Such a conclusion is perhaps surprising (2) in view of the fact that the ground substance is highly mineralized (96). Nevertheless a view of the ground substance as a compliant interface is supported by further experiments. For example, studies of single osteons and osteon groups in torsion have revealed size effects to occur and the osteon to have a higher effective shear modulus than whole bone (97). Similar size effects were observed in torsional and bending experiments upon larger microsamples (98,99). These latter results have been interpreted in light of a generalized form of elasticity theory, known as Cosserat elasticity, which admits both strain of the material and local rotations of microscopic constituents, for example, the osteons (97–99). A twist per unit area or couple stress can occur in addition to a force per unit area or couple stress. By contrast, classical elasticity (Eq. 1), which describes most ordinary materials, involves only strains and stresses. Cosserat elasticity is likely differ significantly from classical elasticity in its predictions of stress and strain around holes, cracks, and interfaces. For large, whole bones, conventional anisotropic elasticity has been shown to be entirely adequate (39).

The interface between osteons also appears to be important in conferring a measure of fracture toughness upon compact bone. In particular, the crack blunting mechanism of Cook and Gordon (100) confers toughness in fibrous media by the action of a *weak* interface between fibers in blunting a propagating crack (101). Evidence for the role of the cement substance as such a weak interface has been presented by Piekarski (102). Pullout of fibers can result in a large energy absorption in the fracture of fibrous composites (103). This toughening mechanism also appears to be operative in compact bone, as seen in micrographs of pullout of osteons in fractured bone specimens (51,104).

Consideration of bone as a composite material may provide insight regarding various phenomena in the mechanics of bone. In particular, stress concentration around holes is significantly less than the predicted value (63). The strength of notched specimens does not decrease with the sharpness of the notch as expected; instead, the strength is independent of notch sharpness (49). Residual strain around holes in bone does not follow the predictions of classical anisotropic elasticity (98,99). The fatigue life of bone specimens in bending exceeds that of specimens in tension by a factor of several thousand (59). Similar effects have been observed in various manmade fibrous composites such as boron-epoxy and graphite-epoxy (104). Such phenomena are not correctly predicted by elasticity theory, but may be accounted for in the context of structural models, composite theories, or generalized continuum models (77).

Polycrystalline elastic properties and single-crystal elastic constants of apatites that are useful In the modeling of calcified tissues elasticity are given in Table 6.

Mechanical Properties of Cancellous Bone

Cancellous or trabecular bone is a highly porous or cellular form of bone. In a typical long bone, the cortex or exterior of the shaft (diaphysis) and flared ends (metaphysis) is composed of compact bone while the interior, particularly near the articulating ends, is filled with cancellous bone. Cancellous bone may also be found to fill the interior of short bones and flat bones as well as in the interior of bony tuberosities under mucle attachments. The structure of cancellous bone is that of a latticework of bars and plates; typical structure is shown in Fig. 3. The volume fraction of solid material can be from 5 to 70%; the interstices are filled with marrow. The compressive strength σ_{ult}(in MPa) depends very much on the density ρ (in $g \cdot cm^{-3}$) and also varies with the strain rate de/dt (in s^{-1}) as follows (105).

$$\sigma_{ult} = 68(de/dt)^{0.06}\rho^2$$

This relation also models the compressive strength of compact bone (221 MPa, at a density of $\rho. = 1.8\ g \cdot cm^{-3}$). To a certain degree of approximation, both compact and cancellous bone may be mechanically viewed as a single material of variable density (105). Density is not, however, the only determinant of the properties of cancellous bone. The microstructure can vary considerably from one part of the body to another (106). For example, in the vertebras and in the tibia (107), a highly oriented, columnar architecture is observed. This kind of trabecular bone is highly anisotropic: the Young's modulus in the longitudinal direction can exceed that in the transverse direction by more than a factor of 10 (107). By contrast, in regions such as the proximal part of the bovine humerus, the cancellous bone can be essentially isotropic (108). This bone is about twice as strong in compression as in tension. In many ways, cancellous bone (109–111) is similar in its behavior to manmade rigid cellular foams (112). For example, in compression, the stress–strain curves contain a linear elastic region, up to a strain of $\sim$0.05, at which the cell walls bend or compress (112). A plateau

region of almost constant macroscopic stress is associated with elastic buckling, plastic yield, or fracture of the cell walls. The compressive failure of the cancellous bone proceeds at approximately constant stress until the cell walls touch each other; at this point any further compression causes the stress to rise rapidly (112). By contrast, fracture of cancellous bone in tension proceeds abruptly and catastrophically (108). The energy absorption capacity of cancellous bone is consequently much less in tension than it is in compression (109). This suggests that tensile and avulsion fractures of cancellous bone observed clinically are associated with minimal energy absorption, and therefore may be precipitated by relatively minor trauma (109). The elastic modulus E for cancellous bone increases as the square of the density ($E = k\rho^2$) if the structure consists of open cells forming a network of rods (110,111). In closed-cell cancellous structures consisting of plates, the modulus E is proportional to the cube of the density ($E = k\rho^3$). Based on study of micrographs and density maps of femora and vertebras, it is suggested that an open cell structure of rods is found when the solid volume fraction is less than ~ 0.13, while a closed cell plate structure occurs at a density of $>$ 350 $kg \cdot m^{-3}$ corresponding to a relative density or solid volume fraction of 0.20 (110). The relationship between density and structure may not be so straightforward in all situations. As for the highly oriented columnar cancellous bone from the human tibia (107), the modulus E in the longitudinal direction is proportional to the density ($E = k\rho$). Such behavior is anticipated on the basis of an axial compressional mode of deformation of the cell walls, in contrast to the bending mode that is expected in rod and plate structures (112).

It is important to distinguish the difference between the elastic moduli of volumetric samples of trabecular bone that are highly porous, low density structures, thus exhibiting low moduli, generally well below 1.0 GPa as obtained by mecanical testing techniques, and that of individual trabeculae, comparable in structure and density to cortical bone,thus exhibiting much higher moduli (10–20 GPa), as obtained in recent years by nanoindentation (19) and SAM (18) in addition to that obtained by mechanical testing). Values of the elastic moduli of both trabecular volumes and individual trabeculae are given in Table 7.

ADAPTIVE PROPERTIES OF BONE

Phenomenology

The relationship between the mass and form of a bone to the forces applied to it was appreciated by Galileo (113), who is credited with being the first to understand the balance of forces in beam bending and with applying this understanding to the mechanical analysis of bone. Wolff (114) published his seminal 1892 monograph on bone remodeling; the observation that bone is reshaped in response to the forces acting on it is presently referred to as Wolff's law. Cowin, in Chapter 25 of Ref. 3, discussed "The Problems with Wolff's Law ". Many relevant observations regarding the phenomenology of bone remodeling have been compiled and analyzed by Frost (115,116). Salient points are as follows:

1. Remodeling is triggered not by principal stress but by "flexure".
2. Repetitive dynamic loads on bone trigger remodelling; static loads do not.
3. Dynamic flexure causes all affected bone surfaces to drift toward the concavity that arises during the act of dynamic flexure.

These rules are essentially qualitative and they do not deal with underlying causes. A critique of these ideas has been presented by Currey (1,2). Additional aspects of bone remodeling may be found in the clinical literature. For example, after complete removal of a metacarpal and its replacement with graft consisting of a strut of tibial bone, the graft becomes remodeled to resemble a real metacarpal; the graft continues to function after 52 years (117). In the standards of the Swiss Association for Internal Fixation it is pointed out that severe osteoporosis can result from the use of two bone plates in the same region as a result of the greatly reduced stress in the bone (118). Pauwels (119) suggested that as a result of bending stresses the medial and lateral aspects of the femur should be stiffer and stronger than the anterior and posterior aspects. Such a difference has actually been observed (120). Large cyclic stress causes more resorption than large static stress (121). Immobilization of humans causes loss of bone and excretion of calcium and phosphorus (122). Long spaceflights under zero gravity also cause loss of bone (123,124); hypergravity induced by centrifugation strengthens the bones of rats (125,126). Studies of stress-induced remodeling of living bone have been performed *in vitro* (127). Recently, *in vivo* studies in pigs (128) were conducted. In this study, strains were directly measured by strain gages before and after remodeling. Remodeling was induced by removing part of the pigs' ulna so that the radius bore all the load. Initially, the peak strain in the ulna approximately doubled. New bone was added until, after 3 months, the peak strain was about the same as on the normal leg bones. *In vivo* experiments conducted in sheep (129) have disclosed similar results. It is of interest to compare the response time noted in the above experiments with the rate of bone turnover in healthy humans. The life expectancy of an individual osteon in a normal 45 year old man is 15 years and it will have taken 100 days to produce it (130,131).

Remodeling of Haversian bone seems to influence the quantity of bone but not its quality, that is, young's modulus, tensile strength, and composition (132). However, the initial remodeling of primary bone to produce Haversian bone results in a reduction in strength (1,2). As for the influence of the rate of loading on bone remodeling, there is good evidence to suggest that intermittent deformation can produce a marked adaptive response in bone, whereas static deformation has little effect (127). Experiments (133) upon rabbit tibiae bear this out. In the dental field, by contrast, it is accepted that static forces of long duration move teeth in the jawbone. In this connection (134), the direction (as well as the type) of stresses acting on the bone tissue should also be considered. Currey (1,2) points out that the response of different bones in the same skeleton to mechanical loads must differ, otherwise lightly loaded

bones such as the top of the human skull, or the auditory ossicles, would be resorbed.

Failure of bone remodeling to occur normally in certain disease states is of interest: for example, osteropetrotic bone contains few if any viable osteocytes and usually contains a much larger number of microscopic cracks than adjacent living bone (135). This suggests that the osteocytes play a role in detecting and repairing the damage. In senile osteoporosis, bone tissue is removed by the body, often to such an extent that fractures occur during normal activities. Osteoporosis may be referred to as a remodeling error (116).

Some theoretical work, notably by Cowin and others (3,136) has dealt with the problem of formulating Wolff's law in a quantitative fashion. In this theory, constitutive equations are developed, which predict the remodeling response to a given stress. Stability considerations are invoked to obtain some constraints on the parameters in the constitutive equation.

Feedback Mechanisms

Bone remodeling appears to be governed by a feedback system in which the bone cells sense the state of strain in the bone matrix around them and either add or remove bone as needed to maintain the strain within normal limits. The process or processes by which the cells are able to sense the strain and the important aspects of the strain field are presently unknown. Bassett and Becker (137) reported that bone is piezoelectric, that is, that it generates electric fields in response to mechanical stress; they advanced the hypothesis that the piezoelectric effect is the part of the feedback loop by which the cells sense the strain field. This hypothesis obtained support from observations of osteogenesis in response to externally applied electric fields of the same order of magnitude as those generated naturally by stress via the piezoelectric effect. The study of bone bioelectricity has received impetus from observations that externally applied electric or electromagnetic fields stimulate bone growth (138). The electrical hypothesis, while favored by many, has not been proven. Indeed, other investigators have advanced competing hypotheses that involve other mechanisms by which the cells are informed of the state of stress around them.

For example, inhomogeneous deformation at the lamellae may impinge on osteocyte processes and thus trigger the osteocytes to initiate bone formation or remodelling (139). Motion at the cement lines was observed and it was suggested that such motion could act as a passive mechanism by which bone's symmetry axes may become aligned to the direction of time averaged principal stresses (99). Stress on bone may induce flow of fluid in channels, (e.g., canaliculi), and such flow could play a role in the nutrition and waste elimination of osteocytes, which may be significant in bone remodeling (140). In a related vein, theoretical arguments have been presented in support of the hypothesis that bone cells are directly sensitive to hydrostatic pressure transmitted to them from the bone matrix via the tissue fluid (141). Although no experimental test of this direct pressure hypothesis has been published, we observe with interest that direct hydrostatic pressure has been observed to alter the swimming behavior of paramecia, possibly by means of action upon the cell membrane (142). Otter and Salman found that a hydrostatic pressure of 68 atm abolishes the reversing of direction of swimming, 170 atm stops swimming, and 400–500 atm irreversibly damages the cell. We observe that 100 atm corresponds to 1400 psi stress, or in bone, a strain of 0.07%, which is in the normal range of bone strain and 500 atm corresponds to 7000 psi or a strain of 0.35%, well above the normal range of bone strain. Stress in bone also results in temperature differences between osteons (143); the cells may be sensitive to sudden temperature changes during human activity. A mechanochemical hypothesis has been advanced, in which the solubility of calcium may be affected by stress in the bone matrix (144). Strain energy in bone might also influence the energetics of bone mineral nucleation (145). It has also been suggested that remodeling may be initiated in response to microcracks generated by mechanical fatigue of bone (146). In summary, many hypotheses have been proposed for the mechanism by which appropriate cells sense the state of strain in bone, but little or no experimental evidence is available to discriminate among them.

Cellular and Biochemical Aspects of Bone Remodeling

The adaptive response of bone to mechanical stimuli is mediated by living cells. A great deal is known concerning bone cell function and its control by ionic and hormonal factors, but little is known concerning the effect of mechanical strain in bone upon the biochemistry of its cells. Rasmussen and Bordier (147) have presented an extensive review of studies of bone cell physiology. Recently, the biochemical consequences of electrical stimulation of bone have been reported (148). Biochemical steps associated with cell activation are as yet poorly understood, but ion fluxes appear to play a role (149). Cyclic nucleotides mediate the effects of extracellular signals (150) and prostaglandins modulate them (149). Prostaglandin E_2 has been hypothesized to mediate bone resorption in trauma, malignancy, and periodontal disease. This prostaglandin, as well as the cellular constituents cyclic AMP and cyclic GMP, has been found in association with regions of bone stimulated electrically (148).

ELECTRICAL PROPERTIES OF BONE

Fukada and Yasuda (151) first demonstrated that dry bone is piezoelectric in the classic sense, that is, mechanical stress results in electric polarization, the indirect effect; and an appplied electric field causes strain, the converse effect. The piezoelectric properties of bone are of interest in view of their hypothesized role in bone remodeling (137). Wet collagen, however, does not exhibit piezoelectric response. Studies of the dielectric and piezoelectric properties of fully hydrated bone raise some doubt as to whether wet bone is piezoelectric at all at physiological frequencies (152). Piezoelectric effects occur in the kilohertz range, well above the range of physiologically significant frequencies (152). Both the dielectric properties (153) and the piezoelectric properties of bone (154) depend strongly on

frequency. The magnitude of the piezoelectric sensitivity coefficients of bone depends on frequency, on direction of load, and on relative humidity. Values up to 0.7 pC/N have been observed (154), to be compared with 0.7 and 2.3 pC/N for different directions in quartz, and 600 pC/N in some piezoelectric ceramics. It is, however, uncertain whether bone is piezoelectric in the classic sense at the relatively low frequencies which dominate in the normal loading of bone. The streaming potentials examined originally by Anderson and Eriksson (155,156) can result in stress generated potentials at relatively low frequencies even in the presence of dielectric relaxation, but this process is as yet poorly understood.

Potentials observed in bent bone differ from predictions based on the results of experiments performed in compression (157). The piezolectric polarization may consequently depend on the strain gradient (157) as well as on the strain. This piezoelectric theory has been criticized as ad hoc by some authors, however, the idea has some appeal in view of Frost's modeling (115,116) and Currey's suggestion (1,2) that strain gradients may be significant in this regard. The gradient theory is not ad hoc, but can be obtained theoretically from general nonlocality considerations (158). The physical mechanism for such effects is hypothesized to lie in the fibrous architecture of bone (26,78). Theoretical analyses of bone piezoelectricity (159–162) may be relevant to the issue of bone remodeling. Recent thorough studies have explored electromechanical effects in wet and dry bone. They suggest that two different mechanisms are responsible for these effects: Classical piezoelectricity due to the molecular asymmetry of collagen in dry bone, and fluid flow effects, possibly streaming potentials in wet bone (163).

Bone exhibits additional electrical properties which are of interest. For example, the dielectric behavior (e.g., the dynamic complex permittivity) governs the relationship between the applied electric field and the resulting electric polarization and current. Dielectric permittivity of bone has been found to increase dramatically with increasing humidity and decreasing frequency (152,153). For bone under partial hydration conditions, the dielectric permittivity (which determines the capacitance) can exceed 1000 and the dielectric loss tangent (which determines the ratio of conductivity to capacitance) can exceed unity. Both the permittivity and the loss are greater if the electric field is aligned parallel to the bone axis. Bone under conditions of full hydration in saline behaves differently: the behavior of bovine femoral bone is essentially resistive, with very little relaxation (164). The resistivity is $\sim$ 45–48 μm for the longitudinal direction, and three to four times greater in the radial direction. These values are to be compared with a resistivity of 0.72 μm for physiological saline alone. Since the resistivity of fully hydrated bone is $\sim$ 100 times greater than that of bone under 98% relative humidity, it is suggested that at 98% humidity the larger pores are not fully filled with fluid (164).

Compact bone also exhibits a permanent electric polarization as well as pyroelectricity, which is a change of polarization with temperature (165,166). These phenomena are attributed to the polar structure of the collagen molecule; these molecules are oriented in bone. The orientation of permanent polarization has been mapped in various bones and has been correlated with developmental events.

Electrical properties of bone are relevant not only as a hypothesized feedback mechanism for bone remodeling, but also in the context of external electrical stimulation of bone to aid its healing and repair (167,168).

The frequency of electrical stimulation influences its effectiveness (169). A frequency band of 20–30 Hz was found from analysis of strain data. Bone growth could be stimulated more easily in the avian ulna at relatively higher frequencies. Electromagnetic stimuli also prevent bone loss due to disuse as revealed in an isolated canine fibula model (170). Bone density may be maintained and increased by weight lifting, which involves no medical intervention. Indeed, bone mineral content values (as determined with dual photon absorptiometry) in the spines of athletes were extremely high and were closely correlated to the amount of weight lifted during training (171).

BIBLIOGRAPHY

Cited References

1. Currey J. The mechanical adaptations of bones. Princeton: Princeton University Press; 1984.
2. Currey J. Bone Structure and Mechanics. Princeton: Princeton University Press; 2002.
3. Cowin S. Bone Mechanics. 2nd ed. Boca Raton, FL: CRC Press; 2001.
4. Park JB, Lakes RS. Biomaterials. 2nd ed. New York: Plenum; 1992.
5. Hancox NM. Biology of Bone. Cambridge, (MA): Cambridge University Press; 1972.
6. LeGeros RZ. Monographs in Oral Science. Karger: 1991.
7. Sokolnikoff IS. Mathematical theory of elasticity. Krieger; 1983.
8. Ferracane JL. Materials In Dentistry. 2nd ed. Philadelphia: Lippincott, Williams & Wilkins; 2001.
9. Briggs A. Acoustic Microscopy. Oxford: Clarendon Press; 1992.
10. Lakes RS. Viscoelastic Solids. Boca Raton, FL: CRC Press; 1998.
11. Eppell SJ, Tong WL, Katz JL, Kuhn L, Glimcher MJ. Shape and size of isolated bone mineralites measured using atomic force microscopy" J Ortho Res 2001;19:1027–1034.
12. Tong W, Glimcher MJ, Katz JL, Kuhn L, Eppell SJ. Size and shape of Mineralites in young bovine bone measured by atomic force microscopy. Calcif Tiss Inter 2003;72:592–598.
13. Reilly DT, Burstein AH. The elastic and ultimate properties of compact bone tissue. J Biomech 1975;8:393–405.
14. Craig RG, Peyton FA. Elastic and mechanical properties of human dentin. J Dental Res 1958;37:710–718.
15. Craig RG, Peyton FA. Compressive properties of enamel, dental cements, and gold. J Dental Res 1961;40:936–945.
16. Ashman RB, Cowin SC, Van Buskirk WC, Rice JC. A continuous wave technique for the measurement of the elastic properties of cortical bone. J Biomech 17:349–361.
17. Reilly DT, Burstein AH. The mechanical properties of cortical bone. J Bone Jnt Surg 1974;56A:1001–1022.
18. Bumrerraj S, Katz JL. Scanning acoustic microscopy study of human cortical and trabecular bone. Ann Biomed Eng 2001;29:1–9.
19. Rho JY, Pharr GM. Effects of drying on the mechanical poroperties of bovine femur measured by nanoindentation. J Mater Sci, Matyer Med 1999;10:1–4.

20. Kapur R. The use of scanning acoustic microscopy to study the microstructural properties of the dentin/enamel junction, B.S. Project (Katz JL, Advisor), Department of Biomedical Engineering, Case Western Reserve University; 1999.
21. Löst C, Irion KM, Nussle JC. Two-dimensional distribution of sound velocity in ground sections of enamel. Endod Dent Traumatol 1992;8:215–218.
22. Van Buskirk WC, Ashman RB. The elastic moduli of bone. In: Mechanical Properties of Bone, Joint ASME-ASCE Applied Mechanics, Fluids Engineering and Bioengineering Conference. Boulder, CO; 1981.
23. Lees S, Rollins FR, Jr. Anisotropy in hard dental tissues. J Biomech 1972;5:557–566.
24. Lees S, Ahern JM, Leonard M. Parameters influencing the sonic velocity in compact calcified tissues of various species. J Acoust Soc Am 1983;74:28–33.
25. Gilmore RS, Pollack RP, Katz JL. The elastic properties of bovine dentin and enamel. Arch Oral Biol 1970;15:787– 796.
26. Katz JL, Lipson S, Yoon HS, Maharidge R, Meunier A, Christel P. The effects of remodeling on the elastic properties of bone, Calc Tiss Inter 1984;36:S31–S36.
27. Yoon HS, Katz JL. Ultrasonic wave propagation in human cortical bone. II. Measurements of elastic properties and microhardness. J Biomech 1976;9:459–464.
28. Yoon HS, Newnham RE. Elastic Properties of fluorapatite. Am Min 1969;54:1193–1197.
29. Gordon JE. Structures. Penguin; 1983.
30. Lemons RA, Quate CF. Acoustic microscopy-scanning version. Appl Phys Lett 1974;24:163–165.
31. Lemons RA, Quate CF. Acoustic microscopy. Phys Acoust 1979;14:1–92.
32. Katz JL, Meunier A. Scanning acoustic microscope studies of the elastic properties of osteons and osteon lamellae. J Biomech Eng 1993;115:543–548.
33. Evans FG, Bang S. Differences and relationships between the physical properties and the microscopic structure of human femoral, tibial, and fibular bone. Am J Anat 1967;120:79–88.
34. Evans FG, Vincentelli R. Relations of the compressive properties of human cortical bone to histological structure and calcification. J Biomech 1974;7:1–10.
35. Lang SB. Ultrasonic method for measuring elastic coefficients of boneand results on fresh and dried bovine bones. IEEE Trans Biomed Eng, BME 1970;17:101–105.
36. Yoon HS, Katz JL. Ultrasonic wave propagation in human cortical bone, I. Theoretical considerations for hexagonal symmetry, J Biomech 1976;9:407–412.
37. Katz JL. Hierarchical modeling of compact Haversian bone as a fiber reinforced material. In: 1976 Advances in Bioengineering. New York City: ASME; 1976. p 17–18.
38. Katz JL. On the anisotropy of young's modulus of bone. Nature (London) 1980;283:106–107.
39. Huiskes R, Janssen JD, Sloof TJ. A detailed comparison of experimental and theoretical stress analyses of a human femur. In: Mechanical Properties of Bone, Joint ASME-ASCE Applied Mechanics, Fluids Engineering and Bioengineering Conference. Boulder, (CO); 1981.
40. Lipson SF, Katz JL. The relationship between the elastic properties and microstructure of bovine cortical bone. J Biomech 1984;17:231–240.
41. Ashman RB, Rosina G, Cowin SC, Fontenot MG. The bone tissue of the canine mandible is elastically isotropic. J Biomech 1985;18:717–721.
42. Ashman RB, Van Buskirk WC, Cowin SC, Sandbornj PM, Wells MK, Rice JC. The mechanical Properties of immature osteopetrotic bone. Calc Tiss Inter 1985;37:73–76.
43. Lakes RS. Materials with structural hierarchy. Nature (London) 1993;361:511–515.
44. Melick RA, Miller DR. Variations of tensile strength of human cortical bone with age. Clin Sci 1966;30:243–248.
45. Hazama H. Study of the torsional strength of the compact substance of human beings. J Kyoto Pref Med Univ 1956;60:167–184.
46. Evans FG, Lebow M. Regional differences in some of the physical properties of the human femur. J Appl Physiol 1951;3:563–572.
47. Sedlin ED, Hirsch C. Factors affecting the determination of the physical properties of femoral cortical bone. Acta Orthop Scand 1966;37:29–48.
48. Burstein AH, Reilly DT, Martens M. Aging of bone tissue: mechanical properties. J Bone Jnt Surg 1976;58A:82–86.
49. Bonfield W, Datta PK. Fracture toughness of compact bone. J Biomech 1976;9:131–134.
50. Behiri JC, Bonfield W. Crack velocity dependence of longitudinal fracture in bone. J Mat Sci 1980;15:1841–1849.
51. Currey JD. Changes in the impact energy absorption of bone with age. J Biomech 1979;12:459–469.
52. Corondan G, Haworth WL. A fractographic study of human long bone. J Biomech 1986;19:207–218.
53. Carter DR, Spengler DM. Biomechanics of fracture. In: Sumner Smith G, editor. Bone in Cinical Orthopaedics. New York: Saunders; 1982. p 305–332.
54. Lanyon LE. Analysis of surface bone strain in the calcaneus of sheep during normal locomotion. J Biomech 1973;6:41– 69.
55. Lanyon LE, Hampson WGJ, Goodship AE, Shah JS. Bone deformation recorded in vivo from strain gauges attached to the human tibial shaft. Acta Orthop Scand 1975;46:256–268.
56. Caler WE, Carter DR, Harris WH. Techniques for implementing an *in vivo* bone strain gage system. J Biomech 1981;14:503–507.
57. Rubin CT. Skeletal strain and the functional significance of bone architecture. Calcif Tissue Int 1984;36:S11–S18.
58. Nunamaker DM, Butterweck DM, Provost MT. Fatigue fractures in thoroughbred racehorses: relationships with age, peak bone strain, and training. J Orthop Res 1990;8:694.
59. Carter DR, Caler WE, Spengler DM, Frankel VH. Uniaxial fatigue of human cortical bone. The influence of tissue physical characteristics. J Biomech 1981;14:461–470.
60. Keller TS, Lovin JD, Spengler DM, Carter DR. Fatigue of immature baboon cortical bone. J Biomech 1985;18:297–304.
61. Devas MB. Stress fractures. Churchill Livingstone, London; 1975.
62. Carter DR, Caler WE, Spengler DM, Frankel VH. Fatigue behavior of adult cortical bone: the influence of mean strain and strain range. Acta Orthop Scand 1981;52:481–490.
63. Brooks DB, Burstein AH, Frankel VH. The biomechanics of torsional fractures: the stress concentration effect of a drill hole. J Bone Jnt Surg 1970;52A:507–514.
64. Burstein AH, Currey JD, Frankel VH, Heiple KG, Lunseth P, Vessely JC. Bone strength: the effect of screw holes. J Bone Jnt Surg 1972;54A:1143–1156.
65. Laurence M, Freeman MA, Swanson SA. Engineering considerations in the internal fixation of fractures of the tibial shaft. J Bone Jnt Surg 1969;51B:754–768.
66. Timoshenko S, Goodier JM. Theory of elasticity, 3rd ed. New York: McGraw Hill; 1983.
67. Lakes RS, Yang JFC. Concentration of strain around holes in a strip of compact bone, Developments in mechanics, Proceedings of the 18th Midwestern Mechanics Conference. Volume 12, Iowa City; 1983. p 233–237.
68. Awerbuch J, Madhukar S. Notched strength of composite laminates: predictions and experiments, a review. J Reinforced Plast Comp 1985;4:3–159.
69. Daniel IM. Strain and failure analysis in graphite/epoxy plates with cracks. Exper Mech 1978;18:246–252.

70. Smith R, Keiper D. Dynamic measurement of viscoelastic properties of bone. Am J Med Electr 1965;4:156.
71. Currey JD. Anelasticity in bone and echinoderm skeletons. J Exper Biol 1965;43:279.
72. Black J, Korostoff E. Dynamic mechanical properties of viable human cortical bone. J Biomech 1973;16:435.
73. Tennyson RC, Ewert R, Niranjan V. Dynamic viscoelastic response of bone. Exp Mech 1972;12:502.
74. Lugassy AA, Korostoff E. Viscoelastic behavior of bovine femoral cortical bone and sperm whale dentin. In: Research in Dental and Medical Materials. New York: Plenum; 1969.
75. Lakes RS, Katz JL, Sternstein SS. Viscoelastic properties of cortical bone: Part 1: Torsional and biaxial studies. J Biomech 1979;12:657.
76. Lakes RS, Katz JL. Interrelationships among the viscoelastic functions for anisotropic solids: application to calcified tissues and related systems. J Biomech 1974;17:259.
77. Park HC, Lakes RS. Cosserat micromechanics of human bone: strain redistribution by a hydration-sensitive constituent. J Biomech 1986;19:385–397.
78. Sasaki N, Nakayama Y, Yoshikawa M, Enyo A. Stress relaxation function of bone and bone collagen. J Biomech 1993;26:1369–1376.
79. Garner E, Lakes RS, Lee T, Swan C, Brand R. Viscoelastic dissipation in compact bone: implications for stress-induced fluid flow in bone. J Biomech Eng 2000;122:166–172.
80. Thompson G, Experimental studies of lateral and torsional vibration of intact dog radii, [dissertation]. Stanford (CA): Stanford University; 1971.
81. Lakes RS. Dynamical study of couple stress effects in human compact bone. J Biomech Eng 1982;104:6–11.
82. Adler L, Cook CV. Ultrasonic parameters of freshly frozen dog tibia. J Acoust Soc Am 1975;58:1107–1108.
83. Currey JD. Three analogies to explain the mechanical properties of bone. Biorheology 1964;2:1–10.
84. Welch DO. The composite structure of bone and its response to mechanical stress. Recent Adv Eng Sci 1970;5:245–262.
85. Katz JL. Hard tissue as a composite material-I. Bounds on the elastic behavior. J Biomech 1971;4:455–473.
86. Piekarski K. Analysis of bone as a composite material. Inter J Eng Sci 1973;11:557–565.
87. Currey JD. The relationship between the stiffness and the mineral content of bone. J Biomech 1969;2:477.
88. Bonfield W, Grynpas MD. Anisotropy of Young's modulus of bone. Nature (London) 1977;270:453–454.
89. Hashin Z, Rosen BW. The elastic moduli of fiber reinforced materials. J Appl Mech 1964;31:223–2xx.
90. Gottesman T, Hashin Z. Analysis of viscoelastic behavior of bone on the basis of microstructure. J Biomech 1979;13:89–yy.
91. Hogan H. Micromechanics modeling of haversian cortical bone properties. J Biomech 1992;25:549–zzz.
92. Crolet JM, Aoubiza B, Meunier A. Compact bone: Numerical simulation of mechanical characteristics. J Biomech 1993;26:677–aaa.
93. Katz JL, Maharidge RL, Yoon HS. The estimation of interosteonal mechanical properties from a composite model for haversian bone. In: Perren SM, Scheider E, editors. Biomechanics: Current Interdisciplinary Research. Dordrecht: Martinus Injhoff, 1985. p 179–184.
94. Katz JL, Maharidge RL, Yoon HS. Calculation of inter-osteonal mechanical properties for haversian bone based on a hierarchical composite model. Biomechanics Symp. AMD-Vol. 68, FED-Vol. 21. New York City: ASME; 1985. p 33–35.
95. Lakes RS, Saha S. Cement line motion in bone. Science 1979;204:501–503.
96. Frasca P. Scanning-electron microscopy studies of 'ground substance' in the cement lines, resting lines, hypercalcified rings, and reversal lines of human cortical bone, Acta Anatomica 1981;109:115–121.
97. Frasca P, Harper R, Katz JL. Strain and frequency dependence of shear storage modulus for human single osteons and cortical bone microsamples-size and hydration effects. J Biomech 1981;14:679–690.
98. Yang JFC, Lakes RS. Transient study of couple stress effects in human compact bone: Torsion. J Biomech Eng 1981;103:275–279.
99. Yang JFC, Lakes RS. Experimental study of micropolar and couple stress elasticity in compact bone in bending. J Biomech 1982;15:91–98.
100. Cook J, Gordon JE. A mechanism for the control of crack propagation in all-brittle systems. Proc R Soc London 1964;A282:508–520.
101. Kelly A. The strengthening of metals by dispersed particles. Proc R Soc London 1964;A282:63–79.
102. Piekarski K. Fracture of Bone. J Appl Phys 1970;41:215–223.
103. Kelly A. Strong solids. London: Oxford University Press; 1966.
104. Wright TM, Barnett DM, Hayes WC. Residual stresses in bone. Volume 3, Recent Advances in Engineering Science. Boston: Scientific Publishers; 1977. p 25–32, Proceedings, 10th meeting, Society of Engineering Science, NC: Raleigh; 1973.
105. Carter DR, Hayes WC. Bone compressive strength: the influence of density and strain rate. Science 1976;194: 1174–1176.
106. Dyson ED, Jackson CK, Whitehouse WJ. Scanning electron microscope studies of human trabecular bone. Nature (London) 1970;225:957–959.
107. Williams JL, Lewis JL. Properties and an anisotropic model of cancellous bone from the proximal tibial epiphysis. J Biomech Eng 1982;104:50–56.
108. Kaplan S, Hayes WC, Stone JL, Beaupre GS. Tensile strength of bovine trabecular bone. J Biomech 1985;18: 723–727.
109. Carter DR, Schwab GH, Spengler DM. Tensile fracture of cancellous bone. Acta Orthop, Scand 1980;51:733–741.
110. Gibson LJ. The mechanical behaviour of cancellous bone. J Biomech 1985;18:317–328.
111. Carter DR, Hayes WC. The compressive behaviour of bone as a two-phase porous structure. J Bone Jnt Surg 1977;59A:954–962.
112. Gibson LJ, Ashby MF. The mechanics of three-dimensional cellular materials. Proc R Soc London 1982;A382:25–42.
113. Galilei G, Discorsi E. Dimostrazioni Matematiche intorna a due nuove Scienze. 1638, Translated by H Crew, A deSalvio, editors. New York: Macmillan; pp 158–172. 1914. p 118–134.
114. Wolff J. Das Gesetz der Transformation der Knochen. Berlin: Hirschwald; 1892.
115. Frost HM. Bone remodelling and its relation to metabolic bone diseases. Springfield, IL: C Thomas; 1973.
116. Frost HM. Bone modelling and skeletal modelling errors. Springfield, IL: C Thomas; 1973.
117. Nathan PA, Fowler A. Remodeling of a metacarpal bone graft in a child. J Bone Jnt Surg 1976;58A:719–722.
118. Muller ME, Allgouer M, Willeneger H. Manual of Internal Fixation—Technique Recommended by the AO Group. Springer Verlag; 1970.
119. Pauwels F. Die Bedeutung des Bauprinzipien des Stütz, und Bewegensapparatens für, die Beanspruchung der Rohrenknochen. Z Anat Entwicklungs 1948;114:129–166.

120. Amtmann E. The distribution of the breaking strength in the femur. J Biomech 1968;1:271–277.
121. Seirig A, Kempko W. Behavior of *in vivo* bone under cyclic loading. J Biomech 1969;2:455–461.
122. Dietrick JE, Whedon G, Shorr E. Effects of immobilization upon various metabolic and physiological functions of bone. Am Jnl Med 1948;4:3–36.
123. Mack PB, La Chance PL. Effects of recumbency and space flight on bone density. Am J Clin Nutrition 1967;20:194–205.
124. Morey ER, Baylink DK. Inhibition of bone formation during space flight. Science 1978;201:1138–1141.
125. Wunder CC, Briney SR, Skangstad CA. Growth of mouse femurs during chronic centrifugation. Nature (London) 1960;188:151–152.
126. Wunder CC, Cook RM, Welch RC, Glade R, Fleming BP. Femur bending properties as influenced by gravity: I. ultimate load and moment for 3-G rats. Aviat Space Environ Med 1977;48:339–346.
127. Glucksmann A. Studies of bone mechanics *in vitro* I-Influence of pressure on orientation of structure. Anat Rec 1938;72:97–115.
128. Goodship AE, Lanyon LE, McFie M. Functional adaptation of bone to increased stress. J Bone Jnt Surg 1979;61A:539–546.
129. Hall BK. Developmental and cellular skeletal biology. New York: Academic Press, 1978.
130. Sumner-Smith G. Bone in clinical orthopaedics. New York: W. B. Saunders; 1982.
131. Lanyon LE, Magee PT, Bagott DG. The relationship of the functional stress and strain to the process of bone remodelling: an experimental study on the sheep radius. J Biomech 1979;12:593–600.
132. Woo SLY, Kuei SC, Amiel D, Gomez MA, Hayes WC, White FC, Akeson WH. The effect of prolonged physical training on the properties of long bone: a study of Wolff's law. J Bone Jnt Surg 1981;63A:780–787.
133. Liskova M, Hert J. Reaction of bone to mechanical stimuli, part 2, periosteal and endosteal reaction of the tibial diaphysis in rabbit to intermittent loading. Folia Morpholog 1971;19:310–317.
134. Wright KWJ, Yettram AL. An analytical investigation into possible mechanical causes of bone remodelling. J Biomed Eng (England) 1979;1:41–49.
135. Frost HM. Osteocyte death in vivo. J Bone Jnt Surg 1960;42A:138–143.
136. Cowin SC, Hegedus DH. Bone Remodeling I: Theory of adaptive elasticity. J Elasticity 1976;6:313–326.
137. Bassett CAL, Becker RO. Generation of electric potentials in bone in response to mechanical stress. Science 1962;137: 1063–1064.
138. Spadaro JA. Electrically stimulated bone growth in animals and man. Clin Orthopaed 1977;122:325–332.
139. Tischendorf F. Das Verhalten der Haversschen Systeme bei Belastung. Arch Entwicklungsmech Org 1951;145:318–332.
140. Piekarski K, Munro M. Transport mechanism operating between blood supply and osteocytes in long bones. Nature (London) 1977;269:80–82.
141. Jendrucko RJ, Hyman WA, Newell PH, Chakraborty BK. Theoretical evidence for the generation of high pressure in bone cells. J Biomech 1976;9:87–91.
142. Otter T, Salman ED. Hydrostatic pressure reversibly blocks membrane control of ciliary motion in paramecium. Science 1979;206:358–361.
143. Lakes RS, Katz JL. Viscoelastic properties and behavior of cortical bone, Part II, relaxation mechanisms. J Biomech 1979;12:689–698.
144. Justus R, Luft JH. A mechanicochemical hypothesis for bone remodelling induced by mechanical stress. Calc Tiss Res 1970;5:222–235.
145. Jendrucko RJ. Energetics of hydroxyapatite nucleation in bone, Proc 30th ACEMB, Los Angeles: 1977.
146. Martin RB, Burr DB. A hypothetical mechanism for the stimulation of osteonal remodelling by fatigue damage. J Biomech 1982;15:137–139.
147. Rasmussen H, Bordier P. The Physiological and Cellular Basis of Metabolic Bone Disease. Williams & Wilkins; 1974.
148. Davidovitch Z, Furst L, Shanfield JL, Montgomery PC, Kelischeck S, Laster L, Korostoff E. Biochemical mediators of electrical stimulation of bone cells. 35th ACEMB. Philadelphia: Sept. 1982. p 217.
149. Rodan GA, Bourret LA, Norton LA. DNA synthesis in cartilage cells is stimulated by oscillating electric fields. Science 1978;199:690–692.
150. Sutherland J, Rall M. The relation of adenosine 3′:5′-phosphate and phosphorylase to the actions of catecholamines and other hormones. Pharmacol Rev 1977;12:265–299.
151. Fukada E, Yasuda I. On the piezoelectric effect of bone. J Phys Soc J 1957;12:1158–1162.
152. Reinish G. Piezoelectric properties of bone as functions of moisture content. Nature (London) 1975;253:626–627.
153. Lakes RS, Katz JL. Dielectric relaxation in cortical bone. J Appl Phys 1977;48:808–811.
154. Burr AJ. Measurements of the dynamic piezoelectric properties of bone as a function of temperature and humidity. J Biomech 1976;1:495–507.
155. Anderson JC, Eriksson C. Electrical properties of wet collagen. Nature (London) 1968;218:167–169.
156. Anderson JC, Eriksson C. Piezoelectric properties of dry and wet bone. Nature (London) 1970;227:491–492.
157. Williams WS. Sources of piezoelectricity in tendon and bone. CRC Crit Rev Bioeng 1974;2:95–117.
158. Lakes RS. The role of gradient effects in the piezoelectricity of bone. IEEE Trans Biomed Eng 1980;BME27:282–283.
159. Korostoff E. Stress generated potentials in bone: relationship to piezoelectricity of collagen. J Biomech 1979;10:41–44.
160. Korostoff E. A Linear piezoelectric Model for characterizing stress generated potentials in bone. J Biomech 1979;12:335–347.
161. Gjelsvik A. Bone remodeling and piezoelectricity II. J Biomech 1973;6:187–193.
162. Guzelsu N. A piezoelectric model for dry bone and tissue. J Biomech 1978;11:257–267.
163. Johnson M, Chakkalakal D, Harper RA, Katz JL. Comparison of the electromechanical effects in wet and dry bone. J Biomech 1980;13:437–442.
164. Chakkalakal DA, Johnson MW, Harper RA, Katz JL. Dielectric properties of fluid saturated bone. IEEE Trans Biomed Eng 1980;BME-27:95–100.
165. Athenstaedt H. Permanent electric polarization and pyroelectric behavior of the vertebrate skeleton. VI, the appendicular skeleton of man. Z Anat Entwickl Gesch 1970;131:21–30.
166. Lang SB. Pyroelectric effect in bone and tendon. Nature (London) 1966;212:704–705.
167. Bassett CAL, Pilla AA, Pawluk RJ. A non-operative salvage of surgically resistant pseudarthrosis and non-unions by pulsing electromagnetic fields: a preliminary report. Clinical Orthop 1977;124:128–143.
168. Brighton CT, Friedenberg ZB, Mitchell EI, Booth RE. Treatment of non-union with constant direct current. Clinical Orthop 1977;124:106–123.

169. McLeod KJ, Rubin CT. The effect of low frequency electrical fields on osteoporosis. J Bone Joint Surg 1992;74-A:920–929.
170. Skerry TM, Pead MJ, Lanyon LE. Modulation of bone loss during disuse by pulsed electromagnetic fields. J Orthop Res 1991;9:600–608.
171. Granhed H, Jonson R, Hansson T. The loads on the lumbar spine during extreme weight lifting., Spine 1987;12(2): 146–149.

Reference List

Gilmore RS, Katz JL. Elastic properties of apatites. J Mat Sci 1982;17:1131–1141.
Katz JL, Ukraincik K. On the anisotropic elastic properties of hydroxyapatite. J Biomech 1971;4:221–227.
Lakes RS, Katz JL. Viscoelastic properties of bone. In Hastings G, Ducheyne P, editors. Natural and Living Biomaterials. Washington, (DC): CRC Press; 1984.
Ortmann R, Perkins JP. Stimulation of adenosine 3′:5′ monophospate formation by prostaglandins in human astrocytoma cells. J Biol Chem 1977;252:6019–6025.

See also BIOMATERIALS FOR DENTISTRY; BONE CEMENT, ACRYLIC; TOOTH AND JAW, BIOMECHANICS OF.

BONE CEMENT, ACRYLIC

CHAODI LI
YAN ZHOU
University of Notre Dame
Notre Dame, Indiana

INTRODUCTION

Many years of intensive research by Rohm led to the development of poly(methyl methacrylate) (PMMA), the basis of bone cement, in 1934 (1,2). This polymer was reportedly first used to close cranial defects in monkeys in a medical application in the late 1930s (2). The use of acrylic bone cement in orthopedic surgery was first advocated in 1951 by Kiaer and Jansen. It was applied as pure anchoring material by fixing acrylic glass caps on the femoral head after removing the cartilage (2–4). At that time, the femoral components in total hip replacements were still implanted by simply press-fitting the prosthesis tightly into the prepared intramedullary (marrow) canal of the femur. Patients were confined to bed for a relatively long period of time after surgery and often experienced pain later from loosening of the implant. In 1958, Sir John Charnley first applied the self-curing bone cement for the fixation of artificial joints (4). In this surgery, the cement filled the free space between the prosthesis and the bone (2). Bone cement served as a mechanical interlock between the metallic prosthesis and the bone and it has been found to be an appropriate material to transfer the load consistently (5). In 1969, bone cement was approved for general use within the United States and since then the number of total hip and knee replacements has increased dramatically (3). Currently, it is the only material used for anchoring cemented arthroplasties to the contiguous bones (6). More than one-half million hip replacement surgeries are performed every year worldwide and 70% of the surgeries are performed using bone cements (7–10).

Bone cement is also extensively used in the fixation of pathological fractures, spinal surgery, and neurosurgery (11). Every year, osteoporosis results in $>310{,}000$ fractures in the United Kingdom alone (12). Vertebral compression fractures occur in 20% of people over the age of 70 years and in 16% of postmenopausal women (12). Although traditional conservative techniques, such as bed rest and the prescription of analgesics may be successful in a proportion of cases, a significant number of sufferers remain in long-term pain. Vertebroplasty is now being used extensively for vertebral compression fracture treatments (12,13). This technique entails the percutaneous injection of bone cement into the fractured vertebra in attempts to stabilize the fracture and reduce pain. Studies have reported excellent pain relief and improved function in most patients (12,13). Bone cements have also been applied as an adjunct to internal fixation for treating fractures. Bone cement fills voids in bone, thereby reducing the need for bone grafts, and may improve the holding strength around the devices in osteoporotic bone. Recently, the technique of PMMA bone cement injection into the osteoporotic proximal femur was proposed (14). Results indicated that cement injection increased the peak fracture load >80 and 20% in the simulated fall and one single limb stance configurations, respectively (14). This technique could become a treatment option to solve the problems with osteoporotic hip fractures in patients at risk. Such treatment may hold a significant impact on the economy and human health as hip fractures result in $\sim 300{,}000$ hospital admissions annually in the United States with an estimated $9 billion in direct medical costs (14). Therefore, bone cements are significantly valuable for biomedical applications.

In many cases, the bone cemented implants perform satisfactorily for years. However, failure of the implants may be linked to bone cement properties. It is well recognized that there are a number of drawbacks that exist with the usage of bone cement, significant ones including its poor mechanical properties and the potential necrosis of bone tissues (6,15,16). For example, bone cement has been called the "weak link" in the prosthesis system and long-term loosening of the prosthesis has been attributed to its mechanical disintegration (10). It is worth pointing out that aseptic loosening of a cemented arthroplasty is a multifactorial phenomenon involving interfacial failure, bone failure, bone remodeling, and cement failure (6). It is not firmly established whether the drawbacks of bone cements contribute to the initiation or are the consequence of aseptic loosening of the implant (6). In spite of the many drawbacks of bone cement, the survival probabilities of recently cemented joint replacements are still high. Currently, cemented total hip arthroplasty shows survival rates of 90% at 15 years and 80–85% at 20 years (10,15,17–21). Increasing bone cement properties along with improving cementing techniques and implantation methods may further extend cemented implant longevity.

In recognizing the drawbacks of bone cement usage, there have been considerable efforts to secure the implants without using cement. Early noncemented prostheses were the press-fitted or screwed-in types (10). The press-fit

techniques must lead to good initial fixation; otherwise, mobility will occur and this will facilitate wear and/or formation of fibrous tissue. Such techniques have limited successes due to a number of reasons, including bone resorption, geometrical constraints, and increased operational fractures. Noncemented porous coated prostheses were also introduced to provoke bony ingrowth for improved fixation. The noncemented prostheses are used predominantly in younger patients (< 60 years), where it is assumed that these prostheses eventually simplify a revision operation (10). Bone ingrowth strategies are not appropriate for older patients. Research indicates that some of the noncemented devices have failed, but others are doing well in the mid- to long- term (10). The clinical applications of cementless total hip and knee arthroplasties induce a new set of problems, including perioprosthetic osteolysis, high thigh pain, and failure of the bone–implant interface (6). Thus, there is currently a resurgence of interest in bone cement (6). Definite conclusions about the ultimate clinical performances of cement or cementless fixation devices require longer term studies (10). The debate is still open as to which fixation method is best. The development of cementless modes of fixation is an area of active research and clinical practice; however, acrylic bone cement continues to be the most commonly used nonmetallic implant material by orthopedic surgeons (6,11).

The literature on bone cement is voluminous with respect to its material development, manufacturing, thermal response, chemical and biological effects to bone, and its short- and long-term physical and mechanical properties (static, fatigue, etc.). This article discusses some of these aspects, especially on cement's material characteristics. For more ample assessment, the readers should refer to other excellent comprehensive reviews [e.g., Krause (3,11), Lewis (6,22), Kuhn (2), Saha (23), Kenny (5), Deb (24), Hasenwinkel (15), Serbetci and Hasirci (16)]. The article is arranged into several sections. It begins with the description of bone cement compositions, followed by a section on cement setting procedure and cementing technique. Then, the thermal and volumetric change effects are discussed. The final section focuses on the physical and mechanical properties of cement. Finally, the article ends with a brief summary.

CEMENT COMPOSITIONS

There are a number of bone cements [> 60 types (2)] available to the orthopedic community. Some popular cements currently available in the United States are given in Table 1 (2,6,11). Most of the present commercial bone cements have similar compositions (Table 2) (2,11,25). The presently used form of bone cement is predominantly the same as the one Sir John Charnley introduced. All acrylic bone cements on the market are chemically based on the identical basic substance: methyl methacrylate (MMA) (2). Pure MMA exhibits a shrinkage of ~21% during polymerization (2), and the polymerization temperature can increase to 100–120°C. Such a high shrinkage is intolerable for use in bone cement (2). For this reason, bone cements are offered as two-component systems in the marketplace: a prepolymerized powder and a liquid mm monomer (2). The MMA in aqueous suspension is prepolymerized in easily cooled reaction boilers. The polymer, obtained in the form of tiny balls (< 150 μm), is easily dissolved in the MMA. By using the prepolymerized polymer powder, both the shrinkage of the sample and the temperature of the reaction can be considerably decreased. In most bone cements on the market, the mixing ratio is two to three parts powder to one part monomer. This reduces the shrinkage and the generation of heat by at least two-thirds, as only the monomer is responsible for these reaction symptoms (2).

Usually, the solid part of bone cement consists of prepolymerized PMMA beads ranging in size from 1 to 150 μm. Other kinds of polymers sometimes are added, including poly(ethyl acrylate), poly(methyl acrylate), poly (styrene), and poly(butyl methacrylate). Free radicals of benzoyl peroxide (BPO) are present within the beads as remnants from the emulsion polymerization process (the process by which most of the beads are manufactured). An additional amount of BPO is mixed with the solid to obtain 1–2.5% by weight benzoyl peroxide. In addition, bone cements generally contain ~10–15% by weight barium sulfate, zirconia, or other additive. The presence of barium sulfate or zirconium dioxide in the powder is necessary for

Table 1. List of Popular Commercial Bone Cements Currently Available in the United States[a]

Bone Cements	Manufacturer or Distributor
CMW1	Depuy, Warsaw, IN
CMW3	Depuy, Warsaw, IN
Palacos R	Smith and Nephew, Memphis, TN
Simplex P	Stryker Howmedica Osteonics, Rutherford, NJ
Osteobond	Zimmer, Warsaw, IN
Zimmer dough-type	Zimmer, Warsaw, IN

[a] See Refs. 2,6, and 11.

Table 2. Compositions of Commercial Bone Cements Currently Available in the United States[a,b]

Liquid component	
Methylmethcrylate (monomer)/ Butylmethacrylate (binding agent)	~98
Activator/Co-initiator: Dimethyl-paratoluidine	0.4–2.75
Stabilizer/inhibitor/radical catcher: Hydroquinone, Ascorbic acid	15–75 ppm
Coloring: Chlorophyll	267 ppm
Powder	
Poly(methyl methacrylate)/copolymer	~90
Initiator: Benzoyl peroxide	0.5–3
Opacifier: $BaSO_4$ or ZrO_2	10–15%
Coloring: Chlorophyllin	200 ppm
Antibiotics: Gentamicin, erythromycin, and colistin	

[a] Compositions are in percent (w/w) except where stated otherwise.
[b] See Refs. 2,6,11 and 25.

a clinical reason. The original bone cements, which did not contain radiopacifiers could not be visualized on radiographs. The main components in the liquid phase are MMA, and, in some bone cements, other esters of acrylic acid or methacrylic acid, one or more amines (as activators for the formation of radicals), a stabilizer and, possibly, a colorant (2). The amine in the bone cement, *N*,*N*-dimethyl-*p*-touluidine (DMPT), acts as an accelerator. The liquid component also consists of 50–100 ppm hydroquinone, which inhibits the polymerization reaction within the monomer and allows for storage of the liquid component. Some cement also contains chlorophyll that gives it a green color. This allows better distinction from body tissues during surgery.

Prosthesis-related infection is described as a devastating failure scenario of a cemented orthopedic implant. Infectious complications of a cemented prosthesis lead to a deterioration of function and increase pain. Buchholz and Engelbrecht first reported on the possibilities of mixing antibiotics in bone cement in 1970 (17). They considered gentamicin sulfate to be the antibiotic of choice because of its wide-spectrum antimicrobial activity, its excellent water solubility, its thermal stability and its low allergenicity. Apart from gentamicin, other antibiotics have also been used as an additive to bone cement. The combination of erythromycin and colistin is an example that made it to a commercial product (17). In most cases, the manufacturers make their antibiotic cements by simply mixing antibiotic to a plain cement version they have (2,15,16).

Currently, bone cement fracture is regarded as a major factor in the mechanical failure of implant fixation (26–28). It is directly related to the mechanical properties of the cement, especially the resistance to fracture of the cement in the mantle at the cement–prosthesis interface or the cement–bone interface. Thus, many investigators have attempted to incorporate second phase materials including polyethylene (29), hydroxyapatite (30), PMMA (31), Kevlar (32,33), carbon (34–36), titanium (37), and steel (38–40) to improve the fatigue properties and fracture toughness of the PMMA. Most of the results regarding the properties of these composite materials have been encouraging; however, the biocompatibility issues regarding some of these fibers are as yet unresolved (6). Consequently, none of the commercial bone cements have incorporated these fibers in cements on the market.

CEMENT SETTING AND CEMENTING TECHNIQUE

Most structural materials are fabricated under controlled conditions at a factory, then transported and assembled on site. Bone cement is one of the few structural materials that are created *in situ*. The surgeon prepares the bone cement directly at the operation table according to the manufacturer's instructions. All of the cements are supplied in sets of the polymer powder and the monomer liquid components packed in two separate containers within a package. At the time of surgery, the liquid monomer and powder are mixed, the DMPT reacts with the BPO to generate free radicals, which in turn are used in the additional polymerization of the MMA monomers to form PMMA. Polymer chains from the PMMA become available for free radical polymerization and entanglements of these chains with newly formed chains lead to an intimate connection between the newly formed PMMA with what was already present. The resulting product is a doughy mixture that later polymerizes to a hard and brittle substance.

Table 3. Four Phases of Bone Cement Polymerization Process[a]

Phases	Time Duration, min	Characteristics
I. Mixing	1–2	Wetting Cement relatively liquid (low viscous)
II. Waiting	2–3	Swelling + polymerization Increase of viscosity Polymer chains, less movable Sticky dough
III. Working	5–8	Chain propagation Reduced movability Increase of viscosity Heat generation
IV. Setting	2–6	Chain growth finished No movability Cement hardened High temperature

[a]See Ref. 2.

The curing process of acrylic PMMA bone cement can be divided into four basis steps (2): the mixing, waiting, working, and hardening phase. The characteristics of these phases, well described by Kuhn (2), are shown in Table 3. The time at which the cement does not stick to the surgeon's glove is referred to as dough time (Fig. 1). The waiting phase ends at this time point. This occurs $\sim$2–3 min after the beginning of mixing for most PMMA cements in an ambient temperature of 23 °C (2,11). The working phase is the time during which the surgeon can easily apply cement to the femur. For manual application, the cement must no longer be sticky during this phase, and the viscosity must not to be too high. With the use of mixing systems, the user needs not to wait until the cement is no long sticky. The working time from the end of dough time to the cement is too stiff to manipulate is usually 5–8 min. The cement will fully polymerize to a hardened mass within 8–12 min after initial mixing (2,11).

The quality of the cement dough produced in the operation room will have considerable influence on the clinical long-term result of a cemented prosthesis. Mixing of cements is an important step, as it has a noticeable influence on the mechanical properties of acrylic bone cements (2,6,24). In the 1970s, loosening of the femoral stem was the most common reason for total hip arthroplasy revision, often occurring 5–10 years postoperatively with early component design and cement technique (21). Early methods of cement preparation involved hand mixing in open air of the MMA monomer with the prepolymerized powder mixture. A slurry was formed that could be hand patted or injected into the femoral canal. Many air bubble voids were

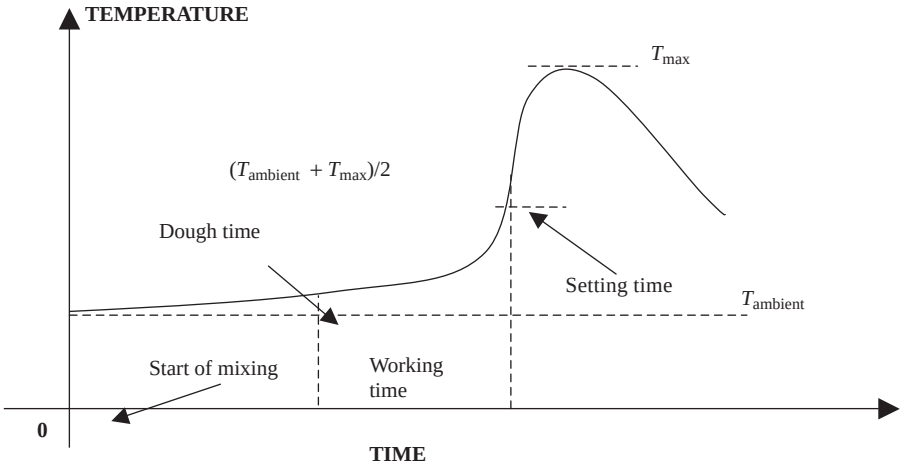


Figure 1. A typical temperature changes with respect to time during cement polymerization. Time zero is designated when the monomer liquid is added to the polymer powder.

created in the mixture during the hand-mixing process. To address this problem in cemented arthroplasty, intensive studies have attempted to improve the technique of cemented stem insertion (21). This results in advancing the cementing techniques from first generation cementing to second and third generation cementing (Table 4). Traditionally, bone cement was mixed using a spatula–bowl arrangement (first generation), which can have the consequence of introducing a high degree of porosity into the cement structure. In addition, the person mixing the cement was exposed to a high level of methyl methacrylate vapors. Second-generation cementing mainly consisted of the use of a canal plug, a cement gun and a high strength cement (21,41,42). Injection guns have been advocated for application of bone cement because long slender injection tips are useful for inserting the cement deep into the cavity, instead of trying to apply it by hand. Gun application also reduces the tendency to form laminations and voids in the cement and also reduces the inclusion of blood into the cement. Substantial improvements in stem survival were reported using such techniques, resulting in stem loosening rates of $< 5\%$ at 5–10 years (21,42). The introduction of porosity reducing measures marked the start of third generation cementing techniques. Cracks may initially occur at the voids in bone cements, which act as stress concentration points (24). Vacuum mixing is a typical ways to achieving pore reduction. Other third-generation developments in cementing techniques are the use of prosthesis positioning devices that ensure correct placement of the prosthesis, pressurization, and enhanced surface finishes (21,43). Attempts to improve cement-bone interlock using techniques such as endosteal preparation, retrograde cement insertion and cement pressurization improve implant survival. Improving cementing techniques have assisted to dramatically decrease the number of failure in the last three decades. Currently, cemented total hip arthroplasty shows survival rates of 90% at 15 years and 80–85% at 20 years (10,17,18,21).

Table 4. Developments in Cementing Techniques

Generations	Techniques	Time
1st Generation	Finger packing No cement restrictor	1960s
2nd Generation	Intramedullary femoral plug Cement gun High-strength cement	1970s
3rd Generation	Pressurization of cement after insertion Porosity reduction (vacuum mixing) Surface roughening or texturing Precoating Avoiding trochanteric osteotomy	mid-1980s–

POLYMERIZATION HEAT

Bone cement generates significant thermal energy during cement setting (an energy of 52 $kJ \cdot mol^{-1}$ of MMA (2)) and this may result in a temperature increase in the cemented system. The exothermic response of the bone cement can be characterized according to American Society for Testing and material (ASTM) standard F451 (44) or international standard ISO 5833. In the ASTM test, a package of cement is mixed at an air conditioned room ($23 \pm 1\,^{\circ}C$, $50 \pm 10\%$ relative humidity) as directed by the manufacturer's instructions. Within 1 min after doughing time, ~ 25 g of the dough is then gently packed into an ASTM specified test mold to achieve a 60 mm diameter with 6 mm thick disk cement mantle (44). The setting curve of temperature change with respect to time (Fig. 1) is recorded with a thermocouple (usually No. 24 gage wire k-type thermocouple) placed at the cement mantle center from the onset of mixing until cooling is observed. The setting time is defined as the time from initial mixing to the time at which the temperature of the polymerising mass has reached half of the maximum temperature (peak temperature) (44). The setting times range from 6 to 14 min and peak temperatures range from 54 to $83\,^{\circ}C$ for the popular commercial bone cements (2,11). A number of variables will affect the measured setting time and peak temperature, including the powder/liquid ratio of the cement, cement preparing method, cement mantle thickness, the initial temperature of the cement, and the ambient conditions (11). By mixing bone cement with Howmedica Mix-Kit I system and the Zimmer Osteobond vacuum system, Dunne and Orr (45)

found peak temperatures increased from 36 to 46 °C and 41 to 59 °C for Palacos R and CMW3 bone cement, specifically. Cold cement in a cold room takes longer to set. Precooling cement prior to use extends the cure time and decreases the peak temperature (46). Using a test mold and cementing techniques that simulated a clinical situation, Iesaka et al. (47) showed that the peak temperatures at the bone–cement interface were 53.1, 50.2, and 48.8 °C, while the polymerization time was 4.1, 8.3 and 11.2 min, when the monomer component of bone cement was initially at 37, 23, and 4 °C, respectively. Research also found that bone cement thickness has significant effects on the thermal responses. Meyer et al. (48) found a setting temperature of 60 °C with 3 mm thick specimens of Simplex R, and 107 °C with 10 mm thick specimens. Sih and Connelly (49) showed that the temperatures were 41, 56, and 60 °C for cement thicknesses of 1, 5, and 6-7 mm, respectively. Vallo (50) and Li et al. (51) demonstrated similar results with finite element modeling. Test mold materials also affect the measured setting time and peak temperature (50,51). According to the ASTM methodology, there is no limit to the setting time. However, the maximum allowable temperature for bone cements is 90 °C (44). There are limits stipulated for setting time and doughing time with ISO 5833 Standard.

The importance of temperature rise is that it may result in thermal necrosis of the bone tissue surrounding the implant (2,11,52–56). *In vitro* studies have shown that maximum temperature of bone cement can be reach higher than 100 °C, varied between 37 and 122 °C in different reports (57). Wang et al. (58) measured four different brands of bone cement (Palaces R, Simplex P, Sulfix, and CMW 1) during polymerization and the peak temperature in the cement was 46–124 °C. Clinical tests show a considerable lower temperature in the body (2,55). The bone-cement interface temperatures have ranged from 35 to 70 °C (2,11,52,3,54,55,57,59). Reasons for this observation are the thin layer of cement (~3–5 mm) and the blood circulation and heat dissipation in the vital tissue connected with it (2). Moreover, further heat dissipation of the system is attained via the prosthesis (2). Belkoff and Molloy (59) found that peak temperatures at the anterior cortex ranged from 44 to 113 °C, in the center ranged from 49 to 112 °C, and 39 to 57 °C at the spinal canal in the *ex vivo* vertebroplasaty tests. Toksvig-Larsen (57) performed tests with 31 total hip replacements and showed that the maximum temperature in the acetabulum ranged from 38 to 52 °C and in the femur from 29 to 56 °C. It has been found that not only the temperature, but also the exposure time plays a significant role in direct thermal cell necrosis (52,60–63). The findings of Moritz and Henriques (60) suggest that epithelial cell necrosis occurs 30 s after exposure to a temperature of 55 °C and 5 h after exposure to 45 °C. Lundskog (61) roughly confirmed these trends for bone cells, although he found a slightly lower threshold level. For example, he observed bone cell necrosis after 30 s at 50 °C. He also established that the regenerative capacity of the bone tissue is only damaged after exposure to temperatures of 70 °C and above. Eriksson and Alberksson (64) found the temperature threshold for impaired bone regeneration to be in the range of 44–47 °C for 1 min exposure. Thermal damage to bone tissue caused by cement polymerization cannot be ruled out and attempts to lower the potential degree of thermal necrosis have been investigated. To reduce the risk of thermal injury it is necessary to avoid exposing bone to temperatures above a certain threshold (62). The thermal responses rely on the quantity of heat produced by the bone cement (cement formulations and cement volume, etc.), rate of heat produced and how the heat conducts (55). The temperature peak can only by influenced to a small extent by adding heat-conducing radiopaque media or by slightly changing the chemical composition of the liquid (2). This, will, however, result in quite different dissolution properties with the polymer, which means it will result in different working properties and, usually, a significant reduction in mechanical stability (2). The thicker the cement mantle, the greater the volume of material and hence the more heat generated. However, simply reducing the thickness of the mantle is not a favorable option as the mechanics of the joint is affected by this. A slightly reduced level of heat generation has been shown by vacuum mixing the bone cement (45,57). Precooling the cement constituents prior to mixing increases the setting time, but has only minor effects on the maximum temperature of the cement (56). The use of a precooled femoral prosthesis did not affect the peak temperature as well (56,57). To avoid local tissue damage, surgeons may irrigate the implant with ice-cold saline during the polymerization process to decrease both the duration and the level of temperature elevation. Without cooling, the temperature was 49 °C (41–67) at the bone–cement interface, while it decreased to 41 °C (37–47) with water cooling in 19 cases of arthroplasties (65). Investigations on the potential tissue thermal necrosis have obtained varied results: while thermal bone necrosis due to cement curing heat has not been widely reported, some studies showed that thermal necrosis of bone did occur (52,54,59). Nevertheless, few would disagree that a clear understanding of the potential thermal necrosis problem requires further investigations.

VISCOSITY

During the working stage of the cement, its viscosity must be low enough to make it easy to force the cement dough through the delivery system and cause it to flow and penetrate into the interstices of the bone surface in a very short time (6,11,22). *In vitro* determination of viscosity is usually accomplished with the use of a capillary extrusion or rotational rheometer (6,11,22). The viscosity of the material can be determined by causing the material to flow at a specific shear rate and measuring the shear stress of the fluid against the stationary instrument (11). Viscosity varies across cements. The dynamic viscosity of its dough during the mixing period has been used to categorize bone cement materials into low viscosity brands (e.g., Osteopal), medium-viscosity brands (e.g., Simplex P), and high viscosity brands (e.g., Palacos R) (22). High viscosity bone cements typically have a doughy consistency; low viscosity cements are similar to viscous oil in consistency. Low viscosity cements have a long liquid

phase, or low viscosity wetting phase. The cement remains sticky for quite some time. Viscosity increases rapidly during the working phase, and the doughy cement becomes warm and sets quickly. High and low viscosity bone cements have different handling characteristics and require different cementing techniques. For optimal results, it is necessary to observe the specific mixing instructions for a given bone cement in combination with a given mixing system. Although some researchers suggest that low viscosity brands may have longer fatigue lives compared to high viscosity ones, some report no significant difference (22).

POROSITY, VOLUMETRIC CHANGES, AND RESIDUAL STRESS

Polymerized bone cement is a porous material, containing macropores (pore diameter > 1 mm) and micropores (pore diameter ≈ 0.1–1 mm) (6). The degree of porosity varies with cement brands and mixing methods. Jasty et al. (66) found that the porosities were 11.99% (CMW), 9.70% (Palacos R), 9.39% (Simplex P), 12.38% (Zimmer Regular), and 5.00% (Zimmer LVC) using manual mixing, while they were 6.00% (CMW), 4.25% (Simplex P), 6.00% (Zimmer Regular) and 4.15% (Zimmer LVC) with centrifugation. For CMW 1 cement, Muller et al. (67) found that the porosity decreased from 6.67 to 1.28% when using vacuum-mixing instead of hand-mixed methods. The pores generated in polymerized bone cement have been attributed to several sources (6,68–72). It may result from the air that is initially present in the powder interstices; entrapped during blending, mixing, transfer, or delivery; or entrained during insertion of the metal stem. Evaporation of the liquid monomer at the high temperatures of polymerization may contribute. Another aspect of porosity development is polymerization shrinkage.

The presence of pores in the polymerized cement may affect its mechanical properties. Pores may act as stress risers and initiation sites for cracks, rendering the cement susceptible to early fatigue failure (6). However, pores also may play a role in blunting crack propagation, thereby prolonging the life of the implant (73). *In vitro* experiments, the static and fatigue strength of bone cement both usually decrease with increased porosity (6,22,23,69,70). Reducing the porosity of both bulk cement and its interfaces should be of clinical benefit. Several methods of reducing the porosity of the bone cement have been developed (66). One commonly used method of void reduction is the centrifugation of a chilled, premixed bone cement prior to insertion into the bone. The other one is the mixing of the bone cement in a vacuum environment. Both techniques have been shown to substantially reduce the porosity of bone cement. It has been observed that mixing procedures plays a significant role in determining the quality of bone cement produced (2,74,75). The influence of vacuum mixing on the pores results in a 15–30% improvement of the bending strength of Palacos R while centrifuging Simplex P cement reduced its porosity from 9.4 to 2.9%. Experiments have shown that relative to hand mixing, centrifugation or vacuum mixing leads to a substantial reduction in porosity (2,6). The extent of such a reduction depends on many mixing variables, including mixing system, monomer storage temperature, vacuum pressure and centrifugation speed, and durations. In the hand-mixing process, air bubbles are mixed into the dough by thorough mixing. The porosity of the material is high and mechanical stability is endangered. Slower mixing of cement over a shorter time decreases air voids within cement and thus improves the strength characteristics. A high degree of porosity is found to exist in cement that is inadequately mixed (74). Monomer bubbles can easily appear, which may develop during the evaporation of the monomer while evacuating the system or later during polymerization under high pressure by the faulty use of vacuum-mixing systems.

Cement porosity distributions, especially at the bone–cement interface, may affect the cemented system. There is strong evidence that cracks in the cement are initiated at voids, particularly at the cement–prostheses interface (69). The preferential formation of voids at this site results from shrinkage during bone cement polymerization and the initiation of this process at the warmer bone–cement interface, which causes bone cement to shrink away from the prosthesis (2,68,69). It is expected that a reversal of polymerization direction would shrink the cement onto the prosthesis and reduce or eliminate the formation of voids at this interface (69). One innovative surgical approach to affect this behavior is to preheat the prosthesis prior to implantation. Results indicated that voids near the cement–prosthesis interface decreased significantly (69,76). The porosity at the cement–prosthesis decreased from 16.4 to 0.1% when the prosthesis was preheated to 37°C from room temperature, 23 °C. Additionally, the residual stress due to such polymerization curing was shown to decrease significantly at the cement–prosthesis interface (77). Studies also showed that preheating the prosthesis prior to implantation is unlikely to produce significant thermal damage to the bone when compared to implanting a prosthesis initially at room temperature (46,69). However, this procedure may induce more voids at the bone–cement interface, and its effects on the damage at this region should be studied (69,78).

The conversion of the monomer molecules into a polymer network is accompanied with a closer packing of the molecules, which leads to bulk contraction (68,79). A number of devices for determining the volumetric changes have been applied, including using a mercury dilatometer or a water displacement dilatometer (79,80). Theoretical calculations predict that bone cement polymerization will produce a volumetric shrinkage of 8% (81). Muller et al. (67) found volume shrinkages of CMW 1 bone cement were 3.43 and 5.99% with hand and vacuum mixing, respectively. Gilbert et al. (68) found shrinkages of Simplex P cement were 5.09 and 6.67%, respectively. The measured volume shrinkage is less than the theoretical prediction. This can be explained by void growth during polymerization (67).

In a situation where a curing material is bonded on all sides to rigid structures or constrained, bulk contraction cannot occur freely, and shrinkage must be compensated for by some kind of volume generation. This can come from a strain on the material and mainly for dislodgement of the bond, increase in porosity or internal loss of coherence (79).

Table 5. Comparison of the Values of Three Mechanical Properties of Six Different Bone Cements Under Same Test Regimes[a]

Cement Brands	ISO 5833 Bending Strength, MPa	Bending Modulus, MPa	Compressive Strength, MPa	DIN53435 Bending Strength, MPa	Impact Strength, $kJ \cdot m^{-2}$
CMW1	67.0	2634	94.4	86.2	3.7
CMW3	70.3	2764	96.3	72.4	2.9
Palacos	72.2	2628	79.6	87.4	7.5
Simplex P	67.1	2643	80.1	70.5	3.9
Osteobond	73.7	2828	104.6	80.1	3.5
Zimmer dough	62.5	2454	75.4	77.0	5.0

[a]See Ref. 2.

Shrinkage of the polymerizing cement *in vivo* (i. e., a constrained state) therefore might result in the development of porosity, both at the interfaces and inside the bulk cement. Thus, polymerization shrinkage may be a significant factor in porosity development (68). On the other hand, polymerization may induce high residual stresses. Shrinkage stress occurs when the material contraction is obstructed and the material is rigid enough to resist sufficient plastic flow to compensate for the original volume (79). The process of cement curing is a complex solidification phenomenon where transient stresses are generated and the residual stresses vary with different initial and boundary conditions during curing. A number of approaches have been used to estimate or measure the level of shrinkage stress in bone cement around femoral replacements, including theoretical model, finite element analysis, strain-gage methods and photoelastic methods (52,77,82–88). Currently, the subject of residual stress has often been neglected because it is assumed that residual stress will relax due to the viscoelastic properties of the cement. However, transient and residual stresses are believed to affect cement mechanical responses. Inclusion of the residual stress at the interface resulted in up to a four-fold increase in the von Mises cement stresses compared to the case without residual stresses (83,84). Recently, Orr et al. proposed that residual stresses are sufficient to initiate crack propagation in the cement before any load is applied (85). Lennon and Prendergarst have experimentally observed that residual stresses in the cement may induce cracking even before weight bearing of the cement (89). The initial residual stresses may have immediate effects influencing the possible initiation of cracks and debonding at the cement–prosthesis interface.

MECHANICAL PROPERTIES

Bone cement fills the space between the prosthesis and the bone; this connection is only a mechanical bond (1). Cement mechanical properties are therefore of particular significance for the performance of acrylic bone cement because cement must endure considerable stresses *in vivo*. There are many physical and mechanical properties of the cement that are considered germane to its clinical performance in the construct, including quasistatic tensile and compressive strength, modulus and ultimate strain, flexural strength and modulus, shear strength and modulus, fatigue properties (e.g., work of fracture, fracture toughness, fatigue resistance, fatigue crack propagation resistance), and creep (6). The mechanical properties of acrylic bone cement have been widely reported in the literature. Kuhn (2) investigated the static bending and compressive characteristics of a number of bone cements under the same test regimes (2). Test results of some of the most commonly used bone cement in United States are listed in Table 5. Osteobond cement was shown to have both the highest ISO 5833 bending strength and compressive strength among these six cement brands, while Zimmer dough has both the lowest ISO 5833 bending strength and compressive strength. All the measured bending strengths of the bone cement using the DIN53435 standard were larger than the ones using ISO 5833 standard, with Palacos R bone cement having the highest bending strength in this case. Harper and Bonfield (90) found a wide range of tensile strength (Table 6) and fatigue failure cycles (Table 7) results. They found that the Palacos R and Simplex P cements were significantly higher in tensile strength compared to the other cements tested with exception of CMW 3. There was no statistical difference between the values obtained for CMW 1 and Osteobond. The differences among the fatigue results for the different cements were much larger than those found with the static tensile results. The highest Weibull median fatigue cycles to failure obtained for Simplex P and Palacos R were considerably higher than found for Zimmer dough type. Harper and Bonfield (90) found that there was some correlation between the static and fatigue strengths, but the ranking of static strength does not exactly follow that of fatigue life. The fatigue results were found to correlated well to the clinical data (91): the order of success of implants with the cement brand was the same as that obtained from the fatigue test.

It has long been recognized that PMMA surgical bone cement undergoes viscoelastic (creep) deformation under physiological loads (92,93). Many studies have been performed to assess the viscoelastic properties of bone cement (6,23). Radiological observations of hip stems have shown subsidence of the stem within the cement mantle without visible cement fractures (94). Creep has been implicated in prosthesis subsidence, in particular subsidence of the femoral stem in hip replacements (6,95). Excessive subsidence can lead to prostheses loosening. Lu and McKellop (92) studied the effects of cement creep on the subsidence of

Table 6. Comparison of Tensile of Six Different Bone Cements[a]

Cement Brands	Ultimate Strength, MPa	Modulus of Elasticity, GPa	Strain at Fracture, %
CMW1	39.1	2.96	1.60
CMW3	44.7	3.53	1.36
Palacos	51.4	3.21	2.25
Simplex P	50.1	3.43	1.87
Osteobond	38.2	3.38	1.41
Zimmer dough	31.7	2.79	1.43

[a]See Ref. 90.

the stem and on the stress within the cement using a cyclic load and three interface bonding conditions (bonded, frictional, and debonded). Results showed that the creep deformation of the cement was accompanied by additional subsidence of the stem and a decrease in the stress components within the cement. The results agreed with an experimental study using stems cemented into cadaver femora (96). Cement creep would accumulate for the frictional stem–cement interface, resulting in 0.46 mm total stem subsidence and a 13% decrease in the stress within the cement (92). Ling et al. (97) concluded that, at least for smooth tapered stems, that substantial hoop and radial creep of the cement not only occurs, but also is essential for the optimum clinical performance of the prosthesis. On the other hand, a limited degree of creep may help in maintaining the cement–implant interface. Also, Harris (98) stated that creep of the cement mantle surrounding a hip prosthesis may be negligible under cyclic physiological loading. The role of creep of the bone cement on the chance of failure of the cement mantle is still a subject of controversy. Due to its viscoelastic nature, cement tested at different strain rates may have changing characteristics (11). There is still lack of tests of bone cement in ways that represent real-life loading patterns that mimic those experienced by the cement *in vivo* (99). The true understanding of the mechanical behavior of bone cement can only be attained if the testing procedure is truly representative (99).

There are considerable differences between the values of physical and mechanical properties of bone cement reported in the literature. This disagreement may be the result of cement type, cement preparation technique, specimen geometry, measurement technique, test parameters, and testing environment (11). Therefore it may be impossible to compare results from different investigations (90). However, results with specific conditions may be found in a number of the excellent comprehensively reviews (1–3,6,11,22,23). It has been shown that cement formulations play a significant role on the cement mechanical properties (Table 5). Also, the test methods will affect the values obtained, which can be easily seen by comparing the bending strength results using ISO 5833 standard to that using DIN 53435 standard (Table 5). Cement mixing methods (hand mixing, vacuum mixing, or centrifugation) have been shown to affect the physical properties of bone cement (74,100,101). Even with vacuum mixing, using different vacuum mixing systems have resulted in different bone cement porosity, static strength, and fatigue strength (75,102). The storage temperature of cement constitutive was shown to have minor effects on porosity and fatigue performance (100). Ishihara et al. (103) showed that the fatigue of bone cements at 1 Hz are shorter by one to two order of magnitude as compared with fatigue lives at 20 Hz. On the other hand, Lewis et al. (104) found that frequency (over the range used) did not exert a statistically significant effect on the fatigue life of cement tested in their investigations.

Table 7. Comparison of Fatigue Test Results of Six Different Bone Cements[a]

	Cycles to Failure	
Cement Brands	Range	Weibull Median
CMW1	3042–8835	4407
CMW3	5996–38262	16441
Palacos	18362–49285	27892
Simplex P	8933–93345	36677
Osteobond	5527–25825	16162
Zimmer dough	153–3978	781

[a]See Ref. 90.

SUMMARY

Acrylic bone cement has been widely used in orthopedic surgery. Currently, the cement is not without its drawbacks as discussed previously. One major research area focusing on modifications of cement formulations may lead to obtain more favorable cement mechanical properties and biological compatibilities. One example of these developments is to incorporate second phase materials (bioactive agents, reinforcement fibers, etc.) into the existing bone cement. Another approach to overcome cement weakness is to manipulate cement processing procedure and the surgical techniques. A good example is to reduce cement porosity by vacuum mixing techniques. It is worth pointing out that considerable research is needed to develop techniques for accurate characterizations of cement properties, monitoring cement curing process and evaluations of potential bone cement failure.

BIBLIOGRAPHY

Cited References

1. Walenkamp G, Murray DW. Bone cements and cementing technique. Berlin Heidelberg New York: Springer-Verlag; 2001.
2. Kuhn KD. Bone cement: up-to-date comparison of physical and chemical properties of commercial materials. Berlin: Springer-Verlag; 2000.
3. Krause W, Mathis RS. Fatigue properties of acrylic bone cements - review of the literature. J Biomed Mater Res-Appl Biomate 1988;22(A1):37–53.
4. Charnley J. Anchorage of the femoral head prosthesis to the shaft of the femur. J Bone Joint Surg 1960;43B:28–30.

5. Kenny SM, Buggy M. Bone cements and fillers: A review. J Mater Sci-Mater Med 2003;14(11): p 923–938.
6. Lewis G. Properties of acrylic bone cement: State of the art review. J Biomed Mater Res 1997;38(2):155–182.
7. NIH. Total hip replacement. NIH Consensus Statement 1994;12(5):1–31.
8. Cristofolini L. A critical analysis of stress shielding evaluation of hip prostheses. Crit Rev Biomed Eng 1997;25(4–5):409–483.
9. Chao E. Orthopaedic Biomechanics. Int Orthop (SICOT) 1996;20:239–243.
10. Huiskes R, Verdonschot N. Biomech Art Joints: the Hip, Basic Orthopaedic Biomechanics. In: Mow VC, Hayes WC, editors. Philadelphia: Lippincott-Raven Publishers; 1997. 395–460.
11. Krause WR. Bone Cement, Acrylic. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation. New York: John Wiley & sons Inc.; 1988. p 491–500.
12. Wilcox RK. The biomechanics of vertebroplasty: a review. Proceedings of the Institution of Mechanical Engineers Part H. J Eng Med 2004;218(H1):1–10.
13. Phillips FM. Minimally invasive treatments of osteoporotic vertebral compression fractures. Spine 2003;28(15):S45–S53.
14. Heini PF, et al. Femoroplasty-augmentation of mechanical properties in the osteoporotic proximal femur: a biomechanical investigation of PMMA reinforcement in cadaver bones. Clin Biomech 2004;19(5):506–512.
15. Hasenwinkel J. Bone Cement. In: Wnek GE, Bowin GL, editors. Encyclopedia of Biomaterials and Biomedical Engineering. New York: Marcel Dekker; 2004. p 170–179.
16. Serbetci K, Hasirci N. Recent developments in bone cements. In: Yaszemski MJ et al. editors. Biomaterials in Orthopedics. New York: Marcel-Dekker; 2004. p 241–286.
17. Hendriks JGE, et al. Backgrounds of antibiotic-loaded bone cement and prosthesis-related infection. Biomaterials 2004;25(3):545–556.
18. Murray DW, Carr AJ, Bulstrode CJ. Which Primary Total Hip-Replacement. J Bone Joint Surg-Br Vol 1995;77B(4): 520–527.
19. Nafei A, et al. Survivorship analysis of cemented total condylar knee arthroplasty: a long-term follow-up report on 348 cases. J Arthroplasty 1996;11:7–10.
20. El-Warrak AO, et al. A review of aseptic loosening in total hip arthroplasty. Veterin Compar Orthopae Traumatol 2001;14(3):115–124.
21. Barrack RL. Early failure of modern cemented stems. J Arthroplasty 2000;15(8):1036–1050.
22. Lewis G. Fatigue testing and performance of acrylic bone-cement materials: State-of-the-art review. J Biomed Mater Res Part B-Appl Biomater 2003;66B(1):457–486.
23. Saha S, Pal S. Mechanical-Properties of Bone-Cement-a Review. J Biomed Mater Res 1984;18(4):435–462.
24. Deb S. A review of improvements in acrylic bone cements. J Biomater Appl 1999;14(1):16–47.
25. Passuti N, Gouin F. Antibiotic-loaded bone cement in orthopedic surgery. Joint Bone Spine 2003;70(3):169–174.
26. Jasty M, et al. The Initiation of Failure in Cemented Femoral Components of Hip Arthroplasties. J Bone Joint Surg Br Vol 1991;73(4):551–558.
27. Hertzberg RW, Manson JA. Fatigue of engineering plastics. London: Academic Press; 1980.
28. Spector M. Biomaterial failure. Orthoped Clin N Am 1992;23(2):211–217.
29. Pourdeyhimi B, Wagner HD. Elastic and Ultimate Properties of Acrylic Bone-Cement Reinforced with Ultra-High-Molecular-Weight Polyethylene Fibers. J Biomed Mater Res 1989;23(1):63–80.
30. Harper EJ, Behiri JC, Bonfield W. Flexural and fatigue properties of a bone cement based upon polyethylmethacrylate and hydroxyapatite. J Mat Sci Mat Med 1995;6:799–803.
31. Gilbert JL, Net SS, Lauthenschlager EP. Self-reinforced composite poly(methylmethacryate): static and fatigue properties. Biomaterials 1995;16:1043–1055.
32. Pourdeyhimi B, Wagner HD, Schwartz P. A Comparison of Mechanical-Properties of Discontinuous Kevlar-29 Fiber Reinforced Bone and Dental Cements. J Mater Sci 1986;21(12):4468–4474.
33. Wright TM, Trent PS. Mechanical-Properties of Aramid Fiber-Reinforced Acrylic Bone Cement. J Mater Sci 1979;14(2):503–505.
34. Saha S. Strain-rate dependence of the compressive properties of normal and carbon-fiber-reinforced bone-cement. J Biomed Mater Res 1983;17(6):1041–1047.
35. Saha S, et al. Biomechanical Evaluation of Bony Defects Repaired with Normal, Carbon-Fiber, and Wire Reinforced Bone-Cement. Biomater Med Devices Artif Organs 1981;9(4):291–291.
36. Pilliar RM, et al. Carbon Fiber-Reinforced Bone Cement in Orthopedic Surgery. J Biomed Mater Res 1976;10(6):893–906.
37. Topoleski LDT, Ducheyne P, Cuckler JM. The Fracture-Toughness of Titanium-Fiber-Reinforced Bone-Cement. J Biomed Mater Res 1992;26(12):1599–1617.
38. Saha S, Kraay MJ. Bending Properties of Wire-Reinforced Bone-Cement for Applications in Spinal Fixation. J Biomed Mater Res 1979;13(3):443–457.
39. Kotha SP, et al. Fracture toughness of steel-fiber-reinforced bone cement. J Biomed Mater Res Part A 2004;70A(3):514–521.
40. Fishbane BM, Pond RB. Stainless steel fiber reinforcement of polymethylmethacrylate. Clin Orthop Rel Res 1977;128:194–199.
41. Mulroy WF, Harris WH. Revision total hip arthroplasty with use of so-called second-generation cementing techniques for aseptic loosening of the femoral component—A fifteen-year-average follow-up study. J Bone Joint Surg—Am Vol 1996;78A(3):325–330.
42. Mulroy WF, Estok DM, Harris WH. Total hip arthroplasty with use of so-called second-generation cementing techniques—A fifteen-year-average follow-up study. J Bone Joint—Am Vol 1995;77A(12):1845–1852.
43. Faulkner A, et al. Effectiveness of hip prostheses in primary total hip replacement: a critical review of evidence and an economic model. Health Technol Asess 1998;2(6):1–146.
44. ASTM, ASTM Standard Specification for Acrylic Bone Cement F451-99a; 1999.
45. Dunne NJ, Orr JF. Curing characteristics of acrylic bone cement. J Mater Sci Mater Med 2002;13(1):17–22.
46. Li CD, Schmid S, Mason J. Effects of pre-cooling and preheating procedures on cement polymerization and thermal osteonecrosis in cemented hip replacements. Med Eng Phys 2003;25(7):559–564.
47. Iessaka K, Jaffe WL, Kummer FL. Effects of the initial temperature of acrylic bone cement liquid monomer on the properties of the stem-cement interface and cement polymerization. J Biomed Mater Res: Appl Biomater 2003;68B:186–190.
48. Meyer PR, Lautenschlager EP, Moore BK. On the setting properties of acrylic bone cement. J Bone Joint Surg 1973;55A:139–156.
49. Sih GC, Connelly GM, Berman AT. The Effect of Thickness and Pressure on the Curing of Pmma Bone-Cement for the Total Hip-Joint Replacement. J Biomech 1980;13(4): 347–352.

50. Vallo CL. Theoretical prediction and experimental determination of the effect of mold characteristics on temperature and monomer conversion fraction profiles during polymerization of a PMMA-based bone cement. J Biomed Mater Res (Appl Biomater) 2002;63:627–642.
51. Li CD, Mason J, Yakimicki D. Thermal characterization of PMMA-based bone cement curing. J Mater Sci—Mater Med 2004;15(1):85–89.
52. Huiskes R. Some fundamental aspects of human joint replacement. Acta Orthopaed Scand 1980; (Suppl.) 185.
53. Mjoberg B. Fixation and Loosening of Hip Prostheses—a Review. Acta Orthopaed Scand 1991;62(5):500–508.
54. Mjoberg B, et al. Bone-Cement, Thermal-Injury and the Radiolucent Zone. Acta Orthopaed Scand 1984;55(6):597–600.
55. Dipisa JA, Sih GS, Berman AT. Temperature Problem at Bone-Acrylic Cement Interface of Total Hip-Replacement. Clin Orthopae Relat Res 1976;121:95–98.
56. Swenson LW, Schurman DJ, Piziali RL. Finite element temperature analysis of a total hip replacement and measurement of PMMA curing temperatures. J Biomed Mater Res 1981;15:83–96.
57. Toksviglarsen S, Franzen H, Ryd L. Cement Interface Temperature in Hip-Arthroplasty. Acta Orthopaed Scand 1991;62(2):102–105.
58. Wang JS, Franzen H, Toksviglarsen S. Does vacuum mixing of bone-cement affect heat-generation - analysis of 4 cement brands. J Appl Biomater 1995;6(2):105–108.
59. Belkoff SM, Molloy S. Temperature measurement during polymerization of polymethylmethacrylate cement used for vertebroplasty. Spine 2003;28(14):1555–1559.
60. Moritz AR, Henriques FC. The relative importance of time and surface temperature in the causation of cutaneous burns. Am J Pathol 1947;23:695–720.
61. Lundskog J. Heat and bone tissue: an experimental investigation of the thermal properties of bone and threshold levels for thermal injury. Scand J Plastic Reconstr Surg 1972;9:1–80.
62. Revie I, Wallace M, Orr J. The effect of PMMA thickness on thermal bone necrosis around acetabular sockets. Proc Instn Mech Eng 1994;208:45–51.
63. Nelson C, Krishnan E, Neff J. Consideration of physical parameters to predict thermal necrosis in acrylic cement implants at the site of giant cell tumors of bone. Med Phys 1986;13(4):462–488.
64. Eriksson AR, Albreksson T. Temperature threshold levels for heat induced bone tissue injury: a vital microscopic study in the rabbit. J Prosthetic Den 1983;50(1):101–107.
65. Wykman AGM. Acetabular Cement Temperature in Arthroplasty - Effect of Water Cooling in 19 Cases. Acta Orthopaed Scand 1992;63(5):543–544.
66. Jasty M, et al. Porosity of various preparations of acrylic bone cements. Clin Orthop Rel Res 1990;259:122–129.
67. Muller SD, Green SM, McCaskie AW. The dynamic volume changes of polymerising polymethy methacrylate bone cement. Acta Orthopaed Scand 2002;73(6):684–687.
68. Gilbert JL, et al. A theoretical and experimental analysis of polymerization shrinkage of bone cement: A potential major source of porosity. J Biomed Mater Res 2000;52(1): 210–218.
69. Bishop NE, Ferguson S, Tepic S. Porosity reduction in bone cement at the cement-stem interface. J Bone Joint Surg-Br Vol 1996;78B:349–356.
70. James SP, et al. A fractographic investigation of PMMA bone cement focusing on the relationship between porosity reduction and increased fatigue life. J Biomed Mater Res 1992;26(5):651–662.
71. James SP, et al. Extensive porosity at the cement-femoral prosthesis interface: A preliminary study. J Biomed Mater Res 1993;27:71–78.
72. Wixson RL, Lautenschlager EP, Novak MA. Vacuum mixing of acrylic bone cement. J Arthroplasty 1987;2:141–149.
73. Topoleski LDT, Ducheyne PI, Cuckler JM. Microstructural pathway of fracture in poly(methyl methacrylate) bone cement. Biomaterials 1993;14(15):1165–1172.
74. Dunne NJ, Orr JF. Influence of mixing techniques on the physical properties of acrylic bone cement. Biomaterials 2001;22(13):1819–1826.
75. Mau H, et al. Comparison of various vaccum mixing system and bone cements as regards reliability, porosity and bending stength. Acta Orthopaed Scand 2004;75(2):160–172.
76. Iessaka K, Jaffe WL, Kummer FJ. Effects of preheating of hip prosthesis on the stem-cement interface. J Bone Joint Surg—Am Vol 2003;85:421–427.
77. Li CD, Wang Y, Mason J. The effects of curing history on residual stresses in bone cement during hip arthroplasty. J Biomed Mater Res Part B-App Biomater 2004;70B(1):30–36.
78. Race A, et al. Early cement damage around a femoral stem is concentrated at the cement/bone interface. J Biomech 2003;36:489–496.
79. Davidson CL, Feilzer AJ. Polymerization shrinkage and polymerization shrinkage stress in polymer-based restoratives. J Den 1997;25(6):435–440.
80. Davies JP, Harris WH. Comparison of diametral shrinkage of centrifuged and uncentrifuged Simplex P bone cement. J Appl Biomater 1995;6:209–211.
81. Hass S, Brauer G, Dickson G. A characterization of polymethyl methacrylate bone cement. J Bone Joint Surg—Am Vol 1975;57:380–391.
82. Ahmed AM, et al. Transient and residual stresses and displacements in self-curing bone cement - part I: characterization of relevant volumetric behavior of bone cement. J Biomech Eng—Trans ASME 1982;104:21–27.
83. Nuno N, Amabili M. Modeling debonded stem-cement interface for hip implants: effect of residual stresses. Clin Biomech 2002;17:41–48.
84. Nuno N, Avanzolini G. Residual stresses at the stem-cement interface of an idealized cemented hip stem. J Biomech 2002;35:849–852.
85. Orr JF, Dunne NJ, Quinn JC. Shrinkage stresses in bone cement. Biomaterials 2003;24(17):2933–2940.
86. Ahmed AM, et al. Transient and residual stresses and displacements in self-curing bone cement-Part II: thermoelastic analysis of the stem fixation system. J Biomech Eng—Trans Asme 1982;104:28–37.
87. Roques A, et al. Quantitative measurement of the stresses induced during polymerization of bone cement. Biomaterials 2004;25:4415–4424.
88. Zor M, Kucuk M, Aksoy S. Residual stress effects on fracture energies of cement-bone and cement-implant interfaces. Biomaterials 2002;23:1595–1601.
89. Lennon AB, Prendergast PJ. Residual stress due to curing can initiate damage in porous bone cement: experimental and theoretical evidence. J Biomech 2002;35:311–321.
90. Harper EJ, Bonfield W. Tensile characteristics of ten commercial acrylic bone cements. J Biomed Mater Res 2000;53(5):605–616.
91. Malchau H, Herberts P. Prognosis of total hip replacement and revision rate in THR: A revision risk study of 148,359 primary operations. 65th Ann Meet Am Acad Ortho Surg 1998. New Orleans.
92. Lu Z, Mckellop H. Effects of cement creep on stem subsidence and stress in the cement mantle of a total hip replacement. J Biomed Mater Res 1997;34:221–226.

93. Verdonschot N, Huiskes R. Creep properties of three low temperature-curing bone cements: a preclinical assessment. J Biomed Mater Res 2000;53B:498–504.
94. Fowler GA, et al. Experience with the Exeter total hip replacement since 1970. Ortho Clin N Am 1988;19:477–489.
95. Morgan RL, et al. Creep behavior of bone cement: a method for time extrapolation using time-temperature equivalence. J Mater Sci: Mater Med 2003;14:321–325.
96. Weightman B, et al. The mechanical properties fo cement and loosening of the femoral component of hip replacements. J Bone Joint Surg Br 1987;69:558–564.
97. Ling RSM. The use of a collar and precoating on cemented femoral stems is unnecessary and detrimental. Clin Orthop Rel Res 1992;285:73–83.
98. Harris WH. Is it advantageous to strengthen the cement-matel interface and use a collar for cemented femoral components of total hip replacement. Clin Orthop Rel Res 1992;285:67–72.
99. Eden OR, Lee AJC, Hooper RM. Stress relaxation modeling of polymethylmethacrylate bone cement. Proc Instn Mech Eng 2002;216H:195–199.
100. Lewis G. Effect of mixing method and storage temperature of cement constituents on the fatigue and porosity of acrylic bone cement. J Biomed Mater Res 1999;48B:143–149.
101. Macaulay W, et al. Difference in bone-cement porosity by vacuum mixing, centrifugation, and hand mixing. J Arthroplasty 2002;17(5):569–575.
102. Dunne NJ, et al. The relationship between porosity and fatigue characteristics of bone cements. Biomaterials 2003;24(2):239–245.
103. Ishihara S, et al. On fatigue lifetimes and fatigue crack growth behavior of bone cement. J Mater Sci Mat Med 2000;11(10):661–666.
104. Lewis G, Janna S, Carroll M. Effect of test frequency on the in vitro fatigue life of acrylic bone cement. Biomaterials 2002;24:1111–1117.

See also BIOMATERIALS, TESTING AND STRUCTURAL PROPERTIES; HIP JOINTS, ARTIFICIAL; ORTHOPEDICS, PROSTHESIS FIXATION FOR; RESIN-BASED COMPOSITES.

BONE DENSITY MEASUREMENT

YIXIAN QIN
ERIK MITTRA
Stony Brook University
New York

INTRODUCTION

Chronic diseases, such as musculoskeletal complications, have a long-term debilitating effect that greatly impacts quality of life. Osteoporosis is a reduction in bone mass or density that leads to deteriorated and fragile bones and is the leading cause of bone fractures in postmenopausal women and in the elderly population for both men and women. About 13–18% of women aged 50 years and older, and 3–6% of men aged 50 years and older, have osteoporosis in the United States alone. These rates correspond to $4 \sim 6$ million women and $1 \sim 2$ million men who suffer from osteoporosis (1). One-third of women over 65 will have vertebral fractures and 90% of women aged 75 and older have radiographic evidence of osteoporosis (2–4). Another 37–50% of women aged 50 years and older, and 28–47% of men of the same age group, have some degree of osteopenia. Thus, approximately a total of 24 million people suffer from osteoporosis in the United States alone, with an estimated annual direct cost of over $18 billion to national health programs. Hence, early diagnosis that can predict fracture risk and result in prompt treatment is extremely important. Early identification of fracture risk, most commonly caused by osteoporosis-induced bone fragility, is also important in implementing appropriate treatment and preventive strategies. Indeed, the ability to accurately assess bone fracture risk noninvasively is essential for improving the diagnostic as well as therapeutic goals (i.e., assessing temporal changes in bone during therapy) for bone loss from such varied etiologies as osteoporosis, microgravity, bed rest, or stress-shielding around an implant.

Assessment of bone mineral density (BMD) has become as essential element in the evaluation of patients at risk for osteopenia and osteoporosis (2,5–8). Bone density was initially estimated from the conventional X ray by comparing the image density of the skeleton to the surrounding soft tissues. Although demineralized bone has an image density closer to soft tissues, dense mineralized skeletal tissues appears relatively white on an X-ray image. Hence, the mineral density of bone can be estimated by the degree of gray color of the X-ray image in the bone region. However, because of its resolution and variations generated in the X-ray image, it has been suggested that bone mineral losses of at least 30% are required before they may be visually measured using a conventional X ray (9,10). Growing awareness of the impact of osteoporosis on the elderly population and the consequent costs of health care, together with the development of new treatments to prevent fractures, have led to a rapid increase in the demand for bone densitometry measurements. Many image modalities and techniques have been developed to improve the quality and the accuracy of the measurement for bone mineral and dense assessment. Two major densitometry techniques are commonly used in assessing bone density, that is, radiography-based densitometry and ultrasound-based assessment.

RADIOGRAPHY-BASED DENSITOMETRY

To improve the sensitivities of X-ray images to bone density changes and assessment, several technologies have been developed. *Bone densitometry* is a term that is defined as a method for imaging density of bone. However, the "true" density is not applicable in the current radiography-based techniques. In the field of densitometry, the term "bone mineral density", referred to as BMD, is related to the mass of bone in the tissue level, which includes both bone and marrow components as well as surrounding soft tissues. Furthermore, most densitometric techniques are projectional for the image formation that provides a two-dimensional image of the three-dimensional (3D) bone volume being measured. Therefore, the BMD defined from the projectional techniques is the mass of bone tissue mass (including marrow and/or soft surroundings) per unit area

Table 1. Radiography-Based Bone Densitometry[a]

Technique	ROI	Unit	Precision, %CV	Effective Dose, μSv
SXA	Total body	BMD ($g \cdot cm^{-2}$)	1	3
QCT	Spine	BMD ($g \cdot cm^{-2}$)	3	50–500
pQCT	Forearm	BMD ($g \cdot cm^{-2}$)	1–2	1–3
RA	Phalanx	BMD ($g \cdot cm^{-2}$)	1–2	10
SPA	Forearm	BMD ($g \cdot cm^{-2}$)	3–4	1–10
DPA	Total body	BMD ($g \cdot cm^{-2}$)	1	1–10
DXA	PA spine	BMD ($g \cdot cm^{-2}$)	1	1–10
	Proximal femur		1–2	1–10
	Total body		1	3

[a]See Refs. 5,6, 12–14.

in the image, not per unit volume of the tissue. Hence, what is actually measured is the apparent bone mineral density, which is defined by the bone mineral content contained in the area scanned, or expressed as gram per squared centimeter in unit. To detect osteoporosis accurately, several methods are developed for the noninvasive measurement of the skeleton for the diagnosis of osteopenia, osteoporosis, and/or the evaluation of an increased risk of fracture (11). These methods include single-energy X-ray absorptiometry (SXA), dual energy X-ray absorptiometry (DXA or DEXA), quantitative computed tomography (QCT), peripheral quantitative computed tomography (pQCT), radiographic absorptiometry (RA), dual photon absorptiometry (DPA), and single photon absorptiometry (SPA). There are two types of BMD measurements, peripheral BMD and central BMD. The peripheral BMD instruments are usually smaller, less expensive, and more portable than the central BMD. Central BMD is capable of measuring multiple skeletal sites, that is, the spine, the hip, and the forearm. Table 1 lists these methods currently available for the noninvasive measurement of the skeleton for the diagnosis of osteoporosis. These techniques differ substantially in physical principles, in the particular physical body sites (e.g., spine, hip, or total body), in the clinical discrimination and interpretation, and in availability of the facility and cost.

Single-Energy Densitometry

This instrumentation passes a beam of radiation through the limb of the body (e.g., forearm) and determines the difference between the incoming (or incident) radiation and the outgoing (or transmitted) radiation, referring to the attenuation. The higher the bone mineral content, the greater the attenuation. Mineral content can be calculated by the attenuation of the radiation. BMD can then be calculated by dividing the mineral content by the detected bone area. The relation between incoming and outgoing X-ray energy can be expressed as

$$I = I_0 \exp(-\lambda d) \tag{1}$$

where I_0 = incoming radiation intensity, I = transmitted radiation intensity, λ = mass attenuation coefficient, and d = area density of the attenuating materials ($g \cdot cm^{-2}$). The mass attenuation coefficient is a physical property that describes how much a given material attenuates and X-ray energy. If the attenuation coefficient can be experimentally determined, the equation becomes explicit, and the area density can be determined by

$$d = k \log(I_0/I) \tag{2}$$

where k is an experimentally determined constant for the attenuation coefficient.This technology is relatively simple and easy to understand. However, biological tissues and body are composed by multiple materials (e.g., bone, muscle, and other soft tissues). The accuracy of the technology is limited.

Single-Photon Absorptiometry

Bone density can be measured by passing a monochromatic or single-energy photon beam through bone and soft tissue. This procedure is referred as SPA. The amount of mineral content can be quantified by the attenuation of the beam intensity. After the photon beam attenuation is calculated, the value of the attenuation can be compared with a calibration parameter derived from a standard mineral content (e.g., using ashed bone of known weight). This procedure can finally determine the BMD with measured attenuation. Iodine-125 at 27 keV, or americium-241 at 59.5 keV, was initially used for generating of the SPA beam. SPA is rarely used in clinical practice today. SPA determined bone mineral content is calculated through uniform thickness of the soft tissue in the path of the beam. The targeted scan site (e.g., limb or forearm) had to be submerged in the water or a tissue-equivalent material, which limited the practical applications of the SPA. The advantages of this technique include a low dose of radiation, portable, and use for particular body sites with relatively precisely measurement. Although the SPA is an approximate method, the limitations of SPA include limited accuracy of the measurement, radiation, and used only on the particular peripheral sites, like forearm and heel.

Dual-Photon Absorptiometry

To overcome the limitations of single-energy or photon densitometry, if a dual radiation source was used, the influence of soft tissues could be eliminated. The basic principle involved in DPA for bone density measurement was similar to SPA. The degree of attenuation of the photon energy beam between incoming and outgoing energy through bone and soft tissue is quantified. As with SPA, the beam source was originally used, but with an isotope

used, which emitted photon energy at two distinct photoelectric peaks. When the beam was passed through a region of the body with both hard and soft tissues, attenuation of the photon beam appeared to both photon energy peaks. The contributions of soft tissue to beam attenuation can be determined by the quantifications of the relative relations between two attenuations (15). Because of its capability to distinct bone from soft tissue, DPA has been used to quantify bone density in deep tissue and large skeletal areas where bone is surrounded by large volume of soft masses (e.g., spine and hip) (16). DPA was considered a major advance from SPA due to its ability to quantify BMD and mineral content in such deep areas like spine and hip, as well as its capability of quantifications of effects from soft tissues. However, DPA has many notable limitations. First, the maintenance of the beam source was expensive, which had to be replaced yearly. Second, the radioactive source decay increased as much as 0.6% per month, which added difficulties for the calibration. These factors may result in the precision of 2–4% for DPA measurements in the region of interests. This precision (e.g., 2%) would limit its clinical application, in which a great change (e.g., 5–6%) from the baseline value had to be observed before one could reach the 95% confidence level for the change of bone density. Nevertheless, the concept of dual-photon densitometry has impacted the development of new technologies such as DXA.

Dual-Energy X-ray Absorptiometry

Perhaps the most popular bone densitometry used in clinical practice is the DXA or DEXA. The basic principles of DXA are the same as DPA. To overcome the major limitation of DPA, it did not take long for manufacturers who originally had the DPA product to replace the decaying isotope beam source with a highly stable dual-energy X-ray tube. There are several advantages of using X-ray sources over radioactive isotopes (i.e., no beam decay concerned in the X-ray tube and no calibration required for correction of the drifting because of the source decay in the DPA). The fundamental basis for DXA is the measurement of the transmission through the body of X rays of two different photon energies. The radiation source is collimated to a pencil beam and aimed at a radiation detector placed directly opposite the objective to be measured (Fig. 1). The patients are positioned on a table in the path of the X-ray beams. Due to the dependence of the attenuation coefficient on atomic number and photon energy, assessment of the transmission factors and attenuations at two energies enables the 2D apparent density, that is, bone density per unit projected area, of two different types of tissues to be inferred (17–19). The X-ray source and detector pair is scanned back and forth across the region of interests in the body, generating annotation images, which the BMD is calculated as the ratio of the bone content to the measured area. Radiation dose to the patients is very low on the order of 1–10 μSv. The DXA system can measure the BMD of the spine, proximal femur, forearm, and the total body. Recent technology uses a fan beam geometry in the DXA scanners that can increase the speed and reduce the acquisition time (GE Medical System Inc.). The image

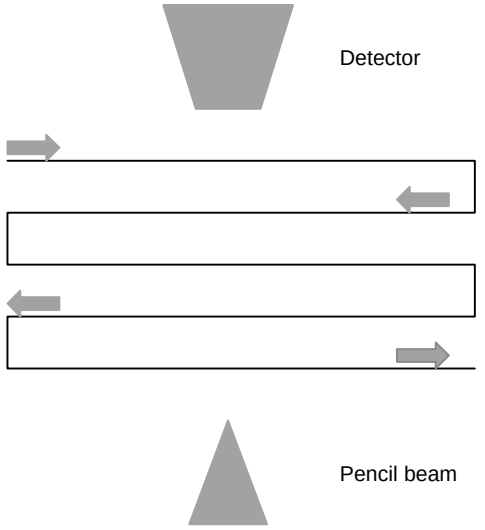


Figure 1. Scan path pattern for DXA densitometer using pencil-beam format.

quality of recent DXA has improved significantly via computational capability for better visualization. In addition to the body DXA scanners, recent technology has also adapted for development of lower cost, small, and particular site densitometries (Fig. 2). These systems are available to the clinic for specific body regions (e.g., spine, hip, leg, arm, and hand). The DXA systems are available for the diagnostic clinical use by many major manufacturers (e.g., GE Medical Systems of Madison, Norland, and Hologic Inc. of Bedford).

The basic working principle of DXA and its ability to reduce the effects of soft tissue is to use two X-ray sources and mathematically solve the bone thickness and soft-tissue thickness (15). By using two X-ray energies, two equations can be derived by scanning the measurement site twice with low (L) and high (H) energies once at each.

$$I^L = I_0^L[\exp - (\lambda_b^L d_b + \lambda_s^L d_s)] \quad (3)$$

$$I^H = I_0^H[\exp - (\lambda_b^H d_b + \lambda_s^H d_s)] \quad (4)$$

where $I_0 =$ incoming radiation intensity, $I =$ transmitted radiation intensity, $\lambda =$ mass attenuation coefficient, $d =$ area density of the attenuating materials ($g \cdot cm^{-2}$), and b and s refer to the bone and soft tissue. Two scans are usually performed simultaneously with either two energies or rapid switching between two energies. When the

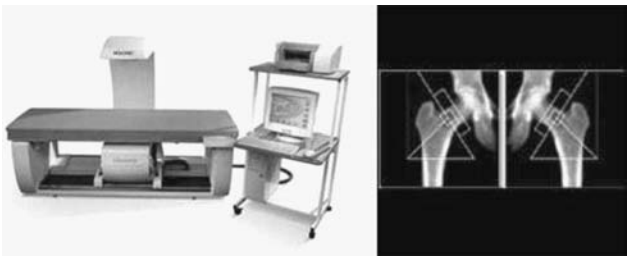


Figure 2. DEXA machine for a whole-body scan (QDR4500 fan-beam scanner, Hologic Inc., Bedford, MA) (left). DEXA bone densitometry is widely used in.

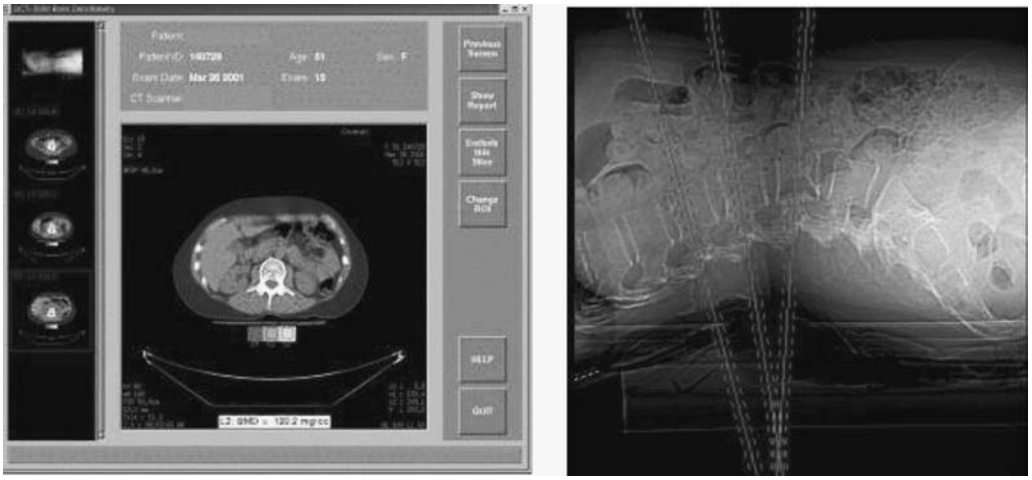


Figure 3. QCT allows selection of the region of interest.

attenuation coefficients for bone and soft tissue are known for both low and high energies, the apparent or area bone mineral density can be calculated as

$$d_b = \frac{(\lambda_s^L/\lambda_s^H)\log(I^H/I_0^H) - \log(I^L/I_0^L)}{\lambda_b^L - \lambda_b^H(\lambda_s^L/\lambda_s^H)} \quad (5)$$

The attenuation coefficient for bone is relative constant but varied to persons. The soft-tissue attenuation coefficient, however, is contributed by fat and other soft tissues and varied greatly in the body. This is the source for the errors generated in the measurement. The manufacturers usually provide phantoms for calibration for the system.

Traditionally, the focus of clinical bone evaluation has been apparent or area BMD as measured by DXA or DEXA (20–23). DEXA provides an effective way to measure BMD in a specific region of interest and is the most widely used diagnostic modality for assessing osteoporosis and osteopenia (8,24–26). However, in particular, DEXA suffers from several shortcomings. Although density (quantity) does positively correlate with strength (27,28) and fracture risk (29–31), anywhere from 10–90% of the variability in bone strength remains unexplained the (32). Additionally, as discussed below, the stereology of trabecular bone is one of its distinguishing features (especially with respect to its mechanical behavior), but because DEXA provides only a 2D image of apparent density, it is inherently limited in this regard. DEXA also suffers from an inability to differentiate trabecular from cortical bone. Although it is true that cortical bone also deteriorates with age (33,34), the effects of bone loss are more prevalent in trabecular bone due to its much higher surface area, and the greater net amount of bone mineral content in cortical bone can conceal small changes in the trabecular bone when measuring only BMD. Nevertheless, as one of the key factors that contribute to the bone's quality evaluation, BMD measured by DEXA is a most popular modality used in assessing the status of bone and the risk of fracture.

Quantitative Computed Tomography (QCT)

QCT provides the true volumetric 3D bone density ($mg \cdot cm^{-3}$) compared with the 2D apparent or aeral density measurement with DEXA (35–43). Because of its high resolution, QCT can provide the measurement in the trabecular region (e.g., femoral neck and vertebral bodies) (39,40,44,45). Compared with DXA the advantage of QCT is the image-based cross-sectional anatomy, which allows for a selection of the region of interests (ROI) and a better assessment of geometrical properties (Fig. 3). Most CT systems provide a software package to automate the placement of the ROI within a particular body volume, e.g., vertebral bodies. QCT scans are generally performed using a single kilovolt setting (single-energy QCT). It is possible to use a dual-energy QCT, which can provide further improvement of the resolution, but at the price of poorer precision and higher radiation dose. New 3D volumetric techniques acquire datasets with which analysis of bone macroarchitecture may be further optimized. Due to its capability of high resolution, geometric and structural parameters determined in QCT may contribute to determine bone strength when integrated with other technology (i.e., finite element analysis). The advantage of spinal QCT is the high responsiveness of the vertebral trabecular bone to aging and disease, whereas the principal disadvantage is the cost of the equipment and the dosage received for the scanning (higher than DEXA).

Peripheral QCT (pQCT)

pQCT systems are available for measuring the forearm. The advantages of these devices are the capability of separating the trabecular and cortical bone of the ultradistal radius and of reporting volumetric density. Several clinical used pQCT devices are available (e.g., the Stratec XCT 2000, that are suitable for use in a physician's office or in primary care).

QUANTITATIVE ULTRASONOMETRY

Quantitative ultrasound (QUS) for measuring the peripheral skeleton has raised considerable interest in recent years. New methods have emerged with the potential to estimate trabecular bone modulus more directly. QUS provides an intriguing method for characterizing the

Table 2. Summary of Current QUS Devices for Calcaneus

Device	Performance	Resolution	Predict Parameter	Cost, $K
Sahara (Hologic)	Index	Nonimage	Z score	20–25
QUS-2 (Metra Biosystem)	Index	Nonimage	Z score	20–25
UBA 575 (Walker Sonix)	Index	Nonimage	Z score	20–25
Achilles (GE-Lunar)	Index + image	Image, 5 mm	Z score	40–50
UBIS 5000 (DMS)	Index + image	Image, 2 mm	Z score	30–35
DTU-one (Osteometer)	Index + image	Image, 2 mm	Z score	25–30
New SCAN	Image + index	Image, 1 mm	Stiffness, BMD, Z	20

material properties of bone in a manner that is noninvasive, nonionizing, nondestructive, and relatively accurate. The primary advantage of QUS is that it is capable of measuring not only bone quantity (e.g., BMD), but also bone quality (i.e., estimation of the mechanical property) of bone. Over the past 15 years, several research approaches have been developed to quantitate bone mass and structural stiffness using QUS (46–48). Preliminary results for predicting osteoporosis using QUS are promising, and it has great potential for widespread applications (including screening for prevention). As such, many QUS machines have been developed, and there are currently many different devices on the market. Most available systems measure the calcaneus using plane waves that use either water or gel coupling [e.g., Sahara (Hologic Inc., MA), QUS-2 (Metra Biosystems Inc., CA), Paris (Norland Inc., WI), and UBA 575 (Walker Sonix Inc., USA)] (Table 2). Recently, an image-based bone densitometry device for calcaneous ultrasound measurement is also made available using an array of plane ultrasound wave (GE-Lunar, Inc., USA). Using several available clinical devices, studies in vivo have shown the ability of QUS to discriminate patients with osteoporotic fractures from age-matched controls (49–51). It has been demonstrated that QUS predicts risk of future fracture generally as well as DEXA (51–54). However, there are several noted limitations, including the tissue boundary interaction, the nonlinear function of density associated with bone ultrasonic attenuation, the single index covering a broad range of tissues (including the cortical and trabecular regions), and the interpolation of the results. Recently, a focused ultrasound sonometer device was developed to obtain the likelihood of a broadband ultrasound attenuation (BUA) image in the human calcaneus region (center frequency 0.5 MHz, focus 50 mm) (55,56) (UBIS 5000, Diagnostic Medical Systems; and DTU-one, Osteometer MediTech). These devices provide ultrasound images in the calcaneus region, in which the parameter compares with DEXA data. Perhaps the major drawbacks of these ultrasound osteometers are low resolution and lack of physical interrelation with meaningful bone strength. Although only showing the correlation between BUA data and BMD, these devices mostly provide qualitative information for assessment of osteoporosis, not the true prediction for bone structural and strength properties. Therefore, QUS remains at a stage as a screening tool (Fig. 4), because of the nonuniformity of the porous structure in the bone tissue and its associated effects in resolution (14). Research attention is focused on developing systems to provide true images reflecting the bone's structural and strength properties at multiple skeletal sites, i.e., in the hip, which can provide a true diagnostic tool (instead of just for screening) that surpasses the radiation based DEXA machines.

If QUS bone densitometry can be developed to provide a "true" bone quality parameter-based diagnostic tool (i.e., directly related to the bone's structural and strength properties) and to target multiple and critical skeletal sites (e.g., hip and distal femur), QUS would have a greater impact on the diagnosis of bone diseases (e.g., osteoporosis) than current available bone densitometry. Research efforts are made in this regard (55–59). As an example, a new QUS modality, called the scanning confocal acoustic diagnostic system, has been developed (57–60), which is intended to provide true images reflecting the bone's structural and strength properties at a particular skeletal site at a peripheral limb and potentially at deep tissue like great trachanter. The technology may further provide both density and strength assessment in the region of interests for the risk of fracture (57–60).

Fundamental QUS Parameters in Bone Measurement

In an effort to use QUS for predicting bone quality, a variety of approaches have been explored with many studies published in the past decade, that have examined the utility of QUS and its potential application as a diagnostic tool for osteoporosis. The physical mechanisms of ultrasound applied to bone may include several fundamental approaches, [i.e., speed of sound (SOS) or ultrasonic wave propagating velocity (UV), sound energy attenuation

Figure 4. A QUS bone densitometry test in a heel region. Reproduced courtesy of GE-Lunar Inc.

(ATT), BUA, and critical angle ultrasound parameters] that closely relate to acoustic transmission in a porous structure. Most commonly, parameters for QUS measurement are BUA and SOS, which can be used to identify those persons at risk of osteoporotic fracture as reliably as BMD (52–54,61,62). It has been shown that both BUA and SOS are decreased in persons with risk factors for osteoporosis, that is, primary hyperparathyroidism (63–66), kidney disease (67), and glucocorticoid use (68,69). The proportion of women classified into each diagnostic category was similar for BMD and QUS. Using the World Health Organization (WHO) criteria to classify osteoporosis for BMD measurement using DEXA and QUS testing, approximately one third of postmenopausal women aged 50+ years with clinical risk factors were diagnosed as osteoporotic compared with only 12% of women without clinical risk factors. This suggests that the measurement of QUS with calcaneal BUA and SOS is to some extent the same as the BMD Z-score measurement.

Background of BUA in Trabecular Bone Measurement

BUA and SOS are currently two commonly used methods for QUS measurements, which make it potentially possible to predict bone density and strength. As an ultrasound wave propagates through a medium, BUA measures the acoustic energy that is lost in bone (unit: dB/MHz). The slope at which attenuation increases with frequency is generally between 0.2 and 0.6 MHz, and it characterizes BUA. The slopes of the frequency spectrum may reflect the density and structure of bone. Although relatively little is known about the fundamental interactions that determine ultrasound attenuation in bone, the potential sources contributing to the attenuation include absorption, scattering, diffraction, and refraction (70–73). Although absorption predominates in cortical bone attenuation, the mechanism of BUA in cancellous bone is believed to be scattering (14,74–76). The importance of scattering has been alluded to in the literature. Scattering is also suggested to contribute to the nonlinear variation in BUA with density observed in cancellous bone and a porous medium (77–79).

Background of SOS or UV for Bone Measurement

The strength of trabecular bone is an important parameter for bone quality. *In vitro* studies have correlated the ultrasound velocity with stiffness in trabecular bone samples (80–82). This indicates that ultrasound has the potential to be advantageous over the X-ray based absorptiometry in assessing the quality of bone in addition to the quantity of bone. The mechanism of SOS in predicting bone strength is believed to be due to the fact that the velocity of an ultrasound wave depends on the material properties of the medium through which it is propagating, but it also depends on the mode of propagation. By determining the wave velocity through a bone, the elastic modulus of bone specimens can be evaluated, or at least be approximated (80,83). When ultrasound travels through a porous material, e.g., trabecular bone, it carries information concerning material properties, such as density, elasticity, and architecture. A relationship exists between the ultrasound velocity (unit: m/s) and the material elasticity E and density ρ (14,80)

$$V = \sqrt{E/\rho} \tag{6}$$

The velocity with which ultrasound passes through normal bone is fast and varies depending on whether the bone is cortical or trabecular. Speeds of 2800–3000 $\mathrm{m \cdot s^{-1}}$ are typical in cortical bone, whereas speeds of 1550–2300 $\mathrm{m \cdot s^{-1}}$ are typical in trabecular bone.

It is demonstrated that trabecular bone strength is highly correlated with elastic stiffness (84). With the introduction of QUS, several new diagnostic parameters and experimental results, both *in vitro* and *in vivo*, have shown potential for evaluating not only bone quantity (i.e., BMD), but also bone quality (i.e., structure and strength). Two principal variables, BUA and UV, have been confirmed to identify those persons at risk of osteoporotic fracture as reliably as BMD from DEXA. However, SOS and BUA are related to bone density and strength as well as to trabecular orientation, the proportion of trabecular bone and cortical shell, the composition of organic and inorganic components, and the conductivity of the cancellous structure. Thus, QUS of trabecular bone depends on a variety of factors that contribute to the measured ultrasound parameters.

Other Bone Status Measurement Methods and Motivation to Assess Bone Quality

Beyond bone quantity, the quality (the integrity of its structure and strength) has become an equally or even more important measure to understand the bone structure and mechanical integrity. Most osteoporotic fractures occur in cancellous bone. Therefore, noninvasive assessment of trabecular bone strength and stiffness is extremely important in predicting the quality of the bone. The strength of the trabecular bone mostly depends on the mechanical properties of the bone at the local and bulk tissue level, and on its spatial distribution (i.e., the microarchitecture). A better understanding of the factors that influence bone strength is a key to developing improved diagnostic techniques and more effective treatments. To overcome the current hurdles, to improve the "quality" of the noninvasive diagnostic instrumentations, and to apply the technology for future clinical application, new clinical modality may concentrate in several main areas: (*1*) increasing the resolution, sensitivity, and accuracy in diagnosing osteoporosis through unique methods for improvement of signal/noise ratio; (*2*) directly measuring bone's strength as one of the primary parameters for the risk of fracture; (*3*) generating real-time compatible imaging to identify local region of interest; (*4*) validating structural and strength properties with new modalities; and (*5*) predicting local trabecular and bulk stiffness and microstructure of bone, and generating a physical relationship between measurement and bone quality. In an attempt to achieve these goals, recent advances of emerging technologies are developed primarily for animal studies at this stage. These include high resolution pQCT, micro-MR-derived measures of structure, micro-CT-based BMD, and combined assessment of strength using geometry, density, and computational simulation. These methods

will further lead to a better understanding of the progressive deterioration of bone in aging populations, and ultimately they may provide early prediction of fracture risk and associated musculo-skeletal complications such as osteoporosis.

ACKNOWLEDGMENT

This work has been kindly supported by the National Space Biomedical Research Institute (TD00207 and TD00405 to Y. Qin) through NASA Cooperative Agreement NCC 9-58.

BIBLIOGRAPHY

Cited References

1. Looker AC, Johnson CL. Prevalence of elevated serum transferrin saturation in adults in the United States. Ann Intern Med 1998;129:940–945.
2. Melton LJI. How many women have osteoporosis now?. J Bone Min Res 1995;10:175–177.
3. Melton LJI. Epidemiology of hip fracture: Implications of the exponential increase with age. Bone 1996;18:121S–125S.
4. Wahner HW, Fogelman I. The Evaluation of Osteoporosis: Dual Energy X-Ray Absorptiometry in Clinical Practice. London: 1994.
5. Genant HK. Current state of bone densitometry for osteoporosis. Radiographics 1998;18:913–918.
6. Kanis JA. An update on the diagnosis of osteoporosis. Curr Rheumatol Rep 2000;2:62–66.
7. Melton LJ III, Atkinson EJ, O'Connor MK, O'Fallon WM, Riggs BL. Bone density and fracture risk in men. J Bone Miner Res 1998;13:1915–1923.
8. Melton LJ III, Orwoll ES, Wasnich RD. Does bone density predict fractures comparably in men and women? Osteoporos Int 2001;12:707–709.
9. Sartoris DJ, Resnick D. Current and innovative methods for noninvasive bone densitometry. Radiol Clin North Am 1990;28:257–278.
10. Sartoris DJ, Resnick D. X-ray absorptiometry in bone mineral analysis. Diagn Imaging (San Franc) 1990;12: 108–113,159,183.
11. Lewiecki EM. Clinical applications of bone density testing for osteoporosis. Minerva Med 2005;96:317–330.
12. Blake G. M, Gluer CC, Fogelman I. Bone densitometry: Current status and future prospects. Br J Radiol 1997;70: Spec No:S177–S186.
13. Blake GM, Fogelman I. Bone densitometry and the diagnosis of osteoporosis. Semin Nucl Med 2001;31:69–81.
14. Njeh CF, Hans D, Fuerst T, Gluer C-C, Genant HK. Quantitative Ultrasound Assessment of Osteoporosis and Bone Status. Munich: 1999.
15. Nord RH. Technical consideration in DPA. In: Genany HK, editor. Osteoporosis Updates 1987. 1987:203–212.
16. Dunn WL, Wahner HW, Riggs BL. Measurement of bone mineral content in human vertebrae and hip by dual photon absorptiometry. Radiology 1980;136:485–487.
17. Blake G. M, Fogelman I. Dual energy x-ray absorptiometry and its clinical applications. Semin Musculoskelet Radiol 2002;6:207–218.
18. Blake G. M, Fogelman I. Methods and clinical issues in bone densitometry and quantitatie ultrasonometry. 1573–1585, 2002.
19. Blake G. M, Fogelman I. Fracture prediction by bone density measurements at sites other than the fracture site: The contribution of BMD correlation. Calcif Tissue Int 2005;76:249–255.
20. Kanis JA. Diagnosis of osteoporosis and assessment of fracture risk. Lancet 2002;359:1929–1936.
21. Kanis JA. Assessing the risk of vertebral osteoporosis. Singapore Med J 2002;43:100–105.
22. Kanis JA, Borgstrom F, Zethraeus N, Johnell O, Oden A, Jonsson B. Intervention thresholds for osteoporosis in the UK. Bone 2005;36:22–32.
23. Kanis JA, Borgstrom F, De Laet C, Johansson H, Johnell O, Jonsson B, Oden A, Zethraeus N, Pfleger B, Khaltaev N. Assessment of fracture risk. Osteoporos Int 2005;16:581–589.
24. Melton LJ III, Atkinson EJ, O'Connor MK, O'Fallon WM, Riggs BL. Determinants of bone loss from the femoral neck in women of different ages. J Bone Miner Res 2000;15:24–31.
25. Melton LJ III, Kanis JA, Johnell O. Potential impact of osteoporosis treatment on hip fracture trends. J Bone Miner Res 2005;20:895–897.
26. Vokes TJ, Favus MJ. Noninvasive assessment of bone structure. Curr Osteoporos Rep 2003;1:20–24.
27. Keaveny TM, Morgan EF, Niebur GL, Yeh OC. Biomechanics of trabecular bone. Annu Rev Biomed Eng 2001;3:307–333.
28. Keaveny TM, Yeh OC. Architecture and trabecular bone—toward an improved understanding of the biomechanical effects of age, sex and osteoporosis. J Musculoskelet Neuronal Interact 2002;2:205–208.
29. Johnston CC Jr, Slemenda CW. Risk assessment: Theoretical considerations. Am J Med 1993;95:2S–5S.
30. Johnston CC Jr, Slemenda CW. Peak bone mass, bone loss and risk of fracture. Osteoporos Int 1994;4 (Suppl 1):43–45.
31. Johnston CC Jr, Hui S. Absolute versus relative fracture risk. J Bone Miner Res 2005;20:704.
32. Hans D, Fuerst T, Lang T, Majumdar S, Lu Y, Genant HK, Gluer C. How can we measure bone quality? Baillieres Clin Rheumatol 1997;11:495–515.
33. Dempster DW, Ferguson-Pell MW, Mellish RW, Cochran GV, Xie F, Fey C, Horbert W, Parisien M, Lindsay R. Relationships between bone structure in the iliac crest and bone structure and strength in the lumbar spine. Osteoporos Int 1993;3:90–96.
34. Dempster DW, Cosman F, Kurland ES, Zhou H, Nieves J, Woelfert L, Shane E, Plavetic K, Muller R, Bilezikian J, Lindsay R. Effects of daily treatment with parathyroid hormone on bone microarchitecture and turnover in patients with osteoporosis: A paired biopsy study. J Bone Miner Res 2001;16:1846–1853.
35. Laib A, Hauselmann HJ, Ruegsegger P. In vivo high resolution 3D-QCT of the human forearm. Technol Health Care 1998; 6:329–337.
36. Lang T, Augat P, Majumdar S, Ouyang X, Genant HK. Noninvasive assessment of bone density and structure using computed tomography and magnetic resonance. Bone 1998;22:149S–153S.
37. Lang TF, Keyak JH, Heitz MW, Augat P, Lu Y, Mathur A, Genant HK. Volumetric quantitative computed tomography of the proximal femur: Precision and relation to bone strength. Bone 1997;21:101–108.
38. Lang TF, Augat P, Lane NE, Genant HK. Trochanteric hip fracture: Strong association with spinal trabecular bone mineral density measured with quantitative CT. Radiology 1998;209:525–530.
39. Lang TF, Li J, Harris ST, Genant HK. Assessment of vertebral bone mineral density using volumetric quantitative CT. J Comput Assist Tomogr 1999;23:130–137.
40. Lang TF, Guglielmi G, Kuijk Cvan, De Serio A, Cammisa M, Genant HK. Measurement of bone mineral density at the spine and proximal femur by volumetric quantitative com-

puted tomography and dual-energy X-ray absorptiometry in elderly women with and without vertebral fractures. Bone 2002;30:247–250.
41. Ruegsegger P, Stebler B, Dambacher M. Quantitative computed tomography of bone. Mayo Clin Proc 1982;57 (Suppl):96–103.
42. Ruegsegger P. Quantitative computed tomography at peripheral measuring sites. Ann Chir Gynaecol 1988;77:204–207.
43. Ruegsegger P, Steiger P, Felder M. Quantitative computed tomography of the rheumatic knee. Clin Rheumatol 1988; 7:486–491.
44. Cann CE, Genant HK, Kolb FO, Ettinger B. Quantitative computed tomography for prediction of vertebral fracture risk. Bone 1985;6:1–7.
45. Cann CE. Quantitative CT for determination of bone mineral density: A review. Radiology 1988;166:509–522.
46. Ashman RB, Cowin SC, Van Buskirk WC, Rice JC. A continuous wave technique for the measurement of the elastic properties of cortical bone. J Biomech 1984;17:349–361.
47. Ashman RB, Corin JD, Turner CH. Elastic properties of cancellous bone: measurement by an ultrasonic technique. J Biomech 1987;20:979–986.
48. Ashman RB, Rho JY. Elastic modulus of trabecular bone material. J Biomech 1988;21:177–181.
49. Cheng S, Tylavsky F, Carbone L. Utility of ultrasound to assess risk of fracture. J Am Geriatr Soc 1997;45:1382–1394.
50. Gregg EW, Kriska AM, Salamone LM, Roberts MM, Anderson SJ, Ferrell RE, Kuller LH, Cauley JA. The epidemiology of quantitative ultrasound: A review of the relationships with bone mass, osteoporosis and fracture risk. Osteoporos Int 1997;7:89–99.
51. Njeh CF, Boivin CM, Langton CM. The role of ultrasound in the assessment of osteoporosis: A review. Osteoporos Int 1997;7:7–22.
52. Bauer DC, Gluer CC, Cauley JA, Vogt TM, Ensrud KE, Genant HK, Black DM. Broadband ultrasound attenuation predicts fractures strongly and independently of densitometry in older women. A prospective study. Study of Osteoporotic Fractures Research Group. Arch Intern Med 1997;157:629–634, 3–24.
53. Hans D, Schott AM, Meunier PJ. Ultrasonic assessment of bone: A review. Eur J Med 1993;2:157–163.
54. Hans D, Schott AM, Arlot ME, Sornay E, Delmas PD, Meunier PJ. Influence of anthropometric parameters on ultrasound measurements of Os calcis. Osteoporos Int 1995;5:371–376.
55. Laugier P, Fournier B, Berger G. Ultrasound parametric imaging of the calcaneus: *In vivo* results with a new device. Calcif Tissue Int 1996;58:326–331.
56. Laugier P, Droin P, Laval-Jeantet AM, Berger G. In vitro assessment of the relationship between acoustic properties and bone mass density of the calcaneus by comparison of ultrasound parametric imaging and quantitative computed tomography. Bone 1997;20:157–165.
57. Qin Y-X, Lin W, Rubin C. Interdependent relationship between Trabecular bone quality and ultrasound attenuation and velocity using a scanning confocol acoustic diagnostic system. J Bone Min Res 2001;16:S470–S470.
58. Qin Y-X, Lin W, Mittra E, Mueller R, Xia Y, Rubin C. Noninvasive assessment of bone quality and quantity using confocal acoustic scanning on *ex-vivo* trabeculae. Ann Biomed Eng. In press.
59. Qin Y-X, Xia Y, Lin W, Chadha A, Gruber B, Rubin C. Assessment of bone quantity and quality in human cadaver calcaneus using scanning confocal ultrasound and DEXA measurements. J Bone Min Res 2002;17:S422–S422.
60. Xia Y, Lin W, Qin Y. The influence of cortical end-plate on broadband ultrasound attenuation measurements at the human calcaneus using scanning confocal ultrasound. J Acoustic Soc Am 2005;118:1801–1807.
61. Frost ML, Blake GM, Fogelman I. Contact quantitative ultrasound: An evaluation of precision, fracture discrimination, age-related bone loss and applicability of the WHO criteria. Osteoporos Int 1999;10:441–449.
62. Frost ML, Blake GM, Fogelman I. Quantitative ultrasound and bone mineral density are equally strongly associated with risk factors for osteoporosis. J Bone Miner Res 2001;16:406–416.
63. Gomez AC, Schott AM, Hans D, Niepomniszcze H, Mautalen CA, Meunier PJ. Hyperthyroidism influences ultrasound bone measurement on the Os calcis. Osteoporos Int 1998;8:455–459.
64. Guo CY, Thomas WE, al Dehaimi AW, Assiri AM, Eastell R. Longitudinal changes in bone mineral density and bone turnover in postmenopausal women with primary hyperparathyroidism. J Clin Endocrinol Metab 1996;81:3487–3491.
65. Minisola S, Scarnecchia L, Carnevale V, Bigi F, Romagnoli E, Pacitti MT, Rosso R, Mazzuoli GF. Clinical value of the measurement of bone remodelling markers in primary hyperparathyroidism. J Endocrinol Invest 1989;12:537–542.
66. Minisola S, Rosso R, Scarda A, Pacitti MT, Romagnoli E, Mazzuoli G. Quantitative ultrasound assessment of bone in patients with primary hyperparathyroidism. Calcif Tissue Int 1995;56:526–528.
67. Wittich A, Vega E, Casco C, Marini A, Forlano C, Segovia F, Nadal M, Mautalen C. Ultrasound velocity of the tibia in patients on haemodialysis. J Clin Densitometry 1998;1:157–163.
68. Blanckaert F, Cortet B, Coquerelle P, Flipo RM, Duquesnoy B, Marchandise X, Delcambre B. Contribution of calcaneal ultrasonic assessment to the evaluation of postmenopausal and glucocorticoid-induced osteoporosis. Rev Rhum Engl Ed 1997;64:305–313.
69. Cortet B, Flipo RM, Blanckaert F, Duquesnoy B, Marchandise X, Delcambre B. Evaluation of bone mineral density in patients with rheumatoid arthritis. Influence of disease activity and glucocorticoid therapy. Rev Rhum Engl Educ 1997;64:451–458.
70. Madsen EL, Dong F, Frank GR, Garra BS, Wear KA, Wilson T, Zagzebski JA, Miller HL, Shung KK, Wang SH, Feleppa EJ, Liu T, O'Brien WD Jr, Topp KA, Sanghvi NT, Zaitsev AV, Hall TJ, Fowlkes JB, Kripfgans OD, Miller JG. Interlaboratory comparison of ultrasonic backscatter, attenuation, speed measurements. J Ultrasound Med 1999;18:615–631.
71. Wear KA, Garra BS. Assessment of bone density using ultrasonic backscatter. Ultrasound Med Biol 1998;24: 689–695.
72. Wear KA. Frequency dependence of ultrasonic backscatter from human trabecular bone: theory and experiment. J Acoust Soc Am 1999;106:3659–3664.
73. Wear KA, Stuber AP, Reynolds JC. Relationships of ultrasonic backscatter with ultrasonic attenuation, sound speed and bone mineral density in human calcaneus. Ultrasound Med Biol 2000;26:1311–1316.
74. Strelitzki R, Evans JA. An investigation of the measurement of broadband ultrasonic attenuation in trabecular bone. Ultrasonics 1996;34:785–791.
75. Strelitzki R, Evans JA. Diffraction and interface losses in broadband ultrasound attenuation measurements of the calcaneum. Physiol Meas 1998;19:197–204.
76. Strelitzki R, Metcalfe SC, Nicholson PH, Evans JA, Paech V. On the ultrasonic attenuation and its frequency dependence

in the os calcis assessed with a multielement receiver. Ultrasound Med Biol 1999;25:133–141.
77. Aindow JD, Chivers RC. Ultrasonic wave fluctuations through tissue: An experimental pilot study. Ultrasonics 1988;26:90–101.
78. Chivers RC. The scattering of ultrasound by human tissues–some theoretical models. Ultrasound Med Biol 1977;3:1–13.
79. Chivers RC, Parry RJ. Ultrasonic velocity and attenuation in mammalian tissues. J Acoust Soc Am 1978;63:940–953.
80. Ashman RB, Rho JY, Turner CH. Anatomical variation of orthotropic elastic moduli of the proximal human tibia. J Biomech 1989;22:895–900.
81. McKelvie ML, Palmer SB. The interaction of ultrasound with cancellous bone. Phys Med Biol 1991;36:1331–1340.
82. Turner CH, Eich M. Ultrasonic velocity as a predictor of strength in bovine cancellous bone. Calcif Tissue Int 1991;49:116–119.
83. Rho JY, Ashman RB, Turner CH. Young's modulus of trabecular and cortical bone material: Ultrasonic and microtensile measurements. J Biomech 1993;26:111–119.
84. Hou FJ, Lang SM, Hoshaw SJ, Reimann DA, Fyhrie DP. Human vertebral body apparent and hard tissue stiffness. J Biomech 1998;31:1009–1015.

See also BONE AND TEETH, PROPERTIES OF; COMPUTED TOMOGRAPHY.

BONE UNUNITED FRACTURE AND SPINAL FUSION, ELECTRICAL TREATMENT OF

AR LIBOFF
Oakland University
Rochester, Michigan

DEFINING THE UNUNITED FRACTURE

Within hours following a fracture in bone and the rapidly resulting hematoma, an endogenous repair process is initiated, characterized by increased cell division in the periosteum and endosteal stem-cell differentiation leading to organization of the hematoma into fibrocartilaginous callus. The latter represents the source of osteogenic potential from which ossification and subsequent bone remodeling occurs. In the ideal case, with proper management, those suffering bone fractures will normally find themselves fully recovered within a few months, with this time varying according to the specific bone involved, the type of fracture, and the age of the patient.

However, a small percentage of fractures fall outside the norm and do not heal as readily. There are upward of 5 million fractures occurring each year in the United States (1). Approximately 5–10% of these remain ununited after a few months. One can identify two types of ununited fractures, those undergoing *delayed* fracture healing, as evidenced by a lack of full healing in 3–6 months, and *nonunions*, where there is a lack of healing 6–12 months after the fracture has occurred. Marsh (2) suggests that the best measure of fracture healing in humans may be recovery of bending stiffness (i.e., the torque measured in Newton-meters that will bend bone by 1°). He defines delayed union as failure to reach a stiffness of of 7 $N \cdot m \cdot deg^{-1}$ at 20 weeks following fracture. Both the periosteum and the *end*osteum are deeply involved in the process of fracture healing, and it has been suggested (2) that delayed healing may be the result of cessation of the periosteal response before bridging has occurred, while nonunion may be indicative of a breakdown of both the periosteal and the endosteal repair mechanisms. A more general term for nonunion is pseudarthrosis, or false joint. Worth noting is that this problem is also infrequently found at birth (congenital pseudarthrosis). Electric and electromagnetic treatment is prescribed for both types of psuedarthroses, those that are the result of ununited fractures, and those that are found at birth. Further, because spinal fusion following back surgery can be problematic, electromagnetic treatment is also being used as an adjunctive procedure to promote spine fusion (3).

THE ELECTRIC CHARACTER OF BONE

Bone has a number of remarkable physical properties, particularly its electric character. Its electrical properties and the intimate relation of these properties to the growth process in bone were brought to light in a series of experimental discoveries, beginning in the 1950s. These revealed

1. A piezoelectric effect in bone.
2. A striking bioelectric signature specifically associated with developing bone.
3. A characteristic signature in adult unstressed bone.
4. A characteristic bioelectric signature following bone fracture.

Piezoelectricity is the rather unique property in which mechanical force is transformed into electric polarization (Fig. 1). Bone was shown to be piezoelectric by Yasuda in the early 1950s, but the work leading to this conclusion was not made generally available until 1957(4). Fukada and Yasuda (5) later found that this property could be traced to the intrinsic collagen component in bone. Since that time, a number of observers (6–8) suggested that this mechanical stress–electric polarization property should more properly be referred to as a stress-generated potential (SGP), reflecting the fact that what actually happens may not be the result of the special sort of crystal or textural structure that underlies the piezoelectric effect, but might instead result from the well-known electrokinetic effect of streaming potential. Streaming potentials, similar to piezoelectric signals, are characterized by the transformation of mechanical stress into a potential difference. However, streaming potentials do not occur because of any intrinsic crystal structure, but rather because fluid displacement through porous materials or tubes results in electric charge separation. It is generally agreed that dry bone indeed exhibits piezoelectricity, but opinions vary on whether this effect actually plays a role when bone is in its usual (i.e., wet) physiological environment. Part of the difficulty in resolving this issue is that the piezoelectric effect is not easily measured in wet bone. Whatever the pros and cons concerning studies on wet bone, it is difficult to put aside the seminal experiment by McElhaney (9). More than 600 silver epoxy electrodes were attached to cover the surface of a dried intact human femur from autopsy, and a vertical

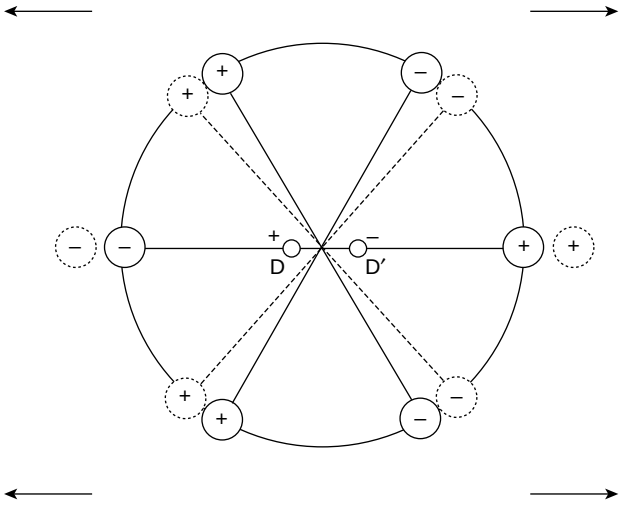


Figure 1. Piezoelectric Effect. Here, a tensile force results in a net electrical polarization in a material that ordinarily does not exhibit any polarization. Note that a compressive force will also result in electrical polarization. The source of the piezoelectric effect in bone is collagen.

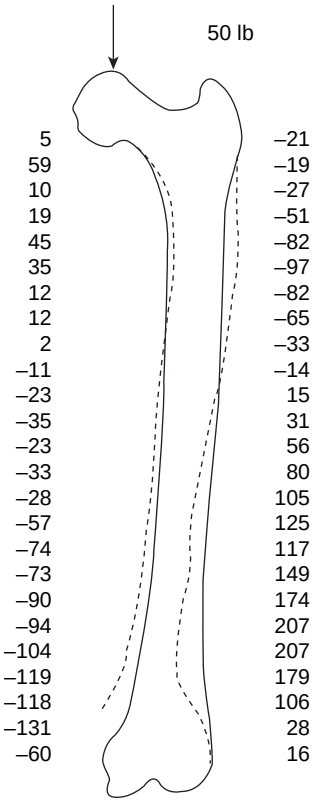


Figure 2. Map of piezoelectric voltages in dry stressed femur. When a 50-lb (220 N) load is impressed on dry human femur, piezoelectric voltages appear over the entire surface. One such set, given in millivolts (mV), is shown (9). The interpretation by Marino and Becker (10) is that the locus of these voltages, as shown by the dotted line for one slice through the femur, corresponds to the way that the bone will remodel under a specific load.

mechanical load was applied to the proximal end of the entire femur, mimicking the femur's weight bearing function. This load produced piezoelectric potentials from each of the electrode points, in effect mapping the piezoelectric response of the entire bone to this load. The voltages obtained varied widely in intensity, and included both positive and negative signs. These results were interpreted by Marino and Becker (10) as showing that, if one assumes that negative potentials tend to activate osteoblasts and positive voltages act to enhance osteoclastic function, then the voltage map (Fig. 2) represents the locus of the new remodeling surface for the femur: Areas of negative polarity are found where the femur needs thickening and areas of positive polarity are located where the bone must be reduced in thickness. Thus the potential remodeling response of the femur to the applied load is related in a very direct way to the polarity and intensity distribution of the piezoelectric signal. The McElhaney experiment showed convincingly that the piezoelectric effect in bone, *in the dry state*, conveys the information necessary to provide a remodeling template for bone under mechanical stress, in effect explaining Wolff's law (11), the empirical statement that bone remodeling follows the distribution of forces applied to the bone. Nevertheless, it is conceivable that the locus of voltages supplied by the piezoelectric effect in bone also requires local electrokinetic potentials to implement the remodeling process at the cellular level, either through cellular differentiation to produce the required osteoblasts and osteoclasts necessary for bone remodeling, or perhaps to separate the osteoblasts and osteoclasts by galvanotaxis (12).

Even in the absence of mechanical stress, bone exhibits a variety of intrinsic electric signals. One such effect is apparently part of the growth and development process. Measuring the electric potential in the same way for the same vertebral element from a group of cadavers covering a wide range of ages, Athenstaedt (13) found that this voltage was clearly connected to the age of the individual, greatest in infancy and ultimately falling to a level voltage plateau with maturity. The implication is that electric polarization in bone plays a role in the growth process. Something similar happens in long bone. One can measure voltage differences, usually referred to as bioelectric potential (BEP) along the length of a long bone (14) (Fig. 3). The BEP is always a relative measurement, where, for example, one can fix one electrode at one end (the epiphysis) and measure the potential difference at various points along the shaft of the bone (the diaphysis). Particular attention has focused on the growth plate, that region between epiphysis and diaphysis where the bone actually is ossifying as it grows. The BEP measured at the growth plate relative to the epiphysis in immature, growing, bone is markedly negative by as much as 5 mV (15), but as growth ceases, this potential difference becomes less pronounced. Furthermore, it has been demonstrated by means of tetracycline labeling (16) that the formation of new bone corresponds closely with the BEP profile.

Bone also exhibits an intrinsic electrical character even if it is not actively growing or under mechanical stress. For example, a BEP profile is also found in adult bone, albeit with a different signature. In measurements of this voltage, it is observed that the proximal metaphysis is always

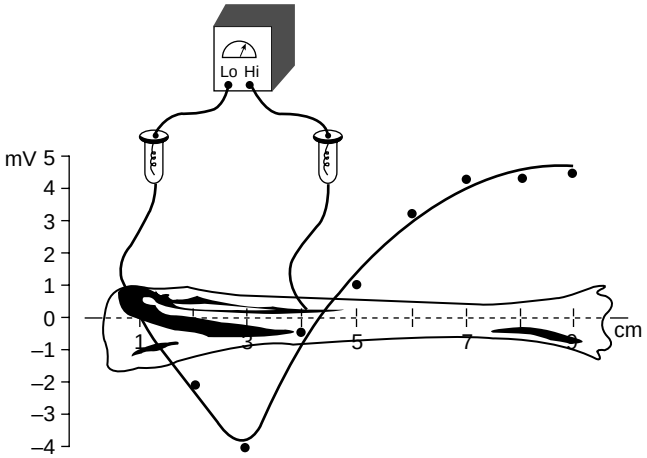


Figure 3. The BEP Profile. Bioelectric potential profile of a rabbit tibia (14). Voltage differences are obtained relative to the proximal end (on the left) by means of salt bridge electrodes.

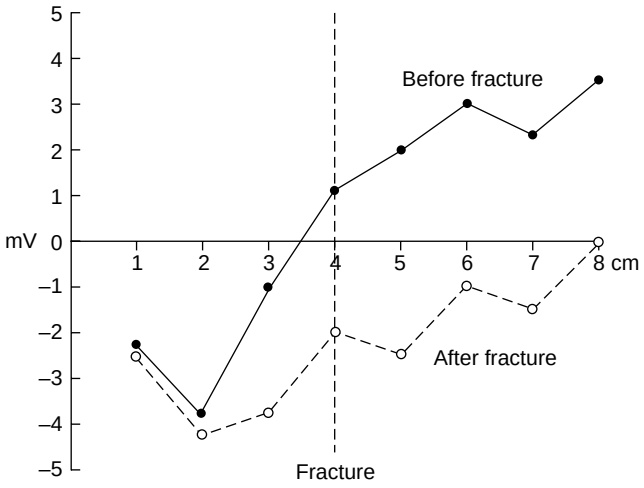


Figure 4. Effect of fracture of an 8 cm rabbit tibia on BEP. Measured potentials are shown before and after the fracture, which is at the 4 cm point.

negative with respect to the midshaft and distal portions of the bone. Because the BEP is unaffected by local nerve denervation or reduced blood flow, but slowly disappears following animal death, it is believed (17) that the origin of this voltage stems from functioning bone cells, acting in concert.

Other than this likely connection to bone cells, it is difficult to pin down a reasonable physical explanation for the ubiquitous potential profile associated with long bone. Electric polarization is readily observed in specimens of mature bone when they are even slightly heated. The origin of this effect is still unclear, but it may reflect a pyroelectric response having a textural origin (18), or perhaps, as Mascarenhas has suggested (19), bone is inherently an electret, a type of material, like many bioploymers, with the interesting property of being capable of storing electric charge. Electrets are the electrical equivalent of magnets, and some observers have suggested that bone exhibits ferroelectric properties. The characteristic property seen in electrets is a slow release of charge when heated. For example, long-term currents on the order of 100 fA can be observed (20) for bone specimens heated to 40 °C. Regardless of the cause, it is most likely the case, as stated by Brighton (21), that: *. . .in living, non-stressed bone, areas of active growth...[are] electronegative when compared with less active areas*.

There is one more impressive electric property associated with living bone, again reflecting this question of the role of negative potentials. Only a few hours following bone fracture, the bone becomes more negative relative to the prefracture BEP (22) (Fig. 4). There is some dispute as to whether this effect is limited to the fracture site or is distributed more widely along the length of the bone (14,23). This uncertainty is in all likelihood due to the fact that there are obvious measurement problems in obtaining a BEP profile for a fractured bone. As an injury current one might expect a more specific and localized expression. However, it is possible that the entire periosteum may be affected in a bone fracture at any point along its length. Worth noting are the experiments by Becker and Murray (24) on fracture healing in amphibian systems indicating a discrete electrical negativity at the fracture site, which led him to characterize the innate ability of bone to heal itself in higher animals as a form of regenerative healing.

Viewed in the context of its other electrical properties, the change in voltage profile associated with bone fracture has to be regarded as consistent with the overarching concept that bone makes extensive use of electricity in all of its growth, repair, loadbearing, and homeostatic processes. Because of this, it is hardly surprising that exogenous electric currents have been widely applied in attempts to grow and/or repair bone.

The FDA-approved devices for electric repair of ununited fractures fall either into invasive or noninvasive categories. The invasive devices make use of implanted direct current (dc) and (ac) electric signal sources, both pulsed and continuously sinusoidal. The noninvasive types are either purely electric (capacitive coupling or CC), or electromagnetic, using pulsed magnetic fields (PMF or PEMF) or ion cyclotron resonance (ICR) tuned magnetic field combinations.

DIRECT CURRENT OSTEOGENESIS

The surgeon who first observed that bone is piezoelectric, Iwao Yasuda, was also the first to demonstrate (25) that electric fields applied to long bone in vivo are capable of producing callus. He wrapped a few turns of wire around rabbit femur, and, maintaining this point at a negative potential, passed a small (1 μA) current to an anode located away from the bone. It was consistently observed that after 3 weeks this current resulted in spicules of osseus callus (called electric callus by Yasuda) (Fig. 5). Surprisingly, these spicules were not directed along the bone itself, but instead along the direction of the current, in some cases actually pointing away from the bone. To the orthopedic surgeon, one of the most positive signs during the course of fracture repair is the appearance of callus. Thus the observation by Yasuda cannot be overemphasized.

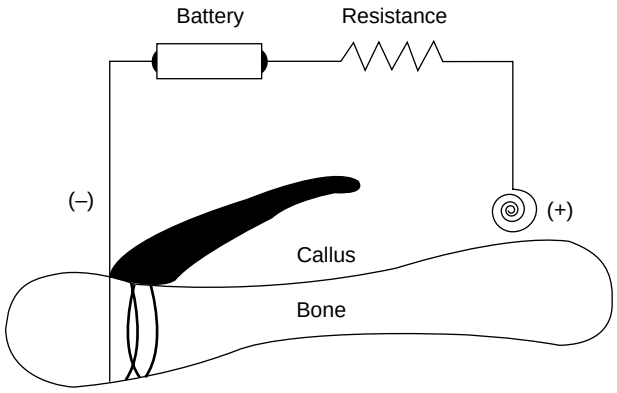


Figure 5. Formation of callus in response to 1 μA dc current. After original sketch by Yasuda (26), showing wires wrapped around the bone.

Although the reasons why electricity is capable of forming callus are still not clear, the implication of Yasuda's work was that electrical stimulation might be of assistance in bringing ununited fractures to closure.

Most of the follow-up experiments to Yasuda's discovery concentrated on determining the effects of electrical signals on normal and fractured bone. A commonly used animal experimental design was to apply an electrical signal to one femur while using the contralateral femur in the same animal as a control. Intrinsic to this approach was the use of dummy electrodes, carrying no current, but serving to affect the contralateral limb by its mere presence in whatever way the electrodes were affecting the activated side. It was in this manner that Bassett et al. (27) used implanted battery packs to deliver microampere-level currents to platinum–iridium wire electrodes extending into the medullary cavities of femora in dogs. The results clearly indicated that more bone was formed in the intermedullary space in the vicinity of the cathodes than near the anodes. Follow-up experiments (28) reinforced the finding that bone growth appeared to be effective at the cathode, but also that bone necrosis occurred at anodes for currents in excess of 20 μA. In this work Friedenberg et al. (28) found that bone growth was most pronounced for currents between 5 and 20 μA. There is some question concerning this optimal current in that the required levels may be dependent on the mode of application. In rabbit femur, circular defects ~ 2.8 mm in diameter were repaired within 3 weeks when subjected to currents ranging between 2.5 and 3 μA applied by two electrodes on either side of the defect (29). Not only was the current lower than that suggested by Friedenberg et al. (28), but there was no particular advantage to either polarity. Similarly, Hambury et al. (30) studying $_{85}$Sr uptake in rabbit femur observed osteogenesis at 3 μA, again with no difference due to polarity. In another attempt (31) to establish the optimal current for repairing bone defects in dog, it was reported that 0.2 μA was more effective than either 2.0 or 20 μA.

Further complicating the issue of what level of current is required to initiate osteogenesis were a number of earlier reports in which callus was formed using currents that were orders of magnitude smaller than microampere levels. Fukada and Yasuda (4) wrapped a charged Teflon electret around bone to initiate callus, work that was later successfully repeated in Japan (32,33). Three different types of current application were employed in the latter experiment: that emitted by an electret, that obtained from the piezoelectric poly-γ-methyl-L-glutamate (PMLG) film, and a battery delivering 8–10 μA. The two current levels for the electret and the film, respectively, were 1 and 10 pA, levels smaller by huge factors of 10^{-7} and 10^{-6} from the "optimal" value of 10 μA. Marino and Becker (34) raised the issue as to whether this enormous difference in currents, both seemingly effective, means that more than one mechanism is involved, with the microampere (μA) results indicative of a nonspecific osteogenic stimulus while the picoamp (pA) currents more closely mimicing the endogenous piezoelectric response.

ELECTROMAGNETIC OSTEOGENESIS

Among his other important discoveries, Michael Faraday was the first to show that voltage is induced in a conductor when a nearby magnetic field is changing rapidly. This phenomenon, often referred to as Faraday's law, can be mathematically expressed by the following expression:

$$dB/dt = -V/A \tag{1}$$

where dB/dt is the time rate of change of the magnetic field B through a region of area A, and V is the voltage induced by dB/dt along the path that is circumferential to A by this rate of change. If B is varying at some frequency f, the product fB is a good measure of the relative effectiveness of dB/dt. When the region in question is electrically conducting, as in living tissue, one can use Ohm's law to rewrite the above expression in terms of the current I instead of V. Thus if R is the resistance of the circumferential path around A, Eq. 1 is changed to read

$$dB/dt = -I(R/A) \tag{2}$$

In this way, one can induce a current in the vicinity of a bone defect by employing a nearby magnetic field that is changing rapidly—the faster the rate of change, the greater the current. One achieves a faster change (i.e., a larger dB/dt) by merely increasing the frequency at which B is changing. Further, it is important to realize that the current so induced is no different from currents that are produced by purely electrical means (Fig. 6). Most important, since the source of the magnetic field can be deployed externally, Faraday's law enables the clinician to generate the required therapeutic currents in a completely noninvasive manner.

In 1974, following 5 years of intensive effort. Bassett and Pilla(35) reported on the successful use of the Faraday induction concept (PMF) to repair fibular osteotomies in beagle. Coils were placed on either side of the leg in such a manner that the magnetic fields from the coils traversed the defect and were additive (Fig. 7). The currents through each coil were pulsed in two ways, at 1 pulse $\cdot$ s^{-1} and at 65 pulses $\cdot$ s^{-1}. As revealed by mechanical testing of the fibula subsequent to treatment, there was greater indication of recovery with 65 pulses $\cdot$ s^{-1}, a finding that was consistent

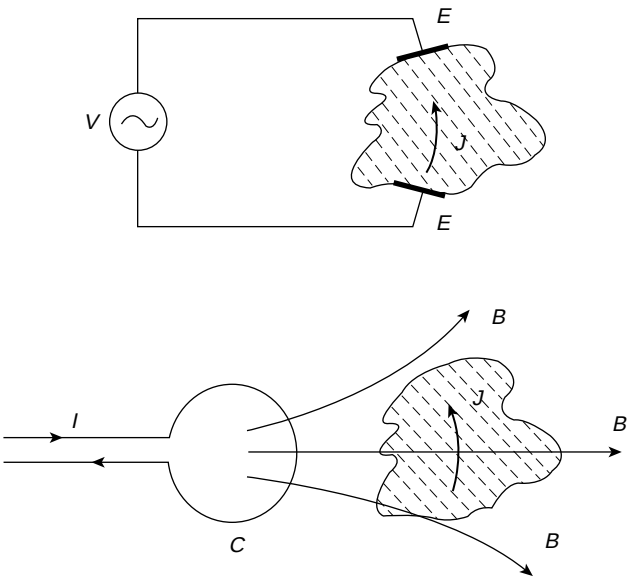


Figure 6. The current density J produced in tissue by a voltage source V acting through electrodes E is no different from the current density induced by a changing magnetic field B according to Faraday's law. The B field is produced by a current I that energizes **a** coil C, whose plane is perpendicular to the page.

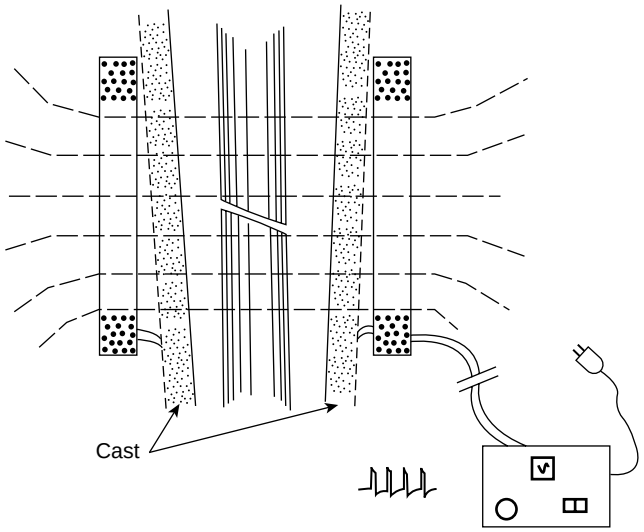


Figure 7. The PMF technique uses two flat coils connected in series to generate a magnetic field (dashed line) through the bone defect. Because the magnetic field is changing rapidly, a voltage is induced, producing a current in the vicinity of the defect. The process is completely noninvasive. In this sketch (36) the two parallel coils, whose planes are perpendicular to the page, are shown outside the cast.

with the prediction from Eq. 2 of a larger current with higher frequency.

INVASIVE (IMPLANTED) ELECTRIC TREATMENTS

Although the use of pulsed magnetic fields provides a means by which one avoids electrode implantation, some surgeons still prefer the extra advantages that come with direct observation of the pseudarthrosis defect. In addition, there is a very lengthy literature background on delivering dc directly to bony defects.

In late 1971, groups at New York University (NYU) and at the University of Pennsylvania independently demonstrated that electrical stimulation using implanted dc devices was successful in repairing pseudoarthrosis defects in humans. In both cases, electrodes and battery were surgically implanted with provisions for percutaneously monitoring the current. Otherwise, however, the methods employed were strikingly different. The NYU group, led by L.S. Lavine (37) used platinum wire electrodes on either side of a congenital pseudarthrosis in the lower tibia of a 14 year old male, in effect allowing the current to pass through the defect (Fig. 8). The polarity of the current was such that the proximal side of the defect was negative. This approach was the same as successfully used in this group's previous experiment (29) to repair defects in rabbit femur. The current was monitored and maintained over the 18-week treatment period at 3.9 μA.

By contrast, Friedenberg et al. (38) in treating a nonunion in the medial malleolus of a 51 year-old woman, used a technique (Fig. 8) that had been previously been found to be successful in producing callus in rabbit fibula (28,39). A stainless steel cathode was located directly in the defect and the anode, an aluminum grid, was taped to the skin. A constant-current power source maintained the current at 10 ± 2 μA over the 9 week treatment period. Again, as with the nonunion treatment employed by the NYU group, the outcome was successful.

These differences in treatment, both leading to repair of the nonunions, remain unresolved. The one treatment (37) is consistent with prior animal work in which the proximal side of bone was found to be intrinsically negative, while the second result (38) fits those observations (24) claiming that fractures are more negative than the rest of the bone. These differences tend to highlight a key difficulty connected to the research on the electric treatment of bone. Apart from the essentially empirical nature of measurements such as the BEP profile, there is no fundamentally sound basis with which to explain the underlying mechanisms, resulting in continuing uncertainties in the clinical techniques.

A number of investigators have attempted to shed light on this question of mechanisms. Almost all such "explanations" have focused on the electrically related regulation of different factors: parathyroid hormone (PTH) (40), adenosine 3', 5'- monophosphate (cAMP) (41,42), insulin-like growth factor II (IGF-II) (43), bone morphogenetic protein (BMP) (44), transforming growth factoř-beta 1 (TGF-β1) (45), and calcium ion channel transport (46). These are contributors, in varying degrees, to the cellular signal transduction pathways controlling bone growth. However, it would be truly surprising if these factors were *not* involved in all types of osteogenic processes, including electrical osteogenesis. At best, such factors must be regarded as merely *indicators* of metabolic activity in bone. At this point in time, they provide little, if any clue as to the reason why bone is responsive to electrical stimulation.

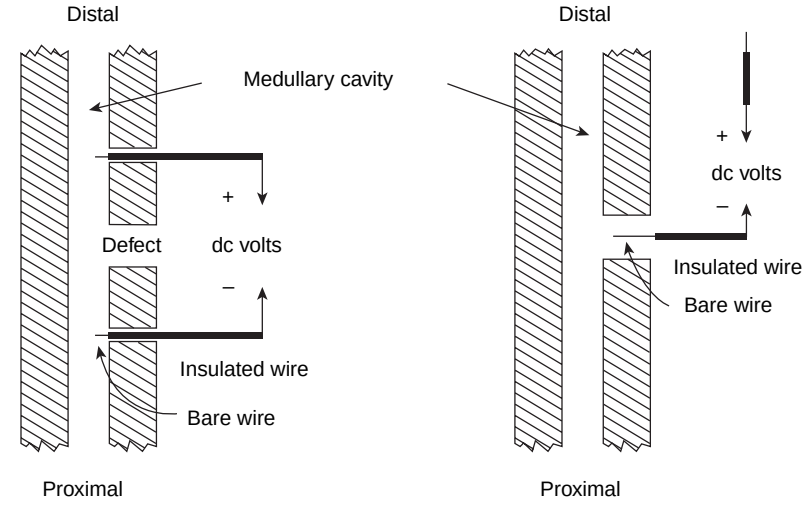


Figure 8. Two ways of using μA-level dc currents to repair defects in bone. In one case, the electrodes are applied so as to bridge the defect. In the other case, the cathode is placed directly into the defect. Both approaches have been successful (37,38) in treating human nonunions.

Presently, there are two FDA-approved implantable direct current devices for treating bony defects, both marketed by ElectroBiology Inc. (Parsippany, NJ). These are shown in Figs. 9 and 10. The Osteogen Bone Growth Stimulator supplies 40 μA through mesh electrodes. Although the cathode is located at the defect, similar to the original placement by Friedenberg et al. (38), the current is far in excess of what was thought to lead to bone necrosis (28). Apparently, the nature of the electrodes used by Friedenberg et al. (28) may have played a role in this discrepancy. The second implantable dc device is the SpF Spinal Fusion Stimulator, prescribed for spinal fusion. The positive and negative leads, in this case carrying 60 μA, are located on either side of the repair site.

NONINVASIVE ELECTRIC TREATMENT: (CC)

One method for applying an electric current to a defect in bone in a noninvasive manner is by means of capacitive coupling (CC). The background for this technique were experiments (47,48) in which 60 kHz sinusoidal voltages were capacitively transferred to bone cell cultures resulting in an electric field within the culture medium of 20 $mV \cdot cm^{-1}$ and a current density of 300 $\mu A \cdot cm^{-2}$. The first clinical use of this was to treat nonunions (49) (Fig. 11). An overall efficacy of 77% was achieved in a group of 22 cases with a mean time to healing of 23 weeks. The FDA-approved version of this technique is marketed by EBI (Fig. 12) As in the earlier studies on cell culture, the pictured device makes use of a 60 kHz alternating electric field that is applied to the skin on either side of the defect using disk electrodes and conducting gel. The current density within the tissue is considerably less than the levels used in cell culture, only $\sim 7\ \mu A \cdot cm^{-2}$. The term *capacitive* may not be warranted for the devices pictured in Figs. 11 and 12. Unlike the lack of ohmic coupling in the earlier, *in vitro* studies, there is a much larger ohmic contribution to the overall impedance when disk electrodes are used. A better description for this device category might be ac (i.e., simply alternating current) instead of CC.

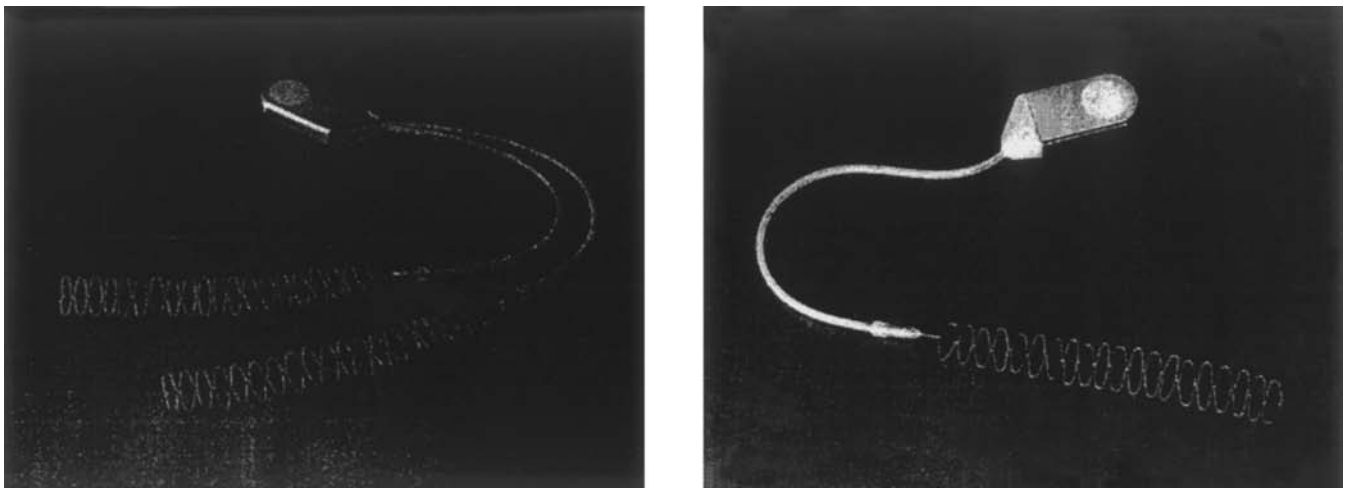


Figure 9. Implantable device for bone growth stimulation. EBI (Electro-Biology, Inc.)

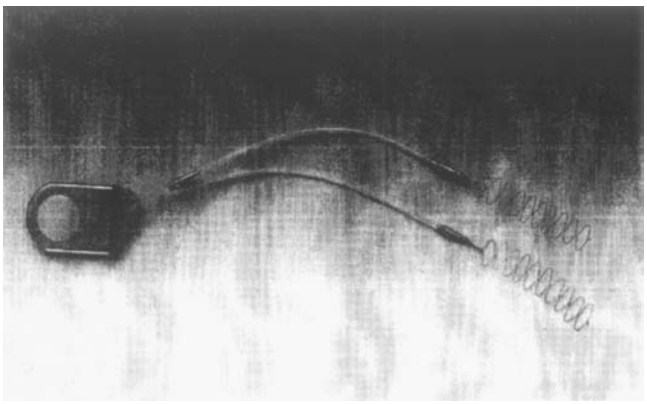


Figure 10. Implantable device for adjunctive treatment of spinal repair (EBI)

A wide range of parameters have been used in studying the clinical and experimental aspects of the CC signal, with various voltages applied to the skin between 1 and 10 V, and frequencies between 20 and 200 kHz. The electric field strengths generated within tissue has ranged from 1 to 100 $mV \cdot cm^{-1}$ and the current densities from 0.5 to 50 $\mu A \cdot cm^{-2}$.

NONINVASIVE ELECTROMAGNETIC DEVICES: (PMF)

The successful use of PMF (also called PEMF) by Bassett et al. (35) to repair bone defects in animals noninvasively led to a number of different devices aimed at applying pulsed magnetic fields. One such early design, successfully applied to the treatment of a tibial nonunion (50,51), made use of an iron-cored electromagnet driven by a square pulse with a repetition rate of 1 pps. However, this design suffered because the large inductive reactance of the iron core acted as a constraint on the repetition rate of the coil current pulses.

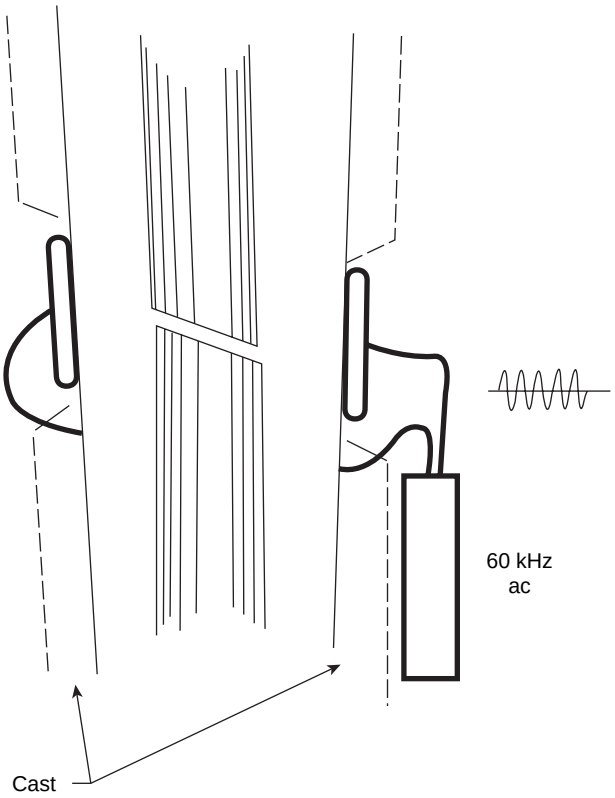


Figure 11. Capacitive Coupling. Electrodes are attached on either side of the bony defect external to skin (here, external to cast) supplying a 60 kHz sinusoidal signal.

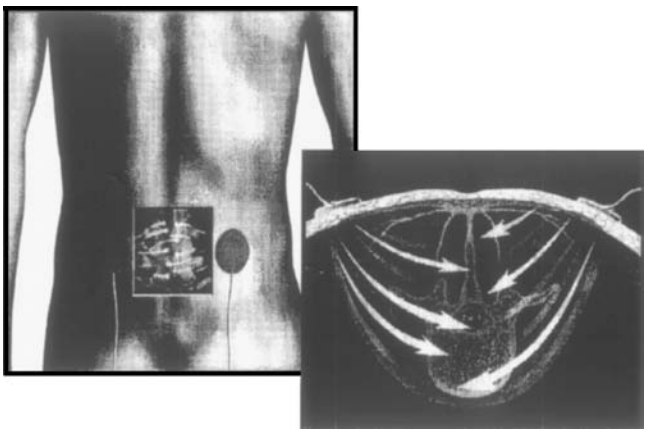


Figure 12. Mode of action of EBI capacitive coupling spinal fusion stimulator (EBI).

With this constraint in mind, Bassett's group succeeded in designing (52) a low inductance air-coil system that could be pulsed at higher frequencies to repair recalcitrant pseudarthroses and nonunions in humans (Figs. 13–16). The success rate that was reported (85%) was greatly in excess of the salvage rate usually obtained by orthopedists using conventional, nonelectrical procedures. However, later (55), reviewing PMF treatments for a wider, all-inclusive group of pseudarthrosis cases, including those with the worst prognosis, Bassett lowered the success rate downward, to 54%.

The pulsed magnetic field that was originally used by Bassett was (and still is) based on the saw-tooth signal common to the fly-back refresher circuit in television receivers. A saw-tooth voltage (Fig. 17) is applied to a pair of many turn coils, creating a current in both coils that generates a single magnetic field. The planes of the coils are roughly parallel, and deployed on opposite sides of the defect (see Fig. 7), creating a commonly directed magnetic field through the defect. The sawtooth signal applied to the coils results in a rapidly changing magnetic field, ~ 10 tesla per second ($T \cdot s^{-1}$), maximized at those times when the voltage applied to the coils is falling sharply. Faraday's law results in the induced voltage shown in Fig. 18. The net induced signal that appears in the vicinity of the defect consists of bursts of 21 pulses, each individual pulse 260 μs in duration, with the bursts repeating at 15 Hz. The magnetic field that actually appears in the area of the defect rises, with each pulse, to ~ 10 G (1 mT), before dropping precipitously in ~ 25 μs. It is this rapid change in B that contributes the most to the induction of current (see Eq. 2). For example, if a sine wave signal of 10 G at 60 Hz were applied to the same region instead of this pulse, the maximum current would be 600 times smaller.

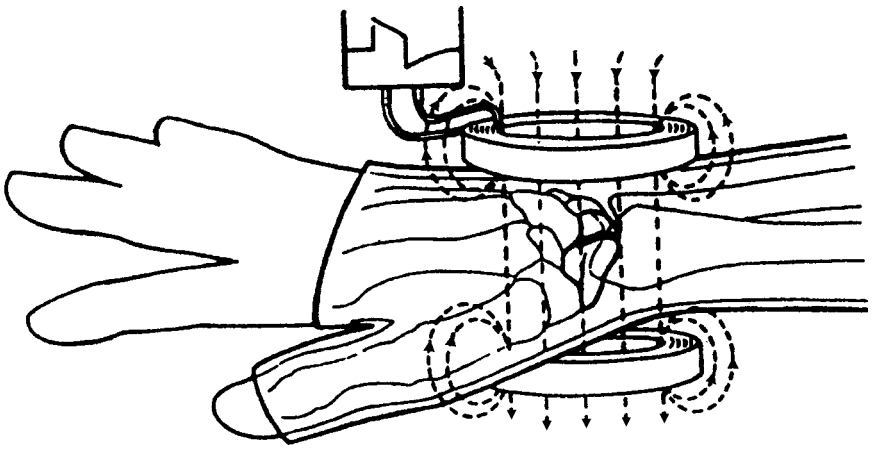


Figure 13. Treatment of ununited scaphoid fracture with PMF (53).

The various PMF clinical and experimental signal repetition rates that have been attempted vary between 1 and 100 Hz, with the maximum magnetic field intensity at the defect site ranging from 0.1 to 30 G, and the induced electric field at the site ranging between 0.01 and 10 $mV \cdot cm^{-1}$.

NONINVASIVE ELECTROMAGNETIC DEVICES (ICR)

Magnetic fields are also used in bone repair in ways that have nothing to do with Faraday induction. It was shown in 1985 (56) that the results embodied in the so-called calcium efflux effect (57,58) were in close agreement with predictions based on the resonance characteristics of certain biological ions subject to the Lorentz force. Specifically, the shape of the nonlinear frequency dependence of calcium binding to chick brain tissue was what might be expected for a particle with the charge-to-mass ratio of the potassium ion moving in combined parallel sinusoidal and dc magnetic fields whose ac frequency and dc intensity corresponded to the ICR condition for K^+. This observation also explained earlier work (59,60) in cell culture demonstrating that weak low frequency magnetic fields enhance DNA synthesis in a manner that is clearly not related to Faraday induction, since the additional DNA synthesis does not scale with either frequency or intensity. Ion cyclotron resonance is a magnetic effect that is fundamentally different from Faraday's law as expressed in Eq. 1. More specifically, as regards possible effects of magnetic fields on bone, it entails a totally different phenomenon than the induction of current in bone using pulsed magnetic fields.

Unlike previous attempts to arrive at the electromagnetic conditions required for electrical osteogenesis, exact predictions are possible using the ICR effect. One can focus on a specific ion and adjust the intensity of the dc magnetic

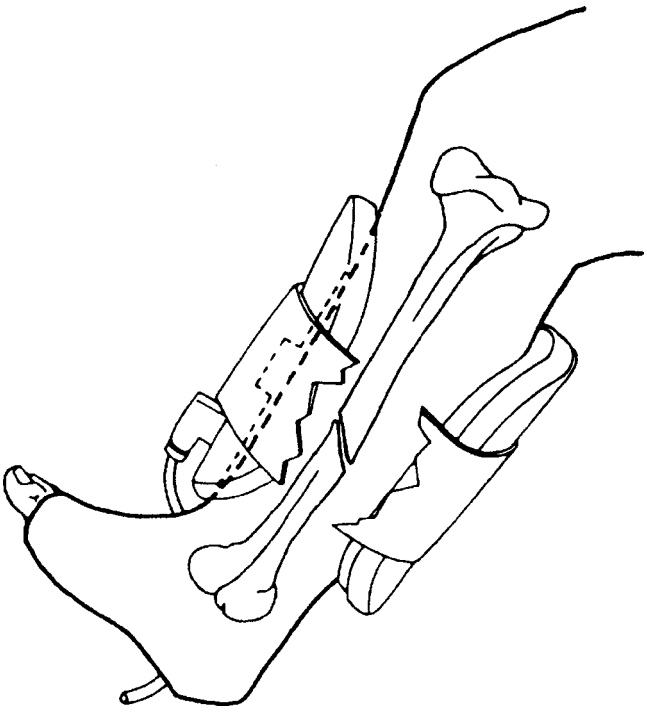


Figure 14. Treatment of congenital pseudarthrosis of the tibia with PMF (54).

Figure 15. PMF Bone healing system (EBI).

Figure 16. PMF device in place on patient (EBI).

field and the frequency of the ac magnetic field to "tune" to this ion. This is because a resonant condition occurs when the ratio of the frequency of the ac field to the intensity of the dc field is equal to the charge-to-mass ratio of the ion. The simple expression governing this resonance is

$$\omega/B = q/m \tag{3}$$

where ω is the (angular) frequency of the ac field, in $\text{rad} \cdot \text{s}^{-1}$, B is the intensity of the dc magnetic field, in Tesla, and, q/m is the mass-to-charge ratio of the ion. For practical applications, the angular frequency ω is replaced by its equivalent, $2\pi f$, where f is the frequency in hertz (Hz). The underlying interaction mechanism for this effect in living tissue is still in question (61), but the most reasonable explanation is that ions in resonance are more likely to stimulate the gating mechanism for ion channel transport.

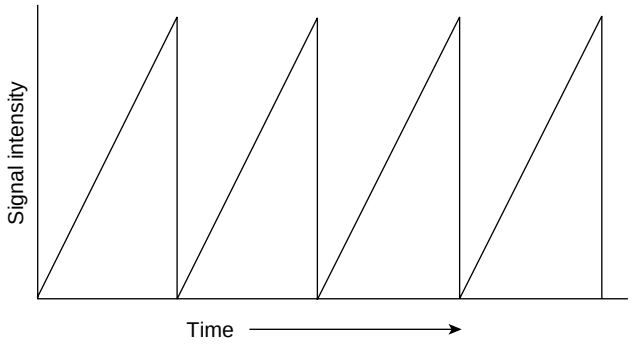


Figure 17. Sawtooth voltage applied to PMF coil.

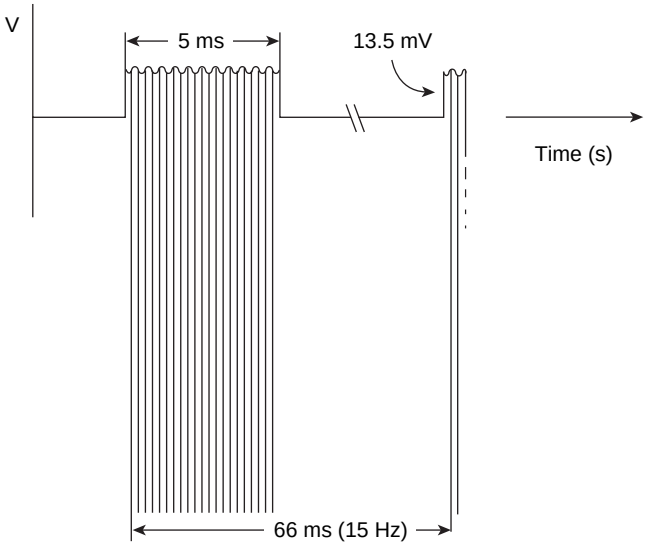


Figure 18. Voltage induced in tissue by PMF.

A great deal of work has been done in examining the effects on biological expression when tuning to Ca^{2+}, Mg^{2+}, and K^{+}, not only in bone cell culture (Fig. 19) (62), but also in neural cell culture, in animal behavior, and in plants (61). It is generally agreed that ICR tuning to these ions can have striking effects on growth. One such example (63) is shown in Fig. 20 illustrating the relative effects on explanted embryonic chick femora cultured under Ca^{2+} and under K^{+} ICR magnetic field conditions.

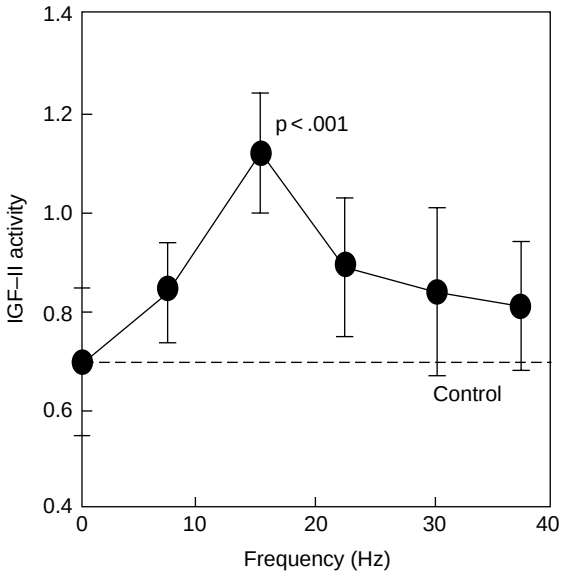


Figure 19. Frequency response of insulin-like growth factor in bone cell culture under combined ac and dc magnetic field exposure 62. The dc field was maintained at 20 μT for each of the points shown. There is a clear peak at 15.3 Hz, corresponding to the predicted ICR condition for Ca^{2+} resonance in Table 1.

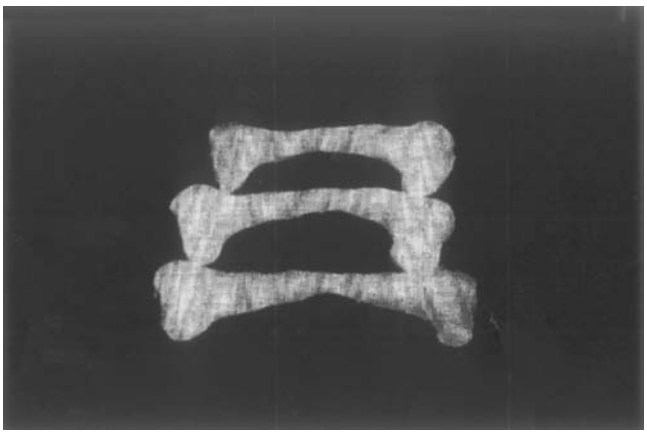


Figure 20. Effect of ICR magnetic exposures on chick embryonic growth (63). The topmost femur, the shortest, was grown under K^+ ICR tuning, while the bottom femur was grown under Ca^{2+} ICR magnetic field conditions. The middle femur was not exposed to any ICR field.

Diebert et al. (64) examined the efficacy of ICR in repairing defects in rabbit fibula, basically reusing the animal model that had been previously employed to study electrical osteogenesis (29), but applying an ICR magnetic field combination instead of a dc current. It was found that the 28-day ICR treatment yielded results equivalent to or better than those employing direct current and pulsed magnetic fields. For animals exposed to Ca^{2+} resonance magnetic fields for as little as 30 min · day^{-1}, there was an average increase in stiffness of 175% over controls, rising to nearly 300% when the exposures were maintained for 24 h. Somewhat smaller increases in stiffness were also observed for exposures tuned to the Mg^{2+} charge-to-mass ratio.

Another aspect of the ICR effect is that one can also use harmonics, that is, multiples of the frequency condition given in Eq. 3. For theoretical reasons (65) only odd harmonics are allowed. Thus, the most general expression for cyclotron resonance frequencies is

$$f_n = (2n + 1)(1/2\pi)(qB/m) \quad n = 0, 1, 2, 3, \ldots \qquad (4)$$

Table 1 lists the frequency/field ratios (f_n/B) for the three ions, Mg^{2+}, Ca^{2+}, and K^+ for the first three harmonics from Eq. 4. Note that some of these ratios are numerically close to one another. The 5th harmonic of Ca^{2+} is slightly $> 1\%$ greater than the 3rd harmonic for Mg^{2+} (3.83 vs. 3.79). This observation led S.D. Smith to suggest that using a frequency/field ratio of 3.8 might be particularly effective in bone where growth is indicated for both Ca^{2+} and Mg^{2+} stimulation (66). This ratio is the basis for a number of bone stimulation devices manufactured by the djOrthopedics

Table 1.

Ion	Fundamental f_0/B, Hz · μT	3rd Harmonic f_1/B, Hz · μT	5th Harmonic f_2/B, Hz · μT
Mg^{2+}	1.26	3.79	6.31
Ca^{2+}	0.77	2.30	3.83
K^+	0.39	1.18	1.97

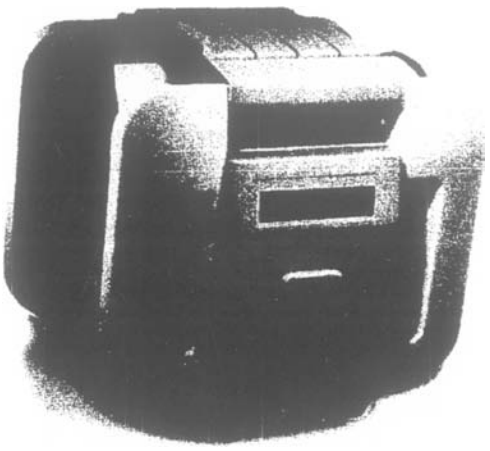


Figure 21. ICR bone repair device for treating nonunions. (djOrthopedics Corp.)

Corporation for treating pseudarthroses and enhancing spinal fusion (Figs. 21,22). The time variation of the magnetic field generated by these devices is shown in Fig. 23. Because the ac and dc magnetic field directions must be maintained parallel to ensure the resonance condition, these clinical devices achieve the frequency/field ratio by fixing the frequency of the applied sinusoidal magnetic field at 76.9 Hz, while using a second coil to continuously adjust for changes in the parallel component of the local dc magnetic field, to maintain this dc level at 20 μT.

Some observers incorrectly use the term *combined magnetic field* (CMF), to characterize this clinical technique. It is important to understand that the fields that are combined are highly specific, following the rules expressed in Eq. 4. In addition, it is possible to achieve the same conditions in tissue with a single magnetic field, using a prepared current derived from an arbitrary waveform generator. For these reasons, the term ICR should be used for all clinical and research techniques that are otherwise termed CMF.

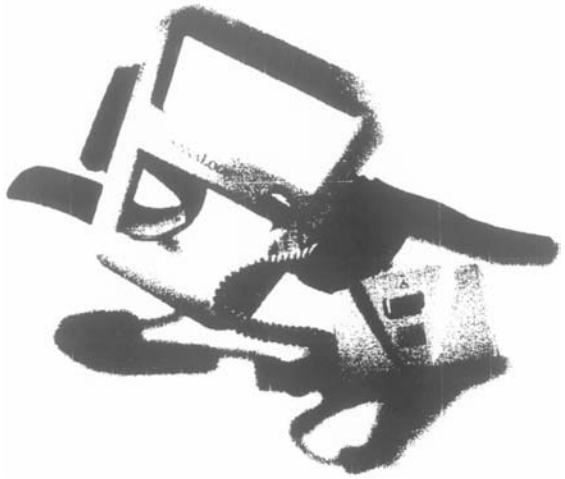


Figure 22. ICR device for adjunctive use in spinal fusion. (djOrthopedics Corp.)

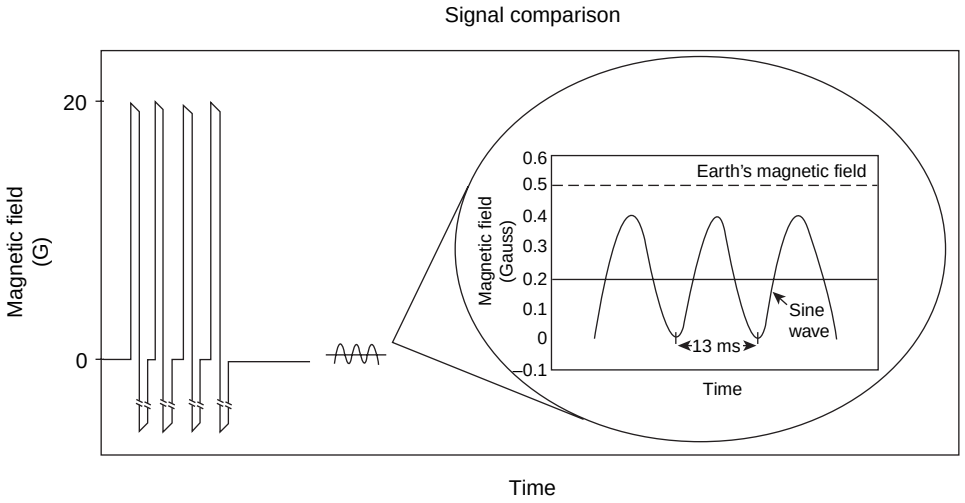


Figure 23. Comparing ICR signal to PMF signal. In the one case, there is a 20 μT peak sinusoidal magnetic field, and in the other a very short magnetic pulse ~ 100 times larger in intensity (Orthologic Corporation).

It is also sometimes incorrectly reported that ICR is an *inductive* procedure. However, the inductive current generated in the ICR device is negligible, approximately a factor of 10^{-5} smaller than the currents induced by PMF devices. While clearly noninductive, the actual ICR interaction mechanism is still in question (45). It is most likely coupled to events occurring at membrane bound ion channels (40), as evidenced that the calcium channel blocker nifedipene prevents the ICR response (67). It has been suggested, in this regard, that the channel gating process may be sensitive to the resonance tuning of specific ions (45).

SUMMARIZING EFFICACIES FOR THE VARIOUS TREATMENT

The three types of noninvasive treatments for bone defects, PMF, ICR, and CC, have each been subjected to randomized, double-blind trials and are shown to be efficacious, with an overall success rate of between 50 and 70%. One reason for this variation is undoubtedly the inclusion, or lack therein, of patients with defects that are intrinsically more difficult to repair. As the gap in a pseudarthrosis extends to widths > 5 mm, the likelihood of successful treatment diminishes. For this reason, some clinicians choose to exclude patients with radiographic gaps > 5 mm from electrical treatment (21,68).

More than 20 years after Bassett's original use (52) of pulsed magnetic fields to repair nonunions, a definitive work on using PMF to treat delayed unions was published by Sharrard (69). A total of 45 fractures of the tibia were examined in a double-blind multicenter trial, with active PMF stimulation in 20 patients and dummy control units in 25 patients for 12 weeks at 12 h/day. The results, 9 unions in the active group compared to only 3 in the control group, were "*very significantly in favour of the active group* ($p = 0.002$)". This effectiveness of PMF stimulation was confirmed for the case of tibial osteomoties in still another randomized, double blind study (70). Similarly, the successful use of CC and ICR, respectively, in treating nonunions, was reported by Scott and King (71) and by Longo (72). Recently, there has been increasing interest in the use of these electromagnetic techniques as an adjunctive to spine fusion. Again, as with the treatment of pseudarthroses, randomized, double-blind trials carried out for the PMF (73), CC (74) and ICR (75) techniques, have indicated that each is also efficacious in the adjunctive treatment of spine fusion.

GENERAL REMARKS

A summary of the electrical and electromagnetic treatments for nonunions and spinal fusion is given in Table 2. It is difficult to make simple comparisons based solely on the relative electrical characteristics, since each modality is based on different types of specifications, including current, current density, time rate of change of magnetic field, frequency, and magnetic intensity. As mentioned above, the levels of current that have been used to achieve osteogenesis extends over a range that has many orders of magnitude, a fact that seems to preclude any mechanism that is simply connected to current alone. It is highly likely that the larger of these successful current levels achieve osteogenesis, as Becker has suggested, by acting as an irritant, and that the smaller levels are perhaps related to the sorts of currents that might occur naturally, perhaps as the result of stress-generated potentials. One measure of

Table 2. Summary of Electrical and Electromagnetic Treatments for Nonunions and Spinal Fusion

	Modality	Designation	Characteristics	Daily Treatment
Invasive	dc electric	dc	1 pA–60 μA	24 h
Noninvasive	ac electric	CC	60 kHz, 7μA · cm^{-2}	24 h
	Pulsed magnetic field	PMF, PEMF	dB/dt=100 T · s^{-1} Repetition rate=15 Hz	3 h
	Sinusoidal magnetic field	ICR, CMF	77 Hz ac frequency 20 μT dc Field	30 min

this potential dichotomy is the remarkable fact that both ICR and PMF techniques are equally successful in treating nonunions, despite the fact that the induced currents differ by a factor of 10^5.

There is undoubtedly room for improvement in the efficacy of the various electromagnetic treatments to repair bony nonunion. Note that both treatments, PMF and ICR, were each adopted for clinical use on the basis of the original designs, with no subsequent studies before or after FDA approval that might have been initiated to search for waveforms and signals that conceivably could be used to optimize treatment. Thus for pulsed magnetic fields, it remains to be seen what roles are played by variables such as pulse width, rise time, repetition rate, and so on, and whether marked improvements in efficacy would follow optimization of these key variables. At least one report (76) claims that peak magnetic fields 100 times smaller than used in the EBI PMF device are just as effective in treating nonunions. Similarly, positive results were obtained in treating tibial osteotomies in rabbit with very different pulse characteristics from that of the EBI clinical device (77). Not only was the magnetic pulse reduced by a factor of 15, but the pulse repetition rate was reduced by a factor of 10, and the frequency components in excess of 20 kHz were filtered from the signal. This lack of optimization is equally true for the djOrthopedics ICR therapeutic signal, based on an approximate simultaneous stimulation of Ca^{2+} and Mg^{2+}ions as well as a very specific ratio of ac to dc magnetic intensities. The ICR device presently approved by the FDA sets this ratio at unity, despite the fact that a number of investigators (78–80) suggested that this ratio may have important consequences for the efficacy of the resonance interaction.

Furthermore, it has been suggested (31,81) that the fundamental reason why some electrical treatments of pseudarthroses are successful may have little to do with the nature of the electrical signal itself, but rather that the initiation of callus formation is known to be tied to local irritants, such as occurs with mechanical, thermal, or chemical sources. It is not inconceivable that the efficacy of treatments such as PMF may result from its role as an irritant. There is evidence (82,83) indicating an increased expression of heat shock proteins in response to low level electromagnetic fields. This type of genetic expression can result from a wide range of stress factors.

The fact that electrical osteogenesis occurs naturally, in growth, homeostasis, and repair and, further, that it can be brought about by exogenous application, begs the question as to whether the present 50–70% repair rate might be substantially improved with further research into the actual underlying mechanism.

BIBLIOGRAPHY

Cited References

1. Ryaby JT. Clinical effects of electromagnetic and electric fields on fracture healing. Clin Orth Rel Res 1998;355S: 205–215.
2. Marsh D. Concepts of fracture union, delayed union, and nonunion. Clin Orthop 1998;355S:22–30.
3. Marone MA, Feuer H. The use of electrical stimulation to enhance spinal fusion. Neurosurg Focus 2002;13: article 6.
4. Fukada E, Yasuda I. On the piezoelectric effect in bone. J Phys Soc Jpn 1957;12:1158–1162.
5. Fukada E, Yasuda I. Piezoelectric effects in collagen. Jpn J Appl Phys 1964;3:117–121.
6. Anderson JC, Eriksson C. Electric properties of wet collagen. Nature (London) 1968;218:166–168.
7. Anderson JC, Eriksson C. Piezoelectric properties of dry and wet bone. Nature (London) 1970;227:491–492.
8. Pienkowski D, Pollack SR. The origin of stress generated potentials in fluid-filled bone. J Orthop Res 1983;1:30–41.
9. McElhaney JH. The charge distribution on the human femur due to load. J Bone Joint Surg 1967;49:1561–1571.
10. Marino AA, Becker RO. Piezoelectric effect and growth control in bone. Nature (London) 1970;228:473–474.
11. Wolff J Das Gesetz der. Transformation der Knochen. Berlin: A. Hirschwohl; 1892.
12. Ferrier J, Moss SM, Kanehisa J, Aubin JE. Osteoclasts and osteoclasts migrate in opposite directions in response to a constant electrical field. J Cell Physiol 1986;129:283–288.
13. Athenstaedt H. Permanent electric polarization and pyroelectric behavior of the vertebrate skeleton III. The axial skeleton of man. Z Zellforsch 1969;93:484–504.
14. McGinnis ME. The nature and effects of electricity in bone, Chapt. 6 in Electric Fields in Vertebrate Repair. In: Borgens RP, Robinson KR, Venable JW, Jr, McGinnis ME, editors. New York: Alan R. Liss; 1989.
15. Friedenberg ZB. Bioelectric potentials in bone. J Bone Joint Surg 1966;48A:915–923.
16. Rubinacci A, Tessari L. A correlation analysis between bone formation rate and bioelectric potentials in rabbit tibia. Calc Tiss Res Int 1983;35:728–731.
17. Friedenberg ZB, Harlow MC, Heppenstall RB, Brighton CT. The cellular origin of bioelectric potentials in bone. Calc Tiss Res 1973;13:53–62.
18. Lang SB. Thermal expansion coefficients and the primary and secondary pyroelectric coefficients of animal bone. Nature (London) 1969;224:798–799.
19. Mascarenhas S. The electret effect in bone and biopolymers and the bound-water problem. Ann NY Acad Sci 1974;238: 36–52.
20. Liboff AR, Furst M. Pyroelectric effect in collagen and structures. Ann NY Acad Sci 1974;238:26–35.
21. Brighton CT. The treatment of non-unions with electricity. J Bone Joint Surg 1981;63A:847–851.
22. Friedenberg ZB, Smith HG. Electric potentials in intact and fractured tibia. Clin Orthop 1969;63:222–225.
23. Becker RO, Spadaro JA, Marino AA. Clinical experiences with low intensity direct current stimulation of bone growth. Clin Orthop Rel Res 1977;124:75–83.
24. Becker RO, Murray DG. The electric control system regulating fracture healing in amphibians. Clin Orthop Rel Res 1970;73:169–198.
25. Yasuda I. Fundamental aspects of fracture treatment. J Kyoto Med Soc 1953;4:395–406 (in Japanese) Translated in *Clin Orthop* 1977;124:5–8.
26. Yasuda I. Mechanical electrical callus. Electrically Mediated Growth mechanisms in Living Systems. Ann NY Acad Sci 1974;238:457–465.
27. Bassett CAL, Pawluk RJ, Becker RO. Effect of electric currents on bone in vivo. Nature (London) 1964;204:652–654.
28. Friedenberg ZB, Andrews ET, Smolenski BI, Pearl BW, Brighton CT. Bone reaction to varying amounts of direct current. Surg Gyn Obstet 1970;127:894–899.
29. Lavine L, Lustrin I, Shamos MH, Moss ML. The influence of electric current on bone regeneration in vivo. Acta Orthop Scand 1971;42:305–314.

30. Hambury HJ, Watson J, Toole A, Sivyer A, Ashley DBJ. Interdisciplinary approaches in electrically mediated bone growth studies. Ann NY Acad Sci 1974;238:508–518.
31. Paterson DC, Carter RF, Tilbury RF, Ludbrook J, Savage JP. The effects of varying current levels of electrical stimulation. Clin Orthop Rel Res 1982;169:303–312.
32. Ohashi T. Electrical callus formation and its osteogenesis. J Jpn Orthop Ass 1982;56:615–633.
33. Inoue S, Ohashi T, Fukada E, Ashihara T. Electric stimulation of osteogenesis in the rat: Amperage of three different stimulation mehods. In: Brighton CT, Black J, Pollack S, editors. Electric Properties of Bone and Cartilage. Philadelphia: Grune and Stratton; 1979. p 199–213.
34. Marino AA, Becker RO. Electrical osteogenesis: an analysis (Let). Clin Orthop Rel Res 1977;123:280–282.
35. Bassett CAL, Pawluk RJ, Pilla AA. Acceleration of fracture repair by electromagnetic fields. A surgically noninvasive method. Ann NY Acad Sci 1974;238:242–262.
36. Werner FW, Spadaro JA. Engineering aspects of medical surgical instruments and devices. In: Barzeley ME, editor. Product Liability. New York: Matthew Bender Publ. Co; 1993.
37. Lavine LS, Lustrin I, Shamos MH, Rinaldi RA, Liboff AR. Electric enhancement of bone healing. Science 1972;175: 1118–1121.
38. Friedenberg ZB, Harlow MC, Brighton CT. Healing of nonunion in the medial malleolus by means of direct current: a case report. J Trauma 1971;11:883–885.
39. Friedenberg ZB, Roberts PG, Jr, Didizian NH, Brighton CT. Stimulation of fracture healing by direct current in the rabbit fibula. J Bone Joint Surg 1971;53A:1400–1408.
40. Luben RA, et al. Effects of electromagnetic stimuli on bone and bone cells in vitro inhibition of responses to parathyroid hormone by low-energy, low-frequency field. Proc Natl Acad Sci USA 1982;79:4180–4184.
41. Norton LA, Rodan GA, Bourret LA. Epiphyseal cartilage cAMP changes produced by electrical and mechanical perturbations. Clin Orthop Rel Res 1977;124:57.
42. Farndale RW, Murray JC. The action of pulsed magnetic fields on cyclic AMP levels in cultured fibroblasts. Biochim. Biphys Acta 1986;881:46–53.
43. Fitzsimmons RJ, Ryaby JT, Mohan S, Magee FP, Baylink DJ. Combined magnetic fields increase IGF-II in TE-85 human bone cell cultures. Endocrinology 1995;136:3100–3106.
44. Bodamyali T, Bhatt B, Hughes FJ, Winrow VR, Kanczler JM, Simon B, Abbott J, Blake DR, Stevens CR. Pulsing electromagnetic fields simultaneously induce osteogenesis and upregulate transcription of bone morphogentic proteins 2 and 4in rat osteoblasts in vitro. Biochem Biophys Res Commun 1998;250:458–461.
45. Zhuang H, Wang W, Seldes RM, Tahemia AD, Fan H, Brighton CT. Electrical stimulation induces the level of TGF-β1 mRNA in osteoblastic cells by a mechanism involving calcium/calmodulin pathway. Biochem Biophys Res Commun 1997;237:225–229.
46. Wang Q, Zhong S, Ouyang J, Jiang L, Zhang Z, Xie Y, Luo S. Osteogenesis of electrically stimulated bone cells mediated in part by calcium ions. Clin Orthop Rel Res 1996;348:259–268.
47. Brighton CT, McCluskey WP. Response of bone cells to a capacitively coupled electric field: inhibition of cyclic adenosine monophosphate response to parathroid hormone. J Orthop Res 1988;6:567–571.
48. Lorich DG, Brighton CT, Gupta R, Corsetti JR, Levine SE, Gelb ID, Seldes R, Pollack SR. Biochemical pathway mediating the response of bone cells to capacitive coupling. Clin Orthop Rel Res 1998;350:246–256.
49. Brighton CT, Pollack SR. Treatment of recalcitrant nonunion with a capacitively coupled electric field. J Bone Joint Surg 1985;67A:577–585.
50. De Haas WG, Morrison DM, Watson J. Non-invasive treatment of the tibia using electrical stimulation. J Bone Joint Surg 1980;62:465–470.
51. Watson J, Downes EM. Light-weight battery-operable orthopaedic stimulator for the treatment of long-bone nonunions using pulsed magnetic fields. Med Biol Eng Comput 1983;21: 509–510.
52. Bassett CAL, Pilla AA, Pawluk RJ. A non-operative salvage of surgically-resistant pseudarthroses non-unions by pulsing electromagnetic fields electromagnetic fields. Clinical Orthop Rel Res 1977;124:128–143.
53. Frykman GK, Taleisnik J, Peters G, Kaufman R, Helal B, Wood VE, Unsell RS. Treatment of nonunited scaphoid fractures by pulsed electromagnetic field and cast. J Hand Surg 1986;11:344–349.
54. Kort JS, Schink MM, Mitchell SN, Bassett CAL. Congenital pseudarthrosis of the tibia: Treatment with pulsing electromagnetic fields. Clin Orthop Rel Res 1982;165:124–136.
55. Bassett CAL, Schink-Ascani M. Long-term pulsed electromagnetic field (PEMF) results in congenital pseudarthrosis. Calc Tiss Int 1991;49:216–220.
56. Liboff AR. Geomagnetic cyclotron resonance in living cells. J Biol Phys 1985;13:99–102.
57. Bawin SM, Kazmarek KL, Adey WR. Effects of modulated VHF fields on the central nervous system. Ann NY Acad Sci 1975;247:74–81.
58. Blackman CF, Benane SG, Kinney LS, Joines WT, House DE. Effects of ELF fields on calcium-ion efflux from brain tissue in vivo. Rad Res 1982;92:510–520.
59. Liboff AR, Williams T, Jr., Strong DM, Wistar R, Jr. Time-varying magnetic fields: effect on DNA synthesis. Science 1984;223:818–820.
60. Takahashi K, Keneko I, Date M, Fukada E. Effect of pulsing electromagnetic fields on DNA synthesis in mammalian cells in culture. Experientia 1986;42:185–186.
61. Liboff AR. The charge-to-mass ICR signature in weak ELF bioelectromagnetic effects. In: Lin JC, editor. Advances in Electromagnetic Fields in Living Systems. Volume 4, New York: Kluwer; 2003.
62. Fitzsimmons RJ, Ryaby JT, Mohan JT, Magee FP, Baylink DJ. Combined magnetic fields increase insulin-like growth factor-II in Te-85 human osteosarcoma bone cell cultures. Endocrinology 1995;136:3100–3106.
63. Smith SD, Liboff AR, McLeod BR. Effects of resonant magnetic fields on chick femoral development in vitro. J Bioelect 1999;10:81–99.
64. Diebert MC, McLeod BR, Smith SD, Liboff AR. Ion resonance electromagnetic field stimulation of fracture healing in rabbits with a fibular ostectomy. J Orthop Res 1994;12:878–885.
65. McLeod BR, Liboff AR. Cyclotron resonance in cell membranes; The theory of the mechanism. In: Blank M, Findl E, editors. Mechanistic Approaches to Interactions of Electric Electromagnetic Fields with Living Systems. New York: Plenum Press; 1987; p 97–108.
66. Smith SD. personal Communication.
67. Rozek RJ, Sherman ML, Liboff AR, McLeod BR, Smith SD. Nifedipene is an antagonist to cyclotron resonance enhancement of 45Ca incorporation in human lymphocytes. Cell Calcium 1987;8:413–427.
68. Barker AT, Dixon RA, Sharrard WJW, Sutcliffe ML. Pulsed magnetic field therapy for tibial non-union. The Lancet 1984;1:994–996.
69. Sharrard WJW. A double-blind trial of pulsed electromagnetic field for delayed union of tibial fractures. J Bone Joint Surg Bri 1990;72B:347–355.

70. Mammi GI, Rocchi R, Cadossi R, Traina GC. Effect of PEMF on the healing of human tibial osteomoties: a double blind study. Clin Orthop 1993;288:246–253.
71. Scott G, King JB. A prospective double blind trial of electrical capacitive coupling in the treatment of non-union of long bones. J Bone Joint Surg 1994;76A:820–826.
72. Longo JA. The management of recalcitrant nonunions with combined magnetic fields. Orthop Trans 1998;22:408–409.
73. Mooney VA. randomized double blind prospective study of the efficacy of pulsed electromagnetic fields for interbody lumbar fusions. Spine 1990;15:708–715.
74. Goodwin CB, Brighton CT, Guyer RD, Guyer RD, Johnson JR, Light KI, Yuan HA. A double blind study of capacitively coupled electrical stimulation as an adjunct to lumbar spinal fusions. Spine 1999;24:1349–1357.
75. Linovitz RJ, Pathria M, Bernhardt M, Green D, Law MD, McGuire RA, Montesano P, Rechtine G, Salib RM, Ryaby JT, Faden JS, Ponder R, Muenz LR, Magee FP, Garfin SA. Combined magnetic fields accelerate increase spine fusion: a double-blind, randomized, placebo controlled study. Spine 2002;27:1383–1389.
76. Satter SA, Islam MA, Rabbani KS, Talukder MS. Pulsed electromagnetic fields for the treatment of bone fractures. Bangladesh Med Res Council Bull 1999;25:6–19.
77. Fredericks DC, Nepola JV, Baker JT, Abbott J, Simon B. Effects of pulsed electromagnetic fields on bone healing in a rabbit tibial osteomoty model. J Orthop Trauma 2000;14:93–100.
78. Lednev VV. Possible mechanism for the influence of weak magnetic fields on biological systems. Bioelectromagnetics 1991;12:71–75.
79. Blanchard JP, Blackman CF. Clarification amplification of an ion parametric resonance model for magnetic field interactions with biological systems. Bioelectromagnetics 1994;15: 217–238.
80. Prato FP, Kavaliers M, Thomas AW. Extremely low frequency magnetic fields can either increase or decrease analgesia in the land snail depending on field and light conditions. Bioelectromagnetics 2000;21:287–301.
81. Becker RO. personal communication.
82. Goodman R, Bassett CAL, Henderson AS. Pulsing electromagnetic fields induce cellular transcription. Science 1983;220:1283–1285.
83. Goodman R, Blank M. A non-thermal low-energy agent that induces stress response proteins: Magnetic fields. Cell Stress Chaperones 1998;3:79–88.

See also BONE AND TEETH, PROPERTIES OF; FUNCTIONAL ELECTRICAL STIMULATION; HUMAN SPINE, BIOMECHANICS OF.

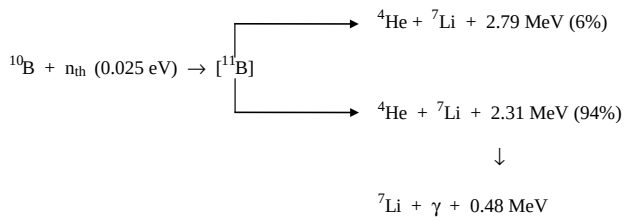
BORON NEUTRON CAPTURE THERAPY

THOMAS E BLUE
The Ohio State University
Columbus, Ohio

INTRODUCTION

High grade gliomas, and specifically glioblastoma multiforme (GBM), are still extremely resistant to all current forms of therapy, including surgery, chemotherapy, radiotherapy, immunotherapy, and gene therapy after decades of intensive research (1–5). Despite aggressive treatment using combinations of therapeutic modalities, the 5 year survival rate of patients diagnosed with GBM in the United States is less than a few percent (6,7). By the time they have had surgical resection of their tumors, malignant cells have infiltrated beyond the margins of resection and have spread into both gray and white matter (8,9). As a result, high grade supratentorial gliomas must be regarded as a whole brain disease (10). Glioma cells and their neoplastic precursors have biochemical properties that allow them to invade the unique extracellular environment of the brain (11,12) and biologic properties that allow them to evade a tumor associated host immune response (13). Chemo- and radiotherapy's inability to cure patients with high grade gliomas is due to their failure to eradicate microinvasive tumor cells within the brain. The challenge facing us is how to develop molecular strategies that can selectively target malignant cells with little or no effect on normal cells and tissues adjacent to the tumor. However, recent molecular genetic studies of glioma suggest that it may be much more complicated than this (14).

In theory, boron neutron capture therapy (BNCT) provides a way to selectively destroy malignant cells and spare normal cells. It is based on the nuclear capture and fission reactions that occur when boron-10, which is a nonradioactive constituent of natural elemental boron, is irradiated with low energy thermal neutrons to yield high linear energy-transfer (LET) alpha particles (^{4}He) and recoiling lithium-7 (^{7}Li) nuclei, as shown below.



In order for BNCT to be successful, a sufficient amount of ^{10}B must be selectively delivered to the tumor ($\sim$20 $\mu g \cdot g^{-1}$ weight or $\sim 10^9$ atoms/cell), and enough thermal neutrons must be absorbed by them to sustain a lethal $^{10}B(n, \alpha)\ ^7Li$ capture reaction. Since the high LET particles have limited boron pathlengths in tissue (5–9 μm), the destructive effects of these high energy particles is limited to cells containing boron. Clinical interest in BNCT has focused primarily on the treatment of high grade gliomas (15), and either cutaneous primaries (16) or cerebral metastases of melanoma (17), and most recently head and neck and liver cancer. Since BNCT is a biologically rather than physically targeted

ROLF F. BARTH
The Ohio State University
Columbus, Ohio
JEFFREY A. CODERRE
Massachusetts Institute of Technology
Cambridge, Massachusetts
M. GRAÇA
H. VICENTE
Louisiana State University
Baton Rouge, Louisiana

type of radiation treatment, the potential exists to destroy tumor cells dispersed in the normal tissue parenchyma, if sufficient amounts of ^{10}B and thermal neutrons are delivered to the target volume. This article covers radiobiological considerations upon which BNCT is based, boron agents and optimization of their delivery, neutron sources, which at this time are exclusively nuclear reactors, past and ongoing clinical studies, and critical issues that must be addressed if BNCT is to be successful. Readers interested in more in-depth coverage of these and other topics related to BNCT are referred to several recent reviews and monographs (15,18–20).

RADIOBIOLOGICAL CONSIDERATIONS

Types of Radiation Delivered

The radiation doses delivered to tumor and normal tissues during BNCT are due to energy deposition from three types of directly ionizing radiation that differ in their LET characteristics: (*1*) low LET γ rays, resulting primarily from the capture of thermal neutrons by normal tissue hydrogen atoms [$^{1}H(n,\gamma)^{2}H$]; (*2*) high LET protons, produced by the scattering of fast neutrons and from the capture of thermal neutrons by nitrogen atoms [$^{10}N(n,p)^{14}C$]; and (*3*) high LET, heavier charged alpha particles (stripped down ^{4}He nuclei) and lithium-7 ions, released as products of the thermal neutron capture and fission reactions with ^{10}B [$^{10}B(n,\alpha)^{7}Li$]. The greater density of ionizations along tracks of high LET particles results in an increased biological effect compared to the same physical dose of low LET radiation. Usually, this is referred to as relative biological effectiveness (RBE), which is the ratio of the absorbed dose of a reference source of radiation (e.g., X rays) to that of the test radiation that produces the same biological effect. Since both tumor and surrounding normal tissues are present in the radiation field, even with an ideal epithermal neutron beam, there will be an unavoidable, nonspecific background dose, consisting of both high and low LET radiation. However, a higher concentration of ^{10}B in the tumor will result in it receiving a higher total dose than that of adjacent normal tissues, which is the basis for the therapeutic gain in BNCT (21). As recently reviewed by one of us (18), the total radiation dose delivered to any tissue can be expressed in photon-equivalent units as the sum of each of the high LET dose components multiplied by weighting factors, which depend on the increased radiobiological effectiveness of each of these components.

Biological Effectiveness Factors

The dependence of the biological effect on the microdistribution of ^{10}B requires the use of a more appropriate term than RBE to define the biological effects of the $^{10}B(n,\alpha)^{7}Li$ reaction. Measured biological effectiveness factors for the components of the dose from this reaction have been termed compound biological effectiveness (CBE) factors and are drug dependent (21–23). The mode and route of drug administration, the boron distribution within the tumor, normal tissues, and even more specifically within cells, and even the size of the nucleus within the target cell population all can influence the experimental determination of the CBE factor. Therefore, CBE factors are fundamentally different from the classically defined RBE, which primarily is dependent on the quality (i.e., LET) of the radiation administered. The CBE factors are strongly influenced by the distribution of the specific boron delivery agent, and can differ substantially, although they all describe the combined effects of alpha particles and ^{7}Li ions. The CBE factors for the boron component of the dose are specific for both the boron-10 delivery agent and the tissue. A weighted gray (Gy) unit [Gy(w)] has been used to express the summation of all BNCT dose components and indicates that the appropriate RBE and CBE factors have been applied to the high LET dose components. However, for clinical BNCT the overall calculation of photon-equivalent [Gy(w)] doses requires a number of assumptions about RBEs, CBE factors, and the boron concentrations in various tissues that have been based on the currently available human or experimental data (24,25).

Clinical Dosimetry

The following biological weighting factors, summarized in Table 1, have been used in all of the recent clinical trials in patients with high grade glioma, using BPA in combination with an epithermal neutron beam. The $^{10}B(n,\alpha)^{7}Li$ component of the radiation dose to the scalp has been based on the measured boron concentration in the blood at the time of BNCT, assuming a blood:scalp boron concentration ratio of 1.5:1 (26,27,29) and a CBE factor for BPA in skin of 2.5 (29). An RBE of 3.2 has been used in all tissues for the high LET components of the beam: protons resulting from the capture reaction with nitrogen, and recoil protons resulting from the collision of fast neutrons with hydrogen (26,27,30). It must be emphasized that the tissue distribution of the boron delivery agent in humans should be similar to that in the experimental animal model in order to use the experimentally derived values for estimation of Gy(w) doses in clinical radiations.

Dose calculations become much more complicated when combinations of agents are used. At its simplest, this could be the two low molecular weight drugs boronophenylalanine (BPA) and sodium borocaptate (BSH). These have been shown to be highly effective when used in combination to treat F98 glioma bearing rats (31,32), and currently are being used in combination in a clinical study in Japan (33). Since it currently is impossible to know the true

Table 1. Assumptions Used in the Clinical Trials of BPA Based BNCT for Calculation of the $^{10}B(n,\alpha)^{7}Li$ Component of the Gy(w) Dose in Various Tissue

Tissue	Boron Concentration[a]	CBE Factor
Blood	measured directly	
Brain	1.0 × blood (26,27)	1.3 (23)
Scalp–skin	1.5 × blood (26–28)	2.5 (29)
Tumor	3.5 × blood (28)	3.8 (21)

[a] An RBE of 3.2 is used for the high LET component of the beam dose: protons from the $^{14}N(n,n)^{14}C$ reaction, and the recoil protons from fast neutron collisions with hydrogen. Literature references are given in parentheses.

biodistribution of each drug, dosimetric calculations in experimental animals have been based on independent boron determinations in other tumor bearing animals that have received the same doses of drugs but *not* BNCT. More recently, the radiation delivered has been expressed as a physical dose rather than using CBE factors to calculate an RBE equivalent dose (34). The calculations are further complicated if low and high molecular weight delivery agents are used in combination with one another. Tumor radiation dose calculations, therefore, are based on multiple assumptions regarding boron biodistribution, which may vary from patient to patient, as well as within different regions of the tumor and among tumor cells. However, normal brain boron concentrations are much more predictable and uniform, and therefore, it has been shown to be both *safe* and *reliable* to base dose calculations on normal brain tolerance.

BORON DELIVERY AGENTS

General Requirements

The development of boron delivery agents for BNCT began ~50 years ago and is an ongoing and difficult task of the highest priority. The most important requirements for a successful boron delivery agent are (*1*) low systemic toxicity and normal tissue uptake with high tumor uptake and concomitantly high tumor/brain (T/Br) and tumor/blood (T/Bl) concentration ratios (>3-4:1); (*2*) tumor concentrations in the range of ~20 μg $^{10}B \cdot g^{-1}$ tumor; (*3*) rapid clearance from blood and normal tissues and persistence in tumor during BNCT. However, at this time *no* single boron delivery agent fulfills all of these criteria. With the development of new chemical synthetic techniques and increased knowledge of the biological and biochemical requirements needed for an effective agent and their modes of delivery, a number of promising new boron agents has emerged (see examples in Fig. 1). The major challenge in their development has been the requirement for selective tumor targeting in order to achieve boron concentrations sufficient to deliver therapeutic doses of radiation to the tumor with minimal normal tissue toxicity. The selective destruction of GBM cells in the presence of normal cells represents an even greater challenge compared to malignancies at other anatomic sites, since high grade gliomas are highly infiltrative of normal brain, histologically complex, and heterogeneous in their cellular composition.

First- and Second-Generation Boron Delivery Agents

The clinical trials of BNCT in the 1950s and early 1960s used boric acid and some of its derivatives as delivery agents, but these simple chemical compounds were nonselective, had poor tumor retention, and attained low T/Br ratios (35,36). In the 1960s, two other boron compounds emerged from investigations of hundreds of low molecular weight boron-containing chemicals, one, (L)-4-dihydroxyborylphenylalanine, referred to as BPA (compound **1**) was based on arylboronic acids (37), and the other was based on a newly discovered polyhedral borane anion, sodium mercaptoundecahydro-*closo*-dodecaborate (38), referred to as BSH (compound **2**). These "second" generation compounds had low toxicity, persisted longer in animal tumors compared with related molecules, and their T/Br and T/Bl boron ratios were > 1. As described later in this article, ^{10}B enriched BPA, complexed with fructose to improve its water solubility, and BSH have been used clinically in Japan, the United States, and Europe. Although these drugs are not ideal, their safety following intravenous (i.v.) administration has been established. Over the past 20 years, several other classes of boron-containing compounds have been designed and synthesized in order to fulfill the requirements indicated at the beginning of this section. Detailed reviews of the state-of-the-art in compound development for BNCT have been published (39–42), and in this overview, only the main classes of compounds are summarized with an emphasis on recently published work in the area. The general biochemical requirements for an effective boron delivery agent are also discussed.

Third Generation Boron Delivery Agents

So-called "third" generation compounds mainly consist of a stable boron group or cluster attached via a hydrolytically stable linkage to a tumor-targeting moiety, such as low molecular weight biomolecules or monoclonal antibodies (MoAbs). For example, the targeting of the epidermal growth factor receptor (EGFR) and its mutant isoform EGFRvIII, which are overexpressed in gliomas and squamous cell carcinomas of the head and neck, also has been one such approach (43). Usually, these low molecular weight biomolecules have been shown to have selective targeting properties and many are at various stages of development for cancer chemotherapy, photodynamic therapy (PDT) or antiviral therapy. The tumor cell nucleus and DNA are especially attractive targets since the amount of boron required to produce a lethal effect may be substantially reduced, if it is localized within or near the nucleus (44). Other potential subcellular targets are mitochondria, lysosomes, endoplasmic reticulum, and the Golgi apparatus. Water solubility is an important factor for a boron agent that is to be administered systemically, while lipophilicity is necessary for it to cross the blood–brain barrier (BBB) and diffuse within the brain and the tumor. Therefore, amphiphilic compounds possessing a suitable balance between hydrophilicity and lipophilicity have been of primary interest since they should provide the most favorable differential boron concentrations between tumor and normal brain, thereby enhancing tumor specificity. However, for low molecular weight molecules that target specific biological transport systems and/or are incorporated into a delivery vehicle (e.g., liposomes) the amphiphilic character is not as crucial. The molecular weight of the boron-containing delivery agent also is an important factor, since it determines the rate of diffusion both within the brain and the tumor.

LOW MOLECULAR WEIGHT AGENTS

Boron-Containing Amino Acids and Polyhedral Boranes

Recognizing that BPA and BSH are not ideal boron delivery agents, considerable effort has been directed toward

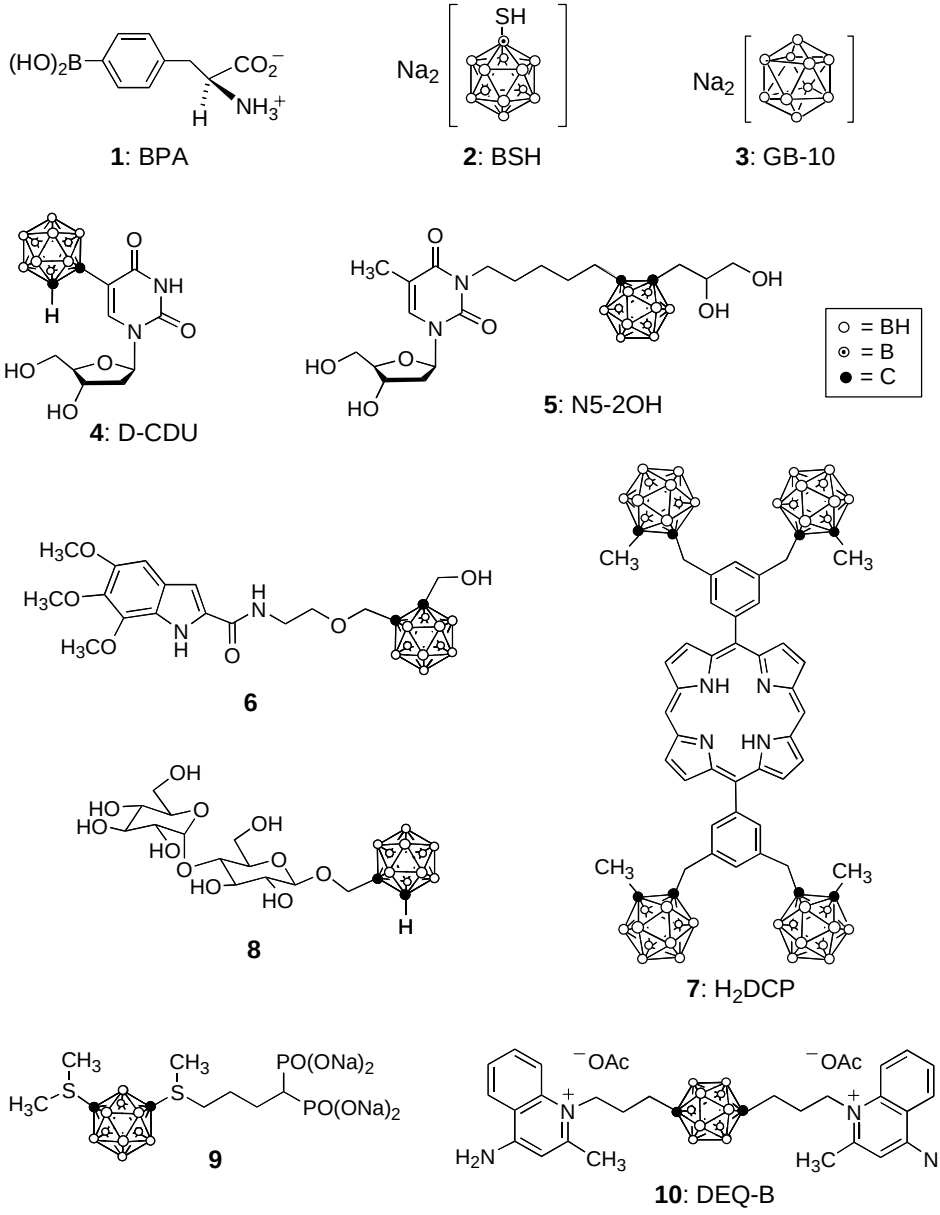


Figure 1. Some low molecular weight BNCT agents under investigation. Compound **1** (BPA) and compound **2** (BSH) are currently in clinical use in the United States, Japan, and Europe. Compound **3** (GB-10) has shown promise in animal models, as have the nucleoside derivatives D-CDU (compound **4**) and N5-2OH (compound **5**). Compound **6**, a trimethoxyindole derivative, has shown promise *in vitro* and compound **7**, a porphyrin derivative, was shown to be tumor selective. The maltose derivative **8** has shown low cytotoxicity and tumor cell uptake *in vitro*, the biphosphonate **9** has tumor targeting ability and the dequalinium derivative DEQ-B (compound **10**) has shown promise in *in vitro* studies.

the design and synthesis of third generation, boron-containing amino acids and functionalized polyhedral borane clusters. Examples include various derivatives of BPA and other boron-containing amino acids (e.g., glycine, alanine, aspartic acid, tyrosine, cysteine, methionine), as well as non-naturally occurring amino acids (45–50). The most recently reported delivery agents contain one or more boron clusters and concomitantly larger amounts of boron by weight compared with BPA. The advantages of such compounds are that they potentially can deliver higher concentrations of boron to tumors without increased toxicity. The polyhedral borane dianions, *closo*-$B_{10}H_{10}^{2-}$ and *closo*-$B_{12}H_{12}^{2-}$ and the icosahedral carboranes *closo*-$C_2B_{10}H_{12}$ and *nido*-$C_2B_9H_{12}^{-}$, have been the most attractive boron clusters for linkage to targeting moieties, due to their relatively easy incorporation into organic molecules, high boron content, chemical and hydrolytic stability, hydrophobic character and, in most cases, their negative charge. The simple sodium salt of *closo*-$B_{10}H_{10}^{2-}$ (GB-10, compound **3**) has been shown to have tumor-targeting ability and low systemic toxicity in animal models (42) and has been considered as a candidate for clinical evaluation (51). Other polyhedral borane anions with high boron content include derivatives of $B_{20}H_{18}^{2-}$, although these compounds have shown little tumor specificity, and therefore may be better candidates for encapsulation into either targeted or non-targeted liposomes (52,53) and folate receptor targeting, boron containing polyamidoamino (PAMAM) dendrimers (54) and liposomes (55). Boron-containing dipeptides also have shown low toxicity and good tumor-localizing properties (56,57).

Biochemical Precursors and DNA Binding Agents

Several boron-containing analogs of the biochemical precursors of nucleic acids, including purines, pyrimidines,

nucleosides, and nucleotides, have been synthesized and evaluated in cellular and animal studies (58–62). Some of these compounds [e.g., β-5-*o*-carboranyl-2'-deoxyuridine (D-CDU, compound **4**] and the 3-(dihydroxypropyl-carboranyl-pentyl)thymidine derivative N5-2OH (compound **5**), have shown low toxicities, selective tumor cell uptake, and significant rates of phosphorylation into the corresponding nucleotides (63–65). Intracellular nucleotide formation potentially can lead to enhanced tumor uptake and retention of these types of compounds (64,65).

Another class of low molecular weight delivery agents are boron-containing DNA binding molecules (e.g., alkylating agents, intercalators, groove binders, and polyamines). Some examples are derivatives of aziridines, acridines, phenanthridines (compound **6**), trimethoxyindoles, carboranylpolyamines, Pt(II)–amine complexes, di- and tribenzimidazoles (66–69). A limitation of boron-containing polyamines is their frequently observed *in vitro* and *in vivo* toxicity, although promising derivatives with low cytotoxicity have been synthesized (70–73). Other nuclear-targeting molecules are *nido*-carboranyl oligomeric phosphate diesters (OPDs). Despite their multiple negative charges, OPDs have been shown to target the nuclei of TC7 cells following microinjection (74), suggesting that the combination of OPDs with a cell-targeting molecule capable of crossing the plasma membrane could provide both selectivity and nuclear binding. Such a conjugate has been designed and synthesized (75), although its biological evaluation has yet to be reported.

Boron-Containing Porphyrins and Related Structures

Several boron-containing fluorescent dyes, including porphyrin, tetrabenzoporphyrin, and phthalocyanine derivatives have been synthesized and evaluated (76–79). These have the advantage of being easily detected and quantified by fluorescence microscopy, and have the potential for interacting with DNA due to their planar aromatic structures. Among these macrocycles, boron-containing porphyrins (e.g., compound **7**: H_2DCP) have attracted special attention due to their low systemic toxicity compared with other dyes, easy synthesis with high boron content, and their remarkable stability (79–82). Porphyrin derivatives have been synthesized that contain up to 44% boron by weight using *closo*- or *nido*-carborane clusters linked to the porphyrin macrocycle via ester, amide, ether, methylene, or aromatic linkages (76–85). The nature of these linkages is believed to influence their stability and systemic toxicity. Therefore, with these and other boron delivery agents, chemically stable carbon–carbon linkages have been preferred over ester and amide linkages that potentially can be cleaved *in vivo*. Boron-containing porphyrins have excellent tumor-localizing properties (76–82) and have been proposed for dual application as boron delivery agents and photosensitizers for PDT of brain tumors (85–91). Our own preliminary data with H_2TCP (tetra[*nido*-carboranylphenyl]porphyrin), administered intracerebrally by means of convection enhanced delivery (CED) to F98 glioma bearing rats, showed tumor boron concentrations of 150 $\mu g \cdot g^{-1}$ tumor with concomitantly low normal brain and blood concentrations (92). Ozawa et al. recently described a newly synthesized polyboronated porphyrin, designated TABP-1, which was administered by CED to nude rats bearing intracerebral implants of the human glioblastoma cell line U-87 MG (93). High tumor and low blood boron concentrations were observed and both we and Ozawa have concluded that direct intracerebral administration of the carboranyl porphyrins by CED is superior to systemic administration. Furthermore, despite the bulkiness of the carborane cages, carboranylporphyrins have been shown to interact with DNA and thereby produce *in vitro* DNA damage following light activation (94,95). Boronated phthalocyanines have been synthesized, although these compounds usually have had decreased water solubility and an increased tendency to aggregate compared to the corresponding porphyrins (76,77,86,87). Boron-containing acridine molecules also have been reported to selectively deliver boron to tumors with high T/Br and T/Bl ratios, whereas phenanthridine derivatives were found to have poor specificity for tumor cells (94–98).

Other Low Molecular Weight Boron Delivery Agents

Carbohydrate derivatives of BSH and other boron-containing glucose, mannose, ribose, gulose, fucose, galactose, maltose (e.g., compound **8**) and lactose molecules have been synthesized, and some of these compounds have been evaluated in both *in vitro* and *in vivo* studies (99–105). These compounds usually are highly water soluble and as a possible consequence of this, they have shown both low toxicity and uptake in tumor cells. It has been suggested that these hydrophilic low molecular weight derivatives have poor ability to cross tumor cell membranes. However, they might selectively accumulate within the glycerophospholipid membrane bilayer and in other areas of the tumor, such as the vasculature.

Low molecular weight boron-containing receptor-binding molecules have been designed and synthesized. These have been mainly steroid hormone antagonists, such as derivatives of tamoxifen, 17β-estradiol, cholesterol, and retinoic acid (106–110). The biological properties of these agents depend on the density of the targeted receptor sites, although to date very little biological data have been reported. Other low molecular weight boron-containing compounds that have been synthesized include phosphates, phosphonates (e.g., compound **9**) phenylureas, thioureas, nitroimidazoles, amines, benzamides, isocyanates, nicotinamides, azulenes, and dequalinium derivatives (e.g., dequalinium-B, compound **10**) (111–113). Since no single chemical compound, as yet synthesized, has the requisite properties, the use of *multiple* boron delivery agents is probably essential for targeting different subpopulations of tumor cells and subcellular sites. Furthermore, lower doses of each individual agent would be needed, which could reduce systemic toxicity while at the same time enhancing tumor boron levels to achieve a therapeutic effect.

HIGH MOLECULAR WEIGHT AGENTS

Monoclonal Antibodies, Other Receptors Targeting Agents and Liposomes

High molecular weight delivery agents (e.g., MoAbs and their fragments), which can recognize a tumor-associated epitope, have been (114–116) and continue to be of interest to us (117,118) as boron delivery agents. Although they can be highly specific, only very small quantities reach the brain and tumor following systemic administration (119) due to their rapid clearance by the reticuloendothelial system and the BBB, which effectively limits their ability to cross capillary vascular endothelial cells. Boron-containing bioconjugates of epidermal growth factor (EGF) (120,121), the receptor which is overexpressed on a variety of tumors, including GBM (122,123), also have been investigated as potential delivery agents to target brain tumors. However, it is unlikely that either boronated antibodies or other bioconjugates would attain sufficiently high concentrations in the brain following systemic administration, but, as described later in this section, direct intracerebral delivery could solve this problem. Another approach would be to directly target the vascular endothelium of brain tumors using either boronated MoAbs or VEGF, which would recognize amplified VEGF receptors. The use of boron-containing VEGF bioconjugates would obviate the problem of passage of a high molecular weight agent across the BBB, but their use would most likely require repeated applications of BNCT, since tumor neovasculature can continuously regenerate. Backer et al. reported that targeting a Shiga-like toxin-VEGF fusion protein was selectively toxic to vascular endothelial cells overexpressing VEGFR-2 (124). Recently, a bioconjugate has been produced by chemically linking a heavily boronated PAMAM dendrimer to VEGF (125). This selectively targeted tumor blood vessels overexpressing VEGFR-2 in mice bearing 4T1 breast carcinoma. There also has been a longstanding interest on the use of boron-containing liposomes as delivery agents (52,53,126,127), but their size has limited their usefulness as brain tumor targeting agents, since they are incapable of traversing the BBB unless they have diameters <50 nm (128). If, on the other hand, they were administered intracerebrally or were linked to an actively transported carrier molecule (e.g., transferin), or alternatively if the BBB was transiently opened, these could be very useful delivery agents, especially for extracranial tumors (e.g., liver cancer).

Recent work of one of us (R.F.B.) has focused on the use of a chemeric MoAb, cetuximab (IMC-C225 also known as Erbitux), produced by ImClone Systems, Inc. This antibody recognizes both wild-type EGFR and its mutant isoform, EGFRvIII (129), and has been approved for clinical use by the U.S. Food and Drug Administration (FDA) for the treatment of EGFR(+) recurrent colon cancer. Using previously developed methodology (114), a precision macromolecule, a polyamido amino (PAMAM or "starburst") dendrimer has been heavily boronated and then linked by means of heterobifunctional reagents to EGF (121), cetuximab (118) or another MoAb, L8A4, which is specifically directed against EGFRvIII (130). In order to completely bypass the BBB, the bioconjugates were administered by either direct intratumoral (i.t.) injection (131) or CED (132) to rats bearing intracerebral implants of the F98 glioma that had been genetically engineered to express either wildtype EGFR (131) or EGFRvIII (133). Administration by either of these methods resulted in tumor boron concentrations that were in the therapeutic range (i.e., $\sim 20\,\mu g \cdot g^{-1}$ wt^{-1} tumor). Similar data also were obtained using boronated EGF, and based on the favorable uptake of these bioconjugates, therapy studies were initiated at the Massachusetts Institute of Technology nuclear reactor (MITR). The mean survival times (MST) of animals that received either boronated cetuximab (134) or EGF (135) were significantly prolonged compared to those of animals bearing receptor negative tumors. A further improvement in MSTs was seen if the animals received BPA, administered i.v., in combination with the boronated bioconjugates, thereby validating our thesis that combinations of agents may be superior to any single agent (32). As can be seen from the preceding discussion, the design and synthesis of low and high molecular weight boron agents have been the subject of intensive investigation. However, optimization of their delivery has not received enough attention, but nevertheless is of critical importance.

OPTIMIZING DELIVERY OF BORON CONTAINING AGENTS

General Considerations

Delivery of boron agents to brain tumors is dependent on (*1*) the plasma concentration profile of the drug, which depends on the amount and route of administration; (*2*) the ability of the agent to traverse the BBB; (*3*) blood flow within the tumor, and (*4*) the lipophilicity of the drug. In general, a high steady-state blood concentration will maximize brain uptake, while rapid clearance will reduce it, except in the case of intraarterial (i.a.) drug administration. Although the i.v. route currently is being used clinically to administer both BSH and BPA, this may not be ideal and other strategies may be needed to improve their delivery. Delivery of boron-containing drugs to extracranial tumors, such as head and neck and liver cancer, present a different set of problems, including nonspecific uptake and retention in adjacent normal tissues.

Intra-arterial Administration with or without Blood–Brain Barrier Disruption

As shown in experimental animal studies (31,32,134–136) Enhancing the delivery of BPA and BSH can have a dramatic effect both on increasing tumor boron uptake and the efficacy of BNCT. This has been demonstrated in the F98 rat glioma model where intracarotid (i.c.) injection of either BPA or BSH doubled the tumor boron uptake compared to that obtained by i.v. injection (31). This was increased fourfold by disrupting the BBB by infusing a hyperosmotic (25%) solution of mannitol via the internal carotid artery. Mean survival times (MST)of animals that received either BPA or BSH i.c. with BBB-D were increased 295 and 117%, respectively, compared to irradiated controls (31). The best survival data were obtained using both BPA and BSH in

combination, administered by i.c. injection with BBB-D. The MST was 140 days with a cure rate of 25%, compared to 41 days following i.v. injection with no long-term surviving animals (32). Similar data have been obtained using a rat model for melanoma metastatic to the brain. BPA was administered i.c. to nude rats bearing intracerebral implants of the human MRA 27 melanoma with or without BBB-D. The MSTs were 104–115 days with 30% long-term survivors compared to a MST of 42 days following i.v. administration (134). A similar enhancement in tumor boron uptake and survival was observed in F98 glioma bearing rats following i.c. infusion of the bradykinin agonist, RMP-7 (receptor mediated permeabilizer-7), now called Cereport (136,137). In contrast to the increased tumor uptake, normal brain boron values at 2.5 h following i.c. injection were very similar for the i.v. and i.c. routes with or without BBB-D. Since BNCT is a binary system, normal brain boron levels only are of significance at the time of irradiation and high values at earlier time points are inconsequential. These studies have shown that a significant therapeutic gain can be achieved by optimizing boron drug delivery, and this should be important for both ongoing and future clinical trials using BPA and/or BSH.

Direct Intracerebral Delivery

Different strategies may be required for other low molecular weight boron-containing compounds whose uptake is cell cycle dependent, such as boron-containing nucleosides, where continuous administration over a period of days may be required. We recently have reported that direct i.t. injection or CED of the borononucleoside N5-2OH (compound **5**) were both effective in selectively delivering potentially therapeutic amounts of boron to rats bearing intracerebral implants of the F98 glioma (61). Direct i.t. injection or CED most likely will be necessary for a variety of high molecular weight delivery agents such as boronated MoAbs (138) and ligands such as EGF (132), as well as for low molecular weight agents (e.g., nucleosides and porphyrins). Recent studies have shown that CED of a boronated porphyrin derivative similar to compound **7**, designated H_2DCP, resulted in the highest tumor boron values and T/Br and T/Bl ratios that have been seen with any of the boron agents that have been studied (92).

NEUTRON SOURCES FOR BNCT

Nuclear Reactors

Neutron sources for BNCT currently are limited to nuclear reactors and in the present section only information that is described in more detail in a recently published review will be summarized (139). Reactor derived neutrons are classified according to their energies as thermal E_n <0.5 eV), epithermal (0.5 eV<E_n<10 keV), or fast (E_n>10 keV). Thermal neutrons are the most important for BNCT since they usually initiate the $^{10}B(n,\alpha)^7Li$ capture reaction. However, because they have a limited depth of penetration, epithermal neutrons, which lose energy and fall into the thermal range as they penetrate tissues, are now preferred for clinical therapy. A number of reactors with very good neutron beam quality have been developed and currently are being used clinically. These include (*1*) MITR, shown schematically in Fig. 2 (140); (*2*) clinical reactor at Studsvik Medical AB in Sweden (141); (*3*) the FRi1 clinical reactor in Helsinki, Finland (142); (*4*) R2-0 High Flux Reactor (HFR) at Petten in the Netherlands (143); (*5*) LVR-15 reactor at the Nuclear Research Institute (NRI) in Rez, Czech Republic (144); (*6*) Kyoto University Research Reactor (KURR) in Kumatori, Japan (145); (*7*) JRR4 at the Japan Atomic Energy Research Institute (JAERI) (146); and (*8*) the RA-6 CNEA reactor in Bariloche, Argentina (147). Other reactor facilities are being designed, notably the TAPIRO reactor at the ENEA Casaccia Center near Rome, Italy, which is unique in that it will be a low-power fast-flux reactor (148), and a facility in South Korea. Two reactors that have been used in the past for clinical BNCT are the Musashi Institute of Technology (MuITR) reactor in Japan and the Brookhaven Medical Research Reactor (BMRR) at the Brookhaven National Laboratory (BNL) in Upton, Long Island, New York (26,27,149). The MuITR was used by Hatanaka (150) and later by Hatanaka and Nakagawa (151). The BMRR was used for the clinical trial that was conducted at the Brookhaven National Laboratory between 1994 and 1999 (27,152) and the results are described in detail later in this section. Due to a variety of reasons, including the cost of maintaining the BMRR, it has been deactivated and is no longer available for use.

Reactor Modifications

Two approaches are being used to modify reactors for BNCT. The first or direct approach, is to moderate and filter neutrons that are produced in the reactor core. The second, the fission converter-plate approach, is indirect in that neutrons from the reactor core create fissions within a converter-plate that is adjacent to the moderator assembly, and these produce a neutron beam at the patient port. The MITR (153), which utilizes a fission converter-plate, currently sets the standard for the world for the combination of high neutron beam quality and short treatment time. It operates at a power of 5 MW and has been used for clinical as well as experimental studies for BNCT. Although the power is high compared to the majority of other reactors that are being used, the treatment time is unusually short, since it utilizes a fission converter-plate to create the neutron beam. All other reactors use the direct approach to produce neutron beams for BNCT. Three examples are the FiR1 reactor in Finland (142), the Studsvik reactor in Sweden (141), and the Washington State University (WSU) reactor in the United States (154), which was built for the treatment of both small and large experimental animals.

Accelerators

Accelerators also can be used to produce epithermal neutrons and accelerator based neutron sources (ABNSs) are being developed in a number of countries (155–161), and interested readers are referred to a recently published detailed review on this subject (28). For ABNSs, one of the more promising nuclear reactions involves bombarding a ^7Li target with 2.5 MeV protons. The average energy of

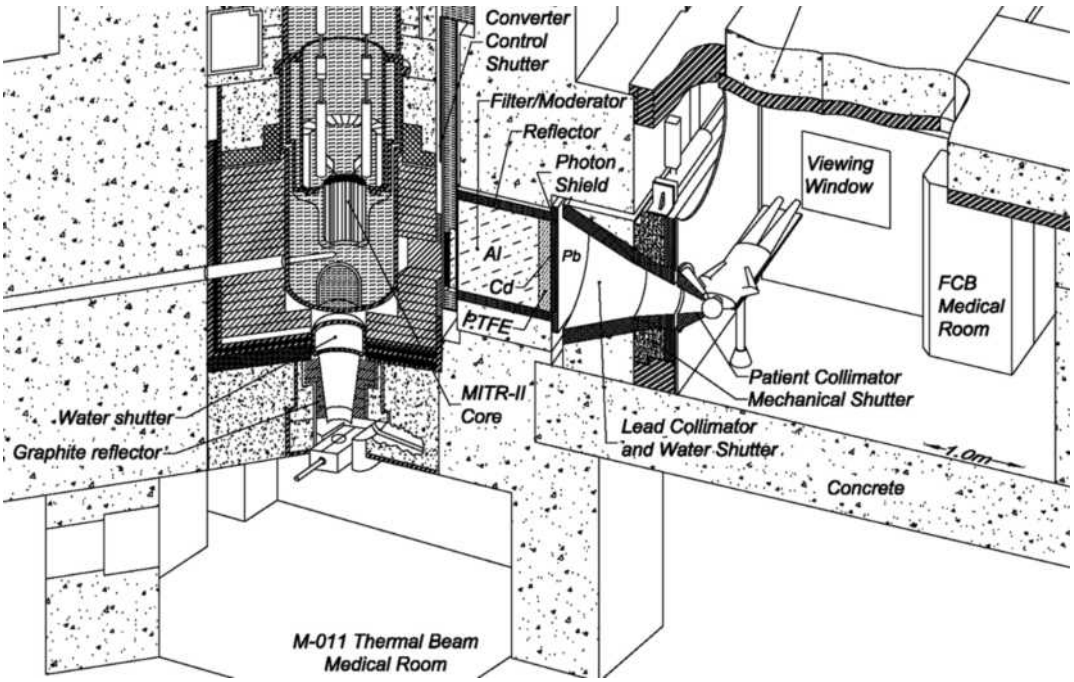


Figure 2. Schematic diagram of the MITR. The fission converter based epithermal neutron irradiation (FCB) facility is housed in the experimental hall of the MITR and operates in parallel with other user applications. The FCB contains an array of 10 spent MITR-II fuel elements cooled by forced convection of heavy water coolant. A shielded horizontal beam line contains an aluminum and Teflon filter-moderator to tailor the neutron energy spectrum into the desired epithermal energy range. A patient collimator defines the beam aperture and extends into the shielded medical room to provide circular apertures ranging from 16 to 8 cm in diameter. The in-air epithermal flux for the available field sizes ranges from 3.2 to 4.6×10^9 n·cm^{-2}·s^{-1} at the patient position. The measured specific absorbed doses are constant for all field sizes and are well below the inherent background of 2.8×10^{-12} RBE Gy·cm^2·n^{-1} produced by epithermal neutrons in tissue. The dose distributions achieved with the FCB approach the theoretical optimum for BNCT.

the neutrons that are produced is 0.4 MeV and the maximum energy is 0.8 MeV. Reactor derived fission neutrons have greater average and maximum energies than those resulting from the ^{7}Li(p,n)^{7}Be reaction. Consequently, the thickness of the moderator material that is necessary to reduce the energy of the neutrons from the fast to the epithermal range is less for an ABNS than it is for a reactor. This is important since the probability that a neutron will be successfully transported from the entrance of the moderator assembly to the treatment port decreases as the moderator assembly thickness increases. Due to lower and less widely distributed neutron source energies, ABNS potentially can produce neutron beams with an energy distribution that is equal to or better than that of a reactor. However, reactor derived neutrons can be well collimated, while on the other hand, it may not be possible to achieve good collimation of ABNS neutrons at reasonable proton beam currents. The necessity of good collimation for the effective treatment of GBM, is an important and unresolved issue that may affect usefulness of ABNS for BNCT. The ABNSs are also compact enough to be sited in hospitals thereby allowing for more effective, but technically more complicated procedures to carry out BNCT. However, to date, no accelerator has been constructed with a beam quality comparable to that of the MITR, which can be sited in a hospital and that provides a current of sufficient magnitude to treat patients in $<$30 min. Furthermore, issues relating to target manufacture and cooling must be solved before ABNS become a reality. The ABNS that is being developed at the University of Birmingham in England, by modifying a Dynamitron linear electrostatic accelerator (155), may be the first facility where patients will be treated, although progress has been slow. Another ABNS being constructed by LINAC Systems, Inc. in Albuquerque, New Mexico (162), and this could be easily sited in a hospital and produce an epithermal neutron beam.

Beam Optimization

For both reactors and ABNSs, a moderator assembly is necessary to reduce the energy of the neutrons to the epithermal range. The neutrons comprising the neutron beam have a distribution of energies and are accompanied by unwanted X rays and gamma photons. A basic tenet of BNCT is that the dose of neutrons delivered to the target volume should not exceed the tolerance of normal tissues, and this applies to neutron beam design, as well as to treatment planning (25). The implications of this for beam design is that the negative consequences of increased normal tissue damage for a more energetic neutron beams at shallow depths, outweighs the benefits of more deeply penetrating energetic neutrons. For fission reactors, the

average energy of the neutrons produced is ~2 MeV, but small numbers have energies as high as 10 MeV. There is generally a trade off between treatment time and the optimum beam for patient treatment in terms of the energy distribution of the neutrons and the contamination of the neutron beam with X rays and gamma photons. Not surprisingly, reactors with the shortest treatment time (i.e., the highest normal tissue dose rate) operate at the highest power, since the number of neutrons that is produced per unit time is proportional to the power, measured in megawatts. Furthermore, high beam quality is most easily achieved using reactors with high power, since a larger fraction of the neutrons can be filtered, as the neutrons traverse the moderator assembly without making the treatment time exceedingly long.

CLINICAL STUDIES OF BNCT FOR BRAIN TUMORS

Early Trials

Although the clinical potential of BNCT was recognized in the 1930s (163), it was not until the 1950s that the first clinical trials were initiated by Farr at the BNL (145,163) and by Sweet and Brownell at the Massachusetts General Hospital (MGH) using the MIT reactor (36,164,165). The disappointing outcomes of these trials, which ended in 1961 and subsequently were carefully analyzed by Slatkin (166), were primarily attributable to (*1*) inadequate tumor specificity of the inorganic boron chemicals that had been used as capture agents; (*2*) insufficient tissue penetrating properties of the thermal neutron beams; and (*3*) high blood boron concentrations that resulted in excessive damage to normal brain vasculature and to the scalp (36,164,165).

Japanese Clinical Trials

Clinical studies were resumed by Hatanaka in Japan in 1967, following a 2 year fellowship in Sweet's laboratory at the MGH, using a thermal neutron beam and BSH, which had been developed as a boron delivery agent by Soloway at the MGH (38). In Hatanaka's procedure (150,151), as much of the tumor was surgically removed as possible (debulking), and at some time thereafter, BSH (compound **2**) was administered by a slow infusion, usually intra-arterially (150), but later intravenously (151). Later (12–14 h) BNCT, was carried out at one or another of several different nuclear reactors. Since thermal neutrons have a limited depth of penetration in tissue, this necessitated reflecting the skin and raising the bone flap in order to directly irradiate the exposed brain. This eliminated radiation damage to the scalp and permitted treatment of more deep-seated residual tumors. As the procedure evolved over time, a ping–pong ball or silastic sphere was inserted into the resection cavity as a void space to improve neutron penetration into deeper regions of the tumor bed and adjacent brain (150,151,167,168). This is a major difference between the procedure carried out by Hatanaka, Nakagawa and other Japanese neurosurgeons and the BNCT protocols that have been carried out in the United States and Europe, which have utilized epithermal neutron beams that have not required reflecting the scalp and raising the bone flap at the time of irradiation. This has made it difficult to directly compare the Japanese clinical results with those obtained elsewhere, and this has continued on until very recently when the Japanese started using epithermal neutron beams (33). Most recently, Miyatake et al. initiated a clinical study utilizing the combination of BSH and BPA, both of which were administered i.v. at 12 and 1 h, respectively, prior to irradiation with an epithermal neutron beam (33). A series of 11 patients with high grade gliomas have been treated, and irrespective of the initial tumor volume, magnetic resonance imaging (MRI) and computed tomography (CT) images showed a 17–51% reduction in tumor volume that reached a maximum of 30–88%. However, the survival times of these patients were not improved over historical controls and further studies are planned to improve the delivery of BPA and BSH, which may enhance survival.

Analysis of the Japanese Clinical Results

Retrospective analysis of subgroups of patients treated in Japan by Hatanaka and Nakagawa (167,168) have described 2, 5, and 10 year survival rates (11.4, 10.4, and 5.7%, respectively) that were significantly better than those observed among patients treated with conventional, fractionated, external beam photon therapy. However, a cautionary note was sounded by Laramore and Spence (169) who analyzed the survival data of a subset of 12 patients from the United States who had been treated by Hatanaka between 1987 and 1994. They concluded that there were no differences in their survival times compared to those of age matched controls, analyzed according to the stratification criteria utilized by Curran et al. (6). In a recent review of Hatanaka's clinical studies, Nakagawa reported that the physical dose from the $^{10}B(n,\alpha)^{7}Li$ reaction, delivered to a target point 2 cm beyond the surgical margin, correlated with survival (168). For 66 patients with GBMs, those who survived <3 years ($n = 60$) had a minimum target point dose of 9.5 ± 5.9 Gy, whereas those who survived >3 years ($n = 6$) had a minimum target point dose of 15.6 ± 3.1 Gy from the $^{10}B(n,\alpha)^{7}Li$ reaction (168). The boron concentrations in brain tissue at the target point, which are required to calculate the physical radiation dose attributable to the $^{10}B(n,\alpha)^{7}Li$ capture reaction, were estimated to be 1.2X that of the patient's blood boron concentration (170).

OTHER RECENT AND ONGOING CLINICAL TRIALS

Beginning in 1994 a number of clinical trials, summarized in Table 2, were initiated in the United States and Europe. These marked a transition from low energy thermal neutron irradiation to the use of higher energy epithermal neutron beams with improved tissue penetrating properties, which obviated the need to reflect skin and bone flaps prior to irradiation. Up until recently, the procedure carried out in Japan required neurosurgical intervention immediately prior to irradiation, whereas the current epithermal neutron-based clinical protocols are radiotherapeutic procedures, performed several weeks after debulking surgery. Clinical trials for patients with brain tumors were initiated at a number of locations including (*1*) the

Table 2. Summary of Current or Recently Completed Clinical Trials of BNCT for the Treatment of Glioblastoma

Facility	Number of Patients	Duration of Administration	Drug	Dose, $mg \cdot kg^{-1}$	Boron Conc.,[a] $\mu g^{10}B \cdot g^{-1}$	Estimated Peak Normal Brain Dose, Gy(w)	Average Normal Brain Dose, Gy(w)	References
HTR, MuITR, JRR, KURR, Japan	>250 (1968–present)	1 h 1 h	BSH BPA	100 250	~20–30	13 Gy-Eq[b] ^{10}B component	Nd	168,169 33
HFR, Petten, The Netherlands	26 (1997–)	$100\ mg \cdot kg^{-1} \cdot min^{-1}$	BSH	100	30[c]	8.6–11.4 Gy-Eq[d] ^{10}B component	Nd	177
LVR-15, Rez, Czech Republic	5 (2001–present)	1 h	BSH	100	~20–30	<14.2	<2	178
BMRR Brookhaven	53 (1994–1999)	2 h	BPA	250–330	12–16	8.4–14.8	1.8–8.5	153,179
MITR-II, M67 MIT	20[e](1996–1999)	1–1.5 h	BPA	250–350	10–12	8.7–16.4	3.0–7.4	176
MITR-II, FCB MIT	6 (2001–2003)	1.5 h	BPA	350	~15			Unpublished
Studsvik AB Sweden	17 (30)[f](2001–2005)	6 h	BPA	900	24 (range: 15–34)	7.3–15.5	3.6–6.1	142
Fir I, Helsinki Finland	18 (1999–present) protocol P-01	2 h	BPA	290–400	12–15	8–13.5	3–6<7	143
Fir 1, Helsinki Finland	3 (2001–present)[g] protocol P-03	2 h	BPA	290	12–15	<8	2–3 <6	143

[a] During the irradiation.

[b] ^{10}B physical dose component dose to a point 2 cm deeper than the air-filled tumor cavity.

[c] Four fractions, each with a BSH infusion, $100\ mg \cdot kg^{-1}$ the first day, enough to keep the average blood concentration at $30\ \mu g\ ^{10}B \cdot g^{-1}$ during treatment on days 2–4.

[d] ^{10}B physical dose component at the depth of the thermal neutron fluence maximum.

[e] Includes two intracranial melanomas.

[f] J. Capala, unpublished, personal communication.

[g] Retreatment protocol for recurrent glioblastoma.

BMR at BNL from 1994–1999 for GBM using BPA with one or two neutron radiations, given on consecutive days (171–173); (*2*) the MITR from 1996–1999 for GBM and intracerebral melanoma (174,175); (*3*) the HFR, Petten, The Netherlands and the University of Essen in Germany in 1997 using BSH (176); (*4*) the FiR1 at the Helsinki University Central Hospital (142) in 1999 to the present; (*5*) the Studsvik reactor facility in Sweden from 2001 to June 2005, carried out by the Swedish National Neuro-Oncology Group (141), and finally (*6*) the NRI reactor in Rez, Czech Republic by Tovarys using BSH (177). The number of patients treated in this study is small and the followup is still rather short.

Initially, clinical studies using epithermal neutron beams were primarily Phase I safety and dose-ranging trials and a BNCT dose to a specific volume or critical region of the normal brain was prescribed. In both the BNL and the Harvard/MIT clinical trials, the peak dose delivered to a $1\ cm^3$ volume was escalated in a systematic way. As the dose escalation trials have progressed, the treatments have changed from single-field irradiations or parallel opposed irradiations, to multiple noncoplanar irradiation fields, arranged in order to maximize the dose delivered to the tumor. A consequence of this approach has been a concomitant increase in the average doses delivered to normal brain. The clinical trials at BNL and Harvard/MIT using BPA (compound **1**) and an epithermal neutron beam in the United States have now been completed.

Analysis of the Brookhaven and MIT Clinical Results

The BNL and Harvard/MIT studies have provided the most detailed data relating to normal brain tolerance following BNCT. A residual tumor volume of $60\ cm^3$ or greater lead to a greater incidence of acute CNS toxicity. This primarily was related to increased intracranial pressure, resulting from tumor necrosis and the associated cerebral edema (152,173,174). The most frequently observed neurological side effect associated with the higher radiation doses, other than the residual tumor volume-related effects, was radiation related somnolence (178). This is a well-recognized effect following whole brain photon irradiation (179), especially in children with leukemia or lymphoma, who have received CNS irradiation. However, somnolence is not a very well-defined radiation related endpoint because it frequently is diagnosed after tumor recurrence has been excluded. Therefore, it is not particularly well suited as a surrogate marker for normal tissue tolerance. In the dose escalation studies carried out at BNL (152,173), the occurrence of somnolence in the absence of a measurable tumor dose response was clinically taken as the maximum tolerated normal brain dose. The volume-averaged whole brain

dose and the incidence of somnolence increased significantly as the BNL and Harvard/MIT trials progressed (175). The volume of tissue irradiated has been shown to be a determining factor in the development of side effects (180). Average whole brain doses greater than ~5.5 Gy(w) were associated with somnolence in the trial carried out at BNL, but not in all of the patients in the Harvard/MIT study (18,152,176). The BNL and Harvard/MIT trials were completed in 1999. Both produced median and 1-year survival times that were comparable to conventional external beam photon therapy (6). Although both were primarily Phase I trials to evaluate the safety of dose escalation as the primary endpoint for radiation related toxicity, the secondary endpoints were quality of life and time to progression and overall survival. The median survival times for 53 patients from the BNL trial and the 18 GBM patients from the Harvard/MIT trial were 13 months and 12 months, respectively. Following recurrence, most patients received some form of salvage therapy, which may have further prolonged overall survival. Time to progression, which would eliminate salvage therapy as a confounding factor, probably would be a better indicator of the efficacy of BNCT, although absolute survival time still is the "gold standard" for any clinical trial. The quality of life for most of the BNL patients was very good, especially considering that treatment was given in one or two consecutive daily fraction(s).

Clinical Trials Carried Out in Sweden and Finland

The clinical team at the Helsinki University Central Hospital and VTT (Technical Research Center of Finland) have reported on 18 patients using BPA as the capture agent (290 mg·kg^{-1} infused over 2 h) with two irradiation fields and whole brain average doses in the range of 3–6 Gy(w) (142). The estimated 1-year survival was 61%, which was very similar to the BNL data. This trial is continuing and the dose of BPA has been escalated to 450 mg·kg^{-1} and will be increased to 500 mg·kg^{-1}, infused over 2 h (H. Joensuu, personal communication). Since BNCT can deliver a significant dose to tumor with a relatively low average brain dose, this group also has initiated a clinical trial for patients who have recurrent GBM after having received full-dose photon therapy. In this protocol, at least 6 months must have elapsed from the end of photon therapy to the time of BNCT and the peak brain dose should be <8 Gy(w) and the whole brain average dose <6 Gy(w). As of August 2005, only a small number of patients have been treated, but this has been well tolerated.

Investigators in Sweden have carried out a BPA-based trial using an epithermal neutron beam at the Studsvik Medical AB reactor (141). This study differed significantly from all previous clinical trials in that the total amount of BPA administered was increased to 900 mg·kg^{-1}, infused i.v. over 6 h. This approach was based on the following preclinical data: (*1*) the *in vitro* observation that several hours were required to fully load cells with BPA (181); (*2*) long-term i.v. infusions of BPA in rats increased the absolute tumor boron concentrations in the 9L gliosarcoma model, although the T:Bl ratio remained constant (182,183), and (*3*) most importantly, long-term i.v. infusions of BPA appeared to improve the uptake of boron in infiltrating tumor cells at some distance from the main tumor mass in rats bearing intracerebral 9L gliosarcomas (184). The longer infusion time of BPA was well tolerated (185–187) by the 30 patients who were enrolled in this study. All patients were treated with two fields, and the average weighted whole brain dose was 3.2–6.1 Gy(w), which was lower than the higher end of the doses used in the Brookhaven trial, and the minimum dose to the tumor ranged from 15.4 to 54.3 Gy(w). At 10 months following BNCT 23 of 29 evaluable patients had died with a median time to progression following BNCT of 5.8 months and a median survival time of 14.2 months. These results are comparable but not better than those obtained with external beam radiation therapy. Furthermore, they emphasize the need to improve the delivery of BPA, as well as BSH. As part of a broader plan to restructure the company, a decision was made by Studsvik AB in June 2005 to terminate operation of both the R2-0 reactor, which was used for this clinical trial, and the R2 reactor.

CLINICAL STUDIES OF BNCT FOR OTHER TUMORS

Treatment of Melanoma

Other than patients with primary brain tumors, the second largest group that has been treated by BNCT were those with cutaneous melanomas. Mishima and co-workers previously had carried out extensive studies in experimental animals with either primary or transplantable melanomas using ^{10}B enriched BPA as the capture agent (188,189). The use of BPA was based on the premise that it would be selectively taken up by and accumulate in neoplastic cells that were actively synthesizing melanin (190). Although it was subsequently shown that a variety of malignant cells preferentially took up large amounts of BPA compared to normal cells (191), nevertheless, Mishima's studies clearly stimulated clinical interest in BPA as a boron delivery agent. Since BPA itself has low water solubility, it was formulated with HCl to make it more water soluble. The first patient, who was treated by Mishima in 1985, had an acral lentigenous melanoma of his right toe that had been amputated (192). However, 14 months later he developed a subcutaneous metastatic nodule on the left occiput, which was determined to be inoperable due to its location. The tumor was injected peritumorally at multiple points for a total dose of 200 mg of BPA. Several hours later, by which time BPA had cleared from normal skin, but still had been retained by the melanoma, the tumor was irradiated with a collimated beam of thermal neutrons. Based on the tumor boron concentrations and the neutron fluence, an estimated 45 RBE-Gy equivalent dose was delivered to the melanoma. Marked regression was noted after 2 months, and the tumor had completely disappeared by 9 months (188,189,192). This successful outcome provided further evidence for *proof-of-principle* of the usefulness of BNCT to treat a radioresistant tumor. Subsequently, at least an additional 18 patients with either primary or metastatic melanomas have been treated by Mishima and co-workers (193). The BPA either was injected peritumorally or administered orally as a slurry (194) until Yoshino et al.

improved its formulation and water solubility by complexing it with fructose, following which it was administered i.v. (195). This important advance ultimately led to the use of BPA in the clinical trials in patients with brain tumors that were described in the preceding section. In all of Mishima's patients, there was local control of the treated primary or metastatic melanoma nodule(s) and several patients were tumor free at 4 or more years following BNCT (193).

Several patients with either cutaneous or cerebral metastases of melanoma have been treated by Busse et al. using BPA fructose as the delivery agent (18,196). The most striking example of a favorable response was in a patient with an unresected cerebral metastasis in the occipital lobe. The tumor received a dose of 24 RBE-Gy and monthly MRI studies revealed complete regression over a 4 month interval (196). As evidenced radiographically, a second patient with a brain metastasis had a partial response. Several other patients with either cutaneous or metastatic melanoma to the brain have been treated at other institutions, including the first in Argentina (197), and the consensus appears to be that these tumors are more responsive to BNCT than GBMs. This is supported by experimental studies carried out by two of us (R.F.B. and J.A.C.) using a human melanoma xenograft model (198,199), which demonstrated enhanced survival times and cure rates superior to those obtained using the F98 rat glioma model (200). In summary, multicentric metastatic brain tumors, and more specifically melanoma, which cannot be treated either by surgical excision or stereotactic radiosurgery, may be candidates for treatment by BNCT.

Other Tumor Types Treated by BNCT

Two other types of cancer recently have been treated by BNCT. The first is recurrent tumors of the head and neck. Kato et al. reported on a series of six patients, three of whom had squamous cell carcinomas, two had sarcomas, and 1 had a parotid tumor (201). All of them had received standard therapy and had developed recurrent tumors for which there were no other treatment options. All of the patients received a combination of BSH (5 g) and BPA ($250\,mg{\cdot}kg^{-1}$ body weight), administered i.v. In all but one patient, BNCT was carried out at the Kyoto University Research Reactor using an epithermal neutron beam in one treatment that was given 12h following administration of BSH and 1 h after BPA. The patient with the parotid tumor, who received a second treatment one month following the first, had the best response with a 63% reduction in tumor volume at 1 month and a 94% reduction at 1 year following the second treatment without evidence of recurrence. The remaining five patients showed responses ranging from a 10–27% reduction in tumor volume with an improvement in clinical status. This study has extended the use of BNCT to a group of cancers that frequently are ineffectively treated by surgery, radio-, and chemotherapy. However, further clinical studies are needed to objectively determine the clinical usefulness of BNCT for head and neck cancers, and another study to assess this currently is in progress at Helsinki University Central Hospital.

The second type of tumor that recently has been treated by BNCT is adenocarcinoma of the colon that had metastasized to the liver (202). Although hepatectomy followed by allogeneic liver transplantation, has been carried out at a number of centers (203,204), Pinelli and Zonta et al. (202) in Pavia, Italy, have approached the problem of multicentric hepatic metastases using an innovative, but highly experimental procedure. Their patient had >14 metastatic nodules in the liver parenchyma, the size of which precluded surgical excision. Before hepatectomy was performed, the patient received a 2 h infusion of BPA fructose ($300\,mg{\cdot}kg^{-1}$ b.w.) via the colic vein. Samples of tumor and normal liver were taken for boron determinations and once it was shown that boron selectively had localized in the tumor nodules with small amounts in normal liver, the hepatectomy was completed (202). The liver then was transported to the Reactor Laboratory of the University of Pavia for neutron irradiation, following which it was reimplanted into the patient. More than 2 years later in October 2004, the patient had no clinical or radiographic evidence of recurrence and CEA levels were low (205). Although it is unlikely that this approach will have any significant clinical impact on the treatment of the very large number of patients who develop hepatic metastases from colon cancer, it nevertheless again provides *proof of principle* that BNCT can eradicate multicentric deposits of tumor in a solid organ. The Pavia group has plans to treat other patients with metastatic liver cancer and several other groups (206–208) are exploring the possibility of treating patients with primary, as well as metastatic tumors of the liver using this procedure.

CRITICAL ISSUES

There are a number of critical issues that must be addressed if BNCT is to become a useful modality for the treatment of cancer, and most specifically, brain tumors. *First* and foremost, there is a need for more selective and effective boron agents, which when used either alone or in combination, could deliver the requisite amounts ($\sim 20\,\mu g{\cdot}g^{-1}$) of boron to the tumor. Furthermore, their delivery must be optimized in order to improve both tumor uptake and cellular microdistribution, especially to different subpopulations of tumor cells (185). A number of studies have shown that there is considerable patient-to-patient as well-intratumoral variability in the uptake of both BSH (209,210) and BPA (184,211,212). At this point in time, the dose and delivery of these drugs have yet to be optimized, but based on experimental animal data (31,32,34,137,183), improvement in dosing and delivery could have a significant impact on increasing tumor uptake and microdistribution.

Second, since the radiation dosimetry for BNCT is based on the microdistribution of ^{10}B (209,213), which is indeterminable on a real-time basis, methods are needed to provide semiquantatative estimates of the boron content in the residual tumor. Imahori and co-workers (214–216) in Japan and Kabalka (217) in the United States have carried out imaging studies with ^{18}F-labeled BPA, and have used to establish the feasibility of carrying out BNCT. This

^{18}F-PET imaging also has been used as a prognostic indicator for patients with GBM who may or may not have received BNCT (214,215). In the former group, it has been used to establish the feasibility of carrying out BNCT based on the uptake and distribution of ^{18}F-BPA within the tumor and in the latter to monitor the response to therapy. The possibility of using MRI for either ^{10}B or ^{11}B has been under investigation (218), and this may prove to be useful for real-time localization of boron in residual tumor prior to BNCT. Magnetic resonance spectroscopy (MRS) and magnetic resonance spectroscopic imaging (MRSI) also may be useful for monitoring the response to therapy (219). Kojimoto and Miyatake et al. recently used MRS to analyze the target specificity of BPA and the effects of BNCT in a group of six patients using multivoxel proton MRS (220). There was a reduction in the choline/creatine ratio without a reduction of the *N*-acetylaspartate/creatine ratio at 14 days following BNCT, strongly suggesting that there was selective destruction of tumor cells and a sparing of normal neurons (220). Noninvasive procedures (e.g., MRSI) may be a powerful way to follow the clinical response to BCNT in addition to MRI. However, in the absence of real-time tumor boron uptake data, the dosimetry for BNCT is very problematic. This is evident from the discordance of estimated doses of radiation delivered to the tumor and the therapeutic response, which would have been greater than that which was seen if the tumor dose estimates were correct (152).

Third, there is a discrepancy between the theory behind BNCT, which is based on a very sophisticated concept of selective cellular and molecular targeting of high LET radiation, and the implementation of clinical protocols, which are based on very simple approaches to drug administration, dosimetry, and patient irradiation. This in part is due to the fact that BNCT has not been carried out in advanced medical settings with a highly multidisciplinary clinical team in attendance. At this time BNCT has been totally dependent on nuclear reactors as neutron sources. These are a medically unfriendly environment and are located at sites at varying distances from tertiary care medical facilities, which has made it difficult to attract patients, and the highly specialized medical team that ideally should be involved in clinical BNCT. Therefore, there is an urgent need for either very compact medical reactors or ABNS that could be easily sited at selected centers that treat large numbers of patients with brain tumors.

Fourth, there is a need for randomized clinical trials. This is especially important since almost all major advances in clinical cancer therapy have come from these, and up until this time no randomized trials of BNCT have been conducted. The pitfalls of nonrandomized clinical trials for the treatment of brain tumors have been well documented (221,222). It may be somewhat wishful thinking to believe that the clinical results with BNCT will be so clearcut that a clear determination of efficacy could be made without such trials. These will require a reasonably large number of patients in order to provide unequivocal evidence of efficacy with survival times significantly better than those obtainable with promising currently available therapy for both GBMs (223,224) and metastatic brain tumors (225). This leads to the issue of conducting such trials, which might best be accomplished through cooperative groups such as the Radiation Therapy Oncology Group (RTOG) in the United State or the European Organization for Research Treatment of Cancer (EORTC).

Finally, there are several promising leads that could be pursued. The upfront combination of BNCT with external beam radiation therapy or in combination with chemotherapy has not been explored, although recently published experimental data, suggest that there may be a significant gain if BNCT is combined with photon irradiation (34). The extension of animal studies, showing enhanced survival of brain tumor bearing rats following the use of BSH and BPA in combination, administered intraarterially with or without BBB-D, has not been evaluated clinically. This approach is promising, but it is unlikely that it could be carried out at a nuclear reactor.

As is evident from this article, BNCT represents an extraordinary joining together of nuclear technology, chemistry, biology, and medicine to treat cancer. Sadly, the lack of progress in developing more effective treatments for high grade gliomas has been part of the driving force that continues to propel research in this field. BNCT may be best suited as an adjunctive treatment, used in combination with other modalities, including surgery, chemotherapy, and external beam radiation therapy, which, when used together, may result in an improvement in patient survival. Clinical studies have demonstrated the safety of BNCT. The challenge facing clinicians and researchers is how to get beyond the current impasse. We have provided a road map to move forward, but its implementation still remains a daunting challenge

ACKNOWLEDGMENTS

We thank Mrs. Michelle Smith for secretarial assistance in the preparation of this manuscript. Text and the two figures, which appear in this article, have been published in *Clinical Cancer Research* with copyright release from the American Association of Cancer Research, Inc.

Experimental studies described in this article have been supported by the National Institutes of Health grants 1R01 CA098945 (to R.F.B.) and 1R01 CA098902 to (M.G.H.V.) and Department of Energy Grants DE-FG02-93ER61612 (to T.E.B.) and DE-FG02-01ER63194 (to J.A.C.) and the Royal G. and Mae H. Westaway Family Memorial Fund at the Massachusetts Institute of Technology (to J.A.C.).

BIBLIOGRAPHY

Cited References

1. Berger MS. Malignant astrocytomas: surgical aspects. Seminars Oncol 1994;21:172–185.
2. Gutin PH, Posner JB. Neuro-Oncology: diagnosis and management of cerebral gliomas—past, present, and future. Neurosurgery 2000;47:1–8.
3. Parney IF, Chang SM. Current chemotherapy for glioblastoma. In: Market J, DeVita VT, Rosenberg SA, Hellman S, editors. Glioblastoma Multiforme, 1st ed., Sudbury (MA): Jones and Bartlett Publishers; 2005. p 161–177.

4. Paul DB, Kruse CA. Immunologic approaches to therapy for brain tumors. Curr Neurol Neurosci Rep 2001;1:238–244.
5. Rainov NG. Gene therapy for human malignant brain tumors. In: Market J, DeVita VT, Rosenberg SA, Hellman S, editors. Glioblastoma Multiforme. 1st ed. Sudbury (MA): Jones and Bartlett Publishers; 2005. p 249–265.
6. Curran WJ, et al. Recursive partitioning analysis of prognostic factors in three radiation oncology group malignant glioma trials. J Nat Cancer Inst 1993;85:704–710.
7. Lacroix M, et al. A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival. J Neurosurg 2001;95:190–198.
8. Hentschel SJ, Lang FF. Current surgical management of glioblastoma. In: Market J, DeVita VT, Rosenberg SA, Hellman S, editors. Glioblastoma Multiforme, 1st ed., Sudbury (MA): Jones and Bartlett Publishers; 2005. p 108–130.
9. Laws ER, Shaffrey ME. The inherent invasiveness of cerebral gliomas: implications for clinical management. Int J Devel Neurosc 1999;17:413–420.
10. Halperin EC, Burger PC, Bullard DE. The fallacy of the localized supratentorial malignant glioma. Int J Radiat Oncol Biol Phys 1988;15:505–509.
11. Kaczarek E, et al. Dissecting glioma invasion: interrelation of adhesion, migration and intercellular contacts determine the invasive phenotype. Int J Devel Neurosc 1999;17:625–641.
12. Huang S, Prabhu S, Sawaya R. Molecular and biological determinants of invasiveness and angiogenesis in central nervous system tumors. In: Zhang W, Fuller GN, editors. Genomic and Molecular Neuro-Oncology. Sudbury (MA): Jones and Bartlette Publishers; 2004. pp 97–118.
13. Parney IF, Hao C, Petruk K. Glioma immunology and immunotherapy. Neurosurgery 2000;46:778–792.
14. Ware ML, Berger MS, Binder DK. Molecular biology of glioma tumorigenesis. Histol Histopathol 2003;18:207–216.
15. Barth RF. A critical assessment of boron neutron capture therapy: An overview. J Neuro-Oncol 2003;62:1–5.
16. Mishima Y. Selective thermal neutron capture therapy of cancer cells using their specific metabolic activities - melanoma as prototype. In: Mishima Y, editor. Cancer Neutron Capture Therapy. New York: Plenum Press; 1996. p 1–26.
17. Busse PM, et al. A critical examination of the results from the Harvard-MIT NCT program phase I clinical trial of neutron capture therapy for intracranial disease. J Neuro-Oncol 2003;62:111–121.
18. Coderre JA, et al. Boron neutron capture therapy: cellular targeting of high linear energy transfer radiation. Technol Cancer Res Treatment 2003;2:1–21.
19. Sauerwein W, Moss R, Wittig A, editors. Research and development in neutron capture therapy. Bologna, Italy: Monduzzi Editore S.p.A., International Proceedings Division; 2002.
20. Coderre JA, Rivard MJ, Patel H, Zamenhof RG. Proceedings of the 11th World Congress on Neutron Capture Therapy. Appl Rad Isotopes 2004; 61s.
21. Coderre JA, Morris GM. The radiation biology of boron neutron capture therapy. Radiat Res 1999;151:1–18.
22. Morris GM, et al. Response of the central nervous system to boron neutron capture irradiation: Evaluation using rat spinal cord model. Radiother Oncol 1994;32:249–255.
23. Morris GM, et al. Response of rat skin to boron neutron capture therapy with *p*-boronophenylalanine or borocaptate sodium. Radiother Oncol 1994;32:144–153.
24. Gupta N, Gahbauer RA, Blue TE, Albertson B. Common challenges and problems in clinical trials of boron neutron capture therapy of brain tumors. J Neuro-Oncol 2003;62:197–210.
25. Nigg DW. Computational dosimetry and treatment planning considerations for neutron capture therapy. J Neuro-Oncol 2003;62:75–86.
26. Coderre JA, et al. Boron neutron capture therapy for glioblastoma multiforme using p-boronophenylalanine and epithermal neutrons: Trial design and early clinical results. J Neuro-Oncol 1997;33:141–152.
27. Elowitz EH, et al. Biodistribution of p-boronophenylalanine in patients with glioblastoma multiforme for use in boron neutron capture therapy. Neurosurgery 1998;42:463–469.
28. Blue TE, Yanch JC. Accelerator-based epithermal neutron sources for boron neutron capture therapy of brain tumors. J Neuro-Oncol 2003;62:19–31.
29. Fukuda H, et al. Boron neutron capture therapy of malignant melanoma using ^{10}B-paraboronophenylalanine with special reference to evaluation of radiation dose and damage to the skin. Radiat Res 1994;138:435–442.
30. Coderre JA, et al. Derivations of relative biological effectiveness for the high-LET radiations produced during boron neutron capture irradiations of the 9L rat gliosarcoma *in vitro* and *in vivo*. Int J Radiat Oncol Biol Phys 1993;27: 1121–1129.
31. Barth RF, et al. Boron neutron capture therapy of brain tumors: enhanced survival following intracarotid injection of either sodium borocaptate or boronophenylalanine with or without blood-brain barrier disruption. Cancer Res 1997;57: 1129–1136.
32. Barth RF, et al. Boron neutron capture therapy of brain tumors: enhanced survival and cure following blood-brain barrier disruption and intracarotid injection of sodium borocaptate and boronophenylalanine. Int J Radiat Oncol Biol Phys 2000;47:209–218.
33. Miyatake S, et al. Modified boron neutron capture therapy (BNCT) for mailignant gliomas using epithermal neutrons and two boron compounds with different accumulation mechanisms-Effectiveness of BNCT on radiographic images. J Neurosurg Dec. 2005 (In Press).
34. Barth RF, et al. Combination of boron neutron capture therapy and external beam X-irradiation for the treatment of brain tumors. Int J Radiat Oncol Biol Phys 2004;58: 267–277.
35. Farr LE, et al. Neutron capture therapy with boron in the treatment of glioblastoma multiforme. Am J Roenthemol 1954;71:279–291.
36. Godwin JT, Farr LE, Sweet WH, Robertson JS. Pathological study of eight patients with glioblastoma multiforme treated by neutron-capture therapy using boron 10. Cancer 1955;8:601–615.
37. Snyder HR, Reedy AJ, Lennarz WJ. Synthesis of aromatic boronic acids, aldehydo boronic acids and a boronic acid analog of tyrosine. J Am Chem Soc 1958;80:835–838.
38. Soloway AH, Hatanaka H, Davis MA. Penetration of brain and brain tumor. VII. Tumor-binding sulfhydryl boron compounds. J Med Chem 1967;10:714.
39. Hawthorne MF. The role of chemistry in the development of boron neutron capture therapy of cancer. Angew Chem Int Ed Engl 1993;32:950–984.
40. Morin C. The chemistry of boron analogues of biomolecules. Tetrahedron 1994;50:12521–12569.
41. Soloway AH, et al. The chemistry of neutron capture therapy. Chem Rev 1998;98:1515–1562.
42. Hawthorne MF, Lee MW. A critical assessment of boron target compounds for boron neutron capture therapy. J Neuro-Oncol 2003;62:33–45.
43. Olsson P, et al. Uptake of a boronated epidermal growth factor-dextran conjugate in CHO xenografts with and without human EGF-receptor expression. Anticancer Drug Des 1998;13:279–289.
44. Gabel D, Foster S, Fairchild RG. The Monte Carlo simulation of the biological effect of the ^{10}B(n,α)^{7}L reaction in cells and

tissue and its implication for boron neutron capture therapy. Radiat Res 1987;111:14–25.
45. Srivastava RR, Singhaus RR, Kabalka GW. 4-Dihydroxyborylphenyl analogues of 1-aminocyclobutanecarboxylic acids: potential boron neutron capture therapy agents. J Org Chem 1999;64:8495–8450.
46. Das BC, et al. Synthesis of a water soluble carborane containing amino acid as a potential therapeutic agent. Syn Letteve 2001;9:1419–1420.
47. Kabalka GW, Yao M-L. Synthesis of a novel boronated 1-amino-cyclobutanecarboxylic acid as a potential boron neutron capture therapy agent. App Organomet Chem 2003;17: 398–402.
48. Diaz S, Gonzalez A, De Riancho SG, Rodriguez A. Boron complexes of *S*-trityl-L-cysteine and *S*-tritylglutathione. J Organomet Chem 2000;610:25–30.
49. Lindström P, Naeslund C, Sjöberg S. Enantioselective synthesis and absolute configurations of the enantiomers of o-carboranylalanine. Tetrahedron Lett 2000;41:751–754.
50. Masunaga S-I, et al. Potential of α-amino alcohol p-boronophenylalaninol as a boron carrier in boron neutron capture therapy, regarding its enantiomers. J Cancer Res Clin Oncol 2003;129:21–28.
51. Diaz A, Stelzer K, Laramore G, Wiersema R. Pharmacology studies of $Na_2\ ^{10}B_{10}H_{10}$ (GB-10) in human tumor patients. In: Sauerwein W, Moss R, Wittig A, editors. Research and Development in Neutron Capture Therapy. Bologna : Monduzzi Editore, International Proceedings Division; 2002. p 993–999.
52. Hawthorne MF, Feakes DA, Shelly K. Recent results with liposomes as boron delivery vehicles from boron neutron capture therapy. In: Mishima Y, editor Cancer Neutron Capture Therapy. New York: Plenum Press; 1996. p 27–36.
53. Feakes DA, Waller RC, Hathaway DK, Morton VS. Synthesis and in vivo murine evaluation of $Na_4[1\text{-}(1'\text{-}B_{10}H_9)\text{-}6\text{-}SHB_{10}H_8]$ as a potential agent for boron neutron capture therapy. Proc Natl Acad Sci USA 1999;96:6406–6410.
54. Shukla S, et al. Evaluation of folate receptor targeted boronated starburst dendrimer as a potential targeting agent for boron neutron capture therapy. Bioconjugate Chem 2003;14:158–167.
55. Sudimack J, et al. Intracellular delivery of lipophilic boron compound using folate receptor-targeted liposomes. Pharm Res 2002;19:1502–1508.
56. Takagaki M, et al. Boronated dipeptide borotrimethylglycylphenylalanine as a potential boron carrier in boron neutron capture therapy for malignant brain tumors. Radiat Res 2001;156:118–122.
57. Wakamiya T, et al. Synthesis of 4-boronophenylalanine-containing peptides for boron neutron capture therapy of cancer cells. Peptide Sci 1999;36:209–212.
58. Lesnikowski ZJ, Schinazi RF. Boron containing oligonucleotides. Nucleosides Nucleotides 1998;17:635–647.
59. Soloway AH, et al. Identification, development, synthesis and evaluation of boron-containing nucleosides for neutron capture therapy. J Organomet Chem 1999;581:150–155.
60. Lesnikowski ZJ, Shi J, Schinazi RF. Nucleic acids and nucleosides containing carboranes. J Organo-met Chem 1999;581: 156–169.
61. Lunato AJ, et al. Synthesis of 5-(carboranylalkylmercapto)-2′-deoxyuridines and 3-(carboranylalkyl)thymidines and their evaluation as substrates for human thymidine kinases 1 and 2. J Med Chem 1999;42:3378–3389.
62. Al-Madhoun AS, et al. Synthesis of a small library of 3-(carboranylalkyl)thymidines and their biological evaluation as substrates for human thymidine kinases 1 and 2. J Med Chem 2002;45:4018–4028.
63. Schinazi RF, et al. Treatment of Isografted 9L rat brain tumors with *b*-5-o-carboranyl-2'-deoxyuridine neutron capture therapy. Clin Cancer Res 2000;6:725–730.
64. Al-Madhoun AS, et al. Evaluation of human thymidine kinase 1 substrates as new candidates for boron neutron capture therapy. Cancer Res 2004;64:6280–6286.
65. Barth RF, et al. Boron containing nucleosides as potential delivery agents for neutron capture therapy of brain tumors. Cancer Res 2004;64:6287–6295.
66. Sjöberg S, et al. Chemistry and biology of some low molecular weight boron compounds for boron neutron capture therapy. J Neuro-Oncol 1997;33:41–52.
67. Tietze LF, et al. Novel carboranes with a DNA binding unit for the treatment of cancer by boron neutron capture therapy. ChemBio-Chem 2002;3:219–225.
68. Bateman SA, Kelly DP, Martin RF, White JM. DNA binding compounds. VII. Synthesis, characterization and DNA binding capacity of 1,2-dicarba-*closo*-dodecaborane bibenzimidazoles related to the DNA minor groove binder Hoechst 33258. Aust J Chem 1999;52:291–301.
69. Woodhouse SL, Rendina LM. Synthesis and DNA-binding properties of dinuclear platinum(II)-amine complexes of 1,7-dicarba-*closo*-dodecaborane(12). Chem Commun 2001; 2464–2465.
70. Cai J, et al. Boron-containing polyamines as DNA-targeting agents for neutron capture therapy of brain tumors: synthesis and biological evaluation. J Med Chem 1997;40:3887–3896.
71. Zhuo J-C, et al. Synthesis and biological evaluation of boron-containing polyamines as potential agents for neutron capture therapy of brain tumors. J Med Chem 1999;42: 1281–1292.
72. Martin B, et al. *N*-Benzylpolyamines as vectors of boron and fluorine for cancer therapy and imaging: synthesis and biological evaluation. J Med Chem 2001;44:3653–3664.
73. El-Zaria ME, Doerfler U, Gabel D. Synthesis of [(amino-alkylamine)-*N*-amino-alkyl] azanonaborane(11) derivatives for boron neutron capture therapy. J Med Chem 2002;45: 5817–5819.
74. Nakanishi A, et al. Toward a cancer therapy with boron-rich oligomeric phosphate diesters that target the cell nucleus. Proc Natl Acad Sci USA 1999;96:238–241.
75. Maderna A, et al. Synthesis of a porphyrin-labelled carboranyl phosphate diester: a potential new drug for boron neutron capture therapy of Cancer. Chem Commun 2002; 1784–1785.
76. Vicente MGH. Porphyrin-based sensitizers in the detection and treatment of cancer: recent progress. Curr Med Chem Anti-Cancer Agents 2001;1:175–194.
77. Bregadze VI, Sivaev IB, Gabel D, Wöhrle D. Polyhedral boron derivatives of porphyrins and phthalocyanines. J Porphyrins Phthalocyanines 2001;5:767–781.
78. Evstigneeva RP, et al. Carboranylporphyrins for boron neutron capture therapy of cancer. Curr Med Chem: Anti-Cancer Agents 2003;3:383–392.
79. Vicente MGH, et al. Syntheses, toxicity and biodistribution of two 5,15-di[3,5-(*nido*-carboranyl-methyl)phenyl] porphyrin in EMT-6 tumor bearing mice. Bioorg Med Chem 2003;11: 3101–3108.
80. Miura M, et al. Evaluation of carborane-containing prophyrins as tumour agents for boron neutron capture therapy. Br J Radiol 1998;71:773–781.
81. Miura M, et al. Biodistribution of copper carboranyltetraphenylporphyrins in rodents bearing an isogeneic or human neoplasm. J Neuro-Oncol 2001;52:111–117.
82. Miura M, et al. Boron neutron capture therapy of a murine mammary carcinoma using a lipophilic carboranyltetraphenylporphyrin. Radiat Res 2001;155:603–610.

83. Gottumukkala V, Luguya R, Fronczek FR, Vicente MGH. Synthesis and cellular studies of an octa-anionic 5,10,15,20-tetra[3,5-(*nido*-carboranylmethyl)phenyl] porphyrin (H_2OCP) for application in BNCT. Bioorg Med Chem 2005;13: 1633–1640.
84. Hao E, Vicente MGH. Expeditious synthesis of porphyrin-cobaltacarborane vonjugates. Chem Commun 2005; 1306–1308.
85. Ongayi O, Gottumukkala V, Fronczek FR, Vicente MGH. Synthesis and characterization of a carboranyl-tetrabenzoporphyrin. Bioorg Med Chem Lett 2005;15:1665–1668.
86. Fabris C, Jori G, Giuntini F, Roncucci G. Photosensitizing properties of a boronated phthalocyanine: studies at the molecular and cellular level. J Photochem Photobiol B: Biol 2001;64:1–7.
87. Giuntini F, et al. Synthesis of tetrasubstitited Zn(II)-phthalocyanines carrying four carboranyl-units as potential BNCT and PDT agents. Tetrahedron Lett 2005;46:2979–2982.
88. Luguya R, Fronckzek FR, Smith KM, Vicente MGH. Synthesis of novel carboranylchlorins with dual application in boron neutron capture therapy (BNCT) and photodynamic therapy (PDT). Appl Rad Isotopes 2004;61:1117–1123.
89. Rosenthal MA, Kavar B, Uren S, Kaye AH. Promising survival in patients with high-grade gliomas following therapy with a novel boronated porphyrin. J Clin Neurosc 2003; 10:425–427.
90. Rosenthal MA, et al. Phase I and pharmacokinetic study of photodynamic therapy for high-grade gliomas using a novel boronated porphyrin. J Clin Oncol 2001;19:519–524.
91. Hill JS, et al. Selective tumor kill of cerebral glioma by photodynamic therapy using a boronated phorhyrin photosensitizer. Proc Natl Acad Sci USA 1995;92:12126–12130.
92. Kawabata S, et al. Evaluation of the carboranyl porphyrin H_2TCP as a delivery agent for boron neutron capture therapy (BNCT). Khamlichi A, editor. 13th World Congress of Neurological Surgery, Marakesh, Morocco. June 19–24, 2005. p 975–979.
93. Ozawa T, et al. *In vivo* evaluation of the boronated porphyrins TABP-1 in U-87 MG intracerebral human glioblastoma xenografts. Mol Pharmaceut 2004;5:368–374.
94. Lauceri R, Purrello R, Shetty SJ, Vicente MGH. Interactions of anionic carboranylated porphyrins with DNA. J Am Chem Soc 2001;123:5835–5836.
95. Vicente MGH, et al. Synthesis, dark toxicity and induction of in vitro DNA photodamage by a tetra(4-nido-carboranylphenyl)porphyrin. J Photochem Photobiol B: Biol 2002;68:123–132.
96. Ghaneolhosseini H, Tjarks W, Sjoberg S. Synthesis of novel boronated acridines and spermidines as possible agents for BNCT. Tetrahedron 1998;54:3877–3884.
97. Gedda L, et al. Cytotoxicity and subcellular localization of boronated phenanthridinium analogs. Anti-Cancer Drug Design 1997;12:671–685.
98. Gedda L, et al. The influence of lipophilicity on binding of boronated DNA-intercalating compounds in human glioma spheroids. Anti-Cancer Drug Design 2000;15:277–286.
99. Giovenzana GB, et al. Synthesis of carboranyl derivatives of alkynyl glycosides as potential BNCT agents. Tetrahedron 1999;55:14123–14136.
100. Tietze LF, et al. Ortho-carboranyl glycosides for the treatment of cancer by boron neutron capture therapy. Bioorg Med Chem 2001;9:1747–1752.
101. Orlova AV, et al. Conjugates of polyhedral boron compounds with carbohydrates. 1. New approach to the design of selective agents for boron neutron capture therapy of cancer. Russ Chem Bull 2003;52:2766–2768.
102. Tietze LF, Bothe U. Ortho-carboranyl glycosides of glucose, mannose, maltose and lactose for cancer treatment by boron neutron-capture therapy. Chem Eur J 1998;4:1179–1183.
103. Raddatz S, et al. Synthesis of new boron-rich building blocks for boron neutron capture therapy or energy-filtering transmission electron microscopy. ChemBioChem 2004;5:474–482.
104. Tietze LF, et al. Novel carboranyl C-glycosides for the treatment of cancer by boron neutron capture therapy. Chem Eur J 2003;9:1296–1302.
105. Basak P, Lowary TL. Synthesis of conjugates of L-fucose and *ortho*-carborane as potential agents for boron neutron capture therapy. Can J Chem 2002;80:943–948.
106. Endo Y, et al. Structure–activity study of estrogenic agonists bearing dicarba-*closo*-dodecaborane. Effect of geometry and separation distance of hydroxyl groups at the ends of molecules. Bioorg Med Chem Lett 1999;9:3313–3318.
107. Lee J-D, et al. A convenient synthesis of the novel carboranyl-substituted tetrahydroisoquinolines: application to the biologically active agent for BNCT. Tetrahedron Lett 2002;43:5483–5486.
108. Valliant JF, Schaffer P, Stephenson KA, Britten JF. Synthesis of Boroxifen, a *nido*-carborane analogue of tamoxifen. J Org Chem 2002;67:383–387.
109. Feakes DA, Spinler JK, Harris FR. Synthesis of boron-containing cholesterol derivatives for incorporation into unilamellar liposomes and evaluation as potential agents for BNCT. Tetrahedron 1999;55:11177–11186.
110. Endo Y, et al. Potent estrogen agonists based on carborane as a hydrophobic skeletal structure: a new medicinal application of boron clusters. Chem Biol 2001;8:341–355.
111. Tjarks W, et al. *In vivo* evaluation of phosphorous-containing derivatives of dodecahydro-*closo*-dodecaborate for boron neutron capture therapy of gliomas and sarcomas. Anticancer Res 2001;21:841–846.
112. Adams DM, Ji W, Barth RF, Tjarks W. Comparative *in vitro* evaluation of dequalinium B, a new boron carrier for neutron capture therapy (NCT). Anticancer Res 2000;20: 3395–3402.
113. Zakharkin LI, et al. Synthesis of bis(dialkylaminomethyl)-*o*- and *m*-carboranes and study of these compounds as potential preparations for boron neutron capture therapy. Pharm Chem J 2000;34:301–304.
114. Barth RF, et al. Boronated starburst dendrimer-monoclonal antibody immunoconjugates: evaluation as a potential delivery system for neutron capture therapy. Bioconjug Chem 1994;5:58–66.
115. Liu L, et al. Critical evaluation of bispecific antibodies as targeting agents for boron neutron capture therapy of brain tumors. Anticancer Res 1996;16:2581–2588.
116. Liu L, et al. Bispecific antibodies as targeting agents for boron neutron capture therapy of brain tumors. J Hematother 1995;4:477–483.
117. Novick S, et al. Linkage of boronated polylysine to glycoside moieties of polyclonal antibody; Boronated antibodies as potential delivery agents for neutron capture therapy. Nuclear Med Biol 2002;29:93–101.
118. Wu G, et al. Site-specific conjugation of boron containing dendrimers to anti-EGF receptor monoclonal antibody cetuximab (IMC-C225) and its evaluation as a potential delivery agent for neutron capture therapy. Bioconjugate Chem 2004;15:185–194.
119. Fallot T, et al. A Phase I study of an anti-epidermal growth factor receptor monoclonal antibody for the treatment of malignant gliomas. Neurosurgery 1996;39:478–483.

120. Carlsson J, et al. Strategy for boron neutron capture therapy against tumor cells with over-expression of the epidermal growth factor receptor. Int J Radiat Oncol Biol Phys 1994;30:105–115.
121. Capala J, et al. Boronated epidermal growth factor as a potential targeting agent for boron neutron capture therapy of brain tumors. Bioconjugate Chem 1996;7:7–15.
122. Sauter G, et al. Patterns of epidermal growth factor receptor amplification in malignant gliomas. Am J Pathol 1996;148: 1047–1053.
123. Schwechheimer K, Huang S, Cavenee WK. EGFR gene amplification-rearrangement in human glioblastoma. Int J Cancer 1995;62:145–148.
124. Backer MV, Backer JM. Targeting endothelial cells overexpressing VEGFR-2: selective toxicity of Shiga-like toxin-VEGF fusion proteins. Bioconjugate Chem 2001;12: 1066–1073.
125. Backer MV, et al. Vascular endothelial growth factor selectively targets boronated dendrimers to tumor vasculature. Mol Cancer Therapeut 2005;4:1423–1429.
126. Feakes DA, Shelly K, Hawthorne M. Selective boron delivery to murine tumors by lipophilic species incorporated in the membranes of unilamellar liposomes. Proc Natl Acad Sci USA 1995;92:1367–1370.
127. Carlsson J, et al. Ligand liposomes and boron neutron capture therapy. J Neuro-Oncol 2003;62:47–59.
128. Pardridge WM. Drug delivery to the brain. J Cerebral Blood Flow Metabol 1997;17:713–731.
129. Mendelsohn, J. Targeting the epidermal growth factor receptor for cancer therapy. J Clin Oncol 2002;20:1s–13s.
130. Nygren P, Sorbye H, Osterland P, Pfeiffer P. Targeted drugs in metastatic colorectal cancer with emphasis on guidelines for the use of bevacizumab and cetuximab. An Acta Oncologica expert report. Acta Oncolog 2005;44:203–218.
131. Wikstrand CJ, Cokgor I, Sampson JH, Bigner DD. Monoclonal antibody therapy of human gliomas: current status and future approaches. Cancer Metastasis Rev 1999;18: 451–464.
132. Barth RF, et al. Molecular targeting of the epidermal growth factor receptor for neutron capture therapy of gliomas. Cancer Res 2002;62:3159–3166.
133. Yang W, et al. Convection enhanced delivery of boronated epidermal growth factor for molecular targeting of EGFR positive gliomas. Cancer Res 2002;62:6552–6558.
134. Barth RF, et al. Neutron capture therapy of epidermal growth factor positive gliomas using boronated cetuximab (IMC-C225) as a delivery agent. App Radiat Isotopes 2004;61: 899–903.
135. Yang W, et al. Boronated epidermal growth factor as a delivery agent for neutron capture therapy of EGFR positive gliomas. App Rad Isotopes 2004;61:981–985.
136. Barth RF, et al. Enhanced delivery of boronophenylalanine for neutron capture therapy of brain tumors using the bradykinin analogue, Cereport™(RMP7). Neurosurgery 1999; 44:350–359.
137. Barth RF, et al. Neutron capture therapy of intracerebral melanoma: Enhanced survival and cure following blood–brain barrier opening to improve delivery of boronophenylalanine. Int J Radiat Oncol Biol Phys 2002;52:858–868.
138. Yang W, et al. Development of a syngeneic rat brain tumor model expressing EGFRvIII and its use for molecular targeting studies with monoclonal antibody L8A4. Clin Cancer Res 2005;11:341–350.
139. Harling O, Riley K. Fission reactor neutron sources for neutron capture therapy—a critical review. J Neuro-Oncol 2003;2:7–17.
140. Harling O, et al. The fission converter-based epithermal neutron irradiation facility at the Massachusetts Institute of Technology Reactor. Nuclear Sci Eng 2002;140:223–240.
141. Capala J, et al. Boron neutron capture therapy for glioblastoma multiforme: clinical studies in Sweden. J Neuro-Oncol 2003;62:135–144.
142. Joensuu H, et al. Boron neutron capture therapy of brain tumors: clinical trials at the Finnish Facility using boronophenylalanine. J Neuro-Oncol 2003;62:123–134.
143. Moss RL, et al. Design, construction and installation of an epithermal neutron beam for BNCT at the High Flux Reactor Petten. In: Allen BJ, et al., editors. Progress in Neutron Capture Therapy for Cancer, New York: Plenum Press; 1992. p 63–66.
144. Marek M, Viererbl M, Burian J, Jansky B. Determination of the geometric and spectral characteristics of BNCT beam (neutron and gamma-ray). In: Hawthorne MF, Shelly K, Wiersema RJ, editors., Neutron Capture Therapy, Vol. I, New York: Kluwer Academic/Plenum Publishers; 2001. p 381–389.
145. Kobayashi T, et al. The remodeling and basic characteristics of the heavy water neutron irradiation facility of the Kyoto University Research Reactor, Mainly for Neutron Capture Therapy. Nucl Technol 2000;131:354–378.
146. Yamamoto K, et al. Characteristics of neutron beams for BNCT. Proceedings of the 9th Symposium on Neutron Capture Therapy, Osaka, Japan, October 2–6, 2000. p 243–244.
147. Blaumann HR, Larrieu OC, Longhino JM, Albornoz AF. NCT facility development and beam characterisation at the RA-6 Reactor. In: Hawthorne MF, Shelly K, Wiersema RJ, editors. Frontiers in Neutron Capture Therapy. Vol. I, New York: Kluwer Academic/Plenum Publishers; 2001. p 313–317.
148. Agosteo S, et al. Design of neutron beams for boron neutron capture therapy in a fast reactor. IAEA Technical Committee Meeting about the Current Issues Relating to Neutron Capture Therapy, June 14–18, 1999, Vienna, Austria.
149. Fairchild RG, et al. Installation and testing of an optimized epithermal neutron beam at the Brookhaven Medical Research Reactor (BMRR). Proceedings of the Workshop on Neutron Beam Design, Development and Performance for Neutron Capture Therapy. MIT, Cambridge (MA), March 29–31, 1989.
150. Hatanaka H. Boron neutron capture therapy for brain tumors. In: Karin ABMF, Laws E, editor. Glioma: Berlin: Springer-Verlag; 1991. p 233–249.
151. Hatanaka H, Nakagawa Y. Clinical results of long-surviving brain tumor patients who underwent boron neutron capture therapy. Int J Radiat Oncol Biol Phys 1994;28:1061–1066.
152. Diaz AZ. Assessment of the results from the phase I/II boron neutron capture therapy trials at the Brookhaven National Laboratory from a clinician's point of view. J Neuro-Oncol 2003;62:101–109.
153. Riley K, Binns P, Harling O. Performance characteristics of the MIT fission converter based epithermal neutron beam. Phys Med Biol 2003;48:943–958.
154. Nigg D, et al. Initial neutronic performance assessment of an epithermal neutron beam for neutron capture therapy research at Washington State University. Research and Development in Neutron Capture Therapy. Proceedings of the 10th International Congress on Neutron Capture Therapy, 2002. p 135–139.
155. Beynon T, et al. Status of the Birmingham accelerator-based BNCT facility. Proceedings of the 10th International Congress on Neutron Capture Therapy, 2002. p 225–228.

156. Burlon A, et al. Optimization of a neutron production target and beam shaping assembly based on the $^{7}Li(p,n)^{7}Be$ reaction. Proceedings of the 10th International Congress on Neutron Capture Therapy, 2002. p 229–234.
157. Kononov O, et al. Investigations of using near-threshold $^{7}Li(p,n)^{7}Be$ reaction for NCT based on in-phantom dose distribution. Proceedings of the 10th International Congress on Neutron Capture Therapy, 2002. p 241–246.
158. Blackburn B, Yanch J, Klinkowstein R. Development of a high-power water cooled beryllium target for use in accelerator-based boron neutron capture therapy. Med Phys 1998;10:1967–1974.
159. Hawk A, Blue T, Woollard J, Gupta N. Effects of target thickness on neutron field quality for an ABNS. Research and Development in Neutron Capture Therapy Proceedings of the 10th International Congress on Neutron Capture Therapy, 2002. p 253–257.
160. Sakurai Y, Kobayashi T, Ono K. Study on accelerator-based neutron irradiation field aiming for wider application in BNCT - spectrum shift and regional filtering. Proceedings of the 10th International Congress on Neutron Capture Therapy, 2002. p 259–263.
161. Giusti V, Esposito J. Neutronic feasibility study of an accelerator-based thermal neutron irradiation cavity. Proceedings of the 10th International Congress on Neutron Capture Therapy, 2002. p 305–308.
162. Starling WJ. RFI Linac for accelerator-based neutrons. Abstracts of the 11th World Congress on Neutron Capture Therapy. Boston, October 11–15, 2004. p 45.
163. Locher GL. Biological effects and therapeutic possibilities of neutrons. Am J Roetgenol Radium Ther 1936;36:1–13.
164. Asbury AK, Ojeann, Nielson SL, Sweet WH. Neuropathologic study of fourteen cases of malignant brain tumor treated by boron-10 slow neutron capture therapy. J Neuropathol Exp Neurol 1972;31:278–303.
165. Sweet WH. Practical problems in the past in the use of boron-slow neutron capture therapy in the treatment of glioblastoma multiforme. Proceedings First International Symposium Neutron Capture Therapy, Brookhaven National Lab Reports 51730. October 12–14, 1983. p 376–378.
166. Slatkin DN. A history of boron neutron capture therapy of brain tumours. Postulation of a brain radiation dose tolerance limit. Brain 1991;114:1609–1629.
167. Nakagawa Y, Hatanaka H. Boron neutron capture therapy: Clinical brain tumor studies. J Neuro-Oncol 1997;33:105–115.
168. Nakagawa Y, et al. Clinical review of the Japanese experience with boron neutron capture therapy and a proposed strategy using epithermal neutron beams. J Neuro-Oncol 2003;62:87–99.
169. Laramore GE, et al. Boron neutron capture therapy: a mechanism for achieving a concomitant tumor boost in fast neutron radiotherapy. Int J Radiat Oncol Biol Phys 1994;28:1135–1142.
170. Kageji T, et al. Pharmacokinetics and boron uptake of BSH ($Na_2B_{12}H_{11}SH$) in patients with intracranial tumors. J Neuro-Oncol 1997;33:117–130.
171. Bergland R, et al. A Phase 1 trial of intravenous boronophenylalanine-fructose complex in patients with glioblastoma multiforme. Cancer Neutron Capture Therapy. Mishima Y, editor. New York: Plenum Press; 1996. p 739–746.
172. Coderre JA, et al. Biodistribution of boronophenylalanine in patients with glioblastoma multiforme: Boron concentration correlates with tumor cellularity. Radiat Res 1998; 149:163–170.
173. Chanana AD, et al. Boron neutron capture therapy for glioblastoma multiforme: interim results from the phase I/II dose-escalation studies. Neurosurgery 1999;44:1182–1193.
174. Busse PM, et al. A critical examination of the results from the Harvard-MIT NCT program phase I clinical trial of neutron capture therapy for intracranial disease. J Neuro-Oncol 2003;111–121.
175. Palmer MR, et al. Treatment planning and dosimetry for the Harvard-MIT phase I clinical trial of cranial neutron capture therapy. Int J Radiat Oncol Biol Phys 2002;53:1361–1379.
176. Wittig A, et al. Current clinical results of the EORTC-study 11961, in: Research and Development in Neutron Capture Therapy. Sauerwein W, Moss R, Wittig A, editors. Bologna: Monduzzi Editore; 2002. p 1117–1122.
177. Burian J, et al. Report on the first patient group of the Phase I BNCT trial at the LVR-15 reactor. Sauerwein W, Moss R, Wittig A, editors. Bologna, Italy: Monduzzi Editore; 2002. p 1107–1112.
178. Coderre JA, et al. Tolerance of normal human brain to boron neutron capture therapy. Appl Radiat Isotopes 2004;61: 1084–1087.
179. Emami B, et al. Tolerance of normal tissue to therapeutic irradiation. Int J Radiat Oncol Biol Phys 1991;21:109–122.
180. Flickinger JC, et al. Development of a model to predict permanent symptomatic postradiosurgery injury for arteriovenous malformation. Arteriovenous Malformation Radiosurgery Study Group. Int J Radiat Oncol Biol Phys 2000;46:1143–1148.
181. Wittig A, Sauerwein WA, Coderre JA. Mechanisms of transport of *p*-borono-phenylalanine through the cell membrane in vitro. Radiat Res 2000;153:173–180.
182. Joel DD, et al. Effect of dose and infusion time on the delivery of *p*-boronophenylalanine for neutron capture therapy. J Neuro-Oncol 1999;41:213–221.
183. Morris GM, et al. Long-term infusions of p-boronophenylalanine for boron neutron capture therapy: evaluation using rat brain tumor and spinal cord models. Radiat Res 2002;158:743–752.
184. Smith DR, Chandra S, Coderre JA, Morrison GH. Ion microscopy imaging of ^{10}B from p-boronophenylalanine in a brain tumor model for boron neutron capture therapy. Cancer Res 1996;56:4302–4306.
185. Dahlström M, et al. Accumulation of boron in human malignant glioma cells *in vitro* is cell type dependent. J Neuro-Oncology 2004;68:199–205.
186. Bergenheim AT, Capala J, Roslin M, Henriksson R. Distribution of BPA and metabolic assessment in glioblastoma patients during BNCT treatment: a microdialysis study. J Neuro-Oncol 2005;71:287–293.
187. Henriksson R, et al. Boron neutron capture therapy (BNCT) for glioblastoma multiforme: A phase 2 study evaluating a prolonged high dose of boronophenylalanine (BPA) at the Studsvik facility in Sweden. Radiother Oncol (Submitted).
188. Mishima Y, et al. New thermal neutron capture therapy for malignant melanoma. Melanogenesis-seeking ^{10}B molecular-melanoma cell interaction from in vitro to first clinical trial. Pigment Cell Res 1989;2:226–234.
189. Hiratsuka J, Kono, Mishima Y. RBEs of thermal neutron capture therapy and $^{10}B(n,\alpha)^{7}Li$ reaction on melanoma-bearing hamsters. Pigment Cell Res 1989;2:352–355.

190. Tsuji M, Ichihashi M, Mishima Y. Selective affinity of ^{10}B-paraboronophenylalanine-HCI to malignant melanoma for thermal neutron capture therapy. Jpn J Dermatol 1983;93: 773–778.
191. Coderre JA, et al. Selective delivery of boron by the melanin precursor analog p-boronophenylalanine to tumors other than melanoma. Cancer Res 1990;50:138–141.
192. Mishima Y, et al. Treatment of malignant melanoma by single neutron capture therapy with melanoma-seeking ^{10}B-compound. Lancet 1989;1:388–389.
193. Mishima Y. Melanoma and nonmelanoma neutron capture therapy using gene therapy: overview. In: Larsson B, Crawford J and Weinreich, editors. Advances in Neutron Capture Therapy Vol. 1, Medicine and Physics. Elsevier; 1997. p 10–25.
194. Madoc-Jones H, et al. A phase-I dose-escalation trial of boron neutron capture therapy for subjects with metastatic subcutaneous melanoma of the extremities. In: Mishima Y, editor. Cancer Neutron Capture Therapy. New York and London: Plenum Press; 1996. p 707–716.
195. Yoshino K, et al. Improvement of solubility of p-boronophenylalanine by complex formation with monosaccharides. Strahlenther Onkol 1989;165:127–129.
196. Busse PM, et al. The Harvard-MIT BNCT Program: overview of the clinical trials and translational research. Proceedings of the 11th International Congress of Radiation Research, Vol 2. Dublin, Ireland, July 18–23, 1999. p 702–709.
197. Gonzalez SJ, et al. First BNCT treatment of a skin melanoma in Argentina: dosimetric analysis and clinical outcome. Appl Radiat Isotopes 2004;61:1101–1105.
198. Barth RF, et al. A nude rat model for neutron capture therapy of human intracerebral melanoma. Int J Radiat Oncol Biol Phys 1994;28:1079–1088.
199. Barth RF, et al. Neutron capture therapy of intracerebral melanoma: Enhanced survival and cure following blood-brain barrier opening to improve delivery of boronophenylalanine. Int J Radiat Oncol Biol Phys 2002;52:858–868.
200. Barth RF, et al. Boron neutron capture therapy of brain tumors: enhanced survival and cure following blood-brain barrier disruption and intracarotid injection of sodium borocaptate and boronophenylalanine. Int J Radiat Oncol Biol Phys 2000;47:209–218.
201. Kato I, et al. Effectiveness of BNCT for recurrent head and neck malignancies. Appl Radiat Isotopes 2004;61:1069–1073.
202. Pinelli T, et al. TAOrMINA: from the first idea to the application to the human liver. In: Research and Development in Neutron Capture Therapy. In: Sauerwein W, Moss R, Wittig A, editors. Bologna, Italy: Monduzzi Editore; 2002. p 1065–1072.
203. Ringe B, Pichlmayr R, Wittekind C, Tusch G. Surgical treatment of hepatocellular carcinoma: experience with liver resection and transplantation in 198 patients. World J Surg 1991;15:27085.
204. Iwatsuki S, et al. Hepatic resection versus transplantation for hepatocellular carcinoma. Ann Surg 1991;214:221–228.
205. Pinelli T. Neutron capture therapy for liver cancer metastases. Abstracts of the Eleventh World Congress on Neutron Capture therapy. Boston, October 11–15, 2004. p 52.
206. Suzuki M, et al. Biodistribution of ^{10}B in a rat liver tumor model following intra-arterial administration of sodium borocaptate (BSH)/degradable starch microspheres (DSM) emulsion. Appl Radiat Isotopes 2004;61:933–937.
207. Koivunoro H, et al. BNCT dose distribution in liver with epithermal D-D and D-T fusion-based neutron beams. Appl Radiat Isotopes 2004;61:853–859.
208. Chou FI, et al. Biological efficacy of BPA in malignant and normal liver cells. Abstract of the Eleventh World Cong on Neutron Capture Therapy. Boston, October 11–15, 2004. p 38.
209. Goodman JH, et al. Boron neutron capture therapy of brain tumors: biodistribution, pharmacokinetics, and radiation dosimetry of sodium borocaptate in glioma patients. Neurosurgery 2000;47:608–622.
210. Hideghéty K, et al. Tissue uptake of BSH in patients with glioblastoma in the EORTC 11961 phase I BNCT trial. J Neuro-Oncol 2003;62:145–156.
211. Coderre JA, et al. Biodistribution of boronophenylalanine in patients with glioblastoma multiforme: boron concentration correlates with tumor cellularity. Radiat Res 1998;149: 163–170.
212. Smith D, et al. Quantitative imaging and microlocalization of boron-10 in brain tumors and infiltrating tumor cells by SIMS ion microscopy: Relevance to neutron capture therapy. Cancer Res 2001;61:8179–8187.
213. Santa Cruz GA, Zamenhof RG. The microdosimetry of the ^{10}B reaction in boron neutron capture therapy: a new generalized theory. Radiat Res 2004;162:702–710.
214. Imahori Y, et al. Positron emission tomography-based boron neutron capture therapy using boronophenylalanine for high-grade gliomas: part 1. Clin Cancer Res 1998;4: 1825–1832.
215. Imahori Y, et al. Positron emission tomography-based boron neutron capture therapy using boronophenylalanine for high-grade gliomas: part 2. Clin Cancer Res 1998;4:1833–1841.
216. Takahasi Y, Imahori Y, Mineura K. Prognostic and therapeutic indicator of fluoroboronophenylalanine positron emission tomography in patients with gliomas. Clin Cancer Res 2003;9:5888–5895.
217. Kabalka GW, et al. The use of positron emission tomography to develop boron neutron capture therapy treatment plans for metastatic malignant melanoma. J Neuro-Oncol 2003;62: 187–195.
218. Bendel P. Biomedical applications of ^{10}B and ^{11}B NMR. NMR Biomed 2005;18:74–82.
219. Bendel P, Margalit R, Salomon Y. Optimized ^{1}H MRS and MRSI methods for the *in vivo* detection of boronophenylalanine. Magn Reson Med 2005;53:1166–1171.
220. Kajimoto Y, et al. Boron neutron capture therapy selectively destroys tumor cells preserving neurons in co-existing tumor lesion of malignant glioma. (Submitted)
221. Perry JR, et al. Challenges in the design and conduct of phase III brain tumor therapy trials. Neurology 1997; 49:912–917.
222. Shapiro W. Bias in uncontrolled brain tumor trials. Can J Neurol Sci 1997;24:269–270.
223. Stupp R, et al. Promising survival for patients with newly diagnosed glioblastoma multiforme treated with concomitant radiation plus temozolomide followed by adjuvant temozolomide. J Clin Oncol 2002;20:1375–1382.
224. Stupp R, et al. Radiotherapy plus concomitant and adjuvant temozolomide for gliomblastomas. N Engl J Med 2005;325: 987–996.
225. Agarwala SS, et al. Temozolomide for the treatment of brain metastases associated with metastatic melanoma: a phase II study. J Clin Oncol 2004;22:2101–2107.

See also IMMUNOTHERAPY; MONOCLONAL ANTIBODIES; RADIONUCLIDE PRODUCTION AND RADIOACTIVE DECAY; RADIOTHERAPY, HEAVY ION.

BRACHYTHERAPY, HIGH DOSAGE RATE

RUPAK DAS
University of Wisconsin
Madison, Wisconsin

INTRODUCTION

Brachytherapy is a form of radiotherapy whereby a radioactive source is used inside or at short distance from the tumor. There are three different forms of brachytherapy: interstitial, intracavitary, and skin therapy. In interstitial brachytherapy, the radioactive sources are implanted inside and throughout the tumor volume; in intracavitary brachytherapy the sources are placed in the body cavities very close to the tumor; while in skin therapy the sources are placed on the skin surface. Conventionally, brachytherapy implants have delivered the radiation at a low dose rate (dose rates of < 1 Gy $\cdot$ h^{-1}). Low dose-rate (LDR) interstitial implants can be temporary (meaning that the radioactive sources are left in place for a period of time, usually a few days, and then removed) or permanent (left in place without removal), while intracavitary implants are temporary. The advent of methods to deliver the dose at a much higher dose rates, in the range of 1–5 Gy $\cdot$ min^{-1}, brought an increase in the use of brachytherapy. All high dose-rate (HDR) brachytherapy treatments are temporary and treatments are administered using discrete fractions.

What Is a Remote Afterloader?

A remote afterloader (RAL) is a computer driven system that transports the radioactive source from a shielded safe into the applicator placed in the patient. Upon termination or interruption of the treatment, the source is driven back to its safe. The device may move the source by one of several methods, most commonly pneumatic air pressure or cable drives.

What is Stepping-Source Remote Afterloader?

A stepping source RAL is a particular design of the treatment unit that consists of a single source at the end of a cable that moves the source through applicators placed in the treated volume. The treatment unit can treat implants consisting of many needles or catheters in the patient. Multiple catheters are often required to cover the target with adequate radiation doses. Each catheter or part of an applicator is connected to the RAL through a channel. The computer drives the cable so that the source moves from the safe through a given channel to the programmed position in the applicator (dwell position) for a specific amount of time (dwell time). In any applicator, there may be many dwell positions. After treating all the positions in a given catheter (channel) the source is retracted to its safe and then driven to the next channel. The dwell positions and the dwell time in each channel are independently programmable, thereby giving a high level of flexibility of dose delivery. All currently available HDR RALs use the stepping-source design.

Currently there are three types of HDR RALs available in the market: MicroSelectron (Fig. 1, vendor Nucletron, Veenendaal, Netherlands), Gamma-Med (Fig. 2), and VariSource (Fig. 3, both marketed by Varian Associates, Palo Alto, CA).

The specific features of the three different RALs are shown in Table 1.

COMPONENTS OF A HIGH DOSE RATE REMOTE AFTERLOADER

While different in detail, all available HDR RALs consist of the same general components. Figure 4 gives an overview of the systems, with the major parts described below.

Shielded Safe

To provide a dose rate in the range of 1–5 Gy $\cdot$ min^{-1} in a RAL requires a ^{192}Ir source of 4–10 Ci. A shielded safe, which is an integrated part of the treatment unit, provides enough radiation shielding to house the source while not in treatment mode. Once in treatment mode, the source is driven out of the safe while it follows the program through the dwell positions. In the event of an interruption or termination of the treatment, the source is driven back to the shielded safe.

Radioactive Source

While delivering the HDR brachytherapy requires an intense source, passing the source through needles placed through a tumor requires one of a small size. The radioactive source in an HDR RAL is usually 3–10 mm in length

Figure 1. The Nucletron MicroSelectron V2 HDR RAL. The RAL wheels allow it to be conveniently positioned near the patient. The treatment head is mounted on a telescopic base that allows the head to be raised or lowered to the required height for treatment without moving the patient.

Figure 2. The Varian GammaMed RAL.

Figure 3. The Varian VariSource RAL.

and < 1 mm in diameter, fixed at the end of a steel cable (Figs. 5 and 6). The Nucletron source is placed in a stainless steel capsule and welded to the cable, while the Varian source is placed in a hole drilled into the cable and closed by welding. The ^{192}Ir radionuclide is now used for all HDR RALs, although early versions of HDR RAL used ^{60}Co. A new source has an activity near 10 Ci. Since ^{192}Ir has a half-life of 74 days, the source should be replaced every 3 months to keep the treatment in the HDR radiobiological regime. A trained medical physicist calibrates the source after each installation using a re-entrant well-type ionization chamber (Fig. 7). The chambers themselves are calibrated by secondary calibration laboratories known as Accredited Dosimetry Calibration Laboratories (ADCL). The resulting source calibration is verified against the manufacturer's source calibration.

Source Drive Mechanism

When the RAL unit receives a command to initiate a treatment, the stepper motor connected to the reel containing the drive cable turns, causing the source cable to advances from the shielded safe along a path constrained by transfer tubes to the first treated dwell position in the applicator attached to the first channel. The source dwells at that position for a predetermined duration (dwell time) as calculated by the treatment planning system (see below). After completing that dwell, it goes on to the subsequent dwell positions. Some units step as the source drives out (MicroSelectron), stopping first at the dwell position most proximal to the afterloader, while the other (VariSource and Gamma-Med) the source travels first to the most distal dwell (toward the tip of the applicator), and a bit farther, and then steps as the source returns toward the safe. Stepping on the outward drive obviates any concern about the effect of slack in the drive mechanism affecting the accuracy of the source position. The unit that steps on the way back into the unit includes correction for slack in the calibration of the source location. Upon completion of the treatment for the first channel, the source is retracted into the safe, and redirected to travel to the second channel. The process is repeated for all the subsequent treatment channels. The programmed movement of the source is verified by means of an optical encoder or other devices that compare the angular rotation of the stepper motor or cable length ejected or retracted with the number of pulses sent to the drive motor. This system is capable of detecting catheter obstruction or constriction as increased friction in the cable movement. Under certain fault conditions, if the stepper motor fails to retract the source, a high torque direct current (dc) emergency motor will retract the source.

The confirmation of the source exit from and return to the safe is carried out by an "optopair", consisting of a pair of light-sensitive detector and infrared (IR) light source, that detects the cable when its tip obstructs the light path. All the currently marketed after-loaders are also equipped with check cables or dummy sources. The check cable is an exact duplicate of the radioactive source along with its cable, except not radioactive. Before the ejection of the radioactive source, the check cable is first ejected to check the integrity of the catheter system. After a noneventful check by this "dry run" with the dummy source, the radioactive source is then sent for treatment.

Indexer

The RALs are equipped with an indexer, shown in Fig. 8. The indexer consists of an S-tube (item 14 in Fig. 4) that directs the source cable from the exit of the safe to one of the exit ports from the unit (channels). The various catheters or applicator parts connect to these channels, usually through connecting guides called transfer tubes. Different units have between 3 and 24 channels available for

Table 1. Specific Features of the Three Currently Marketed HDR RALs

	MicroSelectron V2	Gamma Med+	VariSource 200/200t
Vendor	Nucletron	Varian	Varian
Sources	10 Ci of ^{192}Ir	10 Ci of ^{192}Ir	10 Ci of ^{192}Ir
Source dimension	3.5 mm L, 1.1 mm OD	4.52 mm L, 0.9 mm OD	5 mm L, 0.59 mm OD
Channels	18	3 or 24	20
Source extension	1500 mm	1300 mm	1500 mm
Channel length	Variable	Fixed	Variable
Source movement	Stepping forward	Stepping backward	Stepping backward
Step sizes	2.5, 5 or 10 mm	1–10 mm, 1 mm steps	2–99 mm, 1 mm steps
Dwells/channel	48	60	20
Speed of source	50 cm · s^{-1}	60 cm · s^{-1}	50–60 cm · s^{-1}

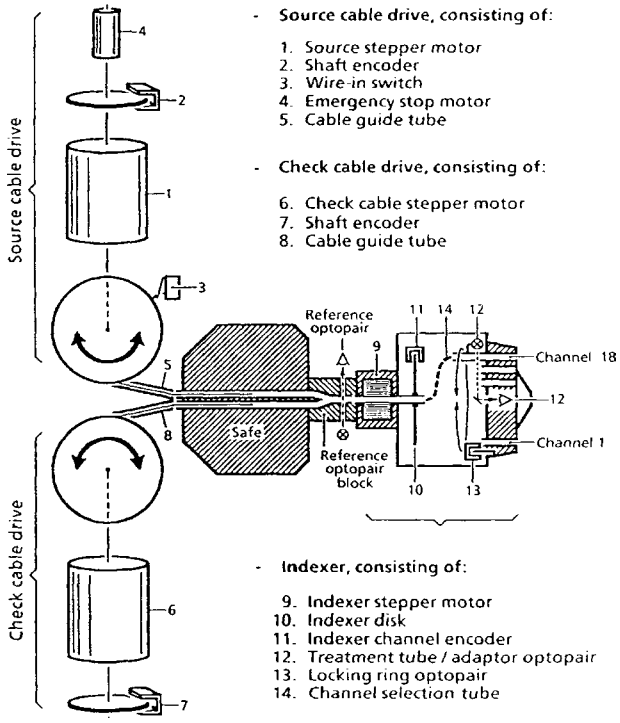


Figure 4. Schematic diagram of a single stepping source RAL. (Courtesy of Nucletron Corporation, Columbia, MD.)

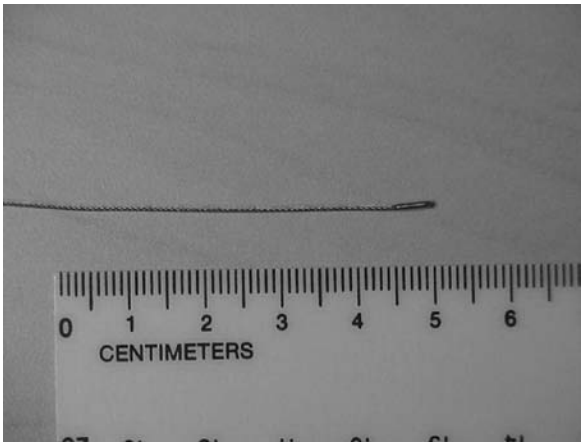


Figure 6. A ^{192}Ir HDR source for the MicroSelectron at the end of a steel drive cable, as shown in Fig. 5.

connection. If a patient's treatment requires more than the number of channels on a given treatment unit, the treatment must be broken into sessions, where the catheters are connected up to the number of channels available and treated Then the transfer tubes are disconnected from the catheters just treated and reconnected to the next set of catheters for continuation of the treatment.

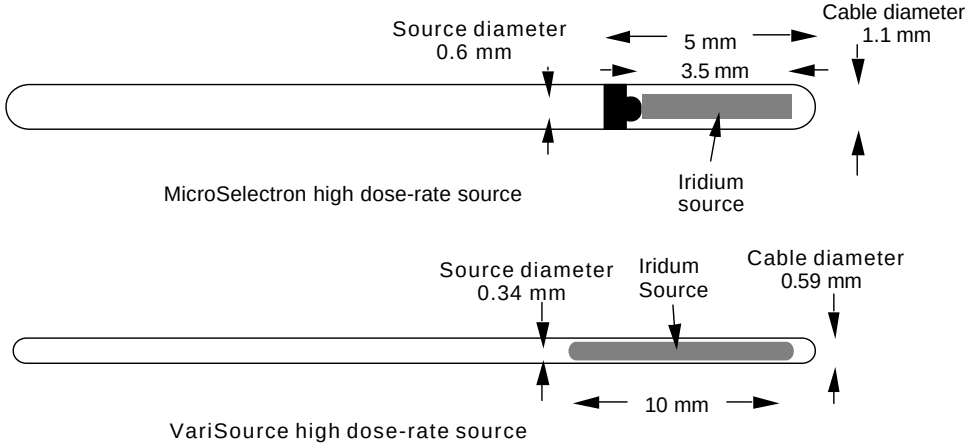


Figure 5. Schematics of the two types of sources used in stepping-source RALs. The VariSource is an earlier version, while the new source has a length of 5 mm.

Figure 7. A re-entrant well-type ionization chamber used for calibration of the HDR brachytherapy sources.

Transfer Tubes

Transfer or guide tubes are long tubes that act as a conduit to transfer the source from the RAL to the applicators or catheters for treatment. One end of the transfer tube is attached to the indexer of the RAL (Fig. 9), while the other end is attached to the interstitial, intracavitary, or transluminal applicators (Fig. 10). The applicator-end of the transfer tube contains spring-loaded ball bearings that block the path through the tube if no applicator is attached. When an applicator is inserted, it pushes aside the ball bearings, opening the path for the source cable. When the

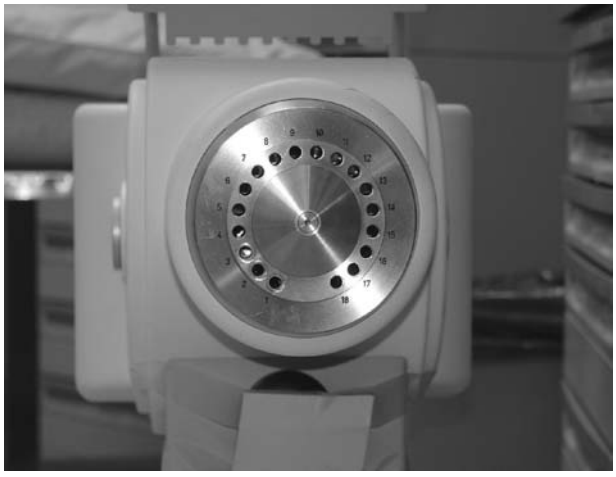


Figure 8. The frontal view of an indexer from the Nucletron, MicroSelectron HDR RAL, consisting of 18 channels.

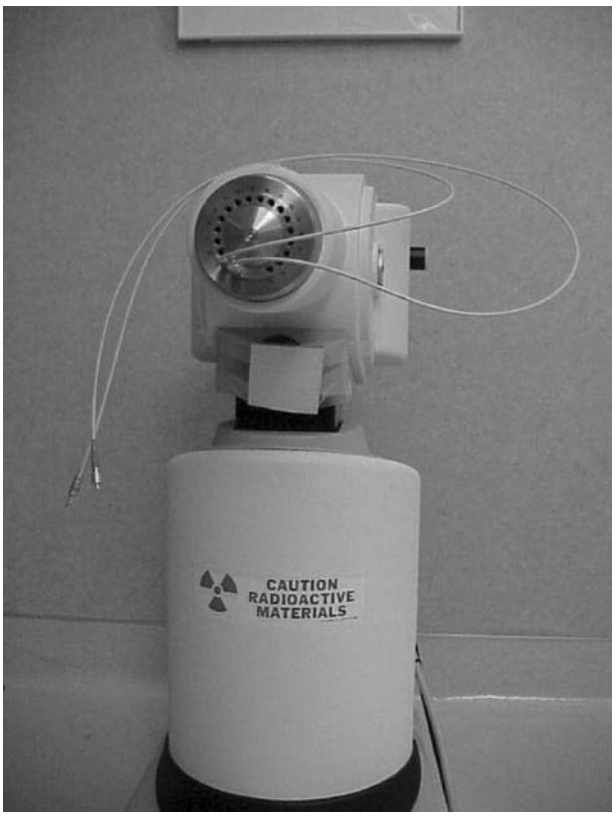


Figure 9. A view of two types of transfer tubes hooked up to the indexer of a RAL.

Figure 10. View of the transfer tubes connected to a gynecological applicator. Ball bearings beneath the gray polymer coating allow verification of the proper connection of the transfer tubes to the applicator. The number 1 and 2 represents that these transfer tubes must connect the channel 1 and 2 of the indexer ring and the similarly numbered parts of the applicator.

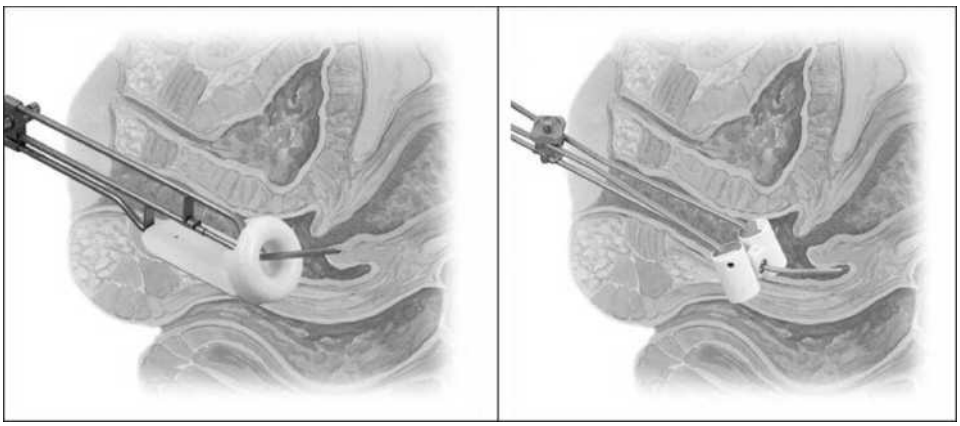


Figure 11. Gynecological applicator used for the treatment of cervical cancer. (Courtesy of Nucletron Corporation, Columbia, MD.)

check cable makes its test run, if no applicator is attached to the transfer tube, the check cable hits the obstacle of the ball bearings, and prevents ejection of the source. Each type of applicator has its own type of transfer tube.

Applicators

An array of applicators for different treatment sites are marketed by each vendor. Each vendor designs their own applicators that can only be used with their transfer tubes and HDR RALs. Figure 11 shows two cervical applicators marketed by Nucletron used for the treatment of cervical cancer.

Treatment Control Station

The treatment control station (Fig. 12) allows the user to select the source travel and dwell sequence to be used in each channel. This can be entered by three ways: (*1*) manually by the keyboard/mouse at the control station; (*2*) recalling a standard plan from the computer and then editing the data without affecting the standard plan from which it originated; or (*3*) by importing the data from a treatment planning system via transfer medium or a network connection to the treatment control station.

Treatment Control Panel

The treatment control station transfers the data to the treatment control panel. A hard or soft START button initiates the execution of the treatment according to the program. In addition, there is an INTERRUPT button, which when pressed retracts the source and stops the timer, allowing the user to enter the treatment room without receiving radiation exposure. A RESUME or START button resumes the treatment from the time and the dwell position where it was interrupted. A master EMERGENCYOFF button initiates the high torque dc emergency motor to retract the source. In the normal course of a successful termination of the treatment, the timer runs to zero and the machine automatically retracts the source. Figure 13 shows an example of the treatment control panel.

SAFETY FEATURES

The HDR RALs are complicated devices containing very high activity radioactive sources. Serious accidents can

Figure 12. A view of the monitor of the RAL treatment control station.

Figure 13. The Treatment Control Panel of the Nucletron, microSelectron HDR RAL. The START button is the white button on the right side of the panel, while the EMERGENCY OFF button is the top button on the left side of the panel.

Figure 14. A view of the access panel of the MicroSelectron treatment unit. The center button is an emergency stop button on the treatment unit. Also showing are the manual retraction of the radioactive source cable (left) and the check cable (right).

happen quickly. All such units have many safety features and operational interlocks to prevent errant source movement or facilitate rapid operator response in the event of a system failure.

Emergency Switches

Numerous EMERGENCY OFF switches are located at convenient places and are easily accessible, in case a situation arises. One EMERGENCY OFF switch is located on the control panel. Another EMERGENCY OFF button is located on the top of the remote afterloader treatment head. Vendors usually install one or two emergency switches in the walls of the treatment room. In the event a treatment is initiated with someone other than the patient in the treatment room, that person can stop the treatment and retract the source by pressing the EMERGENCY OFF button. Figure 14 shows the EMERGENCY OFF switch on the treatment unit.

Emergency Crank

All HDR RALs have emergency cranks to retract the source cable if the source fails to retract normally and the emergency motor also fails to reel in the source. Figure 14 shows such a crank for the MicroSelectron and Fig. 15 for the VariSource. Using the crank requires the operator to enter the room with the source unshielded. Exposure rates for this situation are considered below.

Door Interlock

Interlock switches prevent initiation of a treatment with the door open. While in progress, opening the door interrupts the treatment. This safety feature protects the medical personnel from radiation exposure, in the event somebody enters the treatment room without the knowledge of the operator. If a door is inadvertently opened during the treatment, the treatment is interrupted and

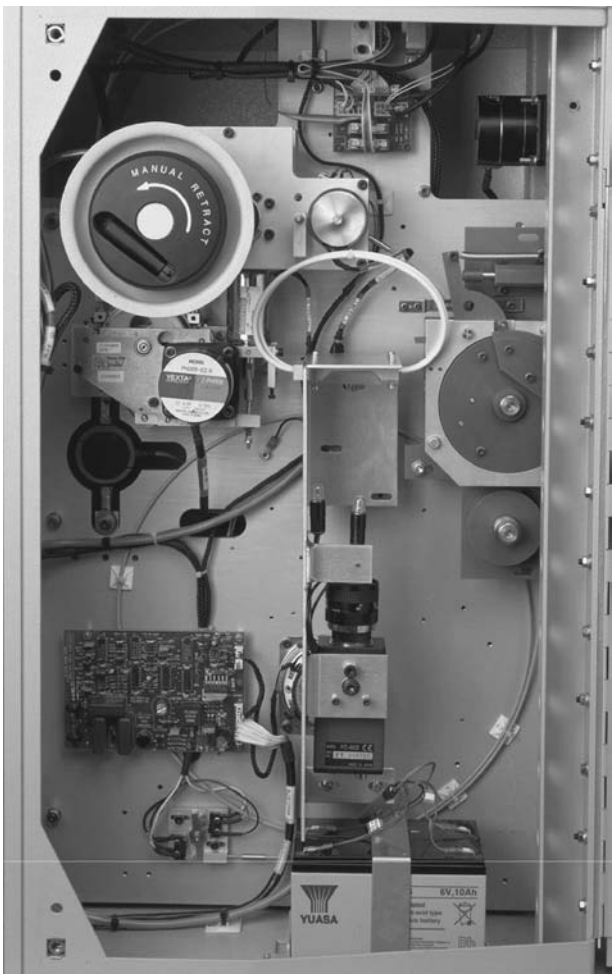


Figure 15. The back panel of the VariSource showing the crank for manual source retraction in an emergency.

the source returns to the safe. The treatment can be resumed at the same point where it was interrupted by closing the door and pressing the START or the RESUME button at the control panel.

Audio–Visual System

All HDR brachytherapy suites are equipped with a closed circuit television system (CCTV) or shielded windows and/or mirrors for observing the patient, and a two-way audio system to communicate with the patient during treatment.

Radiation Monitor and Treatment on Indicator

Three separate independent systems alert personnel when the source is not shielded. One radiation detector is part of the treatment unit and indicates on the control panel when it detects radiation. An independent unit, usually mounted on the treatment room wall with displays both inside and outside the room also alerts the operator and other personnel when the radioactive source is out of the safe. A TREATMENT ON Indicator outside the room, activated when the source passes the reference optical pair discussed above and shown in Fig. 4, also indicates that a treatment is in progress.

Table 2. Exposure Rates from an Exposed 10 Ci ^{192}Ir Source

Typical Situation	Distance, m	Dose Equiv Rates, Sv · h^{-1}	Time, min, to Receive	
			10 Gy (likely injury)	0.5 Sv (annual body limit)
In Patient	0.01	460	1.25 min	0.07 min
Handling with Kelly Clamps, to hands	0.1	4.6	2.1 h	6.5 min 98 min for hand limit
Handling with Kelly Clamps, to body	0.3	0.5	18.8 h	1.0 h
Standing near	1	0.046	8.7 days	11 h
Standing far	2	0.012	34.8 days	43 h

Emergency Service Instruments

In the event the radioactive source fails to retract after termination, interruption, pushing the EMERGENCY SWITCH, or cranking the stepper motor manually, the immediate priority is to remove the source from the patient. Table 2 gives the exposure rates at various distances from a 10 Ci ^{192}Ir source. Table 2 shows that the dose to the patient, with the source in contact, can cause injury in a very short time. On the other hand, the operator, working at a greater distance, is unlikely to receive a dose exceeding regulatory limits for a year, let alone one that would cause health problems. Once the source is removed from the patient and moved to a distance of even a meter, the exposure rate is quite low, and whatever actions need be taken to remove the patient from the room can be performed safely.

The effective annual limit to the body should actually be 10 times < the 0.5 Sv in keeping with the principle to keep exposures as low as reasonable achievable (ALARA), and ideally should not be received in one, short exposure. The allowed exposure to the hands is 15 times that to the body.

The preferred approach to a source that will not retract by any of the methods is to remove the applicator from the patient as quickly as possible, and place the applicator containing the source in a shielded container (Fig. 16). If it is clear that the cable is caught in the transfer tube and not in the applicator itself, the applicator or catheter may be disconnected from the transfer tube and the source pulled from the applicator. In some cases, this will be faster than removing the applicator. The reason to avoid disconnecting the applicator from the transfer tube is that a source may stay in the applicator if the source capsule shatters. In that case, removing the applicator attached to the transfer tube keeps the system closed, while disconnecting the two opens a path for parts of a broken source to fall from the applicator into body cavities or crevices, or roll onto the floor.

A situation may arise when the source needs to be detached manually from the treatment unit. One (still unlikely) scenario would be if the source were stuck out of the treatment unit, the sources or the closed applicator had been removed from the patient, a person were pinned very close to the source so neither they, nor the treatment unit, could be moved, and the source on the cable could not reach the shielded container. In this special situation, the source cable should be cut from the unit and the source placed in the shielded container always present in the room. In cutting the source cable, it must be clear that the cut is *not* through the source capsule. For units with the capsule welded on the cable, the cut must be through the braided cable as opposed to the smooth steel capsule (Fig. 17). For sources imbedded in the cable, a sufficient length of the cable must be seen to assure the cut occurs behind the source. Thus, emergency tools that must be present in the treatment room and always readily accessible include a wire cutter, a pair of forceps, and a shielded service container.

Back-Up Battery

In case of a power failure during the treatment, the machine is equipped with a back-up battery to provide retraction of the source to its safe. The batteries should be tested with each source exchange.

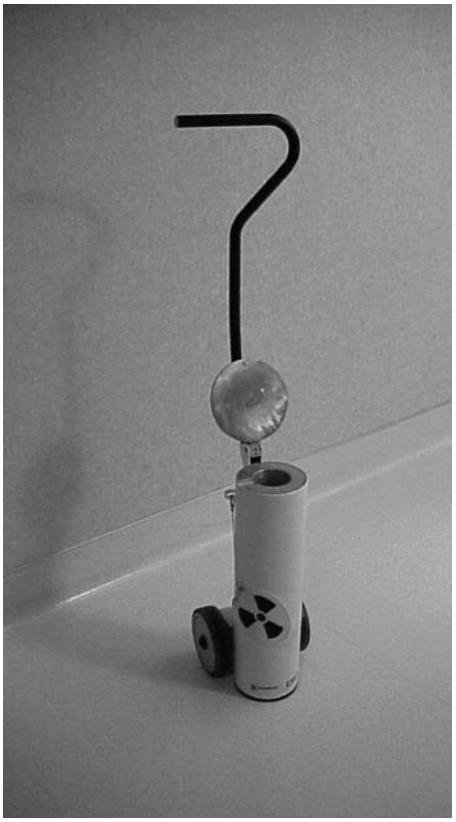


Figure 16. The shielded container for emergency placement of an unretracted source.

Figure 17. Cutting the source cable from a treatment unit. This procedure should only be performed in very special, rare situations as described in the text. Great care must be taken to assure the cut is through the cable and not the source capsule.

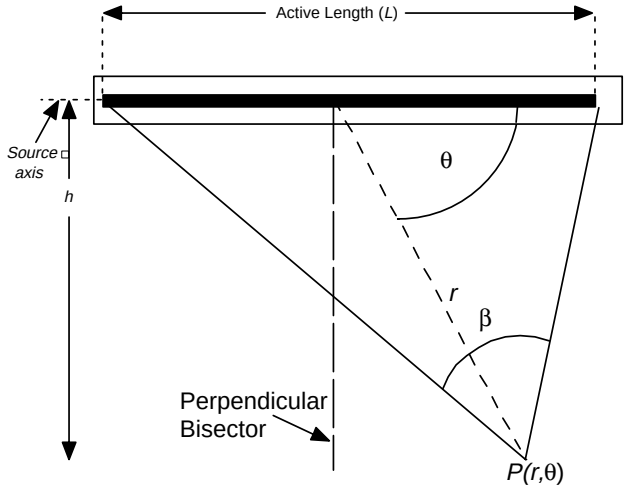


Figure 18. Geometry and legend for the dosimetry of a line source as given in Eq. 1.

TREATMENT PLANNING SYSTEM

Software and hardware for the treatment planning system are provided by the vendor selling the treatment unit. Three-dimensional (3D) patient data [computed tomography (CT), magnetic resonance imaging (MRI)] can be directly transported and loaded in the planning system. Two dimensional (2D) data (e.g., from radiographs) usually are loaded interactively by computer peripherals (scanners, digitizers) although some automated input systems are available. With 2D input, the target information must be inferred since tumors are generally not visible on the images, while the 3D imaging often visualizes tumors as well as surrounding normal tissue structures. With either input, tumor volume is entered on these images, and the treatment-planning volume is constructed by adding some margin to the tumor volume. Various computer algorithms help the planner conform the prescribed dose to the target volume. Data characteristic for the radioactive source are usually supplied by the vendor and included in the software. The medical physicist enters the source strength data both in the planning system and the treatment unit at the time of the installation of the new source in the treatment unit after calibration.

Dose Calculation

The treatment-planning computer calculates the dose distribution for a patient containing an applicator with a given set of dwell positions, each with their own dwell time. In calculating the dose distribution, the computer first calculates the dose to a set of grid points. Usually, the operator wishes to see the results presented as isodose lines. An isodose line on a given plan connects all points receiving the same dose, much like elevation lines on a contour map of part of the Earth connects points with the same altitude. From the dose values at the grid points, the computer interpolates to find the path of the isodose line value specified by the user. The calculation of the dose from one dwell position, identified with the subscript i, to a point $P(r,)$ as shown in Fig. 18, uses the formula (1),

$$D_i(r_i,\theta) = S_K \cdot \Lambda \cdot [G_i(r,\theta)/G(r_o,\theta_o)] \cdot g_i(r_i) \cdot F_i(r_i,\theta) \cdot t_i \quad (1)$$

where

$D_i(r_i,\theta)$ = The radiation dose to water, in units Gy, at position $P(r,\theta)$

S_K = The air kerma strength of the source in $\mu Gy m^2 \cdot h^{-1}$. The strength of photon-emitting brachytherapy sources usually is specified by the intensity of the radiation at some distant outside of the source rather than by the amount of radioactivity contained inside. In this manner, variations in the source encapsulation, which may attenuate varying amounts of the radiation given off by the contained radionuclide, have no effect on the dose delivered to the patient. While the strength of a new HDR source is often quoted as 10 Ci (the approximate activity in the capsule), the actual source strength determination accounts for the energy of radiation from the source, transferred to a mass of air at a given point. Air kerma is the energy transferred from the radiation to kinetic energy in the medium per unit mass, where the medium must be specified. That point for air kerma strength is at 1 m, and a new source would have an air kerma strength of ~ 40 mGy $\cdot$ m$^2 \cdot$ h^{-1}, or in shorthand, 40 kU, where 1 U $= \mu$Gy $\cdot$ m$^2 \cdot$ h^{-1}. A radiation dose of 1 gray (Gy), equals 1 joule per kilogram (1 J $\cdot$ kg).

Λ = The dose rate constant, that is the absorbed dose rate in cGy $\cdot$ h^{-1} at 1 cm from the source in the perpendicular plane that bisects the source axis per unit air kerma strength. For ^{192}Ir sources, $\Lambda = 1.12$ cGy $\cdot$ h^{-1}/U.

$[G_i(r_i,\theta_i)/G(r_o,\theta_o)]$ = The geometry function, which accounts for changes in dose rate due to the relative positions of the source and the calculation point and the

shape of the source. The numerator expresses the geometric dose pattern for the point of calculation while the denominator gives that for the reference condition, where $r_o = 1$ cm and $\theta_o = 90^\circ$. The geometric dose pattern usually is approximated as $1/r^2$ for a point source, and for a line source /(L·h) as shown in Fig. 18.

$g_i(r_i)$ = The radial dose function, variation in the dose rate with distance from the source due to the attenuation and scatter due to the tissue between the source and the point of calculation at distance r, normalized at 1 cm, and not including any effect in dose rate due to geometry (i.e., the geometric function has been removed from the dose at the calculation distance and at 1 cm for the ratio).

$F_i(r_i,\theta)$ = The anisotropy function, which describes the deviation of the shaped of the isodose lines from a circle. The function $F_i(r_i,\theta)$ = the dose at the calculation point $P(r,\theta)$ divided by the dose at the same distance, r_i but on the perpendicular bisector, that is with $\theta_o = 90^\circ$, and, as with the radial dose function, with the geometrical effects removed.

t_i = the dwell time for dwell position i.

Dose Optimization

The treatment planning addresses first which dwell positions will be used. Then for each of the dwell positions, the dwell time must be calculated. The goal is to match the resulting dose distribution with the target volume, a process referred to as optimization. There are several methods to assist the operator in optimizing the dwell times. The methods fall into three main categories:

Analytic methods use relatively simple algorithms to calculate the dwell time for each dwell position. One of the most common, geometric optimization (2,3), weights the dwell time for a position inversely to the sum of the doses to that position from the other dwell positions. For an intracavitary application, where the dose distribution is intended to more or less conform to the shape of the applicator tracks, "distance optimization" is used, where the contributions of all of the other dwell positions are included in the summation of the dose. For interstitial application, where the implant usually treats a volume, the process becomes "volume optimization", with the dose contributions from dwell position along the same track excluded in the summation.

Dose specification methods attempts to calculate the dwell times to deliver a specified dose to designated points (dose points or optimization points) placed throughout the volume or on the surface of the target (4–6). The dose to each specified point described an equation with the dose to the point on one side and the contributions from each of the dwell positions on the other,

$$\text{Dose} = S_K \cdot \Lambda \cdot \sum_{i}^{\text{all dwells}} \{[G_i(r_i,\theta)/G(r_o,\theta_o)] \cdot g_i(r_i) \cdot f_i(r_i,\theta) \cdot t_i\} \quad (2)$$

In the simplest situation, such an approach becomes solving a set of simultaneous equations for the doses to the optimization points for the unknown dwell times. The problem comes when there are more equations than unknowns (more dose points than dwell positions and the set is over determined) or the converse (when the set is underdetermined). In the general case, the set is solved by a least-squares method to find the values for the dwell times that minimizes the square of the difference between the doses desired and the doses resulting from times in the set of equations. However, it is very helpful to add an additional criterion on the dwell times: controlling the fluctuation in dwell times between adjacent positions. The optimization equation becomes

$$X^2 = \sum_{j=1}^{\text{All dose points}} (D_j^{\text{precribed}} - D_j^{\text{calculated}})^2 + V \sum_{i=1}^{\text{dwells}-1} (t_{i+1} - t_i)^2 + u \sum_{i=1}^{\text{all dwells}} t_i \quad (3)$$

where the first term considers the difference between the prescribed dose and that calculated dose for each specified point. The value of the second term depends on the differences in dwell times between each dwell position and its neighbor. The factor, the dwell weight gradient factor, v, determines how important minimizing this fluctuation is. Large values for v (> 1) tend to force the dwell times to be the same, not producing a very conformal dose distribution; small values (< 0.2) permit negative times to result from the optimization (not a physical situation). The last term minimizes the overall exposure time, assuming that the set of dwell times that adequately treats the target volume with the least total time results in the lowest dose to the rest of the patient. The factor u determines how important minimizing dwell time is for the optimization. The optimized set of dwell times minimizes X^2.

Stochastic methods use iterative techniques to find adequate values for the dwell times. Generally, these approaches establish an objective function, such as the difference in dose to a set of point between that prescribed and that achieved. The function may also include penalties for excessive doses to normal tissue structures or lack of homogeneity through an implanted volume. The goal of the optimization is to obtain the best score for the objective function. The process begins with a set of values for the dwell times, evaluates the objective function, and then makes changes in the dwell times. If the new set of times improves the value of the objective function, the new set becomes the current best set. If the old set of times gave a better value for the objective function, it remains the current best set. Obviously, the strategy for how to pick each new set of values in the core of the methodology, and a more complete discussion is beyond the scope of this text. The most common approaches in the literature are simulated annealing (7) and the genetic algorithm (8).

The goal of all the optimization methods is to adequately treat the target tissues while sparing the sensitive normal structures. A resultant plan is shown in Fig. 19.

SHIELDING

The radioactive source in the high dose rate machine starts ~ 10 Ci with an exposure rate at a distance of 1 m from the

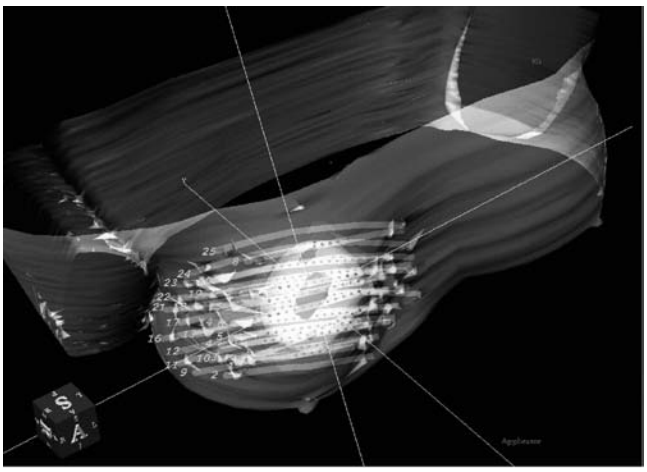


Figure 19. A 3D view of dose distribution of a breast catheter implant with 25 catheters. The inner volume is the tumor volume (lumpectomy cavity), while the outer volume (planning treatment volume) has been generated by adding some margins to the tumor volume. The gray cloud is the dose distribution generated by the treatment planning system.

source of ~ 46 mSv $\cdot$ h^{-1}. According to the rules and regulations of the United States Nuclear Regulatory Commission (USNRC), the annual limit for radiation exposure to the public is 1 mSv and the annual occupational limit is 5 mSv. (The actual limit for occupationally exposure persons is 50 mSv $\cdot$ year^{-1}, but following the principle of maintaining exposures as low as reasonable achievable, the NRC usually holds licensees to exposure 0.1 of the limit.) In addition to the annual limit, NRC requires that in an unrestricted area the dose equivalent rate should not be > 0.02 mSv in any given hour. Thus, the HDR machine needs to be housed in an adequately shielded room. To meet these requirements in a HDR brachytherapy suite, where the walls and the ceiling are at least 1.5 m from the machine head, concrete wall of ~ 43–50 cm (or 4–5 cm of lead) are needed. For larger rooms the concrete wall thickness will be lower since the exposure rate is inversely proportional to the square of the distance from the radioactive source. For details on the procedures for calculating the thickness of barriers for a particular facility, see health physics texts such as Cember (9) or McGinley (10).

QUALITY ASSURANCE

In order to maintain the quality of patient care, a quality management program is required in every facility that provides HDR brachytherapy treatment. Such a program generally follows standards set by professional organizations and intends to minimize untoward events caused by the malfunction of the machine or human error. Such programs become exceedingly important in HDR brachytherapy because the planning and the treatments tend to happen very quickly, increasing the likelihood of accidents and mistakes. Quality Assurance (QA) tests measure some performance aspect of the treatment unit and compare the results with expectations in order to demonstrate proper operation. QA is performed at various intervals: some for each patient, some once each treatment day, and others with each source change. Moreover, for HDR RAL, the USNRC mandates that users meet certain standards, including education and training on operating the machine, emergency procedures, radiation monitoring, pre-treatment safety checks, safe and accurate delivery of the treatment, and monthly/initial calibration of the source. Since the details of quality assurance is outside the scope of this literature, interested readers can refer to the report of Task Groups 59 (11) and 56 (12) of the American Association of Physicists in Medicine and relevant texts (13).

In general, the problem of quality assurance becomes assuring that the treatment will deliver the correct dose, to the correct location, safely. Thus, the tests generally follow the outline below:

Verification of dose variables.

- Checking the strength of the source compared with that projected from the initial calibration based on radioactive decay.
- Checking the proper operation of the controlling timer.

Verification of position control.

- Checking that the source goes to the location programmed.
- Checking coincidence between the programmed positions and the respective positions indicated by imaging markers.
- Checking consistent movement of the source.

Verification of proper operation of safety features.

- Checking operation of the door interlocks.
- Checking the operation of a handheld radiation detector.
- Checking the operation of the on-board and on-wall radiation detectors.
- Checking the operation of the check cable runs and interlocks.
- Checking the operation of the EMERGENCY OFF and TREATMENTINTERRUPT buttons.

COSTS

Currently, two vendors (Nucletron Corporation and Varian Medical Systems) market their RAL treatment unit in the United States. Both the devices requires a capital expenditure of $\sim$\$500,000–750,000, which includes the treatment unit, a variety of transfer tubes, along with the software and hardware for the treatment planning system. Applicators that are needed to be placed in the tumor costs extra. The costs of preparing a shielded room along with ancillary equipment for X-ray imaging and operating room

procedures can be another $500,000–750,000. Hence, the total cost can run in between $1–1.5 million (14).

ADVANTAGES AND DISADVANTAGES

HDR comparing with LDR brachytherapy offers several advantages and disadvantages. Being aware of these permits safe and effective operation and application of HDR brachytherapy.

Advantages of HDR Brachytherapy

Safety. One of the major advantages of a RAL is the reduction or elimination of radiation exposure to the radiotherapy staff. In conventional LDR, manual afterloading, the radiotherapy staff receives radiation exposure while loading the applicators with the radioactive sources, and the nursing personnel are exposed during patient care through the duration of the treatment (1–4 days). With either HDR or LDR remote afterloading, the radiotherapy personnel are outside the shielded room during the treatment, and hence are exposed to minimal radiation.

Optimization. The design of the HDR RAL with the stepping source allows greater flexibility and control over dose distribution. The stepping source allows optimization of the dose distribution by adjustment of the dwell times for each dwell position in each channel. The dwell times can be varied infinitely, permitting very fine control of the dose distribution. In LDR, either manual or using an RAL, the finite number of activities available (usually four at most) and the larger sources used with manual applications impose a restriction on the ability to conform the dose distribution to the target.

Stability. Because HDR intracavitary treatments take so little time ($\sim$ 1 h), applicators can be fixed in place much more stably than for the several day treatments using LDR brachytherapy.

Dose Reduction to Normal Tissue. As with stability, the short duration of HDR intracavitary treatments allows displacement of normal tissue structure (i.e., pushing them away from the source paths) to a greater extent than with LDR treatment.

Applicator Size. The small size of the HDR source permits the use of smaller applicators than those required for the LDR applications, increasing the comfort to the patient.

Outpatient Treatment. Almost all HDR patients are treated on an outpatient basis compared to LDR patients who usually are treated as inpatients. Outpatient treatment is more convenient for the patients and generally results in lower overall costs.

Disadvantages of HDR Brachytherapy

Investment. The initial expense of HDR RAL is very high. Machines and site preparation costs can be anywhere between $0.5 and 1 M.

Complexity. The technological complexity of HDR RAL opens the increased probability of errors, and leads to increased regulatory scrutiny.

Compressed Time Frame. As mentioned above, the rapidity with which procedures progress in HDR brachytherapy increases the probability of executing errors.

Radiobiology. As the dose rate increases, the radiosensitivity (damage per unit dose) increases for both normal tissues and tumors. Unfortunately, the radiosensitivity for the normal tissue increases faster than that for tumors, increasing the likelihood of injuring the patient while controlling the tumor. Overcoming this radiobiological handicap requires the use of the advantages of *optimization*, *stability*, and *dose reduction to normal tissues*, in addition to fractionization. As with external-beam radiotherapy delivered using a linear accelerator, which also operates in the HDR regime, spreading the treatments over many smaller fraction delivered over several days reduces the difference in radiosensitivities between the tumor and the normal tissues.

BIBLIOGRAPHY

Cited References

1. Nath R, Anderson LL, Luxton G, Weaver KA, Williamson JF, Meigooni AS. Dosimetry of interstitial brachytherapy sources: Recommendations of the AAPM Radiation Therapy Committee Task Group No. 43. Med Phys 1995;22:209–234.
2. Edmundson GK. Geometry based optimization for stepping source implants. Activity—The Selectron User's Newsletter 1991;5:22.
3. Edmundson GK. Geometric optimisation: an American view. In: Mould RF, editor. International brachytherapy. Veenendaal, The Netherlands: Nucletron International BV; 1992, p 256.
4. van der Laarse R. Optimization of high dose rate brachytherapy. Activity—The Selectron User's Newsletter 1989;2:14–15.
5. van der Laarse R, Edmundson GK, Luthmann RW, Prins TPE. Optimization of HDR brachytherapy dose distributions. Activity—The Selectron User's Newsletter 1991;5:94–101.
6. van der Laarse, Thomadsen BR, Houdek PV, van der Laarse R, Edmunson G, Kolkman-Deurloo I-KK. Treatment planning and optimization. In: Nag S, editor. High dose rate brachytherapy: a textbook. Armonk, (NY): Futura Publishing Co.,; 1994. p 85–91.
7. Sloboda RS. Optimization of brachytherapy dose distributions by simulated annealing. Med Phys 1992;19:955–964. See also, Sloboda RS, Pearcey RG, Gillan SJ. Optimized low dose rate pellet configuration for intravaginal brachytherapy. Int J Radiation Oncol Biol Phys 1993;26:499–511.
8. Davis L. Handbook of Genetic Algorithms. New York: Van Nostrand Reinhold; 1991.
9. Cember H. Introduction to health Physics. 3rd ed. New York: McGraw-Hill; 1996.
10. McGinley P. Shielding Techniques for Radiation Oncology Facilities 2nd. Madison: Medical Physics Publishing; 2002.
11. Kubo HD, Glasgow GP, Pethel TD, et al. High dose-rate brachytherapy treatment delivery: AAPM Radiation Therapy Committee Task Group No. 59. Med Phys 1998;25:375–403.

12. Nath R, Anderson LL, Meli JA, et al.Code of practice for brachytherapy physics: AAPM Radiation Therapy Committee Task Group No. 56. Med Phys 1997;24:1557–1598.
13. Thomadsen BR. Achieving Quality in Brachytherapy. Bristol: Institute of Physics Press; 1999.
14. Rivard MJ, Kirk BL, Stapleford LJ, Wazer DE. A comparison of the expected costs of high dose rate brachytherapy using ^{252}Cf versus ^{192}Ir. Appl Radiat Isot 2004;61:1211–1216.

See also BRACHYTHERAPY, INTRAVASCULAR; HYPERTHERMIA, INTERSTITIAL; PROSTATE SEED IMPLANTS; RADIATION DOSIMETRY FOR ONCOLOGY.

BRACHYTHERAPY, INTRAVASCULAR

FIRAS MOURTADA
MD Anderson Cancer Center
Houston, Texas

INTRODUCTION

Intravascular brachytherapy (IVB) is a novel treatment modality that delivers ionizing radiation to a coronary artery to prevent renarrowing, that is, restenosis caused by stent placement within the artery. The term *brachy* is the Greek word for near since the radioactive source is placed inside or near the target cells. In general, the field of brachytherapy has been practiced for decades in the field of radiation oncology for treatment of intracavitary (vagina, bronchus, esophagus, rectum, nasopharynx, etc.) and interstitial (muscle sarcoma, prostate, breast, etc.) cancers. As a subspecialty, IVB is relatively new where most of its development took place in the 1990s. Ionizing radiation describes both electromagnetic (γ rays, X rays) and particulate (neutrons, beta, and alpha particles) of sufficient energy to remove electrons from the target atom (thus ionizing). Unlike conventional brachytherapy where the target is mostly centimeters away from the source, IVB targets the adventitia of the vessel wall, located within 1–5 mm from the radioactive source. To obtain accurate dosimetry data in such close range is a challenge. The scope of this article is on the delivery devices for IVB and tools needed to assess the dosimetric properties of such brachytherapy devices. (Dosimetry is a subspecialty of radiation physics that deals with the measurement of the absorbed dose or dose rate resulting from the interaction of ionizing radiation with matter.) Such techniques are useful and can be applied in other future applications that require delivery of ionizing radiation to a target within a few millimeters from a radioactive source.

Mechanisms of Restenosis

A diseased coronary vessel is mainly caused by atherosclerotic plaque formation containing mostly cholesterol and lipids. This condition can lead to a heart attack and chest pain (angina) where the blood flood within the lumen is compromised. Coronary artery bypass surgery (CABG) is the traditional method to alleviate this condition. In the last few decades, minimally invasive procedures have been developed in a field known as *Interventional Cardiology*. Percutaneous transluminal coronary angioplasty (PTCA), first performed by Gruentzig in 1977 (1); and endovascular prosthetic devices (stents), first performed by Dotter et al. (2) and Cragg et al. (3) in 1983, are the most common devices used in interventional cardiology today. Charles Thomas Stent (1807–1885), an English dentist who lent his name to a tooth mould. Charles Dotter used the word "stent" in 1963 to name endoluminal scaffolding devices. However, these interventions have created a new problem, restenosis.

Restenosis is a wound healing process occurring directly at the angioplasty balloon or stent site. It is believed that three processes cause restenosis: elastic recoil, neointimal hyperplasia, and negative vascular remodeling. Elastic recoil, or vessel spasm, occurs in the healthy (plaque-free) portion of the vessel within minutes after balloon expansion (angioplasty balloon or stent-expanding balloon). Elastic recoil causes a luminal cross-sectional area reduction of $\sim 50\%$, but only for a short time after the procedure. The second component of restenosis is neointimal hyperplasia resulting in new tissue growth occupying the microcracks and rapture within the plaque mass, in some patients this process can be overcompensating, filling more tissue within the vessel lumen thus compromising blood flow. The blood vessel wall (any vessel larger than capillaries) has three major layers called tunica, the innermost layer is the intima followed by a middle concentric layer called the media where mostly the smooth muscle cells reside, and then the outer most layer called the adventitia. The adventitia contains a connective tissue with mainly collagen and myofibroblasts. Some controversy still remains as to which cells are responsible for neointimal hyperplasia, media smooth muscle cells, or myofibroblasts migrating from the adventitia. In IVB, the prescription dose should reach the tunica adventitia to insure full therapeutic benefit. The third component of restenosis is negative remodeling. The term negative remodeling refers to contraction of the arterial wall following an arterial injury inflicted by an interventional procedure occurring slowly over the first 3–9 months after the angioplasty. Negative remodeling is believed to play a major factor in restenosis after a PTCA intervention. However, the stent is a mechanical scaffolding device that prevents negative remodeling. Hence, in-stent restenosis appears to derive almost exclusively from neointimal hyperplasia, even more than seen in balloon angioplasty. Schwartz and Holmes (4) provide a detailed discussion on restenosis and remodeling. Hall et al. (5) present on the radiobiological response of vascular tissue to IVB. (5).

Epidemiology and Clinical Trials

Restenosis is a very likely event (within few months of the initial intervention) and has been a frustrating problem in interventional cardiology. For example, > 1 million people worldwide had percutaneous coronary interventions in 2001 and of these $> 85\%$ received a stent. Restenosis occurred in $> 50\%$ of the stented patients ($\sim 425{,}000$ patients worldwide), with the United States share of $\sim 150{,}000$ patients.

Restenosis as measured using quantitative coronary angiography (QCA) is arbitrarily defined as a narrowing

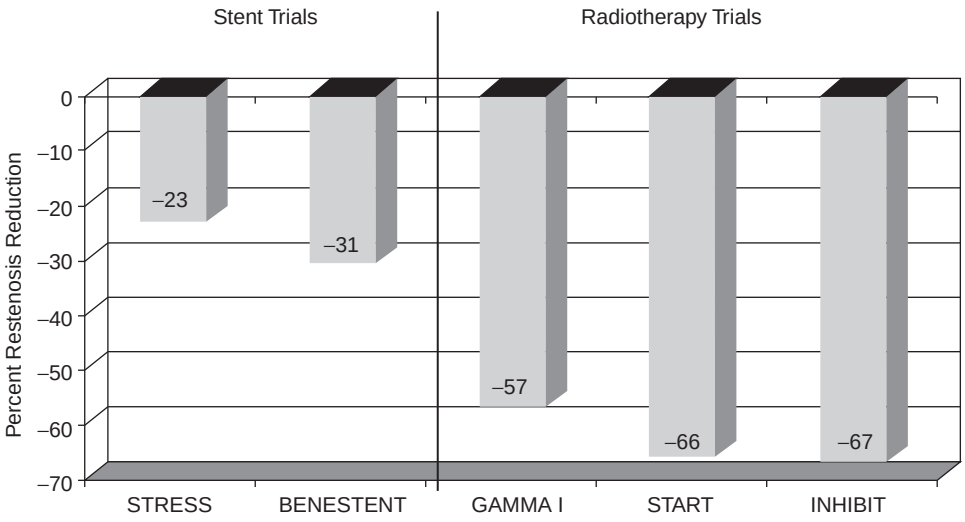


Figure 1. Percent restenosis reduction reported in bare-metal stent (STRESS and BENESTENT), and in intravascular brachytherapy clinical trials Gamma I (^{192}Ir), START (^{90}Sr/^{90}Y), and INHIBIT (^{32}P).

of the vessel lumen of at least 50% relative to the adjacent healthy vessel lumen ratio of minimum lumen diameter (MLD) to the reference lumen diameter (RLD). Mehran et al. (6) studied several risk factors from angiographic patterns of in-stent restenosis (ISR). The MLD, lesion length, and diabetes were found to be important factors predicting in-stent restenosis risk. Risk increased as a function of lesion length (10–40 mm) and decreased as a function of the MLD (2.5–4 mm diameter). Further, the ISR pattern is important where diffused restenosis has a larger risk those with a focal pattern.

The main complications of PTCA are acute vessel occlusion and late restenosis. Two important trials started in 1991, the North American STRESS (STent REStenosis Study) (7) and the European BENESTENT (BElgium Netherlands STENT trial) (8) using metal-bare stents (Palmaz-Schatz stent) transformed the practice of interventional cardiology. (Bare-metal stent is a term used here to make a distinction from the drug-coated stent briefly discussed in this article.) However, in-stent restenosis incidence of $>50\%$ was still observed. The introduction of IVB in the 1990s revamped the hope to eradicate restenosis. Condado et al. (9) in 1997 conducted the first small human trial showing reduction in restenosis over a 5 year follow-up period; this study, however, had some dosimetric issues. Definite multicenter double-blinded randomized clinical trials for IVB for the indication of in-stent restenosis are the GAMMA I (10), the START (Strontium-90 Treatment of Angiographic Restenosis Trial) (11), and the INHIBIT (INtimal Hyperplasia Inhibition with Beta Instent restenosis Trial) (12). With over 5000 patients treated with IVB, this modality has recently proven its safety and efficacy. As shown in Fig. 1, IVB (GAMMA I, START, INHIBIT) clinical trials had about twofold reduction of in-stent restenosis than found from the STRESS and BENESTENT bare-metal stent versus. PTCA trials. Initial problems in the IVB trials, like edge failure due to geographic miss and increased incidence of later thrombosis, were quickly remedied by increasing the radioactive source length and prolonged use of antiplatlet therapy. Table 1 is a detailed summary of these IVB clinical trials. Many other clinical IVB clinical trials were conducted using various isotopes, delivery system, and other clinical indications.

THEORY AND DETAILED DESCRIPTION OF IVB DEVICES

Based on the clinical trials discussed above, the Food and Drug Administration (FDA) granted a premarket approval (PMA) to three IVB devices. The Checkmate system (Cordis Corporation, a Johnson & Johnson Company, Miami Lakes, FL) using ^{192}Ir and BetaCath system using ^{90}Sr (Novoste Corp. Norcross, GA) were both approved on November 3, 2000. About 1 year later, the GALILEO Intravascular Radiotherapy System (Guidant Corp., Santa Clara, CA) using ^{32}P was also approved by the FDA. All of these devices are classified as catheter-based radiation delivery systems using sealed radioactive sources. Catheter-based means the source in the form of a source wire or a train of seeds (ribbon) is placed inside a closed-tip lumen catheter. The catheter is first placed into the target vessel and the source wire or ribbon is delivered or afterloaded using manual or computer-based delivery systems. The radiation safety considerations for these devices have been greatly discussed in the literature (13,14).

Other irradiation techniques were also investigated, including inflation of dilatation balloon catheter with radioactive liquid or gas; insertion of miniature X-ray tubes; implantation of radioactive stents; and postangioplasty external beam irradiation. These techniques did not make it to the market due to variable reasons including suboptimal efficacy, safety, or practicality. Table 2 summarizes several other isotopes and delivery systems investigated for IVB applications. Tables 3 and 4 list a few radiation characteristics for important gamma and beta sources for IVB.

Cordis Checkmate System

Checkmate is indicated by the FDA for the delivery of therapeutic doses of gamma radiation for the purpose of reducing in-stent restenosis. The system is for use in the treatment of native coronary arteries (2.75–4.0 mm in

Table 1. Summary of Pivotal IVB Clinical Trial Used to Obtain FDA Approval for In-Stent Restenosis in Native Arteries Indication

					Angiographic Restenosis		TLR	
Trial	Target Lesion	Source	Dose, Gy	Patients, *n*	Rad., %	Placebo %	Rad. %	Placebo, %
Gamma-1	In-stent (< 45 mm)	^{192}Ir ribbon	8–30	252	22	50 (6 months)	24	42 (9 months)
START	In-stent (< 20 mm)	^{90}Sr/^{90}Y Seed train	16–20 Gy at 2 mm	476	14	41 (8 months)	16	24 (8 months)
INHIBIT	In-stent (< 45 mm)	^{32}P wire	20 Gy at 1 mm into vessel wall	332	16	48 (9 months)	11	29 (9 months)

diameter and lesions up to and including 45 mm in length) with in-stent restenosis following percutaneous revascularization using current interventional techniques. Outside of the FDA approved indication, Waksman et al. (15) also examined the effects of intravascular gamma radiation in patients with in-stent restenosis of saphenous-vein bypass grafts and found favorable results.

Radioactive Source Ribbon. The Checkmate catheter-based brachytherapy system uses ^{192}Ir seeds that are preassembled in 6, 10, and 14 seed strand inside nylon ribbons (Best Medical International, Springfield, VA). Treatment lengths are 23, 39, and 55 mm for the 6-, 10-, and 14-seed ribbons, respectively (16). Iridium-192 has an average energy of 370 keV and a half-life of 73.83 days. The ^{192}Ir radioactive metal (30% Ir, 70% Pt) is 3 mm long and 0.1 mm in diameter encapsulated within a 3 mm long × 0.5 mm diameter stainless steel capsule. The seeds are placed inside a nylon ribbon with an interseed spacing of 1 mm. The overall ribbon length is 230 cm and the outer diameter is 0.76 mm (2.4 F). [1 French (F) = 1/π mm.] At both distal and proximal edges of the seed strand, radiopaque markers are placed for visualization under X rays. A nonradioactive dummy ribbon is preloaded inside the delivery catheter to provide reinforcement during shipping and to improve maneuverability during initial positioning of the catheter across the target lesion. The dummy ribbon has the same length and configuration of the radioactive source ribbon to aid in IVB therapy planning during the procedure. A source lumen plug is used to prevent the movement of the dummy ribbon inside the Checkmate delivery catheter during initial catheter placement via the femoral artery. Dosimetry characterization of the Checkmate source ribbon are discussed in the literature (17).

Table 2. Gamma and Beta Sources with Different Delivery Systems Investigated for Intravascular Brachytherapy Applications

Gamma Delivery Systems	Beta Delivery Systems
^{192}Irseed train	^{32}P - wire, stent, balloon
^{125}I-stent	^{90}Sr/^{90}Y - seed train
^{103}Pd-stent, wire	^{90}Y - wire
^{131}Cs-stent	^{188}W/188Re - wire, balloon-liquid
^{99m}Tc-liposome–liquid	^{186}Re - balloon-liquid
	^{133}Xe - balloon-gas
	^{48}V - stent (positron)[a]
	^{62}Cu - balloon-liquid
	^{106}Ru/^{106}Rh - wire
	^{144}Ce/^{144}Pr - wire
	^{68}Ge/^{68}Ga balloon-liquid (positron)

[a] A positron is an electron with a positive charge.

Table 3. Average Energy and Half-Life of Important Gamma Emitters Investigated for Intravascular Brachytherapy Applications

Isotope	Ave Energy, keV	$T_{1/2}$, day
^{192}Ir	370	73.83
^{125}I	28	59.4
^{103}Pd	21	16.97
^{131}Cs	30	9.69

Delivery Catheter. This is a single lumen catheter with a distal rapid exchange tip and a closed-ended source lumen for isolation from patient blood contact. Both the radioactive source and the nonradioactive dummy ribbon use this lumen to reach the target. A single radiopaque marker at the distal end of the source lumen is to aid in catheter placement under fluoroscopy. A guidewire is used to guide the catheter along the tortuous pathway into the coronary artery. The guidewire exits the catheter 4 mm from the distal tip of the catheter. The overall length of the Checkmate catheter is 230 cm with a usable length of 145 cm. At the distal portion of the catheter, the outer diameter is 3.7 F (0.049 in.), which is deliverable with 7 F (mm) or larger guiding catheter.

Table 4. Maximum Energy and Half-Life of Important Beta Emitters Investigated for Intravascular Brachytherapy Applications

Isotope	Max Energy, keV	$T_{1/2}$
^{32}P	1710	14.3 day
^{90}Sr/^{90}Y	2280	29.1 year
^{90}Y	2280	64 h
^{188}W/^{188}Re	2120	69.4 day
^{186}Re	1090	90.6 h
^{133}Xe	360	5.3 day
^{48}V	696	16 day
^{62}Cu	2930	9.74 min
^{106}Ru/^{106}Rh	3540	371.6 day
^{144}Ce/^{144}Pr	3000	284.9 day

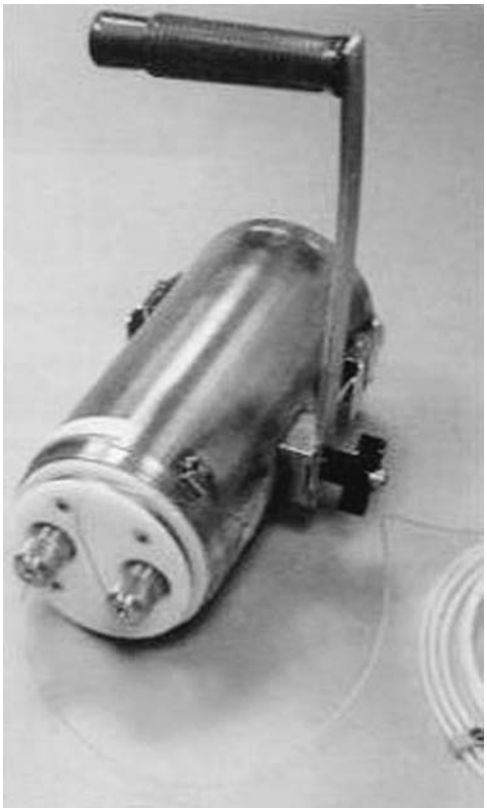


Figure 2. Checkmate delivery device (Cordis Corporation, a Johnson & Johnson Company, Miami Lakes, FL) is mainly a lead shielded cylinder housing both the radioactive ^{192}Ir source ribbon and dummy ribbon.

Delivery Device. As shown in Fig. 2, the Checkmate delivery device is mainly a lead shielded cylinder housing both the radioactive ^{192}Ir source ribbon and the dummy ribbon. This is a simple device where the proximal end of the source ribbon (nonradioactive part) protrudes from the proximal end of the delivery device and is coiled when not in use. When in use, a threaded cap on the distal end of the delivery device is replaced with a luer connector, which is connected to the hub of the delivery catheter. The source ribbon is pushed forward by hand from the proximal end of the delivery device into the delivery catheter.

Dosimetry. The Checkmate IVB system dosimetry was initially based on intravascular ultrasound (IVUS). From these images, the distance from the center of the IVUS catheter to the outer edge of the media tissue, called the external elastic membrane (EEM) is measured. A minimum of three axial images is taken along the stented vessel segment to determine the maximum and minimum distance from the source to the EEM. The dwell time is then calculated to insure that 8 Gy is delivered to the EEM farthest from the source, provided that no >30 Gy is delivered to the closest EEM. The IVUS-based dosimetry was later simplified to prescribe a fixed dose of 14 Gy at a distance of 2 mm from the centerline of the source. This provided a logistic solution to shorten procedure time and to spread the use of this system since many of catheterization labs in the United States do not have IVUS image modality.

Novoste BetaCath

The first generation BetaCath is a 5.0 F (1.59 mm) system. This system was indicated by the FDA to deliver beta radiation to the site of successful percutaneous coronary intervention for the treatment of in-stent restenosis in native coronary arteries with discrete lesions of $\leq$ 20 mm in length using the 30 or 40 mm system and for longer lesions up to 40 mm using a longer source train (60 mm) in a reference vessel diameter ranging from 2.7 to 4.0 mm. The second generation BetaCath is a 3.5 F (1.17 mm) system, which has an equivalent radioactivity to the 5 F (1.59 mm) system, but is smaller in diameter and fits easily inside a (1.91 mm) guide catheter. This system is intended to deliver beta radiation.

The 3.5 F (1.17 mm) BetaCath system has three main components, the ^{90}Sr source train, the β-Rail 3.5 F delivery catheter, and the 3.5 F delivery device.

Radioactive Source. Strontium-90/Yttrium-90 is a pure beta emitter with any energy spectrum with maximum energy of 2.27 MeV and an average of 0.934 MeV. The long half-life (29.1 years) simplifies treatment planning due to the almost unchanged dose rate of the device during the life cycle of the device (6 month). Each seed is 2.5 mm long and 0.38 mm in diameter for the 3.5 F system (0.64 mm seed diameter for the 5 °F system), manufactured by AEA Technology GmbH, Germany or BEBIG Isotopen- und Medizintechnik GmbH, Berlin, Germany. The radioactive source train consists of a wire jacketed "train" of 12 (30 mm source train), 16 (40 mm source train), or 24 (60 mm source train). The jacketed design was a major improvement over the initial design to eliminate seed movement thus providing uniform dose distribution. A radiopaque mark is placed one each side of each source train to provide visualization under fluoroscopy. Dosimetry characterization of the BetaCath sources are discussed in the literature (18,19).

Delivery Catheter. Only details of the second-generation delivery catheter, β-Rail 3.5 F (1.17 mm), will be discussed. This is a closed-end catheter with a total length of 180 cm (Fig. 3). A longer catheter called β-Rail 3.5 F XL delivery catheter has an overall length of 267 cm if desired. The catheter has a guidewire exit port at 1 cm from the distal tip, that is, a rapid exchange design. This catheter accommodates all source train lengths (30, 40, or 60 mm) that reach a most distal radiopaque marker located inside the delivery catheter. The β-Rail delivery catheter is preloaded with an Indicator of Source Train (IST), this Novoste terminology is used for the nonradioactive dummy source to aid in the measurement and positioning of the delivery catheter to insure adequate radiation coverage. The IST includes two radiopaque markings to delineate 30, 40, and 40 mm source lengths. At the proximal end of the delivery catheter, a proprietary connector is provided to insure a secure connection to the delivery device described next.

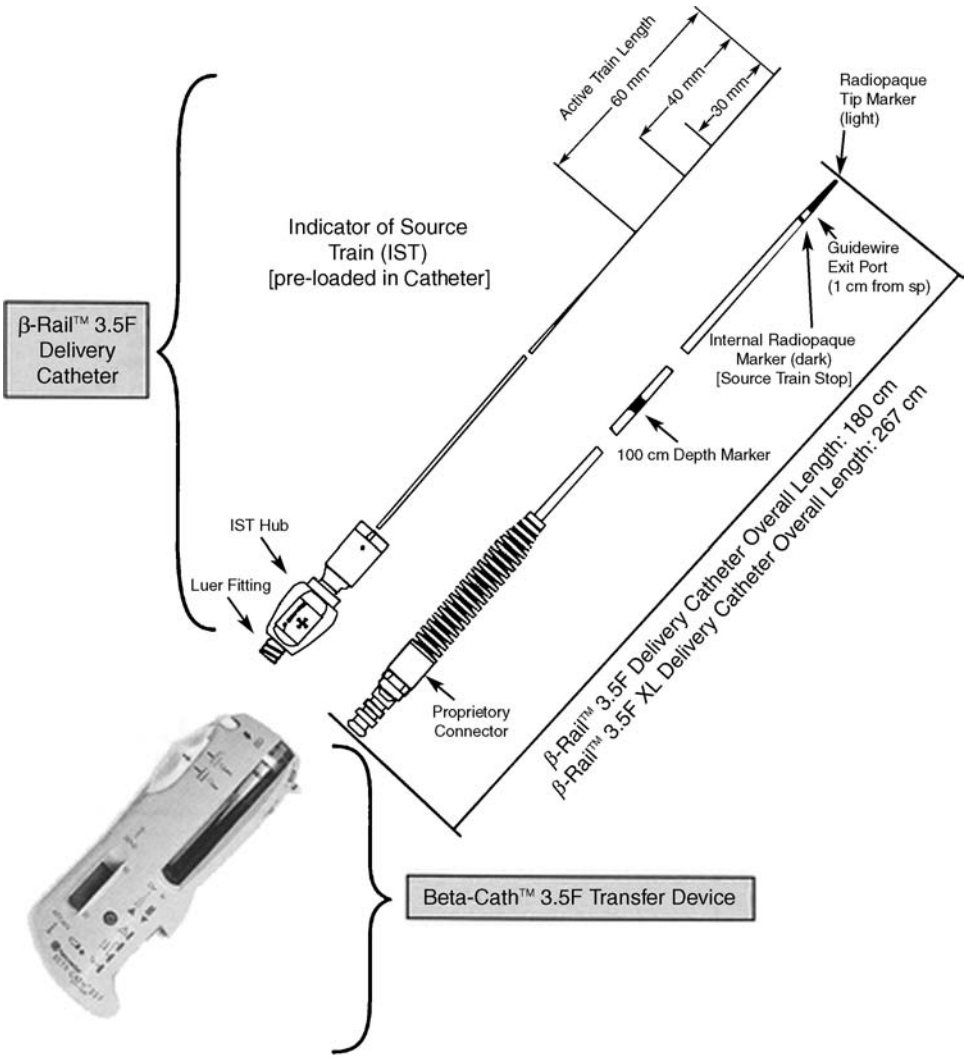


Figure 3. The β-Rail 3.5 F delivery catheter (Novoste, Norcross, GA) is a closed-end catheter with a total length of 180 cm. A longer catheter called β-Rail 3.5 F XL delivery catheter has an overall length of 267 cm is available. (Courtesy of Novoste Corporation.)

Delivery Device. This is a handheld battery-powered device used to store the radioactive source train (only one length). Hence, three separate delivery devices are needed to accommodate the 30, 40, and 60 mm source trains described above. The source train is sent into and returned from the delivery catheter using hydraulic pressure. To insure safe attachment of the delivery catheter, a connector lock latch is provided at the exit port of the delivery device. Several electronic pressure sensors are used to provide the operator with feedback on the pressure required using a saline-filled syringe to send, hold, or return the source train. Two source train position indicator lights (Green: In/Amber: Out), adjacent to the source chamber-viewing window (see Fig. 4). A fluid control lever controls the fluid flow and direction of the source train movement. A treatment counter tracks the number of procedures or test runs, a maximum of 125 transfers is allowed.

Dosimetry. The BetaCath IVB system dose prescription is the simplest out of the three systems discussed. The dose prescription is given relative to the source axis at 2 mm radial distance. The recommended dose is 18.4 Gy for a measured reference vessel diameter < 3.35 mm, but > 2.7 mm; and 23 Gy for a diameter > 3.35 mm, but < 4.0 mm. Vessel diameters < 2.75 mm or > 4.0 mm can be treated with this system; however, this is considered an off-label use of the device as defined by the FDA. The appropriate source train length (30, 40, or 60 mm) is selected after measuring the injured length using angiography and adding a margin on the distal and proximal side of a minimum of 5 mm.

Guidant Galileo

This IVB system is the only computer-based device. The GALILEO Intravascular Radiotherapy System (Guidant Corp., Santa Clara, CA) consists of three main components: the GALILEO ^{32}P Source Wire, the GALILEO Centering Catheter, and GALILEO Source Delivery Unit (SDU). The first generation product used a 27 mm long ^{32}P source and spiral centering catheter. The second generation, called GALILEO III uses a 20 mm long ^{32}P source, a trichannel centering catheter, and an automated high precision stepping algorithm.

The first generation GALILEO was indicated to deliver beta radiation to the site of successful percutaneous coronary intervention for the treatment of in-stent restenosis

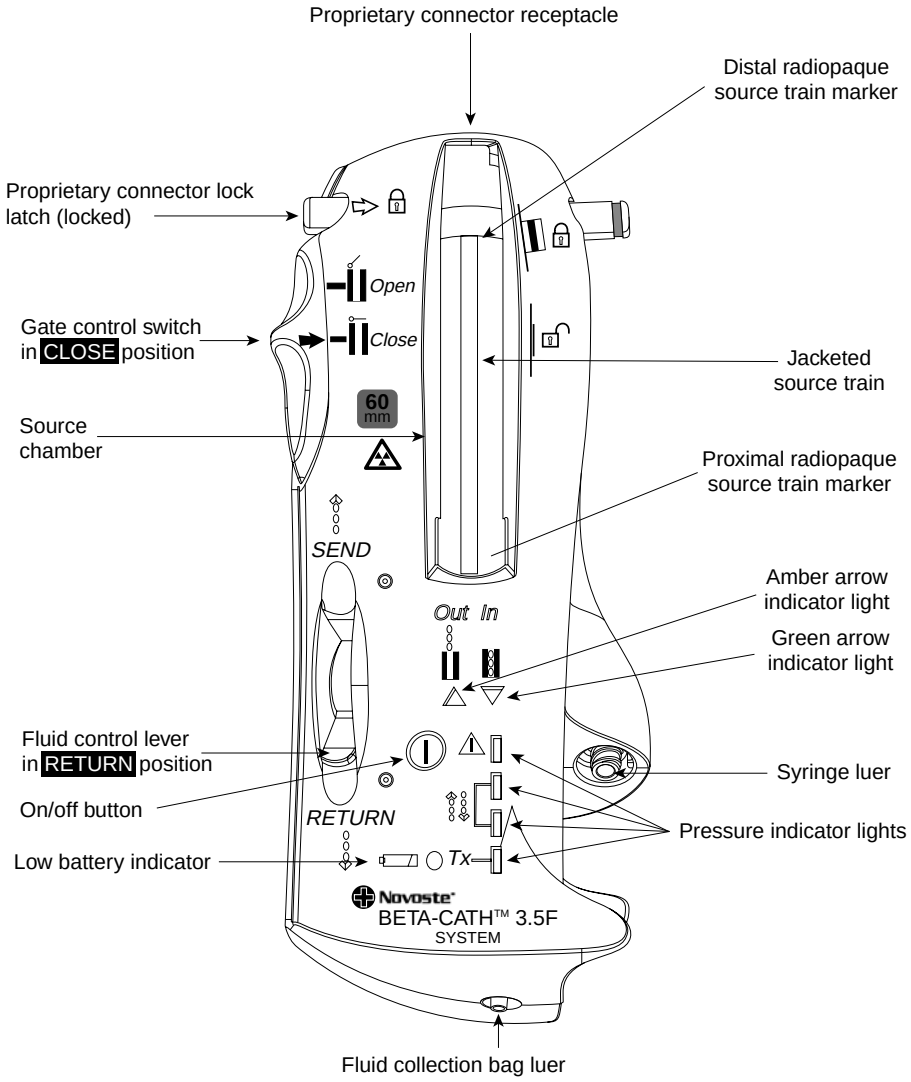


Figure 4. The BetaCath 3.5 F system (Novoste, Norcross, GA) is a handheld battery-powered device used to store the radioactive source train (only one length of 30, 40, or 60 mm source trains). Source train transfer depends on hydraulic pressure. Several electronic pressure sensors are used to provide the operator with feedback on the pressure required using a saline-filled syringe to send, hold, or return the source train. (Courtesy of Novoste Corporation.)

in native coronary arteries with discrete lesions ≤ 47 mm in reference vessel diameter 2.4–3.7 mm. The second generation GALILEO III system extended the indication to treat injured arterial length up to 52 mm.

Radioactive Source. The active wire contains linear solid-form phosphorus 32 (^{32}P) in ceramic glass fiber sealed in the distal end of a flexible nitinol (NiTi) hypotube, which is welded to a nitinol wire (total wire length is 2430 mm). The nominal active length is 27 mm (first generation system) 20 mm (second generation system). Both source wires have an outer diameter of 0.46 mm. Phosphorus-32 is a pure beta-emitting isotope with a maximum energy of 1.71 MeV, an average energy of 0.690 MeV, and a half-life of 14.28 days. The active wire can be used in multiple procedures for ~ 4 weeks (two half-lives). The active wire has two 1 mm tungsten X-ray markers, one proximal and one distal from the source, for visualization (see Fig. 5). A dummy source is used before the active wire is delivered to verify positioning, to check for kinks and catheter obstructions that could prevent the active wire from reaching the treatment site, and to achieve accurate positioning at the treatment site. Similar to the active wire, the dummy source also has two tungsten markers, making it visually identical to that of the active wire. Dosimetry characterization of the Galileo sources are discussed in the literature (20,21).

Delivery Catheter. This is a dual-lumen catheter with a spiral-shaped (first generation) or triloped balloon (second generation) to provide source centering within the lumen to improve dose homogeneity (see Fig. 6a and b). Such balloon profiles are designed to center the source within the lumen and to allow distal and side-branch perfusion during the dwell time that can take up to 10 min. This would make the procedure more tolerable for patients. The design of the second generation was found to provide better perfusion than the spiral design, in particular for the longer balloons to treat longer lesions. One lumen allows the automatic advancement of the source wire. At the proximal end of this lumen, a key connector attaches the delivery catheter to the delivery device. The key connector has a special code that reads using an optical sensor at the entry port to automatically determine the balloon length and the number of dwell positions based on the centering catheter used. The distal end of this lumen is closed to prevent contact of the source wire with the patient blood. The second lumen

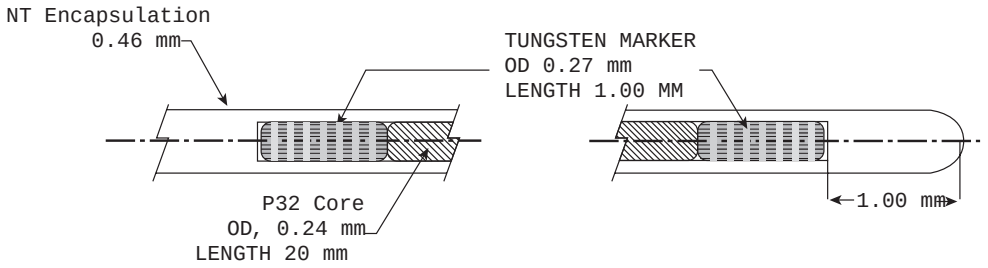


Figure 5. Cross-section of the ^{32}P source wire (Guidant Corporation, Santa Clara, CA) is shown. The radioactive core has a nominal length of 20 mm and 1 mm tungsten X-ray markers, one on each side of the core. All dimensions are in millimeters. (Courtesy of Guidant Corporation: GALIELO system is no longer manufactured or for sale.)

allows inflation (recommended pressure is 4 atm) and deflation of the centering balloon. This lumen terminates in a luer-lock connector allowing the attachment of standard inflation devices. A third lumen is 5 mm long at the most distal tip of the catheter; this is used to place the delivery catheter over a standard 0.014 in. (0.36 mm) coronary guide wire using a Rapid Exchange approach. Radiopaque markers located at the distal and proximal end of the balloon allows proper placement of the delivery catheter under fluoroscopy. Proximal shaft markers are located at 95 and 105 cm to aid in gauging catheter position relative to the tip of a brachial or femoral guiding catheter, respectively. The triloped GALILEO III centering catheter is provided with a balloon diameter of 2.5, 3.0, and 3.5 mm and balloon lengths of 32 and 52 mm. (Balloon length is defined as the distance between radiopaque balloon markers and does not include balloon tapers that extend beyond these markers). The MLD determines the appropriate centering catheter balloon diameter to use in the artery segment being treated (see Table 5). The lesion length determines the appropriate centering catheter length to use (see Table 6).

Delivery Device. The delivery device or the SDU has three main components: the head, base, and cartridge.

The front of the SDU head (Fig. 7) includes the touch screen monitor, status indicator lights, and housing for the cartridge. It also contains the manual retract wheel, the cartridge key port, the catheter key port and catheter eject button, and the red STOP button. On the back of the SDU head are the touch screen tilt lever, the swivel handle, and the system key port. The SDU head houses two motor drives—a primary motor and a battery-operated emergency retract motor. The emergency-retract motor works automatically if the primary motor fails to retract the active wire.

The SDU base provides a stable foundation for the SDU head and allows the unit to be transported easily.

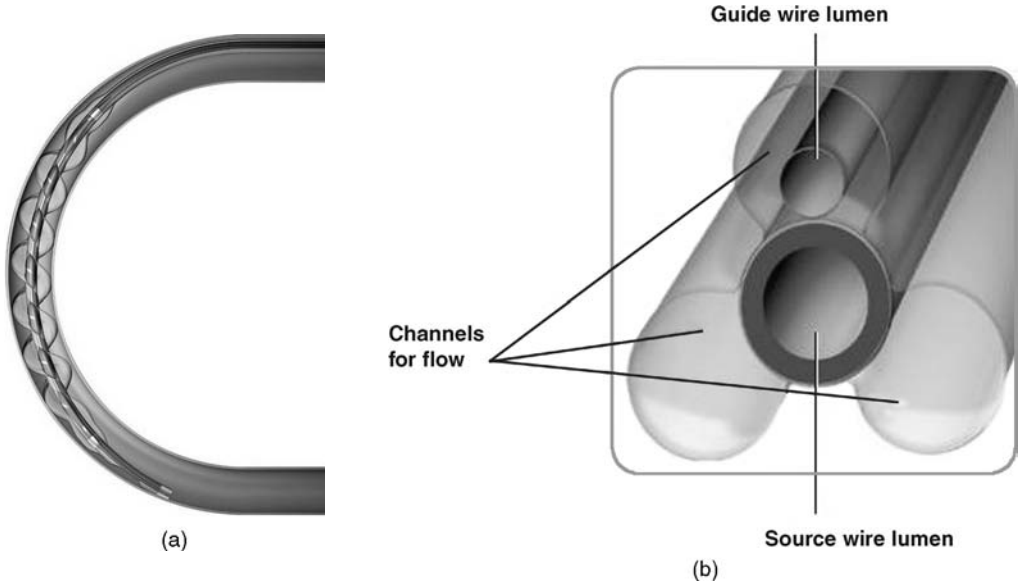


Figure 6. (a) Galileo first generation spiral centering catheter (Guidant Corporation, Santa Clara, CA) is shown inside a 3 mm diameter curved lumen, balloon is inflated to 4 atm with saline to provide optimal source centering and distal blood perfusion (Courtesy of Guidant Corporation—GALIELO system is no longer manufactured or for sale.) (b) Cross-section of a Galileo second generation trilobed centering catheter (Guidant Corporation, Santa Clara, CA) is shown. Source lumen is central and the three lobes are inflated to 4 atm with saline to provide optimal source centering and distal blood perfusion. Note guide wire lumen inside one of the lobes. (Courtesy of Guidant Corporation—GALIELO system is no longer manufactured or for sale.)

Table 5. Balloon Diameter Selection for the Guidant Galileo Centering Catheter as a Function of the Measured MLD

Balloon Diameter, mm	MLD, mm
2.5	2.25–2.75
3.0	2.75–3.25
3.5	3.25–3.7

Table 6. Balloon and Equivalent Source Length Selection

Balloon Length, mm	Injured Arterial Length, mm	Equivalent Source Length, mm
32	≤ 32	40
52	33–52	60

Components of the base include the head release, the handle bar, the emergency compartment, the wheels and wheel locks, and the power cord port. The emergency compartment contains the equipment necessary to handle an emergency, including the emergency safe, the emergency wire cutter, and the emergency tongs.

The cartridge, which is inserted into the SDU head, contains the active (^{32}P source) wire, the dummy wire, and the operating software. It also contains the catheter key port, the tungsten safe, which shields the active wire when not in use, and the wire drive mechanisms. Figure 8 is an example of the GALILEO software screen of the countdown clock.

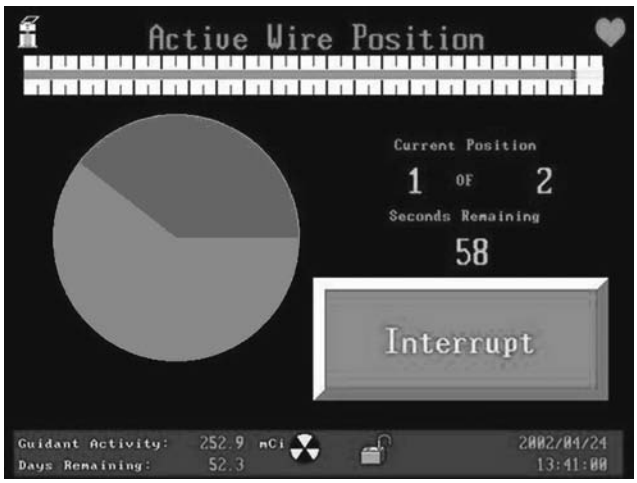


Figure 8. A screen shot of the Galileo software of the countdown clock for a treatment with two-source positions using the automatic source positioning system. (Courtesy of Guidant Corporation—GALIELO system is no longer manufactured or for sale.)

Dosimetry. Based on the measured minimal lumen diameter and lesion length via fluoroscopy, online QCA, or IVUS, a proper centering balloon size is chosen. Also the lumen diameter of the nondiseased vessel is measured immediately proximal and immediately distal to the treatment area. The average of these two diameters is the RLD. The GALILEO prescription point for radiation delivery is 1 mm beyond the RLD. The SDU automatically calculates the dwell time required to deliver the prescribed dose of radiation (20 Gy) at the prescription point. The GALILEO

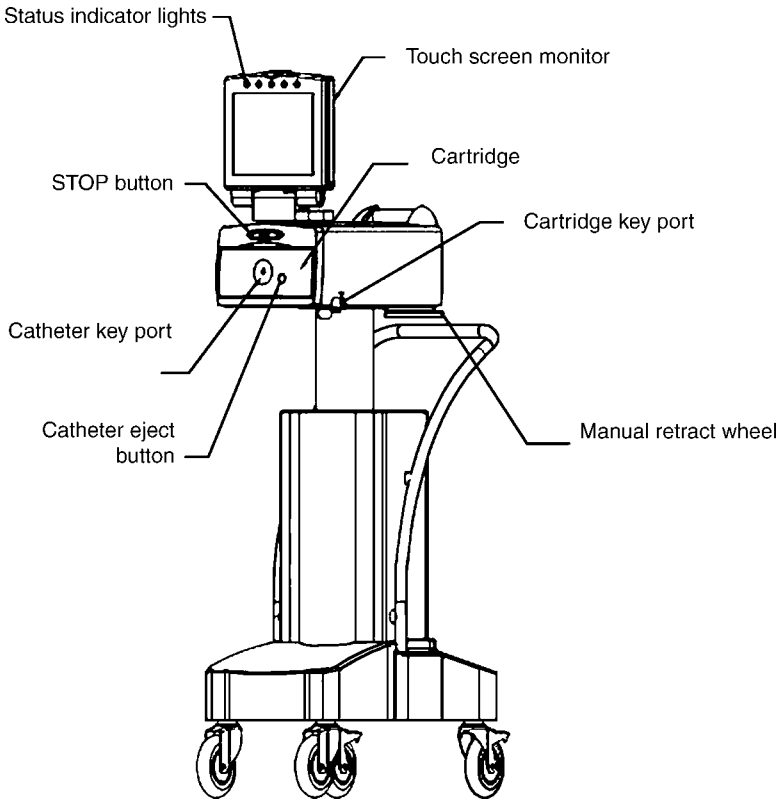


Figure 7. A front view of the Galileo source delivery system. It includes the Touch Screen Monitor, Status Indicator Lights, and housing for the Cartridge. It also contains the Manual Retract Wheel, the Cartridge Key Port, the Catheter Key Port and Catheter Eject Button, and the red STOP Button. (Courtesy of Guidant Corporation–GALIELO system is no longer manufactured or for sale.)

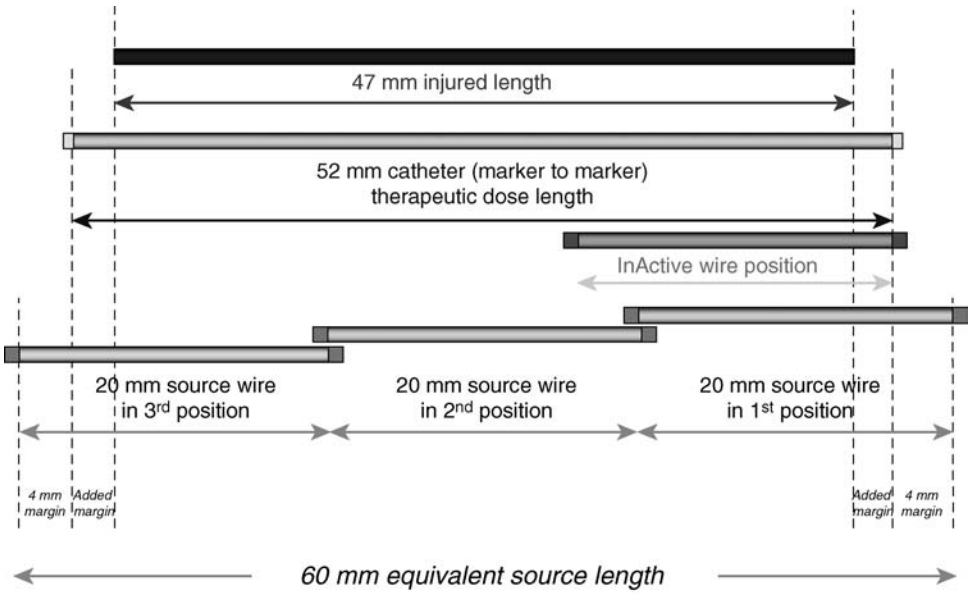


Figure 9. An example of 60 mm equivalent source length resulting from stepping twice a 20 mm ^{32}P source to treat a 47 mm lesion. Note the added margin of 2.5 mm to the minimal margin of 4 mm to provide adequate radiation coverage.

SDU automatically steps the 20 mm ^{32}P source from most distal position to the most proximal position to yield an equivalent source length sufficient to cover the injured length with a minimal margin of 4 mm (Table 6). Equivalent source length (ESL) is defined as the total source length that will be result by stepping the 20 mm active source in tandem, either in two dwell positions (40 mm equivalent source length) or in three dwell positions (60 mm equivalent source length). Figure 9 is an example of 60 mm ESL to treat a 47 mm lesion, note the added margin of 2.5 mm to the minimal margin of 4mm to provide adequate radiation coverage.

DESIGN CONSIDERATION FOR IVB DEVICES AND DOSIMETRY

For beta emitters (^{90}Sr/^{90}Y, ^{32}P) used currently in intravascular brachytherapy, the prescribed dose is greatly influenced by several perturbation factors. These perturbation factors are divided into two categories; the first is *applicator* dependent and the second is *patient anatomy* dependent. Important applicator perturbation factors include the placement of a guidewire during irradiation, the source lumen eccentricity within the applicator and centering capability with the vessel lumen, use of contrast, X-ray markers, and source stepping precision (if injured lengths longer than the source radioactive length are treated). Patient perturbation factors include vessel anatomy (size, curvature, and cross-section eccentricity), plaque morphology (composition, density, and spatial distribution), and stent type. High energy gamma emitters like) ^{192}Ir (J&J Checkmate system) are influenced by such factors as well, but with almost negligible magnitudes (22), hence the next discussion will focus on the does distributions perturbation of the beta sources only.

Applicator Dependent Perturbations

The applicator dependent factors discussed in this report are specific for two commercially available IVB beta source systems, that is, Novoste (^{90}Sr/^{90}Y) and Guidant (^{32}P) systems. Five factors are discussed and considered most important, but are not inclusive, (*1*) guidewire, (*2*) source lumen eccentricity within the applicator and centering capability within the vessel lumen, (*3*) use of contrast, (*4*) X-ray markers, and (*5*) source stepping precision (manual vs. automatic).

Guide Wire Perturbation. The guide wire (GW) is used to navigate through the cardiovascular arteries during common interventional procedures, such as balloon angioplasty, stenting, and IVB. Commonly used guide wires are made of stainless steel and have a solid cross-section with diameter of 0.014 in. (0.36 mm). Several authors have reported on the dose perturbation due to the guide wire in IVB (23–25). Figure 6b depicts the cross-section of 3.0 mm GIII centring catheter with a 0.014 in. (0.36 mm) guide wire inside the upper lope. The distance from the center of the GW and source axis is not fixed in this design and the GW location can vary (shown at closest distance from the source lumen center). Also Fig. 10 depicts the location of the guide wire inside the 5 Fr BetaCath catheter. Shih et al. (24) provides a detailed study on dose perturbation as a function of the GW position relative to the source axis. A dose perturbation

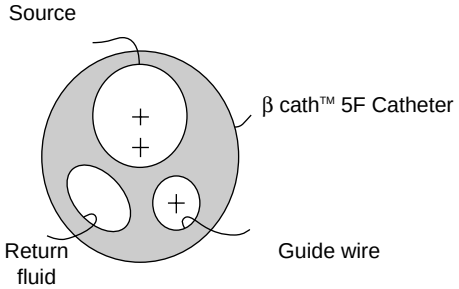


Figure 10. Cross-section of the BetaCath 5 F delivery catheter. Note the location of the guide wore lumen relative to the source lumen. (Courtesy of Novoste Corporation.)

factor (DPF), defined as the ratio of the doses with and without the presence of a guidewire was introduced to quantify the effects.[100] The authors reported a DPF of up to 70% behind a GW for the beta sources. The dose reduction for the beta sources was found to be dependent on the guidewire location. For example, the dose reduction was 10% higher for a stainless steel guidewire located at 0.5 mm than that for the guidewire at 2 mm from the central axis of the source. The portion of the target volume affected (shadowed) dosimetrically by the guidewire was reduced when the guidewire was positioned farther away from the source. The shadow volume (in which the dose reduction occurs) can be reduced by up to 45% as the guidewire is moved away from the source axis from 0.5 to 2 mm.

Fluhs et al. (25) measured the GW [0.014 in. (0.3556 mm), stainless steel] perturbation using a plastic scintillator. The authors pointed that the insertion of a GW into the catheter close to the beta source causes a large angular asymmetry of the radiation emission. For a guide wire positioned eccentrically to the catheter the dose reduction is dominantly limited to a region of some 20° around the angle defined by catheter centerline and guide wire. At the catheter surface the maximal dose reduction in this region was found to be ($30 \pm 2\%$, DPF = 10%). At the typical dose prescription depth of 2 mm from the source axis the shielding effect decreased to ($24 \pm 2\%$) (or DPF = 76%). This value is remarkably larger than the dose reduction caused by any typical stent design.

Source Lumen Eccentricity. Sehgal et al. (26) reports on the dosimetric consequences of source centering (eccentricity) within the arterial lumen as one potentially important factor for the uniform delivery of dose to the arterial tissue. In this study, they have examined the effect of source centering on the resulting dose to the arterial wall from clinical intravascular brachytherapy sources containing ^{32}P and ^{90}Sr/^{90}Y. Monte Carlo simulations using the MCNP code (described in Advanced Topics section below) were performed for these catheter-based sources with offsets of 0.5 and 1 mm from the center of the arterial lumen in homogenous water medium as well as in the presence of residual plaque. Three different positions were modeled and the resulting dose values were analyzed to assess their impact on the resulting dose distribution. The results are shown in Table 7. The debate on the importance of centering of beta emitters used in IVB to treat native coronary vessels has been extensively published (27).

Contrast Perturbation. Contrast agents with high atomic number materials are usually injected into blood vessels to help in the determination of lesion location and to verify source placement during the IVB procedure (small fraction of the entire dwell time). Common contrast agents like Omnipaque and Hypaque are discussed. Omnipaque contains ~25% of iodine (in mass), and Hypaque contains ~23% of iodine. Iodine has an atomic number (Z) of 53. Nath et al. (22) discussed the perturbation factor of these contrast agents on ^{32}P and ^{90}Y IVB sources when the contrast is injected directly in the blood stream, however, this paper did not provide the average DPF over the treatment dwell time. Mourtada used a Monte Carlo simulation of the Galileo III centering catheter (Fig. 6b) to calculate the average dose reduction at 1 mm tissue depth if the three catheter lobes are filled with saline (as recommended by the product instruction for use) and 50:50 Omnipaque contrast. The results are reported in Table 8. The simulations were done for the three different sizes of the GALILEO III centering balloon.

Table 7. Results Are Reported at a Radial Distance of 2 mm from the Coronary Artery Lumen Center[a]

Offset from Center, mm	^{32}P, %	^{90}Sr/^{90}Y, %
0.5	−40 to +70	−30 to +50
1.0	−65 to +185	−50 to +140

[a]Data is from Ref. 26.

Table 8. Average DPF at 1 mm into the Vessel Wall when the Galileo III Centering Catheter is Filled with 50:50 Omnipaque Contrast Agent[a]

Balloon diameter, mm	DPF, %
2.5	−2.9
3.0	−4.8
3.5	−7.6

[a]The contrast remains in the balloon for the entire ^{32}P irradiation dwell time. Data calculated by Mourtada using MCNP Monte Carlo code (unpublished data).

X-Ray Markers Perturbation. The GALILEO III centering catheter has distal and proximal X-ray markers made of 90% Pt and 10% Ir (effective density is 21.6 g cm^{-3}). The X-ray markers are 0.635 mm long and the inner and outer diameter are 0.394 and 0.432 mm, respectively. The GALILEO 20 mm source first position inside the GIII centering catheter is positioned 4 mm beyond the proximal edge of the distal X-ray marker to provide adequate margins. Figure 11a depicts the 20 mm source wire (red) and distal X-ray marker (gold). Using the Monte Carlo simulation MCNP, the dose distribution in water around a 20 mm ^{32}P source was calculated with and without the distal X-ray marker. Figure 11b depicts the two-dimensional (2D) isodose map with the distal X-ray marker in place. As expected due to scattering, the X-ray marker perturbation is reduced quickly as a function of depth. For a 3.0 mm diameter vessel example, the intimal surface maximum DPF is 23% and at the prescription depth of 1 mm into the vessel wall, the maximum DPF is 15%. The effect of the radiopaque marker in the BetaCath ^{90}Sr90/^{90}Y system has not been reported, but it is expected to be minimal since the catheter markers are along the side of the radioactive seed train.

Source Stepping. Lesions that are longer than available IVB sources require stepping the source (i.e., tandem positioned) from mostly distal to most proximal lesion segment to provide adequate radiation coverage. First-generation clinical beta-source systems have gained FDA approval for clinical indication for focal lesions (< 22 mm) (12,28). For injured lesions > 22 mm, but ≤ 47 mm, the manual tandem positioning (MTP) technique was investi-

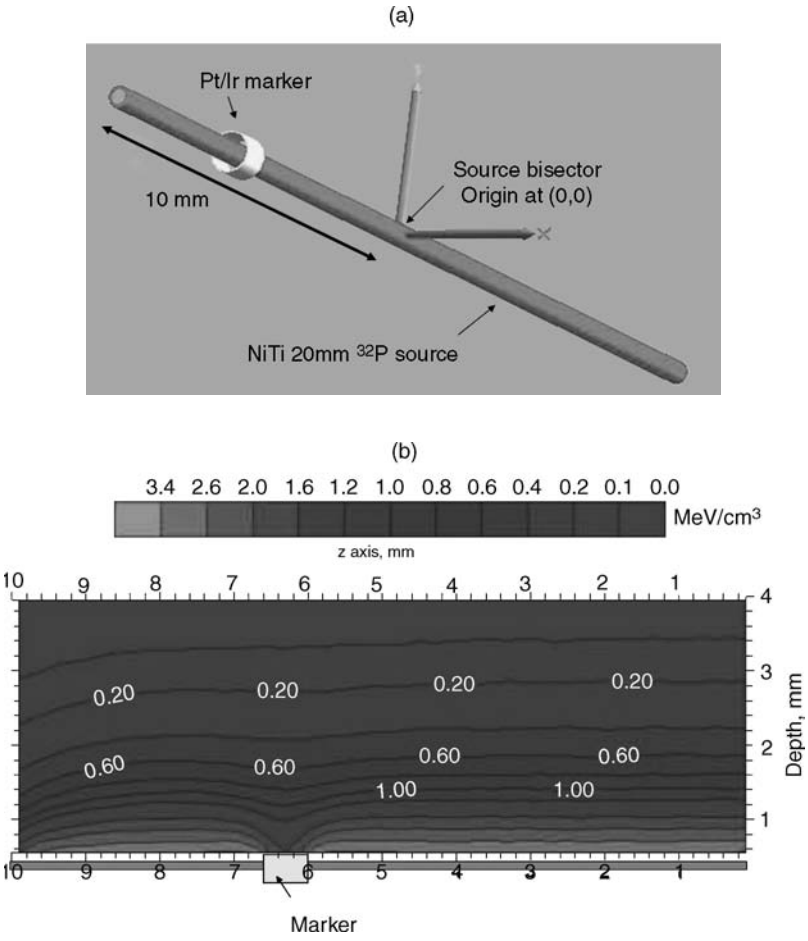


Figure 11. (a) Schematic of the distal X-ray marker of the Galileo second generation centering catheter (GIII) relative to nominal position of the 20 mm ^{32}P source inside the GIII catheter. (b) Energy deposited per unit volume in water (unit: MeV cm^{-3}) calculated using MNCNPX Monte Carlo code for the geometry shown in Fig. 11a. The origin of the coordinate system is located at the bisector of the 20 mm ^{32}P source. The proximal edge of the distal X-ray marker (gold) is located at 6 mm from the source origin. Note the large dose perturbation at closer depth (< 1 mm).

gated in the INHIBIT clinical trial using the 27 mm ^{32}P source (12). From the INHIBIT data analysis; for the 56 patients treated who had a tandem-positioning procedure (and the core lab had reported the size of gap or overlap), 44% had no gap or overlap. But, 19 and 11% of the patients had a 1 (32% increase in dose at junction) and 2 mm (56% increase in dose at junction) overlap respectively. Only one patient (1.9%) had a 1 mm gap (32% decrease in dose at junction). Hence, a 2 mm overlap and 1 mm gap are defined as the upper limits allowed for tandem positioning. Crocker et al. (29) also investigated the MTP procedure with 30 and 40 mm ^{90}Sr/^{90}Y source trains, and concluded from their data that the MTP technique was safe from both a dosimetric and a clinical point of view. However, Coen et al. (30) published a retrospective evaluation of the accuracy of manual multisegmental irradiation with 30 and 40 mm ^{90}Sr/^{90}Y source trains for irradiation of long (re)stenotic lesions in coronary arteries, following PTCA. They concluded that the positioning inaccuracy of MTP caused unacceptable dose inhomogeneities at the junction between source positions, and the procedure was not recommended. Coen et al. (30) suggested using longer line sources or source trains, or preferably an automated stepping source to insure reliable and safer technique for treatment of long lesions. Table 9 lists the dose perturbation at stepping junction due to an overlap or gap at the reference depth of 2 mm in water for GALILEO ^{32}P and Novoste ^{90}Sr/^{90}Y sources.

To reduce MTP dosimetric errors and to allow adequate coverage of radiation to longer injured lengths using the same source, an afterloader can be used to automatically step the radiation source to yield a longer equivalent source length. The only IVB system capable of this is the second generation GALILEO automated stepping system using a 20 mm ^{32}P source wire (33).

Patient Anatomy Dependent Perturbations

Patient perturbation factors discussed in this article include (*1*) vessel geometry (size, curvature, tapering,

Table 9. Dose Perturbation at Junction Due to an Overlap or Gap at the Reference Depth of 2 mm in Water for GALILEO ^{32}P and Novoste ^{90}Sr/^{90}Y Sources

Size of Overlap or Gap, mm	^{32}P[a], %	^{90}Sr/^{90}Y[b], %
0	0	0
0.5	±17	
1	±32	±23
2	±56	±44
3	±75	±60
5	±91	±80

[a] Guidant Corporation Data to Ref. 31.
[b] See Ref. 32.

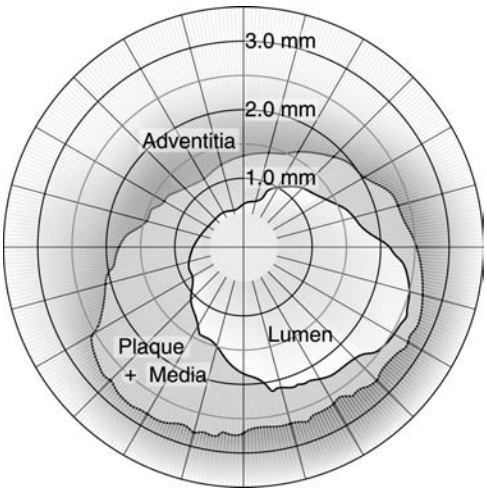


Figure 12. A typical vessel cross-section is eccentric, and the vessel center does not necessarily lie on the lumen center.

and cross-sectional eccentricity); (*2*) plaque morphology (composition, density, and spatial distribution); (*3*) stenting.

Vessel Geometry Perturbation. The native coronary vessel diameter has a range from 2 to 4 mm. Due to plaque formation and tapering, the lumen diameter is a variable. It is expected that the dose perturbation factor to be worse for larger vessels, especially for the BetaCath IVB system whose dose does not have active centering. Also, a typical vessel cross-section is eccentric, and the vessel center does not necessarily lie on the lumen center (Fig. 12). As pointed by Kaluza et al. (27) in reality, plaque thickness can vary from 0.1 to 2.3 mm. The average eccentricity index was 6.38 ± 5.95 in the 59 PREVENT patients. The intimal hyperplasia of in-stent restenosis complicates the vessel cross-section. Mehran et al. (6) developed an angiographic classification of in-stent restenosis mainly under two categories: focal and diffused (6).

Another important vessel geometry factor is curvature. Xu et al. (34) studied the effect of curvature on ^{32}P beta dosimetry. As expected, the curvature causes an increase in dose in the inner surface (concave side) of the coronary vessel and a decrease in dose in the outer surface (convex side). For a maximum theoretical bend of 180°, the dose increases by as much as 20% along the inner radial distance, but decreased by as much as 20% along the outer radial distance compared to the dose along a straight wire. The authors concluded that for curvatures normally encountered in a clinical situation, the dose rate was changed by $< 5\%$.

Plaque Morphology Perturbation. The artery mostly consists of normal healthy tissue, but may also contain plaque, whose density may be unknown. Plaque is a material that develops inside the artery over time and is considered responsible for blockage of the artery. Plaque may range widely in histologic structure, density, and chemical composition. Density is expected to depend on the plaque's collagenous matrix and degree of calcification. Rahdert et al. (35) measured the density and calcium concentration

Table 10. DPF Values at 2 mm Radial Distance for ^{32}P and ^{90}Sr/^{90}Y Sources[a]

Plaque Density, g cm^{-3}	DPF, ^{32}P	DPF, ^{90}Sr/^{90}Y
No plaque	1.0	1.0
1.45	0.93	0.97
1.55	0.91	0.96
3.10	0.70	0.83

[a]The plaque layer has a thick-ness of 0.2 mm for all the different cases. The relative error is within 5% for the given dose rate values. (See Ref. 26).

in 13 cadaveric plaque specimens. This study concluded that based on the plaque calcification, the density range is between 1.25 and 1.5 $g \cdot cm^{-3}$.

Several studies on dose perturbation due to plaque have been reported in the literature for catheter-based beta sources (36–38). Nath et al. (36) assumed a 1 mm thick plaque with cortical bone density of 1.84 $g \cdot cm^{-3}$ and 27% Ca composition. Both values are relatively higher than those reported by Rahdert et al. (35). For this extreme condition, however, Nath et al. (36) reported ~ 0.8 mm reduction in penetration for ^{90}Sr/^{90}Y and ^{32}P beta sources when the calcified plaque was located next to the source, and by ~ 0.9 mm when the plaque was located 1 mm away from the source.

Li et al. reported $\sim 30\%$ reduction due to plaque for the Novoste ^{90}Sr/^{90}Y source (37). The modeled calcified plaque density range in this study was 1.2–1.60 $g \cdot cm^{-3}$ with a nominal density of 1.45 $g \cdot cm^{-3}$ as reported for B100 bone equivalent by ICRU Report 26 (39). Li et al. (37) calculated the DPF as a function of the radial distance from the source axis. Sehgal et al. (26) reported on the perturbation due to concentric plaque with constant thickness of 0.2 mm and three different densities as shown in Table 10 for both the GALILEO ^{32}P and Novoste ^{90}Sr/^{90}Y sources. Comparing with Li et al. (37) the 0.2 mm thick plaque with 1.45 $g \cdot cm^{-3}$ DPF results are similar.

Stent Perturbation. Even though a stent is not part of the actual vessel anatomy, it is assumed that an expanded stent is predislodged into the intima and surrounded by new tissue growth as a result of in-stent neointimal hyperplasia. Hence, the stent becomes part of the diseased vessel segment. In a few instances, a vessel could receive a second or even a third stent, as part of the intervention of a recurrent in-stent restenosis.

For all stent types, a similar behavior can be observed. At a distance of 0.5 mm and more the typical overall dose reduction does not exceed a value of 5–15% (40). In the close vicinity of the stent struts, an increased dose reduction effect reaches up to some 30–40%. The large dose reduction directly behind the stent struts are caused by the absorption effect that is not compensated by scattering contributions from regions outside of the shielded area until a depth of ~ 0.5 mm from the strut surface (41).

Recently, a new stent made of cobalt chromium L-605 alloy (CoCr, $\rho = 9.22$ $g \cdot cm^3$) (MULTI-LINK VISION) was introduced as an alternative to the commonly used 316L stainless steel stent design (SS, $\rho = 7.87$ $g \cdot cm^{-3}$) (MULTI-LINK PENTA). Mourtada and Horton (42) used the Monte

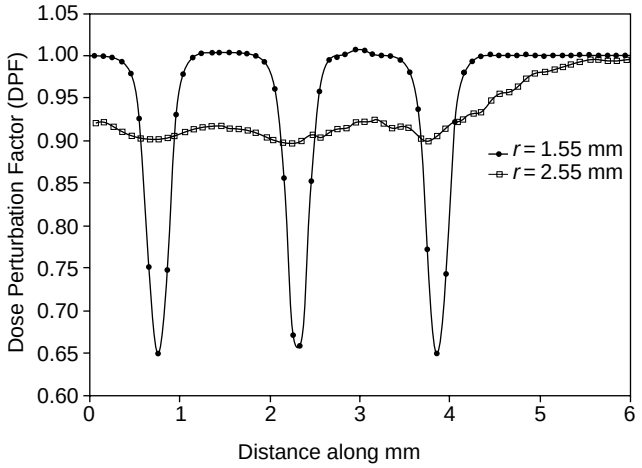


Figure 13. A 3 mm diameter stainless steel stent DPF along the source axis ($x = 0$ is the coordinate system origin at source bisector). The parameter $r = 1.55$ mm is the radial distance from the source axis and is the centroid of the scoring bin directly behind the stent (score bin radial thickness is 0.1 mm). The parameter $r = 2.55$ mm is the centroid of the scoring bin at the 2.5 mm prescription point. Data calculated using MCNPX Monte Carlo code. (Reference 42 with permission from Medical Physics journal.)

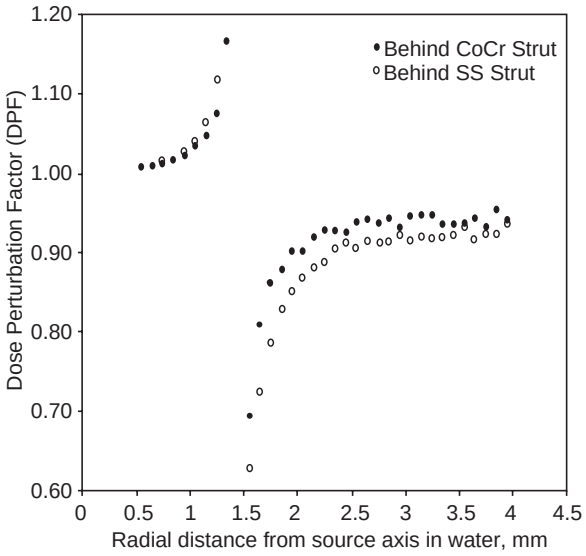


Figure 14. Dose perturbation factor as a function of the radial distance from source axis in water and through a strut located 1.5 mm from the source axis (expanded inside a 3.0 mm vessel model). (Reference 42 with permission from Medical Physics journal.)

Carlo code MCNPX to compare the dose distribution for the ^{32}P GALILEO source in CoCr and SS 8 mm stent models. The DPF, defined as the ratio of the dose in water with the presence of a stent to the dose without a stent, was used to compare results. Both stent designs were virtually expanded to diameters of 2.0, 3.0, and 4.0 mm using finite element models (ABQUS Inc., Pawtucket, RI). The complicated strut shapes of both the CoCr and SS stents were simplified using circular rings with an effective width to yield a metal/tissue ratio identical to that of the actual stents. The mean DPF at a 1 mm tissue depth, over the entire stented length of 8 mm, was 0.935 for the CoCr stent and 0.911 for the SS stent. The mean DPF at the intima (0.05 mm radial distance from the strut outer surface), over the entire stented length of 8 mm, was 0.950 for CoCr, and 0.926 for SS. The maximum DPFs directly behind the CoCr and SS struts were 0.689 and 0.644, respectively. Figures 13 and 14 depict the dose profiles behind the stainless steel stent as an example. The authors concluded that although the CoCr stent has a higher effective atomic number and greater density than the SS stent, the DPFs for the two stents are similar because the metal/tissue ratio and strut thickness of the CoCr stent are lower than those of the SS stent.

ADVANCED TOPICS

Figure 15 is the general dosimetry characterization paradigm of brachytherapy sources. This requires three important steps including the experimental measurement of the dose rate ($cGy^{-1} \cdot s$) as a function of distance from the source using a calibrated dosimeter, a measurement of the source contained activity (SI unit is the Becquerel = Bq = 1 decay s^{-1}). The normalized measure dose rate to the measured contained activity can then be compared to a theoretical calculation such as a Monte Carlo simulation, which is inherently has dose rate units per particle emitted, that is, contained activity. This section will discuss briefly an example of each component that is important in the IVB dosimetry paradigm.

Theoretical Dosimetry: Monte Carlo Simulations

The Monte Carlo technique used for IVB dosimetry is considered the most accurate theoretical tool, particularly suited in handling the complex interactions of the emitted beta particles with the surrounding medium. Fox has published a review of IVB (43), including a rather thorough review of the theoretical dosimetry applied to these sources. Detailed reviews and discussion of the Monte Carlo method can also be found in the literature (44–46). Monte Carlo codes utilized for IVB dosimetry include CYLTRAN from the Integrated Tiger Series (ITS version 3.0, Sandia National Laboratory, Albuquerque, NM), the MCNP series (MCNP4C and MCNPX, Los Alamos

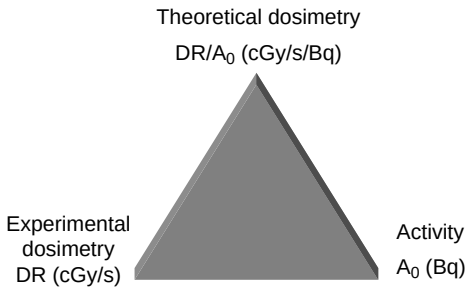


Figure 15. A general dosimetry characterization paradigm of brachytherapy sources.

National Laboratory, Los Alamos, NM), the EGS series (EGS4 Stanford Linear Accelerator Center, Stanford, CA, and EGSnrc National Research Council of Canada, Ottawa, Ontario, Canada), and PENELOPE (University of Barcelona, Barcelona, Catalonia, Spain).

Theoretical modeling is being increasingly used to supplement measurements in IVB. The Monte Carlo method, for example, is based on the idea that if all materials and dimensions of a problem, and all the probabilities of the various possible radiation interactions are known, then emitted particles can be tracked and scored as they are transported from the source through the media of the problem. A random number generator is used to select emitting source element location, emitted particle energy and direction, and results of various interactions and secondary radiations as the particle is tracked to either absorption or out of the problem boundaries. Obviously, the more "histories" (emitted particles) are tracked, the more accurate is the resulting calculation. With the advent of faster and faster computers, Monte Carlo calculations are becoming more and more attractive for determining dose distributions for brachytherapy sources in complex geometries. Often, the combination of a Monte Carlo calculation, which yields dose rate per unit contained activity, and a contained activity measurement, will give better accuracy than a dosimetric measurement. An excellent review of the Monte Carlo method has recently been published (44).

MC simulations of electron transport, for example, beta-particle transport, are usually different from those used in photon or neutron transport, for which the simulated radiation history is followed individually based on conventional methods since uncharged radiation interactions are characterized by relatively infrequent isolated collisions. For example, in photon transport, the distance to the next photon interaction is sampled from the attenuation coefficient distribution; and the change in attenuation coefficient as the photon crosses material boundaries is modeled. The type of interaction is sampled from the appropriate relative probabilities. The history of each photon is continued from collision to collision until the photon either is absorbed, escapes the problem boundary, or its energy falls below a chosen cutoff threshold at which the remaining energy of the photon is locally deposited.

For high energy electrons, such a detailed history is not practical for energies > 100 keV, because many individual elastic and inelastic Coulomb collisions per history are generated through the media resulting in very long computational time. Instead, a "condensed history" is used, where the electron trajectories are divided into many path segments (47). For each path segment, the net angular deflection and the net energy loss are sampled from relevant multiple-scattering distributions. The choice of the step size is important for accuracy and is chosen with conflicting requirements. On the one hand, the steps should be short enough that (*1*) most of the electron history steps are completely inside the boundary of a predefined surface, so that the use of multiple-scattering theories of unbounded media is valid; (*2*) the energy loss is, on average, small within a step; and (*3*) the net angular deflection is, on average, small so that the path within the step is approximated by a straight line. On the other hand, the step size should be large enough to contain a sufficient number of collisions per step to justify the use of the multiple-scattering theories and to limit the number of steps per history to reduce computing time. Further discussion can be found in the literature (45).

Experimental Dosimetry: Radiochromic Film Measurements

Both MD-55 and HD-810 radiochromic dye films (RCF) (GAFChromic type, Nuclear Associates, Carle Place, NY) are widely used in IVB dose-field measurement due to their superior spatial resolution. Also, RCF is used instead of other types of films mainly for its linear dose response. A full description of both film types is reported by AAPM Task Group 55 (48) For beta field measurements, it is recommended that the RCF dosimeter be calibrated using the same $^{90}Sr/^{90}Y$ ophthalmic applicator (New England Nuclear S/N 0258) calibrated at the National Institute of Standards and Technology (NIST), Gaithersburg, Maryland (20). Polystyrene or other tissue-equivalent materials are used to fabricate high precision blocks for the film measurements. Each block has a hole with a diameter slightly larger than the source diameter to reduce positional error (in IVB 0.1 mm could translate to 13% error in the measured dose rate). Several blocks are made with nominal depths (distance from center of hole to block surface) ranging from 0.5 to 5 mm. Actual depths must be verified using a traveling microscope or an optical comparator. At each of these depths, several radiochromic films should be exposed for a range of times to gain a good image of the radiation field. Digitization of film is typically done with a high resolution 2 scanning densitometer (Pharmacia LKB) using a 633 nm laser (HeNe) with a 100 μm diameter spot size and a 40 μm minimum step size. Alternatively, scanning is done using a high resolution (242×375) CCD densitometer (CCD100, Photoelectron, Lexington, MA) with a 665 nm LED array and a 160×200 μm pixel size. To account for optical-density growth as a function of time after exposure, film readout should be done 72 h postirradiation for both the calibration and experimental films. The net optical density measurements of each film were converted into a 2D dose map as shown in Fig. 16. Estimated dose uncertainties for radiochromic film are ± 15.6 % ($k = 2$); the individual components of this uncertainty are shown in Table 11.

Other radiation dosimeters mostly lack the spatial resolution of submillimeters required in IVB. However, recently Amin et al. (49) proposed using a polyacrylamide gel (PAG) dosimeter and a high field 4.7 T MRI scanner for IVB. The get/scanner final in-plane resolution of 0.4×0.2 mm is approaching the film resolution, but not quite. The authors confirmed that both absorbed dose and dose distributions for high gradient vascular brachytherapy sources can be measured using PAG, but the disadvantages of gel manufacture and the need for access to a high resolution scanner suggests that the use of radiochromic film is the method of choice (49).

Determination of Contained Activity. The activity content of this wire must be known in order to relate the measurements and calculations of the absorbed-dose spatial

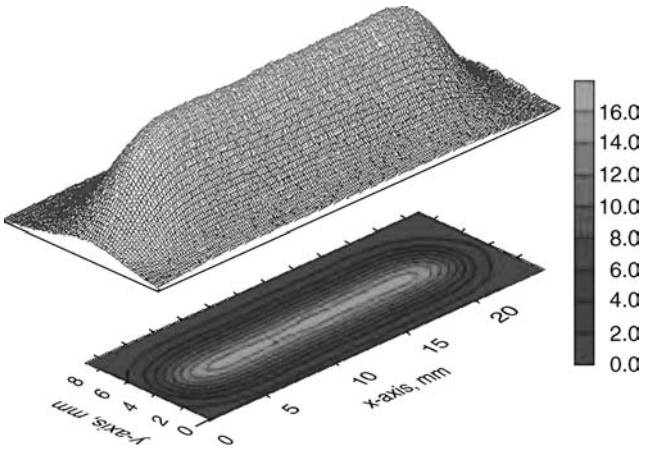


Figure 16. A 2D HD810 radiochromic film image dose (Gy) distribution for a ^{32}P source, measured in a plane parallel to the source's longitudinal axis at 1.97 mm radial distance from the source axis in the polystyrene block.

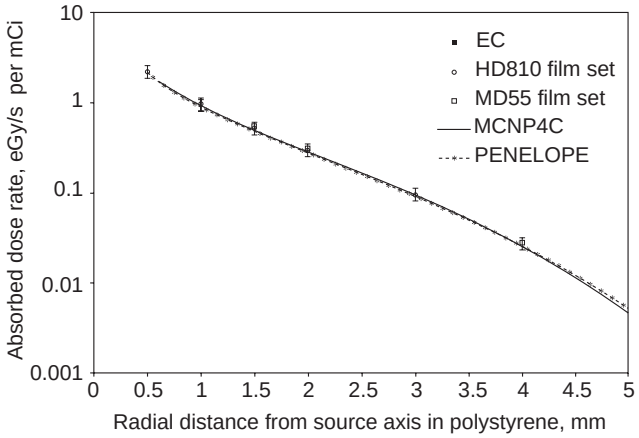


Figure 17. Measured data (Film and EC: extrapolation chamber) depth dose curve plotted along with Monte Carlo estimates from MCNP4C and PENELOPE. The error bars estimate the 95% confidence interval. All data are measured or calculated in polystyrene. (Reference 21 with permission from Medical Physics journal.)

distribution. For the 27 mm ^{32}P source design, Mourtada et al. briefly described the determination of the absolute contained activity from the original work of Collé (50). The contained activity was then related to a calibration factor for the NIST Capintec. The CRC-12 ionization chamber (51). Similar work was done on the Novoste ^{90}S/^{90}Y seed and other beta sources used for IVB to establish radioactivity standards by NIST, Gaithersburg, MA (52,53).

For example, Fig. 17 is the Galileo 20 mm ^{32}P source wire measured radiochromic film (MD55 and HD810) depth dose curve plotted along with Monte Carlo estimates from MCNP4C and PENELOPE. The error bars estimate the 95% confidence interval. All data are measured or calculated in polystyrene (21).

Beyond IVB Treatments of Heart Disease, Other Applications, and Future Roles

In the United States, there are $\sim$8–12 million patients affected with peripheral vascular disease. An estimated 600,000 interventional procedures are performed each year, including percutaneous transluminal angioplasty (PTA), bypass surgery, and amputation. Percutaneous transluminal angioplasty restenosis rates are high with a success rate of $<$ 23% at 6 months follow-up. Intravascular brachytherapy has been investigated to reduce restenosis in the superficial femoral and popliteal arteries after PTA. The main IVB clinical trial for peripheral vessel is the Peripheral Artery Radiation Investigation Study (PARIS) using the Nucletron Ir-192 HDR source (mHDR v2). The PARIS study used the Nucletron micro-Selectron high dose rate (HDR) afterloader and the Guidant PARIS centering catheter (10–20 cm long and 4–8 mm diameter) (54). More recently, the MOBILE clinical trial a ^{90}Sr/^{90}Y source and the Corona gas-filled centering is being investigated (Novoste Corporation) (55).

Other possible applications of IVB include treatment of recurrent narrowing of the arteriovenous (AV) dialysis graft (56), renal artery stenosis, transjugular intrahepatic porto-systemic (TIPS) stenosis, carotid artery (57), and subclavian vein stenosis. Further application of IVB might be for treatment of atrial fibrillation, a most common cardiac arrhythmia. It is proposed that IVB radiation dose can electrically isolate ectopic foci located mostly in the adventitia of the pulmonary vein (PV), which are responsible for atrial fibrillation episodes (58). The IVB approach might alleviate undesirable side effects (PV stenosis due to heating) of rf ablation; a commonly used treatment modality to ablate myocardial tissue.

Intravascular Brachytherapy in the Drug-Eluting Stent Era

The recent introduction of drug-eluting stents (DES) in the interventional cardiology arena has a tremendous impact on the IVB practice. By incorporating antiproliferative agents onto the surface of the stent, neointimal hyperplasia

Table 11. Estimated Relative Uncertainties of the Radiochromic-Film Dose Interpretations per Measured Unit Activity

Uncertainty Component	Relative Standard Uncertainty, %
Calibration of the NIST standard ^{90}Sr/^{90}Y calibration source	6
Response of the film exposed to the calibration source	3
Response of the films exposed to the source under test	3
Activity calibration	2.6
Combined standard uncertainty	7.8
Expanded uncertainty ($k = 2$)	15.6

occurring within the stent is markedly reduced. Stents coated with agents, like Sirolimus, Paclitaxel, Tacrolimus, Everolimus, and so on, when compared to bare-metal stents, had shown remarkable reduction in binary restenosis and target vessel revascularization (TVR) rates in several clinical trials (31,59).

As discussed earlier in this article, IVB has demonstrated its safety and efficacy in limiting recurrence of in-stent restenosis with positive long-term follow-up outcome. However, the utility of IVB is being rethought in relation to its use after the placement of drug eluting stents. It is expected that the role of IVB will decrease primarily due to the simplicity of placement of drug eluting stents and the relative complexity of performing intravascular brachytherapy. The published pivotal clinical restenosis rates from both the Paclitaxel (TAXUS) (32) and Sirolimus (RAVEL, SIRIUS) drug eluting stents suggests that the need to perform an IVB will be in < 1 in 20 patients (60–62). Recently reported registry data from Europe and the United States demonstrates similar efficacy of DES to that of IVB for treatment of in-stent restenosis (63,64). However, DES technology still has to go through further investigations in more complex lesions and higher risk patients in the general population to appreciate its full potential. Other agents with potential benefits like Statins, local gene therapy, and further innovations in polymer technology (biodegradable polymers, multiple-drug release polymers) are in under evaluation (31).

DEFINITION OF TERMS

MACE: Major Adverse Cardiac Events, a composite of death, MI, target, or repeat lesion revascularization.

Angiographic Binary Restenosis: Stenosis of 50% or more of the luminal diameter.

TLR and TVR: Target Lesion and Vessel Revascularization: Describes a rate measuring how many stented lesions had to be retreated, due to clinically driven restenosis, given a specific time period.

MLD: Minimal lumen diameter, the smallest diameter of an artery in a specified segment.

RLD: Reference lumen diameter, the average of two diameters, the lumen diameter of the nondiseased vessel immediately proximal and distal to the target treatment area.

Lumen Diameter: The inner diameter of an artery in a specified segment.

Neointimal Hyperplasia: Wound-healing response to arterial injury that leads to restenosis.

Late Loss: A cardiology term referring to the angiographic measurement of neointimal hyperplasia. It's one of the most important indicators of long-term efficacy in coronary intervention.

Percutaneous Transluminal Coronary Angioplasty (PTCA): A method of treating blood vessel disorders that involves the use of a balloon catheter to enlarge the blood vessel and thereby improve blood flow.

Restenosis: Narrowing of a vessel dilated by angioplasty or other interventional procedure.

In-stent Restenosis (ISR): Narrowing of a vessel after a stent is in place; this may be acute due to a thrombus formation or late (few months) due to wound healing process (neointimal hyperplasia and remodeling).

BIBLIOGRAPHY

Cited References

1. Gruentzig AR. Seven years of coronary angioplasty. Z Kardiol 1984;73 (Suppl.) 2:159–160.
2. Dotter CT, Buschmann RW, Mckinney MK, Rosch J. Transluminal expandable nitinol coil stent grafting: Preliminary report. Radiology 1983;147:259–260.
3. Cragg A, et al. Nonsurgical placement of arterial endoprostheses: A new technique using nitinol wire. Radiology 1983;147: 261–263.
4. Schwartz RS, Homes DR. Restenosis and remodeling. In: Waksman R. editor. Vascular brachytherapy. Armonk, (NY): Futura Publishing Company; 1999.
5. Hall EJ, Miller RC, Brenner DJ. Radiobiological principles in intravascular irradiation. Cardiovasc Radiat Med 1999;1:42–47.
6. Mehran R, et al. Angiographic patterns of in-stent restenosis: Classification and implications for long-term outcome. Circulation 1999;100:1872–1878.
7. Fischman DL, et al. A randomized comparison of coronary-stent placement and balloon angioplasty in the treatment of coronary artery disease. Stent restenosis study investigators. N Engl J Med 1994;331:496–501.
8. Serruys PW, et al. A comparison of balloon-expandable-stent implantation with balloon angioplasty in patients with coronary artery disease. Benestent study group. N Engl J Med 1994;331:489–495.
9. Condado JA, et al. Long-term angiographic and clinical outcome after percutaneous transluminal coronary angioplasty and intracoronary radiation therapy in humans. Circulation 1997;96:727–732.
10. Leon MB, et al. Localized intracoronary gamma-radiation therapy to inhibit the recurrence of restenosis after stenting. N Engl J Med 2001;344:250–256.
11. Popma JJ, et al. Randomized trial of $^{90}Sr/^{90}Y$ beta-radiation versus placebo control for treatment of in-stent restenosis. Circulation 2002;106:1090–1096.
12. Waksman R, et al. Use of localised intracoronary beta radiation in treatment of in-stent restenosis: The inhibit randomised controlled trial. Lancet 2002;359:551–557.
13. Balter S. A health physics perspective. In: WR, editor. Vascular brachytherapy. New York: Futura Publishing Company; 1999.
14. Barish R. Radiation safety considerations for intravascular brachytherapy. In: Balter S, Chan RC, Shope TB, editors. Intravascular brachytherapy and fluoroscopically guided interventions. Madison (WI): Medical Physics Publishing; 2002.
15. Waksman R. Intravascular gamma radiation for in-stent restenosis in saphenous-vein bypass grafts. N Engl J Med 2002;346:1194–1199.
16. Jani S, Massullo V, Tripuraneni P, Teristein P. The ^{192}Ir radioactive seed ribbon. In: Waksman R, Serruys P, editors. Handbook of vascular brachytherapy. London: Martin Dunitz Ltd; 2000.
17. Chiu-Tsao ST, et al. Verification of ^{192}Ir near source dosimetry using gafchromic film. Med Phys 2004;31:201–207.
18. Roa DE, et al. Dosimetric characteristics of the novoste beta-cath $^{90}Sr/Y$ source trains at submillimeter distances. Med Phys 2004;31:1269–1276.

19. Soares CG, Halpern DG, Wang C-K. Calibration and characterization of beta-particle sources for intravascular brachytherapy. Med Phys 1998;25:339–346.
20. Mourtada FA, Soares CG, Seltzer SM, Lott SH. Dosimetry characterization of ^{32}P catheter-based vascular brachytherapy source wire. Med Phys 2000;27:1770–1776.
21. Mourtada F, et al. Dosimetry characterization of a ^{32}P source wire used for intravascular brachytherapy with automated stepping. Med Phys 2003;30:959–971.
22. Nath R, Yue N, Weinberger J. Dose perturbations by high atomic number materials in intravascular brachytherapy. Cardiovasc Radiat Med 1999;1:144–153.
23. Li XA, Shih R. Dose effects of guide wires for catheter-based intravascular brachytherapy. Int J Radiat Oncol Biol Phys 2001;51:1103–1110.
24. Shih R, Hsu WL, Li XA. Dose effect of guidewire position in intravascular brachytherapy. Phys Med Biol 2002;47:1733–1740.
25. Fluhs D, et al. The influence of guiding equipment and stents on the beta dose distribution in the brachytherapy of in-stent restenosis. Cardiovasc Radiat Med 2001;2:241–245.
26. Sehgal V, Li Z, Palta JR, Bolch WE. Dosimetric effect of source centering and residual plaque for beta-emitting catheter based intravascular brachytherapy sources. Med Phys 2001;28:2162–2171.
27. Kaluza GL, et al. Targeting the adventitia with intracoronary beta-radiation: Comparison of two dose prescriptions and the role of centering coronary arteries. Int J Radiat Oncol Biol Phys 2002;52:184–191.
28. Suntharalingam M, et al. Clinical and angiographic outcomes after use of 90strontium/90yttrium beta radiation for the treatment of in-stent restenosis: Results from the stents and radiation therapy 40 (Start 40) registry. Int J Radiat Oncol Biol Phys 2002;52:1075–1082.
29. Crocker I, et al. Treatment of long, diffuse, in-stent restenotic lesions with beta radiation using strontium 90 and sequential positioning "pullback" technique: Procedural details and clinical outcomes. J Invasive Cardiol 2001;13:782–787.
30. Coen VL. Inaccuracy in manual multisegmental irradiation in coronary arteries. Radiother Oncol 2002;63:89–95.
31. Fattori R, Piva T. Drug-eluting stents in vascular intervention. Lancet 2003;361:247–249.
32. Stone GW, et al. One-year clinical results with the slow-release, polymer-based, paclitaxel-eluting taxus stent: The taxus-IV trial. Circulation 2004;109:1942–1947.
33. Waksman R, et al. Beta radiation delivered via an automatic stepping device to inhibit recurrence of diffuse in-stent restenosis: Clinical and angiographic results of the multicenter galileo inhibit clinical study. Circulation (Suppl II) 2001;104:II–509.
34. Xu Z, et al. The investigation of ^{32}P wire for catheter-based endovascular irradiation. Med Phys 1997;24:1788–1792.
35. Rahdert DA, et al. Measurement of density and calcium in human atherosclerotic plaque and implications for arterial brachytherapy. Cardiovasc Radiat Med 1999;1:358–367.
36. Nath R, Yue N, Liu L. On the depth of penetration of photons and electrons for intravascular brachytherapy. Cardiovasc Radiat Med 1999;1:72–79.
37. Li XA, Wang R, Yu C, Suntharalingam M. Beta versus gamma for catheter-based intravascular brachytherapy: Dosimetric perspectives in the presence of metallic stents and calcified plaques. Int J Radiat Oncol Biol Phys 2000;46:1043–1049.
38. Hanefeld C, et al. Dosimetric measurements in isolated human coronary arteries: Comparison of commercially available iridium(192) with strontium/yttrium(90) emitters. Circulation 2002;105:2493–2496.
39. Icru Report 26. International Commission on Radiation Units and Measurements, Bethesda (MD); 1977.
40. Fan P, et al. Effect of stent on radiation dosimetry in an in-stent restenosis model. Cardiovasc Radiat Med 2000;2:18–25.
41. Amols HI, Trichter F, Weinberger J. Intracoronary radiation for prevention of restenosis: Dose perturbations caused by stents. Circulation 1998;98:2024–2029.
42. Mourtada F, Horton JL. Dose perturbation of a novel cobalt chromium coronary stent on ^{32}P intravascular brachytherapy: A monte carlo study. Med Phys 2005;32:268–274.
43. Fox RA. Intravascular brachytherapy of the coronary arteries. Phys Med Biol 2002;47:R1–30.
44. Seltzer SM. Monte Carlo modeling for intravascular brachytherapy sources. In: Shope TB, editor. Intravascular brachytherapy and fluoroscopically guided interventions. Madison (WI): Medical Physics Publishing; 2002.
45. Jenkins TM, Nelson WR, Rindi A, editors. Monte carlo transport of electrons and photons. New York: Plenum Press; 1988.
46. Icru Report 56. International Commission On Radiation Units and Measurements, Bethesda (MA); 1997.
47. Berger MJ. Monte carlo calculations of the penetration and diffusion of fast charged particles. Alder B, Fernabach S, Rotenberg M, editors. Methods in computational physics. New York: Academic Press; 1963.
48. Niroomand-Rad A, et al. Radiochromic film dosimetry: Recommendations of aapm radiation therapy committee task group 55. Med Phys 1998;25:2093–2115.
49. Amin MN, et al. A comparison of polyacrylamide gels and radiochromic film for source measurements in intravascular brachytherapy. Br J Radiol 2003;76:824–831.
50. Collé R. Chemical digestion and radionuclide assay of tini-encapsulated 32p intravascular brachytherapy sources. Appl Rad Isotopes 1999;50:811–833.
51. Colle R, Zimmerman BE, Soares CG, Coursey BM. Determination of a calibration factor for the nondestructive assay of guidant 32P brachytherapy sources. Appl Radiat Isotopes 1999;50:835–841.
52. Collé R. On the radioanalytical methods used to assay stainless-steel-encapsulated, ceramic-based ^{90}Sr/^{90}Y intravascular brachytherapy sources. Appl Radiat Isotopes 2000;52:1–18.
53. Colle R. Activity characterization of pure-beta-emitting brachytherapy sources. Appl Radiat Isotopes 2002;56:331–336.
54. Waksman R, et al. Intravascular radiation therapy after balloon angioplasty of narrowed femoropopliteal arteries to prevent restenosis: Results of the paris feasibility clinical trial. J Vasc Interv Rad 2001;12:915–921.
55. Wang R, Li XA, Lobdell J. Monte carlo dose characterization of a new 90sr/90y source with balloon for intravascular brachytherapy. Med Phys 2003;30:27–33.
56. Bloch P, Bonan R, Wallner P, Lobdell J. Dosimetry for an ^{90}Sr/^{90}Y source train used for intravascular radiation of a hemodialysis graft. Cardiovasc Rad Med 2003;4:90–94.
57. Chan AW, et al. Carotid brachytherapy for in-stent restenosis. Catheter Cardiovasc Interv 2003;58:86–92.
58. Saito T, Waki K, Becker AE. Left atrial myocardial extension onto pulmonary veins in humans: Anatomic observations relevant for atrial arrhythmias. J Cardiovasc Electrophysiol 2000;11:888–894.
59. Degertekin M, et al. Sirolimus-eluting stent for treatment of complex in-stent restenosis: The first clinical experience. J Am Coll Cardiol 2003;41:184–189.
60. Morice MC, et al. A randomized comparison of a sirolimus-eluting stent with a standard stent for coronary revascularization. N Engl J Med 2002;346:1773–1780.
61. Abizaid A, et al. Sirolimus-eluting stents inhibit neointimal hyperplasia in diabetic patients. Insights from the ravel trial. Eur Heart J 2004;25:107–112.

62. Serruys PW, et al. Intravascular ultrasound findings in the multicenter, randomized, double-blind ravel (randomized study with the sirolimus-eluting velocity balloon-expandable stent in the treatment of patients with de novo native coronary artery lesions) trial. Circulation 2002;106:798–803.
63. Bailey SR. Drug-eluting stents have made brachytherapy obsolete. Curr Opin Cardiol 2004;19:598–600.
64. Kaluza GL, Raizner AE. Brachytherapy for restenosis after stenting for coronary artery disease: Its role in the drug-eluting stent era. Curr Opin Cardiol 2004;19:601–607.

Reading List

Waksman R, Serruys P. Handbook of Vascular Bracdhytherapy. 2nd ed. London: Martin Dunitz Ltd; 2000.
Waksman R. Vascular Brachytherapy. 2nd ed. Armonk (NY): Futura Publishing Company; 1999.
Hall EJ. Radiobiology for the Radiologist. 5th ed. Philadelphia: J.B. Lippincott; 2000.
Kutryk MJ, Serruys PW. Coronary Stenting Current Perspective. London: Martin Dunitz Ltd; 1999.
Balter S, Chan RC, Shope TB. Intravascular Brachytherapy, Fluoroscopically Guided Interventions. Madison: Medical Physics Publishing; 2002.
Leon MB, Mintz GS. Interventional Vascular Product Guide. London: Martin Dunitz Ltd; 1999.
Attix FH. Introduction to Radiological Physics and Radiation Dosimetry. New York: John Wiley & Sons, Inc.; 1986.

See also BRACHYTHERAPY, HIGH DOSAGE RATE; CORONARY ANGIOPLASTY AND GUIDEWIRE DIAGNOSTICS.

BRAIN ELECTRICAL ACTIVITY. See ELECTROENCEPHALOGRAPHY.

BURN WOUND COVERINGS. See SKIN SUBSTITUTE FOR BURNS, BIOACTIVE.

BYPASS, CORONARY. See VASCULAR GRAFT PROSTHESIS.

BYPASS, CARDIOPULMONARY. See HEART-LUNG MACHINES.

ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition

VOLUME 2

Capacitive Microsensors for Biomedical Applications – Drug Infusion Systems

ENCYCLOPEDIA OF MEDICAL DEVICES AND INSTRUMENTATION, SECOND EDITION

ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition
Volume 2

Capacitive Microsensors for Biomedical Applications – Drug Infusion Systems

Edited by

John G. Webster
University of Wisconsin–Madison

The *Encyclopedia of Medical Devices and Instrumentation* is available online at http://www.mrw.interscience.wiley.com/emdi

WILEY-INTERSCIENCE

A John Wiley & Sons, Inc., Publication

Published by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Encylopedia of medical devices & instrumentation/by John G. Webster,

editor in chief. – 2nd ed.
p. ; cm.
Rev. ed. of: Encyclopedia of medical devices and instrumentation. 1988.
Includes bibliographical references and index.
ISBN-13 978-0-471-26358-6 (set : cloth)
ISBN-10 0-471-26358-3 (set : cloth)
ISBN-13 978-0-470-04067-6 (v. 2 : cloth)
ISBN-10 0-470-04067-x (v. 2 : cloth)

1. Medical instruments and apparatus–Encyclopedias. 2. Biomedical engineering–Encyclopedias. 3. Medical physics–Encyclopedias. 4. Medicine–Data processing–Encyclopedias. I. Webster, John G., 1932- . II. Title: Encyclopedia of medical devices and instrumentation.

[DNLM: 1. Equipment and Supplies–Encyclopedias–English. W 13

E555 2006]
R856.A3E53 2006
610.2803–dc22
2005028946

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTRIBUTOR LIST

ABDEL HADY, MAZEN, *McMaster University, Hamilton, Ontario Canada*, Bladder Dysfunction, Neurostimulation of

ABEL, L.A., *University of Melbourne, Melbourne, Australia*, Ocular Motility Recording and Nystagmus

ABREU, BEATRIZ C., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

ALEXANDER, A.L., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

ALI, ABBAS, *University of Illinois, at Urbana-Champaign*, Bioinformatics

ALI, MÜFTÜ, *School of Dental Medicine, Boston, Massachusetts*, Tooth and Jaw, Biomechanics of

ALPERIN, NOAM, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

ANSON, DENIS, *College Misericordia, Dallas, Pennsylvania*, Environmental Control

ARENA, JOHN C., *VA Medical Center and Medical College of Georgia*, Biofeedback

ARIEL, GIDEON, *Ariel Dynamics, Canyon, California*, Biomechanics of Exercise Fitness

ARMSTRONG, STEVE, *University of Iowa, Iowa City, Iowa*, Biomaterials for Dentistry

ASPDEN, R.M., *University of Aberdeen, Aberdeen, United Kingdom*, Ligament and Tendon, Properties of

AUBIN, C.E., *Polytechniquie Montreal, Montreal Quebec, Canada*, Scoliosis, Biomechanics of

AYRES, VIRGINIA M., *Michigan State University, East Lansing, Michigan*, Microscopy, Scanning Tunneling

AZANGWE, G., Ligament and Tendon, Properties of

BACK, LLOYD H., *California Institute of Technology, Pasadena, California*, Coronary Angioplasty and Guidewire Diagnostics

BADYLAK, STEPHEN F., *McGowan Institute for Regenerative Medicine, Pittsburgh, Pennsylvania*, Sterilization of Biologic Scaffold Materials

BANDYOPADHYAY, AMIT, *Washington State University, Pullman, Washington*, Orthopedic Devices, Materials and Design for

BANERJEE, RUPAK K., *University of Cincinnati, Cincinnati, Ohio*, Coronary Angioplasty and Guidewire Diagnostics

BARBOUR, RANDALL L., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

BARKER, STEVEN J., *University of Arizona, Tucson, Arizona*, Oxygen Monitoring

BARTH, ROLF F., *The Ohio State University, Columbus, Ohio*, Boron Neutron Capture Therapy

BECCHETTI, F.D., *University of Michigan, Ann Arbor, Michigan*, Radiotherapy, Heavy Ion

BELFORTE, GUIDO, *Politecnico di Torino – Department of Mechanics*, Laryngeal Prosthetic Devices

BENKESER, PAUL, *Georgia Institute of Technology, Atlanta, Georgia*, Biomedical Engineering Education

BENNETT, JAMES R., *University of Iowa, Iowa City, Iowa*, Digital Angiography

BERSANO-BEGEY, TOMMASO, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

BIGGS, PETER J., *Harvard Medical School, Boston, Massachusetts*, Radiotherapy, Intraoperative

BIYANI, ASHOK, *University of Toledo, and Medical College of Ohio, Toledo, Ohio*, Human Spine, Biomechanics of

BLOCK, W.F., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

BLUE, THOMAS E., *The Ohio State University, Columbus, Ohio*, Boron Neutron Capture Therapy

BLUMSACK, JUDITH T., *Disorders Auburn University, Auburn, Alabama*, Audiometry

BOGAN, RICHARD K., *University of South Carolina, Columbia, South Carolina*, Sleep Laboratory

BOKROS, JACK C., *Medical Carbon Research Institute, Austin, Texas*, Biomaterials, Carbon

BONGIOANNINI, GUIDO, *ENT Division Mauriziano Hospital, Torino, Italy*, Laryngeal Prosthetic Devices

BORAH, JOSHUA, *Applied Science Laboratories, Bedford, Massachusetts*, Eye Movement, Measurement Techniques for

BORDEN, MARK, *Director of Biomaterials Research, Irvine, California*, Biomaterials, Absorbable

BORTON, BETTIE B., *Auburn University Montgomery, Montgomery, Alabama*, Audiometry

BORTON, THOMAS E., *Auburn University Montgomery, Montgomery, Alabama*, Audiometry

BOSE SUSMITA,, *Washington State University, Pullman, Washington*, Orthopedic Devices, Materials and Design for

BOVA, FRANK J., *M. D. Anderson Cancer Center Orlando, Orlando, FL*, Radiosurgery, Stereotactic

BRENNER, DAVID J., *Columbia University Medical Center, New York, New York*, Computed Tomography Screening

BREWER, JOHN M., *University of Georgia*, Electrophoresis

BRIAN, L. DAVIS, *Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, Ohio*, Skin, Biomechanics of

BRITT, L.D., *Eastern Virginia Medical School, Norfolk, Virginia*, Gastrointestinal Hemorrhage

BRITT, R.C., *Eastern Virginia Medical School, Norfolk, Virginia*, Gastrointestinal Hemorrhage

BROZIK, SUSAN M., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

BRUNER, JOSEPH P., *Vanderbilt University Medical Center, Nashville, Tennessee*, Intrauterine Surgical Techniques

BRUNSWIG NEWRING, KIRK A., *University of Nevada, Reno, Nevada*, Sexual Instrumentatio n

BRUYANT, PHILIPPE P., *University of Massachusetts, North Worcester, Massachusetts*, Nuclear Medicine, Computers in

BUNNELL, BERT J., *Bunnell Inc., Salt Lake City, Utah*, High Frequency Ventilation

CALKINS, JERRY M., *Defense Research Technologies, Inc., Rockville, Maryland*, Medical Gas Analyzers

CANNON, MARK, *Northwestern University, Chicago, Illinois*, Resin-Based Composites

CAPPELLERI, JOSEPH C., *Pfizer Inc., Groton, Connecticut*, Quality-of-Life Measures, Clinical Significance of

CARDOSO, JORGE, *University of Madeira, Funchal, Portugal*, Office Automation Systems

CARELLO, MASSIMILIANA, *Politecnicodi Torino – Department of Mechanics*, Laryngeal Prosthetic Devices

CASKEY, THOMAS C., *Cogene Biotech Ventures, Houston, Texas*, Polymerase Chain Reaction

CECCIO, STEVEN, *University of Michigan, Ann Arbor, Michigan*, Heart Valve Prostheses, In Vitro Flow Dynamics of

CHAN, JACKIE K., *Columbia University, New York, New York*, Photography, Medical

CHANDRAN, K.B., *University of Iowa, Iowa City, Iowa*, Heart Valve Prostheses

CHATZANDROULIS, S., *NTUA, Athens, Attiki, Greece*, Capacitive Microsensors for Biomedical Applications

CHAVEZ, ELIANA, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CHEN, HENRY, *Stanford University, Palo Alto, California*, Exercise Stress Testing

CHEN, JIANDE, *University of Texas Medical Branch, Galveston, Texas*, Electrogastrogram

CHEN, YAN, *Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, Ohio*, Skin, Biomechanics of

CHEYNE, DOUGLAS, *Hospital for Sick Children Research Institute*, Biomagnetism

CHUI, CHEN-SHOU, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

CLAXTON, NATHAN S., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

CODERRE, JEFFREY A., *Massachus etts Institute of Technology, Cambridge, Massachusetts*, Boron Neutron Capture Therapy

COLLINS, BETH, *University of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

COLLINS, DIANE, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CONSTANTINOU, C., *Columbia University Radiation Oncology, New York, New York*, Phantom Materials in Radiology

COOK, ALBERT, *University of Alberta*, *Edmonton, Alberta, Canada*, Communication Devices

COOPER, RORY, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CORK, RANDALL C., *Louisiana State University, Shreveport, Louisiana*, Monitoring, Umbilical Artery and Vein, Blood Gas Measurements; Transcuta neous Electrical Nerve Stimulation (TENS); Ambulatory Monitoring

COX, JOSEPHINE H., *Walter Reed Army Institute of Research, Rockville, Maryland*, Blood Collection and Processing

CRAIG, LEONARD, *Feinberg School of Medicine of Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

CRESS, CYNTHIA J., *University of Nebraska, Lincoln, Nebraska*, Communicative Disorders, Computer Applications for

CUMMING, DAVID R.S., *University of Glasgow, Glasgow, United Kingdom*, Ion-Sensitive Field-Effect Transistors

CUNNINGHAM, JOHN R., *Camrose, Alberta, Canada*, Cobalt 60 Units for Radiotherapy

D'ALESSANDRO, DAVID, *Montefiore Medical Center, Bronx, New York*, Heart–Lung Machines

D'AMBRA, MICHAEL N., *Harvard Medical School, Cambridge, Massachusetts*, Cardiac Output, Thermodilution Measurement of

DADSETAN, MAHROKH, *Mayo Clinic, College of Medicine, Rochester, Minnesota*, Microscopy, Electron

DALEY, MICHAEL L., *The University of Memphis, Memphis, Tennessee*, Monitoring, Intracranial Pressure

DAN, LOYD, *Linköping University, Linköping, Sweden*, Thermocouples

DAS, RUPAK, *University of Wisconsin, Madison, Wisconsin*, Brachytherapy, High Dosage Rate

DATTAWADKAR, AMRUTA M., *University of Wisconsin, Madison, Madison, Wisconsin*, Ocular Fundus Reflectometry

DAVIDSON, MICHAEL W., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

DE LUCA, CARLO, *Boston University, Boston, Massachusetts*, Electromyography

DE SALLES, ANTONIO A.F., *UCLA Medical School, Los Angeles, California*, Stereotactic Surgery

DECAU, SABIN, *University of Maryland, School of Medicine*, Shock, Treatment of

DECHOW, PAUL C., *A & M University Health Science Center, Dallas, Texas*, Strain Gages

DELBEKE, JEAN, *Catholique University of Louvain, Brussels, Belgium*, Visual Prostheses

DELL'OSSO, LOUIS F., *Case Western Reserve University, Cleveland, Ohio*, Ocular Motility Recording and Nystagmus

DELORME, ARNAUD, *University of San Diego, La Jolla, California*, Statistical Methods

DEMENKOFF, JOHN, *Mayo Clinic, Scottsdale, Arizona*, Pulmonary Physiology

DEMIR, SEMAHAT S., *The University of Memphis and The University of Tennessee Health Science Center, Memphis, Tennessee*, Electrophysiology

DEMLING, ROBERT H., *Harvard Medical School*, Skin Substitute for Burns, Bioactive

DENNIS, MICHAEL J., *Medical University of Ohio, Toledo, Ohio*, Computed Tomography

DESANTI, LESLIE, *Harvard Medical School*, Skin Substitute for Burns, Bioactive

DEUTSCH, STEVEN, *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

DEVINENI, TRISHUL, *Conemaugh Health System*, Biofeedback

DI BELLA EDWARD, V.R., *University of Utah*, Tracer Kinetics

DIAKIDES, NICHOLAS A., *Advanced Concepts Analysis, Inc., Falls Church, Virginia*, Thermography

DOLAN, PATRICIA L., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

DONOVAN, F.M., *University of South Alabama*, Cardiac Output, Indicator Dilution Measurement of

DOUGLAS, WILSON R., *Children's Hospital of Philadelphia, Philadelphia, Pennsylvania*, Intrauterine Surgical Techniques

DRAPER, CRISSA, *University of Nevada, Reno, Nevada*, Sexual Instrumentation

DRZEWIECKI, TADEUSZ M., *Defense Research Technologies, Inc., Rockville, Maryland*, Medical Gas Analyzers

DURFEE, W.K., *University of Minnesota, Minneapolis, Minnesota*, Rehabilitation and Muscle Testing

DYRO, JOSEPH F., *Setauket, New York*, Safety Program, Hospital

DYSON, MARY, *Herts, United Kingdom*, Heat and Cold, Therapeutic

ECKERLE, JOSEPH S., *SRI International, Menlo Park, California*, Tonometry, Arterial

EDWARDS, BENJAMIN, *University of Wisconsin-Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

EDWARDS, THAYNE L., *University of Washington, Seattle, Washington*, Chromatography

EKLUND, ANDERS, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

EL SOLH, ALI A., *Erie County Medical Center, Buffalo, New York*, Sleep Studies, Computer Analysis of

ELMAYERGI, NADER, *McMaster University, Hamilton, Ontario, Canada*, Bladder Dysfunction, Neurostimulation of

ELSHARYDAH, AHMAD, *Louisiana State University, Baton Rouge, Louisiana*, Ambulatory Monitoring; Monitoring, Umbilical Artery and Vein, Blood Gas Measurements

FADDY, STEVEN C., *St. Vincents Hospital, Sydney, Darlinghurst, Australia*, Cardiac Output, Fick Technique for

FAHEY, FREDERIC H., *Childrens Hospital Boston*, Computed Tomography, Single Photon Emission

FAIN, S.B., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

FELDMAN, JEFFREY, *Childrens Hospital of Philadelphia, Philadelphia, Pennsylvania*, Anesthesia Machines

FELLERS, THOMAS J., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

FERRARA, LISA, *Cleveland Clinic Foundation, Cleveland, Ohio*, Human Spine, Biomechanics of

FERRARI, MAURO, *The Ohio State University, Columbus, Ohio*, Drug Delivery Systems

FONTAINE, ARNOLD A., *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

FOUST, MILTON J., JR, *Medical University of South Carolina Psychiatry and Behavioral Sciences, Charleston, South Carolina*, Electroconvulsive Therapy

FRASCO, PETER, *Mayo Clinic Scottsdale, Scottsdale, Arizona*, Temperature Monitoring

FRAZIER, JAMES, *Louisiana State University, Baton Rouge, Louisiana*, Ambulatory Monitoring

FREIESLEBEN DE BLASIO, BIRGITTE, *University of Oslo, Oslo, Norway*, Impedance Spectroscopy

FRESTA, MASSIMO, *University of Catanzaro Magna Græcia, Germaneto (CZ), Italy*, Drug Delivery Systems

FREYTES, DONALD O., *McGowan Institute for Regenerative Medicine, Pittsburgh Pennsylvania*, Sterilization of Biologic Scaffold Materials

FROEHLICHER, VICTOR, *VA Medical Center, Palo Alto, California*, Exercise Stress Testing

FUNG, EDWARD K., *Columbia University, New York, New York*, Photography, Medical

GAGE, ANDREW A., *State University of New York at Buffalo, Buffalo, New York*, Cryosurgery

GAGLIO, PAUL J., *Columbia University College of Physicians and Surgeons*, Liver Transplantation

GARDNER, REED M., *LDS Hospital and Utah University, Salt Lake City, Utah*, Monitoring, Hemodynamic

GEJERMAN, GLEN, *Hackensack University Medical, Hackensack, New Jersey*, Radiation Therapy, Quality Assurance in

GEORGE, MARK S., *Medical University of South Carolina Psychiatry and Behavioral Sciences, Charleston, South Carolina*, Electroconvulsive Therapy

GHARIEB, R.R., *Infinite Biomedical Technologies, Baltimore, Maryland*, Neurological Monitors

GLASGOW, GLENN P., *Loyola University of Chicago, Maywood, Illinois*, Radiation Protection Instrumentation

GLASGOW, GLENN, *University of Wisconsin-Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

GOEL, VIJAY K., *University of Toledo, and Medical College of Ohio, Toledo, Ohio*, Human Spine, Biomechanics of

GOETSCH, STEVEN J., *San Diego Gamma Knife Center, La Jolla, California*, Gamma Knife

GOLDBERG, JAY R., *Marquette University Milwaukee, Wisconsin*, Minimally Invasive Surgery

GOLDBERG, ZELENNA, *Department of Radiation Oncology, Davis, California*, Ionizing Radiation, Biological Effects of

GOPALAKRISHNAKONE, P., *National University of Singapore, Singapore*, Immunologically Sensitive Field-Effect Transistors

GOPAS, JACOB, *Ben Gurion University of the Negev, Beer Sheva, Israel*, Monoclonal Antibodies

GORGULHO, ALESSANDRA, *UCLA Medical School, Los Angeles, California*, Stereotactic Surgery

GOUGH, DAVID A., *University of California, La Jolla, California*, Glucose Sensors

GOUSTOURIDIS, D., *NTUA, Athens, Attiki, Greece*, Capacitive Microsensors for Biomedical Applications

GRABER, HARRY L., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

GRAÇA, M., *Louisiana State University, Baton Rouge, Louisiana*, Boron Neutron Capture Therapy

GRANT, WALTER III, *Baylor College of Medicine, Houston, Texas*, Radiation Therapy, Intensity Modulated

GRAYDEN, EDWARD, *Mayo Health Center, Albertlea, Minnesota*, Cardiopulmonary Resuscitation

GREEN, JORDAN R., *University of Nebraska, Lincoln, Nebraska*, Communicative Disorders, Computer Applications for

HAEMMERICH, DIETER, *Medical University of South Carolina, Charleston, South Carolina*, Tissue Ablation

HAMAM, HABIB, *Université de Moncton, Moncton New Brunswick, Canada*, Lenses, Intraocular

HAMMOND, PAUL A., *University of Glasgow, Glasgow, United Kingdom*, Ion-Sensitive Field-Effect Transistors

HANLEY, JOSEPH, *Hackensack University Medical, Hackensack, New Jersey*, Radiation Therapy, Quality Assurance in

HARLEY, BRENDAN A., *Massachusetts Institute of Technology*, Skin Tissue Engineering for Regeneration

HARPER, JASON C., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

HASMAN, ARIE, *Maastricht, The Netherlands*, Medical Education, Computers in

HASSOUNA, MAGDY, *Toronto Western Hospital, Toronto, Canada*, Bladder Dysfunction, Neurostimulation of

HAYASHI, KOZABURO, *Okayama University of Science, Okayama, Japan*, Arteries, Elastic Properties of

HENCH, LARRY L., *Imperial College London, London, United Kingdom*, Biomaterials: Bioceramics

HETTRICK, DOUGLAS A., *Sr. Principal Scientist Medtronic, Inc., Minneapolis, Minnesota*, Bioimpedance in Cardiovascular Medicine

HIRSCH-KUCHMA, MELISSA, *University of Central Florida NanoScience Technology Center, Orlando, Florida*, Biosurface Engineering

HOLDER, GRAHAM E., *Moorfields Eye Hospital, London, United Kingdom*, Electroretinography

HOLMES, TIMOTHY, *St. Agnes Cancer Center, Baltimore, Maryland*, Tomotherapy

HONEYMAN-BUCK, JANICE C., *University of Florida, Gainesville, Florida*, Radiology Information Systems

HOOPER, BRETT A., *Areté Associates, Arlington, Virginia*, Endoscopes

HORN, BRUCE, *Kaiser Permanente, Los Angeles, California*, X-Rays Production of

HORNER, PATRICIA I., *Biomedical Engineering Society Landover, Maryland*, Medical Engineering Societies and Organizations

HOROWITZ, PAUL M., *University of Texas, San Antonio, Texas*, Fluorescence Measurements

HOU, XIAOLIN, *Risø National Laboratory, Roskilde, Denmark*, Neutron Activation Analysis

HOVORKA, ROMAN, *University of Cambridge, Cambridge, United Kingdom*, Pancreas, Artificial

HUANG, H.K., *University of Southern California*, Teleradiology

HUNT, ALAN J., *University of Michigan, Ann Arbor, Michigan*, Optical Tweezers

HUTTEN, HELMUT, *University of Technology, Graz, Australia*, Impedance Plethysmography

IAIZZO, P.A., *University of Minnesota, Minneapolis, Minnesota*, Rehabilitation and Muscle Testing

IBBOTT, GEOFFREY S., *Anderson Cancer Center, Houston, Texas*, Radiation Dosimetry, Three-Dimensional

INGHAM, E., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

ISIK, CAN, *Syracuse University, Syracuse, New York*, Blood Pressure Measurement

JAMES, SUSAN P., *Colorado State University, Fort Collins, Colorado*, Biomaterials: Polymers

JENSEN, WINNIE, *Aalborg University, Aalborg, Denmark*, Electroneurography

JIN, CHUNMING, *North Carolina State University, Raleigh, North Carolina*, Biomaterials, Corrosion and Wear of

JIN, Z.M., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

JOHNSON, ARTHUR T., *University of Maryland College Park, Maryland*, Medical Engineering Societies and Organizations

JONES, JULIAN R., *Imperial College London, London, United Kingdom*, Biomaterials: Bioceramics

JOSHI, ABHIJEET, *Abbott Spine, Austin, Texas*, Spinal Implants

JUNG, RANU, *Arizona State University, Tempe, Arizona*, Functional Electrical Stimulation

JURISSON, SILVIA S., *University of Missouri Columbia, Missouri*, Radionuclide Production and Radioactive Decay

KAEDING, PATRICIA J., *Godfrey & Kahn S.C., Madison, Wisconsin*, Codes and Regulations: Medical Devices

KAMATH, CELIA C., *Mayo Clinic, Rochester, Minnesota*, Quality-of-Life Measures, Clinical Significance of

KANE, MOLLIE, *Madison, Wisconsin*, Contraceptive Devices

KATHERINE, ANDRIOLE P., *Harvard Medical School, Boston, Massachusetts*, Picture Archiving and Communication Systems

KATSAGGELOS, AGGELOS K., *Northwestern University, Evanston, Illinois*, DNA Sequencing

KATZ, J. LAWRENCE, *University of Missouri-Kansas City, Kansas City, Missouri*, Bone and Teeth, Properties of

KESAVAN, SUNIL, *Akebono Corporation, Farmington Hills, Michigan*, Linear Variable Differential Transformers

KHANG, GILSON, *Chonbuk National University*, Biomaterials: Tissue Engineering and Scaffolds

KHAODHIAR, LALITA, *Harvard Medical School, Boston, Massachusetts*, Cutaneous Blood Flow, Doppler Measurement of

KIM, MOON SUK, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

KIM, YOUNG KON, *Inje University, Kimhae City, Korea*, Alloys, Shape Memory

KINDWALL, ERIC P., *St. Luke's Medical Center, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

KING, MICHAEL A., *University of Massachusetts, North Worcester, Massachusetts*, Nuclear Medicine, Computers in

KLEBE, ROBERT J., *University of Texas, San Antonio, Texas*, Fluorescence Measurements

KLEIN, BURTON, *Burton Klein Associates, Newton, Massachusetts*, Gas and Vacuum Systems, Centrally Piped Medical

KNOPER, STEVEN R., *University of Arizona College of Medicine*, Ventilatory Monitoring

KONTAXAKIS, GEORGE, *Universidad Politécnica de Madrid, Madrid, Spain*, Positron Emission Tomography

KOTTKE-MARCHANT, KANDICE, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Vascular Graft Prosthesis

KRIPFGANS, OLIVER, *University of Michigan, Ann Arbor, Michigan*, Ultrasonic Imaging

KULKARNI, AMOL D., *University of Wisconsin–Madison, Madison, Wisconsin*, Ocular Fundus Reflectometry, Visual Field Testing

KUMARADAS, J. CARL, *Ryerson University, Toronto, Ontario, Canada*, Hyperthermia, Interstitial

KUNICKA, JOLANTA, *Bayer HealthCare LLC, Tarrytown, New York*, Differential Counts, Automated

KWAK, KWANJ JOO, *University of Miami Miller School of Medicine, Miami, Florida*, Microscopy, Scanning Force

LAKES, RODERIC, *University of Wisconsin-Madison*, Bone and Teeth, Properties of

LAKKIREDDY, DHANUNJAYA, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

LARSEN, COBY, *Case Western Reserve University, Cleveland, Ohio*, Vascular Graft Prosthesis

LASTER, BRENDA H., *Ben Gurion University of the Negev, Beer Sheva, Israel*, Monoclonal Antibodies

LATTA, LOREN, *University of Miami, Coral Gables, Florida*, Rehabilitation, Orthotics in

LEDER, RON S., *Universidad Nacional Autonoma de Mexico Mexico, Distrito Federal*, Continuous Positive Airway Pressure

LEE, CHIN, *Harvard Medical School, Boston, Massachusetts*, Radiotherapy Treatment Planning, Optimization of; Hyperthermia, Interstitial

LEE, HAI BANG, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

LEE, SANG JIN, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

LEI, LIU, *Department of General Engineering, Urbana, Illinois*, Bioinformatics

LEI, XING, *Stanford University, Stanford, California*, Radiation Dose Planning, Computer-Aided

LEWIS, MATTHEW C., *Medical College of Wisconsin, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

LI, CHAODI, *University of Notre Dame, Notre Dame, Indiana*, Bone Cement, Acrylic

LI, JONATHAN G., *University of Florida, Gainesville, Florida*, Radiation Dose Planning, Computer-Aided

LI, QIAO, *University of Michigan, Ann Arbor, Michigan*, Immunotherapy

LI, YANBIN, *University of Arkansas, Fayetteville, Arkansas*, Piezoelectric Sensors

LIBOFF, A.R., *Oakland University, Rochester, Michigan*, Bone Ununited Fracture and Spinal Fusion, Electrical Treatment of

LIGAS, JAMES, *University of Connecticut, Farmington, Connecticut*, Respiratory Mechanics and Gas Exchange

LIMOGE, AIME, *The René Descartes University of Paris, Paris, France*, Electroanalgesia, Systemic

LIN, PEI-JAN PAUL, *Beth Israel Deaconess Medical Center, Boston, Massachusets*, Mammography

LIN, ZHIYUE, *University of Kansas Medical Center, Kansas City, Kansas*, Electrogastrogram

LINEAWEAVER, WILLIAM C., *Unive rsity of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

LIPPING, TARMO, *Tampere University of Technology, Pori, Finland*, Monitoring in Anesthesia

LIU, XIAOHUA, *The University of Michigan, Ann Arbor, Michigan*, Polymeric Materials

LLOYD, J.J., *Regional Medical Physics Department, Newcastle-upon-Tyne, United Kingdom*, Ultraviolet Radiation in Medicine

LOEB, ROBERT, *University of Arizona, Tuscon, Arizona*, Anesthesia Machines

LOPES DE MELO, PEDRO, *State University of Rio de Janeiro, Térreo Salas, Maracanã, Thermistors*

LOUDON, ROBERT G., Lung Sounds

LOW, DANIEL A., *Washington University School of Medicine, St. Louis, Missouri*, Radiation Therapy Simulator

LU, LICHUN, *Mayo Clinic, College of Medicine, Rochester, Minnesota*, Microscopy, Electron

LU, ZHENG FENG, *Columbia University, New York, New York*, Screen-Film Systems

LYON, ANDREW W., *University of Calgary, Calgary, Canada*, Flame Atomic Emission Spectrometry and Atomic Absorption Spectrometry

LYON, MARTHA E., *University of Calgary, Calgary, Canada*, Flame Atomic Emission Spectrometry and Atomic Absorption Spectrometry

MA, C-M CHARLIE, *Fox Chase Cancer Center, Philadelphia, Pennsylvania*, X-Ray Therapy Equipment, Low and Medium Energy

MACIA, NARCISO F., *Arizona State University at the Polytechnic Campus, Mesa, Arizona*, Pneumotachometers

MACKENZIE, COLIN F., *University of Maryland, School of Medicine*, Shock, Treatment of

MACKIE, THOMAS R., *University of Wisconsin, Madison, Wisconsin*, Tomotherapy

MADNANI, ANJU, *LSU Medical Centre, Shreveport, Louisiana*, Transcutaneous Electrical Nerve Stimulation (TENS)

MADNANI, SANJAY, *LSU Medical Centre, Shreveport, Louisiana*, Transcutaneous Electrical Nerve Stimulation (TENS)

MADSEN, MARK T., *University of Iowa, Iowa City, Iowa*, Anger Camera

MAGNANO, MAURO, *ENT Division Mauriziano Hospital, Torino, Italy*, Drug Delivery Systems

MANDEL, RICHARD, *Boston University School of Medicine, Boston, Massachusetts*, Colorimetry

MANNING, KEEFE B., *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

MAO, JEREMY J., *University of Illinois at Chicago, Chicago, Illinois*, Cartilage and Meniscus, Properties of

MARCOLONGO, MICHELE, *Drexel University, Philadelphia, Pennsylvania*, Spinal Implants

MAREK, MIROSLAV, *Georgia Institute of Technology, Atlanta, Georgia*, Biomaterials, Corrosion and Wear of

MARION, NICHOLAS W., *University of Illinois at Chicago, Chicago, Illinois*, Cartilage and Meniscus, Properties of

MASTERS, KRISTYN S., *University of Wisconsin, Madison, Wisconsin*, Tissue Engineering

MAUGHAN, RICHARD L., *Hospital of the University of Pennsylvania*, Neutron Beam Therapy

MCADAMS, ERIC, *University of Ulster at Jordanstown, Newtownabbey, Ireland*, Bioelectrodes

MCARTHUR, SALLY L., *University of Sheffield, Sheffield, United Kingdom*, Biomaterials, Surface Properties of

MCEWEN, MALCOM, *National Research Council of Canada, Ontario, Canada*, Radiation Dosimetry for Oncology

MCGOWAN, EDWARD J., *E.J. McGowan & Associates*, Biofeedback

MCGRATH, SUSAN, *Dartmouth College, Hanover, New Hampshire*, Oxygen Analyzers

MEEKS, SANFORD L., *University of Florida, Gainesville, Florida*, Radiosurgery, Stereotactic

MELISSA, PETER, *University of Central Florida NanoScience Technology Center, Orlando, Florida*, Biosurface Engineering

MENDELSON, YITZHAK, *Worcester Polytechnic Institute*, Optical Sensors

METZKER, MICHAEL L., *Baylor College of Medicine, Houston, Texas*, Polymerase Chain Reaction

MEYEREND, M.E., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

MICHLER, ROBERT, *Montefiore Medical Center, Bronx, New York*, Heart–Lung Machines

MICIC, MIODRAG, *MP Biomedicals LLC, Irvine, California*, Microscopy and Spectroscopy, Near-Field

MILLER, WILLIAM, *University of Missouri Columbia, Missouri*, Radionuclide Production and Radioactive Decay

MITTRA, ERIK, *Stony Brook University, New York*, Bone Density Measurement

MODELL, MARK, *Harvard Medical School, Boston, Massachusetts*, Fiber Optics in Medicine

MORE, ROBERT B., *RBMore Associates, Austin, Texas* Biomaterials Carbon

MORE, ROBERT, *Austin, Texas*, Heart Valves, Prosthetic

MORROW, DARREN, *Royal Adelaide Hospital, Adelaide, Australia*, Intraaortic Balloon Pump

MOURTADA, FIRAS, *MD Anderson Cancer Center, Houston, Texas*, Brachytherapy, Intravascular

MOY, VINCENT T., *University of Miami, Miller School of Medicine, Miami, Florida*, Microscopy, Scanning Force

MÜFTÜ, SINAN, *Northeastern University, Boston, Massachusetts*, Tooth and Jaw, Biomechanics of

MURPHY, RAYMOND L.H., Lung Sounds

MURPHY, WILLIAM L., *University of Wisconsin, Madison, Wisconsin*, Tissue Engineering

MURRAY, ALAN, *Newcastle University Medical Physics, Newcastle upon Tyne, United Kingdom*, Pace makers

MUTIC, SASA, *Washington University School of Medicine, St. Louis, Missouri*, Radiation Therapy Simulator

NARAYAN, ROGER J., *University of North Carolina, Chapel Hill, North Carolina*, Biomaterials, Corrosion and Wear of

NATALE, ANDREA, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

NAZERAN, HOMER, *The University of Texas, El Paso, Texas*, Electrocardiography, Computers in

NEUMAN, MICHAEL R., *Michigan Technological University, Houghton, Houghton, Michigan*, Fetal Monitoring, Neonatal Monitoring

NEUZIL, PAVEL, *Institute of Bioengineering and Nanotechnology, Singapore*, Immunologically Sensitive Field-Effect Transistors

NICKOLOFF, EDWARD L., *Columbia University, New York, New York*, X-Ray Quality Control Program

NIEZGODA, JEFFREY A., *Medical College of Wisconsin, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

NISHIKAWA, ROBERT M., *The University of Chicago, Chicago, Illinois*, Computer-Assisted Detection and Diagnosis

NUTTER, BRIAN, *Texas Tech University, Lubbock, Texas*, Medical Records, Computers in

O'DONOHUE, WILLIAM, *University of Nevada, Reno, Nevada*, Sexual Instrumentation

ORTON, COLIN, *Harper Hospital and Wayne State University, Detroit, Michigan*, Medical Physics Literature

OZCELIK, SELAHATTIN, *Texas A&M University, Kingsville, Texas*, Drug Infusion Systems

PANITCH, ALYSSA, *Arizona State University, Tempe, Arizona*, Biomaterials: An Overview

PAOLINO, DONATELLA, *University of Catanzaro Magna Græcia, Germaneto (CZ), Italy*, Drug Delivery Systems

PAPAIOANNOU, GEORGE, *University of Wisconsin, Milwaukee, Wisconsin*, Joints, Biomechanics of

PARK, GRACE E., *Purdue University, West Lafayette, Indiana*, Porous Materials for Biological Applications

PARMENTER, BRETT A., *State University of New York at Buffalo, Buffalo, New York*, Sleep Studies, Computer Analysis of

PATEL, DIMPI, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

PEARCE, JOHN, *The University of Texas, Austin, Texas*, Electrosurgical Unit (ESU)

PELET, SERGE, *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

PERIASAMY, AMMASI, *University of Virginia, Charlottesville, Virginia*, Cellular Imaging

PERSONS, BARBARA L., *University of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

PIPER, IAN, *The University of Memphis, Memphis, Tennessee*, Monitoring, Intracranial Pressure

POLETTO, CHRISTOPHER J., *National Institutes of Health*, Tactile Stimulation

PREMINGER, GLENN M., *Duke University Medical Center, Durham, North Carolina*, Lithotripsy

PRENDERGAST, PATRICK J., *Trinity College, Dublin, Ireland*, Orthopedics, Prosthesis Fixation for

PREVITE, MICHAEL, *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

PURDY, JAMES A., *UC Davis Medical Center, Sacramento, California*, Radiotherapy Accessories

QI, HAIRONG, *Advanced Concepts Analysis, Inc., Falls Church, Virginia*, Thermography

QIN, YIXIAN, *Stony Brook University, New York*, Bone Density Measurement

QUAN, STUART F., *University of Arizona, Tucson, Arizona*, Ventilatory Monitoring

QUIROGA, RODRIGO QUIAN, *University of Leicester, Leicester, United Kingdom*, Evoked Potentials

RAHAGHI, FARBOD N., *University of California, La Jolla, California*, Glucose Sensors

RAHKO, PETER S., *University of Wisconsin Medical School*, Echocardiography and Doppler Echocardiography

RALPH, LIETO, *University of Wisconsin–Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

RAMANATHAN, LAKSHMI, *Mount Sinai Medical Center*, Analytical Methods, Automated

RAO, SATISH S.C., *University of Iowa College of Medicine, Iowa City, Iowa*, Anorectal Manometry

RAPOPORT, DAVID M., *NYU School of Medicine, New York, New York*, Continuous Positive Airway Pressure

REBELLO, KEITH J., *The Johns Hopkins University Applied Physics Lab, Laurel, Maryland*, Micro surgery

REDDY, NARENDER, *The University of Akron, Akron, Ohio*, Linear Variable Differential Transformers

REN-DIH, SHEU, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

RENGACHARY, SETTI S., *Detroit, Michigan*, Human Spine, Biomechanics of

REPPERGER, DANIEL W., *Wright-Patterson Air Force Base, Dayton, Ohio*, Human Factors in Medical Devices

RITCHEY, ERIC R., *The Ohio State University, Columbus, Ohio*, Contact Lenses

RIVARD, MARK J., *Tufts New England Medical Center, Boston, Massachusetts*, Imaging Devices

ROBERTSON, J. DAVID, *University of Missouri, Columbia, Missouri*, Radionuclide Production and Radioactive Decay

ROTH, BRADLEY J., *Oakland University, Rochester, Michigan*, Defibrillators

ROWE-HORWEGE, R. WANDA, *University of Texas Medical School, Houston, Texas*, Hyperthermia, Systemic

RUMSEY, JOHN W., *University of Central Florida, Orlando, Florida*, Biosurface Engineering

RUTKOWSKI, GREGORY E., *University of Minnesota, Duluth, Minnesota*, Engineered Tissue

SALATA, O.V., *University of Oxford, Oxford, United Kingdom*, Nanoparticles

SAMARAS, THEODOROS, *Aristotle University of Thessaloniki Department of Physics, Thessaloniki, Greece*, Thermometry

SANGOLE, ARCHANA P., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

SARKOZI, LASZLO, *Mount Sinai School of Medicine*, Analytical Methods, Automated

SCHEK, HENRY III, *University of Michigan, Ann Arbor, Michigan*, Optical Tweezers

SCHMITZ, CHRISTOPH H., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

SCHUCKERS, STEPHANIE A.C., *Clarkson University, Potsdam, New York*, Arrhythmia Analysis, Automated

SCOPE, KENNETH, *Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

SCOTT, ADZICK N., *University of Pennsylvania, Philadelphia, Pennsylvania*, Intrauterine Surgical Techniques

SEAL, BRANDON L., *Arizona State University, Tempe, Arizona*, Biomaterials: An Overview

SEALE, GARY, *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

SEGERS, PATRICK, *Ghent University, Belgium*, Hemodynamics

SELIM, MOSTAFA A., *Cleveland Metropolitan General Hospital, Palm Coast, Florida*, Colposcopy

SETHI, ANIL, *Loyola University Medical Center, Maywood, Illinois*, X-Rays: Interaction with Matter

SEVERINGHAUS, JOHN W., *University of California in San Francisco*, CO_2 Electrodes

SHALODI, ABDELWAHAB D., *Cleveland Metropolitan General Hospital, Palm Coast, Florida*, Colposcopy

SHANMUGASUNDARAM, SHOBANA, *New Jersey Institute of Technology, Newark, New Jersey*, Polymeric Materials

SHARD, ALEXANDER G., *University of Sheffield, Sheffield United Kingdom*, Biomaterials, Surface Properties of

SHEN, LI-JIUAN, *National Taiwan University School of Pharmacy, Taipei, Taiwan*, Colorimetry

SHEN, WEI-CHIANG, *University of Southern California School of Pharmacy, Los Angeles, California*, Colorimetry

SHERAR, MICHAEL D., *London Health Sciences Centre and University of Western Ontario, London, Ontario, Canada*, Hyperthermia, Interstitial

SHERMAN, DAVID, *The Johns Hopkins University, Baltimore, Maryland*, Electroencephalography

SHI, DONGLU, *University of Cincinnati, Cincinnati, Ohio*, Biomaterials, Testing and Structural Properties of

SHUCARD, DAVID W.M., *State University of New York at Buffalo, Buffalo, New York*, Sleep Studies, Computer Analysis of

SIEDBAND, MELVIN P., *University of Wisconsin, Madison, Wisconsin*, Image Intensifiers and Fluoroscopy

SILBERMAN, HOWARD, *University of Southern California, Los Angeles, California*, Nutrition, Parenteral

SILVERMAN, GORDON, *Manhattan College*, Computers in the Biomedical Laboratory

SILVERN, DAVID A., *Medical Physics Unit, Rabin Medical Center, Petah Tikva, Israel*, Prostate Seed Implants

SINHA, PIYUSH, *The Ohio State University, Columbus, Ohio*, Drug Delivery Systems

SINHA, ABHIJIT ROY, *University of Cincinnati, Cincinnati, Ohio*, Coronary Angioplasty and Guidewire Diagnostics

SINKJÆR, THOMAS, *Aalborg University, Aalborg, Denmark*, Electroneurography

SLOAN, JEFFREY A., *Mayo Clinic, Rochester, Minnesota*, Quality-of-Life Measures, Clinical Significance of

SO, PETER T.C., *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

SOBOL, WLAD T., *University of Alabama at Birmingham Health System, Birmingham, Alabama*, Nuclear Magnetic Resonance Spectroscopy

SOOD, SANDEEP, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

SPECTOR, MYRON, *Brigham and Women's Hospital, Boston, Massachusetts*, Biocompatibility of Materials

SPELMAN, FRANCIS A., *University of Washington*, Cochlear Prostheses

SRINIVASAN, YESHWANTH, *Texas Tech University, Lubbock, Texas*, Medical Records, Computers in

SRIRAM, NEELAMEGHAM, *University of Buffalo, Buffalo, New York*, Cell Counters, Blood

STARKO, KENTON R., *Point Roberts, Washington*, Physiological Systems Modeling

STARKSCHALL, GEORGE, *The University of Texas*, Radiotherapy, Three-Dimensional Conformal

STAVREV, PAVEL, *Cross Cancer Institute, Edmonton, Alberta, Canada*, Radiotherapy Treatment Planning, Optimization of

STENKEN, JULIE A., *Rensselaer Polytechnic Institute, Troy, New York*, Microdialysis Sampling

STIEFEL, ROBERT, *University of Maryland Medical Center, Baltimore, Maryland*, Equipment Acquisition

STOKES, I.A.F., *Polytechniquie Montreal, Montreal Quebec, Canada*, Scoliosis, Biomechanics of

STONE, M.H., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

SU, XIAO-LI, *BioDetection Instruments LLC, Fayetteville, Arkansas*, Piezoelectric Sensors

SUBHAN, ARIF, *Masterplan Technology Management, Chatsworth, California*, Equipment Maintenance, Biomedical

SWEENEY, JAMES D., *Arizona State University, Tempe, Arizona*, Functional Electrical Stimulation

SZETO, ANDREW Y.J., *San Diego State University, San Diego, California*, Blind and Visually Impaired, Assistive Technology for

TAKAYAMA, SHUICHI, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

TAMUL, PAUL C., *Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

TAMURA, TOSHIYO, *Chiba University School of Engineering, Chiba, Japan*, Home Health Care Devices

TANG, XIANGYANG, *GE Healthcare Technologies, Wankesha, Wisconsin*, Computed Tomography Simulators

TAYLOR, B.C., *The University of Akron, Akron, Ohio*, Cardiac Output, Indicator Dilution Measurement of

TEMPLE, RICHARD O., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

TEN, STANLEY, *Salt Lake City, Utah*, Electroanalgesia, Systemic

TERRY, TERESA M., *Walter Reed Army Institute of Research, Rockville, Maryland*, Blood Collection and Processing

THAKOR, N.V., *Johns Hopkins University, Baltimore, Maryland*, Neurological Monitors

THIERENS, HUBERT M.A., *University of Ghent, Ghent, Belgium*, Radiopharmaceutical Dosimetry

THOMADSEN, BRUCE, *University of Wisconsin–Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

TIPPER, J.L., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

TOGAWA, TATSUO, *Waseda University, Saitama, Japan*, Integrated Circuit Temperature Sensor

TORNAI, MARTIN, *Duke University, Durham, North Carolina*, X-Ray Equipment Design

TRAN-SON-TAY, ROGER, *University of Florida, Gainesville, Florida*, Blood Rheology

YENI, YENER N., *Henry Ford Hospital, Detroit, Michigan*, Joints, Biomechanics of

YLI-HANKALA, ARVI, *Tampere University of Technology, Pori, Finland*, Monitoring in Anesthesia

YOKO, KAMOTANI, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

YOON, KANG JI, *Korea Institute of Science and Technology, Seoul, Korea*, Micropower for Medical Applications

YORKE, ELLEN, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

YOSHIDA, KEN, *Aalborg University, Aalborg, Denmark*, Electroneurography

YOUNGSTEDT, SHAWN D., *University of South Carolina, Columbia, South Carolina*, Sleep Laboratory

YU, YIH-CHOUNG, *Lafayette College, Easton, Pennsylvania*, Blood Pressure, Automatic Control of

ZACHARIAH, EMMANUEL S., *University of Medicine and Dentistry of New Jersey, New Brunswick, New Jersey*, Immunologically Sensitive Field-Effect Transistors

ZAIDER, MARCO, *Memorial Sloan Kettering Cancer Center, New York, New York*, Prostate Seed Implants

ZAPANTA, CONRAD M., *Penn State College of Medicine, Hershey, Pennsylvania*, Heart, Artificial

ZARDENETA, GUSTAVO, *University of Texas, San Antonio, Texas*, Fluorescence Measurements

ZELMANOVIC, DAVID, *Bayer HealthCare LLC, Tarrytown, New York*, Differential Counts, Automated

ZHANG, MIN, *University of Washington, Seattle, Washington*, Biomaterials: Polymers

ZHANG, YI, *University of Buffalo, Buffalo, New York*, Cell Counters, Blood

ZHU, XIAOYUE, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

ZIAIE, BABAK, *Purdue University, W. Lafayette, Indiana*, Biotelemetry

ZIELINSKI, TODD M., *Medtronic, Inc., Minneapolis, Minnesota*, Bioimpedance in Cardiovascular Medicine

ZIESSMAN, HARVEY A., *Johns Hopkins University*, Computed Tomography, Single Photon Emission

PREFACE

This six-volume work is an alphabetically organized compilation of almost 300 articles that describe critical aspects of medical devices and instrumentation.

It is comprehensive. The articles emphasize the contributions of engineering, physics, and computers to each of the general areas of anesthesiology, biomaterials, burns, cardiology, clinical chemistry, clinical engineering, communicative disorders, computers in medicine, critical care medicine, dermatology, dentistry, ear, nose, and throat, emergency medicine, endocrinology, gastroenterology, genetics, geriatrics, gynecology, hematology, heptology, internal medicine, medical physics, microbiology, nephrology, neurology, nutrition, obstetrics, oncology, ophthalmology, orthopedics, pain, pediatrics, peripheral vascular disease, pharmacology, physical therapy, psychiatry, pulmonary medicine, radiology, rehabilitation, surgery, tissue engineering, transducers, and urology.

The discipline is defined through the synthesis of the core knowledge from all the fields encompassed by the application of engineering, physics, and computers to problems in medicine. The articles focus not only on what is now useful but also on what is likely to be useful in future medical applications.

These volumes answer the question, "What are the branches of medicine and how does technology assist each of them?" rather than "What are the branches of technology and how could each be used in medicine?" To keep this work to a manageable length, the practice of medicine that is unassisted by devices, such as the use of drugs to treat disease, has been excluded.

The articles are accessible to the user; each benefits from brevity of condensation instead of what could easily have been a book-length work. The articles are designed not for peers, but rather for workers from related fields who wish to take a first look at what is important in the subject.

The articles are readable. They do not presume a detailed background in the subject, but are designed for any person with a scientific background and an interest in technology. Rather than attempting to teach the basics of physiology or Ohm's law, the articles build on such basic concepts to show how the worlds of life science and physical science meld to produce improved systems. While the ideal reader might be a person with a Master's degree in biomedical engineering or medical physics or an M.D. with a physical science undergraduate degree, much of the material will be of value to others with an interest in this growing field. High school students and hospital patients can skip over more technical areas and still gain much from the descriptive presentations.

The *Encyclopedia of Medical Devices and Instrumentation* is excellent for browsing and searching for those new divergent associations that may advance work in a peripheral field. While it can be used as a reference for facts, the articles are long enough that they can serve as an educational instrument and provide genuine understanding of a subject.

One can use this work just as one would use a dictionary, since the articles are arranged alphabetically by topic. Cross references assist the reader looking for subjects listed under slightly different names. The index at the end leads the reader to all articles containing pertinent information on any subject. Listed on pages xxi to xxx are all the abbreviations and acronyms used in the *Encyclopedia.* Because of the increasing use of SI units in all branches of science, these units are provided throughout the *Encyclopedia* articles as well as on pages xxxi to xxxv in the section on conversion factors and unit symbols.

I owe a great debt to the many people who have contributed to the creation of this work. At John Wiley & Sons, Encyclopedia Editor George Telecki provided the idea and guiding influence to launch the project. Sean Pidgeon was Editorial Director of the project. Assistant Editors Roseann Zappia, Sarah Harrington, and Surlan Murrell handled the myriad details of communication between publisher, editor, authors, and reviewers and stimulated authors and reviewers to meet necessary deadlines.

My own background has been in the electrical aspects of biomedical engineering. I was delighted to have the assistance of the editorial board to develop a comprehensive encyclopedia. David J. Beebe suggested cellular topics such as microfluidics. Jerry M. Calkins assisted in defining the chemically related subjects, such as anesthesiology. Michael R. Neuman suggested subjects related to sensors, such as in his own work—neonatology. Joon B. Park has written extensively on biomaterials and suggested related subjects. Edward S. Sternick provided many suggestions from medical physics. The Editorial Board was instrumental both in defining the list of subjects and in suggesting authors.

This second edition brings the field up to date. It is available on the web at http://www.mrw.interscience.wiley.com/emdi, where articles can be searched simultaneously to provide rapid and comprehensive information on all aspects of medical devices and instrumentation.

JOHN G. WEBSTER
University of Wisconsin, Madison

LIST OF ARTICLES

ABBREVIATIONS AND ACRONYMS

AAMI	Association for the Advancement of Medical Instrumentation
AAPM	American Association of Physicists in Medicine
ABC	Automatic brightness control
ABET	Accreditation board for engineering training
ABG	Arterial blood gases
ABLB	Alternative binaural loudness balance
ABS	Acrylonitrile–butadiene–styrene
ac	Alternating current
AC	Abdominal circumference; Affinity chromatography
ACA	Automated clinical analyzer
ACES	Augmentative communication evaluation system
ACL	Anterior chamber lens
ACLS	Advanced cardiac life support
ACOG	American College of Obstetrics and Gynecology
ACR	American College of Radiology
ACS	American Cancer Society; American College of Surgeons
A/D	Analog-to-digital
ADC	Agar diffusion chambers; Analog-to-digital converter
ADCC	Antibody-dependent cellular cytotoxicity
ADCL	Accredited Dosimetry Calibration Laboratories
ADP	Adenosine diphosphate
A-D-T	Admission, discharge, and transfer
AE	Anion exchange; Auxiliary electrode
AEA	Articulation error analysis
AEB	Activation energy barrier
AEC	Automatic exposure control
AED	Automatic external defibrillator
AEMB	Alliance for Engineering in Medicine and Biology
AES	Auger electron spectroscopy
AESC	American Engineering Standards Committee
AET	Automatic exposure termination
AFO	Ankle-foot orthosis
AGC	Automatic gain control
AHA	American Heart Association
AI	Arterial insufficiency
AICD	Automatic implantable cardiac defibrillator
AID	Agency for International Development
AIDS	Acquired immune deficiency syndrome
AL	Anterior leaflet
ALG	Antilymphocyte globulin
ALS	Advanced life support; Amyotropic lateral sclerosis
ALT	Alanine aminotransferase
ALU	Arithmetic and logic unit
AM	Amplitude modulation
AMA	American Medical Association
amu	Atomic mass units
ANOVA	Analysis of variance
ANSI	American National Standards Institute
AP	Action potential; Alternative pathway; Anteroposterior
APD	Anterioposterior diameter
APL	Adjustable pressure limiting valve; Applied Physics Laboratory
APR	Anatomically programmed radiography
AR	Amplitude reduction; Aortic regurgitation; Autoregressive
Ara-C	Arabinosylcytosine
ARD	Absorption rate density
ARDS	Adult respiratory distress syndrome
ARGUS	Arrhythmia guard system
ARMA	Autoregressive-moving-average model
ARMAX	Autoregressive-moving-average model with external inputs
AS	Aortic stenosis
ASA	American Standards Association
ASCII	American standard code for information interchange
ASD	Antisiphon device
ASHE	American Society for Hospital Engineering
ASTM	American Society for Testing and Materials
AT	Adenosine-thiamide; Anaerobic threshold; Antithrombin
ATA	Atmosphere absolute
ATLS	Advanced trauma life support
ATN	Acute tubular necrosis
ATP	Adenosine triphosphate
ATPD	Ambient temperature pressure dry
ATPS	Ambient temperature pressure saturated
ATR	Attenuated total reflection
AUC	Area under curve
AUMC	Area under moment curve
AV	Atrioventricular
AZT	Azido thymidine
BA	Biliary atresia
BAEP	Brainstem auditory evoked potential
BAPN	Beta-amino-proprionitryl
BAS	Boston anesthesis system
BASO	Basophil
BB	Buffer base
BBT	Basal body temperature

BCC	Body-centered cubic
BCD	Binary-coded decimal
BCG	Ballistocardiogram
BCLS	Basic cardiac life support
BCRU	British Commitee on Radiation Units and Measurements
BDI	Beck depression inventory
BE	Base excess; Binding energy
BET	Brunauer, Emmett, and Teller methods
BH	His bundle
BI	Biological indicators
BIH	Beth Israel Hospital
BIPM	International Bureau of Weights and Measurements
BJT	Bipolar junction transistor
BMDP	Biomedical Programs
BME	Biomedical engineering
BMET	Biomedical equipment technician
BMO	Biomechanically optimized
BMR	Basal metabolic rate
BOL	Beginning of life
BP	Bereitschafts potential; Break point
BR	Polybutadiene
BRM	Biological response modifier
BRS	Bibliographic retrieval services
BSS	Balanced salt solution
BTG	Beta thromboglobulin
BTPS	Body temperature pressure saturated
BUN	Blood urea nitrogen
BW	Body weight
CA	Conductive adhesives
CABG	Coronary artery by-pass grafting
CAD/CAM	Computer-aided design/computer-aided manufacturing
CAD/D	Computer-aided drafting and design
CADD	Central axis depth dose
CAI	Computer assisted instruction; Computer-aided instruction
CAM	Computer-assisted management
cAMP	Cyclic AMP
CAPD	Continuous ambulatory peritoneal dialysis
CAPP	Child amputee prosthetic project
CAT	Computerized axial tomography
CATS	Computer-assisted teaching system; Computerized aphasia treatment system
CAVH	Continuous arteriovenous hemofiltration
CB	Conjugated bilirubin; Coulomb barrier
CBC	Complete blood count
CBF	Cerebral blood flow
CBM	Computer-based management
CBV	Cerebral blood volume
CC	Closing capacity
CCC	Computer Curriculum Company
CCD	Charge-coupled device
CCE	Capacitance contact electrode
CCF	Cross-correlation function
CCL	Cardiac catheterization laboratory
CCM	Critical care medical services
CCPD	Continuous cycling peritoneal dialysis
CCTV	Closed circuit television system
CCU	Coronary care unit; Critical care unit
CD	Current density
CDR	Complimentary determining region
CDRH	Center for Devices and Radiological Health
CEA	Carcinoembryonic antigen
CF	Conversion factor; Cystic fibrosis
CFC	Continuous flow cytometer
CFR	Code of Federal Regulations
CFU	Colony forming units
CGA	Compressed Gas Association
CGPM	General Conference on Weights and Measures
CHO	Carbohydrate
CHO	Chinese hamster ovary
CI	Combination index
CICU	Cardiac intensive care unit
CIF	Contrast improvement factor
CIN	Cervical intraepithelial neoplasia
CK	Creatine kinase
CLAV	Clavicle
CLSA	Computerized language sample analysis
CM	Cardiomyopathy; Code modulation
CMAD	Computer managed articulation diagnosis
CMI	Computer-managed instruction
CMRR	Common mode rejection ratio
CMV	Conventional mechanical ventilation; Cytomegalovirus
CNS	Central nervous system
CNV	Contingent negative variation
CO	Carbon monoxide; Cardiac output
COBAS	Comprehensive Bio-Analysis System
COPD	Chronic obstructive pulmonary disease
COR	Center of rotation
CP	Cerebral palsy; Closing pressure; Creatine phosphate
CPB	Cardiopulmonary bypass
CPET	Cardiac pacemaker electrode tips
CPM	Computerized probe measurements
CPP	Cerebral perfusion pressure; Cryoprecipitated plasma
CPR	Cardiopulmonary resuscitation
cps	Cycles per second
CPU	Central Processing unit
CR	Center of resistance; Conditioned response; Conductive rubber; Creatinine
CRBB	Complete right bundle branch block
CRD	Completely randomized design
CRL	Crown rump length
CRT	Cathode ray tube
CS	Conditioned stimulus; Contrast scale; Crown seat
CSA	Compressed spectral array
CSF	Cerebrospinal fluid
CSI	Chemical shift imaging
CSM	Chemically sensitive membrane
CT	Computed tomography; Computerized tomography
CTI	Cumulative toxicity response index
CV	Closing volume

C.V.	Coefficient of variation
CVA	Cerebral vascular accident
CVP	Central venous pressure
CVR	Cardiovascular resistance
CW	Continuous wave
CWE	Coated wire electrodes
CWRU	Case Western Reserve University
DAC	Digital-to-analog converter
DAS	Data acquisition system
dB	Decibel
DB	Direct body
DBMS	Data base management system
DBS	Deep brain stimulation
dc	Direct current
DCCT	Diabetes control and complications trial
DCP	Distal cavity pressure
DCS	Dorsal column stimulation
DDC	Deck decompression chamber
DDS	Deep diving system
DE	Dispersive electrode
DEN	Device experience network
DERS	Drug exception ordering system
DES	Diffuse esophageal spasm
d.f.	Distribution function
DHCP	Distributed Hospital Computer Program
DHE	Dihematoporphyrin ether
DHEW	Department of Health Education and Welfare
DHHS	Department of Health and Human Services
DHT	Duration of hypothermia
DI	Deionized water
DIC	Displacement current
DIS	Diagnostic interview schedule
DL	Double layer
DLI	Difference lumen for intensity
DM	Delta modulation
DME	Dropping mercury electrode
DN	Donation number
DNA	Deoxyribonucleic acid
DOF	Degree of freedom
DOS	Drug ordering system
DOT-NHTSA	Department of Transportation Highway Traffic Safety Administration
DPB	Differential pencil beam
DPG	Diphosphoglycerate
DQE	Detection quantum efficiency
DRESS	Depth-resolved surface coil spectroscopy
DRG	Diagnosis-related group
DSA	Digital subtraction angiography
DSAR	Differential scatter-air ratio
DSB	Double strand breaks
DSC	Differential scanning calorimetry
D-T	Deuterium-on-tritium
DTA	Differential thermal analysis
d.u.	Density unit
DUR	Duration
DVT	Deep venous thrombosis
EA	Esophageal accelerogram
EB	Electron beam
EBCDIC	Extended binary code decimal interchange code
EBS	Early burn scar
EBV	Epstein–Barr Virus
EC	Ethyl cellulose
ECC	Emergency cardiac care; Extracorporeal circulation
ECCE	Extracapsular cataract extinction
ECD	Electron capture detector
ECG	Electrocardiogram
ECM	Electrochemical machining
ECMO	Extracorporeal membrane oxygenation
ECOD	Extracranial cerebrovascular occlusive disease
ECRI	Emergency Care Research Institute
ECS	Exner's Comprehensive System
ECT	Electroconvulsive shock therapy; Electroconvulsive therapy; Emission computed tomography
EDD	Estimated date of delivery
EDP	Aortic end diastolic pressure
EDTA	Ethylenediaminetetraacetic acid
EDX	Energy dispersive X-ray analysis
EEG	Electroencephalogram
EEI	Electrode electrolyte interface
EELV	End-expiratory lung volume
EER	Electrically evoked response
EF	Ejection fraction
EF	Electric field; Evoked magnetic fields
EFA	Estimated fetal age
EGF	Epidermal growth factor
EGG	Electrogastrogram
EIA	Enzyme immunoassay
EIU	Electrode impedance unbalance
ELF	Extra low frequency
ELGON	Electrical goniometer
ELISA	Enzyme-linked immunosorbent assay
ELS	Energy loss spectroscopy
ELV	Equivalent lung volume
EM	Electromagnetic
EMBS	Engineering in Medicine and Biology Society
emf	Electromotive force
EMG	Electromyogram
EMGE	Integrated electromyogram
EMI	Electromagnetic interference
EMS	Emergency medical services
EMT	Emergency medical technician
ENT	Ear, nose, and throat
EO	Elbow orthosis
EOG	Electrooculography
EOL	End of life
EOS	Eosinophil
EP	Elastoplastic; Evoked potentiate
EPA	Environmental protection agency
ER	Evoked response
ERCP	Endoscopic retrograde cholangiopancreatography
ERG	Electron radiography; Electroretinogram
ERMF	Event-related magnetic field
ERP	Event-related potential
ERV	Expiratory reserve volume

ESCA	Electron spectroscopy for chemical analysis
ESI	Electrode skin impedance
ESRD	End-stage renal disease
esu	Electrostatic unit
ESU	Electrosurgical unit
ESWL	Extracorporeal shock wave lithotripsy
ETO, Eto	Ethylene oxide
ETT	Exercise tolerance testing
EVA	Ethylene vinyl acetate
EVR	Endocardial viability ratio
EW	Extended wear
FAD	Flavin adenine dinucleotide
FARA	Flexible automation random analysis
FBD	Fetal biparietal diameter
FBS	Fetal bovine serum
fcc	Face centered cubic
FCC	Federal Communications Commission
Fct	Fluorocrit
FDA	Food and Drug Administration
FDCA	Food, Drug, and Cosmetic Act
FE	Finite element
FECG	Fetal electrocardiogram
FEF	Forced expiratory flow
FEL	Free electron lasers
FEM	Finite element method
FEP	Fluorinated ethylene propylene
FES	Functional electrical stimulation
FET	Field-effect transistor
FEV	Forced expiratory volume
FFD	Focal spot to film distance
FFT	Fast Fourier transform
FGF	Fresh gas flow
FHR	Fetal heart rate
FIC	Forced inspiratory capacity
FID	Flame ionization detector; Free-induction decay
FIFO	First-in-first-out
FITC	Fluorescent indicator tagged polymer
FL	Femur length
FM	Frequency modulation
FNS	Functional neuromuscular stimulation
FO	Foramen ovale
FO-CRT	Fiber optics cathode ray tube
FP	Fluorescence polarization
FPA	Fibrinopeptide A
FR	Federal Register
FRC	Federal Radiation Council; Functional residual capacity
FSD	Focus-to-surface distance
FTD	Focal spot to tissue-plane distance
FTIR	Fourier transform infrared
FTMS	Fourier transform mass spectrometer
FU	Fluorouracil
FUDR	Floxuridine
FVC	Forced vital capacity
FWHM	Full width at half maximum
FWTM	Full width at tenth maximum
GABA	Gamma amino buteric acid
GAG	Glycosaminoglycan
GBE	Gas-bearing electrodynamometer
GC	Gas chromatography; Guanine-cytosine
GDT	Gas discharge tube
GFR	Glomerular filtration rate
GHb	Glycosylated hemoglobin
GI	Gastrointestinal
GLC	Gas–liquid chromatography
GMV	General minimum variance
GNP	Gross national product
GPC	Giant papillary conjunctivitis
GPH	Gas-permeable hard
GPH-EW	Gas-permeable hard lens extended wear
GPO	Government Printing Office
GSC	Gas-solid chromatography
GSR	Galvanic skin response
GSWD	Generalized spike-wave discharge
HA	Hydroxyapatite
HAM	Helical axis of motion
Hb	Hemoglobin
HBE	His bundle electrogram
HBO	Hyperbaric oxygenation
HC	Head circumference
HCA	Hypothermic circulatory arrest
HCFA	Health care financing administration
HCL	Harvard Cyclotron Laboratory
hcp	Hexagonal close-packed
HCP	Half cell potential
HDPE	High density polyethylene
HECS	Hospital Equipment Control System
HEMS	Hospital Engineering Management System
HEPA	High efficiency particulate air filter
HES	Hydroxyethylstarch
HETP	Height equivalent to a theoretical plate
HF	High-frequency; Heating factor
HFCWO	High-frequency chest wall oscillation
HFER	High-frequency electromagnetic radiation
HFJV	High-frequency jet ventilation
HFO	High-frequency oscillator
HFOV	High-frequency oscillatory ventilation
HFPPV	High-frequency positive pressure ventilation
HFV	High-frequency ventilation
HHS	Department of Health and Human Services
HIBC	Health industry bar code
HIMA	Health Industry Manufacturers Association
HIP	Hydrostatic indifference point
HIS	Hospital information system
HK	Hexokinase
HL	Hearing level
HMBA	Hexamethylene bisacetamide
HMO	Health maintenance organization
HMWPE	High-molecular-weight polyethylene
HOL	Higher-order languages
HP	Heating factor; His-Purkinje
HpD	Hematoporphyrin derivative
HPLC	High-performance liquid chromatography
HPNS	High-pressure neurological syndrome
HPS	His-Purkinje system
HPX	High peroxidase activity

HR	Heart rate; High-resolution
HRNB	Halstead-Reitan Neuropsychological Battery
H/S	Hard/soft
HSA	Human serum albumin
HSG	Hysterosalpingogram
HTCA	Human tumor cloning assay
HTLV	Human T cell lymphotrophic virus
HU	Heat unit; Houndsfield units; Hydroxyurea
HVL	Half value layer
HVR	Hypoxic ventilatory response
HVT	Half-value thickness
IA	Image intensifier assembly; Inominate artery
IABP	Intraaortic balloon pumping
IAEA	International Atomic Energy Agency
IAIMS	Integrated Academic Information Management System
IASP	International Association for the Study of Pain
IC	Inspiratory capacity; Integrated circuit
ICCE	Intracapsular cataract extraction
ICD	Intracervical device
ICDA	International classification of diagnoses
ICL	Ms-clip lens
ICP	Inductively coupled plasma; Intracranial pressure
ICPA	Intracranial pressure amplitude
ICRP	International Commission on Radiological Protection
ICRU	International Commission on Radiological Units and Measurements
ICU	Intensive care unit
ID	Inside diameter
IDDM	Insulin dependent diabetes mellitus
IDE	Investigational device exemption
IDI	Index of inspired gas distribution
I:E	Inspiratory: expiratory
IEC	International Electrotechnical Commission; Ion-exchange chromatography
IEEE	Institute of Electrical and Electronics Engineers
IEP	Individual educational program
BETS	Inelastic electron tunneling spectroscopy
IF	Immunofluorescent
IFIP	International Federation for Information Processing
IFMBE	International Federation for Medical and Biological Engineering
IGFET	Insulated-gate field-effect transistor
IgG	Immunoglobulin G
IgM	Immunoglobulin M
IHP	Inner Helmholtz plane
IHSS	Idiopathic hypertrophic subaortic stenosis
II	Image intensifier
IIIES	Image intensifier input-exposure sensitivity
IM	Intramuscular
IMFET	Immunologically sensitive field-effect transistor
IMIA	International Medical Informatics Association
IMS	Information management system
IMV	Intermittent mandatory ventilation
INF	Interferon
IOL	Intraocular lens
IPC	Ion-pair chromatography
IPD	Intermittent peritoneal dialysis
IPG	Impedance plethysmography
IPI	Interpulse interval
IPPB	Intermittent positive pressure breathing
IPTS	International practical temperature scale
IR	Polyisoprene rubber
IRB	Institutional Review Board
IRBBB	Incomplete right bundle branch block
IRPA	International Radiation Protection Association
IRRAS	Infrared reflection-absorption spectroscopy
IRRS	Infrared reflection spectroscopy
IRS	Internal reflection spectroscopy
IRV	Inspiratory reserve capacity
IS	Image size; Ion-selective
ISC	Infant skin servo control
ISDA	Instantaneous screw displacement axis
ISE	Ion-selective electrode
ISFET	Ion-sensitive field effect transistor
ISIT	Intensified silicon-intensified target tube
ISO	International Organization for Standardization
ISS	Ion scattering spectroscopy
IT	Intrathecal
ITEP	Institute of Theoretical and Experimental Physics
ITEPI	Instantaneous trailing edge pulse impedance
ITLC	Instant thin-layer chromatography
IUD	Intrauterine device
IV	Intravenous
IVC	Inferior vena cava
IVP	Intraventricular pressure
JCAH	Joint Commission on the Accreditation of Hospitals
JND	Just noticeable difference
JRP	Joint replacement prosthesis
KB	Kent bundle
Kerma	Kinetic energy released in unit mass
KO	Knee orthosis
KPM	Kilopond meter
KRPB	Krebs-Ringer physiological buffer
LA	Left arm; Left atrium
LAD	Left anterior descending; Left axis deviation
LAE	Left atrial enlargement
LAK	Lymphokine activated killer
LAL	Limulus amoebocyte lysate
LAN	Local area network
LAP	Left atrial pressure
LAT	Left anterior temporalis
LBBB	Left bundle branch block
LC	Left carotid; Liquid chromatography

LCC	Left coronary cusp
LCD	Liquid crystal display
LDA	Laser Doppler anemometry
LDF	Laser Doppler flowmetry
LDH	Lactate dehydrogenase
LDPE	Low density polyethylene
LEBS	Low-energy brief stimulus
LED	Light-emitting diode
LEED	Low energy electron diffraction
LES	Lower esophageal sphincter
LESP	Lower esophageal sphincter pressure
LET	Linear energy transfer
LF	Low frequency
LH	Luteinizing hormone
LHT	Local hyperthermia
LL	Left leg
LLDPE	Linear low density polyethylene
LLPC	Liquid-liquid partition chromatography
LLW	Low-level waste
LM	Left masseter
LNNB	Luria-Nebraska Neuropsychological Battery
LOS	Length of stay
LP	Late potential; Lumboperitoneal
LPA	Left pulmonary artery
LPC	Linear predictive coding
LPT	Left posterior temporalis
LPV	Left pulmonary veins
LRP	Late receptor potential
LS	Left subclavian
LSC	Liquid-solid adsorption chromatography
LSI	Large scale integrated
LSV	Low-amplitude shear-wave viscoelastometry
LTI	Low temperature isotropic
LUC	Large unstained cells
LV	Left ventricle
LVAD	Left ventricular assist device
LVDT	Linear variable differential transformer
LVEP	Left ventricular ejection period
LVET	Left ventricular ejection time
LVH	Left ventricular hypertrophy
LYMPH	Lymphocyte
MAA	Macroaggregated albumin
MAC	Minimal auditory capabilities
MAN	Manubrium
MAP	Mean airway pressure; Mean arterial pressure
MAST	Military assistance to safety and traffic
MBA	Monoclonal antibody
MBV	Maximum breathing ventilation
MBX	Monitoring branch exchange
MCA	Methyl cryanoacrylate
MCG	Magnetocardiogram
MCI	Motion Control Incorporated
MCMI	Millon Clinical Multiaxial Inventory
MCT	Microcatheter transducer
MCV	Mean corpuscular volume
MDC	Medical diagnostic categories
MDI	Diphenylmethane diisocyanate; Medical Database Informatics
MDP	Mean diastolic aortic pressure
MDR	Medical device reporting
MDS	Multidimensional scaling
ME	Myoelectric
MED	Minimum erythema dose
MEDPAR	Medicare provider analysis and review
MEFV	Maximal expiratory flow volume
MEG	Magnetoencephalography
MeSH	Medline subject heading
METS	Metabolic equivalents
MF	Melamine-formaldehyde
MFP	Magnetic field potential
MGH	Massachusetts General Hospital
MHV	Magnetic heart vector
MI	Myocardial infarction
MIC	Minimum inhibitory concentration
MIFR	Maximum inspiratory flow rate
MINET	Medical Information Network
MIR	Mercury-in-rubber
MIS	Medical information system; Metal-insulator-semiconductor
MIT	Massachusetts Institute of Technology
MIT/BIH	Massachusetts Institute of Technology/ Beth Israel Hospital
MMA	Manual metal arc welding
MMA	Methyl methacrylate
MMECT	Multiple-monitored ECT
MMFR	Maximum midexpiratory flow rate
mm Hg	Millimeters of mercury
MMPI	Minnesota Multiphasic Personality Inventory
MMSE	Minimum mean square error
MO	Membrane oxygenation
MONO	Monocyte
MOSFET	Metal oxide silicon field-effect transistor
MP	Mercaptopurine; Metacarpal-phalangeal
MPD	Maximal permissible dose
MR	Magnetic resonance
MRG	Magnetoretinogram
MRI	Magnetic resonance imaging
MRS	Magnetic resonance spectroscopy
MRT	Mean residence time
MS	Mild steel; Multiple sclerosis
MSR	Magnetically shielded room
MTBF	Mean time between failure
MTF	Modulation transfer function
MTTR	Mean time to repair
MTX	Methotroxate
MUA	Motor unit activity
MUAP	Motor unit action potential
MUAPT	Motor unit action potential train
MUMPI	Missouri University Multi-Plane Imager
MUMPS	Massachusetts General Hospital utility multiuser programming system
MV	Mitral valve
MVO_2	Maximal oxygen uptake
MVTR	Moisture vapor transmission rate
MVV	Maximum voluntary ventilation
MW	Molecular weight

NAA	Neutron activation analysis
NAD	Nicotinamide adenine dinucleotide
NADH	Nicotinamide adenine dinucleotide, reduced form
NADP	Nicotinamide adenine dinucleotide phosphate
NAF	Neutrophil activating factor
NARM	Naturally occurring and accelerator-produced radioactive materials
NBB	Normal buffer base
NBD	Neuromuscular blocking drugs
N-BPC	Normal bonded phase chromatography
NBS	National Bureau of Standards
NCC	Noncoronary cusp
NCCLS	National Committee for Clinical Laboratory Standards; National Committee on Clinical Laboratory Standards
NCRP	National Council on Radiation Protection
NCT	Neutron capture theory
NEEP	Negative end-expiratory pressure
NEMA	National Electrical Manufacturers Association
NEMR	Nonionizing electromagnetic radiation
NEQ	Noise equivalent quanta
NET	Norethisterone
NEUT	Neutrophil
NFPA	National Fire Protection Association
NH	Neonatal hepatitis
NHE	Normal hydrogen electrode
NHLBI	National Heart, Lung, and Blood Institute
NIR	Nonionizing radiation
NIRS	National Institute for Radiologic Science
NK	Natural killer
NMJ	Neuromuscular junction
NMOS	N-type metal oxide silicon
NMR	Nuclear magnetic resonance
NMS	Neuromuscular stimulation
NPH	Normal pressure hydrocephalus
NPL	National Physical Laboratory
NR	Natural rubber
NRC	Nuclear Regulatory Commission
NRZ	Non-return-to-zero
NTC	Negative temperature coefficient
NTIS	National Technical Information Service
NVT	Neutrons versus time
NYHA	New York Heart Association
ob/gyn	Obstetrics and gynecology
OCR	Off-center ratio; Optical character recognition
OCV	Open circuit voltage
OD	Optical density; Outside diameter
ODC	Oxyhemoglobin dissociation curve
ODT	Oxygen delivery truck
ODU	Optical density unit
OER	Oxygen enhancement ratio
OFD	Object to film distance; Occiputo-frontal diameter
OHL	Outer Helmholtz layer
OHP	Outer Helmholtz plane
OIH	Orthoiodohippurate
OPG	Ocular pneumoplethysmography
OR	Operating room
OS	Object of known size; Operating system
OTC	Over the counter
OV	Offset voltage
PA	Posterioanterior; Pulmonary artery; Pulse amplitude
PACS	Picture archiving and communications systems
PAD	Primary afferent depolarization
PAM	Pulse amplitude modulation
PAN	Polyacrylonitrile
PAP	Pulmonary artery pressure
PAR	Photoactivation ratio
PARFR	Program for Applied Research on Fertility Regulation
PARR	Poetanesthesia recovery room
PAS	Photoacoustic spectroscopy
PASG	Pneumatic antishock garment
PBI	Penile brachial index
PBL	Positive beam limitation
PBT	Polybutylene terephthalate
PC	Paper chromatography; Personal computer; Polycarbonate
PCA	Patient controlled analgesia; Principal components factor analysis
PCG	Phonocardiogram
PCI	Physiological cost index
PCL	Polycaprolactone; Posterior chamber lens
PCR	Percent regurgitation
PCRC	Perinatal Clinical Research Center
PCS	Patient care system
PCT	Porphyria cutanea tarda
PCWP	Pulmonary capillary wedge pressure
PD	Peritoneal dialysis; Poly-p-dioxanone; Potential difference; Proportional and derivative
PDD	Percent depth dose; Perinatal Data Directory
PDE	Pregelled disposable electrodes
p.d.f.	Probability density function
PDL	Periodontal ligament
PDM	Pulse duration modulation
PDMSX	Polydimethyl siloxane
PDS	Polydioxanone
PE	Polyethylene
PEEP	Positive end-expiratory pressure
PEFR	Peak expiratory now rate
PEN	Parenteral and enteral nutrition
PEP	Preejection period
PEPPER	Programs examine phonetic find phonological evaluation records
PET	Polyethylene terephthalate; Positron-emission tomography
PEU	Polyetherurethane
PF	Platelet factor
PFA	Phosphonoformic add
PFC	Petrofluorochemical
PFT	Pulmonary function testing
PG	Polyglycolide; Propylene glycol

PGA	Polyglycolic add
PHA	Phytohemagglutinin; Pulse-height analyzer
PHEMA	Poly-2-hydroxyethyl methacrylate
PI	Propidium iodide
PID	Pelvic inflammatory disease; Proportional/integral/derivative
PIP	Peak inspiratory pressure
PL	Posterior leaflet
PLA	Polylactic acid
PLATO	Program Logic for Automated Teaching Operations
PLD	Potentially lethal damage
PLED	Periodic latoralized epileptiform discharge
PLT	Platelet
PM	Papillary muscles; Preventive maintenance
PMA	Polymethyl acrylate
p.m.f.	Probability mass function
PMMA	Polymethyl methacrylate
PMOS	P-type metal oxide silicon
PMP	Patient management problem; Poly(4-methylpentane)
PMT	Photomultiplier tube
PO	Per os
Po_2	Partial pressure of oxygen
POBT	Polyoxybutylene terephthalate
POM	Polyoxymethylene
POMC	Patient order management and communication system
POPRAS	Problem Oriented Perinatal Risk Assessment System
PP	Perfusion pressure; Polyproplyene; Postprandial (after meals)
PPA	Phonemic process analysis
PPF	Plasma protein fraction
PPM	Pulse position modulation
PPSFH	Polymerized phyridoxalated stroma-free hemoglobin
PR	Pattern recognition; Pulse rate
PRBS	Pseudo-random binary signals
PRP	Pulse repetition frequency
PRO	Professional review organization
PROM	Programmable read only memory
PS	Polystyrene
PSA	Pressure-sensitive adhesive
PSF	Point spread function
PSI	Primary skin irritation
PSP	Postsynaptic potential
PSR	Proton spin resonance
PSS	Progressive systemic sclerosis
PT	Plasma thromboplastin
PTB	Patellar tendon bearing orthosis
PTC	Plasma thromboplastin component; Positive temperature coefficient; Pressurized personal transfer capsule
PTCA	Percutaneous transluminal coronary angioplasty
PTFE	Polytetrafluoroethylene
PTT	Partial thromboplastin time
PUL	Percutaneous ultrasonic lithotripsy
PURA	Prolonged ultraviolet-A radiation
PUVA	Psoralens and longwave ultraviolet light photochemotherapy
P/V	Pressure/volume
PVC	Polyvinyl chloride; Premature ventricular contraction
PVI	Pressure–volume index
PW	Pulse wave; Pulse width
PWM	Pulse width modulation
PXE	Pseudo-xanthoma elasticum
QA	Quality assurance
QC	Quality control
R-BPC	Reverse bonded phase chromatography
R/S	Radiopaque-spherical
RA	Respiratory amplitude; Right arm
RAD	Right axis deviation
RAE	Right atrial enlargement
RAM	Random access memory
RAP	Right atrial pressure
RAT	Right anterior temporalis
RB	Right bundle
RBBB	Right bundle branch block
RBC	Red blood cell
RBE	Relative biologic effectiveness
RBF	Rose bengal fecal excretion
RBI	Resting baseline impedance
RCBD	Randomized complete block diagram
rCBF	Regional cerebral blood flow
RCC	Right coronary cusp
RCE	Resistive contact electrode
R&D	Research and development
r.e.	Random experiment
RE	Reference electrode
REM	Rapid eye movement; Return electrode monitor
REMATE	Remote access and telecommunication system
RES	Reticuloendothelial system
RESNA	Rehabilitation Engineering Society of North America
RF	Radio frequency; Radiographic-nuoroscopic
RFI	Radio-frequency interference
RFP	Request for proposal
RFQ	Request for quotation
RH	Relative humidity
RHE	Reversible hydrogen electrode
RIA	Radioimmunoassay
RM	Repetition maximum; Right masseter
RMR	Resting metabolic rate
RMS	Root mean square
RN	Radionuclide
RNCA	Radionuclide cineagiogram
ROI	Regions of interest
ROM	Range of motion; Read only memory
RP	Retinitis pigmentosa
RPA	Right pulmonary artery
RPP	Rate pressure product
RPT	Rapid pull-through technique
RPV	Right pulmonary veins
RQ	Respiratory quotient

RR	Recovery room
RRT	Recovery room time; Right posterior temporalis
RT	Reaction time
RTD	Resistance temperature device
RTT	Revised token test
r.v.	Random variable
RV	Residual volume; Right ventricle
RVH	Right ventricular hypertrophy
RVOT	Right ventricular outflow tract
RZ	Return-to-zero
SA	Sinoatrial; Specific absorption
SACH	Solid-ankle-cushion-heel
SAD	Source-axis distance; Statistical Analysis System
SAINT	System analysis of integrated network of tasks
SAL	Sterility assurance level; Surface averaged lead
SALT	Systematic analysis of language transcripts
SAMI	Socially acceptable monitoring instrument
SAP	Systemic arterial pressure
SAR	Scatter-air ratio; Specific absorption rate
SARA	System for anesthetic and respiratory gas analysis
SBE	Subbacterial endocarditis
SBR	Styrene-butadiene rubbers
SC	Stratum corneum; Subcommittees
SCAP	Right scapula
SCE	Saturated calomel electrode; Sister chromatid exchange
SCI	Spinal cord injury
SCRAD	Sub-Committee on Radiation Dosimetry
SCS	Spinal cord stimulation
SCUBA	Self-contained underwater breathing apparatus
SD	Standard deviation
SDA	Stepwise discriminant analysis
SDS	Sodium dodecyl sulfate
S&E	Safety and effectiveness
SE	Standard error
SEC	Size exclusion chromatography
SEM	Scanning electron microscope; Standard error of the mean
SEP	Somatosensory evoked potential
SEXAFS	Surface extended X-ray absorption fine structure
SF	Surviving fraction
SFD	Source-film distance
SFH	Stroma-free hemoglobin
SFTR	Sagittal frontal transverse rotational
SG	Silica gel
SGF	Silica gel fraction
SGG	Spark gap generator
SGOT	Serum glutamic oxaloacetic transaminase
SGP	Strain gage plethysmography; Stress-generated potential
SHE	Standard hydrogen electrode
SI	Le Système International d'Unités
SEBS	Surgical isolation barrier system
SID	Source to image reception distance
SIMFU	Scanned intensity modulated focused ultrasound
SIMS	Secondary ion mass spectroscopy; System for isometric muscle strength
SISI	Short increment sensitivity index
SL	Surgical lithotomy
SLD	Sublethal damage
SLE	Systemic lupus erythemotodes
SMA	Sequential multiple analyzer
SMAC	Sequential multiple analyzer with computer
SMR	Sensorimotor
S/N	Signal-to-noise
S:N/D	Signal-to-noise ratio per unit dose
SNP	Sodium nitroprusside
SNR	Signal-to-noise ratio
SOA	Sources of artifact
SOAP	Subjective, objective, assessment, plan
SOBP	Spread-out Bragg peak
SP	Skin potential
SPECT	Single photon emission computed tomography
SPL	Sound pressure level
SPRINT	Single photon ring tomograph
SPRT	Standard platinum resistance thermometer
SPSS	Statistical Package for the Social Sciences
SQUID	Superconducting quantum interference device
SQV	Square wave voltammetry
SR	Polysulfide rubbers
SRT	Speech reception threshold
SS	Stainless steel
SSB	Single strand breaks
SSD	Source-to-skin distance; Source-to-surface distance
SSE	Stainless steel electrode
SSEP	Somatosensory evoked potential
SSG	Solid state generator
SSP	Skin stretch potential
SSS	Sick sinus syndrome
STD	Source-tray distance
STI	Systolic time intervals
STP	Standard temperature and pressure
STPD	Standard temperature pressure dry
SV	Stroke volume
SVC	Superior vena cava
SW	Standing wave
TAA	Tumor-associated antigens
TAC	Time-averaged concentration
TAD	Transverse abdominal diameter
TAG	Technical Advisory Group
TAH	Total artificial heart
TAR	Tissue-air ratio
TC	Technical Committees
TCA	Tricarboxylic acid cycle
TCD	Thermal conductivity detector
TCES	Transcutaneous cranial electrical stimulation

TCP	Tricalcium phosphate
TDD	Telecommunication devices for the deaf
TDM	Therapeutic drug monitoring
TE	Test electrode; Thermoplastic elastomers
TEAM	Technology evaluation and acquisition methods
TEM	Transmission electron microscope; Transverse electric and magnetic mode; Transverse electromagnetic mode
TENS	Transcutaneous electrical nerve stimulation
TEP	Tracheoesophageal puncture
TEPA	Triethylenepho-sphoramide
TF	Transmission factor
TFE	Tetrafluorethylene
TI	Totally implantable
TICCIT	Time-shared Interaction Computer-Controlled Information Television
TLC	Thin-layer chromatography; Total lung capacity
TLD	Thermoluminescent dosimetry
TMJ	Temporomandibular joint
TMR	Tissue maximum ratio; Topical magnetic resonance
TNF	Tumor necrosis factor
TOF	Train-of-four
TP	Thermal performance
TPC	Temperature pressure correction
TPD	Triphasic dissociation
TPG	Transvalvular pressure gradient
TPN	Total parenteral nutrition
TR	Temperature rise
tRNA	Transfer RNA
TSH	Thyroid stimulating hormone
TSS	Toxic shock syndrome
TTD	Telephone devices for the deaf
TTI	Tension time index
TTR	Transition temperature range
TTV	Trimming tip version
TTY	Teletypewriter
TUR	Transurethral resection
TURP	Transurethral resections of the prostrate
TV	Television; Tidal volume; Tricuspid valve
TVER	Transscleral visual evoked response
TW	Traveling wave
TxB_2	Thrombozame B^2
TZ	Transformation zone
UES	Upper esophageal sphincter
UP	Urea-formaldehyde
UfflS	University Hospital Information System
UHMW	Ultra high molecular weight
UHMWPE	Ultra high molecular weight polyethylene
UL	Underwriters Laboratory
ULF	Ultralow frequency
ULTI	Ultralow temperature isotropic
UMN	Upper motor neuron
UO	Urinary output
UPTD	Unit pulmonary oxygen toxicity doses
UR	Unconditioned response
US	Ultrasound; Unconditioned stimulus
USNC	United States National Committee
USP	United States Pharmacopeia
UTS	Ultimate tensile strength
UV	Ultraviolet; Umbilical vessel
UVR	Ultraviolet radiation
V/F	Voltage-to-frequency
VA	Veterans Administration
VAS	Visual analog scale
VBA	Vaginal blood volume in arousal
VC	Vital capacity
VCO	Voltage-controlled oscillator
VDT	Video display terminal
VECG	Vectorelectrocardiography
VEP	Visually evoked potential
VF	Ventricular fibrillation
VOP	Venous occlusion plethysmography
VP	Ventriculoperitoneal
VPA	Vaginal pressure pulse in arousal
VPB	Ventricular premature beat
VPR	Volume pressure response
VSD	Ventricular septal defect
VSWR	Voltage standing wave ratio
VT	Ventricular tachycardia
VTG	Vacuum tube generator
VTS	Viewscan text system
VV	Variable version
WAIS-R	Weschler Adult Intelligence Scale-Revised
WAK	Wearable artificial kidney
WAML	Wide-angle mobility light
WBAR	Whole-body autoradiography
WBC	White blood cell
WG	Working Groups
WHO	World Health Organization; Wrist hand orthosis
WLF	Williams-Landel-Ferry
WMR	Work metabolic rate
w/o	Weight percent
WORM	Write once, read many
WPW	Wolff-Parkinson-White
XPS	X-ray photon spectroscopy
XR	Xeroradiograph
YAG	Yttrium aluminum garnet
ZPL	Zero pressure level

CONVERSION FACTORS AND UNIT SYMBOLS

SI UNITS (ADOPTED 1960)

A new system of metric measurement, the International System of Units (abbreviated SI), is being implemented throughout the world. This system is a modernized version of the MKSA (meter, kilogram, second, ampere) system, and its details are published and controlled by an international treaty organization (The International Bureau of Weights and Measures).

SI units are divided into three classes:

Base Units

length	meter[†] (m)
mass[‡]	kilogram (kg)
time	second (s)
electric current	ampere (A)
thermodynamic temperature§	kelvin (K)
amount of substance	mole (mol)
luminous intensity	candela (cd)

Supplementary Units

plane angle	radian (rad)
solid angle	steradian (sr)

Derived Units and Other Acceptable Units

These units are formed by combining base units, supplementary units, and other derived units. Those derived units having special names and symbols are marked with an asterisk (*) in the list below:

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
*absorbed dose	gray	Gy	J/kg
acceleration	meter per second squared	m/s^2	
*activity (of ionizing radiation source)	becquerel	Bq	1/s
area	square kilometer	km^2	
	square hectometer	hm^2	ha (hectare)
	square meter	m^2	

[†]The spellings "metre" and "litre" are preferred by American Society for Testing and Materials (ASTM); however, "−er" will be used in the Encyclopedia.

[‡]"Weight" is the commonly used term for "mass."

§Wide use is made of "Celsius temperature" (t) defined $t = T - T_0$ where T is the thermodynamic temperature, expressed in kelvins, and $T_0 = 273.15\,K$ by definition. A temperature interval may be expressed in degrees Celsius as well as in kelvins.

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
*capacitance	farad	F	C/V
concentration (of amount of substance)	mole per cubic meter	mol/m^3	
*conductance	siemens	S	A/V
current density	ampere per square meter	A/m^2	
density, mass density	kilogram per cubic meter	kg/m^3	g/L; mg/cm^3
dipole moment (quantity)	coulomb meter	C·m	
*electric charge, quantity of electricity	coulomb	C	A·s
electric charge density	coulomb per cubic meter	C/m^3	
electric field strength	volt per meter	V/m	
electric flux density	coulomb per square meter	C/m^2	
*electric potential, potential difference, electromotive force	volt	V	W/A
*electric resistance	ohm	Ω	V/A
*energy, work, quantity of heat	megajoule	MJ	
	kilojoule	kJ	
	joule	J	N·m
	electron volt†	eV†	
	kilowatt hour†	kW·h†	
energy density	joule per cubic meter	J/m^3	
*force	kilonewton	kN	
	newton	N	$kg·m/s^2$
*frequency	megahertz	MHz	
	hertz	Hz	1/s
heat capacity, entropy	joule per kelvin	J/K	
heat capacity (specific), specific entropy	joule per kilogram kelvin	J/(kg·K)	
heat transfer coefficient	watt per square meter kelvin	$W/(m^2·K)$	
*illuminance	lux	lx	lm/m^2
*inductance	henry	H	Wb/A
linear density	kilogram per meter	kg/m	
luminance	candela per square meter	cd/m^2	
*luminous flux	lumen	lm	cd·sr
magnetic field strength	ampere per meter	A/m	
*magnetic flux	weber	Wb	V·s
*magnetic flux density	tesla	T	Wb/m^2
molar energy	joule per mole	J/mol	
molar entropy, molar heat capacity	joule per mole kelvin	J/(mol·K)	
moment of force, torque	newton meter	N·m	
momentum	kilogram meter per second	kg·m/s	
permeability	henry per meter	H/m	
permittivity	farad per meter	F/m	
*power, heat flow rate, radiant flux	kilowatt	kW	
	watt	W	J/s
power density, heat flux density, irradiance	watt per square meter	W/m^2	
*pressure, stress	megapascal	MPa	
	kilopascal	kPa	
	pascal	Pa	N/m^2
sound level	decibel	dB	
specific energy	joule per kilogram	J/kg	
specific volume	cubic meter per kilogram	m^3/kg	
surface tension	newton per meter	N/m	
thermal conductivity	watt per meter kelvin	W/(m·K)	
velocity	meter per second	m/s	
	kilometer per hour	km/h	
viscosity, dynamic	pascal second	Pa·s	
	millipascal second	mPa·s	

†This non-SI unit is recognized as having to be retained because of practical importance or use in specialized fields.

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
viscosity, kinematic	square meter per second	m^2/s	
	square millimeter per second	mm^2/s	
	cubic meter	m^3	
	cubic decimeter	dm^3	L(liter)
	cubic centimeter	cm^3	mL
wave number	1 per meter	m^{-1}	
	1 per centimeter	cm^{-1}	

In addition, there are 16 prefixes used to indicate order of magnitude, as follows:

Multiplication factor	*Prefix*	*Symbol*	*Note*
10^{18}	exa	E	
10^{15}	peta	P	
10^{12}	tera	T	
10^{9}	giga	G	
10^{8}	mega	M	
10^{3}	kilo	k	
10^{2}	hecto	h[a]	[a]Although hecto, deka, deci, and centi are
10	deka	da[a]	SI prefixes, their use should be avoided
10^{-1}	deci	d[a]	except for SI unit-multiples for area and
10^{-2}	centi	c[a]	volume and nontechnical use of
10^{-3}	milli	m	centimeter, as for body and clothing
10^{-6}	micro	μ	measurement.
10^{-9}	nano	n	
10^{-12}	pico	p	
10^{-15}	femto	f	
10^{-18}	atto	a	

For a complete description of SI and its use the reader is referred to ASTM E 380.

CONVERSION FACTORS TO SI UNITS

A representative list of conversion factors from non-SI to SI units is presented herewith. Factors are given to four significant figures. Exact relationships are followed by a dagger (†). A more complete list is given in ASTM E 380-76 and ANSI Z210. 1-1976.

To convert from	*To*	*Multiply by*
acre	square meter (m^2)	4.047×10^{3}
angstrom	meter (m)	$1.0 \times 10^{-10\dagger}$
are	square meter (m^2)	$1.0 \times 10^{2\dagger}$
astronomical unit	meter (m)	1.496×10^{11}
atmosphere	pascal (Pa)	1.013×10^{5}
bar	pascal (Pa)	$1.0 \times 10^{5\dagger}$
barrel (42 U.S. liquid gallons)	cubic meter (m^3)	0.1590
Btu (International Table)	joule (J)	1.055×10^{3}
Btu (mean)	joule (J)	1.056×10^{3}
Bt (thermochemical)	joule (J)	1.054×10^{3}
bushel	cubic meter (m^3)	3.524×10^{-2}
calorie (International Table)	joule (J)	4.187
calorie (mean)	joule (J)	4.190
calorie (thermochemical)	joule (J)	$4.184^{\dagger}$
centimeters of water (39.2 °F)	pascal (Pa)	98.07
centipoise	pascal second (Pa·s)	$1.0 \times 10^{-3\dagger}$
centistokes	square millimeter per second (mm^2/s)	$1.0^{\dagger}$

To convert from	*To*	*Multiply by*
cfm (cubic foot per minute)	cubic meter per second (m^3/s)	4.72×10^{-4}
cubic inch	cubic meter (m^3)	1.639×10^{-4}
cubic foot	cubic meter (m^3)	2.832×10^{-2}
cubic yard	cubic meter (m^3)	0.7646
curie	becquerel (Bq)	$3.70 \times 10^{10\dagger}$
debye	coulomb-meter (C·m)	3.336×10^{-30}
degree (angle)	radian (rad)	1.745×10^{-2}
denier (international)	kilogram per meter (kg/m)	1.111×10^{-7}
	tex	0.1111
dram (apothecaries')	kilogram (kg)	3.888×10^{-3}
dram (avoirdupois)	kilogram (kg)	1.772×10^{-3}
dram (U.S. fluid)	cubic meter (m^3)	3.697×10^{-6}
dyne	newton(N)	$1.0 \times 10^{-6\dagger}$
dyne/cm	newton per meter (N/m)	$1.00 \times 10^{-3\dagger}$
electron volt	joule (J)	1.602×10^{-19}
erg	joule (J)	$1.0 \times 10^{-7\dagger}$
fathom	meter (m)	1.829
fluid ounce (U.S.)	cubic meter (m^3)	2.957×10^{-5}
foot	meter (m)	$0.3048^{\dagger}$
foot-pound force	joule (J)	1.356
foot-pound force	newton meter (N·m)	1.356
foot-pound force per second	watt(W)	1.356
footcandle	lux (lx)	10.76
furlong	meter (m)	2.012×10^2
gal	meter per second squared (m/s^2)	$1.0 \times 10^{-2\dagger}$
gallon (U.S. dry)	cubic meter (m^3)	4.405×10^{-3}
gallon (U.S. liquid)	cubic meter (m^3)	3.785×10^{-3}
gilbert	ampere (A)	0.7958
gill (U.S.)	cubic meter (m^3)	1.183×10^{-4}
grad	radian	1.571×10^{-2}
grain	kilogram (kg)	6.480×10^{-5}
gram force per denier	newton per tex (N/tex)	8.826×10^{-2}
hectare	square meter (m^2)	$1.0 \times 10^{4\dagger}$
horsepower (550 ft·lbf/s)	watt(W)	7.457×10^2
horsepower (boiler)	watt(W)	9.810×10^3
horsepower (electric)	watt(W)	$7.46 \times 10^{2\dagger}$
hundredweight (long)	kilogram (kg)	50.80
hundredweight (short)	kilogram (kg)	45.36
inch	meter (m)	$2.54 \times 10^{-2\dagger}$
inch of mercury (32 °F)	pascal (Pa)	3.386×10^3
inch of water (39.2 °F)	pascal (Pa)	2.491×10^2
kilogram force	newton (N)	9.807
kilopond	newton (N)	9.807
kilopond-meter	newton-meter (N·m)	9.807
kilopond-meter per second	watt (W)	9.807
kilopond-meter per min	watt(W)	0.1635
kilowatt hour	megajoule (MJ)	$3.6^{\dagger}$
kip	newton (N)	4.448×10^2
knot international	meter per second (m/s)	0.5144
lambert	candela per square meter (cd/m^2)	3.183×10^3
league (British nautical)	meter (m)	5.559×10^2
league (statute)	meter (m)	4.828×10^3
light year	meter (m)	9.461×10^{15}
liter (for fluids only)	cubic meter (m^3)	$1.0 \times 10^{-3\dagger}$
maxwell	weber (Wb)	$1.0 \times 10^{-8\dagger}$
micron	meter (m)	$1.0 \times 10^{-6\dagger}$
mil	meter (m)	$2.54 \times 10^{-5\dagger}$
mile (U.S. nautical)	meter (m)	$1.852 \times 10^{3\dagger}$
mile (statute)	meter (m)	1.609×10^3
mile per hour	meter per second (m/s)	0.4470

To convert from	*To*	*Multiply by*
millibar	pascal (Pa)	1.0×10^{2}
millimeter of mercury (0 °C)	pascal (Pa)	1.333×10^{2}†
millimeter of water (39.2 °F)	pascal (Pa)	9.807
minute (angular)	radian	2.909×10^{-4}
myriagram	kilogram (kg)	10
myriameter	kilometer (km)	10
oersted	ampere per meter (A/m)	79.58
ounce (avoirdupois)	kilogram (kg)	2.835×10^{-2}
ounce (troy)	kilogram (kg)	3.110×10^{-2}
ounce (U.S. fluid)	cubic meter (m^3)	2.957×10^{-5}
ounce-force	newton (N)	0.2780
peck (U.S.)	cubic meter (m^3)	8.810×10^{-3}
pennyweight	kilogram (kg)	1.555×10^{-3}
pint (U.S. dry)	cubic meter (m^3)	5.506×10^{-4}
pint (U.S. liquid)	cubic meter (m^3)	4.732×10^{-4}
poise (absolute viscosity)	pascal second (Pa·s)	0.10†
pound (avoirdupois)	kilogram (kg)	0.4536
pound (troy)	kilogram (kg)	0.3732
poundal	newton (N)	0.1383
pound-force	newton (N)	4.448
pound per square inch (psi)	pascal (Pa)	6.895×10^{3}
quart (U.S. dry)	cubic meter (m^3)	1.101×10^{-3}
quart (U.S. liquid)	cubic meter (m^3)	9.464×10^{-4}
quintal	kilogram (kg)	1.0×10^{2}†
rad	gray (Gy)	1.0×10^{-2}†
rod	meter (m)	5.029
roentgen	coulomb per kilogram (C/kg)	2.58×10^{-4}
second (angle)	radian (rad)	4.848×10^{-6}
section	square meter (m^2)	2.590×10^{6}
slug	kilogram (kg)	14.59
spherical candle power	lumen (lm)	12.57
square inch	square meter (m^2)	6.452×10^{-4}
square foot	square meter (m^2)	9.290×10^{-2}
square mile	square meter (m^2)	2.590×10^{6}
square yard	square meter (m^2)	0.8361
store	cubic meter (m^3)	1.0†
stokes (kinematic viscosity)	square meter per second (m^2/s)	1.0×10^{-4}†
tex	kilogram per meter (kg/m)	1.0×10^{-6}†
ton (long, 2240 pounds)	kilogram (kg)	1.016×10^{3}
ton (metric)	kilogram (kg)	1.0×10^{3}†
ton (short, 2000 pounds)	kilogram (kg)	9.072×10^{2}
torr	pascal (Pa)	1.333×10^{2}
unit pole	weber (Wb)	1.257×10^{-7}
yard	meter (m)	0.9144†

CAPACITIVE MICROSENSORS FOR BIOMEDICAL APPLICATIONS

D. Tsoukalas
S. Chatzandroulis
D. Goustouridis
NTUA
Athens, Attiki
Greece

INTRODUCTION

The use of reliable, high performance miniature sensors in the medical field is of growing importance for patient health monitoring. Batch sensor fabrication, as this has been introduced by Integrated Circuit (IC) manufacturing, is an efficient way to produce silicon sensors with desirable characteristics. Since microsensors combine small size with electrical and mechanical principles of operation they constitute together with microactuators what is usually called Microelectromechanical Systems (MEMS). These components are mainly made from silicon and other related materials to explore existing know-how and infrastructure of silicon technology. All of the above factors have allowed for fast growth of the microsensor field during the last years. This article focuses on physical microsensors used in the medical field that are based on the capacitive approach.

Historically, silicon piezoresistive devices were first introduced in the early 1960s to monitor pressure related variations (1). Piezoresistive silicon sensors take advantage of the piezoresistance effect observed in Si and Ge crystals (2). During this phenomenon, the value of a resistor realized from these semiconductors changes when mechanical strain is applied.

Piezoresistive devices have since then been investigated and used with catheters or as implantable units to monitor blood pressure variations (3,4). Such devices are already present in the market (5,6). Problems related to the nature of the piezoresistance effect include the relatively important temperature sensitivity of piezoresistive sensors that have to be compensated for with rather sophisticated electronics as well as the increased power consumption when compared to capacitive devices (7). These drawbacks of piezoresistive devices have accelerated the investigation of other device options for pressure measurements. It was in the 1980s that research on capacitive sensors began.

THE CAPACITIVE APPROACH

The capacitive approach is based on the parallel plate capacitor principle. In that structure the capacitance between two parallel electrodes is given by

$$C = \varepsilon_r \varepsilon_0 \frac{A}{d} \tag{1}$$

where ε_r, ε_0 is the relative and vacuum permittivity constant, respectively, A is the plate surface area, and d is the plate distance.

When an applied external force acts in a way to change the distance between the two electrodes, the displacement is detected as a capacitance change. In many applications when a micromechanical structure is realized to detect a capacitance change, one of the two electrodes remains fixed and the other is flexible (Fig. 1a). The flexible electrode that is rigidly supported on its edges deflects so it does not remain parallel to the fixed electrode (Fig. 1b). In that case, the previous simple expression is modified and the capacitance is given by

$$C = \varepsilon_r \varepsilon_0 \int_A \int \left(\frac{1}{d - w(x,y)} \right) \tag{2}$$

where $w(x,y)$ is the displacement of the flexible electrode.

Since the measured capacitance depends on the distance between the two electrodes, it can be used to calculate

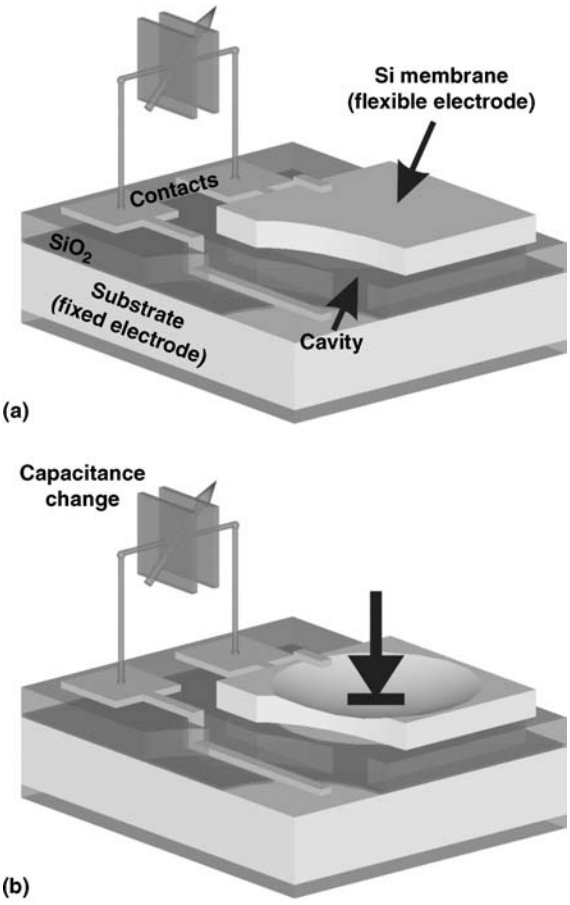


Figure 1. An example of a capacitive sensor is a pressure sensor. In parts a, the thin sensor diaphragm remains parallel to the fixed electrode and in part b, the diaphragm deflects under applied pressure resulting in capacitance change.

the value of the stimulus modifying the interelectrode spacing. Different approaches for capacitive micro devices have appeared that make use of the above principle.

One major application in the medical field for capacitive transducers is pressure measurement. First explorative research of capacitive microdevices made using silicon technology was initiated in the early 1980s (8–11). These initial studies have demonstrated that capacitive devices exhibit superior properties from their piezoresistive competitors in terms of pressure sensitivity, scalability, manufacturing simplicity, as well as process variation tolerance.

APPLICATIONS IN THE MEDICAL FIELD

The manufacturing advantages of micromachined silicon sensors have made them very attractive in the field of minimally invasive therapy. Catheters used in this field need to penetrate into small blood vessels putting stringent requirements on sensor size, which should be one-half of the size of the vessel to be accessed. In small vessels, catheters usually have a diameter of 2 mm, while the smallest catheter at the moment has a diameter of 0.36 mm (12). Miniaturization, therefore, of both sensor and packaging size drives this kind of application. Capacitive devices within the above size requirements have been demonstrated in silicon (13) and are mainly competing with fiber optic pressure sensors (14), or a free hanging strain gauge (15).

Catheters, however, are unsuitable for long-term monitoring of blood pressure. With AAA (abdominal aortas aneurism) and CHF (congestive heart failure) being the major cardiovascular diseases in our days, an implantable pressure monitoring system that would tailor treatment medication by measuring blood pressure appears very attractive. These implantable applications will require a miniature batteryless telemetric sensor able to communicate with an external handheld unit. In such applications, power consumption becomes an additional issue and capacitive devices offer a clear advantage over piezoresistive ones. These types of systems have been under investigation (16,17) and have already demonstrated good performance.

Capacitive pressure sensors have also been successfully applied in the following areas. (1) Intraocular pressure monitoring (10,18). Glaucoma is, for example, a serious disease characterized by an increased pressure in the eye that may result in blindness. In that case, a sensitive capacitive device has been developed for remote sensing of eye pressure (normally 10–20 mmHg 1.33–2.66 kPa above atmospheric pressure, but much increased in glaucoma). (2) Intracranial pressure monitoring (19), as well as for clinical assessment of prosthetic socket fit (20) and pressure distribution in artificial joints (21). Intracranial pressure monitoring is an implant used in the treatment of patients with trauma of the head as well as in neurological patients.

Although pressure is the field where capacitive sensors have been applied more extensively, there are also other applications under development in the medical field. Accelerometers, for example, are used for measuring inclination of body segments and activity of daily living, with application in patient rehabilitation (22), but also register the kind of movements that occur in healthy persons during normal standing (23,24). Physical activity as well as energy expenditure as these can be followed by accelerometers proves to be useful information for personal status monitoring. An accelerometer design includes the fabrication of a proof mass that is displaced in proportion with acceleration. The use of capacitance to measure that displacement significantly improves sensitivity.

Recently, ultrasound imaging technology has also exploited the advantages of capacitive sensors for both transmission and detection purposes (25). Such capacitive devices can be batch fabricated to form a transducer array with array elements that can be as small as 50 μm diameter. Ultrasound devices are made of a thin flexible electrode facing a rigid electrode. For transmission purposes, the membranes are driven into vibration by the electrostatic force exerted between the two electrodes. For reception, the membrane vibration is excited by an impinging acoustic wave that is converted by the capacitive device to electrical signal. This as well as other efforts (26,27) are driven from the need to obtain in the future high resolution images within the body using three-dimensional 3D echographic probes.

Miniature capacitive transducers, known in low frequency applications as condenser microphones, are used in hearing aids and have been reported from different research groups (28,29). In their work, Rombach et al. developed a low noise capacitive microphone with higher sensitivity and broader bandwidth than those used in traditional hearing aids. This device consists of two backplates with an intermediate membrane made of a low stress silicon-rich nitride and B^+ polysilicon multilayer. Pedersen et al., on the other hand, presented an integrated capacitive microphone based on polyimide technology and realized by postprocessing on a CMOS wafer at low processing temperatures.

Capacitive sensors have also been used as humidity sensors for the diagnosis of pulmonary diseases. In this device, a chemically absorbing layer (usually a polymer) is placed between the parallel electrodes of a capacitor. Then, humidity is detected as a change in capacitance because of the dielectric constant change as water molecules are absorbed in the polymer. In a similar configuration, hydrogel has been used between the electrodes of a capacitive sensor to measure body analytes from the capacitance variation occurring due to the swelling of the polymer (30).

FABRICATION TECHNOLOGIES FOR CAPACITIVE SENSORS

Silicon is widely accepted as the material of choice for microsensor fabrication. Known for its good mechanical properties, high mechanical strength, and light weight it's an ideal material for physical sensors (31). More importantly, the existing know-how from IC manufacturing made the use of silicon technology for transducers quite straightforward during the last decades.

This section describes the main silicon technologies that are used for the fabrication of capacitive devices. Most of

the capacitive devices used in medical technology, namely, pressure sensors, accelerometers, and ultrasound sensors, have been developed using two major technology platforms.

Bulk Micromachining Technologies

Bulk micromachining has historically been developed first. It consists of engineering a silicon wafer by a series of lithographic processes followed by wet or dry etching, in order to form thin membranes or other free standing structures that can move upon an external stimulation. In bulk micromachining, the micromechanical silicon structures are fabricated by selectively removing whole sections, or in some cases, all but a small part of the silicon wafer. Thus the structures fabricated in this way are made of single- crystal silicon and have excellent mechanical properties. In bulk micromachining, processing may take place in either the front or the back side of the silicon wafer.

Wet etching of silicon is based on a chemical reaction of silicon with a base solution (KOH) in order to remove silicon material and form the intended 3D structure. Dry etching is a physicochemical process that was initially developed for the etching of thin films. During the last decade, however, this technique has also been used for the removal of thick Si material (32).

The above techniques combined with others usually used in IC manufacturing, like ion implantation, thermal processing, and thin-film (metal or silicon insulator) deposition, constitute the backbones of capacitive sensor fabrication. A detailed description of these processes is beyond the scope of this article and can be found elsewhere (33).

Apart from these processing technologies it is necessary in many applications to use other specific processes. For example, in pressure sensor fabrication it is always desirable to have a reference pressure enclosed in the sensor body. This requires a reliable technology for sealing a cavity with known pressure. For that reason, bulk micromachining techniques are combined with technologies usually referred to as bonding technologies. Two are the most established technologies in this area: anodic bonding and fusion bonding, in chronological order of their discovery and application. Anodic bonding is achieved between silicon and a glass substrate at medium temperature ($\sim 400\,^{\circ}C$) under the application of a direct current (dc) voltage across the two substrates (34). Prior to contact of the two substrates, their surfaces are adequately cleaned to remove any particle that can inhibit their good contact. Bulk micromachining combined with anodic bonding has been successfully used for the realization of medical capacitive pressure sensors by a couple of research groups (35,36). In the process developed at the University of Michigan, boron etch-stop techniques and silicon-glass anodic bonding is used to fabricate a capacitive pressure sensor. In this process, KOH is used to initially form a recess in the surface of a silicon wafer, followed by a deep boron diffusion to define the rim of the transducer, and a shallow diffusion defining the eventual thickness of the diaphragm. The completed silicon wafer is finally electrostatically bonded to a glass wafer. The silicon wafer is then dissolved in EDP letchant, leaving only the silicon transducer islands bonded to the glass. In this way, a thin silicon membrane structure over a sealed cavity is fabricated that exhibits high pressure sensitivity adequate for use in blood pressure monitoring.

A more recently discovered technique that has been applied for sealing of a pressure reference cavity is fusion bonding. This technique does not require the application of a voltage across the bonded substrate. Instead, it includes a high temperature heat treatment. So after a thorough wafer cleaning and drying process of two silicon wafers that renders their surface hydrophilic, they are brought into contact at room temperature. The two wafers are initially drawn together at room temperature by van de Waals forces developed between the hydrogen atoms of water molecules covering their surfaces. This initial attraction is commonly known as prebonding. Prebonding follows a high temperature heat treatment that during which a hydrogen atom is removed transforming the hydrogen bonds to covalent bonds between oxygen atoms thus increasing the bonding strength (37). The temperature necessary to achieve high strength bonding is $> 800\,^{\circ}C$. A successful example of this technology together with bulk micromachining is applied for the realization of a capacitive-type pressure sensor for blood pressure monitoring developed by Goustouridis et al. (13). This simple process results in robust capacitive sensors with low parasistics (Fig. 2). It involves two silicon wafers that are silicon fusion bonded to form the final 3D structure, and a thick oxide in between in which a sealed cavity is formed. The sensor diaphragm is formed by creating a heavily boron-doped region in the cavity bottom. After bonding, the wafer stack is first mechanically ground and then etched in EDP etchant to leave the sensor diaphragm on top of the cavity thus creating the pressure sensor. The sealed cavities are then metalized and packaged. Figure 3 depicts the complete pressure sensing element as it appears after the metallization step.

Hydrophobic bonding is also applied when we need to bond two bare silicon surfaces without a SiO_2 layer on the surface. This technique requires an HF final step to remove any oxide layer from the surface. In hydrophobic bonding, the water molecules necessary to complete the prebonding step are substituted from the HF molecules, while the rest of the bonding process is similar to the hydrophilic one.

More recently, plasma activated bonding (38) or other low temperature bonding ($< 400\,^{\circ}C$) using spin on glass (SOG) (39) have been applied. These techniques are more appropriate for wafer level packaging and have to be more developed in the future for use in sensor fabrication.

Recently, wafer bonding has also been used for demonstration of capacitive ultrasound devices (25), as well as of capacitive accelerometers (40).

Surface Micromachining and SOI Technologies

Surface micromachining technologies have been developed with the primary goal of cointegrating the electronic and mechanical parts on the same silicon chip. Since the active and passive layers of a surface micromachined structure are realized using the same conductive and insulating thin-film layers as in IC manufacturing, it is possible by appropriately designing the mask sequence to realize both

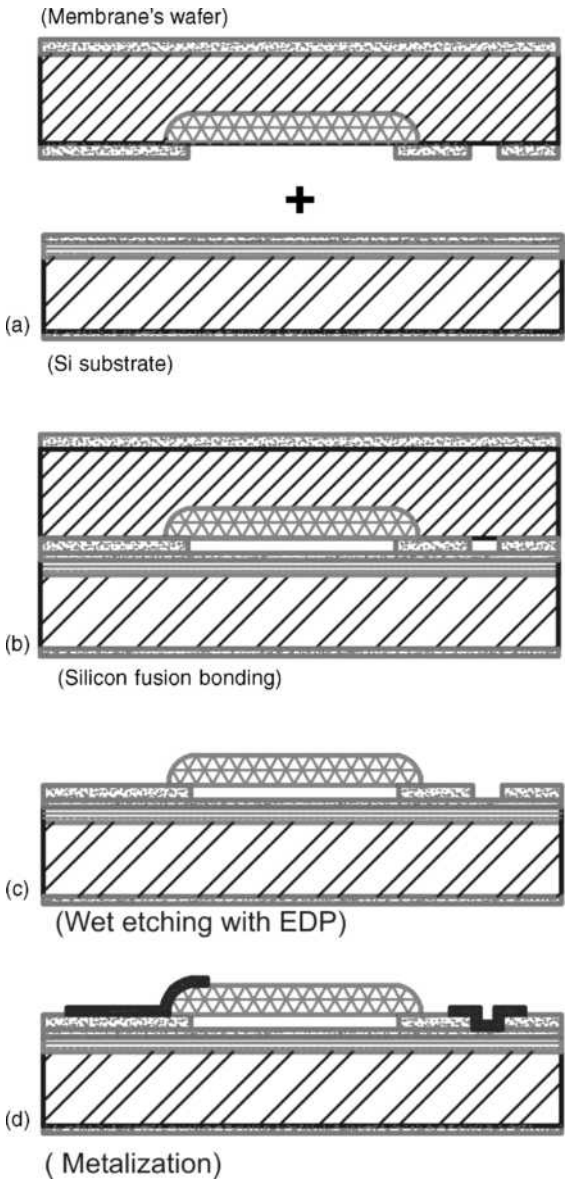


Figure 2. Fabrication of capacitive sensors using bulk micromachining and silicon fusion bonding. Two wafers are used for this purpose. (a) Each wafer is processed independently before the bonding; (b) after the bonding process the two wafers are permanently stacked together; (c) selective wet etching releases the boron doped silicon diaphragms; (d) metallization is performed for electrical connects of the membrane and the substrate.

components in parallel by adding a few mask levels after the completion of the electronic circuit. This particular feature is behind the drive for the development of this technology.

In the case of surface micromachining, all of the processing takes place only on the surface of the front side of the silicon wafer. The micromechanical structures fabricated in this manner are made out of polycrystalline silicon deposited by low pressure chemical vapor deposition (LPCVD) techniques over a sacrificial layer. The sacrificial layer is a deposited oxide, usually phosphorsilicate glass (PSG). This layer is subsequently removed by an HF solution through narrow access channels to release the polysilicon structure (Fig. 4). Sealing of the channels is necessary and it is realized by a deposition step. With surface micromachining, it is possible to fabricate far smaller microelectromechanical devices than with bulk micromachining. Depositing processes allow for very good control over the dimensions of the deposited materials.

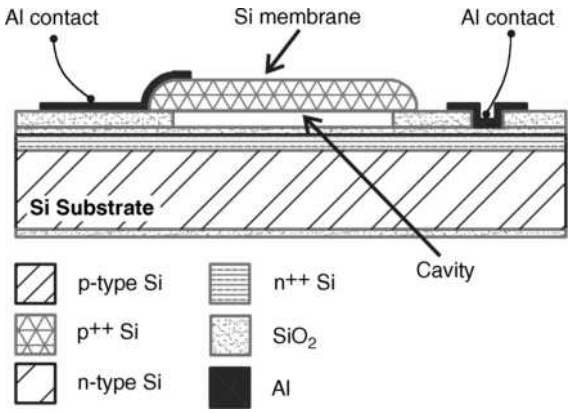


Figure 3. This figure illustrates the final capacitive pressure sensor structure (not to scale).

Surface micromachining is a very important technology with a demonstrated potential. Problems related to stiction of membrane structures on the substrate during aqueous removal of sacrificial layers have been resolved by the application of other etching and drying techniques like gas-phase etching or freeze-drying techniques using cycloexane (41). A typical process sequence developed (42) for capacitive ultrasound transducers is shown in Fig. 5.

There is continuing discussion on the advantages–disadvantages of the cointegration of sensors with electronics. Although it is considered that surface micromachining can result in a higher packing density, and consequently smaller and cheaper components, it appears that yield issues still need to be overcome until this technology can be definitively adopted in preference to hybrid fabrication technologies.

A recent variation of the two technologies employs silicon-on-insulator (SOI) technology with some unique features. The SOI technology offers the possibility of using crystalline silicon as the active part of a capacitive-type structure with more predictable mechanical behavior than the polysilicon film used in surface micromachining. In fact, the internal stresses (either compressive or tensile) developed during the growth of the polysilicon film, which evolved during subsequent thermal treatment because of change of the grain size, is a source of potential uncertainties for the behavior of these films.

Silicon-on-insulator is a mature technology and nowadays can offer crystalline silicon structures of varying thicknesses over a variety of SiO_2 buried layer thicknesses. This has become possible after the discovery of wafer bonding technology, which enables the development of back-etch-silicon-on-insulator (BESOI) structures.

Thin as well as thick silicon structures allow for capacitive pressure sensors development, ultrasound capacitive sensors using rather thin membranes of some micron thickness as well as capacitive-type accelerometers, where

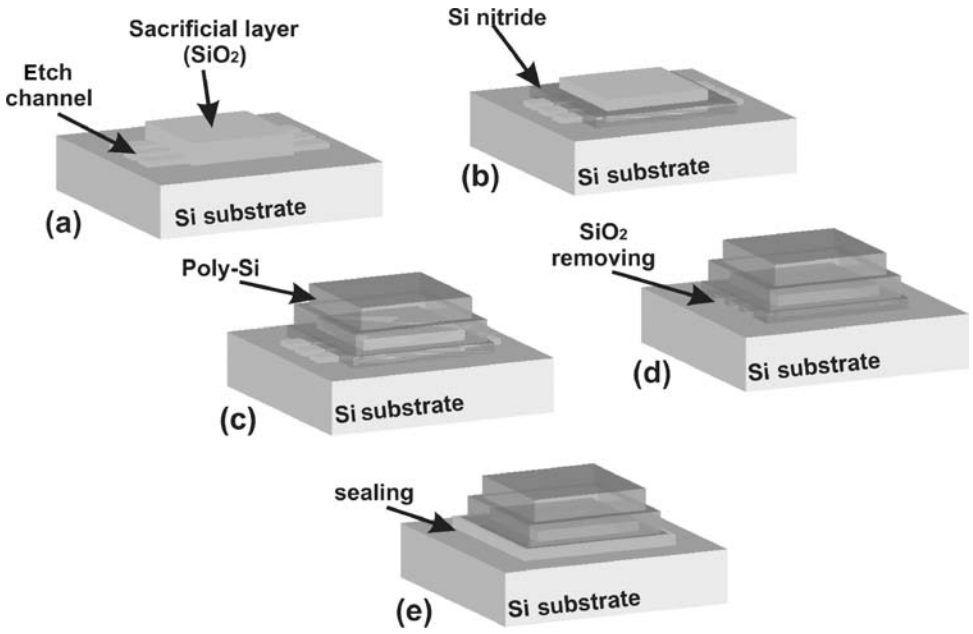


Figure 4. Typical surface micromachining process. (a) Deposition of the sacrificial layer (silicon oxide) and etch channels formation, (b) deposition of silicon nitride ring, (c) deposition of polysilicon film, (d) removal of the sacrificial layer by HF in wet or vapor form, (e) sealing of the device with deposited silicon oxide.

Si structures exceed some hundreds of micrometers. The introduction of deep reactive ion etching (DRIE) in silicon processing especially during the last few years, has enabled the vertical anisotropic etching of Si at the rate of several microns per minute, and the fabrication of thick proof masses on BESOI or glass structures for capacitive accelerometers (43). This accelerometer uses a silicon proof mass of 0.5 mg with 120 μm thickness formed by DRIE and measures in plane (x or y axis) acceleration. It uses a sense gap of only 2 μm between sense fingers and the electrodes (Fig. 6). As the proof mask moves under acceleration, the distance between sense fingers and fixed electrodes change, which consequently modifies the capacitance.

Finally, a different but similar approach results in capacitive-like pressure sensors based on a field effect transistor (44). In this case, the usual dielectric of a MOSFET is replaced by a vacuum cavity. The external pressure variations deflect the gate electrode, and consequently influence the capacitive coupling of the flexible membrane (gate) with the channel thus modifying the current flow in the device. Although these devices present an attractive design, there have not seen any new developments.

OPERATION ISSUES OF CAPACITIVE SENSORS

As introduced in the first paragraph a capacitive sensor is an equivalent parallel plate capacitor with clamped edges where the diaphragm deforms during the application of a deferential pressure across the two sides in the case of capacitive pressure devices.

In the normal operation mode of a capacitive pressure sensor, the diaphragm does not contact the substrate electrode. The capacitance response of a typical pressure sensor is shown in Fig. 7. The output capacitance is nonlinear because of its inverse relationship with the electrode gap d_0-w (eq. 2), which is a function of pressure, P. This nonlinearity becomes more significant for large membrane deflections. At the point when the sensor diaphragm touches the cavity bottom (Fig. 7), the behavior of the sensor changes and enters "touch mode" operation (45). In this operating region, linearity increases and the sensor capacitance is dominated by the area touching the cavity bottom, since the gap there is replaced from the very thin and high dielectric constant SiO_2 layer of the bottom electrode.

Sensitivity of the Capacitive Sensor

Because of the nonlinearity of the response, the sensitivity of a capacitive sensor is not constant. For example, the sensitivity of the capacitive pressure sensor for blood pressure monitoring (defined as $1/C_0(\Delta C/\Delta P)$ with the response shown in Fig. 7 is 1.5 fF/mmHg for the low pressure range and increases to $>$ 18 fF/mmHg for the upper part of the measurement range ($>$ 200 mmHg, 26.66 kPa). In the touch mode operation region, ($>$ 300 mmHg, 39.99 kPa) the sensor has a nearly linear response with a sensitivity of 12 fF/mmHg.

Temperature variation, although not a critical issue for medical application (because of the small variation of the body temperature), can affect the accuracy of the measurements. The temperature influence on the capacitive sensor response is due to either the mismatch of thermal expansion of dissimilar materials used in the fabrication process (usually Si and SiO_2), or to gas expansion in case gas is trapped in a sealed cavity of the sensor. It can of course be due to both of the above reasons. In the case of a sealed cavity, the temperature influence, because of the expansion of the gas trapped inside the cavity, can be eliminated if the cavity is sealed in vacuum. On the other hand, the effect of the different thermal expansions of the materials used is always a problem that requires appropriate design in order to reduce or even eliminate the effect. In the cases of capacitive devices fabricated with surface micromachining,

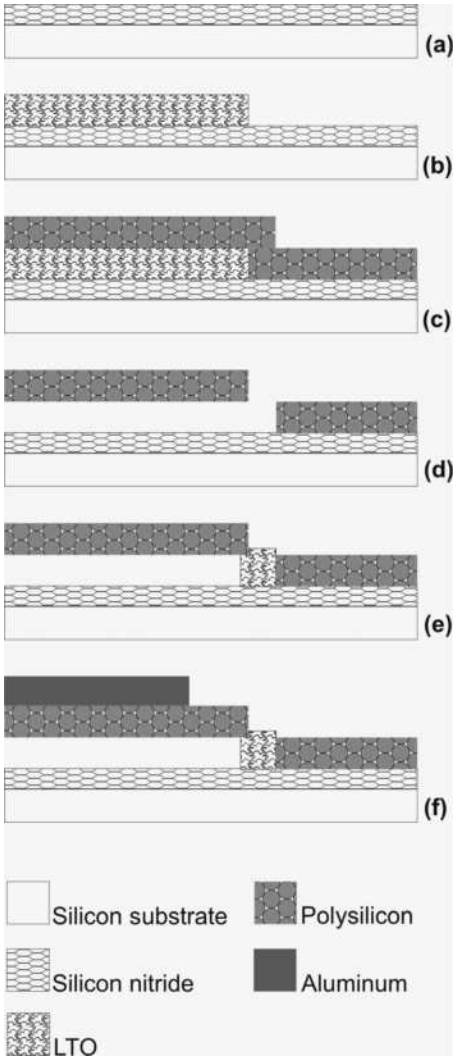


Figure 5. Surface micromachining process sequence for the fabrication of capacitive ultrasonic transducers as taken from X. Jin et al. (42).

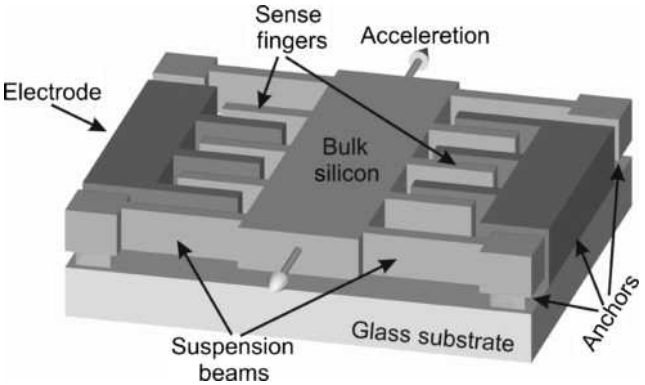


Figure 6. Comb-shaped accelerometer structure fabricated using a combination of processes like bonding silicon with glass substrate and Deep Reactive Ion Etching (43).

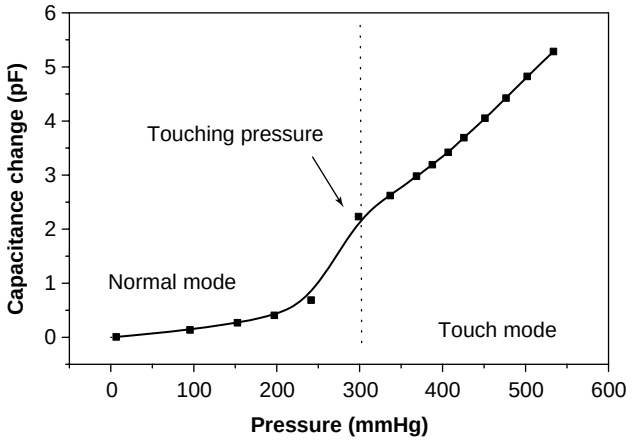


Figure 7. Typical response of a capacitive pressure sensor with 325 μm OD circular diaphragm. The diaphragm touches the cavity bottom at 300 mmHg (∼ 40 kPa). A pressure sensor must be designed to operate either for lower values than 300 mmHg or for higher pressure values. Otherwise an hysteresis phenomenon can be observed.

or by using anodic bonding of silicon with a glass substrate, the influence of temperature becomes more complicated.

Figure 8 shows the influence of temperature on the response of a capacitive pressure sensor, not sealed in a vacuum. The variation of the distance between the curves with temperature represents the temperature coefficient of the pressure sensitivity (TCS) while the slope of the lower curve is the temperature coefficient of zero pressure offset (TCO). The TCS is defined as $\Delta S/S\Delta T$, where S is the pressure sensitivity defined as $\Delta C/C_0\Delta P$ and TCO is defined as $\Delta C/C_0\Delta T$.

This simple configuration of a parallel plate capacitor is used mainly in pressure sensors and ultrasound imaging devices. Accelerometers are usually designed with comb-shaped electrodes that result in increased linearity and sensitivity (40,43).

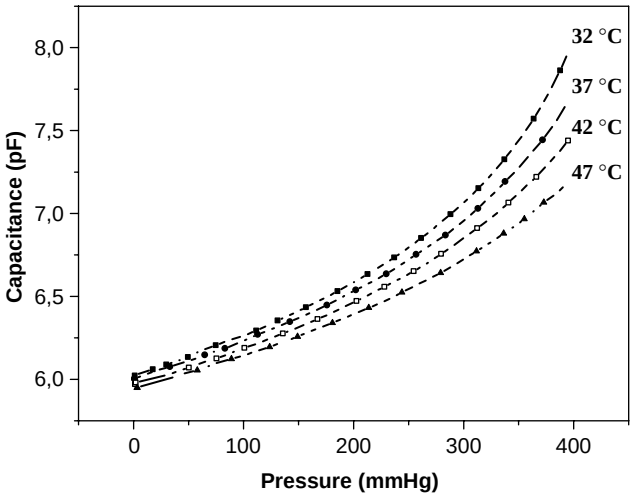


Figure 8. Temperature variation of a capacitive pressure sensor in the range 0–400 mmHg (∼53.3 kPa). The effect is due mainly to trapped gas expansion inside the cavity.

CAPACITIVE SENSOR ELECTRONIC INTERFACES

Although capacitive sensors offer advantages with respect to high sensitivity and low power operation, they also present difficulties in the design of electronic interfaces to convert capacitance changes into electrical signals. Parasitic capacitances often dominate system performance by reducing sensitivity and increasing nonlinearity (46). Therefore, it becomes absolutely essential for sensor systems incorporating capacitive sensors to either integrate the MEMS component with the electronic interface or place the two chips at close proximity to each other to reduce parasitics. At the same time, it is important that the electronic interface is designed to suppress the remaining parasitic capacitances and provide for a large zero capacitance range. A good starting point for the study of capacitance measurement circuits may be found in Ref. 47, where a number of basic circuits are discussed.

Depending on the technological steps used to fabricate the sensor, several parasitic capacitances $C_{p1,2}$ and conductances G_p may be present in the device and increase in importance as devices get smaller and the sensing capacitance C_s gets in the pF range. These parasitic effects may originate from various sources depending on the fabrication process used (i.e., stray capacitances of metal lines and connecting pads in parallel with the sensing capacitance or leakage resitors). Further variations may also be observed, due to technological reasons, between batches of the same sensor. In Fig. 9, a simple electrical model of a typical capacitive sensor, including parasitic effects, is shown. The spread in sensor characteristics (e.g., zero capacitance, sensitivity) resulting from these effects puts a great strain in the design of electronic interfaces for capacitive sensors. For this reason, most of the capacitive interfaces to date have been designed by taking into consideration the particular sensor and application that they are going to be used for. The design is a trade-off between power consumption, interface accuracy, and resolution, or even die size in some applications in the medical field.

A number of architectures have been proposed to date to convert capacitive changes into electrical signal. Some are built around a relaxation oscillator (46,48), others use switch capacitor techniques to convert capacitance changes into voltage, and others interface the sensor directly into a sigma–delta modulator. A few attempts to develop a generic interface have also been reported (49,50). However, the power consumed is too high for remote sensing applications.

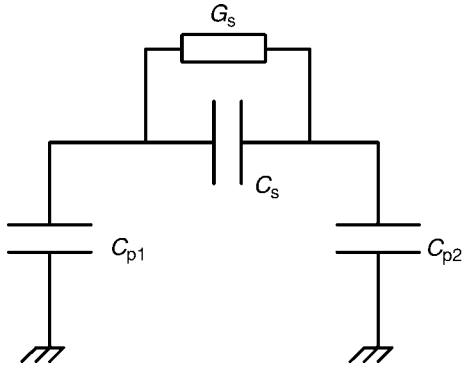


Figure 9. Simple electrical model of a capacitive sensor. Parasitic capacitances is designated by $C_{p1,2}$, conductances by G_p and the sensing capacitance by C_s.

Van Der Goes and Meijer (49) presented a universal transducer interface for the read out of capacitors, platinum resistors, thermistors, resistive bridges, and potentiometers. The circuit uses the three signal technique in which the sensor signal E_x, a reference signal E_{ref}, and the offset E_{off} of the whole interface are measured in a identical way to achieve continuous autocalibration of offset and gain. The two port measurement technique is used to eliminate sensor parasitic capacitance. In this technique, a testing voltage, V, is forced on one capacitor electrode while current I is sensed on the other (51). The current I then depends only on the applied voltage and sensing capacitance. However, the technique is not energy efficient, since it requires four measurement cycles and three external voltage sources to determine the sensor capacitance.

Yazdi et al. (50) also developed a standardized switch capacitor interface for capacitive sensors, as shown in Fig. 10. This interface is capable of interfacing through a standard bus with a microcontroller that collects sensor data through the interface, calibrates data, and either stores or transmits it wirelessly or through a serial port. The readout circuit utilizes a low noise front-end charge integrator to read out the difference between the sensor capacitance and a reference capacitor. An input multiplexer allows for interfacing with up to six capacitive sensors. Finally, the chip can be digitally programmed to operate with one of three external or internal reference laser trimmable capacitors.

Bracke et al. (52) reports on a low power generic switched capacitor interface for capacitive sensors. The circuit uses a special clocking scheme, in which the analog sensor circuit block is clocked at a low 8 kHz frequency, while the sigma–delta modulator is clocked at 128 kHz. The technique ultimately reduces power consumption to 90 μW on the ON-state.

CAPACITIVE ELECTRONIC INTERFACES FOR IMPLANTABLE APPLICATIONS

In medical applications, where the diagnostic device is to be implanted inside the human body, only limited power is available for the electronic circuits and sensor. In such cases, power can only be found via a battery implanted together with the sensing system or through passive telemetry. In passive telemetry, energy may be harvested from a remote electromagnetic field transmitted outside the body. The same field may also be used to receive control data and transmit sensor data to a data logger. In both cases, minimizing power consumption is essential.

A simple low power interface for biomedical applications could be realized by using a simple relaxation oscillator (46,48). A capacitance-to-frequency modulated output was first proposed by Hanneborg et al. (46). This circuit delivers a digital pulse trail with frequency dependent on the sensor

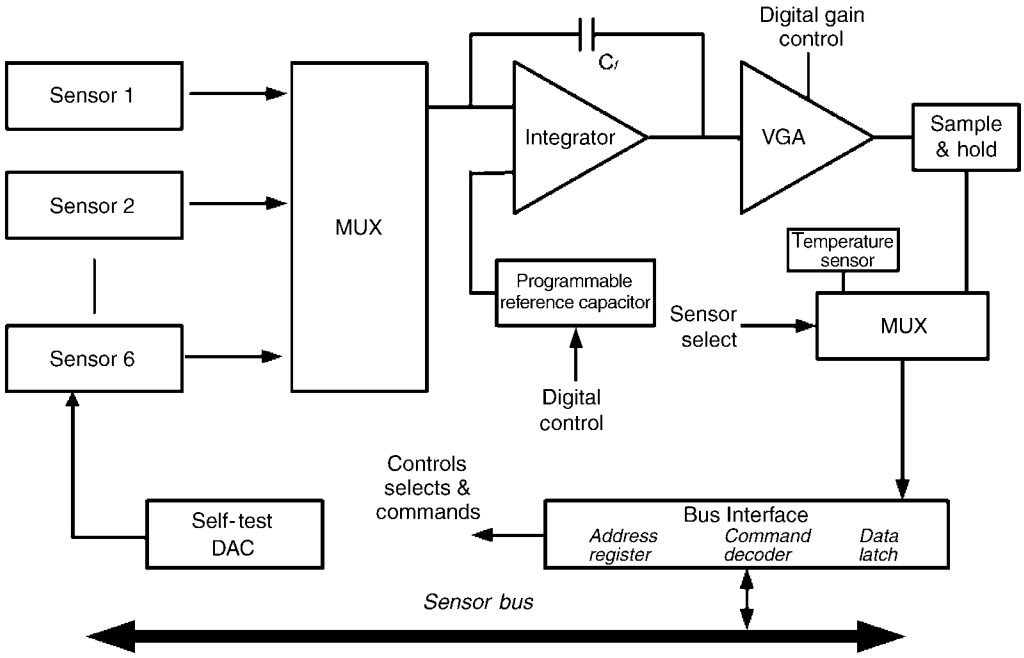


Figure 10. Circuit blocks of the capacitive electronic interface as described by W. Bracke et al. (50).

capacitance. The circuit consists of two current sources: a switch and a Schmitt trigger. The unknown capacitance is periodically charged and discharged by flipping the switch between current $I+$ and $I-$ (Fig. 11).

An implementation of this type of capacitance-to-frequency converter is presented in Fig. 12. The circuit converts capacitance at its input into a frequency signal that can be readily fed to a digital microcontroller for processing. Either a capacitive sensor or a reference may be switched in, and the frequency delivered at the circuit output is dependant on the input capacitance according to

$$f = \frac{I_0}{2C_x V_h} \tag{3}$$

where I_0 is the current by which the sensor capacitance C_x is charged or discharged, and V_h is the hysteresis of the Schmitt trigger. The parameter V_c is a control signal that allows for switching between the unknown sensor capacitance C_x and a reference capacitor C_{ref}. The output frequency when the reference is selected is independent of pressure, thus compensation of temperature and long-term drifts are possible by taking the ratio of the reference frequency and sensor frequency, $F_{\text{ref}}/F_{\text{sens}}$.

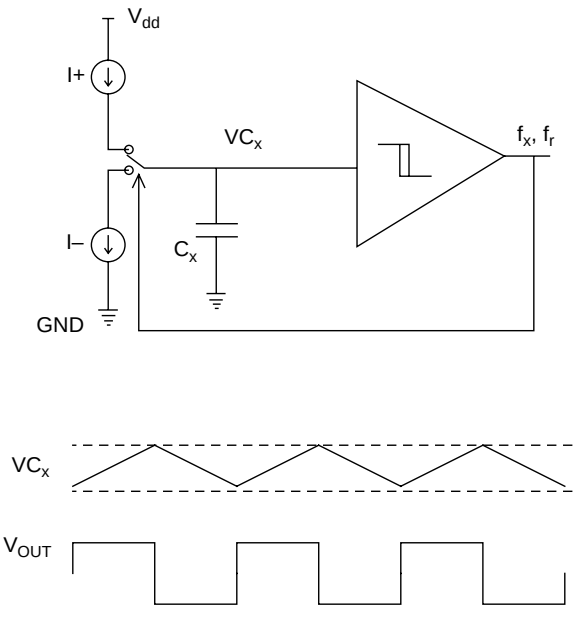


Figure 11. Block diagram of the capacitive to frequency converter proposed by Hanneborg et al. (46). The circuit consists of two current sources, a switch and a Schmitt trigger.

The response of a pressure measuring system based on this circuit realized in 1.2 μm of CMOS technology and a capacitive pressure sensor (53) is shown in Fig. 13. It operates at 4 V and draws 20 μA of average current. The system exhibits a sensitivity of 36 Hz/mmHg and is able to resolve pressure changes of 5 mmHg (0.66 kPa). A photograph of this system is shown in Fig. 14.

In medical science, however, there is often the need for long-term monitoring of vital life parameters. A good example is abdominal aortic aneurysm (AAA), which is a ballooning of the abdominal aorta. Patients who suffer from this condition need to undergo a procedure during which a stent graft is inserted. After the operation, however, it is possible that the aneurysmal sac is not completely isolated, leading to recurrent pressurization of the sac, a complication that, if left undetected, may lead to rupture of the sac and patient death. Long-term, postoperation monitoring of the patients is therefore necessary.

Early efforts for systems for the monitoring of blood pressure (54) used miniature active transmitters to transfer measured data and were battery powered, which limited

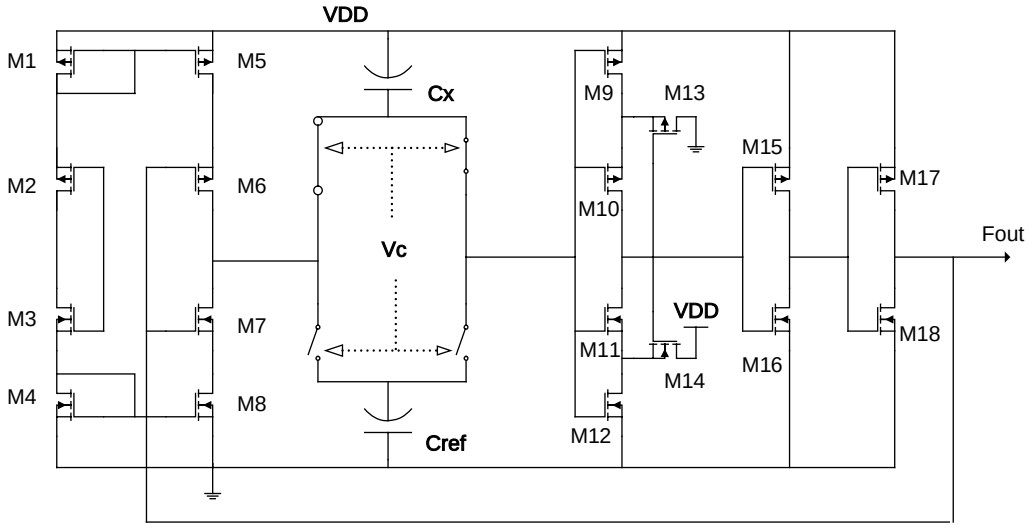


Figure 12. Schematic of Schmitt-trigger based oscillator used as a capacitive interface.

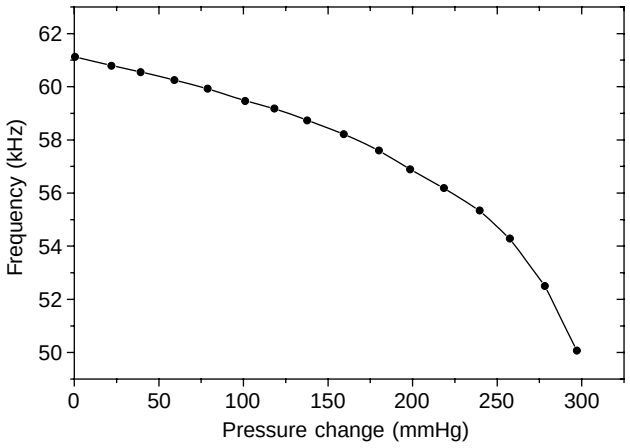


Figure 13. Frequency response of a pressure measuring system consisting of the simple circuit of Fig. 12 and a capacitive pressure sensor in the range 0–300 mmHg (~40 kPa).

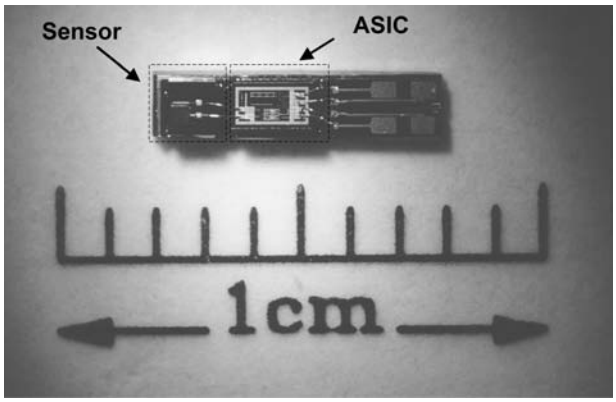


Figure 14. A photograph of a hybrid pressure measuring system consisting of a capacitive sensor and associated signal conditioning electronic circuit.

their lifetime. An alternative approach to power these modules is through induction coupling, while the same radio frequency (RF) field can be used to transfer data out of the implanted module (55). The implanted circuit should be virtually immune to supply fluctuations arising from random misalignment of the implanted and the external coil. A new circuit was developed based on the previous architecture, but in which each circuit block was redesigned.

The block diagram of the passive telemetry system is depicted in Fig. 15. It consists of an external control unit (the base unit) and an implantable transponder. Wireless communication can then be established between the two units, based on an absorption modulation mechanism. The transponder receives power and external control data

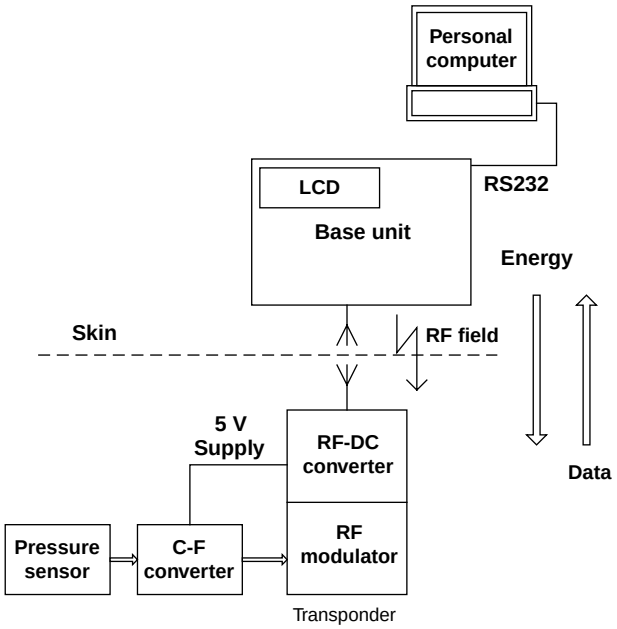


Figure 15. Block diagram of passive telemetry system.

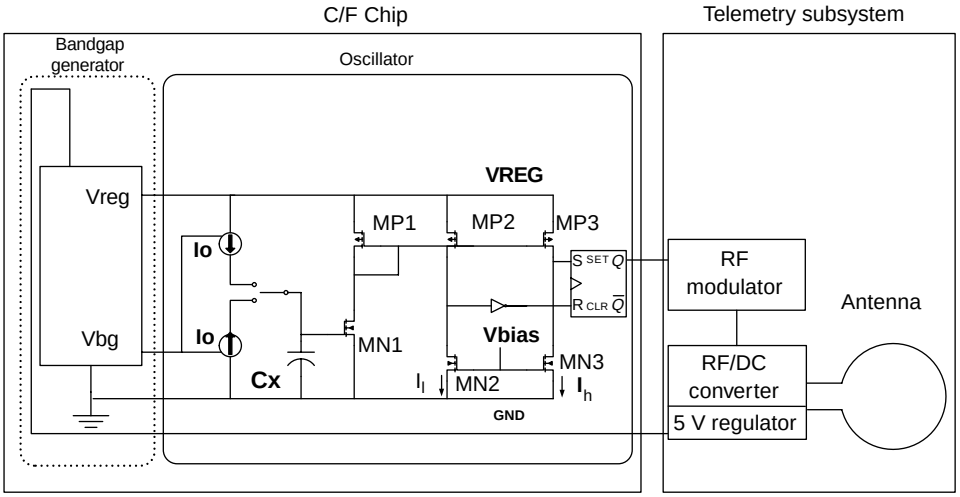


Figure 16. Transponder electronics.

through an RF field, while it can transmit data by modulating the absorption rate. The base unit, on the other hand, demodulates the transmitted data and processes it though a microcontroller to convert the signal into an eight bit unsigned byte array. The resulting byte can then be sent to a PC through a serial output.

The block diagram of the transponder electronics is shown in Fig. 16. The improved capacitance-to-frequency circuit is used to interface a capacitive pressure sensor. The circuit consists of two basic blocks: a bandgap reference voltage generator and an oscillator. In order to achieve independence of the output frequency from the received power, this time the oscillator operates on an internally stabilized voltage generated from a bandgap reference. Voltage regulation is a critical part of telemetric systems as the induced power, and thus the voltage output, of the RF/dc converter in the transponder can greatly fluctuate in an actual implanted system becaused the relative position changes of the two antennas. In addition, to further immune the system from supply fluctuations a current mode comparator is used in the oscillator.

The bandgap reference voltage circuit is capable of operating at a low power supply as it operates on an internally regulated voltage VREG (56). This same node is also used for the supply of the oscillator circuit after the contributions of the extra branches of the oscillator are accounted for.

The oscillator itself is designed around a current mode comparator that results in an output frequency independent of power supply fluctuations and with small temperature drift. Triggering levels of this oscillator are defined by two currents: I_h and I_l. The period of the output pulse can then be shown to be equal to

$$T = \frac{2C_x}{I_0}(V_{\text{bias}} - V_{TN})(\sqrt{n} - 1) \tag{4}$$

where n stands for the ratio of I_h to I_l, and V_t is the threshold voltage. By taking the inverse of eq. 4 and substituting for I_0, we obtain

$$f = \frac{k\frac{W_0}{L_0}(V_{\text{bias}} - V_{TN})}{2C_x(\sqrt{n} - 1)} \tag{5}$$

Equation 5 implies that the output frequency is independent of the supply voltage and is dependent on temperature through the mobility term in k and the threshold voltage V_t. Note also that the bias voltage V_{bias} is chosen to be equal to the bandgap reference voltage produced from the previous stage, and is thus considered independent of voltage and temperature variations.

The *C*/*F* converter was designed and fabricated in 0.8 μm CMOS technology. A hybrid pressure measuring system composed of a capacitive pressure sensor (16) and the *C*/*F* chip. The system remains operational for a supply voltage down to 2.7 V and exhibiting high immunity to voltage variations from 3.7 to 5.5 V (Fig. 17). Simulated pressure pulses as those present in the aorta were measured using passive telemetry (Fig. 18).

CONCLUSIONS

Capacitive microsensors are in progressively increasing use in medical applications because of their advantages, such as small size, high sensitivity, and low power consumption. Silicon is the material of choice for these devices, which are finding applications for measuring pressure and

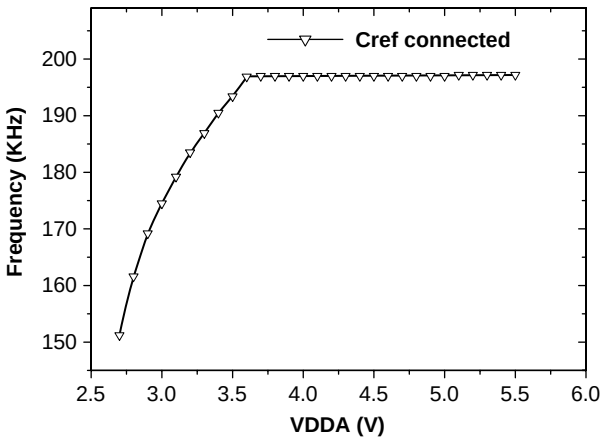


Figure 17. Pressure measuring system frequency output.

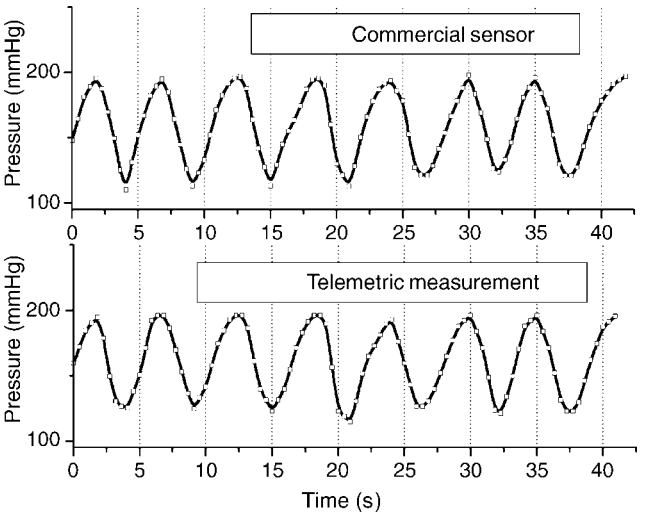


Figure 18. Simulated pressure pulses as those present in the aorta measured using passive telemetry.

acceleration, but also in ultrasound imaging and microphones and for monitoring body analytes. Actually, although most of the existing devices are not integrated together with signal conditioning electronic circuits on the same chip, in the future it is expected that there will be an increasing integration scheme driven by the need of higher level miniaturization. Because of their low power consumption advantage, capacitive microsensors are of interest as implantable monitoring devices. An attractive application appears to be a miniaturized telemetry system that combines techniques for wireless power and data transfer to a capacitive sensor integrated with signal conditioning electronics.

BIBLIOGRAPHY

1. Tufte ON, Chapman PW, Long D. Silicon Diffused-Element Piezoresistive Diaphragm. J Appl Phys 1962;33:3322–3327.
2. Smith CS. Piezoresistance Effect in Germanium and Silicon. Phys Rev 1954;94:42.
3. Samaun X, Wise KD, Angel JB. An IC Piezoresistive Pressure Sensor for Biomedical Instrumentation. IEEE Trans Biomed Eng 1973;BME-20(2):101–109.
4. Ko WH, Hynelek J, Boettcher SF. Development of a miniature Pressure Transducer for Biomedical Application. IEEE Trans Electron Dev 1979;ED-26:1896.
5. Data Science Int., Roseville Minnesota.
6. Konigberg Instruments Inc. Pasadena, California.
7. Clark SK, Wise KD. Pressure sensitivity in anisotropically etched thindiaphragm pressure sensors. IEEE Trans Electron Dev 1979;ED-26(12):1887–1896.
8. Sander GS, Knutti JW, Meindl JD. A monolithic Capacitance Pressure Sensor with Pulse Periodic Output. IEEE Trans Electron Dev 1980;ED-27(5):927–930.
9. Lee YS, Wise KD. A batch-fabricated Silicon Capacitive Pressure Transducer with low temperature Sensitivity. IEEE Trans Electron Dev 1982;ED-29(1):42–47.
10. Backlund Y, Rosengren L, Hök B. Passive Silicon Transensor Intended for biomedical, Remote Pressure monitoring. Sens Actuators 1990;A21(1–3):58–61.
11. Puers R, Peeters E, Van den Bossche A, Sansen W. A capacitive pressure sensor with low impedance output and active supression of parasitic effects. Sens Actuators 1990;A21 (1-3):108–114.
12. Goosen JFL. Design considerations for silicon sensors for use in catheters and guide wires. Smart Mater Struct 2002; 11:804–812.
13. Goustouridis D, Chatzandroulis S, Normand P, Tsoukalas D. A miniature self-aligned pressure sensing element. J Micromech Microeng 1996;6:33–35.
14. Zhu YZ, Wang AB. Miniature fiber optic pressure sensor. IEEE Photonic Technol 2005;17(2):447–449.
15. Melvås P, Kälvesten E, Enoksson P, Stemme G. A free-hanging strain-gauge for ultraminiaturized pressure sensors. Sens Actuators A 2002;97(8):75–82.
16. Chatzandroulis S, Tsoukalas D, Neukomm PA. A Miniature Pressure System with a Capacitive Sensor and a Passive Telemetry Link for Use in Implantable Applications. J MicroElectroMech S 2000;9(1):18–231.
17. Najafi N, Ludominky A. Initial Animal Studies of a Wireless, Batteryless, MEMS Implant for Cardiovascular Applications. Biomed Microdevices 2004;6(1):61–65.
18. Coosemans J, Catrysse M, Puers R. A readout circuit for an intra-ocular pressure sensor, Sens. Actuators A 2004;110 (1-3):432–438.
19. Hierold C, et al. Low power integrated pressure sensor for medical application. Sens Actuators A 1999;73(1–2):58–67.
20. Polliack AA, et al. Laboratory and clinical tests of a prototype pressure sensor for clinical assessment of prosthetic socket fit. Prosth Ortho Inter 2002;26(1):23–24.
21. Müller O, Parak WJ, Wiedemann MG, Martini F. Three-dimensional measurements of the pressure distribution in artificail joits with a capacitive sensor array. J Biomech 2004;37:1623–1625.
22. Luinge HJ, Veltink PH. Inclination measurement of human movement using a 30D accelerometer with autocalibration. IEEE T Neur Sys Reh 2004;12:112–121.
23. Puers R, Reyntjen S. Design and processing experiments of a new miniaturized capacitive triaxial accelerometer. Sens Actuators A 1998;68(1-3):324–328.
24. Lötters JC, Olthuis W, Veltink PH, Bergveld P. Design, realization and characterization of a symmetrical triaxial capacitive accelerometer for medical application. Sens Actuators A 1997;61(1-3):303–308.
25. Huang Y, et al. Fabricating capacitive micromachined ultrasonic transducers with wafer-bonding technology. J Microelectromech S 2003;12(2):128–137.
26. Cianci E, et al. One-dimensional capacitative micromachined ultrasonic transducer arrays for echographic probes. Microelect Eng 2004;73–74:502–507.
27. Knight J, McLean J, Degertekin FL. Low Temperature Fabrication of Immersion Capacitive Micromachined Ultrasonic Transducers on Silicon and Dielectric Substrates. IEEE Trans UFFC 2004;51:1324–1333.
28. Rombach P, Müllenborn M, Klein U, Rasmussen K. The first low voltage, low noise differential silicon microphone, technology development and measurement results. Sens Actuators A 2002;95:196–201.
29. Pederson M, Olthuis W, Bergveld P. High-performance condenser microphone with fully integrated CMOS amplifier and DC-DC voltage converter. J MicroElectroMech S 1998;7(4): 387–394.
30. Strong ZA, Wang AW, McConaghy CF. Hydrogel-actuated capacitive transducer for wireless biosensors. Biomed Microdevices 2002;4(2):97–103.
31. Petersen KE. Silicon as a mechanical material. Proc IEEE 1982;70(5):420–457.
32. de Boer MJ, et al. Guidelines for etching silicon MEMS structures using fluorine high-density plasmas at cryogenic temperatures. J MicroElectroMech S 2002;11(4):385–401.

33. Madou M. Fundamentals of Microfabrication: The Science of Miniaturization. 2nd ed. Boca Raton (FL): CRC Press; 2002.
34. Wallis G, Pomerantz DI. Field assisted glass-metal sealing. J Appl Phys 1969;40:3946–3949.
35. Ji J, Cho ST, Najafi K, Wise KD. An ultaminiature CMOS pressure sensor for a multiplexed Cardiovascular Catheter. IEEE Trans Electron Dev 1992;39:2260–2267.
36. Puers R, Van den Bossche A, Peeters E, Sansen W. An implantable pressure sensor for use in cardiology. Sens Actuators A 1990;23:944–947.
37. Tong QY, Gosele U. Semiconductor Wafer Bonding. New York: Wiley-Interscience; 1999.
38. Henttinen K, Suni I, Lau SS. Mechanically induced Si layer tranfer in hydrogen-implanted Si wafers. Appl Phys Lett 2000;76:2370–2372.
39. Goustouridis D, et al. Low temperature wafer bonding for thin silicon film transfer. Sens Actuators A 2004;110: 401–406.
40. Tsuchiya T, Funabashi H. A z-axis differential capacitive SOI accelerometer with vertical comb electrodes. Sens Actuators A 2004;116:378–383.
41. Kim C, Kim JY, Shridharan B. Comparative evaluation of drying techniques for surface micromachining. Sens Actuators A 1998;64:17–26.
42. Jin X, et al. Fabrication and characterization of surface micromachined capacitive ultrasonic immersion transducers. J Microelectromech S 1999;8:100–114.
43. Chae J, Kulah H, Najafi K. A CMOS-compatible high aspect ratio silicon-on-glass in-plane micro-accelerometer. J Micromech Microeng 2005;15:336–345.
44. Lysko JM, Jachowisz RS, Krzycki MA. Semiconductor pressure sensor based on FET structure. IEEE T Instrum Meas 1995;44:787–790.
45. Ko WH, Wang Q. Touch mode capacitive pressure sensors. Sens Actuators A 1999;75:242–251.
46. Hanneborg A, et al. An integrated capacitive pressure sensor with frequency-modulated output. Sens Actuators 1986; 9(4):345–351.
47. Senturia S. Microsystem Design. Boston: Kluwer Academic Publishers; 2000.
48. Matsumoto Y, Esashi M. Integrated silicon capacitive accelerometer with PLL servo technique. Sens Actuators A 1993; 39:209–217.
49. Van Der Goes FML, Meijer GCM. A universal transducer interface for capacitive and resistive sensor elements. Analog Integr Circuits Signal Process 1997;14:249–260.
50. Yazdi N, Mason A, Najafi K, Wise KD. A generic interface chip for capacitive sensors in low-power multi-parameter Microsystems. Sens Actuators A 2000;84:351–361.
51. Van Der Goes FML, Meijer GCM. A novel low-cost capacitive-sensor interface. Trans Instr Meas 1996;45(2):536–540.
52. Bracke W, Merken P, Puers R, Van Hoof C. On the optimization of ultra low power front-end interfaces for capacitive sensors. Sens Actuators A 2005;117(2):273–285.
53. Chatzandroulis S, Goustouridis D, Normand P, Tsoukalas D. A solid-state pressure-sensing microsystem for biomedical applications. Sens Actuators A 1997;62:551–555.
54. Casadei FW, Gerold M, Baldinger E. Implantable Blood Pressure Telemetry System. IEEE T Biomed Eng 1972;BME-19(5): 334–338.
55. Neukomm PA, Kuendig H. Passive wireless actuator control and sensor signal transmission. Sens Actuators 1990;A21-A23:258–262.
56. Tham KM, Nagaraj K. A low Supply Voltage High PSRR Voltage Reference in CMOS Process. IEEE J Solid-St Circ 1995;30(5):586–590.

See also BIOELECTRODES; INTEGRATED CIRCUIT TEMPERATURE SENSOR.

CARBON. See BIOMATERIALS: CARBON.

CARDIAC CATHETERIZATION. See CORONARY ANGIOPLASTY AND GUIDEWIRE DIAGNOSTICS.

CARDIAC LIFE SUPPORT. See CARDIOPULMONARY RESUSCITATION.

CARDIAC OUTPUT, FICK TECHNIQUE FOR

STEVEN C. FADDY
University of Sydney
Darlinghurst, Australia

INTRODUCTION

Cardiac output (CO) is an important measurement in many medical investigations. It is the amount of blood pumped by the ventricles of the heart and can be defined as the product of stroke volume (SV) and heart rate (HR), where stroke volume is the amount of blood expelled by the ventricle with each contraction and the HR is the number of contractions per minute:

$$CO = SV \times HR$$

Cardiac output gives an indication of ventricular function and is also used in the calculation of a number of flow-dependent parameters, such as cardiac index, systemic vascular resistance, pulmonary vascular resistance, valve areas, and intracardiac shunt ratios.

The Fick technique is the gold standard in CO measurement. It relies on direct measurement of oxygen consumption and expenditure to derive the rate of blood flow throughout the individual.

HISTORY

In 1870, the German physiologist, Adolf Fick (1829–1901), described a novel method of determining cardiac output based on diffusion of respiratory gases in the lungs. This came after almost 30 years of work by Fick and numerous others, who reasoned that diffusion was one of the most essential events within the living organism. In 1855, Fick had published his findings relating to diffusion of gas across a fluid membrane. These became known as Fick's law of diffusion and stated that the rate of diffusion of a gas is proportional to the partial pressures of the gas on either side of the membrane, the area across which diffusion is taking place and the distance over which diffusion must take place. As an aside, Fick also invented contact lenses in 1887.

The 1870 publication by Adolf Fick stated: "It is astonishing that no one has arrived at the following obvious method by which [the amount of blood ejected by the ventricle of the heart with each systole] may be determined directly, at least in animals. One measures how much oxygen an animal absorbs from the air in a given time, and how much carbon dioxide it gives off. During the experiment one obtains a sample of arterial and venous blood; in both the oxygen and carbon dioxide content are measured. The

difference in oxygen content tells how much oxygen each cubic centimeter of blood takes up in its passage through the lungs. As one knows the total quantity of oxygen absorbed in a given time one can calculate how many cubic centimeters of blood passed through the lungs in this time. Or if one divides by the number of heart beats during this time one can calculate how many cubic centimeters of blood are ejected with each beat of heart. The corresponding calculation with the quantities of carbon dioxide gives a determination of the same value, which controls the first (1)."

In simplest terms, cardiac output can be calculated as a ratio of the amount of oxygen consumed through breathing and the rate in which oxygen is taken up by the tissues.

Cardiac Output

= Oxygen consumption/Arteriovenous oxygen dierence

It was not until the 1930s that quantitative measurement of the components allowed confirmation of the Fick equation as a means of calculating cardiac output.

PHYSIOLOGY OF THE FICK TECHNIQUE

Oxygen Consumption (VO_2)

The first step in calculating CO by the Fick technique is to determine the amount of oxygen consumed by the individual over a period of time. This is best done in the resting state so that there is constant oxygen consumption over the collection period. The traditional method is collection of expired gases in a Douglas bag over a period of $\sim$ 3 min. Then, from the volume of expired gas, the oxygen content of the expired gas and the oxygen content of the inspired room air, it is possible to calculate the amount of oxygen taken up by the individual.

Subtracting the oxygen content of expired gas from that of the inspired room air (%v/v or mL of $O_2 \cdot 100\ mL^{-1}$ of the gas) gives the oxygen difference between the inspired and expired gases, expressed in mL of $O_2 \cdot 100\ mL^{-1}$ of expired gas. Applying a factor of 10 gives this figure as milliliters of O_2 per liter of expired gas, which are the units used later in the calculation. Dividing the total volume of expired gas by the collection time gives the minute ventilatory rate, expressed in liters per minute $L \cdot min^{-1}$.

The product of the O_2 difference ($mL \cdot L^{-1}$) and the minute volume ($L \cdot min^{-1}$) is the oxygen consumption (VO_2) expressed in milliliters of oxygen absorbed per minute.

$$VO_2 = (O_{2\,\text{Room air}} - O_{2\,\text{expired}}) \times (\text{volume/time})$$

Example: Inspired $O_2 = 21.0\,mL \cdot 100\,mL^{-1}$ room air
Expired $O_2 = 16.7\,mL \cdot 100\,mL^{-1}$ expired gas
O_2 difference $= 21.0 - 16.7 = 4.3\,mL \cdot 100\,mL^{-1}$
Total volume expired $= 26.1\,L$
Collection time $= 3\,min$
Minute volume $= 26.1\,L \cdot 3\,min^{-1} = 8.7\,L \cdot min^{-1}$
Therefore,
O_2 consumption $= (4.3 \times 10)mL \cdot L^{-1} \times 8.7\,L \cdot min^{-1} = 374\,mL \cdot min^{-1}$

An alternative to the Douglas bag method is the use of a metabolic rate meter with a hood or facemask, a variable-speed blower and a servocontrol loop with an oxygen sensor. This method employs essentially the same principle as the Douglas bag method, but gives a real-time measurement of VO_2. The variable-speed blower maintains a flow of room air through the hood or facemask past the patient into a polarographic oxygen sensor (gold and silver–silver chloride electrode), varying the flow in order to keep the oxygen concentration at the measuring electrode constant. By keeping the oxygen concentration at the measuring electrode constant, the only variable is the flow rate through the system. Under steady-state conditions, this is the only variable determining the oxygen consumption (VO_2).

Although this method provides a real-time measurement of VO_2, thus excluding the need for collection of a Douglas bag, it is still rather time and labor intensive. In addition, it has been suggested that it is difficult to obtain reproducible results and the method gives consistently lower results then the Douglas bag technique.

Arteriovenous Difference

As with oxygen consumption, measurement of oxygen uptake by the body involves measuring blood oxygen content before and after entering the lungs. The arteriovenous oxygen difference (AV_{diff}) is the difference between the content of oxygen (ctO_2) in the oxygenated arterial blood leaving the lungs and the deoxygenated venous blood returning to the lungs (mL O_2 per 100 mL of blood). The AV_{diff} represents the volume of oxygen delivered to meet the body's metabolic demands. Again, this figure is multiplied by 10 to give the AV_{diff} in units of mL O_2 per liter of blood.

$$AV_{diff} = ctO_{2(\text{Arterial})} - ctO_{2(\text{Venous})}$$

Example: Arterial O_2 content $= 19.5\,mL \cdot dL^{-1}$ blood
Venous O_2 content $= 13.2\,mL \cdot dL^{-1}$ blood
$AV_{diff} = 19.5 - 13.2 = 6.3\,mL \cdot dL^{-1} = 63\,mL \cdot L^{-1}$

Typically, a sample from the main pulmonary artery is used for venous blood and a sample from the left ventricle or aorta is used for arterial blood oxygen content measurements.

Cardiac Output

The rate at which oxygen is taken up by the lungs and the rate at which it is taken up by the body is now known from the above calculations. The ratio of these two figures gives the cardiac output. The examples above show that the lungs take up 374 mL of oxygen each minute and that the blood takes up 63 mL of oxygen for each liter that passes through the lungs. How many lots of 63 mL (1 L aliquots of blood) must pass through the lungs to take up 374 mL of oxygen each minute? The answer is 5.9 L of blood must pass through the lungs each minute in order to absorb this amount of inspired oxygen.

Example: $VO_2 = 374\,mL \cdot min^{-1}$
$AV_{diff} = 63\,mL \cdot L^{-1}$
$CO = 374/63 = 5.9\,L \cdot min$

PRACTICAL CONSIDERATIONS FOR USING THE FICK TECHNIQUE

Oxygen Consumption

The measured volume of a gas is affected in part by the ambient temperature and atmospheric pressure in which it is collected. Obviously, these will vary from day to day, leading to a potential source of variation in the calculation of oxygen consumption, and ultimately, cardiac output. The combined gas law (a combination of Boyle's and Charles' law) describes the relationship of pressure, temperature, and volume in a gas. This law can be used to correct measured gas volumes to standard temperature and pressure (STP). This means that the measured volume of gas is standardised to 273 K and 760 mmHg (101.32 kPa).

As well as correcting for variations in atmospheric pressure, it is also necessary to correct for water vapor pressure. Dalton's law tells us that the pressure of a gas mixture is equal to the partial pressures of all of the components of the mixture. Water vapor is present in the atmosphere and in exhaled gas and its partial pressure contributes to the total atmospheric pressure. Water vapor exerts a constant pressure at a given temperature, regardless of the atmospheric pressure. Water vapor pressure is 47 mmHg (6.26 kPa) at normal body temperature and 17.5 mmHg at (2.33 kPa) 20 °C. Before correcting for STP it is necessary to subtract the water vapor pressure from the total atmospheric pressure to obtain the dry gas pressure at the ambient temperature. This is known as 'standard temperature and pressure, dry' (STPD), which is used for the correction.

Example : Atmospheric pressure = 762 mmHg (6.26 kPa)
Ambient temperature = 23 °C
Water vapor pressure at 23 °C = 21 mmHg (2.79 kPa)
Dry gas pressure = 762 − 21 mmHg = 741 mmHg (98.79 kPa)
STPD correction factor for 741 mmHg and 23 °C = 0.8991{from standard tables}
Volume of expired gas = 9.68 L · min^{-1}
STPD corrected volume = 9.68 × 0.8991 = 8.7 L · min^{-1}

By standardizing to STPD, we have removed the effect of water vapor pressure, ambient pressure, and temperature on the volume measurement and, hence, potential sources of day to day variation in measurement of the cardiac output.

Arteriovenous Oxygen Difference

Although many current generation analyzers can calculate oxygen content (ctO_2), earlier models did not. It may be necessary to manually calculate oxygen content from the hemoglobin level (Hb) and oxygen saturation of a sample. Hemoglobin is able to carry 1.36 mL of oxygen per gram of hemoglobin. Therefore, by multiplying the hemoglobin level by 1.36 it is possible to calculate the oxygen carrying capacity of the individual. Simply stated, this is the maximum amount of oxygen that can be carried by 100 mL of the individual's blood and is dependent on the hemoglobin level. Some textbooks have quoted the constant as 1.34 and others add a value of 0.03 to account for oxygen dissolved in plasma, but 1.36 is the generally accepted constant for calculation of oxygen carrying capacity.

$$\text{Oxygen carrying capacity} = \text{Hgb} \times 1.36\,\text{mL} \cdot \text{dL}^{-1}$$

If the total amount of oxygen that the blood is capable of carrying and the saturation of the sample is known, it is possible to calculate the oxygen content of that sample.

$$ctO_2 = \text{oxygen carrying capacity} \times \%\,\text{saturation}$$

The arteriovenous oxygen difference is the difference in oxygen content between arterial and venous blood.

Example : Hb = 14.5 g · dL^{-1}
Oxygen carrying capacity = 14.5 × 1.36 = 19.72 mL · dL^{-1}
Arterial saturation = 98.9%
Arterial oxygen content = 19.72 × 98.9% = 19.5 mL · dL^{-1}
Venous saturation = 66.9%
Venous oxygen content = 19.72 × 66.9% = 13.2 mL · dL^{-1}

Therefore,

$$AV_{diff} = 19.5 - 13.2 = 6.3\,\text{mL} \cdot \text{dL}^{-1} = 63\,\text{mL} \cdot \text{L}^{-1}$$

In this example, each liter of blood leaving the lungs delivers 63 mL of oxygen to the tissues.

Figure 1 shows a complete example of CO measurement using the Fick technique.

ASSUMPTIONS WHEN USING THE FICK TECHNIQUE FOR CARDIAC OUTPUT

Absence of Intracardiac Shunt

The method of calculating cardiac output described above uses the amount of oxygen absorbed by the blood as it travels through the lungs. We then assume that the amount of blood pumped by the right ventricle through the lungs is equal to the amount pumped by the left ventricle through the systemic vessels since the cardiovascular system is a closed system. This assumption does not always hold true and it is sometimes necessary to alter the calculation.

The term shunt describes the condition where a communication exists between the left- and right-sided chambers of the heart. If this condition results in shunting of blood between the venous and arterial circulation, the assumption becomes invalid because some blood is being recirculated through part of the circuit and the two ventricles are pumping unequal volumes. If an intracardiac shunt is known or suspected, it is necessary to collect blood samples at different points than the standard arterial and pulmonary artery sites. Calculation of cardiac output and

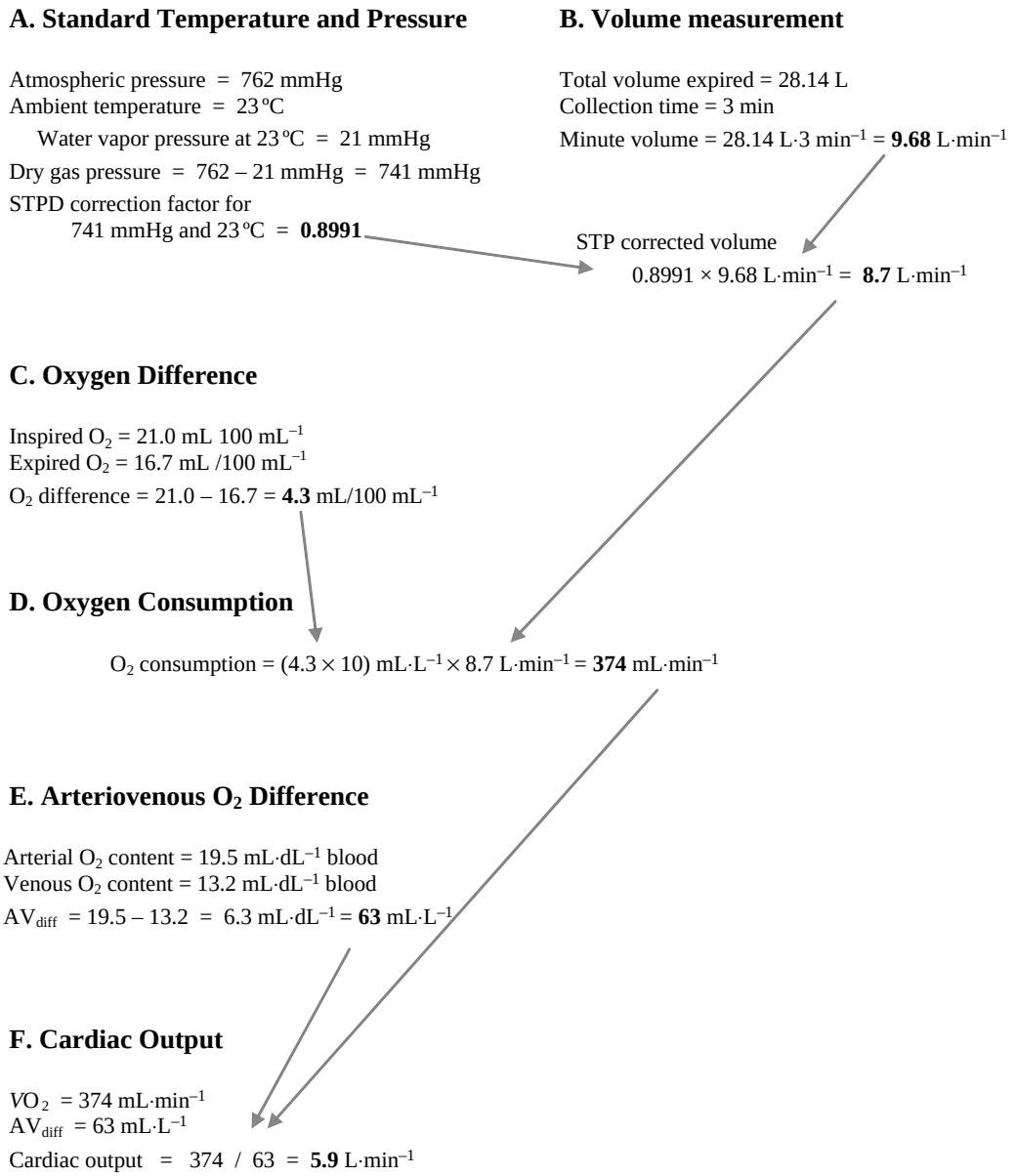


Figure 1. Example of cardiac output measurement using the Fick technique. The arrows indicate how the various parameters discussed in the text are interrelated in the various calculations.

shunt ratios in the presence of an intracardiac shunt is discussed later in this article.

Collection of True Arterial Sample

In a normal heart, it is not easy to gain access to the pulmonary veins to collect an arterial sample as the blood leaves the lungs. As a result, left ventricular or aortic blood is used to measure arterial oxygen content. This method ignores the small amount of venous admixture from bronchial and thebesian venous drainage into the left atrium.

Direct Measurement of Oxygen Consumption

Owing to the time- and labor-intensive methods of measurement of oxygen consumption ($\dot{V}O_2$), there is often a temptation to use an estimate of oxygen consumption rather than direct measurement. Standard formulas and nomograms are used to estimate $\dot{V}O_2$ from height, weight, age, and sex. The body surface area (BSA) is calculated from height and weight and expressed in units of square meters (m^2).

$$\text{BSA} = 0.007184 \times \text{weight}^{0.425} \times \text{height}^{0.725}$$

Age, sex, and basal metabolic rate are used to determine heat production from standard nomograms. Finally, heat production and BSA are used to estimate the oxygen consumption.

$$\dot{V}O_2 = [\text{BSA} \times \text{Heat Production}]/291.72$$

This method estimates the basal oxygen consumption at rest. It does not make allowances for any pathological

conditions, including those being investigated, that may affect the resting oxygen consumption. Studies comparing measured and estimated VO_2 have shown that estimating VO_2 from the various available formulas can lead to large and unpredictable errors in both VO_2 and cardiac output values (2,3). The practice of estimating VO_2 is strongly discouraged.

DETECTION AND ASSESSMENT OF INTRACARDIAC SHUNTS

Earlier in this article it was seen how the oxygen content of blood entering and leaving the lungs was used to calculate the cardiac output. It was assumed that blood flowing through the lungs is equal to blood flowing through the systemic circulation (since the cardiovascular system is a closed circuit). Several conditions may result in blood being recirculated between the left and right sides of the heart, leading to unequal flow in the pulmonary and systemic circulation. These conditions include atrial septal defects, patent foramen ovale, ventricular septal defects, and patent ductus arteriosus. Patent foramen ovale has been estimated to be present in 27.3% of the population (4), but the presence of a defect does not necessarily result in intracardiac shunting.

A communication between the left- and right-sided chambers of the heart can result in blood being shunted from right to left (venous blood being mixed into the arterial circulation), left to right (arterial blood being mixed into the pulmonary circulation), or as a bidirectional shunt (blood moves back and forth across the communication at different stages of the cardiac cycle). Although the method of calculating cardiac output remains essentially the same, the sites of blood collection are different in cases where intracardiac shunting exists. The following passages describe the methods for calculating systemic and pulmonary flow in the presence of different intracardiac shunts.

Left-to-Right Shunt

In a left-to-right shunt, arterial blood is pushed across the defect into the pulmonary circulation. This will artificially elevate the oxygen saturation and oxygen content in the pulmonary artery. To avoid error in the calculation of systemic cardiac output in the presence of a left-to-right shunt, it is necessary to collect the venous blood sample in the chamber immediately proximal to the shunt. In the case of atrial defects, blood is collected from both the inferior and superior vena cavae. Oxygen content (ctO_2) from these sites is used in the calculation of mixed venous oxygen content (MVO_2). The individual values are weighted and averaged according to the relatively higher flow from the superior vena cava and the absence of coronary sinus blood in the measurements. The generally accepted formula used for estimation of mixed venous oxygen content is

$$MVO_2 = [3 \times ctO_{2(SVC)} + ctO_{2(IVC)}]/4$$

MVO_2 becomes the venous component of the arteriovenous difference calculation and cardiac output (systemic flow, Q_s) is calculated as described earlier.

It is also possible to calculate pulmonary flow (Q_p) by using the pulmonary artery oxygen content as the venous component (blood entering the lungs) and left ventricular or aortic oxygen content as the arterial component (blood leaving the lungs). The pulmonary flow is equal to the systemic flow returning to the heart plus the volume being recirculated from the left-sided chambers via the shunt. This is reflected in the CO calculation. Because of the recirculated arterial blood, the venous oxygen content in the pulmonary artery will be elevated, leading to a decrease in the arteriovenous difference and, hence, a higher pulmonary flow.

Right-to-Left Shunt

The opposite occurs in a right-to-left shunt. Venous blood is mixed into the arterial circulation leading to a decrease in systemic arterial oxygen saturation. The calculation of systemic flow (Q_s) uses the arterial and venous (pulmonary artery) samples as usual. The calculation of pulmonary flow (Q_p) requires a sample to be taken after the blood leaves the lungs, but proximal to the shunt. In a right-to-left shunt it is necessary to sample blood from the pulmonary veins. In practical terms, this requires the catheter to pass from the right atrium across the defect to the left atrium and then into a pulmonary vein. Pulmonary vein oxygen content becomes the arterial component of the arteriovenous difference and pulmonary flow is calculated as usual.

In the presence of a right-to-left shunt, systemic flow is equal to the pulmonary flow leaving the lungs plus the amount that passes across the shunt directly from the right-sided chambers. Sampling in the left ventricle or aorta distal to the shunt will therefore give a lower arterial oxygen content than would be measured in the pulmonary veins, leading to an decrease in the arteriovenous difference and, hence, a higher systemic flow compared to the pulmonary flow.

A right-to-left shunt should be suspected in any patient who has an arterial oxygen saturation less than 95%. Investigation of these patients should include assessment for the presence of a bidirectional shunt.

Bidirectional Shunt

The presence of a bidirectional shunt complicates the calculation of pulmonary and systemic flow. Neither of the methods described above is suitable since both assume shunting in only one direction. The systemic blood flow (SBF) is calculated using oxygen contents sampled at the sites where blood enters and leaves the systemic circulation (arterial and mixed venous sites, respectively). Pulmonary blood flow (PBF) is calculated using sampling sites where blood enters and leaves the lungs (pulmonary artery and pulmonary vein, respectively). Finally, effective blood flow (EBF) is calculated using samples taken where the blood enters the heart and leaves the lungs (mixed venous and pulmonary vein oxygen contents). The mixed venous oxygen values should be the same as the pulmonary artery sample if no shunts are present. Similarly, the pulmonary vein should be the same as the left ventricular or aortic

VO_2 = 201 mL·min^{-1}
Hb = 13.9 g·dL^{-1}
Oxygen carrying capacity = 13.9 × 1.36 = 18.9 mL·dL^{-1}

Site	Saturation %	ctO_2 mL·dL^{-1}
Arterial		
Pulmonary vein (PV)	94.1	17.8
Radial artery (RArt)	83.0	15.7
Venous		
Superior vena cava (SVC)	58.6	11.1
Inferior vena cava (IVC)	61.0	11.6
Right atrium (RA)	68.8	13.0
Pulmonary artery (PA)	63.4	12.0
Mixed venous (MV) = (3 × SVC + IVC)/4	59.2	11.2

Pressure measurements
Mean Pulmonary Artery pressure = 43 mmHg (5.73 kPa)
Mean Left Atrial pressure = 4 mmHg (0.53 kPa)

$$\text{PBF} = \frac{VO_2}{ctO_{2\,(PV)} - ctO_{2\,(PA)}} = \frac{201}{17.8 - 12.0} = 3.5\ \text{L·min}^{-1}$$

$$\text{SBF} = \frac{VO_2}{ctO_{2\,(R\,Art)} - ctO_{2\,(MV)}} = \frac{201}{15.7 - 11.2} = 4.5\ \text{L·min}^{-1}$$

$$\text{EBF} = \frac{VO_2}{ctO_{2\,(PV)} - ctO_{2\,(MV)}} = \frac{201}{17.8 - 11.2} = 3.0\ \text{L·min}^{-1}$$

$$\text{Left-to-right shunt} = \text{PBF} - \text{EBF} = 3.5 - 3.0 = 0.5\ \text{L·min}^{-1}$$

$$\text{Right-to-left shunt} = \text{SBF} - \text{EBF} = 4.5 - 3.0 = 1.5\ \text{L·min}^{-1}$$

$$\text{PVR} = \frac{80 \times (PA_m - LA_m)}{Q_p} = \frac{80 \times (43 - 4)}{3.5} = 891\ \text{dyn·s·cm}^{-5}$$

Figure 2. Bidirectional shunt calculation in a 55 year old woman with Eisenmenger's syndrome secondary to an atrial septal defect (ASD). Note the predominantly right-to-left shunt due to increased pulmonary vascular resistance. These results show little deterioration compared to measurements taken six months earlier [$Q_p = 3.7\,\text{L}\cdot\text{min}^{-1}$, mean PA pressure = 39 mmHg (5.19 kPa) and PVR = 800 dyn · s · cm^{-5}].

sample in the absence of any shunts. Therefore, by sampling at these sites, the pulmonary (and hence, systemic) flow that would normally occur if no shunts were present is being calculated

$$\text{SBF} = VO_2/[ctO_{2(Art)} - ctO_{2(MV)}]$$
$$\text{PBF} = VO_2/[ctO_{2(P\,vein)} - ctO_{2(P\,art)}]$$
$$\text{EBF} = VO_2/[ctO_{2(P\,vein)} - ctO_{2(MV)}]$$

Since the systemic flow (SBF), pulmonary flow (PBF), and the flow that would occur in the absence of any shunts (EBF) is known, the size of the shunts can also be calculated

$$\text{Right to left} = \text{SBF} - \text{EBF}$$
$$\text{Left to right} = \text{PBF} - \text{EBF}$$

Figure 2 shows calculations for a bidirectional shunt based on the principles discussed above. Note the changes in oxygen saturation and content as blood passes the atrial defect. In the left heart, oxygen saturation decreases as blood passes from pulmonary veins to the left ventricle due to mixing of deoxygenated blood being shunted across the defect. Conversely, oxygen saturation in the right heart increases as blood passes from the vena cavae (mixed venous) to pulmonary artery due to oxygenated blood being shunted across the defect from the left atrium.

Pulmonary–Systemic Flow Ratio

An alternative to calculating flow across a defect is to calculate the pulmonary/systemic flow ratio (P/S ratio). This value is a ratio of the pulmonary flow relative to

the systemic flow. Calculation of the P/S ratio does not involve calculation of actual flows, so it is not necessary to collect expired gases to calculate the VO_2. The P/S ratio is calculated using only the oxygen content (or saturation) from arterial, pulmonary artery, mixed venous, and pulmonary vein samples.

The AV_{diff} for the pulmonary component is calculated by subtracting mixed venous oxygen content from systemic arterial oxygen content. The systemic AV_{diff} is calculated by subtracting the pulmonary artery oxygen content from the pulmonary vein oxygen content. If a pulmonary vein sample is not possible, use an estimate of 98% unless the arterial saturation is higher. Using arterial oxygen content to estimate the pulmonary vein content will assume that there is no right-to-left shunt. The P/S ratio (or P:S ratio) is the calculated by dividing the pulmonary component by the systemic component.

The P/S ratio is the proportion of flow through the pulmonary circulation relative to the systemic circulation. Therefore, a value > 1.0 indicates left-to-right shunting. An arbitrary value of between 1.5 and 2.0 is often used to determine the need for definitive treatment to correct the defect, in order to avoid late sequelas from prolonged pulmonary vascular overload. A P/S flow ratio < 1 indicates right-to-left shunting and may be a sign of irreversible pulmonary vascular disease.

$$\textit{Example}: \mathrm{Hb} = 14.5\,\mathrm{g}\cdot\mathrm{dL}^{-1}$$

$$\text{Oxygen carrying capacity} = 19.72\,\mathrm{mL}\cdot\mathrm{dL}^{-1}$$

$$\text{Arterial oxygen content} = 98.9\%$$
$$\mathrm{ctO_{2(Art)}} = 19.5\,\mathrm{mL}\cdot\mathrm{dL}^{-1}$$
$$\text{Pulmonary artery oxygen content} = 66.9\%$$
$$\mathrm{ctO_{2(PA)}} = 13.2\,\mathrm{mL}\cdot\mathrm{dL}^{-1}$$
$$\text{Mixed venous oxygen content} = 63.1\%$$
$$\mathrm{ctO_{2(MV)}} = 12.4\,\mathrm{mL}\cdot\mathrm{dL}^{-1}$$

$$\begin{aligned}\text{Pulmonary}: \quad \mathrm{ctO_{2(Art)}} - \mathrm{ctO_{2(MV)}} &= 19.5 - 12.4\\ &= 7.1\,\mathrm{mL}\cdot\mathrm{dL}^{-1}\\ \text{Systemic}: \quad \mathrm{ctO_{2(PV)}} - \mathrm{ctO_{2(PA)}} &= 19.5 - 13.2\\ &= 6.3\,\mathrm{mL}\cdot\mathrm{dL}^{-1}\end{aligned}$$

$$\mathrm{P/S\,ratio}: \ 7.1/6.3 = 1.13$$

FLOW-DEPENDENT PARAMETERS

A number of frequently used parameters in cardiovascular medicine are dependent on knowing systemic or pulmonary flow. Calculation of these parameters, and the effect that the cardiac output has on each, is discussed. Table 1 lists expected normal ranges for a number of common flow-dependent parameters.

Cardiac Index

Cardiac output is often corrected for patient's size, based on body surface area (BSA). Cardiac index (CI) is calculated by dividing the cardiac output by the body surface area.

$$\mathrm{CI} = \mathrm{CO/BSA}\ \mathrm{L}\cdot\mathrm{min}^{-1}\cdot\mathrm{m}^{-2}$$

Table 1. Expected Ranges for Common Flow-Dependent Parameters

$$\text{Cardiac Output} = \frac{\text{Oxygen consumption}}{\text{Arteriovenous oxygen difference}}$$

$$VO_2 = \frac{\mathrm{BSA}\times\text{Heat Production}}{291.72}$$

$$\mathrm{MVO_2} = \frac{3\times\mathrm{ctO_{2(SVC)}} + \mathrm{ctO_{2(IVC)}}}{4}$$

$$\mathrm{SBF} = \frac{\mathrm{VO_2}}{\mathrm{ctO_{2(Art)}} - \mathrm{ctO_{2(MV)}}}$$

$$\mathrm{PBF} = \frac{\mathrm{VO_2}}{\mathrm{ctO_{2}}(P\,\mathrm{vein}) - \mathrm{ctO_{2}}(P\,\mathrm{art})}$$

$$\mathrm{EBF} = \frac{\mathrm{VO_2}}{\mathrm{ctO_{2}}(P\,\mathrm{vein}) - \mathrm{ctO_{2(MV)}}}$$

$$\mathrm{CI} = \frac{\mathrm{CO}}{\mathrm{BSA}}\mathrm{L}\cdot\mathrm{min}^{-1}\cdot\mathrm{m}^{-2}$$

$$\text{Area} = \frac{\left\{\frac{\mathrm{CO}}{(\mathrm{SEP}\times\mathrm{HR})}\right\}}{(44.3\times\sqrt{\text{gradient}})}$$

$$\text{Area} = \frac{\text{Cardiac output}}{\sqrt{\text{Gradient}}}$$

$$\text{Area} = \frac{(\mathrm{CO}\times\mathrm{DFP}\times\mathrm{HR})}{37.7\times\sqrt{\text{gradient}}}$$

$$\mathrm{SVR} = \frac{80\times(A\mathrm{o_m} - \mathrm{RA_m})}{Q_\mathrm{s}}$$

$$\mathrm{PVR} = \frac{80\times(\mathrm{PA_m} - \mathrm{LA_m})}{Q_\mathrm{p}}$$

$$\mathrm{CO} = \frac{V\mathrm{CO_2}}{\mathrm{ctCO_{2(Ven)}} - \mathrm{ctCO_{2(Art)}}}$$

$$\mathrm{CO} = \frac{\Delta V\mathrm{CO_2}}{\Delta\mathrm{ctCO_{2(Art)}}}$$

$$\mathrm{CO} = \frac{\Delta V\mathrm{CO_2}}{S\times\Delta\mathrm{ETCO_2}}$$

Some believe cardiac index is a more useful parameter than cardiac index because it accounts for the patient's size. A large person (as approximated by BSA) would be expected to have a higher cardiac output while a low cardiac output in a smaller person may not necessarily be indicative of a poorly functioning ventricle. Many authors only express cardiac output in terms of cardiac index for this reason.

Valve Areas

Basic fluid dynamic principles state that a fluid exerts pressure equally in all directions. Therefore, when the valves of the heart are open they should allow equalisation of pressure in the two chambers that they separate. Sometimes the valves of the heart become stiff, thickened or do not open properly, inhibiting flow through the valve. This is referred to as valve stenosis. A result of this process is a pressure gradient, a difference in pressure on either side of the valve. Take an example of aortic valve stenosis. When

the left ventricle contracts, it is pushing against an obstruction. The systolic pressure will be higher in the ventricle that in the aorta. The difference in pressure is referred to as a pressure gradient (expressed in mmHg) and can be measured during cardiac catheterisation. The pressure gradient across a valve is often used to determine the severity of a valve stenosis. However, the main parameter that should be considered is the cross-sectional area of the valve. A pressure gradient of 20 mmHg (2.66 kPa) is often considered an indication of mild aortic stenosis. However, in the presence of low cardiac output, it is necessary to have quite a narrow valve orifice to achieve this gradient. Conversely, a less severe stenosis could achieve a gradient of 50 mmHg (6.66 kPa) in a patient with a high cardiac output.

Both the cardiac output and mean pressure gradient are used in the calculation of valve area. The Gorlin formula is used for calculating valve area of the aortic or pulmonary valve:

$$\text{Area} = [\{\text{CO}/(\text{SEP} \times \text{HR})\}/(44.3 \times \sqrt{\text{gradient}})]$$

where HR is the heart rate and SEP is the systolic ejection period (since gradients in these valves are measured during systole). However, a shorter formula is often used as an approximation:

$$\text{Area} = \text{Cardiac output}/\sqrt{\text{Gradient}}$$

Taking the example of aortic stenosis, a mean gradient of 20 mmHg (2.66 kPa) in a patient with a normal cardiac output of $4.5\,\text{L}\cdot\text{min}^{-1}$ would give a valve area of $4.5/\sqrt{20} = 1.00\,\text{cm}^2$. In a patient with a low cardiac output of $3.1\,\text{L}\cdot\text{min}^{-1}$, the valve area ($3.1/\sqrt{20} = 0.69\,\text{cm}^2$) would be much smaller to achieve this gradient. Similarly, our patient with a mean gradient of 50 mmHg (6.66 kPa) would not have such a severe narrowing if a high cardiac output (e.g., $7.1\,\text{L}\cdot\text{min}^{-1}$) is present ($7.1/\sqrt{50} = 1.00\,\text{cm}^2$).

The Gorlin formula for calculating mitral or tricuspid valve area is slightly different:

$$\text{Area} = (\text{CO} \times \text{DFP} \times \text{HR})/(37.7 \times \sqrt{\text{gradient}})$$

where DFP is the diastolic filling period (since gradients in these valves are measured during diastole).

Vascular Resistance

Measurements of vascular resistance are based on principles of fluid dynamics where resistance is defined as the decrease in pressure between two points in a vascular segment divided by the flow through that segment. While this simplification does not account for pulsatile flow, calculation of vascular resistance in this way is useful in a number of clinical settings.

In the past, Wood units ($\text{mmHg}\cdot\text{L}\cdot\text{min}^{-1}$) were used to express vascular resistance. Today, vascular resistance is more commonly expressed in absolute resistance units (dyn.s.cm^{-5}), which are derived from the mean pressure gradient ($\text{dyn}\cdot\text{cm}^{-2}$) divided by the mean flow ($\text{cm}^3\cdot\text{s}^{-1}$). A constant of 80 is used to convert values from traditional units (mmHg and $\text{L}\cdot\text{min}^{-1}$) to absolute resistance units.

Systemic vascular resistance (SVR) is therefore defined as the difference in pressure between blood entering the systemic circulation (mean aortic pressure) and blood leaving the systemic circulation (mean right atrial pressure) divided by the systemic blood flow:

$$\text{SVR} = [80 \times (\text{AO}_\text{m} - \text{RA}_\text{m})/Q_\text{s}]$$

Similarly, pulmonary vascular resistance (PVR) is defined as the difference between mean pulmonary artery pressure and mean left atrial pressure divided by the pulmonary flow:

$$\text{PVR} = [80 \times (\text{PA}_\text{m} - \text{LA}_\text{m})/Q_\text{p}]$$

In the absence of intracardiac shunting both SVR and PVR can be calculated using the standard cardiac output measurement. If intrapulmonary shunting is present, systemic and pulmonary flow must be individually calculated for use in the SVR and PVR calculations, respectively.

There are a number of causes of increased systemic or pulmonary vascular resistance, some reversible and some permanent. The use of serial cardiac output and pressure measurements during drug challenges can assist in identifying management strategies that may be helpful in reducing vascular resistance.

VARIATIONS OF THE FICK METHOD

The Fick principle can be applied to any gas involved in diffusion, including carbon dioxide. Such variations to the classic Fick formula are often referred to as the indirect Fick principle.

By measuring the difference between inspired and expired CO_2 and the minute ventilation volume we can calculate CO_2 production ($V\text{CO}_2$). Arteriovenous CO_2 difference is calculated from the measured values of arterial and venous carbon dioxide content (ctCO_2). The ratio of $V\text{CO}_2$ and the arteriovenous CO_2 difference gives the cardiac output.

Earlier in this article it was seen how the oxygen content (ctO_2) of blood is calculated from the amount of hemoglobin and oxygen saturation of the sample, since nearly all of the oxygen is bound to hemoglobin. A relatively smaller proportion of CO_2 is bound to hemoglobin. About 70% is transported in the blood as bicarbonate. Only 23% is bound to hemoglobin and 7% is transported as dissolved CO_2. Therefore, the calculation of CO_2 content is not dependent on hemoglobin level.

Carbon dioxide content (ctCO_2) is calculated from the formula:

$$\text{ctCO}_2 = 11.02 \times P\text{CO}_2^{0.396}$$

Thus, if we have partial CO_2 pressure of arterial ($P\text{aCO}_2$) and venous ($P\text{vCO}_2$) samples, we can calculate the arteriovenous carbon dioxide difference as

$$\text{ctCO}_{2(\text{Ven})} - \text{ctCO}_{2(\text{Art})} = 11.02(P\text{vCO}_2^{0.396} - P\text{aCO}_2^{0.396})$$

Then cardiac output is calculated with the formula:

$$\text{CO} = V\text{CO}_2/(\text{ctCO}_{2(\text{Ven})} - \text{ctCO}_{2(\text{Art})})$$

There are some advantages to using the carbon dioxide method. When a patient is receiving high concentrations of supplemental oxygen, analysis of inspired and expired

oxygen will give a small difference between two relatively large values. Even a small error in the estimation of either value will yield an inaccurate VO_2. Additionally, some oxygen analysers (e.g., paramagnetic analyzers) have poor accuracy at high oxygen concentrations. Measurement of cardiac output in patients receiving high concentrations of supplemental oxygen may be erroneous and the Fick principle using carbon dioxide may prove more accurate.

Applying the Fick principle to carbon dioxide involves the same steps as using oxygen for the calculation. There is still the requirement for analysis of expired gases, as well as collection and analysis of arterial and mixed venous blood samples. However, there are a number of ways of estimating, rather than directly measuring, the various parameters necessary to calculate cardiac output using the Fick CO_2 technique.

Infrared (IR) light absorption sensors in the breathing circuit can measure inspired and expired CO_2 content. Alternatively, an assumption can be made about the content of CO_2 in the inhaled gas (especially if the patient is being ventilated with 100% oxygen) and only expired CO_2 needs to be measured. Along with an airflow sensor (e.g., as a differential pressure pneumotachometer), these measurements can provide real-time estimation of VCO_2.

There is a logarithmic relationship between cardiac output and end-tidal CO_2 ($ETCO_2$). At normal or high cardiac output the respiratory rate determines the amount of CO_2 that is eliminated by the lungs with each breath. If it is assumed that CO_2 exchange at the alveolar–arterial membrane reaches equilibrium, then $ETCO_2$ can be used to estimate $PaCO_2$. In this way, it is possible to estimate cardiac output without subjecting the patient to unnecessarily invasive procedures.

The critically ill patient presents a number of challenges. These patients are usually intubated and manually ventilated with high concentrations of inspired oxygen. While many will have arterial lines for blood pressure monitoring, those that do not are exposed to added risk of morbidity if arterial access is necessary to determine cardiac output. In addition, placement of a pulmonary artery catheter for the measurement of mixed venous gas tension exposes the patient to significant risk of sepsis, pneumothorax, thrombosis, or pulmonary artery rupture. However, cardiac output is often vital in determining end-organ perfusion.

Recently, a system for noninvasive measurement of cardiac output using the Fick principle and carbon dioxide was developed for use with ventilated patients in the intensive care unit, based on the estimations described above. A number of assumptions allow this system to be used without the need for arterial or mixed venous blood samples. The technique involves measuring changes in carbon dioxide production and arterial CO_2 content between normal breathing conditions and under rebreathing conditions with 10–15% CO_2 and a reservoir with a volume 1.5 times the tidal volume. Carbon dioxide production (VCO_2) is calculated from the minute ventilation and expired CO_2 content under normal breathing conditions. Arterial CO_2 content [$ctCO_{2(Art)}$] is estimated from the end-tidal CO_2 ($ETCO_2$) with adjustments for the slope of the CO_2 dissociation curve and degree of dead space ventilation.

During partial rebreathing, carbon dioxide elimination from the blood is reduced, but $ETCO_2$ increases and reaches a plateau within a few breaths. Studies conducted in anaesthetised dogs have showed that during a brief period of CO_2 rebreathing there is a change in $PaCO_2$ and in calculated VCO_2, but little or no change in venous carbon dioxide content ($ctCO_{2(Ven)}$) (5). It is believed that this finding is due to the quantity of CO_2 stores in the body being large and new equilibrium levels not being attained for 20–30 min. This finding becomes highly important in the noninvasive estimation of cardiac output. Any change in the arteriovenous CO_2 difference during the brief rebreathing period can be attributed to changes in the arterial CO_2 component alone.

If it is assumed that the cardiac output and $ctCO_{2(Ven)}$ remain constant during normal breathing (N) rebreathing (R):

$$\begin{aligned} CO &= VCO_{2(N)}/(ctCO_{2(Ven)(N)} - ctCO_{2(Art)(N)}) \\ &= VCO_{2(R)}/(ctCO_{2(Ven)(R)} - ctCO_{2(Art)(R)}) \end{aligned}$$

From basic algebra it is known that

$$X = A/B = C/D = (A - C)/(B - D)$$

Then,

$$\begin{aligned} CO = (VCO_{2(N)} - VCO_{2(R)})/[&(ctCO_{2(Ven)(N)} - ctCO_{2(Art)(N)}) \\ &- (ctCO_{2(Ven)(R)} - ctCO_{2(Art)(R)})] \end{aligned}$$

Rearranging this equation:

$$\begin{aligned} CO = (VCO_{2(N)} - VCO_{2(R)})/[&(ctCO_{2(Ven)(N)} - ctCO_{2(Ven)(R)}) \\ &- (ctCO_{2(Art)(N)} - ctCO_{2(Art)(R)})] \end{aligned}$$

Since it has been assumed that $ctCO_{2(Ven)}$ does not change during rebreathing ($ctCO_{2(Ven)(N)}$ is equal to $ctCO_{2(Ven)(R)}$), these values cancel each other out and the equation becomes:

$$CO = (VCO_{2(N)} - VCO_{2(R)})/(ctCO_{2(Art)(R)} - ctCO_{2(Art)(N)})$$

In other words, cardiac output is equal to the change in VO_2 divided by the change in arterial CO_2 content between the normal and rebreathing states:

$$CO = \Delta VCO_2/\Delta ctCO_{2(Art)}$$

As $ctCO_{2(Art)}$ is estimated from $ETCO_2$ and the slope (S) of the CO_2 dissociation curve:

$$CO = \Delta VCO_2/S \times \Delta ETCO_2$$

This method of cardiac output estimation gives a measure of the pulmonary capillary blood flow (Q_{PCBF}). Changes in VCO_2 and $ETCO_2$ only reflect the blood flow that participates in gas exchange. An intrapulmonary shunt occurs when venous blood passes through unventilated areas of the lungs and moves into the arterial circulation without taking up oxygen or releasing carbon dioxide. A large intrapulmonary shunt will not be reflected in the changes seen in VCO_2 and $ETCO_2$. Therefore, it is necessary to estimate the degree of shunting and correct the cardiac output estimation accordingly. For example, if only 80% of pulmonary blood flow is participating in gas exchange, the Q_{PCBF} estimated

by this method will be 80% of the total cardiac output. The shunting fraction is calculated using the arterial oxygen saturation (SaO_2, from a pulse oximeter), the fraction of inspired oxygen (FiO_2), arterial oxygen tension (PaO_2) and standard isoshunt tables. The requirement of an arterial blood gas sample for PaO_2 means that this method is not truly noninvasive.

As mentioned previously, this noninvasive method involves a number of assumptions. In summary, these assumptions are the CO_2 exchange at the alveolar–arterial membrane reaches equilibrium. Therefore, $ETCO_2$ is equal to $PaCO_2$; cardiac output remains constant during rebreathing; venous CO_2 content does not change during a brief period of rebreathing; and there is little or no intrapulmonary shunting.

SUMMARY

The Fick method remains the gold standard of cardiac output measurement. While technically challenging, it relies on direct measurement of oxygen consumption and uptake to determine the rate of blood flow through the lungs and around the body. Accurate measurement of cardiac output is necessary for the estimation of several important parameters and assessment of complex congenital cardiac conditions. Variations of the classical Fick principle allow estimation of cardiac output in patients who might otherwise be unsuitable.

BIBLIOGRAPHY

1. Vandam LD, Fox JA, Fick A. (1829–1901), Physiologist: A heritage for anaesthesiology and critical care medicine. Anaesthesiology 1998;88(2):514–518.
2. Kendrick AH, West J, Papouchado M, Rozkovec A. Direct Fick cardiac output: Are assumed values for oxygen consumption acceptable? Eur Heart J 1988;9(3):337–342.
3. Wolf A, et al. Use of assumed versus measured oxygen consumption for the determination of cardiac output using the Fick principle. Catheterization Cardiovascular Diagnosis 1998;43(4): 372–380.
4. Hagen PT, Scholz DG, Edwards WD. Incidence and size of patent foramen ovale during the first 10 decades of life: An autopsy study of 965 normal hearts. Mayo Clinic Proc 1984; 59(1):17–20.
5. Tachibana K, et al. Effect of ventilatory settings on accuracy of cardiac output measurement using partial CO(2) rebreathing. Anesthesiology 2002;96(1):96–102.

Reading List

Grossman W. Blood flow measurement: The cardiac output. In: Baim DS, Grossman W. Cardiac Catheterization, Angiography and Intervention. 5th ed. Baltimore: Williams and Wilkins; 1996: pp 109–120.

Davidson CJ, Fishman RF, Bonow RO. Cardiac Catheterization (Ch 6). In: Braunwald E, editor. Heart Disease: A textbook of cardiovascular medicine. 5th ed. Philadelphia: WB Saunders; 1997: pp 177–203.

Grossman W. Shunt detection and measurement. In: Baim DS, Grossman W. Cardiac Catheterization, Angiography and Intervention. 5th ed. Baltimore: Williams and Wilkins; 1996: pp 167–180.

Feneley MP. Measurement of cardiac output and shunts. In: Boland J, Muller DWM, editors. Cardiology and cardiac catheterisation: The essential guide. Amsterdam: Harwood Academic Publishers; 2001: pp 197–205.

See also BLOOD GAS MEASUREMENTS; CARDIAC OUTPUT, INDICATOR DILUTION MEASUREMENT OF; CARDIAC OUTPUT, THERMODILUTION MEASUREMENT OF; PERIPHERAL VASCULAR NONINVASIVE MEASUREMENTS; RESPIRATORY MECHANICS AND GAS EXCHANGE.

CARDIAC OUTPUT, INDICATOR DILUTION MEASUREMENT OF

F. M. DONOVAN
University of South Alabama

B. C. TAYLOR
The University of Akron
Akron, Ohio

INTRODUCTION

Cardiac output is defined as the volume of blood pumped by the left or right ventricle per unit of time and is normally expressed in $L \cdot min^{-1}$ (1). For the average 70 kg adult male, the cardiac output is $\sim 5\ L \cdot min^{-1}$, however, exercise can cause this figure to increase as much as six times the resting value in well-trained athletes (2). Knowledge of cardiac function is an important tool for determining the hemodynamic status of an individual whether he/she is a trained athlete or a patient in a critical care setting. Accurate direct measurement of cardiac output is a rather difficult task since to obtain a direct measurement would require collecting and measuring all of the blood pumped from the heart into the aortic outflow tract. It is necessary, therefore, to develop indirect methods for the measurement of cardiac output that would provide equivalent accuracies. The Fick (2) and other indicator dilution methods (3) are two of the invasive procedures that provide good reasonable results. More recently, echo cardiography and other noninvasive techniques have been gaining in popularity, yet the Fick and Indicator Dilution methods remain the "Gold Standards" against which all other methods are compared because of their accuracy, safety, reproducibility, and relative simplicity (1).

The Fick principle (4) is based on the fact that the amount of an indicator taken up (or released) by an organ is the product of its blood flow and the difference in concentration of the substance between the organ's arterial and venous blood. Cardiac output can be determined by dividing the amount of oxygen consumed by the arterial-venous oxygen difference (AVO_2 difference). The theory behind this procedure is explained more fully below.

The indicator dilution method became widely accepted after Hamilton, in 1948, demonstrated that this technique agreed with the Fick method. In the indicator dilution method (5) an indicator(dye, thermal, saline) is injected into the venous blood and its concentration is measured continuously in the arterial blood as it passes through the circulatory system The cardiac output is determined by analyzing the resulting time-dependent concentration curve.

INDICATOR DILUTION METHOD FUNDAMENTAL EQUATIONS

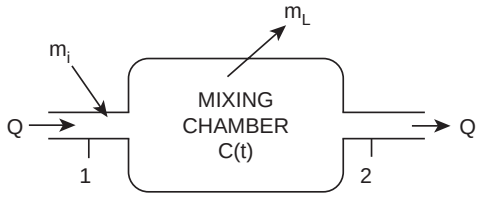


Figure 1. Simple mixing chamber.

Consider the simple mixing chamber shown in Fig. 1, where Q is the constant volumetric flow rate into and out of the chamber, m_i is the mass of indicator injected into the inflow stream, $C(t)$ is the concentration of indictor in the chamber at any instant, and m_L is the mass of indicator that leaves the chamber due to diffusion to the wall and/or metabolism. The differential mass of indicator that leaves the chamber at point 2 per differential time is given by

$$dm_2 = C(t)Qdt$$

where $C(t)$ at point 2 is the same as the concentration of indictor in the mixing chamber assuming complete mixing in the chamber. The total mass of indicator that leaves the chamber is determined by integrating the above equation with the result shown below.

$$m_2 = Q\int_0^\infty C(t)dt$$

The differential mass of indicator that leaves the chamber due to diffusion to the chamber wall and/or metabolism is proportional to the concentration of indicator, $C(t)$, and the surface area of the chamber, A.

$$dm_\mathrm{L} = C(t)ADdt$$

where D is the proportionality constant for diffusion and/or metabolism.

The total mass of indicator leaving the chamber due to diffusion is

$$m_\mathrm{L} = AD\int_0^\infty C(t)dt$$

The total mass of indicator leaving the chamber is equal to the mass of indicator entering the chamber minus the mass loss of indicator due to diffusion

$$m_2 = m_\mathrm{i} - m_\mathrm{L}$$

which leads to

$$Q\int_0^\infty C(t)dt = m_\mathrm{i} - AD\int_0^\infty C(t)dt$$

and subsequently to the equation for volumetric flow rate

$$Q = \frac{m_\mathrm{i}}{\int_0^\infty C(t)dt} - AD$$

The integral of $C(t)dt$ is determined from the area under the indicator dilution curve. Note that if the area of the chamber wall is large and/or the diffusion coefficient is large, then the flow rate will be overestimated unless the diffusion is taken into account. In practice, the effect of diffusion is taken into account by a multiplying calibration factor (K) as shown in equation 1.

$$Q = \frac{m_\mathrm{i}}{\int_0^\infty C(t)dt}K \tag{1}$$

This is the familiar Stewart–Hamilton equation for calculating cardiac output from the indicator dilution curve (6).

INDICTOR DILUTION CURVE FUNDAMENTAL EQUATIONS

The equations for the indicator dilution curve are determined by the indicator mass flow rate conservation, which states that the mass flow rate of indicator into the mixing chamber must equal the mass flow rate of indicator out of the mixing chamber plus the mass rate of removal by diffusion plus the rate of change of indictor stored in the chamber.

$$C_\mathrm{i}Q = C(t)Q + C(t)AD + V\frac{dC(t)}{dt}$$

The parameter V is the volume of the chamber and complete mixing is assumed so that the concentration of indicator leaving the container at point 2 is equal to the concentration of indicator in the chamber at any instant.

Rearranging this equation results in the first-order differential equation

$$\left(\frac{V}{Q+AD}\right)\frac{dC(t)}{dt} + C(t) = \frac{Q}{Q+AD}C_\mathrm{i}$$

which has a time constant of

$$\tau = \frac{V}{Q+AD}$$

After the injection of the indicator is complete, the concentration of indicator flowing into the chamber becomes zero resulting in the following equation during the washout phase.

$$\tau\frac{dC(t)}{dt} + C(t) = 0$$

The solution to this equation is

$$C(t) = C(t_1)e^{-[(t-t_1)/\tau]}$$

where $C(t_1)$ is the indicator concentration at time t_1 on the washout part of the indicator dilution curve.
Taking the natural log of this equation results in

$$\ln C(t) = \ln C(t_1) - [(t-t_1)/\tau]$$

This shows that if the indicator dilution curve is plotted as natural log of C versus t, then the curve will become a straight line during the washout phase and the slope of the curve is the negative reciprocal of the system time constant.

Rearranging this equation yields

$$\tau = \frac{t_1 - t_2}{\ln C(t_2) - \ln C(t_1)}$$

where $C(t_1)$ and $C(t_2)$ are indicator concentrations at time t_1 and t_2, respectively, all located on the washout part of the indicator dilution curve.

This equation can be rearranged to the following:

$$\tau = -C(t_1) \Big/ \left(\frac{dC}{dt}\right)_{t_1}$$

The total area under the indicator dilution curve from t_1 to infinity is given by

$$\int_{t_1}^{\infty} C(t)dt = \tau C(t_1) \int_{t_1}^{\infty} e^{-[(t-t_1)/\tau]} d\left(\frac{t-t_1}{\tau}\right)$$

which results in the equation for the remaining area under the curve.

$$\int_{t_1}^{\infty} C(t)dt = C(t_1)\tau = \frac{(t_1 - t_2)}{\ln C(t_2) - \ln C(t_1)} C(t_1) \qquad (2)$$

This equation can be rearranged to the following:

$$\int_{t_1}^{\infty} C(t)dt = \frac{-[C(t_1)]^2}{[C(t)/dt]_{t_1}} \qquad (3)$$

APPLICATION OF THE INDICATOR DILUTION EQUATIONS AND RECIRCULATION

The following results are from a computer simulation in which 6 mg of indicator are injected into the right atrium of an average male with indicator concentrations being read from the radial artery. The volumes used by the simulation for the chambers involved are shown in Figure 2.

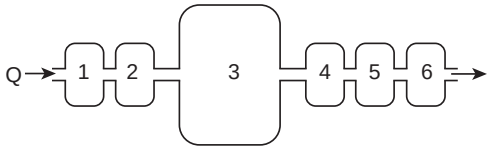


Figure 2. Schematic of simulation system.

Q is cardiac output (6 $L \cdot min^{-1}$)
1 is the right atrium ($V1 = 100$ mL)
2 is the right ventricle ($V2 = 100$ mL)
3 is the pulmonary circulatory system ($V3 = 600$ mL)
4 is the left atrium ($V4 = 100$ mL)
5 is the left ventricle ($V5 = 100$ mL)
6 is the systemic artery volume from the left ventricle to the radial artery ($V6 = 100$ mL)

The total circulatory system volume is taken to be 6 L.

The diffusion coefficient D is taken to be zero in the simulation. The resulting indicator dilution curve as measured in the radial artery is shown in Fig. 3.

The dashed line beginning at $\sim$ 22 s shows what the curve would do if there were no recirculation of the indicator through the circulatory system back to the right atrium. The solid line shows the actual curve with recirculation.

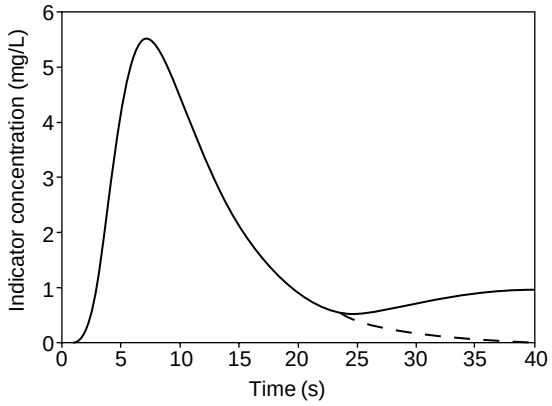


Figure 3. Indicator dilution curve.

The area under the indicator dilution curve can not be determined directly from the dilution curve. The dilution curve is plotted on a semilog graph in Fig. 4.

During the washout phase, the concentration curve approaches a straight line (dashed line) before the recirculation distorts the plot (solid line). This indicates that the system is behaving as a first-order decay, and we can use equation 2 to determine the area under the curve from any chosen time on the straight-line portion of the curve to infinity.

In this simulation, the area under the indicator dilution curve that would occur if there were no recirculation from 0 to 40 s is found to be 59.8 $mg \cdot s \cdot L^{-1}$, which results in a calculated cardiac output of 6.02 $L \cdot min^{-1}$.

With recirculation we determine the area under the curve from 0 to 15 s to be 47.387 $mg \cdot s \cdot L^{-1}$ and use equation 2 to calculate the area from 15 s to infinity.

For example, use the concentrations at 15 and 20 s that are in the straight-line portion of the semilog plot.

$$C_{t=15} = 2.096\,mg \cdot L^{-1} \quad C_{t=20} = 0.908\,mg \cdot L^{-1}$$

Equation 2 gives the area from 15 s to infinity as 12.528 $mg \cdot s \cdot L^{-1}$ so the total area from 0 to infinity is calculated to be 59.915 $mg \cdot s \cdot L^{-1}$. Now using equation 1, the cardiac output is found to be 6.01 $L \cdot min^{-1}$.

Using equation 3 at 15 s yields the same result.

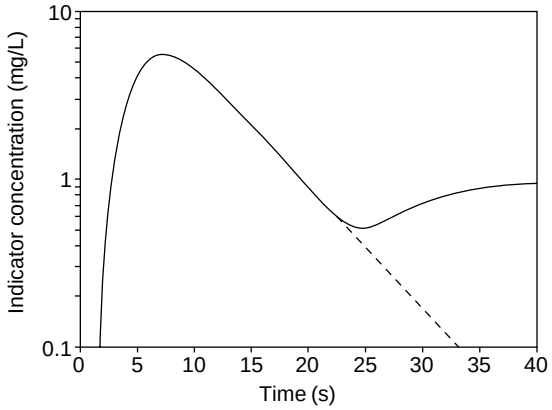


Figure 4. Semilog plot of indicator dilution curve.

FICK PRINCIPLE

If the indicator is supplied to the mixing chamber shown in Fig. 1 at a steady rate until steady state is reached, then the concentration of indicator in the stream leaving the chamber will be constant and is given by the equation

$$m_{f_2} = C_2 Q$$

where m_{f_2} is the mass flow rate of indicator flowing out of the chamber.

Under steady flow conditions the mass flow rate of indicator into the chamber in the inflow stream plus the mass flow rate of indicator entering the chamber by injection will be equal to the mass flow rate of indicator flowing out of the chamber in the outflow stream.

$$m_{f_1} + m_{f_i} = m_{f_2}$$

In terms of inflow and outflow concentration of indicator, this equation is

$$C_1 Q + m_{f_1} = C_2 Q$$

Solving for Q yields the equation on which the Fick method is based.

$$Q = \frac{m_{f_i}}{C_2 - C_1}$$

In practice, the indicator used in the measurement of cardiac output by the Fick method is oxygen so that m_{f_i} is the consumption rate of oxygen in the lungs, C_1 is the concentration of oxygen in the venous blood, and C_2 is the concentration of oxygen in the arterial blood.

THERMAL DILUTION METHOD FUNDAMENTAL EQUATIONS

For thermal dilution measurement of cardiac output, a warm or cold injectate is injected into the right atrium and the temperature of blood in the pulmonary artery is measured by a thermistor as shown in Fig. 5. A warm injectate would need to be considerably warmer than the blood that might be hot enough to denature proteins (60°C). If it were not very warm, the poor signal/noise ratio would render the method unusable. Therefore cold injectate is the only practical thermal indicator (7,8).

A bolus of cold fluid is injected into the right atrium and the resulting temperature is recorded from the pulmonary artery. Conservation of energy requires that the total thermal energy entering the system during the procedure must be equal to the total thermal energy that leaves the system as the system returns to normal temperatures.

The thermal energy carried across a boundary by a differential volume of fluid is given by

$$dE = \rho C_P T(t) Q dt$$

where ρ is the density of the fluid, C_P is the specific heat of the fluid, $T(t)$ is the temperature of the fluid at any instant, Q is the volumetric flow rate of fluid crossing the boundary, and dt is the differential time.

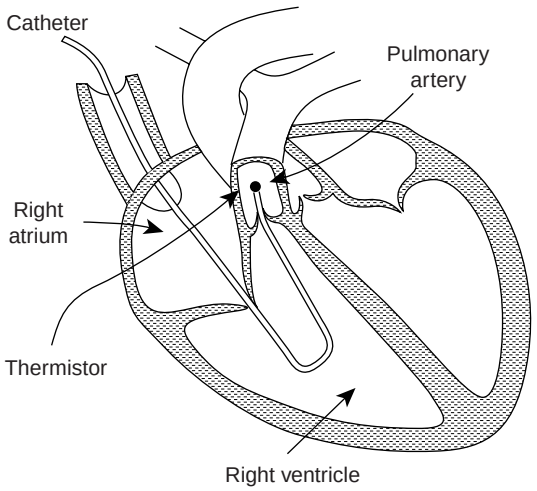


Figure 5. Catheter in place for thermal dilution measurement.

The total thermal energy that crosses a boundary for a constant volumetric flow rate with constant thermal properties is

$$E = \rho C_P Q \int_0^\infty T(t) dt$$

The total thermal energy that enters the system is a combination of energy carried into the system by blood flow and the thermal energy carried into the system by the injectate, where the subscript b refers to blood and subscript i refers to injectate.

$$E_i = \rho_b Q C_{P_b} \int_0^\infty T_b dt + \rho_i C_{P_i} V_i T_i$$

The total thermal energy that leaves the system is a combination of the energy carried out of the system by blood flow and thermal energy loss to the walls of the atrium and ventricle.

$$E_o = \rho_b Q C_{P_b} \int_0^\infty T(t) dt + \rho_i C_{P_i} V_i T_b + hA \int_0^\infty (T - T_b) dt$$

where $T(t)$ is the temperature recorded by the thermistor at any instant, h is the thermal convection coefficient, and A is the internal surface area of the right atrium and right ventricle.

Using these equations in the thermal energy conservation equation

$$E_i = E_o$$

and rearranging gives

$$\rho_b C_{P_b} Q \int_0^\infty (T(t) - T_b) dt = \rho_i C_{V_i} V_i (T_i - T_b) - hA \int_0^\infty (T(t) - T_b) dt$$

Solving for Q yields the thermal dilution equation for volumetric flow rate.

$$Q = \frac{\rho_i C_{V_i}}{\rho_b C_{V_b}} \frac{V_i (T_i - T_b)}{\int_0^\infty (T(t) - T_b) dt} - \frac{hA}{\rho_b C_{P_b}}$$

The term on the right represents the heat loss to the walls of the atrium and ventricle. In practice, there would be an additional heat loss in the catheter. The heat loses are normally accounted for by a correction term (K) that is a function of the catheter type being used as shown below.

$$Q = \frac{\rho_i C_{V_i}}{\rho_b C_{V_b}} \frac{V_i(T_i - T_b)}{\int_0^\infty (T(t) - T_b)dt} K$$

The thermal dilution method has the advantage that recirculation is not a problem due to the large surface available in the circulation to bring the injectate temperature to body temperature. A disadvantage is that the injection site and the sensing site must be close together to avoid large heat losses and the absence of total mixing in the ventricle can cause inaccuracies.

BIBLIOGRAPHY

1. Yang SS, Bentivoglio LG, Maranhao V, Goldberg H. From Cardiac Catheterization Data To Hemodynamic Parameters. 2nd ed. Philadelphia: F.A. Davis Company; 1980. p 55.
2. Geddes LA. Cardiovascular Devices and Their Applications. New York: John Wiley & Sons, Inc.; 1984. p 102–106.
3. Stewart GN. The output of the heart in dogs. Am J Physiol 1921;57:27–50.
4. Fick A. Uber die Messung des Blutstroms in den Herzventrikeln. Verhandl Phys Med Ges Zu Wurzburg 1870;2:XVI.
5. Hamilton W, et al. Comparison of the Fick and dye injection methods of measuring the cardiac output in man. Am J Physiol 153:309–321.
6. Valentinuzzi ME, Posey JA. Fast estimation of the dilution curve area by a procedure based on a compartmental hypothesis. Med Ins Sept-Oct1972;(6)5.
7. Taylor BC, Sheffer DB. Understanding Techniques for Measuring Cardiac Output. Biomedi Inst Technol May/June, 1990; 188–197.
8. Swinney RS, Davenport MW, Wagers P, Sebat F, Johnston W. Iced versus room temperature injectate for thermal dilution cardiac output. Ninth Annual Scientific and Educational Symposium. Soc Crit Care Med, May12–16,1980;137.

See also CARDIAC OUTPUT, FICK TECHNIQUE FOR; CARDIAC OUTPUT, THERMODILUTION MEASUREMENT OF; ECHOCARDIOGRAPHY AND DOPPLER ECHOCARDIOGRAPHY; FLOWMETERS, ELECTROMAGNETIC; TRACER KINETICS.

CARDIAC PACEMAKER. See PACEMAKERS.

CARDIAC OUTPUT, THERMODILUTION MEASUREMENT OF

EDWIN D. TRAUTMAN
RMF Strategies
Cambridge, Massachusetts

MICHAEL N. D'AMBRA
Harvard Medical School
Cambridge, Massachusetts

INTRODUCTION

The amount of blood pumped by the heart each minute, the cardiac output, provides a measure of the body's potential for supplying oxygen and nutrients and is relevant to assessing the condition of the heart. Taken together with various pressures, it is a key clinical indication of the heart's ability to meet the body's needs and an indirect indication of the status of those needs. In the clinical setting, measurement of cardiac output is required to guide drug therapy aimed at manipulating the function of the cardiac muscle (inotropic drugs) and the state of the systemic and pulmonary vascular resistance (vasoconstrictor and vasodilator drugs). Combined with ultrasound velocity data, cardiac output allows precise assessment of the status of mitral and aortic valve stenosis and regurgitation.

G. N. Stewart articulated the basic principle of indicator- dilution measurement of cardiac output in a landmark paper in 1897 (1). Stewart stated that if a substance was introduced at a constant rate into the flowing bloodstream and allowed to mingle with the blood, then the measured steady-state concentration of that substance downstream of the site of introduction would be inversely proportional to the flow rate (cardiac output). Of greater practical importance was his additional observation that if a small amount of the substance was introduced rapidly, then the cardiac output could still be computed. To do that one would divide the average rate at which the substance is introduced (total amount divided by measurement time) by the average concentration. Stewart called this technique the "sudden injection" method. In the late 1920s, W. F. Hamilton and his colleagues further investigated Stewart's sudden injection method (2,3). They found that the concentration curve from timed samples did not simply return to baseline, but exhibited a secondary rise. This was attributed to fast physiologic recirculation of an unknown amount of the indicator. To eliminate the influence of any recirculating indicator on measurement calculation, they proposed extrapolating the original down slope of concentration to zero using an exponential function. This method proved successful in validation studies both in mechanical models and animal experiments, and the sudden injection method with exponential extrapolation is commonly referred to as the Stewart–Hamilton method.

Various indicators have been used to measure cardiac output with the Stewart–Hamilton method, notably saline (detected by its effect on electrical conductivity) and optical dye (detected by its effect on optical absorption), but G. Fegler's proposal in the mid-1950s that heat could be used has proved the most convenient, although initially controversial (4,5). His earliest report "was received with polite incredulity" (6). Fegler rapidly injected a small amount of cold Ringer's solution into the vena cava and recorded the transient decrease in temperature in the aortic arch and in the right ventricle. He computed both left and right heart outputs from these data.

Concerns were voiced regarding the ability to quantify this "negative" indicator, the stability of the baseline temperature, and the background noise (6). These are all valid concerns, but concerns that have been successfully addressed with clinical technology. Current practice is to introduce a small bolus of cold solution into the right atrium (via a venous catheter) and to measure the

consequent transient temperature decrease in the pulmonary artery. With the advent of balloon-flotation pulmonary-artery pressure measurement catheterization techniques in the late 1960s (7), small thermistor sensors could be readily placed in the pulmonary artery, and the thermal dilution measurement became clinically accepted even as validation experiments progressed. The pulmonary-artery catheters provide important hemodynamic pressure information and, for that reason alone, are placed in many patients.

Swan and Ganz are credited with developing and popularizing the pulmonary-artery pressure catheter containing a thermistor and injection port for thermal dilution (8), and these catheters are commonly called Swan–Ganz catheters, although Swan–Ganz, strictly speaking, is a registered trademark of Edwards Lifesciences Corporation. The instrumentation required to process the temperature signal and determine cardiac output is modest, fitting into either a small, battery-operated instrument easily used at the patient's bedside, or into a module component of a bedside workstation in the ICU or operating room, making the method easy and convenient. The additional "invasion" of a thermistor is negligible, and the additional value of cardiac output measurements is great. And since the indicator is a physiologically innocuous solution, thermal dilution measurement of cardiac output has become an important part of clinical care. Today, bolus thermal dilution cardiac output is considered the gold standard against which other methods are compared.

THEORY

Principle of Indicator Dilution

The basic principle of indicator dilution is quite simple: If the concentration of a uniformly dispersed indicator in an unknown volume is measured, then the unknown volume can be simply determined by dividing that concentration into the total amount of indicator. If the volumeis flowing past a sensor, then the volume in any given period of time will equal the amount of indicator in that period of time divided by the concentration. If the rate at which the indicator is flowing past the sensor is controlled and known, then the amount of indicator in a period of time is also known and the volume flow rate can be determined. Alternatively, if the total amount of indicator over a larger period of time is known, such as when a bolus is introduced all at once, then average flow rates can be determined. Each approach has strengths and weaknesses in technique, necessary assumptions, and equipment. We focus on the popular bolus technique but, particularly with thermal techniques, the theory can be extended.

In the case of thermal dilution, the indicator is introduced into the right atrium and its concentration is measured in the pulmonary artery. We assume that all of the indicator introduced, an amount I, eventually passes into the pulmonary artery at some rate $i(t)$. If we assume no indicator recirculates, we may write this as

$$I = \int_0^\infty i(t)\,dt \tag{1}$$

$$= \int_0^\infty F(t)c(t)dt \tag{2}$$

where F is volumetric flow and c is concentration. When the flow is constant, F can be moved out of the integral and we can solve for F as

$$F = \frac{I}{\int_0^\infty c(t)dt} \tag{3}$$

Several assumptions have been made in arriving at this equation. Equation 1 is a statement of conservation and requires that all indicators pass the sensor exactly once. Equation 2 requires that the concentration in the pulmonary artery be uniform across the area where the concentration is being measured, and equation 3 requires that the flow rate be constant. All but the first requirement can be satisfied for the pulmonary artery catheter-based measurements if we consider the right heart to be a perfect mixing chamber, with a competent valve at the outflow, and a pumping rate that is constant over the integration time. The mixing chamber guarantees that the blood will be equivalently labeled at its outflow so that each flow stream is representative of the total, and the valve guarantees that the concentration changes in a stepwise fashion in all flow streams, which allows legitimate averaging of pulsatile variations in flow. A rigorous proof can be found in Perl et al. (9) and the assumptions and necessary conditions are discussed in Trautman and Newbower (10). Each of these articles contains numerous relevant references.

Heat as an Indicator: "Thermal" Dilution

In thermal dilution, the indicator is caloric, introduced as a known volume of a cold physiologic solution whose concentration is measured via the induced temperature change. The relationship between temperature and the concentration of heat in a solution, the amount of heat in a unit volume, involves the specific heat and density of the solution. When two solutions at different temperatures, such as the indicator and blood, are mixed, the temperature of the mixture may be predicted by

$$T = \frac{T_1 C_{p1} m_1 + T_2 C_{p2} m_2}{C_{p1} m_1 + C_{p2} m_2} \tag{4}$$

where C_p and m are, respectively, the specific heat and mass of the solutions. If we take the first solution to be the indicator solution and the second to be the blood, then the difference in temperature due to the indicator will be predicted by

$$T - T_2 = \frac{(T_1 - T_2) C_{p1} m_1}{C_{p1} m_1 + C_{p2} m_2} \tag{5}$$

A very good assumption, at least prior to significant heat exchange with tissue, is that the indicator solution and the temperature transient travel together. In this case,

equation 5 holds for all instances of time, and the mass concentration of the indicator solution may be predicted from the temperature transient by

$$c(t) = \frac{[T(t) - T_2]\rho_1}{T(t) - T_2 - (C_{p1}\rho_1/C_{p2}\rho_2)[T(t) - T_1]} \tag{6}$$

where ρ is the density of the solution. The amount of indicator is equal to the volume of the physiologic solution V times its density ρ_1 (giving its mass) and equation 3 becomes

$$F = \frac{V\rho_1}{\int_0^\infty c(t)\,dt} \tag{7}$$

$$F = \frac{V\rho_1}{\int_0^\infty \dfrac{[T(t) - T_2]\rho_1}{T(t) - T_2 - (C_{p1}\rho_1/C_{p2}\rho_2)[T(t) - T_1]}dt} \tag{8}$$

Equation 8 forms the basis for the thermal dilution method for measuring cardiac output. This equation may be easily programmed, but it generally has been approximated to simplify implementation, The most common approximation is based on the assumption that the indicator solution has no effect on the thermal properties of the blood. In this case, the increment in the amount of heat leaving the heart due to the indicator is equal to

$$H = \rho_2 C_{p2} F \int_0^\infty [T(t) - T_2]dt \tag{9}$$

This is equivalent to equation 2 and requires the same assumptions. The amount of heat added in a certain volume of an indicator solution is equal to

$$H = \rho_1 C_{p1} V(T_1 - T_2) \tag{10}$$

Equating equations 9 and 10 leads to the simpler formula for flow:

$$F = \left(\frac{C_{p1}\rho_1}{C_{p2}\rho_2}\right)\frac{V[T_1 - T_2]}{\int_0^\infty [T(t) - T_2]dt} \tag{11}$$

Equations 8 and 11 are equivalent only when cool blood is used as the indicator. Equation 11 is simpler than equation 8 and can be implemented in an analog circuit. However, it is based on the implausible condition that the indicator solution carries heat (or cold) into the blood and then is either transported completely apart from the thermal transient or has no thermal effect on the blood. Fortunately, the practical difference between flow estimates based on these two equations is small. The ratio of thermal properties for a dextrose-in-water (D5W) indicator solution is approximately equal to 1.08, and for a normal (0.9%) saline solution it is approximately equal to 1.10. Since the expected temperature transient is 0.5–1.0 °C, the expected difference between equations 8 and 11 is only 1–2% for these indicators.

If the indicator is not introduced as a finite bolus, then the conservation statement of equations 1 and 2 needs to be generalized and other assumptions made. The product of flow rate and concentration at the outflow of the mixing chamber will still be equal to the amount of indicator passing by, but its relationship to the input indicator can be more complex. A simple example is where the indicator is infused at a constant rate in which case, absent recirculation, the flow rate will be inversely proportional to measured concentration. If the infusion rate is not constant then the transient response of the heart system needs to be considered.

Heat can be introduced by direct energy transfer, such as from an electrical heater. In this case, the volume factor in the numerator of equation 11 is not relevant and must be replaced by a measure of the amount of heat introduced. It is, however, impractical to introduce a large bolus (impulse) of heat comparable to the 750 W of 10 mL of iced saline: The surface temperature would be dangerously high. Instead, the heater is pulsed at low power, the resulting temperature changes measured with a fast-response thermistor, and sophisticated signal processing used to extract the dilution signal from the baseline. Such techniques have the potential to measure cardiac output continuously and were introduced in the early 1980s (11). Catheters with heating elements (10 cm long filaments in the right ventricle) have been produced since the early 1990s (12). The surface temperature and thus the amount of heat that can be introduced are limited by physiological concerns (4–7 °C), and therefore the technique is sensitive to background thermal noise. The heater is typically pulsed, and the accuracy is dependent on processing the correlation between the heating waveform and the measured temperature response (11–13). These techniques are entering clinical use as a companion to bolus thermal dilution, but are not considered here.

Necessary Conditions

For the preceding development, we assumed that the flow is constant; the volume and temperature of the solution are known, the indicator does not recirculate, and perfect mixing occurs somewhere between injection and sampling. Little can be done to control the variation of flow; it is flow that is being measured. (In those situations where flow is *not* constant, it can be shown that the computed result will be a concentration-weighted average of the true cardiac output over the period of measurement.) The other assumptions are usually reasonable, although in practice it is difficult to have an accurate measure of the volume or temperature of the injected solution since heat will exchange with all material contacting the solution. There is also a lost volume in the dead space of the catheter used to introduce the solution, and it is impossible to eliminate the physiologic recirculation of indicator. In addition, the integrals must be truncated to permit a practical measurement. These and other practical issues are covered next.

Notably absent from these formulas is the time response of the thermal sensor. Although not intuitively appealing, it can be shown that this response is of little importance as long as the curve does not become distorted by the effects of noise and recirculating indicator. The area under the thermal curve is preserved even with slow-responding thermal sensors. The operator should, however, be aware that the thermal curve obtained with a slow thermistor is not necessarily a high fidelity representation of the

temperature transient. The observed or recorded temperature curve will be smoothed and filtered over time.

PRACTICAL APPLICATION OF THE THEORY

The practical application of thermal dilution theory is simple; all that is typically necessary to measure cardiac output is to reset the "cardiac output computer," inject 2–10 mL of an ice-cold or room temperature solution into the catheter port, and wait for the answer to appear on the computer. The thermal sensor is contained on a special pulmonary artery catheter that also provides the injection lumen into the right atrium. The temperature curve may be recorded to reassure the operator that a reasonable signal was obtained. Typical curves for various flow rates are shown in Fig. 1.

When the computer is reset, it samples the baseline temperature, integrates the processed temperature curve until recirculation is detected or assumed, and calculates the cardiac output assuming predefined conditions. It applies a correction for the portion of the curve that is lost by the truncation of the integral (to avoid influence of recirculation). The predefined conditions include the volume and temperature of the injectate, the thermal characteristics of the fluids, and a correction factor that corrects for the physical properties of the particular injection catheter employed. Most computers make approximations and apply corrections. The most common are listed below.

Thermal Properties of Blood Are Approximately Constant

The thermal properties of blood vary with hematocrit. However, the convenience of assuming a normal hematocrit—and thus not requiring knowledge of the actual hematocrit and entering it—far outweighs the importance of the potential error. The specific heat-density product for erythrocytes is $\sim 3.52\ J{\cdot}K^{-1}{\cdot}mL^{-1}$, and for plasma it is $\sim 4.03\,J{\cdot}K^{-1}{\cdot}mL^{-1}$ (14). Therefore, for blood with a hematocrit of 40% this product is $3.83\ J{\cdot}K^{-1}{\cdot}\ mL^{-1}$, with a hematocrit of 30% it would be $3.88\ J{\cdot}K^{-1}{\cdot}\ mL^{-1}$, and with 50% it would be $3.78\ J{\cdot}K^{-1}{\cdot}\ mL^{-1}$. Thus, the nominal value assumed for blood could be in error by 1–2% causing an error in the cardiac output measurement of the same value.

Indicator Does not Affect Thermal Properties of Blood

The specific heat-density product for normal (0.9%) saline is $\sim 4.19\ J{\cdot}K^{-1}{\cdot}\ mL^{-1}$, and that for 5% dextrose-in-water is $\sim 4.11\ J{\cdot}K^{-1}{\cdot}\ mL^{-1}$ (15). These are significantly different from the nominal $3.81\ J{\cdot}K^{-1}{\cdot}\ mL^{-1}$ for blood. The thermal properties of the indicator-blood mixture will thus vary with the level of dilution. However, most computers assume that the indicator does not affect the thermal properties of blood. This assumption leads to the simpler equation [Eq. 11] derived above. Cardiac output computers using this approximation can be expected to overestimate the cardiac output by 1–2%.

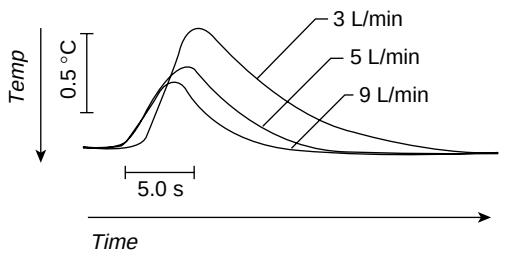


Figure 1. Typical thermal dilution curves, taken at different flows and superimposed to illustrate variations in shape and area.

Heat Loss Is Predictable

When the syringe containing the cold solution is taken from the ice bath it immediately begins to warm. The solution warms further as it is injected through caloric exchange with the walls of the catheter. Only a negligible amount of heat is gained during manipulation of the syringe before injection. However, the exchange with the wails of the catheter can account for several percentages with a 0 °C solution (16). In addition to those conductive losses of indicator, a significant amount of solution is left in the catheter after the injection has been terminated. The typical dead space volume is 0.9 mL so that only 91% of the solution is injected into the bloodstream. However, the solution that filled the dead space prior to injection is pushed into the blood stream and, if not at blood temperature, can add to the effective indicator volume. In addition, some of the "cold" left in the dead space after injection can leak through the catheter wall and add to the injectate.

Empirical studies have shown that the combination of these losses and gains can be grouped into a single correction factor, multiplying the total indicator volume. This correction factor varies only a few percentages with catheter insertion length and other mechanical factors (17). The correction factor does depend on the temperature and volume of the injected solution and on the design of the catheter. Catheter manufacturers generally provide, with their package inserts, a table of values for the correction factor or "computation constant" under various typical conditions. This factor is determined by measuring the average temperature of the injectate, as it emerges from the injectate, lumen of the catheter, while the appropriate length of catheter is immersed in a 37 °C bath. The amount of injectate that emerges is a reasonably constant fraction of the amount introduced.

Devices can be used to measure the temperature of the injected solution as it enters the injection catheter. This reduces the need for precisely controlling the initial temperature of the solution and reduces errors due to warming of the solution during handling. These devices do not improve knowledge about the unknown heat loss during injection. Catheters have also been fabricated with thermistors in the distal port of the injection lumen, to measure true injectate temperature. These catheters demonstrate better reproducibility particularly with room temperature injectates, but have yet to win clinical acceptance due to cost and complexity. Note also that the rate at which the indicator is introduced must be controlled and consistent to allow inferring amount of heat from temperature.

Decay of Dilution Signal Is Exponential

Recirculation of the thermal indicator is relatively small in humans since there is ample opportunity for exchange with the tissue beds. In smaller animals, recirculation is more apparent. In either case, the decay of the temperature signal measured in the pulmonary artery approximates an exponential (the result of the mixing chamber) and, once truncated, the true but obscured curve can be mathematically extrapolated with reasonable accuracy. The relatively small amount of curve area being estimated limits the significance of errors in extrapolation. Some cardiac output computers actually fit an exponential to the uncorrupted curve and use the parameters of fit to extrapolate the curve, while most assume that curves generally have the same shape and integrate to a fraction of the peak temperature and multiply this area by a constant. Either method appears to result in a reliable measure of cardiac output. Certain pathologies can alter the shape of the thermal dilution curve and reduce the effectiveness of these extrapolation procedures.

Baseline Temperature Is Constant

The baseline temperature is not constant, varying with the respiratory cycle and subject to the fluid infusions from other sources (i.e., intravenous fluid administrations). In addition, the heart itself generates heat that can be observed as very small pulsatile variations in temperature in the pulmonary artery. Fortunately, these variations and the baseline shifts are usually small compared to the ~ 0.5–$1.0\,^{\circ}C$ dilution signal. Cardiac output computers thus assume that the baseline acquired prior to the arrival of the dilution signal remains constant during the course of the measurement. (Note that shifts in baseline can adversely affect the extrapolation procedure used to reduce the effects of recirculation.)

Flow Rate Is Constant Throughout the Integral

Cardiac output can vary by as much as 10–20% over the respiratory cycle. Since thermal dilution measurement integrals typically average only 5–10 s of the cardiac output, the measured cardiac output could vary by as much as 10–15% depending on where in the cycle the injection is made. There is really nothing the computer can do about this without information about the phases of the respiratory cycle. Cardiac arrhythmias, which can result from the cold injection, can cause dramatic errors in the measured output. The clinical practice of averaging several separate cardiac output determinations helps to average out some of the potential variation from both of these sources. See section on *Measurement Performance* for more discussion on accuracy and reproducibility.

EQUIPMENT

The thermal dilution method for cardiac output measurement is popular because it is easily performed with a minimum of equipment and little additional invasion of the patient. The basic equipment consists of a pulmonary artery catheter to position a thermistor or other temperature-sensitive element in the pulmonary artery, a means for making thermal indicator (usually saline) injections into the right atrium (usually a syringe), typically through a separate lumen in the catheter, a source of measured volumes of a cold solution, and an electronic instrument to determine the blood temperature from the thermistor signal, to determine and integrate the dilution signal, and to compute a final result. Each of these elements is described separately.

Pulmonary Artery Catheters

The pulmonary artery (PA) catheter generally contains several lumens (channels) that terminate at measured distances from the tip. A balloon, located at or near the tip, is inflated during catheter insertion to carry the tip through the heart and into the pulmonary artery (flow directed). One lumen terminates at the tip and is used to measure the pressure during catheter insertion to follow its position relative to the heart; later it measures pulmonary artery pressure and, intermittently, pulmonary capillary wedge pressure (with the balloon inflated). A second lumen typically terminates in the right atrium and is used to monitor right atrial pressure (central venous pressure). Indicator solutions are injected either through the right atrial port or through a second atrial lumen intended for drug infusion. Catheters can have several additional lumens (e.g., atrial and RV pacing wires) and sensors (e.g., mixed venous oxygen saturation). The pulmonary artery catheter provides important hemodynamic information and may be inserted in patients for that purpose alone.

For use with thermal dilution, the pulmonary artery catheter is augmented by adding a thermistor proximal to (before) the balloon (typically, 4 cm from the tip). A thermal dilution catheter is illustrated in place in Fig. 2. The

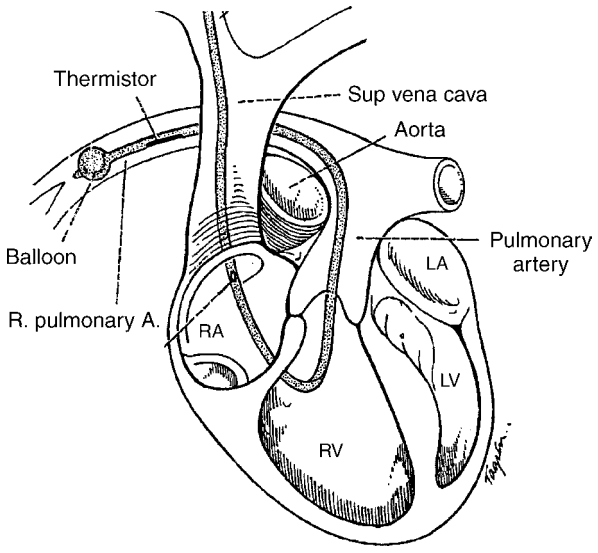


Figure 2. A pulmonary artery catheter in place in the right heart. The balloon, shown inflated here for wedge pressure measurement, normally remains deflated during pressure monitoring and cardiac output measurements. The cold injectate enters the bloodstream through the injection port, and the temperature transient is sensed by the thermistor.

thermistor typically is encapsulated in glass and coated with epoxy to fully insulate it electrically from the blood. The relatively slow time response of this encapsulated sensor does not affect the accuracy of the measurement since the area under a temperature curve is preserved. Wires connecting the thermistor are contained in a separate lumen. The thermistor wires terminate in an external connector that typically contains an electrical resistance used to standardize the response of the thermistor. Thus the catheter contains one-half of a Wheatstone bridge. The overall length is ~ 100 cm, with distance marks every 10 cm to guide insertion. The typical size is 7–8.5 Fr although pediatric catheters may be 5–6 Fr. (One French is equal to one millimeter in circumference.)

Edwards Lifesciences was the original commercial manufacturer of the thermal dilution catheter, basing it on a concept acquired from Swan and Ganz, researchers involved with validation experiments. Their Swan–Ganz catheter was introduced in 1971. Since that time several manufacturers (e.g., Instrumentation Laboratories, Cobe, Abbott, Arrow) produce thermal dilution catheters, disposable items selling for $\sim$ \$50–80 each, although catheters with heating filaments and multiple sensors can cost \$200 and up. A thermal dilution catheter is shown in Fig. 3.

Pulmonary artery catheters are not without clinical complications. The threat of infection is always present. Clots can form on the poly vinyl chloride (PVC) catheter surface, but this complication has been mostly eliminated with anticoagulant coatings. Unfortunately, there are patients who have severe reaction to heparin (i.e., heparin induced thrombosis) and in these patients heparin-coated catheters need to be avoided. The catheter can become knotted in the right heart, a complication that requires a trip to the cardiac catheterization laboratory to resolve.

Most dangerous is the rare complication of the catheter tip puncturing the pulmonary artery. In normal use, the balloon is temporarily inflated, the balloon wedges into a branch of the pulmonary artery, flow through that branch of the pulmonary circulation is stopped and the pressure measured from the distal lumen will approximate the left atrial pressure. It is critical that nurses and physicians understand the waveforms associated with "permanent wedge" position to avoid pulmonary rupture. This complication is almost always associated with erosion of the pulmonary artery from a catheter permanently in the wedge position or with inflating a catheter that is in the distal pulmonary arterial position.

Figure 3. A Swan–Ganz pulmonary artery catheter produced by Edwards Lifesciences. The various lumens are accessed through individual Luer-Lok connectors fanning out from an external divider on the main catheter. The electrical connector for the thermistor wires is also connected to the main catheter at the same point. (Photograph courtesy of Edwards Lifesciences Corporation, Irvine, CA.)

The balloon is an important feature of pulmonary artery catheters since it plays a key role in the acquisition of pressure information in addition to facilitating placement of the catheter. Balloons are generally made from latex and designed to inflate beyond the tip of the catheter while not occluding the distal pressure lumen. This shields the tip reducing the tip trauma to the pulmonary artery. Manufacturers attach the balloon to the base catheter in a way that minimizes rough surfaces and overall size of the catheter while being durable. Latex-free balloon catheters are available for use in patients with latex allergies, but are expensive and have limited functionality, usually having only a single right atrial port. The non-latex balloon is also not as durable and measurements of wedge pressure must be kept to a minimum.

The size and material of the catheters can vary among manufacturers and models, both of which can affect their stiffness and thus the ease of insertion, and the size of the dead space in the injection lumen and, thus, the heat loss correction factor. The injection port may be larger in some catheters, reducing injection effort. The frequency response of the pressure measurement lumens may be different due to attention to details of fluid mechanics. Personal preference, reliability, and economic concerns are also clearly important in purchase decisions. Some catheters offer other capabilities such as continuous cardiac output measurements based on advanced signal processing algorithms, mixed venous oximetry, and the ability to electrically pace right atrium and ventricle.

Cold Solution

The indicator solution is typically an isotonic saline or dextrose solution cooled to 0 °C by placing the bottle or prefilled capped syringes in an ice bath. This makes the injected indicator ~ 37 °C cooler than body temperature. This 10 mL of 37 °C difference injected in 2 s represents extraction of thermal energy from the bloodstream at a rate of ~ 750 W. Cold indicator is most useful in the operating room where patient temperatures can vary rapidly and dramatically. In the non-OR setting, or in operative patients where normothermia is expected, room-temperature solutions are frequently used because they are more convenient, but the variable room temperature must be monitored by the computer. The injected energy is reduced by a factor of ~ 3, reducing the signal-to-noise ratio and, thus, expected measurement performance. Most cardiac output computers provide a temperature probe to measure the actual temperature of the bath or of the room, which is presumably the temperature of the injectate. Manufacturers also produce an optional temperature-measuring probe that attaches to the injection port on the thermal dilution catheter and measures the temperature of the injectate as it enters the catheter. This further reduces concern about the actual room or bath temperature. Some catheters also have thermistors at the injection port itself.

The ice bath is the subject of some unproved concerns regarding infection. Undesirable organisms could remain

Figure 4. A closed injection system for cold injectate, by Edwards Lifesciences. A cooling coil rests in a Styrofoam ice bucket, the syringe serves as a piston pump to draw up a known volume of precooled injectate from the coil and to force it into the injection lumen of the catheter. The closed system reduces the risk of nosocomial contamination associated with traditional injectate delivery methods. CO-Set+ System improves reproducibility and accuracy through its in-line temperature probe and volume-limited syringe. (Photograph courtesy of Edwards Lifesciences, Irvine, CA.)

or grow in the capped syringes when left in the bath for long periods of time. A cleaner alternative to the ice bath is offered by a closed injectate system with a cooling coil, offered as an optional accessory by some manufacturers. One such device is shown in Fig. 4. The coil sits in an ice bath keeping the solution cold. The syringe is used to draw up solution and then immediately introduce it into the catheter port. Fig. 5 shows the complete system.

Although these solutions are generally benign, in some circumstances, such as in small pediatric patients, there is a risk of volume overload from frequent measurements. In these situations, smaller volumes are used (e.g., 3 mL) and fewer measurements are made.

Cardiac Output Computers

When thermal dilution measurements were first introduced in the early 1970s, manufacturers produced catheters with nonstandardized thermistors. Each manufacturer then produced a computer to mate with its catheter. In addition to generic differences in thermistor types, each individual thermistor of a given type can have a different temperature response requiring the operator to enter a calibration constant, idiosyncratic to the specific catheter, its thermistor response, and even the patient's blood temperature. In current products, the thermistor connector contains an electrical resistance selected to match the particular thermistor and complete a half Wheatstone bridge with a standard response. The value of the resistance in this circuit is chosen such that the voltage response of the half bridge will be the same for all thermistors of a given family and also will be linear near 37 °C. The nearly linear range is ~ 20 °C. With these catheters, the catheter is

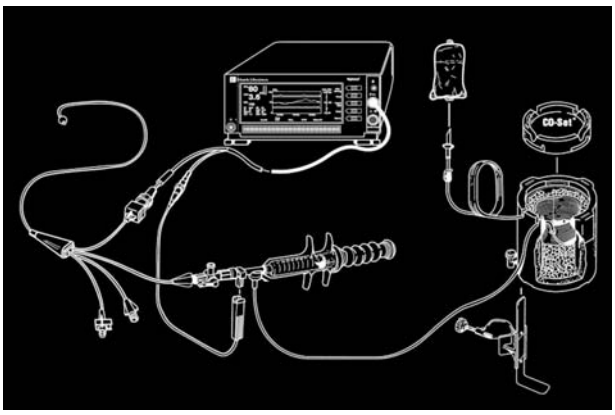


Figure 5. Basic thermal dilution system, with computer, syringe, catheter, and source of cold injectate, in this case from a closed system. (Courtesy of Edwards Lifesciences, Irvine, CA.)

merely connected to the computer's electronics, and the pulmonary artery blood temperature and the cardiac output can be measured. The electronic circuitry used to measure the thermal response is electrically isolated, since the thermistor is in the conductive bloodstream quite close to the heart and insulation failures can conceivably occur. Since the catheter functions as half of a Wheatstone bridge, the electronics merely mimic the other half of the bridge circuit, excite the bridge with a low level of current, and amplify the voltage difference proportional to temperature change.

When the operator signals that a measurement is to be made, the baseline temperature is acquired and the indicator concentration is computed (or approximated) and integrated. When the "end of curve" criterion is reached, the integration is stopped, the integral is adjusted for area lost by truncation, and the final area is inverted and multiplied by the appropriate constants for the measurement conditions. This constant is entered in the computer and only changed when conditions change. Cardiac output computers differ both in the method they use to truncate the integration and in the options they offer, such as, syringe size, injectate temperature, and integration into bedside systems.

Calibration of Equipment

Each catheter is individually calibrated by the manufacturer to give a standard response (as described previously), and the heat loss correction factor is determined also by the manufacturer for the particular catheter model for a variety of measurement conditions. No operator calibrations are necessary or practical.

In certain research settings (e.g., custom-made catheters), it is desirable to add the calibration resistor to the catheter. The value of the resistance is given by

$$R = R_0(\beta - 2T_0)/(\beta + 2T_0) \qquad (12)$$

where R_0 is the thermistor resistance at 37 °C, T_0 is 310 K, and β is the characteristic temperature (a gain constant) for the thermistor, equal to 3500 K for those used in thermal dilution catheters compatible with the Edwards Lifesciences standard.

Example Equipment

Most cardiac output computers are fully integrated into hemodynamic monitoring systems. The cardiac output component is usually part of the temperature-sensing module. Data from the measurements are acquired into the system's data recording and analysis packages and automated calculations of systemic and pulmonary vascular resistances are obtained.

In the typical computer, the preamplifier is fully isolated and uses a conservative 7 μA to sense the thermistor resistance. When a new cardiac output is desired, the operator presses a button and rapidly injects the cold solution. The dilution curve is typically shown as it is measured and the result is displayed once the curve has finished. Analog and digital outputs may be provided for integrating into a larger measurement or workstation system. In a typical computer the temperature difference is integrated from baseline up to its peak, then down to 30% of the peak value, and multiplied by 1.22. This integral is then inverted and multiplied by the computation constant provided by the catheter manufacturer. The computation constant is the product of all constants (e.g., the ratio of thermal constants, the injectate volume, 60 $s \cdot min^{-1}$, and 0.001 $L \cdot mL^{-1}$) and the catheter heat-loss correction factor (e.g., 0.825). Although the method for extrapolating the integral appears overly simple, it is quite effective. As pointed out earlier, the dilution curve has a consistent shape, and amount of area obscured by recirculation is small and comes relatively late in time (in humans). Some computers use more elaborate methods.

A typical computer and display is shown in Fig. 6. This example from Philips Medical Systems integrates into the monitoring system and computes several derived values as well as displaying hemodynamic information.

MEASUREMENT PERFORMANCE

Direct measurement of cardiac output is quite difficult given the location of the measurement site and the necessity to divert flow in some manner. The performance of thermal dilution measurements of cardiac output has been assessed in a number of less direct ways. Validation studies have been performed in mechanical flow models to assure that the measurement theory is sound in practice (and to determine the heat-loss corrections appropriate to specific catheters). Simultaneous thermal dilution and dye- dilution measurements, and also thermal dilution and direct Fick measurements, have been performed in animals and humans. Comparisons have also been made with electromagnetic flowmeters in animal preparations. All of these methods have shown thermal dilution to be effective for measuring cardiac output and as accurate as these other methods. In addition, an important consideration in the clinic is the reproducibility of the measurement over time and from operator to operator. Clinical studies of this sort have shown thermal dilution to be reliable and it is now considered the gold standard against which other measurements are compared.

It is interesting to note that the dye-dilution method was the incumbent standard, using indicators, such as indocyanine green dye measured by withdrawing blood from an artery through an optical sensor. This technique measures somewhat different flows—left-heart output rather than right-heart output, for example—and is subject to other issues of physiology and technique, such as greater recirculation and accumulation of indicator. Nevertheless, dye-dilution was clinically useful and thermal dilution was shown to be better and more convenient. By the early 1980s, thermal dilution was the technique of choice.

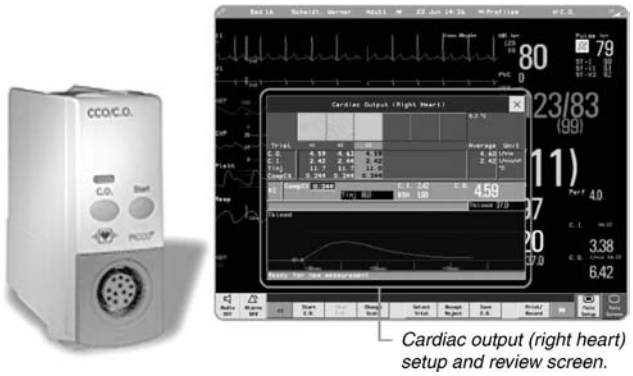


Figure 6. A modern cardiac output computer, integrated into a monitoring system, by Philips Medical Systems. The thermal dilution curve is displayed for inspection, and the cardiac indices can be automatically computed from the cardiac output and body parameters entered by the clinician. (Photograph courtesy of Philips Medical Systems, Andover, MA.)

Accuracy

The accuracy of thermal dilution is degraded by the various assumptions and approximations discussed previously. Thus, even in the absence of physiologic noise, this measurement can only be expected to be within 2–7% of the true value without other measurements and specific corrections relevant only to a research setting. This accuracy is, however, well within a clinically acceptable range and is no worse than that of other methods. In his first trials with this technique, Fegler compared thermal dilution with standard direct Fick measurements in animals, finding a discrepancy of $< 7\%$. Early validation studies with thermal dilution catheters were performed by Ganz and colleagues in the early 1970s (8,17,18). In addition to supporting the overall accuracy of thermal dilution, they determined that the sensitivity of the result to mechanical and technique-dependent factors, such as catheter insertion length and speed of injection, was within 3%. This they considered to be biologically insignificant (17).

Others have since obtained good correlation with simultaneous dye dilution and other methods for measuring flow, if not slopes of identity. It is interesting to note that since the dye concentration is usually measured in a systemic artery, dye dilution will provide a measurement of left heart output that is $\sim 4\%$ higher than the right heart output measured by thermal dilution, due to the bronchiolar circulation bypassing the right heart. In addition, all of these reference methods have some of their own

uncertainty in calibration, and conclusions are thus necessarily limited.

Exploration of the heat loss during injection has yielded interesting information on variability (16,19,20) but has not quantified the systematic loss to the point of accurate prediction of total injected heat (cold). The *in vitro* studies of effective losses and determination of correction constants provide adequate foundation for an accurate measurement.

Reproducibility

Cardiac output need not be known to great accuracy (within 10% is quite adequate) as long as the measurements are reproducible and can be used to track therapies. The reproducibility (variance) of the measurement with 10 mL of iced solution is generally accepted to be in the range of 10–15%. This is higher with room temperature solutions, and with smaller volumes (21). The reproducibility can be improved by 20–40% with a thermistor sensing the injectate temperature at the injection port in the right atrium. (20,22)

Some factors that can affect the reproducibility of this measurement derive from physiology and some from technique.

Physiology. As noted previously, the cardiac output can be expected to vary over the respiratory cycle, particularly with positive-pressure-assisted ventilation. This is shown quite succinctly in a careful study in animals by Jansen et al. (23) where the injection was made at random, but at known phases of the ventilation cycle. When plotted sequentially in time, the results span a range of $\pm 15\%$ and appear randomly distributed. When ordered according to the phase of the ventilator, the cardiac output result varies cyclically over the course of ventilation, Therefore, a determination of cardiac output using an arbitrary injection time could differ from another determination at another arbitrary time by as much as 30% due, presumably, to real physiologic variations, with flow modulated by intrathoracic and intra-abdominal pressure. Some contribution of baseline drift and thermal noise cannot be discounted by this study.

Clinically, the baseline temperature is usually assumed to be constant during the period of the measurement. Yet blood temperature varies by as much as $0.1\,^{\circ}C$ over the ventilatory cycle due to differential blood return from the upper and lower extremities. This fluctuation in baseline temperature is typically small compared with the $\sim 1\,^{\circ}C$ thermal dilution signal obtained with $0\,^{\circ}C$ injectate, but extends over significant time. It is more significant with room temperature injectates and with heated-filament (continuous cardiac output) signals. The magnitude of the baseline drift can be much greater, particularly with patient movement. Of note, intravenous fluid infusions will affect the blood temperature enormously, particularly during flushes of the lines. The heat output from the heart itself returns into the right atrium from the coronary veins in synchrony with the heartbeat. These pulsations in temperature are less pronounced, being smoothed by the mixing volume in the right ventricle, and cause little practical difficulty.

Since the indicator is introduced immediately upstream of the heart, the solution can, conceivably, transit the heart in a single beat (or very few beats) if the ejection fraction is high. (Thermal dilution curves obtained with fast-response thermistors can be used to determine the ejection fraction by quantifying this washout time when the injection is made directly into the ventricle.) Therefore, these few beats must be representative of the average output for the measurement to be useful. An arrhythmia at the time of injection, occasionally caused by the injection, can lead to a single very large ejection with a very good ejection fraction. In this situation, the measured output will be much larger than the true average cardiac output. The method is not in error, but the measured output is not the steady-state output. Therefore, if arrhythmias are suspected, the measurement should be discarded or a very large variation in results anticipated.

Certain pathologies can affect thermal dilution cardiac output measurements. Tricuspid regurgitation will increase the effective mixing volume for the indicator, thus increasing the extent and decreasing the magnitude of the thermal transient. However, it is important to note that TR does not invalidate the fundamental physical principles upon which the measurement is based. If the computer can wait long enough, the CO measurement in the face of TR should be accurate. In order to be sure this is the case, the practitioner must watch the thermal dilution curve as it evolves on the monitor screen.

Very low ejection fractions can have a similar effect. Although the basic assumptions underlying the measurement remain intact, the curve can be distorted to an extent that makes the practical measurement unreliable. More serious problems are caused by an incompetent pulmonic valve. This valve is necessary to minimize the nonlinear averaging effects of the pulsatile flow, and any flow reversal at the thermistor can lead to multiple re-measurement of the thermal transient. Either of these effects degrades the cardiac output determination.

Technique. Thermal dilution measurements are reasonably insensitive to variations in operator technique. As noted above, the injectate will not warm significantly as the syringe is handled briefly prior to injection. And if this is a concern, probes can be used which measure the temperature of the injectate as it is injected. The content of the injection catheter dead space must be considered to achieve a high level of reproducibility. If multiple measurements are made over a short period of time, that is, to average several serial determinations, sufficient time must be allowed for the residual injectate to return to blood temperature. A couple of minutes appears sufficient to warm the dead space as well as to allow the blood temperature to return to a stable baseline. One strategy is to discard the first measurement, using it merely to fill the dead space with a cool solution. Another strategy for assuring a consistent effect from the dead space is to withdraw blood immediately following the injection. The potential for blood clotting, however, limits the applicability of this procedure. As noted earlier, some catheters can measure the temperature of the injectate at the point of injection (20,22) thus minimizing these effects.

In summary, judicious choice of the time of injection can improve reproducibility. Attention should be paid to the phase of ventilation, to changes in any concomitant intravenous fluid infusions, and to any concurrent cardiac arrhythmias. The temperature curve can be recorded from most cardiac output computers. This curve, and the prior baseline, can give the knowledgable operator evidence on which to judge the validity of a particular result. Recall, however, that the time course of the thermal curve is not necessarily the same as the time course of the thermal transient in the flow stream. Most clinicians use a single measurement to guide therapy although in many settings, such as in studies, it is still common practice to use the average of three serial cardiac output determinations or to discard the outlier and average the remaining two.

Ease of Use

The ease and robustness of thermal dilution measurements of cardiac output are probably responsible for its clinical popularity. The equipment is straightforward to operate, and specialized technicians are not needed to acquire reliable data. The right heart catheters may be placed for other clinical reasons without fluoroscopy. When a measurement of cardiac output is indicated, all that is necessary is to attach a computer and inject cold solution.

FUTURE DEVELOPMENTS

This measurement is simple, fundamentally inexpensive, and has remained popular for several decades. It is, however, moderately invasive. If the need for the pressure information from the pulmonary artery catheters was reduced or supplanted, the ease of making a thermal dilution measurement would diminish. There are liabilities and contraindications associated with pulmonary artery catheters and the injection of cold solutions, and this measurement is not always prescribed in critical care. Use of PA catheters is falling somewhat as other methods of assessing cardiac filling and function become more widely available, such as ultrasound-based measurements and central venous lines. However, several million pulmonary artery catheters are used each year in North America and their widespread use is likely to continue. It is interesting that many surgeons who must manage their patients via phone consultations rely heavily on PA catheter measurements, especially when the ICU team does not include experienced physicians.

Indicator-dilution measurements of the sort described in this article are fundamentally intermittent. In many cases, a continuous measurement would be favored. Continuous cardiac output measurement with the heated filament paired with advanced signal processing is becoming popular, and other techniques such as analyzing pulse contours are also becoming more accepted. Thermal dilution with 10 mL of iced solution is the standard against which these techniques are compared, and periodically calibrated (13).

Catheters will continue to improve, with better clot resistance, materials, additional lumens, heating elements and sensing elements as measurements demand. Cardiac output is integrated into measurement systems forming part of derived parameters and important correlations, a trend that will continue to follow medical instrumentation and healthcare information technology in general. And as the reliability of the measurements increases with experience and technology, the long-promised closed-loop therapies may become a reality.

BIBLIOGRAPHY

1. Stewart GN. Researches on the circulation time and on the influences which affect it. IV. The output of the heart. J Physiol (London) 1897;22:159–183.
2. Hamilton WF, Moore JW, Kinsman JM, Spurling RG. Simultaneous determination of the pulmonary and systemic circulation times in man and of a figure related to the cardiac output. Am J Physiol 1928;84:338–344.
3. Kinsman JM, Moore JW, Hamilton WF. Studies on the circulation. I. Injection method; physical and mathematical considerations. Am J Physiol 1929;89:322–330.
4. Fegler G. Measurement of cardiac output in anaesthetized animals by a thermo-dilution method. Q J Exp Physiol Cogn Med Sci 1954;39:153–164.
5. Fegler G. The reliability of the thermodilution method for determination of the cardiac output and the blood flow in central veins. Q J Exp Physiol Cogn Med Sci 1957;42:254–266.
6. Dow P. Estimations of cardiac output and central blood volume by dye dilution. Physiol Rev 1956;36:77–102.
7. Swan HJC et al., Catheterization of the heart with use of a flow-directed balloon-tipped catheter. N Engl J Med 1970;283: 447–451.
8. Ganz W, Swan HJ. Measurement of blood flow by thermodilution. Am J Cardiol 1972;29:241–246.
9. Perl W, Lassen NA, Effros RM. Matrix proof of flow, volume and mean transit time theorems for regional and compartmental systems. Bull Math Biol 1975;37:573–588.
10. Trautman ED, Newbower RS. The development of indicator-dilution techniques. IEEE Trans Biomed Eng 1984;BME-31:800–807.
11. Philip J et al., Continuous thermal measurement of cardiac output. IEEE Trans Biomed Eng 1984;BME-31:393–400.
12. Yelderman ML et al., Continuous thermodilution cardiac output measurement in intensive care unit patients. J Cardiothorac Vasc Anesth 1992;6:270–274.
13. Schmid ER, Schmidlin D, Tornic M, Seifert B. Continuous thermodilution cardiac output: clinical validation against a reference technique of known accuracy. Intensive Care Med 1999;25:166–172.
14. Spector WS, editor. Handbook of Biological Data. Philadelphia: Saunders; 1956; Mendlowitz M. The specific heat of human blood. Science 1948;107:97–98.
15. Diem K, editor. Documenta Geigy, Scientific Tables. Ardsley (NY): Geigy Pharmaceuticals; 1962.
16. Meisner H et al., Indicator loss during injection in the thermodilution system. Res Exp Med 1973;159:183–196.
17. Forrester JS et al., Thermodilution cardiac output determination with a single flow-directed catheter. Am Heart J 1972;83:306–311.
18. Ganz W et al., A new technique for measurement of cardiac output by thermodilution in man. Am J Cardiol 1971;27:392–396.
19. Vliers ACAP, Visser KR, Zijlstra WG. Analysis of indicator distribution in the determination of cardiac output by thermal dilution. Cardiovasc Res 1973;7:125–132.
20. Lehmann KG, Platt MS. Improved accuracy and precision of thermodilution cardiac output measurement using a dual thermistor catheter system. J Am Coll Cardiol 1999;33:883–891.

21. Bourdillon PD, Fineberg N. Comparison of iced and room temperature injectate for thermodilution cardiac output. Cathet Cardiovasc Diagn 1989;17:116–120.
22. Williams JE Jr., Pfau SE, Deckelbaum LI. Effect of injectate temperature and thermistor position on reproducibility of thermodilution cardiac output determinations. Chest 1994; 106:895–898.
23. Jansen JRC et al., Monitoring of the cyclic modulation of cardiac output during artificial ventilation. In: Nair S, editor. Critical Care and Pulmonary Medicine. New York: Plenum; 1980. p 59–68.

See also CARDIAC OUTPUT, FICK TECHNIQUE FOR; CARDIAC OUTPUT, INDICATOR DILUTION MEASUREMENT OF; CORONARY ANGIOPLASTY AND GUIDEWIRE DIAGNOSTICS; MICROPOWER FOR MEDICAL APPLICATIONS; THERMISTORS.

CARDIOPULMONARY BYPASS.

See HEART-LUNG MACHINES.

CARDIOPULMONARY RESUSCITATION

EDWARD GRAYDEN
Mayo Health Center
Albertlea, Minnesota

INTRODUCTION

Cardiopulmonary resuscitation (CPR) may be defined as the emergency restoration of vital functions in a person who has undergone a life-threatening event. The term "cardiopulmonary resuscitation" is actually misleading since the goal of all CPR is to return the victim to appropriate cerebral function; cardiopulmonary resuscitation is the vehicle by which the rescuer attempts to reach this goal. The process of resuscitation may be viewed as a continuum where at one end of the spectrum psychomotor skills of CPR may be initiated by a lay bystander who might be the first rescuer on the scene of an accident, witness to someone choking on food at a restaurant, or perhaps is present when a family member succumbs to a heart attack. Cardiopulmonary resuscitation may also be viewed in a more general and organizational sense to encompass the entire process of the emergency response to victims. The education and training of the public and first responders in basic life support, such as policeman and firefighters, is the cornerstone in an attempt to reduce sudden death through lifesaving skills. Training in basic life support focuses on providing the rescuer with the ability to recognize emergencies, activate the Emergency Medical System (EMS, 911), maintain an airway, provide effective rescue breathing and cardiac circulation. American Heart Association sponsored programs also focus on prevention of risk through education of the public regarding the etiologies of coronary artery disease, myocardial infarction (heart attack), and cerebrovascular disease (stroke). Information presented through these programs attempts to modify lifestyle patterns and behaviors, such as smoking, known to cause or exacerbate these events. The new focus in community emergency response is in the training of laypersons in the use of the Automatic External Defibrillator (AED). Documentation of successful resuscitation in communities with high proportions of the public trained in CPR and use of an AED reach 49% in out-of-hospital victims known to have suffered ventricular fibrillation (a terminal cardiac dysrhythmia that is a common endpoint in the progression toward death) (1,2). Currently, there has been significant progress made in making these automatic defibrillators present in communities and in public places, such as shopping centers, sporting event facilities and mass transportation. The American Heart Association "ABCs" of CPR (airway, breathing, circulation) have now been supplanted with the "ABCDs" (airway, breathing, circulation, defibrillation). The progression of CPR continues into Advanced Cardiac Life Support (ACLS) supervised by a physician and consists of BLS as well as sophisticated adjuncts to provide oxygenation and ventilation, intravenous access with administration of drugs that support circulation, monitoring of cardiac rhythms with rapid interpretation of dysrhythmias and subsequent maneuvers to terminate or suppress these harmful cardiac electrical abnormalities, and postresuscitation care.

This article will first review the history of CPR followed by a detailed analysis of the pulmonary and cardiac physiology relevant to the application of these resuscitative functions. An overview of Emergency Cardiac Care (ECC) will be undertaken to enlighten the reader about the organizational process guiding CPR. The actual mechanism of BLS and ACLS will be then addressed with a brief overview of defibrillators. Finally, the salient points of this article will be summarized and future directions of resuscitation will be explored.

HISTORICAL PERSPECTIVE

Restoration of life to the dying has been a common action from antiquity to the present time. Ancient attempts at artificial respiration have been described by the prophet Elisha in the Bible (3). Galen was able to observe the inflation of a dead animal's lungs in the second century, but there has been no recording of this significant finding applied to early attempts at resuscitation (4). Resuscitation methods during this time were futile—such as applying hot materials to the abdomen or whipping the victim; animal bladders were expanded with smoke and then the outlets of these bladders placed into the dying person's rectum (5). Centuries later Paracelsus, a Swiss physician (1493–1591), first reported the use of a fireplace bellows to ventilate a dying patient. In 1740, the Paris Academy of Sciences recommended the instillation of air into a victim through a mouth-to-mouth technique and within 4 years Tossach used this method successfully to revive a person (4). Ironically, this technique was lost, only to be rediscovered some 200 years later. During the eighteenth century, multiple new attempts at artificial respiration occurred. The "Inversion Method" practiced in Europe and America was used for drowning whereby the victim was hung upside-down in an effort to drain water from the lungs and many successful attempts have been recorded for this maneuver. The "Barrel Method" as well as the "Trotting

Horse Method" were also used at this time consisting of rotating the prone drowning victim over a barrel that alternated chest compression (expiration) and chest relaxation (inspiration) or placing the drowned individual prone on a horse, with the bouncing incurred during the trot inducing the same rhythmic compression and relaxation (5). The realization that alternating compression and relaxation of the chest could induce expiration and inhalation, respectively, led to direct manual efforts by the rescuer. DeHaen in 1783 first described a chest compression, arm-lift combination (6). Leroy reported the first use of the supine victim ventilation position $\sim$1830 and later in this century ($\sim$1860–1870s) Silvester's, Howard's, and Schafer's prone methods of manual compression became popular and persisted into the twentieth century. The familiar Schafer–Emerson–Ivy ventilation method of scapular compression combined with pelvic-lift emerged in the United States at the beginning of this century.

The efficacy of these various methods of manual artificial respiration was resolved in the 1950s by Gordon, who performed experiments upon fresh corpses prior to rigor mortis and then on volunteers who underwent general anesthesia and paralysis by curare. Ventilatory volumes were measured and the "push–pull" maneuvers that caused active inspiration and expiration were at least twice as effective as the Schafer method or other procedures that only produced either active inspiration or expiration (7–9). The Holger–Nielson method (prone back-pressure, arm-lift) for resuscitation became the standard of care.

At the time of these scientific studies attempting to clarify manual methods of artificial respirations, Elam elected to evaluate the physiology of mouth-to-mouth ventilation. As an anesthesiologist, Elam had serendipitously performed mouth-to-mouth ventilation to paralyzed polio patients, for as long as several hours. Though mouth-to-mouth or mouth-to-nose ventilation had been know to have been practiced by midwives for the newborn, the question posed by this physician was, "What was the mechanism involved in the success of exhaled-air ventilation?" (10). The answers to this question came from a series of experiments where volunteers allowed themselves to be paralyzed while awake, and then ventilated by mouth-to-mouth, mouth-to-mask, or mouth-to-endotracheal tube by Elam and his colleagues until the paralyzing agent was allowed to wear off. Blood gas values were analyzed and the conclusion was that normal physiological parameters could be maintained by exhaled-air ventilation (11). This landmark study brought forth the subsequent challenge to the current back-pressure, arm-lift mode of artificial ventilation. In an effort to answer the question of which form of artificial oxygenation and ventilation would prove superior, a series of controlled experiments was then conducted by Elam and Safar. The various lung volumes with blood gas analysis for the back-pressure, arm-lift was compared with mouth-to-mouth ventilations. These two methods were used on awake, paralyzed volunteers and patients without any mask, endotracheal tubes or adjunctive airway support! These experiments also investigated the mechanisms of soft tissue airway obstruction and the effectiveness of head-tilt and jaw-thrust in maintaining the airway in rescue breathing (the jaw-thrust was first described in Germany by Esmarch and Heiberg in the nineteenth century). The data and conclusions of these studies were published and within one year a dramatic change was made within the American and International Red Cross, global medical associations and the Armed Forces. Modern resuscitation through mouth-to-mouth oxygenation and ventilation was born through these landmark investigations (12–20). "Airway, Breathing" of the "ABCs" for current CPR principles had been founded.

The advent of electrical energy production in the eighteenth century made possible the first recorded successful defibrillation by Squires in 1775; a landmark publication came later in 1809 when Burns hypothesized that effective resuscitation would occur with the combination of artificial ventilation and electric shock (6). Even though a primitive "shock instrument" was fabricated by Aldini (6) in the 1830s, there did not appear to be any significant research into electrical cardiac excitation until much later in the century. The miraculous discovery of anesthesia in the 1840s unfortunately led to catastrophic complications. Documentation of the first case of cardiac arrest was reported in 1848 when a child died under chloroform anesthesia while having a superficial procedure completed (21). As this type of complication became more commonplace, research began to focus upon cardiac physiology and mechanisms to restore the normal heart rhythm and function. Open-chest cardiac compression was first reported by Schiff in 1847 during unsuccessful attempts to circulate blood in dogs and 2 years later Niehans reported an emergency attempt at open cardiac compression in a patient who arrested during an induction of general anesthesia using chloroform. Cardiac contractions reoccurred for a brief time prior to the patient's death (21). Interestingly, in 1847 Boehm reported the first study of closed-chest cardiac compressions in cats (22). The chest was compressed with a rhythmic motion and a cardiac pressure was sustained. In the next 10 years, Koenig and Maass reported eight successful closed-chest cardiac compressions in humans (23) secondary to anesthetic-initiated cardiac standstill; one of these resuscitations lasted for more than 1 h (24). Unfortunately, the open-thorax mode of direct cardiac massage was to be the predominant form of attempted circulatory support for the next 60 years despite these reports.

Alternating current, brought forth by the investigations of Tesla, was first reported by Prevost and Batelli to stop dog heart fibrillation in 1899 (25). Intense research into terminal cardiac dysrhythmias and electrical termination of these lethal rhythms was started in the United States by Kouwenhoven, a professor of electrical engineering, in 1928. The funding for this project was undertaken by the Consolidated Edison Company because of the numerous fatalities induced by electrocution of its employees. Termination of ventricular fibrillation through electrical countershock was confirmed and the effects of both alternating and direct current were investigated in the dog open heart model. By 1933, this group had described the principles necessary for successful open heart alternating current (ac) defibrillation (26). In 1939, the Russians Gurvich and Yuniev were the first to describe successful external defibrillation and reported that direct current

(dc) countershock was superior to ac generated currents. They reported that a capacitor discharge applied to the exterior of the dog's chest would stimulate a cardiac rhythm if only applied no later than 1.5 min after the induction of ventricular fibrillation; however, they noted that the time to successful defibrillation could be extended to as long as 8 min by the application of external chest compressions. There was no description as to how these chest compressions were done (27). Unfortunately, their report was not available to western researchers until 1947 and substantiation of the benefits of dc versus ac would not be made for a number of years.

The research of Kouwenhoven at the Johns Hopkins Hospital continued in defibrillation experiments and in 1958 Knickerbocker, a research fellow, made an astute observation; during a defibrillation experiment he noted a pressure wave form being generated by the application of external electrodes on the dog's thorax (28). During a later, but similar study, Knickerbocker had a dog unexpectedly start to fibrillate and since defibrillation electrodes were not immediately available, he employed the same type of pressure upon the dog's sternum that he had found to generate a systolic pressure. After ~5 min of chest compression, the animal was successfully defibrillated into a normal sinus rhythm. A surgeon, Dr. James Isaacs, who was also conducting experiments in the same laboratory, became aware of this incident and had the foresight to encourage new research by this group into the generation of circulatory blood pressures by external cardiac massage (29). During these subsequent studies, arterial-venous pressure gradients were found to be generated and carotid artery flow was documented. Data that was reproducible indicated that if chest compressions were initiated within 1 min of ventricular fibrillation and continued for as long as 20 min, dogs could be resuscitated by defibrillation and appeared to have no deficits in central nervous system function. Further experimentation on dogs led to the conclusion that the optimum location for chest compressions was on the distal one-third of the sternum with a force of between 35 and 45 newtons (30). Even though postmortem studies revealed numerous injuries, such as rib fractures to these animals, the life-saving benefits were very apparent. Soon the practicality of closed-chest compressions became evident when Kouwenhoven and Isaacs made these laboratory observations available to the surgical staff and, in the same year, a 2-year old child was successfully resuscitated in the operating room at Johns Hopkins Hospital. An organized approach directed at patient resuscitation followed, resulting in 118 cases of successful restoration of life by chest compression following documented ventricular dysrhythmia (31).

Further collaboration at this time by Safar, Elam, and Kouwenhoven resulted in the basic tenets of modern CPR. Since external cardiac chest compressions were found not to produce adequate tidal volumes from airway obstruction (32), control of the airway confirmed by head-tilt data became the "A" in the "ABCs" of CPR. Exhaled air ventilation would become the "B" for rescue breathing. The addition of cardiac compressions, the "C" in the rudiments of basic cardiopulmonary resuscitation was then combined to produce what is now the standard protocol of care in basic life support. The final studies determined what ratio for breathing and chest compressions would be used; one rescuer CPR utilized 2 breaths for every 15 compressions while the addition of a second rescuer could increase the ratio to 1 ventilation per 5 chest compressions (33).

While the first open chest defibrillation in an operating room was reported by Beck at Case Western University in 1947, Zoll reported the first successful closed-chest or external defibrillation in humans (34). This early defibrillator utilized 60 Hz ac current of 1.5 A at a range of 120–150 V. A 6:1 isolation step-up transformer converted the 120-V line current to a range of 0–720 V with the duration of current set at 0.15 s by a condenser-relay circuit. The machine was capable of producing 12,000 W during this time interval. The copper electrodes were 7.5 cm in diameter. This paper described the successful countershock for terminating ventricular fibrillation in four patients. The advent of external cardiac defibrillation would now usher in modern cardiopulmonary resuscitation when conjoined with airway manipulation, rescue breathing, and closed cardiac chest compressions.

The historical evolution for understanding the mechanisms of cardiopulmonary resuscitation has been paradoxical; the physiology of rescue breathing appears to have been well understood versus the mechanisms of cardiac flow due to chest compressions. Positive pressure ventilation, of which mouth-to-mouth resuscitation is an example, utilizes different mechanical principles to expand the lungs versus normal breathing, but the gas exchange once in the alveoli is very similar. The action of chest compressions, however, has remained controversial. After the serendipitous finding of increased blood pressure upon application of defibrillator paddles, Kouwenhoven hypothesized that sternum compression of the heart against the spine forced blood out of the ventricles (28), but no hemodynamic studies supported this claim. Further research demonstrated an increased venous pressure equal to arterial pressure during chest compression that brought into question whether the heart ejected blood in the normal manner (35). A study 1 year later actually measured cardiac output in patients being resuscitated utilizing external cardiac compressions. The ejected blood was found to have flows approximately one-quarter of normal even though systolic blood pressures appeared to be adequate (36). An investigation of actual intravascular pressures during external cardiac compressions determined that left atrial (venous) pressure was very close to arterial pressure, which argued against a projectile expulsion of blood by the heart. The hypothesis of this study was that the requisite flow needed for organ perfusion was driven by the action of the cardiac valves. This action was thought to account for the arterial-venous pressure gradient to sustain oxygen delivery (37). The cardiac compression–cardiac flow hypothesis was further contested with a series of studies generated by the observation that coughing by patients sustained blood pressure. Reports of successful resuscitation in documented ventricular fibrillation by coughing led to research that compared arterial pressures produced by chest compressions to that produced by cough. The conclusion was that improved hemodynamic parameters occurred with coughing CPR (38). Further interest into these mechanisms was

induced by a number of reports whereby trauma patients with a flail chest were not able to be resuscitated through closed-chest compressions; a flail chest results when the thoracic cage is compromised during rib fracture. Direct cardiac compression should be easier to produce since the ribs offer no resistance. Evidence appeared to support increased intrathoracic pressure rather than direct cardiac compression as the mechanism producing blood flow (39,40). Echocardiography was also utilized in several studies where CPR was initiated in humans; the cardiac valves were visualized and noted to be in the open position. Additionally, the left ventricle did not appear to be compressed, again lending credence to the "thoracic pump" theory of blood flow (41,42). Unfortunately, this theory could not account for coronary circulation blood flow or as to the mechanism of blood flow during disruption of intrathoracic pressure, such as when a pneumothorax (collapsed lung) occurs. Subsequent research utilizing very sophisticated instrumentation determined that, indeed, pressure gradients were generated with chest compressions in animals relative to aortic and thoracic venous vessels, data not supported by the thoracic pump theory. Contrast dye echocardiography demonstrated typical opening and closure of the mitral valve with projection of the contrast being propelled throughout the heart and then into the aorta (43,44). The momentum changed with these studies in elucidating the exact mechanism for blood flow, resupporting the cardiac compression hypothesis. What is currently hypothesized today is that both mechanisms seem to operate relative to resuscitation–generated cardiac ejection of blood. The key to understanding this paradox is that chest compressions involve two forces: compression and release of pressure upon the sternum. Compression of the heart forces blood through the atria and ventricles with flow generated, as evidenced by arterial and venous pressure gradients. Release of sternum pressure appears to augment venous return, supporting the thoracic pump theory. Therefore, it appears at this time that the current literature supports both mechanisms in CPR generated blood flow (45).

PULMONARY PHYSIOLOGY

Pulmonary function provides for the oxygenation of tissues and the removal of carbon dioxide from cell metabolism; human's survival is dependent on this function. It is by no coincidence that the first two actions of cardiopulmonary resuscitation, airway establishment and then rescue breathing, must be accomplished prior to chest compressions. Resuscitation is hopeless unless oxygenation and ventilation can be established. It is easiest to appreciate pulmonary function as a progression of air transport from the airway into the lungs, with an overview of lung mechanics and the molecular basis for oxygen and carbon dioxide transport.

After a volume of air is breathed through the oral or nasal passages, this inspired gas passes to the lungs by way of the trachea, bronchi, and bronchioles. Muscular tone in the soft palate and pharynx maintain this anatomical area of the airway. The trachea is supported by numerous cartilaginous rings. At the bronchiole and alveolar level, transpulmonary pressures are responsible for patency. Cardiopulmonary resuscitation of the unconscious victim demands that the first action taken by the rescuer is to make sure that the airway is open. The usual cause is obstruction of the airway by the tongue or soft tissues. Maneuvers to open the airway are the first line treatment in CPR when a person is found to be unresponsive.

The lungs function by expanding through a negative pressure pump mechanism causing inspiration of air by two mechanisms. The diaphragm, a large muscle located at the lung bases, contracts increasing the subatmospheric pressure and thus producing a pressure gradient relative to ambient air. Movement of the rib cage acts in conjunction with the diaphragm, as lung expansion occurs during elevation of the ribs. Normally, the ribs are positioned in a superior–inferior dimension; as the thoracic cage is raised, the ribs move in an anterior–posterior direction, increasing the intrathoracic lung compartment by $\sim 20\%$. The lung expansion through this mechanism also acts to produce a subatmospheric gradient, drawing air into the lungs. This occurs because the lung volumes increase at a more rapid rate than gas flow through the airway. As energy is utilized to cause this expansion, expiration during normal breathing is simply the result of the elastic recoil of the lungs and air is expelled, as now the pressure gradient reverses. During episodes of rapid oxygen metabolism, the work of breathing increases and thus the rapidity of chest wall movement requires a forceful expiration. The abdominal musculature functions in this manner to compress the diaphragm. It should be apparent that pressure–volume relationships establish the adequacy of lung mechanics. Transmural pressures, that is, the difference between the interior of the lung minus the lung exterior (or the pleural space, which separates the lung from the chest wall), define the various lung volumes as well as being a measure of elastic forces on the lung (the force tending to cause lung collapse). The slope of the P–V curve at any point represents the lung compliance; in the normal adult lung this averages 200 mL of air/cm of water, that is, when transpulmonary pressure increases by 1 cm of water, the lungs expand by 200 mL. Lung compliance is not only affected by the elastic force of the lung tissue, but also by the forces generated by surface tension in lung and pleural fluids. This surface tension elastic force is reduced in the lung by surfactant, a complex molecule primarily composed of phospholipids, which has hydrophilic and hydrophobic moieties.

When a rescuer determines that a person is unconscious and begins CPR, the airway is first opened and then rescue breathing is attempted. As mouth-to-mouth ventilations are instituted, now the lungs are expanded by positive pressure, quite different than the previously described normal mechanism. The intraalveolar as well as intrapleural pressure will rise above atmospheric pressure. The diaphragm is progressively pushed toward the abdomen in contradistinction to this muscle's upward or cephalad movement with contraction. Upon expiration, the intrapleural pressure, which is positive, decreases to subatmospheric pressure upon end-expiration and the diaphragm also moves away from the abdomen. When

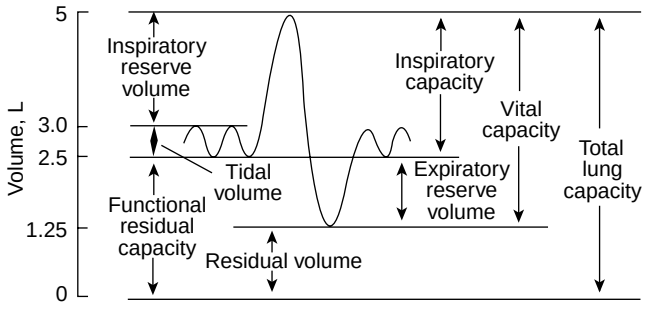


Figure 1. The dynamic lung volumes that can be measured by simple spirometry are the tidal volume, inspiratory reserve volume, expiratory reserve volume, inspiratory capacity, and vital capacity. The static lung volumes are the residual volume, functional residual capacity, and total lung capacity. Reprinted from Anesthesiology, 4th ed., Benumof: Respiratory Physiology and Respiratory Function During Anesthesia, p. 590, 1981, with permission from Elsevier Science.

positive pressure ventilation is employed in other clinical settings, such as with ventilator therapy, a constant concern is that with any damage to the lungs, gases will be propelled into the pleural space. If there is no egress of these gases, a ball-valve mechanism ensues, and the increasing positive pressure in this pleural space will compress the lung, causing hypoxemia and death (pneumothorax).

The lungs are subdivided into four static volumes and four capacities (Fig. 1). A capacity is the combination of two or more lung volumes; capacities are helpful in describing the pulmonary function and disease processes. A device called a spirometer, invented in 1846 by Huntchinson for amusement purposes, is used for these measurements (46). The original machine was a watertight bell emersed in a water tank and connected by tubing to the patient's airway. As this bell moves with inhalation or exhalation, an attached writing instrument marks these volumes on a chart. The current spirometers utilize a bellows or piston with electronic circuitry. All measurements are representative of the average adult man. These volumes and capacities are 20–25% less in women.

1. Tidal Volume: The volume of air either inspired or expired with a normal breath; This is ~500 mL; these are minimum volumes that are typically attempted in rescue breathing.
2. Inspiratory Reserve Volume: This is the maximum volume of air that can be inspired after a normal tidal volume; it is ~3000 mL.
3. Expiratory Reserve Volume: The maximum volume of air that can be ejected after expelling the tidal volume; it is ~1100 mL.
4. Residual Volume: The volume of air remaining in the lungs after a maximal expiration; this volume is ~1200 mL.

The four lung capacities consist of the following:

1. Inspiratory Capacity–Tidal Volume plus Inspiratory Reserve Volume: This volume represents the maximum amount of air that can be inspired after a normal expiration and represents ~3500 mL.
2. Functional Residual Capacity–Expiratory Reserve Volume plus Residual Volume: The volume of air in the lungs after a normal expiration; ~2300 mL.
3. Vital Capacity–Inspiratory Reserve Volume plus Tidal Volume plus Expiratory Reserve Volume: This volume is the maximum amount of air that can be expelled after a maximum inspiration and is ~4600 mL.
4. Total Lung Capacity–Vital Capacity plus Residual Volume: The maximum volume of air that can be expired after greatest possible inspiration.

The minute respiratory volume is equal to the tidal volume as a product of the respiratory rate. Since the normal tidal volume is ~500 mL and the normal respiratory rate is ~12–15 breaths/min, the minute respiratory volume is ~6–7.5 L/min. The inspired and expired volumes are not quite equal since the volume of oxygen absorbed through the alveoli is slightly greater than the volume of carbon dioxide that is expired. Only the inspired air that reaches the alveoli can participate in oxygenating the blood. There is a portion of a normal inspiration that does not reach the alveoli and this volume of gas is referred to as dead space ventilation. Anatomic dead space refers to the volume of gas from the nose, mouth, and trachea to the respiratory bronchioles. This volume averages ~2.2 mL/kg. Thus in a normal tidal volume of 500 mL, only 350 mL of air and thus 72 mL of oxygen, is available for gas exchange.

The tidal volume and the respiratory rate have a profound effect upon the total alveolar ventilation. This fact has been reflected in the revisions of CPR literature over the years. Suppose patients all have the same total minute ventilation of 5000 mL. The first patient has only a small tidal volume of 150 mL and is breathing 33 times/min, producing a minute ventilation of ~5000 mL. Recall that not all of the air in a breath reaches the alveoli; dead space is ~150 mL. The total dead space ventilation would be equivalent to the total minute ventilation. The actual alveolar ventilation would be zero. This patient will become hypoxic very quickly. The second patient has a tidal volume of 250 mL and is breathing at a rate of 20 times/min. The total minute ventilation will be again 5000 mL. The alveolar ventilation will be 2000 mL. The third patient has a tidal volume of 500 mL and a breathing rate of 10 times/min; again the total minute ventilation is 5000 mL, but in this case the actual alveolar ventilation is 3500 mL. The conclusion that should be drawn from these examples is that the efficiency of ventilation is greater when the tidal volume is increased versus the equivalent change in respiratory rate relative to total alveolar ventilation.

The composition of air that one breathes changes significantly from the atmosphere to the alveolus. At sea level, nitrogen produces a partial pressure of ~597 mmHg and composes ~78% of room air. Oxygen has a partial pressure of 159 mmHg and represents almost 21% of the total for atmospheric gas. Carbon dioxide and water make up the remaining partial pressures and percentages. Once the air is humidified by the nasal and oral airways, water vapor

comprises 47 mmHg and increases to ~6% of the mixture with a corresponding reduction for nitrogen and oxygen. The alveolar air has a reduction in both nitrogen (569 mmHg, 75%) and oxygen (104 mmHg and 13%). In the clinical setting, the alveolar oxygen tension is an extremely useful measurement to evaluate the variables in pulmonary mechanics and gas exchange. The ideal alveolar gas equation is useful approximation and is expressed as follows:

$$P\mathrm{A}_{\mathrm{O}_2} = [(P_\mathrm{B} - P_{\mathrm{H}_2\mathrm{O}})(F_\mathrm{I}\mathrm{O}_2)] - \frac{P\mathrm{A}_{\mathrm{CO}_2}}{R} + F$$

Where $P\mathrm{A}_{\mathrm{O}_2}$ is the partial pressure of oxygen in the alveoli; P_B is the barometric pressure; $P_{\mathrm{H}_2\mathrm{O}}$ is the partial pressure of the water vapor in the alveoli at 37 °C; $F_\mathrm{I}\mathrm{O}_2$ is the partial pressure of oxygen; $P\mathrm{ACO}_2$ is the partial pressure of alveolar carbon dioxide; R is the ratio between the volume of carbon dioxide diffusing from the pulmonary blood to the alveoli and the oxygen diffusing from alveoli into pulmonary blood. Approximately 200 mL/min of carbon dioxide versus 250 mL of oxygen exchange, so the ratio 0.8. F is a small correction factor that can be ignored clinically. Therefore, for example, suppose that a patient has been medicated with opioids after a painful operation and the alveolar partial pressure rises to 65 mmHg since these drugs reduce the respiratory sensitivity to carbon dioxide. The barometric pressure is 760 mmHg.

Therefore,

$$P\mathrm{A}_{\mathrm{O}_2} = [(760 - 47)](0.21) - \frac{65}{0.8}$$

$$P\mathrm{A}_{\mathrm{O}_2} = 68 \text{ mmHg}$$

These figures have a profound influence upon oxygenation in resuscitation. A simplified example will enlighten the reader; from the previous review of lung volumes, the total lung capacity is ~5000 mL. If roughly 20% of the atmosphere is oxygen, then 20% of the total lung volume, 1000 cm^3, will contain oxygen. As mentioned earlier, the basal metabolic rate for oxygen consumption is ~250 mL/min. Therefore, the quotient of the 1000 cm^3 relative to the oxygen consumption of 250 mL/min yields 4 min until hypoxia ensues from lack of oxygen. This is reason why time is so critical for the rescuer; unfortunately, the brain is the most oxygen-sensitive organ in the body and cerebral function diminishes rapidly after this critical four minutes. In ACLS, supplemental oxygen is immediately made available to the victim. Given the previous example, if 100% oxygen is administered without entrainment of room air (and nitrogen), now the total lung volume of oxygen would be 5000 cm^3. At the same basal metabolic rate for oxygen utilization, 250 mL/min, theoretically the patient could remain apneic for 20 min before hypoxia would ensue! Practically, this does not occur because of the metabolic byproduct of carbon dioxide diffusing into the alveoli as well as the tremendously increased energy requirements caused by the ventricular dysrhythmias; however, the point to be made here is how the atmospheric composition of gases can easily be altered by the addition of supplemental oxygen to improve the mortality and morbidity of cardiopulmonary resuscitation.

Alveolar ventilation is the ultimate endpoint with respect to lung mechanics. Air must be transmitted throughout the respiratory passages until oxygen can be absorbed by the blood. As a person inspires a normal tidal volume, the contained oxygen reaches the terminal bronchioles. Interestingly there is no organized flow of gas from this point to the alveoli; the oxygen traverses the respiratory bronchiole and alveolar duct into the alveolus for gas exchange by simple diffusion. Once the oxygen reaches the alveolar membrane, the diffusing capacity, which averages 21 mL/min per mmHg, causes the 250 mL of oxygen to traverse the respiratory membrane since the driving oxygen pressure difference is ~12 mmHg. The basic metabolic rate for oxygen utilization is equal to 250 mL/min. Therefore, during quiet respiration, with normal tidal volumes, oxygen intake is appropriate for oxygen utilization. When physical work or exercise increases the metabolic requirements for oxygen, the diffusing capacity can increase threefold in a young healthy adult male. The egress of carbon dioxide through the alveolar membrane is also crucial for survival. The diffusing capacity has never been measured accurately for carbon dioxide due to the rapidity with which this gas passes from red blood cell to alveolus; however, since the diffusion coefficient of carbon dioxide is ~20 times that of oxygen, a range of between 400 and 1200 mL/min per mmHg would be expected for this gas.

OXYGEN AND CARBON DIOXIDE TRANSPORT

Once oxygen diffuses through the alveolar membrane and enters the venous pulmonary blood, it is primarily carried in combination with hemoglobin encased in the red blood cells and secondarily in solution. Hemoglobin is a tetramer molecule consisting of four amino acid polypeptide chains and four heme groups. The globin, or protein portion, consists of two pairs of identical alpha chains and, in the adult hemoglobin, two beta chains. The locus for the alpha chains is located on chromosome 16. The alpha chain is always present; however, there may be some variety in the non-alpha chain. Fetal hemoglobin, for example, has two gamma chains, which increases the hemoglobin binding of oxygen, increasing the efficiency of maternal oxygen transport across the placenta. The four heme moieties are located in the center of each globin molecule. Heme is synthesized from glycine and succinyl coenzyme A to form a tetrapyrrol ring. Subsequent enzymatic reactions produce a protoporphrin and, finally, ferrous iron is inserted into the center of this ring as a function of mitochondrial synthesis. Since there are four heme-combining sites in each hemoglobin molecule, a maximum of four oxygen molecules can attach to the receptors. When all four receptor sites are combined with oxygen, the hemoglobin has a 100% saturation. If only three molecules of oxygen are bound, the hemoglobin is 75%, and so forth. Oxyhemoglobin is hemoglobin that has oxygen bound to the heme sites (HbO_2); unbound hemoglobin is termed "reduced hemoglogin" or "deoxyhemoglobin" (Hb). The key principle to

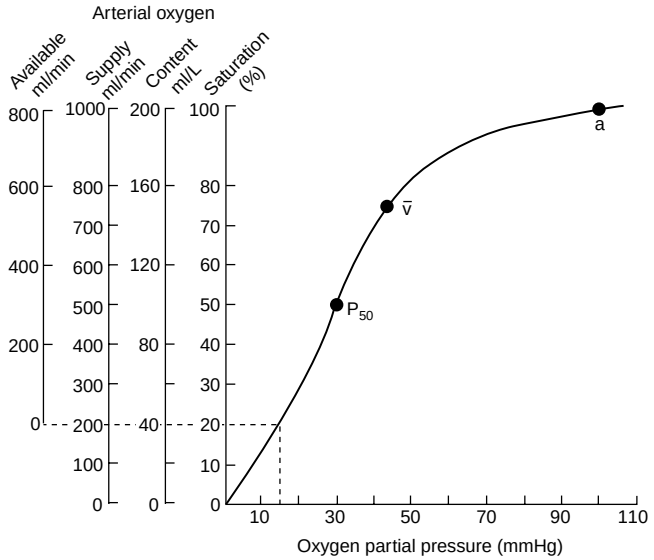


Figure 2. The oxygen–hemoglobin dissociation curve. Four different ordinates are shown as a function of oxygen partial pressure (the abscissa). In order from right to left, they are: saturation (%), O_2 content (mL of O_2/0.1 L) of blood; deoxygen (O_2) supply to the peripheral tissues (mL/min); and O_2 available to the peripheral tissues (mL/min), which is the O_2 supply minus ~200 mL/min that cannot be extracted below a partial pressure of 20 mmHg. Three points are shown on the curve: a, normal arterial: ν, normal mixed venous; and P_{50}, the partial pressure (27 mmHg) at which hemoglobin is 50% saturated. Reprinted from Anesthesiology, 4th ed., Benumof: Respiratory Physiology and Respiratory Function During Anesthesia, p. 596, 1981, with permission from Elsevier Science.

understand is that oxygen binding to hemoglobin is directly related to the partial pressure of oxygen. As the inhaled air reaches the alveoli and participates in gas exchange, hemoglobin becomes fully saturated with oxygen relative to the partial pressure at the alveolar membrane. Oxygen delivery and unbinding occurs at the tissue partial pressure. The initial binding of the first oxygen molecule to hemoglobin facilitates the further binding of the second molecule, and in turn, these first two molecules facilitate further binding of the third oxygen molecule. This interaction occurs until the fourth oxygen molecule is bound, and this characteristic of changing oxygen affinity of hemoglobin is reflected in a sigmoid curve when the percent saturation of hemoglobin is plotted against the partial pressure of oxygen (Fig. 2).

The curve has a steep and flat portion. The steep slope of the curve reflects the rapid combination of oxygen with hemoglobin as the partial pressure increases. Beyond ~60 mmHg, the curve flattens, reflecting very low increases in saturation relative to increases in oxygen partial pressures. The clinical significance of this flat portion of the curve can be observed by noting that a fall from 100 to 60 mmHg only decreases the oxygen saturation from near ~100–90%. This zone of the curve provides for a safe range of minimal saturation and decreases relative to great decreases in partial pressure during oxygen loading. Furthermore, increasing the partial pressure beyond 100 mmHg of O_2 does not really oxygenate the blood to any significant degree; since the hemoglobin is fully saturated, only the dissolved plasma oxygen will increase.

Another significant property of hemoglobin is the fact that the oxygen affinity of this molecule changes with intracellular pH (Bohr effect). As the end product of metabolism, carbon dioxide is present at the tissue level and is converted to a weak acid by the red blood cell catalyst, carbonic anhydrase. This weak acid ionizes to hydrogen ion and lowers the intracellular pH, which decreases the oxygen affinity of hemoglobin, and thus facilitates the unloading of oxygen at the tissue level where it is precisely needed. Since reduced hemoglobin is a weaker acid than hemoglobin, the hydrogen ions are bound and thus deoxyhemoglobin returns to the lungs, where the reverse situation occurs. Carbon dioxide is reconverted in the red blood cell, and with the diffusion of this CO_2 into the alveoli, the pH rises and the affinity of hemoglobin increases for oxygen.

PULMONARY CIRCULATION

Pulmonary blood flow begins with ejection of venous blood from the right ventricle into the pulmonary arteries. Successive arterial branching occurs so that at the level of the alveolar circulation the capillaries lie in intimate contact with the alveoli allowing for a very efficient and exceedingly large surface area for gas exchange. Since the pulmonary arterial pressure is only 20% or so of the systemic circulation, with a mean pressure of ~18 mmHg, these arterioles do not require significant amounts of smooth muscle. Thus the walls of these vessels are extremely thin, allowing for the diffusion of oxygen and carbon dioxide. This characteristic makes these capillaries very susceptible to distortion relative to alveolar pressure. Since the arterial pressure is so low, alveolar pressure may at times exceed pulmonary capillary pressure and this transmural pressure will cause these tiny vessels to collapse. In the upright lung, this situation occurs where pulmonary blood flow pressure is minimal, that is, at the superior aspect of the lungs. This pressure gradient scenario may be observed in Fig. 3.

In *zone 1*, where pulmonary pressure can fall below alveolar pressure, the potential exists for no flow to occur in the capillary. Any situation that decreases systemic blood pressure and thus pulmonary blood flow such as hemorrhage, or increases alveolar transmural pressure, such as might positive pressure ventilation encountered in rescue breathing, might cause this change. The alveolar pressure exceeds pulmonary arterial pressure and, in turn, pulmonary venous pressure.

In *zone 2*, the pulmonary arterial pressure increases due to the elevated hydrostatic pressure as a function of position relative to the column of blood. The alveolar pressure exceeds pulmonary arterial pressure in this zone; however, the pulmonary venous pressure relative to alveolar pressure is low and thus the gradient in this zone is the difference between arterial and alveolar pressure. The analogy to this unique lung region has been described as the vascular waterfall effect (47). The elevation of the river above the dam is described as pulmonary arterial pressure

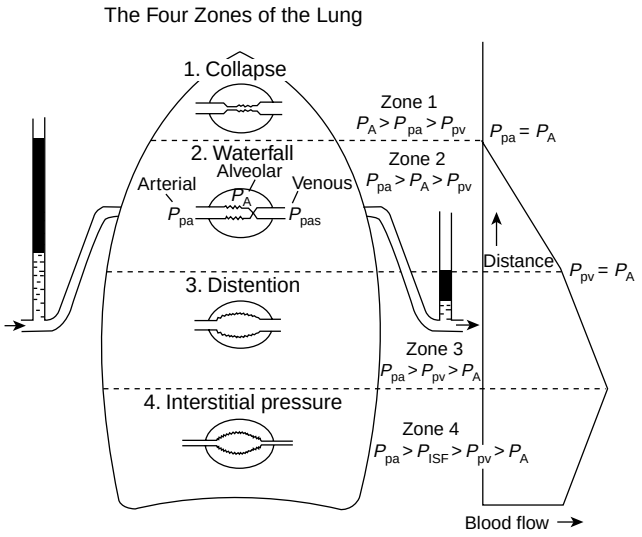


Figure 3. The Four Zones of the Lung. Schematic diagram showing distribution of blood flow in the upright lung. In zone 1, alveolar pressure (P_A) exceeds pulmonary artery pressure (P_{pa}), and no flow occurs because the intraalveolar vessels are collapsed by the compressing alveolar pressure. In zone 2, arterial pressure exceeds alveolar pressure, but alveolar pressure exceeds venous pressure (P_{pv}). Flow in zone 2 is determined by the arterial–alveolar pressure difference (P_{pa}–P_A) and has been likened to an upstream river waterfall over a dam. Since *P*pa increases down zone 2 and P_A remains constant, the perfusion pressure increases, and flow steadily increases down the zone. In zone 3, pulmonary venous pressure exceeds alveolar pressure, and flow is determined by the arterial–venous pressure difference (P_{pa}–P_{pv}), which is constant down this portion of the lung. However, the transmural pressure across the wall of the vessel increases down this zone so that the caliber of the vessels increases (resistance decreases), and therefore flow increases. Finally, in zone 4 pulmonary interstitial pressure becomes positive and exceeds both pulmonary venous pressure and alveolar pressure. Consequently, flow in zone 4 is determined by the arterial interstitial pressure difference (P_{pa}–P_{ISF}). Reprinted from Anesthesiology 4th ed., Benumof: Respiratory Physiology and Respiratory Function During Anesthesia, p. 578, 1981, with permission of Elsevier Science. Diagram modified and reprinted with permission from West JB: Ventilation/Blood Flow and Gas Exchange, 4th ed., Blackwell Scientific Publishers, Oxford, 1970.

and the dam height analogous to alveolar pressure. The downstream river is equivalent to pulmonary venous pressure. Pulmonary blood flow is relative only to the difference between the height of the river upstream and the elevation of the dam. The distance that the water falls over the dam is immaterial to flow rate. Since the alveolar pressure tends to remain constant throughout this zone, but the pulmonary alveolar pressure increases secondary to the gravity, flow increases linearly. *Zone 2* circulation is unique in that ventilation and cardiac changes may alter flow dynamics, shifting these relationships into a momentary *zone 1* or *3* picture.

The dynamics in *zone 3* are straightforward. Here pulmonary venous pressure exceeds alveolar pressure and blood flow is governed by the arterial–venous gradient, which occurs in the systemic circulation. Blood flow never ceases and all capillaries remain patent, with the additional feature of decreasing alveolar pressure maximizing vessel diameters and decreasing pulmonary vascular resistance. The rate of pleural pressure rises as a function of the transmural pressure gradient between lung apex and base; this pressure does not increase as rapidly as the pulmonary artery–venous difference that optimizes blood flow.

Zone 4 is ordinarily not present in normal lung physiology. Some pathological process is required to increase fluid pressure between cells where pulmonary venous and alveolar pressure is exceeded. Conditions such as iatrogenic fluid overload, pulmonary embolism, high levels of negative pleural pressure encountered with airway obstruction in a spontaneously breathing patient, or thoracentesis maneuvers causing profound negative pleural pressures (48,49) may cause this situation. Pulmonary arterial pressures exceed interstitial pressures, which, in turn exceeds venous and alveolar pressures. Since interstitial pressures are greater than venous pressures, regional blood flow is decreased relative to *zone 3*, and flow is governed by the pulmonary arterial-to- interstitial gradient.

In conclusion, it should be evident that both alveolar ventilation and pulmonary blood flow have a variable distribution throughout the lung. The lung base not only receives more blood flow than the apex but, because the compliance of the basal alveoli is greater than the apical alveoli, the lung base receives a greater amount of the tidal volume. Since the blood flow gradient is steeper than the ventilation gradient, the base is relatively overperfused and thus hypoventilated; the reverse situation occurs in the apex where the lung is overventilated and hypoperfused. These conditions have a profound effect upon end-organ oxygen transport. The first scenario refers to physiologic shunt blood flow; should absolutely no ventilation occur, a true shunt occurs. Decreased ventilation relative to perfusion increases alveolar carbon dioxide and thus, as seen in the alveolar gas equation, alveolar oxygen concentration will decrease. The oxygen content of the systemic arterial blood is decreased and thus oxygen transport to the tissue results in hypoxemia. A ventilated alveoli that is not perfused, as in *zone 1*, does not participate in gas exchange. Alveolar carbon dioxide decreases and alveolar oxygen increases due to the absence of blood flow. This situation is termed “alveolar dead space ventilation”. The composition of alveolar gas is essentially equal to atmospheric gas. The extremes of alveolar dead space ventilation and shunt are ends of a continuum in lung ventilation and perfusion dynamics. Ventilation and perfusion ratios will vary throughout the lung both on an anatomical and physiological basis. The total effective gas exchange can thus be seen as the complex interplay between lung mechanics, ventilation, perfusion, and molecular interactions.

CARDIAC PHYSIOLOGY

The heart is an extremely efficient pump, which results in the progressive pulsatile ejection of blood to the organs. The heart is composed of four chambers: two atria and two ventricles. As blood enters the right atrium from the large veins, passive flow continues into the right ventricle. The right atrium then contracts, forcefully ejecting

the remaining 25% of this blood into the right ventricle. After a delay, this right ventricular blood flow is directed into the pulmonary arteries. A progressive reduction in vessel size results in a capillary meshwork intimately in contact with the alveoli whereby the gas exchange mechanisms function. Pulmonary venous blood, now oxygenated and devoid of carbon dioxide, enters the left atrium. This blood is pumped to the left ventricle, and into the systemic circulation where the cycle is continuously repeated.

Cardiac muscle has some similarities to skeletal muscles, but also some very significant differences as well. Cardiac muscle is arranged in a striated latticework with actin and myosin filaments, which lie adjacent to one another and contract in the same manner as skeletal muscle. However, cell membranes separate these fibers yet allow ionic diffusion between these membranes or *intercalated disks*. Thus during a chemical depolarization resulting in an action potential, unimpeded progression of this electrical current flows with minimal resistance throughout the heart. The *intercalated disks* allow for the heart to actually act as two separate systems. The two atria are electrically excited as a unit, as are the ventricles. The anatomical division of atria and ventricles by nonconducting fibrous tissue does not allow conduction to occur between the atrial and muscle in an unorganized manner. A very specialized conduction system ensures that atria and ventricles are depolarized in a progressive manner.

The atrioventricular valves close during ventricular contraction (systole) preventing the backflow of blood into the atria. The tricuspid valve lies between the right atrium and right ventricle; the mitral valve is located between the left atrium and left ventricle. As blood is ejected out of the right and left ventricles, the semilunar valves open; the pulmonary and aortic valves, respectively, then close during cardiac relaxation (diastole) to prevent blood from returning from the pulmonary and systemic circulation. Note that the first arterial branches off the aorta are the coronary arteries.

The specialized conducting system of the heart that produces a progressive, rhythmical contraction of atria and ventricles has several components. The sinus node provides the genesis for cardiac depolarization. This specialized cardiac muscle is located in the right atrium just below and lateral to the superior vena caval ostium. This strip of tissue, ~ 15 mm long, connects directly to the atrial musculature. Generation of action potentials in the sinus node progresses directly to the entire atria causing a unified contraction of all muscle fibers at once. The resting membrane potential of the sinus node fibers is ca. -60 mV compared with the ca. -90 mV for cardiac muscle. This difference in the sinus node electronegativity is due to the fact that sodium ions with their positive charge progressively "leak" intracellularly. A progressive rise in threshold voltage occurs until ca. -40 mV; opening of the rapid sodium channels at this point then produces the initial cardiac depolarization. The sustained contraction of the cardiac muscles is due to the secondary influx of calcium ions, followed by the influx of potassium ions, which exchange with the outward diffusion of the sodium ions.

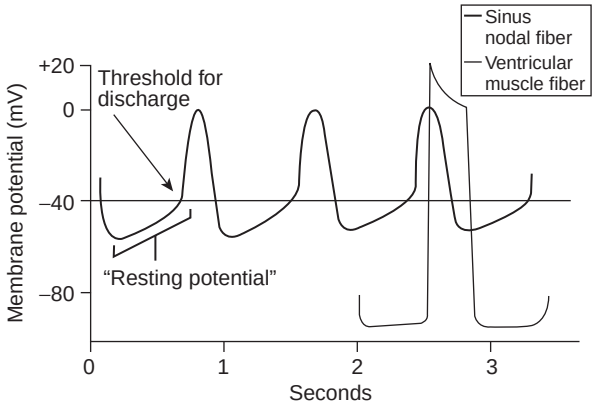


Figure 4. Rhythmical sinus node action potential compared with ventricular muscle fiber. Reprinted from Textbook of Medical Physiology, 10th ed., Guyton and Hall, p. 108, 2000, with permission from Elsevier Science.

This last ion counterexchange of potassium for sodium limits the induced hyperpolarization of the cell allowing repolarization. This phenomenon of "leaky" sodium channels produces the rhythmic excitation, which initiates the cardiac cycle. The rate of this sinus node depolarization is controlled by the autonomic nervous system through the interaction of the para-sympathetic (acetylcholine) and sympathetic (norepinephrine) fibers. Generally, the length of time for this activation is on the order of 10 ms. Drugs utilized in ACLS, such as atropine and epinephrine, affect the firing interval of the sinus node.

Once the atrial muscle fibers are activated, the action potentials cause a generalized contraction of all of these fibers at once, again due to the unique anatomy of the cardiac musculature. Activation of the left atrium occurs through the specialized fibers termed the "anterior interatrial band". The anterior, middle, and posterior internodal pathways transmits the pacemaker impulses to the atrioventricular node in ~ 0.03 s. The AV node is essentially a junction box that has two unique features; a delay in the pacemaker action potential occurs here, which affords a delay in ventricular contraction so that the blood is allowed to empty from the atria to both ventricles and normally action potentials can only travel in one direction. This atrioventricular node is positioned in the right atrial posterior wall just behind the tricuspid valve. The delay in the ventricular depolarizing impulse is ~ 0.13 s.

The final pathway for the activation of the ventricles occurs through the Purkinje fibers, which terminate in the left and right bundle branches. These branches run in the ventricular septum separating the right and left ventricle and then terminate into progressively smaller branches throughout the ventricular muscle. The Purkinje fibers act in contradistinction to the AV node; action potentials are transmitted at a velocity 100-fold allowing rapid excitation and contraction of both ventricles. Transit through the Purkinje fibers is only ~ 0.03 s with the same approximate time necessary for complete ventricular muscle activation.

The electrical activity described in Fig. 5 can be measured at the skin such that electrical potentials are recorded as the ECG. A normal ECG consists of several

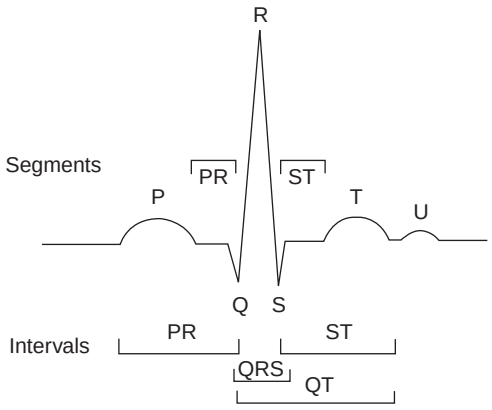


Figure 5. A normal electrocardiogram (ECG) cycle with wave segments and intervals. Reprinted from Anesthesiology, 5th Edition, Hillel and Thys: Electrophysiology, p. 1232, 2000, by permission from Elsevier Science.

waves of depolarization and repolarization. The P wave is produced from the summed action potentials generated during atrial depolarization. It is upright and, after returning to baseline, a pause is observed reflecting the progressive depolarization through the AV node. This P–R interval (actually the P–Q interval, but often the Q-wave is not visualized) begins at the initiation of the P wave and ends at the beginning of the QRS complex. The normal duration of the P–R interval is ~0.12–0.21 s or three to five of the small squares on the ECG graph paper. During the P–R interval atrial depolarization occurs as well as the electrical activity generated in the AV node. The QRS complex, which consists of a Q wave, R wave, and S wave, represents the electrical activity causing ventricular depolarization. The Q wave is seen as a negative deflection from the baseline, which is followed by the large positively deflected R wave. The S wave follows the R wave and, like the Q wave, has a negative deflection. Often the Q wave and S wave may not be observed in the complex. The normal QRS duration is usually no >0.10 s or 2.5 of the small ECG squares. The S–T segment begins at the end of the S wave (commonly termed the "J point") and ends at the onset of the T wave. This interval is usually isoelectric, but can have a normal variance of ca. −0.5 to +2.0 in the precordial leads (see below). The normal duration is <0.12 s and ~2.5 ECG squares. The T wave is a repolarization wave and reflects potentials generated with ventricular recovery. This wave usually has a positive deflection of ~0.5 mV, but often is not observed because of decreased amplitude. The T wave represents a continuum of absolute to a relative refractory period of ventricular depolarization. It is important to note that the electrical activity observed in the electrocardiogram represents electrical activation of the atrial and ventricular muscles and not the actual contractions themselves. A standardized method for recording ECGs consists of graph paper upon that positivity is reflected with upward deflections and negativity, downward deflections. Ten small divisions represent 1 mV. The large vertical lines represent 0.20 s with the smallest intervals representing 0.04 s. These voltages recorded at the skin are very small relative to the actual potentials of ~110 mV at the heart. Proximity of the recording electrodes as well as angular direction from the heart then will affect the size and shape of the ECG, respectively.

The standard electrocardiogram utilizes 12 leads (electrodes) to view the electrical activity of the heart. Six standard limb leads are combined with six chest (precordial) leads. Initial resuscitation events that require dysrhythmia interpretation usually view the bipolar leads I, II, III. These leads utilize two electrodes, one positive and one negative, to monitor the heart and record potential differences. Electrodes are applied through an adherent conductive gel to the left shoulder, right shoulder and left leg essentially forming a triangle (Einthoven's triangle, Fig. 6). The ground lead is usually placed on the right leg. Lines that bisect the sides of the triangle have their origin at the heart, the center of the triangle, which is the zero axis of each side. In lead I, the left shoulder is connected to the positive electrode and the right shoulder electrode is negative; the recording is the potential difference between the left shoulder and right shoulder. In lead II, which is the most common limb lead to be monitored, the negative electrode is placed on the right shoulder and the positive terminal to the left leg. Depolarization in the heart follows this same electrical vector as lead II and thus optimizes P wave height and shape as well as QRS morphology. Lead III places the positive terminal on the left leg and the negative terminal on the left arm. These three leads are very similar to one another in that the P, Q, and T waves are positive. They are excellent for dysrhythmia interpretation; since electrical activity from atria to ventricles is displayed, waveform and time related changes are very apparent, both to the diagnostician and computer (such as in an automatic external defibrillator).

CARDIAC ARREST DYSRHYTHMIAS

The previous section presented information relevant to understanding and interpreting the normal ECG. This

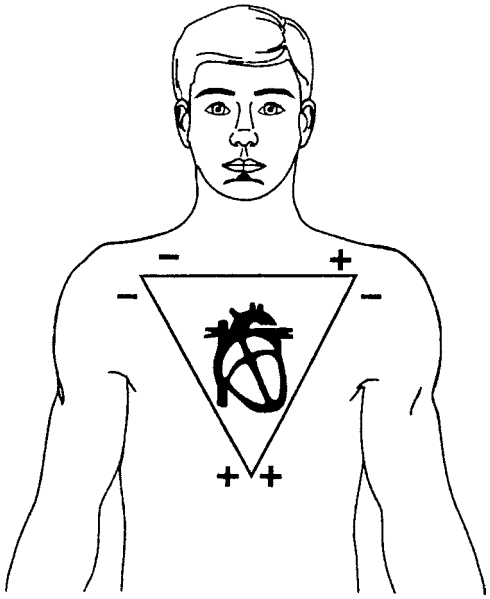


Figure 6. Einthoven's triangle.

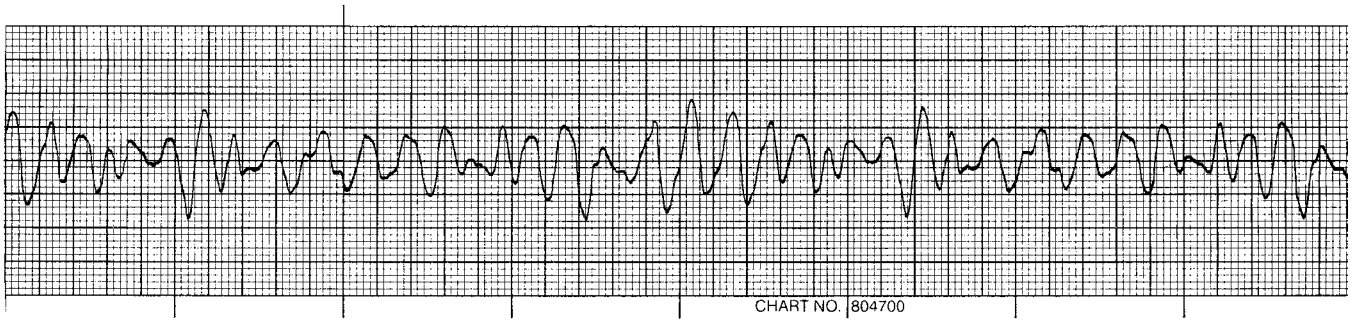


Figure 7. The fatal rhythm of ventricular fibrillation.

primer of basic electrocardiography will enable the reader to now have some ability to differentiate the lethal rhythms that cardiopulmonary resuscitation demands for optimizing treatment.

The fatal dysrhythmia of ventricular fibrillation (Fig. 7) is a common endpoint in cardiac arrest. There are no QRS complexes, P waves or T waves that are identifiable. Scientific evidence over the years has conclusively established that only early defibrillation has any hope of restoring a normal sinus rhythm to the heart and subsequent survival. The fibrillating heart has no ability to eject blood since there is no coordination of heart muscle and no progressive flow of blood from atria to ventricles to systemic circulation. Cardiac depolarization and repolarization within the ventricular muscle occurs in a chaotic manner; ventricular muscle is activated in an unorganized fashion. This electrical activity sustains a vicious cycle of reexcitation, never allowing the return of normal cardiac function. The ventricles neither relax nor contract and in this ventricular fibrillating state consume massive amount of energy. Since there is no ejection of blood, unconsciousness from lack of cerebral blood flow occurs within seconds and death ensues from hypoxia.

After a normal cardiac sequence there is a refractory period (as mentioned previously with the beginning of the T wave, which is at the end of the cardiac cycle), whereby this cardiac impulse cannot reexcite the heart until a new electrical stimulus is generated from the sinoatrial node. However, the underlying etiology of ventricular fibrillation appears to be due to electrical reentry or "circus movements", where the normal termination of depolarization does not occur. These abnormal electrical pathways may be generated in several ways: a shortened refractory period, decreased depolarization velocities, or increased distance for the normal electrical impulse to travel. Recall that the progression of depolarization takes place only in one direction and essentially travels almost in a circle with excitation of the ventricles. If this normal impulse reaches cardiac muscle that has already been depolarized, the refractory time will not allow another depolarization. Stimulation of cardiac muscle will not occur until the entire myocardium is ready to be energized as one unit. Suppose that one of the three abnormal conditions were present; any ventricular muscle that was not refractory could be stimulated to contract in an unorganized manner. In a clinical setting, many individuals have hypertension and develop enlarged hearts. A large ventricular muscle mass would create an increased distance for the normal electrical impulse to follow, thus creating the potential for a "circus movement" to initiate reexcitation of muscle fiber. Rates of depolarization from the sinoatrial node through the AV node may result from blockade of this specialized system from a variety of causes. Electrolyte imbalance, as well as coronary artery disease, are common factors in inducing conduction block. Alterations in the sympathetic nervous system as well as drugs may act in sensitizing the heart, allowing more rapid conduction of impulses and increased susceptibility to fibrillation. Once the ventricular muscle begins this chaotic activity, a chain reaction phenomena begins: conduction velocities throughout the heart decrease, allowing even more time for reentrant depolarizations to occur and the actual muscle refractory time is decreased, increasing the opportunity for these impulses to propagate this dysrhythmia.

The cornerstone of cardiopulmonary resuscitation is early defibrillation. The previous American Heart Association mnemonic of CPR, the "ABCs", which consisted of Airway, Breathing, Circulation, has been changed to "ABCDs" to include defibrillation. After one shock, 60% of all victims succumbing to ventricular fibrillation will survive; after two shocks, 80% survive; after three shocks, 90% will be successfully resuscitated (48). Electrical countershock utilizing high-voltage current can inhibit defibrillation by instantaneously depolarizing all cardiac muscle tissue. The myocardium then is totally refractory to any reentry currents. The electrocardiogram will typically record asystole, or no evidence of electrical activity, from the heart for several seconds. Resumption of the normal cardiac pacemaker will resume, and organized contraction should reoccur. Time is of the essence, since as a heart continues in fibrillation, the rapid utilization of high-energy phosphates depletes this "fuel" for resumption of normal cardiac activity. It is obvious that delay in defibrillation induces a state whereby even successful technique in countershock will be not be able to sustain a normal cardiac rhythm due to the lack of substrate for myocardial energy consumption. The underlying philosophy of cardiopulmonary resuscitation now is early access to defibrillation for the victim.

Pulseless ventricular tachycardia (Fig. 8) is the other malignant dysrhythmia that requires immediate external countershock. Unlike ventricular fibrillation, the electrocardiogram displays a rapid regular rhythm with a widened and abnormal appearing QRS complex. Usually,

Figure 8. Ventricular tachycardia.

the rate is very rapid, with a range of ~ 100–250 beats/min, contrasted to the normal sinus rhythm rate of 60–100 beats/min. Three or more of these bizarre appearing complexes define this dysrhythmia. Usually, the P wave and T wave are obscured and thus the P–R interval cannot be measured. This rapid rate does not allow adequate filling of the ventricles, and causes a ventricular lack synchrony with the atria; the result is dramatic loss of cardiac output and blood pressure. Since there is no effective cardiac output and these dysrhythmias degenerate into ventricular fibrillation, external countershock is mandated. The sinus node functions normally in VT, whereby the atria are properly depolarized. There can be a retrograde depolarization of the atria from the ventricles in some instances of ventricular tachycardia and there will be a definite P wave associated with the abnormal QRS complex. Usually, these retrograde P waves have a negative (downsloping) peak. Another unusual feature of this dysrhythmia is that at certain time intervals the atria may be able to initiate an impulse completely through the AV node and Purkinje system at the instant where the ventricular-initiated depolarizations leave the conduction system vulnerable. The result is termed a "capture beat". If this normally conducted impulse occurs at the same time that a ventricular depolarization is generated, a fusion beat is generated, which appears as a cross between the normal- and ventricular-originated complex. This dysrhythmia may either be paroxysmal or sustained and the shape of these QRS waves either monomorphic or polymorphic. Degeneration into ventricular fibrillation is a common course; for this reason countershock is required. The mechanism for pulseless ventricular tachycardia is hypothesized to be a reenty depolarizing current due to delayed conduction. The site of occurrence for this aberrant mechanism would localize to the Purkinje system and ventricles.

An unusual form of ventricular tachycardia (Fig. 9) is termed "Torsade de Pointes" or twisting of points (49). Notice the QRS morphology viewing the rhythm strip from left to right. On first appearance, this dysrhythmia appears to be ventricular tachycardia. The QRS complexes are wide (versus the normal narrow QRS shape), but constantly changing in shape and amplitude yet there appears to be a rhythmic oscillation about the baseline. The depolarization wave appears to twist around the central axis or helix. This dysrhythmia is triggered by electrical potentials that either occur before or after the normal spontaneous depolarization of the heart or is associated with a prolonged QT interval. Should this dysrhythmia be associated with no evidence of effective blood flow, immediate countershock would be the treatment of choice after the "ABCs" of CPR have been accomplished. However, if Torsade de Pointes is misdiagnosed as ventricular tachycardia and a pulse is present, the potential exists for the wrong treatment and lethal results. This dysrhythmia usually occurs when there is an underlying prolongation of the QT interval. Since many of the antiarrhythmic agents will prolong this interval, it should be apparent that selection of one of these drugs to terminate a misdiagnosed ventricular fibrillation when Torsade is present has the potential to produce a nontreatable dysrhythmia. Antidepressants, antipsychotics, and electrolyte abnormalities (particularly hypokalemia and hypomagnesemia) will produce the underlying QT prolongation that serves as the catalyst for Torsade

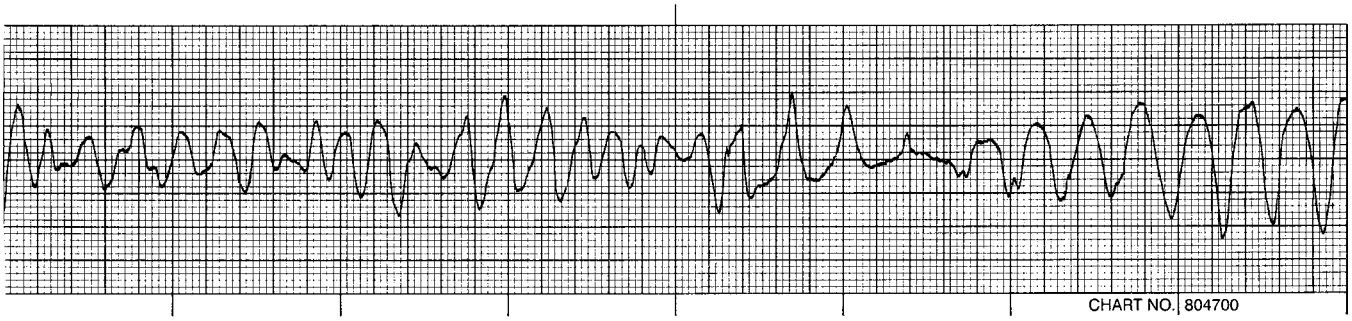


Figure 9. Torsade de Pointes.

Figure 10. Asystole.

to appear (50). Strategies for treating this rhythm are to correct the factors providing the substrate for dysrhythmia; magnesium supplementation, and correction of all electrolyte abnormalities as well as what is termed "overdrive pacing" to externally pace the heart to rates of between 100 and 120 beats/min (51).

Asystole (Fig. 10) is the terminal arrhythmia that represents a dying heart. Observation of the above example shows the absence of electrical activity (flat line), but occasionally P waves, or what is termed "ventricular escape beats", may be present. Usually, the presence of asystole means death for the victim, so rescuers are taught to view this rhythm in more than one lead since fine ventricular fibrillation may actually be the abnormal rhythm rather than asystole. This distinction is extremely important since VF needs to be immediately treated with defibrillation, whereas in asystole the possible causes need to be identified, and defibrillation is contraindicated for resuscitation. Why not shock the heart that is in asystole? The answer lies in the fact that rarely a victim may experience high levels of parasympathetic nerve input to the heart. This parasympathetic stimulation can produce complete cessation of the atrial and ventricular pacemaker. Defibrillation also produces an intense parasympathetic nerve discharge, which, in this specific situation could terminate any chances for the heart to recover normal pacing function (52). Evidence substantiates that defibrillation for asystole does not improve the survival rates in out-of-hospital arrest scenarios (53). There have been rare occasions reported in the literature where victims have positively responded to transcutaneous cardiac pacing for which many defibrillators now have this capability. The heart is stimulated externally to produce an effective cardiac output. The caveat is that this maneuver can only be effective if there are only several minutes of asystole. A specific situation that lends itself to transcutaneous pacing is the asystole that occurs after defibrillation. This witnessed arrhythmia occurs immediately after some countershock attempts and thus there can be a window of < 1 min for pacing to be initiated. There may be certain types of patients who also might respond to asystolic pacing, such as those with Stokes–Adams–Morgagni syndrome, where intermittent atrioventricular block produces asystole or the hypoxia initiated P wave asystole (54).

The last terminal rhythm group that occurs frequently in resuscitation efforts is pulseless electrical activity (PEA). Observe the rhythm strip in Fig. 11; this is an example of a normal sinus rhythm. At first, one would question why a normal ECG would be included in emergency cardiac care and treatment. The key to understanding pulseless electrical activity is that the rhythm may appear to be entirely normal; however, upon further physical assessment of the victim, no pulse is detectable. Historically, cardiologists have referred to this state as electromechanical dissociation, where the normal atrial and ventricular depolarization progresses without any myocardial muscle contraction. Recent scientific evidence utilizing echocardiography and invasive pressure monitoring catheters have found that minute myofibril contractions do occur with depolarization; however, there is not enough effective projectile pressure generated to produce any external signs of perfusion, thus the term pseudoelectromechanical dissociation has been proposed (55,56).

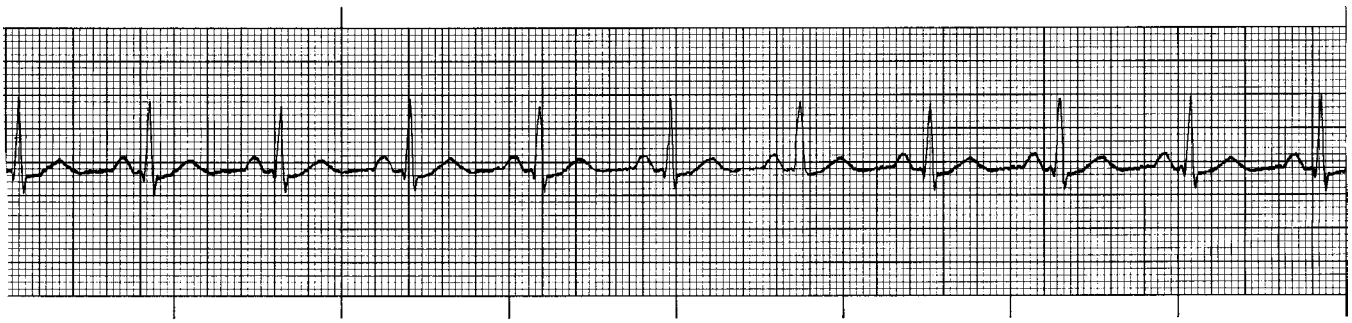


Figure 11. Pulseless electrical activity.

Other rhythms are also included in this group; ventricular escape rhythms, postdefibrillation rhythms, and bradyasystolic rhythms have different QRS morphology than the normal sinus rhythm; however, the same underlying problem of no pulse and no organ perfusion is the defining issue. The key to resuscitation of the victim in PEA is to quickly recognize and treat the underlying problem that has created this scenario. Since there is no perfusing rhythm it should be obvious that the "ABC" basics must be immediately made available to the victim while the differential diagnosis prioritized to the specific situation is reviewed.

There have been a number of mnemonic memory aids suggested to recall the most common conditions provoking pulseless electrical activity. Two common lists are the '5 Hs' and '5Ts' and the "MATCH-HHH-ED" (57,58).

Hypovolemia; Hypoxia; Hydrogen ion excess (acidosis).

Hyper or Hypokalemia (excess or loss of potassium ions).

Hypothermia; "Tablets" (referring to medications or drug overdose).

Tamponade (blood trapped around the heart chamber).

Tension pneumothorax (trapped air causing lung overinflation and resultant loss of cardiac perfusion).

Thrombosis (secondary to myocardial infarction) and Thrombosis (secondary to pulmonary embolism).

M—Myocardial injury or infarction

A—Acidosis

T—Tension pneumothorax

C—Cardiac tamponade

H—Hypothermia

H—Hypo or Hyperkalemia

H—Hypoxia

H—Hypovolemia

E—Embolism

D—Drug overdose or chemical toxins.

The most common cause of PEA is hypovolemia: inadequate intravascular fluid volume. These fluid volume losses may be real, such as in acute hemorrhage or relative, wherein the capacitance vessels (veins) vasodilate, leaving inadequate vascular blood and fluids to compensate. A classic evolution to pulseless electrical activity would ensue from hemorrhage, producing normal electrical complexes, with progression to a tachycardia where the heart rate increases to compensate for pressure loss. Peak pressure for each cardiac ejection of blood (systolic pressure) would decrease, but increased constriction of the circulatory vessels would increase (diastolic pressure). As the blood pressure continues to drop to extremely low levels secondary to the hemorrhage, the sinus tachycardia continues. At this point, the astute clinician would consider this cause for PEA and immediately give the victim intravascular fluid followed by blood products to stabilize the victim. It is obvious that the definitive treatment in this case would be surgery; the goal of the resuscitative response is to maintain perfusion and oxygen delivery to the tissues. Time again is critical since without effective cardiac flow, perfusion of the myocardium is nonexistent and hypoxia develops quickly. Progression toward ventricular fibrillation, asystole, and death are the end result. An example of relative hypovolemia is septic shock. In this situation, a bacterial infection produces endotoxins, which causes the resistance vessels to dilate. There is normal circulating blood volume, but increased vessel capacitance resulting in a progressive blood pressure drop. Due to the underlying pathology, the vessels are unable to adequately constrict and the evolution to cardiac dysfunction secondary to poor or absent coronary perfusion occurs, eventually producing the terminal rhythms of ventricular fibrillation and asystole. Treatment in this scenario consists of increasing the fluid volume as well as to utilize drugs that act to constrict the blood vessels.

Each one of the other conditions mentioned above that can cause pulseless electrical activity also has an immediate treatment response, after the appropriate steps in basic life support have been met. Thus while the victim is being oxygenated and ventilated, and perfusion accomplished through chest compressions, the clinician is rapidly focusing upon the conditions that would be most causative and then initiating treatment. In hypoxia, treatment consists of supplemental oxygen and adequate mechanical ventilation of the lungs. In cardiac tamponade or tension pneumothorax, needle decompression of the pericardial sac or lung is the required emergency management. Electrolyte imbalances require immediate infusion or neutralization of the ion. A preexisting acidosis might require exogenous sodium bicarbonate. An acute myocardial infarction (heart attack) causing PEA would need supplemental oxygen, pain management, fluid resuscitation and immediate infusion of coronary artery thrombolytics. Pulseless electrical activity is always associated with an underlying clinical condition that, when identified and treated early, has the potential to be reversed and thus effective perfusion restored to the victim.

MECHANISMS FOR COMMUNITY CPR

The American Heart Association is a volunteer organization that has as its goal the reduction of mortality and morbidity secondary to cardiovascular and cerebrovascular disease. The AHA established a scientific mission in 1963 to evaluate and promulgate standards of CPR. In 1971, the Emergency Cardiac Care Committee was founded. Since the first conference on CPR in 1966 sponsored by the National Academy of Sciences and the National Research Council, guidelines for CPR have been based upon the available scientific evidence. Since this time there have been six conferences, the last in 2000 consisting of resuscitation experts worldwide. All aspects of CPR have been evaluated, including the actual access for emergency care in the community, education of the layperson, and delivery of basic CPR and ACLS. "Emergency Cardiovascular Care" is the term used to describe the organized response to life-threatening emergencies for the adult, pediatric, and neonatal victim. Other ECC provisions for care include educating the layperson to recognize the signs and symptoms of myocardial ischemia and infarction, stroke, activation of the EMS system (911) as well as early

implementation of basic life support, defibrillation, and advanced cardiac life support with immediate victim transfer to the hospital. The goal of ECC is to save lives and this also includes educating the public regarding cardiovascular risk and how to maintain healthy lifestyles. The cornerstone of all emergency care is the layperson. Effective emergency care can only occur with public awareness of these events as well as prompt administration of CPR and use of the automatic external defibrillator. Once the EMS system has become involved with the victim, progression of care is dictated by a physician and ACLS protocol. The optimization of oxygenation and ventilation through supplemental oxygen and adjunctive respiratory devices, electrocardiographic monitoring with rhythm assessment and treatment, establishment of intravenous access for appropriate medications and post-resuscitation management are all elements of ACLS.

The American Heart Association "Chain of Survival" is a crucial concept to be understood in the context of CPR and is taught to the layperson in the basic lifesaving courses. There are four critical links in this chain that require specific actions on the part of the public. In the event of an emergency the first link involves activating the emergency medical services system. Obviously, the layperson must be able to recognize that a true emergency exists, and this underscores the efforts by the community organizations to educate the public regarding signs and symptoms of heart attack, stroke, and loss of the airway. The key is unresponsiveness; anyone found unconscious should initiate the first link by way of immediately involving EMS through calling 911. Once an emergency medical dispatcher is contacted, immediate aid in the form of paramedicals is sent to the scene. The dispatcher is also taught how to aid the layperson in how to provide basic CPR, which is the second link in the chain. Only until the EMS personnel arrive and begin managing care for the victim does the dispatcher terminate the call. In some communities, "enhanced 911" is available in which a computer will provide the dispatcher with the address from which the call is made in case communication is difficult or there is a premature telephone disconnection. This second link of CPR is the most critical phase in the Chain of Survival since the rescuer provides oxygen to the victim and, if needed, circulates this oxygen by chest compressions to the brain, heart, and vital organs. If bystander CPR is initiated within 4 min after a victim collapses, the odds of survival, after discharge from the hospital, are doubled (59).

The third link in the chain is early defibrillation, either by the public use of the automated external defibrillator or by the paramedic on the scene. Early access to defibrillation for the victim of cardiac arrest will significantly improve the chance for survival: each minute of delay for defibrillation for the victim in ventricular fibrillation decreases the chance of survival by 7–10%, and if defibrillation is provided within the first 5 min, chance of survival is 50% (60). Unfortunately, if >12 min of delay from collapse to initial resuscitation is encountered, the survival rate only ranges between 2 and 5%, and intact neurological function is compromised (61).

Recently, gambling casinos have implemented access for defibrillators and for victims who received a shock within 3 min had a 74% survival rate since a low response time of between 2 and 3 min was documented (62). Since time to defibrillation is so critical, automatic external defibrillators have been made much more accessible to the public. These devices can now be found in large gathering places such as stadiums, golf courses, airports and airplanes, shopping malls, large grocery stores, and other facilities where people in great numbers tend to congregate. So important is the early access to defibrillation that a great majority of states have enacted legislation to encourage use of these devices. The Cardiac Arrest Survival Act provides legal immunity for the layperson and the public business or corporate entity that uses or provides an automatic external defibrillator for resuscitation, which is essentially an expansion of the "Good Samaritan" type legislation. This immunity should encourage active participation by the public for involvement in victim resuscitation. Public access defibrillation has been described as the second most significant advance, compared to CPR, in the pre-hospital rescue scenario.

The final fourth link in the Chain of Survival is early ACLS by highly trained paramedical personnel. Emergency medical technicians expand (EMTs) incorporate basic CPR with interpreting cardiac dysrhythymias and, if required, defibrillation. Emergency medical technicians expand the immediate life-saving care by providing supplemental oxygen, intubation and control of the airway, gaining intravenous access and administering pharmacologic medications while in contact with a physican. This process occurs at the scene, and once the victim is stabilized, advanced cardiac life support continues through transport to the hospital emergency room. The most significant impact of early ACLS is to prevent the catastrophic progression of lack of oxygenation and cardiac arrest rather than to treat the terminal conditions inherent in this process.

The rescuer, whether a lay person of EMT, who begins CPR in the "field" must continue BLS (63) until one of the following events occurs:

1. The victim begins to show signs of spontaneous ventilation and perfusion.
2. Care is transferred to another qualified BLS responder, EMT, or ALS medical providers; or to a physician who makes the determination that resuscitation should be terminated.
3. The rescuer cannot continue resuscitation due to exhaustion or to hazards that may jeopardize the rescuer's life or the lives of others in the team.
4. An authentic no-CPR order is presented to the responders.

The determination to discontinue resuscitation depends on a stepwise evaluation of the efforts made during BLS and ACLS. A review of the process should ensure that each step in the resuscitation has been carried out in a flawless manner. Successful ventilation and intubation, intravenous access and the administration of appropriate medications as well as countershock should be achieved according to ACLS protocol. Electrocardiography evaluation should render a conclusion of no reversibility for the underlying agonal rhythm. Recently, the determination of end-tidal

carbon dioxide has been advocated as a potential predictor of death (64). During resuscitation end-tidal carbon dioxide reflects the adequacy of cardiac output generated during chest compressions. This study suggested that after standard ACLS protocols had been followed for 20 min, a persistent end-tidal carbon dioxide level of 10 mmHg or less predicted nonsurvival in the victim with electrical activity, but without a pulse.

CARDIOVASCULAR DISEASE

Every year ~500,000 people are hospitalized for treatment of chest pain secondary to cardiac origin and 1.5 million victims will experienced a heart attack (65,66).

Some 500,000 people a year who have a myocardial infarction (heart attack) will die from this insult and ~225,000 of these deaths will occur within the first hour after symptoms and prior to reaching a hospital (67,68). In 17% of the victims, chest pain is the first and only symptom (67). It is again significant that time to intervention for the patient experiencing a myocardial infarction is crucial to survival; treatment must be undertaken within the first several hours after the symptoms occur (70,71). Early treatment underscores the necessity for rapid recognition of a cardiac event, followed by rapid CPR, defibrillation, ACLS and transport to the hospital.

The most common cause of a heart attack is ischemic atherosclerotic disease. The essential pathophysiology is the narrowing of the coronary artery lumens by deposits of fat-substrate such as cholesterol and lipids, which eventually retain calcium. The actual process of the accumulation of these plaques occurs very slowly, but has been demonstrated to begin at an early age. This same process affects the cerebral arteries as well and is the etiology for an ischemic stroke. As the coronary artery lumen narrows, a dynamic situation develops where blood flow and thus oxygen supply will not meet with increased demand for oxygen by the cardiac muscle fibers. Typically the coronary artery will have a circumference reduction of 70% for symptoms to occur. This condition of ischemia will produce a characteristic constellation of transient symptoms, referred to as angina pectoris. Chest pain is the most common sign of an acute cardiac ischemic event, occurring in 70–80% of the population (72). This pain appears to have several different components: transmission of dull, poorly localized pain occurs through the sympathetic visceral nerve fibers; a somatic pain generator produces the sharp and dermatomal aspects; and psychological input gives rise to the sense of impending doom (73,74). This cerebral input to the event may significantly exacerbate the ischemia since activation of sympathetic nervous system will increase heart rate and contractile force, further tipping the scale toward more energy consumption and thus oxygen demand.

Paradoxically, the majority of episodes of an acute coronary event (angina and or infarction) occur during periods of rest or mild to moderate exercise; profound physical exercise is associated with the minority of events (75). The victim will experience an intense, dull, crushing pressure sensation in the chest, most commonly behind the breastbone (retrosternally) and/or pain in the back, arms, shoulders, or mandible. Often nausea, vomiting, sweating, and shortness of breath (dyspnea) may accompany the pain. There appears to be a circadian rhythm regarding the occurrence of angina and the progression to infarction. Two daily peaks in incidence have been noted with the first pattern beginning from awakening to about noon, and the second peak occurring in the early evening (76,77). There are atypical presentations to unstable angina or myocardial infarction that will delay access for the victim. This subset of the population may have only vague, mild discomfort, which can be confused with a myriad of medical complaints. Diabetics, women, and the elderly all have a higher incidence of nonclassical presentations for cardiovascular ischemia (78,79). Diabetics are prone to neurological dysfunction and thus may have no sensation of the pain associated with angina. A retrospective review found that 30% of first heart attacks in men and 50% of first infarctions in women did not present with classical signs and symptoms and were clinically not recognized (80). When oxygen demand decreases, such as when the physical activity is discontinued, the decreased oxygen supply secondary to the narrowed lumen will be adequate and the symptoms will usually resolve. Progression of the disease, however, results in a much more severe mechanism for ischemia. The plaque is predisposed to rupture and when this occurs, activation of the coagulation system releases mediators that form a clot or thrombus over the plaque, which can further limit blood flow, or catastrophically, stop all blood flow completely. If the partial occlusion from the thrombus is severe enough, what is termed “unstable angina” develops. Though there have been many definitions of unstable angina, the main characteristic is that this type of angina occurs at rest and is progressive or prolonged in nature. Nocturnal angina, again with the victim at rest, would be classified as unstable angina. The heart is consuming the least amount of oxygen yet there is a lack of supply due to lumen reduction from the plaque. Clot enlargement provides the mechanism for dislodgement of particles or emboli, which then travel downstream, lodging in the microvasculature and these individuals are at a very high risk for progression to irreversible cardiac damage.

Complete occlusion of the coronary artery results in a myocardial infarction. Deprivation of oxygen results in the death of cardiac myofibrils and induces irritability in the cardiac conduction system, setting the stage for the initiation of lethal dysrhythmias such as ventricular fibrillation. Where and how severe the damage is to the heart depends on what coronary artery has the occlusion. If the left main coronary artery has an acute total obstruction, mortality is very high since no blood flow will occur through the two distal branches, the left anterior descending and circumflex arteries. Blood flow will be blocked to the entire left ventricle, the septum between the left and right ventricle, and the bundle branches. Even if the patient receives timely CPR, and clot lysis, a significant amount of heart may be destroyed resulting in scar tissue and a drastic reduction in blood flow, resulting in what is termed “congestive heart failure”. If the thrombus were to lodge and occlude the right coronary artery, hypoxia would occur in

the AV node, right ventricle, and in the majority of individuals, the posterior and inferior aspect of the left ventricle.

The continuum of acute cardiac injury, from unstable angina to myocardial infarction typically presents with characteristic electrocardiographic signs. Recall what the normal elements are to the ECG. During episodes of unstable angina where cardiac muscle demand for oxygen exceeds supply, ischemic changes to the ECG are seen as S–T segment depression, defined as $\geq$ to a 1 mm change from baseline on the standard graphpaper or changes in the T waves. When this S–T segment depression is downsloping, this is a sensitive sign of ischemia. The T waves may appear inverted or enlarged and symmetrical. When actual occlusion of the artery occurs, myocardial injury to tissues ensues and both muscle contraction and conduction are decreased from normal. The ECG change characteristic of injury is S–T segment elevation in contradistinction to changes of ischemia. When there is $\geq$1 mm above baseline for the S–T segment elevation, significant cardiac injury has occurred. This group of patients who exhibit S–T segment elevation on the ECG in two contiguous leads may be salvaged through reperfusion therapy. Restoration of blood flow and the course of injury is very dependent on early administration of thrombolytics or percutaneous transluminal coronary angioplasty (PTCA). The greatest improvement in mortality and morbidity occurs when reperfusion therapy is administered within the first 3 h after onset of symptoms (81,82). Conjoined with early CPR and defibrillaton, reperfusion therapy stands as one of the greatest advances in acute coronary syndromes. Current regimens include the fibrinolytics streptokinase and alteplase, as well as numerous other similar agents that act by inducing fibrinolysis through interactions with tissue plasminogen activator. The PTCA is a mechanical procedure whereby a catheter is guided through the coronary arteries into the area of stenosis, and then a balloon is inflated to expand the vessel. This procedure is restrictive in that only specialized centers have the capability to utilize this regimen, although it may be superior to thrombolytics.

Myocardial infarction defines the actual death of cardiac tissue. This is the end result of the process of ischemia with myocardial cell injury. The infarcted tissue area, again representative for the specific coronary artery occluded, will exhibit characteristics associated with loss of cellular life. Intracellular contents are released after loss of cell wall integrity. Some of these enzymes, such as creatine phosphokinase and the troponins, can be measured in the bloodstream to confirm infarction. The classic ECG changes for myocardial infarction are the presence of abnormal Q is waves. When a Q wave is or $\geq$1 mm in width and the height is >25% of the R wave height, the diagnosis can be made. An abnormal Q wave reveals the existence of dead cardiac tissue, but does not reveal anything about when the infarction happened. The assumption can be made for a recent infarction if the Q wave is associated with S–T segment changes and/or T wave changes. A non-Q wave infarction can also occur: There is myocardial cell wall dissolution with the release of cardiac enzymes, but only accompanied by S–T segment changes or T wave abnormalities. There is a lower mortality rate for non-Q wave heart attacks, but, unfortunately, an increased incidence of future reinfarction or death (83,84). Fibrinolytics are contraindicated in patients with non-Q wave infarction since the clot occlusion may be paradoxically aggravated by release of thrombin, which further activates platelets (85).

Prehospital intervention for the victim suffering an acute cardiac event is based upon the "Chain of Survival". Once the signs and symptoms of a heart attack are recognized, early access to the emergency medical system is imperative. A common problem encountered is denial either from the victim or the rescuer, which impedes response time. Once the EMS personnel arrive at the scene a pertinent medical history is obtained and physical examination completed. A complete 12 lead ECG is obtained and then transmitted to the physician who is dictating care. Oxygen is the first line treatment for anyone complaining of chest pain. It should be recalled that supplemental oxygen substantially increases the oxygen tension in the blood and significantly improves tissue oxygenation. A critical blood flow restriction may be palliated by improving oxygen supply in this manner. The administration of the drug nitroglycerin is quickly administered for the victim symptomatic for angina in conjunction with oxygen. Nitroglycerin is delivered sublingually for rapid absorption into the bloodstream. Nitroglycerin is effective in relieving the symptoms of angina in several ways: relaxation of venous smooth muscle occurs due to binding of specific vascular receptors. As relaxation of the venous capacitance vessels occurs, venous return to the heart is decreased, thereby relieving ventricular wall tension, which ultimately decreases ventricular work and oxygen consumption. The nitrates also dilate the large coronary arteries as well as increasing blood flow through collateral vessels, which improves ischemic blood flow (86,87). Aspirin is the third drug that should be administered immediately by either the BLS provider or EMT when symptoms suggest a cardiac event (and the victim is not allergic to aspirin). A regular tablet of aspirin (325 mg), when ingested, will cause an immediate anticlotting mechanism by way of platelet inhibition. There is evidence that suggests aspirin decreases coronary artery reocclusion and future coronary symptoms, with reduction of death and furthermore, the effects of aspirin appear to be additive to fibrinolyis (88). The fourth drug that is administered during episodes of chest pain secondary to unstable angina or myocardial infarction is morphine. Although morphine is a narcotic analgesic, it produces beneficial hemodynamic effects in addition to profound pain relief. Morphine causes decreased vascular tone in the venous capacitance vessels, thus reducing myocardial wall tension, much like nitroglycerin. The mechanism, however, is different in that the action appears to be mediated through central nervous system reductions in sympathetic tone (89). A convenient mnemonic has been utilized, "MONA", for recall of these four immediate effective therapies for pre-hospital treatment of the acute coronary syndrome.

CEREBROVASCULAR DISEASE

There has been a concerted effort in the Emergency Cardiovascular Care system to improve pre-hospital recognition of

the warning signs of stroke and provide rapid access to the Emergency Medical System. Public awareness of the issues regarding a "brain attack" has lagged relative to the exposure and education afforded cardiovascular disease. Stroke ranks third behind heart disease and cancer for morbity in the United States; 500,000 Americans a year suffer from a cerebrovascular accident and 125,000 of these victims will die (90). Until recently, stroke victims were only offered supportive and rehabilitative therapy for the complications experienced if they survived the initial insult. However, advances in fibrinolytic therapy, as in treatment of cardiovascular disease, dramatically improves outcome for the patient who has experienced an ischemic stroke (91). Fibrinolytic treatment reduces stroke disability and significantly improves quality of life after hospital discharge (92,93). The caveat is that the cerebrovascular accident must be recognized and treatment initiated in a timely manner; fibrinolytics need to be provided within 3 h after the onset of an ischemic stroke (94). Thus there is a narrow window of opportunity to limit cerebral damage, which underscores how important the role is for the public in providing immediate access to the EMS system for the victim.

The underlying cause for an ischemic stroke is comparable to the etiology for myocardial infarction. There is a disruption to cerebral blood flow due to the presence of an occlusive clot. The oxygen supply to the particular area of the brain supplied by the blocked artery does not meet the tissue demand and the same process of ischemia, injury, and cell death will occur. The thrombus, which occludes the vessel, is the end result of atherosclerotic changes to the artery. However, due to the unique anatomical positions of the cerebral arterial system, a blood clot formed elsewhere in the body can embolize to disrupt blood flow to the brain. Approximately 75–85% of all strokes are of this type, and defined as ischemic, and furthermore can be classified as to the arterial system that is affected. The two major arterial conduits to the brain are the carotid arteries and vertebrobasilar arteries, which affect the cerebral hemispheres or brain stem–cerebellum, respectively. Typically, a person who is at risk will develop what is termed "a transient ischemic attack" (TIA) prior to a full-blown stroke. Essentially, a TIA is a reversible mini-stroke that may affect specific brain function or eyesight and will last anywhere from minutes to hours (95). The TIA is a harbinger of a future "brain attack" much like unstable angina will forecast a heart attack. About 5% of those persons presenting with a TIA will end up with a stroke in 1 month; the risk will increase to $\sim$12% after 1 year and an extra 5%/year thereafter (96). Fortunately, the symptoms from a TIA will bring the patient into the medical system for evaluation whereby treatment regimens clearly reduce risk for ischemic stroke. The surgical procedure of carotid endarterectomy in which the carotid artery plaque is removed has been proven very beneficial for patients that have had a recent TIA and a $>$70% stenosis of the carotid artery (97). In those individuals who are not operable candidates, aspirin and the specific platelet inhibitor types of drugs have been shown to be successful in preventing subsequent stroke in patients presenting with TIA (98).

The minority of acute strokes are due to hemorrhage of cerebral artery. The bleeding may occur in the subarachnoid space, which is in the superficial exterior aspect of the brain, or in the brain tissue itself, defined as an intracerebral hemorrhage. The common etiology to a subarachnoid hemorrhage is an aneurysm where the arterial wall weakens, and eventually a disruption occurs (99). In the case of a hemorrhage into the brain tissue itself, high blood pressure appears to be the major causative factor (100). While there are similar signs and symptoms in both types of stroke, there are also distinct differences in findings, which aids in the diagnosis. In general, the presentation for a subarachnoid hemorrhagic stroke is more severe with a very common complaint of an extremely painful headache, which tends to be global, and may have radiation of pain into the face or neck. This headache is often accompanied by mental status changes, nausea, vomiting, photophobia, or cardiac dysrhythmias. In a minority of patients, a prodromal episode of these symptoms may be caused by leakage of the aneurysm offering a warning sign (101). While the victim suffering from an intracerebral hemorrhage may also present with a severe headache, these patients tend to have a greater neurological insult with significantly depressed mental status function. The signs and symptoms of an ischemic versus a hemorrhagic stroke overlap and diagnosis may be difficult based upon the medical history and physical findings. Since radiological imaging offers the greatest aid in differentiating these two types of cerebrovascular accidents, time is very critical to clinch the diagnosis to offer the appropriate treatment. Fibrinolytics would obviously be a catastrophic therapy in the mistaken treatment of what appears to be an ischemic stroke when the etiology is a ruptured blood vessel that requires surgery.

The American Heart Association "Chain of Survival" that has been implemented and associated with cerebrovascular disease has been applied to the pre-hospital care of the stroke victim. Early recognition of a stroke and activation of the EMS system are paramount in initial therapy, which is often problematic, since, unlike a heart attack, stroke may be difficult to detect. While early defibrillation is not ordinarily indicated for the stroke victim, the possibility always exists that coincidental lethal dysrhythmias may be present during the initial presentation (102). The last link in the chain is early hospital care. The common theme regarding out-of-hospital management for cardiac or stroke victims is rapid entry for the victim into advanced life saving. The '7-D' mnemonic has been recommended as an aid for care in the stroke patient: Detection; Dispatch; Delivery: Door; Data; Decision; Drug (103). Early detection, with an accurate recall of the initial signs of a stroke are critical to care and must be accomplished by the immediate family member or layperson, with immediate access to "Dispatch", the EMS personnel. An important point to emphasize is that the majority of strokes occur at home (104). Paramedics who arrive at the scene must confirm a rapid, tentative diagnosis through focused medical history and physical examination, and then "Deliver" the patient rapidly to the hospital. Once the patient is through the Emergency Department "Door" the medical history and physical examination are further refined along with radiography (computerized tomography). A "Decision" is made regarding whether fibrinolytic therapy is

indicated for an ischemic stroke and the "Drug" treatment is initiated. The drug therapy must be initiated within 3 h after the onset of an ischemic stroke.

The changes in mental status and/or sensorimotor function in a cerebrovascular accident may range from minor, almost unrecognizable changes, to loss of consciousness and seizures. A person may exhibit grades of confusion, with a progression to stupor or coma where the airway is obtunded and basic life support is required. Comprehension of language often occurs with inappropriate responses to simple questions. Physical manifestations are often present unilaterally. Paralysis in either the face, upper, or lower extremity may range from slight weakness to frank inability to exercise any muscular control. Since the face is always exposed, muscle weakness is exhibited by loss of tone and sagging of the muscles in facial expression. Difficulty writing (aphasia) or speaking (dysarthria) occurs due to loss of appropriate motor input from the brain. Loss of sensation is another common sign of a stroke (or a TIA). Visual disturbances, including blindness are much more obvious and usually involve only one eye. If the location of the ischemia is in the vertebrobasilar arterial system, centers of the brain controlling coordination are involved and signs such as gait disturbances (gait ataxia) are common. The dilemma of pre-hospital rapid neurological assessment to evaluate the possible stroke victim when the presentation is varied has been improved by several instruments. The Cincinnati Pre-hospital Stroke Scale (104) is very effective in identifying the stroke victim. Three physical findings are assessed: facial droop; arm drift; and speech. Abnormal features in any one category is very predictive for cerebrovascular accident. The Los Angeles Pre-hospital Stroke Screen also is extremely useful for assessment. Six criteria are first evaluated in the medical history: (1) age > 45 years, (2) absent history of seizures or epilepsy, (3) no history of motor loss, (5) serum glucose not < 60 g/dL nor > 400 g/dL, and (6) asymmetry in any one of the three categories of facial musculature, grip strength, and arm strength. If all criteria are positive, there is a 97% chance of an acute stroke (105). These tests have streamlined the response time and have allowed the hospital emergency room to prepare for rapid definitive diagnosis.

THE "ABCs" OF ADULT CPR

The evolution of the current basics of life-support has resulted in a streamlined set of actions that has standardized the initial care for the victim, whether accomplished by a public bystander or emergency room physician. Assessment of the victim always precedes a physical maneuver on the part of the rescuer; constant appraisal of the effectiveness of CPR and the response of the victim is a core principle in American Heart Association Basic Life Support tactics. The initial steps of resuscitation are never bypassed; for example, if the airway is not established, the single rescuer would never start chest compressions, or begin intravenous access. The stepwise process in the algorithm ensures that an orderly process occurs in a situation where chaos and a high degree of emotional turmoil exist for the rescuer. Since there are some basic differences in how resuscitation is administered to the adult versus a child, anyone 8 years or older is considered an adult.

When one encounters a potential victim, the first assessment is to determine unresponsiveness. "Shake and shout" has been a common first action to determine that the victim is really unconscious (there no doubt has been a number of resuscitations initiated upon someone who was sleeping, assuming unconsciousness)! Once there is no doubt that a true emergency exists, the rescuer sends another member of the group to activate the EMS system by phoning 911 and to obtain an AED. If the rescuer is alone, he or she must leave the victim momentarily to call 911 and get the AED; these automated defibrillators have a standardized placement near a telephone. After accessing the EMS and obtaining a defibrillator, the rescuer places the victim in the supine position, and kneels at the head (the left side is suggested when utilizing an AED). The "ABCs" of CPR now are initiated.

A = Airway

The unconscious person has a generalized relaxation of all muscles and in the throat this causes the tongue to move in a posterior direction, occluding the airway. Since the tongue is attached to the mandible, manipulating the jaw and head will retract anteriorly. Two methods are utilized to open the airway; the "head-tilt and chin-lift" or the "jaw-thrust" maneuvers. Tilting the head backward by placing one hand on the forehead and raising the chin with the two fingers of the other hand is the most commonly used technique. An important feature of this technique is to make sure that the fingers are placed on the inferior surface of the mandibular bone and not the soft tissue under the tongue, as the later placement will worsen airway compromise. In a situation where a neck injury is suspected with possible spinal cord compromise, extension of the head is contraindicated; the jaw-thrust is utilized to open the airway. The head is held in the neutral position while applying forward pressure with both hands at the angle of the jaw, just below the ears. In this way, there is no change in head position. At this point inspection of the mouth is important to remove any secretions, vomitus or foreign bodies that may be an impediment to air exchange. Once the airway is opened, the rescuer "looks, listens, and feels" for breathing by placing his or her cheek and ear close to the victim's mouth. The chest is examined for movement while feeling and listening for air passage. In the case of a partial obstruction of the airway, the victim will tend to make high-pitched "crowing" noises that may be accompanied by cyanosis of the skin (due to unoxygenated hemoglobin). Instead of the chest expanding with an inspiration, retraction of thorax or lung compartment will occur. It is imperative that the airway be opened and maintained in this situation since ineffective ventilations will invariably lead to hypoxia.

B = Breathing

Once it is ascertained that the victim is not breathing (this should occur within $\sim$10 s), the rescuer places his mouth

around the victim's mouth and pinches the nose shut with one hand while maintaining chin-lift with the other hand. Two long, extended breaths are given, each ~2 s, with the goal of providing ventilation to the lungs while minimizing the egress of air into the stomach. Since during unconsciousness there is a relaxation of all muscles, the lower esophageal sphincter will relax and thus any air that enters the stomach may force gastric contents into the esophagus and then into the trachea. Aspiration of these highly acidic stomach contents into the lungs may occur. The complex interplay between rescuer, positive pressure ventilation, peak airway pressure, tidal volume, and inspiratory flow rate has had a considerable degree of scientific evaluation (106–109). The consensus supports a tidal volume of between 800 and 1000 mL to maintain adequate oxygenation when only room air is provided in the rescue breathing. This volume is slightly less than the 1992 ECC Guidelines of a rescue tidal volume of 800–1200 mL. A slow prolonged breath over 2 s decreases peak positive pressure and thus entry of air into the stomach while providing the optimum tidal volume. When supplemental oxygen is available, evidence has confirmed that a tidal volume of 500 mL provides effective oxygenation and ventilation in the unintubated patient as long as the inspired oxygen fraction is >40% (110,111). Once rescue breathing is commenced, one should assess effectiveness by noting whether the chest rises with each breath. If there is no change or the rescuer observes that significant effort is required to minimally expand the chest, the airway step has not been optimized and the rescuer has to reopen the airway with additional head extension and chin-lift (or jaw-thrust if a head or neck injury is suspected). If readjustment of the airway does not provide the ability to ventilate, a foreign body lodged in the airway should be suspected and the rescuer should proceed through the algorithm specific for dealing with this issue. Victims with dentures may prove difficult to ventilate; generally dentures should be left in place since it is easier for the rescuer to form a seal around the mouth. However, loose dentures may be aspirated and should be removed if their retention is inadequate. In the case where the rescuer cannot maintain an adequate seal, or if the mouth is unavailable for airway exchange secondary to trauma, mouth-to-nose breathing should be attempted (112). A deep breath should be inhaled by the rescuer prior to respiratory exchange since this maneuver optimizes the maximum amount of oxygen made available for each tidal volume (113). Should the victim only require oxygenation and ventilation, rescue breathing provides one breath every 5 s or 12 breaths/min (114).

There is always the concern regarding exposure to an infectious organism when performing rescue breathing. At this point in time, there has not been any evidence documenting the transmission of human immunedeficiency virus (HIV), hepatitis, or tuberculosis when mouth-to-mouth resuscitation has been instituted in an emergency (115). However, reluctance upon the part of any lay rescue person to perform this action is understandable and there is no moral or legal duty to do so. Barrier devices have been developed that prevent intimate contact with the victim and there are two basic types: face shields and masks. This adjunctive equipment has been made available in the healthcare environments due to the requirements of the Occupational Health and Safety Administration. The face shield has a flexible plastic covering with a one way circular valve that, when placed over the victim, separates the rescuer from contact and from exhaled gases. Mouth-to-mask rescue ventilation provides a better seal and further distance from the victim's mouth than the face shield, which is advantageous should vomiting occur. Some of these masks have a port where supplemental oxygen can be provided and entrained with the rescue breathing. A flow rate of 10 L/min through one of these masks will increase the inspired concentration of oxygen to at least 40% (116). When supplemental oxygen is supplied in this manner, smaller tidal volumes, on the order of 400–600 mL, will maintain oxygenation while decreasing the risk of gastric insufflation (117).

C = Circulation

After delivery of two rescue breaths, the next step in basic CPR is to assess for signs of circulation. For many years the layperson was taught to feel for the presence or absence of a carotid pulse. Research in the 1990s found numerous pitfalls with the pulse check that appeared to have a negative impact on survival and, since 2000, this task is not taught to the lay responder anymore. Significant time delays in trying to determine if a pulse was present delayed time to defibrillation and thus survival (118). The accuracy of the pulse test revealed a sensitivity of only 55% and a specificity of 90% and overall the accuracy was 65% (119). At this time, the lay rescuer is instructed to look for signs of perfusion, such as movement, breathing, or coughing and if unsure, to begin chest compressions.

Correct positioning of the hands and compression skills are easily learned by the layperson. A simplified method for hand placement has been taught for several years and consists of placing the heel of one hand over the center of the breastbone (sternum) between the nipples and then interlocking the fingers of the remaining hand over the first, so that pressure will be transmitted through the heels of both hands. Effective compressions are generated by positioning the rescuer's shoulders over the hands and sternum and depressing the sternum from 1.5 to 2 in. Release after compression must be complete without taking the hands off the chest to prevent "bouncing". Chest compressions should be similar in action to that of a piston in a reciprocating engine with half of the cycle spent in compression and the other half spent in relaxation. The effectiveness of this ratio has been documented with regard to both cerebral and coronary perfusion pressures (120). The recommended rate for chest compressions is 100/min, which has been substantiated by numerous studies (121,122).

The single rescuer initiates chest compressions after providing the victim with two rescue breaths and assessing for signs of effective cardiac blood flow. A ratio of 15 compressions followed by 2 rescue breaths continues for four cycles and then the victim is reassessed for spontaneous circulation. Chest compressions should be resumed within 10 s after noting no signs of perfusion and the 15:2 cycle continued with an interruption for assessment of vital signs in several minutes, followed by the same ratio.

When additional responders are present during resuscitation of a cardiac arrest victim, immediate activation of the EMS system must be accomplished and a defibrillator brought to the scene, if these actions have not already been completed by the lone rescuer. The second rescuer should then assess the adequacy of ventilations and chest compressions and reassess for signs of a pulse and breathing within 10 s while CPR is halted. Though it is not expected that the layperson be able to engage in two-person resuscitation, the process is included here for completeness. Medical professionals as well as the paramedical caregivers should all be able to demonstrate this skill. The compressor is positioned in the normal manner, at the side of the victim. The second rescuer is stationed at the victim's head, maintaining the airway, monitoring for effective compression by carotid artery pulse check, and providing rescue breaths. Previous scientific guidelines utilized a compression:ventilation ratio of 5:2 (123), which has now been changed in light of recent scientific evidence. Currently, a ratio of 15 compressions to 2 ventilations is recommended for both one and two rescuer CPR (124–126) since it appears that improved survival occurs as a result of the higher rate in spite of a decreased number of ventilations. The effectiveness of chest compressions relative to coronary perfusion pressure (the difference between aortic diastolic pressure and the left ventricular end-diastolic pressure) suggests that, as the number of compressions increases, so does the perfusion pressure; therefore, 15 chest compressions improves and sustains blood pressure more effectively than the previous recommendation of 5 compressions to 2 ventilations. The pauses with the previously recommended 5:2 compression:ventilation scheme had more drops in cerebral and coronary perfusion and therefore decreased oxygen delivery compared to the new scheme. Therefore, whether a one- or a two-rescuer resuscitation occurs, the preferred compression/ventilation ratio is 15:2. When advanced life support is initiated and the patient is intubated (a breathing tube place through the mouth and into the trachea) there is no pausing for ventilations; chest compressions continue at 100/min and ventilations are provided at a rate of 12 times a minute (127).

Despite the fact that there has never been evidence to suggest that transmission of disease occurs through mouth-to-mouth exchange of air or secretions, studies have demonstrated a lack of enthusiasm upon the part of both the layperson and professional rescuers to perform this maneuver on strangers (128,129). Current guidelines, as of 2001, now indicate that if the rescuer is unable to perform mouth-to-mouth ventilations, chest compressions should be started for the victim (130). The Cerebral Resuscitation Group of Belgium concluded that there was no difference in outcome for the victim if chest compressions were or were not accompanied by mouth-to-mouth rescue breathing (124). Since any resuscitation attempt utilizing chest compressions without ventilation may provide a better outcome for the victim than no action at all, education regarding this tactic in resuscitation has been made available to the lay responder. While it appears contrary to basic physiological principles that resuscitation could be successful without providing oxygen to the blood, evidence suggests that agonal breathing mechanisms are able to maintain adequate PaO_2 and $PaCO_2$ during CPR without rescue breathing (131). The etiology for this paradox appears to be due to the decreased perfusion from chest compressions; since the cardiac output is only one-fourth that of normal, ventilation perfusion mismatch does not occur due to low rates of blood flow through the lungs. In essence there is a decreased requirement for oxygen; any excess ventilation is wasted due to this decreased perfusion and the lack of oxygen transport by the available red blood cells (132,133). This form of CPR is only recommended for the public rescuer since paramedical personnel should always have adjunctive airway devices available for resuscitation.

AIRWAY OBSTRUCTION

The tongue is the most common cause of airway obstruction and basic life support addresses this issue with various maneuvers to open the airway thus allowing either spontaneous respirations to resume or mouth-to-mouth ventilations to be initiated for the victim. Foreign body airway obstruction is the cause of ~3000 deaths a year (134).

In perspective, there are 198 deaths per 100,000 persons for coronary artery disease, 16.5 deaths per 100,000 individuals for motor vehicle accidents, and 1.2 deaths per 100,000 due to foreign body obstruction (135). The "cafe coronary" (where choking was mistaken for an acute coronary event) appears to be the most common cause of choking in adults since this emergency usually happens during eating and meat seems to be the culprit for most occurrences (136). A foreign body lodged in the airway can either completely occlude or partially occlude any segment of the respiratory passages. The key to distinguishing these two scenarios is that the victim is able to continue to breath, albeit with difficulty, during a partial obstruction, and therefore the rescuer should not attempt any rescue attempt that potentially could convert a partial to a complete obstruction. As in all basic life support, it is crucial to activate the emergency medical system to get assistance. When a victim begins to make high pitched "crowing" sounds, cannot speak, or becomes cyanotic, hypoxia quickly ensues and this person needs immediate aid. The public is taught the universal chocking sign where the neck is clutched with both hands. The first question to ask the choking victim if, in fact, he or she is unable to breathe and if they can speak. The next immediate step is the Heimlich maneuver (137), which should be attempted in anyone between the ages of one and adulthood. This action is not indicated in infants <1 year old (138). Forceful external elevation of the diaphragm utilizes the remaining volume of air in the lungs to expel a foreign body. Placement of the hands is very important to minimize injury to the internal organs; when the victim is standing or sitting the rescuer wraps both arms around the victim's body and clenches the hand to make a fist. This fist, with the thumb compressed against the abdominal skin is placed above the umbilicus and below the xiphoid process (the distal end of the breastbone). A rapid thrust is made in a superior–posterior direction and continued until either the foreign object is displaced or the person becomes unconscious.

Once unconscious the EMS system must be activated by calling 911 and CPR is initiated for the victim. In this circumstance, after the airway is opened, mouth-to-mouth ventilation may be possible due to the muscle relaxation that occurs with unconsciousness, converting a complete obstruction to a partial obstruction, thus allowing rescue breaths to provide oxygenation to the blood. When opening the airway and during subsequent ventilations, the only change in basic CPR is to open the mouth and look for the presence of the offending obstruction, and if visible, grasp the object with the fingers and remove it (139). Blind finger sweeps are prohibited since injury can easily occur to the soft tissues of the mouth and throat. The initiation of chest compressions may also create enough intrathoracic pressure to expel the foreign object (140,141). The recommended maneuver for the obese or pregnant patient is to utilize chest thrust since the increased abdominal girth in either type of victim makes the Heimlich procedure difficult in terms of finding landmarks and avoiding injury, especially to the fetus. The victim is grasped from behind and the arms encircle the chest, just under the armpits, with the fist placed upon the breastbone. Again the thumb of the first hand is placed next to the skin overlying the breastbone and the second hand is then placed over the first, and with the hands interlocked, rapid compressions are performed in a posterior manner.

Once the foreign body is expelled from the unconscious victim, the basic "ABCs" are initiated as in any life-support situation. "Look, listen, and feel" for signs of breathing and if no excursion of air is present, the two rescue breaths are immediately provided for the victim. The next action in the sequence is to observe signs of adequate circulation, and if none are present, chest compressions are initiated and an AED is attached to the patient (142).

THE AUTOMATIC EXTERNAL DEFIBRILLATOR

The key issue in cardiopulmonary resuscitation is rapid initiation of the "Chain of Survival" and the length of time between victim collapse and defibrillation. Public access defibrillation (PAD) has the capability of significantly improving survival rates from cardiac arrest, in some cases to almost 50% (143). Since ventricular fibrillation has the highest frequency of occurrence in cardiac arrest and can only be terminated by countershock, it is clear that early defibrillation is will dramatically improve survival rates. The ECC guidelines of 2000 have a goal defibrillation within five minutes of the EMS activation (911 call).

The first automated cardiac resuscitator was described by Diack et al. in 1979 (144). This 19 lb battery-powered and most importantly, portable device sensed respiratory rate and the ECG. Ventricular fibrillation and asystole could be diagnosed. An oropharyngeal airway utilized a transducer to detect respiratory pressure as well as an electrode, which was applied to the base of the tongue. An electrode was placed over the chest wall (xiphoid area), which completed the electrical circuit for defibrillation. After initial trials in animals, the device was used at St. Vincent Hospital, Portland, Oregon on a 49-year-old male who arrived to the ED in ventricular fibrillation. The patient had failed to respond to CPR, chest defibrillation and medications. The patient was actually seizing from massive doses of lidocaine injected to decrease the automaticity of cardiac condition. The automatic resuscitator converted this lethal rhythm with one 335 J shock via the tongue-epigastric pathway.

There are actually no fully automatic external defibrillators available to the public; this is a misnomer since some actions are required upon the rescuer for the device to work. Once the AED is attached to the patient by adhesive pads and turned on, electronic evaluation occurs of the victim's underlying rhythm, and if a lethal dysrhythmia is evident, the AED will advise the rescuer to activate the "shock" button. These devices first record and then interpret the ECG. Narrow range bandwidth amplifiers first filter out various artifacts such as powerline transmission, high or low frequency radio transmissions, and extraneous "noise" that occurs from poor connections. Successive segments of the filtered ECG are than analyzed through a mathematical algorithm with each segment further tested in a sequential matter. The specifics of these algorithms are not made public due to the competitive nature of the various companies manufacturing these devices, but, essentially the rate, amplitude, waveform, frequency, and baseline variability of the ECG are analyzed and mathematical integration results to identify the rhythm that would or would not be treatable by defibrillation. The accuracy of these devices is extremely high after extensive testing both *in vitro* and *in vivo* to an early skeptical audience of medical personnel. The rare errors encountered with AED function appears to be related to movement of the patient, such as when the patient is being ventilated or repositioned during the analysis mode. Agonal respirations by the victim also create movement artifact that will interfere with rhythm analysis. The errors have generally been of omission in that dysrhythmias that would benefit from defibrillation have not been recognized, such as very coarse or fine ventricular fibrillation (145).

Since the first biphasic waveform was used in an AED in 1996 there has been a progression toward utilizing this new technology in future defibrillators. The current monophasic waveform devices deliver current that is unipolar and either damped sinusoidal or truncated relative to the rate at which the pulsed current decreases to zero. The biphasic defibrillators utilize a biphasic wave form of which each are exactly opposite in polarities. Optimum defibrillation is a balance between producing enough current to terminate VF without extensive damage to the heart. When a monophasic defibrillator is used during CPR, 200 J of energy is recommended for the first shock, with two succeeding shocks of between 200 and 300 J suggested in the AHA VF algorithm. These increases in energy with each shock have been demonstrated to optimize defibrillation success while minimizing tissue damage (146,147). Numerous studies have confirmed that biphasic waveform energy utilizing shocks as low as 115–130 J were as effective in terminating VF as the monophasic 200 J shock (148,149). The current scientific evidence supports conclusions that low-energy biphasic waveforms have at

least the equivalent effectiveness as the monophasic waveforms in defibrillating VF; it appears that biphasic waveform defibrillation will be the standard of care in the future.

Operation of the various AEDs is very straightforward and, although there are different models, the devices have very similar controls and functions. Once the AED is brought to the scene of a victim that is not breathing and has no effective pulse, it should be positioned at the victim's left side for easier electrode placement and to allow the "ABCs" of CPR to continue without interruption on the victim's right side. The machine is turned on, and then a series of prompts by an electronic voice guide the rescuer through the remaining steps. The electrodes are next placed on the skin, with the right pad positioned just under the right collar bone (clavicle) and the left pad placed lateral to the left nipple. The third step requires that everyone involved in the resuscitation desist from any contact with the victim while the AED analyzes the rhythm. This is extremely important in that any movement generated by the rescuers will induce an artifact error. Depending on the particular device, some AEDs will automatically analyze while some machines require manual selection by the operator. If VF or VT is present, an electronic voice will indicate that a shock should be delivered; everyone should clear the victim and there should be no one in contact with the body. Once the operator is sure that everyone is clear of the victim, the "shock" button is depressed and the victim's body will exhibit the generalized musculature contracture observed with a defibrillating current. Cardiopulmonary resuscitation is not resumed since the AED must again be ready to deliver another shock after analyzing the postshock rhythm. In this way, three shocks are successively delivered before resumption of "airway, breathing, and chest compressions". If the "no shock indicated" message is transmitted to the rescuers, either there is now a pulsatile, perfusing rhythm, or there is a lethal dysrhythmia that would not improve survival with a defibrillating shock (such as asystole). If signs of circulation become present, then rescue breathing should continue, unless of course the victim also has a recurrence of adequate ventilations (150).

There are some specific circumstances that must be considered with the use of an AED. Children that are < 8 years of age or victims weighing < 25 kg may not benefit from the current AED since the energy delivered in the monophasic devices are in excess of the recommended 2:4 J/kg. Though it appears that VF can be detected accurately in children, there is insufficient evidence that the algorithms devised for the adult cardiac arrest patient will always determine an arrest rhythm that will benefit from defibrillation, particularly with regard to pediatric tachy-arrhythmias (151,152). However, in the pediatric arrest patient the potential benefit greatly outweighs the risk of AED–individual injury, and therefore this device should be readily available for early application. Another special situation occurs when the victim has been in the water. Even when the victim has been removed from water, there is the risk of induced shock to the rescuer from the moisture present on the victim's body. The other concern, which has the most potential for occurrence, is that the shock would not be conducted along the appropriate pathway from the electrodes, instead bypassing the heart by the water present upon the skin. The third issue regarding AEDs is that many older victims may have a pulse generator present for cardiac rate control or for internal defibrillation. These devices are usually implanted in the superficial tissues of the chest and appear as an elevated, hard lump, which may or may not be easily observed, depending on the obesity of the patient. What can be observed is the scar from implantation; avoidance of pad placement over the implanted device would mitigate any interference for the AED to detect and provide a defibrillating shock. The last situation that may be encountered is the medication patches, which many patients utilize for a variety of conditions. Pad placement over these patches could significantly decrease the energy delivered to the heart; the recommended procedure is to remove the medication patch and wipe the medication away from the skin prior to pad placement (153).

The AED has the potential to be a strategic intervention to increase out-of-hospital survival in the cardiac arrest patient. Public access defibrillation will continue to be more available due to the efficacy already demonstrated in the previously mentioned studies as well as progressive legislation. The continued technology to produce smaller devices at less cost will undoubtedly increase access both in public places and into the homes of those individuals at risk for sudden cardiac death.

PEDIATRIC AND INFANT CARDIOPULMONARY RESUSCITATION

There are significant differences in the etiology for cardiac arrest in the adult compared to children. In adults, a terminal dysrhythmia is usually caused by progression of coronary artery disease where an acute hypoxic event produces hypoxia, injury, and then disruption of organized depolarization of myofibrils. Children are entirely different in that a hypoxic pulmonary event is the usual cause of the progression of a normal sinus rhythm to an extremely slow heart rate (bradycardia) and then to asystole. The majority of these arrests occur from foreign body obstruction, drowning, trauma, or Sudden Infant Death Syndrome (SIDS), poisoning, asthma, or pneumonia (154). This is where the American Heart Association CPR algorithm for the child is different than that of the adult. Since respiratory failure is such a significant etiology for cardiac arrest, the rescuer is instructed to provide CPR first then "phone fast" rather than the adult rescue scenario where activation of the EMS system (phone first) occurs prior to CPR. The remaining steps in the "ABCs" of life support are similar to the adult. Once unresponsiveness is determined for the victim and the rescuer shouts for help, the child should be placed supine, being careful to move the entire body as one if a head or neck injury is suspected. The airway is opened with the "head-tilt, chin-lift" as in the adult, or the "jaw-thrust maneuver" if there is suspicion for head or neck trauma. After opening the airway and inspecting the mouth for a foreign body, "look, listen,

and feel" for the passage of air by the victim within the time frame of no more than ten seconds. If the assessment reveals lack of adequate ventilation, two slow rescue breaths with an adequate pause between them to allow for exhalation are delivered to the victim; the endpoint for tidal volume is a visible rise in the thoracic chest wall. As in the adult, if no air enters the lungs, the most common cause is an inadequate airway lumen secondary to either tongue obstruction or a foreign body. Repositioning of the head should be immediately attempted, and if the rescuer can still not ventilate, entry into the foreign body obstruction algorithm should be the next action undertaken. Should the rescuer provide effective ventilation for the two initial breaths and signs of circulation exist, a rate of 20 breaths/min is recommended for an infant (<1 year of age) and for the child (1–8 years of age.) If no movement, spontaneous breathing, or other signs of circulation exist, chest compressions should be started within 10 s. The depth for compressions in an infant or child should be about one-third to one-half of the total distance from the child's anterior chest wall to the back, and at a rate of 100 times/min. In a child (1–8 years of age), the use of only one hand is recommended, with placement between the nipple line and the bottom of the breastbone (sternum.) As in the adult, the heel of the hand is used. A ratio of five compressions to one ventilation is recommended, with reassessment of the child after 20 cycles of compressions and ventilations have occurred. Chest compressions in an infant (<1 year of age) are similar except that chest compressions are accomplished with two fingers to a depth of one-third to one-half the distance from the anterior chest to the back. The location for compression is one finger-width below an imagined line between the nipples. The rate is 100 times/min and the compression to ventilation ratio is 5:1, as in the child. Reassessment for signs of circulation and spontaneous ventilations should occur after ~ 1 min or 20 cycles of compressions and ventilations (155).

Foreign body obstruction is treated in a similar manner to the adult in a child between 1–8 years of age when assessment by the responder indicates a severe or complete airway occlusion. After immediate activation of the EMS system, the Heimlich maneuver is instituted with the same adult landmarks where the rescuer's fists are place above the umbilicus, being careful to stay away from the inferior aspect of the breastbone, where the xyphoid process is located. Should the child become unresponsive from respiratory arrest, CPR should be initiated, with the initial airway assessment focused upon looking for the foreign object, and, if visible, removing it. Two rescue breaths should be initiated and then chest compressions begun. As in the adult, chest compressions may cause the foreign object to be dislodged where it can be extricated. The infant (under the age of 1 year) is treated quite differently when foreign body aspiration is suspected, since concerns have been raised regarding intraabdominal injury and the Heimlich maneuver (156,157). The infant is cradled with one hand and arm, turned prone, and then five back blows are delivered between the shoulder blades (scapulae) using the heel of the hand. Immediately the victim is turned supine and cradled with the opposite hand and arm. Up to five chest-thrusts are delivered in the same location for chest compressions. This scheme is alternated until either the obstruction is dislodged or the infant becomes unconscious. Once unresponsive, the next action, as in the adult and child, is to look in the airway and determine if the obstruction can be removed. Cardiopulmonary resuscitation is then initiated and activation of the EMS system undertaken (158).

CONCLUSION

Cardiopulmonary resuscitation has made gigantic strides since the scientific community integrated the rudiments of airway maintenance, rescue breathing and external chest compressions in the 1960s. Once these procedures were developed, the foresight by healthcare professionals to implement CPR in the community and educate the public in the early 1970s has contributed greatly to the success of this program. Currently, the strategy is to simplify the training for Basic Life Support as well as to increase the number of lay responders. Since 1973 >40 million people have learned the basic life-saving skills taught in CPR classes (156). The recognition that emergency cardiac care should not only provide an organized structure for resuscitation, but also to incorporate education in the prevention of risk factors for coronary artery disease, stroke, and pediatric mortality has the capability of dramatically reducing future morbidity and death. For example, $\sim 30\%$ of the deaths from atherosclerotic vascular disease are attributable to smoking, and in those individuals who quit smoking, the death rate declines to almost near normal (160,161). Risk factors, such as smoking, that can be changed by providing both education and interventional guidelines, are a continual focus for organizations like the American Heart Association to address at the present and in the future. As previously mentioned, public access for the automated defibrillator has the potential to greatly reduce deaths in the prehospital arrest scenario. As the time element for CPR and defibrillation has been proven to be so critical for reducing morbidity and mortality from cardiac arrest, those communities that have incorporated aggressive public and paramedical training for these two modalities have reported an almost 50% resuscitation rate for victims documented to have had ventricular fibrillation (1,2). The proliferation of AEDs both in public gathering places and into the home, certainly has the capability to improve these statistics. The continued focus upon rapid recognition of stroke victims in lay responder courses will make a dramatic improvement in rapid treatment and neurological salvage for these individuals. It has been stated that the community will be "the ultimate coronary care unit"(162) since the majority of cardiac arrests occur in the out-of-hospital setting and, therefore, public involvement is crucial to survival. The expectation for the future is that the public "coronary care unit" will be expanded to include a "neurological care unit" as well. The challenge for the future will be to continue to expand the public awareness and involvement in these programs as new scientific evidence continues to guide the evolution of cardiopulmonary resuscitation.

HISTORICAL EVENTS IN CPR

Antiquity	Prophet Elisha describes attempts to revive the dead.
Second century	Galen observed the inflation of a dead animal's lungs.
Middle Ages	Hot materials to the abdomen, whipping, rectal smoke.
Paracelsus, sixteenth century	Fireplace bellows ventilated a patient.
Tossach 1744	First recorded mouth-to-mouth resuscitation.
Squires, 1775	First successful defibrillation.
DeHaen, 1783	Chest compression, arm lift technique.
Leroy, 1830	First description of supine ventilation.
Schiff, 1847	Open chest cardiac compression.
Silvester; Howard, twenteeth century	Alternating arm position ventilation; back, abdominal and chest pressure, respectively.
Koenig; Maass, 1850s	Reports of eight successful closed-chest cardiac compressions in humans.
Holger-Nielson, twenty-first century	Prone back pressure, arm-lift.
Gurvich; Yuniev, 1939	First successful external defibrillation by a device.
Kouwenhoven, Knickerbocker, Isaacs, 1950s	Defibrillation experiments coupled with chest compressions.
Elam, 1960s	Physiology of rescue breathing.
Elam, Safar, Kouwenhoven, 1960s	Principles of modern CPR.
Zoll, 1956	First successful external defibrillation in humans.

BIBLIOGRAPHY

1. White RD, Asplin BR, Bugliosi TF, Hankins DG. High discharge survival rate after out of-hospital ventricular fibrillation by police and paramedics. Ann Emerg Med 1996;28: 480–485.
2. White RD, Hankins DG, Bugliosi TF. Seven years experience with early defibrillators by police and paramedics in an emergency medical services system. Resuscitation 1998;39: 145–151.
3. Holy Bible, Kings II 4:34–35 (King James Version).
4. Baker AB. Artificial respiration: the history of an idea. Med Hist 1971;15:336–346.
5. [Anonymaus]. Cardiopulmonary Resuscitation Conference Proceedings. The Ad Hoc Committee on Cardiopulmonary Resuscitation; Division of Medical Sciences of the National Research Council; 1967. p 7.
6. Julian DG. Cardiac resuscitation in the eighteenth century. Heart Lung 1975;4:46–48.
7. Gordon AS. The principles and practice of heart-lung resuscitation. Acta Anaesth Scand Suppl 1961;9:134–147.
8. Gordon AS, Affeldt JE, Sadove M, Raymon F, Whittenberger JL, Ivy AL. Air-flow patterns and pulmonary ventilation during manual artificial respiration on apneic normal adults. J Appl Physiol 1955;4:408–420.
9. Gordon AS, Fainer DC, Ivy AL. Artificial respiration: a new method and a comparative study of different methods in adults. JAMA 1950;144:1455–1464.
10. Elam JO. Rediscovery of expired air methods for emergency ventilation. In: Safar P, Elam JO, editors. Advances in Cardiopulmonary Resuscitation. Chapt. 39, New York: Springer-Verlag; 1977. p 263–265.
11. Elam JO, Brown ES, Elder JD Jr. Artificial respiration by mouth-to-mask method. A study of the respiratory gas exchange of paralyzed patients ventilated by operator's exhaled air. New Engl J Med 1954;250:749–754.
12. Safar P, Elam J. Manual versus mouth-to-mouth methods of artificial respiration. Anesthesiology 1958;19:111–112.
13. Safar P, Escarraga LA, Elam JO. A comparison of the mouth-to-mouth and mouth-to-airway methods of artificial respiration with the chest-pressure arm-lift methods. New Engl J Med 1958;258:671–677.
14. Safar P. Failure of manual respiration. J Appl Physiol 1959;14:84–88.
15. Safar P. Ventilatory efficacy of mouth-to-mouth artificial respiration. Airway obstruction during manual and mouth-to-mouth artificial respiration. JAMA 1958;167:335–341.
16. Safar P, Aguto-Escarraga L, Chang F. Upper airway obstruction in the unconscious patient. J Appl Physiol 1959;14: 760–764.
17. Morikawa S, Safar P, DeCarlo J. Influence of head-jaw position upon upper airway patency. Anesthesiology 1961; 22: 265–270.
18. Safar P, Redding J. The "tight jaw" in resuscitation. Anesthesiology 1959;20:701–702.
19. Elam JO, Greene Dg, Brown ES, Clements JA. Oxygen and carbon dioxide exchange and energy costs of expired air resuscitation. JAMA 1958;167:328–324.
20. Gordon AS, Frye CW, Gittelson L, Sadove MS, Beattie EJ. Mouth-to-mouth versus manual artificial respiration for children and adults. JAMA 1951;147:1444–1453.
21. Jude JR, Kouwenhoven WB, Knickerbocker GG. Cardiac arrest. JAMA 1961;128:1063.
22. Boehm RV. Arbeiten aus dem pharmakologischen Institue der Universitat Dorpat:13. Ueber Wiederbelebung nach Vergiftungen un Asphysie. Arch Exper Path U Pharmakol 1878;8:68–101.
23. Safar P. History of cardiopulmonary resuscitation. In: Kaye W, Bircher N, editors. Cardiopulmonary Resuscitation. New York: Churchill Livingstone; 1989.
24. Maass Die Methode der. Wiederbelebung bei Herztod nach chloroformeinathmung, Berlin Klin. Wochschr 1892;29: 265–268.
25. Crile WG. Surgical Anemia and Resuscitation. New York: D Appleton and Co; 1904. p 220–244.
26. Hooker DR, Kouwenhoven WB, Langworthy OR. The effect of alternating currents on the heart. Am J Physiol 1933;103: 444–454.
27. Gurvich HL, Yuniev GS. Restoration of heart rhythm during fibrillation by condenser discharge. Am Rev Soviet Med 1947;4:252–256.
28. Kouwenhoven WB, Jude JR, Knickebocker GG. Closed-chest cardiac massage. JAMA 1960;173:1064–1067.
29. Safar P. Initiation of closed-chest cardiopulmonary resuscitation basic life support. A personal history. Resuscitation 1989;18:7–20.
30. Kouwenhoven WB, Langworth OR. Cardiopulmonary resuscitation. An account of forty-five years of research. JAMA 1973;226:877–886.

31. Jude JR, Kouwenhoven WB, Knickerbocker GG. Cardiac arrest: report of application of external cardiac massage on 118 patients. JAMA 1961;178:1063–1070.
32. Safar P, Brown TC, Holtey WH. Ventilation and circulation with closed chest cardiac massage in man. JAMA 1961;176: 574–576.
33. Harris LC, Kirimli B, Safar P. Ventilation-cardiac compression rates and ratios in cardiopulmonary resuscitation. Anesthesiology 1967;28:806–813.
34. Zoll PM, et al. Termination of ventricular fibrillation in man by externally placed electric countershock. New Engl J Med 1956;254:727.
35. Weale FE, Rothwell-Jackson RL. The efficiency of cardiac massage. Lancet 1962;1:990–992.
36. Del Guercio LR, Coomraswamy RP, State D. Cardiac output and other hemodynamic variables during external cardiac massage in man. New Engl J Med 1963;269:1398–1404.
37. Thomsen JE, Stenlund RR, Row GG. Intracardiac pressures during closed chest cardiac massage. JAMA 1968; 205:116–118.
38. Criley JM, Blaufuss AH, Kissel GL. Cough induced cardiac compression. JAMA 1976;236:1246–1250.
39. Taylor GJ, Tucker WM, Greene HL, Rudikoff MT, Weisfeldt ML. Importance of prolonged compression during cardiopulmonary resuscitation in man. New Engl J Med 1977;296: 1515–1517.
40. Chandra N, Rudikoff MT, Weisfeldt ML. Simultaneous chest compression and ventilation at high airway pressure during cardiopulmonary resuscitation. Lancet 1980;1:175–178.
41. Werner JA, Greene HL, Janko CL, Cobb LA. Visualization of cardiac valve motion in man during external compression using two-dimensional echocardiography: implications regarding the mechanism of blood flow. Circulation 1981; 63:1417–1421.
42. Rich S, Wix HL, Shapiro EP. Clinical assessment of heart chamber size and valve motion in man during external compression using two-dimensional echocardiography: implications regarding the mechanism of blood flow. Am Heart J 1981;102:368–373.
43. Maier GW, et al. The physiology of external cardiac massage: high impulse cardiopulmonary resuscitation. Circulation 1984;70:86–101.
44. Feneley MP, et al. Sequence of mitral valve motion and transmitral flow during manual cardiopulmonary resuscitation in dogs. Circulation 1987;76:363–375.
45. Porter TR, et al. Transesophageal echocardiography to assess mitral valve function and flow during cardiopulmonary resuscitation. Am J Cardiol 1992;70:1056–1060.
46. West JB. Physiological Basis of Medical Practice. 12th ed. Baltimore: Williams & Wilkins; 1990. p 522.
47. Permut S, Bromberger-Barnea B, Bane HN. Alveolar pressure, pulmonary venous pressure and the vascular waterfall. Med Thorac 1962;19:239.
48. Cummins RO, editor. ACLS Provider Manual. American Heart Association; 2002. p 78.
49. Vukmir RB. Torsades de pointes: a review. Am J Emerg Med 1991;9:250–255.
50. Cummins RO, editor. Advanced Cardiac Life Support. American Heart Association; 1997-1999. p 39–40.
51. American Heart Association in collaboration with International Liaison Committee on Resuscitation and Emergency Cardiovascular Care: International Consesus on Science Part 3: Adult Basic Life Support. Circulation 2000; 102(Suppl I): I–226.
52. Brown DC, Lewis AJ, Criley JM. Asystole and its treatment: the possible role of the parasympathetic nervous system in cardiac arrest. J Am Coll Emerg Phys 1979;8: 448–542.
53. Thompson BM, Brooks RC, Pionkowski RS, Aprahamian C, Mateer JR. Immediate countershock treatment of asystole. Ann Emerg Med 1984;13:827–829.
54. Bocka JJ. External transcutaneous pacemakers. Ann Emerg Med 1989;18:1280–1286.
55. Paradis NA, Martin GB, Goetting MG, Rivers EP, Feingold M, Nowak RM. Aortic pressure during human cardiac arrest: identification of pseudo-electromechanical dissociation. Acta Anaesthesiol Scand 1991;35:253–256.
56. Bocka JJ, Overton DT, Hauser A. Electromechanical dissociation in human beings: an echocardiographic evaluation. Ann Emerg Med 1988;17:450–452.
57. Cummins RO editor. ACLS Provider Manual, American Heart Association. 2001. p 103–104.
58. Rosenberg D, Levin E, Myerburg RJ. A mnemonic for the recall of causes of electro-mechanical dissociation (EMD). Resuscitation 1999;40:57.
59. Cummins RO, Eisenberg MS. Prehospital cardiopulmonary resuscitation: is it effective?. JAMA 1985;253:2408–2412.
60. Cummins RO. From concept to standard-of-care? Review of the clinical experience with automated external defibrillators. Ann Emerg Med 1989;18:1269–1275.
61. Cummins RO, Eisenberg MS. Prehospital cardiopulmonary resuscitation: is it effective?. JAMA 1985;253:2408.
62. Valenzuela TD, Roe DJ, Nichol G, Clark LL, Spaite DW, Hardman RG. Outcomes of rapid defibrillation be security officers after cardiac arrest in casinos. N Engl J Med 2000; 343:1206–1209.
63. Cummin RO.. Advanced Cardiac Life Support, American Heart Association. 1997–1999. p 16–4.
64. Levine RL, Wayne MA, Miller CC. End-tidal carbon dioxide and outcome of out-of hospital cardiac arrest. New Eng J Med 1997;337:301–306.
65. Graves EJ. National hospital discharge survey: annual summary 1991. Vital Health Statistics 1993;13:1–62.
66. 2000 Heart and Stroke Statistical Supplement. Dallas: American Heart Association; 1999.
67. Eisenberg MD, Mengert TJ. Cardiac Resuscitation. New Engl J Med 2001;344:1304–1313.
68. Gillum RF. Trends in acute myocardial infarction and coronary heart disease death in the United States. J Am Coll Cardiol 1994;23:1273–1277.
69. Kannel WB, Schatzkin A. Sudden death: lessons from subsets in population studies. J Am Coll Cardiol 1985;5(Suppl 6); 141B–149B.
70. Mathey DG, Sheehan FH, Schofer J, Dodge HT. Time from onset of symptoms to thrombolytic therapy: a major determinant of myocardial salvage in patients with acute transmural infarction. J Am Coll Cardiol 1985;6: 518–525.
71. Anderson JL, Karagounis LA, Califf RM. Metaanalysis of five reported studies on the relation of early coronary patency grades with mortality and outcomes after acute myocardial infarction. Am J Cardiol 1996;78:1–8.
72. Kannel WB. Prevalence and clinical aspects of unrecognized myocardial infarction and sudden unexpected death. Circulation 1987;75(Suppl 2, pt 2): II–4–II–5.
73. Cummins RO. Advanced Cardiac Life Support, American Heart Association. 1997–1999. p 9–4.
74. Herrick J. Clinical features of sudden obstruction of the coronary arteries. JAMA 1912;59:2015–2020.
75. Smith M, Little WC. Potential precipitating factors of the onset of myocardial infarction. Am J Med Sci 1992;303:141–144.

76. Muller JE, Tofler GH, Stone PH. Circadian variation and triggers of onset of acute cardiovascular disease. Circulation 1989;79:733–744.
77. Peters RW, et al. Identification of a secondary peak in myocardial infarction onset 11 to 12 hours after awakening: the Cardiac Arrhythmia suppression Trial (CAST) experience. J Am Coll Cardiol 1993;22:998–1003.
78. Solomon CG, et al. Comparison of clinical presentation of acute myocardial infarction in patients older than 65 years of age to younger patients: the Multicenter Chest Pain Study experience. Am J Cardiol 1989;63:772–776.
79. Douglas PS, Ginsburg GS. The evaluation of chest pain in women. New Engl J Med 1996;334:1311–1315.
80. Brand FN, Larson M, Friedman LM, Kannel WB, Castelli WP. Epidemiologic assessment of angina before and after myocardial infarction: the Framingham Study. Am Heart J 1996;132(pt 1): 174–178.
81. Mathey DG, Sheehan FH, Schofer J, Dodge HT. Time from onset of symptoms to thrombolytic therapy: a major determinant of myocardial salvage in patients with acute transmural infarction. J Am Coll Cardiol 1985;6:518–525.
82. Newby LK, et al. Time from symptom onset to treatment and outcomes after thrombolytic therapy: GUSTO-1 Investigators. J Am Coll Cardiol 1996;27:1646–1655.
83. Berger CJ, et al. Prognosis after first myocardial infarction: comparison of Q-wave and non-Q-wave myocardial infarction in the Framingham Heart Study. JAMA 1992;268: 1545–1551.
84. Gibson RS. Non-Q-wave myocardial infarcton: pathophysiology, prognosis, and therapeutic strategy. Annu Rev Med 1989;40:395–410.
85. TIMI investigators. Effects of tissue plasminogen activator and a comparison of early invasive and conservative strategies in unstable angina and non-Q-wave myocardial infarction: results of the TIMI IIIB Trial. Thrombolysis in Myocardial Ischemia. Circulation 1994; 89:1545–1556.
86. Cohen MV, et al. The effects of nitroglycerin on coronary collaterals and myocardial contractility. J Clin Invest 1973;52:2836–2847.
87. Malindzak GS. Jr., Green HD, Stagg PL. Effects of nitroglycerin on flow after partial constriction of the coronary artery. J Appl Physiol 1970;29:17–22.
88. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Randomized trial of intravenous streptokinase, oral aspirin, both or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. Lancet 1988;2:349–360.
89. Lee G, et al. Comparative effects of morphine, meperidine and pentazocine on the cardiocirculatory dynamics in patients with acute myocardial infarction. Am J Med 1976; 60:949–955.
90. 2000 Heart and Stroke Statistical Update. Dallas, TX: American Heart Association; 1999.
91. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. New Engl J Med 1995;333:1581–1587.
92. Bendzus M, Urbach H, Ries F, Solymosi L. Outcome after local intra-arterial fibrinolysis compared with the natural course of patients with a dense middle cerebral artery on early CT. Neuroradiology 1998;40:54–48.
93. Kwiatkowski TG, et al., for the NINDS r-tPA Stroke Study Group. Effects of tissue plasminogen activator for acute ischemic stroke at one ear. New Engl J Med 1993;340: 1781–1787.
94. Albers GW, et al. Intravenous tissue-type plasminogen activatior for treatment of acute stroke: the Standard Treatment with Alteplase to Reverse Stroke (STARS) Study. JAMA 2000;283:1145–1150.
95. Barnett HJ. The pathophysiology of transient cerebral ischemic attacks. Med Clin North Am 1979;63:649–679.
96. Viitanen M, Eriksson S, Asplund K. Risk of recurrent stroke, myocardial infarction and epilepsy during long- term follow-up after stroke. Eur Neurol 1988;28:227–231.
97. North American Symptomatic Carotid Endarterectomy Trial Collaborators. Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade stenosis. New Engl J Med 1991;325:445–453.
98. Antiplatelet Trialists' Collaboration. Secondary prevention of vascular disease by prolonged antiplatelet treatment. BMJ 1988;296:320–331.
99. Weir B, editor. Aneurysms Affecting the Nervous System. Baltimore: Williams & Wilkins; 1987.
100. Brott T, Thalinger K, Hertzberg V. Hypertension as a risk factor for spontaneous intracerebral hemorrhage. Stroke 1986;17:1078–1083.
101. Waga S, Otsubo K, Handa H. Warning signs in intracranial aneurysms. Surg Neurol 1975;3:15–20.
102. Korpelainen JT, Sotaniemi KA, Makikallio A. Dynamic behavior of heart rate in ischemic stroke. Stroke 1999;30: 1008–1013.
103. Hazinski MF. Demystifying recognition and management of stroke. Curr Emerg Cardiac Care, Winter 1996;7:8.
104. Lyden PD, Rapp K, Babcock T, Rothcock J. Ultra-rapid identification, triage, and enrollment of stroke patients into clinical trials. J Stroke Cerebrovas Dis 1994;4:106–107.
105. Kidwell CS, et al. Design and retrospective analysis of the Los Angeles prehospital stroke screen (LAPSS). Prehosp Emerg Care 1998;2:267–273.
106. Wenzel V, Idris Ah, Lindner KH. Ventilation with an unprotected airway during cardiac arrest. In: Vincent JL, editor. Yearbook of Intensive Care and Emergency Medicine. Berlin: Springer-Verlag; 1997: 483–492.
107. Wenzel V, Idris AH. The current status of ventilation strategies during cardiopulmonary resuscitation. Curr Opin Crit Care 1997;3:206–213.
108. Stalinger A, et al. Effects of different mouth-to-mouth ventilation tidal volumes on gas exchange during simulated rescue breathing (abstract). Crit Care Med 2001. Forth-coming.
109. Dorges V, et al. Smaller tidal volumes with room air are not sufficient to ensure adequate oxygenation during bag-valve-mask ventilation. Resuscitation 2000;44:37–41.
110. Baskett P, Nolan J, Parr M. Tidal volumes which are perceived to be adequate for resuscitation. Resuscitation. 1996;31:231–234.
111. Wenzel V, et al. Effects of smaller tidal volumes during basic life support ventilation in patients with respiratory arrest: good ventilation, less risk?. Resuscitation 1999;43: 25–29.
112. Heartsaver CPR: A comprehensive course for the lay responder. Dallas: American Heart Association; 2000. p 23–29.
113. Htin KJ, et al. Rescuer breathing pattern significantly affects O_2 and CO_2 received by patient during mouth-to-mouth ventilation. Crit Care Med 1998;26: (Suppl 1) A56.
114. Stapleton ER, Auferheide TP, Hazinski MF. BLS for Healthcare Providers. Dallas: American Heart Association; 2001. p 68–70.
115. Cummins RO, editor. ACLS Provider Manual. Dallas: American Heart Association; 2001. p 208.
116. Johannigman JA, Branson RD. Oxygen enrichment of expired gas for mouth-to mask resuscitation. Respir Care 1991;36:99–103.

117. Idris AH, Wenzel V, Banner MJ, Melker RJ. Smaller tidal volumes minimize gastric inflation during CPR with an unprotected airway. Circulation 1995;92(Suppl I): I–759.
118. Cummins RO, Hazinski MF. Cardiopulmonary resuscitation techniques and instruction: when does evidence justify revision?. Ann Emerg Med 1999;34:780–784.
119. Eberle B, et al. Checking the carotid pulse check: diagnostic accuracy of first responders in patients with and without a pulse. Resuscitation 1996;33:107–116.
120. Handley AJ, Handley JA. The relationship between rate of chest compression and compression:relaxation ratio. Resuscitation 1995;30:237–241.
121. Swenson RD, et al. Hemodynamics in humans during conventional and experimental methods of cardiopulmonary resuscitation. Circulation 1998;78:630–639.
122. Kern KB, et al. A study of chest compression rates during cardiopulmonary resuscitation in humans. Arch Intern Med 1992;152:145–149.
123. Guidelines for cardiopulmonary resuscitation and emergency cardiac care. Emergency Cardiac Care Committee and Subcommittees, American Heart Association. JAMA 1992;268: 2171–2295.
124. Van Hoeyweghen RJ, et al., Quality and efficiency of bystander CPR: Belgian cerebral resuscitation. Resuscitation 1993;26: 47–52.
125. Kern KB, et al. Efficacy of chest compression—only BLS/CPR in the presence of an occluded airway. Resuscitation 1998;39: 179–188.
126. Wik L, Steen PA. The ventilation-compression ratio influences the effectiveness of two rescuer advanced cardiac life support on a manikin. Resuscitation 1996;31:113–119.
127. Cummins RO, editor. ACLS Provider Manual. Dallas: American Heart Association; 2001. p 30.
128. Hew P, Brenner B, Kaufman J. Reluctance of paramedics and emergency medical technicians to perform mouth-to-mouth resuscitation. J Emerg Med 1997;15:279–284.
129. Locke CJ, et al. Bystander cardiopulmonary resuscitation: concerns about mouth-to-mouth contact. Arch Intern Med 1995;155:938–943.
130. Stapleton ER, Aufderheide TP, Hazinski MF, editors. BLS for Healthcare Providers. Dallas: American Heart Association; 2001. p 80.
131. Tang W, et al. Cardiopulmonary resuscitation by precordial compression but without mechanical ventilation. Am J Resp Crit Care Med 1994;150:1709–1713.
132. Weil MH, et al. Differences in acid- base status between venous and arterial blood during cardiopulmonary resuscitation. New Engl J Med 1986;315:153–156.
133. Sanders AB, et al. Acid-base balance in a canine model of cardiac arrest. Ann Emerg Med 1988;17:667–671.
134. National Safety Council. Injury Fact. 1999. Itasca: National Safety Council; 1999.
135. National Safety Council. Accident Facts. 1997. Chicago: National Safety Council; 1997.
136. Ekberg O, Feinberg M. Clinical and demographic data in 75 patients with near-fatal choking episodes. Dysphagia 1992;7:205–208.
137. Heimlich HJ, Hoffmann KA, Canestri FR. Food-choking and drowning deaths prevented by external subdiaphragmatic compression: physiological basis. Ann Thoracic Surg 1975;20:188–195.
138. American Heart Association, International Liaison Committee on Resuscitation (ILCOR). Guidelines 2000 for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. International Consensus on Science. Circulation 2000;102 (Suppl I): I46–I–48.
139. Heartsaver CPR: A comprehensive course for the lay responder. Dallas: American Heart Association; 2000. p 31–33.
140. Langhell A, Sunde K, Wik L, Steen PA. Airway pressure with chest compression versus Heimlich manoeuvre in recently dead adults with complete airway obstruction. Resuscitation 2000;44:105–108.
141. Skullberg A. Chest compressions: an alternative to the Heimlich manoeuvre?. Resuscitation 1992;24:91.
142. Stapleton ER, Auferheide TP, Hazinski MF. BLS for Healthcare Providers. Dallas: American Heart Association; 2001. p 128.
143. White RD, Vukov LF, Bugliosi TF. Early defibrillation by police; initial experience with measurement of critical time intervals and patient outcome. Ann Emerg Med 1996;28:480–485.
144. Diack AW, Welborn WS, Rullman RG, Walter CW, Wayne MA. An automatic cardiac resuscitator for emergency treatment of cardiac arrest. Med Instrum 1979; Mar–Apr; 13(2): 78–83.
145. Cummins RO, Eisenber M, Bergner L, Murray JA. Sensitivity, accuracy, and safety of an automatic external defibrillator. Lancet 1984;2:318–320.
146. Kerber RE, et al. Energy, current, and success in defibrillation and cardioversion: clinical studies using an automated impedence-based method of energy adjustment. Circulation 1988;77:1038–1046.
147. Dahl C, et al. Myocardial necrosis from direct current countershock. Circulation 1974;50:956.
148. Bardy GH, et al. Truncated biphasic pulses for transthoracic defibrillation. Circulation 1995;91:1768–1774.
149. Bardy GH, et al. Transthoracic Investigators. Multicenter comparison of truncated biphasic shocks and standard damped sine wave monophasic shocks for transthoracic ventricular defibrillation. Circulation 1996;94:2507–2514.
150. Auferheide TP, Stapleton ER, Hazinski MF. Heartsaver AED for the Lay Rescuer and First Responder: Adult Cardiopulmonary Resuscitation and Automated External Defibrillation. Dallas: American Heart Association; 2002. p 4–11.
151. Cecchin F, et al. Accuracy of automatic external defibrillator analysis algorithm in young children. Circulation 1999; 100:I–663.
152. Hazinski MF, Walker C, Smith J, Deshpande J. Specificity of automatic external defibrillator (AED) rhythm analysis in pediatric tachyarrhythmias. Circulation 1997;96(Suppl I); I561.
153. Stapleton ER, Auferheide TP, Hazinski MF. BLS for Healthcare Providers. Dallas: American Heart Association; 2001. p 95–98.
154. Sirbaugh PE, et al. A prospective, population-based study of the demographics, epidemiology, management, and outcome of out-of-hospital pediatric cardiopulmonary arrest. Ann Emerg Med 1999;33:174–184.
155. Heartsaver CPR: A comprehensive course for the lay responder. Dallas: American Heart Association; 2000. p 101–107.
156. Hazinski MF, editor. PALS Provider Manual. Dallas: American Heart Association; 2002. p 64.
157. Fink JA, Klein RL. Complications of the Heimlich maneuver. J Pedatr Surg 1989;24:486–487.
158. Heartsaver CPR: A comprehensive course for the lay responder. Dallas: American Heart Association; 2000. p 101–107.
159. Heartsaver CPR: A comprehensive course for the lay responder. Dallas: American Heart Association; 2000. p 39.
160. Gordon T, Kannel WB, McGee D, Dawber TR. Death and coronary attacks in men after giving up cigarette smoking: a report from the Framingham Study. Lancet 1974;2:1345–1348.

161. US Dept of Health and Human Services. Reducing the Health Consequences of Smoking: 25 Years of Progress: A Report of the Surgeon General. US Dept of Health and Human Services, Public Health Service, Centers for Disease Control, Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. 1989. DHHS Publication (CDC); 89–8411.
162. McIntyre KM. Cardiopulmonary resuscitation and the ultimate coronary care unit. JAMA 1980;244:510–511.

See also CARDIAC OUTPUT, THERMODILUTION MEASUREMENT OF; RESPIRATORY MECHANICS AND GAS EXCHANGE; SHOCK, TREATMENT OF; VENTILATORS, ACUTE MEDICAL CARE; VENTILATORY MONITORING.

CARTILAGE AND MENISCUS, PROPERTIES OF

JEREMY J. MAO
ALEXANDER J. TROKEN
NICHOLAS W. MARION
University of Illinois
Chicago, Illinois

LEO Q. WAN
VAN C. MOW
Columbia University
New York, New York

INTRODUCTION

Synovial or diarthrodial joints are created to enable the movement between bones. Articular cartilage on the end of articulating bone, therefore, must accomplish two functions: (1) absorb, distribute, and transmit mechanical loading, and (2) create a low friction and wear surface for movement over decades of mammalian life. Three cartilage phenotypes exist: hyaline cartilage, fibrocartilage, and elastic cartilage. In the older literature, articular cartilage is often referred to as hyaline cartilage due to its glassy appearance; this appearance is derived from its high proteoglycan content. Indeed, articular cartilage has the highest proteoglycan content of all biological tissues, while, at the same time, it has the lowest cellular content. The chondrocytes not only secret and control collagen and proteoglycan contents in the extracellular matrix, but also are responsible for regulating the elaborate molecular architecture of these macromolecules and their ultrastructural organization (1–3). Throughout life, the healthy chondrocytes under normal conditions secrete and elaborate sufficient amounts of the extracellular matrix macromolecules and completely encase themselves in an environment that possesses truly remarkable biomechanical mechanisms that protect them against the mechanical insults associated with joint loading, and thus survive for long periods of time under normal health conditions (4–6).

Cartilage in a small number of joints in humans, such as the knee meniscus, temporo-mandibular joint, and intervertebral discs, is fibrocartilage. The intervertebral disc, besides its complex macromolecular architectural and ultrastructural organization, also has a complex macrostructural organization; the latter is manifested in the macro-layering of the outer rings of collagen-rich annulus fibrosis and an inner core of a proteoglycan-rich "kidney-shaped" nucleus pulposus. The cells of these fibrocartilaginous tissues are fibroblasts and chondrocytes, some of which are called fibrochondrocytes. Whereas the genotype and phenotype of cartilage cells determine the biochemical and molecular properties of cartilage, the mechanical properties of articular cartilage are largely dependent on the constituents of extracellular matrix (3). This divergence in the determination of biological and mechanical properties is attributed to the scarce cellularity in adult cartilage, with chondrocytes that account for only less than 10% of the adult cartilage volume (4,7). Comprehensive reviews of hyaline and fibrocartilage can be found elsewhere (1,3,8).

An average human takes approximately 2 million steps per year. The joints in the lower limbs, therefore, can undergo 1–4 million cyclic loads from physical activities (9,10). These loads can peak 4–5 times body weight (11,12), and can cause both macro- and micro-structural changes in articular cartilage that may ultimately lead to degenerative diseases such as osteoarthritis (4,7,13–15). Arthritis, which encompasses more than 100 diseases and conditions, is recognized as among the leading causes of physical disability worldwide (16). Thus, investigation of the properties of normal and arthritic cartilage is essential not only for the understanding of the etiology of arthritis (6), but also devising possible approaches toward the tissue engineering of cartilage and meniscus for clinical treatment modalities (17).

ARTICULAR CARTILAGE AND MENISCUS: COMPOSITION AND STRUCTURE

Chondrocytes and fibrochondrocytes are responsible for the morphogenesis, matrix synthesis, and maintenance of articular cartilage and meniscus as functional tissues. However, these cells only account for approximately 10% of the total cartilage volume in adults (3,7,18,19). Chondrocytes receive nutrients and shed metabolic waste products largely from convective transport and diffusion, either from/to the synovial fluid or from subchondral bone. In the adult, articular cartilage is generally aneural and avascular. Vasculature is present only in the periphery of mature meniscus (20).

Hyaline Cartilage and Articular Cartilage

Hyaline cartilage is present on the articulating surfaces of the bones in most, but not all, synovial joints. Articular cartilage serves to bear and distribute load and contribute to joint lubrication. It serves these different purposes through varying the amount of water relative to the amounts of Type I collagen and proteoglycans and the molecular and ultrastructural organizations of these structural molecules. Healthy articular cartilage appears smooth, bluish white, glistening, and intact. Osteoarthritic articular cartilage appears dull and coarse and may have tears and frays. Hyaline cartilage is also present in the growth plate at the metaphyseal region of long bones and in the cranial base and serves to enable longitudinal bone growth by endochondral ossification (21–23). A review of

growth plate cartilage is beyond the scope of this chapter, but can be found elsewhere (24–26).

Articular cartilage has two immiscible phases—a solid phase and a fluid phase. Small electrolytes such as Na^+ and Cl^- are dissolved in the fluid phase and are freely mobile by diffusion and convection through the porous-permeable solid phase. The fluid and solid phases have been modeled in the now classic biphasic theory developed by Mow et al. (27). Normal fluid component ranges from 75 to 80% by wet weight, and the remaining 20 to 25% of the organic matrix forms solid material with complex material properties (3,18,19). Up to 65% of the solid ECM by dry weight is made of collagen, whereas proteoglycans constitute up to 25%; other glycoproteins, chondrocytes, and lipids can generally make up 10% (3,28,29). Collagen fibers are classified on the basis of their amino acid composition and molecular structure. Although an assortment of collagens exist in both hyaline cartilage and fibrocartilage, Type II collagen is most prevalent in articular cartilage, whereas Type I collagen is most common in the meniscus (3,30). The collagen fibers are assembled as tight triple-helical structures made from three polypeptide alpha-chains. The triple helices are then arranged as tropocollagen molecules, which are wound in a helical manner to form larger collagen fibers that are, in turn, are organized into a strong cohesive collagen network (3,31,32). This arrangement allows for considerable tensile stiffness and strength (33–38). The collagen also serves to restrain the swelling pressure created from the surrounding embedded proteoglycans (2,18,19,39). Proteoglycans (PGs) are hydrophilic macromolecules with numerous glycosaminoglycans (chondroitin and keratin sulfates) attached to a protein core; the protein core of this bottle-brush-shaped molecular is, in turn, attached to a hyaluronan (mw: molecular weight $\sim 0.5 \times 10^6$) resulting in a supra-macromolecule with an approximate molecular weight ranging from $(200 \text{ to } 300) \times 10^6$ (3,29,40–42). These enormous, negatively charged molecules are trapped in the fine porous meshwork of collagen by frictional and electrostatic forces and by steric exclusion; thus, in the ECM, PGs function largely to generate osmotic pressure (2,39) and to resist the compressive stresses of articulation acting on the cartilaginous surface. Although various PGs exist in cartilage, the one that constitutes up to 80–90% of the total PGs in cartilage is aggrecan (3). As the name implies, aggrecan facilitates the formation of large aggregates. Like collagen, aggrecan, as with all PGs, maintains a structure that is directly correlated with its function. The general structure of PGs occurs through noncovalent bonding of aggrecans to the hyaluronan via link proteins, thus securing firm linkages. Attached to the protein core are the glycosaminoglycan side chains (GAGs) that are vital for biological and biomechanical functions of the tissues; indeed, they are hallmarks of chondrogenic activity in tissue engineering. These GAGs bear the necessary physical properties that ultimately confer onto these tissues their hydrophyllic tendencies and compressive load-carriage abilities (3,18,19). The presence of large numbers of sulfate and carboxyl groups on the GAGs gives rise to a high negative-charge density in the ECM (2). This anionic nature attracts positively charged ions, creating an osmotic pressure, known as Donnan osmotic pressure, that favors tissue hydration (3,39,43). The fixed-negative charges also create intense repulsive forces of the GAGs against each other. This expansion force of the PG molecules causes tensile stresses to be developed within the surrounding collagen network surrounding the PGs. This swelling pressure thus resists the compressive forces against the cartilage without volume loss.

Articular cartilage is organized into three layers or zones (3,44) as shown in Fig. 1. The superficial zone forms the articular surface, whereas the deep zone is anchored to calcified cartilage and subchondral bone; both zones have well-defined collagen architectures. An intermediate zone exists in between with a random collagen fiber ultrastructural organization. These layered structural arrangements have long been hypothesized to be important in cartilage function (45–47). The overall thickness of articular cartilage, including the three zones, varies between joints, age, individuals, and species from less than a millimeter to a few millimeters, with the thickest being measured at the retro-surface of the human patella and femoral trochlea (3,13,48).

The superficial zone has the highest collagen, water content, and chondrocyte density, but the lowest proteoglycan content among all three zones (2,49–51). The abundant collagen fibrils are aligned parallel to the articular surface and provide the superficial zone with substantial tensile strength in an orthotropic manner (37,38,52,53). The chondrocytes in this zone are flattened and are apparently polarized to be parallel to the surface (Fig. 2a) (7,54). Recently, intricate 3D images have been taken to

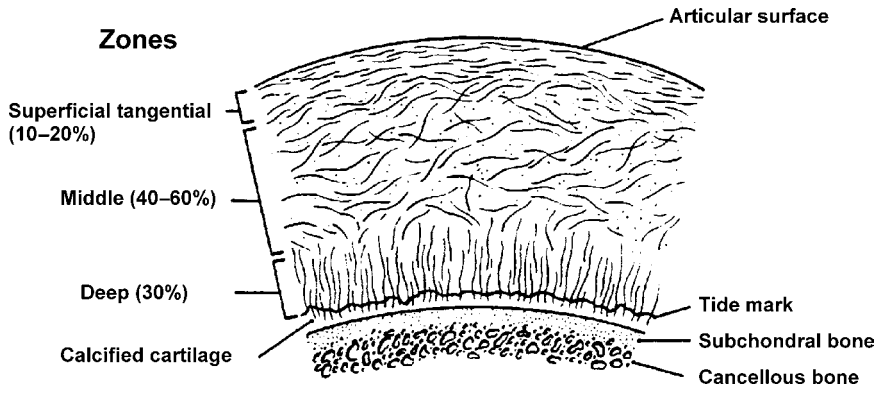


Figure 1. Layered structure of cartilage collagen network showing three distinct regions (3).

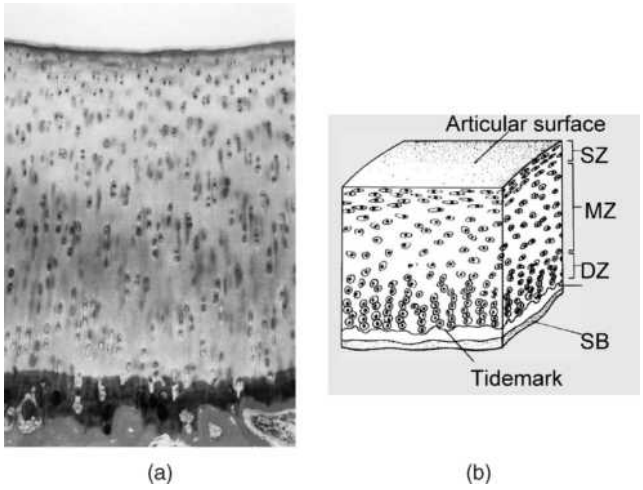


Figure 2. (a) Articular cartilage from a mature rabbit femur showing typical zonal arrangement of chondrocytes (polished saw-cut of resin-embedded tissues, surface-stained with basic fuchsine and toluidine blue) (54). (b) Schema of chondrocyte organization in the superficial zone (SZ), middle zone (MZ), and deep zone (DZ) (7).

view the discoid-shaped cells as they are maintained in this layer. Methods have ranged from digital volumetric imaging (55) to atomic force microscopy (56).

The intermediate, or middle, zone is generally the thickest amongst the three uncalcified zones of articular cartilage. Collagen fibrils, although less dense, have a greater diameter than the superficial zone, but appear to be more randomly oriented (47,57). The intermediate zone also has the highest proteoglycan content (3). Chondrocytes are more rounded, although cell density is not as high as in the superficial zone (3,58) (Fig. 2b).

The deep zone is relatively thin and the collagen are intertwined to form larger fiber bundles, and, from this zone, they insert perpendicularly into the calcified zone, and thus anchor the uncalcified tissue to the bony ends as required by joint articulation. This organization allows the bundles to firmly anchor the articular cartilage to the underlying subchondral bone. In general, chondrocyte density decreases from the middle zone to the deep zone, where they are similarly aligned as the collagen bundles, arranging into columns perpendicular to the uncalcified-calcified cartilage intersurface (3,55).

Several recent studies have investigated the pericellular matrix (PCM) and the interterritorial matrix (ITM) of chondrocytes. Using algorithms to account for fluid flow and differences in the relative stiffness between the PCM, the ITM, and the chondrocyte, different elastic moduli between PCM and ITM have been found to have a significant effect on chondrocyte's mechanical environment (59). Gradient distributions of charges and material densities relative to chondrocyte surface are important in cartilage fluid flow dynamics and deformation behavior (60). Using micropipette isolation of chondrocytes and nuclei, chondrocyte nuclei have been found to be stiffer than intact chondrocytes (59,61,62). Cultured chondrocytes are able to elaborate a PCM rich in Type VI collagen; however, intact chondron pellets accumulate significantly more

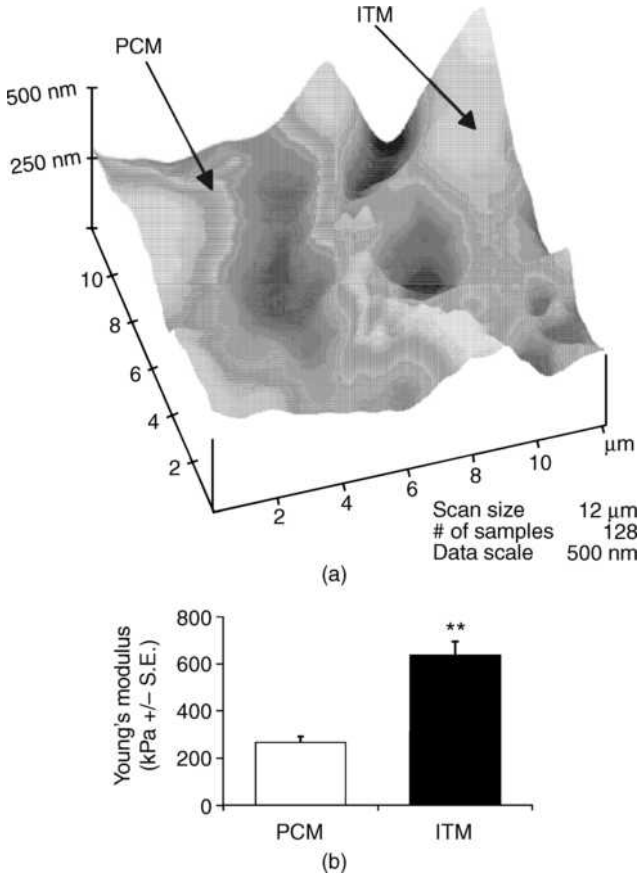


Figure 3. A representative height map of the PCM and ITM chondrocytes obtained through force mode of atomic force microscopy. Qualitatively, the ITM showed greater peak and valley contours than the topographic contour of the PCM (a). (b) presents the average Young's moduli of the PCM and ITM attained via nanoindentation. The average Young's modulus of the ITM (636.1 ± 124.91 kPa) was significantly greater than the PCM (265.1 ± 52.76 kPa) ($p < 0.01$) ($N = 19$) (64).

proteoglycans and Type II collagen than chondrocytes without a native PCM (63). Following a few weeks of accumulation of the ITM and PCM by isolated chondrocytes, a rapid increase in compressive stiffness occurs in both the chondron and the chondrocyte pellets (63). Using atomic force microscopy (AFM), the ITM is found to be stiffer to nanoindentation than the PCM (Figs. 3a and 3b) (64).

Meniscus and Fibrocartilage

The meniscus in the knee joint is a fibrocartilage. The two menisci (lateral and medial) in each knee joint are crescent or semi-lunar shaped and are attached to the joint capsule. The triangular cross section of the meniscus tapers radially inward from the periphery, and the center of the meniscus is thin and unattached. Thus, the cross section of the meniscus is wedge-shaped. The central region is a vascular and has more proteoglycans, hence more hyaline in appearance. The anterior and posterior horns of the meniscus form the tips of the crescents. The anterior horn of the lateral meniscus is attached to the tibia in front of the intercondylar eminence, partially blending with the

anterior cruciate ligament. The posterior horn is attached to the tibia near the intercondylar eminence as well as to the femur via the meniscofemoral ligament. The anterior and posterior horns of the medial meniscus are attached to the tibia near their respective intercondylar fossae. The anterior horns of the lateral and medial menisci are connected by the transverse ligament. The thick peripheral borders and associated horns of the meniscus are vascularized by blood supply predominantly from the genicular arteries surrounding the joint. The thinner central portions of the meniscus are aneural and avascular, a region very much like hyaline cartilage (20). The meniscus is lubricated with synovial fluid (65), probably by the same lubrication mechanisms known to exist in articular cartilage (66). Fibrocartilage is found in a small number of other joints. The disk of the temporomandibular joint (TMJ) and the intervertebral disks are both composed of fibrocartilage, although they are drastically different structures with different distributions of cartilage and PGs and ultrastructual organization.

The fibrocartilagenous structure of the meniscus differs from that of hyaline cartilage in many ways. The cells of the meniscus are sometimes called fibrochondrocytes, although it is probable that some cells are more like fibroblasts, whereas others are more like chondrocytes (65,67). The peripheral two-thirds of the meniscus are primarily composed of a randomly oriented mesh-like, coarse, collagen fibrillar matrix (68–71). In deeper portions, large rope-like collagen fiber bundles are arranged circumferentially, retaining the overall semi-lunar shape of the meniscus and providing tensile strength. Smaller fibers are also found radially and connect to the larger circumferential collagen fiber bundles (72). As mentioned above, the inner portion of the meniscus resembles that of hyaline cartilage, containing a higher percentage of proteoglycans enmeshed within a randomly arranged collagen fibrillar matrix (71,73,74).

The function of the meniscus is to enhance higher congruity of the articulating surfaces of the distal femur and proximal tibia, to accommodate the range of motion, in addition to the same functions of load bearing and load distribution to that of articular cartilage (20). The previous assumption that the menisci are functionless, evolutionary remains of leg muscles is erroneous and that menisectomy (a common clinical procedure) is indeed a common procedure in animal models to study the etiology of osteoarthritis (75–78).

CARTILAGE AND MENISCUS: MECHANICAL PROPERTIES

Articular cartilage and meniscus are both important load-bearing tissues and vital to the maintenance of normal joint functions (3,18,19). Articular cartilage can absorb mechanical shock of joint motion and spread the applied load onto the subchondral bone. It also contributes to the lubrication mechanism and provides a surface with low friction, enabling repetitive gliding motion between articulating surfaces (7,66). The meniscus of the knee has important biomechanical functions such as load transmission at the otherwise highly incongruent tibiofemoral articulation, shock absorption, joint congruity, and stability (18,19). The salient biomechanical functions of articular cartilage and meniscus are dependent on their biological structure, composition, and the intrinsic material properties of the ECM. The knowledge of their material properties such as tensile, compressive, and shear moduli is essential to understand not only their biomechanical functions, but also in the tissue engineering of articular cartilage and meniscus to produce in a biomimetic manner artificial-biological replacements (17,79).

Tensile Properties

When cartilage is tensed, the tensile stress-strain behavior is nonlinear. A typical nonlinear stress-strain (σ-ϵ) curve for cartilage, meniscus, and other soft tissues is depicted in Fig. 4 (3). For small deformations, a 'toe-region' is seen in the stress-strain curve, in which the collagen fibrils will primarily realign in the direction of the externally applied force instead of being stretched (elongation per unit length). For larger deformations, the collagen fibrils are stretched, and a larger tensile stress is generated within the collagen fibers (e.g., 35–38,52,53,80–83). In this linear region, the stress is proportional to the applied strain, and their ratio is known as Young's modulus in tension, or tensile modulus. This tensile modulus is a measure of the stiffness of the collagen-PG solid matrix and is primarily dependent on the density of collagen fibrils, fibril diameter, and type or amount of collagen cross-linking (3,18,19,52). Beyond the linear region, the cartilage strip will rupture abruptly, and the tensile failure stress is a measure of the strength of the collagen fibrillar network.

In general, the tensile modulus of articular cartilage will be in a range of 1–30 MPa, which is much larger than the compressive modulus of cartilage ($\sim$ 0.5 MPa), which is known as tension-compression nonlinear property of the

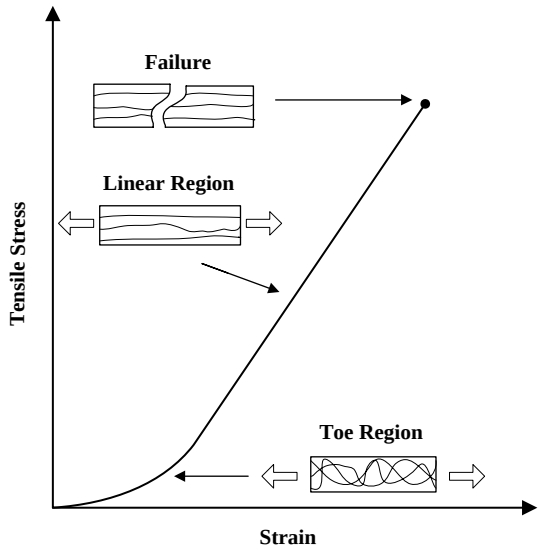


Figure 4. Typical stress-strain curve for articular cartilage and meniscus in a uniaxial and uniform strain rate experiment. The toe region is marked by an increasing slope, whereas the linear region appears to be a straight line (3).

cartilage (84–86). The tensile properties are also known to vary with location, depth, and orientation of test specimens of cartilage and meniscus. Hultkrantz (45) demonstrated the anisotropic organization of the collagen network by puncturing holes in the surface of articular cartilage with a round pin. He found that round puncture holes will form elongated splits, analogous to splits formed in lumber when a large round awl pierces it. The split-line patterns were, to him, evidence of collagen fiber orientation, which is still an enigma today because electron microscopy has not found such surface collagen anisotropy. (Nevertheless, the pattern of split lines is similar to Langer lines formed in the skin in a similar manner.) Much later, Woo et al. (38), Kempson et al. (52), and Roth and Mow (37) showed that the tensile strength and stiffness of the samples cut parallel to the split-line direction were higher than those cut perpendicular to it. The cartilage strips from high weight-bearing areas of human knee joints exhibit larger tensile modulus than those from low weight-bearing areas (53,82) because high weight-bearing areas generally have a relatively higher proteoglycan content. The adult human femoral articular cartilage exhibits a gradual decrease in tensile strength and stiffness as the distance from the articular surface increases (36,81), while this functional dependence was not observed for young bovine humeral joints (38). A dependency of cartilage tensile properties with skeletal maturation was found by Roth and Mow (37). These investigators found that, with the closing of the growth plate (indicative of skeletal maturity), the strength and stiffness of cartilage are much less that those properties of immature cartilage (open-physis). The effects of age on the tensile properties of adult cartilage were extensively studied by Kempson (81), and the results showed that the tensile modulus decreases with age, and that the modulus of the hip cartilage decreases more markedly than that of ankle cartilage. This finding may explain the relatively high occurrence of osteoarthritis in the hip compared with the ankle. Like articular cartilage, the tensile properties of meniscus vary with respect to the location (anterior, central, and posterior) and specimen orientation relative to the predominant collagen fiber direction (circumferential and radial) (18,19,74). Specimens from the posterior half of the medial meniscus have been shown to be significantly less stiff and less strong in tension than specimens from all other regions (87). This experimental result agrees with the ultrastructural findings using polarized light; In the posterior half of the medial meniscus, collagen fiber bundles have significantly reduced circumferential organization (87).

Numerous experiments have shown that the tensile modulus is correlated with the collagen content or the ratio of collagen content to proteoglycan content in articular cartilage (e.g., 82). The tensile modulus of articular cartilage decreases to only 1% after disruption of collagen cross-linking by elastase (88). In contrast, no significant correlations have been found between the tensile property of the cartilage and proteoglycan content (36,89). These findings indicate that collagen content, organization, and cross-linking play significant roles in generating high tensile modulus of articular cartilage. For meniscus, although the tensile properties show significant regional and directional variations, little difference appeared in the biochemical composition with site, and no significant correlation exists between tensile property and chemical contents (74). The variation of tensile properties seems to reflect local differences in collagen ultrastructure and fiber bundle direction as described above.

Compressive Properties

The compressive behavior of cartilage and meniscus has been extensively studied under various configurations, such as confined compression, unconfined compression, and indentation (see Fig. 5). Most of the earliest studies (e.g., 90–94) used the indentation technique to determine the mechanical property of articular cartilage and modeled the cartilage to be a single-phase, elastic body with the assumption of the Poisson's ratio ranging between 0.4 and 0.5 (e.g., 91–96). However, this single-phase elastic model cannot describe the time-dependent viscoelastic behavior of the tissue nor the role played by cartilage's major component (i.e., water). Cartilage and meniscus exhibit a viscoelastic creep in response to a constant load (i.e., its deformation will increase with time). Conversely, if a constant displacement is applied, the force response will decrease gradually with time to a constant value (i.e., a stress-relaxation will be observed).

These viscoelastic behaviors derive from the friction of water flowing through solid matrix (27,97), as well as the flow-independent intrinsic energy dissipation inside the macromolecular solid matrix during mechanical loading (98–101). As mentioned, articular cartilage and meniscus can be regarded as biphasic materials: a fluid phase composed of water and electrolytes, and a solid phase mainly composed of collagen and proteoglycans (27). The solid matrix is considered as being porous and permeable. Water resides in the microscopic pores and flows through the matrix during joint loading. Under a slow ramp loading, the observed viscoelastic behaviors are usually dominated by the large drag forces generated by the flow of interstitial fluid through the porous-permeable solid matrix, and therefore, the flow-independent intrinsic energy dissipation is negligible. However, osteoarthritic cartilage has higher permeability and lower ECM stiffness; in such tissues, the intrinsic viscoelastic behavior becomes the dominating component governing their mechanical behaviors.

The transient behavior of the tissue under compression is primarily determined by the mechanism of fluid pressurization because of high friction between solid and fluid phases, which is also known as flow-dependent viscoelastic behavior (27). Figure 6 (3) shows the stress-relaxation behavior of a tissue specimen under confined compression. In this experiment, before time t_0, the tissue is compressed with a constant rate, and the interstitial fluid inside the tissue will be pushed out through upper porous platen. As a result of the distributive fluid drag force, a larger deformation can be seen at the downstream side (Fig. 6a and 6b). During the relaxation phase (after t_0), no fluid exudation occurs, but the fluid needs to redistribute inside the tissue before the equilibrium is reached (Figs. 6c, 6d, and 6e). Although the velocity of fluid flow is very low, the friction force, or the drag force, could be very large because the pore

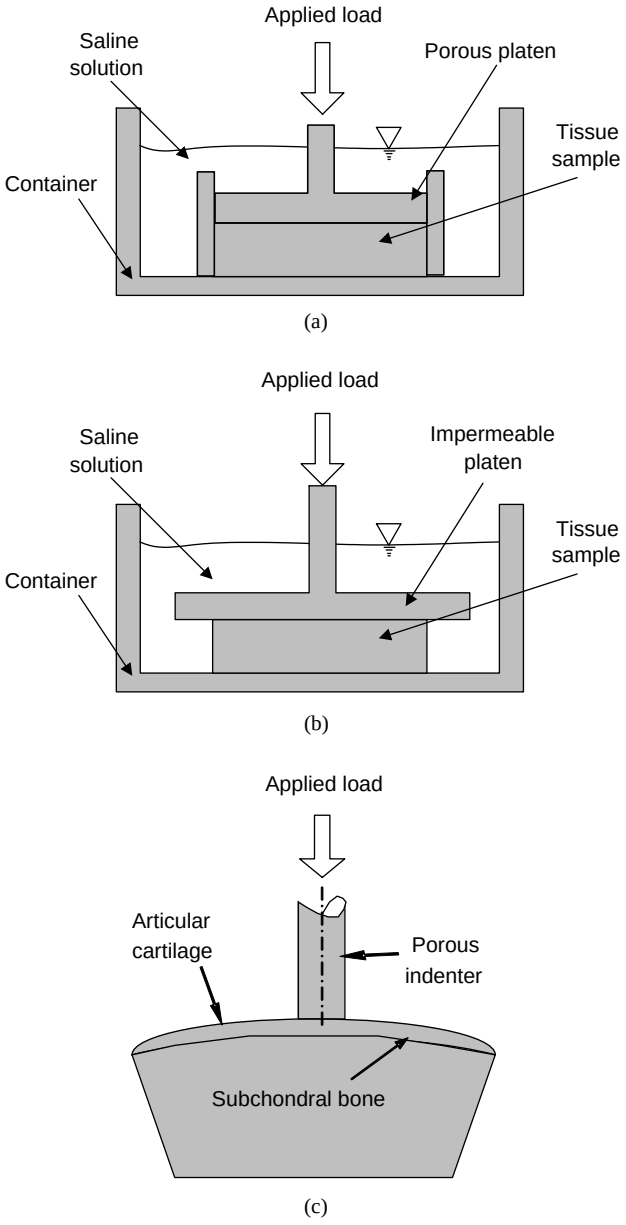


Figure 5. Schema of three configurations frequently used to study the compressive properties of articular cartilage. (a) In the confined compression configuration, a load is applied to the cartilage sample via a rigid porous permeable platen. The side walls are assumed to smooth, impermeable, and rigid, thereby preventing lateral expansion and fluid flow. (b) In the unconfined compression configuration, the cartilage sample is compressed between two rigid, smooth, and impermeable platens. The lateral side allows fluid flow. (c) In the indentation configuration, the cartilage is compressed via a rigid porous permeable indenter. The porous indenter allows the fluid exudation to occur freely into the indenter tip and, therefore, creep of the cartilage layer.

size inside the tissue is very small ($\sim$ 50–65 nm for articular cartilage), and the permeability of the tissue is as low as 10^{-15} N·s/m^4. Therefore, the generated fluid pressure can be remarkably high inside the tissue during the transient state, which also means that chondrocytes encased within the ECM will normally be bathed in a highly

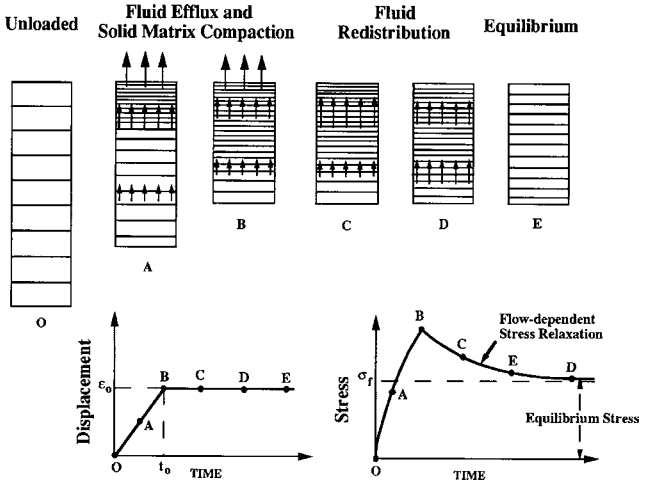


Figure 6. A schematic representation of fluid exudation and redistribution within cartilage during a rate-controlled confined compression stress-relaxation test (lower left). The horizontal bars in the upper figures. indicate the distribution of strain in the tissue. The lower graph (right) shows the stress response during the compression phase (O, A, B) and relaxation phase (B, C, D, E) (3).

pressurized fluid. It has been estimated that this fluid pressure could be 30 times more than the elastic stress generated in the solid matrix of articular cartilage (3). Considering that the equilibration process usually takes several hours, no real equilibrium state occurs in joints under physiological conditions because the joints are moving virtually at all times, even during sleep. Thus, the mechanism for fluid pressurization is likely to be the major physiological load-supporting mechanism in diarthroidal joints, and it plays an important role in shielding the solid matrix from large compressive stresses during the joint function (7).

At equilibrium, the fluid flow stops, no fluid pressure gradient exists inside the tissue, and the applied load is entirely supported by the solid matrix of the tissue. Thus, the compressive property can be obtained from the relations between stress and strain. It has been found that the equilibrium strain is proportional to the applied load. Typically, the equilibrium aggregate modulus (27) for normal articular cartilage ranges from 0.4 to 1.5 MPa, whereas the average equilibrium aggregate modulus for the meniscus is about 0.4 MPa. Table 1 shows the equilibrium aggregate moduli of lateral condyle and patellar groove cartilage and meniscus, showing considerable variation among the species and tissue location (3).

Tissue mechanical properties are highly dependent on their composition and structure. It has been shown that the equilibrium aggregate modulus for human articular cartilage correlates in an inverse manner with water content and in a direct manner with PG content (27,102,103). The highly loaded regions of articular cartilage generally have larger compressive modulus and greater PG content (53,104,105). In contrast, no correlation is found between the compressive stiffness and collagen content. Removal of PGs from articular cartilage samples dramatically decreases the compressive modulus, whereas trypsin

Table 1. Equilibrium Aggregate Modulus of Lateral Condyle, Patellar Groove Cartilage and Meniscus (MPa) (3)

	Human[a]	Bovine[b]	Canine[c]	Monkey[d]	Rabbit[e]
Lateral condyle	0.70	0.89	0.60	0.78	0.54
Patellar groove	0.53	0.47	0.55	0.52	0.51
Meniscus	NA	0.41	NA	NA	NA

[a]Young normal.
[b]18 months to 2 years old.
[c]Mature beagles and greyhounds.
[d]Mature cynomologus monkeys.
[e]Mature New Zealand white rabbits.
[f]Not available.

digestion of collagen fibrils has little effect on compressive modulus (80,106).

The biphasic theory has been the most successful model for the compressive viscoelastic behaviors of cartilage and meniscus under various conditions (27). This theory assumes that (1) the solid matrix and interstitial fluid are immiscible and incompressible; (2) viscous dissipation is due to the fluid flow between water and the porous-permeable solid matrix; and (3) the frictional drag is proportional to the relative velocity and can be affected by ECM compression. This biphasic theory further assumes that the solid matrix experiences infinitesimal strain and that the stress-strain relations can be described by the generalized Hooke's law. Despite its simplification, as biological models typically are, the isotropic form of the linear biphasic theory has been shown to provide an accurate description of the compressive creep and stress relaxation behavior of these tissues. In particular, a numerical algorithm based on this biphasic theory was developed and accurately predicted the aggregate modulus, Poisson's ratio, and permeability of articular cartilage from the indentation creep experiment (13,100,107). The biphasic theory has also been extended by employing higher levels of tissue complexities, including material inhomogeneities (108–110), material symmetries (33,34,85,86,111), and matrix viscoelasticities (33,34,84,98–100).

Shear Properties

The intrinsic viscoelastic properties of the solid matrix of cartilage and meniscus can only be determined in a *pure* shear experiment and under small strain conditions. In pure shear, the kinematics of deformation does not permit volumetric change, and hence, no interstitial fluid flow is possible when no pressure gradients are applied. Under these three conditions, the tissue deforms without change in volume, and therefore, the interstitial fluid pressure and fluid flow are minimal. As a result, the flow-dependent viscoelastic properties are excluded, and the measured physical parameters will be independent on the friction or drag force between fluid phase and solid phase, which often occurs in compressive configurations, thus directly reflecting the intrinsic viscoelastic property of solid matrix. This flow-independent viscoelastic behavior of the collagen-PG matrix derives from the internal friction between collagen and PG molecules (3,101).

The first shear properties measurement was reported by Hayes and Mockros (112), and later, nonlinear viscoelastic and fatigue properties of bovine articular cartilage were investigated (113,114). However, all these tests were performed in a simple shear configuration, and dynamic shear properties of these studies were reported at frequencies (e.g., 20–1000 Hz) much higher than the physiological range (e.g., 1 Hz). Pure shear tests of articular cartilage and meniscus have been performed under transient, equilibrium, and dynamic conditions to characterize the intrinsic or flow-independent viscoelastic behavior (e.g., (101,115–117)). When a circular cartilage specimen is subject to a sudden change of angular displacement, the shear stress will increase instantaneously, followed by a rapid decay before equilibrium is reached. The quasilinear viscoelastic theory (118) has been shown to provide an excellent description of this intrinsic stress-relaxation behavior of normal human patellar cartilage (116). The equilibrium shear modulus for normal human, bovine, and canine articular cartilage has been found to vary in a range of 0.05–0.25 MPa. Values for the magnitude of the dynamic shear modulus $|G^*|$ of normal cartilage are in the range of 0.2–20 MPa and vary with both the frequency and magnitude of the normal stress. The phase shift angle (δ) for cartilage lies between 9° and 20° over a frequency range of 0.01 Hz to 20 Hz (101). Please note that δ is a measure of matrix dissipation, with a loss angle of 0° corresponding to a perfectly elastic material, and 90° to a perfectly dissipative material. The viscoelasticity of meniscus in response to shear is qualitatively similar to that exhibited by articular cartilage, although the magnitudes of the material coefficients of these tissues are significantly different. Meniscus shear properties exhibit an orthotropic symmetry (i.e., the three planes of symmetry defined by its fibrous architecture dominate the shear properties of the meniscus). The equilibrium shear moduli are 36.8 kPa, 29.8 kPa, and 21.4 kPa in the circumferential, axial, and radial directions, respectively (18,19,119). These shear modulus values are ten times less than those observed for articular cartilage. For dynamic tests, the magnitude of the complex shear modulus $|G^*|$ and phase shift angle δ for circumferential, axial, and radial specimens reflect orthotropic collagen fiber organizational symmetry as well (Fig. 7) (3).

The collagen network plays an active mechanical role in contributing to the shear stiffness and energy storage in cartilage (101). Conceptually, the role played by collagen when the specimen is in shear may be visualized as shown in Fig. 8 (101). The tension in the diagonally oriented collagen acts to increase the shear stiffness of the solid matrix. This effect is confirmed by the experimental result that $|G^*|$ is directly and significantly related to the collagen content of articular cartilage and also by the fact that

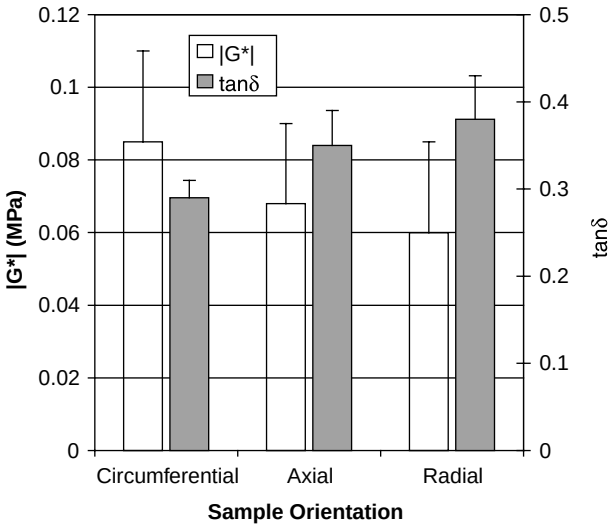


Figure 7. The magnitude of dynamic shear modulus $|G^*|$ and tan δ for meniscal specimens with normal vectors oriented circumferentially, axially, and radially at 1 rad/s (3).

cartilage has a relatively small loss angle and large shear modulus (101) compared with that of PG solutions at physiological concentrations (107,120,121). The depletion of the PG content has been shown to decrease the dynamic shear modulus up to 55% (101,122), which is considered as a result of the decrease of tensile stress inside the collagen fibrils due to the decrease of PG swelling pressure.

Swelling Properties

Swelling in articular cartilage derives from the presence of negatively charged groups (SO_3^- and COO^-) along the GAG chains of PG molecules. For normal and degenerative femoral head cartilage, the fixed-charge density (FCD) ranges from 0.04 to 0.18 mEq/g wet tissue at physiological pH (2,123). These fixed charges will require a high concentration of counter-ions (Na^+) to maintain electroneutrality, and the concentration, along with that of co-ions (Cl^-), is governed by the Donnan equilibrium ion distribution law (124). This excess of freely mobile ions will introduce an imbalance of ion concentration between the fluid compartment inside the tissue and the bathing fluid outside the tissue, giving rise to a higher pressure in the interstitial fluid than the ambient pressure in the external bath, known as the Donnan osmotic pressure. This osmotic pressure decreases with FCD and has a range of 0.1 to 0.25 MPa for normal articular cartilage. With the increase of the external saline concentration, the osmotic pressure will decrease. For the very large external saline concentration (e.g., 2 M), the osmotic pressure is considered to be extremely small, or zero.

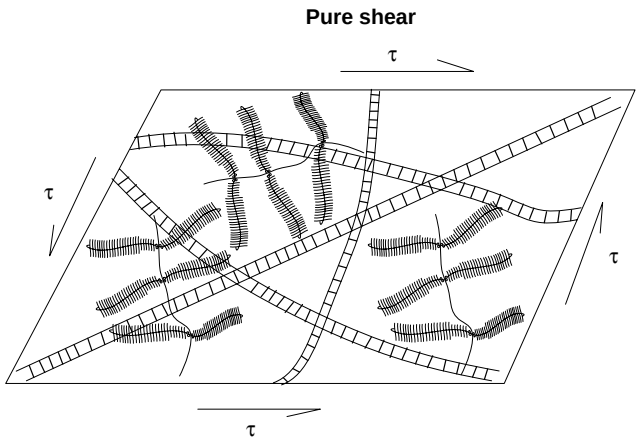


Figure 8. A scheme representation of cartilage in pure shear. The tensile stress inside collagen fibrils provides shear stiffness (101).

This osmotic pressure causes the tissue to swell, as measured by both weight change (2,123) and dimensional change (125,126). These latter experimental results show that the swelling of articular cartilage is inhomogeneous and anisotropic. The swelling ability increases with depth, with the largest dimensional change for the deep zone and almost no change for the superficial layer. The magnitude of swelling is the largest in the thickness direction and the smallest in the split-line direction. In articular cartilage, the osmotic pressure is restrained by the surrounding collagen network. Therefore, residual stress or pre-stress exists inside the solid matrix even before external load is applied. Articular cartilage will warp or curl toward its articular surface upon its removal from the subchondral bone, and the curvature or the extent of curling will decrease with the increase of external saline concentration (126). It has been hypothesized that this curling is caused by the combination effects of swelling pressure and inhomogeneity inside the tissue (126–129). Recently, a three-layer orthotropic model based on triphasic theory (39) has been developed to describe the curling behavior of cartilage strip by considering its layered structure that includes depth-dependent collagen fibril orientation and chemical content distributions (128). The predicted curvature change with external saline concentration agrees well with previously published experimental results. This model has also suggested that the large stiffness of the superficial layer and high swelling pressures play key functional roles in the development of pre-stress in cartilage and in its curling behavior.

Quantification of morphological changes has been extensively used to study changes in cartilage swelling with osteoarthritis (OA) (44). With OA, compositional and microstructural changes will occur, which includes the fibrillation of the superficial zone of articular cartilage, the decrease of the PG concentration, and the imbibitions of water; these items are the earliest indicators of the OA degeneration of cartilage (18,19,29,102). The elevated water content or swelling has been shown to be very sensitive to collagen fibrillation. Experimental results also suggest that the water content of the tissue increases after digestion with collagenase (82,83,101). Physically, collagen fibrillation decreases the stiffness of the solid matrix, specifically the elastic bulk modulus, which allows the tissue to imbibe more water (52,104).

Triphasic Mixture Theory

To account for the swelling behavior, Donnan osmotic effects, ion transport, and electrical potentials inside the

tissue, Lai et al. (39) developed the triphasic theory to incorporate the effects of negatively charged groups on the PGs of solid matrix. In this theory, the electrolytes (mainly Na^+ and Cl^-) within the interstitial fluid are considered as a separate phase, and the solid phase is charged. This triphasic theory was further extended to account for multiple species of ions in the tissue (130). Note that Huyghe and Janssen (131) developed an equivalent theory, named the quadriphasic theory, in which ion species Na^+ and Cl^- were treated as two separate phases.

Using the triphasic theory, the electrokinetic coefficients such as electrical conductivity has also been derived in terms of the physical parameters of charged tissues (39,130,132–134). Furthermore, a theoretical analysis (134,135) showed that the electrical potential inside the tissue comes from two competing sources: a diffusion potential deriving from the FCD inhomogeneity and a streaming potential resulting from the fluid flow within a charged material. These two sources of electrical potential have different polarity and compete against each other. Within the physiological range of material properties of articular cartilage, the polarity of electrical potential inside the tissue depends on the stiffness of the tissue. For softer tissues (such as OA tissue), the diffusion potential tends to dominate, whereas the streaming potential tends to dominate for stiffer tissues (such as normal tissue).

Numerical methods, such as finite difference and finite element formulations, have been developed to demonstrate the contributions of the FCD in mechano-electrochemical (MEC) behaviors of charged, hydrated soft tissues (43,134,136,137). These studies showed that that higher FCD decreases the characteristic time (gel time) and causes the tissue to reach equilibrium in a shorter amount of time, and showed that the osmotic effects can contribute up to 50% of the equilibrium confined compression stiffness (136) and about 30% in unconfined compression (43). With the finite element formation, Lu et al. (138) successfully correlated the predicted FCD with the biochemical measurements, while simultaneously measuring the apparent mechanical properties from the indentation creep experiment.

Recently, the triphasic formulations have been linearized, and the analytic solutions for the MEC response of the tissue under unconfined compression have been obtained both for transient state and at equilibrium (139). With a regular perturbation, simple relations have been derived to describe how the apparent properties, such as Young's modulus and Poisson's ratio, change with the fixed negative-charge density (FCD) at equilibrium (139,140). These relations actually are applicable to various testing configurations, even for steady permeation, and they indicate the correspondence of mechanical properties between an elastic body and a charged triphasic material such as articular cartilage and meniscus (139–142).

CARTILAGE WEAR AND DEGENERATION

Cartilage wear and degeneration have been extensively studied due to their significant roles in physically debilitating diseases such as osteoarthritis and rheumatoid arthritis (1). Two types of wear occur in synovial joints: fatigue wear and interfacial wear (66). Fatigue wear is independent of the lubrication within the joint and is caused by functional activities such as cyclic, repetitive loading. A balance is presumably maintained under the normal physiological condition whereby tissue turnover is maintained by cells in various components of the synovial joint. A number of factors may contribute to cartilage wear and degeneration. Collagen fibers can be severed by excessive functional activities, leading to a compromise in tensile strength. The normally tight collagen fiber bundles can be unwound and loosened (47,143). When inflammatory cytokines are released, proteoglycans are lost rapidly, leading to a breakdown of the ECM (144–146). Fibrillated cartilage from osteoarthritic patients shows an increase of apoptotic chondrocytes deeper than in the normal articular cartilage, which generally has apoptotic cells only near the surface (145,147). Collective loss of chondrocyte and ECM may lead to microcracks and fissures, which may further grow with functional loading. Thus, fatigue wear of cartilage is a mechanism dependent on biological synthesis and mechanical loading (5).

Interfacial wear can result from physical contact loading of articulating surfaces. Interfacial wear has been categorized into two classes: adhesive and abrasive wear (66). Adhesive wear is more common and occurs when a junction is created when the two solids are in contact. As the opposing surfaces continue to move past the junction, fragments from the weaker surface may be torn off and adhere to the stronger material. The concept is analogous to rubber skid marks left on a road from a braking car. The car is able to move past the formed junction only by having elements of the weaker material, the rubber tire, come off and adhere to the stronger material, which is, in this case, the road. Abrasive wear occurs when a harder material comes into contact with a softer material. No junction is formed. While in contact and rubbing against each other, the harder material cuts or plows into the softer material. The harder material can be either one of the opposing surfaces or loose particles caught between two softer opposing surfaces, cutting into both of them (66).

Pain, stiffness, swelling, and reduced range of motion are the common phenotypic characteristics of osteoarthrosis. Typically, clinical diagnosis is only made after significant cartilage deterioration. A number of methods have been formed to monitor the development of osteoarthrosis. Radiography has been the conventional method for both diagnosis and monitoring. The topographical variation in degenerative cartilage has been used in designing an arthroscopic indentation instrument in which osteoarthrosis could be diagnosed *in vivo* and possibly treated before the diminishing qualities of the disease (148). Other manners that have been proposed for detection are chondrocalcin measurement (149) and knee wear particle analysis, which is derived from the concept of abrasive wear (150).

CURRENT CARTILAGE REPAIR STRATEGIES

Cartilage's poor capacity for self-regeneration is well known. The poor regenerative capacity of cartilage has

been contrasted to bone, because bone readily regenerates (unless it is a critical size bone defect) and has a relatively rich blood supply (151). A lack of angiogenesis has been cited as the primary cause for cartilage's poor capacity for regeneration. However, normal cartilage is avascular. Vascular supply to cartilage likely will turn it into bone. Thus, a lack of vascularization is not the direct cause for cartilage's poor regenerative capacity (152).

The regenerative ability of cartilage in response to injury also depends on factors such as joint loading, the degree of injury, the location of injury, and whether it is cartilage lesion alone or osteochondral lesion (153,154). Cartilage responds differently to slowly or rapidly applied loads. For example, loading causes fluid movement in the matrix and may serve to counteract the deformation and to distribute the loads throughout the tissue. However, rapid compressive loading may not allow the fluid to infuse matrix, thus transferring excessive loading to the cells and the ECM macromolecules. Should excessive force be sustained, the chondrocytes and ECM molecules may undergo rupture or degradation. Another factor in determining the cartilage's regenerative capability is the chondrocytes' intrinsic ability to replenish the supply of matrix molecules, as well as the approaches to remove degraded materials (17,153,155).

Articular cartilage injuries can be classified as follows: (1) cartilage matrix and cell injuries without substantial tissue defects; (2) defects, fissures, or ruptures in articular cartilage only; and (3) osteochondral lesions. The first category of cartilage injuries without substantial tissue defects, nonetheless, is associated with a decreased concentration of matrix macromolecules such as PGs and collagen. Albeit without tissue-level defects, loss of PGs and collagen results in a decrease in mechanical strength. Unless repaired, even the first category of cartilage injuries can lead to more substantial defects as described in the second and third categories (153,155).

The second category of articular cartilage injuries is localized within cartilage. They include focused mechanical disruption of the matrix including fissures, tears, incisions, or interruption of the integrity of articular surface. Chondrocytes not only attempt to replace the loss of matrix macromolecules but also proliferate to fill the voids created during injury (153,156). However, the rates of chondrocyte proliferation and matrix synthesis may not be sufficiently high to match the rate of cartilage degradation. As articular cartilage has no nerve supply, except at the very periphery, even substantial cartilage lesions may not elicit pain. A few types of synovial joint injuries likely exist that elicit pain, such as osteochondral injuries, synovial membrane injuries, and injuries to the periphery of articular cartilage. These injuries are usually not repaired by the articular chondrocytes (4).

The third category is osteochondral injuries that involve both articular cartilage and subchondral bone and elicit inflammatory responses such as an influx of blood-borne cells, platelets, and cytokines. Hemorrhaging or fibrin clot formation may occur and later develop into a fibrous mass. The influx of cytokines may induce migration of progenitor cells, although no guarantee exists that these progenitor cells, likely mesenchymal stem cells that are capable of differentiating into all connective tissue lineage cells, will differentiate into chondrocytes. In fact, osteochondral lesions are likely repaired, if reparable, by fibrocartilage or fibrous tissue instead of hyaline cartilage (154), and rarely possess the complex zonal structures of native articular cartilage (153,157). The mechanical strength of fibrocartilage is approximately one-third of the strength of native hyaline articular cartilage, and thus may not be able to fulfill the weight-bearing and load-bearing functions of normal articular cartilage. Over time, osteochondral lesions may undergo further degradation, leading to the exposure of subchondral bone, which results in osteoarthrosis and can lead to joint immobility.

Current clinical treatments for articular cartilage injuries have several deficiencies. Depending on the degree of injury and whether the defect is partial- or full-thickness, the treatments generally involve surgical irrigation, debridement, and tissue augmentation. Partial thickness injuries of the articular cartilage involving clefts and fissures, often in the early stages of osteoarthrosis, are most commonly treated with arthroscopic surgery such as lavage or debridement. Arthroscopic lavage involves the irrigation of the joint, whereas debridement is the arthroscopic removal of damaged tissue. By performing these treatments either alone or in combination, a decrease in joint pain usually results. However, lavage or debridement treatments rarely induce the repair process of cartilage (155,157).

Full-thickness injuries refer to lesions in both articular cartilage and subchondral bone. Although a large number of treatments are empirical, several procedures have taken the advantage to simulate the native repair process. Arthroscopic treatments such as abrasion arthroplasty, Pridie drilling, and microfracture are commonly used, and all include the further perforation of the subchondral bone to induce bleeding and further fibrous tissue formation. Abrasion arthroplasty and microfracture are used in conjunction with debridement to reduce the amount of damaged tissue within the joint. The outcome of these treatments is variable, largely because the healing and repair process within the articular surface are somewhat unpredictable. Furthermore, factors such as the patient's age, postoperative activity level, and overall heath also affect the outcome (158).

Recently, soft tissue grafts such as the transplantation of the periosteum or perichondrium have been used clinically to repair the articular surface for cylindrical, full-thickness defects. The rationale for using periosteum is its observed chondrogenic potential during development and fracture repair (159). The periosteum consists of a fibrous and cambial layer. The cambial layer contains precursor cells that are capable of differentiating into osteoblasts and perhaps chondrocytes. The process of periosteum transplantation involves the creation of a defect spanning the full thickness of articular cartilage and penetrating the subchondral bone, and then placement of the periosteum graft within the defect. However, much debate has occured as to which layer of the periosteum should lay adjacent to the bone and which layer should face the articular surface, as the cambial layer can form cartilage, whereas the fibrous layer forms fibrous tissue. Larger full-thickness defects are generally repaired using allogenic or autogenic

osteochondral tissue plugs, called mosaicplasty, excised from nonload-bearing regions of the joint and inserted into the full-thickness defect. Reports exist of fibrous tissue formation and chondrocyte death at the interface between the plug and surrounding tissue, which may lead to further degeneration of the joint. Furthermore, donor site morbidity remains a drawback of the periosteum graft or mosaicplasty (160).

For substantial osteochondral lesions, total joint replacement using metallic condyle and plastic socket is most commonly in practice. Current modalities of total joint replacements suffer from drawbacks such as donor site morbidity, pathogen transmission, wear and tear, and a limited life span (161). Secondary surgeries are necessary in 10–15% of the cases and suffer from substantial difficulties such as scar tissue formation and loss of host tissue (162). More importantly, current total joint replacement therapies fail to yield biological regeneration.

CURRENT MENISCAL REPAIR STRATEGIES

Prior to the recognition of the importance of the meniscus in the biomechanics of the knee joint in the 1970s, the preferred treatment for meniscal injury such as tear was total excision of the meniscus or open meniscectomy. Despite some reports of temporary relief of symptoms, the long-term outcome of excision or meniscectomy was poor. In the 1980s, understanding of the material properties and biomechanical roles of the meniscus led to more conservative treatments for meniscal tears. Furthermore, the development of arthroscopy enabled more accurate diagnosis of meniscal tears and, subsequently, more precise surgical treatment that has substantially reduced the amount of damage to the surrounding tissue in comparison with open-joint surgery (163).

Partial-thickness split tears and small (< 5 mm) full-thickness split tears, vertically or obliquely, are usually left alone and without surgical intervention. The inner wall of the meniscus must be stable during probing, which is commonly performed arthroscopically during diagnosis. Follow-up arthroscopic examinations are often necessary to monitor tissue healing. These injuries are usually associated with ligament tears such as the anterior cruciate ligament (ACL). Ligament tears in conjunction with meniscal tears drastically reduce the stability of the knee and can lead to further meniscal damage.

Meniscal injuries that require surgical repair or excision are large defects with compromised vascular supply, large meniscal deformations, or damage to the peripheral or circumferential collagen fibers. In need of excision, it is preferable to leave a partial meniscus intact as opposed to full meniscectomy. The surgical approaches are either open-joint surgeries or arthroscopic surgery to suture tears within the meniscus (163).

TISSUE ENGINEERING OF ARTICULAR CARTILAGE AND MENISCUS

The rapidly evolving field of tissue engineering has promised to deliver biological replacements of damaged articular cartilage and meniscus. In comparison with current treatment modalities, tissue engineering represents a shift in the paradigm. Whereas current treatment modalities improve articular cartilage and meniscal injuries by increments, the end goal of tissue engineering is to generate or regenerate articular cartilage and meniscus.

Previous investigations of the structural, biochemical, and mechanical properties of articular cartilage and meniscus serve as the necessary foundation and have set the stage for the tissue engineering of these structures. For example, a commonly stated long-term goal of a biochemical study to investigate the PG distribution in various zones of articular cartilage was to improve the treatment of cartilage defects in arthritis, which was all too familiar for a published study or a grant proposal decades ago.

To tissue-engineer articular cartilage or meniscus, one has the conceptual liberty of selecting cell sources, scaffolds, or growth factors. Another essential choice is whether mechanical stress is to be applied to the tissue-engineered articular cartilage or meniscus prior to or after *in vivo* implantation. Thus, the initial stage of tissue engineering of articular cartilage and meniscus is an optimization process of cells, scaffolds, growth factors, or mechanical stimulus.

Cells Capable of Generating Articular Cartilage and Meniscus

Articular chondrocytes are the obvious choices for articular cartilage regeneration (164–166). From the standpoint of scientific discovery, articular cartilage or meniscal regeneration from articular chondrocytes or meniscal fibrochondrocytes has revealed a wealth of information (e.g., 167–169). From the standpoint of alignment with eventual therapeutic regeneration of synovial joint condyle in arthritis patients, the selection of articular chondrocytes is problematic. The essential problem at this time is that articular chondrocytes are not very expandable *ex vivo*. Thus, relatively large donor site defects are necessary to harvest a large amount of tissue(s) from the patient in order to obtain sufficient numbers of articular chondrocytes or meniscal fibrochondrocytes for healing substantial articular cartilage or meniscal defects. In contrast, mesenchymal stem cells, whose natural progeny includes both chondrocytes and fibrochondrocytes, can be obtained in small quantities (e.g., a few cc of bone marrow content) (170) or from other connective tissue sources such as adipose tissue (170,171), readily expanded *ex vivo* in cell culture and reliably differentiated into chondrocytes cells (170,172). Embryonic stem cells may turn out to be a viable cell source for synovial joint regeneration, especially in consideration of the recent demonstration of the differentiation of embryonic stem cells into osteogenic cells (173,174). Embryonic stem cells are likely to be of greater significance in synovial joint condyle regeneration if the isolation and expansion of adult MSCs encounter substantial difficulties. However, thus far, it does not appear to be the case for synovial joint condyle regeneration (172). Chondrogenic and osteogenic cells derived from MSCs appear to be the logical choices at this time for exploring clinically applicable approaches toward regenerating the synovial joint condyle (152,172,175–180).

Biomaterial Scaffolds Are often Necessary for the Engineering of Structural Tissues such as Articular Cartilage and Meniscus

The optimal scaffolds for articular cartilage and meniscal regeneration are yet to be determined. An increasing number of meritorious studies have reported a wide range of natural and synthetic polymers for articular cartilage regeneration. Many of the tested synthetic polymers are biocompatible and biodegradable, two desirable features for cartilage regeneration (181). A model scaffold should allow effective diffusion of essential nutrients and metabolic wastes, given that chondrocytes and fibrochondrocytes both rely on diffusion for survival.

For cartilage regeneration, natural and synthetic polymers may need to simulate the extracellular matrix environment of chondrocytes that are created by Type II collagen and PGs in a highly aqueous matrix. Several hydrogels simulate cartilage matrix to various degrees, such as alginate, hyaluronate, chitosan, and polyethylene glycol-based polymers (175,176,181–183). The organization of chondrocyte phenotypes in various zones of articular cartilage may also need to be simulated, as demonstrated in recent reports by encapsulating articular chondrocytes from various zones of bovine articular cartilage into different hydrogel layers (184–186). The water content and diffusion properties of hydrogels mimic an ECM to allow tissue-forming cells to obtain systemic nutrients (187). The initial viscous liquid form of several hydrogel materials provides a unique capability to form complicated shapes while maintaining uniform cell distributions. For examples, an aqueous-derived silk scaffold also encompasses hydrogel-type properties and has been shown to support chondrogenesis (188). Pellet culture is a practice devoid of scaffolds that takes advantage of the dense, avascular, and aneural condition of cartilage. This system is done simply by centrifuging chondrocytes or MSCs into a pellet and incubating them in desired conditions. Although pellets do not provide efficient shape retention, they do bestow valuable information *in vitro* as models for chondrogenic MSC differentiation. Despite various levels of reported success with *in vitro* models, cell-hydrogel interactions need to be better understood, along with optimization of hydrogel composition, cross-linking, and degradation behavior as a function of the *in vivo* regenerative outcome.

In engineering an osteochondral construct, it is essential to construct a mold that is specific to the joint. Computer-aided approaches have been developed to construct molds that will both accurately replicate the anatomy of the joint as well as preserve the intricate architectural integrity of the interior of the scaffold. These intricate details can range from pore size to channel orientation to surface texture and can effectively contribute to the synthesis, or lack thereof, of the tissue one is trying to engineer. Common methods for 3D mold fabrication are based on software programs that read and digitize computerized tomography or magnetic resonance imaging. Solid free-form fabrication (SFF) technology can be used to produce an actual 3D scaffold through combining the interior architecture image and the external scaffold image, which is done via a layering process from computer-aided design files. SFF shows excellent promise due to the possibility of controlling the aforementioned necessary parameters needed in scaffold fabrication.

Growth Factors Are Necessary for Modulating Cell Behavior

Growth factors are proteins and polypeptides capable of modulating all aspects of cell behavior such as proliferation, differentiation, and apoptosis. Cells can be regulated by self-released growth factors (autocrine effect), or by growth factors released by other cells (paracrine effect). For chondrogenic differentiation, TGF-β superfamily is frequently used (e.g., 176,189). The de-differentiation and re-differentiation of chondrocytes are regulated by combinations of TGF-β1, fibroblast growth factor-2, and sequential exposure to IGF-I (190). TGF-βs also stimulate GAG synthesis in isolated cultured meniscus cells (191). Platelet-derived growth factor (PDGF) promotes chondrogenesis in chick limbs both *in vitro* and *in vivo* (191). Resting zone chondrocytes treated with PDGF shows hypertrophic activity in forming new cartilage (192). Although many growth factors have provided positive results, the optimal scheme of their application is not yet fully understood. The reader is referred to several recent in-depth reviews of growth factor delivery in cartilage regeneration (154,183,193).

Functional Tissue Engineering of Articular Cartilage and Meniscus

The field of functional tissue engineering has been proposed in response to the need to engineer tissues that have not only the appropriate cellular and matrix structures, but also the necessary physical properties (194,196). The aforementioned mechanical properties of both articular cartilage and meniscus are quite complex and intricate to replicate. The rather severe loading environment of cartilage in a synovial joint attributes to poor repair capabilities as well as degeneration. It has been proposed that a potential solution to engineer tissues with the appropriate physical properties lies in the physical factors, among which mechanical stress is most well studied (194–196). Prior to the conception of functional tissue engineering, it is widely acknowledged that mechanical factors can effectively influence cell behavior such as proliferation, differentiation, and matrix synthesis. Significant evidence exists that physical stress can accelerate or improve tissue regeneration and repair *in vitro*. The preference of chondrocytes to synthesize proper ECM components at an accelerated pace can help engineer cartilage and meniscus in an efficient manner. This concept can be used to enhance engineered grafts to more accurately represent native tissue. In order to better provide these stresses, a movement has begun to construct mechanical bioreactors that increase matrix synthesis through different approaches. Examples in which mechanical stimulation has proven advantageous in synthesizing cartilage has been through fluid flow (197,198), simulated hypogravity (199), simulated microgravity (200), static and cyclic compression (201–203), and hydrostatic pressure (204). Recent studies have suggested that external mechanical loading of cells or cell-polymer constructs may enhance their mechanical strength (e.g., 186,205,206). Although practical in theory, the type, frequency, area, and amount of

loading still need to be determined to furnish the optimal results of engineering each individual tissue.

The mechanical properties of engineered articular cartilage and meniscus can be readily tested, in many ways similar to the mechanical testing of native articular cartilage and meniscus. The results of compressive moduli of native and engineered articular cartilage or meniscus tissue using excised tissue plants and *in vitro* testing can be directly compared. However, loading measurements of native or engineered articular cartilage and meniscus have not been accomplished *in vivo*. Therefore, the ideal mechanical properties of tissue engineered articular cartilage can only be estimated, not accurately determined. Various biomaterials have been investigated for articular cartilage repair and have various mechanical properties that must be considered when designing a suitable scaffold. Engineering a scaffold that is too stiff may detrimentally affect cell viability and matrix synthesis. Stress shielding may occur and surrounding tissue may degrade as physiological loads are transferred to the more mechanically stiff implant. In contrast, a scaffold too soft will not exhibit the needed mechanical stiffness for the applied loads, causing physical breakdown and degradation of the material, leaving a void in the host tissue. Various zones of articular and fibrocartilage have different mechanical properties (56,207). Whether regional differences in the mechanical properties of articular cartilage and meniscus need to be simulated in tissue engineering remains to be explored (79).

SUMMARY AND CONCLUSIONS

Articular cartilage and meniscus are load-bearing and hydrated tissues that are key structures of diarthrodial joints. The similarities between the two structures include their remarkable capacity for resistance to and transmission of mechanical stress and their ability to enable joint lubrication. However, articular cartilage and meniscus have many important differences. Articular cartilage, in the overwhelming majority of human synovial joints, consists of hyaline cartilage, whereas the meniscus is composed of fibrocartilage. The mechanical properties of native articular cartilage and meniscus have been studied extensively and shown to vary with species, location, and even the orientation of test specimens. Motivated by the concept of functional tissue engineering, the mechanical properties of engineered articular cartilage from cells and biomaterials have also been investigated in recent years. As a result of drastically different structural and mechanical properties between articular cartilage and meniscus, the engineering challenges of the two tissues are different. Mesenchymal stem cells, or other stem cells, need to be differentiated into chondrocytes for the engineering of articular cartilage, whereas these stem cells may need to be differentiated into fibroblasts and chondrocytes, or perhaps fibrochondrocytes, for the engineering of meniscus. The optimal biomaterials for the engineering of articular cartilage and meniscus remain to be identified or fabricated. At least, the engineered articular cartilage and meniscus must adapt to possess similar mechanical properties of their native target tissues. The existing knowledge on the biological and mechanical properties of articular cartilage and meniscus provides the necessary foundation for the eventual goal to regenerate or replace diseased or lost articular cartilage and knee meniscus with engineered tissue analogs.

BIBLIOGRAPHY

1. Buckwalter JA, Mankin HJ. Instructional course lectures, the American Academy of Orthopaedic Surgeons—articular cartilage. Part II: Degeneration and osteoarthrosis, repair, regeneration, and transplantation. J Bone Joint Surg 1997;79:612–632.
2. Maroudas A. Physicochemical properties of articular cartilage. In: Freeman MAR, ed. Adult Articular Cartilage. Kent, UK: Pitman Medical Publishing; 1979. pp 215–290.
3. Mow VC, Gu WY, Chen FH. Structure and function of articular cartilage and meniscus. In: Mow VC, Huiskes R, eds. Basic Orthopaedic Biomechanics and Mechano-Biology; Third ed. New York: Lippincott Williams & Wilkins; 2005. pp 181–258.
4. Buckwalter JA, Mow VC. Cartilage repair in osteoarthritis. In: Moskowitz RW, Howell DS, Goldberg VM, Mankin HJ, eds. Osteoarthritis: Diagnosis and Management, 2nd ed. Philadelphia, PA: Saudners; 1992.
5. Howell DS, Treadwell BV, Trippel SB. Etiopathogenesis of osteoarthritis. In: Moskowitz RW, Goldberg VM, Howell DS, Altman RD, Buckwalter JA, eds. Osteoarthritis: Diagnosis and Medical/Surgical Management. 2th ed. Philadelphia, PA: WB Saunders; 1992. pp 233–252.
6. Poole AR, Howell DS. Etiopathogenesis of osteoarthritis. In: Moskowitz RW, Howell DS, Altman RD, Buckwalter JA, Goldberg VC, editors. Osteoarthritis, Diagnosis and Management. 3rd ed. Philadelphia, PA: WB Saunders Publishers; 2001. pp 29–47.
7. Mankin HJ, Mow VC, Buckwalter JA. Articular cartilage structure, composition and function. In: Buckwalter JA, Einhorn TA, Simon SR, eds. Orthopaedic Basic Science: Biology and Biomechanics of the Musculoskeletal System. Rosemont, IL: American Academy of Orthopaedic Surgeons; 2000. pp 443–470.
8. Benjamin M, Evans EJ. Fibrocartilage. J Anat 1990;171: 1–15.
9. Weightman B. Tensile fatigue of human articular cartilage. J Biomech 1976;9:193–200.
10. Seedhom BB, Wallbridge NC. Walking activities and wear of prosthesis. Ann Rheum Dis 1985;44:838–843.
11. Morrison JB. Bioengineering analysis of force actions transmitted by the knee joint. J Biomech Eng 1968;3:164–170.
12. Morrison JB. The mechanics of the knee joint in relation to normal walking. J Biomech 1970;3:51–61.
13. Athanasiou KA, Rosenwasser MP, Buckwalter JA, Malinene TI, Mow VC. Interspecies comparisons of in situ intrinsic mechanical properties of distal femoral cartilage. J Orthop Res 1991;9:330–340.
14. Ewers BJ, Dvoracek-Driksna D, Orth MW, Haut RC. The extent of matrix damage and chondrocyte death in mechanically traumatized articular cartilage explants depends on rate of loading. J Orthop Res 2001;19:779–784.
15. Torzilli PA, Grigiene R, Borrelli J Jr, Helfet DL. Effect of impact load on articular cartilage: Cell metabolism and viability, and matrix water content. J Biomech Eng 1999;121: 433–441.
16. Centers for Disease Control and Prevention (CDC). Prevalence of self-reported arthritis or chronic joint symptoms among adults- United States, 2001. MMWR 2002;51:948–950.
17. Vunjak-Novakovic G, Goldstein SA. Biomechanical principles of cartilage and tissue engineering. In: Mow VC, Huiskes R, eds. Basic Orthopaedic Biomechanics and Mechano-Biology,

3rd ed. New York: Lippincott Williams & Wilkins; 2005. pp 343–407.
18. Mow VC, Ratcliffe A, Chern KY, Kelly MA. Structure and function relationships of the meniscus of the knee. In: Mow VC, Arnoczky SP, Jackson DW, eds. Knee Meniscus: Basic and Clinical Foundations. New York: Raven Press; 1992. pp 37–57.
19. Mow VC, Ratcliffe A, Poole AR. Cartilage and diarthroidal joints as paradigms for hierarchical materials and structures. Biomaterials 1992;13:67–97.
20. Arnoczky SP. Gross and vascular anatomy of the meniscus and its role in meniscal healing, regeneration, and remodeling. In: Mow VC, Arnoczky SP, Jackson DW, eds. Knee Meniscus, Basic and Clinical Foundations. New York: Raven Press; 1992. pp 1–14.
21. Uhthoff HK, Wiley JJ. Behavior of the Growth Plate. New York: Raven Press; 1988.
22. Cohen B, Chorney GS, Phillips DP, Dick HM, Mow VC. Compressive stress-relaxation behavior of bovine growth plate may be described by the nonlinear biphasic theory. J Orthop Res 1994;12:804–813.
23. Williams PL. Gray's Anatomy. New York: Churchill Livingston; 1995.
24. Shimazu A, Nah HD, Kirsch T, Koyama E, Leatherman JL, Golden EB, Kosher RA, Pacifici M. Syndecan-3 and the control of chondrocyte proliferation during endochondral ossification. Exp Cell Res 1996;229:126–136.
25. Mao JJ, Nah HD. More research needed to understand how orthodontists communicate with cells. Am J Orthod Dentofacial Orthop 2004;125:676–689.
26. Tamamura Y, Otani T, Kanatani N, Koyama E, Kitagaki J, Komori T, Yamada Y, Costantini F, Wakisaka S, Pacifici M, Iwamoto M, Enomoto-Iwamoto M. Developmental regulation of Wnt/beta-catenin signals is required for growth plate assembly, cartilage integrity, and endochondral ossification. J Biol Chem 2005;280:19185–19195.
27. Mow VC, Kuei SC, Lai WM, Armstrong CG. Biphasic creep and stress relaxation of articular cartilage in compression: Theory and experiments. J Biomech Eng 1980;102:73–84.
28. Eyre DR. Structure and function of the cartilage collagen: Role of type IX in articular cartilage. In: Brandt KD, ed. Cartilage Changes in Osteoarthritis. Indianapolis, IN: Ciba-Geigy; 1990. pp 12–16.
29. Heinegard D, Bayliss M, Lorenzo P. Biochemistry and metabolism of normal and osteoarthritic cartilage. In: Brandt KD, Doherty M, Lohmander LS, eds. Osteoarthritis. New York: Oxford University Press; 2003. pp 73–82.
30. Eyre DR, Wu JJ. Collagen of fibrocartilage: A distinctive molecular phenotype in bovine meniscus. FEBS Lett 1983;158:265–270.
31. Eyre DR, Oguchi H. The hydroxypyridinium crosslinks of skeletal collagens: Their measurement, properties and a proposed pathway of formation. Biochem Biophys Res Commun 1980;92:402–410.
32. Eyre DR, Dickson IR, Van Ness K. Collagen cross-linking in human bone and articular cartilage. Age-related changes in the content of mature hydroxypyridinium residues. Biochemistry 1988;252:495–500.
33. Huang CY, Mow VC, Ateshian GA. The role of flow-independent viscoelasticity in the biphasic tensile and compressive responses of articular cartilage. J Biomech Eng 2001;123: 410–417.
34. Huang CY, Soltz MA, Kopacz M, Mow VC, Ateshian GA. Experimental verification of the roles of intrinsic matrix viscoelasticity and tension-compression nonlinearity in the biphasic response of cartilage. J Biomech Eng 2003;125: 84–93.
35. Kempson GE, Freeman MA, Swanson SA. Tensile properties of articular cartilage. Nature 1968;220:1127–1128.
36. Kempson GE, Muir H, Pollard C, Tuke M. The tensile properties of the cartilage of human femoral condyles related to the content of collagen and glycosaminoglycans. Biochim Biophys Acta 1973;297:456–472.
37. Roth V, Mow VC. The intrinsic tensile behavior of the matrix of bovine articular cartilage and its variation with age. J Bone Joint Surg 1980;62:1102–1117.
38. Woo SL-Y, Akeson WH, Jemmott GF. Measurements of nonhomogeneous directional mechanical properties of articular cartilage in tension. J Biomechan 1976;9:785–791.
39. Lai WM, Hou JS, Mow VC. A triphasic theory for the swelling and deformation behaviors of articular cartilage. J Biomech Eng 1991;113:245–258.
40. Buckwalter JA, Rosenberg LC. Electron microscopic studies of cartilage proteoglycans: Direct evidence for the variable length of the chondroitin sulfate-rich region of proteoglycan subunit core protein. J Biol Chem 1982;257:8930–8939.
41. Muir H. Proteoglycans as organizers of the intercellular matrix. Biochem Soc Trans 1983;9:613–622.
42. Muir H. The chondrocyte, architect of cartilage: Biomechanics, structure, function and molecular biology of cartilage matrix macromolecules. Bioessays 1995;17:1039–1048.
43. Sun DN, Guo XE, Likhitpanichkul M, Lai WM, Mow VC. The influence of the fixed negative charges on mechanical and electrical behaviors in articular cartilage under unconfined compression. J Biomech Eng 2004;126:1–11.
44. Hunziker EB, Michel M, Studer D. Ultrastructure of adult human articular cartilage matrix after cryotechnical processing. Microsc Res Tech 1997;37:271–284.
45. Hultkrantz W. Ueber die spaltrichtungen der gelenkknorpel. Anat Anzeig verhandl anat Gesellsch 1898;14:248–256.
46. Benninghoff A. Form und bau der gelenkknorpel in ihren beziehungen zu funktion. II. Der aufbau des gelenkknorpel in seinen beziehungen zu funktion. Z Zellforsch 1925;2: 783–862.
47. Broom ND. The collagen framework of articular cartilage: Its profound influence on normal and abnormal load-bearing function. In: Nimni ME, ed. Collagen: Chemistry, Biology and Biotechnology. Boca Raton, FL: CRC Press; 1988. pp 243–265.
48. Ceohn ZA, Henry JH, McCarthy DM, Mow VC, Ateshian GA. Computer simulations of patellofemoral joint surgery. Am J. Sports Med 2003;31:87–98.
49. Lipshitz H, Etheredge R 3rd, Glimcher MJ. Changes in the hexosamine content and swelling ratio of articular cartilage as functions of depth from the surface. J Bone Joint Surg 1976;58:1149–1153.
50. Mitrovic D, Quintero M, Stankovic A, Ryckewaert A. Cell density of adult human femoral condylar articular cartilage. Joints with normal and fibrillated surfaces. Lab Invest 1983;49:309–316.
51. Aydelotte MB, Schumacher BL, Kuettner KE. Heterogeneity of articular chondrocytes. In: Kuettner KE, ed. Articular Cartilage and Osteoarthritis. New York: Raven Press; 1992. pp 237–249.
52. Kempson GE. Mechanical properties of adult articular cartilage. In: Freeman MAR, ed. Adult Articular Cartilage. Kent, UK: Pitman Medical Publishing; 1979. pp 215–290.
53. Akizuki S, Mow VC, Muller F, Pita JC, Howell DS. Tensile properties of human knee joint cartilage. II. Correlations between weight bearing and tissue pathology and the kinetics of swelling. J Orthop Res 1987;5:173–186.
54. Hunziker EB, Tyler JA. Articular cartilage repair. In: Brandt KD, Doherty M, Lohmander LS, eds. Osteoarthritis. Oxford, UK: Oxford University Press; 2003. pp 93–101.

55. Jadin KD, Wong BL, Bae WC, Li KW, Williamson AK, Schumacher BL, Price JH, Sah RL. Depth-varying density and organization of chondrocytes in immature and mature bovine articular cartilage assessed by 3-D imaging and analysis. J Histochem Cytochem DOI: 10.1369/jhc.4A6511. 2005 (In Press).
56. Tomkoria S, Patel RV, Mao JJ. Heterogeneous nanomechanical properties of superficial and zonal regions of articular cartilage of the rabbit proximal radius condyle by atomic force microscopy. Med Eng Phys 2004;26:815–822.
57. Broom ND, Silyn-Roberts H. Collagen-collagen versus collagen-proteoglycan interactions in the determination of cartilage strength. Arthritis Rheum 1990;33:1512–1517.
58. Quinn TM, Hunziker EB, Hauselmann HJ. Variation of cell and matrix morphologies in articular cartilage among locations in the adult human knee. Osteoarthritis Cartilage 2005;13:672–678.
59. Guilak F, Mow VC. The mechanical environment of the chondrocyte: A biphasic finite element model of cell-matrix interactions in articular cartilage. J Biomech 2000;33:1663–1673.
60. Guilak F. The deformation behavior and viscoelastic properties of chondrocytes in articular cartilage. Biorheology 2000;37: 27–44.
61. Guilak F. Compression-induced changes in the shape and volume of the chondrocyte nucleus. J Biomech 1995;28:1529–1542.
62. Knight MM, Ross JM, Sherwin AF, Lee DA, Bader DL, Poole CA. Chondrocyte deformation within mechanically and enzymatically extracted chondrons compressed in agarose. Biochem Biophys Acta 2001;1526:141–146.
63. Graff RD, Kelley SS, Lee GM. Role of pericellular matrix in development of a mechanically functional neocartilage. Biotechnol Bioengin 2003;20:457–464.
64. Allen DM, Mao JJ. Heterogeneous nanostructural and nanoelastic properties of pericellular and interterritorial matrices of chondrocytes by atomic force microscopy. J Struct Biol 2004;145:196–204.
65. Stockwell RA. Biology of Cartilage Cells. Cambridge, UK: Cambridge University Press; 1979.
66. Ateshian GA, Mow VC. Friction, lubrication, and wear of articular cartilage and diarthroidal joints. In: Mow VC, Huiskes R, eds. Basic Orthopaedic Biomechanics and Mechano-Biology, Third ed. New York: Lippincott Williams & Wilkins; 2005. pp 447–494.
67. McDevitt CA, Miller A, Spindler KP. The cells and cell matrix interaction of the meniscus. In: Mow VC, Arnoczky SP, Jackson DW, eds. Knee Meniscus, Basic and Clinical Foundations. New York: Raven Press; 1992. pp 29–36.
68. Bullough PG, Munuera L, Murphy J, Weinstein AM. The strength of the menisci of the knee as it relates to their fine structure. J Bone Joint Surg Br 1970;52:564–567.
69. Aspden RM, Yarker YE, Hukins DW. Collagen orientations in the meniscus of the knee joint. J Anat 1985;140:371–380.
70. Kelly MA, Fithian DC, Chern KY, Mow VC. Structure and function of the meniscus: Basic and clinical implications. In: Mow VC, Ratcliffe A, Woo SLY, eds. Biomechanics of Diarthroidal Joints. New York: Springer-Verlag; 1990. pp 191–214.
71. Yasui K. Three-dimensional architecture of human normal menisci. J Jpn Orthop Assoc 1978;52:391–399.
72. Skaggs DL, Warden WH, Mow VC. Radial tie fibers influence the tensile properties of the bovine medial meniscus. J Orthop Res 1994;12:176–185.
73. Nakano T, Thompson JR, Aherne FX. Distribution of glycosaminoglycans and the nonreducible collagen crosslink, pyridinoline in porcine menisci. Can J Vet Res 1986;50: 532–536.
74. Fithian DC, Kelly MA, Mow VC. Material properties and structure-function relationships in the menisci. Clin Orthop Relat Res 1990;252:19–31.
75. Walker PS, Erkman MJ. The role of the menisci in force transmission across the knee. Clin Orthop Relat Res 1975;109:184–192.
76. Smith MM, Ghosh P. Experimental models of osteoarthritis. In: Moskowitz RW, Howell DS, Altman RD, Buckwalter JA, Goldberg VC, eds. Osteoarthritis, Diagnosis and Management. 3rd ed. Philadelphia, PA: WB Saunders Publishers; 2001. pp 171–199.
77. Moskowitz RW, Davis W, Sammarco J, et al., Experimentally induced degenerative joint lesions following partial meniscectomy in the rabbit. Arthritis Rheum 1973;16:397–404.
78. Dehaven KE. The role of the meniscus. In: Ewing JW, ed. Articular Cartilage and the Knee Joint Function: Basic Science and Arthroscopy. New York: Raven Press; 1990. pp 103–115.
79. Guilak F, Butler DL, Goldstein SA. Functional tissue engineering: The role of biomechanic in articular cartilage repair. Clin Orthop Relat Res 2001; (391 Suppl):S295–S305.
80. Kempson GE. Mechanical properties of articular cartilage and their relationship to matrix degeneration and age. Ann Rheum Dis 1975;34:111–113.
81. Kempson GE. Age-related changes in the tensile properties of human articular cartilage: A comparative study between the femoral head of the hip joint and the talus of the ankle joint. Biochim Biophys Acta 1991;1075:223–230.
82. Akizuki S, Mow VC, Muller F, Pita JC, Howell DS, Manicourt DH. Tensile properties of human knee joint cartilage: I. Influence of ionic conditions, weight bearing, and fibrillation on the tensile modulus. J Orthop Res 1986;4:379–392.
83. Setton LA, Mow VC, Muller FJ, Pita JC, Howell DS. Mechanical properties of canine articular cartilage are significantly altered following transection of the anterior cruciate ligament. J Orthop Res 1994;12:451–463.
84. Setton LA, Zhu W, Mow VC. The biphasic poroviscoelastic behavior of articular cartilage: role of the surface zone in governing the compressive behavior. J Biomech 1993;26: 581–592.
85. Soltz MA, Ateshian GA. A conewise linear elasticity mixture model for the analysis of tension-compression nonlinearity in articular cartilage. J Biomech Eng 2000;122:576–586.
86. Huang CY, Stankiewicz A, Ateshian GA, Mow VC. Anisotropy, inhomogeneity, and tension-compression nonlinearity of human glenohumeral cartilage in finite deformation. J Biomech 2005;38:799–809.
87. Fithian DC, Zhu WB, Ratcliffe A, Kelly MA, Mow VC. Exponential law representation of tensile properties of human meniscus. Proc Inst Mech Eng 1989;c384/058:85–90.
88. Bader DL, Kempson GE, Barrett AJ, Webb W. The effects of leucocyte elastase on the mechanical properties of adult human articular cartilage in tension. Biochemica et Biophysica Acta 1981;677:103–108.
89. Schmidt MB, Mow VC, Chun LE, Eyre DR. Effects of proteoglycan extraction on the tensile behavior of articular cartilage. J Orthop Res 1990;8:353–363.
90. Hirsh 1944.
91. Sokoloff L. Elasticity of aging cartilage. Fed Proc 1966;25: 1089–1095.
92. Kempson GE, Freeman MA, Swanson SA. The determination of a creep modulus for articular cartilage from indentation tests of the human femoral head. J Biomech 1971;4: 239–250.
93. Kempson GE, Spivey CJ, Swanson SA, Freeman MA. Patterns of cartilage stiffness on normal and degenerate human femoral heads. J Biomech 1971;4:597–609.

94. Hoch DH, Grodzinsky AJ, Koob TJ, Albert ML, Eyre DR. Early changes in material properties of rabbit articular cartilage after meniscectomy. J Orthop Res 1983;1:4–12.
95. Parsons JR, Black J. The viscoelastic shear behavior of normal rabbit articular cartilage. J Biomech 1977;10:21–29.
96. Altman RD, Tenenbaum J, Latta L, Riskin W, Blanco LN, Howell DS. Biomechanical and biochemical properties of dog cartilage in experimentally induced osteoarthritis. Ann Rheum Dis 1984;43:83–90.
97. Armstrong CG, Lai WM, Mow VC. An analysis of the unconfined compression of articular cartilage. J Biomech Eng 1984;106:165–173.
98. Mak AF. The apparent viscoelastic behavior of articular cartilage— The contributions from the intrinsic viscoelasticity and interstitial fluid flow. J Biomech Eng 1986;108:123–130.
99. Mak AF. Unconfined compression of hydrated viscoelastic tissues: A biphasic poroviscoelastic analysis. Biorheology 1986;23:371–383.
100. Mak AF, Lai WM, Mow VC. Biphasic indentation of articular cartilage—I. Theoretical analysis. J Biomech 1987;20:703–714.
101. Zhu W, Mow VC, Koob TJ, Eyre DR. Viscoelastic shear properties of articular cartilage and the effects of glycosidase treatments. J Orthop Res 1993;11:771–781.
102. Armstrong CG, Mow VC. Variations in the intrinsic mechanical properties of human articular cartilage with age, degeneration, and water content. J. Bone Joint Surg 1982;64: 88–94.
103. Roth V, Mow VC, Lai WM, Eyre DR. Correlation of intrinsic compressive properties of bovine articular cartilage with its uronic acid and water content. Trans Orthop Res Soc 1981;6:21.
104. Froimson MI, Ratcliffe A, Gardner TR, Mow VC. Differences in patellofemoral joint cartilage material properties and their significance to the etiology of cartilage surface fibrillation. Osteoarthritis Cartilage 1997;5:377–386.
105. Jurvelin J, Kiviranta I, Arokoski J, Tammi M, Helminen HJ. Indentation study of the biochemical properties of articular cartilage in the canine knee. Eng Med 1987;16:15–22.
106. Stahursky TM, Armstrong CG, Mow VC. Variation of the intrinsic aggregate modulus and permeability of articular cartilage with trypsin digestion. Proc Biomech Symp Trans ASME 1981;AMD43:137–140.
107. Mow VC, Zhu W, Lai WM, Hardingham TE, Hughes C, Muir H. The influence of link protein stabilization on the viscometric properties of proteoglycan aggregate solutions. Biochim Biophys Acta 1989;992:201–208.
108. Schinagl RM, Ting MK, Price JH, Sah RL. Video microscopy to quantitate the inhomogeneous equilibrium strain within articular cartilage during confined compression. Ann Biomed Eng 1996;24:500–512.
109. Chen AC, Bae WC, Schnagl RM, Sah RL. Depth- and strain-dependent mechanical and electromechanical properties of full-thickness articular cartilage in confined compression. J Biomechan 2001;34:1–12.
110. Wang CC, Hung CT, Mow VC. An analysis of the effects of depth-dependent aggregate modulus on articular cartilage stress-relaxation behavior in compression. J Biomech 2001;34:75–84.
111. Cohen B, Lai WM, Mow VC. A transversely isotropic biphasic model for unconfined compression of growth plate and chondroepiphysis. J Biomech Eng 1998;120:491–496.
112. Hayes WC, Mockros LF. Viscoelastic properties of human articular cartilage. J Appl Physiol 1971;31:562–568.
113. Spirt AA, Mak AF, Wassell RP. Nonlinear viscoelastic properties of articular cartilage in shear. J Orthop Res 1989;7: 43–49.
114. Simon WH, Mak A, Spirt A. The effect of shear fatigue on bovine articular cartilage. J Orthop Res 1990;8:86–93.
115. Roth V, Schoonbeck JM, Mow VC. Low frequency dynamic behavior of articular cartilage under torsional shear. Trans Orthop Res Soc 1982;7:150.
116. Zhu W, Lai WM, Mow VC. Intrinsic quasilinear viscoelastic behavior of the extracellular matrix of cartilage. Trans Orthop Res Soc 1986;11:407.
117. Setton LA, Mow VC, Howell DS. Mechanical behavior of articular cartilage in shear is altered by transection of the anterior cruciate ligament. J Orthop Res 1995;13:473–482.
118. Fung YC. Mechanical Properties of Living Tissues. New York: Springer-Verlag; 1981.
119. Zhu W, Chern KY, Mow VC. Anisotropic viscoelastic shear properties of bovine meniscus. Clin Orthop Relat Res 1994;306:34–45.
120. Mow VC, Mak AF, Lai WM, Rosenberg LC, Tang LH. Viscoelastic properties of proteoglycan subunits and aggregates in varying solution concentrations. J Biomech 1984;17:325–338.
121. Zhu W, Mow VC, Rosenberg LC, Tang LH. Determinations of kinetic changes of aggrecan-hyaluronan interactions in solution from its rheological properties. J Biomech 1994;27: 571–579.
122. Hayes WC, Bodine AJ. Flow-independent viscoelastic properties of articular cartilage matrix. J Biomechan 1978;11: 407–419.
123. Maroudas A, Muir H, Wingham J. The correlation of fixed negative charge with glycosaminoglycan content of human articular cartilage. Biochim Biophys Acta 1969;177:492–500.
124. Donnan FG. The theory of membrane equilibria. Chem Rev 1924;1:73–90.
125. Myers ER, Lai WM, Mow VC. A continuum theory and an experiment for the ion-induced swelling behavior of articular cartilage. J Biomech Eng 1984;106:151–158.
126. Setton LA, Tohyama H, Mow VC. Swelling and curling behaviors of articular cartilage. J Biomech Eng 1998;120: 355–361.
127. Setton LA, Lai WM, Mow VC. Swelling-induced residual stress and the mechanism of curling in articular cartilage in vitro. In: Tarbell JM, ed. Advances in Bioengineering. 1993. pp 59–62.
128. Wan LQ, Miller C, Guo XE, Mow VC. A three-layer orthotropic model for swelling and curling of articular cartilage. ASME-BED. Vail, Colorado, 2005.
129. Olsen S, Oloyede A. A finite element analysis methodology for representing the articular cartilage functional structure. Comput Meth Biomech Biomed En 2002;5(6):377–386.
130. Gu WY, Lai WM, Mow VC. A mixture theory for charged-hydrated soft tissues containing multi-electrolytes: Passive transport and swelling behaviors. J Biomech Eng 1998;120: 169–180.
131. Huyghe JM, Janssen JD. Quadriphasic mechanics of swelling incompressible prorous media. Int J Eng Sci 1997;35: 793–802.
132. Gu WY, Lai WM, Mow VC. Transport of fluid and ions through a porous-permeable charged-hydrated tissue, and streaming potential data on normal bovine articular cartilage. J Biomech 1993;26:709–723.
133. Gu WY, Yao H. Effects of hydration and fixed charge density on fluid transport in charged hydrated soft tissues. Ann Biomed Eng 2003;31:1162–1170.
134. Lai WM, Mow VC, Sun DD, Ateshian GA. On the electric potentials inside a charged soft hydrated biological tissue: streaming potential versus diffusion potential. J Biomech Eng, 2000;122:336–346.
135. Lai WM, Sun DD, Ateshian GA, Guo XE, Mow VC. Electrical signals for chondrocytes in cartilage. Biorheology 2002;39: 39–45.

136. Mow VC, Ateshian GA, Lai WM, Gu WY. Effects of fixed charges on the stress-relaxation behavior of hydrated soft tissues in a confined compression problem. Int J Solids Structures 1998;35:4945–4962.
137. Sun DN, Gu WY, Guo XE, Lai WM, Mow VC. A mixed finite element formulation of triphasic mechano-electrochemical theory for charged, hydrated biological soft tissues. Int J Num Methods Eng 1999;45:1375–1402.
138. Lu X, Sun DD, Guo XE, Chen FC, Lai WM, Mow VC. Indentation determined mechano-electrochemical properties and fixed charge density of articular cartilage. Ann Biomed Eng 2004;32:370–379.
139. Wan LQ, Miller C, Guo XE, Mow VC. Fixed electrical charges and mobile ions affect the measurable mechano-electrochemical properties of charged-hydrated biological tissues: The articular cartilage paradigm. Mechan Chem Biosyst 2004;1: 81–99.
140. Ateshian GA, Chahine NO, Basalo IM, Hung CT. The correspondence between equilibrium biphasic and triphasic material properties in mixture models of articular cartilage. J Biomechan 2004;37:391–400.
141. Likhitpanichkul M, Miller C, Guo XE, Mow VC. A triphasic model of cell under micropipette aspiration: The osmotic effect on cell mechanical properties. ASME-BED. Vail, Colorado, 2005.
142. Lu X, Miller C, Guo XE, Mow VC. A new correspondence principle for triphasic materials: Determination of fixed charge density and porosity of articular cartilage by indentation. ASME-BED. Vail, Colorado, 2005.
143. Broom ND. Structural consequences of traumatizing articular cartilage. Ann Rheum Dis 1986;45:225–234.
144. Cawston T. Matrix metalloproteinases and TIMPs: Properties and implications for the rheumatic diseases. Mol Med Today 1998;4:130–137.
145. Lotz M, Hashimoto S, Kuhn K. Mechanisms of chondrocyte apoptosis. Osteoarthritis Cartilage 1999;7:389–391.
146. Meredith Jr JE , Fazeli B, Schwartz MA. The extracellular matrix as a cell survival factor. Mol Biol Cell 1993;4:953–961.
147. Hashimoto S, Ochs RL, Komiya S, Lotz M. Linkage of chondrocyte apoptosis and cartilage degradation in human osteoarthritis. Arthritis Rheum 1998;41:1632–1638.
148. Lyyra T, Kiviranta I, Vaatainen U, Helminen HJ, Jervelin JS. *In vivo* characterization of indentation stiffness of articular cartilage in the normal human knee. J Biomed Mater Res 1999;48:482–487.
149. Kobayashi T, Yoshihara Y, Samura A, Tanaka O, Shimmei M. Chondrocalcin as a marker of articular cartilage degeneration in anterior cruciate ligament-deficient knees. Orthopedics 1998;21:773–776.
150. Kuster MS, Posdiadlo P, Stachowiak GW. Shape of wear particles found in human knee joints and their relationship to osteoarthritis. Br J Rheumatol 1998;37:978–984.
151. Hollinger JO, Winn S, Bonadio J. Options for tissue engineering to address challenges of the aging skeleton. Tissue Eng 2000;6:341–350.
152. Mao JJ. Stem-cell driven regeneration of synovial joints. Biol Cell 2005;97:289–301.
153. Buckwalter JA. Articular cartilage injuries. Review Clin Orthop Relat Res 2002;3:257–264.
154. Hunziker EB. Articular cartilage repair: basic science and clinical progress. A review of the current status and prospects. Osteoarthritis Cartilage 2002;10:432–463.
155. Martin JA, Buckwalter JA. The role of chondrocyte-matrix interactions in maintaining and repairing articular cartilage. Biorheology 2000;37:129–140.
156. Redman SN, Oldfield SF, Archer CW. Current strategies for articular cartilage repair. Eur Cell Mater 2005;9:23–32.
157. Johnson LL. Arthroscopic Abrasion Arthroscopy. In: McGinty JB, ed. Operative Arthroplasty. Philadelphia, PA: Lippincott-Raven; 1996.
158. Shapiro F, Koide S, Glimcher MJ. Cell origin and differentiation in the repair of full-thickness defects of articular cartilage. J Bone Joint Surg Am 1993;75:532–553.
159. O'Driscoll SW. Articular cartilage regeneration using perisosteum. Clin Orthop 1999;367:S186–S203.
160. Ahmad CS, Guiney WB, Drinkwater CJ. Evaluation of donor site intrinsic healing response in autologous osteochondral grafting of the knee. Arthoscopy 2002;18:95–98.
161. NIH Consensus Panel. NIH Consensus Statement on total knee replacement December 8–10, 2003. J Bone Joint Surg Am 2004;86-A: 1328–1335.
162. Haydon CM, Mehin R, Burnett S, Rorabeck CH, Bourne RB, McCalden RW, MacDonald SJ. Revision total hip arthroplasty with use of a cemented femoral component. Results at a mean of ten years. J Bone Joint Surg Am 2004;86-A: 1179–1185.
163. DeHaven KE. Meniscectomy versus repair: Clinical experience. In: Mow VC, Arnoczky SP, Jackson DW, eds. Knee Meniscus: Basic and Clinical Foundations. New York: Raven Press; 1992. pp 131–139.
164. Jadlowiec JA, Celil AB, Hollinger JO. Bone tissue engineering: Recent advances and promising therapeutic agents. Expert Opin Biol Ther 2003;3:409–423.
165. Vacanti JP, Langer R. Tissue engineering: The design and fabrication of living replacement devices for surgical reconstruction and transplantation. Lancet 1999;354(Suppl)1: S132–S134.
166. Zhang JY, Doll BA, Beckman EJ, Hollinger JO. Three-dimensional biocompatible ascorbic-acid containing scaffold for bone tissue engineering. Tissue Eng 2003;9:1143–1157.
167. Weng Y, Cao Y, Silva CA, Vacant MP, Vacanti CA. Tissue-engineered composites of bone and cartilage for mandible condylar reconstruction. J Oral Maxillofac Surg 2001;59: 185–190.
168. Niederauer GG, Slivka MA, Leatherbury NC, Korvick DL, Harroff HH, Ehler WC, Dunn CJ, Kieswatter K. Evaluation of multiphase implants for repair of focal osteochondral defects in goats. Biomaterials 2000;21:2561–2574.
169. Freed LE, Grande DA, Lingbin Z, Emmanuel J, Marquis JC, Langer R. Joint resurfacing using allograft chondrocytes and synthetic biodegradable polymer scaffolds. J Biomed Mater Res 1994;28:891–899.
170. Caplan AI. Mesenchymal stem cells. J Orthop Res 1991;9: 641.
171. Gimble J, Guilak F. Adipose-derived adult stem cells: Isolation, characterization, and differentiation potential. Cytotherapy 2003;5:362–369.
172. Alhadlaq A, Mao JJ. Mesenchymal stem cell: Isolation and therapeutics. Stem Cells Develop 2004;13:436–448.
173. Buttery LD, Bourne S, Xynos JD, Wood H, Hughes FJ, Hughes SP, Episkopou V, Polak JM. Differentiation of osteoblasts and in vitro bone formation from murine embryonic stem cells. Tissue Eng 2001;7:89–99.
174. Sottile Thomson VA, McWhir J. In vitro osteogenic differentiation of human ES cells. Cloning Stem Cells 2003;5:149–155.
175. Alhadlaq A, Mao JJ. Tissue-engineered neogenesis of human-shaped mandibular condyle from rat mesenchymal stem cells. J Dent Res 2003;82:950–955.
176. Alhadlaq A, Elisseeff JH, Hong L, Williams CG, Caplan AI, Sharma B, Kopher RA, Tomkoria S, Lennon DP, Lopez A, Mao JJ. Adult stem cell driven genesis of human-shaped articular condyle. Ann Biomed Eng 2004;32:911–923.
177. Gao J, Dennis JE, Solchaga LA, Awadallah AS, Goldberg VM, Caplan AI. Tissue-engineered fabrication of an osteochondral composite graft using rat bone marrow-derived mesenchymal stem cells. Tissue Eng 2001;7:363–371.

178. Gao J, Dennis JE, Solchaga LA, Goldberg VM, Caplan AI. Repair of osteochondral defect with tissue-engineered two-phase composite material of injectable calcium phosphate and hyaluronan sponge. Tissue Eng 2002;8:827–837.
179. Gao J, Caplan AI. Mesenchymal stem cells and tissue engineering for orthopaedic surgery. Chir Organi Mov 2003;88:305–316.
180. Rahaman MN, Mao JJ. Stem cell based composite tissue constructs for regenerative medicine. Biotechnol Bioeng 2005;91:261–284.
181. Lee KY, Mooney DJ. Hydrogels for tissue engineering. Chem Rev 2001;101:1869–1879.
182. Anseth KS, Metters AT, Bryant SJ, Martens PJ, Elisseeff JH, Bowman CN. In situ forming degradable networks and their application in tissue engineering and drug delivery. J Control Release 2002;78:199–209.
183. Randolph MA, Anseth K, Yaremchuk MJ. Tissue engineering of cartilage. Clin Plast Surg 2003;30:519–537.
184. Klein TJ, Schumacher BL, Schmidt TA, Li KW, Voegtline MS, Masuda K, Thonar EJ, Sah RL. Tissue engineering of stratified articular cartilage from chondrocyte subpopulations. Osteoarthritis Cartilage 2003;11:595–602.
185. Kim TK, Sharma B, Williams CG, Ruffner MA, Malik A, McFarland EG, Elisseeff JH. Experimental model for cartilage tissue engineering to regenerate the zonal organization or articular cartilage. Osteoarthritis Cartilage 2003;11:653–664.
186. Williams CG, Kim TK, Taboas A, Malik A, Manson P, Elisseeff J. In vitro chondrogenesis of bone marrow-derived mesenchymal stem cells in a photopolymerizing hydrogel. Tissue Eng 2003;9:679–688.
187. Peppas NA, Huang Y, Torres-Lugo M, Ward JH, Zhang J. Physicochemical foundations and structural design of hydrogels in medicine and biology. Annu Rev Biomed Eng 2000;2: 9–29.
188. Wang Y, Kim U, Blasioli DJ, Kim H, Kaplan DL. In vitro cartilage tissue engineering with 3D porous aqueous-derived silk scaffolds and mesenchymal stem cells. Biomaterials 2005;26:7082–7094.
189. Pittenger MF, Mackay AM, Beck SC, Jaiswal RK, Douglas R, Mosca JD, Moorman MA, Simonetti DW, Craig S, Marshak DR. Multilineage potential of adult human mesenchymal stem cells. Science 1999;284:143–147.
190. Pei M, Seidel J, Vunjak-Novakovic G, Freed LE. Growth factors for sequential cellular de- and re-differentiation in tissue engineering. Biochem Biophys Res Commun 2002;294: 149–154.
191. Collier S, Ghosh P. Effects of transforming growth factor beta on proteoglycan synthesis by cell and explant cultures derived from the knee joint meniscus. Osteoarthritis Cartilage 1995;3:127–138.
192. Lohmann CH, Schwartz Z, Niederauer GG, Boyan BD. Degree of differentiation of chondrocytes and their pretreatment with platelet-derived growth factor. Regulating induction of cartilage formation in resorbable tissue carriers in vivo. Orthopade 2000;29(2):120–128.
193. Almarza AJ, Athanasiou KA. Design characteristics for the tissue engineering of cartilaginous tissues. Ann Biomed Eng 2004;32:2–17.
194. Butler DL, Shearn JT, Juncosa N, Dressler MR, Hunter SA. Functional tissue engineering parameters toward designing repair and replacement strategies. Clin Orthop Relat Res 2004; Suppl: S190–S199.
195. Guilak F, Fermor B, Keefe FJ, Kraus VB, Olson SA, Pisetsky DS, Setton LA, Weinberg JB. The role of biomechanics and inflammation in cartilage injury and repair. Clin Orthop Relat Res 2004;423:17–23.
196. Wang CC, Guo XE, Sun D, Mow VC, Ateshian GA, Hung CT. The functional environment of chondrocytes with cartilage subjected to compressive loading: A theoretical and experimental approach. Biorheology 2002;39:11–25.
197. Freed LE, Martin I, Vunjak-Novakovic G. Frontiers in tissue engineering. In vitro modulation of chondrogenesis. Clin Orthop Relat Res 1999; (367 Suppl):S46–S58.
198. Pazzano D, Mercier KA, Moran JM, Fong SS, DiBiasio DD, Rulfs JX, Kohles SS, Bonassar LJ. Comparison of chondrogenesis in static and perfused bioreactor culture. Biotechnol Prog 2000;16:893–896.
199. Freed LE, Langer R, Martin I, Pellis NR, Vunjak-Novakovic G. Tissue engineering of cartilage in space. PNAS USA 1997;94:13885–13890.
200. Freed LE, Vunjak-Novakovic G. Microgravity tissue engineering. In Vitro Cell Dev Biol Anim 1997;33:381–385.
201. Buschmann MD, Gluzband YA, Grodzinsky AJ, Hunziker EB. Mechanical compression modulates matrix biosynthesis in chondrocyte/agarose culture. J Cell Sci 1995;108(Part 4):1497–1508.
202. Mauck RL, Soltz MA, Wang CC, Wong DD, Chao PH, Valhmu WB, Hung CT, Ateshian GA. Functional tissue engineering of articular cartilage through dynamic loading of chondrocyte-seeded agarose gels. J Biomech Eng 2000;122:252–260.
203. Seidel JO, Pei M, Gray ML, Langer R, Freed LE, Vunjak-Novakovic G. Long-term culture of tissue engineered cartilage in a perfused chamber with mechanical stimulation. Biorheology 2004;41:445–458.
204. Saris DB, Sanyal A, An KN, Fitzsimmons JS, O'Driscoll SW. Periosteum responds to dynamic fluid pressure by proliferating in vitro. J Orthop Res 1999;17:668–677.
205. Simmons CA, Matlis S, Thornton AJ, Chen S, Wang CY, Mooney DJ. Cyclic strain enchances matrix mineralization by adult human mesenchymal stem cells via the extracellular signal-regulated kinase (ERK1/2) signaling pathway. J Biomech 2003;36:1087–1096.
206. Grodzinsky AJ, Levenston ME, Jin M, Frank EH. Cartilage tissue remodeling in response to mechanical forces. Annu Rev Biomed Eng 2000;2:691–713.
207. Hu K, Radhakrishnan P, Patel RV, Mao JJ. Regional structural and viscoelastic properties of fibrocartilage upon dynamic nanoindentation of the articular condyle. J Struct Biol 2001;136:46–52.

Further Reading

Archard JF. Wear theory and mechanisms. In: Peterson MB, Winder WO, eds. *Wear Control Handbook*. New York: ASME Publications; 1980. pp 35–80.

Dowson D. Basic tribology. In: Dowson D, Wright V, eds. *Introduction to the Biomechanics of Joints and Joint Replacement*. London: Mechanical Engineering Publications, Ltd.; 1981. pp 120–145.

Mow VC, Setton LA, Howell DS, Buckwalter JA. Structure-function relationships of articular cartilage and the effects of joint instability and trauma on cartilage function. In: Brandt KD, ed. *Cartilage Changes in Osteoarthritis*. Indianapolis, IN: Ciba-Geigy; 1990. pp 22–42.

Mow VC, Sun DD, Guo XE, Likhitpanichkul M, Lai WM. Fixed negative charges modulate mechanical behavior and electrical signals in articular cartilage under unconfined compression: The triphasic paradigm. In: *Porous Media, Proc Tribute to Professor Reint de Boer*. Berlin: Springer Verlag; 2002. pp 227–247.

See also BIOMECHANICS OF EXERCISE FITNESS; JOINTS, BIOMECHANICS OF; LIGAMENT AND TENDON, PROPERTIES OF.

CATARACT EXTRACTION.

See LENSES, INTRAOCULAR.

CELL COUNTER, BLOOD

YI ZHANG
SRIRAM NEELAMEGHAM
University of Buffalo
Buffalo, New York

INTRODUCTION: NATURE OF BLOOD CELLS (1,2)

Cells compose ~50% of the volume of normal human blood, while plasma constitutes the remaining volume. Generally, cells in blood are divided into three categories: platelets, erythrocytes (or red blood cells, RBCs) and leukocytes (or white blood cells, WBCs) (Table 1) (3). Among these, the platelets or thrombocytes are small, irregular, disk-shaped cells that lack a nucleus. They are of size 2–3 μm in diameter. These cells primarily function to stop bleeding or hemorrhage, and they also participate in coronary artery disease. They do so by being part of the blood coagulation cascade and by aggregating with each other. Platelets are found in blood at a concentration of 0.15–0.5×10^6 cells·mm^{-3}. The second type of cells in blood, erythrocytes, contains a red respiratory protein called hemoglobin. These are disk-shaped, biconcave cells without nuclei. Their diameter ranges from 6 to 8 μm and their thickness is 1.5–2.5 μm. The primary function of erythrocytes is to transport oxygen and carbon dioxide between the lung and body tissues. Erythrocytes are the most numerous blood cells at concentrations of $4–6 \times 10^6$ cells·mm^{-3}. Mature erythrocytes emerge from precursors that are called reticulocytes. Erythrocyte counts are on average ~10% higher in the human adult male population than those in the female population. Lack of iron and hemoglobin in erythrocytes can lead to anemia, a pathological deficiency in the oxygen-carrying component of blood. The third type of blood cells is the leukocytes, whose primary function is to provide the body with immunity and to protect it from infection. Leukocytes are fewer in number than the erythrocytes with a concentration of ~$5–10 \times 10^3$ cells·mm^{-3}. These cells are roughly spherical in shape and they contain nuclei, and considerable internal and cell-surface structures.

Leukocytes are categorized in various ways depending on their function and differentiation pathway. One common method subdivides these cells into myeloid and lymphoid cells. Myeloid cells differentiate into phagocytes, while lymphoid cells primarily produce lymphocytes. The phagocytes include polymorphonuclear granulocytes and monocytes–macrophages. Of these, the former have lobed, irregular shaped (polymorphic) nucleus. They are further subdivided into neutrophils (55–70% of all leukocytes), eosinophils (2–4%), and basophils (0.5–1%), on the basis of how the cellular cytoplasmic granules are stained with acidic and basic dyes. Leukocytes are also commonly characterized based on particular cell-surface receptors that are expressed by them, since these are specific to a particular subpopulation. Many of these receptors are recognized by monoclonal antibodies. A systematic nomenclature has now evolved in which the term CD (Cluster Designation) refers to a group of antibodies that recognize a particular cell-surface antigen. The CD classification is thus often used to classify and identify particular leukocyte subpopulations. A method of blood cell counting called flow cytometry (described below), often uses the fluorescence of labeled antibodies to distinguish between various cell types.

Among the granulocytes, neutrophils are the most abundant cell type. These represent the body's first line of defense during immune response. They are characterized by a number of segmented nucleus lobes connected by fine nuclear strands or filaments. Immature–young neutrophils have a band- or horseshoe-shaped nucleus. Thus, while the younger neutrophils are known as band neutrophils, the mature cells are the segmented neutrophils. Segmented neutrophils are the predominant species in human blood, while band neutrophil levels are elevated following bacterial infection or acute inflammation. The term left shift is used to indicate an increase in the number of circulating immature neutrophils. Condition under which the number of circulating neutrophils is increased is called neutrophilia, while a decrease in this cell type results in neutropenia. Another important morphological characteristic of neutrophils is the virtual lack of endoplasmic reticulum and mitochondria. Mature neutrophils have short lifetimes in circulation and they migrate into tissues to defend against invading microbes during inflammation. Eosinophils, like other granulocytes, possess a polymorphous nucleus, generally with two lobes. They can response to allergy and parasitic infection. They attack large parasites such as helminthes via their C3b receptors. Eosinophils release various substances from

Table 1. Characteristic of Normal Blood Cells[a]

Cell Type	Concentration	Size, μm	Density, g·mL	Shape	Nucleus	Cytoplasm
Platelet(thrombocyte)	$0.15–0.5 \times 10^6 \cdot \mu L^{-1}$	2–3	1.03–1.06	Small disk shape	None	Granular
Erythrocyte (RBC)	$4–6 \times 10^6 \cdot \mu L^{-1}$	6–8	1.09–1.11	Biconcave disk	None	Hemoglobin
Leukocyte (WBC)	$5–10 \times 10^3 \cdot \mu L^{-1}$	8–20	1.05–1.10			
Neutrophil	55–70% of WBC	9–15	1.08–1.10	Various	Lobed	Granular
Eosinophil	2–4% of WBC	9–15	1.08–1.10	Various	Lobed	Granular
Basophil	0.5–1% of WBC	10–16	1.08–1.10	Various	Lobed	Granular
Monocyte	3–8% of WBC	14–20	1.05–1.08	Various	Round	Fine
Lymphocyte	20–40% of WBC	8–16	1.05–1.08	Round	Round	Clear

[a](Adapted from Ref. 3. pp 3–6.)

their eosinophilic granules. These include major basic proteins, plus cationic proteins, peroxidase, phospholipase D and histaminase. The number of eosinophils can augment in blood during allergy (eosinophilia), dermatological disorder and parasitic infection. Basophils have a two-lobe nucleus. They release inflammatory mediators, such as histamine and bradykinin, and prostaglandins and leukotrienes. Basophils play an important role in inflammatory and allergic response. Their number is increased in patients with hypoactive thyroid conditions and during certain malignancies like chronic myeloid leukemia.

Monocytes and macrophages (tissue monocytes) represent the second type of phagocytes and these are relatively large, long-lived cells compared to polymorphonuclear granulocytes. Their cytoplasm is transparent with typically a horseshoe-shape nucleus. Monocytes are involved in both acute and chronic inflammation. The transformation from monocytes to macrophages is controlled by different cytokines. When responding to chemical signals at the inflammation site, monocytes quickly migrate from the blood vessels and start to perform phagocytotic activity. These cells also have an intense secretory activity that results in the production and secretion of chemical mediators such as lysozymes and interferons. The number of monocytes in circulation is increased whenever there is increased amount of cell damage, such as during recovery from infection.

Lymphocytes are mononuclear cells that constitute $\sim$20–40% of all leukocytes. These cells have a round or oval shaped nucleus that is typically large in comparison to the overall cell size. Besides circulating in blood vessels, lymphocytes also populate the lymphoid organs, as well as the lymphatic circulation. The specificity of immune response is due to lymphocytes, since these cells can distinguish between different antigenic determinants. Lymphocytes are subdivided into three main categories: (*1*) T-cells, (*2*) B-cells, and (*3*) natural killer (NK) cells. T-Cells are responsible for cellular immune response and are involved in the regulation of antibody reactions by either helping or suppressing the activation of B lymphocytes. B cells are the primary source of cells responsible for humoral–antibody responses. These are responsible for the production of immune antibodies. The NK cells destroy target cells via nonphagocytic reaction mechanisms that are termed cytotoxic reaction. Lymphocyte number may increase in blood in patients with skin rashes from certain viral diseases such as measles and mumps, in patients with thyrotoxicosis, and in patients recuperating from certain acute infections.

RATIONALE FOR CELL COUNT

Blood cell count is achieved by determining the concentrations and other parameters of different cell types in a unit volume of circulating blood. It can be a complete blood count (CBC, defined later in this article), which examines every blood component, or it may measure only one element. Table 1 demonstrates the characteristic of normal blood cells from a healthy adult. These values vary with age, sex, race, living habit, and health status. Further, under pathological conditions, the distribution of blood cells may be perturbed and thus blood cell counting can aid diagnostics. For example, a typical symptom of common anemia is an inadequate amount of RBCs. The increase or decrease in the numbers of the different types of WBCs may indicate infection and inflammation as discussed above. In addition to the cell numbers, other information regarding blood cells, such as cell size, is also important. For example, in patients with anemia caused by vitamin B_{12} deficiency, the average size of the RBCs is larger than normal and this disease state is called macrocytic anemia. On the contrary, if red blood cells are smaller than normal, as in the case of microcytic anemia, the condition may be indicative of iron deficiency. Therefore, a routine blood test, which includes not only blood cell count but also measurement of other parameters, can aid disease diagnosis and treatment by health professionals.

HISTORY AND BASIC PRINCIPLES FOR BLOOD CELL COUNTING (4–10)

Cell counting has evolved over the centuries from a manual method that heavily relied on microscopic examination, to one where electrical and optical measurement strategies, along with computer automation, are playing an increasingly important role. Indeed, manual methods are still important in research laboratories that study a wide variety of animal systems and cell types, in addition to human blood. On the other hand, automated systems are typically used in clinical studies. In this context, modern technology has automated the process of blood cell counting and the assessment of various blood cell parameters. A CBC, thus, not only provides a panel of tests to quantify the composition of whole blood, it may also include more detailed information regarding the cell profile. Based on technical feasibility and cost, either a simple manually operated cell counter or a more advanced automated blood count platform can be applied to serve the specific medical diagnose need. A brief history and rationale that has lead to current strategies for blood cell counting is outlined below.

Microscopy Coupled with Manual Visualization

Notable, among the early attempts to count blood cells, was the work by Anton van Leeuwenhoek in the seventeenth century who counted the number of chicken erythrocytes in a glass capillary of known dimensions using his microscope. Later, in the nineteenth and early twentieth century, Burker employed a shallow rectangular chamber with a thin coverglass as a counting chamber. Advances in this basic design have now resulted in the laboratory hemocytometer, which is commonly employed for manual cell counting. Ehrlich's classical work on the staining of white blood granules laid the foundations of hematology and differential cell counting. In these studies, he demonstrated that it is possible to distinguish between the various blood cell subgroups using acidic and basic dyes that differentially stained the cellular granules and nucleus.

Hemocytometer

One device that utilizes the light microscope for cell counting is the hemocytometer. This is a commercially

available counting chamber that is used for manual blood counting. It consists of two parts: a microscopic slide with improved Neubauer ruling, and a special thick flat cover slip (Fig. 1a). Both the hemocytometer slide and cover slip must meet specifications of the National Bureau of Standards. The slides have two raised surfaces for duplicate cell counting, each of them bearing square-shaped grids of dimensions 3×3 mm. The two raised surfaces are separated by an H-shaped moat. Each of the 3×3 mm squares has a central area of size 1×1 mm that is further subdivided by 25 groups of 16 smaller squares (Fig. 1b and c). During cell counting, the coverslip is placed on top of the counting surfaces such that the distance between the counting surface and the coverslip is 0.1 mm. Thus, the total volume in the space between the central 1×1 mm area and the coverslip is fixed at 0.1 mm^3. Samples to be studied can be loaded into the chamber using a standard laboratory pipette placed at the point labeled V-slash.

A phase contrast microscope is used to view blood on a hemocytometer slide. In such runs, a sample of diluted blood mixture is placed in a hemocytometer. For a proper count, cells should be evenly distributed. In a white cell count, blood is typically diluted 1:20 in a solution that lyses red cells and stains white cells. Because red cells are so much more numerous than white cells, blood is normally diluted 1:200 for red cell counts. The total number of cells in the central area with fixed volume of 0.1 mm^3 is counted and this measurement is used to estimate the concentration of cells per cubic millimeter (mm^3) according to, cell concentration = number of cells counted $\times$ dilution factor / volume under central grid. For simplicity, instead of counting all the cells in the central 1×1 mm area, counting cells present in a sufficient number of representative squares is also reasonable as long as the acceptable level of accuracy can be ensured. A suitable convention should be applied to avoid counting cells twice, for example, by counting only those cells that touch the top and right-hand margins of a square and omitting cells that touch the bottom and left margins. The World Health Organization (WHO) has recommended methods for the visual determination of WBC count and platelet count using hemocytometer (Recommended methods for the visual determination of WBC count and platelet count. Geneva: World Health Organization, 2000. WHO/DIL/00.3). It describes the detailed sample preparation procedure and counting techniques, and this could be used as a basic protocol for cell counting using the hemocytometer.

Electrooptic Measurements

Advances in electronics and electrooptics in the twentieth century have dramatically simplified blood cell counting and made automation of these processes possible. Some examples of early advances are illustrated in Fig. 2. Panel a describes a method developed in the 1940s where cells

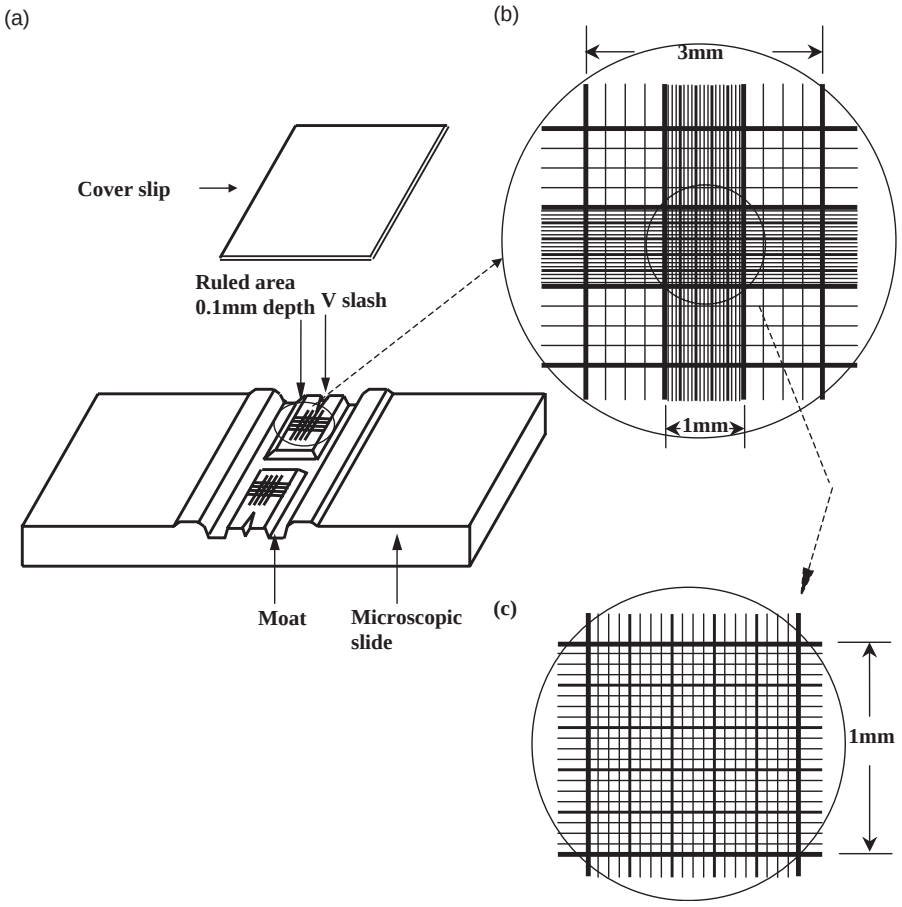


Figure 1. (a) Diagram of hemocytometer with cover slip. (b and c) Expanded view of ruled area as seen under a microscope.

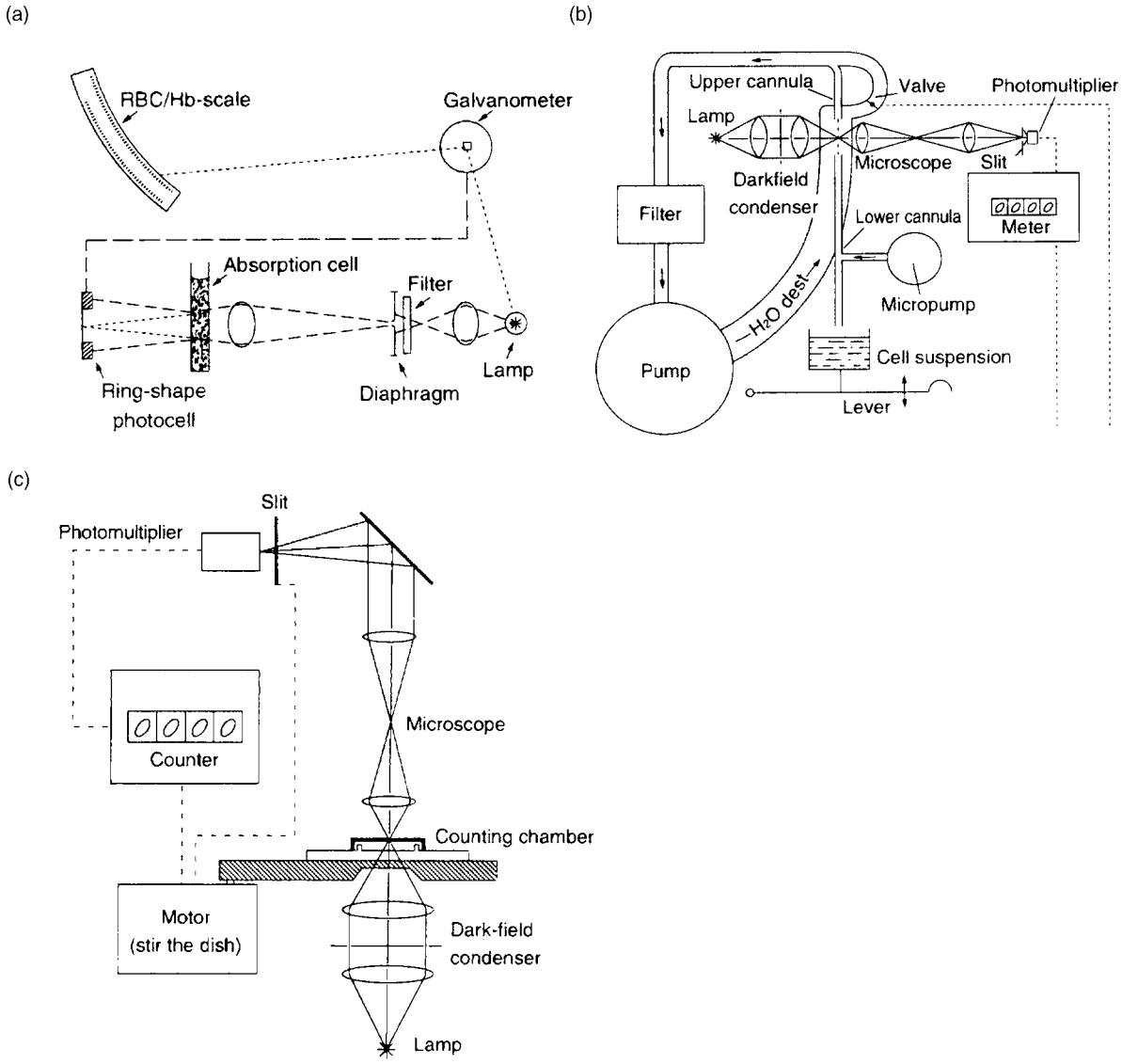


Figure 2. (a) Schematic of an early instrument using the ensemble method to obtain a blood cell count. The intensity of light that is scattered onto a ring-shaped photodetector is measured. The intensity is proportional to cell concentration. (b) Schematic of a photoelectric device that optically counts cells under flow. Here, a fluid stream containing cells is passed through a microscope viewing station. The photomultiplier detects the passage of cells. (c) Schematic diagram of a device using a photoelectric spot scanning method. Mechanical motion is provided to scan cells contained in the counting chamber (3).

were electrooptically measured based on turbidimetry. Here, light lost per millimeter of path length based on scattering or absorbance by blood cells was related to cell concentration. Using cell reference or artificial standards, thus, RBC concentrations could be determined. In this approach, instead of detecting cells individually, the cell concentrations were measured using principles analogous to Beer's law. Panel b illustrates a photoelectric device from 1953, where a thin fluid stream was created such that single cells passed via a microscope viewing station. Images of these cells were magnified and detected using a photomultiplier tube. Panel c illustrates another early instrument where erythrocytes could be counted automatically by means of photoelectric spot-scanning of a thin layer of diluted blood. Here the manual visual counting chamber technique discussed above was improved by introducing a photomultiplier and an electronic counting unit. A motor drives the counting chamber. An instrument based on this principle is the Casella Counter shown in Fig. 2c.

Electronic Cell Counter

In 1950s, Wallace Coulter (Founder of Coulter Company, now Beckman-Coulter Co.) developed a method for cell counting based on electric impedance. This method now forms the basis of most particle size analysis methods in the world. This method, also called low voltage direct

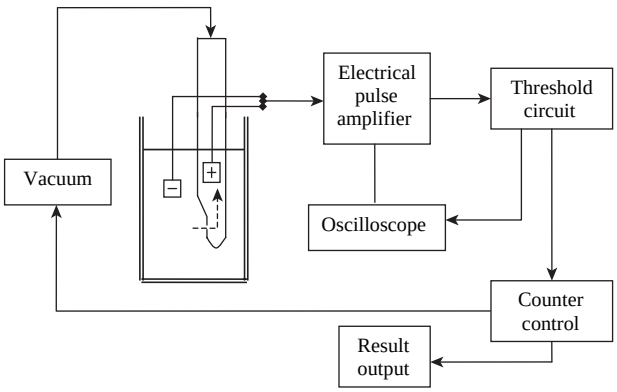


Figure 3. Schematic diagram of electronic cell counter using electric resistance method.

current (dc) method, is based on the measurement of changes in electrical resistance as cells pass through a small orifice that separates two electrodes. In this type of device (Fig. 3), cells are suspended in an electrically conductive diluent, such as saline. Low frequency electrical current is applied between two electrodes; one of them being placed in the cell medium and the second within the aperture tube. The aperture tube has a small orifice or sensing aperture that is typically of size 50–200 μm in diameter. During the measurement, cells are drawn through the aperture using a pressure gradient that is either generated by a mercury manometer or oil displacement pump. Cells are assumed to be non-conducting. Electrical resistance between the two electrodes or impedance in the current occurs as the cells pass through the sensing aperture, causing voltage pulses that are measurable. The number of pulses is proportional to the number of cells counted. The size of the voltage pulse is directly proportional to the size (volume) of the cell. This principle allows discrimination between cells of different sizes. Counting of specific-sized cells is also possible using threshold circuits that cut-off voltage pulses above and below predetermined values. The quantity of suspension drawn through the aperture is precisely controlled to allow the system to count and size particles precisely. Finally, several thousand particles are individually counted within seconds in this device. Measurements are independent of particle shape, color, and density.

Analogous to the above method is the radiofrequency (RF) resistance method where high voltage electromagnetic current is flown between the two electrodes instead of dc. This current circuits the cell membrane lipid layer and penetrates into the cell. While the dc method defines the volume of the cell, changes in conductivity measured using the RF method correlate with the cell's interior structure including the nucleus volume and density, and cytoplasm granule composition. Both dc and RF may be applied simultaneously and this can yield different information about cell size and cellular structure. Such a dual measurement strategy is employed by Sysmex cell counters to quantify the differential leukocyte counts (DLC) as discussed later.

Several factors affect the precision and accuracy of measurements made using the electric impedance methods. First, the aperture size is critical. The instrument is set to count only particles within the proper size range. The upper and lower levels of the size range are called size exclusion limits. Any cell or material larger or smaller than the size exclusion limits will not be counted. Sample must also not contain other material that might erroneously be counted as cells. In practice, erythrocyte and platelet aperture should be smaller than leukocyte aperture in order to increase platelet count sensitivity. Besides the size exclusion limits and aperture size, cell shape and physical properties are also important in determining the shape factor or the ratio of electrically measured volume to the geometric volume. Erythrocytes may result in different signals depending on their orientation with respect to the aperture in the sensing zone. Simultaneous passage of more than one cell at a time through aperture may also cause artificially large pulses, and thus circuits to correct for this coincidence error are required. The magnitude of the coincidence error increases with cell concentration. Correction should be completed by the countercomputer based on the relationship of cell count with cell concentration and aperture size. Finally, an internal cleaning system to prevent or slow down protein buildup in aperture is beneficial in minimizing aperture blockage.

Hydrodynamic focusing as discussed below helps to solve many of the problems above and it provides improved cell counting and characterization. This has been developed and assembled in many cell counters today and this feature dramatically improves the cell volume distribution resolution.

Laser Light Scattering and Fluorescence Detection

Optical scattering can be used alone or in combination with other electrical measurement strategies discussed above for cell counting and characterization. A key feature of such instruments is hydrodynamic focusing where an external sheath flow allows alignment of blood cells one-at-a-time in the path of a light beam, usually within a quartz flow cell (Fig. 4a). Incident light on cells within this flow stream are scattered or redirected in a manner that is dictated by the size of the cell and the intracellular distribution of refractive index. Lasers are generally preferred as the light source since it produces monochromatic light that has a small spot size. Photomultiplier tubes (PMTs) are used to collect the weak signal of scattered light. The light scattered at angles from 5–10° (forward scatter) in general correlates with cell size. Light scattered at 90° is called side-scatter and this is related to the cell shape, orientation, and cellular content. A cell with many complex intracellular organelles will also give a larger side-scatter signal than a cell with fewer intracellular organelles. The design of precise angles where scattered light signals are measured is specific to instrument manufacturers and the cell-enumeration strategy employed. In general, these scatter data together allow reliable identification of distinct populations, such as platelets, RBCs, monocytes, and neutrophils in a mixture. They can also allow enumeration of lymphocyte subsets and reticulocytes. While, optical scattering methods reveal information about cells that is

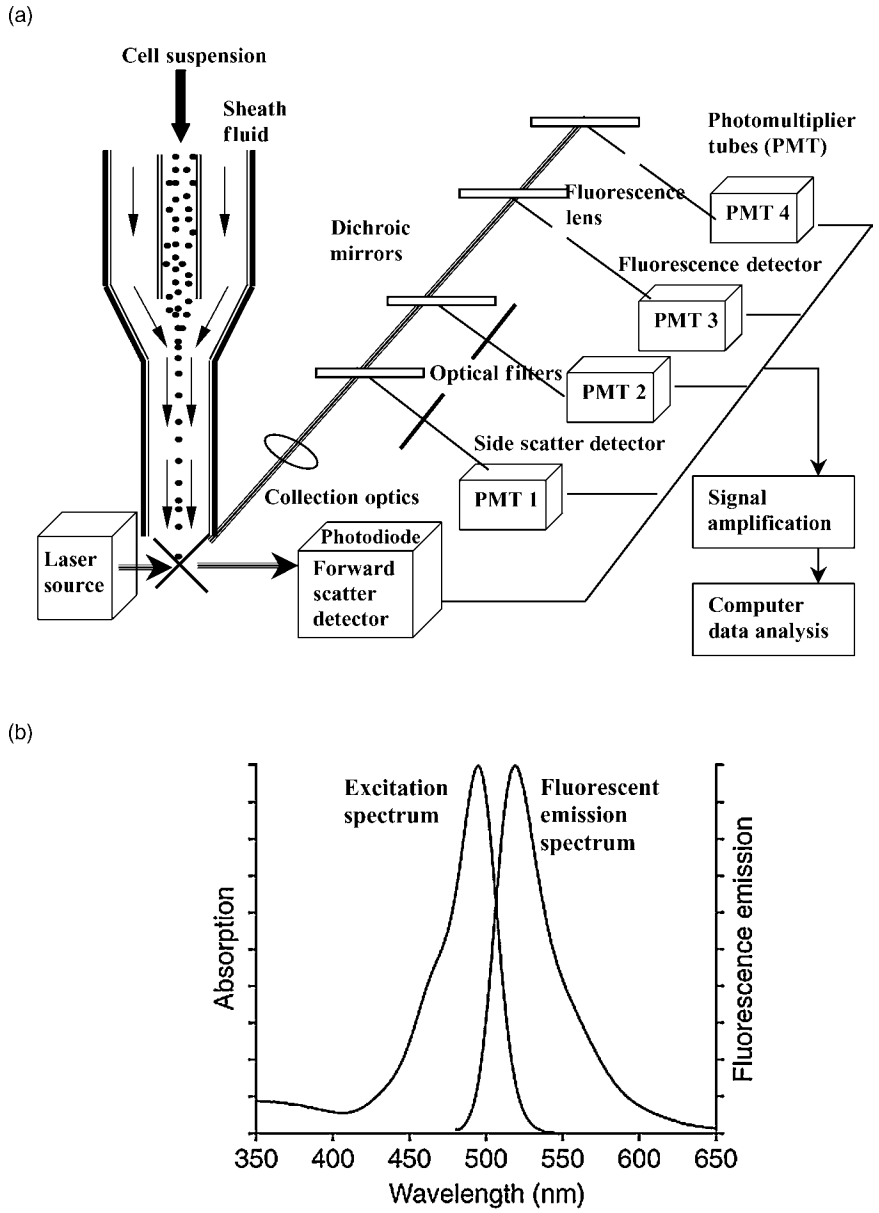


Figure 4. (a) Schematic of flow cytometry showing hydrodynamic focusing of cells by sheath fluid that brings the cells in the path of a laser beam. Light scattered by cell at various angles is collected. These are passed through an arrangement of optical filters to yield measures of forward scatter, side scatter and particle fluorescence. (b) Stokes shift is depicted for the fluorescent probe fluorescein where the wavelength of the absorbed and emitted quanta are shifted, with the emitted wavelength being longer than the absorbed light.

distinct from that obtained from the above electrical methods, their estimates of cell volume are not as accurate as the electrical methods.

A major advantage of optical methods using lasers is that such methods can be readily coupled with fluorescence detection (Fig. 4b). Fluorescent conjugated antibodies to specific cell-surface CD markers or specific ligands can be used not only to identify particular blood cells, but also to label cellular components that may be indicative of disease states. In such work, when the laser light reaches these cells, a fraction of the photons are absorbed by the fluorescent probe, which then reemit the photon at a longer wavelengths. The quantity of this emitted light (both scattered and fluorescent) is measured using photomultiplier tubes that are arranged in conjunction with a series of optical filters and dichroic mirrors as shown in Fig. 4a. The detection and conversion of scattered or fluorescent light into electrical signals is accomplished by photodetectors that capture photons on a light sensitive surface that elicits an electron cascade. The signal output from such detectors is amplified (either linearly or logarithmically) and then converted from analog to digital form for computer analysis. Multidimensional plots of various scattering properties with fluorescent signals can thus be generated to individually characterize each cell in a complex mixture.

COMPLETE BLOOD CELL COUNT (5,10)

Computer Blood Cell Count is a series of tests that result in the quantitation of the number of erythrocytes, leukocytes, and platelets in a volume of blood. The measurements also estimate the hemoglobin content and packed cell volume (or hematocrit) of erythrocytes. This can be done manually using a microscope along with cytochemical dyes, such as Wright–Giemsa stain. The combination of acidic and basic

dyes here can differentially stain the granules, cytoplasm, and nuclei of various blood cell types. Alternatively, in clinical laboratories an automated cell counter can be used to count cells in a given volume. Low end instruments offer RBC and platelet analysis with three-part differential leukocyte count (DLC) while higher end instruments may include a five-part differential count along with reticulocyte analysis. The speed of the instrument and level of automation varies with the class of instrument. The analysis thus obtained is compared with the normal range and assessed for clinical or research purposes. A complete blood cell count mainly includes the following parameters:

Hemoglobin

Hemoglobin concentration (HGB) is reported in grams per deciliter ($g \cdot dL^{-1}$) of blood. This parameter typically varies in proportion to erythrocyte concentration in blood. The normal range for hemoglobin is age and sex dependent. Traditionally, hemoglobin is measured using the cyanmethemoglobin method, as recommended by the International Council for Standardization in Hematology (ICSH). Here, a lysing agent is added to disrupt RBCs and to release cellular hemoglobin. This hemoglobin is converted into a stable form called cyanmethemoglobin (see reaction below), the quantity of which can be measured using a spectrophotometer for absorbance measurement at ~ 540 nm.

$$\mathrm{Hb}(\mathrm{Fe}^{2+}) \xrightarrow{\mathrm{K_3Fe(CN)_6}} \text{methemoglobin}(\mathrm{Fe}^{3+}) \xrightarrow{\mathrm{KCN}} \text{cyanmethemoglobin}$$

Since cyanmethemoglobin measurements contain poisonous cyanide reagent, other more environmentally friendly methods for automated HGB measurement have been developed. Among them, sodium lauryl sulfate-hemoglobin (SLS-Hb) method is used by Sysmex automated cell counters. Here, the lauryl group of the ionic surfactant, which is hydrophobic, binds strongly with hemoglobin. This binding leads to rapid globin molecular conformation change and conversion of hemoglobin from the ferrous (Fe^{2+}) to the ferric (Fe^{3+}) state. The hydrophilic group of SLS now binds with Fe^{3+} to form a stable SLS-Hb. The absorption maximum of SLS-Hb occurs at 535 nm with a shoulder at 560 nm, and this feature is used to determine hemoglobin content. This reaction mechanism is useful since conversion to SLS-Hb occurs rapidly within 10 s.

Platelet Count

Platelet count (PLT) is normally expressed as thousands per microliter (μL) and can be measured manually using the hemocytometer. Care must be taken during such measurements to avoid platelet clumps that can occur in the absence of appropriate anticoagulant. Electronic counting of platelets can also be performed using electric impedance or light scattering methods. Such measurements are typically performed in channels that are designed to discriminate between erythrocytes and platelets. Size distributions resulting from platelet counts can be used to estimate the mean platelet volume (MPV), which is a measure of the platelet volume variation. In general, increased MPV may be expected in regenerative thrombocytopenia, which is accompanied by an increased production of platelets by bone marrow.

Red Blood Cell Count

The RBC or erythrocyte count is expressed in millions per microliter of whole blood. Such counts can be measured manually using the hemocytometer. In hematology analyzers, RBC content is typically measured using either the dc impendence method, light scattering analysis, or a combination of the two. Attention is placed during these measurements to discriminate between small RBCs and platelets. Results of such analysis typically result in a RBC size distribution plot from which other indices can be estimated. These indices include: (*1*) Hematocrit (HCT), which is also called packed cell volume (PCV). This is a measure of the volume fraction of RBCs in whole blood expressed in %vol/vol. Normal adult hematocrit ranges from 35 to 50%, and this is both sex and age dependent. Traditionally, hematocrit is determined by monitoring the height of packed RBCs after centrifugation in a standard microhematocrit tube, relative to the column length. Electronic cell analyzers can also estimate hematocrit by measuring the individual volumes of RBCs (also called MCV as described below) and determining the product of RBC count and MCV. (*2*) Mean corpuscular volume (MCV) is the mean volume of RBCs expressed in femtoliters (fl). The normal range is ~ 80–100 fL. The MCV can be experimentally determined from the RBC size distribution height. Alternatively, if HCT value is known, MCV is calculated based on the ratio of hematocrit and RBC count. This parameter is analogous to MPV, which can be derived from platelet data. When the MCV is low with normal HCT, the blood is said to be microcytic. (*3*) Mean corpuscular hemoglobin concentration (MCHC) is the mean concentration of hemoglobin in the RBCs in grams per deciliter ($g \cdot dL^{-1}$). This is calculated based on the ratio of HGB by HCT. Red cell populations with normal, high, or low values of MCHC are referred to as normochromic, hyperchromic, or hypochromic, respectively. The last case can occur during strongly regenerative anemia, where an increased population of reticulocytes with low HGB content pulls the average value down (an increased MCV would be expected under this scenario). (*4*) Mean corpuscular hemoglobin (MCH) is a measure of the mean mass of hemoglobin (HGB) in RBC, and is expressed in picograms (pg). (*5*) Red cell distribution width (RDW) is an index of the variation in cell volume within the RBC population. It is mathematically determined by (Standard deviation of RBC volume/ MCV) $\times$ 100. The normal range for RDW is 11–15%. While, red cell populations with normal RDW are called homogeneous, those with higher than normal are termed heterogenous. For example, increased number of reticulocytes, which is associated with erythropoiesis, will cause increased RDW values. The RDW index may be an early indicator of changes in red cell population sizes, for example, during anemia caused by iron deficiency. In this case, the presence of few microcytic RBCs may increase the standard deviation of the cell distribution even before marked changes in MCV are observed.

White Blood Cell Count

White blood cell or leukocyte count is measured in thousands per microliter. During manual WBC count, RBCs in blood are lysed and diluted sample is charged into the hemocytometer. Nucleated cells are counted and WBC concentration is determined. Alternatively, impedance-based electronic cell counters can be used to measure WBC count. Besides these basic methods, in automated cell counters, one of many technologies can be applied for WBC differential count. Beckman–Coulter instruments employ the VCS (volume, conductivity, and scattering) technology. In this method, the dc impedance principle is used to physically measure the volume of the cell that displaces the isotonic diluent. Alternating current in the RF range short circuits the bipolar lipid layer of the cell membrane allowing energy penetration into cell. This probe provides information on cell size and internal structure. This data is adjusted by the cell volume measurement to obtain an index called opacity. Finally, coherent light scattering from an incident laser beam is collected to obtain information on cellular granularity and cell surface structure. In Sysmex instruments both dc and RF methods are employed along with differential lysis of cells using lysis solution and temperature treatment. In CELL-DYN instruments from Abbott laboratories, the Multi-Angle Polarization Scattering Separation (M.A.P.S.S.) technology is used to obtain the differential count. Here light scattered by cells localized in a hydrodynamically focused flow stream is measured at three angles (0, 10, and 90°). Polarized light at 90° is also measured. Together these four parameters are used to perform the five-part differential count. Two methods are employed in the Bayer cell counters for differential leukocyte count. In the first method called the peroxidase method, RBCs are lysed and white cells are stained with peroxidase. These cells are counted based on size by forward scatter analysis, and absorbance using dark field optics. The second method, called the basophil method, involves stripping the cells using a non-ionic surfactant in acidic solution. Basophils are resistant to lysis while RBCs and platelets are lysed and other leukocytes are stripped of their cytoplasm. Light scattering analysis distinguishes basophils from other polymorphonuclear and mononuclear cells. The above peroxidase and basophil methods thus provide automated differential cell count by separating the cells into clusters.

Reticulocyte Count (RTC, RET, or RETIC)

Reticulocytes are formed in the last stages of erythropoiesis. These cells spend ~2 days in the bone marrow and 1–2 days in peripheral blood prior to maturing into RBCs. These are nonnucleated RBC, which by definition upon staining with supravital dyes contain two or more particles of blue-stained material that correspond to ribosomal RNA (ribonucleic acid). With new methylene blue, reticulocytes stain bluish-purple. Reticulocyte count as a percentage of RBCs is a measure of the erythropoietic activity in the bone marrow. This is a useful marker of bone marrow suppression following chemotherapy, recovery from anemia, and so on. Reticulocyte counts may be high when the body is replenishing the RBCs in circulation. Reiculocyte counts can be performed using microscopy and supravital stains, such as new methylene blue or brilliant cresyl blue. Reticulocyte counts can also be done using automated instruments. Here light scattering is typically applied to detect cell size and cell fluorescence–absorbance measurements in conjunction with dyes like Auramine O and new methylene blue for quantitation of reticulocytes. Such methods provide good discrimination between reticulocytes and mature RBCs, with greater accuracy than microscopy examination.

AUTOMATED CELL COUNTERS (4,5)

Manufacturers of automated cell counters typically present a vast product line with varying levels of sophistication to meet the market needs. Although the analysis principles may differ, all cell counters have some common basic components, specifically hydraulics, pneumatics and electrical systems (Fig. 5). Among these, the hydraulic

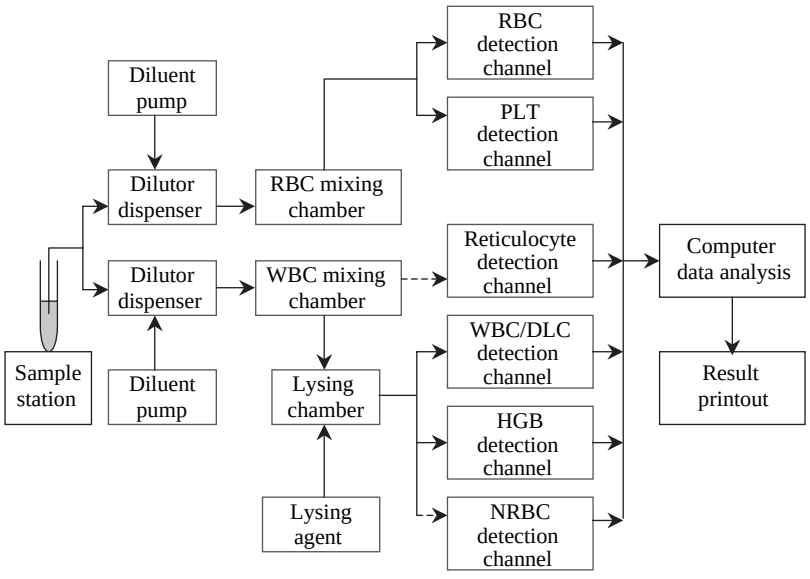


Figure 5. Flow diagram of an automated multichannel cell counter. (Adapted from Ref. 3).

system is designed to dispense, dilute and mix samples prior to analysis. The pneumatic system operates various valves and drives the sample through the hydraulic system. The electrical system controls the operation sequences including optical–electrical detection of signals and computer-assisted data analysis. Instrument electronic analyzers typically have at least two channels. In one channel a diluent is added and RBCs are counted and sized. In the second, lysing agent is added to remove red blood cells and leave WBC intact for counting. These also produce a solution in which hemoglobin can be measured. Platelet count may be performed in either of these two channels or in a different channel. Normally, a separate channel will be required for reticulocyte count measurement. Analysis of a single blood specimen can be performed rapidly within 1 min, and results are presented in the form of numerical tables, histograms, or cytograms. The degree of analysis is both software and user dependent. Upon comparison with standard values, the software may also place flags on the output data that indicate either potential problem with analysis or deviation from cell count characterization of normal controls.

Numerous companies manufacture automated cell counters. Table 2 presents the characteristics of four high end instruments manufactured by some of them.

The Beckman–Coulter LH750 (Beckman Coulter Inc., Fullerton, CA) is a new instrument that provides CBC and five-part DLC. Additionally, it provides automated detection of subpopulations of pathological cells, such as immature granulocytes and atypical lymphocytes. It uses the three-dimensional (3D) Volume, Conductivity, Scatter (VCS) technology to probe hydrodynamically focused cells. A helium–neon laser and multiangle light scattering analysis provide information about cellular internal structure, granularity and surface morphology.

Abbott Cell-DYN 4000 (Abbott Laboratories, Abbott Park, IL) is capable of providing 41 parameters, including fully automated reticulocyte and immature granulocyte count. It uses four-angle argon-laser light scattering (M.A.P.S.S. technology) and two-color fluorescence flow cytometry (two fluorescence emission laser optics) to perform automated leukocyte counts, reticulocyte count, and DLC analysis. Both hydrodynamically focused impedance count and optical method are used for optimal erythrocyte and platelet size distribution analysis. Hemoglobin concentration is measured in a separate sample aliquot based on spectrophotometry. Immature granulocyte and variant lymphocytes are detected by a multiparameter, multiweighted discriminant function: This function generates a flag and reports a confidence fraction (i.e., the probability that these cells are classified correctly).

Sysmex XE-2100 (Sysmex Corporation, Japan) provides analysis of 32 parameters including simultaneous WBC, five-part DLC, human progenitor cell and reticulocyte analysis. Using flow cytometry with a semiconductor laser, RF, and dc measurements, this instrument analyzes the size and the structural complexity of cells. Selective dyes and reagent assist in differentiating the WBC, nucleated RBCs and reticulocyte. The RBC and platelet counts are measured using sheath flow dc detection method. Hemoglobin concentration is measured using a non-cyanide hemoglobin method.

The Bayer ADVIA 120 hematology system (Bayer Diagnostics, Tarrytown, NY) is an automated analyzer with four independent measurement channels. The peroxidase [PEROX and basophil-lobularity (BASO)] channels determine WBC and DLC count. Hemoglobin channel is used to measure HGB. The last channel is the RBC/PLT channel that provides information on platelet activation in addition to measuring PLT and RBC indices. This instrument measures the intensity of light scattered by platelets at low angles (2–3°) to obtain cell volume/size data and high angles (5–15°) for information on internal complexity. From these paired intensities the instrument computes platelet volume (MPV) and platelet component concentration on a cell-by-cell basis. The mean platelet component concentration (MPC) is indicative of platelet activation state. Mean platelet mass can also be computed from the MPV and MPC.

CONCLUDING REMARKS

This article discussed the basic principles of hematology with emphasis on humans. Enumeration of cell population

Table 2. Characteristics of Hematology Analyzers[a]

Instrument	Beckman–Coulter LH 750	Abbott Cell-Dyn 4000	Sysmex XE-2100	Bayer ADVIA 120
Number of parameters	28	41	32	30
HGB	Modified cyanmethemoglobin method	Spectrophotometry	Non-cyanide hemoglobin method	Modified cyanmethemoglobin method
Platelet	VCS	Optical method and impedance count	Hydrodynamic focusing with dc detection	Light scattering
RBC	VCS	Impedance count and optical method	Hydrodynamic focusing with dc detection	Light scattering
WBC and DLC	Five-part DLC VCS technology	Five-part DLC Light scatter and fluorescence flow cytometry	Five-part DLC Flow cytometry, RF and dc detection	Five-part DLC Peroxidase staining optics system, light scattering
Reticulocyte Count	New Methylene blue staining and VCS	Fluorescent dye CD4K530 staining and flow cytometry	Auramine O staining, light scattering, flow cytometry	Oxazin 750 staining and optical scatter

[a]Adapted from Ref. 3.

distribution in peripheral blood is examined using optical, electrooptical and light scattering techniques. As seen, such experimental modalities can be automated and the resulting hematology analyzers can be used for clinical application. Even though the exact strategy of cell counting varies between various manufacturers of automated cell counters, performance standards for such instrumentation have been established by the National Committee for Clinical Laboratory Standard (NCCLS) and the International Council for Standardization in Hematology (ICSH). The parameters evaluated here include (*1*) accuracy in measurement within a single batch and between batches of blood samples; (*2*) carryover of parameters between consecutive samples; (*3*) linearity or the ability to get similar measurements when the sample is diluted to different levels before being read; and (*4*) clinical sensitivity or the specificity and efficiency with which flags are generated during analysis to detect abnormal readouts. In order to evaluate the above and to tune the instrument for higher accuracy and sensitivity, blood count calibrators are also available from instrument manufacturers. Suitable preparations of preserved blood can also be made by individual laboratories as described by WHO document LAB/97.2 (Calibration and control of basic blood cell counters. Geneva: World Health Organization, 1997. WHO/DIL/97.2). Besides automated counting, manual and semiautomated methods are also applied by research laboratories. Establishment of such methods requires optimization of blood anticoagulant [ethylenedramenatetraacetic acid (EDTA), heparin, or sodium citrate typically], definition of appropriate electrolyte for sample dilution and design and optimization of lysis reagents required for specific experimental systems.

While the last 50 years have seen the automation of blood counting using hematology analyzers, a plethora of cell-specific antibodies have also been developed more recently. While some of these reagents are already being applied in the modern blood analyzer, their application may increase in the future. Such development can not only increase the range of parameters measured by the analyzer, they can also improve the accuracy and sensitivity of today's instrumentation.

BIBLIOGRAPHY

1. Armitage JO, editor. Atlas of Clinical Hematology. 2004; Philadelphia: Lippincott Williams & Wilkins; p 266.
2. Stiene-Martin EA, Lotspeich-Steininger CA, Koepke JA, editors. Clinical Hematology: Principles, Procedures, Correlations. 2nd ed. 1998; Philadelphia: Lippincott Williams & Wilkins Publishers; p 817.
3. Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation, 4 Volume Set. 1988; New York: John Wiley & Sons; p 3022.
4. Rodak BF. Hematology: Clinical Principles and Applications. 2nd ed. 2002; Philadelphia: WB Saunders.
5. Bain BJ. Blood Cells A Practical Guide. 3rd ed. 2002; Oxford: Blackwell Science Ltd.
6. Fujimoto K. Principles of Measurement in Hematology Analyzers Manufactured by Sysmex Corporation. Sysmex J Iner 1999;9(1):31–44.
7. Groner W, Kanter R. Optical Technology in Blood Cell Technology. Sysmex J Iner 1999;9(1):21–30.
8. Shapiro HM. Practical Flow Cytometry. 4th ed. 2003; New York: John Wiley & Sons, Inc.; p 736.
9. Tatsumi N et al. Principle of Blood Cell Counter-Development of Electric Impedance Method. Sysmex J Iner 1999;9(1):8–20.
10. Hamaguchi Y. Overview of the Principles of Sysmex's Hemoglobinometry. Sysmex J Iner 1999;9(1):45–51.

Further Reading

Lotspeich-Steininger CA, Stiene-Martin EA, Koepke JA. Clinical Hematology: Principles, Procedures, Correlations. 1992; Philadelphia: Lippincott. xix; p 757.

Carr JH, Rodak BF. Clinical Hematology Atlas. 2nd ed. 2004; St. Louis (MO): Elsevier Saunders.

Brown BA. Hematology: Principles and Procedure. 5th ed. 1988; Philadelphia: Lea & Febiger.

See also ANALYTICAL METHODS, AUTOMATED; BLOOD COLLECTION AND PROCESSING; CYTOLOGY, AUTOMATED; DIFFERENTIAL COUNTS, AUTOMATED.

CELLULAR IMAGING

AMMASI PERIASAMY
University of Virginia
Charlottesville, Virginia

INTRODUCTION

For decades, autoradiography has been used widely to follow the synthesis of macromolecules by using radioactive isotopes (1). Interpretation of autoradiograms depends on knowledge of biochemical pathways and precursors and are carefully chosen so that they are used by the cell to build only one kind of molecule. On the other hand, light microscopy techniques have become a powerful tool for cell biologists to study cells live or fixed noninvasively (2–4). Fixed cells can also be studied using electron microscopy, which provides higher resolution than the light microscopy system (5,6). However, the light microscopy system allows studying live cells in physiological conditions.

The microscope has been an essential tool found in virtually every biological laboratory after the observation and description of protozoa, bacteria, spermatozoa, and red blood cells by Antoni van Leeuwenhoek, in the 1670s (7,8). The ability to study the development, organization, and function of unicellular and higher organisms and to investigate structures and mechanisms at the microscopic level has allowed scientists to better grasp the often misunderstood relationship between microscopic and macroscopic behavior. Further, the microscope preserves temporal and spatial relationships that are frequently lost in traditional biochemical techniques and gives two- (2D) or three-dimensional (3D) resolution that other laboratory methods cannot. The benefits of fluorescence microscopy techniques are also numerous (3,9). The inherent specificity and sensitivity of fluorescence, the high temporal, spatial, and 3D resolution that is possible, and the enhancement of contrast resulting from detection of an absolute rather than relative signal (i.e., unlabeled features do not emit) are

several advantages of fluorescence techniques. Additionally, the plethora of well-described spectroscopic techniques providing different types of information, and the commercial availability of fluorescent probes, many of which exhibit an environment- or analytic-sensitive response, broaden the range of possible applications. Recent advancements in light sources, detection systems, data acquisition methods, and image enhancement, analysis, and display methods have further broadened the applications in which fluorescence microscopy can successfully be applied (2,3). Particularly, the fluorescent probes can be used to target many cellular components to follow the cell signaling in space (nm to m) and time (ns to days).

There are a number of microscopic techniques that have been established for cellular imaging including transmitted light–differential interference contrast microscopy (DIC)–phase contrast, reflection contrast microscopy, polarization microscopy, luminescence microscopy, and fluorescence microscopy (2,3,10). Fluorescence microscopy has been categorized into wide-field fluorescence microscopy, laser scanning confocal microscopy, multiphoton excitation microscopy, Förster (or fluorescence) resonance energy-transfer (FRET) microscopy, fluorescence lifetime imaging (FLIM) microscopy, fluorescence correlation spectroscopy (FCS), total internal reflection fluorescence (TIRF) microscopy, and fluorescence recovery after photobleaching (FRAP) microscopy (2–4,10–15). Some of the other advanced microscopy techniques include near-field microscopy (16,17), atomic force microscopy (18), scanning force–probe microscopy (19), X-ray microscopy (20), and Raman microscopy (21,22). In this article, selected fluorescence microscopy techniques (see Fig. 1) used for cellular imaging such as wide-field, confocal, multiphoton, FRET, FLIM, and CARS microscopy with biological examples are described.

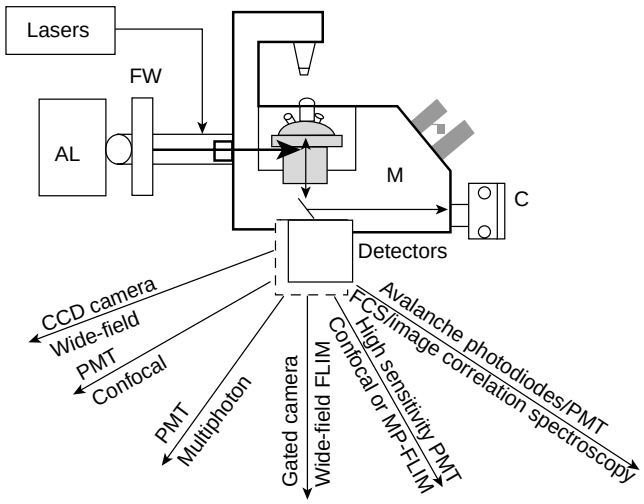


Figure 1. Illustration of various fluorescence microscopy techniques that could be coupled to any upright or inverted epifluorescent microscope. The respective instrumentations are described in the literature. Wide-field (23); confocal (2); multiphoton (3); wide-field FLIM (24); confocal FLIM (25); MP-FLIM (26); FCS and image correlation spectroscopy (27,28).

BASICS OF FLUORESCENCE

Fluorescence is one of the many different luminescence processes in which molecules emit light. Fluorescence is the emission of light from the excited singlet state. Since this type of transition is usually allowed within the molecular orbitals, the emission rates of fluorescence are in the order of 10^8 s^{-1}, and fluorescence lifetimes are in nanoseconds. In contrast, phosphorescence is the emission of light from the triplet excited state and this transition is typically forbidden. The emission rates are much in the order of 10^0–10^3 s^{-1}, and phosphorescence lifetimes are typically milliseconds to seconds.

The excitation of molecules by light occurs via the interaction of molecular dipole transition moments with the electric field of the light and, to a much lesser extent, interaction with the magnetic field. The fluorescence processes following light absorption and emission are usually illustrated by a Jabłoński diagram shown in Fig. 2. Examination of the Jabłoński diagram in Fig. 2 reveals that the energy of the emitted photon is typically less then that of the absorbed photon. Hence, the fluorescence occurs at lower energy (longer wavelength) and this process is called Stokes' shift. The reasons for the Stokes' shift are rapid transition to the lowest vibrational energy level of the excited state S_1, and decay of the fluorophore to a higher vibrational level of S_0. The excess of the excitation energy is typically converted to the thermal energy.

Very intense radiation fields, such as those produced by ultrafast femtosecond lasers, can cause simultaneous absorption of two or more photons (two-photon, three-photon absorption, etc.). This phenomenon was originally predicted by Maria Göppert-Mayer in 1931 (30). It is important to realize that fluorescence intensity resulting from one- and two-photon excitation has a different dependence on excitation light intensity (power). Consequently, the fluorescence intensity depends on the squared laser power for two-photon excitation, the cubed (3rd) power for three-photon excitation, and the fourth power for four-photons excitation. This very strong dependence of fluorescence signal on the excitation power is frequently used to control the mode of excitation.

FLUOROPHORES AND FLUORESCENCE MICROSCOPY

Fluorescence microscopy (FM) plays a vital role in the biological and biomedical sciences, where fluorescence

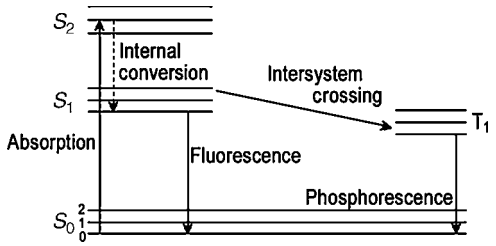


Figure 2. Jabłoński energy level diagram. S_0, S_1, and S_2 are singlet ground, first, and second electronic states, respectively; T1 = triplet state (4,29).

probe specificity and sensitivity can provide important information regarding the biochemical, biophysical, and structural status of cells. The continuing development of fluorescent probes, such as various mutant forms of green fluorescent proteins (GFPs) or fluorophores in conjunction with the strong emergence over the past two decades of confocal and multiphoton microscopy (and specialized applications, such as FRAP, FRET, and FLIM), has been a major contributor to our understanding of dynamic processes in cells and tissue (3,4,12,30–34).

Florescence microscopy can be applied noninvasively to the study of living cells in tissue down to detection levels corresponding to single molecules. Fluorescent probes bound to cellular components with monoclonal antibodies, specific ligand affinities, or covalent bonds allow us to measure chemical properties, such as ion concentrations, membrane potential, and enzymatic activity (35). They allow the experimenter to observe the distribution and function of macromolecules (proteins, lipids, nucleic acids) in living cells and tissues. Techniques have been developed that allow the investigator to place, in cells and tissues, chemically blocked (caged) molecules that can be released or activated (uncaged) by a pulse of light (photolysis) (36). Therefore, a variety of ions, metabolites, drugs, and peptides can be released at carefully controlled times and locations within the specimen. Fluorescent probes and reagents are available from Amersham Pharmacia Biotech, Calbiochem, Fluka, Jackson ImmunoResearch Laboratories, Molecular Probes, Polysciences, Serotec, Sigma-Aldrich, and others. Fluorescent proteins (GFP) and GFP vectors are available from Clontech Laboratories, Quantum Biotechnologies and Life Technologies. Details regarding the selection of fluorophores, labeling, and loading conditions for live cell imaging have been described in the literature (37).

Wide-Field Fluorescence Microscopy

Wide-field fluorescence microscopy is a conventional fluorescence microscope equipped with a movable *xyz*-axis stage that permits imaging of the specimen at different focus and lateral positions, a higher quantum efficiency CCD camera for quantification of the light emitted by the specimen at different spectra, excitation and emission filter wheels, and an appropriate software package that is capable of synchronizing hardware, acquiring images, and correcting them for distortions and information loss inherent in the imaging process (22). To allow simultaneous monitoring of spectral emissions at two or three wavelengths, a dichroic, double, or triple pass filter is used that reflects the respective excitation wavelength to excite the double- or triple-labeled cells and transmit the respective emission bands (www.chromatech.com; www.omegaoptical.com).

Wide-field microscopy is the simplest and most widely used technique. It is used for quantitative comparisons of cellular compartments and time-lapse studies for cell motility, intracellular mechanics, and molecular movement (www.api.com). For example, new fluorescent indicators have allowed the measurement of Ca^{2+} signals in the cytosol and organelles that are often localized (38,39) and nondestructive imaging of dynamic protein tyrosine kinase activities in single living cells (40). This microscope has also been used for localizing protein molecules in living cells (22,41–43). Moreover, it is essential to implement digital deconvolution approaches to remove the out-of-focus information from the images collected in wide-field microscopy (22) (www.api.com).

Laser Scanning Confocal Microscopy (LSCM)

Wide-field microscopy, however, suffers from a major drawback due to the generation of out-of-focus fluorescent signals. Laser scanning confocal microscopy (LSCM) provides the advantage of rejecting out-of-focus information, and also allows associations occurring inside the cell to be localized in three dimensions. A confocal image with improved lateral resolution yields a wealth of spectral information with several advantages over a wide-field image including controllable depth of field and the ability to collect serial optical sections from thick specimens. Owing to its nanometer depth resolution and nonintrusiveness, confocal provides a new approach to measure viscoelasticity and biochemical responses of living cells and real-time monitoring of cell membrane motion in natural environments (2). The LSCM has been widely used in many biological applications, such as calcium, pH, and membrane potential imaging (2,35).

Confocal microscopy was introduced in 1957. Since then, the technique has gained momentum, particularly after the invention of lasers in the 1960s. Commercially available LSCM generate a clear, thin image (512×512) within 1–3 s or less, free from out-of-focus information. A single diffraction-limited spot of laser or arc lamp light is projected on the specimen using a high numerical aperture objective lens. The light reflected or fluorescence emitted by the specimen is then collected by the objective and focused upon a pinhole aperture where the signal is detected by a photomultiplier tube (PMT). Light originating from above or below the image plane strikes the walls of the pinhole and is not transmitted to the detector (see Fig. 3). To generate a 2D image, the laser beam is scanned across the specimen pixel-by-pixel. To produce an image using LSCM, the laser beam must be moved in a regular 2D raster scan across the specimen. Also, the instantaneous response of the photomultiplier must be displayed with equivalent spatial resolution and relative brightness at all points on the synchronously scanned phosphor screen of a CRT monitor. For a 3D projection of a specimen, one needs to collect a series of images at different *z*-axis planes. The vertical spatial resolution is ~ 0.5 μm for a 40×1.3 NA objective; for lenses with higher magnification, the vertical spatial resolution is even smaller. Three-dimensional image reconstruction can be accomplished with many commercially available software systems. Another alternative is a commercially available spinning disk based confocal microscope that can be used for cellular imaging (44) (www.perkinelmer.com).

The LSCM has been widely used in many biological applications and as an example here we describe protein localization using Förster resonance energy transfer (FRET) (4,43,45–48). FRET is a distance-dependent physical process by which energy is transferred nonradiatively from an excited molecular fluorophore (the donor) to

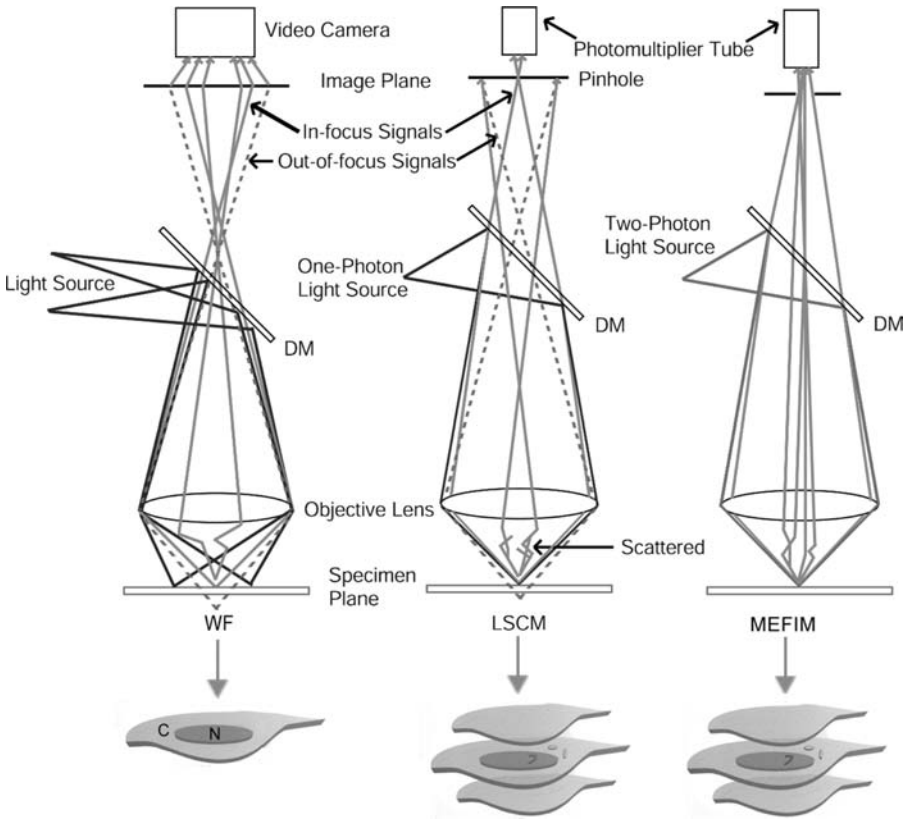


Figure 3. Illumination and detector configuration for wide-field, confocal, and multiphoton microscopy systems. DM = dichroic mirror, WF = wide-field, LSCM = laser scanning confocal microscope, MEFIM = multiphoton excitation fluorescence imaging microscopy, N = nucleus, C = Cytoplasm (22).

another fluorophore (the acceptor) by means of intermolecular long-range dipole–dipole coupling. It can be an accurate measurement of molecular proximity at nanometer distances (1–10 nm) and highly efficient if the donor and acceptor are positioned within the Förster radius (the distance at which half the excitation energy of the donor is transferred to the acceptor, typically 3–6 nm). The efficiency of FRET is dependent on the inverse sixth power of intermolecular separation (29,49,50) making it a sensitive technique for investigating a variety of biological phenomena that produce changes in molecular proximity (51). As an example Fig. 4 shows acquisition and data analysis for localization of CFP- and YFP-C/EBPα proteins expressed in live mouse pituitary GHFT1-5 cell nucleus.

Multiphoton Excitation Microscopy

The instrumentation configuration of multiphoton excitation microscopy (MEM) is generally the same as the LSCM with the exceptions of the excitation light source and the optics. In the LSCM, a visible or ultraviolet (UV) light source is used and an infrared (IR) light source is used for MEM system [see Fig. 3; (52)]. In one-photon (wide-field or confocal) fluorescence microscopy, the absorption of laser energy excites the fluorescent molecules to a higher energy level and results in the emission of one-photon fluorescence. The fluorescence intensity increases at a linear rate with the excitation intensity. Typically, some of the absorbed light energy is dissipated as heat, so the emission wavelength is longer than the absorption wavelength. For example, a fluorophore might absorb one photon at 365 nm and fluoresce at a blue wavelength $\sim$420 nm.

The fluorophores exhibit two-photon absorption at approximately twice (730 nm) their one-photon absorption wavelengths, while two-photon (2p) emission is the same as that of one photon (420 nm), allowing the specimen to be imaged in the visible spectrum. When an IR laser beam is focused on a specimen, it illuminates at a single point and the fluorescence emission is localized to the vicinity of the focal point. The fluorescence intensity then falls off rapidly in the lateral and axial direction. In one-photon (1p) microscopy, illumination occurs throughout the excitation beam path, in an hourglass-shaped path (22). This results in absorption along the excitation beam path, giving rise to substantial fluorescence emission both below and above the focal plane. Excitation from other focal planes contributes to photobleaching and photodamage in the specimen planes that are not involved in imaging. The IR illumination in 2p excitation also penetrates deeper into the specimen than visible light excitation due to its higher energy, making it ideal for cellular imaging involving depth penetration through thick sections of tissue.

Two-photon absorption was theoretically predicted by Göeppert-Mayer in 1931 and was experimentally observed for the first time in 1961 using a ruby laser as the light source (30,53). Denk and others have experimentally demonstrated 2p imaging in a laser scanning confocal microscopy (54). Two-photon excitation occurs when two photons of $h\omega$ and $h\omega'$ are absorbed simultaneously and a molecule is excited to the state of energy $E = h\omega + h'\omega'$ (h = Planck's constant, ω = frequency). The probability that

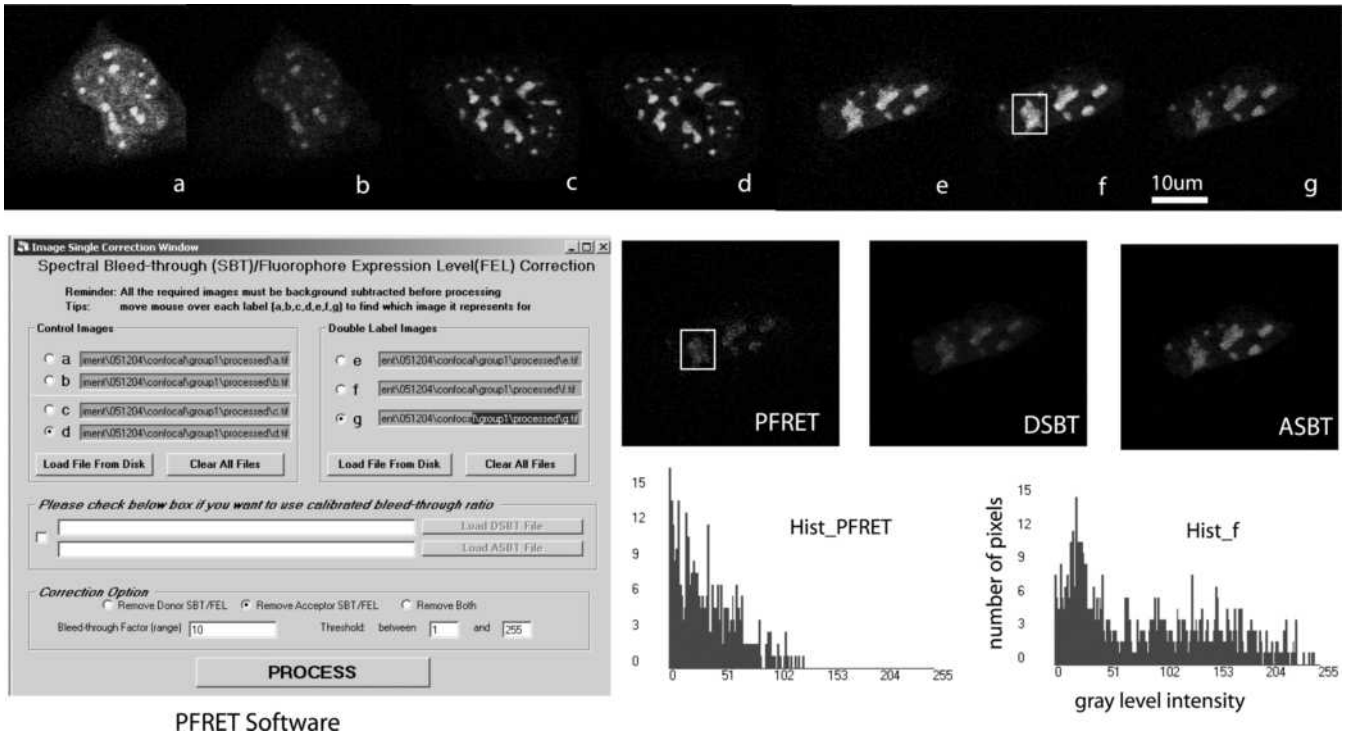


Figure 4. Localization of CFP- and YFP-C/EBPα proteins expressed in live mouse pituitary GHFT1-5 cells studied using confocal-FRET microscopy. Seven images (a–g) are required to remove the contamination in the FRET image (f). The PFRET (processed FRET) image was obtained after removing the donor (DSBT) and acceptor (ASBT) spectral bleedthrough using the PFRET software (shown on the left panel, www.circusoft.com). The spectral bleedthrough varies depending on the excitation power for the donor and acceptor molecules. The respective histogram for the processed (Hist_PFRET) and the contaminated FRET (Hist_f) demonstrates the importance of removing the spectral bleedthrough signals. The energy-transfer efficiency ($E = 20\%$) was estimated after implementing the detector spectral sensitivity correction for the donor and acceptor channel (43).

2p absorption will occur depends on the colocalization of two photons within the absorption cross-section of the fluorophore. The rate of excitation is proportional to the square of the instantaneous intensity. This extremely high local instantaneous intensity is produced by the combination of diffraction-limited focusing of a single laser beam in the specimen plane and the temporal concentration of a femtosecond (fs) mode-locked laser (typically of the order of 10^{-50}–10^{-49} $cm^4 \cdot s^{-1}$/photon^{-1}/molecule) (55). Three- or four photon (or multiphoton) is the extension of two-photon excitation (56).

Two-photon excitation microscopy has been widely used in the area of biomedical sciences including tissue engineering, protein–protein interactions, cell, neuron, molecular, and developmental biology (3,13,22,57–60). Here, we demonstrate as an example the importance of MEM in drug molecule cellular uptake, where MEM is the ideal system for monitoring cellular drug uptake. The separation between excitation and emission wavelengths is considerably more than the 1p (wide-field and confocal) excitation and emission. For example, the excitation for the YK-II-140 drug molecule is 416 nm and emission is at 528 nm and a Stokes shift is $\sim$ 112 nm. In the case of MEM, the excitation for the same drug molecule is 770 nm and the Stokes shift separation is wider than 112 nm. Moreover, in MEM we were able to detect 100 μM drug cellular uptakes compared to 1.0 mM in the wide-field microscopy. The sensitivity of drug detection is improved largely due to the advantage of the MEM (see Fig. 5 for details).

Spectral Imaging Microscopy

Human color vision is a form of imaging spectroscopy, by which we determine the intensity and proportion of wavelengths present in our environment. Spectral imaging improves on the eye in that it can break up the light content of an image not just into red, green, and blue, but into an arbitrarily large number of wavelength classes. Furthermore, it can extend the range to include the invisible UV and IR regions of the spectrum denied to the unaided eye; this type of imaging is usually known as hyperspectral (61). The result of (hyper) spectral imaging is a data set, known as a data cube, in which spectral information is present at every picture-element (pixel) of a digitally acquired image. Integration of spectral and spatial data in scene analysis remains a challenge.

These multispectral imaging approaches have been used to analyze multiple dyes within a sample. Recently,

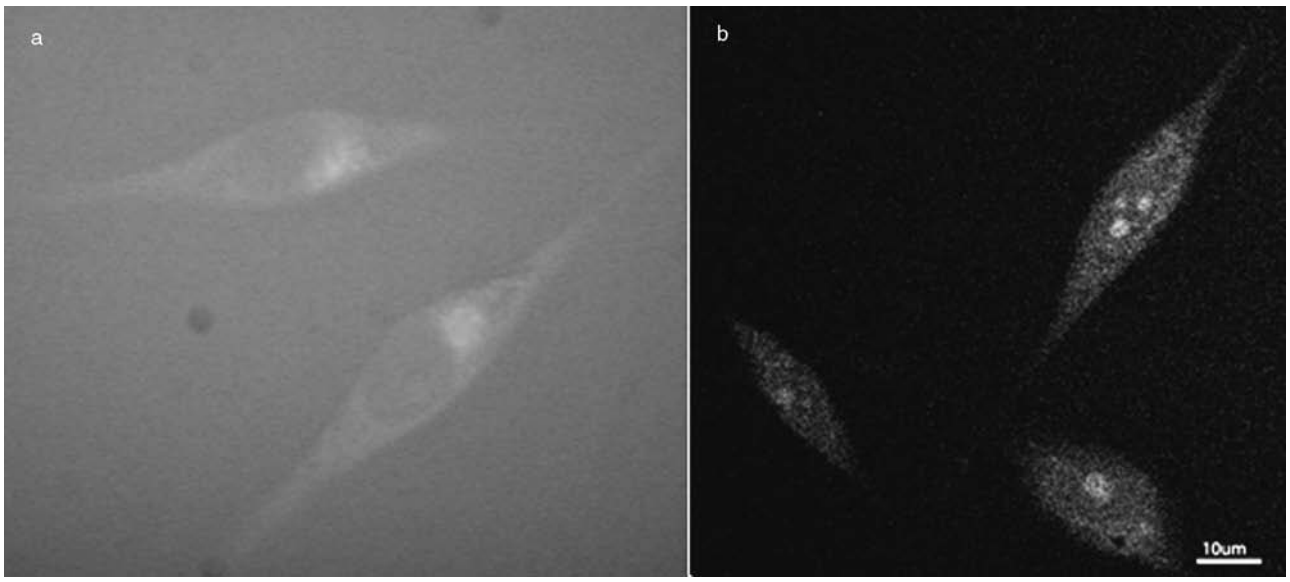


Figure 5. Comparison of one- and two-photon excitation of a living PC-3 cell loaded with YK-II-140 anticancer drug (1.0 mM concentration). Wide-field microscopy provides more autofluorescence from the cell and media (a) compared to the two-photon microscopy (b) The less autofluorescence improves the detection sensitivity of the drug cellular uptake. Wide-field (Ex 416 nm and Em 528 nm). MEFIM or two-photon (Ex 770 nm and Em 528 nm). Biorad Radiance 2100 confocal–multiphoton microscopy was used for the data acquisition.

Carl Zeiss (www.zeiss.de) introduced the Laser Scanning Microscope (LSM) 510 META system with the revolutionary emission fingerprinting technique permitting the clean separation of several even spectrally overlapping fluorescence signals of a specimen (62). The number of dyes that can be used and detected in the experiment is almost unlimited. The new system overcomes the limits of existing detection methods and permits both qualitative and quantitative analyses quickly and precisely *in vitro* and *in vivo*. Furthermore, it is beneficial in many cases for the elimination of unwanted signals, such as background noise or autofluorescence.

The Zeiss 510 Meta system scan head contains two conventional photomultiplier tube detectors (PMT), where the wavelength of the emission light is selected by means of either bandwidth or long passes filters. In the third detector, emission light is passed through a prism and the resulting spectrum is projected onto a detector consisting of a linear array of 32 PMTs, thus enabling the spectral detector to detect a full emission spectrum from a given fluorophore (www.zeiss.com). The advantage of detecting a broad spectrum of emissions is fully realized by the process of linear unmixing (63). This is an image analysis technique that is intrinsic to the LSCM controlling software that compares the experimentally derived emission data to a previously recorded reference spectrum for that fluorophore. In a situation involving samples with multiple overlapping spectra, linear unmixing allows the resolution of fluorophores with closely related emissions, the accurate distinction between GFP and FITC or GFP and YFP being the most often cited example of this feature.

There are other commercial spectral imaging units are available including Leica AOBS (www.leicamicrosystems.com) and Olympus FV1000 (www.olympus.com). The main differences between these three commercial systems are FV1000 based on Grating/slit/PMT two-channel bidirectional scanning mode; Leica system based on Prism/slit/PMT; and the Zeiss system based on Grating/multi-anode PMT.

Here, as an example we provided the data acquired using the Zeiss multiphoton Meta system to measure FRET signals resulting from protein–protein interactions involving C/EBPα. The GHFT1-5 mouse cells that expressed either the CFP- or YFP-C/EBPα fusion protein were used to collect the reference spectra. These reference spectra were used for the linear unmixing of spectra from cells expressing both proteins. Images were then acquired of cells expressing both the CFP- and YFP-C/EBPα bound as dimers to DNA elements in regions of heterochromatin that form clearly defined focal bodies in the nuclei of the mouse cells used here (Fig. 6). Images collected (ex 820 nm; em 545 nm) from cells expressing both the CFP- and YFP-C/EBPα lambda (λ) stacks were spectrally unmixed to reveal the donor bleed-through into the FRET channel (Fig. 6a), allowing the FRET signal to be corrected for the bleed-through signal (Fig. 6b). The emission spectrum for the signal in the FRET channel (Fig. 6c) was determined (b-FRET in panel d) and was then reacquired after selective photobleaching of YFP (a-FRET) using 514 nm, showing the unquenched donor signal. These results demonstrate the power of spectral FRET imaging using the Meta system to detect protein–protein interactions in a single living cell.

Fluorescence Lifetime Imaging Microscopy

Each of the fluorescence microscopy techniques described above uses intensity measurements to reveal fluorophore concentration and distribution in the cell. Recent advances in camera sensitivities and resolutions have improved the

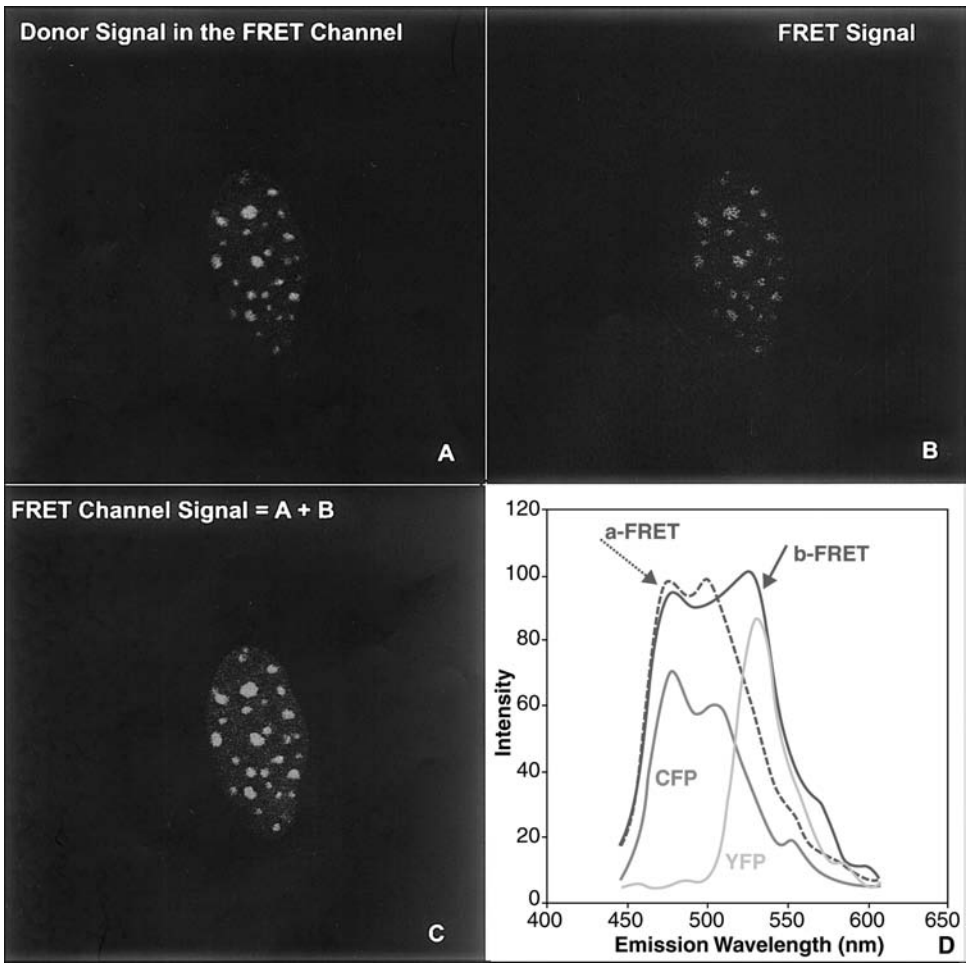


Figure 6. Spectral FRET imaging microscopy. Reference spectra (CFP, YFP) were established using cells expressing either CFP- or YFP-C/EBPα; alone, and the emission spectra are shown in panel D. Images were then collected from cells expression both the CFP- and YFP-C/EBPα, and the signals were spectrally unmixed to reveal the donor bleed-through (a) into the FRET channel (c), the corrected FRET signal is shown in b. The emission spectrum for the signal in the FRET channel (b-FRET) is shown in panel *D*, and was reacquired after selective photobleaching of YFP (a-FRET). CFPex 820 nm; YFPex 920 nm. Zeiss510 META system was used for the data acquisition.

capability of these techniques to detect dynamic cellular events (3). Unfortunately, even with the improvements in technology, these fluorescence microscopic techniques do not have high speed time resolution to fully characterize the organization and dynamics of complex cellular structures. In contrast, the time-resolved fluorescence (lifetime) microscopic technique allows the measurement of dynamic events at very high temporal resolution (nanoseconds). Fluorescence lifetime imaging microscopy (FLIM) merges the information of the spatial distribution of the probe with probe lifetime information to enhance the reliability of the concentration measurements. This technique monitors the localized changes in probe fluorescence lifetime (14,24,25,29,43,64–67) and provides an enormous advantage for imaging dynamic events within the living cells.

The fluorescence lifetime (τ) is defined as the average time that a molecule remains in an excited state prior to returning to the ground state. In practice, the fluorescence lifetime is defined as the time in which the fluorescence intensity decays to $1/e$ of the initial intensity (I_0) immediately following excitation (i.e., 37% of I_0). If a laser pulse excites a large number of similar molecules with a similar local environment and as long as no interaction with another protein or cell organelles occurs, the lifetime is the "natural fluorescence lifetime", τ_0. If energy is transferred, however, the actual fluorescence lifetime, τ, is less than the natural lifetime, τ_0, because an additional path for deexcitation is present (28).

Conventional fluorescence microscopy provides images that reveal primarily the distribution and amount of stain in the cell based on measurements of intensity. In contrast, the time-resolved fluorescence microscopic (or FLIM) technique allows the measurement of dynamic events at very high temporal resolution and can monitor interactions between cellular components with very high spatial resolution, as well. A fluorophore in a microscopic sample may exist, for example, in two environmentally distinct regions and have a similar fluorescence intensity distribution in both regions, but different fluorescence lifetimes. Measurements of fluorescence intensity alone would not reveal any difference between two or more regions, but imaging of the fluorescence lifetime would reveal such regional differences (52).

Instrumental methods for measuring fluorescence lifetimes are divided into two major categories: frequency-domain (29,65) and time-domain (52). With the time-domain method (or pulse method), the specimen is excited with a short pulse and the emitted fluorescence is integrated in two or more time windows (24). The relative intensity captured in the time windows is used to calculate the decay characteristics. The determination of prompt fluorescence with lifetime in the range of 0.1–100 ns requires elaborate fast excitation pulses and fast-gated

detection circuits. As an alternative to the time-domain method, the frequency-domain method uses a homodyne detection scheme and requires a modulated light source and a modulated detector. The excitation light is modulated in a sinusoidal fashion. The fluorescence intensity shows a delay or phase shift with respect to the excitation and a smaller modulation depth (29).

The FLIM system can be coupled to any wide-field microscope (24,29,65) (www.tautec.com; www.lambert-instruments.nl). The lifetime method can also be applied to a laser-scanning confocal microscope (www.coord.nl; www.picoquant.com) and multiphoton microscopy (26) (www.becker-hickl.com). The FLIM techniques measure environmental changes within the living cells and can be used in multilabeling experiments. An important advantage of FLIM measurements is that they are independent of change in probe concentration, excitation intensity, and other factors that limit intensity based steady-state measurements. Additionally, FLIM enables the discrimination of fluorescence coming from different dyes, including autofluorescent materials that exhibit similar absorption and emission properties but show a difference in fluorescence lifetime. The FLIM system is not only used for protein–protein interactions, but also for various biological applications from single cell to single molecule as well as deep tissue cellular imaging (3,26,66,68,69).

The data provided here were collected using the two-photon FLIM system to demonstrate the feasibility of implementing the lifetime imaging technique for drug uptake in live cells. The intensity and the lifetime image are shown in Fig. 7 of a prostate cancer (PC-3) cell after adding the drug (1.0-mM concentration) for ~ 2 min. The data clearly demonstrate that there is a considerable difference in lifetime distribution in the cytoplasmic area versus the nucleus, thus allowing the quantitation of the dynamic process of drug molecule uptake in different cellular organelles. The lifetime distribution in the cytoplasm (2 ns) and nucleus (4 ns) clearly reflects differences in molecule uptake between them and both are considerably reduced compared to the natural lifetime (20 ns) of the drug molecule. The FLIM system would reduce background interference and thus enhance measurement precision to yield more accurate understanding of drug molecule associations involved in living cells. These technologies will significantly improve and expand existing capabilities for understanding the drug molecules interactions and for characterizing their binding properties as an ensemble and at the single molecule level.

Fluorescence Correlation Spectroscopy (FCS)

Fluorescence correlation spectroscopy (FCS) is a technique in which spontaneous fluorescence intensity fluctuations are measured in a microscopic detection volume of $\sim 10^{-15}$ L as defined by a tightly focused laser beam. This spectroscopy is a special case of fluctuation correlation techniques where the laser induced fluorescence from a very small probe volume is autocorrelated in time. Fluorescence intensity fluctuations measured by FCS represent changes in either the number or the fluorescence quantum yield of molecules resident in the detection volume (27). Small,

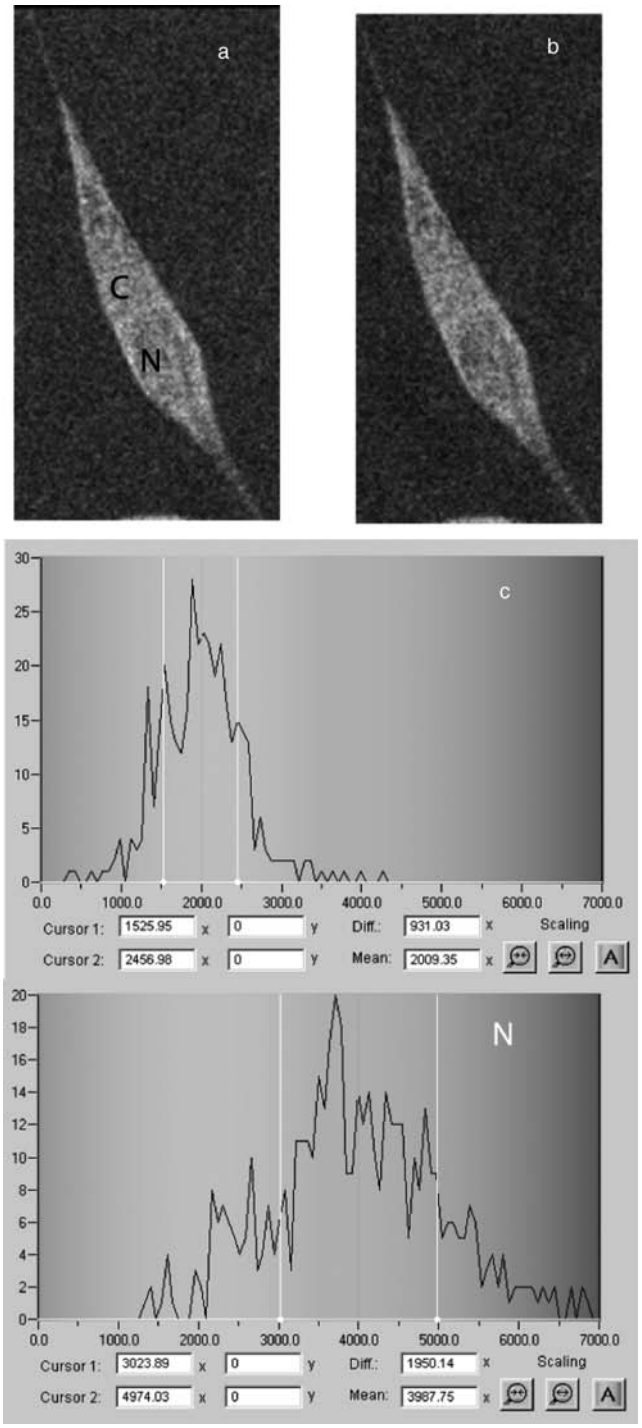


Figure 7. The FLIM microscopy-drug moleculeYK-140 uptake in a single living cell. Multiphoton excitation time-resolved intensity (a) and lifetime (b) images of PC-3 cell after adding the YK-II-140 drug (1.0-mM concentration). There is a clear difference in distribution of lifetime in the nucleus ('N') (mean $\tau_N = 3.967$ ns) versus cytoplasm ('C') (mean $\tau_C = 2.009$ ns). Moreover, considerable amount of quenching of the drug molecule in the cellular environment demonstrates that the drug molecules were interacting with various cell organelles. Consequently, the lifetime was considerably reduced from its natural lifetime 20 ns. Ex–770 nm; Em–528/30 nm. Becker and Hickl board was used in the Biorad Radiance system to acquire the data.

rapidly diffusing molecules produce rapidly fluctuating intensity patterns, whereas larger molecules produce more sustained bursts of fluorescence. If no further effects on fluorescence characteristics are present, fluctuations in the emission light simply arise from occupation number changes in the illuminated region by random particle motion. Excellent article on FCS basics was written by Petra Schwille can be seen in the URL http://www user.gwdg.de/~pschwil/BTOL_FCS.pdf.

Image Correlation Spectroscopy (ICS) was developed as the imaging analog of FCS for measuring protein aggregation in biological membranes. The ICS method entails collecting fluorescence intensity fluctuations as a function of position by using a laser scanning microscope imaging system and analyzing the imaged intensity fluctuations by spatial autocorrelation analysis (28,70). The amplitude of the normalized spatial autocorrelation function is directly related to the absolute concentration of fluorophore in the focal volume and the state of aggregation of the fluorescent entities. Extension of ICS to temporal autocorrelation analysis of image time series also permits measurement of molecular transport occurring on slower time scales characteristic of macromolecules within the plasma membrane. The other related technique, Image Cross-Correlation Spectroscopy (ICCS) allows direct measurement of the interactions of two colocalized proteins labeled with fluorophores having different emission wavelengths. Both ICS and ICCS involve the use of laser scanning confocal microscopy to obtain fluorescence images of fluorescently labeled cell membranes.

RAMAN AND CARS MICROSCOPY

Confocal, multiphoton, and fluorescence lifetime imaging microscopy have become powerful techniques for revealing 3D imaging of molecular distribution and dynamics in living specimens. This followed the development of various natural and artificial fluorophores. For chemical species or cellular components that cannot be fluorescently labeled, Raman microscopy, which measures vibrational properties and does not require molecules to have a fluorescent label, can be used to identify specific signatures of cellular or chemical components (71,72). Raman spectroscopy is an extremely powerful tool for characterizing the physical and chemical properties of the biological molecules. Raman spectroscopy is based upon the Raman effect, which may be described as the scattering of light from a molecule with a shift in wavelength from that of the usually monochromatic excitation wavelength from ultraviolet to infrared light (21). The Raman shifts are thus measures of the amounts of energy involved in the transition between initial and final states of the scattering molecule. Resonance Raman can provide more specific molecular information by working on resonance with particular electronic transitions in the protein (73). Resonant Raman spectroscopy of neutrophilic and eosinophilic granulocytes provided very clear fingerprints of the presence of oxidizing enzymes that these cells require for their functionality (74). With the use of advanced detector technology, single-cell vibrational Raman spectroscopy proved to be sufficiently sensitive to show the typical spectra of the cell nucleus and cytoplasm in human white blood cells (75). The low scattering cross-section of naturally occurring compounds, such as DNA, RNA, and proteins can be overcome by high peak powers in the laser beams used to generate the Raman signal (76,77). Using the principle of Raman microscopy for cellular imaging, several systems have already been realized: Resonant Raman spectroscopy (75), surface-enhanced Raman spectroscopy (SERS) (78), coherent anti-Stokes Raman spectroscopy (CARS) (71,72), and Fourier transform infrared absorption (FTIR) (79).

CARS microscopy relies on the Raman Effect (80). In the spontaneous Raman process, molecules scatter photons, modifying the photon energy with energy quanta that corresponds to the molecule's vibrational modes. Vibrational contrast in CARS microscopy is inherent to the cellular species, thus requiring no endogenous or exogenous fluorophores that may also be prone to photobleaching. For CARS, two optical beams of frequencies ω_p (1064 nm) and ω_s (tunable 770–900 nm) interact in the sample to generate an anti-Stokes optical output at $\omega_{as} = 2\omega_p - \omega_s$ in the phase matched or a specific direction and is resonantly enhanced if $\omega_p - \omega_s$ coincides with the frequency of a Raman active molecular vibration across the entire focal volume. This nonlinear process uses pulsed laser sources. The signal intensity has quadratic and linear dependence on pump and stokes powers, respectively. As a result, it generates signal only within the focus, where the laser intensity is the highest, enabling 3D resolution. The molecular vibrational information obtained by CARS provides a detailed fingerprint of different bonds, functional groups, and conformations of molecules, biopolymers and even microorganisms (71,72). For example, the Raman shift at 2845 cm^{-1} was used to collect the lipid signal (bright red dots, as shown by arrow in Fig. 8a). When the frequency

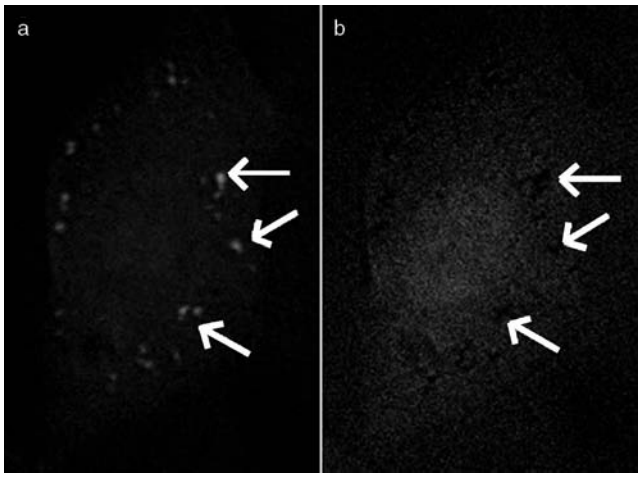


Figure 8. Demonstration of CARS and non-CARS image. (a) CARS image of lipids in 3T3 fibroblast cells excited at the vibrational frequency of 2845 cm^{-1}. (b) The frequency was tuned away from the lipids vibrational modes, 2947 cm^{-1}. No lipid is observed as pointed by the arrows. This CARS image was collected at Prof. Sunney Xie's laboratory (Harvard University) and the CARS microscopy is based on Olympus Fluoview single beam scanning system using synchronizely pumped High Q laser system.

was tuned to 2947 cm^{-1}, the lipid signal disappeared as shown by arrow in Fig. 8b. As described in the Research Activity section above we propose to study the Raman C–H stretching modes and C–C stretching modes in lipid-phase transitions for which we need to tune to different vibrational frequency to calibrate the system. If a particular molecule vibrational frequency is not known, it can be determined by conventional Raman spectrometry. Therefore, vibrational spectroscopy has found wide application in structural characterization of biological materials and in probing interaction dynamics.

CONCLUSION

Multifaceted microscopy technology moved to the center stage in cellular imaging. There is no question that the described microscopy approaches in this paper will continue to increase in all directions, driven by advances in technological development and the growing number of cell biologist researchers who will routinely use this technology for cellular imaging. Even though some of the microscopy techniques are somewhat more complex, they provide an unprecedented level of information about the micromolecular interactions in cells under physiological conditions at a very high temporal and spatial resolution. New fluorophores such as green fluorescent proteins (GFPs) and in particular Quantum Dots will expand the usefulness of cellular imaging qualitatively and quantitatively and that will lead to more detailed insights in studying the cellular dynamics. Raman and CARS microscopy techniques would allow characterizing the physical and chemical properties of the biological without the fluorophore labeling.

ACKNOWLEDGMENTS

The author would like to thank Ms. Ye Chen, Jalan Washington, and Erica Caruso for their help provided in preparation of the manuscript. The author also would like to thank Dr. Milton Brown for providing the drug compounds and Ms. Elise Shumsky, Carl Zeiss for her help in spectral FRET imaging. This work is supported by funds from National Center for Research Resources (NCRR-NIH) and Funds for Excellence in Science and Technology (FEST) at the University of Virginia.

BIBLIOGRAPHY

1. Prescott D. Methods in Cell Physiology. New York: Academic Press; 1968.
2. Pawley JB, editor. Handbook of Biological Confocal Microscopy. 2nd ed. New York: Plenum Press; 1995.
3. Periasamy A, editor. Methods in Cellular Imaging. New York: Oxford University Press; 2001.
4. Periasamy A, Day RN, editors. Molecular Imaging: FRET Microscopy and Spectroscopy. New York: Oxford University Press; 2005.
5. Grimstone A. The Electron Microscope in Biology. New York: St. Martins; 1968.
6. Frank J. Three-dimensional imaging techniques in electron microscopy. Biotechniques 1989;7(2):164–173.
7. Hogg J. The Microscope: History, Construction, and Application. London: George Routledge and Sons; 1871.
8. Jones T. 1997. History of the Light Microscope. Available at http://www.utmem.edu/personal/thjones/hist/c1.htm.
9. Periasamy A, Herman B. Computerized fluorescence microscopic vision in the biomedical sciences. J Computer-Assisted Microsc 1994;6:1–26.
10. Inoue S, Spring KR. Video Microscopy: The Fundamentals. New York: Plenum Press; 1986.
11. Wang XF, Herman B, editors. Fluorescence Imaging Spectroscopy and Microscopy. New York: John Wiley & Sons, Inc.; 1996.
12. Lippincott-Schwartz J, Snapp E, Kenworthy A. Studying protein dynamics in living cells. Nat Rev Mol Cell Biol 2001;2(6): 444–456.
13. Diaspro A, editor. Confocal and Two-photon Microscopy: Foundations, Applications, and Advances. New York: John Wiley & Sons, Inc.; 2002.
14. Marriott G, Parker I. Biophotonics, Part A and B. Methods in Enzymology. San Diego: Academic Press; 2003.
15. Prasad PN. Introduction to Biophotonics. New York: Wiley-Interscience; 2003.
16. Betzig E, Trautman JK. Near-field optics: microscopy, spectroscopy, and surface modification beyond the diffraction limit. Science 1992;257:189–195.
17. Lewis A, Radko A, Ben Ami N, Palanker D, Lieberman K. Near-field scanning optical microscopy in cell biology. Trends Cell Biol 1999;9(2):70–73.
18. Lal R, John SA. Biological applications of atomic force microscopy. Am J Physiol 1994;266(1 Pt. 1):C1–C21.
19. Driscoll RJ, Youngquist MG, Baldeschwieler JD. Atomic-scale imaging of DNA using scanning tunnelling microscopy. Nature (London) 1990;346(6281):294–296.
20. Jacobsen C, Lindaas S, Williams S, Zhang X. Scanning luminescence X-ray microscopy:imaging fluorescence dyes at suboptical resolution. J Microsc 1993;172(2):121–129.
21. Hanlon EB, Manoharan R, Koo TW, Shafer KE, Motz JT, Fitzmaurice M, Kramer JR, Itzkan I, Dasari RR, Feld MS. Prospects for *In vivo* Raman spectroscopy. Phys Med Biol 2000;45(2):R1–59.
22. Periasamy A, Skoglund P, Noakes C, Keller R. An evaluation of two-photon excitation versus confocal and digital deconvolution fluorescence microscopy imaging in Xenopus morphogenesis. Microsc Res Tech 1999;47(3):172–181.
23. Periasamy A, Day RN. Visualizing protein interactions in living cells using digitized GFP imaging and FRET microscopy. Methods Cell Biol 1999;58:293–314.
24. Elangovan M, Day RN, Periasamy A. Nanosecond fluorescence resonance energy transfer-fluorescence lifetime imaging microscopy to localize the protein interactions in a single living cell. J Microsc 2002;205(Pt. 1):3–14.
25. Gerritsen HC, Asselbergs MA, Agronskaia AV, Van Sark WG. Fluorescence lifetime imaging in scanning microscopes: acquisition speed, photon economy and lifetime resolution. J Microsc 2002;206(Pt. 3):218–224.
26. Chen Y, Periasamy A. Characterization of two-photon excitation fluorescence lifetime imaging microscopy for protein localization. Microsc Res Tech 2004;63(1):72–80.
27. Berland KM, So PT, Gratton E. Two-photon fluorescence correlation spectroscopy: method and application to the intracellular environment. Biophys J 1995;68(2):694–701.
28. Petersen N. FCS and spatial correlations on biological surfaces. In: Rigler R, Elson EL, editors. Fluorescence Correlation Spectroscopy. Berlin: Springer Verlag; 2001. p 2–35.
29. Lakowicz JR. Principles of Fluorescence Spectroscopy. 2nd ed. New York: Plenum Press; 1999.
30. Göppert-Mayer M. Ueber Elementarakte mit Quantenspreungen. Ann Phys 1931;9:273–295.

31. Haugland RP. Molecular Probes: Handbook of Fluorescent Probes and Research. Eugene, Oregon: Molecular Probes Inc.; 1989.
32. Tsien RY, Waggoner A. Fluorophores for confocal microscopy: photphysics and photochemistry. In: Pawley JB, editor. Handbook of Biological Confocal Microscopy. New York: Plenum Press; 1995. p 267–279.
33. Tsien RY. Imagining imaging's future. Nat Rev Mol Cell Biol 2003;(Suppl):SS16–SS21.
34. Wallrabe H, Periasamy A. Imaging protein molecules using FRET and FLIM microscopy. Curr Opin Biotechnol 2005;16: 19–27.
35. Lemasters JJ, Qian T, Trollinger DR, Muller-Borer BJ, Elmore SP, Cascio WE. Laser scanning confocal microscopy applied to living cells and tissues. In: Periasamy A, editor. Methods in Cellular Imaging. New York: Oxford University Press; 2001. p 66–87.
36. Corrie JET, Katayama Y, Reid GP, Anson M. The development and application of photosensitive caged compounds to aid time-resolved structure determination of macromolecules. Philos Trans R Soc London A Ser 1992;340:233–236.
37. Harper IS. Flurophores and their Labeling Procedures for Monitoring Various Biological Signals. In: Periasamy A, editor. Methods in Cellular Imaging. New York: Oxford University Press; 2001. p 20–39.
38. Miyawaki A, Llopis J, Heim R, McCaffery JM, Adams JA, Ikura M, Tsien RY. Fluorescent indicators for Ca^{2+} based on green fluorescent proteins and calmodulin. Nature (London) 1997;388(6645):882–887.
39. Miyawaki A, Griesbeck O, Heim R, Tsien RY. Dynamic and quantitative Ca^{2+} measurements using improved cameleons. Proc Natl Acad Sci USA 1999;96(5):2135–2140.
40. Ting AY, Kain KH, Klemke RL, Tsien RY. Genetically encoded fluorescent reporters of protein tyrosine kinase activities in living cells. Proc Natl Acad Sci USA 2001;98 (26):15003–15008.
41. Day RN. Visualization of Pit-1 transcription factor interactions in the living cell nucleus by fluorescence resonance energy transfer microscopy. Mol Endocrinol 1998;12(9): 1410–1419.
42. Chen Y, Periasamy A. Time-correlated single photon counting (TCSPC) FLIM-FRET microscopy for protein localization. In: Periasamy A, Day RN, editors. Molecular Imaging: FRET Microscopy and Spectroscopy. New York: Oxford University Press; 2005. Chapt.13.
43. Chen Y, Elangovan M, Periasamy A. FRET data analysis: The algorithm. In: Periasamy A, Day RN, editors. Molecular Imaging: FRET Microscopy and Spectroscopy. New York: Oxford University Press; 2005. Chap. 7.
44. Maddox P, Desai A, Salmon ED, Mitchison TJ, Oogema K, Kapoor T, Matsumoto B, Inoue S. Dynamic confocal imaging of mitochondria in swimming Tetrahymena and of microtubule poleward flux in Xenopus extract spindles. Biol Bull 1999;197(2):263–265.
45. Elangovan M, Wallrabe H, Chen Y, Day RN, Barroso M, Periasamy A. Characterization of one- and two-photon excitation fluorescence resonance energy transfer microscopy. Methods 2003;29(1):58–73.
46. Mills JD, Stone JR, Rubin DG, Melon DE, Okonkwo DO, Periasamy A, Helm GA. Illuminating protein interactions in tissue using confocal and two-photon excitation fluorescent resonance energy transfer microscopy. J Biomed Opt 2003;8(3):347–356.
47. Sekar RB, Periasamy A. Fluorescence resonance energy transfer (FRET) microscopy imaging of live cell protein localizations. J Cell Biol 2003;160(5):629–633.
48. Wallrabe H, Elangovan M, Burchard A, Periasamy A, Barroso M. Confocal FRET microscopy to measure clustering of ligand-receptor complexes in endocytic membranes. Biophys J 2003; 85(1):559–571.
49. Forster T. Delocalized excitation and excitation transfer. In: Sinanoglu O, editor. Modern Quantum Chemistry Part III: Action of Light and Organic Crystals. New York: Academic Press; p 93–137.
50. Clegg RM. Fluorescence resonance energy transfer. In: Wang XF, Herman B, editors. Fluorescence Imaging Spectroscopy and Microscopy. Volume 137, New York: John Wiley & Sons, Inc.; 1996. p 179–251.
51. dos Remedios CG, Miki M, Barden JA. Fluorescence resonance energy transfer measurements of distances in actin and myosin: A critical evaluation. J Muscle Res Cell Motil 1987;8: 97–117.
52. Periasamy A, Wodnicki P, Wang XF, Kwon S, Gordon GW, Herman B. Time-resolved fluorescence lifetime imaging microscopy using a picosecond pulsed tunable dye laser system. Rev Sci Instrum 1996;67(10):3722–3731.
53. Kaiser W, Garrett CGB. Two-photon excitation in CaF2: Eu^{2+}. Phys Rev Lett 1961;7:229–231.
54. Denk W, Strickler JH, Webb WW. Two-photon laser scanning fluorescence microscopy. Science 1990;248(4951):73–76.
55. Denk W, Piston DW, Webb WW. Two-photon molecular excitation in laser-scanning microscopy. In: Pawley JB, editor. Handbook of Biological Confocal Microscopy. New York: Plenum Press; 1995. p 445–458.
56. Szmacinski H, Gryczynski I, Lakowicz JR. Three-photon induced fluorescence of the calcium probe Indo-1. Biophys J 1996;70(1):547–555.
57. Svoboda K, Helmchen F, Denk W, Tank DW. Spread of dendritic excitation in layer 2/3 pyramidal neurons in rat barrel cortex *In vivo*. Nat Neurosci 1999;2(1):65–73.
58. Bacskai BJ, Hickey GA, Skoch J, Kajdasz ST, Wang Y, Huang GF, Mathis CA, Klunk WE, Hyman BT. Four-dimensional multiphoton imaging of brain entry, amyloid binding, and clearance of an amyloid-beta ligand in transgenic mice. Proc Natl Acad Sci USA 2003;100(21):12462–12467.
59. Konig K, Riemann I. High-resolution multiphoton tomography of human skin with subcellular spatial resolution and picosecond time resolution. J Biomed Opt 2003;8(3): 432–439.
60. Soeller C, Jacobs MD, Jones KT, Ellis-Davies GC, Donaldson PJ, Cannell MB. Application of two-photon flash photolysis to reveal intercellular communication and intracellular Ca^{2+} movements. J Biomed Opt 2003;8(3):418–427.
61. Farkas D. Spectral microscopy for quantitative cell and tissue imaging. In: Periasamy A, editor. Methods in Cellular Imaging. New York: Oxford University Press; 2001. Chap. 20.
62. Dickinson ME, Simbuerger E, Zimmermann B, Waters CW, Fraser SE. Multiphoton excitation spectra in biological samples. J Biomed Opt 2003;8(3):329–338.
63. Nashmi R, Dickinson ME, McKinney S, Jareb M, Labarca C, Fraser SE, Lester HA. Assembly of alpha4beta2 nicotinic acetylcholine receptors assessed with functional fluorescently labeled subunits: effects of localization, trafficking, and nicotine-induced upregulation in clonal mammalian cells and in cultured midbrain neurons. J Neurosci 2003;23(37): 11554–11567.
64. Gadella TWJ, Jovin TM, Clegg RM. Fluorescence Lifetime Imaging Microscopy (Flim) - Spatial-Resolution of Microstructures on the Nanosecond Time-Scale. Biophys Chem 1993;48(2):221–239.
65. Gratton E, Breusegem S, Sutin J, Ruan Q, Barry N. Fluorescence lifetime imaging for the two-photon microscope: time-domain and frequency-domain methods. J Biomed Opt 2003;8(3):381–390.
66. Krishnan RV, Masuda A, Centonze VE, Herman B. Quantitative imaging of protein-protein interactions by multiphoton fluorescence lifetime imaging microscopy using a streak camera. J Biomed Opt 2003;8(3):362–367.

67. Redford G, Clegg RB. Real-Time Fluorescence Lifetime Imaging and FRET using Fast Gated Image Intensifiers. In: Periasamy A, Day RN, editors. Molecular Imaging: FRET Microscopy and Spectroscopy. New York: Oxford University Press; 2005. Chapt. 11, in press.
68. Bastiaens PI, Squire A. Fluorescence lifetime imaging microscopy: spatial resolution of biochemical processes in the cell. Trends Cell Biol 1999;9:48–52.
69. Hohng S, Joo C, Ha T. Single-Molecule Three-Color FRET. Biophys J 2004;87(2):1328–1337.
70. Petersen NO, Hoddelius PL, Wiseman PW, Seger O, Magnusson KE. Quantitation of membrane receptor distributions by image correlation spectroscopy: concept and application. Biophys J 1993;65(3):1135–1146.
71. Cheng JX, Volkmer A, Book LD, Xie XS. Multiplex coherent anti-Stokes Raman scattering microspectroscopy and study of lipid vesicles. J Phys Chem B 2002a 106:8493–8498.
72. Cheng JX, Jia YK, Zheng G, Xie XS. Laser-scanning coherent anti-Stokes Raman scattering microscopy and applications to cell biology. Biophys J 2002b;83(1):502–509.
73. Carey PR. Raman spectroscopy, the sleeping giant in structural biology, awakes. J Biol Chem 1999;274(38):26625–26628.
74. Salmaso BL, Puppels GJ, Caspers PJ, Floris R, Wever R, Greve J. Resonance Raman microspectroscopic characterization of eosinophil peroxidase in human eosinophilic granulocytes. Biophys J 1994;67(1):436–446.
75. Puppels GJ, de Mul FF, Otto C, Greve J, Robert-Nicoud M, Arndt-Jovin DJ, Jovin TM. Studying single living cells and chromosomes by confocal Raman microspectroscopy. Nature (London) 1990;347(6290):301–303.
76. Volkmer A, Cheng JX, Xie XS. Vibrational imaging with a high sensitivity via epidetected coherent anti-Stokes Raman scattering microscopy. Phys Rev Lett 2001;87:023901.
77. Uzunbajakava N, Lenferink A, Kraan Y, Volokhina E, Vrensen G, Greve J, Otto C. Nonresonant confocal Raman imaging of DNA and protein distribution in apoptotic cells. Biophys J 2003;84(6):3968–3981.
78. Hawi SR, Rochanakij S, Adar F, Campbell WB, Nithipatikom K. Detection of membrane-bound enzymes in cells using immunoassay and Raman microspectroscopy. Anal Biochem 1998;259(2):212–217.
79. Diem M, Chiriboga L, Lasch P, Pacifico A. IR spectra and IR spectral maps of individual normal and cancerous cells. Biopolymers 2002;67(4–5):349–353.
80. Nan X, Yang WY, Xie XS. CARS microscopy: lights up lipids in living cells. Biophoton Int 2004;11(8):44–47.

See also CYTOLOGY, AUTOMATED; MICROSCOPY, CONFOCAL; MICROSCOPY, ELECTRON.

CEREBROSPINAL FLUID. See HYDROCEPHALUS, TOOLS FOR DIAGNOSIS AND TREATMENT OF.

CHEMICAL ANALYZERS. See ANALYTICAL METHODS, AUTOMATED.

CHEMICAL SHIFT IMAGING. See NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY.

CHROMATOGRAPHY

THAYNE L. EDWARDS
University of Washington
Seattle, Washington

INTRODUCTION

Chromatography is the process of separating a mobile phase mixture into its individual components using the relative interactions of the components with a stationary phase. Chromatography is often used as a method of purification, even when the components are closely related. When used in conjunction with a concentration or mass-based sensor, it can also be a power analytical method. This chapter explores the basic processes of the various types of chromatography and the closely related techniques of field-flow fractionation (FFF) and electrophoresis. Some of the basic theory of chromatography will be given in terms of retention mechanism and separation performance. In addition, a description of the basic types of chromatography, electrophoresis, and FFF will be given in relation to their theory and application.

Chromatography literally means "color writing" because of an observation in the early 1900s by Mikhail Tswett in separating pigments of plants into various color bands using $CaCO_3$ (1). Although Tswett is considered the father of chromatography, he was not the first to observe the chromatographic process in an experimental setting. Pliny the Elder (ca. 79 AD) recorded a crude paper chromatography experiment. Several others between this time and the realization of the usefulness of chromatography by Tswett also observed the process in the laboratory. It was not until the 1930s that it was recognized generally as an analytical technique (2). Since then, many advances, discoveries, and inventions have made chromatography an indispensable laboratory and industrial technique.

The purpose of chromatographic methods falls under either purification or analysis. Gas chromatography is almost invariably an analytical method whereas liquid chromatography is used for either. As a result of the extensive research into various methods of sample retention, even samples with relatively little difference in their physical or chemical structure can be separated. Chromatographic systems also range in size and complexity from a simple, gravity-fed packed column to a complex high pressure industrial-sized system with integrated components for sample introduction and fraction detection and collection. Often, chromatography is used in conjunction with another analytical method for determination of fraction purity and composition. For medical device instrumentation, chromatography is an invaluable technique that may find its application in a wide variety of ways such as protein purification and sample contamination detection.

Numerous journal articles and books have been written on the subject, as well as reviews (3) and a host of information published on the Internet (2,4–6). Several journals dedicated to chromatography and related fields also exist (Advances in Chromatography, Chromatographia, Biomedical Chromatography, Journal of Chromatography, Journal of Liquid Chromatography and Related Technologies, Journal of Planar Chromatography, Journal of Separation Science). This information is well accessible to anyone who has interest in pursuing the methods and techniques described in this chapter. No attempt has been made here to review all this material. For this reason, this chapter will not focus on the details and extensive literature, but on the

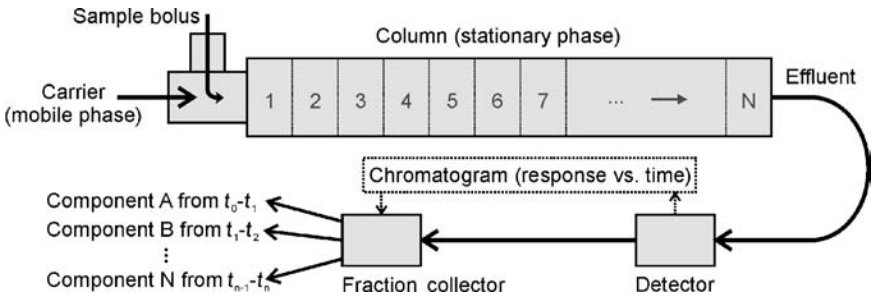


Figure 1. Depiction of a basic chromatographic system with the major components: carrier or mobile phase, sample bolus, column containing the stationary phase, concentration or mass-based detector, and the effluent. Also shown are the theoretical equilibrium plates, each width being one plate height.

basics and presenting what is generally possible with chromatography.

Chromatographic systems have a mobile phase and a stationary phase. The mobile phase is used to transport the samples through the stationary phase. The mobile phase is inert and does not interact with the sample or stationary phase. The stationary phase is unique for each separation because its purpose is to provide selective interaction with the samples. A simple chromatographic system is depicted in Fig. 1. Apart from the two fundamental required components just mentioned, a system must typically contain a sample injection port and detector in order to be useful as an analytical system. If it is to be used as a purification method, then it must also contain some method of recovering the fractions in the effluent.

The basic operating principle is that samples in the mixture, which interact to a higher degree with the stationary phase, travel slower and are retained longer in the system. This interaction, speed, and retention is relative to samples that interact to a lesser degree and so travel faster and are retained for a shorter duration. The spatial separation occurs in the mobile phase flow direction.

Chromatographic methods are grouped either into *liquid* or *gas* chromatography, based on the carrier phase and further identified by its particular method of sample retention. For example, in size-exclusion chromatography, the stationary phase is a bed of porous beads. The pores allow only samples smaller than a particular size to enter and then exit again. As a result of the pores tortuosity, the travel distance for these smaller samples is longer and thus increases the residence time relative to the larger samples. Some other methods include gas-adsorption, gas-liquid, capillary gas, liquid-adsorption, liquid-liquid, supercritical fluid, ion exchange, and affinity. In addition, the closely related fields of electrophoresis and FFF need to be mentioned as they typically complement chromatographic methods. In order to more fully understand the methods mentioned here, the basic theory will be mentioned first and then a more detailed description of the methods will follow.

GENERAL THEORY

The basic measurement in chromatography is the retention factor. This measurement is calculated from the measured retention times (t_0, t_A, t_B, ..., t_n) in the chromatogram output, as shown in Fig. 2. The void time t_0 is the elution time of an unretained sample. It is the time required to sweep one-column volume. All retention values are calculated relative to this time. The retention factor for sample A, R_A, is then

$$R_A = \frac{t_A - t_0}{t_0}$$

where t_A is the sample residence time. The ability of the separation column to distinguish between two components is quantified in the selectivity term S

$$S_{AB} = \frac{R_S}{R_A}$$

.

For mixtures, the degree of separation between two components is termed the resolution R_S. In most cases, the resolution is the most important factor because that is the purpose of performing chromatographic separations. It is calculated from the retention times t and the peak widths w for the two samples as

$$R_S = \frac{2(t_B - t_A)}{W_B + W_A}$$

The relative distance in the system at which a degree of equilibrium between the two phases occurs is referred to as a theoretical plate. Chromatographic systems usually have many theoretical plates. By dividing the length of the column L with the number of plates N gives the plate height, or height of an equivalent theoretical plate H. These numbers characterize the quality of the retention and can be determined from a chromatogram by assuming

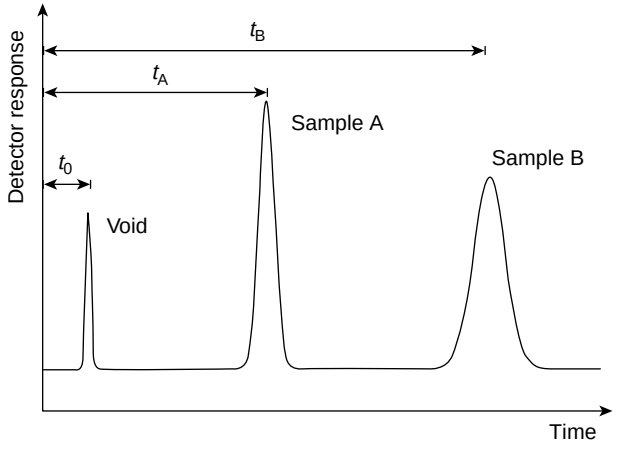


Figure 2. Chromatogram of separation of a two-component mixture. The retention times are measured from the injection time.

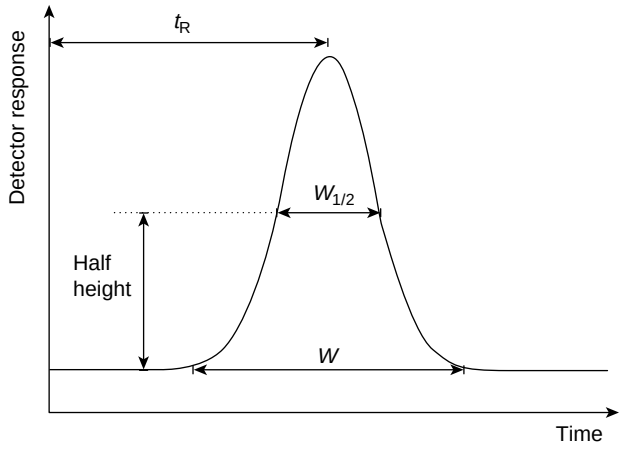


Figure 3. Chromatogram of single peak from response of mass- or concentration-based detector showing measurements for determining the number of theoretical plates.

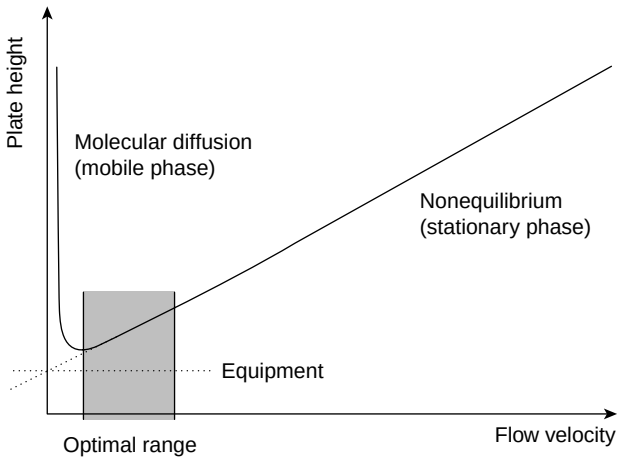


Figure 4. The van Deemter plot, based on the van Deemter equation for determining the optimal flow rate for a given chromatographic separation.

that the peaks are Gaussian shaped, as shown in Fig. 3. It is noted that this method is for an ideal situation and that other methods are available for peaks of other shapes. The sample peak width is measured either at baseline w or at half-height, $w_{1/2}$. In practice, it is more convenient and accurate to measure at half-height

$$N = 5.54\left(\frac{t_R}{W_{1/2}}\right)^2$$

and

$$H = \frac{L}{N}$$

Plate height is a summation of three effects, two of which are dependent on flow rate. This theory was proposed by van Deemter et al. (7). The first effect is due to equipment and users, such as column-packing and injection variabilities, and is not a function of the flow rate. This term can be minimized through careful design and manufacturing of the column. The use of automated sample handling also helps to reduce this effect. At low flow velocities u, molecular diffusion of the sample in the carrier dominates and peak broadening rises quickly. At higher flow velocities the sample plug broadens due to nonequilibrium effects such as eddy diffusion and multiple paths in the stationary phase. The van Deemter equation is

$$H = H_0 + H_1(u^{-1}) + H_2(u)$$

This equation is easily graphed (Fig. 4) for a visual indication of the optimal flow rate. In practice, it is best to have the flow rate slightly higher than at the optimum to avoid the region of rapid plate height increase. Each of the terms is dependent on the type of chromatography used.

Now, the resolution of the separation can be put in terms of plate height or the equivalent number of theoretical plates,

$$Rs_{AB} = \frac{\sqrt{N}}{4}\left[\frac{S_{AB} - 1}{S_{AB}}\right]\left[\frac{1 + R_B}{R_B}\right]$$

This form is useful for optimizing a separation based on the three groups in the equation. Resolution can be increased by increasing the number of theoretical plates, which can be accomplished by both increasing the column length and minimizing the plate height using the van Deemter plot. However, increasing the column length will also cause proportional band broadening, and so it is not always ideal. The second term involves adjusting the selectivity through column modifications such as changing either or both of the phases and the temperature. The temperature also plays an important role in the last term as well. By adjusting and tuning these parameters, nearly every difficult separation can be successfully resolved. Theoretically, choosing the proper stationary phase for the column and an appropriate mobile phase, any two materials can be separated. These phases are at the heart of chromatography.

TYPES OF CHROMATOGRAPHY AND RELATED TECHNOLOGIES

The mobile phase is either a gas, a liquid, or a supercritical fluid, whereas the stationary phases can be either a solid or a liquid. The types of chromatography are typically named by their phases or interaction process and are categorically divided by their stationary phase into either gas chromatography (GC) or liquid chromatography (LC). Another variation on LC is high performance liquid chromatography (HPLC), in which the carrier is driven by pressure. Some common subclasses of LC and GC are based on ion-exchange, phase change, adsorption, size exclusion, partitioning, and absorption. Specific types and hybrid systems also exist that fall under each of these categories. Each of these complements the others to build a broad spectrum of types of chromatographic separation technique available (8–11).

Gas Chromatography (GC)

Gas chromatography makes use of a pressurized gas cylinder and a carrier gas, such as helium, to carry the solute

through the column. GC can be used for both purification and analysis, when a detector is used in tandem. Common detectors used in GC are thermal conductivity and flame ionization detectors. Many more types of detectors exist, each with its advantages and disadvantages. Three types of GC that are among the more common methods are gas adsorption, gas-liquid, and capillary gas chromatography.

Gas Adsorption. Gas adsorption chromatography has a solid stationary phase packed bed. The samples selectively adsorb and desorb to the stationary phase, effectively increasing each sample's retention time based on its isotherm. Some of the more common adsorbents used are silica, zeolite, and activated alumina. This method is the primary method for separating mixtures of gases.

Gas-Liquid. Separation in gas-liquid chromatography is based on the gaseous samples partitioning with a viscous liquid stationary phase. This liquid is supported in the column by coating a solid, most commonly diatomaceous earth. The solid support to the stationary phase liquid is inert to the samples. The sample retention time is governed by the rate at which it dissolves into and vaporizes out of the liquid. Thus, the relative partitioning of each of the samples in the liquid stationary phase is the basis of the separation.

Capillary Gas. In this method, the stationary phase is a capillary coated with a liquid (wall-coated open tubular) or a solid-coated capillary onto which the liquid is adsorbed (support-coated open tubular), as has been described in the previous two methods. Liquid or gum temperature stable polymers are used as the stationary phase. Most common polymers used are poly-ethylene glycol or poly-siloxanes. Also used are molecular sieves and alumina particles. Unlike the previous methods, the stationary phase has a small volume due to the capillary geometry and is thus limited to the amount of sample that can interact. However, because of the small column geometry, the partitioning or adsorption of the sample is relatively fast. The capillary is typically glass or fused silica coated with polyimide for support. The tubing (column) can be long and also be wound into tight areas for compactness and good temperature control. It is the most common gas chromatography analytical method.

Liquid Chromatography (LC) and High Performance Liquid Chromatography (HPLC)

As the carrier phase in LC is a liquid, it is naturally more amenable for biological separations and analysis, such as the purification of proteins. However, it is also amenable to any sample dissolved in a liquid. In LC, the carrier is driven by gravity through the column. These columns, made of glass or plastic and sometimes disposable, are typically used for lab-scale preparative work. For analysis of samples, the carrier is pressurized for increased speed and sample resolution. This variation is termed high performance liquid chromatography or HPLC. These systems are much more complex and costly. The columns are made of steel to withstand high pressures and are reused a number of times. Detectors are also placed inline with these columns for analysis, although HPLC is also used in preparative work as well (3,6,11–14).

Liquid Adsorption. Liquid adsorption, also termed liquid-solid chromatography, uses a solid stationary phase made of particles such as alumina or silica. In particular, this method is used in separating isomers. The retention is based on the adsorption/desorption kinetics of each sample onto the particles. Liquid adsorption is often found in large-scale applications because the adsorbent beds are relatively inexpensive.

Liquid-Liquid. In liquid-liquid chromatography (LLC), also called partition chromatography, the stationary phase is a liquid-coated solid surface. This liquid is immiscible with the liquid solvent mobile phase. Retention is based on partitioning of the sample between the two phases. LLC can be accomplished in either *normal* phase or *reverse* phase. Normal phase has a nonpolar mobile phase and polar stationary phase. Reverse phase is the opposite of having a polar mobile phase and nonpolar stationary phase. It is used primarily in separating nonvolatile components of mixtures and is similar to a chemical extraction process.

Size Exclusion. This method was described briefly in the introduction. It is somewhat unique because the stationary phase is inert to the sample. The increased path length due to tortuous pores that exclude large samples causes an increased retention time for samples smaller than the cut-off size. It is also referred to as filtration, gel permeation, or molecular-sieve chromatography. This method is useful for protein separation and purification such as in antibody production and buffer exchange applications.

Supercritical Fluid. Unlike the other methods, supercritical fluid chromatography is characterized by its unique carrier fluid. Supercritical fluids used to carry the sample have very high viscosities and molecular diffusivities compared with liquids but with densities on the same order. One type of supercritical fluid used is a mixture of carbon dioxide and modifiers. Implementation of this technique is difficult because of the high temperature and pressures to reach the supercritical fluid state.

Ion-Exchange. Ion-exchange chromatography is commonly used in the purification of biological materials, such as amino acids and proteins, and also ions in solution. This method is capable of quantifying samples in the ppb to ppm concentration range. The stationary phase is an ion-exchange resin that is either cationic or anionic. Charged atoms or molecules in the liquid phase sample bind to the stationary phase as they are passed through the column. The sample is released by adjusting the carrier pH or ionic strength. Separation by this method is highly selective and especially useful for anions in which separations are typically slow. The resins are typically high capacity and inexpensive.

Affinity. Affinity chromatography has a stationary phase that is highly selective to one particular sample.

Unlike other chromatographic methods, the sample is highly bound to the stationary phase until the carrier solution is changed and the sample released. To accomplish this selective release, the stationary phase is engineered using an inert affinity matrix, such as agarose or cellulose derivative, and infused with ligand molecules that are design to bind only the sample of choice. Immunologic interactions of specific antibody-antigen pairs are particularly useful because of the high specificity that can be obtained and the reversibility of the binding event. The addition of a high salt concentration or low pH to the stationary phase reverses the selectivity, similar to ion-exchange chromatography, and allows the release of the sample after the other components of the mixture have been washed away. Care must be taken to ensure that impurities do not foul the matrix. Some preprocessing is typically accomplished prior to the separation to remove the potential fouling components. This method is used often with biological samples.

Electrophoresis

Electrophoresis is a separation method using the transport of electrically charged compounds in a conductive liquid environment under the influence of an electric field (15,16). Positively charged molecules migrate toward a negative electrode, and negatively charged molecules migrate toward a positive electrode. It is regularly applied in analytical chemistry to determine the constituent molecules of a compound. It is also widely used in medical diagnostics and other biological areas to determine molecules within biological samples, such as protein and DNA. From the various modes of electrophoresis, capillary electrophoresis (CE) is the most widely used separation method used in a modern analytical laboratory (17). High separation speed, excellent resolution power, and low consumption of buffer and sample are some of the advantages. Typically, samples are injected into a capillary tube with diameters ranging between 25 and 100 μm, and an electrical field is applied along the capillary tube to separate compounds based on the differences in charge to mass ratio. Negatively ionized surface silanol group of the capillary creates an electrical double layer at the solid/liquid interface to preserve electroneutrality, and this mobile layer is pulled toward the negatively charged electrode when an electric field is applied. These ion layers drag the bulk buffer solution, causing an electro-osmotic flow. Compared with HPLC, which has a parabolic flow profile due to the laminar flow inside the channel, the flow profile is flat in the electro-osmotic flow, which helps the detection peak to be very sharp, increasing its sensitivity. Laser-induced fluorescence detection is the most widely used method for detecting the separated molecules (18). Over 100 review articles exist covering capillary electrophoresis, and Beale wrote an excellent review categorizing these articles (19).

Development in microfabrication technologies and the lab-on-a-chip concept in the early 1990s further expanded the role of this powerful analysis technique (20–22). Smaller sample injection into a microchannel and higher electric field result in short analysis time with excellent resolution. Applying higher electric field is possible due to the high surface-to-volume ratio of a microchannel that can dissipate the heat produced during electrophoresis faster. Automated sample injection and capability to perform the separation on arrays of microchannel in conjunction with the short analysis time enables high throughput analyses. Low manufacturing cost due to the batch fabrication capability of the microchips is another advantage over conventional separation technologies. Fig. 5 shows a typical channel configuration for a capillary electrophoresis microchip. High voltages are applied between reservoir 1 and 2 so that the sample in reservoir 1 fills the injection channel. Once the injection channel is filled, the high voltages are switched to reservoir 3 and 4, and the sample plug in the cross section gets injected into the separation channel. As the sample plug moves down the separation channel, separation occurs depending on the charge to mass ratio of the compounds being analyzed. In fluorescence detection, detection typically occurs at the end of the separation channel by illuminating the fluorescence-tagged sample with laser or UV light, followed by light detection using a photomultiplier tube (PMT).

Typical channels are several tens to hundreds of microns wide, a couple tens of microns deep, and the separation channel lengths are in the centimeter scale. Electric fields applied to these channels range in kV/cm scale. This high electric field and small sample size enables separation within several seconds compared with tens of minutes required in conventional chromatography, which can be of great benefit when monitoring time-dependent reactions, such as conducting an enzyme kinetic study (23). This system also enables automatic sample injection because voltages can be simply switched between reservoirs to inject samples into the separation channels without the need for manual sample loading, which further reduces the time associated with sample analyses. Several other sample injection schemes have been also studied. Instead of using a simple cross-channel injector, twin-T injectors can be used to further control the sample plug size (24). Electrical biasing of the different reservoirs such as

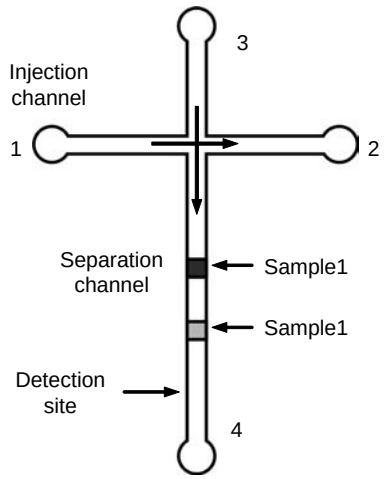


Figure 5. Typical channel configuration for a capillary electrophoresis microchip including loading or injection channel and separation channel.

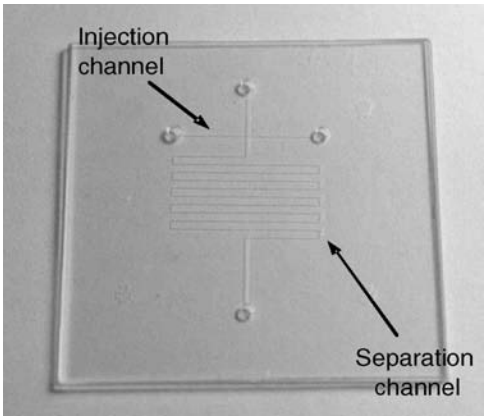


Figure 6. CE microchip fabricated in borosilicate glass. Serpentine separation channel can be observed.

pinch injection can reduce band broadening of the sample plugs caused by leakage at the intersection of the injection channel (25). Fig. 6 shows a CE microchip made in borosilicate glass. Serpentine separation channel was created to have a longer separation channel within a compact geometry.

The most common material used for the microchip is glass. Some of the earliest capillary electrophoresis microchips were fabricated in glass due to the good optical properties, a surface that enables electro-osmotic flow, and well-developed microfabrication techniques. To further reduce the microchip fabrication cost, polymer materials have been used due to the low material cost and easy mass-fabrication processes. Injection molding or hot embossing of plastics and polymer casting of polydimethylsiloxane (PDMS) are some of the methods used (26–28).

The application of this technology has been expanded from simple chemical analysis to biological applications, such as DNA sequencing, immunoassay, and biological particle separation (viruses, bacteria, and eukaryotic cells) (22,29–31). DNA sequencing on microchips was first reported by Mathies in 1995 (32). Single base resolution reached 150–200 bases in 10–15 min. Some of the advantages on top of the excellent resolution and short analysis time are the capability for high throughput. The compact size of the separation channels enable a large number of channels to be placed close together, which also facilitates fast detection using optical imaging or scanning lenses. To pack even more channels into a small area, a 6 inch circular glass plate carrying 96 radical channels conversing at the center of the chip was also developed (33). Detection occurred using a spinning confocal system. Separation of antigen and antibody from the corresponding antibody-antigen complex can be separated using microchip CE for immunoassays (34). Automatic sample injection capability can possibly eliminate the need of conventional robotic sample injection, which is commonly used in life science laboratories. More complex operations, by using electro-osmotic flow in conjunction with other operations such as lysing and concentration, have been demonstrated to show transport and analysis of biological particles (35–37). Dual injection has been also used where sample and reagents can be mixed on-column and analyzed to provide information about reaction kinetics, to perform on-column derivatization for improved separation and termination, and to develop methods for simultaneous analysis of anionic and cationic compounds (38). Electro-osmotic flow in microchannels has been used in numerous applications to transport and mix fluid and particles but is beyond the subject of this chapter (39,40).

Electrophoresis with Other Separation/Detection Methods

One of the advantages of capillary electrophoresis is that this technique is easy to combine with different separation and detection methods to provide even more versatile, powerful, and efficient analysis tools. Isoelectric focusing (IEF) is a separation technique to resolve amphoteric molecules based on their isoelectric points (pI) (41). Isoelectric point is the pH at which a molecule carries no net electric charge. In capillary isoelectric focusing (CIEF), the capillary is first filled with a mixture of ampholytes and samples (42). When an electric field is applied to the capillary, a pH gradient is formed inside the capillary and the sample molecules migrate and stop at a position where the pH equals the pI of the sample molecules due to the loss of their net charges. In a one-step process, the entire capillary is illuminated to obtain images of the separation. In a two-step process, the separated samples are mobilized to the detection point using chemical, hydrodynamic, or electro-osmotic flow mobilization to simplify the detection equipments. When analyzing mixture of peptides in a microchip-based CIEF, focusing time of less than 30 seconds and total analysis time as short as 5 minutes is possible. As a result of the high resolving power, this method is most commonly used for studying peptides, proteins, recombinant products, cell lysates, and other complex mixtures (43,44).

Although capillary electrophoresis can provide excellent resolution, it is challenging to identify unknown substances. Mass spectrometry is a technique used for separating ions by their mass to charge ratios that enables identification of compounds by the mass of one or more elements in the compounds and enables determination of isotopic composition of one or more elements in the compound. By coupling capillary electrophoresis with mass spectrometry, direct identification of analytes by molecular mass, selectivity enhancement, and insight into the molecular structures are possible (43,44). The most prominent application of this combination is in proteomics (45). Interfacing these two techniques is of great importance because mass spectrometry requires ionized gas as samples whereas the output from a CE system is fluid (46). Electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) are some of the most widely used ionization methods. ESI, first developed in the 1980s, is the softest ionization technique currently available. It transforms ions in solution into ions in gas phase based on the electrostatic effects in solutions. An electric potential applied to an electrospray tip breaks the solution containing mixture of samples and solvents into small charged droplets. The shrinkage of the charged droplets by solvent evaporation further disintegrates the drops and forms gas-phase ions. MALDI uses laser beams to ionize samples located inside a crystallized bimolecular matrix

that is used to protect the sample from being destroyed by direct laser beam. For microchip-based capillary electrophoresis systems, efforts have been focused on developing on-chip electrospray ionization techniques so that coupling to MS systems are efficient (47). The ultimate goal of such a coupled system is to use the microchip for fast and convenient sample preparation followed by online sample introduction for MS analysis. Beyond proteomics, the CE-ESI-MS combination has been widely used for drug analysis, food analysis, achiral and chiral solutes analysis, glycoscreening, and metabolic disorder screening (48–52).

Other detection methods used include electrochemical detection (53), electrochemiluminescence, nuclear magnetic resonance (NMR), ultraviolet resonance Raman spectroscopy (54,55).

Field-Flow Fractionation

Another technique similar in many ways and complementary to chromatography is field-flow fractionation (FFF). It is relatively young compared with chromatography, proposed by Giddings in 1968. This technique is always performed in an open channel (no packing or coatings) that is usually, but not restricted to, a wide, flat geometry with the breadth to height ratio being greater than 100 (Fig. 7). The purpose of this geometry is to take advantage of the laminar velocity parabolic flow profile of the carrier while minimizing secondary dispersion effects from the sidewalls. Circular channels can also be used in this way but are difficult to implement a uniform field in. Just as in chromatography, the samples spatially separated along the length of the channel are eluted discretely at the outlet, if enough column length and separating power are provided. For this reason, chromatographic principles and basic theory also apply to FFF. For example, FFF system retentions are often characterized by the number of theoretical plates and the van Deemter theory and the separations are characterized by resolution. However, some distinct differences exist that demonstrate the unique and complementary characteristics of FFF.

The first difference between FFF and chromatography is that a force field is applied to affect the samples instead of a stationary phase. The second difference is that the field is applied normal to the flow and separation direction. In chromatography, this field always acts opposite to the direction of the separation. The third difference is that, in most simple FFF systems, a direct mathematical relationship exists between the field and sample elution time.

The mechanism of retention and separation in FFF systems is based on compartmentalizing the various samples in the mixture to velocity zones in the parabolic flow profile of the carrier. The samples are selectively perturbed using a field applied normal to the carrier flow that are then concentrated at the accumulation wall. Normal diffusion opposes this movement until an equilibrium condition is established. Each sample will form a layer thickness based on its degree of perturbation. The sample specific velocity is then obtained from the first moment of the concentration and velocity profiles. The exact mathematical relationship among the variables in each of these steps allows for the determination of specific sample properties based solely on the elution time. A simple concentration or mass-based detector is sufficient without the need for calibration as in chromatography. In practice, this type of analysis is complicated by other factors leading to the need for calibration or more complicated detectors.

Any type of field or combination of fields can, in theory, be used to drive the separation. Some of the common fields used are sedimentary, flow, thermal, and electric. Some less common fields include magnetic, acoustic, dielectric, and others.

Recent advancements in FFF have included miniaturization of FFF channels. Miniaturization not only reduces sample and carrier volume requirements but also enhances

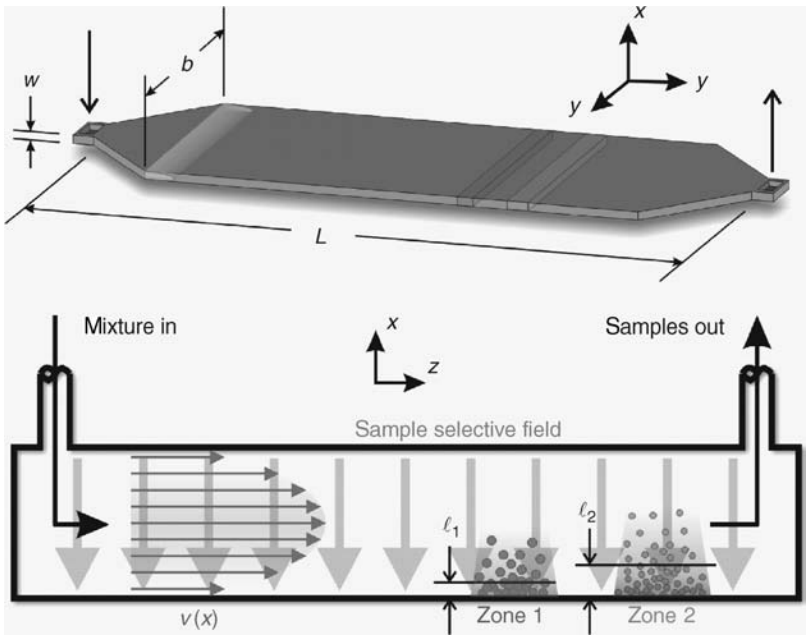


Figure 7. *Top*: Representation of a typical FFF channel. *Bottom*: Cross-sectional slice of channel in the center along the length and showing the operational principles of a separation in FFF. The required components for retention are a parabolic velocity profile and sample selective field.

retention and separation in some cases. Some of the systems that benefit through field enhancement from reducing the scale are thermal, electric, magnetic, and dielectrophoretic. In addition, microscale and nanoscale fabrication technologies have also made possible the implementation of systems that were previously impossible, difficult, or unreasonable to try, such as acoustic, magnetic, and dielectrophoretic FFF, as well as combinations of these systems.

Several microscale systems have been successfully implemented: micro-electric FFF (56–60); micro-thermal FFF (61–66); and AcFFF by Edwards and Frazier (67). Gale also integrated an electrical impedance detector within the channel to minimize the plate height due to extracolumn volumes in the detector and between the column and detector (56). Both Gale and Edwards also developed sample injection methods to minimize plate height further (56,61).

Systems manufactured using microelectricalmechanical system (MEMS) technologies are not only suitable for creating an inexpensive, disposable analysis device, but also for integrating with other methods such as chromatography, sensors, fluid handling devices, and actuators to create a total analysis system, or lab-on-a-chip. FFF can be used as a sample preparation tool, analytical device, or both in these systems.

CONCLUSION

Chromatography and the closely related fields of FFF and electrophoresis have proven to be valuable methods over the last century for separation, purification, and analysis. As a result of the wide variety and combinations of phases used in chromatography and fields applied in FFF, the number of types of samples that have been or can be retained and separated in these systems appears to be endless. Further advances in column, carrier, and detector technology, such as miniaturization, will continue to push the limits outward and make available faster and higher quality separations. In turn, researchers and industries in fields such as chemical engineering, bioengineering, chemistry, and pharmaceutics that rely on these techniques will also advance.

BIBLIOGRAPHY

1. Tswett M. Physical chemical studies on chlorophyll adsoptions. Berichte der Deutschen botanischen Gesellschaft 1906;24: 316–323.
2. Lesney MS. A brief history of "color writing." Today's Chemist at Work 1998;7(8):67–72.
3. Dorsey JG, et al. Liquid chromatography: Theory and methodology. Analyt Chem 1998;70(12):591R–644R.
4. Carrier R, Yip K. Intro to Chromatography. Available http://www.rpi.edu/dept/chem-eng/Biotech-Environ/CHROMO/chromintro.html.
5. Hardy JK. Analyt Chem. Available http://ull.chemistry.vakron.edu/analytical/.
6. Kazakevich Y, McNair H. Basic liquid chromatography. Available http://hplc.chem.shu.edu/NEW/HPLC_Book/.
7. van Deemter J, et al. Longitudinal diffusion and resistance to mass transfer as causes of nonideality in chromatography. Chem Eng Sci 1956;5:271–289.
8. Brede C, Pedersen-Bjergaard S. State-of-the art of selective detection and identification of I^-, Br^-, Cl^-, and F-containing compounds in gas chromatography and liquid chromatography. J Chromatogr A 2004;1050(1):45.
9. Gonzalez FR. Application of capillary gas chromatography to studies on solvation thermodynamics. J Chromatog A 2004; 1037(1-2):233.
10. Roubani-Kalantzopoulou F. Determination of isotherms by gas-solid chromatography: Applications. J Chromatogr A 2004;1037(1-2):191.
11. He L, Beesley TE. Applications of enantiomeric gas chromatography: A review. J Liquid Chromatogr Related Technol 2005; 28(7-8):1075.
12. Abraham MH, et al. Hydrogen bonding. 42. Characterization of reversed-phase high-performance liquid chromatographic C-18 stationary phases. J Phys Organ Chem 1997;10(5):358–368.
13. Sherma J. High-performance liquid chromatography/mass spectrometry analysis of botanical medicines and dietary supplements: A review. J AOAC Int 2003;86(5):873.
14. Petrovic M, et al. Liquid chromatography-tandem mass spectrometry for the analysis of pharmaceutical residues in environmental samples: A review. J Chromatogr A 2005;1067(1-2):1.
15. Kuhn R. Hoffstetter-Kuhn S. Capillary Electrophoresis: Principles and Practice. Berlin, Germany: Springer-Verlag; 1993.
16. Baker DR. Capillary Electrophoresis. New York: Wiley-Interscience; 1995.
17. Horvath C, Nikelly JG. Capillary Electrophoresis and Chromatography. Washington, DC: American Chemical Society; 1990.
18. Nouadje G, et al. Capillary electrophoresis with laser-induced fluorescence detecion: Optical design and applications. Progress in HPLC-HPCE 1997;5:49–72.
19. Beale SC. Capillary electrophoresis. Analyt Chem 1998;70:279R–300R.
20. Effenhauser CS. Integrated chip-based microcolumn separation systems. Topics Curr Chem 1998;194:51–81.
21. Regnier FE, et al. Chromatography and electrophoresis on chips: critical elements of future integrated, microfluidic analytical systems for life science. Trends Biotechnol 1999;17(3): 101–106.
22. Dolnik V, et al. Capillary electrophoresis on microchip. Electrophoresis 2000;21:41–54.
23. Starkey DE, et al. A fluorogenic assay for b-glucuronidase using microchip-based capillary electrophoresis. J Chromatogr B 2001;762:33–41.
24. Effenhauser CS, et al. Glass chips for high-speed capillary electrophoresis separations with submicrometer plate heights. Analyt Chem 1993;65:2637–2642.
25. Jacobson SC, et al. Effects of injection schemes and column geometry on the performance of microchip electrophoresis devices. Analyt Chem 1994;66:1107–1113.
26. Effenhauser CS, et al. Integrated chip-based capillary electrophoresis. Electrophoresis 1997;18:2203–2213.
27. Martynova L, et al. Fabrication of plastic microfluidic channels by imprinting methods. Analyt Chem 1997;69:4783–4789.
28. McCormick RM, et al. Microchannel electrophoretic separations of DNA in injection-molded plastic substrates. Analyt Chem 1997;69:2626–2630.
29. Colyer CL, et al. Clinical potential of microchip capillary electrophoresis systems. Electrophoresis 1997;18:1733–1741.
30. Effenhauser CS, et al. Integrated capillary electrophoresis on flexible silicone microdevices: Analysis of DNA restriction fragments and detection of single DNA molecules on microchips. Analyt Chem 1997;69:3451–3457.
31. Carrilho E. DNA sequencing by capillary array electrophoresis and microfabricated array systems. Electrophoresis 2000;21: 55–65.
32. Woolley AT, Mathies RA. Ultra-high-speed DNA sequencing using capillary electrophoresis chips. Analyt Chem 1995;67: 3676–4086.

33. Simpson PC, et al. High-throughput genetic analysis using microfabricated 96-sample capillary array electrophoresis microplates. Proc Nat Acad Sci USA 1998;95:2256–2261.
34. Chiem N, arrison HDJ. Microchip-based capillary electrophoresis for immunoassays: analysis of monoclonal antibodies and theophylline. Analyt Chem 1997;69:373–378.
35. Li PCH, Harrison DJ. Transport, manipulation, and reaction of biological cells on-chip using electrokinetic effects. Analyt Chem 1997;69:1564–1568.
36. Cabrera CR, Yager P. Continuous concentration of bacteria in a microfluidic flow cell using electrokinetic techniques. Electrophoresis 2001;22:355–362.
37. Kremser L, et al. Capillary electrophoresis of biological particles: Viruses, bacteria, and eukaryotic cells. Electrophoresis 2004;25:2282–2291.
38. Priego-Capote F, Castro MDLd. Dual injection capillary electrophoresis: Foundations and applications. Electrophoresis 2004;25:4074–4085.
39. Debesset S, et al. An AC electro-osmotic micropump for circular chromatographic applications. Lab Chip 2004;4:396–400.
40. Glasgow I, et al. Electro-osmotic mixing in microchannels. Lab Chip 2004;4:558–562.
41. Hofmann O, Che D, Cruickshank KA, Muller UR. Adaptation of capillary isoelectric focusing to microchannels on a glass chip. Analyt Chem 1999;71:678–686.
42. Tan W, Fan ZH, Qiu CX, Ricco AJ, Gibbons I. Miniaturized capillary isoelectric focusing in plastic microfluidic devices. Electrophoresis 2002;23:3638–3645.
43. Shimura K. Recent advances in capillary isoelectric focusing: 1997–2001. Electrophoresis 2002;23:3847–3857.
44. Kilar F. Recent applications of capillary isoelectric focusing. Electrophoresis 2003;24:3908–3916.
45. Simpson DC, Smith RD. Combining capillary electrophoresis with mass spectrometry for applications in proteomics. Electrophoresis 2005;26:1291–1305.
46. Schmitt-Kopplin P, Frommberger M. Capillary electrophoresis-mass spectrometry: 15 years of development and applications. Electrophoresis 2003;24:3837–3867.
47. Sung W-C, Makamba H, Chen S-H. Chip-Based microfluidic devices coupled with electrospray ionization-mass spectrometry. Electrophoresis 2005;26:1783–1791.
48. Smyth WF. Recent applications of capillary electrophoreis-electrospray ionization-mass spectrometry in drug analysis. Electrophoresis 2005;26:1334–1357.
49. Simo C, Barbas C, Cifuentes A. Capillary electrophoresis-mass spectrometry in food analysis. Electrophoresis 2005;26:1306–1318.
50. Shamsi SA, Miller BE. Capillary electrophoresis-mass spectrometry: Recent advances to the analysis of small achiral and chiral solutes. Electrophoresis 2004;25:3927–3961.
51. Zamfir A, Peter-Katalinic J. Capillary electrophoresis-mass spectrometry for glycoscreening in biomedical. Electrophoresis 2004;25:1949–1963.
52. Senk P, Kozak L, Foret F. Capillary electrophoresis and mass spectrometry for screening of metabolic disorders in newborns. Electrophoresis 2004;25:1447–1456.
53. IV WRV, Pasas-Farmer SA, Fischer DJ, Frankenfeld CN, Lunte SM. Recent developments in electrochemical detection for microchip capillary electrophoresis. Electrophoresis 2004; 25:3528–3549.
54. Qiu H, Yan J, Sun X, Liu J, Cao W, Yang X, Wang E. Microchip capillary electrophoresis with an integrated indium tin oxide electrode-based electrochemiluminescence detector. Analyt Chem 2003;75:5435–5440.
55. Wolters AM, Jayawickrama DA, Webb AG, Sweedler JV. NMR detection with multiple solenoidal microcoils for continuous-flow capillary electrophoresis. Analyt Chem 2002;74:5550–5555.
56. Gale BK, et al. A micromachined electrical field-flow fractionation (mu-EFFF) system. IEEE Trans Biomed Eng 1998;45(12):1459–1469.
57. Gale BK, et al. Geometric scaling effects in electrical field flow fractionation. 1. Theoretical analysis. Analyt Chem 2001;73 (10):2345–2352.
58. Gale BK, et al. Geometric scaling effects in electrical field flow fractionation. 2. Experimental results. Analyt Chem 2002;74(5):1024–1030.
59. Gale BK. Novel techniques and instruments for field flow fractionation of biological materials. Abstr Papers Am Chem Soc 2003;225:U138–U138.
60. Gale BK. Miniaturized field flow fractionation systems. Abstr Papers Am Chem Soc 2004;227:U116–U116.
61. Edwards TL, et al. A microfabricated thermal field-flow fractionation system. Analyt Chem 2002;74(6):1211–1216.
62. Schimpf ME, Polymer analysis by thermal field-flow fractionation. J Liquid Chromatogr Related Technol 2002;25 (13-15):2101–2134.
63. Janca J. Micro-channel thermal field-flow fractionation: High-speed analysis of colloidal particles. J Liquid Chromatogr Related Technol 2003;26(6):849–869.
64. Janca J, Ananieva IA. Micro-thermal field-flow fractionation in the characterization of macromolecules and particles: Effect of the steric exclusion mechanism. E-Polymers 2003.
65. Janca J, et al. Effect of channel width on the retention of colloidal particles in polarization, steric, and focusing micro-thermal field-flow fractionation. J Chromatogr A 2004;1046 (1-2):167–173.
66. Bargiel S, et al. A micromachined system for the separation of molecules using thermal field-flow fractionation method. Sens Actuators A-Phys 2004;110(1-3):328–335.
67. Edwards TL. Microfrabricated acoustic and thermal field-flow fractionation systems. Electrical and computer engineering. Atlanta, GA: Georgia Institute of Technology; Ph.D. thesis, 2005. p 300.

See also ANALYTICAL METHODS, AUTOMATED; PHARMACOKINETICS AND PHARMACODYNAMICS; TRACER KINETICS.

CO_2 ELECTRODES

JOHN W. SEVERINGHAUS
University of California in San Francisco
San Francisco, California

METHODS OF MEASURING BLOOD pCO_2 BEFORE DISCOVERY OF THE pCO_2 ELECTRODE

Bubble Equilibration Methods

Carbon dioxide in blood is largely in the form of the bicarbonate ion, which could be converted to CO_2 gas by adding acid and extracting the gas in a vacuum. The concept of partial pressures gradually stimulated interest in measuring pCO_2 in the late nineteenth century. Gas analysis had been developed earlier, so the first method was to equilibrate a small gas bubble with a large volume of blood sample at body temperature, and then remove the bubble for gas analysis. Pflüger developed a tonometer for this purpose in the 1870s, and August Krogh used this

method in fish in the early twentieth century. It was developed into a clinical and laboratory method by Richard Riley, using a specially adapted syringe with a capillary attached, which was invented by F. J. W. Roughton and P. Scholander during World War II. Riley's bubble method worked well for pCO_2, but poorly for pO_2 especially when blood was saturated with oxygen.

The Henderson–Hasselbalch Method

The most accurate early method was made possible by L. J. Henderson's discovery of buffering and his equation in 1908, its logarithmic modification by K. A. Hasselbalch in 1916, and P. T. Courage's design of a glass pH electrode (1925) in which blood could be measured with little loss of CO_2 to air. Blood pH was determined, usually at room temperature, there being no thermostated electrodes, and the Rosenthal temperature correction (-0.0147 pH units/°C) was used to compute pH at 37 °C. Plasma CO_2 content was determined in the Van Slyke manometric apparatus that used 1 mL of plasma (after carefully centrifuging blood under a gas tight seal, a floating cork). This method reached a precision of 0.3 mmHg pCO_2 in studies of the arterial to alveolar pCO_2 difference during surgical hypothermia at The National Institute of Health (NIH) (5).

Astrup and The Equilibration Method

Hundreds of patients with polio needed artificial ventilation in the communicable disease hospital in Copenhagen during epidemics in 1950–1952. Poul Astrup, M.D. (Professor of Clinical Chemistry, University of Copenhagen, Copenhagen, Denmark, and Director of the Clinical Laboratory, Rigshospitalet, Copenhagen, Denmark) and his associates, particularly Ole Siggaard Andersen, Ph.D., M.D. (Professor of Clinical Chemistry, University of Copenhagen, and Director, Clinical Chemistry Laboratory, Herlev Hospital, Copenhagen, Denmark), devised a way of determining blood pCO_2 using only a pH electrode to measure pH before and after equilibration of a blood sample with two known concentrations of pCO_2(6). Astrup made use of the little known fact that, as pCO_2 is changed, the relationship of pH to pCO_2 in a given blood sample is semilogarithmic (Fig. 1). By plotting the two measured values of pH at the known equilibrated pCO_2, he could graphically interpolate the pCO_2 from the original sample pH.

From 1954 until the mid-1960s, Astrup's method was made widely available by the Radiometer Co. of Copenhagen. The device had a thermostated capillary pH electrode, reference electrode, and tiny shaking equilibrator through which humidified gas flowed. Astrup's apparatus and method became obsolete with the introduction of the CO_2 electrode.

Ole Siggaard Andersen, Astrup, and others used the values obtained for pH and pCO_2 to calculate bicarbonate, total CO_2, and base excess, a term they introduced as a quantitative measure of the nonrespiratory or metabolic abnormality in a whole blood sample. Base excess proved to be the first accurate index of the nonrespiratory component of acid–base balance (7). Its first application was only for blood, but by 1966, it was shown to apply to the extracellular fluid of the entire body if one assumed an average extracellular fluid hemoglobin concentration of 5 g/dL.

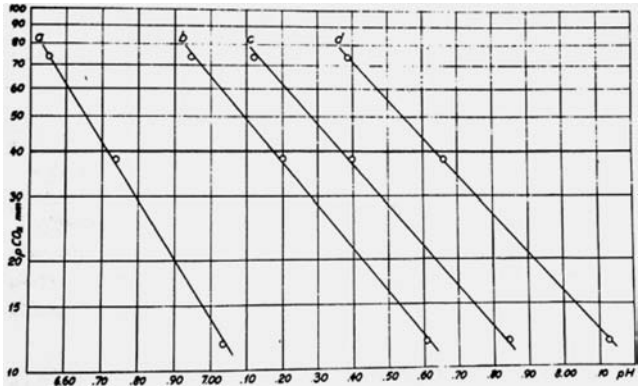


Figure 1. Equilibration method for measuring arterial pCO_2 introduced by Astrup during the Copenhagen polio epidemic, 1952–1954. Log pCO_2 plotted versus pH results in straight lines with varying pCO_2, and shifts of pH and slope when blood is acidified or alkalinized. The shift gave rise to the concept of base excess.

THE CO_2 ELECTRODE

History

A carbon dioxide (CO_2) electrode was first described by physiologists Gesell and McGinty at the University of Michigan in 1926, for use in expired air, but not in blood (8). It used the effect of CO_2 on the pH of a film of peritoneal membrane wet with a salt solution. Their paper was rediscovered 40 years later by M. Laver at Massachusetts General Hospital who informed Trubohovich of this effect (9).

In August 1954, Richard W. Stow, Ph.D. (Associate Professor of Physical Medicine, Ohio State University, Columbus, Ohio) (Fig. 2), a physical chemist, reported the design of a CO_2 electrode at the fall meeting of the American Physiologic Society in Madison, Wisconsin (10).

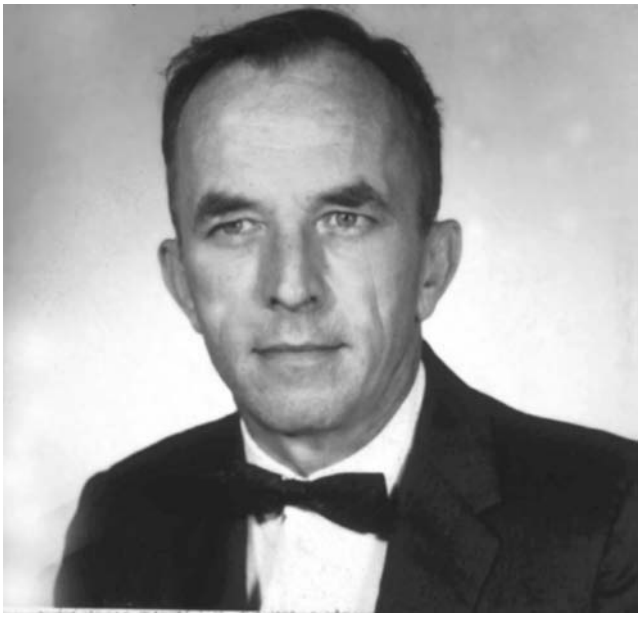


Figure 2. Richard Stow, invented the CO_2 electrode in 1954 to assist in managing polio patients on ventilators (10).

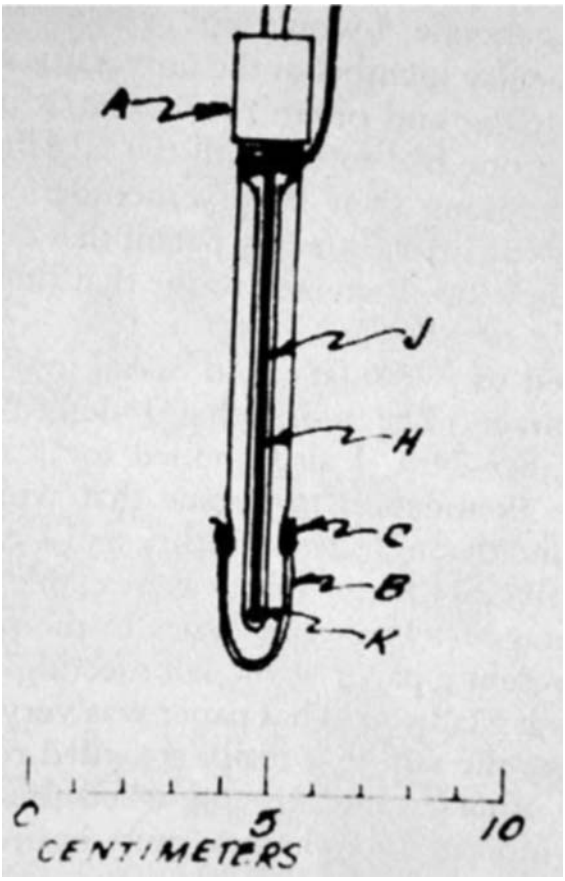


Figure 3. Stow's sketch of his 1954 CO_2 electrode. (a) Cable connection enclosure. (b) Rubber membrane. (c) Retaining O ring. (h) Chamber for internal pH electrolyte. (j) Reference electrode of silver chloride, not in contact with internal electrolyte, but opening to exterior through port K.

The polio epidemic was raging at the time, and as part of the physical therapy faculty he had sought some way to measure pCO_2 in the victims. He read in the library about specific ion electrodes, and conceived the electrode idea. He had wrapped a thin rubber membrane wet with distilled water over a homemade combined pH and reference electrode (Fig. 3). When he changed gas pCO_2 outside the device, the pH inside changed as a log function of gas pCO_2. However, he was unable to get stable readings and said he doubted it could be made useful.

After his talk, Severinghaus asked him why he did not try adding sodium bicarbonate ($NaHCO_3$) to the water film in the electrode. He replied that he believed this would abolish the signal because bicarbonate would buffer the effect of pCO_2 on pH. Severinghaus replied that he was confident that bicarbonate would not block the sensitivity. Stow agreed that Severinghaus would further investigate this idea. In September, 1954, after returning from Madison to the National Institutes of Health, Severinghaus confirmed the advantage of adding bicarbonate ions. A schematic diagram of his modification of Stow's electrode is shown in Fig. 4. He used a Beckman bulb-type pH electrode, a chloride-coated silver wire reference, and a Beckman pH meter. He tied a film of cellophane over the

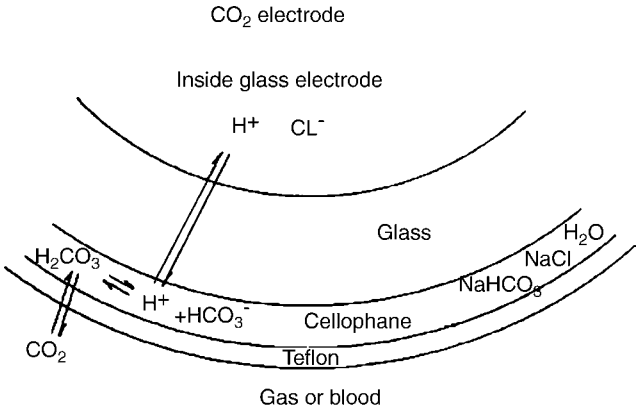


Figure 4. The concept of a pCO_2 electrode. The pH sensitive surface of a pH glass electrode is covered by a layer of electrolyte (here in cellophane), and then by a thin layer of a membrane permeable to CO_2, but not to hydrogen ions (here Teflon). The pH in that film is controlled by the partial pressure of CO_2 on the outside of the outer membrane as CO_2 dissolves and reacts with water to form carbonic acid. The carbonic acid dissociates into hydrogen and bicarbonate ions. Because the electrolyte has 5–20 m*M* HCO_3^- ions, changes of pCO_2 have no measurable effect on HCO_3^-. The mass law then requires H^+ to change in direct proportion to change in pCO_2. A doubling of pCO_2 doubles H^+ concentration, which is seen as a 0.3 pH unit fall.

pH electrode soaked in 25 m*M* $NaHCO_3$ and then covered the entire tip with a thin rubber dam, later from a surgical glove. The bicarbonate not only made the device stable, but doubled the pCO_2 sensitivity compared with an electrolyte of distilled water (or 1% NaCl). Salt was added to help stabilize the silver chloride reference electrode.

In 1957, Stow. Baer and Randall (11) published their discovery of the CO_2 electrode without mentioning the need to add bicarbonate ion, and took no further interest in this idea. Stow had no interest in a patent, thinking it would distract him from his job, and also because his university only allowed inventors 10% of royalties. As a U.S. government employee, Severinghaus was not permitted to patent it, certainly not with a reluctant coinventor.

Severinghaus and co-worker A. Freeman Bradley proceeded to investigate and optimize the electrode design and to test its performance, linearity, drift, and response time. They constructed electrodes for laboratory use by several colleagues, but unfortunately made no attempt at commercial development for 4 years.

Between 1958 and 1960 several other investigators constructed and published similar CO_2 electrodes, in several instances without being aware of the Stow–Severinghaus electrode (12–14).

CO_2 ELECTRODE DESIGN DETAILS

A CO_2 electrode consists of a slightly spherical surfaced glass pH electrode and a silver chloride reference electrode. Both are mounted in a glass or plastic sleeve holding a Teflon or silicone rubber membrane, typically 12 μm thick, over the glass surface, in some cases with a spacer of very thin lens-cleaning paper between membrane and glass to

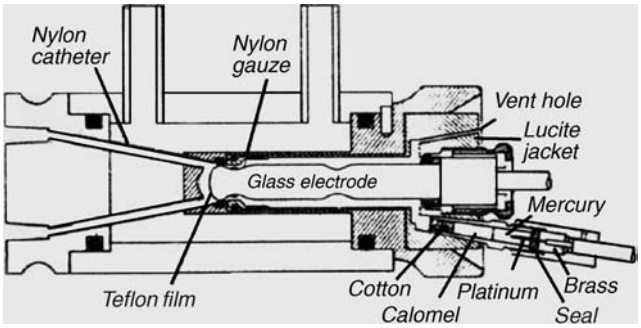


Figure 5. Cuvette with blood inlet and outlet connections in a thermostated water jacket made for the Stow–Severinghaus pCO_2 electrode (National Welding Co, San Francisco, 1959).

insure a uniform distribution of the electrolyte that is NaCl or KCl with ~5–20 mequiv/L of $NaHCO_3$. For use in blood, the electrode is mounted in a 37 °C cuvette into which a small sample of blood can be injected (typically 50 μL) (Fig. 5).

The electrode output voltage is a logarithmic function of pCO_2, ~60 mV for a 10-fold change of pCO_2, which induces a pH change of ~1 pH unit. Sensitivity is defined as Δ pH/$\Delta \log pCO_2$, where S reaches nearly the ideal maximum value of 1.0 with HCO_3^- concentrations of 5–25 mM (Fig. 6). At higher bicarbonate levels, carbonate acts as a buffer, and reduces both sensitivity and speed of response. Response is faster at lower bicarbonate concentration, but carbonic acid pK' is 6.1, resulting in some change of bicarbonate as pCO_2 changes, reducing sensitivity. As bicarbonate concentration is lowered, sensitivity falls to 30 mV/decade pCO_2 change, or $S = 0.5$, at zero bicarbonate.

The log response is almost linear from 5 to 700 mmHg pCO_2. The response time to a step change of pCO_2 is exponential with a 95% response time of ~30 s, depending on the membrane thickness and material, bicarbonate ion concentration and the thickness of the electrolyte layer over the glass electrode surface. It can be made to respond

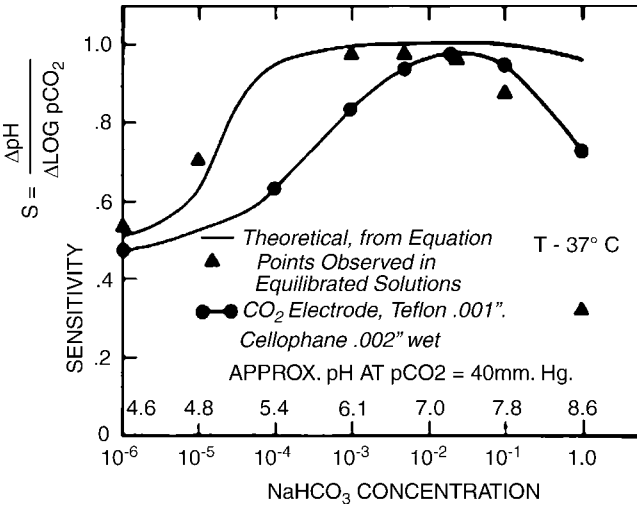


Figure 6. The first blood gas apparatus, with the Clark pO_2 electrode (below) in a stirred cuvette, and the Stow–Severinghaus pCO_2 electrode above, tilted to keep the internal air bubble of the pH electrode away from the tip (1957).

in < 1 s by using thin silastic (silicone rubber) membrane, low bicarbonate concentration (i.e., 1 mequiv/L), and adding carbonic anhydrase to the electrolyte, but the downside is loss of stability and signal amplitude.

The CO_2 electrode is usually calibrated to read in millimeters of mercury. It reads the same value for gas and liquid equilibrated with that gas at the electrode temperature, usually 37 °C.

A useful test of a leaking membrane is to equilibrate a dilute solution (e.g., 1 mequiv/L) of HCl, or lactic acid with a known calibration gas, and test its reading. Any leak will permit acid entry and an erroneously high pCO_2.

Maintenance requires replacement of the membrane and electrolyte when errors are detected or when drift has driven the electrode beyond the ability of the apparatus to compensate its potential. The pH glass may become so impermeable to hydrogen ions that it shows low sensitivity or slow response after years of use.

The amplifier circuit must be electrically isolated from the ground because any ground path leakage will draw current through the silver chloride reference and changes its potential causing drift. The input impedance of all modern pH and pCO_2 meter amplifiers is $>10^{11}\,\Omega$.

The Combined Blood Gas Analysis Apparatus

In 1956, Leland Clark disclosed his invention of the oxygen electrode at a meeting in Atlantic City to which Severinghaus had invited physiologists interested in measuring pO_2. That invention made a huge difference in blood gas analysis.

While Severinghaus completed his anesthesia residency at the University of Iowa, with help from the physiology workshop, he constructed a thermostat into which he mounted both the Stow–Severinghaus CO_2 electrode and the Clark O_2 electrode in a stirred cuvette with a small blood tonometer. That apparatus was exhibited at the meeting of the American Society of Anesthesiologists in October 1957 and at the meeting of the Federation of American Societies of Experimental Biology in Atlantic City in the spring of 1958 and published in 1958 (15) (Fig. 7).

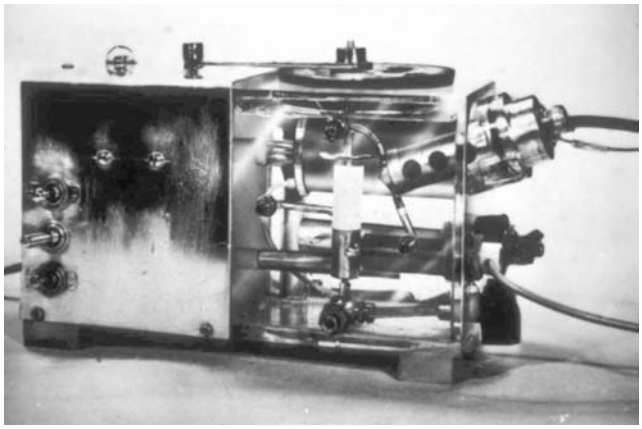


Figure 7. The first three-function blood gas analyzer, using a McInnes Belcher pH electrode (1930) with the pCO_2 and pO_2 electrodes in a 37 °C bath.

The Three-Function Blood Gas Analyzer

In 1958, after moving from the National Institutes of Health to the University of California, San Francisco, Severinghaus and Bradley added a pH electrode to the blood gas electrode waterbath, making the first three-function blood gas apparatus (Fig. 8). Forrest Bird, Ph.D., M.D. (President, Bird Corporation, Palm Springs, California) had designed popular positive-pressure ventilators, manufacturing them at the National Welding Co. in San Francisco. He proposed to manufacture the CO_2 electrode and to make it commercially available. From 1959–1961 the National Welding Co. sold the only available pCO_2 electrode. The design concept was soon copied and marketed by Beckman, Radiometer, Instrumentation Labs and later by several other firms.

Impact of Blood Gas Analysis

During the 1960s, blood gas analysis became widely available in anesthesia, intensive and critical care facilities, and cardiorespiratory research laboratories. For several years, the Severinghaus paper (15) was among the most quoted articles in biologic literature, and blood gases were called the most important laboratory test for critically ill patients. Blood gas apparatus now uses automatic self-calibration and automatic transport of sample and washing of cuvettes, printing of results, and often sending the values to remote terminals. In the United States, regulations have been used by pathologists to require that these automated instruments can only be used by licensed technicians, usually meaning that the income flows to pathologists. Gone are the days when students, nurses, residents, and faculty all took part in doing blood gas analysis.

A more complete history of the CO_2 electrode and related blood gas technology is available in References (9,16).

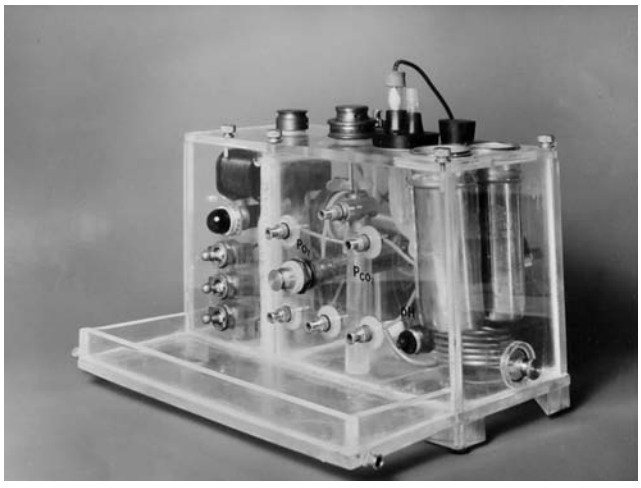


Figure 8. Relationship of pCO_2 electrode sensitivity to its internal electrolyte bicarbonate ion concentration. Maximum sensitivity occurs at $\sim 20\,mM$ HCO_3^-, but for faster response, most electrodes operate at 5–10 mM.

HISTORY AND THEORY OF TRANSCUTANEOUS BLOOD OXYGEN MONITORING

From 1951 to 1952, the discovery of oxygen related blindness in premature infants created an urgent need for continuous noninvasive monitoring of blood oxygen. A new solution to the problem came from physiologists studying skin respiration. Human skin breathes, taking up oxygen and giving off CO_2 to the air. If skin is covered (as by a flat unheated pCO_2 electrode) the surface tcpO_2 falls to zero in a few minutes. However, in 1951 Baumberger and Goodfriend showed that if skin blood flow is greatly increased by the highest tolerable heat (45 °C), the surface pO_2 rises to about paO_2 (arterial blood) (17).

Within a year after Clark's invention of the membrane covered platinum polarographic electrode (18,19), Rooth used polarography to confirm the Baumberger report (20). Researchers tried unsuccessfully to use chemical vasodilators to make skin pO_2 a monitor of paO_2. Kwan and Fatt (21) noted that pO_2 of the palpebral conjunctiva measured with an unheated tiny Clark electrode mounted facing outward on a contact lens over the cornea simulated paO_2. This device was briefly marketed a decade later, but discontinued due to the danger of infection.

In Marburg, Germany, Professor of Physiology Dietrich Lübbers and students, especially Renate Huch, pursued the concept of heating the skin under an oxygen electrode by heating the electrode itself to as high as 45 °C. They were joined by Patrick Eberhard, and the group soon found ways of making electrically heated, thermostated oxygen surface electrodes. By 1972, they had shown a good relationship between heated skin and arterial blood pO_2 in infants (22). Several firms began to design electrodes for this purpose.

DEVELOPMENT OF METHODS AND UNDERSTANDING OF THEORY

By 1977, the Marburg group had published at least 11 papers documenting the validity of transcutaneous oxygen measurement. At least three commercial tcpO_2 electrode systems were available (Helige, Roche, Radiometer). In November 1977, some 18 research teams joined for a workshop on transcutaneous blood gas methods in San Francisco, assessing the theory, problems, possibilities, and progress (23–30). The following summer (1978) many of these workers joined the Marburg team and others for the first international congress on transcutaneous blood gas monitoring, establishing the technology as an essential tool in neonatology and as useful in many other fields (31,32).

The agreement of tcpO_2 with paO_2 proved to be a cancellation of two opposing effects illustrated in Fig. 9 (27,33).

1. Heating of desaturated blood raises its pO_2 by 7%/°C, or 50% at 43 °C, but in saturated blood, as in water, pO_2 rises only 1.3%/°C (35);
2. Skin metabolism at the high temperature consumes O_2 as it diffuses outward from capillaries through living cells, reducing the value to about paO_2.

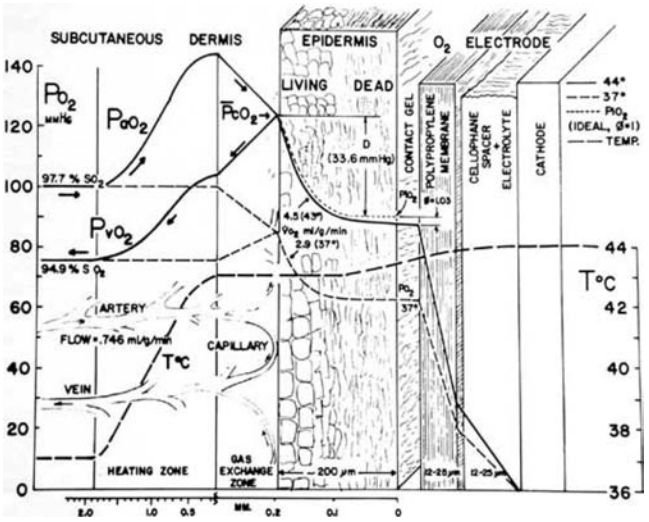


Figure 9. A schema of the effect of both heating of skin surface by a transcutaneous electrode, and of local metabolism, on the tissue internal oxygen tension from the arteries out past the capillaries and the living and dead epidermis to the surface, and through the electrode membrane into the cathode that keeps its surface $pO_2 = 0$ by its electrical negative potential (39).

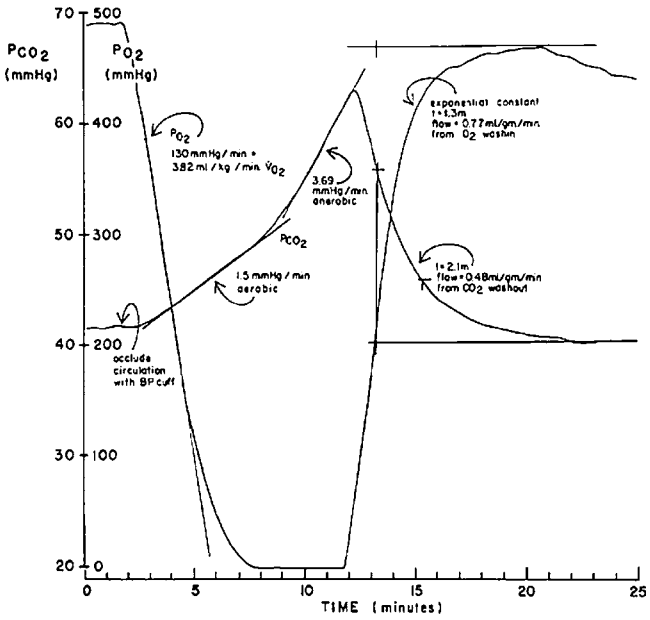


Figure 10. The time course of skin pO_2 and pCO_2 on an arm after sudden circulatory occlusion with a blood pressure cuff. The rate of fall of pO_2 from a high level is a measure of the skin metabolic rate. The pCO_2 rises at first from metabolic CO_2 production, but later at a steeper rate as skin generates lactic acid when skin pO_2 reaches zero. With release of occlusion, the electrode recovery time is delayed by both the skin washin and washout, and by electrode equilibration (34).

The outward oxygen diffusion is facilitated by heat that proved to "melt" some skin diffusion barriers (33,36). Skin O_2 conductivity C (adult volar forearm) was determined by two groups by comparison of flux with two membranes (teflon and mylar) of very high and low conductivity. With a large gold cathode Clark electrode, $C = 15\,\text{nL} \cdot \text{cm}^{-2} \cdot \text{s}^{-1} \cdot \text{atm}^{-1}$(37) and with a mass spectrometer $C = 10\,\text{nL} \cdot \text{cm}^{-2} \cdot \text{s}^{-1} \cdot \text{atm}^{-1}$(38).

Skin O_2 consumption (VO_2) was determined after thermal vasodilation by the rate of fall of tcpO_2 with circulatory occlusion (arm cuff) (Fig. 10) (27). Relative skin blood flow under the heated electrode was estimated by measuring the required heating power (39). Analysis of data collected at two levels of pO_2 and two temperatures permitted calculation of blood flow, capillary temperature under a heated electrode, and diffusion gradient from capillary to surface (40). Mean adult volar forearm skin VO_2 was $4.2 \pm 0.4\ \mu\text{L} \cdot \text{g}^{-1} \cdot \text{min}^{-1}$ at 44 °C and $2.8 \pm 0.3\ \mu\text{L} \cdot \text{g}^{-1} \cdot \text{min}^{-1}$ at 37 °C. At 44 °C, skin blood flow averaged $0.64 \pm 0.17\,\text{mL} \cdot \text{g}^{-1} \cdot \text{min}^{-1}$, capillary temperature was 43 °C and the diffusion gradient was 32 ± 7 mmHg.

TRANSCUTANEOUS CO_2

In 1959, Severinghaus constructed a 37 °C thermostated open tipped CO_2 electrode to determine pCO_2 of various tissue surfaces in animals (Fig. 11) (41). Without heating the skin well above body temperature, skin pCO_2 at 37 °C climbed steadily over one-half of an hour to > 80 mmHg. Dog intestinal mucosa and liver surfaces were very high.

Fifteen years later, the success in transcutaneous measurement of oxygen led to design and testing of electrodes to measure tcpCO_2 by Beran et al. (42,43), Huch et al. (44) and Severinghaus et al. (45,46). Combined tcpO_2–tcpCO_2 electrodes were initially described by Parker et al. (47) and Severinghaus (48). Figure 12 schematically shows the internal design of an early Radiometer combined electrode. Figure 13 shows the electrode with a membrane mounted.

When a heated combined pO_2–pCO_2 electrode is first attached to skin, the time needed to equilibrate is ~ 5 min for both electrodes, although the pO_2 electrode may show later small changes as thermal vasodilation slowly develops (Fig. 14). The response to step changes in alveolar and arterial pCO_2 is slower as seen in Fig. 15. Here the response is delayed both by the washout or washin of CO_2 into the tissue by blood flow, and the electrode's own delay. The response is pseudoexponential, a combination of the two delays, resulting in a 95% response times of ~ 10 min.

Without correction, tcpCO_2 is not similar to $paCO_2$. Heating of blood (and water) raises pCO_2 $\sim 4.6\%$/°C

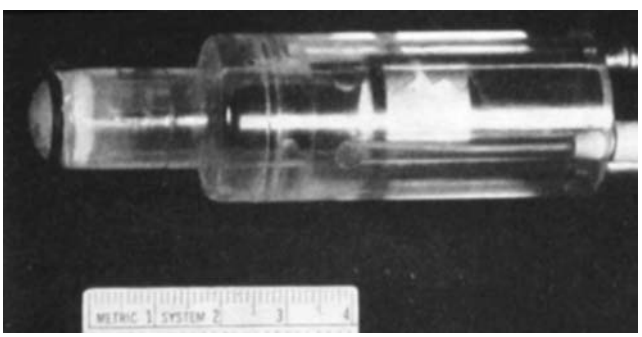


Figure 11. The first tissue surface pCO_2 electrode (41) with a circulating temperature controlled water jacket.

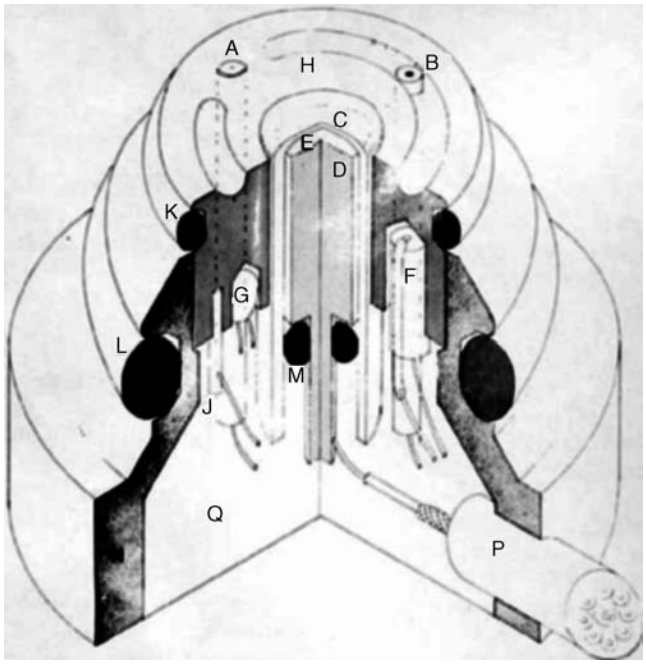


Figure 12. Schema of the design of a combined tcpO_2–tcpCO_2 electrode. (a) pO_2 cathode, the end of a 12 μm platinum wire fused in glass. (b) A silver wire reference electrode. (c) pH glass electrode surface. (d) solid silver internal pH electrode (used to improve heat transfer to skin). (e) Internal pH electrolyte. (f) Heater Zener diode. (g) Thermistor. (h) Silver body, and reference electrode. J, K, L, M: O rings. N: Lexan jacket. Q: epoxy. P: Cable (48).

Figure 13. Photograph of combined pO_2–pCO_2 electrode with teflon membrane (48).

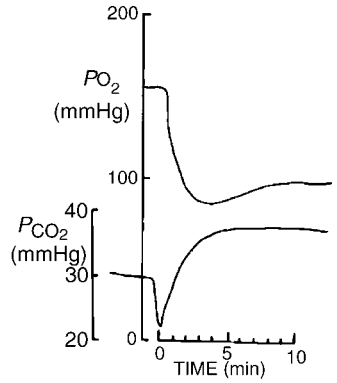


Figure 14. Initial responses of a combined pO_2–pCO_2 electrode when first mounted on skin. Both electrodes need ~5 min to equilibrate, the pO_2 showing a small late rise as skin hyperemia develops from the heating (49).

(41), metabolism adds ~3 mmHg pCO_2, and the cooling by skin and blood of the electrode surface further raises the electrode reading. The effect of heating on blood pCO_2 may be computed as $\Delta pCO_2 = \exp(0.046[T - 37])$ (51). The net effect at 43 °C was found to be tc$pCO_2 = 1.33paCO_2 +$ 4 mmHg (48,52) or tc$pCO_2 = 1.4paCO_2$ (53). This form of temperature-dependent correction factor was later incorporated in most commercial transcutaneous blood gas monitoring apparatus.

With this correction factor, the relationship of tcpCO_2 to $paCO_2$ is excellent, as shown in Fig. 16. The previous correction factors appear to have become incorrect for a second generation of the Radiometer tcpCO_2 electrodes, due to a design change in the internal temperature coefficient of the glass pH electrode. The additive factor of 4 mmHg changed to ~8 mmHg in the newer instruments (Kagawa S, personal communication).

Although tcpCO_2 appears to work at 42–43 °C, Tremper et al. showed that 44 °C was a better temperature when

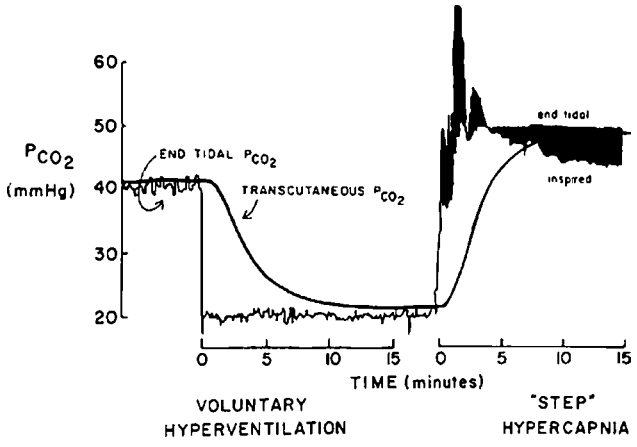


Figure 15. Transcutaneous pCO_2 electrode response to a step increase of ventilation adjusted by the subject to reduce end tidal pCO_2 suddenly from 40 to 20 mmHg and hold it at a constant level for 18 min, followed by addition of enough CO_2 to inspired gas to raise end tidal pCO_2 as quickly as possible to ~50 mmHg. The response time constants (63%) are ~5 min (45).

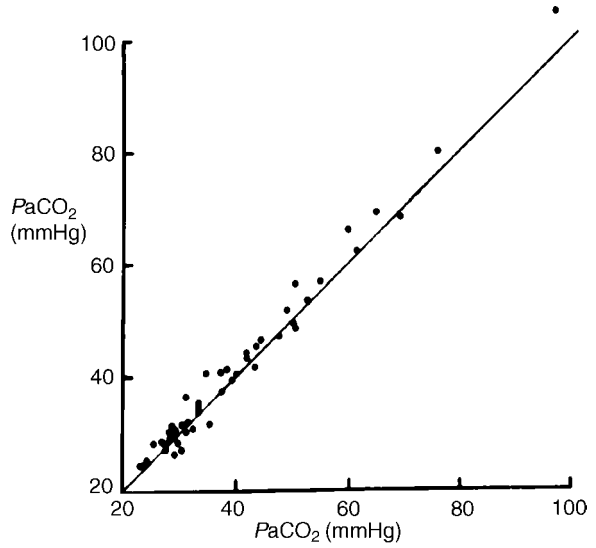


Figure 16. Transcutaneous pCO_2 correlates well with arterial pCO_2 in patients during anesthesia or intensive care (48).

blood pressure was or had been low (54). The tcpCO_2 value was better than the $P_{ET}CO_2$ value (end-tidal or end-expired air) in predicting $paCO_2$ (bias and s.d. -1.6 ± 4.3 mmHg) in anesthetized adults ($n = 24$) (55).

A special advantage of tcpCO_2 is that it averages out breath-by-breath variations, and has almost no inherent "noise" or variability, such that it often is found to be the best trend monitor for detecting small changes in $paCO_2$ such as those induced by experimental variations (anesthesia, ventilatory settings, posture, F_IO_2, F_ICO_2, blood pressure, pharmacologic agents, etc).

APPLICATIONS

Transcutaneous technology is used in many ways, some of which are discussed in accompanying papers:

1. Neonatology: Guidance of O_2 therapy remains the most common use of transcutaneous monitoring (56–58). The suspected etiologic role of hyperoxia (tc$pO_2 > 80$ mmHg) in retinitis of premature infants has been confirmed in a cohort study (59). The tcpO_2 value can be measured above and below the ductus to demonstrate closure (60). In low birth weight infants, tcpCO_2 (at 40 °C!) is the best available monitor of ventilation (61).
2. Fetal Monitoring: Using specially designed electrodes attached to the fetal scalp, intrapartum monitoring revealed some important new pathophysiologic understanding (62–65). As hoped, changes in tcpO_2 rapidly reflected changing maternal and fetal conditions (66). The tcpO_2 value fell and tcpCO_2 rose with contractions during the second stages of labor (67). The tcpCO_2 value closely followed fetal $paCO_2$(68). When there were signs of fetal distress, fetal scalp tcpO_2 was < 15 mmHg (69). Surprisingly, O_2 administration to mothers with fetal distress did not alter fetal pCO_2 or raise pO_2(70). During maternal hypocapnia, fetal tcpO_2 fell due to the Bohr effect, whereas it rose during hypercapnia (71). Fetal tcpO_2 was considered influenced by local scalp blood flow (72). Repeated episodes of asphyxia were reported to express catecholamines, which reduced blood flow to the fetal skin, artifactually reducing tcpO_2(73,74). Fetal tcpCO_2 may have failed to disclose severe acidosis or circulatory impairment (75).
3. Sleep Studies: Combined pO_2–pCO_2 electrodes are used in sleep studies in combination with pulse oximetry, because nostril sampling of end-tidal pCO_2 is somewhat annoying and more apt to become plugged or dislodged (76–83). The combined tcpO_2–tcpCO_2 electrode made it possible to show that the ventilatory response to induced mild hypoxia in sleeping infants changes with age from acute depression at 1–5 days, to stimulation at 4–8 weeks, and mild or no stimulation at 10–14 weeks (84). A method was designed for estimating the ventilatory response to CO_2 during sleep using capnography and tcpCO_2(79).
4. Peripheral Circulation: The tcpO_2 electrodes are extensively used in evaluating arterial disease in the peripheral circulation (85–88). A test of adequacy of peripheral circulation, "initial slope index" (ISI) was suggested by Lemke and Lübbers (89). Blood flow is stopped by an arm cuff above the electrode and restarted when tc$pO_2 = 0$. The initial rate of rise should be a slope per min of at least 75% of the preocclusion tcpO_2.
5. Skin Circulation: Monitoring the viability of skin after injury or transplant or flap movement (90,91).
6. Ventilatory Control: In intensive care, transcutaneous electrodes greatly increased the safety and simplicity of PEEP optimization and respiratory management of adults with respiratory distress syndrome (92). They are widely used simply to reduce arterial blood sampling.
7. Hyperbaric Oxygen: Monitoring and guiding hyperbaric oxygen therapy, primarily for infections and wound healing (93,94). The tcpO_2 tracked paO_2 up to 4-atm hyperbaric pressure in normal subjects (95). Surprisingly, no one has reported using tcpO_2 in hyperbaric treatment of CO poisoning despite the demonstration by Barker and Tremper in experimental CO administration that transcutaneous pO_2 falls linearly as COHb increases, and reaches about one-fifth of its initial value at the highest COHb levels despite the maintenance of constant arterial pO_2(96). It is thus unknown whether HBO can normalize tissue pO_2 in the presence of high levels of COHb.
8. Clinical Physiology: Transcutaneous monitoring has found use in exercise tolerance studies (97,98). End-tidal CO_2 is not exactly equal to $paCO_2$ and the difference between them varies with posture and inspired oxygen concentration. When testing hypoxic ventilatory responses by monitoring $P_{ET}CO_2$, tcpCO_2 helps to correct these small errors (99).

9. Pharmacologic Research: Transcutaneous monitoring may be the simplest monitor of the depressant effects of opiates, sedatives, and anesthetics especially in awake children (100).
10. Animal Studies: Intestinal or other tissue animal experimental ischemia has been found to be better detected by the rise of the organ or tissue surface pCO_2 using tcpCO_2 electrodes at body temperature than by gastric tonometry (101). Both tcpO_2 and tcpCO_2 have been widely used in small and large animal studies (102) and to assess the effect of cardiopulmonary resusitation (CPR) (103).

ACCURACY

With the widespread use of tcpO_2 and tcpCO_2 came concern about its accuracy and the possible sources and effects of errors, especially with severe hypotension (28,104). Peabody et al.(25) identified two groups of infants in whom tcpO_2 was lower than paO_2. These were infants receiving an intravascular infusion of tolazoline and infants with mean arterial blood pressures >2.5 s.d. below the predicted average value. Vasoconstrictors also lower tcpO_2(105). Both of these situations represent extreme alterations in peripheral blood flow. Mild hypotension, hypothermia, anemia, radiant warmers, and bilirubin lights did not adversely affect transcutaneous accuracy (106). In a large multiinstitutional study of 327 patients older than 1 month, when paO_2 was between 80 and 220 mmHg, Palmisano found the mean bias $\pm$ s.d. of tcpO_2 was -43 ± 40 mmHg, and the slope of the regression was 0.65 (107). It was determined that tcpCO_2 correlated far better with $paCO_2$: $R = 0.929$, slope 1.052, bias and s.d. $= 1.3 \pm 4.0$ mmHg ($n = 756$).

Defining a tcpO_2 index as tcpO_2/paO_2, Tremper and Shoemaker (108) studied the effect of shock. For 934 data sets taken on 92 patients not in shock, there was a correlation coefficient (r) of 0.89 and a tcpO_2 index 0.79 ± 0.12 (SD). In five patients with moderate shock, the r was 0.78 and the tcpO_2 index was 0.48 ± 0.07. In nine patients with severe shock, there was no correlation between tcpO_2 and paO_2 and the tcpO_2 index was 0.12 ± 0.12.

LIMITATIONS

Skin burns may occur after an electrode has been in one place over several hours at 44–45 °C, and sometimes even at 43 °C. Long-term monitoring requires site changes, or a dual electrode alternating system (109). There may be problems with drift of calibration, membrane failure, partial loss of skin contact giving errors in both O_2 and CO_2 readings. Maintenance of these electrodes requires training and some technical proficiency.

IMPACT OF PULSE OXIMETRY

Pulse oximetry came into widespread use in 1985–1987, and quickly replaced transcutaneous blood gas analysis in many situations. However, after an initial switch to oximetry, neonatologists found that oximetry failed to detect hyperoxia adequately (110) and now mostly use both technologies (111–115). In neonatology, a significant problem is that the inherent errors of pulse oximetry are $\sim 3\%$, which could fail to warn of $paO_2 > 80$ unless a set point of $\sim 90\%$ S_pO_2 is chosen (116). Some have arbitrarily dismissed transcutaneous monitoring as "...plagued by technical problems, ...Its use in efforts to prevent retinopathy of prematurity, an eye disease of preterm newborns often leading to blindness, proved disappointing" (117). To them, the transcutaneous field served as a model of problems in medical innovation, new technology, and personnel training. Not everyone agrees with this pessimism. Most technical problems have been solved, and the occurrence of blindness in very premature infants is now believed to be multifactorial, not just due to hyperoxia. Therefore when it occurs, it is not appropriate to attribute it to failed transcutaneous methodology.

CONCLUSIONS

The enthusiasm for transcutaneous blood gas analysis of the period 1976–1986 was followed by a decrease due to the advent of pulse oximetry. The number of papers per year listing medline keywords "transcutaneous blood gas" reached an early peak of 75 in 1979, when the first international symposium was devoted to this field, in Marburg and 200 in 1987. However, after 1986 many papers used the keywords "transcutaneous blood gas" when writers meant to refer to pulse oximetry.

Transcutaneous technology is inherently somewhat complicated. Users must change membranes and calibrate, change skin sites periodically to avoid burns, beware of drift or error due to poor circulation or poor skin attachment, and take account of the slower response than given by oximetry. Nonetheless, transcutaneous blood gas measurement continues to be used because of its unique ability to meet many special situations needing its characteristics of noninvasively and continuously determining partial pressures of O_2 and CO_2. Several professional organizations have published guidelines for use of these monitors (118,119).

BIBLIOGRAPHY

1. Severinghaus JW. The current status of transcutaneous blood gas analysis and monitoring. Blood Gas news (Radiometer house organ) 1998;7:4–9.
2. Severinghaus JW. The Invention and Development of Blood Gas Apparatus. Anesthesiology 2002;97:253–256.
3. Severinghaus JW. Severinghaus electrode. In: JR Maltby, editor. Notable Names in Anaesthesia. London: Royal Society of Medicine Press Ltd.; 2002.
4. Severinghaus JW, Astrup P, Murray J. Blood gas analysis and critical care medicine. Am J Respir Crit Care Med 1998;157:S114–S122.
5. Severinghaus JW, Stupfel MA, Bradley AFJ. Accuracy of blood pH and pCO_2 determinations. J Appl Physiol 1956;19: 189–196.
6. Astrup P. A simple electrometric technique for the determination of carbon dioxide tension in blood and plasma, total

content of carbon dioxide in plasma and bicarbonate content in "separated" plasma at a fixed carbon dioxide tension. Scand J Clin Lab Invest 1956;8:33–43.
7. Siggaard-Andersen O, Engel K, Jorgensen K, Astrup P. A micro method for determination of pH, carbon dioxide tension, base excess and standard bicarbonate in capillary blood. Scand J Clin Lab Invest 1960;12:172–176.
8. Gesell R, McGinty DA. Regulation of respiration: VI. Continuous electrometric methods of recording changes in expired carbon dioxide and oxygen. Am J Physiol 1926;79:72–90.
9. Trubuhovich RV. History of pCO_2 electrodes. Br J Anaesth 1970;42:360–362.
10. Stow RW, Randall BF. Electrical measurement of the pCO_2 of blood (abstract). Am J Physiol 1954;179:678.
11. Stow RW, Baer RF, Randall B. Rapid measurement of the tension of carbon dioxide in blood. Arch Phys Med Rehabil 1957;38:646–650.
12. Gertz KH, Loeschcke HH. Elektrode zur bestimmung des CO_2 drucks. Naturwissenschaften 1958;45:160–161.
13. Hertz CH, Siesjo B. A rapid and sensitive electrode for continuous measurement of pCO_2 in liquids and tissue. Acta Physiol Scand 1959;47:115–123.
14. Snell FM. Electrometric measurement of carbon dioxide and bicarbonate ion. J Appl Physiol 1960;15:729–732.
15. Severinghaus JW, Bradley AF. Electrodes for blood pO_2 and pCO_2 determination. J Appl Physiol 1958;13:515–520.
16. Severinghaus JW, Astrup P. History of blood gas analysis. Int Anesthesiol Clin 1987;25:69–95.
17. Baumberger JP, Goodfriend RB. Determination of arterial oxygen tension in man by equilibration through intact skin. Fed Proc 1951;10:10.
18. Clark LC. Monitor and control of tissue O_2 tensions. Trans Am Soc Artif Intern Organs 1956;2:41–48.
19. Clark LC, Clark EW. Personalized history of the Clark oxygen electrode. Inter Anesthesiol Clin 1987;25:1–30.
20. Rooth G, Sjostedt S, Caligara F. Bloodless determination of arterial oxygen tension by polarography. Sci Tools LKW Instr J 1957;4:37.
21. Kwan M, Fatt I. A noninvasive method of continuous arterial oxygen tension estimation from measured palperal conjunctival oxygen tension. Anesthesiology 1971;35:309–314.
22. Huch R, Lübbers DW, Huch A. Quantitative continuous measurement of partial oxygen pressure on the skin of adults and new-born babies. Pflügers Arch 1972;337:185–198.
23. Vesterager P. Transcutaneous pCO_2 electrode. Scand J Clin Lab Invest 1977;37:27–30.
24. Friis Hansen B. Transcutaneous measurement of arterial blood oxygen tension with a new electrode. Scand J Clin Lab Invest 1977;37:31–36.
25. Peabody JL, Willis MM, Gregory GA, Tooley WH, Lucey JF. Clinical limitations and advantages of transcutaneous oxygen electrodes. Acta Anaesthesiol Scand Suppl 1978;68:76–82.
26. Tremper KK, Huxtable RF. Dermal heat transport analysis for transcutaneous O_2 measurement. Acta Anaeshesiol Scand Suppl 1978;68:4–8.
27. Severinghaus JW, Stafford MJ, Thunstrom AM. Estimation of skin metabolism and blood flow with $tcPo_2$ and $tcPco_2$ electrodes by cuff occlusion of the circulation. Acta Anaesth Scand Suppl 1978;68S:9–15.
28. Versmold HT, Linderkamp O, Holzmann M, Strohhacker I, Riegel KP. Limits of $tcpO_2$ monitoring in sick neonates: Relation to blood pressure, blood volume, peripheraal blood flow and acid base status. Acta Anaesthesiol Scand Suppl 1978;S68:88–90.
29. Kimmich HP, Kreutzer F. Model of oxygen transport through skin as basis for absolute transcutaneous measurement of PaO_2. Acta Anaesthesiol Scand Suppl 1968;S68:16–19.
30. Fatt I. Transmucosal measurement of blood pH at the palpebral conjunctiva. Acta Anaesthesiol Scand Suppl 1978;S68: 142–144.
31. Huch A, Huch R. The development of the transcutaneous pCO_2 technique into a clinical tool. In: Huch R, Huch A, Lucey JR, editors. Continuous Transcutaneous Blood Gas Monitoring, Birth Defects: Original Article Series. Volume XV-No. 4. New York: A.R.Liss; 1979.
32. Lübbers DW, Cutaneous and Transcutaneous pO_2 and pCO_2 and their measuring conditions. In: Huch R, Huch A, Lucey JF, editors. Continuous Transcutaneous Blood Gas Monitoring, Birth Defects: Original Article Series. Volume XV-No. 4. New York: A. R. Liss; 1979.
33. Lübbers DW. Theoretical basis of the transcutaneous blood gas measurements. Crit Care Med 1981;9:721–733.
34. Severinghaus JW. Transcutaneous Blood Gas Analysis. Respir Care 1982;27:152–159.
35. Severinghaus JW. Simple, accurate equations for human blood O_2 dissociation computations. J Appl Physiol 1979;46: 599–602.
36. Lübbers DW. Theory and development of transcutaneous oxygen pressure measurement. Int Anesthesiol Clin 1987; 25:31–65.
37. Eberhard P, Severinghaus JW. Measurement of heated skin O_2 diffusion conductance and pCO_2 sensor induced O_2 gradient. Acta Anaesthesiol Scand Suppl 1978;68:1–3.
38. Hansen TN, Sonoda Y, McIlroy MB. Transfer of oxygen, nitrogen and carbon dioxide through normal adult human skin. J Appl Physiol 1980;49:438–443.
39. Parker D, Delpy D, Reynolds EOR, St. Andrew D. A transcutaneous pO_2 electrode incorporating a thermal clearance local blood flow sensor. Acta Anaesthesiol Scand Suppl 1978; S68:33–39.
40. Thunstrom AM, Stafford MJ, Severinghaus JW. A two temperature, two pO_2 method of estimating the determinants of $tcpO_2$. In: Huch R, Huch A, Lucey JR, editors. Continuous Transcutaneous Blood Gas Monitoring, Birth Defects: Original Article Series. Volume XV-No. 4, New York: A. R. Liss; 1979.
41. Severinghaus JW. CO_2 Spannung und Perfusion in Gewebe. Anaesthetist 1960;9:50–55.
42. Beran AV, Huxtable RF, Sperling DR. Electrochemical sensor for continuous transcutaneous pCO_2 measurement. J Appl Physiol 1976;41:442–447.
43. Beran AV, Shigezawa GY, Yeung HN, Huxtable RF. An improved sensor and a method for transcutaneous CO_2 monitoring. Acta Anaesthesiol Scand Suppl 1978;S68:111–117.
44. Huch A, Seiler D, Meinzer K, Huch R, Galster H, Lübbers DW. Transcutaneous pCO_2 measurement with a miniaturised electrode. Lancet 1977;1:982–983.
45. Severinghaus JW, Stafford M, Bradley AF. $tcpCO_2$ electode design, calibration and temperature gradient problems. Acta Anaesthesiol Scand Suppl 1978;68:118–122.
46. Severinghaus JW, Bradley AF, Stafford MJ. Transcutaneous pCO_2 electrode design with internal silver heat path. In: Huch A, Huch R, Lucey JF, editors. Continuous Transcutaneous Blood Gas Monitoring, Birth Defects: Original Article Series. Volume XV-No. 4, New York: A.R. Liss, Inc.; 1979.
47. Parker D, Delpy D, Reynolds EOR. Single electrochemical sensor for transcutaneous measurement of pO_2 and pCO_2. In: Huch R, Huch A, Lucey JF, editors. Continuous Transcutaneous Bloo Gas Monitoring, in Birth Defects: Original Article Series. Volume XV-No. 4, New York: A. R. Liss; 1979.
48. Severinghaus JW. A combined transcutaneous pO_2—pCO_2 electrode with electrochemical HCO_3- stabilization. J Appl Physiol 1981;51:1027–1032.
49. Severinghaus JW. Transcutaneous monitoring of arterial pCO_2. Resp Monit Int Care 1982; 85–91.

50. Gothgen I. Heat-indued changes in pO_2 and pCO_2 of blood. Acta Anaesthesiol Scand 1984;28:447–451.
51. Jacobsen E, Gothgen I. Relationship between arterial and heated skin surface carbon dioxide tension in adults. Acta Anaesthesiol Scand 1985;29:198–202.
52. Hazinski TA, Severinghaus JW. Transcutaneous analysis of arterial pCO_2. Med Instrum 1982;16:150–153.
53. Wimberley PD, Pedersen KG, Thode J Fogh-Andersen, Sorensen AM, Siggaard-Andersen O. Transcutaneou and capillary pCO_2 and and pO_2 measurements in healthy adults. Clin Chem 1983;29:1471–1473.
54. Tremper KK, Mentelos RA, Shoemaker WC. Effect of hypercarbia and shock on transcutaneous carbon dioxide at different electrode temperatures. Crit Care Med 1980;8:608–612.
55. Phan CQ, Tremper KK, Lee SE, Barker SJ. Noninvasive monitoring of carbon dioxide: A comparison of the partial pressure of transcutaneous and end-tidal carbon dioxide with the partial pressure of arterial carbon dioxide. J Clin Monit 1987;3:149–154.
56. Hoppenbrouwers T, Hodgman JE, Arakawa K, Durand M, Cabal LA. Transcutaneous oxygen and carbon dioxide during the first half year of life in premature and normal term infants. Pediatr Res 1992;31:73–79.
57. Huch R. Review: Perinatal monitoring. Acta Anaesthesiol Scand Suppl 1995;S107:91–94.
58. Huch A. Transcutaneous blood gas monitoring. Acta Anesthesiol Scand Suppl 1995;107:87–90.
59. Flynn JT, et al., A cohort study of transcutaneous oxygen tension and the incidence and severity of retinopathy of prematurity [see comments]. New Engl J Med 1992;326: 1050–1054.
60. Schmidt S, Kakatschikaschwili T, Langner K, Dudenhausen JW, Saling E. [Circulatory adaptation of the newborn infant immediately post partum by biolocal measurement of transcutaneous pCO_2]. Z Geburtshife Perinatol 1984;188: 21–23.
61. Binder N, Atherton H, Thorkelsson T, Hoath SB. Measurement of transcutaneous carbon dioxide in low brithweight infants during the first two weeks of life. Am J Perinatol 1994;11:237–241.
62. Huch A, Huch R, Schneider H. Fetal transcutaneous pO_2—current knowledge. In: Huch R, Huch A, Lucey JF, editors. Continuous Transcutaneous Blood Gas Monitoring, Birth Defects: Original Article Series. Volume XV-No. 4, New York: A.R.Liss; 1979.
63. Huch R, Huch A. Fetal and maternal $PtcO_2$ monitoring. Crit Care Med 1981;9:694–697.
64. Lofgren O. Continuous transcutaneous carbon dioxide monitoring in the fetus during labor. Crit Care Med 1981;9:750–751.
65. Okane M, Shigemitsu S, Inaba J, Koresawa M, Kubo T, Iwasaki H. Non-invasive continuous fetal transcutaneous pO_2 and pCO_2 monitoring during labor. J Perinat Med 1989;17:399–410.
66. Antoine C, Young BK, Silverman F. Simltaneous measurement of fetal tissue pH and transcutaneous pO_2 during labor. Eur J Obstet Gynecol Reprod Biol 1984;17:69–76.
67. Schmidt S, Langner K, Dudenhausen JW, Saling E. Reliability of transcutaneous measurement of oxygen and carbon dioxide partial pressure with a combined pO_2–pCO_2 electrochemical sensor in the fetus during labor. J Perinat Med 1985;13:127–133.
68. Bergmans MG, van Geijn HP, Weber T, Nickelsen C, Schmidt S, van den Berg PP. Fetal transcutaneous pCO_2 measurements during labour. Eur J Obstet Gynecol Reprod Biol 1993;51:1–7.
69. Kaneoka T, Kobayashi H, Uchida K, Shirakawa K. [Continuous fetal biochemical monitoring and cardiotocography]. Nippon Sanka Fujinka Gakkai Zasshi 1988;40:721–728.
70. Bartnicki J, Langner K, Harnack H, Meyenburg M. The influence of oxygen administration to the mother during labor on the fetal trasncutaneously measured carbon-dioxide partial pressure. J Perinat Med 1990;18:397–402.
71. Aarnoudse JG, Oeseburg B, Kwant G, Zwart A, Zijlstra WG, Huisjes HJ. Influence of variations in pH and pCO_2 on scalp tissue oxygen tension and carotid arterial oxygen tension in the fetal lamb. Biol Neonate 1981;40:252–263.
72. Smits TM, Aarnoudse JG, Zijlstra WG. Fetal scalp blood flow as recorded by laser Doppler flowmetry and transcutaneous pO_2 during labour. Early Hum Dev 1989;20:109–124.
73. Jensen A, Kunzel W, Kastendieck E. Fetal sympathetic activity, transcutaneous pO_2, and skin blood flow during repeated asphyxia in sheep. J Dev Physiol 1987;9:337–346.
74. Paulick R, Kastendieck E, Wernze H. Catecholamines in arterial and venous umbilical blood: placental extraction, correlation with fetal hypoxia, and transcutaneous partial oxygen tension. J Perinat Med 1985;13:31–42.
75. Braems G, Kunzel W, Lang U. Transcutaneous pCO_2 during labor—a comparison with fetal blood gas analysis and transcutaneous pO_2. Eur J Obstet Gynecol Reprod Biol 1993;52: 81–88.
76. Fukui M, Ohi M, Chin K, Kuno K. The effects of nasal CPAP on transcutaneous pCO_2 during non-REM sleep and REM sleep in patients with obstructive sleep apnea syndrome. Sleep 1993;16:S144–5.
77. Manning DJ, Stothers JK. Sleep state, hypoxia and periodic breathing in the neonate. Acta Paediatr Scand 1991;80:763–769.
78. Morielli A, Desjardins D, Brouillette RT. Transcutaneous and end-tidal carbon dioxide pressures should be measured during pediatric polysomnography. Am Rev Respir Dis 1993;148: 1599–1604.
79. Naifeh KH, Severinghaus JW. Validation of a maskless CO_2-response test for sleep and infant studies. J Appl Physiol 1988;64:391–396.
80. Naughton M, Benard D, Tam A, Rutherford R, Bradley TD. Role of hyperventilation in the pathogenesis of central sleep apneas in patients with congestive heart failure [see comments]. Am Rev Respir Dis 1993;148:330–338.
81. Naughton MT, Benard DC, Rutherford R, Bradley TD. Effect of continuous positive airway pressure on central sleep apnea and nocturnal pCO_2 in heart failure. Am J Respir Crit Care Med 1994;150:1598–1604.
82. Schafer T, Schafer D, Schläfke ME. Breathing, transcutaneous blood gases, and CO_2 response in SIDS siblings and control infants during sleep. J Appl Physiol 1993;74:88–102.
83. Schlaefke ME, Schaefer T, Kronberg H, Ullrich GJ, Hopmeier J. Transcutaneous monitoring as trigger for therapy of hypoxemia during sleep. Adv Exp Med Biol 1987;220:95–100.
84. Milerad J, Hertzberg T, Lagercrantz H. Ventilatory and metabolic responses to acute hypoxia in infants assessed by transcutaneous gas monitoring. J Dev Physiol 1987;9: 57–67.
85. White RA, Nolan L, Harley D, Long J, Klein S, Tremper K, Nelson R, Tabriski J, Shoemaker W. Noninvasive evaluation of peripheral vascular disease using transcutaneous oxygen tension. Am J Surg 1982;144:68–75.
86. Kram HB, Shoemaker WC. Diagnosis of major peripheral arterial trauma by transcutaneous oxygen monitoring. Am J Surg 1984;147:776–780.
87. Padberg FT, Back TL, Thompson PN, Hobson RW. Transcutaneous oxygen ($Tcp O_2$) estimates probability of healing in the ischemic extremity. J Surg Res 1996;60:365–369.
88. Wutschert R, Bounameaux H. Determination of amputation level in ischemic limbs. Reappraisal of the measurement of $Tcp O_2$. Diabetes Care 1997;20:1315–1318.

89. Lemke R, Klaus D, Lübbers DW, Oevermann G. Noninvasive $ptCO_2$ initial slope index and invasive $ptCO_2$ arterial index as diagnostic criterion of the state of peripheral circulation. Crit Care Med 1988;16:353–357.
90. Keller HP, Klaue P, Hockerts T, Lübbers DW. Transcutaneous pO_2 measurement on skin transplants. In: Huch R, Huch A, Lucey JF, editors. Continuous Transcutaneous Blood Gas Monitoring, Birth Defects: Original Article Series. Volume XV-No. 4, New York: A.R.Liss; 1979.
91. Lübbers DW. Transcutaneous measurements of skin O_2 supply and blood gases. Adv Exp Med Biol 1992;316:49–60.
92. Tremper KK, Waxman K, Shoemaker WC. Use of transcutaneous oxygen sensors to titrate PEEP. Ann Surg 1981;193:206–209.
93. Dooley J, Schirmer J, Slade B, Folden B. Use of transcutaneous pressure of oxygen in the evaluation of edematous wounds. Undersea Hyperb Med 1996;23:167–174.
94. Wattel F, Pellerin P, Mathieu D, Patenotre P, Coget JM, Schoofs M, Leps P. [Hyperbaric oxygen therapy in the treatment of wounds, in plastic and reconstructive surgery]. Ann Chir Plast Esthet 1990;35:141–146.
95. Huch A, Huch R, Hollmann G, Hockerts T, Keller HP, Seiler D, Sadzek J, Lübbers DW. Transcutaneous pO_2 of volunteers during hyperbaric oxygenation. Biotelemetry 1977;4: 88–100.
96. Barker SJ, Tremper KK. The effect of carbon monoxide inhalation on pulse oximetry and transcutaneous pO_2 [see comments]. Anesthesiology 1987;66:677–679.
97. Sridhar MK, Carter R, Moran F, Banham SW. Use of a combined oxygen and carbon dioxide transcutaneous electrode in the estimation of gas exchange during exercise. Thorax 1993;48:643–647.
98. Breuer HW, Skyschally A, Alf DF, Schulz R, Heusch G. Transcutaneous pCO_2-monitoring for the evaluation of the anaerobic threshold. Comparison to lactate and ventilatory threshold [see comments]. Int J Sports Med 1993;14:417–421.
99. Sato M, Severinghaus JW, Powell FL, Xu FD, Spellman MJJ. Augmented hypoxic ventilatory response in men at altitude. J Appl Physiol 1992;73:101–107.
100. Alswang M, Friesen RH, Bangert P. Effect of preanesthetic medication on carbon dioxide tension in children with congenital heart disease. J Cardiothorac Vasc Anesthesiol 1994;8: 415–419.
101. Rozenfeld RA, Dishart MK, Tønnessen TI, Schlichtig R. Methods for detecting intestinal ischemic anaerobic metabolic acidosis by local pCO_2. J Appl Physiol 1996;81:1834–1842.
102. Keller HP, Klaue P, Lübbers DW. Transcutaneous pO_2 measurements on rats and rabbits. In: Huch R, Huch A, Lucey JR, editors. Continuous Transcutaneous Blood Gas Monitoring, Birth Defects: Original Article Series. Volume XV-No. 4, New York: A.R.Liss; 1979.
103. Tremper KK, Shoemaker WC. Continuous CPR monitoring with transcutaneous oxygen and carbon dioxide sensors. Crit Care Med 1981;9:417–418.
104. Versmold HT, Linderkamp O, Holzmann M, Strohhacker I, Riegel K. Transcutaneous monitoring of pO_2 in newborn infants: where are the limits? Influence of blood pressure, blood volume, blood flow, viscosity, and acid base state. In: Huch R, Huch A, Lucey JF, editors. Continuous Transcutaneous Blood Gas Monitoring, in Original Article Series. Volume XV-No. 4, New York: A.R. Liss; 1979.
105. Wendling P, Fussinger R, Schmidt HD, Stosseck K. [Validity of the transcutaneous pO_2-measurement during pharmacologically induced changes of skin perfusion (author's transl)]. Anaesthesist 1982;31:135–138.
106. Ewald U, Huch A, Huch R, Rooth G. Skin reactive hyperemia recorded by a combined $TcpO_2$ and laser Doppler sensor. Adv Exp Med Biol 1987;220:231–234.
107. Palmisano BW, Severinghaus JW. Transcutaneous pCO_2 and pO_2: a multicenter study of accuracy. J Clin Monit 1990;6: 189–195.
108. Tremper KK, Shoemaker WC. Transcutaneous oxygen monitoring of critically ill adults, with and without low flow shock. Crit Care Med 1981;9:706–709.
109. Fallenstein F, Ringer P, Huch R, Huch A. A new system for $tcpO_2$ long-term monitoring using a two-electrode sensor with alternating heating. Adv Exp Med Biol 1987;220: 285–289.
110. Paky F, Koeck CM. Pulse oximetry in ventilated preterm newborns: reliability of detection of hyperoxaemia and hypoxaemia, and feasibility of alarm settings. Acta Paediatr 1995;84:613–616.
111. Baeckert P, Bucher HU, Fallenstein F, Fanconi S, Huch R, Duc G. Is pulse oximetry reliable in detecting hyperoxemia in the neonate?, Adv Exp Med Biol 1987;220:165–169.
112. Bragiroli A, Sacco C, Carone M, Donner CF. Pulse oximeter and transcutaneous O_2 monitoring: criteria for a choice. Eur Respir J Suppl 1990;11:515s–517s.
113. Fallenstein F, Baeckert P, Huch R. Comparison of in-vivo response times between pulse oximetry and transcutaneous pO_2 monitoring. Adv Exp Med Biol 1987;220:191–194.
114. Wimberley PD, Helledie NR, Friis-Hansen B, Fogh-Andersen N, Olesen H. Pulse oximetry versus transcutaneous pO_2 in sick newborn infants. Scand J Clin Lab Invest Suppl 1987;188:19–25.
115. Wimberley PD. Oxygen monitoring in the newborn. Scand J Clin Lab Invest Suppl 1993;214:127–130.
116. Poets CF, Southall DP. Noninvasive monitoring of oxygenation in infants and children: practical considerations and areas of concern [see comments]. Pediatrics 1994;93:737–746.
117. Mike V, Krauss AN, Ross GS. Doctors and the health industry: a case study of transcutaneous oxygen monitoring in neonatal intensive care. Soc Sci Med 1996;42:1247–1258.
118. American Academy of Pediatrics Committee on Drugs: Guidelines for monitoring and management of pediatric patients during and after sedation for diagnostic and therapeutic procedures. Pediatrics 1992;89:1110–1115.
119. Wimberley PD, Burnett RW, Covington AK, Maas AHJ, Mueller-Plathe O, Siggaard-Andersen O, Weisberg HF, Zijlstra WG. Guidelines for transcutaneous pO_2 and pCO_2 measurement. IFCC document. Ann Biol Clin 1990;48:39–43.

See also BLOOD GAS MEASUREMENTS; CARDIOPULMONARY RESUSCITATION; RESPIRATORY MECHANICS AND GAS EXCHANGE.

COBALT-60 UNITS FOR RADIOTHERAPY

JOHN R. CUNNINGHAM
Camrose, Alberta, Canada

INTRODUCTION

Cobalt is a metal, between iron and nickel, in the periodic table. It resembles them and occurs fairly commonly in iron and nickel ores, such as those found near Sudbury, Ontario, Canada. Cobalt as a substance has been known since about the mid-1700s. It was discovered in 1735 by a Swedish chemist named Brandt and was named after Kobald, a goblin from Germanic legends, known for stealing silver. Its salts were used in ancient days for making pigments, which produced brilliant blue colors in pottery.

The ancient Egyptians used it in painting murals in tombs and temples. It is necessary, in trace amounts, for proper nutritional balance.

The nucleus of ^{59}Co, which is the only isotope of cobalt found in Nature, has 27 protons and 32 neutrons. It happens to have an unusually large neutron capture cross-section, which means that bombardment with neutrons turns many of its atoms into ^{60}Co, which is very highly radioactive. ^{60}Co has a relatively long half-life (5.26 years) and it decays to ^{60}Ni by the emission of a beta particle (an electron). ^{60}Ni is also radioactive and emits two energetic gamma rays with energies 1.17 and 1.33 MeV. Million electron volts = MeV. An electron volt is the amount of energy an electron has when it is accelerated through a voltage of 1 MV. It is very small: $1\ \text{MeV} = 1.602 \times 10^{-13}$ J. These gammma-rays are produced in almost equal number and the pair of them can be approximated by their average 1.25 MeV, to form radiation that has high penetration in matter.

Cobalt has an atomic weight of 58.933 atomic mass units (amu), a mass density of 8900 kg/m^3, and melts at $\sim 1500\ ^\circ$C. All of these properties combine to make it unique as a practical source of radiation for cancer treatment, industrial radiography, sterilization of food, and other purposes requiring intense but physically small sources of radiation.

It was not isolated as a metal until early in the eighteenth century and was not used for its metallic properties until the twentieth century. Its most important modern use is in the production of alloys of steel that are very hard and very resistant to high temperatures. These alloys find their uses in cutting tools and such diverse products as jet engines and kitchen cutlery.

HISTORY

Sampson et al. (1) noted the interesting radioactive properties of ^{60}Co at least as early as 1936. Livingood and Seaborg (2) described its properties in 1941. W.V. Mayneord, of the Royal Cancer Hospital in London (later the Royal Marsden Hospital), and A. J. Cipriani, then Head of the Biology Division at Chalk River, Ontario, Canada, described its production by neutron bombardment of ^{59}Co in a nuclear reactor in 1947 (3).

In June of 1949, H.E. Johns, then professor of physics at the University of Saskatchewan, Canada, and physicist to the Saskatchewan Cancer Commission, visited the NRX nuclear reactor at Chalk River, Ontario, to discuss, with Cipriani and others, the possibilities of irradiating a sample of cobalt in order to produce a ^{60}Co source. The theoretical advantages of using the energetic gamma rays of ^{60}Co to destroy cancer cells had been known for some time, but practical problems of source production centered on the availability of a reactor with a sufficiently high neutron flux combined with a facility to handle and prepare the resulting highly radioactive source. Earlier, in 1945, J.S. Mitchell of Cambridge and J.V. Dunworth of Chalk River had discussed the possibilities of producing ^{60}Co using the high neutron flux expected to be available from NRX, a nuclear reactor being built at the Chalk River site. At that time NRX was not yet operating, but in 1949, when Johns visited it, it was. The NRX is a heavy water reactor and at that time had the highest available neutron flux in the world ($\sim 3 \times 10^{13}$ neutrons/cm^2/s). The reactor was heavily involved in a program of radioisotope production and the irradiation of cobalt was taken to be part of this program.

Arrangements were made to irradiate three samples of cobalt and they were placed in the reactor in the fall of 1949. They were removed ~ 1.5 years later. The first source was destined for a cobalt unit being designed and built by Dr. Johns and his students in Saskatoon (4). It was delivered there in July of 1951 and on the 18th of August it was installed in the cobalt unit that had been prepared for it. The second source was sent to the Victoria Hospital in London, Ontario, where it was installed on the 23rd of October 1951 in a unit that had been designed and built by Eldorado Mining and Refining Company (later Atomic Energy of Canada Ltd.). Dr. Ivan Smith treated the first patient in London on 27th of October 1951, just 4 days after the installation of the source. The first patient treated on the Saskatoon unit, by Dr. T.A. Watson, was on the 8th of November 1951. Some mystery surrounds the details of the third source. There is some evidence that it was originally intended to go to Mayneord in England, but that in 1951 it was considered that postwar reconstruction was not yet sufficiently advanced there so it was diverted to the M.D. Anderson Hospital in Houston, Texas. It was to be installed in a unit designed by L.G. Grimmett, who had recently been hired by Dr. Gilbert Fletcher largely for this task (5). Part of the mystery concerns the fact that it was delayed in its irradiation and was actually removed from the reactor for a time and later replaced. Some have suggested that this may have been related to the outbreak of the Korean War and the general sensitivity concerning nuclear matters. Whatever the reason, it was not actually shipped until July of 1952, almost a full year later than the other two sources. The M.D. Anderson unit was then at Oak Ridge Tennessee for experimental purposes and was transferred, with its source, to the M.D. Anderson Hospital in Houston in September of 1953. The first patient was treated, in Houston, on the 22nd of February 1954. Pictures of these three cobalt units are given in Fig. 1. Roger F. Robinson has told an informative and interesting history, which includes many details about the original sources, as well as stories about a number of the people involved(6).

Each of these three sources was used in cobalt units for the treatment of cancer for many years. The two Canadian units became prototypes for units that were subsequently sold commercially. The unit in London, built by Atomic Energy of Canada Ltd., was the first of a long series of machines manufactured by them. The first series was known as the “Eldorado” series. A later series of units went under the name “Theratron”. The descendant of that company: MDS Nordion, is still building and selling cobalt units. The Saskatoon unit, designed by H. E. Johns and several of his students at the University of Saskatchewan, was made by John MacKay of the Acme Machine and Electric Co. Ltd. in Saskatoon (7) and later commercially by Picker X-ray of Cleveland, Ohio. Each of these units is pictured in Fig. 1 near the times of their source installations.

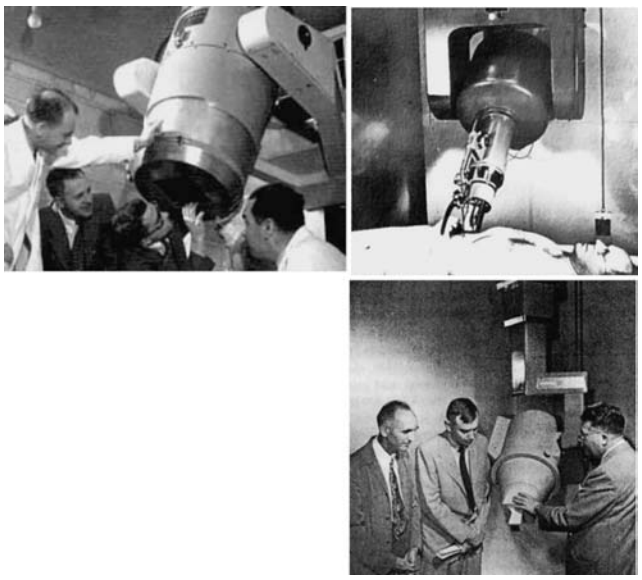


Figure 1. The worlds first three cobalt units. Clockwise from above London, Ont., Canada, Saskatoon, Sask., Canada and Houston, Texas.

Before 1951, radiation therapy had been carried out almost exclusively by X-ray machines operating at tube voltages of 400,000 V or less. Such machines produce X-ray beams having a broad spectrum of X-ray energies with an average of one-third or less of the maximum. Thus, a 400 kV machine would correspond to a single energy of ~ 133 keV. Cobalt-60, with its average photon energy of 1.25 MeV, is the equivalent of an X-ray machine operating about six times the old value. As will be seen later, cobalt units are also mechanically and electrically simple devices and, following their introduction, rapidly became the standard machine for treating nearly all cancers other than that of the skin. Cobalt units have now been almost completely replaced by linear accelerators, which produce X rays having still greater penetration and higher outputs allowing shorter treatment times.

THE PHYSICS OF ACTIVATION: EXPOSURE AND DOSE

Only the physics directly related to the description of ^{60}Co sources and units will be discussed here. More detailed information can be found in standard textbooks such as those of Attix (8), Greening (9), and Johns and Cunningham (10).

Almost any material placed within the neutron radiation field of a nuclear reactor will become radioactive. The probability of this happening is determined by the cross-section of the material for capturing a neutron. The cross-section is the equivalent of a probability, although it is usually expressed as an area. Many atoms have neutron capture cross-sections, of the order of 10^{-24} cm^2 around 1935, Enrico Fermi, then in Rome, was measuring these cross-sections. When he found one of about this size he exclaimed, "That's as big as a barn!" 1 barn $= 10^{-24}$ cm^2 is the common measure of nuclear cross-section and its use is permitted by the International System (SI) of units, and if a neutron passes through this area it is "captured" by the nucleus to form a new nuclear species, which usually is radioactive.

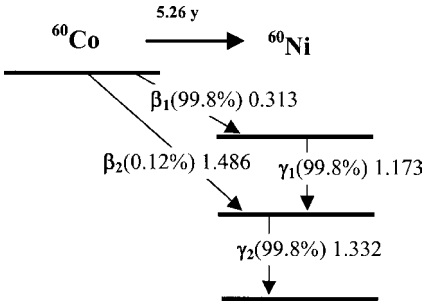


Figure 2. The decay schemes of ^{60}Co and ^{60}Ni, showing the beta particle energies of ^{60}Co and the gamma-ray energies from ^{60}Ni.

The interaction of the neutron with a nucleus is quite complex, and a number of different products may be formed. The nucleus may capture the neutron to produce a new species that is stable, or the neutron may be re-emitted at the same or a different energy. In the latter case, we refer to the process as neutron scattering. The production of ^{60}Co is an example of neutron capture. A nucleus of ^{59}Co absorbs a neutron and forms ^{60}Co, which is radioactive and decays with a half-life of 5.26 years by the emission of an electron that turns it into an isotope of nickel, ^{60}Ni. The decay scheme of ^{60}Co and ^{60}Ni is shown in Fig. 2. The two gamma rays mentioned earlier are actually emitted by the Nickel nucleus ^{60}Ni. Some properties of cobalt and its radiation are given in Table 1.

The ^{60}Co activity produced is determined by the neutron flux density in the reactor, the neutron capture cross-section, the amount of ^{59}Co inserted into the reactor, and the length of time it is left there. The rate of production of radioactive atoms can be expressed as

$$\frac{N}{t} = N\,\sigma\,\phi \tag{1}$$

where N is the number of ^{59}Co atoms placed in the reactor, σ is the neutron capture cross-section per atom, ϕ is the flux density of neutrons, and Δt is a time interval. The

Table 1. Properties of Cobalt and Its Radiation

Property	Value
Cobalt-59	
Atomic number	$Z = 27$
Atomic weight	$A = 58.933$ amu
Mass density	$\rho = 8900$ kg/m^3
Melting point	1500 K
Neutron capture cross-section	$\sigma = 37\times10^{-24}$ cm^2
Cobalt-60	
Half-life	$T_{1/2} = 5.26$ years
Bata energies	0.313 MeV (99.8%)
	1.486 MeV (0.12%)
Nickel-60	
Photon energies	$\gamma_1 = 1.733$ MeV
	$\gamma_2 = 1.332$ MeV
Interaction coefficient in water	$(\mu/\rho) = 0.0698$ cm^2/g
Average Energy Absorbed in water	$E_{ab} = 0.456$ MeV
Half-value layer in Pb	$X_{1/2} = 11$ mm

parameter ΔN will be the number of activations that take place in this time interval.

As an illustrative numerical example, consider a sample of 15 g of ^{59}Co to be located in a nuclear reactor at a point where the neutron flux density is 10^{14} cm^{-2}/s. This represents a source that is 1.5 cm in diameter and $\sim$1 cm high and is fairly representative of sources and neutron fluxes that have been used. The original two Canadian sources were 2.54 cm in diameter and composed of $\sim$26 disks each 0.5 mm thick. The American source was square in cross-section. From Eq. 1, and with the use of some of the information given in Table 1, we calculate the number of atoms of ^{59}Co that are converted to ^{60}Co during a period of time Δt. We also require a value for Avogadro's Number N_A, so that we can calculate the number of atoms (at) of ^{59}Co in 1 g of the substance.

$$N_A = 6.023 \times 10^{23}\ \text{atoms/mol}$$

The number of ^{59}Co atoms in our 15 g sample is

$$N_{59_{Co}} = 15\,\text{g} \times \frac{6.023 \times 10^{23}\text{at}}{\text{mol}} \times \frac{1\,\text{mol}}{58.933\,\text{g}} = 1.533 \times 10^{23}\,\text{at}$$

From Table 1, we see that the cross-section for neutron capture in ^{59}Co is 37×10^{-24} cm^2/atom. If the 15 g of cobalt were left in the reactor at this location for a period of 1 h, the number of atoms (at) converted to ^{60}Co, following Eq. 1, would be

$$N = 1.533 \times 10^{23} \times \frac{37 \times 10^{-24}\text{cm}^2}{\text{at}} \times \frac{10^{14}\,\text{cm}^{-2}}{\text{s}} \times \frac{3600\,\text{s}}{\text{h}} = 2.042 \times 10^{18}\,\text{at}$$

Although this appears to be a very large number of atoms it represents only $\sim$0.2 mg of ^{60}Co. It does, however, represent a considerable amount of radioactivity and would be easy to measure.

The most fundamental parameter for the specification of the strength of a radioactive source is activity. Activity is defined as the number of decay processes that occur per second and its special unit is the bequerel (Bq), which is defined to be an average of one nuclear disintegration each second. Activity is easy to describe theoretically, but is very difficult to determine experimentally. It can be inferred from the number of atoms of the substance and the value of its half-life, which for ^{60}Co is given in Table 1 as 5.26 years.

Activity can be calculated from the simple relation

$$A = N\lambda \tag{2}$$

where λ is a constant of proportionality known as the transformation constant. It is related to the half-life $T_{1/2}$, of the radioactivity by

$$\lambda = \frac{0.693}{T_{1/2}} \tag{3}$$

where the number 0.693 is the natural logarithm of 2.

For example, the activity of ^{60}Co that would result from the above irradiation of 15 g of ^{59}Co would be

$$A = 2.04 \times 10^{18} \times \frac{0.693}{5.26\,\text{year} \times 3.1557 \times 10^6\,\text{s/year}} = 0.0852 \times 10^{12}\,\text{s}^{-1} = 85.2 \times 10^9/\text{s} = 85.2\,\text{GBq} \tag{4}$$

where the half-life $T_{1/2}$ has been expressed in seconds. The activity that is actually produced in a reactor irradiation is considerably less than this theoretical amount. This is largely due to attenuation of the neutron flux by the considerable mass of the cobalt.

The more traditional unit of activity has been the curie (Ci), which corresponds to 3.7×10^{10} nuclear decays/s. The activity of the above source, stated in curies would be

$$A = \frac{85.2 \times 10^9}{\text{s}} \times \frac{1\,\text{Ci}}{3.7 \times 10^{10}/\text{s}} = 2.30\,\text{Ci}$$

The specification of a commercial source of radiation in terms of activity is not very practical because activity does not uniquely relate to the radiation output when an individual source is loaded into a treatment unit. The output will depend on the physical size and configuration of the source and the design of the collimator of the treatment unit.

This problem was solved by the use of a quantity called exposure. Exposure is defined in terms of the amount of ionization that is produced in air by the radiation. The special unit is the roentgen (R). One roentgen corresponds to the release of 2.58×10^{-4} C/kg of air.

For gamma-ray emitters, such as this one, a quantity known as the exposure rate constant (Γ), has been defined that relates the activity in curies to the exposure rate in roentgen/hour at a point in air 1 m from the source. It is calculated from the gamma-ray spectrum using the interaction coefficients of air (the required data are given in Table 1). For a ^{60}Co source, Γ, is

$$\Gamma = 1.29\,\text{R} \cdot \text{m}^2/\text{h} \cdot \text{Ci}^{-1}$$

This allows calculation of the parameter that is frequently used to specify source strength: the "roentgens per hour at a meter" (Rmm). For our 2.30 Ci source it is

$$\text{Rmm} = 2.30\,\text{Ci} \times \frac{1.29\,\text{R} \cdot \text{m}^2}{\text{h} \cdot \text{Ci}} \frac{1}{1\,\text{m}^2} = 2.97\,\text{R} \cdot \text{h}$$

A much more practical quantity, from the point of view of radiotherapy, is the absorbed dose rate produced at some agreed distance. To explain this, it will be useful to first define absorbed dose and to go through some approximate calculations connecting activity and absorbed dose rate.

Absorbed dose is the physical quantity that most closely correlates with the biological effect of the radiation and it is defined (11) as the amount of energy absorbed per unit mass of an irradiated material. The special unit of absorbed dose is the gray (Gy), which is defined as 1 joule (J) of energy imparted to 1 kg of matter.

A ^{60}Co activity of 85.2×10^9 Bq, as derived above, would give rise to the following photon fluence rate at a distance of 1 m.

$$\psi = 2\,\frac{A}{4}\,\frac{1}{100^2} = \frac{85.2 \times 10^9\,\text{Bq}}{2\ 10^4\,\text{cm}^2} = 1.356 \times 10^6\,\text{cm}^{-2}/\text{s} \tag{5}$$

The rate of photon interactions with a mass M of the water is given by

$$N' = \sum_i \psi_i \left(\frac{\mu}{\rho}\right)_i M \tag{6}$$

where ψ_i is the fluence (number crossing an area equal to 1 cm^2) of each of the photon energies, $(\mu/\rho)_i$ is the mass interaction coefficient for each of them. The parameter(μ/ρ) expresses the cross-section, or probability of interaction of photons with 1 g of material and M is the mass of the material in grams. The summation in Eq. 6 is over the two components of the photon spectrum as depicted in Fig. 2.

Since the photon energies are so close together, we can use the average value of the interaction coefficients, which is given in Table 1 as 0.0698 cm^2/g. The rate of photon interactions, calculated from Eq. 6, would then be

$$N' = \frac{1.356 \times 10^6}{\text{cm}^2\text{s}} \times 0.0698\ \frac{\text{cm}^2}{\text{g}} \times 1\,\text{g} = 94.6 \times 10^3/\text{s} \qquad (7)$$

Each photon that interacts imparts an average of 0.456 MeV (Table 1) of energy so the rate of energy absorbed E', from this irradiation would be

$$E' = \frac{94.6 \times 10^3}{\text{s}} \times 0.456\,\text{MeV} = 43.1 \times 10^3\ \frac{\text{MeV}}{\text{s}} \qquad (8)$$

This is a very tiny amount of energy. It was deposited in 1 g of water. Its value can be converted to a more familiar energy unit by using the relation $1\,\text{MeV} = 1.6022 \times 10^{-13}\,\text{J}$. The absorbed dose rate from these photons would then be

$$D' = 43.1 \times 10^3\ \frac{\text{MeV}}{\text{g s}} \times \frac{1.6022 \times 10^{-13}\text{J}}{1\,\text{MeV}} \times \frac{10^3\,\text{g}}{\text{kg}} \qquad (9)$$

$$= 69.1 \times 10^{-7}\ \frac{\text{J}}{\text{kg s}} = 6.91 \times 10^{-6}\,\text{Gy/s}$$

$$D = 6.91 \times 10^{-6}\ \frac{\text{Gy}}{\text{s}} \times 3600\ \frac{\text{s}}{\text{h}} = 0.025\,\text{Gy} \qquad (10)$$

A simple radiation treatment for cancer typically involves an absorbed dose at the tumor of 2.0 Gy (in the old units; 200 rad), and because of attenuation in the tissues, and various other factors, this implies, for say a 2 min treatment, an activity almost 5000 times stronger than in our example source. The distance from the source to the tumor has typically been 80 cm. This would call for a source activity of $\sim 25 \times 10^{13}$ Bq or 250 TBq or $\sim$7500 Ci.

To attain this, the cobalt must be left in the reactor for a much longer time than in our example above. With a longer activation, one must note that while ^{60}Co is being formed it is also decaying. The resulting activity would be the sum of that which is being produced, as described by Eq. 1, and the amount that decays. This can be written as

$$\frac{dN}{dt} = N_0\,\sigma\,\phi - \lambda N \qquad (11)$$

where N_0 is the initial number of ^{59}Co atoms present and λ is the transformation constant (see Eq. 2) for the ^{6O}Co decay. The other symbols have the same meaning as for Eq. 1. The solution to this equation, expressed in terms of activity, is

$$A(t) = A_{\max}(1 - e^{-\lambda t}) \qquad (12)$$

where $A_{\max} = N_0\,\sigma\,\phi$ is the maximum activity attainable for an infinitely long irradiation. For the neutron irradiation conditions of our example, the maximum activity attainable is

$$A_{\max} = 1.533 \times 10^{23}\,\text{at} \times 37 \times 10^{-24}\ \frac{\text{cm}^2}{\text{at}} \times \frac{10^{14}}{\text{cm}^2 \cdot \text{s}} \qquad (13)$$

$$= 56.72 \times 10^{13}/\text{s} = 567.2\,\text{TBq} = 15,000\,\text{Ci}$$

It would require 5 years in the reactor to produce a source half this strong, that is, 7500 Ci.

This is not strong enough for modern treatment requirements and a higher neutron flux is required. As time has passed, reactor fluxes have increased considerably, and this has allowed both the irradiation times to be shortened and the sources to be made smaller.

There are a number of advantages to making cobalt sources as small as possible. One of these has to do with the sharpness of the edges of the radiation beam. This is known as penumbra and will be discussed later under that topic. It will be seen that a small diameter source is desirable. Another reason for a small source has to do with the amount of self-absorption and photon scattering that will take place within it. The source that we have been considering was a cylinder 1.0 cm in height, and for the gamma rays of cobalt this is almost a half value layer even in lead (Table 1), let alone in cobalt. It must be expected that the radiation emitted by such a source would be accompanied by considerable attenuation and would include an appreciable component of scattered photons. Because of the attenuation and scatter that takes place in the source, the dose rate is greatly overestimated in the calculations made above. The larger the source physically, the greater the activity required to give a desired dose rate.

SPECIFICATION OF SOURCE STRENGTH

In actual practice, the strength of the source is stated in terms of exposure rate at 1 m (Rmm). This is a measured quantity and is determined by the vendor of the source. Sources delivering up to 250 R/min at a meter are now available.

One way of judging the "efficiency" of the neutron irradiation is by stating the specific activity of the source produced. This is the activity, expressed in becquerel (or curie) per gram of cobalt. The specific activity of a 7500 Ci source that weighed 15 g would be 7500 Ci/15 g = 500 Ci/g. In modern reactors, the neutron flux density can be greater than the $10^{14}/\text{cm}^{-2}\cdot\text{s}^{-1}$ that we assumed, sources can be irradiated for longer times than in the example. Specific activities of up to 500 Ci/g have been produced. Cost goes up linearly with irradiation time, but, as can be seen, activity does not, and source strengths actually produced are decided by economic considerations. In actual practice, sources are not irradiated as solid cylinders, as has been assumed for this example, but rather they are made up into a capsule on demand from stocks or pellets that were preirradiated to a selection of specific activities. Pellets are shown in Fig. 3 along with a pair of stainless steel containers into which they will be placed. The pellets will be loaded into the cylinder shown in the center of the picture, then spacers, such as those shown on the right, are inserted to hold the pellets in position, and finally this cylinder, when capped, is inserted into the cylinder shown on the left and cold-welded shut. All

Figure 3. Cobalt pellets and source capsule with components.

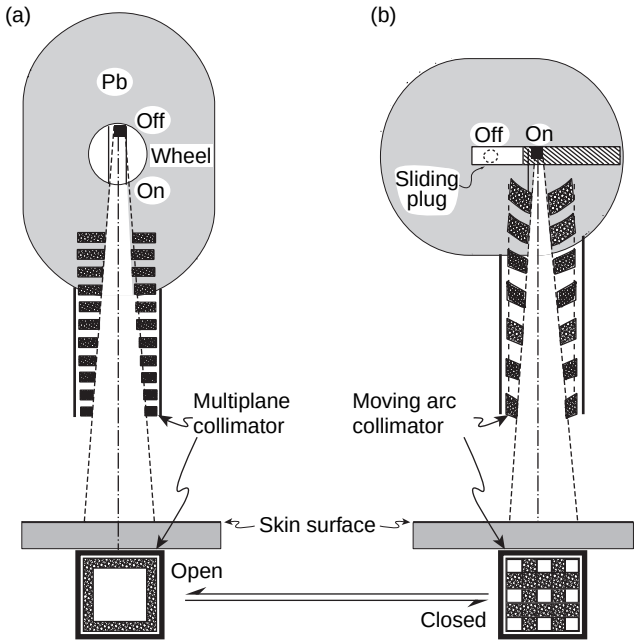


Figure 4. Two designs for cobalt unit heads. (a) A rotating wheel carries the source to the "on" position. A multiplane collimator controls the size of the rectangular beam. (b) A sliding drawer moves the source and a multileaf collimator moves on an arc to control the beam.

of these operations are carried out remotely in a hot cell. Finally, the source is shipped in a well-protected and shielded container to be loaded into a cobalt unit.

COBALT UNIT DESIGN

The first cobalt units went into operation in 1951. Very soon after that they became available commercially, and the production of cobalt sources and cobalt units expanded to such an extent that, for 30 years, more radiotherapy was carried out with ^{60}Co than with all other types of radiation combined. Cobalt machines have the tremendous advantages of producing a completely predictable, steady, reliable beam of relatively high energy radiation, being mechanically simple, rarely needing repair, and being easy to repair when required.

HEAD DESIGN

In all cobalt units, the source is placed near the center of a large, lead-filled steel container. A device is provided for moving the source from a position where it is "Off", because it is shielded in all directions, to a position opposite an opening through which the useful beam may emerge. A number of mechanisms have been devised for moving the source, and two of them are shown in Fig. 4. In Fig. 4a, the source is mounted in a heavy metal (mostly tungsten) wheel that may be rotated through 180° to carry it from the Off position to the On position. In Fig. 4b, the source is mounted in a sliding plug or drawer that carries it from the Off to the On position. In one of the first cobalt units (the Eldorado A), the source did not move at all. The beam opening was filled with a tank of mercury that was pumped out of the way by air pressure to turn the machine On and then the mercury returned by gravity to turn the beam Off. The sliding drawer mechanism shown in Fig. 4b has tended to be the more commonly used.

All machines must be arranged so that they fail "safe". That is, the source must be held in the On position by the continuous application of a force so that if the power fails, it must return quickly to the Off position. For both a and b in Fig. 4, this is provided by a strong spring. The lead-filled container, or "head" of the unit, must be of the order of 25 cm thick in all directions from the source. The design criteria will depend on the regulations in force where it is to be used, but basically it must be such that the leakage radiation emerging from the shield would not cause an overexposure to anyone staying at its surface for prolonged periods of time. This would imply, for example, a yearly equivalent dose of not >5 mSv (or - 500 mrem) at a distance of 1 m from the source. This exposure level is greater than the average in low natural background areas, but is less than the exposure in many other regions of the world where people live. The sievert (Sv) is the special unit of equivalent dose. One sievert will result in the same biological effect as 1 (Gy) gray of conventional X rays. If we assume a maximum source strength of 10,000 Ci, and again use the exposure rate constant of 1.29 R $m^2/h{\cdot}Ci^{-1}$, and assume that 1 R corresponds to an equivalent dose of 0.01 Sv, this would imply a thickness of 20–30 half-value layers. The half-value layer in lead for cobalt radiation is $\sim$1.1 cm (Table 1), and this calculation would imply a thickness of $\sim$30 cm. In actual practice a much more detailed calculation would be done, augmented by measurement.

This simple calculation can serve as a guide only. The half-value layer for a broad beam of radiation, such as in

this case, would be >1.1 cm. On the other hand, it is unlikely that anyone would remain for a whole year just beside the head of the cobalt unit. In fact, 20–25 cm is about the thickness of the heads of most cobalt units.

Figure 4 also shows two types of collimators. Both consist of sets of bars that can be adjusted to produce a radiation beam with a rectangular cross-section. The diagrams at the bottom of Fig. 4 show an end-on view of the appearance of both collimator bars in the open and the closed positions.

MOUNTS

There are only two basic ways of mounting and "porting" radiation treatment units. One of the two oldest designs is illustrated in Fig. 5 and is an example of the so-called SSD mount. The head of the unit was held in a yoke, which was suspended by a column from a set of rails attached to the ceiling. It could be moved up and down or back and forth and the head could be rotated about the horizontal axis seen. The unit was also equipped with a treatment applicator, which in this case was mounted on the end of the collimator. The motions of the mount allowed the unit to "point" over a wide range of directions and enabled the operator to place the end of the treatment applicator against the skin of the patient at a prescribed location. The floor was left clear to allow easy and full movement of the treatment couch. The distance from the source to the skin of the patient (SSD) was thus a fixed quantity, usually 80 cm, and the focus of the "setup" was the surface of the patient. The size of the beam was defined there, and the reference point for dosimetry was just under the skin.

Figure 5. A Picker cobalt unit at the Ontario Cancer Institute, Toronto in the 1960s–1980s. The unit was mounted on a column suspended from rails on the ceiling leaving the floor clear. A protractor allows the angle to be set carefully using the "SSD" technique. The rack on the wall holds "wedge filters" that shape the beam intensity.

Figure 6. An isocentric mounted cobalt unit of the Theratron series produced by Atomic Energy of Canada Ltd., installed at the Ontario Cancer Institute in Toronto in the 1970s and 1980s.

The alternative mount is the so-called isocentric or fixed source-axis-distance (SAD) mount. An example, dating from the 1970s and 1980s is shown in Fig. 6. The head, encased in a streamlined plastic cover, is mounted on a gantry that can rotate about a horizontal axis. The patient lies on a couch as shown and is raised, lowered, moved sideways, or lengthways so that the tumor is positioned on the intersection of the gantry axis and the collimator axis. This means that for any angle of the gantry, the beam will pass through the tumor. This point is called the isocenter and was a fixed distance from the source, usually 80 cm, in later units 100 cm. The beam is specified by its size at the isocenter. The focus of attention is now at the tumor rather than the surface. In addition, the couch can usually be rotated about a vertical axis, also passing through the isocenter. Virtually all modern treatment units are mounted in the isocentric manner.

The procedures for treatment planning and dosimetry are somewhat different for each of these two types of mount. Treatment planning is discussed in several standard textbooks such as those by Bentel (12), Johns and Cunningham (10), and Kahn (13).

In 1956, an early and innovative symposium was held at Oak Ridge Institute of Nuclear Studies, just before the Eighth International Congress of Radiology, which was held in Mexico City. Problems of source production, machine design and installation, dosimetry, and source specification were discussed. The title of the publication that resulted from this symposium "Roentgens, rads and Riddles", largely reflected the uncertainties of the day in dosimetry. It also includes some history to that time and descriptions of a variety of cobalt units that had been made experimentally and by commercial suppliers.

Cobalt units are inherently simple machines and can be designed and constructed by relatively unsophisticated

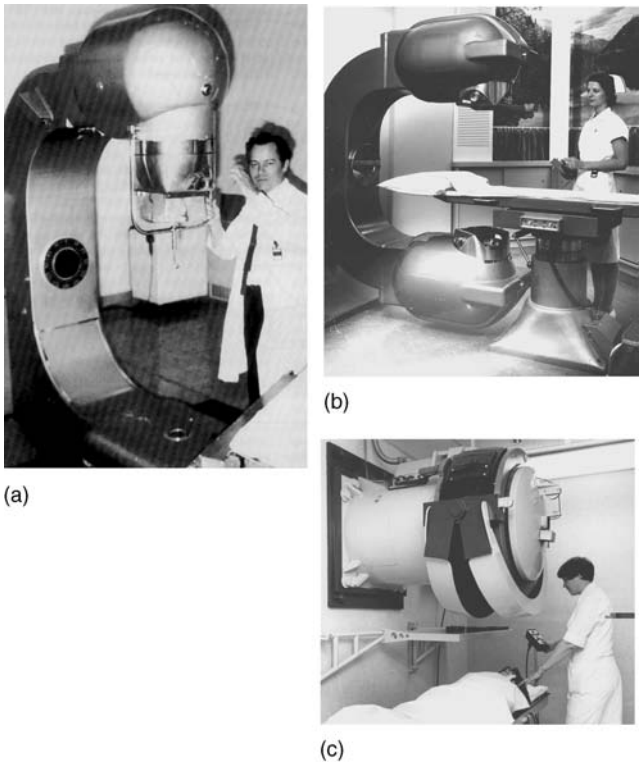


Figure 7. Three experimental cobalt unit designs: (a) a unit with a number of special features, (b) a double-headed unit, and (c) a unit for half-body irradiation.

engineering facilities. This is illustrated by Fig. 7, which shows three quite different units that were designed and built at the Ontario Cancer Institute in Toronto. The unit in (a) was built in 1959 and had a number of special experimental features (14). These included a diagnostic X-ray tube installed in the head of the unit so that good quality placement films could be taken of patients undergoing treatment. This facility is now standard equipment in all modern radiation treatment machines. The films are called "port films". It was isocentrically mounted and was capable of full 360° rotation about the patient. This allowed continuous rotation during treatment or easy set up for the use of several fixed fields from different angles. The latter feature too, is standard on modern machines. The unit also had a large (95 cm) source-to-axis distance, which improved the depth dose characteristics (see the following section). This unit also had an ionization chamber in the counterweight so that the effective thickness of the patient could be determined. This did not prove to be as useful as expected and was not adopted by unit manufacturers.

The unit in Fig. 7b contained two sources and was called the Double-Header (15). The sources were arranged to be very nearly equal in strength and the beams were directed opposite to each other. This provided an automatic "parallel pair" of beams, which forms a component of many multiple field treatments. The real reason for the two sources, however, was to extend their useful life. The Ontario Cancer Institute had, at different times, as many as eight other cobalt units and two of the sources, after each had been used for ~5 years (approximately one half-life) in one or another of them were transferred to the Double-Header for another 5 years of use.

The third cobalt unit depicted in Fig. 7c, was especially designed for "half-body" treatments. It was equipped with a special collimator to provide radiation fields up to 150 cm long and 50 cm wide. It was fitted with a compensating filter so that a uniform dose distribution could be achieved (16).

CHARACTERISTICS OF THE RADIATION BEAM

The decay scheme for ^{60}Co is shown in Fig. 2. There are two γ rays of photon energies 1.17 and 1.33 MeV, respectively. These energies are very close to each other, so ^{60}Co is almost a monoenergetic emitter with energy 1.25 MeV. The actual beam from a cobalt source also contains lower energy photons, which come from the scattering processes that take place within the source. It is also inevitably contaminated with photons scattered from the mechanism that holds the source in position as well as from the various collimator components that are "in view" of the source. That the beam is not purely that from ^{60}Co is attested to by the fact that the linear attenuation coefficient for 1.25 MeV photons in water is 0.0698 cm^{-1} (Table 1), while the experimentally determined coefficient for a cobalt unit beam in water is closer to 0.063 cm^{-1}. A rather more "realistic" spectrum of the radiation for a cobalt unit has been determined by Rogers et al. (17) by Monte Carlo calculations. The low energy components contribute up to ~15% to the dose received by the patient.

In a patient, the intensity of a radiation beam falls off approximately exponentially. This can be seen from the data plotted in Fig. 8, where percentage depth doses for cobalt-60 radiation, and a few other radiations used in radiotherapy, are shown plotted against depth. Percentage depth dose is the single most important quantity in choosing a radiation for radiotherapy. The radiations shown vary from that produced by 100 kV X rays to 25 MV. The depth at which the percentage depth dose falls to

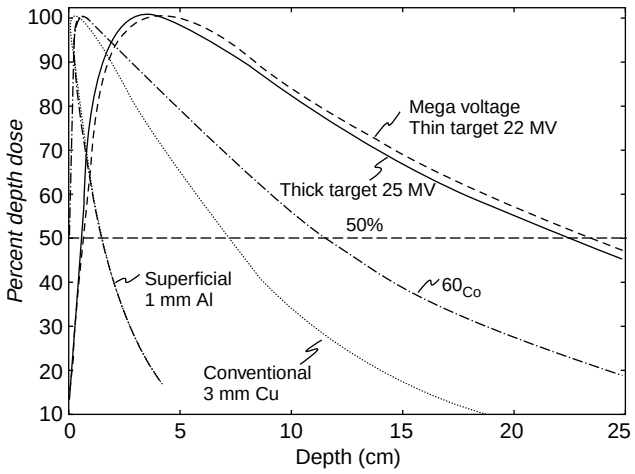


Figure 8. Percentage depth doses plotted against depth for a series of beam energies from superficial (low energy X rays) to megavoltage radiation. All curves are for a 10 × 10 cm field and the depth to 50% dose can be easily determined.

50% can be seen for each radiation by reference to the horizontal dashed line. It varies from < 2 cm for the superficial radiation through ~ 7 cm for "conventional" or 250 kV radiation, ~ 12 cm for ^{60}Co radiation, to > 22 cm for the 26-MV radiation. Cobalt-60 is right in the middle of this range. The graphs in Fig. 8 also show that for the higher energies, the dose at the surface is low and rises as penetration increases. For ^{60}Co radiation, it reaches its maximum at a depth of 0.5 cm and falls off relatively slowly from there. This low dose on the surface, the so-called skin sparing effect, was one of the important properties cobalt radiation had for radiotherapy.

When the cross-sectional area of a radiation beam is small, the dose received at a point below the surface is due almost entirely to primary radiation. As the area of the field is increased, the doses will increase due to an increase in scattered radiation. The greater the depth, the greater the increase, with the result that percentage depth dose increases with field size.

CALIBRATION

Calibration of the output of a cobalt unit is normally done with the use of an ionization chamber that has been calibrated against a standard exposure reference at a standardization laboratory. A calibration factor N_X, is determined by the laboratory and its meaning is that $N_X = X/M$, where X is a known exposure and M is the reading of the electrometer monitoring the ionization produced in the chamber by the radiation.

The traditional and simplest method for calibrating the output of a cobalt unit has been to measure exposure rate in air at a chosen distance and field size, and to derive from this the absorbed dose rate that would occur at the center of a small mass of tissue-like material located at this point. An alternative, but equivalent, method is to determine the dose at a chosen position at a specified depth in a water phantom, again for a specified beam size.

Procedures for calibration, and the mathematical formalism required, to determine absorbed dose from exposure measurements are given in textbooks (9,10), as well as in various dosimetry protocols, both national and international. Examples are those of the American Association of Physicists in Medicine (18) and the International Atomic Energy Agency (19). Since the calibration procedures will only be outlined here, these sources should be consulted for more detailed procedures.

Calibration in Air

A number of physical arrangements for making measurements in a radiation beam are illustrated in Fig. 9. The diagram on the left can be used to refer to calibration in air. An ionization chamber, which has been calibrated in terms of exposure, is placed at point P′, free in air, and a reading, M, is taken for a specified "source-on" time T. This exposure time must be the actual exposure time; that is, it must be exclusive of a time, if any, taken for the source mechanism to move the source from the off to the on position. The reading, M, must also include any adjustment required for atmospheric conditions if the temperature and pressure

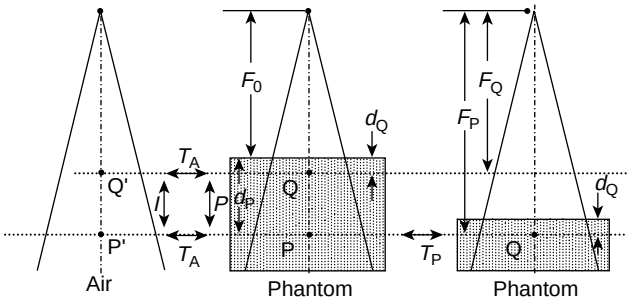


Figure 9. Diagrams showing the meaning of a number of functions used for calibration and dose calculation for treatment planning.

differ from those that pertain to the exposure calibration factor. This would normally be 22°C and 101.3 kPa (equivalent to 1 atm, or 760 mmHg). The parameter M must also be corrected for any small loss of charge that might occur due to charge recombination in the ion chamber during the exposure. Methods for making all of these corrections are discussed in Ref. 9,10,14, and 15. The ion chamber must also have been fitted with a buildup cap, if this is required to make its walls sufficiently thick to provide electronic equilibrium in them. The buildup cap must be made of water-like material. With these precautions, the exposure rate at the point designated in Fig. 9 as P′ would be

$$X = N_X \frac{M}{T} \tag{12}$$

If the cobalt unit is "isocentric" in mount, point Q′ would be on the axis of rotation of the gantry, a distance F_P from the source and the field size would be specified at this point. If the unit were operated in an SSD mode, the calibration point would be the one shown as Q′ in Fig. 9 and would be at a distance F_Q from the source. The absorbed dose rate, free in air, may be calculated from the exposure by the following relationship:

$$\dot{D}_{P'} = N_X \left(\frac{M}{T}\right) \left[0.00876 \frac{\text{J}}{\text{kg R}}\right] \left(\frac{\bar{\mu}_{en}}{\rho}\right)_{\text{air}}^{\text{wat}} k(d_Q) \tag{13}$$

The term in square brackets is derived from the definition of the roentgen, which is the release of a certain electrical charge per kilogram of air, and the average energy required to release 1 C of this charge. (One roentgen is defined as the release of 2.58×10 C/kg of air, and each coulomb released requires an average 33.85 J. Thus, 1 R corresponds to 0.00876 J/kg of air.) The next term is the ratio of mass energy absorption coefficients averaged over the radiation spectrum for water to air, and the final term is a correction factor to account for the fact that in order to characterize a dose rate at a point in air, it must be surrounded by at least enough phantom (water-like) material to produce electronic equilibrium. This material will attenuate and scatter radiations, and $k(d_Q)$, the allowance for this, is estimated to be 0.985.

Although the size of the beam at point P′ is larger than it is at point Q′, the collimator opening is the same for both,

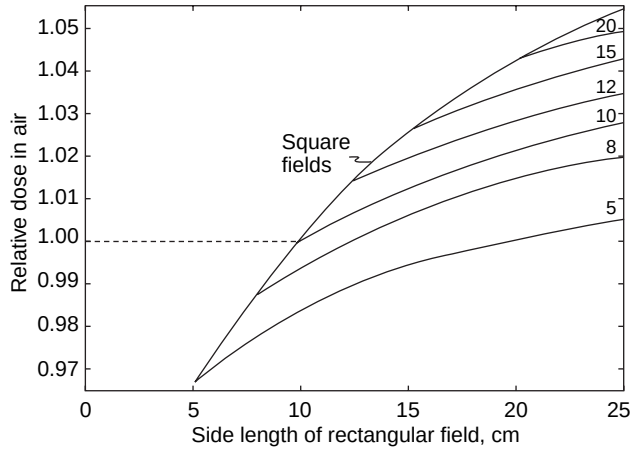


Figure 10. Graphs showing relative output data for a cobalt unit. The output is measured in air and is expressed relative to that of a 10×10-cm field.

and so the source self-absorption and scatter, and collimator scatter, would be expected to be essentially the same. Consequently, the dose rate at P′ should be related to that at Q′ by the inverse square law. For any given cobalt unit this must be tested experimentally, but would be expected to be valid except for distances F, of Fig. 9, that are < 50 cm or so

This is indicated by I in Fig. 9 and by the relation

$$\frac{\dot{D}_{Q'}}{\dot{D}_{P'}} = \frac{F_P^2}{F_Q^2} \tag{14}$$

On the other hand, if the collimator opening is changed, the dose rate at points such as P′ or Q′ will change, due principally to a change in the amount of collimator scatter reaching them. The way this output changes for an example cobalt unit is shown in Fig. 10, where relative dose rates measured on the axis (point P′ of Fig. 9) of an isocentric cobalt unit are plotted against the side length of a rectangular field. The data are normalized to 1.00 for a 10×10 cm field. From this diagram, it can be seen that the dose rates differ by $> 8\%$ from a small, 5×5 cm field to a large 25×25 cm field. The family of curves shown represents rectangular fields, and it can be seen that a rectangular field gives approximately the same relative dose rate, as does a square field of the same area. For example, a 5×20 cm field shows a relative dose rate of almost exactly 1.00, as does the square field, 10×10 cm, of the same area. Curves such as these are specific to a particular collimator design and must be determined as part of the procedure of commissioning a new treatment unit.

Calibration in a Phantom

The right-hand diagram in Fig. 9 shows the arrangement for calibration in a phantom. The procedure is essentially the same as that for calibration in air; Q in this diagram has the same location and field size as does P′. The same precautions must be taken with the ion chamber reading and the same calibration factor, N_X, is used. The dose rate at depth d_Q in a water phantom is given by an expression that is very similar to that in Eq. 13:

$$\dot{D}_{P'} = N_X \left(\frac{M}{T}\right) \left[0.00876 \frac{\text{J}}{\text{kg} \cdot \text{R}}\right] \left(\frac{\bar{\mu}_{en}}{\rho}\right)_{air}^{wat} k(c) \tag{15}$$

$(\bar{\mu}_{en}/\rho)_{wat}^{air}$ is, as before, the ratio of averaged mass-energy absorption coefficients, but in this case they should be averaged over the photon spectrum that is present in the phantom. Values for this ratio are given in Table 2. It is generally assumed to be the same in the phantom as in air, although this cannot be quite correct, as shown by Cunningham et al. (20), Eq. 12. The factor $k(c)$ is very similar to $k(d_Q)$ of Eq. 13, except that c is the radius of the ion chamber as it was configured when the calibration factor was obtained. This factor will be the same whether or not a buildup cap is actually in place in the phantom. The dose rate in a phantom, like that in air, varies with the field size, and a set of data like that shown in Fig. 10 can be compiled. The variation is greater, however, because the beam intensity incident on the phantom changes with collimator opening, as discussed previously, but in

Table 2. Dosimetry Factors for ^{60}Co Radiation[a]

	$(\bar{\mu}_{en}/\rho)_{med}^{wat}$				$(\bar{\mu}_{en}/\rho)_{air}^{med}$			
Spectrum	Graphite	Bakelite	Lucite	Polystyrene	Water	Muscle	Fat	Bone
Ratios of averaged mass energy absorption coefficient for a few materials								
Primary[b]	1.111	1.051	1.029	1.032	1.112	1.103	1.113	1.061
Primary plus scatter[c]	1.116	1.055	1.032	1.037	1.111	1.102	1.107	1.105
Ratios of averaged mass stopping powers								
Primary[b]	1.009	1.071	1.099	1.105	1.129			
Primary plus scatter[c]	1.011	1.073	1.101	1.109	1.131			
Average energy required to cause ionization in air, $W = 33.85$ (dry air) $= 33.7$ (ambient air)								

[a]From Ref. 10, page 230.
[b]Assuming monoenergetic 1.25-MeV photons.
[c]Spectrum derived by Monte Carlo calculation for depth 10 cm in a 20×20 cm beam.

addition, the scatter generated within the phantom changes with a change in irradiated volume.

General Calibrations

Radiation beams of energy lower than that of ^{60}Co are most frequently calibrated in air. Radiation beams higher in energy should always be calibrated in a phantom. Cobalt units, because of their energy and constancy of output, form a natural reference for all radiotherapy calibration procedures.

RELATIVE DOSE FUNCTIONS THAT ARE USED IN TREATMENT PLANNING

Over the years, a set of functions has been defined that make possible accurate point dose calculations as part of treatment planning. These are "tissue air ratio", "percentage depth dose", "backscatter factor", and "tissue phantom ratio". They are also used with radiations other than that from Co-60, but several of them were derived or refined for use with cobalt therapy. They will be discussed briefly. They can all be clarified by reference to Fig. 9.

Tissue Air Ratio (TAR)

Tissue air ratio, first called "tumor air ratio", was introduced by Johns to facilitate the calculation of tumor dose for rotation therapy. This type of treatment uses the isocentric mode of operation in that the tumor is placed on the axis of rotation of the treatment unit and the beam may be pointed toward the tumor from a selection of angles. The tissue air ratio, which may be defined by referring to Fig. 9, is the quotient formed by the dose, as determined for point P, on the central ray of the beam in a water phantom to the dose determined at the same point P′, with the water phantom removed. The dose at point P would be determined from Eq. 13 and the dose at P′ by Eq. 15, both exposures being made for the same time interval. In practice, it is assumed that all factors except the ion chamber readings will cancel, and tissue air ratios are actually taken to be

$$T_{\mathrm{a}}(d, W_d) = M_{\mathrm{P}}/M_{\mathrm{P}'} \tag{16}$$

In this expression d, is the depth below the surface of the phantom and W_{d} is the field size at that depth. Tissue air ratio is an expression of the way the radiation beam is attenuated and scattered by the material of the phantom. It is the most fundamental of the relations discussed, and all of the others can be derived from it. Numerical data for this quantity for Co-60 are readily available.

Backscatter Factor

The ratio of doses determined from points Q and Q′ of Fig. 9 is a special value of the tissue air ratio. The depth, d_{Q}, is the special depth just needed to produce electronic equilibrium at the point of dose measurement. At this point primary attenuation is the same in the phantom at Q and in the small mass of phantom-like material placed at Q′ in order to make the measurement. Most of the scattered radiation reaching point Q is scattered backward from within the phantom. For the range of X rays that were in use before the advent of ^{6O}CO, the depth d_{Q}, was very small and the point Q, was considered to be on the surface, hence the name backscatter factor. This quantity is also called "peak scatter factor" because the depth at which electronic equilibrium is attained also tends to be the depth of peak dose in the phantom. For ^{60}Co radiation, the depth of electronic equilibrium is taken to be 0.5 cm.

Percent Depth Dose

Whereas tissue air ratios relate doses in the phantom to doses free in air, percent depth doses interrelate doses at points within the phantom. Again referring to Fig. 9, the dose at point P is related to that at point Q by the percentage depth dose.

$$P(d, d_{\mathrm{Q}}, W, F_0) = 100 M_{\mathrm{P}}/M_{\mathrm{Q}} \tag{17}$$

For this quantity, the field size is defined at the surface, and the distance F_0 from the source to the surface must be stated. The doses at points P and Q should be determined from ion chamber measurements by the factors indicated in Eq. 15, and, as for tissue air ratios, it is generally assumed that all factors, except for instrument readings, cancel between the numerator and denominator.

Since point P is farther from the source than is Q, part of the falloff in dose with depth is due to the inverse square attenuation. Because of this, percentage depth doses increase with SSD. For example, the most common source–surface distance in use for Co-60 has been 80 cm. This was chosen as a compromise between increasing percentage depth dose and decreasing output. If the surface distance is increased from 80 cm to 1 m, the percentage depth dose at 10 cm in a 10×10 cm beam will increase from 55.6 to 57.8. This change is just slightly less than would be entirely accounted for by the inverse square law.

Tissue Phantom Ratios

For radiation of energy higher than that of cobalt, the dosimeter must be equipped with thick walls, and its size makes it inconvenient for use in air—particularly for small field sizes. It becomes convenient, therefore, to make the reference measurement in a phantom rather than in air. This is indicated in the right side of Fig. 9 by the point indicated by Q, which is the same distance from the source as is P (and P′), but is in a phantom at some chosen reference depth. The tissue phantom ratio is then the ratio $D_{\mathrm{Q}}/D_{\mathrm{P}}$ and is entirely analogous to tissue air ratio and has many of the same properties. This quantity is, for example, also independent of distance from the source.

Tissue phantom ratios were introduced by Karzmark et al. (21) for use with high energy radiation, but can be applied equally well to Co-60 radiation.

Relationships between the Dose Calculation Functions

From Fig. 9, one can easily see the relationships between the various doses. For example, D_{P} can be related to D_{Q} directly by a percentage depth dose. It could also be

expressed by means of two tissue air ratios and the inverse square law:

$$D_{\mathrm{P}} = D_{\mathrm{Q}} \frac{T(d_{\mathrm{P}}, W_{\mathrm{d_P}})}{T(d_{\mathrm{Q}}, W_{\mathrm{d_Q}})} \left(\frac{F_{\mathrm{Q}}}{F_{\mathrm{P}}}\right)^2 = \frac{D_{\mathrm{Q}}}{100} P(d_{\mathrm{P}}, d_{\mathrm{Q}}, W_{\mathrm{d_Q}}, F_0) \quad (18)$$

The tissue phantom ratio is a combination of two tissue air ratios:

$$T\mathrm{p} = \frac{T(d_{\mathrm{Q}}, W_{\mathrm{Q}})}{T(d_{\mathrm{P}}, W_{\mathrm{P}})} \quad (19)$$

PENUMBRA

All of the previous considerations of dosimetry have been for points on the axis of the beam. Treatment planning is a 3D process, and regions not on the axis must also be considered. The behavior of the dose at points off the beam axis can be discussed by referring to Fig. 11.

In Fig. 11a, the radiation beam is incident on a point X′, in air. The conditions are the same as for the left side of Fig. 9. Consider a small dosimeter to be moved laterally across the beam from A to F. At A it is shielded by the collimator, while at X′ it is in the middle of the beam, in full "view" of the source. The dose will be at its greatest value at X′. At C it would still be in full view of the source, but it is slightly further away from the source than it is at X′ and the dose will be slightly lower. The expected doses at A, X′ and C, as well as other points on the line are shown by the dashed lines in Fig. 11b. At point D, the collimator blocks off half of the source and the dose would be expected to be one-half of its value at C. The point at E is just out of view of the source, and ideally the dose here should sink to zero. The portion of the line A–F between C and E is called the geometric penumbra. It is dependent on the diameter of the source, the distance f_c, from the source to the end of the collimator, and the distance $(f–f_c)$, from the end of the collimator to the line A–F. The geometrical penumbra is given by the very simple relation:

$$p = s \frac{(f - f_{\mathrm{c}})}{f_{\mathrm{c}}} \quad (20)$$

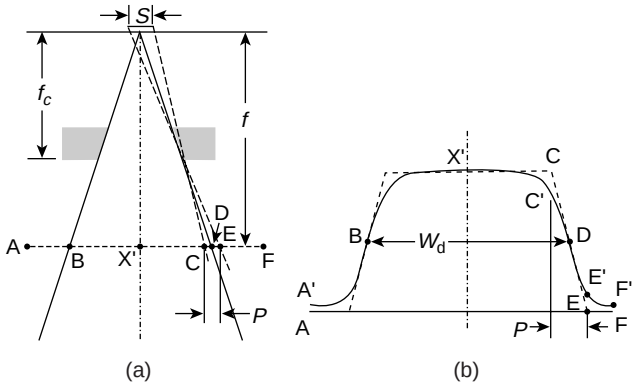


Figure 11. Diagrams showing the geometrical considerations involved in describing the shape of a cross-beam profile for a cobalt unit. (a) Shows the source and the collimator and (b) side shows a dose profile line A–F.

The actual measured penumbra differs somewhat from this and is always a little larger. The source does not behave like a sharp, well-defined disk because of its volume, and the radiation therefore scattered within it and the radiation scattered from the structures that hold it in place, and from the beam collimating apparatus. There is also, inevitably, some transmission through the collimator and some scattering from its lower end. The result is that the dose outside of the beam at points A and F is not zero, and the real dose profile is rounded off as depicted by the solid curve in Fig. 11b.

The shape of the dose profile in a phantom for ^{60}Co radiation is only slightly different from that observed in air. The penumbral region is broadened somewhat by the transport of energy along the tracks of the electrons that are set into motion by the photons near the edge of the beam.

The meaning of field size can also be derived from Fig. 11. It is, by convention, taken to be the distance between points B and D. It is indicated as W_{d} in that diagram. This is the distance between the points that are at 50% of the dose on the axis at the same depth. It is also the full width at half maximum (fwhm) of the dose profile. Normally, the measurement of field size would be made in a phantom.

ISODOSE CHARTS

A more complete description of the dosage pattern of the beam is by means of an isodose chart. An isodose chart is a map of the distribution of the dose in a plane. Such charts are found in many books and papers in the literature and only one example will be given here. In Fig. 12, a small

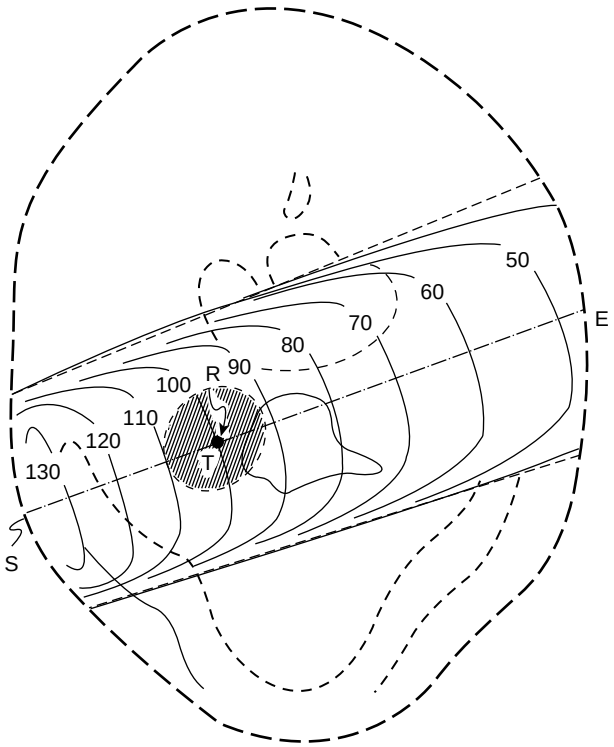


Figure 12. Diagram showing an isodose chart for a beam from a cobalt unit treating a tumor in the neck of a patient. The target and some structures are shown.

beam from an isocentric cobalt unit is treating a tumor in the neck region of a patient. This is an application for which cobalt radiation is still ideal. The target area, which is shown by the cross-hatched region, has been chosen by a radiation oncologist. A safety region has been allowed for and the beam is planned to be directed as shown. The dose at the target will be calculated from the calibration information as described above and (in this case) a tissue air ratio. On the diagram, it has been given the value 100. The solid lines in Fig. 12 show the distribution of percentages of the dose planned for the center of the target. In this case, the dose distribution has not been corrected for air cavities that might be in the path of the beam but modern treatment planning methods, carried out by computers, would include such considerations.

The dose at the center of the tumor is $\sim 77\%$ of the maximum dose, and a single beam like this would not be deemed suitable. The planning process would be carried further by the addition of at least one more beam from another direction, so that the two would cross at the tumor and produce the maximum dose there. Such a treatment plan might even call for four or even more beams, all of which would be arranged to cross at the target.

Complete isodose distributions drawn for individually designed treatments for individual patients are part of the normal procedures of treatment planning. Dose calculation functions that have been discussed in this article have been incorporated into computer programs and enhanced into procedures that allow calculation of dose distributions for complicated treatment conditions. These calculation procedures have been refined to the extent that the 3D shape of the patient and tissue inhomogeneities can be accounted for. The development of these calculation methods has a lengthy history. Suffice it to say that current methods of calculation use Monte Carlo procedures and are quite precise.

NOTABLE FIRST CLINICAL APPLICATIONS

The use of large irregularly shaped radiation beams was introduced for use with cobalt units in the late 1950s. The task was irradiation of chains of lymph nodes for treatment of Hodgkin's disease. This was so successful that a diagnosis of this disease went from a virtual "death sentence" to one that was highly curable. These "mantle" fields, frequently shaped from low melting point lead alloys for individual patients, are still in use. One of the earliest developments of a computer program for the calculation of the dose distribution was introduced with these treatments in mind (22). This program too, is still in use.

A precursor to today's intensity modulation radiation therapy was instituted in the 1960s in Japan by Takahashi who described the use of multileaf collimators and dynamic treatment delivery with a cobalt unit in 1963 (23).

A group under A. Green at the Royal Northern Hospital in London, England pioneered conformal radiation therapy by developing cobalt machines in which the patient was automatically positioned during rotational therapy by moving the treatment couch and machine gantry by electromechanical systems. It was given the name "The Tracking Cobalt Project" because it attempted to make the dose distribution conform to the spread of the disease. With a similar intent Proimos in Patras, Greece and later Rawlinson and Cunningham in Toronto (24), described the use of synchronous shielding in a Co-60 beam to make the radiation beam conform to the target while avoiding critical normal tissues.

SUMMARY AND CONCLUSIONS

It is still likely, even now, that more cancer patients have been treated by radiation from cobalt units than by any other kind of radiation. The number is estimated to be > 30 million (25). The cobalt unit was the backbone of radiation therapy for over four decades. The cobalt unit is mechanically simple and its output is totally predictable and reliable. Sources with sufficient strength to enable practical, short treatment times can easily be produced. Because of the source decay, sources must be renewed at intervals of 5 years or so, but this procedure is quite straightforward and its expense is more than offset by the low maintenance cost of the machine.

The beam characteristics are well known and relatively easy to measure. It is also easy to make special filters and beam modifiers for individual treatment needs. The energy is high enough to provide skin sparing. The most important single parameter in choosing a radiation energy for therapy is depth dose and the depth dose of cobalt radiation is quite satisfactory for treating tumors that are within 10 cm or so of the surface. This includes head and neck tumors and all but deep-seated lesions in very large patients. With respect to this quantity ^{60}Co is in the middle ground. It remains the unit of choice as a first unit in a developing department and is a must as part of the equipment for any large radiotherapy department. Cobalt units are still being manufactured and sold at about half the rate that obtained at the peak of their use. Modern cobalt units include many of the technological innovations, such as computer control, that are part of the more modern treatment machines. An excellent chapter on Co-60 and its role in modern times has been written by Glenn Glasgow (26). This is recommended to the interested reader.

BIBLIOGRAPHY

1. Sampson M, Ridenouri LN, Bleakney W. A long lived radio-cobalt produced by irradiating cobalt with neutrons. Phys Rev 1936;50:382.
2. Livingood JJ, Seaborg GT. Radio-active isotopes of Cobalt. Phys Rev 1941;60:913.
3. Mayneord WV, Cipriani AJ. The absorption of gamma-rays from ^{60}Co. Can J Res Sec A: Phys Sci 1947;25:303.
4. Johns HE, Bates LM, Watson TE. 1000 curie cobalt units for radiation therapy. The Saskatchewan cobalt-60 unit. Br J Radiol, 1952;25:296.
5. Grimmett LG, Kerman HD, Brucer M, Fletcher GH, Richardson JE. Design and construction of a multicurie cobalt teletherapy unit. A preliminary report. Radiology (Easton Pa) 1952;59:19.
6. Robinson RF. The race for Megavoltage. Acta Oncol 1995; 34:1055.

7. Johns HE, MacKay JA. A collimating device for ^{60}CO teletherapy units. J Fac Radiol, London 1953-1954;5:239.
8. Attix FH. Introduction to Radiological Physics and Radiation Dosimetry. New York: John Wiley and Sons Inc.; 1986.
9. Greening JR. Fundamentals of Radiation Dosimetry. Medical Physics Handbook 6. Bristol, England: Adam Hilger; 1981.
10. Johns HE, Cunningham JR. The Physics of Radiology. 4th ed. Springfield, (IL): Charles C. Thomas; 1983.
11. ICRU Report 33. Radiation Quantities and Units. Bethesda, (MD): International Commission on Radiation Units and Measurements; 1980.
12. Bentel GC. Radiation Therapy Planning. 2nd ed. New York: McGraw-Hill; 1996.
13. Khan FH. The Physics of Radiation Therapy. 3rd ed. Philadelphia: Lippincott Williams and Wilkins; 2003.
14. Johns HE, Cunningham JR. A precision cobalt 60 unit for fixed field and rotation therapy. Am J Roentgenol 1959;81:4.
15. Cunningham JR, Ash CL, Johns HE. A double headed cobalt 60 teletherapy unit. Am J Roentgenol 1964;92:202.
16. Leung PM, Rider WD, Webb HP, Aget H, Johns HE. Cobalt-60 therapy unit for large field irradiation. Int J Radiat Oncol Biol Phys 1981;7:705.
17. Rogers DWO, Bielajew AF, Ewart GM. Co beam contamination from the source capsule (Abstr.). Med Phys 1984;11:401.
18. American Association of Physicists in Medicine (AAPM), Task Group 51, A protocol for the determination of absorbed dose from high energy photon and electron beams. Med Phys 1983;120:741.
19. A Code of Practice for Absorbed Dose Determination in Photon and Electron Beams. Vienna: International Atomic Energy Agency (IAEA); 1987.
20. Cunningham JR, Woo M, Rogers DWO. The dependence of mass energy absorption coefficient ratios on beam size and depth in a phantom. Med Phys 1986;13:496.
21. Karzmark CJ, Deubert A, Loevinger R. Tissue-phantom ratios-an aid to treatment planning. Br J Radiol 1965;38:158.
22. Cunningham JR, Shrivastava PN, Wilkinson JM. Program IRREG–Calculation of dose from irregularly shaped radiation beams. Comp Prog Biomed 1972;2:192.
23. Takahashi S. Conformation radiotherapy-rotation techniques as applied to radiography and radiotherapy of cancer. Acta Radiol 1965;242 (Suppl): 1.
24. Rawlinson JA, Cunningham JR. An Examination of Synchronous Shielding in 60-Co Rotation Dose Distributions. Radiology. 1972;102:667.
25. Battista JJ. Cobalt-60 Radiation Therapy: Fifty Years Review and More. London: Ontario; October 27th 2001.
26. Glasgow GP. Cobalt-60 Teletherapy. Chapt. 10. In: Van Dyk J, editor. The Modern Technology of Radiation Oncology. Madison (WI): Medical Physics Publishing; 1999.

See also PHANTOM MATERIALS IN RADIOLOGY; RADIATION DOSIMETRY FOR ONCOLOGY; RADIOTHERAPY TREATMENT PLANNING, OPTIMIZATION OF; X-RAY THERAPY EQUIPMENT, LOW AND MEDIUM ENERGY.

COCHLEAR PROSTHESES

FRANCIS A. SPELMAN
University of Washington
Seattle, Washington

INTRODUCTION

Cochlear prostheses (also called *cochlear implants*) bypass acoustic processing of sound by the cochlea and convert acoustic signals into electrical currents. These currents are delivered via intracochlear electrodes, which directly stimulate the auditory nerve fibers that connect the cochlea to the central nervous system. Cochlear prostheses convert auditory signals into minute electrical currents that stimulate auditory nerve cells via electrodes placed near viable nerve cells. Cochlear implants differ profoundly from acoustic hearing aids. They stimulate the cells of the auditory nerve directly, bypassing the hair cells of the organ of Corti. Acoustic aids increase the mechanical signals that are delivered to the hair cells, aiding their depolarization and the delivery of signals to the auditory nerve. Since the introduction of commercial implants nearly 30 years ago, cochlear prostheses have become one of bioengineering's prominent success stories: $>60{,}000$ people use cochlear implants worldwide. The devices provide patients with a means to overcome deafness. Their success is such that, since the time that the article was written about cochlear implants in the first edition of this Encyclopedia, the cochlear implant has been recommended for people who are severely deaf, rather than reserving the implant for the profoundly deaf (1). Cochlear prostheses provide the standard treatment for people who are profoundly deaf.

In addition to cochlear prostheses, some prostheses are implanted surgically in the central nervous system as auditory brainstem implants, in the cochlear nucleus, or as mid-brain implants in the inferior colliculus.

This article is an update of the article *Cochlear Prosthesis* in the 1st ed. of this Encyclopedia (2).

CANDIDATES FOR IMPLANTS

Hearing loss can occur in either one or both ears. The common classifications of hearing impairment are mild (21–40 dB), moderate–severe (61–70 dB), severe (71–81 dB), and profound (90+ dB) (3). Here dB ($20 \log_{10}[P_2/P_1]$) is the sound pressure, P_2, referenced to normal hearing thresholds, P_1, usually measured at 500, 1000, and 2000 Hz. It refers to the increase of sound pressure that must be used for a subject to reach hearing threshold. Blanchfield et al. number the severely to profoundly deaf between 464,000 and 738,000, all of whom are candidates for cochlear implants (4).

Some prostheses are implanted surgically in the cochlear nucleus. The numbers of patients receiving those devices are much smaller than those who receive implants in the cochlea, ~ 300 people (5). The candidates come primarily from subjects with neurofibromatosis (6–8). The morbidity and mortality with central implants is small, and the success is reasonable. The subjects do not do as well as those with the cochlear prostheses described below, but are able to decode speech (6). The emerging field of central auditory implants will not be covered further in this article because the numbers of users are relatively small at this time.

THE AUDITORY SYSTEM

A complete description of the functioning of the peripheral auditory system is beyond the scope of this article. However, to understand the operation of the prosthesis, one

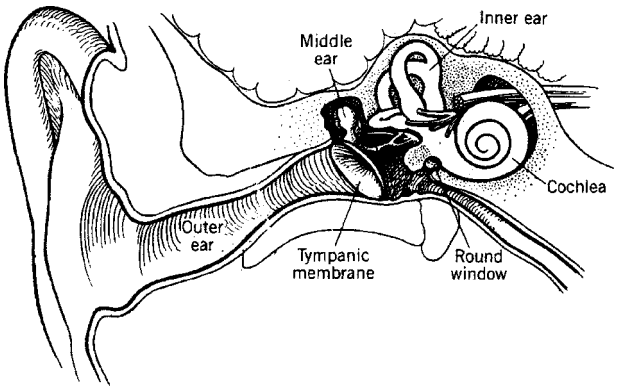


Figure 1. A sketch of the peripheral auditory system, external ear, ear canal, eardrum, middle ear, and inner ear (2).

must know a little about the anatomy and physiology of the peripheral auditory system, which consists of the external ear, the middle ear, and the inner ear (9). Figure 1 shows the auditory system in a simplified form. Sound impinges on the external ear and is guided by way of the ear canal to the tympanic membrane (eardrum). The tympanic membrane vibrates with a relatively large displacement and low pressure. The ossicles (bones) of the middle ear act as an acoustic impedance transformer to change the vibration to relatively small displacement and high pressure at the oval window. The cochlea, the spiral-shaped organ of the inner ear, contains the cells that convert mechanical motion into the electrochemical signals that are recognized by the nervous system (9,10). Several sites on the World Wide Web provide animations of the operations of the components of the auditory system. One such site may be found at, http://www.neurophys.wisc.edu/animations. Other animations and data are maintained in a "virtual library" that has been assembled by the Association for Research in Otolaryngology at its web site http://www.aro.org. Geisler refers to both sites in his work, From Sound to Synapse (9).

THE AUDITORY PERIPHERY

The peripheral auditory system (Fig. 1) consists of the external ear, the middle ear and the inner ear (9). The external ear guides acoustic waves through the external auditory meatus to the tympanic membrane, which vibrates in response to air moving in the ear canal. The middle ear acts as a mechanical transformer, a system of levers and pistons, to match the air-driven tympanic membrane to the fluid-filled inner ear, the cochlea (9).

Figure 2 shows a cutaway view of the inner ear and its three chambers or scalae, that is, the scala vestibuli and the scala tympani, which communicate via the helicotrema, an opening at the apical end of the cochlea, and the scala media, which is isolated from the other two scalae by membranes (9,10). The stapes (stirrup) of the middle ear drives the fluids of the scala vestibuli and in doing so deflects the membranes of the scala media (10). One of these membranes, the basilar membrane, bears the hair cells, the motion-sensitive cells that excite the VIII cranial nerve (9,10).

The inner ear acts as a transduction and signal processing mechanism. Auditory information is decomposed into its fundamental frequencies by the frequency-sensitive basilar membrane. Amplitude, phase, and frequency information is carried by the cells of the auditory (VIII cranial) nerve. Simplistically, sounds are decomposed into their spectral peaks (11).

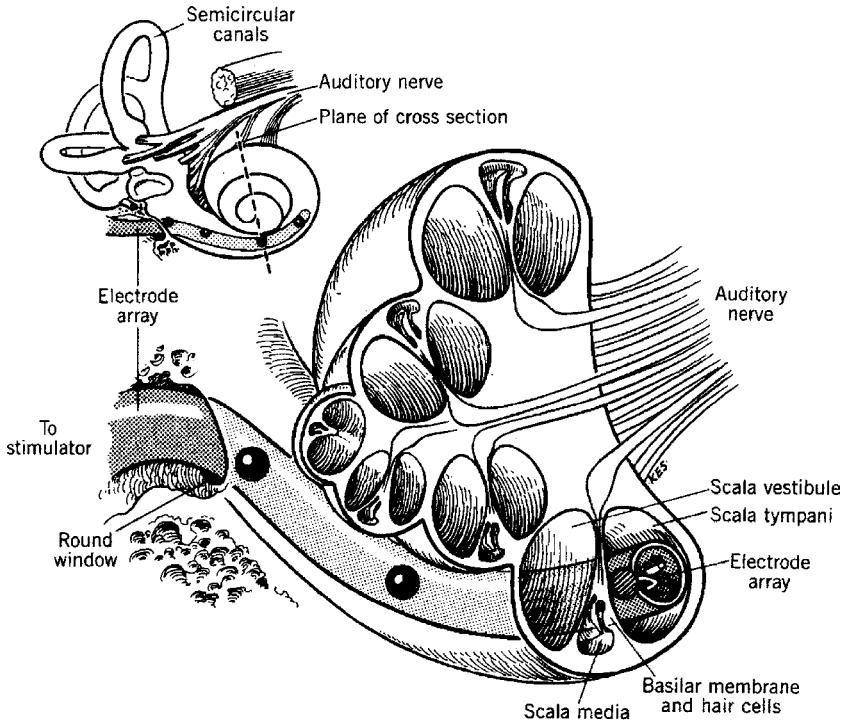


Figure 2. Cutaway view of the cochlea of the inner ear, showing the three chambers or scalae of the ear, and an artist's conception of a cochlear electrode array inserted into the scala tympani of the cochlea (2).

Each of the 30,000-odd fibers of the auditory nerve has an auditory threshold function that is sensitive to a small range of frequencies. All threshold minima lie within 10–15 dB; fibers have dynamic ranges that can be as much as 30–40 dB at their characteristic frequencies (9,12). The rate at which a single peripheral fiber fires is a monotonically increasing function of the acoustic stimulus at its characteristic frequency. The dynamic range of a fiber depends on a number of factors, including its threshold and its spontaneous firing rate, the latter of which can be as large as 100 spikes/s (9).

The responses of auditory nerve fibers are nonlinear. At low intensities, the responses of single nerve fibers mimic the frequency spectra of the complex sounds that stimulate the ears of experimental animals (13). At higher intensities, the spectra produced by the responding fibers are dominated by the low frequency component of the speech sound (its first formant) and the distortion products of that frequency (13). Recent evidence provides strong support for nonlinear system to preserve speech sounds at low and high intensity, in quiet and in noise (9).

In summary, the auditory system has a number of features that enable it to decode sound: (1) specific cells are excited at threshold by specific acoustic frequencies; (2) increasing intensity of an acoustic signal causes an increasing spread of influence from cells for which it is the best frequency, to cells that respond at threshold to other frequencies; (3) the intensity of a particular signal appears to be coded both by the rate at which cells fire and by the numbers of cells excited by a particular stimulus; (4) nonlinear properties of the auditory system cause the suppression of one cell's response to one frequency by stimulation with another frequency, by saturation of rate and by the production of distortion products in the system's response to high intensity excitation; and (5) frequency information contained in complex stimuli is preserved in the temporal responses of auditory neurons.

HISTORY OF COCHLEAR PROSTHESES

The first report of electrical stimulation of the ear is attributed to Volta, in a paper read to the British Royal Society in July of 1800 (14,15). He reported that his approach, using perhaps 50 V excitation, was uncomfortable, sounding like the boiling of fluid. He did not repeat the study. More recently, Djuorno and Eyries (16) reported the first attempt to excite the auditory nerve directly with electricity. Later, Doyle et al. reported results with electrical stimulation of the auditory nerve (17). Simmons performed an experiment a year later in which he went further, stimulating the VIII auditory nerve and the inferior colliculus of a human patient, showing that it was possible for the subject to distinguish frequencies well below 900 Hz, but not >1000 Hz (18). Simmons demonstrated that both peripheral and central stimulation of the auditory system was possible. In 1964, the House Ear Institute began an extensive series of surgeries to implant cochlear prostheses, reporting on their long-term effects in 1973 (19). The first experiments on multichannel cochlear prostheses were initiated by Simmons et al. in 1979 (20). Their results were promising, and now multichannel implants are the standard of the industry.

Since the first experiments, cochlear prostheses have been built and applied worldwide, receiving approval from governmental agencies and remarkable success in $>60{,}000$ patients. Indeed, cochlear prostheses are considered the standard treatment for profoundly and severely deaf adults. Three commercial firms, Cochlear Corp. (Sydney, Australia), Advanced Bionics Corporation (Valencia, CA; recently purchased by Boston Scientific Corporation), and Med-El Corporation (Innsbruck, Austria) produce cochlear implants successfully. The early cochlear implants were single-channel devices (21), but all of the cochlear prostheses that are implanted today are multichannel devices (22).

THEORY OF OPERATION

The cochlear implant operates on the premise that, if the hair cells of the auditory system are damaged, they can be bypassed and that neurons can be driven directly with very small electrical signals. Figure 3 shows a greatly simplified block diagram of a cochlear implant. Acoustic signals are transduced by a microphone, whose small electric signal is amplified. An external processor decomposes the electrical analogue of the acoustic signal. In the processors that are produced today, processing is digital, with the processor analyzing the instantaneous frequency content of the acoustic signal in the frequency domain. The signals are sent across the skin via a radio frequency link (in the VHF band) that transmits both information and power from the outside of the subject to the inside. These transcutaneous signals are shown with bidirectional paths. Data can be transferred in both directions, providing information to therapists about the condition of the electrodes and the state of the auditory system of the patient. The data flowing to and from the external signal processor are serial bit streams.

The transcutaneous bit streams can have rapid rates: consider that the sampling rate of the audio signal can

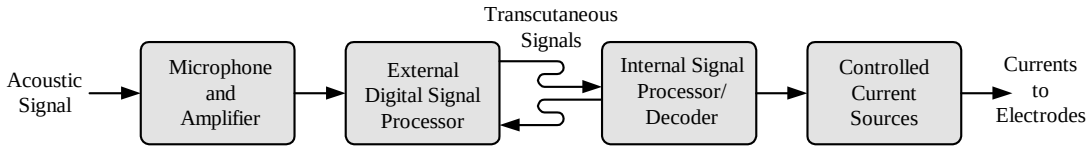


Figure 3. Simplified block diagram of a cochlear implant. Four blocks are shown, a microphone and amplifier, external digital signal processor; internal signal processor/decoder; and, controlled current sources (see text; after Ref. 23).

exceed 20,000 samples/s, and that updates of information delivered to the internal processor may present data at 5000 or more data points per second. The data must include the electrode(s) that are being driven, the amplitude and the pulse width of the current pulse that is applied. Rates can exceed 80,000 pulses/s (25). The internal signal processor–decoder decodes the incoming bit stream. It distributes drive signals to the current sources, selecting specific sources to drive, and setting the amplitude and duration of the control signals.

The current sources drive the electrodes of the cochlear implant's electrode array. Those electrodes are placed in the scala tympani of the cochlea, and direct the current drive signals to the neurons of the auditory nerve, the VIII cranial nerve. The electrodes of the array are placed in proximity to the neurons, in order to reduce the threshold currents necessary to excite the cells, and to reduce current spread within the inner ear (26).

Figure 4 shows one of the cochlear electrode arrays that is produced commercially by Cochlear Corporation (Sydney, Australia). There are 24 contacts in all, two of which are placed outside of the cochlea, leaving 22 contacts that may be driven to excite neurons of the auditory nerve. The contacts may be driven as monopoles (single internal current sources, referenced to an return contact external to the inner ear), as dipoles (pairs of internal current sources) or as combinations of three or more contacts. The contacts are placed along the inside of the spiraling, silicone carrier. The carrier is shaped to fit snugly against the modiolus of the scala tympani. The contacts can be driven singly or in combinations, for example, as dipoles or as multiple sources.

The three manufacturers of cochlear implants use scala tympani arrays that are similar to the array shown in Fig. 4. However, other approaches to stimulate the auditory nerve cells are possible. Normann and his colleagues have tested monolithic electrode arrays that are designed to penetrate the auditory nerve directly (27). Like others before him, Normann realized that bringing electrode contacts near the neurons will reduce thresholds and limit the spread of excitation (28–30). The concept has not been tested chronically in human subjects.

Figure 4. Picture of the Nucleus 24 Contour electrode array, showing 24 contacts and a shape that is designed to appose the modiolar wall of the scala tympani (24).

The concept of an *information channel* is critical to the understanding of the cochlear prosthesis. A channel may drive current to one electrode, but it often distributes drive to two or more electrodes. Field shaping and steering techniques suggest the use of multiple electrodes for each channel (31–33). Indeed, demonstrations by Bierer and Middlebrooks (33,34) showed that the quadrupolar configurations (called tripolar in the Bierer paper) produce more focused stimuli than either monopolar or bipolar excitations. Recent experiments in cats have upheld the finding, showing that multipolar stimulation allows two triads of electrodes to be driven simultaneously without significant crosstalk (35). It is clear that a channel may involve several electrodes driven simultaneously, and cannot be defined as the information conveyed by a single contact on an electrode array. While one contact may be driven at a time, bipolar stimulus configurations are common and multipolar configurations may emerge soon.

Signal processing techniques have changed dramatically since the time that the first version of this article was written. The number of available electrodes has more than doubled. In common to most processors is a bank of filters, analogue or, commonly, digital. The filtered signals are decomposed into time-varying envelope signals that are compressed and delivered as either amplitude modulated pulses or width modulated pulses. The pulses are delivered with a variety of strategies.

Continuous interleaved stimulation (CIS) is a technique by which a single electrode is stimulated at a time in order to eliminate field interactions between and among channels when electrodes are driven as monopoles (31). The electrodes receive signals from specific filters. The signals are converted to symmetrical, rectangular, biphasic current pulses whose amplitudes may be proportional to the envelope of the filter signal and whose width is invariant. Conversely, amplitude can be held constant and width can be varied. More recently, Advanced Bionics Corporation has used a processor whose repetition rate can be 5800 pulses/s per channel, to develop rapid updates of channels in the CIS paradigm. In a recent processor, HiRes, stimulation rates can be as much as 5800 pps when two widely spaced channels are driven simultaneously, and drops by one-half when the two channels are driven sequentially (36).

The n-of-m strategy employs a larger number of filters, m, than there are electrodes, n (37). Depending on which filters contain the maximum acoustic energy, pulses are delivered to appropriate electrodes. The cochlea is organized tonotopically along the basilar membrane. Hence, each electrode's field excites a specific group of characteristic frequencies in perceptual space. Those filters that exhibit the maximum energy determine the electrodes that will be driven by a given temporal sample of the acoustic signal. Biphasic, symmetrical, rectangular pulses are delivered to specific electrodes, n, at particular sample times. Because of the field interactions between electrodes no more than two channels are driven during a given sample.

Other techniques include simultaneous analog stimulation (SAS), in which widely separated electrodes

are driven simultaneously to increase the rate at which information is transferred to the auditory nerve. The field interactions are reduced by driving electrodes that are separated by several millimeters in the inner ear (38). Simultaneous analog stimulation is a special case of the "filters with compression" technique described by Eddington > 20 years ago (39). Today, fewer electrodes are driven simultaneously, but they are updated more rapidly (40). Thus, SAS is a variation of both Eddington's filters and CIS. Eddington described a means by which electrodes were assigned the compressed analog outputs of filters. Those analogue signals were delivered continuously to the electrodes.

A potentially exciting new technique of stimulation takes advantage of the stochastic behavior of auditory neurons. If a stimulator provides high rate conditioning pulses to its electrode array, it is possible to simulate the stochastic firing frequencies of the cells of the auditory nerve (41). This approach has been tested in small numbers of European patients with what appears to be dramatic success, particularly with auditory signals in noise (Rubinstein, personal communication; see below).

Despite the richness of the processing techniques that have been employed, there are still hurdles to be overcome. The number of true, simultaneous channels is too small. It should be at least 16; there is often a mismatch between the frequency assigned to an electrode and its position in the cochlea; the signals that are delivered to the neurons do not contain fine temporal information; the phase information between channels is not preserved; and, there may be neurons missing, causing some electrodes and critical frequencies to be missing as well (42). Future implants may be able to address some of the concerns that are raised here.

EVALUATION OF COCHLEAR PROSTHESES

When human subjects first used cochlear implants, the numbers of subjects were small and tests were not standardized. As the devices improved, standard tests were developed and used across the centers at which implantation was being done (15). The tests include materials that are open and closed set. The test subjects do not review open set materials prior to the test, whereas closed set materials are reviewed before testing takes place. Subjects participate in word tests and sentence tests. In the former, single words are presented while in the latter sentences are presented and the subjects can deduce parts of the sentence logically.

In addition to providing word and sentence tests, consonant (C) and vowel (V) discrimination tests are included in the test batteries. In these tests, nonsense utterances, CVCs or VCVs, are presented and the subject must identify the appropriate vowel or consonant.

Open word tests are difficult while sentence tests are relatively easy. For example, implant users have steadily increased their comprehension of sentences from much < 10% with early single channel devices to 80% or above with today's multichannel devices (22). Many implant users are able to converse on the telephone, a significant result, since they cannot rely on the cues presented by lip reading in that situation. Still, word comprehension from open-word sets remains relatively low, between 40 and 50%, and most users dislike listening to music (43). Clearly, the context that comes from sentence structure and content is important to comprehension, and the complex spectral content of music makes it difficult.

Cochlear implants are a great bioengineering success. Wilson used an aviation metaphor recently, likening the cochlear implant to a DC-3, a reliable workhorse of an aircraft without the sophistication of a twenty-first century transport airplane (43). The implant has advanced from the single-channel stage of the Wright flyer, but has yet to reach its pinnacle.

THE BENEFITS AND RISKS OF IMPLANTATION

Cochlear implants provide clear benefits to their users. For example, hearing-impaired children learn language more rapidly with cochlear implants than they do with hearing aids (44). Adults do well and benefit from their implants, particularly when they are dealing with speech in quiet. However, for patients to achieve the greatest benefits from the device, their prostheses should be adjusted individually for the minimum and maximum stimulation levels for each electrodes in the array, the stimulation rate, and the speech processing strategy (45). Skinner suggests that for best results the parameters should be adjusted for the maximum dynamic range: from quiet sounds to maximum sounds that are "...not too loud..." (45).

A recent survey of patients from Toronto, Ontario, Canada, was taken of 42 early deafened adult users. Of the 30 who responded, > 96% said that they were satisfied with the implant, > 93% would undergo the procedure again, and 90% said that they would recommend the implant to another person in the same situation (46). The subjects were encouraged by family and peer support and bolstered by having a positive attitude before, during and after the process of implantation and therapy.

There are risks associated with the surgery, but they are quite small. Cunningham et al. (47) reviewed the cases of 462 adults and 271 children in a private tertiary care center for the years 1993–2002. They found that the overall incidence of infection postoperatively was 4.1%. Major infectious complications occurred in 3.0% of the cases; those complications required surgical intervention (47). Bacterial meningitis was found in 26 of 4264 children receiving cochlear implants in the United States (48). That was found to be associated with a particular electrode array that used a positioner to place it near the modiolar wall. The array was subsequently withdrawn from the market (http://www.fda.gov/cdrh/safety/cochlear.html), and there have been no other reports of the occurrence of meningitis. Cunningham et al. (47) recommended that children undergo vaccination before implantation to prevent bacterial infections.

THE COST OF IMPLANTATION

A recent article cited the cost of cochlear implant treatment as > $40,000.00, of which $20,000.00 is the approximate

cost of the device itself (49). Despite the high cost of the device and the surgery, the cochlear prosthesis is beneficial when compared with the long-term costs of other medical device procedures (50,51). Garber et al. (49) asked why the cochlear implant has limited access despite its success and the likely market, and surveyed 25 of 231 practices and 96 of 213 hospitals to try to learn what caused the limits of availability. They concluded that both the practitioners and hospitals lose money when they provide cochlear implants, limiting access to the devices. The cochlear implant is approved in the United States for Medicare, Medicaid, and insurance reimbursement.

THE FUTURE OF COCHLEAR PROSTHESES

In a recent review, Wilson et al. (52) suggested that the future held combined acoustic and electrical stimulation, bilateral implants, new electrode designs and closer mimicking of processing in the normal cochlea. This article discusses electrode designs, combined acoustic, and bilateral stimulation and the closer mimicking of processing in the normal cochlea.

HIGH DENSITY ELECTRODE ARRAYS

Electrode arrays have remained much the same for more than a decade. They are built manually on substrates of silicone, using Pt–Ir (90–10%) alloyed electrodes. The group of Dr. Kensall Wise at the University of Michigan has proposed the use of high density arrays that are made on silicon substrates using IrO contacts (54,55). If such arrays can be built for human use, they will reduce the cost of building electrode arrays while increasing the specificity of excitation of cells. Another approach to the problem is to build electrode arrays on multilayered polymer substrates (Fig. 5). Sample arrays have been used to demonstrate the use of high density arrays in animal studies, with clear independence of channels driven in the first turn of the scala tympani (Snyder, Corbett, Bonham, Rebscher, and Johnson, personal communication).

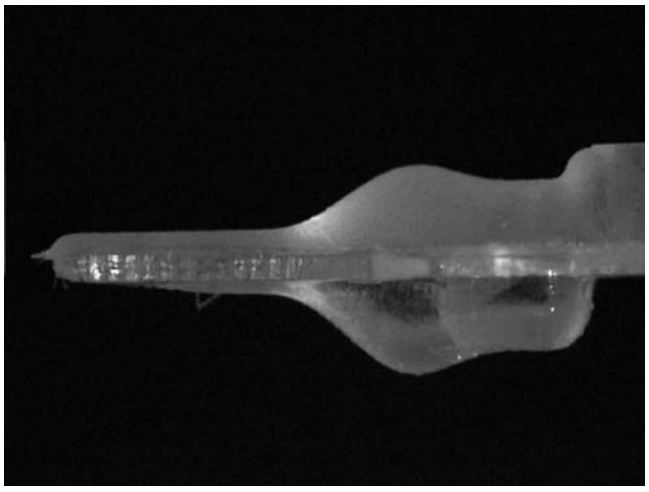


Figure 5. Photograph of a 12-site sample array made by Advanced Cochlear Systems (Snoqualmie, WA) to insert into the scala tympani of a cat (53). The width of each gold electrode contact is 100 μm

The goal driving the development of these high density arrays is to increase the specificity of stimulation and to allow several independent groups of cells to be driven simultaneously (32,56). The work of Jolly (32) and Bierer and Middlebrooks (33,57) suggested that this might be the case. More recent work has confirmed the earlier results and extended them (35,34). The benefit of focused multipolar stimulation and of simultaneous excitation of several independent groups of neurons is not without cost. More driven electrodes require greater current consumption. Current consumption is increased with focusing, since focused stimuli require more applied current to reach the same potential fields in conducting media (58). It is likely that high density electrode arrays will be a part of cochlear implants, but there are engineering challenges to be met before it will happen.

COMBINED ACOUSTIC AND ELECTRICAL STIMULATION

Preliminary studies of combined electrical and acoustic stimulation have been done successfully in both Europe and the United States (52,59,60). The subjects come from the substantial population of people who preserve some hearing for frequencies < 1 kHz, but who are severely impaired for frequencies > 1 kHz. Two questions arise immediately. (1) Can low frequency hearing be preserved after an electrode array has been placed in the high frequency regions of the inner ear? (2) Can acoustic and electrical stimuli be applied simultaneously and successfully?

The likelihood of success is great, particularly if patients have short electrode arrays implanted, avoiding damage to the delicate structures of the inner ear. That concern is critical in the case of the hybrid stimulation scheme, since low frequency information will come via the normal, albeit amplified, acoustic pathway. Two manufacturers, Cochlear Corporation (Sydney, Australia) (60) and Med-El (Innsbruck, Austria) (59) have produced electrode arrays for the purpose and have tested them in clinical settings. The Med-El array has an implanted length of 31.5 mm (59), while the Cochlear Corporation array's length is 10 mm in its latest version (60). Both have had extensive laboratory tests and have been used clinically. Clinical tests confirm the initial hypothosis: when patients suffer primarily from high frequency hearing loss, the use of hybrid stimulation is likely to provide great benefit, and may well increase the numbers of people who can have near-normal hearing (52). Electrical and acoustic stimuli can be combined by implanting one ear with a cochlear prosthesis and using a hearing aid in the contralateral ear. This approach has had some reports of success and is currently under study in research laboratories.

NORMAL PROCESSING: CONDITIONING PULSES

In the 1990s, investigators began to consider the issue of the stochastic behavior of neurons (61) and that high rate

conditioning stimuli might improve the behavior of cells in the auditory nerve, decreasing thresholds and increasing dynamic ranges (52). They proposed to use electric currents with 5 kHz pulse trains of brief pulses, biphasic rectangular pulses of 40 μs duration for each phase (62). Rubinstein and various colleagues pursued the idea further, suggesting that high rate stimuli might mimic stochastic resonance in neurons and improve signal processing in cochlear implants (60). Computer models validated the concept, as did initial tests in a human subject (63). An extensive neurophysiological study confirmed the idea in experimental animals (62).

Rubinstein and Frijns did preliminary tests for the use of high rate, low amplitude conditioning pulses in the processors of some human subjects, reporting success in the majority of their subjects (Rubinstein, personal communication). The concept is certainly a logical and promising idea; whether it will provide a dramatic improvement to cochlear implants is something that will be learned from further experiments in human subjects.

NORMAL PROCESSING: FINE STRUCTURE

Present cochlear implants impose low pass filter functions on the acoustic signals that they decode. Signals are filtered, and their envelopes detected, with a concomitant loss of fine structure. Fine structure is defined as information spanning frequencies from 500 to 10 kHz (22). Speech can be well understood in quiet environments. The users of present-day cochlear implants rarely enjoy music. Some of that may be improved by increasing the fine structure of the signals delivered to the ear via the cochlear implant. The Hilbert transform provides a potential approach to providing both amplitude information and fine structure (22,64,65). Smith et al. (63) determined that the envelope of the transform was important for speech perception, while the fine structure determines localization and pitch.

Processors that employ Hilbert transforms have yet to be produced in quantity. Although prototypes exist, they have not made their way into cochlear implants (64). The development of implants that can reproduce the fine structure of signals is likely to improve cochlear prostheses.

BILATERAL IMPLANTS

Binaural hearing is critical to sound localization and the extraction of auditory signals in noise. In addition, binaural implants may allow listeners to employ the "head shadow" benefit to hear a specific voice in the face of sounds produced by a competing crowd of people (52). Wilson notes promising results from several centers at which patients have received bilateral implants (52). He reports improvements in speech comprehension, as well as the results of several careful psychophysical studies that were focused on the balance between the prostheses that were implanted. Wilson and his colleagues concluded that bilateral implants are likely to provide clear benefits. While users are tolerant of some timing and amplitude mismatches, the careful matching of stimulus sites, that is, electrode locations, may be necessary for success (52). Another issue to consider is the cost of bilateral implantation. Bilateral implantation incurs the cost of two cochlear prostheses and two surgeries. Does the benefit accrued by the patient double? That remains to be seen at the time of this writing.

CONCLUSION

Cochlear prostheses are a clear bioengineering success story. More than 60,000 patients have benefited worldwide. Many users can talk on the telephone and communicate effectively without visual aids, like lipreading. The design of the cochlear prosthesis is likely to improve, even as the number of implantees grows rapidly, indeed, at double-digit rates. With that rich background and rapid growth, there are opportunities for bioengineers to produce even better cochlear prostheses.

ACKNOWLEDGMENTS

This work was sponsored by grants R43DC000531 and R43DC04614 of the National Institutes of Health.

BIBLIOGRAPHY

1. NIH NIH Consensus Statement: cochlear Implants in Adults and Children. Bethesda, MD, National Institutes of Health; 1995.
2. Webster JG. Encyclopedia of medical devices and instrumentation. New York: John Wiley & Sons; 1988.
3. Blanchfield BB, Feldman JJ, et al. The severely to profoundly hearing impaired population in the United States: Prevalence and demographics. Policy Anal Brief H Ser 1999; 1 (October): 1–4.
4. Blanchfield BB, Feldman JJ, et al. The severely to profoundly hearing-impaired population in the United States: prevalence estimates and demographics. J Am Acad Audiol 2001;12(4): 183–189.
5. Kuchta J. Neuroprosthetic hearing with auditory brainstem implants. Biomed Tech (Berlin) 2004;49(4):83–87.
6. Otto SR, Brackmann DE, et al. Multichannel auditory brainstem implant: update on performance in 61 patients. J Neurosurg 2002;96(6):1063–1071.
7. Schwartz MS, Otto SR, et al. Use of a multichannel auditory brainstem implant for neurofibromatosis type 2. Stereotact Funct Neurosurg 2003;81(1–4):110–114.
8. Kanowitz SJ, Shapiro WH, et al. Auditory brainstem implantation in patients with neurofibromatosis type 2. Laryngoscope 2004;114(12):2135–2146.
9. Geisler CD. From Sound to Synapse: Physiology of the Mammalian Ear. New York: Oxford University Press; 1998.
10. Dallos P, Popper AN, Fay RR, editors. The Cochlea. Springer Handbook of Auditory Research. New York: Springer; 1996.
11. Sachs MB, Young ED. Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate. J Acoust Soc Am 1979;667(2):470–479.
12. Liberman MC. Auditory-nerve response from cats raised in a low-noise chamber. J Acoust Soc Am 1978;63(2):442–455.
13. Sachs MB, Young ED. Effects of nonlinearities on speech encoding in the auditory nerve. J Acoust Soc Am 1980; 68:858.
14. Volta A. On the electricity excited by mere contact of conducting substances of different kinds. R Soc Philos Trans 1800;90: 403–431.

15. Clark GM. Cochlear Implants: Fundamentals and Applications. New York: Springer-Verlag; 2003.
16. Djourno A, Eyries C. Prothese Autitive par excitation electrique a distance du nerf sensoriel a l'aide d'un bobinage inclus a demcure. Presse Med 1957;35:14–17.
17. Doyle JB, Doyle HD, et al. Electrical stimulation in eighth nerve deafness. Bull LosAngeles Neurol Soc 1963;18:148.
18. Simmons FB, Mongson CJ, et al. Electrical stimulation of the acoustic nerve and inferior colliculus in man. Arch Otolaryngol Head Neck Surg 1964;79:559.
19. House WF, Urban J. Long term results of electrode implantation and electronic stimulation of the cochlea in man. Ann Otolaryngol Rhinol Laryngol 1973;82:504.
20. Simmons FB, Mathews RG, et al. A functioning multichannel auditory nerve stimulator. A preliminary report on two human volunteers. Acta Otolaryngol 1979;87(3–4):170–175.
21. House WF, Berliner KI. Cochlear Implants: from idea to clinical practice. In: Cooper H, editors. Volume 1, Cochlear Implants: A Practical Guide. San Diego, CA: Singular Publishing Group, Inc.; 1991. p 9–33.
22. Zeng F-G. Trends in cochlear implants. Trends Amplif 2004;8(1):1–34.
23. Spelman F. Cochlear Prostheses. In: Ratner BD, Hoffman AS, Schoen FJ, Lemons JE, editors. Volume 1, Biomaterial Science: An Introduction to Materials in Medicine. Amsterdam: The Netherlands Elsevier Academic Press; 2004. p 658–669.
24. Anonymous. Nucleus 24 Contour, Cochlear Americas. 2004.
25. Kessler DK. The CLARION Multi-Strategy Cochlear Implant. Ann Otol Rhinol Laryngol Suppl 1999; 177(Apr.):8-16.
26. Jolly CN, Gstöttner W, et al. Principles and outcome in perimodiolar positioning. Ann Otolaryngol Rhinol Laryngol Suppl 2000;185(12):20–23.
27. Hillman T, Badi AN, et al. Cochlear nerve stimulation with a 3-dimensional penetrating electrode array. Otolaryngol Neurotol 2003;24(5):764–768.
28. Simmons FB. Electrical Stimulation of the Auditory Nerve in Man. Arch Otolaryngol 1966;84 (July, 1966):24–76.
29. White MW, Merzenich MM, et al. Multichannel cochlear implants: Channel interactions and Processor design. Arch Otolaryngol 1984;110:493–501.
30. Arts HA, Jones DA, et al. Prosthetic stimulation of the auditory system with intraneural electrodes. Ann Otolaryngol Rhinol Laryngol Suppl 2003;191:20–25.
31. Wilson BS, Finley CC, et al. Better speech recognition with cochlear implants. Nature (London) (July 18, 1991); 352:236–238.
32. Jolly CN, Spelman FA, et al. Quadrupolar stimulation for cochlear prostheses: Modeling and experimental data. IEEE Trans Biomed Eng 1996;43(8):857–865.
33. Bierer JA, Middlebrooks JC. Cortical responses to cochlear implant stimulation: Channel interactions. J Assoc Res Otolaryngol. 2004;5(1):32–48.
34. Snyder RL, Bierer JA, et al. Topographic spread of inferior colliculus activation in response to acoustic and intracochlear electric stimulation. J Assoc Res Otolaryngol 2004;5(3): 305–322.
35. Bonham B, Snyder RL, et al. The neurophysiological effects of simulated auditory prosthesis stimulation: channel interaction, current steering and channel morphing. San Francisco, CA: University of California at San Francisco; 2004.
36. Anonymous. New methodology for fitting cochlear implants. Valencia, CA: Advanced Bionics Corporation. 1–5; 2003.
37. McDermott HJ, McKay CM, et al. A new portable sound processor for the University of Melbourne/Nucleus Limited multielectrode cochlear implant. J Acoust Soc Am 1992;91: 3367–3371.
38. Anonymous. Clarion S-Series. Sylmar, CA, Advanced Bionics, Inc.; 1997.
39. Eddington DK. Speech discrimination in deaf subjects with cochlear implants. J Acoust Soc Am 1980;68:885–891.
40. Anonymous. PULSARci Cochlear Implant, Med-El; 2004.
41. Rubinstein JT, Hong R. Signal coding in cochlear implants: Exploiting stochastic effects of electrical stimulation. Ann Otol Rhinol Laryngol 2003;112(9, Part 2):14–19.
42. Moore BCJ. Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implants. Otolaryngol Neurotol 2003;24(2):243–254.
43. Wilson BS. The History of Cochlear Implants. Neural Interfaces Workshop, Hyatt Regency Bethesda Hotel, Bethesda, MD: NIDCD, National Institutes of Health; 2004.
44. Skinner MW. Cochlear implants in children: What direction should future research take? 2001 Conference on Implantable Auditory Prostheses, Pacific Grove CA; 2001.
45. Skinner MW. Optimizing cochlear implant speech performance. Ann Otolaryngol Rhinol Laryngol Suppl 2003;191: 4–13.
46. Chee GH, Goldring JE, et al. Benefits of cochlear implantation in early-deafened adults: the Toronto experience. J Otol 2004;33(1):26–31.
47. Cunningham CD, 3rd, Slattery WH, 3rd, et al. Postoperative infection in cochlear implant patients. Otolaryngol Head Neck Surg 2004;131(1):109–114.
48. Reefhuis J, Honein MA, et al. Risk of bacterial meningitis in children with cochlear implants. N Engl J Med 2003;349(5): 435–445.
49. Garber S, Ridgely MS, et al. Payment under public and private insurance and access to cochlear implants. Arch Otolaryngol Head Neck Surg 2002;128(10):1145–1152.
50. Cheng AK, Niparko JK. Cost-utility of the cochlear implant in adults. Arch Otolaryngol Head Neck Surg 1999;125(11): 1214–1218.
51. Niparko JK, Kirk KI, et al. Cochlear Implants: Principles and Practices. Baltimore MA: Lippincott Williams & Wilkins; 2000.
52. Wilson BS, Lawson DT. Ann Rev Biomed Eng 2003;5:207–249.
53. Corbett SS, III, Johnson T, Rebscher S, Carson M, Ketterl J, Snyder R. unpublished results.
54. Weiland JD, Anderson DJ. Chronic neural stimulation with thin-film, iridium oxide electrodes. IEEE Trans Biomed Eng 2000;47(7):911–918.
55. Weiland JD, Anderson DJ, et al. *In vitro* electrical properties for iridium oxide versus titanium nitride stimulating electrodes. IEEE Trans Biomed Eng 2003;49(12): 1574-1579.
56. Clopton BM, Spelman FA. Technology and the future of cochlear implants. Ann Otolaryngol Rhinol Laryngol Suppl 2003;191:26–32.
57. Bierer JA, Litvak L, et al. Effects of electrode configuration on psychophysical measures of channel interaction in cochlear implant subjects. Soc Neurosci 2003.
58. Spelman FA, Pfingst BE, et al. The effects of electrode configuration on potential fields in the electrically-stimulated cochlea: models and measurements. Ann Otol Rhinol Laryngol 1995;104(Suppl. 166):131–136.
59. Adunka O, Kiefer J, et al. Development and evaluation of an improved cochlear implant electrode design for electric acoustic stimulation. Laryngoscope 2004;114(7):1237–1241.
60. Gantz BJ, Turner C. Combining acoustic and electrical speech processing: Iowa/Nucleus hybrid implant. Acta Otolaryngol 2004;124(4):344–347.
61. Rubinstein JT, Abbas PJ, et al. Stochastic Resonance: Can it be exploited by speech processors? Conference on Implantable Auditory Prostheses, Pacific Grove, CA; 1997.

62. Runge-Samuelson CL, Abbas PJ, et al. Response of the auditory nerve to sinusoidal electrical stimulation: effects of high-rate pulse trains. Hear Res 2004;194(1–2):1–13.
63. Rubinstein JT, Wilson BS, et al. Pseudospontaneous activity: stochastic independence of auditory nerve fibers with electrical stimulation. Hear Res 1999;127(1–2):108–118.
64. Clopton BM, Lineaweaver SKR, et al. Method of processing auditory data. United States Patent and Trademark Office. Advanced Cochlear Systems.
65. Smith ZM, Delgutte B, et al. Chimaeric sounds reveal dichotomies in auditory perception. Nature (London) 2002; 416: 87–90.

See also AUDIOMETRY; COMMUNICATIVE DISORDERS, COMPUTER APPLICATIONS FOR.

CODES AND REGULATIONS: MEDICAL DEVICES

MORRIS WAXLER
PATRICIA J. KAEDING
Godfrey & Kahn S.C.
Madison Wisconsin

INTRODUCTION

The U.S. Food and Drug Administration (FDA or agency) regulates medical devices according to specific definitions, classifications, requirements, codes, and standards. The FDAs authority and framework for medical device regulation are specified in the Federal Food, Drug, and Cosmetic Act of 1938, as amended (FDCA). The FDCA is codified at Title 21, Chapter 9, United States Code (21 USC) (1). For purposes of medical device regulation, several acts of Congress amending the FDCA are especially significant: the Medical Device Amendments of 1976, the Safe Medical Devices Act of 1990, the Food and Drug Administration Modernization Act of 1997, and the Medical Device and User Fee and Modernization Act of 2002. The FDA has promulgated regulations for the efficient enforcement of the FDCA. These regulations, which generally have the force of law, are codified in Title 21 of the Code of Federal Regulations (21 CFR or the regulations) (2). The agency also has issued guidances and guidelines to assist in the regulation of medical devices (3).

Pursuant to the FDCA, the FDA determines the entities subject to regulation (e.g., manufacturers, specifications developers), evaluates whether products and regulated entities are in compliance, and initiates appropriate regulatory and enforcement actions to impose penalties for violations. The FDA's requirements affect each stage of a medical device's lifecycle. Some FDA requirements apply to particular periods of a medical device's lifecycle. Others apply more broadly. Design, technical development, preclinical testing, clinical study, market authorization, market approval, postmarket assessment, modification, obsolescence, redesign, and labeling requirements are part of this regulatory framework for medical devices. The FDA's Center for Devices and Radiological Health (CDRH) is the FDA component with primary responsibility for medical device regulation.

WHAT IS A MEDICAL DEVICE?

The FDCA contains definitions for the various product areas the FDA regulates, including medical devices. Under the FDCA, a "device" must be

- "an instrument, apparatus, implement, machine, contrivance, implant, *in vitro* reagent, or other similar or related article, including any component, part, or accessory"
- which is either "intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease, in man or other animals," or "intended to affect the structure or any function of the body of man or other animals," and
- "which does not achieve its primary intended purposes through chemical action within or on the body of man or other animals and which is not dependent upon being metabolized for the achievement of its primary intended purposes" [21 USC § 321(h)].

To be a medical device, a product must achieve its "primary intended purpose" without chemical or metabolic action within or on the body. This characteristic distinguishes "devices" from "drugs". For example, perfluorocarbon gas is injected into the human eye to hold a detached retina in place. The gas has no metabolic reaction with the body and thus is regulated as a medical device. But determining whether the FDA would consider a product, a "device", or a "drug" can be difficult. Products can be medical devices even if there is some chemical or metabolic reactions within or on the body. For example, the body often reacts metabolically to hip and other implants. Because these reactions are side effects rather than the primary intended purpose of these implants, the products are medical devices.

The FDCA's definition of medical device includes a concept that is a key part of the FDA's regulatory framework: A medical device is both the physical product and its intended use or uses. "Intended use" is sometimes described as the express and implied claims made for a product. This concept means, for example, that a manufacturer (and his representatives) cannot, without penalty, label, or promote a laser for refractive correction eye surgery if it is legally marketed only for cardiac surgery. The manufacturer must apply to the FDA for authorization or approval to use the laser for a new indication. Changes in indications or uses can create regulatory hurdles for a manufacturer.

MEDICAL DEVICE CLASSIFICATION

Prior to 1976, the FDCA did not contain any specific provisions for medical device regulation. The Medical Device Amendments (MDA) of 1976 greatly expanded the FDA's statutory authority over medical devices and established a comprehensive regulatory scheme for medical devices. The MDA established three classes of medical devices based on the potential risk of the device to patients

or users. Devices with greater potential risks are subject to more regulatory controls.

Since 1976, the FDA has established classification regulations for >1700 different generic types of devices, and grouped them into 16 medical specialties, such as cardiovascular, respiratory, general hospital, infection control, and restorative (4). Each of these generic types of devices is assigned to one of three regulatory classes depending on the level of controls needed to provide a reasonable assurance of the devices' safety and effectiveness. Unclassified devices and new devices are automatically Class III medical devices. But not all medical devices that a layperson likely would understand to be new remain "new" for purposes of the FDCA. If a manufacturer can show that its device is "substantially equivalent" to a device that was legally marketed in 1976, often referred to as a "predicate device", then the device becomes subject to the classification and requirements that apply to that predicate device.

Class I devices are those posing the least amount of risk. Examples include elastic bandages, examination gloves, and hand-held surgical instruments. Class I devices do not require FDA review prior to marketing. However, Class I devices are subject to the FDCA's general controls for all medical devices. These general controls are the regulatory common denominator for all medical devices, and include do not distribute adulterated or misbranded devices; register the commercial establishment with the FDA; list the marketed devices with the agency; label the devices in accordance with applicable labeling regulations; manufacture the devices in accordance with the quality system and good manufacturing practices regulations (many Class I devices, however, are exempt from this requirement); permit FDA inspection. The FDA has the authority to ban medical devices under appropriate circumstances; restrict the sale, distribution, or use of some devices; and require the submission of records and reports.

Class II medical devices have an intermediate level of risk. General controls alone are not sufficient to address the risks of Class II devices. Examples include powered wheelchairs, infusion pumps, and surgical drapes. Class II devices are subject to special controls that are developed to control risks specific to particular devices. Examples of the types of special controls used by FDA include performance standards, guidelines, postmarket surveillance, and patient registries. Most Class II devices require 510(k) premarket notification. The "510(k)" refers to FDCA section 510(k), codified at 21 USC § 360(k). A 510(k) submission contains information and data to show that the device is "substantially equivalent" to a legally marketed predicate device. Clinical data is usually not required for the FDA to clear a 510(k) submission for marketing. Some Class II devices are exempt from 510(k) clearance.

Class III medical devices are those presenting the greatest risks. Examples include replacement heart valves, silicone gel-filled breast implants, and implanted brain stimulators. In general, Class III devices are subject to premarket approval prior to marketing. General and special controls alone are insufficient to provide a reasonable assurance of the devices' safety and effectiveness. Class III devices are usually devices that are life sustaining, life supporting, or implantable, or have the potential for serious injury (e.g., sight threatening). New devices that are not substantially equivalent to a legally marketed device also are usually subject to premarket approval. A premarket approval application (PMA) contains extensive scientific and technical evidence that demonstrates that a reasonable assurance of safety and effectiveness exists for the device. Clinical studies are usually required to support FDA approval of a PMA.

Under the 1997 amendments to the FDCA, manufacturers of certain devices that have been found to be not substantially equivalent can request immediate reclassification into Class I or II based on the device's low risk level. This process is called *de novo* classification. If the FDA agrees, then the device becomes subject to the requirements of either Class I or II, and a PMA is not required.

FDA-REGULATED ENTITIES

The FDA regulates manufacturers, specification developers, distributors, contract manufacturers, sterilization facilities, importers, exporters, contract research organizations, and clinical researchers of medical devices. In addition, the FDA regulates, and otherwise influences, the use and nonuse of voluntary standards by these organizations and individuals to support their regulatory activities and submissions to the agency. The manner in which parties are regulated depends on their role in the distribution of the device and on the stage of the device's lifecycle. For example, the FDA requires preapproval of medical device clinical trials that present significant risks to patients. On the other hand, establishments must register with the FDA only after the FDA authorizes marketing of the device.

USE OF STANDARDS

The FDA recognizes that a device's conformance with recognized consensus standards can be used to support a PMA, 510(k), or other submissions to the agency (5). The FDA maintains a list of officially recognized standards (6). Some domestic and international standards focus on specific medical devices (e.g., respirators). Others characterize an important aspect of many medical devices, (e.g., electrical safety). The former is sometimes called a "vertical" standard. The latter is called a "horizontal" standard. The agency also issues guidance documents for specific devices that refer to the FDA-recognized standards or to other standards. Standards should be used consistent with FDAs guidances because there can be a considerable delay between the development of consensus standards and the agency's recognition of them.

ENFORCEMENT AND PENALTIES

The FDCA authorizes civil and criminal penalties for violations (21 U.S.C. §§ 331-337). The statute, for example, prohibits the adulteration or misbranding of medical devices as well as the introduction or delivery for introduction into interstate commerce, or the receipt in interstate commerce, of any adulterated or misbranded device. The

FDCA also prohibits the submission of false or misleading information to the agency, including the withholding of material or relevant information. For example, a failure to report to the FDA all device failures that occurred during the clinical trial of a Class III medical device is a violation. Such actions can lead to not only disapproval or withdrawal of the PMA for the device, but also civil and criminal penalties on manufacturer.

The FDCA authorizes the FDA to pursue some remedies administratively, including clinical investigator disqualifications, temporary detention of medical devices, and certain civil money penalties. Other remedies, including product seizures, injunctions, criminal charges, and some civil money penalties, require judicial proceedings in federal court. The FDA refers judicial enforcement actions to the U.S. Department of Justice, and works closely with the Justice Department to prosecute these actions. The FDCA is a strict liability statute, which means that a company's management may be prosecuted for a failure to detect, prevent, or correct violations. Knowing and following the rules is important.

REQUIREMENTS GENERALLY

Marketing safe and effective medical devices in the United States requires an understanding of FDA requirements that govern the entire life cycle of the device. These include requirements for conducting nonclinical laboratory studies and clinical trials, bringing a product to market, manufacturing practices, labeling, reporting device problems and patient injuries, carrying out recalls and corrective actions, and making modifications to the device.

NONCLINICAL LABORATORY STUDIES

Manufacturers and other entities must comply with the FDA's Good Laboratory Practices (GLP) regulations when conducting nonclinical laboratory studies that are going to be used to support any regulatory submission to the FDA (21 CFR Part 58). Good Laboratory Practices regulate the organization and personnel of the laboratory as well as the facilities, equipment, test operations and study protocols, and records and reporting. Failure to comply with these regulations may invalidate data submitted to the agency. Contract research organizations used to obtain data for regulatory submissions must comply with GLP regulations.

In addition, the study should conform to FDA-recognized standards that are relevant to particular aspects of the studies, for example, laser safety, toxicity, and biocompatibility. Also, the study's documentation should specifically identify and conform to those parts of FDA performance standards and guidance documents relevant to the device rather simply state overall compliance with the standard or guidance. Whenever particular laboratory study practices will not conform to relevant guidance, the manufacturer or study sponsor should, prior to conducting the studies, discuss the discrepancies with knowledgeable FDA staff, obtain a variance from the GLP regulations if necessary, and document the reasons for the discrepancies.

CLINICAL TRIALS

The FDA regulates clinical trials of medical devices under its investigational device provisions [21 USC § 360j(g), 21 CFR Part 812]. Also important are the regulations for institutional review boards [21 CFR Part 56] and the protection of human subjects [21 CFR Part 50], and the consolidated guidance for good clinical practice [ICH E6]. Different Part 812 procedures apply depending on whether the device study presents "significant risk" or "nonsignificant risk" (7). A significant risk device presents a potential for serious risk to the health, safety, or welfare of a subject. Significant risk devices can include implants, devices that support or sustain human life, and devices that are substantially important in diagnosing, curing, mitigating or treating disease, or in preventing impairment to human health. Examples include sutures, cardiac pacemakers, hydrocephalus shunts, and orthopedic implants. Nonsignificant risk devices are devices that do not pose a significant risk to human subjects. Examples include most daily-wear contact lenses and lens solutions, ultrasonic dental scalers, and urological catheters. Although these latter devices generally are nonsignificant risk devices, the FDA could consider a particular clinical trial using these devices to be a significant risk study and regulate the trial accordingly.

An institutional review board (IRB) may approve a nonsignificant risk device study, and the study may proceed without FDA approval. But clinical studies involving significant risks must receive FDA approval prior to IRB approval. Sponsors, usually investigators or manufacturers, apply for this FDA approval through submission of an Investigational Device Exemption (IDE) application. Although IRBs are to evaluate whether a study is a nonsignificant risk, the FDA has final authority and does determine, from time to time, that an FDA-approved IDE is needed even though an IRB approved a clinical trial protocol as being a nonsignificant risk study.

The FDA's IDE regulations set forth the requirements for submitting IDEs and conducting device clinical trials. These regulations are first and foremost designed to protect human subjects from unnecessary risk. In addition, the IDE regulations are designed to guide the development and documentation of evidence needed to evaluate a device's safety and effectiveness in a PMA application, or a device's substantial equivalence in a 510(k) submission. An IDE is a request for an exemption from the restriction that only legally marketed medical devices can be distributed.

The FDA has a pre-IDE meeting program that can be extremely valuable (8). These meetings usually include FDA review of some portions of a planned IDE submission. Pre-IDE meetings can be requested in a variety of circumstances, and are intended to provide the sponsor with preliminary FDA input related to the device. For example, the pre-IDE meeting should help clarify whether any additional preclinical or technical data are needed, what concerns FDA reviewers may have, whether the proposed protocols are adequate from the FDA's perspective, and the appropriate regulatory path to market for the device. Sponsors planning to conduct nonsignificant risk studies

sometimes request a pre-IDE meeting to whether deficiencies exist in the protocols that might preclude marketing approval. Other sponsors find it useful to discuss issues related to ongoing preclinical testing.

An IDE sponsor must submit a detailed description of the device, including its intended use and indication for use, that is, what does the device do and in what kind of patients or user. The sponsor must submit an investigational plan and a detailed protocol for the proposed clinical trial, including proposed informed consent documents. An IDE application also requires other documentation, including results from all laboratory and animal studies conducted with the medical device proposed for the clinical study. These laboratory and animal studies must be conducted in conformity with GLPs. The sponsor must report all relevant published studies, both nonclinical and clinical, regarding the device. Information on all medical uses of the device, and on any clinical trials conducted outside the United States may also be required. If consensus standards exist for the device, the sponsor must identify them and explain whether the device conforms with them. If previous clinical trials were conducted under IRB-only approval, then that data must be submitted in the IDE.

The IDE regulations include an IDE application template (21 CFR 812.20). The FDA also has issued a number of guidance documents on IDE processes and specific types of medical devices (3). Prior to submitting an IDE application to FDA, agency guidance documents relevant to the medical device at issue should be reviewed, and relevant aspects of those guidances implemented. These guidances often recommend specific preclinical tests for categories of devices and can include template investigational plans. But these recommendations and templates are not always suitable for particular devices. Also, guidances are not binding on the FDA and may not fully reflect current agency thinking. Consultation with the FDA may be appropriate where a sponsor believes that modifications are needed for its device.

Once a sponsor submits a complete IDE, FDA must make a decision regarding the IDE submission no later than 30 calendar days from the stamped date of arrival of the IDE application at CDRH headquarters. The FDA's initial decision letter usually lists deficiencies in the IDE, even when FDA approves the IDE. A disapproval letter is rare, especially if the sponsor had a pre-IDE meeting with the FDA. Sponsors receiving a disapproval letter may find it useful to seek assistance from an experience regulatory affairs professional to help evaluate and resolve these deficiencies. If the FDA conditionally approves the IDE, but with deficiencies that have major impact on the clinical trial or the device's indications, these deficiencies should be resolved with the FDA before the clinical trial is started. The FDA usually "conditionally" approves an IDE application, meaning that the applicant may start the clinical trial immediately, but that the applicant must answer the deficiencies satisfactorily within a short time period (e.g., 30–45 days). When the FDA perceives a high risk to human subjects, it will initially approve the IDE for a limited number of subjects and study sites, and then approve expansion of the study after the preliminary data demonstrates reasonable safety. The FDA also typically provides a list of deficiencies that do not have to be answered to conduct the clinical trial, but must be responded to in the marketing application [e.g., 510(k) or PMA]. If any aspect of the FDA's response letter is unclear, clarification should be sought from the FDA or an experienced regulatory affairs professional, or both.

Responsibilities of a clinical trial sponsor, include, but are not limited to, obtaining IRB approval, providing adequate informed consent, and ensuring that the investigators are trained and follow the approved protocol. Adequate record keeping, especially of adverse events, and study site monitoring are critical to success. Annual reports of the clinical study must be submitted to the FDA on the anniversary of the FDA's initial approval of the IDE. Also, serious adverse events must be reported to the FDA within five working days of their occurrence. All adverse events must be reported to the FDA even if the sponsor does not believe the event is related to use of the medical device being studied. Sponsors should also consult medical practice specialty standards and international standards that may be relevant to the study.

Although IDE sponsors (and their agents) may conduct limited advertising for subjects, they must not claim or suggest that the device is safe and effective for the uses it is being studied for. When discussing the device with potential investors, issuing reports on the company, and conducting similar activities, sponsors must carefully avoid making any conclusory statements regarding the device's safety and effectiveness. These restrictions continue until the FDA authorizes or approves the device for marketing. Sponsors also may not charge subjects, investigators, hospitals, or other entities a price for the device that is larger than that necessary to recover costs for manufacture, research, development, and handling. These costs should be documented in the event of an FDA inspection or audit.

Although clinical investigations of medical devices generally must comply with IDE requirements, some limited exemptions exist. For example, a diagnostic device that is noninvasive, does not require an invasive sampling procedure that poses significant risk to the subject, does not introduce energy into a subject, is not used as a diagnostic procedure without confirmation by another medically established diagnostic device, and meets certain other requirements, is exempt from IDE requirements. But the study must still comply with IRB and informed consent requirements.

REGULATORY PATHWAYS TO MARKET

Some medical devices require clearance through premarket "510(k)" notification, some medical devices require premarket approval, and others are exempt from premarket notification and premarket review. The majority of devices—more than 75%—have entered the market through 510(k) premarket notification.

Premarket notification is a process under which the FDA decides whether the evidence demonstrates substantial equivalence between a new device and a legally marketed (predicate) device. If the FDA decides that the device is substantially equivalent to the predicate device, then the

device is "cleared" for market. If the FDA decides that the device is not substantially equivalent, it is sometimes appropriate for a manufacturer to request *de novo* classification into Class I or II based on the device's low potential risks. But if the FDA denies that request, the only pathway to market is the PMA approval process. Typically, the FDA will determine, in discussions with a manufacturer or sponsor, which of the three pathways to market is required: (1) 510(k) → substantial equivalence; (2) 510(k) → nonequivalence → *de novo*; (3) PMA. But, as noted, some medical devices are exempt from even 510(k) requirements.

The Medical Device User Fee and Modernization Act of 2002 (MDUFMA) authorizes user fees for premarket reviews of PMAs, PDPs, certain supplements, 510(k)s, and certain other submissions (21 USC §§ 379i-379j). The MDUFMA also set agency performance goals for many types of premarket reviews. These goals become more demanding on the FDA over time. User fees must be paid at the time a submission is sent to the agency or the agency will not file or review it. The MDUFMA includes some fee exemption, waiver, and reduction provisions, including a fee waiver for the first premarket application by a small business.

PREMARKET NOTIFICATION EXEMPTIONS

Class I medical devices are exempt from 510(k) notification unless the FDA has by regulation stated that a particular medical device type is not exempt, or has specified conditions under which it is exempt. But the exemption applies only where the device is intended and indicated for the use or uses specified in the applicable regulation. If the device is to be marketed for a different use or medical condition, then the device is not exempt from 510(k) notification. If the new use presents extremely high risks or involves particularly vulnerable patients, a PMA may be required instead of a 510(k).

The same basic exemption rules apply to Class II devices, except that few Class II devices are exempt from premarket review. For devices exempt from premarket review by regulation, some changes in uses or indications do not require premarket review of the device because certain uses or indications are sufficiently similar to legally marketed intended uses. But in other instances, the FDA decides that an otherwise exempt device must receive premarket notification even though the uses or indications seem very similar. Although the FDCA provides a means for manufacturers to obtain a formal opinion from the FDA where uncertainty exists about the regulatory status of a device, an informal opinion may be sufficient, and preferable, in some situations. Manufacturers should consult an experienced regulatory affairs professional to evaluate how best to proceed in these circumstances.

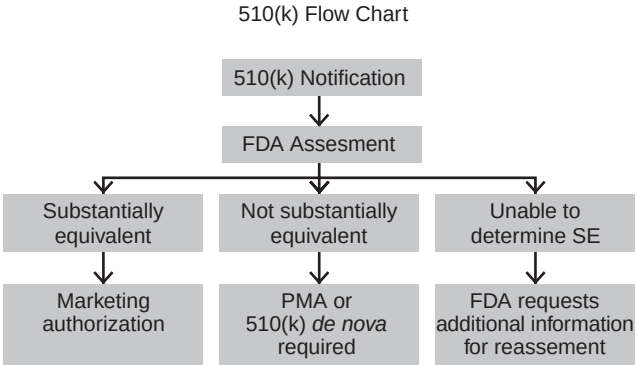
PREMARKET "510(K)" NOTIFICATION

A 510(k) submission → substantial equivalence decision requires a determination by the FDA that

1. The intended use of the sponsor's device is the same as that of the predicate device(s). Predicate devices may be any Class I or II device with the same intended use. (A limited number of Class III devices marketed before 1976 also can be predicate devices if the FDA has not yet called for a PMA.)
2. The technological characteristics of the sponsor's device must be either
   (a) The same as the predicate device.
   (b) Have performance characteristics that demonstrate that it is as safe and effective as the predicate device.

A substantially equivalent device is not "approved" for market. Instead, a 510(k) "clearance" decision is based on the FDA's evaluation of whether the device is substantially equivalent to a legally marketed device for which a reasonable assurance of safety and effectiveness exists. "Substantial equivalence" is a term of art, and does not require that a sponsor's device look or even operate the same as a predicate device. Two devices that visually appear dissimilar can be substantially equivalent under the FDCA. For example, the FDA cleared laser-light and water-jet microkeratomes as equivalent to vibrating steel blades to cut the cornea even though the former products use completely different cutting mechanisms than the latter.

The FDA has issued many guidance documents on various medical device types requiring premarket notification (3). The agency also has issued guidance documents for the 510(k) notification process. The FDA will provide prenotification consultation in telephone or in-person conferences to discuss a sponsor's medical device and answer questions regarding written guidance documents and applicable standards. The FDCA requires the FDA to consider, in consultation with a sponsor, the "least burdensome", appropriate means of evaluating a device (9). To maximize this requirement, a sponsor should understand, as much as possible, the requirements, guidances, and standards that apply to its medical device before meeting with FDA staff. As noted, guidances do not "bind" the FDA. But they can provide valuable information on the agency's thinking on particular topics. Also, a sponsor should try to understand how similar devices have been regulated by the FDA.



The 510(k) submission → nonequivalence → *de novo* process use the same 510(k) processes to try to establish that substantial equivalence exists and obtain FDA

clearance for marketing. But when the sponsor is unable to do so, despite thorough efforts to do so, then the objective becomes convincing the FDA that a PMA is not necessary for regulatory control of the device. This requires a showing that the risks from the device are minimal, that the device is effective for its intended use, and that general controls and, in some cases, special controls will be sufficient to mitigate the product's risks. A request for *de novo* classification must be made within 30 days of receiving a not substantially equivalent determination, describe the device in detail, and provide a detailed recommendation for classification. The FDA then has 60 days to respond to that request with a written order classifying the device and identifying any special controls that may be needed if the device is in Class II. The device is then considered cleared and may be marketed. If the FDA keeps the device in Class III, PMA approval will be required before marketing.

MARKETING APPROVAL

Class III medical devices generally are high risk devices that cannot be regulated adequately by general and special controls alone. In other words, the FDA must review the safety and effectiveness data for these devices to determine if they should be approved for the treatment or diagnosis of diseases or other conditions in humans, and under what conditions. Class III devices may be approved for marketing under the humanitarian use device exemption (HDE), product development protocol (PDP), or premarket approval application (PMA) requirements.

HUMANITARIAN USE DEVICES

The FDCA's humanitarian use device exemption provision is narrow in that the objective is to provide rapid access to new therapeutic or diagnostic devices for patients with rare diseases or conditions, that is, so-called "orphan" devices (21 USC § 360j(m), 21 CFR Part 814, Subpart H). The humanitarian use device (HUD) process is relatively rapid because the applicant does not have to conduct clinical trials to demonstrate reasonable assurance of safety and effective, and the statute allows the FDA significantly less time to act on an HUD application than the agency has for a PMA. Rather than provide data to determine the safety and effectiveness of the device, the applicant has only to satisfactorily explain to the FDA why the probable benefit of the device outweighs the risks to patients in the context of other treatments for the disease. However, this regulatory pathway has many requirements, including the disease or condition affects fewer than 4000 patients/year, the device would not otherwise be available for persons with this disease or condition, the device and will not expose patients to unreasonable or significant risks, and the benefits to health from the device's use must outweigh the risk. Because of the provision's narrow scope and limitations, the humanitarian use device exemption is not used frequently. But it can be very valuable in some instances.

PRODUCT DEVELOPMENT PROTOCOL

The product development protocol (PDP) is an alternative to the PMA process, but is rarely used [21 USC § 360e(f)]. The PDP's distinguishing feature is that it involves a close relationship between the FDA and the sponsor in designing appropriate preclinical and clinical investigations to establish the safety and effectiveness of a device. The PDP requires multiple levels of review and approval of study protocols. The requirements for proof of safety and effectiveness are the same as for a PMA. The PDP process thus offers few advantages for a manufacturer over premarket approval processes, particularly for a device that has undergone significant evaluation and investigation. The PDPs also have required much more FDA staff time than PMA processes.

PREMARKET APPROVAL (PMA)

The FDCA's requirements for PMA approval apply to most Class III medical devices, except for a few devices marketed before the 1976 MDA and those being used consistent with an investigational device exemption (IDE) in order to obtain clinical data to establish the device's safety and effectiveness (21 USC § 360e). The FDA has promulgated regulations on PMA requirements and processes (21 CFR Part 814). These regulations include the FDA's procedures for reviewing and acting on a PMA application. Other important sources for information on PMA issues include general guidances, guidances for specific devices, meetings with the agency and advisory panels, and correspondence from the agency.

The regulations specify and describe the general categories of required information in a PMA [21 CFR 814.20(b)]. These categories include an "indication for use" statement, a device description, and data from nonclinical and clinical studies of the device. The foreign and U.S. marketing history, if any, of the device by the applicant or others must be described in the PMA, including a list of countries in the device has been withdrawn from marketing.

INDICATION FOR USE

A PMA's "indication for use" statement must provide a general description of "the disease or condition the device will diagnose, treat, prevent, cure, or mitigate" and "the patient population for which the device is intended" [21 CFR 814.20(b)(3)]. The "indication for use" statement is key to the device's labeling and, if the device is approved, the uses for which it can be legally marketed. In addition to this statement, the application must include a separate description of existing alternative procedures and practices for the indicated use.

DEVICE DESCRIPTION

The device must be described in summary form and then in detail, including manufacturing and trade secret

information where necessary, to allow FDA specialists to evaluate the risks associated with the device. The summary must explain "how the device functions, the basic scientific concepts that form the basis for the device, and the significant physical and performance characteristics of the device" [21 CFR 814.20(b)(3)]. The full device description must include detailed drawings, and details of each functional component or ingredient of the device, all properties of the device relevant to the indication for use, the scientific and technical principles of operation of the device, and the quality control methods (good manufacturing practices) used in the manufacture, processing, packing, storage, and installation of the device [21 CFR 814.20(b)(4)]. In addition, the applicant must reference any standard, mandatory or voluntary, that is relevant to the device for the indicated use. If applicable, the applicant must identify how the device deviates from the standard and demonstrate, to the FDA's satisfaction, how the applicant resolves these deviations.

NONCLINICAL STUDIES

A PMA must include summaries of nonclinical laboratory studies appropriate to the device, including, but not limited to, microbiological, toxicological, immunological, biocompatibility, stress, wear, shelf life studies. The PMA must also include a statement that each study was conducted in accordance with the FDA's good laboratory practices regulations, or explanations as to why not. The study summaries must include descriptions of the objectives, experimental design, data collection and analysis, and results of each study. The results should be described as positive, negative, or inconclusive with regard to the objectives of each study. After each of the studies is summarized, it must be described in sufficient detail to enable the FDA to determine the adequacy of the information for FDA review of the PMA.

CLINICAL STUDIES

Clinical studies involving human subjects with the device must be conducted in accordance with IDE regulations or, if they are conducted outside the United States without an FDA-approved IDE, they must be conducted in accordance with special requirements discussed with the FDA before the PMA is submitted. IRB and human subjects protection requirements and the ICH guidance for good clinical practice also apply. The results of these clinical studies must be summarized first and then discussed in sufficient detail to enable the FDA to determine the adequacy of the information for FDA approval of the PMA. Clinical trial summaries must include the following:

> "...a discussion of subject selection and exclusion criteria, study population, study period, safety and effectiveness data, adverse reactions and complications, patient discontinuation, patient complaints, device failures and replacements, results of statistical analyses of the clinical investigations, contraindications and precautions for use of the device, and other information from the clinical investigations as appropriate...." [21 CFR 814.20(b)(3)].

Discussion of the results of the clinical investigations must include details regarding:

> "...the clinical protocols, number of investigators and subjects per investigator, subject selection and exclusion criteria, study population, study period, safety and effectiveness data, adverse reactions and complications, patient discontinuation, patient complaints, device failures and replacements, tabulations of data from all individual subject report forms and copies of such forms for each subject who died during a clinical investigation or who did not complete the investigation, results of statistical analyses of the clinical investigations, device failures and replacements, contraindications and precautions for use of the device, and any other appropriate information from the clinical investigations...." [21 CFR 814.20(b)(6)].

The applicant must identify any investigation conducted under an FDA-approved IDE and provide a written statement with respect to compliance with IRB requirement, or explain the noncompliance. In addition to submitting the data for all the studies conducted by the applicant (or on the applicant's behalf), the applicant is responsible for submitting a bibliography of all studies (nonclinical as well as clinical) relevant to the device and copies of any studies requested by the FDA or the advisory panel. Also, the applicant must identify, discuss, and analyze:

> "...any other data, information, or report relevant to an evaluation of the safety and effectiveness of the device known to or that should reasonably be known to the applicant from any source, foreign or domestic, including information derived from investigations other than those proposed in the application and from commercial marketing experience." [21 CFR 814.20(b)(8)].

LABELING

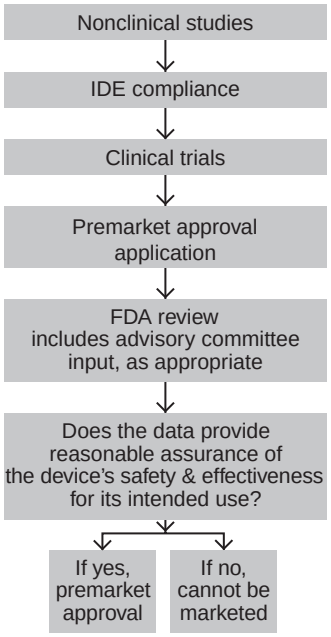
The applicant must submit copies of all proposed labeling for the device, including contraindications, warnings, precautions, and adverse reactions. Labeling typically includes, but is not limited to, physician instructions, an operation manual, a patient brochure, and all applicable information, literature, or advertising materials that constitutes labeling [21 CFR 814.20(b)(10)]. The FDA reviews and revises the proposed labeling prior to PMA approval.

REVIEW STANDARD

The applicant must demonstrate that the nonclinical, clinical, and technical data submitted in the PMA embody valid scientific evidence of reasonable assurance that the device is safe and effective for its intended use. In addition,

the applicant must discuss the benefits and risks (including any adverse effects) of the device, and describe any additional studies or surveillance the applicant intends to conduct following approval of the PMA. In evaluating safety and effectiveness, the FDA defines "valid scientific evidence" broadly and retains final authority on what is acceptable [21 CFR 860.7(c)]. The agency considers a variety of factors in deciding whether reasonable assurance of safety and effectiveness has been submitted for a medical device, including intended use (indication), use conditions, benefit–risk considerations, device reliability, and generally requires well-controlled clinical investigations (21 CFR 860.7).

PMA flow chart



UPDATE REPORT REQUIREMENTS

While the FDA is reviewing a PMA application, the applicant must update it. Such updates are required 3 months after the PMA filing date, following the applicant's receipt of an FDA letter stating the PMA is "approvable", and at any other time as requested by the FDA. An "approvable" letter is a decision by the FDA that the PMA will be approved after the applicant resolves minor deficiencies.

After a device is approved, periodic and other reports are required. The owner of an FDA-approved PMA device is responsible for periodically updating any safety and effectiveness information on the device that may reasonably affect the FDA's evaluation of the device's safety or effectiveness, or that may reasonably affect statements of contraindications, warnings, precautions, and adverse reactions. If a PMA owner becomes aware of off-label (unapproved) uses of its device that may be unsafe or ineffective, then it is responsible for reporting these unauthorized uses to the FDA, especially if adverse events are associated with them.

POSTMARKET RULES

Major postmarket requirements include adequate labeling, medical device reporting, corrections and removals, and device modifications integrated into a system for manufacturing quality medical devices.

LABELING OVERVIEW

Labeling of medical devices is one of a manufacturer's key postmarket responsibilities. Each device must comply with general labeling requirements, and with the specific requirements and limits identified in the FDA's authorization or approval to market the device.

LABELING: GENERAL REQUIREMENTS

Many general labeling requirements exist for medical devices (21 CFR Part 801, Subpart A). The regulations include details on issues such as how a manufacturer's name is to be listed on a device package label. This article focuses on key concepts in the FDA's regulation of device labeling. These concepts include the FDCA's definition of "labeling", the regulation's definition of "intended use", and "adequate directions for use" requirements.

Under the FDCA, "labeling means all labels and other written, printed, or graphic matter (1) upon any article or any of its containers or wrappers, or (2) accompanying such article" [21 USC § 321(m)]. This definition is very broad and includes promotional and advertising materials and oral statements about the device. The FDA's regulation of advertising and promotion presents many challenges for medical device companies. Three basic principles are critical: materials must be truthful and not misleading, must contain a fair balance of benefits and risks, and must provide full disclosure for use.

The "intended use" of a medical device is the objective intent of the product as expressed by the manufacturer or distributor of the device (21 CFR 801.4). It includes all conditions, uses, or purposes stated by the manufacturer or distributor orally or in written form. As discussed earlier, "intended use" is an integral part of the FDCA's "medical device" definition. If a manufacturer or distributor promotes an "intended use" different from the one authorized by the FDA, then the device is adulterated and misbranded until and unless the FDA authorizes the new use. This is often referred to as "off-label" use. The regulation further provides that "if a manufacturer knows, or has knowledge of facts that would give him notice that a device introduced into interstate commerce by him is to be used for conditions, purposes, or uses other than the ones for which he offers it, he is required to provide adequate labeling for such a device which accords with such other uses to the article is to be put." "Intended use" is how the manufacturer intends the device to be used. "Indication for use" is a subset of "intended use" that usually represents a narrowing of the intended use to a specific patient population. In short, why a patient, or a practitioner on a patient's behalf, would use a particular device. Indications for use include a general description of

the disease or condition the device will diagnose, treat, prevent, cure, or mitigate, including a description of the patient population for which the device is intended. If differences related to gender, race, ethnicity, age, or other factors exist, they should be reflected as well in the product's labeling.

The FDCA requires device labeling to bear "adequate directions for use," unless the FDA has promulgated regulations exempting a particular device [21 USC § 352(f)]. Under the regulations, "[a]dequate directions for use means directions under which the layman can use a device safely and for the purposes for which it is intended..." (21 CFR 801.5). These directions include specification of all applicable use conditions, dose quantity, use frequency, use duration, time of use, method of use, and use preparation. Adequate directions for use on over-the-counter devices must include a statement of indication for use [21 CFR 801.61(b)].

Prescription devices are exempt from the adequate directions for use requirement because, by definition, such directions cannot be prepared for a prescription device (21 CFR 801.109). However, prescription devices must have adequate instructions for the device's use by practitioners, including, but not limited to information on its use and indications, and any adverse events, contraindications, and side effects that may accompany the use of the device. In addition, to qualify for the adequate directions for use exemption for prescription devices, the device must meet other conditions, such as being in the possession of the practitioner. The regulations authorize other exemptions from the adequate directions for use requirement, including ones for medical devices that have common uses known to ordinary individuals, for medical devices used in certain teaching not involving clinical research, and for medical devices used in manufacturing, processing, and repacking (21 CFR Part 801, Subpart D).

LABELING: SPECIFIC DEVICES

Sources for labeling requirements for particular devices include labeling regulations for a few specific kinds of devices, classification regulations that provide "indications for use" statements for most Class I and Class II devices, guidance documents on specific devices, the FDA marketing authorization and approval letters, and approved labeling for PMA-approved devices.

The FDA has issued specific labeling regulations for dentures, eyeglasses and sunglasses, hearing aids, menstrual tampons, latex condoms, and devices that contain natural rubber (21 CFR Part 801, Subpart H). It also has specific labeling regulations for *in vitro* diagnostic devices (21 CFR Part 809, Subpart B). Each approved PMA includes labeling requirements for the device specified in the approval letter, in the summary of safety and effectiveness, and in written instructions for physicians (and other appropriate professionals) and patients.

REPORTING, CORRECTIONS, AND REMOVALS

Entities that manufacture, prepare, process, package, and/or distribute medical devices are subject to certain requirements regarding device reporting, corrections, and removals. They must track, document, investigate, take action on, and report on events associated with their medical devices. Device user facilities (e.g., hospitals) and importers of medical devices also have responsibilities for reporting certain medical device events (21 CFR Part 803, Subparts A-B and C-D). This article focuses on FDA reporting requirements for device manufacturers (21 CFR Part 803, Subparts A-B and E).

Device manufacturers must report medical device reportable (MDR) events to the FDA with five workdays of becoming aware of a reportable incident if remedial action to prevent an unreasonable risk of substantial harm to the public health, or the event is of a type that the FDA has designated as requiring a report within five work days. Otherwise, MDR events must be reported to the FDA within 30 calendar days. An MDR event is any information that a manufacturer becomes aware of that reasonably suggests that the device marketed by the manufacturer may have "caused or contributed to a death or serious injury" or "malfunctioned...and would be likely to contribute to a death or serious injury, if the malfunction were to recur" [21 CFR 803.3(r), 803.50(a)]. By "any information", the regulations mean all information in the manufacturer's possession or that the manufacturer could obtain from user facilities, distributors, initial reporters of the information, or by analysis, testing, or evaluation of the device. The FDA's regulations specify that manufacturers "become aware" of a reportable event when any employee and any manager or supervisor of employees with responsibility for MDR events acquires information reasonably suggesting that a reportable adverse event has occurred. Moreover, MDR events include any information that necessitates "remedial action to prevent an unreasonable risk of substantial harm to the public health", including, but not limited to, trend analysis [21 CFR 803.3(c)].

Manufacturers should be very inclusive of potential MDR reportable events because the regulations define "caused or contributed" factors very broadly to include events due to user error and labeling misunderstandings in addition to manufacturing and design problems, and device failure and malfunction. In addition, the regulations define malfunction to mean the failure of the device to meet performance specifications of the device for the labeled intended use of the device. "Remedial action" means "any action other than routine maintenance or servicing, of a device where such action is necessary to prevent recurrence of a reportable event" [21 CFR 803.3(z)]. For MDR purposes, the regulations define "serious injury" more broadly than a life-threatening illness or injury. Serious injuries are also those that produce permanent functional impairment, or damage to, body structure or that requires treatment to preclude such impairment [21 CFR 803.3(bb)].

Manufacturers should have written procedures in place to identify, evaluate, and document potential MDR reportable events so that reports can be submitted to the agency accurately and within the required timeframes. A manufacturer must maintain files and records of all events associated with its medical devices whether the manufacturer decided that such events were MDR reportable, and

the FDA must be given access to these records upon request. A manufacturer must maintain records of MDR reportable events for ready access by FDA inspectors, and also coordinate these files with the complaint files required by the FDA's Quality System Regulations.

In order to comply with MDR reporting and general record keeping requirements (21 CFR 803.17), a manufacturers must have a system of written procedures to identify, communicate, and evaluate events subject to MDR reporting requirements that is timely and effective; transmit medical device reports to the FDA that are complete and timely; document and record all information that was evaluated in determining if an event was MDR reportable, submitted to the FDA (including MDR reports), used in preparing semiannual reports or certifications to the FDA, and ensure that this documentation and record keeping is readily and promptly accessible to the FDA upon inspection.

Manufacturers also must submit reports to the FDA about medical devices that the manufacturer has corrected in, or removed from, the marketplace to reduce the risk to public health (21 CFR Part 806). A "corrected" medical device is one that the manufacturer has repaired, modified, destroyed, adjusted, relabeled, or inspected at the user location. This includes patient monitoring. A "removed" medical device is one that the manufacturer has physically moved from the user facility to repair, modify, destroy, adjust, relabel, or inspect. Corrections or removals do not have to be reported for devices that have not been distributed to users (stock recovery) or for routine maintenance. However, corrections or removals must be reported for "repairs of an unexpected nature, replacement of parts earlier than their normal life expectancy, or identical repairs or replacements of multiple units" of the device [21 CFR 806.2(k)]. The manufacturer must explain to the agency the reasons for, and estimate the risk to public health of, each correction and removal action within 10 days of initiating the action. The manufacturer must keep records of all corrections and removals, including those not reportable to the FDA, such as those for routine maintenance and stock recovery.

MODIFICATIONS TO MEDICAL DEVICES

Manufacturers must ensure that modifications to their marketed devices are made using design control requirements of the FDA's Quality System Regulations, including, but not limited to, verification and validation processes and updates of the design history file. Manufacturers should also have procedures in place to evaluate whether particular device modifications need to be reported to the FDA. All device modifications must be documented in the company's design and device history files. But some device modifications require prior approval by the agency, some require the opportunity for FDA disapproval prior to implementation, and still others may be reported after the company has implemented the changes. Because a medical device is defined as the physical apparatus and its intended use, significant changes to the product's intended use can require prior authorization from the FDA, even if no physical modification is made to the apparatus; the claim is only implied by the physical modification made to the apparatus; the manufacturer does not make the change but is aware that an entity to which the company sold the device is making additional substantial claims for the device. In other words, the FDA authorized manufacturer of a medical device can be responsible for the device it manufactures for the entire life cycle of the device.

The FDA's guidances for reporting device modifications for 510(k)-cleared devices and PMA-approved devices are summarized in Table 1 (10,11). Manufacturers should establish policies and principles for the company's medical devices based on these guidance documents and agency guidances specific to the company's devices.

QUALITY SYSTEM REGULATIONS

The two main objectives of the FDA's Quality System Regulations (QSR) (21 CFR Part 820) are to ensure (1) that quality in designed into medical devices, and (2) that management is responsible for the device throughout its life cycle and will be held accountable for shortcomings. The QSR sets forth the agency's current good manufacturing practices (cGMP) requirements for medical devices. Each manufacturer must integrate processes for controlling device modifications, labeling, and actions, reports, and record keeping regarding MDR events, corrections and removals into a quality system that is compliant with QSR. The QSR requires manufacturers to integrate all events associated with the manufacture and distribution of the medical devices into a corrective and preventive action (CAPA) subsystem linked to a record keeping subsystem that includes complaint files. The manufacturer must establish standard operating procedures that define, for example, the criteria for MDR reportable events for each kind of medical device that are manufactured, what actions are required, and the processes that must be followed. The manufacturer is responsible not only for maintaining complaint files and device history files, but for actively evaluating this information to maintain the medical device quality. This system involves using diverse information, including device maintenance, modifications, malfunctions, and failures with complaints from users, off-label (unapproved) use, and adverse reactions for continuous evaluation to ensure that the device is performing as designed. Corrective actions are to be taken as appropriate.

The corrective and preventive action subsystem is only one subsystem in a manufacturer's quality system. A quality system should be formed during the establishment of a company's management responsibilities and reviewed and revised during the initial design phase of device development. In addition to management and design control requirements, the QSR requires systems to control documents, purchasing, identification, traceability, production, processing, acceptance, nonconforming products, labeling, packaging, handling, storage, distribution, installation, servicing, and statistics. The regulations give a manufacturer the flexibility to develop a quality system for its medical devices that is tailored to the characteristics of these medical devices. However, the manufacturer's management team must justify and document the quality system, usually in

Table 1. Device Modification Reporting

Types of Modification	Premarket Notice [510(k)]	PMA Supplement
Changes due to recall or corrective action	Recall or corrective actions imply a safety or effectiveness problem with the device. Therefore, the FDA usually requires submission of a 510(k) notice if a device modification is necessary as part of the corrective action	Submit a "180-Day PMA Supplement" for design changes due to recall or corrective action even if the device still meets design specifications. Submit a "Special PMA Supplement-Changes Being Effected" for manufacturing changes that result from the corrective action
Changes that significantly affect safety or effectiveness	Use quality system, especially design controls, to determine if changes that significantly affect safety or effectiveness and if they do then submit a 510(k) notice	Submit a "180-Day PMA Supplement" for changes in, but not limited to, indications for use, labeling, new facilities, sterilization method, packaging, performance or design specifications, and the expiration date that affect safety or effectiveness. Use the quality system, especially design controls, to determine if changes affect safety or effectiveness. The FDA may issue a formal opinion that permits certain changes to be submitted in a "30-Day Supplement" rather than a "180-Day Supplement"
Labeling changes	Most, but not all, changes in intended use/ indication for use require submission of a 510(k) notice. For example, if the device will be indicated for use in a subset of patients for which the device is already cleared, then a 510(k) may not have to be submitted. Or no notice may be needed if a risk analysis demonstrates no additional risk by expanding the patient population being treated	Submit a "180-Day PMA Supplement"
Technology or performance specifications	Changes in a device's control mechanism, principles of operation, or energy source usually requires submission of a 510(k) notice. Changes in sterilization method usually do not require 510(k) notification if design verification and validation is adequate	Submit a "180-Day PMA Supplement"
Materials changes	Evaluate the effects of materials changes on the performance characteristics of the device. If the performance characteristics are changed significantly or new labeling must be added then perhaps a 510(k) notice should be submitted to the FDA	Submit a "180-Day PMA Supplement"
Minor incremental changes or changes that do not affect safety or effectiveness	Use design controls to evaluate risks associated with "minor" evolutionary changes in the device. Proactively develop a decision rule about when these incremental changes should be reported to the FDA	Usually does not require FDA approval prior to implementation but describe the modifications in the Annual Report required for the PMA
Minor changes to the manufacturing process	Notice to the FDA not required	File a "30-Day Notice" to the FDA describing the changes in detail. Implement the changes at the end of the 30-day period unless the changes require submission of a "135-Day Supplement" because the 30 day notice to the FDA was inadequate
Changes that improve the safety of the device	Notice to the FDA not required	File a clearly marked "Special PMA Supplement—Changes Being Effected." The changes that enhance safety include, but are not limited to, changes that strengthen a contraindication, an instruction, or quality controls. They must be described in detail

a quality system manual that specifies each of the subsystems as identified in the QSR and any deviations from it.

IMPLEMENTATION

Bioengineers and informed specialists developing innovative medical devices must understand the regulatory implications of their scientific and technical innovations in order to develop a realistic business plan for their product. Sometimes the innovations are considerable yet the agency regulatory pathways remain simple. For example, as discussed earlier, CDRH cleared laser-light and water-jet microkeratomes as equivalent to vibrating steel blades to cut the cornea even though the former products use completely different cutting mechanisms than the latter. Similarly, FDA decided that a manufacturer's microscopic dermal fragments should be regulated as human tissues under the same tissue bank rules used to regulate its macroscopic sheets of dermis. The FDA could have decided to regulate microscopic dermis as a medical device because of the additional processing (a decision that would have required requiring premarket authorization of the dermal fragments), but instead decided both were human tissues from a regulatory point of view. On the other hand, innovative products can be subject to profoundly different regulatory pathways. For example, external kidney dialysis products have almost always been regulated as medical devices by the CDRH using the 510(k) process, an efficient process. However, the FDA decided to use the drug–biologics review process (IND/NDA) to regulate an external kidney dialysis filter using human cells, a more complex and costly review process than for devices.

The following analysis of a hypothetical medical device illustrates some of the regulatory implications of innovative medical devices. The hypothetical device is an implanted artificial kidney that can continuously dialyze the human body. Currently, 90% of patients that require kidney dialysis are treated with an external device in which the patient's blood is dialyzed outside the body, an external kidney dialysis device. Some patients are treated with an external kidney dialysis device that infuses the dialysate (the dialysis solution) into the abdominal cavity (the peritoneum) and then drains the waste products out of the peritoneum ~45 min later or continuously overnight. As mentioned above, the FDA currently is regulating an external kidney dialysis product using more burdensome drug–biologic regulatory requirements rather than the simpler 510(k) process used for other dialysis machines. Therefore, if metabolic interaction and/or cells are used in an implanted artificial kidney devices, it is likely that either CDER or CBER will lead the review of the combination product through the FDA's drug–biologics approval process. However, if the implanted artificial kidney device achieved its intended use of dialysis without primarily biochemical or metabolic interaction with the human body, then the implanted artificial kidney likely would be regulated as a medical device by the CDRH. Filter material and microscopic control elements such as valves and motors are likely critical components. This example illustrates that issues imbedded in the scientific and technical characteristics of an innovative medical product could result in a regulatory pathway that is more complex, and costly, than already marketed alternative products.

Regardless of whether the innovative medical product, an implanted artificial kidney in this example, is reviewed by the FDA as a device, drug, or biologic, or a combination product, agency reviewers may or may not have expertise or knowledge directly relevant to the critical science. In fact, it is unlikely. Therefore, very early in product development the manufacturer should engage FDA reviewers in a dialogue about the cutting edge science or technology used in the device so that a common understanding evolves about key safety and effectiveness issues. This approach should help reduce misunderstandings about necessary nonclinical laboratory studies, animal study protocols, key safety and effectiveness endpoints, and fail-safe mechanisms so that agency reviewers will be comfortable with the risks associated with initial pilot study in humans. Also, the manufacturer should dialogue with agency reviewers about the scientific, clinical, and ethical issues associated with an initial clinical study in humans. In order to maximize control, manufacturers should take the initiative in making study proposals to the FDA rather than simply ask the agency for advice.

REGULATORY CHALLENGES

Medical device developers, academic researchers and engineers, start-up companies, research and development departments of large manufacturers, and other innovators are at the leading edge of scientific, technological, and medical product development, not the FDA. They therefore should be proactive with regard to the issues critical to the development and eventual marketing of the medical device. Developers of medical devices should take advantage of the opportunities to establish conditions for efficient FDA regulation of their devices before making regulatory submissions to the agency by developing a detailed quality system manual tailored to development and manufacture of the company's medical device; implementing good laboratory practices and specific standard operating procedures for nonclinical studies; identifying existing technical standards that are applicable to manufacturing quality devices, developing applicable standards where none exist; identifying the best clinical practices for clinical trials with the device; communicating the science and technology of the device to FDA reviewers; proposing a specific regulatory pathway to the agency based on a risk-benefit analysis of the device; incorporating feedback from discussions with the FDA. These proactive steps are particularly important for devices that are very innovative, and where scientific consensus may not exist on procedures and new standards needed to verify and validate the design of the device.

BIBLIOGRAPHY

1. Online access to the United States Code. Available at http://www.gpoaccess.gov/uscode/index.html. Accessed 2005 Feb 11.
2. Online access to U.S. Food and Drug Administration regulations. Available at http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm. Accessed 2005 Feb 11.

3. U.S. Food and Drug Administration Guidance Documents. Available online at http://www.fda.gov/cdrh. A searchable database of FDA guidances involving medical devices is available at http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfggp/search.cfm. Accessed 2005 Feb 11.
4. A searchable database of U.S. Food and Drug Administration Medical Device Classification Regulations. Available at http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPCD/PCDSimpleSearch.cfm. Accessed 2005 Feb 11.
5. U.S. Food and Drug Administration, Center for Devices and Radiological Health (2001, June 20). Recognition and Use of Consensus Standards; Final Guidance for Industry and FDA Staff. [Online version]. USFDA. http://www.fda.gov/cdrh/ost/guidance/321.html. Accessed 2005 Feb 11.
6. Standards recognized by the U.S. Food and Drug Administration. Available in a searchable online database at http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfStandards/search.cfm. Accessed 2005 Feb 11.
7. U.S. Food and Drug Administration, Office of the Commissioner (2001, April 18). Information Sheets: Guidance for Guidance for Institutional Review Boards and Clinical Investigators: Medical Devices. [Online] USFDA. Available at http://www.fda.gov/oc/ohrt/irbs/devices.html. Accessed 2005 Feb 11.
8. U.S. Food and Drug Administration, Center for Devices and Radiological Health. (1999, March 25). IDE Guidance Memorandum-Pre-IDE Program: Issues and Answers. [Online version]. USFDA. Available at http://www.fda.gov/cdrh/ode/d99-1.html. Accessed 2005 Feb 11.
9. U.S. Food and Drug Administration. Center for Devices and Radiological Health. (2002, October 4). The Least Burdensome Provisions of the FDA Modernization Act of 1997: Concept and Principles; Final Guidance for FDA and Industry. [Online version]. USFDA. Available at http://www.fda.gov/cdrh/ode/guidance/1332.html. Accessed 2005 Feb 11.
10. U.S. Food and Drug Administration, Center for Devices and Radiological Health. (1997, Jan 10). Deciding When to Submit a 510(k) for a Change to an Existing Device. [Online version]. USFDA. Available at http://www.fda.gov/cdrh/ode/510kmod.html. Accessed 2005 Feb 11.
11. U.S. Food and Drug Administration, Center for Devices and Radiological Health. (1998, Feb 19). 30-Day Notices and 135-Day PMA Supplements for Manufacturing Method or Process Changes. [Online version]. USFDA. Available at http://www.fda.gov/cdrh/modact/daypmasp.html. Accessed 2005 Feb 11.

See also CODES AND REGULATIONS: RADIATION; HOME HEALTH CARE DEVICES; HUMAN FACTORS IN MEDICAL DEVICES; SAFETY PROGRAM, HOSPITAL.

CODES AND REGULATIONS: RADIATION

BRUCE THOMADSEN
GLENN GLASGOW
BENJAMIN EDWARDS
RALPH LIETO
University of Wisconsin-Madison
Madison, Wisconsin

INTRODUCTION

Every country develops its own regulations governing radiation. Because this text is coming from the United States, the regulations considered here mostly apply to that country. "Radiation" in this article always refers to electromagnetic radiation and to energies higher than that used for communications (radio and microwave): particularly to higher-energy, directly and indirectly ionizing radiation [referred to as ionizing radiation (IR) hereafter] and medium-energy, nonionizing radiation (NIR). The discovery of ionizing radiations (i.e., those forms of radiation with sufficient energy to directly or indirectly *ionize* atoms by stripping away one or more electrons, thereby producing an ion pair consisting of the freed electron and charged atom) at the end of the nineteenth century (c. 1895) was followed quickly by observations of radiation injury.The first *recommendations* on IR dose limitation and personnel protection appeared shortly thereafter. The first general *regulations* for ionizing radiation came with the advent of the program to develop nuclear weapons during World War II. Most of this article deals with IR only because those regulations are more complex and voluminous. In addition, because this Encyclopedia focuses on biomedical applications, this text will concentrate most on those regulations most pertinent to medical settings, particularly those that have changed since the original edition of the Encyclopedia (1).

Although the electrical and magnetic fields associated with NIR were well known long before the discovery of ionizing radiation, a lack of significant observable health effects and the scarcity of powerful NIR sources delayed the development of NIR exposure standards until much later. The development of NIR standards was further complicated by the very wide range of wavelengths and photon energies covered by the NIR designation, and by the consequently wide variety of NIR tissue interaction mechanisms associate with each spectral region. Despite these obstacles, a comprehensive framework of NIR safety guidance now exists, but generally with less regulatory rigor and compulsion than for IR. Efforts to harmonize the exposure limits offered by various standard setting organizations have improved consistency, although disparities remain in some spectral regions.

ORGANIZATIONS INVOLVED IN IONIZING RADIATION PROTECTION RECOMMENDATIONS AND REGULATIONS

Sources of Guidance

The U.S. government relies on guidance from scientific organizations in the development of regulations. None of these organizations have any regulatory authority in the United States, but supply information and recommendations for the regulation-making processes. The most important organizations include the following:

- International Commission on Radiation Protection (ICRP): an international organization founded in 1928 under the International Congresses of Radiology (currently called the International Society of Radiology) that occasionally establishes panels to review the published literature on an issue concerning radiation protection and make recommendations.
- International Commission on Radiation Units and Measurements (ICRU): An international organization

organized in 1925 also under the International Congresses of Radiology that, like the ICRP, occasionally establishes panels to review the published literature on an issue concerning radiation units, measurement or dosimetry, and make recommendations.

National Council on Radiation Protection and Measurement (NCRP): A committee organized in 1929 as an informal gathering of radiation scientists to represent radiation-related organizations in the United States, and then formally chartered by Congress in 1964. As with the two international commissions, the NCRP establishes panels and writes reports on radiation related topic, and serves as the main source for guidance to the US government in the formulation of radiation regulations.

International Atomic Energy Agency (IAEA): An agency of the United Nations, the IAEA provides guidance documents and expert consultation on radiation safety issues, particularly for developing countries.

United Nations Committee on the Effects of Atomic Radiation (UNSCEAR): A committee under the United Nations established in 1955 to study the biological effects of radiation. Periodically this committee publishes report on their findings.

National Academy/Board on Radiation Effects Research (BRER): The BRER was established in 1981 to coordinate activities of the National Research Council involving the biological effects of radiation. Periodically, the BRER establishes panels to review the literature on the Biological Effects of Ionizing Radiation (BEIR) and issue reports bearing that acronym.

Joint Commission on Accreditation of Healthcare Organizations (JCAHO): A commission established by many medical organizations, such as the American Hospital Association and the American Medical Association. The Joint Commission establishes some standards for the use of radioactive materials and radiation in medical settings. Their standards, as of this writing, tend to be broad and vague statements on quality.

Professional Organizations: Organizations of professionals that may make recommendations, guidance documents or standards for various aspects of their profession. Often these documents form the basis for regulations. Some of the major organizations that influence radiation regulations include: The American Association of Physicists in Medicine; The American College of Interventional Cardiologist/The American College of Cardiology; The American College of Medical Physics; The American College of Nuclear Physicians; The American College of Radiology; The American Nuclear Society; The Health Physics Society/American Academy of Health Physics.

International Electrotechnical Commission (IEC)/ American National Standards Institute (ANSI): organizations that establish standards mostly pertaining to industry and manufacturers, their recommendations sometimes find their way into U.S. regulations aimed toward manufacturers of radiation-producing equipment.

Divisions of the U.S. Government Regulating Ionizing Radiation

In the United States, no one governmental agency regulates radiation and radioactive materials. Rather, aspects of radiation regulation fall under several agencies. Some of the major agencies are listed below, although the list is not exhaustive.

Nuclear Regulatory Commission. The Nuclear Regulatory Commission(NRC) is headed by a five-member Commission appointed by the President. The authority for the NRC comes from the Atomic Energy Act of 1954 (as the Atomic Energy Commission), as amended. The NRC was established by the Energy Reorganization Act of 1974. Because of the historical development of radiation regulations, the NRC formerly only exercised control over reactors and reactor byproduct materials. Thus, naturally occurring radioactive material, radioactive materials produced in particle accelerators and machine produced radiation fell outside the purview of the NRC. By these acts, the NRC regulates:

Special nuclear material, which is uranium-233, or uranium-235, enriched uranium, or plutonium.

Source material, which is natural uranium or thorium or depleted uranium that is not suitable for use as reactor fuel.

Byproduct material, which is, generally, nuclear material (other than special nuclear material) that is produced or made radioactive in a nuclear reactor.

Most recently, the Energy Policy Act of 2005 extended NRC authority to include naturally occurring and accelerator-produced radioactive materials (NARM). Before this time, the individual States regulated NARM with a somewhat non-uniform array of regulations.

The relevant NRC rules governing the authorized use of radioactive materials for medical applications are found in Title 10 Code of Federal Regulations. The specific divisions of Title 10 with a significant impact on medical uses are the regulations in Part 19—Notices, instructions and reports to workers: inspection and investigations; Part 20—Standards for protection against radiation; Part 21—Reporting of defects and noncompliance; Part 30—Rules of general applicability to domestic licensing of byproduct material; Part 31— General domestic licenses for byproduct material; Part 32— Specific domestic licenses to manufacture or trade certain items containing byproduct material; Part 33— Specific domestic licenses of broad scope for byproduct material; Part 35— Medical use of byproduct material; Part 71— Packaging and transportation of radioactive material.

The rules in Part 19, Part 20, and, most of all, Part 35 dominate the daily activity of medical licensees (2–5).

Since the late 1990s, NRC regulation changes have been performance-based rather than risk-based only. This was largely in response to the wide criticism by the medical

community of regulations and enforcement activity. This resulted in an Institute of Medicine–National Academy of Science report (6) that made several recommendations for improvement to the agency, and the subsequent NRC Strategic Assessment and Rebaselining Initiative. These initiated a major revision of the medical use rules of Part 35 (2). The NRC regulations attempt to protect workers and patients while minimizing its imposition on the practice of medicine. The last major change was completed in March 2005 that addressed training and experience of users, which demonstrates the lengthy federal rulemaking process (5).

Department of Transportation. The Department of Transportation (DOT) regulates (in Title 49 of the Code of Federal Regulations) shipping or carrying radioactive materials, be it by air or surface, including any radioactive materials on public streets or highways.

Environmental Protection Agency. The Environmental Protection Agency (EPA) regulates the allowed levels of radioactive materials in the air, water, and landfills, as well as radiation exposures to the public outside nuclear power reactors. In some cases their regulations also covers occupational exposures to radiation. The rules enforced by the EPA do not all come from a single section of the Code of Federal Regulations.

Department of Energy. The Department of Energy (DOE) is charged with leading the energy development in the United States. A large part of their work involves reactors, and for DOE funded projects and facilities, particular radiation regulations apply.

Department of Defense. The Department of Defense (DOD) establishes radiation regulations for DOD facilities.

Food and Drug Administration. Department of Health and Human Services (DHHS) enters into the radiation regulation field mostly through one of its 10 agencies, the Food and Drug Administration (FDA). The FDA is responsible for protecting the public health by assuring the safety, efficacy, and security of human and veterinary drugs, biological products, medical devices, our nation's food supply, cosmetics, and products that emit radiation, either ionizing or nonionizing. Accordingly, by Title 21 of the Code of Federal Regulations, Food and Drugs, Revised April 1, 2004, the FDA approves and regulates the testing, manufacture, and approved use of a radioactive drug, also called radiopharmaceutical, or a medical device containing a radioactive source. However, the radiation safety regulations of who is authorized to use such drugs or devices and the conditions for use are the responsibility of the NRC or its Agreement States. Regulations in Title 21 can be found on line at http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm [5 October 2005]

Radiopharmaceuticals. Radiopharmaceuticals are used for diagnostic purposes such uptake, dilution, or imaging, or for therapy applications. The relevant FDA regulations applicable to approving a radioactive drug are found in 21CFR 200–680 (References to the Code of Federal Regulations are usually written with the number of the Title before "CFR" followed by the part number, so this reference is Title 21 of the Code of Federal Regulations, Parts 200 through 680.):

Subchapter C—Drugs: General

- Part 201: Labeling
- Part 211: Current Good Manufacturing Practice for Finished Pharmaceuticals

Subchapter D—Drugs for Human Use

- Part 310: New Drugs
- Part 312: Investigational New Drug Application
- Part 361: Prescription Drugs for Human Use Generally Recognized as Safe and Effective and Not Misbranded: Drugs Used in Research

Subchapter F—Biologics

- Part 600: Biological Products: General
- Part 601: Licensing
- Part 610: General Biological Products Standards
- Part 660: Additional Standards for Diagnostic Substances For Laboratory Tests

A rapidly increasing aspect of nuclear medicine is the use of radioactive drugs employing positron emitters for diagnostic imaging. Positron emitters have a physical characteristic of very short half-lives (less than a few hours). The dominant radiopharmaceutical is F-18 tagged to fluorodeoxyglucose (FDG). They are used to perform positron emission tomography (PET). A problem with the production of PET drugs is meeting the FDA current good manufacturing practices (CGMP) regulation, which ensures that PET drug products meet safety, identity, strength, quality and purity requirements. The cause is their short half-lives prevent completing the current good manufacturing practices (CGMP) in a manner to allow distribution and administration. Current good manufacturing practices (CGMP) covers items such as control of ingredients used to make drugs, production procedures and controls, recordkeeping, quality system and product testing. The FDA and professional societies, such as the Society of Nuclear Medicine, are working to achieve a resolution. (http://www.fda.gov/cder/regulatory/pet/default.htm).

Machines and Devices. It is the responsibility of the FDA to determine if a submission is a device or a drug. With the increasing complexity and miniaturization of technology this is becoming increasingly difficult. Nevertheless, the FDA must approve any device that will be used on–in humans. Examples of such radioactive medical devices are high dose rate (HDR) remote afterloaders, intravascular brachytherapy devices, and radioactive-liquid filled balloons for the treatment of brain tumors. In addition, either the NRC or an Agreement State must perform engineering and radiation safety evaluations of the ability

of the device to safely contain radioactivity under the conditions of their possession and use. If deemed satisfactory, the regulatory authority issues a registration certificate. The evaluations are summarized in the registration that the NRC maintains in the National Sealed Source and Device Registry (NSSDR). The registration certificates contain detailed information on the sources and devices, such as how they are permitted to be distributed and possessed (specific license, general license, or exempt), design and function, radiation safety, and limitations on use. Either the NRC or Agreement States can issue a registration certificate for distributors and manufacturers within their jurisdiction, but only the NRC is responsible for devices distributed as exempt products (i.e., smoke detectors) and issues those registration certificates.

Analogous to drugs, the Radiation Control for Health and Safety Act of 1968 established the requirements and responsibility for the FDA to administer an electronic product radiation control program to protect the public health and safety. As part of that program, FDA has authority to issue regulations prescribing radiation safety performance standards for electronic products, most importantly including diagnostic X-ray systems. This gives the FDA the authority to promulgate regulations on the manufacture and assembly of such machines. Again, it is the individual state that regulates who can operate such machines and the radiation safety conditions of use. The exception to this is diagnostic mammography. For mammography, under the Mammography Quality Standards Act (MQSA) of 1992, the FDA approves the accrediting bodies that accredit the facilities to be eligible to perform screening or diagnostic mammography services. It also establishes minimum national quality standards for mammography facilities to ensure safe, reliable, and accurate mammography. These standards address the physician interpreters, the radiologic technologists performing the imaging, the medical physicists performing the testing, and machine performance and testing for mammography *only*. The Center for Devices and Radiological Health (CDRH) is the agency within the FDA that has responsibility for radiation machines and machines (http://www.fda.gov/cdrh/). The relevant FDA regulations applicable to the manufacture and performance of radiation machines are found in 21 CFR 900-1050:

Subchapter I—Mammography Quality Standards Act (MQSA)
Part 900-Mammography
Subchapter J-Radiological Health
Parts 1000–1050

For radioactive pharmaceuticals, implantable radioactive sources, radiation producing machines, or computer software that may be used with humans, approval must first be obtained by the vendor from the FDA. Before such approval, any use must be performed under an Investigational Drug Exemption (IDE) from a facility's Institutional Review Board (IRB), and if the drug or device poses significant risk, by the FDA also. After demonstration of the safety of the investigational drug or device, the FDA may approve general use following the manufacturer's instructions as given in the premarket approval (PMA) documentation. Use other than as described is considered "off-label," and, while allowed, imposes increased liability on the institution should something go wrong.

Occupational Safety and Health Administration. The Occupational Safety and Health Administration (OSHA) administers the Occupational Safety Health Act to assure safe and healthful working conditions. The health standard 29CFR 1910.1096 governs employee exposure to ionizing radiation from X-ray equipment, accelerators, accelerator-produced materials, electron microscopes, betatrons, and technology-enhanced naturally occurring radioactive materials not regulated by the NRC (7,8). The OSHA encourages states to develop and operate their own programs, which OSHA approves and monitors.

Their rules have not been revised since 1971 (9). and essentially reflect the NRC regulations at that time. At the time of writing, OSHA is considering revising its regulations.

States. Regulation of radioactive materials and radiation producing machines that are not covered by any federal rules fall to the individual states to regulate. However, the states often enter into agreements with federal agencies to assume the federal regulatory role. This is discussed in greater detail in the section on Regulatory Standards for Radioactive Byproduct Material.

Conference of Radiation Control Program Directors. There is one organization that needs to be noted especially with regard to the establishment of regulations by the individual states. This organization is the Conference of Radiation Control Program Directors (CRCPD). In the early 1960s many states were developing radiation control programs. Such programs included, but were not limited to, regulating the use of diagnostic and therapeutic X ray, environmental monitoring, and regulating the use of certain radioactive materials including NARM. Simultaneous to the development of these early state and local radiation control programs were similar activities at the federal level. Many of these and varied state, local, and federal programs and activities in radiation control were being developed independent of each other.

A need for uniformity was identified to avoid inconsistencies and conflicts of rules and regulations throughout the country regarding radiation users. As a result, the CRCPD was established in 1968 to (*1*) serve as a common forum for the many governmental radiation protection agencies to communicate with each other; and (*2*) promote uniform radiation protection regulations and activities. To achieve these purposes, the CRCPD developed the Suggested State Regulations (SSR) for radiation control, which it regularly updates as federal or industry changes occur at its websitw (10). The SSR address both radioactive materials and radiation machines in medicine and industry. Although the SSR are only recommendations, their importance is that many states have, and continue to, adopt them as their state regulations giving them the force of law. These suggested rules are discussed in detail below. The

primary membership of CRCPD is radiation professionals in state and local government who regulate the use of radiation sources. But it works closely with all relevant federal agencies, (NRC, FDA, DOT, EPA, etc).

U.S. Divisions Regulating Nonionizing Radiation

The key U.S. government agencies involved in regulating NIR are listed in Table 1. U.S. regulatory guidance specifically addresses some kinds of NIR sources in some spectral regions while omitting direct mention of other sources and spectral regions. For example, Curtis (11) acknowledges that the exposure standards in the OSHA are dated, noting the following weaknesses: the construction industry standard does not include laser classification and controls; the radio frequency (RF) exposure limit is from the 1966 ANSI standard (it has no frequency dependence and does not address induced current limits); The RF Safety Program Elements are incomplete.

However, the obligation of employers under the General Duty Clause of OSHA [Occupational Safety and Health Act of 1970, 29 USC 654, section 5(a)(1)] to protect workers from recognized hazards compels the control of all potentially harmful NIR hazards, whether specifically regulated or not. Various government agencies also provide a wealth of guidance beyond the requirements specified in the regulations. As noted in Table 1, the FDA regulations apply primarily to manufacturers, so although much FDA guidance clearly pertains to the end users, the FDA typically does not inspect healthcare providers or enforce compliance with FDA guidance by healthcare facilities. However, other organizations that do routinely audit healthcare providers, including in particular the JCAHO, refer to and hold hospitals accountable for compliance with FDA guidance. Table 2 summarizes the requirements of those states having comprehensive laser safety regulations, adapted and updated from Ref. (12). Many of these states have also passed regulations for the control of other NIR hazards as well (see e.g. Article 14 in Chapter 1 of Title 12, Arizona Administrative Code). Several nonregulatory organizations have established exposure limits covering the entire NIR spectrum. The primary industry consensus

Table 1. U.S. Government Agency Nonionizing Radiation Regulations

Agency	Role	NIR Related Regulations/Guidance	Created by
FAA	Responsible for the safety of civil aviation; includes the safe and efficient use of navigable airspace	14cfr91.11: prohibits interference with aircrew FAA Order 7400.2 Part 6 Chapter 29 [Outdoor Laser Operations]; limits laser exposure levels near airports	Federal Aviation Act (1958)
FCC	Responsible for regulating interstate and international communications by radio, television, wire, satellite, and cable	To comply with the National Environmental Policy Act of 1969 (NEPA), established limits for Maximum Permissible Exposure (MPE) to RF radiation based on NCRP and ANSI/IEEE criteria, in 1996 Report and Order, and 1997 Second Memorandum Opinion and Order. In addition, per 47cfr18, industrial, scientific, and medical equipment manufacturers must comply with requirements designed to reduce electromagnetic interference	Communications Act (1934)
FDA and CDRH	Protecting the public health by assuring the safety, efficacy, and security of human and veterinary drugs, biological products, medical devices, our nation's food supply, cosmetics, and products that emit radiation	The following regulations apply primarily to manufacturers: 21cfr1040.10 and 11—laser products 21cfr1040.20—sunlamp products and ultraviolet lamps intended for use in sunlight products 21cfr1040.30—high intensity mercury vapor discharge lamps. 21cfr1030.10—microwave ovens	Food and Drugs Act (1906)
OSHA	Ensure the safety and health of America's workers by setting and enforcing standards; providing training, outreach, and education; establishing partnerships; and encouraging continual improvement in workplace safety and health	29cfr1910.97 nonionizing radiation 29cfr1910.268 telecommunications	Occupational Safety and Health Act (1970)

Table 2. Representative Sample of State Laser Regulations[a]

State	Regulation	Exemptions	Training Required	Warning Signs Required	Controls Required	Registration Required	ANSI or FDA Based	Outdoor or Light Show Requirements
AK	18AAC35, Art. 7, Sec. 670-730	Stored, Inoperable, Enclosed and below MPE (e.g., Class 1)	No	Yes	Yes	No	No	Yes
AZ	AAC Title 12, Chpt. 1, Article 14, Sec. R12-1-1421-1444	None, but focus on control of Class 3b and 4	Yes	Yes	Yes	Yes	ANSI/FDA	Yes
FL	FL Code: Chap. 64-E4	Stored, Class 1, 2, and 3a	Yes	Yes	Yes	Yes	ANSI/FDA	Yes
IL	Chapter 420 ILSC 56	Transported, negligible hazard	No	No	No	Yes	FDA	No
MA	105 CMR 121	Transit and storage	Yes	Yes	Yes	Yes	ANSI	Yes
NY	Title 12 NYCRR Part 50	Non-R&D Class 1, 2, and 3a	Yes	Yes	Yes	Yes	FDA	Yes
TX	Title 25 TAC Part 1 Rule 289.301	Transit, stored, inoperable	Yes	Yes	Yes	Yes	ANSI/FDA/IEC	Yes
WA	WAC 296-62-09005	None, but focus on control of Class 3b and 4	Yes	Yes	Yes	No	ANSI/FDA	No
HI	HAR 12-201-3	None	Yes	Yes	Yes	No	No	Yes

[a]Adapted and updated from Ref. 12.

standard organizations and international standard-setting agencies appear in Table 3. Some of these voluntary standards carry more weight than others, especially internationally. For example, all member countries of the European Union are required to adopt the laser safety standard, IEC/EN 60825-1, of the International Electrotechnical Commission (IEC), which has also been adopted by Japan, Australia, Canada, and nearly every other nation that publishes a laser standard (13). In addition, the FDA now accepts conformance with the IEC/EN 60825-1 in lieu of conformance with most (but not all) of the requirements of the U.S. Federal Laser Product Performance Standard (14). Similarly, the FDA, OSHA, and JCAHO all reference the ANSI Z136 series of standards.

REGULATORY STANDARDS FOR RADIOACTIVE BYPRODUCT MATERIAL

Use of IR in medical, dental, and veterinary facilities is governed by either federal (e.g., NRC, OSHA) or state regulations. The NRC, drawing its authority from the Atomic Energy Act of 1954, regulates byproduct material, source material, and special nuclear material, and their uses. Here OSHA controls IR sources (X-ray equipment, accelerators, accelerator-produced materials, electron microscopes, betatrons, and technology-enhanced naturally occurring radioactive materials) not covered by the Atomic Energy Act of 1954 and not regulated by the NRC. A 1989 "Memorandum of Understanding. . ." defined responsibilities and authorities of each agency (7). Each agency has arrangements with some states for regulatory enforcement. NRC has an Agreement State Program, by which a State can sign a formal agreement with the NRC to assume NRC regulatory authority and responsibility over certain byproduct, source, and small quantities of special nuclear material. There are 33 States, listed in Table 4, with two (Pennsylvania and Minnesota) in the process of becoming Agreement States. The Atomic Energy Act of 1954 provides a statutory basis under which NRC relinquishes to the states portions of its regulatory authority to license and regulate byproduct materials (radioisotopes); source materials (uranium and thorium); and certain quantities of special nuclear materials. The mechanism for the transfer of authority to a state is an agreement signed by the Governor of the State and the Chairman of the Commission. The NRC has established compatibility obligations with the Agreement State regarding its current rules and future regulations that it may promulgate. Because two-thirds of the states have assumed Agreement status, the NRC has provided them increasing voice in their activities. This is done through the NRC Office of Tribal and State Programs and the independent Organization of Agreement States (OAS). Both can be accessed via the URL, http://www.nrc.gov/what-we-do/state-tribal/agreement-states.html. The NRC regulations apply in federal facilities directly holding federal licenses and in the nonagreement states. Agreement states have certain periods (3 years or more) within which state regulations must become compliant, at certain levels of compliance, with NRC regulations. During this transition period state regulatory agencies enforce their current state regulations, based on NRC regulations in force prior to the regulatory changes, as they prepare new state regulations compliant with the recent revisions changes in federal codes. Twenty-six states, also in Table 4, have OSHA-approved state plans with their individual state standards and enforcement policies.

Table 3. Selected Organizations Publishing Voluntary NIR Safety Standards

Organization	Role	Significant NIR Standards
American Conference of Governmental Industrial Hygienists (ACGIH)	Professional society devoted to the administrative and technical aspects of occupational and environmental health	TLVs and BEIs (Threshold Limit Values for Chemical Substances and Physical Agents; Biological Exposure Indices)
American College of Radiology (ACR)	Maximize radiology value by advancing science, improving patient care quality, providing continuing education and conducting research	White Paper on MR Safety
American National Standards Institute (ANSI)	Promoting and facilitating voluntary consensus standards and conformity assessment systems	Z136.1—Safe Use of Lasers Z136.2—Safe Use of Optical Fiber Communication Systems Utilizing Laser Diode and LED Sources Z136.3 Safe Use of Lasers in Health Care Facilities Z136.5 Safe Use of Lasers in Educational Institutions Z136.6 Safe Use of Lasers Outdoors B11.21 Machine Tools Using Lasers—Safety Requirements for Design, Construction, Care and Use ANSI/IESNA RP-27.1 Recommended Practice for Photobiological Safety for Lamps and Lamp Systems—General Requirements; RP-27.3 Risk Group Classification and Labeling ANSI/IEEE 95.6 Safety Levels With Respect to Human Exposure to Electromagnetic Fields, 0–3 kHz ANSI/IEEE C95.1 Safety Levels with Respect to Human Exposure to Radio Frequency Electromagnetic Fields, 3 kHz–300 GHz
International Commission on Non-Ionizing Radiation Protection (ICNIRP)	Disseminate information and advice on the potential health hazards of exposure to nonionizing radiation to everyone with an interest in the subject	Guidelines on Limits of Exposure to Ultraviolet Radiation of Wavelengths Between 180 nm and 400 nm (Incoherent Optical Radiation) Guidelines on Limits of Exposure to Laser Radiation of Wavelengths between 180 nm and 1 mm Revision of the Guidelines on Limits of Exposure to Laser radiation of wavelengths between 400 nm and 1.4 μm Guidelines on Limits of Exposure to Broad-Band Incoherent Optical Radiation (0.38–3 μm) Guidelines for Limiting Exposure to Time-Varying Electric, Magnetic, and Electromagnetic Fields (up to 300 GHz) Guidelines on Limits of Exposure to Static Magnetic Fields
International Electrotechnical Commission (IEC); European Committee for Electrotechnical Standardization. (CENELEC)	Prepares and publishes international standards for all electrical, electronic and related technologies; these serve as a basis for national standardization and as references when drafting international tenders and contracts	60601-2-33: Particular Requirements for the Safety of Magnetic Resonance Equipment for Medical Diagnosis 60825-1: Equipment Classification, requirements, and user's guide 60825-2: Safety of Optical Fibre Communication Systems 60825-3: Guidance for laser displays and shows 60825-4: Laser guards 60825-5: Manufacturer's checklist for IEC 60825-1 60825-6: Safety of products with optical sources, exclusively used for visible information transmission to the human eye 60825-7: Safety of products emitting infrared optical radiation, exclusively used for wireless 'free air' data transmission and surveillance 60825-8: Guidelines for the safe use of medical laser equipment 60825-9: Compilation of maximum permissible exposure to incoherent optical radiation 60825-10: Application guidelines and explanatory notes to IEC 60825-1 60825-12: Safety of Free Space Optical Communication Systems used for the Transmission of Information TR60825-14: A User's Guide

In a few instances, the DOE operates research programs at national laboratories under DOE supervision and governs occupational exposure under 10CFR 835 (Occupational Radiation Protection). Because of its limited role, DOE regulations are not further discussed. Table 5 lists the web sites of these federal agencies and other organizations with interests in regulation of radiation in its many forms.

Table 4. NRC Agreement States[a] and States, Commonwealths, and Territories[b] with OSHA-Approved State Plans[c]

Alaska (O)	**Iowa** (A,O)		
Alabama (A)	Kansas (A)	***New Hampshire (A)***	Rhode Island (A)
Arizona(A,O)	**Kentucky**(A,O)	New Jersey (O)	**South Carolina** (A,O)
Arkansas (A)	Louisiana (A)	**New Mexico**(A,O)	**Tennessee**(A,O)
California(A,O)	Maine (A)	***New York***(A,O)[d]	Texas (A)
Colorado (A)	**Maryland**(A,O)	**North Carolina** (A,O)	**Utah**(A,O)
Connecticut (O)[d]	Massachusetts (A)	North Dakota (A)	Vermont (O)
Florida (A)	Michigan (O)	Ohio (A)	*Virgin Islands*(O)[d]
Georgia (A)	**Minnesota**(A[e],O)	Oklahoma (A)	Virginia (O)
Hawaii (O)	Mississippi (A)	**Oregon** (A,O)	**Washington** (A,O)
Illinois (A)	Nebraska (A)	Pennsylvania (A[e])	Wisconsin (A)
Indiana (O)	**Nevada** (A,O)	Puerto Rico (O)	Wyoming (O)

[a]Designated A.
[b]Designated O.
[c]Those that are both (A,O) are in bold print.
[d]These in italics have plans that cover public sector (State and local government) employment only.
[e]These are not yet agreement states, but have filed intent to become agreement states.

NRC Regulations—Summary of Changes to U.S. Regulations

The NRC has actively revised their governing regulations. The most recent revisions (4-24-02) were to four parts of the federal code: *10 CFR 19 (Notices, Instructions, and Reports to Workers; Inspections); 10 CFR20 (Standards for Protection Against Radiation); 10CFR32(Specific Domestic Licenses to Manufacture or Transfer Certain Items Containing Byproduct Material), and 10CFR35 (Medical Use of Byproduct Material)*(2–5). The Occupational Safety and Health Act enforces the terms of the OSHA promulgated in the federal code *29 CFR 1910.1096 (Ionizing Radiation)* which appear to date to 1970 (9). Indeed, current OSHA standards are based on original terms, definitions, and units historically used by the NRC in 1971, many of which were changed in the 2002 NRC revisions. While OSHA websites allude to potential code revisions under active internal review, none are posted for public comment. Hence, we focus on a synopsis of NRC revisions.

10 CFR 19 (Notices, Instructions, and Reports to Workers; Inspections). This long-standing regulation (2), with 14 sections, issued 12/18/1981, unrevised, remains in force. Table 6 lists seven important sections with brief comments about their content. Insuring that all current members of a constantly changing workforce receive initial and timely recurrent annual instruction is a significant regulatory compliance challenge for radiation safety officers (RSO) charged with their instruction.

10 CFR20 (Standards for Protection Against Radiation). These standards, consisting of 69 sections, are, with one exception, discussed later, mostly unchanged from the 5/21/1991 release (3). Tables 7a,7b lists 10 key headings with brief comments about their content.

Three sections, *10CFR20.1002/Scope; -0.1003/Definitions, and -.1301/Dose Limits for Individual Members of the Public* (4) were revised. 20.1002/Scope now states [conventional radiation units deleted] that "...limits in this part do not apply ...to exposures from individuals administered radioactive materials (RAM) and released under §35.75..." 20.1003/ Definitions adds "Occupational dose does not include...dose... dose... from individuals administered RAM and released under §35.75..." "Public dose does not include...dose... from individuals administered RAM and released under §35.75..."

20.1301/ Dose Limits for Individual Members of the Public now adds to the exclusion of dose from RAM in sanitary sewers, the following: "...does not exceed ... 1 mSv in a year

Table 5. Useful Web Sites with Information About Radiation, Regulations, and Regulatory Issues

Agency	Internet Address, http://www.	Electronic Mail Address
Conference Radiation Control Program Directors	crcpd.org	Not given on web page
Department of Energy	energy.gov	Not given on web page
Department of Transportation	dot.gov	dot.commentsost.dot.gov
Environmental Protection Agency	epa.gov	Not given on web page
Food and Drug Administration	fda.gov	Not given on web page
Health Physics Society	hps.org	hpsBurkInc.com
Idaho State University	physics.isu.edu/radinf/rso toolbox	Not given on web page
International Atomic Energy Agency	iaea.org	official.mailiaea.org
International Commission on Radiological Protection	crp.org	scient.secretaryicrp.org
International Commission Radiation Units and Measurements	icru.org	icruicru.org
National Council on Radiation Protection and Measurements	ncrp.com	Not given on web page
Nuclear Regulatory Commission	nrc.gov	Not given on web page
Occupational Safety and Health Administration	osha.gov	Numerous information-specific links on web page

Table 6. Partial Contents of 10 CFR 19[a]

Section Major	Major contents of section
0.3/Definitions	Workers, licenses, restricted areas defined
0.11/Postings notices to workers	(a) Post regulations, (i) license and its conditions; (ii) operating procedures; (iii) violations; (b) Documents, forms must be conspicuous
0.12/Instructions to workers	Inform about: (a) storage, use RAM; (b) health protection problems; (c) procedures to reduce exposures; (d) regulations; (e) report conditions, violations; (f) response to warnings; (g) their exposures
0.13/Notification and reports to individuals	(a) Written exposure reports; (b) annual exposure reports per workers request; (c) other provisions not stated here
0.14/Presence of licensee's and workers representatives during inspections	(a) Licensee to allow inspections; (b) inspectors may meet workers; (c) reps may accompany inspectors during inspections; (d) other provisions not stated here
0.15/Consultations with workers during inspections	(a) Inspectors may consult privately with workers; (b) workers may consult privately with inspectors
0.16/Requests by workers for inspections	Workers may request inspections without retribution

[a]Notices, Instructions, & Reports to Workers; Inspections.

exclusive of the dose contributions from background radiation, from any medical administration to the individual, from individuals administered RAM and released under §35.75, from voluntary participation in medical research programs. . ."

Also, added: ". . .a licensee may permit visitors to an individual . . .to receive a radiation dose greater than . . . 1 mSv if 1. the radiation dose . . .does not exceed . . . 5 mSv and 2. the authorized users has determined *before the visit* that it is appropriate."

Security of RAM is addressed in §20.1801. A new international and national concern is the security of byproduct sources in medical facilities. Most medical licensees have small (tenths of GBq) quantities of long-lived byproduct

Table 7a. Unchanged Components of CFR 20[a]

Section	Major contents of section
0.1101/Radiation Protection Program (RPP)	(a) RPP must be developed, documented, implemented, commensurate with extent and scope of licensed activities; (b) ALARA for occupational and public doses; (c) Annually review RPP content and implementation
0.120/Occupational Dose Limits Dose Equivalent (DE); Deep Dose Equivalent (DDE); Cumulative Dose Equivalent (CDE); Total Effective Dose Equivalent (TEDE)	(a) Annual TEDE 0.05 Sv; sum of DDE and CDE of organs 0.5 Sv; eye DE 0.15 Sv; shallow skin or extremity DE 0.5 Sv; (b) Excess DEs must be planned; (c) Other provisions not stated here
0.1208/Dose to an Embryo/Fetus	(a) 5 mSv dose to embryo/fetus, entire pregnancy, occupational exposure of mother; (b) Avoid variations in uniform monthly doses; (c) Dose is sum of DDE of mother and radionuclides in mother and embryo/fetus; (d) Other provisions not stated here
0.1502/Individual Monitoring of External/Internal Occupational Doses Cumulative Effective Dose Equivalent (CEDE)	(a) Those likely DE 10% of limits; (b) Those in high and very high radiation areas; (c) Those likely to receive CEDE of 10% from radionuclides; (d) Other provisions not stated here
0.1801/Security of radioactive materials	(a) Secure from unauthorized removal or access licensed material stored in controlled or unrestricted areas; (b) Licensed material not in storage shall have control and constant surveillance

[a]Standards for Protection Against Radiation.

Table 7b. Unchanged Components of CFR 20[a]

Section Major	Major contents of section
0.1901/Caution Signs	Radiation symbol (trefoil) color schema (magenta, purple, black) on yellow and design defined;
0.1904/Labeling Containers Radioactive Materials	(a) Containers of RAM must be marked either "CAUTION" or "DANGER", RADIOACTIVE MATERIAL; (b) Label must identify quantity, date, radiation levels, kind of material; (c) Remove/deface labels on empty containers
0.1906/Receiving/Opening Packages	(a) Package receipt and monitoring procedures; (b) Carrier notified if wipe test or radiation levels exceed limits; (c) Package opening procedures; (d) Other provisions not stated here
0.1501/Surveys and Monitoring	(a) Make necessary surveys; (b) Equipment used for surveys calibrated; (c) Excluding direct/indirect pocket dosimeters, NVLAP accreditation for badge processor
0.2001/Waste Disposal	(a) By transfer to authorized recipient; (b) By decay in storage; (c) By effluent release within limits; (d) Others provisions not stated here

[a]Standards . . .Protection . . .Radiation.

materials (^{137}Cs, ^{60}Co, etc.), ideal components for dispersal "dirty bomb". The IAEA has developed an action plan to combat nuclear terrorism (15,16). These international efforts likely will lead to new national and state regulations requiring greater security for radioactive sources.

Listed in Table 7b as unchanged is signage. Radiation areas and places or contains that hold radioactive materials must be posted as such. Fig. 1 shows a typical radiation area sign, and gives the criteria for each of the types of signs required.

The 10CFR32 (Specific Domestic Licenses to Manufacture or Transfer Certain Items Containing Byproduct Material) revisions (4-24-2002) are only notational bookkeeping, changing the paragraphs numbers and sections in Part 32 to correspond with the corresponding sections of the revised 10CFR35.

Figure 1. A typical "Caution: Radioactive Materials" sign. The wording on the sign follows the criteria: Rooms containing more than the quantity listed in Part 20 Appendix C–"CAUTION: RADIOACTIVE MATERIALS"; 2. Areas with exposure rates >0.05 mSv in 1 h, 30 cm from a source (or surface that radiation penetrates) – "CAUTION: RADIATION AREA" or "DANGER: RADIATION AREA" ; 3. Areas with exposure rates greater than 1 mSv in 1 h, 30 cm from a source (or surface that radiation penetrates) - "HIGH RADIATION AREA"; 4. Areas with exposure rates greater than 5 GY in 1 h, 1 METER from a source (or surface that radiation penetrates)– "GRAVE DANGER: RADIATION AREA"; 5. Areas where the derived air concentrations exceeds values in appendix B, to 20.1001–20.2401, or where an individual without respiratory protection could exceed, during the hours an individual is present in a week, an intake of 0.6% of the annual intake limit —"DANGER: AIRBORNE RADIOACTIVITY AREA".

Revisions in 10CFR 35 (Medical Use of Byproduct Material). With 126 sections, we focus only on those of direct interest or applicability to medical byproduct material. Tables 8a–e summarizes, using some shorthand notations, the major contents of the important sections. The bulk of regulatory changes relative to byproduct material occur in these sections.

Components of CFR 35 Applicable to All Forms of Brachytherapy (Tables 8a, b). A new term, Authorized Medical Physicist (AMP), and the training thereof, is defined, as well as types (low dose rate, LDR; pulsed dose rate, PDR; and high dose rate, HDR) of remote afterloading units (RAU), including medium dose rate (MDR). Mobile services and medical events are new additions. Roles of management, the RSO, and authorized users (AU) supervision of individuals are explained. Dose prescriptions, or written directives (WD) details and procedures are enumerated.

Table 8b notes source inventories are now at 6 month intervals. §35.75 explains new release criteria for patients (4). Some requirements for mobile medical services are in this section, as well as rules for decay-in-storage of RAM.

Some Components of CFR 35 (F) Applicable to Manual Brachytherapy (Table 8c). One major change is a requirement to decay output or source activities in 1% intervals. Another section adopts AAPM good practices, per various protocols, for quality assurance of therapy planning systems, as a regulation.

Some components of 10CFR 35 (H) for Photon-Emitting Remote Afterloaders (Tables 8d, e). In the nine sections, the most significant change is the requirements for MDR

Table 8a. Components of CFR 35 (A, B) Applicable to All Forms of Brachytherapy

Section	Major Contents of Section
0.2/Definitions	(a) Authorized medical physicist defined; (b) LDR, MDR, HDR, PDR defined; (c) Mobile medical service defined; (d) Medical event (no more misadministration's! explained; (e) Manual prescribed dose (total sources strength and time, or dose per WD) given; (f) Remote prescribed dose (total dose and dose per fraction per WD) given
0.24/Authority Radiation Protection Program	(a) Defines a stronger management role; (b) Defines and strengths RSO role
0.27/Supervision	Explains role of authorized user (AU) and supervised individuals with respect to process and procedures with RAM
0.40/Written Directives (WD)	(a) Written directives required or oral directives with 48 h for written; (b) HDR: radionuclide; site, fx dose, #fxs, total dose; (c) Others; before tmt: radionuclide; site, dose; before finish: # sources, total source strength and time (or total dose); revisions allowed during treatment.
0.41/Procedures. . .written directives	(a) ID patient; (b) administration per WD; (c) Check manual, computer dose calculations; (d) confirm console data
0.51/Training authorized medical physicist	(a) Board certifications; (b) degrees +1 year training + 1 y experience; (c) preceptor's written statement regarding training

Table 8b. Some Components of CFR 35 (C) Applicable to All Forms of Brachytherapy

Section	Major Components of Section
0.67/Requiremenst for possession	(a) Leak tests (5 nCi) before 1st use, 6 mos.; (b) exempt Ir-192 seeds in ribbons and unused sources; (c) 6 months. inventory
0.75/Release. . .patients containing. . .RAM	(a) OK if others TEDE $< 5\,\text{mSv}\cdot\text{year}^{-1}$; (b) Instruction if others TEDE $> 1\,\text{mSv/year}$;
0.80/Mobile medical services	(a) Facility agreement letters; (b) on-site, before use survey meter checks; (c) Post-treatment surveys; (d) possession licenses required for all sites
0.92/Decay in storage	(a) $T_{1/2} < 120$ day; decay to background level; (b) remove labels; keep records

Table 8c. Some Components of CFR 35 (F) Applicable to Manual Brachytherapy

Section	Major Contents of Section
0.404/Surveys after. . . implant and removal	(a) After implant; source accountability; (b) After source removal; keep records
0.406/Source accountability	(a) . . .at all times. . .in storage and use; record
0.410/Safety instructions	(a) Initially, annually. . .to caregivers; (b) Size, type, handling, shielding, visitor
0.415/Safety precautions	(a) No room sharing with regular patients; (b) Post-room (RAM) and visitor limits; (c) Emergency equipment for source retrieval from or in patient
0.432/Source calibrations (post-10/24/04)	(a) Determine output or activity; (b) positioning in applicators per "protocols"; (c) decay outputs/activities at 1% intervals; keep records
0.433/Decay Sr-90 sources	Only AMP shall calculate decayed activity and keep records
0.457/Therapy-related computer systems	(a) Acceptance testing per "protocols"; (b) Source input parameters; (c) accuracy of dose/time at points; isodose and graphics plots; (d) localization image accuracy

Table 8d. Some Components of 10CFR 35 (H) for Photon. . .Remote Afterloaders

Section	Major Components of Section
0.604/Surveys of patients	Before releasing patient. . .survey patient and RAU to confirm . . .returned to safe
0.605/Installation, . . ., repair	(a) Certain source work, that is install, adjust, and so on, by licensed person; (b) For LDR RAU, licensed person or AMP can do certain source work; record
0.610/Safety procedures	(a) Secure unattended RAU; (b) only approved individuals present in room; (c) No dual operations; (d) written procedures for abnormal situations; posted copies; initial/annual instructions with drills; records
0.615/Safety precautions	(a) Control access with interlock; (b) area monitors; (c) CCTV/audio for all except LDR RAU; (d) for MPD/PDR an AMP and AU **or** operator-emergency response MD present at initiation and immediately available during treatments; (e) for HDR an AU and AMP physically present at initiation, but, during continuation, AMP and AU or operator-emergency response MD; (f) emergency equipment for unshielded source or source in patient.
0.657/Therapy-related computer system	(a) Acceptance testing per "protocols"; (b) Source input parameters; (c) accuracy of dose/time at points; isodose and graphics plots; (d) localization image accuracy; (e) electronic transfer to RAU accuracy

Table 8e. Some Components of 10CFR 35 (H) for Photon. . .Remote Afterloaders

Section	Major Contents of Section
0.630/Dosimerty system (DS) equipment	(a) Except for LDR RAUs, NIST/ ADCL calibrated DS; (b) 2 year and after service; or, (c) 4 year, if intercom pared with calibrated DS within 18–30 month.and $< 2\%$ change
0.633/Full calibrations (FC) of RAUs	(a) Before 1st use; at source exchanges and/or repairs to exposure assembly; (b) for $T_{1/2} > 75$ days, excluding LDR RAUs, quarterly; (c) LDR RAUs yearly; (d) FC: 5% output/1 mm positions, source retraction, timer accuracy/linearity; (e) tube lengths and functions; (f) quarterly autoradiographs of LDR RAU sources; (g) decay outputs/activities at 1% intervals; (h) FC and decay by AMP; keep records; for LDR RAU can use manufacturer's data for FC
0.643/Periodic spot-checks (SC) of RAUs	(a) For LDR RAUs, before 1st treatment; for other RAUs 1st use daily; (b) per WP by AMP; (c) AMP review by 15 day; (d) SC includes: interlocks, status lights, audio and CCTV, emergency equipment, source position monitors, timer, clocks, decayed source activity
0.647/Additional requirements. . .mobile RAUs	(a) Survey meter checks; (b) source inventory; (c) all 0.643 checks; (d) interlocks, status lights, radiation monitors, source positioning, before 1st use, simulated treatment at each address

and PDR units. Physicians other than AUs, trained in MDR and PDR operation, emergency procedures, and source removal, may work under the supervision of an AU. Note: We denote them as substitute authorized users (SAU). For the initial treatment, the AMP and AU *or* SAU must be present; during subsequent (continuation) treatments, the AMP, AU, *or* SAU must be *immediately available*. These requirements are less

onerous than the prior requirements of the AU *always* being present during all treatments. Another section adopts AAPM good practices, per various protocols, for quality assurance of RAU therapy planning systems, as a regulation.

Requirements for dosimetry systems (DS), full calibrations (FC), and spot-checks (SC) are described, including those for mobile services.

Some Components of 10CFR35 (J)(Recognition of Specialty Boards). The 2002 revisions in 10CFR 35 did not address personnel training. On March 30, 2005, the NRC published the final rule (5) regarding specialty boards and personnel training. The rule identifies (on the NRC web site, *not* in the published rule) various approved specialty boards and describes pathways for approvals of RSOs, AMPs, authorized nuclear pharmacists, and physicians using many forms of byproduct materials. The rule offers multiple pathways by which individuals may achieve authorizations to perform various tasks or assume authorized titles, (e.g., RSO, AMP, authorized nuclear pharmacist, or physician authorized user). One pathway is the educational degree -> experience -> specialty examination -> certification path. Another pathway is the supervised experience -> preceptor statement path. For example, Table 9 shows five ways an individual, depending on their education, experience, and certification status, can qualify to be an RSO. This flexible approach offers individuals multiple pathways to achieve authorization, which maintaining the integrity of the approval process. For those not physicians, the education requirements are either (a) a bachelor or graduate degree in physical science, or, engineering or biologic science with 20 college credits in physical science, or, (b) a master's degree or PhD in physics, medical physics, or physical science, engineering, or applied mathematics. Experience requirements vary from 1 to 5 years depending on the authorization, and are shorter for those with higher degrees. Generally, experience must be gained under a certified medical physicist or authorized individual, and documented. Preceptors must document the successful completion of any structured training programs and attest to the individual's competency and ability to perform learned tasks independently. In some instances, structured didactic training programs including classroom and laboratory training are allowed. Table 10 show similar requirements for becoming an AMP or ANP.

Training requirements for physicians, Tables 11 and 12, generally offers physicians two options: Completing requirements for medical specialty board certification and passing a certification examination, or, completing a structured educational program with a specific number of classroom and laboratory hours and work experience. In some instances, a preceptor must provide a written statement attesting to the satisfactory completion of the requirements and to the individual's "...competency sufficient to function independently..." (5) In other instances, a certain number of cases must be performed. The classroom and laboratory training requirements are specific to each specialty. Tables 11 and 12 only broadly describe training requirements; details of each specialty training program as described in USNRC 2005.

Table 9. Training Requirements for Radiation Safety Officers

Person	Degree	Experience	Examination	Classroom Laboratory Training	Preceptor Statement	Special Training
(1) Radiation Safety Officer	*and* B or GD in PS; or, E or BS w 20 cc in PS;	*and* 5 or more years in HP including 3 years in AHP	*and* Passes Exam			
or, (2) Radiation Safety Officer	*and* M or PhD in P, MP, or PS, E, AM	*and* 2 years full-time training in MP under supervision by CMP, *or*, in CNM, by physician AU	*and* Passes Exam			
Or, (3) Radiation Safety Officer		1 year full-time RS under supervision by RSO		200 h in topical areas		
Or (4) Radiation Safety Officer	*and* is a CMP	*and* applicable experience			*and* has written attestation by preceptor	*and* training in RS, regulatory issues, and emergency procedures
or (5) Radiation Safety Officer	*and* is AU, AMP, or ANP on license	*and* applicable experience			*and* has written attestation by preceptor	*and* training in RS, regulatory issues, and emergency procedures

ANP = Authorized nuclear pharmacist; PS = Physical science; B = Bachelor's degree; CMP = Certified medical physicist; BS = Biological science; RS = Radiation safety; CC = College credits; AU = Authorized user; E—Engineering; CNM = Clinical nuclear medicine; GD = Graduate degree; MP = Medical Physicist; M = Master's degree; RSO = Radiation safety officer; Ph.D = Doctoral degree; AMP = Authorized medical physicist.

Table 10. Training Requirements for Authorized Medical Physicist and Nuclear Pharmacist

Person	Degree	Experience	Examination	Classroom Laboratory Training	Preceptor Statement	Special Training
(1) Authorized Medical Physicist	*and*; M or Ph.D . in P, MP, or PS, E, AM	*and* 2 years under supervision by CMP, *or*, ...	*and* Passes			
or (2) Authorized Medical Physicist	*and*; M or Ph.D. in P, MP, or PS, E, AM	*and* 2 yrs in CRF under supervision by AU eligible physician	*and* Passes			
or (3) Authorized Medical Physicist	*and*; M or Ph.D. in P, MP, or PS, E, AM	*and* 1 year full-time training in MP and 1 year full-time experience by AMP eligible MP			*and* has written attestation of "competency and independency" by MP preceptor	*and* training in device operation, clinical use, and treatment planning systems
(1) Authorized Nuclear Pharmacist	Pharmacy; or, passed FPGEC exam	4000 h in nuclear pharmacy	*and* Passes			Current, active license
or (2) Authorized Nuclear Pharmacist				700 h in structured program with 200 h in topical areas	*and* has written attestation of "competency and independency" by preceptor ANP	

AM = Applied mathematics; ANP = Authorized nuclear pharmacist; PS = Physical science; CMP = Certified medical physicist; AMP = Authorized medical physicist; RS = Radiation safety; FPGEC = Frgn pharm.grad exam comm.; AU = Authorized user; CRF = Clinical radiation facility; E = Engineering; P = Physics; MP = Medical physicist; M = Master's degree.

Licensure. There are two categories of NRC licenses: General License and Specific License. General licenses have been issued for non-medical uses, such as fixed gauges containing sealed radioactive sources. Medical licenses are for specific uses of a licensed material in a medical program, for example, diagnostic nuclear medicine program. Specific licenses control manufacture, production, acquisition, receipt, possession, preparation, use, and transfer of byproduct material for medical uses. A Type A license of broad scope, often held by university medical facilities, exempts the licensee from certain requirements of a specific license, but requires the facility to assume responsibility by certain administrative processes for the radiation protection program. NRC license requirements, applications, renewals, amendments, notifications, exemptions, and issuances are described in *10CFR 35.11–19* (2). Licenses categories exists for the use of unsealed byproduct material for uptake, dilution, excretion studies without a written directive (§35.100), unsealed byproduct material for imaging and localization studies without a written directive (§35.200), unsealed byproduct material requiring a written directive (§35.300), manual brachytherapy sources (§35.400), sealed sources in teletherapy units, and stereo tactic radiosurgery units (§35.600), and for other uses of byproduct materials (§35.1000).

Some components of 10CFR 35 (L) (Record retentions) (Table 13). Table 13 summarize the duration (for license, for program, and for 5 and 3 years) requirements for the retention of records.

Some components of 10CFR35 (M) (Reports ...Medical Events ... Sources) (Table 14). The term *misadministration* is replaced with the term *medical event* (ME). The ME depends, in some cases, on the *difference* (presumably lower or higher) in delivered dose and prescribed dose (PD), and in other cases, in *exceeding* the PD. Moreover, the definitions are not in medical physics terms of absorbed dose in gray (Gy); rather, they are in health physics terms of effective dose equivalent (EDE), shallow dose equivalent (SDE), in sievert (Sv). Recall that in partial organ irradiation in health physics, organ or tissue weighting factors apply in calculating DE. As a brachytherapy ME will likely involve adjacent organs, some judgment may be required in deciding on the correct DE in an ME.

Table 14 summarizes the reporting of medical events; Reporting requirements are similar to pre-2002 regulations.

Transport

Every day thousands of packages containing radioactive material move via public transportation routes—roads, airplane, and railway. Of all the hazardous material shipments, it is estimated that ~1%, nearly 3 million packages annually, involve radioactive materials (17). These packages are needed for medicine, industry, and research.

Shipments can be made only to persons who are licensed by the Nuclear Regulatory Commission (NRC) or appropriate Agreement State to receive radioactive materials.

Table 11. Some Training Requirements for Physicians Using Sealed Sources and Medical Devices

Person	Certification Examination	Education	Experience	Laboratory Training	Preceptor Statement	Authorized User
(1) Physician (Manual brachytherapy and sources), *or*,	Passes examination by medical specialty board	3 year MR in Rad Onc				
(2) Physician (Manual brachytherapy and sources)		Structured educational program with 200 h topical classroom and laboratory and 500 h work experience	3 years clinical supervision by AU in Rad Onc		*and* has AU preceptor's written attestation of competency sufficient to function independently	
(1) Physician (Ophthalmic use Sr-90)		Active practice and 24 h classroom and laboratory training applicable to medical use of Sr-90,	*and* AU supervised clinical training		*and* has AU preceptor written certification of completed requirements *and* attestation of competency sufficient to function independently	
(1) Physician, dentist, or podiatrist (Sealed sources for diagnosis), *or*,	Passes examination by medical specialty board					
(2) Physician (Sealed sources for diagnosis)		8 h classroom and laboratory training applicable to medical use of Sr-90,		Has completed training in use of device for uses requested		
(1) Physician (RA, T, GSR units, TMD), *or*	Passes examination by medical specialty board	3-year MR in Rad Onc				
(2) Physician (RA, T, GSR units, TMD)		Structured educational program with 200 h topical classroom and laboratory and 500 h work experience	3 year clinical supervision by AU in Rad Onc		*and* has AU preceptor written certification of completed requirements *and* attestation of competency sufficient to function independently	

RA = Remote afterloader; MR = Medical residency; T = Teletherapy unit; AU = Authorized user; GSR = Gamma stereotactic radiosurgery Unit; TM = Therapeutic medical device

The shipment must be made in accordance with procedures established by the recipient. Prior to shipping radioactive materials, a copy of the recipient's radioactive materials license should be on file with the shipper's Radiation Safety Office to document what radionuclides, forms, and quantities the recipient is authorized to receive.

There are five categories of radioactive material packages. Development of the technical criteria for each packaging category is correlated to certain general and performance requirements. The categories include (1) excepted or limited quantity packaging; (2) type A packaging; (3) type B packaging; (4) industrial packaging; (5) fissile material packaging. All medical shipments occur in the first two categories. Figure 2 illustrates the "spectrum" of increasing package hazard with activity.

Both the Department of Transportation (DOT) and the Nuclear Regulatory Commission are responsible for the regulations governing a package containing hazardous materials that are intended for transport on public routes(18). The DOT regulations are found in 49 CFR 107, 172-178. The NRC regulations are found in Title 10 CFR Part 71. In 1979, the DOT and NRC agreed to a Memorandum of Understanding under which the DOT regulates Type A packages and below, carriers, and has authority for international shipments. The NRC regulates Type B and fissile packages, investigates incidents and accidents, and provides technical advise to DOT.

The transportation requirements were revised in October 2004 to bring U. S. standards into consistency with the latest international transportation safety regulations (19). The

Table 12. Some Training Requirements for Physicians Use of Radiopharmaceuticals

Person	Certification Examination	Education	Experience
(1) Physician (Uptake, Dilution, and Excretion Studies), *or,*	Passes examination by medical specialty board	Satisfies board education requirement	
(2) Physician (Uptake, Dilution, and Excretion Studies), *or*,		40 h topical classroom and laboratory	*and*, 20 h clinical supervised by AU
(3) Physician (Uptake, Dilution, and Excretion Studies)		Successfully completed 6 month NM training	
(1) Physician (Imaging and Localization Studies), *or*,	Passes examination by medical specialty board	Satisfies board education requirement	
(2) Physician (Imaging and Localization Studies), *or*		200 h classroom and laboratory training applicable to medical use and 500 h supervised work	*and* 500 h AU supervised clinical training
(3) Physician (Imaging and Localization Studies)		Successfully completed 6 month NM training	
(1) Physician (Therapy use Unsealed Byproduct Material), *or*,	Passes examination by medical specialty board	Satisfies board education requirement	
(2) Physician (Therapy use Unsealed Byproduct Material)		80 h topical classroom and laboratory	*and, c* linical supervision by AU for specific number of cases
Physician (Only I-131 for Hyperthyroidism, Thyroid Ca)		Special experience and 80 h classroom and laboratory training	*and*, c linical supervision by AU for specific number of cases
(1) Physician (Sealed Sources for Diagnosis), *or*,	Passes examination by medical specialty board	Satisfies board education requirement	
(2) Physician (Sealed Sources for Diagnosis)		8 h classroom and laboratory training	

international regulations follow the International Atomic Energy Agency (IAEA) report Safety Series ST-1-R, which most foreign countries have adopted (20). This is important because most radioactive materials for medical use are produced outside U.S. borders, for example sealed sources, $^{99}Mo/^{99m}Tc$ generators.

There are four essential elements that are the shipper's responsibility to properly providing packages for transport that contain radioactive, or any other hazardous, materials. These are proper containment, labeling/marking, documentation, and training. The major factors affecting these requirements for these elements are the radionuclide, physical form, and quantity (activity). The specific requirements for packaging containment, labeling, and documentation are in the relevant sections of 49CFR 172-177. This information can be found at the website http://hazmat.dot.gov.

Table 13. Some Components of 10CFR 35 (L) (Record Retentions)

Record Retention Requirement	Section
Duration of license	0.2024/RPP (b) RSO authority
Duration of program (device)	0.2610/Safety procedures for device
5 years	0.2041/Procedures for WP; 0.2026/RPP changes
3 years	0.2040/WDs;.2061/Meter calibrations; 0.2067/Leak tests and inventories; 0.2070/Surveys; 0.2075/Patient release; 0.2080/Mobile services; 0.2092/Decay in storage; 0.2310/Safety instructions; 0.2404/Implants and source removals;0.2406/Source accountability; 0.2432/Source calibrations; 0.2433/Sr-90 decays; 0.2605/RAU installation, repairs;0.2632/Full calibrations; 0.2643/Spot checks; 0.2647/Additional mobile records;

Table 14. Some Components of 10 CFR35 (M)[a]

Section	Major Contents of Section
0.3045/Report/notification medical event (excluding patient intervention) (1)	Dose differs from PD >0.05 Sv EDE, 0.5 Sv organ/tissue and SDE skin, and, TD, *and*, TD delivered differs from PD by $+20\%$ or falls outside PD range; or single fraction delivered dose differs from single fraction PD $+50\%$
0.3045/Report/notification medical event (excluding patient intervention) (2)	Dose *exceeds* 0.05 Sv EDE, 0.5 Sv organ/tissue and SDE skin, and, TD from wrong: (a) byproduct material; (b) administration route; (c) person; (d) treatment mode; (e) leaking source
0.3045/Report/notification medical event (excluding patient intervention) (3)	Excluding migrating permanent implant seeds, dose to skin/organ/tissue *other* than treatment site that exceeds 0.5 Sv organ/tissue and $+50\%$ dose expected from WD
0.3045/Report/notification medical event (excluding patient intervention) (3) (b)	Report any patient interventions producing permanent/physiological damage
0.3045/Report/notification medical event (excluding patient intervention) (3) (c, d)	Notify NRC next calendar day after ME with written report in 15 days; notify referring MD and patient unless referring MD chooses not to for medical reasons; details of reports omitted here
0.3067/Report leaking source	Report >5 nCi removal contamination within 5 days

[a]Reports . . .Medical Events . . . Sources.

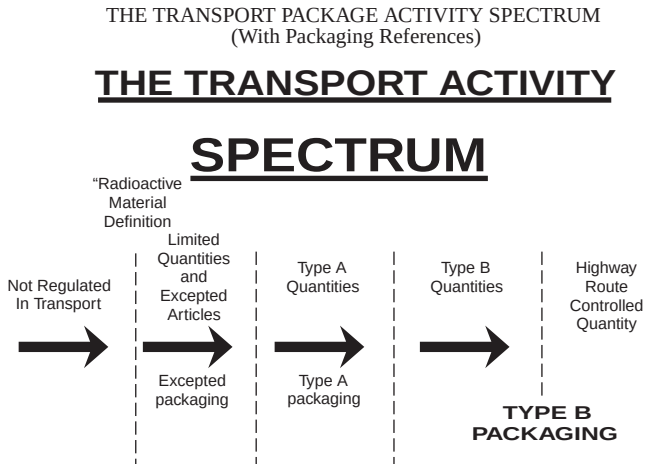


Figure 2. The "spectrum" of increasing radioactive package hazard with activity.

All shipments of radioactive material, with the exception of those containing very small, limited quantities must have labels bearing the word "Radioactive" and affixed to opposite sides of the outer package. There are three different labels: White-I, Yellow-II, or Yellow-III, as shown in Fig. 3. The criteria for the three labels are given in Table 15.

Training must be provided for those that prepare for transport, transport, or receive packages of hazardous materials. The training must be commensurate with the duties involved. For medical facilities, the training involves the proper receipt of radioactive packages and preparing packages for return to the vendor or manufacturer. For shippers, the training must be done at least every three years and be certified by the employer.

The receipt of labeled radioactive packages must be handled according to the procedures in NRC regulations (10 CFR 20.1906). This requires assessing radiation levels and removable contamination within 3 h of taking possession. Examples of returned packages are residual radiopharmaceuticals in syringes or vials from nuclear medicine studies or sealed sources after radiation therapy use. Any returned package must be prepared for transport in accordance with DOT requirements.

It is important to note that *anyone shipping radioactive materials must receive training from an approved program beforehand!*

Figure 3. Labels for radioactive packages based on the activity contents and radiation levels outside as given in Table 15.

Table 15. Shipping Label Criteria[a]

Label	Transportation Index, TI	Maximum Radiation Level on Surface, X
White–I	$TI < 0.05$	$X < 5\ \mu Sv \cdot h^{-1}$
Yellow–II	$0.05 \leq TI < 1$	$5\ \mu Sv \cdot h^{-1} \leq X < 500\ \mu Sv \cdot h^{-1}$
Yellow–III	$1 \leq TI < 10$	$500\ \mu Sv \cdot h^{-1} \leq X < 2\ Sv \cdot h^{-1}$
Yellow–III exclusive use of vehicle	$TI \geq 10$	$2\ mSv \cdot h^{-1} \leq X < 10\ mSv \cdot h^{-1}$

[a]Transportation index $= 100$ the maximum reading in $mSv \cdot h^{-1}$ at 1 m from the surface.

Disposal of Radioactive Material

Radioactive materials used in medicine can be solid, liquid, or gaseous. Some solids are specially encapsulated and called sealed sources. When the material is no longer useful or in a form or presence that is undesirable, the radioactive material is considered waste, and the licensee must dispose of it. All waste generated from medical use is categorized as low level radioactive waste. Depending on various factors, radioactive waste can be disposed by (1) decay-in-storage (DIS); (2). discharge into the environment (3) transfer for land burial; (4) return to the vendor/manufacturer.

Disposal by Decay-in-Storage. This is the dominant disposal method for radioactivity used in nuclear medicine. The container of radioactivity is stored and simply allowed to radioactively decay to background level. Currently, this disposal method is only available for radionuclides with a physical half-life <120 days. Depending on the size of the operations, the limiting factor with this method is adequate space for the waste volumes generated during the time that the waste is segregated to "decay". Shielding of the container(s) may be another consideration depending on the radiation levels involved from the waste during radioactive decay. Procedures must be developed to comply with 10 CFR 35.92 for the decay-in-storage (DIS) of waste. Release into the general medical waste stream requires that the radiation level be at background level when measured at the surface of the unshielded waste container. The survey meter or instrument used to measure the radiation level must be capable of detecting the radiation being emitted from the radionuclide(s) being stored.

Disposal into the Environment. In some circumstances, especially with medical research, the disposal of liquids into the sanitary sewer or by evaporative release to the atmosphere or the discharge of volatile gases may intentionally occur. This is permissible and safe providing that compliance is maintained with other regulations regarding toxic or hazardous properties of these materials.

Disposal in the sanitary sewer or into the atmosphere must comply with 10 CFR 20, Subpart K. The NRC release or discharge limits have a concentration and annual aggregate activity limit. Records of each release is kept, such that a periodic (at least annual) assessment can be performed to confirm compliance with release limits required in Part 20. These limits apply at the facility boundary of the radioactive material licensee. The release limits are radionuclide specific such that exposure or intake would not exceed the applicable occupational or general public dose limits. The current release limits are derived from the scientific basis presented in 1990 by the ICRP (21). Patient excreta containing readily dispersible forms of radioactivity are exempt from these release limits.

Waste from *in vitro* laboratory kits that use radioactive materials under a general licensed pursuant to 10 CFR 31.11 is exempt from waste disposal regulations. Radioactive labels must be defaced or removed, but there is no requirement to keep a record of release or make any measurement. A standard of good practice is to do a radiation survey of any general waste from an area where radioactive materials are used, such as a nuclear medicine radiopharmacy or an inpatient therapy room, to confirm that it is at background levels before release into the general waste.

Transfer for Burial. For some medical facilities, radioactive material may require disposal by transfer to a burial site because the volume of waste generated requires removal or the waste contains long-lived (>120 day half-life) items that cannot be decayed in storage. Medical facilities use a broker licensed by the NRC or Agreement State to receive the material. Packaging will follow instructions received from the broker and the burial site operator. Records of the transfer to the broker must be maintained to comply with 10 CFR 20. Because this is the most expensive means of disposal, most generators of waste also employ volume reduction, (e.g. compaction) to reduce costs.

At the time of writing, there are only three burial sites for low level radioactive waste in the United States Richland, Washington, Barnwell, South Carolina, and Tooele, Utah. All are commercially operated and regulated by the respective state. The facilities are designed, constructed, and operated to meet safety standards. The operator of the facility must also extensively characterize the site on which the facility is located and analyze how the facility will perform for thousands of years into the future. In 1985, the Low-level Radioactive Waste Policy Amendments Act gave the states responsibility for the disposal of their low-level radioactive waste by encouraging the states to enter into compacts that would allow them to dispose of waste at a common disposal facility. While most states have entered into compacts, but no new disposal facilities have been built since the Act was passed, or are any expected to be.

Since the 1985, the volume of medical low level radioactive waste shipped for burial has dropped dramatically because of the cost of disposal, employment of volume reduction methodologies, and the conversion to short half-life or nonradioactive agents.

Return Sources to the Vendor or Manufacturer. For solid or sealed sources especially, a viable means of disposal for a medial facility is return to the vendor or a manufacturer. This is common with items that have exceeded their useful activity or shelf-life, such as $^{99}Mo/^{99m}Tc$ generators or brachytherapy sealed sources (e.g., ^{192}Ir, ^{125}I) or quality control calibration sources (e.g., ^{57}Co, ^{153}Gd). For such package, the packaging, labeling, and surveys must comply with the instructions of the vendor/manufacturer and 10 CFR 71 (NRC) and 49 CFR 173 (DOT) regulations. Currently, there is no distinct time at which a sealed source might be considered waste. The licensee determines when a material is no longer usable and becomes considered part of the radioactive waste stream. For solid sources with >120 day half-life, because land burial is very expensive, many licensees choose to simply "store" sources under their control. Such sources require routine inventory and periodic leak-testing. The current standards for stored, *unused* sealed sources require inventory every 6 months and leak test within 10 years.

COMMUNICATIONS FROM THE NRC

The NRC issues to licensee's bulletins, directives, guidance's, information notices, newsletters, and regulatory summaries as new issues not covered in regulations arise and must be addressed. In some cases, these documents endure for many years, and may actually be incorporated by agreement states into their regulatory statutes.

Bulletins

Bulletins provide information to NRC licensees. Apparently there are no recent bulletins pertaining to medical Applications; the last one was in 1997 (22).

Directives

Directives appear in several forms. FC86-4, Revision 1—*Information Required for Licensing Remote Afterloading Devices*, a long-standing (1986) policy and guidance directive, explained the contents for NRC license applications for RAUs. While it is not current on the NRC web site, some states have adopted it, with some changes, into their licensing process for RAUs. FC83-20, Revision 2-*Facility Interlocks and Safety Devices for High, Medium, and Pulsed Dose-Rate Afterloading Units*, is not on the NRC web site. As the title implies, this release clarified the requirements for interlocks and safety devices. It appears that issues raised are addressed in the 2002 10CFR 35 revisions.

Guidances

Guidance's often discuss evolving technologies. For example, as intravascular brachytherapy developed, the NRC issued several guidance documents (23,24). These were necessary as the new 10CFR35 applies only to photon-emitting RAUs; beta-emitting RAUs fall into the emergent technology category evaluated on a case-by-case basis.

Information Notices

Information Notices advise licenses of recent concerns usually arising from medical events reported to the NRC. A recent notice discussed failures of HDR RAUs (25).

Newsletters

Newsletters, notable, Nuclear Materials Safety and Safeguards (NMSS), announce medical events and enforcement actions against those who violate regulations. A recent one reported on a hospital's failure "...to secure...licensed material..." (26).

Regulatory Summaries

Regulatory Summaries often clarify issues about the interpretation of regulations, such as the calibration measurements for brachytherapy sources (4).

Recent NRC Activities

Recent or current NRC activities are posted on the website www.nrc.gov. For those interested in commenting on proposed NRC regulations, a site, www.ruleform.llnl.gov, is available.

REGULATORY STANDARDS : OSHA

Tables 16–18 offer a limited synopsis of the major components of the OSHA regulations. As noted earlier, current enforceable OSHA regulations, *29 CFR 1910.1096 (Ionizing Radiation)*, dating from the 1970s, are now at variance with the recent NRC regulations.

A recent supporting statement (Fed Register [07/23/2004]) for information-collection requirement offers some insight into OSHA regulation terms, definitions, and their application.

As with the NRC, over the years OSHA has issued Directives, Standard Interpretations, and Compliance Letters regarding regulations. They are available on the website www.osha.gov.

While a few cover general radiation topics, most relate to non-medical (nuclear power plant) radiation issues. There appear to be no releases within the last 5 years that relate to medical uses of radiation under OSHA standards.

NONBYPRODUCT MATERIALS AND MACHINE-PRODUCED RADIATION

As noted above, the NRC was authorized only to oversee the use of fissile and byproduct material. Regulation of naturally occurring or accelerator produced radionuclides, or of radiation from machines fell to the individual states. Since every state develops their own regulations, the depth and coverage of those regulations very widely. Often, states with smaller populations and small non-federal radionuclide programs tended to have less complete or in-depth regulations than states with larger populations and programs. Two developments have been working to change the wide variations in regulations between states.

Table 16. Partial Contents of 29 CFR 1910.1096(a), (b), and (c) of OSHA Regulations[a]

Section	Major Contents of Section
(a) Definitions—Radiation and areas	(1) Radiation; (2) Radioactive materials; (3) Restricted area; (4) unrestricted areas;
(a) Definitions—Quantities and equivalencies	(5) Dose; (6) Rad; (7) Rem; (1 R X or γ- ray; 1 rad X- or γ- ray or beta particle; 0.1 rad high energy proton; (8) Air dose
(a) Definitions—Neutron flux or equivalent	Neutron flux dose equivalency table
(b) (1) Exposure to employed individuals 18 years age or older in restricted areas (Rem/calendar quarter)	Whole body; Head and truck; active blood forming organs; eye lens, gonads: 1.25
	Hands and forearms; feet and ankles: 18.75 Skin of whole body: 7.5
(b)(2) Greater quarterly whole body doses allowed based on individual's age"N"	Whole body dose shall not exceed 3 rem per quarter *and* shall not exceed 5(N-18)
(b) (3) Exposure to employed individuals *under* 18 years age in restricted areas (Rem/calendar quarter)	Quarterly calendar dose limited to 10% of that allowed those 18 years of age
(c) Exposure of employed individuals 18 years age or older to airborne radioactive material in restricted areas shall not exceed	Limits in 10CFR Part 20 Table I, Ax.B (1971); for 40 h workweeks, 7 consecutive days; time proportionately applicable
(c) Exposure of employed individuals *under* 18 years age to airborne radioactive material in restricted areas shall not exceed	Limits in 10CFR Part 20 Table II, Ax.B (1971); for 40 h workweeks, 7 consecutive days; time proportionately applicable

[a]The use of conventional (old) units in this table reflects the fact that these regulations are outdated and lag behind the NRC regulations.

Agreement State Status

The first unifying factor is the growing trend toward agreement state status. An agreement state enters into an agreement with the NRC to take over for the NRC regulation and control of byproduct material. Before doing so, the state must demonstrate that the state regulations for byproduct material are compatible with those of the NRC. "Compatibility" varies based on guidelines from the NRC as to how important the NRC feels that the state regulations agree with the federal, according to the following scale (27):

A Basic radiation protection standard or related definitions, signs, labels or terms necessary for a common understanding of radiation protection principles. The State program element should be essentially identical to that of NRC;

B Program element with significant direct transboundary implications. The State program element should be essentially identical to that of NRC;

C Program element, the essential objectives of which should be adopted by the State to avoid conflicts, duplications or gaps. The manner in which the essential objectives are addressed need not be the same as NRC, provided the essential objectives are met;

D Not required for purposes of compatibility.

For example, occupational exposure limits fall under category A, requiring congruence between the state and federal regulations. On the other extreme, most application and recording regulations are left to the states' discretion. For the most part, the laxer categories are those with less impact. Thus, as states have adopted agreement status, the variation in regulations between states has decreased. Table 4 lists the agreement states as of 2005.

Conference of Radiation Control Program Directors

Established in 1968, the Conference of Radiation Control Program Directors (CRCPD) is an organization of representatives of state radiation control programs. The organization shares information useful to state radiation control agencies, and has educational meetings focused on topics of current interest to state regulators. The CRCPD also distributes to its members model state regulations, so when states revamp their respective radiation safety codes, they need not start from nothing(28). The contents of the model regulations are discussed below. Because many state agencies use these models as a guide for their radiation regulations, increasingly the various states' regulations have been converging. Still, many important aspects of regulations remain, for example, the allowed radiation limit to the general public. While most states follow the federal rules, some use more restrictive levels based (sometimes erroneously) on recommendations of the ICRP.

CRCPD Model Regulations

Since the CRCPD model program serves as the basis for many of the state rules, we will consider the provisions here for regulations dealing with ionizing radiation not from byproduct material. Because of the compatibility

Table 17. Partial Contents of 29 CFR 1910.1096(d) and (e) of OSHA Regulations

Section	Major Contents of Section
(d) (1) Definition of a survey	"An evaluation . . .radiation hazards. . .production, use, release, disposal, or presence . . .radioactive material or . . .radiation. . ."
(d) (2) Employer responsibility for monitors	". . .shall provide. . .personnel monitoring equipment. . ."
(d) (2) (i) 18 year age or older employee use of monitors in restricted areas;	". . .employeee . . .restricted area likely to receive a quarterly dose > 25% that in (b)(1); or, enters a high radiation area
(d) (2) (ii) *Under* 18 year age employee use of monitors in restricted areas	". . .employeee . . .restricted area likely to receive a quarterly dose > 5% that in (b)(1)
(d) (3) Personnel monitoring equipment	"e.g., film badges & rings, pocket chambers and dosimeters
(d) (3) Area definitions	Radiation area. . .could receive > 5 mrem in any 1 h; or, > 100 mrem in 5 consecutive days; High radiation area. . . could receive > 100 mrem in any 1 h; Airborne radioactivity area. . .concentrations in excess 10CFR Part 20 Table I, column 1, Ax.B (1971)
(e) Caution signs, labels, signals	Radiation symbol(trefoil) described;
(e) (2) Radiation area posting	Radiation caution symbol with "Caution–Radiation Area"
(e) (3) (i) High radiation area posting	Radiation caution symbol with "Caution–High Radiation Area"
(e) (3) (ii) High radiation area control	". . .equipped with control device . . .cause radiation levels to be reduced < 100 mrem in 1 h, or, . . .energize . . .alarm system. . . individual entering . . .supervisor. . .made aware of entry."
(e) (4) Airborne radioactivity area posting	Radiation caution symbol and "Caution–Airborne Radioactivity Area"
(e) (5) (i) Radioactive materials posting (excluding natural uranium or thorium)	Areas/rooms > 10 times quantities in 10CFR Part 20 Apx.C (1971)
(e) (5) (ii) Radioactive materials posting for natural uranium or thorium	Areas/rooms > 100 times quantities in 10CFR Part 20 (1971)
(e) (6) (i) Container labeling (excluding natural uranium or thorium)	Containers . . .transported, stored, used. . .> quantities in 10CFR Part 20 Apx.C (1971). . .Radiation symbol and "Caution-Radioactive Materials"
(e) (6) (ii) Container labeling for natural uranium or thorium	Containers . . .transported, stored, used . . . > 10 times quantities in 10CFR Part 20 Apx.C (1971). . .Radiation symbol and "Caution-Radioactive Materials"

The use of conventional (old) units in this table reflects the fact that these regulations are outdated and lag behind the NRC regulations.

requirement to become an agreement state, those parts of the CRCPD model regulations that deal with material under NRC oversight follow the federal rules as discussed above. Thus, these need not be considered here. The FDA does impose some requirements on the *manufacturers* of radioactive materials and radiation-producing machines intended for human use, but that leaves the *use* of machine-produced radiation and naturally occurring and accelerator-produced radionuclides only under the control of individual states.

The model regulations fall into many sections, with each section covering a particular part of radiation safety. General rules that apply to all applications and follow the NRC notably Parts 19 and 20 come in sections in the beginning. In addition to the general provisions, each of the parts that deal with particular applications all have sections addressing shielding and survey requirements for the modality (such that the radiation levels satisfy Part 20 limits), safety requirements for operation (such as door interlocks to prevent walking in during irradiation), ventilation if airborne radionuclide production is possible, record retention requirements and training and experience.

Machine-Produced Radiation

While much of machine-produced radiation is covered by state regulations, when used on humans applications manufacture of the units falls under the auspices of the FDA. The FDA rules can be found in 21 CFR 1020. For the most part, the state regulations follow the FDA guidances when applicable, but sometimes with a sizable delay.

Mammography forms a notable exception to the general lack of federal control over machine-produced radiation in medicine. Based on the MQSA, as noted above, the FDA sets requirements for practitioners on mammography, and failure to satisfy the requirements prevents providers from obtaining reimbursement from government sources. The requirements for mammography equipment are given below in the section on Diagnostic Units. In addition, there are considerable requirements placed on the training and experience of the persons involved: the radiologist, the radiographer, and the medical physicist [21 CFR 900.12 (a)].

Radiation producing machines fall into three main categories discussed in the following sections.

Table 18. Partial Contents of 29 CFR 1910.1096(f), (g), (h), (i), (j), and (k) of OSHA Regulations[a]

Section	Major Contents of Section
(f) Immediate evacuation warning signal	34 subsections regarding the signal characteristics, design, and testing requirements
(g) (i) Exceptions from posting requirements for sealed sources	Room/area with sealed source. . .not required. . .if radiation levels $< 5\,\text{mrem}\cdot\text{h}^{-1}$ at 12 in from source container/housing
(g) (ii) Exceptions from posting requirements for rooms housing radioactivity patients	Rooms. . .not required to be posted. . .personnel in attendance who shall . . .prevent individual exposure above limits
(g) (iii) Exceptions from posting requirements for rooms containing radioactive materials	Cautions signs not required for rooms containing radioactive materials for < 8 h and provided materials constantly attended. . .
(h) Exemptions for radioactive materials packaged for shipment	Radioactive materials packaged and labeled per DOT 49CFR Chp. I are exempt provided containers inside properly labeled
(i) (2) Instruction of personnel, postings	Individuals working in or frequenting any portion of a radiation area shall be informed of radioactive materials and radiation; instructed in safety. . .; instructed in applicable provisions of regulations. . .; advised of radiation exposure reports
(i) (3) Posing regulations and operating procedures	Employer. . .shall post. . .current copy of regulations and operating procedures
(j) Storage of radioactive materials	. . .shall be secured against unauthorized removal. . ..
(k) Waste disposal	. . .by transfer to an authorized recipient. . .
(l) (i) Notification (immediate) of incidents	. . .any incident involving radiation which may have caused . . .> 25 rem whole body, 150 rem skin, or 375 rem to feet, ankles, hands, or forearms, or, release of radioactive materials > 5000 applicable limits averaged over 24 h
(l) (ii) 2Notification (24 h) of incidents	. . .any individual . . .5 rem or more total body; 30 rem skin, 75 rem to feet, ankles, hands, forearms,
(m) Reports of overexposure and excessive levels and concentrations	. . ..written report in 30 days. . .to OSHA; notification of individual exposed
(n) Records	Advise employees of annual exposures; provide employees exposure records
(p) Definitions of agreement states	List of current agreement states

[a]The use of conventional (old) units in this table reflects the fact that these regulations are outdated and lag behind the NRC regulations.

Radiotherapy Units. Radiotherapy units consist of two major categories: orthovoltage X-ray units (i.e., conventional X-ray machines that treat with bremsstrahlung beams produced with tube potentials from 10 kVp to 300 kVcp) and those from accelerators (from electron beams with energies from 2 to 45 MeV). The regulations use as a diving line between the modalities a photon beam energy of 500 kV, which clearly delineates units since no machines currently in use run close to that specification. Table 19 lists the requirements for an orthovoltage unit, and Table 20 those for an accelerator.

Regardless of the machine type, the regulations require the output of the unit be determined using dosimeters calibrated at either the National Institute of Standards and Technology or at one of the Accredited Radiation Dosimetry Calibration Laboratories. The calibration procedure must follow a protocol established by a recognized national professional society. Also for either type of unit (except for contact therapy units), the facility design requires: the ability to monitor the patient aurally and visually; interlocks on the door to prevent entry during irradiation; beam-on indicators; and emergency power cutoffs by the control panel or door.

Radiography (Imaging) Units. Regulations for diagnostic radiographic units actually exceed those for the therapy units, even though the latter produce much greater quantities of radiation. Tables 21a,b and 22 give *highlights* of the regulations for radiographic and fluoroscopic units. The regulations also contain many points on how the specifications should be measured as well as cover other aspects not included in the tables. Table 21b gives values referred to in Table 21a. As an important factor in patient dose, the regulations also address exposure control for the various types of equipment.

In addition to the regulations for the radiographic and fluoroscopic units, there are also sections on radiotherapy simulators; computed tomography units; mammography units; mobile units; and veterinary units.

As noted above, mammography units have special requirements according to the MQSA. The requirements for these units are given in Table 23, and the special quality assurance requirements in Table 24. The quality assurance summary greatly simplifies the actual requirements, which have undergone some modifications to adapt better to various imaging systems and practice conditions. All persons involved in mammography, including the radiologist, radiographer and the medical physicist performing the quality measurements, must satisfy specific training and experience guidelines, as well as continuing education requirements.

Nonmedical Radiation-Producing Equipment. Nonmedical radiation producing equipment actually finds its way

Table 19. Requirements for Orthovoltage X-Ray Units

Leakage Radiation [air kerma rates]	< 1 mGy·h^{-1} 5 cm from housing
5–50 kV Systems	< 10 mGy·h^{-1} 1 m from target;
> 50 and < 500 kV Systems	< 300 mGy·h^{-1} 5 cm from housing
Permanent Beam Limiting Devices	Same attenuation as housing
Adjustable or Removable Beam Limiting Devices	Transmission $< 5\%$ of useful beam Opening indicated by light beam
Beam Filter System	Cannot be displaced Interlocked to prevent beam use with filter absent Slot provides same shielding as housing Filters clearly identified
Tube Immobilization	Cannot move when locked
Source Marking	Indicated to within 5 mm
Contact units beam blocking	Equivalent to 0.5 mm Pb at 100 kV
Timer	Unit has presetable timer and show elapsed or remaining time Retains reading with interruptions Terminates exposure after set time Precision of at least 1 s or 1% Prevents exposures with zero time Begins with shutter or is compensated for lags
Control Panel Functions	Displays indicate ac power, X-rays possible, X-rays on, shutter condition and tube potential and filter Termination button Locking device
Multiple Tubes	Only one used at a time Indication of which is in use
Target-to-Skin Distance	Accurate to within 1 cm, reproducible to within 2 mm
Shutters	Required if beam takes > 5 s to come on
Low Filtration X-ray Tubes	Permanent warning label

into medical application, for example, as cyclotrons making radioactive materials for imaging or analytic X-ray units to assess kidney stones. Much of the operation of such equipment would be covered by the general radiation safety provision of the regulations. Most of the additional rules deal with preventing the accidental irradiation of a person in a high radiation area.

Particle Accelerators (e.g., Cyclotrons). The main concern for a particle accelerator would be a staff member being in either the accelerator room or one of the rooms served by the beam lines. To prevent such occurrences, the rules require: interlocks on doors to prevent accidental entry with the beam on or inhibit the beam initiation with the door open; buttons to stop the beam from within the room; radiation-detector warning devices in the room; and hand-held Geiger counters carried when entering the room. The regulations also include the requirement for periodic testing of the safety devices to assure proper function.

Analytic X-Ray Units. The X-ray units covered under this heading usually are small devices (often fitting on a desktop), used for analysis of small samples, such as for crystallography or pathologic X rays of surgical samples. These devices are usually enclosed within a shielding box. While small, accidents that involve an operator's hand being in the box during beam production frequently lead to loss of fingers or hands. Thus, similarly to the particle accelerator, the rules try to keep hands out with the beam on, or prevent the beam if the doors are open. Rules include interlocks to prevent beam with doors open; warning lights indicating the status of the beam and shutters; and warning labels.

Nonbyproduct Radionuclides

State regulation of byproduct material must follow closely the NRC regulations. However, before the 2005 agreement, the states have been responsible originating their own regulations for NARM. Much of the suggested regulations (Part C) define quantities of NARM below regulatory concern. Table 31 gives a brief, and not nearly complete listing of some exemptions as examples. Appendix A of Part C of the suggested regulations gives air and water concentrations exemptions.

The regulations go on to exempt devices such as static eliminators containing less than specified amounts of radioactive materials (on the order of 20 MBq for heavy nuclides of 2 MBq for tritium). For clinical laboratories, small quantities of material (generally ~ 0.4 MBq except tritium at 1.85 MBq and ^{59}Fe at 0.7 MBq) used in assay kits and 1.85 MBq check sources are also exempt. The remainder of Part C addresses licensing and labeling. Medical use of radioactive materials is covered under Part G, which mostly mirrors the federal 10CFR35.

What is not clearly addressed in the model regulations is regulation on accelerator-produced radioactive materials. Some states have taken the tack that the same regulations

Table 20. Requirements for a Radiotherapy Accelerator

Leakage Radiation	Maximum $< 0.2\%$, average $< 0.1\%$ useful beam -2 m radius of central ray at isocenter $< 0.5\%$ 1 m from electron path Neutron dose compliant with IEC standard
Collimator leakage	$< 2\%$ of useful beam for photon beams Maximum $< 2\%$, average $< 0.5\%$ useful beam for electron beams, outside 7 cm of beam $< 10\%$ 2 cm outside of field
Filters/Wedges	Identification clearly marked Interlocked requiring selection Panel indicates wedge identification
Stray Radiation	Compliant with IEC standards
Beam Monitors	Redundant independent systems required Both systems show on control panel until reset Retrievable in case of power failures Count up
Beam Symmetry Monitor	Can detect asymmetry $> 10\%$ Terminates beam with asymmetries $> 10\%$
Beam Control	Beam initiation requires monitor setting Preset displayed Reinitiation requires clearing of setting Monitor unit rate is displayed Provides termination for excess dose rate
Termination of beam	By monitor systems at respective preset Manually at panel By timer after preset time with reset necessary
Radiation selection	Type of radiation must be selected (if more than one available) Interlocks prevent simultaneous types Type displayed on panel Interlocks prevent inappropriate beam type and accessories Special mode allows X-rays for imaging with electron applicators
Energy Selection	Energy selection required Energy displayed on panel Interlock prevents beam without appropriate mechanical conditions
Stationary or moving beam	Selection required Mode indicated on panel Interlocks prevent beam in inappropriate condition
Moving beams	Beam controlled for dose per degree Interlocks stop beam if dose per degree off

Table 21a. Highlights of Requirements for All diagnostic X-Ray Units[a–c]

Warning label	Attached to the control panel containing main power switch
Battery charge indicator	Visual on control panel if relevant
Source leakage radiation	< 1 mGy/m^2 at maximum technique for 1 h
Radiation other than tube	< 20 μGy/h 5 cm
Half-value layer[d]	$>$ values in Table 2dx including all material between tube and patient For variable filter units, control prevents incorrect selection
Multiple tubes	Selection of tube clearly indicated
Mechanical Support of the tube head	Hear remains stable during exposure (except dynamic studies)
Technique indication	Technique factors shown before exposure
Locks	Function properly

[a]Table by Tim Burns and Mark Geurts.
[b]For new units. Older units have some allowances made for regulations in effect at manufacture.
[c]Details for such units should be found in 21CFR1020 or in the particular state's regulations.
[d]Based in FDA regulations in 21CFR 1020.

Table 21b. Half–Value Layer Requirements

Operating Range	Measured Potential, kVp	Half-Value Layer in mm Aluminum: Diagnostic X-Ray Systems	Half-Value Layer in mm Aluminum: Dental Intraoral
<51	30	0.3	N/A[a]
	40	0.4	N/A[a]
	50	0.5	1.5
51–70	51	1.2	1.5
	60	1.3	1.5
	70	1.5	1.5
>70	71	2.1	2.1
	80	2.3	2.3
	90	2.5	2.5
	100	2.7	2.7
	110	3.0	3.0
	120	3.2	3.2
	130	3.5	3.5
	140	3.8	3.8
	150	4.1	4.1

[a]Not available = NA.

apply to all radioactive materials regardless of their origin. Others recognize that most accelerator-produced radionuclides tend to have shorter half-lives, and therefore require less control. Thus, when dealing with accelerator-produced material, consultation with the particular state's regulations becomes imperative.

REGULATIONS FOR NONIONIZING RADIATION

Understanding and applying NIR regulatory standards and guidance requires careful attention to the spectral characteristics of the radiation source(s) involved. The situation is probably most clearly described by dividing

Table 22. Highlights of Requirements for Fluoroscopic Units[a–c]

Primary barrier	Primary barrier intercepts entire beam Transmission $\leq 0.2\%$
Beam limitation	Beam not exceed visible area by >3%SID[d] Sum of excess < 4%SID Beam < largest spot-film size Units with visible area $> 300\,cm^2$ shall have continuously adjustable collimators, down to $5 \times 5\,cm^2$, or, if fixed SID, to $125\,cm^2$
Spot-film beam limitation	Beam automatically limited to film size Beam adjustable to fields smaller than film size down to $5 \times 5\,cm^2$ Beam not exceed visible area by >3%SID[d] Sum of excess < 4%SID Misalignment of the centers of beam and film < 2% SID
Activation of fluoroscopy	Requires continuous press on switch Serial exposures may be terminated at any time
Entrance exposure rates	≤ 50 mGy/minute, except 1. if unit has no high mode for automatic exposure control (AEC) units, then ≤ 100 mGy·min^{-1}; 2. during image recording
Indications	Panel shows kVp and mA during exposure
Source-to-skin distance	$\leq$38 cm for stationary units$\leq$30 cm for mobile units $\leq$20 for mobile, special surgical units
Fluoro timer	Maximum time without resetting $\leq$5 min Signals during fluoroscopy after time until reset
Control of scatter	Unit and table design prevent exposure of persons to scatter, except extremities, without ≥ 0.25 mm Pb equivalent attenuation or 1.2 m from beam

[a]Table by Tim Burns and Mark Geurts.
[b]For new units. Older units have some allowances made for regulations in effect at manufacture.
[c]Details for such units should be found in 21CFR1020 or in the particular state's regulations.
[d]SID is source to image intensifier distance.

Table 23. Characteristics of a Mammography System as Required by the Mammography Quality Standards Act

Item	Criterion
Type of equipment	Specially designed for mammography
Motion of tube-image receptor	Tube-image receptor may be fixed and remain so if power fails.
Image receptor size and grid	i. Screen-film units shall have a minimum of $18 \times 24\,cm^2$ and $24 \times 30\,cm^2$ and moving grids. ii. Magnification units can operate without the grid.
Light fields	Units with light fields shall have an average illumination of not less than 160 lux at 100 cm or the maximum source-image receptor distance, whichever is less.
Magnification	i. Units used for non-interventional problem solving shall have radiographic magnification capability ii. Units with magnification shall provide at least one magnification value between 1.4 and 2.0
Focal Spot selection	Unit indicates focal spot size and material selected prior to exposure, unless determined by algorithm during, where displayed after.
Compression	Unit shall have compression: i. power driven by hands-free controls on each side of the patient, including fine control;[a] ii. Compression paddle size shall match the full-field receptor size, and shall be level with the breast-support table to < 1 cm, except when designed otherwise; iii. The chest-wall edge shall be strain and parallel to the edge of the receptor, and may be curved for comfort if out of the field.
Technique factor selection and display	Has manual selection of mAs; ii. technique factors set display before exposure; iii. In automatic exposure control mode the technique used for the exposure displays afterwards.
Automatic exposure control (AEC)	i. Screen-film systems shall provide an AEC mode that is operable in all combinations of equipment; ii. positioning of detector shall permit flexibility in the placement under the target with the size and available positions of the detector marked on the input surface of the paddle, and the selected position of the detector indicated; iii. there shall be means to vary the selected optical density from the normal.
Film-Intensifying screens, if used	i. Film shall be designed for mammography; ii. screens shall be designed for mammography and the film used; iii. processing chemicals use as per manufacturer; iv. hot-lights and film masking devices shall be available;

[a]Applies to units built after 10/02.

NIR into three spectral regions: optical: ultraviolet (UV), visible, and infrared (IR); microwave RF; extremely low frequency (ELF) and static fields. The relationship between wavelength, frequency, and photon energy for these spectral regions is shown in Fig. 4.

Emission limits for specific optical sources combine with exposure limits to protect workers and the general public. These limits are further organized according to the type of source: lamps and other optical sources; and lasers. This separate consideration of non-laser sources and lasers necessarily reflects the different qualities of these sources. While lamps and other optical sources typically present a broad spectrum (i.e., the radiation is spread over many wavelengths) and widely divergent emission, lasers emit just one or at most a few discrete wavelengths in a very narrow, highly collimated beam.

The exposure limits for broad band non-laser optical sources are generally expressed in terms of some sort of spectral weighting scale to account for the fact that some wavelengths more efficiently cause injury than others. In the UV region, the International Commission on Non-ionizing Radiation Protection (ICNIRP) (29,30) and the National Institute for Occupational Safety and Health (NIOSH) (31) support the spectral weighting function and exposure limits set forth by the American Conference of Governmental Industrial Hygienists (ACGIH) (32) several decades ago. In this scheme, the weighting function is normalized to the peak of the spectral effectiveness curve at 270 nm, with an effective spectrally weighted limit of $30\,J \cdot m^{-2}$ over the region from 180 to 400 nm. Eye and skin exposure limits for monochromatic UV sources can be found in tables provided by ACGIH, ICNIRP, or NIOSH : The

Table 24. Quality Assurance Required for a Mammography System by the Mammography Quality Standards Act

	Daily Quality Control Tests for Film Systems
Film processor control	i. Base plus fog density within 0.03 of the established level; ii. mid-density on test strip within 0.15 of the established level; iii. density difference within 0.15 of the established level.
	Weekly Quality Control Tests for Film Systems
Phantom density and contrast	With approved phantoms, optical density$\geq$1.2 from typical exposure, varies <0.2 from normal, passes imaging for phantom, and contrast changes <0.05 with standard test.
	Quarterly Quality Control Test for Film Systems
Film fixer clearance	Residual fixer in film <5 μgm$\cdot$cm^{-2};
Reject analysis	Repeat or reject rate changes $<2\%$ of the total films in analysis (otherwise determine the reason for change, corrective actions recorded and the results assessed).
	Semiannual Quality Control Tests for Film Systems
Darkroom fog	Darkroom fog shall not exceed 0.05 for 2 min exposure on the counter top
Screen-film contact	40 mesh screen on cassette shows no appreciable blurring.
Compression device performance	Device provides >111 Nt force (between 111 and 200 Nt[a])
	Annual Quality Control Tests for Film Systems
Automatic exposure control performance	i. The AEC maintains optical density within 0.30 (or 0.15[a]) of mean (thickness varied from 2 to 6 cm and kVp varied appropriately for thicknesses); ii. optical density in center of phantom image >1.2.
kVp accuracy and reproducibility	i. Indicated kVp accurate within 5% at: the lowest clinical kVp that can be measured by a kVp test device, the most commonly used clinical kVp, and the highest available clinical kVp; ii., the coefficient of variation of reproducibility of the kVp ≤ 0.02 at the most commonly used clinical settings
System resolution	High contrast pattern resolves 11 line pair/mm with bars perpendicular to anode–cathode axis, and 13 when parallel; pattern 4.5 cm above breast support, centered, 1 cm of chest edge; test performed for each focal spot and target material.
Half-value layer (HVL)	HVL in mm$\geq$kVp/100
Breast entrance air kerma and AEC reproducibility	Coefficient of variation for both air kerma and mAs$\leq$0.05
Dosimetry	Average glandular dose (cranio-caudal view standard breast) $\leq$3 mGy/exposure.
X-ray field/light field/image receptor/compression paddle alignment	i. System has beam-limiting devices that allow the entire edge field to extend to the chest wall edge of the receptor and assure that the x-ray field does not extend beyond any edge of the receptor $>2\%$ of the SID; i. misalignment of the light and x-ray field $\leq 2\%$ of the SID; iii. chest wall edge of compression paddle $<1\%$ of the SID beyond the chest wall edge of the receptor.
Uniformity of screen speed	i. Difference between the maximum and minimum optical densities of all screens$\leq$ 0.30; ii. screen artifacts shall also be evaluated during this test.
System artifacts	System artifacts shall be evaluated for all available focal spot sizes and target filter combinations with a sheet of homogeneous material to cover the, for all cassette sizes used.
Radiation output	System[a] can produce >7 mGy$\cdot$s^{-1} air kerma at 28 kVp in Mo target/Mo filter mode at any SID with a detector 4.5 cm above the breast support surface with the compression paddle in place, over 3 s.
Automatic decompression, if included	System provides override capability to allow maintenance of compression, a continuous display of the override status, and a manual emergency compression release that can be activated in the event of power or automatic release failure.

[a]For units after built 10/02.

Table 25. Some Exemptions from Regulation Control for Naturally Occurring Radionuclides

Incandescent gas mantles	Marine compasses
Vacuum tubes	Timepieces, dials (various limits)
Welding rods	Lock illuminators
Electric lamps (< 50 mg thorium)	Precision balances
Glassware (< 10%t source material)	Automobile shift quadrant
Outdoor or industrial germicidal lamps, sunlamps, lamps (< 2 gm thorium)	Thermostat dials
Glazed ceramic tableware (glaze <20% source material)	Uranium as counterweights in aircraft, rockets, projectiles, missiles (labeled)
Piezoelectric ceramic(< 2% source material)	Electron tubes
Photographic film, negatives, and prints	Spark gap irradiators
Finished optical lenses (< 30% thorium)	Gas and aerosol detectors
Aircraft engine parts (nickel-thoria alloy < 4% thorium)	

limits are very similar for all three organizations. In addition to this general exposure limit, standard setting groups have established several hazard-specific limits spanning the entire optical region. The standard of ANSI and the Illumination Engineering Society of North America (IESNA) provides a typical treatment (33). This standard describes the application of exposure limits for the following hazards:

200–400 nm Skin and Eye Exposure Limit: very similar to the ACGIH, ICNIRP, and NIOSH ultraviolet limit.

320–400 nm Eye Exposure Limit: 1 $mW \cdot cm^{-2}$ for exposure durations >1 000 s, and 1 $J \cdot cm^{-2}$ for shorter exposure durations.

400–1400 nm Retinal Thermal Hazard Exposure Limit: an exposure duration dependent limit based on an associated burn hazard spectral weighting function table.

400–700 nm Retinal Blue Light Hazard Exposure Limit: a limit that is time dependent for exposure durations <10,000 s, and exposure rate dependent for longer durations, with an associated blue light hazard spectral weighting function table.

Retinal Blue Light Hazard Exposure Limit—Small Source: basically the Retinal Blue Light Hazard limit modified to accommodate sources subtending an angle <11 mrad

Aphakic Eye Hazard Exposure Duration: extends the Retinal Blue Light Hazard down to 305 nm and replaces the blue light hazard spectral weighting function with an aphakic hazard weighting function.

770–3,000 nm Infrared Hazard Exposure Limit: 0.1 $W \cdot cm^{-2}$ for exposure durations <1000 s and $1.8t^{-0.75}$ $W \cdot cm^{-2}$ for shorter exposure durations.

770–1400 nm Infrared Radiation Hazard Exposure Limit – Weak Visual Stimulus: where the visual stimulus may not activate the aversion response.

400– 3,000 nm Skin – Thermal Hazard Exposure Limit: provides limits for thermal skin exposure.

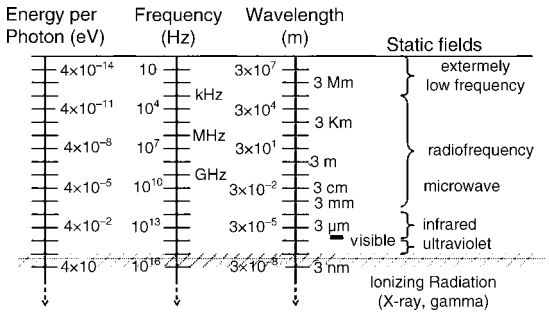


Figure 4. Nonionizing radiation: Wavelength, frequency, and photon energy for optical, microwave, radio frequency, and ELF spectral regions.

The various references caution users that to recall that several compounds (e.g., tetracycline and its congeners, porphyrins, Retin-A.) can make individuals more susceptible to the photochemical damage associated with UV radiation and blue light. Some common drugs that promote photochemical reactions are listed in Table E1 of the ANSI Z136.1-2000 standard. (34)

Lasers present a unique set of optical radiation hazards and consequently detailed exposure limits have been developed. Laser exposure limits are based on wavelength and various beam characteristics (e.g., beam diameter and divergence, plus pulse characteristics for pulsed lasers). While the OSHA occupational safety and health standards (Part 1910 of Title 29 CFR) do not specifically address lasers, Safety and Health Regulations for Construction (29 CFR Part 1926.54) provides some very general requirements (e.g., documented training of laser operators, posting of laser use areas, laser protective eyewear for workers in areas where the laser radiation level could exceed 5 $mW \cdot cm^{-2}$). The OSHAs construction standard also specifies non-wavelength specific (though presumably visible "light") laser radiation exposure limits of 0.001 $mW \cdot cm^{-2}$ for direct staring, 1 $mW \cdot cm^{-2}$ for incidental viewing, and 2.5 $mW \cdot cm^{-2}$ for diffuse reflections. Figure 5 illustrates the strong wavelength dependence of the ANSI Z136.1-2000[k] eye maximum permissible exposure (MPE) irradiance for continuous wave lasers under default exposure durations. Pulsed laser MPEs are more complicated to calculate, especially for repetitively pulsed lasers. See Thomas et al. for guidance on performing these repetitively pulsed laser MPE calculations(35). The ANSI Z136.1-2000 standard provides a framework for calculating the MPE for lasers occupying the wavelength region between 0.18 μm and 1 mm. The time domain covered extends to pulse

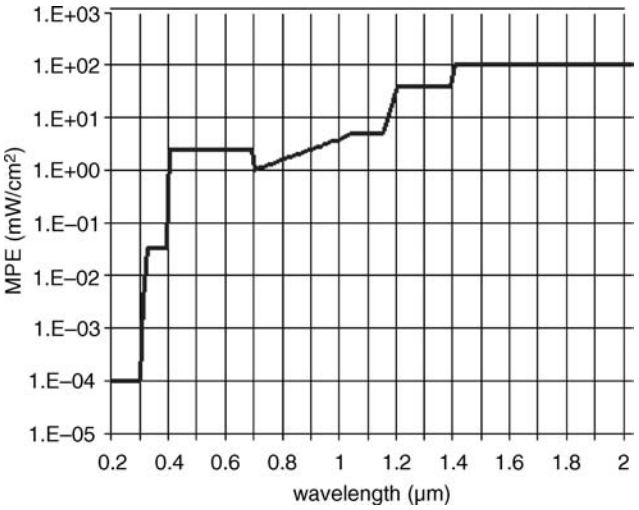


Figure 5. Maximum permissible exposure (MPE) versus wavelength for continuous wave lasers, assuming extended viewing duration for UV wavelengths (<0.4 μm), only accidental viewing (<0.25 s) for visible wavelengths (0.4–0.7 μm), and 10 s exposure for wavelengths (>0.7 μm). For clarity this chart extends only to 2 μm, but the MPE remains 100 mW·cm^{-2} for wavelengths from 1.4 to 1 mm.

durations as short as 10^{-9} s for UV (<0.4 μm) and far IR (<1.4 μm), and down to 10^{-13} s for the visible and near-IR regions (i.e., 0.4–1.4 μm). However, laser technology has pushed beyond these boundaries, with pulse durations in the 10^{-15} s range and shorter wavelength laser-like sources operating near the threshold of ionizing radiation now attainable in the laboratory. Subsequent editions of the Z136.1 standard are expected to address these gaps.

The FDA Laser Product Performance Standard (21CFR1040.10&11) requires laser manufacturers to classify laser products sold in the United State into one of four hazard classes according to the laser's ability to cause injury. The ANSI Z136 series of standards then specifies the required control measures appropriate for each class of laser. The IEC/EN 60825 series of standards establishes a similar laser hazard classification scheme for both manufacturers and application of control measures. Table 26 summarizes, in very general terms, the classes and subclasses for each of these organizations. In an apparent nod to pending harmonization, the recent Z136.3-2005 Standard for the Safe Use of Lasers in Health Care Facilities utilizes the IEC classification designations instead of the ANSI/FDA scheme (36). The laser hazard classes established by each of these three standard setting organizations also have associated accessible emission limits (AEL), but the AEL for each hazard class may not be identical for each organization.

The microwave region of the frequency spectrum is generally considered to extend from 300 MHz to 300 GHz, while the RF portion covers from 3 kHz to 300 MHz. (32,37) An opportunity for nomenclature confusion arises from the fact that some organizations, including ICNIRP (38), consider this microwave region to be a subset of an RF range extending from 300 Hz to 300 GHz. See Figure 4 for a sense of relative position of these regions in the electromagnetic spectrum. The FDA/CDRH Microwave Oven standard (21CFR1030.10) defines a microwave oven as a device designed to heat, dry, or cook food using electromagnetic radiation from 890 to 6000 MHz (most commercial microwave ovens operate at either 915 or 2450 MHz), and requires microwave oven manufacturers to limit radiation levels to <5 mW·cm^{-2} at any point 5 cm or more from any external surface. OSHAs Construction standard (29 CFR 1926.54) limits microwave exposure to 10 mW·cm^{-2} (no averaging), but offers no other guidance in this region. The NIR section of OSHA's General Industry standard (29 CFR 1910.97) specifies a 10 mW·cm^{-2}, 6 min average for exposure limit for the 10 MHz–100 GHz range, with no spatial averaging, and prescribes RF warning sign appearance. This OSHA section applies to all radiations originating from radio stations, radar equipment, and other possible sources of electromagnetic radiation (e.g., as used for communication, radio navigation, and industrial and scientific purposes), but does not apply to the deliberate exposure of patients by health care providers. In addition, section 29 CFR 1910.268 [Telecommunications] provides guidance for "lock out tag out" type securing and grounding prior to working on 3–30 MHz radio station broadcast antennas, prohibits employers from allowing employees to look into energized microwave waveguides, requires compliance with the 29 CFR 1910.97 exposure limits, and requires posting warning signs with specific wording in accessible areas where those limits could be exceeded. Table 27 gives some sample exposure limits.

Several standards setting organizations have developed exposure limits for microwave and radiofrequency radiation. These standards show considerable consistency for the spectral region extending from 300 GHz (i.e., near the boundary between optical and microwave radiation) to ~15 GHz, where most standards recognize an exposure limit of 10 mW·cm^{-2}. At lower frequencies, additional considerations come into play, as indicated in Table 28. At frequencies below ~10 MHz, the longer wavelengths necessitate separate limits for magnetic and electric field exposures. This distinction arises because at longer wavelengths the emitted energy from an antenna passes first through a *near field* region, in which the electric and magnetic fields are not directly coupled as they are under the *far field* conditions normally associated with electromagnetic radiation. In the far field, the electric field strength is always directly proportional to the magnetic field strength, and the radiation levels generally decrease with the inverse square of the distance. In the near field, the electric and magnetic fields are not coupled, more complicated relationships with distance exist, and many far field assumptions no longer hold. The distance that the near field extends from the emitter is often defined as the wavelength divided by 2π, although ICNIRP defines the near field simply as the distance within one wavelength from a radiating antenna (38). By using the ICNIRP expression and consulting Fig. 3, we see, for example, that while at 10 MHz the near field extends only 30 m, at 100 kHz it is 3 km. The ICNIRP exposure limits (reference levels) appear in Table 34 (38,39), expressed as separate electric and magnetic field strengths (the ICNIRP guidance actually expresses exposure limits in terms of

Table 26. Summary of Laser Classification Schemes

FDA/CDRH 21CFR1040.10	ANSI Z136	IEC/EN 80625
Class I—levels of laser radiation are not considered hazardous	Class 1—no hazard; exempt from all control measures	Class 1—no risk, even with viewing instruments Class 1M[a]—no risk except perhaps to eye when viewed through viewing instruments (eye loupes or binoculars)
Class IIa—levels of laser (applies to visible only) radiation are not considered hazardous if viewed $\leq$1000 s but are considered a chronic viewing hazard for any period of time >1000 s Class II—levels of (visible only) laser radiation considered a chronic viewing hazard	Class 2—visible (0.4–0.7μm) lasers not considered hazardous for momentary viewing (<0.25 s), but for which the Class 1 accessible emission limit may be exceeded for longer exposure durations; avoid prolonged staring	Class 2—no eye risk for short term exposures, even with viewing instruments; no risk to skin (applies to visible lasers only) Class 2M[a]–no eye risk for short term exposures, except perhaps with viewing instruments; no risk to skin (visible only)
Class IIIa—levels of laser radiation are considered, depending upon the irradiance, either an acute intrabeam viewing hazard or chronic viewing hazard, and an acute viewing hazard if viewed directly with optical instruments	Class 3a—with "Caution" label: does not exceed the appropriate irradiance MPE, except perhaps when viewed through collecting optics (e.g., microscopes, telescopes)—with "Danger" label: may exceed the appropriate irradiance MPE	Class 3R[b]—low risk to eyes, no risk to skin
Class IIIb—levels of laser radiation are considered to be an acute hazard to the skin and eyes from direct radiation	Class 3b—emit greater than Class 3a limits and pose an acute eye hazard; more rigorous controls are required to prevent exposure of the unprotected eye	Class 3B—medium to high risk to eyes, low risk to skin
Class 4—levels of laser radiation are considered an acute hazard to the skin and eyes from direct and scattered radiation	Class 4—acute eye and skin hazard, plus ignition source (fire) and laser-generated airborne contaminants hazards; strict control measures required	Class 4—high risk to eyes and skin

[a]The "M" designation in the IEC classification scheme is derived from "magnifying" optical viewing instruments.
[b]The "R"designation in the IEC classification scheme is derived from reduced or relaxed requirements for manufacturers (no key switch or interlock connector required) and users (usually no eye protection required).

quantities that are not easily measured directly, such as currents induced in the body; the references levels then extrapolate those limits via models to the directly measurable field strengths given).

The microwave and RF exposure limits are generally based on protection against known adverse health effects of NIR exposure, but these limits may not be sufficiently protective for individuals with implanted medical devices (e.g., pacemakers, insulin pumps, cochlear implants, etc.) (38). Also, in the ELF region (defined here as by the ICNIRP as <300 Hz (40), although some organizations and regulations specify slightly different spectral boundaries), considerable ambiguity marks the basis underlying the generally accepted exposure limit values. As Petterson and Hitchcock note, exposure guideline derivation should ideally stem from accepted mechanisms of interaction, dose response studies in animals, and epidemiological evidence of similar effects in humans; none of this has occurred for ELF fields (41). This view was reiterated in a 2002 update of the American Industrial Hygiene Association (AIHA) White Paper on ELF Fields (42). Finally, accurately assessing non-ionizing radiation hazards is complicated by the difficulty in properly measuring RF and ELF field strength. The NIOSH *Manual for Measuring Occupational Electric and Magnetic Field Exposures* provides some insight into this subject (43).

Because of its shallow tissue penetration and ease of control, ELF electrical fields are generally not of much concern below $\sim 15\,\mathrm{kV \cdot m^{-1}}$. However, individuals with implanted medical devices (e.g., cardiac pacemakers

Table 27. Sample of Microwave and Radiofrequency Radiation Exposure Limits, 10 MHz–300 GHz[a]

Source	Reference	Frequency	Limit, $mW \cdot cm^{-2}$		Additional Criteria
			Occupational	*Public*	
OSHA	29 CFR 1926.54(l) [Construction]	0.3–300 GHz	10		(no averaging)
OSHA	29 CFR 1910.97(a)(2)(i) [General Industry]	0.01–100 GHz	10		Power (duration $\geq$ 0.1 h): 10 $mW \cdot cm^{-2}$ Energy (duration < 0.1 h): 10 $mW \cdot hr \cdot cm^{-2}$ in any 0.1 h (no spatial averaging)
ICNIRP	Guidelines up to 300 GHz (1998)[n]	2–300 GHz	50	10	Averaged over any 20 cm^2 exposed area and, above 10 GHz, any $68/f^{1.05}$ (f = frequency in GHz) minute period; Limit peak exposures averaged over 1 cm^2 < 20 times these limits
		400–2,000 MHz	$f/40$	$f/200$	f = frequency in MHz
		10–400 MHz	10	2	Between 100 kHz and 10 GHz, averaged over any 6 minute period

[a]We thank the International Commission on Non-Ionizing Radiation Protection, ICNIRP, for the permission to reprint part of its guidelines in the present article. We also thank Health Physics, where the guidelines were first published.

Table 28. ICNIRP Electrical and Magnetic Field Exposure Limits, Static—10 MHz[a]

Reference	Frequency	Reference Level (Exposure Limit)				Additional Criteria
		Occupational		*Public*		
		Electrical $V \cdot m^{-1}$	Magnetic $A \cdot m^{-1}$	Electrical $V \cdot m^{-1}$	Magnetic $A \cdot m^{-1}$	
ICNIRP Guidelines up to 300 GHz (1998)[b]	1–10 MHz	$610/f$	$1.6/f$	$87/f$	$0.73/f$	f = frequency in MHz
	0.065–1 MHz	610	$1.6/f$			f = frequency in MHz
	0.15–1 MHz			$87/f$	$0.73/f$	f = frequency in MHz
	0.82–65 kHz	610	24.4			
	3–150 kHz			87	5	
	0.025–0.82 kHz	$500/f$	$20/f$			f = frequency in kHz
	0.025–0.8 kHz			$250/f$	$4/f$	f = frequency in kHz
	8–25 Hz	20,000	$20{,}000/f$	10,000	$4{,}000/f$	f = frequency in Hz
	1–8 Hz	20,000	$163{,}000/f$	10,000	$32{,}000/f$	f = frequency in Hz
	0– Hz	Use electrical safety procedures to avoid electric shock	163,000	Avoid spark discharges	32,000	
ICNIRP Guidelines for Static Magnetic Fields (1994)[t]	Static		Average: 160,000 Maximum: 1,600,000 Limbs: 3,980,000		General: 33,000 Electronic medical implants: 400	

[a]We thank the International Commission on Non-Ionizing Radiation Protection, ICNIRP, for the permission to reprint part of its guidelines in the present article. We also thank Health Physics, where the guidelines were first published.
[b]1 tesla (T) = 796,000 amperes per meter (A/m) in air, vacuum and biological materials.

and insulin pumps) should avoid ELF electric fields above 1 $kV \cdot m^{-1}$. The static electric and magnetic field designation is generally applied to any radiator with a frequency below 1 Hz (37). Static electrical fields are relatively simple to shield with grounded conducting enclosures. Static magnetic fields are more difficult to shield. The NIR environment surrounding nuclear magnetic resonance imaging facilities (primarily a high strength static magnetic field, but also RF radiation) poses a particular concern. The American College of Radiology's White Paper on MR Safety (44) offers some useful guidance on control of the hazards associated with MRI systems. While the ACR White Paper has served to increase awareness and foster discussion, the user community has not achieved consensus on these recommendations (45,46).

The FDAs 21CFR892.1000 identifies MRI systems as Class II medical devices, meaning among other things that institutions must implement safety programs rather than merely following the manufacturer's recommendations. In addition, MRI manufacturers generally must provide the customers with information, for example, indicating the extent of the static magnetic field (i.e., the location of the 5 G line). Attachment B of the FDAs *Guidance for the Submission of Premarket Notifications for Magnetic Resonance Diagnostic Devices* lists the elements of a MRI

safety program, including patient screening, appropriate levels of patient monitoring and supervision, emergency procedures and shutdowns, noise control measures, access restrictions, control of cryogenic hazards, adherence to the IEC operating mode guidelines, use of MRI compatible equipment, fire precautions, and so on (47). The 5 G line around MRI facilities is generally posted with warning signs to avoid harmful effects on medical device wearers. There is no generally accepted format for MRI warning signs, though Shellock offers some helpful suggestions (48). The FDA guidance indicates that a MRI procedure performed under any of the following conditions constitutes a significant risk as defined in 21CFR812.3(m)(4), triggering the requirement for a FDA investigational device exemption, as well as the institutional review board (IRB) approval specified in 21CFR56 and the informed consent rules of 21CFR50 (49,50): procedures utilizing >8 tesla (T) for adults, children and infants older than 1 month, or >4 T for neonates (infants <1 month old); specific absorption rate (SAR) $\geq 4\ W \cdot kg^{-1}$ for 15 min or more for the whole body, $3\ W \cdot kg^{-1}$ for 10 min or more for the head, $8\ W \cdot kg^{-1}$ for 5 min or more per gram of tissue for the head or torso, or $12\ W \cdot kg^{-1}$ for 5 min or more per gram of tissue for the extremities; or any time rate of change of gradient magnetic fields sufficient to produce severe discomfort or painful nerve stimulation.

The IEC 60601-2-33 standard specifies three modes of operation relating to RF energy-induced heating and gradient magnetic field (51):

Normal operating mode: suitable for all patients and requires only routine monitoring;

First level controlled (FLC): may cause undue physiological stress; requires medical supervision and operator confirmation to enter this mode.

Second level controlled (SLC): may produce significant risk; requires IRB approval, and manufacturers are required to restrict operator access to this mode (e.g., password, key lock).

Although the FDA does not currently require MRI manufacturers to incorporate provisions for these three IEC modes, in the current global market all MRI manufacturers already include such provisions, and the FDA has indicated their intention to adopt these IEC 60601-2-33 guidelines (47,50).

CONCLUSIONS

Understanding codes, regulations, and license conditions can be the least exciting but most challenging part of a medical physicist's job! The federal codes may be the basis for state codes, but state codes are not necessarily identical to federal codes, even in NRC agreement states or OSHA-approved states. Compliance with myriad regulations and license conditions is a challenge. However, by knowing the codes and regulations one can write a better license or structure a radiation safety program with which it is easier for one to comply. To be forewarned is to be forearmed! We encourage readers to thoroughly study those codes and regulations applicable to their own facilities and to stay current with continuingly changing codes and regulations.

ACRONYMS AND DEFINITIONS

Most acronyms are discussed in the text.

AAPM. The American Association of Physicists in Medicine.

ACGIH. American Conference of Governmental Industrial Hygienists.

ACMP. American College of Medical Physics.

ACR. American College of Radiology.

Agreement States. Those states that have entered into a formal agreement with the NRC to take over certain of its regulatory authority within that state.

AIHA. American Industrial Hygiene Association.

AMP. Authorized medical physicist.

ANSI. American National Standards Institute, the voluntary standards-setting organization in the United States and its representative to the International Organization for Standardization.

AU. Authorized user, a physician authorized to use radioactive materials in the health professions.

BEIR. Biological Effects of Ionizing Radiation, often referring to reports bearing that acronym from the BRER.

BRER. Board on Radiation Effects Research, a board of the National Academy to coordinate activities of the National Research Council involving the biological effects of radiation.

Byproduct material. Any radioactive material (except special nuclear material) yielded in or made radioactive by exposure to the radiation incident to the process of producing or utilizing special nuclear material.

CDRH. Center for Devices and Radiological Health, and agency of the Food and Drug Administration.

CGMP. Current good manufacturing practices.

CRCPD. Coference of Radiation Control Program Directors.

Department of Defense

DOE. Department of Energy.

DOT. Department of Transportation.

DS. Dosimetry system.

EDE. Effective dose equivalent.

ELF. Extremely low frequency.

EPA. Environmental Protection Agency.

FC. Full calibration.

FDA. Food and Drug Administration.

HDR. High dose-rate brachytherapy.

HPS. Health Physics Society.

IAEA. International Atomic Energy Agency, an arm of the United Nations.

ICNIRP. International Commission on Nonionizing Radiation Protection.

ICRP. International Commission on Radiation Protection.

ICRU. International Commission on Radiation Units and Measures.

IEC. International Electrotechnical Commission, an organization that establish standards mostly pertaining to industry and manufacturers.

IR. Ionizing radiation.

IRB. Institutional Review Board, an institutions committee that evaluates new drugs, procedures or devices for proposed human use.

JCAHO. Joint Commission on Accreditation of Healthcare Organizations.

LDR. Low dose-rate brachytherapy.

MDR. Medium dose-rate brachytherapy.

ME. Medical event (see Table 14).

MQSA. Mammography Quality Standards Act.

NARM. Naturally occurring and accelerator-produced radioactive materials.

NCRP. Nation Council on Radiation Protection and Measurement.

NIOSH. National Institute for Occupational Safety and Health.

NIR. Non-ionizing radiation.

NRC. Nuclear Regulatory Commission.

NSSDR. National Sealed Source and Device Registry, a registry maintained by the NRC of information on approved devices.

OAS. Organization of Agreement States.

OSHA. Occupational Health and Safety Administration.

PD. Prescribed dose.

PDR. Pulsed dose-rate brachytherapy.

PMA. Premarket approval.

RAM. Radioactive materials.

RAU. Remote afterloading unit for brachytherapy.

RF. Radiofrequency.

RSO. Radiation safety officer.

SAU. Substitute authorized user, a physician working under the supervision of an authorized user.

SC. Spot check.

SDE. Shallow dose equivalent.

Source Material. Uranium, thorium or any combination thereof, in any physical or chemical form, or ores that contain by weight one-twentieth of 1% of them, excluding special nuclear material.

Special nuclear material (SNM). Plutonium, uranium-233, uranium enriched in isotopes 233 or 235, or any material artificially enriched by any of these.

SSR. Suggested State Regulations, guidelines for state radiation rules assembled by the CRCPD.

UNSCEAR. United Nations Committee on the Effects of Atomic Radiation, a committee under the United Nations to study the biological effects of radiation.

UV. Ultraviolet.

BIBLIOGRAPHY

1. Deye JA. Codes and Regulations, Radiation. In: Webster JG, editor Encyclopedia of Medical Devices and Instruments. New York: John Wiley & Sons, Inc; 1988.
2. U.S. Nuclear Regulatory Commission. 10 CFR 19 (Notices, Instructions, and Reports to Workers; Inspections), 1981 [Online]. Nuclear Regulatory Commission. Available at http://www.nrc.gov/reading-rm/doc-collections/cfr/part 0.19. Accessed 2004, December 4.
3. U.S. Nuclear Regulatory Commission. 10 CFR 20 (Standards for Protection Against Radiation; Final Rule,) 1991. [Online]. Nuclear Regulatory Commission. Available at http://www.nrc.gov/reading-rm/doc-collections/cfr/part020. Accessed 2004, December 4.
4. U.S. Nuclear Regulatory Commission. 10 CFR 20, 32, and 35 (Medical use of byproduct material: final rule.) Washington DC, Federal Register; Vol.67, No.79 (April 24): 20250-20397, 2002; [Online]. Nuclear Regulatory Commission. Available at http://www.nrc.gov/reading-rm/doc-collections/cfr/part020/part032/035. Accessed 2004, December 4.
5. U.S. Nuclear Regulatory Commission. 10 CFR 35 (*Medical use of byproduct material-Recognition of Specialty Boards; Final Rule*) Washington DC, Federal Register; Vol.70, No.60 (March 30): 16366-16367, 2005; [Online]. Nuclear Regulatory Commission. Available at http://www.nrc.gov/reading-rm/doc-collections/cfr/part35. Accessed 2005, April 19.
6. Institute of Medicine. Radiation Medicine: A Need for Regulatory Reform. In: Gottfried K-LD, Penn G. editors. Committee for Review and Evaluation of the Medical Use Program of the Nuclear Regulatory Commission, Institute of Medicine: 1996 ISBN-0-309-58875-8.
7. Memorandum of Understanding Between the OSHA and U.S. Nuclear Regulatory Commission. CPL 02-00-086–CPL 2.86 (1989, December 22) [Online]. Occupational Safety & Health Administration. Available at http://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=DIRECTIVE&p_id=1658. Accessed 2005, April 29.
8. Federal Register [07/23/2004] 69: 44068-44069; Supporting Statement for the Information-Collection Requirement for the Ionizing-Radiation Standard (29CFR1910.1096) OMB Control No. 1218-0103(2004)(June2004).
9. Occupational Safety and Health Administration. 29 CFR 1910.1096 (Ionizing Radiations), [Online] Occupational Safety and Health Administration. Available at http://www.osha.gov/pls/oshaweb/owadisd.show. Accessed 2005, April 14.
10. Conference of Radiation Control Program Directors. Available at http://www.crcpd.org/free_docs.asp. Accessed 3 October 2005.
11. Curtis R. Non-ionizing radiation: standards and radiation (PowerPoint presentation). OSHA-Salt Lake Technical Center. [Online] Available at http://www.osha-slc.gov/SLTC/radiation_lectures/nir_stds_20021011.ppt. Accessed April 22, 2005.
12. Rockwell RJ, Parkinson J. State and local laser safety requirements. J Laser Appl 1999;11:225–231.
13. Henderson R, Schulmeister K. Laser Safety Philadelphia: Institute of Physics Publishing; 2004.

14. Center for Devices and Radiological Health. Laser Products – Conformance with IEC 60825-1, Am. 2 and IEC 60601-2-22; Final Guidance for Industry and FDA (Laser Notice 50), 2001.
15. News and Notices IAEA action plan to combat nuclear terrorism Health Phys 2002;82:908–909.
16. News and Notices "IAEA and UPU join forces to protect mail" Health Phys 2003;84:129–130.
17. Nuclear Energy Institute, Fact Sheet, August 2004.
18. U.S. Nuclear Regulatory Commission, U.S. Department of Transportation, U.S.-Specific Schedules of Requirements for Transport of Specified Types of Radioactive Material Consignments, RAMREG-002/U.S. Nuclear Regulatory Commission, NUREG-1600, January 1999.
19. Hazardous Materials Regulations; Compatibility With the Regulations of the International Atomic Energy Agency, RSPA-99-6283 (HM-230); Final rule; Published 01/26/2004; Effective Date: Oct 1, 2004; 69 FR 3631.
20. International Atomic Energy Agency, Regulations for the Safe Transport of Radioactive Material, 1996 Edition, Safety Standards Series/Requirements, ST-1, Vienna, Austria; International Atomic Energy Agency, 1996.
21. International Commission on Radiation Protection, 1990 Recommendations of the ICRP: ICRP Publication 60, Ann of the ICRP 21: 1-3 Oxford: Pergamon Press; 1991.
22. USNRC Potential for Erroneous Calibration, Bulletin 97-01. Dose Rate, or Radiation Exposure Measurements with Certain Victoreen Model 530 and 530SI Electrometer/Dosemeters (April 30, 1997).
23. Glasgow GP. Nuclear Regulatory Commission regulatory status of approved intravascular brachytherapy systems. Cardiovascular Rad Med 2002;3:1–11.
24. USNRC, 2004.
25. U.S. Nuclear Regulatory Commission . NRC Information Notice 2003-21: High-dose rate remote afterloader equipment failure. November 24, [Online] Available at http://www.nrc.gov/materials/miau/med-use-toolkit/info-notices.html. Accessed December 4, 2004.
26. U.S. Nuclear Regulatory Commission. Newsletter NUREG/BR-0117/04-2: Nuclear Material Safety and Safeguards [Online] Available at http://www.nrc.gov/reading-rm/doc-collections/nureg/brochures/br0117/04-2.pdf. Accessed December 4, 2004.
27. USNRC, Compatibility Categories and Health and Safety Identification for NRC Regulations and Other Program Elements—Procedure No. SA-200, 2004 [Online] Available at http://www.hsrd.ornl.gov/nrc/procintro.htm. Accessed 3 October 2005.
28. Conference of Directors of Radiation Control Programs, Suggested State Radiation Control Regulations. Avaliable at http://www.crcpd.org/SSRCRs. Accessed 3 October 2005.
29. International Commission on Non-Ionizing Radiation Protection. Guidelines on Limits of Exposure to Ultraviolet Radiation of Wavelengths Between 180 nm and 400 nm (incoherent optical radiation Health Phys 2004;87:171–186.
30. International Commission on Non-Ionizing Radiation Protection. Guidelines on UV Radiation Exposure Limits. Health Phys 1996;71:978.
31. National Institute for Occupational Safety and Health. Criteria for a Recommended Standard–Occupational Exposure to Ultraviolet Radiation; DHHS (NIOSH) Publication No. 73–11009, 1972.
32. American Conference of Governmental Industrial Hygienists. 2004 TLVs® and BEIs®. Cincinnati, OH: ACGIH, 2004.
33. American National Standards Institute/Illuminating Engineering Society of North America. RP-27.I-96: Recommended Practice for Photobiological Safety for Lamp & Lamp Systems - General Requirements. New York: IESNA; 1996.
34. American National Standards Institute, Inc., ANSI Z136.1-2000, American national standard for safe use of lasers. Orlando: (FL): Laser Institute of America; 2000.
35. Thomas RJ, et al. A procedure for multiple-pulse maximum permissible exposure determination under the Z136.1-2000 American National Standard for the Safe Use of Lasers. J Laser Appl 2001 13:134–139.
36. American National Standards Institute, Inc., ANSI Z136.3-2005, American national standard for safe use of lasers in health care facilities. Orlando, FL: Laser Institute of America; 2005.
37. Hitchcock RT. Radio-Frequency and Microwave Radiation. In: DiNardi SR. editor. The Occupational Environment—Its Evaluation and Control. Fairfax, VA: AIHA Press; 1998.
38. International Commission on Non-Ionizing Radiation Protection. Guidelines for Limiting Exposure to Time-Varying Electric, Magnetic, and Electromagnetic Fields (Up to 300 GHz). Health Phys 1994;74:494–522.
39. International Commission on Non-Ionizing Radiation Protection. Guidelines on Limits of Exposure to Static Magnetic Fields. Health Phys 1994;66:100–106.
40. International Non-Ionizing Committee of the International Radiation Protection Association. Review of Concepts, Quantities, Units and Terminology for Non-ionizing Radiation Protection. Health Phys 1985;49:1329–1362.
41. Patterson RM, Hitchcock RT. Extremely Low Frequency (ELF) Fields. In: DiNardi SR, editor. The Occupational Environment–Its Evaluation and Control. Fairfax, (VA): AIHA Press; 1998.
42. American Industrial Hygiene Association. AIHA White Paper on Extremely Low Frequency Fields, Fairfax, VA: AIHA; 2002.
43. Bowman JD, Kelsh MA, Kaune WT. Manual for Measuring Occupational Electrical and Magnetic Field Exposures [DHHS (NIOSH) Publication No. 98-154]. Cincinnati, (OH): National Institute for Occupational Safety and Health Publications Dissemination; 1998.
44. Kanal E, et al. American College of Radiology white paper on MR safety. Am J Radiol 2002;178:1335–1347.
45. Shellock FG, Crues JV. MR Safety and the American College of Radiology White Paper. Am J Radiology 2002;178:1349–1352.
46. Kanal E, et al American College of Radiology White Paper on MR Safety: 2004 Update and Revision. Am J Radiol 2004;182:1111–1114.
47. Center for Devices and Radiological Health. Guidance for the Submission of Premarket Notifications for Magnetic Resonance Diagnostic Devices. CDRH, [Online]. Available at http://www.fda.gov/cdrh/ode/95.html. Accessed 5 October 2005.
48. Shellock FG. MR Safety Signs. RT Image 16(3): 2003. Available at http://www.rt-image.com/content=8702J84E489CAE9040A-240441. Accessed 30 Sept 2005.
49. Center for Devices & Radiological Health. Guidance for Industry and FDA Staff - Criteria for Significant Risk Investigations of Magnetic Resonance Diagnostic Devices. CDRH, 2003.
50. International Commission on Non-Ionizing Radiation Protection. Medical magnetic resonance (MR) procedures: protection of patients. Health Phys 2004;87:197–216.
51. Zaremba LA. FDA Guidance for magnetic resonance system safety and patient exposures: current status and future G considerations. In: Shellock FG, editor. Magnetic resonance procedures: health effects and safety, Boca Raton, (FL): CRC Press LLC; 2001.

See also CODES AND REGULATIONS: MEDICAL DEVICES; IONIZING RADIATION, BIOLOGICAL EFFECTS OF; NONINONIZING RADIATION, BIOLOGICAL EFFECTS OF; RADIATION PROTECTION INSTRUMENTATION.

COGNITIVE REHABILITATION. See REHABILITATION, COMPUTERS IN COGNITIVE.

COLORIMETRY

LI-JIUAN SHEN
National Taiwan University
Taipei, Taiwan

RICHARD MANDEL
Boston University
Boston, Massachusetts

WEI-CHIANG SHEN
University of Southern California,
Los Angeles, California

INTRODUCTION

Light can be characterized as a wave with frequency and wavelength λ. This wave has energy E, which is proportional to its frequency

$$E = hc/\lambda = h\nu$$

where h is Planck's constant and c is the velocity of light. When such a wave of light encounters a molecule, it will either be absorbed (i.e., its energy will be transferred to the molecule) or scattered (i.e., its direction of propagation will be changed). The probability of occurrence of each process will depend on the nature of the molecule encountered. If the electromagnetic energy is absorbed, the molecule is said to be excited. A molecule or part of a molecule that can be excited by absorption is called a chromophore. The absorption of light generally excites the electrons of the molecule or chromophore from its ground electronic state to one of its excited electronic states. Absorption of light is likely to occur if the energy of the light is equal to the difference in the energy between the ground (E°) and excited (E^{*}) electronic state.

$$\lambda = hc/(E^{*} - E^{\circ})$$

These transitions have rather diffuse energies because excited states occur in both the vibrational and rotational as well as electronic states, giving rise to broadened energy ranges. A plot of the probability of absorption versus wavelength is called an absorption spectrum. The excitation energy is usually released as radiant energy in the form of kinetic energy and heat. Under certain conditions, some molecules rapidly reemit their energy as visible or ultraviolet (UV) light. This is known as fluorescence.

The wavelengths that give rise to electronic transitions in molecules are generally in the visible (700–400 nm) and UV (400–200 nm) region. These transitions are often characteristic of specific molecules and can be used to assay biological and biochemical samples. Transitions at longer wavelengths and lower energies in the infrared (IR) or near-infrared (NIR) generally correspond to excitation of vibrational states alone and are characteristic of specific functional groups such as carbonyl oxygen bonds or carbon–carbon bonds. Even though this region of the absorption spectrum does not find many biomedical applications, IR and NIR spectroscopy have been used for qualitative and quantitative analysis of different forms (crystal and amorphous) of pharmaceutical solids (1,2). It is further developed into an on-line process monitoring and control entitled process analytical technology (PAT) by the collaboration of U. S. Food and Drug Administration (FDA) and the pharmaceutical industry (3). On the other hand, transitions at shorter wavelengths and higher energies, in the vacuum UV and X-ray region, cannot be carried out on biological samples in aqueous solution and have limited usefulness in biomedical assays.

The probability of light absorption at a single wavelength is described by the Beer–Lambert law. It has been observed that the passage of light through any given thickness of any substance results in the absorption of a constant fraction of that incident light. In differential form, this equation can be written

$$dI/I = -KCdl$$

where dI/I is the fraction of light absorbed by a layer of thickness dl, K is a constant that depends on the properties of the substance, and C is the concentration of the absorbing substance. Thus, for a given beam of light passing through a sample of finite thickness, the amount absorbed at each point in that sample is proportional to the incident intensity at that point. If, for example, in the first millimeter of passage through a sample 50% of the light is absorbed, then in the second millimeter of passage 50% of the remaining light or 25% of the initial intensity will be absorbed, and in the third millimeter 50% of the remainder or 12.5% of the initial intensity will be absorbed. In total, 87.6% of the incident light beam will be absorbed by the 3 mm path length. Mathematically, this is calculated by integrating the above equation, which yields

$$ln(I_0/I) = K'Cl \quad \text{or} \quad I/I_0 = 10^{-K'Cl}$$

where $K' = K/2.303$, l is the path length, I_0 is the initial intensity, and I the final intensity of the light beam after having passed the sample. The left-hand side of the equation defines the optical density and is a useful quantity, because it is directly proportional to the concentration of the absorbing substance and the path length. Absorbance, also called optical density (OD), is designated by A. The transmittance of a solution, designated T, is the fraction (I/I_0) and is related to the absorbance by

$$A = -\log T$$

In the above example, the absorbance through a single millimeter path length is $-\log 0.5 = 0.301$, while the total absorbance through 3 mm is $3 \times 0.301 = 0.903 = -\log 0.125$.

A plot of absorbance as a function of concentration (called a Beer's law plot) is ideally a straight line, while a plot of transmittance is a hyperbolic curve (Fig. 1). Deviations from linearity will be discussed below. Absorbance is a unitless quantity. Therefore, the constant *(K′)* in the above equation must be in units of reciprocal concentration and reciprocal length (typically $cm^{-1} \cdot M^{-1}$). The absorbance is proportional to both concentration and

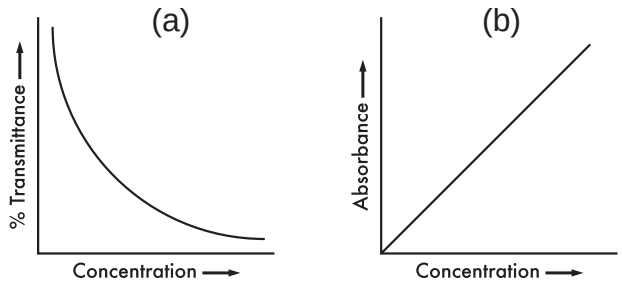


Figure 1. Plots of concentrations versus (a) transmittance (A) and (b) absorbance (B).

length of light path, while the constant specifies a characteristic property of the absorbing chromophore. This constant, which is a function of wavelength, is called the absorption coefficient or extinction coefficient of the material. Absorption coefficients may be expressed on either a weight or a molar basis. The molar extinction coefficient, designated by e specifies the total absorption of a 1 M solution through a 1 cm path length. A molar extinction coefficient of 10,000 or greater is characteristic of a strongly absorbing substance.

Plots of absorbance as a function of concentration are linear if the absorbing chromophore remains the same over the concentration range studied, the chromophore is uniformly distributed, and the orientation of dichroic absorbers (i.e., substances having directional asymmetry) is random. Changes in the nature of the chromophore with concentration, such as ionization, hydration, aggregation, or disaggregation, will alter the absorption spectrum of the solution. There are no general rules as to whether the absorption will increase or decrease. An example of an absorption change resulting from molecular interactions is that of deoxyribonucleic acid (DNA). When the absorption of a solution of native DNA is compared with that of the same concentration of DNA bases, the DNA shows a 30–40% lower extinction coefficient at the 260 nm peak. Heating of the DNA to disrupt the orderly stacking of the bases in the double helix raises the extinction coefficient to that of its constituent bases. The DNA is said to be hypochromic with respect to its bases, while heating of the DNA is said to display a hyperchromic effect. This principle is also widely applied to determine the thermodynamic parameters of nucleic acids using colorimetric measurements. Often some parts of the spectrum will show increased absorbance and other parts will show decreased absorbance. In such a case and when there is an equilibrium between two different molecular forms, there will always exist a wavelength that shows invariant absorption. This point, known as an isobestic point, can be used to quantify concentration of compounds that can exist as different molecular forms. For example, phenol red changes from a protonated to an ionized form between pH 6 and 8, with a switch of the absorption wavelength from 432 (yellow) to 559 nm (red). The absorption at 479 nm, the isosbestic point, is constant regardless the molecular form of this compound and, therefore, this wavelength can be used to measure the concentration of phenol red without a concern to the pH of the solution (Fig. 2).

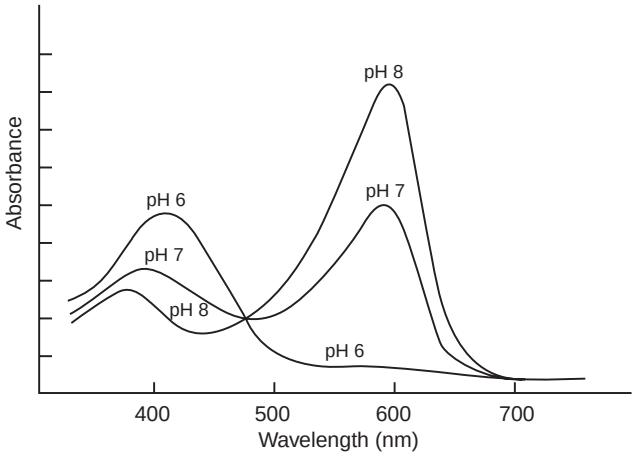


Figure 2. Absorption wavelength of phenol red under various pH conditions. A constant absorption wavelength is shown here at 479 nm, which is the isosbestic point.

Non-uniform distributions of chromophore in solution will lead to decreased absorption. This is most readily understood by considering a situation in which one-half of the cross-section of a light beam goes through a transparent solvent, while the other one-half traverses an absorbing solution. Even if the solution is totally absorbing, at least one-half of the light will be transmitted through the sample leading to a limiting absorbance of $\log(I/I_0) = \log 2 = 0.301$. Thus, a plot of absorbance versus concentration will lead to a curve asymptotically reaching 0.301 as the concentration increases. An example of nonuniform distributions leading to changed absorbance is illustrated by the absorption of Acridine Orange in the presence of cellular organelles. Acridine Orange is a weak base that accumulates in acidic cellular compartments, such as endosomes, lysosomes, and Golgi. When it is concentrated in such organelles, the apparent absorbance of the solution decreases. Since the acid gradient is generated by a membrane-bound enzyme that requires adenosine triphosphate (ATP) as its energy source, addition of ATP to a subcellular suspension containing Acridine Orange will cause a decrease in absorption. In this way, changes in the absorbance of Acridine Orange can be used to monitor the activity of the ATPase (4).

The most common biochemical or clinical use of the absorption properties of chromophores is to measure concentration. In brief, the absorption spectrum, that is, a graph of absorbance as a function of wavelength, is recorded from which a suitable wavelength is chosen. Generally, the wavelength chosen is at or near the peak in absorption of the chromophore. This maximizes the sensitivity of the absorption measurement and decreases possible error resulting from incorrect wavelength calibration of the instrument. Other wavelengths may be chosen if another chromophore present in the solution interferes with the measurement. The extinction coefficient is determined, or a calibration curve of the absorbance is run at a number of different concentrations of chromophore. Concentration is determined by weighing a standard or from the extinction coefficient, if it is accurately known. The plot

is examined for linearity. Subsequent concentrations can be determined either directly from the extinction coefficient or from the calibration curve of absorbance versus concentration. For example, the concentration of DNA can be readily determined by measuring the absorbance of a solution at 260 nm. At this wavelength, an absorbance of 1.0 corresponds to a concentration of 50 and 33 $\mu g \cdot mL^{-1}$ for double- and single-stranded DNA, respectively, at a 1 cm path length.

Different substances can often be distinguished by the use of spectrophotometric measurements if they exhibit different absorption spectra. In an ideal solution, we can assume that the total absorption at any wavelength, is equal to the sum of the individual absorptions. This principle has been used to determine the protein and nucleic acid content of cell extracts and cell fractions or to assess the purity of nucleic acids. At 260 nm, ribonucleic acid (RNA) has a peak absorption of 50.8 $mL \cdot mg^{-1} \cdot cm^{-1}$, while at 280 nm, its absorption is 24.8. Proteins containing an average proportion of aromatic amino acids have a small peak of 2.06 $mL \cdot mg^{-1} \cdot cm^{-1}$ at 280 nm, which decreases to 1.18 $mL \cdot mg^{-1} \cdot cm^{-1}$ at 260 nm (5). By comparing the ratio of absorbance at these two wavelengths, the relative RNA and protein concentration can be determined. The method assumes that the total absorbance at 260 and 280 nm is due to the sum of the absorbances of the constituent protein and nucleic acid. Solving the simultaneous equations yields the concentrations as follows:

Protein concentration (mg/mL)

$$= 0.674 \times A_{280} - 0.33 \times A_{260}$$

Nucleic acid concentration (mg/mL)

$$= -0.016 \times A_{280} + 0.027 \times A_{260}$$

This relationship varies for different proteins and nucleic acids. For accurate work, a calibration must be carried out to determine the actual coefficients for the samples studied. For an accurate determination, specific reagents can selectively react with DNA, RNA, and protein, respectively, for its own quantification to avoid the interference from each other (6,7). For example, the bicinchoninic acid (BCA) assay, Coomassie Blue-G 250 dye binding assay (the Bradford), and the Lowry method are commonly used colorimetic methods for protein quantification (7).

Sometimes samples consist of light-absorbing particles in suspension rather than molecules in solution. For example, while many bacteria do not have chromophores that absorb visible light, suspensions consisting of intact bacteria display apparent absorption due to the scattering of light (Rayleigh scattering). Rayleigh equation demonstrates the reciprocal forth-power relationship between light scattering intensity and wavelength. This apparent absorption has been used to quickly measure bacterial or viral concentrations. A calibration curve is determined by comparing absorbance with viable cell count. For example, the concentration of *Salmonellatyphimurium* in suspension can be determined from the absorption at 440 nm, since the absorption is linear with concentration such that 1.07×10^8 bacteria mL^{-1} yields an apparent absorbance of 1.0, for a 1 cm path length. Methods that utilize the apparent absorption of scattering solutions are often referred to as turbidimetry. Turbidimetric methods can also be used to monitor the kinetics of enzymatic processes that cause changes in the level of light scattering. For example, the enzyme rennin will coagulate milk, leading to an increase in its scattering (monitored at 600 nm) (8). The initial rate of change of apparent absorbance can therefore be measured and used to determine the concentration of active enzyme added. In some cases, the material being measured consists of light-absorbing particles in suspension with chromophores. For example, the absorbance of bacteriophages is primarily due to DNA. Since scattering always begins at wavelengths removed from the absorption, its contribution can be subtracted by measuring the absorbance at longer wavelengths than that of the chromophore and by linearly extrapolating the scattering from the absorption peak. For kinetic measurements on scattering solutions, difference spectra are often used, where the wavelengths are chosen such that one is at the chromophore maximum, while the other is at wavelengths where the only absorption is due to scattering. Specialized instruments that allow for absorbance measurement at two different wavelengths simultaneously are required for such measurements.

Most of the approaches to detect specific DNA sequence are expensive and time-consuming, such as fluorescent microarrays. An inexpensive and rapid assay for the identification of DNA sequence and single nucleotide polymorphisms has been recently described by using a colorimetric method with nanotechnology (9). In this assay, the color of negatively charged gold nanoparticles (Au-nps) is very sensitive to the degree of aggregation. Single-stranded DNA can stabilize Au-nps and prevent the salt-induced aggregation. This method has been applied into clinical samples and successfully identified the relationship between a specific gene and a fatal arrythmia (9).

Due to the contribution of nanotechnology, Au-nps can be utilized as ideal color reporting groups as the colorimetric biosensors. For example, it can provide the qualification and quantification of Pb^{2+}(10), and the detection of polynucleotides (11).

INSTRUMENTATION

Presently, most colorimetry is done using spectrophotometers rather than colorimeters. In principle, spectrophotometers and colorimeters are similar, with the latter being very simple and inexpensive versions of spectrophotometers. Both instruments consist of a light source, a means of selecting wavelength, a sample compartment, and a detector of transmitted light, as shown in a single-beam spectrophotometer (Fig. 3). The source generally consists of an incandescent tungsten lamp for measurements in the visible range between 350 and 700 nm and a hydrogen deuterium, or xenon arc lamp for measurements in the UV range down to 190 nm. Thermal lens spectrophotometry, a new high sensitivity method utilizes a laser as its source (12). A stabilized current is generally provided, especially for measurements in the UV range to

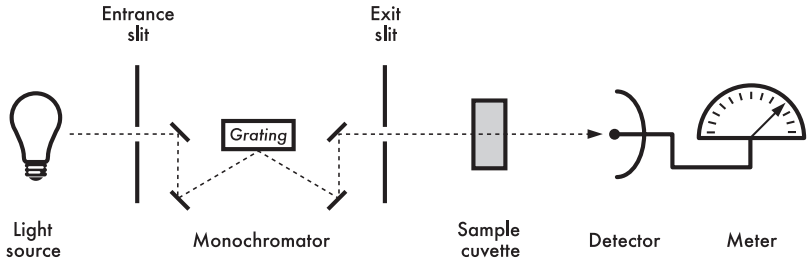


Figure 3. A single-beam spectrophotometer.

prevent fluctuations in the intensity of the lamp. Light from the lamp is collimated or focused by a lens onto a wavelength selector. In the case of the simplest colorimeters, the wavelength selector is a broadband filter that yields a distribution of wavelengths with a width of ~ 40 nm. In spectrophotometers and more sophisticated colorimeters, the wavelength is chosen with a monochromoter by either reflecting the beam off a grating or by passing it through a prism. In a spectrophotometer, the bandwidth of light can be set by adjusting the exit slit. As the slit opening decreases, the light incident on the sample decreases, but the spectral resolution improves. Excessively wide bandwidth can lead to decreased peak absorption and interference from other substances absorbing at nearby wavelengths. In a colorimeter, the bandwidth is fixed, depending on the properties of the filters. While it is possible to obtain narrow bandpass filters if necessary, they may cost as much as the colorimeter.

The monochromatic light is then focused on the sample, contained in a transparent rectangular cuvette, a test tube, or a microplate with 6–384 wells. In some spectrophotometers, the sample compartment has been designed to accommodate with 96–1536-well microplate for high throughput screening (13). Colorimeters generally use test tubes made of any clear materials, such as glass or plastics, as long as they allow any required chemical reactions to be carried out in the same disposable container in which the measurement is carried out. Spectrophotometers and automated colorimeters designed to determine large numbers of samples usually hold rectangular cuvettes and microplates. These are generally made of glass or plastics for measurements within visible light wavelengths, and of quartz for measurements < 350 nm in the UV range. The transmitted light is incident on a phototube that records the intensity of the light reaching it. Some spectrophotometers are double-beam instruments, which electronically subtract an absorption blank. There are two types of the double-beam spectrophotometers. One is called the double -beam in-time spectrophotometer (Fig. 4). In this instrument, a single beam of light from the monochromator is alternately switched between the sample and the reference cuvettes. The two alternate beams then reach a single detector in separate time to provide an alternating signal with an amplitude that is proportional to the difference of the light intensities between the sample and the reference. The other type is called the double-beam in-space spectrophotometer (Fig. 5). In this instrument, two separate light pathways are created by a beam splitter and mirrors. One beam passes through the sample cuvette, and the other beam the reference cuvette. The light intensities of the sample and the reference cuvettes are measured by two separate detectors and the difference is recorded as the absolute photometric measurement.

Many spectrophotometers are still designed as single-beam instruments (Fig. 3); the solvent absorption is separately determined, stored, and subtracted by an interfaced microcomputer. In colorimeters and other single-beam instruments, the absorption is set to zero or the transmission is set to 100% when the reference blank is placed in the sample compartment. The detector is designed to give a direct reading of either transmission or absorbance. The newer spectrophotometer is a multifunctional instrument with a triple-mode cuvette port and microplate reading capability. The detection modalities include absorbance, fluorescence intensity, fluorescence polarization, time-resolved fluorescence, and luminiscence. With a dual-monochromator, the filters for specific wavelengths are not required in these spectrophotometers. (e.g., SpectraMax M5 by Molecular Devices Corp.). The function of spectrophotometer also has improvement in the rate of data acquisition. It can be as fast as 50 μs, which is beneficial in rapid kinetics.

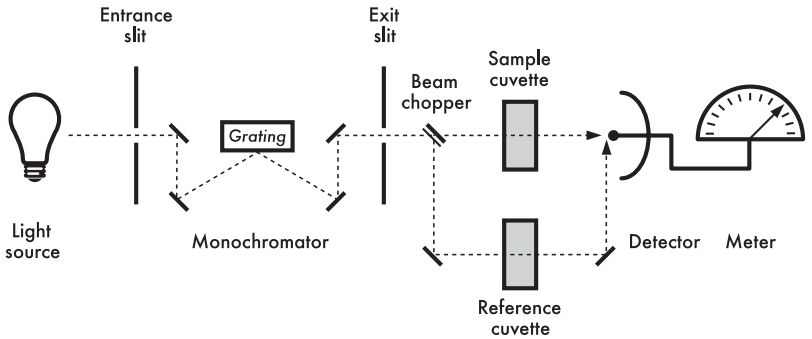


Figure 4. A double beam in-time spectrophotometer.

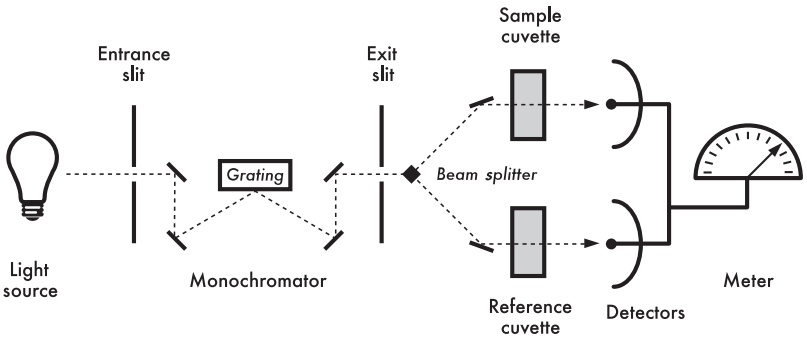


Figure 5. A double-beam in-space spectrophotometer.

CLINICAL APPLICATIONS OF COLORIMETRIC METHODS

Colorimetry has seen wide applications in clinical and medicinal chemistry and in biochemistry. Colorimetric methods are used to measure concentration or enzyme activity. Lately, colorimetric methods are also applied in genomics and proteomics research, such as the single nucleotide polymorphism analysis (14), the detection of protein microarrays (15), and so on. The concentration of a substance can be determined directly and immediately if the unknown has a distinct absorption band that does not overlap other substances in the assay mixture. If not, the unknown may be analyzed indirectly on completion of one or more chemical reactions, yielding a colored compound with a definite stoichiometric relationship to the unknown compound. The reaction must be rapid and specific and must react completely with the unknown to be of use. In contrast to quantitative measurements, kinetic measurements determine initial reaction kinetics by the rate of appearance of an absorbing product or by the rate of disappearance of an absorbing reactant. Kinetic measurements are carried out either directly by monitoring changes in the absorption of substrates or cofactors or indirectly by coupling a second, third, or additional enzyme reactions to the primary one. Each of these cases will be discussed and illustrated later.

Due to the development of new automated diagnostic instruments, the manual colorimetric assays have become less useful in most clinical laboratories. However, most tests performed in automated instruments are based on simple colorimetry. Therefore, the principles in colorimetry are essential for the improvement and development of new automated assays. Furthermore, manual colorimetric methods are still used in handling small numbers of samples, such as in small clinical laboratories or for special diagnostic purposes.

There are several general precautions that should be taken in colorimetric measurements of clinical samples. These are as follows:

1. Sample preparation. Most clinical samples (e.g., blood and urine) contain a large amount of components other than the analyte. The results of the test are only meaningful if the background reading of the sample has been subtracted. The background must be obtained by using proper controls. Usually, when a measurement is made by the addition of a reagent, an identical sample without the reagent can be used as a control or blank. When an enzyme activity is measured, a control or blank can be obtained by excluding the specific substrate or by including a specific inhibitor for the enzymatic reaction. Because of instability, many specimens must be handled immediately after they reach the laboratory and cannot be stored. The turbidity of a sample will cause false readings in colorimetry. This problem generally can be eliminated either by centrifugation, filtration, or comparing with identical controls.
2. Buffer solutions. The absorbances, especially in the visible wavelength range, are usually pH dependent. Furthermore, the type and concentration of ions in the buffer may also influence the absorbance of a molecule or the rate of reaction. Therefore, careful attention must be paid to the solvent and buffer system. In addition, the rates of enzymatic reactions are highly pH dependent. It is therefore important to choose a buffer with an ionization potential (pK) appropriate to the optimal pH for the assay. For example, imidazole or triethanolamine buffer is commonly used in the pH range of 6.5–7.0.
3. Temperature. Kinetic measurements use enzymes to catalyze the reactions of intensely colored substances. Enzymatic reactions are highly temperature dependent. It is therefore necessary to use isothermal control when carrying out such assays. If not, it is necessary to know precisely the reaction temperature. If the reaction is carried out in the colorimeter, care must be taken to ensure that the temperature of the reaction mixture is not affected by the heat produced in the sample chamber.
4. Inhibitors. Enzymatic reactions are subject to competitive inhibition or to the action of specific inhibitors leading to apparent decreases in the measured activity or concentration of the unknown. It is therefore extremely important to eliminate these interfering substances or to include the proper controls during the assay. In cases of product inhibition, the initial rate should be used for the determination of the enzymatic activity.

In the following sections, several colorimetric measurements of clinical importance will be discussed. They are used primarily as examples to indicate the standard procedures of colorimetry in clinical laboratories. The number

of applications of colorimetry is too vast to present an extensive survey in this article.

Quantitative Measurement

Direct Measurement. Direct colorimetric measurement for quantitative assays are limited to very few cases in which the compound of interest itself has an intense absorption and in which there is no interference from other components in the samples. One example is the determination of hemoglobin in blood. Since oxygenated and deoxygenated hemoglobins, carboxyhemoglobin, and hemoglobin derivatives have different absorption spectra, the concentration of a specific type of hemoglobin, as well as the oxygen-binding capacity, can be determined by measuring the ratio of absorption (A) at two different wavelengths (16) (e.g., A_{562}: A_{540} for carboxyhemoglobin and A_{560}: A_{506} or A_{650}: A_{825} for oxygen-binding capacity), similar to that described above for the protein–nucleic acid measurement.

Bilirubin measurement in blood samples can also be done by direct measurement at 460 nm (17). The absorbance is compared with that of a standard solution of potassium dichromate and is reported in term of units of icterus index. One unit of icterus is equivalent to a 1:10,000 solution of potassium dichromate. This determination is only approximate, because there are many other components in blood, mostly carotenoid pigments, which will interfere with the readings.

Indirect Measurements

Complex Formation. One of the most common applications of colorimetry is in the determination of protein concentration. Colorimetric determination of the total serum protein by the Biuret reaction is still used in many clinical laboratories. In this method, proteins react with copper sulfate in alkaline solution to form a Biuret complex that can be determined by the intensity of the violet color with an absorption at 555 nm (18). The Biuret reaction is a relatively straightforward, precise, and accurate method. Usually, 0.1 mL serum is added to 5 mL of Biuret Reagent solution, which contains sodium, potassium tartrate, potassium iodide, and copper sulfate in 0.2 M NaOH. After incubation at 30–32 °C for 10 min or at room temperature for 20 min, the reaction mixture is read at 555 nm. The sample absorbance is generally compared to human or bovine serum albumin (BSA) standards. The color of the Biuret complex is stable for several hours after it reaches the maximum intensity. The measurement of protein by Biuret measurement is subject to interference by other substances in serum. In serum separated from moderately hemolyzed blood or serum with high bilirubin content potassium cyanide is usually included in the reaction mixture to correct the difference. Dextrans may also interfere by causing turbidity. For these determinations, appropriate blanks should be used instead of just saline or water, as described in most general procedures.

Many other methods use nearly colorless chemical reagents that complex with the protein to form highly colored compounds. One of them, the Lowry method (19), is the single most utilized assay in biochemistry research and is the most cited reference in the biochemical literature. However, it is not widely used in clinical applications. A variety of dyes, which form strong binding complexes with protein, is also used to quantitate protein concentration. For example, Coomassie Brilliant Blue R-250 is commonly used to visualize and quantify proteins in gels separated by electrophoresis and to quantify proteins eluted from gels (20). To visualize proteins, gels are stained with dye in 50% methanol–10% acetic acid solvent, after which excess dye is removed by the solvent. To accurately quantitate protein concentration, dye bound to protein is removed by electroelution or chemical elution, and absorbance of the solutions at 595 nm is measured. Coomassie Brilliant Blue stained electrophoresis gels are quantitated by direct scanning colorimetry at the same wavelength. This dye is most useful to quantify proteins present in the amount range of 0.5–50 μg. The binding of silver to protein has been used to visualize and to quantitate proteins separated in gels (21). While this method is ~100 times more sensitive than Coomassie Brilliant Blue staining, it is also much more expensive due to the cost of the silver reagent. Recently, several visible dyes have been developed for the staining of protein in electrophoresis (22). Many of them can reach the sensitivity at the ng level, and some of them, such as 3, 3′-diethyl-9-methyl-4,5,4′,5′-dibenzothiacarbocyanine (Stain-All), can simultaneously stain DNA and RNA with the appearance of different colors.

A complex between the dye, Methylene Blue, and protein has been used to quantitate the growth of cells in culture and their inhibition by cytotoxic drugs. Cultures are stained, destained, then solubilized overnight in a 1% aqueous Sarkosyl solution, and quantified at 620 nm. This assay can be automated by using 96-well microtiter plates and commercially available spectrophotometers designed for enzyme-linked immunosorbent assays (ELISA) (23). The dye complex can also show specificity and can be used to assay a single protein in a specimen. For example, serum albumin can be determined by its high affinity binding to Bromocresol Green (24).

Another important colorimetric method by complex formation in clinical laboratories is the measurement of metal ions by chelators. For example, serum iron levels are measured by the complex formation between ferrous and ferrozine. Serum iron can be determined either with or without the removal of proteins by trichloroacetic acid precipitation. Ferric is usually reduced to ferrous by ascorbic acid. The complex of Fe^{2+}-ferrozine has a dark violet color with an absorbance at 562 nm. Advantages of this method are its simplicity, sensitivity (molar absorbance = 27,900), and the constant absorbance in a wide range of pH (from pH 4–10) (25).

Colorimetry is commonly used for water testing, including specific impurities, such as ammonia, calcium, chlorine, copper, iron, nitrite, phenol, phosphate, sulfate, and sulfide, or total hardness. These tests are available in kit form and can be used with portable colorimeters for field testing. The methods generally utilize either complex formation or derivatization of the inorganic impurity to yield a highly colored compound.

Derivatization. Besides complex formation, substances can be detected colorimetrically after chemical modification to give colored products. One example is the diazotization of bilirubin. As mentioned previously, the direct measurement of bilirubin by its yellow color is inaccurate, because there are many pigments in serum that can interfere with the detection. For a more reliable assay, bilirubin is reacted with diazotized sulfanilic acid to give azobilirubin, which is red-violet in moderately acid solution and blue in strongly acid or alkaline solution. Most of the methods currently used in clinical laboratories (e.g., the Jendrassik–Grof method and the Malloy–Evelyn method) are based on this diazotization reaction.

Bilirubin exists in serum in two forms, the glucuronide-conjugated and the free form. The conjugated form (clinically known as "direct bilirubin") is more soluble in water and can be detected by a direct reaction with the diazo reagent. The free form (clinically known as "indirect bilirubin") is less soluble in water and can be detected by diazo reagent only if other reagents are also included. In the Malloy–Evelyn method, ethanol is added to increase the solubility of free bilirubin. In the Jendrassik–Grof method, caffeine–benzoate is added to displace serum protein-bound bilirubin. Therefore, both free and conjugated bilirubin can be detected by the diazo reagent. The Jendrassik–Grof method is generally considered the method of choice for the measurement of bilirubin in most clinical laboratories (26). The diazo reagent in this method is a mixture of sulfanilic acid and sodium nitrite in hydrogen chloride. This reagent should be prepared within 30 min of use. Serum is reacted with the diazo reagent for exactly 10 min, and the reaction is stopped by the addition of ascorbic acid solution. A strongly alkaline solution (tartrate in 1 *N* NaOH) is then added, and the color developed is compared with a standard curve of absorbance at 600 nm. Bilirubin is extremely sensitive to light. Sunlight can markedly decrease the bilirubin content in samples (as much as 50%/h). However, serum specimens can be stored for many weeks or months without appreciable change of bilirubin content if they are kept in the dark in a freezer. Other examples of derivatization include the determination of cholesterol by Liebermann–Burchard reaction. In this reaction, cholesterol is reacted with a mixture of acetic anhydride, acetic acid, and sulfuric acid to give a bluish green product that can be measured colorimetrically.

Enzymatic Conversion. Enzymatic conversion can be used to determine concentrations by the measurement of either the loss of substrate, the creation of product, or the change of the cofactor (coenzyme). Nicotinamide adenine dinucleotide (NAD) and nicotinamide adenine dinucleotide phosphate (NADP) and their respective reduced forms, NADH and NADPH are cofactors that have been extensively utilized for analytical purposes in colorimetric assays. These cofactors serve as the natural oxidizing and reducing agents in a wide variety of enzyme systems. With the appropriate enzyme, they can selectively oxidize or reduce a single substrate in the presence of innumerable other compounds, making possible the analysis of a single compound in a complex mixture. Both NADH and NADPH have identical absorption bands with peak absorption at 340 nm. The compounds NAD and NADP do not absorb at this wavelength. Therefore, changes in oxidation or reduction can be measured colorimetrically. In addition, the reduced forms can be completely destroyed at acidic pH without affecting the oxidized forms and the oxidized forms can be completely destroyed at basic pH without affect the reduced forms (27).

The measurement of glucose-6-phosphate in the range of 20–200 μM can be done in one step by monitoring the reduction of NADP to NADPH. This reaction is carried out by the oxidation of glucose-6-phosphate to 6-phosphate gluconolactone by the enzyme glucose-6-phosphate dehydrogenase, for which NADP is the coenzyme and is reduced to NADPH. The reaction is quantitated by measuring the 340 nm absorption of NADPH after completion of the reaction. Biological substances that do not react directly with nicotinamide nucleotides can be analyzed in this system by one or more additional enzymes. For example, glucose can be determined by the following two-step reaction that convert NADP to NADPH.

$$\text{Glucose} + \text{ATP} \xrightarrow{\text{hexokinase}} \text{Glucose-6-phosphate} + \text{ADP}$$

$$\text{Glucose-6-phosphate} + \text{NADP} \xrightarrow[\text{dehydrogenase}]{\text{glucose-6-phosphate}} \text{6-Phosphate gluconolactone} + \text{NADPH} + \text{H}^+$$

An example of a three-step assay is the enzymatic analysis of inorganic phosphate (P_i), carried out with three simultaneous reactions, ends in the conversion of NADP to NADPH (28).

$$\text{Glycogen} + P_i \xrightarrow{\text{phosphorylase}\,\alpha} \text{Glucose-1-phosphate}$$

$$\text{Glucose-1-phosphate} \xrightarrow{\text{phosphate glucomutase}} \text{Glucose-6-phosphate}$$

$$\text{Glucose-6-phosphate} + \text{NADP} \xrightarrow{\text{glucose-6-phosphate dehydrogenase}} \text{6-Phosphate gluconolactone} + \text{NADPH}$$

Hydrogen peroxide can be readily quantitated colorimetrically by its peroxidase-catalyzed reaction with a colorless chromogenic oxygen acceptor to form an intensely colored product. This reaction can therefore be used to measure the concentration of any organic compound that, on reaction, will produce hydrogen peroxide. For example, glucose levels in serum can be determined by indirect detection using the enzyme glucose oxidase. This enzyme catalyzes the oxidation of glucose to gluconic acid and hydrogen peroxide. The amount of hydrogen peroxide produced can be detected by reaction with peroxidase, and most commonly *o*-dianisidine. The oxidized product of *o*-dianisidine has a strong absorbance ~ 540 nm. The overall reactions are as follows:

$$\text{Glucose} + O_2 + H_2O \xrightarrow[\text{oxidase}]{\text{glucose}} \text{Gluconic acid} + H_2O_2$$

$$H_2O_2 + o\text{-Dianisidine} \xrightarrow{\text{peroxidase}} H_2O + \text{Oxidized } o\text{-dianisidine}$$

Since this measurement is dependent on the color formation of the oxidized chromogen, other substances in the sample can interfere with this reaction by competing with chromogens for hydrogen peroxide and by reducing the final color intensity. Some of the interfering substances in

serum are creatine, uric acid, ascorbic acid, bilirubin, and glutathione. Therefore, results obtained directly from serum tend to be lower than the true values. Serum samples, especially with red cells, or extensive hemolysis require precipitation of the protein to remove interfering enzymes, with measurements carried out on the protein-free filtrates. Direct measurements of glucose from urine specimens cannot be done by the glucose oxidase method because of the presence of enzyme inhibitors.

Glucose oxidase is highly specific to β-d-glucose. In aqueous solutions, glucose exists 36% in the α form and 64% in the β form. The same ratio of the two forms is also found in serum. Crystalline glucose, however, can be either the α or β form, depending on the conditions of crystallization. In order to correct the difference in standard solutions, some commercial preparations of glucose oxidase contain another enzyme, mutarotase, which accelerates the conversion of α form to β form during the assay. Alternatively, standard solutions from crystalline glucose can be prepared 2 h before the determination to allow the mutarotation to reach equilibrium. The final assay solution contains the following components: glucose oxidase ($5\,U{\cdot}mL^{-1}$), peroxidase ($16\,U{\cdot}mL^{-1}$), and *o*-dianisidine ($0.6\,\mu mol{\cdot}mL^{-1}$) in a pH 7.0 phosphate buffer. Under these assay conditions, glucose can be measured at a range of up to 250 $mg{\cdot}dL^{-1}$. Samples with higher glucose concentrations should be diluted before the determination. Note that most chromogens used in the peroxidase reaction (e.g., *o*-dianisidine and *o*-tolidine) are potential carcinogens, and precautions should be taken when handling these compounds.

Colorimetric assays have found substantial applications in determination of concentration by ELISA. For example, the previously described assay to measure glucose concentration has been adapted as follows (29). The production of hydrogen peroxide from glucose is catalyzed by glucose oxidase, an enzyme consisting of apoglucose oxidase and flavin adenine dinucleotide (FAD) cofactor. If a ligand is bound to FAD and antibodies to that ligand are added to the solution, the enzyme will be inactive. Competition by free ligand will make conjugated FAD available for apoglucose oxidase. Since this reaction is enzymatic, excess glucose can be added, which leads to amplification of the signal. Therefore, very low concentrations of ligand can be assayed colorimetrically by the oxidation of glucose.

Besides glucose, the production of hydrogen peroxide as an intermediate has been employed to measure a number of clinically important substances including lecithin, high density lipoprotein cholesterol, phospholipids, digoxin, and triglycerides in human serum (30). In the last example, triglycerides are hydrolyzed to glycerol and fatty acid with lipase. The resulting glycerol is phosphorylated by l-α-glycerophosphate oxidase to produce hydrogen peroxide, which reacts in the peroxidase-catalyzed coupling of 4-aminoantipyrine and sodium 2-hydroxy-3,5-dichlorobenzene sulfonate to form an intense red product.

Kinetic Measurement

Kinetic measurements are generally used to determine the activity of enzymes. The expression of enzyme activity units (U) has been extremely arbitrary and inconsistent over the years. Historically, activity has been measured colorimetrically, and often activity for many common enzymes is expressed as an initial rate of absorbance change of a particular chromogen at the appropriate wavelength per unit time. More recently, enzyme activities have been expressed on a molar basis (i.e., the molar rate of destruction, or creation of substrates or products per unit time). Kinetic measurement of enzyme activity requires more control than quantitative measurements, since activity is usually a function of pH, ionic strength, temperature, substrate concentration, and the presence of activators and inhibitors.

Direct Measurement. An enzymatic reaction can be monitored by the measurement of either the rate of loss of the substrate, the formation of the product, or the rate of change of the cofactor (coenzyme) concentration. All of these methods are commonly used to determine enzyme activity.

Substrate. In an enzymatic reaction, which converts a substrate to a colored product or a colored substrate to a colorless product, enzymatic activity can be detected colorimetrically. An example of this type of measurement is prostatic acid phosphatase determination (31). Acid phosphatase is present in many tissues, including bone, liver, kidney, erythrocytes, and platelets. Its level in serum is measured as a diagnostic for the detection of metastatic, prostatic carcinoma. Therefore, substrates specific to prostatic acid phosphatase isoenzyme are most desirable for this assay. One of the most commonly used substrates is thymolphthalein monophosphate, which, upon hydrolysis by the enzyme, produces thymolphthalein, a compound with intensive absorbance at 590 nm in alkaline solution.

$$\text{Thymolphthalein monophosphate} \xrightarrow[\text{phosphatase}]{\text{acid}} \text{Thymolphthalein} + HPO_4^{2-}$$

The course of this reaction can be monitored by a direct colorimetric measurement of product formation. Note that thymolphthalein monophosphate is not completely specific for prostatic acid phosphatase. However, unlike acid phosphatase from other sources, prostatic acid phosphatase can be inhibited by tartrate. Thus, from the assays of the enzyme activity in the presence or absence of tartrate, one can determine the phosphatase activity specific to the prostatic secretion. Acid phosphatase is extremely labile at the neutral pH of normal serum. Therefore, specimens for the enzyme measurement should be handled on ice and delivered to the laboratory as rapidly as possible. Serum separated in the laboratory should be acidified by acetate buffer to a pH range (pH 5–6) to stabilize the enzyme. Samples should be stored in the freezer if they are not to be assayed on the same day.

Acid phosphatase activity is highly dependent on pH and the specific ion in the buffer. Small changes in these parameters may cause a large difference in the enzyme activity measurements. To avoid a discrepancy between assays, a standardized procedure has to be followed carefully. Additional factors, such as the ratio of serum sample to the volume of final mixture, surfactant (Brij-35) for activation

of enzyme, and the source and concentration of the substrate, must be consistent. For example, thymolphthalein monophosphate obtained from different commercial sources has been found to give as much as a 40% deviation in the enzyme activity by the colorimetric measurement. A typical assay procedure uses a substrate solution consisting of 0.6 mL of 0.15 *M* acetate buffer (pH 5.4), 1 m*M* thymolphthalein monophosphate and 1.5 $g \cdot L^{-1}$ Brij-35, a sample volume of 50 μL serum, a 30 min incubation at 37 °C, and final color development in 1 mL of 0.1 *M* NaOH and 0.1 *M* Na_2CO_3. The absorbance at 590 nm is then measured. In controls, the substrate solution is incubated without serum sample. Acid phosphatase activity is determined by subtracting the reading of the control from that of the sample.

The activity of a large number of proteolytic enzymes can be measured directly using specific synthetic substrates designed for that purpose. For example, the activity of leucine aminopeptidase can be directly measured by the hydrolysis of l-leucyl-β-naphthylamide to β-naphthylamine, which can be read at 560 nm. The clinically important γ-glutamyl transpeptidase can be measured by its reaction with glutamyl nitroanilide to form nitroaniline, which can be read at 405 nm.

Cofactors. The cofactors NAD and NADP find wide use in the kinetic measurement of enzyme activity, as well as in the previously discussed quantitative measurement of substrate concentration. One of the most important enzymes to be measured in clinical laboratories is lactate dehydrogenase. Marked increase of the enzyme level has been found following myocardial infarction. Elevations have also been reported in leukemia, in anemias, and in some liver diseases. This enzyme catalyzes the interconversion of lactate and pyruvate in the presence of NAD or NADH.

$$\text{Lactate} + \text{NAD} \underset{\text{dehydrogenase}}{\overset{\text{lactate}}{\rightleftharpoons}} \text{Pyruvate} + \text{NADH}$$

In clinical laboratories, the reaction can be detected in either direction (i.e., using either lactate or pyruvate as substrate). A simple colorimetric method to detect the change of absorbance at 340 nm is sufficient to measure the enzyme activity (32). Generally, a small aliquot of serum sample is added to a cuvette and is mixed with a substrate solution containing either lactate and NAD or pyruvate and NADH. The reaction mixture in the absence of enzyme is set up, and the absorbance at 340 nm is measured. A dilution of enzyme is added and the OD_{340} is either monitored continuously or measured every minute or two for an appropriate time interval. The initial rate of increase in absorbance will be linear in time and proportional to the quantity of enzyme. The enzyme activity will therefore be proportional to the rate of change of the absorbance. A standard assay procedure is performed at 30°C. When the absorption change is measured in a spectrophotometer without constant temperature equipment, the effect of temperature on the enzyme activity must be considered. For example, the enzyme activity measured at 30 °C is 1.44-fold higher than that at 25 °C. Any enzymes that require the NADs as cofactors can be assayed in this manner. In clinical laboratories, lactate dehydrogenase is usually determined with samples of serum. Normally, serum contains no inhibitors or interfering substances for this enzyme assay. However, blood samples with marked hemolysis should be avoided, because they may give false elevations of this enzyme in the serum. It has been reported that commercial NADH contains inhibitors for the dehydrogenase reactions. Therefore, when pyruvate is used as the substrate for lactate dehydrogenase determination, inhibitors in the NADH solution should be considered. Many other enzymes can be determined by direct colorimetric measurement of NAD–NADH or NADP–NADPH conversion. Some of these enzymes with clinical importance are isocitrate dehydrogenase, glutamate dehydrogenase, and glucose-6-phosphate dehydrogenase.

Indirect Measurements. The activity of an enzyme that is not NAD or NADP dependent can nonetheless be measured with this system if its products are substrates for other enzymatic reactions requiring pyridine nucleotides. For example, phosphate glucomutase can be determined as follows

$$\text{Glucose-1-phosphate} \xrightarrow{\text{phosphate glucomutase}} \text{Glucose-6-phosphate}$$

$$\text{Glucose-6-phosphate} + \text{NADP} \xrightarrow[\text{dehydrogenase}]{\text{glucose-6-phosphate}} \text{6-Phosphate gluconolactone} + \text{NADPH}$$

To prevent interfering side reactions, this assay is often carried out in two steps. The phosphate glucomutase is allowed to react for a measured time period, without added NADP or glucose-6-phosphate dehydrogenase after which the reaction is stopped with heat. The quantity of glucose-6-phosphate produced during that time period is then determined by carrying out the second reaction. Even though NADPH can be directly monitored at 340 nm, it can also be measured indirectly by the interaction with a tetrazolium salt, 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT), to reduce the latter to a more intensely colored form that absorbs in the visible rather than the UV range. The reducing reaction occurs only in the presence of an intermediate electron carrier, such as phenazine methosulfate. The final color of the reduced tetrazolium salt has an absorption ~520 nm. This same reagent is now used in place of [^{3}H]thymidine to measure cell proliferation or complement mediated cytotoxicity in lymphocytes. 3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide is cleaved to a dark blue formazan product only if the cell has active mitochondria while even newly dead cells show no color change.

The enzyme creatine kinase is measured in clinical laboratories for the detection of diseases related to skeletal or heart muscles. In the presence of magnesium ions, this enzyme catalyzes the reaction of creatine phosphate with adenosine diphosphate (ADP):

$$\text{Creatine phosphate} + \text{ADP} \underset{}{\overset{\text{creatine kinase}}{\rightleftharpoons}} \text{Creatine} + \text{ATP}$$

This reaction can be detected by coupling with two other enzymes (33). The first enzyme, hexokinase, can use ATP, one of the products of the previous reaction, to convert d-glucose to d-glucose-6-phosphate, which is subsequently

detected with the second enzyme, d-glucose-6-phosphate dehydrogenase, by measuring the conversion of NADP to NADPH as the increase of absorbance at 340 nm. The reactions catalyzed by the auxiliary enzymes are as follows:

$$\text{ATP} + \text{D-glucose} \xrightarrow{\text{hexokinase}} \text{ADP} + \text{glucose-6-phosphate}$$

$$\text{Glucose-6-phosphate} + \text{NADP} \xrightarrow{\text{glucose-6-phosphate}} \text{D-Glucose-6-phosphate lactone} + \text{NADPH}$$

Since creatine kinase in serum is rapidly inactivated by oxidation, reducing agents are required to reactivate the enzyme in each assay. The most commonly used reactivating compound is *N*-acetylcysteine. Other sulfhydryl compounds can also be used if they do not interfere with the measurement (i.e., absorbance at 340 nm, solubility, and odor). Serum with extensive hemolysis should be avoided, because it may give falsely elevated values. One of the factors from hemolyzed red cells is the enzyme, adenylate kinase, which can generate additional amounts of ATP from ADP without creatine phosphate. To eliminate this artifact, 3 m*M* of adenosine monophosphate (AMP) should be included in the assay solution to inhibit adenylate kinase activity. Higher concentrations of AMP are required in markedly hemolyzed serum samples. However, at high concentrations, AMP can also inhibit creatine kinase activity. The false activity contributed from enzymes other than creatine kinase can be detected by running an appropriate control, without creatine phosphate in the final assay solution. The coupling of three reactions in this manner leads to a complicated assay that requires 10 different reagents added to the sample. These include imidazole buffer at pH 6.7 (100 m*M*), creatine phosphate (30 m*M*), ADP (2 m*M*), Mg^{2+} (10 m*M*), d-glucose (20 m*M*), NADP (2 m*M*), AMP (5 m*M*), *N*-acetylcysteine (20 m*M*), glucose-6-phosphate dehydrogenase ($1.5\,\text{U}\cdot\text{mL}^{-1}$), and hexokinase ($2.5\ \text{U}\cdot\text{mL}^{-1}$).

FUTURE DEVELOPMENTS

Colorimetric methods will continue to be important analytical tools in research and clinical laboratories. Applications that replace the need for radioactive compounds will continue to be developed because of the increasing costs and difficulties associated with the processing and disposal of radioactive materials. In the same way that the ELISA has replaced radioimmunoassay (RIA), colorimetric methods will continue to replace applications requiring radioisotopes. For example, in recent developments of microarray technology, colorimetry is still one of the major detection methods for the measurement of the signals (14,34). However, the sensitivity of colorimetric methods is limited by the absorption of the chromophores, new reagents with very high extinction coefficients will need to be developed for direct colorimetric measurements (35). This limitation is also overcome by the coupling of a second reaction to the primary measurement. Such amplifying systems will be increasingly used as a means to increase the colorimetric sensitivity. Modified colorimetric methods, such as the kinetic spectrophotometric method, can also markedly improve the sensitivity of the determination. For example, trace concentrations of iron can be determined by the oxidation of *o*-tolidine (36). The rate of oxidation, which can be measured colorimetrically at 440 nm for the oxidized *o*-tolidine derivative, is proportional to the concentration of ferric ion. This method can detect as low as $4 \times 10^{-7} M$ iron in acidic solutions. The use of lasers as light sources in colorimetry has opened a new frontier of analytical chemistry. The method, called "thermal lens spectrophotometry" is based on the direct measurement of the absorbed radiant energy by the "thermal lens effect" and can detect a very small absorption by increasing the power of the heating laser (12). This method can determine accurately an absorbance that is lower than 0.001.

Since the sensitivity of fluorimetry is often greater than that of colorimetry, it will probably find more new applications than colorimetry especially in biomedical measurements. However, the instrumentation and the stability of fluorimetric measurements are generally more expensive and more complex than colorimetry. In addition, new colorimetric instruments have been developed that are capable of measuring several wavelengths simultaneously, of processing large numbers of samples automatically, and of reading the results of samples contained in multiwell or multicuvette containers. Furthermore, colorimeters with computer interfaces will automatically calculate results, will average multiple determinations, and will print and store the final data. This kind of instrumentation, already widely used in ELISA and microarrays, will be increasingly borrowed for use in other colorimetric assays. Therefore, colorimetry will continue to be used because of its speed, simplicity, versatility, and low cost. It can be anticipated that more sophisticated methods of colorimetric measurement, in combination with other advanced technologies such as nanotechnology (11), will continue to be developed into new analytical methods with various biomedical and clinical applications.

BIBLIOGRAPHY

1. Bugay DE. Characterization of the solid-state: spectroscopic techniques. Adv Drug Deliver Rev 2001;48:43–65.
2. Stephenson GA, Forbes RA, Reutzel-Edens SM. Characterization of the solid state: quantitative issues. Adv Drug Deliver Rev 2001;48:67–90.
3. Yu LX, Lionberger RA, Raw AS, D'Costa R, Wu H, Hussain AS. Applications of process analytical technology to crystallization processes. Adv Drug Deliver Rev 2004;56:349–369.
4. Stone DK, Xie X-S, Racker E. An ATP driven proton pump in clathrin-coated vesicles. J Biol Chem 1983;258:4059.
5. Warburg O, Christian W. Isolation and crystallization of enolase. Biochem Z 1941;310:384.
6. Morozkin ES, Laktionov PP, Rykova EY, Vlassov VV. Fluorometric quantification of RNA and DNA in solutions containing both nucleic acids. Anal Biochem 2003;322:48–50.
7. Sapan CV, Lundblad RL, Price NC. Colorimetric protein assay techniques. Biotechnol Appl Biochem 1999;29:99–108.
8. McMahon DJ, Brown RJ. Milk coagulation time—linear relationship with inverse of rennet activity. J Dairy Sci 1983;66: 341.
9. Li H, Rothberg LJ. Label-free colorimetric detection of specific sequences in genomic DNA amplified by the polymerase chain reaction. J Am Chem Soc 2004;126:10958–10961.

10. Liu J, Lu Y. Accelerated color change of gold nanoparticles assembled by DNAzymes for simple and fast colorimetric Pb^{2+} detection. J Am Chem Soc 2004;126:12298–12305.
11. Elghanian R, Storhoff JJ, Mucic RC, Letsinger RL, Mirkin CA. Selective colorimetric detection of polynucleotides based on the distance-dependent optical properties of gold nanoparticles. Science 1997;277:1078–1081.
12. Long ME, Swofford RL, Abrecht AC. Thermal lens technique: A new method of absorption spectroscopy. Science 1976;191: 183.
13. Sittampalam GS, Kahl SD, Janzen WP. High-throughput screening: advances in assay technologies. Cur Opin Chem Biol 1997;1:384.
14. Ihara T, Tanaka S, Chikaura Y, Jyo A. Preparation of DNA-modified nanoparticles and preliminary study for colorimetric SNP analysis using their selective aggregations. Nucleic Acids Res 2004;32:e105.
15. Liang RQ, Tan CY, Ruan KC. Colorimetric detection of protein microarrays based on nanogold probe coupled with silver enhancement. J Immunol Methods 2004;285:157–163.
16. Van Kampen EJ, Zijlstra WG. Spectrophotometry of hemoglobin and hemoglobin derivatives. Adv Clin Chem 1983;23:199.
17. Henry RJ, Golub OJ, Berkman S, Segalove M. Critique on the Icterus index determination. Am J Clin Pathol 1953;23:841.
18. Kingsley GR. Procedure for serum protein determinations. Stand Methods Clin Chem 1972;7:199.
19. Lowry OH, Rosebrough NS, Farr AL, Randall RI. Protein measurement with the Folin phenol reagent. J Biol Chem 1951;193:265.
20. Rylatt DB, Parish CR. Protein determination on an automatic spectrophotometer. Anal Biochem 1982;121:213.
21. Peats S. Quantitation of protein and DNA in silver stained gels. Anal Biochem 1984;140:178.
22. Jin L-T, Choi J-K. Usefulness of visible dyes for the staining of protein or DNA in electrophoresis. Electrophoresis 2004;25: 2429.
23. Finlay GJ, Baguley BC, Wilson WR. A semiautomated microculture method for investigating growth inhibiting effects of cytotoxic compounds on exponentially growing carcinoma cells. Anal Biochem 1984;139:272.
24. Webster D. The immediate reaction between bromcresol green and serum as a measure of albumin content. Clin Chem (Winston-Salem, NC) 1977;23:663.
25. Stookey LL. Ferrozine—a new spectrophotometric reagent for iron. Anal Chem 1970;42:779.
26. Jendrassik L, Grof P. Simplified photometric methods for the determination of the blood bilirubin. Biochem Z 1938; 297:81.
27. Lowry O, Passonneau J. A Flexible System of Enzymatic Analysis. New York: Academic Press; 1972. Chapt. 5.
28. Fawaz EN, Roth L, Fawaz G. The enzymatic estimation of inorganic phosphate. Biochem Z 1966;344:212.
29. Morris DL, Ellis PB, Carrico HJ, Yeager RM, Schroeder HR, Schroeder JP, Albarella JP, Boguslaski RC. Flavin adenine dinucleotide as a label in homogeneous colorimetric iminunoassays. Anal Chem 1981;53:658.
30. Fossati P, Prencipe L. Serum triglycerides determined colorimotrically with an enzyme that produces hydrogen peroxide. Clin Chem (Winston-Salem, NC) 1982;28:2077.
31. Ewen LM, Spitzer RW. Improved determination of prostatic acid phosphatase (sodium thymolphthalein monophosphate substrate). Clin Chem (Winston-Salem, NC) 1976;22:627.
32. Babson AL, Philips GE. A rapid colorimetric assay for serum lactic dehydrogenase. Clin Chim Acta 1965;12:210.
33. Szasz G, Gruber W, Bernt E. Creatine kinase in serum. 1. Determination of optimum reaction conditions. Clin Chem (Winston-Salem, NC) 1976;22:650.
34. Hoever M, Zbinden P. The evolution of microarrayed compound screening. Drug Discover Today 2004;9:358.
35. Hargis LG, Howell JA, Sutton RE. Ultraviolet and light absorption spectrometry. Anal Chem 1996;68:169R–183R.
36. Rooze H. Kinetic-spectrophotometric determination of trace levels of iron. Anal Chem 1984;56:601.

Further Reading

Evenson MA. Chapt. 3: Spectrophotomtric Techniques. In: Burtis CA, Ashwood ER, editors. Tietz Textbook of Clinical Chemistry. 3rd ed. Philadelphia (PA): W.B. Saunders; 1999. p 75–93. A widely used reference and textbook for clinical chemists. Chapter 3 covers both the concepts and the instrumentation of spectrophotometry.

Hargis LG, Howell JA, Sutton RE. Ultraviolet and light absorption spectrometry. *Anal Chem* 1996;68:169R–183R. A review includes a comprehensive list of reagents used for the spectrophotometric analysis of organic and inorganic compounds. It also includes a brief description of data processing and instrumentation. A total of 440 literatures are cited in this article.

Mikkelsen SR, Corton E. Bioanalytical Chemistry. Hoboken (NJ): John Wiley & Sons Inc., 2004. Chapt. 1. Spectroscopic Methods for Matrix Characterization. Chapt. 2. Enzyme. These two chapters focus on the biochemical and biomedical applications, especially to the spectroscopic measurement of protein concentrations and enzyme kinetics.

Pesce AJ, Frings CS, Gauldie J. Chapt. 4. Spectral Techniques. In: Kaplan LA, Pesce AJ, Kazmierczak SC, editors. Clinical Chemistry: Theory, Analysis, Correlation. 4th ed. St. Louis (MO): Mosby; 2003. p 83–106. This chapter includes a good description on the design of various types of spectrophotometers.

Skoog DA, West DM, Holler FJ, Crouch SR. Fundamentals of Analytical Chemistry. Brooks/Cole: 2004. A textbook in analytical chemistry with extensive coverage of both the principles and applications. Part V, Spectrochemical Analysis, includes: Chapt. 24. Introduction to Spectrochemical Analysis; Chapt. 25. Instruments for Optical Spectroscopy; and Chapt. 26: Molecular Absorption Spectroscopy.

See also ANALYTICAL METHODS, AUTOMATED; FLUORESCENCE MEASUREMENTS.

COMPUTERS IN CARDIOGRAPHY.

See ELECTROCARDIOGRAPHY, COMPUTERS IN.

COLPOSCOPY

MOSTAFA A. SELIM
ABDELWAHAB D. SHALODI
Cleveland Metropolitan
General Hospital
Palm Coast, Florida

Today, fewer women die annually from carcinoma of the cervix then in any period in modern history. Increased utilization of cervical cytology (Pap smears) and better assessment and evaluation of patients with abnormal Pap smears are important reasons for the progress in this field.

The Papanicolaou stained cytology smear is an invaluable tool for detecting cervical carcinoma. Such a test

depends on collection of exfoliating cells from the cervix, spreading the material on a glass slide, and staining with special stains in order to identify the microscopic structures in nuclei and the cytoplasm of the cell. An abnormal smear is called positive and a normal smear is called negative for malignancy. However, in recent years the incidence of false-negative smears has been recognized to be significant, varying between 15 and 22% (1–4). This high incidence of false-negative smears emphasizes the need for histological study of the abnormal cervix before any therapy is started, regardless of the findings from the cytology smear.

In order to fill the serious gap in the screening of cervical cancer and its precursors created by the high incidence of false-negative smears, colposcopy is utilized (3–6).

THE COLPOSCOPE

The word colposcopy simply means viewing the vagina. The colposcope is a binocular, long-focal-length, wide-field microscope with which it is possible to examine the epithelium and subepithelial vascular pattern of the cervix at magnifications varying from 6 to 40×. Figure 1 shows a typical instrument. A beam of light is projected between the objectives so that the cervix is well illuminated. To help the visualization of blood vessels, a green filter may be fitted to the light source. This adds contrast in viewing microvessels against the tissue. Photographic apparatus consisting of a prefocused lens system, a 35 mm camera, and a flash tube are attached to the colposcope in order to document the findings. The photographs obtained using 35 mm film are two dimensional (2D). To obtain three-dimensional (3D) images, the instrument may be fitted with a stereocamera that produces image pairs. These pairs give a stereoscopic effect when viewed in a special viewer. The colposcope, the camera, and the flash tube are mounted on a fixed, sturdy base with rack-and-pinion drives for fine positioning and focusing. Modern coloposcope is equipped with digital video camera interfaced microcomputer.

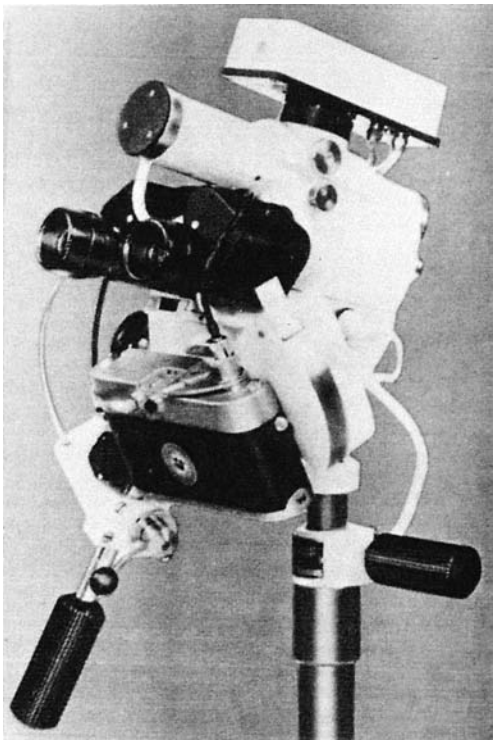


Figure 1. Colposcope with camera and flash attached.

FINDINGS

A colposcopic examination is considered satisfactory whenever it is possible to view all of the critical portions of cervical anatomy, and if an abnormal lesion is seen, the upper margin of this lesion must be adequately visualized. Specifically, it is necessary to view the entire transformation zone. The transformation zone is the lower segment of the endocervical canal. This area is normally covered by columnar epithelium, although through the process of metaplasia some areas or the entire zone can be transformed into squamous epithelium. The normal transformation zone is recognized by the presence of small glandular openings, nabothian cysts, and a normal pattern.

The procedure can be performed quickly and easily on out-patients without anesthesia. However, the gynecologist needs intensive training in order to master the interpretation of what they sees.

The colposcope was first devised by Hinselmann in 1925 at Hamburg, Germany. The instrument became popular in the German speaking and Latin nations. It was accepted in the English speaking nations in the 1960s. One of the reasons for the delay in accepting colposcopy was that all of the original studies and the terminology were written in German (7,8).

INDICATIONS FOR COLPOSCOPY (3–6)

Ideally all patients, during their gynecological examination, at one time or another ought to be examined colposcopically. However, since colposcopy is time consuming and requires specialized training, this is usually not possible, and thus the following indications are recommended:

1. All patients with abnormal cervical cytology.
2. All patients with an abnormal lesion on the cervix, vulva, or vagina.
3. Cases of persistent cervicitis, vulvitis, or vaginitis.
4. Pregnant patients with unexplained vaginal bleeding.
5. All offspring of diethylstilbestrol-exposed pregnancies.

Colposcopy plays an essential role in evaluating these patients. The colposcope helps to pinpoint the abnormal area. However, the final diagnosis has to await histopathological diagnosis. This tissue diagnosis is obtained through endocervical curettings, punch biopsy, and/or cone biopsy according to the individual case, as illustrated in Fig. 2.

Recent studies confirmed that colposcopy could be utilized in follow-up in conservative management of low grade

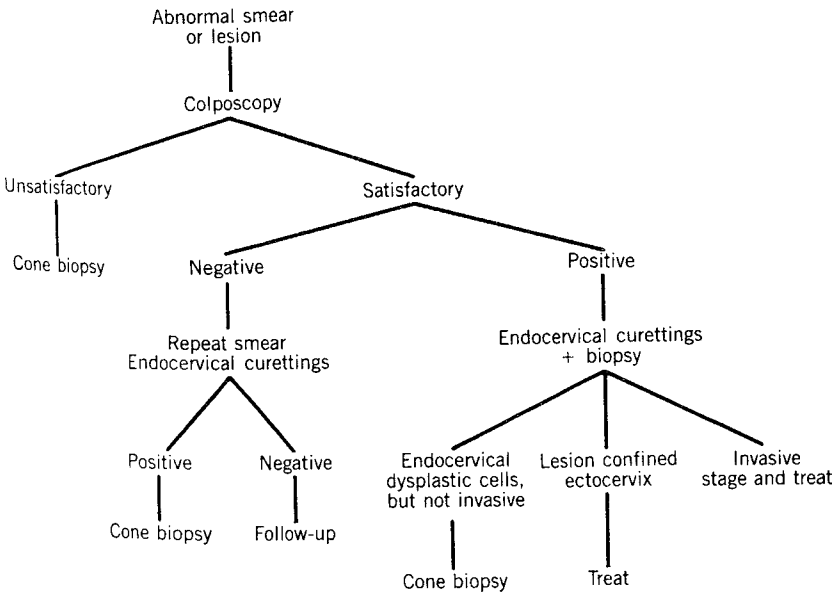


Figure 2. Protocol for management of patients.

precancerous lesions, as most of them regress on its own. In addition, the colposcope can help in detecting and guiding treatment of genital infections.

A movable, counterweighted arm is used for rough positioning of the colposcope. The arm is fixed to the examining table or to a heavy wheeled base (5,6).

There are several commercially available colposcopes. Manipulation, magnification, length, intensity, and type of green filter vary from one instrument to another. For detailed inspection of the vascular pattern, the magnification ought to be not $<14\times$ and preferably $16\times$ (7,8).

TECHNIQUE OF COLPOSCOPY

1. The cervix and the upper vagina are examined, at a magnification of not $< 14\times$, after excess mucus is removed by cotton swab moistened by physiological saline, which allows the subepithelial architecture to be seen in greater detail. To enhance visibility the green filter should be used.
2. Acetic acid (3%) is gently applied by a cotton swab. Abnormal epithelium will become whitish and sharply demarcated. The normal squamous epithelium appears pink, and columnar epithelium will have a grape-like appearance. The effect of the acetic acid will last $\sim$ 30–40 s.
3. Endocervical curetting and punch biopsies should be done for all unusual-appearing areas. Specimens should be put in separate formalin-containing bottles.
4. Bleeding is usually minimal and does not need any packing. However, during pregnancy, when vascularity is increased, Oxycel, Surgicel packing, or the use of Munsell's solution may be necessary.

Normal pattern of blood vessels. An example of a satisfactory colposcopic finding is shown in Fig. 3. The sketch identifies the various features of this view of the cervix. An example of an unsatisfactory view of the cervix is shown in Fig. 4. In this case the transformation zone as indicated in the sketch is not fully visualized; the columnar epithelium is not seen and/or the upper margin of the lesion extends into the endocervical canal and thus cannot be fully visualized.

In evaluating satisfactory colposcopic findings, many factors must be considered. It is beyond the scope of this article to provide a detailed description of the different patterns seen in colposcopy, but a general idea of what is looked for can be presented. In order to reach an adequate diagnosis, the colposcopist must consider the following criteria in observing the uterine cervix: (*1*) vascular pattern, (*2*) intercapillary space, (*3*) color and texture, (*4*) surface pattern, and (*5*) sharpness of the line of demarcation between the lesion and the rest of the cervix.

The lower segment of the endocervical canal that extends to the visible part of the cervix is the transformation zone. This area is originally and normally covered by columnar epithelium. Through a process of metaplasia, some areas or the whole zone can be transferred into squamous epithelium. The normal transformation zone is recognized by the presence of small glandular openings, nabothian cysts, and a normal pattern of blood vessels (Fig. 3).

Vascular Pattern

Normal epithelium vessels appear as fine dots or a network of capillaries. However, in abnormal pathology, the individual vessel becomes prominent, leading to punctuation (Fig. 5). A mosaic pattern is due to increased communication between the individual vessels, arranged parallel to the surface epithelium (Fig. 6). Atypical vessels are capillaries that are irregular in shape, size, course, and arrangement (Fig. 7).

Intercapillary Space. Due to rapid proliferation of the abnormal epithelium, the intercapillary distance between the vessels is increased. The intercapillary distance of normal capillaries varies between 50 and 250 μm with an average of 100 μm. The intercapillary distance in early

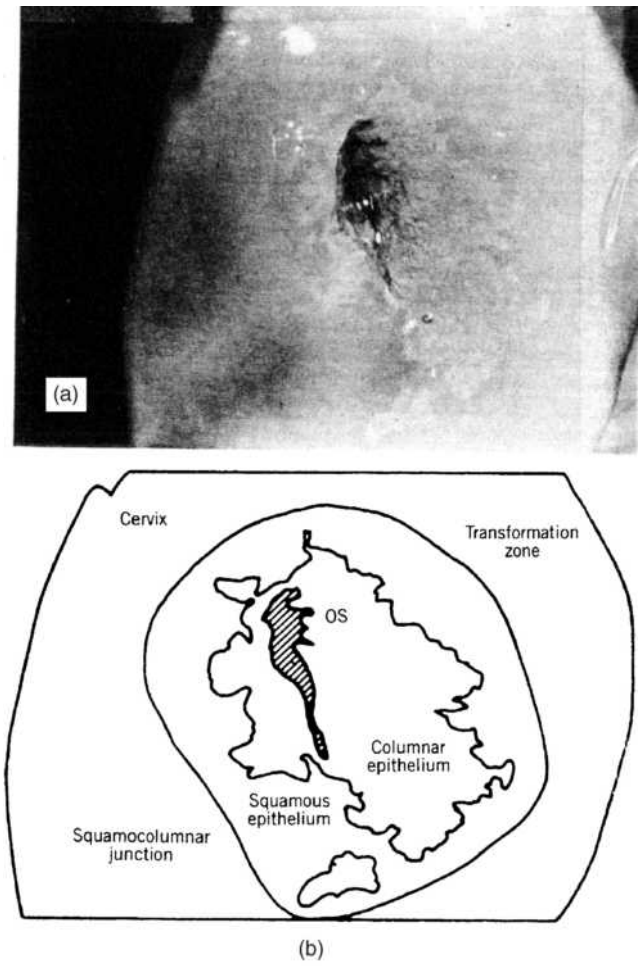


Figure 3. (a) Areas of glandular epithelium extending to ectocervix became whitish upon addition of acetic acid. Note the grape-like appearance of the glandular epithelium; no abnormal vessels or lesions are seen, and the lower end of the cervical canal is adequately visualized. (b) Sketch of the transformation zone, consisting of columnar and squamous epithelium. The squamous–columnar junction is part of the transformation zone. In the lower left corner of the diagram there is an island of glandular epithelium in the middle of the squamous epithelium. Note that the color of the squamous epithelium in the transformation zone is identical to that of the original squamous epithelium.

intraepithelial neoplasia (CIN 1) is 200 μm, while in severe intraepithelial neoplasia (CIN 3) it is 450–550 μm (6) (Figs. 5 and 6).

Color Tone. Abnormal epithelium is darker than normal epithelium, and upon addition of acetic acid it temporarily becomes whitish (Figs. 5 and 6).

Surface Epithelium. Abnormal lesions are uneven, granular, papillomatous, or nodular (Figs. 5–7). Normal squamous epithelium is smooth, whereas columnar epithelium has a grape-like appearance (Fig. 3).

Line of Demarcation. The abnormal epithelium is usually raised and well demarcated from the normal epithelium, especially after addition of acetic acid (Figs. 3–7).

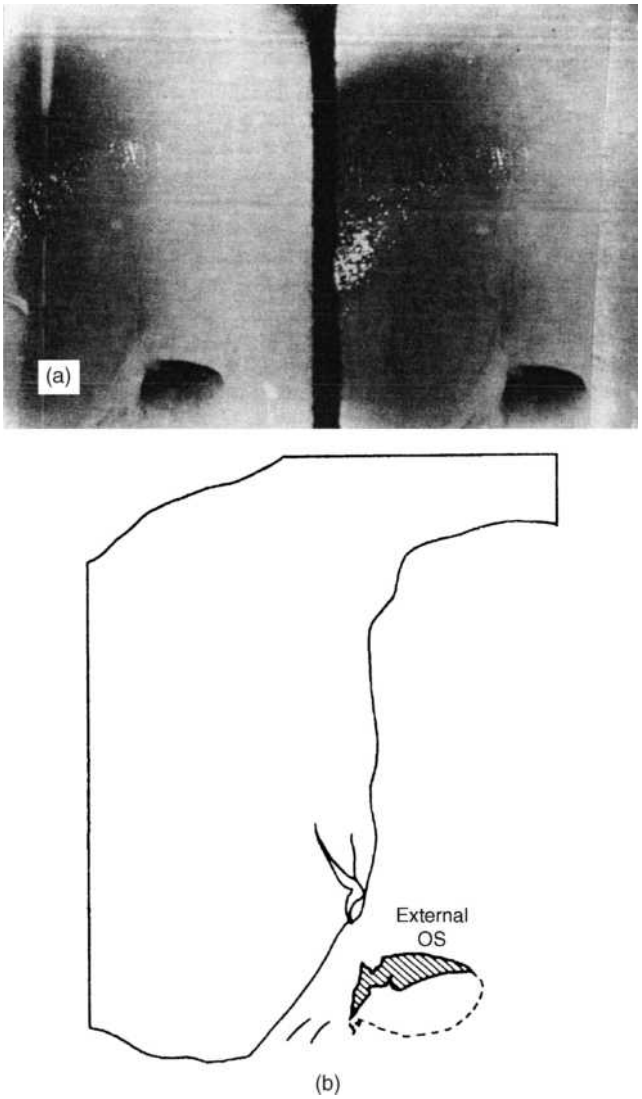


Figure 4. (a) Unsatisfactory view of the cervix and (b) accompanying sketch. This cervix is covered completely by squamous epithelium; no glandular epithelium is seen, including the lower endocervical canal. The left side of the cervix has a large red region. This region is covered by newer squamous epithelium, which is more transparent so that the normally arranged vessels are seen though it. The entire field of the figure is covered by squamous epithelium. Neither glandular epithelium nor the squamocolumnar junction can be seen. This represents an unsatisfactory colposcopic examination.

ACCESSORY INSTRUMENTS TO COLPOSCOPY

The best examination table is one that can be adjusted for height and tilt, is in the lithotomy position, and is preferably electrically operated. Punch biopsy forceps ought to be sharp with square jaws (Fig. 8) in order to obtain well-oriented, adequate tissue for pathological examination. In addition, in order to prevent loss of fragments of tissue obtained by endocervical curettings, the curette must be a closed-sided instrument to create negative suction (Fig. 9).

To aid in visualization of the endocervical canal, which cannot be seen by a colposcope, a Hamou microcolpohysteroscope can be utilized (Fig. 10). The technique utilizes

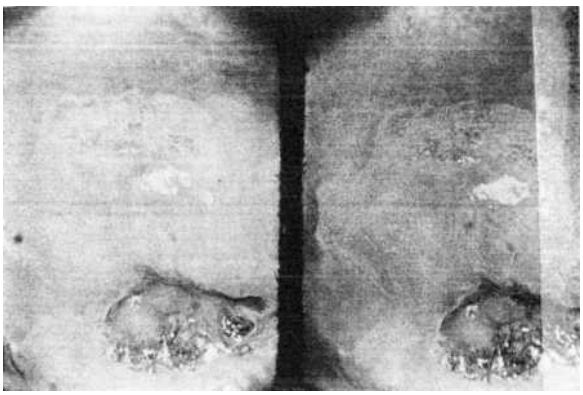


Figure 5. Punctation. Large, well-defined areas of the transformation zone became white after addition of acetic acid. The areas are covered by punctate of blood vessels with wide intercapillary distances. These abnormal lesions do not extend into the canal, and the lower canal columnar epithelium is seen and is normal. Some glandular openings are seen in the lower right part of the figure.

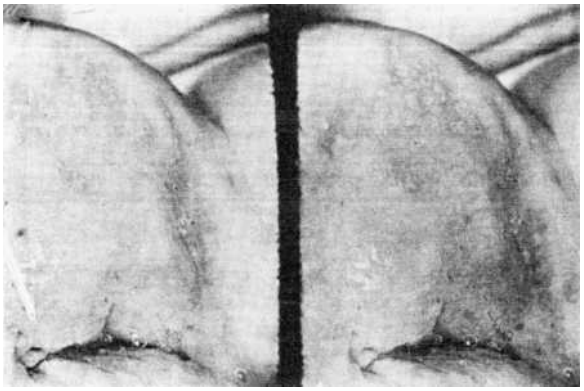


Figure 6. Large transformation zone covered by a mosaic pattern, inside well-defined borders. The areas became whitish after addition of acetic acid. The upper margin of the abnormal lesion can be adequately seen. The columnar epithelium is seen and appears to be normal. These findings are typical for carcinoma *in situ*, a precancerous lesion.

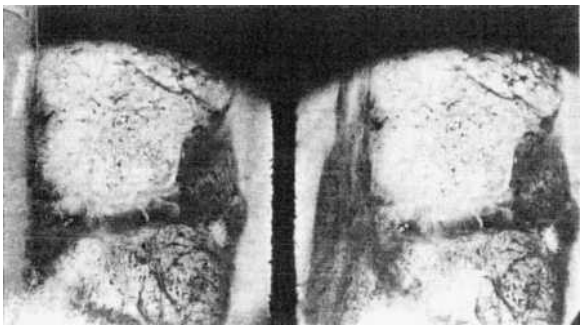


Figure 7. Highly abnormal cervix covered by atypical vessels with increased intercapillary distance. It is nodular and shows an exophytic growth pattern with a whitish and glossy appearance. These findings are indicative of invasive carcinoma.

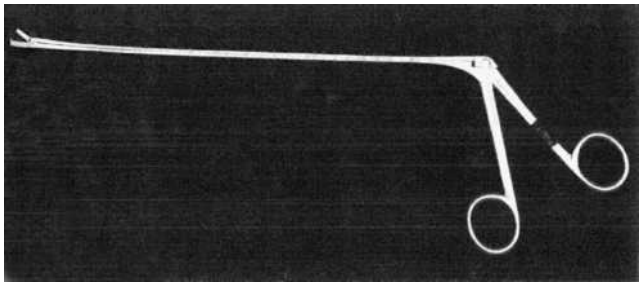


Figure 8. Punch biopsy forceps. Note the square jaw necessary to obtain an adequate and well-oriented biopsy for pathological examination.

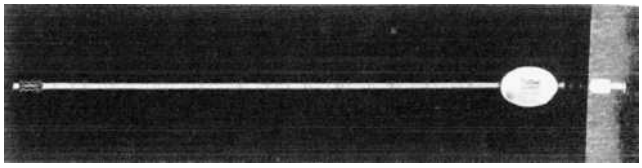


Figure 9. Curette. Note that it is serrated and open only from one side in order to prevent loss of tissue. The stem is hollow to create a suction effect.

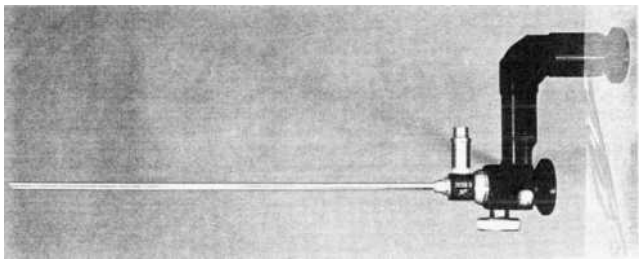


Figure 10. Hamou microcolpohysteroscope. Note that it has the same caliber as the curette and has an outlet for a fiberoptic light source.

the contact technique, thereby achieving $\sim 60\times$ magnification, and does not need any transmission media other than the normal mucus secretion of the endocervix. This method can aid in assessing the extent of the disease and may aid in defining the extent of cone biopsy, whenever it is needed to investigate or to treat the disease (9).

Digital Colposcopy (10–12)

Digital colposcopy is an improvement on regular colposcopy. Integrating video camera interfaced microcomputer and using real time image achieve this. Such arrangement allows computerized manipulation of the image signal. In order to manipulate the image by the computer it needs to be converted into matrix number. Each number is representing one point in image matrixes, called pixel. Each pixel has a value that corresponds to discrete gray level. The computer then converts the number back to analogue for display on the video.

Regular colposcope uses green filter to enhance visualization of blood vessels and discoloration. However, this has its limitation. Digital colposcopy has different ways to

improve on this (e.g., spatial filtering the image). These are pixel-by-pixel process that creates a new image, where the pixels values are determined, taking into account the value of neighboring pixels.

A very significant advantage to digital colposcopy is that it allows storage of colposcopic images in digital format on different media. This gives an unlimited ways of storage and filing.

Digital colposcopy is useful in metric measurement of the lesions and accuracy of measurement improved by the new techniques and software. This measurements and the improvement in digital filing helps the clinician in accurate follow-up of the lesions for regression or progression. Recent clinical studies confirmed that it is possible to transmit the images for consultation: which proved of help to areas lacking specialists in the field of oncology. In addition digital colposcopy helps in quality control, follow-up and in teaching.

SUMMARY

The manuscript describes the history, the instrument of colposcope and the advances into the digital age. The significant clinical application of this instrument in diagnosis and treatment of precancerous,cancerous lesions of the cervix is described. The terminology used to explain the vascular changes in the cervix is defined.

BIBLIOGRAPHY

1. Selim MA, So-Bosita JL, Blair OM, Little BA. Cervical biopsy versus conization. Obstet Gynecol (NY) 1973;41:177–182.
2. Selim MA, So-Bosita JL, Neuman MR. Carcinoma *in situ* of the cervix uteri. Surg Obstet Gynecol 1974;139:697–700.
3. Selim MA, Vasquez HH, Masri R. Indications and experience in colposcopy in management of cervical neoplasia Surg Obstet Gynecol 1977;149:529–532.
4. Selim MA, Razi A. Cryosurgery for intraepithelial neoplasia of the cervix. Cancer (Philadelphia) 1980;46:2315–2318.
5. Sootra-Gartaux, Carter I, Jourdau-DeSilva N, Decremax P. Regression of low grade epithelial neoplasia. Obst Gynecol 2004;104:751–755.
6. Norman JE, et al. An evaluation of economic and suitability of screening for chlamydia trachomatis infection in women attending antenatal, abortion, colposcopy and family planning clinic in Scotland UK BJOG 2004;111:1261–1268.
7. Kolshad P, Stafl A, editors. Atlas of Colposcopy. Baltimore: University Park Press; 1972.
8. Jordan JA. Colposcopy in the diagnosis of cervical cancer and precursor. Clin Obstet Gynecol 1985;12:67–76.
9. Soutter WP, Fenton DW, Gudgeon P, Shoup P. Quantitative microcolpohysteroscope assessment of the extent of endocervical involvement by cervical intraepithelial neoplasia. Br J Obstet Gynecol 1984;91:712–715.
10. Craine BL, Craine ER. Digital imaging colposcopy. Basic concepts and application. Obs Gynecol 1993;82:869–873.
11. Craine BL, Craine ER, O'Toole CJ, Ji Q. Digital imaging colposcopy: corrected are measurements using shape- from -shading. IEEE Trans Med Imaging 1998;17:1003–1010.
12. Schadell D, et al. Suitability of digital colposcopy for telematic applications. Biomed Tech 2004;49:157–162.

See also CRYOSURGERY; CYTOLOGY, AUTOMATED; SEXUAL INSTRUMENTATION.

COMMUNICATION AIDS FOR THE BLIND.

See BLIND AND VISUALLY IMPAIRED, ASSISTIVE TECHNOLOGY FOR.

COMMUNICATION DEVICES

ALBERT COOK
University of Alberta

INTRODUCTION

Augmentative and alternative communication (AAC) systems supplement, but do not replace other modes of communication such as speech, gestures, vocalizations, or facial expressions. The need for AAC, may be *congenital in utero* or *perinatal* (e.g., cerebral palsy or developmental disability) or a*cquired* (neurological conditions). The severity of need varies from mild to moderate to severe based on physical, cognitive and linguistic involvement. The overall prevalence from mild to severe for AAC needs is 0.2–0.6% of the total population (1). The age range encompasses three distinct ranges: infant to preschool, school age to teenage, and teenage to adult. The selection of systems to meet the needs of individuals in these age ranges is influenced by the experience that the individual brings to the use of AAC and the degree to which language and speech have been developed prior to the need for AAC. If the person has developed speech and language, and then subsequently lost those abilities, it is very different from the child born without speech and language who has never had the opportunity to develop those skills. Alternative communication needs may also change over time. For example, the needs of children change as they develop cognitive and language skills. In contrast, some disorders are degenerative and result in loss of function and decrease in skills (e.g., amyotrophic lateral sclerosis, multiple sclerosis). Using current technologies, we are able to meet the AAC needs for children who have cerebral palsy or developmental disabilities, individuals with good cognitive skills, and adults with degenerative diseases. Our current approaches are less effective for individuals who have mental retardation, are ambulatory, have dual sensory impairment, traumatic brain injury or are elderly.

NEEDS SERVED BY AAC

There are two basic communication needs that lead to the use of augmentative and alternative communication systems: conversation and graphics (2). These two needs differ in many important aspects. Conversational needs are those that would typically be accomplished using speech if it were available. Examples are an informal conversation with a friend, a formal oral presentation to a group of people, a telephone conversation, or a small group discussion. Much of conversational use focuses on interaction between two or more people. Light (3) describes four types of communicative interaction: (1) expression of needs and wants, (2) information transfer, (3) social closeness, and (4) social etiquette. Expression of needs and wants is the most

basic of AAC needs and allows requests for objects or people to be made. Information transfer allows expression of ideas, discussion and meaningful dialogue. Social closeness refers to the ways in which communication serves to connect individuals to each other, regardless of the content of the conversation. Social etiquette is used to describe those formalities that we adapt to our listener. For example, students will speak differently to their peers than to their teacher. Graphic communication describes all the things that we normally do using a pencil and paper, computer, calculator, and other similar tools, and it includes writing, drawing, mathematics, and Internet access.

Rates of communication using speech vary between 150 and 250 words/min (4). In contrast, many AAC devices use a keyboard to generate messages. This can result in significantly lower rates of communication than for speech. For example, a trained, nondisabled typist can generate typed text during transcription at a rate of nearly 100 words/min. However, this is still only about two-thirds of the rate of speech. If this same typist is asked to compose rather than transcribe, then their rate will drop by 50% to a maximum of 50 words/min (5). Many people who have disabilities must rely on single-finger typing, and they may only be able to type at a maximum rate of 10–12 words/min. Scanning (see below) reduces the maximum rates to as low as 3–5 words/min. The great disparity in rates of communication between a speaking person and an AAC system user often results in the speaking person's dominating a conversation with a nonspeaking person. This renders the individual using an AAC device to a passive role in the conversation.

There are three types of graphic communication: writing, mathematics, and drawing/plotting. Since each of these types of graphic communication serves a different need, AAC devices designed to meet each type also have different characteristics. Writing results in an electronic (soft copy) or paper (hard copy) output. However, writing does not need to be via letters combined into words on a page. Other symbols (e.g., line drawings, pictures) can also be printed on paper and used, in place of written output. Since some devices allow the selection of whole words, which are then output to a printer, spelling is not a prerequisite for the generation of written output. Alternatively, a nonspeaking person can point to letters on a board and have an attendant write down the letters to accomplish writing tasks.

There are three types of writing: note taking, messaging, and formal writing, the requirements of which all differ (6). Portability is important for note taking since the needs may be in many different locations (e.g., home, library, school, job, or meetings). Note taking may require writing at a high rate in order to keep up with the speaker. Just as nondisabled persons typically use abbreviations and other shorthand notations, so do users of AAC writing devices. Words and phrases stored in an AAC device may further increase the rate of text entry. The difference between note taking and messaging is the recipient of the written output. Messaging typically results in a note made for another person to read. This affects the types of abbreviations and shorthand notations that are used. The writing rate may also be slower than note taking since the person receiving the message is not present and waiting for it. Individuals who have intellectual disabilities and use symbols for communication can also use messaging. For example, individuals living in group homes who wish to send messages to their families, but are unable to use voice communication over the telephone can use rubber stamps with the appropriate AAC symbols (7). These stamped messages are then FAXed to their significant others as a message. The family can respond by using a second set of rubberstamps with the symbols on them.

The most demanding type of writing is formal writing, including reports, school homework, writing for publication, and similar applications. Here, accuracy is of prime importance, with rate becoming secondary. Word processing allows accurate entry of written material, but some users may take up to several hours to create one page of written material. However, with an entry rate of 1.5 to 5 words/min, it can take 2.5 h or more, and the use of abbreviations and other input acceleration techniques is necessary to allow an individual to keep up with the demands of work or school.

The AAC systems for mathematics focus on the written manipulation of numbers required for arithmetic and higher mathematics, as opposed to calculator functions. The goal is to for the user to learn mathematics in a manner that is as close as possible to that used by nondisabled peers. Cursor movement is a key difference between written English (the cursor always moves left to right and moves down one line at the right margin) and mathematics, where the cursor moves left to right as we enter numbers to be added, but once we have a column of numbers, the cursor moves right to left as we enter the sum. Carry or borrow are concepts when learning to add or subtract, and children are taught to cross out the number at the top of the adjacent column and substitute the borrowed or carried value. These types of cursor movement are also required in a math worksheet for children who cannot use pencil and paper. Algebra requires special symbols (e.g., Greek letters) and the use of superscripts and subscripts. Statistics or calculus adds the need for special symbols such as summation signs and integral signs and the formatting of problems. Some commercial AAC devices include some or all of these mathematical functions.

The final type of graphic output is drawing used to convey information, to help us clarify our thinking, and for creative expression. Typically, we use pencils, paint and brush, or computer programs to draw. Many people who are unable to use these conventional means of drawing because of motor limitations utilize computer programs designed for drawing or plotting

CHARACTERISTICS OF AUGMENTATIVE COMMUNICATION SYSTEMS

The characteristics of AAC devices can be grouped into three major components shown in Fig. 1: (1) control interface, (2) processor, and (3) activity output (6). The control interface links to a selection method, selection set, and an optional user display. The processor is further broken down into components of (1) selection technique, (2) rate enhancement and vocabulary expansion, (3) vocabulary storage, (4) text

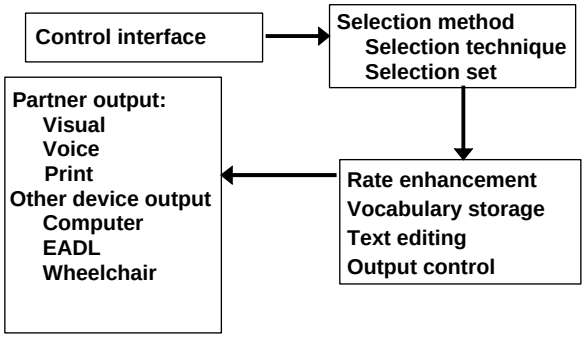


Figure 1. Characteristics of AAC systems.

editing, and (5) output control. The activity outputs to the communication partner include visual display, speech, and printer. A control port for external devices (e.g., computers or electronic aids to daily living) is sometimes included. Not every device includes all the individual functions shown in Fig. 1. In some cases, the functions shown in Fig. 1 are implemented in software using portable computers. Others are based on special purpose computers designed specifically for use as AAC devices.

Control Interface

The control interface is the way in which the user is able to make entries into the AAC device to generate a communication utterance or output. There are various types of interfaces based on the number of independent choices that can be made by an anatomical site or a combination of anatomical sites (e.g., hand, arm, head, chin, leg, foot). Keyboards, single or dual switches, and joysticks or multiple switch arrays are the most commonly used control interfaces for augmentative communication devices. A variety of switches used with AAC systems are shown in Fig. 2 (8).

Various types of keyboards are typically used for AAC devices. When a standard keyboard is not accessible due to limited fine motor control, enlarged keyboards are used with gross muscle movements (e.g, hand, foot, elbow activation). When range of motion prevents reaching all the keys, a contracted keyboard is used with finger or mouth stick activation. Keyboards may also be modified by the use of key guards that prevent accidental activation of keys. When keyboards of any type cannot be used, single or multiple switches are used. Some examples of these are shown in Fig. 3.

Selection Set

The selection set of the augmentative communication device presents the symbol system and possible vocabulary selections to the user. One type of selection set is the label on the keys of a keyboard. The selection set may include individual letters, words, phrases, or other symbols. The mode of presentation to the user may be display-based, chart-based, or memory-based (9). Display- and chart-based approaches present a list of vocabulary choices (e.g., words, letters, symbols), and the user chooses the item of interest from that list. Display-based methods are electronically generated and built into the device. Chart-based approaches are typically on a separate sheet that is used for prompting and often not needed after training. Both of these types require only recognition memory, that is, the ability to recognize the correct item when a list is provided. Memory-based presentation of the selection set does not include a prompting list and relies on recall memory. This is a more difficult task than recognition

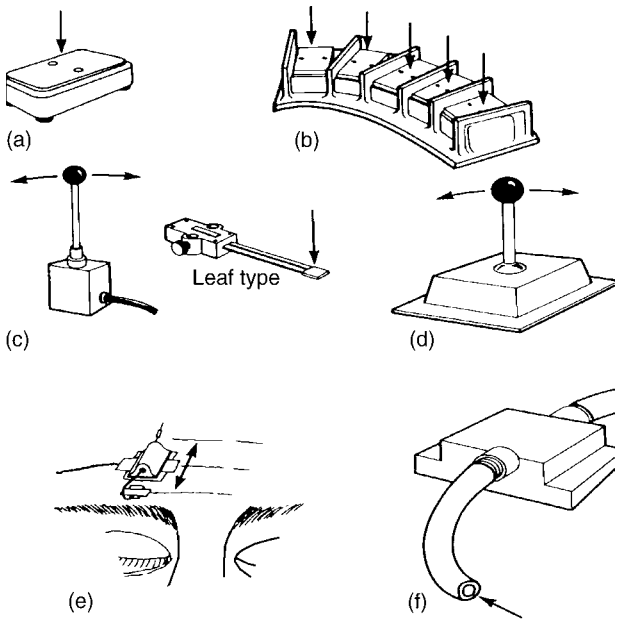


Figure 2. Typical interfaces for augmentative communication systems, (a) paddle (tread) switch; (b) array of paddle switches (slot switch); (c) wobble switches; (d) joystick; (e) brow wrinkle switch; (f) sip and puff switch (pneumatic). (From Electronic Devices for Rehabilitation, New York: John Wiley & Sons, 1984, with permission.)

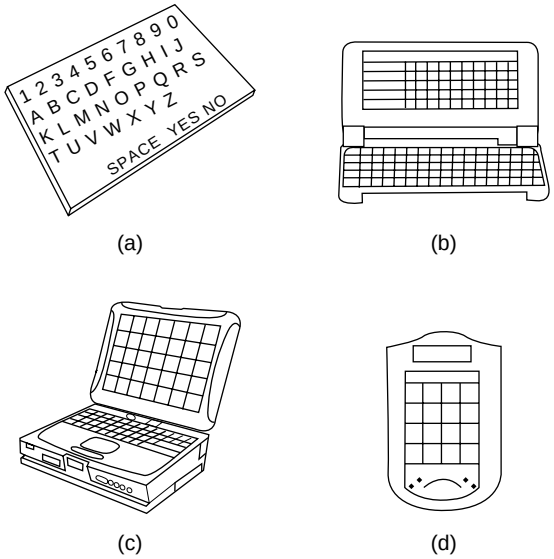


Figure 3. Direct selection communication devices. (a) Pointing boards with letters or other symbols, (b) hand-held keyboard, notebook-sized keyboard, and (d) palm-sized keyboards for high portability when good fine motor control exists.

for individuals with cognitive limitations due to developmental disabilities, stroke, brain injury, or similar conditions. Even in the absence of such disabilities, only ~ 200–300 vocabulary items and corresponding recall codes can be easily remembered and used. Thus, the use of display-based selection sets is desirable. However, display-based selection sets are often static, that is, there is one set of elements from which the user can choose. While this set of elements can be combined to access a larger set of vocabulary through codes or abbreviations, this converts a display-based selection set to a memory-based one.

One way of avoiding the problem of limited vocabulary choices associated with memory-based systems is the use of dynamic communication displays. An AAC device that has a dynamic display is shown in Fig. 2d. Dynamic displays change the displayed selection set when a new level is selected. Since the user's selection set is always updated on the display panel, it can be altered easily depending on previous choices. For example, a general selection set may consist of categories such as work, home, food, clothing, greetings, or similar classifications. If one of these is chosen, either by touching the display surface directly, using a mouse-driven cursor or by switch access, then a new selection set is displayed. This retains the display-based selection set while dramatically increasing the functional size of the selection set. For example, a variety of food-related items and activities (eat, drink, ice cream, pasta, etc.) would follow the choice of "foods" from the general selection set. The symbols on the display can be varied, and this changes the targets for the user. Since each new selection set is displayed, the user can depend on recall, not recognition memory.

A significantly different approach to the presentation of vocabulary choices to the Individual who uses AAC is implemented in Visual Scene Displays (VSDs) (10). Visual scene displays capture events in person's life on a screen. Hotspots are then liked to text messages that describe events, invite discussion or serve as prompts for conversational use. These VSDs offer a greater degree of contextual information to the Individual who uses AAC and communication partner information in order to support interaction. They enable communication partners to participate more actively in an interactive conversation and may represent either a generic context (e.g., a person's home) or a very specific personalized event (e.g., a birthday party). The screen contains pictures of the activity, place or event. A caregiver typically enters the vocabulary associated with the screen elements, although it could be the user who enters the text. A comparison of characteristics of a traditional grid AAC display containing vocabulary elements and a VSD is shown in Table 1. The biggest advantages of the VSD are the type of material that can be included in the display, the degree of personalization, the management of the display, and the methods available for concept retrieval. The functional uses of a traditional display focus on communication of needs, wants, and information exchange. The VSDs provide greater conversational support by allowing an interaction to be a shared activity and a potential learning environment for the user. While also including communication needs, wants, and information exchange, VSDs add a real element of social closeness. The VSDs can be applied to stimulate conversation between interactants, support play, shared experiences, and telling of stories. They also facilitate active participation of interactants during shared activities and can provide instruction to both the user and the communication partner through specific information or prompts. Specific populations that can be served by VSDs include those with cognitive limitations (e.g., Down's syndrome) and those with language limitations (e.g., aphasia, autism).

Another approach that allows dependence on recognition memory though a display-based approach is word prediction or word completion. These approaches can be used with any selection technique (11). In this case, there is a window on the screen that displays the most likely words based on the letters entered. To complete the word, the user selects the code (e.g., a number listed next to the word). If the desired word is not listed, the user continues to enter letters and the displayed words change as more letters are entered. For example, if the word "what" is entered, a word completion device may list "time", "is" "are" "can", etc., as choices to follow "What". The display is a display-based selection set that is dynamic based on user input. There are two approaches to the storage of items to be presented in a predictive system. Fixed dictionaries use a preselected stored word vocabulary that never changes. We can have different vocabularies for different contexts such as school, work or recreation. Other systems offer a menu of words using an adaptive vocabulary that is altered based on the user's own selections and is constantly updated. Character prediction phrase prediction are similar approaches using non-orthographic symbols or phrases instead of letters or words.

Selection Method

Two basic selection methods, direct selection or indirect selection are used in AAC systems (6). Direct selection is the fastest and easiest selection method to understand and use because each possible choice in the selection set is

Table 1. Comparison of Standard Grid AAC Display and Visual Scene Displays

Variable	Typical AAC Grid	VSD
Type of representation	Symbols, TO, line drawings	Digital photos, line drawings
Personalization	Limited	High
Amount of context	Low	High
Layout	Grid	Full or partial screen, grid
Display management	Menu, pages	Menu pages, navigation Bars
Concept retrieval	Select gird space, pop ups	Hotspots, speech key, select grid space

available at all times and the user merely chooses the one that he or she wants. Several examples of direct selection AAC devices are shown in Figure 2. Indirect selection is used to provide access for individuals who lack the motor skills necessary to use select directly and involves one or more intermediate steps between the user's action and the entry of the choice into the device. A variety of indirect selection AAC devices are shown in Figure 4. There are three types of indirect selection are scanning, directed scanning and coded access. All scanning-approaches rely on the basic principle of presenting the selection set choices to the user sequentially and having the user indicate when his or her choice is presented. Typically, the indication is by the activation of a switch by a single movement of any body part. A hybrid method called directed scanning allows the user to first activate the control interface (typically a joystick or other array of switches) to select a direction for cursor movement (vertically or horizontally) and then to stop the cursor movement with a switch at the desired element. Several types of scanning are used. In step scanning, the scan advances one step each time the switch is pressed. With autoscanning, the scan steps automatically until the switch is hit. An inverse scan reverses the process and advances continuously as long as the switch is depressed and stops when the switch is released. Each of these has advantages for different types of users.

Scanning is inherently slow, and there have been a number of approaches that increase the rate of selection (6). Just by placing the most frequently used characters near the beginning of the scan, the rate can be increased by as much as 30%. Many of the rate enhancement methods involve selecting groups of characters first to narrow the choices, then selecting the desired item from the selected group. Several types of adaptations are employed with scanning to increase the rate of selection. With group item scanning, items are clustered in groups. The first scan is through the groups and a switch press selects one. Then the items in the group are scanned sequentially. A row column scan is a group-item scan with the elements arranged in a matrix of rows and columns. The rows are scanned top to bottom, and then the columns are scanned left to right. A halving scan presents each half of the display until a switch is hit, then scans within that half.

In coded access, the individual uses a single switch or an array of switches to generate a unique sequence of movements to select a code that corresponds to each item in the selection set. The most common form of coded access used in AAC devices is the Morse code. Here the selection set is the alphabet, and an intermediate step [e.g., holding longer (dash) or shorter (dot)] is necessary in order to make a selection. Two-switch Morse code is also used in which one switch sends dots and the other sends dashes. Morse code was developed to be efficient and fast, and these features are exploited in its AAC applications. However, Morse code is a memory-based technique in general (although some display-based approaches have been used) and this can result in additional constraints on the user. The memory-based nature also makes it useful for individuals who have visual limitations that prevent them from using a display-based approach.

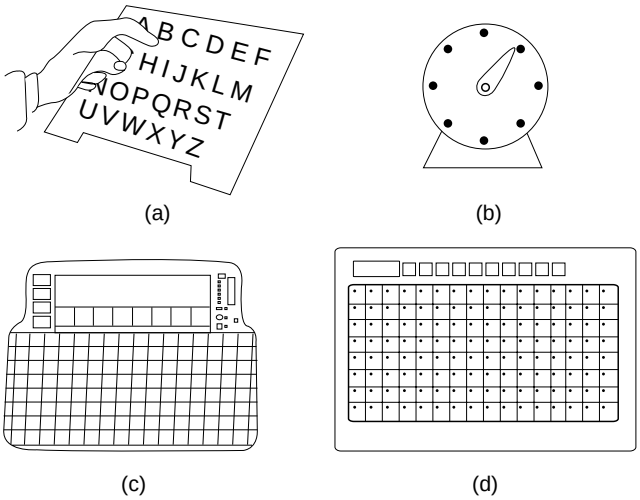


Figure 4. Scanning communication systems. (a) Simple pointing boards in which the listener points to each element and waits for a response, (b) circular scanning with up to 32 elements, (c) electronic row-column matrix scanning with picture display and voice output, (d) alternative electronic row-column matrix scanning with voice output. Parts (c) and (d) can have a variety of symbols (letters, words, pictures, line drawings) in the scanned elements.

Output Formats

Most AAC devices and some assistive technology applications rely on voice output. There are two types of speech output, differing in the manner by which the speech is electronically produced. These are (1) digital recording, and (2) speech synthesis. Digitized Speech is similar to a tape recorder. The speech is electronically compressed with up to a few minutes of speech (2–3 s/utterance) stored at any one time. A care provider can record any voice (male, female, child) or sound (e.g., laughter) and the user can play it back using either direct or indirect selection. One drawback is that all vocabulary to be spoken must be recorded and the user cannot produce a new or novel utterence.

Voice synthesis is electronic generation of speech using a mathematical model of the vocal tract realized in software or hardware. The two types of sounds in speech are voiced and unvoiced (a hissing sound similar to unvoiced sounds such as s or f). Both of these types of speech signal are used as sound sources for the vocal tract model. The parameters of the model are varied to produce speech in a manner similar to the variation of the tongue, teeth, lips, and throat during human speech. The parameters are either derived from actual speech samples (e.g., linear predictive coding-base synthesis) or from a set of parameters representing each phoneme ($\sim$64 phonemes are required for intelligible speech synthesis) or morpheme (combinations of phonemes-over 1500 are required for intelligible synthesis). Text can also be converted directly to speech by using a set of rules. The conversion of text characters into the parameters required by the vocal tract model in the speech synthesizer is accomplished by text-to-speech software. The synthesizer combines the phonemes

or morphemes into words. There are ~400 rules necessary for letter-to-phoneme conversion. Some systems also use morphonemic rules. About 8000 morphemes can generate 95% of all words. The major advantage of speech synthesis is that there is unlimited vocabulary as long as it can be represented in text strings.

Internet Access

Another common output for AAC devices is Internet access using the AAC device. While some AAC devices include basic computer functions (word processing, spreadsheet, presentation software, web browser), many do not. In the latter case, the AAC device is connected to a computer wither via hard wire or infrared link. In either case, any stored vocabulary or special access methods used with the AAC device are then available for on-line use. Further, many commercial AAC systems utilize portable computers with AAC software, and they can also function as Internet workstations. One of the most important communication functions accomplished via the internet is e-mail. Aside from the benefits that we all receive from e-mail (global access, low cost) people who have disabilities also have the benefit of composition independently and at a slower speed than face-to-face communication since the recipient reads it at a later time. The major advantage is that the person's disability is not immediately visible, and individuals who use AACs report that they enjoy establishing relationships with people who experience them first as a person and then learn of their disability (12). Information retrieval, socialization (e.g., chat rooms), development of literacy through large amounts of reading and writing, booking of airline or theater reservations, and general conduct of business from home without traveling to a place of business are other advantages of Internet access for individuals who use AACs.

AAC ASSESSMENT

Effective use of an augmentative communication device requires skills in several domains. These include gross and fine motor control required to make selections; visual, auditory, and tactile sensory capabilities; and cognitive and language abilities (e.g., the use of some symbolic representation).

Needs Assessment

The first step in an assessment aimed at the selection of an AAC system is the documentation of the individual's communication needs. The second step is to determine how many of these can be met through the individual's current communication methods. Finally, an AAC system is selected that will reduce as many of the unmet communication needs as possible through a systematic intervention.

The Participation Model (4) is a framework that focuses on identification of opportunity and access barriers to AAC use. Opportunity barriers are classified as policy, practice, attitude, knowledge, or skill (of support personnel). Access barriers include natural supports, environmental considerations, and user competencies. These are evaluated in terms of their potential to increase natural abilities (speech, gestures, vocalizations), improvement of communication through environmental adaptations or strategies, and profiles of the individual's abilities and skills. These are all measured and related to AAC use. Interventions are developed to address barriers resulting from opportunity, natural abilities, environment, or lack of an AAC system. Once identified, an intervention plan is derived for each identified barrier. For example, a policy may need to be changed, a teacher's attitude toward AAC altered, or the staff of a facility needs to be trained to develop their skills and knowledge of AAC. Natural ability interventions may involve speech-language pathology services to increase loudness or intelligibility of speech. Environmental adaptation interventions may result in a change in the layout of a classroom to place the individuals who use AAC in the center of the class rather than the periphery. The AAC system interventions determine the skills the user has and identify a system or systems that will be of use to that individual.

Physical–Motor Assessment

Physical assessment consists of identifying an anatomic site that can be used to indicate choices from the selection set. The primary sites are the upper limb (L/R), head and/or neck, eyes, leg, foot, or arm. These may be used directly or with a hand pointer to increase precision, or a mouth or head stick. Adaptations such as a keyguard or a hand splint may also be used. A variety of methods for using head control are shown in Fig. 5.

Sensory Assessment

Visual, auditory, and tactile abilities must also be assessed. Visual skills include acuity, tracking, and visual scanning. Auditory thresholds and speech perception should also be evaluated. Tactile response may be limiting if it is too sensitive or not sensitive enough. In the former case, the

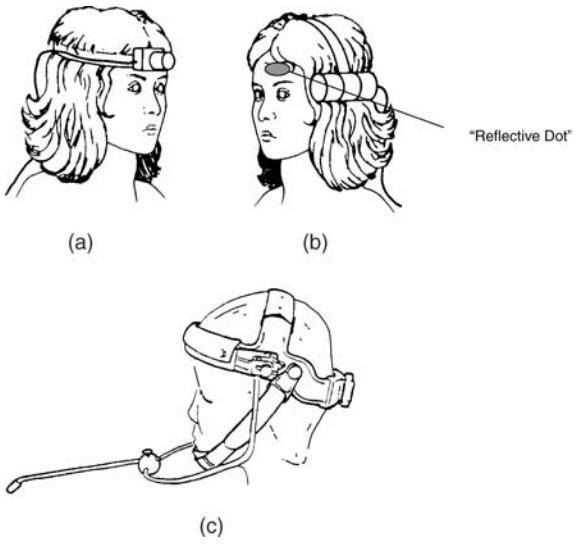


Figure 5. The head may be used as a control site by attaching a light beam (a): a reflective dot used with an infrared transmitter/receiver on the device (b), of a mechanical head pinter (c). part a (From Electronic Devices for Rehabilitation, New York: John Wiley & Sons, 1984, with permission, part c from Zygo Industries, Inc.)

person may be tactually defensive and not be able to use a switch or other interface that has a rough surface. If the person is insensitive, then they may not be able to sense when a switch or key is activated.

Cognitive and Language Assessment

A variety of symbols are used in AAC selection sets. These include real objects (including miniatures of object such as doll house furniture), colored drawings, line drawings, words, or letters. The symbol type for any given individual is selected through a formal assessment process (4). Language assessment related to AAC use is very difficult due to lack of expressive ability (i.e., the very reason for the need for AAC). Two approaches are employed. The first is single word vocabulary testing that measures vocabulary comprehension in relation to the individual's level of functioning. This assessment includes relationship concepts that have no real-world referent as well as traditional language sampling. The second type of language evaluation is a literacy assessment. This can include a reading evaluation in which both word recognition and reading comprehension are used (4). Several types of spelling tests are employed. Recognition spelling requires the user to pick the correct choice from a list of options. This requires only recognition memory and is easily accommodated into AAC devices. The second approach relies on first letter of word spelling and is related to AAC word completion techniques. Spontaneous spelling is the typical letter-by-letter text generation we all learn in school. This is the most flexible and powerful spelling skill since it results in the generation of vocabulary limited only by the persons knowledge, not by the features of the device.

A cognitive assessment may also be conducted to determine how the individual understands the world and how communication can be best facilitated within this understanding. No formal tests predict the ability of an individual to meet the cognitive requirements of various AAC techniques (4). Many cognitive tests require expressive language via AAC itself in order to accurately assess cognitive ability. For this reason, the individual's cognitive ability must be estimated to assess the probability of an AAC device being successful. There are several basic cognitive skills relevant to the use of an AAC system. These include: alertness, attention span, cause and effect, vigilance (the ability to visually and auditorially attend to a task and process information), expression of preferences, making choices, symbolic representation, and understanding of object and/or pictorial permanence.

Educational Assessment

The SETT model was developed to aid in the effective selection and use of assistive technologies, including AAC, in education (13). It consists of four elements: student, environment, task, tools. Each element Includes a set of questions that focus on the interrelationship between the elements to enhance classroom use of AAC. Student-related questions include What does the student need to do?, What are the student's special needs?, What are the student's current abilities? Environment questions comprise: What materials and devices are available?, What is the physical layout? Are there special issues?, What is the instructional arrangement?, Are changes planned?, What supports are available to the student?, What resources are available to those supporting the student? Task questions focus on: What activities take place?, What activities support the student's curriculum?, What are critical elements of activities?, How can activities be modified to meet student needs?, How can technology support the student's participation in the activities? Finally, tools questions include: What strategies might be used to increase student performance?, What no-tech, low tech, and high tech options should be considered? How might tools be tried with student in environments where they will be used? The SETT framework promotes team building. Consensus is built by using clearly understood language, requiring broad-based participation, and valuing input from all perspectives. Exploring environments and tasks strengthens the links between assessment and intervention. It is also necessary to develop a system of tools to enhance the student's abilities to address the tasks and build competency. The SETT framework can address and overcome many of the obstacles that lead to marginal student inclusion, general dissatisfaction, and device abandonment. It can also increase opportunities for success that come with assistive technology systems that are well matched to the student's needs and abilities.

TRAINING INDIVIDUALS WHO USE AAC FOR COMMUNICATIVE COMPETENCE

There are some disturbing statistics regarding AAC use. For example, Magilei and Sandoval (14) reported that 87% of parents reported that their child had access to assistive technology, but they had no training or technology assistance. Also, 33% of all assistive devices are abandoned, largely due to failure to meet the user's needs in community settings. Thus, it is essential that individuals who use AAC devices receive adequate training.

There are four basic types of AAC competence that must be developed in order for the user to be an effective communicator (15). Operational competence refers to understanding of the mechanics of making a selection, turning the device on or off, battery charging, and so on. This involves both the user and the caregiver. Linguistic competence is the understanding of the language elements (symbols, rules of language) that are included in the system. Social competence is the effective use of the AAC system in a functional manner to convey a message to a listener in a given context. Strategic competence deals with the strategies necessary to determine when to use one AAC mode rather than another and how to use it most effectively with a given partner. Communication partners must also be also trained. For children, the training of parents to recognize communication attempts and to understand the operational, linguistic, strategic, and social competencies is also important.

Both physical and communication skills are required for the use of an augmentative communication device. Communicative competence can be developed once sufficient skills are available in both of these domains to allow basic

communication. Physical skills are required to make choices from the selection, and these skills may need to be developed in order for the individual to use the device effectively.

For operational competence development, training may be through tutorials built into a device, supplied separately by a manufacturer on a CD/DVD or made available through a company website. A provider (speech-language pathologist, teacher, rehabilitation engineer) may also provide this training face-to-face. In order for the user to develop linguistic competence the AAC device user must understand the symbol system and rules of organization. The individual who uses AAC often must be competent in two languages: the spoken language of the community and the language of their AAC device (15). Development of linguistic competence may require many hours of practice, often built around a functional task.

Many users of AAC devices have little or no experience in social discourse, and training in social competence is required. Rules of conversation are altered for AAC use, and the perception of the individual by their communication partners is also different than conversations between two speakers. In order to be socially competent, the user must have knowledge, judgment, and skills in both sociolinguistic (e.g., turn taking, initiating a conversation, conversational repair) and sociorelational areas (e.g., understanding of interaction between individuals) (15). These skills are best taught in the contexts in which they are to be used, that is, training should occur in the community at school, work, shopping mall, rather than in a clinic setting. Training should be motivational, educational, and functional through the use of age and environmentally appropriate activities, such as playing a board game, which allows participation and multiple communication turns around a topic. The incorporation of creativity and fun activities into therapy sessions can lead to carryover of desired skills–goals and limit the amount of drill-like exercises.

Every user of an AAC device develops strategies that make use more effective. Strategic competence describes the degree to which the user of the device is able to develop adaptive methods to make the most of the device. For example, a child's speech may be better understood at home than at school. He will rely on the electronic AAC device more in school, but they will also develop strategies to make maximum use of both systems. One approach to strategic competence training is to simulate a situation, model the types of interaction likely to occur, and have the user "practice" the strategies and skills necessary to make it a success, followed by an actual situation in which the user goes into the community. If the provider accompanies the user they can then prompt, encourage, and help to clarify when necessary. This combination of clinic-based practice and community-based skill development is often very effective.

A seven-step process for developing communication competence in individuals who use AACs has been developed (16). These seven steps are (1) specify the goal, do baseline observations, (2) select vocabulary, (3) teach the facilitators how to support the individual who uses AAC in developing the target skill, (4) teach the skill to the individual who uses AAC, (5) check for generalization, (6) evaluate outcomes, and (7) complete maintenance checks.

VOCABULARY SELECTION

There are a variety of types of vocabulary that serve various needs. For conversational messages, greetings, small talk, information sharing, wrap-ups, farewells, and conversational repair are important types of vocabulary to have available. Small talk is used for initiating and maintaining conversations and is transition between the greeting and information sharing. It builds social closeness where content is less important than connection to another person. Sometimes scripts (complete dialogues or stories that are stored and replayed bit by bit) are stored in the device so they can be used over and over. For an adult, a story might be about a film that the user saw and enjoyed and wants to talk about. For a child, the story could be about a trip to the circus. Generic small talk is more general and can be used in different conversations with different people. The needs of the AAC vocabulary vary by context, communication mode, and user characteristics.

For preliterate users, a coverage vocabulary is needed to communicate essential messages including greetings, requests for objects and information, comments (e.g., "that's cool", "wow", "that is terrible"), emotional states (e.g., happy, sad, angry), and needs (e.g., "I feel sick", help me"). In order to increase linguistic competence, developmental vocabulary is also required (4). This vocabulary includes words and concepts that are not yet understood. The vocabulary is not selected for functional purposes, but rather to encourage language and vocabulary growth. In some cases, the individual may learn or memorize the location of utterances and become more functionally communicative than their cognitive or linguistic skills would suggest.

For literate users, there are a variety of vocabulary resources (4). Core vocabulary is used by a variety of individuals and occurs frequently. There are word lists based on successful patterns as well as those based on a specific user. In one study, Individuals who use AACs who were operationally and socially competent used a list of 500 words that covered 80% of total utterances for all users (4). Fringe vocabulary refers to words and messages that are unique to the individual. These include names of people, places, activities, and preferred expressions. Fringe vocabulary personalizes the AAC system and compliments the core vocabulary list. Items for this list are identified by informants, usually the user, family, and friends. Fringe vocabulary is selected based on initial items of high interest to the user that have the potential for frequent use. This vocabulary provides ease of production by the user and interpretation by the partner. One method of selecting vocabulary is to use environmental inventories. These inventories attempt to document the individual's experiences by noting precipitating events and subsequent consequences in communicative interactions. A pool of vocabulary items is reduced to a list of the most critical words that the user can manage. Another approach is communication diaries and checklists in which informants record the words and phrases needed by an Individual who uses AAC.

Privacy Issues in AAC

When collecting data on vocabulary used by an individual there is a real risk of invading their privacy, since stored information may include intimate utterances or vocabulary, personal data, or confidential information (16). The use of automatic monitoring devices that capture all of what the user generates by collecting AAC device output in an electronic form during use to assess compliance with a training plan is also a potential violation of privacy. On the other hand, the collection of data reflecting the experience of the individual user of AAC can provide important information related to the individual and more broadly to AAC needs. The user should be able to make the choice as to whether their data is collected or not, and should be able to turn off a data logger when a private conversation is desired.

BIBLIOGRAPHY

1. Blackstone S. Populations and practices in AAC. Augment Commun News 1990;3(4): 1–3.
2. Cook AM. Communication devices. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation. New York: John Wiley & Sons; 1988.
3. Light J. Interaction involving individuals using augmentative and alternative communication systems: State of the art and future directions. Augment Altern. Commun 1988; 4(2): 66–82.
4. Beukelman DR, Mirenda P. Augmentative and alternative communication, management of severe communication disorders in children and adults. 2nd ed. Baltimore MD: Paul H. Brooks; 1998.
5. Foulds RA. Communication rates for non-speech expression as a function of manual tasks and linguistic constraints. Proc Int Conf Rehab Eng 1980;83–87.
6. Cook AM, Husswy SM. Assistive Tecnologies: Principles and Practice. St. Louis: Mosby; 2002.
7. Brodin J. Facsimile transmission for graphic symbol users. Euro Rehab 1992;2:87–92.
8. Electronic Devices for Rehabilitation. New York: John Wiley & Sons; 1984.
9. Vanderheiden GC, Lloyd LL. Communication systems and their components. In: Blackstone S, Bruskin D, editors. Augmentative communication: an introduction. Rockville, MD: American Speech-Language and Hearing Association; 1986.
10. Blackstone S. Visual Scene Displays. Augment Commun News 2004;16(2):1–5.
11. Swiffin AL, Arnott JL, Pickering AA, Newell AF. Adaptive and predictive techniques in a communication prosthesis. Augment Altern Commun 1987;3(4):181–191.
12. Blackstone S. The Internet: what's the big deal? Augment Commun News 1996; 9(4):1–5.
13. Zabala J. The SETT Framework: Critical Areas to Consider When Making Informed Assistive Technology Decisions. Proc Closing Gap Conf 1995.
14. Magilei A, Sandoval L. Creative Therapy Activities Using AAC for Adolescents and Adults. Proc CSUN Conf 2003.
15. Light J. Toward a definition of communicative competence for individuals using augmentative and alternative communication systems. Augment Altern Commun 1989;5(2):137–144.
16. Light JC, Binger C. Building communicative competence with individuals who use augmentative and alternative communication. Baltimore, MD: Paul H. Brookes Publishing; 1998.
17. Wiliams M. Privacy and AAC. Alt Speaking 2000;5(2):1–2.

See also COMMUNICATIVE DISORDERS, COMPUTER APPLICATIONS FOR; LARYNGEAL PROSTHETIC DEVICES.

COMMUNICATION DISORDERS, COMPUTER APPLICATIONS FOR

CYNTHIA J. CRESS
JORDAN R. GREEN
University of Nebraska-Lincoln
Lincoln, Nebraska

INTRODUCTION

A variety of relative inexpensive, portable, and easy-to-use computer devices are now available to educators, researchers, clinicians, clients, and other independent users. This article includes discussion of software and hardware applications for reprogrammable computer devices as well as discussion of devices with fixed programs that evolved from programmable computer technology. A communication disorder is defined as an impairment in the ability to (1) receive and/or process a message or symbol system; (2) represent concepts, messages or symbol systems; and/or (3) transmit and use messages or symbol systems (1,2). This impairment is observed in disorders of speech (use of the oromotor system to produce the movements and sounds of speech), language (comprehension and/or use of spoken, written, or other symbol systems), and hearing (detection, perception and processing of auditory system information, and associated language, speech, or interpersonal factors).

Types of communication disorders may affect specific areas of speech, language, or hearing, as well as broader aspects of communication and social interaction. Persons may have multiple disorders within these areas, as well as closely related disabilities such as swallowing impairments, physical impairments, cognitive impairments, and/or sensory impairments. Disorders may be congenital (present from birth) or acquired (through illness, injury, or late-onset disorders). Some disorders that are presumed to be associated with factors present at birth may not be demonstrated until later childhood. Children may experience speech and/or language delays in early childhood that are resolved with experience and not diagnosed as a disability. Some communication disorders are considered organic (associated with a physical or neurological cause), while other disorders may be functional (attributed to experiential or unknown causes). Communication disorders that are associated with systemic and/or neurological symptoms may continue to affect communication skills throughout the lifespan, although many persons develop compensatory skills to effectively produce, process, and comprehend spoken and written information.

Speech impairments tend to be grouped into one of three categories, depending on the aspects of speech most affected: disorders of articulation (difficulty producing speech sounds), fluency (unusual interruption in speaking rhythm, rate and/or repetition), or voice (difficulty producing appropriate loudness, pitch, quality or nasality) (3). Specific speech disorders that can affect multiple aspects of speech in children and/or adults include the following:

Apraxia: Neuromuscular deficits in planning and programming speech movements; if seen in children, usually called Childhood Apraxia of Speech.

Cleft Palate–Lip: Incomplete fusion of oral structures prenatally that may result in articulation and/or nasality problems.

Dysarthria: Speech problems that are caused by abnormal function of the muscles or nerves that control articulation, phonation, resonance, and/or respiration. Dysarthria is often associated with weakness or imprecision of movements; with severe dysarthria, as in some persons with cerebral palsy, the person may be considered "nonspeaking" and rely primarily on alternative forms of communication to speech *(Augmentative and Alternative Communication, or AAC).*

Laryngeal Pathology: Changes in voice quality due to injury, nodules, or tumors; a removal of the larynx (or laryngectomy) requires alternative forms of voicing and/or communication.

Language disorders are commonly distinguished as either developmental (congenital), or acquired. Three general categories of acquired language disorders in adults and older children include the following (3):

Aphasia: Impairment in understanding and/or producing language after language is developed, resulting from brain damage such as due to a stroke.

Dementia: General loss in mental functions, including language, due to pathological deterioration of the brain; this includes Alzheimer's disease.

Brain Injury: Acquired injury to the brain from trauma (Traumatic Brain Injury) or illness (e.g., encephalitis, meningitis) that results in a variety of communication, language, and/or cognitive processing disorders (e.g., memory, attention).

Developmental language disorders are presumed to result from conditions present at birth, and are usually manifested in early childhood. Common developmental disorders that affect language and may affect other cognitive or social processes include the following (3):

Specific Language Impairment: Significant deficits in language comprehension and use that are not attributable to hearing, cognitive, or motor disabilities.

Learning Disability: Difficulties in acquisition and use of listening, speaking, reading, writing, reasoning, and/or mathematical skills.

Dyslexia: Specific reading disorder with difficulties representing and analyzing sounds and print representations of words.

Central Auditory Processing Disorder: Difficulty identifying interpreting or organizing auditory information with normal hearing acuity skills.

Mental Retardation: Impaired cognitive function that affects a variety of skills including communication, self-help skills, and independence.

Autism Spectrum Disorders: Impairments in social interaction, communication, and restricted interests/routines that affect skills at relating appropriately to people, events and objects.

Categories of hearing impairment relate to the types and severity of hearing functions affected (4). Dysfunction of the outer or middle portions of the ear is considered conductive hearing loss, and dysfunction of the cochlea or auditory nerve is called sensorineural hearing loss. Hearing loss from either cause can be slight, mild, moderate, severe, or profound, depending on the sensitivity of the person's hearing in decibels. Persons with mild-to-severe hearing impairments in their better ear are considered hard of hearing, and often benefit from hearing aids to help amplify spoken language. Persons with profound hearing loss in both ears are usually called deaf, and may have speech or language delays secondary to the hearing impairment. Some deaf persons may communicate primarily through signed languages (such as American Sign Language) and learn English as a second language through visual and/or residual auditory input. Some persons with severe to profound hearing loss may elect to have a cochlear implant, a surgically implanted electromagnetic device to stimulate sensory input to the cochlea.

This article is organized according to the general type of computer application, then by specific applications to disorders of speech, language, and hearing. Types of computer applications in communication disorders include clinic administration and management, analysis of normal function, assessment, intervention, and assistive devices. Tables in each section present potential benefits and limitations of computer applications for each type of application discussed. These benefit and limitation tables are intended to briefly highlight some of the issues involved in computer applications and are not organized thematically or listed in any hierarchical order of significance. For consistency across different sections, the general term *client* has been used to refer to the user of the computer applications, since many programs developed for one population (such as children) can also be used with other populations in limited circumstances.

Computer applications listed are intended to overlap somewhat between different sections. For example, many of the techniques developed for analysis of normal function are also used to assess atypical function across all areas of communication disorders. Also, similar techniques have been developed in computer-aided assessment of some speech and hearing functions. As far as possible, programs and devices are described as types of applications rather than specific programs, which are updated and revised too frequently be considered in a reference chapter. Specific examples of software and hardware applications can be derived from the references and websites listed.

ADMINISTRATION–INFORMATION PROCESSING

Overview

Computer software designed for administration may apply to other functions within communication disorders; for example, word processing programs are potential tools for language intervention, and statistical programs may be linked to other analysis functions in research. This section will concentrate on devices and programs that

facilitate, but do not directly accompany clinical, teaching, or research processes.

Clinical Management Applications

Programs are available to collect, organize, store, and report clinical information. Examples of functions addressed by available management software include the following (5–7): (1) maintaining mailing and billing logs, (2) collecting client case history, (3) maintaining correspondence and form letters, (4) graphing and organizing client performance data, (5) retrieving or compiling cumulative information, (6) analyzing cost effectiveness or distribution of clinic activities, (7) coordinating inventory control or budget management, (8) maintaining attendance or scheduling for activities, (9) creating electronic spreadsheets, (10) managing a calendar or clinical team interaction, (11) generating individualized evaluation plans (IEPs), and (12) writing and editing reports.

Computer-based applications are also used to create or adapt clinical materials for speech-language pathology (5). Many of these applications use standard multimedia tools such as digital resource libraries, authoring programs, and integrated drawing, video and/or web resources (8). For example, clinicians might create customized articulation or vocabulary exercises for their clients to practice. Clients might interact with commercially available story programs and interface to individualized prompts for additional narrative retelling or speech tasks. Adults and children might create self-prompting materials or applications to remind them of sequences or strategies for difficult tasks (9). Persons with literacy difficulties might create visual and/or auditory supplements for reading and writing materials to provide multisensory information.

Teaching Application

Programs assisting educators in the field of communication disorders supplement textbook materials, provide shared electronic information files, or simulate aspects of communication disorders for demonstration and practice. In a classroom, computers can be used to visually demonstrate lecture points by producing speech waveforms, demonstrating language analysis, or manipulating audio signals during a lecture. Many programs that simulate or demonstrate testing procedures or communication problems can be utilized in the classroom or for individual student practice. Students may individually respond to group tasks during lectures, allowing for personalized feedback on concept acquisition (10). University networking systems allow for web-based student input and discussion, creation and posting of assignments, feedback and dialog with instructors, and regular updating of course materials and progress (11–13). These networking systems have created opportunities for interactive learning through distance education courses and programs in communication disorders (14,15)

Research Applications

Research-based software can collect or manage information through electronic databases, statistical analysis, data manipulation or integration, and grant or report writing assistance. Large bibliographic databases such as Medline or ERIC search for references for a requested topic, title, or description. The Child Language Data Exchange System (CHILDES) collects and maintains electronic transcripts of typical and disordered language samples available for open access and analysis by language researchers (16). Community or professional listservs and websites can share software, information, messages, or reports, usually for access to all interested persons.

Also, computer networks can link researchers as well as clinicians or clients to other computer users, information centers, or databases. Improved computer conferencing technologies allow real-time shared work environments at a distance, including joint editing of documents, spreadsheets, or data analysis. Web-based video conferencing technologies allow real-time interactions among researchers at multiple locations.

Summary for Administration–Information Management Functions

Potential Benefits for Computer Applications

Allows easy storage, recall, and modification of information.

Provides standard format for reports.

Facilitates documentation of results for report or grant writing.

Sorts information across categories or levels.

Can coordinate or code information for multiple variables of interest.

Provides clean copy of output.

Links computer users in real time or through listservs.

Allows shared workspaces and interactions among persons at multiple locations.

Enables quick access to large bases of information.

Allows shared research databases and remote control of computer functions.

Allows students to practice analysis procedures independently.

Computer-based coding requires researchers to convert subjective decisions into rules.

Potential Limitations of Computer Applications

May require more total time on clerical tasks instead of passing routine functions to clerical staff.

Large data systems have more to lose if the system "crashes".

Potential for data corruption or transmission problems, particularly during periods of high web traffic or network overload.

Information may be entered into computer by users with little clinical knowledge.

Limited monitoring and/or maintenance of many web-based listservs or websites.

Large databases may be difficult to manage and access.

Standard format may discourage originality in report writing.

Computer organization may promote biased or poorly designed research.

Clinicians may design activities to utilize new technologies rather than primarily to meet a client's communication goals.

Students may need more direct interaction or coaching for some types of activities.

Potential for completion of distance education class assignments that do not reflect independent student work.

ANALYSIS OF NORMAL FUNCTION

Overview

Many advances in the assessment and intervention of communication disorders have resulted from extensions of the study of normal speech, hearing, and language functions. Computer technology has expanded the scope of acoustic, aerodynamic, and physiological studies of typical speech. Computer technology has also expanded the scope, accuracy, and speed of analysis of some aspects of language and hearing performance. Many of the applications discussed for analysis of normal function are also used to detect atypical functions, and may overlap with information in the Assessment section.

Speech Analysis Applications

Generate Stimuli. In psychoacoustic studies, digitized speech signals can be created to experimentally manipulate the acoustic features of a stimulus (17). Programs can also randomize the presentation order of acoustic stimuli manipulations, or synchronize stimuli with other experimental. Digitized signals are also converted to analog signals for synthesizing non-naturally occurring signals, or controlling peripheral devices such as pure tone or waveform generators. Adaptive techniques have been devised to randomly present ordered stimuli, organize intervals between events, and modify stimulus levels according to previous values and subject responses. Such programs can respond differentially to indications of subject fatigue or performance variation.

Record Data. Multichannel digital recorders are used to simultaneously capture multiple analog signals from a speaker including acoustic, airflow, air pressure, lung volume, muscle activity, and oral force and movement. The number of channels that can be recorded is only limited by the recorder and available storage memory. Specialized analog transducers and amplifiers convert each signal type into a voltage that can be digitized. These signals provide an objective means to study the physiologic processes that underlie speech production and to establish normative reference data, which can be used to gauge the degree of impairment in individuals with disordered speech. Most modern speech laboratories rely heavily on analog-to-digital conversion. Digitized signals can be easily accessed for editing, analysis, modification, and integration with other data.

Digital audio recordings of speech sounds are the most widely studied aspects of speech. Commercial sound cards provide a sufficient sampling rate and signal/noise ratio for most analysis applications. Typically, a preamplifier is used to increase signal gain prior to digitization.

Computer applications have also been essential for recording speech and swallowing movements. Most knowledge of speech movements has been derived from strain gauge devices, that were mounted directly to the upper lip, lower lip, and jaw [e.g., (18,19)]. Optically based motion capture (OBMC) systems register facial motion in three dimensions (3D) during speech and are rapidly replacing strain gauge techniques. These systems use computer-based visual pattern recognition to extract the motions of passively illuminated markers that are attached to the face. The OBMC systems offer a number of significant advantages over other methods for registering speech motion. Specifically, subjects are not required to maintain a specific posture while speaking nor are they encumbered by wires or metal beams extending from the mouth. At present, OBMC systems provide the only suitable method for studying orofacial movements in young children (20). One significant limitation of OBMC is, however, that they can only record the motions of superficial facial structures.

Most of the information regarding the performance of the articulators inside the mouth during speech and swallowing (i.e., tongue, velum, and the pharynx) has been obtained using four technologies: cineradiography, videofluroscopy, X-ray microbeam, and Electromagnetic Midsagittal Articulography (EMA). Cineradiography is a filmed X-ray and is typically no longer performed because of concerns regarding radiation exposure. Videofluroscopy is an X-ray technique that reduces radiation exposure and is usually only performed on individuals undergoing a clinical assessment of swallowing (21,22). X-ray microbeam minimizes X-ray exposure by using predictive algorithms to maintain a focused X-ray beam on pellets attached to moving structures such as the tongue (23). Electromagnetic Midsagittal Articulography tracks the motion of electromagnetic sensors through a calibrated magnetic field surrounding the midsagittal plane of the head (24).

Manipulate Data. Once digitized, the data representing the various aspects of speech performance are displayed and analyzed using custom or commercially available software. Computer programs can extract spatial, temporal, and spectral information from digitized signal input. Signal conditioning techniques are used to filter, rectify, and differentiate digital signals. These manipulations can be performed by dedicated hardware or computer software.

Several manufactures have developed commercial software for capturing speech recordings and for synthesizing speech (25–27). For example, Barlow and colleagues (28,29) have developed commercial software and hardware for recording and analyzing speech airflow, orofacial force, and orofacial reflexes. In contrast, commercial products for analyzing speech movements and muscle activity have not been developed. Presently, programming environments such as MATLAB (30) and Lab View (31) provide researchers who only have a moderate degree of programming

experience a means to develop custom analyses to meet their specific needs.

Hearing Analysis Applications

Research in hearing science utilizes many of the same techniques of stimulus generation, recording, manipulation, storage, and control as speech science. For example, computer-generated tones or speech can be used in audiometry to maximize control over stimulus characteristics such as frequency or timing (32,33). Computer models of auditory processes can simulate relative effects or differences in intelligibility in response to systematic changes in the auditory system (34,35). Computer systems can produce randomly ordered sounds that are necessary for some types of auditory physiology and psychoacoustic research, as well as store and index massive databases of data.

Signal-averaging is essential for computer analysis techniques, such as acoustic emittance or auditory evoked response (AER), that require separation of a signal from background responses (36). The AER systems analyze brainstem responses to sound by averaging EEG measurements for repeated sound presentations. Modifications can include nonlinear gating of signal initiation that is not possible with analog signals.

Language Analysis Applications

Analysis of normal linguistic function with computers utilizes descriptive techniques to track developmental progress and maintain databases for language clients, or databasing systems to compare language development patterns observed to those of other populations (37). Descriptive analyses currently available summarize lexical and syntactic functions in normal or disordered language development, such as the Systematic Analysis of Language Transcripts (SALT) (38). Over the past 20 years, SALT analysis has provided standards for language transcript coding and analysis in comparison to normative expectations, including school-based language assessments (39).

The relative speed and standardization of these transcript and analysis systems enables the collection of large and accessible databases of language behavior and variation. One of the largest databases in the study of language development is the Child Language Data Exchange System (CHILDES) (16,40). Continuing development of databases provides accessible digitized audio and/or video transcripts for comparison of performance of children and adults, not only with typical and multicultural development, but Down syndrome, specific language impairment, aphasia, or focal brain injury (37).

Summary for Analysis Applications

Potential Benefits of Computer Analysis

Makes possible complex transformation of a speech signal.

Enables precise timing manipulations for stimulus intervals and randomization.

Allows precise editing, analysis, modification, and recombination of signal or subarrays of signals, without splicing of tapes.

Easier to incorporate probability theory and statistics in analysis process.

Enables selective analysis of separate muscle groups.

Can be used to record speech movements.

Can compare multiple sequences, continuous data points, or multiple plots of data directly with point-by-point analysis.

Can compute online analysis as study is conducted.

Digitally generated speech and/or auditory signals are less distorted and more easily controlled.

Digitally stored data does not degrade as fast as analog.

Signal averaging is necessary for evoked potential study because potentials of interest are small and contaminated by noise.

Possibility for richer and more flexible stimuli.

Analysis techniques can more easily borrow models from other disciplines.

Standardization of transcript collection and coding makes analysis across individuals easier, more reliable.

Computer databases can be collected and shared across research facilities.

Quicker analysis of routine elements of language sample frees more time for other types of analysis.

Language transcripts can be searched for items of interest quickly.

Potential Limitations of Computer Analysis

Limited and nonintelligent algorithms may reduce research questions to analyses of narrow aspects of behavior.

Assumptions for analyzing, storing, or manipulating signals or behaviors are implicit in programs and difficult to retrieve.

Time savings in analysis may be offset by increased transcription or data-entry time.

Created stimuli may distort relevant features of the natural speech event.

Analog-to-digital converters code information via interpretive process, producing an averaged rather than exact representation of continuous signal.

Limited normal or disordered databases restrict interpretive power.

Promotes tendency for researchers to limit field of study to only what the computer can analyze.

Faster analysis may only provide faster mistakes.

ASSESSMENT

Overview

Assessment for communication disorders involves (1) describing an individual at a given time with measures of communication functions, (2) comparing an individual to normative data, (3) extrapolating information to predict behavior over time, (4) profiling abilities or deficits to aid in planning intervention, and (5) providing an objective means of following and recording progress. The format

for most computer-administered assessment programs follows the sequence of presenting the stimulus, accepting and evaluating the response, and providing a detailed report of results with optional storage of baseline scores. More complicated programs will offer variable depth testing in which the program evaluates client responses on line and adjusts test content to either extend testing in a problem area or skip questions in areas of high success. Some analysis programs do not test client performance directly but accept information on test performance and provide detailed analysis of that information, often with recommendations for further evaluation or intervention.

Speech

Dedicated speech devices (see also assessment procedures in the section Assistive Devices) have been used to collect acoustic, aerodynamic, and physiological data for dysarthric and apraxic speech (28,29). These types of data are used for both assessment of current function and tracking of intervention progress in children and adults (41,42). Recently, semiautomated algorithms (43) and automatic speech recognition technologies have been used to identify speaking characteristics of typical or disordered speech in children (44).

The vibratory characteristics of the vocal folds are assessed using acoustic voice analysis software (45) or direct imaging (46). Acoustic voice analysis systems have become commonplace in speech laboratories and clinics (42). Increased affordability combined with the continued need to objectify treatment outcomes ensures that these systems will be even more prevalent in the near future. Common acoustic voice measures include (1) fundamental frequency, (2) jitter (short-term variations in the period of the fundamental frequency), (3) shimmer (short-term variations in the amplitude of the fundamental frequency), and (4) a measure of glottal noise (47).

Many of the types of biofeedback used for articulation and voice assessment have also been applied to fluency disorders. For example, computer-based technology can track client speaking rate, pitch, intensity, stuttering frequency and duration, and percentage of utterances without stuttering (48). Other programs provide a user interface for clinicians to score perceptual judgments such as stuttering severity or number of syllables with stuttering [see (5)]. Available software to address fluency issues is listed at www.stutteringhomepage.com.

Both EMG and motion analysis technology can be used to quantify the movement deficits associated with impaired speech. Some abnormal features that have been identified using this technology include problems with force, endurance, movement displacement and velocity, interarticulator coordination, and movement pattern stability. The clinical implementation of motion analysis technology to assess speech motor problems has significantly lagged behind its application to assess gait and posture problems by physical therapists. Obtaining useful normative data for speech movements has been challenging because speech performance is highly variable across individuals. The cost and maintenance of this equipment are also problematic. These present challenges are not insurmountable, and it is likely that, with more empirical research, motion analysis will become an integral part of speech assessment in the not-too-distant future.

Many commercially available speech assessment programs provide computer analysis of clinician-entered data rather than direct assessment of function. Computer-based measures such as the Sentence Intelligibility Test (49) present word or sentence stimuli and record client response, but the clinician judges the intelligibility of units within the test items relative to nondisabled individuals. From this input, the program analyzes and presents the percentage and severity of intelligibility impairment. Other programs, such as the Logical International Phonetic Programs (LIPP) to analyze prelinguistic vocalizations of infants, interface between listener judgments and presentations of digitized infant vocal samples (50). Automatic computer analysis of infant vocalization samples can analyze relative complexity and variety of infant sound production, but does not yet directly correspond to phonological categories of listener-perceived judgments (51).

Hearing

Computer advances in hearing assessment (see also hearing aid evaluation in the section Assistive Devices) include improvement of test signal generation, different emphasis in audiological tests, and improved automated measurement of hearing (52). Computerized assessments of hearing provide direct measurement of acoustic characteristics across separate frequency components of a complex signal, including hearing performance while a hearing aid is worn (53,54). Computer-based audiometers can measure tinnitus or pitch problems, loudness problems, reaction times, and masking properties. Speech intelligibility under various listener conditions can be estimated from digital speech samples and automatically scored (55,56). Information gathered from these various audiologic tests can be used to optimize a client's hearing aid performance.

Other audiologic tests that are slow or impractical with traditional audiometry can be facilitated with computers. Evoked potential audiometry to screen or test thresholds of hearing or other features of the auditory system is made possible with computer signal averaging of the brain wave signals recorded at sound presentations (52). Automatic screening of such otoacoustic emissions has become standard universal practice for infants, particularly newborns (57–59). Automated visual response audiometry provides information on user spontaneous response to sounds for young children and persons with cognitive disabilities (60–62). Some kinds of auditory response tests have been computerized for more consistent administration and scoring with older children and adults (63).

Language

Computerized language-assessment tools may probe client linguistic production, comprehension, or problem-solving skills, elicit or describe language production, administer or score language tests, or analyze syntactic, semantic, lexical, or pragmatic elements of language samples. Many language assessments are designed for particular aspects

of specific language impairment or language–learning disabilities. However, language measures designed for one population can be applied to persons not in the target population who exhibit similar language characteristics if comparable normative data are available.

Several protocols are available for analyzing features of language transcripts; these protocols differ according to input format, number and type of linguistic analyses, speed, costs, and profile, training, editing, and search capabilities (37,64). Language sample tools can assist with two primary functions: data retrieval and tallying of codes entered by the clinician, and symbiotic programs that use algorithms to mark linguistic elements, with corrections by the clinician as needed (5). Most of these analysis programs are intended to profile language development and are standardized only for typically developing children, but some extensions of norms have been developed for children with language delays and disorders.

Each of the computerized language sample analyses has a standard input format and a range of options for language computations commonly used in the field, such as a Mean Length of Utterance (65). For example, various programs provide information on language characteristics such as conversational functions, grammar, semantic relations, vocabulary, narrative, and phonological patterns (66). Computer analysis can dramatically reduce assessment time while maintaining acceptable range of accuracy in clinical decisions to human-generated language analyses (67). Language sample analysis programs include the following: additional programs and reviews of features are available in Cochran (5):

1. Child Language Analysis (CLAN) (68).
2. Systematic Analysis of Language Transcripts (SALT) (38).
3. Computerized Profiling (69).

Computer-based phonological assessment programs provide analyses of standardized test results, systematic analysis of phonology, or phonological errors in a spoken transcript. Available phonology analysis programs fall into three types: analysis of phonemes and errors in different positions, distinctive feature analysis, or phonological process examination. These computer-based analyses are directly comparable to standard hand-scored measures, but can be substantially faster even for relatively novice users (64). Most programs require input of results of a particular language sample, and provide summary sheets of numbers or types of errors and characteristics of the systems impaired. Some also provide suggestions for target sounds in intervention. Features of available software addressing these goals, such as Programs to Examine Phonetic and Phonologic Evaluation Records (PEPPER) (71), Computerized Profiling (CP) (70), or the Interactive System for Phonological Analysis (ISPA) (71) are listed in Masterson, Long and Buder (72), and Cochran (5).

Other computer-based language tools have adapted strategies for collecting samples and/or scoring standardized paper assessments. Computers can facilitate assessment of language competence during conversational interaction by providing a dynamic context with shared reference to elicit spontaneous language (5). Programs have been developed to facilitate language sample analysis, test administration/scoring, and reading/writing skills in adults (41). Assessment techniques utilizing the computer for administration and scoring of paper language tests including vocabulary inventories, receptive vocabulary probes, and targeted probes of language fundamentals (5). Also, literacy assessments such as reading or spelling tests can be administered and/or scored by computer, and used to assist planning in language and literacy intervention (5).

Several standardized language assessments have been converted to computer-based administration and tested for equivalency to traditional administration. Results show variable influence of the computer interface, potentially related to the variables associated with the disability, task, access, and presentation. Some test administrations with simple recognition behaviors show equivalent results with preschool and school-aged children, using a variety of input devices (73). Other test adaptations indicate that computer administration may introduce variability and cognitive load from the task and input requirements, and that separate norms may be needed for computerized administrations of standardized tests (74,75). Selection of computerized language measurements should be consistent with the assessment principles of the clinician, and integrated with other direct measurements and observations of client language skills (76). Dynamic capabilities allow the potential for computer-based assessment of active language learning, rather than sampling of language skills already achieved. Jacobs (77,78) has developed a computer-based screening test to identify risk of language delay in children from multicultural backgrounds. Because the program uses video and audio stimuli to present and test language concepts in an invented language, it both samples children's dynamic potential to learn new language uses and avoids cultural bias of experience in their familiar languages. Continued progress in statistical analysis techniques for dynamic assessment and developmental data will facilitate the ability to base language assessment measures on predictions of language change over time (79).

Computer-assisted neuropsychological testing has been implemented with adults, for skills such as visual attention, memory, response speed, and motor tracking (41). Most language functions have not been sufficiently studied in typical communicators to identify language deficits associated with damage or dysfunction in specific neural regions (80,81). Additional integration of neural imaging and function measures may improve the diagnostic and descriptive potential of language-based assessments (82).

Summary of Assessment Issues

Potential Benefits of Computer Assessment

Computers prompt for missing information without making assumptions about nontested variables.

Standardizes presentation and analysis of assessment procedures.

Simplifies longitudinal assessment of function or learning.

Direct input of signals can improve accuracy and precision of measurement.

Computers are more reliable than humans at determining boundaries in acoustic or physiologic signals.

Possible to modify assessment techniques in response to change in client behavior response or test conditions.

Systems encourage accountability for decision making.

Analysis and assessment can be completed in real time, continuously as the program operates.

Larger databases can be accessible for test outcomes.

Easier to store data and track client progress.

Can temporally analyze and compare verbal and nonverbal aspects of transcript.

Signals can be generated, stored, and ordered for assessment presentation.

Evoked potential research is particularly useful for hard-to-test individuals, such as infants.

Simplifies routine assessment functions to encourage more clinician time spent in other types of assessment.

Dedicated specific language analysis tools can adapt rapidly to changing theories of language processes.

Enables assessment of comprehension with nonverbal stimuli and responses.

Comprehension of dynamic concepts can be assessed dynamically.

Computer can react to subject variables difficult to recognize, like fatigue, by pattern of response.

Artificial intelligence models can consolidate expertise from multiple clinicians and situations.

Computers can do multifactorial comparisons impossible by hand.

Enables noninvasive measures of internal functions.

Increase quantification of perceptual judgments.

Quantitative analysis is accessible to a wider range of clinicians with computer-based techniques.

Potential Limitations of Computer Assessment

Computer transfers experimental biases to implicit level, which is often inaccessible, and subject to misinterpretation.

Humans are better pattern recognizers than computers and are able to integrate a greater variety of factors.

Many assessment programs ignore or average out variability that may be one of the most interesting or useful aspects.

Need to determine if assessment tools themselves are valid before introducing computer for test administration and analysis.

Standardized tests are often designed to administer only one task.

Computerized assessments may overlook important factors or variations that are observed by clinicians.

Measurement techniques may be too time consuming to be clinically useful.

Techniques may require time, equipment, skills, and money not available to clinicians in general.

Clinicians may avoid responsibility for interpreting data and simply report computer results to clients.

Clinicians may be overwhelmed by the volume of data output from assessments, and potentially misinterpret results.

Comparative analysis is limited without normative database for the population assessed.

Danger for undetected errors in program's computations and output.

Clinicians may be tempted to entirely substitute computerized tests for informal tests.

If input to the assessment protocol is only estimated, computer will erroneously interpret that as fact.

Uniform assessment techniques may be uniformly mediocre without benefit of clinical intuition.

INTERVENTION

Overview

Note that relevant applications for clients with hearing impairments are distributed across three different sections according to intervention techniques addressing speech, language, or specific communication aspects of hearing impairment. Habilitation of hearing itself with hearing aids is discussed in the Assistive Devices section.

Applications of computers to intervention in communication disorders range from direct training and modification of measurable behaviors, to practice and feedback of speech/language functions, to presentation and teaching of facts or concepts, many of which utilize software and techniques originally targeted for regular education. Potential roles for the computer in fulfilling these tasks include the following [adapted from (83)]:

1. Tutor: Present information or tasks in sequences designed to achieve a desired behavior, provide drill, and practice.
2. Eyeglasses: Provide tools for amplifying and extending abilities (e.g., learning to influence the environment).
3. Mirror: Provide sensitive feedback on behavior.
4. Stimulus: Provide stimulating material for learning.
5. Access tool: Provide materials, real or simulated capabilities, or activities appropriate for tasks.
6. Communication: Input or output of spoken or written communication.

Tasks of the teacher or therapist in providing computer assisted instruction (CAI) include assessment of current and past performance and learning style, task analysis of what and how to teach, identification of instruction level, and evaluation of ongoing progress (84). Specific elements of tutorial programs can include (84,85): (1) pretest of requisite knowledge, (2) presentation of stimulus (text, graphics, videos, and/or speech), (3) acceptance and evaluation of response, (4) provision of prompts or cues for

responses, (5) feedback or reinforcement of correct responses, (6) change in difficulty based on evaluation of response, and (7) record keeping and string of data on client's performance.

Knowledge-based or intelligent tutoring systems utilize artificial intelligence models for describing skills and rules for deriving given behaviors, which are applied to normal and deviant behavior and modified through cumulative knowledge and experience. Research and therapeutic applications of artificial intelligence include the following (86): (1) simulation of perceptual, communicative, and cognitive processes to examine the nature of normal and disordered processes and recommend and test assessment or intervention techniques; (2) development of interactive systems which can evaluate, creatively adapt to responses, and vary stimulus or measurement techniques; and (3) development and testing of models of behavior that attempt to objectify rules and systematic influences of subjective behaviors.

Speech Intervention

Except for a few specific techniques or tests targeted at a given disability, computer applications will be discussed by general qualities of the speech signal addressed in intervention. Many of the techniques discussed here have been applied to various disorders, including voice, articulation, fluency, dysarthria, apraxia, and speech consequences of hearing disorders or physical handicaps.

Visible Speech. Techniques derived for the analysis and display of speech characteristics are used not only for research but also for intervention for speech disorders. Clients receive visible feedback about the qualities of their speech production as a cue to improve specific qualities of the production. Types of information displayed include the following: (1) analysis and display of acoustic parameters; fundamental frequency, intonation, amplitude, spectrum, and nasality (42,87,88); (2) direct representation of speech waveform characteristics (89); (3) interpretation of perceptual analyses or other classifications of speech parameters (90); (4) information on the progress and efficacy of the swallowing response (89,92), and (5) information on closeness of fit of client production to a template along acoustic or perceptual parameters (88,90,93).

Speech Feedback. Delayed auditory feedback, in which a speaker's voice is replayed to headphones after a preset interval, has been implemented in both stand-alone and in-the-ear devices to improve fluency (93). Other modifications of feedback to improve fluency include altered frequency, amplitude, or speech masking of the client's speech (93). Some types of speech feedback may be supplemented by tactile feedback sensors, for clients with limited hearing. Feedback on a wide variety of speech qualities can also presented in game or simple visual presentations for young children and other clients with basic communication skills (90,95).

Speech Movement Indicators. Real-time displays of the speech acoustic signal are the most widely used method to deliver feedback to patients regarding their speech performance. These displays are used to training features of speech such as loudness, rate, rhythm, intonation, or pitch (42). Acoustic voice analysis systems provide a convenient and noninvasive method to measure and track vocal characteristics throughout treatment (96,97).

A number of devices display aspects of speech movement in real-time, which can be used to train appropriate speech behaviors. Strain gauge and optically based motion transduction systems can be used to provide feedback to a patient of their lips and jaw movements (98). Coordination of speech breathing can be visualized using devices such as the Inductotrace (2) that transduce movements of the rib cage and abdomen. Dynamic palatography provides a means to display tongue to palate contact patterns in real-time during connected speech (99,100). Movement of the vocal folds can be obtained real-time through high speed stroboscopy (26,89). EMG is used to provide patients information about muscle activity patterns and may be used to increase or decrease muscle activity levels.

Speech Cueing. For clients with rate difficulties in speech, including fluency disorders, computer-based metronomes and cueing systems may facilitate fluent speech (93). These rate programs may be integrated with reading highlighting tasks, to reduce the difficulty of both the speech and language aspects of reading aloud. The feedback from a voice output assistive device may also provide cueing for clients with limited speech to elicit more frequent and/or complex speech signals (102,103).

Counseling and Support. Behavioral and emotional responses can directly affect speech disorders, and technology has been used to facilitate client counseling. For example, in cases of psychogenic aphonia (loss of voice for nonphysical reasons), direct feedback of the vocal function has been helpful in altering clients' perceptions of their own voice functions (89). Similarly, a biofeedback program was used in combination with behavioral treatment activities to increase client perception of control over stuttering episodes and reduce likelihood of relapse after treatment (102).

Hearing Intervention

While intervention with hearing-impaired persons utilizes techniques of both speech and language training (52,103,105), some communicative function training is unique to hearing-impaired persons. For example, several different programs, CDs, or websites use computer graphics and animation to teach sign language, finger spelling, speech reading, or simultaneous communication (8). Websites, such as www.deafed.net, provide coordinated resources for specialized interventions for hearing and associated speech/language skills.

Some intervention programs have utilized dynamic graphic displays to support practice or drill of specific speech and language skills associated with hearing impairments. Intervention systems have been established to reinforce therapy drills, provide remote-accessible tutorial programs, encourage speechreading practice and interaction with other deaf and hearing persons, and practice constructing language output with dynamic, visual input

(52,104). Aural rehabilitation for children with cochlear implants can be presented with specialized video games targeting specific listening and/or speech tasks. For persons with some hearing, computer speech output devices have been used to provide stimuli for auditory perceptual training. Direct intervention of hearing function through technology tends to be provided with assistive listening devices (see later section), although some systems have been constructed to target specific hearing features in listeners (107).

Balance and dizziness intervention can be assisted by computer-based presentation and evaluation of response to specific tasks. For example, technologies are becoming available to directly sample video images of a person's eye movements, standing balance, and walking (105). In therapy, computer technology can present virtual reality of environmental situations that gradually increase the balance challenge in response to the user's reactions in target behaviors.

Language Intervention

Language intervention utilizes many procedures and software programs in ways similar to computer-assisted instruction in general education. Programs designed to address general functions, such as problem solving, teaching school subjects, or playing video games, can be applied to language intervention as specific visual and auditory perception, reading, or sequencing practice tasks. The following are some of the ways computers can stimulate language development (5,83):

1. Language as subject matter: A wide variety of general and special education software is available which provides drills, tutorials, or examples of language skills such as spelling, grammar, vocabulary, and writing.
2. Language as currency: Language can be required as input or output in order to perform other activities such as games, problem-solving activities, or simulations; with some computer programs, the clinician can structure the program to accept either more or less complex language input, depending on the client's language skills.
3. Computer as a tool for linguistic communication: Clients with limited experience with either spoken or written communication can use computer systems such as word processors, computer networks, interactive terminals, speech output devices, or any of the augmentative communication devices to practice language and communication skills.
4. Computer as a tool for literacy: Computers can provide multisensory input and output to facilitate client reading and writing, as well as providing structure and expert feedback for problem-solving tasks such as editing grammar or creating a narrative.
5. Computer as a topic to generate conversation or interaction: Computers can serve not only as an instructional tool but as a source of problem solving and conversational topics embedded in the social context.
6. Computer as an instructor: Computer-assisted instruction is not limited to interactions between clients and computers alone in tutorial or drill-and-practice programs, but also as a facilitator of instruction as a triad with the clinician or educator.
7. Computer as an interactant: With some artificial intelligence programs, computers can react creatively to some aspects of client input and can function as an interactive learning tool; the programs vary in how much control the client has over learning and how the program adapts its linguistic responses or tasks.
8. Computer as a tool or reinforcer: Some types of video and accessory programs can present or manage tasks in ways that support the user's memory, cognition, language skills, and/or motivation.

Many general education software packages can be adapted for intervention with language disabilities. These educational software programs can be integrated with other technological resources such as websites and search engines to develop theme-based instruction with language support around an academic topic (11). Some modifications of standard educational software recommended for language and learning disabilities are (106) (1) expanding application of software beyond stated purpose, such as using alphabet programs for visual recognition or story programs for retelling a sequential narrative; (2) individualizing program characteristics for clients, such as level of difficulty and duration of stimulus, multimodal input or output, use of graphics or printed text, or tasks subdivision; (3) providing supplemental activities such as retelling a story or answering written or spoken questions to apply program information; and (4) individualizing content of entire programs for specific practice or focus by using authoring programs.

Some dedicated programs have been developed to practice specific domains of language skill in relatively focused decontextualized environments. Tasks addressed by software targeted for direct language practice include drills or tutorials for teaching vocabulary, figurative language, visual memory, phonological awareness, and grammar (90,109–111). Programs have been developed to modify characteristics of auditory speech input that are proposed to enhance skills in auditory and language comprehension, but data on the efficacy of these programs are mixed (112–114).

Most applications of technology with young children rely on computers or other technology as one element of an interactive or academic experience. For example, computers may assist in presenting or reinforcing concepts, supporting access to task materials, searching for information, or presenting stories or other written information to nonreaders (115). Children may develop cognitive concepts such as the connection between causative actions and effects through operation of a variety of switches that simplify the motor activities necessary to activate environmental devices or toys (116). Dedicated programs for early language intervention with preschool children address concepts similar to other language programs, usually embedded into teacher–child or child–child interactions: vocabulary,

following directions, differentiating letters–shapes–colors, and numerical concepts (117,118). Early intervention software and devices for infants and toddlers have been designed to engage the child in meaningful and cognitively appropriate play, with maximum technical support to allow easy access to social interaction, play, and tasks such as speech or writing (116,119).

Computer technology can provide ongoing support for reading and writing as well as tool for learning a wide variety of literacy skills. Early literacy skills can be promoted with programs verbalizing letters, words, and phrases as they are typed, highlighting text as it is read, and animating stories and words (120,121). Programs can provide scaffolding of conventional writing tasks to simplify writing or typing tasks and compensate for poor grammar and spelling, thereby encouraging maximum language-output capabilities (6,122–124). Voice capabilities of programs can support reading of books and text out loud for multisensory output, as well as specific word prompts for literacy tasks (125,126). Voice recognition programs can provide limited spoken input to writing programs, although most still rely on user correction of errors, even when combined with standard or custom text analysis programs (121,127). Programs to support and teach spelling and/or phonological awareness skills tend to use a combination of prompts, voice feedback, and systematic cues to identify and associate letters and phonemes appropriately (88,90,128).

Many persons with language impairments have additional congenital or acquired disabilities in cognition, memory, social interaction, and/or organization. Programs designed or adapted for clients with other developmental disabilities can address skills such as perceptual-motor skills, cognition, vocational skills, creative arts, self-help, memory-cuing skills, or social skills (129–132). Recommended ways of structuring computer-based tasks for clients with developmental disabilities include (133): (1) providing proper incentives for the client, (2) structuring programs to adapt a task to the client's behavior, (3) adjusting speed of presentation and amount of repetition, and (4) evaluating progress longitudinally rather than by initial performance. Because persons with cognitive or other developmental disabilities have difficulty generalizing experiences from one context to another, it is important to embed technology experiences into real life functional and social interactions, particularly for persons with social impairments, such as Autism Spectrum Disorders (134,135).

Available rehabilitative software designed specifically for adults with language and cognitive impairments addresses impairments of functions similar to other language disabilities in different instructional formats and styles. Some of these programs utilize alternative input systems when visual/perceptual or motor skills interfere with performance, such as for supported reading and writing (136,137). Examples of specific areas addressed for cognitive rehabilitation include concept training, memory, personal organization, planning, and perceptual processing (138–140). Technology can also serve as a memory or organizational tool for persons with brain injury, using common off-the-shelf devices such as handheld personal organizers and integrated video, and/or cell phone technology.

Summary of Intervention Issues

Potential Benefits of Computer Intervention

Computers are objective and reliable.

Potential for undivided attention, client-paced instruction from program.

Reinforcement immediate, contingent on user response.

Possible to present information multimodally.

Context and tool for exploration, problem solving, networking concepts.

Motivates interaction with computer as dynamic context in group interaction.

Provides experiences not otherwise possible, such as voice output of original text.

Intervention elements can be regulated or altered by clinician.

Allows intervention with using nonlinguistic and visual stimuli.

Can present dynamic stimuli difficult to represent in other ways, such as verbs or sign language.

Programs can respond immediately to subtle variations in speakers that are difficult for listeners to detect.

Can incorporate more than one instructional strategy, simultaneously or in sequence.

Collects, analyzes performance data on line.

Allows additional practice time outside of therapy, even at a remote distance.

Intelligent programs can have accumulated experience across clients and experts.

For responses that are variable, the user's best production can be the target production rather than a set template.

Potential Limitations for Computer Intervention

Danger that computers oversimplify a task and are used as electronic flashcards.

May artificially separate memorization from integration of knowledge rather than memory through integration.

Automaticity of drill responses is not optimal for facilitating language development.

Computer interaction may detract from social interaction, particularly in persons with specific social impairments, such as in Autism Spectrum Disorders.

Computer may prompt clinicians or educators to teach things that children do not need to learn.

Programs can make mistakes more interesting.

Programs must introduce enough variability to build generalization skills.

Clients must understand program instructions, particularly if using self-paced exercises.

Computers and access strategies can introduce independent cognitive load to task.

Input and output modalities may divide perceptual resources.

More difficult problems in everyday life are solved by analysis rather than drill.

Software may legitimately do what is advertised, but concept taught may not be valid or valuable.

Predetermined programs may limit originality of clinician interpretation of behavior.

Computer tasks may introduce methods or cognitive loads that exceed the purpose of the original therapeutic goals.

Should teach rewards of communication and interaction for its own sake, not for computer reinforcement alone.

Users may expect technology to serve as a stand-alone system for client–computer interaction, rather than embedding technology use into interaction and functional use.

Users may expect that intervention technology can replace language, speech, or hearing functions that are inherently irreplaceable.

May not be able to simplify task enough to overcome difficulties of interacting with computer or symbolic mode.

May perceive that computer-based techniques are more effective than they really are Computer program may teach skills according to philosophies or techniques that are contrary to the clinician's or the client's values.

Programs may be designed to rehabilitate skills that are best served by support that bypasses impaired functions.

Users may adapt educational or intervention goals to suit the capabilities of the technology, rather than using technology to support intervention.

Computers cannot be equivalent in function to clinicians, and cannot be used to justified reduced clinical staff.

ASSISTIVE DEVICES

Overview

Computer technology has dramatically improved the functional capabilities of physically, sensory, and communicatively impaired individuals. Functions addressed by assistive computer technology include communicative output (speech or written), communicative input (visual or auditory), environmental control, and access to standard computer capabilities (both personal and public computer devices). Many aspects of adapted computer access and universal design have already been incorporated into standard computers, software, and consumer devices. Key resources for further information about assistive devices and augmentative and alternative communication include the following: Beukelman and Mirenda (141), Cook and Hussey (142), the AAC Web site (http://aac.unl.edu), and the Co-Net assistive device inventory (http://trace.wisc.edu/tcel/). Topics addressed within this section will be organized according to communicative input, output, and other applications for language, hearing, and speech–physical impairments.

Language

Input (Language Learning). Persons who rely on augmentative communication (AAC) may face additional difficulties over typical communicators in both developing and using language skills through these alternative means. Early application of an augmentative system can help structure linguistic intervention by providing a motivating means for language practice, alternative modes of communicative output, and a vocabulary access and expansion system which can translate individual word units into acceptable grammatical form (143–145). Persons with physical or visual impairments often have limited access to standard written materials, and computers can provide access to reading and text production to promote receptive and expressive language skills (146,147). Still, young children who are nonspeaking have multiple additional sources of cognitive and interactive difficulty that influence early language development, such as difficulty of parents and children recognizing atypical movements as communicative or linguistic (148). Current research has addressed ways to adapt AAC language input techniques to be more cognitively accessible (149–151).

Output (Language Use). Conversational language output using AAC is slower than spoken language and often relies on users switching back and forth between spelled and prestored vocabulary to complete messages (152). Older AAC users may alter linguistic features of their output such as grammar and word choice because of the limitations of their available system; access to fully generative language systems is a critical element of language development in AAC (153). Fully competent language users who rely on AAC may still modify their language output because of rate and partner limitations (154). An important part of being linguistically competent in AAC is being able to determine strategies for when and how to use different language content and modalities for different purposes (155). An essential part of intervention with AAC devices, particularly for young children, involves strategies for partner training and support of the augmented communicator (156).

Most effective assistive technology applications for persons with acquired language and/or cognitive impairments focus on control and support of existing skills, rather than direct rehabilitation of impaired skills (157,158). This support includes low tech output such as gestures or communication boards, high tech devices for message formulation and retrieval, and partner-supported techniques such as written choice or partner drawing (158–160). Computerized linguistic cuing and speech output devices have been applied to clients with acquired language disabilities with only limited success (137,161). For most users, continued technology use after an acquired injury depends on the environmental and partner support as much or more than the type of technology used (162).

Other. With the current proliferation of web access and simulation programs, persons with language and other disabilities can also have access to language and functional learning experiences for which they are physically not suited, like laboratory problems or driving simulation. Computers and other distance technology may also be used to supplement language therapy for homebound, rural, or physically disabled clients who

cannot attend regular therapy sessions at a clinic (87,163). Most telecommunication systems require written language and computer access skills, which are problem areas for many language-impaired individuals with perceptual and/or symbolic deficits.

Hearing

Input. Digitized hearing aids have become the industry standard, with capacities to maximize high-frequency gain, filter out nonspeech distortion or noise, and program aids to mirror a patient's hearing loss (107,164,165). There have been significant advances in the use of computer technology for the direct evaluation and improvement of hearing aids and cochlear implants (53,54). New developments in hearing aid technology can allow users to alter the characteristics of the aid remotely to match various listening situations. Similarly, cochlear implants that provide electrocochlear stimulation have advanced signal-processing capabilities to improve speech intelligibility (33). Further research is needed to determine more accurate predictions of the types of benefits or limitations of cochlear implants expected for different populations (166).

Output. Other assistive devices for speech–language output for persons with hearing impairments provide dynamic translation of spoken information in face-to-face as well as remote interactions. For instance, interpreter services now offer real-time captioning, in which a transcriber records auditory information in a classroom or meeting to be immediately transmitted to the user's own display or a whole-room display. Standard text messaging and video transmission technologies allow quick access to direct interaction that can substitute for telephone access for deaf or hard of hearing users. If telephone services are used, TTY (teletypewriter) connections allow typed interactions between telecommunication devices using standard phone lines. If only one user has access to a TTY, telephone relay services are available to provide voice translations to and from TTY signals as needed. Similarly, fiberoptic Polycom video systems can be used either for direct signed conversations or remote ASL interpretation of spoken input when the interpreter is not physically present at the location. Text to speech and voice recognition systems are not yet sophisticated enough to translate spoken to written communication accurately in real time (167), but are used for limited purposes for persons with hearing impairments.

Other. Computer-based environmental control systems can be established in homes or workplaces to give a visual, automatic signal for telephones, doorbells, alarms, or emergency signals. See Cook (this volume, COMMUNICATION DEVICES) for more information on access to environmental control devices.

Speech and other Physical Impairments

Individuals with severely impaired control, coordination, or strength of oral-motor systems may require an assistive device to augment or replace functional speech. Types of disability or disease that may result in this condition include cerebral palsy, spinal cord injury, muscular dystrophy, multiple sclerosis, amyotrophic lateral sclerosis, stroke, traumatic brain injury, and spastic dysphonia. Individuals with motor impairments affecting their hands and arms may also require assistive devices to facilitate writing or access to a computer. Augmentative aids are designed to fit a client's motor, linguistic, cognitive, social, and educational skills and needs, and may address several different communication or access functions. Computer-based programs may be used to determine likely matches between technology and user needs, available either as stand-alone software or integrated software within dedicated communication devices (141,142). Many applications of technology for these users are covered under the chapter on computer access (Cook, this volume, COMMUNICATION DEVICES). This section will concentrate on technology to support speech input and output from persons who rely on AAC.

Input (Voice Recognition). Two types of uses are common for voice input systems with present technology. For persons who have good use of speech but severely limited physical movement, such as persons with high spinal cord injuries, voice input may be a primary strategy for controlling computers for a wide variety of writing, computing, and environmental control functions (167,168). Available voice input systems rely on voice recognition systems that are calibrated to the client's speaking style, and users typically must speak at a reduced rate and monitor the system for errors in recognizing words. Future development is necessary before voice recognition systems can be used seamlessly for writing or distance communication without error correction. For users with some speech with limited intelligibility, such as persons with cerebral palsy, voice recognition systems have been adapted to recognize and respond to atypical speech productions and provide support for writing and other communication tasks (169). Such systems need improvement in algorithms for recognizing subtle variations in vocalizations as well as triangulating intended utterances of users with high variability in speech production.

Output (Controlling Voice Output). Electronic voice output is an alternative or supplement to speech for many people with severe speech and/or physical impairments, who have difficulty being understood using their own voice. For individuals who cannot use standard keyboards or input devices, two general techniques are used to activate voice output devices: direct selection and scanning, both with options for encoding the input signal. The choice of input techniques depends on the range and type of behavior that the person can control, individual preferences, and communicative needs. Most of these input techniques could operate devices for speech, writing, and/or computer access functions. For direct selection, a variety of peripheral devices are available which utilize different motions to select communication units from an array of choices, including adapted keyboards, pointing devices, light pointers, switches, eyegaze, and speech input (141). For individuals who only have a small set

of discrete motions, encoding techniques can be used to provide a larger selection set. For example, persons who can operate only a limited number of keys or a joystick can select a larger number of words or letters by using sequences of two or three movements or selections for each word.

If an individual has only a single controllable movement, fatigues easily, or does not use encoding techniques, a scanning input method may be appropriate. A series of items or groups of items are presented sequentially, and the individual makes a single signal at the proper time to select the desired item. More sophisticated scanning techniques systematically narrow down the field of choice to speed the process. Since the user of a scanning aid must wait until a desired item is highlighted by the device, scanning is often a slow method (170). For users with additional cognitive and/or language limitations, scanning can introduce cognitive load that may require specific training or interfere with skilled technology access (171,172).

Since most alternative input techniques are slower than speech or typing, special acceleration programs have been developed that can provide the client with the ability to input the same text with fewer necessary input selections or keystrokes. Whole words and phrases can be coded, by the user and/or manufacturer, which can be accessed by abbreviated letter and number codes, picture symbol codes, or word prediction. For conversation needs, computer devices can provide visible correctable displays and voice synthesizers. For writing, correctable displays can provide printed output, with the potential for portable printing and writing aids as well as handwriting facsimile output. See Beukelman and Mirenda (90) for more detailed information on these techniques.

Other. Functions other than speaking or writing that computer devices can provide for individuals with severe disabilities include environmental control and access to standard computers. A variety of computer-based devices can adjust controls of appliances or lights, often linked through the same device used for speech and writing. Devices such as robotic arms can potentially handle light books, turn pages, insert computer disks, and provide other manipulative functions. Current efforts to improve universal design for easy access of environments to all users will improve the usefulness of technology-based solutions for environmental control and access (173). See Cook (this volume, COMMUNICATION DEVICES) for more information on these resources.

Another rapidly growing need for persons with disabilities is the ability to access and use standard computers and computer programs in the daily living environment. Individualized adaptive devices for accessing a personal communication aid will not necessarily allow direct access to any standard computer, particularly if the client must use an alternative input device. Interfacing units (including keyboard emulators) can allow the user to operate virtually any computer software program with their specially chosen adaptive devices by making the output of the personal adaptive device look like the information that the computer program expects to receive from the standard keyboard. However, the ability to use most software with one's own adaptive computer system at home does not necessarily mean that the same adaptive equipment can be used to operate other computers, bank teller machines, or library terminals. Limited compatibility between different dedicated and nondedicated hardware and software limits the range of technology a person can access with any given tool. See Cook (this volume, COMMUNICATION DEVICES) for more information on these resources.

Summary of Assistive Device Issues

Potential Benefits of Assistive Devices

Provides voice output to supplement or substitute for impaired speech skills.

Provides language input as well as output to children learning language.

Provides support for the development of language, writing, and cognitive skills.

Voice output allows more complex interaction between speaking and nonspeaking children who are preliterate.

Enables user-programmable vocabularies.

Communication aids can be made portable with speech, writing, or computer access capabilities.

Allows access and control of a wide variety of activities via computer.

Allows privacy of writing without need for translator.

Enables user to alternate between communicative and computer modes.

Potentially greater vocabulary storage and access techniques than nonautomated aids.

Provides mechanism to reach larger audience or range of interactants.

Augmentative communication over a distance (e.g., listservs) can compensate for unequal spoken to nonspoken speed in face-to-face interaction.

Distance technologies allow for complete writing and vocational tasks to be completed at the user's pace as a standard feature of that interaction mode.

Less expensive for some control and communication functions than personal aide.

Provides means for organizing thought visually.

Provides speech systems with intelligible and variable output.

Can expand simple mechanical operation such as a switch activation to perform complex functions.

Potential Limitations of Assistive Devices

Augmented communication is not as fast, flexible, or as varied as standard spoken or written communication.

Technology cannot replace original function and is always less easily adaptable.

Difficulty with compatibility of various devices, particularly between standard technology and dedicated software and hardware for communication.

Computer technology may present additional cognitive load for some users.

Short product life; computer devices are quickly out of date.

Other nonautomated techniques may fulfill same function at less cost with greater control over modifications.

Nonautomated techniques may be more interactive with the listener.

Time, money, effort expenditures may outweigh benefits.

People who have difficulties processing symbolic information will also have difficulties using a symbolic communication device.

Objective measures of relative efficiency and effectiveness of communication aids for individuals are currently limited.

Nonautomated aids may be more portable.

Sensory aids, particularly cochlear implants, can affect physiological function.

Problems with client, family, clinician acceptance, and/or understanding of device.

Difficulties with listener acceptance of AAC within communicative interaction.

Success with technology is the degree to which technology adapts to client needs instead of client adapting to technology capability.

FUTURE DIRECTIONS

Research, development, and clinical application of computer technology will continue in the field of communication disorders. Some of the directions of future development are suggested by progress in current research or clinical applications. New developments that are being addressed for computer applications in speech, hearing, and language disorders, and assistive devices are listed below.

New Directions in Speech

Improved databases of quantitative indices of typical and disordered speech performance.

Clinical implementation of motion analysis technology to assess and/or predict speech motor problems in children and adults with disabilities.

Improved commercial products for analyzing speech movements and muscle activity.

Modeling of the relationship between oral movements and speech sounds in early development.

Use of imaging technology such as magnetic resonance imaging (MRI) to assess and monitor vocal tract anatomy and physiology.

Clinical implementation of automatic analysis of vocal samples for variety and complexity of sounds produced.

Computerized intelligibility assessments appropriate for children as well as adults.

Inexpensive and clinically feasible acoustic analysis for assessment of voice and speech production.

Instrumental feedback to reinforce vocal productions of young children.

Instrumental feedback for treating problems with vocal loudness and unintelligible speech.

Instrumental feedback of "easy onset" productions by clients to improve fluency.

Better integration of speech and other behavioral or language intervention technologies.

Genetic testing of congenital speech disorders.

Electrical stimulation of neuromotor centers and pathways using direct current or magnetic induction.

New Directions in Hearing and Hearing Aids

Modeling of highly complex auditory processes, such as stochastic processing of speech (different possible outputs from the same input) at the auditory nerve level.

Automated computer assessment of routine aspects of hearing tests, and better data on user interface effects for different populations (e.g., elderly users).

Computer enhancements of neurosurgery for hearing, including robotics, microsurgery, and interoperative monitoring of surgery with technology.

Customized presentation and adaptation of therapy activities for balance, such as matching user head or eye movement to computer-based targets.

User-adaptable hearing aids.

Improved text to speech and voice recognition systems for real-time translation of spoken to written communication for persons with hearing impairments.

Increased standardization and improvement of video capabilities of telecommunication devices for persons with hearing impairments.

Enhancements of unintelligible auditory signals in hearing aids using more complex signal detection theory.

Implantable hearing aids that are user adjustable to improve fidelity of sound in selective environments.

Brainstem implants for persons who have eighth nerve damage to their hearing (e.g., neurofibromatosis).

Implantable prosthetic devices for improving vestibular function, including mixed sensory input such as tactile feedback for the position of the user's head in space.

New Directions in Language

Improved databases of language development and patterns of children and adults with various language disabilities.

More extensive use of imaging techniques such as fMRI to determine brain systems involved in different language and communication tasks.

Simultaneous analysis of multiple aspects of language behaviors in samples.

Dependable fully automated language sample analysis programs, using artificial intelligence algorithms to avoid common errors in current programs.

Computerized techniques for describing and predicting relationships between linguistic variables observed.

Computer-prompted sampling of language behaviors in targeted contexts and topics, and integration with cognitive responses to those language concepts.

Voice recognition of input to assist with some aspects of broad transcription of language samples.

Clinical consultation and language sampling through distance technology, including personal handheld devices and video/audio conference interactions.

Universal design of word processing software with literacy support, analogous to the integration of computer access technology supplied with standard computers.

Improvement of voice input technologies for text creation and editing by persons with language and/or literacy impairments.

Customizable language and grammar correction systems that can recognize particular nonstandard error patterns for clients with language impairments.

Improvements in scanning and voice output technology for seamless multimedia access and control of written information.

Simple and small personal reminders and navigation tools for persons with head injuries or other developmental impairments, that minimize the need for complex user interface.

New Directions in Assistive Devices

Improvements in synthesized voices for voice output devices.

Improvements in speed and organization of vocabulary access for persons relying on AAC.

Fully portable and durable voice output technology for young children.

AAC devices for young children and persons with cognitive impairments that are organized by conceptual or visual association rather than linguistic categories.

Digital photo and virtual scene interfaces for children and persons with cognitive impairments.

Better understanding of the language development processes in children who rely on AAC throughout their lifespan.

Better understanding of the grammatical and conversational modifications to provide maximum clarity, speed, and naturalness in persons using AAC.

Voice recognition for input by severely dysarthric speakers.

Strategies for storing voice samples from persons with degenerative diseases to use for later personalized voices in their AAC systems.

Integrated small units of voice output that can be incorporated singly into activities and natural interaction, as well as gathered into a single communication device.

Use of data transfer and distance access technology to support telework for persons with disabilities, including the potential for shared positions through telework.

Intelligent agents within AAC devices that can prompt communication and social tasks, such as for persons with autism.

ACKNOWLEDGMENTS

Part of the data collection for this article and the previous edition was the interviewing of persons involved with the application of computer technology in communication disorders. These persons were asked to comment on functions within their areas of specialty that are addressed by computer technology. We greatly appreciate the cooperation and expertise of the following persons (in alphabetical order by chapter edition): *Second Edition:* David Beukelman, T. Newell Decker, Malinda Eccarius, Charles Healey, Karen Hux, and Neil Shepard. *First Edition:* Claudia Blair, Diane Bless, Robin Chapman, Michael Collins, Stanley Ewanowski, Robert Goldstein, Linda Hesketh, Carol Hustedde, Raymond Karlovich, Jesse Kennedy, Raymond Kent, Marilyn Kertoy, Mikael Kimelman, Pamela Mathy-Laikko, Vicky Lord-Larson, Richard Lehrer, Malcolm McNeil, Jon Miller, Linda Milosky, Lois Nelson, Katharine Odell, Mary Jo Osberger, Barry Prizant, Ann Ratcliff, John Peterson, Margaret Rosin, Lawrence Shriberg, Dolores Vetter, Francisco Villarruel, Gary Weismer, and Terry Wiley.

BIBLIOGRAPHY

1. Coleman JG. The Early Intervention Dictionary. Bethesda (MD): Woodbine House; 1993.
2. Nicolosi L, Harryman E, Kresheck J. Terminology of Communication Disorders: Speech-Language-Hearing. 4th ed. Philadelphia: Lippincott, Williams & Wilkins; 1996.
3. Gillam RB, Marquardt TP, Martin FN. Communication Sciences and Disorders: From Science to Clinical Practice. San Diego: Singular; 2000. p 85–98.
4. Herer GR, Knightly CA, Steinberg AG. Hearing: Sounds and silences. In: Batshaw ML, editor. Children with Disabilities. 5th ed. Baltimore: Paul H. Brookes; 2002.
5. Cochran PS. Clinical Computing Competency for Speech-Language Pathologists. Baltimore: Paul H. Brookes; 2004.
6. Cochran PS, Bull GL. Integrating word processing into language intervention. Top Lang Dis 1991;11(2):31–49.
7. Smith SW, Kortering LJ. Using computers to generate IEPs: Rethinking the process. J Special Educ Technol 1996;13(2): 81–90.
8. Lieberth AK, Martin DR. Authoring and hypermedia. Language Speech Hearing Ser Schools 1995;26(3):241–250.
9. Epstein JN, Willis MG, Conners CK, Johnson DE. Use of a technological prompting device to aid a student with attention deficit hyperactivity disorder to initiate and complete daily tasks: An exploratory study. J Special Educ Technol 2001;16(1): 19–28.
10. Foegen A, Hargrave CP. Group response technology in lecture-based instruction: Exploring student engagement and instructor perceptions. J Special Educ Technol 1999; 14(1):3–17.
11. Gardner JE, Wissick CA, Schweder W, Canter LS. Enhancing interdisciplinary instruction in general and special education: Thematic units and technology. Remedial Special Educ 2003;24(3):161–172.

12. Glaser CW, Rieth HJ, Kinzer CK, Colburn LK, Peter J. A description of the impact of multimedia anchored instruction on classroom interactions. J Special Educ Technol 1999;14(2): 27–43.
13. Roberson L. Integration of computers and related technologies into deaf education teacher preparation programs. Am Ann Deaf 2001;146(1):60–66.
14. Hasselbring TS. A possible future of special education technology. J Special Educ Technol 2001;16(4):15–21.
15. Masterson JJ. Future directions in computer use. Lang Speech Hearing Ser Schools 1995;26(3):260–262.
16. Sokolov JL, Snow CE. Transcript analysis using the Child Language Data Exchange System. In: Sokolov JL, Snow CE, editors. Handbook of Research in Language Development using CHILDES. Hillsdale (NJ): Lawrence Erlbaum Associates; 1994.
17. McGuire RA. Computer-based instrumentation: Issues in clinical applications. Lang Speech Hearing Ser Schools 1995;26(3):223–231.
18. Barlow SM, Cole KJ, Abbs JH. A new head-mounted lip-jaw movement transduction system for the study of motor speech disorders. J Speech Hearing Res 1983;26:283–288.
19. Müller EM, Abbs JH. Strain gauge transduction of lip and jaw motion in the midsagittal plane: Refinement of a prototype system. J Acoust Soc Am 1979;65:481–486.
20. Green JR, Moore CA, Higashikawa M, Steeve RW. The physiologic development of speech motor control: Lip and jaw coordination. J Speech Language Hearing Res 2000; 43:239–255.
21. American Speech-Language-Hearing Association, Guidelines for speech-language pathologists performing videofluoroscopic swallowing studies. ASHA Suppl 2004;24:
22. Logemann JA. Manual for the Videofluoroscopic Study of Swallowing. 2nd ed. Austin (TX): ProEd; 1993.
23. Hirose H, Kiritani S, Ushijima T, Yoshioka H, Sawashima M. Patterns of dysarthric mocments in patients with Parkinsonism. Folia Phoniatrica 1981;33:204–215.
24. Perkell J, Cohen M, Svirsky M, Matthies M, Garabieta I, Jackson M. Electro-magnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements. J Acoust Soc Am 1992;92:3078–3096.
25. Boersma P, Weenink WD. Pratt, (Version 4.2) [Computer software]. Amsterdam: University of Amsterdam; 2004.
26. Kay Elemetrics Corp, 2 Bridgewater Lane, Lincoln Park, NJ, 07035. 2004. Tel: (973) 628-6200.
27. Milenkovic P. CSpeech [Computer software]. 118 Shiloh Drive, Madison, WI, 53705, 2004. Tel: (608) 833-7956. Available at http://userpages.net.chorus/cspeech.
28. Barlow SM, Suing G. Aerospeech: Automated digital signal analysis of speech aerodynamics. J Comput Users Speech Hearing 1991;7:211–227.
29. Barlow SM, Suing G, Andreatta RD. Speech aerodynamics using AEROWIN. In: Barlow SM, editor. Handbook of Clinical Speech Physiology. San Diego: Singular; 1999. p 165–189.
30. The Mathworks, 3 Apple Hill Drive, Natick, MA, 01760. 2004. Tel: 508-647-7000. Available at www.mathworks.com.
31. National Instruments, 11500 N. Mopac Expressway, Austin, TX, 78759. Tel: (888) 280-7645. Available at www.ni.com.
32. Chan JS, Spence C. Presenting multiple auditory signals using multiple sound cards in Visual Basic 6.0. Behav Res Methods Instrum Comput 2003;35(1):125–128.
33. Henry JA, Flick CL, Gilbert A, Ellingson RM, Fausti SA. Reliability of computer-automated hearing thresholds in cochlear-impaired listeners using ER-4B Canal Phone earphones. J Rehabil Res Dev 2003;40(3):253–264.
34. Walsh T, Demkowicz L, Charles R. Boundary element modeling of the external human auditory system. J Acoust Soc Am 2004;115(3):1033–1043.
35. Yao J, Zhang YT. The application of bionic wavelet transform to speech signal processing in cochlear implants using neural network simulations. IEEE Trans Biomed Eng 2002;49(11): 1299–1309.
36. Schneider U, Schleussner E, Haueisen J, Nowak H, Seewald HJ. Signal analysis of auditory evoked cortical fields in fetal magnetoencephalography. Brain Topog 2001;14(1):69–80.
37. Evans JL, Miller J. Language sample analysis in the 21st century. Sem Speech Lang 1999;20(2):101–116.
38. Miller J, Chapman R. 2000. SALT: Systematic Analysis of Language Transcripts (SALT). Version 6.1 [Computer software]. Madison (WI): Language Analysis Laboratory. Waisman Center. University of Wisconsin. Available at http://waisman.wisc.edu/salt/.
39. Miller J, Freiberg C, Rolland MB, Reeves M. Implementing computerized language sample analysis in the public school. Top Lang Dis 1992;12(2):69–82.
40. MacWhinney B. The CHILDES Project: Tools for Analyzing Talk. 3rd ed. Mahwah (NJ): Lawrence Erlbaum; 2000a.
41. Hallowell B, Katz R. Technological applications in the assessment of acquired neurogenic communication and swallowing disorders in adults. Sem Speech Lang 1999; 20(2):149–168.
42. Violin RA. Microcomputer-based systems providing biofeedback of voice and speech production. Top Lang Dis 1991; 11(2):65–79.
43. Green JR, Beukelman DR, Ball LJ. Algorithmic estimation of pauses in extended speech samples. J Med Speech Hearing Res 2004;12:149–154.
44. Hosom JP, Shriberg L, Green JR. Diagnostic Assessment of Childhood Apraxia of speech using automatic speech recognition (ASR) methods. J Med Speech Hearing Res 2004; 12.
45. Titze IT. Summary Statement for the Workshop on Acoustic Voice Analysis. National Center for Voice and Speech: Iowa City; 1995.
46. American Speech-Language-Hearing Association, Vocal tract visualization and imaging. ASHA 1992;34(March, 7 Suppl): 37–40.
47. Case JL. Technology in the assessment of voice disorders. Sem Speech Lang 1999a;20(2):169–184.
48. Bakker K. Technical solutions for quantitative and qualitative assessments of speech fluency. Sem Speech Lang 1999a;20(2): 185–196.
49. Yorkston K, Beukelman D, Tice R. Sentence Intelligibility Test (Version 1.0). [computer software]. Lincoln (NE): Communication Disorders Software, 1996.
50. Masterson JJ, Oller KD. Use of technology in phonological assessment: Evaluation of early meaningful speech and prelinguistic vocalizations. Sem Speech Lang 1999; 29(2):133–148.
51. Fell JH, MacAuslan J, Ferrier LJ, Chenausky K. Automatic babble recognition for early detection of speech related disorders. Behav Inf Technol 1999;18(1):56–63.
52. Mendel LL, Wynne MK, English K, Schmidt-Troike A. Computer applications in educational audiology. Lang Speech Hearing Schools 1995;26(3):232–240.
53. Johnson CE, Danhauer JL, Krishnamurti S. A holistic model for matching high-tech hearing aid features to elderly patients. Am J Audiol 2000;9(2):112–123.
54. Newman CW. Digital signal processing hearing aids: Determining need on an individual basis. Arch Otolaryngol Head Neck Surg 2000;126(11):1397–1398.
55. Bradley S. A spreadsheet for calculating the articulation index and displaying the unaided and aided speech spectrum. J Comput Speech Hearing 1991;7:357–361.
56. Leavitt R, Flexer C. Speech degradation as measured by the rapid speech transmission index (RASTI). Ear Hearing 1991; 12:115–117.

57. Iley KL, Addis RJ. Impact of technology choice on service provision for universal newborn hearing screening within a busy district hospital. J Perinatol 2000;20:S122–S127.
58. Johnson MJ, Maxon AB, White KR, Vohr BR. Operating a hospital-based universal newborn hearing screening program using transient evoked otoacoustic emissions. Sem Hearing 1993;14:46–55.
59. Stone KA, Smith BD, Lembke JM, Clark LA, McLellan MB. Universal newborn hearing screening. J Family Practice 2000;49:1012–1016.
60. Allen B, Lambert G. Computerization of V.R.O.A: A double blind hearing screening technique. Aust J Audiol 1990;1 2:11–15.
61. Eilers RE, Widen JE, Urbano R, Hudson TM, Gonzales L. Optimization of automated hearing test algorithms: A comparison of data from simulations and young children. Ear Hearing 1991;12:199–204.
62. Schmida MJ, Peterson JH, Tharpe AM. Visual reinforcement audiometry using digital video disc and conventional reinforcers. Am J Audiol 2003;12(1):35–40.
63. McCullough JA, Cunningham LA, Wilson RH. Auditory-visual word identification test materials: Computer application with children. J Am Acad Audiol 1992;3:208–214.
64. Long SH. About time: A comparison of computerized and manual procedures for grammatical and phonological analysis. Clin Linguistics Phonetics 2001;15(5):399–426.
65. Miller J, Chapman R. The relation between age and mean length of utterance in morphemes. J Speech Hearing Res 1981;24:154–161.
66. Long SH. Technology applications in the assessment of children's language. Sem Speech Lang 1999;20(2):117–132.
67. Long SH, Channell RW. Accuracy of four language analysis procedures performed automatically. Am J Speech-Langauge Pathol 2001;10:180–188.
68. MacWhinney B. Child Language Analysis (CLAN) Manual. Available at http://childes.psy.cmu.edu/pdf/clan.zip. Accessed 2000b.
69. Long SH, Fey ME, Channell RW. Computerized Profiling (CP). Version 9.2.7. [Computer program]. Cleveland, OH: Department of Communication Sciences, Case Western Reserve University. Also available at http://www.cwru.edu/artsci/cosi/cp.htm. Accessed 2000.
70. Shriberg L. Program to Examine Phonetic and Phonological Evaluation Records (PEPPER). Version 4.0. Hillsdale (N.J.): Lawrence Erlbaum Associates; 1986.
71. Masterson J, Bernhardt B. Computerized articulation and phonology evaluation system (CAPES) [computer program]. San Antonio: The Psychological Corporation, 2001.
72. Masterson JJ, Long SH, Buder EH. Instrumentation in clinical phonology. In: Bernthal JE, Bankson NW, editors. Articulation and Phonological Disorders. 4th ed. Boston: Allyn & Bacon; 1998.
73. Haaf R, Duncan B, Skarakis-Doyle E, Carew M, Kapitan P. Computer-based language assessment software: The effects of presentation and response format. Lang Speech Hearing Ser Schools 1999;30(1):68–74.
74. Shriberg LD, Kwiatkowski J, Snyder T. Articulation testing by microcomputer. J Speech Hearing Dis 1986;51(4):309–324.
75. Wiig EH, Jones SS, Wiig ED. Computer-based assessment of word knowledge in teens with learning disabilities. Lang Speech Hearing Ser Schools 27:21–28.
76. Cochran PS, Masterson JJ. NOT using a computer in language assessment/intervention: In defense of the reluctant clinician. Lang Speech Hearing Ser Schools 1995;26(3):213–222.
77. Jacobs EL. The effects of adding dynamic assessment components to a computerized preschool language screening test. Commun Dis Quart 2001;22(4):217–226.
78. Jacobs EL, Coufal KL. A computerized screening instrument of language learnability. Commun Dis Quart 2001;22(2):67–75.
79. van Geert P, van Dijk M. Focus on variability: New tools to study intra-individual variability in developmental data. Infant Behav Dev 2002;25:340–374.
80. Clancy B, Finlay B. Neural correlates of early language learning. In: Tomasello M, Bates E, editors. Language Development: The Essential Readings. Malden (MA): Blackwell; 2001.
81. Neville H, Mehler J, Newport E, Werker J, McClelland J. Special issue: The developing brain. Section 4: Language. Dev Sci 2001;4(3):293–312.
82. Bhatnagar SC, Andy OJ. Diagnostic techniques and neurological concepts. In: Bhatnagar SC, Andy OJ , editors. Neuroscience for the Study of Communicative Disorders. Baltimore: Williams & Wilkins; 1995. p 314–332.
83. Goldenberg EP. Computers in the special education classroom: What do we need, and why don't we have any?" In: Mulick J, Mallory B, editors. Transitions in Mental Retardation. Vol. 1. Norwood (NJ): Ablex Press; 1984.
84. Behrmann M. Handbook of Microcomputers in Special Education. San Diego: College Hill Press; 1984.
85. Rushakoff G. Clinical applications in communication disorders. In: Schwartz A, editor. Handbook of Microcomputer Applications in Communication Disorders. San Diego: College Hill Press; 1984.
86. Mahaffey R. An overview of computer applications. Top Lang Dis 1985;6:1–10.
87. Katz RC, Hallowell B. Technological applications in the treatment of acquired neurogenic communication and swallowing disorders in adults. Sem Speech Lang 1999; 20(3): 251–270.
88. Ruscello DM. Visual feedback in treatment of residual phonological disorders. J Commun Dis 1995;28:279–302.
89. Case JL. Technology in the treatment of voice disorders. Sem Speech Lang 1999b;20(3):281–295.
90. Masterson JJ, Rvachew S. Use of technology in phonological intervention. Sem Speech Lang 1999;20(3):233–250.
91. Crary MA, Groher ME. Basic concepts of surface electromyographic biofeedback in the treatment of dysphagia: A tutorial. Am J Speech-Language Pathol 2000;9(2):116–125.
92. Logemann JA, Kahrilas PJ. Relearning to swallow after stroke—application of non-invasive biofeedback. A case study. Neurology 1990;40:1136–1140.
93. Bakker K. Clinical technologies for the reduction of stuttering and enhancement of speech fluency. Sem Speech Lang 1999b;20(3):271–280.
94. Stuart A, Xia S, Jiang Y, Jiang T, Kalinowski J, Rastatter MP. Self-contained in-the-ear device to deliver altered auditory feedback: Applications for stuttering. Ann Biomed Eng 2003;31(2):233–237.
95. Fell H, Cress C, MacAuslan J, Ferrier L. VisiBabble for reinforcement of early vocalization. Presentation at the ASSETS '04 Conference, Atlanta, GA, 2004, October.
96. Hirano M, Kurita S, Sakaguchi S. Ageing of the vibratory tissue of human vocal folds. Acta Oto-Laryngolog 1989; 107:428–433.
97. Kasuya H, Ogawa S, Kikuchi Y. An acoustic analysis of pathological voice and its application to the evaluation of laryngeal pathology. Speech Commun 1986;5:171–181.
98. Fletcher SG, Hasegawa A. Speech modification by a deaf child through dynamic orometric modeling and feedback. J Speech Hearing Dis 1983;48:178–185.
99. Fletcher SG. Visual articulation training through dynamic orometry. Volta Rev 1989;91:47–64.
100. Ambulatory Monitoring, Inc., 731 Saw Mill River Road, Ardsley, N.Y. 10502. Tel: (800) 341-0066. Available at www.ambulatory-monitoring.com/index.html. Accessed 2004.

101. Hardcastle WJ, Gibbon FE, Jones W. Visual display of tongue-palate contact: Electropalatography in the assessment and remediation of speech disorder. Br J Dis Commun 1991;26:41–74.
102. Blischak DM. Increases in natural speech production following experience with synthetic speech. J Special Educ Technol 1999;15(2):44–53.
103. Hustad KC, Shapley KL. AAC and natural speech in individuals with developmental disabilities. In: Light JC, Beukelman DR, Reichle J, editors. Communicative Competence for Individuals who use AAC: From Research to Effective Practice. Baltimore: Brookes; 2003. p 41–62.
104. Blood GW. A behavioral-cognitive therapy program for adults who stutter: Computers and counseling. J Commun Dis 1995;28:165–180.
105. Pratt SR, Heintzelman AT, Deming SE. The efficacy of using the IBM speech viewer vowel accuracy module to treat young children with hearing impairment. J Speech Hearing Res 1993;36(5):1063–1074.
106. Tye-Murray N. Laser videodisc technology in the aural rehabilitation setting: Good news for people with severe and profound hearing impairments. Am J Audiol 1992; 1(2):33–36.
107. Uziel A, Mondain M, Hagen P, Dejean F, Doucet G. Rehabilitation for high-frequency sensorineural hearing impairment in adults with the symphonix vibrant soundbridge: A comparative study. Otol Neurotol 2003;24(5):775–783.
108. Shepard NT, Solomon D, Ruckenstein M, Staab J. Evaluation of the vestibular (balance) system. In: Ballenger JJ, Snow JB, editors. Otorhinolaryngology Head and Neck Surgery. 16th ed. San Diego: Singular; 2003.
109. Steiner S, Larson VL. Integrating microcomputers into language intervention with children. Top Lang Dis 1991; 11(2): 18–30.
110. Nippold MA, Schwarz LE, Lewis M. Analyzing the potential benefit of microcomputer use for teaching figurative language. Am J Speech Lang Pathol 1992;1:36–43.
111. Rose MO, Cochran PS. Teaching action verbs with computer-controlled videodisc vs. traditional picture stimuli. J Comput Users Speech Hearing 1992;8:15–32.
112. Diehl SF. Listen and Learn? A software review of Earobics. Lang Speech Hearing Ser Schools 1999;30(1):108–116.
113. Friel-Patti S, DesBarres K, Thibodeau L. Case studies of children using Fast Forword. Am J Speech-Language Pathol 2001;10:203–215.
114. Gillam RB, Crofford JA, Gale MA, Hoffman LM. Language change following computer-assisted language instruction with Fast Forword or Laureate Learning Systems software. Am J Speech-Language Pathol 2001;10:231–247.
115. Judge SL. Computer applications in programs for young children with disabilities: Current status and future directions. J Special Educ Technol 2001;16(1):29–40.
116. Kinsley TC, Langone J. Applications of technology for infants, toddlers, and preschoolers with disabilities. J Special Educ Technol 1995;12(4):312–324.
117. Cochran PS, Nelson LK. Technological applications in intervention for preschool-age children with language disorders. Sem Speech Lang 1999;20(3):203–218.
118. Howard J, Greyrose E, Kehr K, Espinosa M, Beckwith L. Teacher-facilitated microcomputer activities: Enhancing social play and affect in young children with disabilities. J Special Educ Technol 1996;13(1):36–47.
119. Cress CJ, Marvin CA. Common questions about AAC services in early intervention. Augment Alter Commun 2003;19 (4):254–272.
120. Howell RD, Erickson K, Stanger C, Wheaton JE. Evaluation of a computer-based program on the reading performance of first grade students with potential for reading failure. J Special Educ Technol 2000;15(4):5–14.
121. Wood LA, Masterson JJ. The use of technology to facilitate language skills in school-age children. Sem Speech Lang 1999;20(3):219–232.
122. Daiute C, Morse F. Access to knowledge and expression: Multimedia writing tools for students with diverse needs and strengths. J Special Educ Technol 1994;12(3):221–256.
123. Montgomery DJ, Karlan GR, Coutinho M. The effectiveness of word processor spell checker programs to produce target words for misspellings generated by students with learning disabilities. J Special Educ Technol 2001;16(2):27–40.
124. Sturm JM, Rankin JL, Beukelman DR, Schutz-Meuhling L. How to select appropriate software for computer-assisted writing. Intervention School Clinic 1997;32:148–162.
125. Wise BW. Computer speech and the remediation of reading and spelling problems. J Special Educ Technol 1994;12(3): 207–220.
126. Wood LA, Rankin JL, Beukelman DR. Word prompt programs: Current uses and future possibilities. Am J Speech-Language Pathol 1997;6(3):57–65.
127. Higgins EL, Raskind MH. Speaking to read: The effects of continuous vs. discrete speech recognition systems on the reading and spelling of children with learning disabilities. J Special Educ Technol 2000;15(1):19–30.
128. Edwards BJ, Blackhurst AE, Koorland MA. Computer-assisted constant time delay prompting to teach abbreviation spelling to adolescents with mild learning disabilities. J Spec Educ Technol 1995;12(4):301–311.
129. Hutinger P, Johanson J, Stoneburner R. Assistive technology applications in educational programs of children with multiple disabilities: A case study report on the state of practice. J Special Educ Technol 1996;13(1):16–35.
130. Morgan RL, Gerity BP, Ellerd DA. Using video and CD-ROM technology in a job preference inventory for youth with severe disabilities. J Special Educ Technol 2000;15(3):25–33.
131. Nelson KE, Heimann M, Tjus T. Theoretical and applied insights from multimedia facilitation of communication skills in children with autism, deaf children, and children with other disabilities. In: Adamson LB, Romski MA, editors. Communication and Language Acquisition: Discoveries from Atypical Development. Baltimore: Brookes; 1997. p 295–325.
132. Norman JM, Collins BC, Schuster JW. Using an instructional package including video technology to teach self-help skills to elementary students with mental disabilities. J Special Educ Technol 2001;16(3):5–18.
133. Rostron A, Sewell D. Microtechnology in Special Education. Baltimore: Johns Hopkins University Press; 1984.
134. Light JD, Roberts B, Dimarco R, Greiner N. Augmentative and alternative communication to support receptive and expressive communication for people with autism. J Commun Dis 1998;31:158–180.
135. Mirenda P, Wilk D, Carson P. A retrospective analysis of technology use patterns of students with autism over a five-year period. J Special Educ Technol 2000;15(3):5–16.
136. Katz RC. Computer applications in aphasia treatment. In: Chapey R, editor. Language Intervention Strategies in Aphasia and Related Neurogenic Communication Disorders. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2001. p 718–741.
137. Katz RC, Wertz RT. The efficacy of computer-provided reading treatment for chronic aphasic adults. J Speech Lang Hearing Res 1997;40(3):493–507.
138. Herrmann D, Yoder CY, Wells J, Raybeck D. Portable electronic scheduling/reminding devices. Cog Technol 1996;1: 19–24.

139. Kaasgaard K, Lauritsen P. The use of computers in cognitive rehabilitation in Denmark. Am J Speech-Language Pathol 1995;4:5–8.
140. Robinson I. Does computerized cognitive rehabilitation work? A review. Aphasiology 1990;4:381–405.
141. Beukelman DR, Mirenda P. Augmentative and Alternative Communication: Management of Severe Communication Disorders in Children and Adults. Baltimore: Brookes; 1998.
142. Cook AM, Hussey SM. Assistive Technologies: Principles and Practice. St. Louis: Mosby; 1995.
143. Goossens C, Kraat A. Technology as a tool for conversation and language learning for the physically disabled. Top Lang Dis 1985;6:56–70.
144. Reichle J, Hidecker MJC, Brady NC, Terry N. Intervention strategies for communication: Using aided augmentative communication systems. In: Light JC, Beukelman DR, Reichle J, editors. Communicative Competence for Individuals who use AAC: From Research to Effective Practice. Baltimore: Brookes; 2003. p 441–477.
145. Romski MA, Sevcik RA, Hyatt AM, Cheslock M. A continuum of AAC language intervention strategies for beginning communicators. In: Reichle J, Beukelman DR, Light JC, editors. Exemplary Practices for Beginning Communicators: Implications for AAC. Baltimore: Brookes; 2002. p 1–24.
146. Justice LM, Chow SM, Capellini C, Flanigan K, Colton S. Emergent literacy intervention for vulnerable preschoolers: Relative effects of two approaches. Am J Speech-Lang Pathol 2003;12:320–332.
147. Sandberg AD. Reading and spelling: Phonological awareness, and working memory in children with severe speech impairments: A longitudinal study. Augment Alter Commun 2001;17:11–26.
148. Cress CJ. Expanding children's early augmented behaviors to support symbolic development. In: Reichle J, Beukelman DR, Light JC, editors. Exemplary Practices for Beginning Communicators: Implications for AAC. Baltimore: Brookes; 2002. p 272–291.
149. Fallon KA, Light J, Achenbach A. The semantic organization patterns of young children: Implications for augmentative and alternative communication. Augment Alter Commun 2003;19(2):74–85.
150. Light JD, Drager KDR, Nemser JG. Enhancing the appeal of AAC technologies for young children: Lessons from the toy manufacturers. Augmentative and Alter Commun 2004;20 (3):137–149.
151. Wilkinson KM, Jagaroo V. Contributions of principles of visual cognitive science to AAC system display design. Augment Alter Commun 2004;20(3):123–136.
152. File P, Todman J. Evaluation of the coherence of computer-aided conversations. Augment Alter Commun 2002;18: 228–241.
153. Blockberger S, Sutton A. Toward linguistic competence: Language experiences and knowledge of children with extremely limited speech. In: Light JC, Beukelman DR, Reichle J, editors. Communicative Competence for Individuals who use AAC: From Research to Effective Practice. Baltimore: Brookes; 2003. p 63–106.
154. Smith MM, Grove NC. Asymmetry in input and output for individuals who use AAC. In: Light JC, Beukelman DR, Reichle J, editors. Communicative competence for individuals who use AAC: From research to effective practice. Baltimore: Brookes; 2003. p 163–195.
155. Mirenda P, Bopp KD. Playing the game: Strategic competence in AAC. In: Light JC, Beukelman DR, Reichle J, editors. Communicative Competence for Individuals who use AAC: From Research to Effective Practice. Baltimore: Brookes; 1993. p 401–437.
156. Cress CJ. AAC and language: Understanding and responding to parent perspectives. Top Lang Dis 2004;24(1): 28–38.
157. Garrett KL, Kimelman MDZ. AAC and aphasia: Cognitive-linguistic considerations. In: Beukelman DR, Yorkston KM, Reichle J, editors. Augmentative and Alternative Communication for Adults with Acquired Neurologic Disorders. Baltimore: Brookes; 2000. p 339–374.
158. Hux K, Manasse N, Weiss A, Beukelman DR. Augmentative and alternative communication for persons with aphasia. In: Chapey R, editor. Language Intervention Strategies in Aphasia and Related Neurogenic Communication Disorders. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2001. p 675–687.
159. Garrett KL, Beukelman DR. Augmentative communication approaches for persons with severe aphasia. In: Yorkston K, editor. Augmentative Communication in the Medical Setting. Tucson: Communication Skill Builders; 1992. p 245–321.
160. Lasker J, Hux K, Garrett K, Moncrief E, Eischeid T. Variations on the written choice communication strategy for individuals with severe aphasia. Augment Alter Commun 1997;13:108–116.
161. Colby K, Christinaz D, Parkinson S, Graham S, Karpf C. A word-finding computer program with a dynamic lexical-semantic memory for patients with anomia using an intelligent speech prosthesis. Brain Lang 1981;14:272–281.
162. Lasker JP, Bedrosian JL. Acceptance of AAC by adults with acquired disorders. In: Beukelman DR, Yorkston KM, Reichle J, editors. Augmentative and Alternative Communication for Adults with Acquired Neurologic Disorders. Baltimore: Brookes; 2000. p 107–136.
163. Vaughn GR, Kramer JO, Ozley CF. Tel-communicology for clinician and computer outreach. Commun Dis 1983;8: 75–88.
164. Taylor RS, Paisley S, Davis A. Systematic review of the clinical and cost effectiveness of digital hearing aids. Br J Audiol 2001;35(5):271–288.
165. Yueh B. Digital hearing aids. Arch Otolaryn Head Neck Surg 2000;126(11):1394–1397.
166. Wilson BS, Lawson DT, Muller JM, Tyler RS, Kiefer J. Cochlear implants: Some likely next steps. Ann Rev Biomed Eng 2003;5:207–49.
167. Koester HH. User performance with speech recognition: A literature review. Assis Technol 2001;13(2):116–130.
168. Noyes J, Frankish C. Speech recognition technology for individuals with disabilities. Augment Alter Commun 1992;8: 297–303.
169. Ferrier L, Shane H, Ballard H, Carpenter T, Benoit A. Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. Augment Alter Commun 1995;11:165–174.
170. Ratcliff A. Comparison of relative demands implicated in direct selection and scanning: Considerations from normal children. Aug Alter Commun 1994;10:67–74.
171. (a) Light JC. Teaching automatic linear scanning for computer access: A case study of a preschooler with severe physical and communication disabilities. J Special Educ Technol 1993;12(2):125–134.
172. Wehmeyer ML. Assistive technology and students with mental retardation: Utilization and barriers. J Special Educ Technol 1999;14(1):48–58.
173. Rose D. Walking the walk: Universal design on the web. J Special Educ Technol 2000;15(3):45–49.

See also COMMUNICATION DEVICES; ENVIRONMENTAL CONTROL; REHABILITATION, COMPUTERS IN COGNITIVE.

COMPOSITES, RESIN-BASED. See RESIN-BASED COMPOSITES.

COMPUTED RADIOGRAPHY. See DIGITAL RADIOGRAPHY.

COMPUTED TOMOGRAPHY

MICHAEL J. DENNIS
Medical University of Ohio
Toledo, Ohio

INTRODUCTION

"Computed tomography . . . measures the attenuation of x-ray beams passing through sections of the body from hundreds of different angles, and then, from the evidence of these measurements, a computer is able to reconstruct pictures of the body's interior." That is the basic description of Computed Tomography (CT) as given by Sir Godfrey Hounsfield in his 1979 Nobel Lecture (1).

Computed tomography was a breakthrough in the implementation and acceptance of digital computers into clinical diagnostic imaging. It's commercial birth in the early 1970s was an amalgamation of X-ray imaging, detector development, mathematical methods, along with the developing computer capabilities of the time to produce a whole new way of peering into the human body.

The attenuation properties of X and γ rays are well known. The logarithmic scaled values of transmission measurements through a body yields a line integral or summation of the attenuation properties along the path of the beam. The linear attenuation coefficient, or the probability of interaction per microscopic distance traveled, is directly related to the density of the material and the effective atomic number of the material. A computer utilizes a mathematical algorithm to determine what the distribution of attenuation coefficients within the body must be to produce the measured set of transmission values. By unfolding the data in this way, tissues of interest within the patient are not obscured by anatomy above and below it. Consequently, structures may be accurately localized within the body and small, previously invisible differences in density or attenuation ($< 1\%$) were seen.

Generally, the reconstructed data is calculated and presented as a series of cross-sectional slices. Each slice in the computer is represented by a two-dimensional (2D) matrix of numbers. The numbers in this array are scaled values of the linear attenuation coefficient, referred to as CT numbers or Hounsfield units. The individual data elements in a CT image are referred to as pixels or picture elements. The measurements through the body, however, are not along infinitely thin planes. The X-ray beam and resultant measurements have a finite width or thickness. A 2D picture element corresponds to a box shaped volume within the patient, referred to as a volume element or voxel (Fig. 1).

Advances in diagnostic medical imaging over the past half century have been phenomenal, in particular with the development and implementation of computed tomography and magnetic resonance imaging. The ability to virtually slice and dice a living human body to see the internal anatomical structures has eliminated the previously common practice of exploratory surgery, and has enabled more accurate diagnosis and improved effectiveness of medical treatment.

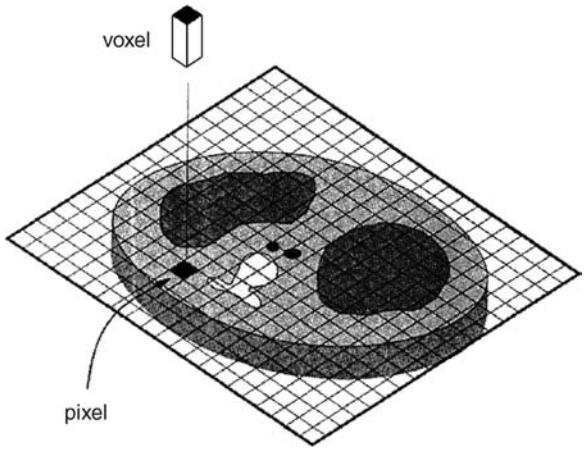


Figure 1. An axial CT image is composed of a 2D array of CT numbers with each picture element or pixel corresponding to a volume element or voxel within the patient.

BASIC PRINCIPLES OF COMPUTED TOMOGRAPHY

The basic technique of computed tomography as illustrated in Fig. 2 is to probe a thin slice of the patient with a thin beam of radiation, which is attenuated as it passes through the patient. The fraction of the X-ray beam that is attenuated is directly related to the density, thickness, and composition of the material through which the beam has traveled and to the energy of the X-ray beam. Computed

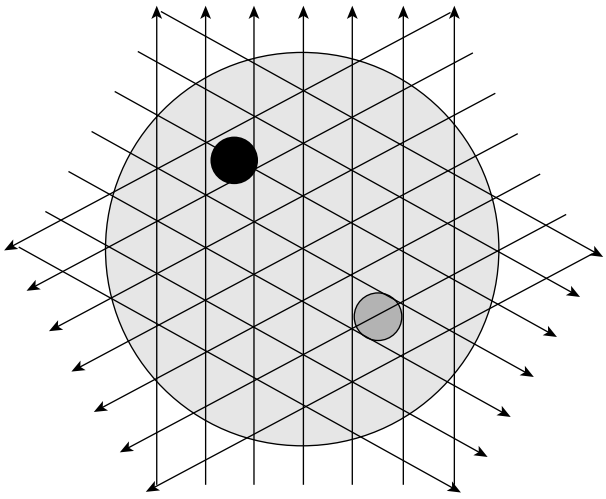


Figure 2. Basic principle of CT is that X-ray transmission measurements are taken along many rays through a thin slice of the patient from many different angles. The measured values are then used to map the distribution of attenuating material that produced the measured transmission values.

tomography utilizes this information from many different angles, to determine cross-sectional configuration with the aid of a computerized reconstruction algorithm. This reconstruction algorithm quantitatively determines the point-by-point mapping of the relative radiation attenuation coefficients for the set of transmission measurements.

The CT scanning system contains a radiation source and radiation detector along with precision mechanics to scan a cross-sectional slice through the patient. The X-ray detector is usually a linear array of detectors, that is, a series of individual X-ray sensors arranged in a line. Current multidetector CT (MDCT) systems utilize multiple rows of detectors in order to acquire the data in less time. The X-ray source is collimated to form a thin fan beam that is wide enough to expose the detector array. In a single-slice CT, the narrow beam thickness defines the thickness of the cross-sectional slice. The MDCT system slice thickness is determined by detector widths or the grouping of the linear arrays of detectors. The data acquisition system (DAS) reads the signal from the individual detectors, converts these measurements to numeric values, and transfers the data to a computer to be process. This processed is repeated as the X-ray source is rotated around the patient to acquire a full set of transmission measurements (2).

The CT image reconstruction algorithm generates 2D images from the set of measured transmission measurements. There are a number of reconstruction algorithms that can be used to generate the CT image. These mathematic algorithms can be divided into two general categories: analytical or transform techniques, and iterative reconstruction techniques. The transform techniques are generally based on the theorem of Radon (3), which states that any 2D distribution can be reconstructed from the infinite set of its line integrals through the distribution. The line integrals in CT are the sums of the linear attenuation coefficients along a line through the patient determined from the X-ray transmission measurements. The filtered-backprojection 2D reconstruction techniques used in most clinical CT scanners, as well as the cone-beam volumetric reconstruction algorithms based on Feldkamp's method (4) are analytical methods.

Iterative methods are rarely used in medical X-ray computed tomography, but are commonly used in nuclear medicine single-photon-emission computed tomography (SPECT) and positron emission tomography (PET) imaging. These methods are often more tolerant of limited or irregular data, and may use additional *a priori* information to improve the reconstructed results. Iterative techniques are generally algebraic methods that reconstruct the image by performing a series of iterative corrections on a guess of the image distribution (5–7).

EVOLUTION OF THE TECHNOLOGY

Although the mathematical principle of computed tomography was developed early in the twentieth century by Radon, application of the technology occurred much later. Techniques were independently developed in the 1950s for radioastronomy (8) and experimental work progressed through the 1960s, primarily in nuclear tracer imaging and electron microscopy (9,10). Cormack addressed the problem relative to determine X-ray attenuation coefficient information, with the interest of using this information for improved radiation therapy calculations (11). In the late 1990s and early 1970s, Hounsfield at EMI, Ltd in England developed the first commercial X-ray CT system, also known as computer assisted tomography or CAT scanning (12). The initial prototype head scanner was installed in 1971 at Atkinson Morley's Hospital in Wimbleton, England, and commercial systems began delivery the following year.

Due to its unique capability of demonstrating anatomical information the medical interest and demand for CT grew rapidly in spite of the high costs and technical challenges. Numerous manufacturers entered the market with designs to decrease the scan time and to expand the use of CT to body imaging.

First Generation: Translate–Rotate

The initial clinical systems utilized an X-ray beam collimated to a small pencil beam mechanically linked to a detector on the opposite side of the patient. The mechanics translates the tube and detector across the full width of the patient, and then rotates one degree. This process is repeated until a full 180° of data is acquired (Fig. 3a). Two detectors were utilized in the initial EMI scanner in order to acquire two slices simultaneously, which was useful since each scan took over four minutes. One of the innovations utilized by Hounsfield to reduce the necessary dynamic range of the radiation detector, and also minimize certain artifacts, was to have the patient's head push into a elastic membrane into a water-filled box. The box was linked to the X-ray source and detector such that the X-ray beam always traversed through 24 cm of water and anatomy. This was quite effective, but impractical for expanding into body imaging.

Second Generation: Multidetector Translate–Rotate

To reduce the time to acquire the data, additional detectors lying within the scan plane were added and a narrow fan beam was used to cover this detector array. The system translates and rotates like the first generation systems, however, the rotation may be 20 or 30° between translations (Fig. 3b). In this way, the scan time could be as low as 20 s, and body size scan could be performed. While not used any more for medical CT scanners, translate–rotate data acquisition provides considerable flexibility regarding scan field of view and sample spacing, but at the cost of longer scan times. This approach is still used for some research and industrial testing systems which may be designed for samples of several millimeters, or of several meters (13).

Third Generation: Rotate–Rotate

A faster scan approach, which is still the basis for most current clinical scanners, is to utilize a linear array that fully encompasses the width of the patient. Mechanically the tube and detector rotates around the patient to acquire a series of fan beam views $> 360°$ (Fig. 3c). An data set at a particular angle or view with this approach resulted in a fan shaped set of rays with the apex at the X-ray source.

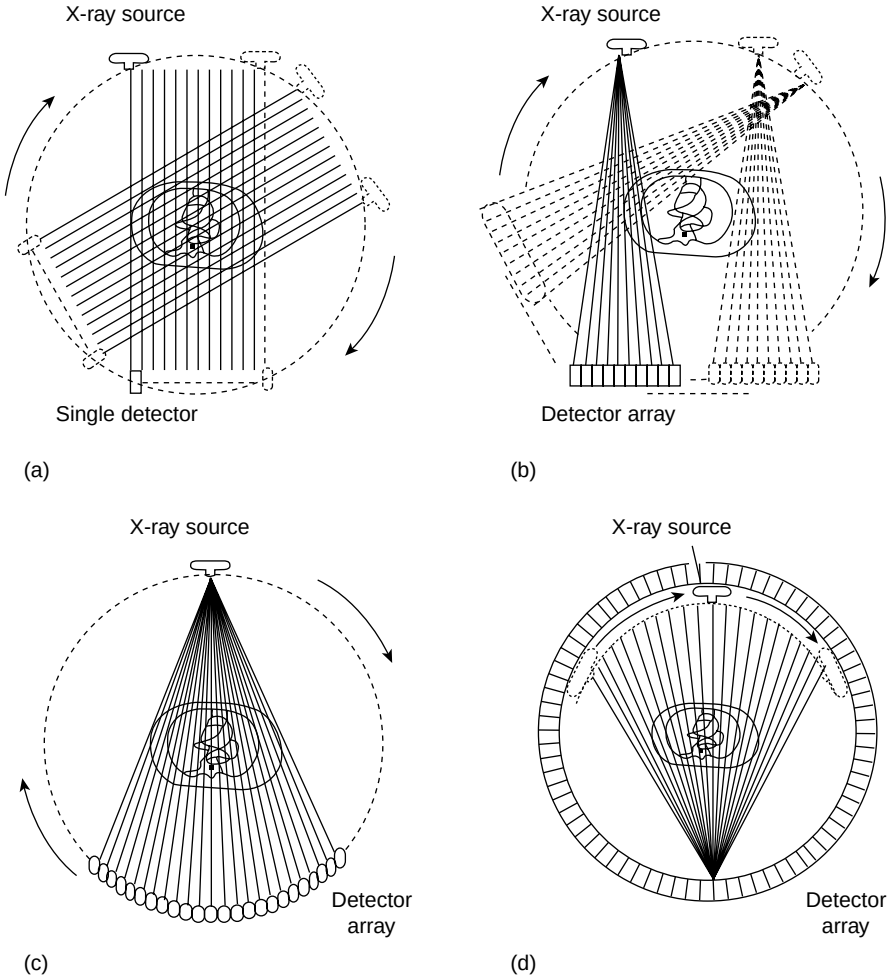


Figure 3. Data acquisition configurations or geometries used in CT. (a) First generation translate–rotate, (b) Second generation narrow fan beam translate–rotate, (c) Third generation rotate–rotate, and (d) Fourth generation fixed–rotate scanning systems. Most clinical systems utilize a rotate–rotate design.

The data sampling flexibility is restricted since the ray spacing is determined in large part by the size and spacing of the detector elements in the linear detector array. The number of views acquired, however, is determined by the number of samples taken over the 360° rotation. To distinguish these systems from the translate–rotate data acquisition systems, manufacturers labeled these rotate–rotate systems as third generation scanners.

Fourth Generation: Fixed–Rotate

Around the same time frame in the mid-to-late 1970s a data acquisition approach using a fixed ring of detectors was used. This requires the X-ray tube to rotate within the circle of detectors, or the use of a mechanism to tilt the detector out of the way of the X-ray beam (Fig. 3d). Usually the acquired data is rebinned or grouped such that a view data set or projection set consists of all transmission measurements made by a single detector as the X-ray tube rotates around the patient. This results in a fan shaped data set, but with the detector at the apex of the fan. Using this scheme the number of detectors determines the number of views acquired, but the ray spacing between views is determined by the data sampling rate. Predictably, the manufacturers of these fixed-rotate scanners labeled them as fourth generation. Further developments including nutating or oscillating ring of detectors, steerable electron beams, 2D detector arrays and helical data acquisition are sometimes given generation numbers, but not in a consistent manner.

Electron Beam CT: Fixed–Fixed

These third and fourth generation rotate-only systems reduced the scan time initially to 10 s, with current scanners capable of rotating around the patient in < 0.5 s. In order to reduce scan times further, especially for rapid dynamic and cardiac imaging, electron beam cine CT system (EBCT) was developed by Imatron Corporation (Fig. 4) (14). This system uses a fixed detector system, but has the X-ray tube target encircling the patient. The unique X-ray tube uses an electron gun and deflection electronics to steer the electron beam within a large cone shaped vacuum enclosure to one of four target rings partially encircling the patient. The X-ray tube ring is opposed by a 240° double ring of fixed detectors. The system has no moving parts since the X-ray source location is changed by the steering of the electron beam. The X-ray source can rapidly move around the patient, and the data for an image acquired in 50 ms or less. The use of four separate target rings and two detector rings permitted the acquisition of eight separate axial planes without moving the patient.

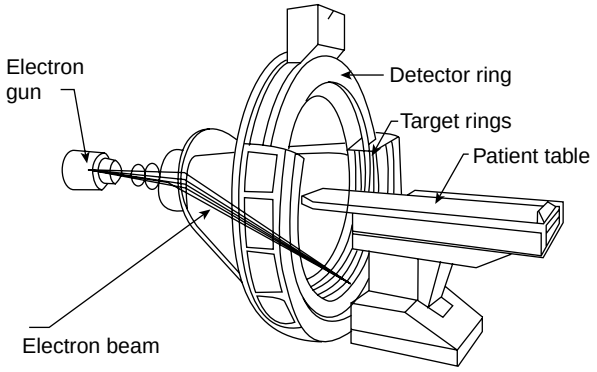


Figure 4. The electron beam CT (EBCT) scanner developed by Imatron (San Francisco) requires no moving parts, but rapidly moves the location of the X-ray source by steering the electron beam in the large cone-shaped X-ray tube to the desired source location.

It should be noted that an alternative research approach to cardiac imaging was developed at Mayo Clinic called the Dynamic Spatial Reconstructor. This system utilized a series of 14 X-ray tube-image intensifier pairs rotating around the patient to rapidly acquire the volume data. The system was designed to enable the use of 28 imaging system pairs (15).

Helical CT

The CT systems through the 1970s and 1980s were generally limited to a single rotation of the X-ray tube around the patient per data acquisition due to the need to have the high voltage cables connected to the tube. In the early 1990s, this changed with the advent of systems that utilized slip rings to transfer the power to the tube, permitting continuous rotation of the source. This continuous rotation allowed for more rapid dynamic scanning where a series of images of a single slice are sequentially acquired, allowing the characterization of motion or to evaluate the flow of a highly attenuating contrast material flowing into the tissue. More importantly, this continuous rotation permitted the ability to rapidly acquire a series of images covering a volume of the patient (16–18).

Normal axial scanning is performed in a step-and-shoot fashion, where the tube rotates around the patient within the plane to be imaged. The acquired data set is reconstructed to form the axial image at this location. The slice location and slice thickness are well defined by the X-ray beam. The patient table is incremented to the next location to be imaged and the process is repeated. The average time per image is the scan time plus the time to increment the table.

With the continuously rotating capability data can be acquired in a helical data acquisition mode. In this mode the table is moved continuously while the tube rotates around the patient. Since there is not a full set of X-ray views through a specific plane of the patient, the data for each angular position around the patient is interpolated from the nearby data acquired at that angle (Fig. 5) (19). Not only is the data acquisition faster, but one can also

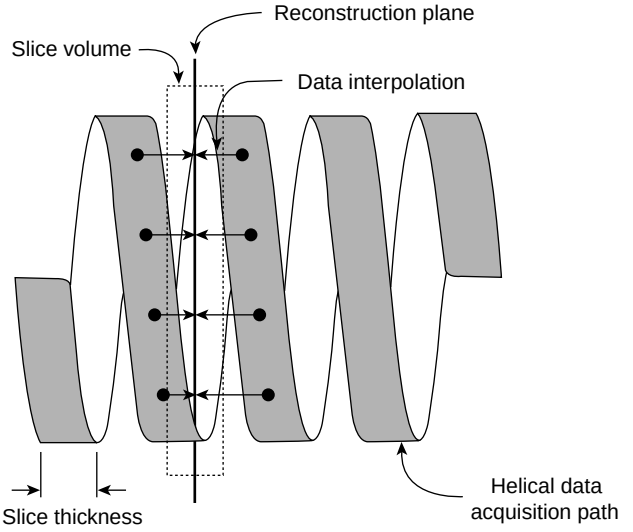


Figure 5. During helical CT data acquisition the patient is moved through the scanner while the X-ray source continuously rotates around the patient. To reconstruct a particular axial slice through the patient, the data at each angular position are interpolated to create a 360° data set corresponding to that slice.

arbitrarily select the locations of the planes to be reconstructed since the data is not fixed to a particular acquisition plane. For example, one could have a collimated slice thickness of 5 mm and generate contiguous or adjacent images every 5 mm, or one could reconstruct images from the same data set every 3 mm (or other arbitrary spacing), however the slice thickness would remain 5 mm.

Multidetector CT

In the latter part of the 1990s, systems were being marketed that contained more than one linear array of detectors. In these multidetector CT systems (MDCT), the slice width of the measured data does not correspond with the overall X-ray beam width, but on the width of the linear detector arrays used to acquire the data. The detector array generally consists of a number of narrow width or thin slice detectors that may be grouped together to generate a thicker effective slice. This detector array allows the acquisition of multiple slices in the axial mode. In the helical mode the overall X-ray beam width is larger than the image slice thickness defined by the detector rows. This permits the acquisition of more data in less time, allowing for faster scan times and the practical scanning with thin slice thicknesses (20,21).

As the number of rows of detectors increase from a few to 64 or 256 and beyond, the data acquired per scan rotation becomes a significantly sized volume. The diverging rays from the X-ray source form a cone of radiation striking the 2D area detector. Reconstruction algorithms developed to deal with these volume reconstructions, as opposed to the axial slice approach on earlier scanners, are sometimes referred to as cone beam scanning and reconstruction. Cone beam scanning can provide rapid information on a volume and is particularly useful for acquiring a rapid sequence of images of a volume to evaluate dynamic processes.

Table 1. Typical CT Number Values

Tissue	CT Number
Air	**−1000**
Fat	−60
Water	**0**
Cerebral spinal fluid	10
Brain edema	20
Brain white matter	30
Brain gray matter	38
Blood	42
Muscle	44
Hemorrhage	80
Dense bone	∼1000

CT SCANNER COMPONENTS

The CT scanners are a union of several component systems to provide clinical imaging capability. Outward mechanics include the table system and the gantry located in a radiation shielded scan room. The gantry contains the X-ray source and detector system. Computers are needed to control data acquisition, reconstruction and display of the images, and for the user interface or control console to allow operation of the system. This may be augmented with additional display and archival capabilities with a picture archiving and communication system (PACS), and workstations for additional display processing and print or filming capabilities Table 1.

Table 1 that the patient lays upon is a fairly basic component. It is typically a cantilevered design with the tabletop extending out from the pedestal, so that only the patient and the tabletop are in the X-ray beam. The tabletop must be strong enough to hold large patients, yet should not provide much attenuation of the X rays. Carbon composite materials are typical used for their strength and radiation transmission properties. Extensions to the table are used for a patient headholder, or for mounting of test and calibration phantoms.

Gantry

The gantry is the donut shaped main body of the computed tomography system and contains the x-ray source and detector system, as well as the mechanics for moving these devices as needed to perform the scan. The patient is extended on the table into the gantry aperture or hole in the gantry so that the X-ray source may rotate around the areas to be scanned (Fig. 6). The scannable region within the gantry is somewhat smaller than the gantry aperture or hole size. Typical is a 50 cm scan field of view within a 70 cm gantry aperture.

The entire gantry is usually pivoted to allow the top of the gantry to tilt toward or away from the table by 30° or more. This allows acquisition of images that are aligned or oriented to specific anatomy, such as the aligned with the disk in the lumbar spine. This feature is being used less, however, with the increasing use of thin slice data acquisition with MDCT systems permitting high quality computer generated images of alternate planes. The gantry also has localizer lights or lasers, and the table and gantry tilt controls to assist the technologist in positioning the patient. The mechanics within the gantry include a large turret bearing, larger than the gantry aperture, to permit rotation of the rotating components of the system, and motor drives and controllers to actuate the scanning motions. Slip rings and data transponders are used to transmit power and data between the stationary and rotating system components. The gantry may also include active or passive cooling devices to prevent heat buildup.

Figure 6. The major system components in the scan room are the patient table, and the scanner gantry which houses the X-ray source, detector, and mechanical drive components.

X-Ray Source

An X-ray tube and generator are needed to produce the radiation for the scan. In the X-ray tube a negatively charged hot filament or cathode emits electrons that are accelerated by a high voltage. The high energy electron strike a target that is part of the positively charged anode and produce X rays along with a considerable amount of heat. The X-ray technique is characterized by the specifying the tube current or mA and the tube voltage or kilovolts, which determine the amount and energy of the X-ray photons emitted. The generation of X rays is the same as is found in other radiographic imaging systems. A notable difference is the workload these tubes endure in clinical imaging. Consequently, the X-ray tubes in CT scanners are often the big-brother to the tubes found in general radiography, with a super-sized anode capable of holding the considerable heat developed during the scans. X-ray tubes designed for CT systems may have other features to prevent anode wobbling, which can cause artifacts, or to be more effective at removing the heat generated. Important parameters for the X-ray tube include its focal spot size, the heat capacity and the cooling rate of the anode. A small focal spot or X-ray source size can provide better image resolution, but a small size may limit the X-ray output that can be obtained. Since the X-ray tubes utilize a rotating anode, it is important that the axis of this anode is parallel to the axis of rotation of tube around the patient, otherwise considerable gyroscopic torque would be placed on the tube.

The X-ray generator includes the high voltage transformer used to create the high voltages necessary for X-ray production. A key requirement for CT systems is to have a highly stable voltage with little ripple or variation. Older systems often used bulky three-phase transformers and voltage rectifiers in order to produce a constant high voltage. Current systems tend to use high frequency single-phase generators. These systems take the utility supplied power and process it to produce a high frequency electrical source with frequencies typically in the range from 1000 to 2000 Hz. The higher frequency has several advantages. High voltage transformer efficiencies are much better at high frequency, and since it is single phase, only one pair of coils is required, making for a much smaller transformer package. Single-phase power is normally associated with 100% ripple as the voltage varies from zero to its peak value. At high frequencies, however, a minimal amount of capacitance in the system smoothes this voltage ripple to produce a nearly uniform voltage. This transition to high frequency transformers has been an enabling technology for helical scanning. In order to continuously rotate the tube around the patient, the high voltage X-ray power cables had to be eliminated. With helical imaging systems, a low voltage of a couple of hundred volts is transferred to the rotating portion of the gantry through an electrical slip ring. The high voltage transformer is mounted on the rotating portion and circles the patient along with the X-ray tube, thereby eliminating the constraint of a single rotation on the older systems. Even with the smaller and lighter generator package, there is considerable mass rotating around the patient, and considerable G forces on these components, especially with the subsecond rotation times.

Collimation and Beam Filtration

Since high energy X rays cannot readily be focused like light, a collimator blocks the X rays coming from the X-ray tube that are not directed at the detector. The X-ray beam is shaped by tungsten or lead plates into its fan beam shape. The width of the fan beam may be varied allowing the technologist to select the slice thickness. On single slice CT scanners, with a single linear array of detectors, the tube side collimation determines the slice thickness. On MDCT systems, the width of the detector or the averaged grouping of detectors determines the slice width. The nominal slice width or thickness is the thickness of the reconstructed voxel at the center of the scanner.

Additional X-ray beam filtration is also in the X-ray beam. Beam filtration is material the X rays pass through before getting to the patient. Legally a certain amount of filtration is required in order to remove the soft or low energy X rays that contribute significantly to the patient dose with little chance of passing through the patient contribute to the transmitted signal. The CT scanner beams are generally heavily filtered, not only to reduce patient dose, but it also reduces beam-hardening artifacts. Most scanners also utilize a bowtie or compensating X-ray filter. This is a filter that has a variable thickness along the length of the fan beam, being thinner at the center of the field and thicker toward the edges of the scan field, thus looking like a bowtie. This filter helps reduce the peripheral dose to the patient and also can help reduce beam hardening variations by adding attenuating material to the portions of the beam that are going through thinner portions of the patient (Fig. 7).

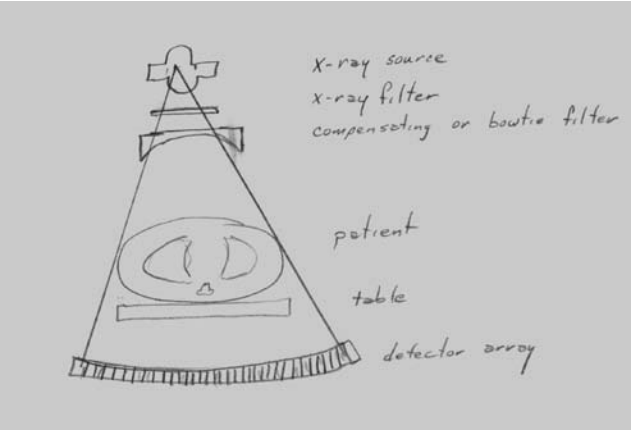


Figure 7. The X-ray beam is filtered to reduce the low energy X rays, passes through a bowtie shaped compensating filter that reduces peripheral dose to the patient, passes through the patient to the radiation detector array.

X-Ray Detector and Data Acquisition System

The X-ray detector is a critical component of the scanner system. It should be efficient at absorbing the X-ray beam energy, and converting the X rays into the detected signal, and should have a rapid response time to allow for rapid data acquisition. The detector size, along with the X-ray tube focal spot size, limits the potential image resolution (22).

Scintillation Detectors. The detector found in Hounsfield's original scanner was a sodium iodide (NaI) scintillation crystal linked to a photomultiplier tube (PMT). These types of devices are commonly used in nuclear medicine counting systems. The X rays are absorbed in the scintillation crystal where they are converted into a number of light photons. The PMT is a very sensitive detector of light and measures the light output. In nuclear medicine counting, the number of high energy photons entering the scintillation crystal is limited and each photon is analyzed and counted separately. With X-ray systems the rate at which photons are entering the detector is much faster than the ability of the system to detect separate distinguishable scintillations or flashes of light. The CT scintillation detectors are operated in a current mode rather than a pulse mode and measure the overall intensity of light produced instead of individual pulses of light.

A number of different scintillating or fluorescent materials have been used in CT scanners including cesium iodide (CsI), cadmium tungstate ($CdWO_4$), and fluorescing materials using rare earth elements, such as gadolinium and ytterbium. Important characteristics of the detector material include its X-ray absorption efficiency, the energy conversion efficiency, and its temporal response. The X-ray absorption efficiency depends on the density and atomic

number of the material, as well as the thickness of the detector. Conversion efficiency is the ability of the fluorescing material to take the energy that is absorbed and convert it to light that can be measured by the light sensitive detectors. When the X ray is absorbed the light is emitted over a short period of time. If this time to emit the light is too long, then this afterglow may influence subsequent measurement. This is one of the reasons that NaI(Tl) is not used in current fast scanners.

Additional factors affecting the detector efficiency is the effectiveness of getting the produced light to the light detecting element and the efficiency of this light detector. The photomultiplier tubes of early scanners have been replaced by photodiode arrays. These components do not have the inherent amplification found in PMTs, but they enable the manufacture of small, closely spaced detectors and the implementation of 2D or multirow arrays.

Gas-Filled Detectors. Gas-filled ionization detectors are another type of detector system that was widely used in CT systems. These detectors operated on the principle that the X rays passing through matter, such as the gas in the detector, causes ionizations or free electrons. A voltage can be placed across the gas to collect the electrons and determine the number of ionizations and the amount of radiation. This type of detector is used in many X-ray survey meters. In order to increase the fraction of the radiation that interacted with the gas and increase the signal level, high pressure xenon gas is used. The electrodes are tungsten plates that are oriented toward the position of the X-ray source. This directional chamber limits the detector sensitivity to radiation coming at an angle from these tungsten plates, thereby providing a capability to reject some of the scatter radiation entering the detector. These systems have been supplanted by the solid-state, scintillation detector systems, especially with the advent of MDCT.

Multiple Row and Area Detectors. The scintillation material in the MDCT detectors is mounted onto a photodiode array chip. The scintillation crystal is diced or sawed to form a series of individual elements. The sawed surfaces, or with the assistance of a reflective coating, help direct the light produced to the light sensitive component directly beneath this element. The width of the detector, in the slice thickness direction, is typically ~ 0.5 mm. The number of rows of data that may be acquired is often limited by the data acquisition system (DAS). The signal from a series of rows may be combined to produce an effective slice thickness that is some multiple of this value. This may be done prior to the digitization, allowing for a thicker slab of tissue to be scanned per rotation, or may be done as a postprocessing technique to reduce image noise.

An example may be a four slice CT scanner with a series of 1.0 mm detector rows covering a total width of 20 mm. It is limited to acquiring four rows of data by its data acquisition system. A scan may be performed with 4×1 mm detectors for a total beam width of 4 mm, or 4×2 mm for a beam width of 8 mm, up to a 4×5 mm for a beam width of 20 mm (Fig. 8). The first approach would give the best interslice resolution, while the latter would allow one to scan a given volume in less time.

The DAS must provide a highly accurate digitization of the signal and is handling a tremendous amount of data. As an example a 64-slice scanner may have 64-active rows of detectors each containing 1000 elements. As the scanner rotates around the patient in 0.5 s, 1000 measurements are made from each of these elements. That results in 128 million precision measurements made each second. This value increases as more rows of detectors are added and area array detectors for cone beam scanning are used. Data acquired during the scan is transmitted by a telemetry system to the fixed portion of the gantry. The data is sent to a computer that utilizes array processors for rapid data

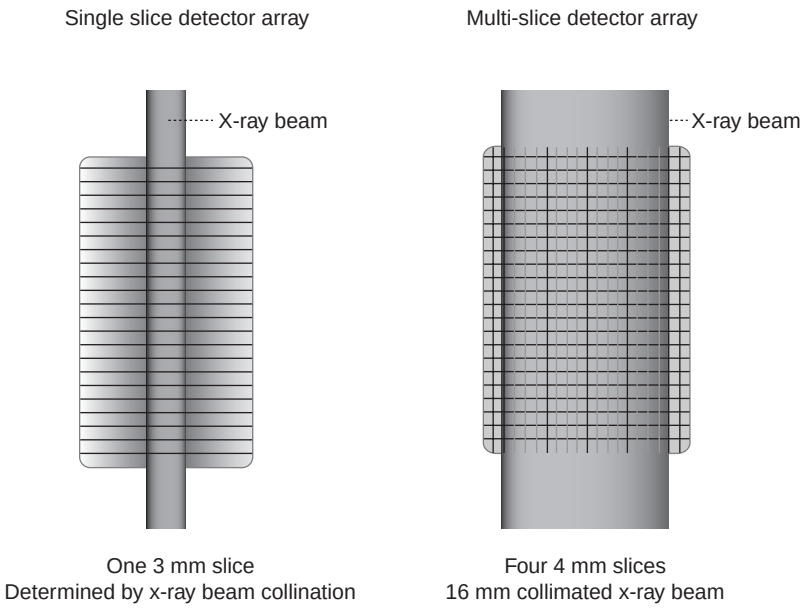


Figure 8. (a) In a single slice CT scanner the entire width of the detector is active and the slice width is determined by the collimated width of the X-ray beam. (b) Multidetector CT slice width is determined by the effective detector width. Individual detector elements may be grouped to yield a larger effective slice thickness. In this example a detector array consists of 20 rows each 1 mm wide. A four slice CT system may use groupings of four rows yielding 4 rows of 4 mm wide detectors as shown, or can use other groupings of the 20 rows.

reconstruction, and manages the storage and display of the resultant images.

Computer and Operator Console

The operator console utilizes an interactive computer display and dedicated function buttons to allow the procedure setup, scan initiation, image display, and data storage. The demographic information for the patient may be received from the facilities radiology or hospital information system (RIS or HIS) or entered by the technologist. Editable routine scan protocols facilitate scan setup and preview radiographic type image is used to identify the specific volume of the patient to be scanned. The images are displayed and some image processing and measurement features are available. A limited amount of the raw transmission data is stored on the system and may be used to for additional reconstructions from the data with alternate parameters. The reconstructed images may be filmed and the image data archived on the scan computer system, or transferred to a central PACS system for storage and for remote image display.

SCAN PITCH AND EFFECT ON PATIENT DOSE

One of the technique parameters set when performing a helical scan is the scan pitch, which is like the pitch on a screw. This refers to the ratio of the distance the table moves per 360° rotation of the X-ray source to the thickness of the X-ray beam. If the table moves the same distance as the beam width each rotation, then the scan pitch is one. The radiation dose with a pitch of one is similar to that obtained in a step-and-shoot axial mode where the table incrementation between scans equals the slice or beam thickness. With these contiguous axial images the entire surface of the patient within the area scan is struck once with the primary, unattenuated X-ray beam. Having a pitch < 1 indicates that overlapping data is acquired with a commensurate higher average dose, and a pitch greater than one results in gaps between primary exposed areas and a lower average radiation dose. A pitch < 1 requires less data interpolation and yields sharper slice profiles, while pitch > 1 reduces dose, but may blur the slice thickness profile and is more subject to certain image artifacts. With some CT systems this change in the effective technique and average dose as a result of the selected pitch is reported as the effective mAs. The effective mAs is the X-ray tube current (mA) times the time per rotation divided by the pitch.

The concept of pitch gets a little more complicated with MDCT systems. With single-slice CT systems the slice thickness corresponded to the detector width. In MDCT systems, there are multiple rows of detectors covering the width of the X-ray beam. This leads to two separate, but related pitch values. There is the collimator pitch that relates the table motion to the overall X-ray beam width, and the detector pitch that relates the table motion to the width of the individual detector rows (or their combined width when rows are combined prior to digitization) (23).

Consider an example with the four slice scanner with 20 rows of 1.0 mm wide detectors described above with a scan time per 360° tube rotation of 1 s. If the data acquisition mode is 4×2 mm, that is to simultaneously acquire four sets of data from detectors having a detector width of 2 mm, then four pairs of 1.0 mm physical detectors will be combined to produce four detector rows each with an effective 2 mm detector width, and the overall collimated beam width is 4×2 mm or 8 mm. If the table incrementation speed is $6\,\text{mm} \cdot \text{s}^{-1}$ or 6 mm/rotation, then the collimated pitch is

$$\text{Collimator pitch} = \frac{\text{Table travel per tube rotation}}{\text{Collimated beam width}} \tag{1}$$

$$\begin{aligned} &\text{Collimator pitch} \\ &= \frac{\text{Table travel per tube rotation}}{\text{Number of detector rows} \times \text{Detector width}} \\ &\text{Collimator pitch} = \frac{6\,\text{mm/rotation}}{4 \times 2\,\text{mm}} \\ &\text{Collimator pitch} = 0.75 \end{aligned} \tag{2}$$

A collimator pitch < 1 indicates that the radiation fields are overlapping, which will result in a patient radiation dose higher than an equivalent set of contiguous slices or a pitch of one. The detector pitch in this example is given by

$$\begin{aligned} \text{Detector pitch} &= \frac{\text{Table travel per tube rotation}}{\text{Detector width}} \\ \text{Detector pitch} &= \frac{6\,\text{mm/rotation}}{2\,\text{mm detector width}} \\ \text{Detector pitch} &= 3 \end{aligned} \tag{3}$$

CT SCAN TECHNIQUES

Preview Digital Radiograph

There are several scan modes or types of data acquisition that a CT scanner may be to acquire data. One commonly used technique is the acquisition of a scout or preview scan. These scans are basically a digital radiographs that are used to set up the tomographic imaging sequence, or may be used to visually locate the position of an axial slice on a radiographic reference image. To acquire the preview image the X-ray source and detector remain stationary. The detector sees a single line of an X-ray transmission image. As the table and patient are moved through the X-ray fan beam, the series of transmission lines acquired generate the radiographic image.

From the preview scan the technologist can define a range or volume within the patient to be scanned. Lateral preview scans can be used to determine the proper gantry angulation to orient the tomographic slices with desired anatomical structures, such as to the intervertebral disks in the spine.

Axial CT Scan

An axial scan is a basic CT scan, normally implying the data being acquired without the table moving during data

acquisition (Although an axial image also refers to any image orientated transverse across the patient, as opposed to sagittal or coronal plane orientations.) Prior to the advent of the continuously rotating helical scanners, all CT scans were acquired with a stationary table. For a single-slice scanner the slice thickness or slice profile is defined by the collimation of the X-ray beam with the detector width being somewhat larger than this beam. For a multidetector CT the effective detector width of the rows of detectors tends to be the primary factor in determining the slice thickness. The effective detector width may be the summation of several physical rows of detectors. The grouping of detector rows may be to form thicker slices in order to reduce the image noise and number of images generated, or may be due to data acquisition system limitations.

The MDCT systems may be limited in their ability to acquire axial images due to the divergence of the fan beam. With a single detector row all of the transmission rays passed through a particular plane within the patient. The beam divergence along the slice thickness orientation caused some variation in the detected slice profile, but it was relatively minor. With the MDCT systems the data seen by the row of detectors on the ends is not consistently within a single plane due to the angulation of the diverging X-ray beam. This can cause inconsistencies in the data and may cause image artifacts or errors.

Helical CT Scan

The primary advantage to the continuously rotating source and detector is the ability to do helical or spiral CT scanning. Data is acquired as the patient is moved through the beam. There is no set of measurements where one has transmission data from all angles around the patient, but adjacent measurements are used to estimate the data corresponding to a particular plane. This mode allows for the rapid acquisition of data though a patient, and the ability to reconstruct images at any location within this volume. This rapid scanning allows procedures to be done quicker, often allows data to be acquired within a single breathhold, minimizing motion blurring, and facilitates the ability to perform contrast enhanced angiography studies to evaluate major blood vessels.

Dynamic Scan, Fluoro CT, and Triggered Scan Start

Another mode of data acquisition is dynamic scanning. In this mode a series of images are sequentially obtained at a single location. This can be used to analyze motion, or more commonly to evaluate the flow of contrast material into a tissue. This capability prior to continuously rotating systems was limited to one scan every few seconds since the tube had to stop and reverse motion between scans. Continuously rotating systems not only can acquire a sequence of images with no time gap between them, but also can obtain images overlapping in time where the time spacing between images is shorter than the data acquisition time for the image. Dynamic scanning can produce a series of images to assist in evaluating a tumor or mass by how it enhances or changes as iodine contrast material flows into the tissue, or it may be used for quantitative analysis of the tissue perfusion.

A variation of this is fluoro or fluoroscopy mode CT. Here a series of images at a location are dynamically scanned and rapidly reconstructed to allow the technologist or physician to see the image in real time. This may be used to assist in a CT guided invasive procedure. Note that another approach is to have a conventional X-ray fluoroscopy system adjacent to the CT where the fluoroscopy is used for needle or catheter guidance and the CT is used to verify and evaluate results. Computed tomography fluoroscopy may also be used to visualize the arrival of injected contrast material into a vessel. This information may be used to initiate a scan sequence to catch the maximum concentration of the contrast media in the vessels of interest for CT angiography. Angiography scan starts may also be assisted using a feature where the computer evaluates the transmission data through a defined vessel and triggers scan start when a sufficient attenuation increase is detected.

Cardiac Gated CT

Physiologic motion can degrade the image quality. Fast helical and MDCT techniques allow for single breathhold studies. With the exception of the electron beam CT systems, a full set of data cannot be acquired of the heart without motion. In order to freeze the cardiac motion, the data is acquired and characterized relative to the cardiac cycle and is selectively grouped to obtain images without the typical motion blurring. This gated imaging requires an electrocardiogram (EKG) or similar input from the patient to define the cardiac cycle (24). Since the heart is relatively stationary during the longer diastolic rest phase than during systolic contraction, the gating may also be used to eliminate or minimize the systolic data to produce a diastolic only image. Alternatively, data can be acquired over many cardiac cycles and binned to produce images for various portions of the cardiac cycle. This multi-phase imaging process is similar to what is done in gated nuclear medicine and MRI studies. The series of images may be viewed in a movie mode to visualize the beating heart, and may be analyzed regarding wall motion and cardiac output.

CT NUMBERS

The linear attenuation coefficient is scaled into an integer pixel value. Medical systems utilize an offset scale that is normalized to water. This scale assigns air a CT value of –1000, water is at 0 and a material twice as attenuative as water has a CT value of +1000, and so on. The CT number is an integer relating to the attenuation properties of the tissue by the following formula.

$$\text{CT number} = \frac{(\mu_{\text{tissue}} - \mu_{\text{water}})}{\mu_{\text{water}}} \times 1000$$

Where μ_{tissue} and μ_{water} are the linear attenuation coefficients for the tissue in the particular voxel, and of water, respectively. Typical CT number values for some common tissues and test objects are listed in Table 1.

Display Window and Level

The CT numbers are commonly stored in the computer as 12 bit integers covering a CT number range from -1000 to $+3000$ (or -1023 to 3072). To display the full possible range of data one needs >4000 shades of gray or displayed intensity. The human visual system, however, is limited, and we generally can discern something closer to 30 different shades. A common technique with all of the digital imaging methods is to use a viewer selectable mapping of the digital numbers representing the image to the various displayable intensities. There are a number of variations and processing methods that can be applied, but one of the most basic and most used methods is to define a display window level and window width.

The window level value defines the CT value that will be mapped as the middle gray intensity. The window width is the range of CT numbers that will have a range of gray values from black to white. Everything below the lower range value (the window level minus one-half of the window width) will be black and everything above the upper range level (the window level plus one-half of the window width) will be white. By adjusting these levels one can ignore the air-like CT densities, and display all the dense structures, such as bone as white, while obtaining a relatively high contrast view of the a narrow range of CT numbers corresponding to the soft tissue densities within the body. On the computer display one can easily vary these settings to look at low density structures in the lung or the high density detail of the bone if desired. Example of the effect of display window settings on the displayed image is seen in Fig. 9.

Other variations to this gray scale mapping function can also be performed, such as histogram equalization, where the resultant display will have an equal number of pixels for each gray level. The display may also be done in color where each CT number is mapped to a particular color. This is sometimes referred to as psuedo-color to emphasize that the displayed color is not that of the object, but some arbitrary assigned color. Clinical CT generally does not use color for basic cross-sectional image viewing. Color is commonly used, however, for processed data displays, such as 3D surface imaging where one views the surface of organ structures or of the vascular tree, or may be used as an overlay over the gray scale anatomical image with the color representing some functional feature, such as blood perfusion.

DISPLAY TECHNIQUES

Film and Soft-Read Workstations

Traditionally the cross-sectional images generated in a clinical procedure are windowed as appropriate for the tissues of interest, and then photographed or printed onto a large 14×17 in. (356×531 mm) transparent film. If necessary, two sets of films may be made to have the window level and width adjusted for two different CT number ranges, such as for soft tissue and for the lower densities within the lung. The films provided a highly portable record of the study that can be illuminated with any X-ray film viewbox, and provides a medical record of the procedure. This was a manageable process producing a handful of films when used with single slice scanners acquiring relatively thick slices (3–10 mm) through a volume of interest.

With the fast MDCT systems one can rapidly scan through the same volume of the patient with thin slices. This results in hundreds to thousands of images for a single procedure. This would result in many dozens of films per study, which is not only expensive, but also unwieldy for physician review. This has been one of the drivers to implement a picture archiving and communication system (PACS) (see PACS topical entry), which enables the use of computerized soft-read workstations for the primary analysis of the image set.

Analyzing the images from the computer display provides a number of interactive tools for the reviewer. The ability to interactively change the window level and width is a powerful function for evaluating subtle features. One can measure the area and average CT number within a region-of-interest (ROI), measure distances and angles, and magnify regions of the image. One can rapidly page through a stack of images providing a better view of the continuity of structures from slice-to-slice.

Alternative Image Plane Display

A number of processing techniques are available for analyzing and presenting the volume information contained in a stack of axial slices (Fig. 10a). An alternative plane through the patient can be generated through this volume. This can be a coronal (frontal) plane, a sagittal (lateral) plane (Fig. 10b and c), or an arbitrary oblique plane. A series of parallel oblique planes may be reconstructed in a batch mode using the multi-planar reformation feature of the scanner or workstation, or the location may be interactively defined and displayed. The reformatted slice may be generated with a definable slice thickness down to the voxel size of the data set.

Maximum Intensity Projection

The displayed data in the oblique plane display is an average of the voxels contributing to each of the reformatted pixels. An alternative is to display an intensity value that corresponds with the largest voxel value in these contributing pixels. This type of display is referred to as maximum intensity projection (MIP). The slab thickness for the MIP image can include the entire volume scanned or a thinner slab (Fig. 10d). The MIP image gives a 3D type presentation for viewing dense structures such as bone or contrasted blood vessels, especially when rotating the viewing angle.

3D Surface Imaging

Another volume viewing technique is the 3D surface imaging. If structures can be characterized by their CT number range, the contours of the structure can be defined and surface view formed. These images have much in common with the visualization techniques used by computer-aided design or in computerized animation in the entertainment

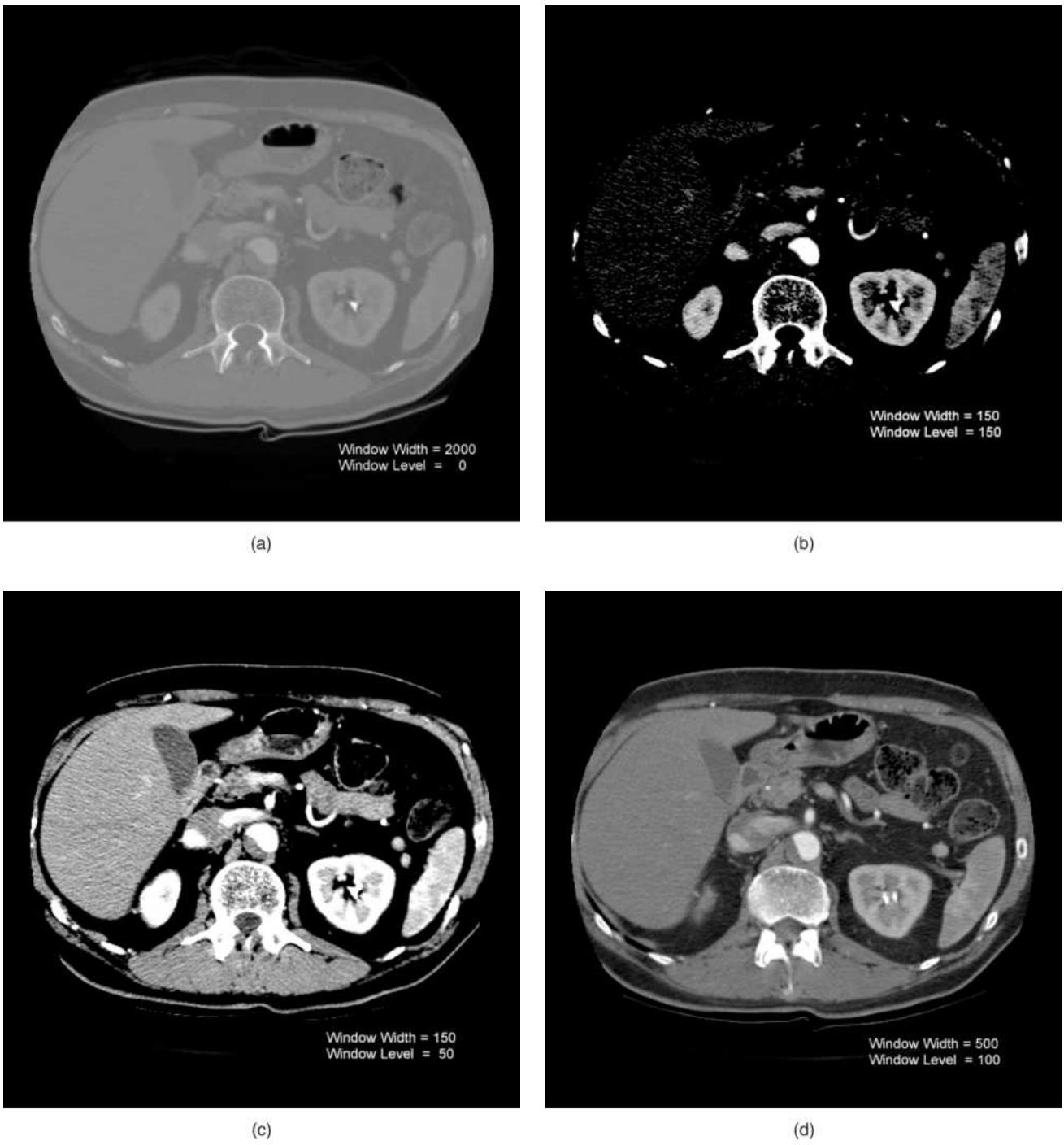
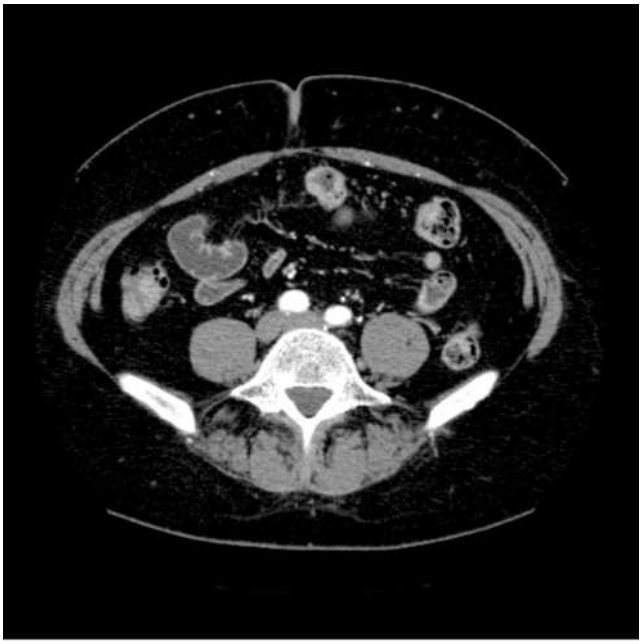


Figure 9. Various display window width and window level settings for the same abdominal CT image. (a) A wide window (2000 CT numbers) shows nearly the full range of CT numbers, but without discernable contrast between soft tissue structures. (b) A narrow window (WW = 150) yields high contrast between structures, but the window level centered at 150 results in most soft tissue being black because they are below the window range, with only the bone and structures containing iodine contrast media being seen. (c) A narrow window (WW = 150) centered at 50 results in a high contrast visualization of the soft tissue. (d) A typical display window (WW = 500, WL = 100) may compromise to provide good contrast while displaying a wider range of structures with lower displayed noise.

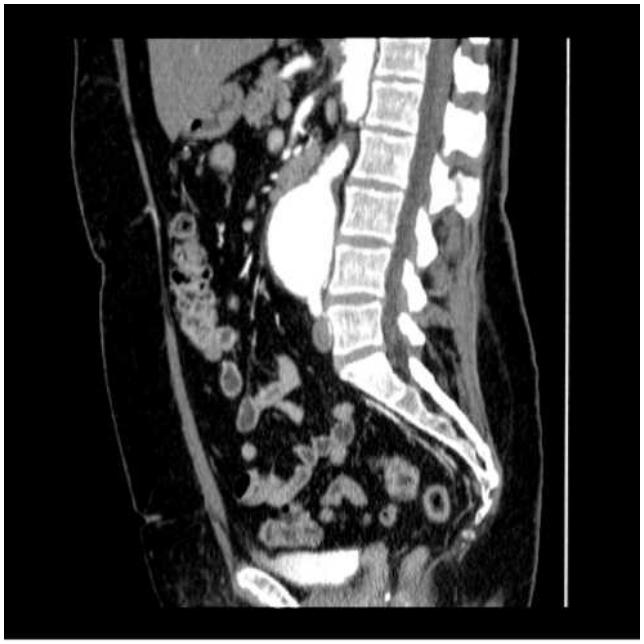
industry. The surface defining points are connected and plated with tiles or surface segments (25). The display software will project the nearest surfaces to the displayed image and use features such as distance from the viewer and angulation of the surface to define the brightness intensity. Light source position and coloration may also add to the display. Several different structures with differing CT number ranges can be simultaneously displayed

with different color schemes for each structure. In this way, bone may be shades of white while a tissue or vascular structure may be red (Fig. 10e and f). The structure may also be given the property of transparency allowing visualization of deeper features, such as visualizing the ventricles of the brain through a visible but transparent skull. With specialized displays a stereoscopic pair of images may be viewed, enhancing the 3D effect, however, the ability to use motion and rotate the structures on the display is very effective at producing a 3D view.

Three-dimensional views may be enhanced with some computerized surgery. The user can select certain structures to be eliminated from the image. This selection can be made by defining cut planes or surfaces from various views and erasing structures outside of a volume of interest. Connectivity tools may also be used where the cursor is used to identify a structure, and then all surface points that are contiguously connected to the seed point are either selected or erased. In this way, one may select a vascular tree that is otherwise obscured by bony structures with

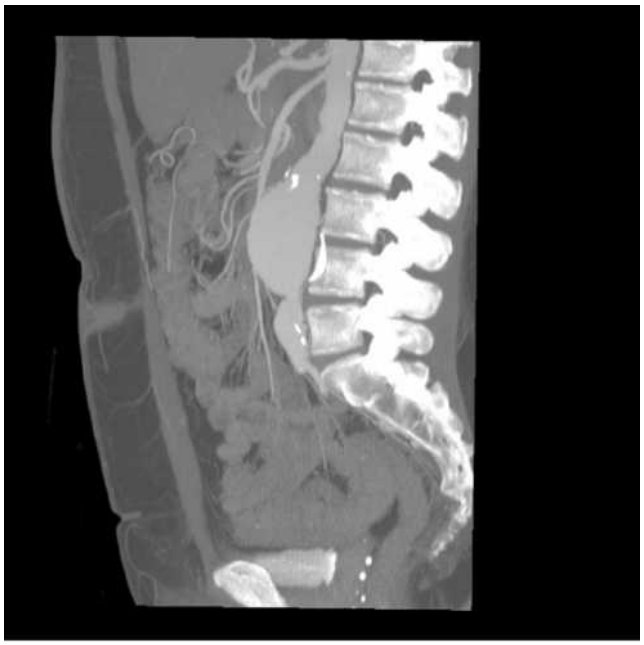


(a)



(b)

(c)



(d)

Figure 10. Alternative image sets can be generated from a series of closely spaced axial images. The image in (a) is one of 266 2.5 mm thick axial images. Iodine contrast media has been injected into the patient to make the major blood vessels visible, including the balloon shaped aortic aneurysm. From this data set the computer can generate (b) a sagittal image, or (c) a coronal image, or (d) a maximum intensity projection (MIP) image of a sagittal slab containing the spine and aneurysm. Structures may be identified by their CT number range to generate 3D surface images of bone and contrasted vessels (e), and structures may be removed to produce a vascular tree image.

similar CT numbers. Problems with this approach occur when the two structures touch making a connectivity bridge between them.

SPECIAL CLINICAL FUNCTIONS

Surgical Planning

The ability to produce a 3D visualization of structure surfaces can be used in several ways. It is useful for general viewing and obtaining an overview of certain structures. This may be useful in seeing areas that should be scrutinized more closely, and appears useful in communicating anatomical findings to surgeons and other physicians that are more familiar with the physical anatomy rather than a series of cross-sectional slices through it. In some cases data may be obtained from these images to assist in surgical planning, including the repositioning of bone fragments or the appropriate type and size prosthetic hardware to use.

CT Angiography, Virtual Colonoscopy, and CT Perfusion Imaging

Computed tomography angiography (CTA) is the procedure used to visualizing the blood vessels (26–28). Iodine contrast media is injected into the patient to increase the attenuation and increase the CT number of the blood within the vessels (Figs. 10e, f, and 11). Timing of the

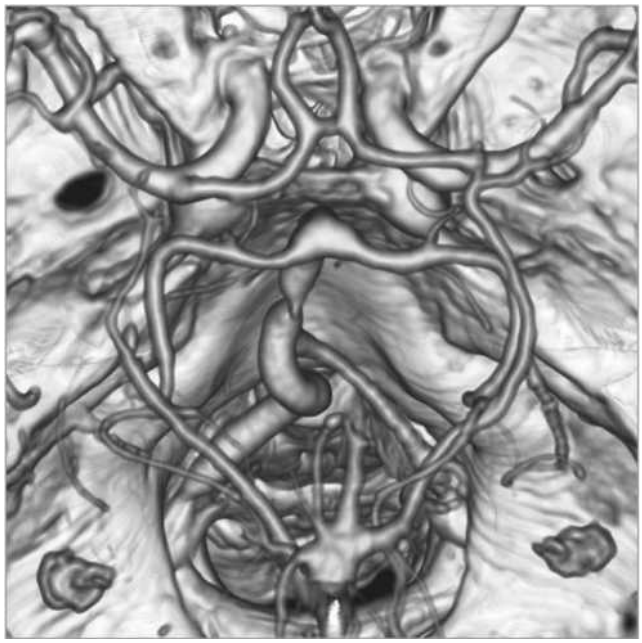


Figure 11. Three-dimensional view of a CT angiogram of the brain arteries including the Circle of Willis along with the skull structures as viewed from the top of the head.

CT scans is important in these procedures since the contrast media will return through the venous system and obscure the visualization of the arterial system, hence the use of some of the previously described scan start techniques. Besides general 3D viewing of a vascular tree to produce a CTA, other related techniques may provide additional information. Two points in a vascular tree may be identified and the computer can locate the line within the scan volume corresponding to the center of the vessels connecting them. The vessel along this line may be analyzed producing a plot of the vessel diameter or cross-sectional area. A stenosis appears as a reduce area, while an aneurysm may be seen as a greatly enlarged area. Since the vessel wall is defined as the surface between the high X-ray densities within the vessel to the water-like tissue densities outside the vessel, one can use 3D visualization techniques with the viewer located inside of the vessel. The viewer may travel or fly-through the vessel and visualize the structure of the lumen surface.

This fly-through technique is also the basis of virtual colonoscopy. In a clinical colonoscopy, the bowel is prepped to remove residual feces. An endoscope is inserted into through the anus into the colon and a camera and light source allows visualization of intestinal surface. If suspicious polyps or lesions are located, devices may be guided through the endoscope to remove or sample the tissue. In virtual colonoscopy bowel preparation is still needed and the colon inflated with air to produce a well-defined interface at the wall surface. A series of CT slices are acquired and 3D visualization and fly-through techniques are used to view the structure of the colon surface (Fig. 12).

Besides visualizing the vessel by CTA, the vascular condition of the tissue may be analyzed using a CT perfusion procedure, especially in evaluating the brain. Acquiring the data in a perfusion study requires obtaining a series of images of selected slices over a short period of time as iodine contrast media or inhaled xenon gas in the blood flows into and washes out of the tissue. Various parameters may be measured, such as mean transit time (MTT) showing how fast the blood reaches the tissue. This may provide some indication of blood shunting or obstruction. The enhancement curve may be analyzed to obtain a relative cerebral blood volume (rCBV) and relative cerebral blood flow (rCBF) images. This information may be useful in evaluating strokes and obstructive disease.

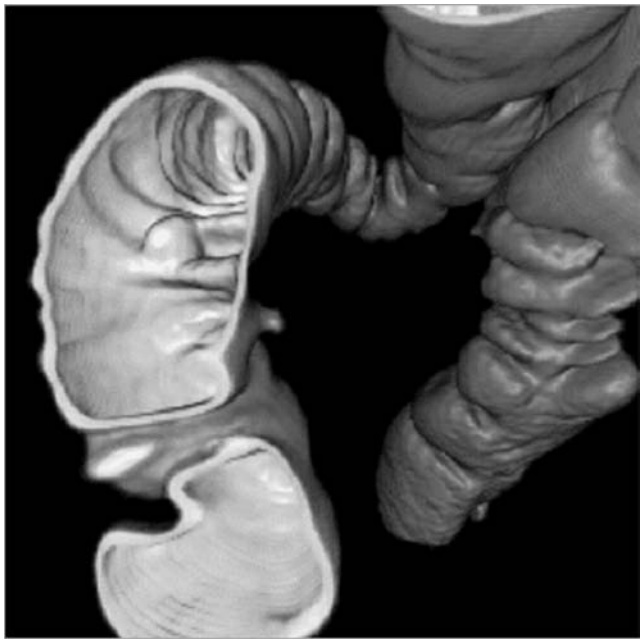


Figure 12. A 3D surface image of the interior wall of the colon allows a virtual colonoscopy fly-through to inspect the intestinal wall for polyps.

Quantitative Analysis: Bone Density and Calcium Scoring

In general clinical CT scanners are not designed to produce highly accurate attenuation data, but rather high quality diagnostic images with minimal artifact. A number of factors can affect the calculated CT number value within a pixel, including its location and the size of the patient. With this caution, however, there are several applications where the analysis of the CT numbers is valuable. In general image interpretation, the CT value of a tissue lesion may be made to help determine if it is a mass, a cyst, edema, or a hemorrhage. Other scans may be performed specifically for the quantitative analysis. One screening procedure is calcium scoring. The calcium plaques in blood vessels will increase the CT value of the corresponding pixels. An evaluation of the amount of calcium in coronary arteries is an indicator of cardiac risk and may be measured by CT (29).

Osteoporosis is the loss of calcium bone mass, especially prevalent in postmenopausal women. The CT technique may be used to analyze the calcium content of the bone. Usually the trabecular bone in the middle of the spinal vertebra is analyzed due to their large surface area and sensitivity to bone loss. In order to produce reproducible data, the measurements made of the bone are compared to other reference densities within the image or in a comparable image. This may be done by having the patient lie on a phantom containing known reference materials, or comparing the measurements to other tissue in the image with known CT values (30–32). Most bone mineral densitometry, however, is performed with dedicated systems rather than using CT scan procedures.

Radiation Therapy Treatment Planning

Another group that would like quantitative information is the radiation oncologist for use in radiation therapy treatment planning. Some CT scanners are dedicated for radiation oncology use and these CT simulators may have special features and software to assist in radiation therapy simulation. In radiation therapy treatment planning it is important to be able to define the target tissue to be irradiated and the adjacent sensitive tissues, and to have them in the same position and orientation that they will be at the time of treatment. Since most linear accelerators used for treatment have flat tables, a hard flat table pad should be used for the corresponding CT scan. Likewise the body position, such as the position of the arms, should be as

it will be during treatment. Skin markers and CT visible fiducial markers may be used to orient and register the images with the treatment plan.

Besides seeing the pathology and anatomy to identify the targets for the treatment planning, obtaining information on the attenuating properties of the various tissues to the high energy photon and electron beams is useful for accurate treatment planning. The problem is that diagnostic CT scans are acquired at relatively low photon energies as compared to that used in therapy. The diagnostic CT X-rays are much more sensitive to the atomic number of the materials within the voxels than are the high energy therapy beams. Characteristics of known tissues are used along with the measured CT numbers to estimate the physical or electron density of the tissue and its high energy attenuating properties.

Dual-Energy Scanning

One approach that can be used for quantitative imaging, in particular to determine effective atomic number and density of the tissue is dual energy scanning. Using two different X-ray beams will produce data corresponding to the attenuating properties at the two separate effective energies. At diagnostic X-ray energies the primary attenuation processes are photoelectric absorption and Compton scattering. Photoelectric absorption is highly dependent on the atomic number of the material and the probability of interaction falls off rapidly with increasing photon energy. Compton scattering is relatively independent of the atomic number and falls off at a much slower rate. That is, the probability of photoelectric absorption is proportional to Z^3/E^3, while Compton scattering falls off with $1/E$. This information may be used to take the CT measurements and calculate an alternate pair of basis images, such as effective atomic number and density (33). This is preferably done with the transmission data, but may be performed with the reconstructed images. Challenges exist in these calculations, however, due to other factors in the imaging process, and methods used to correct for other systemic errors.

Stereotactic Surgery Planning

Stereotactic surgery utilizes a hard fixed frame to the head to direct a needle to a very particular location in the brain. The base frame usually is attached to the skull with screws or pins. During CT scanning a localizing frame is attached to the base. When the CT scans are analyzed, the target location is identified in the images. The location of various frame components are also identified and recorded. This information is used to localize the target in the 3D frame space. During the surgical procedure the localizing frame is replaced with a needle guide that can be set for insertion to the target spot. The same approach is used for stereotactic radiosurgery where thin radiation therapy beams are used to irradiate specific targets in the brain.

PET–CT Image Fusion and PET-CT Scanners

Positron emission tomography (PET) scanning is a specialized nuclear medicine technique for generating cross-sectional images of the distribution of positron emitting radioactive tracers in the body. In this task, it is very sensitive at presenting this information. It is, however, relatively poor at presenting high resolution detailed anatomy. The PET images may be fused with corresponding CT images to delineate the structures containing the radioactive tracer. This is usually displayed as a color PET image overlaid onto a grayscale CT image. The alignment of the two data sets may be performed manually or automatically by various computer algorithms. A key aspect of this image fusion is the patient being in identical positions for both data sets. This can present problems including different table shapes, arm position, flexure of neck and back, or changes in the patient between the scans.

Much of the difficulties in image fusion are eliminated by the use of a specialized system that contains both the PET scan capability and CT scan capability (34,35). Typically these are two relatively independent scanners with a connected gantry and utilizing a single patient table system. One of the steps in PET scanning is to acquire transmission measurements through the patient in order to perform accurate attenuation correction of the data. On a stand-alone PET scanner, this is acquired by use of a radioactive source that emits similar photon energies. With PET–CT the CT image data may be used to determine the PET attenuation correction. Note that the CT scan represents X-ray attenuation properties at diagnostic X-ray energies, which are much lower than the 0.511 MeV photons from the PET radionuclides, and appropriate corrections must be applied.

RADIATION DOSE

Computed tomography is an X-ray procedure with an associated radiation dose. X rays are ionizing radiation, meaning that the X-ray photons have sufficient energy to rip orbital electrons from atoms. As a consequence, small amounts of absorbed energy can cause biochemical actions that may have biological consequences. Radiation dose is the amount of energy absorbed per mass of tissue at a defined location and is measured in rads or preferably in the SI unit of grays, where

$$1\,\mathrm{Gy} = 1\,\mathrm{J}\cdot\mathrm{kg}^{-1} = 100\,\mathrm{rads} \tag{4}$$

Related are units of effective dose, the rem and sievert, which estimate the whole-body dose that has an equivalent long-term risk as an actual dose to just part of the body (36). Risk from radiation exposure can be divided into a couple of categories. Nonstochastic or deterministic effects are those that will happen if a certain radiation dose is received. Most relevant to diagnostic imaging are skin effects, such as erythema, the reddening of the skin. These effects require several gray of dose, which is significantly higher than doses normally encountered in CT. Stochastic or statistical biological effects are of some concern and should be part of the risk–reward evaluation for the procedure. The principal stochastic effect is the increase risk of getting cancer as a consequence of the radiation exposure. The risks are relatively small, but

unwarranted radiation exposures should be avoided. Since developing embryos and fetuses are especially sensitive to radiation, special cautions are often taken to minimize *in utero* exposures.

Computed tomography is a bit different from standard radiographs relative to the total dose received. The maximum entrance dose to the skin may be quite similar between a CT scan and a radiograph, but in radiography the intensity of the radiation decreases due to attenuation as it passes through the patient. Consequently, the dose to deep structures is much less than the surface dose, and the dose at the exit surface can be orders of magnitude less than the entrance dose (37,38).

In CT, the X-ray source rotates around the patient, such that the entrance surface is not just on one side of the patient. This results in a more uniform dose and considerably more total energy deposited in the patient. In a typical head scan, the dose across the imaged slices is fairly uniform, and for body sections the midline dose is approximately half that of the surface dose. This results in a much higher effective dose to the patient. It is estimated that CT accounts for $\sim 10\%$ of the radiology imaging procedures, but amounts to around two-thirds of the total effective dose patients receive, and these values are likely to increase with the increasing utilization of CT.

Measuring radiation dose in CT presents some challenges. The X-ray beam is a narrow fan beam and may not even be constant across its width. Bow-tie compensating filters may further vary the beam intensity along the length of the fan beam. We see that slice width and spacing, and helical scan pitch, as well as the patient size, are also factors affecting the average dose.

Dose across the slice thickness, either in air or within a plastic phantom that simulates the patient, may be measured with a stack of thermoluminescent dosimeter (TLD) chips, with a radiation sensitive dosimetry film, or with a photoluminescent dosimeter strip. This data can be useful in characterizing the dose profile and the amount of scatter radiation present. Acquiring this data is cumbersome, and the use of this information to estimate a dose from a series of scans can be complex.

An alternative is to measure the CT dose index or CTDI. If one considers a series of contiguous slices, where the distance between the centers of adjacent slices is equal to the slice thickness, then the dose to a particular point in the patient is equal to the primary dose from the slice containing that point, plus the scatter radiation from the other slices. The CTDI is effectively this multiple slice average dose. It is measured with a long thin cylindrical chamber, typically 100 or 140 mm in length, about the size and shape of a pencil. It is exposed with a single axial scan. If the slice thickness is 5 mm, then the center 5 mm of the chamber receives the primary exposure. The adjacent 5 mm segments encounter the exposure for the adjacent slices, and the next 5 mm the scatter dose two slices away, and so on for the full length of the chamber. If one normalizes the measurement for to the 5 mm primary segment length, the overall measurement is the exposure from the primary beam plus scatter this location would receive from CT scans of the surrounding slices. In general, the CTDI is given by

$$\begin{aligned}\text{CTDI} = (\text{measured exposure}) \times (f - \text{factor}) \\ \times \frac{\text{chamber length}}{N \times \text{slice thickness}}\end{aligned} \tag{5}$$

The *f*-factor is the Roentgen exposure to the rad or gray dose conversion factor for the material being exposed. Alternatively, the ionization chamber measurement may be calibrated in air kerma or air dose and the appropriate air kerma to dose conversion factor for the material is used. The value N is the number of detector rows being used, and N times the slice thickness is the width of the X-ray beam. The ratio of chamber length to the beam width is one over the fraction of the chamber that is exposed with the primary beam.

The measurement of CTDI is a straightforward measurement with the proper equipment. The phantoms used to simulate the attenuation by the patient are typically acrylic cylinders with a diameter of 16 cm for the head phantom and a diameter of 32 cm for the body phantom. The phantom has a hole in the center and at four locations near the periphery of the phantom for insertion of the pencil shaped CTDI ionization chamber.

A composite measurement of the center and peripheral value is the weighted CTDI or CTDI_w. It is given by

$$\text{CTDI}_\text{w} = \left(\frac{1}{3} \times \text{CTDI}_\text{center}\right) + \left(\frac{2}{3} \times \text{CTDI}_\text{peripheral}\right) \tag{6}$$

These CTDI values should correspond to the multiple slice average dose from a series of contiguous axial scans. This value should also correspond to the dose from an equivalent helical scan with a collimator pitch of one. If the axial slices overlap or the helical pitch is less than one, the dose will be higher. If the axial slices have gaps between them or the helical pitch is > 1, then the average dose will be lower. The volume CTDI or CTDI_vol dose estimate adjusts the CTDI_w for the slice spacing or pitch.

$$\begin{aligned}\text{CTDI}_\text{vol} &= \text{CTDI}_w \\ &\quad \times \frac{\text{detector width} \times \text{number of detector rows}}{\text{table increment per } 360^\circ \text{ tube rotation}} \\ \text{CTDI}_\text{vol} &= \frac{\text{CTDI}_\text{vol}\text{CTDI}_w}{\text{pitch}(7)}\end{aligned} \tag{7}$$

A number of factors affect the patient dose. Some are defined by the design of the scanner, such as tube to patient distance, and the X-ray beam filtration. The patient size affects the attenuation and the subsequent dose for a given technique. Others are selectable by the technologist, (e.g., the kVp, mA, rotation speed, and pitch or slice increment). Since the X-ray output is directly proportional to the tube current or mA, then the total output per rotation, hence the dose, is directly proportional to the mA and the time per rotation. The dose is also related to the tube accelerating voltage or kV, but proportional to the kV to a power of ~ 3. Note that even though the dose goes up with kV, often some of the other dose factors can be reduced for comparable image quality, especially for large patients. The average patient dose is inversely proportional to the collimator pitch in helical (39). For comparable image quality, a thick-body section

requires a higher X-ray technique than does a thin-body section since a larger percentage of the incident X-rays are absorbed. Techniques are typically reduced for pediatric cases due to the smaller body size and higher concern for radiation exposure. Instead of selecting one technique to be used for all slices, the scanner may have a type of dose modulation or automatic exposure control that reduces the mA for less attenuating body sections. This may be performed based on data from the preview scan, or may be determined by the attenuation found in the previous rotation (40,41).

CT RECONSTRUCTION METHODS

There are several different reconstruction methods or mathematical algorithms that can be used to estimate the cross-sectional distribution of attenuation coefficients that results in the measured set of X-ray transmission values (5,6,39). Knowledge of the basic elements of the reconstruction method can help determine elements relating to the image quality, and artifacts. Reconstruction methods can be categorized into two basic approaches: analytic and iterative reconstruction techniques. The primary reconstruction method in medical CT systems is the filtered backprojection method, an analytic reconstruction technique.

Projection Data

What is the relationship between the measured transmission data and the CT data values? It is necessary to normalize the measured transmission data and convert these values into projection values that correspond to the objects attenuation values. The projection data values for a narrow, monoenergetic beam of X radiation can be determined by considering Lambert's law of absorption

$$I = I_0 e^{-\mu s}, \qquad \text{or} \tag{8}$$

$$I/I_0 = e^{-\mu s} \tag{9}$$

where I is the intensity of the transmitted beam, I_0 is the initial intensity or intensity of the beam with no attenuating material present, μ is the linear attenuation coefficient of the absorber material, and s is the thickness of the absorber. The linear attenuation coefficient corresponds to the fraction of the radiation beam that a thin absorber will absorb or scatter. This coefficient is dependent on the atomic number of materials present, the physical density, and the energy of the X-ray beam.

If instead of a single homogenous absorber there are a series of absorbers, each with thickness s, the overall transmitted intensity is

$$I/I_0 = e^{-\mu_1 s} \times e^{-\mu_2 s} \times e^{-\mu_3 s} \times e^{-\mu_4 s} \times \cdots \tag{10}$$

$$I/I_0 = e^{-(\mu_1+\mu_2+\mu_3+\mu_4+\cdots)s} \tag{11}$$

$$I/I_0 = e^{-\sum \mu_i s_i} \tag{12}$$

where μ_i is the linear attenuation coefficient of the ith absorber.

Considering a 2D section through an object of interest, the linear attenuation coefficients of the material distribution in this section can be represented by the function $\mu(x,y)$, where x and y are the Cartesian coordinates specifying the location within the section. The integral equivalent to the above equation is then

$$I/I_0 = e^{-\int \mu(x,y)ds} \tag{13}$$

integrated along the line, s, from the X-ray source to the detector.

The objective of the reconstruction program in a CT system is to determine the distribution of $\mu(x,y)$ from a series of intensity measurements through the section.

Inverting both sides of the equation to eliminate the negative sign, and taking the natural log of both sides to eliminate the exponential yields what is called the projection value, given by

$$p = \ln(I_0/I) = \int \mu(x,y)ds \tag{14}$$

This equation is the basis of the Radon transformation that is fundamental to the CT process. The inversion of this transform, going from the projection data to the 2D distribution was solved in 1917 by Radon (3). He showed that the distribution could be determined analytically from an infinite set of line integrals through the distribution.

The projection values are based on several assumptions that are not necessarily true in making practical measurements. This may require certain corrections to the data for these systemic errors, or may result in artifacts or degradations in the image. Some of these will be discussed relative to image quality and image artifacts.

Iterative Reconstruction Techniques

One of the broad categories of reconstruction is the iterative reconstruction techniques. With this approach an initial guess is made of the density distribution of the object. The computer then calculates the projection data values that would be measured for this assumed object in a process referred to as forward-projection. Each calculated value is compared to the corresponding measured projection data value, and the difference between these values is used to adjust the assumed density values along this ray path. This correction to the assumed distribution is applied successively for each measured ray. An iteration is completed when the image has been corrected along all measured rays, yielding an improved estimate of the object. The process is repeated and with each iteration the estimated object or reconstructed image improves its correspondence to the object distribution.

There are numerous variations of iterative processing that may be used. One of the most popular is the Algebraic Reconstruction Technique (7), or ART, which in itself is an offshoot of the Kaczmarz technique for inverting large ill-conditioned matrices (42). The variations include additive or multiplicative error correction, weighted or unweighted data, restricted or unrestricted values, the order in which one corrects the rays, and whether to apply the error corrections along a ray after each ray, or all at once for all rays.

Iterative techniques are rarely used in X-ray computed tomography. They require all data to be collected before

completion of even the first iteration, and they are very process intensive. Iterative techniques may be useful for selected situations where the data is limited or distorted, working with incomplete data sets, or with irregular data collection configurations. Known information on the object or object values may be incorporated into these techniques and reconstruction dependent corrections, such as for beam hardening, may be incorporated into the process. While not normally used for medical X-ray computed tomography, iterative techniques are commonly used in nuclear medicine SPECT and PET imaging. Variations may also be used in X-ray tomosynthesis that is a partial angle data acquisition technique used to generate planar images through an object, but without complete elimination of overlying structures (43).

Analytical Reconstruction Techniques

If one can analyze and solve a series of equations directly, it is an analytical technique. Radon in 1917 mathematically determined that a solution existed for determining the distribution of an object from a series of line integrals through it. Interesting though, is that Radon's work was not utilized in the development of CT, but it was noted afterwards that it encompassed the analytic reconstruction methods. Applied developments of the principles used were often driven outside of X-ray imaging, including radio astronomy (8), electron microscopy (9), and nuclear medicine (10), and discovered methods were often not implemented due to computational requirements in the precomputer age.

Direct Fourier Reconstruction

The Fourier transform is a mathematical operation that converts the object distribution defined in spatial coordinates into an equivalent distribution of sinusoidal amplitude and phase values in spatial frequency. The one-dimensional (1D) Fourier transformation of a set of projection data at a particular angle θ is given as

$$P(\rho, \theta) = \int p(r, \theta) e^{-2\pi\rho r} dr \qquad (15)$$

where r is the spatial position along the set of projection data and ρ is the corresponding spatial frequency variable.

The direct Fourier reconstruction technique, as well as the filtered backprojection method, are based on a mathematical relationship know as the central projection theorem or central slice theorem. This theorem states that the Fourier transform of a 1D projection through a 2D distribution is mathematically equivalent to the values along a radial line through the 2D distribution of the original distribution.

Taking the Fourier transform of one set of projection data measurements through an object at a particular angle provides data values along one spoke in the object's 2D Fourier transform frequency space. Repeating this process for a number of angles defines the 2D Fourier transform of the object distribution in polar coordinates (Fig. 13). Taking the inverse Fourier transform of this data yields the reconstructed image of the object.

Direct Fourier technique is potentially the fastest method for image reconstruction, however, it generally does not achieve the image quality of the filtered backprojection method due to data interpolation difficulties. Typical computer methods and display systems are based on rectangular grids rather than polar distributions, and direct Fourier reconstructions generally require an interpolation of the data from polar coordinates to a Cartesian grid, usually performed in the frequency domain. Consequently, these methods are generally not used in commercial medical scanners.

Convolutions and Filters

The filtered-backprojection technique is the most commonly used CT reconstruction algorithm. Before discussing this method, a brief review of filtering and simple backprojection methods is in order.

A convolution is a mathematical operation in which one function is smeared by another function. A common example is a presentation of a blurry out-of-focus projection of a text slide. A small dot, such as a period, instead of being small, sharp, and dark gets blurry with smooth edges and less contrast or darkness. This blurry spot is effectively the point spread function of the image. All of the lines and characters in the original slide can be considered as being made up of many points. Replacing each of these points with the

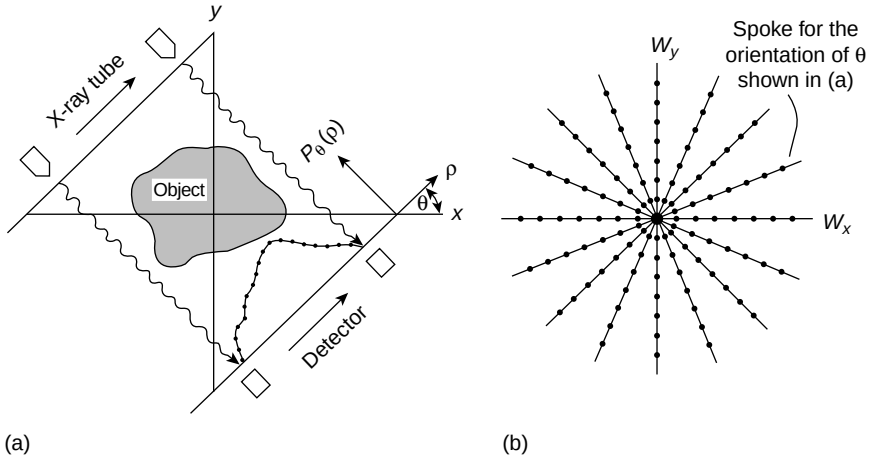


Figure 13. According to the central-slice theorem, the 1D Fourier transform of the projection values measured through an object in the spatial domain (a), correspond to the values of the 2D Fourier transform of the object along a diagonal line in the frequency domain. A series of transformed projection values yield the Fourier transform of the object, but in polar coordinates.

out-of-focus point results yields the overall blurry slide that is seen. The process of applying this blurry point to all points in the images is a convolution of the point spread function with the original object. Mathematically this is defined as

$$g(x) = f(x) \otimes h(x) \tag{16}$$

$$g(x) = \int_{-\infty}^{\infty} f(x)h(x-u)du \tag{17}$$

for 1D, or for 2D

$$g(x,y) = f(x,y) \otimes h(x,y) \tag{18}$$

$$g(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)h(x-u,y-v)dudv \tag{19}$$

where the symbol $\otimes$ is the convolution operator between two functional distributions. If the system is a digital system with a discrete number of samples, the corresponding equation is

$$g_i = \sum_{k=-\infty}^{\infty} f_i h_{i+k} \tag{20}$$

or in 2D

$$g_{i,j} = \sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} f_{i,j} h_{i+k,j+l} \tag{21}$$

In the blurry slide example above, f may represent the original text (or image) distribution, h is the blur or point spread function, and g is the resultant blurred image.

Convolution operations can be used in image processing to smooth an image, as previously described, or to sharpen an image. Smoothing convolution filters are typically square (averaging) or bell shaped, while sharpening convolution filters often have a positive central value with adjacent negative tails.

Convolution Theorem

Reference was made to Fourier transforms in Eq. 15 and their ability to transform spatial data into corresponding spatial frequency data. Filtering operations, such as smoothing and sharpening, can readily be performed on the data in the spatial frequency domain. According to the convolution theorem, convolution operations in the spatial domain correspond to a simple functional multiplication in the spatial frequency domain (Fig. 14). This states that

$$g(x) = f(x) \otimes h(x) \tag{22}$$

is equivalent to

$$G(k_x) = F(k_x)H(k_x) \tag{23}$$

where $G(k_x)$, $F(k_x)$, and $H(k_x)$ are the Fourier transformed functions of $g(x)$, $f(x)$, and $h(x)$, where k_x is the spatial frequency conjugate of x. The functional multiplication in Eq. 23 is simply the multiplication of values of $F(k_x)$ and $G(k_x)$ at all values of k_x. The 2D convolution of Eq. 18 has its counterpart to Eq. 23 where $G(k_x,k_y)$, $F(k_x,k_y)$, and $H(k_x,k_y)$ are the 2D Fourier transforms of $g(x,y)$, $f(x,y)$and $h(x,y)$. Note that the spatial frequency variables, k_x and k_y have units of 1/distance.

Since the convolution process is represented by a simple functional multiplication in the spatial frequency domain, the blurred image conceptually can be easily restored to the original object distribution. This can be accomplished by multiplying the Fourier transform of the blurred image, $G(k_x,k_y)$, by the inverse of the blurring function, that is $1/H(k_x,k_y)$. The result is the original object frequency

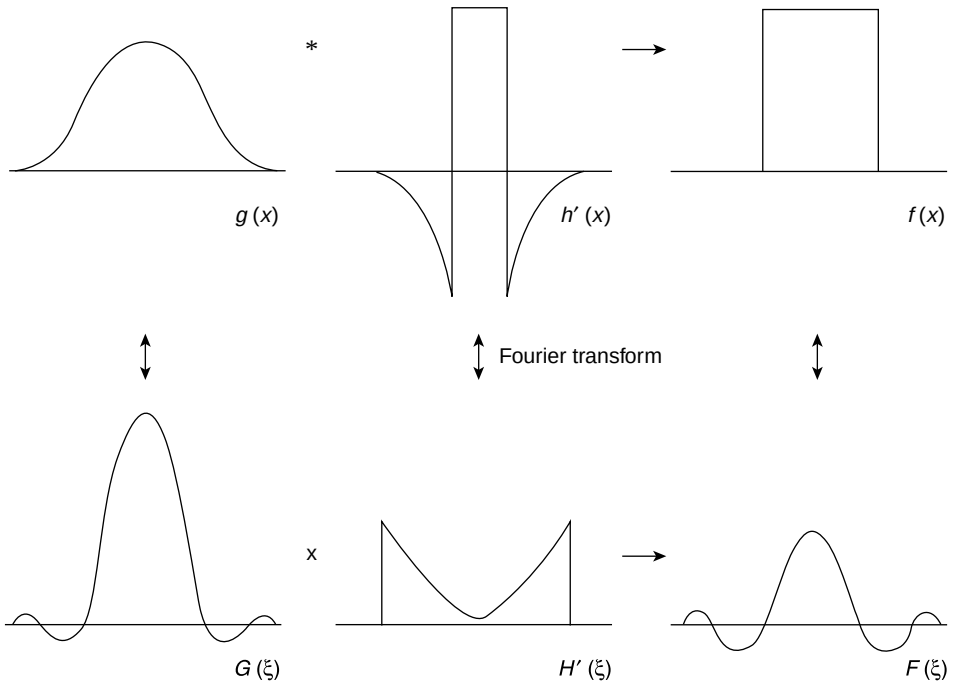


Figure 14. The convolution theorem states that a convolution operation in the spatial domain is equivalent to the functional multiplication of the Fourier transform of the functions in the spatial frequency domain. The filtering of the projection data for CT reconstruction may be performed by the convolving the projection data with a sharpening filter in the spatial domain, or by taking the Fourier transform of the projection data, multiplying by a ramp shaped filter, and taking the inverse Fourier transform to yield the filtered projection data.

distribution, $F(k_x,k_y)$. In practice this restoration is limited by the frequency limits of $H(k_x,k_y)$ leading to division by zero, and by the excessive enhancement of noise along with the signal at frequencies with small $H(k_x,k_y)$ values. This restoration process or deconvolution of the blurring function can likewise be performed as a convolution in the spatial domain.

Backprojection

Backprojection is the mathematical operation of mapping the 1D projection data back into a 2D grid. This is done intuitively by radiologists in interpreting X-ray films. If a high density object is visible in two or more radiographs taken at different angles, the radiologist mentally backprojects along the corresponding ray paths to determine the intersection of the rays within the patient and the location of the object.

Mathematically, this is done by taking each point on the 2D image grid and summing the corresponding projection value from each angular projection view. For that high density object the result is a line projected through the image from each view (Fig. 15a). This backprojection process yields a maximum density at the location of the object where the lines cross, but the lines form a star artifact emanating from the object. If an infinite number of views were used, the lines would merge and the density of the object would be smeared across the image with its amplitude decreasing with $1/r$ where r is the distance from the object. This simple backprojected image, f_b,can be represented by the convolution of the true image, f, with the blurring function $1/r$, or

$$f_b(r,\theta) = f(r,\theta)\otimes(1/r) \tag{24}$$

With ideal data, this blurring function can be removed by a 2D deconvolution or filtering of the blurred image. The appropriate filter function can be determined by using the convolution theorem to transform Eq. 24 into it frequency domain equivalent, or

$$F_b(\rho,\theta) = F(\rho,\theta)(1/\rho) \tag{25}$$

where the function $(1/\rho)$ is the Fourier transform of $(1/r)$ in polar coordinates. Dividing both sides by $(1/\rho)$ yields

$$F(\rho,\theta) = \rho F_b(\rho,\theta) \tag{26}$$

The corrected image, f, can be obtained by determining a simple backprojection, f_b, taking its Fourier transform, filtering with the ρ function, and taking the inverse Fourier transform. Likewise this operation may be performed as a 2D convolution operation in the spatial domain. Equation 25 and 26 get more complicated when evaluated in rectangular coordinates rather than polar coordinates.

This approach of making a very blurred image through backprojection, and then attempting to sharpen the image tends to produce poor results with actual data. Filtering out this blurring function from the projection data prior to backprojecting, however, is quite effective and is the basis for the filtered backprojection reconstruction technique used in medical CT systems.

Filtered Backprojection Reconstruction Technique

According to the central slice theorem the Fourier transform of the 1D projection data is equivalent to the radial values of the 2D distribution Fourier transform of the distribution. Consequently, the filtering operation performed in Eq. 26 and illustrated in Fig. 14 can be performed on the projection data prior to backprojection. This is the conceptual basis for the filtered-backprojection reconstruction technique as illustrated in Fig. 15b (44,45).

As with other filtering operations, this correction can be implemented as a convolution in the spatial domain or as a functional multiplication in the frequency domain. Fourier filtered backprojection is performed by taking the measured projection data, Fourier transforming it into the frequency domain, multiplying by the ramp-shaped ρ filter, taking the inverse Fourier transform, and then backprojecting this filtered projection data onto the 2D grid.

If the filtering is performed in the spatial domain by convoluting the measured projection data with a spatial filter that is equivalent to the inverse Fourier transform of ρ, the process is often referred to as the convolution

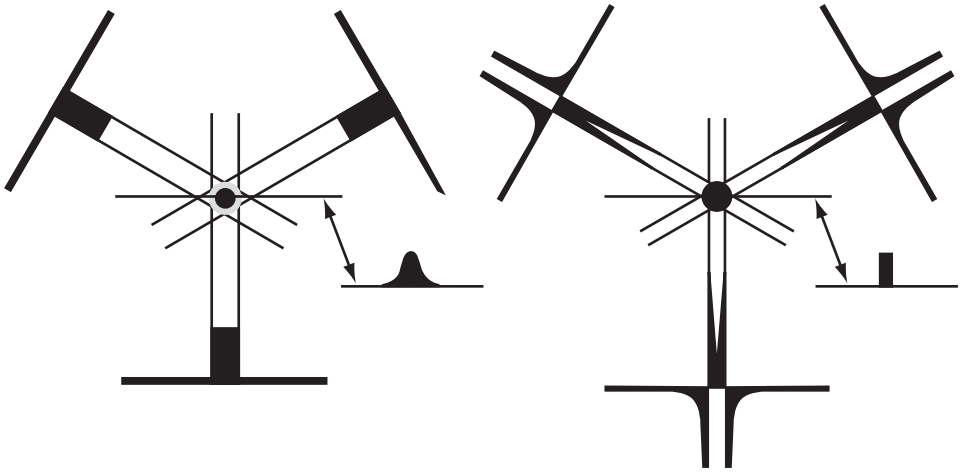


Figure 15. (a) A simple backprojection from three views results in a highly blurred reconstruction. (b) Filter the projection data prior to backprojection corrects the data for the backprojection induced blurring.

filtered backprojection reconstruction method. The frequency filter, ρ, has the shape of a ramp and enhances high spatial frequencies of the projection data. The convolution function, or kernel, has the expected shape of a sharpening filter with a positive central value surrounded by negative tails that diminish in magnitude with distance from the center. Convolution and Fourier filtering techniques are mathematically equivalent and the general term filtered backprojection reconstruction technique may refer to either filtering approach.

Variations from this ideal ramp filter are normally used. Some of these may develop from applying the finite quantity of data and boundary assumptions to the mathematical derivation. Other variations, in particular frequency windowing, are applied on a more empirical basis. The high spatial frequency component of the projection data contains noise variations due to photon statistic along with diminishing amounts of signal data. The ramp filter greatly enhances these high frequency values, in particular the noise. Consequently, use of the ramp filter results in high resolution, but very noisy images. They are also more susceptible to image artifacts. If one wishes to see structures with only small differences in attenuation values, such as white versus gray brain matter, the noise must be reduced.

Medical CT scanners offer a variety of reconstruction algorithms or kernels from which to choose. In actuality they are not changing the reconstruction method, but the filter function used in the filtered backprojection technique. The ramp filter is modified by a windowing or apodizing filter that reduces the amplification of the higher frequency values. This has the same effect as smoothing the image. This smoothing is especially effective for CT imaging since the reconstruction process results in the noise frequency spectrum in the image following the reconstruction filter function, with most of the noise at the high spatial frequencies. In nuclear medicine they sometimes use the mathematical name for the windowing filter, such as cosine filter, Butterworth filter, or Hannings window. In X-ray CT the equipment manufacturers utilize different naming conventions for these filters kernels. Typically, they will have descriptive names, such as smooth, standard, sharp, bone, edge, or will have numerical values relating to its shape. The filter selection may also enable other features in the reconstruction process to minimize certain artifacts, such as motion in the body scans, or implement other needed data corrections.

The filtered backprojection reconstruction technique is the general method used in medical CT systems. This method is more tolerant of measured data imperfections than some of the other analytical techniques. This method provides relatively fast reconstructions and permits processing the data as projection views are obtained.

Cone Beam CT

Computed tomography scanning has progressed from a single row of detectors to the ability to use hundreds of rows of detectors for data acquisition. The CT reconstruction algorithms discussed above have generally considered all of the projection values being contained within a single plane through the patient. As one uses multiple rows of detectors, the projection data from the rows away from the center row pass through the patient at an angle. Objects within a view from one side may be outside to the view from the opposite view angle (Fig. 16). Generally, this angulation is ignored in the reconstruction, but data inconsistencies from the angulated data can cause artifacts in the image, especially around high density structures. Other algorithms are continuing to be developed that take into account the actual ray paths of the data (4). These algorithms are referred to as cone beam reconstruction methods.

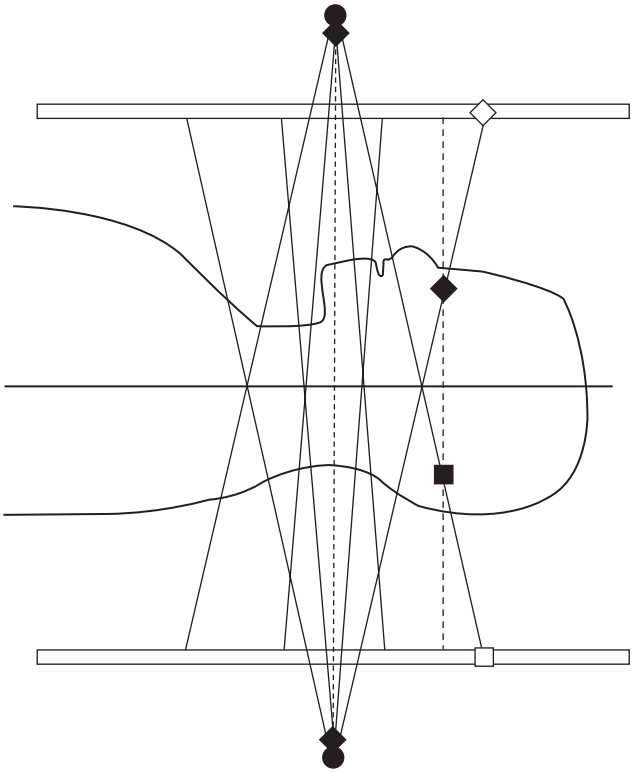


Figure 16. With a multislice detector or area detector, an axial (nonhelical) data acquisition results in rays outside of the center to pass through the patient at an angle rather than within a particular plane. This can cause artifacts in planar reconstructions. Specialized cone beam reconstruction algorithms may be used to help account for the angulated data.

With data acquired in an axial mode, with no table motion during data acquisition, large angulations cause significant degradation of the image quality, but may be useful in particular for high contrast structures such as bone or contrasted blood vessels. Considering the relationship between the object and the projection data, a simple rotation does not fully sample the Radon space needed for reconstruction. A helical data acquisition and certain other motion schemes do acquire a sufficient set of data for a potentially accurate cone beam reconstruction.

CT IMAGE QUALITY

A particular image will have limitations on the quality of the image and the types of structures that are visible. Two primary measures of image quality are the image resolution and noise (46). Resolution is the ability to see two separate small structures as two structures. It is somewhat different

that detection. A very small dense structure considerably smaller than the voxel size may change the average attenuation within the voxel sufficiently to detect that a foreign object is present, however, the system would not have sufficient resolution to distinguish multiple objects within the voxel.

Resolution may be measured by using a test pattern of holes that are spaced with the center-to-center distance between holes being equal to twice the hole diameter. Alternatively, the pattern may be a series of lines or bars, or a pie shaped wedge of bars that get smaller toward the apex of the pattern. The resolution is usually stated as the hole or bar size of the smallest pattern clearly visible, or as a spatial frequency value in line pairs per millimeter ($lp \cdot mm^{-1}$). The resolution may be more precisely described by determining the modulation transfer function or MTF. This function indicates the level of signal loss for each of the spatial frequencies. The MTF is the normalized Fourier transform of the point spread function, and can be calculated by analyzing the image of a small dense object, or from measurements across a sharp boundary (47).

Geometric Blurring

A number of factors can affect the resolution in an image, some inherent in the design and construction of the system, and others selectable by the user. Limits on resolution for a system are generally determined by the detector width and the focal spot size of the X-ray source. Since the patient is approximately midway between the source and detector, there is necessarily some magnification of the object structures onto the detector. As a result of this magnification a sharp edge in the object will produce a blurry edge in the image due to the size of the X-ray source. This blurry area where only part of the X-ray source illuminates the detector is called the penumbra from the Latin for partial shadow, and the effect is referred to as geometric blurring. This sharpness can be improved by reducing the magnification, which is done by having the X-ray source farther from the patient and the detector as close to the patient as possible. It can also be improved by use of a smaller X-ray focal spot. Most X-ray tubes utilize a dual focal spot, which means they have two focal spot sizes from which to choose. Most scanning is done with the large focal spot. This allows operation at higher tube currents or mA, thereby allowing shorter scan times and less image noise. This is especially true for large field-of-view objects where this geometric blurring is not the limiting factor for resolution. Scanners will typically select the smallest focal spot that can be used with the mA and time parameters selected.

Ray Spacing

The detector size may also affect the resolution. For a rotate–rotate scan system, the in-plane detector-to-detector spacing defines the ray spacing within a view. Note that this ray spacing is variable due to the fan shape of the beam, but consider it at the center of rotation, which is normally the center of the patient. The detector size has decreased with the evolution of the scanners, resulting in the need for an increase in the number of detector channels and the amount of data gathered and processed. Since the physical detector spacing cannot readily be changed for a given detector, a couple of alternative approaches have been used to change the effective ray spacing. An approach that has been used in the past maintains a fixed source–detector distance, but moves the X-ray tube closer to the patient, and the detector farther away when scanning smaller objects, such as the head. This increases the magnification and reduces the ray spacing, but puts more of a resolution burden on the size of the focal spot. No commercial systems still use this approach.

A common approach in rotate–rotate scanners is to improve sampling by using quarter–quarter offset detector shift. First generation translate–rotate scanners acquired data only $> 180^\circ$ of rotation since the view at 0° is of the same data as the 180° view. If the ray spacing is not symmetric about the center of rotation, but the middle ray being displaced one-quarter of a detector width above the center of rotation, then the opposite view would have the ray one-quarter of a detector width below the center of rotation and the rays would be interleaved. It is not as obvious on a rotate–rotate scanner, but the same type of interleaving can be utilized. A rotate–rotate scanner only needs 180° plus a fan angle to sufficient set of data for a reconstruction. Certain fast scan modes will utilize this partial rotation data set, but typically the full 360° data set is used. In high resolution reconstructions, the full 360° data set should be used.

Pixel Size

The CT image is a finite array of values. Parameters selectable setting up the image reconstruction include the reconstructed field of view (FOV), and the image matrix size. Images are routinely reconstructed with a 512×512 matrix. The large FOV that would be used for a large body section is ~ 50 cm. Therefore the generated pixels will be squares with sides of 50 cm/512 or ~ 1 mm. In this case, the image would not be effective at resolving or differentiating objects smaller than this pixel size. Medical CT systems have the capability to provide better resolution than can be displayed with 1 mm pixels. To reduce the pixel size the image matrix size must be increased, or the reconstructed FOV reduced. The reduction in FOV is standard for imaging smaller body sections, such as the head where a 25 cm FOV may be used. This yields a 0.5 mm pixel size. If higher resolutions are desired, such as to evaluate the bones of the inner ear, then a sharp reconstruction filter is needed along with an even smaller pixel size in order to maximize the system resolution. Unfortunately these parameter changes also increase the image noise.

Image Magnification and Targeted Reconstructions

A number of display tools are available when an image is displayed. These include such useful features as the window level and width adjustments. One tool is image magnification where a portion of the original image can be magnified to fill the display or a digital magnifying glass can be moved around the screen. This function uses the data from the image and interpolates additional pixels to yield smaller pixel spacing. Since the source data is the original image, the magnified image does not contain any

additional information, but the viewer may find it easier to see and interpret the image.

This image magnification is in contrast to a targeted reconstruction that goes back to the measured projection data and performs a reconstruction with new parameters such as reconstruction filter and pixel size. Consequently, a targeted reconstruction may yield information that was not present in the original image. In the case of helical scans and multislice detectors, the z-axis location of the reconstructed slices and the slice thickness, down to the acquisition detector row thickness, may also be specified.

z-Axis Resolution

The z-axis resolution is the resolution in the direction of the table motion or perpendicular to the axial image plane. The acquisition slice thickness and slice spacing dominate this resolution. The slice spacing limitation on z-axis resolution follows the same arguments as the pixel size does for the x–y resolution. The slice thickness is determined by the focal spot size and either the collimated beam thickness or the z-axis height of the detector. For multislice detectors, the detector rows may be utilized individually, or ganged together to yield a thicker slice, or both with the averaging of detector rows performed in the software. A major advantage to MDCT and cone beam CT is the ability to rapidly acquire many thin slices through the patient. This often results in isotropic resolution where the resolution in the z axis is equivalent to the in-plane resolution. These thinner slices and isotropic resolution improves the contrast of small high contrast structures, such as the blood vessels of the lung. The thin closely spaced slices greatly improve the quality of images that contain the z dimension, such as sagittal and coronal images, or 3D surface reformations.

Image Noise

Another significant limitation to image quality is image noise or graininess. Noise is caused by variations in the measured signal. These may be due to systemic causes, such as electronic noise, however, a well-designed system will not be limited by these sources. The primary source of noise in medical CT systems is due to quantum mottle or photon statistics. Due to the random nature of photon emission and absorption, repeated identical measurements will vary with a percent standard deviation proportional to one over the square root of the number of photons detected, that is

$$\sigma = \sqrt{N} \tag{27}$$

$$\%\sigma = 100\% \times \frac{\sqrt{N}}{N} \tag{28}$$

$$\%\sigma = 100\% \times \frac{1}{\sqrt{N}} \tag{29}$$

where σ is the standard deviation and N is the number of photons detected. The number of photons detected is determined by the output of the radiation source, the attenuation in the patient, and the efficiency of the detector in absorbing the radiation and converting it into a measurable signal. The source output is directly proportional to the selected tube current or mA, and the scan time per rotation and to the kVp to some power ~ 3. The beam is diminished by attenuation, hence noise is more prevalent scanning a large body section. Higher kVp settings are more penetrating and allow a larger fraction of the photons through the patient, but may also reduce the contrast of some structures.

Another factor is the size of the detector, in particular the detector height or the slice thickness. Thinner slices result in fewer X rays for a given technique. Helical scanning interpolates data from adjacent rotations, thereby increasing the effective slice thickness and reducing the noise to some degree. The image resolution and pixel size also can have an effect. One can consider the image noise to be related to the number of photons detected per voxel element. Reducing the voxel or resolution element size increases the noise.

Noise can be expressed as the standard deviation in the image of a uniform object. A more complete method of characterizing noise is to determine the noise power spectrum. The noise power spectrum defines the noise content in the image versus spatial frequency. The measure transmission data ideally contains white noise, that is, having a noise power spectrum at all frequencies. As with other imaging systems this ideal spectrum is modified by the modulation transfer function (MTF) or ability of the system to transfer signal (and noise) of various frequencies in the object to the image. The resolution limitations due to detector size and sampling reduces or eliminates some high frequency signal components. In CT, there is the additional step of the image reconstruction which modifies the noise power spectrum. The projection data is filtered with a ramp or a windowed ramp filter, reducing the low frequency content and enhancing the medium and high frequency content. For signal data where there is a correspondence in the data from various angles, the low frequency component is restored and a typical MTF response is generated. The noise content, however, is random and does not have a direct correspondence between angles. Consequently, the noise power spectrum will mimic the shape of the filtering function, enhancing noise at the higher spatial frequencies. Smoothing or the use of a windowed filter is especially effective in CT since there is a high level of noise relative to signal at these frequencies (48).

Object Contrast

Contrast is the difference in the measured value of a structure from its surroundings. The ability to distinguish structures that have attenuation differences of a fraction of a percent is one of the key imaging benefits of CT. The contrast between structures in CT is the fractional difference in attenuation coefficient, or more commonly the difference in attenuation coefficient relative to the attenuation coefficient of water. This is given by

$$\%\ \text{Contrast} = \frac{|\mu - \mu_{\text{background}}|}{\mu_{\text{water}}} \times 100\% \tag{30}$$

When CT values have an offset of 1000 in order to normalize water to a value of zero, the comparable equation using the CT values is

$$\% \text{ Contrast} = \frac{|\text{CT} - \text{CT}_{\text{background}}|}{1000} \times 100\%$$

For example, if brain gray matter has a CT number of 40 and white matter 30, then the fractional contrast is 10/1000 or 1%.

One of the factors that often limits contrast is the partial volume effect. Voxels or resolution volumes that correspond to a particular pixel may contain a mixture of tissues or structures. The resultant CT number is generally a volume average of the contents of the voxel. Reducing the voxel size with thinner slice thickness or higher image resolution may reduce this volume averaging. For example, a 1 mm piece of bone that is twice as attenuating as water (CT number of + 1000) is surrounded by water (CT number of 0). If scanned with a 10 mm slice thickness, the bone would occupy 10% of the voxel volume and the resultant image would show a CT number for this volume of ~ 100. Reducing the slice thickness to 2 mm would increase the CT number to $\sim$500. This is not an error or image artifact, but a limitation in the image quality and content. If structures are positioned in the voxel such that part of the ray goes through one material, and other portions of the ray go through other material, this may cause a measurement error and a partial volume artifact.

The linear attenuation coefficient values are a function of the density of the material and effective atomic number of the material, as well as the effective energy of the X-ray beam. Sometimes there is insufficient contrast between a structure of interest and its surroundings, or certain structures are to be highlighted. This may be done through the use of contrast agents or contrast media. This is a material that will change the X-ray absorption properties and the visibility of the structure. In X-ray studies the primary contrast agent used is iodine, which may be given orally to highlight the intestinal track, or injected to make blood vessels or highly vascular tissue visible. The iodine does not actually change the density of the blood significantly, but its higher atomic number of 56 versus 7 or 8 for tissue and water, and its k shell electron binding energy of 34 keV make it much more effective at stopping the X rays. Using contrast media the radiologist may determine the vascularity of a mass to help determine the type of tumor, or contrast agents can be used to enhance the blood vessels to evaluate blockages or aneurysms in CT angiography procedures.

Low Contrast Detectability

The ability to see small differences in contrast is a key feature of CT. What limits this ability, however, is noise. When the noise variations are of the same magnitude as the contrast between the structures of interest, the structure will not be visible. This is especially true for small objects, in that the human vision will effectively average the signal over an area and one may be able to distinguish larger low contrast objects. The limit to low contrast detectability is noise and resolution has relatively little impact (Fig. 17c–e). Conversely, high contrast resolution is not greatly affected by noise since the structures contrast is much larger than the noise.

ARTIFACTS

The ideal imaging system reproduces a faithful image of the object. Limitations in resolution, noise, and low contrast detectability are not errors, but are definable limitations of particular imaging systems. Errors do occur, however, and structures or patterns may appear in the image that do not correspond to the patient or object being scanned. These false structures in an image are referred to as artifacts (49). Artifacts often are readily identifiable because of their characteristics, such as a bright line extending through and beyond the boundaries of the patient. These artifacts may cause problems by obscuring parts of the image. Less common, but of significant concern, are artifacts that can appear similar to pathologies.

Artifacts occur in all imaging systems, but the reconstruction process of CT enhances the opportunity for producing artifacts. Mathematically a perfect reconstruction of the object should be obtainable. This can be done with an infinite amount of perfect data, however this is not available in a practical system. In general, image artifacts are caused by an insufficient amount of data, or insufficient quality of data.

Insufficient Data Quantity

Modern CT scanners acquire millions of transmission measurements and insufficient data quantity is not a limiting factor for routine studies, but still may pose challenges in studies attempting to reduce the data acquisition time or those pushing the image resolution. Nyquist sampling theorem was mentioned as a requirement for sampling an object in order to capture the high spatial frequencies. When the object contains higher frequencies than the sampling rate can characterize, the high frequencies reappear in the data and take on the alias of a lower spatial frequency. The resulting error is referred to as aliasing. It can take on the appearance of ripples in the reconstructed object parallel to a sharp edge.

Aliasing may occur due to insufficient sampling of rays within a view, or may be due to an insufficient number of views. Ray aliasing may appear as oscillations or ripples parallel to high contrast. View aliasing typically appears a considerable distance from a high contrast object, and appear as a radial pattern of light and dark bands emanating from the high contrast object.

The occurrence of aliasing is greatly reduced by the data smoothing effect that occurs due to the finite size of the X-ray detector. The transmission data is not measuring a series of infinitely thin rays, but columns through the patient. The variations in intensity between the various paths within a single ray or detector measurement get averaged out and lost, and are not available to cause aliasing.

A related artifact is the Gibb's phenomena. The range of the frequency domain is limited by the sample spacing. Using the full ramp function results in an abrupt cutoff of the filter function or transfer function at this limit, or the

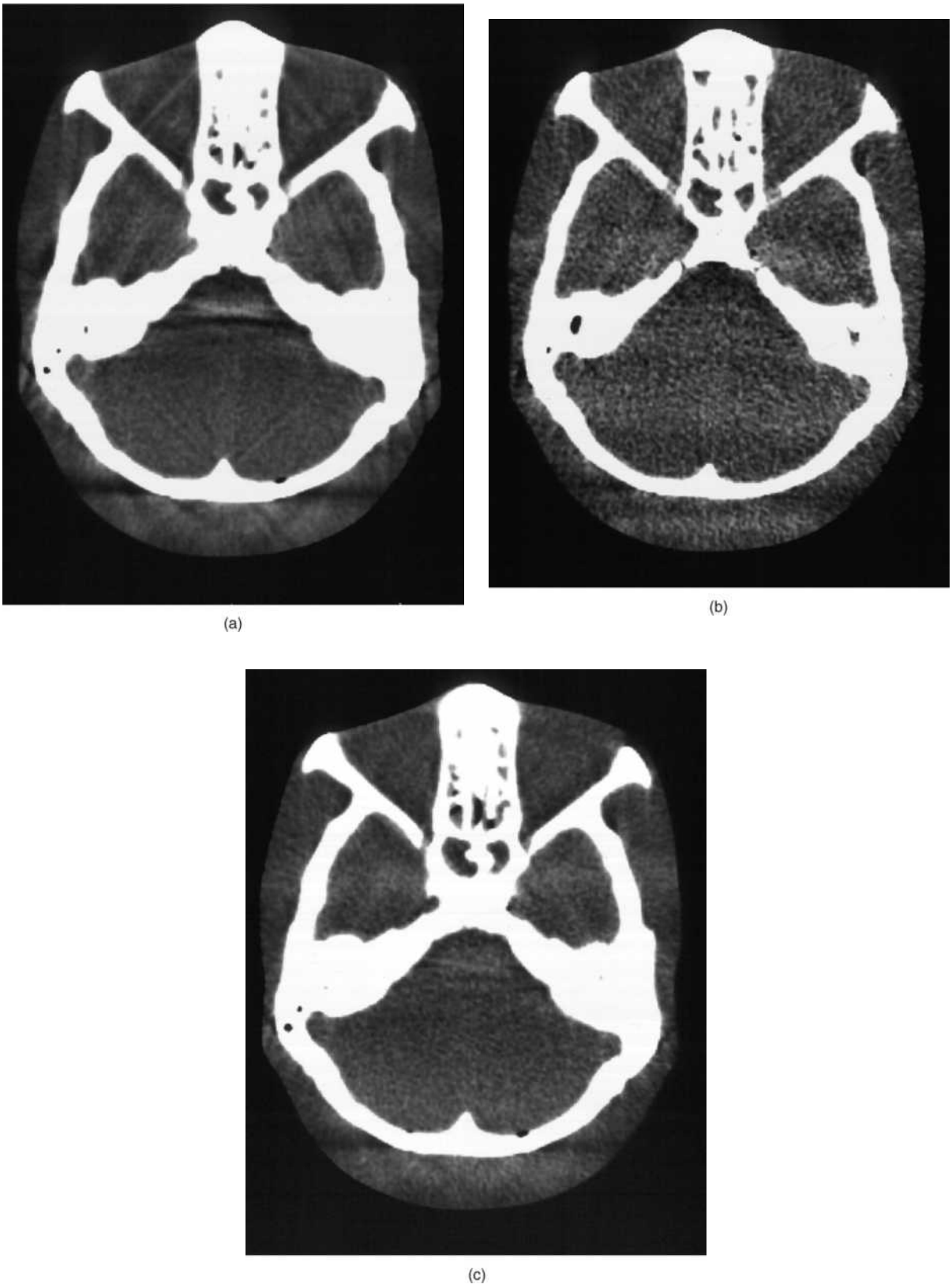


Figure 17. (a) A 5 mm thick CT slice through the posterior fossa can cause beam hardening and partial volume artifacts as seen in this acrylic phantom containing a skull. (b) Use of a thinner 1 mm slice thickness reduces the partial volume artifact from the angled bony structures, but increases the image noise or graininess. (c) Adding five 1 mm slices together (or the five sets of projection data after the logarithmic scaling) results in reduced artifact, but comparable image noise as the 5 mm slice.

filter may be designed with a sharp cutoff. This sharp edge in the frequency domain causes an overshoot and ringing artifact along edges in the image. This can be reduced by windowing the filter function with a function that smoothly goes to zero without an abrupt change.

Inconsistent Data

The CT reconstruction process is based on obtaining a consistent set of transmission views of the patient from various angles. The reconstruction algorithms effectively correct for errors caused in the backprojection from one view with corrections in other views. If the object or data is not consistent, then improper corrections are applied resulting in artifacts. There are several factors that can cause data inconsistencies and result in artifacts.

Motion

One of the most obvious inconsistencies is patient motion, whether squirming, coughing, fluid level movements in the stomach, breathing, or even the beating of the heart. Here the views from different angles are not of the same cross-sectional object. Motion artifacts are less of a problem in head imaging since with a cooperative patient the head can remain motionless for a considerable period of time. Body imaging can be more problematic causing image artifacts in addition to the motion blurring in the image.

The artifact is most pronounced when there is an abrupt change in the data from one view to the next usually resulting in streaks across the image. For a 360° data acquisition this abrupt change in the continuity of the data is likely to occur between the first and last view of the rotation. There are several ways to minimize this. One can overscan, collecting $> 360^\circ$ of data, and perform a weighted average of the data in the overlap region. Partial angle scanning, scanning 180° plus a fan angle, reduces this interface effect and reduces scan time, but also reduces image quality. The data may use a variable weighting of the data at the first–last view interface region reducing artifacts, but may increase noise some and result in the noise structure having a directional pattern toward this start–stop angle. This may be seen on some body scans but typically would not be used or present in head scans.

One of the obvious improvements in reducing motion artifacts is the use of shorter scan times. In early scanners it took tens of seconds to minutes to scan a single image, much less a volume through the patient. With MDCT and cone beam scanning, one can scan the entire chest within a breathhold. If one is looking at cardiac structures, however, the motion is rapid and less controllable. At times the motion may cause the image of a structure at one point in time overlaid on the image of the structure at a different point in time. This can cause artifacts that may mimic a pathology and special care is needed regarding this type of artifact. One example is imaging of the aorta as it changes diameter with the beating of the heart may create the image of a double vessel wall, which could look similar to a dissecting aneurysm (50). The electron beam CT systems were developed to provide very fast, freeze-action scans, and the use of cardiac gating to selectively collect data through the heart during defined portions of the cardiac cycle also reduce motion artifacts and blurring.

Partial Volume Artifact

Equation 12 in the discussion of the attenuation of X rays shows that a beam passing through a series of objects yields a relative transmission value that is the exponential of a sum of μs values. Taking the logarithm of this value yields the projection data that is the sum or integral of attenuation values along a line through the object. The X-ray beam that strikes a single detector is not a single infinitely slender beam, but a beam of some finite cross-sectional area. If part of the beam passes through one structure, and the other part of the beam passes through another structure, then the resulting transmission values corresponds to a sum of two or more exponentials, rather than an exponential of a sum. Taking the logarithm of this does not yield the same results. This can be seen by the example of a beam passing through a series of alternating dense and radiolucent layers will eventually get infinitesimally small. The same beam, however, having half of the beam transverse the dense material, and half the beam travel through the radiolucent material will always transmit at least 50% of the beam through the radiolucent path. This difference in attenuating material for different parts of a measured ray is commonly present along long linear boundaries. If the difference in density of the two materials is small, then there is little effect, but if there is a large difference in attenuation between the materials, such as for bone or a metal structure, then these inconsistencies result in streaks emanating as an extension of the edge (Fig. 17a).

This partial volume artifact is also produced by structures that penetrate only part way through the slice thickness, and may cause streaks between such partially penetrating structures. Partial volume streak artifacts may be reduced by reducing the slice thickness, which reduces the cross-sectional area of the measured ray. This can be especially useful in high resolution imaging of structures with bony prominences (Fig. 17b and c), or in the presence of dense metal structures (Fig. 18). The routine use of thinner detectors in MSCT systems, even when the detectors are averaged together as long as it is after the logarithmic scaling, reduces this artifact.

A related artifact is the windmill artifact due to insufficient sampling in the z axis in helical scanning. This presents as fan shaped lines emanating from edges. Increased sampling or lower pitch can reduce this artifact. Alternatively, view spacing in the axial direction may be reduced by using a flying focal spot approach that rapidly moves the position of the X-ray source in the X-ray tube. This uses an oscillating magnetic field to alter the path of the electrons and the location on the X-ray tube target where the electrons strike.

Beam Hardening Artifact

The attenuation parameters used in Eqs. 10–14 assumed a monoenergetic X-ray source yielding consistent values of μ for a given material. However, X-ray sources produce

Figure 18. Dense metal structures in the body can produce pronounced streaking artifacts. Partial volume edge artifacts, beam hardening, scatter, and motion can all contribute to inconsistencies in the data causing the artifacts.

radiation with a range of photon energies up to the maximum energy of the electrons striking the X-ray tube target. This maximum photon energy corresponds to the kV or kVp voltage applied to the tube. The lower energy or soft photons are attenuated to a greater degree than the more penetrating high energy photons. Consequently, the effective energy of a beam passing through a thick object section is higher than that of a beam going through less material. This preferential transmission of the higher energy photons and the resulting increase in effective energy of the X-ray beam is referred to as beam hardening.

The linear attenuation coefficient is a function of the beam energy and changes in beam energy cause variations in the CT number for a material. Likewise, if rays passing through an object from different angles have different effective energies, then the inconsistencies can also cause artifacts. X-rays that pass through the center of a cylindrical object will have higher effective energy than those traversing through the edges. This additional attenuation and beam hardening of the central rays result in lower CT number values in the center, corresponding to the higher energy beam. This appearance of lower CT numbers in the center of an object is referred to as a cupping artifact.

Another type of beam hardening artifact occurs between two dense structures. The rays passing through two dense structures have increased beam hardening, and tissues between these structures will appear to have a lower CT value. This appears as a dark band between the structures, and may be present along with partial volume artifacts appearing as fine streaking from edges. This is often seen in head scans as dark bands between the petrous ridges (Fig. 17a).

The effect of beam hardening can be reduced by several techniques. The original EMI scanner used a constant-length water bath that resulted in a relatively uniform degree of attenuation from the center to the periphery. The addition of X-ray beam filtration reduces the soft X rays and reduces the degree of beam hardening (51). Bow-tie shaped compensating X-ray filters are often used and attenuate the beam more toward the periphery. Normalizing the data with a cylindrical object of similar size and material also reduces beam hardening as well as minimizing detector variation errors.

Beam hardening is also compensated for in the processing software. If the material being scanned is known, the measured transmission value can be empirically corrected. Difficulty occurs when the object consists of multiple materials with different effective atomic numbers, such as the presence of bone or metal in the tissue. Iterative beam hardening corrections can be used, where the initial reconstruction identifies the dense structures, and the rays through these structures are corrected for a second reconstruction. Dual energy scanning techniques can also eliminate beam-hardening effects.

Scatter Radiation

Detected scatter radiation produces a false detected signal that does not correspond to the transmitted intensity along the measured ray. Scatter is a factor in all radiographic measurements. The amount of scatter radiation detected in CT is much lower than encountered with large area radiographs due to the thin fan beam normally used in CT. Most of the scatter is directed outside of the fan beam and is not detected. The sensitivity of computed tomography, and the need for consistent data makes even the low level of scatter detected a potential problem. The amount of scatter and this problem becomes more challenging as the collimated beam width gets larger with MDCT and cone beam systems.

The scatter contribution across the detector array tends to be a slowly varying additive signal. The effect on the measured data is most significant for highly attenuated rays where the scatter signal is relatively large compared to the primary signal. The additional scattered photons detected make the materials along the measured ray appear less attenuating, in a manner similar to beam hardening. Because of the similarity of these effects, scattering artifacts are similar to and are often associated with beam hardening. Likewise, the basic beam hardening correction performed provides some degree of compensation for scatter. Using thin beams and large distances between the patient and detector can minimize scatter. The use of directional dependent detectors such as the xenon gas ionization detectors with their focused tungsten plates can also reduce the detection of scatter. As detectors get larger with cone beam CT applications, scatter is significant and antiscatter radiographic grids can be used to reduce the detected scatter. Since scatter varies slowly with distance, special reference detectors outside of the primary radiation beam can be used to measure the level of the scatter signal for more effective correction.

Cone Beam or Divergence Errors

The X-ray fan beam does not only diverge or fan out in the x–y plane, perpendicular to the rotational axis, but also a slight divergence in the z direction, across the detector row (Fig. 16). A high density structure toward the edges of the patient may be in the x-ray beam from one angle, but be missing the beam from the opposite angle. This inconsistency causes an artifact that may include diffuse streaking or smearing of the density from the edges of the object. This may also occur in a helical data acquisition that interpolates the data to produce a data set corresponding to a given slice. Alternative variations in the reconstruction process that utilize a cone beam reconstruction approach or backproject along the actual ray paths reduce this effect.

Note that a single axial rotation around the patient with a widely diverging beam with an area detector does not contain a complete set of projection data sufficient to perform a reconstruction. Mathematically, it does not fully sample the Radon space. Reconstructions can be performed with various means to estimate the missing data, with the difficulty increasing as the divergence perpendicular to the plane of rotation gets larger.

Other Systemic Errors

There are a lot of things that can happen to the signal between the X-ray source and the image. Computed tomography scanners are complex electromechanical devices that are sensitive to relatively small variations in the measurements. A number of factors can cause accuracy or inconsistency errors and result in artifacts, and considerable effort is taken by the manufacturers to minimize these problems and artifacts. These inconsistencies may be in the measured signal or may be the result of poor characterization of the spatial position or the measured ray path. Geometric errors can result in wobble of the center of rotation due to the mechanical limitations of the large turret bearing holding the rotating mechanism, or may be due to small changes in the focal spot position as the X-ray tube gets hot, or variations in the spacing or position of the individual detector elements.

Signal variations occur due to fluctuations in the X-ray tube output, and differences in response characteristics among the individual detector elements. Periodic calibration of the detectors is required and radiation sensors are used to monitor variations in the X-ray tube output. Typically the reference detectors are at the ends of the linear detector array. A large patient, or body part outside of the normal scan field of view can block the reference detector, causing an inaccurate normalization of the measurements. Alternatively, an X-ray sensor may be placed adjacent to the X-ray tube where the measurement cannot be blocked. Vibrations in components can also affect the readings, as well as a number of other factors.

Detector calibration and characterization is particularly important with rotate-rotate data acquisition systems. The ray paths for a single detector all are tangential to a circle around the center of rotation. A relatively small consistent error in a detector can accumulate to form a ring artifact in the image (Fig. 19). The commercial X-ray CT systems are designed to be as stable and consistent as practical, however characterization of the various components and software correction of the measured transmission data is an important step in clinical CT imaging.

Figure 19. Ring artifacts centered on the scanner's center of rotation can be caused by the miscalibration or error in a single radiation detector on a rotate–rotate data acquisition CT scanner.

BIBLIOGRAPHY

1. Hounsfield GN. Computed medical imaging. J Computer Assist Tomogr 1980;5:665–674.
2. Brooks R, DiChiro G. Principles of computer assisted tomography (CAT) in radiographic and radioisotopic imaging. Phys Med Biol 1976;21:689–732.
3. Radon J. Uber die bestimmung von funktionen durch ihre Integralwerte langs gewisser Mannigfaltigkeiten. Ber Verhandlung 1917;69:262–277.
4. Feldkamp LA, Davis LL, Kress JW. Practical cone-beam algorithm. J Opt Soc Am 1984;1:612–619.
5. Herman GT. Image Reconstruction from Projections: Implementation and Applications. New York: Springer-Verlag; 1979.
6. Parker JA. Image Reconstruction in Radiology. Boca Raton: CRC Press; 1990.
7. Gordon R, Bender R, Herman GT. Algebraic reconstructioni techniques (ART) for three-dimensional electron microscopy and x-ray photography. J Theor Biol 1970;29:471–481.

8. Bracewell RN. Strip integration in radio astronomy. Aust J Phys 1956;9:198–217.
9. DeRosier DJ, Klug A. Reconstruction of three-dimensional structures from electron micrographs. Nature London 1968; 217:130–134.
10. Kuhl DE, Edwards RQ. Image separation radioisotope scanning. Radiology 1963;80:653–662.
11. Cormack AM. Representation of a function by its line integrals with some radiological applications. J Appl Phys 1963;34:2722–2727.
12. Hounsfield GN. Computerized transverse axial scanning tomography: Part I: Description of system. Br J Radiol 1973;46:1016–1022.
13. Dennis MJ. Industrial Computed Tomography. Metals Handbook. Metals Park, (OH): ASM International; 1989. p 358–386.
14. Boyd DB, Lipton MJ. Cardiac computed tomography. Proc IEEE 1983;71:298–307.
15. Ritman EL, Robb RA, Harris LD. Imaging physiological functions: Experience with the dynamic spatial reconstructor Philadelphia: Praeger; 1985.
16. Mori I. Computerized tomographic apparatus utilizing a radiation source. US Patent 4,630,202. 1986.
17. Nishimura H, Miyazaki O. CT system for specially scanning subject on a moveable bed synchronized to x-ray tube revolution. US Patent 4,789,929. 1988.
18. Kalender WA, Klotz W, Vock E. Spiral volumetric CT with single breath-hold technique, continuous transport, and continuous scanner rotation. Radiology 1990;176:181–183.
19. Kalender WA. Computed Tomography: Fundamentals, System Technology, Image Quality, Applications. Munich: Publicis MCD Verlag; 2000.
20. Hu H. Multi-slice helical CT: Scan and reconstruction. Med Phys 1999;26(1):5–18.
21. Fishman EK, Jeffrey RB, editors. Multidetector CT: Principles, techniques, and clinical applications. Philadelphia: Lippincott, Williams & Wilkins; 2003. p 560.
22. Yester MW, Barnes GT. Geometrical limitations of computed tomography (CT) scanner resolution. Appl Opt Instr Med VI Proc SPIE 1977;127:296–303.
23. IEC. International Electrotechnical Commission: Medical electrical equipment–60601 Part 2-44: Particular requirements for the safety of X-ray equipment for computed tomography. Geneva, Switzerland: 1999.
24. Kachelriess M, Kalender WA. ECG-correlated image reconstruction from subsecond spiral CT scans of the heart. Med Phys 1998;25(12):2417–2431.
25. Cline HE, et al. Two algorithms for the three-dimensional reconstruction of tomograms. Med Phys 1988;15(3):320–327.
26. Schoepf UJ, et al. Multislice CT angiography. Eur Radiol 2003;13(8):1946–1961.
27. de Feyter PJ, Kresin GP, editors. Computed Tomography of the Coronary Arteries. New York: Taylor & Francis Group; 2004. p 208.
28. Vrtiska TJ, Fletcher JG, McCollough CH. State-of-the-art imaging with 64-channel multidetector CT angiography. Perspect Vasc Surg Endovasc Ther 2005;17(1):3–10.
29. Ulzheimer S, Kalender WA. Assessment of calcium scoring performance in cardiac computed tomography. Eur Radiol 2003;13(3):484–497.
30. Kalender WA, Klotz W, Suss C. Vertebral bone mineral analysis: an integrated approach with CT. Radiology 1987;164: 419–423.
31. Lang TF, et al. Assessment of vertebral bone mineral density using volumetric quantitative CT. J Computer Assisted Tomogr 1999;23(1):130–137.
32. Braillon PM. Quantitative computed tomography precision and accuracy for long-term follow-up of bone mineral density measurements: a five year in vitro assessment. J Clin Densitom 2002;5(3):259–266.
33. Alvarez RE, Macovski A. Energy selective reconstructions in x-ray computerized tomography. Phys Med Biol 1976;21: 733–744.
34. Vogel WV, et al. PET/CT: Panacea, redundancy, or something in between? J Nucl Med 2004;45(Suppl 1): 15S–24S.
35. Bockisch A, et al. Positron emission tomography/computed tomography—imaging protocols, artifacts and pitfalls. Mol Im Biol 2004;6(4):188–199.
36. Limitation of Exposure to Ionizing Radiation. NCRP Report No. 91. National Council on Radiation Protection; 1993.
37. Morin RL, Gerber TC, McCollough CH. Physics and dosimetry in computed tomography. Cardiol Clinics 2003;21(4):515–520.
38. McCollough CH, Schueler BA. Calculation of effective dose. Med Phys 2000;27(5):828–837.
39. Barrett HH, Swindell W. Radiological Imaging: The Theory of Image Formation, Detection, and Processing. New York: Academic Press; 1981.
40. Greess H, et al. Dose reduction in computed tomography by attenuation based on-line modulation of tube current: Evaluation of six anatomical regions. Eur Radiol 2000;10(2):391–394.
41. Kalender WA, et al. Dose reduction in CT by on-line tube current control: principles and validation on phantoms and cadavers. Eur Radiol 1999;9(2):323–328.
42. Kaczmarz S. Angenaherte Auflosung von Systemen linearer Gleichungen. Bull Acad Polonaise Sci Lett 1937;A35:355–357.
43. Grand DG. Tomosynthesis: A three-dimensional radiographic imaging technique. IEEE Trans Biomed Eng 1972;BME-19(1): 20–28.
44. Ramachandran GN, Lakshminarayanan. Three-dimensional reconstruction from radiographs and electron micrographs: III. Description and application of the convolution method. Indian J Pure Appl Phys 1971;9:997.
45. Shepp LA, Logan BF. The Fourier reconstruction of a head section. Trans IEEE 1974;NS-21:21–43.
46. McCollough CH, et al. The phantom portion of the American College of Radiology (ACR) Computed Tomography (CT) accreditation program: Practical tips, artifact examples, and pitfalls to avoid. Med Phys 2004;31(9):2423–2442.
47. Blumenfeld SM, Glover G. Spatial resolution in computed tomography. In: Newton TH, Potts DG, editors. Radiology of the Skull and Brain, Vol 5, Technical Aspects of Computed Tomography. New York: C.V. Mosby Company; 1981.
48. Joseph PM. Image noise and smoothing in computed tomography (CT) scanners. Opt Eng 1978;17:396–399.
49. Joseph PM. Artifacts in computed tomography. Phys Med Biol 1978;23:1176–1182.
50. Parry CK, Rajagopalan B. Characterization of artifact simulating aortic dissection in computed tomography imaging. J Digital Imaging 2001;14(2 Suppl. 1):220–221.
51. McDavid WD, et al. Spectral effects on three-dimensional reconstruction from x-rays. Med Phys 1975;2(6):321–324.

See also IONIZING RADIATION, BIOLOGICAL EFFECTS OF; MAGNETIC RESONANCE IMAGING; ULTRASONIC IMAGING.

COMPUTED TOMOGRAPHY SCREENING

DAVID J. BRENNER
Columbia University Medical Center
New York, New York

INTRODUCTION

Computed tomography (CT), developed by Hounsfield and colleagues in the 1970s (1,2) has revolutionized much of

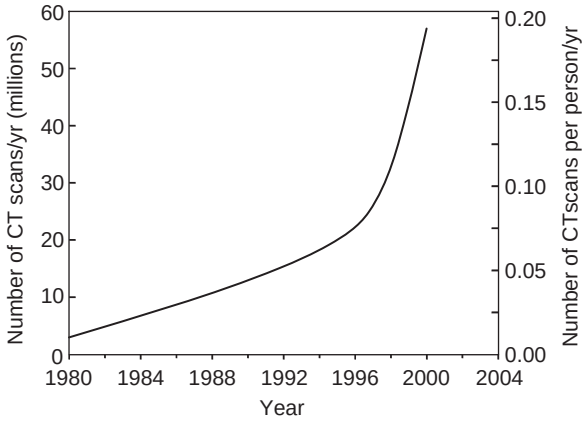


Figure 1. Graph shows the increase in the estimated number of CT scans performed in the United States between 1980 and 2000. (Based on data in Refs. 86–89.)

medical imaging by allowing a three dimensional (3D) view of the organ or part of the body of interest. Since its inception, the use of CT has increased very rapidly, both in the United States and other countries. At present, ~60 million CT scans are being performed each year in the United States. As seen in Fig. 1, this increase has occurred roughly over 20 years. It has largely been driven by the major technical advances in CT technology, in particular the development of helical multidetector CT scanners, as discussed below, which allow CT scans to be made in 1 s or less.

The basic principle of helical, or spiral, CT scanning is shown in Fig. 2. Essentially, the patient is moved through a continuously rotating X-ray source–detector combination. A more modern version is the multidetector CT, which gives the advantage of short scan times, coupled with potentially very thin slice widths. The result is a series of many images of "slices" of the organ or part of the body of interest, which can then be combined by computer-based mathematical techniques, to provide 3D views.

The use of CT for mass screening is a more recent innovation, driven in part by the increased availability and convenience of CT scans. Three applications, each of which will be discussed, have been suggested for CT-based screening: for early stage lung cancer in smokers and exsmokers; for lesions in the colon (virtual colonoscopy); for general screening for many diseases in the whole body (full-body screening).

All three modalities, as of 2005, are quite new, and a general consensus has not yet been reached about the efficacy of any of them. The general issues regarding efficacy of these new modalities are, in essence, the same as for all other potential mass screening modalities, for example, mammography, pap smear screening, and colonoscopy. However, as discussed, there is an added issue for CT-based screening modalities, namely, the significant X-ray radiation exposures involved.

The more general issues of screening relate to (a) whether the screening modality truly produces a stage shift (i.e., allows detection of more early-stage cancers and less late-stage cancers), (b) whether the screening

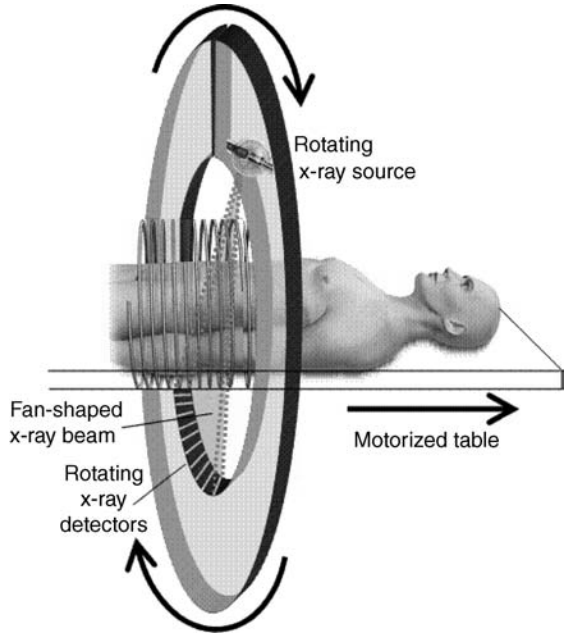


Figure 2. Schematic of helical (spiral) CT scanning. Both the X-ray source and, on the other side of the patient, the X-ray detectors, rotate around the patient. If the table were not moving, a single slice of the patient would be imaged (axial CT). Because the table is moving at the same time as the source–detector combination is rotating, the result is a helical or spiral CT scan of the patient, as depicted here. Shown in this schematic is a single row of detectors; modern multidetector scanners have several rows of detectors alongside each other, which allow both for thinner slice widths, and shorter scan times.

modality produces overdiagnosis (identifying lesions that the individual would die with, rather than die of); and (c) false positives, the possibility of mistakenly identifying a tumor, with the attendant possibility of subsequent unnecessary procedures. All these issues feed in to the general question of whether the overall mortality rate from the disease in question is significantly reduced by the screening test.

By contrast, the radiation exposure issues that relate to CT-based mass screening are unique. It is, of course, true that mammography also involves the use of X rays, but, as we will discuss, the radiation doses involved are generally much higher for CT-based screening compared to mammography. Thus the potential benefits of any CT-based screening procedure must, in addition to the more general efficacy issues discussed above, have to significantly outweigh any potential harm from repeated low dose X-ray exposures. In the next section, what is currently known about the hazards of low doses of X rays is reviewed.

CANCER RISKS ASSOCIATED WITH EXPOSURE TO LOW X-RAY DOSES

Some typical radiation doses associated with common radiological examinations are shown in Table 1. The biological effects of low dose X-ray exposures have been investigated and debated for more than a century (3). There is little

Table 1. Typical Organ Doses from Various Radiological Examinations[a]

Examination	Relevant Organ	Relevant Organ Dose, mSv
Dental X ray	Brain	0.005
PA Chest X ray	Lung	0.01
Lateral chest X ray	Lung	0.15
Screening mammogram	Breast	3
Adult abdominal CT	Stomach	10
Neonate abdominal CT	Stomach	25

[a]Radiation dose, a measure of ionizing energy absorbed per unit mass, has units of Gy (gray) or mGy ($1\,Gy = 1\,J{\cdot}kg^{-1}$); it is often quoted as an equivalent dose, in units of Sv (Sievert) or mSv. For X rays, which are the radiations produced in CT scanners, 1 mSv = 1 mGy.

question that intermediate and high doses of ionizing radiation, say above 100 mSv, given acutely or over a prolonged period, produce deleterious effects in humans, the most significant being cancer induction (4). At lower doses, however, the situation is less clear. Compared to higher doses, the risks associated with low doses of radiation are lower, and progressively larger epidemiological studies are required to quantify the cancer risk to a useful level of precision. The reason is, as the dose goes down, the signal (radiation risk) to noise (natural background risk) ratio decreases.

Most of the quantitative information that we have regarding radiation-induced cancer risks comes from studies of A-bomb survivors. A-bomb survivor cohorts are generally used as the basis for predicting radiation-related risks to a general population because (a) they are the most thoroughly studied (over many decades) large exposed population; (b) the cohorts are not selected for disease; (c) all age groups are covered; and (d) a substantial subcohort of ~25,000 survivors, typically those who were ~2–3 km from the explosion hypocenters (5), received radiation doses comparable to those of concern here.

Key questions here are as follows: What is the lowest dose of X rays for which there is convincing evidence of significantly elevated cancer risks in humans? What is the most appropriate way to extrapolate these risks to still lower doses? What is the dependence of cancer risks on age at exposure? These issues have recently been extensively reviewed (3).

Effects of Radiation Dose on Cancer Risk

In summary, there is good epidemiological evidence of increased cancer risk for children exposed to an acute dose of 10 mSv (or higher), and for adults exposed to acute doses of 50 mSv (or higher) (3). As discussed below, relevant organ doses for CT exams are of the order of 15 mSv or less.

Extrapolation of Risks to Lower Radiation Doses

The issue here is how to estimate risks at doses somewhat (though not a great deal) lower than those for which there is statistically significant evidence of increased cancer risks. The current consensus (6) is that the measured risks can reasonably be linearly extrapolated to somewhat lower doses, though as the dose of interest becomes progressively lower, the uncertainties inherent in this extrapolation become progressively greater. Relatively small extrapolations from epidemiological data are required (e.g., from 50 down to 15 mSv), however, to estimate cancer risks at the doses relevant to single CT examinations.

Effect of Dose Fractionation

If individuals receive multiple CT screenings over a period of years, the radiation dose will, of course, increase proportionately. The most likely case is that any radiation risks will also increase proportionately. Specifically, at high doses, theory (7), animal data (8), and epidemiological data (9), suggest that fractionating a radiation exposure decreases the overall risk at a given dose, but at the low doses of relevance here, both theory (7) and animal data (8) suggest that the risks are roughly independent of fractionation.

Effect of Age at Exposure

Regarding age at exposure, as can be seen in Fig. 3, radiation risks generally decrease markedly with age. The reason is because (a) sensitivity is related to the proportion of dividing cells in an organ, which decreases with increasing age; and (b) other competing risks play an increasing role with increasing age.

RADIATION DOSES FROM SCREENING CT

The radiation dose from CT scans depends on a number of factors: The most important are the tube current; the scan time; the pitch [For helical CT scans, the speed that the patient table moves relative to the rotation speed of the X-ray tubes–detectors will be an important determinant of the radiation dose; it is defined through the pitch, which is the linear table motion feed per 360° rotation, divided by the total beam width (the slice width × the number of detectors).]; the tube voltage; the number of detectors; and the particular scanner design (10). For a given CT scanner operating at a given voltage, the organ dose is

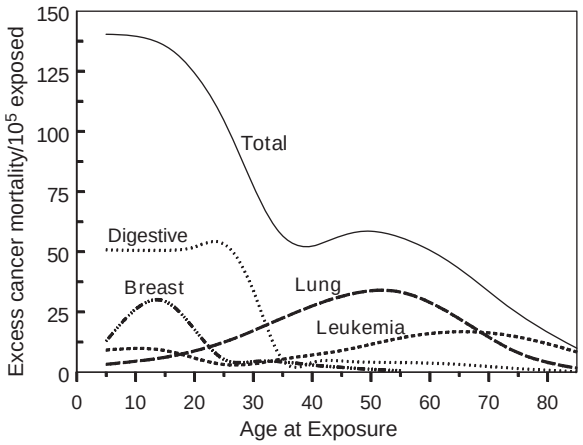


Figure 3. Estimated radiation-related absolute cancer mortality risk per 10^5 individuals in the United States exposed at different ages to a whole-body dose of 10 mSv (63).

proportional to the mAs [current (mA) × rotation time], and is inversely proportional to the pitch. It is always the case, however, that the relative noise in CT images will increase as the radiation dose decreases; thus there will always be a tradeoff between the need for low noise images and the desirability of using low radiation doses (11). As discussed later, the amount of noise that can be tolerated depends very much on the application. Thus, for example, more noise (i.e., lower doses) can probably be tolerated for virtual colonoscopy because of the radiological contrast of colonic polyps projecting into an air filled lumen), compared with whole-body screening (where most of the potential lesions show less radiologic contrast).

A relatively new and very promising radiation dose reduction technique for CT is automatic tube current modulation (Fig. 4) (12–15), now available from all the major scanner manufacturers: these systems continuously lower or raise the X-ray tube current to compensate for different instantaneous levels of attenuation of the X-ray beam by the patient. For example, when the beam is aimed in the posterior–anterior direction, fewer X rays are needed (for the same image quality) compared to the lateral–medial direction; or when the beam is passing through the region of the transverse colon, fewer X rays are needed compared to the pelvic bone region.

COMPUTED TOMOGRAPHY COLONOGRAPHY (VIRTUAL COLONOSCOPY)

There is no doubt (a) that colonoscopy-driven polypectomy can result in a significantly decreased incidence of colorectal cancer (16,17), and (b) that there is suboptimal compliance with current guidelines for colorectal cancer screening (18,19). Screening using CT colonography, often referred to as "virtual colonoscopy" (VC), was first suggested in 1983 (20), but has only recently become a potential option for mass screening (21–23).

In the most common current usage of VC, after bowel preparation, the colon is inflated with air or CO_2, and the colon is CT scanned. The resulting data can then be analyzed for polyps, based on two-dimensional (2D) images, or using a 3D endoluminal view. Virtual colonoscopy is an excellent application of CT because of the radiological contrast exhibited by colonic polyps projecting into a gas-filled lumen (20,21,24), and a National CT colonography trial is underway in the United States.

Virtual colonoscopy may well have the potential to increase colorectal cancer screening compliance, largely because of the possibility that it can be performed with noncathartic preexamination bowel preparation. Current compliance with screening guidelines is clearly poor: At most, about one-third of adults >50 in the United States have had an endoscopic examination within the past 10 years (18,19).

From a technological perspective, VC is not quite ready for use in mass-screening programs. The three main outstanding issues, all of which seem relatively close to solution, are as follows:

1. The sensitivity and specificity of VC for detecting lesions in the size range from 5 to 10 mm: VC sensitivity and specificity for lesions >10 mm in diameter are generally well over 90% (about as good as those for conventional optical colonoscopy) (25). There is evidence that a well-designed VC screening program can achieve at least 90% sensitivity and specificity in the size category from 7 to 10 mm (22,26), but not all studies have achieved this (27).
2. The use of noncathartic preexamination bowel preparation regimens: In general it may be less the invasive nature of conventional colonoscopy that results in poor compliance, but more the necessity for cathartic bowel preparation (28–32). Virtual colonoscopy offers the potential for noncathartic bowel preparation, through the use of barium or iodinated tagging agents, which impart a high density to both stool and residual fluid, allowing increased contrast with soft-tissue polyps. Recent results with noncathartic VC have been very encouraging (23,33–35).

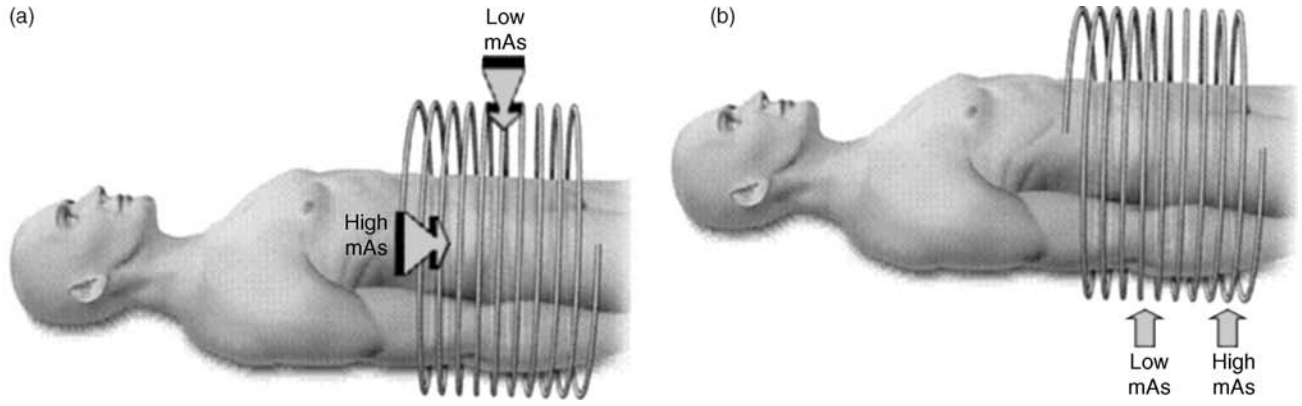


Figure 4. Principles of automatic tube current modulation: (a) Angular modulation, where the X-ray tube current is lowered as the X rays are aimed in the anterior–posterior directions, and increased when the X rays are aimed in the lateral–medial directions, when there will be more X-ray attenuation. (b) z-axis modulation where, for example, fewer X rays are required in the abdominal region superior to the pelvic bones, compared with the pelvic region.

3. Optimization and standardization of CT parameters: Just as mammographic examinations are now well standardized (36) and regulated (37), so VC should be optimized and standardized, if it is to be used for mass screening. Particularly until Points 1 and 2 are settled, it is probably premature to consider standardizing VC scanner parameters.

If VC were to become a standard screening tool for the >50s, the potential "market" in the United States will soon be >100 million people. Even if the recommended VC frequency were to be that currently recommended for optical colonoscopy (every decade), this would imply that several million VC scans might be performed each year. Should the relative simplicity of the VC tests result in the recommended examination frequency being increased, then several tens of millions of these VC scans might be expected to be performed in the United States each year. It is pertinent, therefore to consider the radiation exposure and any potential radiation risk to the population from such a mass screening program.

Because of the advantageous geometry of a VC scan, the dose–noise tradeoff can be very much weighted toward low dose, higher noise images (24,25,38–40). Several studies have come to the conclusion that more noise (and thus a lower dose) can be accepted in a VC scan compared to other CT scans, while still maintaining sensitivity and specificity, at least for polyps greater than ~7 mm in diameter (25,26,38,41,42).

It is important to note that, in general, paired VC exams are given, one in the supine and one in the prone position. Several studies have suggested that this technique improves colonic distention (43–45), decreasing the number of collapsed colonic segments.

Table 2 (46) shows estimated organ doses for one of the more common CT scanners (GE LightSpeed Ultra). The scanner parameters were taken from a recent Mayo Clinic study by Johnson and co-workers (26), and are toward the low dose end of published VC protocols (38). To provide an estimate of scanner-to-scanner dose variations, Table 3 (46) shows the radiation dose to the colon estimated for five of the more common CT scanners in use today, using identical scanner parameters in each case; the coefficient of variation of the dose to the colon is ~20%.

It can be seen from Table 2 that typical organ doses are <20 mSv, even for organs directly in the X-ray beam, for example, the colon, stomach, bladder, and kidneys. The subcohort of ~25,000 A-bomb survivors (5) that received comparable radiation doses (A-bomb dose range 5–50 mSv, mean 20 mSv) does show a slight increase in cancer mortality compared to the control population (4), but this increase is of marginal statistical significance ($p = 0.15$). It is also pertinent to point out that this A-bomb subcohort consists of individuals covering all age groups, and thus it is reasonable to assert that there is no direct statistically significant evidence from A-bomb survivor data that a single VC exam increases cancer risks in adults. It does not follow, of course, that the radiation risk is zero, rather that it is likely to be small. It also follows that there is persuasive evidence that a *series* of virtual colonoscopy examinations would result in increase in cancer risk due to the radiation exposure: The issue being how much of an increase and how it compares with the potential benefit of the VC screening.

Table 2. Typical Organ Doses and Estimated Additional Absolute Lifetime Cancer Risks Associated with a Paired VC Screening Examination of a Healthy 50 Year Old[a,b]

	Organ Dose from Paired CTC Scans,[b] mSv	Additional Absolute Lifetime Cancer Risk because of Paired CTC Scans at Age 50, %
Colon (male)	13.2	0.044
Colon (female)	13.2	0.038
Bladder (male)	16	0.025
Bladder (female)	16	0.016
Stomach (male)	14.8	0.013
Stomach (female)	14.8	0.031
Kidney (male)	16.1	0.012
Kidney (female)	16.1	0.017
Liver (male)	13.8	0.016
Liver (female)	13.8	0.005
Leukemia (male)	6.6	0.032
Leukemia (female)	6.6	0.018
Lung (male)	2.2	0.006
Lung (female)	2.2	0.008
Total (male)		*0.15*
Total (female)		*0.13*

[a]See Ref. 46.
[b]Paired VC examinations at 65 mAs, 120 kVp, 10 mm collimation, pitch 1.35.

Table 2 also shows the estimated absolute lifetime cancer risks associated with the radiation exposure from paired VC scans in a 50 year old (46). As expected, the main organs at risk are the colon, stomach, and bladder, as well as the leukemic cancers. All the estimated absolute radiation risks are relatively small, the largest being <0.05% (1 in 2000). Summed over all the organs at risk, the estimated absolute lifetime risk of cancer induction from a pair of VC scans (with the scanner parameters from Table 2) in a 50 year old is ~0.14%, ~1 in 700. Estimated risks for cancer mortality would, of course, be considerably less.

Several points need to be considered regarding the estimated risks in Table 2:

1. The risks are highly dependent on the scanner settings used, particularly the mAs and the pitch. The settings used in Table 2 are on the low dose side of

Table 3. Estimated Colon Doses from Paired VC Scans Using the Same Machine Settings with Different CT Scanners[a]

Scanner	Colon Dose from Paired CTC Scans,[b] mSv
GE LightSpeed Ultra	13.2
GE QX/I, LightSpeed, LightSpeed Plus	11.6
Phillips Mx8000	9.0
Siemens Volume Zoom, Access	8.6
Siemens Sensation 16	7.6

[a]Ref. 46.
[b]Paired VC examinations at 65 mAs, 120 kVp, 10 mm collimation, pitch 1.35.

those used in current reported studies (38), but there is good evidence (23,26,40) suggesting that the mAs and thus the dose could be decreased further, by at least a factor of 5 (and perhaps as much as a factor of 10) from these settings, while still maintaining sensitivity and specificity for polyps larger than ~5 mm. Still further reductions of up to 50% in VC doses may be possible through the use of automatic tube current modulation (Fig. 4) (14,15), now available from all the major CT scanner manufacturers (14).

2. The estimated absolute cancer risks are highly age dependent. Thus, for example, the estimated radiation-associated absolute lifetime risk for colon cancer induction decreases from 0.044% for a VC scan at age 50 to 0.022% for a scan at age 70.
3. There are quantifiable uncertainties involved in the radiation risk estimates shown in Table 2. The largest is the uncertainties in "transferring" risk estimates from a Japanese population to a U.S. population, but there are also uncertainties associated with the extrapolation of risks from somewhat higher doses, where the risks are statistically significant, and uncertainties associated with the reconstructed dosimetry estimates at Hiroshima and Nagasaki (47). Based on Monte Carlo simulations of the various uncertainties (48), the upper and lower 90% confidence limits of the radiation risk estimates are about a factor of 3 higher and lower, respectively, than the point estimates.

In summary, because the geometry for VC is highly advantageous (soft-tissue polyps projecting into an air or CO_2 filled lumen), it can be performed using lower radiation doses than almost any other CT examination. The cancer risks associated with the radiation exposure from VC are unlikely to be zero, but they are small. A best estimate for the absolute lifetime cancer risk associated with the radiation exposure using typical current scanner techniques is ~0.14% for paired VC scans for a 50 year old, and about one-half of that for a 70 year old. These values could probably be reduced by factors of 5 or 10, with optimized protocols. Thus it seems clear that, in terms of the radiation exposure, the benefit/risk ratio is potentially large for VC.

LOW DOSE CT SCREENING FOR EARLY STAGE LUNG CANCER IN SMOKERS AND EXSMOKERS

Lung cancer is the number one cancer killer in the United States. Thus, there is increasing interest in the possibility of using low dose CT scans for annual screening of smokers and former smokers for early-stage lung cancer. In part, this is the result of the failure of earlier attempts to screen this population with conventional chest X rays (49). The logic is that these earlier screening modalities failed because of their inability to detect sufficiently small (typically <10 mm) lesions—and low dose lung CT has been demonstrated to have a greater sensitivity for detecting small pulmonary lesions (50). A National Lung Cancer Screening Trial is now underway (51).

As with virtual colonoscopy, the geometry for lung CT is quite advantageous, and this allows the use of a relatively low dose (i.e., noisier) image, while still maintaining good sensitivity for detecting small pulmonary lesions (52).

The potential mortality benefits of lung cancer screening have been much debated (53–56), and it is fair to say that, at the very earliest, the issue will not be resolved until the completion of the National Lung Cancer Screening Trial in 2009. Several relatively small pilot studies have already taken place (50,56–60), suggesting that low dose lung CT does have the potential for detecting more early stage tumors than other lung screening modalities. Whether this represents a meaningful shift toward earlier detection of potentially fatal tumors, or whether it is largely associated with an overdiagnosis of nonfatal lesions, is yet to be established—as have the potential risks of invasive procedures resulting from false positives (61).

Less attention has been paid to the potential radiation risks, specifically radiation-induced lung cancer, associated with radiation from these CT scans. In part, this is because the screening technique involves "low dose", rather than standard, CT lung scans, and in part this is because excess relative risks of radiation-induced cancer generally decrease markedly with increasing age (62).

There are, however, several indications that the radiation risk to the lung associated with this screening technique may not be insignificant:

1. Cancer risks from radiation are generally multiplicative of the background cancer risk (63), which is, by definition, high for lung cancer in the target population here; this general observation has been born out in terms of the interaction between radiation and smoking, which most authors have suggested is near-multiplicative (64–70) although an intermediate interaction between additive and multiplicative has also been suggested for radon exposure (71), and there is one report of an additive interaction (72).
2. As shown in Fig. 3, while radiation-related cancer risks generally decrease markedly with increasing age at exposure, radiation-induced lung cancer does not apparently show this decrease in risk with increasing age (62,63).

These considerations suggest that risk of radiation-induced lung cancer associated with the radiation from repeated low dose CT scans of the lung in smokers may not be negligible. A recent estimate (73) suggests that a 50 year old smoker planning an annual lung screening CT would incur an estimated radiation-related lifetime lung-cancer risk of 0.5%, in addition to their otherwise expected lung cancer risk of ~14% (the radiation-associated cancer risk to any other organ is far lower). If 50% of the ever-smoking current U.S. population aged between 50 and 75 received annual CT screens, the estimated number of lung cancers associated with the radiation from these scans would be ~14,000. These estimated risks set a baseline of benefit that annual CT screening must substantially exceed. This risk–benefit analysis suggests that mortality benefits from annual CT screening

of considerably $> 3\%$ would be necessary to outweigh the potential radiation risks (73).

FULL-BODY CT SCREENING

There is increasing interest, particularly from independent radiology clinics, in the use of full-body CT screening of healthy adults (74–76). The technique is intended to be an early detection device for a variety of diseases including lung cancer, coronary artery disease, and colon cancer. At present, the evidence for the utility of this technique is anecdotal, and there is considerable controversy (77–79) regarding its efficacy: to date, no studies have yet been reported indicating a life-prolonging benefit (80). Because of the nature of the scan, the false positive rate is expected to be high (81), and a small study on full-body CT screening (80) found that 37% of those screened were recommended for further evaluation, whereas the overall evaluable disease prevalence is probably $\sim$2% (79). Other estimates suggest that the false positive rate may be as high as 90% (79).

Another aspect that is important in assessing the technique is the potential risk from the radiation exposure associated with full-body CT scans. Typical doses from a single full-body scan are $\sim$9 mGy to the lung, 8 mGy to the digestive organs, and 6 mGy to the bone marrow (82). The effective dose, which is a weighted average of doses to all organs (83), is $\sim$7 mSv. If, for example, five such scans were undertaken in a lifetime, the effective dose would be $\sim$35 mSv, that is, five times larger. Note that even with the same settings different scanners will produce somewhat different organ doses. In particular, the estimated dose to the lung is 8.9 mGy for the Siemens scanner, 9.2 mGy for the Philips, and 12.2 mGy for the GE scanner (82). To put these doses in perspective, a typical screening mammogram (see Table 1) produces a dose of $\sim$2.6 mGy to the breast (36), with a corresponding effective dose of $\sim$0.13 mSv (a factor of $\sim$50 times less).

It is important to note that these CT doses and the corresponding risk estimates are based on a particular published protocol (84). Even for the same CT settings, different scanners will produce different doses and therefore risks (varying by up to 50%). Full-body scan protocols are by no means standardized at this time, and larger mAs settings will result in correspondingly larger doses and therefore larger risks.

The estimated lifetime cancer mortality risks from a single full-body scan are $\sim 4.5 \times 10^{-4}$ ($\sim$1 in 2200) for a 45 year old, and $\sim 3.3 \times 10^{-4}$ ($\sim$1 in 3000) for a 65 year old. To put these values in perspective, the odds of an individual dying in a traffic accident in the United States during the single year 1999 were $\sim$1 in 5900 (85).

Of course, there is uncertainty in the radiation risk estimate: It is estimated (82) that the 95% credibility limits for the radiation risk estimate are about a factor of 3.2 in either direction: thus the lifetime risk from a full-body scan to a 45 year old could be as low as 1.4×10^{-4} or as high as 1.4×10^{-3}. The dominant potential radiation-induced cancer is of the lung; this is not unexpected because, as illustrated in Fig. 3, while radiation-related cancer risks generally decrease markedly with increasing age at exposure, radiation-induced lung cancer does not apparently show this decrease in risk until approximately age 55 (62,63,82).

The risk estimates for multiple scans, which would be necessary if full-body CT screening was to become a useful screening tool, are correspondingly larger. For example, a 45 year old who plans on undergoing 10 three-yearly full-body scans would potentially accrue an estimated lifetime cancer mortality risk of 0.33% ($\sim$1 in 300) (82). Again to give a comparison risk, this is comparable to the lifetime risk that a healthy 45 year old faces of dying of a brain tumor.

CONCLUSIONS

The increased availability and ease of use of CT scanners makes them attractive options for screening. The three CT screening modalities discussed here, virtual colonoscopy, lung screening, and full-body screening, are all comparatively new, and none have yet undergone definitive trials to assess their efficacy. However, both virtual colonoscopy, and low dose CT lung screening, but not full body screening, are currently undergoing national clinical trials in the United States and elsewhere. These trials should provide insight into the efficacy of virtual colonoscopy and low dose CT lung screening in terms of the potential mortality gain and the false positive rate. The trials will not, however, be able to assess the potential radiation risks, because of the long latency period between radiation exposure and development of a clinically recognizable malignancy. Nevertheless, because of the nontrivial doses associated with CT screening, the potential radiation risks will need to be factored into the overall risk–benefit analysis.

BIBLIOGRAPHY

1. Hounsfield GN. The E.M.I. scanner. Proc R Soc London B Biol Sci 1977;195:281–289.
2. Hounsfield GN. Computerized transverse axial scanning (tomography). 1. Description of system. Br J Radiol 1973;46: 1016–1022.
3. Brenner DJ, et al. Cancer risks attributable to low doses of ionizing radiation: assessing what we really know. Proc Natl Acad Sci USA 2003;100:13761–13766.
4. Preston DL, et al. Studies of mortality of atomic bomb survivors. Report 13: Solid cancer and noncancer disease mortality: 1950–1997. Radiat Res 2003;160:381–407.
5. Preston DL, et al. Effect of recent changes in atomic bomb survivor dosimetry on cancer mortality risk estimates. Radiat Res 2004;162:377–389.
6. NCRP, Evaluation of the linear-nonthreshold dose-response model for ionizing radiation, Report No. 136, NCRP; 2001.
7. NCRP. Influence of dose and its distribution in time on dose-response relationships for low-LET radiations, in: NCRP Report No. 64, National Council on Radiation Protection and Measurements, Washington (DC); 1980.
8. Ullrich RL, Jernigan MC, Satterfield LC, Bowles ND. Radiation carcinogenesis: time-dose relationships. Radiat Res 1987; 111: 179–184.
9. Howe GR. Lung cancer mortality between 1950 and 1987 after exposure to fractionated moderate-dose-rate ionizing

radiation in the Canadian fluoroscopy cohort study and a comparison with lung cancer mortality in the Atomic Bomb survivors study. Radiat Res 1995;142:295–304.
10. McNitt-Gray MF. AAPM/RSNA Physics Tutorial for Residents: Topics in CT. Radiation dose in CT. Radiographics 2002;22:1541–1553.
11. Martin CJ, Sutton DG, Sharp PF. Balancing patient dose and image quality. Appl Radiat Isot 1999;50:1–19.
12. Lehmann KJ, Wild J, Georgi M. Clinical use of software-controlled x-ray tube modulation with "Smart-Scan" in spiral CT. Aktuelle Radiol 1997;7:156–158.
13. Hundt W, et al. Dose reduction in multislice computed tomography. J Comput Assist Tomogr 2005;29:140–147.
14. Keat N. CT scanner automatic exposure control systems. Medicines and Healthcare Products Regulatory Agency, Report 05016, London; 2005.
15. Kalra MK, et al. Techniques and applications of automatic tube current modulation for CT. Radiology 2004;233:649–657.
16. Winawer SJ, et al. Prevention of colorectal cancer by colonoscopic polypectomy. The National Polyp Study Workgroup. N Engl J Med 1993;329:1977–1981.
17. Citarda F, et al. Efficacy in standard clinical practice of colonoscopic polypectomy in reducing colorectal cancer incidence. Gut 2001;48:812–815.
18. Seeff LC, et al. Patterns and predictors of colorectal cancer test use in the adult U.S. population. Cancer 2004;100:2093–2103.
19. Subramanian S, Amonkar MM, Hunt TL. Use of colonoscopy for colorectal cancer screening: evidence from the 2000 national health interview survey. Cancer Epidemiol Biomarkers Prev 2005;14:409–416.
20. Coin CG, et al. Computerized radiology of the colon: a potential screening technique. Comput Radiol 1983;7:215–221.
21. Hara AK, et al. Detection of colorectal polyps by computed tomographic colography: feasibility of a novel technique. Gastroenterology 1996;110:284–290.
22. Pickhardt PJ, et al. Computed tomographic virtual colonoscopy to screen for colorectal neoplasia in asymptomatic adults. N Engl J Med 2003;349:2191–2200.
23. Iannaccone R, et al. Computed tomographic colonography without cathartic preparation for the detection of colorectal polyps. Gastroenterology 2004;127:1300–1311.
24. Hara AK, et al. Reducing data size and radiation dose for CT colonography. AJR Am J Roentgenol 1997;168:1181–1184.
25. Macari M, et al. Colorectal neoplasms: prospective comparison of thin-section low-dose multi-detector row CT colonography and conventional colonoscopy for detection. Radiology 2002; 224:383–392.
26. Johnson KT, et al. CT colonography: determination of optimal CT technique using a novel colon phantom. Abdom Imaging 2004;29:173–176.
27. Cotton PB, et al. Computed tomographic colonography (virtual colonoscopy): a multicenter comparison with standard colonoscopy for detection of colorectal neoplasia. JAMA 2004;291: 1713–1719.
28. Weitzman ER, Zapka J, Estabrook B, Goins KV. Risk and reluctance: understanding impediments to colorectal cancer screening. Prev Med 2001;32:502–513.
29. Ristvedt SL, McFarland EG, Weinstock LB, Thyssen EP. Patient preferences for CT colonography, conventional colonoscopy, and bowel preparation. Am J Gastroenterol 2003;98: 578–585.
30. Akerkar GA, Yee J, Hung R, McQuaid K. Patient experience and preferences toward colon cancer screening: a comparison of virtual colonoscopy and conventional colonoscopy. Gastrointest Endosc 2001;54:310–315.
31. Harewood GC, Wiersema MJ, Melton LJ, 3rd. A prospective, controlled assessment of factors influencing acceptance of screening colonoscopy. Am J Gastroenterol 2002;97:3186–3194.
32. Gluecker TM, et al. Colorectal cancer screening with CT colonography, colonoscopy, and double-contrast barium enema examination: prospective assessment of patient perceptions and preferences. Radiology 2003;227:378–384.
33. Callstrom MR, et al. CT colonography without cathartic preparation: feasibility study. Radiology 2001;219:693–698.
34. Lefere PA, et al. Dietary fecal tagging as a cleansing method before CT colonography: initial results polyp detection and patient acceptance. Radiology 2002;224:393–403.
35. Zalis ME, Perumpillichira J, Del Frate C, Hahn PF. CT colonography: digital subtraction bowel cleansing with mucosal reconstruction initial observations. Radiology 2003;226: 911–917.
36. Kruger RL, Schueler BA. A survey of clinical factors and patient dose in mammography. Med Phys 2001;28:1449–1454.
37. Monsees BS. The Mammography Quality Standards Act. An overview of the regulations and guidance. Radiol Clin N Am 2000;38:759–772.
38. van Gelder RE, et al. CT colonography at different radiation dose levels: feasibility of dose reduction. Radiology 2002;224: 25–33.
39. Hara AK, et al. CT colonography: single- versus multi-detector row imaging. Radiology 2001;219:461–465.
40. Iannaccone R, et al. Detection of colorectal lesions: lower-dose multi-detector row helical CT colonography compared with conventional colonoscopy. Radiology 2003;229:775–781.
41. Taylor SA, et al. Multi-detector row CT colonography: effect of collimation, pitch, and orientation on polyp detection in a human colectomy specimen. Radiology 2003;229:109–118.
42. Wessling J, et al. CT colonography: Protocol optimization with multi-detector row CT—study in an anthropomorphic colon phantom. Radiology 2003;228:753–759.
43. Chen SC, Lu DS, Hecht JR, Kadell BM. CT colonography: value of scanning in both the supine and prone positions. AJR Am J Roentgenol 1999;172:595–599.
44. Morrin MM, et al. CT colonography: colonic distention improved by dual positioning but not intravenous glucagon. Eur Radiol 2002;12:525–530.
45. Fletcher JG, et al. Optimization of CT colonography technique: prospective trial in 180 patients. Radiology 2000;216: 704–711.
46. Brenner DJ, Georgsson MA. Mass screening with CT colonography: Should the radiation exposure be of concern? Gastroenterology 2005;129(1):328–337.
47. NCRP. Uncertainties in fatal cancer risk estimates used in radiation protection. Report 126, National Council on Radiation Protection and Measurements, Bethesda, MD; 1997.
48. Land CE, Gilbert E, Smith JM. Report of the NCI-CDC Working Group to Revise the 1985 NIH Radioepidemiological Tables. NIH Publication 03-5387. See also, available at www.irep.nci.nih.gov, NIH, Bethesda (MD); 2003.
49. Fontana RS. The Mayo Lung Project: a perspective. Cancer 2000;89:2352–2355.
50. Henschke CI, et al. Early lung cancer action project: Overall design and findings from baseline screening. Lancet 1999;354: 99–105.
51. Vastag B. Lung screening study to test popular CT scans. JAMA 2002;288:1705–1706.
52. Rusinek H, et al. Pulmonary nodule detection: low-dose versus conventional CT. Radiology 1998;209:243–249.
53. Aberle DR, et al. A consensus statement of the Society of Thoracic Radiology: screening for lung cancer with helical computed tomography. J Thorac Imaging 2001;16:65–68.

54. Miettinen OS, Henschke CI. CT screening for lung cancer: coping with nihilistic recommendations. Radiology 2001;221: 592–596.
55. Patz EF, Jr. Black WC, Goodman PC. CT screening for lung cancer: not ready for routine practice. Radiology 2001;221: 587–591.
56. Swensen SJ, et al. Screening for lung cancer with low-dose spiral computed tomography. Am J Respir Crit Care Med 2002;165:508–513.
57. Sone S, et al. Results of three-year mass screening programme for lung cancer using mobile low-dose spiral computed tomography scanner. Br J Cancer 2001;84:25–32.
58. Nawa T, et al. Lung cancer screening using low-dose spiral CT: results of baseline and 1-year follow-up studies. Chest 2002; 122:15–20.
59. Garg K, et al. Randomized controlled trial with low-dose spiral CT for lung cancer screening: Feasibility study and preliminary results. Radiology 2002;225:506–510.
60. Sobue T, et al. Screening for lung cancer with low-dose helical computed tomography: anti-lung cancer association project. J Clin Oncol 2002;20:911–920.
61. Mahadevia PJ, et al. Lung cancer screening with helical computed tomography in older adult smokers: a decision and cost-effectiveness analysis. JAMA 2003;289:313–322.
62. Thompson DE, et al. Cancer incidence in atomic bomb survivors. Part II: Solid tumors, 1958–1987. Radiat Res 1994;137: S17–67.
63. NRC. Health effects of exposure to low levels of ionizing radiation: BEIR V. Washington (DC): National Academy Press; 1990.
64. Gilbert ES, et al. Lung cancer after treatment for Hodgkin's disease: focus on radiation effects. Radiat Res 2003;159: 161–173.
65. Tokarskaya ZB, et al. Interaction of radiation and smoking in lung cancer induction among workers at the Mayak nuclear enterprise. Health Phys 2002;83:833–846.
66. Melloni B, Vergnenegre A, Lagrange P, Bonnaud F. Radon and domestic exposure. Rev Mal Respir 2000;17:1061–1071.
67. Morrison HI, Villeneuve PJ, Lubin JH, Schaubel DE. Radon-progeny exposure and lung cancer risk in a cohort of Newfoundland fluorspar miners. Radiat Res 1998;150:58–65.
68. Neugut AI, et al. Increased risk of lung cancer after breast cancer radiation therapy in cigarette smokers. Cancer 1994; 73:1615–1620.
69. Pershagen G, et al. Residential radon exposure and lung cancer in Sweden. N Engl J Med 1994;330:159–164.
70. Samet JM, et al. Lung cancer mortality and exposure to radon progeny in a cohort of New Mexico underground uranium miners. Health Phys 1991;61:745–752.
71. Hornung RW, Deddens J, Roscoe R. Modifiers of exposure-response estimates for lung cancer among miners exposed to radon progeny. Environ Health Perspect 1995;103(Suppl 2): 49–53.
72. Pierce DA, Sharp GB, Mabuchi K. Joint effects of radiation and smoking on lung cancer risk among atomic bomb survivors. Radiat Res 2003;159:511–520.
73. Brenner DJ. Radiation risks potentially associated with low-dose CT screening of adult smokers for lung cancer. Radiology 2004;231:440–445.
74. FDA. Full-body CT scans: What you need to know. DHHS Publication FDA (03)-0001. Available at www.fda.gov/cdrh/ct/ctscansbro.html, U.S. Food and Drug Administration, Rockville (MD); 2003.
75. Brant-Zawadzki M. CT screening: why I do it. AJR Am J Roentgenol 2002;179:319–326.
76. Illes J, et al. Self-referred whole-body CT imaging: current implications for health care consumers. Radiology 2003;228: 346–351.
77. Holtz A. Whole-body CT screening: Scanning or scamming? Oncol Times 2003;25:5–7.
78. Berland LL, Berland NW. Whole-body computed tomography screening. Semin Roentgenol 2003;38:65–76.
79. Beinfeld MT, Wittenberg E, Gazelle GS. Cost-effectiveness of whole-body CT screening. Radiology 2005;234:415–422.
80. Casola G, et al. Whole body CT screening: Spectrum of findings and recommendations. Radiology 2002;225(Suppl.):317.
81. Casarella WJ. A patient's viewpoint on a current controversy. Radiology 2002;224:927.
82. Brenner DJ, Elliston CD. Estimated radiation risks potentially associated with full-body CT screening. Radiology 2004 232:735–738.
83. ICRP. 1990 Recommendations of the International Commission on Radiological Protection: Publication 60. Oxford: Pergamon; 1991.
84. Fishman EK, Horton KM. What application should you offer in a whole body CT screening center? Available at www.screeningctisus.com/articles/screeningctisus.html.
85. Hoyert DL, et al. Deaths: final data for 1999. Natl Vital Stat Rep 2001;49:1–113.
86. Evens RG, Mettler FA. National CT use and radiation exposure: United States 1983. AJR Am J Roentgenol 1985;144: 1077–1081.
87. Bahador B. Trends in diagnostic imaging to 2000. London: Financial Times Pharmaceuticals and Healthcare Publishing; 1996.
88. UNSCEAR. Sources and effects of ionizing radiation: United Nations Scientific Committee on the Effects of Atomic Radiation: UNSCEAR 2000 report to the General Assembly. United Nations, New York; 2000.
89. Linton OW, Mettler FA, Jr. National conference on dose reduction in CT, with an emphasis on pediatric patients. AJR Am J Roentgenol 2003;181:321–329.

See also BONE DENSITY MEASUREMENT; COMPUTER-ASSISTED DETECTION AND DIAGNOSIS.

COMPUTED TOMOGRAPHY SIMULATOR

XIANGYANG TANG
GE Healthcare Technologies
Waukesha, Wisconsin

GE WANG
University of Iowa
Iowa City, Iowa

INTRODUCTION

The development of conformal radiation therapy (RT) and computerized treatment planning dates back to the late 1950s (1,2). The first milestone of computerized treatment planning is the invention of beam's eye view (BEV) display in the late 1970s (3,4). Up to now, along with surgery and chemotherapy, conformal RT has become the most effective measure for curative or palliative management of cancers at various anatomic sites (5). The biological mechanism underlying RT is that radiation damages crucial structures, such as deoxyribonucleic acid (DNA), of a cell, resulting in cell death, whether the cell is cancerous or normal, when the biologically damaged cell cannot be repaired by itself (5). The radiation dose delivered by RT is toxic to

normal biological tissues and organs while it kills abnormal cells to cure cancer or palliate the local symptom of a malignant tumor (5). If the radiation dose distribution is made sufficiently concentrated on a targeted cancerous volume and tolerable over nearby normal anatomic structures, the cells within the cancerous volume may be killed while those within the normal organs or tissues survive. Hence, the ultimate goal of radiation therapy is to cure or control the cancer by precisely delivering an adequate and a homogeneous radiation dose to the targeted cancerous volume while maintaining the unavoidable dose to surrounding biological structures, particularly those critical organs or tissues that are very sensitive to radiation dose, such as eye, testis, lung, spinal cord, brain, and so on, below the biological toxicity tolerance. To improve the therapeutic ratio while decreasing the occurrence of acute or chronic side effects caused by radiation toxicity as much as possible, an administration of fractionated radiation therapy process has been clinically proven to be more efficient than a "lump sum" radiation delivery (5).

A typical conformal RT process consists of diagnostic data acquisition, simulation, treatment planning, treatment verification, and treatment delivery, although its implementation can be customized based upon available personnel and resources. Intuitively, a malignant tumor cannot be cured or its local symptom cannot be controlled if it misses the adequate radiation dose prescribed by a radiation oncologist. Meanwhile, the normal cells of surrounding tissues or organs can be fatally damaged if it absorbs the radiation dose supposed to be delivered to a targeted cancerous tumor. Consequently, among all factors compromising the success of a radiation therapy process, the geometrical imprecision or inaccuracy caused by patient localization and immobilization, as well as inadequate dose during treatment delivery, play dominant roles. Moreover, since a radiation therapy process is usually administrated in a fractionated manner, the patient position reproducibility between treatment delivery fractions is also of crucial importance. The maintenance of geometrical precision, accuracy, and reproducibility, which can never be overemphasized, are the tasks of simulation, while the warranty of delivering an adequate radiation dose at a homogenous distribution are the tasks of radiation treatment planning.

Conventionally, being carried out by a physician on a simulator in which two-dimensional (2D) imaging techniques are usually utilized, the simulation in conformal RT is an interactive and iterative process. Instead of using photons at the MeV energy level that are utilized in radiation treatment delivery, photons at the keV energy level, that is, X-ray, are utilized in the simulation to provide fluoroscopy for beam designing, in which an image intensifier or flat panel imager is usually employed. Moreover, X-ray source and film-based radiography are usually utilized for portal verification. In the beginning of the simulation, a radiation oncologist identifies the clinical target volume (CTV) by initially specifying the gross target volume (GTV) with a clinical margin added by taking the possible surrounding metastases and spreads into account (5). A planning target volume (PTV) is eventually defined by taking an extra margin into account to compensate for systematic geometrical mismatch between the simulator and the treatment machine, as well as geometrical error caused by the beam setup uncertainty and internal organ motion (6). It is important to state that, such an extra margin is crucial to the success of a radiation therapy, because the internal organ of a patient is always moving, no matter how perfect the patient localization, immobilization, and the geometric match between simulator and treatment machines can be achieved. By mimicking the geometry of a radiation therapy machine, such as a linear accelerator (LINAC), and with the availability of those 2D imaging techniques mentioned above, the conventional simulation generates a delineation of PTV and layout of radiation fields, including the number and orientation of beams, aperture of collimator, shape and size of field, as well as the specification of beam blocker, wedge, or compensator and markers. Since the geometry of a conventional simulator is exactly the same as that of a treatment machine, a spatial and geometrical integrity between them can be achieved if radioopaque markers are placed on a patient's skin or at appropriate anatomic landmarks, for example, the sternum for a patient with breast cancer. The conventional simulation in conformal radiation therapy is an on-line iterative process. Such an on-line process is inefficient in terms of patient throughput, since it requires a patient to remain in the simulator couch during the entire simulation process. It is very tough, if not impossible, to obtain an optimized conventional simulation in conformal radiation therapy, because this tedious and time-consuming simulation process is greatly dependent on the skill or experience of the physician committed to the task.

Intrinsically, the anatomic structure of a human being is three-dimensional (3D), and can be mapped to 3D models representing its geometric, pathologic, and physiologic characteristics, respectively. The advent of clinical X-ray computed tomography (CT) in the early 1970s was a revolution over 2D imaging technology, making the 3D modeling of a patient a reality. However, the X-ray CT scanner in its early stage was not capable of providing high quality, especially high spatial resolution, tomographic images of a patient volume within an acceptable or tolerable time. Nevertheless, the potentiality offered by X-ray CT technology for RT was well recognized then. In the 1980s, CT technology evolved dramatically while great progress simultaneously had been accomplished in 3D computer visualization technologies. By combining the state-of-the-art CT technology with modern 3D computer visualization, the concept of CT simulator or virtual simulator was formed in the late 1980s (7–9). Since then, enormous effort and resources were invested in the research and development (R&D) of CT or virtual simulator, resulting in numerous CT or virtual simulators commercially available in the market. For convenience, a CT simulator hereafter refers to either a CT simulator or a virtual simulator unless otherwise specified.

As shown in Fig. 1, a CT simulator consists of a radiation therapy dedicated CT scanner (viz., RT-dedicated CT scanner) and a software workstation. The RT-dedicated CT scanner is to provide a 3D model of the patient to be treated by acquiring contiguous tomographic images over a volume of interest. The software workstation is to carry out simulation based on the 3D anatomic structures and clinical

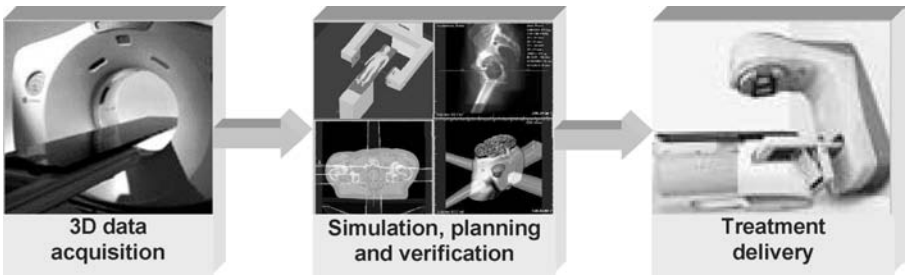


Figure 1. A schematic diagram showing the process of 3D conformal RT using a CT simulator: (a) Three-dimensional patient data acquisition through a RT dedicated CT scanner; (b) CT simulation, treatment planning and verification; (c) Treatment delivery via a LINAC.

information revealed by the 3D model of the patient. In general, the simulation software of a CT simulator can be either stand-alone or an embedded part of a treatment planning system. By integrating an RT-dedicated CT scanner and simulation software, a CT-simulator generally conducts the following tasks: patient positioning, patient immobilization, 3D patient data set acquisition by CT scanning, identification of a target volume (GTV, CTV, and PTV) and surrounding vital normal tissues and organs, placement of beams (number, orientation, isocenter, and collimator aperture), field design (beam shaper, blocker, wedge and compensator), as well as generation of portal images, such as digitally reconstructed radiography (DDR) and other instructions exported to a treatment planning and RT machine (10–17). In the following sections, these tasks are described in detail.

THREE-DIMENSIONAL COMPUTED TOMOGRAPHY PATIENT DATA ACQUISITION

A CT scanner is the cornerstone of a CT simulator for treatment planning and delivery using 3D conformal RT. With a focus on diagnostic imaging, CT technology has been substantially improved over the past three decades and became the most popular tomographic imaging modality in clinics (18,19). With the impressive progresses in CT technology, particularly the image reconstruction methods for multidetector row CT and volumetric CT (20–22), a state-of-the-art diagnostic CT scanner has well satisfied the requirement posed by RT. Unfortunately, however, a start-of-the-art diagnostic CT scanner usually cannot be readily used in a CT simulator for RT treatment planning, because the gantry aperture diameter of a diagnostic CT scanner is usually ~70.0 cm, which cannot guarantee a smooth accommodation of a patient and necessary immobilization devices, for example, the arm holder utilized in breast cancer radiation therapy (14). As a result, RT-dedicated CT scanners have been developed by major medical CT scanner manufacturers and are currently available on the market. The gantry aperture diameter of an RT-dedicated CT scanner ranges from 80 to 85 cm with a display field of view (DFOV) between 60 and 65 cm, which can handle virtually all clinical situations. Technical parameters of a typical RT-dedicated CT scanner are listed in Table 1. The performance of an RT-dedicated CT scanner should be periodically calibrated and verified (14,17). It is interesting to note that, although its overall performance is inferior to that of a state-of-the-art diagnostic CT scanner, a currently available RT-dedicated CT scanner can serve radiation therapy very well (12). One of the major reasons for such a technical delay is a relatively small market volume of RT-dedicated CT scanners in comparison to that of diagnostics CT scanners. Considering the fast evolution of the diagnostic CT scanner and the technical catching up of the RT-dedicated CT scanner in X-ray detector z coverage and scanning speed, it is very hopeful for a future RT-dedicated CT scanner to have the overall performance of a current diagnostic CT scanner.

In general, the protocols used for diagnostic CT imaging can be used correspondingly by an RT-dedicated CT scanner to acquire a 3D patient data set for CT simulation, although a trade-off between the spatial resolution along the longitudinal direction of a patient (image slice thickness) and the total number of tomographic images has to be made in practice. As shown in Table 1, whereas the thinnest available slice thickness in a commercial RT-dedicated CT scanner is between 1.0 and 2.0 mm, an image slice thickness thinner than 3.0 mm is usually acceptable to render DRR for portal verification (23).

PATIENT POSITIONING AND IMMOBILIZATION

It has to be emphasized that, to guarantee a geometrical match between the CT scanning and the treatment delivery, the patient has to be in exactly the same position with all the immobilization devices, such as foam body casts, thermoplastic head masks (24), and stereotactic frames (25), in place. In addition to an enlarged gantry aperture to accommodate a patient and immobilization devices, an RT-dedicated CT scanner has a flat top patient couch that is exactly the same as the one used in a treatment machine so that an exact geometrical match and reliable patient position reproducibility can be achieved between the RT-dedicated CT scanner and the treatment system. As illustrated below, all the tasks accomplished by a CT simulator are solely dependent on the spatial integrity of the 3D patient data set acquired by an RT-dedicated CT scanner. Also, the importance of geometric accuracy and match between the CT scanner and RT machine, as well as the reproducibility of patient position, can never be overstated.

In principle, the orthogonal laser beams existing in an RT-dedicated CT scanner gantry can be employed to mark the isocenter of a target volume. However, to assure a convenient and efficient marking process on the patient's skin or other anatomic landmarks, it is preferable to have external laser devices installed on the side wall and ceiling of the room where the RT-dedicated CT scanner is installed (14). The laser beams on the side wall (transaxail and

Table 1. Primary Technical Parameters of a Typical Radiation Therapy Dedicated CT Scanner

Scan mode	Axial; Helical	
Scan speed (s/360° gantry rotation)	1.0, 2.0, 3.0, 4.0	
Number of detector row	4	
Diameter of gantry bore (mm)	800.00	
Display FOV (mm)	650.0	
X-ray tube	Current: 10–440 mA, 5mA increments kVp: 80, 100, 120, 140 Power: 0.8–53.2 (kW) Heat capacity: 6.3 MU	
Gantry tilt	±30°	
Image slice thickness (mm)	0.625, 1.25, 2.5, 3.75, 5.0, 7.5, 10.0	
Dose (mGy/100 mAs)	CTDI:	
		Head: 16.5 (center); 17.2 (surface)
		Body: 5.6 (center); 11.0 (surface)
	$CTDI_{100}$:	
		Head: 17.0 (center); 18.4 (surface)
		Body: 5.4 (center); 11.8 (surface)
	$CTDI_{vol}$:	
		Head: 17.9 Body: 9.7
In-plane spatial resolution (lp/cm)	Standard	Hi-resolution
	4.2 at 50% MTF	10.5 at 50% MTF
	6.8 at 10% MTF	13.9 at 10% MTF
	8.5 at 0% MTF	15.4 at 0% MTF
Low contrast resolution [on 8 in. CATPHAN phantom]	5mm at 0.3% at 13.3 mGy 3mm at 0.3% at 37.6 mGy	
Noise(on AAPM water phantom)	0.32% +/−0.03% at 25.2 mGy (2.52 rad)	
Image generation speed (s/frame)	0.17	
Image matrix	512×512	
CT number scale	− 31743–31743 HU	

coronal) are fixed to identify the isocenter by moving the patient couch longitudinally and vertically. The laser beam on the ceiling (sagittal) has to be moveable laterally to reach the isocenter of a target volume since the patient couch of a CT scanner is generally not movable laterally. A schematic diagram showing the deployment of external laser beams on an RT-dedicated CT scanner is illustrated in Fig. 2.

There exist two ways to identify the isocenter of a target volume (14). The first way is to place marks on a patient's skin to obtain the absolute location of the isocenter in the coordinate system of the RT-dedicated CT scanner. Since the coordinate system of the RT-dedicated CT scanner is aligned with that of a treatment machine by geometrical calibration and verification, the isocenter coordinates obtained in such a way are ready for utilization by the treatment machine as long as the patient is appropriately marked. The primary advantage of this manner is the avoidance of any errors due to intermediate geometrical transforms. However, this method requires the involvement of radiation oncologists and physicists simultaneously on site while the patient remains in the couch of the RT-dedicated CT scanner. Such a process is called a synchronized mode because the identification of a target volume has to be accomplished in real-time as described above. The second way is relatively flexible and can be carried out off-line. The patient with a few radioopaque marks placed on the skin or preferably solid anatomic landmarks goes through a regular CT scanning, and then can immediately leave the hospital. The radioopaque landmarks can be easily placed with the availability of external laser beams of an RT-dedicated CT scanner as introduced above. In the simulation, the isocenter of a target volume can be readily determined by a physicist under the guidance of a radiation oncologist. This way is in an asynchronous mode and provides substantial flexibility for physicians to improve patient throughput and patient comfort.

COMPUTED TOMOGRAPHY SIMULATION

With an RT-dedicated CT scanner, a set of transaxial images covering a volume of interest are obtained and

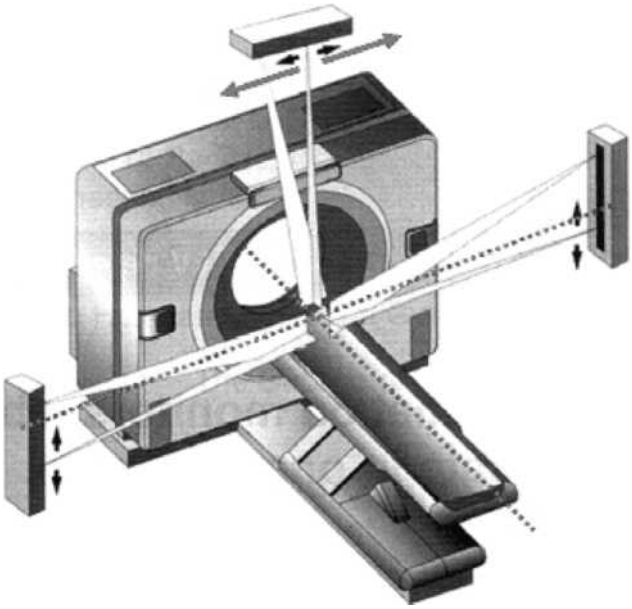


Figure 2. A schematic diagram showing an RT-dedicated CT scanner and its associated external laser beams for patient positioning in 3D conformal RT using a CT simulator for treatment planning (Courtesy Robert L. Steinnhauser, Gammex rmi.)

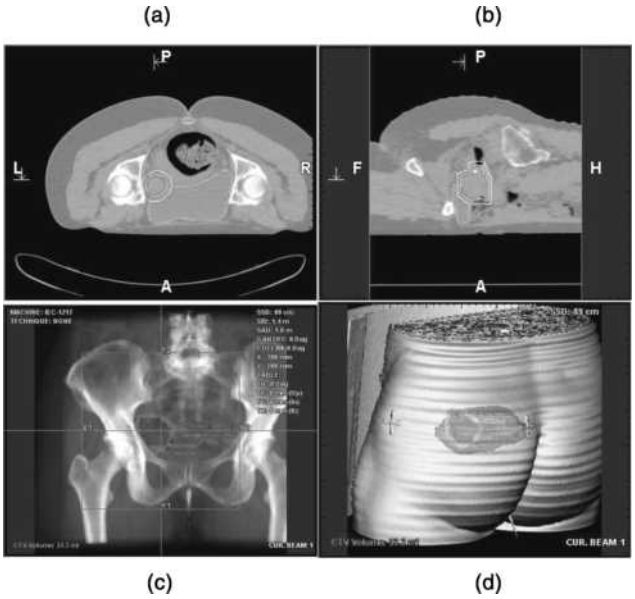


Figure 3. A target volume and its nearby normal critical organ are identified using slice-wise contouring in transaxial image plane (a), sagittal image plane (b), beam's eye view (c) and operator's eye view via segmented semitransparent surface rendering (d). In baseline image planes, the CTV is outlined by red, PTV by yellow and the nearby critical organ at risk by green contours, respectively, while colors fill each volume correspondingly in beam's eye view and operator's eye view. [Courtesy Dr. Georgios Sakas and Dr. Evelyn A. Firle, Fraunhofer Institute for Computer Graphics (IGD).] Please see online for color figure.

stored in the simulation workstation. Generally speaking, the 3D data set contains much more information about the patient than can be viewed in the transaxial plane only. For example, with the availability of multiplanar reformatting in other baseline planes (coronal, sagittal, and even oblique), physicians can attain more freedom in visualizing a targeted tumor volume and its geometrical relationship with adjacent normal anatomic structures. The 3D patient data set can be conceived as a virtual patient or a 3D model of the patient to be treated, and a 3D model fiducially represents both anatomic and physical properties of the patient. With modern 3D image processing and visualization techniques that will be described below, the anatomic and physical information of the patient, can be exploited in much more details (13).

Structure Identification

The purpose of a CT simulator is not only to identify the extent of a tumor, but also to provide an unambiguous delineation of the relationship between the targeted tumor volume and its neighboring normal tissues or organs. The basic, but most effective, tool in a CT simulator to identify a tumor is the use of contouring to define the GTV, CTV, and PTV of a targeted tumor volume, in contiguous transaxial tomographic images. The GTV of a cancerous tumor is initially specified by a radiation oncologist via outlining the boundaries of the GTV manually or semiautomatically. As mentioned above, during the target volume defining process, it is necessary to expand the GTV into a CTV by including clinical margins surrounding the GTV. The aim of adding clinical margins is to assure that possible spreading or metastases be included in the targeted volume. Furthermore, recognizing the imperfect geometrical reproducibility over treatment fractions, such as systematic geometric errors caused by beam setup uncertainty and geometrical mismatch between an RT-dedicated CT scanner and a treatment machine, the random and/or periodic geometrical errors caused by patient and organ motions, such as breathing and swallowing, it is crucial to include an adequate margin into the planed target volume (5,6).

A manual contouring process is conducted by defining the boundary of the GTV, CTV, and PTV, respectively, in every image slice. The image slice can be in any baseline plane, though the transaxial slice is preferably utilized in practice. An example of such a process is illustrated in Fig. 3. The manual contouring operation can also be done in several other slices sequentially, and the boundaries in interim images can be automatically obtained by linear or nonlinear interpolation techniques, as long as the gap between adjacent contoured image slices are within certain threshold. An even more efficient way to carry out the manual contouring operation is to define the boundaries of the target volume on the central transaxial, coronal, and sagittal planes, respectively, and then the target volume can be obtained by automatic 3D linear or nonlinear interpolation techniques available in a CT simulator. On the other hand, a semiautomatic contouring can be accomplished by just placing a seed within the targeted volume, and the boundaries are then determined by thresholding over CT numbers in Housfield units or other characteristics of the voxels in an image. Subsequently, the targeted volume can be obtained through automatic image processing techniques, such as volume growing over contiguous image slices.

Beam Design

Once the targeted cancerous tumor is defined by the techniques presented above, the next step in the simulation is beam design: to determine the number and orientation of beams, aperture of collimators, as well as shape and size of the beam field. Prior to the beam design, a physician has to specify a coordinate system convention based on which simulation is to be conducted, the treatment machine type, and its associated characteristics, such as modality (photon or electron) and energy level of the beam intended for radiation treatment. Usually, a CT simulation workstation provides versatile and powerful 2 and 3D visualization tools and utilities for beam manipulation, and a typical beam designing process in central baseline planes is illustrated in Fig. 4. In the process, the orientation or deployment of beams is specified by machine angle, collimator angle, and patient couch angle, with the beam manipulation tools, such as creating, adding, deleting, mirroring, duplicating, or renaming, in the CT simulator. The isocenter of an identified target volume can be adjusted, and the resultant coordinate system shifts are recorded and exported to a treatment machine. With versatile 2 and 3D visualization tools and utilities, the collimator aperture can be adjusted interactively by clicking and dragging the *X*- and *Y*-jaw of a collimator on the display screen of the simulation workstation. The beam field can conform the boundaries of the identified target volume by manipulating various beam shapers (either aperture or shield). A shaper can be either available in the standard set offered by a CT simulator or defined by a physician according to the structure to be conformed. The definition of a new shaper

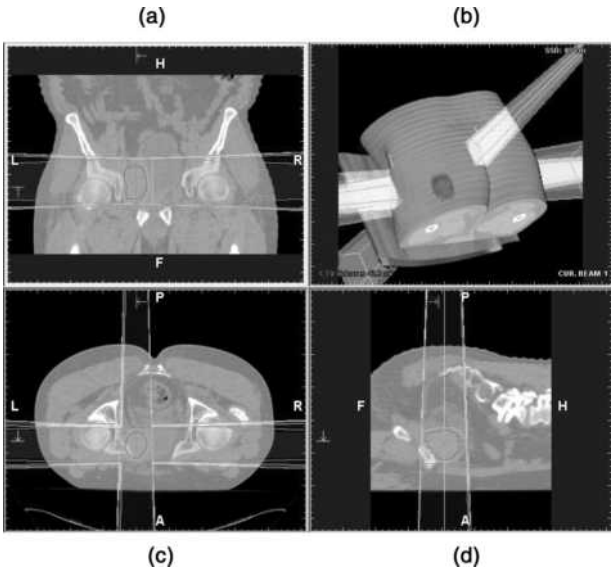


Figure 4. Beams are designed to deliver radiation treatment to the target volume identified through the process shown in Fig. 3, with the beam manipulating functions provided by a CT simulator in coronal plane (a), operator's eye view via 3D segmented semitransparent surface rendering (b),transaxial plane (c), and sagittal plane (d). [Courtesy Dr. Georgios Sakas and Dr. Evelyn A. Firle, Fraunhofer Institute for Computer Graphics (IGD).] Please see online for color figure.

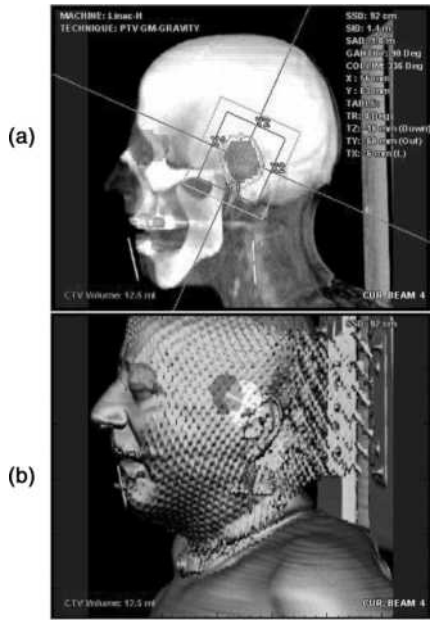


Figure 5. Modern 3D visualization techniques are extensively utilized in CT simulator to design beams via beam shaper for 3D conformal RT: (a) Beam's eye view, in which the CTV is filled in red, the PTV is outlined by green contour, the collimator aperture is defined by red rectangle, and the organ at risk (eye) is filled with green; (b) Operator's eye view using segmented semitransparent surface rendering, in which the PTV is in red, the beam field on the surface in yellow, and the organ at risk in green. [Courtesy Dr. Georgios Sakas and Dr. Evelyn A. Firle, Fraunhofer Institute for Computer Graphics (IGD).] Please see online for color figure.

can be either manual by clicking and dragging on the display screen or automatic by conforming the shaper to the boundaries of the targeted volume. Markers defining the corner of the field can be placed on a patient's skin or anatomic landmarks via clicking and dragging or typing into corresponding entries on the display screen. Furthermore, a typical CT simulator usually provides numerous utility functions for geometry measurement, such as distance, angle, area and volume, and the measurement results are displayed on the screen while textual annotations can be manipulated simultaneously.

Alternatively, the procedures introduced above can be carried out in the so-called BEV: the most fundamental and powerful 3D visualization technique employed in the simulation. Being illustrated in Fig. 5, the BEV is obtained via DRR by mimicking an X-ray source emanating from exactly the same location corresponding to the radiation focal spot of a treatment machine. With the availability of BEV, the following operations can be readily carried out: adjustment of the beam's isocenter, adding makers to the corners of a beam field, and manipulating the leaves of a multileaf collimator (MLC). Note that a number of modern image processing techniques can be combined with 3D visualization to highlight interested anatomic features while the interference from noninterested anatomic structures is being removed from visualization. For example, as shown in Fig. 6, by thresholding voxel gray values in an appropriate range, DRRs of the lungs, fat, muscle, other

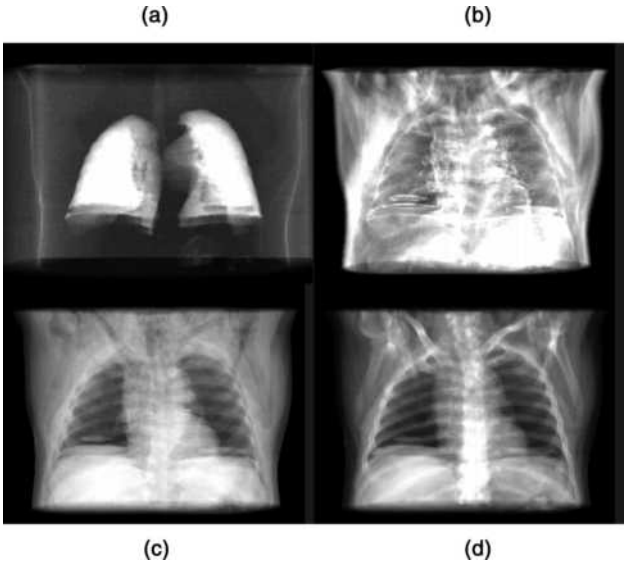


Figure 6. Other examples of modern 3D visualization techniques provided by a typical CT simulator for treatment planning in 3D conformal RT, in which only the interested organs or tissues are displayed: DRR of lung (a); DRR of fat (b); DRR of muscle (c); DRR of all anatomic structure and tissues (d). [Courtesy Dr. Georgios Sakas and Dr. Evelyn A. Firle, Fraunhofer Institute for Computer Graphics (IGD).]

tissues and organs can be exclusively displayed to facilitate target volume identification and contouring, beam deployment, and field designing. Another example of 3D visualization techniques for CT simulation is depth controlling, in which only a slab containing the target volumes and their surrounding structures are displayed. All these DRR-based 3D visualization techniques can be combined wherever they can make a simulation process more productive and reliable. Moreover, other modern 3D visualization techniques, such as surface rendering to generate operator's eye view or room's eye view, enable an even smoother and more efficient beam designing process. It is interesting to mention that the transparency in surface rendering can be adjusted to facilitate the simulation (e.g., Figs. 3 and 4).

In practice, various beam blockers to shape the beam field can be made of lead, depleted uranium, or low melting point lead alloy (5,26). Usually, these shapers are customized for individual patients. Such a customization process is time consuming, inefficient, and expensive. Instead of customizing beam shapers for each target volume of each patient, a beam field can be shaped by a multileaf collimator in which two banks of leaves are installed oppositely with each leaf being driven by a computer-controlled step motor (27–29) (see Fig. 7). Through a coordinated movement of these leaves, the beam field can conform various target volumes, provided that the thickness of a leaf is sufficiently small, such as 0.5 or 1.0 cm. Thus, the multileaf collimator is a much more general and efficient beam shaping technique, since it can be repeatedly and unlimitedly utilized in 3D conformal radiation therapy for any targeted volume at any location within a patient. Consequently, the facility for beam shaper fabrication, such as block cutting and compensator making, are no longer

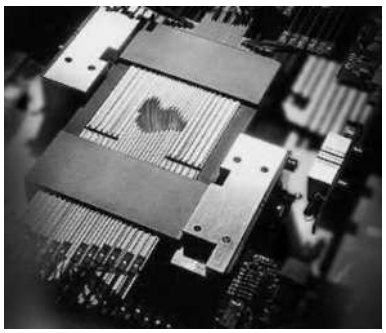


Figure 7. A multileaf collimator used for beam shaping in 3D conformal radiation therapy consists of two banks of metal leaves that are installed oppositely, and each leaf is driven by a computer-controlled step motor or pneumatic device individually.

needed. To suppress the interleaf radiation leakage as much as possible, there usually exists interlocking between adjacent and opposite leaves of a multileaf collimator, which is implemented in a tongue–groove structure. All the beam design techniques introduced above can be readily employed with a multileaf collimator. An example is shown in Fig. 8, in which the deployment of all leaf is illustrated.

It is not an exaggeration to state that, without the aforementioned 3D visualization-based functionalities,

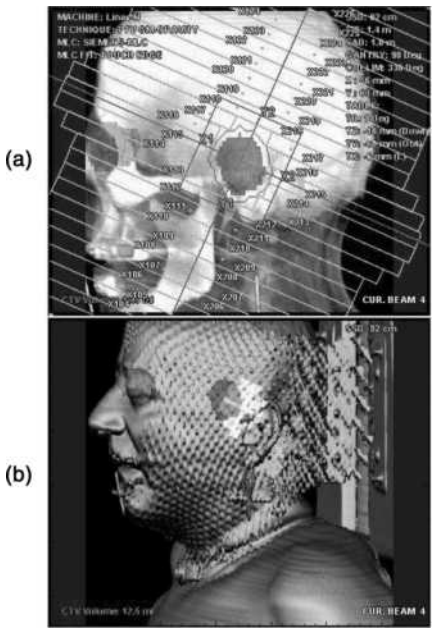


Figure 8. Modern 3D visualization techniques are extensively utilized in CT simulator to design beams via multileaf collimator for 3D conformal radiation therapy: (a) Beam's eye view, in which the CTV is filled in red, the PTV is outlined by green contour, the collimator aperture is defined by red rectangle, and the organ at risk (eye) is filled with green; (b) Operator's eye view using segmented semitransparent surface rendering, in which the PTV is in red, the beam field on the surface in yellow, and the organ at risk in green. [Courtesy Dr. Georgios Sakas and Dr. Evelyn A. Firle, Fraunhofer Institute for Computer Graphics (IGD).] Please see online for color figure.

the beam design using a multileaf collimator would be very difficult, if not impossible. Moreover, the optimization of the beam design, in which many plans have to be investigated and compared, via either beam shapers or a multileaf collimator can only be possible with the aid of these functionalities that are provided by a CT simulator.

Protocol Export

Once targeted volumes are identified, beams and their fields are designed, the last step is to export the simulation results to a treatment machine. The simulation results consist of both textual and pictorial information. The textual information usually includes a list of identified structures and their types (targeted cancerous volume, surrounding critical normal organ or tissue, and, etc.), a list of beam setting up parameters corresponding to each targeted volume (machine angles, patient couch angles, isocenter coordinates, field markers, and etc.), and a list of sequential actions to be executed in the treatment delivery process. The pictorial information refers to the pictures and images, such as DRRs, that can be used in the treatment process for planning verification by comparing them with the portal images acquired on the treatment machine.

COMPUTED TOMOGRAPHY SIMULATION FOR ADVANCED RADIATION THERAPY

The methodology for designing radiation beams in a CT simulator introduced so far is quite straightforward. For example, with respect to a PTV, all beams are coplanar, and the intensity distribution within the field of each beam is constant. Furthermore, usually only two to three beams are engaged in treatment delivery. As shown in the left column of Fig. 9, such a very limited number of coplanar beams with uniform intensity distribution are generally not able to conform a 3D target volume tightly, no matter how accurately the beam and its field are designed. Hence, the conventional 3D conformal radiation therapy is at most a coarse approximation to the ideal 3D conformity.

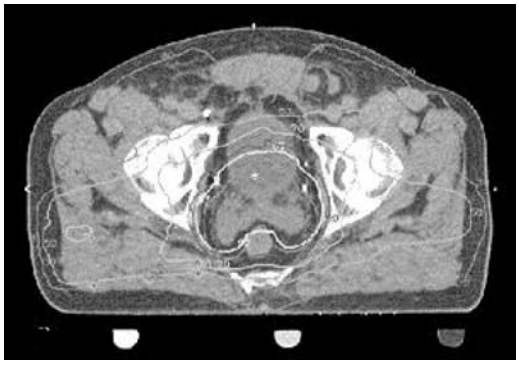


Figure 9. Schematic diagrams illustrates the mechanisms underlying conventional 3D conformal RT (a) and modern intensity modulated RT (b). The volume of the targeted tumor is concave, and the critical organ (spine cord) is within the concavity. The spine cord would undertake almost the same radiation dose as the target tumor volume if the conventional 3D conformal RT is administrated. However, with intensity modulation RT, the radiation dose to the spine cord can be substantially reduced to be below tolerable toxicity level. (Courtesy of Sabbe Mollio, Ph.D., University of California at Irvin.) Please see online for color figure.

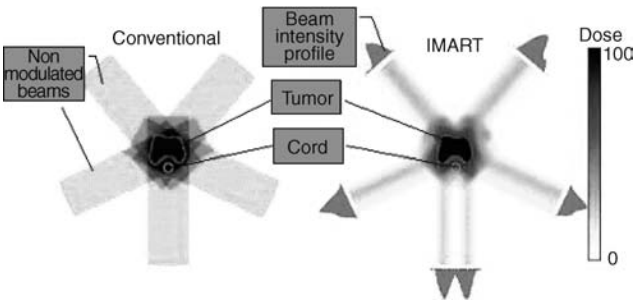


Figure 10. A CT image in transaxial plane shows a concave CTV outlined by the yellow boundary, the corresponding PTV contoured by the red boundary, and the nearby critical organ (spine cord) within the concavity. (Courtesy of Michelle Lee, Eleka AB.) Please see online version for color figure.

In reality, a targeted cancerous tumor can be much more complicated than just a regular and convex volume. For example, a target tumor can be irregular, concave with nearby normal critical organs or tissues within its cavity (see Fig. 10), or the targeted cancerous tumor even wraps itself around nearby normal organs or tissues. To make sure all cancerous tumors be included in a PTV, a physician has no choice but to take extra margins surrounding the identified CTV into account. Consequently, the PTV can be unfortunately quite large due to the conservative strategies exercised, resulting in an undesirable radiation dose to the surrounding normal organs or tissues, and significantly compromising the success of radiation therapy.

Given the complexity of the 3D shape of a targeted malignant tumor and its geometrical relationship with nearby normal structures, true 3D conformity is desired in RT to deliver the needed radiation dose accurately and avoid unnecessary damages to the cells of surrounding normal organs or tissues. Actually, the combination of the modern CT technology and 3D visualization methods has paved the way toward true 3D conformal radiation therapy. Aiming at improving the therapeutic index, tremendous efforts have been devoted in the radiation therapy community to exploring more effective RT solutions. What follows is a brief introduction to the exciting progresses made up to date.

Computed Tomography Simulation For Intensity Modulated Radiation Therapy

One of the innovations over conventional 3D conformal RT is the so-called intensity modulated radiation therapy (IMRT) (30–33). The two primary differences between the IMRT and conventional 3D conformal RT are (a) beams do not have to be coplanar; (b) the radiation intensity distribution within a beam is not uniform. In practice, noncoplanar beams can be readily realized by horizontally rotating the patient couch of a treatment machine. With respect to beam intensity, there exist two ways to modulate the distribution. One way is to partition the beam field into a few subfields, and various radiation blockers, such as

wedge, partial blocker, or compensator, are employed to modulate the intensity of each subfield. Such an IMRT implementation is called the "step-and-shoot" mode. The other way is to move the leaves of a computer-controlled multileaf collimator dynamically and individually. The intensity distribution within each beam field is modulated by appropriately opening and shutting the leaves of a multileaf collimator. Such an implementation is called the "sliding window" mode, because each leaf of a multileaf collimator is continuously sliding during the treatment delivery. The radiation beams in these two modes can three-dimensionally encompass an identified target volume with improved conformity, and deliver a radiation dose distribution with a significantly improved accuracy than that allowed by the conventional 3D conformal radiation therapy, no matter how complicated the shape of the identified target volume is.

To illustrate the mechanism of 3D conformity in IMRT, the 2D in-plane conformity achieved by intensity-modulated coplanar beams is shown in the right column of Fig. 9, while the 3D conformity using noncoplanar intensity-modulated beams is not hard to deduce. As the conformity significantly is improved, either an escalated radiation dose can be delivered to the target volume at the same dose level absorbed by surrounding normal organs or tissues, or the same dose can be delivered to the target volume with less radiation toxicity to surrounding structures. It means that, IMRT not only conforms a high dose delivery to the targeted tumor volume, but also minimizes the radiation dose to the nearby structures.

To achieve an improved 3D conformity, the IMRT requires accurate 3D geometrical delineation of an identified target volume and its surrounding structures, beam orientation, field partition, intensity modulation via blockers, wedges, compensators, or a multileaf collimator. It is not difficult to imagine that, without the previously described functionalities offered by a CT-simulator, the implementation of IMRT would be out of the question. In fact, the dependence of the IMRT upon the CT simulator is so profound that the CT simulator is no longer separable from a radiation treatment planning system for which the IMRT is designed and verified. At present, the boundary between a CT simulator and a treatment planning system is fading. More and more radiation treatment planning systems are incorporating the functionalities that used to be provided by a stand-alone CT simulator as an integrated part in the IMRT-based 3D conformal radiation therapy.

With an improved 3D conformity, the success of the IMRT-based 3D conformal RT relies on the geometry accuracy and reproducibility to a significantly larger extent than that in conventional 3D conformal RT. Any geometry uncertainty (6) due to beam setup (mechanical and optical allowances), imperfect reproducibility, or organ motion induced errors can significantly compromise the outcome of the radiation therapy, as the intensity modulated beams conform the target volume very compactly. In addition to 3D conformity, the homogeneity of the dose distribution delivered to a target volume is also a requirement of a successful radiation therapy. However, with an emphasis on the 3D conformity, the dose homogeneity is usually compromised, posing more challenges to the RT process. Meanwhile, the verification of the radiation dose distribution in IMRT becomes difficult, demanding much more complicated devices and phantoms to accomplish quality assurance (34,35). As a result, caution or discretion must be exercised in the clinic while a decision is being made on adopting IMRT for 3D conformal radiation therapy, although clinical reports showing favorable results of IMRT over conventional 3D conformal RT have been growing in the literature. In all fairness, not all cancer sites are suitable for the application of IMRT, and a general rule is that for a relatively regular target tumor, conventional 3D conformal RT is preferable, whereas IMRT is preferable in circumstances where the target tumor is irregular, concave with critical organs or tissues within the concavity, or even wraps itself around the nearby normal critical organs or tissues. Meanwhile, there may exist too much freedom in IMRT to conform an irregularly shaped target volume, leading to more difficulties in the optimization with respect to candidate treatment plans to achieve various objective functions while satisfying certain constraints. At present, numerous basic and clinical investigations are under way in the RT community to attain IMRT protocols for 3D conformal RT.

Computed Tomography Simulation For Tomotherapy

With the freedom in increasing the number of beams, specifying the beam shape and orientation, and modulating the beam intensity distribution, IMRT technology can be advanced even further (30,31). A more aggressive implementation of IMRT is the so-called tomotherapy (36–38). As implied by its name, tomotheraphy is inspired by the scanning mode of X-ray CT that has been extensively utilized in diagnostic imaging over the past three decades. Interestingly, the evolution of tomotherapy is similar to that of diagnostic CT. The geometry of an initial tomotherapy system is virtually the same as that of a single slice CT scanner prior to the introduction of helical–spiral CT. In such a tomotherapy system, the target volume is irradiated by a small fan beam moving along an arc trajectory (37). Consequently, the number of beams is increased dramatically while the field size of each beam is accordingly decreased. At each angular position, the radiation intensity of the small fan beam is dynamically modulated by a binary multileaf collimator that is driven pneumatically. In a "slice-by-slice" fashion, a target volume can be irradiated with a significantly improved 3D conformity, provided that the thickness of the small fan beam is sufficiently small. Similar to the CT evolution from the circular scanning mode to the helical–spiral scanning mode, tomotherapy has evolved into helical–spiral tomotherapy, in which the patient couch proceeds at a constant speed while the radiation source is rotated in the gantry around a patient (36,38). At each angular position, the intensity of the small fan beam is temporally modulated by a pneumatically driven multileaf collimator using an appropriate opening and shutting scheme. Through helical–spiral scanning, the radiation dose can conform a targeted volume very tightly, regardless of whether it is regular or irregular, concave or convex, delivering an adequate treatment dose to the targeted volume while depositing a minimum dose to the surrounding normal organs and tissues.

As in conventional IMRT, tomotherapy, either in the arc or the helical–spiral mode, requires an accurate geometrical delineation of a target volume and a reliable reproducibility of the patient positioning. The latest helical–spiral tomotherapy system has even evolved into an "all-in-one" system: a radiation therapy machine that includes all the needed modules: CT simulator, treatment planning, treatment delivery, as well as a real CT scanner (called tomoimage CT scanner) using therapeutic photons as the tomographic imaging means. Although its contrast resolution is significantly degraded in comparison to that of a diagnostic CT scanner or an RT-dedicated CT scanner, the tomoimage CT scanner can provide acceptable spatial resolution for accurate patient repositioning, playing a role of portal imaging, such as DRR, used in the conventional 3D conformal RT to assure geometry accuracy and reproducibility. With self-consistency among anatomic imaging, treatment planning, treatment delivery, and verification, such an "all-in-one" tomotherapy system can significantly improve the geometrical integrity, which is one of the most critical challenging tasks of IMRT (36,38).

Computed Tomography Simulation For Stereotactic Radiosurgery

In addition to 3D conformal RT, a state-of-the-art stereotactic radiosurgery (39–41) also relies entirely on the capabilities of a CT simulator in which either protons or high energy X-ray photons are employed. Rather than being administrated in a fractionated manner, stereotactic radiosurgery is a one-session or "lump sum" radiation treatment delivery procedure just like a surgery operation. Because of single treatment delivery, the radiation intensity must be elevated to the highest tolerable degree, and such a greatly elevated radiation intensity can damage cells in surrounding normal organs or tissues. Hence, stereotactic radiosurgery is usually employed in the management of small (2–3 cm) cancerous tumors. Moreover, with a radiation intensity that is greatly larger than that of 3D conformal RT, stereotactic radiosurgery demands a much more accurate geometric delineation of targeted tumors, and any geometrical uncertainty due to patient position, immobilization, and organ motion could result in treatment disaster. Consequently, stereotactic radiosurgery is only employed to cure small tumor in patient's head or neck, where patient immobilization devices can be applied reliably. In the early stage of stereotactic radiosurery, 2D imaging techniques, such as fluoroscopic angiography, were employed to identify target volumes and determine the number and orientation of radiation beams. Recently, similar to the evolution of 3D conformal RT, state-of-the-art stereotactic radiosurgery procedure increasingly rely on 3D patient data acquired by a CT scanner and the treatment planning in which the functionalities of a CT simulator are integrated.

THE FUTURE OF COMPUTED TOMOGRAPHY SIMULATOR

To achieve a full 3D conformity in radiation therapy, an ultimate solution would be the utilization of a very large number of pencil beams with their intensity being modulated instantly (42–44). However, a full 3D conformity demands perfect geometrical accuracy, since the dose gradient at the boundary of the 3D conformed target volume would be extremely sharp. Any geometrical uncertainties due to inaccurate patient position, unreliable immobilization, beam set up errors, misregistration between 3D patient data acquisition and treatment delivery, and particularly organ motion, could lead to treatment failure: underdosing-to-target volumes or overdosing to nearby normal critical organs or tissues, or both.

All the 3D conformal RT techniques introduced thus far are implemented in a "forward" way that is, once a set of 3D patient data is acquired, the simulation, treatment planning, planning verification, and delivery are implemented sequentially. In the forward mode, the simulation and treatment planning processes are conducted in a semiautomatic and iterative way with the involvement of radiation oncologists, physicists, therapists, and dosmetrists. To a great extent, the success of such a strategy is dependent on the experience or skill of the physicians engaged. Encouraged by the vigorous technological progresses made over the past three decades, it is believed that modern 3D conformal RT will continue progressing with the development of medical imaging and 3D visualization technologies: delivering treatment to target volumes with a full 3D conformity and adequate dose while surrounding normal structures are minimally damaged. It is not hard to imagine that future RT would not achieve an optimized solution if the treatment planning is only performed in the forward mode, because the parameter space to be exhaustively searched in the optimization process is really large.

Recognizing the complexity of modern IMRT, a more reasonable way to accomplish radiation treatment planning is to treat it as an inverse problem. Starting from specifications on a desirable radiation dose distribution and certain constraints, such as the avoidance of critical anatomic structures, inverse treatment planning can be performed based on a set of patient CT data with the help of a CT simulator (30,45). There exist two major tasks in the inverse treatment planning: (a) selection of an objective function and constraints; (b) development of an efficient algorithm to solve the optimization problem (46). A number of objective functions and constraints, such as dose (47), dose-volume (48,49), equivalent uniform dose (50,51), generalized equivalent dose (52), biological indices (53), and their combinations, have been proposed to date (54). Meanwhile, a few optimization approaches, such as simulated annealing (55), gradient (48), active set (56), genetic (57), maximum likelihood (58), dynamically penalized likelihood (59), and fuzzy logic (60) algorithms, have been investigated (54). It is underlined that inverse treatment planning is mathematically an ill-posed optimization problem, and that certain regularization techniques ought to be used in the optimization process (46). Considering the inhomogeneous dose distribution associated with a full 3D conformity, the simulation process for the inverse treatment planning is significantly different from that for the forward treatment planning, but the basic functionalities, such as target volume defining and 3D visualization of a CT simulator, are still the same. Currently, inverse radiation

treatment planning is still an open research and development area that has attracted the major attention of the RT community.

The ultimate solution to guarantee geometry accuracy and reproducibility is to track targeted volumes and their motion dynamically for a full 3D conformity during the radiation treatment delivery. A tracking process can be realized through gating techniques employing various mechanical or optical sensors while a patient is scanned by an RT dedicated CT scanner (61–63). This is related to the concept of four-dimensional (4D) CT, by which variation of a 3D model of a patient is revealed instantly. Such a 4D planning strategy can be implemented either off- or on-line. In the off-line mode, a target volume and its surrounding normal organs or tissues are tracked by gating during the data acquisition by an RT-dedicated CT scanner, and the motion of the patient and organs are recorded and exported to a treatment machine for radiation delivery. In the on-line mode, a patient is scanned by an RT-dedicated CT scanner during the treatment delivery. Apparently, the on-line mode needs to integrate an RT-dedicated CT scanner into a treatment machine (64,65). It is believed that with the development of the 4D CT technology, the 3D conformal RT can be administrated in an adaptive and well-controlled manner, leading to a significantly improved therapeutic index.

The majority of simulation in treatment planning for RT is currently implemented based up on 3D patient datasets acquired by a CT scanner, because of its merits in data acquisition, image generation, superior spatial resolution, as well as the capability of estimating the electronic density for dose calculation and verification. In addition to CT scanners, other modern 3D imaging modalities, such as magnetic resonance imaging (MRI), positron emission tomography (PET), and ultrasonic imaging, can also be incorporated into a virtual simulator for 3D conformal RT. There is no doubt that, in the predictable future CT will remain the modality of chioce for 3D conformal RT. However, with the development of modern image registration and data fusion techniques, images acquired by MRI, PET, and ultrasound are becoming more and more relevant, suggesting chances for them to be utilized in 3D conformal RT. Finally, it should be pointed out that all the CT simulation and virtual simulation techniques we have covered in this article are also applicable for RT using other high energy particles, such as protons and neutrons (5).

BIBLIOGRAPHY

1. Wright K, et al. Field shaping selective protection in megavolt radiation therapy. Radiology 1959;72:101.
2. Tsien K. The application of automatic computing machines to radiation treatment planning. Br J Radiol 1955;28:432.
3. Reinstein Le, et al. A computer-assisted three-dimensional treatment planning system. Radiology 1978;127:259–264.
4. McShan DL, et al. A computerized three-dimensional treatment planning system utilizing interactive colour graphics. Br J Radiol 1979;52:478–481.
5. Smith Rp, Mckenna WG. The Basics of Radiation Therapy. In: Abeloff MD, et al. editors. Clinical Oncology 3rd ed. Elsevier; Churchill Livingstone: 2004.
6. Jones B, et al. United Kingdom radiation oncology 1 conference (UKRO 1): Accuracy and uncertainty in radiotherapy. Br J Radiol 2002;75:297–305.
7. Sherouse GW, et al. Virtual simulation: concept and implementation. The 9th International Conference of the use of computers in radiation therapy (ICCR). North Holland Publishing Co.; The Netherland: 1987.
8. Sherouse GW, Novins K, Chaney EL. Computation of digitally reconstructed radiographs for use in radiotherapy treatment design. Int J Radiat Oncol Biol Phys 1990;18:651–658.
9. Sherouse GW, Bourland JD, Reynolds K. Virtual simulation in the clinical setting: some practical considerations. Int J Radiat Oncol Biol Phys 1990;19:1059–1065.
10. Mutac S, et al. Quality assurance for computed-tomography simulators and the computed-tomography-simulation process: Report of the AAPM radiation therapy committee Task Group No. 66. Med Phys 2003;30:2762–2792.
11. Gerber RL, Purdy JA. Quality assurance procedures and performance testing for CT-simulators. In: Purdy JA, Starkschall G, editors. A Practical Guide to 3-D Planning and Conformal Radiation Therapy. Advanced Medical Publishing, Inc.; Middleton (WI): 1999.
12. Conway J, Robinson MH. CT virtual simulation. Br J Radiol 70 (Suppl.):1997;S106–S118.
13. Aird EGA, Conway J. CT simulation for radiotherapy treatment planning. Br J Radiol 2002;75:937–949.
14. Fraass B, et al. American Association of Physicists in Medicine Radiation Therapy Committee Task Group 53: Quality assurance for clinical radiotherapy treatment planning. Med Phys 1998;25:1773–1829.
15. Van Dyk J, Taylor JS. CT-simulators. In: Van Dyk J. editor. The Modern Technology for Radiation Oncology: A Compendium for Medical Physicists and Radiation Oncologists. Medical Physics Publishing; Madison (WI): 1999.
16. Kutcher GJ, et al. Comprehensive QA for radiation oncology: report of AAPM Radiation Therapy Committee Task Group 40. Med Phys 1994;21:581–618.
17. AAPM, Report No. 39. Specification and Acceptance Testing of Computed Tomography Scanners. American Institute of Physics; New York: 1993.
18. Kalender WA. Computed Tomography: Fundamentals, System Technology, Image Quality, Applications. 2nd ed. Wiley; New York: 2004.
19. Hsieh J. Computed Tomography: Principles, Design, Artifacts, and Recent Advances. SPIE Press; Bellingham (WA): 2003.
20. Wang G, Crawford CR, Kalender WA. Multi-row-detector and cone-beam spiral/helical CT. IEEE Trans Med Imag 2000; 19:922–929.
21. Tang X, Hsieh J. A filtered backprojection algorithm for cone beam reconstruction using rotational filtering under helical source trajectory. Med Phys 2004;31:2949–2960.
22. Tang X, Hsieh J, Nilsen RA, Dutta S. A helical cone beam filtered backprojection (CB-FBP) reconstruction algorithm using three-dimensional (3D) view weighting. SPIE Proc 2004;5535:577–587.
23. Langmack KA. Portal Imaging. Br J Radiol 2001;74:789–804.
24. Verrellen D, Linthout N, Berge DVD, Bel A, Storme G. Initial experience with intensity modulated therapy for treatment of the head and neck region. Int J Radio Oncol Bio Phys 1997;39:99–114.
25. Grant W, Woo SY. Clinical and financial issues for intensity-modulated radiation therapy delivery. Semin Radia Oncol 1999;9:99–107.
26. Korba A, et al. Pseudoblocks and portal localization. Radiology 1977;122:260–261.
27. Convery DJ, Rosenbloom ME. The generation of intensity modulated fields for conformal radiotherapy by dynamic collimation. Phys Med Biol 1992;37:48–59.

28. Stein J, Bortfeld T, Dorschel B, Schlegel W. Dynamic X-ray compensation for conformal radiotherapy by means of multileaf collimation. Radiother Oncol 1994;32:163–173.
29. Boyer AL, Yu CX. Intensity-modulated radiation therapy with dynamic multi-leaf collimators. Semin Radiat Oncol 1999;32:48–59.
30. Nutting C, Dearnaley DP, Webb S. Intensity modulated radiation therapy: a clinical review. Br J Radiol 2000;73:459–469.
31. Intensity Modulated Radiation Therapy Collaborative Working Group, Intensity-modulated radiotherapy: current status and issues of interest. Int J Radiat Oncol Biol Phys 2001; 51:880–914.
32. Purdy JA. Advances in three-dimensional treatment planning and conformal dose delivery. Semin Oncol 1997;24:655–672.
33. Boyer AL, Xiang L, Xia P. Beam shaping and intensity modulation in modern technology of radiation oncology. Van Dyk J, editors. Modern Technology of Radiation Oncology. Medical Physics Publishing; Madison (WI): 1999.
34. Wang X, et al. Dosimetric verification of intensity-modulated fields. Med Phys 1996;23:317–327.
35. Xing L, et al. Dosimetric verification of a commercial inverse treatment planning system. Phys Med Biol 1999;44:463–478.
36. Mackie TR, et al. Tomotherapy: a new concept for the delivery of conformal radiotherapy. Med Phys 1993;20:1709–1719.
37. Yu CX. Intensity-modulated arc therapy with Dynamic multileaf collimation: an alternative to tomotherapy. Phys Med Biol 1995;40:1435–1449.
38. Mackie TR, et al. Tomotherapy. Semin Radiat Oncol 1999; 9:108–117.
39. Svenssen R, Lind BK, Brahme A. A new compact treatment unit design combining narrow pencil beam scanning and segmental multileaf collimation. Radiothr Oncol 1999;51:S21.
40. Schweikard A, Tombropoulos R, Adler JR. Robotic radiosurgery with beams of adaptable shape. In: Ayache N, editor. Computer Vision and Robotics in Medicine. Springer-Verlag; Berlin (Heidelberg): 1995.
41. Webb S. Conformal intensity-modulated radiotherapy (IMRT) delivered by robotic linacs—testing IMRT to the limit? Phys Med Biol 1999;44:1639–1654.
42. Woo SY, et al. A comparison of intensity modulated conformal therapy with a conventional external beam steoreotactic radiosurgery system for the treatment of single and multiple intracranial lesions. Int J Radia Oncol Biol Phys 1996; 35:593–597.
43. Kramer BA, et al. Dosmetric comparison of stereotactic radiosurgery to intensity modulated radiotherapy. Radia Oncol Invest 1998;6:18–25.
44. Cardinale RM, et al. A comparison of three stereotactic radiosurgery techniques: arcs vs non-coplanar fixed fields vs intensity modulation. Int J Radia Oncol Bio Phys 1998;42: 431–436.
45. Botfeld T. Optimized planning using physical objects and constraints. Semin Radiat Oncol 1999;9:20–34.
46. Chvetsov AV, Calvetti D, Sohn JW, Kinsella T. Regularization of inverse planning for intensity-modulated radiotherapy. Med Phys 2005;32:501–514.
47. Sauer OA, Shepard DM, Mackie TR. Application of constrained optimization to radiotherapy planning. Med Phys 1999;26:2359–2366.
48. Spirou SV, Chui CS. A gradient inverse planning algorithm with dose-volume constraints. Med Phys 1998;25:321–333.
49. Hristov DH, Stavrev P, Sham E, Fallone BG. On the implementation of dose-volume objectives in gradient algorithms for inverse treatment planning. Med Phys 2002;29:848–856.
50. Wu Q, Mohan R, Niemierko A, Schmidt-Ullrich R. Optimization of intensity-modulated radiotherapy plans based on the equivalent uniform dose. [comment]. Int J Radia Oncol Biol Phys 2002;52:224–235.
51. Das S, et al. Beam orientation selection for intensity-modulated radiation therapy based on target equivalent uniform dose maximization. Int J Radia Oncol Biol Phys 2003;55:215–224.
52. Choi B, Deasy JO. The generalized equivalent uniform dose function as a basis for intensity-modulated treatment planning. Phys Med Biol 2002;47:3579–35894.
53. Wang XH, et al. Optimization of intensity 3D conformal treatment plans based on biological indices. [comment]. Radiothre Oncol 1995;37:140–152.
54. Baydush AH, Marks LB, Das SK. Penalized likelihood fluence optimization with evolutionary components for intensity modulated radiation therapy treatment planning. Med Phys 2004;31:2335–2343.
55. Webb S. Optimization by simulated annealing of three-dimensional conformal treatment planning for radiation fields defined by a multileaf collimator. Phys Med Biol 1991;36:1201–1226.
56. Hristov DH, Fallone BG. An active set algorithm for treatment planning optimization. Med Phys 1997;24:1455–1464.
57. Ezzell GA. Genetic and geometric optimization of three-dimensional radiation therapy treatment planning. Med Phys 1996;23:293–305.
58. Olivera GH, et al. Maximum Likelihood as a common computational framework in tomotherapy. Phys Med Biol 1998;43: 3277–3294.
59. Llacer J, Solberg TD, Promberger C. Comparative behavior of the dynamically penalized likelihood algorithm in inverse radiation therapy planning. Phys Med Biol 2001;46:2637–2663.
60. Yan H, Yin F, Guan H, Kim JH. Fuzzy logic guided inverse treatment planning. Med Phys 2003;30:2675–2685.
61. McKezie AL. How should breathing motion be combined with other errors when drawing margins around clinical target volumes? Br J Radiol 2000;73:973–977.
62. Bergstrom P, Lofroth PO, Widmark A. High precision conformal radiotherapy (HPCRT) of prostate cancer—a new technique for exact positioning of the prostate at the time of treatment. Int J Radiat Oncol Biol Phys 1998;42:305–311.
63. The BS, et al. Intensity modulated radiation therapy (IMRT) following prostateectomy: more favorable acute genitourinary toxicity profile compared to primary IMRT for prostate cancer. Int J Radiat Oncol Biol Phys 2001;49:465–472.
64. Matsinos E. Current status of the CBCT project at Varian. Proc SPIE 2005.
65. Colbeth RE, Roos PG, Mollov IP. Flat panel CT detector for sub-second volumetric scanning. Proc SPIE 2005.

See also PHANTOM MATERIALS IN RADIOLOGY; RADIATION THERAPY SIMULATOR.

COMPUTED TOMOGRAPHY, SINGLE PHOTON EMISSION

FREDERIC H. FAHEY
Children's Hospital Boston

HARVEY A. ZIESSMAN
Johns Hopkins University

INTRODUCTION

Single photon emission computed tomography (or SPECT) provides a three-dimensional (3D) representation of the distribution within a patient's body of a radiopharmaceutical that was given as part of a diagnostic nuclear

medicine study. Diagnostic nuclear medicine provides a unique way of imaging physiology and function. One can administer to a patient a pharmaceutical with a radioactive marker, and then use external detectors to determine where that "radiopharmaceutical" has distributed within the patient's body. The distribution of the radiopharmaceutical in the body depends on its specific biology. For example, suppose a patient is given a small amount of radioactive iodine (e.g., iodine-131 or ^{131}I). Since the thyroid naturally metabolizes iodine, some portion of the radioactive iodine will be preferentially incorporated into the thyroid with the rest going to other organs with lower concentrations. The amount of ^{131}I that will go to the thyroid will depend on whether the thyroid is functioning in a normal, hyperactive, or hypoactive fashion. By acquiring an image from this patient, one can determine whether certain regions of the thyroid are more active than others. For example, there may be hyperactive nodules within the thyroid. In this manner, nuclear medicine provides a unique opportunity to view the patient's physiology and not just the anatomy. Devices known as gamma cameras are used to provide images of the *in vivo* distribution (i.e., the distribution within the body) of the radiopharmaceutical. From these image data, the patient's specific physiology can be inferred.

The images produced by gamma cameras are two dimensional (2D) representations of a 3D object. In some cases, this is adequate to interpret the study. However, in many cases, the ambiguity introduced by activity in the overlying and underlying tissue can make it very difficult to infer the *in vivo* distribution of the radiopharmaceutical appropriately. In these cases, a 3D representation is necessary. Single photon emission computed tomography provides such a 3D representation. For example, SPECT can make it easier to determine whether the activity is reduced in the basal or apical aspect of the inferior wall of the myocardium or whether the activity in a tumor seen in the chest is in a rib or in the lung. As will be discussed, the 3D SPECT images are generated by a computer from a series of images taken about the patient at different angles. Since the photons used to generate the images are emitted from within the patient's body, SPECT is considered *emission* computed tomography, which is in contrast to *transmission* computed tomography (CT) where the X-ray photons emanate from an X-ray tube and are transmitted through the patient. In the early development of SPECT, the word "single" was used to distinguish SPECT from positron emission tomography (PET), which uses two photons to localize each event and requires different instrumentation. For these reasons, the generation of a 3D representation of the *in vivo* distribution of radiopharmaceutical that is not a positron emitter is referred to as single photon emission computed tomography or SPECT.

In 1963, a method of nuclear tomographic imaging was developed by Kuhl and Edwards (1). This method used a specially designed scanning device along with a processing method called simple backprojection to generate its 3D images. This early research in SPECT predates Hounsfield's work in CT by ~10 years. Kuhl and Edwards (2) subsequently developed a dedicated SPECT device for imaging the brain known as the Mark IV scanner. During the 1970s, several devices were developed that were dedicated to SPECT, specifically for brain imaging. However, techniques were also being investigated to use the gamma camera that was used for all other nuclear imaging applications for SPECT as well. This required the gamma camera to rotate around the patient's body in order to acquire different angular views. Keyes et al. (3) developed the first prototype of the rotating gamma camera by mounting a standard gamma camera head to the gantry of a decommissioned cesium-137 radiation therapy unit. At about the same time, Jaszczak et al. (4) developed a commercial version of the rotating gamma camera. Although the development of dedicated SPECT devices continues, by far the majority of devices used clinically are rotating gamma cameras. For this reason, this chapter will devote most of its attention to the use of the rotating gamma camera.

SPECT DATA ACQUISITION AND PROCESSING

For SPECT data acquisition, a series of images are obtained at a number of different viewing angles. Typically, these images are acquired in a circular arc about the patient with the axis of rotation parallel to the long axis of the patient's body. This acquisition geometry is shown in Fig. 1. The images acquired at each viewing angle are referred to as "projection images" or "projections". These projections are acquired at a number of evenly spaced viewing angles over either a 360 or a 180° arc. There is a minimum number of projections that will assure adequate angular sampling at the periphery of the object being imaged which will lead to high quality, tomographic images. For SPECT, this number is between 50 and 150, depending on the size of the object and the spatial resolution of the system. The gamma camera is usually equipped with a multihole, parallel-hole collimator that assures that the photons interacting in the radiation detector of the gamma camera traveled from the patient to the detector on a ray that is perpendicular to the detector surface. Thus, if one knows where the photon interacted in the detector, one can assume that the photon was emitted from a point that was along this ray. In Fig. 1, we refer to this ray as the "line of origin".

Consider a small, high contrast feature in the center of a cylindrical object that contains some radioactivity as shown in Fig. 2. We can acquire some number of projections (3 in the example shown in Fig. 2). If we consider the tomographic plane passing through the center of the tumor, each of the projections will look reasonably similar with a single intensity corresponding to the tumor. The process by which we generate a tomographic image from this series of projections is referred to as "tomographic reconstruction". We can generate a tomographic image of this simple object by filtering and adding the projections, each oriented at the angle associated with that specific projection. This reconstruction technique is referred to as "filtered backprojection". In many cases, additional smoothing is applied using a windowing filter that can control the sharpness and noise associated with

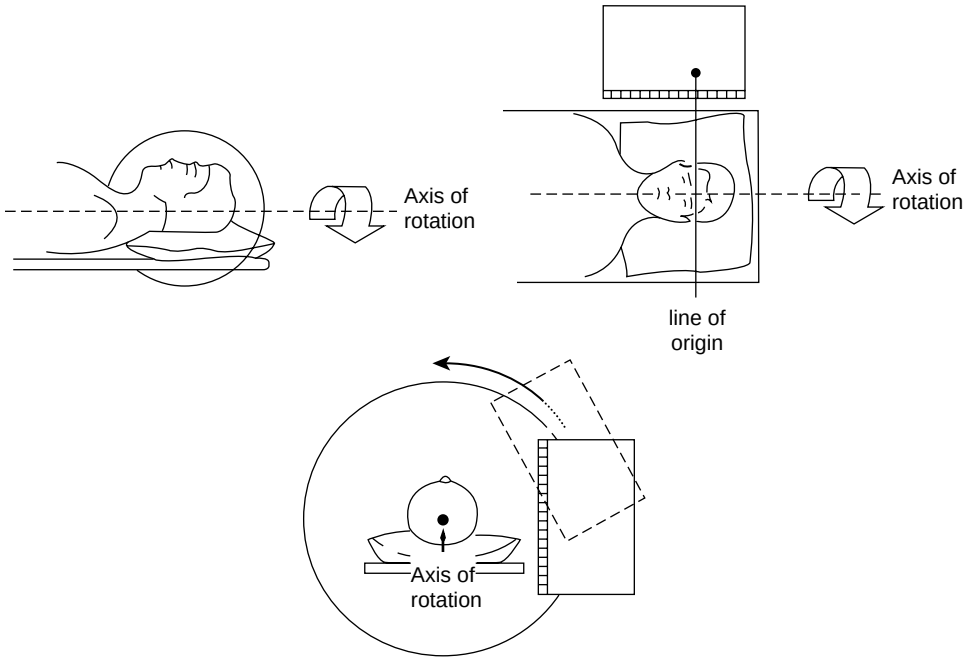


Figure 1. The geometry for SPECT data acquisition with a rotating gamma camera is shown. The camera rotates acquiring projection images at different angles about the patient. These data are subsequently reconstructed into cross-sectional slices that indicate the 3D *in vivo* distribution of the radiopharmaceutical within the patient. (Reprinted from Henkin R., editor, Nuclear Medicine. P 235, Copyright © 1996 with permission from Elsevier.)

the reconstructed image. The windowing functions that are typically used include the Hanning, Hamming, Shepp-Logan, and Butterworth filters. Depending on the signal and noise content of the underlying projection data, one can choose an appropriate windowing filter for the best image quality. Figures 3a–c shows three images that are filtered backprojection reconstructions of the same raw data (in this case, a brain study) using three different windowing filters (sharp filter, moderate filter and smooth filter, respectively). One can note the differences in image quality that one can attain by simply varying the filtering.

In addition, different clinical applications (e.g., brain vs. cardiac SPECT) may require different reconstruction filters. Therefore, it is very important to select the most appropriate reconstruction filter for each clinical application of SPECT.

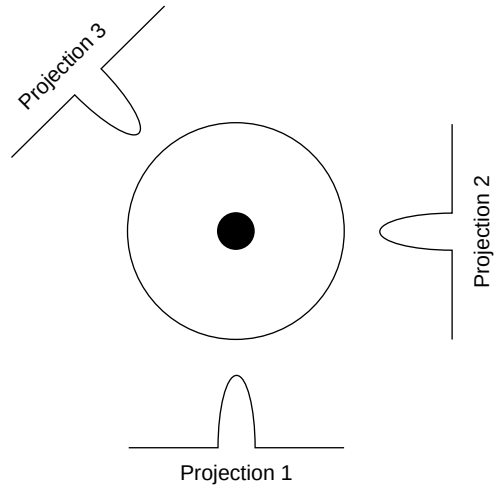


Figure 2. The object in the middle has a central region of high intensity. Three projections about the object are also shown. In a typical SPECT study, 50–150 projections are acquired about the object over 180 or 360°.

Although filtered backprojection has traditionally been the most common means of SPECT reconstruction, iterative reconstruction methods are currently available on newer SPECT systems. These methods utilize a "feedback" approach to generate the tomographic data. These methods start with an initial estimate of the object. This estimate may assume the object is totally uniform or it may use a filtered backprojection reconstruction of the object. From this estimate, a series of projections are calculated along the same viewing angles as the "real" projections, that is the acquired, raw data. If the estimate is close to the true object, then the calculated projections would be very similar to the real projections. If they are not similar, then the variations between the two are determined (either as a ratio or a difference) and used to alter the initial estimate. The process is then repeated. A new set of calculated projections are generated and again compared to the real projections. Presumably, each iteration provides a better estimate of the true object. In other words, the estimates should "converge" to a good representation of the true object. Typically, some statistic such as the "likelihood" or the "entropy" is used to determine how well the method is converging or when to stop the iterative process. In many cases, it can take tens or even hundreds of iterations before the reconstructed estimate converges. Thus, these methods have traditionally been quite slow, much slower than filtered backprojection. However, an advantage of these methods is the ability to take into account the physics of data collection and the statistics of the noise in the images to provide a more accurate reconstruction. Another advantage is that these methods are not as susceptible to some of the artifacts that one encounters in filtered backprojection. Much of the research in this area has centered on the development of methods that are more efficient or converge more

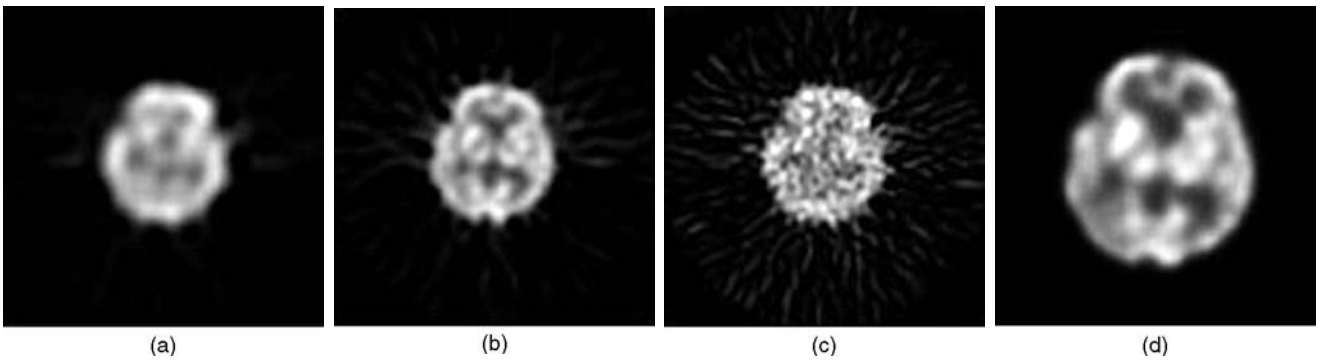


Figure 3. The SPECT brain study with four different reconstructions. Figures 3a–c are reconstructed with three different filters (smooth, moderate and sharp filter, respectively). Figure 3d is reconstructed with an iterative method known as OSEM.

quickly. With the increasing speed of computers and the development of more efficient iterative algorithms, the clinical application of these methods has become feasible and many newer SPECT systems provide iterative reconstruction methods as an option. Figure 3 shows a comparison with the same raw SPECT data reconstructed with both an iterative method (3d) and filtered backprojection (3a–c).

Consider a radiopharmaceutical that basically distributes uniformly within the body. Those photons that are emitted from deep within the body will have to travel through more tissue to reach the gamma camera than those emitted on the periphery of the body. In turn, those photons that must traverse more tissue are more likely to be absorbed within the body and, therefore, are less likely to be detected than those emitted at the periphery. In other words, the signal from deep-seated tissues is "attenuated" as compared to the surface tissues due to self-absorption of the photons within the object. To measure the amount of activity at different locations within the object accurately, one must apply a correction for photon attenuation. There are two basic approaches to attenuation correction in SPECT, one that assumes that the attenuation is uniform within the object, and one that considers the fact that different tissues may have varying absorption characteristics making the attenuation nonuniform.

In much of the body, the composition and density of the soft tissue is basically constant and thus one can assume a single attenuation property for all of the tissue. This assumption is apt, for example, in brain imaging. To apply uniform attenuation correction, one need only know where the body outline is located relative to where the radiopharmaceutical has distributed. For a particular point of interest within the object, the mean distance to the body outline is estimated and, using this estimate, the expected amount of photon attenuation from that point is determined. By multiplying the amount of radioactivity at that point by the reciprocal of the calculated photon attenuation at that point, one can estimate the amount of signal one would have received from that location if there were no photon attenuation. This correction can, in turn, be applied for every location within the object. If this correction is applied to the object with a uniform radioactivity distribution, a uniform signal throughout the object would be obtained. This method, referred to as the first-order Chang correction for photon attenuation, is the most common approach to applying uniform attenuation correction in clinical SPECT (5).

This uniform assumption, however, is not at all appropriate for the thorax where there is lung tissue, spine, and mediastinum in addition to soft tissue, all of which have very different attenuating properties. If a uniform attenuation assumption was used, one would overcorrect the regions of the lungs and undercorrect the regions near the spine and mediastinum. This is of particular concern for myocardial SPECT. For these reasons, no attenuation correction was applied traditionally for cardiac SPECT since no correction was considered better than a poor correction. However, over the past 10 years, a number of investigators have implemented various approaches to non-uniform attenuation correction. In these cases, one needs to not only know the body outline, but also must know the types of tissues within that outline and their attenuation characteristics. To determine this, one acquires a transmission image using an external, photon-emitting source. These data indicate which regions within the body are highly attenuating and which yield less attenuating. This knowledge is incorporated into the SPECT reconstruction process to correct for the nonuniform attenuation. A variety of ways have been developed for the acquisition of the transmission data, several of which will be discussed in the section on instrumentation. In addition to attenuation correction, several other corrections have been developed for SPECT in order to provide more quantitative reconstructed data including those for scatter and resolution recovery.

SPECT INSTRUMENTATION

SPECT requires the ability to acquire projection images from a number of viewing angles about the patient. Thus the gamma camera must be mounted onto a gantry that allows the camera to rotate about the patient in a circular or elliptical orbit. One of the limitations of the rotating gamma camera is the inherently low sensitivity of the system. Since the cameras must utilize absorptive collimation to determine the directionality of the interacting

Figure 4. Dual-detector SPECT camera. With this camera, the two detectors can be oriented at 180° to each other for whole body imaging (a) or 90° for cardiac imaging (b).

photons, only a very small fraction of the photons emitted from within the patient will actually be detected by the gamma camera. More stringent collimation can be used to improve the spatial resolution but at a cost of lower sensitivity leading to a higher level of noise in the images. One straightforward approach to improving the sensitivity is to increase the number of detectors. Thus, both dual-detector and triple-detector SPECT systems have been developed. Figure 4a shows a modern dual-detector SPECT system.

Since the heart is located in the left anterior portion of the chest and low energy radiopharmaceuticals, such as thallium-201, are routinely used in this application, it is common to acquire cardiac SPECT data only over 180° (from right anterior oblique to left posterior oblique) rather than over 360°, since most of the data that is acquired in the right posterior projection only adds noise and poor resolution to the image. This being the case, the use of two opposing detectors does not reduce the total imaging time, since one will still need to rotate the gantry over 180°. For this reason, a number of manufacturers have designed their dual-detector cameras such that the data can be acquired with the detectors either opposing each other (180° orientation) and in a 90° orientation as shown in Fig. 4b. This allows for the same amount of SPECT data to be acquired in half the time. Increasing the number to three detectors improves the sensitivity of the SPECT device even further. Such triple-detector devices are excellent for acquiring SPECT but lack the flexibility for other nuclear imaging and thus tend to be less popular than the dual-detector systems.

As discussed previously, a transmission scan must be acquired in order to perform non-uniform attenuation correction. Several different approaches have been developed for the acquisition of the transmission image. Collimated, radioactive line sources can be scanned over an area of the patient during the acquisition of the emission data. Since the radionuclide in the sources (^{153}Gd) emits a gamma ray with a slightly different photon energy than the radiopharmaceutical administered to the patient, the two data sets (emission and transmission) can be acquired simultaneously. In an alternate method, a series of smaller line sources are used. These also contain ^{153}Gd and thus again the emission and transmission data can be acquired simultaneously.

In the past several years, hybrid SPECT–CT systems have been developed. In these devices, a helical CT study is acquired in conjunction with the SPECT study on the same device. The CT scan is used to characterize the material within the body such that a non-uniform attenuation correction can be applied. It typically requires a transformation to be performed between the CT values and the attenuating coefficients for SPECT because the energies of the photons in CT are different that those for SPECT. In some cases, the device provides a diagnostic quality CT that can be interpreted either in conjunction with the SPECT study or independently. In the case of one SPECT–CT device, the CT provided is not of diagnostic quality, and it is used only for attenuation correction and gross anatomical correlation.

With the newest developments in molecular medicine has come the desire to image small animals such as rodents. Thus a number of investigators have developed methods of performing SPECT imaging in rodents with very high spatial resolution. One approach that is straightforward is the use of very small pinhole collimators. The size of pinhole is only ~1 mm as compared to the 4–6-mm pinholes that are typically used for clinical imaging. This method can provide SPECT images that have spatial resolution that is better than what is typical in clinical SPECT by almost a factor of 10. These high resolution approaches cannot be applied in clinical SPECT because these very small pinholes are very inefficient and thus an inordinate image time on the order of several hours would be required in order to obtain human images of sufficient image quality.

CLINCIAL APPLICATIONS IN SPECT

In this section, we will review several of the most common clinical applications of SPECT.

Cardiac

The most common clinical indication for SPECT is myocardial (cardiac) perfusion imaging. Atherosclerotic heart disease is manifested by coronary artery narrowing, which limits blood flow to the region of the heart supplied by that artery, producing chest pain or myocardial infarction. SPECT can confirm or exclude significant coronary disease. If abnormal, invasive coronary angiography may be performed in anticipation of intervention, for example, coronary angioplasty or bypass surgery.

The radiopharmaceutical, thallium-201 or newer technetium-labeled cardiac radiotracers, is delivered by the individual coronary arteries to the myocardium where it is extracted. The SPECT images depict the 3D blood flow to each region of the myocardium supplied by its coronary artery.

The study is performed in two stages, at rest and with stress. Treadmill exercise is the usual stress, although pharmacologic methods are used in those unable to exercise. In a normal heart, SPECT will show good blood flow at rest and stress. In a patient with a prior myocardial infarction, no blood flow will be seen in the nonviable region at either stage. A patient with significant coronary artery stenosis, without infarction, will have a normal rest study, but an abnormal exercise study (Fig. 5). Adequate blood flow and oxygen can be delivered to a resting heart even with a high grade coronary artery obstruction, however, the increased demand for oxygen and blood flow required at stress cannot be met and myocardial uptake will be decreased in the myocardium fed by that artery.

Tumors

SPECT with various radiopharmaceuticals provide valuable information regarding the extent of disease and distribution in the body. This information is used for initial diagnosis, staging, preoperative localization, and evaluating response to therapy.

Gallium-67 citrate (^{67}Ga) has been used for several decades for tumor imaging. It binds nonspecifically to a variety of tumors. Gallium-67 SPECT is used most commonly for imaging malignant lymphoma, a disease of lymphatic tissue seen in adults and children. Images depict the extent of disease and the effectiveness of therapy. This SPECT imaging is more accurate than CT or magnetic resonance imaging (MRI) for determining the effectiveness of therapy.

^{111}In OctreoScan is a somatostatin receptor peptide imaging agent, most useful for imaging tumors of neuroendocrine origin, (carcinoid, gastrinoma, neuroblastoma, pheochromocytoma, etc.) SPECT cross-sectional imaging makes it possible to see small tumors that are often not detected with CT or MRI.

^{111}In ProstaScint is radiolabeled monoclonal antibody directed against antigens on the surface of prostate cancer cells. It can detect the site of recurrence of prostate cancer,

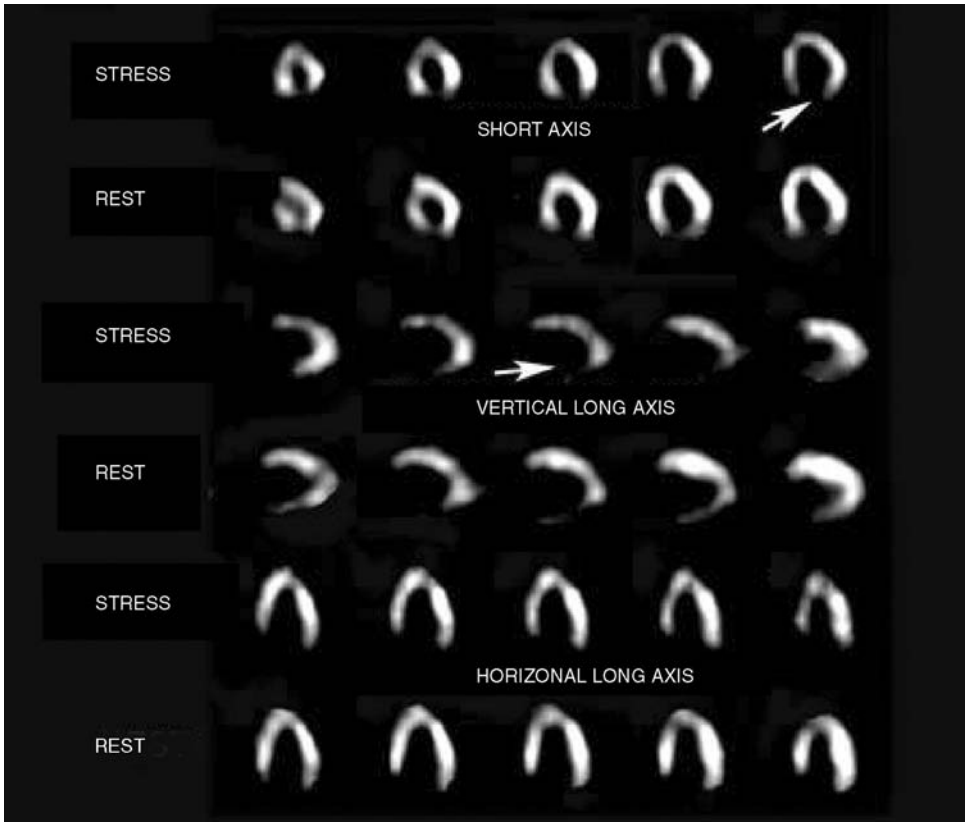


Figure 5. The SPECT exercise and rest cardiac study. Short axis, vertical long axis, and horizontal long axis cross-sectional sequential slices are shown. Arrows point to the inferior wall where there is no perfusion. Since this finding is unchanged between stress and rest, this is diagnostic of an inferior wall myocardial infarction.

suggested by a rising prostate serum antigen (PSA) level. Adequate imaging is not possible without SPECT because of the relatively high background activity.

Brain Imaging

SPECT is mandatory for brain imaging in order to visualize the various overlying and underlying convoluted regions of the brain.

Seizures are often caused by small regions of scar tissue in the brain. Patients not responding to drug therapy for their seizure disorder may be helped with resection of the seizure focus. Proper localization of the seizure site is critical for effective surgery. Brain wave studies (electroencephalogram or EEG) are only moderately successful in locating a seizure site. Even then, additional studies are needed for confirmation. The traditional method is to place electrodes on the surface of the brain and record electrical activity. However, this requires a neurosurgical operation and is associated with some morbidity.

SPECT brain blood flow radiopharmaceuticals, ^{99m}Tc hexamethyl propylene amine oxime (^{99m}Tc HMPAO) and ^{99m}Tc ethyl cysteinate dimer (^{99m}Tc ECD), show the distribution of blood flow in the brain. During seizure activity, the small area of the brain responsible has increased metabolism and increased blood flow. Between seizures, the abnormal site has decreased metabolism and blood flow. SPECT imaging can localize the seizure site by detecting these focal abnormal blood flow patterns. An example of a SPECT study in a seizure patient is shown in Fig. 6.

The second indication is to determine the cause for dementia. Characteristic patterns of abnormal perfusion are seen with certain types of dementia, for example, Alzheimer's disease, frontal lobe dementias, and multi-infarction dementia (strokes).

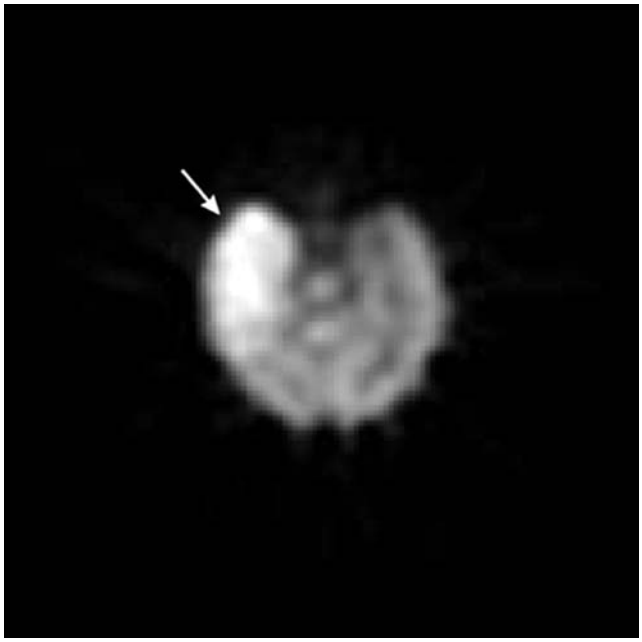


Figure 6. Brain SPECT study acquired during an epileptic seizure. The arrow indicates a region of high blood flow that corresponds to the part of the brain where the seizure originated.

Bone

SPECT is used for a variety of indications. It can help detect small sites of tumor, fracture, or infection not easily seen or localized with traditional two-dimensional bone scanning. Precise localization of a bone radiopharmaceutical uptake (e.g., ^{99m}Tc diphosphonate), can help differentiate benign from malignant processes, confirm small fractures as the cause of pain, and detect sites of infection.

SUMMARY

SPECT is a 3D imaging approach for evaluating the physiology and function of the patient. The patient is injected with a small amount of a radiopharmaceutical and a series of projection images are acquired at different angles about the patient. These data are reconstructed into a series of cross-section views of the *in vivo* distribution of the radiopharmaceutical. Depending on the radiopharmaceutical used, the nuclear medicine physician can infer essential information regarding the patient's physiologic condition. The rotating gamma camera is the most common device used to acquire SPECT data. Recent developments include the incorporation of CT into a hybrid SPECT–CT imaging device and the use of very small pinhole collimators for the imaging of small animals. With the advancements of molecular approaches to medicine, SPECT will continue to be a very important approach to medical imaging.

BIBLIOGRAPHY

1. Kuhl DE, Edwards RQ. Image separation radioisotope scanning. Radiology 1963;80: 653–661.
2. Kuhl DE, Edwards RQ. The Mark 3 Scanner: a compact device for multiple-view and section scanning of the brain. Radiology 1970; 96:563–70.
3. Keyes JW, Jr., Orlandea N, Heetderks WJ, Leonard PF, Rogers WL. The Humongotron—a scintillation-camera transaxial tomograph. J Nucl Med 1977; 18:381–387.
4. Jaszczak RJ, Murphy PH, Huard D, Burdine JA. Radionuclide emission computed tomography of the head with ^{99m}Tc and a scintillation camera. J Nucl Med 1977; 18:373–380.
5. Chang LT. A method for attenuation correction in radionuclide computed tomography. IEEE Trans Nucl Sci 1978; NS-25: 638–643.

Further Reading

Tsui BM. The AAPM/RSNA physics tutorial for residents. Physics of SPECT. *Radiographics* 1996; 16:173–183.

Miller TR. The AAPM/RSNA physics tutorial for residents. Clinical aspects of emission tomography. *Radiographics* 1996; 16:661–668.

Tsui BM, Frey EC, LaCroix KJ, Lalush DS, McCartney WH, King MA, Gullberg GT. Quantitative myocardial perfusion SPECT. *J Nucl Cardiol* 1998; 5:507–522.

Madsen MT. The AAPM/RSNA physics tutorial for residents. Introduction to emission CT. *Radiographics* 1995; 15:975–991.

King MA, Tsui BM, Pan TS. Attenuation compensation for cardiac single-photon emission computed tomographic imaging: Part 1. Impact of attenuation and methods of estimating attenuation maps. *J Nucl Cardiol* 1995; 2:513–524.

King MA, Tsui BM, Pan TS, Glick SJ, Soares EJ. Attenuation compensation for cardiac single-photon emission computed tomographic imaging: Part 1. Attenuation compensation algorithms. *J Nucl Cardiol* 1996; 3:55–64.

Groch MW, Erwin WD. SPECT in the year 2000: basic principles. *J Nucl Med Technol* 2000; 28:233–244.

Cherry SR, Sorenson JA, Phelps ME. Physics in Nuclear Medicine. 3rd ed. Philadelphia: Saunders: 2003. p 299–324.

See also ANGER CAMERA; NUCLEAR MEDICINE INSTRUMENTATION; POSITRON EMISSION TOMOGRAPHY; RADIOPHARMACEUTICAL DOSIMETRY.

COMPUTER-AIDED RADIATION DOSE PLANNING.

See RADIATION DOSE PLANNING, COMPUTER-AIDED.

COMPUTER-ASSISTED DETECTION AND DIAGNOSIS

ROBERT M. NISHIKAWA
The University of Chicago
Chicago, Illinois

INTRODUCTION

Since their discovery, X rays have been used to make images (radiographs) that allow the internal condition of the human body to be examined. Reading or interpreting these images has been without exception performed solely by humans until very recently. As our society depends more and more on automation or assistance from automated systems, so too has the interpretation of radiographs, although still in a very limited way at the present time.

In the early 1960s, researchers attempted to automate the interpretation of radiographs. The first published study was by Winsberg et al. who developed an automated computerized scheme to diagnosis breast cancer from a radiograph of the breast (mammogram) (1). This attempt, like others from that time period, was largely unsuccessful. Compared to current technology, these studies suffered from poor quality film digitizers (all images were recorded on film), insufficiently powered computers that had severely limited memory and storage space, and only a rudimentary armament of image processing, pattern recognition and artificial intelligence techniques. Clearly, the goal of automating the interpretation of radiographs was beyond the technical capabilities of that era.

After those initial attempts, there was a period of inactivity. In the late 1980s, a new approach was developed called computer-aided diagnosis (CAD) (2–5). The goal here was not to automate the interpretation of radiographs, but to give assistance to radiologists when they read images. The seminal paper was published in 1990 by Chan et al. (6). They conducted an observer study, where radiologists read mammograms once without the computer aid and once with the computer aid. For this study, the computer aid was a computer-aided detection (CADe) scheme that detected microcalcifications on mammograms. Microcalcifications are tiny deposits of calcium that can be an early indicator of breast cancer. Chan et al. (2) found a statistically significant improvement in the performance of radiologists in detecting microcalcifications when the radiologists used the computer aid. This study was the first CAD algorithm of any kind shown to be a benefit to radiologists and it validated the concept of the computer as an aid to the radiologist. It has spurred CAD research in mammography and in other organs, and in other imaging modalities.

Since that study, the field has grown rapidly. There are now at least four commercial systems available [two for detecting breast cancer from mammograms, and one each for detecting lung cancer from computed tomography (CT) scans and from chest radiographs]. Further, CAD is being developed for a wide variety of body parts (breast, thorax, colon, bone, liver, brain, heart, vasculature, and more) and several different imaging modalities [principally, radiography, magnetic resonance imaging (MRI), CT, and ultrasound]. Since mammography is the most mature area of CAD development, most of the examples used in this article to illustrate the principals of CAD will be drawn from mammographic CAD.

Mammography can detect breast cancer before it appears clinically (i.e., before it can be palpated or some other physical sign appears, such as a nipple discharge or breast pain). It has been shown to reduce breast cancer mortality. Typically, when a woman receives a mammogram, two X-ray images from two different angles are taken of each breast. One view is taken in the head to toe direction and the other at a 45° angle to the first. To detect breast cancer mammographically, radiologists look principally for two types of lesions. The first are masses, which are typically 0.5–5 cm in size. They appear slightly brighter than the surrounding tissue, but often have low contrast. Further, their appearance can be obscured by the normal structures of breast making them even more difficult to detect. Compounding their detection are normal tissues that appear to be a mass-like. This can occur if overlapping tissues in the breast are projected from a three-dimensional (3D) volume into the two-dimensional (2D) image. Taking two different views of each breast can help the radiologist to differentiate actual masses from a superposition of normal tissues, but actual masses are sometimes only seen in one view. The second type of lesion are microcalcifications, which are tiny deposits of calcium salts ranging in size from 10 μm up to 1 mm in size, however, mammographically microcalcifications < 200 μm are usually not detectable in a screening mammogram. Microcalcifications are difficult to detect because of their small size. Unlike masses that appear with low contrast, the principal limitation to detecting microcalcifications is the presence of image noise (i.e., a low signal/noise ratio).

Computer-aided diagnosis is well suited to mammography for several reasons. First, the only purpose for mammography is to detect breast cancer: Up to as many as 100 different abnormal conditions can be discovered from a chest X ray. Second, mammography is used as a screening modality, so the number of mammograms to be read is high. It is one of the most common radiological procedures performed. Third, breast cancer is difficult to

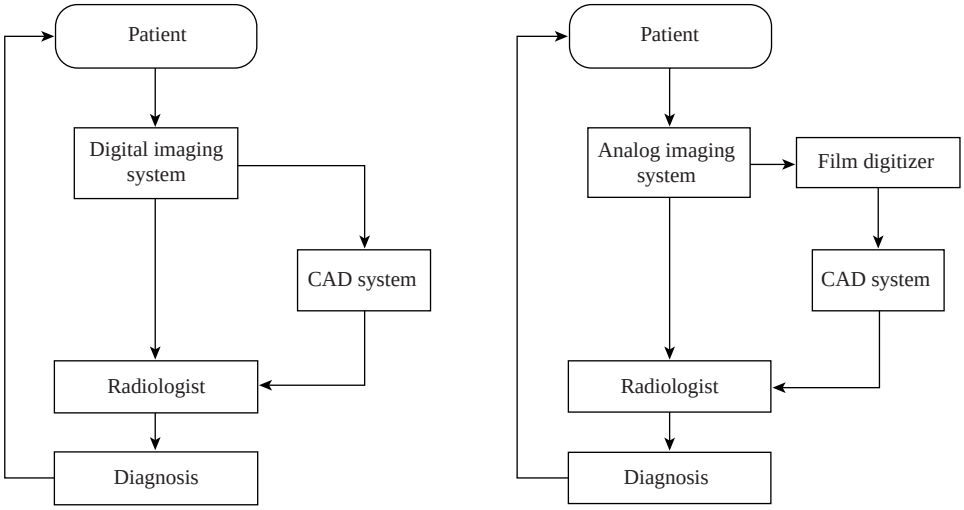


Figure 1. Schematic representation for the clinical implementation of CAD. If the image is acquired digitally (e.g., using CT, MRI, ultrasound, nuclear medicine, or a digital X-ray detector) the image can be used directly as input to a CAD scheme. If the patient is radiographed using an analog system (i.e., a screen-film system), then the image needs to digitized before it can be analyzed by the CAD system. In either case, the radiologist views the image or images and then views the output of the CAD systems, after which the radiologist makes their decision.

detect at an early stage mammographically. It requires that the radiologist careful check each image with a magnifying glass. Fourth, the prevalence is low: Only ~4 in every 1000 women screened has breast cancer. If a radiologist read 50 mammographic exams a day, they would see on average only 1 patient with breast cancer every week. Under such conditions, it is difficult to remain vigilant and alert. Fifth, mammography is one of the most common sources of malpractice lawsuits in the United States, so that missing a breast cancer can have severe consequences both for the patient and for the radiologist.

COMPUTER-AIDED DETECTION AND COMPUTER-AIDED DIAGNOSIS

What is Computer-Aided Diagnosis

The formal definition of CAD as first stated by Doi et al. is a diagnosis made by a radiologist who incorporates the output of a computerized analysis of the radiograph when making their decision (7). This definition emphasizes the distinction between CAD and automated diagnosis. The CAD is used as an aid to the radiologist, whereas the goal in automated diagnosis is to replace the radiologist.

There are in general two main types of CAD schemes. The first is CADe, where suspicious regions in the image are located. The second is CADx, where a suspicious region is classified (e.g., malignant vs. benign). Unfortunately, there is a possible confusion in nomenclature between the field as a whole and a specific type of CAD algorithm for distinguishing between different disease states (characterization or classification). To avoid this problem, the term CADe will be used for computer-aided detection, CADx for computer-aided diagnosis (classification), and CAD when referring to the whole field of study, which encompasses both CADe and CADx.

How Does Computer-Aided Diagnosis Work?

In CAD, the computer is used as an aid to the radiologist. It provides a second opinion to the radiologist. Double reading of mammograms has shown to increase the detection of breast cancer by as much as 15% (8,9). However, double reading is expensive since two highly trained individuals must read the images instead of one, and it is logistically difficult to implement efficiently. As a result, double reading is not commonly practiced, especially in the United States. It is believed that CAD could be an effective and efficient method for implementing double reading (10).

Figure 1 shows schematically how CAD can be implemented clinically. A radiograph is made of the patient. If the radiograph was acquired digitally, it can be sent directly to a CAD system for analysis. If the radiograph was recorded on film, then the image needs to be digitized first. After the computer analyzes the image, the output of the CAD system is displayed to the radiologist. The radiologist then views this information and considers their personal opinion with the computer output before giving a diagnosis. In all cases, the radiologist has the final diagnostic decision; the computer acts only as an aid. The CADe schemes have been compared to a spell checker tool in word processing software. In situations where double reading is employed, that is, each case is reviewed by two radiologists in a sequential manner, each radiologist may use the computer aid independently. Even in situations where double reading by two radiologists is employed, CADe can still detect cancers that both radiologists missed (11). It is also possible to modify the double reading paradigm when CAD is used. The first reader after reading with CADe can assign only cases that are not either clearly normal or clearly abnormal for the second reader (12).

There are several different methods of displaying the CAD output to the radiologist. Figure 2 shows a CT scan that is annotated with a circle indicating an area that the computer deemed suspicious. This is typical for a CADe system. Different symbols are used by different commercial systems. Figure 3 shows a chest radiograph that is annotated with different symbols to indicate different types of disease states. The symbols also have different sizes indicating the severity of the disease. This type of display can be used for both CADe and CADx systems. Figure 4 shows a simple interface that can be used by a CADx system in

Figure 2. Typical output of a CADe system. The circle indicates a region that the computer deemed suspicious, in this case for the presence of a lung cancer in a slice from a CT scan of the thoracic.

which the likelihood of malignancy is shown. Figure 5 shows a more advanced interface that can be used by CADx systems. Here the probability that a lesion is malignant is given numerically and in graphically form what this number means in terms of likelihood of being a malignant or benign lesion. In addition, this interface shows lesions similar to the unknown lesion that are selected from a library of lesions with known pathology. This provides a pictorial representation of the computer

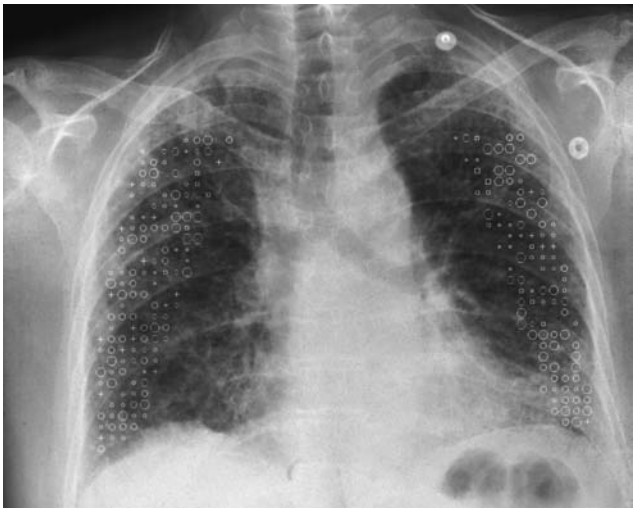


Figure 3. This chest radiograph is annotated with symbols indicating the presence of different appearances or patterns associated with interstitial chest disease. Crosses indicate a normal pattern; squares indicate a reticular pattern, circles represent a nodular pattern; and hexagons are honeycomb or reticulonodular patterns.

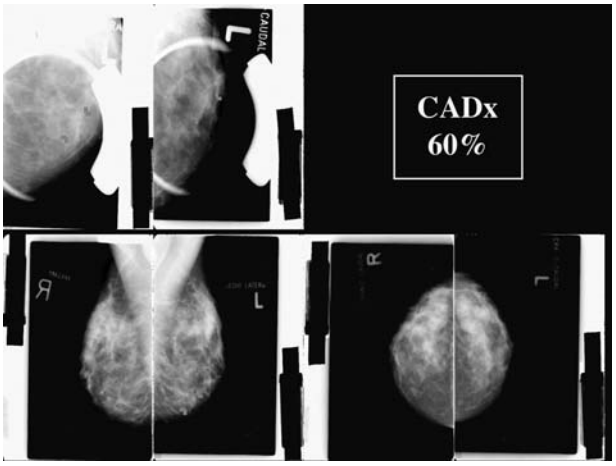


Figure 4. An example of a simple interface that could be used by a CADx scheme. Along with the images showing the lesion being analyzed, the display also shows the computer's estimated likelihood that the lesion is malignant. The images in the top left corner are spot magnification views and the images along the bottom are the standard screening views.

result that to the radiologist may be more intuitive than just a number.

In CADe, lesions or disease states are identified in images. The output of CADe schemes is typically locations of areas that the computer considers suspicious. These areas can then be annotated on the digital image. For CADe, the computer can detect actual lesions (true-positive detection), miss an actual lesion (false-negative detection), or detect something that is not an actual lesion (false-positive detection). The objective of the computer algorithm is to find all the actual lesions while minimizing the

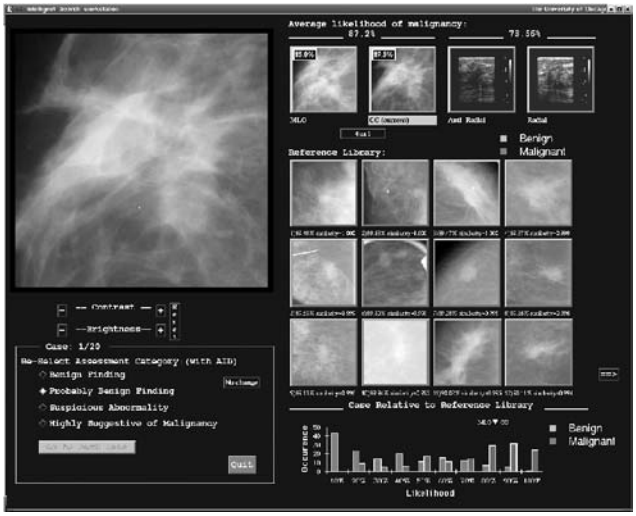


Figure 5. An example of an advanced CADx display that shows in addition to the likelihood of malignancy, a graphical representation and a selection of lesions similar to the lesion being examined recalled from a reference library containing lesions with known pathology. The red frame around a lesion indicates a malignant lesion and a green frame indicates a benign lesion.

number of false-positive detections. In practice, there exists a tradeoff between having high sensitivity (high true-positive detections) and the number of false detections per image. This tradeoff is important because the clinical utility of CADe depends on how well the algorithm works. In particular, if the false-positive rate is high (the exact definition of high is unknown) then the radiologist will spend unneeded time examining computer false-detections. Further, there is a possibility that a computer false detection could cause the radiologist to make an incorrect interpretation, if the radiologist should incorrectly agree with the computer. If the sensitivity of the CADe scheme is not high enough, then the radiologist could lose confidence in the CADe scheme's ability to detect actual lesions, reducing the potential effectiveness of the CADe scheme.

In CADx, the computer classifies different pathologies or different types of diseases. The most common application is to distinguish between benign and malignant lesions, to assist radiologists in deciding which patients need biopsies. The output of a CADx scheme is the probability that a lesion is malignant. If one sets a threshold probability value above which the lesion is considered malignant, the computer classification can be a true-positive (an actual malignant lesion is classified as malignant), a false-negative (an actual malignant lesion is classified as benign), a false positive (an actual benign lesion is classified as malignant), or a true negative (an actual benign lesion is classified as benign). The goal is to maximize the true-positive rate while minimizing the false-positive rate. Both false positives and false negatives could have serious consequences. A computer false positive could influence the radiologist to recommend a biopsy for a patient who does not need one. A computer false negative could influence the radiologist not to recommend a biopsy for a patient who has cancer.

There are important differences between CADx and CADe. In CADe, the results are presented as annotations on an image, so the radiologist examines image data, which they are trained and accustomed to do. In CADx, the output can be in numerical form, which radiologists are neither trained nor accustomed to using. Therefore, it may be more difficult for radiologists to use CADx effectively compared to using CADe. In addition, the consequences of inducing an error by the radiologist is much more severe in CADx than CADe. In most situations, the next step in a positive detection (CADe case) is more imaging, whereas in the CADx case, the next step is usually a biopsy or some other invasive procedure.

Why Is Computer-Aided Diagnosis Needed?

The goal of CAD is threefold: (*1*) to improve the accuracy of radiologists; (*2*) to make a radiologist more consistent (reduce intrareader variability) and to reduce discrepancies between radiologists (reduce interreader variability); and (*3*) to improve the efficiency of radiologists.

Since the interpretation of a radiograph is subjective, even highly trained radiologists make mistakes. The radiographic indication for the presence of a disease is often very subtle because variation in appearance of normal tissue often mimics subtle disease conditions. Another reason a radiologist may miss a lesion is that there is some other feature in the image that attracts their attention first. This is known as satisfaction of search and can occur when, for example, the radiologist's attention is focused on identifying pneumonia in a chest radiograph so that they do not notice a small lung cancer. Furthermore, in many instances the prevalence of disease in a population of patients is low. For example, in screening mammography only 0.4% of women who have a mammogram actually have cancer. It can be difficult to be ever vigilant to find the often-subtle indications of malignancy on the mammogram. Consequently, the missed cancer rate in screening mammography is between 5% and 35% (13–15) and, in chest radiography, the missed rate for lung cancer is ~30% (16).

Radiologists often have to decide whether an abnormal area is an indication of malignancy or of some benign process, or they may have to differentiate between different types of diseases that could give rise to the abnormality. This differentiation is often difficult because the radiographic indications between different disease types are not distinct. In diagnosing breast cancer, radiologists will recommend between 2 and 10 breast biopsies of benign lesions or normal tissue for every 1 cancer biopsied (17,18). The 50–90% of women who do not have breast cancer, but undergo a breast biopsy, suffer physical and mental trauma, and valuable medical resources were used unnecessarily. It is estimated that there is a 18.6% chance of receiving a biopsy for a benign condition after 10 years of screening for women in the United States (19). In general, the European false-positive rates are lower than in the United States (20). This is due in part to having a national programs and screening policies in Europe (21) and the higher likelihood of having malpractice lawsuit for a false-negative diagnosis in the United States.

Variability in performance of radiologists will reduce the effectiveness of the radiographic exam and, further, it will undermine the confidence that the public and medical community has in the technology. Again, because of differences in ability and differences in opinions, there exists variability between radiologists. In mammography, the variability between radiologists is well documented and can be very large (up to 40% variation in sensitivity in one national study) (22). Even a given radiologist is not always internally consistent. That is, the same radiologist may give a different interpretation when reading the same case a second time (23).

As the cost of the health care system increases, there is growing pressure to improve the efficiency of the system. In radiology, technology improvements have been introduced to improve workflow. Two examples are digital imaging systems, which can acquire images faster than conventional film systems and picture archive and retrieval systems (PACS), which can streamline the process of storing and recalling images. These and other technology improvements have increased the radiologists' workload. While not yet proven, it is hoped that CAD will allow the radiologist to read faster without reducing their accuracy. This will probably occur when the CAD schemes reach a certain level of accuracy (the exact level is unknown), so that the radiologist is not spending too much

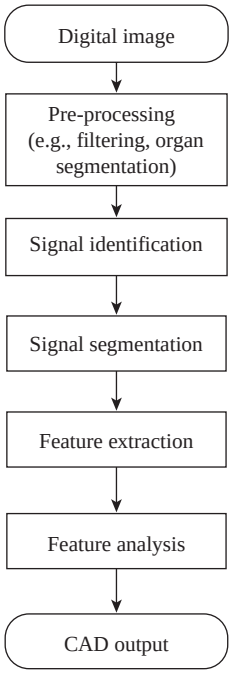


Figure 6. A flowchart of a generic CAD scheme. Most CADx and CADe schemes follow this template, although there are widely varying techniques for implementing each step.

time examining– considering computer false- positives. The CAD may also increase the radiologists' confidence in their interpretation, which should lead to faster reading times.

COMPUTER-AIDED DIAGNOSIS ALGORITHMS

There are probably thousands of publications on CAD, with hundreds of different techniques being developed. It is not practical to describe them all. There is, however, commonality between most approaches. Most techniques, whether they are for CADe or for CADx, can be described generically, as outlined in Fig. 6, as consisting of five steps: preprocessing, identification of candidate lesions (signals), segmentation of signals, feature extraction, and classification to distinguish actual lesions from false lesions or to differentiate between different types of pathologies or diseases (e.g., benign and malignant).

Radiographs can either be acquired using digital technology or screen-film systems. In digital systems, the image is acquired as a 2D array of numbers. Each element of the array is a pixel and corresponds to the amount of X-ray energy absorbed in the detector at that pixel location. This in turn is related to the number of X rays transmitted through the patient. Thus, a radiograph is a map of the X-ray attenuation properties of the patient. Different tissues in the body and different tissue pathologies have different attenuation properties, although the differences can be small. In a screen-film system, the image is recorded on X-ray film. The screen converts the X rays into visible light and the light is recorded by the film. For CAD purposes, this film must be digitized so that the image can be analyzed. A film digitizer basically shines light through the film and measures the amount of light transmitted. The resulting image is again a 2D array of numbers related to the X-ray attenuation properties of the patient.

One complicating property of screen-film systems is that they respond nonlinearly to X-ray exposure: digital systems respond linearly. Figure 7 shows the characteristic curve for screen-film system and a digital system. The contrast of objects in the image is proportional to the slope of the characteristic curve. For a screen-film system, at high and low exposures, slope approaches zero. Thus, for screen-film systems the contrast in the image is reduce in bright and dark areas of the image.

In this article, one approach to accomplishing these steps will be illustrated. In general, the techniques described are relatively simplistic, but are used to illustrate the concepts involved. More advanced and effective techniques are described in the literature; in particular,

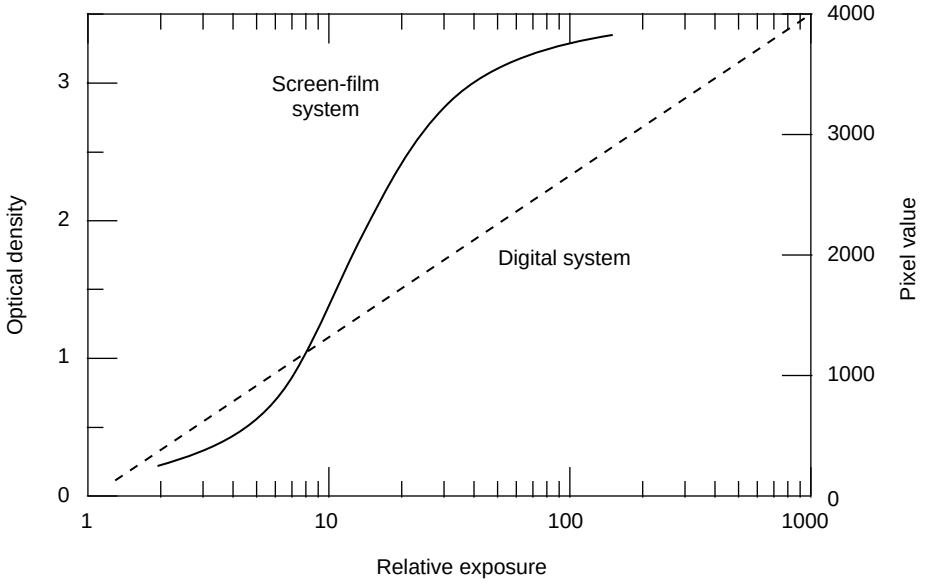


Figure 7. Comparison of a linear and a nonlinear detector. In a nonlinear detector (e.g., a screen-film system, the response of the system is X-ray exposure dependent. That is, the amount of darkening on the film (called film-optical density) depends on the X-ray exposure to the detector. At low exposures and high exposures, the image contrast (difference in film optical density) is lower than at optimal exposures. For a linear detector (e.g., most digital X-ray detectors), the response of the system is linearly dependent on the X-ray exposure to the detector. As a result, the contrast is independent of X-ray exposure.

review articles (24–29), and conference proceedings, such as the Proceedings of the SPIE Medical Imaging Conference, International Workshop on Digital Mammography (30–35), and the Proceedings of Computer Applications in Radiology and Surgery (CARS).

Using a mammogram, the steps outlined in Fig. 6 are illustrated in Figs. 8–14 using a somewhat simple procedure. Figure 8a shows the mammogram with an arrow indicating the location of a cluster of calcifications. An enlargement of the cluster is shown in Fig. 8b. In a radiograph, dark regions correspond to areas where there are more X-rays incident on the image and these are assigned high pixel values; and bright areas correspond to areas with fewer X-rays incident and they are assigned low pixel values. Within the breast area of the image, dark areas correspond to predominately fatty areas of the breast and bright areas correspond to fibroglandular tissues (those involved in milk production and the physical support of the breast to the chest wall). In the following sections, this image will be used to illustrate some of the basic concepts in developing CAD schemes.

Most of the initial CAD research was applied to 2D images, in which the 3D body part is projected into a 2D plane to produce an image. While being relatively simple, fast, and inexpensive, 2-D imaging methods are limited by the superposition of tissue. Lesions can be obscured by overlapping tissues or the appearance of a lesion can be created where no lesion actually exists. Three-dimensional imaging [ultrasound, CT, magnetic resonance imaging (MRI), positron emission tomography (PET), and single-photon-emission competed tomography (SPECT)] produces a 3D image of the body, which is often viewed as a series of thin slices through the body. These images are more costly to produce and take significantly longer to complete the exam, however, they can produce vastly superior images. These image sets also take longer to read. In some cases, the radiologist needs to review up to 400 image slices for each exam. With such large datasets, it is believed that CAD may be beneficial for radiologists.

In developing these techniques, researchers use information that a radiologist uses, information that cannot be visualized by radiologists, and information based on the radiographic properties of the tissue and of the imaging system.

Preprocessing

The first step in a CAD scheme is usually preprocessing. Preprocessing is employed mainly for three reasons: (*1*) To reduce the effects of the normal anatomy, which acts as a camouflaging background to lesions of interest: (*2*) To increase the visibility of lesions or a particular feature of a lesion: (*3*) To isolate a specific region within the whole image to analyze (e.g., the lung field from a chest radiograph). This reduces the size of the image that needs to be analyzed reducing computation time, reducing the chances

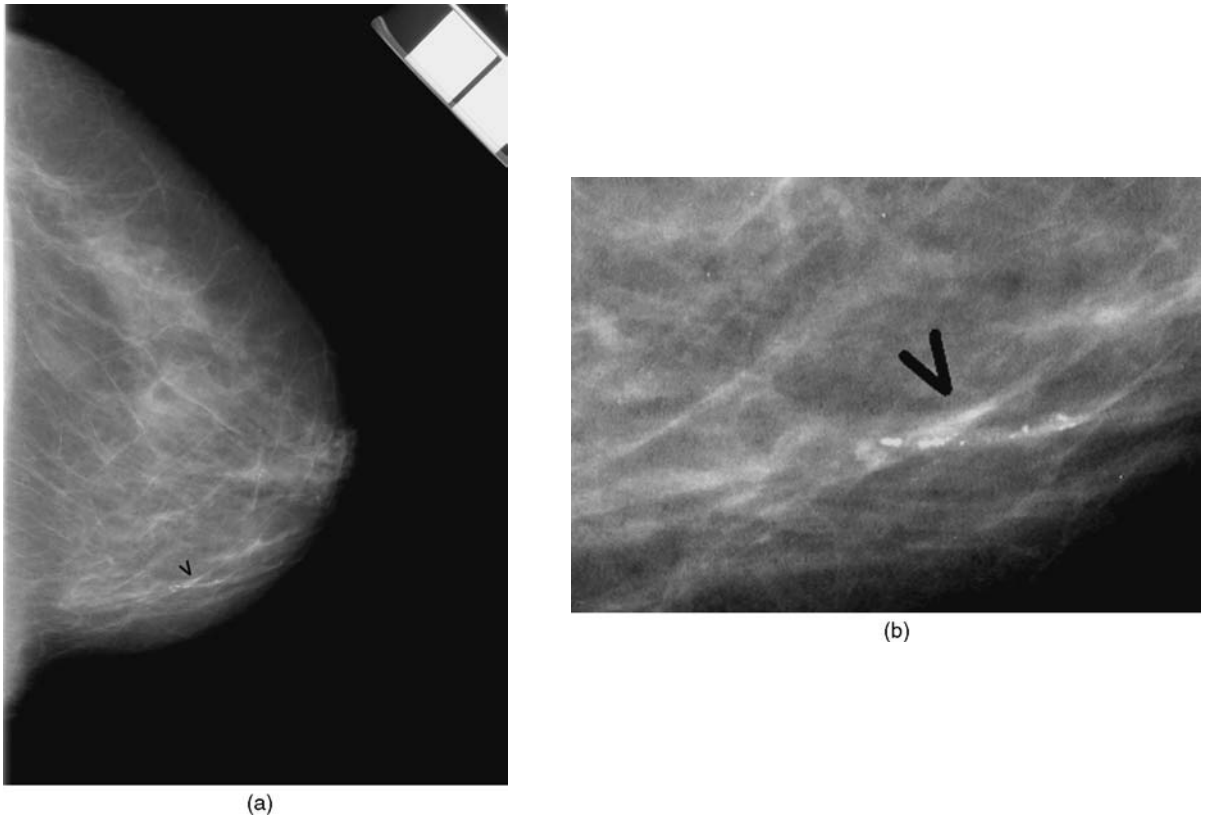


Figure 8. (a) A portion of a digitized mammogram. The arrow indicates the location of a cluster of calcifications. (b) An enlargement of the area containing calcifications.

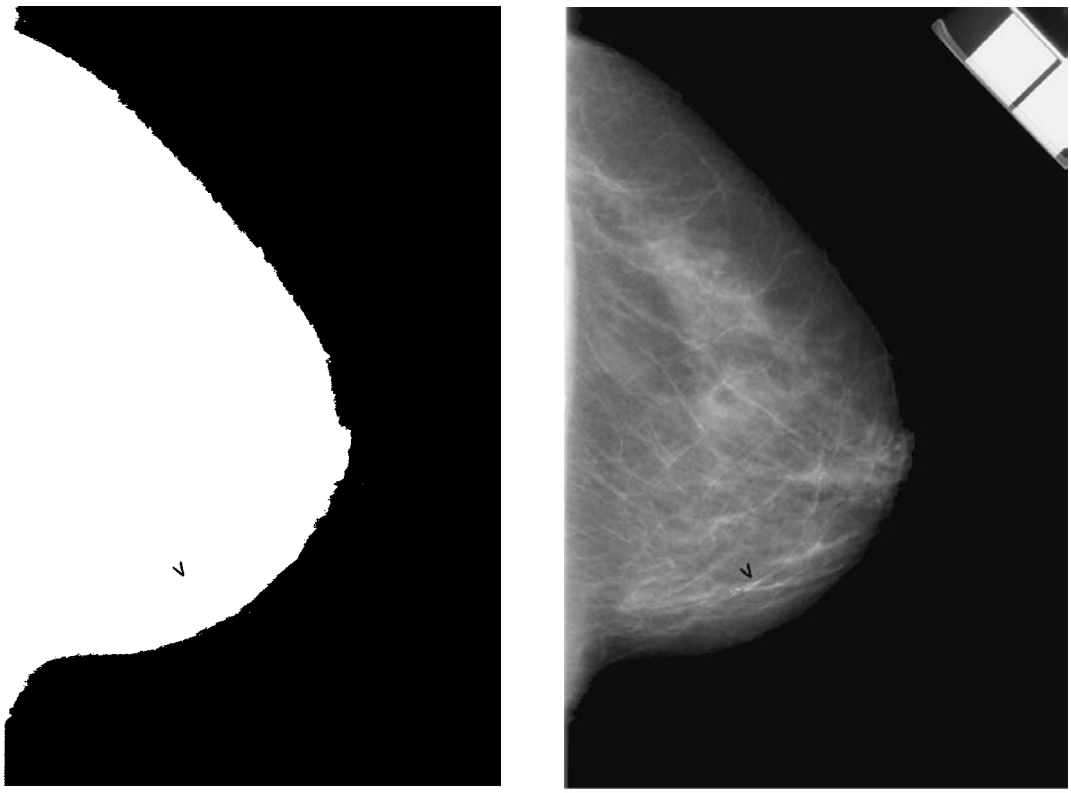


Figure 9. (a) A gray-level threshold was applied to the image shown in Fig. 8 in order to determine the breast boundary. Above the threshold value the image is made white and below the boundary the image is made black. The interface between the two areas is the skinline or breast boundary. (b) The determined breast boundary superimposed on the original image.

of a false detection, and reducing the complexity of the analysis, since only the area of interest is analyzed and extraneous image data are eliminated.

In our example CAD scheme, two procedures are performed. The first is to identify the border of the breast and the second is to process the image to reduce the background structure of the breast so as to highlight small bright areas that could be calcifications. The border of the breast was determined, so as to include only breast area in subsequent analysis. Outside of the breast, the image is black, corresponding to a large number of X rays incident on the film. Within the breast, X rays are absorbed or scattered so that the number of X rays incident on film is decreased and the image appears brighter. In this simple example, the breast outline is determined by thresholding the image so that below a threshold pixel value, all the pixel values are black and equal to above the threshold all pixels are white (Fig. 9a). The border between the black and white pixels is chosen as the border of the breast. The result is shown in Fig. 9b. This method produces a suboptimal result. At the top left corner of the image, there is an area included that does not contain breast tissue that was considered to be part of the breast. When the film was digitized, where there was a sharp transition from bright to dark, the response of the digitizer was slow, so that some of the bright area "bleeds" into the dark area.

In our example, the image is next processed by using a technique called the difference of Gaussians (DoG). In this technique, the image, which is typically 18×24 cm, is filtered twice by Gaussian filters, one with a small width (fwhm value, of 0.155 mm) and the other with a wider width (fwhm of 0.233 mm) (36). The first filter keeps small bright signals in the image (Fig. 10a) and the second eliminates the small bright signals (Fig. 10b). The two images are subtracted producing an image where large structures are eliminated and small signals are retained (Fig. 10c). Another useful outcome of this difference method is that the background of the subtracted image is more uniform compared to the original mammogram. This is important for the next step where potential calcifications are identified using a gray-level threshold method.

Note that for images that are recorded not on film but are acquired in digital format (i.e., a digital mammogram) now undergo preprocessing before the radiologist views the image (37). The goal is to improve the visibility of abnormalities for the human viewer. The techniques used in this type of preprocessing can be similar to those used in preprocessing for CAD purposes. The main difference is that a CAD preprocessed image would be considered overprocessed to the radiologist because they are designed to highlight only one type of lesion, whereas a radiologist needs to look for several to many different types of abnormalities in the image.

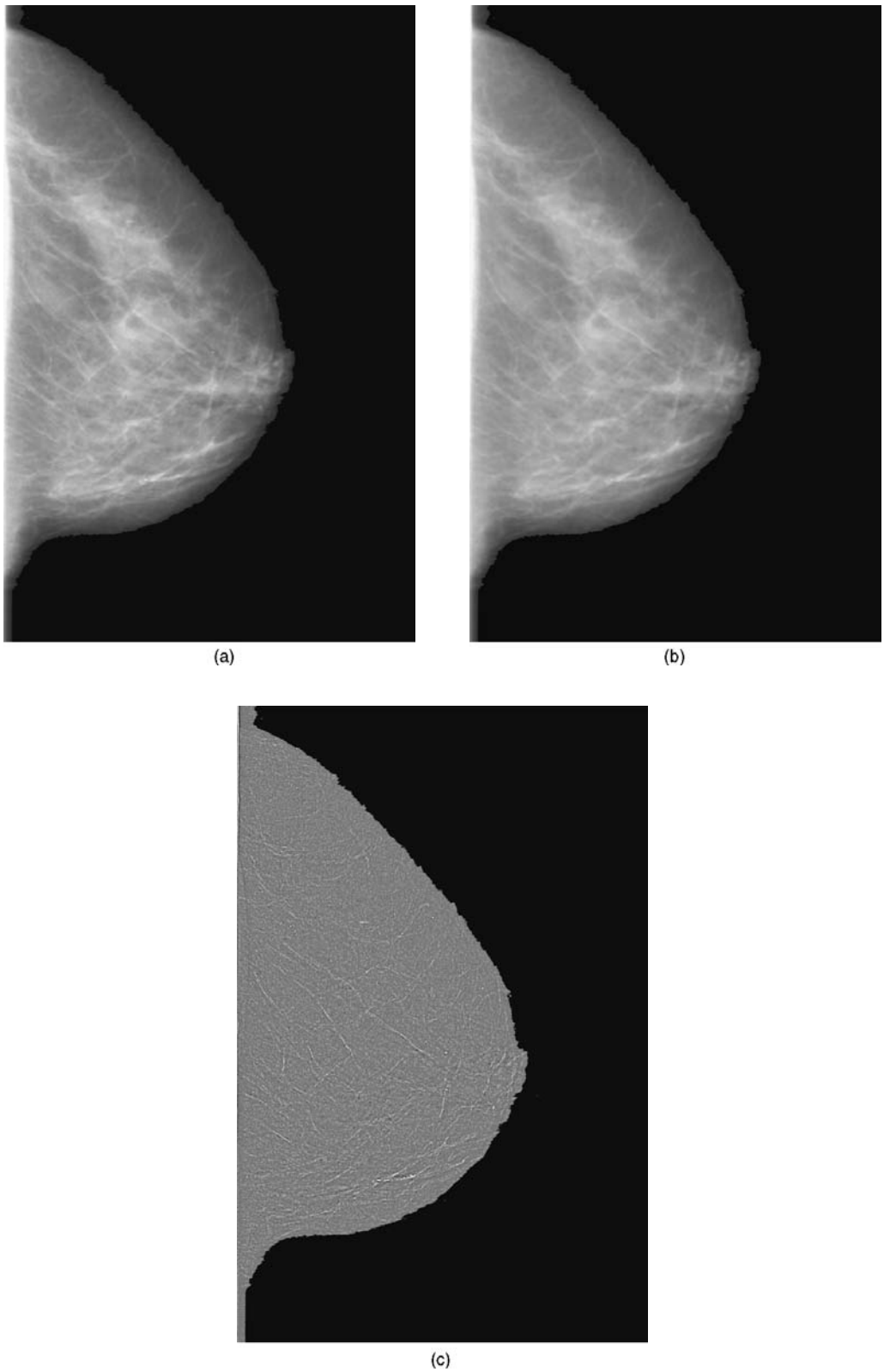


Figure 10. An example of a preprocessing method using the difference of Gaussian technique. (a) The image shown in Fig. 8 is filtered by a Gaussian filter that has a full width at half-maximum (fwhm) value of 0.155 mm. This filter enhances small objects on the order of 0.5 mm in diameter. (b) The image shown in Fig. 8 is filtered by a Gaussian filter that has a fwhm value of 0.233 mm. This filter degrades small objects on the order of 0.5 mm in diameter. (c) The image in Fig. 10b is subtracted from Fig. 10a producing an image with enhancement of small objects on the order of 0.5 mm in diameter and a reduction in the normal background structure of the breast.

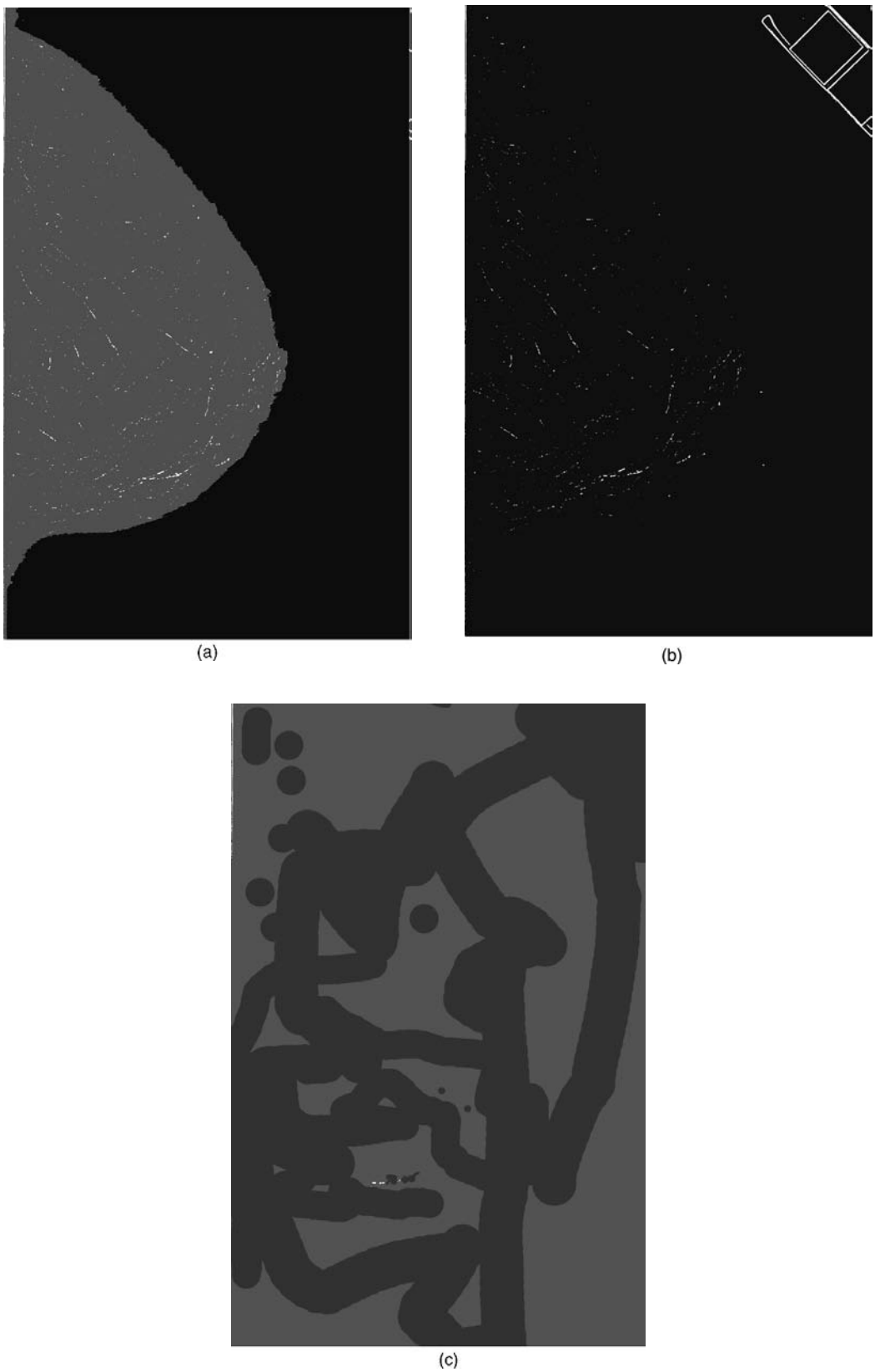


Figure 11. An illustration of signal identification using gray-level thresholding. Using the preprocessed image (Fig. 10c), a gray-level threshold is chosen so as to keep only a fraction of the brightest pixels. Three different threshold levels are shown in a–c. If the threshold value is too low as in part a, too many false signals (noncalcifications) will be kept. If the threshold is too high, as in part c, some of the actual calcifications are lost, even though most of the false signals are eliminated.

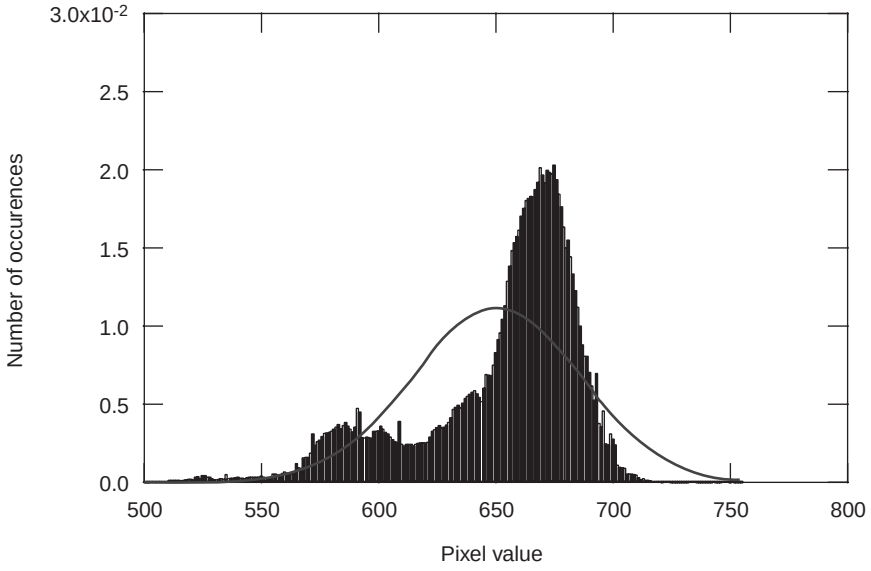


Figure 12. A histogram of pixel values from a hypothetical lesion. From a histogram like this one, features can be determined (e.g., mean pixel value and kurtosis). See text for details.

If the system does not respond linearly to X-ray exposure, then, the background level will affect the image contrast of lesions (see Fig. 7). In bright or dark areas of the image, the contrast is reduced compared to regions that are optimally exposed. This reduces the effectiveness of pixel-value-based and contrast-dependent techniques.

A limitation of these types of filtering is that lesions can often range in size. For example, lung nodules can as small as 0.5 cm or less and 5 cm or larger. A fixed-sized filter cannot be optimal for the full range of sizes possible. To accommodate a large size range, many researchers have developed multiscaled approaches. Wavelets are one class of multiscale filters. Multiple numbers of bandpass filters can be chosen based on which daughter wavelet decomposition are used in reconstructing the image. The principle of applying wavelets to medical images is discussed by Merkle et al. (38). In mammography, a weighted sum of the different levels in the wavelet domain is performed to enhance calcifications (39). Different approaches differ in their choice of wavelets and the selection of which levels to use in the reconstruction. A list of different wavelets used for processing microcalcifications on mammograms is given in Table 1.

Identification of Lesions

Once the image has been preprocessed, candidate lesions or signals need to be identified. The goal in CADe is to maximize the number of actual lesions identified even if a large number of false signals are detected. The false detections are reduced in the feature analysis step.

Signal identification is sometimes accomplished simultaneously with lesion segmentation (see next section). However, there are circumstances in which it is not. The

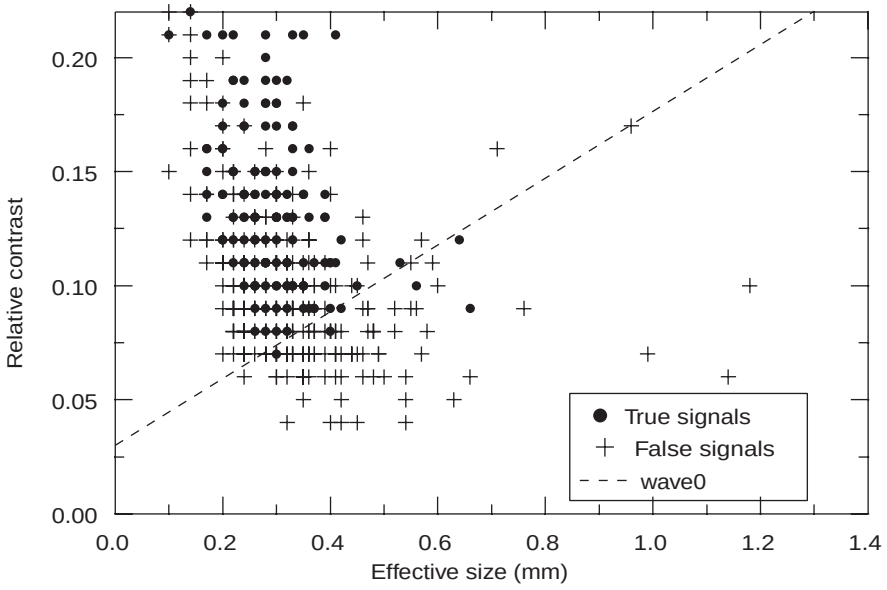


Figure 13. An example of feature analysis in which two features, contrast and size, are used to differentiate actual calcifications from computer-detected false detections. A threshold can applied to reduce the number of false detections, without eliminating many actual calcifications. In this example, the broken straight line shows the threshold values.

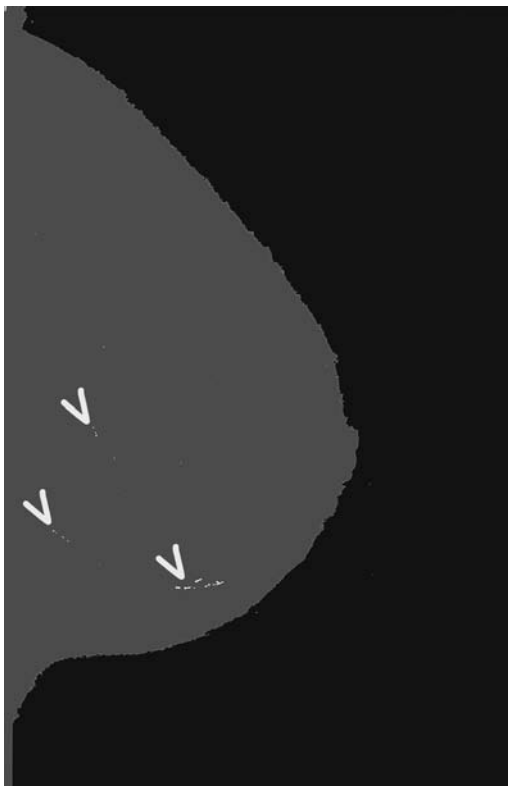


Figure 14. An example of the output of a simple CADe method illustrated in Figs. 8–13. This simple method detected the actual cluster of calcifications, but also two false clusters. More sophisticated methods can have much better performance over a wide variety of cases.

most obvious case is when a human indicates the location of signals. When functioning clinically a CADx scheme needs to know the location of the lesion that needs to be classified. This can be done using a CADe scheme. However, the CADe scheme may not detect a lesion that the radiologist is scrutinizing because either the CADe scheme may have failed to detect the lesion or the area that the radiologist is examining does not contain a real lesion. In either case, the radiologist would have to mark the location of the lesion in order for the CADx scheme to analyze it. When this occurs, the process is no longer automated.

Table 1. List of Different Mother Wavelets Used for Processing Mammograms Containing Microcalcifications

Lead Investigator	Reference	Wavelet
Brown	40	Undecimated spline
Chen	41	Morlet
Chitre	42	Daubechies' 6- and 20- coefficient
Kallergi	43	12-coefficient Symmlet
Laine	44	Dyadic
Lo	45	Daubechies 8-tap
Strickland	46	Biorthogonal B-spline
Wang	47	Daubechies' 4- and 20- coefficient
Yoshida	48	8-tap least asymmetric Daubechies

In X-ray imaging, the presence of a disease state is usually detected because the lesion is either more attenuating or less attenuating than the surrounding tissue. That is the lesion that will appear as a bright or a dark spot in the image. This makes gray-level thresholding a potentially useful method for segmentation. In Fig. 11, our example image has undergone gray-level thresholding using different thresholds. As the threshold is increased the total number of detections decrease. If the threshold is too high, some actual calcifications are lost. The signals that are detected that do not correspond to actual calcifications are caused by primarily image noise (due to the statistical fluctuations in the number of X rays incident on the patient per unit area) and normal breast tissue. Unfortunately, radiographs are a 2D projection of a 3D object, so that summation of overlapping tissues can produce areas in the image that mimic actual lesions. While gray-level thresholding can be effective on a given image, over a large number of images this simple thresholding method is not optimal. Calcifications can appear differently in different images, due in part to differences in the X-ray exposure used to make the image. Therefore, a single threshold applied to a cross-section of images will not be effective.

One method to improve the gray-level thresholding method is to make it adaptive. For example, one can use the statistics of a small region (e.g., 5 mm^2) centered on the calcification to select the appropriate threshold (2). A threshold can be set to the mean pixel value within the region plus a multiple (typically 3.0–4.0) of the standard deviation within the region. In this way, differences in X-ray exposure and differences in image noise can be accounted for.

Lesion Segmentation

Once the location of a lesion has been determined, the exact border of the lesion needs to be determined. This is necessary to extract features of the lesion. Again, gray-level threshold can be employed. One simple method is to find the highest pixel value near the identified pixel (e.g., in a 0.3×0.3 mm region) and then to find the mean pixel value in a local area surrounding the signal (e.g., in a 1×1 cm region). A threshold can be taken to be the mean pixel value, plus 50% of the difference between the maximum pixel value and the mean pixel value. Finally, all pixels above the threshold that are connected to the pixel with the maximum value are considered to be part of the calcification. Starting from a seed point, in this case the pixel with the highest value, a region is grown that contains all connected pixels that are above threshold. Connected pixels can be defined as the pixel above, below, to the right, and to the left of a given pixel, so-called four-point or eight-point connectivity, the same four pixels plus the four corner pixels.

This region growing method can be improved by first correcting the appearance of the calcification in the image for the degradation caused by the imaging system. The imaging will blur the image and further, the imaging system can response nonlinearly to X-ray exposure. Jiang et al. (49), and Veldkamp and Karssemeijer (50) indepen-

dently developed a segmentation technique based on background-trend correction and signal dependent thresholding. In these two approaches, corrections for the nonlinear response and the blurring of the calcification by the screen-film system and film digitizer are performed. At low and high X-ray exposures to the screen, the contrast, which is proportional to the slope of the characteristic curve, is reduced (see Fig. 7). That is, the inherent contrast of the calcifications is reduced when recorded by the screen-film system. Therefore, the image or radiographic contrast will depend on the background intensity. If a correction is not made for this nonlinearity, then it becomes extremely difficult to segment accurately calcifications in dense and fatty regions of the image simultaneously with calcifications in other regions of the breast. Similarly, the smaller the calcification, the more that its contrast is reduced due to blurring. This can be corrected based on the modulation transfer function of the screen-film system (51).

The above segmentation method will work well as long as there is sufficient separation between the calcifications. If two or more calcifications are too close, then they may be segmented as one large calcification. In such situations, more sophisticated methods are more effective. Besides the pixel value, thresholding can be applied based on other features of the image, such as the texture and gradients. These more advanced techniques are also important in applications where the border of the lesions is not well-defined visually in the image.

Feature Extraction

To reduce the number of false detections (i.e., to differentiate true lesions from false detections) or to classify a lesion (e.g., benign versus malignant) features are extracted and subsequently used by a classifier. The strategy in CADe is to segment as many actual lesions as possible. This will include a large number of false detections.

There are probably thousands of features that can be used and these are dependent on the imaging task. Further, the optimum set of features is not known for any imaging task. Therefore, a large number of different features are being used by different investigators. The features are based on those that a radiologist would use and those that a radiologist would not use. As an example, a radiologist uses the brightness of the lesion to determine whether a lesion is present in the image. This can be determined conceptually by plotting a histogram of pixel values (see Fig. 12). The brightness is related to the mean pixel value, M, and is given by

$$M = \frac{1}{N}\sum_{i=1}^{N} f(i)\,p(i) \tag{1}$$

where f is the frequency of occurrence of pixel value $p(i)$, N is the total number of pixels in the region being analyzed, and i is the index in the histogram, $f(i)$. An example of a feature not used by a radiologist is kurtosis, K, which is defined as

$$K = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(f(i)-M\right)^4 p(i)}{\left[\frac{1}{N}\sum_{i=1}^{N}\left(f(i)-M\right)^2 p(i)\right]^2} \tag{2}$$

The kurtosis describes the flatness of the histogram. Histograms that are very peaked have high kurtosis. Kurtosis can also be thought of as comparing the histogram to a Gaussian distribution. A value of 3.0 indicates that the histogram has a Gaussian distribution.

A large, but incomplete, list of features used by different investigators for the detection of calcifications in mammograms is given in Table 2. The features used for other applications of CADe and CADx will differ than the ones in Table 2, but the categories will be the same: pixel intensity-based, morphology-based, texture-based, and others. For example, for 3D images, such a CT scan of the thorax, instead of a circularity feature, the sphericity of a lesion would be calculated. The drawback of having a large number of features to choose from is that the selection of the optimum set of features is difficult to do, unless a very large number of images are available for feature selection (69). These images are in addition to the images needed for training and the images needed for testing the technique.

Table 2. List of Different Features Used for Distinguishing Actual Calcifications from False Detections

Pixel-Value Based	Morphology Based	Derivative Based	Other
Contrast (52–57)	Area (52,54,56,58,59)	Mean edge gradient (45,54,55,57,60–64)	Number of signals per cluster (52,59,65)
Average pixel value (53,66)	Area/maximum linear dimension (61)	Standard deviation of gradient (62)	Density of signals in cluster (52,58,59)
Maximum value (45,67)	Average radius (67)	Gradient direction (64)	Distance to nearest neighbor (65)
Moments of gray-level histogram (68)	Maximum dimension (62)	Second derivative (55)	Distance to skin line (65)
Mean background value (45)	Aspect ratio (65)		Mean distance between signals (59)
Standard deviation in background (45,57,58)	Linearity (63)		First moment of power spectrum (6)
	Circularity (54,59,67)		Effective thickness (49)
	Compactness (61,53,55)		Peak contrast/area (67)
	Sphericity (contrast is the third dimension) (67)		Mean distance from center of mass (42)
	Convexity (68)		

Most of the features listed in Table 2 use standard techniques for determining their value. One feature, effective thickness of the calcification developed by Jiang et al., is calculated using a model of image formation (49). That is, what thickness of calcification will give rise to a given measured contrast in the digital image? To do this, corrections for the blurring of the digitizer and the screen-film system are performed, along with corrections for the characteristic curves of the digitizer and the screen-film system. The assumption is that, in general, calcifications are compact, so their diameter and thickness should be comparable. Film artifact (e.g., dust on the screen), will have a very high thickness value compared to its size, and therefore can be eliminated. Similarly, detections that are thin compared to their area are likely to be false positives due to image noise.

Highnam and Brady take this one step further. For every pixel in the image they estimate the corresponding thickness of nonfatty tissue (essentially fibroglandular tissue) by making the corrections described in the preceding paragraph and in addition corrections for X rays that are scattered within the breast, for the energy and intensity distribution of the X-ray beam, and for other sources (70). This in principal produces an effective image that is independent of how the image was acquired or digitized.

Most features are extracted from either the original image or a processed image that has sought to preserve the shape and contrast of the calcifications in the original image. Zheng et al. used a series of topographical layers ($n = 3$) as a basis for their feature extraction. The layers generated by applying a 1, 1.5, and 2% threshold using equation 2. This allows for features related to differences between layers (e.g., shape factor in layer 2 and shape factor in layer 3) and changes between layers (e.g., growth factor between layers 1 and 2) to be used.

Classification

Once a set of features has been identified, a classifier is used to reduce the number of false detections, while retaining the majority of actual calcifications that were detected. Several different classifiers are being used: simple thresholds (2,52–54,60–62,71), artificial neural networks (40,72–74), nearest-neighbor methods (54,75,76), fuzzy logic (45,68), linear discriminant analysis (77), quadratic classifier (40), Bayesian classifier (55), genetic algorithms (78), and multiobjective genetic algorithms (79).

The objective of the classifier is to find the optimal threshold that separates the two classes. Shown in Fig. 13 is an example problem where two features of candidate lesions have been extracted and plotted (2D problem). The classifier determines the boundary between the two classes (normal and abnormal). The optimal boundary is one that maximizes the area under the receiver operating characteristic (ROC) curve (see the section CADx Schemes). There are different types of classifiers: linear, quadratic, parametric, and nonparametric. A linear classifier will produce a boundary that is a straight line in a 2D problem, as shown in Fig. 13. Quadratic classifiers can produce a quadratic line. More complex classifiers (e.g., k nearest neighbor, artificial neural networks, and support vector machines) can produce very complex boundaries.

Alternatives to Feature Extraction

In lieu of feature extraction, or in addition to feature extraction, several investigators have used the image data as input to a neural network (80–83). The difficulty with this approach is that the networks are usually, quite complex (several thousand connections). Therefore, to properly train the network and to determine the optimum architecture of the network requires a very large database of images.

Another limitation of these and similar approaches is that extremely high performance is needed to avoid having a high false-positive rate. The vast majority of mammograms are normal and the vast majority of the areas of mammograms that are abnormal do not contain calcifications. Mammographically, the average breast is $\sim 100\,\text{cm}^2$ in area. For a 50 μm pixel, there will be 4 million pixels to be analyzed. If there are 13 calcifications of 500 μm in diameter, then only 0.01% of pixels will belong to a calcification. Therefore, a specificity of 99.9% will give rise to 10 false ROIs per image, which is more than 100 times higher than a radiologist.

Computation Times

The computation times are not often stated by most investigators, perhaps under the belief that this is not an important factor since computers will always get faster. In general, times range from 20 s (84) up to several tens of minutes (inferred from description of other published techniques), depending on the platform and pixel size of the image. For any of the technique to be used clinically, they must be able to process images at a rate that is useful clinically. For real-time analysis (e.g., for diagnostic mammography) there is approximately one patient approximately every 20 min per X-ray machine. This means there is 5 min available per film, including the time to digitize the film. However, many clinics have several mammography units, and some clinics have mobile vans that image up to 200 women off-site. In these situations, computation times of < 1 min per film may be necessary or multiple CADe systems would need to be employed. This also assumes that only one detection scheme is run. In mammography, at least one other algorithm, for the detection of masses, will be implemented, cutting the available time for computation in half.

In most centers, screening mammograms are not read to the next day. This allows for processing overnight and computation time becomes less critical, but still important. To analyze 20 cases (80 films) in 15 h (overnight) is 11 min per film for at least two different algorithms. For a higher volume center (40 cases), this gives < 3 min for each algorithm.

In other applications, the results of the computer analysis are needed immediately. For example, if a radiologist marks in the image an area that they would like analyzed by the computer (e.g., to determine the likelihood of malignancy of a lesion) then the radiologist, who is busy and under pressure to read the images quickly and accurately, does not want to wait to see the computer results. If the

computer takes more than a few seconds to return a result to the radiologist, then the radiologist may choose not to use the computer because it is reducing down their productivity. Computation time is not a trivial matter.

EVALUATION OF CAD ALGORITHMS

CADx Schemes

The performance of a CADx scheme can be measured in terms of sensitivity (the fraction of actual abnormal cases that are called true) and specificity (the fraction of actual normal cases called normal). This pair of values is easy to compute. However, in general, as the sensitivity increases, the specificity decreases. This can give rise to the problem of comparing two CADx schemes, one with high sensitivity, but low specificity, and the other with lower sensitivity, but higher specificity. This problem is solved by using ROC analysis (85,86).

The CADx schemes that involve a differentiation between two categories or classes (e.g., benign and malignant) can be evaluated using ROC analysis. In a two-class classification, objects are separated into one of two classes. This is usually done based on feature vector that characterized the object. Conceptually, the problem reduces to sorting lesions into one of two classes based on the feature vector as illustrated in Fig. 15. In any problem of interest, the population of feature vectors for actual negatives (e.g., benign or computer-detected false-positives) overlaps with the population of feature vectors for actual positives (e.g., malignant or actual lesions). Depending on the selection of a threshold based on a given feature vector value, some actual negatives will be classified as positive and vice versa. Depending on the selected threshold, different pairs of true-positive fraction (the fraction of actual positive cases that are classified as positive) and false-positive fraction (the fraction of actual negative cases that are classified as positive) will be obtained. By selecting all possible thresholds, a set of TPF and FPF pairs will be obtained and they will form an ROC curve, as shown in Fig. 16. In practice, the populations shown in Fig. 15 are not smooth, because only a finite number of cases are used in evaluating a CAD scheme. Therefore, there is some uncertainty as to the true shape of the ROC curve. Fortunately, statistical methods exist to fit a curve to the TPF-FPF pairs (86). The most common summary metric to compare different ROC curves is the area under the curve (AUC). The AUC is the average TPF for all possible FPFs. Since there is uncertainty to the exact shape of the ROC curve, there is uncertainty in the true AUC value. Again, statistical methods have been developed to allow hypothesis testing between curves (86).

The most advanced of the ROC analysis methods can allow generalization of the results to any population of cases that are represented by the testing cases and to a pupulation of readers (87). This is, so-called multiple-reader-multiple-case (MRMC) model is important when analyzing observer studies (see the section Observer Studies).

If the CADx scheme is not a two-class problem, then currently there is no validated method for evaluating the CADx scheme in general (88). One can, however, reduce the problem to a series of two-class problems or make some assumptions to simplify the problem to allow two-class ROC analysis to be used (89–91).

While the AUC is a useful metric, it assumes that all TPF are equally desirable. Often, this is not the case. In diagnostic mammography, where the radiologist must

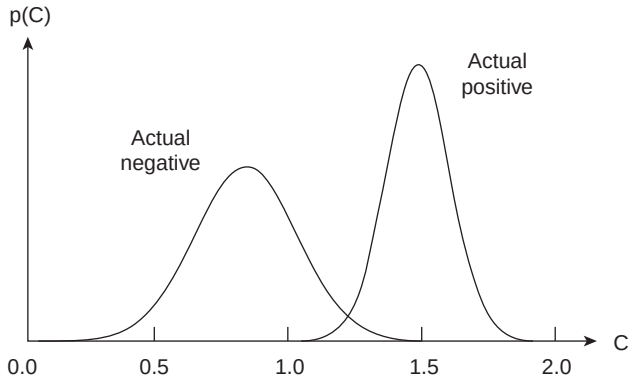


Figure 15. Conceptual representation of a two-class classification problem. The task is to classify lesions correctly into either actually negative lesions (e.g., benign or computer-detected false positive) or actually positive lesions (e.g., malignant or an actual lesion). The two probability distributions represents the probability of an positive or a negative lesion having a given value of a feature C (contrast). For a small range of contrasts (1.1–1.4), the two distributions overlap and misclassification can occur. To reduce the overlap regions, multiple features, instead of a single feature, are used.

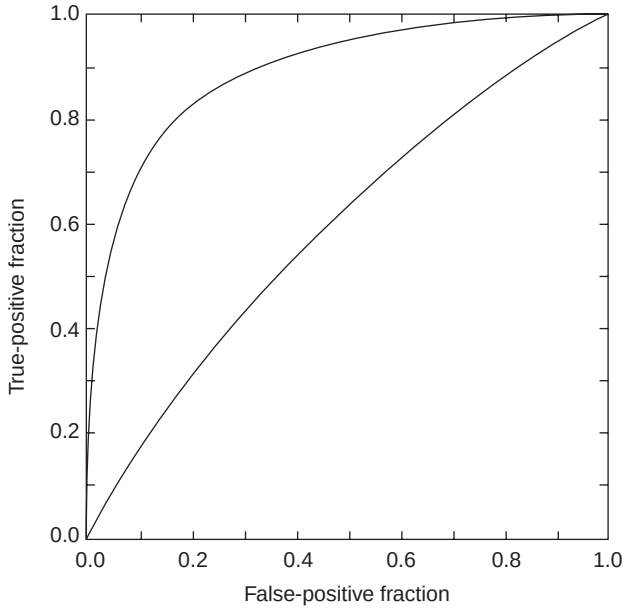


Figure 16. A typical ROC curve. The higher curve indicates higher performance than the lower curve. The upper curve represents the performance of radiologists in classifying masses as benign or malignant and it has an area under the curve of 0.91. The lower curve represents the performance of radiologists in classifying calcifications and it has an area under the curve of 0.62.

decide whether a known breast lesion is benign or malignant, the "penalty" for missing a cancer (classifying a malignant lesion as benign) is considered greater than classifying a benign lesion as malignant. In such circumstances, it may be more appropriate to consider only the region of the ROC curve in the high sensitivity (TPF) region (92,93). Partial AUC can be estimated, for example, considering the area under the curve for TPF > 0.80. This would imply that operating at a TPF < 0.80 is clinically unacceptable.

Finally, radiologists, when reading clinically, operate at one point on an ROC curve. In fact, two radiologists could have a difference of opinion on a number of cases, but operate on the same ROC curve. If one radiologist is more aggressive than the other, than the aggressive radiologist will have a higher TPF and a higher FPF than a more conservative radiologist and thus would be operating at a point higher up on the curve. Given that AUC is used as a figure of merit, then two radiologists having the same ROC curve have the same performance, although one may detect more disease than the other, because the other radiologist will have a lower false-positive rate (i.e., higher specificity).

CADe Schemes

The CADe schemes are not amenable to ROC analysis because CADe schemes address detection of lesions not the classification of a known lesion as with CADx schemes. That is, lesion classification is often a binary problem (e.g., benign or malignant), and therefore ROC analysis can be used. The CADe schemes involve the identification of the location of a lesion in the image and thus it is not a binary problem. The CADe schemes in general involve a tradeoff between TPF and false-positive rate (the number of false detections per image). Under such situations, free-response operating characteristic (FROC) curves are plotted (94). An example curve is shown in Fig. 17. For the task of detecting a lesion, it is important to identify the location of the lesion correctly. This raises the issue of how to score a CADe detection as correct or a miss (see the section Scoring Criteria).

In an ROC plot, the two axes range between 0 and 1.0, since they represent the true and false positive fractions (i.e., they are the ratio of two numbers). In an FROC plot, the y axis ranges from 0 and 1.0, but the x axis can range between 0 and 10 or higher, since there can be multiple detections on one image. That is, the upper limit for the average number of false positives per image is not fixed. Because of this difference between FROC and ROC curves, the statistical methods to analysis ROC data cannot be applied to FROC data. Recently, Chakraborty developed a statistical method for comparing two FROC curves (95), although the method still has some limitations.

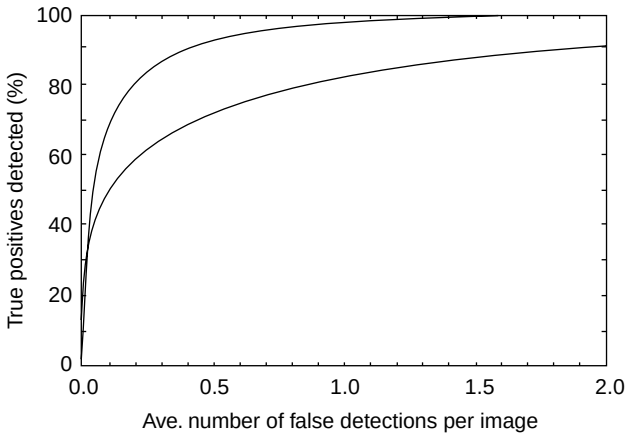


Figure 17. Typical FROC curves. The upper curve represents better performance.

In general, CADe schemes are not as accurate as radiologists, because the computer's false-positive rate is much higher than that of radiologists. Radiologists recall $\sim 10\%$ patients who receive a screening mammogram. Since only $\sim 0.5\%$ of the screened population will have breast cancer, the radiologist has a false-positive rate of ~ 0.0125–0.025 per image (the higher number is assuming that the lesion was seen in both views). The CADe schemes have ~ 10 times higher false-positive rates.

Technical Issues in Evaluation

The measured performance of a CAD scheme is important to compare different CAD schemes. The measured performance of a CAD scheme depends on several factors: (*1*) the inherent performance of the scheme; (*2*) the cases used to evaluate the scheme; (*3*) the scoring criteria used; (*4*) the quality of the truth data; (*5*) the method used to develop, train, and evaluate the scheme.

Ideally, one would like factors 2–5 above to be identical in all measurements of performance, but this does not exist at present, since there are no universal standards for these factors. Factors 2–5 are described below.

Database. The measured performance of a CAD scheme depends of the difficulty of cases used to evaluate performance. For a detection scheme, if the cases contain a high fraction of obvious abnormalities, then the measured performance will be high (96). In fact, differences in the cases used for evaluation can easily mask difference in true performance between two different CADe schemes.

The same problem exists for CADx schemes. Here an obvious case would be one that belongs clearly to one of the two classes. Often researchers will use a consecutive series of cases that went to biopsy as an evaluation dataset. These in general, represent difficult cases, especially for the actually benign cases, since at least one radiologist thought the cases was suspicious. However, since different radiologists have different skill levels, a set of consecutive biopsied cases from one institution may be different in difficulty from a set of consecutive biopsied cases from another institution. Currently, there is no method available to compare the difficulty of cases.

Public databases are becoming available (97). There are two websites where images can be downloaded: one for mammograms (http://marathon.csee.usf.edu/Mammography/Database.html) and the other for CT scans of the lungs (http://imaging.cancer.gov/reportsandpublications/reportsandpresentations/firstdataset). These in principal allow a common set of cases to be used in evaluating CAD schemes. This development is important in allowing

comparison of different CAD schemes, but is insufficient in of itself.

Truth. Truth is elusive in CAD research, but vitally important. To optimize and to train a CAD scheme properly, accurate truth information is needed. Truth is needed at several levels. First, the pathology of actual lesions needs to be known (i.e., benign, malignant, or normal tissue). If a biopsy is performed, then it is possible to determine whether a lesion is malignant. However, there is some variability between pathologists, particularly on "borderline" lesions (those on the line between being malignant and premalignant). Further, studies have shown that pathologists, while being highly accurate, make mistakes $\sim 5\%$ of the times. Researchers usually accept this level of accuracy since it is too time consuming and costly to have the pathology of all biopsies redone.

There are lesions that are potentially malignant, but do not go to biopsy. In these patients, the lesion is followed and either a biopsy is performed in the future when the lesion becomes more suspicious, or the lesion is judged to stable after following the patient with more imaging in subsequent years. However, in a number of instances, the patient moves or changes hospitals and their follow-up is then incomplete. These patients are usually eliminated from CAD research use.

The next level of truth is where is the exact boundary of a lesion. It is not possible to know the answer to this question. Even though the boundary can be well defined by a pathologist, it is not possible to correlate the pathology boundary from the excised tissue to the boundary of the lesion in the image. To optimize (e.g., the segmentation of a lesion), the boundary of the lesion needs to be known exactly, that is, whether a given pixel in the image is part of the lesion or not. The current method to determine "truth" for the boundary of a lesion is to have a radiologist draw the outline of the lesion while viewing the image on a computer screen. Since this is subjective, the variability can be reduced by getting a number of radiologists to outline the lesion.

The last level of truth is whether a lesion is present or not. There are two notable examples of this problem: microcalcifications in a mammogram and lung nodules in a CT scan. It is possible to know with certainty that a cluster of microcalcifications is present in a mammogram: The tissue containing the cluster that can be seen mammographically is excised and a pathologist can verify that calcifications are present in the tissue. It is not possible, however, to determine the location of every individual calcification in the cluster, even if the pathologist were able to count the number of calcifications in the cluster. When cancer metastasizes to the lungs, it is not possible to determine the exact number of nodules present, because some of the nodules are very subtle and difficult to detect in the images. With metastatic lung cancer, it is not necessary clinically to biopsy every nodule present.

To deal with this uncertainty, a number of radiologists, often called a panel of experts, can mark the location of individual microcalcifications or lung nodules. However, there are a number of problems with using a panel of experts. First, there is wide variability between radiologists, so that a large number of radiologists is needed to form the panel. The minimum number, if not known, is likely > 3. Panels can also be biased. Members of the panel can discuss their markings to reduce variability. However, one member of the panel could unduly influence other members of the panel, for example, if one member is a world-renowned expert (98).

Scoring Criteria. For CADx schemes, scoring the computer output is straightforward when ROC analysis is used. The truth for each case is known (either benign or malignant). Then it is a matter of selecting a threshold (applied to the CADx output) for which a lesion is considered malignant. For a given threshold, if the output of the CADx scheme is greater than the threshold value, then the lesion was classified as malignant.

For CADe output, scoring is much more problematic. There are many different criteria used for judging whether the computer found the lesion or whether the detection is a false positive (a computer detection not corresponding to an actual lesion). Most scoring criteria rely on some combination of the location of the center of the computer-detected signal and the actual lesion and the border of the computer-detected signal and the actual lesion. Figure 18 illustrates

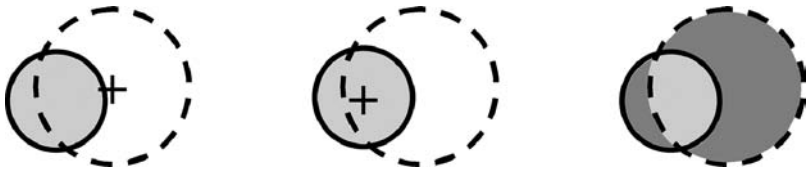


Figure 18. An illustration of some different criteria used to score computer detections of actual lesions. The actual lesion is shaded light blue and the computer detection is given by the circle with the dashed line. In the right most figure, the computer is scored as a false detection, since the center of the computer-detected lesion is not within the boundary of the actual lesion. In the center part of the figure, the computer detection is scored as a true positive, since the center of the actual lesion is within the boundary of the computer detection. In this method, there is no penalty when the computer detects a very large lesion, even though the actual lesion may be small. Both of these methods are sensitive to the size of the actual lesion and the computer detected lesion. The method on the right corrects for these effects. In this method the area of the overlap of the computer-detected lesion and the actual lesion (light blue area) is divided by the area of the union of the two lesions (the gray and light blue areas). If this value is greater than a threshold (e.g., 40%), then the computer detection is scored as a true positive.

several different scoring methods. There is no consensus on which is the best method. In addition to the objective methods illustrate in the figure, many clinical studies rely on the subjective opinion of a radiologist to determine whether the computer correctly detected the lesion. Given the variability among radiologists, this method has severe limitations.

Methodology. As the field of CAD matures, evaluation methodology is becoming more refined, reducing possible biases that were common in earlier work. It is now known that different sets of images are needed to develop a technique, to train the technique, and to test the technique (69,99,100). However, unlike some other type of image analysis problems, medical images that have verified truth are not easily obtainable. Therefore, while it is desirable to have three independent datasets to develop, train and test, it is usually not possible.

When images are scarce, training and testing can be accomplished through cross-validation or bootstrapping (101–105). These methods divide a single dataset into two subsets: one for training and the other for testing. If the cases are divided such that all images from a given patient are all in one or the other subset, then bias due to commonality of images between testing and training can be eliminated. When dividing the dataset into subsets there are two contrasting goals: maximize the number of training images to properly train the algorithm versus maximizing the number of test images to minimize the variance in the measured performance.

There are three common cross-validation methods used in CAD research:

1. K-fold. The dataset is divided into K subsets of equal size. One of the K subsets is reserved for testing while the other $K - 1$ subsets are used for training. The process is repeated until each of the K subsets is used for testing. The mean performance and the variance can then be computed.
2. Leave-one-out. This is the extreme of the K-fold, where K equals the number of cases.
3. Jackknife. This is a generalization of the leave-one-out, where the number left out can be one to half the number of cases.

Many researchers believe that the 0.632+ bootstrap is the optimum method for evaluating CAD schemes (105). In this method, training samples are selected at random from the N cases in the dataset. After each sample is selected, a new sample is selected again from the full dataset; this is called random selection with replacement. After a total of N samples are selected, a fraction of the dataset will not have been selected. These are used to testing. Statistically speaking, 63.2% of the cases will be selected for training while 36.8% will have not been selected for training and form the testing set. This is repeated many times and the mean performance and variation can be computed. This technique is believed to minimize both the selection bias and the variance in the measured performance compared to other techniques.

While all these methods are effective in reducing some of the biases associated with measuring the performance of a CAD scheme, they have one limitation. In all these methods, the CAD scheme is trained and tested multiple times using a different mixture of cases. This permits the mean performance (and the standard deviation in performance) to be calculated. However, this mean performance does not correspond to any specific CAD scheme, since it is an average of many trained schemes. Researchers must arbitrarily choose one trained scheme, often taking the best performing scheme, but that scheme may not have the best performance on different set on cases.

Evaluation of CAD Schemes as an Aid

The ultimate test of CAD efficacy will be whether patient outcomes are improved when CAD is used and whether it is cost-effective in doing so. Since CAD is still a new technology, it has not been widely use for a sufficient amount of time to address this question properly. Clinical studies are a prerequisite to patient outcome studies. Recently, a few clinical studies have been reported examining the effect of CADe on screening mammography with mixed results. However, intermediate measures of diagnostic efficacy can be measured. Studies have been conducted to show that CADe can find cancers that are missed clinically and observer studies, which simulate clinical reading conditions, have been performed. These are described below.

Missed Lesions. In theory, the actual performance of CAD on a set of cases is not important. Since CAD is to act as an aid to the radiologist, in principle, CAD only needs to be correct on cases the radiologist makes an error. For example, in CADe for mammography, the CADe scheme need only detect those cancers that the radiologist misses. Since radiologists miss between 5–35% of cancers, the sensitivity of the CADe scheme could be as low as 35% and still be an effective aid. Similarly, in theory, the CADe scheme could have 95% sensitivity and be of no use to the radiologist in detecting missed cancers. Therefore, it is necessary to measure the ability of a CADe scheme to detected clinically missed disease. In reality, if the CADe scheme does not have high sensitivity, the radiologist will lose confidence in the ability of the CADe to find cancer, thus reducing the CADe scheme's effectiveness as an aid.

There have been several studies in mammography and one in thoracic CT of the effectiveness of CADe to detected missed cancers. These studies show that between 50 and 80% of mammographically missed cancers can be detected by a CADe scheme (106–109) and 84% of missed lung nodules can be detected by a CADe scheme on lung CT scans (110).

These studies by themselves do not prove that CADe will be an effective clinical tool. It is still necessary to show that a radiologist will recognize an area detected by CADe as a diseased area. It is possible that the cancers detected by CADe will not be above the radiologist's threshold for what is a cancer. To test this, observer studies are conducted.

Observer Studies. Observer studies are conducted to measure the ability of a CAD scheme to improve radiologists' performance. They are designed to simulate clinical

reading conditions and thus are considered an indication of how CAD may affect radiologists in their clinical work. In a typical CAD observer study, between 6 and 15 radiologists read a set of cases under two different reading conditions: without the computer aid and with the computer aid. The number of cases ranges from 60 to up to 1000. There are two different types of CAD observer studies: independent reading and sequential reading. In independent reading, each case is read twice, once under each of the two reading conditions, with each reading separate by time period, typically a few weeks, to reduce the chance that the radiologist will remember the case. With a large number of cases, multiple reading sessions are required and readers are required to read under both conditions in each session. Half the cases are read first with aid and then without and in the other half the order is reversed. In sequential reading, the without aid condition is always first and after the radiologist has given their opinion without, the computer result is shown and the case is reassessed by the radiologist. In this way, the radiologist views each image once. The sequential reading method more closely resembles how CAD is used clinically. In addition, the power of the sequential method is higher because the two reading (with and without aid) are correlated.

The first CAD observer study was conducted by Chan et al. in 1990 (6). They showed the potential clinical benefits of CAD for the first time. Using 15 radiologists and 60 mammograms, half with a cluster of calcifications and half without, they showed that a CADe scheme design to detected clustered microcalcifications improved the radiologists performance at a statistically significant level: the AUC increase from 0.94 without aid to 0.97 with aid ($p < 0.001$). Their CADe scheme had a sensitivity of 85% with four false-positive detections per image.

The first observer study involving CADx was conducted by Getty et al. (111). In their study, 6 radiologists read 150 cases containing a benign lesion and containing a malignant lesion. They found that when using CADx radiologist, performance in classifying breast lesions increased significantly. The radiologists' sensitivity increased from 0.51 to 0.69 at a false-positive fraction of 0.1. Further, they showed that general radiologists (those who did not read mammograms full-time) when using the computer aid had comparable performance to expert radiologists (those who read mammograms full-time). In their technique, the radiologists subjectively extracted information from the image and this information was used as input to the CADx scheme.

A similar technique was developed by Wu et al. (112). They should that the CADx scheme could outperform the radiologists in classifying breast lesions. That is, using the information extracted by the radiologists, the CADx scheme was more accurate than the radiologists were. This suggests that radiologists can extract useful information from the mammogram, but they cannot synthesize the information to produce the correct classification. This was the first study to show that a CAD scheme could outperform a radiologist.

The first observer study involving CAD where the computer extracted the features was performed by Jiang et al. (113). In their study, 10 radiologists read 104 cases containing clustered calcifications (54 malignant and 60 benign). They found that the radiologist increased their AUC from 0.61 to 0.75 when they used the computer aid ($p < 0.0001$). In practical terms, they found that on average each radiologist recommended biopsies on 6.0 more malignant clusters ($p = 0.0006$) and 6.4 fewer benign clusters ($p = 0.003$). They also found that their CADx scheme outperformed the unaided radiologists (AUC of 0.80 vs. 0.61, $p < 0.0001$).

Using this observer study, Jiang et al. showed that the variability between radiologists was reduced when they used the CADx scheme (114). They further showed, theoretically, that the radiologists using CADx outperformed independent double reading with two radiologists. In independent double reading, two radiologists read each case independently and if either considers the case abnormal, then it is called abnormal. Another form of double reading is either having the two radiologist discuss the case if they initially disagree and reach a mutual decision or have a third radiologist break the tie. Jiang et al. also simulated double reading with a tiebreaker where they assumed that when two radiologists disagreed, the third radiologist always made the correct decision. This represents a theoretical upper limit on the performance of double reading. When compared to this type of double reading, a single radiologist using CADx had nearly comparable performance. This provides evidence that CADx could be used effectively to implement double reading.

There have been several other observer studies, using ROC analysis, for mammography, breast ultrasound and chest radiography that show the potential for CAD to improve radiologists' performance (115–127).

Clinical Studies. While observer studies provide evidence for the benefits of CAD, actual clinical studies need to be performed to prove that CAD is useful clinically.

The first reported study was conducted by Freer and Ulissey. Using a CADe system for screening mammography, they read 12,860 cases first without aid and then immediately after with aid (sequential reading). They found an additional eight cancers after viewing the computer detections: a 19% increase. This increase was not statistically significant, however, in part because of the smaller number of patients with cancer in the screened population. The use of CADe also increased the callback rate (the number of women consider abnormal) from 6.5 to 7.7% (statistically significant). Since the prevalence of cancer is low, most of the recalls are false positives. Gur et al. conducted the largest clinical study to date (128). They reported on 115,571 screening exams, 56,432 read before the implementation of CADe and 59,139 read after the implementation of CADe. In their study, they compared two time periods before and after CADe was implemented. They found that the number of cancers detected per 1000 women screen when using CADe increased only slightly from 3.49 to 3.55 (not statistically significant) and the callback rate also increased only slightly from 11.39 to 11.40% (not statistically significant). In their study, 17 radiologists were reading clinically. When the data were analyzed by the number of cases read,

those reading fewer cases, and by definition less experienced readers, had a increase in cancer detection from 3.05 to 3.65 (not statistically significant) while their callback rate changed from 10.5 to 12.0% (statistically significant, $p < 0.001$) (129).

Clinical evaluation of CAD is difficult, particularly when examining screening mammography. Since the prevalence of cancer in screening mammography is only 4/1000 women, large numbers of women need to be screened to collect enough cancer cases to have a statistically meaningful study. For example, if 25,000 women were screened, ~ 100 cancers would be present, of which ~ 20 maybe missed mammographically. If CADe can detect 50% of the missed cancers and the radiologist when using CADe recognize 100% of those as a miss, then 10 additional cancers would be detected. Measuring this small number of additional cancers is difficult statistically because there are several sources of variability. The number of cancers in screened population, the number of missed cancers, and the number of misses detected by the radiologist when using CADe can all fluctuate statistically speaking. To increase the statistical power of the study, a large number of cancers need to be detected. This can be accomplished by including more women in the study. As the number of women screened increases, it is likely that the results from more than one radiologist need to be included in the study. However, since there is a large variation between radiologists, having more than one observer will introduce another source of statistical uncertainty into the results. These issues need to be carefully addressed when planning a clinical study.

The ultimate endpoint for measuring the benefits of CAD are in a reduction in mortality and morbidity. Randomized controlled clinical trials are ideal for measuring this type of endpoint. For a CADe study, half the women would have their mammograms read without CADe and half would them read with CADe. The women would then be followed for a number of years and the number of deaths from breast cancer in the two groups can be measured and compared. Unfortunately, to have enough statistical power to measure a statistically significant decrease in mortality when using CADe is a very large number of women would need to participate in the trial (on the order of 50,000 or more) and these women would need to be followed for at least 5–10 years. Such a study is cost prohibited. Therefore, studies that do not require long-term follow up are the more likely ones to be preformed to measure the benefits of CAD.

ADVANCED APPLICATIONS

Radiologists rarely consider only a single image in making an interpretation. Most radiologic procedures involve multiple views of the body part of interest. Often there are images from more than one type of modality. Correlation of CT, chest X ray, and a nuclear medicine scan; and of multiple mammograms, ultrasound and breast MRI are just two examples of multimodality imaging. Further, images are often correlated with any clinical findings or patient history to improve the accuracy of the diagnosis. To be a more integrated tool, CAD schemes must be designed to help radiologist with multiimage, multimodality imaging.

Multiimage CAD

The classic example of multi-image CAD is in mammography, where there are two views taken of each breast. Radiologists compare images from the left and right breasts, since the breasts have a natural symmetry and normally they look similar. In addition, any finding in one image is correlated to possible findings in the corresponding second view. This helps the radiologist to determine if a finding is a superposition of normal tissue (a possibility if it is seen in only one view) or a real finding (if seen in both views).

Paquerault et al. developed a method to compare CADe detections between the craniocaudal and the mediolateral oblique views (130). They use a combination of geometric location and morphological and textural features to correlate computer detections between the two views. A major problem with this approach is that actual masses are not always detected in both views because they are not visible in both views or the computer misses the lesion in one of the two views.

For CADx combining results is somewhat simpler, since the output of a CADx scheme is a number that is related to the likelihood that the lesion is malignant. One could show the likelihood value for each image or combine them in some manner. There are different strategies for combining the two values, for example, averaging or choosing the maximum or minimum. The optimum strategy has been studied by Liu et al. (131). They found that the maximum performance in combining two estimates depends on the accuracy of the ROC curves that underlie the two estimates.

In CADe for chest imaging, Li et al. compare the right side to the left side of the chest radiograph, since, with some exceptions, there is symmetry between the two sides of the thorax (132). The comparison is done by subtracting the left and right sides of the chest radiograph, after correcting for the patient not being perfectly upright. The image is also warped to improve the match between structure in the left and right sides. This technique can be used to find any type of disease that appears as an asymmetry (e.g., lung cancer, pneumothorax, pneumonia, and emphysema).

Multimodality CAD

Currently, multimodality CAD only exists for breast imaging and only for classifying breast lesions, CADx. If a woman has an abnormal screening mammogram or some physical symptoms that indicate that she may have breast cancer, she receives a diagnostic work-up, which may include additional specialized mammograms, ultrasound, and a breast MRI. The radiologist interprets these images along with any clinical findings and patient data (e.g., family history) to decide whether or not to recommend the patient have a breast biopsy. The CADx, to be fully integrated into this process and provide the maximum amount of assistance, needs to be multimodality analyzing the lesion as it appears mammographically, sonographically,

and from the magnetic resonance (MR) images that usually includes both spatial and temporal information.

CAD Server

The current paradigm for clinical implementation of CAD is for each site to have a CAD system or multiple systems. Since most radiographs are acquired in digital form, the possibility exists for CAD to be implemented over the internet. A central CAD server could serve a number of sites. This scenario offers a number of possibilities over the current paradigm:

1. The CAD schemes from multiple vendors or multiple versions from the same vendor could be made available and the user could specify which algorithm they wanted for a specific image. For example, a radiologist may prefer to have the mass detection scheme from vendor A and the microcalcifications scheme from vendor B run on the mammograms.
2. The CAD server could offer archiving services. This would be appealing to small clinical sites that do not want to or cannot afford to purchase and maintain their own image archive (known as a Picture Archiving and Communication System, PACS).
3. More powerful computer systems could be employed. This offers the possibility of using more power image processing and artificial intelligence techniques that would be time prohibited on a common desktop computer.
4. Patient confidentiality issues not withstanding, a centralized CAD server would analyze hundreds or even thousands of images daily. With some effort, it would be possible to use these images to provide additional training images for the CAD algorithms.

A project called the National Digital Mammography Archive (NDMA) developed such a concept and demonstrated an initial test bed (133). The NDMA was a scalable large-scale image storage and retrieval system providing clinical service for digital mammograms. It provided a number of other services, such as CAD functionality, image retrieval for research purposes (automatically de-identifying patient data), multilevel security, and a radiology teaching component for tele-education in mammography. This was done in real-time, using virtual private networks over the high bandwidth Next Generation Internet (Internet 2). Similar projects are underway in Europe: the eDiamond project (134) and the GPCALMA project (Grid Platform for Computer Assisted Library for MAmmography) (135).

FUTURE STUDIES

Computer-aided diagnosis is still an emerging field, far from being mature. However, much progress has been made in the past 20 years to point where commercial systems are now available for detecting breast cancer from mammograms and detecting lung nodules from CT scans of the thorax or a chest radiograph. The CADe systems for detecting colon polyps from colonography (CT of the colon, as called virtual colonoscopy) are being developed. Not too surprisingly, all three applications are being used to screen for cancer. In mass screening, very few people screened have the diseased that is being screened for ($< 1\%$). Further, in mammography the radiologist views four views in total, while in CT of the chest and colon up to 400 slices of the body are obtained and viewed for each person. With this amount of information and with subtlety that a cancer can appear with, it is difficult for the radiologist to be ever vigilant. The CADe could play an important role in these situations.

While CADe schemes are available clinically, no clinical CADx system is available. In mammography, there is good evidence that CADx can be a useful aid to the radiologist, especially for differentiating between benign and malignant calcifications. As CT screening for lung cancer becomes more prevalent, there will be an increase in the number of lung biopsies that will need to be performed. Unlike a breast biopsy, however, a lung biopsy carries a bigger risk for complications. Therefore, a CADx scheme for distinguishing benign and malignant lung nodules would be very valuable. Two studies have shown that a CADx system can improve the performance of radiologists in deciding whether a lung nodule is malignant (136,137). It is anticipated that clinical CADx systems will be available in the future and these will need to undergo clinical evaluation.

BIBLIOGRAPHY

1. Winsberg F, et al. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. Radiology 1967;89:211–215.
2. Chan H-P, et al. Image feature analysis and computer-aided diagnosis in digital radiography. 1. Automated detection of microcalcifications in mammography. Med Phys 1987;14: 538–548.
3. Fujita H, Doi K, Fencil LE, Chua KG. Image feature analysis and computer-aided diagnosis in digital radiography. 2. Computerized determination of vessel sizes in digital subtraction angiography. Med Phys 1987;14:549–556.
4. Giger ML, Doi K, MacMahon H. Image feature analysis and computer-aided diagnosis in digital radiography. 3. Automated detection of nodules in peripheral lung fields. Med Phys 1988;15:158–166.
5. Katsuragawa S, Doi K, MacMahon H. Image feature analysis and computer-aided diagnosis in digital radiography. Detection and characterization of interstitial lung disease in digital chest radiographs. Med Phys 1988;15:311–319.
6. Chan H-P, et al. Improvement in radiologists' detection of clustered microcalcifications on mammograms: The potential of computer-aided diagnosis. Invest Radiol 1990;25:1102–1110.
7. Doi K, et al. Computer-aided diagnosis (CAD): Development of automated schemes for quantitative analysis of radiographic images. Semin Ultrasound CT MR 1992;13:140–152.
8. Anderson ED, Muir BB, Walsh JS, Kirkpatrick AE. The efficacy of double reading mammograms in breast screening. Clin Radiol 1994;49:248–251.
9. Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994;191:241–244.
10. Kopans DB. Double reading. Radiol Clin N Am 2000;38:719–724.
11. Destounis SV, et al. Can computer-aided detection with double reading of screening mammograms help decrease the false-negative rate? Initial experience. Radiology 2004;232: 578–584.

12. Astley S, et al. CADET: The computer-aided detection evaluation trial. In: Pisano E D, editor. Digital Mammography, 2004. Madison (WI): Medical Physics Publishing; 2005 (in press).
13. Lewin JM, et al. Comparison of full-field digital mammography to screen-film mammography for cancer detection: results of 4945 paired examinations. Radiology 2001;218:873–880.
14. Poplack SP,et al. Mammography in 53,803 women from the New Hampshire mammography network. Radiology 2000;217: 832–840.
15. Smith-Bindman R, et al. Physician predictors of mammographic accuracy. J Natl Cancer Inst 2005;97:358–367.
16. Forrest JV, Friedman PJ. Radiologic errors in patients with lung cancer. West J Med 1981;134:485–490.
17. Barlow WE, et al. Performance of diagnostic mammography for women with signs or symptoms of breast cancer. JNCI 2002;94:1151–1159.
18. Kopans DB. The positive predictive value of mammography. AJR 1992;158:521–526.
19. Elmore JG, et al. Ten-year risk of false positive screening mammograms and clinical breast examinations. N Engl J Med 1998;338:1089–1096.
20. Elmore JG, et al. International variation in screening mammography interpretations in community-based programs. JNCI 2003;95:1384–1393.
21. de Wolf CJ, et al. European guidelines for quality assurance in mammography screening. 2nd ed Luxembourg: European Commission, Europe Against Cancer Programme; 1996.
22. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. Arch Intern Med 1996;156:209–213.
23. Ciccone G, Vineis P, Frigerio A, Segnan N. Inter-observer and intra-observer variability of mammogram interpretation: a field study. Eur J Cancer 1992;28A:1054–1058.
24. Giger ML, Huo Z, Kupinski MA, Vyborny CJ. Computer-aided diagnosis in mammography. In: Sonka M, Fitzpatrick JM, editor. Handbook of Medical Imaging. Bellingham (WA): The Society of Photo-Optical Instrumentation Engineers; 2000.
25. Karssemeijer N. Detection of masses in mammograms. In: Strickland RN, editor. Image-Processing Techniques in Tumor Detection. New York: Marcel Dekker; 2002.
26. Karssemeijer N, Hendriks JH. Computer-assisted reading of mammograms. Eur Radiol 1997;7:743–748.
27. Nishikawa RM. Detection of microcalcifications. In: Strickland RN, editor. Image-Processing Techniques in Tumor Detection. New York: Marcel Dekker; 2002.
28. Sampat MP, Markey MK, Bovik AC. Computer-aided detection and diagnosis in mammography. In: Bovik AC, editor. The Handbook of Image and Video Processing. New York: Elsevier; 2005.
29. Astley SM, Gilbert FJ. Computer-aided detection in mammography. Clin Radiol 2004;59:390–399.
30. Gale AG, Astley SM, Dance DR, Cairns AY. Digital Mammography Amsterdam: Elsevier; 1994.
31. Doi K, Giger ML, Nishikawa RM, Schmidt RA. Digital Mammography '96. Amsterdam: Elsevier Science; 1996.
32. Karssemeijer N, Thijssen M, Hendriks J, van Erning L. Digital Mammography Nijmegen 98. Amsterdam: Kluwer Academic Publishers; 1998.
33. Yaffe MJ. Digital Mammography 2000 Madison (WI): Medical Physics Publishing; 2000.
34. Peitgen HO. Digital Mammography IWDM 2002 Berlin: Springer-Verlag; 2003.
35. Pisano ED. Digital Mammography 2004 Madison (WI): Medical Physics Publishing; 2005.
36. Zheng B, et al. Computer-aided detection of clustered microcalcifications in digitized mammograms. Acad Radiol 1995;2: 655–662.
37. Pisano ED, et al. Radiologists' preferences for digital mammographic display. Radiology 2000;216:820–830.
38. Merkle R, Laine AF, Smith SJ. Evaluation of a multiscale enhancement protocol for digital mammography. In: Strickland RN, editor. Image-Processing Techniques for Tumor Detection. New York: Macel Dekker; 2002.
39. Zhang W, Yoshida H, Nishikawa RM, Doi K. Optimally weighted wavelet transform based on supervised training for detection of microcalcifications in digital mammograms. Med Phys 1998;25:949–956. Yoshida H, et al. An improved computer-assisted diagnostic scheme using wavelet transform for detecting clustered microcalcifications in digital mammograms. Acad Radiol 1996;3:621–627.
40. Brown S, et al. Development of a multi-feature CAD system for mammography. In: Karssemeijer N, Thijssen M, Hendriks J, van Erning L, editors. Digital Mammography Nijmegen 98. Amsterdam: Kluwer Academic Publishers; 1998.
41. Chen CH, Lee GG. On digital mammogram segmentation and microcalcification detection using multiresolution wavelet analysis. Graphical Models Image Processing 1997;59:349–364.
42. Chitre Y, et al. Classification of mammographic microcalcifications using wavelets. Proc SPIE 1995;2434:48–55.
43. Kallergi M. Computer-aided diagnosis of mammographic microcalcification clusters. Med Phys 2004;31:314–326.
44. Laine A, Song S, Fan J. Adaptive multiscale processing for contrast enhancement. Proc SPIE 1993;1905:521–532.
45. Lo S-CB, et al. Detection of clustered microcalcifications using fuzzy modeling and convolution neural network. Proc SPIE 1996;2710:8–15.
46. Strickland RN, Hahn H. Wavelet transforms for detecting microcalcifications in mammograms. IEEE Trans Med Imaging 1996;15:218–229.
47. Wang TC, Karayiannis NB. Detection of microcalcifications in digital mammograms using wavelets. IEEE Trans Med Imaging 1998;17:498–509.
48. Zhang W, Yoshida H, Nishikawa RM, Doi K. Optimally weighted wavelet transform based on supervised training for detection of microcalcifications in digital mammograms. Med Phys 1998;25:949–956. Yoshida H, et al. An improved computer-assisted diagnostic scheme using wavelet transform for detecting clustered microcalcifications in digital mammograms. Acad Radiol 1996;3:621–627.
49. Jiang Y, et al. Method of extracting microcalcifications' signal area and signal thickness from digital mammograms. Proc SPIE 1992;1778:28–36.
50. Veldkamp WJ, Karssemeijer N. Accurate segmentation and contrast measurement of microcalcifications in mammograms: a phantom study. Med Phys 1998;25:1102–1110.
51. Barnes GT. Radiographic mottle: a comprehensive theory. Med Phys 1982;9:656–667.
52. Lefebvre F, et al. A fractal approach to the segmentation of microcalcifications in digital mammograms. Med Phys 1995;22: 381–390.
53. Spiesberger W. Mammogram inspection by computer. IEEE Trans Biomed Eng 1979;26:213–219.
54. Carman CS, Eliot G. Detecting calcifications and calcification clusters in digitized mammograms. In: Doi K, Giger ML, Nishikawa RM, Schmidt RA, editors. Digital Mammography '96. Amsterdam: Elsevier Science; 1996.
55. Bankman IN, et al. Automated recognition of microcalcification clusters in mammograms. Proc SPIE 1993;1905:731–738.
56. Chan H-P, et al. Computer-aided detection of microcalcifications in mammograms: Methodology and preliminary clinical study. Invest Radiol 1988;23:664–671.
57. Kobatake H, Takeo H, Nawano S. Microcalcification detection system for full-digital mammography. In: Karssemeijer N, Thijssen M, Hendriks J, van Erning L, editors. Digital

Mammography Nijmegen 98. Amsterdam: Kluwer Academic Publishers; 1998.
58. Zhao D, Shridhar M, Daut DG. Morphology on detection of calcifications in mammograms. Proc SPIE 1993;1905:702–715.
59. Fukuoka D, et al. Automated detection of clustered microcalcifications on digitized mammograms. In: Karssemeijer N, Thijssen M, Hendriks J, van Erning L, editors. Digital Mammography Nijmegen 98. Amsterdam: Kluwer Academic; 1998.
60. Kobatake H, Jin H-R, Yoshinaga Y, Nawano S. Computer diagnosis of breast cancer by mammogram processing. In: Lemke H, Inamura K, Jaffe C, Felix R, editors. Computer Assisted Radiology. Berlin: Springer-Verlag; 1993.
61. Davies DH, Dance DR. Automatic computer detection of clustered calcifications in digital mammograms. Phys Med Biol 1990;35:1111–1118.
62. Fam BW, Olson SL, Winter PF, Scholz FJ. Algorithm for the detection of fine clustered calcifications on film mammograms. Radiology 1988;169:333–337.
63. Ema T, et al. Image feature analysis and computer-aided diagnosis in mammography: Reduction of false-positive clustered microcalcifications using local edge-gradient analysis. Med Phys 1995;22:161–169.
64. Fujita H, et al. Automated detection of masses and clustered microcalcifications on mammograms. Proc SPIE 1995;2434:682.
65. Mascio LN, Hernandez JM, Logan CM. Automated analysis for microcalcifications in high resolution digital mammograms. Proc SPIE 1993;1898:472–479.
66. Shen L, Rangayyan RM, Desautels JEL. Detection and classification of mammographic calcifications. Int J Pat Recog Artif Intell 1993;7:1403–1416.
67. Bottema MJ, Slavotinek JP. Detection of subtle microcalcifications in digital mammograms. In: Karssemeijer N, Thijssen M, Hendriks J, van Erning L, editors. Digital Mammography Nijmegen 98. Amsterdam: Kluwer Academic; 1998.
68. Magnin IE, El Alaoui M, Bremond A. Automatic microcalcifications pattern recognition from X-ray mammographies. Proc SPIE 1989;1137:170–175. Cheng H-D, Lui YM, Freimanis RI. A novel approach to microcalcification detection using fuzzy logic technology. IEEE Trans Med Imaging 1998;17:442–450.
69. Kupinski M, Giger ML. Feature selection with limited datasets. Med Phys 1999;26:2176–2182.
70. Highman R, Brady M. Mammographic Image Analysis. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2000.
71. Zheng B, et al. Computer-aided detection of clustered microcalcifications in digitized mammograms. Acad Radiol 1995;2:655–662.
72. Diahi JG, et al. Evaluation of a neural network classifier for detection of microcalcifications and opacities in digital mammograms. In: Karssemeijer N, Thijssen M, Hendriks J, van Erning L, editors. Digital Mammography Nijmegen 98. Amsterdam: Kluwer Academic Publishers; 1998.
73. Nagel RH, Nishikawa RM, Doi K. Analysis of methods for reducing false positives in the automated detection of clustered microcalcifications in mammograms. Med Phys 1998;25:1502–1506.
74. Lure FYM, Gaborski RS, Pawlicki TF. Application of neural network-based multi-stage system for detection of microcalcification clusters in mammogram images. Proc SPIE 1996;2710:16–23.
75. Davies DH, Dance DR. The automatic computer detection of subtle calcifications in radiographically dense breasts. Phys Med Biol 1992;37:1385–1390.
76. Hojjatoleslami SA, Kittler J. Detection of clusters of microcalcifications using a K-nearest neighbour rule with locally optimum distance metric. In: Doi K, Giger ML, Nishikawa RM, Schmidt RA, editors. Digital Mammography '96. Amsterdam: Elsevier Science; 1996.
77. Cernadas E, et al. Detection of mammographic microcalcifications using a statistical model. In: Karssemeijer N, Thijssen M, Hendriks J, van Erning L, editors. Digital Mammography Nijmegen 98. Amsterdam: Kluwer Academic Publishers; 1998.
78. Anastasio MA, et al. A genetic algorithm-based method for optimizing the performance of a computer-aided diagnosis scheme for detection of clustered microcalcifications in mammograms. Med Phys 1998;25:1613–1620.
79. Anastasio MA, Kupinski MA, Nishikawa RM. Optimization and FROC analysis of rule-based detection schemes using a multiobjective approach. IEEE Trans Med Imaging 1998;17:1089–1093.
80. Wu YZ, Doi KN, Giger ML, Nishikawa RM. Computerized detection of clustered microcalcifications in digital mammograms - Applications of artificial neural networks. Med Phys 1992;19:555–560.
81. Sajda P, Spence C, Pearson J. Learning contextual relationships in mammograms using a hierarchical pyramid neural network. IEEE Trans Med Imaging 2002;21:239–250.
82. Stafford RG, Beutel J, Mickewich DJ. Application of neural networks to computer-aided pathology detection in mammography. Proc SPIE 1993;1898.
83. El-Naqa I, Yang Y, Nishikawa RM, Wernick MN. A support vector machine approach for detection of microcalcifications. IEEE Trans Med Imaging 2002;21:1552–1563.
84. Nishikawa RM, et al. Computer-aided detection of clustered microcalcifications on digital mammograms. Med Biol Engin Comput 1995;33:174–178.
85. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283–298.
86. Metz CE. Fundamental ROC Analysis. In: Beutel H, Kundel J, Van Metter RL, editors. Handbook of Medical Imaging. Bellingham (WA): SPIE; 2000.
87. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. Invest Radiol 1992;27:723–731.
88. Edwards DC, et al. Estimating three-class ideal observer decision variables for computerized detection and classification of mammographic mass lesions. Med Phys 2004;31:81–90.
89. Chan H-P, et al. Design of three-class classi ers in computer-aided diagnosis: Monte carlo simulation study. Proc SPIE 2003;5032:567–578.
90. Dreiseitl S, Ohno-Machado L, Binder M. Comparing three-class diagnostic tests by three-way ROC analysis. Med Decis Making 2000;20:323–331.
91. Mossman D. Three-way ROCs. Med Decis Making 1999;19: 79–89.
92. Jiang Y, Metz CE, Nishikawa RM. An ROC partial area index for highly sensitive diagnostic tests. Radiology 1996;201:745–750.
93. McClish DK. Analyzing a portion of the ROC curve. Med Decis Making 1989;9:190–195.
94. Chakraborty DP. The FROC, AFROC and DROC Variants of the ROC analysis. In: Beutel J, Kundel H, Van Metter R, editors. Handbook of Medical Imaging. Bellingham (WA): SPIE; 2000.
95. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: modeling, analysis, and validation. Med Phys 2004;31:2313–2330.
96. Nishikawa RM, et al. Effect of case selection on the performance of computer-aided detection schemes. Med Phys 1994;21:265–269.
97. Nishikawa RM. Mammographic databases. Breast Dis 1998; 10:137–150.
98. Revesz G, Kundel HL, Bonitatibus M. The effect of verification on the assessment of imaging techniques. Invest Radiol 1983;18:194–198.

99. Fukunaga K, Hayes RR. Effects of sample size on classifier design. IEEE Trans Pattern Anal Machine Intell 1989;11: 873–885.
100. Sahiner B, et al. Feature selection and classifier performance in computer-aided diagnosis: the effect of finite sample size. Med Phys 2000;27:1509–1522.
101. Chan HP, Sahiner B, Wagner RF, Petrick N. Effects of sample size on classifier design for computer-aided diagnosis. Proc SPIE 1998;3338:846–858.
102. Chen DR, et al. Use of the bootstrap technique with small training sets for computer-aided diagnosis in breast ultrasound. Ultrasound Med Biol 2002;28:897–902.
103. Zheng B, Chang YH, Good WF, Gur D. Adequacy testing of training set sample sizes in the development of a computer-assisted diagnosis scheme. Acad Radiol 1997;4: 497–502.
104. Tourassi GD, Floyd CE. The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis. Med Decis Making 1997;17:186–192.
105. Efron B, Tibshirani RJ. An Introduction to the Bootstrap New York: Chapman and Hall; 1993.
106. Brem RF, et al. Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial. AJR 2003;181:687–693.
107. Burhenne LJW, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000;215:554–562.
108. Moberg K, et al. Computed assisted detection of interval breast cancers. Eur J Radiol 2001;39:104–110.
109. te Brake GM, Karssemeijer N, Hendriks HCL. Automated detection of breast carcinomas not detected in a screening program. Radiology 1998;207:465–471.
110. Armato SG III, et al. Performance of automated CT nodule detection on missed cancers from a lung cancer screening program. Radiology 2002;225:685–692.
111. Getty DJ, Pickett RM, D'Orsi CJ, Swets JA. Enhanced interpretation of diagnostic images. Invest Radiol 1988;23:240–252.
112. Wu Y, et al. Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer. Radiology 1993;187:81–87.
113. Jiang Y, et al. Improving breast cancer diagnosis with computer-aided diagnosis. Acad Radiol 1999;6:22–33.
114. Jiang Y, et al. Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications. Radiology 2001;220: 787–794.
115. Awai K, et al. Pulmonary nodules at chest CT: effect of computer-aided diagnosis on radiologists' detection performance. Radiology 2004;230:347–352.
116. Brown MS, et al. Computer-aided lung nodule detection in CT: results of large-scale observer test. Acad Radiol 2005;12:681–686.
117. Chan HP, et al. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. Radiology 1999;212:817–827.
118. Fraioli F, et al. Evaluation of effectiveness of a computer system (CAD) in the identification of lung nodules with low-dose MSCT: scanning technique and preliminary results. Radiol Med (Torino) 2005;109:40–48.
119. Hadjiiski L, et al. Improvement in radiologists' characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: an ROC study. Radiology 2004;233:255–265.
120. Huo Z, Giger ML, Vyborny CJ, Metz CE. Effectiveness of computer-aided diagnosis: Observer study with independent database of mammograms. Radiology 2002;224:560–568.
121. Kakeda S, et al. Improved detection of lung nodules on chest radiographs using a commercial computer-aided diagnosis system. AJR Am J Roentgenol 2004;182:505–510.
122. Kegelmeyer WP Jr, et al. Computer-aided mammographic screening for spiculated lesions. Radiology 1994;191:331–337.
123. Kobayashi T, et al. Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. Radiology 1996;199:843–848.
124. Li F, et al. Radiologists' performance for differentiating benign from malignant lung nodules on high-resolution CT using computer-estimated likelihood of malignancy. AJR Am J Roentgenol 2004;183:1209–1215.
125. MacMahon H, et al. Computer-aided diagnosis of pulmonary nodules: results of a large-scale observer test. Radiology 1999;213:723–726.
126. Shah SK, et al. Solitary pulmonary nodule diagnosis on CT: results of an observer study. Acad Radiol 2005;12:496–501.
127. Shiraishi J, Abe H, Engelmann R, Doi K. Effect of high sensitivity in a computerized scheme for detecting extremely subtle solitary pulmonary nodules in chest radiographs: observer performance study. Acad Radiol 2003;10:1302–1311.
128. Gur D, et al. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. JNCI 2004;96:185–190.
129. Feig SA, Sickles EA, Evans WP, Linver MN. Re: Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. JNCI 2004;96:1260–1261.
130. Paquerault S, et al. Improvement of computerized mass detection on mammograms: fusion of two-view information. Med Phys 2002;29:238–247.
131. Liu B, Metz CE, Jiang Y. An ROC comparison of four methods of combining information from multiple images of the same patient. Med Phys 2004;31:2552–2563.
132. Li Q, et al. Contralateral subtraction: a novel technique for detection of asymmetric abnormalities on digital chest radiographs. Med Phys 2000;27:47–55.
133. Schnall MD, et al. National digital mammography archive. In: Karellas A, Giger ML, editors. RSNA 2004 Categorical Course: Advances in Breast Imaging: Physics, Technology, and Clinical Applications. Oak Brook (IL): Radiological Society of North America; 2004.
134. Lloyd S, et al. Digital mammography: a world without film? Methods Inf Med 2005;44:168–171.
135. Cerello P, et al. GPCALMA: a Grid-based tool for mammographic screening. Methods Inf Med 2005;44:244–248.
136. Shah S, et al. Computer-aided diagnosis of the solitary pulmonary nodule. Acad Radiol 2005;12:570–575.
137. Aoyama M, et al. Automated computerized scheme for distinction between benign and malignant solitary pulmonary nodules on chest images. Med Phys 2002;29:701–708.

See also BIOTELEMETRY; COMPUTERS IN THE BIOMEDICAL LABORATORY; MEDICAL EDUCATION, COMPUTERS IN; TELERADIOLOGY.

COMPUTERS IN CARDIOGRAPHY.

See ELECTROCARDIOGRAPHY, COMPUTERS IN.

COMPUTERS IN THE BIOMEDICAL LABORATORY

GORDON SILVERMAN
Manhattan College

INTRODUCTION

Activities of daily living as well as within the industrial and scientific world have come to be domynated by machines

that are designed to reduce the expenditure of human effort and increase the efficiency with which tasks are completed. The technology underlying these devices is heavily dependent on computers that are designed for maximum flexibility, modularity, and "intelligence". A familiar example of flexibility comes to us from the automobile industry where a car manufacturer may use a single engine design for a large variety of models. These automobiles may also contain microcomputers that can sense and "interpret" road surfaces and adjust the car's suspension system to maximize riding comfort. This type of design is characterized by the development of "functional" components that can be reused. The content of these elements, or "*objects*" as they are called, can be changed as technology improves, as long as access (or use) is not compromised by such changes. The characteristics of flexibility, modularity, and intelligence also exemplify instrumentation in the biomedical laboratory where, in recent years, laboratory sciences have become a matter of data and information processing. A user's understanding of the tools of data handling is as much a part of the scientist's skill set as performing a titration, conducting a clinical study, or simulating a biomedical model. Specific tasks in the biomedical (or other) laboratories have become highly automated through the use of intelligent instruments, robotics, and data processing systems. With the existence of >100 million internet-compatible computers throughout the world together with advanced software tools, the ability to conduct experiments "at-a-distance" opens such possibilities as remote control of such experiments and instruments, sharing of instruments among users, more efficient development of experiments, and report or publication of experimental data.

A new vocabulary has emerged to underscore the new computer-based biomedical laboratory environment. Terms such as *bioinformatics* (the scientific field that deals with the acquisition, storage, sharing and optimal use of information, data and knowledge) has come to characterize the biomedical laboratory culture. To understand contemporary biomedical laboratories and the role that the computer plays, it is necessary to consider a number of issues: the nature of data; how data is acquired; computer architectures; software formats; automation; and artificial intelligence.

Computers are instruments that process information, and within the biomedical laboratory they aid the analyst in recording, classifying, interpreting, and summarizing information. These outcomes encompass a number of specific instrument tasks:

- Data handling: acquisition, compression, reduction, interpretation, and record keeping.
- Control of laboratory instruments and utilities.
- Procedure and experiment development.
- Operator interface: control of the course of the experiment.
- Report production.

The biomedical laboratory environment is summarized in Fig. 1. Information produced by the experimental source

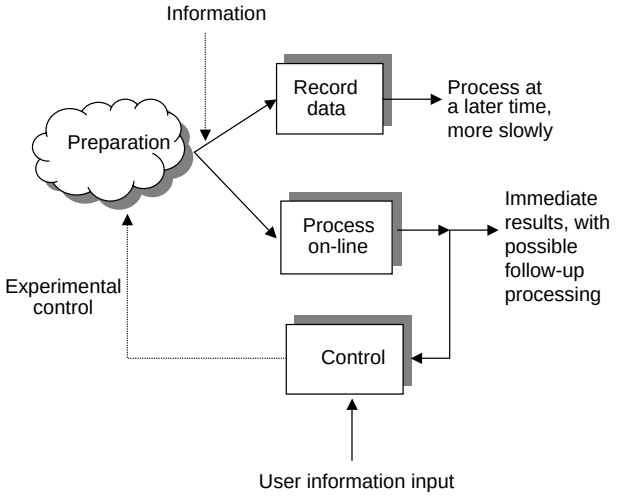


Figure 1. Information processing in laboratory environments.

may simply be recorded and the data analyzed at a later time. Alternatively, the data may be processed while the experiment proceeds and the results, possibly in reduced form, used to modify the experimental environment.

HISTORICAL ORIGINS

Scientific methodologies involving Biomedical laboratory research trace their origins to the experiments of Luigi Galvani in the 1789s with the study of "animal electricity" (1). This productive line of scientific investigation signals the start of the study of electrophysiology that continues to the present time. The "golden age" of electrophysiology began in the twentieth century (in particular starting ~1920) and was led by such (Nobel recognized) scientists as Gasser, Adrian, Hodgkin, Huxley, Eccles, Erlanger, and Hartline, to name but a few (2). These scientists introduced the emerging vacuum tube technologies (e.g., triodes) to observe, record, and subsequently analyze the responses of individual nerve fibers in animal neural systems. The ingenious arrangements that the scientists used were based on an "analog computer" model. A sketch of such arrangements (3) for measuring action potentials is shown in Fig. 2. Of particular value was the cathode ray tube that could display the "rapid" electrochemical changes

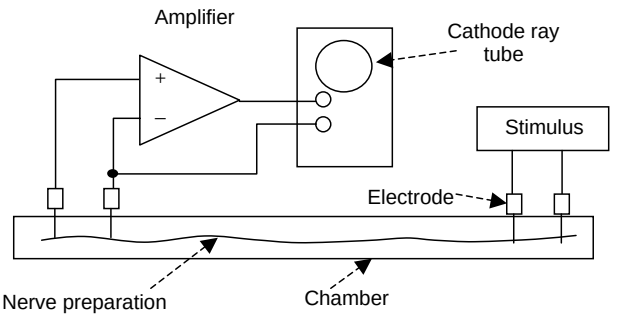


Figure 2. Sketch of an early laboratory arrangement for measuring response of nervous tissue.

from nervous tissue; this was a vast improvement over the string galvanometer that was previously used. A "highly automated" equivalent of this model with a digital computer architecture has been developed by Olansen et al. (4). Numerous electronic advances were made during the 1920s and 1930; Jan Toennies was one of the first bioengineers to design and build vacuum tube-based cathode followers and differential amplifiers. These advances found their way into the military technology of radar and other electronic devices of World War II. During the conflict, the electrophysiologists were pressed into service to develop the operational amplifier circuits that formed the basis of "computation" in the conduct of the war (e.g., fire- control systems). While numerous advances in speed, and instrumental characteristics (e.g., increased input impedance, noise reduction) characterize wartime developments, equipment available in 1950 to continue electrophysiological work precluded rapid analysis of results; it took many weeks to calculate experimental results, a task that was limited to "pencil and paper" computations aided by electromechanical calculators. All this changed with the introduction of digital technology and the digital computer. H.K. Hartline was one of the first of the Nobel Laureates to automate the electrophysiological laboratory with the use of the digital computer. A highly schematic representation of Hartline's experimental configuration is shown in Fig. 3 (5–7).

The architecture suggested in Fig. 3 became a fundamental model for information processing in the biomedical laboratory. However, within the digital computer, a number of (architectural) modifications have been introduced since the 1950s in order to improve informational *throughput*: the ability to complete data processing from acquisition to analysis to recording (as measured in "jobs/s").

DATA IN THE LABORATORY

An appreciation of the data underlying experiments in the biomedical laboratory is essential to successful implementation of a computer-based instrument system. The processing of experimental data is heavily dependent on the amount of information generated by the experimental preparation and the rate at which such data are to be processed by the computer. Both the quantity of data (e.g., the number of samples to be recorded) and the rate at which the data are to be processed can be estimated.

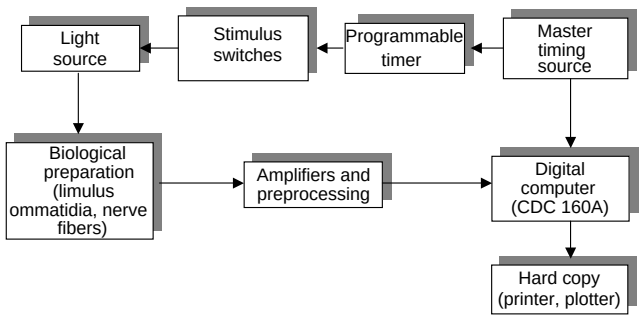


Figure 3. Early architecture of computer-automated biomedical laboratory environment. (After Hartline.)

These factors have important influence on the characteristics of the computer that is to be used in the design. A unit of information is the *bit* (as defined below), and the rate at which information is to be processed is determined by the *capacity* of the experimental environment (including any communications between the preparation and the rest of the system). The units of *system capacity* are bits/s. The following formulas are used to calculate these parameters:

$$\text{Information(bits)} = H = -\sum_i p_i \log_2(p_i)$$
$$\text{Capacity(bits/s)} = C = H/T$$

In these formulas, p_i is the probability of experimental outcome i (the experiment may have a finite number of N outcomes) and $\log_2$ is the logarithm to the base 2. The "T" parameter is the time required to transmit the data from the preparation to the instrumental destination. (It must include any required processing time as well.) If all outcomes are equally likely (i.e., probable) then it can be readily shown that $2^H = N$, where N is the total number of possible outcomes (and H is the information content in bits). The informational bit (H) is not to be confused with the term bit associated with binary digits. [Correlation of information (H) and the number of required binary processing digits may be correlated after appropriate coding.]

As an example of an appropriate calculation, consider one scientific temperature-measuring system that can report temperatures from -50 to $150\,^{\circ}$C in increments of $0.01\,^{\circ}$C. The system can thus produce 20,000 distinct results. If each of these outcomes is equally likely or probable, then the amount of information that must be processed amounts to 14.29 bits. Further, if the data processing system requires 0.1 s to generate a reading, then the capacity of the system is 142.9 bits/s. (As one cannot realistically subdivide a binary digit, 15 binary bits would be needed within the processing system.)

The formulas noted above do not describe the format of the data. For example, how many decimal places should be included. Generally, experiments are designed either to confirm or refute a theory, or to obtain the characteristics of a biological element (for purposes of this discussion). Within the laboratory there may be many variables that affect the data. In experimental environments, the results may often depend on two variables: one is the independent variable and the second, which is functionally related to the independent variable, is specified as the dependent variable. Other variables may act as parameters that are held constant for any given experimental epoch. Examples of independent variables include time, voltage, magnetizing current, light intensity, temperature, frequency (of the stimulating energy source), and chemical concentration. Dependent variables may also come from this list in addition to others.

Two possibilities exist for the experimental variables: Their domains (values) may either be continuous or discrete (i.e., having a fixed number of decimal places) in nature. As a consequence, there are four possible combinations for the independent and dependent variables within the functional relationship; these are shown in Fig. 4, where the format "independent/dependent" applies to the axes. In Fig. 4, the solid lines represent actual values

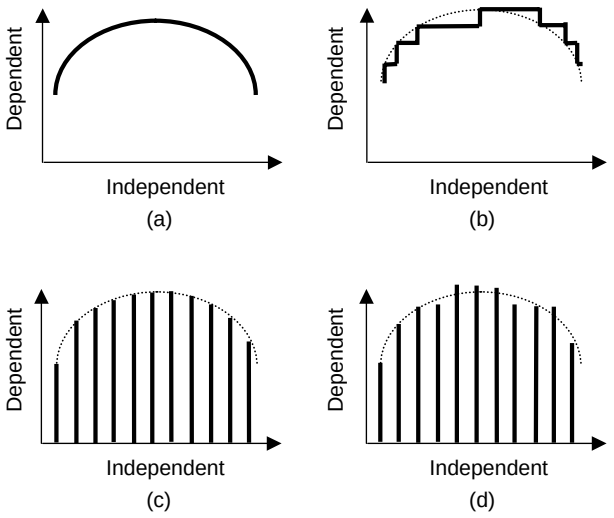


Figure 4. Underlying information formats for biomedical laboratory data: (a) continuous–continuous; (b) continuous–discrete; (c) discrete–continuous; (d) discrete–discrete.

generated by the experimental preparation. The bold vertical lines emphasize the fact that measurements are taken only at discrete sampling times. "Staircase-like" responses indicate that only discrete values are possible for the dependent variable. Although any of the formats are theoretically possible, contaminants (noise for the dependent variable and bandwidth for the independent variable) usually limit the continuum of values. Thus, all instances of information in the biomedical laboratory are ultimately discrete–discrete (Fig. 4d) in nature. (In Figs. 4b–d, the dotted line reflects the original signal.)

The ability of a computer-based instrument system to discriminate between datum points is measured by its *resolution* and reflects the number of distinct values that a variable can assume. In the temperature-measuring example, there were 20,000 distinctly possible readings and consequently the resolution was 0.01 °C. When resolution is combined with the range of values that the variable can assume, the number of distinct experimental outcomes can be computed:

$$\text{Number of unique experimental outcomes} = \text{range}/\text{resolution}$$

Summarizing the temperature-measuring system example in light of this relationship we conclude: resolution = 0.01 (°C); range = 200 (°C); number of outcomes = 20,000. Within a computer-based data acquisition (DAQ) system, the values may appear as coded representations of the underlying outcomes. The codes might be related to the equivalent decimal values of the original data. However, one could also assign an arbitrary code to each outcome. Because computer-based systems are designed to interpret codes that have two distinct states (or symbols), binary coding systems are normally used in laboratory applications. A variety of binary coding schemes are possible. A simple, but effective, code employs the binary number system to represent experimental outcomes. A number in this system is a weighted combination of the two

Table 1. Binary Coded Outcomes of Experimental Data

Outcome	Binary Code
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111
8	1000
9	1001
10	1010
11	1011
12	1100
13	1101
14	1110
15	1111

symbols that are recognized in the binary number system, namely, 0 and 1. A complete representation of a binary number is given by

$$a_n 2^n + a_{n-1} 2^{n-1} + \cdots + a_1 2^1 + a_0 2^0 \cdot a_{-1} 2^{-1} + a_{-2} 2^{-2} + \cdots$$

where all coefficients (a values) are either 1 or 0. Starting with the least significant digit (2^0), the positional weights for the positive powers of 2 are 1, 2, 4, 8, 16, and so on. For the negative powers of 2, the weights in increasingly smaller values are 1/2, 1/4, 1/8, and so on. Table 1 contains a list of 16 possible outcomes from a (low resolution) laboratory experiment including both the binary and decimal equivalents. The outcomes might represent values of voltage, time, frequency, temperature, or other experimental variables.

The elements of a computer-based information processing system for biomedical laboratories are generally compatible with the binary system previously discussed. However, human users of such machines are accustomed to the decimal number system (as well as the alphanumeric characters of their native language). Within the system, internal operations are carried out using binary numbers and calculations. Binary results are often translated into decimal form before presentation to a user; numerical inputs, if in decimal format, are translated (by the computer) into binary format before use within the computer. Other number systems may be found in a computer application. These include octal systems (base 8) and the hexadecimal number system (base 16), where the base symbols include 0, 1,. . ., 9, A, B, C, D, E, F. A user may also be required to enter other forms of information such as characters that represent a series of instructions or a program. Several widely accepted codes exist for alphanumeric data, and some of these together with their characteristics are shown in Table 2.

Coded information such as that shown in Tables 1 and 2, may be passed (i.e., transmitted) between different elements of a computer-based laboratory information processing system. The communication literature provides a

Table 2. Partial List of Alphanumeric Codes

Name of Code	Number of Bits	Number of Available Code Combinations
Extended Binary code		
Decimal Interchange Code (EBCDIC)	8	256
American Standard Code		
For Information Interchange (ASCII)	7	128
ASCII-8		
8-bit extension of ASCII	8	256
Hollerith	12	4096

rather complete description of the technology (8,9), and while it is not immediately germane to many circumstances of this discussion, some elements need to be mentioned. For example, the "internet" should be noted as an emerging development in computer-based biomedical laboratories.

There are two general protocols for transmitting laboratory data from the information source to its destination. Each part of the coded information (i.e., the bit) may be passed via a single communication channel. The channel element is the media and it might be wire, fiber optic cable, or air (as in wireless). Since there is only one channel, the data is passed in serial fashion, one bit at a time. An alternative arrangement permits the bits to be passed all at once (in parallel), but this requires an independent path for each bit such as a multiwire architecture. Parallel transmissions have inherently greater capacity than serial schemes. For example, if it requires 1 μs (i.e., 10^{-6} s) to transmit a bit, then a parallel transmission, using an 8-bit code, can pass 8×10^6 bits/s (i.e., 8 Mbits/s). An equivalent serial system would only have a capacity of 1 Mbit/s as it would require 8 μs for complete transmission of the code representing one of the possible experimental outcomes. Note that serial systems have a decided economic advantage over parallel schemes.

There are circumstances when two-way communication between elements of a computer-based data processing system is necessary. One element (e.g., the computer) may initiate a measurement instruction to a remotely located laboratory instrument (e.g., a spectrophotometer); the remote unit, in turn, responds with a set of measurements. Instructions, data, and parameters may need to pass from the computer, and status information (e.g., a busy signal) or results must be able to pass from the instrument to the computer: all of this over a serial path. Serial architectures can occur over a one-way (i.e., single-lane highway) or a two-way (i.e., two-lane highway) link. Figure 5 summarizes these communication alternatives.

In the half-duplex case, a single path must suffice for two-way communication. For proper transmission, the path must be made ready for communication in an appropriate direction before communication starts.

OVERALL ARCHITECTURE OF A COMPUTER-BASED BIOMEDICAL LABORATORY SYSTEM

The hardware for contemporary computer-based laboratory instrument systems reflects an information-processing model as shown in Fig. 1. A broad representative computer-based DAQ, and processing system is shown in Fig. 6. The elements depicted in the figure can be divided into several categories: the computer [PC, laptop, personal digital assistant (PDA)], sensors (or transducers), signal conditioning components, DAQ, software, and other elements for other aspects of computer-based environments (remote instrumentation, external processors, vision/imaging equipment, and motion control apparatus).

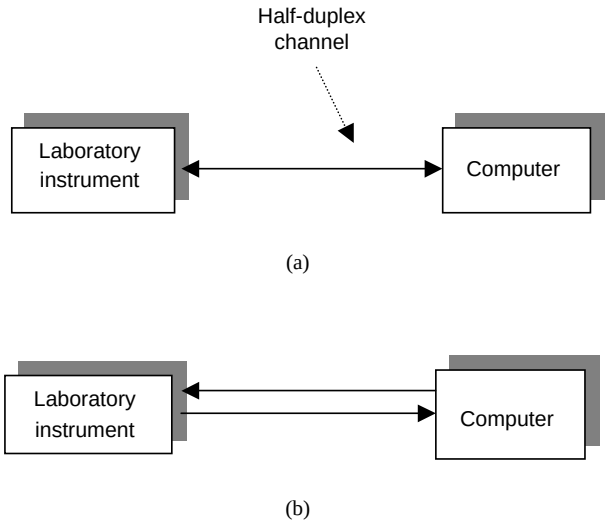


Figure 5. Serial communication alternatives: (a) half-duplex; (b) full-duplex.

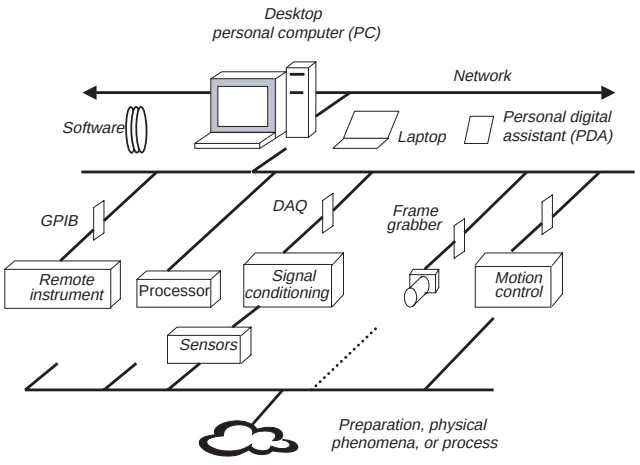


Figure 6. General arrangement of computer-based information processing system for biomedical laboratories.

Table 3. Sampling of Sensors for Biomedical Laboratory Applications

Application	Sensor Technology
Position	Resistive (potentiometric, goniometric), shaft Encoder, linear variable differential transformer (LVDT), capacitive, piezoelectric
Velocity	LVDT
Acceleration	LVDT, strain gauge, piezoelectric (attached to an elastic flexure), vibrometer
Force	Strain gauge, piezoelectric, LVDT, resistive, capacitative (all making use of a flexible attachment)
Pressure	Strain guage, piezoelectric, also LVDT and capacitative.
Flow	Measure pressure drop which is correlated to flow rate via a calibration function (Venturi tube, pitot tube)
Temperature	Thermoresistive (Seebeck), thermistor (semiconductor), resistive (platinum wire)
Light	Photocell, photoresistor, photodiode, phototransistor.

The computer in such systems has considerable impact on the maximum throughput and, in particular, often limits the rate at which one can continuously acquire data. New bus (communication) facilities in the modern computer have greatly increased speed capabilities. A limiting factor for acquiring large amounts of data is often the hard drive (secondary storage system in the computer). Applications requiring "real-time" processing of high frequency signals often make use of an external (micro)processor to provide for preprocessing of data. (The term *real-time* refers to a guaranteed time to complete a series of calculations.) With the rapid development of new technologies, and the reduction is size, laptops and PDAs have found their way into the laboratory, particularly when the data is to be accumulated in isolated sites as with many biological experiments. With appropriate application software, laptops can act as data loggers [simple recorders of source (raw) data are collected] and further processing subsequently completed on a PC (or other) computer. Even greater miniaturization now permits PDAs to collect and transmit data as well. With wireless technology, data can be e-mailed to a base station. Also noted in Fig. 6 is the potential to connect the computer to a network [local area network (LAN), wide area network (WAN), or Internet] with the possibility of conducting experiments under remote control.

Sensors (see sensors) or transducers (10,11) sense physical phenomena and produce electrical signals that DAQ components can ultimately accept after suitable signal conditioning. Transducers are grouped according to the physical phenomena being measured. These devices have an upper operating frequency above which they produce a signal that is no longer independent of the frequency of the source (phenomena) frequency. Position is the most common measurement and sensors normally translate such physical distortions into changes in the electrical characteristics of the sensor component. For example, capacitive transducers rely on the fact that the capacitance depends on the separation (or overlap) of its plates. Since the separation is a nonlinear function of the separation, capacitive sensors are usually combined with a conditioning circuit that produces a linear relationship between the underlying phenomena and the potential delivered to the DAQ. Changes in the dimensions of a resistor alter its resistance. Thus, when a thin wire is stretched, its resistance changes. This can be used to measure small displacements and generate a measure of strain. Inductive principles are employed to measure velocity. When a movable core passes through the center of a coil of wire, the (electromagnetic) coupling is altered in a way that can be used to determine its velocity. Table 3 summarizes several applications.

Electrical signals generated by the sensors often need to be modified so that they are suitable for the DAQ circuitry. A number of conditioning functions are carried out in the signal conditioning system: amplification (to increase measurement resolution and compatibility with the full-scale characteristic of the DAQ); linearization (to compensate for nonlinearities in the transducer such as those of thermocouples); isolation (of the transducer from the remainder of the system to minimize the possibility of electric shock); filtering (to eliminate noise or unwanted interference such as those frequencies that are erroneous (e.g., high frequency or those from the power lines); excitation provides external signal source requirements for the transducer (such as strain gages that require a resistive arrangement for proper operation).

The DAQs normally include an analogue-to-digital converter (ADC) for converting analog (voltage) signals into a binary (digital) quantity that can be processed by the computer. There are several well-developed techniques for performing the conversion (12). As a general principle, the concept shown in Fig. 7 can be used to explain the conversion process.

Resistors in a high-tolerance network are switched in a predetermined manner resulting in an output voltage that is a function of the switch settings. This voltage is compared to the unknown signal and when this reference equals the unknown voltage the switch sequence is halted.

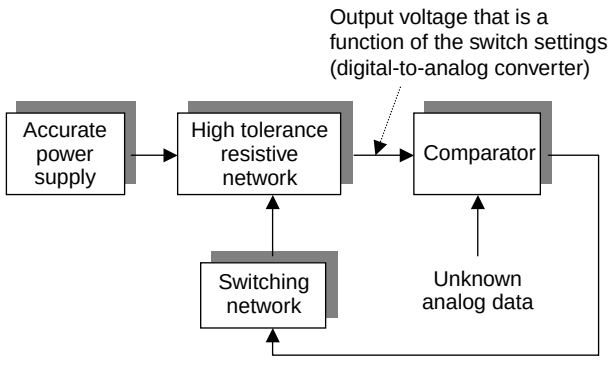


Figure 7. Principle of analog-to-digital conversion.

The pattern of switches then represents a digital quantity equivalent to the unknown analogue signal. (Other schemes are possible.)

Several characteristics of the DAQ need to be considered when specifying this part of a computer-based processing system.

Range: The spread in the value of the measurand (experimental input) over which the instrument is designed to operate.

Sensitivity: The change in the DAQs output for a unit change in the input.

Linearity: The maximum percentage error between an assumed linear response and the actual nonlinear behavior. (The user should be assured that the DAQ has been calibrated against some recognized standard.)

Hysteresis: Repeatability when the unknown is first increased from a given value to the limit of the DAQ (range) and then decreased to the same (given) value.

Repeatability: Max difference of the DAQ reading when the same input is repeatedly applied (often expressed as a percentage of the DAQs range).

Accuracy: Maximum degree to which an output differs from the actual (true) input. This summarizes all errors previously noted.

Resolution: Smallest change in the input that can be observed.

Time Constant: Time required for the DAQ to reach 63.2% of its final value from the sudden application of the input signal.

Rise Time: time required to go from 5 to 95% of its final output value.

Response Time: time that the DAQ requires reaching 95% of its final value.

Settling Time: Time that the DAQ requires to attain and/or remain within a given range of its final value (e.g., $\pm 2\%$ of its final value).

Delay Time: Time taken for the DAQ to reach 50% of its final value (not normally considered important).

Other sensor characteristics include: natural frequency, output impedance, mass, size, and cost.

Existing instruments may also be integrated into the computer-based environment if they include compatibility with a standard known variously as the General Purpose Interface Bus (GPIB) or IEEE 488: Originally developed by Hewlett-Packard (now Agilent) in 1965 to connect commercial instruments to computers (13). The high transfer rates (1MB/s) led to its popularity and it has evolved into an ANSI/IEEE Standard designated as 488.1, and subsequently as 488.2. The GPIB devices communicate with other such devices by sending device-dependent messages across the interface system (bus). These devices are classified as "Talkers", "Listeners", and/or "Controllers." A Talker sends data messages to one or more Listeners that receive the data. The Controller manages the flow of information on the bus by sending commands to all devices. For example, a digital voltmeter has the potential to be a Talker as well as Listener. The GPIB Controller is akin to the switching center of a telephone system. Such instruments may be connected to the computer system as seen in Fig. 6 as long as an appropriate component (card) is installed within the computer.

Images may be gathered from the biomedical laboratory using a (digital) camera and a suitable card within the computer. (See Olansen and Rosow in the *Reading List*.) Machine vision may be viewed as the acquisition and processing of images to identify or measure characteristics of objects. Successful implementation of a computer-based vision system requires a number of steps including:

Conditioning: Preparing the image environment including such parameters as light and motion.

Acquisition: Selecting image acquisition hardware (camera and lens) as well as software to be able to capture and display the image.

Analysis: Identification and interpretation of the image.

Computer-based (software) analysis of images takes into consideration the following:

Pattern Matching: Information about the presence or absence, number and location of objects (e.g., biological cells).

Positioning: Determining the position and orientation of a known object by locating features (e.g., a cell may have unique densities).

Inspection and *Examination*: Detecting flaws (e.g., cancer cells).

Gauging: Measuring lengths, diameters, angles and other critical dimensions. If the measurements fall outside a set of tolerance levels, then the object may be "discarded".

In addition to the acquisition of information from the laboratory, the computer system may be used to control a process such as an automated substance analysis using a robotic arm. (The motion control elements in Fig. 6 support such applications.) A sketch of such a system is shown in

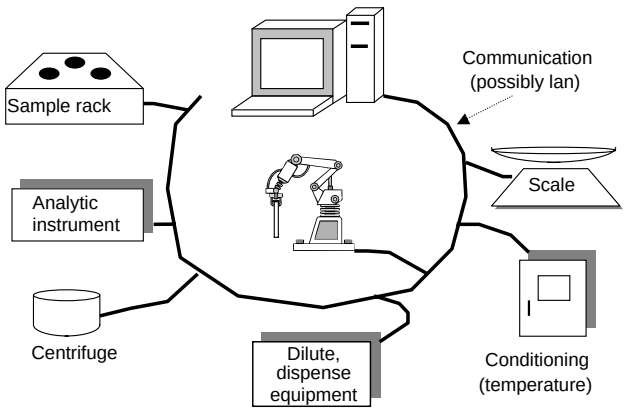


Figure 8. Computer-based architecture for an automated substance analysis system.

Fig. 8. The robotic arm would normally have several degrees of freedom (i.e., axes of motion) (14).

The system includes: a centrifuge (e.g., for analysis of blood samples); an analytic instrument (e.g., a spectrophotometer or chromatograph); a rack to hold the samples; a balance; a conditioning unit (possibly a stirrer or temperature oven); and instrumentation for dispensing, extracting and/or diluting chemicals. Various application programs within the computer could be used to precisely define the steps taken by the robotic arm to carry out a routine test. This program must also take into consideration the tasks to be carried out by each instrument (i.e., the *drivers*). When the computer does not obtain ongoing, continuous, status information from an instrument, the resultant arrangement is referred to as *open-loop* control of the particular instrument. In such cases, the program must provide for appropriate time delays (such as the time needed to position the robotic arm). Alternatively, the computer can receive signals from the various components that advise the program of their status; this is referred to as *closed-loop* control and is a generally more desirable mode of operation than the open-loop configuration.

COMPUTER BASICS

Personal computers are organized to carry out tedious, repetitive tasks in a rapid and error-free manner. The computer has four principal functional elements:

- Central processing unit (CPU) for arithmetic and logical operations, and instruction control.
- Memory for storage of data, results, and instructions (programs)
- Input/output components (I/O) for interaction between the computer and the external environment.
- Communication bus: or simply the bus, that allows the functional elements to communicate.

These elements are shown in Fig. 9 that comprises the functional architecture of the PC. Detailed descriptions, and operation of the PC and its components (e.g., secondary storage system—*hard drive*) are readily available (15). The architectures of computer-based instrument systems for laboratory environments generally fall into one of four categories; these are noted in Fig. 10.

A single-purpose (fully dedicated) instrument is shown in Fig. 10a. This arrangement is convenient because it is

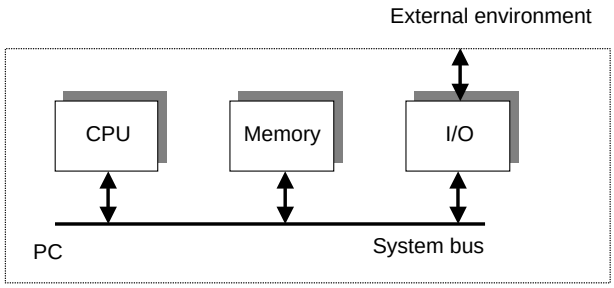


Figure 9. Basic architecture of the PC.

consistent with such things as existing building wiring, particularly the telephone system although emerging developments also lend this architecture to a wireless arrangement. Standard communication protocols (previously noted) allow manufacturers to develop instruments to accepted standards. This arrangement may be limited to a single PC and a single instrument, and the distance between the host (PC) and the instrument may also be constrained. By adding additional communication lines, other instruments can be added to the single PC.

Remote control of instruments is depicted in Fig. 10b and is accomplished by adding devices within the PC that support communication over a traditional (standard) telephone line (including use of the internet). Real-time operation in such circumstances may be limited because time is required to complete the communications between the PC and the remote instrument placing significant limits on the ability of the system to obtain complete results in a prescribed time interval. Delays produced by the PCs operating system (OS) must also be factored into information processing tasks.

With the development of, and need for, instruments with greater capacity, new architectures emerged. One configuration is shown in Fig. 10c and includes a single PC together with multiple instruments coupled via a standard (i.e., IEEE 488) or proprietary (communication) bus. While such architectures are flexible and new instruments can be readily added, the speed of operation can deteriorate to the point where the capabilities of the PC are exceeded. Speed is reduced because of competition for (access to) PC resources (e.g., hard disk space).

The arrangement shown in Fig. 10d is referred to as "*tightly coupled*". In such cases, the instruments are integral to the PC itself. Communication between the PC and the instrument is rapid. Real-time (on-line) operation of the instrument is facilitated by a direct communication path (i.e., system bus) between the instrument and other critical parts of the PC such as its memory. No (external) PC controller is necessary and consequently the time delays associated with such functional elements do not exist. By varying the functional combinations, the system can be reconfigured for a new application. For example, functional components might include: data acquisition resources, specialized display facilities, and multiport memory for communication (message-passing) between the other elements.

Each of the arrangements in Fig. 10 includes a single PC. Additional operating speeds are possible (at relatively low cost) if more than one PC is included in the instrumental configuration. Such architectures are called multiprocessor-based instrument systems. Each processor carries out program instructions in its own right (16–18).

With the increasing complexity and capabilities of new software, a more efficient arrangement for computer-based laboratory systems has emerged. This is the *client–server* concept as shown in Fig. 11. The server provides services needed by several users (database storage, computation, administration, printing, etc.) while the client computer (the users) manage the individual laboratory applications (e.g., DAQ), or local needs (graphical interfaces, error checking, data formatting, queries, submissions, etc.).

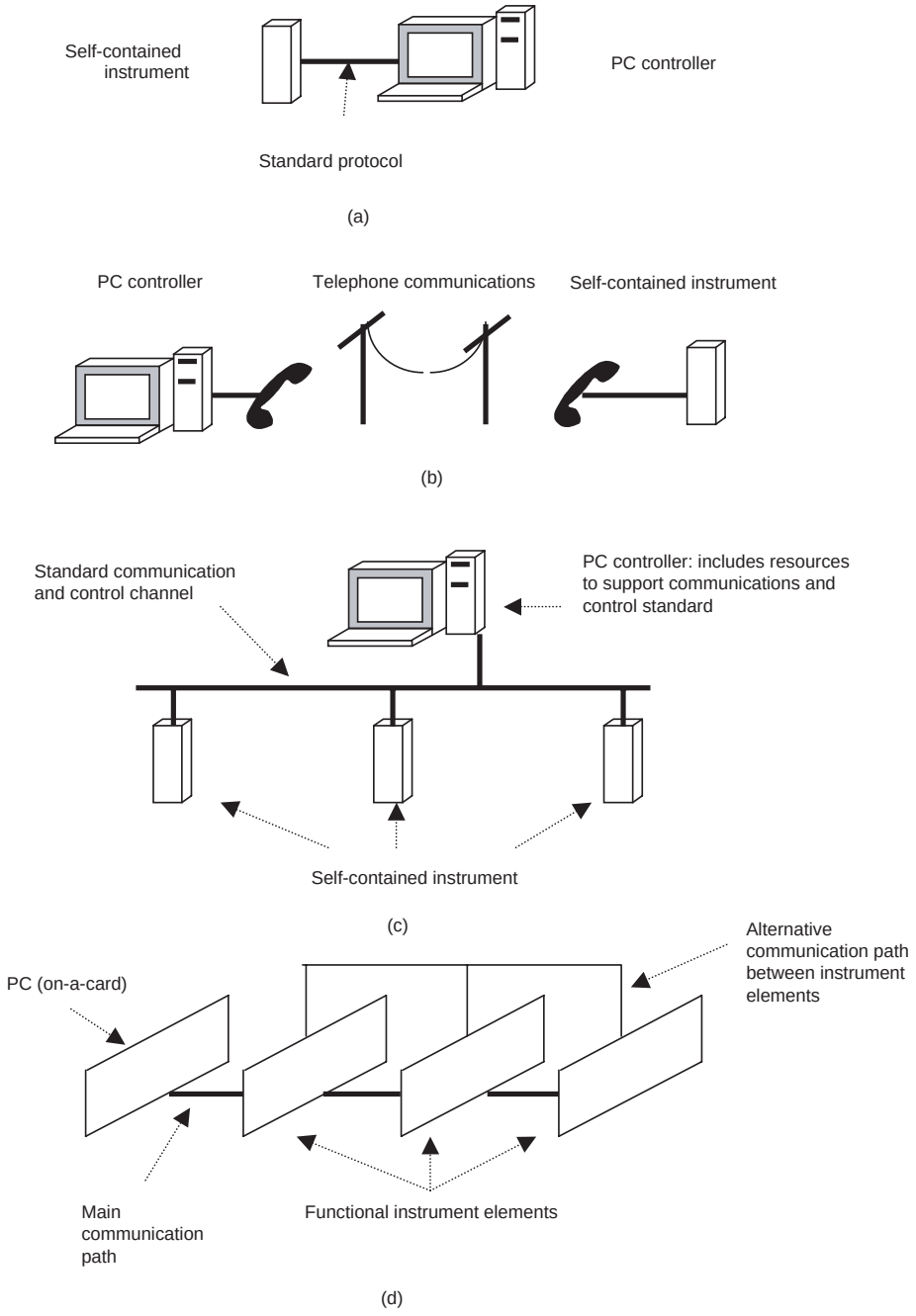


Figure 10. Single-PC instrument architecture (a) Dedicated system. (b) One form of remotely controlled arrangement. (c) Multiple instrument arrangement. (d) Tightly coupled architecture.

SOFTWARE IN COMPUTER-BASED INSTRUMENT SYSTEMS

Programming consists of a detailed and explicit set of directions for accomplishing some purpose, the set being expressed in some "language" suitable for input to a computer. Within the computer, the components respond to two signals: +5 V, or 0 V (ground). These potentials are interpreted as the equivalent of two logical conditions; normally the +5 V is viewed to mean logically true (or logical 1), and a 0 V is interpreted as logically false (or logical 0). (Note that this is not universally true, and in

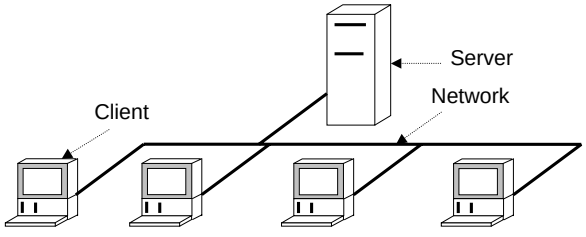


Figure 11. The client–server architecture for biomedical laboratory environments.

some situations the logical 0 signals "true" while the logical 1 signals "false", but this is normally an exceptional case.) During the early 1950s, laboratory computers were programmed by entering a series of logical 1s and 0s directly into the computer using a series of switches on the computer's front panel. A breakthrough occurred when English-like phrases could be used in place of these binary numbers. A series of programs within the machine called an *assembler* could be employed to interpret the instructions underlying the binary numbers. Within assembly language programs, English-like mnemonics are used in place of the numbers previously used to designate an instruction. The following example represents a series of assembly language instructions that might be used to add to quantities and store the results in one of the memory locations of the computer. (Text after the semicolon is considered to be a comment and not an instruction.)

```
MOV ACC, A      ;move the augend into the arithmetic
                  unit
ADD ACC, B      ;add the addend to the sum
MOV C, ACC      ;store the result in location "C"
```

During the 1950s, greater abstraction was introduced when text-based programming languages such as FORTRAN and COBOL were developed. Such languages are referred to as high level languages (HLLs). When individual versions are taken into account, there are literally hundreds of HLLs currently viable with languages such as C, C++, and JAVA being prominent. By using HLLS, the three lines of code shown above could effectively be replaced by a single instruction:

$$C = A + B$$

Statements such as these made problem solving and programming more abstract, readable, and reduced the time it took to develop software applications. The statements are entered into the computer using a program called an *editor*; the code is then *compiled* and *assembled* (translated using a *compiler* program and an *assembler* program) to reduce the original text to the binary numbers needed to control the computer: the only "instructions" that a computer really "understands".

Rather than having to "rewrite" a program each time it was required, HLLs provided a means to develop highly abstract "application programs". A key development of such powerful resources was the introduction of Visicalc, the first (primitive) spreadsheet program: It is progenitor of such widely used programs as Excel, LOTUS, and others. Development of automated spreadsheet programs was motivated by the need for them in business applications, but they have come to find considerable utility in biomedical laboratory environments, particularly for data and statistical analysis as well as for data acquisition.

Increasing levels of abstraction in which programming details are hidden have continued to drive developments in software. A most important transition was made when software entered the age of "visual" programming. Arrangements and interconnections of functional icons have come to replace text when developing software for the biomedical laboratory. A key example of this architecture is the "graphic programming language" (GPL) called LabVIEW, which stands for Laboratory Virtual Instrument Electronic Workbench. This programming scheme provides work areas (windows) that the programmer uses to develop the software. In particular, the windows include a "Front Panel" and a "Diagram". This software enables a user to convert the computer into a software instrument that carries out real tasks (when coupled to appropriate elements as shown in Fig. 6). The Panel displays the indicators and controls that a user would find if the investigator had obtained a separate instrument for the experimental setup. Figures 12 and 13 are representative of a LabVIEW (front) panel and diagram. They are suitably annotated to indicate, controls, indicators, and symbols to replace traditional programming constructs (19).

The HLL programming is characterized by a *control flow* model in which the program elements execute one at a time in an order that is coded explicitly within the program. The narrative-like statements of the program describe the sequential execution of "Procedure A" followed by "Procedure B", and so on. In contrast, a visual programming paradigm functions as a *dataflow* computer language. This depends on *data dependency*; that is, the object in the block (node) will begin execution at the moment when all of its inputs are available. (This reflects a "parallel" execution scheme and is consistent with a multitasking model.) After completing its internal operations, the block will present processed results at its output terminals. While one node waits for events, other processes can execute. This is in

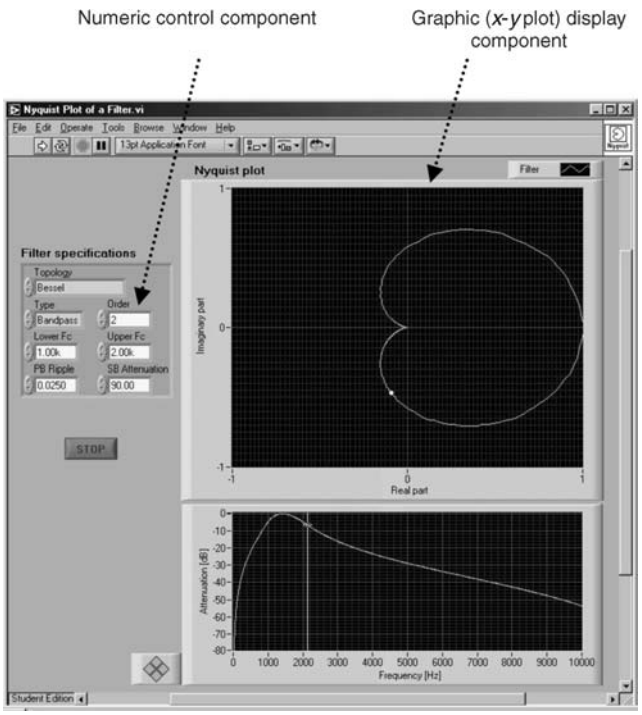


Figure 12. Front panel of a virtual instrument that obtains the attenuation and Nyquist characteristics (plot) of a filter that could be used in a Biomedical laboratory DAQ system. Control and display components are identified.

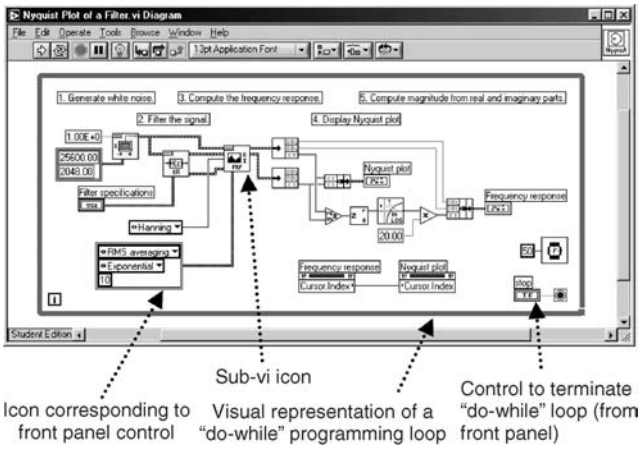


Figure 13. Diagram (Data Flow) of the virtual instrument for determining the frequency response and Nyquist plot of a filter.

some contrast to the control flow case where any waiting periods can create "dead time," and may reduce the system throughput (20).

The virtual instrument (vi) depicted in Figs. 12 and 13 computes the frequency response of a digital filter and displays the attenuation versus frequency (the independent variable), as well as the imaginary part of the response versus the real part (i.e., *Nyquist* plot). In this example, a white noise signal is used as the stimulus of the filter and the vi returns the frequency response of the filter (21).

EMERGING AND FUTURE DEVELOPMENTS FOR COMPUTER-BASED SYSTEMS IN BIOMEDICAL LABORATORIES

Data processing in the biomedical laboratory is coming to rely increasingly on artificial intelligence (AI) for analysis, pattern recognition, and scientific conclusions. The development of "artificially intelligent" systems has been one of the most ambitious and controversial uses of computers in the biomedical laboratory. Historically, developments in this area (biomedical laboratory) were largely based in the United States (22–24). Artificial intelligent can support both the creation and use of scientific knowledge within the biomedical laboratory. Human cognition is underscored by a complex and interrelated set of phenomena. From one perspective, AI can be implemented with computer systems whose performance is, at some level, indistinguishable from those of human beings. At the extreme of this approach, AI would reside in "computer minds" such as robots or virtual worlds like the information space found in the Internet. Alternatively, AI can be viewed as a way to support scientists to make decisions in complex or difficult situations. For example, anesthesiology requires the health provider to monitor and control a great many parameters at the same time. In such circumstances, dangerous trends may be difficult for the anesthesiologist to detect in "real time;" AI can provide "intelligent control." In science, AI systems have the capacity to learn, leading to the discovery of new phenomena and the creation of scientific knowledge. Modern computers and their associated application software tools can be used to analyze large amounts of data, searching for complex patterns, and suggesting previously unexpected relationships (see Coiera in the Reading List). Simply stated, the goal of AI is to develop automata (machines) that function in the same way that a human would function in a given environment with a known complement of stimulants. In 1938, the British mathematician Alan Turing showed that a simple computational model (the Turing Machine) was capable of universal computation. This was one basis for the stored program model used extensively in modern computers.

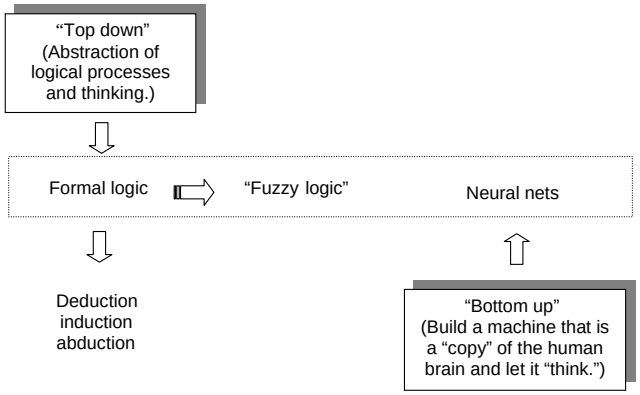


Figure 14. Classification of artificial intelligence: Formal logic, Fuzzy Logic, Neural Nets.

An attempt to build an automaton that imitates human behavior falls into three broad categories: formal logic, "fuzzy" logic, and neural net technologies. These are depicted in Fig. 14.

While there is considerable overlap between human cognitive activities and machine technologies, in its most generic sense the relationships can be summarized in Table 4.

Historically, the first attempts at machine intelligence reflected formal logical thinking of which there are three kinds: deductive, inductive, and abductive. These are all built on a system of rules, some of which may be probabilistic in nature. Deductive reasoning is considered to be perfect logic : you cannot prove a false predicate to be true, or a true predicate to be true. The logic is built on the following sequence of predicates:

If p then q
p is true
Therefore q is true

By using the classification of beats in the ECG signal based on QRS duration and RR interval, we can develop a simple

Table 4. Human Cognitive Activities and Corresponding Machine Technologies

Human Activity	Machine Technology
Pattern recognition	Neural Nets
Belief System and control	Fuzzy logic
Application of logic	Expert Systems: rules and generic algorithms

example of deductive reasoning:

All{beats (RR interval) falling between 1.0 and 1.5s having a QRS interval between 50 and 80 ms}are normal

Patient's {60th QRS complex occurs 1.25 s after the 59th complex with a duration of 60 ms}Patient's 60th QRS complex is normal

Inductive conclusions, which can be imperfect and produce errors, follow from a series of observations. This logic is summarized by the following series of predicate statements:

From : (P a), (P b), (P c), . . .
Infer : [forall(x)(P x)]

(P a), (P b), (P c), and so on, all signify that entities whose properties are a, b, c, and so on, belong to the category identified as P. We therefore conclude that any object whose properties are similar to those of a, b, c, and so on, belong to the category identified as P. For example, a physician may observe many patients who have had fevers and some of who have subsequently died. Postmortem examination may reveal that they all had a lung infection (labeled "pneumonia"). The physician may (erroneously) conclude that "all fevers must imply pneumonia", because he/she has a number of observations in which fever was associated with pneumonia.

Using cause–effect statements, abductive reasoning gather all possible observations (effects) and reaches conclusions regarding causes. For example, both pneumonia and septicaemia may both cause fever. The physician would then use additional observations (effects) to single out the "correct" cause. Abductive logic may also lead to false conclusions. Abductive logic follows from the argument that follows:

If p then q
q is true
Therefore p follows(i.e., is true)

Using the circumstances just cited, a physician may (erroneously) conclude that having observed a fever, the patient is suffering from septicaemia. (It may, of course, also be due to pneumonia.)

Machine-Based Expert Systems

These systems require machine-based reasoning methods just noted and are depicted in Fig. 15. In addition, they must include stylized or abstracted versions of the world. Each of the representations in the database must be able to act as a substitute or surrogate for the underlying object (or idea). In addition, these tokens may have metaphysical features that reflect how the system intends to "think about the world". For example, in one type of representation known as a script, a number of predicates may appear that describe what is to be expected for a particular medical test. An Expert System for "detecting" an asystole in an ECG (electrocardiogram) might invoke the following rules (see Coiera in Further Reading):

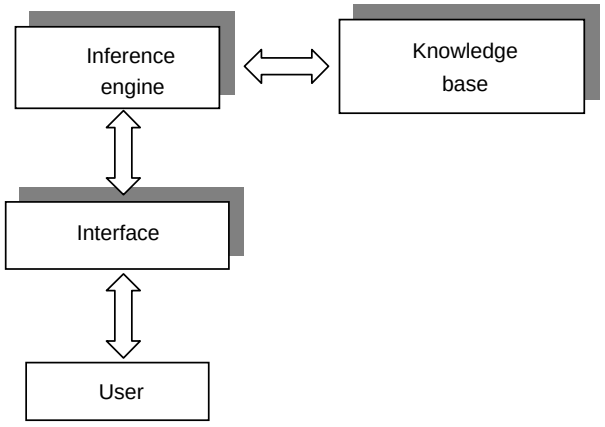


Figure 15. Block diagram of a typical Expert System.

Rule 1:

If heart rate = 0
Then conclude asystole

Rule 2:

If asystole and(blood pressure is pulsatile and in the normal range)
Then conclude retract asystole

Where knowledge is less certain, the rules might be modified. For example, for Rule 1, the conclusion might become, ''conclude asystole with probability (0.8)''.

A somewhat more informative example can be drawn from an interactive fragment from a contemporary medical Expert System (with similarities to the historical MYCIN software) (25):

The fragment does not represent a complete interactive session. The user would need to supply additional information to generate a potential diagnosis. An excellent demonstration of such systems can be found on the Internet:

http://dxplain.mgh.harvard.edu/dxp/dxp.sdemo.pl?/login=dems/cshome

While Expert System technologies have produced useful applications, they are confronted with a fundamental problem: How to determine what is "true" and what is "false". Contemporary systems address this in a variety of ways (e.g., providing a probabilistic result). This remains a problem for application software.

Fuzzy Logic Systems

These systems attempt to overcome the vagaries of truth and falsity and thus better reflect human thinking and may have some advantage over Expert Systems, where predicates are either true or false (or have some fixed probability of truth or falsity). Such systems were pioneered by Loti Zadeh in 1966 although exploitation began in earnest during the 1990s. [These are currently well over 2000 patents (many from Japan where this

[An asterisk (*) indicates physician responses. What follows ";" are explanatory comments]

```
Please enter findings
* sex male
* race white
* alcoholism chronic
* go
Disregarding:
     Exposure to rabbits
     Leg weakness
     Creatinine blood increased
Considering
     Age 26–55
Ruleout:
     Hepatitis chronic
     Alcoholic hepatitis
Abdomen pain generalized?
* no
Abdomen pain right quadrant?
```

```
;The program asks for facts about the
;patient
;There is a fixed vocabulary of symptoms
;that must be followed
;This starts processing in the Expert System
;The system finds a set of suspected diseases
;Symptoms not explained by these diseases
;are put aside.

;The system explains its reasoning

;and rules out certain disease


;It requests additional information to
;further refine its findings
;Fragment ends here.
```

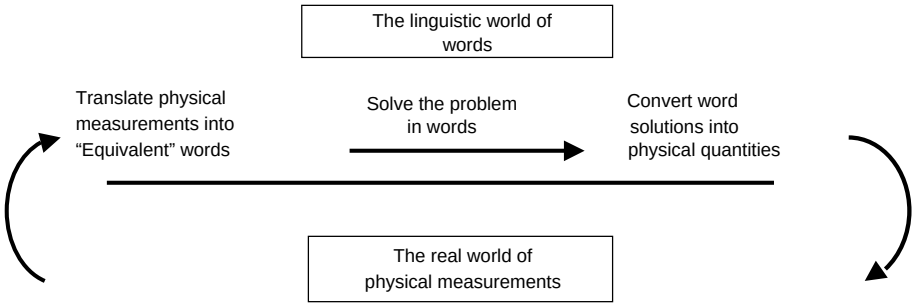


Figure 16. Paradigm for fuzzy logic control.

technology was first embraced) and billions of dollars of sales of fuzzy products.] The concept underlying fuzzy logic is shown in Fig. 16 (26). Measurements in the real world are translated into equivalent linguistic concepts; the resulting "word" problems are solved in the linguistic world and conclusions are reconverted into physical entities that control elements in the real world.

Translation from physical measurements is accomplished by using a "belief system" (so called "*membership functions*") that reflects the degree to which we accept the particular measurement. A given measurement will then determine the extent to which we interpret its meaning. A representative set of membership functions is shown in Fig. 17, shown together with outcomes for a particular physical measurement (input).

A similar set of functions is also constructed to represent outputs, or rules, as summarized below. The output functions determine the extent to which we should set a particular control parameter in the physical system. In Fig. 17, we interpret the measurement to mean that we have a 50% belief that it has a low value; we also believe that this measurement could represent a "medium" quantity, but we only have a 25% confidence in this value. (One might say, "The measurement is somewhere between a low and a medium value. Membership functions provides a measure of such informality.)

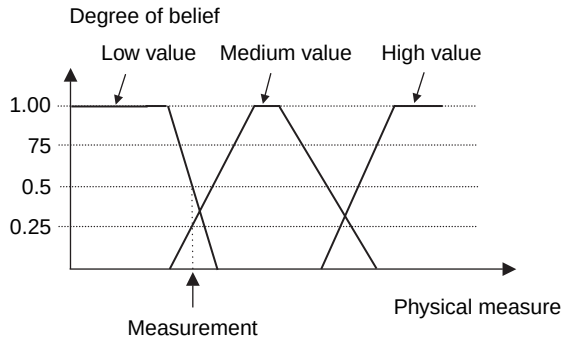


Figure 17. Fuzzy logic membership functions.

Table 5. Operations on Fuzzy Relations

Operation	Fuzzy Application[a]
Parameter 1 AND Parameter 2	Min (m1, m2)
Parameter 1 OR Parameter 2	Max (m1, m2)
NOT Parameter 1	1–m1

[a]The m stands for the membership or belief value as described in the narrative.

A system of rules also forms part of fuzzy control facilities. These rules have the following form:

If (physical parameter 1 is low) AND (physical parameter 2 is high) THEN (apply Rule 1 with a low intensity).

If (physical parameter 1 is medium) OR (physical parameter 2 is high) THEN (apply Rule 1 with a medium intensity)

The measurements and the logical operators (e.g., AND, OR, NOT) are employed according to the following set of (fuzzy) rules (as developed by Zadeh) (see Table 5).

For a given control problem, the measurement, application of the rules, and the invocation of actions leads to an overall profile of action; a typical result is shown in Fig. 18. From this a control value can be returned to the physical system: generating what is referred to as a *crisp result* (*Defuzzification*). There are several methods for obtaining this value from the curve noted in Fig. 18. Shown is the Center-of-Gravity method wherein the control value to be applied to the variable under control is the "balance point" of the curve.

Fuzzy control in a biomedical environment is exemplified by control of oxygen delivery to ventilated newborns in a neonatal intensive care environment (27). A sketch of the system is shown in Fig. 19.

For newborns requiring mechanical ventilation, oxygen toxicity is a potential danger that could result in chronic lung disease. Oxygen levels are also implicated in the development of retinopathy. Inspired oxygen concentration is commonly adjusted on an acute basis to control oxygen delivery and maintain patient saturation levels. The design of classic (engineering) control systems in such cases presents a significant challenge because of "transportation" delays. The alternative of manual control is also

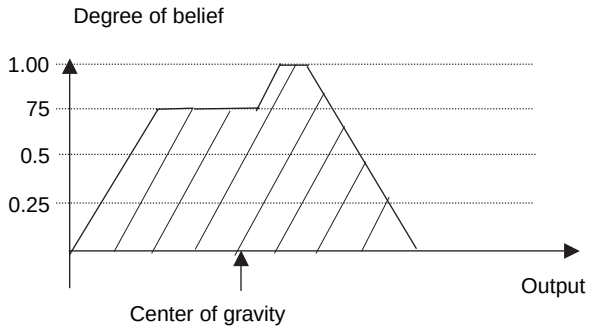


Figure 18. Defuzzification or generation of crisp results for a fuzzy logic control system.

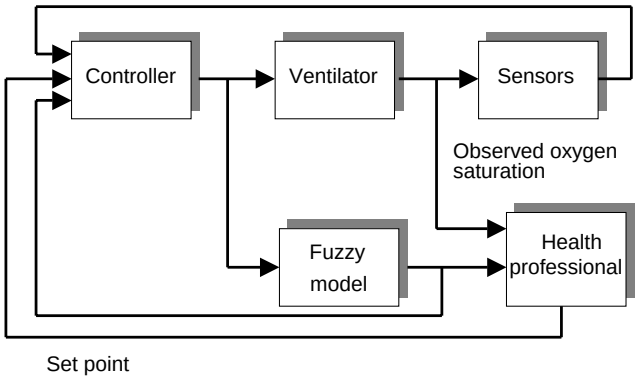


Figure 19. Fuzzy control of inspired O_2 concentration in ventilated infants.

unsatisfactory for two reasons: the patient might require increased O_2 requirement but the manual increase is delayed due to delay in human response (e.g., the clinician is not present); the patient has decreased O_2 requirement (clinical conditions improve) but the amount of O_2 is not immediately decreased (because there is a perception that the patient is "doing well" and does not require intervention).

The Fuzzy model included some 35 rules, of which the following is one example:

If {change in oxygen saturation is small-negative} AND {rate of change in oxygen saturation is medium-negative}
THEN
{increase inspired oxygen concentration by a medium-positive amount}

The membership curves for the various parameters will impose specific amounts for each rule that is invoked. Each rule yields an "action" value according to a membership class or extent to which the rule should be applied. A weighted mean of all rule outputs produces a single value for inspired oxygen concentration. This system maintained a target oxygen saturation (the set point determined by the health professional) better than routine manual control. It reduced overall oxygen exposure. No complex mathematical models were required as might be the case for traditional control with predictive capabilities. The rules are easy to understand and modify; expert knowledge about the problem was utilized. The controller was easily designed for nonlinear system responses: a goal that is daunting for traditional control technologies.

Neural Net (NN)

These systems employ a combination of circuits that approximate the behavior of neurological cells. While not limited to such applications, NNs are particularly useful for pattern recognition (28,29). Figure 20 shows the model of a single neuronal element and as well as a network of neuronal elements (NN).

The design of NNs is definitely not a precise enterprise; it is decidedly an art. A NN is "trained" to recognize patterns and while there are a great variety of NN

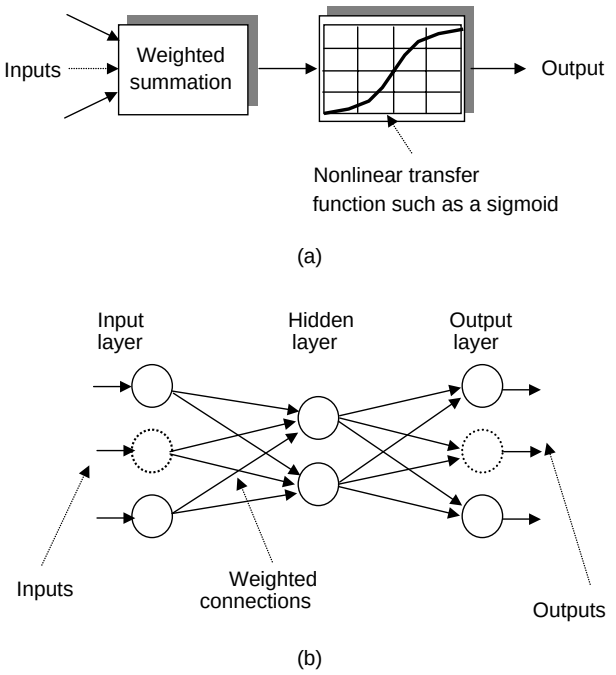


Figure 20. Neural Net architecture. (a) A single neuron. (b) A NN showing input, hidden, and output layers with weighted interconnections.

architectures and training paradigms, a general scheme works as follows:

- A NN is presented (at the input layer) with a set of "test" signals (e.g., samples of the input as a function of time). This is the so-called *training set*. If the NN produces an incorrect output (i.e., recognizing the wrong pattern), the weighted interconnections are automatically readjusted using an algorithm designated as *back propagation* (of error correction). The next time that the particular input sample is presented, the NN will *tend* to produce the correct answer.
- Training continues until the NN satisfactorily recognizes the members of the training set. This recognition does not have to be perfect: just as human experts might disagree on the interpretation of a pattern.
- The NN is then put into operation with inputs that it has not necessarily seen before; it will recognize whether or not the inputs have the same characteristics of the training set that it has been taught to recognize.
- This process may be iterative; if the operational results are unsatisfactory, additional samples may be added to the training set and retraining instituted.

The NNs have been used to advantage in an ever-growing set of circumstances (30–32). In one instance, the NN (*Hypernet*) used a multi-layer architecture in which subjects' anamnestic data, and 24-h diastolic and systolic blood pressure measurements were used as the parameters. System outputs included four 24-item arrays, whose values specify the hourly dosage to be administered to a patient for each of the most common antihypertensive drugs. The presence and degree of hypertension (diagnosis) could be inferred from the drug dosage. No treatment was required for normal subjects whose records formed part of the training set. The system was evaluated on the basis of accuracy in both diagnosis and prescriptions. Hypernet correctly diagnosed 33 subjects out of 35, and 82% of the system's prescribed treatment was deemed correct or acceptable by a group of medical experts.

SUMMARY

Biomedical laboratory instrument technology has changed markedly over the last 50 years. During the 1970s, a few daring scientific investigators began using the computer for instrument control and data acquisition. Such features are now commonplace add-ons to most instrumentation. The modern desktop computer system is a window into computing resources available anywhere in the world. Increasing levels of abstraction within the computer and its application software signal the emergence of a data processing model for experimental design within the Biomedical laboratory. This will only accelerate in the future as new computational algorithms are developed. Artificial intelligence as a knowledge-based tool is certain to grow manifestly in this emerging culture.

BIBLIOGRAPHY

1. Nebeker F. Golden accomplishments in Biomedical Engineering. Charting the milestones of Biomedical. Engineering. Engineering in Medicine and Biology Society; 2003.
2. Schoenfeld RL. From Einthoven's Galvanometer to Single-Channel Recording. Charting the milestones of Biomedical Engineering, Engineering. Medicine and Biology Society; 2003.
3. Gasser HS, Erlanger J. A study of the active currents of nerve cells with the cathode ray oscilloscope. Am J Physiol 1922; 62: 406–524.
4. Olansen JB, Ghorbel F, Clark JW, Bidani A. Using virtual instrumentation to develop a modern biomedical engineering laboratory. Int J Eng Educ 2000; 16(3):244–254.
5. Schoenfeld RL. The role of a digital computer as a biological instrument. Ann New York Acad Sci 1964; 125(2):915–942.
6. Schoenfeld RL. How we know what the eye and the mind's eye sees. IEEE Eng Med Biol Mag 1992; 11:47–71.
7. Silverman G, Eisenberg L. Programmable parallel timing system. IEEE Trans Biomed Eng 1971; 18(3):201–205.
8. Stallings W. Data and computer communications. 5th ed. New York: Prentice Hall; 1997.
9. Rappaport TS. Wireless communications; principles and practice. New York: Prentice Hall; 1999.
10. Carr JJ. Sensors and circuits. New York: Prentice Hall; 1993.
11. Norton HN. Handbook of transducers. New York: Prentice Hall; 1989.
12. Tyler M. The ABCs of ADCs. Scientific Computing & Instrumentation, April 2001; p 32–34.
13. Institute of Electrical and Electronics Engineers. IEEE Standard Digital Interface for Programmable Instrumentation, Standard IEEE 488.1, IEEE 488.2.
14. Liscouski J. Laboratory and scientific computing: A strategic approach. New York: John Wiley & Sons; 1995.

15. Clements A. Principles of computer hardware. PWS-Kent; 1991.
16. Silverman G. Automation in the biomedical laboratory. IEEE Trans Biomed Eng 1884; BME-31:748–752.
17. Stromquist BR, Pavlides C, Zelano JA. On-line acquisition, analysis and presentation of neurophysiological data based on a personal microcomputer system. J Neurosci Methods 1990;35:215–222.
18. Beavis BC, Chait BT. Rapid, sensitive analysis of protein mixtures by mass spectrometry. Proc Natl Acad Sci USA 1990;87:6873–6877.
19. Wald M. Int J Eng Educ 2000;16:3.
20. Essick J. Advanced LabVIEW Labs. New York: Prentice Hall; 1999.
21. Ziemer RE, Traner WH, Fannin DR. Signals and Systems. 4th ed. New York: Prentice Hall; 1998.
22. Szolovits P. Artificial intelligence in medicine. AAAS Selected Symposia Series. Colorado: Westview Press; 1982.
23. Clancy WJ, Shortliffe EH editors. Readings in Medical Artificial Intelligence—The First Decade. Reading, MA: Addison-Wesley; 1984.
24. Miller PL. Selected Topics in Medical Artificial Intelligence. New York: Springer Verlag; 1988.
25. Buchanan BG, Shortliffe EH, editors. Rule-based expert systems: The MYCIN experiments of the Stanford heuristic programming project. Reading, Massachusetts: Addison-Wesley Publishing Company; 1984.
26. Ross TJ. Fuzzy logic with engineering applications. New York: McGraw-Hill; 1995.
27. Sun Y, Kohane I, Stark A. Fuzzy logic control of inspired oxygen concentration in ventilated newborn infants. www.chip.org/chip/projects/avent/sun-flpaper.html.
28. Fausett L. Fundamentals of neural networks. New York: Prentice-Hall; 1994.
29. Kohonen T. An introduction to neural computing. Neural Networks 1988;1:3–16.
30. Poli R, Cagnoni S, Livi R, Coppini G, Valli G. A nn expert system for diagnosing and treating hypertension. Computer March 1991.
31. Silverman G, Brudny J, Gage P. Artificial intelligence in rehabilitation medicine: an emerging technology. Proc 6th Annu Meet Am Telemedicine Assoc, Ft. Lauderdale, FL; June 3–6, 2001.
32. Hirsch S, Frank PI, Shapiro JL. Use of an artificial neural network in estimating prevalence and assessing underdiagnosis of asthma. Neural Computing and Appl 1997;5(2):134–128.

Further Reading

Liscouski J. Laboratory and scientific computing: A strategic approach. New York: John Wiley & Sons; 1995.
Webster JG, editor. Bioinstrumentation. New York: John Wiley & Sons; 2004.
Carr JC, Brown JM. Introduction to biomedical equipment technology. 4th ed. New York: Prentice Hall; 2001.
Olansen JB, Rosow E. Virtual bio-instrumentation. New York: Prentice Hall PTR; 2002.
Keener J, Sneyd J. Mathematical physiology. New York: Springer; 1998.
Paton BE. Sensors, transducers & LabVIEW. New York: Prentice Hall PTR; 1999.
Silverman G, Silver H. Modern Instrumentation: A Computer Approach. IOP Publishing; 1995.
Coiera E. Guide to Medical Informatics, the Internet and Telemedicine. Oxford University Press; 1997.

See also ANALYTICAL METHODS, AUTOMATED; COMPUTER-ASSISTED DETECTION AND DIAGNOSIS; CYTOLOGY, AUTOMATED; DIFFERENTIAL COUNTS, AUTOMATED.

COMPUTERS IN MEDICAL EDUCATION.

See MEDICAL EDUCATION, COMPUTERS IN.

COMPUTERS IN MEDICAL RECORDS.

See MEDICAL RECORDS, COMPUTERS IN.

COMPUTERS IN NUCLEAR MEDICINE.

See NUCLEAR MEDICINE, COMPUTERS IN.

CONFOCAL MICROSCOPY.

See MICROSCOPY, CONFOCAL.

CONFORMAL RADIOTHERAPY.

See RADIOTHERAPY, THREE-DIMENSIONAL CONFORMAL.

CONTACT LENSES

ERIC R. RITCHEY
The Ohio State University
Columbus, Ohio

INTRODUCTION

Contact lenses are prescription medical devices applied to the anterior surface of the cornea for the temporary correction of refractive error. Contact lenses can be used successfully to correct a number of refractive error conditions, such as myopia, hyperopia, presbyopia, and aphakia. Currently, there are an estimated 38 million contact lens wearers in the United States and 125 million contact lens wearers worldwide, making contact lenses one of the most commonly prescribed medical devices available (1). The ability to wear contact lenses successfully is dependent on numerous factors, including but not limited to the contact lens design and materials, corneal health and physiology, the lens to epithelium interface, proper use of lens care solutions, patient compliance, professional fitting and follow-up care. Unacceptable contact lens fits can lead to deleterious results ranging from poor lens comfort to microbial keratitis and permanent loss of visual acuity.

HISTORY

The earliest origins of the theoretical application of a device to the anterior cornea for the correction of vision can be traced to Leonardo da Vinci in the early sixteenth century (2). Leonardo da Vinci described an experiment where the subject immerses his face in a transparent globe filled with water that effectively neutralizes the subject's refractive error. In 1636, Rene Descartes described the neutralization of refractive error through the use of a long water filled tube, called a hydrodiascope that was held against the anterior surface of the cornea (3). The ideas proposed by da Vinci and Descartes, however, could not be used in practical application for the correction of refractive error on a daily basis. The first true contact lens that could be worn on the eye was the scleral contact lens. The scleral contact lens is a large diameter lens (>13 mm) that features a scleral (also known as a haptic) segment of

the lens that rests on the conjunctiva and a central corneal section that arched over the cornea. The first scleral contact lens was created Frederick Muller in 1887. Although the lens was a nonoptical lens, it showed that the placement of a lens on the eye was an achievable goal. In 1888, Adolph Fick of Germany and Euegene Kalt of France each developed optically corrective glass scleral contact lenses independently of one another (4). In 1936, William Feinbloom created a hybrid scleral lens with a poly(methylmethacrylate) (PMMA) scleral section and a glass corneal section (5). The change to PMMA as the lens material of choice was due to its superior durability compared to glass. In 1947, Kevin Touhy created the first PMMA corneal contact lens. The corneal contact lens featured an overall diameter smaller than the corneal diameter and lacked a scleral–haptic section that rests on the cornea. The Touhy design was the precursor to the modern rigid gas permeable contact lens used today (4,5).

In 1954, Czechoslovakias Otto Wichterle and Drahoslav Lim developed the first soft contact lens polymer called hydroxyethyl methacrylate (HEMA). The development of HEMA became the chemical backbone for soft hydrogel contact lenses and made the development of a soft contact lens possible (5,6). In 1971, Bausch and Lomb received the U.S. Food and Drug Administration (FDA) approval for the SofLens soft contact lens, the first soft contact lens released to the public for daily wear. The FDA granted the first approval for soft contact lenses overnight wear, also known as extended wear, in 1981 (6). Improvements in lens manufacturing technologies in the late 1980s allowed soft contact lens manufacturers to produce lenses, at a cost where frequent replacement, or disposable lenses were introduced to the marketplace. These disposable lens were developed and promoted for a number of replacement schedules including every 3 months, every month, every 2 weeks, and every day throughout the late 1980s and the 1990s (7).

In 1998, the first silicone hydrogel lens was released for public distribution. The development of a soft contact lens that successfully incorporates silicone into the lens matrix started with work with silicone elastomers in the 1950s (4). Silicone elastomers have the advantage of high oxygen permeability compared to HEMA-based contact lenses. The early silicone elastomer lenses had problems with wettability and comfort. Silicone hydrogel lenses provide the oxygen benefits of a silicone elastomer with the comfort of hydrogel materials. Currently, the FDA has approved some silicone hydrogel design contact lenses for continuous extended wear ranging from 1 week to 1 month dependent on the lens design (8,9).

CONTACT LENS OPTICS AND DESIGN

Contact lenses are designed to correct the refractive error of the patient by converging or diverging light entering the visual system. In myopia, or nearsightedness, the refractive power of the eye converges light excessively, causing incoming light to be focuses in front of the retina creating blur. To correct for myopia, a lens that diverges light, indicated with a minus power, in combination with the optics of the eye will allow the incoming light to focus on the retina providing clear vision. In hyperopia, or farsightedness, the refractive power of the eye lacks sufficient convergence to focus incoming light on the retina, causing light to be focused behind the retina (10). To correct for hyperopia, a lens that causes light to converge, indicated with a plus power, in combination with the optics of the eye will allow the incoming light to focus on the retina providing clear vision. The contact lens must change the vergence of the incoming light to correct the patient's refractive error while maintaining positional stability, corneal health, and patient comfort.

POLY(METHYL METHACRYLATE)–RIGID GAS PERMEABLE CONTACT LENS DESIGN

Rigid contact lenses were initially manufactured from PMMA, a nonpermeable plastic material used for the first "hard" contact lens (4). Today rigid gas permeable (RGP) lenses, also referred to simply as gas permeable (GP) lenses, are manufactured from a number of gas permeable silicone acrylate or fluorosilicone acrylate materials. Silicone acrylate and fluorosilicone acrylate materials, although not as durable as PMMA lenses, have superior oxygen permeability and reduce corneal hypoxia associated with long term PMMA use (11,12). The manufacturing process for PMMA and rigid gas permeable lenses is essentially identical. The PMMA and gas permeable lenses are cut from a button of the lens material that is lathed with one of two basic curve designs. The traditional contact lens design is a tricurve or multicurve design with each peripheral curve flatter than the proceeding curve. The central curve is designated the base curve (BC), which is typically designed to match the central flat corneal curvature measured for the patient. The next two curvatures, referred to as the secondary curve and the peripheral curve, are each flatter than the preceding curve. The transitions between the curvatures, as well as the edge of the lens, are rounded and polished to prevent the development of sharp edges that will decrease patient comfort and cause physiological damage to the cornea (13). Rigid lenses may also be lathed in an aspheric curve design, which has a central base curve and a continuously flattening peripheral curvature without distinct secondary or peripheral curves. The purpose behind flattening the curvature of the lens as you move from the center of the lens for each design is to allow the gas permeable lens to match the progressively flattening aspheric corneal surface, promoting patient comfort and an improved physiological response to contact lens wear (12).

The rigid contact lens diameter can be subdivided into different zones that correspond to the curvature system described above. The central zone of the contact lens is referred to as the optic zone and the width of this zone is referred to as the optic zone diameter (OZD). The optic zone is the portion of the contact lens that contains the refractive power of the lens and is used for the correction of the patient's refractive error. The curvature of the lens at the optic zone is the base curve of the lens. The optic zone

diameter constitutes the majority of the overall lens diameter (OAD). The width of the secondary and peripheral curves are designated the secondary curve width and the peripheral curve width (12,13). For aspheric designs, an optic zone diameter is typically not specified and measurement of the peripheral curve width is impractical due to the gradually changing curvature of this region. The typical gas permeable contact lens is 8–11 mm in diameter with an optic zone diameter between 6 and 9 mm (13).

SOFT CONTACT LENS DESIGN

Soft hydrogel contact lenses are produced by one of three techniques: lathe cut, spin casting, and cast molding. Lathe cut lenses are produced in a method similar to the production of rigid gas permeable lenses. A button of dehydrated plastic is cut to the desired shape using a computerized lathe. The lens is then hydrated in a saline solution bath, where it will expand into its final shape. Lathe cut soft contact lenses can be made in a wide variety of powers and curvatures. These lenses are typically replaced annually due to the cost of each lens produced. Because of the lathing and hydrations process, there can be some variability in the optics and curvatures observed between lenses (14,15).

Spin casting techniques utilize a spinning mold to distribute a liquid polymer in a thin consistent layer without the use of a lathe. The spinning cast mold features a concave surface that will hold the liquid polymer, and therefore determines the shape of the anterior surface of the lens. The posterior lens surface is determined by the centrifugal force generated by the spinning mold and the surface tension of the polymer. Once the liquid polymer has been distributed by the spinning mold, the lens is polymerized while the mold spins using ultraviolet (UV) light. Polymerization will solidify the liquid polymer and the lens is removed from the cast. Spin casting can be used to produce a number contact lens powers without the marks associated with lathing techniques (14,15).

A third method of soft contact lens production is the cast molding technique. Cast molding uses a two piece die or cast that when placed together will form the anterior and posterior surface of the lens. The lens polymer is injected between the two cast pieces and is then polymerized. The two cast pieces are then separated and the contact lens is finished and packaged for distribution (14,15). The majority of disposable contact lenses produced for the contact lens market in the United States are cast molded or spin casting designs.

FITTING PHILOSOPHY: RIGID GAS PERMEABLE CONTACT LENSES

Gas permeable rigid contact lenses are typically fit according to one of two fitting philosophies. One philosophy is the interpalpebral (IP) lens fit. Interpalpebral lens fit contact lenses feature small diameters lenses with the central base curve that parallels the flat central corneal curvature. This philosophy is designed to keep the contact lens centered on the corneal and promotes minimal interaction with the upper lid when the patient blinks (16). The second philosophy is the lid attachment contact lens fit. The lid attachment fit philosophy utilizes large diameter contact lenses with a base curve that is equal or slightly flatter than the flattest corneal meridian measured by central keratometry. The lens edge will lie underneath the superior lid and will move in tandem with the lid while blinking (16). The peripheral curve is designed to maximize the contact of the eyelid with the lens. Lid attachment philosophy proponents argue that this design promotes superior tear exchange behind the contact lens and provides superior patient comfort as it keeps the lid from chronically moving over the edge of the contact lens (17). The mid-peripheral curve design for each technique aims to closely follow the curvature of the peripheral cornea to provide adequate lens movement and comfort. Adjustments are made to the final power of the contact lens to compensate for the tear lens created behind the lens with each fitting philosophy (16).

SOFT CONTACT LENS FITTING

The fitting of soft contact lenses is driven by the material properties of the lens. Due to the water content of soft contact lenses, a lens applied to the cornea will drape over the epithelial surface and assume the shape of the cornea. The goal of the practitioner fitting the lens is to avoid any adverse effects from lens wear while providing the patient with a comfortable lens fit and good visual acuity. The three primary considerations for the practitioner when evaluating the interaction of the contact lens and the cornea are the coverage of the cornea by the contact lens, the centration of the contact lens over the cornea, and the movement of the contact lens with the blink response. A successful soft contact lens fit consists of a lens that covers the entire corneal surface, referred to as paralimbal coverage, is centered over the corneal apex, and moves ~ 0.25–1.00 mm with the blink (18). Soft contact lenses that consistently fail to cover and stay centered on the cornea or lack adequate lens movement may lead top potential adverse events with contact lens wear. The adverse events experienced from an inadequate soft contact lens fit can range in severity from poor lens comfort to microbial infection and a severe reduction in the patient's best corrected visual acuity. The success of a soft contact lens fit is also determined by the visual acuity obtained with the lens on eye and a patient history of lens comfort and adequate wear time throughout the day (7,18).

CONTACT LENS WEAR SCHEDULES

Contact lenses are prescribed by the practitioner with a specified wearing schedule. Wearing schedules can be divided into four categories: Daily wear, flexwear, overnight wear, and extended wear. Daily wear contact lenses are to be worn during waking hours and removed before sleep. The lenses are cleaned and stored overnight or discarded as determined by the recommended replacement schedule. Flexwear is a term used for contact lenses that are typically worn for daily wear but may be worn on a 24 h

basis 1–2 days per week. Overnight wear contact lenses are to be worn while sleeping and removed upon awakening. The most common example of overnight wear are reverse geometry gas permeable contact lenses used for orthokeratology. Extended wear, also referred to as continuous wear, is where a lens is worn on a 24 h basis for a predetermined period of time such as 1 week or 1 month. Currently, the maximum allowable extended wear time permitted by the FDA is 30 continuous day dependent on the lens design and material. Soft hydrogel, silicone hydrogel, and rigid gas permeable contact lens materials have been approved by the FDA may be worn for varying lengths of extended wear. Contact lenses designed for extended wear are typically made from high oxygen permeability materials to compensate for 24 h lens wear.

CONTACT LENS CARE

Cleaning and disinfection of contact lenses is one of the most critical elements for successful contact lens wear. Failure to properly clean and disinfect the contact lens can lead to visually threatening adverse events. Lens care begins with proper hygiene. Hand washing is required before the handling of contact lens (19). Once the patient has washed their hands, lens cleaning and disinfection may begin. A number of different commercially available lens cleaning systems have been distributed for consumer use (20). The cleaning and disinfection of lens has traditionally included some form of digital cleaning (a.k.a lens rubbing), rinsing, and overnight storage of the lens in a preserved saline or disinfecting solution. Digital cleaning of the lens is important to free accumulated deposits such as proteins, lipids, and microbes from the lens surface. Rinsing the lens further removes the accumulated deposits and overnight storage in the contact lens solution disinfects and hydrates the contact lens (21). More recently, solutions have been developed to clean and disinfect contact lenses that do not require digital cleaning.

Cleaners are divided into surfactant cleaners and enzymatic cleaners. Surfactant cleaners are used to remove lipids, oils, and environmental pollutants. Enzymatic cleaners are used to remove proteins on the surface of the contact lens and may be derived from plants, animals, or bacteria. Currently, the most popular method for the cleaning and disinfection of soft contact lenses is the multipurpose contact lens solution (MPS). Multipurpose contact lens solutions are used as the cleaning solution during digital lens cleaning, the rinsing solution, and the overnight storage and disinfection solution. Several multipurpose contact lens solutions have been designated by the FDA as "no rub" cleaning solutions, where the patient simply rinses the lens for a designated period of time and then stores the lens in the solution overnight after removal from the eye. Regardless of no rub approval by the FDA, digital cleaning is recommended with the use of silicone hydrogel lens materials. Other cleaning and disinfection systems available for use with soft contact lenses include hydrogen peroxide based systems, thermal disinfection units, ultraviolet or microwave radiation disinfection units, or ultrasonic mechanical agitation units (19,20,22).

Cleaning and disinfection of rigid gas permeable contact lenses is similar to the care of soft contact lenses. The rigid gas permeable lens is removed from the eye, cleaned digitally, rinsed with tap water or saline solution, and stored in a preserved conditioning solution. Two bottle cleaning systems consist of a contact lens cleaning solution and a lens conditioning solution. The cleaning solution typically contains a surfactant cleaner and may also have an abrasive component, such as silica beads, to help remove deposits from the lens surface (23). After cleaning, the cleaning solution is rinsed from the lens surface using tap water or saline solution. The contact lens is then placed in conditioning or disinfecting solution for overnight storage. The use of tap water for rinsing rigid gas permeable contact lenses is a subject of much debate and should be avoided in situations where the quality of the water is suspect (24,25). The use of tap water should be avoided in patients with an overnight or extended wear schedule. Multipurpose one bottle cleaning systems are available for use with rigid gas permeable contact lenses. As with soft contact lens care systems, the multipurpose lens care solution is used as the cleaning, rinsing, and storage solution. Dependent on the formulation, the multipurpose solution may have to be rinsed from the lens surface with tap water or saline solution prior to lens insertion. Unlike multipurpose solutions used in soft contact lens care, multipurpose solutions used with rigid gas permeable lenses are not indicated for no rub usage (20).

CORNEAL PHYSIOLOGY AND RESPONSE TO CONTACT LENS WEAR

The cornea is a multilayered, avascular tissue that receives oxygen from the atmosphere through diffusion of oxygen in the preocular tear film. The nutrient supply for the cornea comes from the anterior chamber of the eye (26). When a contact lens is placed on eye, oxygen from the atmosphere must pass through the lens matrix or must be transported to the corneal epithelium by the tears pumped underneath the lens by the blink response. A chronic lack of oxygen, referred to as corneal hypoxia, will lead to swelling of the cornea called corneal edema (27). The permeability of a contact lens material, referred to as the Dk of the lens, is the ability of oxygen to permeate through a contact lens material. The Dk is determined by the chemical composition of a lens polymer. A more relevant measure the oxygen reaching the cornea through a contact lens is the oxygen transmissibility, referred to as Dk/t, where Dk is the oxygen permeability of the lens and t is equal to the average thickness of the contact lens (28). The first contact lens material, PMMA, had a Dk of 0 and relied solely on the pumping of tears (a.k.a the tear pump) beneath the contact lens with the blink to carry oxygen to the corneal surface (11). Insufficient oxygen to the cornea can lead to a number of adverse events, such as corneal edema or epithelial breakdown (27). Therefore much emphasis has been placed on the oxygen permeability of contact lenses used on a daily wear and an extended wear basis.

The ultimate goal traditionally has been the development of a contact lens material with oxygen permeability

sufficient enough to mimic the conditions that occur when no lens is worn. In 1984, Holden and Mertz reported that a contact lens must have an oxygen transmissibility of $87.0 \pm 3.3 \times 10^{-9}$ $(cm \cdot mL\,O_2)/s \cdot mL \cdot mmHg)$ to limit the cornea to the 4% corneal edema noted after sleeping with no contact lens wear (29). A number of different hydrogel lens designs were utilized in the study with water content from 38.6 up to 75% as well as the Silsoft silicone elastomer lens. In 1999, Harvitt and Bonanno revisited oxygen permeability and corneal swelling to compensate for the effect of acidosis on oxygen consumption. Harvitt and Bonanno found that with decreasing contact lens oxygen transmissibility there was an increase in corneal stroma acidosis effectively reducing the amount of oxygen available to the cornea. The oxygen transmissibility required to prevent stromal anoxia after compensation for acidosis in closed eye conditions was found to be 125 barrer$\cdot mm^{-1}$ (1 barrer $=10^{-10}$ $cm^2 \cdot s^{-1} \cdot cmHg^{-1}$ or 7.5005×10^{-18} $m^2 \cdot s^{-1} \cdot Pa^{-1}$) (30). Oxygen transmissibility in traditional hydrogel lenses is limited the water content of the lens and the overall lens thickness. Oxygen transmissibility in hydrogel lenses is increased by increasing the water content while making a thinner contact lens. The theoretical best oxygen permeability for a hydrogel lens would be 80 barrer, the oxygen permeability of water (8). Thus, hydrogel lenses will never be able to obtain a level of oxygen transmissibility set forth by Holden and Mertz or Harvitt and Bonanno.

With the development of the Holden–Mertz criteria and subsequent adjustments by Harvitt and Bonanno, there was a renewed interest in silicone contact lens technology. The original patents for the siloxane hydrogel were filed in the late 1970s. The advantage of a marriage of silicone with hydrogel technology would be dramatically improved comfort and wettability compared to pure silicone elastomer lenses. The permeability of pure dimethylsiloxane is 600 barrer compared to 80 barrer for water (8). Thus, the incorporation of silicone into the hydrogel material greatly increases the oxygen permeability of the contact lens. The first silicone hydrogel lens approved for use in the United States was the Bausch and Lomb Purevision (Balafilcon A; 36% water) silicone hydrogel lens. The lens was approved for daily wear and later for up to 7 days of extended wear. The oxygen transmissibility of the lens was measured at 110. The CibaVision Focus Night and Day lens (Lotrafilcon A, 24% water) was approved for sale in the United States in October 2001. The CibaVision Focus Night and Day lens was the first contact lens approved for 30 days of continuous extended wear since 30 day extended wear approval was rescinded by the FDA in 1989. Purevision lenses received 30 day continuous wear approval one month later.

SAFETY

Contact lenses, when used properly under the care and supervision of a licensed practitioner, have been proven to be a safe and effective form of vision correction for millions of patients. However, as with any medical device, the risk of adverse events may occur with their use. With the approval of hydrogel lenses for extended wear in the 1980s, there were a number of severe adverse events reported. The most alarming were reports of increased incidence of corneal ulcers, also known as microbial keratitis or ulcerative keratitis, associated with extended wear schedules. Public debate on the safety of extended wear contact lenses grew as reports of serious complications rose in the media. The Contact Lens Institute sponsored two landmark studies on the incidence and relative risk of microbial keratitis with daily wear and extended wear contact lenses.

Poggio et al. reported in 1989 that the incidence of ulcerative keratitis was 4.1 cases per 10,000 people for daily wear hydrogel lenses. The incidence of ulcerative keratitis for extended wear hydrogel lenses was reported at 20.9 cases per 10,000 people per year (31). This compares to an incidence of 2.0 per 10,000 people per year for hard PMMA lenses and 4.0 per 10,000 people per year for RGP wearers. Poggio et al. reported a trend for decreasing incidence of ulcerative keratitis with extended wear lenses with increasing age that was not statistically significance ($p = 0.07$). Schein et al. (32) published a case-control study on the relative risk of ulcerative keratitis with the use of daily wear and extended wear of hydrogel contact lenses. They found that wearing extended wear lenses overnight produced 10–15 times the risk of ulcerative keratitis compared to wearing daily wear lenses. The race, sex, and age of the patient were not related to the relative risk of ulcerative keratitis. Smokers were found to have about three to four times the risk of developing ulcerative keratitis compared to nonsmokers. The study showed that the risk of ulcerative keratitis increased significantly with the number of consecutive days lenses were worn (32). Due to the results of these studies, the FDA rescinded the approval for 30 days of extended wear and applied a limit of 7 days maximum extended wear with hydrogel lens. The 7 day limit for extended wear was reexamined with the development of silicone hydrogel contact lens materials. Silicone hydrogel lenses were utilized in Europe prior to their approval for use in the United States. A number of articles were published on the clinical performance and safety with the use of these lenses. Iruzubieta et al. published a report in 2001 on the clinical performance of the CibaVision Night and Day lens. Seven patients discontinued lens wear due to lens discomfort and seven discontinued lens wear due to positive slit lamp findings out of 85 patients dispensed lenses. There were two cases of sterile peripheral ulcers and two cases of superior epithelial arcuate lesions. There were no cases of microbial keratitis in the study (33). Nilsson reported on a study of the Purevision contact lens used for 7 day verses 30 day extended wear. Nilsson randomized 504 patients into 7 day extended wear or thirty day extended wear. There was no statistically significant difference between the 7 day and 30 day group in the prevalence of objective findings and there were no incidents of microbial keratitis reported over a 12 month period (34).

Although microbial keratitis is the most severe potential adverse event possible with the use of contact lenses, a number of conditions may result from the use of contact lenses. Corneal changes from chronic oxygen deprivation, such as endothelial polymegethism, corneal striae, and

epithelial microcysts have been associated with the development of microbial keratitis. Contact Lens-Induced papillary conjunctivitis (also known as Giant papillary conjunctivitis), sterile contact lens peripheral ulcers, contact lens acute red eyes (CLARE), and epithelial staining or abrasions are other significant adverse events associated with the use of contact lenses, particularly on an extended wear basis (27,35).

The use of contact lenses while swimming is contraindicated. If swimming in contact lenses is desired or unavoidable the patient should use of watertight goggles to prevent water from coming into contact with the lens. Rigid gas permeable contact lenses typically will displace from the eye during swimming as they lack adhesion to the corneal surface. Soft hydrogel and silicone hydrogels, while having the adhesion required to prevent lens loss while swimming, are prone to absorbing substances present in the water. Choo et al. published a study examining the establishment of bacterial colonies on soft hydrogel and silicone hydrogels after exposing lenses to chlorinated water while swimming. Of 28 lenses examined, 27 revealed bacterial colonization. The most prominent bacterial colony observed on the lenses was Staphylococcus epidermidis, which was found to be the most common bacteria in the water. Sixteen lenses examined that were not worn during swimming revealed only three lenses with bacterial colonies. No differences were observed in bacterial colonization for hydrogel lenses verses silicone hydrogel lenses (36). Given the potential for bacterial colonization and possible subsequent microbial infection, contact lenses should not be worn in conditions where the lens may be exposed to contaminated water.

ASTIGMATIC DESIGNS

Astigmatism is a condition where the cornea has a toric surface. A toric surface features two different curvatures located in meridians that are 90° apart. As a result, light is focused into two different line foci which causes blur (10). Astigmatism can be corrected with the use of rigid gas permeable or soft hydrophilic/silicone hydrogel lenses.

There are three rigid gas permeable lens designs utilized in the correction of astigmatism. Spherical gas permeable lens designs can correct astigmatism in patients where the astigmatism observed, as determined by refraction, matches the corneal astigmatism, determined by keratometry or corneal topography. The lens, when placed on the eye, corrects the patient's astigmatism through the use of a "tear lens". The tear lens is generated when the space beneath the contact lens created by the difference in curvature between the cornea and the contact lens is filled by the tear film. The tear lens refracts light to focus light from each meridian onto the retina providing the patient astigmatic correction. Spherical rigid gas permeable contact lenses can correct patients with two diopters or less corneal toricity (37). With corneal astigmatism of greater than two diopters, a spherical lens may flex, decenter, or fall out of the eye when the patient blinks. For patients with large amounts of corneal astigmatism, a spherical gas permeable contact lens design is not feasible. A rigid gas permeable lens can be designed with two different base curvatures, anterior curvatures and refractive powers in each meridian (38). These lenses, referred to as back surface toric lenses or bitoric lenses are custom designed for each individual patient. Patients with minimal or no corneal toricity who require astigmatic refractive correction with rigid gas permeable contact lenses can be corrected with a lens that has a spherical base curve and toric anterior surface curvatures to correct the patient's refractive error (37). These lenses, referred to front surface toric prism ballasted gas permeable lenses, are used infrequently due to the popularity of soft toric contact lenses.

Soft hydrogel and silicone hydrogel contact lenses can be used to correct astigmatic refractive error in patients. Patients with low amounts of refractive astigmatism, typically under -0.75 diopters cylinder, can be corrected using the spherical equivalent refraction power (sphere power + ($0.5 \times$ cylinder power)) in a spherical contact lens. Patients with more than -0.75 diopters of refractive astigmatism should be corrected with a soft toric contact lens (37). The soft toric contact lens design may have toric front or back surfaces with the appropriate refractive power in each meridian. Soft toric contact lenses will assume the shape of the cornea; therefore there is no tear lens present to correct the astigmatism. Soft toric contact lenses must stay properly oriented on the cornea to correct the patient's astigmatism. The meridians of the soft toric contact lens should correspond with the meridians of the refractive error of the eye. Rotation of the contact lens will improperly align the contact lens meridians with the patient's refractive error and lead to a reduction of visual acuity. With increasing amounts of astigmatism, stability of lens rotation becomes more critical to ensure clear vision (39).

Lens stabilization can be achieved by a number of techniques. The most popular method of stabilization is the use of prism ballasting. A prism ballasted contact lens integrates prism into the shape of the lens. The contact lens is designed with a varying thickness profile where the top of the contact lens is thinner than the bottom of the contact lens. The interaction of the eyelids with the prism moves the lens into the proper orientation by a principle called the "watermelon seed" effect. Another method of soft toric lens stabilization is the dual thin zones, or double slab off, design. Lenses that use a dual thin zone design have thin superior and inferior portions of the lens with a thicker central area. The interaction of the eyelids with the dual thin zones stabilizes the lens and holds the lens in the proper orientation through the watermelon seed effect. Soft toric contact lens rotation stabilization methods, such as lens truncation, where a small portion of there inferior portion of the lens is removed so that the lens is stabilized by the lower lid margin and eccentric lenticularization may be used to ensure that the power meridians in the soft toric lens remain oriented in their proper position on the eye to ensure proper astigmatic correction (38,39).

PRESBYOPIA AND MONOVISION

Presbyopia is the reduction of accommodation that occurs with aging. Accommodation allows the intraocular lens to

change shape, thus increasing the amount of plus power and allowing the patient to see near objects (10). Most patients will start to have a reduction in accommodation in their fourth decade. In spectacles, presbyopia is corrected through the use of bifocals. In contact lenses, presbyopia can be corrected with a number of multifocal contact lens designs or through monovision. Prior to the development of multifocal contact lenses the method of presbyopic vision correction with contact lenses was monovision. In monovision, one eye is corrected with the full refractive error needed to provide distance vision. The other eye is corrected with a lens that focuses the eye for a set near point. Monovision is most effective with patients that have low to moderate amounts of presbyopia (40). With increasing amounts of presbyopia, the disparity in refractive correction between the distance eye and the near eye becomes more significant with patients reporting more difficulty with binocular vision and stereopsis (41,42). Despite the development of multifocal contact lenses, monovision is still utilized with a single vision contact lens in one eye and a multifocal contact lens in the other eye in a technique referred to as modified monovision (41).

Translating bifocal, also known as alternating vision contact lenses correct presbyopia by having a distance and near corrective sections in the lens. The principle behind this correction is similar to bifocal spectacle lenses, where the upper portion of the lens provides distance correction and the lower bifocal portion provides the patient with near correction. The lower portion of the lens translates with eye movement. As the patient looks down, the bifocal section of the lens is pushed up over the pupil by the lower eyelid allowing the patient to see near objects (42). The orientation of the bifocal is maintained through the use of prism ballasting and the watermelon seed effect. As in spectacles, a translating trifocal design is available for patients who need more presbyopic correction. The majority of translating bifocal lens designs are rigid gas permeable lenses. Translating bifocal lenses work well with patients requiring high bifocal addition powers and have good eyelid apposition with the eye that will force the lens to translate with downgaze. The positioning of the lens optics is critical for translating bifocal designs. Translating bifocal lenses that do not translate in down gaze will provide no multifocal effect and poor near vision (41).

Another class of bifocal contact lenses is referred to as simultaneous vision bifocal lenses. Simultaneous vision lenses encompasses a number of lens designs, including aspheric lenses, concentric ring designs, and less commonly, diffractive optics (42). The unifying concept for each simultaneous vision bifocal design is the placement of images of the intended near target and the distance target on the retina at the same time. Simultaneous vision bifocal contact lens designs require the patient to ignore the image that is not relevant for the intended task. Simultaneous vision designs are best suited to patients who do not require high bifocal powers (41). Simultaneous vision contact lenses are particularly well suited for patients requiring intermediate vision correction for tasks such as computer use. Simultaneous vision lenses are the preferred bifocal contact lens design for patients who require near vision correction in straightforward gaze. Simultaneous vision lenses work well for these tasks because translation of the lens is not required to provide the multifocal vision effect (42). The majority of soft hydrogel contact lenses are simultaneous vision designs.

ORTHOKERATOLOGY

With the advent of the corneal gas permeable contact lens, practitioners observed that patients would report spectacle blur at night upon removal of the contact lens. Use of the rigid gas permeable lens would change the shape of the cornea during daily wear and cause a subsequent change in the patient's refraction and spectacle blur after removing the contact lens (43). In 1962, Jessen described the Orthofocus technique, the first published report of a controlled attempt to change the refractive error of a patient through the use of rigid contact lens (44). The procedure of using a rigid contact lens to change the refractive error of a contact lens patient was given the name orthokeratology. Orthokeratology was defined by as "the reduction, modification, or elimination of refractive anomalies by the programmed application of contact lenses or other related procedures" (45). Early attempts to reshape the cornea were accomplished by fitting the contact lens much flatter than the corneal curvature of the patient. Orthokeratology failed to gain widespread acceptance over the next 20 years due to limitations in corneal lens lathe technology. In 1989, a major advancement in the field of orthokeratology took place. Wlodyga and Stoyan presented the concept of reverse geometry lenses for "accelerated" orthokeratology (43). Most previous fitting philosophies featured traditional lens designs with peripheral curves flatter than the base curve of the lens. Wlodyga and Stoyan proposed using a reverse geometry contact lens with a central base cure that was flatter than the midperipheral curve. The advantage of such a system is improved lens centration, a decrease in the time to reach maximum effect, a more consistent treatment effect and a possibility to correct patients with higher degrees of myopia. Despite the advancement in orthokeratology lens designs, orthokeratology was still limited by the lack of FDA approval for the overnight use of this technology. With the development of hyperpermeable contact lens materials, the use of reverse geometry orthokeratology contact lens for overnight wear became practical. In 2002, the FDA approved the Paragon Corneal Refractive Therapy (CRT) overnight orthokeratology lens for overnight wear to temporarily correct myopia up to -6.00 diopters (46). As of January 2005, overnight orthokeratology lenses approved by the FDA are produced by Paragon Vision Sciences (Mesa, AZ) and Bausch and Lomb/Polymer Technologies (Rochester, NY) along with their approved manufacturers.

SUMMARY

Contact lenses have been utilized by millions of people around the world to correct a variety of refractive error conditions including myopia, hyperopia, presbyopia, and astigmatism. The study of contact lenses is a dynamic pursuit as advances in manufacturing technologies and

lens materials are incorporated into state of the art lens designs. Despite an excellent safety profile demonstrated through years of clinical practice, contact lenses remain medical devices that require expert fitting and monitoring by a licensed contact lens practitioner and patient compliance to ensure a successful outcome for the contact lens patient.

BIBLIOGRAPHY

1. Barr JT. 2004 Annual Report. Contact Lens Spectrum 2005; January.
2. Hofstetter H, Graham R. Leonardo and contact lenses. Am J Optom Arch Am Acad Optom 1953;41.
3. Enoch J. Descartes' contact lens. Am J Optom Arch Am Acad Optom 1956;33:77.
4. Barr JT. History and Development of contact lenses. In: Bennett ES, Weissman BA. editor Clinical Contact Lens Practice. Philadelphia: Lippincott Williams & Wilkins; 2005.
5. Mandell R. Historical development. In: Contact Lens Practice. 4th ed. Springfied: Charles C Thomas; 1988.
6. Barr JT, Bailey NJ. History of contact lenses. In: Bennett ES, Weissman BA, editors. Clinical Contact Lens Practice, revised ed. 1997 Philadelphia: Lippincott-Raven; 1997; chapt 11.
7. Chun M, Fox L, Zhou A. Disposable and frequent replacement hydrogel contact lenses. In: Bennett ES, Weissman BA, editors. Clinical Contact Lens Practice. Philadelphia: Lippincott Williams&Wilkins; 2005.
8. Nicolson P, Vogt J. Soft contact lens polymers: an evolution. Biomaterials 2001;22:3273–3283.
9. Landers R, Rixon A. Contact lens materials update: options for most prescriptions. Contact Lens Spectrum 2005(March).
10. Cline D, Hofstetter H, Griffin J, editors. Dictionary of Visual Science. 4th ed. Boston: Butterworth-Heinemann; 1997.
11. Cannella A, Bonafini J. Polymer chemistry. In: Bennett ES, Weissman BA, editors. Clinical Contact Lens Practice. Philadelphia: Lippincott Williams & Wilkins; 2005.
12. Phillips A. Rigid gas permeable corneal lens fitting. In: Phillips A, Speedwell L, editors. Contact Lenses. 4th ed. Oxford: Butterworth Heinemann; 1997.
13. Mandell R. Basic principles of rigid lenses. In: Contact Lens Practice. Springfield: Charles C Thomas; 1988.
14. Loran D. The verification of hydrogel contact lenses. In: Phillips A, Speedwell L, editors. Contact Lenses. 4th ed. Oxford: Butterworth Heinemann; 1997.
15. Yeung K, Weissman BA. Soft contact lens application. In: Bennett ES, Weissman BA, editors. Clinical Contact Lens Practice. Philadelphia: Lippincott Williams & Wilkins; 2005.
16. Mandell R. Fitting methods and philosophies. In: Mandell R, editor. Contact Lens Practice. Springfield: Charles C Thomas; 1988.
17. Bennett ES. Basic fitting. In: Bennett ES, Weissman BA, editors. Clinical Contact Lens Practice. Philadelphia: Lippincott Williams & Wilkins; 2005.
18. Uras R, Rah M. Sperical hydrophilic soft contact lenses. In: Mannis M, Zadnik K, Coral-Ghanem C, Kara-Jose N, editors. Contact Lenses in Ophthalmic Practice. New York: Springer; 2004.
19. Coral-Ghanem C, Bailey M. Maintenance and handling of contact lenses. In: Mannis M, Zadnik K, Coral-Ghanem C, Kara-Jose N, editors. Contact Lenses in Ophthalmic Practice. New York: Springer; 2004.
20. Tran L, Myung E. Contact lens care update. Contact Lens Spectrum 2005(April).
21. Atkinson K, Port M. Patient management and instruction. In: Phillips A, Speedwell L, editors. Contact Lenses. 4th ed. Oxford: Butterworth Heinemann; 1997.
22. Weisbarth R, Henderson B. Hydrogel lens care regiments and patient education. In: Bennett ES, Weissman BA, editors. Clinical Contact Lens Practice. Philadelphia: Lippincott Williams and Wilkins; 2005.
23. Bennett ES, Wagner H. Rigid lens care and patient education. In: Bennett ES Weissman BA, editors. Clinical Contact Lens Practice. Philadelphia: Lippincott Williams & Wilkins; 2005.
24. Koenig S, Solomon J, Hyndiuk R, Sucher R, Gradus M. Acanthamoeba keratitis associated with gas-permeable contact lens wear. Am J Ophthalmol 1987;103(6):832.
25. Shovlin J. Acanthamoeba keratitis in rigid lens wearers: the issue of tap water rinses. Int Contact Lens Clin 1990;17:47.
26. Mandell R. Anatomy and physiology of the cornea. In: Contact Lens Practice. 4th ed. Springfield: Charles C Thomas; 1988.
27. Kara-Jose N, Coral-Ghanem C, Joslin C. Complications associated with contact lens use. In: Mannis M, Zadnik K, Coral-Ghanem C, Kara-Jose N, editors. Contact Lenses in Ophthalmic Practice. New York: Springer; 2004.
28. Mandell R. Oxygen and the cornea. In: Contact Lens Practice. Springfield: Charles C Thomas; 1988.
29. Holden BA, Mertz G. Critical oxygen levels to avoid corneal edema for daily and extended wear contact lenses. Invest Ophthalmol Vis Sci 1984;25:1161–1167.
30. Harvitt D, Bonanno J. Re-evaluation of the oxygen diffusion model for predicting minimum contact lens Dk/t values needed to avoid corneal anoxia. Optom Vis Sci 1999;76(10):712–719.
31. Poggio E, et al. The incidence of ulcerative keratitis among users of daily-wear and extended-wear soft contact lenses. N Eng J Med 1989;321(12):779–783.
32. Schein O, Glynn R, Poggio E, Seddon J, Kenyon K. The relative risk of ulcerative keratitis among users of daily-wear and extended-wear soft contcat lenses. N Eng J Med 1989;321(12): 773–778.
33. Iruzubieta JM, et al. Practical experience with a high Dk Lotrafilcon A fluorosilicone hydrogel extended wear contact lens in Spain. Clao J 2001;27(1):41–46.
34. Nilsson SE. Seven-day extended wear and 30-day continuous wear of high oxygen transmissibility soft silicone hydrogel contact lenses: a randomized 1-year study of 504 patients. Clao J 2001;27(3):125–136.
35. Binder PS. Complications associated with extended wear of soft contact lenses. Ophthalmology 1979;86(6):1093–1101.
36. Choo J, et al. Bacterial populations on silicone hydrogel and hydrogel contact lenses after swimming in a chlorinated pool. Optom Vis Sci 2005;82(2):134–137.
37. Twa M, Moreira S. Astigmatism and toric contact lenses. In: Mannis M, Zadnik K, Coral-Ghanem C, Kara-Jose N, editors. Contact Lenses in Ophthalmic Practice. New York: Springer; 2004.
38. Lindsay R, Westerhout D. Toric contact lens fitting. In: Phillips A, Speedwell L, editors. Contact Lenses. 4th ed. Oxford: Butterworth Heinemann; 1997.
39. Epstein A, Remba M. Hydrogel toric contact lens correction. In: Bennett ES, Weissman BA, editors. Clinical Contact Lens Practice. Philadelphia: Lippincott Williams & Wilkins; 2005.
40. Mandell R. Presbyopia. Contact Lens Practice. 4th ed. Springfield: Charles C Thomas; 1988.
41. Schornack M, Coral-Ghanem C, Pena AdS. Presbyopia and contact lenses. In: Mannis M, Zadnik K, Coral-Ghanem C, Kara-Jose N, editors. Contact Lenses in Ophthalmic Practice. New York: Springer; 2004.
42. Bennett ES, Jurkus J. Presbyopic correction. In: Bennett ES, Weissman BA, editors. Clinical Contact Lens Practice. Philadelphia: Lippincott Williams & Wilkins; 2005.
43. Winkler TD, Kame RT. Orthokeratology handbook. Boston: Butterworth-Heinemann; 1995; p 113.
44. Jessen G. Orthofocus techniques. Contacto 1962;6:200–204.

45. Kerns RL. Research in orthokeratology. Part I: Introduction and background. J Am Optom Assoc 1976;47(8):1047–1051.
46. Barr JT. Contact Lenses 2002: Annual Report. Contact Lens Spectrum. 2003(January).

See also BIOMATERIALS: POLYMERS; BLIND AND VISUALLY IMPAIRED, ASSISTIVE TECHNOLOGIES; LENSES, INTRAOCULAR; VISUAL PROSTHESES.

CONTINUOUS POSITIVE AIRWAY PRESSURE

DAVID M. RAPOPORT
NYU School of Medicine
New York, New York

RON S. LEDER
Universidad Nacional Autonoma de Mexico
Mexico, Distrito Federal

INTRODUCTION

Beginning in the 1970s, positive-end expiratory pressure (PEEP) began to be added to the pressure applied during inspiration in patients undergoing mechanical ventilation. The rationale was that when a patient had loss of alveolar surfactant, the alveoli tended to collapse during expiration. "Holding them open" by offsetting the increased elastic recoil with a pressure that did not return to atmospheric at the end of expiration was beneficial to gas exchange because it prevented complete collapse with resultant shunting of blood past airless lung. This process was applied to both infant lungs (neonatal respiratory distress syndrome, RDS) and to adult lungs (adult respiratory distress syndrome, ARDS) with improved oxygenation as the main endpoint.

As PEEP was more widely used, it was observed that at the time of removal of respiratory support, some patients (especially newborns) benefited from PEEP for oxygenation despite being able to ventilate. This suggested that the strategy of providing a constant distending pressure to the lung (CPAP) during BOTH inspiration and expiration without increasing the pressure during the inspiratory phase (i.e., ventilation) provided some transient benefit during the period before extubation. In addition, it proved possible to apply CPAP via a nose or face mask after extubation, with continued benefit to the lung (Table 1).

CPAP IN OBSTRUCTIVE SLEEP APNEA/HYPOPNEA SYNDROME (OSAHS)

Basic Circuit and Rationale for Use

CPAP was first introduced in 1981 as a treatment for obstructive sleep apnea/hypopnea syndrome (OSAHS). The concept was initially proposed by Collin Sullivan (Australian Patent AU-B83901/82) as a pneumatic upper airway splint and later shown to work even in the presence of chronic respiratory failure (chronic hypercapnia) by David Rapoport (U.S. Patent 4,655,213). In this application of CPAP, the effect of interest is that of continuous positive airway pressure and not its effect on the lung (as with PEEP), although this is necessarily always present. The critical rationale is the effect a positive pressure in the AIRWAY has on the collapsible upper airway (i.e., the posterior pharynx and hypopharynx), which is not relevant in the intubated patient. The pressure is applied via nose or mouth mask and distends the area that extensive physiologic work has shown to have a tendency to collapse during sleep (especially during the negative pressure of inspiration).

The original circuit proposed consisted of a nose mask to which was attached either a pressure-dissipating threshold valve or a restrictor that created a roughly constant backpressure due to a constant bias flow provided by a blower or other source of compressed air. Early in development, it became clear that fans and blowers had better characteristics than piston-type high-pressure compressors, due to their ability to deliver high flow rates to the mask with control via motor speed, little dependence of delivered pressure on the backpressure, low cost, and quieter operation.

The original concept described by Sullivan was that the CPAP (pressure) was needed continuously to "hold" the airway open against a natural tendency of the walls of the airway to collapse due to loss of active muscle tone during sleep and the suction caused by inspiration. There is a tendency for some degree of airway collapse during sleep in everyone. Patients with the obstructive sleep apnea/hypopnea syndrome tend to collapse their airway to excess. In all cases, the collapse and airway obstruction occurs because of loss of tone in the airway muscles, whose role is to stiffen the walls against the suction created by breathing during inspiration. Although there has been much debate, most models of this process of collapse and its treatment with CPAP have suggested that the treatment pressure needs to be relatively constant at the point of collapse unless the patient changes body position, head and neck position, sleep state, or wakes up. The point of collapse is usually found to be at the back of the throat or at the level of the soft palate.

Leak Circuit Modification

Until 1985, CPAP was delivered by means of a restrictor or mechanical valve that was placed on the patient's nose mask. This valve, through its design and its passive mechanical properties, held the pressure at a value that was as constant as the mechanics of the valve could achieve (a so-called "threshold" valve, which opens more to discharge air when pressure rises). It also provided a vent for exhaled CO_2 and excess humidity, as it was located near the patient; a side-effect of the constant dissipation of pressure was venting of the circuit, including the exhaled gas from the patient. In 1985, Rapoport proposed that the valve used to set the pressure in the circuit could be removed from the mask to increase patient comfort. However, this required that the "venting" function (removal of exhaled CO_2 from the circuit) be performed separately. The modification consisted of a small controlled leak deliberately introduced near the mask that did not significantly dissipate the pressure (previously this had been a large leak or a threshold valve). This modified circuit is the most widely used hose circuitry in both CPAP and noninvasive mask ventilation.

A further improvement was instituted in the mid-1980s, when it was observed that the use of a threshold valve was

Table 1. CPAP Definitions (From www.cpap.com)

IPAP	This stands for inhalation positive airway pressure. This is the pressure setting that is used when you inhale.
EPAP	This stands for exhalation positive airway pressure. This is the pressure setting that is used when you exhale. This setting is always lower than the IPAP, making exhalation easier or more comfortable.
Bi-Flex	This setting establishes a level of pressure relief that takes place at the end of inhalation and at the start of exhalation. Settings of "1", "2", or "3" will progressively provide increased pressure relief. You can adjust this setting to suit your comfort level.
Spontaneous	Spontaneous means the patient breathes without assistance from a respiratory rate set on the bilevel.
4–20 cm	4–20 cm is the pressure range that can be delivered to the patient. A CPAP (constant positive air pressure) uses one constant pressure from 4 cm to 20 cm. This pressure is measured in centimeters of water pressure (H_2O).
Optional DC Cable	A cable that plugs directly into a dc port on the CPAP machine. This allows the you to plug into a dc power source, such as a battery or car cigarette lighter.
Direct Battery Operation	This feature means the machine has a dc port on the back of the machine in which you can use the Puritan Bennett Battery Pack or a deep cycle marine battery.
Auto Altitude Adjustment	Auto altitude adjustment is the CPAP machine's ability to compensate for changes in altitude automatically.
Ramp	The ramp feature allows the user to start treatment at a lower pressure, and as they fall asleep, the pressure slowly rises. This is a comfort setting and can be from 0 to 45 minutes on most CPAP machines
Hour/Session	This feature records the hours of usage and the sessions the machines is used longer than 4 hours.
Optional Software	Software is an option on some CPAP machines. The software can give details, compliance, and performance. The patient or physician downloads the data from the CPAP machine and uses it to determine how long a patient has used their machine each night and how well the machine is working to stop apnea/hypopnea events.
Leak Compensation	The CPAP machine compensates for mask leak, to keep the CPAP pressure accurate.
Heated Humidifier	This is an optional feature that can be added to the machine. Some machines have heated humidifiers designed to integrate with the machine while all can be used with stand-alone heated humidifiers like the Fisher & Paykel HC150.
Passover Humidifier	This is an optional feature that can be added to the machine. The Passover humidifier is a chamber filled with cool water. The CPAP machine tubing is routed through this chamber, and cool humidity soothes your nasal passages.
Data Card	A data card is a small card the same size as a credit card, that stores information to be placed into a Data Card Reader, downloaded to a computer, and read with optional software. Depending on the model of machine, the data card will hold either compliance data, performance data, or both.
Auto ON/OFF	This feature turns the machines OFF and ON when putting on or taking off the mask. When you put your mask on, the machines senses you breathing and turns itself ON. Take off the mask, and the machine turns OFF.

optional. This was because the blower could be designed to have a sufficiently flat flow-to-pressure relationship at a given speed of rotation to maintain a near-constant pressure during the increased flow of inspiration and decreased flow of expiration and changing amount of mask leak. Since then, CPAP blowers have either been entirely passive (set at one blower speed for each prescribed CPAP) or had some type of speed control that adjusted the speed in response to sensed pressure feedback. A few devices still use a threshold valve, but these have tended to replace the passive mechanical valves with active electronically controlled stepper motor-driven valves.

Variations in Delivered Pressure

Because active control of pressure is necessitated by removal of the threshold valve from the mask, there has been gradually increasing attention to modifying the pressure contour provided to the patient interface. In particular, various techniques have been used to keep a particular pressure constant. Conceptually, two distinct targets for stabilization of the pressure are either pressure at the blower or pressure at the mask. Initially, CPAP systems were designed to have a constant pressure at the blower, neglecting that this constant pressure at the blower would cause fluctuations at the patient mask (see above). More recently, attention has been directed to maintaining a constant pressure at other points in the circuit.

As air flows through a closed tube, it is driven by the pressure drop, which occurs progressively in the direction of flow. This implies that in any system with a nonzero resistance, there will always be a difference in pressure as one travels in the direction of flow along the tube. Specifically, as one travels from the blower along the tubing toward the patient's most collapsible airway point, the airway pressure will fall from that set at the blower and will differ depending on the rate of flow through the system and on where it is measured. Pressure differences between points along this route also depend on the direction of airflow (inspiration and expiration) and the size of the bias flow (e.g., through intentional or unintentional leaks at the mask. Thus, during inspiration, pressure is always higher by some small amount at the machine end of the tubing than it is at the patient's nose, and during expiration, it is often lower at the machine end of the tubing than at the nose if flow reverses. The amount of pressure difference between the machine end of the tubing and the patient depends on the resistance of the tubing connecting the two

and on the flow through the system, which is the sum of the patient's breathing airflow and any leak that occurs at the mask.

As the purpose of the CPAP is to maintain a therapeutic pressure that prevents upper airway collapse, a strategy to control the variations in this pressure must be established. Different approaches have been taken by different devices intended to deliver what is called CPAP. In the earliest CPAP devices, the valve located at the mask controlled the pressure; this intrinsically adjusted for changes in leak and reversal of flow from inspiration to expiration, and the only requirement of the blower was to provide an excess (not necessarily constant) flow to the valve located near the patient. When the valve was removed from the mask, pressure control shifted away from the patient to a point in the circuit near the blower. At least under some conditions, pressure can differ considerably from the desired therapeutic pressure as felt by the patient. The following is a list of some strategies adopted by current CPAP devices to deal with this (in terms of the original therapeutic intent, constant pressure at the mask during inspiration is key):

1. The controller sets a constant pressure at the machine (constant blower speed). This pressure must be slightly in excess of the patient's need; i.e., it must be sufficiently high to allow some fall during inspiration under maximal leak conditions, or the patient will be under-treated at this critical time in inspiration. This strategy necessarily implies that pressure at the patient will be in excess of the required therapeutic pressure at all other times, and this may contribute to patient discomfort.
2. The controller is driven by active feedback from the pressure as measured in the mask. This feedback will cause the blower to continuously vary pressure (at the blower) so as to maintain it constant at the mask. Either blower speed or valve opening may be varied, but pressure as sensed at the mask is the controlled variable. Until leak at the mask becomes enormous, this will be the closest to the original concept of a mask CPAP proposed by Sullivan and implied by the ventilator uses of PEEP and CPAP.
3. Control of pressure as exerted at the machine is based on assumptions about how the pressure will change as it travels along the tubing that connects the blower to the patient. Some devices assume a known pressure drop across the tubing and just add this to the desired therapeutic pressure. Other devices use the flow (or some estimate of flow such as blower speed) to calculate a predicted drop in pressure between machine and patient, creating a deliberate but variably higher pressure than prescribed—in an attempt to deliver the constant therapeutic target.

The above strategies handle changes in flow through the system (e.g., changing leaks), but they may not adequately address changes in backpressure during each breath related to breathing. This is because the intrinsic properties of blowers (fans) are such that at a fixed rotational speed, these devices tend to produce a flow (not pressure) that is heavily influenced by backpressure (e.g., the tubing resistance and the difference in magnitude and direction of flow between inspiration and expiration). As a result, fans tend to produce a relatively constant pressure against a wide range of loads (because of the changes in delivered flow). Thus, blowers result in a pressure profile during breathing that is close in their behavior to that of a circuit with a threshold valve. The result of this pattern of response to varying backpressures is that setting a constant blower speed results in a system that, to a first approximation, maintains a pressure that is relatively constant at the blower, independent of the patient's breathing pattern. However, blowers are not perfect in this regard. Blowers (fans), when kept at a constant speed within each breath, necessarily produce slight changes in delivered pressure (higher during expiration and lower during inspiration). Greater variability in breath size, and large leaks through the mask will all result in progressively greater pressure swings at the blower. Because of tubing resistance, even greater pressure changes will occur at the patient if the system is entirely passive. Specifically, pressure in the system and at the patient will fall during inspiration and rise during expiration to a value different from the treatment pressure.

The latest CPAP machines (U.S. Patent Application 2005/0188989) have begun to address these intrabreath pressure variations by modifying the pressure they deliver within individual breaths as a function of the instantaneous flow. The assumption is that this will improve pressure (exhalation) induced discomfort, which is reported by many users of CPAP, by limiting unnecessary rises in pressure above therapeutic during expiration. The simplest way to accomplish this pressure stabilization is to vary the blower speed in response to fluctuations detected in measured pressure. This type of control is a classic feedback system and is used to keep pressure constant at the blower by responding to any deviations or perturbations that occur in the desired constant pressure. Detected changes in pressure result in the controller changing the speed of the blower. Typically, pressure in the circuit varies as a result of changes in the patient's breathing (inspiration vs. expiration) or changes in the leak from the system at the mask, both of which produce changes in the backpressure felt by the blower.

As pointed out, varying the blower speed within a breath in response to the sensed instantaneous flow can also be done to vary the blower pressure profile such that it is maintained constant (without measurement) at the mask. An alternative to this is to reinsert an active threshold valve at the blower that accomplishes a similar function based on sensing flow in the circuit or some other measured variable that allows prediction of pressure in the mask. Much like the original CPAP circuit, pressure control is provided by driving the blower to produce a pressure in excess of that needed, and diverting ("bleeding off") some pressure in the system by variably opening the valve at a "threshold" pressure. However, instead of targeting a constant blower pressure, the valve is instructed to produce a pressure profile predicted to cause a constant mask pressure, by adjusting the opening and closing of the valve.

"BiLevel" PAP. Introducing a valve under microprocessor control provided an interesting opportunity to create patterns other than a constant pressure in the system. As soon as the pressure delivered to the patient begins to be significantly higher during inspiration than during expiration, however, this nonconstant pressure is fundamentally different from CPAP. In fact, this is similar to the behavior of artificial ventilation devices (ventilators). If the control system is made aware of when inspiration and expiration begin, the valve used in venting pressure can be adjusted to rapidly achieve higher and lower pressures in synchrony with the patient; this can assist or even fully replace patient breathing efforts and is the essence of mechanical ventilation. Whereas CPAP is the imposition of a control algorithm targeting a near-constant pressure in the system or at least at the patient, ventilation (sometimes referred to as "bilevel ventilation") is the imposition of a nonconstant waveform of pressure on the output of the blower so as to raise inspiratory pressure above expiratory pressure at the patient level. However, as the pressure profile (constant or variable at the patient) depends only on the programming of the valve controller, much confusion exists in the literature about whether a "bilevel" device is being used for "CPAP" or assisted ventilation.

In concept, patients with obstructive sleep apnea have no ventilatory control abnormality once the airway is open. Thus, assistance with ventilation (once the airway is splinted open) is not indicated. The original proposal for bilevel "CPAP" was not targeted at ventilation, but to date, there has been little in the published literature to support its use for "comfort" in patients with OSAHS alone. However, as a noninvasive ventilator, bilevel devices are very effective and deliver what is essentially a combination of PEEP and pressure support ventilation. Their use is clearly indicated in chronically hypercapnic patients and in those with nocturnal hypoventilation. Not only is there little logic to the use of this type of device for intermittent obstructive apnea, but also recent publications have suggested that they can exaggerate central apnea—presumably because increasing breath size (pressure support) will increase plant gain in the patient's respiratory control loop and tend to produce increased overshoot of the size of compensatory ventilatory efforts whenever there is instability of breathing, creating a classic "ringing" system.

Although the above discussion shows that bilevel ventilation is completely different in purpose and application from CPAP, current devices are such that they can deliver both modes with little change in their circuitry if they contain the means to rapidly change the pressure according to a prescribed algorithm. As a result, there continues to be confusion about what is being done when a physician prescribes a treatment for a patient. Clarification as to the algorithm being used by a setting on the machine requires a decision as to whether the device targets

- A pressure that is constant at the blower (the controller removes fluctuations at the blower). When the pressure is measured at the patient, there will necessarily be small fluctuations throughout breathing. Pressure will be lower at the patient during inspiration and higher during expiration than at the blower. This is passive CPAP.
- A pressure that is constant at the patient (the controller removes fluctuations at the patient). To accomplish this, the pressure when measured at the blower will be slightly higher during inspiration and lower during expiration. This is the purest form of classic CPAP.
- A pressure that is higher during inspiration than during expiration both at the blower and the patient. This type of pressure oscillation has as a purpose to actively assist the patient in magnifying his breathing efforts. The pressure changes assist the patient's spontaneous muscular breathing efforts when these are weak. This is active ventilatory support.

In the last two above cases, the expiratory pressure has been lowered from the value it would have achieved during expiration if the system was left to behave passively in response to the patient's breathing backpressure. The difference between the two algorithms above is not in the direction of change applied to the output expiratory pressure, but in the purpose for which it is lowered and the amount that pressure at the output of the blower is made to fall during expiration through active control. If the pressure at the blower is not lowered, i.e., forced to be constant, then the pressure as measured at the patient will rise during expiration. If the pressure at the blower is forced to fall slightly during expiration, the pressure may remain constant at the patient. Finally, if the pressure is forced to fall sufficiently at the blower during expiration, pressure will also fall at the patient during expiration. Unlike the first two algorithms, this pattern at the patient of a fall in pressure during expiration when compared with inspiration produces actual assistance to the patient's breathing efforts, and it defines assisted ventilation; this type of ventilatory assistance is fundamentally different from CPAP, whose purpose is only to hold the airway open.

MONITORING/TITRATION ISSUES

Recording the Pressure

The clinical prescription of CPAP is usually given as a single therapeutic pressure value. Typically, this is derived from some type of titration in a recorded sleep study. As should be evident from the earlier discussion of pressure gradients, this prescription pressure should to be related to how pressure was measured during the titration, as well as to how it will be implemented by the patient's CPAP machine, but this is often overlooked. If it assumed that the prescription is a generic one for a therapeutic pressure to be delivered in the mask, then the mask pressure should be the one measured during the titration study. However, most CPAP machines used in the laboratory do not provide this pressure as an easily available electronic output because they do not measure it. Instead they measure the pressure at the blower, which may differ by up to 1–2 cm H_2O from that at the patient and vary with respiration actively or passively. Furthermore, this gradient, as

discussed above, varies with the uncontrolled leak conditions at the mask and the amount and type of tubing circuitry, including whether a humidifier is in line. Furthermore, because many CPAP machines output an internally measured pressure as an analog or digital signal to facilitate laboratory recording during the sleep study, it is critical to know whether the actual mask pressure is being output or whether the output is a calculated estimate of pressure of a CPAP machine to that assumed to be present at the patient interface. In our laboratory, we prefer the actual measurement of pressure in the mask of the patient and provide this to the patient as his "prescription pressure." This should be independent of the brand of CPAP chosen for chronic use by the patient in its relation to adequacy of pressure if measured in the mask.

Algorithm for Deciding on the "Therapeutic" Pressure

When a patient undergoes a "CPAP titration," the pressure in the system during the period of monitoring is gradually increased until all evidence of upper airway obstruction disappears. Different laboratories titrate to different indices, but in principle most include trying to abolish evidence of both severe and partial obstruction as below:

Apneas (complete cessation of airflow caused by obstruction for at least 10 s) usually disappear first, so that at pressures above 8 to 10 cm H_20, it is rare to find obstructive apneas. Central apneas (failure of respiratory effort to occur, but usually without obstruction) may appear, especially at higher pressures. These are usually distinguished from obstructive apneas by the absence of persistent respiratory movements (rib and abdomen movements) during the apnea.

Hypopneas (significant reductions in airflow lasting at least 10 s) tend to predominate once CPAP has been applied at low levels. These are easily identified as obstructive by the presence of a flattened inspiratory flow/time contour, which differs from the sinusoidal shape of normal inspiration and breaths with unobstructed reductions in effort ("central" hypopneas). This "flow limited" behavior of obstructive hypopneas is explained by a Starling resistor model of the upper airway where dynamic collapse of the airway occurs due to the negative intraluminal pressure of inspiration. Transient appearance and disappearance of the flattened contour of groups of individual breaths indicates recurrent obstructive apnea and indicates the need for increased CPAP. Most laboratory titrations will strive to eliminate these events by raising CPAP.

Evidence of sustained elevated upper airway resistance (in contrast to discrete "events") may remain after all apneas and hypopneas disappear. This evidence can consist of stable snoring (upper airway vibration induced by unstable airway tissue), sustained runs of breaths with an inspiratory contour suggesting Starling behavior ("flow limitation"), or other direct measures of elevated airway resistance (e.g., direct measurement of intrathoracic pressure, from an esophageal catheter probe, divided by flow). It is currently often assumed that this evidence of high upper airway resistance must be completely relieved by elevating CPAP, but there is controversy as to the benefits of this form of titration. In some cases, raising CPAP to eliminate all such evidence of elevated upper airway resistance results in further improvement of sleep structure and decrease in daytime sleepiness. By contrast, in other subjects, few if any symptoms occur when the patient has sustained elevated resistance, provided this occurs without causing repetitive arousal. In this latter setting, raising the CPAP is difficult to justify, although often done. Very limited studies attempting to justify this titration approach have not to date supported any benefit of one approach over another.

During CPAP titration, in addition to defining the events that should prompt raising the pressure, it is important to consider when the pressure may be too high (and thus needs to be lowered). Although it is generally assumed that the lowest effective pressure is most comfortable and excess pressure will disrupt sleep, this has not been shown by controlled trials. However, most titration studies should include periodic reductions in CPAP once breathing and sleep have been stabilized to test for the lowest pressure at which evidence of airway instability (see above) recurs. This pressure may be different at different times in the night, and it is almost always different in the supine position and during REM sleep. These observations challenge the concept of a single prescription of CPAP.

Auto-Titration

The above discussion has assumed that a single therapeutic pressure at which the upper airway is effectively splinted exists for a given patient, and that this pressure remains relatively unchanged over time (within each night and across nights). There is ample evidence that this is NOT true. Where it has been studied, it is strongly suggested that for many patients in the supine position, upper airway obstruction is more severe and/or takes more CPAP to treat (although these are not synonymous). There may also be differences in the CPAP needed during REM and non-REM sleep. Thus, when a single pressure is prescribed, most practitioners use the highest pressure needed during a prolonged period of titration (e.g., at night), knowingly over-treating during the rest of the time.

Beginning in about 1990, several investigators began to automate the titration algorithm for choosing a pressure. The concept evolved of a feedback loop that constantly adjusted the CPAP based on sensing either frank apnea, hypopnea, or indices of upper airway abnormality like snoring and/or the contour of the inspiratory airflow. These devices were called auto-titrating CPAP or Auto-CPAP. Two conflicting goals were suggested for optimizing their function—maximizing the efficacy of CPAP and improving patient compliance by reducing pressure to the minimal need at all times. The first of these was to respond to unexpected increases in need in order to prevent undertreatment. The latter was to prevent unnecessarily high values of pressure at a time they were not needed. As both the signal driving feedback and the time constants of the systems developed varied greatly, it is difficult to address the whole group of Auto-CPAP devices in a single study. In particular, the effectiveness of the decision process for raising and lowering the CPAP will dictate whether the final pressure profile is high or low compared with CPAP.

This is not the logical target by itself, and only an outcome such as quality of sleep, improved hours of use by patients, and ultimately, improved daytime function and reduced sleepiness, can be used to evaluate the punitive value of Auto-CPAP over constant pressure. To date, however, limited data support this in large groups of patients. Some data suggest improved compliance with specific devices.

Having said that, no data suggest that the more reasonable of these devices is any LESS effective than CPAP, but some Auto-CPAP devices occasionally show changes in pressure that do not bear any logical relationship to the patient's breathing (runaways), and there is every reason to assume these will impair sleep.

A logical approach to evaluating such devices needs to address several questions before beginning to ask whether long-term use is effective or better than traditional CPAP:

1. Which signal is driving the response of Auto-CPAP? Possible signals include the flow signal amplitude (apnea and hypopnea), shape (detection of starling resistor behavior in the form of "flow limitation shape" as described above), vibrations (e.g., snoring and airway instability), breathing pattern on a longer timescale, direct sound measurements, and direct measures of airway abnormality (e.g., measurement of impedance via forced oscillation technique). The existing devices are driven by different signals, and new devices appear frequently. When compared head-to-head these devices have different responses to breathing test-waveforms, and both bench and patient testing is not yet standardized.
2. What makes the pressure rise? Is a response sought to each abnormal breath or detection of abnormal impedance? Is the pressure adjusted after a "testing" protocol—e.g. a periodic deliberate lowering of the effective pressure to induce some endpoint of abnormality?
3. When is pressure lowered, and after how long? Is continuous testing possible (as with forced oscillations to measure impedance) to which pressure can be lowered when the control variable is low, or is "normal" a condition that, once achieved, provokes a prolonged period of constant pressure (e.g., what is the response to "normal breathing" when detected)? When pressure is lowered, is this a provocative test, or an attempt to detect over-treatment? How frequent are pressure decreases? The implications of these decreases and their endpoint are physiologically significant—"testing" too frequently with a nonsubtle endpoint (e.g., an apnea or an arousal) will disrupt sleep. Testing too infrequently for decreasing pressure will produce ever increasing therapeutic pressure because there will be insufficient compensation for unavoidable errors in the algorithm's detection of a need to raise pressure.

Furthermore, a constant tradeoff exists between the need to optimally set CPAP for a stable physiologic state (e.g., in stable stage 2 sleep in the supine position, a pressure of x cm H_2O may be appropriate for long periods) and the need to respond with a rapid change in CPAP to state changes affecting the airway (e.g., awakening, entering REM, or rolling from the lateral to the supine position). Each machine currently on the market and in development has made different decisions about the way to balance these needs, and the resultant behavior, although it can be described, is not clearly better or worse by simple criteria. Large numbers of patients are needed to show benefit in terms of daytime outcome or compliance with therapy, and these trials are not widely available, nor are the results from one machine easy to apply to another machine or even to a slightly modified algorithm.

This field is still in evolution, but there has been disappointment in the advantage of the approach as reflected in better therapy. Despite this, automation of titration may still have large benefits for patients, even if it is not "better" titration, or even "more comfortable" CPAP. This arises from a trusted algorithm being able to replace the costly CPAP titration study, which is currently usually done in an attended fully monitored laboratory setting. To date, only a few machines on the market have sufficiently reliable "auto-titration" that they can be left unattended and monitored on a first-time user of CPAP, with the resulting pressure behavior assumed to represent an accurate reflection of the patient's need for CPAP. Even the best available machines still over-treat and under-treat some patients, and it seems advisable to recommend that evaluation of the results of a titration study be reviewed (at least off-line) by an expert with more than an assessment of the pressure profile the machine chose.

Our laboratory chooses to review all Auto-CPAP titrations by examining the flow profile and looking to see if overall we agree with the induced rises in pressure. We also review the pressure profile for rapid uncontrolled and unexpected rises in pressure that end with an arousal of the patient, and usually assume these are erroneous.

Finally, if Auto-CPAP is used for titrating a patient's need with the intent of using a single pressure as a prescription, yet another "algorithm" must be invoked to translate a constantly fluctuating pressure into a single prescription. Review of the pressure and or flow tracings rarely results in a single pressure that is constant for much of the night. One must, on subjective grounds, discard excesses and ignore periods of inadequate therapy during the fluctuations. One proposal is to discard the highest pressures achieved during 5% to 15% of the night. There has been no testing of this approach by objective criteria of long-term benefit.

INDICATIONS FOR USE OF CPAP IN OSAS

Stated simply, CPAP is currently the first line of treatment and is indicated for reversal of sleep-induced abnormal upper airway behavior, provided it is severe and results in disruption of sleep with negative daytime consequences. When obstructive apneas and hypopneas occur very frequently and result in severe blood oxygen desaturations, it seems obvious that CPAP is needed. Formal trials of the benefits of CPAP have relatively conclusively shown benefit when more than 30 apnea/hypopneas occur per hour of sleep. This benefit is mostly in the form of reduced daytime sleepiness, although small studies have suggested

reductions in blood pressure, improvement in daytime cognitive performance, or reaction time after weeks to months of therapy. CPAP is now near universally accepted as the most effective therapy (better than surgery or oral appliances) but not always as the most acceptable therapy from the patient. This has resulted in compliance rates among moderate–severe apneics (see above definition), which range from 50% to 80%, leaving many patients suboptimally treated, or anxious to change to other treatments as they become available.

However, a more contentious issue is how mild can the physiological abnormality be before treatment with CPAP is either unnecessary or unacceptable to patients. Two relatively large clinical trials are currently underway to answer this question, but no definitive statement can be made at present. A therapeutic trial of CPAP may answer the question in individual patients who show abnormal respiratory events during sleep and have an overt complaint (such as excessive daytime sleepiness). The trial is considered successful provided that patients see a noticeable improvement in symptoms. Better documentation of the validity of this approach is urgently needed as recent studies have shown a very large number of subjects in the general population who have apnea–hypopnea indices ranging between 10 and 30 events per hour (up to 25% of the population), some of whom are asymptomatic, and others who have significant symptoms that might be due to this pathology. Anecdotally, many patients improve on CPAP, but many cannot adapt to chronic therapy. Some of these may benefit from alternative therapy, but CPAP may remain the most effective and definitive way to perform a therapeutic trial for all treatments for OSAHS.

ISSUES IN COMFORT/COMPLIANCE FOR OSAHS AND ANCILLARY TREATMENT ISSUES

Interfaces/Masks

As comfort is the most perceived issue affecting patient compliance, it is clear that the mask must be an important contributor to the patient's willingness to use the device. Although this is accepted dogma, compliance rates over the years in which CPAP has been available are not clearly changing, and much of the willingness to use CPAP may also be affected by subjective patient perceptions of improvement (cost/benefit) and the reinforcement they get from the care provider. It is rare that a patient will use CPAP if the prescribing physician does not believe it works. Many types of nasal, oral, and full-face interfaces have been developed to maximize comfort. Nasal masks are currently most used, and details of material, shape, supporting extensions to relieve pressure points, and so on are beyond the scope of this discussion. Non-mask nasal interfaces also exist ("pillows" or "prongs") and may help address issues of claustrophobia, variant facial anatomy preventing a good seal with a mask, and personal preference. Oral interfaces are less common, but they have a devout following by some patients. Finally, for those with large leakage out of the mouth when the nose is pressurized, full-face masks may provide an alternative. Chin straps are frequently used to reduce mouth leak. It is clear that the technologist who knows the available masks and spends time trying multiple ones with a new patient will have greater success than one using the "one size fits all" approach.

Headgear

Like masks, a variety of headgear exist. These affect fit of the mask, pressure on the nasal bridge, tension of the straps, and even appearance. There is little published on the relative effect these have on compliance or patient preference, but it seems this is an important area.

Oxygen

Some patients (a minority) who use CPAP have a concomitant or related need for supplemental oxygen. As the oxygen is being delivered into a larger air stream, the rate of infusion (typically 2 to 10 L/min) may need to be different from that prescribed for a patient just breathing supplemental O_2. Furthermore, simple examination of the circuit will show that the leak (intentional and unintentional at the mask) will have a large effect on the delivered concentration of the O_2 bled into the air stream. A larger leak will change by a factor of 2–4 the final concentration of O_2 at the patient's nose. As CPAP masks are intrinsically leaky and variable, so it is predictable that the need for O_2 will change. In patients without evidence of hypoventilation and central regulatory abnormalities (usually marked by daytime hypercapnea, or arterial PCO_2 >45 mm Hg), giving too much is not a problem other than expense, so titration to the highest level needed to keep the oxygen saturation during all of sleep (including REM) $>90\%$ is the usual goal. However, in patients who tend to hypoventilate, excessive O_2 will worsen CO_2 retention and may lead to accumulation of serum bicarbonate, further depressing ventilatory drive even in the daytime. Thus, it is desirable to try to minimize O_2 use.

Finally, it is not often appreciated that the location at which O_2 is inserted into the CPAP circuit has a large effect. If the bleed is into the hose near the blower, the tubing promotes mixing and acts as a reservoir of a relatively fixed but lower O_2-enriched gas. Pattern of breathing, i.e., time in inspiration and expiration and tidal volume, may have less effect, but the degree of leak will still play a large role. In contrast, if the O_2 is bled directly into the mask, especially if this is beyond the leak in the circuit, the leak may have less effect. However, small changes in timing of breathing and mask size will have enormous effects on the inspired O_2 concentration as buildup of a small volume of near pure O_2 can accumulate during pauses and part of inspiration, whereas there is little volume to act as a reservoir and mixing chamber. This issue should be addressed by providing the location of O_2 connection in any prescription so that it will at least match the titration technique.

Humidity

Although at first glance it is not clear why humidity should be needed if breathing occurs through the normal nasal mucosal humidifying mechanisms, drying of the nose and nasal reactive obstruction are common complaints in CPAP users. Recent literature suggests that humidifying the

inspired air is helpful for these complaints, and anecdotally this may improve compliance with CPAP. Several mechanisms may be involved, but the most likely is that any leak out of the mouth will result in a constant desiccating flow through the nose with air below 100% relative humidity, rather than the usual bidirectional flow of normal breathing that replenishes the humidity in the nasal mucosa on expiration that was lost on inspiration. This situation is most prominent in mouth leak exacerbated by high pressure, after palatal surgery, and certain anatomical variants that promote mouth leak. Humidity does address this, but the degree of humidification of air is proportional to temperature as well as to the efficacy of the humidifier. Thus, simple cold pass over humidification is rarely sufficient, and heated humidification has been shown to have advantages in many studies. Because cooling of the air as it travels down the CPAP circuit may cause "rain-out" and a water hazard, some advocate the use of insulation or heating of the CPAP tubing as well as heated humidification of the air before delivery.

WRAP UP

Obstructive sleep apnea has probably been around for as long as there has been sleep, although it has been treated as a clinical disorder and syndrome only in the last 25 years. Epidemiological studies prove it is a major health hazard.

The National Institutes of Health recognize that most OSAHS patients remain undiagnosed and that the principal therapeutic apporach is CPAP. Even though the medical device industry has produced a variety of CPAP technologies and enough different makes and models to ravial the automobile industy, the therapy remains somewhat cumbersome, and so it is not associated with optimal compliance rates in the long term. It is, however, relatively noninvasive, efficient, and definitive at demonstrating therapeutic value for a particular patient.

The advantages of CPAP are that if tolerated it provides a relatively risk-free route to symptomatic relief of a serious disorder. An alternative in severe cases is a surgical procedure to place a tracheostomy. CPAP units require a prescription from a physician, however, several Internal vendors will accept a facsimile via fax or e-mail and will send CPAP machines and related equipment via mail order. Patients can select from model features on websites. Some vendors show over 250 styles and sizes of masks.

As a result of the need to improve patient compliance with CPAP therapy, manufacturers have constantly been improving the comfort and self-regulating capability of the machines to delivery therapy as needed, to match the changing conditions that occur during a night of sleep and over time. To accomplish this, devices include diagnostic capabilities to tailor the pressure of the therapy to the patient's needs. Just as pulse oximeters were the first wave of medical diagnostic devices to incorporate advanced data handling and storage capabilities, it seems that CPAP machines have incorporated technology as it has become available and serve as a test bed for applications such as improved control algorithms, performance, usability features, comfort, and compliance records. All of this information can be transmitted to the physician via the Internet or a small piece of flash memory. CPAP machines are available in the United States at prices ranging from $450 to $3000. The higher priced units can function as ventilation assist devices.

The result is that the same technology being developed for the growing CPAP industry can be used in other medical devices that play a role in self-regulating home therapy and objective tracking of patient compliance and/or progress.

The search continues for a less cumbersome method to splint the airway open either with surgery, an implanted stent, or a drug that stiffens the upper airway during sleep. CPAP will probably remain the first choice for a therapeutic trial.

See also RESPIRATORY MECHANICS AND GAS EXCHANGE; VENTILATORY MONITORING.

CONTRACEPTIVE DEVICES

MOLLIE KANE
Madison, Wisconsin

INTRODUCTION

Contraceptives and contraceptive devices are possibly the most widely used medical devices today. In the United States, 42 million women, or 7 in 10 women ages 15–44, are currently sexually active and do not want to become pregnant (1). In 2002, 98% of women who had ever had intercourse had used at least one method of contraception. Ninety percent had used a condom at least once. Of reproductive age women (15–44 years), 62% were currently using contraception. Over 10 million women in the United States had undergone female sterilization (2).

In fact, from a population perspective, there may be no other medical devices that have such a profound impact on quality of life. The implications of a person's ability to control when, or whether, they have a child are profound. For the individual, it will forever influence their health, educational and work options, and income. The typical U.S. woman, who desires two children, will use contraception for $\sim$30 years of her life in an attempt to control both the timing and number of her pregnancies (1).

What is especially unique about contraceptive devices, compared to most other medical devices, is that they are used by healthy people. They are neither for the detection nor treatment of disease. Therefore, patients and clinicians alike may have different thresholds for the acceptability of side effects and complications of contraceptives compared to other medical devices. The effects of contraceptives are often incorrectly compared to the effect of no treatment, rather than to the effect of pregnancy.

Despite the wide variety of contraceptive methods available in the United States and the high rate of women who use them, the number of unintended pregnancies remains

alarming. The Pregnancy Risk Assessment Monitoring System (PRAMS) study from the U.S. Centers for Disease Control and Prevention (CDC) looked at live births in 17 States in 1999 and found that 34–52% of live births were the result of unintended pregnancy (3). One-half of unintended pregnancies in the United States occur among women who were not using contraception (4). The other one-half is the result of contraceptive failures because of incorrect or inconsistent use. Of all unintended pregnancies, one-half end in therapeutic abortion (5). Worldwide ~190 million pregnancies occurred in 1995 and 1 in 3 of these ended in therapeutic abortion (6).

The high rate of unintended pregnancy could potentially be curtailed via changes in the contraceptive industry. A 2004 report by the National Academy of Sciences states that "those millions of women who choose to terminate a pregnancy, many submitting to unsafe and illegal procedures that can be life threatening, attest to the need for improved access to and utilization of existing contraceptive methods and the need for new and improved contraceptive options" (6). Research is beginning to focus on how to make contraceptives easier to obtain and simpler to use. Methods that are available over the counter may be more accessible to some women. There are experts who feel that many contraceptive methods are appropriate for sale without prescription (7). Research could determine whether this type of accessibility could lead to increased contraceptive success and decreased rates of unintended pregnancy. Emergency contraception could be dispensed prophylactically. To prevent pregnancy between the time that a method is dispensed and utilized, emergency contraception should be used as soon as possible. Marketing can continue to focus on "positive side effects" of contraceptives, such as decreased acne or increased sexual function.

Other new and creative methods of dispensing contraceptives may increase use and continuation rates. Planned Parenthood of Columbia/Willamette, OR is offering prescription of hormonal contraceptives via the internet. Women fill out an online questionnaire and are then contacted by phone by a nurse practitioner. Appropriate candidates may then receive pills, patches, or rings via overnight mail (7). Pharmacy access is another new method for prescription of contraceptives. Several states now allow pharmacists to prescribe emergency contraception, which may be expanded to other contraceptive types in the future.

In addition, the intrinsic properties of a contraceptive device affect whether a woman will be able to use it with every episode of intercourse or whether she will choose to discontinue use. Safety profiles and side effects affect a woman's willingness to start or continue a given method. For example, breakthrough bleeding, weight gain, nausea, or mood changes are side effects that are not dangerous, but that are likely to result in method discontinuation. Ease of use and disruption of intercourse may also affect a woman's willingness to continue use of a given contraceptive.

There is no perfect, completely effective contraceptive method. The best contraceptive choice for an individual or a couple will depend on many factors including, among others, their sexual attitudes and behavior. A thorough sexual history will assist the clinician in guiding a patient in choosing an appropriate contraceptive. Considerations include a woman's level of comfort with her own body, her likelihood of exposure to sexually transmitted infections (STIs), the degree of cooperation from the patient's partner, and the patient's need to have contraception occur separate from intercourse (8). Frequency of intercourse may have an effect on contraceptive choice. Those who have intercourse rarely may not want a method that requires daily action. However, those who have intercourse rarely are most likely to have intercourse unexpectedly and find themselves unprepared.

Individuals who will have a new partner or multiple partners may desire a barrier method of contraception to decrease risk of STI transmission. However, those who would find an unintentional pregnancy completely unacceptable may prefer the higher effectiveness rates of hormonal contraceptives or sterilization. Using both types of contraception, barrier, and hormonal–sterilization, provides the highest level of effective pregnancy prevention and provides protection from STIs. However, an inverse relationship is seen between the use of effective noncoital contraceptives and condoms. In one study of 12,000 U.S. high school students, pill use was the strongest predictor of failure to use condoms. It even had a more profound effect than use of alcohol or drugs or having multiple partners (9).

This article reviews concepts involved in contraceptive efficacy, informed consent, and contraceptive research and development. A brief description is provided of those contraceptives that can be considered medical devices. Much of the information in this article comes from the book Contraceptive Technology, 18th Revised ed. (10), as well as from peer-reviewed literature, online information from the CDC (http://www.cdc.gov/), literature from the Alan Guttmacher Institute, a nonprofit organization focused on sexual and reproductive health research, policy analysis and public education (http://www.agi-usa.org/index.html), and documentation from medical device manufacturers via the internet. This article will not address nondevice methods of contraception such as oral contraceptives, injectables, and natural family planning.

CONTRACEPTIVE EFFICACY

Contraceptive efficacy is based on many factors. Inherent efficacy of the method is only one aspect of whether pregnancy will occur. Factors that facilitate or interfere with proper use of the method also contribute to efficacy. Such factors can be method related, such as breakthrough bleeding, or interruption of spontaneity, or user related. Study populations that have high rates of intercourse will have higher rates of pregnancy, and therefore lower contraceptive effectiveness rates, than study populations with low rates of intercourse. Age of the user or of the study population will also affect efficacy rates. Fertility declines with age, thus reducing unintended pregnancies. Therefore, study populations containing higher populations of older women will find higher contraceptive efficacy rates. In addition, frequency of intercourse declines with age. Women with regular menstrual cycles will have higher rates of pregnancy and thus, higher rates of contraceptive failure,

than women with irregular cycles (11). For example, one study of women using the Reality female condom found that women with regular menstrual cycles were 7.2 times more likely to experience a pregnancy than women whose cycles were < 17 days or > 43 days (12).

Typical use contraceptive failure rates reflect the rate of pregnancy among women who report that a contraceptive is her method. It does not mean that she uses the method every time, uses it correctly, or even uses it at all. Perfect use reflects the rate of pregnancy that will occur among women who use a contraceptive correctly and with every episode of intercourse. These rates are usually estimated by researchers or are based on one or two studies.

Contraceptive failure rates are often calculated using the PEARL index, calculated as the number of pregnancies occurring divided by the number of woman years of exposure. The PEARL index is highly affected by the length of the study. Pregnancy rates are higher during the early stages of use of a method. Therefore, a study of 100 women using a given method for 1 year will have a higher PEARL index than a study of 10 women using the same method for 10 years.

Studies that administer pregnancy tests each month will find more pregnancies due to early detection of conceptions that end in spontaneous abortion prior to being recognized. Patients lost to follow up may be more likely to have experienced pregnancy than those who remain in the study.

Approximately 85 out of 100 sexually active U.S. couples that do not use contraception will become pregnant in 1 year based on estimates from populations who have low contraception use or who are actively trying to conceive (13). Conversely, the definition of infertility is the absence of conception after 1 year of unprotected intercourse. Contraceptive effectiveness rates should be regarded in comparison to what the pregnancy rate would have been if no contraceptive technique was used. For example, if use of a given contraceptive results in 15 pregnancies/100 women in the first year, then $\sim$70 pregnancies were prevented (85 minus 15).

INFORMED CONSENT

Informed consent for contraceptive device use should always be obtained. Individuals may not be willing to tolerate the same level of risk with contraceptive use as they would for other medical devices. The Department of Health and Human Services provides regulations regarding what constitutes appropriate informed consent for sterilization. It contains seven basic principles. These seven principles can help to guide appropriate informed consent for all types of contraceptives (14,15). The seven principles follow: (*1*) The patient should understand benefits of the method. (*2*) The patient should understand risks of the method (including risks of method failure). (*3*) The patient should know alternatives to the method (including abstinence and no method). (*4*) The patient should know that they have the right and responsibility to ask questions about the method. (*5*) The patient should know that they have the ability to withdraw from the method at any time. (*6*) All of the above issues must be explained in a way the patient understands. (*7*) Documentation that the caregiver has ensured understanding of each of the first six points should occur.

The term "informed choice" has been gaining popularity in the field of family planning. It implies the idea that all contraceptives have side effects and risks and that an individual must choose which side effects and risks are acceptable to them. An individual should not use a product if they find any one of the side effects or risks unacceptable.

MALE CONDOM

The male condom is a thin sheath of latex, lambskin, or polyurethane, placed over the shaft and glans of the penis. Condoms prevent pregnancy by blocking the passage of sperm into the vaginal canal. Condoms provide the only well-documented method for prevention of transmission of sexually transmitted infections and human immunodeficiency virus (HIV) by blocking the exchange of blood, semen, and vaginal secretions.

Latex condoms are by far the most popular and widely used type of male condoms. Latex condoms are available in the United States in a wide variety of styles and brands. They are available with or without reservoir tips or nipple ends, straight or tapered, smooth or ribbed, transparent or in a variety of colors, odorless, scented, or flavored. Average condom length is $\sim$7.5 in. (19 cm), with ranges from $\sim$6.5 (16.5 cm) to 9.5 in. (24 cm). Widths of $<$2 in. (5.1 cm) are considered snug and width $>$2.125 in. (5.4 cm) is "baggy". Thickness averages 0.0027 in. (0.0069 cm). Condoms with thickness $<$0.0019 in. (0.0048 cm) are considered extra sensitive. Those with thickness $>$0.0027 in. (0.0069 cm) are considered extra strong and may help with premature ejaculation. "Climax control" condoms are available with a small amount of benzocaine in them to aid with premature ejaculation. A Taiwanese company, SakuNet International provides condoms in 55 different sizes ranging from 3 (7.6 cm) to 9.4 in. (24 cm) in length and from 1.6 (4.1 cm) to 2.5 in. (6.4 cm) in diameter. Men log onto the SakuNet website and print out a measurement card that assists them in ordering the appropriate sized condom.

Lambskin condoms are manufactured from the intestinal lining of lambs. They contain pores that allow for the passage of small particles including HIV, hepatitis B virus, and herpes simplex virus (OO). Lambskin condoms are effective at pregnancy prevention because sperm are too large to penetrate the pores. Lambskin condoms should be used for contraception only.

There are four synthetic (polyurethane) condoms that are U.S. Food and Drug Administration (FDA) approved and available for purchase in the United States. These include two Avanti condoms (Durex Consumer Products), Trojan Supra (ARMKEL), and the eZ-on (Mayer Laboratories). Synthetic condoms have many advantages over latex condoms. Compared to latex condoms they are thinner and better at transmitting, and offer a less restrictive fit. They are stronger and more resistant to deterioration. They are compatible with water- or oil-based lubricants

and are acceptable for individuals with latex allergy. However, polyurethane condoms have not been well studied for their effectiveness in the prevention of STIs and HIV (16).

The majority of male condom research has been done on latex condoms. Latex condoms that are used correctly and consistently are effective at the prevention of pregnancy. Method failure (condom breakage) is quite rare. With perfect use as few as 2% of couples using condoms for 1 year will experience pregnancy (17). As much as 24–65% of condom breakage occurs prior to intercourse, therefore not increasing the risk of pregnancy (18). Reported rates of condom breakage or slippage during intercourse or withdrawal are variable, but low. In a study of 353 latex condoms used by sex workers in Nevada brothels, none broke or fell off during intercourse. Two (0.6%) slipped off during withdrawal (18).

Actual pregnancy rates with condom use are much higher than perfect use rates. This is due to the high likelihood of incorrect or inconsistent condom use. For typical condom use, ~15% of couples will experience a pregnancy over 1 year of use (17). This includes couples who fail to use the condom for every episode of intercourse. Other causes of condom failure include failure to use the condom throughout intercourse, partial or complete condom slippage, poor withdrawal technique, incorrect placement of the condom, and use of oil-based lubricants that degrade the condom.

Research has shown latex condoms to be extremely effective at preventing transmission of HIV during vaginal and anal intercourse. In a 1994 study published in the *New England Journal of Medicine* of 124 sero-discordant heterosexual couples, no sero-conversion occurred in 20 months of consistent and correct condom use. With inconsistent condom use, 10% of HIV-negative partners sero-converted >20 months and with no condom use 15% of partners sero-converted (19).

Variable data is available regarding the effectiveness of condoms in reducing the transmission of each STI. After reviewing all available data, the CDC concluded that "the lack of data about the level of condom effectiveness indicates that more research is needed—not that latex condoms do not work" in the prevention of STI transmission (20). They found that latex condoms, when used consistently and correctly, can reduce the risk of transmission of gonorrhea, chlamydia, and trichomoniasis. In addition, they reduce the risk of transmission of genital herpes, syphilis, chancroid, and human papillomavirus when the infected areas are covered by the condom. They also found that use of latex condoms reduces the risk of HPV associated diseases such as cervical cancer (20).

Condoms are regulated as medical devices by the FDA. Every condom manufactured in the United States is electronically tested for holes or weak areas. In addition, manufacturers are required to test samples from each lot of finished, packaged condoms. Tests that must be performed on the sampled condoms include the water leak test, the air burst test, and the tensile property test (21). Should failure rates of the sample condoms be unsatisfactory, the entire lot will be destroyed.

Condoms pose very few risks or side effects for their users. Latex condoms may not be used by individuals with latex allergy. Individuals who do not have latex allergy, but experience an allergic reaction to a condom, may be reacting to condom specific components such as lubricant, perfume, or other agents used in the manufacturing process (22). Some studies have found an increased risk of urinary tract infections in women using condoms with spermicide.

Latex condoms are vulnerable to heat and sunlight. They must be stored in a cool, dark place. They may be stored in wallets for up to 1 month. They must be used within 5 years of their manufacture date, or within 2 years if they are lubricated with spermicide (23).

THE FEMALE CONDOM

The female condom (FC), Reality (Female Health Company, UK, http://www.femalehealth.com/index.htm), is a 17 cm long, 7.8 cm wide, 0.05 mm thick sheath made of polyurethane. It has a ring at each end. The inner ring is at the closed end of the sheath and is used to insert the FC and to hold it in place behind the public bone. The outer ring is at the open end of the sheath and remains outside the vagina. The sheath loosely lines the vagina and covers some of the vulva. The FC is prelubricated with a nonspermicidal silicone lubricant. Additional oil- or water-based lubricant may be added. The FC prevents pregnancy and also protects the vagina, cervix, and external genitalia against STI. The same device is available in other countries under different names.

The polyurethane that makes up the FC is soft and odorless. It is stronger than the latex used to make male condoms. It conducts heat well, making sexual intercourse feel more natural. It does not contain latex and does not deteriorate at high temperatures or require any special type of storage. The expiration date is 5 years from the day of manufacture.

The FC may be placed up to 8 h prior to intercourse and removed at any time following intercourse. It does not require an erect penis to hold it in place. The FC should not be used together with a male condom as friction between the products could result in product failure.

The FC has been found to be effective in preventing STI transmission both *In vitro* (24) and *In vivo* (25,26). Clinical studies in the United States, Latin America, and Japan revealed pregnancy rates similar to those found with the use of other barrier methods. In a study looking at the use of the FC as the sole means of contraception among 328 monogamous couples at six sites in the United States, Latin America, and Japan, cumulative failure rates were ~20%. However, with "perfect" correct and consistent use, <10% of the subjects experienced an accidental pregnancy (27). The FC has no serious side effects, and it does not alter the original flora or cause significant skin irritation or vaginal trauma.

The FC is approved for single use only. A single Reality condom costs ~$2.00. The World Health Organization states that single use is preferable. However, in situations where female condoms are not available or affordable,

evidence suggests that the FC may be used safely up to five times.

VAGINAL SPERMICIDES

Vaginal spermicides contain benzalkonium chloride, octoxynol 9, or nonoxynol 9. They are available without prescription as films, suppositories, gels, creams, foam, and on condoms. Currently, only Nonoxynol-9 is available in the United States. It is a surfactant that works by destroying the sperm cell membrane. The spermicide is inserted into the vagina prior to any genital contact.

Vaginal spermicides used alone are fairly poor at preventing pregnancy. Pregnancy rates of 10–28% in 6 months of use have been found in recent studies. The higher pregnancy rates were found in a population of young women who had frequent coitus (28). Spermicide used together with condoms or other barrier methods can significantly increase the efficacy of these methods.

Spermicide may be placed up to 1 h prior to intercourse and must be placed again for each repeated episode of intercourse. Suppositories, foaming tablets, and films require time to dissolve prior to intercourse. Gel, tablets, suppositories, and film need to make contact with the cervix in order to be effective. Appropriate placement may be challenging for some women. Spermicides have not been associated with an increased risk of vaginitis. They may increase the risk of urinary tract infection in some cases. No adverse effects to spermicides have been reported, but toxicology data is limited. Spermicides should not be used together with any other vaginal medications or with vaginal cleansing products or douches. Spermicides should usually be avoided in the presence of sores on the genitals or when there is vaginal irritation and should not be used in the presence of cervical cancer.

Spermicides should not be used for the prevention of STI or HIV transmission. In 2000, a letter by Helen Gayle, the Director of the National Center for HIV, STD, and TB Prevention of the CDC, to clinicians reviewed several studies including that of the Joint UN Program on AIDS (UNAIDS). It concluded that Nonoxynol-9 does not prevent the transmission of HIV or other STIs. In addition, it may increase the risk of HIV transmission by causing irritation or ulceration of the female genital mucosa. Increased HIV transmission with use of Nonoxynol-9 was seen among women with very frequent high risk exposure. The effects of Nonoxynol-9 on HIV transmission among lower risk women are not known (29).

CERVICAL CAP

The cervical cap is a small, firm latex or silicon dome that creates a barrier over the cervix and holds spermicide in place. The Prentif Cavity Rim Cervical Cap (Lamberts Ltd, UK) was FDA approved in 1988, although it has been used in Europe for >60 years. However, Cervical Cap Ltd. (http://www.cervcap.com/), the sole U.S. distributor, dissolved itself in March of 2005, and as of June 2005, there are no other U.S. distrubutors. The Prentif Cap is made of latex in the shape of a thimble. A small groove on the inside creates suction to hold the cap in place over the cervix. The dome of the cap may be filled one-third full of spermicide prior to insertion, providing both a mechanical and a chemical barrier to the entry of sperm into the upper genital tract.

The Prentif Cap is available in four sizes, 22, 25, 28, and 31 mm, based on the diameter of the caps interior. It must be fit by a trained provider and is available only by prescription. It costs $49–68 when obtained directly from the distributor.

Approximately 20% of women are poor candidates for the cervical cap. A good fit is not possible in those with a very long or very short cervix, or when the cervix is asymmetrical. The Prentif Cap is also contraindicated in women with a history of cervical laceration, current cervicitis or infection, history of toxic shock syndrome, latex rubber allergy, a vaginal septum, cervical or uterine malignancy, or unresolved abnormal Pap smear.

Fitting should be performed near the time of ovulation, when the cervix is most small and firm. The bowel and bladder should be empty. Several sizes should be tested in ensure the best fit. With proper fitting, the cervix is completely covered. There should be 1–2 mm of space between the cap and the cervix. A firm tug on the dome should not displace the cap. When the cap is left in place for 10 min or more, a suction rim should be visible and palpable after removal.

Initial fitting should be performed at least 6-weeks postpartum and at least 2 weeks after therapeutic or spontaneous abortion. Refit should be performed annually, after childbirth, therapeutic or spontaneous abortion, after cessation of lactation, following a weight gain or loss of 10 lb (4.53 kg) or more, or after any reported accidental dislodgement.

The patient may leave the cap in place for up to 48 h. It should be in place for at least 8 h following the last episode of intercourse. Additional spermicide does not need to be placed for more than one episode of intercourse. Overfilling of the cap with spermicide may lead to slipping and dislodgement. If the cap is used with a condom, the condom must be lubricated to prevent "grabbing" between the two devices, which may result in dislodgement of the cap. The patient should use a back-up method of contraception for the first menstrual cycle of use and should check the cap for good placement after intercourse.

There may be degeneration of the latex over time. The FDA recommends annual replacement of the Prentif Cap. Exposure to heat, light, or petroleum-based products will accelerate deterioration of the cap. Early signs of deterioration include dimpling or thinning of the dome of the cap, a sticky texture, or a mosaic pattern forming on the dome. After use, the cap should be washed in soapy water, dried, dusted with cornstarch, and stored in its original container.

A second cervical cap, the FemCap, was FDA approved in 2003. It has been available in Europe since 1999. It is a silicon cap that covers the cervix and forms a seal to prevent the entrance of sperm into the upper genital tract. The FemCap is shaped like a sailor's cap with a brim, a dome, a groove between the brim and the dome, and a strap to facilitate removal. The dome covers the cervix, the rim

fits against the vaginal fornices and forms a seal with the vaginal wall. The groove stores spermicide and traps sperm. Pushing on the dome breaks the suction, allowing removal of the cap by hooking a finger around the strap.

Prior to insertion one-quarter teaspoonful of spermicide should be spread into the dome of the FemCap. An additional one-half teaspoon is placed in the groove. Because a majority of the spermicide remains in the groove and because the groove faces the vaginal side, exposure of cervical epithelium to spermicide is minimized.

The FemCap is available in three sizes. Sizing is based on parity alone, which eliminates the need for fitting by a healthcare provider. However, the patient must be trained to use the cap and a prescription is required. Sizing is based on the inner diameter of the rim of the cap. The 22 mm cap is for women who have never been pregnant. The 26 mm cap is for women whose pregnancies have resulted in spontaneous abortion, therapeutic abortion, or Cesarean section. Any woman who has delivered one or more full-term infants vaginally should use a 30 mm cap.

A smooth, symmetrical cervix is required for FemCap use. Therefore, it is contraindicated in women with a history of cervical laceration. It should not be used in the presence of infection of the upper or lower genital tract. The FemCap may be used by couples with latex allergy. It should not be used during menstruation.

The FemCap should be placed prior to any sexual arousal to allow optimal formation of a seal. It must be in for at least 6 h after the last episode of intercourse and may be in place for up to 48 h at a time. Additional spermicide should be inserted for repeat episodes of intercourse.

The Vimule cap was previously available in the United States, but was recalled by the FDA in 1983 due to a high incidence of vaginal lesions. It remains available in Europe and elsewhere. It is a bell-shaped cap with a flanged rim. It is available in 3 sizes: 42, 48, and 52 mm by external diameter.

Two additional caps are available in Europe, but are not FDA approved. The Dumas cap, by Lamberts (Dalston) Ltd., is a shallow, bowl-shaped latex cap available in five sizes: 50, 55, 60, 65, and 75 mm by external diameter. The Oves cap, by Veos, UK Ltd., is a clear, disposable, silicone cap with a loop on the rim to aid in removal. It comes in three sizes: 26, 28, and 30 mm.

LEA'S SHIELD

In 2002, the FDA approved Lea's Shield (YAMA, Inc.; Millburn, NJ, http://www.leasshield.com/about.htm), a reusable vaginal barrier contraceptive made of medical grade silicone rubber. The Lea's Shield is shaped like an elliptical bowl that covers the cervix. The posterior surface is thickened to improve fit in the posterior fornix of the vagina. There is an anterior loop to assist in removal. The bowl contains a centrally located valve to allow the passage of cervical secretions and air. Lea's Shield comes in one size designed to fit most women. It is available by prescription from Planned Parenthood and from the manufacturer.

A phase II clinical trial published, in 1996 by Mauck et al. (30), found efficacy rates for the Lea's Shield to be similar to those found in other studies of the diaphragm and cervical cap. Efficacy data were available for 146 women who used the Lea's Shield as their only method of contraception for 6 months. The adjusted 6 month life table pregnancy rates were 5.6% for users of Lea's Shield with spermicide and 9.3% for users of Lea's Shield alone (not a statistically significant difference with $p = 0.086$). This corresponds to 1-year failure rates of 9 and 14%, respectively. However, the study included a high percentage (84%) of parous women. For unknown reasons, barrier contraceptives in general are less effective among parous women. There were no pregnancies among the small number of nulliparous women in this study.

Studies are not available directly comparing the Lea's Shield's safety or efficacy to the diaphragm or cervical cap. It is not known whether uterine position effects efficacy. No adverse events have been reported with the use of Lea's Shield. Data on the effect of the Lea's Shield on cervical or vaginal epithelium are not currently available. In the study by Mauck et al. (30), discontinuation rates for device related reasons were low, with 84% of women, but only 55% of their partners, reported liking the Lea's Shield. This device may be used by individuals with latex allergy.

A clinician should check the initial fit of Lea's Shield. The device covers the entire cervix and the strap should be behind the symphysis. If these two criteria are met and the device feels comfortable to the woman, then it is considered a proper fit.

Prior to each use the bowl of the device should be filled with spermicide. The device may be inserted at any time prior to intercourse and may be left in for up to 48 h at a time. Additional spermicidal jelly is not needed for additional episodes of intercourse. Lea's Shield should be left in place for at least 8 h after the last episode of intercourse. It should not be used during menses due to a theoretical increased risk of toxic shock syndrome. After removal, the Lea's Shield should be washed with soap and water, air dried, and stored in its original silk pouch. The manufacturer recommends annual replacement of the device.

THE CONTRACEPTIVE SPONGE

The Today sponge (Allendale Pharmaceuticals, Allendale, NJ, http://todaysponge.com/.), is a disk shaped device made of disposable polyurethane foam infused with 1 g of nonoxynol-9 spermicide. The sponge was initially marketed in 1983, and then removed from the market in 1995. The discontinuation of the sponge was not based on any problems with safety or efficacy. Rather, routine inspection of the manufacturing plant revealed non-TSS causing bacteria in the water used to make the Today Sponge. The company decided that the cost of the modifications needed to continue production were prohibitive. However, these modifications have now been undertaken and the Today Sponge became available again in Canada and via the Internet in early 2005. The FDA approval, leading to U.S. sales, is expected in 2005. Approval in the United Kingdom is also imminent.

The proximal side of the Today sponge is concave to allow tighter fit over the cervix that decreases the chance of

dislodgement. The opposite side has a woven polyester loop to aid in removal. The polyurethane foam creates a texture that mimics vaginal tissue. The sponge's main mechanism of action is to release spermicide and 125–150 mg are released into the vaginal vault during 24 h of use. It also functions as a physical barrier to entry of sperm into the upper genital tract, and it absorbs sperm into its polyurethane matrix. The Today Sponge has a pH 6–8. It expires 18 months after the manufacturing date.

Studies in the United States reveal a 12 month failure rate of ~10% with perfect use. Actual use 12 month failure rates are higher, and accedental pregnancy will occur in 15–20% of couples using the sponge (31).

The Today Sponge is one size that fits all common. It is an over-the-counter device and requires no fitting or other clinician involvement. Prior to insertion the sponge is activated by generously wetting it with water, then squeezing it until it produces suds. It is then inserted deeply into the vagina to cover the cervix. It is effective immediately upon insertion. It may be left in for up to 30 h and may be used for more than one episode of intercourse during this time with no additional spermicide needed. It should be left in place for at least 6 h after the last episode of intercourse. After removal, it is discarded. There is a theoretical risk of toxic shock syndrome if the sponge is left in place >24–30 h.

The Today Sponge is held in place by the vaginal muscles. It is possible to experience expulsion with straining. If the sponge is found at the vaginal entrance within 6 h after intercourse, it should be pushed back in. If it is expelled entirely, a new sponge should be wetted and inserted immediately.

The Today Sponge may tear upon removal. Any residual pieces within the vagina must be removed. The Today Sponge will not dissolve over time. Retention of a piece of the sponge may increase the risk of toxic shock syndrome or vaginal infections.

The Today Sponge is contraindicated in women with a current upper or lower genital tract infection, and in women with a history of toxic shock syndrome. It should not be used during menstruation due to a theoretical increase in the risk of toxic shock syndrome. It should not be used during the first 8 postpartum weeks. Following spontaneous or elective abortion, the sponge should not be used until an examination has been performed by a clinician to ensure that cervical and vaginal tissues appear normal. The Today Sponge should not be used underwater (in pools or hot tubs), because large quantities of water entering the vagina may dilute the spermicide. It should not be used together with any vaginal medications as interactions have not been studied.

The Today Sponge is equally efficacious in both parous and nulliparous women (31,32). The Today Sponge cannot be used if either partner has an allergy to one of its components. One study found that 4% of U.S. women discontinued use due to allergic type symptoms (33). The Today Sponge contains a small amount of metabisulfite. No allergic reactions to this sulfa component of the Today Sponge have been reported. However, couples where either partner has a sulfa allergy should not use the Today Sponge.

The Today Sponge has not been shown to increase the risk of vaginal infections with normal use. Increased risk of vaginal candidiasis has been found with prolonged use of an individual sponge. Some women experience a white vaginal discharge during use of the Today Sponge. This is normal, but may be confused with vaginal candidiasis. Vaginal dryness may occur due to the sponge soaking up vaginal secretions. Adequate wetting of the Today Sponge prior to insertion should minimize this.

CONTRACEPTIVE DIAPHRAGM

The contraceptive diaphragm is a silicone or latex cup shaped device with a firm, yet flexible, rim and a dome that covers the cervix. The diaphragm provides a mechanical barrier to the entrance of sperm into the upper genital tract. In addition, the cup of the diaphragm provides a chemical barrier by acting as a receptacle for spermicide.

Reported effectiveness rates vary widely, from 70 to 99 % (34). With perfect use, the 1 year pregnancy rate has been found to be as low as 6% per 100 women (34). Diaphragm failure may include failure due to improper fitting.

Prior to insertion of the diaphragm, spermicide should be distributed over the dome of the device. The device is effective immediately after insertion and may be placed up to 6 h prior to intercourse. If >6 h have elapsed, and for more than one episode of intercourse, additional spermicide must be inserted in the vagina (without removing the diaphragm). The diaphragm must remain in place at least 6 h after the last episode of intercourse and should not be left in place for >24 h. After removal the diaphragm should be washed in warm, soapy water, and stored in a clean, dry container.

Fitting rings or fitting diaphragms may be obtained from the manufacturers. In order to fit the diaphragm, the patient is placed in dorsal lithotomy position. The clinician inserts the gloved middle and index fingers into the vagina until the middle finger reaches the posterior fornix. The spot where the index finger touches the inferior pubic arch should be marked with the thumb or an instrument. The fingers are withdrawn in this position. The fitting ring or diaphragm is then placed with the rim over the end of the middle finger and the opposite rim touching the mark of the inferior pubic arch.

The appropriate fitting ring or diaphragm should then be inserted. Fitting is appropriate if the anterior rim lies just behind the symphysis pubis, the posterior rim lies at the vaginal fornix, the rim touches both lateral walls, and the cervix can be felt through the dome. At least one size smaller and one size larger should always be fitted to check for the optimal fit. The best size is the biggest one that meets all fitting criteria, but cannot be felt by the patient. With a proper fit, the diaphragm will not dislodge or fall out with straining and will not be felt by the patient, even with ambulation.

Contraindications to the diaphragm include a history of toxic shock syndrome, current upper or lower genital tract infection or allergy to any of the components or to spermicide. Adequate fit may not be possible with a markedly

anteverted cervix. A shallow vaginal shelf or poor vaginal tone may lead to dislodgement. The diaphragm should be used with caution in the presence of a rectocele or cystocele.

Adverse effects from the diaphragm are minimal. There is an increased risk of recurrent urinary tract infection. The risk may be reduced with the softer rim or flat spring diaphragms. In addition, there is an increased risk of toxic shock syndrome. This can be minimized by avoiding use during menses or immediately postpartum and limiting use to 24 h at a time.

There are multiple types of diaphragms available in the United States. The ALL-FLEX Arcing Spring Diaphragm (Ortho-McNeil Pharmaceutical, Inc.) is a latex, buff-colored dome with a flexible rim. The rim contains a spring that forms an arc no matter where the rim is compressed. It is available in sizes 55–95 mm with 5 mm increments. It is the most widely used diaphragm and has been available in the United States since 1940. The firm rim of this diaphragm makes it the easiest type to insert. It may be the best choice for women with decreased pelvic tone, rectocele, or cystocele.

The Ortho Coil Spring Diaphragm, Ortho-McNeil Pharmaceutical, Inc, is a latex diaphragm with a flexible rim that contains a tension-adjusted, cadmium-plated coil spring. This allows for compression in one plane only. It is also available in sizes 55–95 mm with 5 mm increments.

The Milex Wide Seal is a silicone diaphragm with an arcing spring. A small skirt around the inner rim of the device is meant to hold spermicide in place and to improve the seal. It is available in eight sizes, with diameters of 60–95 mm, in 5 mm increments. In the United States, it is available only from the manufacturer.

The Reflexions Flat Spring diaphragm is made of latex with a rim similar to the coil spring. It is thinner and more delicate than the other diaphragms. It is available in seven sizes from 65- to 95 mm diameter with 5 mm increments.

The SILCS (pronounced "silks") intravaginal barrier is being developed by SILCS, Inc., Middlesex, N. J., in collaboration with the Contraceptive Research and Development Program (CONRAD) and The Program for Appropriate Technology in Health (PATH). One size will fit all women. It is a silicone device with a unique shape to allow for easier insertion and removal. The device is to be used with spermicide. It is currently recommended that it be left in for at least 8 h post-coitus. A maximum period of time per use has not yet been determined. The device is currently in phase II and III clinical trials. Once it is FDA approved it is intended to be made available over the counter.

TRANSDERMAL CONTRACEPTIVE PATCH

Ortho-Evra is a combination transdermal contraceptive patch containing 6.00 mg of norelgestromin and 0.75 mg of ethinyl estradiol (EE). Norelgestromin 150 μg and EE 20 μg are released into the bloodstream per 24 h of use. Norelgestromin is the primary active metabolite produced following oral administration of norgestimate. Ortho-Evra is available in cartons of one cycle (three patches) as well as in cartons containing a single patch to be used if a patch is lost or damaged.

The medication is administered via a triple layer patch with a contact surface area of 20 cm^2. The patch consists of the backing layer, an adhesive middle layer, and a release liner. The backing layer is made of flexible beige colored film with a polyethylene outer layer and a polyester inner layer. It has the appearance of a band-aid. It provides structural support and protects the adhesive middle layer. The adhesive layer contains polyisobutylene–polybutene adhesive, crospovidone, polyester, and lauryl lactate. The active ingredients, norelgestromin and EE, are in this layer. The release liner is a transparent polyethylene terephthalate film with a polydimethylsiloxane coating that protects the adhesive layer during storage and is removed prior to application of the device.

Following application of the device, both norelgestromin and EE rapidly appear in the blood stream. They reach a plateau level at ~48 h and then remain at steady-state levels for the remaining 5 days. Half-lives of norelgestromin and EE are 28 and 17 h, respectively. The FSH, LH, and Estradiol levels, suppressed during treatment, return to normal by 6 weeks after discontinuation. Release of norelgestromin has been found to be unaffected by exposure to saunas, whirlpools, exercise, and cold bath water. Release of EE is slightly increased by the sauna, whirlpool, or exercise, but levels remain within the desired range. Ortho-Evra bypasses the need for GI absorption. Therefore, it may be a good choice for women with some types of GI absorption disease.

In an open label study of 1672 reproductive age women using the patch for 10,994 cycles there were five pregnancies resulting from method failure and one resulting from user failure. This gives a PEARL index of 0.59 for method failure and 0.71 for user failure (F). A second study comparing Ortho-Evra to oral contraceptives found a PEARL index of 1.24 for Ortho-Evra versus 2.18 for the oral contraceptives (this difference was not statistically significant) (35). Ortho-Evra users also completed more cycles with perfect use compared to oral contraceptive users (88 and 78%, respectively) (35).

Studies have found that discontinuation of Ortho-Evra due to side effects is relatively low. A study by Smallwood et al. (32) found discontinuation rates of 1–2% each for side effects including skin irritation, nausea, emotional liability, headache, dysmenorrhea, and breast discomfort (35,36). Break through bleeding may be more common with Ortho-Evra, compared to oral contraceptives, during the initial cycles of use.

Ortho-Evra uses a 28-day cycle. A new patch is applied once a week for 3 weeks (21 days), followed by 7 days with no patch in place. Withdrawal bleeding occurs during the patch-free week. "Patch change day" should always be the same day of the week for a given woman. To start the patch, a woman should wait for her menses to begin. She may either place the patch on the first day of her menses or do a Sunday start. The day she applies her first patch will be her "patch change day". The patch must be applied to clean, dry, intact skin on the buttock, abdomen, upper outer arm, or upper torso (but not breasts). No lotions or other topical products should be in place on the skin where the patch will be placed.

To place the patch, it is first removed from its foil pouch. One-half of the release liner is then removed. This half of

the patch is applied to the skin and the second half of the release liner is removed. The woman should then press down on the entire patch with the palm of her hand for 10 s. She should then check to make sure that all of the edges are well adhered. The patch should then be checked every day to make sure it is sticking well.

Detachment of Ortho-Evra occurs rarely, with an occurrence of <3% at the time in one study (35). If a patch becomes partially or completely detached, it should be reapplied to the same spot immediately. If it will not reattach or it has become soiled, a new patch should be placed right away. No back-up contraception is needed and the "patch change day" remains the same. If more than 1 day has lapsed since detachment or if length of detachment is unknown, the current patch cycle should be discarded. The first patch from a new cycle should be placed. This will now be the new "patch change day" for this woman. In this case, a back-up method of contraception should be used for the first 7 days of the cycle. Other adhesives should not be used to hold a patch in place.

Contraindications and risks of Ortho-Evra are the same as those for the use of other combination hormonal methods. These are discussed in depth in multiple other references (see resources). In addition, Ortho-Evra is unique among contraceptives in causing significant rates of skin irritation, redness, or rash. Should skin problems arise, the patch may be removed and a new one applied in a different location until the next "change day". Patients with skin abnormalities, including eczema, psoriasis, or sunburn should not use Ortho-Evra. Ortho-Evra may have a lower efficacy in women weighing >198 lb (89.8 kg).

CONTRACEPTIVE VAGINAL RING

NuvaRing (Organon USA, Inc., Roseland, NJ, http://www.muvaring.com/Consummer/) is a nonbiodegradable, flexible vaginal ring containing 11.7 mg of etonogestrel and 2.7 mg of ethinyl estradiol (11). When placed in the vagina each ring releases $\sim$0.015 mg·day^{-1} of ethinyl estradiol and 0.120 mg·day^{-1} of etonogestrel per day. Etonogestrel is the biologically active metabolite of desogestrel. The NuvaRing is colorless and is composed of ethylene vinylacetate copolymers and magnesium stearate. The outer diameter is 54 mm. Each NuvaRing is packaged in a reusable aluminum laminate sachet. They are available in boxes of either one or three sachets.

NuvaRing is left in place for 21 days, followed by a 7 day hormone free period to allow for menstruation. Serum hormone levels required to suppress ovulation are achieved within the first 24 h of NuvaRing use, so there is no delay in the onset of contraception. Serum hormone levels are lower than those achieved by oral contraceptives and the contraceptive patch. The ring has a steady release rate, so serum hormone concentrations do not vary throughout the day as with oral contraceptives. Maximum serum drug concentrations are reached at 59 h for EE and at 200 h for etonogestrel. Half-lives are 29.3 and 44.7 h for etonogestrel and EE, respectively. If NuvaRing is broken it does not release a higher concentration of hormones, making overdose unlikely.

Contraindications and adverse reactions for NuvaRing are the same as those for combination oral contraceptives. In addition, local reactions specific to the ring can occur. These include vaginitis (5.6%), leukorrhea (4.6%), and vaginal discomfort (2.4%). One major clinical trial revealed a withdrawal rate of 15.1% due to problems such as the sensation of a foreign body, coital problems, and expulsion (37).

NuvaRing should be inserted between day 1 and 5 of the menstrual cycle. It should not be started later than day 5 even if bleeding is still occurring. Once inserted, NuvaRing should remain in place continuously for 21 days. The NuvaRing is inserted by folding it in half and inserting it high into the vagina. Exact positioning of the NuvaRing is not important for effectiveness.

NuvaRing is removed 3 weeks after insertion on the same day of the week that it was inserted. After removal, it should be placed in the foil pouch and discarded. After a 1 week break, a new ring is inserted. This will be on the same day of the week as the ring was inserted during the previous cycle. A withdrawal bleed will usually occur on day 2 or 3 after removal of the ring. The new ring must be inserted 7 days after removal of the previous ring to maintain contraceptive effectiveness. It should be inserted even if menstrual bleeding has not ended. A back-up method of contraception should be used until day 7 of use of the first NuvaRing.

If the NuvaRing is expelled or inadvertently removed, it should be washed in cool or warm (not hot) water, and reinserted within 3 h. If > 3 h have lapsed between expulsion and reinsertion a back-up method of contraception should be used until the NuvaRing has been in place for at least 7 continuous days. If the ring-free period has been 7 or more days, the possibility of pregnancy should be considered.

NuvaRing is meant to be in place for no >21 continuous days. If it is inadvertently left in place for up to 7 additional days it should be removed and left out for 1 week. A new ring should then be inserted. If NuvaRing has been left in place for > 4 weeks, then the risk of pregnancy should be considered. A new ring should be inserted, but a back-up method of contraception should be used until the new NuvaRing has been in place for 7 continuous days.

NuvaRing can be stored at room temperature (77 °F) for up to 4 months with excursions permitted from 59–86 °F. For longer storage it must be kept refrigerated at 36–46 °F.

CONTRACEPTIVE IMPLANTS

As of June 2005, there are currently no contraceptive implants available in the United States. Norplant (Wyeth, Madison, NJ, http://www.norplantinfo.com/) has been removed from the market, but Implanon (Organon USA, Inc., Roseland, NJ, http://www.organon.com/authfiles/index.asp) may soon become available in the United States. Implants function via the continuous release of a very low dose of progesterone. As with all progesterone only contraceptives, the implants inhibit ovulation via inhibition of the midcycle peaks of LH and FSH. In addition, progesterone only methods lead to thickening of the cervical mucous to

prevent penetration by sperm and to development of an atrophic endometrium to inhibit implantation.

Contraceptive implants are not subject to user error. Once in place, they are effective for their life expectancy or until removed. Use of Norplant in the United States has shown that it has a higher continuation rate than other contraceptive methods (I). Other benefits of contraceptive implants are similar to those of any progesterone only methods. Most notably, the implants do not contain estrogen, eliminating the risk of estrogen related complications and side effects. Progesterone only methods lead to amenorrhea or scanty periods for some women. The contraceptive effect of implants is rapidly reversed with removal.

Risks and side effects of implants are similar to those of all progesterone only methods of contraception. In addition, because implants release such a low dose of progesterone, drug interactions become a more significant concern. Also unique to implants is the risk for local skin problems. With Norplant one analysis found that 0.8% of users experienced local infection, 0.4% experienced expulsion of a capsule, and 4.7% had local skin irritation. The majority of these complications occur soon after insertion, but a significant number of women will also experience them after two or more months of use (38,39).

Norplant was an implant of six flexible capsules of silicon rubber tubing containing 26 mg of levonorgestrel. Each capsule was 2.4 mm in diameter and 34 mm in length. After insertion, Norplant initially, released 85 $\mu g \cdot day^{-1}$ of levornorgestrel. This decreased to 30 $\mu g \cdot day^{-1}$ over time. Norplant was very effective, however, removal was quite difficult in many cases.

In August of 2000, Wyeth recalled Norplant System Kits distributed beginning October, 1999 due to an atypically low level of levonorgestrel release in routine shelf-life stability tests. In July of 2002, Wyeth announced that it did not plan to reintroduce Norplant due to limitations in the availability of components of the product. Because the system is meant to be in place for a maximum of 5 years there should be very few women left with Norplant in place at this time. Ideally, all Norplant Systems will be removed by August of 2005.

Implanon is a single rod contraceptive implant containing 68 mg of etonogestrel, which is released gradually into the bloodstream for 3 years. The rod is nonbiodegradable and semirigid. It is 4 cm long and 2 mm in diameter and is made of ethylene vinylacetate copolymer. The etonogestrel is released at an average rate of 40 $\mu g \cdot day^{-1}$. In clinical trials Implanon took just 2.2 min to insert and 5.4 min to remove. Implanon can migrate, complicating removal (40). In the clinical trials no pregnancies occurred (41–43).

INTRAUTERINE DEVICES

Today's intrauterine devices (IUD) are highly effective, very convenient, and safe. However, the intrauterine device is used in fairly low rates in the United States, likely because of persisting misunderstandings on the part of many women and clinicians (44). Public opinion still reflects concern about the safety of intrauterine devices left over from the Dalkon Shield. The Dalkon Shield (A. H. Robins) was released in 1971 but episodes of septic abortions and other infections were reported, and the FDA recommended removal from the market in 1974. Litigation against A. H. Robins caused the company to declare bankruptcy in 1985. The cause of the infections was likely a multifilament string used only by the Dalkon Shield that allowed the product to be removed more easily. Bacteria could migrate up the string and cause the serious infections (45). In fact, clinical studies on current IUD models show excellent safety and effectiveness profiles.

The mechanism of action of IUDs is not completely understood. The two broad possibilities are preventing fertilization and destroying the early embryo, with prefertilization effects providing the contraception nearly all of the time (46–48). There are a number of proposed mechanisms of the IUD to prevent fertrilization of the ovum. The IUD distorts the hormonal and enzymatic environment of the female reproductive tract. It causes a chronic, sterile inflammatory environment of the endometrium. The copper found in IUDs is particularly toxic to sperm. Sperm are damaged by the environment, and are less likely to have the potential to fertilize the egg. The IUD also protects against ectopic pregnancy. One office visit is required for many years of contraception. Fertility returns quickly upon removal of the device. Women who are poor surgical candidates for sterilization are often excellent candidates for the intrauterine device.

Increasing evidence indicates that intrauterine devices protect against endometrial cancer. Of six studies examining the relationship between intrauterine devices and endometrial cancer, five found a protective effect, although only in two of the studies was this statistically significant (49,50). The mechanism of protection is unknown, but is thought to be related to effects of the IUD on the endometrium.

Initial cost of the intrauterine devices is high. In addition, there is a charge for insertion. However, this is a one time fee. Amortized over the life of the IUD it becomes one of the least expensive forms of contraception.

Pain and cramping may occur at the time of insertion of the IUD, but usually resolve within 15 min. Up to 10% of IUD users will spontaneously expel the IUD, sometimes without realizing it. Once a woman has expelled an IUD she has a 30% chance of expelling future IUDs as well (51). There is approximately a 1 in 1000 risk of uterine perforation associated with insertion of the IUD.

Early IUD research indicated significant increases in the risk of upper genital tract infection and tubal infertility associated with IUD use. It is now recognized that this research was flawed in several respects. Newer data and reanalysis of old data both indicate that risks of upper genital tract infection and tubal infertility are very low. There is a transient increase in the risk of upper genital tract infection for $\sim$20 days after the insertion procedure (it is the insertion, and not the IUD, that causes the increased risk). Afterward, the risk returns to low levels. It is not known whether upper genital tract infection or tubal infertility are more common in women with STI who receive an IUD versus women with STI and no IUD (52).

Antibiotic prophylaxis prior to intrauterine device insertion is not currently recommended. A large randomized controlled trial in Los Angeles County found no

benefit to the use of prophylactic azithromycin for IUD placement. The same study found that only about 1 woman in 100 developed salpingitis during the first months of IUD use (53).

The IUDs may be used by women who are nulliparous, but have had a previous spontaneous or elective abortion. Women who are nulligravid may have increased difficulty with IUD insertion and retention. However, The World Health Organization considers nulligravidity to be eligibility criteria category 2, implying that the benefits of the method generally outweigh any theoretical or proven risk.

Intrauterine device users should use condoms with new partners or whenever there is a risk of acquiring an STI. Should chlamydia or gonorrhea infection occur, there is no evidence to suggest that the IUD needs to be removed. Standard treatment of the STI is indicated.

Two types of IUDs are currently available in the United States. The Copper T 380A (ParaGard, FEI Women's Health LLC, New York) has been available since 1988. It is made of polyethylene in a T shape with barium sulfate to give X-ray visibility. Copper is wound around the polyethylene resulting in a copper surface area of $380 \pm 23\,mm^2$. The device is 36 mm tall and 32 mm wide with a 3 mm bulb at the bottom to which a monofilament polyethylene string is attached. The Copper T 380 is approved for use for up to 10 years. Studies have found that it remains highly effective for up to 12 years (54).

The copper on the CopperT 380 increases the presence of copper ions, enzymes, prostaglandins, and white blood cells in the uterine and tubal fluids. This impairs sperm function, which prevents fertilization. The device is extremely effective. In World Health Organization trials the cumulative 12-year pregnancy rate with the CopperT 380 was 2.2 pregnancies per 100 women. The highest risk time for pregnancy was during the initial year of use (54). A disadvantage of the CopperT 380 is an increase in menstrual blood loss.

The LNG-IUS (Mirena, Berlex Laboratories, Montville, NJ) was approved for use in the United States in 2000, although it has been available in Europe for > 10 years. It is a T-shaped device with a polyethylene frame and a cylinder composed of a polydimethylsiloxane-levonorgestrel mixture molded around the vertical arm. The cylinder is covered by a membrane to regulate release of the hormone. The device is 32 mm in height and 32 mm in width. There is a dark monofilament polyethylene thread at the base to assist with removal.

The LNG-IUS releases levonorgestrel directly into the endometrial cavity. The initial release rate is $20\,\mu g \cdot day^{-1}$, which diminishes to $14\,\mu g \cdot day^{-1}$ after 5 years. The system is approved for 5 years of use. Data shows that it actually remains effective for at least 7 years (55). Some levonorgestrel is absorbed, leading to a mean plasma concentration of 5%. This is lower than the plasma concentrations reached by other progesterone only methods.

Effectiveness rates for the LNG-IUS have been determined in several studies. These have found a first year cumulative failure rate of 0.14 per 100 women, a 5-year cumulative failure rate of 0.71 per 100 women and a 7-year cumulative failure rate of 1.1 per 100 women (55).

The LNG-IUS causes a significant decrease in menstrual blood loss. In addition, $\sim 20\%$ of users will experience complete cessation of menses. This effect is especially important in women for whom anemia is a concern. The LNG-IUS can be used to treat heavy menses, sometimes even in place of endometrial ablation or hysterectomy.

BILATERAL TUBAL STERILIZATION

Tubal sterilization may be performed postpartum, after spontaneous or elective abortion, or as an interval procedure. Interval sterilizations may be performed at any time during the menstrual cycle, however, current pregnancy must first be ruled out. Postpartum sterilization is performed by minilaparotomy prior to any involution of the uterus. Postabortion sterilization may be done via minilaparotomy or laparoscopically. Tubal sterilization performed as an interval procedure is done via laparoscope. Sterilization may also be done using a transcervical or transvaginal approach.

Laparoscopic sterilization can be performed on an outpatient basis. It leaves barely visible scars and return to normal activity is rapid. The cost and upkeep of laparoscopic equipment is expensive. Trocar insertion may result in injury to the bowel, bladder, or major blood vessels. General anesthesia is required.

Minilaparotomy requires a 2–3 cm incision placed in relation to the location of the uterine fundus. Only basic instruments and training are required. Minilaparotomy may be done under local anesthesia with sedation, with regional anesthesia, or with general anesthesia.

In November, 2002 the FDA approved the use of Essure (Conceptus Inc., San Carlos, CA, http://www.essure.com/consumer/c_homepage.aspes), a sterilization device that is placed using hysteroscopy through a transcervical approach. No entry into the peritoneal cavity is required. Back-up contraception is then used for 3 months followed by a hysterosalpingogram to confirm bilateral tubal occlusion. Data on this method is promising, but long-term efficacy rates are not yet available.

The transvaginal approach to sterilization is rarely used. It is contraindicated in the presence of major pelvic adhesions or enlarged uterus. There are many potential complications including cellulites, pelvic abscess, hemorrhage, proctotomy, or cystotomy.

Methods of tubal occlusion include electrocoagulation, mechanical occlusion, ligation, and chemical sclerosing. Bipolar electrocoagulation is performed via the laparoscopic approach only. At least 3 cm of the isthmic fallopian tube is coagulated using at least 25 W delivered in a cutting waveform. A current meter most accurately indicates complete coagulation.

Ligation is the most common method of occlusion used during laparotomy or minilaparotomy. There are multiple methods for ligation including the Pomeroy, modified Pomeroy, Parkland, Uchida, and Irving methods. Chemical sclerosing agents are under investigation but none are approved for use in the United States (56).

Several mechanical occlusion devices are available in the United States. Because less of the fallopian tube is destroyed, microsurgical reversal is more likely to succeed

following these methods. Mechanical occlusion requires a normal fallopian tube. If there are adhesions, thickening, or dilation of the tube proper application is more difficult and failure is more likely. Mechanical occlusion devices include the Falope ring, a silicon rubber band, the Hulka-Clemens clip, a spring-loaded clip, and the Filshie clip, a titanium clip lined with silicone rubber. Each device has its own type of applicator and each requires special training to use.

Efficacy for tubal sterilization is high. The CREST study found a 5-year cumulative life-table probability of pregnancy to be 13 per 1000 procedures for all types of tubal ligation in aggregate. Rates for each method include 5-year cumulative pregnancy rates of 6.3 per 1000 procedures for postpartum partial salpingectomy, 16.5 per 1000 for bipolar coagulation, 10 per 1000 for silicone band methods, and 31.7 per 1000 for spring clips. The risk of sterilization failure persists for many years. Therefore, 10-year cumulative pregnancy risk is >5 year cumulative risk for all methods. Women who are younger at the time of sterilization are more likely to experience method failure because of their greater fecundity (56).

When tubal sterilization failure occurs, the risk of ectopic pregnancy is significant. For the women in the CREST study, one-third of poststerilization pregnancies were ectopic (57). Other risks of tubal sterilization are rare and are mainly related to the need for general anesthesia.

CONTRACEPTION RESEARCH AND DEVELOPMENT

In 2004, the National Academy of Sciences found that "while the existing array of contraceptive options represents a major contribution of science and industry to human well-being, it fails to meet needs in significant populations and the costs of that failure are high, for societies, for families, and for individuals" (6). Dozens of contraceptive devices utilizing a wide array of mechanisms of action are currently under development. At least 14 new contraceptive methods have been approved by the FDA since 1998. Nevertheless, research and development of new contraceptive methods is difficult. The Institute of Medicine Committee on New Frontiers in Contraception estimates that development of a new contraceptive takes 10–14 years and $400–800 million. Large companies are reluctant to assume the risk involved in long-term development of devices that potentially will not be approved or will result in minimal income. Small companies lack sufficient resources for such large undertakings. The success of development of any given contraceptive is unpredictable and economically risky. The risk of liability is also a strong disincentive to contraceptive research and development. Therefore, as other fields of medicine make research advances in genetics, molecular biology, and immunology, contraceptive technology has been unable to keep up. Given the overwhelming effect of unintended pregnancy on people's lives, and the very high rates of unintended pregnancy in the United States, such low resources for contraceptive development must be taken seriously (58).

Future success in the research and development of contraceptives may be improved by collaborations between small and large industry and not-for-profit organizations. To this end, CONRAD supported the creation of the Consortium for Industrial Collaboration in Contraceptive Research (CICCR). The CICCR identifies potential project leads under investigation by not for profit organizations and encourages industry to collaborate with these not for profits. In addition, they provide additional funds to not-for-profit investigators.

Research and development has traditionally focused on creating contraceptives with high rates of efficacy intrinsic to the device and better safety profiles. It is now recognized that it is also important to design contraceptives that are easier to use consistently and correctly. More consumer-based research will help to determine what types of contraceptive development would be most desirable to women of different ages, races, or life styles. In addition, the need for more dual protection methods is great. Other than condoms, there is no single method that provides high rates of efficacy for both pregnancy and STI prevention. Current understanding of how existing methods affect STI transmission, including effects on the immune system, needs to be improved (59). "Dual packaging" of contraceptives is being investigated. For example, oral contraceptives could be packaged together with emergency contraceptives or condoms.

Many microbicides are currently under varying stages of development. A perfect microbicide would destroy STI pathogens and prevent pregnancy, without causing vaginal tissue damage. Spermicides could inactivate HIV or other pathogens, interfere with cell attachment or entry, or prevent viral replication. Pregnancy prevention could be accomplished via effects on sperm motility.

There are several vaginal barrier contraceptives under development. The Today Sponge has been reintroduced in Canada and on the internet and is awaiting FDA approval. Protectaid, containing nonoxynol-9 and benzalkonium chloride, and Pharmatex, containing benzalkonium chloride alone, are both contraceptive sponges that are available in Canada and Europe and may become FDA approved. Oves is a disposable silicone cervical cap. Research is also underway on self-fitting diaphragms and new types of female condoms as well.

Jadelle is another hormonal implants that is FDA approved, but not available in the U.S. market. Jadelle is an improved two-rod version of Norplant. Hormone receptor blocking agents, such as mifepristone, are being investigated for emergency contraception, low daily dosing, or monthly dosing. Research is being done on non-hormonal systemic methods for women. For example, a contraceptive could specifically inhibit implantation by blocking leukemia inhibitory factor, perimplantation factor, or leptin.

Immunocontraceptives offer a new and innovative approach to contraception as well as STI prevention. Possible areas of immunologic research include pursuit of immunogens to sperm, reproductive hormones, hormone receptors, and sexually transmitted pathogens, as well as research on the local immune response of the female reproductive tract (6). Immunocontraceptives under research include a vaccine to HCG and to sperm antigens. At this time, the HCG vaccine requires frequent boosters and has a potential cross-reaction with pituitary hormones. Boosting

mucosal immunity to sperm may be possible via oral or vaginal vaccine administration.

Research is underway on several systemic methods for males. Efforts have been targeting the production of LH and FSH by the pituitary. Lack of LH and FSH blocks hormonal support for testicular cell function. This results in decreased sperm production, but also decreases testosterone. Low testosterone levels lead to a lack of libido. Human studies are in progress to look at drug combinations to suppress LH while replacing testosterone losses. Several forms of testosterone are being examined, but many require daily or weekly injections. Because high levels of testosterone cause multiple side effects, it is challenging to find the perfect balance.

Other systemic male methods are also under investigation. Gossypol, a derivative of cottonseed oil, is effective in sperm suppression while not effecting testosterone levels. However, in current protocols it is up to 20% irreversible. Some chemotherapy drugs, including lonidamine, are also being considered for male contraceptives.

BIBLIOGRAPHY

1. Facts in Brief: Contraceptive Use. 2004. Available at http://www.agi-usa.org/pubs/fb_contr_use.html.
2. Mosher WD, et al. Use of contraception and use of family planning services in the United States: 1982–2002. Adv Data 2004;350:1–36.
3. Beck LF, et al. Prevalence of selected maternal behaviors and experiences, pregnancy risk assessment monitoring system (PRAMS), 1999. MMWR 2002;51(SS02):1–26.
4. Burnhill MS, Contraceptive use: the US perspective. Int J Gynaecol Obstet 1998;62(Suppl 1):S17–S23.
5. Sharing Responsibilities: Women, Society and Abortion Worldwide. 1991. Available at http://www.agi-usa.org/pubs/sharing. pdf.
6. Harrison PF, Rosenfield A, editors, Contraceptive Research and Development: Looking to the Future. Washington (DC): National Academy Press; 2004. p 1–28.
7. The Unfinished Revolution in Contraception: Convenience, Consumer Access and Choice. 2004. Available at http://www.guttmacher.org/pubs/2004/09/20/UnfinRevInContra.pdf.
8. Haffner DW, Styton WR. Sexuality and Reproductive Health. In: Hatcher RA et al. editors. Contraceptive Technology 18th Revised Edition. New York: Ardent Media, Inc.; 2004. p 20.
9. Guest F. HIV/AIDS and Reproductive Health, in Contraceptive Technology 18th Revised Edition. In: Hatcher RA, et al. editors. New York: Ardent Media, Inc.; 2004. p 160.
10. Hatcher RA, et al. Contraceptive Technology, 18th Revised Edition. New York: Ardent Media, Inc.; 2004.
11. Trussell J. The essentials of contraception: efficacy, safety, and personal considerations. In: Hatcher RA, et al. editors. Contraceptive Technology 18th Revised Edition. New York: Ardent Media, Inc.; 2004. p 230.
12. Steiner MJ, et al. Influence of cycle variability and coital frequency on the risk of pregnancy. Contraception 1999; 60(3): 137–143.
13. Trussell J. Contraceptive efficacy. In: Hatcher RA, et al. editors. Contraceptive Technology 18th Revised Edition. New York: Ardent Media, Inc.; 2004. p 774.
14. Guest F. Education and counseling. In: Hatcher RA, et al. editors. Contraceptive Technology 18th Revised Edition. New York: Ardent Media, Inc.; 2004. p 264.
15. Sterilization of persons in federally assisted family planning projects. Fed Reg 1978;43:52146–52175.
16. Hatcher RA, et al. editors. Contraceptive Technology, 17th revised edition, New York: Ardent Media, Inc.; 1998. p 326.
17. Warner L, Hatcher RA, Steiner MJ. Male condoms. contraceptive technology 18th Revised ed. In: Hatcher RA, et al. editors. New York: Ardent Media, Inc.; 2004. p 334.
18. Hatcher RA, et al. editors. Contraceptive Technology, 17th Revised Edition. New York: Ardent Media, Inc.; 1998. p 329.
19. de Vincenzi I. A longitudinal study of human immunodeficiency virus transmission by heterosexual partners. European study group on heterosexual transmission of HIV. N Engl J Med 1994;331(6):341–346.
20. Latex Condoms and Sexually Transmitted Diseases- Prevention Messages. 2001. Available at http://www.metrokc.gov/health/apu/std/condomefficacy.htm.
21. Workshop Summary: Scientific Evidence on Condom Effectiveness for Sexually Transmitted Disease (STD) Prevention. 2001 July 20, 2001 Available at http://www.niaid.nih.gov/dmid/stds/condomreport.pdf.
22. Warner L, Hatcher RA, Steiner MJ. Male condoms. In: Hatcher RA, et al. editors. Contraceptive Technology 18th Revised Edition. New York: Ardent Media, Inc.; 2004. p 346.
23. Warner L, Hatcher RA, Steiner MJ. Male condoms. In: Hatcher RA, et al. editors. Contraceptive Technology 18th Revised Edition. New York: Ardent Media, Inc.; 2004. p 348.
24. Drew WL, Blair M, Miner RC, Conant M. Evaluation of the virus permeability of a new condom for women. Sex Transm Dis 1990;17(2):110–112.
25. The Female Condom: A Review. 1997, World Health Organization.
26. Macaluso M, et al. Efficacy of the female condom as a barrier to semen during intercourse. Am J Epidemiol 2003;157(4): 289–297.
27. Farr G, Gabelnick H, Sturgen K, Dorflinger L. Contraceptive efficacy and acceptability of the female condom. Am J Public Health 1994;84(12):1960–1964.
28. Cates WJ, Raymond EG. Vaginal spermicides, In: Hatcher RA, et al. editors. Contraceptive Technology 18th Revised Edition. New York: Ardent Media, Inc.; 2004. p 356.
29. Available at http://www.cdc.gov/hiv/pubs/mmwr/mmwr11aug 00. htm.
30. Mauck C, et al. Lea's Shield: a study of the safety and efficacy of a new vaginal barrier contraceptive used with and without spermicide. Contraception 1996;53(6):329–335.
31. McClure DA, Edelman DA. Worldwide method effectiveness of the Today vaginal contraceptive sponge. Adv Contracept 1985;1(4):305–311.
32. Edelman DA, North BB. Updated pregnancy rates for the Today contraceptive sponge. Am J Obstet Gynecol 1987; 157(5):1164–1165.
33. Edelman DA, McIntyre SL, Harper J. A comparative trial of the Today contraceptive sponge and diaphragm. Am J Obstet Gynecol 1984;150(7):869–876.
34. Trussell J, Strickler J, Vaughan B. Contraceptive efficacy of the diaphragm, the sponge and the cervical cap. Fam Plann Perspect 1993;25(3):100–105,135.
35. Audet MC, et al. Evaluation of contraceptive efficacy and cycle control of a transdermal contraceptive patch vs an oral contraceptive: a randomized controlled trial. JAMA 2001;285(18): 2347–2354.
36. Smallwood GH, et al. Efficacy and safety of a transdermal contraceptive system. Obstet Gynecol 2001;98(5 Pt. 1):799–805.
37. Hatcher RA, Nelson A. Combined hormonal contraceptive methods. In: Hatcher RA, et al. editors. Contraceptive Technology 18th Revised Edition. New York: Ardent Media, Inc.; 2004. p 449.
38. Klavon SL, Grubb GS. Insertion site complications during the first year of NORPLANT use. Contraception 1990;41(1):27–37.

39. Hatcher RA. Depo-Provera injections, implants, and progestin-only pills (minipills). In: Hatcher RA, et al. editors. Contraceptive Technology. New York: Ardent Media, Inc.; 2004. p 471.
40. Le J, Tsourounis C. Implanon: a critical review. Ann Pharmacother 2001;35(3):329–336.
41. Kiriwat O, et al. A 4-year pilot study on the efficacy and safety of Implanon, a single-rod hormonal contraceptive implant, in healthy women in Thailand. Eur J Contracept Reprod Health Care 1998;3(2):85–91.
42. Croxatto HB. Clinical profile of Implanon: a single-rod etonogestrel contraceptive implant. Eur J Contracept Reprod Health Care 2000;5(Suppl. 2):21–28.
43. Croxatto HB, Makarainen L. The pharmacodynamics and efficacy of Implanon. An overview of the data. Contraception 1998;58(6 Suppl.):91S–97S.
44. Espey E, Ogburn T. Perpetuating negative attitudes about the intrauterine device: textbooks lag behind the evidence. Contraception 2002;65(6):389–395.
45. Cheng D. The intrauterine device: still misunderstood after all these years. South Med J 2000;93(9) 859–864.
46. Ortiz ME, Croxatto HB, Bardin CW. Mechanisms of action of intrauterine devices. Obstet Gynecol Surv 1996;51(12 Suppl.): S42–S51.
47. Stanford JB, Mikolajczyk RT. Mechanisms of action of intrauterine devices: update and estimation of postfertilization effects. Am J Obstet Gynecol 2002;187(6):1699–1708.
48. Rivera R, Yacobson I, Grimes D. The mechanisms of action of hormonal contraceptives and intrauterine contraceptive devices. Am J Obstet Gynecol 1999;181(5 Pt. 1):1263–1269.
49. Grimes DA. Intrauterine devices (IUDs). In: Hatcher RA, et al. editors. Contraceptive Technology 18th Revised Edition. New York: Ardent Media, Inc.; 2004. p 499.
50. Hubacher D, Grimes DA. Noncontraceptive health benefits of intrauterine devices: a systematic review. Obstet Gynecol Surv 2002;57(2):120–128.
51. Bahamondes L, et al. Performance of copper intrauterine devices when inserted after an expulsion. Hum Reprod 1995;10(11):2917–2918.
52. Grimes DA. Intrauterine devices (IUDs), In: Hatcher RA, et al. editors. Contraceptive Technology 18th Revised Edition. New York: Ardent Media, Inc.; 2004. p 501.
53. Walsh T, et al. Randomised controlled trial of prophylactic antibiotics before insertion of intrauterine devices. IUD Study Group. Lancet 1998;351(9108):1005–1008.
54. Long-term reversible contraception. Twelve years of experience with the TCu380A and TCu220C. Contraception 1997; 56(6):341–352.
55. Sivin I, et al. Prolonged intrauterine contraception: a seven-year randomized study of the levonorgestrel 20 mcg/day (LNg 20) and the Copper T380 Ag IUDS. Contraception 1991; 44(5): 473–480.
56. Peterson HB, et al. The risk of pregnancy after tubal sterilization: findings from the U.S. Collaborative Review of Sterilization. Am J Obstet Gynecol 1996;174(4):1161–1168; discussion 1168–1170.
57. Peterson HB, et al. The risk of ectopic pregnancy after tubal sterilization. U.S. Collaborative Review of Sterilization Working Group. N Engl J Med 1997;336(11):762–767.
58. Stewart F, Gabelnick HL. Contraceptive research and development, In: Hatcher RA, et al. editors. Contraceptive Technology 18th Revised Edition. New York: Ardent Media, Inc.; 2004. p 606.
59. Stewart F, Gabelnick HL. Contraceptive research and development, In: Hatcher RA, et al. editors. Contraceptive Technology 18th Revised Edition. New York: Ardent Media, Inc.; 2004. p 603.

See also COLPOSCOPY; SEXUAL INSTRUMENTATION.

CORONARY ANGIOPLASTY AND GUIDEWIRE DIAGNOSTICS

RUPAK K. BANERJEE
ABHIJIT SINHA ROY
University of Cincinnati
Cincinnati, Ohio

LLOYD H. BACK
California Institute of Technology
Pasadena, California

INTRODUCTION

Percutaneous Transluminal Coronary Angioplasty (PTCA) is an invasive procedure, where a blocked coronary artery is opened by inserting a pressurized balloon. Since its inception in the year 1964 by Dotter and Judkins (1), coronary angioplasty has undergone much development and is commonly performed in Cardiac Catheterization today. A statistic provided by the Center for Disease Control (CDC) shows that nearly one-half of a million PTCA procedures were conducted in United States alone in the year 2002. A typical coronary angioplasty procedure includes these basic components:

Guiding Catheter

A guiding catheter serves three broad purposes: It provides support and passage to the introduction of smaller diameter guidewires. It provides a conduit to the administration of drugs and external agents, such as contrast agent for angiography. It provides damping, due to heart motion, to guidewires inserted through them.

For passage of guidewires, a catheter, having a diameter at least twice that of the guidewire, is recommended. Most guidewires are made up of soft material in the tip so that instance of vessel injury is as minimal as possible. In modern practice, guiding catheter are available in different shapes and sizes, such as Judkins, Amplatz curves, pig tail. Selection of a suitable guiding catheter depends on the application. Small sized guiding catheters of size 6F and 7F are most commonly used in PTCA as their size is well suited for guidewires of size 0.014 in. (0.355 mm) and for passage of balloon catheters, appropriate for coronary dimensions in humans. Some guiding catheters may be designed with side holes at the end, which enables them to engage the coronary ostium, while maintaining continuous administration of fluoroscopy agent. The length of guiding catheters is $\sim$ 90–100 cm. Figure 1 shows some of the guiding catheters being used today in coronary angioplasty.

Guidewires

Modern day guidewires are designed for tip stiffness, easy maneuverability, location control and visibility to angiography. A typical guidewire has a solid core, usually made up of stainless steel or nitinol and has a gradual taper from the proximal to distal end. This core is encapsulated in a spring coil and platinum in the distal section for improved radiographic visibility. The spring coil, which is a teflon coated stainless steel, is usually welded to directly or through a band to the tapered end of a guidewire so that

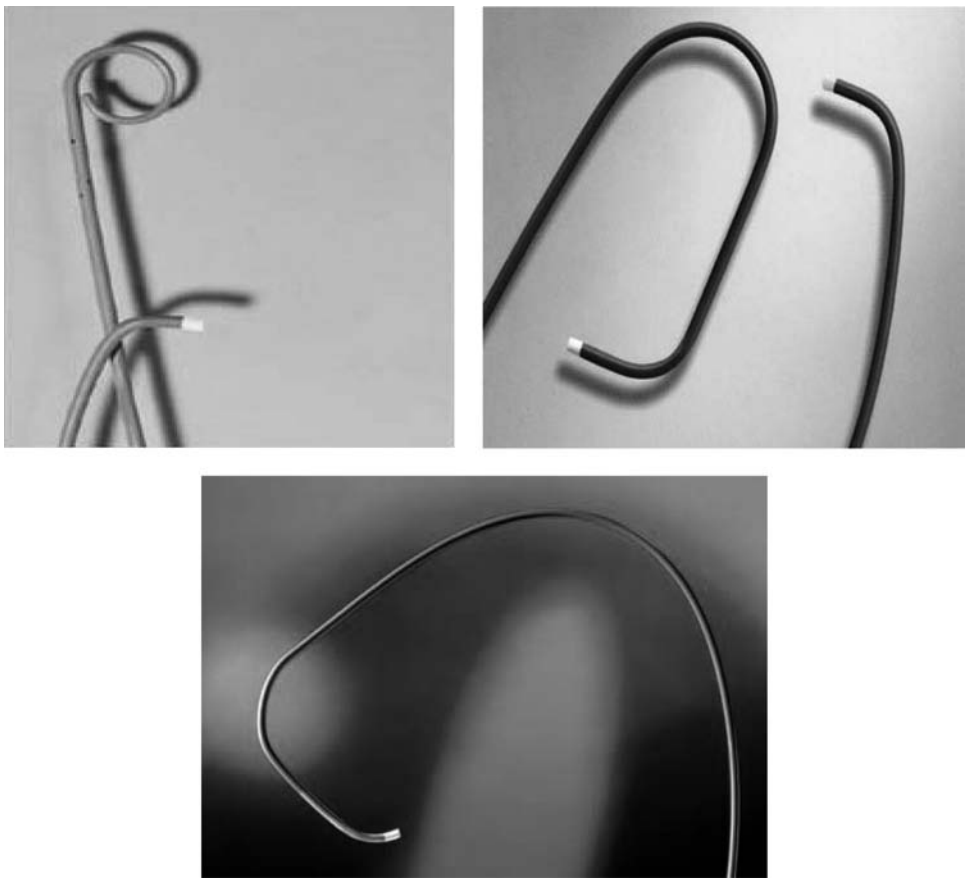


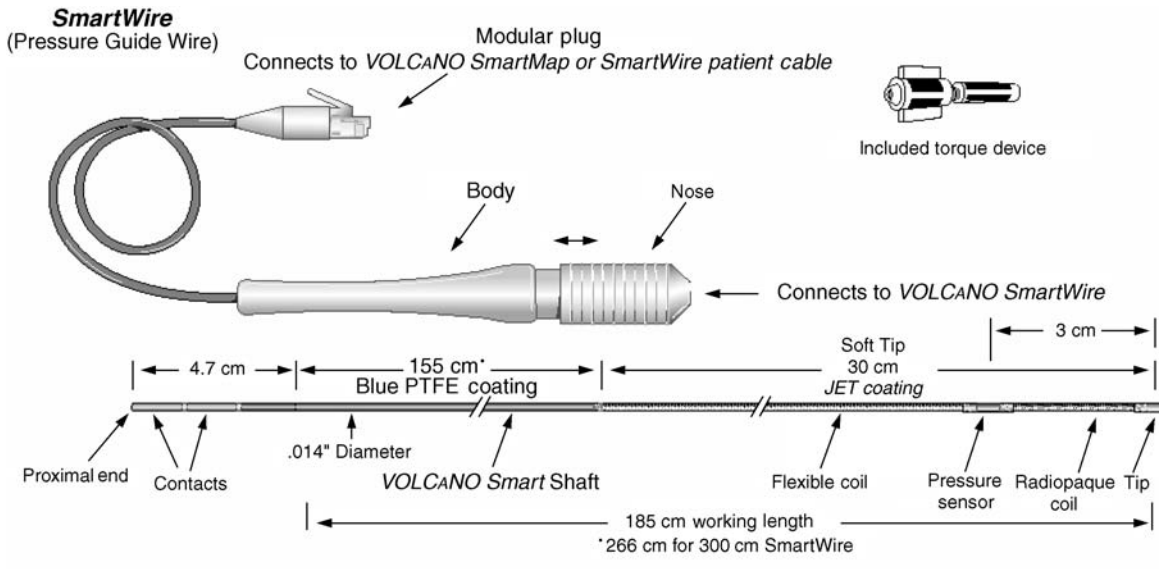
Figure 1. A few coronary angioplasty guiding catheters. (Courtesy of Boston Scien., MA.)

the user may bend the tip to access the desired artery. Guidewires are available in a wide range of sizes from 0.010 in. (0.254 to 0.457 mm) to 0.018 in. In modern day PTCA, 0.014 in. (0.355 mm) is the size most widely used. In some cases, double length (∼300 cm) guidewires may also be used. These enable access to the diseased vessel while other devices, such as stent, dilation catheters, are being deployed, with minimal risk of vessel injury (2).

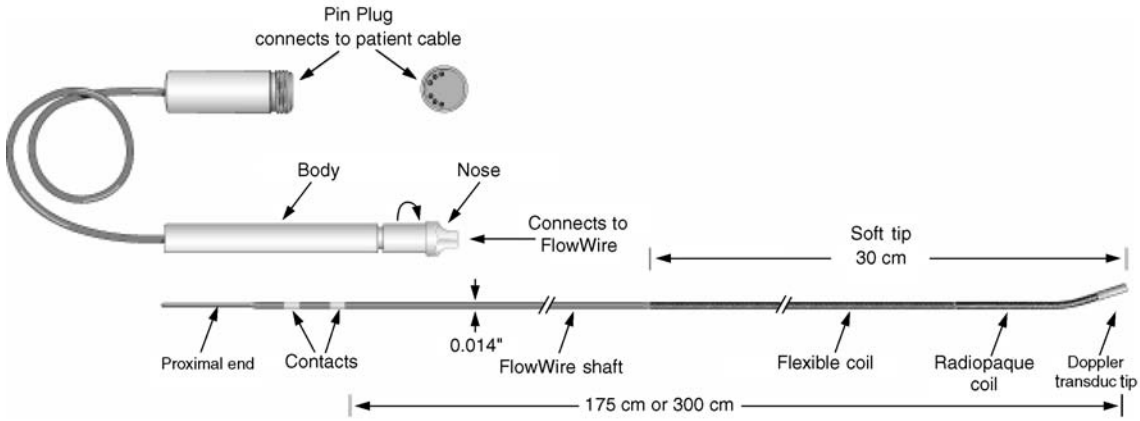
Several specialized guidewires are also available. These guidewires are designed for measurement of arterial pressure and flow to assess the ischemic severity of a stenosis. For pressure measurement, a piezoelectric pressure transducer is placed around the inner solid core of the 0.014 in. (0.355 mm) 3 cm from the tip (Fig. 2). This transducer facilitates measurement of transstenotic pressure drop and Myocardial Fractional Flow Reserve (FFR_{myo}) (3). These pressure sensors can measure pressure in the range of −30–300 mmHg (−3.9–39.9 kPa) with an accuracy of ±1 mmHg (0.133 kPa). For phasic flow measurement, the technology most widely used is a Doppler-based flow sensor. The Doppler flow sensor is placed at the tip of the 0.014 in. (0.355 mm) guidewire (∼175 cm long). For coronary flow measurements, usually 40 MHz piezoelectric transducer is used (Fig. 2). As a general rule, the smaller the vessel size, the larger is the frequency of the sensor. The ultrasound beam describes a conical beam in the distal vessel, thus obtaining a small sample volume. Doppler-based guidewire are capable of measuring translesional velocity, which is reported as Coronary Flow Reserve (CFR = coronary flow at hyperemia / coronary flow at basal flow). Currently, however, Doppler guidewires are designed to measure average peak velocity, mean velocity for a cycle as well as diastolic/systolic flow ratio. A unit that measures both flow and pressure simultaneously in coronary vessels is shown in Fig. 3. Another technique used to measure CFR is based on coronary thermodilution. In this method, the wire has a microsensor at a location 3 cm from the floppy tip, which enables simultaneous recording of coronary pressure measurement and temperature, with an accuracy of 0.02 °C (4,5). The shaft of this wire can be used as a second thermistor, which provides the input signal at the coronary ostium of any fluid injection at a temperature different from blood. With this method, CFR is expressed as the ratio of mean transit time at basal flow to mean transit time at hyperemic flow. Experimentally, it has been shown that $CFR_{doppler}$ and CFR_{thermo} differ by ∼20% (5). To facilitate simultaneous evaluation of epicardial and microvascular diseases, a single wire, having both pressure and flow sensor, is also available.

Balloon Catheter

Appropriate selection of a balloon catheter is a must for the success of coronary angioplasty. Beginning from the "over the wire" design, balloons have undergone many improvements. Present balloon catheters are both strong and flexible enough to handle tortuous vessel segment, with minimal intimal injury. Most balloon catheters have a silicone or hydrophilic coating, such as polyethylene, to



A guidewire for measuring pressure.



A guidewire for measuring flow.

Figure 2. Guidewires for measuring pressure and flow in diseased coronary arteries. (Courtesy of Volcano Therapeutics Inc., CA.)

reduce friction. To dilate a balloon catheter, pressures upto 20 atm can be used. Figure 4a and b show a balloon catheter, having an *over the wire* design, with and without a stent. To generate the pressure, an indeflator is used. It consists of a cylinder, one end of which has a movable plunger and the other end is connected to the dilation catheter (Fig. 4c). The indeflator is partially filled with liquid and has an attached pressure gauge. Dilation catheters have a wide range of inflation diameters ranging from 1.5 to 4 mm depending on the artery dimension. Length of balloons varies from 10 to as much as 40 mm depending on the length of the atherosclerotic lesion. Materials used for manufacturing balloons can be polyethylene, polyolyfin, or nylon to name some, with the wall thickness varying from 0.0003 to 0.0005 in. (0.007 to 0.0127 mm). Additionally, all balloon catheters are sold with a *rated burst pressure*.

Besides the traditional design in which the catheter passes over the entire length of the guidewire, a design known as *monorail* is one in which the catheter passes just on its tip such that quick removal and insertion can be done. Some balloon catheters (*perfusion balloons*) are equipped with side holes at their tips in the shaft proximal and distal to the balloon to allow the blood to flow from the proximal to the distal vessel, when the balloon is inflated. This reduces the risk of ischemia in the heart. Nowadays, PTCA is usually combined with implantation of a stent in the clogged artery to keep the artery open, after balloon removal and reduce the risk of restenosis. The path of approach of catheters and guidewires into the coronaries is usually done via the femoral artery and vein with direction guidance being provided by fluoroscopy. Other paths of approarch can be done via the axillary, brachial, or radial artery. For insertion of balloon as well as normal catheters and guidewires,

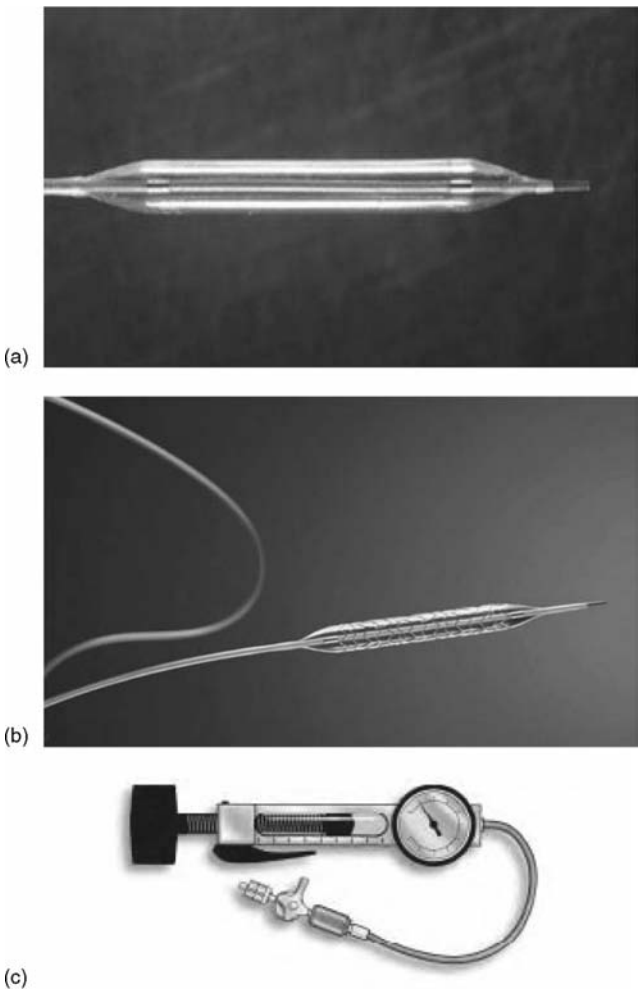


Figure 3. (a) An example of coronary dilation catheter; (b) with stent over it. This is an "over the wire design". (Courtesy of Boston Scientific, MA); (c) An indeflator, which is used to inflate the balloon. (Courtesy of Guidant Corp., IN.)

percutaneous needles, such as seldinger needle for femoral artery, Potts–Cournand needle (which is hollow from inside so that the user can know when the artery has been punctured) and, vascular sheaths are used. In case there is difficulty in detecting the arterial or venous pulse, a smart needle (PSG, Mountain View, CA) may be used, which has a Doppler crystal to direct the needle to the center of the vessel.

Vessel Closure Devices

On completion of PTCA, the punctured vessel needs to be closed to prevent any postprocedural bleeding followed by coagulation. The most commonly used device is a collagen plug applied to skin outside the outer wall of the vessel (6). Another device, called the Hemostatic Puncture Closure Device, also known as Angio-Seal Vascular Closure device, (Kensey Nash, Exton, PA), uses an anchor on the inner wall of the vessel and uses an attached suture to raise a collagen plug to the outer wall of the vessel (7). Figure 5 shows the

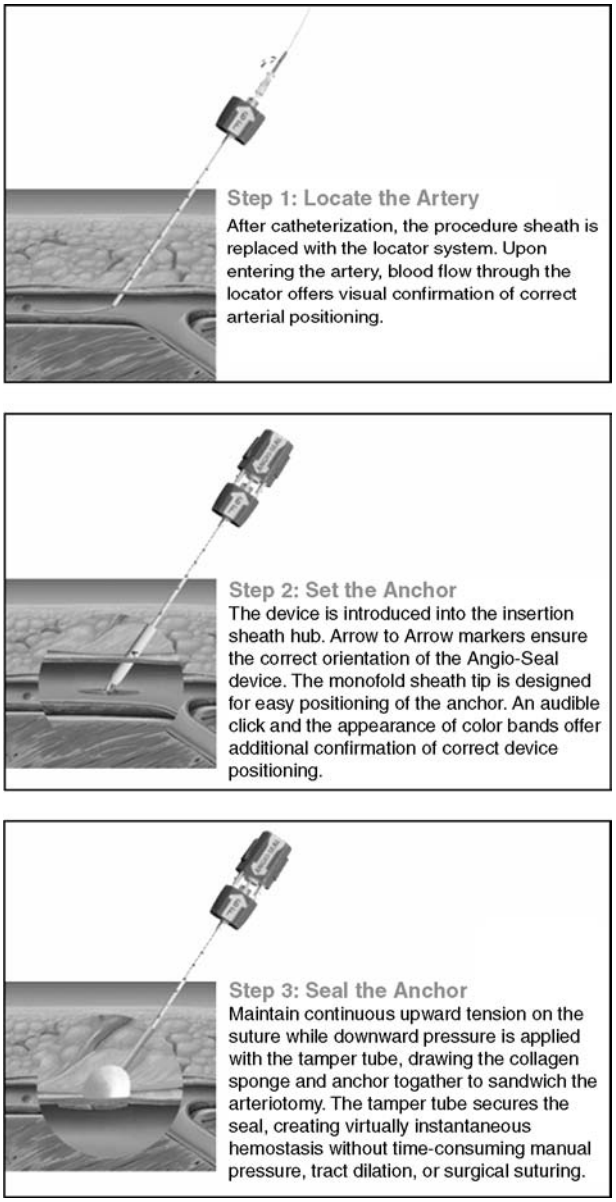


Figure 4. Vascular closure device, known as Angio-Seal. (Courtesy of Kensey Nash., PA.)

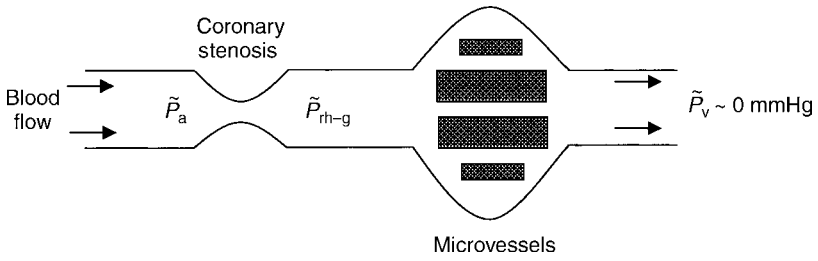


Figure 5. Measurement of myocordial fractional flow reserve is a coronary stenosis.

Table 1. Lesion Geometry and Hemodynamics Before, Intermediate, and After PTCA[a]

Lesion	A_m, mm^2	d_m, mm	% Area Stenosis	l_m, mm	CFR	$\tilde{p}_a$ mmHg	$\Delta\tilde{p}_h$ mmHg	$\tilde{p}_{rh}$ mmHg	FFR_{myo}
Before-PTCA	0.7 ± 0.1	0.95	90	0.75	2.3 ± 0.1	89 ± 3	34	55	0.62
Intermediate	1.43	1.35	80	0.75	3.3	86	14.3	70.4	0.82
After-PTCA	2.5 ± 0.1	1.80	64	3	3.6 ± 0.3	84 ± 3	7.4	75.2	0.89

[a]The parameters A_m, d_m, and l_m are the area, diameter, and length of the narrowest region of the stenoses; $\tilde{p}_a$ and $\tilde{p}_{rh}$ are the mean arterial pressure in the coronary ostium and distal to the stenosis under hyperemia; $\Delta\tilde{p}_h$ is the hyperemic pressure drop across the stenosis.

procedure followed to close the vessel using Angio-Seal. Another device, called the Prostar device (Perclose, Redwood City, CA) uses a sheath-like device to pass a suture around the puncture through the skin to close the puncture.

DIAGNOSTICS WITH GUIDEWIRES

With widespread use of PTCA, there has been a surge in the use of guidewires for evaluation of ischemic severity of focal and diffuse lesions before and after angioplasty. Since their inception, several authors (10–13) have validated the usefulness of guidewires in clinical procures, including PTCA. However, during coronary intervention, introduction of guidewire itself produces an additional resistance to blood flow that has not been well documented.

Issues

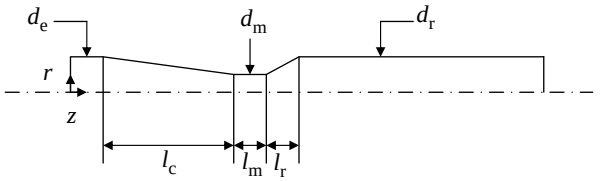
Current usage of guidewires for coronary stenoses diagnostics is for measurement of FFR_{myog} and CFR_g. By definition (4,12), FFR_{myog} is the ratio of mean distal pressure ($\tilde{p}_{rhg}$) to mean pressure proximal to the stenosis ($\tilde{p}_a$) at hyperemia, which is induced by administration of vasodilator agents (e.g., adenosine). Since pressure drop in normal epicardial vessels is very small, $\sim\tilde{p}_a$ mean aortic pressure. Likewise, CFR_g is the ratio of mean coronary flow at hyperemia ($\tilde{Q}_{hg}$) to mean coronary flow at basal (i.e., rest) ($\tilde{Q}_b$) (13).

A value of $FFR_{myog} = 0.75$ is assumed to accurately discriminate stenosis whether or not associated with inducible ischemia (4,12). Both FFR_{myog} and CFR_g increase after coronary angioplasty, thereby signifying an enhanced and normal blood supply to distal myocardium. However, in the presence of microvascular disease, FFR_{myog} and CFR_g measured alone cannot dissociate an epicardial coronary lesion from distal microvascular disease (12,14). To address the nonuniformity of microvascular circulation, application of Relative Coronary Flow Reserve, $rCFR_g$ (13,15) was proposed. However, in patients in whom a stenotic artery supplies an area of myocardial infarction, neither CFR_g nor $rCFR_g$ can differentiate flow impairment due solely to a stenosis.

Currently, guidewires of size 0.014 in. (0.355 mm) are capable of measuring both flow (CFR_g) and mean pressure drop, $\Delta\tilde{p}$ (and FFR_{myog}) across a stenosis. However, the introduction of a guidewire causes an obstructive effect, creating an "artifactual" stenosis (16–19). The threshold limit of $FFR_{myog} = 0.75$ is a measured value with guidewire. However, this measured value of 0.75 must be attributed to FFR_{myog} and not FFR_{myo}, the value for the lesion without guidewire obstruction. Limited information is available on what degree of flow blockage exists with currently used guidewires, although clinical investigators have acknowledged the limitations of mean pressure drop and flow measurements because of flow obstruction produced by guidewires (20,21). In the following sections, the authors present a summary of their past studies on, (1) quantifying the flow obstruction effect of guidewires of diameter 0.014 in. (0.35 mm) and 0.018 in. (0.46 mm), which results in enhanced $\Delta\tilde{p}_h$ and reduced $\tilde{Q}_h$ in a significant, intermediate and moderate focal stenoses; and (2) corrections to be applied to FFR_{myog} and CFR_g to get true values of FFR_{myo} and CFR without guidewire, thus improving the diagnosis of focal coronary lesions.

To evaluate the flow obstruction effect, stenoses geometry of a focal pre-PTCA (22,23) and post-PTCA lesion (24,25) were obtained from the *in vivo* data set of Wilson et al. (21) in a 32 patient group. The patients had single-vessel, single-lesion coronary artery disease with unstable or stable angina pectoris. Dimensions and shape of the coronary stenosis before and after angioplasty were obtained from quantitative biplanar X-ray angiography. Biplane angiography of each lesion in orthogonal projections (60° left anterior oblique and 30° right anterior oblique) resolved vessel widths with cross-sectional area calculated from the equation for an ellipse, which were converted to mean diameters. The measured mean values ± SD of minimal area stenosis (A_m), mean pressure measured proximal to the stenoses at the ostium ($\tilde{p}_a$), CFR by Wilson et al. (21) and dimensions are summarized in Table 1 and Fig. 6. Patients with abnormalities that might affect the vasodilator capacity of the arteriolar vasculature were excluded from the study (21). Measured values of CFR with a 3F pulsed Doppler ultrasound catheter ($d_i = 1.0$ mm) with tip positioned proximal to the lesions (with minimal flow blockage) increased from $2.3 + 0.1$ to $3.6 + 0.3$ in the procedure; mean arterial pressure, measured in the coronary ostium, decreased from $89 + 3$ to $84 + 3$ mmHg (21). In the flow analysis, the residual composite lesion was assumed to have a smooth, rigid plaque wall, and round concentric shape. Additional dimensional data on the shape of similar size lesion are from Back and Denton (26).

Additionally, an intermediate stenosis (27) having maximal area blockage based on minimal diameter = 80% was used in this study [the dimensions of which are given in Table 1 and were obtained from Back and Denton (26)] to include a wide range of lesion sizes for obtaining the correlations. However, this intermediate lesion size was not measured by Wilson et al. (21). Further, for guidewire analyses, the guidewire was placed concentrically within the lesion. The concentric configuration of the guidewire within the lesion may give the largest pressure drop



(a)

	d_e	d_r	d_m	l_c	l_m	l_r	% Area Stenosis
Post-Angioplasty	3	3	1.8	6	3	1.5	64
Intermediate	3	3	1.35	6	0.75	1.5	80
Pre-Angioplasty	3	3	0.95	6	0.75	1.5	90

(b)

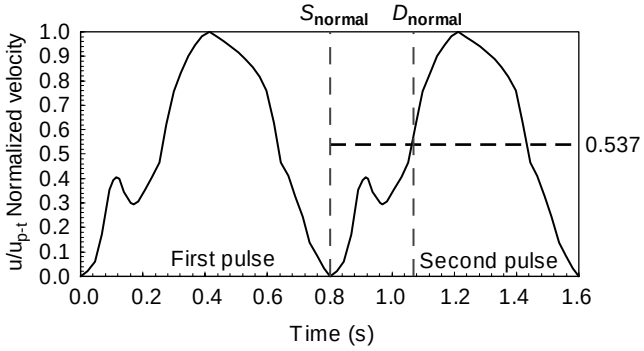


Figure 6. (a) Stenoses geometry showing the shape and dimensions in mm; (b): Normal coronary flow waveform $\tilde{u}/\tilde{u}_{\text{p-t}}$ versus t, where S_{normal} indicates the beginning of systole and D_{normal} indicates the beginning of diastole.

(1). The geometry of the native, intermediate, and moderate lesion with and without guidewire is shown in detail in Fig. 7a–c .

The coronary velocity waveform $\overline{u}(t)$ (spatially averaged at each time across the cross-sectional area) used in the flow analyses was obtained in our laboratory from *in vitro* calibration (28), smoothing the fluctuating Doppler signal, and phase shifting the normal pattern for the proximal left anterior descending (LAD) artery. In Fig. 5b, the peak diastolic velocity $\overline{u}_{\text{p-t}}$ corresponds to a normalized velocity of 1.0, so that the mean peak velocity ratio $\tilde{\overline{u}}/\overline{u}_{\text{p-t}}$ is 0.537, as shown by the dashed line.

With guidewire inserted concentrically, in the proximal vessel, the spatial velocity profile in the annular gap was taken to be the analogous Poiseuille flow relation for the axial velocity u (22,25):

$$u/2\overline{u} = [(1-(r/r_o)^2)\ln(r_o/r_i) + (1-(r_i/r_o)^2)\ln(r/r_o)]/ [(1+(r_i/r_o)^2)\ln(r_o/r_i) - (1-(r_i/r_o)^2)] \quad (1)$$

where u is a function of r and t. Without the guidewire in the proximal vessel, the spatial velocity profile was initially taken to be the Poiseuille flow relation for the axial velocity u:

$$u/2\overline{u} = (1-(r/r_o)^2) \quad (2)$$

The Carreau model, given by Eq. 3, was used for shear rate dependent non-Newtonian blood viscosity with the local shear rate (Eq. 4) calculated from the velocity gradient through the second invariant of the rate of strain tensor (35).

$$\eta = \eta_\infty + (\eta_o - \eta_\infty)\left[1 + (\lambda\dot{\gamma})^2\right]^{(n-1)/2} \quad (3)$$

$$\dot{\gamma} = \sqrt{\frac{1}{2}\left[\sum_i \sum_j \dot{\gamma}_{ij}\dot{\gamma}_{ji}\right]} \quad (4)$$

where $\eta_\infty = 0.00345$ Pa·s $\eta_o = 0.056$ Pa·s, $\lambda = 3.313$ s and $n = 0.3568$.

Details of the numerical method used to calculate the pulsatile hemodynamics in coronary artery and lesions with and without guidewire were previously described by Banerjee et al. (22–24). A typical basal physiological value $\tilde{Q}_b = 50\,\text{mL}\cdot\text{min}^{-1}$ for a coronary vessel of 3 mm size was used (30). The cycle time of 0.8 s and density of blood $\rho = 1.05$ g·cm^{-3} was used. In the Reynolds number (Re), a kinematic viscosity of $v = 0.035$ cm^2·s was used, a value near the asymptote in the Carreau model for blood ($\eta_\infty \to 0.00345\,\text{Pa}\cdot\text{s}$ as $\dot{\gamma} \to \infty$), which gives $v_\infty \to 0.033$ cm^2·s. The Womersley number varied from 1.9 with guidewire size 0.35–2.25 mm in the pathophysiological scenario without guidewire.

CFR and $\tilde{p}_{\text{rh}}$ without guidewire (21–25) and computed values of CFR and $\tilde{p}_{\text{rh-g}}$ for the different lesions with guidewire (0.35 and 0.46 mm) were used to construct the maximal vasodilation-distal perfusion pressure curve, also known as CFR $-\,\tilde{p}_{\text{rh}}$ relationship. The maximal CFR $-\,\tilde{p}_{\text{rh}}$ curve was plotted by joining the measured (21), and computed values of CFR and $\tilde{p}_{\text{rh}}$ at hyperemia for the native, intermediate, and residual lesions after angioplasty since blood was supplied to the same distal vasculature, which was originally with marked arteriolar dilation. The CFR $-\,\tilde{p}_{\text{rh}}$ relationship was then used to construct the correlations between FFR_{myo} and FFR_{myog}, and CFR and CFR_g in native, intermediate, and residual lesions for the two guidewires. The linear CFR $-\,\tilde{p}_{\text{rh}}$ was also extrapolated toward its origin to estimate zero-coronary flow mean pressure ($\tilde{p}_{\text{zf}}$) for the Wilson et al. (21) patient group.

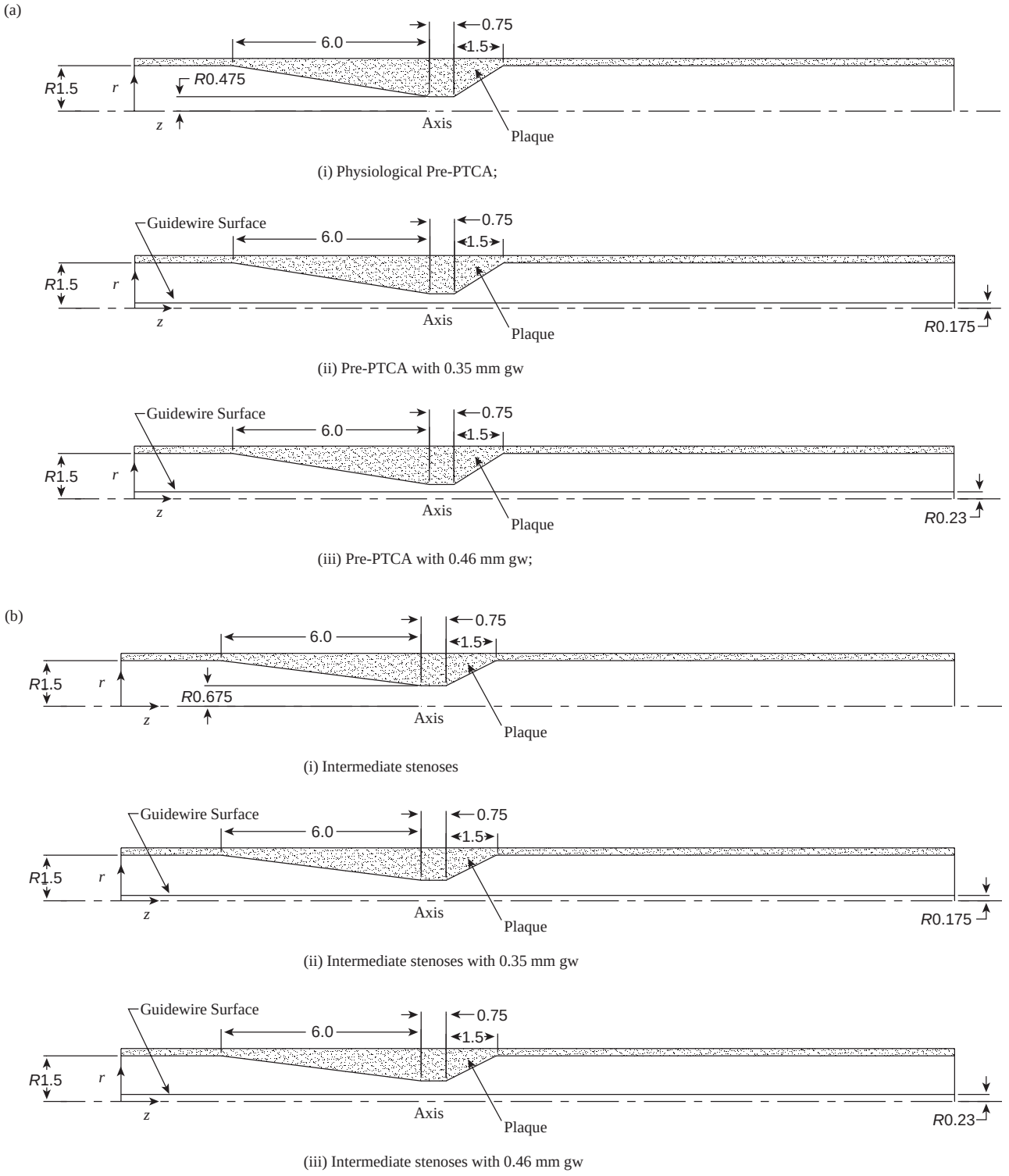
Diagnostics with Angioplasty Catheters

As an initial step, the authors calculated the $\Delta\tilde{p} - \tilde{Q}$ relationship post-PTCA (Curve J in Fig. 8), in conjunction with the pressure measurements, using the angioplasty catheter ($d_i = 1.4$ mm) before the development of small guidewire sensors (31). For resting conditions with the catheter present, flow was believed to be ~40% of normal basal flow in the absence of the catheter, and for hyperemia, ~20% of elevated flow in the patient group. Also, $\Delta\tilde{p}$ was significantly elevated in the tighter artifactual stenoses during the measurements. The above diagnostic measurements were compared with the pathophysiologic scenario, having no angioplasty catheter. The results of pathophysiologic conditions cannot be measured in lesions, and are descriptive of the unperturbed conditions that may have existed on average in the patient group after PTCA. In the absence of angioplasty catheter, the calculated $\Delta\tilde{p}$ was only ~1 mmHg (0.133 kPa) at basal

flow, and increased moderately to ~7.4 mmHg for hyperemic flow measured proximally (CFR = 3.6) with minimal blockage. On the other hand, with the catheter, $\Delta\tilde{p}$ was ~28.7 mmHg for the basal flow showing an order of magnitude increase in $\Delta\tilde{p}$.

Increased Pressure Drop and Reduced Hyperemic Flow Due to the Presence of Guidewire

Tables 1 and 2 give the mean pressure drop $\Delta\tilde{p}_{\mathrm{h}}$ and distal mean pressure $\tilde{p}_{\mathrm{rh}}$ at hyperemic condition in native,



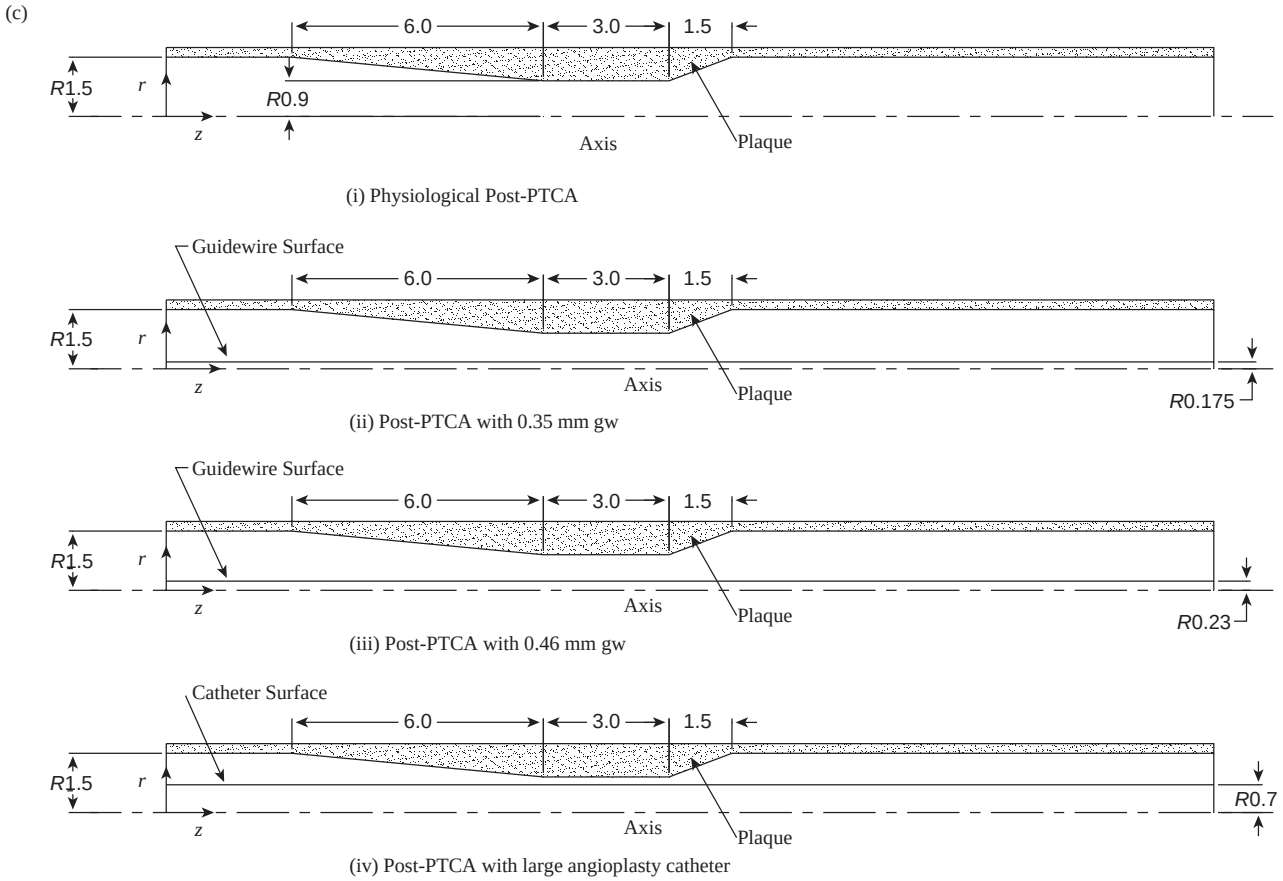


Figure 7. (a) Stenoses geometries with guidewire inserted: Pre-PTCA; (b): Stenoses geometries with guidewire inserted: Intermediate; (c) Stenoses geometries with guidewire inserted: Post-PTCA.

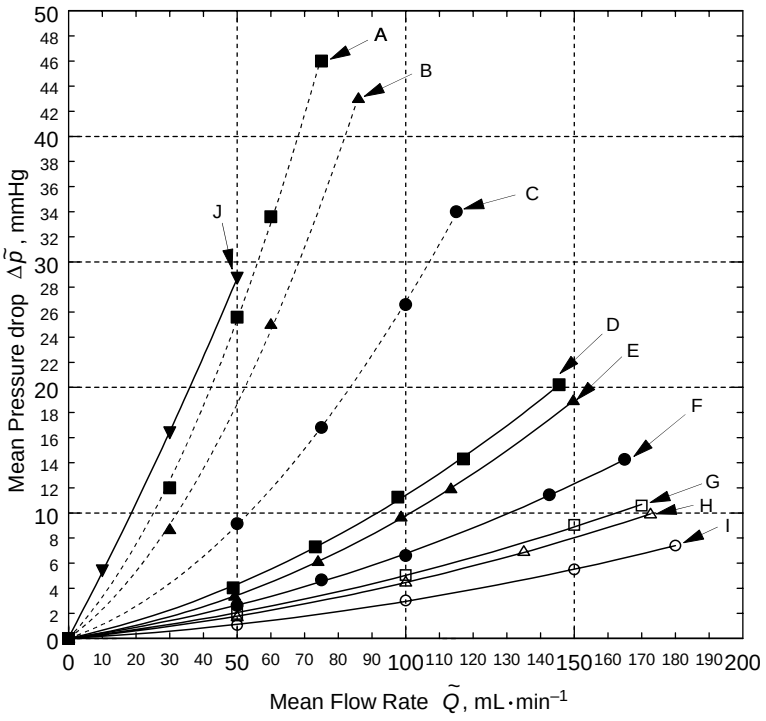


Figure 8. Mean pressure drop $\Delta\tilde{p}$ versus mean flow rate $\tilde{Q}$ in pre-PTCA, intermediate, and post-PTCA lesion with (0.35 and 0.46 mm) and without guidewire.

Table 2. Hyperemic Mean Pressure Gradients $\Delta\tilde{p}_{gh}$ and Distal Mean Pressure $\tilde{p}_{rh}$ Before, Intermediate, and After Intervention with Guidewire[a]

		Guidewire size: $d_i = 0.35$ mm				Guidewire size: $d_i = 0.46$ mm			
Lesion	$\tilde{p}_a$, mmHg	$\frac{d_i}{d_m}$	$\tilde{Q}_{gh}$ mL·min^{-1}	$\Delta\tilde{p}_{gh}$, mmHg	$\tilde{p}_{rh}$, mmHg	$\frac{d_i}{d_m}$	$\tilde{Q}_{gh,}$ mL·min^{-1}	$\Delta\tilde{p}_{gh,}$ mmHg	$\tilde{p}_{rh,}$ mmHg
Before-PTCA	89	0.37	86.0	43.0	46.0	0.48	75.0	46.0	43.0
Intermediate	86	0.26	149.7	18.9	65.8	0.34	145.6	20.2	64.5
After-PTCA	84	0.19	172.5	9.9	72.8	0.26	170.0	10.6	72.1

[a]Values obtained from linear CFR-$\tilde{p}_{rh}$ curve.

intermediate and residual lesions with and without guidewire (22–26). The ratio of guidewire diameter d_i to throat diameter d_m for different stenoses are given in Table 2. From Tables 1 and 2, overall guidewires of size 0.35 and 0.46 mm caused 30–45% overestimation in $\Delta\tilde{p}_h$ in the three stenoses as hyperemic flow was reduced by 5–37% from the physiologic condition without guidewire obstruction. While diagnosis of severely blocked arteries is relatively easier and additional pressure drop due to the guidewire does not affect the diagnosis that much, it can be seen that additional pressure drop due to the guidewire could have an important role in clinical evaluation of moderate and intermediate stenoses. Further, it can be seen that with guidewire of size 0.35 and 0.46 mm, $FFR_{myog} < FFR_{myo}$ and $CFR_g < CFR$ at hyperemia. The maximum decrease was observed in the native stenoses. The residual stenoses (i.e., post-PTCA) showed the least decrease. In summary, Fig. 8 shows the mean pressure drop from basal to hyperemic for the three stenoses with and without 0.35 and 0.46 mm guidewire.

Figure 9 shows the maximal CFR $-\tilde{p}_{rh}$ for the native, intermediate, and residual lesions after angioplasty with and without guidewire (27). Distal mean perfusion pressure for hyperemic flow $\tilde{p}_{rh}$ increased from 55 mmHg in pre-PTCA without guidewire to ~75 mmHg (9.99 kPa) in post-PTCA without guidewire. A $\tilde{p}_{rh}$ of 55 mmHg (7.33 kPa) is indicative of subendocardium ischemia.

Extrapolation of the nearly linear CFR $-\tilde{p}_{rh}$ relation toward its origin gave a zero-coronary flow mean pressure ($\tilde{p}_{zf}$) of ~20 mmHg (2.66 kPa). This value is near a measured value (32) of 18 mmHg (2.39 kPa), where myocardial blood flow ceased in the subendocardium layer of dog hearts, which were maximally dilated by infusion of adenosine.

FFR_{myo}-$FFR_{myo\text{-}g}$ and CFR–CFR_g Correlations

Table 3 provides the values of CFR, CFR_g, FFR_{myo}, and FFR_{myog} for the different stenosis configurations (27). Figures 10 and 11 show the correlation plots between

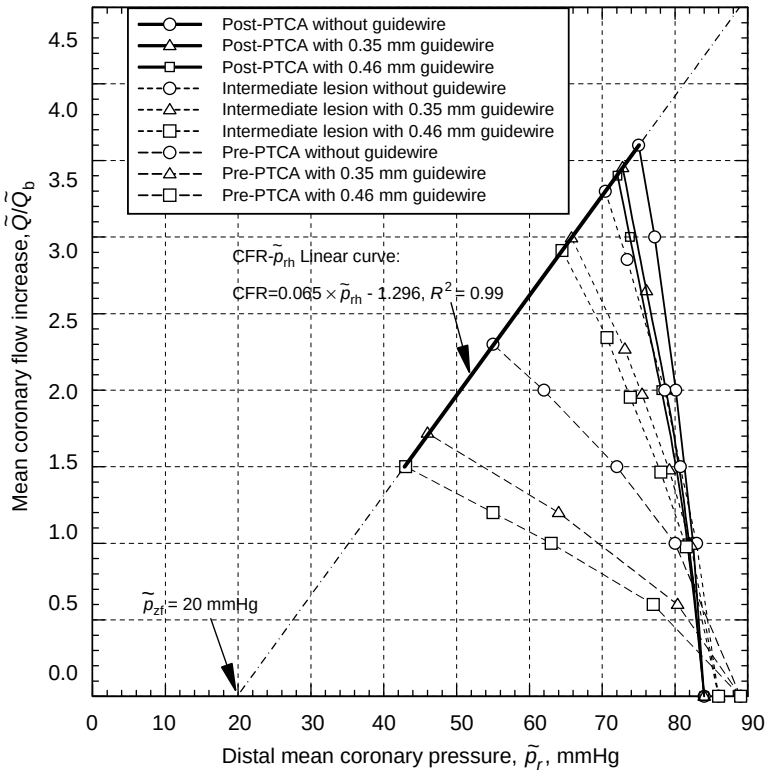


Figure 9. Relative mean coronary flow rate increase ($\tilde{Q}/\tilde{Q}_b$) versus distal mean coronary pressure $\tilde{p}_r$ before and after intervention. The maximum vasodilation-distal perfusion pressure relation (CFR- $\tilde{p}_{rh}$) is shown by the nearly linear solid line.

Table 3. Effect of the Presence of Guidewire on the Values of CFR and FFR_{myo} at Hyperemic Conditions in Comparison to Values of CFR and FFR_{myo} Under Pathophysiological Condition for Different Guidewire Sizes $d_i = 0.35$ mm (0.014 in.) and 0.46 mm (0.018 in.)[a]

	Physiological		0.35 mm Guidewire		0.46 mm Guidewire	
Lesion	CFR	FFR_{myo} with $\tilde{p}_v \sim 0$	CFR_g	$FFR_{myo\text{-}g}$ with $\tilde{p}_v \sim 0$	CFR_g	$FFR_{myo\text{-}g}$ with $\tilde{p}_v \sim 0$
Before-PTCA	2.3	0.62	1.72	0.52	1.50	0.48
Intermediate	3.3	0.82	2.99	0.76	2.91	0.75
After-PTCA	3.6	0.89	3.45	0.87	3.40	0.86

[a] $FFR_{myo\text{-}g}$ is FFR_{myo} measured with guidewire.

CFR and CFR_g, and between FFR and FFR_g with their *linear* regression lines. In Fig. 10, CFR was related to CFR_g by the equation: with guidewire size 0.46 mm, $CFR = CFR_g \times 0.689 + 1.271$ ($R^2 = 0.99$), and with guidewire size 0.35 mm, $CFR = CFR_g \times 0.757 + 1.004$ ($R^2 = 0.99$). With central venous pressure ($\tilde{p}_v$) ~ 0, as used in present clinical practice, FFR_{myo} and FFR_{myog} correlated equally well (Fig. 11): with guidewire size 0.46 mm, $FFR_{myo} = FFR_{myog} \times 0.737 + 0.263$ ($R^2 = 0.99$), and with guidewire size 0.35 mm, $FFR_{myo} = FFR_{myog} \times 0.790 + 0.210$ ($R^2 = 0.99$), which gave $FFR_{myo} = 0.8$ for a measured $FFR_{myog} = 0.75$. The study showed that the correlations for FFR and CFR for guidewire size 0.35 mm are closer to the ideal (expected) relationship, that is, $CFR = CFR_g$ and $FFR = FFR_g$ due to relatively lesser flow obstruction effect than guidewire size 0.46 mm.

Usefulness of $FFR_{myo\text{-}g}$-FFR_{myo} and CFR-CFR_g Correlations

Though there are uncertainties associated with measurements of CFR with guidewires in diagnostic procedures as these measurements are made distal to the stenosis, these corrections could be useful in measuring CFR and FFR simultaneously. This could provide useful information about the status of both the epicardial stenosis and distal microcirculation during the procedure. In light of this, CFR measurement proximal to the stenosis will be more accurate as this is the region of more stable flow. Further, the heterogeneity of stenoses could produce variations in the slope of the correlations. However, the correlations were obtained based on stenosis shape and size measured invasively by Wilson et al. (21) in a group of patients, and have shown a direct and stronger relation to the minimal stenosis area as compared to the overall length and shape of the stenosis. Further clinical evaluation with a larger patient group will be needed to confirm the dominant effect of minimal stenosis area on these correlations.

While several studies have focused on the relationship between CFR measured with guidewire and CFR measured with a noninvasive flow probe, for examples a Doppler cuff in animal studies, no similar study has been done to distinguish the effect of guidewire on FFR_{myo} as pressure

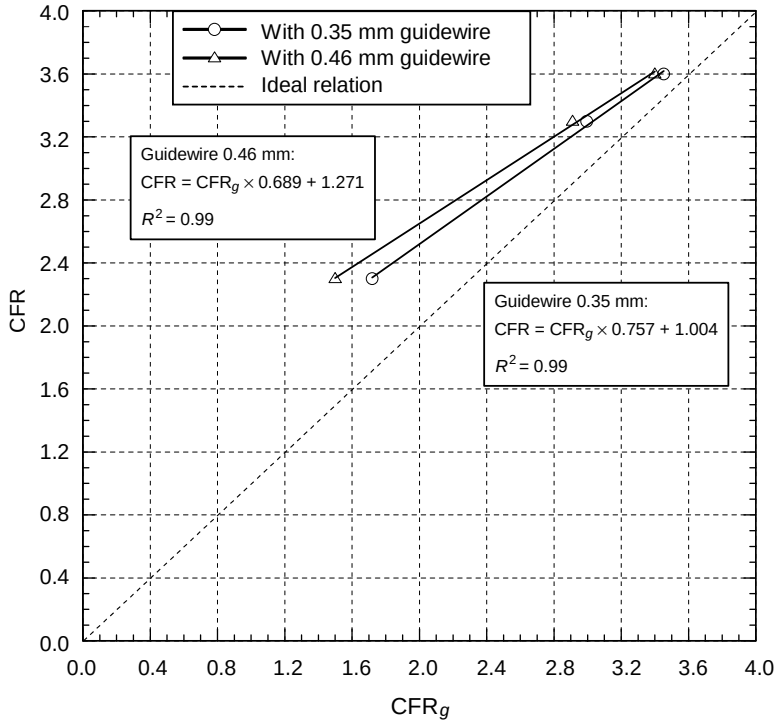


Figure 10. CFR versus CFR_g correlation. Dotted line shows the ideal CFR versus CFR_g relation.

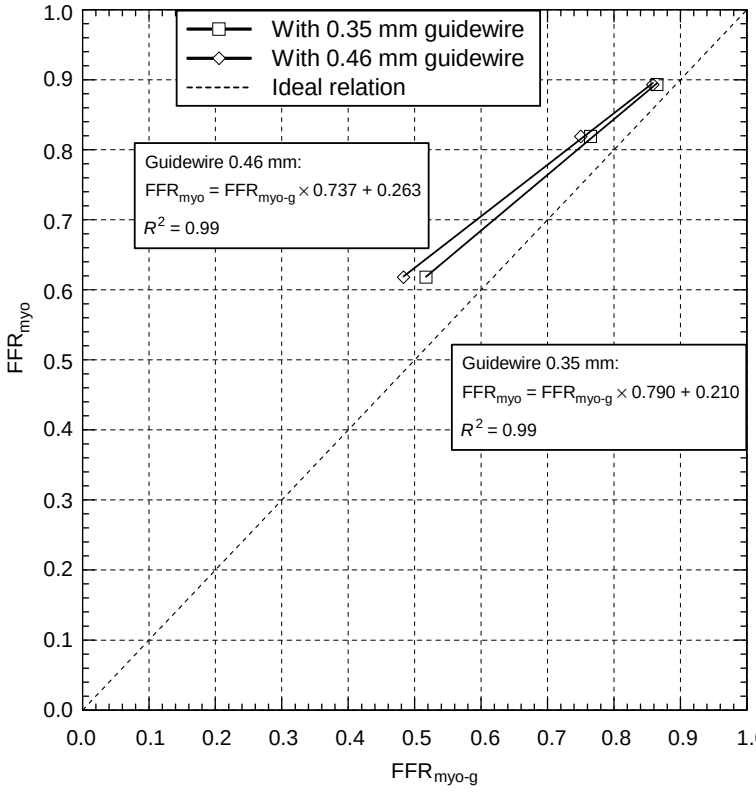


Figure 11. FFR_{myo} versus FFR_{myog} correlation (with $\tilde{p}_v \sim 0$). Dotted line shows the ideal FFR_{myo} versus $FFR_{myo\text{-}g}$ relation.

can be measured invasively using guidewires only. Earlier *In vitro* measured Δp data reported by DeBruyne et al. (33) for steady flow of a saline solution through various blunt hollow plug stenosis models (50–90% area stenosis; $d_e = d_r = 4$ mm; $l_m = 10$ mm) with and without a small guidewire pressure sensor ($d_i = 0.38$ mm) also indicated appreciable overestimation of the true Δp as the severity of stenoses increased. In particular, for a 90% area stenosis ($[d_i/d_m] = 0.30$) at a flow rate $Q = 120$ mL·min^{-1}, the measured ratio ($\Delta p_g/\Delta p$) increased by the factor 23/16 mmHg (3.06/2.13 kPa) ($\sim 40\%$) due to the increased flow resistance with the guidewire spanning the model stenosis in the rigid plastic tube $d_e = 4$ mm. Our detailed hemodynamic analysis in a smaller coronary vessel ($d_e = 3$ mm; 90% area stenosis; larger ($[d_i/d_m] = 0.48$) also indicated significant increases in flow resistance with the guidewire present (Tables 1 and 2). Similarly, the stenosis resistance $\tilde{R}_h$, defined as ($\Delta p_h/\tilde{Q}_h$), decreased considerably from 0.29 to 0.04 mmHg·mL·min^{-1} from pre- to post-PTCA, respectively (27). With guidewire size 0.35 mm, $\tilde{R}_{hg}$ decreased from 0.5 for pre-PTCA to 0.06 for post-PTCA mmHg·mL·min^{-1} (27). While the 32 patient group in Wilson et al. (21) had normal microvascular function, the presence of microvascular impairment could produce different correlations between FFR_{myo} - FFR_{myog} and CFR–CFR_g with epicardial stenoses, and could be the focus of future studies on guidewire effect and microcirculation.

BIBLIOGRAPHY

1. Dotter CT, Judkins MP. Transluminal treatment of arteriosclerotic obstruction. Description of a new technic and a preliminary report of its application. 1964. Radiology 1989;172:904–920.
2. Baim DS. Percutaneous transluminal coronary angioplasty. Cardiac catheterization, angiography, and intervention. 5th ed. (MA): Williams & Wilkins; 2000. pp 537–580.
3. Dervan JP, McKay RG, Baim DS. The use of an exchange guide wire in coronary angioplasty. Cathet Cardiovasc Diagn 1985;11:207–212.
4. Pijls NH, et al. Measurement of fractional flow reserve to assess the functional severity of coronary-artery stenoses. N Engl J Med 1996;334:1703–1708.
5. Pijls NH, et al. Coronary thermodilution to assess flow reserve: validation in humans. Circulation 2002;105:2482–2486.
6. De Bruyne B, et al. Coronary thermodilution to assess flow reserve: experimental validation. Circulation 2001;104:2003–2006.
7. Siebes M, et al. Single-wire pressure and flow velocity measurement to quantify coronary stenosis hemodynamics and effects of percutaneous interventions. Circulation 2004;109: 756–762.
8. Ernst SM, et al. Immediate sealing of arterial puncture sites after cardiac catheterization and coronary angioplasty using a biodegradable collagen plug: results of an international registry. J Am Coll Cardiol 1993;21:851–855.
9. Aker UT, et al. Immediate arterial hemostasis after cardiac catheterization: initial experience with a new puncture closure device. Cathet Cardiovasc Diagn 1994;31:228–232.
10. Gruentzig AR, Senning A, Siegenthaler WE. Nonoperative dilation of coronary artery stenosis: Percutaneous Transluminal Coronary Angioplasty. N Engl J Med 1979;301:61–68.
11. Ganz P, Harrington DP, Gaspar J, Barry WH. Phasic pressure gradients across coronary and renal artery stenoses in humans. Am Heart J 1983;106:1399–1406.
12. Ganz P, et al. Usefulness of transstenotic coronary pressure gradient measurements during diagnostic catheterization. Am J Cardiol 1985;55:910–914.

13. Anderson HV, et al. Measurement of transstenotic pressure gradient during percutaneous transluminal coronary angioplasty. Circulation 1986;73:1223–1230.
14. Gould KL, Kirkeeide R, Buchi M. Coronary flow reserve as a physiologic measure of stenosis severity, part 1: relative and absolute coronary flow reserve during changing aortic pressure and cardiac workload; part II: determination from arteriographic stenosis dimensions under standardized conditions. J Am Coll Cardiol 1990;15:459–474.
15. Gould KL, Lipscomb K, Hamilton GW. Physiologic basis for assessing critical coronary stenosis: instantaneous flow response and regional distribution during coronary hyperemia as measures of coronary flow reserve. Am J Cardiol 1974;33: 87–94.
16. Gould KL. Coronary artery stenosis and reversing atherosclerosis. 2nd ed. London: Arnold Publishers; 1999.
17. Pijls NHJ, De Bruyne B. Coronary pressure. 2nd ed. The Netherlands: Kluwer Publishers; 1999.
18. Baumgart D, et al. Improved assessment of coronary stenosis severity using the relative flow velocity reserve. Circulation 1998;98:40–46.
19. Pijls NH, et al. Fractional flow reserve: a useful index to evaluate the influence of an epicardial coronary stenosis on myocardial blood flow. Circulation 1995;92:3183–3193.
20. De Bruyne B, et al. Fractional flow reserve in patients with prior myocardial infarction. Circulation 2001;104(2):157–162.
21. Meuwissen M, et al. Intracoronary pressure and flow velocity for hemodynamic evaluation of coronary stenoses. Expert Rev Cardiovasc Ther 2003;1:471–479.
22. Back LH. Estimated mean flow resistance increase during coronary artery catheterization. J Biomech 1994;27:169–175.
23. Kern MJ, et al. Translesional pressure—flow velocity assessment in patients: part I. Cathet Cardiovasc Diagn 1994;31:49–60.
24. Segal J, et al. Alterations of phasic coronary artery flow velocity in humans during percautaneous coronary angioplasty. J Am Coll Cardiol 1992;20:276–286.
25. Doriot P, Dorsaz P, Dorsaz I, Chatelain P. Accuracy of coronary flow measurements performed by means of doppler wires. Ultra Med Biol 2000;26:221–228.
26. Wilson RF, Laxson DD. Caveat emptor: a clinician's guide to assessing the physiologic significance of arterial stenoses. Cathet Cardiovasc Diagn 1993;29:93–98.
27. Banerjee RK, Back LH, Back MR. Effects of diagnostic guidewire catheter presence on translesional hemodynamic measurements across significant coronary artery stenoses. Biorheology 2003;40(6):613–635.
28. Banerjee RK, Back LH, Back MR, Cho YI. Physiological flow analysis in significant human coronary artery stenoses. Biorheology 2003;40(4):451–476.
29. Banerjee RK, Back LH, Back MR, Cho YI. Physiological flow simulation in residual human stenoses after coronary angioplasty. ASME J Biomech Eng 2000;122:310–320.
30. Sinha Roy A, Back LH, Banerjee RK. Guidewire flow obstruction effect on pressure drop-flow relationship in moderate coronary artery stenosis. Accepted for publication in J Biomech January, 2005.
31. Wilson RF, et al. The effect of coronary angioplasty on coronary flow reserve. Circulation 1988;77:873–885.
32. Back LH, Denton TA. Some arterial wall shear stress estimates in coronary angioplasty. Adv Bioeng ASME BED 1992;22:337–340.
33. Sinha Roy A, et al. Delineating the guidewire flow obstruction effect in assessment of fractional flow reserve and coronary flow reserve measurements. Accepted for publication in Am J Physiol: Heart Circ Physiol February, 2005.
34. Cho YI, Back LH, Crawford DW, Cuffel RF. Experimental study of pulsatile and steady flow through a smooth tube and an atherosclerotic coronary artery casting of man. J Biomech 1983;16:933–946.
35. Cho YI, Kensey KR. Effects of non-Newtonian viscosity of blood on flows in a diseased arterial vessel: Part 1. Steady flows. Biorheology 1991;28:241–262.
36. Back LH, Radbill JR, Crawford DW. Analysis of pulsatile viscous blood flow through diseased coronary arteries of man. J Biomech 1977;10:339–353.
37. Womersley JR. Method for the calculation of velocity, rate of flow and viscous drag in arteries when the pressure gradient is known. J Physiol 1955;127:553–563.
38. Banerjee RK, Back LH, Back MR, Cho YI. Catheter obstruction effect on pulsatile flow rate-pressure drop during coronary angioplasty. ASME J Biomech Eng 1999;121:281–289.
39. Brown BG, Bolson EL, Dodge HT. Dynamic mechanisms in human coronary stenosis. Circulation 1984;170:917–922.
40. Bache RJ, Schwartz JS. Effect of perfusion pressure distal to a coronary stenosis on transmural myocardial blood flow. Circulation 1982;65:928–935.
41. De Bruyne B, et al. Transstenotic coronary pressure gradient measurement in humans: in vitro and in vivo evaluation of a new pressure monitoring angioplasty guide wire. J Am Coll Cardiol 1993;22(1):119–126.
42. Fearon WF, Yeung AC. Evaluating intermediate coronary lesions in the cardiac catheterization laboratory. Rev Cardiovasc Med 2003;4:1–7.

See also ARTERIES, ELASTIC PROPERTIES OF; BRACHYTHERAPY, INTRAVASCULAR; HEMODYNAMICS; INTRAAORTIC BALLOON PUMP.

CPR. See CARDIOPULMONARY RESUSCITATION.

CRYOSURGERY

ANDREW A. GAGE
State University of New York at Buffalo
Buffalo, New York

INTRODUCTION

Cryosurgery is a method of therapy that uses freezing temperatures to achieve effects on tissue. The term cryotherapy, often used interchangeably with cryosurgery, has broader connotations, including, for example, the application of cold packs to prevent tissue swelling after injury. Cryosurgery is one form of cryotherapy. The term cryoablation is commonly used also, especially in relation to the treatment of tumors.

Cryosurgery may be applied for several different purposes, some related to the adhesion of super-cold metal to tissue and some related to the response of tissue to freezing. For example, in the extraction of cataracts of the eye, a cold instrument or probe is used only to secure a hold on the lens of the eye and facilitate removal as the lens adheres to the probe, which functions as a handle. This technique can be used to facilitate extraction of tumors of diverse sites, including the brain, the eye, and the liver. Cryosurgery also can be used to produce an inflammatory response. For example, in the treatment of retinal detachment, fast freezing of the tissue for a few seconds will damage the tissue, rather

than destroy it, and the resultant inflammatory response is expected to heal the detachment. Cryosurgery also can be used for the destruction of tissue, which might be selective, as in the treatment of non-neoplastic disease, or which can be complete, as is needed in the treatment of tumors. The destructive response is the major use of cryosurgery, and this type of response is emphasized in this article.

Cryosurgery is commonly used to destroy tissue by freezing *in situ*. The technique requires the use of cryosurgical apparatus to produce tissue temperatures in the freezing range. The freezing of the tissue is accomplished by the direct application of cryogenic agents or by the use of closed-probe systems in which the cryogen circulates and is not released on the tissue. The diverse techniques of cryosurgery range from the rather easy surface application of the cryogen for the treatment of skin disease to the more complex use by percutaneous application, as for prostatic disease. The diseases that can be treated by cryosurgery range from minor conditions, such as warts, to serious conditions such as advanced cancer. Wherever the disease, the goal of treatment by cryosurgery is controlled production of a predictable area of tissue necrosis.

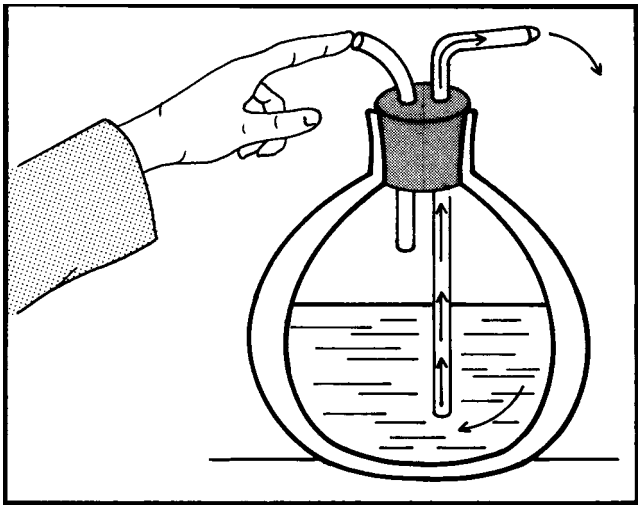


Figure 1. Dr. White's "wash bottle" spray method of ~1900, drawn as described in his article of that year. When the finger is placed over the air outlet, the pressure in the container produced by the boiling of liquid air in the container forces liquid air through the longer glass tube. The spray was then directed at the skin lesion.

HISTORICAL DEVELOPMENTS

Local tissue freezing for the treatment of cancer was first used by Dr. James Arnott of London, who described his technique in the year 1850. Using salt solutions containing ice (ca. $-12\,^{\circ}C$), he produced local freezing by irrigation of advanced cancers in accessible sites, such as the breast and uterine cervix. He described diminution of the size of the tumor and amelioration of pain and drainage. Following his reports, some enthusiasm was generated for the use of cold as an anesthetic agent. Of course, the use of cold for relief of pain had been known since ancient times (1). In Arnott's time, general anesthesia had just been described in America, and its rapidly widening use sharply reduced the usefulness of cold as an anesthetic agent.

In the years from 1870 to 1900, the natural gases were liquefied and Dewar developed a vacuum flask to store cryogenic fluids. These advances permitted development of tissue-freezing techniques. In 1899, A. Campbell White of New York City described the use of liquid air to treat diverse types of skin lesions. Treatment was given by dipping a cotton swab into liquid air and applying the fluid quickly to the skin lesion, using repetitive application to freeze the entire lesion. White also suggested a wash bottle device that sprayed liquid air on skin lesions (Fig. 1).

In 1907, H. H. Whitehouse used freezing techniques for the treatment of a variety of disorders of the skin, including skin cancer. In the same year, Pusey reported similar varied uses of solidified carbon dioxide ($-78.5\,^{\circ}C$), which was easier to handle and more easily obtained than the liquefied gases. For these reasons, carbonic snow remained in clinical use in succeeding years. In the 1930s, liquid oxygen ($-182.9\,^{\circ}C$) had some use in the treatment of skin disease, but flammability and related safety considerations precluded general use.

In the 1930s and 1940s, the usefulness of solid carbon dioxide was increased by the development of new instruments, generally used to cool metal probes, sometimes with a mixture of solid carbon dioxide and acetone. These devices had little or no advantage over the use of the simple stick of solid carbon dioxide. Fluorinated hydrocarbons, commonly known as Freons, came into clinical use in the 1940s, but their freezing capacity was limited.

During these years, beginning in the 1940s with the extensive experiments of Temple Fay in Philadelphia with localized freezing of cancers by irrigation techniques, a number of reports dealt with the experimental production by local freezing of lesions in tissues, such as the liver or brain. In part, these experiments were made for purposes of physiological studies, but they required that an effort be made to develop instrumentation suitable for those purposes, and they showed the possibility of producing destruction of tissue by localized freezing, using diverse cryogenic agents.

In 1950, liquid nitrogen was introduced into dermatological practice by Allington, who described the use of cotton swabs dipped in this cryogen for the treatment of skin disease. As the availability of liquid nitrogen increased, cryosurgery gained in popularity. Nevertheless, cryosurgery remained a rather unimportant therapeutic modality because the freezing capability of cryogenic agents applied topically with cotton swabs was limited. Experimental studies showed that the depth of freezing with the swab techniques was ~2 mm when liquid nitrogen was used as the cryogen, and carbon dioxide was even less effective. This depth of freezing was about the same as the thickness of normal skin, so the technique was suitable only for superficial lesions of the skin.

Cryosurgery as a therapeutic technique received a major stimulus through the development of cryosurgical apparatus by Cooper and Lee in 1961 (2). The apparatus used liquid nitrogen ($-196\,^{\circ}C$) in a closed system, which permitted continuous and rapid extraction of heat from

tissues. The apparatus was originally designed to produce a cryogenic lesion in the brain for the treatment of Parkinsonism and other neuromuscular disorders. However, it was obvious that the apparatus had wider usefulness and, therefore, it was modified quickly and applied to the treatment of other types of diseases in diverse sites. In the following years, several types of cryosurgical apparatus using liquid nitrogen and other cryogenic agents were developed and found areas of usefulness in diverse benign and neoplastic conditions.

In the 1970s, some uses of cryosurgery were virtually abandoned, in part because of competition with other methods of local treatment, such as lasers and electrocoagulation. Other uses of cryosurgery became standard practice, especially in easily accessible areas, such as the skin and uterine cervix. In the 1980s, cryosurgical techniques for cardiac tachyarrhythmias were developed, but progress in other areas was slow.

In the 1990s, renewed interest in cryosurgery followed the development of intraoperative ultrasound, the improvement in cryosurgical apparatus, and the availability of percutaneous access techniques. The ultrasound image identified the site of the lesion, guided the placement of the probe into the lesion, and monitored the process of freezing. More types of cryosurgical apparatus were available, were well suited to percutaneous or endoscopic use, and permitted the simultaneous use of multiple probes. As a result, cryosurgical techniques have evolved into new applications, such as the treatment of visceral and other deep tumors. A recently written history of cryosurgery provides greater detail of its evolution, including the pertinent references (3).

EFFECT OF FREEZING ON TISSUE

All types of cells can be devitalized by freezing. The mechanisms of injury are related to crystallization of water, solute concentration in the cells, and irreversible changes in cell membranes. In the absence of cryoprotective agents, which are used when freezing is used to preserve cells, ice crystal formation produces damage that makes cell survival unlikely. Intracellular ice formation, which occurs when cells are frozen rapidly, is lethal. Extracellular ice formation, which occurs when cells are frozen slowly and water has sufficient time to leave the cell, is not considered as certainly destructive, but the cellular water loss does result in hypertonic damage to the cells. Under cryosurgical conditions, both mechanisms of injury are operative.

Damage from direct cellular injury is enhanced by the effect of freezing on the circulation. After freezing and thawing, the involved area becomes congested, and effective circulation through small vessels ceases within ~30 min. With stagnation of the microcirculation, the hypoxic cells die and necrosis follows. These effects are well known from studies of frostbite, in which the relative importance of direct cellular injury and vascular stasis have long been debated. The full scope of the effects of freezing on tissue is described in recent reviews (4–6). Though direct cell injury is important, in cryosurgery, microcirculatory failure clearly is a major factor in cell death: The loss of blood supply deprives all cells in frozen tissues of any possibility of survival. This results in a uniform necrosis of the tissue, except at the periphery of the previously frozen area. At the periphery of the cryogenic lesion, where the freezing temperature is not sufficiently cold to kill all of the cells, some cells survive and some cells linger between life and death for days and may die showing signs of apoptosis, that is, gene-regulated cell death (7,8).

The cryogenic lesion is characterized by sharply circumscribed necrosis. As thawing takes place, the previously frozen area becomes edematous and discolored due to congestion and perivascular hemorrhage. At the periphery of the dark red area, which closely corresponds to the margins of the previously frozen tissue, a narrow, bright red zone due to hyperemia appears. Further evidence of injury develops slowly, and the later tissue response depends on the severity of freezing. If only the tissue is subjected to mild superficial freezing, as might be done for benign disease of the skin, then the response will range from inflammatory reaction to superficial necrosis. The more extensive freezing required by large tumors is followed by greater destruction of tissue. Sharply demarcated necrosis becomes apparent in ~2 days. The time required for slough of the necrotic tissue depends in part on its stroma. Cellular tissue sloughs quickly, but skin and other tissues with large quantities of fibrous stroma resist structural change and the necrotic tissue requires many days for separation. In the skin, the eschar requires two more weeks for separation, leaving a clean, granulating wound, which heals at a normal rate. The delay in healing that is characteristic of cryosurgical wounds is due to the time required for separation of the necrotic tissue. Healing is commonly favorable with rather little scarring. However, whenever full thickness of skin is lost, some scarring is inevitable. Hyperplastic scars are unusual. Depending on the severity of injury, the pigmentation of the treated area may be diminished or lost. Sometimes increased skin pigmentation at the periphery of the injured area is seen as a result of increased melanoblastic activity, but this is only temporary. In deep tissues, such as the viscera, the necrotic tissue is slowly absorbed over weeks or even longer, depending upon the volume of tissue frozen. Scarring is minimal.

Though tissues are devitalized by freezing, the matrix or structure of the tissue may be little changed, and this preservation of the framework is important in later repair. The resistance of the collagen fibers in the skin to damage by freezing is important to the reparative process (9). It is manifest in the favorable healing frequently seen in the treatment of skin disease and in the peripheral nerves after freezing. Though degeneration of axons and Schwann cells occurs, the perineurium is preserved, and this serves as a pathway for regrowth of axons, leading to eventual return of nerve function. Similar effects follow the freezing and repair of major blood vessels. Larger blood vessels, such as the aorta, femoral arteries, carotid artery, and portal vein, are devitalized by freezing *in situ*. With thawing, the previously frozen vessel is slightly dilated due to loss of tone, but the function as a blood conduit is unimpaired. The endothelium is lost, but the stroma of the vessel wall serves

as the matrix for repair, commonly with some intimal thickening (10).

The effect of freezing on bone is of particular pertinence. Bone devitalized *in situ* by freezing is slowly resorbed and simultaneously replaced with new bone, a lengthy healing process that may take many months, depending on the volume of bone, similar to that which occurs with autogenous bone grafts. During repair, the devitalized bone maintains form and continues function, though bones subjected to considerable stress (as the femur) are susceptible to fracture in the first month or two when bone resorption is maximal (11). This favorable reparative response has permitted extensive freezing of bone tumors in order to avoid excision.

APPARATUS

A wide variety of cryosurgical apparatus, using diverse cryogens, such as liquid nitrogen, nitrous oxide, argon, and carbon dioxide, is available. Various Freons, which were used for cryosurgery in past years, are no longer used because of environmental concerns. The types vary from electronically controlled automated apparatus with probe heaters to inexpensive hand-held devices that are little more than thermos bottles with controls for cryogen flow. The cooling is produced by change in the phase of the cryogen, that is, evaporation of a liquid or solid, or by expansion of compressed gas through a small orifice [Joule–Thomson (J–T) effect]. Thermoelectric cooling (Peltier effect), produced by passing direct current through dissimilar metal junctions (thermocouples), has not been useful in cryosurgery.

Currently, most apparatus use liquid nitrogen ($-195.8\,^{\circ}C$), argon ($-185.9\,^{\circ}C$), or nitrous oxide ($-89.5\,^{\circ}C$). Carbon dioxide, which sublimes at $-78.5\,^{\circ}C$, though commonly used in past years, is little used in current times. Liquid nitrogen cools by change in phase, that is, changing from a liquid to a gas. Argon, nitrous oxide, and carbon dioxide are used as pressurized gases that cool by the J–T effect. The freezing capability of these cryogenic agents varies substantially, and this determines the choice of equipment for a particular disease. The cryogenic agents may be applied directly to the tissue, typically as a spray of liquid nitrogen, though nitrous oxide has been used in this way also. Freezing with liquid nitrogen applied via a cotton swab is another example of direct use. The cryogens may also be used in probes, which are a means of confining the cryogen in a closed system. At its tip, the metal probe has a heat-exchange surface that is applied to the tissue to be frozen. At this contact area, heat transfer to the probe results in tissue cooling. The freezing capacity varies with the size of the probe, the temperature of the probe, and the areas of the contact with the tissue. The heat removing capacity of probes varies from 10 to 100 W, depending on the features of probe construction and the cryogen that cools it.

The coldest cryogenic agent in clinical use is liquid nitrogen. It has the greatest freezing capability and is the best agent for destruction of large volumes of tissue, as is required in the treatment of cancer. Nevertheless, in current practice, pressurized argon is commonly used for the treatment of neoplastic disease. The use of multiple probes simultaneously compensates for the somewhat lesser freezing capability of argon. The other cryogens are useful for less serious lesions, such as nonneoplastic or benign neoplastic disease, for which lesser degrees of freezing will suffice (Table 1).

Liquid nitrogen is a clear fluid that is odorless and nonflammable. Its boiling point is $-195.8\,^{\circ}C$ at atmospheric pressure. The liquid will expand to 750 times its volume under normal atmospheric pressure. It must be stored in a double-walled, vacuum-insulated container with provision for pressure relief and liquid nitrogen withdrawal. The most popular containers for office or clinic use have a capacity of 15–35 L, which provides a holding time of 60–90 days and, depending on the rate of use, will require refilling every 4–6 weeks. Liquid nitrogen will evaporate at a rate of a few percentage points per day, depending on the quality of the container. Withdrawal devices, basically spigots, are used to transfer the liquid nitrogen from the storage container to the cryosurgical instrument.

Hand-held cryosurgical apparatus, weighing $\sim$1 kg when empty, are commonly used, especially in dermatological practice (Fig. 2). They are basically small containers (thermos bottle construction) with storage capacities of 250–500 mL and with suitable on–off controls to initiate and control the spray of liquid nitrogen. Some devices allow pressure to build up in the container by means of a heat exchanger in the wall or top. Most have Luer lock fittings in the nozzle so that a variety of spray apertures, needles, or nozzles can be attached. These range in size from 24 to 15 gauge. The smaller aperture sizes are hindered by a tendency to become occluded by the development of frost in use, but this can be alleviated by bypass nozzle systems. Problems that must be solved in the construction of hand-held devices include the design of a delivery tube that permits adequate heat exchange and prevents drip of liquid nitrogen from the delivery system nozzle. Most hand-held devices can be fitted with probes of diverse shapes and sizes, including those suitable for treatment of oral or gynecological diseases. The hand-held devices, relatively heavy when the container is filled with liquid nitrogen, are somewhat more difficult to use with a probe than with a spray because it is cumbersome to hold the weight steady in the hand while the probe is adherent to the tissue. Motion may cause fracture of the bond between probe and tissue, and ineffective freezing may result. With the spray technique, the small movement caused by the weight in the hand is not important because there is no direct bond to tissue.

The automated apparatus, cooled by liquid nitrogen, available from several companies, is almost always used with probes. The feed lines leading to the probes are insulated, usually by vacuum (Fig. 3). The control of the flow of liquid nitrogen is achieved by regulators. Most systems require pressurization, which is facilitated and speeded with an internal heating device. A heater in the probe tip speeds release from the tissue at the conclusion of freezing.

The automated apparatus first available in the early 1960s provided for control of the probe temperature in the

Figure 2. Modern handheld liquid nitrogen cryosurgical unit of the type commonly used in the treatment of skin diseases. The taller unit is 11 in. (280 mm) in height, has a capacity of 16 oz (500 mL), and weighs 30 oz (846 g) when filled. The shorter unit is 8.5 in. (215 mm) in height, has a capacity of 10 oz (300 mL), and weighs 24 oz (618 g) when filled. The devices may be fitted with a large selection of cryogen spray tips with apertures ranging from 0.4 to 1 mm (shown in Fig. 2b) and probes ranging in size from 1 mm to 4 mm in diameter at the tip. The probes allow precise control of freezing and may be used for cutaneous or mucosal lesions. (Photographs by courtesy of Brymill Cryogenic Systems, Ellington, CT 06029.)

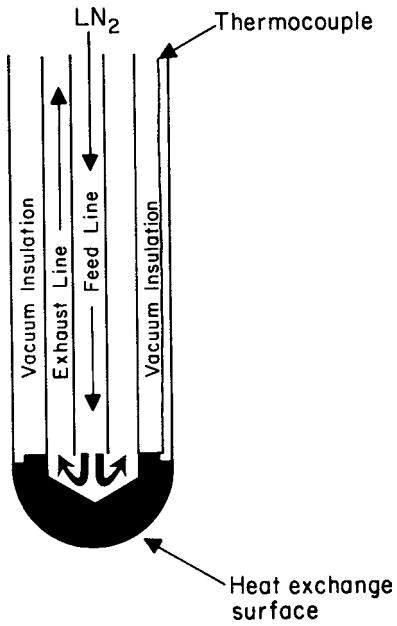


Figure 3. Diagram of a typical probe cooled by liquid nitrogen. The liquid nitrogen passes from a reservoir in the console, down the vacuum-insulated feed line, to the heat exchange surface at the probe tip. The probe tip is cooled by a change in phase (liquid to gas), then the gas is returned to the console. Probes of many sizes and shapes are available. A thermocouple may be placed in the probe tip.

range of +36 to −160 °C (2). The apparatus was soon modified to enable treatment of diverse tumors (12). In the 1990s, an advance in cryosurgical equipment technology featured the development of a technique to use liquid nitrogen super cooled to ca. −209 °C in new equipment. The super cooling was achieved by passing pressurized liquid nitrogen through a heat exchanger immersed in a liquid nitrogen chamber (−209 °C) held under vacuum. The supercold liquid nitrogen was circulated to the probe, producing probe surface temperatures in the range of −165 to −185 °C, depending on the diameter of the probe (13). Used with multiple probes, this apparatus had substantially greater tissue-freezing capability than the earlier technology. Nevertheless, this new apparatus did not improve to a significant extent on the probe cooling rate which was only ∼20 °C/min at ∼5 mm from the probe (14). The time required to reach a tissue temperature of −50 °C at a distance of 1 cm from the probe was 5–10 min (15). In comparison, the argon gas J–T apparatus cooled a probe much faster (16).

Recently, liquid nitrogen-cooled apparatus based on new technology equals the fast cooling rate of argon J-T apparatus and produces probe temperatures in the range of −170 to −180 °C. This system uses a submersible nitrogen pump to generate the operating pressure to cool the probes with liquid nitrogen. Up to six vacuum-insulated probes may be cooled simultaneously (Cryo6. Erbe Co., Tubingen, Germany).

The cryosurgical apparatus that cool by the J–T effect are lightweight, portable, and quickly responsive in cooling or warming. Pressurized gas is passed through a small

nozzle and expands, cooling the probe. This type of apparatus, using diverse kinds of gases, has been available for many years, but the cryogens commonly used in the modern apparatus are argon and nitrous oxide.

Argon, a colorless, odorless gas that boils at −185 °C, has been used in J–T type apparatus since late in the 1960s. The gas in current devices is stored in steel cylinders that are pressurized at 3000 psi. In use, the probe temperature is ca. −130 °C at coldest. The cooling efficacy is pressure dependent. As the pressure in the cylinder falls, the cooling capacity diminishes. Argon permits the use of probes of very small diameter, as small as 17-gauge needles. In treatment, the use of such small probes requires that multiple probes be placed in the lesion. A larger probe, such as 3 mm in diameter, permits greater gas flow in the conduit, and will freeze a larger volume of tissue than the needle structures. Therefore, fewer probes may be used for the same volume of tissue. Since argon is a noble gas and is normally in the atmosphere, venting the gas from the apparatus into the operating area is not a safety concern.

Nitrous oxide (−89.5 °C) is a colorless, nonflammable gas, commonly available in clinics and hospitals in the familiar "E" cylinders, which hold 2.72 kg (6 lb) of N_2O at a pressure of 5.1 MPa (740 psig). The withdrawal of the nitrous oxide gas depletes the gas pressure in the cylinder, which affects the rate of freezing. The gas cylinder may be enclosed in a warming jacket to provide some heat in order to maintain gas pressure at an appropriate level. There is a safety consideration. Cryosurgical units using nitrous oxide that do not provide for venting of the exhaust to the outside air may expose personnel to some ill effects, such as impaired performance and cognition. Such units should be used only in well-ventilated rooms. Older devices, which may exhaust into the room 20–90 L of N_2O/min, may be hazardous. Most new devices provide for gas scavenger systems to safely exhaust the nitrous oxide (Fig. 4).

In clinical use, the differences in cooling rate are important. Argon will cool a probe to −100 °C in ∼1 min, and to −130 °C in ∼2 min. Nitrous oxide will cool a probe to ca. −80 °C in the same time. In contrast, liquid nitrogen cools more slowly but will become colder, the probe temperature reaching −160 to −180 °C in ∼5 min, the final temperature depending on the engineering features of the apparatus (14–16). However, the new type of liquid nitrogen apparatus with a design based on a reciprocating bellows pump will produce fast freezing of a probe, perhaps to −180 °C in < 1 min.

Sprays of cryogen can also be provided by nitrous oxide apparatus. If the fine droplets of liquid nitrous oxide are released on a surface, the droplets, instead of vaporizing to gas, recrystallize into solid nitrous oxide particles that lie on the surface or fly in all directions. To keep this partial-pressure effect from occurring, the droplets must be surrounded by pure nitrous oxide gas, necessitating an inverted-cup shield around the applicator tip.

Carbon dioxide is a colorless gas that is used as a cryogen in solid and gaseous form. Solid carbon dioxide (−75 °C) has been used for direct application to tissue for ∼100 years. Carbon dioxide is also available as a compressed gas contained in E cylinders. In J–T apparatus, it provides probe temperatures of ca. −60 °C.

Figure 4. Modern cryosurgical device, cooled by nitrous oxide, used in cardiac cryosurgery for the treatment of atrial fibrillation and other arrhythmias. The console houses the primary gas supply in cylinders, the gas circuits, and the controlling electronics and software. The device cools the catheter probe by passing the pressurized gas through a restricting orifice at the probe tip. The probe is a steerable cryoablation catheter with leads to connect to the console for cryogenic gas flow, pressure monitoring at the catheter tip, and cardiac electrical signal recording (Fig. 3b). (Photographs by courtesy of CryoCor Inc., San Diego, CA 92121.)

Devices that cool by thermoelectric principles have limited freezing capability and can provide a probe tip temperature of ca. −20 to −30 °C at a low cooling rate. These devices have been satisfactory for ophthalmic use, but not for the treatment of tumors. The efficiency of these devices may be improved by combining with the technology of heat pipes, which then could provide probe temperatures of −50 to −60 °C (17). Nevertheless, the applicability of thermoelectric cooling to cryosurgery is limited.

TECHNIQUE

The two basic techniques for the use of cryosurgical apparatus are the direct application of the cryogen to the tissue, as in a spray of liquid nitrogen, or use of the cryogen in a closed system with probes. The choice of technique is in part a matter of personal preference and in part a matter of fitting the technique to the nature, size, and location of the disease. Probe techniques in general provide an easily controllable and a greater depth of freezing. The use of a probe is essential in freezing in less accessible areas of the body. Spray techniques, widely used in dermatological practice, permit easy application to accessible surfaces to achieve the superficial, and perhaps wide, freezing usually desired. Variants in these techniques blur the apparent sharp distinction between spray and probe techniques (Fig. 5).

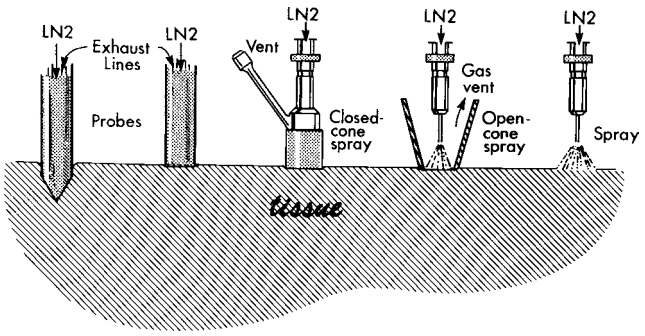


Figure 5. Diagrammatic representation of differences and similarities in probe, closed-cone, open-cone, and spray techniques, commonly used in the treatment of skin diseases. The two probes shown on the left are closed systems in which the cryogen does not come into direct contact with the tissues. In the closed systems, the liquid nitrogen, after change of phase, is vented as a gas somewhere along the return line, usually in the console of the apparatus. In the open systems, the three devices on the right, the cryogen is sprayed directly on the tissues. Surface freezing techniques by pressing a probe against the tissue or by spraying cryogen are commonly used. However, the pointed probe shown on the left may be inserted into the tissue to achieve deeper freezing. This penetration of the tissue causes a wound, which may cause some problems. In the cone techniques, the dispersion of the spray is limited and the effect of confining the spray is to produce an open probe. In the cones, the heat-exchange surface is on the skin and the venting of the gas is at the side or top of the device. The open-spray technique is effective in treating wide areas with irregular outline, but care has to be taken to avoid dispersion of the spray to unwanted areas.

The Spray Technique

Spraying of the cryogen, usually liquid nitrogen, from the nozzle of an appropriate apparatus is an efficient method of producing rapid freezing of tissue. When used as a spray, the cryogen is used at its coldest temperature, and it produces superficial freezing rather quickly and easily over wide areas if required. The more superficial the lesion, the more suitable is the use of a spray, especially if the surface is irregular, as over a bony surface, or if the area of disease is extensive. The spray often is used intermittently, especially for small lesions. Steady spraying, as in an effort to freeze deeply in one site, causes the liquid nitrogen droplets to strike the tissue without vaporizing and run off the frozen tissues to freeze in undesired areas (unless spray-limiting devices are used), especially if excessive pressure is being used in the apparatus. Other methods of control are to reduce the pressure in the spraying device, if possible, or to use a smaller nozzle size, but reduced cryogen flow results in lessened capability of freezing the tissues. Problems with the dispersion of the spray are best corrected by the use of spray-confining devices such as hollow cones placed over the lesion so that the liquid nitrogen may be sprayed into the hollow device (Fig. 6). The use of the cone devices also has the effect of creating an open probe and is a useful technique of improving depth penetration of spray techniques, as required for the treatment of invasive skin cancers. Similar devices, such as funnels and hollow cylinders, have been used to confine liquid nitrogen as it is poured over bone tumors after removal by curettement.

The Probe Technique

The cryogen is circulated in a closed system, using metal probes for contact with the tissue to provide a heat sink or heat-transfer surface. Various sizes and shapes of probes, ranging from rod-like probes 1 cm in diameter to 18 gauge needle size and catheter shapes, to fit diverse anatomical sites and disease dimensions are available. The probes may have flat surfaces for contact with the tissue or may be pointed to allow insertion into the tissue. Slip-on metal end pieces, which fit over the end of the probe to modify the freezing surface, increase the versatility but also slow the freezing capability. The lines that feed the liquid nitrogen to the probe tip are vacuum insulated or insulated with appropriate materials, which increases the efficiency of the cryogen and provides for the safety of the user. The J–T types of apparatus have thin cryogen feed conduits that require no insulation.

In use, the physician selects one or more probes as appropriate to the disease. The manner of use is to apply the freezing surface of the cryoprobe to the lesion and allow the cryogen to flow. The cold probe acts as a heat sink and produces tissue freezing by removing heat from the tissue faster than blood supply and conduction restores it. Surface contact freezing is performed by pressing the freezing surface of the probe firmly on the tissue in order to ensure a good contact for heat exchange. This contact is improved by the use of water-soluble hospital lubricating jelly between the probe and the tissue. Greater depth of freezing may be achieved by increased pressure on the probe or by penetration of a sharp, pointed probe into the tumor. The

Figure 6. Canine liver being frozen with a large probe, 1 cm in diameter cooled with liquid nitrogen to −160 °C. The white frosted area is the tissue frozen after 3 min of contact. The depth of freezing is about the same as the lateral spread of frost from the side of the probe. With proper technique, the area of necrosis will closely approximate the visibly frozen area. (Reprinted with permission from J. Am. Med Assoc., **204**, 566, 1968.)

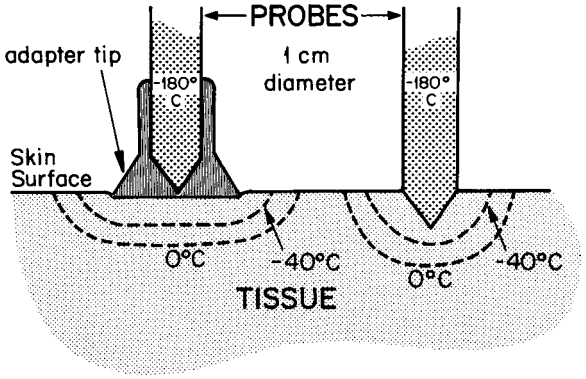


Figure 7. Diagram showing the effect of probe configuration on the shape of the frozen tissue. The probe on the right held with some pressure on the tissue causes indentation of the tissue, or minimal penetration, and produces a roughly hemispheric frozen zone. The same probe, fitted with a freezing tip adapter to form a wide, flat freezing surface, produces a wider but less deep frozen area. In each frozen zone, the −40 °C isotherm is shown to delineate its approximate location. The location of this isotherm varies slightly with the rate of cooling the tissue. Rapid cooling moves the isotherm slightly more toward the periphery.

penetration technique has the disadvantage that a wound is produced, and this may later bleed. Care must be taken to avoid motion of the probe during freezing because this might fracture the bond with the tissue and interfere with heat exchange. Heat exchange also depends on the gradient of temperature between the tissue and the probe, so the probe is always used as cold as possible when tissue destruction is sought. The larger the gradient, the faster the rate of freezing.

Freeze–Thaw Cycles

As freezing progresses, the tissue turns white (frosted in appearance) and hard, a change that begins at the area of contact with the cryogen and extends to incorporate an increased volume of tissue as time passes (Fig. 6). As treatment continues, in the case of lesions in easily accessible areas, such as the skin, the extent of freezing is judged by inspection and palpation. Estimation of depth of freezing may be difficult, but physicians experienced in the techniques can make reasonably accurate estimates of depth. With surface contact freezing by probes, the shape of the frozen volume of tissue is roughly hemispheric, so the depth of freezing can be judged to be about the same as the lateral spread of freezing from the probe. Significant modification in the shape of the frozen area may be achieved by the selection of different probe shapes (Fig. 7). The shape of the frozen area is also modified by the amount of pressure placed on the probe because increased pressure compresses the tissue and increases the depth of freezing. Another factor influencing the shape of the frozen area is the presence of major blood vessels, which provide a source of heat. In freezing with the spray techniques, a more superficial freezing may be expected and often is desired. However, if the spray is confined with a cone, the effect on the tissue is similar to probe freezing.

Freezing with the probe or spray continues until the desired volume of tissue, including a margin of apparently normal tissue, is treated or until the frozen area no longer enlarges. This is easy to observe in accessible tissues, such as the skin. In the use of a spray of liquid nitrogen, the nozzle of the device is moved about to distribute the cryogen evenly over the target, and this freedom of movement allows wide areas to be treated. On the other hand, probe freezing takes place from a selected site of application, and the rate of expansion of the frozen area slows as an equilibrium is established between the heat loss from the tissue via the probe, and heat brought to the area by the circulation of blood. For this reason, it is difficult to freeze tissue to a distance $>$ 2–3 cm from the probe. This means that large lesions cannot be frozen completely in a single application of the probe. In these circumstances, the plan of treatment must include freezing from multiple sites with successive applications of the probe, which improves the width of the frozen area and also increases the depth of freezing in overlapped frozen areas to a slight extent. For large cancers, the simultaneous use of multiple probes is advantageous and time saving.

When used for the destruction of tissue, as for the treatment of tumors, cryosurgery must be performed in a manner that produces a predictable area of necrosis. The techniques to achieve this goal stress the rapid freezing of the tissue, slow thawing without assistance from warming devices, and immediate repetition of the freezing process in order to maximize destruction. Some modifications of these basic factors in technique are necessary for the diverse diseases in different parts of the body, especially if the intent is to destroy some cells while preserving others, but, in general, the cited basic technique provides the basis of effective therapy.

Rapid freezing promotes the formation of intracellular ice crystals, which are considered almost certainly lethal for cells. This occurs only in the tissue close to the freezing probe or in contact with the spray of cryogen. The rate of freezing of the tissue varies inversely with the distance from the site of application of the cryogen (Fig. 8). Close to the probe, temperature changes of the order of $> 10\,^{\circ}C$/min may be achieved, and this is considered rapid freezing. However, $\sim$2 cm from the probe, the cooling rate is slow, perhaps of the order of 2–5 $^{\circ}C$/min. The possibility of tissue survival is enhanced with slow freezing rates, but fortunately even slow freezing rates have lethal potential because of cellular dehydration and related deleterious effects.

The thawing rate should be slow and unassisted because this increases the time that the tissues spend in a phase when recrystallization phenomena can add to the cell injury. Tissue often thaws in about the same time as was required for freezing, but this depends on the volume of frozen tissue. Large volumes of frozen tissue warm slowly. In some therapeutic applications, a probe heater is used to speed release from the tissues after freezing. This practice also slightly speeds thawing of the tissues, but this probably does not increase the chance of cell survival if the technique is otherwise correct. Certainly, no warming solutions should be used to thaw the tissue, because it is well established that quick thawing of tissues reduces injury from freezing, whether in cryopreservation, cryosurgery, or frostbite.

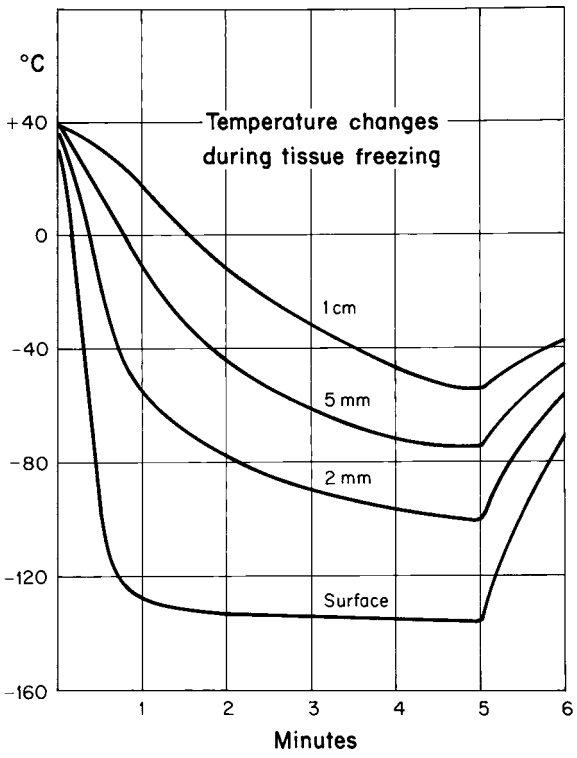


Figure 8. Graph showing temperature changes recorded from thermocouples inserted in the tissue at diverse distances from a probe cooled by liquid nitrogen. The tracing identified as "surface" is from a thermocouple placed in the contact area between the probe surface and the tissue. The other tracings show the temperature changes at 2, 5, and 7 mm from the edge of the probe. The most rapid and deepest cooling of the tissue is at the point of contact. The greater the distance from the probe, the slower the cooling, the less the depth of freezing, and the shorter the duration of freezing.

Repetition of freezing maximizes damage by subjecting the tissues again to the same mechanism of thermal injury. To take full advantage of the lethal effect of repetitive freezing, it is necessary that the frozen tissues be completely thawed ($> 0\,^{\circ}C$) before the next freezing cycle because then the entire volume of the tissue will pass through recrystallization phases with their attendant injurious effects. The second freezing cycle is usually faster and more effective in cooling the tissue than the first cycle and achieves a slightly greater depth of freezing because the latent heat in the previously frozen tissue is reduced and because the microcirculation has begun to fail in the interval between freezings, reducing the heat supply. The lethal isotherm is then deeper in the frozen tissue. For this reason, it is advisable to wait a few minutes between freezing–thawing cycles.

The freezing of tissue in cryosurgery is a heat-transfer process, and the mechanics of extraction of heat from the tissue directly influence cryosurgical technique. The duration of each freezing–thawing cycle of treatment is dependent on factors that affect the rapidity of tissue freezing, such as efficiency of heat exchange, gradient of temperature, blood supply to the area, attainment of a desired temperature goal, and on the size of the lesion. The temperature of the cryogen is also important: the larger the gradient between cryogen and tissue, the faster the cooling. It is also a function of the contact area between cryogen and tissue: a wide area for heat exchange cools tissues faster.

Thermal modeling techniques have been used to predict the size of the frozen area that may be produced by a probe. The blood flow, the probe temperature, the area of contact with the tissue, and the duration of application are known. This work has contributed to the understanding of the functions and cooling capacity of cryosurgical equipment, has provided a method of estimating the effect on the tissues, and has shown direction in equipment design and in the planning of treatment. The modeling techniques have been useful in the several aspects of cryobiology, including the mechanisms of injury to cells. Nevertheless, before clinical use, it is necessary for the physician to practice with the cryoprobes or with a spray of cryogen in test materials in order to develop the ability to predict the size and shape of the frozen field as a function of time, temperature gradient, and heat exchange surfaces. The heat diffusion equations that must consider the frozen area and a moving solid–liquid interface as freezing progresses are complex and are best reviewed in source material (5,18).

MONITORING TECHNIQUE

Many cryosurgical procedures, especially those for benign diseases and small cancers of the skin, are performed using only observation and palpation of the frozen tissue to determine the progress of treatment and judge its adequacy. In

easily accessible sites, those physicians with considerable experience in cryosurgery can achieve satisfactory results without the use of monitoring techniques to guide therapy. However, the temperature of frozen tissue cannot be determined by its appearance: Frosted tissue looks the same at any freezing temperature. Equally important, the depth of freezing is difficult to judge in many situations, although the relationship between the depth of freezing and the lateral spread of freezing from a probe is an important clinical aid. The clinical evaluation, if not perfect, may produce an error in treatment that may be of critical importance in the treatment of malignant disease. Effective cryosurgery must yield predictable and certain necrosis.

To guide cryosurgical procedures and permit reasonable certainty of the death of tissues, methods of monitoring the freezing process have evolved. These methods include (1) the measurement of tissue temperature by thermocouples; (2) the measurement of electrical impedance on resistance in tissue; (3) the measurement of heat lost from the tissue by a heat flowmeter; (4) thermography and; (5) the imaging techniques, which are ultrasound, computerized tomography (CT); and magnetic resonance (MR).

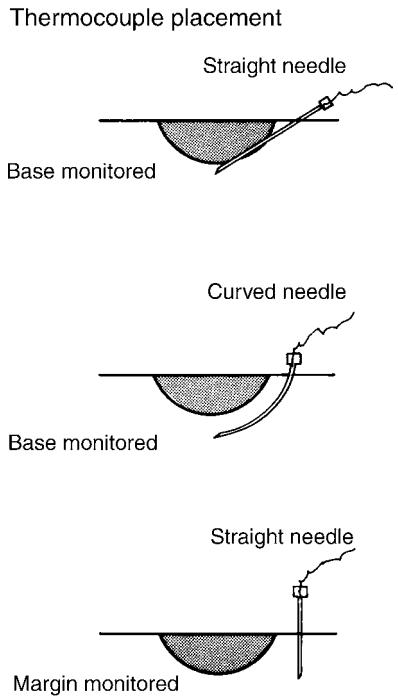


Figure 9. Diagram showing typical sites of thermocouple insertion. Three useful methods are shown. At the top is a thermocouple inserted at an angle from the side of the lesion through normal tissue. The thermocouple tip rests directly beneath the lesion and shows the temperature at the depth of the tissue. It probably is the most common method of thermocouple usage. In the center is shown a curved needle used for base monitoring. This method avoids passing the shaft of the thermocouple through the frozen tissue. At the bottom is shown an alternative method of thermocouple use. The thermocouple is inserted at the border of the lesion. With this technique, the temperature registered at this thermocouple is interpreted as being the same temperature that would be measured at the border beneath the tumor. This method assumes that the depth of freezing is approximately the same as the lateral spread of freezing from the probe. The advantage of this technique (and the use of a curved thermocouple) is that the thermocouple shaft remains outside of the frozen area until the advancing ice front incorporates the tip. More than one thermocouple may be used.

Thermocouples

The most commonly used method of monitoring is by the insertion of needle-mounted thermocouples into the tissue at appropriate sites to measure tissue temperature (Fig. 9). Thermocouples are formed by the junction of two dissimilar conductors, commonly iron and constantan, or copper and constantan, in a closed electric circuit (Fig. 10). When the junction of the conductors is held at a temperature, an electromotive force (emf) proportional to the temperature difference will be generated. An instrument, such as a potentiometer or pyrometer, is used to measure this emf and provide a readout in terms of temperature.

Thermocouples perform several important functions in cryosurgical treatment. The insertion of thermocouples into the tissue in appropriate locations monitors the progress of the freezing treatment and ensures that temperatures destructive for tissues are attained. Thermocouples can also be used to ensure that tissues adjacent to a diseased area are not frozen and are hence safe from freezing injury. If used with a recorder, thermocouple measurements on the tracing provide written evidence of treatment. Thermocouples often are built into the probe tips, so that the temperature of the probe can be monitored. Though this provides useful information, confirming that the apparatus is working, probe temperature tells nothing about tissue temperature, except by estimations based on experience with the performance of the probe.

In cryosurgical treatment, especially for cancer, it is important to know that destructive temperatures are produced in the tissues. In the freezing–thawing cycles used in cryosurgery, tissue destruction is a multifaceted process involving the freezing rate, the tissue temperature, the duration of freezing, the thawing rate, and repetition of the freezing–thawing cycle. Separation of the relative destructive effects of these components is difficult, but the easiest to control and measure is the coldest temperature attained

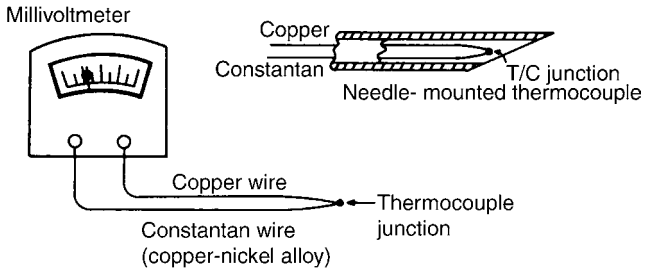


Figure 10. Diagram of a basic thermocouple-millivoltmeter circuit consisting of a copper and a constantan wire welded together to form a measuring junction, the thermocouple. The free ends are connected to the millivoltmeter, which measures the emf associated with the temperature at the thermocouple. In the upper right of the illustration is shown the mounting of the thermocouple in a hypodermic needle, which can be inserted in the tissue at appropriate sites to measure temperature changes during freezing–thawing cycles.

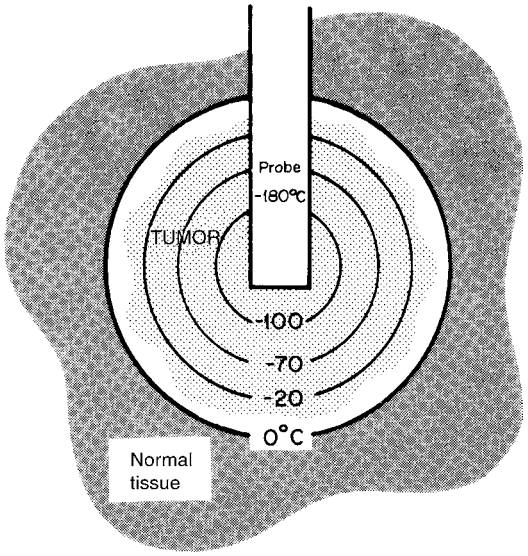


Figure 11. Drawing demonstrating the gradients in temperature that exist in tissue being frozen with a probe at −180 °C. The incorporation of a tumor in a frozen mass is depicted. The gradients are substantial, ranging from ∼0 °C at the edge of the frozen volume to nearly −180 °C adjacent to the probe. Cell Death would not occur in the entire frozen area. If treatment were to cease at this stage, the tissue in the 0 to −20 °C range would probably survive and the tumor would grown again.

in the cryosurgical treatment. Hence, a lethal temperature goal is used in cryosurgery.

Early in cryosurgical experience, it was thought that −15 or −20 °C was a proper lethal temperature goal. Substantial tissue damage results from a tissue temperature of −20 °C, but this temperature is not safe for the treatment of malignant disease. In nonneoplastic disease, usually conservative freezing is wise and, therefore, tissue temperatures of the order of −20 to −30 °C are satisfactory. Temperatures of −40 °C are satisfactory for superficial skin cancers. The treatment of more aggressive cancers, as in the oral cavity, requires repetitive freezing to tissue temperatures of −50 °C or colder if cure is to be achieved (Fig. 11).

The accuracy of thermocouples is a matter of interest since treatment depends in part on this measurement. Some minor error is inherent in the temperature recording system, which consists of the thermocouple, the electrical leads, and the readout device. The readout device is commonly a potentiometer with a digital readout, which is sufficiently accurate for the purpose of cryosurgery. An important source of error is from conductance of heat along the thermocouple needle shaft. If the needle shaft passes through a frozen area, the reading from the tip may be falsely low (Fig. 12). Under other conditions, heat may be added to the temperature-measuring system from extraneous sources anywhere between the thermocouple needle tip and the recorder. However, with proper use, the error due to heat conductance is minimal and does not interfere with the thermocouple function of supplementing clinical judgment. Based on experimental data, temperatures in the range of −20 to −50 °C, produced in short freezing cycles, may be associated with an error of ∼2 °C due to heat

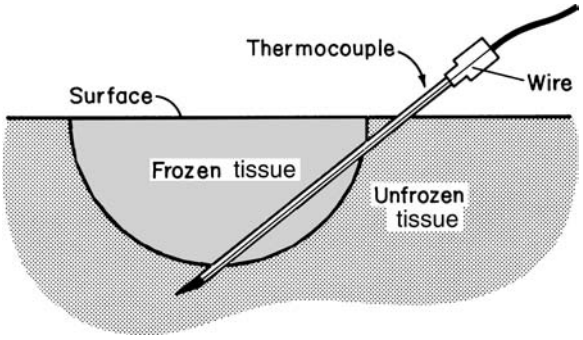


Figure 12. Diagram illustrating a mechanism for thermocouple error due to heat conduction. The thermocouple needle enters the tissue at an angle, just as commonly done in clinical practice. During freezing, the expanding ice front incorporates the needle shaft. The cooling of the shaft affects the thermocouple at the needle tip, and this may produce a falsely low reading. Also, the hub of the needle and the adjacent needle shaft in unfrozen tissue may be warmed by the tissue or the air, which may affect the reading in the opposite way. The magnitude of this possible error is a source of concern, but proper thermocouple use can avoid the error.

conductance (19). This differential is of little importance in cryosurgical techniques. The error due to heat conductance in thermocouples is less significant than the error produced by the positioning of the thermocouples in the tissue. A 1-mm variation in thermocouple placement in the tissue represents ∼10–15 °C difference in the temperature recorded in usual cryosurgical freeze–thaw cycles. Therefore, the important errors in thermocouple use are produced by the accuracy of placement rather than by heat conduction from extraneous sources.

Impedance–Resistance Measurements

Another method of quantification of freezing injury is a technique that measures the impedance or resistance changes to the passage of a small electric current through the tissues being frozen. Unfrozen tissue is a conductor of electricity because of the electrolyte content of the tissue fluid. During freezing, the formation of ice crystals in tissue and the removal of water from the tissue results in decreased electrical conductivity. When practically all of the extracellular water is crystallized, electrical impedance or resistance rises to the high levels. This change is interpreted as being associated with tissue death.

The techniques of impedance–resistance measurements require the insertion of needle electrodes in or about the tumor (Fig. 13). These conduct the small electric current from the line or battery-powered device to the tissue. The measurement is made between two electrodes. Both electrodes may be placed in the treated area, but it is preferable to measure between one electrode in the treated area and a distant reference electrode. The initial impedance–resistance in unfrozen tissue is of the order of 1–2 kΩ. When the entire measuring electrode is incorporated in frozen tissue, the impedance–resistance rises to megohm levels. This change seems to occur quickly, at least in comparison to the changes in temperature (Fig. 14).

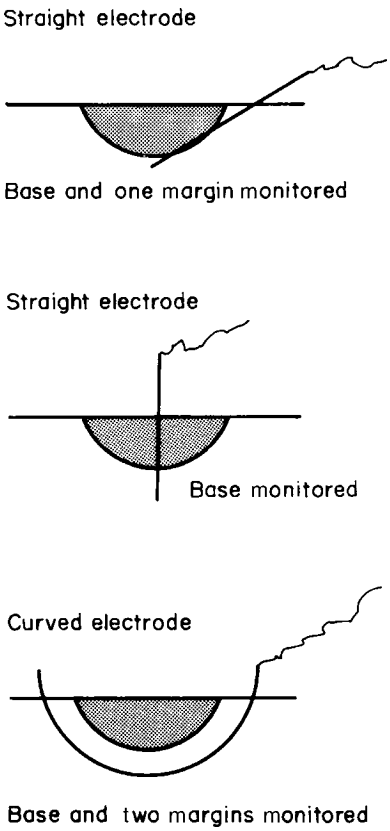


Figure 13. Diagram of electrical impedance–resistance monitoring system showing alternative methods of placement of the electrode. The entire electrode must be incorporated in the frozen tissue before the change in impedance–resistance rises to the megaohm level. Multiple electrodes may be used to increase the monitoring sites. The second electrode, which is placed in a more remote site to complete the electrical circuit, is not shown.

Impedance measurements were introduced into cryosurgical techniques by Le Pivert et al. (20), who measured electrical impedance in the tissue to the passage of a low frequency (1000 Hz) alternating current between electrodes. Impedance values of 0.5–1.0 MΩ were equated to −40 °C and were associated with tissue destruction. This destructive effect on the tissue was reported to be independent of repetition of the freezing–thawing cycles.

Other investigators, using Le Pivert's instrument in experiments on canine skin, found that an impedance of 1 MΩ corresponded to a tissue temperature of ca. −30 °C, but the range of temperatures about each impedance value was sufficiently great to cause concern about the possibility of tissue survival at the 1 MΩ value. Using a line powered 1000 Hz low current impedance meter, the relationship between tissue impedance and temperature and the subsequent development of tissue necrosis was investigated. An impedance of 10 MΩ was not always associated with tissue death, and the range of temperatures about any impedance value was considerable (Fig. 15). Comparison of the electrical characteristics of the tissue and the tissue temperature with the border of necrosis that was obvious in a few days showed that the tissue temperature was the more accurate predictor (21).

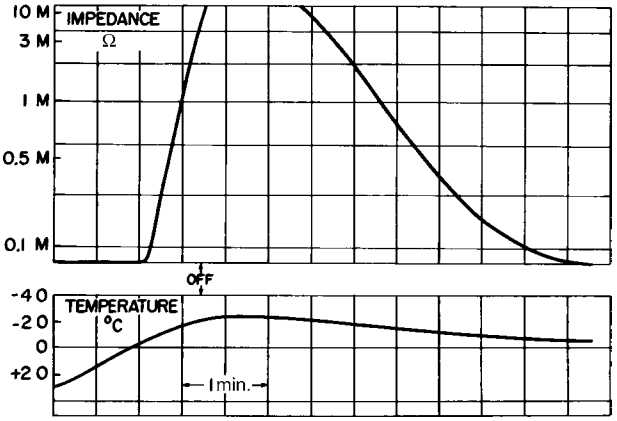


Figure 14. Graph illustrating simultaneous temperature and impedance changes. The temperature in the tissue was cooled to −22 °C at which time the cryosurgical apparatus was turned off. The tissue cooled further to −25 °C because the cold probe continued to function as a heat sink. Then the warming cycle began. The impedance rose steadily after the tissue cooled to −8 °C. When the impedance reached 10 MΩ ~40 s later, the cryosurgical apparatus was turned off. The impedance continued to rise and the trace passed off the chart, returning ~1 min later in the warming cycle. The temperature and impedance changes in thawing were more gradual in thawing than during freezing.

At this time, devices to measure tissue impedance–resistance during cryosurgical procedures are little used. The incorporation of an impedance electrode into the freezing surface of an endoscopic probe confirms that the probe is cooling properly, but yields no information about tissue temperature (22). Nevertheless, such a device would confirm fixation of the probe to the tissue. An impedance-measuring device could be useful in the determination of depth of freezing. The use of multiple electrodes about the circumference of a tumor to monitor cryosurgery is a use that requires further testing.

Whatever method of quantification of cryogenic injury is used to supplement clinical judgment in cryosurgery, whether thermocouples or electrodes are inserted in the

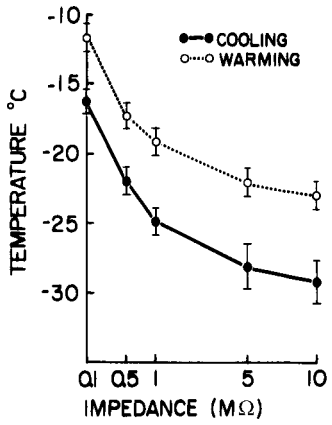


Figure 15. Graph showing the relationship between impedance and temperature. The data for the graph are taken from mean values during the cooling cycle. The standard error of the mean is shown.

tissues, it is important to recognize that their placement in itself may be a source of error. It is equally difficult to determine the relationship of the electrode or thermocouple to the margin of the disease. Both thermocouple and impedance devices should be viewed as adjunctive to clinical judgment in therapy.

Heat Flowmeter

The heat lost from the tissue can be measured by means of a heat flowmeter (23). This is done by attaching a differential thermocouple to the probe tip. As heat passes from the tissue into the probe, the temperature electromotive force, proportional to the temperature differential, develops across the disk. This may be recorded so that total heat exchange is used as the monitor of the cryosurgical procedure. Though the technique was used in the treatment of patients, it has not proven useful generally.

Thermography

The use of thermography to monitor the progress of freezing tissue is best considered as a research method. The thermograms, recorded in monochrome or in color, define isotherms down to $-40\,^{\circ}$C. Bradley's experiments with thermography, comparing the freezing characteristics of different freezing techniques, have shown the faster freezing capability of the spray devices in comparison with the probe techniques and have confirmed the importance of pressure on the tissue from the probe in modifying the shape of the frozen area. Though *in vitro* tissue testing permits evaluation of depth of freezing, clinical use provides only surface freezing evaluation in most circumstances (24,25).

Ultrasound

Ultrasound used during cryosurgery provides a real-time image of the frozen volume of tissue. Since ultrasound provides a more global view of the frozen tissue than do thermosensors, this imaging technique has come into wide use for monitoring the freeze–thaw cycle, especially for visceral tumors. Ultrasonography offers the possibility of matching the extent of the neoplasm with the volume of tissue frozen in treatment. Frozen tissue is hypoechoic, so the ultrasonic image is black. The edge of the frozen tissue is hyperechoic and appears as a bright line. As freezing continues and the volume of frozen tissue increases, the hyperechoic rim moves away from the probe, leaving the hypoechoic zone behind it. Therefore the process of tissue freezing can be observed during the cryosurgical procedure. Experience with ultrasound monitoring in hepatic and prostatic cryosurgery is now substantial. The correlation between the ultrasound image and the actual diameter of the frozen tissue is excellent, though it has become evident that ultrasound overestimates the volume of tissue frozen. Much of the frozen volume is obscured by distortions and reflections of the image. Sonography does not provide an image beyond the near edge of the ice because of complete posterior acoustic shadowing (26,27). Some compensation for this limitation can be obtained by viewing the frozen volume from another angle. Three-dimensional (3D) ultrasound is in development (28).

The correlation between tissue temperature and the ultrasound image is of considerable importance because effective treatment depends on achieving an appropriate tissue temperature goal. The temperature of frozen tissue cannot be determined from its ultrasonic image. However, the hyperechoic rim of the image is $\sim 0\,^{\circ}$C and the probe temperature is generally known, so inferences about the steep gradients of temperature in the tissue can be made. For example, using a probe at $-160\,^{\circ}$C on tissue for 5 min, the $-20\,^{\circ}$C isotherm is $\sim$70% of the distance from the probe to the periphery of the frozen tissue. The $-40\,^{\circ}$C isotherm, which is commonly used as a goal in tumor therapy, is $\sim$60% of the distance from the probe to the frozen boundary. A rapid cooling rate moves the isotherm toward the periphery.

Computerized Tomography and Magnetic Resonance

Other investigational methods of monitoring are radiological. The absorption of X rays in biological tissue is proportional to the tissue density. Water, which is the major component of most biological tissues, changes density upon freezing and can be visualized. Computerized tomography (CT) can show the entire cross-section of the lesion. The development of the frozen volume of tissue can be seen on a series of CT images taken at frequent intervals. Magnetic resonance (MR) will provide a 3D image of the frozen volume of tissue. When used with thermal models, the temperature changes in the tissue can be quantified (29,30). Magnetic resonance requires the use of MR-compatible probes, which have been developed and used to a limited extent. Electrical impedance tomography (EIT) has been proposed as a method to provide real time imaging and provides a global image by introducing low amplitude alternating current (ac) into the body and thereby measurs the electrical potentials on the surface of the body. These potentials are analyzed to create a tomographic image (31). At present, the high cost of apparatus and logistical considerations will limit the usefulness of these imaging techniques in cryosurgery. However, it seems likely that image guidance will achieve greater applicability as monitoring techniques during the freezing process.

To summarize, clinical judgment, the prime factor in the control of the freezing of tissue, requires assistance from the monitoring techniques. Depending on the clinical circumstances, including the nature and the site of the disease, either tissue temperature measurement by thermosensors or ultrasound imaging is commonly used. The limitations of these techniques in providing thermal information during cryosurgery are well defined. To compensate for the limitations, wherever practical, the use of both thermosensors and imaging techniques are advised.

CLINICAL USES

Cryosurgery has established a firm place in medical practice for the treatment of diverse diseases in different parts of the body. Lesions in accessible sites, such as the skin or

oral cavity, may be treated with little need for anesthesia and without risk of hemorrhage. Treatment for many conditions can be given under local anesthesia in an office or outpatient clinic, avoiding the need for hospitalization, with its attendant inconvenience, cost, and risk. Cryosurgical treatment of viscera and other deep-seated disease is more difficult because of accessibility, but these techniques have become well developed in the past 15 years, benefiting from percutaneous technology and ultrasound imaging. Cryosurgical techniques deserve emphasis as an excellent choice of therapy on high surgical risk patients, especially those who have problems difficult to manage by other methods of treatment.

Skin Diseases

Cryosurgery is widely used in the treatment of skin diseases. Many non-neoplastic lesions, and a variety of benign tumors and cancers of the skin are successfully treated commonly with a hand-held device containing liquid nitrogen. Non-neoplastic skin lesions are easily treated by cryosurgery with little risk, but skin cancers require further comment.

Cryosurgery has become a standard technique of treatment for skin cancer in any location, and joins a variety of treatment methods, including surgical excision or radiotherapy as well as other special techniques. Most skin cancers are small, commonly not > 3 mm in depth and < 1 cm in surface diameter. These are easily treated by cryosurgery, and a cure rate in excess of 97% has been achieved. Certain types of skin cancer, called sclerosing or morphea type of basal cell carcinomas and tumors of the scalp, are difficult to cure by cryosurgery, perhaps because of difficulty in determining the extent of disease, or because of biologic behavior, or because of the richness of blood supply. These require aggressive freezing therapy in order to achieve a cure. In general, however, the histologic type of the cancer is not important, and either squamous cell or basal cell cancers can be treated. Melanomas are also easily destroyed by freezing, but most primary melanomas, with the possible exception of lentigo maligna, are better treated by excision in order to establish the diagnosis and to permit staging of the extent of the disease. Cryosurgery is well chosen for certain skin cancers in special situations. These include the following:

1. Multiple superficial small skin cancers, as commonly seen from excessive exposure to sunlight. A large number of cancers can be treated quickly and easily in comparison to multiple surgical excisions.
2. Cancers about the ears, nose, and eyes. The irregular contours of the head in these sites and the tightness of the skin over the underlying bone make treatment difficult at times. In these locations, cryosurgery offers ease of treatment and improved cosmetic results.
3. Cancers arising in irradiated skin. Such cancers require conservative management because new cancers will develop elsewhere in the damaged skin.
4. Cancers that persist after radiotherapy excision or other methods of treatment. These are often advanced cancers that present problems in management.

In these special situations, the results of cryosurgery compare favorably with surgical excision or radiation therapy.

Oral Diseases

Benign and malignant tumors and precancerous conditions of the oral cavity are well suited to treatment by cryosurgery (Fig. 16). An important use is for the treatment of leukoplakia, which is the general descriptive term for white patches in the mouth and includes a number of different pathological conditions. Some of these, such as epithelial dysplasia and papillary epithelial hyperplasia, are precancerous. A biopsy, preliminary to definitive treatment, is necessary to differentiate between those lesions that are relatively innocuous, those that are precancerous, and those that are carcinoma *in situ*. All may be treated by cryosurgery. Such disease is superficial, and either nitrous oxide or liquid nitrogen apparatus may be used because the tissue needs to be frozen only to a depth of $\sim$3 mm. The results are excellent, and most patients will remain free of disease. Persistent disease or recurrent disease usually results from failure to correct the etiologic factors, that is, continuation of the use of alcohol and tobacco or continued irritation of the oral mucosa for other reasons. In dysplastic disease or carcinoma *in situ*, $\sim$10% of patients will require additional treatment over a 2-year period.

Oral Cancer

Malignant tumors of the oral cavity are aggressive cancers that are difficult to cure by conventional treatment, that is,

Figure 16. Photograph showing the treatment of a benign blood vessel tumor (hemangioma) by freezing *in situ*. A cryoprobe, 5 mm in diameter, is applied. The frozen zone is extending over the lesion. Freezing will continue until the entire tumor is encompassed. No thermocouple monitoring is necessary with this benign lesion.

surgical excision and/or irradiation. In general, the 5-year survival rate is only ~30% of all patients who acquire the disease. Cryosurgery can be used in special circumstances, especially for high surgical risk patients whose extensive associated disease would make general anesthesia and extensive operation prohibitive in risk. Cryosurgical techniques are also suitable when the cancer is on or adjacent to bone because the underlying bone limits the depth of penetration of a tumor and facilitates its destruction by freezing. In selected patients, the survival rate achieved by cryosurgical treatment appears comparable to that provided by surgical excision. However, the result can be achieved at a lessened cost to the patient in terms of operative mortality and postoperative functional disability. Nevertheless, cryosurgery is seldom used for oral cancer in current surgical practice.

The further away from direct vision, as is possible in the oral cavity, the more difficult is the application of cryosurgery. Cryosurgery has been used for cancers in the pharynx, larynx, bronchi, and esophagus with treatment given via endoscopy. Vision is limited, so accurate and extensive freezing is difficult. In the pharynx and larynx, occasional successes in attempts at cure have been achieved, but current experience provides little reason to choose cryosurgery in preference to radiotherapy or excision. Experience in the trachea and esophagus is limited to a few patients treated for palliation of symptoms. Obstruction can be relieved and tumor size kept under control by cryosurgery repeated every few months, but invasive growth continues. The presence of necrotic tissue in the larynx is a threat to the airway. In general, palliation of incurable cancer is provided better by radiotherapy or chemotherapy.

Bronchial Tumors

Considerable experience has matured in recent years with the cryosurgical treatment of bronchial tumors, including cancers, which produce symptoms by obstruction of the air passages. The treatment is via endoscopy, freezing the tumor with nitrous oxide-cooled probes, which then opens the airway and relieves the symptoms. With the removal of the obstructing tumor, irradiation of the remaining tumor may be used more safely (32).

Nose and Throat Diseases

Cryosurgery has been used for a wide variety of diseases of the nose and throat, including mucosal dysplasia, tonsillitis, nasal polyps, and rhinitis. However, generally other methods of treatment are chosen for these conditions. Special probes and techniques are necessary, and commonly the treatment can be done under local anesthesia. In chronic nasal airway obstruction due to hypertrophy of the nasal turbinates, freezing of the excess nasal tissue is followed by slough of the hypertrophic tissues, which improves the nasal airway and decreases the secretions. The treatment should be conservative so that normal tissue is not unnecessarily frozen. Healing occurs over a 2- or 3-week period and results in a reduced amount of turbinate tissue with a normal appearing mucosa and relief of symptoms.

Cryosurgery may be used for tonsillectomy. Cryosurgery is well chosen for adult patients who have blood dyscrasias or who are high surgical risks, because cryosurgery may be performed with little or no blood loss and low risk of postoperative bleeding. Nitrous oxide or liquid nitrogen apparatus may be used. The chance that tonsillar remnants may remain after freezing, which may lead to persistent symptoms, means that careful attention must be given to technique to ensure that all tonsillar tissue is frozen. Repetitive treatments may be used to eliminate tonsillar remnants.

Gynecological Diseases

Cryosurgical treatment has become common in gynecological practice in recent years. A wide variety of inflammatory and neoplastic diseases of the vulva, vagina, and cervix may be treated by freezing *in situ*. Since tissue diagnosis is important, it must be recognized that advances in colposcopy have improved the diagnostic ability of the physician to differentiate between inflammatory and neoplastic diseases.

Chronic cervicitis, which is inflammatory disease of the uterine cervix, is a principal use for cryosurgery. Careful evaluation of the extent of the disease must exclude the possibility of invasive carcinoma. Cryosurgery is effective treatment for dysplastic disease and carcinoma *in situ*. The differentiation between inflammatory disease and premalignant or malignant disease must be made by pap smear or biopsy before treatment is begun. The results of therapy are excellent, but if the disease was carcinoma *in situ*, persistent disease must be expected in ~9% of patients.

Cryosurgical techniques are seldom used in the treatment of invasive malignant disease of the vulva, vagina, and cervix. Though it has been used for cancer in some potentially curable patients who are not candidates for one reason or another for conventional therapy, sometimes with excellent results, too few patients have been treated to permit evaluation of its ability to cure cancer in these sites. More commonly, cryosurgery has been used to relieve pain due to cancer after other methods of treatment have failed. Under these conditions, cryosurgery will diminish the size of the cancer, reduce malodorous discharges, control bleeding, and ameliorate pain.

Proctological Diseases

Cryosurgery has been used to treat a variety of proctological disorders, ranging from relatively minor conditions such as anal fissures and hemorrhoids to large, incurable cancers. However, anal fissures and hemorrhoids may be treated successfully by a variety of different surgical methods. For example, injection of sclerosing agents into small to moderate sized hemorrhoids is a commonly used effective therapeutic method. These alternative techniques are good therapeutic methods, so cryosurgery has little place in the treatment of non-neoplastic proctological diseases.

Cancers of the anus and rectum may be treated by cryosurgery in selected patients who are at high surgical risk for conventional surgical excision. Best suited for treatment are carcinomas that are exophytic rather than ulcerated and infiltrating. In the rectum, the ideal cancer

for cryosurgery is within reach of the examining finger and on the posterior and lateral walls where freezing is safer. Similar criteria have been used for selection of patients for therapy by electrocoagulation. It is best to reserve cryosurgery for patients who are difficult problems in management by excisional surgery.

Diseases of the Prostate Gland

Cryosurgical treatment of prostatic disease was introduced into clinical practice in the mid-1960s, had extensive trial in the 1970s, and then fell into disuse in the 1980s. However, after intraoperative ultrasound became available and when improved cryosurgical equipment was developed, then cryosurgery for prostatic cancer became a viable alternative to excisional surgery. The technique requires the passage of multiple thin probes, as few as six, as many as twenty, through the skin of the perineum into the prostate gland. The probes are placed in an appropriate spatial relationship to include the entire gland during the freezing. Then the probes are cooled, using pressurized argon in a J–T apparatus or using liquid nitrogen. The process of freezing is monitored by ultrasound imaging and commonly also by placement of thermosensors in the peripheral areas of the prostate (33). The cryosurgical treatment is suitable for most stages of prostatic cancer, including those patients who have persistent disease after irradiation. The long-term beneficial results are similar to those of excisional surgery.

The effect of prostatic cryosurgery on metastatic disease is controversial. Regression of metastatic deposits of prostatic cancer following repeated freezing of the prostate has been attributed to the release of tumor antigens from the frozen tissue and the subsequent development of prostatic tumor-specific antibodies. In patients with bone metastases, relief of bone pain often is achieved for several months, but radiological evidence of reduction of metastatic tumor is uncommon. In experimental animals bearing tumors, cryosurgery has been shown to produce an immune response, manifest in a lower incidence and a slower growth of recurrent tumor. Though objective evidence of benefit from a humoral or a cell-mediated immunological response in men following prostatic cryosurgery is lacking, considerable interest in the possibility of benefit by this mechanism is evident.

Visceral Tumors

In addition to the prostate gland, tumors in diverse organs have been treated by cryosurgery. In general, the tumor requires operative exposure, which can be acquired by minimally invasive percutaneous techniques in some cases. The freezing is done with ultrasound monitoring whenever feasible.

Tumors of the liver, either primary in the liver or metastatic from cancers in other organs, have received a wide clinical trial of cryosurgery. Conventional surgery of liver tumors is associated with two major problems, which are the technical difficulty of the operation, chiefly related to hemorrhage, and the fact that the majority of hepatic tumors are not suitable for surgical excision. With treatment by cryosurgery, prolonged survival may be achieved in $\sim$25% of patients, tumor recurs in the liver in $\sim$25%, and extra-hepatic disease becomes manifest in 50% of patients. Recurrence in the cryo-treated site occurs in 20% of patients (34). When one considers the fact that the patients selected for cryosurgery are those considered inoperable or unsuited for excisional surgery, the increased survival and chance of cure represents good results.

The application of cryosurgical techniques to kidney tumors is in clinical trial. The successful use of ultrasound to monitor the freezing process in tumors of the prostate and the liver has encouraged use in tumors of the kidney, especially when conservation of kidney function is critical to the patient. At the present time, in patients with marginal kidney functional reserve, partial nephrectomy is an option in therapy. In a similar manner, cryosurgical ablation offers the possibility of conserving renal tissue. The techniques and tissue effects of renal cryosurgery are similar to those in other viscera, such as the liver. Early clinical results indicate that cryosurgery is useful in the treatment of small tumors of the kidney in patients with associated disease (35).

Bone Tumors

Cryosurgical techniques have achieved a small place in the management of bone tumors. Cryosurgery is a conservative method of management and is better suited to the management of benign bone tumors rather than malignant tumors. The best technique requires that cryosurgery be used in combination with curettage, removing most of the soft tumor by scraping with a curette. This forms a cavity in the bone, and cryosurgery is used to freeze the cavity, destroying any residual tumor. In bone tumors, the freezing is performed with diverse techniques, including use of probes or by spraying or pouring liquid nitrogen into the bone cavity. When the cryosurgical treatment is finished, the bone cavity is filled with bone grafts or acrylic cement for added support. Cryosurgery is not yet used widely for bone tumors, but the results achieved by a few surgeons have been outstanding and have shown that amputation of selected long bones can be avoided (36).

Rhythm Disorders of the Heart

In recent years, cryosurgical techniques have been adapted to the treatment of arrhythmias of the heart, which is particularly useful for disabling abnormally fast heart rates that begin in the atrium. To prevent the ventricle from following the fast atrial rate, cryosurgical ablation of the atrioventricular node (AV node) is used to produce heart block. Then a pacemaker is implanted to maintain the cardiac rate. Similar techniques can be used to inactivate irritable foci in the ventricle or accessory pathways of conduction that produce dangerous fast heart rhythms. Cardiac cryosurgery is commonly done with catheter cryoprobes, passed into the heart from a peripheral blood vessel. The control of the freezing depends on electrophysiological monitoring that confirms that the desired freezing has been produced (37). Recent research related to cardiac disease has been in the direction of using warmer freezing temperatures to elicit a specific tissue response. The object is to stimulate angiogenesis as treatment of myocardial ischemia or to inhibit myogenesis as needed to

control the smooth muscle response to the injury caused by balloon angioplasty (38). To achieve these therapeutic goals may require the use of adjunctive chemotherapy.

COMPARISON WITH OTHER METHODS OF TREATMENT

The advantages of cryosurgery are most evident when used to treat an accessible lesion without excising any tissue. When used in this way, the technique is quick, relatively painless, and associated with little or no blood loss. Since usually no tissue is excised, there is no opportunity for tumor cell implantation in the open wound. Conservation of tissue, especially bone, is possible. The lethal effect of freezing extends into bone and destroys any tumor cells that may be present, while the devitalized matrix remains as a structural basis for later repair. Most patients have little discomfort after surgery because of the desensitizing effect of cold on sensory nerves. Nerve function, lost as a consequence of freezing, commonly returns after several months. Wound healing is surprisingly good, but is delayed by the need to await necrosis and sloughing of devitalized tissue. Soft tissue wounds heal in a month. The healing of soft tissue is favorable, and extensive scarring is rare. Bone repair may require a year or longer for completion.

The principal disadvantages are that an entire specimen is not available for histologic examination and healing is slower than with excision and closure. The lack of a specimen is a circumstance shared by other methods of treatment, such as electrocoagulation and radiotherapy. In fact, only complete surgical excision yields the specimen for study as a whole. Slow healing is not entirely disadvantageous, since it provides time to study the open wound and is associated with favorable healing.

The chief limitation with cryosurgery, especially when dealing with cancer, is the difficulty in freezing sufficient tissue. The best presently available apparatus is barely adequate for large cancers because tissues are poor conductors of heat and are provided with a source of heat that limits the extension of freezing. The amount of tissue that can be frozen in a single application of a probe is small in comparison with the size of many cancers. Depth of freezing beyond 2 cm is difficult to achieve. Multiple applications of the probe, or the use of multiple probes and repetition of freezing are the methods of compensating for the difficulties of freezing sufficient tissue.

As a form of producing local necrosis of tissue, cryosurgery must be compared with the other methods of producing similar effects, including conservative local excision and electrocoagulation. Freezing *in situ* requires less anesthesia than electrocoagulation, and the freezing process is more easily controlled and yields a necrotic lesion of more predictable size. The scar after healing is less with cryosurgery than with electrosurgery. Conservative local excision is applicable to many cancers, but has greater risk of bleeding and infection than cryosurgery. Cryosurgery is better for lesions that rest on underlying bone because the extensive destruction can be achieved without excising bone and producing undesirable and unnecessary postoperative disability. In some advanced skin cancers that are difficult problems in management, the results with Moh's technique of serial excision and immediate histologic monitoring are strongly competitive with cryosurgery.

Laser therapy is competitive with other methods of tissue destruction. The advantages of laser therapy closely resemble those of cryosurgery from the standpoint of healing, speed of treatment of multiple lesions, and usefulness in many diseases. The limitations of laser therapy are related to the expense of the equipment, considerations of safety in use, and the difficulty in determining the needed depth of treatment.

Radiation therapy may also be used for selected problems in management of tumors and has advantages in difficult areas, such as about the eyelids, the nose, and the anus. The results in these sites indicate that radiotherapy is a legitimate alternative choice in many patients; nevertheless, cryosurgery can achieve the same results, yet faster. Radiotherapy is not a good choice for lesions with bone invasion.

In comparison to many other tools used by the surgeon, cryosurgical apparatus is simple to use and the techniques of cryosurgery are easily learned. Nevertheless, as with any other technique that physicians use to cure disease, the key to success in therapy is in the selection of patients suitable to cryosurgical treatment and in the proper use of the technique. Considerable attention must be placed on technique that ensures good heat transfer from the tissues and an appropriate amount of destruction by freezing. Physicians who use cryosurgery for cancer should be familiar with alternate accepted methods of treatment so that an appropriate choice of therapy can be made and a change to a different method of therapy will be made when appropriate. Under these conditions, in the hands of an experienced physician, cryosurgery can be used to solve many difficult problems in cancer therapy and will yield surprisingly good results in carefully selected patients (39).

FUTURE DIRECTIONS

Cryosurgery has achieved a modest stature in medical, dental, and veterinarian practice. In the past 15 years, considerable progress has been made in cryosurgical techniques as a result of improved technology. With the availability of percutaneous and endoscopic probes, visceral and other deep lesions can be treated. Nevertheless, the available apparatus, even with the use of multiple probes, still has limited freezing capability. Improvements in equipment are needed. Current efforts are focused also on improvements in the imaging techniques, which will facilitate the control of the tissue freezing process.

The current direction of research explores cryosurgical techniques that are combined with other methods of treatment, such as radiotherapy and cancer chemotherapy, for the treatment of invasive cancers (40). Rather little experience has been reported concerning combinations of cryosurgery with radiotherapy, but the methods are complementary, with the effects of radiotherapy longer lasting in contrast to the quick destruction by freezing *in situ*. In advanced cancers, cryosurgery has also been used in combination with the systemic or local

administration of cancer chemotherapeutic drugs. This use is based on the hypothesis that cells that are injured, but not necessarily killed by freezing *in situ*, might be more susceptible to complete destruction by an antineoplastic agent.

Current research shows continued interest in the specific immunological response against antigens of frozen tissue. The practical clinical benefit is in the possibility that freezing a cancer in its primary site would produce an immunological response that would destroy cancer in distant sites to which it had spread. Some reports suggest that such benefit has been achieved, especially in the treatment of advanced cancer of the prostate gland. Furthermore, a specific immunological benefit has been shown by several groups of investigators working with experimental tumor systems in animals. Unfortunately, evidence of clinical benefit remains unclear. Therefore, one can conclude only that the potential use of specific immunotherapy is an attractive feature of cryosurgery that requires further investigation.

Differences in the sensitivity to cold injury of the several types of cells are also of importance to the future development of cryosurgery (38). The cells of bone, osteocytes, are very susceptible to freezing and die at minimal subfreezing temperatures. In skin, the pigment-bearing cells, melanocytes, are highly sensitive to cold injury, and selective destruction of melanocytes can be achieved at temperatures of -4 to $-20\,^{\circ}C$ in single short exposures. At this temperature range, the damage to other epidermal cells is minimal. Squamous cells of skin resist freezing injury at temperatures as cold as $-20\,^{\circ}C$. The importance of these differences is in the possibility that selective destruction of cells may be possible and some therapeutic advantage may be gained. The recent demonstration of the occurrence of apoptosis in the periphery of the cryogenic lesion points the way to molecular-based optimization of therapeutic freezing techniques, whether in the direction of selective destruction, partial preservation, or more complete destruction of cells. This appears to be the most potentially rewarding direction of cryosurgical research.

BIBLIOGRAPHY

1. Henderson A. Cold: Man's assiduous remedy. Med Ann DC 1971;40:583–588.
2. Cooper I, Lee A. Cryostatic congelation: A system for producing a limited controlled region of cooling and freezing of biologic tissues. J Nerv Dis 1961;133:259–263.
3. Gage AA. History of cryosurgery. Sem Surg Oncol 1998;14: 99–109.
4. Gage AA, Baust JG. Mechanisms of tissue injury in cryosurgery. Cryobiology 1998;37:171–186.
5. Rubinsky B. Cryosurgery. Annu Rev Biomed Eng 2000; 2:157–187.
6. Hoffmann NE, Bischof JC. The cryobiology of cryosurgical injury. Urology 2002;60:40–49.
7. Clarke DM, Hollister WB, Baust JG, VanBuskirk RG. Cryosurgical modeling: sequence of freezing and cytotoxic agent application affects cell death. Mol Urol 1999;3:25–31.
8. Yang WL, Addona T, Nair DG, Qi L, Ravikumar TS. Apoptosis induced by cryo-injury in human colorectal cancer cells is associated with mitochondrial dysfunction. Int J Cancer 2003;103:360–369.
9. Li AK, Ehrlich HP, Trelstad RL, Koroly MJ, Schattenkerk ME, Malt RA. Differences in healing of skin wounds caused by burn and freeze injuries. Ann Surg 1980;191:224–248.
10. Gage AA, Fazekas G, Riley EE. Freezing injury to large blood vessels in dogs. Surgery 1967;61:748–754.
11. Gage AA, Greene GW, Neiders ME, Emmings FG. Freezing bone without excision—an experimental study of bone cell destruction and manner of regrowth. JAMA 1966;196:770–774.
12. Gage AA. Cryosurgery in the treatment of cancer. Surg Gynecol Obstet 1992;174:73–92.
13. Baust JG, Gage AA, Ma H, Zhang CM. Minimally invasive cryosurgery—technological advances. Cryobiology 1997;34: 373–384.
14. Saliken JC, Cohen J, Miller R, Rothert M. Laboratory evaluation of ice formation around a 3-mm Accuprobe. Cryobiology 1995;32:285–295.
15. Popken F, Seifert JK, Englemann R, Dutkowski P, Nassir F, Junginger T. Comparison of iceball diameter and temperature distribution achieved with 3-mm Accuprobe cryoprobes in porcine and human liver tissue and human colorectal liver metastases *in vitro*. Cryobiology 2000;40:302–310.
16. Hewitt PM, Zhao J, Akhter J, Morris DL. A comparative laboratory study of liquid nitrogen and argon gas cryosurgery systems. Cryobiology 1997;35:303–308.
17. Hamilton A, Hu J. An electronic cryoprobe for cryosurgery using heat pipes and thermoelectric coolers. J Med Eng Technol 1993;17:104–109.
18. Diller KR. Engineering-based contributions in cryobiology. Cryobiology 1997;34:304–314.
19. Gage A, Caruana J, Garamy G. A comparison of instrument methods of monitoring freezing in cryosurgery. J Dermatol Surg Oncol 1983;9:209–214.
20. Le Pivert P, Binder P, Ougier T. Measurement of intra tissue bio-electrical low frequency impedance: A new method to predict preoperatively the destructive effect of cryosurgery. Cryobiology, 1977;14:245–250.
21. Gage AA, Augustynowicz S, Montes M, Caruana J, Whalen D. Tissue impedance and temperature measurements in relation to necrosis in experimental cryosurgery. Cryobiology, 1985;22: 282–288.
22. Homasson JP, Thiery JP, Angebault M, Outracht L, Maiwand O. The operation and efficacy of cryosurgical nitrous oxide-driven cryoprobe. Cryobiology 1994;31:290–304.
23. Harly S, Aastrup J, Elbrand O. Heat exchange in cryosurgery of Meniere's disease: Experimental and clinical studies. Cryobiology 1977;14:609–613.
24. Bradley P. Thermography as an aid to cryosurgery. Acta Thermographica 1977;2:83–90.
25. Pogrel MA, Yen CK, Taylor R. A study of infrared thermographic assessment of liquid nitrogen cryotherapy. Oral Surg Oral Med Oral Path 1996;81:396–401.
26. Brewer WH, Austin RS, Capps GW, Neifeld JP. Intraoperative monitoring and postoperative imaging of hepatic cryosurgery. Sem Surg Oncol 1998;14:129–155.
27. Saliken JC, Donnelly BJ, Rewcastle JC. The evolution and state of modern technology for prostate cryosurgery. Urology 2002;60:26–33.
28. Chin JL, Downey DB, Mulligan M, Fenster A. Three-dimensional transrectal ultrasound guided cryoablation for localized prostate cancer in nonsurgical candidates: A feasibility study and report of early results. J Urol 1998;159: 910–914.
29. Gilbert JC, Rubinsky B, Wong ST, Brennan KM, Pease GR, Leung PP. Temperature determination in the frozen region during cryosurgery of rabbit liver using MR image analysis. Magn Reson Imaging 1997;15:657–667.
30. Mala T, Edwin B, Tillung T, Kristian HP, Soreide O, Gladhaug I. Percutaneous cryoablation of colorectal liver metastases

potentiated by two consecutive freeze–thaw cycles. Cryobiology 2003;46:99–102.
31. Otten DM, Onik G, Rubinsky B. Distributed network imaging and electrical impedance tomography of minimally invasive surgery. Tech Cancer Res Treatment 2004;3:125–133.
32. Maiwand MO, Asimakopoulos G. Cryosurgery for lung cancer: Clinical results and technical aspects. Tech Cancer Res Treatment 2004;3:143–150.
33. Bahn DK, Lee F, Bodalament R, Kumar A, Greski J, Chernick M. Targeted cryoablation of the prostate: 7-year outcomes in the primary treatment of prostate cancer. Urology 2002;60:3–11.
34. Seifert JK, Junginger T. Cryotherapy for liver tumors: current status perspectives, clinical results, and review of the literature. Tech Cancer Res Treatment 2004; 3:151–163.
35. Spaliviero M, Moinzadeh A, Gill I. Laparoscopic cryotherapy for renal tumors. Tech Cancer Res Treatment 2004;3:177–180.
36. Bickels J, Meller I, Shmookler BM, Malawer MM. The role and biology of cryosurgery in the treatment of bone tumors. A review Acta Orthop Scand 1999;70:308–315.
37. Rodriguez LM, Timmermans C. Transvenous cryoablation of cardiac arrythmias. Tech Cancer Res Treatment 2004;3: 515–524.
38. Gage AA. Selective cryotherapy. Cell Pres Tech 2004; 2:3–14.
39. Gage AA, Baust JG. Cryosurgery for tumors—a clinical overview. Tech Cancer Res Treatment 2004;3:187–199.
40. Baust JG, Gage AA. Progress toward optimization of cryosurgery. Tech Cancer Res Treatment 2004;3:95–101.

See also MINIMALLY INVASIVE SURGERY; TISSUE ABLATION.

CRYOTHERAPY. See HEAT AND COLD, THERAPEUTIC.

CT SCAN. See COMPUTED TOMOGRAPHY.

CUTANEOUS BLOOD FLOW, DOPPLER MEASUREMENT OF

LALITA KHAODHIAR
ARISTIDIS VEVES
Harvard Medical School
Boston, Massachusetts

INTRODUCTION

The necessity for measuring the skin blood flows occurs in many areas of physiology, pharmacology, and clinical medicine. Although the measurement of blood flow in the large blood vessels in the human body has been performed for centuries, the use of techniques to explore microcirculation have just evolved over the past 30 years. Available tests for assessing the skin microcirculation include tissue pH measurement, radioactive isotope clearance, capillary microscopy, plethysmography, transcutaneous oxygen tension, ultrasonic Doppler flowmetry, and laser Doppler flowmetry (1–5). Each method relies on different physiology principles and has its own advantages and disadvantages (Table 1). Currently, there is no gold standard test for the evaluation of skin blood flow and clinical observation remains the most acceptable method for assessing blood flow in the skin in clinical practice (6,7). The ideal blood flow measurement technique should be simple, noninvasive, reproducible, and able to provide a continuous measurement of skin blood flow.

LASER DOPPLER FLOWMETRY

Laser Doppler flowmetry is the most widely accepted technique currently used for evaluating blood flow in the skin microcirculation. The basic technology underlying laser Doppler was introduced in 1975 by Stern, who demonstrated that the use of laser Doppler shifted light to measure the moving blood cell in the skin microcirculation. This technique has been in clinical use since 1977 (8,9) and since then it has been extensively studied, particularly in the field of vascular surgery, rheumatology, and dermatology. Although the laser Doppler flowmetry technology and data processing have continued to evolved, it has yet to gain the widespread acceptance for clinical applications. In this article, the principle, instrumentation (laser probe and laser scanning), and the clinical applications are discussed.

PRINCIPLES

This technique depends on the Doppler principle, which is the alteration in the frequency of light that is emitted or reflected by a moving object. The Doppler frequency shift can be calculated using the following equation:

$$df = v/c\, f$$

where df is magnitude of the frequency shift; v is the velocity of the moving object with respect to the observer; c is the velocity of light, and f is the frequency of unshifted light. This means when light hits a moving object, it undergoes a frequency shift that is proportional to the velocity of the moving object (10).

Because of the movement of red blood cells in the skin microvascular network, low power light from a monochromatic stable laser is scattered and as a result is frequency shifted. Since the velocity of the red blood cells is ~ 10 orders of magnitude smaller than the speed of light, it is impossible to measure this frequency-shifted light directly. The laser Doppler flowmetry, however, provides an indirect measurement of red cell velocity as follows: when the coherent laser light hits a surface, the light scattered from the red blood cells undergoes a frequency shift, but the light from the surrounding area remains at the same frequency as the transmitted light. The mixing of these two different light frequencies produces a beat frequency, that is, an oscillation of the measured light intensity. This beat frequency can be detected by the laser Doppler machine, and then analyzed to provide a skin blood flow measurement (4,11,12).

The term commonly used to describe blood flow measured by the Doppler techniques is flux, which is the amount relative to the product of the number of moving red cells in a given volume and their mean net velocity. The flux can be calculated using analogue circuitry or high speed digital processing. All laser measurements are usually expressed in volts and depend on the voltage difference created by the returned light to the computer. Higher blood flow at the skin

Table 1. Techniques for Assessing Skin Microcirculation

Technique	Advantages	Disadvantages
Photoplethysmography	Noninvasive, accurate, reproducible	Estimate based on changes in blood volume, not blood flow. Not useful in darkly pigmented skin
Transcutaneous oxygen tension	Noninvasive, provides physiologic and nutritional microcirculation assessment	Indirect estimate of blood flow, the measurement is affected by the affinity of blood for oxygen and the change in skin temperature
Thermometry	Noninvasive, correlated closely with capillary density	Flow estimated based on skin temperature, which can be influenced by ambient temperature, pain, anxiety
Capillary microscopy	Noninvasive, provides information on capillary size and number	Provides a relative estimate of blood flow based on visual characteristics
Ultrasonic doppler flowmetry	Noninvasive, measure nutritive perfusion	Probe is highly sensitive to motion,
Laser doppler flowmetry	Noninvasive, provides continuous real-time measurement of skin perfusion	Probe-poor reproducibility

level results in a higher amount of light picked by the laser Doppler and a higher voltage recorded by the computer.

Generally, the laser Doppler flowmetry measures blood flow in the very small blood vessels of the microvasculature, such as flow in the underlying arterioles and venules that regulate skin temperature and the low speed flows associated with nutritional blood flow in capillaries close to the skin surface. Thus, this technique does not differentiate between nutritional and non-nutritional skin perfusion (13).

INSTRUMENTATION

There are two types of laser Doppler flowmetry devices; a single-point laser probe, which evaluates the microvascular blood flow at one point of the skin, or a real-time laser scanner, which evaluates the blood flow in an area of skin.

Single-Point Laser Probe

In a single-point laser probe, laser light is transmitted to the tissue surface via optic fiber. The optic fiber terminates in an optic probe, which can be attached to the tissue surface. One or more light collecting fibers also terminate in the probe head and these fibers transmit a proportion of the scattered light to a photodetector and the signal processing electronics. Normal fiber separations in the probe head are a few tenths of a millimeter, so consequently blood flow is measured in a tissue volume of typically 1 mm^3 or smaller. The measuring volume (depth) of laser penetration is generally $\sim$ 1–1.5 mm, but it is dependent on many factors, such as probe configuration (14), laser light wavelength (3), and skin pigmentation. A light source with wavelength 543 nm has less penetration depth than 633 nm, which has less penetration depth than 780 nm (15). In clinical medicine, a wavelength of 633 nm, is generally used. The distance between the transmitting and receiving fibers (fiber separation) also influences the penetration depth, with the increasing depth with greater fiber separation.

The single-point laser probe is mainly used for evaluating the hyperemic response to a heat stimulus, or for evaluating the nerve-axon-related hyperemic response.

Heat-Related Hyperemic Response. To assess heat-related hyperemic response, the baseline blood flow is first measured. The skin is then heated to 44 °C for 20 min using a small brass heater. The measurement of the maximum blood flow is subsequently repeated to evaluate the magnitude of change from baseline.

Nerve-Axon-Related Hyperemic Response. The nerve-axon related-hyperemic response evaluates the integrity of the neurovascular function. In healthy subjects, the ability to increase blood flow depends on the existence of normal neurogenic vascular response, which is conducted through the C nociceptive nerve fibers. Stimulation of these nerve fibers leads to antidromic stimulation of adjacent C fibers, which secrete substance P, calcitonin gene-related peptide (CGRP) and histamine, causing vasodilatation and increased blood flow to the injured tissues, thereby promoting wound healing (Lewis' triple flare response) (Fig. 1). For this measurement, two single-point laser probes are applied (Fig. 2). One probe measures the blood flow to an area of skin, which is exposed directly to

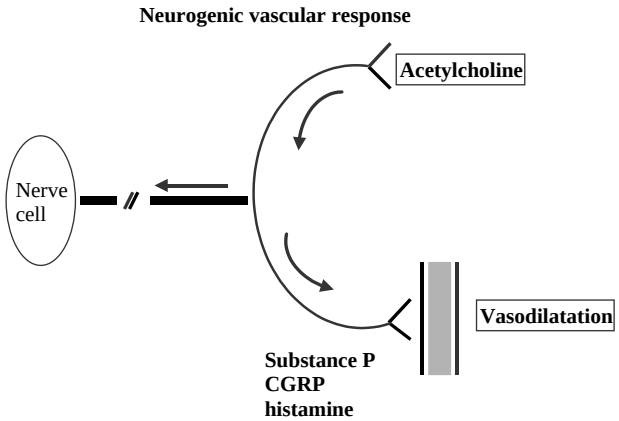


Figure 1. Stimulation of the C-nociceptive nerve fibers leads to antidromic stimulation of the adjacent C fibers, which secrete substance P, calcitonin gene related peptide (CGRP), and histamine that cause vasodilatation and increased blood flow.

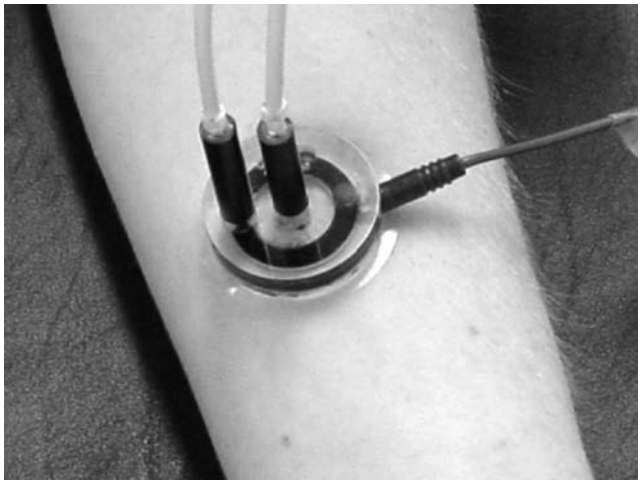


Figure 2. Measurements of direct and indirect effect of vasoactive substance using single-point laser probes: One probe is used in direct contact with the iontophoresis solution chamber (colored ring) and measures the direct response. The center probe measures the indirect response (nerve axon-related effect). A small quantity (<1 mL) of 1% acetylcholine chloride solution or 1% sodium nitroprusside solution is placed in the iontophoresis. A constant current of 200 mA is applied for 60 s achieving a dose of 6 mC cm^{-2} between the iontophoresis chamber and a second nonactive electrode placed 10–15 cm proximal to the chamber (black strap around the wrist). This current causes a movement of solution to be iontophorized toward the skin.

acetylcholine (Ach). The second probe, placed in close proximity (5 mm), measures the indirect effect of applied Ach. This indirect effect results from stimulation of C-nociceptive nerve fibers of the adjacent area and reflects the stability of the nerve-axon-related reactive hyperemia.

Our lab has examined the contribution of the nerve-axon reflex-related vasodilation to the total endothelium-dependent vasodilation at the forearm and the foot level in healthy adults and patients with diabetes mellitus (16). In healthy adults, the nerve-axon reflex-related response is approximately equal to one-third of the total response to Ach at both the forearm and the foot level. In diabetic patients with microvascular complications including diabetic neuropathy, Charcot arthropathy, and peripheral vascular disease, this contribution was significantly diminished. Another study demonstrated that the nerve-axon reflex-related vasodilatation is directly related to the function of the C-nociceptive fibers and is significantly associated with other nerve function measurements (17). As this method is an objective measurement, it is potentially useful as an alternative to currently employed techniques to evaluate small nerve fiber function.

Single-point measurements give a high temporal resolution (40 Hz data rates are typical) enabling rapid blood flow changes to be recorded. However, there are several limitations of the conventional laser Doppler probe. Because the probe is directly contacted to the skin, it can only measure the restricted area ~ 1 mm^2 at one time. Its pressure on the skin itself may also alter the skin blood flow (18). In addition, the probe is very sensitive to motion and vibration while it has poor reproducibility (3).

The Laser Scanning Method

The laser Doppler scanner–imager has been developed in response to the limitation of the laser Doppler probe. In the laser scanning method, a larger area of the skin can be studied while avoiding the contact between the scanner and the tissue being assessed. This technique is based on the same principle of measuring blood flow as the laser probe, but instead of the fiber optic probes, a system of mirrors and light-collecting lenses are used (14). This technique, the low intensity laser beam, is scanned across tissue surface in a raster fashion using a moving mirror. The scanner can scan the area from 5×5 cm to up to 50×50 cm. Light reflected back from the skin is then detected by a photodetector, which is connected to the computer enabling a mapping and a display of color-coded images of the blood flow. Regions of interest can then be defined and statistical data are calculated and recorded. This technique is also useful for the study of the skin microcirculation in response to various vasoactive substances.

To evaluate the endothelium-dependent and the endothelium-independent microvascular reactivity, the laser scanning method is used through the iontophoresis technique. The conditions associated with endothelial dysfunction are listed in Table 2.

The term iontophoresis denotes the introduction of soluble ions into the human skin by applying electric current. Using this technique, vasoactive substances can be applied to a localized area of the skin. The delivered dose depends on the current flowing and its duration. The test is noninvasive and avoids any systemic effects of the used drugs. By applying Ach chloride, the endothelium-dependent vasodilatation can be measured, while the use of sodium nitroprusside (SNP) measures the endothelium-independent vasodilatation.

In this technique, a delivery vehicle device is attached firmly to the skin with double-sided adhesive tape. The device contains two chambers that accommodate two single-point laser probes. A small quantity of (< 1 mL) of 1% Ach solution or 1% of SNP solution is placed in the iontophoresis chamber and a constant current of 200 mA is applied for 60 s, achieving a dose of 6 mC·cm^{-2} between the

Table 2. Conditions Associated with Impaired Endothelial Function

Atherosclerosis
Hypertension
Dyslipidemia; high LDL-C, low HDL-C, small dense LDL-C
Diabetes mellitus and impaired glucose tolerance
Metabolic syndrome
Obesity
Congestive heart failure
Preeclampsia
Vasculitis
Renal failure
Menopause
Family history of coronary heart disease
Family history of diabetes
Smoking
Inactivity

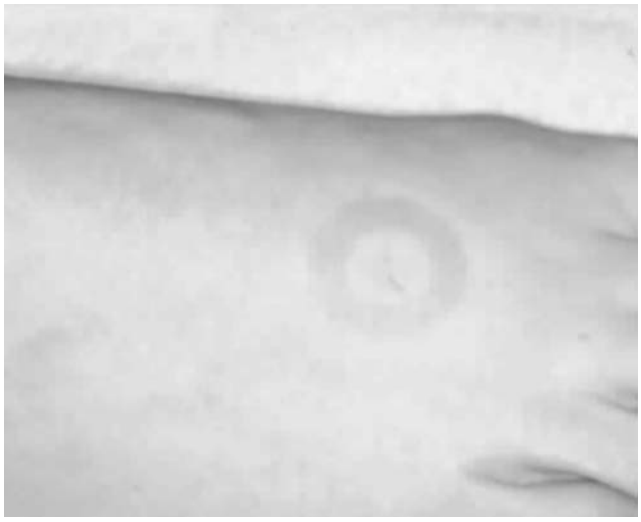


Figure 3. A normal response of blood flow in a skin to iontophoresis technique. Vasodilatation occurs in both the area that contact with the iontophoresis solution and area adjacent to, but not in direct contact with, the solution.

iontophoresis chamber and a second nonactive electrode placed 10–15 cm proximal to the chamber. This current causes a movement of solution to be iontophoresed toward the skin, resulting in vasodilatation (Fig. 3).

After the adhesive device has been removed, the localized area exposed to the vasoactive substances is scanned. The laser Doppler perfusion imager employs a 1 mW helium–neon laser beam of 633 nm wavelength, which sequentially scans an area of skin (Fig. 4). The maximum number of measured spots is 4096, and the apparatus produces a color-coded image of skin erythrocyte flux on a computer monitor. The scanner is set up to scan up to 32×32 measurement points over an area $\sim 4 \times 4$ cm.

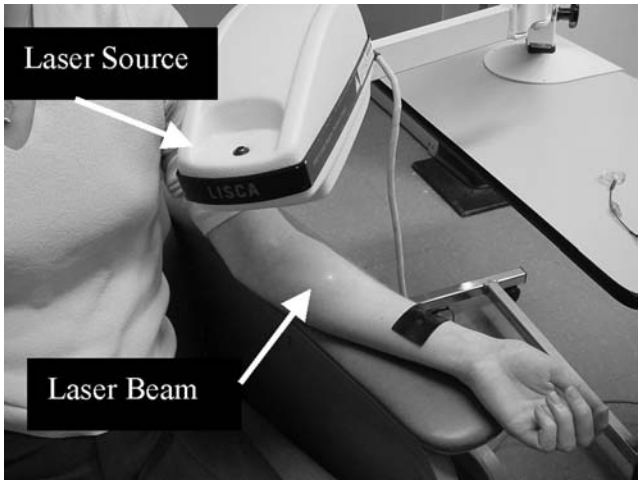


Figure 4. Laser doppler flowmetry: A helium–neon laser beam is emitted from the laser source to sequentially scan the circular hyperemic area (seen surrounding the laser beam) produced by the iontophorized vasoactive substance to a small area on the volar surface of the forearm.

Generally, the laser Doppler scanner is not a reliable method for a quantification of the nerve-axon reflex related vasodilatation. However, Krishnan and Rayman recently published a technique assessing the axon-reflex related vasodilatation using the laser Doppler imager (LDI) called LDIflare. This method involved skin heating to 44 °C to evoke flare followed by scanning the site using a laser Doppler imager. The LDIflare was markedly reduced in healthy control subjects after topical administration of anesthesia, confirming its neurogenic nature. Similarly, LDIflare was decreased in patients with diabetic neuropathy when compared to diabetic patients with no neuropathy or healthy controls. The authors suggested that the LDIflare may be a simple objective method to detect early neuropathy and may be useful in assessing therapeutic interventions aimed at preventing or reversing C-fiber function.

In summary, the laser Doppler scanning offers a more global assessment of skin perfusion than the laser Doppler probe, while avoiding a direct contact to the skin surface. The scanning technique is best suited for studying the relative changes in flow induced by variety of physiological maneuvers or pharmaceutical intervention procedures.

VALIDATION

Single-Point Laser Probe

The technique has been validated against direct measurements of the capillary flow velocity (19). The day-to-day reproducibility of the technique was evaluated in healthy subjects who were repeatedly tested at their foot and arm for 10 consecutive working days in our lab. The coefficient of variation (CV) for the baseline blood flow measurement obtained with the laser probe evaluating the response to heat was 44.0%, while that for the maximal response to heat was 27.9%. The indirect response to Ach, measured by a single-point laser probe, had a CV of 60.6% for the baseline measurements and 35.2% for the maximal hyperemic response after the iontophoresis. Note that the variability of the technique is mostly a spatial one, that is, the variability is mainly related to the high heterogeneity of the skin microcirculation and not to the technique itself. The reproducibility of the single-point laser Doppler can be improved when one pays attention to place the laser probe approximately in the same skin area for repeated measurement.

The Laser Scanner Method

The laser scanner method has a significantly better reproducibility than a single-point laser probe. The CV at the foot and forearm level is ranged between 14 and 19%. For the laser Doppler perfusion imaging measurements before and after the iontophoresis of Ach, the reproducibility of the technique was evaluated in our lab in five healthy subjects (four males and one female, age 23–39 years) who were repeatedly tested at the forearm for 10 consecutive working days. The coefficient of variation of the baseline measurement before the iontophoresis of Ach was 14.1% and during maximal hyperemic response after the iontophoresis it was 13.7% (16,20).

CLINICAL APPLICATIONS

Laser Doppler flowmetry have several clinical applications. It has been used to monitor to ischemic tissue (2), to follow the progress of atherosclerotic disease, to evaluate therapeutic effects of drugs or operations (21), and to study skin under both normal physiologic states and the pathogenetic mechanisms underlying pathologic skin disorders (22,23).

Atherosclerosis. Peripheral arterial disease (PAD) of the lower extremity is a common manifestation of the atherosclerotic process. As many as 10 million people in the United States have PAD with a prevalence $> 10\%$ in people aged > 60 years (24). Up to three-quarters of patients with PAD are asymptomatic (25,26). Those with symptoms usually present with intermittent claudication, resting ischemia or foot ulceration. Kvernebo et al. reported laser Doppler flowmetry values decreased progressively from healthy controls to patients with intermittent claudication to those with critical ischemia (21). The test was reproducible on a given population particularly with the local heating technique, but the reproducibility was poor for individual subjects. The possible explanations include changes in sympathetic vascular tone and different vascular architecture in the measuring volumes, which are only some few cubic millimeters (mm^3). These patients with PAD also demonstrate a reduced skin hyperemic response to iontophoresis of ACh and to SNP (27), suggesting both endothelium and smooth muscle dysfunction. In addition, when they exercise, a reduction of the leg skin hyperemic response to ACh delivery is more pronounced at peak of claudication than prior exercise. The authors suggested acute endothelial dysfunction occurs with exercise-induced leg muscle ischemia and threshold of claudication should not be exceeded during rehabilitation programs for PAD patients.

Laser Doppler flowmetry may also be useful for follow-up after therapeutic intervention in patients with PAD. Ray et al. (28) studied 41 patients who underwent technically successful revascularization for severe leg ischemia. Laser Doppler flowmetry performed before the procedure was better than toe and ankle systolic pressures in identifying patients who continued to have ischemic symptoms or required amputation after surgery. Another study reported the prevalence of high frequency flux waves, which increased in peripheral ischemia decreased after successful percutaneous angioplasty (29). One randomized, double-blind, placebo-controlled trial examining the effect of glutathione in patients with peripheral artery disease reported improvement in pain-free walking distance (PFWD) and microciculatory (laser Doppler flowmetry) test (30). Another study, the START-trial: STimulation of ARTeriogenesis using subcutaneous application of GM-CSF as a new treatment for peripheral vascular disease (31), is underway. In this study, the primary endpoint will be the change walking distance from day 0 to day 14 as assessed by an exercise treadmill test, while the cutaneous microcirculation alterations assessed by laser Doppler flowmetry will be one of the secondary endpoints.

Diabetes Mellitus. Over the last two decades, it has become apparent that metabolic alterations in diabetes cause both structural and functional changes in multiple areas within the arteriolar and capillary systems. The most characteristic structural changes of the capillary circulation in diabetic patients are a reduction in the capillary size and thickening of basement membranes, while the major functional change includes a marked limitation of microvascular vasodilation (32–34). Studies have shown that the structural changes are essentially responsible for the impairment of microvascular function. Microcirculatory test in patients with diabetes mellitus demonstrated a reduction in hyperemic response to heat stimulus and minor skin trauma, suggesting the role of endothelial dysfunction as the cause of the impaired vascular reactivity at microcirculatory level (35). Such dysfunction occurs early in the course of diabetes and may even predict diabetic micro- and macrovascular complications (36,37).

Several therapeutic interventions have been tested both *in vitro* and/or *in vivo* aiming to improve endothelial function in patients with diabetes mellitus. This includes insulin sensitizers (thiazolidenediones), angiotensin-converting enzyme (ACE) inhibitor, lipid lowering medications (statins), antioxidant, and exercise. To date, the studies have shown these interventions had no effects on the skin microcirculation in patients with diabetes, including both resting skin blood flow and blood flow after the iontophoresis of acetylcholine and sodium nitroprusside (38–41).

Skin Disease. Several skin diseases are associated with abnormal cutaneous blood flow, for example, psoriasis, skin inflammatory reaction, skin cancers, and scleroderma. Laser Doppler flowmetry has been used to determine differences in blood flow within psoriatic plaques, to identify the location of their leading edge, and to monitor the treatment (42,43). Laser Doppler perfusion imaging may also allow differentiation between different types of skin tumor. Stucker et al. found that malignant melanomas were significantly more perfused than basal cell carcinomas and tended also to be more so than melanocytic naevi (44,45).

Skin Burns. Laser Doppler flowmetry has been used to determine burn depth. Niazi et al. (46) demonstrated the Laser Doppler flowmetry flux values at 24 h after burn correlated with clinical assessment and histology reports, with low flux seen in deep dermal or full-thickness wound but high flux in superficial dermal wound (46). Park et al. (47) showed that laser Doppler flow measurements obtained within 72 h of burn injury correlated well with the depth of burn wound. The accuracy of laser Doppler in the assessment of burn depth is as high as 97%, compared with 60–80% for standard clinical methods (48). In addition, this technique has been used in concurrence with clinical judgment to objectively determine the need for excision of burns of indeterminate depth (49).

Skin Ulcer. Both ischemic and venous ulcers and their adjacent skin are associated with disorders in microcirculation. Together the deeper, subpapillary (thermoregulatory) network and the superficial (nutritive) network are affected. These network can be studied with laser Doppler flowmetry (subpapillary area) and capillary microscopy

(superficial area). Gschwandtner et al. (50–53) examined the microcirculatory characteristics of ulcers. They described the remarkable local differences in subpapillary and nutritive perfusion in ischemic and venous ulcers and their surrounding skin (51,52). Ulcer areas without granulation tissue (sign of ulcer without healing) demonstrated low laser Doppler area flux and the very low capillary density. The lack of capillary in this area suggested the breakdown of nutritive microcirculation as an underlying cause of ulcers. Ulcer with granulation tissue (sign of wound healing) had high laser Doppler area flux and an intermediate capillary density, implying both thermoregulatory, and nutritive networks are essential for wound healing of ulcers. The skin area adjacent ulcers where healing process nearly completed had an intermediate laser Doppler area flux and highest capillary density, indicating an adequate thermoregulatory and nutritive perfusion (53). Bornmyr et al. (54) reported the use of laser Doppler perfusion imaging and digital photograph for postoperative evaluation of vascularized grafts and monitoring of treatment of chronic skin ulcers.

CONCLUSION

Laser Doppler flowmetry has already been an important tool for studying blood flow in the skin. The technique provides continuous, noninvasive real-time assessment of skin perfusion. It has been successfully used in many studies of the cutaneous blood flow in patients with atherosclerosis, diabetes, skin diseases, or skin ulcers.

BIBLIOGRAPHY

1. Chittenden SJ, Shami SK. Microvascular investigations in diabetes mellitus. Postgrad Med J 1993;69:419–428.
2. Furnas H, Rosen JM. Monitoring in microvascular surgery. Ann Plast Surg 1991;26:265–272.
3. Schabauer AM, Rooke TW. Cutaneous laser Doppler flowmetry: applications and findings. Mayo Clin Proc 1994;69:564–574.
4. Choi CM, Bennett RG. Laser Dopplers to determine cutaneous blood flow. Dermatol Surg 2003;29:272–280.
5. Fagrell B. Microcirculatory methods for the clinical assessment of hypertension, hypotension, and ischemia. Ann Biomed Eng 1986;14:163–173.
6. Hellner D, Schmelzle R. Laser Doppler monitoring of free microvascular flaps in maxillofacial surgery. J Craniomaxillofac Surg 1993;21:25–29.
7. Hirigoyen MB, Urken ML, Weinberg H. Free flap monitoring: a review of current practice. Microsurgery 1995;16:723–726; discussion 727.
8. Stern MD, Lappe DL, Bowen PD, Chimosky JE, Holloway GA Jr, Keiser HR, Bowman RL. Continuous measurement of tissue blood flow by laser-Doppler spectroscopy. Am J Physiol 1977;232:H441–H448.
9. Holloway GA Jr, Watkins DW. Laser Doppler measurement of cutaneous blood flow. J Invest Dermatol 1977;69:306–309.
10. Fischer JC, Parker PM, Shaw WW. Laser Doppler flowmeter measurements of skin perfusion changes associated with arterial and venous compromise in the cutaneous island flap. Microsurgery 1985;6:238–243.
11. Rendell M, Bergman T, O'Donnell G, Drobny E, Borgos J, Bonner RF. Microvascular blood flow, volume, and velocity measured by laser Doppler techniques in IDDM. Diabetes 1989;38:819–824.
12. Zinser G. Scanning laser Doppler flowmetry. In: Pillunat L, Harris A, Anderson D, Greve E, editors. Current concepts on ocularblood flow in glaucoma. The Hague: Kugler Publications; 1999: 197–204.
13. Rossi M, Carpi A. Skin microcirculation in peripheral arterial obliterative disease. Biomed Pharmacother 2004;58:427–431.
14. Essex TJ, Byrne PO. A laser Doppler scanner for imaging blood flow in skin. J Biomed Eng 1991;13:189–194.
15. Bonner RF, Nossal R. Principles of laser Doppler flowmetry. In: Shepherd AP, Oberg PA, editors. Laser Doppler Blood Flowmetry. Kluwer Academic Publishers; 1990.
16. Hamdy O, Abou-Elenin K, LoGerfo FW, Horton ES, Veves A. Contribution of nerve-axon reflex-related vasodilation to the total skin vasodilation in diabetic patients with and without neuropathy. Diabetes Care 2001;24:344–349.
17. Caselli A, Rich J, Hanane T, Uccioli L, Veves A. Role of C-nociceptive fibers in the nerve axon reflex-related vasodilation in diabetes. Neurology 2003;60:297–300.
18. Obeid AN, Barnett NJ, Dougherty G, Ward G. A critical review of laser Doppler flowmetry. J Med Eng Technol 1990;14:178–181.
19. Tooke JE, Ostergren J, Fagrell B. Synchronous assessment of human skin microcirculation by laser Doppler flowmetry and dynamic capillaroscopy. Int J Microcirc Clin Exp 1983;2:277–284.
20. Veves A, Akbari CM, Primavera J, Donaghue VM, Zacharoulis D, Chrzan JS, DeGirolami U, LoGerfo FW, Freeman R. Endothelial dysfunction and the expression of endothelial nitric oxide synthetase in diabetic neuropathy, vascular disease, and foot ulceration. Diabetes 1998;47:457–463.
21. Kvernebo K, Slagsvold CE, Stranden E, Kroese A, Larsen S. Laser Doppler flowmetry in evaluation of lower limb resting skin circulation. A study in healthy controls and atherosclerotic patients. Scand J Clin Lab Invest 1988;48:621–626.
22. Braverman IM. The cutaneous microcirculation. J Investig Dermatol Symp Proc 2000;5:3–9.
23. Braverman IM. The cutaneous microcirculation: ultrastructure and microanatomical organization. Microcirculation 1997;4: 329–340.
24. Criqui MH. Peripheral arterial disease—epidemiological aspects. Vasc Med 2001;6:3–7.
25. Hooi JD, Kester AD, Stoffers HE, Overdijk MM, van Ree JW, Knottnerus JA. Incidence of and risk factors for asymptomatic peripheral arterial occlusive disease: a longitudinal study. Am J Epidemiol 2001;153:666–672.
26. Stoffers HE, Rinkens PE, Kester AD, Kaiser V, Knottnerus JA. The prevalence of asymptomatic and unrecognized peripheral arterial occlusive disease. Int J Epidemiol 1996;25: 282–290.
27. Rossi M, Cupisti A, Perrone L, Mariani S, Santoro G. Acute effect of exercise-induced leg ischemia on cutaneous vasoreactivity in patients with stage II peripheral artery disease. Microvasc Res 2002;64:14–20.
28. Ray SA, Buckenham TM, Belli AM, Taylor RS, Dormandy JA. The predictive value of laser Doppler fluxmetry and transcutaneous oximetry for clinical outcome in patients undergoing revascularisation for severe leg ischaemia. Eur J Vasc Endovasc Surg 1997;13:54–59.
29. Bollinger A, Hoffmann U, Franzeck UK. Evaluation of flux motion in man by the laser Doppler technique. Blood Vessels 1991;28(1 Suppl): 21–26.
30. Arosio E, De Marchi S, Zannoni M, Prior M, Lechi A. Effect of glutathione infusion on leg arterial circulation, cutaneous microcirculation, and pain-free walking distance in patients with peripheral obstructive arterial disease: a randomized, double-blind, placebo-controlled trial. Mayo Clin Proc 2002; 77:754–759.

31. van Royen N, Piek JJ, Legemate DA, Schaper W, Oskam J, Atasever B, Voskuil M, Ubbink D, Schirmer SH, Buschmann I, Bode C, Buschmann EE. Design of the START-trial: STimulation of ARTeriogenesis using subcutaneous application of GM-CSF as a new treatment for peripheral vascular disease. A randomized, double-blind, placebo-controlled trial. Vasc Med 2003;8:191–196.
32. Jaap AJ, Shore AC, Stockman AJ, Tooke JE. Skin capillary density in subjects with impaired glucose tolerance and patients with type 2 diabetes. Diabetes Med 1996;13:160–164.
33. Jaap AJ, Pym CA, Seamark C, Shore AC, Tooke JE. Microvascular function in type 2 (non-insulin-dependent) diabetes: improved vasodilation after one year of good glycaemic control. Diabetes Med 1995;12:1086–1091.
34. Jaap AJ, Tooke JE. Pathophysiology of microvascular disease in non-insulin-dependent diabetes. Clin Sci (London) 1995; 89:3–12.
35. Sandeman DD, Shore AC, Tooke JE. Relation of skin capillary pressure in patients with insulin-dependent diabetes mellitus to complications and metabolic control. N Engl J Med 1992;327: 760–764.
36. Williams SB, Cusco JA, Roddy MA, Johnstone MT, Creager MA. Impaired nitric oxide-mediated vasodilation in patients with non-insulin-dependent diabetes mellitus. J Am Coll Cardiol 1996;27:567–574.
37. Stehouwer CD, Fischer HR, van Kuijk AW, Polak BC, Donker AJ. Endothelial dysfunction precedes development of microalbuminuria in IDDM. Diabetes 1995;44:561–564.
38. Economides PA, Caselli A, Tiani E, Khaodhiar L, Horton ES, Veves A. The effects of atorvastatin on endothelial function in diabetic patients and subjects at risk for type 2 diabetes. J Clin Endocrinol Metab 2004;89:740–747.
39. Hamdy O, Ledbury S, Mullooly C, Jarema C, Porter S, Ovalle K, Moussa A, Caselli A, Caballero AE, Economides PA, Veves A, Horton ES. Lifestyle modification improves endothelial function in obese subjects with the insulin resistance syndrome. Diabetes Care 2003;26:2119–2125.
40. Economides PA, Caselli A, Zuo CS, Sparks C, Khaodhiar L, Katsilambros N, Horton ES, Veves A. Kidney oxygenation during water diuresis and endothelial function in patients with type 2 diabetes and subjects at risk to develop diabetes. Metabolism 2004;53:222–227.
41. Caballero AE, Saouaf R, Lim SC, Hamdy O, Abou-Elenin K, O'Connor C, Logerfo FW, Horton ES, Veves A. The effects of troglitazone, an insulin-sensitizing agent, on the endothelial function in early and late type 2 diabetes: a placebo-controlled randomized clinical trial. Metabolism 2003;52: 173–180.
42. Speight EL, Essex TJ, Farr PM. The study of plaques of psoriasis using a scanning laser-Doppler velocimeter. Br J Dermatol 1993;128:519–524.
43. Speight EL, Farr PM. Calcipotriol improves the response of psoriasis to PUVA. Br J Dermatol 1994;130:79–82.
44. Stucker M, Hoffmann M, Memmel U, von Bormann C, Hoffmann K, Altmeyer P. [In vivo differentiation of pigmented skin tumors with laser Doppler perfusion imaging]. Hautarzt 2002;53:244–249.
45. Stucker M, Horstmann I, Nuchel C, Rochling A, Hoffmann K, Altmeyer P. Blood flow compared in benign melanocytic naevi, malignant melanomas and basal cell carcinomas. Clin Exp Dermatol 1999;24:107–111.
46. Niazi ZB, Essex TJ, Papini R, Scott D, McLean NR, Black MJ. New laser Doppler scanner, a valuable adjunct in burn depth assessment. Burns 1993;19:485–489.
47. Park DH, Hwang JW, Jang KS, Han DG, Ahn KY, Baik BS. Use of laser Doppler flowmetry for estimation of the depth of burns. Plast Reconstr Surg 1998;101:1516–1123.
48. Pape SA, Skouras CA, Byrne PO. An audit of the use of laser Doppler imaging (LDI) in the assessment of burns of intermediate depth. Burns 2001;27:233–239.
49. Jeng JC, Bridgeman A, Shivnan L, Thornton PM, Alam H, Clarke TJ, Jablonski KA, Jordan MH. Laser Doppler imaging determines need for excision and grafting in advance of clinical judgment: a prospective blinded trial. Burns 2003;29:665–670.
50. Gschwandtner ME, Koppensteiner R, Maca T, Minar E, Schneider B, Schnurer G, Ehringer H. Spontaneous laser doppler flux distribution in ischemic ulcers and the effect of prostanoids: a crossover study comparing the acute action of prostaglandin E1 and iloprost vs saline. Microvasc Res 1996;51: 29–38.
51. Gschwandtner ME, Ambrozy E, Fasching S, Willfort A, Schneider B, Bohler K, Gaggl U, Ehringer H. Microcirculation in venous ulcers and the surrounding skin: findings with capillary microscopy and a laser Doppler imager. Eur J Clin Invest 1999;29:708–716.
52. Gschwandtner ME, Ambrozy E, Schneider B, Fasching S, Willfort A, Ehringer H. Laser Doppler imaging and capillary microscopy in ischemic ulcers. Atherosclerosis 1999;142: 225–232.
53. Gschwandtner ME, Ambrozy E, Maric S, Willfort A, Schneider B, Bohler K, Gaggl U, Ehringer H. Microcirculation is similar in ischemic and venous ulcers. Microvasc Res 2001;62:226–235.
54. Bornmyr S, Martensson A, Svensson H, Nilsson KG, Wollmer P. A new device combining laser Doppler perfusion imaging and digital photography. Clin Physiol 1996;16:535–541.

See also BLOOD RHEOLOGY; HEMODYNAMICS.

CYSTIC FIBROSIS SWEAT TEST

WARREN J. WARWICK
University of Minnesota
Minneapolis, Minnesota

INTRODUCTION

Cystic fibrosis (CF, cystic fibrosis of the pancreas, mucoviscidosis) is an autosomal recessive genetic clinical condition that occurs in ~ 1 in 2500 Caucasian newborn infants. The frequency ranges from $\sim 1{:}200$ in genetically isolated populations to 1:30,000 in certain ethnic and racial groups.

The CF mutations occur in a gene located in region q31.2 on the long (q) arm of human chromosome 7. The gene contains $\sim 250{,}000$ base (amino acids) pairs in 27 exons. These base pairs code for a 1480 amino acid molecule, which serves both as a transmembrane channel for the transfer of water and salt and other substances across the cell membrane and has been given the name: cystic fibrosis transmembrane regulator (CFTR) protein.

One mutation, the ΔF508 mutation, accounts for two-thirds of the mutations in Caucasian populations. Over 1338 mutations of the CFTR gene have been found in patients with one or more of the CF associated diseases although >100 mutations have been found in patients with as yet none of the CF associated diseases. A few patients have been found with increased numbers of CF associated diseases, but with no mutations in the CFTR gene (1). These patients are presumed to have similar risk factors caused by other, as yet not identified, transmembrane channels. Such mutations or combinations of gene

mutations, even in a large population, occur very rarely by chance (2). The most complete source of information about the CFTR gene is in the Cystic Fibrosis Mutation Database at http://www.genet.sickkids.on.ca/cftr/.

The CFTR protein controls sweat gland chloride ion transport as well as regulating other chloride secretory channels. This abnormality of chloride transport, which has been observed with almost all mutations, is an increased excretion of salt in the sweat. This elevation of chloride in sweat is seen in $>98\%$ of Caucasian CF patients. The clinical heterogeneity of CFTR expression is less obvious in the lung and other organs producing the diversity of clinical expression (phenotypes) of CF (3).

The reduced salt absorption in cystic fibrosis sweat glands is due primarily to poor chloride absorption with secondarily poor sodium absorption (4). A cyclic adenosine monophosphate (cAMP)-mediated sweating rate test has been developed that demonstrates a quantitative discrimination of CFTR function. This function may help distinguish between homozygous CF, CF carrier, and non-CF (5).

The measurement of the increased chloride in the sweat is the most reliable diagnostic test for the presence of two mutations of the CFTR mutation. This excessive sweat chloride is found in $>98\%$ of patients with the genetic potential to develop the clinical diseases associated with these CFTR mutations.

HISTORY

Although the folklore of many Caucasian peoples record that an infant who tastes salty will die young, the first modern confirmation of the excessive amount of salt in the sweat of children with CF was of seven patients with cystic fibrosis admitted to a New York City hospital with heat prostration (6). Although five other children also had heat prostration the unusual association with cystic fibrosis of the pancreas was investigated by di SantAgnese et al. (7) who found abnormal levels of sodium, chloride, and potassium in the sweat of these children. Their crude techniques for collection of sweat were inconsistently applied and produced nonstandard sweat samples for analysis, still all analyses consistently showed an increased amount of ions in the sweat.

Normal values were soon developed for children and adults without CF and parents of children with CF were shown to have slight, but significantly elevated, chloride and sodium in their sweat (8,9). The large amount of sweat required for these tests was obtained by sweating an arm or a leg of an adult or large child and required whole body sweating of small children. Such total body sweating led to some deaths due to heat exhaustion of some infants with genetic CF.

DEVELOPMENT OF A GOLD STANDARD

This uniform finding that children with CF of the pancreas have 10 times the amount of salt in their sweat created the need for a simple, cheap, rapid, and precise way to determine the salt content of sweat. Six years later L.C. Gibson, working in RE Cooke's laboratory, developed a technology and technique (10) for sweat stimulation that (1) is simple and quick, (2) almost always produces a sufficient (>70 mg) amount of sweat, (4) has virtually zero risk of injury with equipment, and (5) could be built by any electrician associated with a hospital laboratory (11). Soon, laboratories worldwide adopted this technology and showed that $>95\%$ of normal patients had sweat chlorides <30 mmol/L, whereas almost all patients with CF had sweat chloride values >60 mmol/L. This method has been named the Gibson–Cooke Sweat Test (GCST) honoring the innovators and the Quantitative Pilocarpine Iontophoresis Test (QPIT) identifying and focusing on the key elements of the technology. Both names may be regarded as interchangeable and equally appropriate abbreviations of the excessively long "Quantitative Gibson–Cooke Pilocarpine Iontophoresis Sweat Test" (GCST/QPIT).

Over the subsequent 40 years the (GCST/QPIT) has been validated and confirmed as the only to be trusted sweat test technology for the laboratory diagnosis of CF. This GOLD Standard label has persisted despite the extreme care that must be taken to assure accurate results. The basis for that consensus is founded on three factors; (1) a known amount of sweat, (2) the chloride concentration provides the greatest discrimination, and (3) the arithmetic difference between unit measurements provides the same visual distance throughout the physiological clinical range of sweat ion concentrations.

Because of the many sources of technical error that exist with the GCST/QPIT, efforts have been made to develop alternative technology that are easier to do, have fewer risks of errors, and are simple enough to be used in general hospitals, clinics, and even physicians offices. So far, because of the potential for missed diagnoses, the best of these have only reached approval and are recommended for use only as screening tests (12).

The GCST/QPIT is the most valuable laboratory test for the diagnosis of CF. The primary indication for the GCST/QPIT is the presence of one or more of the common CF associated diseases, which include malabsorption, failure to thrive, recurrent pulmonary infection, chronic obstructive lung disease, nasal polyps, chronic sinusitis, male infertility, gallstones, unexplained cirrhosis, arthritis, diabetes, bleeding due to vitamin K deficiency, asthma, rectal prolapse, intussusception, meconium ileus, night blindness due to vitamin A deficiency, hyponatremia, bowel obstruction, volvulus, acute pancreatitis, and the child who tastes salty. Other required reasons for performing the sweat test include immediate family history (to include first cousins) of a patient having CF, a positive or suspicious newborn screening for cystic fibrosis, and the request of the parent for a sweat test.

The purity of the GCST/QPIT technology has been and continues to be maintained (13) by many CF Center Directors, national and international organizations, including Cystic Fibrosis Foundations, Directors of Clinical Laboratories, Cystic Fibrosis Center Directors, medical specialties including Pediatricians, Pulmonologists, Gastroenterologists, Clinical Biochemists and Pathologists. Their concerns to keep the GCST/QPIT technology pure and accurate have been the object of many papers and publications with the most complete being the 97 page *Guidelines*

for the Performance of the Sweat Test for the Investigation of Cystic Fibrosis in the UK, Report from the Multi-Disciplinary Working Group with Representation from the Association of Clinical Biochemists, British Paediatric Respiratory Society, British Thoracic Society, Cystic Fibrosis Trust, Royal College of Paediatrics & Child Health, Royal College of Pathologists. UK National External Quality Assessment Schemes. This guideline has been formally appraised and endorsed by The Royal College of Paediatric and Child Health (http://www.acb.org.uk/Guidelines/sweat.htm) November 2003.

THE STATE OF THE ART OF SWEAT TESTS

While CF Center Directors are focusing on the precision of this test, as it is the most constant abnormality identifying CF, there have been many attempts to help the physician in practice to screen patients with some of the classic CF symptoms, and so to avoid the need, cost, and inconvenience of referring such patients to CF Centers for the approved GCST/QPIT best test.

Harry Shwachman, doyen of CF Center Directors, made the first screening test using agar plates filled with silver chromate (14). He used these agar plates on ward rounds and in clinics by placing the child's hand firmly on the silver chromate filled agar. Any salt on the hands would produce a strong white silver chloride image of the hand. This dramatic and immediately apparent test could give a false negative test if a CF patient had newly washed hands and could give a false positive test when a non-CF patient had been eating salty hand food such as potato chips. Because patients with positive screening tests and negative screening tests in patients who had classic CF related diseases still needed to be referred to the sweat test laboratory for the GCST/QPIT, this screening test is no longer used.

Over the 40 years the physiological basis for the GCST/QPIT has been known, efforts to develop a screening test that might be accepted as being as accurate as the GCST/QPIT, which might be done in a non-CF Center Laboratory, have received little support from CF physicians and Foundations.

Never-the-less three such tests have reached the attention of enough CF Center Directors to warrant a multi-CF Center study of their efficiency. These tests were (1) the CF Indicator System (15) a compact configuration of manufactured electrodes that dispense pilocarpine and a manufactured chloride sensor patch that collected a standard amount of sweat and was read as "normal", "CF" or "questionable" (16); (2) the chloride electrode *in situ* measurement of pCl after sweat stimulation by pilocarpine iontophoresis (17); and (3) the Macroduct system that used pilocarpine iontophoresis with visible collection of sweat in a plastic tube followed with conductivity or osmolarity measurements to match with a comparable concentration of sodium chloride (17).

The conclusions of the Cystic Foundation were that in the hands of community laboratories the potential for errors was so large that these tests should be considered only as screening sweat tests and that diagnostic GCST/QPIT tests for diagnosis should be done only at CF Center sweat testing laboratories.

Never-the-less the pressure to generalize sweat testing so CF diagnosis could be done before referral to CF Centers continued to be a powerful pressure for improving these three technologies. The CF Indicator was redesigned into a new integrated system that has been built, patented, and tested in one study (18). This study showed that compared to the GCST/QPIT the Quantum Patch (19) had equal sensitivity (94%) and specificity (99%), but differed in rate of failed tests, 1% for the Quantum Patch as compared to 15% for the GCST/QPIT. The Quantum test was faster, calculated the required amounts of sweat (3–10 mg) compared to the extra step of weighing and the larger amount of sweat (70 mg) required for the GCST/QPIT, was simpler to perform, required less equipment, less expensive equipment, and was less operator dependent. As of February 2005 the Quantum Patch technology, which includes a stimulator with disposable electrode, the Quantum Patch, and a scanner that simultaneous scans the patch and calculates the weight of the sweat and the chloride concentration, has not yet received FDA approval. The manufacturing standardization of the Quantum Patch test eliminates the substantial requirements for laboratory control and supervision and shortens the time for analysis to < 5 min compared to the GCST/QPIT required time of > 1 h. The simplification and the brief time needed may make this a diagnostic sweat test that could be done in a doctor's office or clinic. This rapid, simple, and quantitative pilocarpine iontophoresis sweat test has yet to be vetted by CF Scientists who have no financial interest in the product.

The chloride electrode (20–23) measures the pCl providing a true and immediate measurement of the chloride content of the sweat, but without a measure of the weight of the sweat. This technique might be resuscitated if an authoritative organization would mandate strict and inviolate guidelines for performance. If this could happen and if the CF Physicians and Organizations would accept such *in situ* measurements without knowing the weight of sweat or the rate of sweating, then diagnostic results could be known immediately. Given the strictness of the mandated strict and inviolate guidelines for performance this test might still be confined to CF Centers and other authorized Sweat Test Laboratories.

The Macroduct system is able to measure two abnormalities of the mixed-ion content of cystic fibrosis sweat. Both osmolarity and conductivity measure the ignored abnormality, the increased amount of all electrolytes in the sweat, and so have ignored what might be proven as an equal or better way to discriminate CF from non-CF subjects. Unfortunately, instead of adding new science to the study of this CF sweat abnormality, the manufacturers have reported the conductivity, or the osmolarity, of the sweat chloride concentration of a salt solution with that amount of conductivity or osmolarity. While such adjusted sweat chloride values can be used to discriminate between CF and non-CF subjects (24), the numbers are nonphysiological and so are offensive to and rejected by CF Center Directors and Sweat Test Laboratory Directors. In addition, the sweat chloride numbers confuse some practicing physicians who misdiagnosis some patients as carriers or

even patients when the MacroDuct pseudosweat chloride numbers are in the GCST/QPIT intermediate or low CF ranges. Fortunately, two groups (25,26) have demonstrated that either conductivity or osmolarity can be used as new demonstrations of diagnostic alterations of sweat gland secretion. Such efforts should be encouraged.

THE POTENTIAL FOR A GCST/QPIT SUCCESSOR

The excellent work done so far has not closed the possibility that an improved sweat test cannot be developed (27). In a multistep procedure, nonstimulated sweat was collected for 10 min from the surfaces of either thumb for 10 min, while the rate of sweating was measured from the other thumb. The thumb collections are reversed and repeated a second time. Calculations "estimate the chloride concentration by dividing the amount of chloride per unit area of one finger by the amount of sweat per unit area of the other finger". The sweat chloride values were similar to GCST/QPIT sweat chloride values. Most laboratories will find this technique more tedious and filled with potential for errors than the GCST/QPIT method which, because of its reliability, accuracy, and dependability, remains the gold standard for the diagnosis of CF. The QPIT/GCST is and has been the worldwide standard because of its accuracy. However, because of the many sources of potential error and the detailed steps that are needed to ensure that accuracy the search will continue to find an equally accurate replacement.

At this time there are three candidates.

1. The Quantum Patch has been designed to have "all" sources of errors from the preparation of iontophoresis solutions, stimulation time, collection quantities, and computerized measurement of sweat weight and chloride concentration standardized by the manufacturer to give virtually 100% successful tests with sensitivity and specificity equal to the GCST/QPIT. Unfortunately, at this time the only published paper was published by the developers who have a financial interest in the product and the product is awaiting FDA 510K approval. If this product is well vetted by CF Centers and Sweat Test Laboratories it will be worthy of large scale tests by others. The Quantum Patch meets all the QPIT standards.
2. The chloride electrode test has had mixed reviews in the literature. It has the potential to deliver chloride concentrations within minutes after pilocarpine iontophoresis. However, because of inadequate attention to the details of performance the technology has done poorly in many hands. It has one major defect in that the weight of sweat or the rate of sweating are unknown. If standards for maintenance of equipment and of performing the test were to be simplified and standardized to CF Center and Sweat Test Laboratories requirements this technology could supersede the GCST/QPIT. Unfortunately, at this time there is little interest in such development. The chloride electrode test does not meet all of the QPIT standards.
3. The MacroDuct system is the only system currently manufactured, supported, and maintained. It is reliable and has acquired some supporters among CF Center Directors despite that it measures only conductivity or osmolarity which, with rare exceptions, is converted to the chloride content of NaCl solutions of similar conductivity or osmolarity. The novel collection system is excellent and provides the potential for quantifying both sweat rate and weight. The MacroDuct system is the only method that takes advantage of the initial description of the sweat abnormality, the increased sweat concentrations of Na, K, and Cl, but loses that advantage by reporting the mixed ion content of sweat by specimen, conductivity, or osmolarity, results equal to a solution of NaCl. The manufacturers and the CF Centers that favor this test, with two published exceptions, have failed to develop ion content as an alternative and as an equal or more specific abnormality characterizing CF. The pseudochloride concentration carries the risk of misdiagnosis, therefore this test will probably continue to carry the label of screening test in most CF physician's minds. New science, study of total ion content of sweat, might make this method a suitable replacement for the GCST/QPIT. There is no indication that the manufacturers are willing to make such an effort. As marketed and at this writing the MacroDuct system does not meet all of the QPIT standards.

BIBLIOGRAPHY

1. Groman JD, Meyer ME, Wilmott RW, Zeitlin PL, Cutting GR. Variant Cystic Fibrosis Phenotypes in the Absence of CFTR Mutations. N Eng J Med Aug 8, 2002;347:401–407.
2. Boyle MP. Nonclassic cystic fibrosis and CFTR related diseases. Curr Opin Pulm Med 2003;9(6):498–503.
3. Jiang Q, Engelhardt JF. Cellular heterogeneity of CFTR expression and function in the lung: implications for gene therapy of cystic fibrosis. Eur J Hum Genet 1998;6:12–31.
4. Reddy MM, Light MJ, Quinton PM. Activation of the epithelial Na(+) channel (ENaC) requires CFTR CI(−) channel function. Nature (London) 1999;402:301–304.
5. Callen A, Diener-West M, Zeitlin PL, Rubenstein RC. A simplified cyclic adenosine monophosphate-mediated sweat rate test for quantitative measure of cystic fibrosis transmembrane regulator (CFTR) function. J Pediat 2000;137:849–855.
6. Kessler WR, Andersen DH. Heat prostration in fibrocystic disease of the pancreas and other conditions. Pediatrics 1951;8:648–56.
7. di Sant' Agnese PA, Darling RC, Perera GA, et al. Abnormal electrolyte composition of sweat in cystic fibrosis of the pancreas: clinical implications and relationship to the disease. Pediatrics 1953;12:549–563.
8. Darling RC, diSant'Agnese PA, Perera GA, Andersen DH. Electrolyte abnormalities of sweat in fibrocystic disease of pancreas. J Med Sci 1953;225:67–70.
9. Di Sant' Agnese PA, Darling RC, Perera GA, et al.Abnormal electrolyte composition of sweat in cystic fibrosis of the pancreas: clinical implications and relationship to the disease. Pediatrics 1953;12:549–563.
10. Gibson LE, Cooke RE. A test for concentration of electrolytes in sweat in cystic fibrosis of the pancreas utilizing pilocarpine electrophoresis. Pediatrics 1959;23:545–549.

11. Gibson LE, Cooke RE. A test for concentration of electrolytes in sweat in cystic fibrosis of the pancreas utilizing pilocarpine electrophoresis. Pediatrics 1959;23:545–549.
12. Smalley CA, Addy DP, Anderson CM. Does that child really have cystic fibrosis? Lancet 1978 Aug. 19; 2(8086):415–417.
13. Baumer JH. Evidence based guidelines for the performance of the sweat test for the investigation of cystic fibrosis in the UK. Arch Dis Childhood 2003;88:1126–1127.
14. Shwachman H, Mahmoodian A. Reappraisal of the chloride plate test as screening test for cystic fibrosis. Arch Dis Child 1981;56(2):137–139.
15. Warwick WJ, Huang NN, Waring WW, Cherian AG, Brown I, Stejskal-Lorenz E, Yeung WH, Duhon G, Hill JG, Strominger D. Evaluation of a cystic fibrosis screening system incorporating a miniature sweat stimulator and disposable chloride sensor. Clin Chem 1986;32(5):850–853.
16. Warwick WJ, Hansen LG, Werness ME. Quantification of chloride in sweat with the Cystic Fibrosis Indicator System. Clin Chem 1990;36(1):96–98.
17. Denning CR, Huang NN, Cuasay LR, Shwachman H, Tocci P, Warwick WJ, Gibson LE. Cooperative study comparing three methods of performing sweat tests to diagnose cystic fibrosis. Pediatrics 1980;66(5):752–757.
18. Warwick WJ, Hansen LG, Brown IV, Laine WC, Hansen KL. Sweat chloride: quantitative patch for collection and measurement. Clin Lab Sci 2001;14(3):155–159.
19. Warwick WJ, Hansen LG, Brown IV, Laine WC, Hansen KL. Sweat chloride: quantitative patch for collection and measurement. Clin Lab Sci 2001;14(3):155–159.
20. Hansen L, Buechele M, Koroshec J, Warwick WJ. Sweat chloride analysis by chloride ionspecific electrode method using heat stimulation. Am J Clin Pathol 1968;49(6):834–841.
21. Warwick WJ, Hansen LG. Measurement of chloride in sweat by use of a selective electrode and strip-chart recorder. Clin Chem 1978;24(2):381–382.
22. Warwick WJ, Hansen L. Measurement of chloride in sweat with the chloride-selective electrode. Clin Chem 1978;24(11): 2050–2053.
23. Warwick WJ, Hansen L, Viela I, Matheson J. Comparison of the chloride electrode and gravimetric chloride titration sweat tests. Am J Clin Pathol 1979;72(2):142–145.
24. Hammond KB, Turcios NL, Gibson LE. Clinical evaluation of the macroduct sweat collection system and conductivity analyzer in the diagnosis of cystic fibrosis. J Pediatr 1994; 124(2): 255–260.
25. Lezana JL, Vargas MH, Karam-Bechara J, Aldana RS, Furuya ME. Sweat conductivity and chloride titration for cystic fibrosis diagnosis in 3834 subjects. J Cyst Fibros 2003;2(1): 1–7.
26. Barben J, Ammann RA, Metlagel A, Schoeni MH. Conductivity determined by a new sweat analyzer compared with chloride concentrations for the diagnosis of cystic fibrosis. J Pediatr 2005;146(2):183–188.
27. Naruse S, Ishiguro H, Suzuki Y, Fujiki K, Ko SB, Mizuno N, Takemura T, Yamamoto A, Yoshikawa T, Jin C, Suzuki R, Kitagawa M, Tsuda T, Kondo T, Hayakawa T. A finger sweat chloride test for the detection of a high-risk group of chronic pancreatitis. Pancreas 2004;28(3):e80–e85.

Further Reading

Warwick WJ. Cystic fibrosis sweat test for newborns. *JAMA* 1966 Oct 3; 198(1):177–180.
Warwick WJ, Hansen LG. Measurement of chloride in sweat by use of a selective electrode and strip-chart recorder. Clin Chem 1978;24(2):381–382.
Warwick WJ, Hansen L, Viela I, Matheson J. Comparison of the chloride electrode and gravimetric chloride titration sweat tests. Am J Clin Pathol 1979;72(2):142–145.
Warwick WJ, Viela I, Hansen LG. Comparison of the errors due to the use of gauze and the use of filter paper in the gravimetric chloride titration sweat test. Am J Clin Pathol 1979;72(2):211–215.
Denning CR, Huang NN, Cuasay LR, Shwachman H, Tocci P, Warwick WJ, Gibson LE. Cooperative study comparing three methods of performing sweat tests to diagnose cystic fibrosis. Pediatrics 1980;66(5):752–757.
Warwick WJ, Huang NN, Waring WW, Cherian AG, Brown I, Stejskal-Lorenz E, Yeung WH, Duhon G, Hill JG, Strominger D. Evaluation of a cystic fibrosis screening system incorporating a miniature sweat stimulator and disposable chloride sensor. Clin Chem 1986;32(5):850–853.
Warwick WJ, Hansen LG, Werness ME. Quantification of chloride in sweat with the Cystic Fibrosis Indicator System. Clin Chem 1990;36(1):96–98.
Warwick WJ, Hansen LG, Brown I. Improved correlation of sweat chloride quantification by the CF Indicator System and the Gibson-Cooke Sweat Test. Clin Chem 1993;39(8):1748.
Warwick WJ, Hansen LG, Brown IV, Laine WC, Hansen KL. Sweat chloride: quantitative patch for collection and measurement. Clin Lab Sci 2001;14(3):155–159.

See also BIOHEAT TRANSFER; FLAME ATOMIC EMISSION SPECTROMETRY AND ATOMIC ABSORPTION SPECTROMETRY.

CYTOLOGY, AUTOMATED

HARRY W. TYRER
University of Missouri-Columbia
Columbia, Missouri

INTRODUCTION

Cytology is the scientific study of cells, which includes their origin, structure, and function. Automated cytology arose from the important medical problem of classifying and enumerating cell types using instrumented means primarily for speed improvements. Automated cytology has also contributed to elucidating cell origin, structure, and function. It provides quantitative methodology directed to establishing relationships between variables to predict cellular behavior.

Understanding automated cytology requires knowledge of cells, the fundamentals of measurements on cells, and subsequent processing of these data. Present and future applications in automated cytology justify the value of automated cytology as well as define its usefulness and promise.

Cytology (Greek: *kytos*, hollow vessel; *logos*, word, reason) as a discipline primarily focuses on cell structure. Cell structure deals with those factors that define the shape and spatial distribution of the components within a cell. It is synonymous with cell morphology, which historically has been the primary method to describe cells. More generally, cytology must also encompass cellular function, which describes a cell's operational characteristics. For example, a red blood cell is a bag of hemoglobin that carries oxygen to all parts of the body, a lymphocyte produces antibodies as a result of stimulation, and squamous cells, which are on the

surface of the skin, become leathery to protect the exposed skin from mechanical assault.

Automated cytology is an area of multidisciplinary specialization initially oriented toward the automated identification and classification of cells. The resulting measurement and computational methodologies provide a database from which relationships between the variables measured can be expressed. These relationships harbor the promise of establishing predictive relationships so that future structural and functional cellular characteristics can be determined. By extension then, the behavior of organs and organisms can then be determined.

CELLS

If we examine unstained cells with a microscope it is apparent that the cell is encased in a membrane. Within that membrane is a second structure encased within its own membrane. Thus the cell is subdivided into the nucleus and the cytoplasm. The nucleus is the control center and the cytoplasm carries out the function to which the cell differentiated. The work described here deals with nucleated or eukaryotic cells; incidentally cells without nuclei are called prokaryotic.

We now add a nucleophylic dye, such as hematoxylin, which binds preferentially to the nucleus. The nucleus is selectively stained and becomes much more apparent than the cytoplasm (Fig. 1). Note that both the nucleus and cytoplasm show spatial variation in light intensity. The cause of these variations can be analyzed by increasing resolution, which may include the use of electron microscopy. As we increase our ability to see into the cell either by optical or electron microscopy, we identify inclusions within the cells, which are called organelles. From Fig. 2, we consider some well-defined organelles.

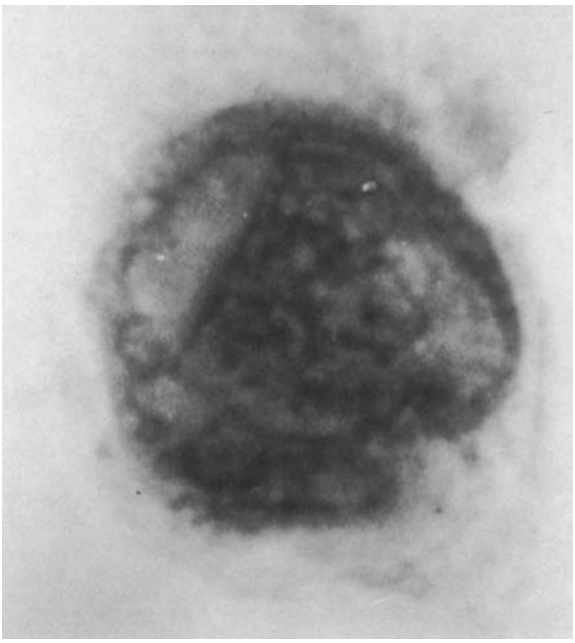


Figure 1. A cell, the lighter outside area of which is the cytoplasm and the darker of which is the nucleus.

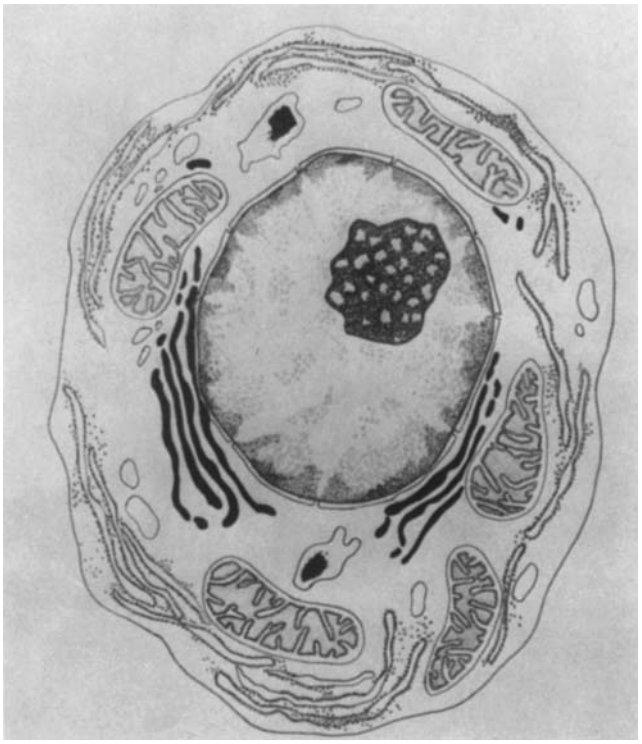


Figure 2. Drawing of a cell showing cytoplasmic and nuclear organelles. Cytoplasm: Drawing an imaginary line starting at 4 and ending at 10 o'clock, one traverses the following: cytoplasmic membrane, a granulated endoplasmic reticulum body, a single mitochondrion, a set of Golgi bodies, the nuclear membrane, and finally the nucleus. Continuing on simply reverses the order. Nucleus: On a line from 1 to 7 o'clock, one begins at the nuclear membrane to identify the nuclear elements: the nuclear membrane (note the infoldings at irregular intervals), the inner membrane, chromatin structure, nucleolus, the center of the nucleus, and finally the inner nuclear membrane.

The nucleus, usually the largest organelle, consists almost entirely of nucleic acids. Both deoxyribonucleic (DNA) and ribonucleic acid (RNA) are present. By far, the major constituent is DNA. Within the nucleus and under appropriate conditions DNA molecules replicate giving rise to DNA synthesis and then to cell division. An organelle within the nucleus is the nucleolus. In Fig. 2, it is the darkened mottled body encased in the nucleus. It consists mostly of RNA. In the initial steps of transcription to make protein, DNA produces messenger RNA (mRNA) according to the genetic code. The mRNA then passes through the nuclear envelope into the cell cytoplasm.

The cytoplasm is the entity that performs the cell's function. These functions are carried out by several organelles. The same organelle of different cells may be substantially different so that the cell may focus its energy to carry out its usually single function.

The mytochondrion (in Fig. 2, the ovoid body with the complex curved inclusions) is an organelle that is responsible for producing the energy required by the cell. It has its own DNA and is able to replicate itself within the cell.

The Golgi body complex (the long dark strands) provides temporary storage of secretory substances and connects to the endoplasmic reticulum.

The endoplasmic reticulum (light strand) is of two types in the same cell: with granules (the dots) and without granules. The granules are ribosomes. In the final stages of transcription (making protein from DNA), the mRNA travels from the nucleus to the ribosomes (probably in the endoplasmic reticulum) and attaches to the ribosome. The ribosome then binds the transfer RNA (tRNA) corresponding to the genetic code expressed in the mRNA. Each value of the genetic code has one tRNA, which in turn corresponds to one amino acid. One ribosome houses two tRNAs that connect the proper amino acids together to the protein chains. The resulting protein (called a polypeptide) passes through the Golgi complex to be secreted from the cell. Thus the function of a particular cell is to produce a particular protein. Additionally, binding a signal protein will alter the state of the cell, causing it to produce a different protein or to divide and function in a different manner.

As just shown, we can identify cell structure by binding chemicals to the structure to allow its observation. On the other hand, cellular products are determined by several methods. First, one can selectively poison a particular functional entity and compare the results that are produced from nondestroyed entities. Second, one can separate the entities and selectively return them until the tested functionality has been restored. These techniques are routinely performed on samples containing many cells. Unfortunately, the average population behavior masks the individual cell behavior, which may be of interest as a rare event.

Consequently, automated cytology has evolved to enumerate and provide quantitative information on the structural features as well as the functional capabilities of a single cell. Work is done on an intact single cell, which may be viable or nonviable. Viable means that the cell is able to perform its functional activities, whereas a nonviable cell cannot. A nonviable cell is preserved in an appropriate fixative. Such a fixed cell can withstand many of the rigors of experimental treatment compared to a viable cell. Unfortunately, fixation may alter properties that are to be measured. Thus, choice of fixation (including no fixation) requires careful consideration.

CYTOCHEMICAL PROBES

Cytochemistry is the chemical organization and activity of a cell. Numerous chemical probes have been reported for use in determining cell structure and function. The probe binds to the molecule to be detected. If a stochiometric relationship between the probe and detected molecule is maintained (i.e., conditions for binding are fixed so that the relationship between the amount of probe and detected molecule are fixed), then relative quantities of the detected molecule can be obtained. Furthermore, if the amount of detected molecule to quantity of probe molecule can be established, absolute standardization is achieved.

Biochemical probes are used in automated cytology to detect a wide range of structural and functional characteristics of cells. For example, let us consider the steps involved in measuring DNA in single cells. Specifically we want to see the activation of DNA synthesis over several days. White blood cells withdrawn from fresh human blood, can be stimulated to replicate by placement in an appropriate tissue culture system. We can remove several cells from the tissue culture system every 24 h, and determine the relative amount of DNA in each cell. A cytochemical probe specific for DNA is propidium iodide (PI), which fluoresces in the red when excited by a 488-nm (blue) light. The fluorescence intensity of each cell is proportional to the amount of PI bound to the DNA, assuming care has been taken to minimize non-DNA binding of PI. From this, we form a histogram of the fluorescence intensity of the cells; that is, we plot for each value of fluorescence the number of cells (frequency) expressing that fluorescence value. The series of histograms in Fig. 3 shows that at first there is a

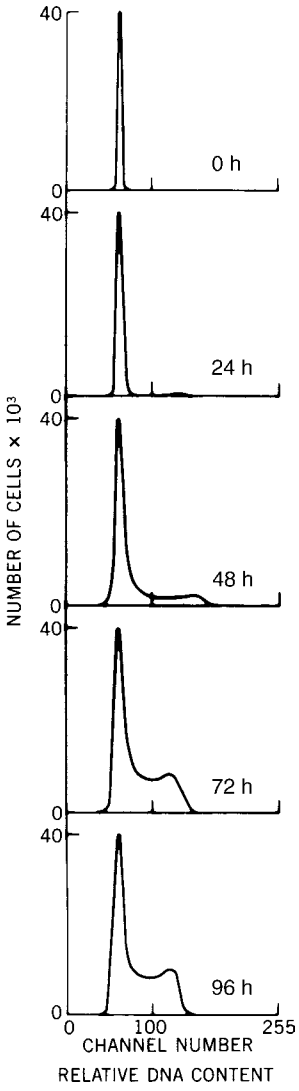


Figure 3. Each of the graphs displays the DNA distribution of cells in tissue culture. Cells were removed from culture and analyzed every 24 h. The top chart displays the DNA distribution of cells initially placed in culture. The succeeding charts show increasing numbers of cells with increasing amounts of DNA over the 4-day intervals sampled.

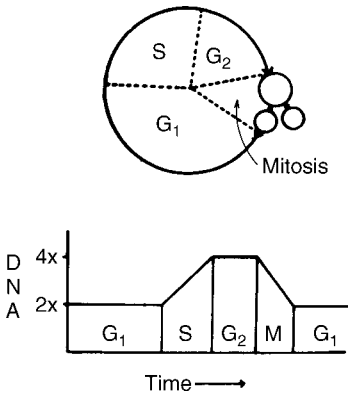


Figure 4. Cell cycle diagram and the corresponding graph of DNA quantity versus time. The Parameter G_1 is gap 1, a resting state; S is when the cell synthesizes DNA; G_2, gap 2, another resting state; and M is mitosis in which the cell divides.

single population of cells in the blood with a remarkably narrow range of DNA value. Within 24 h there is a barely visible set of cells with twice the normal amount of DNA; by 48 h and beyond, it is apparent that there are also cells with values intermediate to the two values. These distributions demonstrate DNA synthesis and cellular replication.

It is commonly known that cells replicate by dividing, creating daughter cells. In the process of division, from some resting state, the cell synthesizes DNA until two times the normal amount of DNA is present. At this point, the two nuclei are formed and the cell subsequently divides into two presumably identical parts. Figure 4 graphically illustrates the cycling of the cells. At the portion of the circle labeled G1 (resting state, Gap 1), a cell has two times the DNA shown in the lower part of Fig. 4. (Since cell DNA quantity is variable between species, we list the amount of DNA as *X*; furthermore, egg and sperm have 1*X* DNA, but most mature cells in the body normally have 2*X* DNA.) As time increases, the cell goes into the S phase of the cycle where it synthesizes DNA. Finally, the cell has 4*X* DNA and waits some amount of time at G2 (Gap 2) until it divides. The cell divides (M, mitosis), generating two daughter cells in the G1 phase of the cycle.

These observations are based on total DNA of each cell. Had we also observed a cell in mitosis, under proper microscope and staining conditions, we would have observed the nucleolus condensing into well-defined bodies called chromosomes. Preparation of chromosomes is made from cells undergoing division. Such preparations are used to study genetic observations.

The same cytochemical principles are used to classify human chromosomes. DNA consists of chains of adenosine—thiamine (AT) or guanine—cytosine (GC) base pairs. A strategy to increase the information content in measuring the DNA of chromosomes is to use two dyes simultaneously; one with selective affinity to the AT bases the other with affinity to GC bases. Under proper excitation, each chromosome will produce a fluorescence intensity value from each of the two dyes. These two intensities are each proportional to the number of AT or GC bases of each chromosome. Histograms of these two parameters are two dimensional, and each point in the histogram is the number of chromosomes (or particles) with the same pair of fluorescence intensity values. For example, Chromomycin A3 (CA3) has an affinity for AT bases, whereas Hoechst bis-benzimidizole 33258 binds preferentially to GC bases. The increased information provided by the use of the second fluorescent stain increased the resolution for separating the 23 human chromosomes compared to each stain alone.

We have used DNA quantity as an example of the use of cytochemical probes. There are a large number of DNA probes in current use as shown in Table 1. The bibliography can be consulted for additional information on their properties and use.

Cytochemical probes are, of course, useful for detection and quantization of other cellular structures and functions. Table 2 displays a representative set of such probes.

Table 1. Biochemical Probes with DNA Specificity

Probe[a]	Base-Pair Enhancement	Spectrum Maximum, nm		Extinction Coefficient, wavelength, pH[b]
		Excitation	Fluorescence	
Acridine Orange	None	490	530	50 (492, 7)
Ethidium bromide	None	546	610	517 (480, 7)
33258	AT	365	450	4 (338, 7)
Daunomycin	AT	500	600	14.8 (480, 7)
EK-4	AT	530	560	36 (570, 7)
LL-585	AT	500	600	32 (510, 7)
Proflavin	AT	460		34.9 (454, 7)
Quinacrine	AT	436	525	8.9 (424, 7)
VL-772	AT	500	600	32 (510, 7)
7-AAD	GC	546	610	21.5 (502, 7)
Chromomycin A3	GC	440	585	8 (490, 7)

[a]Abbreviations: 33258 = Hoechst bis(benzimidizole) No. 33258; 7-AAD = 7-aminoactinomycin D; A = adenine; C = cytosine; EK-4 = 26-diphenyl-4-(4-dimethylaminophenyl) pyriliumbisulfate; G = guanine; LL-585 = 6-benzothiazolyl-3-ethyl-2-(dimethylamino-stytyl)-benzothiazolium,-*p*-toluenesulfonate; T = thymine (T); VL-772 = 6-(dimethylamino)-2-[2,5-dimethyl-1-pheny-1*H*-pyrrol-3-yl ethenyll]-1-methyl-quinolinium methosulfate.
[b]Extinction coefficient $\times 10^{-3}/(M\,\mathrm{cm})$; in parentheses are the values of measurement wavelength (in nanometers) and the pH, respectively.

Table 2. Fluorescent Probes to Assay Cell Structure and Function

Cell structure	Probe
RNA content	Acridine Orange, pyronin
Total protein	Fluorescein isothyocyanate
Cell cytoskeleton	NBD-phallacidin
Cell mitochondria	Fluorescent tetracyclenes
Cell function	
Membrane permeability (cell viability)	Fluorescein diacetate, propidium iodide, 33342[a]
Membrane potential (cytoplasm and mitochondria)	3,3′-Dihexyloxacarbocyanine
Calcium (Ca^{2+})	
Membrane bound	Chlortetracycline
Cytoplasmic	Quin-2
Enzyme activity	Fluorescein derivatives, methylumbelliferyll derivative, napthol derivatives
Intracellular pH	1,4-Diacetoxy-2,3-dicyanobsenzene, carboxyfluorescein
Phagocytosis	Fluorescent beads, fluorescent-stained bacteria and other particles

[a]Hoechst bis(benzimidizole) No. 33342.

Finally, the biochemical constituents of the cells have physical properties that are directly measurable. A well-known example is the pigment that makes blood cells red, hemoglobin. Furthermore, it is well known that the important metabolic constituent reduced nicotinemide adenine dinucleotide (NADH) fluoresces as does riboflavin, one of the 13 vitamins. These compounds have optical properties that are useful because of the convenience of use with existing equipment. This is not true of the majority of cell constituents whose physical properties may not be easily detected or measured. For example, proteins with aromatic amino acids, such as tryptophan, will fluoresce when excited by 257 nm light. Such a short wavelength is not conveniently handled because its propagation attenuation in standard optical materials is so severe.

MEASUREMENT OF CELLULAR PARAMETERS

Automated cytology implies that cellular parameters are to be measured. Since the visual quasiquantitative analyses described in the previous section cannot provide the accuracy and discrimination required, technologies have arisen that measure physical phenomena to describe cellular parameters. Commercially available instrumentation use optical or electrical phenomena whereas experimental devices have been used with acoustic, magnetic, and a variety of spectroscopic techniques.

Probes for nucleic acids abound. The probe binding mechanics to the nucleic greatly influences their use and application. Table 3 shows a list of such probes along with their excitation and emission wavelengths.

Protein probes tend to be nonspecific and usually stick to the protein. The stain can be removed by (sometimes) vigorous rinsing. Table 4 shows the excitation and emission wavelengths of various probes used for proteins. The probes are also bound to specific antibodies so that multi wavelength emission due to multiple antibody binding provides a multiparametric analysis. Common pairs are flouresceim isothyocianate (FITC) and Rhodamine (see Tables 4 and 13).

Table 3. Fluorescent Probes for Nucleic Acids with Excitation and Emission Wavelengths in nm

Probe[a]	Excitation	Emission
Hoechst 33342 (AT rich) (UV)	346	460
DAPI (UV)	359	461
POPO-1	434	456
YOYO-1	491	509
Acridine Orange (AO) (RNA)	460	650
Acridine Orange (DNA)	502	536
Thiazole Orange (vis)	509	525
TOTO-1	514	533
Ethidium Bromide	526	604
PI (UV/VIS)	536	620
7-Aminoactinomycin D (7AAD)	555	655

[a]Abbreviations: DAPI = 4′,6-diamidino-2-phenylindole; POPO-1, YOYO-1,TOTO-1 are cyanine dimers available from Molecular Probes, Inc.; PI: Propidium Iodide; UV = ultraviolet, vis = visible.

There are a variety of probes of importance to identifying cell parameters. These require care in their use. The following three tables (Tables 5–7) list such probes along with their excitation and fluorescent emission wavelengths. Ion probes are listed in Table 5, pH sensitive indicators appear in Table 6, and probes for oxidation states along with the oxidant appear in Table 7.

Finally, just as there is a dichromatic display for RNA and DNA with acridine orange, where AO can be seen to bind to nucleolus as a red fluorescence and to the nucleus as green fluorescence, so several probes have an affinity for specific organelles. Specifically, the Golgi bodies and mitochondria can be identified with the appropriate stain as Table 8 shows. Finally, the lipid stains help to identify the cell's membrane and other lipids.

Optical Measurements

By far, most measurements on cells are optical. Spectra obtained from a multiplicity of physical properties are

Table 4. Fluorescent Probes for Proteins with Excitation and Emission Wavelengths in nm

Probe[a]	Excitation	Emission
FITC	488	525
PE	488	575
APC	630	650
PerCP	488	680
Cascade Blue	360	450
Coumerin-phalloidin	350	450
Texas Red	610	630
Tetramethylrhodamine-amines	550	575
CY3 (indotrimethinecyanines)	540	575
CY5 (indopentamethinecyanines)	640	670

[a]Abbreviations: PE = Phycoerythrin; APC = allophycocyanin; PerCP = peridinin chlorophyll.

Table 5. Fluorescent Probes for Ions with Excitation and Emission Wavelengths in nm

Probe[a]	Excitation	Emission
INDO-1	350	405/480
QUIN-2	350	490
Fluo-3	488	525
Fura-2	330/360	510

[a]**INDO-1** = 1*H*-Indole-6-carboxylic acid, 2-[4-[bis[2-[(acetyloxy)methoxy]-2-oxoethyl]amino]-3-[2-[2-[bis[2- [(acetyloxy)methoxy]-2-oxoetyl]amino]-5-methylphenoxy]ethoxy]phenyl]-, (acetyloxy)methyl ester [$C_{47}H_{51}N_3O_{22}$], **FLUO-3** = Glycine, *N*-[4-[6-[(acetyloxy)methoxy]-2,7- dichloro-3-oxo-3*H*-xanthen-9-yl]-2-[2-[2- [bis[2-[(acetyloxy)methoxy]-2- oxyethyl]amino]-5-methylphenoxy]ethoxy]phenyl]-*N*-[2- [(acetyloxy)methoxy]-2-oxyethyl]-, (acetyloxy)methyl ester.

useful in identifying and quantitating cellular biochemical and physical parameters. The most popular optical measurements, namely, light scatter, absorption spectroscopy, and fluorescence spectroscopy, and their use in automated cytology are discussed.

Light Scattering. Light scattering from a single cell has been used to characterize cell volume, shape, internal structures, and other cell properties that produce changes in index of refraction. Light scattering includes diffraction, refraction, and reflection of light from a single cell. From Maxwell's equations, a series expression was developed by Mie for plane wave propagation perturbed by a solid nonconductive homogeneous sphere with a diameter on the order of the wavelength of illuminating light. Although some cells are spherical and their diameter approximates the wavelength of visible light, cells are not homogeneous. In fact, it may be their optically heterogeneous structures that are important and, therefore, cannot be neglected. There is a large body of literature devoted to light scattering and some of it is directed to the problem of aerosol detection.

Measurements have established the relationship between the amount of light scattered and the size of the cell. For small-angle scattering ($< 2^\circ$ numerical aperture), the signal intensity is proportional to the diameter cubed (in agreement with the Mie solutions). This proportionality is linear and monotonically increasing over a restricted size range. As the numerical aperture increases the relationship is no longer linear and is not monotonic over the entire possible range of sizes. Cell shape also influences scattering. Since most cells are not spherical,

Table 6. Fluorescent pH Sensitive Indicators with Excitation and Emission Wavelengths in nm

Probe[a]	Excitation	Emission
SNARF-1	488	575
BCECF	488	525/620
BCECF	440/488	525

[a]**SNARF-1** = Benzenedicarboxylic acid, 2(or 4)-[10-(dimethylamino)-3-oxo-3*H*- benzo[c]xanthene-7-yl]-, **BCECF** = Spiro(isobenzofuran-1(3*H*),9′-(9*H*) xanthene)-2′,7′-dipropanoic acid, ar-carboxy-3′,6′-dihydroxy-3-oxo-.

Table 7. Fluorescent Probes for Oxidation States with Excitation and Emission Wavelengths in nm

Probe[a]	Oxidant	Excitation	Emission
DCFH-DA	(H_2O_2)	488	525
HE	(O_2^-)	488	590
DHR 123	(H_2O_2)	488	525

[a]DCFH-DA = dichlorofluorescin diacetate, HE = - hydroethidine 3,8-Phenanthridinediamine, 5-ethyl-5,6-dihydro-6-phenyl-, DHR-123 = dihydrorhodamine 123 Benzoic acid, 2-(3,6-diamino-9*H*-xanthene-9-yl)-, methyl ester.

the relative orientation of the cell to the source and detector can produce different scatter fields. Furthermore as the numerical aperture of detection increases the internal structure is increasingly detected. On some instrumentation, simultaneous detection of 0 and 90° scattering is used to discriminate cells based on internal structures. Light scattering under a highly coherent light source, such as a laser, is a very sensitive indicator of variations of index of refraction. Under normal conditions the cell is encapsulated in some medium and the index of refraction between air and the cellular medium contribute to the entire signal. Furthermore, under abnormal conditions, local differences in index of refraction, as a result of improper fluid mixing, contribute signals that approximate those produced by a cell.

Measurements of light scattering from particle suspensions provide a single datum from which estimates of total particles are obtained. Such measurements produce a signal that is dependent on the number, size, and shape of the suspended particles. Systems measuring either scattering or absorption of suspensions cannot provide discriminatory information for individual particles.

Optical Spectroscopy. When energy interacts with matter, a sequence of events occurs that is explained in terms of the atomic or molecular behavior. This interaction provides quantitative information as well as identification of the species involved. The well-known proportionality of energy E to electromagnetic frequency v is related by Planck's constant (h), which underscores the discrete nature of

Table 8. Fluorescent Probes for Organelle with Excitation and Emission Wavelengths in nm

Probe[a]	Organelle	Excitation	Emission
BODIPY	Golgi	505	511
NBD	Golgi	488	525
DPH	Lipid	350	420
TMA-DPH	Lipid	350	420
Rhodamine 123	Mitochondria	488	525
DiO	Lipid	488	500
diI-Cn-(5)	Lipid	550	565
diO-Cn-(3)	Lipid	488	500

[a]BODIPY = borate-dipyrromethene complexes; NBD = nitrobenzoxadiazoxadiazole; DPH = diphenylhexatriene; TMA = trimethylammonium; DiO = diI-Cn-(5);diO-Cn-(3) = Carbocyanines (DiI, DiA, DiO), e.g., DiI 1,1′-dioctadecyl-3,3,3′,3′-tetramethylindocarbocyanine perchlorate.

atomic behavior:

$$h = 6.6 \times 10^{-34}\ \mathrm{J \cdot s}$$
$$E = hv \qquad (1)$$

For example, gas discharge tubes supply energy by means of a high voltage between electrodes onto a sealed container with small amounts of gas. The resulting arc has an emission spectrum consisting of lines of light energy at discrete wavelengths that are characteristic of the electronic structure of that gas. The first correct explanation of this phenomenon arose from the Bohr model of the hydrogen atom. The calculated values of energy from the various discrete orbits that the electron could have about the nucleus correspond to the spectral lines obtained from the gas discharge series.

This characteristic spectrum can be used to identify and quantitate the amount of element or compound. The interaction of the constituent atoms of a molecule produces a spectrum that is substantially different from that of each atom. The spectrum is characteristic of the molecule.

The response of the material to light excitation may by considered to occur in two general steps (Fig. 5). In a molecule excited by light, an electron is elevated in energy to an excited state: This results in light absorption. The relaxation of this excited electron back to a lower energy state may result in (1) light being emitted (fluorescence and phosphorescence), (2) heating of the substance, or (3) some combination of the two. The light is not only absorbed but may be scattered (Rayleigh scattering) or be modulated to a different wavelength by the rotational and vibrational motion of the illuminated molecules (Raman scattering).

The emitted light frequency is different from the absorbed frequency ν_1, except for Rayleigh scattering. In Raman scattering, it is possible for the emitted frequency to exceed the exciting (or absorbed) frequency. In the case of fluorescence or phosphorescence, the loss of energy in the molecule (due to collisions or vibration) results in a lower frequency of emission compared to the excitation (absorbed) frequency. Similarly, because of energy decrease, phosphorescence frequency is less than fluorescence frequency. For the remainder of this article we will no longer deal with frequency of light, but with the more commonly used parameter, wavelength.

Absorption is obtained from the loss of light in a material at a particular wavelength. A homogenous material is irradiated by light of intensity I_r (r, reference). As a result of transmission through the material there is a loss of light resulting in light intensity I_s (s, sample) emanating from the material (see Fig. 6). The optical density or absorbance can be defined from these two quantities, which is

$$\log_{10}(I_r/I_s) = A \qquad (2)$$

The absorbance A (for this discussion, always an upper case A) is directly related to the light path thickness d, the concentration of the material C, and by the proportionality constant ε called the extinction coefficient, which is a property of the material. The resulting equality is the Beer–Lambert law, which may be expressed as

$$A = \varepsilon dC \qquad (3)$$

The self-absorption of the sample material reduces the sensitivity to values $0.1 < A < 1.2$. Thus the material is suspended in nonabsorbing solution to disperse it to satisfy the homogeneity assumption. In practice, the

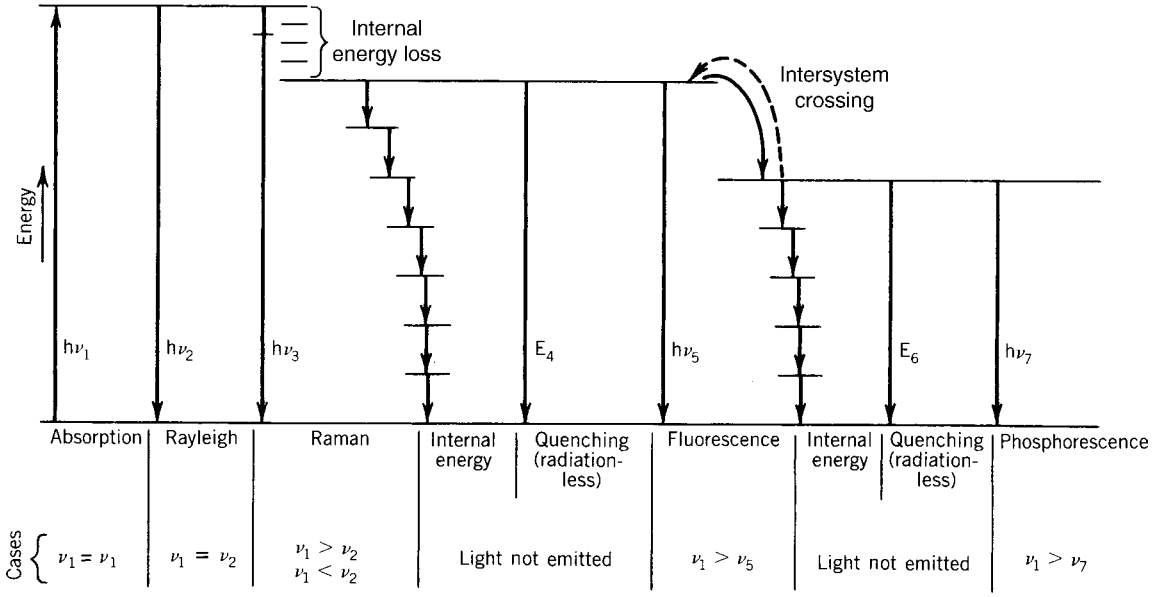


Figure 5. An energy diagram of light absorption by an atom. Light (energy) is provided into the system and is absorbed. It is then scattered (Rayleigh or Raman), dissipated in the system, or emitted by fluorescence or phosphorescence (h, Plank's constant; v, frequency of light).

concentration of the material in solution can be obtained by using a known concentration of the sample (standard) and determining the ratio of absorbance of the unknown to the standard:

$$\frac{A_{\text{unknown}}}{A_{\text{standard}}} = \frac{C_{\text{unknown}}}{C_{\text{standard}}} \tag{4}$$

For species differentiation, a spectrum is usually sufficient to uniquely identify that species.

It is useful to define the quantity "transmissivity" t, which is

$$t = I_s/I_r \tag{5}$$

In practice, the input intensity I_r is fixed and the sample modulates I_s. So, if the sample does not absorb light, $t = 1$, and if the sample completely blocks out the light, $t = 0$. Thus, it is convenient to speak of the transmission T, which is a percentage, as

$$T = 100t \tag{6}$$

The entire range of intensities is mapped between 0 and 100%.

A homogeneous mixture of particles and solutions can be expected to arise in conditions in which the measured geometries are very large compared to the absorbing particle size. In the microscopic environment, particle sizes are in the order of the measurement geometries, invalidating the homogeneous sample assumption. This is referred to as the distributional error.

The distributional error can be analyzed using Fig. 6 as follows: We wish to determine the true value of absorbance of the sample in solution. Assume that only the light passing through the area $[a_T = (l_1 + l_2)W]$ under consideration reaches a detector. Let us begin by assuming that the

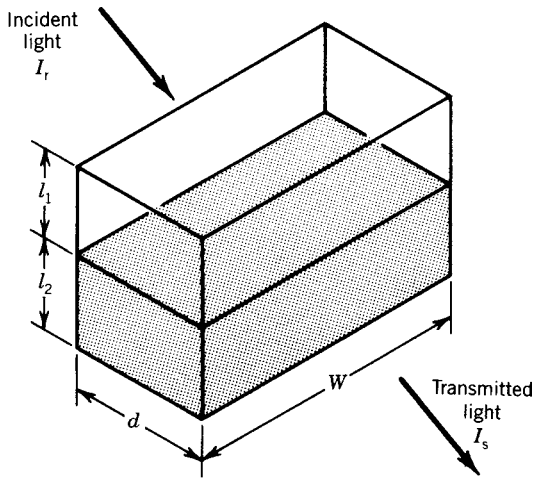


Figure 6. A uniform beam of light passing through a nonhomgeneous medium in a clear container. A material with extinction coefficient ε is suspended in a nonabsorbing medium. The material is uniformly distributed in the lower part of the container so it absorbs light by an amount $A_2 = dC_2$ over the surface l_2W. A different amount of the material is also uniformly distributed over the upper part of the container, absorbing light in the amount $A_1 = dC_1$ over l_1W.

entire mass m_T is distributed throughout the total volume (da_T). Express Eq. 3 as

$$A_T = \varepsilon d\left(\frac{m_T}{dWl}\right)$$

Now, if we assume, as shown in Fig. 6, an amount m_2 constitutes most of the absorbing material in the lower part of the volume and a different amount m_1 is in the upper part, we can substitute and cancel to get

$$A_T = \varepsilon\left(\frac{m_1 + m_2}{a_T}\right)$$

Now, we observe that the dry mass of the dissolved material is determined by $m = (A/)a$; that is, the viewing area a determines the measured dry mass, so that with further substitution and cancelation,

$$A_T = \frac{1}{a_T}(A_1a_1 + A_2a_2)$$

We can generalize: First, our derivation deals only with areas so that any arbitrary area could be used; second, we can expand the preceding equation to get

$$A_T = \frac{1}{a_T}\sum_{\text{all}\,i} A_i a_i \tag{7}$$

Thus, the solution to the distributional error problem in quantitative microscopy is to measure the absorption over a small area and sum each measurement so that the entire object is covered. It is assumed that each area is so small that it is homogeneous. Thus, the absorption of a cell at a given wavelength is obtained by measuring the absorbance of contiguous areas over the boundaries of the cell.

Finally, we observe that had we summed the individual transmittances (see Eq. 2), a different and incorrect value would have been obtained. Summing the absorbencies is based on the conservation of mass: total absorbance depends on total mass regardless of its distribution.

Fluorescence properties, wavelength, and intensity also provide information on molecular species and concentration, respectively. The fluorescence intensity F is proportional to the absorbed light intensity, I_a (a, absorbed), which is the difference between incident I_r and exiting I_s light intensity. The constant of proportionality q is related to the efficiency of the conversion of absorbed energy to emitted energy and is called the quantum efficiency,

$$q = \frac{\text{number of quanta emitted}}{\text{number of quanta absorbed}}$$

So that we get

$$F = I_a q$$

Now inserting Eq. (3) into Eq. (2) and solving for I_s, we get

$$I_s = I_r\, 10^{-\varepsilon dC}$$

so that

$$F = qI_r(1 - 10^{-\varepsilon dC}) \tag{8}$$

Again we have a parameter of material property and quantity relating to a measurable optical parameter. This relation using the McLaurin expansion, can be linearized by which assumes that $A < 0.05$ (e.g., small values of absorption). Now we can write

$$F = qI_r(2.3\,\varepsilon\,dC) \tag{9}$$

Again, if we have a standard and an unknown sample, the following relation holds, since we now assume that fluorescence is linearly related to concentration

$$\frac{F_{\text{unknown}}}{F_{\text{standard}}} = \frac{C_{\text{unknown}}}{C_{\text{standard}}}$$

where both values of F are measured, and C_{standard} is determined by the experimenter.

We have covered the fundamental issues of absorption and fluorescence spectroscopy directed to identifying and quantifying various molecular species. There are numerous other optical techniques that use biochemical probes to provide mechanistic information of cellular function. We will discuss only two: fluorescence polarization and resonance energy transfer.

Fluorescence polarization has been used to assess molecular motion with respect to the cell to which the molecule is bound. In general, the exciting light is polarized in a given direction and the absorbing probe is polarized in some other direction making an angle ϕ. The probability of absorption is proportional to the $\cos^2\phi$, so that maximum absorption occurs in those probes parallel to the exciting light. The fluorescence emission is detected in two directions, parallel and perpendicular to the excitation source. These intensities can be used to form the quantity called emission anisotropy R,

$$R = \frac{I_\parallel - I_\perp}{I_\parallel + 2I_\perp}$$

We introduce the notion of an absorption and emission vector, each an independent directed quantity in space. In a rigid system, it can be surmized that the relative motion between the absorption and the emission vector is very small. However, in nonrigid systems there is motion between the absorption and emission vectors within the lifetime of the fluorescence emission. Thus, material probe motion occurring within the fluorescence lifetime is detected as an anisotropic increase in fluorescence.

Resonance energy transfer is used to assess the so-called nearest-neighbor distance. Energy transfer occurs between two resonating probes, that is, a donor transfers energy to an acceptor probe by nonradiative energy transfer. The prime condition is that there be a reasonable overlap between the donor emission spectrum and the acceptor absorption spectra. Obviously, quantum yields and donor–acceptor orientations must be satisfactory. This technique is used to determine the separation of the donor and acceptor probes. A critical parameter is R_0, which is defined as the distance for 50% energy transfer. Table 9 shows various acceptor and donor combinations and their respective values of R_0 that have been reported in the literature (4–7).

Table 9. Energy-Transfer Combinations[a]

Donor	Acceptor	Separation[b]
Fluorescein isothyocyanate	Rhodamine isothiocyanate	5.6
Quinacrine	Ethidium bromide	2.2
Quinacrine	7-Aminoactinomycin	3.0
33258[c]	Ethidium bromide	3.1
33258[c]	Daunomycin	4.0
33258[c]	Chromomycin A3	8.3
Chromomycin A3	Ethidium bromide	2.0

[a]For more information, see Refs. 4–7.
[b]Value of R_0 for 50% energy transfer in nanometers.
[c]Hoechst bis(benzimidizole) No. 33258.

Electrical Resistance

The Coulter effect is the name given to the phenomenon (Fig. 7) whereby current through a small aperture in an aqueous conducting medium is modulated by the passage of particles through it. This was first used to develop cell counters; later, it was observed that the change in resistance was related to cell volume. The large amount of data produced by the many blood cells is easily reduced if one uses a histogram of cell size, similar to those discussed previously. The substantially increased number of cells improved the counting statistics. This along with the improved speed measurements made such blood cell counting and sizing instrumentation the methodologies of choice.

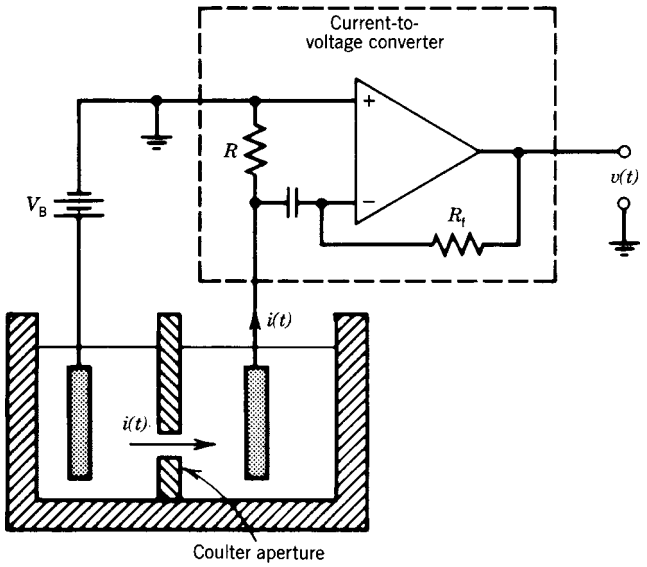


Figure 7. Coulter principle. A conducting physiologic fluid supports the current $i(t)$, resulting from the potential difference V_B in the immersed conducting plates. A particle passing through the Coulter aperture interrupts $i(t)$ by an amount depending on the relative size of the particle the feedback resistor R_f and the input resistor R convert the current $i(t)$ to the output voltage waveform $v(t)$. Processing of $i(t)$ provides data for counting and sizing particles.

Other Measurements

Various physical measurements have been (or should be) adapted for use in automated cytology. These include mechanical techniques such as acoustics, electromagnetic techniques (e.g., nuclear magnetic resonance, NMR), and other spectroscopic techniques. Acoustic microscopes have been constructed that are used for identifying cellular microstructure. On the other hand, NMR has been used to identify metabolic constituents in cells. Raman spectroscopy has been suggested to be useful for identifying chemical constituents in intact cells if the signal-to-noise ratio problems can be overcome. Finally, optical rotatory dispersion and circular dichroism have been used to study nucleic acids. Table 10 summarizes the measurement systems in common use and the cell features they measure.

Devices used in Automated Cytology for Cell Measurement and Isolation

Automated cytology makes use of instrumentation for cell analysis and for cell isolation. Cell analysis instrumentation is classified further into three types according to the information content obtained from the cell: (1) zero-resolution instrumentation, which generates a single datum for each parameter from a cell; (2) high resolution instrumentation, which generates a very large number of data from a single parameter measured from a cell; and (3) low resolution instrumentation, which obtains more information than zero-resolution devices from each cell as a result of a slight increase in resolution.

Instrumentation for isolation is of two general types. The first type is physical placement of the cell in a desired location in space. The second type identifies and localizes the cell with respect to some origin on a fixed medium such as a microscopic slide; cell identity and position data are stored in a computer.

Table 10. Correlation of Selected Cell Measurements and Features

Measurement	Feature
Resistance (Coulter) orifice	Cell volume
Light scattering	
Low angle (2–20°)	Size and shape
Large angle (to 90°)	Size, shape, internal structure, and viability
Polarized	Macromolecular conformation
Acoustic energy	Cell compressibility and deformation
Pulse shape analysis	
Slit scan	Particle shape information; distribution of stain within the cell
Time of flight	Double-cell detection; particle shape information; particle diameter; resolution of cell structure
Time	Kinetics
Fluorescence intensity	Amount of fluoregen
Fluorescence depolarization	Macromolecular viscosity
Energy transfer	Nearest-neighbor detection (spatial separation); molecular mobility

Flow Cytometry and Sorting. Flow cytometry and cell sorting have been developed for the rapid identification and isolation of cells. By causing cells in a suspending medium to flow past detection devices, identification is accomplished. With appropriate electronics and data processing devices, cell analysis is effected. Based on the analysis, cells are isolated by physically moving the suspension medium.

Cell analysis includes detection of the parameter desired, appropriately converting the detected energy to some electrical signal, and processing that signal. A cell riding in the stream passes by the laser beam. The cell causes light to scatter and depending on its preparation may fluoresce. If the Coulter effect is implemented, the cell flows through the Coulter aperture and the modulating current flow is detected. Each cell causes a signal to appear at the output of the various detectors. In a sense, the cell has been identified according to the value of the detected parameter. This data is in essence real-time data; it can be used with further treatment for sorting (see below) or stored in the appropriate form for data reduction. In practice, all detection schemes have involved optical parameters or cellular resistance to measure cell size, cellular fluorescence, and other properties.

Cell sorting begins with identifying a detected cell as the one to sort. Essentially the suspending medium or stream is an electronically conducting fluid jet. Fortunately, the ions that imbue electronic conduction properties in the medium also provide the fluid with physiologically useful properties. The stream breaks up into droplets that contain the desired cells. The droplets are charged, and in turn are acted on by an electrostatic field that deflects the charged droplets containing the cells to appropriate containers. Sorting the droplets containing the desired cells and no other plays a major role in determining the purity of the sorted fraction. This is guaranteed by using acoustic energy vibrating the nozzle to create instability on the stream so that the stream breaks into droplets at a predictable point.

We demonstrate these principles with the aid of Fig. 8. The conducting physiologic fluid (sheath) flows into the nozzle and is ejected as a jet through a circular orifice of $\sim$50 μm. A sample consisting of cells in suspension flows into the nozzle, is injected into the sheath, and is also ejected with the jet. The nozzle is designed to establish laminar flow conditions; this enables the sample to be accurately centered on the stream. If the radius of the sample in the sheath is on the order of the cell size, the cell is highly localized in the center of the stream. The acoustic drive assures the predictability of the location of droplet separation from the stream.

Table 11 lists lasers and their emission lines in common use. Argon ion lasers are capable of delivering several watts of laser energy in the blue-green region of the visible spectrum. Since they require high power input, and a very high discharge current, this laser is very bulky and emits a large amount of heat. Similarly, krypton ion lasers deliver several watts of laser energy in the visible spectrum, and

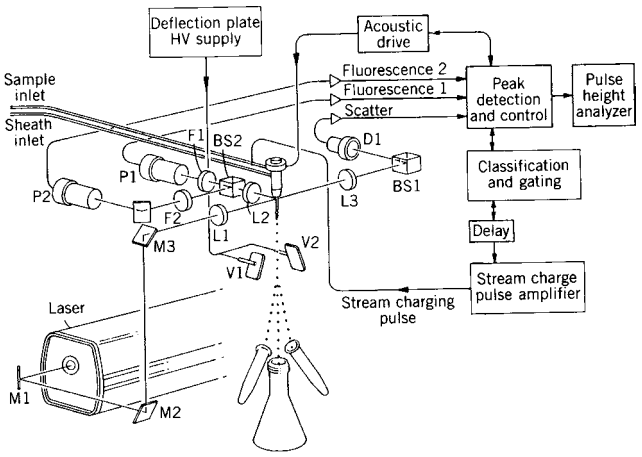


Figure 8. A simplified block diagram of a cell sorter. Mirrors directing the laser beam, M1, M2, and M3; lens to focus the laser beam, L1; lenses to image the laser illuminated cell on the detectors L2 and L3; beam splitters to deflect the light, BS1 and BS2; filters to separate the laser excitation from the emitted light of the particle; F1 and F2; photomultipliers to detect fluorescence, P1 and P2; and diode to detect forward scatter, D1. Courtesy FACS Division, Becton, Dickinson Electronics Laboratories, Mountain View, California.

require high power input and a very high discharge current. Mixtures of Argon and Krypton are sometimes referred to as "white light" lasers because of the coverage over the visible spectrum the combination of the two gases produces.

For a zero-resolution device implemented as shown in Fig. 8, each signal produces a pulse 10 μs wide, whose height is proportional to the total light energy detected. Thus, for each cell, the peak of three pulses is obtained: light scatter and two fluorescence wavelengths. Typically, 50,000 cells can be analyzed in 1 min. With 150,000 data points, powerful data reduction capability is required. For sorting, the cell requires ~250 μs to arrive at the point in the stream just prior to breaking up into droplets. This distance is programmed into the electronics causing the delay required for the identified cell to reach the end of the stream. When the cell reaches that point, the entire stream is charged by the conduction of the ions in the buffer (investigations indicate that <0.01 M salt solution can conduct satisfactorily). The drop containing the cell is separated from the stream and carries the charge imparted to it. The charged drop traverses between electrostatic plates; and due to the interaction of the electric field on the charge, the drop is deflected.

An interesting variation is to replace the electronics with a computer. The data placed into the computers will consist of scattered light intensity, two fluorescence intensities, and relative time of detection. The computer is used to analyze the data and within the 250 μs limit, outputs the sort word. If the sort word is variable so is the charge on the stream. Consequently, one can sort the droplets containing the cells at a defined location on a microscope slide. If the cell's data and position are stored in the computer, that cell can be retrieved by a computerized microscope (described later). The ability to correlate flow cytometric data with visual data is important in identifying the properties of rare cells, such as cancer cells. Analytical studies of the sort trajectory under conditions of drag show that a "knee" is produced when the horizontal velocity is zero (Fig. 9).

The cytometer just described illuminates and detects cells orthogonally to their traversing path. A different implementation is to place the illumination and detection optical path along the same axis as the path of cell traversal. In this way, the cell passes through the source and detector focal plane. After the cell has passed through the focal point, it is directed away from the optical axis.

Table 11. Lines of Laser Emission[a]

Helium–Cadmium	Argon Ion	Krypton Ion	Violet Diode	Krypton–Argon	Helium–Neon
325		350.7			
	351.1	356.4			
	364.8		405		
450					
	457.9				
	476.4				
	488.0			488	
	496.5				
	514.5				
		530.9			
		568.2		568.2	
					632.8
		648.1		648.1	
		752.5			

[a]Wavelengths of light in nanometers.

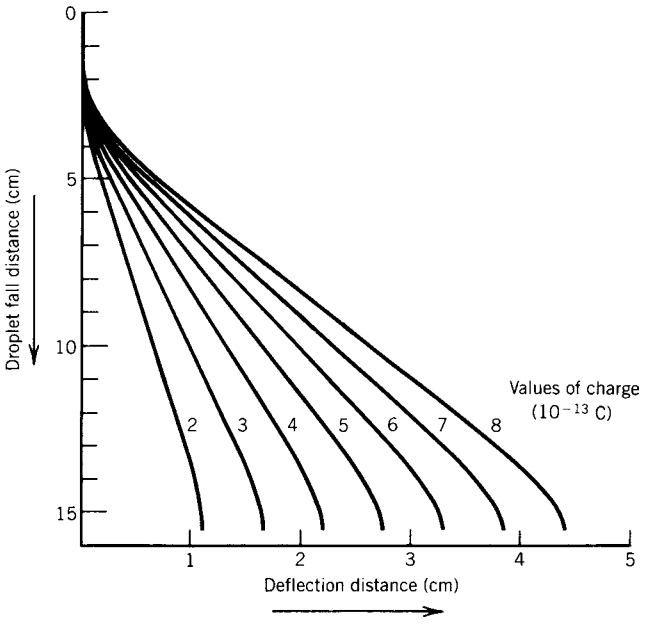


Figure 9. Trajectories of charged droplets for seven values of charge. The trajectories were obtained by solving the equations of motion, assuming that the drag on the droplet was a linear function of velocity.

Since the cell passes through the focal plane of the optical system, errors due to focusing and misalignment are reduced, particularly if the same objective is used for excitation and detection.

Most fluorescence spectra has uninteresting shape and the single value obtained by zero resolution systems is adequate. This limits the usefulness of spectral analysis of the fluorescent moiety. Nevertheless some effort has been carried out in this direction. The easiest is the use of a variable color filter and to select the resulting wave for shape (see slit scan systems below). Also some have used white light pulses and processing to identify the spectra. By far the solution has been to use multiple lasers for excitation and to detect the different emission on separate detectors. There are technical limitations as well, the need for UV light means expensive powerful lasers or substantial Xenon light sources, furthermore the speed of processing has now gone up an order of magnitude from the time between cells to the time that the cell is in the laser beam.

An interesting application is the use of fluorescence lifetimes or the so-called phase detection systems. One selects probes with overlapping spectra and different singlet excited states. Using a modulated laser beam and measuring the phase shift in the fluorescence signals gives a measure of differential lifetime.

Low Resolution Systems. In the flow system just described, the cell flowed past a laser beam and the detected energy resulted in a pulse that was peak detected. Clearly, any other parameters could be obtained from the pulse, including pulse area and pulse width, and each descriptor is a single number. Cells are analyzed and sorted based on a single value of a descriptor; thus, each parameter of a zero-resolution system takes a single value from each cell.

There is potential for greater information available from the pulse. For example, if the laser is flattened into a ribbon whose thickness is small compared to the length of the traversing cell, the resulting pulse will contain information concerning the shape of the cell. Proper analysis of the pulse can provide increased information about the cell because of the slight increase in resolution. This technology is referred to as "slit-scan". The very narrow laser beam is 4 μm in width and the flow system is slower than the normal 10 m/s.

Figure 10 shows the schematic diagram of a cell traversing through the laser beam. The figure on the left shows a zero-resolution wave form in which only the peak-detected value is used to represent the data from the cell. The low resolution system (right) shows some structure as the cell traverses through the laser beam: First the cytoplasm (C) is detected, then the nucleus (N), and then again the cytoplasm.

The development of automated cytology was largely motivated by the requirement to diagnose cancer by finding cancer cells in a sample. Cancer cells can be distinguished from normal cells based on their DNA content (as distinguished by, e.g., AO) and other parameters such as size and RNA. Patient samples for the purpose of detecting early cancers contain highly variable and usually very small numbers of diagnostic cells. Consequently, the problem is not in distinguishing cancer cells from normal cells, which is easily accomplished, but in distinguishing particles whose values fall in the space consistent with cancer cells, but are in fact not cancer cells. This requires additional information provided by low resolution systems, but not by zero-resolution systems.

High Resolution Cytometry. We now consider cell image analyzers usually implemented as a computer-controlled microscope. This high resolution instrumentation is characterized by the ability to obtain large amounts of data on a few cells and to perform complex analyses on those cells.

A microscope is connected to a computer so that data reduction and control of the object in the microscope can be performed. These data are used for automating cell recognition.

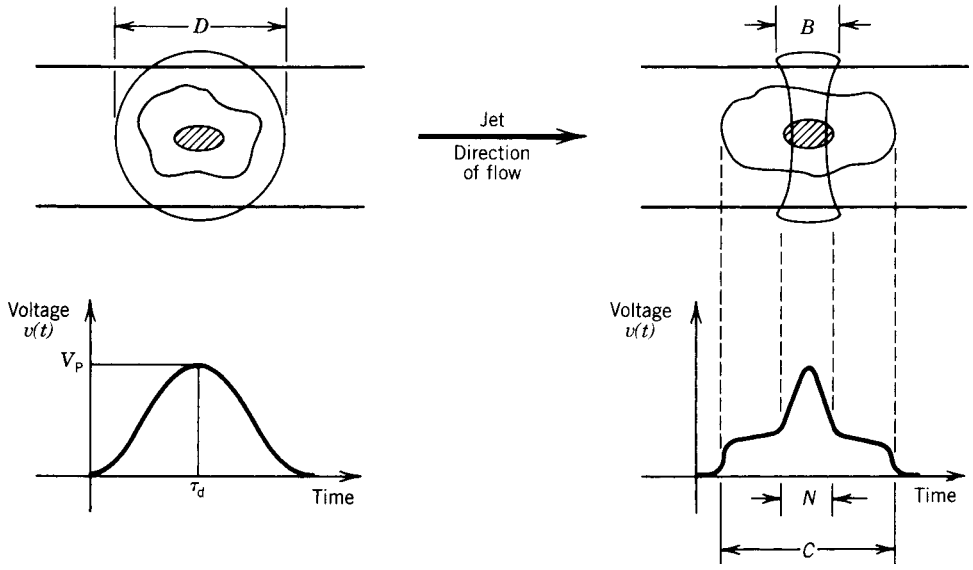


Figure 10. Axial view of a circular laser beam of diameter D and cylindrical laser beam of width B, both intersected by a cell. The circular laser beam produces a pulse signal from which a single parameter of information, such as peak height, can be obtained. For example, at time T_d, $v(t)$ has a peak value V_p corresponding to the cell in the center of the laser beam. The cylindrical-shaped laser beam is a ribbon of light inter- sected by a portion of the cell. The output signal, which is the convolution of cell structure with laser light, allows cell structure information to be retrieved easily. For example, the cytoplasm in the laser beam for time C is easily distinguished from the nucleus in the laser beam for time N.

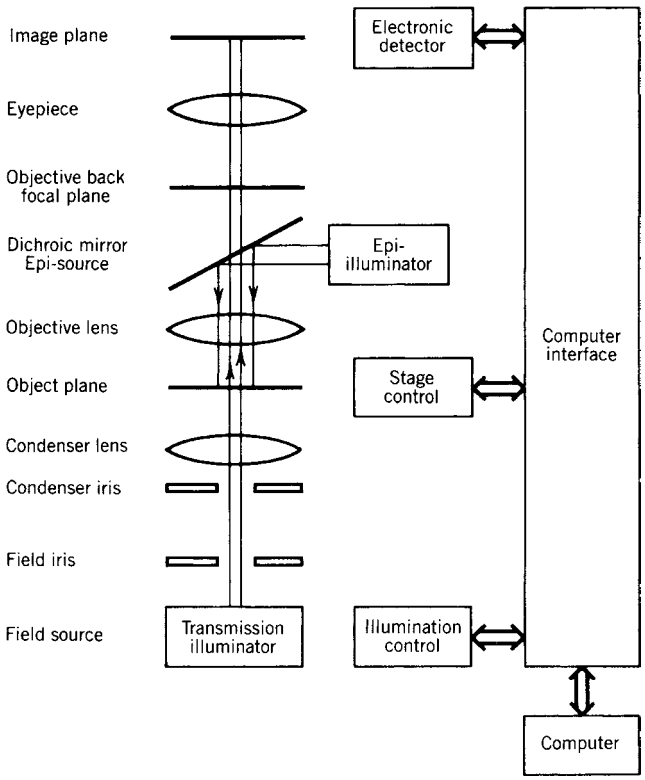


Figure 11. Schematic diagram of a microscope. Light from the transmission source traverses through to the image plane. Epi-illumination is reflected to the object that in turn emits fluorescence, which passes through the dichroic mirror onto the image plane. The computer provides scanning and illumination control and reads the electronically detected data.

The essential features of a computerized microscope are shown in Fig. 11. This is a schematic diagram of a modern-day microscope supporting transmission and reflection (fluorescence) microscopy. For transmission microscopy, the light path from the source is treated so that parallel light is transmitted through the object plane. The objective and eyepiece lenses function together to provide the visually detected image at the image plane. Fluorescence excitation (epi-illumination) is by short-wavelength light (violet), which is reflected by the dichroic mirror, then passes through the objective lens onto the object. The object's fluorescence emission passes through the dichroic mirror, which allows the long-wavelength light to pass; it is then detected in the image plane.

The computer interface provides data exchange between the computer and the controllers of stage motion, illumination, and detection. A photodetector at the image plane measures the intensity of light emanating from the object and a common implementation passes light through a pinhole on to the optical access. In this case, the moving stage controls the object so that its spatial variation of light intensity is acquired. The entire object is imaged one point at a time. Each point is a picture element (pixel) emitting light. Light from each object point produces a datum point, which is stored in a matrix such that the value of the ijth element is the intensity of the light at the ijth pixel.

We abstract the microscope to consist of a light source, an object plane, and detector. Any one moving with respect to the other two can produce the scanning required to image the object. Table 12 shows the characteristics of the scanning components of a microscope.

These systems have high degrees of flexibility since performance is controlled by means of software. In such a system, a cell is located for analysis, appropriately illuminated, and contiguous absorbance values of the image are detected. The image is made up of pixels (say 1×1 μm squares) and the pixel's intensities are digitized and stored. Software controls data acquisition, performs mathematical operations, and formats the data to a convenient form. For example, the absorbance values of a cell are converted to digital values, gray scale histograms are formed from which cluster analysis in a multiparameter space is performed. Consequently, statistical decision rules can be invoked to assess characteristics of a cell within some statistical confidence interval.

DATA ACQUISITION, PROCESSING, AND MODELING

In the previous section, we described cell measurement systems. Now, we concern ourselves with the acquisition and manipulation of the data produced by these systems. We will begin with data lists produced by zero-resolution devices and proceed to high resolution images produced by cell-scanning devices.

Zero-Resolution Systems

In zero-resolution systems, a cell produces a single value for each parameter measured. If only light scattering is measured, each cell produces only single values of scattering; if additionally two values of fluorescence are

Table 12. Scanning Components on Microscope

Component Performing Scan	Source	Characteristics of Components	
		Object	Detector
Source	Point scanner (laser illumination)	Stationary	Stationary
Object	Stationary (uniformly illuminated field)	Scanning stage	Stationary (pinhole)
Detector	Stationary (uniformly illuminated field)	Stationary	Point detector (television camera)

measured, each cell produces three data points. The acquired data are listed one point after the other as

Cell Number	Parameter 1	Parameter 2	Parameter 3
Cell 1	255	10	128
Cell 2	128	210	197
•	•	•	•
•	•	•	•
•	•	•	•
Cell n	37	196	212

Such list mode data can be stored and reduced into histograms of the data.

Histograms. A common form of data reduction is the production of histograms of the data. An n-parameter histogram is in $n+1$ dimensions, where the additional dimension is frequency of events.

A single-parameter histogram substantially reduces the amount of data compared to the list mode. For example, a 256-bin histogram (2^8) requires 256 words of storage compared to the 25,000 words required for storing the 50,000 bytes of data obtained from a single run. Examples of single-parameter histograms are shown in Fig. 3.

Dual-Parameter Histograms. Data from two parameters, P_1 and P_2, can be displayed so that each pair (p_1, p_2) contains the number of cells that express that pair of values. For data storage, the multiplicative effect of the number of dimensions reduces the histogram's value. A two-parameter histogram of 256 channels requires 256^2 (64 k) locations, which is approximately the number of data points acquired. Furthermore, the data will not cluster well when insufficient numbers of cells have been acquired. A common way around this problem is to reduce the resolution from 8 bits to 6 bits (256 to 64). Now, only 4096 channels are required for that histogram display. The increased ability to visually cluster the data is achieved with a loss in additional information.

Three-Dimensional Histograms. Although this is a four-dimensional system, a common method of displaying three-parameter data is to plot each triple for a given frequency. For example, Gaussian-distributed data in three parameters (with equal variances) is displayed as a spherical cloud in the three-parameter space at each frequency. Since the storage space for such a histogram is 256^3 (16 million), clearly storage of the list data is more efficient than storing the data in histogram form.

A useful method of analyzing data displays one or more parameters as a function of a specified set of values of another parameter. For example, it is convenient to analyze two fluorescence parameters resulting from the largest cells. The set of values representing the largest cells is called a window. For each value in the window, there corresponds one or more parameters that can be displayed. The values within the window act as a gate to display the other parameters.

Finally, a physical interpretation of the histogram occurs when we consider its mean value. Assume we acquire data from a multiplicity of cells. Further, we wish to measure the presence of some molecule by causing a radioactive tag to bind to it. The resulting measurement of radioactivity is the sum of the contributions of the individual cells. If this is normalized to the number of cells, the result is a mean value. Now the histogram, of course, displays the number of cells that contain each value of data for all the values of data. So it provides substantially more information than the single mean-value datum. Furthermore, the mean of the histogram corresponds to the mean-value result. For example, the distribution of DNA in a cycling population is very uneven, a fact that would not be elucidated from mean values over the population.

Histogram Analysis. Parametric and Nonparametric Analysis of Histograms. Parametric analysis assumes a model or distribution is used to compare, or analyze, the data. Nonparametric analysis assumes no such model.

Nonparametric Analysis. It is useful to compare histograms so that the effect of different treatments can be statistically assessed; two techniques are generally used. The first one requires three or more identical histograms and sums up the individual channels to obtain a mean and variance of each channel. Based on this model a channel-by-channel t test is made of the sample histograms to decide if statistical significance is valid.

A second technique uses the Kolmogorov–Smirnoff technique (see Fig. 12). In this case the histogram [Fig. 12a] is normalized, then summed to a cumulative distribution [Fig. 12b]. The two cumulative distributions are compared. The comparison is in the form of the absolute value of the difference between the two histograms. If this value exceeds a certain critical value that determines the level of confidence, the two histograms are statistically significantly different.

Parametric Analysis. The majority of work in this area has been to develop models that will aid in finding the number of cells in the various compartments of the cell cycle from DNA histograms. It is assumed that the "true" DNA histogram consists of two impulses with DNA values at $2n$ (for G0/G1 cells), the other at $4n$ (for G2/M), and cells in S are in between. As a result of the imprecision of measurements, the impulses at $2n$ and $4n$ values of DNA have been broadened. More generally, we write the expression used to model single-parameter DNA distributions.

$$F(k) = F_1(2\pi\sigma_1)^{-1/2}\ \exp[-(k-\mu_1)^2/2\sigma_1^2] + F_2(2\pi\sigma_2)^{-1/2}\exp[-(k-\mu_2)^2/2\sigma_2^2] + s(k)$$

In parametric analyses of the DNA histograms, the broadened pulses are assumed to be Gaussian with means μ_1, μ_2 and variances σ_1, σ_2, respectively. The critical issue is the shape of the s-phase distribution, $s(k)$. In the earliest modeling systems a second-degree polynomial

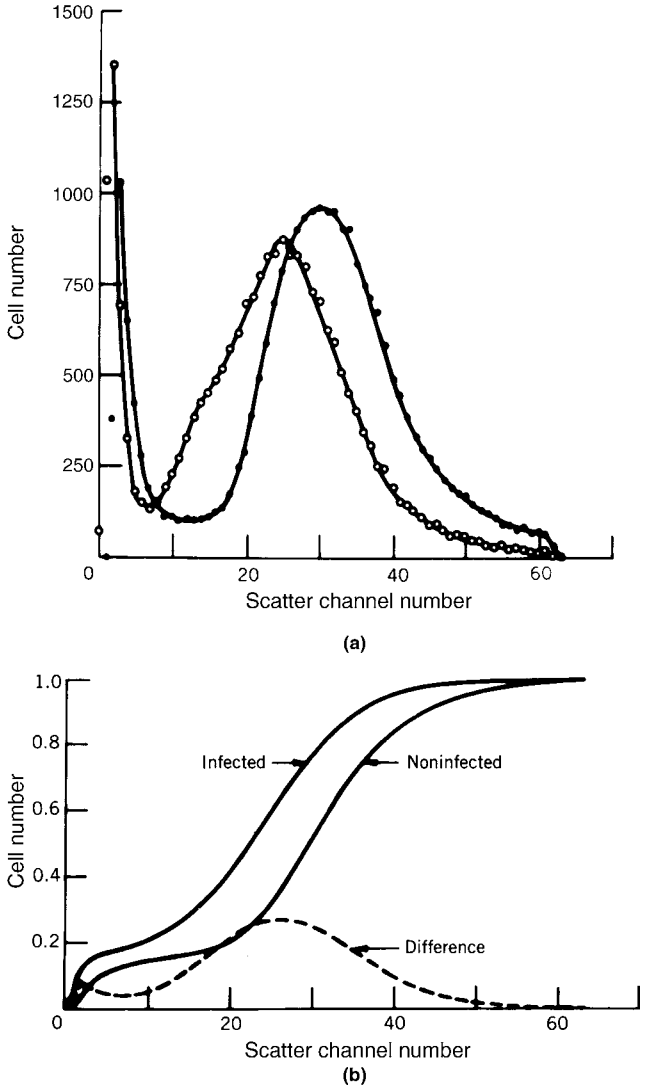


Figure 12. (a) Histograms of the size distribution of cells either infected (○) or not infected (●) with herpes virus. (b) The cumulative distribution of each of the above histograms.

was assumed to be the fit. Other models have assumed Gaussian curves of various values to try and fit the data to the *s* values. The distinction between the cells that are in G1 and S is not sharp and these populations overlap each other. Consequently, methods that distinguish between G1 and S without models (e.g., graphical) usually underestimate G1 compared to S by 20%. The number of cells in S is small compared to the number of cells in G1; furthermore, the width of the G1 peak is also very small; consequently a small error in defining the G1 compartment can result in substantial errors in estimating S and G2.

Unfortunately, these models are good only for well-behaved populations whose value of DNA does not exceed $4n$, such as tissue culture cells. The DNA distribution of cells from cancer patients are complex due to cell clones arising with increased amounts of DNA, increased number of chromosomes (aneuploidy), and the presence of multinucleated cells.

Finally, these models seek only to fit curves to the values of DNA. They are not intended to elucidate the production of such distributions from fundamental principles, although some authors have tried such derivations, and achieved limited success.

Image Analysis

Cell data acquired by high-resolution systems consist of a list of photometric values of light intensity and the corresponding source point address. Scanning is sequential so the source addresses (or locations of the source points) need not be explicit. Each source point (pixel) on the cell is really an area of the cell that produces a uniform intensity. The pixel size determines the resolution of the image.

The light-intensity list can be manipulated, resulting in classification of the object by techniques referred to as pattern recognition. This is distinct from other operations on the image such as restoration. Image restoration uses the mathematical properties of the image generation system to obtain an inversion that will remove the errors introduced during image generation. Pattern recognition uses image features to distinguish between objects and to express that distinction in statistical terms. The general sequence of events for pattern recognition is shown in Fig. 13.

Pattern Recognition. The preprocessor operates on the digitized image, which is the list of pixel values. In preparation for feature extraction, sections of the image are defined (automatically or by human interaction) using edge detection filters to distinguish boundaries.

A second operation is the formation of a gray scale (and other) histograms. The gray scale histogram lists the number of pixels for each gray value (intensity value) for the entire dynamic range of gray values in the system. Image characteristics cluster about some value of gray in the histogram. For example, a cell typically displays a cluster value representing the nucleus and second cluster about the lesser dense values of the cytoplasm. Thus, a separation of the nucleus and cytoplasm can be effected depending on the degree of separation between the two clusters. The cell shown in Fig. 14 has had all its cytoplasmic values truncated to a particular gray value; this gives prominence to the highly variable but more dense values of the nucleus.

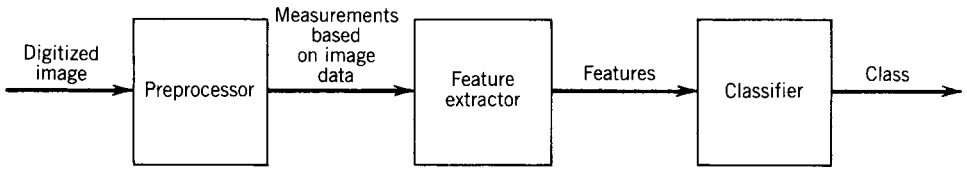


Figure 13. Sequence of events in pattern recognition.

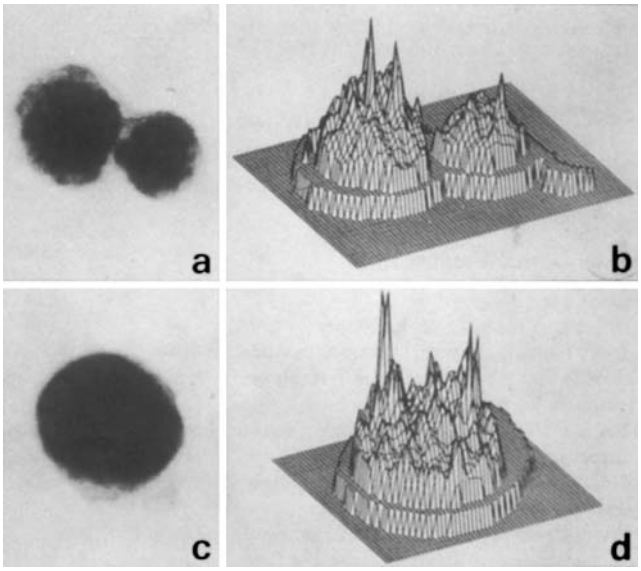


Figure 14. Display of a digitized cell showing the truncated values of the cytoplasm.

A second operation is equalization of the gray scale histogram. This is a nonlinear transformation resulting in a histogram with reduced number of gray levels at approximately equal frequency. Since the frequency of gray level occurrences contains no information about the texture of the image, this permits images to be placed into a consistent format prior to comparisons.

Feature Extraction. After preprocessing the image data to yield measures based on image data, it is possible to obtain cell characteristics, called features, that are based on geometric parameters, texture parameters, optical density, or gray level parameters.

For example, statistical texture analysis, such as is typically used in medical imaging, uses conditional probability matrices to obtain features. Picture analysis proceeds by forming a transition probability matrix from the image data. From this matrix the texture features used for classifying objects in the image are determined. Each element in the matrix, $P_L(I/J)$, is defined as the conditional probability of gray level occurring L picture points after gray level J occurs. Each of L matrices is square of dimension n, where n is also the number of gray values. Operations on these matrices can be performed to produce parameters, such as the second moment of the matrix or the moment of inertia about the diagonal of the matrix, for all values of I and J. Clearly, a very large number of parameters can be chosen in this manner. Texture analysis has been used successfully to classify cytological samples.

Classifier. The last step in the pattern recognition chain is classifying the cell based on the measured features. Each feature is a dimension in feature space. A range of values of each feature are obtained from a training set, that is, a set of objects whose classifications are known. This feature space is sectioned by boundaries defined by the class or members of the training set. These boundaries separate the objects into the different classification regions. The features of an unclassified or unknown object can then be compared to the classification obtained from the training set. The probability of the unknown object belonging to a particular class as then determined. An interesting problem is that the vectors produced by the features may not be orthogonal. If such vectors are correlated, the feature provides little new information in addition to its correlative feature. Consequently it is important to have features with different information content.

CURRENT USAGE

DNA Measurements

In previous examples, we discussed the measurement of DNA in intact cells, which are now providing new and useful information in both biology and medicine. We briefly mention three applications of interest: bacterial analysis, chromosome analysis, and sperm analysis.

Flow systems can be used to classify bacteria types by the difference in dye uptake by the DNA base pairs. Fluorescent CG- or AT-specific dyes are used to generate a two-dimensional histogram that can be used to differentiate bacterial species. This may be useful in a biology laboratory for rapid identification of bacterial types. In medicine, such a detection system would be useful in urine samples. However, with blood its usefulness is limited. First, the level of infection in blood is produced by very few bacterial cells. Their small size and the presence of overwhelming numbers of red blood cells makes detection extremely difficult. A second major difficulty is that the information of importance in medical practice is sensitivity: which and how much antibiotic kills the bacterium, not which bacterial cell is causing the infection.

Chromosome Analysis. As yet neither zero-resolution systems nor low resolution systems provide adequate separation between the human chromosomes to completely perform a human karyotype. It is interesting, though, that errors in karyotyping due to chromosomal particles produced by disease are constant and repeatable and thus can be expected to show exceptions to the normal.

Flow systems have been used to analyze sperm cells, primarily to measure the DNA of the cell types. Spermatids, which are spherical precursors to the mature sperm, are easily analyzed by flow cytometry and produce a separation between the X and Y chromosome. This separation implies that a distinction can be made between spermatids whose sperm will give rise to females (X chromosome) or males (Y chromosomes). These data were obtained with an axial flow analysis device and with a very small coefficient of variation (0.9%). Unfortunately, the mature sperm cells are not well-behaved spheres, but flattened bags of DNA with a comparatively long tail. Thus, an orientation artifact blurs the 2% difference in DNA resulting from the X–Y chromosome mismatch. However, flow analysis can be used for fertility determinations and where the number of sperm is an important parameter.

Hematology

An important area is hematology, the discipline concerned with the study of blood and its components.

Differential blood cell counters place white blood cells into five basic classes: lymphocyte, eosinophil, basophil, monocyte, and neutrophil. Since red blood cells outnumber the white cells by a thousandfold, they are usually excluded by preparation. These six-cell types have been classified by several commercial devices that use pattern recognition and high resolution image analysis for classification. Such classifications have occurred with varying levels of reliability. A major problem is a category required for "others", that is, for cells that cannot be classified into any of the six classes. This category reflects the fact that at any one time, blood cells are maturing into the different classes. Furthermore, the body's reaction to some diseases or insults is to produce increased numbers of cells from the bone marrow into the blood stream at various levels of maturation of the cells. Consequently, some workers have attempted to classify nucleated blood cells into 17 different types. Efforts at white blood cell classification using flow systems have not had the commercial success that scanning systems have had.

Immunology

An important application of flow systems has been to enumerate and classify the cells in the body that are part of the immune system. Briefly, the body produces antibodies to antigens. The antigens are usually foreign substances such as a bacterial cell surface; but in abnormal situations the body produces antibodies to its own antigens. The covalent binding of antibody to antigen initiates a set of reactions that results in the destruction of the foreign substance.

A specific kind of antibody, called a monoclonal antibody, reacts only with a single antigenic determinant. This is in contrast to the multiplicity of antigenic determinants that produce heteroclonal antibodies, which are usually produced by the body. Fluorescent stains such as FITC and PE are covalently bound to the monoclonal antibody of choice. Thus, simultaneous green and yellow fluorescence to distinguish the different immunological cell types can be produced (see Table 13).

One application of this technology is to differentiate the various classes of lymphocytes. Lymphocytes are divided into two general classes: B cells and T cells. In general, the B cells produce antibodies, whereas the T cells are direct protagonists in the immune response. By the use of monoclonal antibodies, T cells have been divided into many subsets. Some of the most important are T4 (helper cells) and T8 (suppressor cells). These cells are distinguished by T4 and T8 monoclonal antibodies. Table 14 shows the ligand-binding applications, including lymphocyte cell surface characteristics.

Table 13. Excitation and Emission of Selected Fluorescence Labels

	Excitation, nm	Emission, nm
Fluorescein isothyocyanate	488	530
(S)-Phycoerythrin	490	570
(R)-Phycoerythrin	498	575
Bodipy	503	511
Tetramethyl rhodamine	560	580
L-Rhodamine	572	590
B-Phycoerythrin	540	575
Texas Red	590	620
CY-5	649	666

Table 14. Ligand Binding

Immunologic
Cell surface receptor on lymphocytes
Cells identified by surface immunoglobulin
T cells identified by various T receptors including T4 for helper cell and T8 for killer cells
DNA synthesis: Antibody to bromodeoxyuridine
Indirect immunologic binding
Goat-anti-rabbit antibody with fluorescent tag binds to rabbit–anti-receptor molecule
Avidin tagged with fluorescent molecule binds to biotin attached to the antibody against the detected antigen
Hormonal
Estrogen receptor analog (17-fluorescein estradiol)

In medicine, determining the number of helper and suppressor cells is valuable in transplantation and cancer. The ratios of these cells and the changes in ratio as the patient undergoes treatment is an indicator of the patient's ability to respond to the treatment. Indeed the absence of helper cells due to viral destruction is the primary problem in acquired immune deficiency syndrome (AIDS).

Oncology

A major application in both flow and cell scanning systems has been in oncology. The primary thrust has been to classify cell types and to distinguish the malignant cells from nonmalignant cells. For example, nuclear texture has been used to classify uterine cervical cells into cancer cells and noncancer cells.

Studies involving the four major types of lung cancers have been performed to determine the ability of flow systems to first find and then sort the malignant cells. The criterion for malignancy is based on the morphological features used by pathologists to diagnose cancer. Cytological examination of lung cells begins with the sample (sputum). The cells in the sample are fixed to minimize disease contagion and for preservation. After fixation, the cells are stained with acridine orange, analyzed, and sorted for the highest value of green and red fluorescence. The Acridine Orange analysis of cancer cells provides information about the relative amount of RNA and DNA in cells. Cancer cells are found in regions showing the highest green fluorescence. By sorting, a substantial enrichment (65-fold) for cancer cells can be obtained. Concomitantly, there is a sharp reduction in normal cell types, such as lymphocytes and squamous cells.

A similar technique using acridine orange staining of cells has been applied to urinary cytology. This has resulted in a system that is demonstrably superior to human cytological analysis for earlier detection of recurrent cases of bladder cancer. Data on cells are placed in list

mode for scatter, green fluorescence, and red fluorescence. The pulse width of scattering is used to find debris and eliminate it. The resulting "clean" data list then produces RNA and DNA histograms to find whether the number of cells in S>20%; if so, a recurrence is said to have occurred.

Many studies have been made that identify an aneuploid (aberrant number of chromosomes) population of cells. Such aneuploid cells are considered by some pathologists to be diagnostic of cancer.

In a clinical setting, a machine approach to cancer detection has been to analyze a very large number of cells without false positive indications. This is not possible with zero-resolution systems; however, low resolution systems have been developed with sufficient information content to identify the false alarms. These have been used in clinical trials and the results have been dramatic. High resolution instrumentation using cell-scanning techniques and texture analysis have been used to identify and classify cells according to cell type and denote the malignant cells. These systems have had success in identifying various normal as well as disease cell types.

The medical and biology literature abounds with examples of applications in flow cytometry and high resolution image analysis. The bibliography lists samples of this literature.

FUTURE PROSPECTS

Historically, the motivation for new instrumentation is the increased information. In automated cytology, instrumentation has increased information by improved optical resolution and the increased number of cells for improved reliability of statistics. Furthermore, the development of quantitative techniques to analyze cells makes possible the basis for a quantitative theory of biological phenomena.

In one sense these are learning tools; they provide basic measurements. In another sense, their increased speed is a basis for economic value since productivity is improved.

A further advantage is the use of computers. The measurement instrumentation can easily act as an input to a computer. The data handling power of computers makes possible analysis of enormous numbers of cells thereby establishing repeatable quantitative relationships.

The increase in speed and information and objectivity of this instrumentation makes it an economic force in the market. The present high expense of illumination equipment (primarily lasers) limits its economic practicality. Nevertheless, the present use of this equipment in diagnosis can be expected to increase. Furthermore, its use in therapy and prediction of disease are just now starting. With expected cost reduction, use of this instrumentation will increase.

BIBLIOGRAPHY

1. Mellors RC. Analytical Cytology. 2nd ed. New York: McGraw-Hill; 1959.
2. Melamed MR, Mullaney PF, Mendelsohn ML. Flow Cytometry and Sorting. New York: John Wiley & Sons; 1979.
3. Melamed MR, Lindmo T, Mendelsohn ML. Flow Cytometry and Sorting. 2nd ed. New York: Wiley-Liss; 1994.
4. Shapiro HM. Practical Flow Cytometry. New York: Alan R. Liss; 1985.

Further Reading

Cytometry, New York: original publisher Alan R. Liss. This is the journal of the International Society for Analytic Cytology. First issued in July 1980, it is a forum for automated cytology, with a thrust to basic measurements in the biology of single cells and particles. It is now published monthly by Wiley-Liss, Inc. Clinical Cytometry Started in 1994, and in 1997 produced the first issue of Current Protocols in Cytometry.

Analytical and Quantitative Cytology and Histology. St. Louis, MO: Science Printers and Publishers, Inc. This journal is sponsored by the International Academy of Cytology and the American Society of Cytology. It focuses primarily on automation and quantitative aspects of cytology and is also a forum for automated cytology.

See also ANALYTICAL METHODS, AUTOMATED; COMPUTERS IN THE BIOMEDICAL LABORATORY; DIFFERENTIAL COUNTS, AUTOMATED.

D

DECAY, RADIOACTIVE. See RADIONUCLIDE PRODUCTION AND RADIOACTIVE DECAY.

DECOMPRESSION SICKNESS, TREATMENT. See HYPERBARIC MEDICINE.

DEFIBRILLATORS

BRADLEY J. ROTH
Oakland University
Rochester, Michigan

INTRODUCTION

Ventricular fibrillation is a lethal malfunction of the heart. Normally the heart beats about once a second, and is controlled by electrical signals that occur in a predictable, periodic way. The heart's electrical activity, called the electrocardiogram (ECG), can be measured on the surface of the body. A normal ECG is shown in Fig. 1a. If the heart is in a state of ventricular fibrillation, the electrical control of the heart becomes disorganized and chaotic. Instead of producing a normal ECG, the fibrillating heart produces an ECG that looks more like random noise, as shown in Fig. 1b. Rather than contracting in unison, different regions of the heart contract independently, resulting in a quivering that is not effective in pumping blood.

Once the ventricles of the heart start to fibrillate, death follows in minutes. The American Heart Association estimates that in the United States 335,000 people die each year of sudden cardiac death, with most of the deaths attributed to ventricular fibrillation (1). The most effective way to prevent these deaths is to apply a strong electric shock to the heart within the first few minutes after the onset fibrillation (Fig. 2) (2,3). Devices that deliver such shocks are called defibrillators, and come in two types: external and internal. A physician, a paramedic, or even an untrained bystander can use an external defibrillator to apply a shock to an unconscious victim of ventricular fibrillation. The more sophisticated of these devices are automated so that the user need do little more than follow some simple instructions; such devices are called Automated External Defibrillators (AEDs). Internal defibrillators are similar to cardiac pacemakers, and are implanted

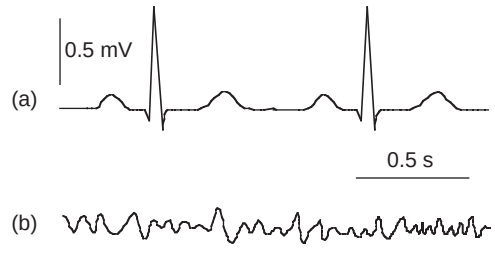


Figure 1. (a) A normal electrocardiogram (ECG). (b) The ECG during fibrillation.

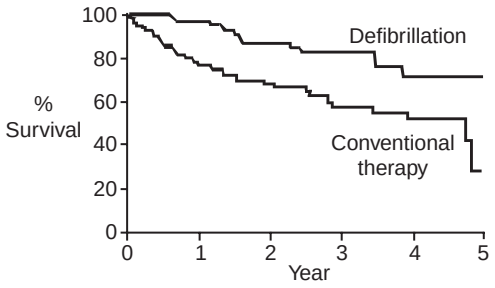


Figure 2. The percent survival for high risk patients when treated with conventional drug therapy or with an implanted defibrillator. [Modified and Reproduced with permission from Moss et al., Improved survival with an implanted defibrillator in patients with coronary disease at high risk for ventricular arrhythmia. *N. Engl. J. Med.*, 335: 1933–1940, 1996. see Ref. 2].

into patients who are at risk for ventricular fibrillation. They monitor the electrical activity of the heart and deliver a shock when necessary. Modern defibrillators can also function as pacemakers, and are called Implantable Cardioverter Defibrillators (ICDs).

EXTERNAL DEFIBRILLATORS

An external defibrillator works by applying a shock through electrodes on the surface of the body. Automated external defibrillators are becoming common in schools, on airplanes, and at other public places. A typical AED is shown in Fig. 3. Each electrode has an area of at least $50\,cm^2$ and is attached to the skin by a self-adhesive pad. A conducting gel should always be placed between the skin and the electrode to reduce the skin resistance. The current passes from the electrodes through the entire torso, with only a fraction of it reaching the heart.

A defibrillator works by charging a capacitor to a high voltage and then discharging it through the patient's body (Fig. 4). When the switch S is to the left, a capacitance of about $200\,\mu F$ is charged to $\sim 1500\,V$, implying a stored charge of $0.3\,C$ and a stored energy of $225\,J$. Move the switch to the right, and the capacitor discharges through the resistance of the patient's body ($50\,\Omega$ or more), generating a peak current of 30 A that decays exponentially with a time constant of 10 ms.

An actual defibrillator circuit is more complicated than shown in Fig. 4. For example, the battery pack used to power an AED typically has a voltage of $\sim 12\,V$. A high voltage power supply is needed to raise this voltage to the level necessary to charge the capacitor. Also, many defibrillators use a biphasic, truncated-exponential waveform, which is more effective for defibrillation than a monophasic wave form (Fig. 5). The biphasic waveform is produced by discharging the capacitor part way, then reversing the polarity of the leads, followed by further discharge. Switching circuitry that functions at high voltages is required.

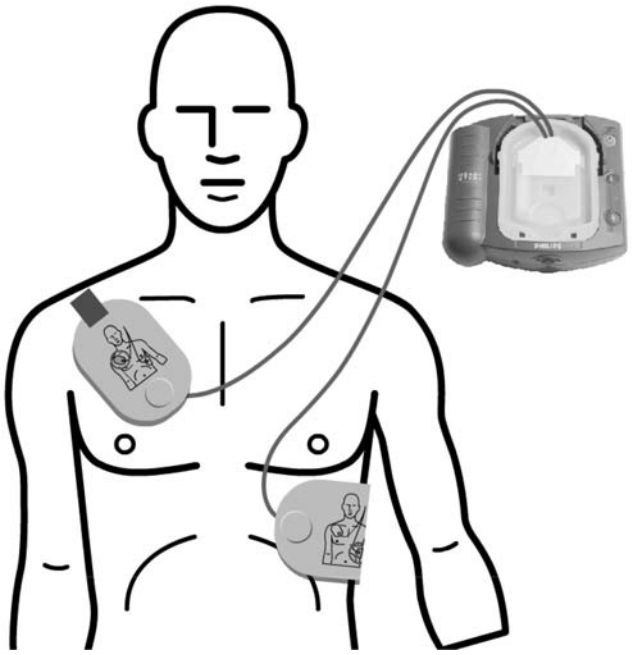


Figure 3. An automated external defibrillator (AED). This figure appears courtesy of Philips Medical Systems.

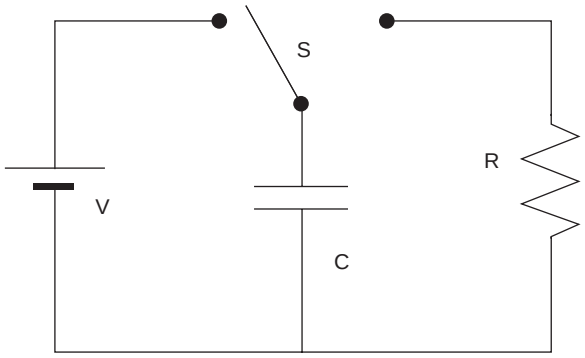


Figure 4. A simplified defibrillator circuit, where V is the voltage of the power supply, C is the capacitor, R is the resistance of the body, and S is a switch.

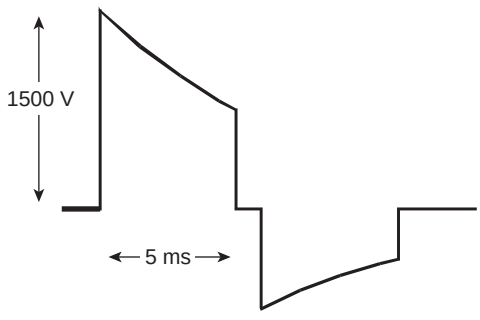


Figure 5. A typical biphasic, truncated exponential waveform used in many defibrillators.

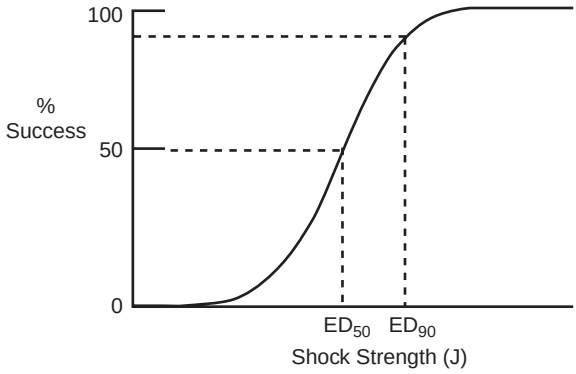


Figure 6. Defibrillation success follows a sigmoidal probability curve. The shock strength corresponding to a 50% success rate is called ED_{50} (for effective dose, 50%).

The word "automatic" in the term Automatic External Defibrillator means that the device can decide for itself if defibrillation is necessary. The AED monitors the electrocardiogram, and enough memory is included in the machine to store the ECG data. Also present are electronics that allow the device to analyze the ECG and decide if the ventricles are fibrillating. If they are, the AED will tell the caregiver to shock the patient. Most AEDs provide both written and oral instructions about how to attach the electrodes and operate the device. In theory, minimal training is required.

The success rate of defibrillation follows a probability curve like that shown in Fig. 6: the higher the shock energy, the higher the probability of defibrillation. A shock strength corresponding to a 50% success rate is known as "ED_{50}". To reduce the probability of a failed shock, physicians often use strengths of about ED_{90}. Unwanted side effects also increase with shock energy, so an AED usually shocks with a relatively low energy first, say 200 J. If that fails, it delivers shocks of increasing energy up to a maximum of ~ 360 J.

External defibrillators used in hospitals and ambulances are similar to AEDs, except that they are not automatic (the physician decides when to shock a patient rather than the device) and they can often be powered by plugging into the hospital's electric power grid instead of, or in addition to, relying on batteries.

IMPLANTABLE DEFIBRILLATORS

An implantable cardioverter defibrillator resembles a pacemaker, but its circuitry is similar to that in an AED. The battery, capacitor, and electronics are enclosed in a metal case (titanium or stainless steel), which is implanted under the skin in the chest (Fig. 7). The typical size of the case, or "can ", is $\sim 50 \times 50 \times 15$ mm. The can often serves as one of the ICD electrodes.

The capacitors in an ICD are only slightly smaller ($\sim 125\,\mu$F) than in an AED, but in an ICD the capacitor is charged to a voltage of only ~ 600 V, implying a charge of 0.075 C and an energy of 23 J. An ICD delivers about one-tenth the energy that an AED does, but in an ICD the shock is delivered through electrodes placed within the heart and is therefore just as effective for defibrillation. Tissue impedance for an ICD is at least 50 Ω, implying a discharge

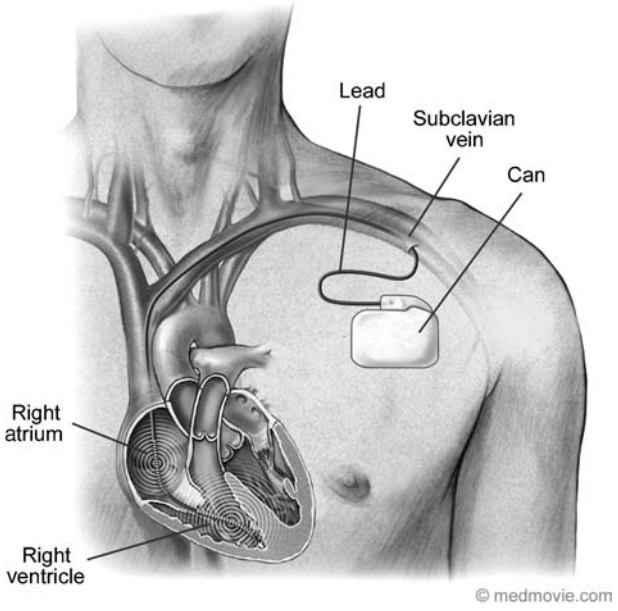


Figure 7. An implantable cardioverter defibrillator (ICD).

time constant of ∼5 ms. Many ICDs contain two 250 μF capacitors charged in parallel, to give a total capacitance of 500 μF. When discharged, the connection of the capacitors is changed so they are in series, resulting in the capacitance of 125 μF mentioned earlier. One advantage of this technique is that when in parallel each capacitor needs to be charged to a voltage of only 300 V, which becomes a total voltage of 600 V when placed in series. Most ICDs have a maximum shock energy of ∼30 J.

Lithium-type batteries, often lithium silver vanadium oxide, power ICDs. Two such batteries in series provide ∼6 V. Since the capacitor voltage is ∼600 V, the batteries are used to power a high voltage power supply. They are implanted in the patient's body, so changing them requires surgery, implying that battery lifetime is important. Lifetime is often measured in ampere-hours (A · h) (equivalent to a charge of 3600 C), and a typical battery is rated at ∼3 A · h. If each time the capacitor is charged uses 0.075 C, the battery should be able to deliver thousands of shocks. However, the battery performance begins to decay before its total charge is exhausted, and also it must provide power for continuous monitoring of the ECG and other functions, so its observed lifetime is ∼5 years. Another important property of a battery is the time required to charge the capacitor. Typically, the battery takes ∼10–20 s to generate a full charge. If this time increased significantly, it would delay the delivery of the shock. The voltage decays gradually and predictably throughout the lifetime of a lithium silver vanadium oxide battery, so that the voltage can be used as an indicator of the battery's remaining useful life.

The electrodes and their leads are critical components of an ICD. Unlike the electrodes in an AED, ICD electrodes are implanted inside a beating heart and must continue to function there for years. Many ICD malfunctions arise because of problems with the leads. Like pacemaker leads, ICD leads are made from coils of wire to make them flexible and avoid breaks (Fig. 8). They are insulated, except at the

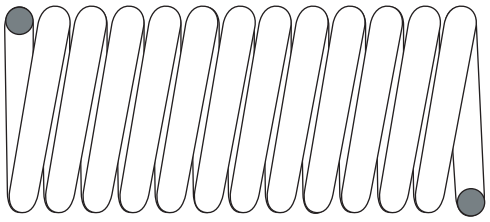


Figure 8. The conductor in the lead is often in the form of a coil to increase its flexibility and reduce mechanical stresses in the metal.

electrodes, by silicone rubber or polyurethane. A typical lead contains three electrodes: one for pacing and sensing, and two larger ones for defibrillation. An ICD lead is affixed to cardiac tissue on the inner (endocardial) surface of the heart. Often the tissue is damaged (inflammation, followed by fibrosis) in the area in contact with the lead tip. Steroid eluting leads minimize the tissue damage by slowly releasing the corticosteroid dexamethasone sodium phosphate. The ICD lead must be attached to the endocardial surface to prevent it from becoming dislodged. Some leads use a "passive" fixation technique consisting of plastic tines on the lead tip that become entangled in the trabeculae on the endocardial surface of the right ventricle (Fig. 9a). Other leads use an "active" fixation technique consisting of a metal helix, similar to a corkscrew, that is screwed into the endocardium (Fig. 9b). The defibrillation shock is delivered through a large electrode located many millimeters back from the lead tip. In some cases, current is passed through two electrodes (one in the right ventricle and one in the right atrium, as shown in Fig. 7), and in other cases the shock is delivered between one electrode and the defibrillator can.

The ICD recording lead senses the several-millivolt ECG signal within the heart. Two parameters that the ICD uses to detect abnormal arrhythmias are heart rate and arrhythmia duration. The ICDs use sophisticated algorithms to determine from the ECG if an arrhythmia is present, and these algorithms differ between manufacturers. Sufficient memory is included in the ICD to store ECGs before, during, and after a shock. Figure 1 shows that the ECG from ventricular fibrillation has a smaller amplitude than the normal ECG, making detection of fibrillation challenging. Information about the defibrillator,

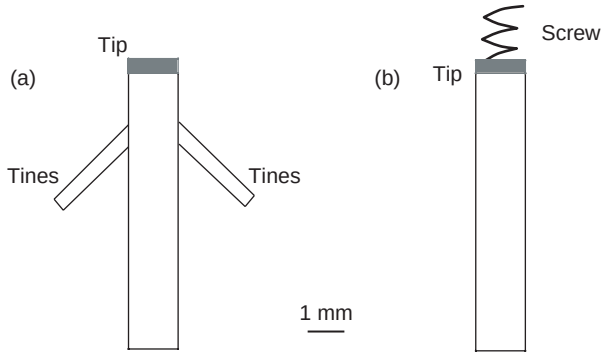


Figure 9. Distal end of a lead. (a) Passive fixation for attaching the lead to the endocardial surface of the heart using plastic tines. (b) Active fixation using a metal helix that is screwed into the heart wall.

e.g., the status of the battery and the lead impedance, as well as ECG traces, can be obtained through telemetry. Modern ICDs use the medical implant communications system radio frequency band (402–405 MHz). Most ICDs can be reprogrammed using telemetry.

When implanting an ICD, the physician must choose between a single-chamber and a dual-chamber defibrillator. Single-chamber devices have a single lead with the sensing electrode placed in the right ventricle. Their advantage is simplicity, longevity, and fewer complications. Dual-chamber devices have two leads: one sensing the right atrium and one sensing the right ventricle. Patients who rely on the ICD for pacing as well as defibrillation may benefit from the dual-chamber design. For example, a patient with a problem in the sinus node, which is located in the right atrium and serves as the heart's natural pacemaker, may respond best to atrial pacing. The atrial lead would then be for pacing, and the ventricular lead for defibrillation.

A cardiologist usually implants an ICD with the patient under local, not general, anesthesia. Typically, implantation requires an overnight hospital stay, although sometimes it is performed as an outpatient procedure. The ICD can is placed in a "pocket" under the skin in the upper chest, in the pectoral region. The lead is introduced into the subclavian vein, often by puncturing the vein with a needle, and then advanced into the right atrium under fluoroscopic view. A ventricular lead passes through the tricuspid valve between the right atrium and the right ventricle, and then is placed in contact with the endocardium near the apex of the heart (Fig. 7). After the implantation, the cardiologist tests the device by inducing fibrillation and then shocking the heart to check that defibrillation is successful.

The "C" in the term ICD stands for "cardioversion," which is a type of shock therapy for treating any rapid arrhythmia other than ventricular fibrillation. Typically, a physician uses cardioversion to treat atrial fibrillation or a ventricular tachycardia (a rapid but still organized beating of the ventricles), both serious abnormalities but neither immediately life threatening. The shock strength used during cardioversion may be weaker than or similar to that used for defibrillation, depending on the type of abnormality. Cardioversion often uses information from the ECG to time the shock optimally, whereas defibrillation shocks are delivered without any such timing.

MECHANISMS OF DEFIBRILLATION

Defibrillators have been developed empirically without a complete understanding of the mechanism of defibrillation. Many researchers are studying this important problem (4). The mechanism of defibrillation is closely tied to the mechanism of arrhythmia induction by an electrical shock. In order to initiate an arrhythmia, a shock must be given during the vulnerable period (during the time of repolarization of the ventricular action potential). A very weak shock during the vulnerable period has little effect. A stronger shock can induce an arrhythmia; the lowest strength that induces an arrhythmia is called the "lower limit of vulnerability". An even stronger shock will not cause an arrhythmia; the lowest strength above the lower limit of vulnerability for which an arrhythmia is not induced is called the "upper limit of vulnerability". In order to defibrillate, a shock must be at least as strong as the upper limit of vulnerability. If not, a shock that would otherwise successfully defibrillate the heart could restart a new arrhythmia that might decay quickly into fibrillation.

One of the more difficult issues in defibrillation research is determining exactly where, when, and how a shock affects cardiac tissue (5). The crucial question is how the shock alters the transmembrane potential (the voltage across the cell membrane), because it is the transmembrane potential that opens and closes voltage gated ion channels, thereby causing an action potential and wave front propagation. A simple, one-dimensional (1D) model of current flow through the heart wall suggests that the transmembrane potential caused by a shock is large only within a few length constants ($\sim$1 mm) of the heart's surface. This model cannot be correct, because fibrillation occurs throughout the heart and the shock must affect a large fraction of the cardiac tissue—not just a thin surface layer—if it is to defibrillate. What then is the mechanism by which a shock alters transmembrane potential deep in the heart wall? This question has not yet been answered definitively, but recent evidence suggests that tissue anisotropy and fiber curvature plays a key role (6).

Scientists continue to study defibrillation even as engineers improve defibrillator designs empirically. How increased fundamental knowledge about defibrillation will improve defibrillators is an open question. Nevertheless, defibrillators contribute crucially to the treatment of cardiac arrhythmias. They represent the best, and often only, option for patients with ventricular fibrillation.

BIBLIOGRAPHY

1. American Heart Association. Heart Disease and Stroke Statistics—2005 Update. Dallas, TX.: American Heart Association; 2004.
2. Moss AL, et al. Improved survival with an implanted defibrillator in patients with coronary disease at high risk for ventricular arrhythmia. Multicenter Automatic Defibrillator Implantation Trial Investigators. N Engl J Med 1996;335 (26):1933–1940.
3. The Antiarrhythmias Versus Implantable Defibrillators (AVID) Investigators, A comparison of antiarrhythmic-drug therapy with implantable defibrillators in patients resuscitated from near-fatal ventricular arrhythmias. N Engl J Med 1997; 337:1576–1583.
4. Ideker RE, Chattipakorn TN, Gray RA. Defibrillation mechanisms: the parable of the blind men and the elephant. J Cardiovasc Electrophysiol 2000;11:1008–1013.
5. Roth BJ, Krrassowska W. The induction of reentry in cardiac tissue. The missing link: How electric fields alter transmembrane potential. Chaos 1998;8:204–220.
6. Trayanova NA. Concepts of ventricular defibrillation. Philos Trans R Soc London A 2001;359:1327–1337.

Further Reading

Jeffrey K. Machines in our Hearts: The Cardiac Pacemaker, The Implantable Defibrillator, and American Health Care. Baltimore: The Johns Hopkins University Press; 2001. An excellent history of the pacemaker/defibrillator industry.

Zipes DP, Jalife J. Cardiac Electrophysiology, From Cell to Bedside, 4th ed. Philadelphia: Saunders; 2004. A comprehensive, multi-author reference book that covers the entire field of cardiac electrophysiology, including chapters on defibrillators. New editions have been appearing about every 4 years.

Hayes DL, Lloyd MA, Friedman PA. Cardiac Pacing and Defibrillation: A Clinical Approach. Armonk, NY: Futura Publishing Co.; 2000. An excellent introduction to both pacing and defibrillation from the point of view of a medical doctor.

Bronzino JD. The Biomedical Engineering Handbook, 2nd ed. Boca Raton, FL: CRC Press; 2000. The definitive source for information about biomedical engineering, with chapters on pacemakers and defibrillators. New editions have been appearing about every 5 years.

See also ARRHYTHMIA ANALYSIS, AUTOMATED; CARDIOPULMONARY RESUSCITATION; PACEMAKERS.

DENTISTRY, BIOMATERIALS FOR.

See BIOMATERIALS FOR DENTISTRY.

DIATHERMY, SURGICAL.

See ELECTROSURGICAL UNIT (ESU).

DIFFERENTIAL COUNTS, AUTOMATED

DAVID ZELMANOVIC
JOLANTA KUNICKA
Bayer HealthCare LLC
Tarrytown, New York

INTRODUCTION

Blood is a tissue composed of a fluid medium called serum, which contains suspended formed elements called blood cells. These include red blood cells (RBCs), white blood cells (WBCs), and platelets. The white blood cells are subcategorized as neutrophils, lymphocytes, monocytes, eosinophils, and basophils. The relative concentrations of the white blood cell types, commonly referred to as white blood cell differential counts, or simply differentials, can provide important diagnostic information regarding the blood donor. In fact, the differential is one of the standard diagnostic tests, ordered by physicians frequently. Originally, differential counts were obtained by microscopic evaluation of 100 or 200 WBCs at 500- or 1000-fold magnification. In fact, the microscopic differential counting method remains the recognized reference method, as per NCCLS H20-A (1).

The first automated hematology analyzer was a cell counter based on the Coulter Principle, as described in Ref. 2. According to this principle, a suspension of particles diluted in an electrolyte-containing aqueous medium is drawn through minute apertures on either side of which charged electrodes are positioned. The electrolytic medium and electrodes are part of an electrical circuit, which can be direct current (dc) or radio frequency (RF). As a particle passes through the aperture it raises the impedance of the circuit because of its insulating properties. The momentary change in impedance is recorded as a signal pulse in the form of a voltage. The number of pulses is proportional to the particle concentration. Additional information about the particles can be obtained from pulse height analysis. White blood cell signals may be subcategorized based on pulse heights as lymphocytes + basphils, mid-range cells (monocytes + other large mononuclear forms), and granulocytes (neutrophils + eosinophils) (Fig. 1) in order to provide a so-called three-part differential counting method. This method is widely used in small laboratories around the world. Hematology analyzers providing three-part differentials are available from all major manufacturers of hematology instrumentation, and from many small manufacturers.

Since the advent of the Coulter principle, other techniques have been developed that permit the automated determination of the full five-part differential. Innovative technology of the current hematology analyzers permits assessment of the patient's clinical status through a combination of numeric results, morphology flags, cytograms and histograms. Significant technological improvements occurred to the automated analysis of WBCs since the first edition of the EMD (3). Additional automated information on abnormal blood cells combined with microscopic smear review provides an aid in disease diagnosis and patient monitoring. There is a continuing development of more sophisticated analysis to perform extended differential counts that would include differentiation of immature cells, but methods for standardization are not finalized yet.

This article includes a discussion of the cellular properties that are used to automatically analyze white blood cells and describes the techniques used to measure these properties. Examples are provided of the implementation of these techniques on commercial hematology analyzers. Finally, there is a brief discussion of the relative merits of the analysis methods in terms of result accuracy and laboratory efficiency.

MEASURABLE PROPERTIES OF WHITE BLOOD CELLS

Morphology

Cell Size. White blood cells are distinguishable from red blood cells and platelets based on size. In normal individuals all WBCs are larger than red cells or platelets. The lymphocytes and basophils are usually smaller than neutrophils or eosinophils, which are in turn smaller than monocytes. Cell size is used by all automated hematology analyzers to distinguish WBCs from other cells in blood samples, and to further subcategorize WBC types.

Nuclear Size and Shape. White blood cell nuclei may be classified as mononuclear (single lobed) or polymorphonuclear (multilobed). Lymphocytes and monocytes are mononuclear; neutrophils, eosinophils, and basophils are polymorphonuclear. In normal samples lymphocyte nuclei are round and monocyte nuclei are kidney shaped and larger than lymphocyte nuclei. The polymorphonuclear cells typically have two to five nuclear lobes. The nuclei of immature neutrophils may be band-shaped instead of having well-defined lobes. These properties are also used in all automated hematology analyzers.

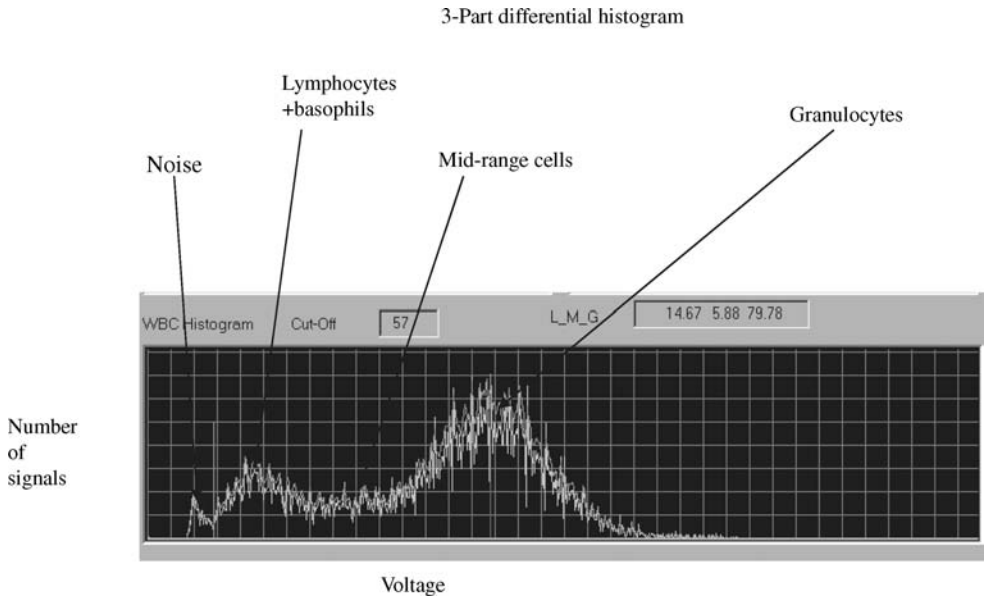


Figure 1. Coulter Principle—Three-part differential histogram.

Granule Number and Size. Cytoplasmic granularity is more pronounced for polymorphonuclear WBCs than mononuclear WBCs. Also, eosinophil granules are larger and more numerous than neutrophil or basophil granules. These properties are used by some automated analyzers to subcategorize WBC types.

Cytochemistry

RNA/DNA Staining. White blood cell nuclei and granules contain ribonucleic acid (RNA) and deoxyribonucleic acid (DNA), which are stained by nucleic acid staining dyes. These properties are used by some automated analyzers to subcategorize WBC types.

Enzyme Activity

Peroxidase. Granules in eosinophils, neutrophils, and to a lesser extent, monocytes, contain peroxidase enzyme, which catalyzes reactions between peroxide and various substrates that can result in deposition of dyes on the granules (4,5). Lymphocyte and basophil granules are peroxidase negative. This property is used in some automated hematology analyzers for subcategorization.

Esterase. Monocytes and neutrophils contain esterase enzymes, and these enzymes can catalyze reactions that result in deposition of dyes (6,7). This method can be used to distinguish monocytes and neutrophils from other cell types. This property is not currently used in automated hematology analysis.

Differential Lysis

Basophils. Basophils are more resistant to lysis than the other WBC types under certain pH conditions (8,9) and this property is used commercially to distinguish them from the other WBC types.

Eosinophils. Eosinophils are more resistant to lysis than other WBC types under certain pH conditions in the presence of polyoxyethylene nonionic surfactants (10) and this property is used commercially to distinguish them from the other WBC types.

Immunocytochemistry

Multidimensional fluorescence flow cytometry uses the measurement of cell physical properties of cell size and cytoplasmic complexity, in addition to surface marker phenotyping using fluorescence-labeled monoclonal antibodies. Monoclonal antibodies have been produced that are specific to WBC generally, such as, CD45, and to specific WBC types, such as, CD3 for T-lymphocytes. These may be labeled with absorbent or fluorescent dyes for identification.

MEASUREMENT TECHNIQUES

Commercial analyzers use some combination of the following six measurement techniques to count WBCs and to distinguish among their subtypes:

Aperture Impedance

This technique uses the electrical insulation properties of WBCs in conjunction with the Coulter principle, described above. As stated, the amplitudes of WBC signals depend on size and to some extent, intracellular properties. Therefore, at least three distinct signal amplitude populations form on impedance signal frequency histograms.

Light Scattering

This technique uses the optical refraction properties of WBCs. White blood cells, suspended in a medium whose

refractive index is significantly different (normally lower) than those of the cells, pass essentially in single file through a narrow aperture in a clear glass block. The suspension column is sheathed in a fluid whose refractive index matches that of the medium. A collimated beam of typically monochromatic visible light is incident on the glass cube in a direction that is perpendicular to the direction of the cell stream. As the cells interrupt the beam of light, they scatter it in a manner that is characteristic of their size and refractive index, and to a lesser extent of their internal properties, including granularity and nuclear configuration. Each signal pulse corresponds to an enumerated WBC. Light scattering intensities are a function of the scattering angle, and intensity measurements are usually made over two, three, or even four different scattering-angle intervals. The scattering intensity patterns are characteristic of the WBC subtypes. Two- or higher-dimensional scattering intensity plots form signal clusters that are associated with different WBC subtypes. These scatter–scatter plots are called scattergrams or cytograms, since they are multidimensional plots of the scatter signals generated by cells.

Optical Absorption

This technique uses the cytochemical properties of WBCs that permit the cells to be stained or to accept dye in a manner characteristic of the cell subtype, in conjunction with pretreatment of the cells. The measurement process is similar to that for light scattering, except that a characteristic decrease in light transmission is measured.

Fluorescence

Fluorescence is the reemission, at a lower frequency, of light absorbed at a given frequency. Fluorescence signals are generally larger than absorption signals. This technique uses cytochemical properties, as above, except that the stain is fluorescent rather than absorptive. This technique for obtaining WBC differentials should be distinguished from the immunofluorescence technique that relies on the lineage-specific expression of cell surface antigens to produce distinct so-called clusters of designation (CDs). Although immunofluorescence can be highly specific and therefore very accurate, it is not used in commercial hematology analyzers because of the high costs of the fluorescent-labeled monoclonal antibodies required to specifically tag the cell surface antigens.

Automatic Pattern Recognition

This method relies upon computerized pattern recognition algorithms to classify WBCs. Stained blood film slides are mounted on a microscope stage with motor-driven advancement of the slide in the plane perpendicular to the optical axis. Cellular images are captured on CCD arrays and the images are analyzed by pattern recognition techniques. A WBC differential can be reported based on the classifications. The basic technique was developed in the 1960s by a company called Geometric Data, in a product called the Hematrak (11). A current version of this technology, called Cellavision, is available from a company called CellaVision, Inc. (1555 Jupiter Park Dr., Suite 6, Jupiter, FL 33458, www.cellavision.com.

This technique is not widely used because the instrumentation, which provides a WBC differential, but not an absolute count, is expensive. The per-sample cost is also high. Further, until recently the throughput was low. Currently, Cellavision claims a throughput of 100 samples per hour, which is comparable to the throughput of major automated hematology analyzers.

One advantage of this technique is that the fixed, stained blood films, which are actual whole blood samples, can be stored for years. Also, since the positions of the imaged cells can be recorded, the cells can be recalled for future manual review. This is not possible with routine automated hematology analyzers, where the analyzed cell suspensions are disposed of immediately.

Image-in-Flow

This technology combines flow cytometry and image analysis. As suspended stained cell flow through a narrow aperture in an optical flow cell, their images are captured on CCD arrays. Pattern recognition algorithms analyze the images and classify the cells. In addition, these instruments act as flow cytometers and provide light scattering intensities and fluorescence intensities. The scattering and/or fluorescence data can be displayed on cytograms. As a result, it is possible to select a point on a cytogram and display the associated image. The Amnis Corporation ImageStream 100 is an example of such a device (12). Sysmex Corporation has reported on an experimental version of such a device, as well (13).

SAMPLE STABILITY

The physical and chemical properties of WBCs are subject to change *in vitro*. The extents of these changes depend on both storage time and temperature. First, WBC swell, then their membranes become leaky and they release their granules, and ultimately they autolyse. In addition, WBC nuclei undergo subtype-dependent configuration changes *in vitro*. Degradation is more rapid at ambient temperatures than under refrigeration. The rate of degradation differs according to the WBC subtype.

Swelling affects both cell size and refractive index. Also, since the granules are denser than their cytoplasmic medium, the cells become less dense upon granule release, even without consideration of swelling. Further, granule release affects the cytochemical properties of WBCs in a subtype-specific manner.

Generally, cell morphology is less stable over time *in vitro* than enzymatic properties, such as peroxidase and esterase activity. It is also less stable than nucleic acid staining capability. Therefore, automated analyzers that use only morphological properties to determine WBC differentials are more limited in terms of the *in vitro* age of samples that they can accept for analysis than are analyzers that use cytochemical properties. For example; automated analyzers that use light-scattering patterns to distinguish among WBC subtypes based on differences in morphology, use either fixed gates based on typical cell patterns to define

cell populations, or combinations of gates and pattern recognition techniques, such as cluster analysis. Analyzers that use fixed gates provide inferior discrimination to those that include cluster analysis, because cell clusters shift as a result of morphological changes. Even analyzers using cluster analysis cannot distinguish well among cell populations once these begin to merge due to cell degradation. To the extent that cell cluster positions can be maintained, these limitations are overcome. Cytochemical staining based on enzymatic activity of cells, along with absorption or fluorescence measurements, provides added cluster position stability because of the relative stability of these cellular properties.

SAMPLE PREPARATION

White blood cell concentrations in peripheral blood samples normally range from $4 \times 10^3 - 11 \times 10^3$ per microliter (μL). Red blood cell concentrations normally range from $4 \times 10^6 - 5.5 \times 10^6 \cdot \mu L^{-1}$, and platelet concentrations from $150 \times 10^3 - 400 \times 10^3 \cdot \mu L^{-1}$. Individually, WBCs, RBCs, and platelets are mutually distinguishable by size alone. However, in undiluted whole blood samples the concentration of RBC is so large that electronic sensors can detect only a single prolonged signal due to the ever-present red cells. Given typical signal processing conditions for automated analyzers, this remains true even at 50-fold sample dilutions. It is not until approximately a 500-fold dilution that the signal interference from RBCs in a sample becomes manageable, but at this dilution only 100 or so WBCs are counted in a typical cycle in which 50,000 or so RBCs are counted. This is statistically inadequate for automated WBC differential determinations. Automated analyzers deal with this issue by selectively destroying (lysing) the RBCs in a whole blood sample, by adding surfactant and/or by reducing the osmolality of the suspension medium. This is usually done at a dilution ratio of 40–100:1 to maintain adequate WBC concentrations for counting purposes. In the absence of RBCs, at a dilution ratio of 50:1 as many as 100,000 events can be automatically analyzed within 10–20 s.

SIGNAL GENERATION

The number of WBC events analyzed by automated hematology systems during a measurement cycle must be controlled. Typically, 5000–10000 WBC events are counted during the cycle. In addition to controlling the number of events counted, the systems must control the quality of the observations made. This is necessary because in both optical and aperture impedance measurements the signal generated by a particle in the sensing zone depends on the position of the particle within the zone. To minimize particle position variability, the stream of cells in a WBC suspension is centered within the zone. The cell suspension is constricted to the center of the zone by enveloping it in a fluid cladding called a sheath. This constriction is often referred to as hydrodynamic focusing. Sheathing also serves to control event frequency.

The signals generated in aperture impedance systems and in light scatter systems are measures of cellular properties, such as size, density, granule content, nuclear size, and shape. The signals in systems using light absorption or fluorescence measurements are based on the labeling of cellular components by various dyes/stains.

Measurement artifacts that can interfere with signal generation include spurious signals resulting from high frequency electronic noise; often due to improper electrical grounding of electrical components. Spurious signals arising from light output instability, often due to uncontrolled switching from one laser light emission mode to another. Truncation of signals or short-term signal intensity variations associated with either poor hydrodynamic focusing of cells in the flow stream or misalignment of the stream in the signal generation path. Signals associated with particulate matter other than cells. These are often due to impurities in reagent containers or to precipitation of reagent components due to mishandling or improper storage.

SIGNAL DETECTION

Aperture Impedance

In both dc current and RF current versions, signals appear as voltages. Direct currents primarily probe cell size, whereas RF currents probe cell features, such as granularity, nuclear lobularity, and cell density. Impedance measurements are by themselves adequate for automated three-part differentials, but for automated five-part differential analysis they must be combined with at least one other measurement.

Light Scattering

Four scatter regions are usually associated with optical detectors for automated hematology analysis: axial or forward scatter (0–1°), low angle scatter (1–5°), high angle scatter (5–45°), and very high angle scatter (45–90°). The axial- or forward-scatter detectors are considered to be sensitive mainly to cell size. Low angle scatter is also associated mainly with cell size. The higher angle regions are associated with internal structure, mainly as a result of multiple scattering from numerous granules and/or from multilobed nuclei. Although there is an association of low angle scatter with cell size and high angle scatter with internal cell properties, both types of scatter depend on size, refractive index, and internal cell properties. Also, the angle cutoffs in parentheses are only by way of example and are not definitive, since smaller segments within these regions may be selected for optical detection, or the collection angle range may bridge the regions.

Light Absorption

Light absorption is detected as a loss of transmitted light. Detectors usually encompass up to $\pm 20^\circ$ of forward light scatter. The absorption is the difference between the light transmitted in the absence of and in the presence of a cell. This measurement technique is use in association with the selective uptake of an absorptive dye as a result of a distinguishing chemical feature of a WBC subtype.

Light loss is not identical to light absorption because it also involves light that is not absorbed, but that is instead

scattered outside the collection cone of the detector. This is referred to as pseudoabsorption. In practice, pseudoabsorption contributes significantly to axial light loss for cells with large, numerous, and closely spaced granules that cause multiple light scattering to occur, with subsequent scatter at relatively large angles.

Light absorption signals are usually smaller than fluorescence signals and may also be smaller than low angle light scattering signals because the combination of dye extinction coefficient and the short path traversed through a cell usually results in only minor absorption. Certain combinations of concentration of granular material and absorptive dyes, such as eosinophil granules and 4-chloronaphthol, provide exceptions. Also, light absorption measurements have low signal/noise ratios because they are detected as fractional (1% range) reductions in light transmission values, and can have cell-specific pseudoabsorption contributions.

Fluorescence

Fluorescence detectors are normally placed at 90° to the incident beam to eliminate the contribution from the light source. This is not formally required, since the fluorescence signal is at a longer wavelength than that of the incident light. Therefore, a wavelength-selective beam splitter may be placed directly in the transmitted-light optical path, with the transmitted component following this path and the fluorescence component a diverted path.

Fluorescence signals are normally larger than absorption signals by their nature. Also, they have relatively high signal/noise because they are detected at different wavelengths than that of the incident radiation so that there is no interference from incident radiation. Also, they do not ride on much larger signals, as in the case of light absorption (Table 1).

EXAMPLES OF WBC DIFFERENTIAL ANALYSIS ON HEMATOLOGY SYSTEMS

The systems are listed alphabetically, by manufacturer name.

Abbott Cell-Dyn 4000 Hematology System and Bayer ADVIA 70 Hematology System

The Cell-Dyn 4000 and ADVIA 70 Systems use similar methodologies, and will therefore be described under the same heading. The ADVIA 70 will be described first, because it is of somewhat simpler design.

The ADVIA 70 Hematology Analyzer (Bayer HealthCare LLC, Diagnostics Division, Tarrytown, NY) combines the results of an optical channel and a dc electrical impedance channel to determine the WBC differential count.

In the impedance channel, a blood sample reacts with a basic reagent that contains a surfactant. The reagent lyses RBCs while maintaining the integrity of the WBCs. As stated previously, the WBC types that may be distinguished on this basis are lymphocytes + basophils, which generate the smallest WBC signals, so-called midsized cells, which usually contain monocytes, and granulocytes that include neutrophils and eosinophils. Although monocytes are expected to produce the largest impedance signals based on their size, they do not do so in this reaction system. This is because the reagent compromises the cellular integrity of the monocytes more than those of the neutrophils or eosinophils.

In the optical channel, a whole blood sample is diluted in a reagent that contains a surfactant to lyse RBCs and a fixative to maintain the integrity of WBCs. The reaction mixture is analyzed in the optical flow cell similar in design to the one described bwlow for the ADVIA 2120 Nuclear Density Channel. The light source is a laser diode emitting radiation at 633 nm. The analysis channel includes four silicon photodetectors, to determine the following: (*1*) Light extinction (Ex), which is the loss in transmission along the axis of incident radiation, (*2*) small-angle scattering intensity (Sa), (*3*) wide-angle scattering intensity (Wa), and (*4*) Super-wide-angle scattering intensity (Swa).

The analyzer displays two cytograms that are based on the signals from the four detectors: the Size cytogram is the graph of the small-angle versus wide-angle signal pairs and it is used to distinguish among neutrophils + eosinophils, monocytes, lymphocytes, and basophils. The distinctions are based on a combination of cell size, to which the *Y*-axis signals (small-angle) are most sensitive, and refractive index, to which the *Y*-axis signals (wide angle) are most sensitive. The other is called the Structure cytogram and is used to distinguish eosinophils from the other white cell types based on the eosinophils' internal structure. Eosinophils have larger, more numerous granules than the other white cell types. These granules scatter more light into larger angles than the other white cell types do because of multiple scattering of incident radiation among the granules. This diffuses the scattered radiation into larger angles than would result from single scattering.

Table 1. Measurement Techniques Used by the Major Manufacturers of Automated Hematology Analyzers for Routine Five-Part WBC Differential Determinations[a]

Hematology System/Manufacturer	dc Impedance	RF Impedance	Scatter	Absorption	Fluorescence
CD4000/Abbott	X		X	X	
Pentra 120/ABX	X		X	X	
ADVIA 2120/Bayer			X	X	
ADVIA 70/Bayer	(X)[a]		X		
LH750/Beckman Coulter	X	X	X		
XE-2100/Sysmex	X	X	X		X

[a]Techniques used for other than routine analyses are not included.

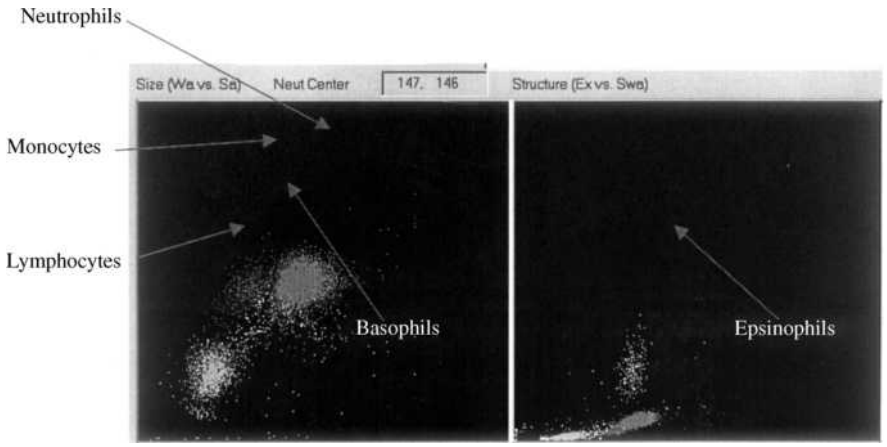


Figure 2. Differentiation of WBC populations based on impedance measurements—ADVIA 70 cytograms.

The ADVIA 70 optical channel analyzes 10,000 cells for each sample and it provides the WBC percentages only. In combination with the absolute WBC count from the impedance channel, the system provides the absolute white cell type counts. The system also compares the lymphocyte and granulocyte percentages from the two channels as a cross-check of results validity and to flag for the presence of abnormal WBC types.

Figure 2 shows associated Size and Structure cytograms. The populations in each figure are labeled and they correspond to lymphocyte signals, mid-range cell signals, and granulocyte signals, respectively. The lymphocyte signals occupy the region closest to the origin since they are the smallest cells and have relatively low refractive index values. The monocyte signals appear higher along the *Y*-axis, but only slightly to the right of lymphocytes because they are the largest cells, but also have the lowest refractive index values. The Neutrophil + eosinophil signals appear highest along the *Y*-axis and furthest along the *X*-axis because they are larger than lymphocytes and have higher refractive index values than either monocytes or lymphocytes. The basophil signals occupy the region below and to the right of monocyte signals, based on a combination of their sizes and refractive index properties. In the second panel of Fig. 2, the eosinophil signals are located higher along the *Y*-axis, which corresponds to super-wide angle signal intensity, than signals from the other cell types because of the multiple-scattering properties of the eosinophil granules.

The Cell-Dyn 4000 Hematology System (Abbott Laboratories, Abbott Park, Ill.), is similar in its routine WBC differential analysis technology to the ADVIA 70. It uses both impedance measurements and optical measurements for counting and differentiating WBC types (14–16).

The system differs from the ADVIA 70 system in two ways. First, it uses a 488 nm argon-ion laser instead of a 633 nm diode laser. Scattering patterns are sensitive to wavelength, so that scattering intensity at a given angle is different for 488 nm illumination than 633 nm illumination (16). Second, the CD4000 system distinguishes eosinophils from neutrophils based on 90° depolarized-light scatter "90D" versus 90° polarized-light scatter "90" (Figs. 3–4). According to the manufacturer, 90°D measurement is especially sensitive to granularity, whereas 90° polarized light is sensitive to lobularity. Since eosinophil granules are generally larger and more numerous than those of neutrophils, they are expected to scatter more light into 90D than neutrophils.

The underlying concept in 90D measurement is the multiple scattering experienced within eosinophils. As the initially polarized light scattered from the first granule encountered is incident on subsequent granules, the polarizarion state of the light is progressively scrambled, so that the depolarized component of light increases at the expense of the polarized component.

ABX Pentra 120 Hematology System

The Pentra 120 Hematology System (ABX, Montpellier, France) uses a combination of dc impedance and light absorbance to determine the routine WBC differential count (17–21). Two reaction mixtures are prepared by

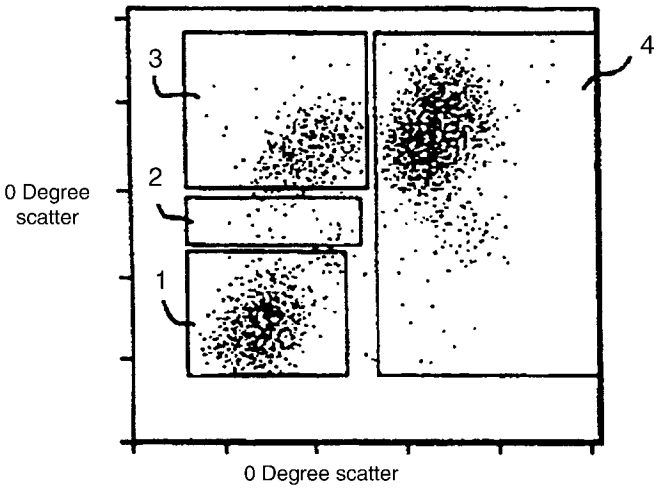


Figure 3. Differentiation of WBC populations based on light-scattering intensity Cell-Dyn 4000 0-degree vs. 10-degree scatter Cytogram.

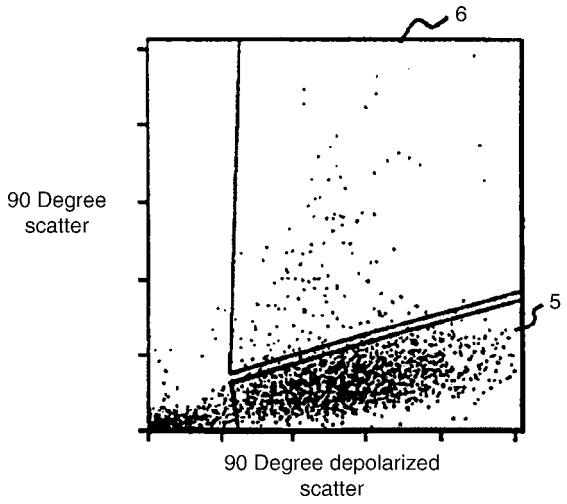


Figure 4. Differentiation of eosinophils from other WBC populations based on polarized vs. depolarized light scattering intensity-Cell-Dyn 4000 polarized light vs. depolarized light Cytogram.

the system. In one, the sample is first mixed with a reagent called Eosinofix, that lyses RBC and stains WBC differentially with chlorazol black, based on cell granularity. The mixture is subsequently diluted with a reagent that stops the reaction process in the first mixture. The suspension passes through a single flow cell that measures both the dc impedance of the WBC and their absorption of incident 488 nm radiation from the argon–ion laser. The resulting signals are displayed on a cytogram, where the *X*-axis corresponds to cell volume based on dc impedance, and *Y*-axis corresponds to light absorption (Fig. 5). On this cytogram, lymphocyte signals are clustered nearest to the bottom of the *X*-axis since they are typically the smallest WBC type. Neutrophil signals are higher along the *X*-axis since they are larger than lymphocytes. Monocyte and eosinophil signals are higher still, and at roughly the same height, based on their size. Lymphocytess signals are closest to the bottom of the *Y*-axis because they effectively do not stain with chlorazol black. Monocyte signals are slightly higher because monocytes stain weakly. Neutrophil signals are higher because they stain more heavily, and eosinophil signals are the highest because they stain the most.

This cytochemical staining method is similar to that used on the ADVIA 2120 system. Indeed, this cytogram is similar in appearance to the Peroxidase Channel cytogram of the ADVIA 2120 system when rotated through 90 and viewed in reflection.

In the second reaction mixture RBCs, as well as all WBCs except for basophils are lysed in a reagent called Basolyse (Roche Diagnostics). The resultant reaction mixture passes through the same flow cell as used for the first reaction mixtures and the dc impedance signals of the WBC are recorded (Fig. 6). The intact basophils produce distinctly larger signals than the other WBC types, and are enumerated on this basis. The basophil count is subtracted from the lymphocyte count obtained from the Eosinofix reaction mixture. This method of enumerating basophils is chemically similar to those used in both the Bayer ADVIA 2120 and Sysmex XE-2100 systems.

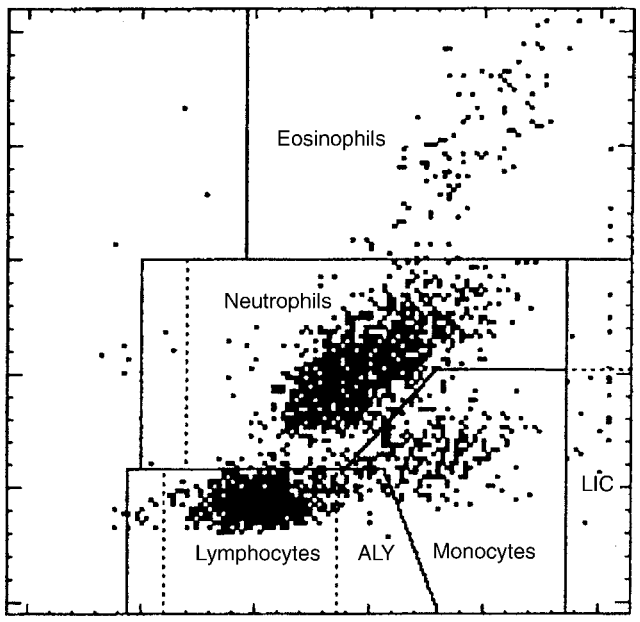


Figure 5. Differentiation of WBC populations based on DC impedance (to yield volume) and light absorption due to staining by chlorazol black-ABX Pentra 120 volume vs. light absorption cytogram.

Bayer ADVIA 2120 Hematology System

In the ADVIA 2120 Hematology System (Bayer HealthCare LLC, Diagnostics Division, Tarrytown, NY), the results of two optical channels, the Peroxidase Channel and the Lobularity/Nuclear Density Channel, are combined to produce the white blood cell differential count (8,9,22,23).

The Peroxidase Channel measures the peroxidase activity inherent in the WBC types, along with differences in cell type size, to distinguish among the cell types. A whole blood samples is mixed first with a reagent that lyses the sample's RBCs and fixes the WBCs, and two additional reagents that contain hydrogen peroxide and the dye 4-chloronaphthol are added to the mixture. The cells' native peroxidase enzyme catalyzes the reaction of the peroxide with the naphthol, resulting in the precipitation of the dye on the cells' granules. After a few seconds of incubation, the cell suspension is then passed through an optical flow cell for analysis. Two silicon photodetectors

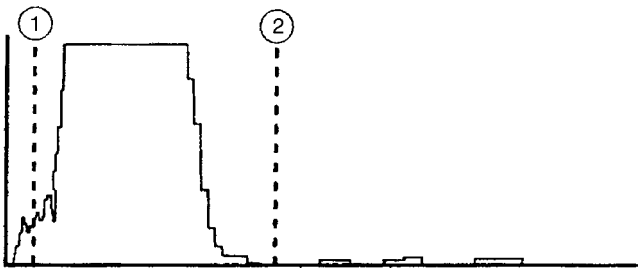


Figure 6. Differentiation of basophils from other WBC populations based on DC impedance (to yield volume)- ABX Pentra 120 volume frequency histogram.

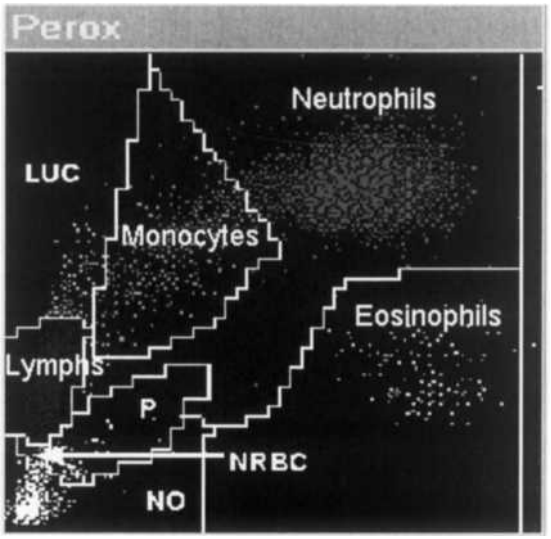


Figure 7. Differentiation of WBC populations based on cytochemistry – ADVIA 2120 Peroxidase Channel Cytogram.

are located such that one collects a component of the small-angle scattering intensity, and the other senses the drop in light transmission due to light absorption by the 4-chloronaphthol deposited on the cells' granules. A fixed volume at known dilution passes through the flow cell, so that an absolute concentration of cells can be determined. A peroxidase cytogram for a normal blood sample appears in Fig. 7.

The cell types are labeled, including "noise", which consists mainly of platelets. The positions of cells in certain regions of the peroxidase cytogram reflects cell morphology. The lymphocytes are typically the smallest cells and are also peroxidase-negative. Therefore, they appear along the lower part of the *Y*-axis, which represents the low angle scattering intensity and to the left along the *X*-axis, which represents light absorption and is sensitive to transmission loss. The LUCs (large unstained cells) are also peroxidase negative, but are larger than normal lymphocytes, so that their signals appear higher along the *Y*-axis, but at about the same position along the *X*-axis. Monocyte signals appear above and to the right of normal-sized lymphocytes. They are larger than lymphocytes and slightly peroxidase-positive. Neutrophil signals appear slightly above monocyte signals and to their right. They are typically smaller than monocytes but more peroxidase-positive. Finally, eosinophils, which are typically as large as or larger than neutrophils, have signals that appear to the right of and below the neutrophils.

The Lobularity/Nuclear Density Channel, also called Basophil channel uses the following two WBC features to distinguish among the cell types:

1. Basophils are significantly more resistant to lysis under acidic conditions than the other white cell types.
2. Mononuclear cell nuclei (MNs) scatter light in a different manner than polymorphonuclear nuclei (PMNs). The PMN scattering intensity depends on the number of nuclear lobes; the more lobes per nuclear volume, the greater the low angle scattering intensity.

In this channel, a whole blood sample is mixed with a reagent that is acidic and contains a surfactant. The reagent lyses RBC and platelets, and strips all white cell types of their cytoplasm except for basophils. The reaction mixture is passed through an optical flow cell for analysis. The measurement includes light scattered at ± 2–3° off the axis of the incident beam, and the other measuring light scattered at ± 5–15°.

The Lobularity/Nuclear Density Channel cytogram with cell types labeled is shown in Fig. 8. The basophils, which remain intact, scatter significantly more light than the much smaller nuclei of the other white cell types. Therefore they appear in the upper region of the *Y*-axis, which corresponds to 2–3° scattering. The nuclei appear near the bottom of the *Y*-axis due to their small size. However, since the nuclei have a higher refractive index than that of the intact basophils, at least some of them appear to the right of the basophils along the *X*-axis, which corresponds to 5–15° scatter. As noted above, scattering intensity is a nonlinear function of both size and refractive index. Further, it depends on the number of scattering particles encountered at one time in the sensing zone. The lobes of MNs are single scatterers, whereas the lobes of PMNs behave to a first approximation as multiple scatterers. The scattering pattern of the MNs is due to a combination of their single-lobed nature and their relatively low refractive index. The scattering pattern of the PMNs is due to a combination of their relatively high refractive index and the number of nuclear lobes.

The ADVIA 2120 system uses the Peroxidase Channel results to determine the percentages and absolute numbers of neutrophils, eosinophils, monocytes, lymphocytes + basophils, and LUCs, and to provide a WBC count for comparison with the Nuclear Density Channel WBC count.

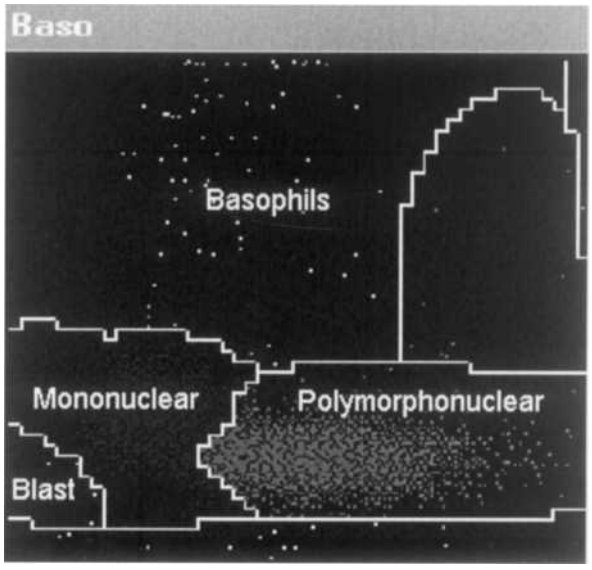


Figure 8. Differentiation of WBC populations based on scatter measurements – ADVIA 2120 Nuclear Density Channel Cytogram.

The primary WBC count determined and the number and percentage of basophils is determined in this channel. The absolute white cell subtype counts are computed as the product of the differential percentages and the primary WBC count. The system uses the additional information provided by the MN and PMN percentages and counts, and the Peroxidase Channel WBC count to cross-check the validity of the results of the two channels and to test for the presence of abnormal cells.

Beckman Coulter LH 750 Hematology System

The LH 750 Hematology System (Beckman Coulter, Hialeah, FL) uses VCS technology for routine WBC differential analysis (24–27). In the VCS technology (Volume/Conductivity/Scatter), the RBCs in an aliquot of blood are lysed in one reagent and then the WBCs are stabilized in a second reagent. The stabilized WBC suspension passes through a single quartz flow cell that is used for both electrical and optical measurements.

Cell volume (V) is determined based on dc impedance signals. The RF impedance signals provide information about internal structure, such as nuclear volume and internal chemistry. Light scattering signals from 10 to 70°, called median angle light scattering (MALS) provide information about cell granularity and lobularity. Since both RF signals and light scattering signals depend on cell size as well as on internal structure, whereas dc impedance is considered to be a function of size only, the RF signals and scattering signals are corrected for the contribution due to cell size, based on the dc signals. The resulting RF signals are called "opacity" because the signals are considered to be sensitive primarily to the density of internal components. For example, opacity is used to distinguish normal lymphocytes from variant lymphocytes, because of characteristic differences between the two cell subtypes in nuclear/cytoplasmic ratio. The volume-compensated MALS signals are called Rotated Light Scatter (RLS) signals. Volume compensation serves to better separate eosinophils from neutrophils on the one hand, and monocytes from lymphocytes on the other hand.

The WBC differential data are displayed as V (DC impedance) signal intensity versus RLS (compensated MALS) intensity signals (Fig. 9). In this cytogram, lymphocytes and monocytes appear to the left along the RLS axis, since they are mononuclear. Monocytes appear above lymphocytes along the *V* axis since they are larger. In fact, for normal samples they are highest along the *V* axis since they are the largest cells. Neutrophils and eosinophils appear to the right along the RLS axis, since their nuclei are polymorphonuclear and they have granularity. Eosinophils appear further to the right than neutrophils, because of their larger, more numerous granules. Basophils appear to the right of and somewhat above lymphocytes, based on size, lobularity, and granularity.

Sysmex XE-2100 Hematology System

The XE-2100 Hematology System (Sysmex Corporation, Kobe, Japan) combines fluorescence, forward and side scatter, and dc and RF impedance to determine the five-part WBC differential (Sysmex 28-36). It also provides

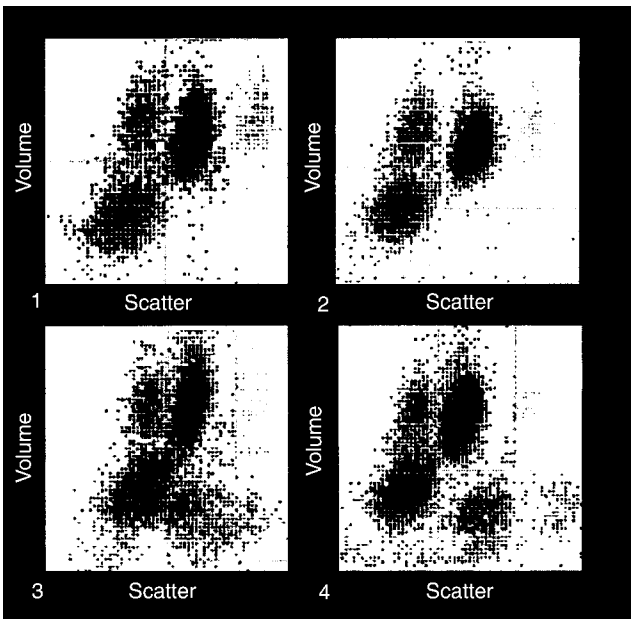


Figure 9. Differentiation of WBC populations based on DC impedance (to yield volume) and light scattering intensity-Beckman Coulter LH 750 volume vs. RLS cytograms.

absolute and differential counts of immature WBC, including bands, metamyelocytes, myelocytes, promyelocytes, and myeloblasts as part of the routine differential analysis (28). Three aliquots of whole blood are separately reacted and analyzed.

One aliquot is diluted with a reagent that lyses RBCs and compromises the integrity of WBC membranes, except for basophils. The suspension is passed through an optical flow cell, the suspended cells interrupt a beam of red light from a laser diode, and the forward scattering intensity and side scattering intensity are measured (Fig. 10). The basophils in the suspension produce larger forward scatter and side scatter signals than the other WBC types because they are larger, having retained their cellular integrity and their granules. This method for determining the basophil differential count is similar to that used by the ADVIA 2120 system.

A second aliquot is first diluted with a reagent that lyses the red blood cells and permeabilizes the membranes of the white blood cells to the passage of a red-fluorescent polymethine dye that stains RNA/DNA (29,30). A second reagent containing the dye is then added. As the suspended cells pass through the optical flow cell and interrupt the red laser-diode beam, the side fluorescence and side scatter signal intensities are measured (Fig. 11). Lymphocytes and monocytes produce larger fluorescence signals than neutrophils, basophils, or eosinophils in this reaction channel. This is presumably because the dye preferentially stains RNA, and lymphocytes and monocytes contain more cytoplasmic RNA than neutrophils, basophils, or eosinophils. On the other hand, lymphocyte and monocytes produce smaller side scatter signals than neutrophils and basophils, which in turn produce smaller side scatter signals than eosinophils. The mononuclear cells scatter least because they are less refractile than the polymorphonuclear

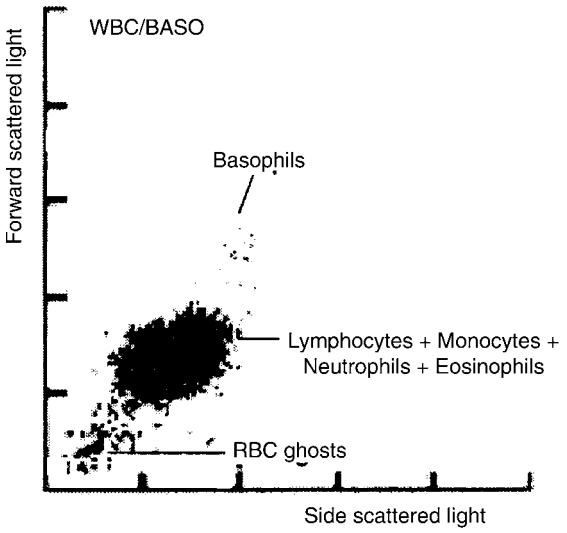


Figure 10. Differentiation of basophils from other WBC populations based on light scattering intensity. Sysmex XE-2100 forward scatter vs. side scatter cytogram.

cells. Neutrophils and basophils scatter less than eosinophils because they lack the side scatter component provided by eosinophils' large, numerous granules.

A third aliquot of blood is diluted with a reagent that lyses RBCs and maintains the cellular integrity of immature WBC types in preference to that of mature WBC. The cell suspension is passed through a narrow aperture on either side of which is an electrode. The electrical circuit, completed by the electrodes and the conductive reaction suspension medium, carries both a dc and RF current. The RF and dc impedance signals are measured for each cell as it passes through the aperture (Fig. 12). The mature WBCs and immature white cells produce RF signals of similar magnitude, but the immature cells produce larger dc signals. In this reaction mixture their cellular membrane

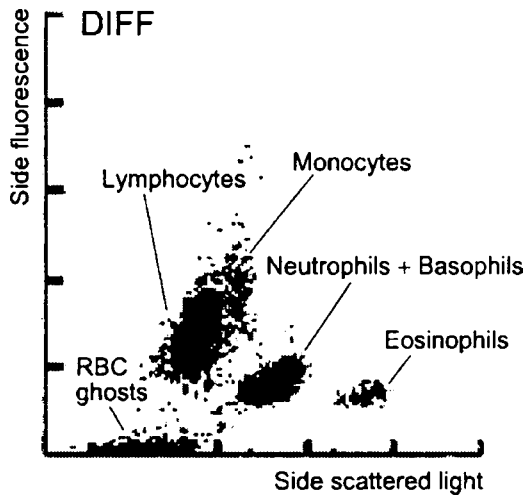


Figure 11. Differentiation of WBC populations based on fluorescence intensity and side scatter intensity. Sysmex XE-2100 side fluorescence vs., side scatter cytogram.

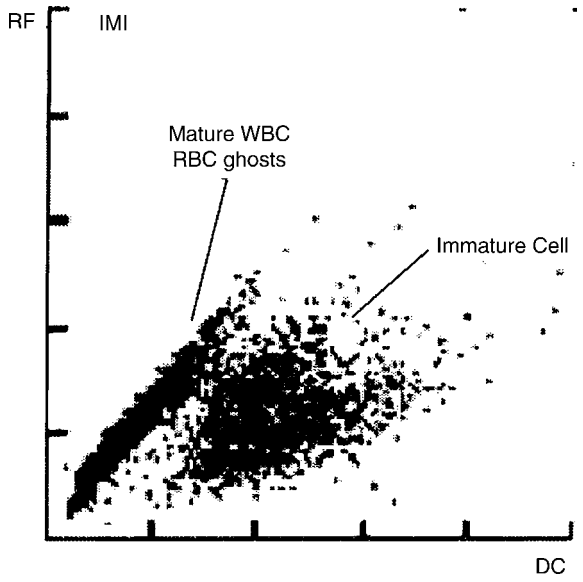


Figure 12. Differentiation of immature WBC populations from mature WBC populations based on RF and DC impedance Sysmex XE-2100 RF vs. DC cytogram.

integrity is superior to that of the mature white cells, and the conductivity of the mature cells is reduced because of the compromised cell membrane integrity.

Abbott Cell-Dyn 1700, ABX Micros 60, Bayer ADVIA 60, Beckman Coulter Ac•T Diff, Sysmex KX-21N

The above are examples of aperture impedance devices for determining three-part differentials, as described above.

MEASUREMENT TECHNIQUE VERSUS ACCURACY OF RESULTS AND LABORATORY EFFICIENCY

All of the automated five-part differential analyzers described above provide an accurate five-part differential counts for fresh, normal samples as per the specifications listed in the respective operator's manuals. Differences in performance arise for samples stored *in vitro* longer than 8 h at room temperature or >24 h at 4° centigrade, as discussed above. Even larger differences may arise for samples with abnormal morphological and/or cytochemical properties. If the cells in the blood samples are morphologically abnormal, but retain their cytochemical properties, then analyzers using cytochemistry will provide accurate enumeration while analyzers using morphological properties will not. On the other hand, retention of cytochemical properties may mask an important underlying cause of abnormal morphology that morphological analysis may reveal through abnormal cytograms.

Laboratory efficiency can be evaluated in terms of throughput and cost. All of the major analyzers produce at least a hundred five-part differential results per hour. Also, the per-test cost, which includes reagents and instrumentation, as well as associated laboratory overhead, such

as space and lab personnel, does not vary significantly among manufacturers. However, throughput is affected by factors other than analyzer speed. Even though an analyzer can report 100 results per hour, the laboratory routinely does not immediately release the results without review. The reviews invariably trigger a re-analysis of some fraction of the results. Reanalysis rates, also called review rates, may vary from 10 to 60% depending on a number of factors, listed below. Effective management of review rates optimizes laboratory throughput by creating the proper balance between throughput and accuracy of results. This balance in turn bounds the cost of laboratory operations.

Since the review rate significantly affects laboratory efficiency, it is important to list the factors affecting the review rate, in order to compare the performance of different types of analyzers with respect to these factors. The factors include type of laboratory; donor population; sample age; review criteria.

If the laboratory is in a hospital, then the samples it receives will usually be fresh (<8 h postvenipuncture), so that differences in analyzer performance associated with sample stability will not be a factor. If it is a reference laboratory, which typically receives samples that are 4–48 h old, then these differences may significantly impact the review rate. Analyzers that use cytochemical properties of cells such as enzyme activity or nuclear staining to determine differentials may provide more reliable results and therefore reduce review rates than analyzers that use cell morphology alone to determine differentials.

If the donor population is comprised of mostly normal donors, such as occurs in labs that perform screens for insurance companies and the like, then the review rates can be expected to be low and any of the automated analyzers will provide good results. In this donor population, differences in the low review rate will depend more on differences in analyzer reliability than on differences in method accuracy. If the donors are from a general hospital population where a wide range of conditions apply, then technologies that are robust with respect to hematologic variations can be expected to produce a lower review rate than technologies that are suited to only narrow ranges of hematologic conditions. Hematologic variations may include wide swings in WBC concentration, wide swings in differential ratios, cell morphology abnormalities including size variations, changes to nuclear properties and changes to cytochemical properties. If the donor populations are well defined, such as oncology patients, newborns, end-stage-renal dialysis patients, thalassemics, sickle cell disease sufferers and so on, then analyzers whose measurement techniques are best suited to the given population should be selected, in order to optimize review rate. Although sample age correlates strongly to laboratory type, it also comes into play when samples that are normally run fresh must be stored for extended periods before being analyzed. In this case, analyzers that use more stable cellular properties for analysis and that are also accurate over a wide range of hematologic conditions are preferred.

Although review criteria are expected to vary widely based on differences in sample age and donor population, they can still vary widely even among laboratories of like type and among laboratories that handle the same types of patient populations. The reason is that there is wide latitude among laboratories in what is considered an accurate result and in what is considered an abnormal result. Therefore, variability in review rate is probably attributable more to the lab's choice of criteria than to any methodology-related factors.

BIBLIOGRAPHY

1. Reference Leukocyte Differential Count (Proportional) and Evaluation of Instrumental Methods: Approved NCCLS Document H20-A. Villanova (PA): National Committee for Clinical Laboratory Standards; 1992.
2. Wallace Coulter: Means for counting particles suspended in a fluid. US Patent 2656508. 10/20/1953.
3. Eggert AA. Differential counts, automated. In: Encyclopedia of Medical Devices, Webster, JG editor. John Wiley & Sons Inc.; 1988. pp 944–956.
4. Mansberg HP, Saunders AM, Groner W. The Hemaolog –D white cell differential system. J Histochem Cytochem 1974;22: 711–724.
5. Saunders AM. Development of automation of differential leukocyte counts by use of cytochemistry. Clin Chem 1972;18: 783–788.
6. Gomori G. Histochemical differentiation between esterases. Proc Soc Exp Biol Med 1945;67:4.
7. Gomori G. Chloroacyl esters as histochemical substrates. J Histochem Cytochem 1953;1:469.
8. Cremins et al. Method for the determination of a differential white blood cell count. US patent 4,801,549. 1989 Jan 31.
9. Cremins et al. Leukocyte differentiation method. US Patent 5,518,928. 1996 May 21.
10. Hamaguchi et al. Reagent and Method for Measuring Leukocytes and Hemoglobin in Blood. US Patent 5116539. 1992 May 26.
11. Miller MN, et al. Pattern recognition system for generating hematology profile. US Patent 4307376. 1981 Dec 22.
12. George TC, et al. Distinguishing modes of cell death using the ImageStream multispectral imaging flow cytometer. Cytometry A 2004; Jun: 59(2):237–245.
13. Wang FS, Kubota F. A Novel Apoptosis Research Method with Image-Combined Flow Cytometry and HITC or IR-125 Staining. Cytometry (Clini Cytom) 2002;50:267–274.
14. Marshall PN. (to Abbott Laboratories); Flow Cytometric Lytic Agent and Method Enabling 5-Part Leukocyte Differential Count. US Patent 5,510,267. 1996 Apr 23.
15. Uptmore C, et al. Comparison of the Sysmex XE-2100 to the Abbott Cell-Dyn 4000, Automated Hematology Analyzer. Sysmex J Inter 2001;11(1):22–26.
16. CELL-DYN 4000 System Operation Manual, Revision 3-03.doc, Abbott Laboratories.
17. Pentra 120 SPS User Manual. Section 2: Description and Technology. P/N RAB 106 CA. ABX Horiba Diagnostics.
18. Lefevre et al. (to ABX); Apparatus For Counting And Determining At Least One Leucocytic Sub-Population. US Patent 5,196,346. 1992 Aug 11.
19. Lefevre et al. (to ABX); Reagent And Method Of Using Same For Automatically Counting Basphilic Leukocytes In The Blood In Resistivity Variation Measuring Apparatus. US Patent 5,196,346. 1993 Mar 23.
20. Lefevre et al. (to ABX); Reagent For Use In Automatic Analyzers For Distinguisher Leukocyte Sub-Populations In Blood Samples. US Patent 5,282,857. 1993 Aug 3.

21. Kass L. Staining of Granulocytic Cells by Chlorazol Black E. Am J Clin Pathol 1981;76:810–812.
22. ADVIA 2120 Operator's Guide V1.0.1.00, 2004. Bayer HealthCare LLC, Diagnostics Division, Tarrytown, NY.
23. Harris N, Kunicka J, Kratz A. The ADVIA 2120 Hematology System: Flow-cytometry-based analysis of blood and body fluids in the routine hematology laboratory. Lab Hematol 2005;11(1):47–61.
24. Fernandez T, et al. Performance Evaluation of the Coulter LH 750 Hematology Analyzer. Lab Hematol 2001;7:217–228.
25. Aulesa C, et al. Validation of the Coulter LH 750 in a Hospital Reference Laboratory. Lab Hematol 2003;9:15–28.
26. Coulter VCS Technology: Clinical Case Studies. Beckman Coulter Bulletin No. 3008.
27. Coulter Gen S System Enhanced VCS Technology: Clinical Case Studies. Bulletin 9165.
28. Fujimoto K. Principles of Measurement in Hematology Analyzers Manufactured by Sysmex Corporation. Sysmex J Inter 1999;9(1):31–44.
29. Sakata et al. (to Toa Medical Electronics Co, Ltd.); Reagent And Method For Classifying Leukocytes By Flow Cytometry. US Patent 5,928,949. 1999 July 27.
30. Uchihashi et al. (Sysmex Corporation); Reagent For Measurement of Leukocytes And Hemoglobin Concentration In Blood. US Patent 5,968,832. 1999 Oct 19.

Further Reading

Shibata et al. (to Toa Medical Electronics Co, Ltd.); Method And Apparatus For Determining A Particle Criterion And Particle Analyzer Using The Criterion. US Patent 5,690,105. 1997 Nov 25.

Sakata et al. (to Sysmex Corporation); Method For Classifying And Counting Immature Leukocytes. US Patent 5,958,776. 1999 Sept 28.

Shibata et al. (to Sysmex Corporation); Reagent and Method For Classification And Counting Of Leukocytes. US Patent 6,004,816. 1999 Dec 21.

Ruzicka K, et al. The New Hematology Analyzer Sysmex XE-2100: Performance Evaluation of a Novel White Blood Cell Differential Technology. Arch Pathol Lab Med 2001;125: 391–396.

Walters J, Garrity P. Performance Evaluation of the Sysmex XE-2100 Hematology Analyzer. Lab Hematol 2000;6:83–92.

Briggs C, et al. Performance Evaluation of the Sysmex XE-2100 Automated Haematology Analyzer. Sysmex J Inter 1999;9(2): 113–119.

See also BLOOD COLLECTION AND PROCESSING; CELL COUNTERS, BLOOD; CYTOLOGY, AUTOMATED.

DIFFERENTIAL TRANSFORMERS.

See LINEAR VARIABLE DIFFERENTIAL TRANSFORMERS.

DIGITAL ANGIOGRAPHY

JAMES R. BENNETT
University of Iowa
Iowa City, Iowa

INTRODUCTION

The term "angiography" is derived from the Greek *angeio-*, meaning blood vessel, and *graphein*, meaning representation of a specified object (1). Thus, a very general definition of angiography would be "the representation of blood vessels". Angiography is a technique that allows visualization of any aspect of the human circulatory system. The principle of angiography is to increase the conspicuity of blood vessels during imaging by displacing the blood within the vessels of interest with a contrast medium, although this definition does not necessarily hold true for all imaging modalities. Whether it is the venous system, cardiac arteries, or the abdominal aorta, whenever diagnostic information is needed on a patient's vasculature, some form of angiography is likely to be employed. This article will be structured as follows: a brief history of angiography, the advent of digital subtraction angiography (DSA), noncatheter/noninvasive angiographic techniques, and finally a discussion on the future of vascular imaging.

A BRIEF HISTORY AND OVERVIEW

The first angiography was performed shortly after Wilhelm Roentgen's discovery of the X ray. In January of 1896, Mr. Hascheck and Dr. Lindenthal, of the Physicochemical Institute in Vienna, produced the first angiogram by injecting Teichmann's mixture into the arteries of a cadaver's hand and imaging the hand using X rays, shown in Fig. 1 (2). Born from this experiment was the field of vascular imaging, and for the next 90 years, angiography was the field's principal technique. In this section, the explanation of angiography are described in terms of X-ray imaging for the sake of simplicity. The overall principles of X-ray angiography generally hold true for all modalities. Yet, there are important deviations when utilizing other imaging technology, which are noted and explained in subsequent sections.

The principle behind all radiographic imaging is differential X-ray attenuation. As an X-ray beam passes through an object, the intensity of the beam is attenuated, or diminished, in proportion to the density and thickness of the object. This attenuation can be modeled by the following equation:

$$I = I_0 e^{-\mu t}$$

where I is the X-ray intensity after the original X-ray beam, with intensity I_0, passes through an object with a thickness of t and a linear attenuation coefficient of μ. For this article, assume that beam attenuation is logarithmically proportional to the attenuation coefficient, which in turn is linearly proportional to the material density.

Returning to radiographic imaging, materials with different density attenuate X-rays to differing degrees. Thus, a radiographic image is formed when an X-ray beam passes through an object, and is then captured on the other side by a radiosensitive fluorescent screen. The screen changes based on the X-ray intensity: The more intense the X-ray beam is, the greater the screen fluoresces and visa versa. However, there must be significant differences in material density to produce an image. For example, the femur (bone) is easily identified in an X-ray study of the leg, but differentiation between skin and muscle is nearly impossible. This is due to the fact that bone is significantly denser than the surrounding tissue: Skin and muscle have similar densities, and therefore attenuate X rays in a

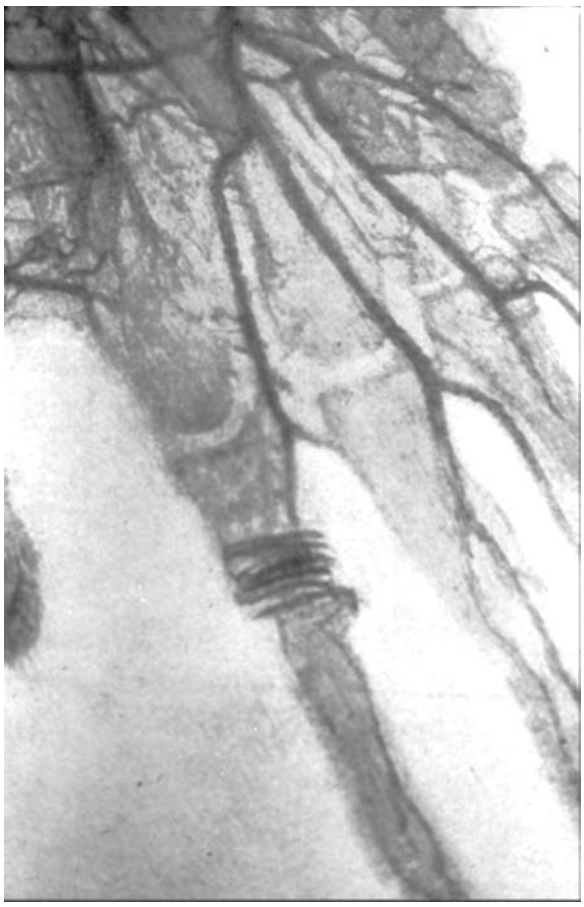


Figure 1. Radiograph of cadaver hand with contrast injection. (Copyrighted © by Radiology Centennial, Inc.)

nearly identical fashion. This attenuation homogeneity is true for many tissues in the body, including blood and blood vessels; hence the need for a dense contrast medium to discriminate the blood vessels from the adjacent tissues.

In Fig. 1, Teichmann's mixture was used as a contrast medium to image the blood vessels of a cadaver's hand. Teichmann's mixture is a dense liquid containing lime, mercury, and petroleum (2). Injection of the mixture displaced the blood within the vessels; as the X-ray beam passes through the hand, the material within the vessels significantly attenuates the beam, and results in the visualization of the vessels in the resultant image. For the next 90 years, the main changes to this technique would be the contrast material and its delivery technique. Iodinated agents have become the main contrast medium for X-ray angiography. Furthermore, catheter guided injections became the predominate means to deliver a contrast injection in a localized region of interest (ROI). This technique utilizes a small puncture to a main artery, where a guide wire is inserted and routed through the arterial system and parked at a location proximal (upstream) of the ROI, where the contrast agent is delivered. This technique will be explored further in the subsequent section.

The field of radiology has arguably incurred the greatest paradigm shift of any medical field resulting from the introduction of practical computing technology. Research in digital imaging began in the late 1960s and early 1970s and clinical applications were realized in the 1980s. Two-dimensional (2D) X-ray imaging was the first digital application on the clinical radiology scene. In a general sense, film was replaced with the digital capture of X-ray images. This change to digital imaging brought vast and immediately benefits, essentially falling into three categories: time, money, and quality. Digital imaging created near instantaneous X-ray images, eliminating the costly and time consuming process of developing film. Relative to angiography, this allowed radiologists to immediately ascertain whether or not the contrast injection was adequately captured during imaging, which was previously delayed by the lengthy film developing process. Postprocessing of the digital images had a revolutionary effect on image quality. Although the resolution of the digital images was about a fivefold decrease relative to film, the ability to digitally manipulate images far outweighed this loss in resolution. Radiologists could instantly adjust the contrast and brightness of an image, yet the most significant gain in the quality of angiographic imaging was the advent of digital subtraction angiography.

DIGITAL SUBTRACTION ANGIOGRAPHY

Digital subtraction angiography (DSA) undoubtedly revolutionized angiography and the field of vascular imaging as a whole. This is currently the gold standard against which all emerging vascular imaging techniques are compared, however, subtraction imaging is not a new technique to the field of radiology. It was first theorized in the 1930s that if a radiograph were taken before a contrast injection and another as the contrast passes through the imaging window, it would be possible to subtract out everything but the contrast within the vessels (3). The goal of subtraction is to remove the nondiagnostic artifacts in the X-ray image; organs, bones, surgical staples, metallic implants, and other fairly dense objects that can overshadow the contrast enhanced vessels as they intersect within the imaging plane. Essentially, when the enhanced vessel crosses a dense object, it becomes indistinguishable from the artifact on the radiographic image and can significantly impede interpretation. Manual film-subtraction angiography proved to be extremely valuable to the field of vascular imaging. Yet this process is expensive and time consuming: a single subtracted image involves complicated film development that usually requires multiple attempts before a usable image is produced.

The key principle behind image subtraction is the acquisition of a preinjection image, or mask image. Essentially, if two images are taken of the same motionless object, and the first image is subtracted from the second, the resultant image should be blank. If, however, there is any change to the object between the two image captures, only the difference will be visible when the first image is subtracted from the second. For example, in Fig. 2a, there is an image of an apple. Figure 2b represents the same apple, taken with the same camera in the same position, but with an

interval of several days in between the two images. Within this time interval, the apple became spotted, while its shape did not change. Each image is a grayscale matrix of 200×200 pixels; each pixel has a grayscale value from 0 to 255. The first image matrix, or mask matrix, will be denoted by M and the second image will be denoted by I. To subtract the two images, it is necessary to subtract each matrix entry in the second image from its corresponding entry in the mask image, giving the subtracted image matrix, S, represented by this formula:

$$S = I - M$$

The resultant image, S, displays only the spots on the apple, while everything else in the image is black. Black represents zero on the grayscale image, thus all points in the two image matrices that did not become zero when the images are subtracted. The next step is to equalize the histogram of the resultant image.

A histogram is a plot representing the relative occurrence of pixel values in an image; the x axis contains the pixel values, 0–255, and the y axis contains the number of times that pixel value occurs in the image. An image that is very dark will have a histogram that is weighted toward the origin, as black would be the predominant pixel value and is represented by zero on grayscale images. Very light images are weighted toward the outer boundary because white is represented by 255. Upon close examination, the subtracted image in Fig. 2c does not exactly match the change in the two previous images. White spots were added to the second image and the subtracted image shows similar spots, but not perfectly white spots. It is necessary to adjust the histogram of the image after subtraction of the two images, due to the fact that subtraction does not result in an image with an equalized histogram. An image with an equalized histogram has pixel values that are spread throughout the histogram. Figure 3a represents the histogram of the original apple image (Fig. 2a). The background of the image is white, and therefore the histogram is weighted toward the outer boundary. However, there is also a fairly equal spread of pixel values throughout the entire range. In comparison, Fig. 3b shows the histogram of the subtracted apple image (Fig. 2c). This histogram weighted heavily toward the origin and does not have pixel values throughout the entire range. The reason for this compressed histogram is beyond the scope of this article, but it is a general rule that the resultant image from subtraction will have a compressed histogram. The next step is to utilize an algorithm that equalizes the histogram by spreading the pixel values out over the entire range. Figure 3d shows the subtracted image after its histogram (Fig. 3c) was equalized. Now, the subtracted image is a true representation of the change in the original image.

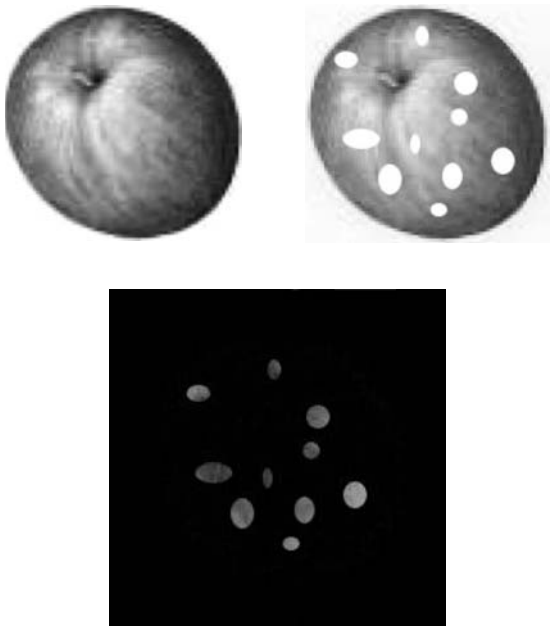


Figure 2. (a) Original picture of apple. (b). Apple after time elapsed and became spotted. (c) Digital subtraction of Fig. 2a from 2b.

Digital subtraction works well in optimal conditions where the camera and object do not move within the interval that the two images are taken. However, real-world implementation presents a host of issues, the most prevalent being patient motion. Imaging equipment is generally very precise in its positioning, that is, it does not deviate from its expected location. Patients tend to shift position in between the contrast and mask image capture, the side effect of which is the introduction of artificial artifacts in the subtracted image. An example of such artifact introduction is given in Fig. 4, where one of the two images has been shifted to the upper right quadrant. In some cases, the patient motion is due to discomfort, as the contrast medium tends to displace oxygenated blood within the vessels that may result in a burning sensation. Another source of patient motion can be organ movement, which can similarly introduce motion. In either case, motion results in image artifacts that can impede diagnostic interpretation. It is possible to compensate for these movements, given that the movement is a linear translation within the imaging plane. Simply shifting the contrast and mask images can realign the objects within the image, which will minimize the induced artifacts. In film-based image subtraction, this shifting can require multiple attempts before a usable image is produced because there is much time involved with developing the subtracted image. However, with digital imaging, it is possible to instantly manipulate the position of the two images. This technique is termed pixel shifting, as one generally shifts the images, pixel by pixel, until the objects are aligned within the images. Unfortunately, a significant amount of patient movement does not occur linearly within the imaging plane. Movement is nonlinear if the patient rolls their body, extends their limb, or any movement that is not translational within the imaging plane (the patient table can be representative of the imaging plane in most cases). A technique has recently been developed to adjust for such nontranslation movement. Although the details are beyond the scope of this article, essentially a map is created of objects both within the mask and contrast images. An algorithm compares the two maps and adjusts, or stretches, the images based on differences within the maps.

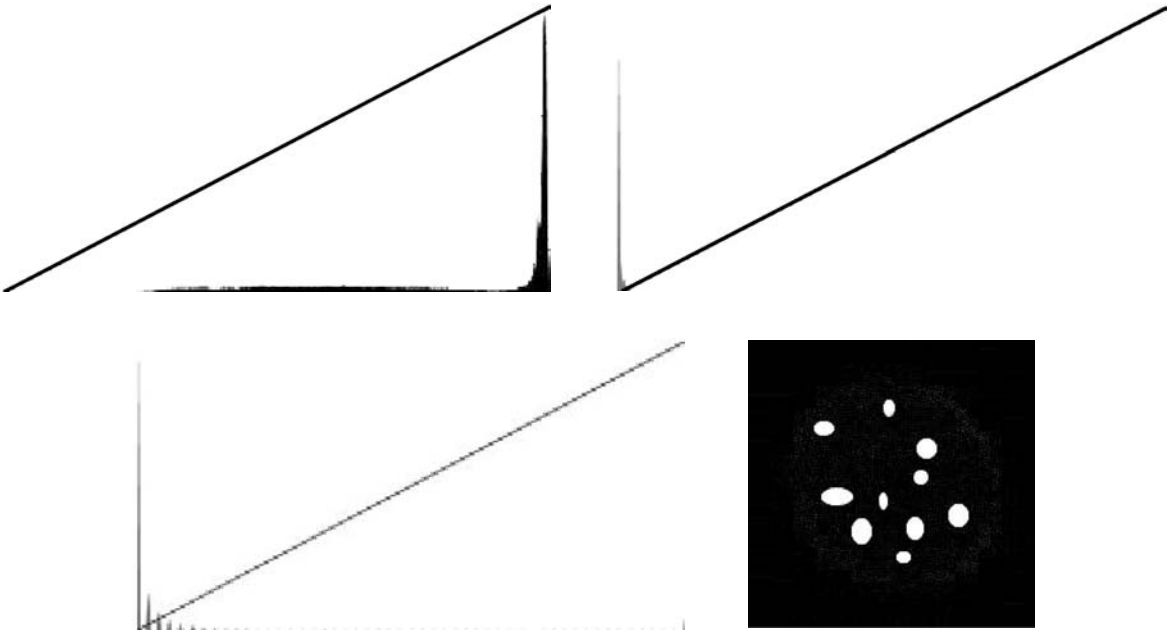


Figure 3. Histogram of Fig. 2a. (b) Histogram of Fig. 2c. (c) Histogram of Fig. 3d. (d) Figure 2c after histogram equalization algorithm applied.

Now that digital subtraction imaging has been covered, a brief explanation of DSA follows. A typical DSA procedure begins with the insertion of a catheter in the femoral artery, as shown in Fig. 5. A guidewire is inserted through the catheter and routed to a position proximal (upstream) to the vessels of interest, through which a contrast delivery sheath is inserted. The next step depends on the purpose of the study: if the clinician solely wishes to view one particular area, for example, the abdominal aorta, they will take a single mask image, inject and image the contrast, and then perform digital subtraction. If, however, the clinician wishes to perform a runoff sequence, where the peripheral vasculature is studied, they will program the imaging equipment to follow a series of imaging stations. These stations are required because the desired image is significantly larger than the actual imaging window. Thus, many frames, or stations, overlap each other and follow the vessels of interest until their termination at the extremities. Once the stations are programmed, the imaging apparatus steps through the programmed sequence and captures a mask image at each station. Next, the machine returns to the first station, and when the clinician is ready to inject the contrast,

Figure 4. Example of digital subtraction when object shifts between original and mask frame.

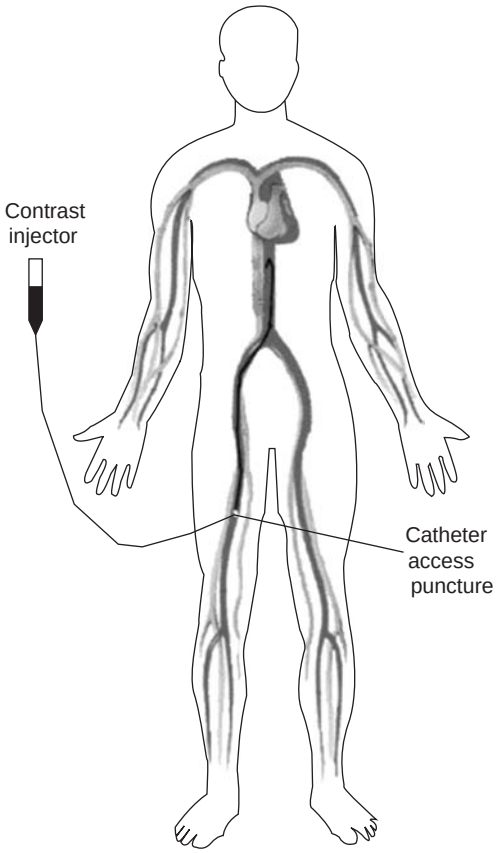


Figure 5. Diagram of DSA contrast delivery technique.

the machine begins imaging the first station at a set capture rate (typically 3–5 frames · s). The clinician injects the contrast and watches the viewing monitor for the contrast arrival at the first station. When the contrast reaches the lower portion of the first station, they triggers the machine to advance to the next station, which is always distal (downstream) to the previous station. Once the contrast passes through the second station, the clinician triggers the machine to the next station, and this process is repeated until the contrast reaches the final station. Digital subtraction is performed at each station, and the final step is merging the subtracted images from each station into a single image of the patient's peripheral vasculature. One of the most recent advances in DSA has come through flat panel technology. In traditional X-ray imaging systems, the X-ray beam is passed through the patient and captured with a scintillator, which is a screen that fluoresces in proportion to X-ray intensity. Typically, a camera captures the luminescence from the screen and transfers this image digitally to a computer. Flat panel technology replaces this system with complementary metal oxide semiconductor (CMOS) detector panels, which can be thought of as radiosensitive charge coupled device (CCD) chips. Essentially, the panel is a matrix of radiosensitive pixels, which monitor the X-ray beam intensity at each pixel. The intensity level from each pixel is converted to a digital signal, which is then reconstructed to form an X-ray image. This system bypasses the need for the scintillator screen and camera setup by capturing the X-ray beam directly. The benefits of flat panel technology are an increase in resolution and elimination of geometric distortion resulting from the nature of the optical camera system. Another advance in DSA technology is three-dimensional (3D) imaging, where the imaging system rotates around the patient table, capturing images at different angles, from which 3D images are reconstructed.

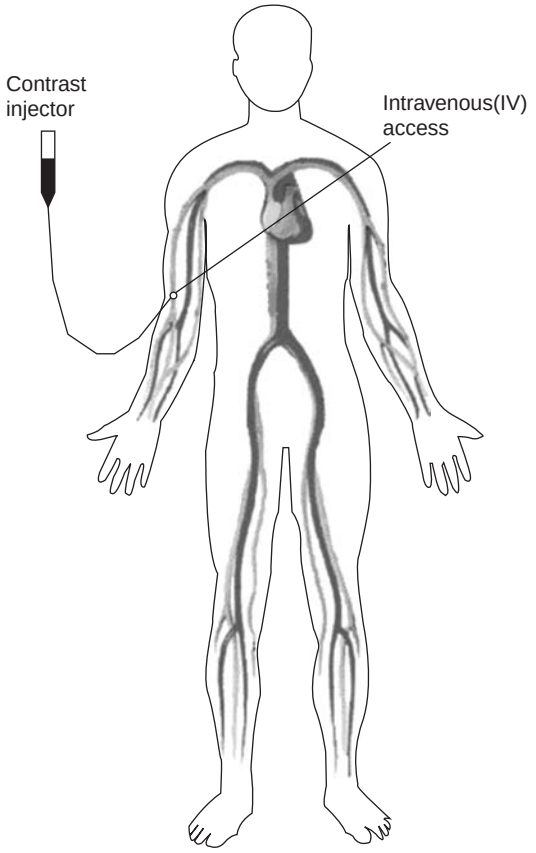


Figure 6. Diagram of computed tomography angiography (CTA) and gadolinium enhanced magnetic resonance angiography (MRA) contrast delivery technique.

NON-CATHETER/NONINVASIVE ANGIOGRAPHY

Non-catheter/noninvasive imaging technologies, especially 3D technologies, have created the next paradigm shift in vascular imaging. The principles of such imaging technologies will not be discussed in this article, as such information can be found in this Encyclopedia within the respective modality articles. This section will begin with single detector-row computed tomography (CT) X-ray imaging, which became commercially available in the 1980s. This technology was able to reconstruct 2D axial slices (images) of the patient. In a similar technique to DSA, iodinated contrast agents were injected into the vasculature during CT imaging, which produced basic 2D contrast enhanced vascular images. This technique could be considered fairly rudimentary as compared to today's standards. Yet, it did allow clinicians to view patient's vasculature in relation to the rest of the body, instead of the enhanced vasculature being superimposed upon the rest of the body, as in DSA. As CT advanced, so did CT technology. Soon, multirow detector arrays were introduced to CT, which revolutionized vascular imaging. Multirow detectors increased longitudinal coverage and image resolution. It allowed for 3D images to be produced, along with increased visualization of small vessels. However, the greatest contribution of multirow detectors was the increased scanning speed. As scan time decreased, conspicuity increased because patient motion, and therefore the resulting artifacts, was less prevalent. Further benefits resulted from the noninvasive nature of this technique. Instead of catheter based contrast delivery in DSA, the iodinated contrast is injected intravenously, as shown in Fig. 6. The contrast travels from the venous system to the right side of the heart, through the pulmonary system, and then back to the left side of the heart where it is pumped throughout the arterial system. The contrast can be imaged wherever diagnostic information of vessels is needed. This technology greatly decreased the time and cost associated with an angiographic study, along with providing superior vascular images. There are several drawbacks associated with CT imaging. It requires an increased amount of radiation and iodinated contrast dose when compared with 2D angiography. The U.S. Food and Drug Administration (FDA) has recently classified ionizing radiation, used to create X-ray images, as a carcinogen. Computed tomography imaging delivers a substantial dose of ionizing radiation, especially in a full-body scan. The iodinated contrast material is nephrotoxic. In cases where a patient has borderline kidney function, the toxicity of a large dose of iodinated contrast may induce renal failure.

Another drawback to CT imaging is that metallic implants, such as hip prostheses, create artifacts in the images. Finally, calcium within the vasculature can limit evaluation of vessel patency (openness).

Magnetic resonance (MR) imaging can also be used to perform diagnostic imaging of blood vessels. To fully comprehend the nature of this modality, and therefore MR angiography, it would be recommended to review the article on MR imaging in this Encyclopedia. There are many methods for imaging the vasculature with this modality: time of flight (tof) angiography, phase-contrast (PC) angiography, and gadolinium enhanced angiography. In this section, tof and gadolinium enhanced MR angiography will be discussed, as they are the most prevalent in clinical use. Time of flight angiography utilizes constant radio frequency (RF) pulses within a presaturation slab. The RF pulses do not allow the atomic spins within that slab to realign with the magnetic field. Thus, these saturated atoms give off a very weak signal. However, when fresh blood from the arteries enters the imaging slice, the RF pulses have not saturated the spins, and therefore it gives off a much stronger signal when passing through the imaging slice. This technique works well for arterial visualization, the only drawback is that this technique does not work when the vessel, and therefore blood flow, is parallel with the imaging slice or if the blood flow is slow or turbulent. This is due to the fact that the blood becomes saturated as it travels within the saturation slab, and therefore does not give a strong signal, unlike the fresh blood (4). Gadolinium enhanced MRA does not have any of these drawbacks associated with tof angiography. Gadolinium enhanced MRA works in a similar fashion as DSA or CTA; however, the physics behind its principles are quite dissimilar. Gadolinium contrast agents work by shortening the $T1$ time of the blood and blood vessels (4). This enhances the blood and blood vessel visualization as the contrast enhanced blood passes through the imaging slice. Recently, there has been an emergence of advanced MRA techniques, such as fresh blood imaging and time-resolved MRA, which are well beyond the scope of this article. Magnetic resonance angiography overcomes many of the drawbacks associated with CTA; however, it is not without its own issues. Patients with pacemakers, surgical clips, metallic prostheses, or foreign bodies cannot undergo an MRA examination. Also, patients with claustrophobia cannot undergo MR imaging. It has a longer scan time as compared to CT imaging, and finally it does not produce images with adequate vessel wall definition.

Ultrasound (US) plays an important role in vascular imaging. Although most applications of this technique would not generally be classified as angiography, it is an extremely important tool nonetheless. The latest wide-scale medical application of ultrasound is termed triplex imaging. Triplex imaging combines grayscale, color, and Doppler information to form images of the vasculature with color coded Doppler blood flow data overlaid on the grayscale image. One of many recent developments in US imaging is intravascular ultrasound (IVUS). This technique works by inserting an extremely small US transducer into the blood vessel of interest via a catheter. The IVUS produces the most accurate vessel wall characterization of any currently available modality. Another recently developed technique involves injecting microbubbles into the bloodstream. This US technique would be the most likely to be classified as angiography. The principle behind microbubble US is that these tiny bubbles will create differences in density in the blood, which can be easily detected by a US transducer. Both IVUS and the microbubble technique are currently in their infantile stages, but have vast potential to improve vascular imaging. There are many advantages utilizing US as a vascular imaging modality, including the low cost and portability relative to MR and CT scanners. Also, vascular US imaging does not involve any harmful contrast medium or ionizing radiation.

FUTURE OF VASCULAR IMAGING

The need for vascular imaging will not diminish in the foreseeable future. The United States alone has seen obesity and diabetes climb to a near epidemic scale, factors that can significantly increase the incidence of circulatory disease. According to the World Health Organization, in 1997 there were 15.3 million deaths due to circulatory disease worldwide. It is unknown whether these deaths could have been prevented with medical intervention, but the main point is that circulatory disease will not be subsiding in the near future. Thus, it is probable that there will be an increasing need for diagnostic vascular imaging. There are many promising emergent modalities and techniques that will be included in the next generation of vascular imaging. Improvements such as 3D dimensional ultrasound and 256-row CT scanners are not far from implementation. One improvement that could have a vast impact on vascular imaging would be a nonnephrotoxic radioopaque contrast agent. The optimal solution would be a modality that combines the resolution and detail provided with CTA with the relatively risk-free imaging found in ultrasound and MRA.

BIBLIOGRAPHY

1. Merriam-Webster Online Dictionary. (2005). Merriam Webster Online. [Online]. Merriam-Webster. Available at http://www. merriam-webster.com Accessed 2005 April 18.
2. Sprawls P. 1996 Feb 1. The X-ray Century. [Online]. Emory University. Available at http://www.emory.edu/X-RAYS/century.htm. Accessed 2005; May 3.
3. Ziedses des Plantes B. Plantinigraphie en subtractie Roentgenographische differentiatiemethoden. Ph. D. dissertation, University of Utrecht; 1934.
4. Bakal CW, Silberzweig JE, Cynamon J, Sprayregen S. Vascular and Interventional Radiology: Principles and Practice. New York: Thieme Medical Publishers; 2002.

Reading List

Hagspiel KD, Matsumoto AH. The Radiological Clinics of North America. Philadelphia: W. B. Saunders, 2002.
Gonzalez RC, Woods RE. Digital Image Processing. Upper Saddle River (NJ): Prentice Hall; 2002.
World Health Organization. The World Health Report 1998. Geneva, Switzerland: World Health Organization; 1998.

DIVING PHYSIOLOGY.

See HYPERBARIC MEDICINE.

DNA SEQUENCING

SOTIRIOS A. TSAFTARIS
AGGELOS K. KATSAGGELOS
Northwestern University
Evanston, Illinois

INTRODUCTION

The DNA molecule is one of the most important molecular structures in our planet. As an information carrier molecule it is used to encode the role and function of proteins that are later used to create complex organic structures (e.g., human cells).

Information is encoded using four nucleotides adenine, guanine, thymine, and cytosine, abbreviated, respectively, as A, G, T, and C. Nucleotides are joined together to form sequences, which encode certain functions. These sequences are translated into proteins, which dictate certain actions. It is therefore critical when examining those sequences to directly determine the exact sequence of nucleotides. Such an effort has been more publicly acknowledged with the Human Genome Project. The goal of this national effort was to extract deoxyribonucleic acid (DNA) sequences from human cells and decode the exact nucleotide sequence.

This process is called DNA sequencing and is commonly used in major research laboratories. There are many commercially available automated machines that can sequence DNA and output in a human readable format, usually through a computer, the exact nucleotide sequence of DNA.

Understanding how DNA sequencing works is the objective of this article, which is organized as follows. First, a very short introduction into the chemistry of the DNA molecule is provided. Subsequently, some of the basic principles used by common sequencing techniques are presented. This analysis is followed by a presentation of the most commonly available techniques and equipment. This article conclude with the presention of some of the most promising techniques for DNA sequencing in the future.

THE DNA MOLECULE

A double helix of DNA is made from two single strands of DNA, each of which is a chain of nucleotides (1). A nucleotide is an organic molecule made up of three basic parts: a phosphate group, a five-carbon sugar group, and a nitrogenous side group, which is more commonly called a base. Four different nucleotides occur in DNA: adenine, guanine, thymine, and cytosine. Nucleotides can be joined together in a linear chain to form a single strand of DNA.

A short single strand of DNA consisting of up to 100 or so nucleotides is called an oligonucleotide or oligo. It has a backbone of alternating sugar and phosphate groups with one of the four bases bound to each sugar group. The backbone gives an oligonucleotide a polarity, that is, it has two distinct ends, the 5′ and the 3′ end.

The chemical structure of the bases allows for the unique pairing between A-T (double hydrogen bond) and G-C (triple hydrogen bond). Each base in DNA has its unique Watson–Crick complement, which is formed by replacing every A with a T and vice versa, and every G with a C, and vice versa. Every oligonucleotide has a complementary sequence with opposite polarity (e.g., the complementary sequence of 5′-ATG-3′ is 3′-TAC-5′).

If two complementary sequences meet in a solution under appropriate conditions (temperature, pH, sequence length), they will attract each other and form a double-stranded structure. This process is called hybridization or annealing. Through hydrogen bonds and Van der Waals forces, these pairings are the basis for the exquisite molecular recognition, which allows DNA to act as an information-carrying molecule. There are two types of hybridization: (*1*) specific hybridization, which refers to cases where the two single strands are perfectly complementary at every position and the double-stranded molecule that is formed is perfect; and (*2*) nonspecific hybridization, for which the sequence may not be completely complementary, and thus it may contain mismatched base pairs. DNA melting or denaturation is the opposite of hybridization. When the temperature is raised, the chemical bonds break and the duplex breaks into the two single-stranded parts.

Ligation is the process of joining together double-stranded DNA with compatible sticky ends with the use of DNA ligase. A double-stranded DNA molecule can either have blunt ends or it can have single-stranded overhanging ends (called sticky ends) at one or both of its extremities. The enzyme DNA ligase, joins together, or ligates, the end of a DNA molecule to another molecule.

Restriction enzymes (endonucleases), recognize a specific short sequence of DNA, known as a restriction site and cut any double-stranded DNA at that location. Using enzymes called exonucleases, either double- or single-stranded DNA molecules may be selectively degraded from the ends in.

GEL ELECTROPHORESIS GENERICS

Gel electrophoresis is a technique used for the separation of nucleic acids and proteins (1). Separation of large (macro) molecules depends on two elements: charge and mass. When a biological sample (e.g., proteins or DNA) is mixed in a buffer solution and applied to a gel, these two factors act together. The electrical current from one electrode repels the molecules while the other electrode simultaneously attracts the molecules. The frictional force of the gel material acts as a molecular sieve, separating the molecules by size. During electrophoresis, macromolecules are forced to move through the pores when the electrical current is applied. Their rate of migration through the electric field depends on the strength of the field, size, and shape of the molecules, the relative hydrophobicity of the samples, and the ionic strength and temperature of the buffer in which the molecules are moving. After staining, the separated macromolecules in each lane can be seen in a series of bands spread from one end of the gel to the other, as seen, for example, in Fig. 1.

Some of the concepts of gel electrophoresis are used even in some of the most advanced commercially available

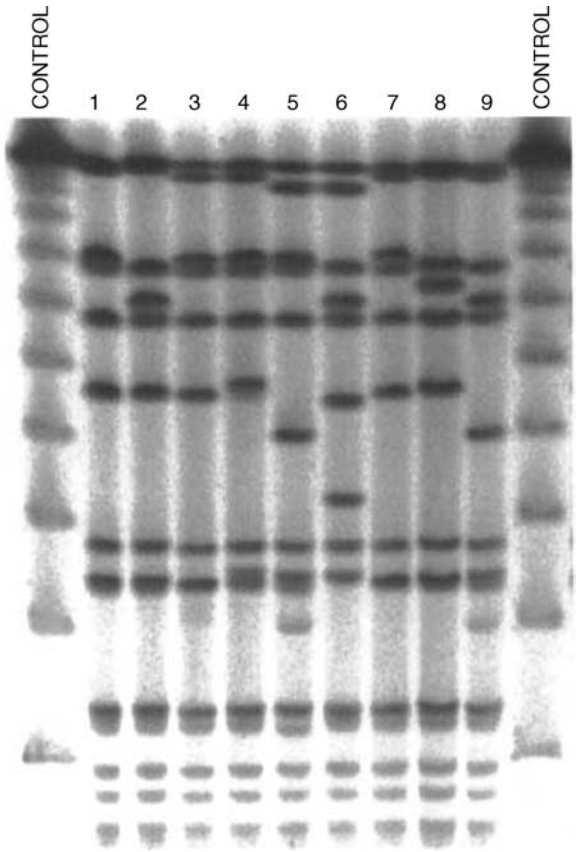


Figure 1. A photograph of a gel from an electrophoresis experiment. Spots higher on columns represent lighter molecules. The lanes at the left and right end are control lanes where the DNA used had known length.

techniques for DNA sequencing. The role of gel electrophoresis will be made clear in describing later Sanger's method.

TYPES OF DNA SEQUENCING

The inherent meaning of the word sequencing translates into finding the sequence of nucleotides of an unknown DNA strand. This type of DNA sequencing is usually referred to as de novo sequencing (de novo in Latin means from the beginning).

On the other hand DNA detection refers to the process of identifying known sequences of DNA within a sample. There are many laboratory techniques commonly used to perform such a task. Polymerase Chain Reaction (PCR) is used in amplifying (multiplying the concentration of) DNA sequences that contain certain primers (1). DNA microarrays is a technology used in gene expression profiling; it is a high throughput DNA detection mechanism where multiple DNA probes are simultaneously detected. These microarrays will be further analyzed in the following paragraphs. Biotin–streptavidin bead-based detection is a process that permits single-stranded DNA molecules containing a given subsequence to be filtered out from a heterogeneous pool of other DNA molecules (1). Strands complementary to the subsequence are attached with biotin to streptavidin coated magnetic beads. The heterogeneous solution is passed over the beads and strands containing the subsequence anneal to the complementary sequence and are retained, while strands not containing it, pass through.

In many cases the DNA sequence under examination is largely known, but only small regions are of interest. This is the case in genotyping when it is desired to detect small variations in a whole genome (or gene) when compared with a known DNA sequence. Much of the variation in organisms originates from single-base changes in genes. These small changes can significantly affect the translation, and hence the role of the gene. This type of variation termed single nucleotide polymorphism (SNP, pronounced "snips") is of extreme interest in molecular biology. When a genome is examined for certain SNPs usually it is desired to detect subsequences of the form xxx. . .xYxxx. . .x, where xxx. . .x indicates known DNA bases and Y can be any base of A,T, G, or C. For a variation to be considered a SNP, it must occur in at least 1% of the population. The SNPs, which make up ~90% of all human genetic variation, occur every 100–300 bases along the 3-billion-base human genome. There exist >100 techniques for detecting known forms of SNPs. Many SNPs have no effect on cell function, but scientists believe others could predispose people to disease or influence their response to a drug. For more information on SNPs their significance and detection methods interested readers are directed to (3).

DNA SEQUENCING PRINCIPLES: THE SANGER METHOD

The foundations of DNA sequencing were laid in 1974. Two groups, a British headed by Sanger et al. (4) and an American lead by Maxam and Gilbert (5), independently discovered a technique that enables to break a fragment of DNA into smaller nested subfragments. Both groups shared the 1980 Nobel Prize in chemistry for their discovery. The method from the American team was based on a chemical cleavage protocol and used toxic chemicals and large amounts of radioactivity, whereas Sanger's method essentially mimics DNA replication as it takes place in cells. Sanger's method was eventually adopted by the industry and a form of it is still used today since it was simpler to implement in large-scale production sequencing.

For the Sanger method the following items are needed: (*1*) the unknown DNA; (*2*) a primer; (*3*) DNA polymerase; (*4*) a mixture of dNTPs (deoxynucleotide triphospates) and ddNTPs (di-deoxynucleotide triphosphates).

The unknown DNA, termed here template, is the fragment of DNA that needs to be sequenced. The fragment needs to be in a single-stranded form in the 3′-5′ direction. If in double-stranded form a single-stranded sequence can be obtained by melting (denaturing) the duplex. We also assume that the unknown template contains a known subsequence usually ~12–24 bases long. The complement of this subsequence in the 5′-3′ direction is called the primer. The primer is chemically synthesized. Once the primer is inserted in the solution containing the unknown DNA it will anneal (bind) to its complementary sequence

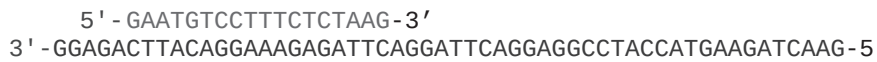


Figure 2. The primer (in red) is annealed to the template at the primer binding site.

with hydrogen bonds (Fig. 2). The primer needs to be long enough to ensure that the annealing site is unique, but not very long such that the annealing is unstable.

Once the primer and the template have annealed the DNA polymerase starts reacting and catalyzes the DNA and extending from the 3′ end of the primer starts filling-in nucleotides (dNTPs) that are complementary to the template at each position. This serial addition of nucleotides is dependent on the bases of the template. The incoming nucleotide forms a covalent bond with the 3′ end of the previous sugar using its 5′ end. Under normal conditions, nucleotides are filled-in till the end of the template is reached, that is, the strand is fully extended. Sanger's idea was to modify this process such that it ceases before it reaches the end of the template.

By using a simple chemical modification, nucleotides can be transformed such that they prohibit the addition of another nucleotide in their 3′ end. The necessary chemical alteration is the substitution of the hydroxyl (OH) group on the 3′ end of the nucleotide with a hydrogen (H). Such modified nucleotides are called ddNTPs or dideoxynucleotide triphosphates and are usually termed as terminators. When such terminators are incorporated in the extension the replication stops, as shown in Fig. 3.

The DNA polymerase will stop extending when a ddNTP is incorporated. Now, if in the solution a certain mixture of dNTPs and ddNTPs is present at the end, DNA polymerase will create a mixture of strands of various lengths terminated by ddNTPs. To distinguish between the different ddNTPs (A,T,G, or C) a unique fluorescent label is attached to each one of them. In some implementations the label is attached to the primer, but assumes that the reaction is run in parallel in four tubes where each tube contains only one type of ddNTP. The relative concentrations of the dNTPs and ddNTPs are adjusted in such a way that we end up with about the same number of copies of fragments between 100 bp and 500 bp long, and a smaller number of shorter and longer fragments.

At the end the solution contains a random mixture of partially terminated double-stranded sequences. If the ddNTPs were labeled, and hence only one test tube was used, the sequences are denatured into single-stranded DNA molecules and are run on a polyacrylamide–urea gel in a single lane. The gel is dried onto chromatography paper (to reduce its thickness and keep it from cracking) and exposed to X-ray film. Since the template strand is not radioactively labeled, it does not generate a band on the X-ray film.

The fragments will be ordered on the gel lane according to length. A laser (for stimulating the emission of radiation) and a detector (for collecting the stimulated radiation) are placed at a certain distance away from the initial position. When a fragment is scanned by the laser, the fluorescent label attached to the terminator is excited, and a signal at a certain wavelength depending on the label will be emitted and sensed by the detector, as shown in Fig. 4. Multiple copies of each fragment will ensure high signal strength, which will hopefully be strong enough to be detected. By examining the peaks of the time sequence of the fluorescence intensity at different wavelengths, the bases of the unknown sequence can be determined.

The above procedure is similar if the label was attached on the primer or in the dNTPs and four tubes and separate reactions were run. In this case, the results of each tube correspond to a specific ddNTP. Each tube's contents are placed in a different lane (four in total) as seen in Fig. 5. With this setup only one fluorescent label is used, hence the excitation and detection mechanism is much more simplified. By examining the peaks of the intensity at each lane and working from bottom to top the base path or the "base ladder" can be determined.

It is evident that the resolution of the gel electrophoresis is rather critical and a single base resolution is usually a prerequisite. To improve the resolution of gel assays, the gels must be much large so that the molecules migrate further and are better resolved. They must contain a high concentration of urea (7–8 m) to prevent folding of the

Figure 3. A mixture of the products of synthesis for the G ddNTP reaction.

Figure 4. An example of a gel where labels are attached on the ddNTPs, and hence a single-lane gel is only used.

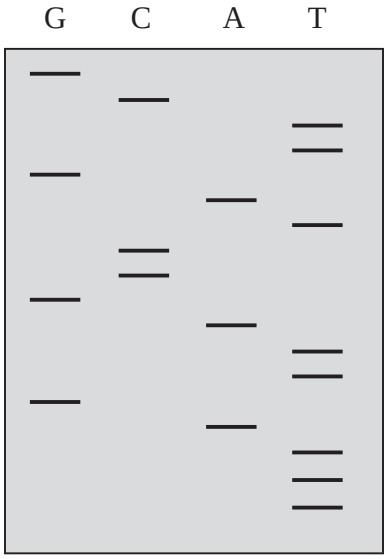


Figure 5. An example of a gel where labels are attached on the primer, and hence four lanes are used.

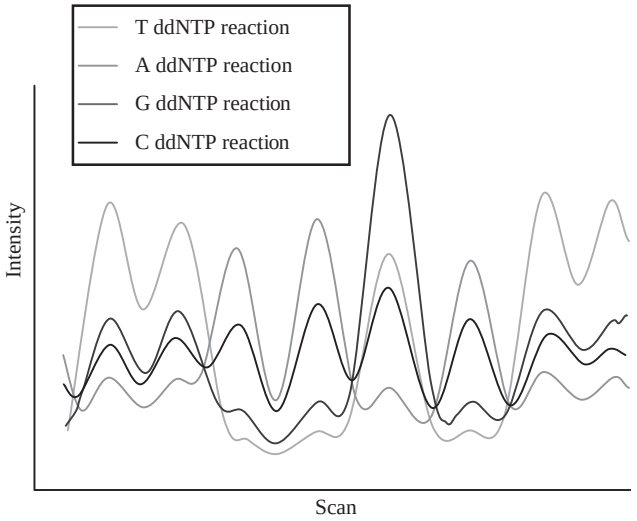


Figure 6. A chromatogram example from an experiment with four dyes. The curves correspond to intensity measurements of fluorescent emission at different wavelengths corresponding to the dyes used for each ddNTP reaction.

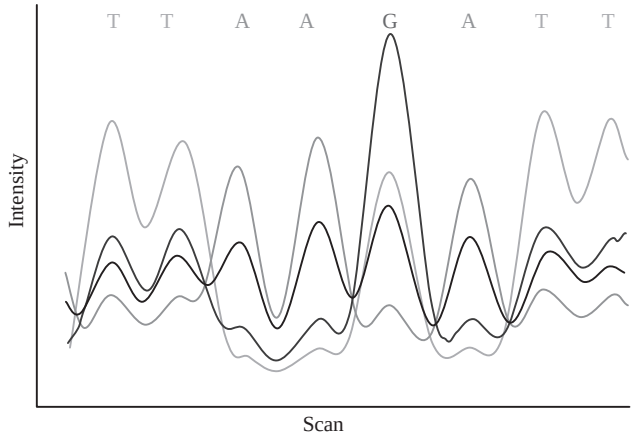


Figure 7. A decoded sequence from the chromatogram example of Fig. 6.

molecules and formation of DNA secondary structures by hydrogen bonding that would alter the mobility of the molecule. Similarly, the samples are denatured before they are loaded. The gels must run at higher temperature (∼50 °C), to prevent hydrogen- bond formation.

From the above analysis, it is clear that in order to extract the bases from the gel reactions the peaks have to be identified. Traditionally the laser scanners output the fluorescence intensity into chromatograms (Fig. 6). Prior to the development of computers the chromatograms were interpreted by humans. The peaks were assigned to bases in a procedure known as base calling (Fig. 7). Ideally a periodic peak detection scheme would have been adequate if the signal was noiseless and perfect. There are certain errors and limitations that make the base calling aspect of DNA sequencing a rather challenging task. Some of the sources of error are

1. Errors in fragment formation: (a) Abnormalities in primer extension (false stops, terminator is not incorporated, or conversely, several terminators accumulate at the same position). (b) Poor choice of relative concentrations of ddNTPs and dNTPs resulting in too many short or long fragments. (c) While DNA moves down the gel, secondary (e.g., hairpin) structures may form and change the mobility properties of the DNA fragments.
2. Convolution: Due to the stochastic nature of the DNA migration in the gel, the time scale of the chromatograms changes resulting in more elongated and less discreet peaks.

Table 1. Common Parameters Used when Evaluating DNA Sequencing Equipment

Parameter	Explanation
Technology used	Electrophoretic or nonelectrophoretic, capillary, and so on affects many of the other parameters in a DNA sequencing system
Length of the gel	Applies only to electrophoretic systems and refers to the length of the gel material. The longer the better since it increases resolution
Throughput	Measured in bases per cycle (or per day), illustrates the processing capability of the system
Cycle speed	The equipment may need replenishing of reagents after a run. The number of runs per day define the cycle speed
Read length	Maximum length of the DNA template that can be sequenced
Capacity	Number of different DNA templates that can be sequenced simultaneously
Sample volume	Defined as the volume of the template needed for a certain outcome quality (the less the better)
Error rates	Number of bases in error out of 1000 usually defines the error rate. Error rates are tightly bound to base calling quality assessment
Maintenance and operation cost	Number of dyes and labeling method used, have a direct impact on maintenance and operation cost

3. Intensity cross-talk: Due to the overlapping of the fluorescent response spectra of the fluorophores employed in the four-dye sequencing strategy there is a need for a transformation to recover the relative concentrations of the four dyes from the fluorescence intensities measured at four different wavelengths.
4. Measurement errors: White noise can originate from several sources, including background, detector, and other noise from the operating environment. Another type of noise encountered is low frequency variation due to slow changes in the background light level during collection. Such variations may be caused by deformation of the gel due to heating, the formation of bubbles in the path of the laser, variations in laser output power and other systematic changes in the environment.

Nowadays advances in statistics, signal processing, electronics, laser optics and software have lead to automated DNA sequencing and base calling capable of sequencing many different DNA templates.

EVALUATING DNA SEQUENCING TECHNIQUES

When deciding on DNA sequencing equipment, a prospective buyer has to evaluate certain aspects of the DNA sequencing scheme offered by the vendor. The buyer has to consider the traits shown in Table 1. All these parameters are critical, but their importance is weighted differently according to the application sought after by the buyer.

CURRENT COMMERCIAL STATE OF THE ART

Since the development of the early DNA sequencing methods the capabilities of the DNA sequencing equipment have improved dramatically. This change can be attributed to the radical advances in the fields of DNA chemistry, laser and optics, statistics, robotics, automation, and software. In many of the laboratories involved in the human genome project, the high throughput DNA sequencing machines that were employed used robotic arms to move samples in and out of the machines and heavy automation to perform those tasks with minimal human intervention. Advances in laser optics led to even finer scanning and detection resolution with lower error rates. As seen in the previous section, one of the most critical aspect in DNA sequencing is the analysis of the chromatograms to determine the bases. Nowadays this task is performed by sophisticated software packages that employ statistics, digital signal processing, and adaptive algorithms that can identify the bases from the fluorescence graphs. A comparison of some of the most commonly used packages can be found in Table 2.

In the following section, sequencing devices are first presented that rely on electrophoretic principles followed by those that do not.

Electrophoretic-Based Methods

Slab-Gel. It was expected that the first automated DNA sequencers would be based on the Sanger method. Acrylamide slab gel electrophoresis until recently was the most widespread method of *de novo* sequencing (6). The Prism 373 by Applied Biosystems (ABI) Prism (Foster City, CA) was the first sequencer that could scan and detect such gels using a procedure very similar to the one described in the previous sections. Some of the drawbacks of slab gel instruments are gel casting (preparing the gel), gel loading (loading the gel into the device), and lane tracking (detecting lanes on the gel).

The Prism 373 underwent many changes in order to increase throughput and read length before it was replaced by ABI PRISM 377. The PRISM 377 is based on a four dye chemistry coupled with a CCD (charged couple device) imaging detector and can process up to 96 samples per cycle (9–11 h) with read lengths of 650–750 bases. Despite their drawbacks, slab gel systems are still preferred for applications with low throughput requirements but large read lengths. Reviews of experimental and commercial systems based on slab gels can be found in (7,8). Of such systems the following are worth noting since they are still used due to their unique properties.

Table 2. A Comparison of Base Calling and Sequence Analysis Software[a]

Name	Publisher	License	Short Description
Phred	University of Washington, Phil Green Laboratory	Free, Open Source	The phred software reads DNA sequencing trace files, calls bases, and assigns a quality value to each called base (1,3,4)
Autoseq	Reece Hart	Free, Open Source	Autoseq is a small package of base-calling software for ABI automated DNA sequencers (5)
Sequence Analyzer	GE Healthcare	Commercial	Usually bundled with MegaBASE sequencers (6)
Lasergene	DNAstar	Commercial	Comprehensive suite of easy-to-use sequence analysis software (7)
Sequencher	Gene Codes Corp	Commercial	Allows SNP detection (8)
Staden	R. Staden and other contributors	Free, Open Source	A suite of sequence assembly, editing, etc. (9)
Sequencing Analysis Software	Applied Biosystems	Commercial	Usually accompanies ABI Prism sequencers (10)
TraceTuner	Paracel	Discontinued	Another base calling application

The DNA 4300 System by LI-COR Inc (Lincoln, NE) uses two dyes at near-(IR) frequencies. Based on this innovation, the ability to operate and sequence from both ends of the template in parallel and coupled with an excellent software suite, the higher end version of the 4300 has read length of up to 1250 bases thus making it ideal for applications with large read length requirements.

The BaseStation from MJ Research Inc (Waltham, MA) uses a 75 μm thick polyacrylamide gel to improve heat dissipation in the gel thus reducing significantly the run time. Armed with robotic gel loading, a four-color photomultiplier with high sensitivity and a 100 sample capacity, the instrument offers a nice alternative to capillary systems when long read lengths are needed.

Capillary Systems. The persistent drawbacks of slab gel electrophoresis and the desire for faster sequencing runs and higher throughput led to the development of capillary array electrophoresis (CAE). Electrophoresis works in a way similar to slab gels, except that each capillary contains a single sample, and therefore tracking problems are eliminated. Furthermore, the high surface/volume ratio of a capillary allows for more rapid heat dissipation than is possible in slab gels, thus allowing higher operating voltages and faster run times.

Capillary electrophoresis uses capillaries usually 50 μm in diameter. Capillaries are very narrow tubes that based on the capillary action can draw liquid against gravity. Similarly to the technique used for manufacturing fiber optics, the capillaries are made from highly pure fused silica.

As seen in Fig. 8 the instrumentation is rather simple. The sample is injected into the capillary and high electrical field is applied to advance the sample into the capillary. Subsequently, the sample is replaced by a buffer solution and the field is reapplied to migrate the samples through the capillary. Since the capillary is filled with a sieving medium it allows for the separation of the DNA sequences according to length. The fragments pass through a laser-induced fluorescence detector, which can record at different frequencies the fluorescence response of the four dyes similar to the slab gel techniques.

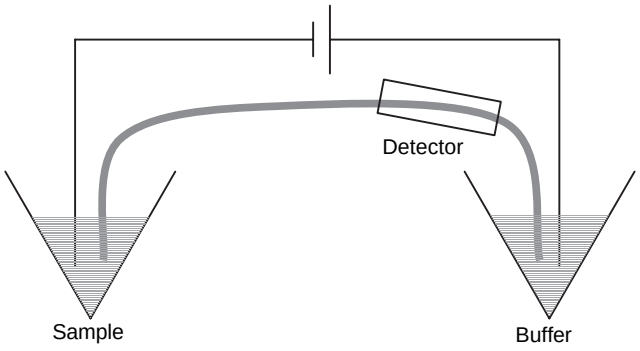


Figure 8. A single capillary electrophoresis device where a fused-silica capillary is used for the separation. The left end of the capillary is submerged into the sample solution while the other is dipped into a buffer-filled holding tank. At some point the capillary goes through a detection apparatus. High voltage is applied at each end using platinum electrodes.

High voltage (double compared to slab gel techniques) allows for rapid separation, but certain phenomena limit the resolution at high read lengths (9). Although the ability to use high voltages is rather attractive, the most interesting aspect of capillaries is their flexibility, allowing them to be incorporated easily into automated systems.

Capillary array electrophoresis uses a collection of capillaries each of which is injected with a different sample. In some instruments the detector sequentially moves from capillary to capillary while in most advanced ones each capillary is scanned simultaneously using detectors attached to each one (10,20–22).

Some of the sequencers currently available in the market are Applied Biosystems PRISM 310 and 3730, Hitachi, BioRad/MJ Research (Hercules, CA) BaseStation, GE

Healthcare (Piscataway, NJ) MegaBASE 4000, Beckman Coulter (Fullerton, CA) CEQ 8000, SpectruMedix (State College, PA) Aurora, and RIKEN (Tsukuba and Wako, Japan) RISA.

Nonelectrophoretic-Based Methods

During the last 20 years, several techniques for sequencing have been discovered that do not rely on electrophoretic principles. Although these techniques have now reached the performance of CAEs for de novo sequencing they are mostly used for other types of sequencing described in a previous section.

Pyrosequencing. Pyrosequencing is a sequencing technique developed around real-time monitoring of the release of pyrophosphate (PPi) during polymerase assisted DNA synthesis (23). Similarly to electrophoretic methods a sequencing primer is hybridized to a single-stranded DNA template, and incubated with the enzymes, DNA polymerase, ATP sulfurylase, luciferase, and apyrase, and the substrates, adenosine 5′-phosphosulfate (APS) and luciferin. One of the four dNTPs is added to the solution. Assisted by polymerase the correct dNTP will be incorporated in the chain resulting in the release of PPi at a concentration analogous to the amount of incorporated nucleotides. The amount of PPi is constantly monitored by a coupled enzymatic reaction where PPi is converted to ATP by ATP sulfurylase. The ATP subsequently assists in the conversion of luciferin to oxyluciferin by firefly luciferase, which results in light emission. The process is repeated iteratively for the other dNTPs. A critical component for the success of the method is the removal of excess dNTP and ATP prior to a new dNTP addition. These can be achieved by attaching the template sequence on solid support that is washed prior to a new dNTP addition or by solution enzymatic reaction where apyrase is added to catalyze the remaining dNTPs.

The read length of pyrosequencing is smaller when compared to electrophoretic methods thus making pyrosequencing less advantageous for de novo sequencing. This is not true, however, for applications, such as genotyping of SNPs, resequencing, tag sequencing, microbial typing, and many others where pyrosequencing shines.

The technique, although in its infancy, can claim the first automated sequencer, the PSQ HS 96 from Pyrosequencing/Biotage (Uppsala, Sweden). It uses a disposable inkjet cartridge for precise delivery of small volume (200 nL) of six different reagents into a temperature-controlled microtiter plate and is widely used for SNP detection with a throughput of 96 samples per hour.

Sequencing by Hybridization: DNA Arrays, Microarrays. Hybridization arrays or microarrays or DNA arrays were originally developed for *de novo* sequencing. Sequencing by Hybridization (SBH) requires annealing a labeled unknown DNA fragment to a complete array of short oligonucleotides (e.g., all 65,336 combinations of 8-mers) and decoding the unknown sequence from the annealing pattern (24). The array could be imaged using laser scanners and CCD devices or photomultiplier tubes. The computational complexity of decoding the annealed pattern limited the popularity of such systems for de novo sequencing. Nowadays, the key applications of DNA arrays are SNP and expression analysis (25).

The DNA microarrays are small, solid supports onto which the sequences from thousands of different genes are immobilized, or attached, at fixed locations. The supports themselves are usually glass microscope slides, of various sizes, but can also be silicon chips or nylon membranes. The DNA can be printed, spotted, or actually synthesized directly onto the support. It is important that the gene sequences in a microarray are attached to their support in an orderly or fixed way, because the location of each spot in the array identifies a particular gene sequence. The spots themselves can be DNA, cDNA, or oligonucleotides. In each microarray experiment two samples are tested simultaneously, each labeled with a different fluorescent dye. The control sample is labeled by the Cy5 dye and the test sample by the Cy3. Both samples are introduced in the microarray simultaneously and when excited by different laser frequencies each dye returns a distinct response, which can be recorded as intensity measurements. This process produces two images; the green image, which corresponds to intensity measurements of the Cy5 dye, and the red image, which corresponds to intensity measurements of the Cy3. These fluorescence intensities correspond to the levels of hybridization of the two samples to the DNA sequences spotted on the slide. An example of a microarray image is shown in Fig. 9.

The list of manufacturers of DNA arrays is rather exhaustive with >30 entries. Of those, the pioneer and first in market Affymetrix (Santa Clara, CA), Agilent (Palo Alto, CA), and Nimblegen (Madison, WI) should be mentioned. The war on density and throughput of microarrays is everlasting. Nimblegen, for example, can produce microarrays with >40,000 genes, with each spot being ~80 μm in diameter.

GENOME SEQUENCING

As of now, the maximum read length permitted by today's commercial and experimental techniques does not exceed 2000 bases. The human genome is currently estimated to be >3 billion bases with 20,000–25,000 genes, while organisms such as the *Escheuchia coli* bacterium has ~4.6 million bases. It is clear that in order to sequence whole genomes of organisms with the currently available techniques a method for combining sequencing results of smaller reads is needed (26).

Most of the techniques rely on shotgun sequencing, which is based on the idea of sequencing overlapping

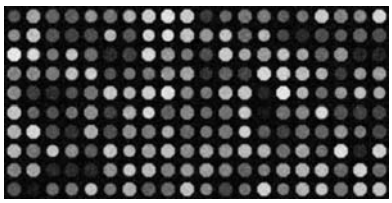


Figure 9. An example of a microarray image.

fragments of DNA (27). The genomic segment is sheared into overlapping fragments of DNA ~ 500 bases long and each fragment is then sequenced. The fragments are assembled into continuous sequences called contigs using complicated computer algorithms where they examine the overlapping sequences and try to order the fragments. There are two issues with this technique: gaps and errors. Since the shearing technique is random, and to avoid laboratory errors usually multiple shearing experiments are performed and the fragments are sequenced to assemble contigs. Another source of errors are repeats. If sequences appear in multiple positions throughout the genomic segment it will lead to errors when the fragments are overlapped. In some cases during contig assembly nonoverlapping sequences are formed that create gaps. Gaps are resolved with directed sequencing experiments with primers derived from the contigs that surround the gap.

The difference between the Human Genome Project (HGP) and Celera Corp was the origin of the target sequence. Human Genome Project used a directed sequencing method (also seen as hierarchical shotgun sequencing) for which the whole genome is first broken into long fragments. The fragments are then mapped into the genome, which is equivalent to finding their order (location) within the genome. Each fragment then is sequenced using shotgun sequencing. The advantage of this approach is the relatively easy assembly, while the disadvantages are the difficulty of building the library, mapping the long fragments, and the need for redundant sequencing.

Celera Corp relied on whole genome shotgun sequencing, which is essentially shotgun sequencing applied directly on the whole genome. The challenge is to assemble the whole genome from small 500 base fragments. While this technique overcomes the shortcomings of hierarchical sequencing and is faster and less expensive, the assembly is rather complicated and resolving the repeats requires sequencing of many clones.

THE FUTURE OF DNA SEQUENCING

Microscale Systems

Micro capillary Systems are microfabricated systems that in principle work similarly to CAE systems. Such systems have the potential of reducing cost while increasing speed and throughput. Due to their small size, lab-on-chip solutions are even considered where most of the sample preparation, amplification, and sequencing is all taking place on a single glass surface (chip). Their unique manufacturing methods allow for a large number of capillaries at low cost with arbitrary geometries, which are not possible with standard capillaries.

The Mathies group at University of California at Berkeley (28) published a method for fabricating capillary systems with complex architecture and layout on glass substrates using photolithography. One of their systems can achieve a read length of 500 bases with 99% accuracy and a cycle speed of 20 min (29).

A rather interesting technique is Massive Parallel Signature Sequencing (MPSS) (30). Using a microarray structure and beads, targets are sequenced iteratively at each cycle using a type IIs restriction enzyme that cleaves (cuts) within a target sequence, exposing a four-base-pair overhang. The overhang is identified using a sequence-specific ligation of a fluorescent linker. The method can read up to 20 bases (in 4–5 cycles) making it well suited for expression analysis.

Mass Spectrometry Based DNA Sequencing

Matrix Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry (MALDITOF–MS) is the first MS based DNA sequencing technique (31). The method can replace the electrophoretic molecule separation step in DNA sequencing by a MS component. The mass of the molecule is estimated by measuring the time to travel of gas-phase DNA molecules within a flight tube that connects an excitation source (ultraviolet, UV laser) and an ion-to-electron conversion detector. The molecules collide at the detector thus registering the time to travel, which is analogous to molecular mass. The technique has the advantages of allowing the fast and parallel separation of a heterogeneous mixture of molecules without being affected by possible secondary structures that the DNA molecules have fallen into. Although the read lengths remain small, the availability of fast and autonomous MALDITOF–MS DNA sequencers (Sequenom Corporation in San Diego, CA) makes them good candidates for precise resequencing of small fragments useful in SNP detection.

DNA Sequencing at the Nanoscale

Most of the methods described below are based on manipulating properties of DNA at the nanoscale or utilizing properties of other materials at the nanoscale. Since such methods work with very low concentrations they can also be viewed as single-molecule sequencing methods and are suitable for applications when the template DNA is in very low concentration and amplification techniques could not be applied efficiently.

DNA Detection with Nanoparticles. Nanosphere Inc. (Northbrook, IL) has developed a method for rapid and low concentration detection of proteins and nucleic acids (32). Their technology is based on attaching oligonucleotide probes on nanoparticles. The probes attach to the target DNA and due to the unique properties of the used nanoparticles the event can be detected electrically, optically, or magnetically without amplification of the target sequence. The concentrations needed are below the operational threshold of PCR reactions. Although the techniques have not been extended to *de novo* sequencing the unique detection characteristics are proving very useful in detection scenarios.

In another effort from the founder of Nanosphere, Dr. Mirkin at Northwestern University (Evanston, IL), the electrical detection of DNA was first proposed (33). With this protocol the imaging aspect of microarray applications can be eliminated using gold nanoparticles that once hybridized onto the DNA probes and deposited with silver can close an electric circuit thus enabling detection of the hybridization event with electrical signals.

Sequencing with Atomic Force Microscopy. Atomic Force Microscopy (AFM) was invented at IBM Zurich Labs in 1986 and has completely revolutionized research at the nanoscale (34). A nanoscopic tip that is attached at the end of the cantilever interacts with the surface of the target material and records the tips deflections to create a topographic map of the surface. The AFM can be used to study the surface of duplex DNA and detect mismatches or it can be used as force measuring tool to study the mechanical properties of DNA. One application of particular interest is the AFM assisted unzipping of the DNA duplex, where a DNA duplex is suspended between a solid support and the AFM tip. Pulling the AFM tip further causes the duplex to unzip. The force needed to unzip depends on the percentage of the Gas Chromatography (GC) content, and hence can be used to estimate the GC content of an unknown target if needed in a more large-scale sequencing function (35). A very similar idea was proposed in Ref. 36 where optical traps are used to stretch DNA molecules and to measure force.

Nanopore Sequencing. Another interesting technique that uses features at the nanoscale is nanopore sequencing. As DNA passes through an 1.5 nm nanopore, different base pairs hinder the pore to different degrees, altering the electric conductivity of the pore (37). The pore conductance can be measured and monitored to identify the DNA sequence. The accuracy of base calling ranges from 60% for single events to 99.9% for 15 events. The technique has only been shown to work experimentally on certain sequences but exhibits a big potential for super fast sequencing without amplification of the target. It is evident that the evolution of this technique depends on nanopore engineering. To break apart from this restriction Visigen (Houston, TX) and Li-cor (Lincoln, NE: U.S. Patent 6,306,607) are in the process of engineering DNA polymerases or fluorescent labeled nucleotides that can provide real-time, base-dependent signals during the natural DNA synthesis process.

Sequencing by Fluorescence Microscopy. This is a new class of DNA sequencing methodologies, for which the fluorescence emitted during single molecule interactions is detected (38). The interactions most commonly referred to are single nucleotide incorporation during DNA polymerase replication or nucleotide digestion from an exonuclease. The change in fluorescence emission is detected using microscopes and CCDs. An enabling technology is fluorescence resonance electron transfer (FRET), where the fluorescence emission of two molecular dyes can be affected by their proximity. The research in the area is vast and already three companies, Nanofluidics (Menlo Park, California), Solexa (Essex, UK), and GenoVoxx (Lubeck, Germany), are developing products based on this technology for high throughput DNA detection and genotyping.

DNA Computing Based DNA Sequencing

Up to this point instruments and electronic computers were assigned the task of analyzing and processing DNA sequences. In 1994, the roles were reversed by the first proof of concept experiment by Adleman of using DNA to perform computations (39). This development led to the birth of the field of DNA computing [for a short introduction see (40)].

Landweber and Lipton were the first to suggest that DNA computing can be used to improve the performance of DNA sequencing (41). Their approach is based on DNA^2DNA computations, where nucleotides of an unknown sequence are translated into a new DNA sequence using a unique mapping transformation. A library of DNA oligonucleotides is synthesized and mixed in the solution containing the template DNA. The oligonucleotides then anneal to complementary parts. The partially double-stranded sequences are ligated and hybridized on a DNA chip. The reconstruction of the encoded sequence is achieved by analyzing the DNA array image. Although the technique was never implemented in large scale it points to potential future applications where instruments can be assisted by DNA computers.

The first proof of such development came a few years later in an announcement by Dr. Suyama from the University of Tokyo and Olympus Corp. (Japan), where they developed the first DNA-computer-assisted gene expression instrument (42). The instrument is a hybrid of a molecular computer and an electronic–digital computer. The molecular computer is in charge of DNA input–output, DNA reactions, capture of DNA results, and DNA detection while the electronic is responsible for information processing by means of DNA reaction calculations and result analysis.

In Ref. 43, a new laboratory protocol is proposed that assists in the faster sequencing of genomes. The distance between primers (probes) that have annealed on a target sequence can be estimated by measuring the intensity and color of light emission of specialized hybridization array. Although the method is not intended for de novo sequencing it is proposed as an alternative method of comparing genomes.

Bio-informatics, a subfield of computational biology, refers to processing, analyzing or storing DNA sequencing data with computers. The field of analyzing DNA sequences using digital signal processing theory has been known as genomic signal processing (44). The idea is to process the sequence of DNA as a digital signal and find certain characteristics. Recently, the application of DNA computing in digital signal processing, termed as DNA-based Digital Signal Processing, has been suggested (45). A future is envisioned where a DNA based digital signal processor can process DNA sequences and output certain characteristics in the form of DNA sequences that can be subsequently detected (or sequenced). This will allow researchers to process a vast amount of DNA sequences without prior sequencing.

BIBLIOGRAPHY

1. Watson JD, et al. Molecular biology of the gene. 5th ed. San Francisco: Pearson/Benjamin Cummings; 2004.
2. Shopsin B, Kreiswirth BN, Molecular Epidemiology of Methicillin-Resistant *Staphylococcus aureus*, [serial on the Internet], 2001; 7(2) Accessed 2005 July 14. Available at http://www.cdc.gov/ncidod/eid/vol7no2/shopsin.htm.

3. Weiner MP, Hudson TJ. Introduction to SNPs: discovery of markers for disease. Biotechniques 2002;32(Suppl.) S4–S13.
4. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977;74:5463–5467.
5. Maxam AM, Gilbert W. A new method of sequencing DNA. Proc Natl Acad Sci USA 1977;74:560–564.
6. Studier FW. Slab-gel electrophoresis. Trends Biochem Sci 2000;25(12):588–590.
7. Meldrum D. Automation for genomics. Part Two: Sequencers, microarrays, and future trends. Genome Res 2000;10:1288–1303.
8. Huang GM. High-throughput DNA sequencing: a genomic data manufacturing process. DNA Seq 1999;10:149–153.
9. Viovy JL, Duke T. DNA electrophoresis in polymer solutions: Ogston sieving, reptation and constraint release. Electrophoresis 1993;14(4):322–329.
10. Zagursky RJ, McCormick RM. DNA sequencing separations in capillary gels on a modified commercial DNA sequencing instrument. Biotechniques 1990;9(1):74–79.
11. Green P (No date). Phrep, Phrap and Consed [Online]. University of Washington. Available at http://www.phrap.org/phredphrapconsed.html. Accessed 2005, June 29.
12. Ewing B, Green P. Basecalling of automated sequencer traces using phred. II. Error probabilities. Genome Res 1998;8:186–194.
13. Ewing B, Hillier L, Wendl M, Green P. Basecalling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 1998;8:175–185.
14. Hart C. (1997, August 1). Autoseq home page. [Online]. In-Machina. Availabel at http://www.in-machina.com/~reece/autoseq/ Accessed 2005, June 29.
15. Software Sequencing (No date). GE Healthcare - formerly Amersham Biosciences - Sequencing [Online]. GE Healthcare. Available at http://www5.amershambiosciences.com/aptrix/upp01077.nsf/Content/autodna_software_sequencing. Accessed 2005, June 29.
16. Lasergene (No date). DNASTAR. [Online]. DNASTAR, Inc. Available at http://www.dnastar.com/web/index.php. Accessed 2005, June 29.
17. Sequencher (No date). Gene Codes Corporation: Sequencher. [Online]. Gene Codes Corporation. Available at http://www.genecodes.com/sequencher/. Accessed 2005, June 29.
18. Staden Package (No date). Staden Package Home Page [Online]. SourceForge. Available at http://staden.sourceforge.net/ Accessed 2005, June 29.
19. Applied Biosystems Product Information Page (No date). Sequence Analysis Software [Online]. Applied Biosystems. Available at http://www.appliedbiosystems.com/. Accessed 2005, June 29.
20. Huang XC, Quesada MA, Mathies RA. DNA sequencing using capillary array electrophoresis. Anal Chem 1992;64(18):2149–2154.
21. Kambara H, Takahashi S. Multiple-sheathflow capillary array DNA analyzer. Nature(London) 1993;361(6412):565–566.
22. Crabtree HJ. Capillary array DNA sequencer based on a micromachined sheath-flow cuvette. Electrophoresis 2000;21:1329–1335.
23. Ronaghi M. Pyrosequencing Sheds Light on DNA Sequencing. Genome Res 2001;11:3–11.
24. Drmanac R, et al. DNA sequence determination by hybridization: a strategy for efficient large-scale sequencing. Science 1993;260:1649–1652; Erratum, Science 1994;163(5147):596.
25. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270(5235):467–470.
26. Venter JC, et al. The Sequence of the Human Genome. Science 2001;291:1304–1351.
27. Sanger F, et al. Nucleotide sequence of bacteriophage lambda DNA. J Mol Biol 1982;162(4):729–773.
28. Woolley AT, Mathies RA. Ultra-High-Speed DNA Fragment Separations Using Microfabricated Capillary Array Electrophoresis Chips. Proc Natl Acad Sci USA 1994;91:11348–11352.
29. Simpson PC. High-throughput genetic analysis using microfabricated 96-sample capillary array electrophoresis microplates. Proc Natl Acad Sci USA 1998;95:2256–2261.
30. Brenner S, et al. *In vitro* cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. Proc Natl Acad Sci USA 2000;97:1665–1670.
31. Cantor CR, et al. DNA sequencing after the Human Genome Project. Nucleosides Nucleotides 1997;16:591–598.
32. Nam J-M, Park S-J, Mirkin CA. Bio-barcodes based on oligonucleotide-modified nanoparticles. J Am Chem Soc 2002;124:3820–3821.
33. Park SJ, Taton TA, Mirkin CA. Array-Based Electrical Detection of DNA Using Nanoparticle Probes. Science Feb. 2002;295(5559):1503–1506.
34. Binnig G, Quate CF, Gerber C. Atomic force microscope. Phys Rev Lett 1986;56:930–933.
35. Essevaz-Roulet B, Bockelmann U, Heslot F. Mechanical separation of the complementary strands of DNA. Proc Natl Acad Sci USA 1997;94:11935–11940.
36. Wang MD, et al. Stretching DNA with optical tweezers. Biophy J 1997;72:1335–1346.
37. Deamer DW, Branton D. Characterization of nucleic acids by nanopore analysis. Acc Chem Res 2002;35:817–825.
38. Braslavsky I, Hebert B, Kartalov E, Quake SR. Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci USA 2003;100:3960–3964.
39. Adleman L. Molecular computation of solutions to combinatorial problems. Science Nov. 1994;266:1021–1024.
40. Tsaftaris SA, Katsaggelos AK, Pappas TN, Papoutsakis ET. DNA computing from a signal processing viewpoint. IEEE Sig Proc Mag 2004;21(5):100–106.
41. Landweber LF, Lipton RJ. DNA2DNA Computations: A potential 'killer app'? Proceedings of the 24th International Colloquium on Automata, Languages and Programming (ICALP). Lecture Notes in Computer Science. New York: Springer-Verlag; 1997. 672–683.
42. Normile D. DNA-Based Computer Takes Aim at Genes. Science 2002;295(5557):951.
43. Mishra B. Comparing Genomes. Comp Sci Eng 2002;4(1):42–29.
44. Anastassiou D. Genomic Signal Processing. IEEE Sig Proc Mag 2001;18(4):8–20.
45. Tsaftaris SA, Katsaggelos AK, Pappas TN, Papoutsakis ET. How can DNA-Computing be applied in Digital Signal Processing?. IEEE Sig Proc Mag 2004;21(6):57–61.

Further Reading

The following two articles provide a well-rounded review of commercially available and experimental sequencing techniques.

Marziali A, Akeson M. New DNA sequencing methods. Annu Rev Biomed Eng 2001;3:195–223.

and

Shendure J, Mitra RD, Varma C, Church GM. Advanced sequencing technologies: methods and goals. Nature Rev Genet 2004;5(5):335–344.

The reader is suggested to study the projections into the future of sequencing technology of the first article and compare it to the presentation of the current status of the second article. The advancement in technology is rather interesting given that the papers are only three years apart.

A very informative presentation of the development of capillary array electrophoresis can be found in the following references.

Dovichi NJ, Zhang J. How capillary electrophoresis sequenced the human genome. Angew Chem Int Ed Engl 2000;39(24):4463–4468.

For a review of lab-on-chip methods for sequencing and genotyping please see the following reference.

Kan CW, Fredlake CP, Doherty EA, Barron AE. DNA sequencing and genotyping in miniaturized electrophoresis systems. Electrophoresis Nov. 2004;25(21–22):3564–3588.

An excellent review paper on DNA microarray technology is by the following references.

Venkatasubbarao S. Microarrays—status and prospects. Trends in Biotechnology 2004;22 (12):630–637.

Stears RL, Martinsky T, Schena M. Trends in microarray analysis. Nature Med 2003;9(1).

This article gives an interesting view of how microarrays can change the future of diagnostics.

Sauer S, et al. Miniaturization in functional genomics and proteomics. Nat Rev Genet 2005;6(6):465–76.

Finally, this article provides an overall picture on the effect of miniaturization of diagnostic and laboratory techniques in biology and medicine.

See also BIOINFORMATICS; MICROARRAYS; POLYMERASE CHAIN REACTION.

DOPPLER ECHOCARDIOGRAPHY.

See ECHOCARDIOGRAPHY AND DOPPLER ECHOCARDIOGRAPHY.

DOPPLER ULTRASOUND.

See ULTRASONIC IMAGING.

DOPPLER VELOCIMETRY.

See CUTANEOUS BLOOD FLOW, DOPPLER MEASUREMENT OF.

DOSIMETRY, RADIOPHARMACEUTICAL.

See RADIOPHARMACEUTICAL DOSIMETRY.

DRUG DELIVERY SYSTEMS

DONATELLA PAOLINO
MASSIMO FRESTA
University of Catanzaro Magna Græcia
Germaneto (CZ), Italy

PIYUSH SINHA
MAURO FERRARI
The Ohio State University
Columbus, Ohio

PRINCIPLES OF CONTROLLED DRUG DELIVERY

A perspective drug delivery systems can be defined as mechanisms to introduce therapeutic agents into the body. Chewing leaves and roots of medical plants and inhalation of soot from the burning of medical substances are examples of drug delivery from the earliest times. However, these primitive approaches of delivering drugs lacked a very basic need in drug delivery; that is, consistency and uniformity (a required drug dose). This led to the development of different drug delivery methods in the later part of the eighteenth and early nineteenth century. Those methods included pills, syrups, capsules, tablets, elixirs, solutions, extracts, emulsions, suspension, cachets, troches, lozenges, nebulizers, and many other traditional delivery mechanisms. Many of these delivery mechanisms use the drugs derived from plant extracts.

The modern era of medicine development started with the discovery of vaccines in 1885 and techniques for purification of drugs from plant sources in the late nineteenth century, followed by the introduction of penicillin after its discovery in 1929, and a subsequent era of prolific drug discovery. The development and production of many pharmaceuticals involves the genetic modification of microorganisms to transform them into drug-producing factories. Examples are recombinant deoxyribonucleic acid (DNA), human insulin, interferon [for the treatment of acquired immunodeficiency syndrome (AIDS) related Kaposi's sarcoma, Hairy cell leukemia, Hepatitis B and C, etc.], interleukin-2 (Renal cell and other carcinomas), erythropoietin (for the treatment of anemia associated with chronic renal failure/AIDS/antiretroviral agents, chemotherapy-associated anemia in nomnyloid malignancy patient), and tissue plasminogen activator (1). It is now possible to produce oligonucleotide, peptide, and protein drugs in large quantities, while gene therapies also appear to be clinically feasible. Each of these therapeutic agents, by virtue of size, stability, or the need for targeting, requires a specialized drug delivery system (2). While the conventional drug delivery forms are simple oral, topical, inhaled, or injections, more sophisticated delivery systems need to take into account pharmacokinetic principles, specific drug characteristics, and variability of response from one person to another and within the same person under different conditions.

The efficacy of many therapeutic agents depends on their action on target macromolecules located either within or on the surface of particular cells types. Many drugs interact with enzymes or other macromolecules that are shared by a large number of cell types, while most often a drug exerts its action on one cell type for the desired therapeutic effect. Certain hormones, for example, interact with receptor mechanisms that are present in only one or a few cell types. An ideal gene delivery system should allow the gene to find its target cell, penetrate the cell membrane, and enter into the nucleus. Further, genes should not be released until they find their target and one has to decide whether to release the genes only once or repeatedly through a predetermined way (2). Thus, the therapeutic efficacy of a drug can be improved and toxic effects can be reduced by augmenting the amount and persistence of drugs in the vicinity of the target cells, while reducing the drug exposure to the nontarget cells.

This basic rationale is behind controlled drug delivery. A controlled drug delivery system requires simultaneous consideration of several factors, such as the drug property, route of administration, nature of delivery vehicle, mechanism of drug release, ability of targeting, and biocompatibility. These have been summarized in Fig. 1.

It is not easy to achieve all these in one system because of extensive independency of these factors. Further,

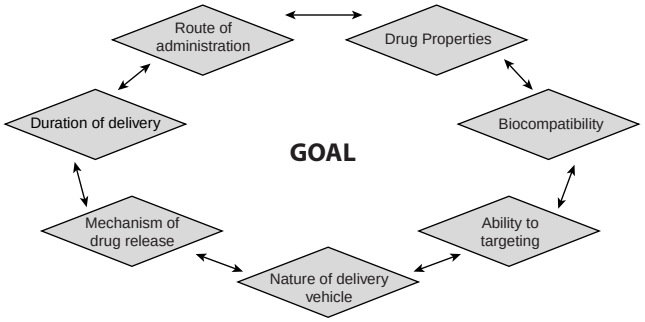


Figure 1. Design requirement for a drug delivery systems.

reliability and reproducibility of any drug delivery systems is the most important factor while designing such a system. The emphasis here is on the need for precision of control and to minimize any contribution to intraand intersubject variability associated with the drug delivery system. There are many different approaches for controlled drug delivery applications (3). They are summarized in the following section.

Overview of the Development of Drug Delivery Systems

To obtain a given therapeutic response, the suitable amount of the active drug must be absorbed and transported to the site of action at the right time and the rate of input can then be adjusted to produce the concentrations required to maintain the level of the effect for as long as necessary. The distribution of the drug-to-tissues other than the sites of action and organs of elimination is unnecessary, wasteful, and a potential cause of toxicity. The modification of the means of delivering the drug by projecting and preparing new advanced drug delivery devices can improve therapy. Since the 1960s, when silicone rubber was proposed as an implantable carrier for sustained delivery of low molecular weight drugs in animal tissues, various drug delivery systems have been developed.

At the beginning of the era of controlled drug delivery systems, a controlled release system utilizes a polymer matrix or pump as a rate-controlling device to deliver the drug in a fixed, predetermined pattern for a desired time period (4). These systems offered the following advantages compared to other methods of administration: (*1*) the possibility to maintain plasma drug levels a therapeutically desirable range, (*2*) the possibility to eliminate or reduce harmful side effects from systemic administration by local administration from a controlled release system, (*3*) drug administration may be improved and facilitated in underpriviledged areas where good medical supervision is not available, (*4*) the administration of drugs with a short *in vivo* half-life may be greatly facilitated, (*5*) continuous small amounts of drug may be less painful than several large doses, (*6*) improvement of patient compliance, and (*7*) the use of drug delivery systems may result in a relatively less expensive product and less waste of the drug. The first generation of controlled delivery systems presented some disadvantages, that is possible toxicity, need for surgery to implant the system, possible pain, and difficulty in shutting off release if necessary. Two types of diffusion-controlled systems have been developed. The reservoir is a core of drug surrounded with a polymer film. The matrix system is a polymeric bulk in which the drug is more or less uniformly distributed.

Pharmaceutical applications have been made in ocular disease with the Ocusert, a reservoir system for glaucoma therapy that is not widely used, and in contraception with four systems: (*1*) subdermal implants of nonbiodegradable polymers, such as Norplant (6 capsules of 36 mg levonorgestrel); (*2*) subdermal implant of biodegradable polymers; (*3*) steroid releasing intrauterine device (IUD); and (*4*) vaginal rings, which are silicone coated. Other applications have been made in the areas of dentistry, immunization, anticoagulation, cancer, narcotic antogonists, and insulin delivery. Transdermal delivery involves placing a polymeric system containing a contact adhesive on the skin.

Since the pioneering work in controlled drug delivery, it was demonstrated that when a pharmaceutical agent is encapsulated within, or attached to, a polymer or lipid, drug safety and efficacy may be greatly improved and new therapies are possible (5). This concept prompted active and intensive investigations for the design of degradable materials, intelligent delivery systems, and approaches for delivery through different portals in the body. Recent efforts have led to development of a new approach in the field of controlled drug delivery with the creation of responsive polymeric drug delivery systems (6). Such systems are capable of adjusting drug release rates in response to a physiological need. The release rate of these systems can be modulated by external stimuli or self-regulation process.

Different Approach for Controlled Drug Delivery

Localized Drug Delivery. In many cases, it would be desired to deliver drugs at a specific site inside the body to a particular diseased tissue or organ. This kind of regional therapy mechanism would reduce systemic toxicity and achieve peak drug level directly at the target site. A few examples of drugs that require this kind of therapy are anticancer drugs, antifertility agents, and antiinflammatory steroids. These drugs have many severe unintended side effects in addition to their therapeutic effects.

Targeted Drug Delivery. The best controlled mechanism would be delivery of drug exclusively to the targeted cells or cellular components. That means the development of delivery mechanisms that would equal or surpass the selectivity of naturally occurring effectors (e.g., peptide hormones). As in the case of hormone action, drug targeting would probably involve a recognition event between the drug carrier mechanism and specific receptors at the cell surface. The most obvious candidates for the targetable drug carriers are cell-type specific immunoglobulins. The concept of targeted drug delivery is different than localized drug delivery. The latter simply implies localization of the therapeutic agent at an organ or

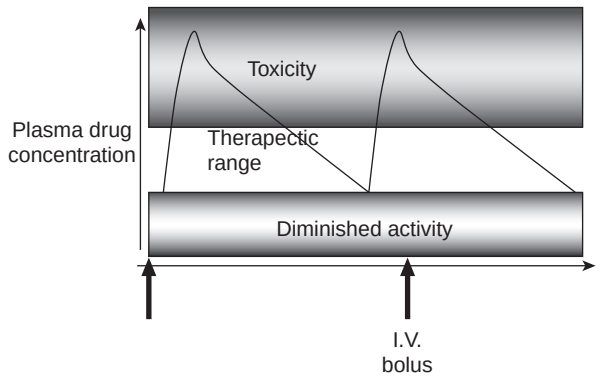


Figure 2. Plasma concentration versus time curve for intravenous (IV) drug administration showing first-order kinetic.

tissue site, while the former implies more subtle delivery to specific cell types.

Sustained Drug Delivery (Zero Order Release Profile). Injected or ingested drugs follow first-order kinetics, with initial high blood levels of the drug after initial administration, followed by an exponential fall in blood concentration. Toxicity often occurs when blood levels peak, while efficacy of the drug diminishes as the drug levels fall below the therapeutic range. This profile is shown in Fig. 2. and the drug kinetics is undesirable, especially in the case where the margin between toxicity and required therapeutic concentration levels is small. The importance of controlled-release drug delivery systems may be argued with reference to the goal of achieving a continuous drug release profile consistent with zero-order kinetics, wherein blood levels of drugs would remain constant throughout the delivery period. The therapeutic advantages of continuous-release drug delivery systems are thus significant, and encompass: *in vivo* predictability of release rates on the basis of *in vitro* data; minimized peak plasma levels, and thereby reduced risk of toxic effects; predictable and extended duration of action; reduced inconvenience of frequent dosing, thereby improving patient compliance (7,8).

Figure 3 illustrates the constant plasma concentration that is desired for many therapeutic agents.

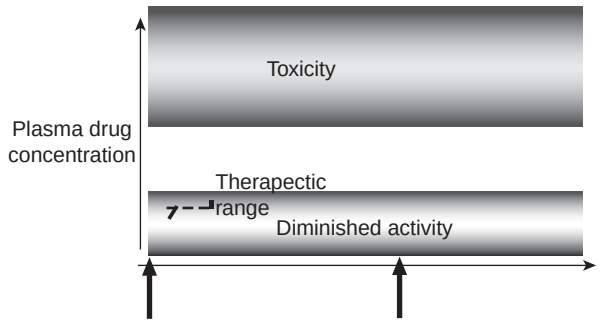


Figure 3. Plasma concentration versus time curve for sustained release profile of zero-order kinetics and pulsatile release profile.

The controlled release aspect of sustained drug delivery systems pertain to a reliable and reproducible system whose rate of drug delivery is independent of the environment in which it is placed. This requirement emphasizes the need for precision of control and elimination of undesired contribution associated with the drug delivery system.

Modulated Drug Delivery (Nonzero-Order Release Profile). A significant challenge in drug delivery is to create a delivery system that can achieve manipulable nonzero-order release profile. This could be pulsatile or ramp or some other pattern. In some cases it is also required that the release should be immediate. A pulsatile release profile within the therapeutic window is shown in Fig. 3.

Feedback Controlled Drug Delivery. The ideal drug delivery system is the feedback controlled drug delivery system that releases drug in response to a therapeutic marker. This can be classified into two classes: modulated and triggered device. A modulated device involves the ability to monitor the chemical environment and changes drug delivery rate continuously in response to the specific external marker, while in a triggered device no drug release takes place until it is triggered by a marker.

These different approaches of drug delivery can have different routes of administration. Some of the most preferred routes are oral, pulmonary inhalation, transdermal, transmucosal, and implantable systems.

Implantable Controlled Drug Delivery Devices. Although most controlled drug delivery systems are designed for transdermal, subcutaneous, or intramuscular uses, implantable devices are very attractive for a number of classes of drugs, particularly those that cannot be delivered via the oral route or are irregularly absorbed via the gastrointestinal (GI) tract (9). Implantable systems are designed to deliver therapeutic agents into the bloodstream. This replaces the repeated insertion of IV catheters. The basic idea behind this device is simple: The treatment of certain diseases that require the chronic administration of drug could benefit from the presence of implantable devices. These systems can also be used to deliver drug to the optimum physiological site. These systems are particularly suited for drug delivery requirements of insulin, steroids, chemotherapeutics, antibiotics, analgesics, contraceptives, and heparin. Implantable systems are placed completely under the skin (usually in a convenient, but inconspicuous location). Benefits include the reduction of side effect (drug delivery rate within the therapeutic window) caused by traditional administration techniques, and better control. Ideally an implantable system will have a feedback controlled release mechanism and will be controlled by electronics with a long-life power source to achieve zero-order or manipulable nonzero-order release profiles in a manner similar to a physiological release profile.

The focus of this research is on two major requirements of an implantable controlled drug delivery device:

1. One of the major requirements for implantable drug delivery devices is to allow controlled-release of therapeutic agents, especially biological molecules, continuously over an extended period of time. The goal here is to achieve a continuous drug release profile consistent with zero-order kinetics where the concentration of drug in the blood remains constant throughout the delivery period. As mentioned earlier, the therapeutic advantages of continuous release of drug by implantable delivery devices are significant: minimized adverse reactions by reducing the peak levels, predictable and extended duration of action, reduced inconvenience of frequent dosing and thereby improved patient compliance.
2. The second, and more important requirement, is to achieve a manipulable nonzero-order release profile, such as pulsatile or any other pattern required for applications in therapeutic medicine. Vaccines and hormones are examples that require pulsatile delivery (10,11). Gonadotropin releasing hormone, for example, is most effective when delivered in a pulsatile manner to female patients undergoing treatment for infertility.

A sequence of two implantable systems was developed to achieve the above mentioned goals. The first device that addresses the first goal is named nanochannel delivery system I (or nDSI), while the device that addresses the second goal is called nanochannel delivery system 2 (or nDS2).

The Economics of Drug Delivery Devices

The fact that drug delivery technology can bring both therapeutic and commercial value to healthcare products cannot be neglected. Big pharmaceutical companies have recently started losing their market share to generic competitors after their patents expired, and therefore they have started recognizing the importance of drug delivery companies. Pharmaceutical companies are looking to extend their patents lifetimes by making strategic alliances with drug delivery technology companies, by presenting old drugs in new forms. Most of the drug delivery products therefore reach the market as a result of strategic alliance between drug delivery companies and pharmaceutical companies. Pharmaceutical companies provide the drug that may not be delivered efficaciously with a conventional delivery mechanism, while the drug delivery companies provide the cutting edge technology to administer the drug more effectively. The joint venture not only offers considerable advantages over the R&D efforts to bring new drug into the market as drug delivery systems provide means to reformulate existing products, but it also protects the drugs from erosion by generics in the case of patented drugs. As a result, drug delivery technology companies seem to enjoy a good return on their investments in the form of increased revenues and market share (9,12).

The global drug delivery market grew between 1998 and 2002, with a compound annual growth rate (CAGR) of 13.7%, increasing from \$39.6 billion to slightly $>$ \$66 billion. The market is expected to grow at a slightly lower CACR of 11.6% between 2002 and 2007 corresponding to a market value of \$114.3 billion by 2007. One of the contributing factors in this growth is the use of drug delivery systems as strategy to expand the shelf-life of products (particularly blockbusters), enabling pharmaceutical companies to sustain the revenue streams from their best sellers.

The largest market for drug delivery systems in the world is in the United States, having captured 47.9% of the global market's revenue generation in 2002. This figure is forecast to fall to 41.9% by 2007 although the U.S. market will retain its position as the leading market. The U.S. market for drug delivery systems was worth \$31.7 million in 2002, having experienced a CAGR of 12.6% during 1998–2002. Oral drug delivery systems had the largest market share, taking 47.7vo of the total market share. Transmucosal, injectable, and implantable systems together had 8.8% of the market share in 2002. The U.S. market value for drug delivery systems is expected to grow at a rate of 8.5% annually, reaching a value of \$48 billion by 2007.

MICROELECTRO-MECHANICAL SYSTEMS

A number of devices have been developed to achieve controlled drug delivery. These devices utilize a different route of administration and different materials for device fabrication. Typically, each of these devices is targeted toward delivering one or a few of the therapeutics. The factors that need to be considered when designing a drug delivery device were previously discussed in great details (Fig. 1). This article begins with a brief history of implantable drug delivery devices. These include polymeric devices, osmotic pumps, micropumps, and microelectro-mechanical systems (MEMS) based devices. Since the drug delivery devices developed in this research are based upon MEMS technology, a good understanding of MEMS fabrication technology is needed, and therefore under the section MEMS for drug delivery devices, it is digressed from the topic implantable drug delivery devices and a more in-depth description on the use of MEMS for different drug delivery devices is presented. This includes MEMS for transdermal, oral, injectable, and *implantable* drug delivery. This article concludes with a critical analysis of implantable drug delivery devices.

A History of Implantable Drug Delivery Devices

The history of implantable devices goes back to May 1958 when the first implantable cardiac pacemaker was placed in an experimental animal (13). Later that year the first pacemaker was implanted in a human that operated for 3 h and then failed (14). The second unit operated for 8 h before failing, and the patient went unstimulated for 3 years before receiving a satisfactory implantable unit. The record shows that this patient was alive in 1991 and was using a pacemaker (15). The development of an implantable pacemaker revolutionized the field of biomedical science and

engineering over the last 30 years providing many different implantable biomedical devices to the medical professionals for therapeutic and diagnostic use. Today, implantable cardioverter–defibrillators, drug delivery systems, neurological stimulators, bone growth stimulators, and other implantable devices make possible the treatment, of a variety of diseases.

Extensive research has been done on implantable drug delivery devices over the last 30 years. Different technologies have been developed with many breakthroughs in clinical medicine. The first such device that saw extensive clinical use was reported in the 1970s (15–18). This system used a bellows-type pump activated by partially liquefied Freon. The Freon was reliquefied with each transcutaneous refill of the implantable device, and the administration was constant. Later, extensive research started to develop more sophisticated devices that could offer better control and more clinical options. Another device was developed by Medtronic Company that has a peristaltic pump to deliver the drugs (19). The device was controlled by electronics. Another system developed by MimiMed Technologies employs a solenoid pump, a reservoir, and advanced electronic control (20). The Infusaid Company developed an advanced programmable implantable pump that employed a bellows-type pump and a solenoid valve set to control drug flow (21). Other technologies developed to achieve this goal are summarized in the following sections.

Polymeric Implants. Polymers have been used extensively in controlled drug delivery systems. These can be classified as (*1*) nondegradable polymeric reservoirs and matrices, and (*2*) biodegradable polymeric devices. The first kind of polymeric devices are basically silicone elastomers. This kind of drug delivery system is based upon the research conducted in the 1960s, when researchers recognized that certain dye molecules could penetrate through the walls of silicone tubing (22–24). This lead to the development of reservoir-based drug delivery system, which consisted of hollow polymer tubes filled with a drug suspension. The drug is released by dissolution into the polymer and then diffusion through the walls of the polymeric device. The two most commonly used nondegradable polymers are silicone and poly(ethylene-covinyl acetate) (EVAc). The Norplant 5 year contraceptive drug delivery system is based upon this technology. Some of the implantable reservoir systems are simple cylindrical reservoir surrounded by a polymeric membrane. The other variety in this first category is constructed of a solid matrix of nondegradable polymers. These systems are prepared by homogeneous dispersement of drug particles throughout the matrix (25). Drug release occurs by diffusion through the polymer matrix or by leaching or a combination of both (26). The matrix may be composed of either a lipophilic or hydrophilic polymer depending on the properties of the drug and the rate of release desired. However, it is difficult to achieve constant rates of drug release with nondegradable matrix systems, for example, the rate of release of carmustine from an EVAc matrix device drops continuously during incubation in buffered water (27). Constant release can sometimes be achieved by making the matrix as a reservoir surrounded by a shell of rate-limiting polymeric membrane. In some cases, water soluble, cross-linked polymers can be used as matrices. Release is then activated by swelling of the polymer matrix after exposure to water (28). One other kind is a magnetically controlled system where magnetic beads are dispersed within the matrix (25). Drug is released by diffusion with a concentration gradient. The addition of an externally oscillating magnetic field causes the physical structure of the polymer to alter, creating new channels, and thus leading to further drug release.

Biodegradable polymeric devices are formed by physically entrapping drug molecules into matrices or microspheres. These polymers dissolve when implanted (injected) and release drugs. Examples of biodegradable polymers are poly(lactide-*co*-glycolide) (PLGA), and poly (*p*-carboxyphenoxypropane-*co*-sebacic acid) (PCPP-SA) (24). Some of the commercially available polymeric devices are Decapeptyl, Lupron Depot (microspheres), and Zoladex (cylindrical implants) for prostate cancer and Gliadel for recurrent malignant glioma. The half-life of therapeutics administered by microspheres is much longer than free drug injection. Polymers are also being investigated for treating brain tumors (29), and delivery of proteins and other macromolecules (30).

The above mentioned polymeric implants are utilized for sustained drug delivery. Methods have been developed to achieve controlled drug delivery profiles with implantable polymeric systems (31,32). These technologies include preprogrammed systems, as well as systems that are sensitive to (triggered or modulated by) modulated enzymatic or hydrolytic degradation, pH, magnetic fields, ultrasound, electric fields, temperature, light, and mechanical simulation. Researchers are also exploring the use of nontraditional MEMS fabrication techniques and materials that could be used to form microwell- or microreservoir-based drug delivery devices. For example, microwells of varying sizes (as small as 3fL/well) have been fabricated by micromolding of poly(dimethylsiloxane) (PDMS) on a photoresist-coated silicon wafer that is photolithographically patterned (33).

Osmotic Pumps. Osmotic pumps are energy modulated devices (9). These are usually capsular in shape. When the system is exposed to an aqueous environment, such as that after subcutaneous implantation, water is drawn to the osmotically active agent through a semipermeable membrane and pressure is supplied to the collapsible drug reservoir and drug is released through an orifice with precise dimension. The delivery mechanism is dependent on the pressure created and is independent of drug properties. The ALZET pumps (only for investigational purpose at this time, not for humans) have been used in thousands of studies on the effects of controlled delivery of a wide range of experimental agents, including peptides, growth factors, cytokines, chemotherapeutic drugs, addictive drugs, hormones, steroids, and antibodies (34). The ALZA Corporation built the DUROS implant based upon the foundation of the ALZET osmotic pump, the system of choice for implant drug delivery in research laboratories around the world for >20 years. Viadur, a once-yearly implant for the palliative treatment of advanced prostate cancer, is the first

approved product to incorporate ALZAs proprietary DUROS implant technology. A single Viadur implant continuously delivers precise levels of the peptide leuprolide for a period of 1 full year, providing an alternative to frequent leuprolide injections. Although most of the osmotic pumps are designed for sustained release profile, research is being conducted to modify this design for different patterns (9). Further, a catheter was attached to the exit port of an implantable osmotic pump to achieve site specific drug delivery at a location distant from site of implantation (35).

Micropumps. Micropumps have been actively investigated for drug delivery applications. Some micropumps are nonmechanical that utilizes electrohydrodynamic, electroosmotic, ultrasonic, or thermocapillary forces (36). However, most of the micropumps are mechanical, composed of mechanically moving membranes. A number of mechanical micropumps have been developed using various mechanisms, including piezoelectric (37), electrostatic (38), thermopneumatic (39), electromagnetic (40), bimettalic (41), shape memory alloy (SMA) (42), ionic conducting polymer films JCPF (43), and surface tension driven actuators (36). One example is the silicon piezoelectric micropump based on silicon bulk micromachining, silicon pyrex anodic bonding, and piezoelectric actuation (37). This can be used for application requiring low (typically $1\,\mu L \cdot min^{-1}$), precisely controlled flow rate. The whole system includes the refillable reservoir, control, and telemetry electronics and battery. This can be implanted in the abdomen and a catheter can be brought to the specific site. The SynchroMed pump is an implantable, programable, battery-powered device commercially available by Medtronics (44). A large number of other implantable drug delivery devices have been developed in last decade utilizing the silicon microfabrication technology that was developed in integrated circuits (ICs) industries.

MEMS for Drug Delivery

Since the invention of silicon microfabrication technology in early 1960s, the IC has changed our world. During last 40 years, the semiconductor industry has come up with a fastest growing industry in our history. From a modest beginning, which allowed few transistors on a chip, we have reached an integration level of tens of millions of components in a square centimeter of silicon. The minimum feature size on silicon is reducing and thus the number of devices per square centimeter is increasing. Since the observation made in 1965 by Gordon Moore (45), co-founder of Intel, the number of transistors per square inch on integrated circuits had doubled every year since the integrated circuit was invented. Moore predicted that this trend would continue for the foreseeable future. In subsequent years, the pace slowed down a bit, but data density has doubled approximately every IS months, and this is the current definition of Moore's law.

This silicon fabrication technology was later extended to machining mechanical microdevices, which was later called MEMS. The pioneer work was done by Nathanson et al. in 1965 when they demonstrated the first micromachined structure to fabricate a free-standing gold beam electrode used in a resonant gate transistor (46). By late 1970s, there was an immense interest in silicon as a mechanical material (47,48). During 1980s and 1990s, many MEMS devices were fabricated, for example, micrometers (49–51), deformable mirrors (52,53), accelerometers (54–58), and comb-drive actuators (59).

In recent years, this fabrication technology has been extensively used for the development of microfluidic devices for biological and biochemical applications (these are called bio-MEMS) (60,61). Further, the integration of microfluidic devices and integrated circuits over the last decade has revolutionized the chemical and biological analysis systems, and has opened the possibility of fabricating devices with increased functionality and complexity for these applications (62–64). These tiny devices hold promise for precision surgery with micrometer control, rapid screening of common diseases and genetic predispositions, and autonomous therapeutic management of allergies, pain and neurodegenerative diseases (7). The development of retinal implants to treat blindness (65), neural implants for stimulation and recording from the central nervous system (CNS) (66), and microneedles for painless vaccination (67), are examples in which MEMS technology has been used. With microfabrication technology it is also possible to produce the novel drug delivery modalities with capabilities not present in the current systems. A variety of microfabricated devices, such as microparticles, microneedles, microchips, nanoporous membranes, and micropumps, have been developed in recent years for drug delivery applications (68–71). This section reviews various microfabricated devices. These have been categorized and described below as microfabricated devices for transdermal, oral, IV, and implantable drug delivery devices.

Microneedles for Transdermal Drug Deliver. Transdermal drug delivery is probably the most favored way of drug delivery since it avoids any degradation of molecules in the GI tract and first-pass effects of the liver, both of which are associated with the oral drug delivery, and eliminates the pain associated with IV injection (72–76). However, the major barrier for the transdermal delivery is the stratum corneum, the outermost dead layer of the skin. 1n human, it is 10–20 μm thick. A number of different approaches have been studied with two common goals: first is to disrupt stratum corneum structure in order to create "holes" big enough for molecules to pass through and the second goal is to develop microneedles that are long enough to provide transport pathways across the stratum corneum and short enough to reach nerves found in deeper tissues. These approaches include chemical–lipid enhancers (77,78), electric fields employing iontophoresis and electroporation (79), and pressure waves generated by ultrasound or photoacoustic effects (80,81).

MEMS technology has provided an alternative approach to transdermal drug delivery. The development of microneedles for transdermal drug delivery enhances the poor permeability of the skin by creating microscale conduits for transport across the stratum corneum (69,76). Needles of micron dimensions can pierce into the skin surface to create holes large enough for molecules to enter, but small enough to avoid pain or significant damage.

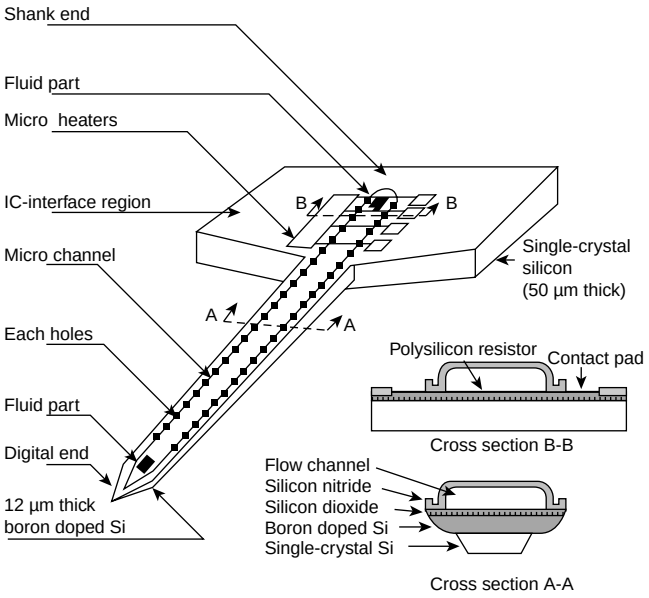


Figure 4. Schematic diagram of a silicon processed microneedles by Lin and Pisano (84).

Although the microneedles concept was proposed in the 1970s (82), it was not demonstrated experimentally until the 1990s (83). Since then, many different kinds of microneedles have been fabricated in several materials (e.g., silicon, glass, and metal). Further, these microneedles can be fabricated in-plane, where the needle lumen (flow channel) is parallel to the substrate surface, or out-of-plane, where the lumen is normal to the substrate. Some of these are summarized below.

Lin and Pisano (84) fabricated microneedles in silicon (Figs. 4 and 5). The primary structural material of these microneedles was silicon nitride, forming the top, and a bulk micromachined boron doped silicon base defined by etching the substrate in ethylenediamine pyrocatechol (EDP). This layer of silicon, which varied in thickness from $\sim 50\,\mu m$ at the shank to $12\,\mu m$ near the tip improved the structural strength. The lumen was defined by a sacrificial layer of phosphorous doped glass. These microneedles were 1–6 mm in length with lumens $9\,\mu m$ high and 30–50 μm wide.

The proximal ends of the microstructures had integrated polycrystalline silicon heater strips. The heater could generate bubbles, which were useful in pumping fluid down the lumen. Authors suggested that electrodes could also be patterned along the length of the needle by a slight process modification for the measurement of neural activity.

Other microneedles made out of polysilicon molding process were reported by Talbot and Pisano (85) (Fig. 6). The two halves of the mold are produced by bulk micromachining of silicon wafers followed by deposition of a $2\,\mu m$ phosphosilicate glass (PSG) release layer. The two halves are temporarily bonded together under nitrogen ambiance at 1000 °C. After bonding, a $3\,\mu m$ layer of amorphous silicon is deposited by LPCVD through access holes in the top mold wafer. The mold along with the deposited film was then annealed at 1000 °C. Deposition and annealing steps were repeated until the desired thickness of 12–18 μm was

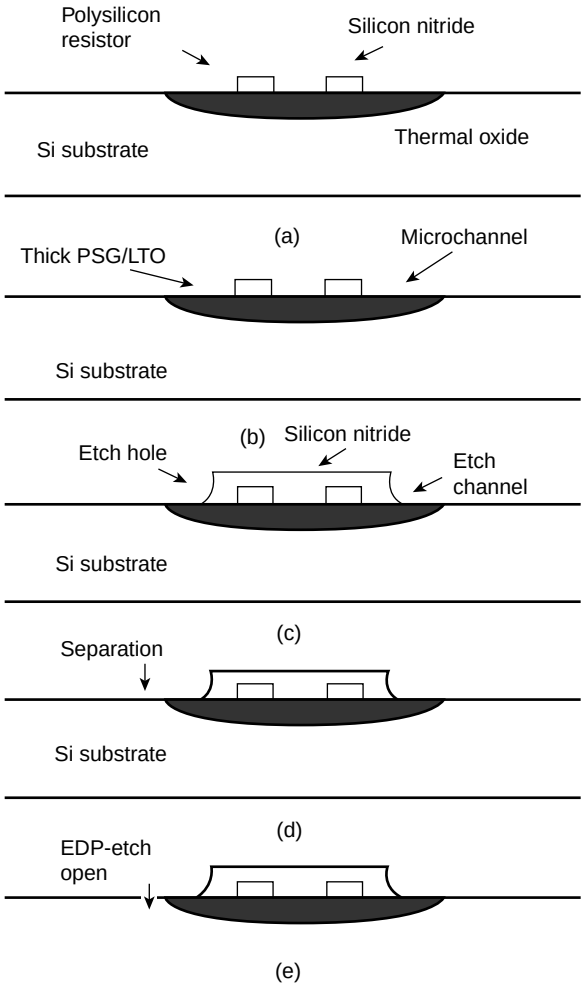


Figure 5. Process sequences of a silicon processed microneedles by Lin and Pissano (84).

obtained. Plasma etching was used to remove the polysilicon coating the funnel-shaped access holes in the top mold layer. The devices were released from the mold by etching in concentrated hydrofluoric acid, which selectively attacks the PSG. The mold could be used repeatedly by redepositing PSG, the release layer in order to minimize the cost. The resulting polysilicon microneedles are 1–7 μm long, 110–200 μm rectangular cross-section, and submicrometer tip radii.

Brazzle et al. (86–88) fabricated metal microneedles using a micromolding process. The fabrication process of the microneedles developed by Papautsky is shown in Fig. 7. A P+ etch stop layer was formed and backside anisotropic etching in KOH was performed to define a thin membrane. The lower wall of the microneedles consisted of deposited and patterned metal layers. A thick layer (5–50 μm) of positive photoresist was then spin coated and lithographically patterned on the top of the lower metal walls.

The dimensions of this sacrificial layer precisely defined the cross-section of the lumen. After sputter deposition of a Pd seed layer, the thick metal structure walls and top of the microneedles were formed by electrodeposition. The sacrificial photoresist was removed with acetone and the P+

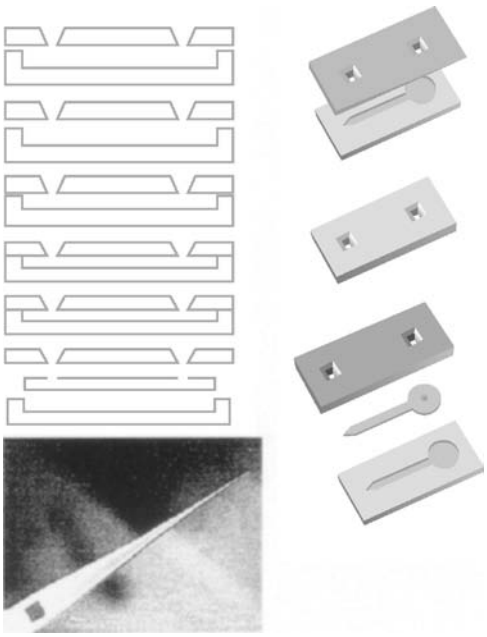


Figure 6. Microneedles fabricated from a polysilicon molding process using two silicon wafers (85).

membrane was etched away in an S176 plasma, resulting in a one-dimensional (1D) array of hollow microneedles released from the substrate.

Out-of-plane array of microneedles were fabricated by Stoeber and Liepmann (89,90). The fabrication process is summarized in Fig. 8. A double-sided polished wafer was oxidized. The lumen was etched through the wafer by plasma etching following a mask patterned at the backside. A silicon nitride film was then deposited across the backside and into the etched holes. Needle locations were photolithographically defined on the top surface on the wafer. The microneedle shaft was created by isotropic and etching on the silicon substrate. The isotropic etching forms a microneedle with a gradually increasing diameter along the shaft. By displacing the circular pattern for isotropic etching from the center of the lumen, a pointed needle shape was obtained. These microneedles were 200 μm tall, with a base diameter of 425 μm tapering to a 40 μm lumen. Individual needles were 750 μm apart. Fluid injection was demonstrated by delivering under the skin of a chicken thigh, a depth of ~100 μm.

Solid microneedles with no lumen were demonstrated by Henry et al. (76,91). The fabrication steps are shown in Fig. 9. A chrome mask was deposited on a silicon wafer and patterned into dots that have a diameter approximately equal to that of the base of the desired needles. A deep reactive ion etching was performed. Etching proceeded until the mask fell off from undercutting. The region protected by chromium remained and eventually became the microneedles. The tapering on the microneedles were controlled by adjusting the degree of anisotropy in the etch process. The resulting microneedles were 150 μm tall, and could be fabricated in dense arrays.

Gardeniers et al. (92) fabricated out-of-plane microneedles that employed reactive ion etching from both sides on

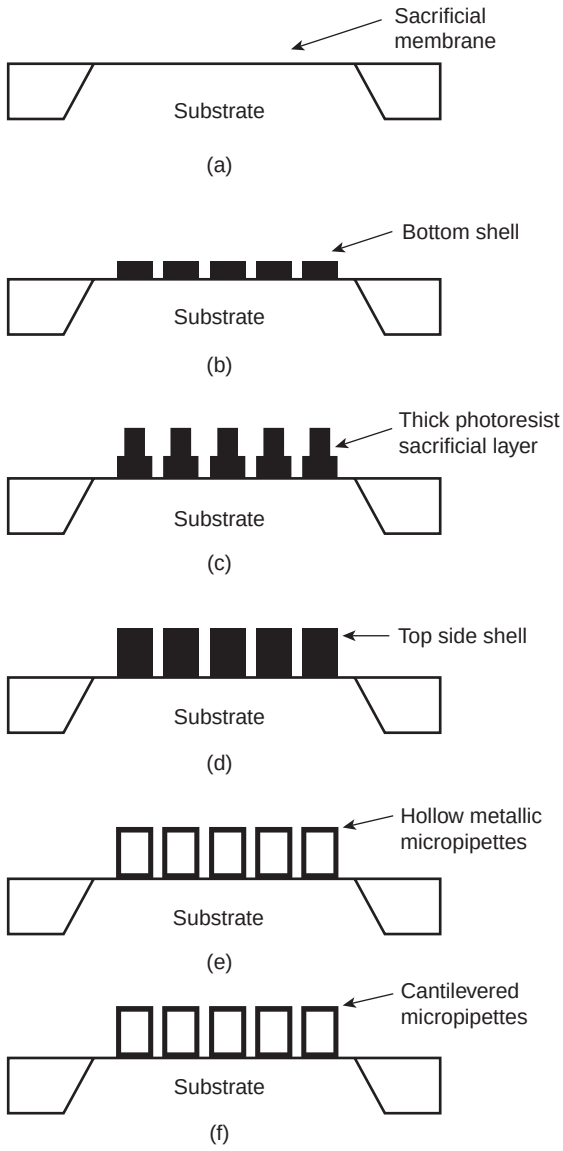


Figure 7. Fabrication process of a hollow in-plane microneedles (86).

a (100) oriented silicon wafer (Fig. 10). A hole (feature a in Fig. 10), which becomes lumen and a slot (feature b) that defines the position of the needles tip and needle sidewalls, was etched at the top surface. These structures were aligned to the crystallographic planes of silicon so that anisotropic etching performed later produces the slanted structure. The connecting lumen (feature c) was etched from the back side. The substrate, including the sidewalls of the etched features were coated with the chemically vapor deposited silicon nitride. The nitride was removed form the top surface of the wafer and etched in KOH. The etch left a structure defined by (111) plane in the areas where the nitride slot walls were concave, but where the mask was convex, the etch found all of the fast etching planes. The nitride mask was stripped at the end of the process.

Microneedles have also been developed for gene delivery. One such structure was fabricated by Dizon et al. (93).

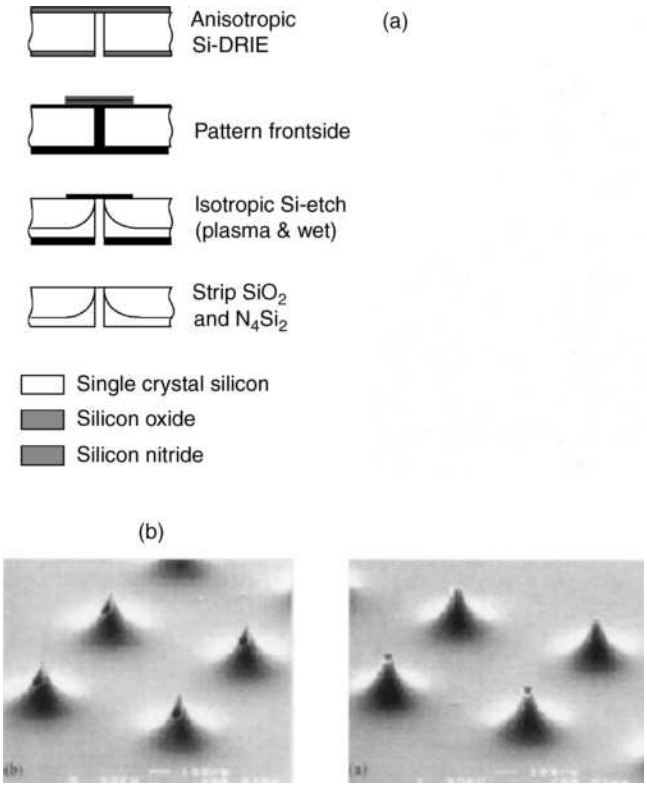


Figure 8. Out-of-plane array of microneedles. (a) Fabrication step, (b) Symmetric and asymmetric needles (90).

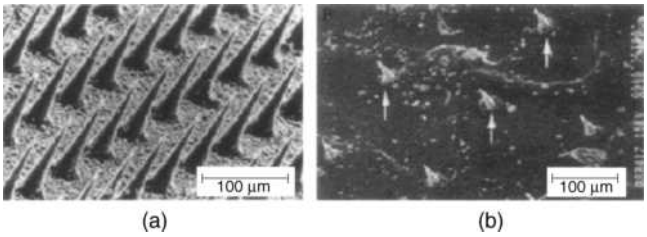


Figure 9. (a) Scanning electron micrograph (SEM) of microneedles made by reactive ion etching technique. (b) Micro-needle tips inserted across the epidermis. The underside of the epidermis is shown, indicating that the microneedles penetrated across the tissue and that the tips were not damaged. Arrows indicate some of the microneedle tips (91).

This structure was fabricated in dense array using a silicon bulk micromachining technique (Fig. 11), called Microprobes. The microprobes were ~80 μm high topped by a wedge-shaped tip with a radius of curvature <0.1 μm. The facets of the microstructure were fabricated utilizing fast etching (411) planes, produced by convex-corner undercutting in an anisotropic etching solution and a square mask. These microprobes can be coated with genes and pressed into cells or tissues. The sharp tips penetrate into cells and affect the transport of genetic material. Successful expression of foreign genes using this technique has been demonstrated in the nematode *Caenorhabditis elegans* (94), tobacco leafs (95), and mammalians cells (96).

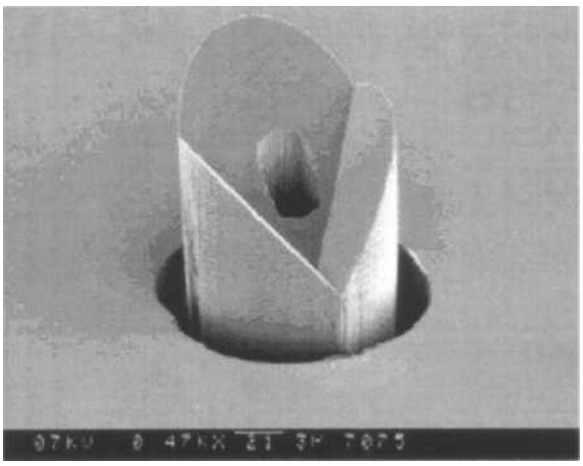


Figure 10. Out-of-plane microneedles were fabricated that employed reactive ion etching from both sides on a (100) silicon wafer (92).

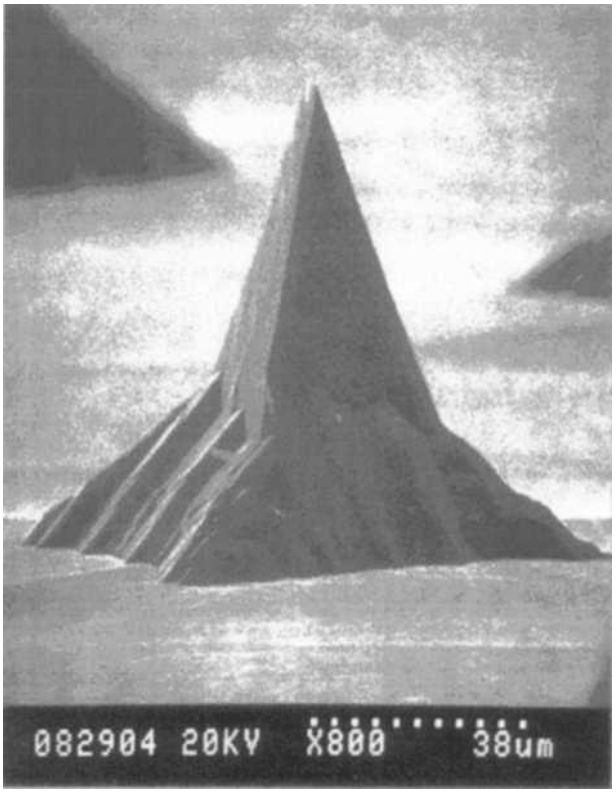


Figure 11. Solid silicon microprobe for gene delivery (93).

Mikszta et al. (67) used silicon micromachining technology for DNA and vaccine delivery to the epidermis. Figure 12 shows the microstructure, which they call microenhancer arrays (MEAs), that was fabricated by isotropic chemical etching of silicon wafers.

On the whole, existing microneedle-based drug delivery devices offer several advantages, such as the ability to inject drugs directly through the stratum corneurn at reproducible and accurate depth of penetration, minimal

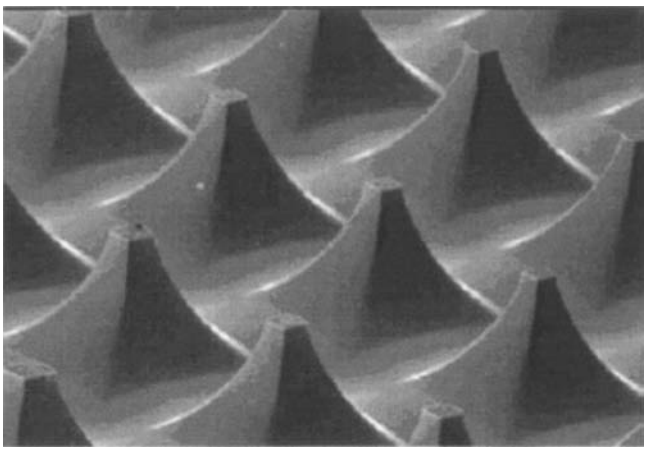


Figure 12. Silicon microenhancer arrays (MEAs) for DNA and vaccine delivery (67).

pain, and on-board ability to probe or sample the same device. Nevertheless, local irritation and low mechanical stability are some of the potential drawbacks that demand further investigation for alternate fabrication techniques and materials. Furthermore, improved fluid flow models that determine the most effective structural. fluidic, and biological design considerations for a given delivery application continue to be required.

Microparticles for Oral Drug Delivery. Oral route is a preferred method of drug delivery because of its ease of administration and better patient compliance. However, oral delivery of peptides and proteins has remained an illusive goal to date. The two main reasons why it is currently impossible are (*1*) destruction or inactivation due to enzymatic action, and the acidity of the upper GI tract; and (*2*) physiological permeation barrier, opposing penetration of large biological molecules through intestinal walls (71). These are mucosal layers and the tight junctions connecting intestinal epithelial cells, which restrict the possible passageways to be transcellular, and thus expose the diffusing biomolecule to enzymatic degradation. This method of drug delivery, therefore, leads to unacceptably low oral bioavailability. Consequently, various approaches based on the use of protective coatings (97), targeted delivery (98), permeation enhancers (99), protease inhibitors (100), and bioadhesive agents (101–103) have been explored in recent years. While all of these methods have been shown to increase the oral bioavailability of drug molecules, none of them offer a complete solution for adequate and safe oral delivery of peptides and proteins.

Microfabrication technology may address the shortcomings of the current oral drug delivery systems by combining the aforesmentioned approaches in a single drug delivery platform. Fabrication of microparticles of silicon and silicon dioxide has been conceptualized and demonstrated to achieve this (104–106). Unlike other spherical drug delivery particles, microfabricated devices may be designed to be flat, thin, and disk–shaped to maximize contact area with the intestinal lining and minimize the side areas exposed to the constant flow of liquids through the intestines (107). The size of the particles (within thickness of 0.1–5 nm and diameters of 1–100 μm) can be selected to have good contact with the undulations of the intestinal wall and large enough to avoid endocytosis of the entire particle. Permeation enhancers, such as bile salts and metal chelating agents, can be added to loosen the tight junctions of the intestinal epithelium. Aprotinin, or other enzyme inhibitors, can also be added to protect the macromolecule from intestinal degradation. In addition, one can selectively attach bioadhesive agents onto the device surface using relatively simple surface chemical modification strategies. By replacing the specific markers attached to the microparticles, specific cell types and tissues can be targeted for therapy as well as imaging. This would allow for the high concentration of drug to be locally delivered while keeping the systemic concentration at a low level. Finally, these devices can have multiple reservoirs of desired size to contain not just one, but also many drugs–biomolecules of interest (108).

iMEDD Inc. in collaboration with Ferrari et al. (109) developed Oral MEDDS (Oral Micro-Engineered Delivery Devices), novel porous silicon particles that can be used as oral drug delivery vehicles. The microparticle dimensions ranged from $150 \times 150 \times 25$–$240 \times 240 \times 25$ μm with a pore distribution of 20–100 nm (Fig. 13). Once prepared, the particles could be loaded with a liquid drug formulation through simple capillary action. Interstitial air is removed by vacuum aspiration, and the formulation is dried completely using vacuum or freeze-drying. OralMEDDS particles have been designed to target intestinal epithelial cells, adhere to the apical cell surface, and deliver a drug formulation containing a permeation enhancer that would open the local tight junctions of the paracellular transport pathway. The absorption of macromolecules and hydrophilic drugs, which are unable to undergo transcellular transport across lipid membranes, is largely restricted to this paracellular route. Therefore, the intestinal absorption of orally administered water-soluble drugs can be greatly enhanced through the utilization of OralMEDDs particles (110).

Micromachined silicon dioxide and PMMA microparticles designed by Desai and co-workers (70,111) can be best described as microparticles with reservoirs (Figs. 14 and 15). These microparticles are adaptable for use as a bioadhesive controlled release oral drug delivery system. Silicon dioxide microparticles were created by growing a thermal oxide under wet conditions followed by low pressure chemical vapor deposition to deposit a sacrificial layer of

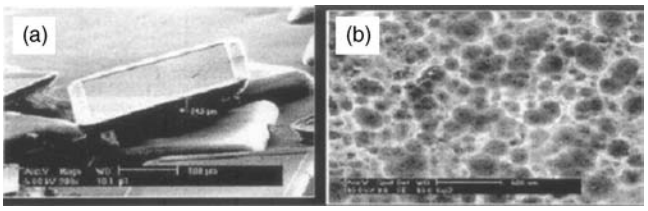


Figure 13. Scanning electron microscopy images of a porous silicon particle: (a) Demonstrating the thickness. (b) Particle demonstrating the pore size distribution of −20–100 nm (110).

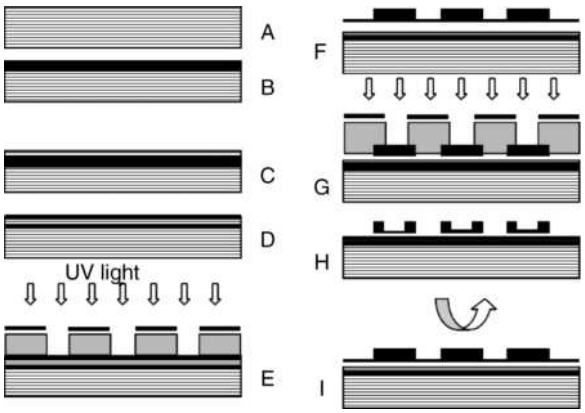


Figure 14. Process flow of the silicon dioxide microparticles (111).

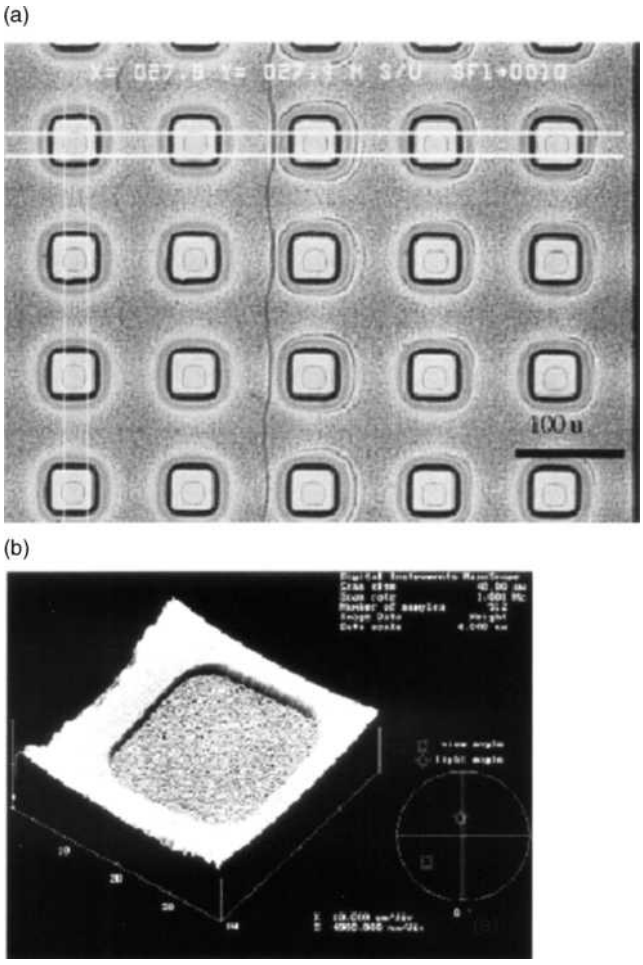


Figure 15. (a) 50 μ particles with 25 μ reservoirs. (b) AFM image of the particles (25 reservoir, 50 m particles) (111).

polycrystalline silicon (111). Next, a layer of low temperature silicon dioxide (LTO) was deposited to form the device layer. Positive lithography was carried out to define the shape of the device reservoir. A reactive ion etch (RIE) with S176 and 02 was used to fabricate the actual reservoir in the LTO device layer and any remaining photoresist was then removed in negative photoresist remover. Negative lithography was carried out to define the device bodies. Reservoir features on the mask were aligned to the photomask features using front-side alignment. The unmasked area of the LTO layer was etched using RIE and the remaining photoresist was removed. These microdevices were then released into KOH solution by etching the sacrificial polysilicon layer. The particles were uniform and semitransparent due to their polycrystalline nature. Later, a lectin–biotin–avidin complex suited for binding these microparticles to the intestinal mucosa was developed. The Caco-2 cell line was used to examine the bioadhesive properties of microparticles *in vitro*. Bioadhesive silicon dioxide microparticles demonstrated greater adherence to Caco-2 cells as compared to unmodified particles.

Poly(methyl methacrylate) (PMMA) particles were fabricated by spinning PMMA (device layer) on to a clean silicon wafer (70). Positive lithography was carried out to define the device bodies followed by a reactive ion etching to carve the devices. Then a second mask positive photolithography was carried out to carve the device reservoir. The process flow of the device fabrication is shown in Fig. 16. The dimensions of the reservoir can be altered by changing the masked area and their depth can be modified by changing the time and/or flow rate of plasma in the RIE. By creating smaller reservoirs, a series of multiple reservoirs can be etched into the particles to create separate reservoirs for a combination of drugs or permeation enhancers. Since the PMMA is adherent to the surface of silicon by linkage to the native oxide layer, the wafer was soaked in basic solution to break this bonding and immediately release the particles. Bioadhesive properties were introduced to microfabricated PMMA microdevices by attachment of lectins, a group of proteins capable of specifically targeting cells in the GI tract. In this process, the PMMA microdevices were chemically modified by aminolysis to yield amine-terminated surfaces. Avidin molecules were covalently bound to the surface of the particles using a hydroxysuccinimide-catalyzed carbodiimide reagent and then incubated in an aqueous solution of biotinylated lectin. The bioadhesive characteristics of lectin-modified microdevices were successfully demonstrated *in vitro*.

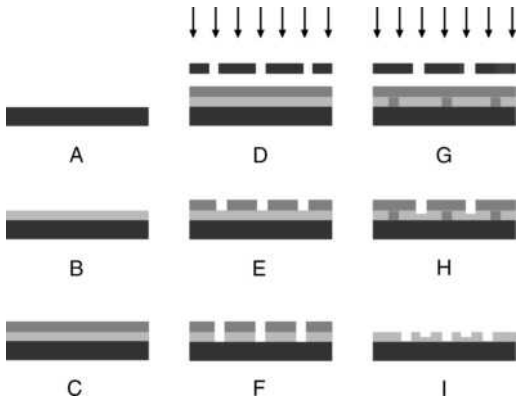


Figure 16. Process flow of PMMA microdevices with reservoir for oral drug delivery (70).

Microparticles for Intravenous Drug Delivery. The same microfabrication technology that has been used quite extensively for the fabrication of particles for oral drug delivery can be employed to develop precisely sized and shaped microparticles with high specific targeting abilities for IV delivery, especially for the treatment of diseases where oral and transdermal delivery are not effective. As an example, systemic chemotherapy using cytotoxic or biological treatment is the only treatment available for many patients with advanced metastatic cancer. While many tumors respond to initial courses of chemotherapy, after multiple courses and drugs, cancer cells become resistant to further therapy. In addition, growth of metastatic tumors is supported by factors, that are secreted by tumor cells themselves and cause angiogenic leaky vessels to grow. One strategy for preventing or treating metastatic tumors is to intervene in the process of angiogenesis by destroying the blood vessels that supply tumor cells rather than the tumor cells themselves (112). In such cases, precisely sized and shaped microparticles especially designed for IV delivery of cytotoxic biomolecules–drugs to the microvasculature of tumors with an improved safety profile could be employed. These have been described below.

Nonporous Microparticles. First generation of nonporous (solid) microparticles of silicon and silicon dioxide suitable for IV drug delivery (16,113), were rectangular shaped with thickness of 0.9 μm, and varied from 1 to 3 μm in length and width (Fig. 17). These microparticles were treated with amino- and mercaptosilanes, followed by coupling to human antibody (IgG) by using the heterobifunctional cross-linker succinimidyl 4-(*N*-maleimidolmethyl)-cyclohexane-l-carboxylate, to demonstrate their capability toward specific attachment of bioadhesive agents. These solid microparticles and their next generations are currently being explored for drug delivery and bioimaging applications (114).

Nanoporous Microparticles. Currently, porous silicon has begun to receive significant attention for biomedical usage. Nano- and microparticulates of this material have immense potential to be clinically and diagnostically significant both *in vivo* and *ex vivo* (115,116). Li et al. (113) demonstrated the incorporation, characterization, and release of cisplatin [*cis*-diammine dichloroplatinum(Il)], carboplatin [cis-diammine (cyclobutane-1,1-dicarboxylato) platinum(Il)], and Pt(en)C12 [ethylenediamminedichloro platinum(Il)] within layers of calcium phosphate on porous Si–Si substrates for bone cancer treatment.

Superior control over particle dimensions, pore size, pore shape, and loading capacity is critical for microparticles for IR drug delivery (17,117). iMEDD Inc. has developed nanoporous microparticles (called IV-MEDDS or NK-MEDDS, where NK denotes the fact that the particles mimic Natural Killer cells) to treat systemically accessible solid tumors, specifically the multiple lesion sites associated with metastatic disease (71). The approach here is to kill the circulatory accessible endothelial cells that support the existing tumor capillaries using micromachined asymmetrical particles, that is, the top face of the particle contains a pore loaded with cytotoxic drugs, which is plugged with an erodible gelatinous material and layered with chemically grafted ligand (including growth factors, e.g., FGF, EGF and VEGF to bind endothelial or tumor cell receptors or folate and tumor-targeting RGD peptides to bind $\alpha_v\beta_3$ with high affinity) for targeting and protection. Designed to mimic the behavior of NK cells, a potent cytolytic agent, such as bee venom-derived melittin, can be plugged with a material designed to erode in 1–48 h. After injection, the particles circulate within the bloodstream for several minutes to several hours after that they are removed from the body's immune system. Bound particles should release their contents in the vicinity of the tumors and cause lysis and death of the target endothelial cells. Melittin peptides released by particles elsewhere in the body and not bound to endothelial target, are inactivated by binding to albumin and thus are not toxic to normal cells (71).

Based on the above-mentioned concept, Cohen et al. (118) prepared micron-sized particles with nanometer-sized pores out of porous silicon and porous silicon dioxide. The fabrication steps are shown in Figs. 18–20. The particles were fabricated with precise shapes and sizes. The size and thickness of these particles could be altered by changing the dimensions of the photolithography mask, the anodization time, and the electropolishing time. The

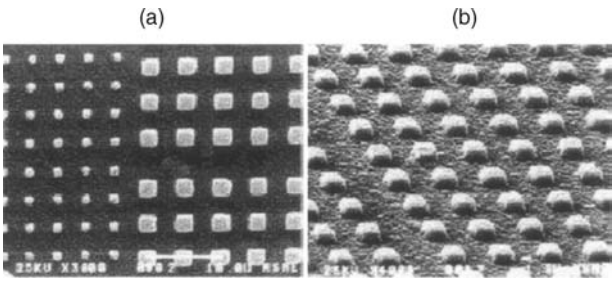


Figure 17. Scanning electron micrographs of microparticles. (a) Dimensions are 2.2×2.1 (± 0.1) mm for the larger particles, and 1.2×1.1 (± 0.05) mm for the smaller ones. (b) Shows tilted view of larger microparticles (104).

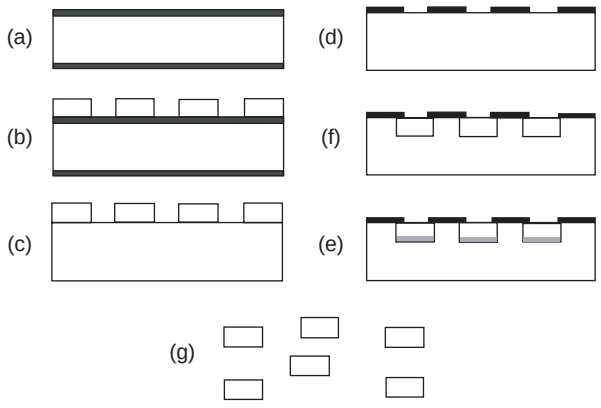


Figure 18. Fabrication details for porous silicon particles. (a) LPCVD silicon nitride deposition. (b) Photolithography. (c) Dry etch silicon nitride. (d) Piranha. (e) Anodization of silicon. (f) Electropolishing. (g) Particle release (118).

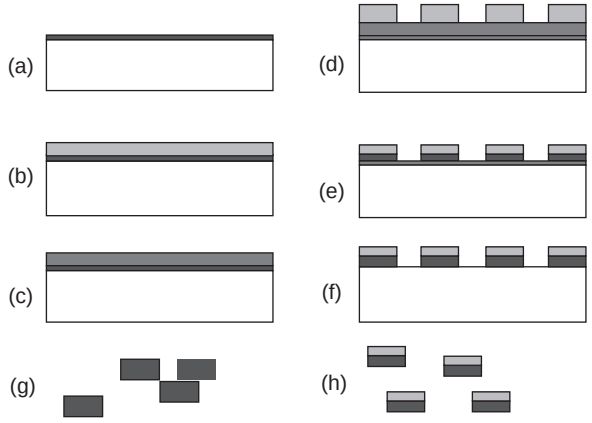


Figure 19. Process flow of porous SiO_2 particle fabrication. (a) Aluminum deposition. (b) Spun on mesoporous oxide film. (c) Baked mesoporous oxide film. (d) Photolithography. (f) Particle release in pirana. (g) Uncapped particles. (h) Particles capped with photoresist (118).

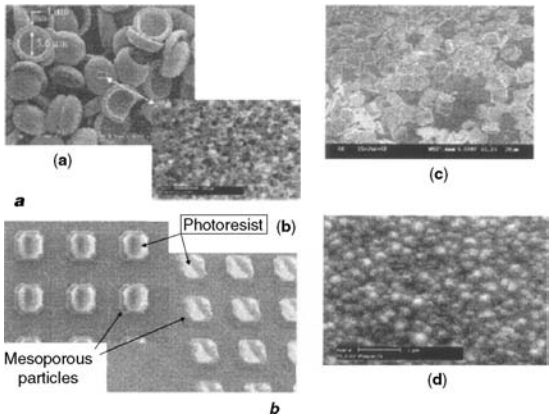


Figure 20. (a) SEM images of released porous silicon particles: Top image shows the shape and size of the particles. Bottom image demonstrates pores in the size range of 20–100 nm. (b) SEM images of mesoporous silicon oxide particles on wafer: (*a*) flat. (*b*) 45 tilt. (c) SEM images of released porous silicon dioxide particles. (d) SEM images of released porous silicon particles (118).

porous silicon dioxide particles were 4.7 μm squares with a thickness of 1.0 μm. The porosity of silicon dioxide particles was 52.5%. In order to determine safe particle size and concentration for IV drug delivery, a safety study was performed using solid silicon particles with various shapes, squares and circles, and varying sizes, 2, 5, and 10 μm. Results indicated that at concentrations of 1×10^7 particles per mouse, particles of size 2 and 5 μm safely circulate throughout the vasculature. No mice survived for any length of time when they were injected with 10 μm particles. Work is underway to demonstrate the coupling of EGF to porous dioxide particles that will allow for the particles to bind to the cells that express EGF receptors.

Smith et al. (114) prepared novel, controllably dual-sided, symmetric particulates of porous silicon from a polysilicon precursor. These particulates are precisely monodisperse on the scale of 1 μm (diameter and thickness) and may enable unidirectional flow of transported drugs, proteins–peptides, nucleic acids, and so on. They may also facilitate controllably different intraparticle surface chemistries, and therefore potentially different types of antibodies, proteins, and so on, can be present on the same particle.

MEMS for Implantable Drug Delivery Devices. Implantable devices are preferred for the therapies that require many injections daily or weekly. The requirement and advantages of an implantable drug delivery device has been discussed above in greater detail. These devices can either be implanted into the human body or placed under the skin, consequently reducing the risk of infection by eliminating the need for frequent injections. Most of the implantable microsystems are expected not to cause pain or tissue trauma owing to their small size and are often virtually invisible. The advances in microfabricated implantable drug delivery device have been reviewed below.

Microreservoirs. Silicon microfabrication technology has been used to develop drug delivery device consisting of an array of microreservoirs (68,119,120) (Fig. 21). This device is currently being developed by MicroCHIPS, Inc., for use as external and implantable systems for the delivery of proteins, hormones, pain medications, and other pharmaceutical compounds (117). Each dosage is contained in a separate reservoir that is covered with a gold membrane. The membrane gets dissolved in the presence of chloride ions when anodic voltage is applied to the membrane of interest. This causes the membrane to weaken and rupture, allowing the drug within the reservoir to dissolve and diffuse into the surrounding tissues. This device allows the release of a potent substance in a pulsatile manner. Each microreservoir can be individually filled, so multiple substances can be delivered from a single MEMS device. Release of fluorescent dye and radiolabeled compounds has been demonstrated from these microreservoir devices *in vitro* in saline solution and serum (68).

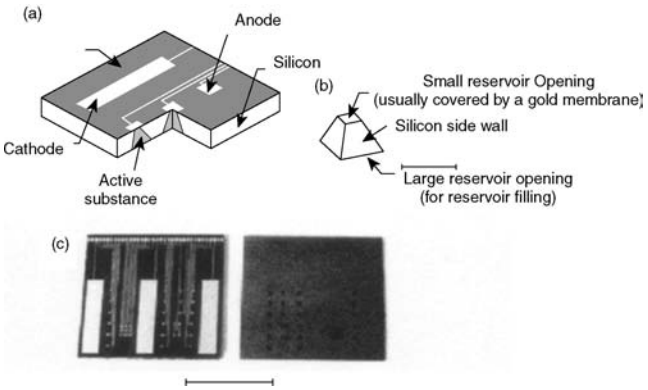


Figure 21. A schematic of a silicon microchip for controlled release. (a) Cut-away section showing anodes, cathodes, and reservoirs. (b) Shape of an individual reservoir. (c) Photograph of a prototype microchip: the electrode-containing frontside and the backside with openings for filling the reservoirs (15).

The release studies from this device demonstrated that the activation of each reservoir could be controlled individually, creating a possibility for achieving many complex release patterns. Varying amounts of chemical substances in solid, liquid, or gel form could be released into solution in either a pulsatile, a continuous, or a combination of both manners, either sequentially or simultaneously from a single device. Such a device has additional potential advantages including small size, quick response times, and low power consumption. In addition, all chemical substances to be released are stored in the reservoirs of the device itself, creating a possibility for the future development of autonomous devices. A microbattery, multiplexing circuitry, and memory could be integrated directly onto the device, allowing the entire device to be mounted onto the tip of a small probe, implanted, swallowed, integrated with microfluidic components to develop a laboratory-on-a-chip, or incorporated into a standard electronic package, depending on the particular application. Proper selection of biocompatible device materials may result in the development of an autonomous, controlled-release implant or a highly controllable tablet for drug delivery applications (68).

Nanoporous Silicon Membranes. Silicon nanopore membranes were developed by Ferrari and co-workers for application as immunoisolating biocapsules, and for molecular filtration (121–123). These membranes were shown to be sufficiently permeable to oxygen, insulin, and glucose, while at the same time impermeable to larger proteins, such as immunoglobulin G (IgG), which might lead to destruction of the transplanted cells (124). Since the diffusion through these membranes is linear, they can also be used for sustained drug delivery. This is currently being developed by iMEDD, Inc. (71,109). Over the years, nanopore technology has undergone continued improvements. Nevertheless, the basic structure and fabrication protocol for the nanopores has remained the same. The membrane area is made of thin layers of polysilicon, silicon dioxide, and/or single crystalline silicon depending on the design employed. The strategy used to make nano-size pores was based on the use of a sacrificial oxide layer sandwiched between two structural layers, for the definition of the pore pathways. The first design of nanoporous membranes consisted of a bilayer of polysilicon with L-shaped pore paths. The flow path of fluids and particles through the membrane is shown in (Fig. 22a) (125). As shown, fluid enters the pores through openings in the top polysilicon layer, travel laterally through the pores, make a 90° turn, and exit the pores through the bottom of the pore where both the top and bottom polysilicon layers lay on the etch stop layer. While this design performed well for preventing the diffusion of the larger, unwanted immune system molecules, its L-shaped path slowed down and, in some cases, prevented the diffusion of the smaller molecules of interest. The pores in this design were fairly long, which led to the slow diffusion of the desired molecules. Also, because of the large area per pore, it was difficult to increase the pore density and thus the diffusion rate. The next design had an improvement in the production of short, straight, vertical pores through a single-crystal base layer (Fig. 22b and c). This design had the advantage of direct flow paths. This direct path allows the smaller molecules of interest to diffuse much quicker through the membrane, while still size-separating the larger molecules. To further improve the reliability of the nanoporous membranes, several basic changes were made in the fabrication protocol from the previous membrane design to eliminate problems with the diffused etch stop layer (126). This design also incorporated a shorter diffusion path length, based on the thicknesses of the two structural layers. The design of a new membrane fabrication protocol incorporated several desired improvements: a well-defined etch stop layer, precise control of pore dimensions, and a lower stress state in the membrane. The new protocol also increased the exposed pore area of the membranes. The nanoporous membranes have been studied extensively for the use of drug delivery and the results are very encouraging.

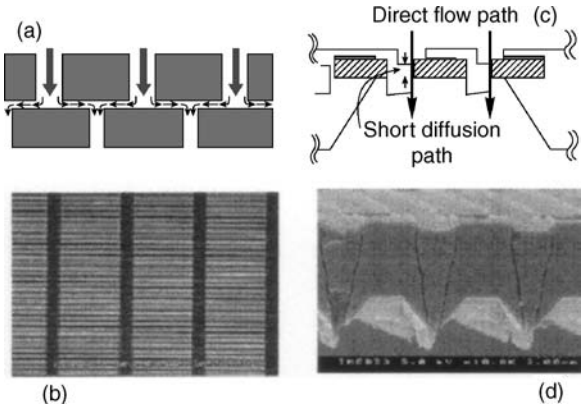


Figure 22. (a) Flow path through MI filters, with lateral diffusion through the nanopores defined by sacrificial oxide. (b) Cross-section of M2 design showing direct flow path. Scanning electron micrographs of microfabricated membrane: (c) top view detail; (d) side view detail (126).

Zero-Order Kinetics through Nanoporous Membrane. *In vitro* bovine serum albumin (BSA) release data through 13 nm pore is shown in Fig. 23. The experimental results show zero-order release profile (zero-order kinetics). Note that the zero-order kinetics does not follow Fick's law. Fick's laws are usually adequate to describe diffusion kinetics of solutes from a region of higher concentration to a region of lower concentration through a thin, semi-permeable membrane. But, when the size of the membrane pores approaches that of the solute, an unexpected effect may occur, which deviate substantially from those predicted by Fick's laws. Diffusion of molecules in microporous media, such as zeolites, has led to experimental evidence of such unusual phenomena as molecular traffic control and single file diffusion (SFD) (127,128). Theoretical treatments and simulations suggest that in the case of SFD, solute molecules of equal size cannot pass each other in pores that approximate the dimensions of the molecule itself, regardless of the influence of concentration gradient, and thus their initial rate of movement (or flux) is underestimated by Fick's law (129–133).

The microfabricated nanopore channels are of molecular size in 1D, and therefore non-Fickian diffusion kinetics

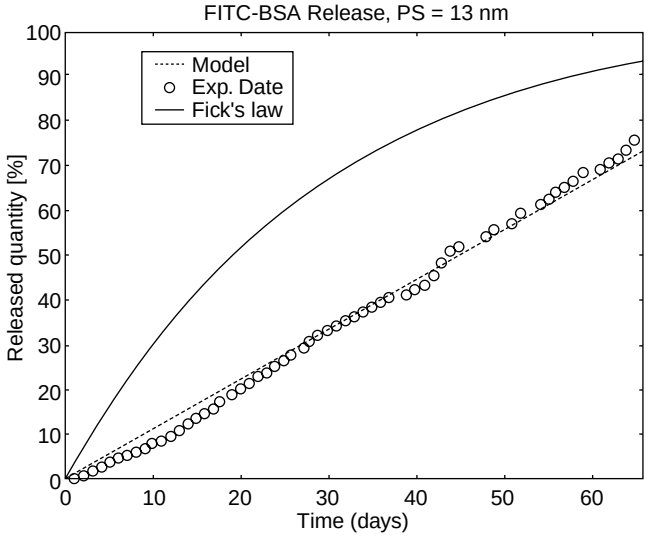


Figure 23. *In vitro* diffusion kinetics of fluorescein isothiocyanate (FITC) labeled BSA through 13 nm pore size: experimental data (o), Fick's law prediction (–), model-based simulation (. . .).

is observed. The observations are consistent with the diffusion reported for colloidal particles confined in closed 1D channels of micrometer scale where particle self-diffusion is non-Fickian for long time periods and the distribution of particle displacements is a Gaussian function (128). Zero-order flux is observed when a chamber filled with a solute is separated from a solute-free external medium by channels that are only several times wider than the hydrodynamic diameter of the individual molecules. The basic principle of diffusion as a mixing process with solutes free to undergo Brownian motion in three dimensions (3D) does not apply since in at least 1D solute movement within the nanopore is physically constrained by the channel walls. Experimental observations of colloidal particles in a density matched fluid confined between two flat plates reveal that particle diffusion becomes anisotropic near the interface; in this case leading to hindered diffusion as a consequence of constrained Brownian motion and hydrodynamic drag effects at distances close to the walls (134). In the case of nanoporous membranes, it is not entirely certain that the ordering of solutes imposed by the nanopore geometry will be as strict as true cylindrical pores, nor that the sequence of particles passing through the nanopores under the influence of the concentration gradient will remain unchanged over the time required to travel the 4 m length of the channel; particles could conceivably pass each other laterally. Whether a consequence of a SFD-like phenomenon or drag effects (or a combination of both), the nanopore membrane is rate limiting and, if properly tuned, restricts solute diffusion to a point that flux rates across the membrane are entirely independent of concentration gradient and follow zero-order kinetics.

In order to achieve further insight in the mechanisms involved in nanochannel diffusion, an experimental phenomenon in mathematical terms, thus yielding to the creation of a dynamical model, which makes it possible to simulate the diffusion experiments and fit the related data,

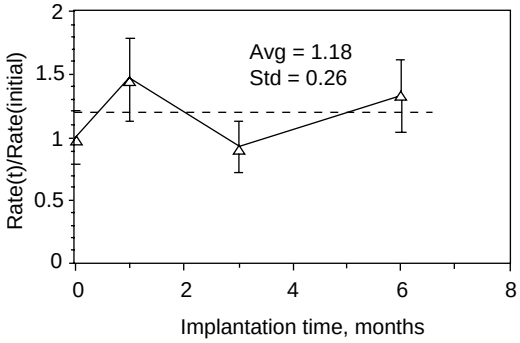


Figure 24. Ratio of post-preimplantation glucose diffusion rates.

is being investigated. A detailed description of such model is presented in Ref. 135.

Biocompatibility of Nanoporous Membranes. *In vivo* membrane biocompatibility was evaluated using glucose as a model molecule. Figure 24 shows the ratio of post-explantation glucose diffusion rate compared to its initial value. There was no noticeable change in glucose diffusion rates pre- and postimplantation illustrating that the silicon membranes did not foul over a 6-month implantation period. The membrane was placed on a titanium capsule and the entire assembly was placed subcutaneously in mice. The assembly was removed after 7 days and examined visually. There was no visible evidence of tissue binding to the surface. Figure 25 shows a photograph of the implant site after 30 days of implantation. As can be seen, only a thin vascular capsule forms around the implant as opposed to the avascular fibrous capsule. This minimal tissue response is supposed to be responsible for the comparable pre- and postimplantation glucose diffusion rates observed in this investigation.

Sandwich Design Filter. Nanochannels fabricated between two directly bonded silicon wafers were also developed for the applications as immunoisolating biocapsules, and molecular filtration (125,136–139). These devices possess high mechanical strength since the filtration occurs at the interface of two bonded silicon wafers instead of through a 1–10 μm thick membrane (in the case of silicon nanopores membrane). Well-developed bulk microfabrication

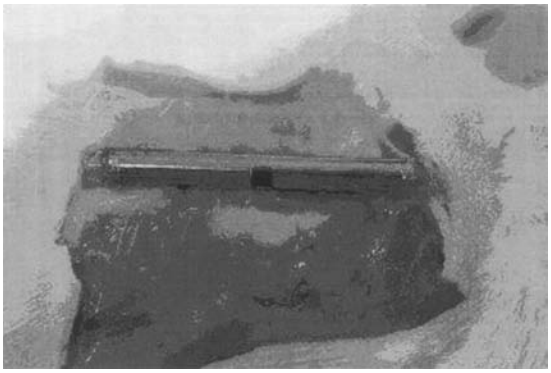


Figure 25. Photograph of implantation site after 30 days *in vivo*.

technology was used to fabricate these devices. With the use of a silicon dioxide sacrificial layer, pore sizes as small as 40 nm were fabricated with size variations $< 4\%$. It was already established in the case of silicon nanopore membranes that the diffusion of molecules though nanopores is constant, and therefore the sandwich design filter can also be used for sustained drug delivery applications.

MOLECULAR DRUG DELIVERY SYSTEMS

This type of carrier, including cryptands, calixarenes (140), cyclophanes (141), spherands, cyclodextrins, and crown ethers, carry out chemical reactions that involve all intramolecular interactions where covalent bonds are not formed between interacting molecules, ions or radicals. Most of these reactions are of host–guest type.

Cyclodextrins: General Information

Between the several drug delivery systems, the molecular carrier have aroused great interest in the scientific world. Compared to all the molecular hosts mentioned above, cyclodextrins (CDs) are most important. As a result of molecular complexation phenomenon CDs are widely used in many industrial products, technologies, and analytical methods.

The CDs represent the more important molecular carrier today, in fact they are already strongly present in commerce for various types of drugs. Cyclodextrins have been discovered in the nineteenth century. They were produced for the first time by Villiers in 1881 by digesting the starch with Bacillus amylobacter, but only in 1903 was the cyclic structure of these compounds demonstrated by Schardinger.

Chemically, CDs are cyclic oligosaccharides, consisting of (α-1,4)-linked α-D-glucopyranose units. They are produced as a result of an intramolecular chain splitting reaction from degradation of starch by enzymes called cyclodextrins glucosyltransferases (CGTs) (142). In times past, only small amounts of CDs were generated and high production costs prevented their industrial application, but now most of the CGT genes have been cloned making the large scale production of this kind of carrier low cost.

The CDs are characterized by the presence of a lipophilic central cavity and a hydrophilic outer surface. The glucopyranose units are in the form of a chair and, for this reason, the CDs may be represented as a truncated cone. The OH groups are oriented with the primary hydroxyl groups of the various units of glucose on the narrow side of the cone and the secondary OH groups at the larger edge. The lipophilic character of the central cavity is determined by skeletal carbons and ethereal oxygens.

The CDs may contain even >15–16 glucopyranose units, but the must abundant natural CDs are α-cyclodextrin (α-CD), β-cyclodextrin (β-CD), and γ-cyclodextrin (γ-CD) containing six, seven, and eight glucopyranose units, respectively.

The CDs are chemically stable in alkaline solutions, but are susceptible to hydrolytic cleavage under strong acidic conditions, however, they are more resistant toward acid-catalyzed hydrolysis than linear dextrins and the hydrolytic rate decreases with decreasing cavity size (143). The rate of both the nonenzymatic and enzymatic hydrolysis is decreased when the cavity is occupied by drug molecule.

Table 1. The Principal Physical Chemical Characteristics of Natural CDs

	α-CDs	β-CDs	γ-CDs
Molecular weight	972	1135	1297
Unites of glucopyranose	6	7	8
Internal diameter, Å	5	6	8
Solubility, $mg \cdot 100\,mL^{-1}$, 25 °C	14.2	1.85	23.2
Melting point, °C	250–255	250–265	240–245

Table 1 reports the principal physical–chemical characteristics of natural CDs. Natural CDs, in particular β-CD, have aqueous solubility much lower with respect to comparable linear or branched dextrins. This is probably due to the relatively strong binding of the CDs molecules in the crystal state (i.e., relatively high crystal lattice energy). Moreover, β-CD form intramolecular hydrogen bonding between the secondary hydroxyl groups that reduces the number of hydroxyl groups capable of forming hydrogen bonds with the surrounding water molecules (142). This low aqueous solubility may cause precipitation of solid CDs complexes.

The most important characteristics of CDs is their ability to form inclusion complexes both in solution and in the solid state, in which the guest molecule places its self in the hydrophobic cavity hiding from the aqueous environment. This leads to a modification of physical, chemical, and biological properties of the guest molecules, but principally of the aqueous solubility.

The β-CD is the most useful pharmaceutical complexing agent principally because of its low cost and easy production. It contains 21 hydroxyls groups, of which 7 are primary and 14 are secondary (Fig. 26). All the OH groups are reactive enough to be used as points of reaction for structural modifications, allowing the introduction of several functional groups in to the natural macrocyclic

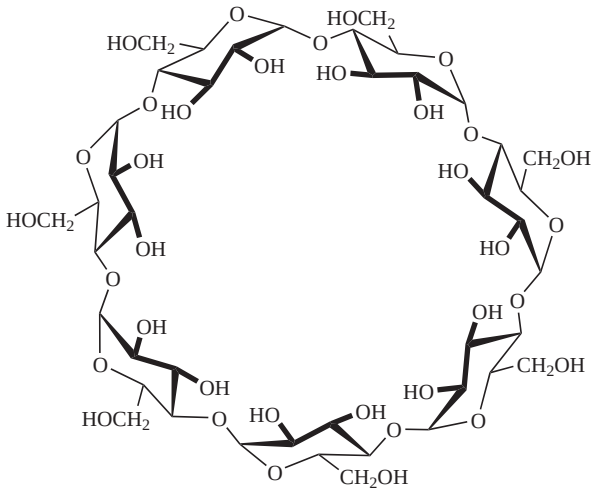


Figure 26. Structure of β-Cyclodextrin.

Table 2. Pharmaceutically Useful β-CD Derivatives[a]

Derivative	Position of Substituent	Substituent
Hydrophilic Derivatives		
Methylated β-CD	2,6-; 2,3,6-	–O–CH_3
Hydroxyalkylated β-CD	Random	–O–CH_3 –CH(OH)–CH_3
Branched β-CD	6-	–Glucosyl, –maltosyl
Hydrophobic Derivatives		
Ethylated β-CD	2,6-; 2,3,6-	–O–C_2H_5
Peracylated β-CD	2,3,6-	–O–$CO(CH_2)_n$–CH_3
Ionizable Derivatives		
Carboxyalkyl β-CD	Random	–O–(CH_2)–COONa
Carboxymethyl; ethyl	2,6-; 3-	–O–CH_2–COONa; –O–C_2H_5
Sulfates	Random	–O–SO_3Na
Alkylsulfonates	Random	–O–$(CH_2)_n$–SO_3Na

[a]Obtained by substitution of the OH groups located on the edge of the CD ring. From Ref. 144.

ring. In the past few years, research has led to a great a number of modified CDs having better characteristics with respect to natural CDs (Table 2).

To date, ~100 different CD derivatives are commercially available, but only few of those derivatives have gone through toxicological evaluations and are available as bulk chemicals for pharmaceutical use. In particular, CDs currently used in drug formulation, derived from natural CD, include 2-hydroxypropyl-β-CD (HP-β-CD), randomly methylated βCD (RM-β-CD), sulfobutylether βCD (SBE-β-CD), maltosyl βCD (ML-β-CD), and (hydroxyethylβCD (HE-β-CD). The aqueous solubility of all these cited derivatives is $>50\,g \cdot 100\,mL^{-1}$.

Inclusion Complex Formation. The most important feature of CDs is their ability to form solid inclusion complexes (of the host–guest type) with a very wide range of solid, liquid, and gaseous lipophilic compounds by a phenomenon of molecular complexation (145). In these complexes, a guest molecule is kept within the cavity of the CD. Complex formation is a dimensional fit between the host cavity and the guest molecule (146). The lipophilic cavity of CDs molecules supplies a microenvironment where an appropriately sized lipophilic moietie can enter to form an inclusion complex (147). The formation of the complex never involves the formation or the breaking of covalent bonds (148). The main driving force for this kind of process is the replacement of enthalpy-rich water molecules contained in the cavity with more hydrophobic guest molecules present in the solution to attain an apolar–apolar association and decrease of CD ring strain resulting in a more stable lower energy state (142). The binding of guest molecules within the host CD is not permanent and is characterized by a dynamic equilibrium, whose strength depends on how well the host–guest complex fits together and on specific local interactions between surface atoms. More specifically, in aqueous solution the CD–drug complexes are constantly being formed and broken.

In particular, considering a 1:1 complexation, the association is usually described by the following equilibrium:

$$D_f + CD_f \overset{K}{\rightleftharpoons} DCD$$

The most important parameters that influence the inclusion process are the complexation strength or stability constant (K) defined by this equilibrium and equation 1, where (CD_f) and (D_f) are the concentrations of free CD and free drug, respectively; the other parameter is the lifetime (t) of the complex and equation 2, measured when the equilibrium is perturbed. The constants k_f and k_r are the forward and reverse rate constants, respectively, and k_{obs} is the observed rate constant for the reestablishment of the equilibrium after its perturbation.

$$K = \frac{k_f}{k_r} = \frac{[DCD]}{[D_f][CD_f]} \quad (1)$$

$$K_{obs} = \frac{l}{\tau} = k_f([CD_f] + [D_f]) + k_r \quad (2)$$

The CDs are able to complex the lipophilic substances both in solution, in this case water is the solvent of choice, or in the crystalline state. In some particular cases, the complexation may be performed also in the presence of any nonaqueous solvent, even if in this case a competition drug–solvent for the complexation may happen.

The inclusion of a drug in CDs lead to a profound change of its physicochemical properties as it is temporarily blocked within the host cavity giving rise to beneficial modifications of guest molecules, which are not achievable otherwise (149). In particular, the more influenced properties are enhanced solubility in water of highly insoluble guests, stabilization of labile guests against the degradative effects of oxidation, visible or ultraviolet (UV) light and heat, control of volatility and sublimation, physical isolation of incompatible compounds, chromatographic separations, taste modification by masking off flavors, unpleasant odors, and controlled release of drugs and flavors. Therefore, cyclodextrins may be used in several field: in food, pharmaceuticals, cosmetics, environment protection, bioconversion, packing, and textile industry.

The substances that may be complexed in CDs are quite varied and includes such compounds as straight- or branched-chain aliphatics, aldehydes, ketones, alcohols, organic acids, fatty acids, aromatics, gases, and polar compounds (e.g., halogens, oxyacids, and amines) (149).

Main Methods of Preparation of Drug–Cyclodextrins Complex. The CD complexes may be prepared with various methods. In solution, the complexes are prepared by addition of an excess amount of the drug to an aqueous CD solution. The suspension formed is equilibrated (for periods of up to 1 week at the desired temperature) and then filtered or centrifuged to form a clear CD–drug complex solution. Then, the water is removed from this solution by evaporation or sublimation, for examples, spray or freeze drying to obtain a solid complex.

Other methods applied to prepare solid CD–drug complex include kneading and slurry methods, coprecipitaion,

neutralization, and grinding techniques (150). In some cases, the complexation efficiency is not very high, and therefore relatively large amounts of CDs are needed to complex small amounts of a given drug. Moreover, various vehicle constituents, such as surfactants, lipids, organic solvents, buffer salts, and preservatives, often reduce the efficiency. However, it is possible to enhance the efficiency through formation of multicomponent complex systems (151). For example, recent research demonstrated that water-soluble polymers are able to enhance the complexation efficacy of a wide variety of guest molecules, through stabilization of the CD–drug complex, and to increase the aqueous solubility of the natural cyclodextrins (152).

Analytical Methods Used to Detect the Complex Formation. Following the preparation of a drug–CD complex, a fundamental step is to verify this complexation, the stoichiometry of the complex, and its stability constant. All these parameters may be clarified by mean of several technique: thin-layer chromatography (TLC) (153), high performance liquid cromatography (HPLC) (154); gas chromatography–mass spectrometry (GC–MS) for the appraisal of the pattern of substitution (155); nuclear magnetic resonance (NMR) (156); circular dicroism (CD) (157) differential scanning calorimetry, X-ray diffraction, ultraviolet (UV) spectrometry (158), capillary zone electrophoresis (159); electrokinetic chromatography (160); GC stationary phase (161); light scattering and cryoelectronic microscopy (162).

Toxicological Profile. An important limitation for the pharmaceutical application of a substance (both drug or excipient) is the appearence of toxicity after its administration. For this reason, a fervent field of research has been the evaluation of a toxicological profile of CDs. Recently, a review has been published showing the adverse effects from CDs (163).

In general, oral administration of CDs does not show any toxic effect, due to lack of absorption from the GI tract.

Natural CDs, α- and β-CDs, as well as many of their alkylated derivatives, show significant renal toxicity, and for this reason are not used for parenteral use (164). A number of safety evaluations have shown that HP-β-CD, SB-E-βCD, ML-β-CD, γ-CD, and HP-β-CD appear to be suitable in parenteral, as well as oral formulations (164,165). However, the lack of available toxicological data will, more than anything else, hinder pharmaceutical applications of CDs.

Cyclodextrins Elimination. Both HP-β-CD and SBE -β-CD are quantitatively cleared unmodified by renal filtration. Following IV administration, these cyclodextrins have an half-life of 1 h (for humans, this value is specie dependent) with the major amount present in the urine between 1 and 4 h after administration.

The elimination of an unmodified CD–drug complex may lead to an increase in renal clearance of unchanged drug. The drug elimination may occur also following another mechanism. Water reabsorption physiologically occurs in the proximal and distal tubules leading to about a 100-fold increase in the concentration of filtered molecules. In this process, lipophilic drugs normally undergo passive reabsorption while polar molecules are only concentrated. This concentration, encourages the complex formation between the renal cleared cyclodextrins and any lipophilic molecule remaining in the kidney tubules. Since the complex is polar, the presence of the CDs is able to inhibit passive reabsorption of lipophilic drugs physiologically present resulting in greater renal clearance of lipophilic molecules.

Cyclodextrins as Drug Delivery Systems

The most classic application of CDs is in drug delivery. The CDs offer significant advantages over standard formulation. The CD–drug complexes can stabilize, enhance the solubility, bioavailability, and diminish the adverse effects of a drug. The bioadaptability and multifunctional characteristics of CDs make them able to minimize the undesirable properties of drugs in various routes of administration including oral, rectal, nasal, ocular, transdermal, and dermal. In Table 3, the role of CDs in drug formulation and delivery is reported in detail (166).

Cyclodextrins in Nasal Drug Delivery. Nasal administration of drugs is an important route of administration for several classes of drugs. Unfortunately, mucosa present both physical and metabolic barriers to drug permeation, restricting this therapeutically approach. The CDs may be used to overcome these obstacles. It has been demonstrated that CDs are able to reduce or to minimize the enzymatic activity of nasal mucosa (167).

Some morphological studies have shown that the methylated β-CDs are useful carriers for nasal drug delivery. Their effects on the mucosa are not significantly different to that of physiological saline and smaller than those of benzalkonium chloride, a worldwide used preservative for nasal drug formulations.

After nasal administration of a drug–CD formulation, only the drug permeates through the nasal epithelium, but not the highly water-soluble CD or its complex. In humans, DM-β-CD is hardly absorbed after nasal administration of a solution containing $\sim 5\%$ of dimethyl-β-CD. Four percent of the nasally administered dose was recovered in the urine (168).The fraction of the CD dose that is not absorbed from the nasal cavity is removed by the nasal mucociliary clearance system.

Moreover, studies of permeation of various lipophilic drugs complexed in CDs demonstrate that they can largely improve the permeation through nasal mucosa of these substances both in the case of polypeptides and proteins (169).

Cyclodextrins in Ophthalmic Drug Delivery. Tear fluid contains a large variety and amount of enzymes that influence the permeation of topical applied drugs. Numerous studies have shown that CDs are useful additives in ophthalmic formulations because they are able to increase the aqueous solubility, stability of ophthalmic drugs, and to decrease drug irritation (170).

The CD complexation of water-soluble drugs (in order to modify an adverse property, e.g., increase their chemical stability or to decrease ophthalmic drug irritation) generally

Table 3. Role of Cyclodextrins in Drug Delivery[a]

Improved Drug Functions by CD Complexation	Example Drug	Type of CD
Increase in bioavailability (by increased solubility and stability)	Thalidomide	Natural CDs
As above	Nimuselide	β-CD, 2HP-β-CD
As above	Prednisolone	SBE-7-β-CD
As above	Oteprednol etabonate	γ-CD
As above	Sulfhamethazole	β-CD and HP-CD
As above	Tacrolimus	Natural and hydrophilic CDs
As above	Artemisin	β- and γ-CD
As above	Prostaglandin E1	Sulfobutly ether β-CD
Increase in solid-stability of amorphous drug	Quinapril	β-CDs
Increased absorption		
Oral delivery	Ketoconazole, testosterone	β-CD and HPβ-CD
Rectal delivery	Flurbiprofen, carmafur biphenyl acetic acid	2 HPβ-CD
Nasal delivery	Morphine, antiviral drug and insulin	2HPβ-CD
Transdermal delivery	Prostagalndin E1	6-*O*-(carboxymethyl) *O*-ethylβ-CD
Ocular delivery	Dexamethasone, Carbonicanhydrase inhibitors	2HPβ-CD β-CD
Protein and peptide delivery	Growth hormone, interleukin-2, aspartame, albumin and MABs	Different modified CDs
Reduction of local irritancy and toxicity	Pilocarpine, phenothiazine euroleptics, *all-trans*-retenoic acid	2 HPβ-CD (2,6-diOmethyl) β-CD and β-CD

[a]Modified from Ref. 162.

decrease the ophthalmic bioavailability (171); in the case of water-poor soluble drugs, the bioavailability strongly depends on the amount of CDs present: It is fundamental to use only the minimum quantity to form a complex (to solubilize) of the drug (172). In fact, the amount of CDs in excess reduces the free drug and, as a result the ocular permeation will decrease. An example is given in the research of Jarho et al. 1996 (173), which measured the permeability of arachidonyl-ethanolamide through isolated rabbit cornea. As reported in Fig. 27 the maximum value of permeability was found when the minimum amount of CD was used to maintain the drug in solution.

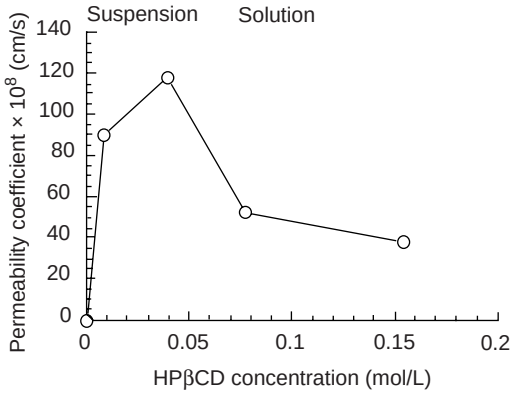


Figure 27. Permeability of arachidonylethanolamide through isolated rabbit cornea as a function of HP-β-CD concentration. The vehicle consisted of $0.5\,mg \cdot mL^{-1}$ suspension or solution of the drug in water containing 0–1.155 *M* HP-β-CD (173).

Moreover, this reduced bioavailability may be attributed to a too rapid ocular clearance (few minutes), and it has been demonstrated that by increasing the viscosity of the ophthalmic formulation, this obstacle may be reduced (174). In addition, a lot of substances used to increase the viscosity of aqueous solutions have been shown good ability to increase the complexation efficacy of CDs and, thus, the amount of CD needed to obtain adequate drug solubility can be decreased significantly when water-soluble polymer is present in the formulation (175).

Cyclodextrins in Dermal and Transdermal Drug Delivery. The CDs are relatively large molecules, and consequently both they and their complexes are not able to permeate through intact skin easily. Only 0.02% of the applied dose of radiolabeled HP-β-CD permeated through human skin. The principal barrier to the permeation of CDs is represented by the stratum corneum, since by stripping it, it is possible to enhance the percutaneous permeation by 24% (176). Lipophilic CDs (as DM-β-CD and RM-β-CD) are absorbed to a greater extent, but this absorption is still of little significance (0.3% of the applied dose).

The CDs are able to interact with some components of skin lipids. In particular, it has been demonstrated that pure aqueous solution of β-CDs and RM-β-CDs ad HPβCD are able to extract the lipids present in stratum corneum (177). Various studies (178,179) have shown that excess CDs, more than needed to complex the lipophilic drug, lead to a decrease of drug permeation through the skin (Fig. 28). When the drug (hydrocortisone) was in suspension, the increase of the cyclodextrin concentration lead to an increase of the flux through the skin. When all hydrocortisone was in solution,

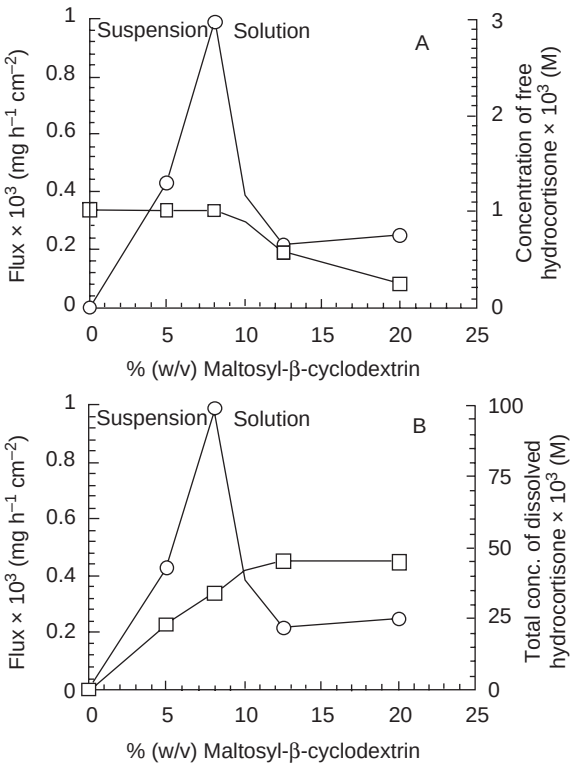


Figure 28. The effect of the maltosyl-β-cyclodextrin (ML-β-CD) concentration on the flux (○) of hydrocortisone through hairless mouse skin. The hydrocortisone concentration was kept constant at 0.045*M*. (a) The flux in relation to the amount of free hydrocortisone (□) in the donor phase. (b) The flux in relation to the total amount of dissolved hydrocortisone (□) in the donor phase. The donor phase consisted of aqueous hydrocortisone suspension at ML-β-CD concentrations $< 8\%$ (w/v), but hydrocortisone solution at higher ML-β-CD concentrations (175).

the increase of the cyclodextrin content led to a decrease in flux. In all cases, maximum flux through the skin was obtained when just enough cyclodextrin was added to the vehicle to keep all hydrocortisone in solution. The mechanism has been already elucidated in a previous section.

A great number of topical drugs have been complexed with CDs (Tables 4 and 5). In every case, it has been demonstrated that CDs can markedly enhance the dermal delivery of lipophilic drugs (e.g., corticoids and NSAIDS). In particular, from a comparative evaluation (181) among the compounds DM-β-CD, β-CD, and HP-β-CD, resulted in HP-β-CD being more able to increase the percutaneous permeation (Fig. 29).

The effects of CDs on the permeation rates of drugs through the skin may be determined by both the increase of thermodynamic activity of drugs in a vehicle (in particular it is referred to the escaping tendency of drugs, and it is supposed that increasing this activity will lead to the augmentation of the permeation rate of drugs through the skin; moreover, the thermodynamic activity is proportional to the solubility of drugs in its vehicle and is maximal just in the saturated solution), the extraction of skin component, and the partition coefficient of the drug between skin and vehicle.

Recently, it has been evaluated for transdermal use in peracylated CDs with medium alkyl chain length (C_4–C_6) and in particular 2,3,6-tri-*O*-valeryl-β-CD (TV-β-CD). This particular type of CD shows the property of forming a film, and for this reason is very promising in transdermal preparations.

Cyclodextrins in Rectal Drug Delivery. The CDs have been studied to optimize the rectal delivery of drugs for systemic use. Table 6 reportes the CDs and drugs investigated for rectal application. The effects of CDs on the rectal delivery of drugs depends on vehicle type (hydrophilic or oleaginous), physicochemical properties of the complexes, and an existence of excipients, such as viscous polymer. The enhancement of rectal permeation of lipophilic drugs made by CDs is generally attributed to the improvement of the release from vehicles and the dissolution rates in rectal fluids (Fig. 30). In the case of inabsorbable drugs, such as antibiotics, peptide, and proteins, the rectal delivery is based on the direct action of CDs on the rectal epithelial cells.

On the other hand, the prolonging effects of CDs on the drug levels in blood are caused by several factors: sustained release from the vehicles, slower dissolution rates in the rectal fluid, and retardation in the rectal absorption of drugs by an inabsorbable complex formation.

Another important aspect that has been evaluated is the stabilization of drugs in rectal delivery. The complexation of drugs with CDs has been used to improve the chemical stability in suppository bases according to a stabilizing effects (principally of β-CD and DM-β-CD) attributable to a poor solubilization of drugs in the oleaginous suppository base; this may lead to a difficult interaction of drugs with the base.

The CDs may inhibit the bioconversion of drugs in the rectum (182) leading to the alleviation of the rectal irritancy of the some drugs as NSAIDS (Fig. 31).

Cyclodextrins in Oral Drug Delivery. An important parameter to be considered for the oral administration of a drug is its water solubility. For poorly aqueous soluble drugs, CDs are able to increase the aqueous solubility and thus enhance its dissolution rate and the biopharmaceutical parameters. An example was observed with the celecoxib (a nonsteroidal antinflammatory drug). The complex celecoxib DM-β-CD showed an increased permeation respect to the free drug. This observed enhanced permeation was due to the fast dissolution rate of the included drug and to a destabilizing action exerted by the CD on the biomembrane (Fig. 32) (183).

The CD derivatives may be used in order to modify drug release of oral preparations. Table 6 reports some application of CDs.

The hydrophilic CDs are able to give an immediate release of the complexed drug, while hydrophobic CDs are useful for the prolonged release formulations. The use of *O*-carbossimethyl-*O*-ethyl-β-CD (CM-β-CD) gives a delayed release formulation.

The immediate release formulation is required in an emergency situations and in a particular way in the administration of analgesics, coronaric antipyretics, and

Table 4. The Use of Parent Cyclodextrins in Transdermal Route[a]

CDs	Abbreviation	Improvement	Drugs
α-Cyclodextrin	α-CD	Release and/or permeation	Miconazole
		Stability	Tixoxortol 17-butyrate 21-propionate
			Betamethasone
			4-biphenylacetic acid
			Chloramphenicol
			Ciprofloxacin
			Ethyl 4-biphenylyl acetate
			Flurbiprofen
			Hydrocortisone
β-Cyclodextrin	β-CD	Release and/or permeation	Indomethacin
			Nitroglycerin
			Norfioxacin
			Piroxicam
			Prednisolone
			Prostaglandin E_1
			Sulfanilic acid
		Local irritation	Chlorpromazine hydrochloride
			Tretinoin
γ-Cyclodextrin	γ-CD	Release and/or permeation	Beclomethazone dipropionate
			Betamethasone
			Menadione
			Predonisolone

[a]Adapted from Ref. 180.

Table 5. The Use of Cyclodextrin Derivatives in Transdermal Route[a]

CD Derivatives	Abbreviation	Improvment	Drugs
Dimethyl-β-cyclodextrin	DM-β-CD	Release and/or permeation	4-Biphenylacetic acid
			Ethyl 4-biphenylyl acetate
			Indomethacin
			Predonisolene
			Sufanilic acid
		Local irritation	Chlorpromazine
Random methyl-β-cyclodextrin	RM-β-CD	Release and/or permeation	Acitretin
			Hydrocortisone
			Piribedil
			S-9977
Hydroxypropyl-β-cyclodextrin	HP-β-CD	Release and/or permeation	4-Biphenylacetic acid
			Dexamethasone 17β-estradiol
			Ethyl 4-biphenylyl acetate
			Hydrocortisone
			Liarozole
			Miconazole
Maltosyl-β-cyclodextrin	G_2-β-CD	Release and/or permeation	Hydrocortisone
β-Cyclodextrin polymer	β-CD polymer	Release and/or permeation	Tolnaftate
			Indomethacin
Diethyl-β-cyclodextrin	DE-β-CD	Release and/or permeation	Nitroglycerin
Carboxymethyl-β-cyclodextrin	CM-β-CD	Release and/or permeation	Hydrocortisone
Carboxymethyl-ethyl-β-cyclodextrin	CME-β-CD	Release and/or permeation	Prostaglandin

[a]Adapted from Ref. 180.

vasodilatators. Hydrophilic CDs have been used in order to improve the oral bioavailability of the previous mentioned drugs (184). The improvement is mainly dependent on the increase of solubility and wettability of drugs through the formation of inclusion complexes (185).

The oral bioavailability of a lipophilic drug from the CD complex may be optimized varying several factors, that influence the equilibrium of dissociation of the complex (166,186). The maximum improvement of the absorption is obtained when a sufficient amount of CD is used to complex all the molecules of the drug present in suspension, and the further addition of CDs lead to a reduction the free fraction of the drug and, therefore, reduces the bioavailability of the drug. Moreover, drug

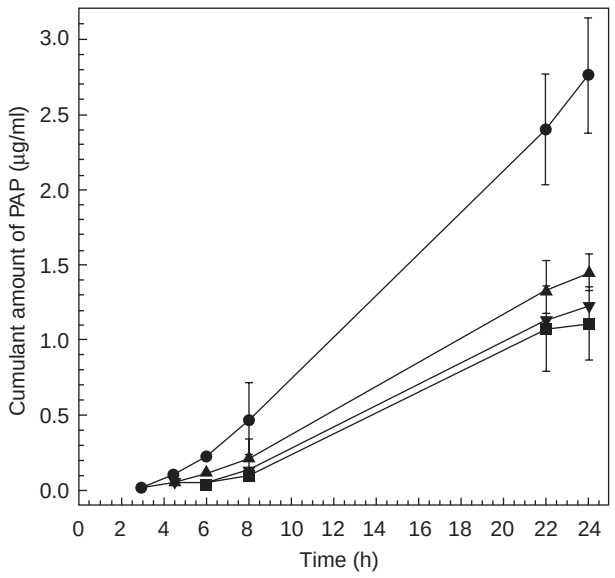


Figure 29. Total amount of free or complexed papaverine permeated through abdominal rat skin. Symbols: ν free PAP alone; ▲ PAP-HP-β-CyD; ● PAP-β-CyD; λ PAP-DM-β-CyD (179).

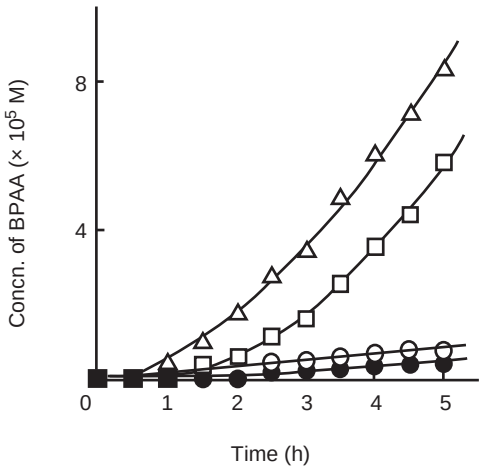


Figure 30. Profiles of BPAA permeation through isolated rat rectum after applications of suppositories containing EBA or its β-CD complexes in isotonic phosphate buffer (pH 7.4) at 37 °C. ● without β-CDs; ○ with β-CD;□ with DM-β-CD; Δ with HP-β- CD (182).

formulations contain a certain amount of excipients that may compete with the drug for the cavity of the CD. Competition may also happen with the endogenous substances present in the absorption site. The replacement of the drug from the cavity of the CD from both endogenous and exogenous substances lead to an acceleration of the absorption of the drug (187,188). Early studies showed that in some cases the improvement of oral bioavailability is principally due to a stabilizing effect of CDs on labile drugs (189).

Table 6. The Use of Cyclodextrins in Rectal Delivery[a]

CDs	Improvement	Drugs
	Stability	Morphine hydrochloride
α-CD	Release and/or permeation	Cefmetazole
		G-CSF
		Morphine hydrochloride
	Stability	AD1590
		Carmoful
		Ethyl 4-biphenylyl acetate
β-CD	Release and/or permeation	4-Biphenylacetic acid
		Ethyl 4-biphenylyl acetate
		Naproxen
		Phenobarbital
		Piroxicum
γ-CD	Release and/or permeation	Diazepam
		Flurbiprofen
	Release and/or permeation	4-Biphenylacetic acid
		Carmoful
		Diazepam
DM-β-CD		Ethyl 4-biphenylyl acetate
		Flurbiprofen
		Insulin
	Local irritation	4-Biphenylacetic acid
		Ethyl 4-biphenylyl acetate
TM-β-CD	Release and/or permeation	Carmoful
		Diazepam
		Flurbiprofen
HP-β-CD	Release and/or permeation	4-Biphenylacetic acid
		Diazepam
		Ethyl 4-biphenylyl acetate
β-CD Polymer	Selective transfer into lymphatics	Carmofur

[a]Adapted from Ref. 180.

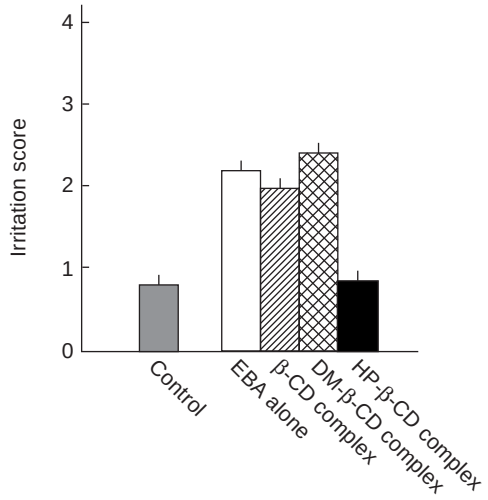


Figure 31. Irritation effect of suppositories containing EBA or its β-CD complexes (equivalent to BPAA 10 mg · kg^{-1}) on rectal mucosa in rats 12 h after multiple (four times at 12 h intervals) administration (182).

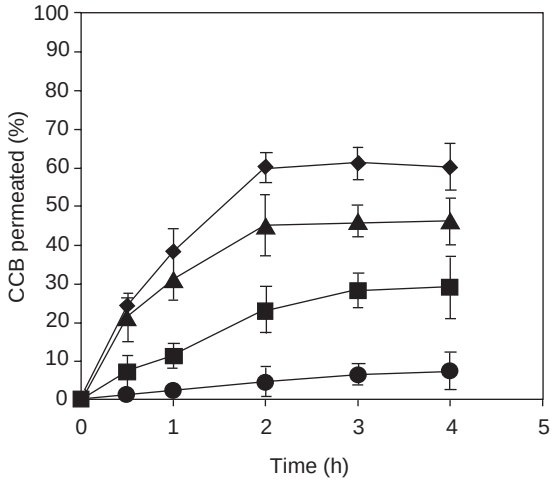


Figure 32. Permeation profiles of CCB alone or in the presence of the CCB-DM-β-CD complex in a different molar ratio. (●) CCB alone; (●) CCB-DM-β-CD 1:2 complex as suspension; (▲) CCB-DM-β-CD 1:5 complex as solution;(◆) CCB-DM-β-CD 1:10 complex as solution (181).

Others positive results may be obtained for the sublingual administration of drugs, complexed with CD (190,191). In this application, not only the drug permeates rapidly giving an immediate response, but it also avoid hepatic first pass metabolism.

The preparations characterized by a slow release are planned for having a zero-order release to guarantee a constant blood level for a long period of time. This type of formulation has many advantages, like the reduction of the administration frequency with an extension of the efficacy of the drug and the reduction of the toxicity associated with the administration of a simple dose. Hydrophobic CDs, as alkylated and acylated derivatives, are used as slow-release carriers for hydrophilic drugs. Between the alkylated CDs, 2,6-di-*O*-ethyl-β-CD (DE-β-CD) and 2,3,6-tri-*O*-ethyl-β-CD (TE-β-CD) were the first used as slow-release carriers (192).

Another type of CD useful in oral formulation is 2,3,6-tri-*O*-butyryle-β-CD (TB-β-CD), whose bioadhesive property make it very advantageous in oral and transmucosal formulations.

Cyclodextrins in Parenteral Administration. Modern technology is trying to obtain semisynthetic CDs that have the following characteristics to use as parenteral drug delivery systems. For this application, the drug toxicity at high doses will need to be improved for chronic treatment; its inability to react with cholesterol, phospholipids, or others members of the cellular membrane, and its biodegradability in circulation and elimination of small molecular metabolites.

In general, for this kind of application only hydrophilic CDs, in particular HP-βCD are used. This has been, carefully studied by means of innumerable toxicological experiments and has been the object of numerous clinical tests on human. One formulation, based on the carrier Sporanox by Jassen (193) has been approved by the U.S. Food and Drug Administration (FDA). Another hydrophilic derivative is used for parenteral use is β-CD sulfobutylether. It is used under the name of Captisol.

The sulfate CD represent another class of soluble CD in water with a characteristic biological activity. It shows an antiangiogenic power that may be useful in new therapies against cancer. A few studies have demonstrated that the sulfate CD does not have any hemolytic properties at all, are not toxic, and protect against the nephrotoxicity induced by gentamicin without even reducing renal accumulation of this active principle (194).

Cyclodextrins in Anticancer Therapy. Cyclodextrins also play a vital role in the drug formulation design for cancer therapy. Bekers et al. (195) in 1991 studied the effect of cyclodextrins on the chemical stability of mitomycin C, a clinically useful anticancer drug able to generate severe dermatological problems after administration. The complexation of this drug with CD reduced the skin necrosis observed after the treatment with the free drug.

Real advantages were demonstrated in the delivery of paclitaxel, an anticancer agent used in breast, ovarian, lung, head and neck cancers, characterized by very low water solubility. For this reason, it must be formulated as a micellar solution made up of polyoxyethylated castor oil and 50% absolute ethanol. This formulation triggers severe acute adverse effects in both animals and humans. The complexation of paclitaxel in CDs, β-CD, DM-β-CD, and TM-β-CD, showed a modulation of the maintainance of the anticancer activity (196).

Cyclodextrins as Carrier for Biological Drugs. Besides drugs, different peptides and proteins (197), oligosaccharides, and oligonucleotides (198) are also delivered by the formation of inclusion complexs with cyclodextrins because of CDs ability of interacting with cellular membranes and giving rise to improved cellular uptake. The most recent usage of cyclodextrins lies in the ability of these agents to

deliver agents, such as plasmids, viral vectors, and antisense constructs. The *in vitro* stability of antisense molecules is increased by binding to CDs, such as hydroxypropyl b-CD. A two- to threefold increase in the cellular uptake of antisense constructs by hydroxyalkylated b-CD has been noted in human T-cell leukemia H9 cells (199). Certain CDs modulate the intracellular distribution or activity of antisense molecules and they may be used for reversal of atherosclerosis (200). Cyclodextrins are also used to formulate the enhancement of the physical stability of viral vectors for gene therapy by suspending the adenovirus and adeno-associated virus in blends of CD, complex carbohydrates, and various surfactants (201). Three native CDs (α, β, and γ) were observed to improve the antiviral effect of ganciclovir on two human cytomegalovirus strains (202). Use of CDs as carriers of antiviral drugs appears to be a good alternative to traditional treatments as it allows the administration of lower doses and reduces the toxic effect of drug molecules.

Cyclodextrins in Colon Targeting. Colon targeting may be classified as a delayed release with a fairly long time because the time required to reach the colon is ~ 8 h in humans (203). When a formulation is administered orally, it will dissociate in the GI fluid and for this reason CD complexes are not suitable for colon delivery. For this reason, it was proposed to use CD–drug conjugates (a prodrug) that were able to survive the passage through the stomach and small intestine. In particular, the linkage of CD to biphelylyacetic acid (BPAA) has been investigated. It is interesting to note that the solubility of this type of prodrug is strictly related to the cavity size of the CD. Moreover, in the case of ester-type conjugates, drug release is the case of ester-type conjugates, drug release is triggered by the ring opening of CDs, which consequently provides site-specific drug delivery to the colon. On the other hand, the amide conjugates do not release the drug even in the cecum and colon, despite the ring opening of CDs. The amide linkage of the small saccharide–drug conjugates may be resistant to bacterial enzymes and poorly absorable from the intestinal tract due to high hydrophilicity. Therefore, the ester-type conjugate is preferable as a delayed release-type prodrug that can release a parent drug selectively in the cecum and colon (204).

SUPRAMOLECULAR AGGREGATES FOR DRUG DELIVERY

General Characteristics of Surfactants

Surfactants are molecules characterized by a polar head and an apolar tail region, the latter occupies the larger molecular volume, in a particular way for ionic surfactants. When dispersed in water, surfactants self-associate into a variety of equilibrium phases, the nature of which stems directly from the interplay of the various (forces inter- and intermolecular), as well as entropy evaluations. Surfactants also self-associate in nonaqueous solvents, particularly apolar liquids, such as alkanes. In this case, the orientation of the surfactant molecules are reversed compared to those observed in aqueous solution. This reorientation lead to a lowering of the free energy of the system overall. When surfactants are incorporated into two immiscible solvents as oil and water, the surfactant molecules locate themselves at the oil–water interface. This arrangement is thermodynamically favorable.

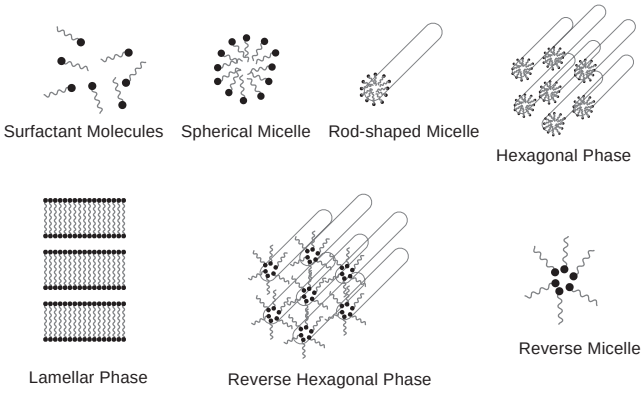


Figure 33. Schematic representation of the most commonly self association structures in water, oil or a combination thereof (205).

Figure 33 reports a number of possible self-association structures that surfactant may form when placed in a oil and water (205).

Microemulsions

The microemulsion concept was introduced as early as the 1940s by Hoar and Schulman, who generated a clear single-phase solution by titrating a milky emulsion with hexanol (206). Later, in 1959, Schulman coined the term microemulsion (207). Today microemulsions are defined as A mixture of water, oil, and amphiphile substances forming a single optically, isotropic and thermodynamically stable liquid solution. The stability is the most important difference between emulsions and microemulsions. In fact, emulsions are fundamentally thermodynamically unstable and, even if they show an excellent kinetic stability, may undergo phase separation (208). Another important difference is related to their appearance. Emulsions are milky while microemulsions are clear or translucent. In addition, there is a noticeable difference in their method of preparation, since emulsions require a large input of energy while microemulsions do not. Microemulsions are dynamic systems in which the interface is continuously and spontaneously fluctuating (209).

Schematic representations of the three types of microemulsions are most likely formed are reported in Fig. 34. The structures shown are very different, but in each there is an interfacial surfactant monolayer separating the oil and water domains.

Three approaches have been proposed to explain the spontaneous microemulsion formation and their consequent stability: interfacial or mixed-film theories (210); solubilization theories (211); and thermodynamic treatments (212). In particular, the free energy of microemulsion formation reported in equation 3 is dependent on the extent to which surfactant is able to lower the surface tension of the oil–water interface and the change in

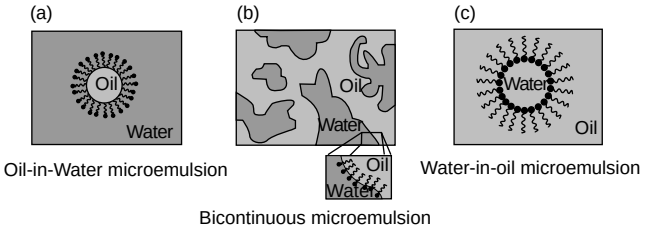


Figure 34. Schematic representation of three type of microemulsion microstructures: (a) oil-in-water, (b) bicontinuous, and (c) water-in-oil microemulsion (205).

entropy of the system such that,

$$\Delta G_f = \gamma \Delta A - T\Delta S \tag{3}$$

where ΔG is the free energy of formation, γ is the surface tension of the oil–water interface, ΔA is the change in interfacial area on microemulsification, ΔS is the change in entropy of the system, and T is the temperature.

When a microemulsion is formed, the change in ΔA is very large due to the formation of a great number of very small droplets generated. Originally, it was proposed that to form a microemulsion a negative value of γ was required. It is now accepted that this value of g is always positive, but it is very small (of the order of fractions of $\mathrm{mN \cdot m^{-1}}$), and is offset by the entropic component. The dominant favorable entropic contribution is the very large dispersion entropy arising from the mixing of one phase in the other in the form of large numbers of small droplets. Thus a negative free energy of formation is achieved when large reductions in surface tension are accompanied by significant favorable entropic change. In such cases, microemulsification is spontaneous and the resulting dispersion is thermodynamically stable.

The phase behavior of simple microemulsion systems comprising oil, water, and surfactant can be studied with the aid of a ternary phase diagram in which each corner of the diagram represents 100% of that particular component. More commonly, however, and in a special way in the case of microemulsions for pharmaceutical applications, the microemulsion contains additional components, such as a cosurfactant and/or drug. The cosurfactant is also amphiphilic with an affinity for both the oil and aqueous phases and partitions to an appreciable extent into the surfactant interfacial monolayer present at the oil–water interface. It has three functions: to provide very low interfacial tensions required for the formation of microemulsions and their thermodynamic stability; to modify the curvature of the interface based on the relative importance of their apolar groups; and to act on the fluidity of the interfacial film. If the film is too rigid, it prevents the formation of microemulsion and results in a more viscous phase. The existence of unsaturated bounds on the hydrocarbon chain of the surfactants equally increases the fluidity of the film. The cosurfactants used are small molecules, generally alcohols with the length of the carbon chain ranging from C2 and C10, or amines with short chains can also be used as cosurfactants. Moreover, a large number of drug molecules are themselves surface active and influence phase behavior.

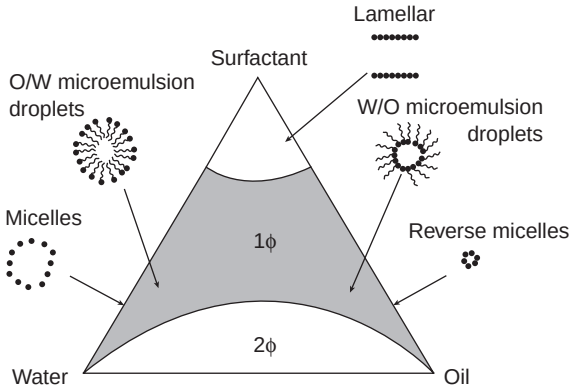


Figure 35. A hypothetical pseudo-ternary phase diagram of an oil–surfactant system with emphasis on microemulsion and emulsion phases. Within the phase diagram, existence fields are shown where conventional micelles or water–oil microemulsion, and oil–water microemulsions are formed along with the bicontinuous microemulsions. At very high surfactant concentrations two-phase systems are observed (205).

In the case where four or more components are present, pseudoternary phase diagrams are used where a corner will typically represent a binary mixture of two components, such as surfactant–cosurfactant, water–drug, or oil–drug. A highly schematic (pseudo) ternary-phase diagram illustrating various phases is presented in Fig. 35. Not every combination of various components produce microemulsions over the whole range of possible compositions, in some instances the extent of microemulsion formation may be very limited. The procedure most often employed to construct the phase diagrams is to prepare a series of (pseudo) binary compositions and titrate with the third component, evaluating the mixture after each addition. The temperature must be accurately controlled and the observations must not be made on metastable systems (213). Transitions between the various phases pictured in these phase diagrams can be driven by the further addition of one of the components, addition of a new component (drug or electrolyte), or by changing the temperature. Transitions from water/oil (w/o) to oil/water (o/w) microemulsions may occur via a number of different structural states including bicontinuous, lamellar, and also multiphase systems. In particular, microemulsions stabilized by nonionic surfactants are very susceptible to an increased temperature, leading to the phase inversion temperatures (PIT). The presence of PIT may cause problems especially when formulations are for parenteral application and must be sterilized by means of an autoclave. On the other hand, the presence of PIT may be use for the drug delivery directed to a specific site.

Advantages of Microemulsions as Drug Delivery Systems. Microemulsions present some important characteristics that make themselves very versatile carriers. In particular, they present a thermodynamic stability, optical clarity, and ease of preparation. The existence of microdomains of different polarity within the same single-phase solution allow the solubilization both water soluble and at the same time if this is so desired. Furthermore it is also possible to

incorporate amphiphilic drugs into the microemulsion. It must be emphasized that the use of o/w microemulsions in drug delivery is more straightforward than it is with w/o microemulsions. The reason is because the droplet structure of o/w microemulsions is not broken following the dilution by a biological aqueous phase; this aspect make possible the oral as well as parenteral administration. The process of dilution will result in the gradual desorption of surfactant present at the droplet interface. This process is thermodynamically driven by the requirement of surfactant to maintain an aqueous phase concentration equivalent to its critical micelle concentration while maintaining temperature, pH, and ionic strength. The use of w/o microemulsions for oral or parenteral drug delivery is complicated by the fact that they are destabilized when diluted by biological aqueous fluids.

Applicative Potentialities of Microemulsions

Transdermal Application. Microemulsions represent an ideal vehicle for the topical administration of drugs because they combine the emulsion properties with those of solution. It is well known that surfactants produce stratum corneum dehydration and barrier compromise (214,215), and consequently the high levels of surfactant–cosurfactant present in the microemulsions may cause a disruption of the stratum corneum. Consequently, there is an enhancement in the permeation of drugs. However, the choice of the component is very important to minimize the alteration of the stratum corneum and the appearance of toxic effects. The choice of biocompatible components can guarantee an increased skin tolerability. For this reason, the potential application of highly biocompatible o/w microemulsions as topical drug carrier systems for the percutaneous delivery of antiinflammatory drugs (i.e., ketoprofen) was investigated (216). The components were triglycerides as the oil phase, a mixture of lecithin, and *n*-butanol as a surfactant–cosurfactant system, and an aqueous solution as the external phase. The topical carrier potentialities of lecithin-based o/w microemulsions were compared with respect to conventional formulationsn (i.e., a w/o emulsion, a o/w emulsion, and a gel).

The percutaneous adsorption of ketoprofen, evaluated through healthy adult human skin, delivered with microemulsions, showed an enhancement with respect to conventional formulations. No significant percutaneous enhancer effect was observed for ketoprofen-loaded oleic acid–lecithin microemulsions (Fig. 36). Moreover microemulsions showed a good human skin tolerability (Table 7).

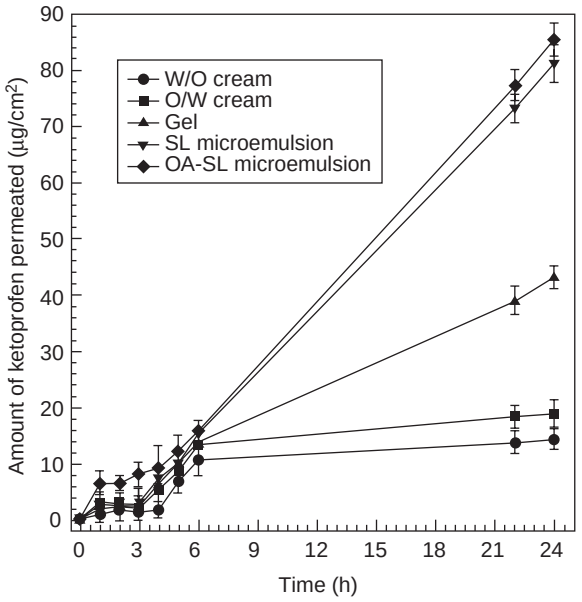


Figure 36. Permeation profiles of ketoprofen through human skin from various topical formulations. Each value is the mean value of three different experiments ± S.D (216).

Several reports have demonstrated that microemulsive vehicles may increase transdermal delivery of both lipophilic and hydrophilic drugs, compared to conventional formulations, depending on the constituents used for the microemulsive vehicle (217–219). These research papers suggested that microemulsion formulations may increase cutaneous drug delivery by means of the high solubility potential for both lipophilic and hydrophilic drugs, which creates an increased concentration gradient toward the skin and/or by using constituents with penetration enhancer activity (211). The incorporated ratio of the respective constituents influence in a significant way the percutaneous and transdermal drug delivery potential of the microemulsions. In every case, the enhancement of the drug delivery mechanism seems to be related to the drug mobility in the vehicle, and that measurement of self-diffusion

Table 7. Human Skin Irritancy Test of Various Topical Formulations After 24 h of Treatment

Sample	Irritation Evidence at 24 h								
	Number of Cases[b]								Score[a]
	Vesicles	Edema	Erythema	Flakiness	Dryness	Wrinkling	Glazing	No Visible Reaction	
OA 1%			3	7		2		18/30	10.17 ± 2.08
w/o						2	3	25/30	6.20 ± 2.77
O/W						2	1	27/30	4.67 ± 2.52
Gel						1	2	27/30	4.33 ± 1.15
SL-ME						1	1	28/30	3.50 ± 1.39
OA-SL-ME						2	1	27/30	4.67 ± 2.08

[a]Nonparametric variable Kruskal–Wallis test provided: $P < 0.001$ for OA (1% w/w) aqueous dispersion vs. all other samples; $P < 0.05$ for w/o cream versus all other samples; $P < 0.05$ for SL microemulsions versus o/w cream, gel, and OA-SL microemulsion.

[b]The value reported in each column represent the number of subjects who showed the skin reaction symptom.

coefficients is valuable to optimize the formulation of a given microemulsion vehicle, in order to maximize drug delivery.

Ophthalmic Application. The drug delivery system used in the ophthalmic field must overcome the disadvantages present in traditional formulations (e.g., a very low bioavailability, 1–10% of the drugs, and consequently frequent administrations are required during the day). Microemulsions represent an interesting alternative because their industrial production and sterilization are relatively simple and inexpensive; they are thermodynamically stable and permit us to solubilize both lipophilic and hydrophilic drugs.

With ophthalmic use, the choice of the various components is fundamental more than with any other topical application. The ionic surfactants are generally too toxic to be used for this application, therefore, nonionic surfactants are preferred (220). These surfactants are easily soluble in water due to the presence of either functional groups. The most used surfactants in the preparation of microemulsions are the poloxamers and polysorbates.

The choice of the oily phase is important because it conditions both the existance of the microemulsion and the solubilization of the drug. Polar oils, such as triglycerides with medium or long chains, are preferred instead of nonpolar oils, based on their solubility. The most often used consist of vegetable oils, such as soja oil, castor oil, or triglycerides, for which 95% of the fatty acids are made up of 8–10 carbon atoms, Myglyol 812s (triesters of glycerol, capric, and caprylic acids), isopropyl myristate, fatty acids, such as oleic acid, and esters of saccharose, such as mono-, di-, or tripalmitates of saccharose. As these excipients are well tolerated by the eye, their degree of purity must be high in order to prevent any contamination with potentially irritating substances.

Several additives, such as buffers, antibacterial, and isotonic agents, contained in the aqueous phase may affect the area of existence of the microemulsions, and therefore they must be studied in the presence of other constituents of the microemulsions. For example salinity influences the phase diagrams when ionic surfactants are added and decreases the phase inversion temperature (PIT) of the nonionic surfactants. Thiomersal and chlorobutanol are preservatives that are usually used in eye drop formulations, with concentrations of 0.01–0.2%, can be used without altering microemulsions structure (221).

The main advantage of the microemulsions is the increase in the solubilization of poorly soluble drugs. In a recent work, different o/w microemulsions containing indomethacin (an antiinflammatory drug). were evaluated *in vivo* by determining both the tolerability (Draize test) and the ocular drug bioavailability. This investigation showed that the colloidal carrier has a certain tolerability, eliciting only a slight irritation at the level of the conjuctiva. A positive effect regarding tolerability was exerted by jaluronic acid. In fact, by increasing the concentration of jaluronic acid present in the formulation up to 1% (w/v), an improved microemulsion ocular tolerability was observed with a substantial reduction of conjunctiva irritation (Table 8). *In vivo* ocular bioavailability of the microemulsion formulation containing indomethacin was evaluated by means of the Draize test. At various time intervals, the rabbits were killed, aqueous humor samples were collected and indomethacin content was determined by high performance liquid chromatography (HPLC). Indomethacin-loaded microemulsion was compared with an aqueous dispersion of the drug, containing the same drug concentration. The microemulsion-encapsulated indomethacin formulation showed a significant ($P > 0.005$) increase of drug levels compared with the free drug (Fig. 37). High colloidal properties of microemulsions may achieve a better interaction with the corneal epithelium in terms of paracellular transport or passage, thus leading to a greater drug transport into the ocular tissues. The microemulsion controlled drug release showed by ocular pharmacokinetics was probably elicited by the colloidal carrier mucoadhesion on the cell surface, thus allowing a prolonged ocular permanence and a release of the content directly into the cell (222).

Table 8. Effect of Microemulsions on Ocular Structures[a]

Ocular Structure	Without Hyaluronic Acid	With Hyaluronic Acid
Conjuctiva Irritation	1.8	0.4
Conjuctiva Edema	0.4	0.2
Fluorescein Adsorption	0.8	0.2

[a]The scores were calculated awarding a value on scales from 0 to 3 at each observed reaction. All the assigned values were added and divided for the number of subjects.

Lecithin Organogel

A particular type of self-aggregate is represented from lecithin organogel. They were seen for the first time in 1988 by Scartazzini and Luisi (223), who noted that an addition of trace amounts of water into nonaqueous solutions of lecithin caused a sudden increase in the viscosity (~ 100 times) producing a transition of the initial nonviscous solution into a jelly-like state. In succeeding years, it was demonstrated that lecithin, when dissolved in at nonpolar solvent, forms spherical reversed micelles. The addition of water induces an uniaxial growth of the micelles. As a result, at the end of the preparation one will find cylindrical aggregates instead of the initial spherical ones. After reaching threshold length, the extended micelles begin overlapping, forming a temporal 3D network. This

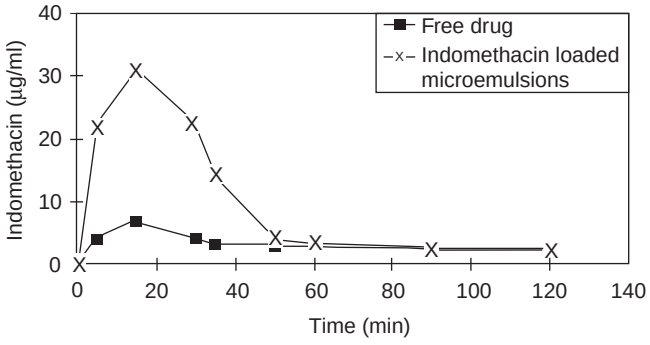


Figure 37. Bioavailability of free indomethacin or loaded microemulsions.

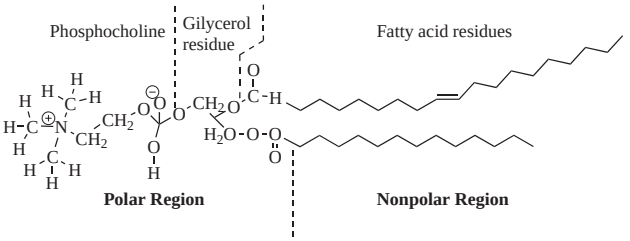


Figure 38. Structural formula of lecithin.

supramolecular structure from entangled micellar aggregates bears resemblance with that of uncrossed polymers in semidilute or concentrated solutions. For this reason, they are often called polymer-like micelles, wormlike, or threadlike micelles, or spaghetti-like structures.

The transition to polymer-like micelles is accompanied with a formation of hydrogen bonds between the phosphate group of a lecithin molecule (Fig. 38) and water.

The lecithin organogel is an optically transparent isotropic phase, appearing as the initial solution before the addition of water. The only difference between them is in the increased viscosity. This aspect is strictly dependent on oil, water, and lecithin concentrations, as well as on temperature (224). The amount of water needed to obtain the gel-like structure is a peculiar properties of any organic solvent (225). An important parameter for the organogel structure formation is the purity of the lecithin solution, in fact, commercial low purity lecithin is not able to form gels (226). The last component for the formation of lecithin organogel is water. This solvent can be substituted by polar organic solvents, such as glycerol, ethylene glycol, and formammide, or by a mix of ethanol–water in different ratios (227). The physical–chemical characteristics of the incorporated drug noticeably influence its release from organogel (144).

An important characteristic of this aggregate is its thermoreversibility, in fact, at 40 °C they become fluid, but by reducing the temperature they again reassemble a gel-like structure.

In this kind of carrier, it is possible to load hydrophilic drugs (localize themselves in aqueous, internal compartment), lipophilic drugs (in the hydrophobic environment), and amphiphilic substances (at the interface w/o).

The principal application of this carrier is its transdermal delivery, as first proposed in the early 1990s by Luisi's research group (225,226). Scopolamine, broxaterol, and propranolol were incorporated into lecithin organogels (containing cyclohexane, isooctane, or IPM as the oily phase). The permeation rates increased 10 fold compared to a solvent drug solution used as a control (180). The utility of lecithin organogels has been supported by *in vivo* human skin tolerability studies by means of a noninvasive technique as spectrophotometry of reflectance (228). *In vivo* percutaneous tolerability results showed no appearance of erythema even after 48 h of application. Certain amphiphilic lipids are characterized by lyotropic and thermotropic aggregation-phase transition. These supramolecular aggregates are under investigation to evaluate their potentialities as drug delivery systems (229).

COLLOIDAL DRUG DELIVERY DEVICES

The main scope of colloidal drug delivery systems is the modulation of the phamacokinetics and/or the tissue distribution of a drug in a beneficial way. The properties of colloidal drug delivery systems to target specific sites of action (organs or tissues) are related to the physicochemical and morphological properties of the carriers, namely, these parameters determine the destination and the fate of the drug entrapped within the carrier system, provided that a drug is released from the system at a suitably controlled rate (230,231). By using colloidal carriers, drugs can be selectively directed to specific sites by applying passive or active strategies of delivery, rather than allowing a free drug diffusion throughout the body by using conventional dosage forms. The carrier physicochemical properties (i.e., size and surface properties) are the main determining factors in passive targeting of colloidal drug carriers. On the other hand, the possibility to achieve a colloidal carrier with active targeting capacity is related to the possibility of inserting specific ligands on the carrier surface so as to achieve a specific receptor-mediated interactions with target cells (232,233).

The potential use of colloidal drug carriers in clinical therapy is strongly related to their *in vivo* fate. In particular, the rapid uptake (following a phagocytosis mechanism) of these carriers by the reticulum endothelial systems (RES), that is abundantly present at the level of the liver, spleen bone marrow, and lungs, is the only fate after their IV administration, thus leading to rapid removal from blood circulation. The phenomenon of opsonization, that is based on binding of some plasma proteins (opsonines) onto the surface of colloidal carriers, is the first step allowing the carrier recognition and binding promotion by phagocytes (234). Therefore, the opportunity to avoid the carrier opsonization is often translated into a deep change of the carrier biodistribution patterns.

In this attempt, colloidal carriers with the ability to avoid RES uptake have been developed, thus achieving long circulating properties (235). The so-called Stealth colloidal carriers are obtained by grafting their surface with hydrophilic macromolecules, mainly poly(ethylene glycols), that hamper the opsonization.

Colloidal drug delivery systems are not able to extravasate, except in tissues and/or organs in which the endothelium is discontinuous (i.e., liver, spleen, and bone marrow) or defective, such as in the case of tumors or in the sites of infection and/or inflammation. Therefore, the therapeutic uses of IV administered colloidal drug delivery devices can be grouped into three cases (235,236): (*1*) drug accumulation in macrophages; (*2*) *in vivo* drug distribution away from the sites of toxicity; (*3*) circulating reservoirs of labile or short blood half-life drugs.

The use of colloidal drug delivery systems has the following advantages: protection, duration, direction, internalization, and amplification.

Protection. Drugs entrapped within colloidal carriers can be protected against both environmental factors (i.e., temperature, UV radiation, moisture) and the

action of detrimental factors of the host (i.e., degradative enzymes). Also, the patient can be protected against toxic effects of administered drugs.

Duration. These carriers can be suitably projected and prepared to achieve a perfectly controlled drug release to fulfill the therapeutic requirements, thus allowing the maintenances of therapeutic (but nontoxic) drug levels in the bloodstream or at the level of local administration sites for a prolonged time. This situation leads to a reduction of administration frequency, and hence to enhanced clinical safety and increased patient compliance.

Direction. As mentioned above, drugs may be passively or actively targeted to specific sites of action by colloidal delivery devices, thus providing an improvement of the drug therapeutic efficacy. These carriers can also provide a site-avoidance delivery, namely, the drug delivery away from sites of their toxicity.

Internalization. Colloidal carriers may be able to promote the intracellular delivery of drugs by ensuring different interaction pathways with target cells in comparison with the free drug that may not be able to reach the inner-cell due to unfavorable physicochemical parameters.

Amplification. In the case of antigen delivery, colloidal drug delivery systems can act as immunological adjuvant in vaccine formulations.

General Colloidal Carrier Classification

By considering the carrier features, colloidal drug delivery devices can be classified into conventional, long circulating and actively targeted systems (Fig. 39).

Conventional colloidal carriers (liposomes and nanoparticles) can be characterized by a wide differences both in terms of composition and physicochemical properties (i.e., size, size distribution, surface charge, number, and fluidity of phospholipid bilayers), in the case of liposomes, matrix compactness, in the case of nanoparticles. The modulation of these properties can influence technological properties, such as colloidal stability, drug loading, drug release rate, and to a certain extent the *in vivo* behavior of conventional colloidal carriers (i.e., blood stability, clearance, and distribution). However, some *in vivo* features are very consistent among different types of conventional colloidal carriers, by presenting a short blood circulation time when parenteraly administered due to a rapid RES uptake. A consequential successful therapeutic use of conventional colloidal carriers characterized by the accumulation at the level of the mononuclear phagocyte system is the delivery of antimicrobial agents to infected macrophages (237,238). Conventional colloidal carriers are also very effective as vaccine adjuvants against viral, bacterial, and parasitic infections (239).

Long-circulating colloidal delivery systems allow the therapeutic treatment of a wide range of diseases involving tissues other than liver and spleen (240). A common characteristic of all long-circulating systems is the presence along the surface of the colloidal carrier of hydrophilic macromolecular moieties, such as polyethylene glycol (PEG). Highly hydrated macromolecular moieties determine a steric barrier against interactions with molecular and cellular components in the biological environment, thus avoiding the opsonization phenomenon, and hence the RES organ uptake (235,240).

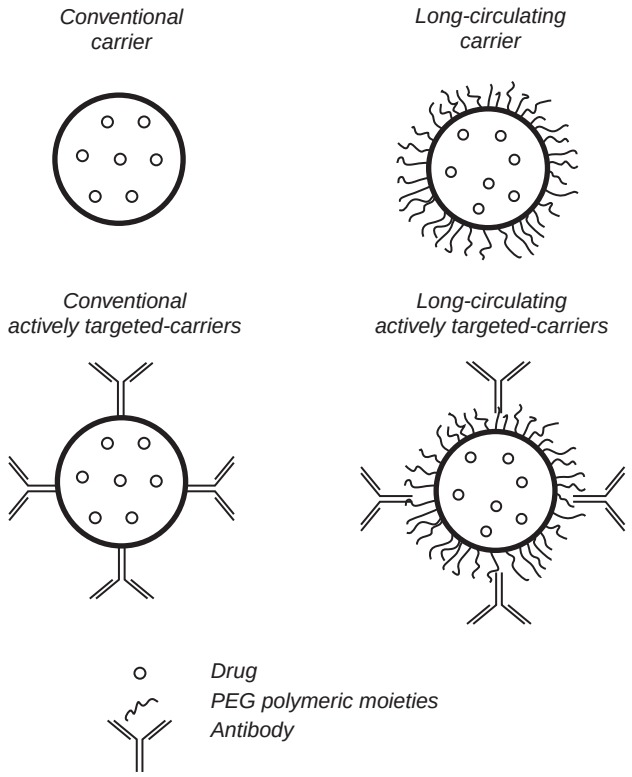


Figure 39. Schematic representation of the various kind of colloidal drug delivery devices. Conventional carriers are made up of a body matrix (phospholipid bilayers in vesicles or polymeric network in nanoparticles) with a hydrophilic colloidal surface (neutral, negatively, or positively charged). Long-circulating systems (the so-called Stealth devices) are coated by hydrophilic polymeric moieties (i.e., PEG) that provide a certain steric stability, and hence reduction of opsonization. Actively targeted carriers (i.e., antibody targeted) can be of conventional (targeting-agent conjugated directly to colloidal carrier surface) or sterically stabilized type (targeting-agent conjugated with a PEG moiety anchored to the surface of the colloidal carrier).

To obtain a certain specificity, actively targeted carriers can be obtained by conjugation of a colloidal drug delivery systems to specific antibodies, antibody fragments (e.g., Fab or single-chain antibodies), or small targeting agents (peptides, hormons, specific ligands), thus increasing target site binding and the delivery of the encapsulated drug. In the first generation of these kind of colloidal carriers, the active targeting agent was conjugated directly to their surface. This strategy led to a successful *in vitro* recognition and activity, but to a failure in *in vivo* applications due to the RES uptake. The last generation of actively targeted carriers is represented by long-circulating colloidal carriers with the PEG moieties conjugated with the targeting agent, thus presenting suitable *in vivo* features.

Colloidal Carrier Characterization

For routine measurements of particle sizes, two techniques are commonly used. Photon correlation spectroscopy (PCS) (also know as dynamic light scattering), which measures the fluctuation of the intensity of light when it is scattered by particles movements. The particle diameter range goes from a few nanometers to $\sim 3\ \mu m$, so PCS is not useful in for lipid particles $>3\ \mu m$. In these cases, a laser diffraction (LD) technique is used. This method is based on the relation between the diffraction angle and the particle radius, so that smaller particles cause more intense scattering at high angles compared to the larger ones.

In general, it is recommended to use both techniques simultaneously in order to obtain precise data. However, it should be kept in mind that both PCS and LD do not measure particle sizes directly, they only correlate light scattering to particle size.

To obtain direct information on particle sizes and shapes, electron microscopy (EM) is used. Electron microscopy extracts structural information carried by the scattered electrons; the most commonly used EM techniques are transmission electron microscopy (TEM) and SEM. Atomic force microscopy (AFM) is another microscopic technique that is getting increasing attention. This method is based on the interactive forces between a surface and a probing tip that leads to the imaging of particles. This technique has the clear advantage of simplicity of sample preparation, so that it is possible to conduct analysis directly on the hydrated, solvent containing samples (241).

The field-flow fractionation (FFF) is a technique recently used for measurements of solid lipid nanoparticle sizes. It is based on the different effect of a perpendicular applied field on particles in a laminar flow (242); the characterization of particles is based on the different nature of perpendicular fields, for example, sedimentation size (cross-flow FFF) or charge (electric field FFF). All these principles can be used combined together in order to obtain unique resolution.

The determination of a zeta potential is predictive of the storage stability of colloidal dispersions (243). In general, the greater the zeta potential value of a nanoparticulate system, the better the colloidal suspension stability due to a repulsion effect between charged nanoparticles. Nanoparticle stability can also be obtained by the addition of some polymers, such as PEG, which adhere to the particle surface stabilizing it. Surface characteristics are also important for the *in vivo* fate and the interaction with biological systems of colloidal carriers.

The characterization of the physical state of colloidal carriers (particularly vesicles and lipid-based particles) can be efficiently carried out by two techniques, DSC and X ray. The DSC method is based on the fact that different material polymorphic form possess different melting points and melting enthalpies (244) and that changes in thermotropic parameters of a systems are usually evidence of different spontaneous and/or induced arrangements. X-ray techniques allow the characterization of polymorphic forms and the determination of large and small spaces in an ordered matrix, such as the lipid grid of a solid lipid nanoparticle (245). The advantages of these two techniques are the possibility of particle suspension analysis without drying the solvent, thus avoiding possible modifications of the carrier structure.

Also, NMR and electron spin resonance (ESR), are used for the investigation of dynamic phenomena in colloidal lipid dispersions. Nuclear magnetic resonance is based on the different proton relaxation times in the liquid and semisolid–solid state (246). The NMR technique can also be used to determine lamellarity in vesicular carriers (247). The ESR technique uses a paramagnetic spin probe to give a noninvasive characterization of the distribution of the spin probe between hydrophilic and hydrophobic phases. Both NMR and ESR are noninvasive methods and allow repeated measurements of the same sample.

Vesicular Drug Carriers

Drug delivery systems composed of lipidic compounds have gained great importance in medical, pharmaceutical, cosmetic, and alimentary fields. Formulations based on phospholipids and other excipients represent an interesting field of application in the novel research for delivery models.

Lipidic materials are characterized by their possibility to self-organize in different supramolecular arrangements as a function of some environmental factors (i.e., temperature, lipid concentration, type of medium, ionic strength, pH value, and presence of other compounds). Among the various supramolecular forms of aggregation, the bilayer structure, and hence the formation of vesicles (defined as a lipid bilayer surrounding an aqueous space) represents the most suitable device in terms of drug delivery. In fact, vesicles are boundary structures (Fig. 40), in which it is possible to have at the same time vari'ous microenvironments characterized by different physicochemical properties, namely, a highly hydrophilic region made up of the intravesicular aqueous compartment, a highly hydrophobic region of the bilayer core made up of the alkyl chains of the lipid constituent, and an amphipatic region at the level of the vesicular surface made up of the polar lipid headgroups. These peculiarities make vesicular systems a very versatile drug carrier being able to entrap and delivery hydrophilic (in the intravesicular aqueous compartment), hydrophobic (in the core of vesicular bilayers), and amphipatic (at the level of vesicular boundary zone) drugs.

An important feature that make vesicles a unique drug delivery system is the biomimetism of having the same supramolecular lipid organization of natural membrane living cells.

Therefore, the possibility to create a structure similar to the biological membrane for carrying out the delivery of drugs has represented an interesting challenge for a number of researchers. In particular, liposomes, ethosomes, transfersomes and niosomes have been extensively investigated and are up to now the main vesicular systems used in drug delivery.

Liposomal Carrier. The appearance of Banghman's vesicle in the mid-1960s, the so-called Liposome, represented a milestone in the field of innovative drug delivery. Liposomes are mostly made up of phospholipids, and for

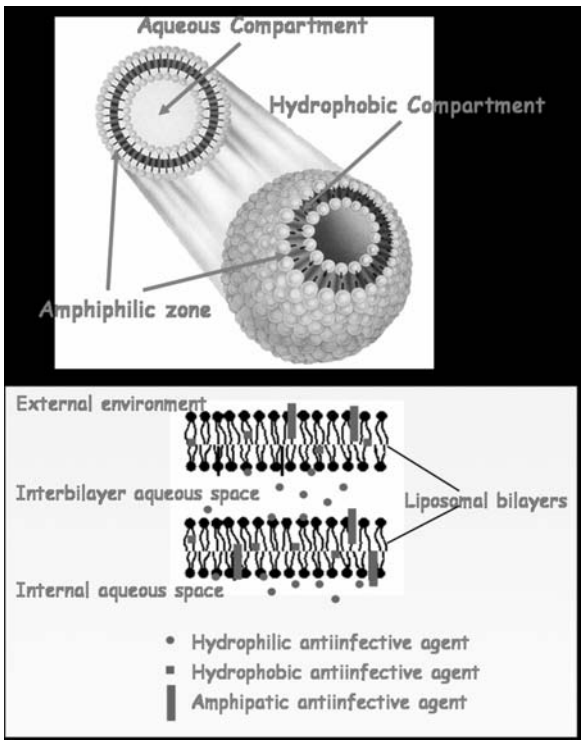


Figure 40. Schematic representation of a liposomal structure with the characteristic microenvironments.

this reason they are highly biocompatible and biodegradable. The liposomal carrier has the advantage of also being able to deliver macromolecules, such as enzymes, proteins, and genetic material (248).

From the morphological point of view (Table 9), liposome systems can be classified as a function of the number of bilayers and the mean size of the carrier in unilamellar, oligolamellar or multilamellar vesicles, and in small ($<$100 nm), medium (100–500 nm) and large ($>$1 μm) vesicles, respectively.

Lipid Component Used in Liposomal Formulations. Lecithins and cholesterol (Chol) are the lipids most commonly used in the preparation of liposomes. Other components can be used in the liposome preparation, that is, steroid molecules, charged phospholipids, ganglioside, and polymeric material to modulate the carrier properties as a function of the therapeutic requirements to be achieved (249). In fact, different components can modify the biodistribution, the surface charge, the release, and the clearance rate of the liposomal drug delivery system (249,250). The circulation lifetime of a liposome is also altered by the charge of the liposome surface that can influence the pharmacokinetic of the system (251).

It was demonstrated that the use of negatively charged lipids [i.e., phosphatic acid (PA), phosphatidylserine (PS), phosphatydilglycerol (PG)] are able to elicit a rapid clearance of the liposomal system from the blood stream mediated by the RES uptake (249,252).

Cholesterol plays a fundamental role in liposome formulations being, able to act as a vesicle membrane modulator as concern membrane fluidity. It has a stabilizing function on the liposome bilayers both *in vitro* and *in vivo*, allowing the protection of the vesicular structure by the action of blood high density lipoproteins (HDL) and hence the possibility of having a prolonged circulation of intact liposomes (253).

Similarly to cholesterol, some phospholipid components are also able to influence the physicochemical behaviors of liposomes to obtain a more rigid vesicular structure that is much more resistant to the phospholipid extraction effect mediated by blood HDL. In this attempt, both 1,2-distearoyl-3-*sn*-phasphatidylcholine (DSPC) and sphingomyelin (SM) have been used to maintain a certain vesicular carrier integrity following IV administration. A rigid vesicular structure of liposomes hampers an effective adsorption of opsonine and prolongs the plasmatic level of the drug carrier, that is, liposomes made-up of 1,2-distearoyl-3-*sn*-phasphatidylcholine–cholesterol (DSPE–Chol) showed higher half-time than liposomes prepared with phosphatidylcholine (PC) or 1,2-dipalmitoyl-3-*sn*-phasphatidylcholine (DPPC) (253). In particular, SM has an additional stabilizing effect on the liposome formulations when used together with Chol (254). In this case, SM can interact with cholesterol by forming intermolecular hydrogen bonds and eliciting an increased compactness of the liposomal bilayers that leads to an improved serum stability (249).

Since the appearance of liposomes, positively charged lipids were introduced in liposome composition to obtain a vesicular system characterized by a net positive charge

Table 9. Main Characteristics of the Various Liposome System

Liposome Type	Abbreviation	Properties
Multi-lamellar vesicles	MLVs	Vesicles constituted by 7–15 bilayers with a mean size $>1.5\,\mu m$
Multi-vesicular vesicles	MVVs	Vesicles constituted by 3–5 vesicles contained within a bigger one. The mean size is $>1.5\,\mu m$
Oligo lamellar vesicles	OLVs	Vesicles constituted by 2–5 bilayers with a mean size $\sim 1\,\mu m$
Giant unilamellar vesicles	GUVs	Vesicles constituted by only one bilayers with mean size $\geq 1\,\mu m$
Large unilamellar vesicles	LUVs	Vesicles constituted by only one bilayers with mean size ranging from 400 to 800 nm
Medium unilamellar vesicles	MUVs	Vesicles constituted by only one bilayers with mean size ranging from 200 to 400 nm
Small unilamellar vesicles	SUVs	Vesicles constituted by only one bilayers with mean size ranging from 30 to 100 nm

along the liposomal surface. In the last decade, positively charged liposomes have gained much more interest than in the past due to their potential application as carriers for genetic material delivery (255). In this attempt, the most frequently used cationic lipids are DMRIE, *N*-(2-hydroxyethyl)-*N*,*N*-dimethyl-2,3-bis(tetradecyloxy)-1-propanaminium bromide; dioctadecyl amino glycyl spermine (DOGS); dioleoylphosphatidylethanolamine (DOPE); 2,3-dioleyloxy-*N*-[2(spermine carboxaminino)-ethyl]-*N*,*N*-dimethyl-1-propanaminium trifluoroacetate (DOSPA); 1,2-dioleoyl-3-trimethylammonium propane (DOTAP); 2,3-bis(oleyl)oxipropyl-trimethylammonium chloride (DOTMA). Cationic liposomes composed of DOTMA and DOPE became commercially available as a transfection reagent designated Lipofectin.

The above mentioned cationic lipid components of this particular kind of liposomes are able to interact with, and neutralize, the negatively charged DNA or ribonucleic acid (RNA). This interaction leads to a genetic material condensation into a more compact structure. The resulting lipid–genetic material complexes (lipoplexes), rather than DNA or RNA encapsulation within liposomes, provide protection and promote cellular internalization and expression of the condensed plasmid (255).

Most recently, amphiphilic polymeric materials have been introduced in the composition of vesicles to cover their surface by inserting their hydrophobic domain in the liposomal bilayers (anchor moiety) and facing the hydrophilic domain toward the aqueous environment (shield moiety). This advance allowed a further modularity of the liposomal carrier by conjugating together the advances of colloidal drug delivery devices (carrier capacity) with those of macromolecules (fine chemical approach and infinite modulation potentiality) (249). The principal polymer used to cover the surface of liposome formulation was polyethylene glycol. This is a flexible-chained hydrophilic polymer of different molecular weight (i.e., PEG-750, PEG-2000, PEG-5000) conjugated to phosphatidylethanolamine (PE) or distearoylphosphatidylethanolamine (DSPE) (256). Liposomes containing PEG in their structure (the so-called pegylated liposomes) represented an important class of vesicular delivery systems that started the new generation of liposome carriers. The presence of this hydrophilic polymer on the surface of liposomes not only is able to reduce the RES uptake and to increase the blood circulation time (249), but it can also modulate some pattern of interaction with cultured cells, such as the intracellular drug delivery (257).

Another important aspect for drug delivery by liposomes is the possibility to achieve a triggered release of the encapsulated agent from the carrier following certain stimuli. Targeted drug delivery is based on the fact that upon attachment to the target site, or delivery into the target cell, the therapeutic agent must be released from the carrier to exert its action. When liposomes are taken up by the target cell through endocytosis, they come into contact with acidic conditions. For some drugs and biothecnological products (e.g., peptides and genetic material) it could be essential to escape from liposomes and endosomes, thus entering the cytosol before reaching the lysosomal structures with their highly efficient degradation machinery. Liposome destabilization under acidic conditions and bilayer fusogenic properties are required to achieve lysosome escape. Besides the pH-dependent liposome release, other triggered releases may be accomplished for certain drug selectivity, namely, bilayer composition controlled release, destabilization by removal of bilayer components, complement-induced leakage, and temperature-induced destabilization of the liposomal bilayer structure. Therefore, to have a triggered liposomal carrier release, some compounds that are stimuli responsive must be introduced in the liposomal bilayer composition (e.g., DOPE, cholesteryl hemisuccinate, oleic acid, fusogenic peptides) (258,259).

Main Methods to Prepare Liposomes. As reported in Table 9, various types of liposomes exist, each of those with specific peculiarities that make them suitable for certain therapeutic applications. Although aqueous dispersions of phospholipids spontaneously lead to a self-aggregation into closed bilayers, vesicles, particular procedures must be carried out if a certain type of liposome has to be obtained. In fact, this type of liposome is mainly determined by the preparation procedure, and for these reason the main preparation methods are reported below.

Thin-Layer Evaporation (TLE). This method allows the formation of multilamellar vesicles. Basically, a mixture of lipid compounds is dissolved by an organic solvent (chloroform) or a mixture of two organic solvents (chloroform–methanol) in a round-bottomed flask. Other hydrophobic components (e.g., drugs) can be cosolubilized with the liposome-forming materials. The complete evaporation of the organic solvent by a rotavapor lead to the formation of a thin lipid film along the surface of the glass wall. This lipid film is then hydrated with an aqueous solution buffered to the desired pH value and solubilizing any hydrophilic component that should be entrapped within liposomes (e.g., water-soluble drugs). The hydration temperature is normally higher than the highest transition temperature (T_m) of lipids used in the film preparation. In some cases, to increase the surface of film deposition, and hence the surface undergoing buffer hydration, glass beads can be added during the TLE preparation procedures (260).

Reverse-Phase Evaporation Vesicles (REVs). This method allows us to obtain large unilamellar, oligolamellar, and multilamellar vesicles. A lipid film, formed as reported in the TLE method, is dissolved in an organic solvent (diethyl ether) and an aqueous solution is added. This two-phase mixture is energetically sonicated, thus obtaining an w/o emulsion. The organic solvent constituting the external phase of the w/o emulsion is gradually removed by a rotavapor up to the reversion of the phases with the appearance of an external hydrophilic phase. The total removal of the organic solvent leads to the formation of a gel-like highly concentrated liposome suspension that can be suitably diluted with a suitable aqueous buffer solution. This method represents the first approach used in the attempt to increase the amount of drug entrapped within vesicles (261,262).

Freeze and Thawed Multilamellar Vesicles (FAT–MLVs). A multilamellar liposome formulation obtained with the TLE method is subjected to a serious of cycles of freezing in liquid nitrogen and thawing in warm water ($\sim 40\,^{\circ}C$). At the end of the procedure, liposomes are kept at room temperature to stabilize the bilayer. This procedure is carried out to obtain a multilamellar liposomes with a homogeneous distribution of solutes throughout the various multilamellar aqueous compartment (263,264).

Dehydration Rehydration Vesicles (DRVs). Multilamellar liposomes obtained with one of the previous methods are submitted to a freezing-drying process. The product of liophylization is resuspended in an aqueous solution (265). This method leads to the formation of oligolamellar or multilamellar liposomes with an high drug entrapment efficiency.

Vesicles by Extrusion Technique (VET). The reduction of the mean size of a colloidal liposomal suspension characterized also by a narrow size distribution can be achieved with the extrusion of multilamellar liposomes through polycarbonate membranes of different sizes (from 400 to 50 nm). Usually, 10 cycles of extrusion are carried out to obtain an homogeneous formulation. Both LUV and SUV are obtained following the VET method (266).

pH Gradient Loading Method. This method is used to increase the loading capacity of liposomes in regard to ionizable drugs. This method is based on the formation of a pH gradient between the inner-liposomal aqueous phase and the external environment. This situation promotes the protonation or deprotonation of an entrapped drug thus favoring its accumulation within the vesicular carrier due to the incapability of a ionized molecule to freely diffuse through a lipid bilayer (Fig. 41) (267). Ammonium sulfate or ammonium citrate are used to obtain an acid pH environment while calcium acetate to have basic conditions (250,268). The efficiency of liposome drug loading using the method of pH gradient is influenced by the drug partition coefficient between the aqueous phase and the lipid bilayer (269).

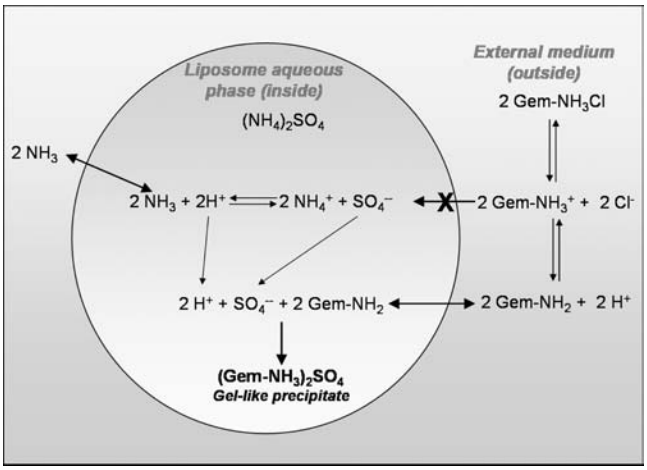


Figure 41. A schematic representation of the encapsulation of Gemcitabine in multilamellar liposomes by using a pH gradient method in the presence of ammonium sulfate 250 m*M* (267).

One of the most important parameters for an ideal drug delivery system is the drug loading capacity. The amount of drug encapsulated in liposome formulation is influenced by a serious of parameters, such as the preparation method, the size of the liposome, and the type of lipid used to form the lipid film (263). Therefore, to have a colloidal liposome system with particular carrier properties, it is often necessary to carry out two or more preparation procedures. Namely, the DRV or FAT procedure can be carried out to improve the liposome encapsulation capacity, and then the VET method to obtain a small mean size with a narrow size distribution. These two aspects (carrier capacity and mean size) are very important for liposomes to be proposed for certain therapeutic application (i.e., antitumoral chemotherapy).

The removal of untrapped drug is the last step in the preparation of a drug-loaded liposome colloidal suspension. Many lipophilic drugs exhibit a high affinity to the bilayer and are completely liposome associated. For compounds with an encapsulation $< 100\%$, the nonencapsulated fraction of the drug may determine unacceptable side effects. The removal of the untrapped drug can be carried out by the following techniques: dialysis, ultracentrifugation, ultrafiltration, gel permeation chromatography, and ion exchange reactions.

Liposome Stability. An ideal drug delivery system should maintain its physicochemical characteristics during storage, that is, mean size, size distribution, thermotropic parameters, no lipid degradation (hydrolysis and/or peroxidation), no appearance of microbial flora, to be considered for practical applications. Liposomes are self-assembled colloidal carriers, and hence their stability can be strongly influenced by the component used for their preparation, considering that the presence of foreign molecules in the liposomal bilayers deeply influence their mode and strength of aggregation in a concentration-dependent manner. For this reason, in the case of drugs to be delivered by liposomes and characterized by liposomal bilayer localization, particular attention should be paid to the drug/lipid ratio. This is a very important parameter because the payload of the drug can be increased with a consequential reduction of the system stability. In some cases, the segregation of the lipid bilayer components in various microdomains can be observed (270).

The osmolarity of liposomes seems to be a very important factor to achieve a stable liposomal system. Some studies (271) showed that hypertonic conditions triggered a rapid drug release from Ara-C-loaded liposomes and that the release kinetic is characterized by a biphasic profile with a first step of very rapid and massive Ara-C release followed by a second phase of slow drug release (249).

The chemical stability of liposome formulations mainly depends on the chemical characteristics of both drugs and lipid component used for the carrier preparation (272). The

presence in the phospholipid bilayers of polyunsatured fatty acid moieties, such as arachidonic, linoleic, or linolenic acid, can favor the occurrence of peroxidation processes at the level of single or conjugated double bonds. The membrane lipid peroxidation can destabilize liposomal bilayers due to the formation of secondary oxidation products that can change the integrity of the liposome structure (249).

Main Therapeutic Applications of Liposome. The following criteria should be taken into account to evaluate the possibility of delivering a drug by using the liposomal carrier: (*1*) the chosen drug should be sufficiently active; (*2*) the drug should be efficiently entrapped within liposomes; (*3*) the drug must be compatible with the liposomal carrier.

A basic concept for the success of liposome drug delivery is the fact that the encapsulated agent may be released at a suitable rate to become bioavailable upon arrival at the action site. Liposomes protect drugs from metabolism and inactivation in plasma and also allow a reduction of the drug accumulation in healthy tissues and/or organs, due to size restrictions in the transport of large macromolecules and carriers across healthy endothelium (271). A number of pathologies (i.e., cancer, stroke, infections, and some methabolic diseases) are characterized by direct or mediated inflammation, which elicits discontinuities in the endothelium vasculature of the diseased zone. This thus increases the extravasation of colloidal carriers and, in combination with an impaired lymphatics and a high value of interstitial pressure, the accumulation of the therapeutic agent-loaded liposomes at the level of the diseased site (passive targeting). This situation, referred to as enhanced permeation and retention (EPR) phenomenon, consequently elicits an increase of the drug therapeutic index (249).

A successful therapeutic approach of the liposomal passive targeting is the efficacious delivery both *in vitro* and *in vivo* of various anticancer drugs (249). As shown in Fig. 42, the use of pegylated liposomes (Stealth liposomes) with a mean size of $\sim 100\,\text{nm}$ allows the passage of the carrier in the tumor tissue and a local accumulation of the encapsulated drug. Furthermore, liposomal chemotherapeutic agents display distinctive pharmacokinetic characteristics, because they possess longer elimination half-lives, reduced clearance, and smaller volume of distribution with respect to corresponding free drugs. Taken together, these features lead to the highest levels of cytotoxic agents in tumors, as demonstrated in preclinical models and clinical trials, whereas healthy tissues are spared from toxicity. Liposomal anticancer drugs lead to improved clinical effectiveness and better toxicity profile with respect to corresponding free drugs when they are used for the treatment of metastatic tumors (e.g., breast and ovarian cancers). A successful example of antitumoral agent-loaded long-circulating liposomes is Doxil, a doxorubicin-loaded pegylated liposomes with a 100 nm mean size.

This innovative liposomal formulation is currently approved for use in AIDS-related Kaposi's sarcoma and refractory ovarian cancer. It has also shown activity in other tumors, including metastatic breast cancer. A preclinical toxicology study of IV administered doxorubicin-loaded stealth liposomes compared to the free drug showed that the drug liposomal formulation was less toxic (LD50 $32\,\text{mg}\cdot\text{kg}^{-1}$) than the free doxorubicin (LD50 $17\,\text{mg}\cdot\text{kg}^{-1}$). The organ specific toxicities seen with Doxil were qualitatively similar to those of free doxorubicin, but less severe (273). In addition, Doxil accumulates in tumor tissues to a large extent with respect to the free drug due to its capacity to escape macrophagic uptake (274). Reduced toxicity and selectivity are the reasons of the improvement of doxorubicin therapeutic index.

A recent and very active field of research in the liposomal anticancer chemotherapy is the active targeting of long-circulating liposomes (249,275,276). A high density of the targeting moiety on the surface of liposomes is very important to have an efficient binding to the target site, a specific antigen or receptor expressed on the surface of target cell. This interaction increases the amount of drug in the target site and it decreases the systemic side effects (275).

Antibodies, particularly monoclonal, are the more versatile ligands that can be conjugated on the liposome surface (the so-called immunoliposomes). In the past, an obstacle in the use of immunoliposomes was the antigenicity of murine antibodies that were easily available, however, the more recent availability of humanized forms should contribute to overcome this problem. Important

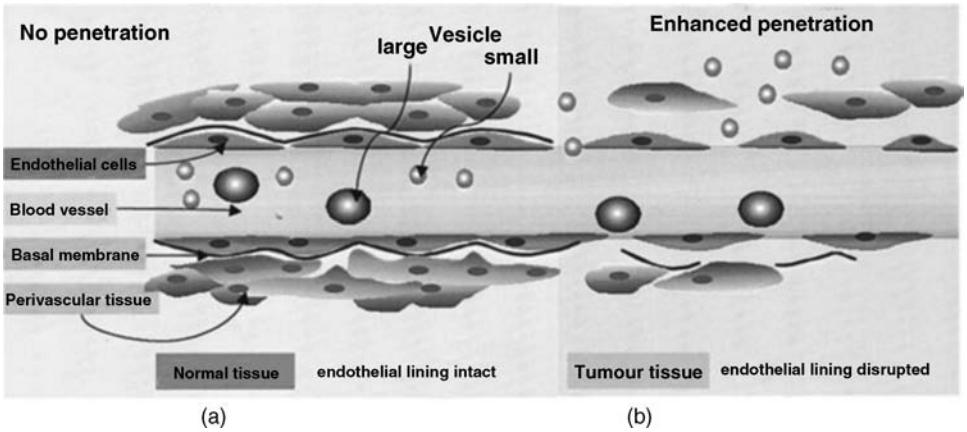


Figure 42. Schematization of the accumulation mechanism of long-circulating small unilamellar liposomes in solid tumor. Extravasation of liposomes through vascular endothelium of the tumor site (a); behaviour of liposomes in a normal tissue (b).

parameters for immunoliposomes are the ability to become selectively cell associated and the ability to deliver the loaded drug within target cells. In the case of immunoliposomes endocytosis seems to be the predominant mode of delivery to the cells, and hence has an efficient intracellular delivery. The mean size of immunoliposomes should be ≤ 100 nm.

Given a suitable antibody with high specificity and affinity for the target antigen, the critical factor is the *in vivo* accessibility of target cells to the immunoliposomes. To have an efficient target binding of the injected immunoliposomes, target cells should be located in the intravascular compartment and/or in accessible tissues and organs characterized by leaky vascular structures. Thus, in terms of targeting drug delivery by immunoliposomes, two anatomical compartments can be considered. One is a readily accessible intravascular site, such as the vascular endothelial surface, T cells, B cells, or a thrombus. The other is a much less accessible extravascular site, such as a solid tumor, an infection site, or an inflammation site, where the vascular structure is leaky (277).

Antibiotics encapsulation in liposomes is of great utility in the case of very potent drugs that can be administered intravenously and present a certain toxicity (i.e., nephro- and neurotoxicity). The toxicity of antibiotics limits their dosing, and hence the drug efficacy. Antimicrobial agent-loaded liposomes were used for the treatment of various obligate and facultative bacterial infections (i.e., *Salmonella*, *Listeria*, *Brucella*, *Mycobacterium*, *Staphylococcus* and *Escherichia coli*) (278). Obligate microbes are more difficult to eradicate due to the fact that they can multiply only within host cells, while facultative bacteria can be reached by the drug in the extracellular compartment. The conventional liposome biodistribution properties represent a noticeable advantage for treatment of infections in which bacteria are taken up and/or reside in the cells of the phagocytic systems. Another advantage of the liposome carrier is the capability to facilitate the entrance within infected cells of antimicrobial agents that are not able to cross cell membranes with a consequential intracellular drug accumulation (279) (Fig. 43). An intrabacterial antibiotic drug accumulation was also observed (280), thus showing that liposome formulations can contribute to overcome bacterial resistance phenomena due to drug impermeability (Fig. 44). In particular, in the case of intracellularly infected phagocytic cells (e.g., *Legionella pneumophila*, *Mycobacterium tuberculosis*, *Listeria monocytogenes*, and *Staphylococcus aureus*) a 10–100 times increased efficacy has been reported for the antibacterial agent-loaded liposome formulation compared to the free drug (278) both *in vitro* and *in vivo*. The specific macrophage targeting of liposomes can be further improved by grafting the surface of liposomes with carbohydrate moieties whose receptors are expressed along the surface of macrophages. This possibility may lead to an additional efficacy of the liposome delivery device in the *in vivo* treatment of intramacrophagic infections.

Long-circulating liposomes with a reduced size have the opportunity to accumulate in the infection site according to the mechanism reported in Fig. 42. In an experimental *in vivo* model, a large accumulation of long-circulating liposomes in the infected lung was observed; while no presence

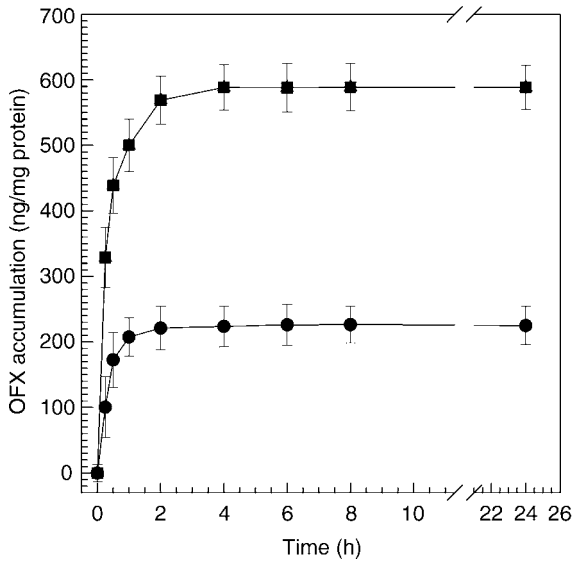


Figure 43. Accumulation profiles of ofloxacin into McCoy cells as a function of time. The biological assay was carried out at room temperature (20 °C) by adding 5.7 $\mu g \cdot mL^{-1}$ of free (●) or liposome entrapped (■) ofloxacin into confluent McCoy cells. Each point represents the average of nine different experiments ± standard deviation. Data from Ref. 279.

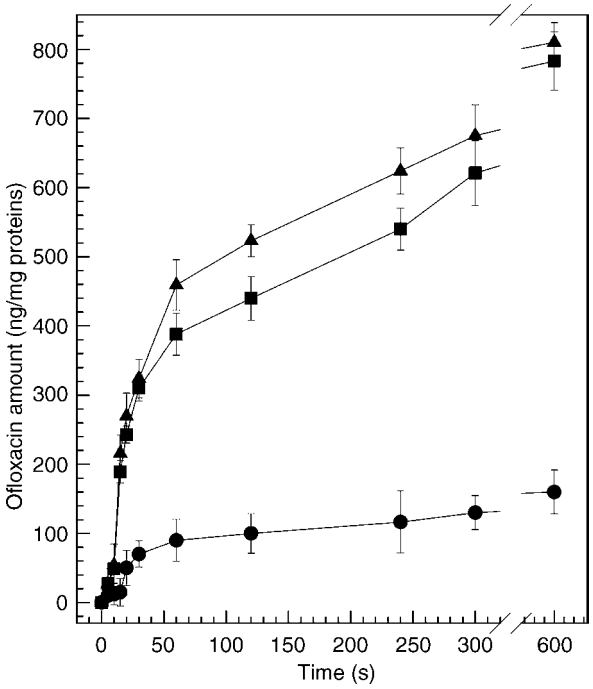


Figure 44. Intrabacterial accumulation of ofloxacin-loaded MC-Chol-DP (4:3:4 *M* ratio) unilamellar liposomes within *E.coli* ATCC 25922 (■) and *E.coli* ATCC 35218 (▲) versus the free drug (*E.coli* ATCC 35218 accumulation) (●) as a function of time. Free drug accumulation within both *E.coli* strains is very similar (data not reported). The experiments were carried out at 37 °C. Each point represents the mean value of five different experiments ± S.D. Data from Ref. 280.

of long-circulating liposomes was noted in the noninfected lung. Interestingly, the accumulation extent seemed to be a function of the severity of the infection (281). Therefore, Stealth liposomes offer targeting to the deep tissues, which can harbor *Mycobacterium avium intercellulare*. The chance of using Stealth liposomes containing some new and potent antibacterial agents, (e.g., fluoroquinolones) can represent a real improvement in the therapy for the eradication of infections situated in organs and tissues other than the RES.

Liposomes can be suitable delivery devices in antiviral chemotherapy (282) due to their capability of delivering entrapped drugs across cell membranes (257,279). This aspect is of fundamental importance in antiviral chemotherapy, because the nature of virus action and proliferation is intracellular. In particular, the liposomal therapy of viral infections can be accomplished by two different approaches: (*1*) the encapsulation of the antiviral drug having a liposome-mediated antiviral activity; (*2*) the encapsulation of immunomodulators, such as lymphokines (macrophage activation factor, MAF), thus achieving an activation of the macrophages.

Liposomal antiviral chemotherapy can offer special targeting possibilities due to the natural ability of viruses to fuse with cellular membranes. In this case, various antiviral therapeutic approaches can be achieved by the following strategies: (*1*) the administration of drug-loaded long-circulating liposomes bearing cellular antigens that attract and destroy viruses; (*2*) the saturation of the cell receptor by binding other antigens delivered with liposomes; (*3*) the reconstitution of viral glycoproteins onto liposomes (the so-called virosomes), which are characterized by a very strong fusogenic activity depending also on the vesicle lipid composition. Such virosomes can bind to and fuse selectively with the infected cells. Therefore, this particular carrier can ensure a very effective and specific intracellular antiviral therapy.

Liposomal antiviral chemotherapy, for example, can be efficaciously used for the treatment of HIV infection. The encapsulation of gelonin (a plant toxin) allowed a selective killing of human immunodeficiency virus (HIV) infected cells (283). Another success with respect to HIV therapy was observed in the case of treatment with liposomes containing fragment A of diphtheria toxin, which was toxic to HIV infected cells, but not to uninfected cells (284).

Another application of liposomes in antiinfective chemotherapy can be the treatment of fungal infection. Invasive fungal infections are among the most important causes of morbility and mortality in immunocompromised patients. Amphotericin B and nystatin are the most widely used drugs in the treatment of systemic fungal infections (285). These two drugs show some drawbacks when used *in vivo* in the treatment of mycosis (i.e., nephrotoxicity and side effects at the level of the CNS). In this case, liposomes are a suitable colloidal carriers for amphotericin B, not being able to accumulate in the kidneys (e.g., of the site avoidance mode of liposome action) and the nervous system and providing a smart system to efficaciously solubilize amphotericin B. As for other pathological situations, the most important advantage of the liposomal carrier is its ability to accumulate at the level of the same cells where fungi are localized (mainly the RES). The improved selectivity and the reduced toxicity determined the noticeable increase of the amphotericin B therapeutic index. Considering the consistent therapeutic advantages of amphotericin B-loaded liposomes (286), a new liposomal formulation was produced and commercialized by Vestar, Inc., with the name of AmBisome. This pharmaceutical formulation is made up of phosphatidylcholine, cholesterol, distearoylglycerol, and amphotericin B (2:1:0.8:0.4 molar ration) with a 9.5 lipid/drug ratio. The mean size of these small unilamellar liposomes ranges from 45 to 80 nm.

Infective diseases caused by parasites are a great problem for developing countries. In these particular infections, especially for those pathologies where the infection agent is closely associated to the RES, the possibility of delivering already existing drugs by liposomes can represent a very attractive strategy. In fact, due to poor drug membrane penetration, *in vivo* treatments of these pathologies are often poorly effective, despite the *in vitro* effective activity of the drug. An interesting example of effective liposome treatment of protozoal diseases is leishmaniasis.

The parasites of leishmaniasis live almost exclusively in fixed macrophages at the level of the RES (liver, spleen, and the rest of the visceral). Antimonial derivatives (therapeutic index approaching 1) are the most effective drugs for this pathology. Liposomal formulations of these drugs can improve the therapeutic effectiveness up to a thousand times with respect to the free drug. Experiments showed that doses close to the lethal level of free potassium antimony tartrate were ineffective, but a single dose (40% of the previous dose; $20\,mg \cdot kg^{-1}$) of drug-loaded liposomes completely eliminated the parasites (287).

Liposomes can also be used as immunoadjuvants for vaccines (288) and as macrophage activators against tumoral, viral, and microbial cells. For both applications, a substance is delivered to macrophages thus triggering immunization, immunomodulation, or activation by means of antigens. The presence in the liposome structure of a nonliposomal adjuvant, that is, muramyl tripeptide covalently coupled to phosphatidylethanolamine, can enhance the antibody response induced by liposome-associated antigens.

Liposomes can be efficaciously used to deliver to the CNS. Under some pathological conditions (i.e., tumors, ischemia, and traumatic shocks) a hypermeabilization of the blood–brain barrier can occur, thus allowing the passage of very small aggregates (< 100 nm). The CDP–choline loaded very small (50 nm) long circulating liposomes were used to treat successfully the cerebral ischemia (289,290). The drug-loaded liposome was able to increase the amount of drug that reached the brain and the survival rate of rats submitted to ischemia and reperfusion (Fig. 45). The liposomal formulation is also able to efficaciously antagonize the phenomenon of postischemic damage maturation that is the main reason of a poor neuronal recovery and hence of an enlargement of the damaged (291).

Liposome formulations resulted effective not only in systemic administration, but also in topical administration (e.g., dermal, mucosal, ocular, pulmonary).

The potential application of liposomes as dermal delivery systems has been extensively investigated, with regard

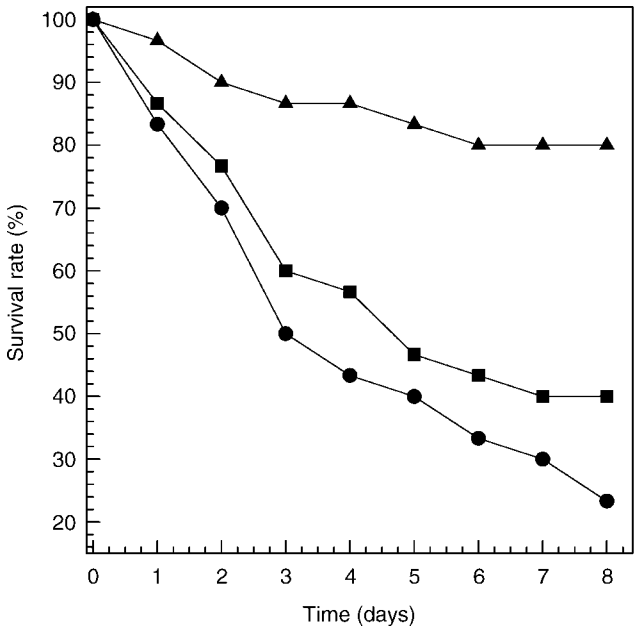


Figure 45. Survival rate of postischemic reperfused Wistar rats (320–350 g). The duration of the ischemic event was 30 min. The rats were treated with saline ●, with the free drug ■, or with CDP–choline loaded liposomes ▲. Unloaded liposomes showed no effect on rat survival (data not reported). The results are expressed as the percentage of the total number of animals in each group which survived ischemia as a function of time. Data from Ref. 290.

to vesicle composition and size (292–294). It was proposed (295) that the main advantages of using liposomes as topical drug formulations were due to their demonstrated ability: (*1*) to reduce serious drawbacks and incompatibilities that may arise from undesirably high systemic absorption of drugs; (*2*) to enhance accumulation of drugs at administration sites due to the high substantivity of liposomes with biological membranes; and (*3*) to the possibility to incorporate both hydrophilc and hydrophobic drugs. In addition, liposomes can be readily prepared on a large scale. The requirement of smart drug delivery systems for skin application comes from the necessity to have a modulation of both the administration rate and the skin permeation properties, namely, a sustained drug release strictly confined at the level of the skin with no systemic absorption or an enhanced transdermal effect to deliver the drug to some inner structures (e.g., joints) or to achieve a systemic effect are required as a function of the disease to be treated (296). By the use of quantitative skin autoradiography, it was demonstrated that small liposomes allowed the localization of a greatest amount of caffeine (hydrophilic drug) in the epidermis and a lowest amount in the dermis and appendages (297). In this case, liposomes ensured a drug skin accumulation three times greater that that observed for an aqueous drug solution prepared in the presence of penetration enhancers.

The liposome lipid composition and the thermodynamic state of the liposomal bilayers play a crucial role in the effect of this vesicular carrier on drug transport rate across the skin. In particular, incorporation of drugs in the liquid-state liposomes provides a higher skin permeation rate than that observed for drug-loaded gel-state (the so-called solid) liposomes (298). Liposomes made up of the same lipids usually present in the skin were prepared and referred to as skin-lipid liposomes (299). These kind of liposomes are able to provide a drug dermal delivery of the highest drug disposition within the deeper skin layers, that is, in the epidermis and dermis, while avoiding systemic drug adsorption (299). For example, skin-lipid liposomes can be a suitable topical carrier for chronic topical applications of corticosteroids by optimizing drug concentration at the site of action while minimizing systemic absorption and, as a consequence, possible side effects (300). In the case of transdermal drug delivery requirements, the high deformability of vesicular carriers seems to be a fundamental feature to achieve the intact vesicles penetration, thus also favoring the delivery of encapsulated drugs across the skin. Special liposomes characterized by an high bilayer elasticity have been developed, namely, ethosomes and transfersomes. Ethosomal systems are different from transfersomes by their structure and mechanism of action. As an example of different behavior, occlusion has no effect on skin permeation of molecules from ethosomes, while transfersomes are unable to enhance drug delivery under the occluded conditions. Ethosomal systems contain vesicles with fluid bilayers (soft vesicles) in a hydroethanolic milieu. Both components have a crucial role in the delivery of the active agent (301,302).

Liposomal colloidal carriers also can be applied as ophthalmic drug delivery devices to increase the bioavailability and the efficacy of drugs (303). Liposomes can enhance the ocular drug absorption and prolong the precorneal retention time (303), thus increasing drug effectiveness. In particular, the ocular application of positively charged small oligolamellar liposomes seems to be promising, considering that positively charged delivery devices may ensure a suitable bioadhesivity with the negatively charged corneal epithelium. As shown in Fig. 46, the acyclovir -loaded liposome showed a significant ($P < 0.005$) and noticeable increase of drug levels in the aqueous humor compared to the liposome–acyclovir physical mixture and the free drug (260). Several mechanisms can be proposed to elucidate the ocular effects of liposomes, but adsorption and/or lipid exchange seem to be most probably involved (303). Cornea permeability alteration due to liposomes may be discarded as a plausible explanation for enhanced drug penetration, since the presence of empty lipid vesicles added to drug solutions does not enhance the availability of the drug. Last, but not least, liposomes present a very good ocular tolerability showing no evidence of ocular inflammation or discomfort (260).

Niosomal Carrier. Niosomes are nonionic surfactant self-assembled vesicles that presents a structure similar to liposome (Fig. 47) and hence they can represent alternative vesicular systems with respect to liposomes, due to the niosome ability to encapsulate different type of drugs within their multienvironmental structure (304). The first application of niosomes was the cosmetic field followed by their use as drug delivery systems (305).

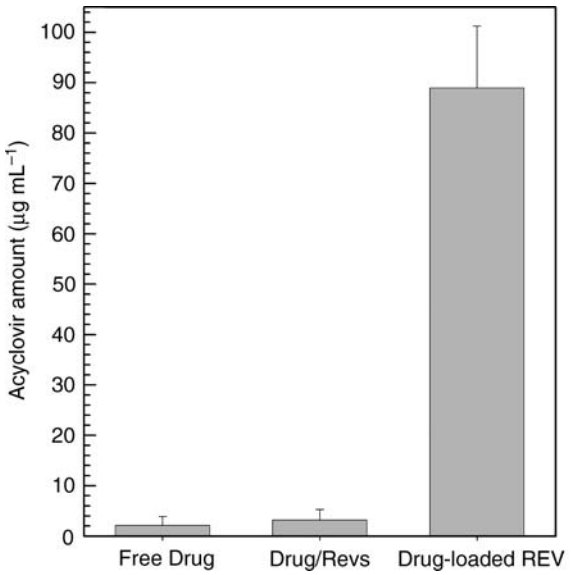


Figure 46. Aqueous humor concentrations of acyclovir at 30 min following topical instillation (50 μL) of acyclovir-loaded positively charged REVs (oligolamellar) liposomes (DPPC-Chol-DDAB 7:4:1 molar ratio), acyclovir–liposomes physical mixture and aqueous solution. Each bar represents mean values ± S.D. of four experiments. Data from Ref. 260.

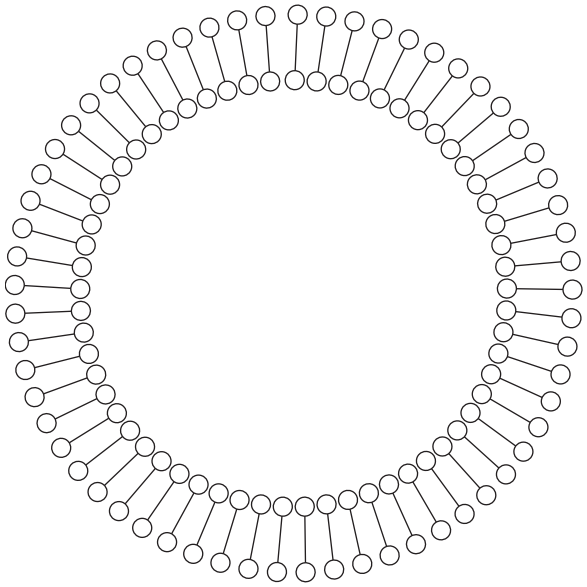


Figure 47. Schematic representation of a niosome structure. ○, hydrophilic head group; —, hydrophobic tail (305).

Components Used in Niosome Preparation. The main components of niosomes are nonionic surfactants. Different types of self-assembling nonionic surfactant were proposed as starting material to prepare niosomes (i.e., the SPAN and the Brij series). The type of surfactant can influence the stability of the vesicular system being able to influence the fluidity of bilayer structures. In particular, the nonionic surfactant can influence the leakiness of the entrapped drug from niosomes with the following increasing order, SPAN80 < SPAN20 < SPAN40 < SPAN60.

High niosomal concentration of soluble surfactant agents can influence the solubility of this vesicular colloidal carrier and determine the formation of micelles or complex aggregates. This phenomenon is observed with the presence of actylglucoside in the niosome formulation. This compound can destabilize the niosome bilayer and start a micellization phenomenon (305).

Another fundamental component for the preparation of niosomes is cholesterol. This molecule is used as an additive compound both to reduce the temperature of the vesicular gel to the liquid-crystal phase transition (305) and to decrease the overall HLB value of the surfactant mixture used for the preparation (306,307). Thus, cholesterol allows a more efficient aggregation of the nonionic surfactant component into a closed bilayer structure, and then a higher stability of the niosomal vesicles. The inclusion of cholesterol into niosomal formulation can reduce the leakiness of the membrane. A 1:1 molar ratio of cholesterol and nonionic surfactant is generally used for niosome preparation.

A parameter that should be taken into account in the choice of the niosome component is the physicochemical property of the encapsulated drug, due to a series of possible interactions occurring with the nonionic surfactant component leading to the formation of homogeneous dispersion or aggregate structure (305).

Methods of Niosome Preparation. Niosomes are prepared through the hydration of a mixture of nonionic surfactant–helper lipid (cholesterol) (1:1 molar ratio) at a temperature ranging from 40 to 70 °C followed by a suitable sizing process to obtain the required colloidal dispersion characteristics. The methods used to reduce the niosome mean size and to achieve an homogenous size distribution are similar to those used for liposomes, that is, extrusion through decreasing pore size polycarbonate filters, cycles of sonication, and high pressure homogenization (305,308). Similarly to liposomes, the mean size of niosome formulations is very important to reduce the RES uptake (305).

As concern the hydration of the nonionic surfactant–helper lipid mixture, some procedures reported for liposomes also can be used (e.g., the TLE method). In addition, other specific preparation methods have been developed for niosomes (305):

1. Injection of an organic solution (ether) of surfactant and cholesterol in a drug aqueous solution and heating of this mixture above the boiling point of the organic solvent;
2. Formation of an o/w emulsion between a drug aqueous solution and an organic solution of surfactant–cholesterol. Then, the organic phase is evaporated off and an aqueous niosomal colloidal dispersion is obtained;
3. Injection of the melted surfactant–cholesterol mixture in an aqueous heated solution of the drug under continuous stirring or vice versa injection of a

warmed aqueous drug solution into the niosomal component mixture.

The niosomal formulations obtained with the previous mentioned methods are generally micro size.

Considering the importance of the drug loading parameter, some procedures can be carried out to increase the amount of the ecapsulated drug within niosomes. There is evidence (305) that the DRV method, originally developed for the preparation of multilamellar liposomes with a high entrapment efficiency of water-soluble drugs (309), can also be used for niosomes with an increaes of their loading capacity from 3.3 to 64.4%. Another method successfully used to increase the amount of drug entrapped in niosomes is based on the formation of a pH gradient (305).

At the end of the preparation procedures, the excess of nonencapsulated drug is removed by dialysis, centrifugation, or filtration.

Toxicological Aspects of Niosomes. Considering that niosomes are made up of at least 50% synthetic nonionic surfactant, the toxicological profile of this carrier is very important for its application as a drug delivery system. Unfortunately, there are not many studies on niosome toxicity. An *in vitro* investigation, made on a model of ciliotoxicity to evaluate the influence of alkyl polyoxyethylene moiety of niosomes on the nasal mucosa, showed that increasing of the alkyl chain length of the nonionic surfactant determined a reduction of toxicity while the increase of the polyoxyethylene chain length pronounced the carrier ciliotoxicity. These findings seems to be correlated with the thermotropic state of niosomes, considering that the longer the alkyl chain the higher the transition temperature from gel-to-liquid phase, while the longer the polyoxyethylene chains the lower the transition temperature. This findings concluded that gel-state niosomes are less ciliotoxic than the liquid-state vesicles. On the contrary experiments on human keratinocytes showed on toxic activity related to both the alkyl chain length and the length of polyoxyethylene chain (310).

For the parenteral administration of niosomes, usually through the IV route, the evaluation of the vesicular system hemocompatibility is very important. The incubation of $C_{16}G_2$ and Span 60 niosomes with rat erythrocytes showed $<5\%$ hemolysis after 5 h. This level of hemolysis is not significant, considering that $<2\%$ of an injected dose of C16G2 niosomes is still present in the blood stream 5 h after dosing (305).

In the case of niosomal soluble surfactant components, a dose-dependent effect was observed. When low concentrations are used, the soluble surfactant is totally incorporated in the niosome structure and a drastic reduction of its intrinsic toxicity is achieved. The situation changes when the amount of soluble surfactants (e.g., Solulan C_{24}) is increased, because the formation of micelles occurs, and then the free monomers and/or micelles may exert their toxic action on cultured cells (311). Therefore, the whole niosomal carrier should be investigated for potential toxicity rather than the single components.

The issue of niosome toxicity is quite complex due to the fact that the presence of a drug can change the toxicological profiles of the unloaded carrier. For example, the inclusion of doxorubicine in $C_{16}G_2$ niosomes produce a severe dose-dependent inflammatory effect at the level of the lung within 24 h following intraperitoneal administration (305). After intraperitoneal administration of empty $C_{16}G_2$ niosomes or the free drug, such an effect on lungs is not observed. A possible explanation is the fact that doxorubicin-loaded niosomes are transported away from the peritoneum by the lymphatics via the thoracic duct allowing a higher dose in the main veins emptying into the heart. This hypothesis can be supported by the fact that 56% of a methotrexate-loaded niosome formulation is found in the thoracic lymph following intraperitoneal administration with respect to 12% observed for a free drug solution (312).

The modulation of drug toxicological effect is an important aim of the niosomal carrier. The encapsulation of vincristine in noisomes can reduce the free drug toxicological profile and improve the drug antitumoral activity in S-180 sarcome and Erlich ascites mouse models (313).

Niosomes in Complex Systems. The need for a more precise controlled drug release prompted the research of new and more sophisticated delivery systems. For this reason, niosomes based on Span surfactants were used to prepare a v/w/o (vesicle in water in oil system) niosomal formulation (314). The release rate of carboxyfluorescein, a hydrophilic fluorescet probe, showed the following increasing trend: v/w/o $<$ w/o emulsions $<$ niosome dispersion. Also, the nature of surfactant can influence the release of the fluorescent probe according to the following decreasin order: Span 20 $>$ Span 40 $>$ Span 60. The presence of Span 80 in the v/w/o system can drastically increase the probe release from the system due to its unsaturation in the alkyl chain, which generate a more leaky bilayer structure. While, the crystallization of Span 60 in the oil phase elicit the formation of an oil gel phase that can noticeable reduce the release rate from this vesicular system (314). A temperature-dependent release can be obtained in Span 60 v/w/o by adding Span 20 as a stabilizer, thus providing a faster probe release at 37 °C (305).

Niosome colloidal dispersion can be easily viscosized by the addition of hydrocolloids.

The addition of Solulan C24 in C16G2 niosomes determined the formation of the discome phase, that is a large vesicle ($\sim 60\,\mu m$) able to encapsulate hydrophilic compounds. These giant vesicles were found to be of two types: large vesicles that appear ellipsoid in shape and large vesicles that are truly discoid (305). The features of the discome structure prompt the used of this particular niosomal system as an ophthalmic drug delivery.

Therapeutic Applications of Niosomes. Niosomes can be used as a fine drug delivery systems being able to confer a certain selectivity to the entrapped drug as a function of their composition and physicochemical properties. After IV administration, niosomes show a high liver tropism (304,305). However, a niosomal formulation containing doxorubicin, composed of palmitoyl muramic acid, cholesterol, Solulan C_{24}, can escape from the liver uptake (305).

At the same time, a iopromide-loaded niosomal formulation extruded through a 220 nm filter and with the presence of stearylamine in its composition is able to accumulate in the kidneys (315). These findings showed that the presence of a positive charge on the surface of the niosomes can improve the targeting to the kidneys. The intraperitonel administration of niosomes with Span 80 in their formulation (312) can produce a lymphatics targeting, while $C_{16}G_2$ niosomes (305) after intraperitoneal administration can act as a depot system.

The presence in niosomal formulations of surfactant characterized by ester bonds can support the enzymathic degradation by esterases present in plasma, thus influencing the biodegradability, the residence time, and the stability of the system in the plasma. Moreover, the nature of the entrapped drug can influence the structure of the niosomal surface and the biodistribution of the system.

The first application of niosomes was as antiparasitic vesicular system for the treatment of leishmaniasis. The administration of a niosamal formulation containing stibogluconate was very useful to reduce the parasite disease because niosomes acted as a drug depot in the liver. In this case, the antiparasitic activity of niosomes regarding to the liver leishmania donovani can be correlated to the rapid uptake of the formulation in the liver after IV administration. However, this formulation cannot eradicate the parasite in the spleen and bone marrow. For this reason, different types of polyoxyethylene niosomes ($C_{16}EO_2$, $C_{16}EO_4$, $C_{16}EO_6$) are used to suppress the parasite in the spleen and bone marrow (305).

The IV administration of 100 nm of $C_{16}G_3$ niosomes containing methotrexate can improve the hepatic levels of the drug with serum levels of the drug higher than when it is administered in solution (316). In particular, a 23-fold increase in the area under the curve of metotrexate plasma level as a function of time is observed after IV administration of niosomes (4.5 μm mean size) containing Span 60 to tumor bearing mice (317), this finding is probably due to the great size of this vesicular system. Span 60 niosomes can further increase the plasma level of methotrexate if they are administered following the macrophages activation with mramyl dipeptide-gelatin derivatives (317). The oral and IV administration of $C_{16}G_3$ niosomal formulation encapsulating methotrexate can cross the blood–brain barrier and provide a substained release of this drug at the level of the CNS (316). However, the delivery of drug to the brain with niosomes has not been successful.

The administration of doxorubicin-loaded $C_{16}G_3$ niosomes (850 nm mean size) in tumor bearing mice determined a high drug level in the tumor site, serum, and lung, but not in the liver (305,318). While, doxorubicin-loaded 240 nm niosomes made up of Span 60 increased plasma, liver, and tumor levels. The reduction of proliferation of the S-180 sarcoma in NMRI mice after IV administration of niosomal formulation containing doxorubicin demonstrated an increased drug anticancer activity after encapsulation in niosomes (Fig. 48). At the same time the side effects, in particular cardiotoxic activity, are reduced following entrapment in niosomal formulations (314). Niosomes can improve the antitumoral effect of vincristine in S-180 sarcoma well as other anticancer drugs (313).

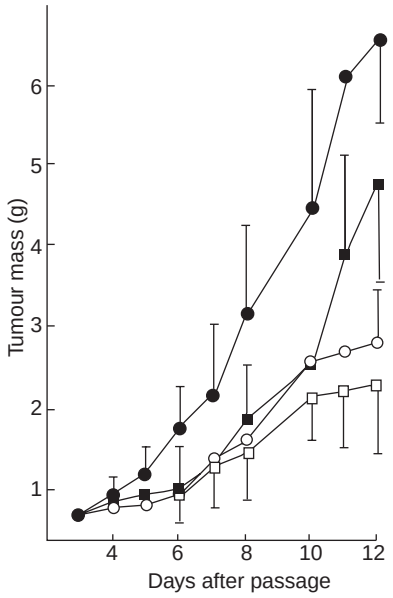


Figure 48. The growth in the mass of implanted tumor as a function of time after IV injection of (●) phosphate buffered saline pH 7.4, (■) doxorubicin solution ($5\,mg \cdot kg^{-1}$), (○) doxorubicin ($5\,mg \cdot kg^{-1}$) $C_{16}G_3$ niosomes, (□) doxorubicin ($5\,mg \cdot kg^{-1}$) $C_{16}G_3$:cholesterol (50:50) (305).

A diclofenac-loaded niosomal formulation composed by Span 60, cholesterol and DCP (22:73:5) produces a noticeable reduction of inflammatory processes in rat more efficaciously than the free drug. The improved activity of the drug can be determined by an increase in the area under the plasma time curve. Similar findings were obtained for niosomes-containing flurbiprofen, which showed an improved drug effect and bioavailability with a reduction of side effects produced by the free drug (305).

Niosomes can be used as agents for diagnostic imaging. Iopromide radioparque agent encapsulated in niosomes made up of $C_{16}G_3$, cholesterol, and stearylamine, can be concentrated in the kidney after IV administration (315). As mentioned above, the kidney targeting action is mediated by the positive charge on niosome surface.

Niosomes also can be effectively used for the oral delivery of drugs. The first application in this field was carried out with methrotrexate-loaded $C_{16}G_3$ niosomes characterized by a mean size of 100 nm (316). This investigation showed higher levels of methrotrexate in serum, liver, and brain after oral delivery using the niosomal formulation with respect to the free drug. A certain interest is focused on the possibility of using niosomes as carrier for the oral delivery of peptides and proteins. For example, ovoalbumine-loaded niosomes are able to increase the production of specific antibodies after oral administration (305).

Other successful applications of niosomes as delivery systems concern the topical administration of drugs and particularly the transdermal and ophthalmic delivery of drugs.

Niosomal formulations can increase the amount of drug permeated through the stratum corneum (319), even if the

exact mechanism involved in the drug and/or carrier passage has to be investigated and elucidated in a more detailed way. A hypothetical mechanism of skin penetration is related to a possible reorganization of the niosomal membrane at the level of the stratum corneum (320). *In vitro* data showed an efficacious transdermal delivery of oestradiol when it is entrapped in $C_{18}EO_7$ and $C_{12}EO_7$ niosomes. The improved drug passage through the outer skin layer seems to be mediated by the high flexibility of the bilayer structure of some niosomal formulatons (319). Similarly, a niosomal formulation made-up of glyceryl dilaurates ($C_{16}EO_7$) and cholesterol can increase the passage through the stratum corneum and the penetration of cyclosporine A into the inner layer of the skin (305). Then, niosome can be used as a transdermal drug delivery system for both hydrophobic and hydrophilic drugs.

Niosomes were proposed as a potential ophthalmic drug delivery system. Cyclopentolate-loaded niosomes made-up of Span 20 and cholesterol can pass through the cornea in a pH dependant manner, that is, pH value 5.5 is optimal for the cyclopentolate penetration, while at pH 7.4 a decreased permeation was observed. However, the *in vivo* mydriatic response is irrespective of the pH of the niosomal formulation. The explanation of the increased corneal adsorption of cyclopentolate may be due to a niosome-induced modification of the permeability characteristics of the conjunctival and scleral membranes (321).

Similar to liposomes, niosomes can be used as a vaccine adjuvant. A niosomal formulation composed by 1-monopalmitoyl glycerol, cholesterol, diacetyl phosphate can be used to encapsulate antigenic compounds and this result is fundamental for the adjuvanticity (305). A v/w/o niosomal system containing Span 80 and cotton seed oil was evaluated as an immunological adjuvant using the antigen tetanus toxoid (314). An increased secondary response (level of IgG1) was observed when the v/w/o formulation was administered by the intramuscular route in comparison with the vesicle formulation and the free antigen.

Ethosomal Carrier. Ethosomes have been invented by Touitou (322–324). The low toxicity and the property of ethanol as a permeation enhancer (325) as well as the possibility to include ethanol in the liposomal formulation, has brought to the realization of a new vesicular system for transdermal delivery: ethosome (301).

Ethosomes presents interesting features correlated with its ability to permeate intact through the human skin due to its high deformability. In fact, ethosomes are soft, malleable vesicles tailored for enhanced delivery of active agents. It has been shown that the physicochemical characteristics of ethosomes allow this vesicular carrier to transport active substances more efficaciously through the stratum corneum into the deeper layers of the skin than conventional liposomes (326). This aspect is of great importance for the design of carriers to be applied topically both for topical and systemic drug administration. Furthermore, the ethosomal carrier is also able to provide an effective intracellular delivery of both hydrophilic and lipophilic molecules (327) and also the penetration of an antibiotic peptide (i.e., bacitracin) within fibroblast cells was facilitated (328).

Formulative Aspects of Ethosomes. Ethosomes are a vesicular system made up of a phospholpid component, ethanol, and water. Phospholpid is the lipid component that confers the shape of vesicle to the delivery system. Ethanol is an important component in ethosome due to its destabilizing action regarding the packed-ordered structure of conventional liposomes (326), thus conferring the characteristic elasticity and deformability to this vesicular carrier. There are a number of methods that can be used to prepare stable ethosomal formulations depending on drug and the target of drug delivery (322–324). Among these, a frequently used method to prepare ethosomes is based on the dissolution of phospholipids in ethanol (20–50% w/v). Then, an aqueous solution is added to the lipidic solution under stirring thus allowing the formation of ethosomes (327–329).

The ethanol/phospholipid ratio used for the preparation of ethosomes is a crucial factor influencing the mean size and size distribution of ethosomes (Fig. 49). Usually, ethosomes prepared with a great amount of ethanol ($\geq$ 40% v/v) show a narrow vesicle size distribution. The size of ethosomes decreases with increasing ethanol concentration, while the concentration of phospholpid influenced the vesicle mean size in a different way, namely, the higher the phospholpid concentration the larger the ethosome mean size (301,330). The amount of ethanol used in the formulation can modify the superficial charge of ethosomes and the skin interaction (301). Normally, the presence of drugs have no significant influence on both mean size and size distribution. Ethosome composition can also influence the lamellarity as shown by electron transmission microscopy (Fig. 50), since the formation of either unilamellar or multilamellar ethosomes is a multifactor process.

Ethosomes can entrap hydrophobic and hydrophilic molecules in their structure. With respect to liposomes, where hydrophilic drugs are entrapped in the aqueous

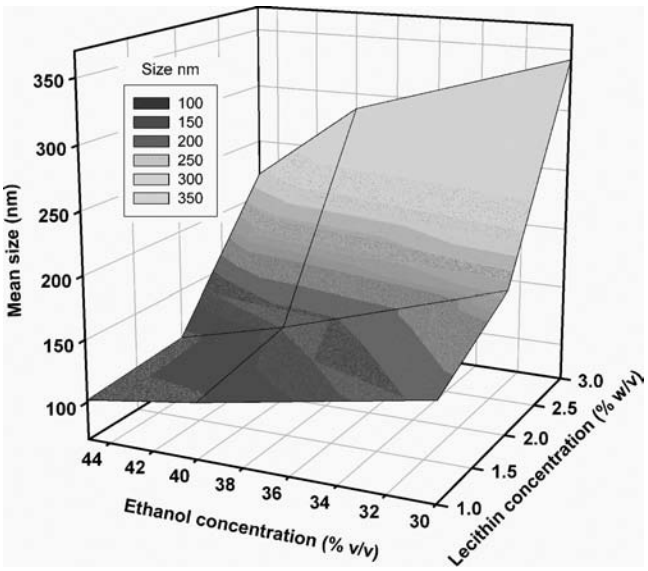


Figure 49. Influence of the amount of ethanol and lecithin used for the preparation of ethosomes on vesicle suspension mean size and colloidal polydispersity index. Data from Ref. 330.

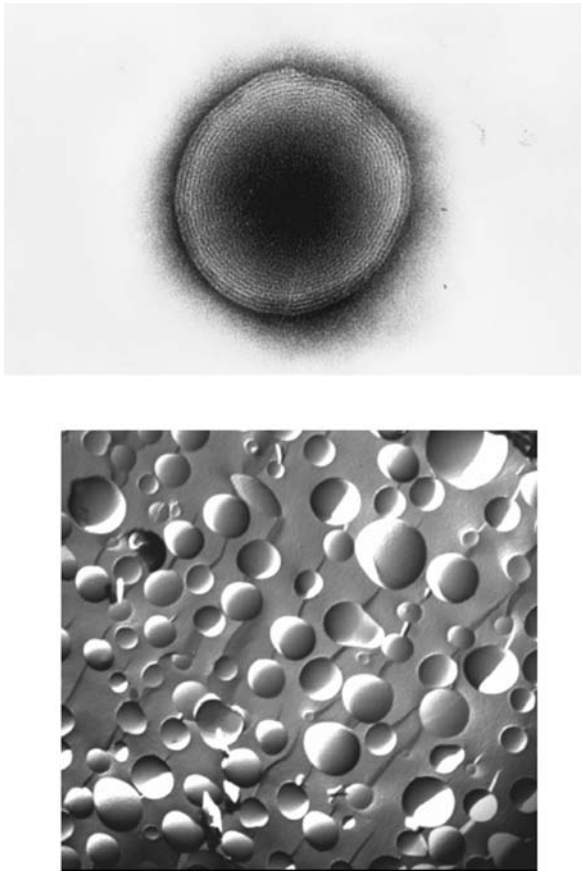


Figure 50. Transmission electron microscopy of ethosomal vesicles composed of 2% lecithin and 30% ethanol (a) (301). Freeze-fracture electron micrographs of ethosomes composed of 45% ethanol and 2% lecithin (b) (330).

compartment and hydrophobic drugs are in the lipid bilayer core, in ethosomal formulations drugs are homogenously present in ethosome structures in spite of drug physicochemical properties (301,327) (Fig. 51). This finding can be explained by the multilamellarity of the ethosomal vesicles as well as by the presence of ethanol in the ethosome, which allows for better solubility of the lipophilic and amphiphilic probes (301). The ethosome composition can also influence the drug entrapment efficiency, that is, the amounts of ethanol and phospholpid used for ethosome preparation positively influence the loading capacity of the colloidal carrier. Namely, the higher the amount of ethanol and phospholpid the greater the drug entrapment within ethosomes (301,326,330), the values of drug entrapment efficiency are often higher than those expected for a conventional vesicle formulations (330). This fact can be explained by the presence of ethanol, which increases the drug solubility in the polar phase of ethosomes (301).

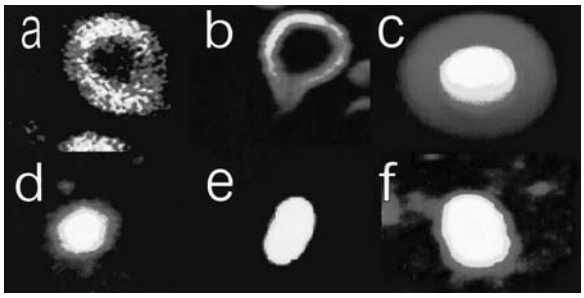


Figure 51. Entrapment of fluorescent probes by phospholipid vesicles determined by confocal scanning laser microscopy. Liposomes (a–c) or ethosomes (prepared with 2% lecithin and 30% ethanol) which (d–f) were prepared with one of three following fluorescent probes: rhodamine red, a highly lipophilic molecule (a,d); D-289, an amphiphilic molecule, (b,e); calceine, a hydrophilic molecule (c,f). White represents the highest concentration of a probe, followed by yellow, with red being the lowest probe concentration (301).

Therapeutic Potentialities of Ethosomes. The enhanced percutaneous permeation capability of ethosomes is due to the unique feature of this carrier that is able to interact with the stratum corneum and to elicit a reversible disorganization of the stratum corneum lipid packing order, thus increasing the skin permeability to drugs and vesicles (329).

An important characteristic to be evaluated before the proposal of a drug carrier as a potential topical drug delivery system is its *in vivo* skin tolerability on human subjects. *In vivo* reflectance spectrophotometry data (330) on volunteers showed that ethosomes elicit no induction of skin erythema, while a hydroethanolic solution with an equal water/ethanol ratio of ethosomes induces a remarkable skin erythema (Fig. 52). These results demonstrate that ethanol present in the ethosomal formulation is not

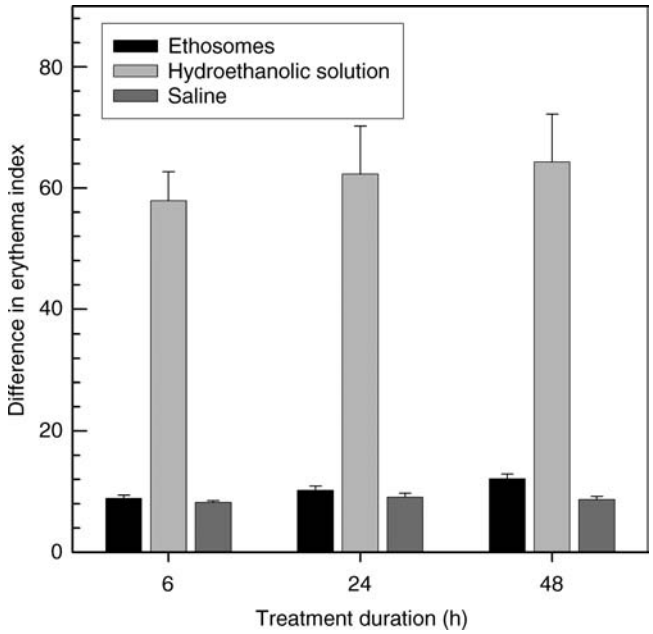


Figure 52. *In vivo* human skin tolerability of various topical formulations after 6, 24, or 48 h of treatment. Results are expressed as a mean value of the variation of the erythema index ($n = 6$) ± standard deviation. Legend keys: ethosomes, formulation containing 2% (w/v) Phospholipon 90 and 45% (v/v) ethanol; hydroethanolic solution, solution of water, and ethanol at a volume ratio of 55:45; saline, control saline (0.9% w/v NaCl in water) solution (330).

able to act as a skin erythema-inducing agent, even though it is present at a high concentration.

A wide range of drugs have been formulated in ethosomal carriers and tested *in vitro*, *in vivo*, and in clinical studies. These molecules comprise steroid hormones, antivirals, antibiotics, vitamins, peptides, and cosmeceutical agents. Moreover, ethosomes are very efficient carriers for targeting molecules to the pilosebaceous units and could be used for acne and alopecia treatment. Carrier consists of materials approved for pharmaceutical and cosmetic use (316,327–332).

An interesting example of the ethosome potential application as innovative topical carriers is represented by the transdermal delivery of cannabidiol (333), a new drug candidate for treatment of rheumatic diseases, that presents a number of drawbacks when administered orally. The ethosomal formulation is able to prevent the inflammation and edema induced by subplantar injection of carrageenan in ICR mice.

Often, the skin permeation enhancement observed for all ethosome-based formulations is much greater than that can be expected from ethanol alone. This behavior can be due to a synergistic mechanism between ethanol, phospholipid vesicles, and skin lipids (301).

Two different research groups reported an *in vivo* sustained release effect of ethosomes with a prolongation of the drug therapeutic activity, which can be related to an accumulation in the skin (330,333).

Then, recent findings on ethosomes are very encouraging and confirm that this carrier is very promising for the topical administration due to the enhanced delivery of drugs through the skin, thus prompting various opportunities for the development of suitable therapeutic strategies through the topical route.

Ultradeformable Vesicular Carrier. It is believed that liposomes, when administered on the skin, first disintegrate their structure, and then diffuse through the barrier in the form of small fragments or lipid monomers (334). Conventional rigid liposomes were shown to be unsuitable vesicular carriers to cross the skin barrier (335). Highly deformable vesicles were developed, the so-called transfersomes or ultradeformable liposomes were invented by Cevc (336). It was shown that the high deformability of vesicular carriers could allow them to penetrate intact skin if applied nonocclusively *in vivo* (336,337), thus favoring the delivery of encapsulated drugs across the skin (338,339). Ultradeformable vesicles seem to cross the skin without irreversible disruption, probably because they are much more elastic, and hence more deformable respect to classic liposomes. For the preparation of ultradeformable vesicles, the so-called edge activators were incorporated into the phospholipid bilayers at suitable amounts, namely, bile salts were often used for this purpose (340).

Formulative Aspects of Ultradeformable Liposomes. Lecithins and a bile salt at different molar ratios are the main components of ultradeformable liposomes that can be prepared with the TLE method or any other used for liposome preparation (the section main methods to prepare liposomes) (336–340). Small amounts of ethanol ($\leq 7\%$ v/v) are normally used for the preparation of ultradeformable liposomes.

Similarly to ethosomes, the ratio between the various components is a crucial factor for the determination of the physicochemical and drug-loading capacity properties of ultradeformable liposomes.

The DSC studies have demonstrated that the amount of the edge activator is related to the increased fluidity of the vesicular bilayers up to a certain values, beyond this value the formation of mixed micelles and other kinds of colloidal aggregates were observed (341). For this reason, large amounts of edge activator beyond a certain value hinder the transdermal drug delivery; in fact, mixed micelles and aggregates are much less effective transdermal carriers than ultradeformable liposomes. The formation of a coexistence region characterized by various phospholipid/bile salt aggregates (i.e., mixed vesicles, opened vesicles, mixed micelles, and rod-like mixed micelles) is evidenced by a reduction of the mean size and a concomitant increase of the polydispersity index values, thus showing the presence of a wide size distribution (341). The presence of bile salts in the composition of ultradeformable liposomes leads to a negative zeta-potential due to the increase of negative charge (carboxylate group of bile salts) along the surface of vesicle bilayers.

The ratios between the components and the type of edge activator used to prepare ultradeformable liposomes can influence the amount of drug entrapped within this carrier. When nonionic surfactants (i.e. Span and Tween) are used instead of bile salts, a decrease of the drug entrapment efficiency of the carrier is observed. In any case, a high concentration of edge activators cause a drastic reduction of the drug loading capacity due to the presence of other forms of aggregation than vesicles. The above mentioned aggregates have a poor drug loading capacity. The higher hydrophilic form of ultradeformable liposomes than conventional liposomes and their high flexibility avoids the aggregation and fusion of the transfersomal system providing a stable vesicular structure (342).

Similarly to other vesicular carriers, ultradeformable liposomes can entrap different types of molecules (i.e., lipophilic, hydrophilic, and amphipatic drugs). The release rate of entrapped drug from ultradeformable liposomes is mainly influenced by the carrier composition and the drug physicochemical properties. Generally, the release of water soluble drugs from ultradeformable liposomes is modulated by the concentration gradient between the inner and the outer compartment. The water gradient through the skin can trigger the release of the entrapped drug. The release of hydrophobic drug is slower than that of the hydrophilic drug, and it is confined to the contact and lipid exchange between ultradeformable liposomes and biological membranes (342). The slower release of the hydrophobic drugs is due to a strong interaction between drugs and lipid bilayers (343). Amphipatic drugs have intermediate release characteristic between hydrophobic and hydrophilic drugs.

Therapeutic Potentialities of Ultradeformable Liposomes. Ultradeformable liposomes are characterized by a deformable structure that can pass intact through the skin using a

water active gradient, thus favoring the drug delivery through the skin without modifying the integrity of the cutaneous barrier (342,343).

In vitro and *in vivo* tests with ultradeformable liposomes showed that this vesicular system does not produce any toxic effects after topical application and it is well tolerated by the skin tissue (342).

Skin is a nanoporosus barrier that permits only the passage of small (nanometer size molecule) compounds (334,342). The nonocclusive topical application of ultradeformable liposomes undergoes water evaporation from the formulation and the consequent dying out of the vesicles (343). The elastic and hydrophilic properties of ultradeformable liposomes determine the movement of the vesicle through the skin pores by following the transdermal water gradient. The topical application of ultradeformable liposomes can increase the size of skin nanopore (78,344). After ultradeformable liposome percutaneous permeation, these vesicles can distribute in cells and after the bypassing of cutaneous capillary they can reach the subcutaneous tissue.

The most important example of the application of ultradeformable liposomes as transdermal drug delivery is represented by Transfenac, formulation of diclofenac in ultradeformable liposomes (338). Transfenac mediates the agent transport through intact skin and into the target tissues. Therapeutically meaningful drug concentrations in the target tissue are reached even when the administered drug dose in Transfenac is $< 0.5\,\mathrm{mg \cdot kg^{-1}}$ body weight. Diclofenac association with ultradeformable carriers permits it to have a longer effect and to reach 10 times higher concentrations in the tissues under the skin in comparison with the drug from a commercial hydrogel. The relative advantage of diclofenac delivery by means of ultradeformable liposomes increases with the treated muscle thickness and with decreasing drug dose, as seen in mice, rats, and pigs (338); this can be explained by assuming that the drug associated with carriers is cleared less efficiently by the dermal capillary plexus.

Transfenac, hence promises to be a useful formulation for the treatment of diseases of superficial tissues, such as muscles or joints, having the potentiality to replace combined oral–topical diclofenac administration in humans.

Particle Drug Carriers

Micro- and nanoparticles are solid colloidal suspensions in which the mean particle size is > 1 or $< 1\,\mu\mathrm{m}$, respectively. Under the morphological point of view, two different types of particles can be distinguished: capsules and spheres (Fig. 53). Sphere systems are usually characterized by a porous matrix in which drugs are contained, while capsule systems are formed by a core containing drugs surrounded by a shell.

Polymeric Particles. These colloidal carriers are prepared from natural or synthetic polymers and, in dependence of the preparation method and of the polymer used, micro- or nanocapsules and micro- or nanospheres can be distinguished. Polymeric particles have become very important because of their ability to deliver a variety of drugs to different areas of the body for sustained periods of time (345–347). Up to the end of 1980s, microparticles were extensively investigated. Now great interests are focused on the much smaller carriers (e.g., nanoparticle) that are able to ensure fine drug delivery opportunities both in terms of efficacy and selectivity. For this reason, this section will be mainly focused on polymeric nanosystems.

Figure 53. Schematic representation of spheres and capsules as potential drug delivery devices.

Concerning the materials used for the preparation of these colloidal carriers, natural polymers (i.e., proteins, polysaccharides, waxes) are not widely used because they present a huge variability in their purity and defined physicochemical properties. Furthermore, they often require a cross-linking procedure that may cause an alteration of the encapsulated drug. For these reason, mainly synthetic polymers have received attention and have been largely investigated for potential use in drug delivery devices.

The use of a large series of polymers is restricted and limited by their bioacceptability, which is also influenced by colloidal particle mean size. In fact, the diameter of the smallest blood capillaries is $\sim 4\,\mu\mathrm{m}$, thus nanoparticles should have a smaller diameter than this to traverse all capillaries. As a general consideration, for the suitable choice of the appropriate macromolecular polymer to be used as a nanoparticle matrix, the colloidal particle size and the preparation method will first depend on the biocompatibility of the polymer, second on the physicochemical properties of the drug, as well as on the therapeutic goal to be reached. In this colloidal carrier, drugs can be adsorbed, attached, dissolved, entrapped, and/or encapsulated (348). Micro- and nanoparticles can be used to deliver both hydrophobic and hydrophilic molecules, proteins, vaccines, biological macromolecules. They can also be formulated for targeted delivery to all organs or made for long-term systemic circulation (235). Thus, a lot of synthesis procedures exists.

Preparation Methods and Formulative Aspects. Drugs can be incorporated into nanoparticles in a number of ways: (*1*) drug can be entrapped in the polymeric matrix; (*2*) it can be encapsulated in a nanoparticle core; (*3*) it can be chemically conjugated to the polymer; (*4*) it can be surrounded by a shell-like polymer membrane; (*5*) it can be adsorbed on particle surface.

Polymeric nanoparticles can be prepared using a lot of different techniques. One of the most used preparation methods is the emulsification–solvent evaporation technique. The polymer and the drug are solubilized in an organic

solvent and an emulsion is prepared by adding water and a surfactant. Liquid nanodroplets are produced by sonication or homogenization, and then the organic solvent is evaporated in order to achieve the nanoprecipitation of the polymeric material in solid nanoparticles (345,349). Obviously, this procedures can be used only for hydrophobic drugs. To allow the encapsulation of hydrophilic molecules, a modification of this procedure led to the multiple emulsion technique (350).

Another method is the phase-inversion nanoencapsulation (PIN), which has been used to encapsulate insulin for oral administration (351). A limitation of these two techniques is the use of toxic and fluorinated solvents, which may cause drug degradation. For these reasons, other techniques, that do not compromise drug stability have been developed.

One of these is the emulsification–diffusion method. In this case, the polymer and the active compound are dissolved in a partially water-soluble solvent. This organic solution is added, and then emulsified in an aqueous phase containing a surfactant. To favor the precipitation of nanoparticles, additional water is added to the emulsion under stirring. At the end of the process, the solvent can be removed by centrifugation or dialysis.

The nanoprecipitation method involves the dissolution of the polymer and the drug in a freely water-miscible organic solvent (e.g., acetone) and then the addition of this organic solution into a water phase containing a nonionic surfactant (e.g., Pluronic F68). The organic solvent is then removed under reduced pressure by a rotavapor (352). This procedure leads to the formation of nanospheres, but if a biocompatible oily component is added in the organic solution nanocapsules are formed (353).

The formation of nanospheres or nanocapsules also can be achieved by an *in situ* polymerization process. The emulsion or micellar polymerization is the most used approach to achieve nanocapsules and nanospheres by starting from the polymeric monomer, respectively. In the case of micellar polymerization, reactions take place in the solvent phase. The following polymers can be prepared as nanosphere colloidal suspensions following this preparation procedure: PMMA, poly(ackyl cyanoacrylate), and acrylic copolymer. When the polymerization is carried out at the interface (interfacial polymerization) between an oil phase and an aqueous solution, nanocapsules are formed (354).

The determination of the loading capacity of nanoparticle colloidal suspensions can be carried out by separation of the untrapped material with ultracentrifugation followed by the drug analysis after dissolution of the pelleted polymeric matrix. Other reliable separation methods are ultrafiltration and gel permeation chromatography (345,346). The drug loading capacity also can be calculated by determining the drug content in the supernatant or in the filtrate. In fact, the amount of drug entrapped in nanoparticle colloidal systems can be obtained by subtraction of the untrapped drug amount from the total amount of drug present in the suspension.

The mechanisms of drug release from nanoparticle colloidal suspensions depends on the characteristics of the colloidal suspension, as well as on physicochemical properties of the drug. In particular, the release of a drug may occur by one of the following mechanisms or a cooperation of more than one of them: (*1*) drug desorption from the colloidal surface (both for nanospheres and nanocapsules); (*2*) drug diffusion through the polymeric network of the nanospheres; (*3*) drug diffusion through the polymeric shell of nanocapsules; (*4*) polymeric matrix erosion of nanoparticles. The drug release rate is dependent on the release mechanism, the diffusion coefficient, and polymer biodegradation rate. The nanoparticle drug release is also greatly influenced by the type of interaction with the biological substrate (345).

Besides the previous mentioned drug release mechanisms, it should be considered that the drug delivery function of nanoparticles also can be accomplished by a direct contact with the biological membranes, thus leading to an enhanced drug delivery through membranes with respect to a simple drug solution (355). As a consequence of this behavior, it may happen that the *in vitro* drug release profiles are poorly related to the *in vivo* drug delivery and release situation (356).

Size and zeta potential are important physicochemical parameters to be determined to achieve a suitable colloidal carrier. Nanoparticle size is influenced by the preparation technique and by the polymer used, that is, low molecular weight polymers form small-sized nanoparticles, but this fact reduces the amount of encapsulated drug. An increase of polymer concentration usually elicits an increase of both nanoparticle size and encapsulation efficiency (357,358).

The zeta potential is a measure of the surface electrical charge of the particles. As the zeta potential increases, the repulsion phenomenon between particles will be greater, thus leading to a more stable colloidal dispersion. The minimum zeta potential value to prevent particle aggregation and to have a stable nanosuspension was defined to be ± 30 mV (359).

Therapeutic Applications of Nanoparticles. The oral administration is one of the promising application of nanoparticles that have been administered either for achieving a systemic uptake or for having a local residence within the GI tract. The polymers used for peroral application are nondegradable polymers (cellulose, acrylate derivatives, etc.) and are designed not to be adsorbed (359).

Polymer nanoparticles for oral treatment may be formulated as an aqueous suspension or incorporated into traditional dosage forms. A lot of different nanoparticle formulations have been incorporated into tablets or capsules, and then compared with traditional dosage forms. In all cases, nanoparticles maintained the advantages of a colloidal carrier, such as an enhanced dissolution of lipophilic drugs and a prolonged and sustained release (360).

An innovative nanoparticle application in this field is oral chemotherapy, which can be a valid alternative, because it allows a continuous exposure of the cancer cells to anticancer drugs at a lower concentration, thus reducing or avoiding side effects. In addition, it is more convenient and better tolerated by the patients, especially for those with advanced metastatic cancers. Unfortunately, most anticancer drugs cannot be administered orally because of their poor solubility, stability, and permeability. It has

been found that anticancer drug encapsulation into an oral formulation of nanoparticles has been able to play a key role in drug adhesion and interaction with cancer cells. For example, PEG-coated nanoparticles are able to adhere to intestinal cells and subsequently to escape from the multidrug resistance pump proteins (361).

Chitosan-coated nanoparticles are used for colon targeted drug delivery of diclofenac. Chitosan nanoparticles are microencapsulated in Eudragit L-100 or S-100 to form a gastroresitent reservoir system in which the drug release is triggered only in the basic (pH 8) environment of colon (362).

For parenteral delivery, nanoparticles can be formulated as aqueous dispersions or they are converted in lyophilized powders to be resuspended just before their administrations (345,359).

Cancer therapy is one of the most important applications of polymeric nanoparticles. Nowadays, the aim of any anticancer research is to improve patient survival after chemo- or radiotherapy. Unfortunately, traditional anticancer therapy is affected by a lot of side effects that involve healthy cells leading to an unsuitable quality of life for cancer patients. So the effectiveness of a treatment is related to the ability to target the cancer cells while affecting as few healthy cells as possible. Nanoparticles can provide an alternative solution for the site-specific delivery of anticancer drugs due to their small size and the possibility to escape RES recognition and uptake, thus leading to a prolonged blood circulation time.

Biodegradable nanoparticles made of PLGA have been used to incorporate paclitaxel, a microtubule-stabilizing agent that causes cell death by promoting the polymerization of tubuline during cell division. Paclitaxel was encapsulated to a very large extent ($\sim$100% encapsulation efficiency) and this paclitaxel-loaded colloidal system showed a 70% loss of viability of human small-cell lung cancer cells at a drug concentration as low as 0.025 $\mu g \cdot mL$. Paclitaxel also has been incorporated in poly(ethylene oxide) modified poly(β-amino ester) nanoparticles to obtain a sustained release into most solid tumors (363). Also, Tamoxifen (364) and verteporfin (365) have been encapsulated into PLGA or poly (ε-caprolactone) particles for *in vivo* studies against breast cancer. Doxorubicin, a widely used anticancer drug, has been encapsulated into PLGA nanoparticles (366) that presented the ability to release the drug up to 1 month. In addition, this carrier system avoids a lot of the undesiderable effects of doxorubicin (e.g., cardiotoxicity).

If a sustained release of the drug in the tumor site is required, then the nanoparticle surface must be modified in order to avoid RES macrophages, which recognize hydrophobic particles as foreign. To escape RES, the surface of nanoparticles is modified with hydrophilic molecules that form a steric barrier on the particle surface. Polyoxypropylene–polyoxyethylene (POP/POE) surfactants are suitable macromolecules to prevent nanoparticles from sticking to the blood vessel endothelium and to inhibit RES recognition. Indeed, among the various copolymer members, poloxamine and poloxamer have the best prolonged circulation time of nanoparticles (367). Unfortunately, poloxamer and poloxamine do not exhibit prolonged circulation times when nanoparticles are made up of PLGA (368). Recently, the most common moiety used for nanoparticle surface coating to obtain the so-called Stealth nanoparticles is PEG and its derivatives (369,370). Attachment of PEG on the nanoparticle surface can be performed in different ways: (*1*) by adsorption, (*2*) by incorporation during the preparation process, (*3*) by covalent linkage with the nanoparticle polymeric matrix.

In the passive targeting, as with stealth liposomes, long circulating nanoparticles escape from the blood circulation through the fenestrations of capillaries perturbed by inflammatory processes or by tumors (235). Inflamed vessels present fenestration sized up to 700 nm, so improved colloidal nanoparticles (< 200 nm) are able to pass across, thus accumulating and releasing the drug just in the site of action. These particular characteristic of nanoparticles leads to an increased therapeutic index of the incorporated drug. Nowadays, chitosan-based particulate systems are attracting the most attention as potential long-term drug delivery systems because of mucoadhesive and long circulating properties of chitosan. Doxorubicin–dextran conjugates were encapsulated into chitosan nanoparticles to minimize the cardiotoxicity of the drug. This system reduced not only the drug side effects, but also improved the therapeutic efficacy of doxorubicin in the case of solid tumors (371).

The strategy of nanoparticle passive targeting is widely used for cancer therapy, but it presents a limitation due to a high resistance factor of some solid tumors that cannot be circumvented by PEG-coated nanoparticles. Therefore, an alternative approach is the use of temperature-sensitive nanoparticles, which are able to release its drug content only in hypertermic zones (235). Other approaches may involve the use of biochemical triggers, such as the pH-sensitive lipid-anchored copolymers, to generate fusogenic particles (249).

The strategy of active targeting increases the probability of having a selective direction of nanoparticles to a designed site. In this case, ligands that specifically bind to surface receptors of target sites are coupled to long-circulating particulates. Among polymers suitable for coupling to specific ligands, poloxamers and PEG have received the most attention. A derivatization is achieved between the end group of the poloxamer chains with pyridyl disulfide. After a disulfide exchange with a thiol-containing moiety on the peptide or antibody to be attached on the surface of the nanoparticle, long-circulating actively targeted particles are obtained (372).

The use of nanoparticles as drug delivery systems can overcome the barriers to the penetration of antiinfective drugs into cells, that is, strong protein binding, an unfavorable lipid–water distribution coefficient, an unfavorable pH gradient between different cellular compartments, and the existence of active transport pump mechanisms that prevent the accumulation of sufficient antibiotic concentrations in the interior of the infected cells (278). In particular, a suitable nanoparticle coating can promote the permeation of hydrophilic drugs through membranes (Fig. 54). In fact, the free drug, able to freely diffuse in the aqueous medium and to interact with the outer-hydrophilic zone of the membrane model, has to pass

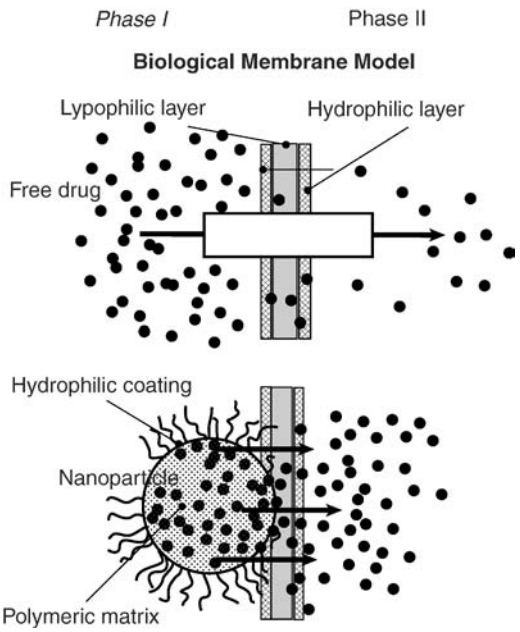


Figure 54. A model of the interaction between the aqueous phase containing a free hydrophilic antibiotic, or the drug-loaded PECA nanospheres and the biological membrane model. The permeation driving force is the different drug concentration between phases I and II (355).

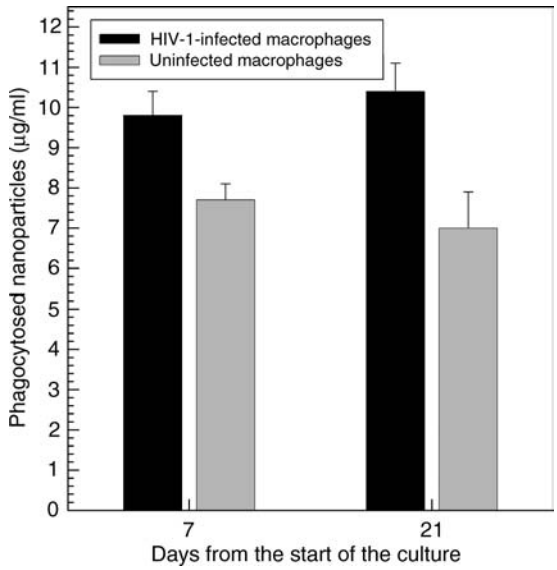


Figure 55. Influence of HIV infection on phagocytosis of poly(butylcyanoacrylate) nanoparticles by human macrophages. Cell cultures were infected with HIV-1 at day 1 after start of the culture. At day 7 or 21, the nanoparticles (200 nm diameter) were added at a final concentration of $0.5\,mg \cdot ml^{-1}$ to the infected cultures and incubated for 6 h. Data from Ref. 374.

through the lypophilic layer of the same membrane. This process could represent the limiting step in the diffusion through the membrane model of a hydrophilic molecule. In the case of nanoparticles, the outer hydrophilic shell of the particles (coated with nonionic surfactant) could ensure an interaction with the hydrophilic layer of the membrane, while the internal lypophilic core of the particle can ensure a close interaction with the hydrophobic layer of the membrane, providing a high permeation of the drug (354). The nanosphere-mediated increase of drug membrane permeation leads to an improvement of the antibacterial activity (373) and to an intrabacterial drug accumulation. The entrapment of antibiotics in nanospheres may prevent the bacterial resistance to drugs due to pleiotropic drug resistance and changes in the bacterial outer membrane leading to a decrease in OmpF porin, which probably causes decreased drug permeation.

Nanoparticles can be used for the treatment of various viral infection diseases localized at the level of the RES. For example, RES cells can be infected by both strains of the HIV, namely, HIV-1 and HIV-2. Monocytes and macrophages seem to have a fundamental role in the immunopathogenesis of the HIV infection, by behaving as virus reservoirs from which HIV can disseminate throughout the body and brain. Because nanoparticles can be easily phagocytosed by macrophages, they may represent a suitable and promising drug delivery system for the treatment of HIV infection persisting in these cells.

In vitro studies showed that human macrophages are able to phagocyte different types of polyacrylic and albumin nanoparticles (374). Interestingly, HIV-infected macrophages seem to have a higher phagocytotic activity concerning nanoparticle uptake than noninfected macrophages (Fig. 55). This phenomenon can be due to an activated state of these infected cells, leading potentially to a preferential phagocytosis of drug-loaded nanoparticles, and hence to a targeted delivery of antiviral drugs to these cells.

The RES tropism of conventional nanoparticles also can be efficiently used for the treatment of protozoa infection (i.e., leishmaniasis) (348). Contrary to liposomes, empty poly(isobutylcyanoacrylate) nanospheres exhibit a certain *in vitro* and *in vivo* antiparasitic activity against (375). Probably, this action could be attributed to peroxide production following nanosphere phagocytosis, which led to a respiratory burst, more pronounced in infected than uninfected macrophages.

One of the major problems connected with ophthalmic therapy is drug loss after instillation of eyedrops. To improve ocular bioavailability, mainly nanoparticles have been used. Colloidal systems have the convenience of a drop and are able to maintain drug concentration and activity at its site of action, probably due to an improved ocular mucoadhesion. In fact, poly(butylcyanoacrylate) nanospheres were able to improve the amikacin ocular delivery (376) by increasing the corneal and aqueous humor concentration of amikacin with respect to the free drug and to other formulations. The surface coating of nanospheres can also be used for ophthalmic application and can be a crucial factor to achieve a well-tailored and efficient drug delivery. In spite of the nanoparticle polymer and the method of coating, the presence of PEG on the surface of nanoparticles improves the ocular drug permanence time and increases the drug level in various ocular structures compared with both conventional ocular formulations and uncoated nanoparticles (352,377). Ocular gene therapy is also possible with polymeric nanoparticles (378).

In recent years, mucosal surfaces (nasal, pulmonary, buccal, and ocular) have received much attention as alternative routes of systemic administration (379). Chitosan-coated nanoparticles present mucoadhesive properties that can be useful to enhance mucosal drug adsorption. An example is the enhanced nasal absorption of insuline-loaded chitosan nanoparticles that do not damage the biological system.

The buccal adsorption of vaccine encapsulated into nanoparticles is an alternative to the parenteral route of administration of vaccines. The oral or nasal delivery of ovalbumine from chitosan microparticles enhances the systemic and local immune response against diphteria toxoid vaccine (380).

Lipid-Based Nanoparticles. Lipids instead of polymers can be used to obtain colloidal drug carriers. Solid lipid nanoparticles (SLN), nanostructured lipid carriers (NLC), and lipid drug conjugates (LDC) are nanoparticles with a solid lipid matrix and present an average diameter in the nanometer range. These innovative colloidal carriers have attracted increasing attention in recent years. They are regarded as an alternative carrier system to traditional colloidal systems (e.g., polymeric micro- and nanoparticles). General ingredients include solid lipids, emulsifiers, and water. Lipids include triglycerides, partial glycerides, fatty acids, steroids, and waxes. All excipients are generally recognized as safe (GRAS) substances, so a wide variety of compounds can be used for formulation purposes.

Solid Lipid Nanoparticles. Thse nanoparticles are colloidal systems made up of solid lipids and are stabilized by surfactants. During the 1950s, lipid nanoemulsions were introduced for the parenteral nutrition and later they were used as carriers for lipophilic drugs. The major problem with nanoemulsions was the loss of drugs related to their liquid form. The SLN were developed to overcome this problem. In fact, the use of solid lipids instead of liquid oils is a very attractive idea to achieve controlled drug release, because drug mobility is much slower into a solid lipid than in a liquid oil.

The SLN can be prepared using different procedures. The main preparation process is the high pressure homogenization (HPH) method, in which a dispersion of the drug in the melted lipid is constricted through a narrow gap (in the range of few microns) under very high pressure, thus disrupting lipid particles down to the submicron range. Other methods are the solvent emulsification–evaporation method and the microemulsion-based preparation (381). The first is a method to prepare nanoparticles by precipitation in o/w emulsions. The second production method is based on the preparation and the subsequent dilution in cold water (2–3 °C) of a microemulsion made up of a low melting lipid, an emulsifier, a coemulsifiers, and water. Following preparation, as for other colloidal carriers, the determination of the mean size, size distribution, and zeta potential of SLN is necessary to define their physicochemical properties.

Formulative Aspects of SLN. Evaluations on the degree of crystallinity and lipid modifications are very important for SLN. In fact, the organization of lipids into a crystalline reticulate is fundamental for the determination of the drug encapsulation and release rates. In general, if the lipid matrix is made-up of pure molecules (i.e., tristearin or tripalmitin), a perfect crystal with few imperfections, is formed. As incorporated drugs are located between fatty acid chains, in crystal imperfections, and between the lipid layers, a highly ordered crystal form cannot accommodate large amounts of drug (382). Also, the lipid modification influences the encapsulation and release degree of a drug. In fact, glycerides exist in three different polymorphic forms: α, β', and β. The degree of reticulate imperfection of the lipid decreases in this sequence: $\alpha < \beta' < \beta$. Lipid nanoparticles recrystallize at least partially in the α-form, but with increasing formation of the more stable β'/β modifications, the lattice is becoming perfected, the number of imperfections decreases, and drug is expelled from nanoparticles.

The coexistence of additional colloidal structures, such as micelles, liposomes, and supercooled melts, has to be taken into account after the preparation of SLN; the quantification of these additional structures is a serious challenge because of their size similarities with those of the SLN. Therefore, it would be desirable to use methods that are sensitive to the simultaneous presence of different colloidal species. Both NMR and ESR techniques meet these requirements.

Physical stability of SLN dispersions is generally >1 year (383), up to 3 years in the case of SLN made of glyceryl palmitate. The storage stability of SLN depends on two factors: (*1*) the physical modification of lipid structure ($\alpha \rightarrow \beta'/\beta$); (*2*) the presence of additional colloidal structures (liposomes, micelles, drug nanoparticles).

Gelation phenomena, that is, increase of particle sizes and drug expulsion from the lipid carrier, are the major problems of storage stability. Gelation is the transformation of a low viscosity SLN colloidal dispersion into a viscous gel. It occurs when SLN is put in contact with other surfaces and shear forces, and it is connected with crystallization processes. This destabilizing phenomenon can be retarded or prevented by the addition of coemulsifying surfactants with high dynamic mobility, such as glycocholate (381).

The increase of particle sizes is a consequence of particle aggregation and it is less significant when SLN have a zeta potential value of -25 mV, while it becomes an important phenomenon when the zeta potential is -15 mV or less (381).

Drug expulsion is related to crystallization of lipids and their modification to the β'/β form, in which the lipid lattice is packed in a more ordered way with a reduction of imperfections.

To have an optimal storage conditions, SLN can be lyophilized or spray-dried. During lyophilization, a colloidal lipid dispersion is deprived of its solvent to guarantee a better chemical and physical stability. However, two transformations in the formulation occur during and after this process, which might be the source of additional stability problems: (*1*) passage of SLN from an aqueous dispersion to a powder with possible changes of osmolarity and pH; (*2*) resolubilization that favors particle aggregation. To overcome these problems, a cryoprotector (e.g., trehalose,

sorbitol, mannose, and glucose) is added to the SLN dispersion before lyophilizing. These protective agents are used in a 10–15% (w/v) concentration and act to decrease the osmotic activity of water and favour a glassy state of the frozen sample (384).

Spray drying is an alternative method to transform an aqueous SLN dispersion into a dry product. It is cheaper and simpler than lyophilization, but has the disadvantage of needing high temperatures, which can cause particle aggregation. Therefore, it is recommended to use lipids with melting points $>70\,^{\circ}C$ for spray drying. Also, in this case the addition of cryoprotective agents may be useful to prevent particle aggregation.

Drug incorporation into SLN is related to crystalline modification of the lipids and is inversely proportional to the β'/β modification of lipids. Depending on the drug/lipid ratio and solubility, the drug is located in the core of the particles, in the shell, or dispersed throughout the matrix, so that drug loading capacity of conventional SLNs is generally from 25 up to 50% (381). The drug-loading capacity is higher for lipid mixtures with different acyl chain lengths than for lipids that form a perfect crystal with few imperfections and cannot accommodate large amount of drug.

The release profiles could be modulated showing a burst release followed by a prolonged release, or generating systems without any burst release at all. The release kinetic can be controlled by modification of the preparation procedures and the type of surfactant and lipid material.

Therapeutic Applications of SLN. Due to their small sizes, SLN may be administered through every route: oral, transdermal, and IV administration can be possible.

The SLN for oral administration may include aqueous dispersions or conventional dosage forms (e.g., tablets, capsules, and pellets). Camptothecin-loaded particles are an example of orally administered SLN. This is a stearic acid/Poloxamer 188 formulation of SLN that present a zeta potential value of -69 mV and an encapsulation efficiency of 99.6%. The incorporation of camptothecin into SLN provided drug protection from hydrolysis (381). A better bioavailability, prolonged plasma levels, and lack of nephrotoxicity are observed for orally administered SLN encapsulating drugs, thus leading to the conclusion that SLN are a promising sustained release system for the oral administration of lipophilic drugs (381,385).

The SLN are formulated into creams, hydrogels, or ointments before their application onto the skin and form an adhesive film upon the skin, which is able to restore the protective action of the naturally occurring hydrolipidic skin film when it is damaged. Many different cosmetic ingredients have been encapsulated into SLN (i.e., coenzyme Q_{10}, vitamin E, and retinal) (386). A modern approach to an intelligent release of the drug from SLN to the skin is related to lipid modification from the α to the β form. An intelligent drug-loaded SLN is a colloidal system that maintains itself into the more energetic α-form during storage, while transforming into the β-form after application onto the skin, thus releasing its incorporated drug by expulsion from the lipid crystalline reticule. The SLN per se also have a sun protective effect (387) that is due to their particulate nature and their ability to scatter UV light. In this case, SLN show a synergistic effect if formulated with a molecular sunscreen, also showing better skin protection and a reduction in side effects.

It is possible to use SLN for pulmonary drug delivery since they maintain their particle size and polydispersion index after nebulization. Only very little aggregation could be detected, which is of no significance for pulmonary administration (388). In addition, SLN may be used as a powder for inhalation. The use of SLN instead of polymeric nanoparticles has many advantages, such as high tolerability, faster degradation, and passive targeting toward lung macrophages.

The SLN can be injected intravenously and used to target drugs to particular organs. They also can be administered intramuscularly or subcutaneously. When administered subcutaneously, SLN act as a depot of the drug, while they are cleared from the circulation by RES (liver and spleen) when administered intravenously. The incorporated drug is released upon erosion by diffusion from the particles or by enzyme degradation. Obviously, in the case of IV administration, the SLN size must be $< 5\,\mu m$ to avoid the possibility of embolism into the fine capillaries. Similar to any colloidal drug with a small mean size (≤ 200 nm), SLN have been coated by polyoxyethylene in order to achieve long circulating colloidal lipid particles (389). Stealth SLN can be prepared by using Pluronic F188, a block poly(oxyethylene/polyoxypropylene) copolymer that anchors its hydrophobic portion in the SLN lipid matrix, while the hydrophilic portion forms the hydrophilic coating of SLN. Stealth SLN can be used in tumor and antibacterial therapy. Paclitaxel-loaded SLN provided a higher and prolonged plasma level of the drug with respect to the cremophor EL-based commercial formulation with a consistent reduction of side effects (381). In all the investigated cases, SLN containing an anticancer drug showed higher blood levels with respect to the relative commercial drug formulations after IV injection.

The therapy at the level of the brain is always difficult due to the presence of the blood–brain barrier, and hence the possibility of having a suitable drug delivery system that is able to reach the brain can be extremely useful. Stealth SLN allow brain delivery of drugs that are not capable of passing through the blood–brain barrier (390).

The potential SLN toxicity has to be considered as a function of the administration route. Topical and oral administration are absolutely nonproblematic because all excipients used in SLN formulations are those currently employed for the formulation of traditional dosage forms or cosmetic products. The situation is slightly different for parenteral administration. In this case, only surfactants accepted for parenteral administration can be used (e.g., lecithin, Tween 80, PVP, Poloxamer 188). Up to now, there is no SLN product for parenteral use on the market. However, SLN show a very good tolerability both *in vitro* and *in vivo* (381).

Nanostructured Lipid Carrier. The NLC were introduced at the end of the 1990s in order to overcome some limitations of SLN: (*1*) too low payload for a number of drugs, (*2*) drug expulsion during storage, (*3*) high water content of SLN dispersions.

Figure 56. Schematization of the three types of NLCs.

The NLC are made up of very different lipid molecules mixed together, that is, a blend of solid with liquid lipids (oils). The resulting matrix of the lipid particles shows a melting point depression compared to the original solid lipid, but the matrix is still solid at body temperature. There are three different models of NLC, depending on the way of production and the composition of the lipid blend (Fig. 56). In the first model, called the imperfect type, spatially different lipids lead to larger distances between the fatty acid chains of the glycerides and to general imperfections in the crystal structure, thus allowing a greater extent of drug entrapment. The second type of NLC, the so-called amorphous type, contains liquid oil nanocompartments within the lipid particle matrix. In this way, crystallization can be avoided and the solid character of NLC can be maintained as shown by NMR and DSC measurements (391). The third type of NLC, the so-called multiple type, are produced by mixing a solid lipid with a high amount of oil. In this way, a phase separation occurs between solid lipids and oil molecules thus forming nanocompartments. This phenomenon occurs during the cooling process after the hot homogenisation method.

The NLC are produced successfully by the high pressure homogenization method and it is possible to obtain particle dispersions with a solid content of 50 or 60% (392). The particle dispersions thus produced have a high consistency with a cream-like or almost solid appearance.

Because of their high particle concentrations, NCL can be used for granulation or as wetting agents in pellet production. In addition, NLC can easier be processed to traditional oral dosage forms, such as tablets or capsules (393).

These carriers have been used for dermal delivery. Similarly to SLN, they can be incorporated in to existing products or formulated in a final product containing only NLC. When incorporated into an o/w emulsion saturated with the drug, the NLC disordered structure is preserved and the drug remains inside the particles. Following skin application, and then increased temperature, water loss and NLC lipid transition to a more stable polymorphic form triggers drug expulsion. The drug expelled from NLC is supersaturated with the drug already present in the emulsion. The supersaturation phenomenon can be used to increase the skin drug permeation as observed for cyclosporine (394).

In the future, an area of particular interest can be the prolonged release of drugs after subcutaneous or intramuscular injection (e.g. erythropoietin). Also, NLC dispersions for IV injection appear feasible (395).

To overcome the important limitation of SLN and NLC to incorporate only lipophilic drugs or very low concentrations of highly potent hydrophilic drugs, lipid drug conjugates were recently developed with drug loading capacities of up to 33% (396). In this type of lipid nanoparticle, an insoluble drug–lipid conjugate bulk is prepared using two different methods: (*1*) salt formation with a hydrophobic moiety (e.g., with a fatty acid) and (*2*) covalent linking (e.g., to ethers or ester).

FUTURE PRESPECTIVES

Drug delivery technology has had considerable advances, bringing many clinical products to the market. However, the major needs for drug delivery devices are still unmet and important classes of drugs have yet to benefit from these technological successes. The central focus of any controlled delivery devices is "control." This can b achieved if: (*1*) the size of the device can be modulated as accurately as possible; (*2*) the device can be produced with a certain reproducibility; (*3*) the device is stable enough for administration purpose; (*4*) the device is biocompatible; (*5*) the rate of drug delivery should be independent of the surrounding environment. To date, it is possible to have recourse to a lot of innovative approachs characterized by different features as a fucntion of the physicochemical properties of the delivered drug, the administration route, and therapeutic aims.

Many aspects of these drug delivery devices require an improvement to be applied to clinical use, but considering the *in vitro* and *in vivo* results, they seem to be very interesting.

From the appearance of the first drug delivery device to date, a constant improvement has been made, that is, from microtechnology we have passed to nanotechnology, and from an aspecific drug delivery, we have passed to a selective drug delivery. New challenges for the future are the feasibility of scaling-up processes to bring to the market quickly innovative therapeutic entities and the possibility of obtaining multifunctional devices that will be able to fulfill the different biological and therapeutic requirements.

BIBLIOGRAPHY

1. Park K. Controlled Drug Delivery: Challenges and Strategies. Washington (DC): American Chemical Society; 1997.
2. Park K, Mersny RJ. Controlled drug delivery: present and future. Controlled Drug Delivery: Designing Technology for the Future, 2–13. Washington (DC): American Chemical Society; 2000.
3. Juliano RL. Controlled delivery of drugs: an overview and prospectus. Drug Delivery Systems: Characteristics and Biomedical Application 310. Oxford University Press; 1980.
4. Langer R. Implantable controlled release systems. Pharmacol Ther 1983;21:35–51.
5. Langer R. Drug delivery and targeting. Nature (London) 1998;392:5–10.
6. Kost J, Langer R. Responsive polymeric delivery systems. Adv Drug Deliv Rev 2001;46:125–148.
7. Breimer DD. Future challenges for drug delivery. J Control Rel 1999;62:3–6.
8. Klausner EA, et al. Novel levodopa gastroretentive dosage form: in-vivo evaluation in dogs. J Control Rel 2003;88:117–126.

9. Ranade VV, Hollinger MA, editors. Drug Delivery Systems. CRC Press; 1996.
10. Powell MF. Drug delivery issues in vaccine development. Pharm Res 1996;13(12):1777–1785.
11. Grayson ACR, Shawgo RS, Li Y, Cima MJ. Electronics mems for triggered delivery. Adv Drug Deliv Rev 2004; 56:173–184.
12. Global drug delivery: Industry profile. Datamonitor; Aug 2003.
13. Schwan HP, Webb GN, editors. IRE Transactions on Biomedical Engineering.
14. Schwan HP, Webb GN. Biophys J 1967;7:978.
15. Greatbatch W, Holmes CF. History of implantable devices. IEEE Eng Med Biol Mag 1991;10(3):38–49.
16. Blackshear PJ, et al. A permanently implantable self-recycling low flow constant rate multipurpose infusion pump of simple design. Surg Forum 1970;21:136–137.
17. Blackshear PJ, Rohde TD, Prosl F, Buchwald H. The implantable infusion pump: a new concept in drug delivery. Med Prog Technol 1979;6:149–161.
18. Rupp WM, et al. The use of an implantable insulin pump in the treatment of type II diabetes. NEJM 1982;307:265–270.
19. Salkind AJ, et al. Electrically driven implantable prostheses. New York: Plenum Press; 1986.
20. Saudek CD, et al. A preliminary trial of the programmable implantable mediation system for insulin delivery. NEJM 1989;321:574–579.
21. Fogel H, et al. Treatment of iddm with a totally implantable programmable insulin infusion devise. Rev Euro Technol Biomed 1990;12:196.
22. Folkman J, Long DM, Rosenbaum R. Silicone rubber: a new diffusion property useful for general anesthesia. Science 1966;154:148–149.
23. Folkman J, Long D. The use of silicone rubber as a carrier for prolonged drug therapy. J Surg Res 1964;4:139–142.
24. Fung LK, Saltzman WM. Polymeric implants for cancer chemotherapy. Adv Drug Deliv Rev 1997;26:209–230.
25. Danckwerts M, Fassihili A. Optimization and development of a core-in-cup tablet for modulated release of theophylline in simulated gastrointestinal fluids. Drug Dev Ind Pharm 1991;17:1465–1502.
26. Hall EAH. The developing biosensor arena. Enz Microb Technol 1986;8:651–658.
27. Yang MB, Tamargo RJ, Brem H. Controlled delivery of 1,3-bis(2chloroethyl)-l-nitrosourea from ethylene-vinyl acetate copolymer. Cancer Res 1989;49:5103–5107.
28. Kim SW, Bae YH, Okano T. Hydrogels: swelling, drug loading, and release. Pharm Res 1992;9:283–290.
29. Gutman RL, Peacock G, Lu DR. Targeted drug delivery for brain cancer treatment. J Control Rel 2000;65:31–41.
30. Raiche AT, Puleo DA. Association polymers for modulated release of bioactive proteins. IEEE Eng Med Biol Mag 2003;22(5):35–41.
31. Emanuele AD, Staniforth JN. An electrically modulated drug delivery device. J Pharm Res 1991;8:913–918.
32. Sershen S, West J. Implantable, polymeric systems for modulated drug delivery. Adv Drug Deliv Rev 2002;54: 1225–1235.
33. Jackman RJ, et al. Fabricating large arrays of microwells with arbitrary dimensions and filling then using discontinuous dewetting. Anal Chem 1998;70:2280–2287.
34. Available at http://www.alzet.com.
35. Urquhart J, Fara J, Willis KL. Rate controlled delivery systems in drug and hormone research. Annu Rev Pharmacol Toxicol 1984;24:199–236.
36. Yun KS, et al. A surface-tension driven micropump for low-voltage and low-power operations. J Microelectromec Syst 2002;11:454–461.
37. Maillefer D, Van Lintel H, Rey-Mermet G, Hirschi R. A high-performance silicon micropump for an implantable drug delivery system. Proc 12th IEEE MEMS 1999; 541–546.
38. Zengerle R, Kluge S, Richter M, Richter A. A bidirectional silicon micropump. Proc IEEE MEMS 1995; 19–24.
39. Jeong OC, Yang SS. Fabrication and test of a thermopneumatic micropump with a corrugated p+ diaphragm. Sens Actuators A Phys 2000;83:249–255.
40. Zhang W, Ahn CH. A bidirectional magnetic micropump on a silicon wafer. Proc IEEE Solid-State Sensor Actuator Workshop 1996; 94–97.
41. Yang Y, Zhou Z, Ye X, Jiang X. A bimetallic thermally actuated micropump. J Microelectromec Syst 1996;59:351–354.
42. Benard WL, Kahn H, Heuer AH, Huff MA. Thin-film shape memory alloy actuated micropumps. J Microelectromec Syst 1998;7:245–251.
43. Guo S, Nakamura T, Fukuda T, Oguro K. A new type of micropump using icpf actuator. IEEEIASME Inter Conf Adv Intelligent Mechatronics 1997; 16–20.
44. Available at http://www.medtronic.com.
45. Moore G. VLSI, what does the future hold. Electron Aust 1980;42:14.
46. Nathanson HC. A resonant-gate silicon surface transistor with high-q band-pass properties. Appl Phys Lett 1965;7:84.
47. Petersen KE. Silicon as mechanical materials. Proc IEEE 1982;70(5):420.
48. Howe RT. Polycrystalline silicon micromechanical beams. I Electrochem Soc 1983;130(6):1420.
49. Mehregany M, Gabriel KJ, Trimmer WSN. IEEE Trans El Dev 1988;35:719.
50. Bart SF, et al. Design considerations for micromachined electric actuators sensors. Actuators 1988;14:269–292.
51. Fan LS, Tai YC, Muller RS. Ic-processed electrostatic micromotors. Sensors Actuators 1989;20:41–48.
52. Jaecklin VP, et al. Line-addressable torsional micromirrors for light modulator arrays. Sensors Actuators A 1994;41–42:324.
53. Fischer M, Graef H, von Munch W. Electrostatically deflectable polysilicon torsional mirrors. Sensors Actuators A 1994;44:83–89.
54. Tsang WK, Core TA, Sherman SJ. Fabrication technology for an integrated surface-micromachined sensor. Solid State Technol 1993; 36–39.
55. Chau KHL, et al. An integrated force-balanced capacitive accelerometer for low-g applications. Int Conf Solid-State Sensors Actuators 1995; 593.
56. Hierold C, et al. A pure cmos surface micromachined integrated accelerometer. IEEE Workshop Micro Electro Mech Syst 1996; 174.
57. Offenberg M, et al. Novel process for a monolithic integrated accelerometer. Int Conf Solid-State Sensors Actuators 1995; 589.
58. Burns DW, et al. Resonant microbeam accelerometers. Int Conf Solid-State Sensors Actuators 1995;2:659.
59. Hirano T, Furuhata T, Gabriel KJ, Fujita H. Design, fabrication, and operation of submicron gap comb-drive microactuators. J Microelectromec Syst 1995;1:52–59.
60. Suzuki H. Advances in the microfabrication of electrochemical sensors and systems. Electroanalysis 2000;12(9):703–715.
61. Wang J. Electrochemical detection for microscale analytical systems: a review. Talanta 2002;56:223–231.
62. Simpson PC, Woolley AT, Mathies A. Microfabrication technology for the production of capillary array electrophoresis chip. Biomed Microdev 1998;11:7–26.
63. Liu S, Guttman A. Electrophoresis microchips for dna analysis. Trends Anal Chem 2004;23(6):422–431.
64. Woolley AT, Lao K, Glazer AN, Mathies RA. Capillary electrophoresis chips with integrated electrochemical detection. Anal Chem 1998;70:684–688.

65. Liu W. Retinal implant: bridging engineering and medicine. Electron Devices Meeting, IEDM '02 Digest Inter 2002; 492–495.
66. Kipke DR, Vetter RJ, Williams JC, Hetke JY JF. Silicon-substrate intracortical microelectrode arrays for long-term recording of neuronal spike activity in cerebral cortex. IEEE Trans Neural Sys Rehab Eng 2003;11(2):151–155.
67. Mikszta JA, et al. Improved genetic immunization via micromechanical disruption of skin-barrier function and targeted epidermal delivery. Nature Med 2002;8(4):425419.
68. Santini Jr. JT, et al. Microchips as controlled drug delivery devices. Angew Chem Int Ed 2000;39:2396–2407.
69. McAllister DX, Allen MG, Prausnitz MR. Annu Rev Biomed 2000;2:289.
70. Tao S, Lubeley M, Desai TA. Bioadhesive poly(methyl methacrylate) microdevices for controlled drug delivery. J Control Rel 2003;88:215–228.
71. Lewis JR, Ferrari M. BioMEMS for drug delivery application, Lab-on-achip: Chemistry in miniaturized synthesis and analysis system. New York: Elsevier Science; 2003. pp 373–389.
72. Polla DL, et al. Microdevices in medicine. Annu Rev Biomed 2000;2:551–576.
73. Hadgraft J, Guy RH, editors. Transdermal Drug Delivery: Developmental Issues and Research Initiatives. New York: Marcel Dekker; 1989.
74. Smith EW, Maibach HI, editors. Percutaneous Penetration Enhancers. CRC Press; 1995.
75. Amsden BG, Goosen MFA. Transdermal delivery of peptide and protein drugs: an overview. AIChE J 1995;41:1972–1997.
76. Henry S, McAllister DX, Allen MG, Prausnitz MR. Microfabrication microneedles: a novel approach to transdermal drug delivery J Pharm Sci 1998;87:922–925.
77. Barry B, Williams A. Penetration enhancers. Adv Drug Deliv Rev 2003;56:603–618.
78. Cevc G. Lipid vesicles and other colloids as drug carriers on the skin. Adv Drug Deliv Rev 2004;56:675–711.
79. Preat V, Vanbever R. Skin electroporation for transdermal and topical delivery. Adv Drug Deliv Rev 2004;56:659–674.
80. Doukas A. Transdermal delivery with a pressure wave. Adv Drug Deliv Rev 2004;56:559–579.
81. Mitragotri S, Kost J. Low-frequency sonophoresis: a review. Adv Drug Deliv Rev 2004;56:589–601.
82. Gerstel MS, Place VA. Drug delivery device, US Patent 3,964,482, 1976.
83. Reed ML. Microsystems for drug and gene delivery. Proc IEEE 2004;92(1):56–75.
84. Lin L, Pisano AP. Silicon-processed microneedles. J Microelectromec Syst 1999;8(l).
85. Talbot NH, Pisano AP. Polymolding: Two wafer polysilicon micromolding of closed-flow passages for microneedles and microfluidic devices. Tech Dig Solid-State Sensor and Actuator Workshop 1998; 265–268.
86. Brazzle JD, Papautsky L, Frazier AB. Micromachined needle arrays for drug delivery or fluid extraction. IEEE Eng Med Biol Mag 1999;18:53–58.
87. Brazzle JD, Papautsky L, Frazier AB. Fluid-coupled hollow metallic micromachined needle arrays. Proc SPIE, Microfluidic Devices Systems 1998;3515:116–124.
88. Brazzle JD, Mohanty S, Frazier AB. Hollow metallic micromachined needles with multiple output ports. Proc SPIE, Microfluidic Devices Systems 1999;11, 3877:257–266.
89. Stoeber B, Liepmann D. Two-dimensional arrays of out-ofplane needles. Proc ASME Int Mechanical Engineering Congr Exposition 2000; 355–359.
90. Stoeber B, Liepmann D. Fluid injection through out-of-plane microneedles. Proc Ist Annu Int IEEE-EMBS Special Topic Conf Microtechnologies Medicine Biology 2000; 224–228.
91. Henry S, McAllister DV, Allen MG, Prausnitz MR. Micromachined needles for the transdermal delivery of drugs. Proc IEEE 11th Annu Int Workshop Micro Electro Mechanical Systems 1998; 494–498.
92. Gardeniers JGE, et al. Silicon micromachined hollowmicroneedles for transdermal liquid transfer. Proc IEEE Conf MEMS 2002; 141.
93. Dizon R, Han H, Russell AG, Reed ML. An ion milling pattern transfer technique for fabrication of three-dimensional micromechanical structures. J Microelectromech Syst 1993;2(4): 151–159.
94. Ling P, et al. Genetic transformation of nematodes using arrays of micromechanical piercing structures. J Microelectromec Syst 1995;19(5):766–770.
95. Trimmer W, et al. Injection of dna into plant and animal tissues with micromechanical piercing structures. Proc 8th Int Workshop Micro Electra Mechanical Systems 1995; 111–115.
96. Feldman MD, et al. Stent-based gene therapy. J Long-Term Effects Med Implants 2000;10:47–68.
97. Saffran M, et al. Biodegradable azopolymer coating for oral delivery of peptide drugs. Biochem Soc Trans 1990;18:752–754.
98. Fara JW, Myrback RE, Swanson DR. Evaluation of oxprenolol and metoprolol Oros systems in the dog: comparison of in vivo and in vitro drug release, and of drug absorption from duodenal and colonic infusion sites. Br J Clin Pharmacol 1985;19:91–95.
99. Fasano A, Uzzau S. Modulation of intestinal tight junctions by Zonula occludens toxin permits enteral administration of insulin and other macromolecules in an animal model. J Clin Invest 1997;99:1158–1164.
100. Schwarz UI, et al. P-glycoprotein inhibitor erythromycin increases oral bioavailability of talinolol in humans. Int J Clin Pharmacol Ther 2000;38:161–167.
101. Arangoa MA, et al. Bioadhesive potential of gliadin nanoparticulate systems. Eur J Pharm Sci 2000;11:333–341.
102. Lehr CM. Lectin-mediated drug delivery: the second generation of bioadhesives. J Control Rel 2000;65:19–29.
103. Bies C, Lehr CM, Woodley JF. Lectin-mediated drug targeting: history and applications. Adv Drug Deliv Rev 2004;56: 425–435.
104. Nashat AH, Moronne M, Ferrari M, Detection of functional groups and antibodies on microfabricated surfaces by confocal microscopy. Biotechnol Bioeng 1998;60:137–146.
105. Ferrari M. Therapeutic microdevices and methods of making and using same, US Patent 6,107,102, 2000.
106. Ferrari M, et al. Particles for oral delivery of peptides and proteins, US Patant 6,355,270 131, 2000.
107. Tao SL, Desai TA. Microfabricated drug delivery systems: from particles to pores. Adv Drug Deliv Rev 2003;55:315–328.
108. Martin FJ, Grove C. Xlicrofabricated drug delivery systems: concepts to improve clinical benefit. Biomed Microdev 2001;3:97–108.
109. Smith BR, et al. Nanodevices in Biomedical Applications. Lee A, Lee J, Ferrari M, editors. BioMEMS and Biomedical Nanotechnology, Vol. I: Biological and Biomedical Nanotechnology. New York: Springer; 2005.
110. Foraker AB, et al. Microfabricated porous silicon particles enhance paracellular delivery of insulin across intestinal Caco-2 cell monolayers. Pharm Res 2003;20:110–116.
111. Ahmed A, Bonner C, Desai TA. Bioadhesive microdevices with multiple reservoirs: a new platform for oral drug delivery. J Control Rel 2002;81:291–306.
112. Ishida O, Maruyama K, Sasaki K, Iwatsuru M. Size-dependent extravasation and interstitial localization of polyethyleneglycol liposomes in solid tumor-bearing mice. Int J Pharm 1999;190:49–56.

113. Li X, et al. Porosified silicon wafer structures impregnated with platinum anti-tumor compounds: Fabrication, characterization, and diffusion studies. Biomed Microdev 2000; 2:265–272.
114. Smith BR, et al. A biological perspective of particulate nanoporous silicon. Mat Technol 2004;19:16.
115. Canham LT, et al. Derivatized mesoporous silicon with dramatically improved stability in simulated human blood plasma. Adv Mater 1999;11:1505–1507.
116. Stewart MP, Buriak JM. Chemical and biological applications of porous silicon technology. Adv Mater 2000;12:859–869.
117. Microchips Inc. Available at www.mchips.com.
118. Cohen MH, et al. Microfabrication of silicon-based nanoporous particulates for medical applications. Biomed Microdev 2003;5:253–259.
119. Santini Jr. JT, Cima MJ, Langer R. A controlled-release microchip. Nature (London) 1999;397:335–338.
120. Richards AC, et al. A biomems review: Mems technology for physiologically integrated devices. Proc IEEE 2004; 82:6–21.
121. Ferrari M, et al. Proc Mat Res Soc. 1995;414:101–106.
122. Chu WH, Ferrari M. Micromachined filter and capsule having porous membranes and bulk support, US Patent 5,570,076, 1996.
123. Kellar CG, Ferrari M. Microfabricated particle filter, US Patent 5,651,900, 1997.
124. Desai TA, Hansford D, Ferrari M. Characterization of micromachined silicon membranes for immunoisolation and bioseparation applications. J Membr Sci 1999;159:221–231.
125. Desai TA, et al. Nanoparous anti-fouling silicon membranes for implantable biosensor applications. Biosensors Bioelectronics 2000;15:453–462.
126. Leoni L, Boiarski A, Desai TA. Characterization of nanoporous membranes for immunoisolation: diffusion properties and tissue effects. Biomed Microdev 2002;4(2):131–139.
127. Clark LA, Ye GT, Snurr RQ. Molecular traffic control in a nanoscale system. Phys Rev Let 2000;84:2893–2896.
128. Wei Q, Bechinger C, Leiderer P. Single-file diffusion of colloids in onedimensional channels. Science 2000;287:625–627.
129. Auerbach SM. Theory and simulation of jump dynamics, diffusion and phase equilibrium in nanopores. Int Rev Phys Chem 2000;19:155–198.
130. Mao Z, Sinnott SB. A computational study of molecular diffusion and dynamic flow through carbon nanotubes. J Phys Chem B 2000;104:4618–4624.
131. Nelson P, Auerbach S. Self-diffusion in single-file zeolite membranes is fickian at long times. J Chem Phys 1999; 110:9235–9243.
132. MacElroy JMD, Suh SH. Self-diffusion in single-file pores of finite length. J Chem Phys 1997;106:85–95.
133. Levitt DG. Dynamics of a single-file pore: Non-fickian behavior. Phys Rev A Gen Phy 1973;8:30–50.
134. Lin B, Yu J, Rice S. Direct measurements of constrained brownian motion of an isolated sphere between two walls. Phys Rev E 2000;62:3909–3919.
135. Cosentino C, et al. A dynamic model of biomolecules diffusion through two-dimensional nanochannels. J Phys Chem B 2005;109: to be published.
136. Tu JK, Huen T, Szema R, Ferrari M. Filtration of sub-100 run particles using a bulk micromachined, direct-bonded silicon filter. Biomed Microdev 1999;1:113–119.
137. Desai TA, et al. Nanopore technology for biomedical applications. Biomed Microdev 1999;2:11–40.
138. Leoni L, Desai TA. Micromachined biocapsules for cell-based sensing and delivery. Adv Drug Deliv Rev 2004;56:211–229.
139. Tu JK, Ferrari M, Microfabricated particle filter, US Patent 5,938,923, 1999.
140. Kim SJ, Kim BH. Syntheses and structural studies of calix[4]arene-nucleoside and calix[4]arene-oligonucleotide hybrids. Nucleic Acids Res 2003;31:272–274.
141. Murakami Y, Hayashida O. Supramolecular effects and molecular discrimination by macrocyclic hosts embedded in synthetic bilayer membranes. Proc Nat Acad Sci 1993;90: 1140–1145.
142. Szetjli J. Introduction and general overview of cyclodextrin chemistry. Chem Rev 1998;98:1743–1753.
143. Hirayama F, et al. Utilization of diethyl–ß cyclodextrin as a sustained-release carrier for isosorbide dinitrate. J Pharm Sci 1989;77:233–236.
144. Paolino D, Puglisi G, Ventura CA, Fresta M. Lecithin Organogels: Effect of Drug Physico-Chemical Characteristics on Matrix Release. European Conference on Drug Delivery and Pharmaceutical Technology, 97; 2004.
145. Eastburn SD, Tao BY. Applications of modified cyclodextrins. Biotech Adv 1994;12:325–339.
146. Muňoz-Botella S, del Castillo B, Martin MA. Cyclodextrin properties and applications of inclusion complex formation. Ars Pharm 1995;36:187–198.
147. Loftsson T, Brewster ME. Pharmaceutical applications of cyclodextrins: Drug solubilisation and stabilization. J Pharm Sci 1996;85:1017–1025.
148. Schneiderman E, Stalcup AM. Cyclodextrins: a versatile tool in separation science. J Chromatogr B 2000;745:83–102.
149. Schmid G. Cyclodextrin glucanotransferse production: yield enhancement by overexpression of cloned genes. Trends Biotechnol 1989;7:244–248.
150. Loftsson T, Fridriksdottir H, Sigurdardottir AM, Ueda H. The effect of water-soluble polymers on drug-cyclodextrin complexation. Int J Pharm 1994;110:169–177.
151. Redenti E, Szente L, Szejtli J. Drug/cyclodextrin/hydroxyacid multicomponent systems. Properties and pharmaceutical applications. J Pharm Sci 2000;89:1–8.
152. Loftsson T, Másson M, Sigurjónsdóttir JF, Methods to enhance the complexation efficiency of cyclodextrins. STP Pharma Sci 1999;9:237–242.
153. Jindrich J, et al. Regioselectivity of alkylation of cyclomaltoheptaose and synthesis of its mono-2-*O*-methyl, -ethyl, -allyl, and propyl, derivatives. Carbohydr Res 1995;266:75–80.
154. Gazpio C, et al. HPLC and solubility study of the interaction between pindolol and cyclodextrins. J Pharm Biomed Anal 2005;37:487–492.
155. Koizumi K, Kubota Y, Utamura T, Horiyama S. Analysis of heptakis 2,6,-di-*O*-methyl-ß-cyclodextrin by thin-layer chromatography, HPLC and gas chromatography/mass spectrometry. J Chromatogr 1986;368:329–337.
156. Uccello-Barretta G, et al. Combining NMR and molecular modelling in a drug delivery context: investigation of the multi-mode inclusion of *trans-n*-{4-[*n'*-(4-chlorobenzoyl)-hydrazinocarbonyl] cyclohexylmethyl}-4-romobenzenesulfonamide, a new chemotype of npy-5 antagonist, into β-cyclodextrin. Bioorg Med Chem 2004;12:447–458.
157. Bakirci H, Zhang X, Nau WM. Induced circular dichroism and structural assignment of the cyclodextrin inclusion complexes of bicyclic azoalkanes. J Org Chem 2005;70: 39–46.
158. Puglisi G, et al. Preparation and Physico-Chemical Study of Inclusion Complexes between Idebenone and Modified β-Cyclodextrins. J Inclus Phenom 1996;24:193–210.
159. Beaufour M, Morin P, Ribet JP. Chiral separation of the four stereoisomers of a novel antianginal agent using a dual cyclodextrin system in capillary electrophoresis. J Sep Sci 2005;28:529–533.
160. Wang Z, et al. Enantioseparation of chiral allenic acids by micellar electrokinetic chromatography with cyclodextrins as chiral selector. Electrophoresis 2005;26:1001–1006.

161. Liang M, Qi M, Zhang C, Fu R. Peralkylated-beta-cyclodextrin used as gas chromatographic stationary phase prepared by sol–gel technology for capillary column. J Chromatogr A 2004;1059:111–119.
162. Geze A, et al. Long-term shelf stability of amphiphilic beta-cyclodextrin nanosphere suspensions monitored by dynamic light scattering and cryo-transmission electron microscopy. J Microencapsul 2004;21:607–613.
163. Irie T, Uekama K. Pharmaceutical applications of cyclodextrins. III. Toxicological issues and safety evaluation. J Pharm Sci 1997;86:147–162.
164. Thompson DO. Cyclodextrins-enabling excipients: their present and future use in pharmaceuticals.Crit Rev Ther Drug Carrier Syst 1997;14:1–104.
165. Rajewski RA, et al. Preliminary safety evaluation of parenterally administered sulfoalkyl ether β-cyclodextrin derivatives. J Pharm Sci 1995;84:927–932.
166. Stella VJ, Rajewski RA. Cyclodextrins: their future in drug formulation and delivery. Pharm Res 1997;14:556–567.
167. Lopez RFL, Collett JH, Bentley MVLB. Influence of cyclodextrin compl exation on the *in vitro* permeation and skin metabolism of dexamethasone. Int J Pharm 2000;200:127–132.
168. Reeuwijk HJEM, et al. Liquid chromatographic determination of b-cyclodextrin derivatives based on fluorescence enhancement after inclusion complexation. J Chromatogr 1993;614:95–101.
169. Kublik H, Bock TK, Schreier H, Muller BW. Nasal absorption of 17b-estradiol from different cyclodextrin inclusion formulations in sheep. Eur J Pharm Biopharm 1996;42: 320–324.
170. Loftsson T, Masson M. Cyclodextrins in topical formulations: Theory and practice. Int J Pharm 2001;225:15–30.
171. Keipert S, Fedder J, Bohm A, Hanke B. Interactions between Cyclodextrins and pilocarpine-as an example of a hydrophilic drug. Int J Pharm 1996;142:153–162.
172. Davies NM, Wang G, Tucker IG. Evaluation of a hydrocortisone/hydroxypropyl-b-cyclodextrin solution for ocular drug delivery. Int J Pharm 1997;156:201–209.
173. Jarho P, et al. Increase in aqueous solubility, stability and in vitro corneal permeability of anandamide by hydroxypropyl-β-cyclodextrin. Int J Pharm 1996;137:209–217.
174. Loftsson T, Fridriksdottir H. The effect of water-soluble polymers on the aqueous solubility and complexing abilities of β-cyclodextrin. Int J Pharm 1998;163:115–121.
175. Sigurdardottir AM, Loftsson T. The effect of polyvinylpyrrolidone on cyclodextrin complexation of hydrocortisone and its diffusion through hairless mouse skin. Int J Pharm 1995;126:73–78.
176. Tanaka M, et al. Effect of 2-hydroxypropyl-β-cyclodextrin on percutaneous absorption of methyl paraben. J Pharm Pharmacol 1995;47:897–900.
177. Vitória M, et al. Characterization of the influence of some cyclodextrins on the stratum corneum from the hairless mouse. J Pharm Pharmacol 1997;49:397–402.
178. Loftsson T, Sigurjónsdóttir AM. The effect of polyvinylpyrrolidone and hydroxypropyl methylcellulose on HP-β-CD complexation of hydrocortisone and its permeability through hairless mouse skin. Eur J Pharm Sci 1994;2:297–301.
179. Masson M, Loftsson T, Masson V, Stefansson E. Cyclodextrins as permeation enhancers: some theoretical evaluations and in vitro testing. J Control Rel 1999;59:107–118.
180. Bhatnagar S, Vyas SP. Organogel-based system for transdermal delivery of propanolol. J Microencapsul 1994;11:431–438.
181. Ventura CA, et al. Biomembrane model interaction and percutaneous absorption of papaverine through rat skin: effects of cyclodextrins as penetration enhancers. J Drug Target 2001;19:379–393.
182. Arima H, Kondo T, Irie T, Uekama K. Enhanced rectal absorption and reduced local irritation of the anti-inflammatory drug ethyl 4-biphenylylacetate in rats by complexation with water-soluble b-cyclodextrin derivatives and formulation as oleaginous suppository. J Pharm Sci 1992;81:1119–1125.
183. Ventura CA, et al. Celecoxib-Dimethyl-β-Cyclodextrin inclusion complex. Characterization and in vitro permeation study. Eur J Med Chem 2005;40:624–631.
184. Brewster ME, Anderson WR, Estes KS, Bodor N. Development of aqueous parenteral formulations for carbamazepine through the use of modified cyclodextrins. J Pharm Sci 1991;80:380–383.
185. Sridevi S, et al. Enhancement of dissolution and oral bioavailability of gliquidone with hydroxy propyl-β-cyclodextrin. Pharmazie 2003;58:807–810.
186. Uekama K, Otagiri M. Cyclodextrins in drug carrier systems. CRC Crit Rev Ther Drug Carrier Syst 1987;3:1–40.
187. Tokumura T, et al. Enhancement of bioavaiability of cinnarizine from its ß-cyclodextrin complex on oral administration with D,L-phenylalanine as a competing agent. J Pharm Sci 1986;75:391–394.
188. Nakanishi K, Masada M, Nadai T, Miyajima K. Effect of the interaction of drug–ß-cyclodextrin complex with bile salts on the drug absorption from rat small intestinal lumen. Chem Pharm Bull 1989;37:211–214.
189. Uekama K, et al. Improvement of the oral bioavailability of digitalis glycosides by cyclodextrin complexation. J Pharm Sci 1983;72:1338–1341.
190. Stuenkel CA, Dudley RE, Yen SS. Sublingual administration of testosterone–hydroxypropyl–ß-cyclodextrin inclusion complex simulates episodic androgen release in hypogonadal men. J Clin Endocrinol Metab 1991;72:1054–1059.
191. Fridriksdottir H, et al. Design and in vivo testing of 17b-estradiol–hydroxypropyl–ß-cyclodextrin sublingual tablets. Pharmazie 1996;51:39–42.
192. Hirayama F, Uekama K. Cyclodextrin-based controlled drug release system. Adv Drug Deliv Rev 1999;36:125–141.
193. Szejtli J. 1997; Cyclodextrin News 11, Budapest Cyclolab.
194. Shiotani K, Irie T, Uekama K, Ishimaru Y. Cyclodextrin sulfates in parenteral use: Protection against gentamicin nephrotoxicity in the rat. Eur J Pharm Sci 1995;3:139–151.
195. Bekers O, et al. Effect of cyclodextrins on the chemical stability of mitomycins in alkaline solution. J Pharm Biomed Anal 1991;9:1055–1060.
196. Alcaro S, et al. Preparation, Characterization, Molecular Modeling and In Vitro Activity of Paclitaxel-Cyclodextrin Complexes. Bioorg Med Chem Lett 2002;12:1673–1641.
197. Irie T, Uekama K. Cyclodextrins in peptide and protein delivery. Adv Drug Deliv Rev 1999;36:101–123.
198. Redenti E, Pietra C, Gerloczy A, Szente L. Cyclodextrin in oligonucleotide delivery. Adv Drug Deliv Rev 2001;53:235–244.
199. Zhao T, Temsamani J, Agarwal S. Use of cyclodextrin and its derivatives as carriers for oligonucleotide delivery. Antisense Res 1995;5:185–192.
200. Dass CR, Jessup W, Apolipoprotiens AI. Cyclodextrins and liposomes as potential drugs for the reversal of atherosclerosis. J Pharm Pharmacol 2000;52:731–761.
201. Croyle MA, Cheng X, Wilson JM. Development of formulations that enhance physical stability of viral vectors for gene therapy. Gene Ther 2001;8:1281–1290.
202. Nicolazzi C, et al. Effect of the complexation with cyclodextrins on the in vitro antiviral activity of ganciclovir against human cytomegalovirus. Bioorg Med Chem 2001;9:275–282.
203. Hovgaad L, Broendsted H. Current applications of polysaccharides in colon targeting. CRC Crit Rev Ther Drug Carrier Syst 1996;13:185–223.

204. Minami K, Hirayama F, Uekama K. Colon-specific drug delivery based on a cyclodextrin prodrug: release behavior of biphenylylacetic acid from its cyclodextrin conjugates in rat intestinal racts after oral administration. J Pharm Sci 1998;87:715–720.
205. Lawrence MJ, Rees GD. Microemulsion-based media as novel drug delivery systems. Adv Drug Deliv Rev 2000;45: 89–121.
206. Hoar TP, Schulman JH. Transparent water-in-oil dispersions: the oleopathic hydro-micelle. Nature (London) 1943;152:102–103.
207. Schulman JH, Stoeckenius W. Prince LM. Mechanism of formation and structure of microemulsions by electron microscopy. J Phys Chem 1959;63:1677–1680.
208. Shinoda K, Lindman B. Organised surfactant systems: microemulsions. Langmuir 1987;3:135–149.
209. Lam AC, Schechter RS. The theory of diffusion in microemulsions. J Colloid Interface Sci 1987;120:56–63.
210. Kriwet K, Muller-Goymann CC. Diclofenac release from phospholipid drug systems and permeation through excised human stratum corneum. Int J Pharm 1995;125:231–242.
211. Schmalfuss U, Neubert R, Wohlrab W, Modification of drug penetration into human skin using microemulsions. J Control Rel 1997;46:279–285.
212. Osborne DW, Ward AJ, O'Neill KJ. Microemulsions topical drug delivery vehicles: in-vitro transdermal studies model hydrophilic drug. J Pharm Pharmacol 1991;43:450–454.
213. Rosano HL, Cavello JL, Chang DL, Whittham JH. Microemulsions: a commentary on their preparation. J Soc Cosmet Chem 1988;39:201–209.
214. Grunewald AM, Gloor M, Gehring W, Kleesz P. Damage to the skin by repetitive washing. Contact Dermatitis 1995;32:225–232.
215. Effendy I, Maibach HI. Surfactants and experimental dermatitis. Contact Dermatitis 1995;33:217–225.
216. Paolino D, et al. Lecithin microemulsions for the topical administration of ketoprofen: percutaneous adsorption through human skin and in vivo human skin tolerability. Int J Pharm 2002;244:21–31.
217. Trotta M, Morel S, Gasco MR. Effect of oil phase composition on the skin permeation of felodipine from o/w microemulsions. Pharmazie 1997;52:50–53.
218. Delgado-Charro MB, et al. Delivery of a hydrophilic solute through the skin from novel microemulsion systems. Eur J Pharm Biopharm 1997;43:37–42.
219. Kriwet K, Muller-Goymann CC. Diclofenac release from phospholipid drug systems and permeation through excised human stratum corneum. Int J Pharm 1995;125:231–242.
220. Attwood D. Colloydal Drug Delivery Systems. New York: Marcel Dekker; 1994.
221. Benita S, Muchtar S. Ophthalmic Compositions. Eur. Patent 0521,799,A1, 1992.
222. Vandamme TF. Microemulsions as ocular drug delivery systems: recent Developments and future challenges. Prog Retinal Eye Res 2002;21:15–34.
223. Scartazzini R, Luisi PL. Organogels from Lecithins. J Phys Chem 1988;92:595–596.
224. Yurtov EV, Murashova NM. Lecithin Organogels in Hydrocarbon Oil. Colloid J 2003;65:114–118.
225. Willimann HL, Luisi PL. Lecithin organogels as matrix for transdermal transport of drugs. Biochem Biophys Commun 1991;177:897–900.
226. Willimann HL, et al. Lecithin Organogels as Matrix for Transdermal Transport of Drugs. J Pharm Sci 1992;81: 871–874.
227. Shchipunov YA, Shumilina EV. Lecithin bridging by hydrogen bonds in the organogel. Mat Sci Eng 1995;3:43–50.
228. Paolino D, et al. *In Vivo* Evaluation of Lecithin Organogels for Transdermal Application, 30th Annual Meeting & Exposition of the Controlled Release Society. Glasgow (UK); 2003.
229. Shah JC, Sadhale Y, Chilukuri DM. Cubic phase gels as drug delivery systems. Adv Drug Deliv Rev 2001;47:229–250.
230. Chasin M, Langer R. Biodegradable Polymers as Drug Delivery Systems. New York: Marcel Dekker; 1990.
231. Rolland A. Pharmaceutical Particulate Carriers: Therapeutic Applications. New York: Marcel Dekker; 1993.
232. Schwab G, et al. Antisense oligonucleotides adsorbed to polyalkylcyanoacrylate nanoparticles specifically inhibit mutated Ha-ras-mediated cell proliferation and tumorigenicity in nude mice. Proc Natl Acad Sci USA 1994;91:10460–10464.
233. Benns JM, Kim SW. Tailoring new gene delivery designs for specific targets. J Drug Target 2000;8:1–12.
234. Pastan I, Chaudhary V, Fitzgerald DJ. Recombinant toxins as novel therapeutic agents. Annu Rev Biochem 1992;61:331–354.
235. Moghimi SM, Christy Hunter A, Murray JC. Long-Circulating and Target-Specific Nanoparticles: Theory to Practice. Pharmacol Rev 2001;53:283–318.
236. Allen TM, Moase EH. Therapeutic opportunities for targeted liposomal drug delivery. Adv Drug Deliv Rev 1996;21:117–133.
237. Bakker-Woudenberg IAJM. Delivery of antimicrobials to infected tissue macrophages. Adv Drug Deliv Rev 1995; 17:5–20.
238. Ten Hagen TL, Van Vianen W, Bakker-Woudenberg IAJM. Modulation of nonspecific antimicrobial resistance of mice to Klebsiella pneumoniae septicemia by liposome-encapsulated muramyl tripeptide phosphatidylethanolamine and interferon-gamma alone or combined. J Infect Dis 1995;171: 385–392.
239. Russell-Jones GJ. Oral vaccine delivery. J Control Rel 2000;65:49–54.
240. Woodle MC, Storm G. Long Circulating Liposomes: Old Drugs, New Therapeutics. Berlin: Springer-Verlag; 1998.
241. B Ruozi, et al. Atomic force microscopy and photon correlation spectroscopy: Two techniques for rapid characterization of liposomes. E J Pharm Sci 2005;25:81–89.
242. Gimbert LJ, Haygarth PM, Beckett R, Worsfold PJ. Comparison of centrifugation and filtration techniques for the size fractionation of colloidal material in soil suspensions using sedimentation field-flow fractionation. Environ Sci Technol 2005;39:1731–1735.
243. K Thode, Muller RH, Kresse M. Two-time window and multi-angle photon correlation spectroscopy size and zeta potential analysis-highly sensitive rapid assay for dispersion stability. J Pharm Sci 2000;89:1317–1324.
244. Ford JL, Timmins P. Pharmaceutical thermal analysis techniques and applications. Southampton: John Wiley & Sons, Inc.
245. Venkateswarlu V, Manjunath K. Preparation, characterization and in vitro release kinetics of clozapine solid lipid nanoparticles. J Control Rel 2005;95:627–638.
246. Bower PV, et al. Solid-state NMR structural studies of peptides immobilized on gold nanoparticles. Langmuir 2005;21: 3002–3007.
247. Frohlich M, Brecht V, Peschka-Suss R. Parameters influencing the determination of liposome lamellarity by ^{31}P-NMR. Chem Phys Lipids 2001;109:103–112.
248. Lasic DD. Novel applications of liposomes. TIBTECH 1998; 16:307–321.
249. Drummond DC, et al. Optimizing liposomes for delivery of chemotherapeutic agents to solid tumors. Pharmacol Rev 1999;51:691–743.
250. Waterhouse DN, et al. Preparation, characterization and biological analysis of liposomal formulations of vincristine. Methods Enzymol 2005;391:140–157.

251. Park JW. Liposome-based drug delivery in breast cancer treatment. Breast Cancer Res 2002;4:95–99.
252. Senior JH. Fate and behaviour of liposomes in vivo: A review of controlling factors. CRS Crit Rev Ther Drug Carrier Syst 1987;3:123–193.
253. Senior J, Gregoriadis G. Stability of small unilamellar liposomes in serum and clearance from the circulation: the effect of the phospholipid and cholesterol components. Life Sci 1982;30:2123–2136.
254. Allen TM, et al. Liposomes containing synthetic lipid derivates of poly (ethylene glycol) show prolonged circulation half-live in vivo. Biochim Biophys Acta 1991;1066:29–36.
255. Pedroso de Lima MC, et al. Cationic lipid–DNA complexes in gene delivery: from biophysics to biological applications. Adv Drug Deliv Rev 2001;47:277–294.
256. Crosasso P, et al. Preparation, characterization and properties of sterically stabilized paclitaxel-containing liposomes. J Control Rel 2000;63:19–30.
257. Paolino D, et al. Tolerability and improved protective action of idebenone-loaded pegylated liposomes on ethanol-induced injury in primary cortical astrocytes. J Pharm Sci 2004;93:1815–1827.
258. Sudimack JJ, Guo W, Tjarks W, Lee RJ. A novel pH-sensitive liposome formulation containing oleyl alcohol. Biochim Biophys Acta 2002;1564:31–37.
259. Torchilin VP, Rammohan R, Weissig V, Levchenko TS. TAT peptide on the surface of liposomes affords their efficient intracellular delivery even at low temperature and in the presence of metabolic inhibitors. Proc Natl Acad Sci USA 2001;98:8786–8791.
260. Fresta M, et al. Characterization and in vivo ocular absorption of liposome-encapsulated acyclovir. J Pharm Pharmacol 1999;51:565–576.
261. Szoka Jr F, Papahadjopoulos D. Procedure for preparation of liposomes with large internal aqueous space and high capture by reverse-phase evaporation. Proc Natl Acad Sci USA 1978;75:4194–4198.
262. Duzgunes N. Preparation and quantitation of small unilamellar liposomes and large unilamellar reverse-phase evaporation liposomes. Methods Enzymol 2003;367:23–27.
263. Fresta M, Wehrli E, Puglisi G. Neutrase entrapment in stable multilamellar and large unilamellar vesicles for the acceleration of cheese ripening. J Microencapsul 1995;12:307–325.
264. Mayer LD, Hope MJ, Cullis PR, Janoff AS. Solute distributions and trapping efficiencies observed in freeze-thawed multilamellar vesicles. Biochim Biophys Acta 1985;817: 193–196.
265. Alino SF, et al. High encapsulation efficiencies in sized liposomes produced by extrusion of dehydration-rehydration vesicles. J Microencapsul 1990;7:497–503.
266. Hunter DG, Frisken BJ. Effect of extrusion pressure and lipid properties on the size and polydispersity of lipid vesicles. Biophys J 1998;74:2996–3002.
267. Celano M, et al. Cytotoxic effects of gemcitabine-loaded liposomes in human anaplastic thyroid carcinoma cells. BMC Canc 2004;4:63.
268. Clerc S, Barenholz Y. Loading of amphipatic weak acids into liposomes in response to transmembrane calcium acetate gradients. Biochim Biophys Acta 1995;1240:257–265.
269. Cullis PR, et al. Influence of pH gradient on the transbilayer transport of drugs, lipids, peptides and metal ions into large unilamellar vesicles. Biochim Biophys Acta 1997;1331:187–211.
270. Fresta M, Ventura CA, Mezzasalma E, Puglisi G. A calorimetric study on the idebenone-phospholipid membrane interaction. Int J Pharm 1998;163:133–143.
271. CB Hansen TM, Lopes de Menezes DE. Pharmacokinetics of long-circulating liposomes. Adv Drug Deliv Rev 1995;16:267–284.
272. Barenholz Y, et al. Stability of liposomal doxorubicin formulations: problems and prospects. Med Res Rev 1993;13:449–491.
273. Kanter PM, et al. Preclinical toxicology study of liposome encapsulated doxorubicin (TLC D-99): comparison with doxorubicin and empty liposomes in mice and dogs. *In Vivo* 1993;7:85–95.
274. Charrois GJ, Allen TM. Multiple injections of pegylated liposomal Doxorubicin: pharmacokinetics and therapeutic activity. J Pharmacol Exp Ther 2003;306:1058–1067.
275. Maruyama K, Ishida O, Takizawa T, Moribe K. Possibility of active targeting to tumor tissue with liposomes. Adv Drug Deliv Rev 1999;40:89–102.
276. Iden DL, Allen TM. *In vitro* and *in vivo* comparison of immunoliposomes made by conventional coupling techniques with those made by a new post-insertion approach. Biochim Biophys Acta 2001;1531:207–216.
277. Forssen E, Willis M. Ligand-targeted liposomes. Adv Drug Deliv Rev 1998;29:249–271.
278. Fresta M, Puglisi G. Colloidal drug delivery systems in antiinfective chemotherapy. Pandalai SG, editor. Recent Research Developments in Antimicrobial Agents and Chemotherapy Trivandrum-8. India: Research Signpost; 2000.
279. Fresta M, et al. Intracellular accumulation of ofloxacin-loaded liposomes in human synovial fibroblasts. Antimicrob Agents Chemother 1995;39:1372–1375.
280. Furneri PM, Fresta M, Puglisi G, Tempera G. Ofloxacin-loaded liposomes: their in vitro activity and effect on Bacterial drug accumulation. Antimicrob Agents Chemother 2000; 44:2458–2464.
281. Bakker-Woundenberg IAJM, Lokerse AF, Ten Kate MT, Storm G. Enhanced localization of liposomes with prolonged blood circulation time in infected lung tissue. Biochim Biophys Acta 1992;1138:318–326.
282. Duzgunes N, et al. Delivery of antiviral agents in liposomes. Methods Enzymol 2005;391:351–373.
283. Cudd A, et al. Specific interaction of CD4-bearing liposomes with HIV-infected cells. J Acquir Immune Defic Syndr 1990;3:109–114.
284. Lee JT, et al. Evaluation of cationic liposomes for delivery of diphtheria toxin A-chain gene to cells infected with bovine leukemia virus. J Vet Med Sci 1997;59:169–174.
285. Fielding RM, et al. Comparative pharmacokinetics of amphotericin B after administration of a novel colloidal delivery system, ABCD, and a conventional formulation to rats. Antimicrob Agents Chemother 1991;35:1208–1213.
286. Lopez-Berestein G, et al. Liposomal amphotericin B for the treatment of systemic fungal infections in patients with cancer: a preliminary study. J Infect Dis 1985;151: 704–710.
287. Proulx ME, et al. Treatment of visceral leishmaniasis with sterically stabilized liposomes containing camptothecin. Antimicrob Agents Chemother 2001;45:2623–2627.
288. Gregoriadis G. DNA vaccines: a role for liposomes. Curr Opin Mol Ther 1999;1:39–42.
289. Fresta M, Wehrli E, Puglisi G. Enhanced therapeutic effect of cytidine-5^{I}-diphosphate choline when associated with G_{M1} containing small liposomes as demonstrated in a rat ischemia model. Pharm Res 1995;12:1769–1774.
290. Fresta M, Puglisi G. Survival rate improvement in a rat ischemia model by long circulating liposomes containing CDP-choline. Life Sci 1997;61:1227–1235.
291. Fresta M, Puglisi G. Reduction of maturation phenomenon in cerebral ischemia with CDP-choline-loaded liposomes. Pharm Res 1999;16:1843–1849.
292. Mezei M, Gulusekharam V. Liposomes, a selective drug delivery system for the topical route of administration: Gel dosage form. J Pharm Pharmacol 1982;34:473–474.

293. Touitou E, et al. Liposomes as carriers for topical and transdermal delivery. J Pharm Sci 1994;83:1189–1203.
294. Mezei M, Gulusekharam V. Liposomes, a selective drug delivery system for the topical route of administration. Life Sci 1980;26:1473–1477.
295. Egbaria K, Weiner N. Liposomes as a topical drug delivery system. Adv Drug Deliv Rev 1990;5:287–300.
296. Foong WC, Harsanyi BB, Mezei M. Biodisposition and histological evaluation of topically applied retinoic acid in liposomal cream and gel dosage forms. Haning I, Pepeu G, editors. Phospholipids. New York: Plenum Press; 1990.
297. Touitou E, et al. Modulation of caffeine skin delivery by carrier design: liposomes versus permeation enhancers. Int J Pharm 1994;103:131–136.
298. Bouwstra JA, Honeywell-Nguyen PL. Skin structure and mode of action of vesicles. Adv Drug Deliv Rev 2002;54:41–55.
299. Fresta M, Puglisi G. Application of liposomes as potential cutaneous drug delivery. *In vitro* and *in vivo* investigation with radioactively labelled vesicles. J Drug Target 1996; 4:95–101.
300. Fresta M, Puglisi G. Corticosteroid dermal delivery with skin-lipid liposomes. J Control Rel 1997;44:141–151.
301. Touitou E, et al. Ethosomes–novel vesicular carriers for enhanced delivery: characterization and skin penetration properties. J Control Rel 2000;65:403–418.
302. Cevc G, Blume G. New, highly effcient formulation of diclofenac for the topical, transdermal administration in ultradeformable drug carriers, Transfersomes. Biochim Biophys Acta 2001;1514:191–205.
303. Gregoriadis G, Florence AT. Liposomes in drug delivery. Clinical, diagnostic and ophthalmic potential. Drugs 1993; 45:15–28.
304. Bouwstra JA, van Hal DA, Hofland HEJ, Junginger HE. Preparation and characterization of nonionic surfactant vesicles. Colloids Surface A 1997;123:71–80.
305. Uchegbu IF, Vyas PS. Non-ionic surfactant based vesicles (niosomes) in drug delivery. Int J Pharm 1998;172:33–70.
306. Carafa M, et al. Preparation and properties of new unilamellar non-ionic: ionic surfactant vesicles. Int J Pharm 1998;160:51–59.
307. Santucci E, et al. Vesicles from polysorbate-20 and cholesterol-α simple preparation and a characterisation. STP Pharm Sci 1996;6:29–32.
308. Gupta PN, et al. Non-invasive vaccine delivery in trasferosomes, niosomes and liposomes: a comparative study. Int J Pharm 2005;293:73–82.
309. Kirby C, Gregoriadis G. Dehydration-rehydration vesicles: a simple method for high yield drug entrapment in liposomes. Biotechnology 1984; 979–984.
310. Hofland HEJ, et al. Safety aspects of non-ionic surfactant vesicles-a toxicity study related to the physicochemical characteristics of non-ionic surfactants. J Pharm Pharmacol 1992;44:287–294.
311. Dimitrijevic D, et al. The effect of monomers and of micellar and vesicular forms of non-ionic surfactants (Solulan C24 and Solulan 16) on Caco-2 cell monolayers. J Pharm Pharmacol 1997;49:611–616.
312. Jain CP, Vyas SP. Lymphatic delivery of niosome encapsulated methotrexate. Pharmazie 1995;50:367–368.
313. Parthasarathi G, Udupa N, Umadevi P, Pillai GK. Niosome encapsulated of vincristine sulfate-improved anticancer activity with reduced toxicity in mice. J Drug Target 1994; 2:173–182.
314. Yoshioka T, Florence AT. Vesicle (niosome)-in-water-in-oil (v:w:o) emulsions-an in-vitro study. Int J Pharm 1994;108: 117–123.
315. Erdogan S, et al. In-vivo studies on iopromide radiopaque niosomes. STP Pharma Sci 1996;6:87–93.
316. Azmin MN, et al. The effect of non-ionic surfactant vesicle (niosome) entrapment on the absorption and distribution of methotrexate in mice. J Pharm Pharmacol 1985;37:237–242.
317. Chandraprakash KS, Udupa N, Umadevi P, Pillai GK. Effect of macrophage activation on plasma disposition of niosomal ^{3}H-Methotrexate in sarcoma-180 bearing mice. J Drug Target 1993;1:143–145.
318. Rogerson A, Cummings J, Willmott N, Florence AT. The distribution of doxorubicin in mice following administration in niosomes. J Pharm Pharmacol 1988;40:337–342.
319. Vanhal D, et al. Diffusion of estradiol from non-ionic surfactant vesicles through human stratum-corneum *in vitro*. STP Pharm Sci 1996;6:72–78.
320. Junginger HE, Hofland HEJ, Bouwstra JA. Liposomes and niosomes: interactions with human skin. Cosmet Toilet 1991; 106:45–50.
321. Saettone MF, et al. Non-ionic surfactant vesicles as ophthalmic carriers for cyclopentolate a preliminary evaluation. STP Pharm Sci 1996;6:94–98.
322. Touitou E. Compositions for Applying Active Substances to or through the skin, US Patent 5,540,934, 1996.
323. Touitou E. Compositions for Applying Active Substances to or through the skin, US Patent 5,716,638, 1997.
324. Touitou E. Compositions and Methods for Intracellular Delivery. PCT/IL02/00516, 2002.
325. Berner B. editors. Percutaneous Penetration Enhancers. Boca Raton (Fl): CRC Press; 1995.
326. Dayan N, Touitou E. Carriers for skin delivery of trihexyphenidyl HCl: ethosomes vs. liposomes. Biomaterials 2000; 21:1879–1885.
327. Touitou E, et al. Intracellular delivery mediated by an ethosomal carrier. Biomaterials 2001;22:3053–3059.
328. Godin B, Touitou E. Mechanism of bacitracin permeation enhancement through the skin and cellular membranes from an ethosomal carrier. J Control Rel 2004;94:365–379.
329. Godin B, Touitou E. Ethosomes: new prospects in transdermal delivery. Crit Rev Ther Drug Carrier Syst 2003;20:63–102.
330. Paolino D, et al. Ethosomes for skin delivery of ammonium glycyrrhizinate: *in vitro* percutaneous permeation through human skin and *in vivo* anti-inflammatory activity on human volunteers. J Control Rel 2005;106:99–110.
331. Horwitz E, et al. A clinical evaluation of a novel liposomal carrier for aciclovir in the topical treatment of recurrent herpes labialis. Oral Surg Oral Med Oral Pathol Oral Radiol. Endod 1999;87:700–705.
332. Touitou E, Godin B, Weiss C. Enhanced delivery of drugs into and across the skin by ethosomal carriers. Drug Dev Res 2000;50:406–415.
333. Lodzki M, et al. Cannabidiol—transdermal delivery and antiinflammatory effect in a murine model. J Control Rel 2003; 93:377–387.
334. Touitou E, Drug delivery across the skin. Expert Opin Biol Ther 2002;2:723–733.
335. Schreier H, Bouwstra J. Liposomes and niosomes as topical drug carriers: dermal and transdermal drug delivery. J Control Rel 1994;30:1–15.
336. Cevc G, Blume G, Schatzlein A. Transfersomes-mediated transepidermal delivery improves the regio-specificity and biological activity of corticosteroids *in vivo*. J Control Rel 1997;45:211–226.
337. Cevc G, et al. The skin: a pathway for systemic treatment with patches and lipid-based agent carriers. Adv Drug Deliv Rev 1996;18:349–378..
338. Cevc G, Blume G. New, highly effcient formulation of diclofenac for the topical, transdermal administration in ultradeformable drug carriers, Transfersomes. Biochim Biophys Acta 2001;1514:191–205.

339. Kim A, Lee EH, Choi SH, Kim CK. *In vitro* and *in vivo* transfection effciency of a novel ultradeformable cationic liposome. Biomaterials 2004;25:305–313.
340. Cevc G, et al. Ultraflexible vesicles, Transferosomes, have an extremely low pore penetration resistance and transport therapeutic amounts of insulin acrossthe intact mammalian skin. Biochim Biophys Acta 1998;1368:201–215.
341. Hildebrand A, et al. Solubilization of negatively charged DPPC/DPPG liposomes by bile salts. J Coll Interf Sci 2004; 279:559–571.
342. Cevc G, Transfersomes®-Innovative transdermal drug carriers. In: modified release drug delivery technology. New York: Marcel Dekker; 2002.
343. Cevc G, Blume G. Hydrocortisone and dexamethasone in ultradeformable drug carriers, Transfersomes®, have an increased in biological potency and reduced therapeutic dosages. Biochim Biophys Acta 2004;1663:61–73.
344. Cevc G. Transfersomes, liposomes and other lipid suspensions on the skin: permeation enhancement, vesicle penetration, and transdermal drug delivery. Crit Rev Ther Drug Carrier Syst 1996;13:257–388.
345. Vauthier C, et al. Poly(alkylcyanoacrylates) as biodegradable materials for biomedical applications. Adv Drug Deliv Rev 2003;55:519–548.
346. Couvreur P, et al. Nanocapsule technology: a review. Crit Rev Ther Drug Carrier Syst 2002;19:99–134.
347. Cui Z, Mumper RJ. Microparticles and nanoparticles as delivery systems for DNA vaccines. Crit Rev Ther Drug Carrier Syst 2003;20:103–137.
348. Kreuter J, editor. Nanoparticles, Colloidal Drug Delivery Systems. New York: Marcel Dekker; 1994.
349. Feng SS, Huang G. Effects of emulsifiers on the controlled release of paclitaxel (Taxol) from nanospheres of biodegradable polymers. J Control Rel 2001;71:53–69.
350. Li YP, et al. PEGylated polycyanoacrylate nanoparticles as tumor necrosis factor-alpha carriers. J Control Rel 2001;71: 287–296.
351. Carino GP, Jacob JS, Mathiowitz E. Nanosphere based oral insulin delivery. J Control Rel 2000;65:261–269.
352. Giannavola C, et al. Influence of preparation conditions on acyclovir-loaded poly-*d*,l-lactic acid nanospheres and effect of PEG-coating on ocular drug bioavailability. Pharm Res 2003;20:584–590.
353. Teixeira M, Alonso MJ, Pinto MM, Barbosa CM. Development and characterization of PLGA nanospheres and nanocapsules containing xanthone and 3-methoxyxanthone. Eur J Pharm Biopharm 2005;59:491–500.
354. Fresta M, et al. Preparation and characterization of polyethyl-2-cyanoacrylate nanocapsules containing antiepiletic drugs. Biomaterials 1996;17:751–758.
355. Cavallaro G, et al. Entrapment of β-lactames antibiotics on polyethylcyanoacrylate nanoparticles. Studies on the possible *in vivo* application of this colloidal delivery system. Int J Pharm 1994;111:31–41.
356. Harmia T, et al. Enhancement of the myotic response of rabbits with pilocarpine-loaded polybutylcyanoacrylate nanoparticles. Int J Pharm 1986;33:187–193.
357. Blanco MD, Alonso MJ. Development and characterization of protein-loaded poly(lactide-*co*-glycolide) nanospheres. Eur J Pharm Biopharm 1997;43:287–294.
358. Bala I, Hariharan S, Kumar MN. PLGA nanoparticles in drug delivery: the state of the art. Crit Rev Ther Drug Carrier Syst 2004;21:387–422.
359. Muller RH, Jacobs C, Kayser O. Nanosuspensions as particulate drug formulations in therapy. Rationale for development and what we can expect for the future. Adv Drug Deliv Rev 2001;47:3–19.
360. Ubrich N, et al. Oral evaluation in rabbits of cyclosporin-loaded Eudragit RS or RL nanoparticles. Int J Pharm 2005; 288:169–175.
361. Xu Z, et al. *In vitro* and *in vivo* evaluation of actively targetable nanoparticles for paclitaxel delivery. Int J Pharm 2005;288:361–368.
362. Lorenzo-Lamosa ML, et al. Design of microencapsulated chitosan microspheres for colonic drug delivery. J Control Rel 1998;52:109–118.
363. Potineni A, Lynn DM, Langer R, Amiji MM. Poly(ethylene oxide)-modified poly(beta-amino ester) nanoparticles as a pH-sensitive biodegradable system for paclitaxel delivery. J Control Rel 2003;86:223–234.
364. Shenoy DB, Amiji MM. Poly(ethylene oxide)-modified poly-(epsilon-caprolactone) nanoparticles for targeted delivery of tamoxifen in breast cancer. Int J Pharm 2005;293:261–270.
365. Konan-Kouakou YN, Boch R, Gurny R, Allemann E. *In vitro* and *in vivo* activities of verteporfin-loaded nanoparticles. J Control Rel 2005;103:83–91.
366. Yoo HS, Lee KH, Oh JE, Park TG. *In vitro* and *in vivo* antitumor activities of nanoparticles based on doxorubicin-PLGA conjugates. J Control Rel 2000;68:419–431.
367. Redhead HM, Davis SS, Illum L. Drug delivery in poly(lactide-co-glycolide) nanoparticles surface modified with poloxamer 407 and poloxamine 908: *in vitro* characterisation and *in vivo* evaluation. J Control Rel 2001;70:353–363.
368. Stolnik S, et al. Surface modification of poly(lactide-*co*-glycolide) nanospheres by biodegradable poly(lactide)-poly-(ethyleneglycol) copolymers. Pharm Res 1994;11:1800–1808.
369. Peracchia MT, et al. Stealth PEGylated polycyanoacrylate nanoparticles for intravenous administration and splenic targeting. J Control Rel 1999;60:121–128.
370. De Jaeghere F, et al. Cellular uptake of PEO surface-modified nanoparticles: evaluation of nanoparticles made of PLA:PEO diblock and triblock copolymers. J Drug Target 2000;8:143–153.
371. Mitra S, Gaur U, Ghosh PC, Maitra AN. Tumor targeted delivery of encapsulated dextran-doxorubicin coniugate using chitosan nanoparticles as carrier. J Control Rel 2001;74:317–323.
372. Nobs L, Buchegger F, Gurny R, Allemann E. Surface modification of poly(lactic acid) nanoparticles by covalent attachment of thiol groups by means of three methods. Int J Pharm 2003;250:327–337.
373. Fresta M, et al. Pefloxacine mesilate- and Ofloxacin-loaded polyethylcyanoacrylate nanoparticles. Characterization of the colloidal drug carrier formulation. J Pharm Sci 1995;84:895–902.
374. Schäfer V, et al. Phagocytosis of nanoparticles by human immunodeficiency virus (HIV)-infected macrophages: a possibility for antiviral drug targeting. Pharm Res 1992;9:541–546.
375. Lherm C, et al. Unloaded polyisobutylcyanoacrylate nanoparticles: efficiency against bloodstream trypanosomes. J Pharm Pharmacol 1987;39:650–652.
376. Losa C, et al. Improvement of ocular penetration of amikacin sulphate by association to poly(butylcyanoacrylate) nanoparticles. J Pharm Pharmacol 1991;43:548–552.
377. Fresta M, et al. Ocular tolerability and in vivo bioavailability of PEG-coated polyethyl-2-cyanoacrylate nanosphere-encapsulated acyclovir. J Pharm Sci 2001;90:288–297.
378. Gomes Dos Santos AL, Bochot A, Fattal E. Intraocular delivery of oligonucleotides. Curr Pharm Biotechnol 2005; 6:7–15.
379. Agnihotri SA, Mallikarjuna NN, Aminabhavi TM. Recent advances on chitosan-based micro- and nanoparticles in drug delivery. J Control Rel 2004;100:5–28.
380. van der Lubben IM, et al. Chitosan microparticles for mucosal vaccination against diphtheria: oral and nasal efficacy studies in mice. Vaccine 2003;21:1400–1408.

381. Muller RH, Mader K, Gohla S. Solid lipid nanoparticles (SLN) for controlled drug delivery—a review of the state of the art. Eur J Pharm Biopharm 2000;50:161–177.
382. Jenning V, Gohla SH. Encapsulation of retinoids in solid lipid nanoparticles (SLN). J Microencapsul 2001;18:149–158.
383. Westesen K. Novel lipid-based colloidal dispersions as potential drug administration systems-expectations and reality. Coll Polym Sci 2000;278:609–618.
384. Cavalli R, Gasco MR, Barresi AA, Rovero G. Evaporative drying of aqueous dispersions of solid lipid nanoparticles. Drug Dev Ind Pharm 2001;27:919–924.
385. Hu L, Tang X, Cui F. Solid lipid nanoparticles (SLNs) to improve oral bioavailability of poorly soluble drugs. J Pharm Pharmacol 2004;56:1527–1535.
386. Wissing SA, Muller RH. Cosmetic applications for solid lipid nanoparticles (SLN). Int J Pharm 2003;254:65–68.
387. Wissing SA, Muller RH. Solid lipid nanoparticles as carrier for sunscreens: *in vitro* release and *in vivo* skin penetration. J Control Rel 2002;81:225–233.
388. Videira MA, et al. Lymphatic uptake of pulmonary delivered radiolabelled solid lipid nanoparticles. J Drug Target 2002;10:607–613.
389. Zara GP, et al. Intravenous administration to rabbits of non-stealth and stealth doxorubicin-loaded solid lipid nanoparticles at increasing concentration of stealth agent: pharmacokinetics and distribution of doxorubicin in brain and other tissues. J Drug Target 2002;10:327–335.
390. Muller RH, Keck CM. Drug delivery to the brain-realization by novel drug carriers. J Nanosci Nanotechnol 2004;4:471–483.
391. Jenning V, Thünemann AF, Gohla SH. Characterization of a novel solid lipid nanoparticle carrier system based on binary mixtures of liquid and solid lipids. Int J Pharm 2000;199: 166–177.
392. Muller RH, Radtke M, Wissing SA. Solid lipid nanoparticles (SLN) and nanostructured lipid carriers (NLC) in cosmetic and dermatological preparations. Adv Drug Deliv Rev 2002; 54:131–155.
393. Bummer PM. Physical chemical considerations of lipid-based oral drug delivery-solid lipid nanoparticles. Crit Rev Ther Drug Carrier Syst 2004;21:1–20.
394. Zhang Q, et al. Studies on the cyclosporin A loaded stearic acid nanoparticles. Int J Pharm 2000;200:153–159.
395. Cavalli R, Caputo O, Gasco MR. Preparation and characterization of solid lipid nanospheres containing paclitaxel. Eur J Pharm Sci 2000;10:305–309.
396. Olbrich C, Geßner A, Kayser O, Müller RH. Lipid-drug conjugate (LDC) nanoparticles as novel carrier system for the hydrophilic antitrypanosomal drug diminazenediaceturate. J Drug Target 2002;10:387–396.

See also DRUG INFUSION SYSTEMS; PHARMACOKINETICS AND PHARMACODYNAMICS.

DRUG INFUSION SYSTEMS

SELAHATTIN OZCELIK
Texas A&M University
Kingsville, Texas

INTRODUCTION

An infusion system can be described as the process of delivering fluids and medications in solution to patients by way of an infusion device. Intravenous route is generally used for drug delivery; however, subcutaneous, epidural, and enteral routes are also used for special drug administration (1). Administration of medication into the patient by way of some kind of drug infusion device provides the desired level of medication in the patient and allows direct control over pharmacological variables, such as onset of drug effects and peak serum drug concentrations (1). This type of drug administration has been the choice especially for specific conditions, including the use of antibiotics for severe infection, chemotherapy for malignant conditions, cardiac medication in critical cases, and analgesics for relief of severe pain.

When curing all common medical disorders that justify therapeutic intervention, pharmacological therapy is the preferred and effective method of treatment. Device-based drug delivery systems for the administration of effective pharmacological therapy can be grouped as injection–infusion, transdermal patch-based, and inhalation systems (2). Among the established methods of injection and infusion systems are the needle and jet injection, intravascular, intraspinal, intraoperative site, and intraperitoneal–transperitoneal infusion systems. Major applications of drug infusion are anesthesia delivery, antibiotic–antiviral therapy, nutritional support, pain management, cardiovascular disease therapy, chemotherapy, diabetes management, hydration therapy, bone marrow and organ transplant support therapy, and transfusion therapy (2).

The use of powered infusion devices has grown enormously in the last two decades. Infusion pumps together with an appropriate administration set provide an accurate flow of fluids over a prescribed time period. The simplest devices are the gravity controllers, in which the flow of liquid under the force of gravity is regulated by clamping action. More complex infusion systems include a positive pumping action for infusion (3). Volumetric pumps posses a linear peristaltic pumping mechanism. Syringe pumps work by pushing the plunger of a disposable syringe along at a predetermined rate. The type of pump used depends on the required volume and speed of infusion (3).

Medication errors are a major concern of healthcare professionals and medical institutions and have been reported to contribute to between 7000 and 140,000 deaths only in the United States each year (4–9). The impact of medication errors was found to be more severe in pediatric patients (4). There are a wide variety of reasons for medication errors. Reports suggest that the most significant factor is the user errors; however, the contribution of device-related problems to medical errors cannot be underestimated. Many reports of incidents have also been received involving infusion pumps. These incidents are primarily due to over infusions and may result in patient harm or death (3). In practice, some of the common problems in drug infusion systems originate from syringe pumps. Most of the patient morbidity and mortality, being the most significant among all the problems, has happened when using syringe pumps. Another common problem with drug delivery is venous air embolisms, which can be caused by air ingress due to improper drug delivery technique, damaged equipment and tubing, leaking or loose tubing connectors, or failure to stop delivery prior to complete evacuation of IV bags. Venous air embolisms have been

observed in central venous cannulation and pressurized intravenous infusion systems (3).

Today it has been proved that technology is essential in reducing risk in medication delivery. Computerized physician order entry (CPOE) and bar code applications for drug administration are such technologies that are capable of reducing medication errors. Unfortunately, most hospitals have not yet implemented these systems; therefore, many errors that otherwise might be eliminated continue to put patients at risk (10). Computerized intravenous (IV) infusion devices, so-called smart pumps include software that incorporates dosage limits established by the medical institution, warnings when dosage limits are exceeded, configurable settings by patient type, and access to transaction data for quality improvement efforts. Such systems make it possible to provide an additional verification at the point of care to help prevent IV medication errors (11). The Institute for Safe Medication Practices and the Emergency Care Research Institute recognize safety systems for IV medication as vital to reducing medication-related errors (10,12). A couple of examples utilizing a new technology is MEDLEY from Alaris Medical Systems and COLLEAGUE CX by Baxter Healthcare Corporation. These infusion systems allow hospitals to enter various drug infusion protocols into a drug library with predefined dose limits. For example, if a dose is outside the programmed range or clinical parameters, the pump halts or informs the physician by providing an alarm. Some pumps are even capable of integrating patient monitoring and other parameters, such as patient's age or clinical condition. More and more manufacturers are bringing similar devices to market (10).

The aim of this section is to provide a review on drug infusion systems, basic operational principles of pumps used in such systems, new infusion devices that are being developed, and recent developments for the control of infusion devices. With this goal in mind, this section is organized as follows: The section Common Infusion Systems presents most commonly used infusion systems and their operational principles. In the section, New Developments in Drug Infusion Systems, smarter, smaller, reliable, and cost effective new infusion devices, which are currently under development, are reviewed. Since the performance of any automated system highly depends on its controller structure, most recent research work on the control of drug infusion devices is reviewed in the section Recent Advancements in Controller Design for Drug Infusion Systems.

COMMON INFUSION SYSTEMS

Any drug infusion system requires some kind of control unit in which necessary parameters are monitored continuously so that the drug is delivered to the patient in a desired manner. Infusion devices range from very simple mechanical devices based on elastic containers, springs, and flow restrictors to sophisticated microprocessor controlled pumps. The choice of an infusion device depends on both the type of therapy to be applied and patient characteristics. Some IV infusions can safely and effectively be delivered via gravity drip systems, while others require sophisticated microprocessor controlled pumps for more precise control, positive pressure, and greater flow rate range. Traditionally, these devices are used in hospitals for the controlled delivery of drugs and fluids. However, as these devices become smarter with the use of new technology, more and more patients are using them for home therapy.

Therefore, the use of infusion pumps is increasing in the community. To ensure that a patient receives the correct dose, the appropriate infusion device should be chosen for the drug. Syringe pumps are commonly used at low rates of infusion, but may not be suitable for drugs that require constant blood levels (13). It is required that any infusion system be able to reliably deliver the prescribed drug dose–volume to the patient, at pressures that overcome all baseline and intermittent resistance, while causing no harm to the patient. Additive resistances, such as the small bore and kinking potential of connecting tubing, cannula, needles, and patient vessels, make infusion flow difficult. Filters, viscous solutions, and syringe stiction can also adversely affect the infusion flow. Therefore, infusion pumps are required to overcome these resistances and accurately deliver prescribed drugs to patients. These pumps must be capable of delivering infusions at pressures of between 100 and 500 mmHg (2–10 psi or 13.79–68.95 kPa) (13). Ideally, pumps should also reliably detect the infusion pressure and the presence of air in the line close to the patient vessel being infused. Table 1 provides pressure ranges for IV pump pressure settings. Infusion devices can be classified according to their power source as gravity controllers and infusion pumps.

Table 1. Pressure Ranges for IV Pump Pressure Setting[a,b]

Pressure, mmHg	Example	Pressure, psi
2–20	Central venous pressure range	0–0.4
10–30	Peripheral venous pressure range	0.2–0.6
100	Extravasation risk	2
100–150	Systolic arterial pressure range	2–3
75	Gravity pressure of fluid 100 cm above cannulation site	1.5
500	Highest probable pressure required by an infusion pump	10
1000	Maximum modern–Ambulatory pump occlusion pressure setting	20
3000	Common max. pressures form older peristaltic pumps	60

[a]See Ref. 14.
[b]1 mmHg = 1 psi = 6.89 kPa.

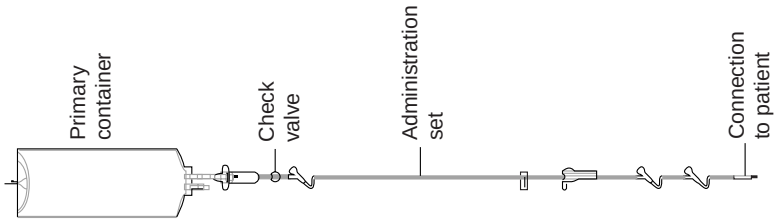


Figure 1. Common gravity drip infusion device.

Gravity Drip Systems

The simplest infusion device is the gravity drip system in which a bag or bottle is hung on a hook of a pole sufficiently high from the level of the patient. Figure 1 shows a typical gravity drip infusion system. The fluid flows by gravitational force down the line and into the catheter. They are quite suitable for lower risk applications, including fluid replacement therapy, provided that the required flow rate is achieved by the delivery pressure of the device (4). Gravity controllers are based on gravity to provide the infusion pressure. Therefore, in order to achieve the desired flow rate, the fluid container is placed sufficiently high above the patient's heart. A drop sensor monitoring the drip rate is attached to the drip chamber of the administration set. The rate of flow in a simple gravity drip system is controlled by a special clamp or valve on the line that can be manually adjusted to permit the prescribed amount of fluid to flow through (usually described in drops per minute).

These devices range in complexity and ease of operation from roller and slide clamps to more sophisticated rotating valves. Compared with slide and roller clamps, rotating valves are less awkward to manipulate and provide a more consistent flow rate. However, even the most sophisticated manual drip valve cannot offer precise flow control, due the viscosity of the solution being infused. Another factor to be taken into account is the second flow control caused by the size of the needle at the end of the line through which the fluid flows into the catheter. The smaller the needle is the slower the maximum rate of flow into the body (15). These types of devices are quite effective in controlling overinfusion; however, control of underinfusion would not be satisfactory due to increased resistance to flow. One way to avoid this problem is to use a drip rate controller with a visible flow status system (14). The pressure available from a bag of saline is equal to the height that the bag is above the patient's heart. Drip rate controller is a type of gravity controller, in which the desired flow rate is set in drops per minute and controlled by occlusion valves powered by electricity. All models of drip controllers have a drop sensor. More advanced models incorporate a flow status system, which gives a visual indication of resistance to flow (4). The required number of drops delivered by gravity controllers is controlled by the drop counting mechanism that is quite accurate. However, the actual amount of volume delivered to patients may vary because of error involved when converting the number of drops to milliliters (mL). Conversion chart values for drops $\cdot$ mL^{-1} are approximate and a small error made for each drop may result in a large difference in the entire volume of drug delivered.

The volume of fluid in a drop depends on several factors, some of which are the fluid's composition, temperature and surface tension, the drip rate set, the size, shape, and condition of the drop-forming orifice (3). Expected nominal volume for a drop is 20 drops $\cdot$ mL^{-1}. This nominal rate can easily be achieved for most simple aqueous solutions of electrolytes, lactates, or dilute sugars. However, due to the viscosity characteristics of parenteral nutrition mixtures, fat soluble vitamins and solutions containing alcohol, the drop volume will be lower than nominal resulting in longer infusion time (3). Naturally, with all fluids, the drop volume decreases as the delivery rate increases. These variations are acceptable for the majority of infusions. However, if the volumetric accuracy is critical, then an infusion pump must be used (3).

There is a standard formula for calculating the flow rate on any type of IV tubing as follows:

$$(V \times df)/t = \text{drops min}^{-1}$$

t = time to be infused (in min); V = volume of solution to be infused; df = drop factor of solution set (drops mL^{-1}); (mL $\times$ drop factor)(min^{-1}) = drops min^{-1}.

If the result of the calculation includes a decimal point, round-off to the nearest.

Electronic controllers provide better accuracy for the regulation of flow by controlling uneven or runaway flow of fluid in a gravity drip system. These electronic devices are equipped with a drop sensor to monitor flow rate and can detect infiltrations and mal-positioning of the catheter or IV tubing by measuring backflow. An alarm sounds when flow rate is altered or when backflow is detected.

The gravity drip is conceptually simple, inexpensive, and requires less equipment than most other infusion systems. In the home setting, however, it has some limitations. First, it is difficult to maintain a constant infusion rate in a gravity drip system due to factors, such as the decreasing volume of fluid in the bag (i.e., the infusion rate will decrease as the bag empties) and changes in the shape of the tubing around the clamp. Consequently, a gravity system may provide insufficient flow control for drugs that require a very slow, very precise, or very long infusion time. Second, errors in using the gravity drip that remain unnoticed can result in serious complications (15). In addition, a gravity drip system may be an inappropriate choice for certain patients due to functional limitations of the patients or their caregivers. Because the IV bag is suspended well above the catheter site in this system, patients with decreased mobility may have difficulty changing the bag. Ambulatory patients on continuous infusion may also find gravity drip frustrating because the system is not easily portable. Despite the drawbacks of this traditional method of IV administration, it does maintain some important functional advantages over more expensive electronic infusion devices discussed below. Because the drugs are

forced into the vein under the pressure of gravity alone, there may be less irritation at the catheter site, especially peripheral catheter sites. Gravity drip systems may also be preferred for patients who are confused by and resistant to learning how to use more complex, computerized drug delivery systems (15).

Infusion Pumps

An electronically controlled device that could deliver constant and precise amounts of fluid over a specified time period was a major technological advance in infusion therapy. Although many therapies can be delivered safely and effectively via gravity drip systems, others require the highly precise and constant flow rate offered by electronic infusion devices (15). For example, intraarterial infusions usually require positive pressure pumps because the back pressure is higher in arteries than in veins (15). Volumetric or syringe pumps are the most common. Other methods include elastomeric, pneumatic, implantable, clockwork, or spring (3). They are used to accurately administer intravascular drugs, fluids, whole blood, and blood products. These pumps can administer up to 2000 mL of fluid (normally from a bag or bottle) at flow rates of 0.1–2000 $mL \cdot h^{-1}$.

Volumetric Pumps. Most volumetric pumps will perform satisfactorily at rates as low as $5\,mL \cdot h^{-1}$. However, these pumps are generally not used for delivering drugs at rates $<1\,mL \cdot h^{-1}$, even though the device can be set to such low rates. The rate is in milliliters per hour ($mL \cdot h^{-1}$) or micrograms per kilogram per hour ($\mu g \cdot kg^{-1} \cdot h^{-1}$) (3).

Most volumetric pumps have the feature of automatic alarm and shut down in case air enters the system, an occlusion is detected, or the reservoir is empty. The device controls the total volume to be infused and provides digital read-out of volume infused. Some of the other features include automatic switching to keep the vein open (KVO) rate at the end of infusion; switch to internal battery operation automatically if the mains supply fails; micro and macro delivery modes; computer interface; operator call alarm; a drop sensor—used for monitoring and alarm purposes (e.g., as an empty container) rather than as a control of the delivery rate; primary and secondary infusion capability; technical memory log for incident analysis. Features, such as air-in-line detection or a mechanism that cannot pump air and comprehensive alarm systems, make IV infusion much safer (3).

Most infusion pumps work by peristaltic action, which is achieved by alternately squeezing and releasing the tube containing the fluid to force the fluid through at a predetermined rate. There are two types of volumetric pumps: peristaltic and dedicated cassette. Peristaltic mechanisms can be further classified as linear peristaltic and rotary peristaltic. Both mechanisms consist of fingers, cams, or rollers that pinch off a section of the set. In linear peristaltic mechanisms as seen in Fig. 2, cams are located on a camshaft. Required volume is delivered to the patient by pinching off each section as the shaft rotates. These mechanisms are commonly used. In rotary peristaltic mechanisms, as seen in Fig. 3, rollers are placed on a

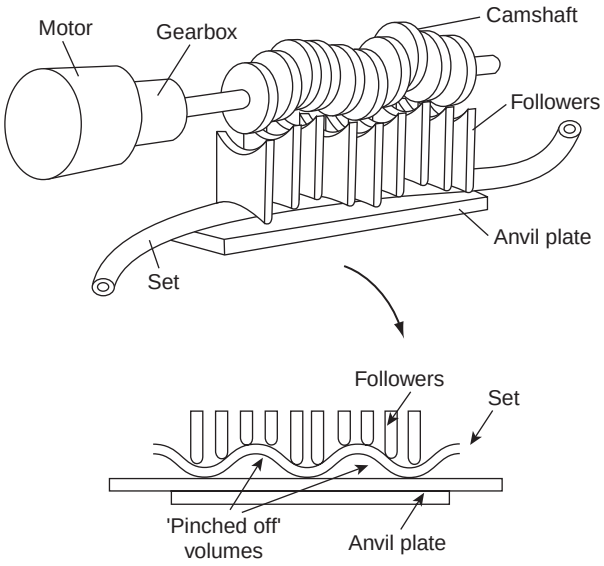


Figure 2. Schematic of a linear peristaltic pump (3) © CROWN COPYRIGHT.

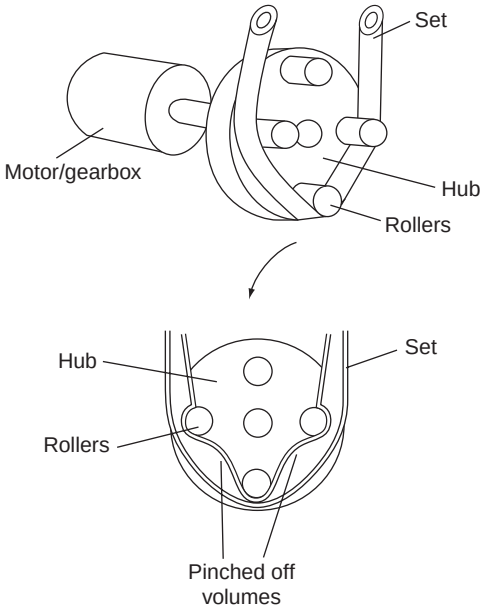


Figure 3. Schematic of a rotary peristaltic pump (3) © CROWN COPYRIGHT.

hub and as it rotates the volume in each pinched off section is delivered to the patient. The volume delivered varies according to the size of the cams, rollers, the tube, and the speed at which they rotate. These mechanisms are usually designed for a particular administration set (3).

Another mechanism used in infusion pumps is the dedicated cassette mechanism. Commonly, these types of pumps consist of a cassette body in which a valve and cylinder are placed, a piston, valve actuator, and a crank mechanism. This type of pump is depicted in Fig. 4. Drug is sucked to the cylinder from a bag or a container as the piston moves down and it is pumped to the patient through

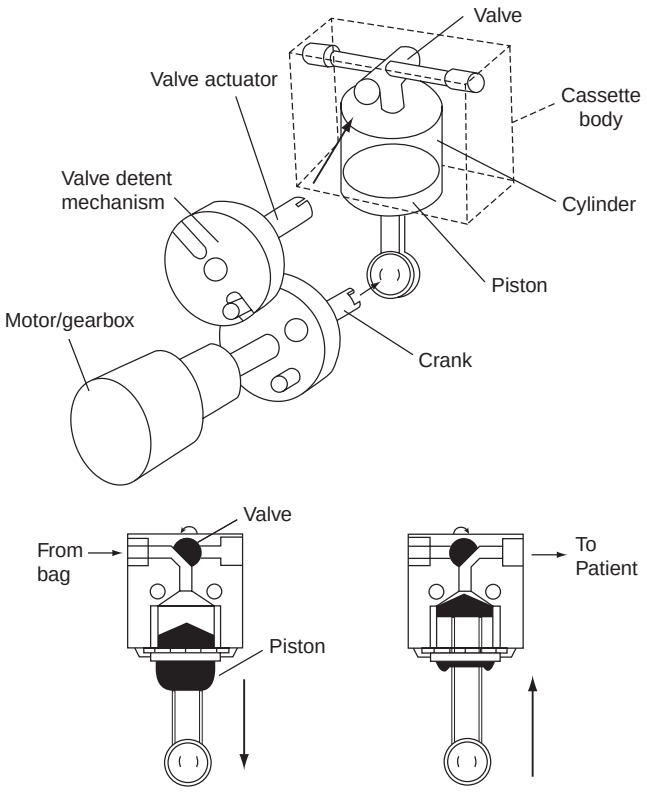


Figure 4. Dedicated cassette set (3) © CROWN COPYRIGHT.

a valve as the piston moves up (3). Although volumetric pumps can develop high pressures, they generally have a preset default value. In determining what pressure level is to be set, one needs to determine the factors of pressure raisers and calculate the needed pressure. However, it is important that the occlusion pressure should be set to the lowest possible value in order to observe early warning of occlusions.

A specific type of infusion set is required when using volumetric infusion pumps in order to achieve satisfactory drug delivery and to detect occlusion pressure. If an infusion set other than the required one is considered to be used, then extra care must be given when configuring the pump for that infusion set. Although using incorrect sets might seem to operate satisfactorily, this may be misleading and the actual performance and accuracy of drug infusion would be far from the desired level. This would lead to severe consequences. Air-in-line detectors use ultrasonic or optics for detecting air bubbles in the line. Air-in-line and occlusion detectors are designed for use with a particular infusion set. Therefore, these detectors may not be working properly if an incorrect set is used. Some other unwanted results are underinfusion due to very small inner-diameter of tube; overinfusion due to tubing material that is not flexible enough; and wear or rupture of tube from pumping action due to tubing material that is not strong enough. It is therefore important to use recommended sets for infusion. Specifications for testing of pumps at maximum flow rates are currently not given by the international standard for infusion devices. Therefore some fall-off in performance at high flow rates should be expected (3).

Most infusion pumps used today are modern, sophisticated versions of one of these two types of pumps. With the development of small, portable pumps with specialized uses for particular types of therapies and adaptations, these pumps are being used commonly by nonprofessionals as part of home therapy. Because computerized pumps can deliver medication at a wide range of dose frequencies and intensities, they broaden the scope of therapies that can be safely and effectively administered at home.

Pumps specifically for the infusion of narcotics to treat cancer-related pain, for example, may have adaptations that provide a low level of ongoing infusion, but also permit patients to dose themselves with bursts of medication when pain becomes intense, up to a preprogrammed number of such extra doses per day. Other pumps, designed for the volume of fluid typical of most antibiotic therapy, can be preprogrammed to deliver infusions at standard intervals (e.g., four times per day), thus enabling patients to sleep undisturbed while receiving therapy. Pumps used for long-term IV nutrition administration, on the other hand, may be designed to administer the large volume of fluid required for the overnight infusions typical of patients receiving this therapy. Infusion pumps currently available range in complexity and sophistication. These pumps can range from very simple, single-medication stationary infusion pumps to fully programmable, ambulatory pumps.

Sophisticated pumps can deliver multiple medications and are equipped with a variety of alarms, bells, and other warning mechanisms. While stationary pumps may be appropriate for patients who are bedridden or whose medications are delivered over shorter periods of time, ambulatory pumps provide greater independence for patients on continuous, frequent, or long-term therapy regimens. Many pumps also have automatic piggyback mechanisms that control secondary infusions at an independent rate, decreasing the nursing time required for multiple infusions (15). Besides these benefits and advancements, infusion pumps do have certain disadvantages. If patients, caregivers, or even health professionals find the level of sophistication of these pumps confusing, the patients' safety could be jeopardized through misuse of equipment. Many patients, and the nurses who instruct and care for them, might prefer simpler models that are easier to operate. Even many hospital nurses are unfamiliar with or unaware of sophisticated features of pumps they use on a regular basis. Highly sophisticated pumps cost more and often require considerably more training for both the health professional and the patient than simpler models. New types of electronic infusion pumps are constantly evolving, widening the menu from which providers must choose and from patients and health professionals must learn to operate.

Syringe Pumps. In this type of pump, drug is pumped forward in the tubing by a syringe-type pushing action. Schematic drawing of a syringe pump is depicted in Fig. 5. The syringe is placed in a housing of the pump, while syringe plunger is attached to a moving carriage. The carriage is attached to the lead screw through a nut,

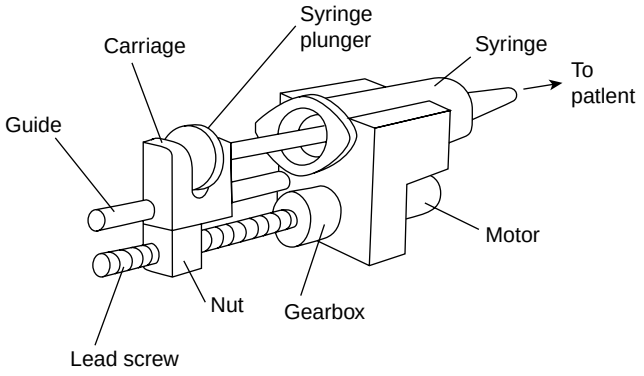


Figure 5. Typical syringe pump (3) © CROWN COPYRIGHT.

and the lead screw is attached to the motor through a gearbox. As the motor rotates, lead screw forces the carriage to slide on the guide, resulting syringe plunger to move forward in the syringe. This forward action delivers the drug to patient and empties the syringe. Controlled rate for the delivery of a drug is ensured by controlling the motor speed and rotation. This controlled rate may be in steps or continuous.

Advanced syringe pumps permit the simultaneous administration of several different therapies at different intervals, with dosages and administration regimens preprogrammed on a microchip that fits in the back of the pump. Syringe pumps can deliver small volumes of drugs at low flow rates. Single-use syringes are inexpensive and mass manufactured items, and are not meant to be highly accurate. When used in a syringe pump at low plunger speeds, the friction between the syringe plunger and the barrel causes a jerking effect and the fluid is delivered as a series of small boluses. The fit between the plunger and the barrel may vary from batch to batch and, consequently, the jerking effect may also vary. This problem is commonly known as stiction. In general, the bigger the syringe and the lower the flow rate, the more pronounced the stiction. Stiction may not be a problem with drugs having a long half-life or that do not require steady blood levels in the short term, such as heparin or insulin. In contrast, the delivery of powerful drugs with short half-life, like catecholamines, at rates under $5\,\mathrm{mL}\cdot\mathrm{h}^{-1}$ from large syringes (>30 mL) is not recommended.

Some currently available peristaltic pumps provide reasonably smooth flows at low delivery rates and should be considered as an alternative. Occasionally, the dimensions of a particular model of disposable syringe may be changed by the manufacturer. As a safe practice, only the syringe recommended by the manufacturer should be used with a syringe pump. These pumps are suitable for lower volume and low flow rate infusions. It is important to note that the actual drug delivered at the beginning of infusion process may be considerably less than the preset value. Due to the backlash, especially at low flow rates, it takes some time for the flow rate to reach steady-state regime.

Syringe pumps vary according to their functionality and so do their features. More advanced and expensive models have many features, including delivery pressure displays and in-line pressure monitoring. In most recent advanced pumps, one can set occlusion alarm pressures to very low values. This feature helps patients to prevent hazards due to occlusions by the alarm signal from the system in shorter times. These advanced pumps may also prevent the delivery pressure rising to unwanted high values. Since these devices are powered externally, placing them approximately at the patient's level will suffice for the pump to work satisfactorily. In fact, if the pump is placed well above patient's level, some draining could result.

Implantable Pumps. Some therapies that require very small drug dosages can be administered by way of totally implantable pumps. Insulin delivery, continuous epidural morphine administration for chronic pain management, and continuous venous antineoplastic therapy infusion for liver cancer patients are some examples where implantable pumps are used. The only service directly related to infusion therapy for these devices is refilling of the pump's reservoir, which may be done weekly or even less frequently in a medical outpatient or home setting (15).

Patient Controlled Analgesia (PCA) Pumps. These pumps are designed specifically for use in PCA. Unlike a general-purpose infusion pump, these pumps allow the patients to deliver the drug on their own by operating a switch or pressure pad connected by a cord to the pump. It is important that free-flow is prevented. These pumps can be connected to a computer or printer and have a memory, where data in terms of usage is stored. This feature allows the clinician to review when, how often, and how much of drug infused by the patient. The PCA pumps are typically syringe pumps, since the required drug to be infused can usually be supplied in a single-use syringe. Some PCA pumps are based on volumetric designs, in which a battery powered volumetric pump has a disposable internal fluid reservoir. The PCA pumps can be disposable (pneumatic and elastomeric) or nondisposable (3). The PCA pumps can be programmed by clinical staff in different ways. Options include loading dose, continuous infusion (basal rate), continuous infusion with bolus on demand, bolus on demand only, with choice of units (mL or $\mu\mathrm{g}\cdot\mathrm{mL}^{-1}$, etc.) and variable lockout time, drug concentration. Once programmed, a key or software code is needed to access control of the pump. In some cases, patients are given limited access in order to change some parameters.

Elastomeric Infusers. Elastomeric infusers are devices that can be used as substitutes for infusion pumps. These infusers consist of disposable containers with inner-elastic bladders that can be filled with the medication. The devices are sold empty and are filled by the pharmacist through a port at the top of the bladder. The drug flows through an opening at the base of the bladder membrane and into the tube leading to the patient. The force of the flow, and thus the rate of infusion, is determined by the elasticity of the bladder and the concentration of the drug, regardless of whether the bladder is above, below, or on level with the IV site. Different drugs and dosages require devices of differing size and bladder membrane composition. Most devices currently on the market are designed for either antibiotic or antineoplastic therapy administration. They can be used

for IV, intraarterial, and subcutaneous administration of drugs.

A patient on a twice-a-day regimen of home IV antibiotics would use two infusers per day, while a patient on continuous antineoplastic therapy might use a single device for several days at a time. Some devices allow patient-controlled administration of bolus doses above and beyond the continuous infusion rate. A disadvantage to the use of these devices for patient-controlled analgesia is the lack of a memory function that can record the frequency of patient-requested bolus doses, like that found in some electronic infusion pumps. Bladder devices are also not appropriate for multiple drug regimens. According to one home infusion provider, the availability of disposable elastomeric infusion devices has increased the feasibility of home-based care for disabled elderly patients. Like sophisticated electronic infusion pumps, these devices can deliver a precise dose over a specific period of time. However, because they are self-contained and much simpler to operate, they may be less confusing for patients who are uncomfortable with high tech equipment. The patient or caregiver need only hook the device to the catheter at dosing time and disconnect and dispose of it when the dose has been completed.

Anesthesia Pumps. These are also syringe-type pumps designed particularly for anesthesia infusion. Operating of these pumps is limited to theaters and high dependency areas. It is possible that the rate and other functions can be adjusted during infusion. Their flow rates are normally much higher than the typical syringe pumps rate. It therefore allows quick delivery in a single operation. These pumps can be interfaced with a computer and have built-in drug libraries. They are embedded with a drug-specific smart card and can be programmed for drug concentration and patient's body weight. The pump is automatically configured for the drug being infused. If the pump is to be used for other applications, automatically built-in features for specific application must be disabled.

Ambulatory Pumps. These pumps are designed so that patients can continue their drug therapy away from the hospital. These pumps allow patients to continue their normal life while treatment is being given. For ease of use and carry they are light and small in size, and are powered by battery. Alarming features of these pumps are not fully provided due to limitation in their size; therefore, their use in therapies in which precise flow is required for critical drugs is not recommended. The main mechanism used in these pumps is the syringe or cassette type.

Therapies that can be administered by ambulatory pumps include analgesia, continuous and PCA, antibiotic or antiviral infusions, chemotherapy, and hormone delivery. Back pressure, temperature of the flow-limiting element, temperature, and viscosity of the fluid are such factors that determine the accuracy of ambulatory pumps. One type of ambulatory pump manufactured by Baxter is given in Fig. 6. Depending on the pumping mechanism used, flow rates can range between 0.01 and 1000 $mL \cdot h^{-1}$. Different models have different features. Flow rates can be set in millimeters per hour or day, milliliters per hour or day. They can also be programmed for different delivery modes (3). Ambulatory pumps are generally powered by electricity. Accuracy and alarming features will be limited if the pump is not powered by electricity. The pumping mechanism is generally the same mechanism used in volumetric and syringe pumps.

Some ambulatory pumps are reusable. They consist of a syringe that is operated by pressurized gas, usually carbon dioxide or a precompressed spring. In the case of a pressurized-gas-operated system, the force generated by the pressurized gas pushes the syringe plunger. As the syringe plunger moves forward, it infuses the drug to the patient. The infusion rate is determined by the pressure of the gas as well as the rate at which pressurized gas is released. When the infusion is completed, syringe and gas cartridges are thrown away and the rest of the device is kept for future use. Infusers and bolus-only analgesia devices controlled by the patient are of disposable devices. They

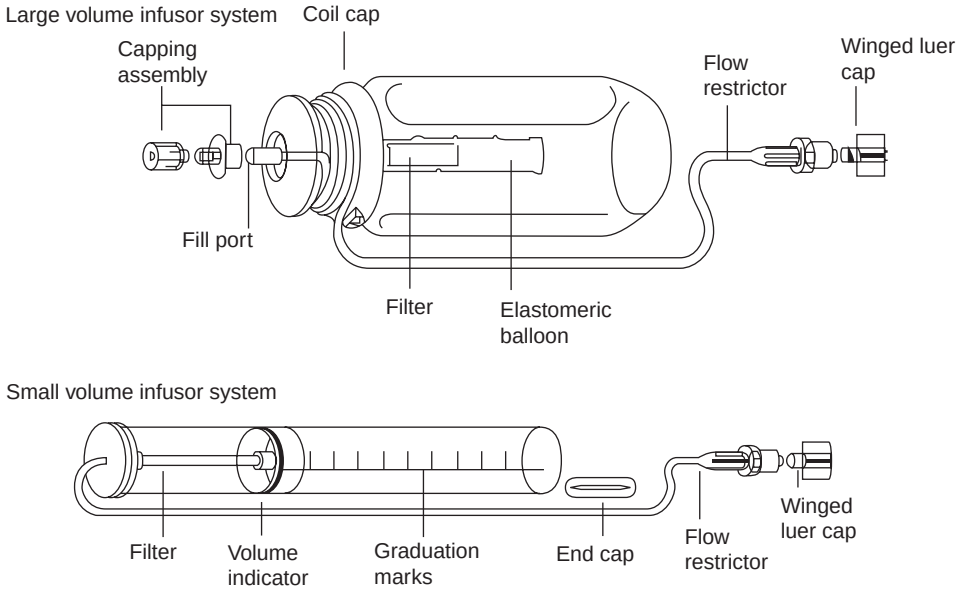


Figure 6. Ambulatory pumps manufactured by Baxter.

Figure 7. Signature Edition Gold infusion system by ALARIS. A range of infusion programs can be selected to save nursing time and meet sophisticated administration requirements including: Loading Dose, Multi-Dose and Multi-Step.

consist of a calibrated bolus chamber that is filled from an elastomeric reservoir or syringe by a capillary tube (3).

There are a number of companies that manufacture a variety of drug infusion systems. Some of the state-of-the-art products are given in Fig. 7–11. Abbott Laboratories' hospital products business (now Hospira), introduced the Plum A+ IV drug delivery medication management system. This system is an innovative infusion system for electronic control of intravenous medication administration. It is used for standard, piggyback, or concurrent delivery and are suitable for a wide range of medical–surgical and critical care applications

Figure 8. The Medley medication Safety System by ALARIS is a modular point-of-care computer that integrates infusion, patient monitoring and clinical best practice guidelines in a single platform for optimal outcomes.

NEW DEVELOPMENTS IN DRUG INFUSION SYSTEMS

Advances in science and technology result in new materials and devices. These materials and in particular electronic devices allow engineers to develop smarter, better performed, smaller, reliable, and cost-effective products. Therefore, new drug infusion systems are being developed and increasingly used in hospitals as well as in home therapies. Some of these developments are summarized in the following list.

***Automated Syringe-Filling System*:** Stanford Research Institute's (SRI) drug delivery system expertise is to develop a compact, home device for diabetics that would help them fill their insulin syringes accurately. The system needed to handle both long- and fast-acting insulin formulations and needed to be easy to use and reliable for elderly and vision-impaired patients. The SRI developed an automated system that stored both types of insulin, automatically resuspended the long-acting insulin, checked for adequate drug supply, dispensed the proper amount of medication, and kept a dosage record. The system is under test and evaluation (16).

***Disposable Drug Infusion Pump*:** Medical devices and pharmaceutical companies working with SRI to reengineer a disposable drug infusion pump design that reduced the number of parts by 30%, and reduced cost (16).

***Tiny Drug Infusion System*:** A tiny meter in a belt will someday monitor dosages of up to 12 drugs needed around the clock by patients with diabetes, cancer, or acquired immune deficiency syndrome (AIDS). The dime-sized device is being developed by Integrated Sensing Systems Inc. The device will make sure patients are receiving the correct drug in the right volume at the right flow rate. It will hook into a drug controller that attaches to a patient's belt. The controller, $\sim 2.5 \times 1$ in., (6.35×2.54 cm) will deliver drugs from an attached reservoir to the patient. The system, as envisioned, will

Figure 9. Outlook Safety Infusion System with DoseScan and DoseGuard by B|BRAUN technologies helping to ensure that the Right patient receives the Right medication in the Right dose from an authorized clinician at point of care.

Figure 10. The Ipump Pain Management System by BAXTER can be programmed for epidural, IV, or subcutaneous delivery. The PCA doses can be set per hour. Control flow rates can be set in 0.1 $mL \cdot h^{-1}$ increments for maximum flexibility with continuous flow rates up to 90 $mL \cdot h^{-1}$.

Figure 11. This AITECS by EO Systems is a multipurpose syringe pump with flow rates from 1 to 1500 $mL \cdot h^{-1}$, can be used for any nuclear cardiology and nuclear medicine infusion.

simultaneously deliver up to 12 drugs in units as small as nanoliters, or billionths of a liter (17).

***Bar-Coded Infusion System*:** B. Braun Medical has introduced the Horizon Outlook IV Safety Infusion System. Braun notes that the most common source of human error is inaccurate manual programming of intravenous pumps. The Braun infusion system uses bar code technology to ensure the right patient is receiving the right drug in the right dose from an "authorized" clinician. Its patented DoseScan bar code technology creates a primary level of safety, with automated checks and balances that augment the manual procedures in use today. Secondary protection is provided by its DoseGuard software, which notifies clinicians if institution defined dose limits are exceeded (18).

***Coronory Micro-Syringe*:** EndoBionics has created the first micromedical device to inject safely through vessel walls. Using standard interventional procedures, physicians will position the EndoBionics μSyringe (MicroSyringe) in coronary or peripheral vessels. While the μSyringe is closed, the microneedle is hidden and does not injure vessel walls as it is maneuvered into place. When the μSyringe is opened, the microneedle slides through the vessel wall to inject drugs directly to the surrounding tissue. The drugs are deposited around the outside of the vessel and diffuse inward through the vessel layers. The microscopic puncture is so small that it heals almost immediately, limiting trauma and bleeding (19).

***Needleless Injection*:** PowderJect Technologies has developed a technology that could be considered a hybrid of transdermal and parenteral (injection): a needleless injection. The company's device propels powder drugs with a supersonic jet of helium gas. A high pressure ampule of helium within the device is broken open, the gas flows through a cassette that is holding the powder between two membranes. The membranes rupture and the gas stream picks up the particles. The particles are propelled fast enough to penetrate the stratum corneum, the outer layer of the skin. The drug is targeted to the boundary between the epidermis and the dermis. Drugs then dissolve and either reach systemic circulation or exert a local effect. Vaccines can be picked up by antigen-presenting cells in the epidermis or by the lymph system (20).

Alza is another company that is developing technologies to deliver drugs through the skin. One of these technologies, called E-Trans, uses electrical current to deliver drugs across the skin, a process known as iontophoresis. The lead product is for the on-demand delivery of fentanyl, an opioid analgesic used for the treatment of acute pain. When a patient pushes a button on the device, current flows between two electrodes. As current flows, we get a predetermined amount of drug injected into the body. That gives a very reliable way of delivering a particular amount of drug into the body (20).

Alza is also developing what it calls Macroflux technology, which incorporates a thin titanium screen with microprojections to create mechanical pathways for drug transport. It expands the range of drugs amenable to transdermal delivery to include small hydrophilic molecules and macromolecules. It can be incorporated with the E-Trans technology or more traditional transdermal patches. One simple prototype in early exploration involves a Macroflux system where the projections have been coated with the therapeutic agent, such as a macromolecule. After application, the agent is rapidly absorbed into the skin (20).

Elan Pharmaceutical Technologies have a technology known as Medipad worn by the patient on the chest, back, or abdomen. The device is a small, plastic gas-driven pump with an adhesive backing. The adhesive is used to attach the device to the patient's body, and a button is pressed. A needle is deployed, which enters the subcutaneous space and then delivers the drug at a constant rate until the entire content of the reservoir is expended. The first applications for this device will be in chronic pain management and in the delivery of macromolecules that have inherently short biological half-lives (20).

***A Novel Device for Flow Monitoring*:** A novel device for blockage detection in catheters during drug delivery is designed. This device consists of a low cost disposable microfluidic chip and a nondisposable detection unit. The microfluidic chip consists of a microstructured silicon layer bonded between two glass covers using anodic bonding technology. The flow monitoring is performed by a robust light transmission method. The main component of the microfluidic chip is a movable element coupled with a spring to a base. Depending on the drug flow state the element is blocking or vacating an optical path through the chip (21).

***A High Performance Silicon Micropump*:** A new, low cost, high performance silicon micropump has been developed for a disposable drug delivery system (22). It is reported that the pump demonstrated linear and accurate (±5%) pumping characteristics for flow rates up to 2 $mL \cdot h^{-1}$ with intrinsic insensitivity to external conditions. The stroke volume of 160 nL was maintained constant by the implementation of a double limiter acting on the pumping membrane. The chip is a stack of three layers, two Pyrex wafers anodically bonded to the central silicon wafer. The technology is based on the use of Silicon On Insulator (SOI) technology, silicon Deep Reactive Ion Etching (DRIE), and the sacrificial etch of the buried oxide in order to release the structures (22).

***An Implantable Microfabricated Drug Delivery System*:** A fully implantable drug delivery system capable of delivering hundreds of individual doses has been developed by MicroCHIPS (23). This product is intended for the controlled release of potent therapeutic compounds that might otherwise require frequent injections. The device is capable of storing therapeutic drugs in solid, liquid, or gel form. It allows individual storage of discrete doses for multiple-drug regimens. Device monitoring and therapy modification can be achieved via wireless communication with an external controller. Currently, a fully implantable device contains 100 individual doses. A future device intended for human clinical trials will contain 400 doses, enough for a daily release of drug for >1 year (23).

***A Water-Powered Microdrug Delivery System*:** A plastic microdrug delivery system has been designed by utilizing the principle of osmosis without any electrical power consumption. The system has an osmotic microactuator and a polydimethylsiloxane (PDMS) microfluidic cover compartment consisting of a reservoir, a microfluidic channel, and a delivery port. The typical dimension of the microfluidic channel is 1 cm in length with a cross-sectional area of 30–100 μm^2 to minimize the diffusive drug flow while pressure drop remains moderate. Employing the net water flow induced by osmosis, the prototype drug delivery system has a measured constant delivery rate of 0.2 $\mu L \cdot h^{-1}$ for 10 h, with an accumulated delivery volume of 2 μL. Both the delivery rate and volume could be altered by changing the design and process parameters for specific drug delivery applications up to a few years (24).

***Microflow Regulator for Drug Delivery Systems*:** A micromachined flow regulator has been designed to provide a constant liquid flow rate of 1 $mL \cdot h^{-1}$ within a pressure difference of 100–600 mbar (0.01–0.06 kPa). At pressures >600 mbar (0.06 kPa) the device is designed to block the flow, preventing an overdelivery of medicine. One application of this device is the replacement of the flow restrictor in an elastomeric infusion system, which will increase the accuracy and safety of the drug delivery system. This pressure compensating flow regulator is passive; hence it needs no external energy source. The device is small, lightweight, and relatively inexpensive; therefore, it could be used as a disposable unit in a microfluidic system (25).

***Nanoengineered Device for Drug Delivery*:** A high precision device has been developed to yield long-term zero-order release of drugs for therapeutic applications. The device contains nanochannels that were fabricated in between two directly bonded silicon wafers, and therefore poses high mechanical strength. Diffusion through the nanochannels is the rate-limiting step for the release of drugs (26).

***Smartdose by PRO-MED AG*:** This device is a safe, accurate, and simple infusion system. It is a disposable prefilled drug delivery system for enteral or parenteral controlled infusion. SmartDose is equipped with its own source of energy (chemical reaction) to dispense liquid over a specific time with a predetermined administration rate. The administration accuracy and safety is comparable with those of electronic pumps, yet the ease of use is similar to a simple infusion bag. The system is especially convenient for emergency, ambulatory, and homecare therapy, as well as for hospitalized patients (27).

***Biodegradable Polymeric Drug Delivery Systems*:** These systems are increasingly being used for the design of temporary drug delivery systems. As these polymers hydrolyze in the body into low molecular degradation products, which are either metabolized or excreted, biodegradable delivery systems do not have to be removed after completion of release. Poly(DL-lactide-*co*-glycolide) (PLGA) is the most widely investigated biodegradable polyester and is widely used as a carrier polymer in parenteral sustained release formulations, either as microspheres, microparticulates or injectable gels (28).

RECENT ADVANCEMENTS IN CONTROLLER DESIGN FOR DRUG INFUSION SYSTEMS

Closed-loop system control is a technological concept that may be applicable to several aspects of critical care practice. This is a technology in the early stages of evolution and much more research and data are needed before its introduction into usual clinical practice. Furthermore, each specific application and each device for each application are sufficiently different in terms of hardware and computer algorithms (29). Studies have shown that closed-loop infusion systems may have a role in critical care

practice, improve clinical outcomes, eliminate errors due to poor performance of automated infusion devices, and provide precise, error-free drug administration. Some of the most recent works in advanced controller designs are reviewed below:

Huzmezan et al. (30) states that feedback control of drug administration is well suited to anesthetized surgical patients as well as the critically ill patients because of drugs with rapid onset times, short duration of action and small margins of safety are frequently used. The application of an adaptive predictive process control technology to drug administration will assist physicians in avoiding both overdosages and underdosages in their patients. An adaptive controller would avoid overdosing and underdosing by compensating for nonlinear drug responses as well as inter- and intrapatient variation (30).

Linkens has proposed the design of a fuzzy control for patient muscle relaxation (31). With advancements in sensor and instrumentation technology, automated drug infusion systems are also evolving into hierarchical systems. Research in this area has led to a variety of control strategies ranging from simple linear controllers to complex adaptive and rule-based schemes to handle inter- and intrapatient variability in drug responses (31).

For the assessment of depth of anesthesia, an intelligent system has been developed, which utilizes auditory evoked brain potentials, heart rate, and blood pressure measurements (32). Using wavelet analysis, the features within the auditory evoked signals are extracted and then fed to a learning neurofuzzy system, which in turn classifies the depth of anesthesia. In addition, the heart rate and blood pressure signals are used as a second measure based on a rule-based fuzzy logic system. The two measures are then fused to give a final indication of anesthetic depth. This is then fed back to a target controlled infusion (TCI) system for regulating the infusion of the drug Propofol for the maintenance phase of anaesthetic state (32).

A control strategy is developed by Bequette to regulate blood pressure and cardiac output during surgery (33). Adaptation is incorporated through a multiple model predictive control (MMPC) approach. A Bayesian-based estimator recursively updates weighting functions to find the best combination of models that describes the current input–output behavior; this weighted model is used for the output prediction (33).

A robust direct model reference adaptive controller (DMRAC) is developed by Palerm et al. (34) for plants with uncertainty in both the time delay elements and in the transfer function coefficients. The control of hemodynamic variables, particularly mean arterial pressure (MAP) and cardiac output (CO), is a challenging problem. A good controller is difficult to design, due to the complex, nonlinear behavior of the system. Adding to this are the significant changes in dynamics from one patient to another, and even variations in the patient's response to the drugs as his condition evolves (34).

A model predictive control strategy is developed and tested on a nonlinear canine circulatory model for the regulation of hemodynamic variables under critical care conditions (35). Different patient conditions, such as congestive heart failure, postoperative hypertension, and sepsis shock are studied in closed loop simulations. The model predictive controller, which uses a different linear model depending on the patient condition allows constraints to be explicitly enforced. The controller is initially tuned based on a linear plant model, then tested on the nonlinear physiological model; the simulations demonstrate the ability to handle constraints, such as drug dosage specifications, commonly desired by critical care physicians (35).

To evaluate the use of intelligent systems in the improvement of patient care, an agent was developed to regulate ICU patient sedation by Moore et al. in (36). A temporal differencing form of reinforcement learning was used to train the agent in the administration of intravenous propofol in simulated ICU patients. The agent utilized the well-studied Marsh–Schnider pharmacokinetic model to estimate the distribution of drug within the patient. A pharmacodynamic model then estimated drug effect. The agent demonstrated satisfactory control of the simulated patient's consciousness level in static and dynamic set-point conditions. It also satisfactorily demonstrated superior stability and responsiveness when compared to a well-tuned PID controller, which is a method of choice in closed-loop sedation control literature (36).

Advanced model-based controllers that can take into account the model of the patient and constraints on the state of the patient and the drug infusion rates have been developed (32). Delivery of insulin to type 1 diabetics, control of anesthesia, and chemotherapy for cancer patients are typical examples of drug delivery systems. The main objective of a drug delivery system is to provide effective therapy while minimizing the side effects. These controllers are based upon the theory of multiparametric programming. This theory allows an optimal division of the multidimensional space of the state of the patient into a set of regions and each region is characterized by a unique drug infusion law that is an explicit function of the state in the corresponding region. These developments simplify controller implementation and result in tighter control of drug infusion rates and better lifestyle for patients (37).

Parker et al. in (38) discusses closed-loop blood glucose regulation algorithms that use the intravenous route for insulin delivery to insulin-dependent diabetic patients. Classical control methods and advanced algorithms using implicit knowledge or explicit models (empirical, fundamental, or gray-box) of the diabetic patient are examined in (38). Current research on characterizing patient variability is presented, in the context of a model predictive controller able to adjust to changes in patient glucose and insulin sensitivity (38).

Linkens in (39) presents the control of on-line drug infusions to patients in an operating theater for regulating their muscle relaxation according to necessary surgical procedures. It is stated that fuzzy logic control (FLC) offers the advantages of model-free controller design for systems that are dynamically nonlinear, uncertain, and possibly time varying (39).

Rao et al. discusses the design of two different control methodologies for automated regulation of hemodynamic variables in (40). These controllers are designed to regulate

MAP and CO in critical care subjects using inotropic and vasoactive drugs. Both controllers account for inter – and intrapatient variability and handle drug infusion constraints. The first approach is a multiple model predictive controller (MMPC). The algorithm uses a multiple model adaptive approach in a model predictive control framework to account for variability and explicitly handle drug rate constraints. The second approach, a robust direct model reference adaptive controller (DMRAC) is developed for plants with uncertainty in both the time delay elements and in the transfer function coefficients, such as the drug infusion process. The controllers are experimentally evaluated on canines that are pharmacologically altered to exhibit symptoms of hypertension and depressed cardiac output (40).

Bequette (41) discusses the development of an artificial pancreas and current efforts in the control of complex systems. It is stated that advances in continuous glucose sensing, fast-acting insulin analogues, and a mature insulin pump market allow commercial realization of a closed-loop artificial pancreas. Model predictive control is discussed in-depth as an approach that is well suited for a closed-loop artificial pancreas (41).

Target controlled infusion (TCI) systems are discussed by Van Poucke et al. (42). In their work, a novel mathematical algorithm is proposed for controlling the effect site concentration using a TCI device. The algorithm limits the peak plasma concentration, thereby slowing the onset of anesthetic drug effect, but potentially ameliorating side effects. Simulations are used to examine the delay in time to peak effect for fentanyl, alfentanil, sufentanil, remifentanil, and propofol when the peak plasma concentration is limited by the algorithm. Results showed that the plasma overshoot can be reduced by 60% with only $\sim 20\%$ delay in the onset of drug effect (42).

McKinley et al. (43) compares the effectiveness of a new method of closed-loop control of blood pressure with usual manual control. In their work, closed-loop and manual drug administrations were studied. The target and observed MAP and drug infusion rate were recorded electronically. Time taken to achieve initial control; fidelity of control, and average drug dose administered were all measured. Results showed that closed-loop achieved faster initial control and greater fidelity as compared to manual control. There was no difference in average drug dose administered. It was concluded that the new closed-loop system is more effective than the usual manual control in managing acute blood pressure disturbances in the seriously ill patients (43).

The bispectral index (BIS) was used for automatic control of propofol anesthesia, using a proportional-integral-differential control algorithm (44). The performance of the controlled system was measured in patients undergoing minor surgery under propofol and remifentanil anesthesia. Anesthesia was manually induced with target-controlled infusions (TCI) of propofol and remifentanil. After the start of surgery, when anesthesia was clinically adequate, automatic control of the propofol TCI was commenced using the closed-loop system. The system provided adequate operating conditions and stable cardiovascular values in all patients during closed-loop control. The system was able to provide clinically adequate anesthesia in all patients (44).

Brock et al. in (45) presents a study to determine the relative advantage of computer-controlled couch movement versus manual repositioning to correct patient setup error measured using an electronic portal imaging device (EPID). The speed of setup adjustment and accuracy of corrected setup were determined. Computer-controlled setup adjustment was determined to be faster and slightly more accurate than manual correction (45).

Another comparison study between computer and manual control is presented in (46) by Hoeksel et al. They investigated the effects of computer-controlled blood pressures on hemodynamic stability when compared to conventional manual control. Systemic artery blood pressures were managed either by computer or by a well-trained anesthesiologist. Hemodynamic stability was determined from the standard deviation of the MAP samples and from the percentages of time that arterial pressure was hypertensive or hypotensive. The average standard deviation of the MAP samples was smaller for the computer-controlled than for the manually controlled group. The systemic artery pressure was less hypertensive and less hypotensive in the computer-controlled than in the manually controlled group. It was concluded that, compared with manual control, computer control of systemic hypertension significantly improved hemodynamic stability during cardiac surgery (46).

The clinical applicability of administering sodium nitroprusside by a closed-loop titration system compared with a manually adjusted system was evaluated. The MAP was registered and the results were then analyzed. It was reported that the computer-assisted therapy provided better control of MAP, was safe to use, and helped to reduce nursing demands (47).

Chitwood et al. in (48) states that hypertension is common after a cardiac operation and has been treated using manually controlled doses of intravenous sodium nitroprusside. To evaluate the clinical impact of an automated closed-loop administration system on patients after cardiotomy, a prospective trial was conducted. Patients with hypertension were managed by either manual nitroprusside titration or a closed-loop automated titration system. The automated group showed a significant reduction in the number of hypertensive episodes per patient. At the same time, the number of hypotensive episodes per patient was reduced with automated closed-loop titration. Chest tube drainage, percentage of patients receiving transfusion, and total amount transfused were all reduced significantly by the use of an automated titration system (48).

A nonprogrammable and programmable insulin external pump using regular insulin on glycemic stability, the risk of severe hypoglycemia, and metabolic control in type 1 diabetic patients was compared (49). The results of the study suggest that programmable external insulin pumps, although more complex and more expensive than nonprogrammable insulin pumps, significantly reduce fasting glycemia during the day without increasing the risk of severe hypoglycemia and are safer during the night (49).

In Ref. 50, it was argued that continuous improvements in microelectronics, as well as in the development of biomaterials and stable insulin solutions, led to the availability of implantable pumps that are able to infuse insulin by the peritoneal route, in a continuous and programmable way, for several years. These systems represent the most efficient and physiological mode of insulin therapy at the present time. It was demonstrated during clinical trials that intravascular, implantable, glucose sensors using glucose oxidase were able to measure with good accuracy real-time blood glucose for several months. In their study, they performed the first trials of closed-loop insulin delivery according to sensor signal for periods of 48 h in type 1 diabetic patients. This mode of functioning appeared to be feasible and able to establish glucose control closer to physiology than the use of implantable pumps in open loop (50).

Tamborlane et al. (51) states that while treatment of Type 1 diabetes mellitus (T1DM) in children and adolescents is especially difficult, recent technological advances have provided new therapeutic options to clinicians and patients. The urgency to achieve strict diabetes control and the introduction of new and improved insulin pumps have been accompanied by a marked increase in use of continuous subcutaneous insulin infusion (CSII) therapy in youth with diabetes. Results of clinical outcome studies indicate that CSII provides a safe and effective alternative to MDI therapy, even when employed in a regular clinic setting in a large number of children (51).

BIBLIOGRAPHY

1. Draft for Public Comment, Australian/New Zealand Standard. 2005. DR04547.
2. Medtech, Inc., U.S. Markets for Drug and Fluid Delivery Devices. 2001. Report No. RP-485004.
3. Infusion Systems, Medical Devices Agency. 2003. MDA DB 2003(2).
4. Sparks DR, et al. Preventing medication infusion errors and venous air embolisms using a micro-machined specific gravity sensor. Adv Deliv Devices 2004.
5. Kohn L, et al., editors. Report of Institute of Medical Committee on Quality Health Care in America, To err is human: building a safer health system. Washington (D.C.): National Academy Press; 2000.
6. Phillips J, et al. Retrospective analysis of mortalities associated with medication errors. Am J Health-Syst Pharm 2001;58:1831–1841.
7. Gorman R, et al. Prevention of medication errors in the pediatric inpatient setting. Pediatrics 2003;112(2):431–436.
8. Taxis K, Barber N. Ethnographic study of incidence and severity of intravenous drug errors. BMJ 2003;326:684–696.
9. Parshuram C, et al. Discrepancies between ordered and delivered concentrations of opiate infusions in critical care. Cri Care Med 2003;31(10):2483–2487.
10. Smart infusion pumps join CPOE and bar coding as important ways to prevent medication errors. ISMP Me Safety Alert 2002.
11. Wilson K, Sullivan M. Preventing medication errors with smart infusion technology. Am J Health-Syst Pharm 2004; 61(2):177–183.
12. Steingass SK. Beyond pumps: Smart infusion systems. Nursing Management 2004;35:10–10.
13. Ferrari R, Beech DR. Infusion pumps: guidelines and pitfalls. Aust Prescr 1995;18:49–51.
14. Davis WOM. Types of pumps in use, Infusion devices training tutorial (Online). Available at http://www.ebme.co.uk/arts/art11.htm. 2004.
15. Home drug infusion therapy under medicare. Congress of the United States, Office of Technology Assessment. 1992.
16. Intravenous therapy monitoring and discontinuing, (online) http://www.laras-lair.com/nursing/IVtherapy.pdf. RPN Self Learning Package. SRI International-Medical Development (Online). Available at http://www.sri.com/esd/med_devel/altdrugdel.html. 2005.
17. Infusion device could improve drug delivery (Online). Available at http://www.detnews.com/2004/technology/0402/16/b04-64095.htm. 2005.
18. Braun Medical introduces bar-coded infusion system (Online). Available at http://www.uspharmacist.com. 2002.
19. Endobionics, Inc. (online) http://www.endobionics.com/technology.html. 2005.
20. Henry C. Special delivery. Sci/Technol 2000;78(38):49–65.
21. Richter M, et al. A novel device for low flow monitoring in drug delivery systems. 8th Int Conf New Actuators 2002; 223–226.
22. Maillefer D, et al. A high performance silicon micropump for disposable drug delivery systems. Micro Electro Mechanical Systems (MEMS). The 14th IEEE International Conference on MEMS, 2001; p 413–417.
23. Maloney JM. An implantable microfabricated drug delivery system. Proc IMECE 2003;1–2.
24. Su Y, Lin L. A water-powered micro drug delivery system. IEEE J Microelectromechan Systems 2004;13(1):75–82.
25. Cousseau P, et al. Improved mico-flow regulator for drug delivery systems. IEEE Int Conf MicroElectroMechanical Syst 2001;527–530.
26. Sinha PM, et al. Nanoengineered device for drug delivery application. Inst Phys Publ Nanotechnol 2004;15:S585–S589.
27. Pro-Med AG. SmartDose controlled infusion systems. Business Briefing Pharmatech 2004;1–2.
28. Steendam R. SynBiosys: Biodegradable polymeric drug delivery system. Bus Briefing: Pharma Outsourcing 2005; 59–61.
29. Jastremski M, et al. A model for technology assessment as applied to closed loop infusion systems. Technol Ass Task Force Soc Crit Care Med 1995;23(10):1745–1755.
30. Huzmezan M, Dumont GA, Zikov T, Bibian S. Advances in automatic drug delivery for general anesthesia. Peter Wall Institute Exploratory Workshop, Automation and Robotics: The Key For Computer Integrated Health Care Delivery. University of British Columbia. 2002.
31. Linkens DA. Fuzzy control for patient muscle relaxation. Erudit News Letter. Vol. 2, No. 1 (online). Available at http://www.erudit.de/erudit/newsletters/news_21/page_5.htm, 2003.
32. Linkens DA. An intelligent system for drug delivery in control of anesthesia. Peter Wall Institute Exploratory Workshop, Automation and Robotics: The Key For Computer Integrated Health Care Delivery. University of British Columbia. 2002.
33. Bequette BW. A multiple model approach for adaptation during drug delivery. Peter Wall Institute Exploratory Workshop, Automation and Robotics: The Key For Computer Integrated Health Care Delivery. University of British Columbia. 2002.
34. Palerm C, Bequette BW, Ozcelik S. Robust control of drug infusion with time delays using direct adaptive control: experimental results. Am Control Conf 2000.
35. Rao R, Huang J, Bequette BW, Kaufman H. Control of a non-square drug infusion system-a simulation study. Biotechnol Prog ACS Pub 1999;15(3):556–564.
36. Moore B, Sinzinger E, Quasny T, Pyeatt L. Intelligent control of closed-loop sedation in simulated ICI patients. Am Assoc Artificial Intelligence 2004.

37. Dua V. Explicit model based control for drug delivery systems, CoMPLEX seminar (Online). Available at http://www.ucl. ac.uk/CoMPLEX/dua_abstract.html. 2004.
38. Parker R, Doyle F, Peppas N. The intravenous route to blood glucose control. IEEE Eng Med Bio Mag 2001;20(1):65–71.
39. Linkens DA. Fuzzy control for patient muscle relaxation. Erudit News Letter Vol. 2, No. 1, (Online). Available at http://www.erudit.de/erudit/newsletters/news_21/page_5.htm. 2003.
40. Rao RR, Palerm CC, Aufderheide B, Bequette BW. Automated regulation of hemodynamic variables. IEEE Eng Med Bio Mag 2001;20(1):2438.
41. Bequette BW. A Critical Assessment of Algorithms and Challenges in the Development of a Closed-Loop Artificial Pancreas. Diab Technol Therapeut 2005;7:28–47.
42. Van Poucke GE, Bravo LJB, Shafer SL. Target controlled infusions: targeting the effect site while limiting peak plasma concentration. IEEE Trans Biomed Eng 2004;51:1869–1875.
43. McKinley S, et al. Clinical evaluation of closed loop control of blood pressure in seriously ill patients. Crit Care Med 1991;19(2):166–170.
44. Absalom AR, Kenny GNC. Closed-loop control of propofol anesthesia using bispectral index: performance assessment in patients receiving computer-controlled propofol and manually controlled remifentanil infusions for minor surgery. Br J Anesthesia 2003;90:737–741.
45. Brock KK, McShan DL, Balter JM. A comparison of computer-controlled versus manual on-line patient setup adjustment. J App Clin Med Phys 2002;3(3):241.
46. Hoeksel SAAP, Blom JA. Computre control versus manual control of systemic hypertension during cardiac surgery. Acta Anesthesiol Scand 2001;45:553–557.
47. Bednarski P, et al. Use of a computerized closedloop sodium nitroprusside titration system for antihypertensive treatment after open heart surgery. Crit Care Med 1990;18(10):1061–1065.
48. Chitwood WR, III DMC, Lust RM. Multicenter trial of automated nitroprusside infusion for postoperative hypertension. Ann Thorac Surg 1992;54:517–522.
49. Catargi B et al. A randomized study comparing blood glucose control and risk of severe hypoglycemia achieved by nonprogrammable versus programmable external insulin pumps. Diab Metab 2001;27:323–328.
50. Renard E, Costalat G, Bringer J. From external to implantable insulin pump, can we close the loop? Diab Metab 2002;28(4 Pt. 2):19–25.
51. Tamborlane W, Bonfig W, Boland E. Recent advances in treatment of youth with type 1 diabetes: Better care through technology. Diab Metab 2001;18:864–870.

See also DRUG DELIVERY SYSTEMS; NUTRITION, PARENTERAL.

ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition

VOLUME 3

Echocardiography and Doppler Echocardiography – Human Spine, Biomechanics of

ENCYCLOPEDIA OF MEDICAL DEVICES AND INSTRUMENTATION, SECOND EDITION

ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition
Volume 3

Echocardiography and Doppler Echocardiography – Human Spine, Biomechanics of

Edited by

John G. Webster
University of Wisconsin–Madison

The *Encyclopedia of Medical Devices and Instrumentation* is available online at http://www.mrw.interscience.wiley.com/emdi

WILEY-INTERSCIENCE

A John Wiley & Sons, Inc., Publication

Published by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Encylopedia of medical devices & instrumentation/by John G. Webster,

editor in chief. – 2nd ed.
p. ; cm.
Rev. ed. of: Encyclopedia of medical devices and instrumentation. 1988.
Includes bibliographical references and index.
ISBN-13 978-0-471-26358-6 (set : cloth)
ISBN-10 0-471-26358-3 (set : cloth)
ISBN-13 978-0-470-04068-3 (v. 3 : cloth)
ISBN-10 0-470-04068-8 (v. 3 : cloth)

1. Medical instruments and apparatus–Encyclopedias. 2. Biomedical engineering–Encyclopedias. 3. Medical physics–Encyclopedias. 4. Medicine–Data processing–Encyclopedias. I. Webster, John G., 1932- . II. Title: Encyclopedia of medical devices and instrumentation.

[DNLM: 1. Equipment and Supplies–Encyclopedias–English. W 13

E555 2006]
R856.A3E53 2006
610.2803–dc22
2005028946

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTRIBUTOR LIST

ABDEL HADY, MAZEN, *McMaster University, Hamilton, Ontario Canada*, Bladder Dysfunction, Neurostimulation of

ABEL, L.A., *University of Melbourne, Melbourne, Australia*, Ocular Motility Recording and Nystagmus

ABREU, BEATRIZ C., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

ALEXANDER, A.L., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

ALI, ABBAS, *University of Illinois, at Urbana-Champaign*, Bioinformatics

ALI, MÜFTÜ, *School of Dental Medicine, Boston, Massachusetts*, Tooth and Jaw, Biomechanics of

ALPERIN, NOAM, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

ANSON, DENIS, *College Misericordia, Dallas, Pennsylvania*, Environmental Control

ARENA, JOHN C., *VA Medical Center and Medical College of Georgia*, Biofeedback

ARIEL, GIDEON, *Ariel Dynamics, Canyon, California*, Biomechanics of Exercise Fitness

ARMSTRONG, STEVE, *University of Iowa, Iowa City, Iowa*, Biomaterials for Dentistry

ASPDEN, R.M., *University of Aberdeen, Aberdeen, United Kingdom*, Ligament and Tendon, Properties of

AUBIN, C.E., *Polytechniquie Montreal, Montreal Quebec, Canada*, Scoliosis, Biomechanics of

AYRES, VIRGINIA M., *Michigan State University, East Lansing, Michigan*, Microscopy, Scanning Tunneling

AZANGWE, G., Ligament and Tendon, Properties of

BACK, LLOYD H., *California Institute of Technology, Pasadena, California*, Coronary Angioplasty and Guidewire Diagnostics

BADYLAK, STEPHEN F., *McGowan Institute for Regenerative Medicine, Pittsburgh, Pennsylvania*, Sterilization of Biologic Scaffold Materials

BANDYOPADHYAY, AMIT, *Washington State University, Pullman, Washington*, Orthopedic Devices, Materials and Design for

BANERJEE, RUPAK K., *University of Cincinnati, Cincinnati, Ohio*, Coronary Angioplasty and Guidewire Diagnostics

BARBOUR, RANDALL L., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

BARKER, STEVEN J., *University of Arizona, Tucson, Arizona*, Oxygen Monitoring

BARTH, ROLF F., *The Ohio State University, Columbus, Ohio*, Boron Neutron Capture Therapy

BECCHETTI, F.D., *University of Michigan, Ann Arbor, Michigan*, Radiotherapy, Heavy Ion

BELFORTE, GUIDO, *Politecnico di Torino – Department of Mechanics*, Laryngeal Prosthetic Devices

BENKESER, PAUL, *Georgia Institute of Technology, Atlanta, Georgia*, Biomedical Engineering Education

BENNETT, JAMES R., *University of Iowa, Iowa City, Iowa*, Digital Angiography

BERSANO-BEGEY, TOMMASO, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

BIGGS, PETER J., *Harvard Medical School, Boston, Massachusetts*, Radiotherapy, Intraoperative

BIYANI, ASHOK, *University of Toledo, and Medical College of Ohio, Toledo, Ohio*, Human Spine, Biomechanics of

BLOCK, W.F., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

BLUE, THOMAS E., *The Ohio State University, Columbus, Ohio*, Boron Neutron Capture Therapy

BLUMSACK, JUDITH T., *Disorders Auburn University, Auburn, Alabama*, Audiometry

BOGAN, RICHARD K., *University of South Carolina, Columbia, South Carolina*, Sleep Laboratory

BOKROS, JACK C., *Medical Carbon Research Institute, Austin, Texas*, Biomaterials, Carbon

BONGIOANNINI, GUIDO, *ENT Division Mauriziano Hospital, Torino, Italy*, Laryngeal Prosthetic Devices

BORAH, JOSHUA, *Applied Science Laboratories, Bedford, Massachusetts*, Eye Movement, Measurement Techniques for

BORDEN, MARK, *Director of Biomaterials Research, Irvine, California*, Biomaterials, Absorbable

BORTON, BETTIE B., *Auburn University Montgomery, Montgomery, Alabama*, Audiometry

BORTON, THOMAS E., *Auburn University Montgomery, Montgomery, Alabama*, Audiometry

BOSE SUSMITA,, *Washington State University, Pullman, Washington*, Orthopedic Devices, Materials and Design for

BOVA, FRANK J., *M. D. Anderson Cancer Center Orlando, Orlando, FL*, Radiosurgery, Stereotactic

BRENNER, DAVID J., *Columbia University Medical Center, New York, New York*, Computed Tomography Screening

BREWER, JOHN M., *University of Georgia*, Electrophoresis

BRIAN, L. DAVIS, *Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, Ohio*, Skin, Biomechanics of

BRITT, L.D., *Eastern Virginia Medical School, Norfolk, Virginia*, Gastrointestinal Hemorrhage

BRITT, R.C., *Eastern Virginia Medical School, Norfolk, Virginia*, Gastrointestinal Hemorrhage

BROZIK, SUSAN M., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

BRUNER, JOSEPH P., *Vanderbilt University Medical Center, Nashville, Tennessee*, Intrauterine Surgical Techniques

BRUNSWIG NEWRING, KIRK A., *University of Nevada, Reno, Nevada*, Sexual Instrumentatio n

BRUYANT, PHILIPPE P., *University of Massachusetts, North Worcester, Massachusetts*, Nuclear Medicine, Computers in

BUNNELL, BERT J., *Bunnell Inc., Salt Lake City, Utah*, High Frequency Ventilation

CALKINS, JERRY M., *Defense Research Technologies, Inc., Rockville, Maryland*, Medical Gas Analyzers

CANNON, MARK, *Northwestern University, Chicago, Illinois*, Resin-Based Composites

CAPPELLERI, JOSEPH C., *Pfizer Inc., Groton, Connecticut*, Quality-of-Life Measures, Clinical Significance of

CARDOSO, JORGE, *University of Madeira, Funchal, Portugal*, Office Automation Systems

CARELLO, MASSIMILIANA, *Politecnicodi Torino – Department of Mechanics*, Laryngeal Prosthetic Devices

CASKEY, THOMAS C., *Cogene Biotech Ventures, Houston, Texas*, Polymerase Chain Reaction

CECCIO, STEVEN, *University of Michigan, Ann Arbor, Michigan*, Heart Valve Prostheses, In Vitro Flow Dynamics of

CHAN, JACKIE K., *Columbia University, New York, New York*, Photography, Medical

CHANDRAN, K.B., *University of Iowa, Iowa City, Iowa*, Heart Valve Prostheses

CHATZANDROULIS, S., *NTUA, Athens, Attiki, Greece*, Capacitive Microsensors for Biomedical Applications

CHAVEZ, ELIANA, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CHEN, HENRY, *Stanford University, Palo Alto, California*, Exercise Stress Testing

CHEN, JIANDE, *University of Texas Medical Branch, Galveston, Texas*, Electrogastrogram

CHEN, YAN, *Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, Ohio*, Skin, Biomechanics of

CHEYNE, DOUGLAS, *Hospital for Sick Children Research Institute*, Biomagnetism

CHUI, CHEN-SHOU, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

CLAXTON, NATHAN S., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

CODERRE, JEFFREY A., *Massachus etts Institute of Technology, Cambridge, Massachusetts*, Boron Neutron Capture Therapy

COLLINS, BETH, *University of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

COLLINS, DIANE, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CONSTANTINOU, C., *Columbia University Radiation Oncology, New York, New York*, Phantom Materials in Radiology

COOK, ALBERT, *University of Alberta*, *Edmonton, Alberta, Canada*, Communication Devices

COOPER, RORY, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CORK, RANDALL C., *Louisiana State University, Shreveport, Louisiana*, Monitoring, Umbilical Artery and Vein, Blood Gas Measurements; Transcuta neous Electrical Nerve Stimulation (TENS); Ambulatory Monitoring

COX, JOSEPHINE H., *Walter Reed Army Institute of Research, Rockville, Maryland*, Blood Collection and Processing

CRAIG, LEONARD, *Feinberg School of Medicine of Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

CRESS, CYNTHIA J., *University of Nebraska, Lincoln, Nebraska*, Communicative Disorders, Computer Applications for

CUMMING, DAVID R.S., *University of Glasgow, Glasgow, United Kingdom*, Ion-Sensitive Field-Effect Transistors

CUNNINGHAM, JOHN R., *Camrose, Alberta, Canada*, Cobalt 60 Units for Radiotherapy

D'ALESSANDRO, DAVID, *Montefiore Medical Center, Bronx, New York*, Heart–Lung Machines

D'AMBRA, MICHAEL N., *Harvard Medical School, Cambridge, Massachusetts*, Cardiac Output, Thermodilution Measurement of

DADSETAN, MAHROKH, *Mayo Clinic, College of Medicine, Rochester, Minnesota*, Microscopy, Electron

DALEY, MICHAEL L., *The University of Memphis, Memphis, Tennessee*, Monitoring, Intracranial Pressure

DAN, LOYD, *Linköping University, Linköping, Sweden*, Thermocouples

DAS, RUPAK, *University of Wisconsin, Madison, Wisconsin*, Brachytherapy, High Dosage Rate

DATTAWADKAR, AMRUTA M., *University of Wisconsin, Madison, Madison, Wisconsin*, Ocular Fundus Reflectometry

DAVIDSON, MICHAEL W., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

DE LUCA, CARLO, *Boston University, Boston, Massachusetts*, Electromyography

DE SALLES, ANTONIO A.F., *UCLA Medical School, Los Angeles, California*, Stereotactic Surgery

DECAU, SABIN, *University of Maryland, School of Medicine*, Shock, Treatment of

DECHOW, PAUL C., *A & M University Health Science Center, Dallas, Texas*, Strain Gages

DELBEKE, JEAN, *Catholique University of Louvain, Brussels, Belgium*, Visual Prostheses

DELL'OSSO, LOUIS F., *Case Western Reserve University, Cleveland, Ohio*, Ocular Motility Recording and Nystagmus

DELORME, ARNAUD, *University of San Diego, La Jolla, California*, Statistical Methods

DEMENKOFF, JOHN, *Mayo Clinic, Scottsdale, Arizona*, Pulmonary Physiology

DEMIR, SEMAHAT S., *The University of Memphis and The University of Tennessee Health Science Center, Memphis, Tennessee*, Electrophysiology

DEMLING, ROBERT H., *Harvard Medical School*, Skin Substitute for Burns, Bioactive

DENNIS, MICHAEL J., *Medical University of Ohio, Toledo, Ohio*, Computed Tomography

DESANTI, LESLIE, *Harvard Medical School*, Skin Substitute for Burns, Bioactive

DEUTSCH, STEVEN, *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

DEVINENI, TRISHUL, *Conemaugh Health System*, Biofeedback

DI BELLA EDWARD, V.R., *University of Utah*, Tracer Kinetics

DIAKIDES, NICHOLAS A., *Advanced Concepts Analysis, Inc., Falls Church, Virginia*, Thermography

DOLAN, PATRICIA L., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

DONOVAN, F.M., *University of South Alabama*, Cardiac Output, Indicator Dilution Measurement of

DOUGLAS, WILSON R., *Children's Hospital of Philadelphia, Philadelphia, Pennsylvania*, Intrauterine Surgical Techniques

DRAPER, CRISSA, *University of Nevada, Reno, Nevada*, Sexual Instrumentation

DRZEWIECKI, TADEUSZ M., *Defense Research Technologies, Inc., Rockville, Maryland*, Medical Gas Analyzers

DURFEE, W.K., *University of Minnesota, Minneapolis, Minnesota*, Rehabilitation and Muscle Testing

DYRO, JOSEPH F., *Setauket, New York*, Safety Program, Hospital

DYSON, MARY, *Herts, United Kingdom*, Heat and Cold, Therapeutic

ECKERLE, JOSEPH S., *SRI International, Menlo Park, California*, Tonometry, Arterial

EDWARDS, BENJAMIN, *University of Wisconsin-Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

EDWARDS, THAYNE L., *University of Washington, Seattle, Washington*, Chromatography

EKLUND, ANDERS, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

EL SOLH, ALI A., *Erie County Medical Center, Buffalo, New York*, Sleep Studies, Computer Analysis of

ELMAYERGI, NADER, *McMaster University, Hamilton, Ontario, Canada*, Bladder Dysfunction, Neurostimulation of

ELSHARYDAH, AHMAD, *Louisiana State University, Baton Rouge, Louisiana*, Ambulatory Monitoring; Monitoring, Umbilical Artery and Vein, Blood Gas Measurements

FADDY, STEVEN C., *St. Vincents Hospital, Sydney, Darlinghurst, Australia*, Cardiac Output, Fick Technique for

FAHEY, FREDERIC H., *Childrens Hospital Boston*, Computed Tomography, Single Photon Emission

FAIN, S.B., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

FELDMAN, JEFFREY, *Childrens Hospital of Philadelphia, Philadelphia, Pennsylvania*, Anesthesia Machines

FELLERS, THOMAS J., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

FERRARA, LISA, *Cleveland Clinic Foundation, Cleveland, Ohio*, Human Spine, Biomechanics of

FERRARI, MAURO, *The Ohio State University, Columbus, Ohio*, Drug Delivery Systems

FONTAINE, ARNOLD A., *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

FOUST, MILTON J., JR, *Medical University of South Carolina Psychiatry and Behavioral Sciences, Charleston, South Carolina*, Electroconvulsive Therapy

FRASCO, PETER, *Mayo Clinic Scottsdale, Scottsdale, Arizona*, Temperature Monitoring

FRAZIER, JAMES, *Louisiana State University, Baton Rouge, Louisiana*, Ambulatory Monitoring

FREIESLEBEN DE BLASIO, BIRGITTE, *University of Oslo, Oslo, Norway*, Impedance Spectroscopy

FRESTA, MASSIMO, *University of Catanzaro Magna Græcia, Germaneto (CZ), Italy*, Drug Delivery Systems

FREYTES, DONALD O., *McGowan Institute for Regenerative Medicine, Pittsburgh Pennsylvania*, Sterilization of Biologic Scaffold Materials

FROEHLICHER, VICTOR, *VA Medical Center, Palo Alto, California*, Exercise Stress Testing

FUNG, EDWARD K., *Columbia University, New York, New York*, Photography, Medical

GAGE, ANDREW A., *State University of New York at Buffalo, Buffalo, New York*, Cryosurgery

GAGLIO, PAUL J., *Columbia University College of Physicians and Surgeons*, Liver Transplantation

GARDNER, REED M., *LDS Hospital and Utah University, Salt Lake City, Utah*, Monitoring, Hemodynamic

GEJERMAN, GLEN, *Hackensack University Medical, Hackensack, New Jersey*, Radiation Therapy, Quality Assurance in

GEORGE, MARK S., *Medical University of South Carolina Psychiatry and Behavioral Sciences, Charleston, South Carolina*, Electroconvulsive Therapy

GHARIEB, R.R., *Infinite Biomedical Technologies, Baltimore, Maryland*, Neurological Monitors

GLASGOW, GLENN P., *Loyola University of Chicago, Maywood, Illinois*, Radiation Protection Instrumentation

GLASGOW, GLENN, *University of Wisconsin-Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

GOEL, VIJAY K., *University of Toledo, and Medical College of Ohio, Toledo, Ohio*, Human Spine, Biomechanics of

GOETSCH, STEVEN J., *San Diego Gamma Knife Center, La Jolla, California*, Gamma Knife

GOLDBERG, JAY R., *Marquette University Milwaukee, Wisconsin*, Minimally Invasive Surgery

GOLDBERG, ZELENNA, *Department of Radiation Oncology, Davis, California*, Ionizing Radiation, Biological Effects of

GOPALAKRISHNAKONE, P., *National University of Singapore, Singapore*, Immunologically Sensitive Field-Effect Transistors

GOPAS, JACOB, *Ben Gurion University of the Negev, Beer Sheva, Israel*, Monoclonal Antibodies

GORGULHO, ALESSANDRA, *UCLA Medical School, Los Angeles, California*, Stereotactic Surgery

GOUGH, DAVID A., *University of California, La Jolla, California*, Glucose Sensors

GOUSTOURIDIS, D., *NTUA, Athens, Attiki, Greece*, Capacitive Microsensors for Biomedical Applications

GRABER, HARRY L., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

GRAÇA, M., *Louisiana State University, Baton Rouge, Louisiana*, Boron Neutron Capture Therapy

GRANT, WALTER III, *Baylor College of Medicine, Houston, Texas*, Radiation Therapy, Intensity Modulated

GRAYDEN, EDWARD, *Mayo Health Center, Albertlea, Minnesota*, Cardiopulmonary Resuscitation

GREEN, JORDAN R., *University of Nebraska, Lincoln, Nebraska*, Communicative Disorders, Computer Applications for

HAEMMERICH, DIETER, *Medical University of South Carolina, Charleston, South Carolina*, Tissue Ablation

HAMAM, HABIB, *Université de Moncton, Moncton New Brunswick, Canada*, Lenses, Intraocular

HAMMOND, PAUL A., *University of Glasgow, Glasgow, United Kingdom*, Ion-Sensitive Field-Effect Transistors

HANLEY, JOSEPH, *Hackensack University Medical, Hackensack, New Jersey*, Radiation Therapy, Quality Assurance in

HARLEY, BRENDAN A., *Massachusetts Institute of Technology*, Skin Tissue Engineering for Regeneration

HARPER, JASON C., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

HASMAN, ARIE, *Maastricht, The Netherlands*, Medical Education, Computers in

HASSOUNA, MAGDY, *Toronto Western Hospital, Toronto, Canada*, Bladder Dysfunction, Neurostimulation of

HAYASHI, KOZABURO, *Okayama University of Science, Okayama, Japan*, Arteries, Elastic Properties of

HENCH, LARRY L., *Imperial College London, London, United Kingdom*, Biomaterials: Bioceramics

HETTRICK, DOUGLAS A., *Sr. Principal Scientist Medtronic, Inc., Minneapolis, Minnesota*, Bioimpedance in Cardiovascular Medicine

HIRSCH-KUCHMA, MELISSA, *University of Central Florida NanoScience Technology Center, Orlando, Florida*, Biosurface Engineering

HOLDER, GRAHAM E., *Moorfields Eye Hospital, London, United Kingdom*, Electroretinography

HOLMES, TIMOTHY, *St. Agnes Cancer Center, Baltimore, Maryland*, Tomotherapy

HONEYMAN-BUCK, JANICE C., *University of Florida, Gainesville, Florida*, Radiology Information Systems

HOOPER, BRETT A., *Areté Associates, Arlington, Virginia*, Endoscopes

HORN, BRUCE, *Kaiser Permanente, Los Angeles, California*, X-Rays Production of

HORNER, PATRICIA I., *Biomedical Engineering Society Landover, Maryland*, Medical Engineering Societies and Organizations

HOROWITZ, PAUL M., *University of Texas, San Antonio, Texas*, Fluorescence Measurements

HOU, XIAOLIN, *Risø National Laboratory, Roskilde, Denmark*, Neutron Activation Analysis

HOVORKA, ROMAN, *University of Cambridge, Cambridge, United Kingdom*, Pancreas, Artificial

HUANG, H.K., *University of Southern California*, Teleradiology

HUNT, ALAN J., *University of Michigan, Ann Arbor, Michigan*, Optical Tweezers

HUTTEN, HELMUT, *University of Technology, Graz, Australia*, Impedance Plethysmography

IAIZZO, P.A., *University of Minnesota, Minneapolis, Minnesota*, Rehabilitation and Muscle Testing

IBBOTT, GEOFFREY S., *Anderson Cancer Center, Houston, Texas*, Radiation Dosimetry, Three-Dimensional

INGHAM, E., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

ISIK, CAN, *Syracuse University, Syracuse, New York*, Blood Pressure Measurement

JAMES, SUSAN P., *Colorado State University, Fort Collins, Colorado*, Biomaterials: Polymers

JENSEN, WINNIE, *Aalborg University, Aalborg, Denmark*, Electroneurography

JIN, CHUNMING, *North Carolina State University, Raleigh, North Carolina*, Biomaterials, Corrosion and Wear of

JIN, Z.M., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

JOHNSON, ARTHUR T., *University of Maryland College Park, Maryland*, Medical Engineering Societies and Organizations

JONES, JULIAN R., *Imperial College London, London, United Kingdom*, Biomaterials: Bioceramics

JOSHI, ABHIJEET, *Abbott Spine, Austin, Texas*, Spinal Implants

JUNG, RANU, *Arizona State University, Tempe, Arizona*, Functional Electrical Stimulation

JURISSON, SILVIA S., *University of Missouri Columbia, Missouri*, Radionuclide Production and Radioactive Decay

KAEDING, PATRICIA J., *Godfrey & Kahn S.C., Madison, Wisconsin*, Codes and Regulations: Medical Devices

KAMATH, CELIA C., *Mayo Clinic, Rochester, Minnesota*, Quality-of-Life Measures, Clinical Significance of

KANE, MOLLIE, *Madison, Wisconsin*, Contraceptive Devices

KATHERINE, ANDRIOLE P., *Harvard Medical School, Boston, Massachusetts*, Picture Archiving and Communication Systems

KATSAGGELOS, AGGELOS K., *Northwestern University, Evanston, Illinois*, DNA Sequencing

KATZ, J. LAWRENCE, *University of Missouri-Kansas City, Kansas City, Missouri*, Bone and Teeth, Properties of

KESAVAN, SUNIL, *Akebono Corporation, Farmington Hills, Michigan*, Linear Variable Differential Transformers

KHANG, GILSON, *Chonbuk National University*, Biomaterials: Tissue Engineering and Scaffolds

KHAODHIAR, LALITA, *Harvard Medical School, Boston, Massachusetts*, Cutaneous Blood Flow, Doppler Measurement of

KIM, MOON SUK, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

KIM, YOUNG KON, *Inje University, Kimhae City, Korea*, Alloys, Shape Memory

KINDWALL, ERIC P., *St. Luke's Medical Center, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

KING, MICHAEL A., *University of Massachusetts, North Worcester, Massachusetts*, Nuclear Medicine, Computers in

KLEBE, ROBERT J., *University of Texas, San Antonio, Texas*, Fluorescence Measurements

KLEIN, BURTON, *Burton Klein Associates, Newton, Massachusetts*, Gas and Vacuum Systems, Centrally Piped Medical

KNOPER, STEVEN R., *University of Arizona College of Medicine*, Ventilatory Monitoring

KONTAXAKIS, GEORGE, *Universidad Politécnica de Madrid, Madrid, Spain*, Positron Emission Tomography

KOTTKE-MARCHANT, KANDICE, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Vascular Graft Prosthesis

KRIPFGANS, OLIVER, *University of Michigan, Ann Arbor, Michigan*, Ultrasonic Imaging

KULKARNI, AMOL D., *University of Wisconsin–Madison, Madison, Wisconsin*, Ocular Fundus Reflectometry, Visual Field Testing

KUMARADAS, J. CARL, *Ryerson University, Toronto, Ontario, Canada*, Hyperthermia, Interstitial

KUNICKA, JOLANTA, *Bayer HealthCare LLC, Tarrytown, New York*, Differential Counts, Automated

KWAK, KWANJ JOO, *University of Miami Miller School of Medicine, Miami, Florida*, Microscopy, Scanning Force

LAKES, RODERIC, *University of Wisconsin-Madison*, Bone and Teeth, Properties of

LAKKIREDDY, DHANUNJAYA, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

LARSEN, COBY, *Case Western Reserve University, Cleveland, Ohio*, Vascular Graft Prosthesis

LASTER, BRENDA H., *Ben Gurion University of the Negev, Beer Sheva, Israel*, Monoclonal Antibodies

LATTA, LOREN, *University of Miami, Coral Gables, Florida*, Rehabilitation, Orthotics in

LEDER, RON S., *Universidad Nacional Autonoma de Mexico Mexico, Distrito Federal*, Continuous Positive Airway Pressure

LEE, CHIN, *Harvard Medical School, Boston, Massachusetts*, Radiotherapy Treatment Planning, Optimization of; Hyperthermia, Interstitial

LEE, HAI BANG, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

LEE, SANG JIN, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

LEI, LIU, *Department of General Engineering, Urbana, Illinois*, Bioinformatics

LEI, XING, *Stanford University, Stanford, California*, Radiation Dose Planning, Computer-Aided

LEWIS, MATTHEW C., *Medical College of Wisconsin, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

LI, CHAODI, *University of Notre Dame, Notre Dame, Indiana*, Bone Cement, Acrylic

LI, JONATHAN G., *University of Florida, Gainesville, Florida*, Radiation Dose Planning, Computer-Aided

LI, QIAO, *University of Michigan, Ann Arbor, Michigan*, Immunotherapy

LI, YANBIN, *University of Arkansas, Fayetteville, Arkansas*, Piezoelectric Sensors

LIBOFF, A.R., *Oakland University, Rochester, Michigan*, Bone Ununited Fracture and Spinal Fusion, Electrical Treatment of

LIGAS, JAMES, *University of Connecticut, Farmington, Connecticut*, Respiratory Mechanics and Gas Exchange

LIMOGE, AIME, *The René Descartes University of Paris, Paris, France*, Electroanalgesia, Systemic

LIN, PEI-JAN PAUL, *Beth Israel Deaconess Medical Center, Boston, Massachusets*, Mammography

LIN, ZHIYUE, *University of Kansas Medical Center, Kansas City, Kansas*, Electrogastrogram

LINEAWEAVER, WILLIAM C., *Unive rsity of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

LIPPING, TARMO, *Tampere University of Technology, Pori, Finland*, Monitoring in Anesthesia

LIU, XIAOHUA, *The University of Michigan, Ann Arbor, Michigan*, Polymeric Materials

LLOYD, J.J., *Regional Medical Physics Department, Newcastle-upon-Tyne, United Kingdom*, Ultraviolet Radiation in Medicine

LOEB, ROBERT, *University of Arizona, Tuscon, Arizona*, Anesthesia Machines

LOPES DE MELO, PEDRO, *State University of Rio de Janeiro, Térreo Salas, Maracanã, Thermistors*

LOUDON, ROBERT G., Lung Sounds

LOW, DANIEL A., *Washington University School of Medicine, St. Louis, Missouri*, Radiation Therapy Simulator

LU, LICHUN, *Mayo Clinic, College of Medicine, Rochester, Minnesota*, Microscopy, Electron

LU, ZHENG FENG, *Columbia University, New York, New York*, Screen-Film Systems

LYON, ANDREW W., *University of Calgary, Calgary, Canada*, Flame Atomic Emission Spectrometry and Atomic Absorption Spectrometry

LYON, MARTHA E., *University of Calgary, Calgary, Canada*, Flame Atomic Emission Spectrometry and Atomic Absorption Spectrometry

MA, C-M CHARLIE, *Fox Chase Cancer Center, Philadelphia, Pennsylvania*, X-Ray Therapy Equipment, Low and Medium Energy

MACIA, NARCISO F., *Arizona State University at the Polytechnic Campus, Mesa, Arizona*, Pneumotachometers

MACKENZIE, COLIN F., *University of Maryland, School of Medicine*, Shock, Treatment of

MACKIE, THOMAS R., *University of Wisconsin, Madison, Wisconsin*, Tomotherapy

MADNANI, ANJU, *LSU Medical Centre, Shreveport, Louisiana*, Transcutaneous Electrical Nerve Stimulation (TENS)

MADNANI, SANJAY, *LSU Medical Centre, Shreveport, Louisiana*, Transcutaneous Electrical Nerve Stimulation (TENS)

MADSEN, MARK T., *University of Iowa, Iowa City, Iowa*, Anger Camera

MAGNANO, MAURO, *ENT Division Mauriziano Hospital, Torino, Italy*, Drug Delivery Systems

MANDEL, RICHARD, *Boston University School of Medicine, Boston, Massachusetts*, Colorimetry

MANNING, KEEFE B., *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

MAO, JEREMY J., *University of Illinois at Chicago, Chicago, Illinois*, Cartilage and Meniscus, Properties of

MARCOLONGO, MICHELE, *Drexel University, Philadelphia, Pennsylvania*, Spinal Implants

MAREK, MIROSLAV, *Georgia Institute of Technology, Atlanta, Georgia*, Biomaterials, Corrosion and Wear of

MARION, NICHOLAS W., *University of Illinois at Chicago, Chicago, Illinois*, Cartilage and Meniscus, Properties of

MASTERS, KRISTYN S., *University of Wisconsin, Madison, Wisconsin*, Tissue Engineering

MAUGHAN, RICHARD L., *Hospital of the University of Pennsylvania*, Neutron Beam Therapy

MCADAMS, ERIC, *University of Ulster at Jordanstown, Newtownabbey, Ireland*, Bioelectrodes

MCARTHUR, SALLY L., *University of Sheffield, Sheffield, United Kingdom*, Biomaterials, Surface Properties of

MCEWEN, MALCOM, *National Research Council of Canada, Ontario, Canada*, Radiation Dosimetry for Oncology

MCGOWAN, EDWARD J., *E.J. McGowan & Associates*, Biofeedback

MCGRATH, SUSAN, *Dartmouth College, Hanover, New Hampshire*, Oxygen Analyzers

MEEKS, SANFORD L., *University of Florida, Gainesville, Florida*, Radiosurgery, Stereotactic

MELISSA, PETER, *University of Central Florida NanoScience Technology Center, Orlando, Florida*, Biosurface Engineering

MENDELSON, YITZHAK, *Worcester Polytechnic Institute*, Optical Sensors

METZKER, MICHAEL L., *Baylor College of Medicine, Houston, Texas*, Polymerase Chain Reaction

MEYEREND, M.E., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

MICHLER, ROBERT, *Montefiore Medical Center, Bronx, New York*, Heart–Lung Machines

MICIC, MIODRAG, *MP Biomedicals LLC, Irvine, California*, Microscopy and Spectroscopy, Near-Field

MILLER, WILLIAM, *University of Missouri Columbia, Missouri*, Radionuclide Production and Radioactive Decay

MITTRA, ERIK, *Stony Brook University, New York*, Bone Density Measurement

MODELL, MARK, *Harvard Medical School, Boston, Massachusetts*, Fiber Optics in Medicine

MORE, ROBERT B., *RBMore Associates, Austin, Texas* Biomaterials Carbon

MORE, ROBERT, *Austin, Texas*, Heart Valves, Prosthetic

MORROW, DARREN, *Royal Adelaide Hospital, Adelaide, Australia*, Intraaortic Balloon Pump

MOURTADA, FIRAS, *MD Anderson Cancer Center, Houston, Texas*, Brachytherapy, Intravascular

MOY, VINCENT T., *University of Miami, Miller School of Medicine, Miami, Florida*, Microscopy, Scanning Force

MÜFTÜ, SINAN, *Northeastern University, Boston, Massachusetts*, Tooth and Jaw, Biomechanics of

MURPHY, RAYMOND L.H., Lung Sounds

MURPHY, WILLIAM L., *University of Wisconsin, Madison, Wisconsin*, Tissue Engineering

MURRAY, ALAN, *Newcastle University Medical Physics, Newcastle upon Tyne, United Kingdom*, Pace makers

MUTIC, SASA, *Washington University School of Medicine, St. Louis, Missouri*, Radiation Therapy Simulator

NARAYAN, ROGER J., *University of North Carolina, Chapel Hill, North Carolina*, Biomaterials, Corrosion and Wear of

NATALE, ANDREA, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

NAZERAN, HOMER, *The University of Texas, El Paso, Texas*, Electrocardiography, Computers in

NEUMAN, MICHAEL R., *Michigan Technological University, Houghton, Houghton, Michigan*, Fetal Monitoring, Neonatal Monitoring

NEUZIL, PAVEL, *Institute of Bioengineering and Nanotechnology, Singapore*, Immunologically Sensitive Field-Effect Transistors

NICKOLOFF, EDWARD L., *Columbia University, New York, New York*, X-Ray Quality Control Program

NIEZGODA, JEFFREY A., *Medical College of Wisconsin, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

NISHIKAWA, ROBERT M., *The University of Chicago, Chicago, Illinois*, Computer-Assisted Detection and Diagnosis

NUTTER, BRIAN, *Texas Tech University, Lubbock, Texas*, Medical Records, Computers in

O'DONOHUE, WILLIAM, *University of Nevada, Reno, Nevada*, Sexual Instrumentation

ORTON, COLIN, *Harper Hospital and Wayne State University, Detroit, Michigan*, Medical Physics Literature

OZCELIK, SELAHATTIN, *Texas A&M University, Kingsville, Texas*, Drug Infusion Systems

PANITCH, ALYSSA, *Arizona State University, Tempe, Arizona*, Biomaterials: An Overview

PAOLINO, DONATELLA, *University of Catanzaro Magna Græcia, Germaneto (CZ), Italy*, Drug Delivery Systems

PAPAIOANNOU, GEORGE, *University of Wisconsin, Milwaukee, Wisconsin*, Joints, Biomechanics of

PARK, GRACE E., *Purdue University, West Lafayette, Indiana*, Porous Materials for Biological Applications

PARMENTER, BRETT A., *State University of New York at Buffalo, Buffalo, New York*, Sleep Studies, Computer Analysis of

PATEL, DIMPI, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

PEARCE, JOHN, *The University of Texas, Austin, Texas*, Electrosurgical Unit (ESU)

PELET, SERGE, *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

PERIASAMY, AMMASI, *University of Virginia, Charlottesville, Virginia*, Cellular Imaging

PERSONS, BARBARA L., *University of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

PIPER, IAN, *The University of Memphis, Memphis, Tennessee*, Monitoring, Intracranial Pressure

POLETTO, CHRISTOPHER J., *National Institutes of Health*, Tactile Stimulation

PREMINGER, GLENN M., *Duke University Medical Center, Durham, North Carolina*, Lithotripsy

PRENDERGAST, PATRICK J., *Trinity College, Dublin, Ireland*, Orthopedics, Prosthesis Fixation for

PREVITE, MICHAEL, *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

PURDY, JAMES A., *UC Davis Medical Center, Sacramento, California*, Radiotherapy Accessories

QI, HAIRONG, *Advanced Concepts Analysis, Inc., Falls Church, Virginia*, Thermography

QIN, YIXIAN, *Stony Brook University, New York*, Bone Density Measurement

QUAN, STUART F., *University of Arizona, Tucson, Arizona*, Ventilatory Monitoring

QUIROGA, RODRIGO QUIAN, *University of Leicester, Leicester, United Kingdom*, Evoked Potentials

RAHAGHI, FARBOD N., *University of California, La Jolla, California*, Glucose Sensors

RAHKO, PETER S., *University of Wisconsin Medical School*, Echocardiography and Doppler Echocardiography

RALPH, LIETO, *University of Wisconsin–Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

RAMANATHAN, LAKSHMI, *Mount Sinai Medical Center*, Analytical Methods, Automated

RAO, SATISH S.C., *University of Iowa College of Medicine, Iowa City, Iowa*, Anorectal Manometry

RAPOPORT, DAVID M., *NYU School of Medicine, New York, New York*, Continuous Positive Airway Pressure

REBELLO, KEITH J., *The Johns Hopkins University Applied Physics Lab, Laurel, Maryland*, Micro surgery

REDDY, NARENDER, *The University of Akron, Akron, Ohio*, Linear Variable Differential Transformers

REN-DIH, SHEU, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

RENGACHARY, SETTI S., *Detroit, Michigan*, Human Spine, Biomechanics of

REPPERGER, DANIEL W., *Wright-Patterson Air Force Base, Dayton, Ohio*, Human Factors in Medical Devices

RITCHEY, ERIC R., *The Ohio State University, Columbus, Ohio*, Contact Lenses

RIVARD, MARK J., *Tufts New England Medical Center, Boston, Massachusetts*, Imaging Devices

ROBERTSON, J. DAVID, *University of Missouri, Columbia, Missouri*, Radionuclide Production and Radioactive Decay

ROTH, BRADLEY J., *Oakland University, Rochester, Michigan*, Defibrillators

ROWE-HORWEGE, R. WANDA, *University of Texas Medical School, Houston, Texas*, Hyperthermia, Systemic

RUMSEY, JOHN W., *University of Central Florida, Orlando, Florida*, Biosurface Engineering

RUTKOWSKI, GREGORY E., *University of Minnesota, Duluth, Minnesota*, Engineered Tissue

SALATA, O.V., *University of Oxford, Oxford, United Kingdom*, Nanoparticles

SAMARAS, THEODOROS, *Aristotle University of Thessaloniki Department of Physics, Thessaloniki, Greece*, Thermometry

SANGOLE, ARCHANA P., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

SARKOZI, LASZLO, *Mount Sinai School of Medicine*, Analytical Methods, Automated

SCHEK, HENRY III, *University of Michigan, Ann Arbor, Michigan*, Optical Tweezers

SCHMITZ, CHRISTOPH H., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

SCHUCKERS, STEPHANIE A.C., *Clarkson University, Potsdam, New York*, Arrhythmia Analysis, Automated

SCOPE, KENNETH, *Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

SCOTT, ADZICK N., *University of Pennsylvania, Philadelphia, Pennsylvania*, Intrauterine Surgical Techniques

SEAL, BRANDON L., *Arizona State University, Tempe, Arizona*, Biomaterials: An Overview

SEALE, GARY, *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

SEGERS, PATRICK, *Ghent University, Belgium*, Hemodynamics

SELIM, MOSTAFA A., *Cleveland Metropolitan General Hospital, Palm Coast, Florida*, Colposcopy

SETHI, ANIL, *Loyola University Medical Center, Maywood, Illinois*, X-Rays: Interaction with Matter

SEVERINGHAUS, JOHN W., *University of California in San Francisco*, CO_2 Electrodes

SHALODI, ABDELWAHAB D., *Cleveland Metropolitan General Hospital, Palm Coast, Florida*, Colposcopy

SHANMUGASUNDARAM, SHOBANA, *New Jersey Institute of Technology, Newark, New Jersey*, Polymeric Materials

SHARD, ALEXANDER G., *University of Sheffield, Sheffield United Kingdom*, Biomaterials, Surface Properties of

SHEN, LI-JIUAN, *National Taiwan University School of Pharmacy, Taipei, Taiwan*, Colorimetry

SHEN, WEI-CHIANG, *University of Southern California School of Pharmacy, Los Angeles, California*, Colorimetry

SHERAR, MICHAEL D., *London Health Sciences Centre and University of Western Ontario, London, Ontario, Canada*, Hyperthermia, Interstitial

SHERMAN, DAVID, *The Johns Hopkins University, Baltimore, Maryland*, Electroencephalography

SHI, DONGLU, *University of Cincinnati, Cincinnati, Ohio*, Biomaterials, Testing and Structural Properties of

SHUCARD, DAVID W.M., *State University of New York at Buffalo, Buffalo, New York*, Sleep Studies, Computer Analysis of

SIEDBAND, MELVIN P., *University of Wisconsin, Madison, Wisconsin*, Image Intensifiers and Fluoroscopy

SILBERMAN, HOWARD, *University of Southern California, Los Angeles, California*, Nutrition, Parenteral

SILVERMAN, GORDON, *Manhattan College*, Computers in the Biomedical Laboratory

SILVERN, DAVID A., *Medical Physics Unit, Rabin Medical Center, Petah Tikva, Israel*, Prostate Seed Implants

SINHA, PIYUSH, *The Ohio State University, Columbus, Ohio*, Drug Delivery Systems

SINHA, ABHIJIT ROY, *University of Cincinnati, Cincinnati, Ohio*, Coronary Angioplasty and Guidewire Diagnostics

SINKJÆR, THOMAS, *Aalborg University, Aalborg, Denmark*, Electroneurography

SLOAN, JEFFREY A., *Mayo Clinic, Rochester, Minnesota*, Quality-of-Life Measures, Clinical Significance of

SO, PETER T.C., *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

SOBOL, WLAD T., *University of Alabama at Birmingham Health System, Birmingham, Alabama*, Nuclear Magnetic Resonance Spectroscopy

SOOD, SANDEEP, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

SPECTOR, MYRON, *Brigham and Women's Hospital, Boston, Massachusetts*, Biocompatibility of Materials

SPELMAN, FRANCIS A., *University of Washington*, Cochlear Prostheses

SRINIVASAN, YESHWANTH, *Texas Tech University, Lubbock, Texas*, Medical Records, Computers in

SRIRAM, NEELAMEGHAM, *University of Buffalo, Buffalo, New York*, Cell Counters, Blood

STARKO, KENTON R., *Point Roberts, Washington*, Physiological Systems Modeling

STARKSCHALL, GEORGE, *The University of Texas*, Radiotherapy, Three-Dimensional Conformal

STAVREV, PAVEL, *Cross Cancer Institute, Edmonton, Alberta, Canada*, Radiotherapy Treatment Planning, Optimization of

STENKEN, JULIE A., *Rensselaer Polytechnic Institute, Troy, New York*, Microdialysis Sampling

STIEFEL, ROBERT, *University of Maryland Medical Center, Baltimore, Maryland*, Equipment Acquisition

STOKES, I.A.F., *Polytechniquie Montreal, Montreal Quebec, Canada*, Scoliosis, Biomechanics of

STONE, M.H., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

SU, XIAO-LI, *BioDetection Instruments LLC, Fayetteville, Arkansas*, Piezoelectric Sensors

SUBHAN, ARIF, *Masterplan Technology Management, Chatsworth, California*, Equipment Maintenance, Biomedical

SWEENEY, JAMES D., *Arizona State University, Tempe, Arizona*, Functional Electrical Stimulation

SZETO, ANDREW Y.J., *San Diego State University, San Diego, California*, Blind and Visually Impaired, Assistive Technology for

TAKAYAMA, SHUICHI, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

TAMUL, PAUL C., *Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

TAMURA, TOSHIYO, *Chiba University School of Engineering, Chiba, Japan*, Home Health Care Devices

TANG, XIANGYANG, *GE Healthcare Technologies, Wankesha, Wisconsin*, Computed Tomography Simulators

TAYLOR, B.C., *The University of Akron, Akron, Ohio*, Cardiac Output, Indicator Dilution Measurement of

TEMPLE, RICHARD O., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

TEN, STANLEY, *Salt Lake City, Utah*, Electroanalgesia, Systemic

TERRY, TERESA M., *Walter Reed Army Institute of Research, Rockville, Maryland*, Blood Collection and Processing

THAKOR, N.V., *Johns Hopkins University, Baltimore, Maryland*, Neurological Monitors

THIERENS, HUBERT M.A., *University of Ghent, Ghent, Belgium*, Radiopharmaceutical Dosimetry

THOMADSEN, BRUCE, *University of Wisconsin–Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

TIPPER, J.L., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

TOGAWA, TATSUO, *Waseda University, Saitama, Japan*, Integrated Circuit Temperature Sensor

TORNAI, MARTIN, *Duke University, Durham, North Carolina*, X-Ray Equipment Design

TRAN-SON-TAY, ROGER, *University of Florida, Gainesville, Florida*, Blood Rheology

YENI, YENER N., *Henry Ford Hospital, Detroit, Michigan*, Joints, Biomechanics of

YLI-HANKALA, ARVI, *Tampere University of Technology, Pori, Finland*, Monitoring in Anesthesia

YOKO, KAMOTANI, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

YOON, KANG JI, *Korea Institute of Science and Technology, Seoul, Korea*, Micropower for Medical Applications

YORKE, ELLEN, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

YOSHIDA, KEN, *Aalborg University, Aalborg, Denmark*, Electroneurography

YOUNGSTEDT, SHAWN D., *University of South Carolina, Columbia, South Carolina*, Sleep Laboratory

YU, YIH-CHOUNG, *Lafayette College, Easton, Pennsylvania*, Blood Pressure, Automatic Control of

ZACHARIAH, EMMANUEL S., *University of Medicine and Dentistry of New Jersey, New Brunswick, New Jersey*, Immunologically Sensitive Field-Effect Transistors

ZAIDER, MARCO, *Memorial Sloan Kettering Cancer Center, New York, New York*, Prostate Seed Implants

ZAPANTA, CONRAD M., *Penn State College of Medicine, Hershey, Pennsylvania*, Heart, Artificial

ZARDENETA, GUSTAVO, *University of Texas, San Antonio, Texas*, Fluorescence Measurements

ZELMANOVIC, DAVID, *Bayer HealthCare LLC, Tarrytown, New York*, Differential Counts, Automated

ZHANG, MIN, *University of Washington, Seattle, Washington*, Biomaterials: Polymers

ZHANG, YI, *University of Buffalo, Buffalo, New York*, Cell Counters, Blood

ZHU, XIAOYUE, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

ZIAIE, BABAK, *Purdue University, W. Lafayette, Indiana*, Biotelemetry

ZIELINSKI, TODD M., *Medtronic, Inc., Minneapolis, Minnesota*, Bioimpedance in Cardiovascular Medicine

ZIESSMAN, HARVEY A., *Johns Hopkins University*, Computed Tomography, Single Photon Emission

PREFACE

This six-volume work is an alphabetically organized compilation of almost 300 articles that describe critical aspects of medical devices and instrumentation.

It is comprehensive. The articles emphasize the contributions of engineering, physics, and computers to each of the general areas of anesthesiology, biomaterials, burns, cardiology, clinical chemistry, clinical engineering, communicative disorders, computers in medicine, critical care medicine, dermatology, dentistry, ear, nose, and throat, emergency medicine, endocrinology, gastroenterology, genetics, geriatrics, gynecology, hematology, heptology, internal medicine, medical physics, microbiology, nephrology, neurology, nutrition, obstetrics, oncology, ophthalmology, orthopedics, pain, pediatrics, peripheral vascular disease, pharmacology, physical therapy, psychiatry, pulmonary medicine, radiology, rehabilitation, surgery, tissue engineering, transducers, and urology.

The discipline is defined through the synthesis of the core knowledge from all the fields encompassed by the application of engineering, physics, and computers to problems in medicine. The articles focus not only on what is now useful but also on what is likely to be useful in future medical applications.

These volumes answer the question, "What are the branches of medicine and how does technology assist each of them?" rather than "What are the branches of technology and how could each be used in medicine?" To keep this work to a manageable length, the practice of medicine that is unassisted by devices, such as the use of drugs to treat disease, has been excluded.

The articles are accessible to the user; each benefits from brevity of condensation instead of what could easily have been a book-length work. The articles are designed not for peers, but rather for workers from related fields who wish to take a first look at what is important in the subject.

The articles are readable. They do not presume a detailed background in the subject, but are designed for any person with a scientific background and an interest in technology. Rather than attempting to teach the basics of physiology or Ohm's law, the articles build on such basic concepts to show how the worlds of life science and physical science meld to produce improved systems. While the ideal reader might be a person with a Master's degree in biomedical engineering or medical physics or an M.D. with a physical science undergraduate degree, much of the material will be of value to others with an interest in this growing field. High school students and hospital patients can skip over more technical areas and still gain much from the descriptive presentations.

The *Encyclopedia of Medical Devices and Instrumentation* is excellent for browsing and searching for those new divergent associations that may advance work in a peripheral field. While it can be used as a reference for facts, the articles are long enough that they can serve as an educational instrument and provide genuine understanding of a subject.

One can use this work just as one would use a dictionary, since the articles are arranged alphabetically by topic. Cross references assist the reader looking for subjects listed under slightly different names. The index at the end leads the reader to all articles containing pertinent information on any subject. Listed on pages xxi to xxx are all the abbreviations and acronyms used in the *Encyclopedia.* Because of the increasing use of SI units in all branches of science, these units are provided throughout the *Encyclopedia* articles as well as on pages xxxi to xxxv in the section on conversion factors and unit symbols.

I owe a great debt to the many people who have contributed to the creation of this work. At John Wiley & Sons, Encyclopedia Editor George Telecki provided the idea and guiding influence to launch the project. Sean Pidgeon was Editorial Director of the project. Assistant Editors Roseann Zappia, Sarah Harrington, and Surlan Murrell handled the myriad details of communication between publisher, editor, authors, and reviewers and stimulated authors and reviewers to meet necessary deadlines.

My own background has been in the electrical aspects of biomedical engineering. I was delighted to have the assistance of the editorial board to develop a comprehensive encyclopedia. David J. Beebe suggested cellular topics such as microfluidics. Jerry M. Calkins assisted in defining the chemically related subjects, such as anesthesiology. Michael R. Neuman suggested subjects related to sensors, such as in his own work—neonatology. Joon B. Park has written extensively on biomaterials and suggested related subjects. Edward S. Sternick provided many suggestions from medical physics. The Editorial Board was instrumental both in defining the list of subjects and in suggesting authors.

This second edition brings the field up to date. It is available on the web at http://www.mrw.interscience.wiley.com/emdi, where articles can be searched simultaneously to provide rapid and comprehensive information on all aspects of medical devices and instrumentation.

JOHN G. WEBSTER
University of Wisconsin, Madison

LIST OF ARTICLES

ABBREVIATIONS AND ACRONYMS

AAMI	Association for the Advancement of Medical Instrumentation
AAPM	American Association of Physicists in Medicine
ABC	Automatic brightness control
ABET	Accreditation board for engineering training
ABG	Arterial blood gases
ABLB	Alternative binaural loudness balance
ABS	Acrylonitrile–butadiene–styrene
ac	Alternating current
AC	Abdominal circumference; Affinity chromatography
ACA	Automated clinical analyzer
ACES	Augmentative communication evaluation system
ACL	Anterior chamber lens
ACLS	Advanced cardiac life support
ACOG	American College of Obstetrics and Gynecology
ACR	American College of Radiology
ACS	American Cancer Society; American College of Surgeons
A/D	Analog-to-digital
ADC	Agar diffusion chambers; Analog-to-digital converter
ADCC	Antibody-dependent cellular cytotoxicity
ADCL	Accredited Dosimetry Calibration Laboratories
ADP	Adenosine diphosphate
A-D-T	Admission, discharge, and transfer
AE	Anion exchange; Auxiliary electrode
AEA	Articulation error analysis
AEB	Activation energy barrier
AEC	Automatic exposure control
AED	Automatic external defibrillator
AEMB	Alliance for Engineering in Medicine and Biology
AES	Auger electron spectroscopy
AESC	American Engineering Standards Committee
AET	Automatic exposure termination
AFO	Ankle-foot orthosis
AGC	Automatic gain control
AHA	American Heart Association
AI	Arterial insufficiency
AICD	Automatic implantable cardiac defibrillator
AID	Agency for International Development
AIDS	Acquired immune deficiency syndrome
AL	Anterior leaflet
ALG	Antilymphocyte globulin
ALS	Advanced life support; Amyotropic lateral sclerosis
ALT	Alanine aminotransferase
ALU	Arithmetic and logic unit
AM	Amplitude modulation
AMA	American Medical Association
amu	Atomic mass units
ANOVA	Analysis of variance
ANSI	American National Standards Institute
AP	Action potential; Alternative pathway; Anteroposterior
APD	Anterioposterior diameter
APL	Adjustable pressure limiting valve; Applied Physics Laboratory
APR	Anatomically programmed radiography
AR	Amplitude reduction; Aortic regurgitation; Autoregressive
Ara-C	Arabinosylcytosine
ARD	Absorption rate density
ARDS	Adult respiratory distress syndrome
ARGUS	Arrhythmia guard system
ARMA	Autoregressive-moving-average model
ARMAX	Autoregressive-moving-average model with external inputs
AS	Aortic stenosis
ASA	American Standards Association
ASCII	American standard code for information interchange
ASD	Antisiphon device
ASHE	American Society for Hospital Engineering
ASTM	American Society for Testing and Materials
AT	Adenosine-thiamide; Anaerobic threshold; Antithrombin
ATA	Atmosphere absolute
ATLS	Advanced trauma life support
ATN	Acute tubular necrosis
ATP	Adenosine triphosphate
ATPD	Ambient temperature pressure dry
ATPS	Ambient temperature pressure saturated
ATR	Attenuated total reflection
AUC	Area under curve
AUMC	Area under moment curve
AV	Atrioventricular
AZT	Azido thymidine
BA	Biliary atresia
BAEP	Brainstem auditory evoked potential
BAPN	Beta-amino-proprionitryl
BAS	Boston anesthesis system
BASO	Basophil
BB	Buffer base
BBT	Basal body temperature

BCC	Body-centered cubic
BCD	Binary-coded decimal
BCG	Ballistocardiogram
BCLS	Basic cardiac life support
BCRU	British Commitee on Radiation Units and Measurements
BDI	Beck depression inventory
BE	Base excess; Binding energy
BET	Brunauer, Emmett, and Teller methods
BH	His bundle
BI	Biological indicators
BIH	Beth Israel Hospital
BIPM	International Bureau of Weights and Measurements
BJT	Bipolar junction transistor
BMDP	Biomedical Programs
BME	Biomedical engineering
BMET	Biomedical equipment technician
BMO	Biomechanically optimized
BMR	Basal metabolic rate
BOL	Beginning of life
BP	Bereitschafts potential; Break point
BR	Polybutadiene
BRM	Biological response modifier
BRS	Bibliographic retrieval services
BSS	Balanced salt solution
BTG	Beta thromboglobulin
BTPS	Body temperature pressure saturated
BUN	Blood urea nitrogen
BW	Body weight
CA	Conductive adhesives
CABG	Coronary artery by-pass grafting
CAD/CAM	Computer-aided design/computer-aided manufacturing
CAD/D	Computer-aided drafting and design
CADD	Central axis depth dose
CAI	Computer assisted instruction; Computer-aided instruction
CAM	Computer-assisted management
cAMP	Cyclic AMP
CAPD	Continuous ambulatory peritoneal dialysis
CAPP	Child amputee prosthetic project
CAT	Computerized axial tomography
CATS	Computer-assisted teaching system; Computerized aphasia treatment system
CAVH	Continuous arteriovenous hemofiltration
CB	Conjugated bilirubin; Coulomb barrier
CBC	Complete blood count
CBF	Cerebral blood flow
CBM	Computer-based management
CBV	Cerebral blood volume
CC	Closing capacity
CCC	Computer Curriculum Company
CCD	Charge-coupled device
CCE	Capacitance contact electrode
CCF	Cross-correlation function
CCL	Cardiac catheterization laboratory
CCM	Critical care medical services
CCPD	Continuous cycling peritoneal dialysis
CCTV	Closed circuit television system
CCU	Coronary care unit; Critical care unit
CD	Current density
CDR	Complimentary determining region
CDRH	Center for Devices and Radiological Health
CEA	Carcinoembryonic antigen
CF	Conversion factor; Cystic fibrosis
CFC	Continuous flow cytometer
CFR	Code of Federal Regulations
CFU	Colony forming units
CGA	Compressed Gas Association
CGPM	General Conference on Weights and Measures
CHO	Carbohydrate
CHO	Chinese hamster ovary
CI	Combination index
CICU	Cardiac intensive care unit
CIF	Contrast improvement factor
CIN	Cervical intraepithelial neoplasia
CK	Creatine kinase
CLAV	Clavicle
CLSA	Computerized language sample analysis
CM	Cardiomyopathy; Code modulation
CMAD	Computer managed articulation diagnosis
CMI	Computer-managed instruction
CMRR	Common mode rejection ratio
CMV	Conventional mechanical ventilation; Cytomegalovirus
CNS	Central nervous system
CNV	Contingent negative variation
CO	Carbon monoxide; Cardiac output
COBAS	Comprehensive Bio-Analysis System
COPD	Chronic obstructive pulmonary disease
COR	Center of rotation
CP	Cerebral palsy; Closing pressure; Creatine phosphate
CPB	Cardiopulmonary bypass
CPET	Cardiac pacemaker electrode tips
CPM	Computerized probe measurements
CPP	Cerebral perfusion pressure; Cryoprecipitated plasma
CPR	Cardiopulmonary resuscitation
cps	Cycles per second
CPU	Central Processing unit
CR	Center of resistance; Conditioned response; Conductive rubber; Creatinine
CRBB	Complete right bundle branch block
CRD	Completely randomized design
CRL	Crown rump length
CRT	Cathode ray tube
CS	Conditioned stimulus; Contrast scale; Crown seat
CSA	Compressed spectral array
CSF	Cerebrospinal fluid
CSI	Chemical shift imaging
CSM	Chemically sensitive membrane
CT	Computed tomography; Computerized tomography
CTI	Cumulative toxicity response index
CV	Closing volume

C.V.	Coefficient of variation
CVA	Cerebral vascular accident
CVP	Central venous pressure
CVR	Cardiovascular resistance
CW	Continuous wave
CWE	Coated wire electrodes
CWRU	Case Western Reserve University
DAC	Digital-to-analog converter
DAS	Data acquisition system
dB	Decibel
DB	Direct body
DBMS	Data base management system
DBS	Deep brain stimulation
dc	Direct current
DCCT	Diabetes control and complications trial
DCP	Distal cavity pressure
DCS	Dorsal column stimulation
DDC	Deck decompression chamber
DDS	Deep diving system
DE	Dispersive electrode
DEN	Device experience network
DERS	Drug exception ordering system
DES	Diffuse esophageal spasm
d.f.	Distribution function
DHCP	Distributed Hospital Computer Program
DHE	Dihematoporphyrin ether
DHEW	Department of Health Education and Welfare
DHHS	Department of Health and Human Services
DHT	Duration of hypothermia
DI	Deionized water
DIC	Displacement current
DIS	Diagnostic interview schedule
DL	Double layer
DLI	Difference lumen for intensity
DM	Delta modulation
DME	Dropping mercury electrode
DN	Donation number
DNA	Deoxyribonucleic acid
DOF	Degree of freedom
DOS	Drug ordering system
DOT-NHTSA	Department of Transportation Highway Traffic Safety Administration
DPB	Differential pencil beam
DPG	Diphosphoglycerate
DQE	Detection quantum efficiency
DRESS	Depth-resolved surface coil spectroscopy
DRG	Diagnosis-related group
DSA	Digital subtraction angiography
DSAR	Differential scatter-air ratio
DSB	Double strand breaks
DSC	Differential scanning calorimetry
D-T	Deuterium-on-tritium
DTA	Differential thermal analysis
d.u.	Density unit
DUR	Duration
DVT	Deep venous thrombosis
EA	Esophageal accelerogram
EB	Electron beam
EBCDIC	Extended binary code decimal interchange code
EBS	Early burn scar
EBV	Epstein–Barr Virus
EC	Ethyl cellulose
ECC	Emergency cardiac care; Extracorporeal circulation
ECCE	Extracapsular cataract extinction
ECD	Electron capture detector
ECG	Electrocardiogram
ECM	Electrochemical machining
ECMO	Extracorporeal membrane oxygenation
ECOD	Extracranial cerebrovascular occlusive disease
ECRI	Emergency Care Research Institute
ECS	Exner's Comprehensive System
ECT	Electroconvulsive shock therapy; Electroconvulsive therapy; Emission computed tomography
EDD	Estimated date of delivery
EDP	Aortic end diastolic pressure
EDTA	Ethylenediaminetetraacetic acid
EDX	Energy dispersive X-ray analysis
EEG	Electroencephalogram
EEI	Electrode electrolyte interface
EELV	End-expiratory lung volume
EER	Electrically evoked response
EF	Ejection fraction
EF	Electric field; Evoked magnetic fields
EFA	Estimated fetal age
EGF	Epidermal growth factor
EGG	Electrogastrogram
EIA	Enzyme immunoassay
EIU	Electrode impedance unbalance
ELF	Extra low frequency
ELGON	Electrical goniometer
ELISA	Enzyme-linked immunosorbent assay
ELS	Energy loss spectroscopy
ELV	Equivalent lung volume
EM	Electromagnetic
EMBS	Engineering in Medicine and Biology Society
emf	Electromotive force
EMG	Electromyogram
EMGE	Integrated electromyogram
EMI	Electromagnetic interference
EMS	Emergency medical services
EMT	Emergency medical technician
ENT	Ear, nose, and throat
EO	Elbow orthosis
EOG	Electrooculography
EOL	End of life
EOS	Eosinophil
EP	Elastoplastic; Evoked potentiate
EPA	Environmental protection agency
ER	Evoked response
ERCP	Endoscopic retrograde cholangiopancreatography
ERG	Electron radiography; Electroretinogram
ERMF	Event-related magnetic field
ERP	Event-related potential
ERV	Expiratory reserve volume

ESCA	Electron spectroscopy for chemical analysis
ESI	Electrode skin impedance
ESRD	End-stage renal disease
esu	Electrostatic unit
ESU	Electrosurgical unit
ESWL	Extracorporeal shock wave lithotripsy
ETO, Eto	Ethylene oxide
ETT	Exercise tolerance testing
EVA	Ethylene vinyl acetate
EVR	Endocardial viability ratio
EW	Extended wear
FAD	Flavin adenine dinucleotide
FARA	Flexible automation random analysis
FBD	Fetal biparietal diameter
FBS	Fetal bovine serum
fcc	Face centered cubic
FCC	Federal Communications Commission
Fct	Fluorocrit
FDA	Food and Drug Administration
FDCA	Food, Drug, and Cosmetic Act
FE	Finite element
FECG	Fetal electrocardiogram
FEF	Forced expiratory flow
FEL	Free electron lasers
FEM	Finite element method
FEP	Fluorinated ethylene propylene
FES	Functional electrical stimulation
FET	Field-effect transistor
FEV	Forced expiratory volume
FFD	Focal spot to film distance
FFT	Fast Fourier transform
FGF	Fresh gas flow
FHR	Fetal heart rate
FIC	Forced inspiratory capacity
FID	Flame ionization detector; Free-induction decay
FIFO	First-in-first-out
FITC	Fluorescent indicator tagged polymer
FL	Femur length
FM	Frequency modulation
FNS	Functional neuromuscular stimulation
FO	Foramen ovale
FO-CRT	Fiber optics cathode ray tube
FP	Fluorescence polarization
FPA	Fibrinopeptide A
FR	Federal Register
FRC	Federal Radiation Council; Functional residual capacity
FSD	Focus-to-surface distance
FTD	Focal spot to tissue-plane distance
FTIR	Fourier transform infrared
FTMS	Fourier transform mass spectrometer
FU	Fluorouracil
FUDR	Floxuridine
FVC	Forced vital capacity
FWHM	Full width at half maximum
FWTM	Full width at tenth maximum
GABA	Gamma amino buteric acid
GAG	Glycosaminoglycan
GBE	Gas-bearing electrodynamometer
GC	Gas chromatography; Guanine-cytosine
GDT	Gas discharge tube
GFR	Glomerular filtration rate
GHb	Glycosylated hemoglobin
GI	Gastrointestinal
GLC	Gas–liquid chromatography
GMV	General minimum variance
GNP	Gross national product
GPC	Giant papillary conjunctivitis
GPH	Gas-permeable hard
GPH-EW	Gas-permeable hard lens extended wear
GPO	Government Printing Office
GSC	Gas-solid chromatography
GSR	Galvanic skin response
GSWD	Generalized spike-wave discharge
HA	Hydroxyapatite
HAM	Helical axis of motion
Hb	Hemoglobin
HBE	His bundle electrogram
HBO	Hyperbaric oxygenation
HC	Head circumference
HCA	Hypothermic circulatory arrest
HCFA	Health care financing administration
HCL	Harvard Cyclotron Laboratory
hcp	Hexagonal close-packed
HCP	Half cell potential
HDPE	High density polyethylene
HECS	Hospital Equipment Control System
HEMS	Hospital Engineering Management System
HEPA	High efficiency particulate air filter
HES	Hydroxyethylstarch
HETP	Height equivalent to a theoretical plate
HF	High-frequency; Heating factor
HFCWO	High-frequency chest wall oscillation
HFER	High-frequency electromagnetic radiation
HFJV	High-frequency jet ventilation
HFO	High-frequency oscillator
HFOV	High-frequency oscillatory ventilation
HFPPV	High-frequency positive pressure ventilation
HFV	High-frequency ventilation
HHS	Department of Health and Human Services
HIBC	Health industry bar code
HIMA	Health Industry Manufacturers Association
HIP	Hydrostatic indifference point
HIS	Hospital information system
HK	Hexokinase
HL	Hearing level
HMBA	Hexamethylene bisacetamide
HMO	Health maintenance organization
HMWPE	High-molecular-weight polyethylene
HOL	Higher-order languages
HP	Heating factor; His-Purkinje
HpD	Hematoporphyrin derivative
HPLC	High-performance liquid chromatography
HPNS	High-pressure neurological syndrome
HPS	His-Purkinje system
HPX	High peroxidase activity

HR	Heart rate; High-resolution
HRNB	Halstead-Reitan Neuropsychological Battery
H/S	Hard/soft
HSA	Human serum albumin
HSG	Hysterosalpingogram
HTCA	Human tumor cloning assay
HTLV	Human T cell lymphotrophic virus
HU	Heat unit; Houndsfield units; Hydroxyurea
HVL	Half value layer
HVR	Hypoxic ventilatory response
HVT	Half-value thickness
IA	Image intensifier assembly; Inominate artery
IABP	Intraaortic balloon pumping
IAEA	International Atomic Energy Agency
IAIMS	Integrated Academic Information Management System
IASP	International Association for the Study of Pain
IC	Inspiratory capacity; Integrated circuit
ICCE	Intracapsular cataract extraction
ICD	Intracervical device
ICDA	International classification of diagnoses
ICL	Ms-clip lens
ICP	Inductively coupled plasma; Intracranial pressure
ICPA	Intracranial pressure amplitude
ICRP	International Commission on Radiological Protection
ICRU	International Commission on Radiological Units and Measurements
ICU	Intensive care unit
ID	Inside diameter
IDDM	Insulin dependent diabetes mellitus
IDE	Investigational device exemption
IDI	Index of inspired gas distribution
I:E	Inspiratory: expiratory
IEC	International Electrotechnical Commission; Ion-exchange chromatography
IEEE	Institute of Electrical and Electronics Engineers
IEP	Individual educational program
BETS	Inelastic electron tunneling spectroscopy
IF	Immunofluorescent
IFIP	International Federation for Information Processing
IFMBE	International Federation for Medical and Biological Engineering
IGFET	Insulated-gate field-effect transistor
IgG	Immunoglobulin G
IgM	Immunoglobulin M
IHP	Inner Helmholtz plane
IHSS	Idiopathic hypertrophic subaortic stenosis
II	Image intensifier
IIIES	Image intensifier input-exposure sensitivity
IM	Intramuscular
IMFET	Immunologically sensitive field-effect transistor
IMIA	International Medical Informatics Association
IMS	Information management system
IMV	Intermittent mandatory ventilation
INF	Interferon
IOL	Intraocular lens
IPC	Ion-pair chromatography
IPD	Intermittent peritoneal dialysis
IPG	Impedance plethysmography
IPI	Interpulse interval
IPPB	Intermittent positive pressure breathing
IPTS	International practical temperature scale
IR	Polyisoprene rubber
IRB	Institutional Review Board
IRBBB	Incomplete right bundle branch block
IRPA	International Radiation Protection Association
IRRAS	Infrared reflection-absorption spectroscopy
IRRS	Infrared reflection spectroscopy
IRS	Internal reflection spectroscopy
IRV	Inspiratory reserve capacity
IS	Image size; Ion-selective
ISC	Infant skin servo control
ISDA	Instantaneous screw displacement axis
ISE	Ion-selective electrode
ISFET	Ion-sensitive field effect transistor
ISIT	Intensified silicon-intensified target tube
ISO	International Organization for Standardization
ISS	Ion scattering spectroscopy
IT	Intrathecal
ITEP	Institute of Theoretical and Experimental Physics
ITEPI	Instantaneous trailing edge pulse impedance
ITLC	Instant thin-layer chromatography
IUD	Intrauterine device
IV	Intravenous
IVC	Inferior vena cava
IVP	Intraventricular pressure
JCAH	Joint Commission on the Accreditation of Hospitals
JND	Just noticeable difference
JRP	Joint replacement prosthesis
KB	Kent bundle
Kerma	Kinetic energy released in unit mass
KO	Knee orthosis
KPM	Kilopond meter
KRPB	Krebs-Ringer physiological buffer
LA	Left arm; Left atrium
LAD	Left anterior descending; Left axis deviation
LAE	Left atrial enlargement
LAK	Lymphokine activated killer
LAL	Limulus amoebocyte lysate
LAN	Local area network
LAP	Left atrial pressure
LAT	Left anterior temporalis
LBBB	Left bundle branch block
LC	Left carotid; Liquid chromatography

LCC	Left coronary cusp
LCD	Liquid crystal display
LDA	Laser Doppler anemometry
LDF	Laser Doppler flowmetry
LDH	Lactate dehydrogenase
LDPE	Low density polyethylene
LEBS	Low-energy brief stimulus
LED	Light-emitting diode
LEED	Low energy electron diffraction
LES	Lower esophageal sphincter
LESP	Lower esophageal sphincter pressure
LET	Linear energy transfer
LF	Low frequency
LH	Luteinizing hormone
LHT	Local hyperthermia
LL	Left leg
LLDPE	Linear low density polyethylene
LLPC	Liquid-liquid partition chromatography
LLW	Low-level waste
LM	Left masseter
LNNB	Luria-Nebraska Neuropsychological Battery
LOS	Length of stay
LP	Late potential; Lumboperitoneal
LPA	Left pulmonary artery
LPC	Linear predictive coding
LPT	Left posterior temporalis
LPV	Left pulmonary veins
LRP	Late receptor potential
LS	Left subclavian
LSC	Liquid-solid adsorption chromatography
LSI	Large scale integrated
LSV	Low-amplitude shear-wave viscoelastometry
LTI	Low temperature isotropic
LUC	Large unstained cells
LV	Left ventricle
LVAD	Left ventricular assist device
LVDT	Linear variable differential transformer
LVEP	Left ventricular ejection period
LVET	Left ventricular ejection time
LVH	Left ventricular hypertrophy
LYMPH	Lymphocyte
MAA	Macroaggregated albumin
MAC	Minimal auditory capabilities
MAN	Manubrium
MAP	Mean airway pressure; Mean arterial pressure
MAST	Military assistance to safety and traffic
MBA	Monoclonal antibody
MBV	Maximum breathing ventilation
MBX	Monitoring branch exchange
MCA	Methyl cryanoacrylate
MCG	Magnetocardiogram
MCI	Motion Control Incorporated
MCMI	Millon Clinical Multiaxial Inventory
MCT	Microcatheter transducer
MCV	Mean corpuscular volume
MDC	Medical diagnostic categories
MDI	Diphenylmethane diisocyanate; Medical Database Informatics
MDP	Mean diastolic aortic pressure
MDR	Medical device reporting
MDS	Multidimensional scaling
ME	Myoelectric
MED	Minimum erythema dose
MEDPAR	Medicare provider analysis and review
MEFV	Maximal expiratory flow volume
MEG	Magnetoencephalography
MeSH	Medline subject heading
METS	Metabolic equivalents
MF	Melamine-formaldehyde
MFP	Magnetic field potential
MGH	Massachusetts General Hospital
MHV	Magnetic heart vector
MI	Myocardial infarction
MIC	Minimum inhibitory concentration
MIFR	Maximum inspiratory flow rate
MINET	Medical Information Network
MIR	Mercury-in-rubber
MIS	Medical information system; Metal-insulator-semiconductor
MIT	Massachusetts Institute of Technology
MIT/BIH	Massachusetts Institute of Technology/ Beth Israel Hospital
MMA	Manual metal arc welding
MMA	Methyl methacrylate
MMECT	Multiple-monitored ECT
MMFR	Maximum midexpiratory flow rate
mm Hg	Millimeters of mercury
MMPI	Minnesota Multiphasic Personality Inventory
MMSE	Minimum mean square error
MO	Membrane oxygenation
MONO	Monocyte
MOSFET	Metal oxide silicon field-effect transistor
MP	Mercaptopurine; Metacarpal-phalangeal
MPD	Maximal permissible dose
MR	Magnetic resonance
MRG	Magnetoretinogram
MRI	Magnetic resonance imaging
MRS	Magnetic resonance spectroscopy
MRT	Mean residence time
MS	Mild steel; Multiple sclerosis
MSR	Magnetically shielded room
MTBF	Mean time between failure
MTF	Modulation transfer function
MTTR	Mean time to repair
MTX	Methotroxate
MUA	Motor unit activity
MUAP	Motor unit action potential
MUAPT	Motor unit action potential train
MUMPI	Missouri University Multi-Plane Imager
MUMPS	Massachusetts General Hospital utility multiuser programming system
MV	Mitral valve
MVO_2	Maximal oxygen uptake
MVTR	Moisture vapor transmission rate
MVV	Maximum voluntary ventilation
MW	Molecular weight

NAA	Neutron activation analysis
NAD	Nicotinamide adenine dinucleotide
NADH	Nicotinamide adenine dinucleotide, reduced form
NADP	Nicotinamide adenine dinucleotide phosphate
NAF	Neutrophil activating factor
NARM	Naturally occurring and accelerator-produced radioactive materials
NBB	Normal buffer base
NBD	Neuromuscular blocking drugs
N-BPC	Normal bonded phase chromatography
NBS	National Bureau of Standards
NCC	Noncoronary cusp
NCCLS	National Committee for Clinical Laboratory Standards; National Committee on Clinical Laboratory Standards
NCRP	National Council on Radiation Protection
NCT	Neutron capture theory
NEEP	Negative end-expiratory pressure
NEMA	National Electrical Manufacturers Association
NEMR	Nonionizing electromagnetic radiation
NEQ	Noise equivalent quanta
NET	Norethisterone
NEUT	Neutrophil
NFPA	National Fire Protection Association
NH	Neonatal hepatitis
NHE	Normal hydrogen electrode
NHLBI	National Heart, Lung, and Blood Institute
NIR	Nonionizing radiation
NIRS	National Institute for Radiologic Science
NK	Natural killer
NMJ	Neuromuscular junction
NMOS	N-type metal oxide silicon
NMR	Nuclear magnetic resonance
NMS	Neuromuscular stimulation
NPH	Normal pressure hydrocephalus
NPL	National Physical Laboratory
NR	Natural rubber
NRC	Nuclear Regulatory Commission
NRZ	Non-return-to-zero
NTC	Negative temperature coefficient
NTIS	National Technical Information Service
NVT	Neutrons versus time
NYHA	New York Heart Association
ob/gyn	Obstetrics and gynecology
OCR	Off-center ratio; Optical character recognition
OCV	Open circuit voltage
OD	Optical density; Outside diameter
ODC	Oxyhemoglobin dissociation curve
ODT	Oxygen delivery truck
ODU	Optical density unit
OER	Oxygen enhancement ratio
OFD	Object to film distance; Occiputo-frontal diameter
OHL	Outer Helmholtz layer
OHP	Outer Helmholtz plane
OIH	Orthoiodohippurate
OPG	Ocular pneumoplethysmography
OR	Operating room
OS	Object of known size; Operating system
OTC	Over the counter
OV	Offset voltage
PA	Posterioanterior; Pulmonary artery; Pulse amplitude
PACS	Picture archiving and communications systems
PAD	Primary afferent depolarization
PAM	Pulse amplitude modulation
PAN	Polyacrylonitrile
PAP	Pulmonary artery pressure
PAR	Photoactivation ratio
PARFR	Program for Applied Research on Fertility Regulation
PARR	Poetanesthesia recovery room
PAS	Photoacoustic spectroscopy
PASG	Pneumatic antishock garment
PBI	Penile brachial index
PBL	Positive beam limitation
PBT	Polybutylene terephthalate
PC	Paper chromatography; Personal computer; Polycarbonate
PCA	Patient controlled analgesia; Principal components factor analysis
PCG	Phonocardiogram
PCI	Physiological cost index
PCL	Polycaprolactone; Posterior chamber lens
PCR	Percent regurgitation
PCRC	Perinatal Clinical Research Center
PCS	Patient care system
PCT	Porphyria cutanea tarda
PCWP	Pulmonary capillary wedge pressure
PD	Peritoneal dialysis; Poly-p-dioxanone; Potential difference; Proportional and derivative
PDD	Percent depth dose; Perinatal Data Directory
PDE	Pregelled disposable electrodes
p.d.f.	Probability density function
PDL	Periodontal ligament
PDM	Pulse duration modulation
PDMSX	Polydimethyl siloxane
PDS	Polydioxanone
PE	Polyethylene
PEEP	Positive end-expiratory pressure
PEFR	Peak expiratory now rate
PEN	Parenteral and enteral nutrition
PEP	Preejection period
PEPPER	Programs examine phonetic find phonological evaluation records
PET	Polyethylene terephthalate; Positron-emission tomography
PEU	Polyetherurethane
PF	Platelet factor
PFA	Phosphonoformic add
PFC	Petrofluorochemical
PFT	Pulmonary function testing
PG	Polyglycolide; Propylene glycol

PGA	Polyglycolic add
PHA	Phytohemagglutinin; Pulse-height analyzer
PHEMA	Poly-2-hydroxyethyl methacrylate
PI	Propidium iodide
PID	Pelvic inflammatory disease; Proportional/integral/derivative
PIP	Peak inspiratory pressure
PL	Posterior leaflet
PLA	Polylactic acid
PLATO	Program Logic for Automated Teaching Operations
PLD	Potentially lethal damage
PLED	Periodic latoralized epileptiform discharge
PLT	Platelet
PM	Papillary muscles; Preventive maintenance
PMA	Polymethyl acrylate
p.m.f.	Probability mass function
PMMA	Polymethyl methacrylate
PMOS	P-type metal oxide silicon
PMP	Patient management problem; Poly(4-methylpentane)
PMT	Photomultiplier tube
PO	Per os
Po_2	Partial pressure of oxygen
POBT	Polyoxybutylene terephthalate
POM	Polyoxymethylene
POMC	Patient order management and communication system
POPRAS	Problem Oriented Perinatal Risk Assessment System
PP	Perfusion pressure; Polyproplyene; Postprandial (after meals)
PPA	Phonemic process analysis
PPF	Plasma protein fraction
PPM	Pulse position modulation
PPSFH	Polymerized phyridoxalated stroma-free hemoglobin
PR	Pattern recognition; Pulse rate
PRBS	Pseudo-random binary signals
PRP	Pulse repetition frequency
PRO	Professional review organization
PROM	Programmable read only memory
PS	Polystyrene
PSA	Pressure-sensitive adhesive
PSF	Point spread function
PSI	Primary skin irritation
PSP	Postsynaptic potential
PSR	Proton spin resonance
PSS	Progressive systemic sclerosis
PT	Plasma thromboplastin
PTB	Patellar tendon bearing orthosis
PTC	Plasma thromboplastin component; Positive temperature coefficient; Pressurized personal transfer capsule
PTCA	Percutaneous transluminal coronary angioplasty
PTFE	Polytetrafluoroethylene
PTT	Partial thromboplastin time
PUL	Percutaneous ultrasonic lithotripsy
PURA	Prolonged ultraviolet-A radiation
PUVA	Psoralens and longwave ultraviolet light photochemotherapy
P/V	Pressure/volume
PVC	Polyvinyl chloride; Premature ventricular contraction
PVI	Pressure–volume index
PW	Pulse wave; Pulse width
PWM	Pulse width modulation
PXE	Pseudo-xanthoma elasticum
QA	Quality assurance
QC	Quality control
R-BPC	Reverse bonded phase chromatography
R/S	Radiopaque-spherical
RA	Respiratory amplitude; Right arm
RAD	Right axis deviation
RAE	Right atrial enlargement
RAM	Random access memory
RAP	Right atrial pressure
RAT	Right anterior temporalis
RB	Right bundle
RBBB	Right bundle branch block
RBC	Red blood cell
RBE	Relative biologic effectiveness
RBF	Rose bengal fecal excretion
RBI	Resting baseline impedance
RCBD	Randomized complete block diagram
rCBF	Regional cerebral blood flow
RCC	Right coronary cusp
RCE	Resistive contact electrode
R&D	Research and development
r.e.	Random experiment
RE	Reference electrode
REM	Rapid eye movement; Return electrode monitor
REMATE	Remote access and telecommunication system
RES	Reticuloendothelial system
RESNA	Rehabilitation Engineering Society of North America
RF	Radio frequency; Radiographic-nuoroscopic
RFI	Radio-frequency interference
RFP	Request for proposal
RFQ	Request for quotation
RH	Relative humidity
RHE	Reversible hydrogen electrode
RIA	Radioimmunoassay
RM	Repetition maximum; Right masseter
RMR	Resting metabolic rate
RMS	Root mean square
RN	Radionuclide
RNCA	Radionuclide cineagiogram
ROI	Regions of interest
ROM	Range of motion; Read only memory
RP	Retinitis pigmentosa
RPA	Right pulmonary artery
RPP	Rate pressure product
RPT	Rapid pull-through technique
RPV	Right pulmonary veins
RQ	Respiratory quotient

RR	Recovery room
RRT	Recovery room time; Right posterior temporalis
RT	Reaction time
RTD	Resistance temperature device
RTT	Revised token test
r.v.	Random variable
RV	Residual volume; Right ventricle
RVH	Right ventricular hypertrophy
RVOT	Right ventricular outflow tract
RZ	Return-to-zero
SA	Sinoatrial; Specific absorption
SACH	Solid-ankle-cushion-heel
SAD	Source-axis distance; Statistical Analysis System
SAINT	System analysis of integrated network of tasks
SAL	Sterility assurance level; Surface averaged lead
SALT	Systematic analysis of language transcripts
SAMI	Socially acceptable monitoring instrument
SAP	Systemic arterial pressure
SAR	Scatter-air ratio; Specific absorption rate
SARA	System for anesthetic and respiratory gas analysis
SBE	Subbacterial endocarditis
SBR	Styrene-butadiene rubbers
SC	Stratum corneum; Subcommittees
SCAP	Right scapula
SCE	Saturated calomel electrode; Sister chromatid exchange
SCI	Spinal cord injury
SCRAD	Sub-Committee on Radiation Dosimetry
SCS	Spinal cord stimulation
SCUBA	Self-contained underwater breathing apparatus
SD	Standard deviation
SDA	Stepwise discriminant analysis
SDS	Sodium dodecyl sulfate
S&E	Safety and effectiveness
SE	Standard error
SEC	Size exclusion chromatography
SEM	Scanning electron microscope; Standard error of the mean
SEP	Somatosensory evoked potential
SEXAFS	Surface extended X-ray absorption fine structure
SF	Surviving fraction
SFD	Source-film distance
SFH	Stroma-free hemoglobin
SFTR	Sagittal frontal transverse rotational
SG	Silica gel
SGF	Silica gel fraction
SGG	Spark gap generator
SGOT	Serum glutamic oxaloacetic transaminase
SGP	Strain gage plethysmography; Stress-generated potential
SHE	Standard hydrogen electrode
SI	Le Système International d'Unités
SEBS	Surgical isolation barrier system
SID	Source to image reception distance
SIMFU	Scanned intensity modulated focused ultrasound
SIMS	Secondary ion mass spectroscopy; System for isometric muscle strength
SISI	Short increment sensitivity index
SL	Surgical lithotomy
SLD	Sublethal damage
SLE	Systemic lupus erythemotodes
SMA	Sequential multiple analyzer
SMAC	Sequential multiple analyzer with computer
SMR	Sensorimotor
S/N	Signal-to-noise
S:N/D	Signal-to-noise ratio per unit dose
SNP	Sodium nitroprusside
SNR	Signal-to-noise ratio
SOA	Sources of artifact
SOAP	Subjective, objective, assessment, plan
SOBP	Spread-out Bragg peak
SP	Skin potential
SPECT	Single photon emission computed tomography
SPL	Sound pressure level
SPRINT	Single photon ring tomograph
SPRT	Standard platinum resistance thermometer
SPSS	Statistical Package for the Social Sciences
SQUID	Superconducting quantum interference device
SQV	Square wave voltammetry
SR	Polysulfide rubbers
SRT	Speech reception threshold
SS	Stainless steel
SSB	Single strand breaks
SSD	Source-to-skin distance; Source-to-surface distance
SSE	Stainless steel electrode
SSEP	Somatosensory evoked potential
SSG	Solid state generator
SSP	Skin stretch potential
SSS	Sick sinus syndrome
STD	Source-tray distance
STI	Systolic time intervals
STP	Standard temperature and pressure
STPD	Standard temperature pressure dry
SV	Stroke volume
SVC	Superior vena cava
SW	Standing wave
TAA	Tumor-associated antigens
TAC	Time-averaged concentration
TAD	Transverse abdominal diameter
TAG	Technical Advisory Group
TAH	Total artificial heart
TAR	Tissue-air ratio
TC	Technical Committees
TCA	Tricarboxylic acid cycle
TCD	Thermal conductivity detector
TCES	Transcutaneous cranial electrical stimulation

TCP	Tricalcium phosphate
TDD	Telecommunication devices for the deaf
TDM	Therapeutic drug monitoring
TE	Test electrode; Thermoplastic elastomers
TEAM	Technology evaluation and acquisition methods
TEM	Transmission electron microscope; Transverse electric and magnetic mode; Transverse electromagnetic mode
TENS	Transcutaneous electrical nerve stimulation
TEP	Tracheoesophageal puncture
TEPA	Triethylenepho-sphoramide
TF	Transmission factor
TFE	Tetrafluorethylene
TI	Totally implantable
TICCIT	Time-shared Interaction Computer-Controlled Information Television
TLC	Thin-layer chromatography; Total lung capacity
TLD	Thermoluminescent dosimetry
TMJ	Temporomandibular joint
TMR	Tissue maximum ratio; Topical magnetic resonance
TNF	Tumor necrosis factor
TOF	Train-of-four
TP	Thermal performance
TPC	Temperature pressure correction
TPD	Triphasic dissociation
TPG	Transvalvular pressure gradient
TPN	Total parenteral nutrition
TR	Temperature rise
tRNA	Transfer RNA
TSH	Thyroid stimulating hormone
TSS	Toxic shock syndrome
TTD	Telephone devices for the deaf
TTI	Tension time index
TTR	Transition temperature range
TTV	Trimming tip version
TTY	Teletypewriter
TUR	Transurethral resection
TURP	Transurethral resections of the prostrate
TV	Television; Tidal volume; Tricuspid valve
TVER	Transscleral visual evoked response
TW	Traveling wave
TxB_2	Thrombozame B^2
TZ	Transformation zone
UES	Upper esophageal sphincter
UP	Urea-formaldehyde
UfflS	University Hospital Information System
UHMW	Ultra high molecular weight
UHMWPE	Ultra high molecular weight polyethylene
UL	Underwriters Laboratory
ULF	Ultralow frequency
ULTI	Ultralow temperature isotropic
UMN	Upper motor neuron
UO	Urinary output
UPTD	Unit pulmonary oxygen toxicity doses
UR	Unconditioned response
US	Ultrasound; Unconditioned stimulus
USNC	United States National Committee
USP	United States Pharmacopeia
UTS	Ultimate tensile strength
UV	Ultraviolet; Umbilical vessel
UVR	Ultraviolet radiation
V/F	Voltage-to-frequency
VA	Veterans Administration
VAS	Visual analog scale
VBA	Vaginal blood volume in arousal
VC	Vital capacity
VCO	Voltage-controlled oscillator
VDT	Video display terminal
VECG	Vectorelectrocardiography
VEP	Visually evoked potential
VF	Ventricular fibrillation
VOP	Venous occlusion plethysmography
VP	Ventriculoperitoneal
VPA	Vaginal pressure pulse in arousal
VPB	Ventricular premature beat
VPR	Volume pressure response
VSD	Ventricular septal defect
VSWR	Voltage standing wave ratio
VT	Ventricular tachycardia
VTG	Vacuum tube generator
VTS	Viewscan text system
VV	Variable version
WAIS-R	Weschler Adult Intelligence Scale-Revised
WAK	Wearable artificial kidney
WAML	Wide-angle mobility light
WBAR	Whole-body autoradiography
WBC	White blood cell
WG	Working Groups
WHO	World Health Organization; Wrist hand orthosis
WLF	Williams-Landel-Ferry
WMR	Work metabolic rate
w/o	Weight percent
WORM	Write once, read many
WPW	Wolff-Parkinson-White
XPS	X-ray photon spectroscopy
XR	Xeroradiograph
YAG	Yttrium aluminum garnet
ZPL	Zero pressure level

CONVERSION FACTORS AND UNIT SYMBOLS

SI UNITS (ADOPTED 1960)

A new system of metric measurement, the International System of Units (abbreviated SI), is being implemented throughout the world. This system is a modernized version of the MKSA (meter, kilogram, second, ampere) system, and its details are published and controlled by an international treaty organization (The International Bureau of Weights and Measures).

SI units are divided into three classes:

Base Units

length	meter[†] (m)
mass[‡]	kilogram (kg)
time	second (s)
electric current	ampere (A)
thermodynamic temperature§	kelvin (K)
amount of substance	mole (mol)
luminous intensity	candela (cd)

Supplementary Units

plane angle	radian (rad)
solid angle	steradian (sr)

Derived Units and Other Acceptable Units

These units are formed by combining base units, supplementary units, and other derived units. Those derived units having special names and symbols are marked with an asterisk (*) in the list below:

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
*absorbed dose	gray	Gy	J/kg
acceleration	meter per second squared	m/s^2	
*activity (of ionizing radiation source)	becquerel	Bq	1/s
area	square kilometer	km^2	
	square hectometer	hm^2	ha (hectare)
	square meter	m^2	

[†]The spellings "metre" and "litre" are preferred by American Society for Testing and Materials (ASTM); however, "−er" will be used in the Encyclopedia.

[‡]"Weight" is the commonly used term for "mass."

§Wide use is made of "Celsius temperature" (t) defined $t = T - T_0$ where T is the thermodynamic temperature, expressed in kelvins, and $T_0 = 273.15\,K$ by definition. A temperature interval may be expressed in degrees Celsius as well as in kelvins.

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
*capacitance	farad	F	C/V
concentration (of amount of substance)	mole per cubic meter	mol/m^3	
*conductance	siemens	S	A/V
current density	ampere per square meter	A/m^2	
density, mass density	kilogram per cubic meter	kg/m^3	g/L; mg/cm^3
dipole moment (quantity)	coulomb meter	C·m	
*electric charge, quantity of electricity	coulomb	C	A·s
electric charge density	coulomb per cubic meter	C/m^3	
electric field strength	volt per meter	V/m	
electric flux density	coulomb per square meter	C/m^2	
*electric potential, potential difference, electromotive force	volt	V	W/A
*electric resistance	ohm	Ω	V/A
*energy, work, quantity of heat	megajoule	MJ	
	kilojoule	kJ	
	joule	J	N·m
	electron volt†	eV†	
	kilowatt hour†	kW·h†	
energy density	joule per cubic meter	J/m^3	
*force	kilonewton	kN	
	newton	N	$kg·m/s^2$
*frequency	megahertz	MHz	
	hertz	Hz	1/s
heat capacity, entropy	joule per kelvin	J/K	
heat capacity (specific), specific entropy	joule per kilogram kelvin	J/(kg·K)	
heat transfer coefficient	watt per square meter kelvin	$W/(m^2·K)$	
*illuminance	lux	lx	lm/m^2
*inductance	henry	H	Wb/A
linear density	kilogram per meter	kg/m	
luminance	candela per square meter	cd/m^2	
*luminous flux	lumen	lm	cd·sr
magnetic field strength	ampere per meter	A/m	
*magnetic flux	weber	Wb	V·s
*magnetic flux density	tesla	T	Wb/m^2
molar energy	joule per mole	J/mol	
molar entropy, molar heat capacity	joule per mole kelvin	J/(mol·K)	
moment of force, torque	newton meter	N·m	
momentum	kilogram meter per second	kg·m/s	
permeability	henry per meter	H/m	
permittivity	farad per meter	F/m	
*power, heat flow rate, radiant flux	kilowatt	kW	
	watt	W	J/s
power density, heat flux density, irradiance	watt per square meter	W/m^2	
*pressure, stress	megapascal	MPa	
	kilopascal	kPa	
	pascal	Pa	N/m^2
sound level	decibel	dB	
specific energy	joule per kilogram	J/kg	
specific volume	cubic meter per kilogram	m^3/kg	
surface tension	newton per meter	N/m	
thermal conductivity	watt per meter kelvin	W/(m·K)	
velocity	meter per second	m/s	
	kilometer per hour	km/h	
viscosity, dynamic	pascal second	Pa·s	
	millipascal second	mPa·s	

†This non-SI unit is recognized as having to be retained because of practical importance or use in specialized fields.

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
viscosity, kinematic	square meter per second	m^2/s	
	square millimeter per second	mm^2/s	
	cubic meter	m^3	
	cubic decimeter	dm^3	L(liter)
	cubic centimeter	cm^3	mL
wave number	1 per meter	m^{-1}	
	1 per centimeter	cm^{-1}	

In addition, there are 16 prefixes used to indicate order of magnitude, as follows:

Multiplication factor	*Prefix*	*Symbol*	*Note*
10^{18}	exa	E	
10^{15}	peta	P	
10^{12}	tera	T	
10^{9}	giga	G	
10^{8}	mega	M	
10^{3}	kilo	k	
10^{2}	hecto	h[a]	[a]Although hecto, deka, deci, and centi are
10	deka	da[a]	SI prefixes, their use should be avoided
10^{-1}	deci	d[a]	except for SI unit-multiples for area and
10^{-2}	centi	c[a]	volume and nontechnical use of
10^{-3}	milli	m	centimeter, as for body and clothing
10^{-6}	micro	μ	measurement.
10^{-9}	nano	n	
10^{-12}	pico	p	
10^{-15}	femto	f	
10^{-18}	atto	a	

For a complete description of SI and its use the reader is referred to ASTM E 380.

CONVERSION FACTORS TO SI UNITS

A representative list of conversion factors from non-SI to SI units is presented herewith. Factors are given to four significant figures. Exact relationships are followed by a dagger (†). A more complete list is given in ASTM E 380-76 and ANSI Z210. 1-1976.

To convert from	*To*	*Multiply by*
acre	square meter (m^2)	4.047×10^{3}
angstrom	meter (m)	$1.0 \times 10^{-10\dagger}$
are	square meter (m^2)	$1.0 \times 10^{2\dagger}$
astronomical unit	meter (m)	1.496×10^{11}
atmosphere	pascal (Pa)	1.013×10^{5}
bar	pascal (Pa)	$1.0 \times 10^{5\dagger}$
barrel (42 U.S. liquid gallons)	cubic meter (m^3)	0.1590
Btu (International Table)	joule (J)	1.055×10^{3}
Btu (mean)	joule (J)	1.056×10^{3}
Bt (thermochemical)	joule (J)	1.054×10^{3}
bushel	cubic meter (m^3)	3.524×10^{-2}
calorie (International Table)	joule (J)	4.187
calorie (mean)	joule (J)	4.190
calorie (thermochemical)	joule (J)	$4.184^{\dagger}$
centimeters of water (39.2 °F)	pascal (Pa)	98.07
centipoise	pascal second (Pa·s)	$1.0 \times 10^{-3\dagger}$
centistokes	square millimeter per second (mm^2/s)	$1.0^{\dagger}$

To convert from	*To*	*Multiply by*
cfm (cubic foot per minute)	cubic meter per second (m^3/s)	4.72×10^{-4}
cubic inch	cubic meter (m^3)	1.639×10^{-4}
cubic foot	cubic meter (m^3)	2.832×10^{-2}
cubic yard	cubic meter (m^3)	0.7646
curie	becquerel (Bq)	$3.70 \times 10^{10\dagger}$
debye	coulomb-meter (C·m)	3.336×10^{-30}
degree (angle)	radian (rad)	1.745×10^{-2}
denier (international)	kilogram per meter (kg/m)	1.111×10^{-7}
	tex	0.1111
dram (apothecaries')	kilogram (kg)	3.888×10^{-3}
dram (avoirdupois)	kilogram (kg)	1.772×10^{-3}
dram (U.S. fluid)	cubic meter (m^3)	3.697×10^{-6}
dyne	newton(N)	$1.0 \times 10^{-6\dagger}$
dyne/cm	newton per meter (N/m)	$1.00 \times 10^{-3\dagger}$
electron volt	joule (J)	1.602×10^{-19}
erg	joule (J)	$1.0 \times 10^{-7\dagger}$
fathom	meter (m)	1.829
fluid ounce (U.S.)	cubic meter (m^3)	2.957×10^{-5}
foot	meter (m)	$0.3048^{\dagger}$
foot-pound force	joule (J)	1.356
foot-pound force	newton meter (N·m)	1.356
foot-pound force per second	watt(W)	1.356
footcandle	lux (lx)	10.76
furlong	meter (m)	2.012×10^2
gal	meter per second squared (m/s^2)	$1.0 \times 10^{-2\dagger}$
gallon (U.S. dry)	cubic meter (m^3)	4.405×10^{-3}
gallon (U.S. liquid)	cubic meter (m^3)	3.785×10^{-3}
gilbert	ampere (A)	0.7958
gill (U.S.)	cubic meter (m^3)	1.183×10^{-4}
grad	radian	1.571×10^{-2}
grain	kilogram (kg)	6.480×10^{-5}
gram force per denier	newton per tex (N/tex)	8.826×10^{-2}
hectare	square meter (m^2)	$1.0 \times 10^{4\dagger}$
horsepower (550 ft·lbf/s)	watt(W)	7.457×10^2
horsepower (boiler)	watt(W)	9.810×10^3
horsepower (electric)	watt(W)	$7.46 \times 10^{2\dagger}$
hundredweight (long)	kilogram (kg)	50.80
hundredweight (short)	kilogram (kg)	45.36
inch	meter (m)	$2.54 \times 10^{-2\dagger}$
inch of mercury (32 °F)	pascal (Pa)	3.386×10^3
inch of water (39.2 °F)	pascal (Pa)	2.491×10^2
kilogram force	newton (N)	9.807
kilopond	newton (N)	9.807
kilopond-meter	newton-meter (N·m)	9.807
kilopond-meter per second	watt (W)	9.807
kilopond-meter per min	watt(W)	0.1635
kilowatt hour	megajoule (MJ)	$3.6^{\dagger}$
kip	newton (N)	4.448×10^2
knot international	meter per second (m/s)	0.5144
lambert	candela per square meter (cd/m^2)	3.183×10^3
league (British nautical)	meter (m)	5.559×10^2
league (statute)	meter (m)	4.828×10^3
light year	meter (m)	9.461×10^{15}
liter (for fluids only)	cubic meter (m^3)	$1.0 \times 10^{-3\dagger}$
maxwell	weber (Wb)	$1.0 \times 10^{-8\dagger}$
micron	meter (m)	$1.0 \times 10^{-6\dagger}$
mil	meter (m)	$2.54 \times 10^{-5\dagger}$
mile (U.S. nautical)	meter (m)	$1.852 \times 10^{3\dagger}$
mile (statute)	meter (m)	1.609×10^3
mile per hour	meter per second (m/s)	0.4470

To convert from	*To*	*Multiply by*
millibar	pascal (Pa)	1.0×10^{2}
millimeter of mercury (0 °C)	pascal (Pa)	1.333×10^{2}†
millimeter of water (39.2 °F)	pascal (Pa)	9.807
minute (angular)	radian	2.909×10^{-4}
myriagram	kilogram (kg)	10
myriameter	kilometer (km)	10
oersted	ampere per meter (A/m)	79.58
ounce (avoirdupois)	kilogram (kg)	2.835×10^{-2}
ounce (troy)	kilogram (kg)	3.110×10^{-2}
ounce (U.S. fluid)	cubic meter (m^3)	2.957×10^{-5}
ounce-force	newton (N)	0.2780
peck (U.S.)	cubic meter (m^3)	8.810×10^{-3}
pennyweight	kilogram (kg)	1.555×10^{-3}
pint (U.S. dry)	cubic meter (m^3)	5.506×10^{-4}
pint (U.S. liquid)	cubic meter (m^3)	4.732×10^{-4}
poise (absolute viscosity)	pascal second (Pa·s)	0.10†
pound (avoirdupois)	kilogram (kg)	0.4536
pound (troy)	kilogram (kg)	0.3732
poundal	newton (N)	0.1383
pound-force	newton (N)	4.448
pound per square inch (psi)	pascal (Pa)	6.895×10^{3}
quart (U.S. dry)	cubic meter (m^3)	1.101×10^{-3}
quart (U.S. liquid)	cubic meter (m^3)	9.464×10^{-4}
quintal	kilogram (kg)	1.0×10^{2}†
rad	gray (Gy)	1.0×10^{-2}†
rod	meter (m)	5.029
roentgen	coulomb per kilogram (C/kg)	2.58×10^{-4}
second (angle)	radian (rad)	4.848×10^{-6}
section	square meter (m^2)	2.590×10^{6}
slug	kilogram (kg)	14.59
spherical candle power	lumen (lm)	12.57
square inch	square meter (m^2)	6.452×10^{-4}
square foot	square meter (m^2)	9.290×10^{-2}
square mile	square meter (m^2)	2.590×10^{6}
square yard	square meter (m^2)	0.8361
store	cubic meter (m^3)	1.0†
stokes (kinematic viscosity)	square meter per second (m^2/s)	1.0×10^{-4}†
tex	kilogram per meter (kg/m)	1.0×10^{-6}†
ton (long, 2240 pounds)	kilogram (kg)	1.016×10^{3}
ton (metric)	kilogram (kg)	1.0×10^{3}†
ton (short, 2000 pounds)	kilogram (kg)	9.072×10^{2}
torr	pascal (Pa)	1.333×10^{2}
unit pole	weber (Wb)	1.257×10^{-7}
yard	meter (m)	0.9144†

E

ECG. See ELECTROCARDIOGRAPHY, COMPUTERS IN.

ECHOCARDIOGRAPHY AND DOPPLER ECHOCARDIOGRAPHY

PETER S. RAHKO
University of Wisconsin Medical School
Madison, Wisconsin

INTRODUCTION

Echocardiography is a diagnostic technique that utilizes ultrasound (high frequency sound waves above the audible limit of 20 kHz) to produce an image of the beating heart in real time. A piezoelectric transducer element is used to emit short bursts of high frequency, low intensity sound through the chest wall to the heart and then detect the reflections of this sound as it returns from the heart. Since movement patterns and shape changes of several regions of the heart correlate with cardiac function and since changes in these patterns consistently appear in several types of cardiac disease, echocardiography has become a frequently used method for evaluation of the heart. Echocardiography has several advantages over other diagnostic tests of cardiac function:

1. It is flexible and can be used with transducers placed on the chest wall, inside oral cavities such as the esophagus or stomach, or inside the heart and great vessels.
2. It is painless.
3. It is a safe procedure that has no known harmful biologic effects.
4. It is easily transported almost anywhere including the bedside, operating room, cath lab, or emergency department.
5. It may be repeated as frequently as necessary allowing serial evaluation of a given disease process.
6. It produces an image instantaneously, which allows rapid diagnosis in emergent situations.

The first echocardiogram was performed by Edler and Hertz in 1953 (1) using a device that displayed reflected ultrasound on a cathode ray tube. Since that time multiple interrogation and display formats have been devised to display reflected ultrasound. The common display formats are

M-mode: A narrow beam of reflected sound is displayed on a scrolling strip chart plotting depth versus time. Only a small portion of the heart along one interrogation line ("ice pick" view) is shown at any one time.

Two-Dimensional Sector Scan (2D): A sector scan is generated by sequential firing of a phased array transducer along different lines of sight that are swept through a 2D plane. The image (a narrow plane in cross-section) is typically updated at a rate of 15–200 Hz and shown on a video monitor, which allows real time display of cardiac motion.

Three-Dimensional Imaging (3D): Image data from multiple 2D sector scans are acquired sequentially or in real time and displayed in a spatial format in three dimensions. If continuous data is displayed, time becomes the fourth dimension. The display can be shown as a loop that continuously repeats or on some systems in real time. Software allows rotation and "slicing" of the display.

Doppler: The Doppler effect is used to detect the rate and direction of blood flowing in the chambers of the heart and great vessels. Blood generally moves at a higher velocity than the walls of cardiac structures allowing motion of these structures to be filtered out. Blood flow is displayed in four formats:

Continuous Wave Doppler (CW): A signal of continuous frequency is directed into the heart while a receiver (or array of receivers) continuously processes the reflected signal. The difference between the two signals is processed and displayed showing direction and velocity of blood flow. All blood velocities in the line of sight of the Doppler beam are displayed.

Pulsed Wave Doppler (PW): Many bursts of sound are transmitted into the heart and reflected signals from a user defined depth are acquired and stored for each burst. Using these reflected signals an estimate is made of the velocities of blood or tissue encountered by the burst of sound at the selected depth. The velocity estimates are displayed similar to CW Doppler. The user defined position in the heart that the signal is obtained from stipulate the time of the acquisition of the reflected signal relative to transmission and length of the burst, respectively. The difference between transmitted and reflected signal frequencies is calculated, converted to velocity, and displayed. By varying the time between transmission and reception of the signal, selected velocities in small parts of the heart are sampled for blood flow.

Duplex Scanning: The 2D echo is used to orient the interrogator to the location of either the CW or PW signal allowing rapid correlation of blood flow data with cardiac anatomy. This is done by simultaneous display of the Doppler positional range gate superimposed on the 2D image.

Color Doppler: Using a complex array of bursts of frequency (pulsed packets) and multiple ultrasound acquisitions down the same beam line, motion of blood flow is estimated from a cross-correlation among the various acquired reflected waves. The

data is combined as an overlay onto the 2D sector scan for anatomical orientation. Blood flow direction, and an estimate of flow velocity are displayed simultaneously with 2D echocardiographic data for each point in the sector.

CLINICAL FORMATS OF ULTRASOUND

Current generation ultrasound systems allow display of all of the imaging formats discussed except 3D/4D imaging that is still limited in availability to some high end systems.

Specialized transducers that emit and receive the ultrasound have been designed for various clinical indications. Four common types of transducers are used (Fig. 1):

- Transthoracic: By far the most common, this transducer is placed on the surface of the chest and moved to different locations to image different parts of the heart or great vessels. All display formats are possible (Fig. 2).
- Transesophageal: The transducer is designed to be inserted through the patient's mouth into the esophagus and stomach. The ultrasound signal is directed at the heart from that location for specialized exams. All display formats are possible.
- Intracardiac: A small transducer is mounted on a catheter, inserted into a large vein and moved into the heart. Imaging from within the heart is performed to monitor specialized interventional therapy. Most display formats are available.
- Intravascular: Miniature sized transducers are mounted on small catheters and moved through arteries to examine arterial pathology and the results of selected interventions. Limited 2D display formats are available some being radial rather than sector based. The transducers run at very high frequencies (20–30 MHz).

Ultrasound systems vary considerably in size and sophistication (Fig. 3). Full size systems, typically found in hospitals display all imaging formats, accept all types of

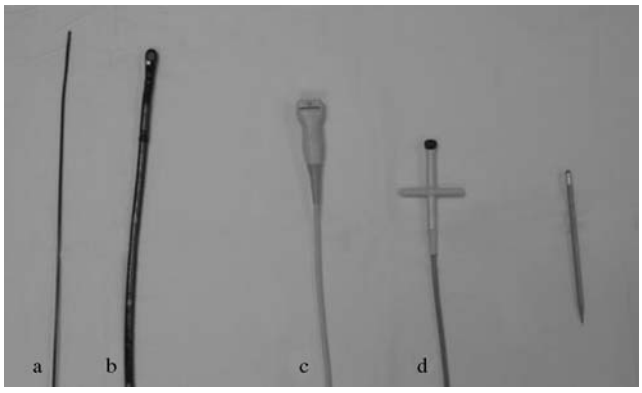
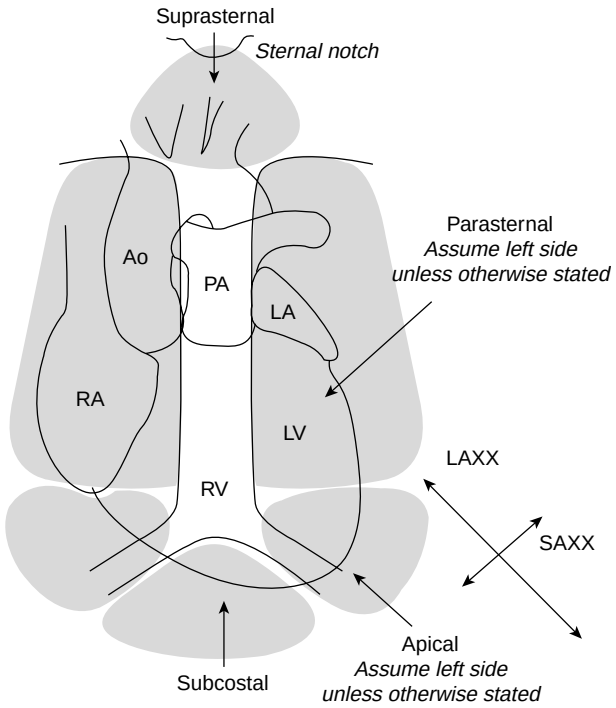


Figure 1. Comparative view of commonly used ultrasound transducers: (a) intracardiac, (b) transesophageal, (c) transthoracic, and (d) Pedof (special Doppler) transducer. A pencil is shown for size reference.

(a)



(b)

Figure 2. (a) Phased array transducer used for transthoracic imaging. The patient is shown, on left side, on exam table, typical for a transthoracic study. The transducer is in the apical position. (b) Diagram of the chest showing how the heart and great vessels are positioned in the chest and the locations where transthoracic transducers are placed on the chest wall. The four common positions for transducer placement are shown. The long axis (LAXX) and short axis (SAXX) orientations of the heart are shown for reference. These orientations form a reference for all of the views obtained in a study. Abbreviations are as follows: aorta (Ao), left atrium (LA), left ventricle (LV), pulmonary artery (PA), right atrium (RA), and right ventricle (RV).

Figure 3. Picture showing various types of ultrasound systems. At left a large "full-size" system is shown that can perform all types of imaging. At center is a miniaturized system that has most of the features of the full service system, but limited display, analysis and recording formats are available. The smallest "hand-held" system shown at right has basic features and is battery operated. It is designed for rapid screening exams integrated into a clinical assessment at the bedside, clinic, or emergency room.

transthoracic and transesophageal transducers, allow considerable on line image processing and support considerable analytic capacity to quantify the image.

Small imaging systems accept many but not all transducers and produce virtually all display formats, but lack the sophisticated array of image processing and analysis capacity found on the full systems. These devices may be used in ambulatory offices or other specialized circumstances requiring more basic image data. A full clinical study is possible.

Portable hand-held battery operated devices are used in limited circumstances, sometimes for screening exams or limited studies. Typically transducers are limited, image processing is rudimentary and analysis capacity very limited.

PRINCIPLES OF ULTRASOUND

Prior to discussing the mechanism of image production some common terms that govern the behavior of ultrasound in soft tissue should be defined. Since ultrasound is propagated in waves, its behavior in a medium is defined by

$$\lambda = \frac{c}{f}$$

where f is the wave *frequency*, λ is the *wavelength*, and c is the *acoustic velocity* of ultrasound in the medium. The acoustic velocity for most soft tissues is similar and remains constant for a given tissue no matter what the frequency or wavelength (Table 1). Thus in any tissue frequency and wavelength are inversely related. As frequency increases, wavelength decreases. As wavelength decreases, the minimum distance between two structures, that allows them to be characterized as two separate structures, also decreases. This is called the *spatial resolution* of the instrument. One might conclude that very high frequency should always be used to maximize resolution. Unfortunately, as frequency increases, penetration of the ultrasound signal into soft tissue decreases. This serves to limit the frequency and, thus, the resolving power of an ultrasonic system for any given application.

Sound waves are emitted in short bursts from the transducer. As frequency rises, it takes less time to emit the same number of waves per burst. Thus, more bursts of sound can be emitted per unit of time, increasing the spatial resolution of the instrument (Fig. 4). The optimal image is generated by using a frequency that gives the highest possible resolution and an adequate amount of penetration. For transthoracic transducers expected to penetrate up to 24 cm into the chest, typical frequencies used are from 1.6 to 7.0 MHz. Most transducers are broadband in that they generate sound within an adjustable range of frequency rather than at a single frequency. Certain specialized transducers such as the intravascular transducer may only need to penetrate 4 cm. They may have a frequency of 30 MHz to maximize resolution of small structures.

The term *ultrasonic attenuation* formally defines the more qualitative concept of tissue penetration. It is a complex parameter that is different for every tissue type and is defined as the rate of decrease in wave amplitude per distance penetrated at a given frequency. The two important properties that define ultrasonic attenuation are reflection and absorption of sound waves (2). Note in Table 1 that ultrasound easily passes through blood and soft tissue, but poorly penetrates bone or air-filled lungs.

Acoustic impedance (z) is the product of acoustic velocity (c) and tissue density (ρ); thus this property is tissue specific but frequency independent. This property is important because it determines how much ultrasound is

Table 1. Ultrasonic Properties of Some Selected Tissues[a]

Tissue	Velocity of Propagation, 10^3 m·s^{-1}	Density, g·mL^{-1}	Acoustic Impedance, 10^6 rayl[b]	Attenuation at 2.25 MHz, dB·cm^{-1}
Blood	1.56	1.06	1.62	0.57
Myocardium	1.54	1.07	1.67	3
Fat	1.48	0.92	1.35	1.7
Bone	~3–4	1.4–1.8	4–6	37
Lung (inflated)	0.7	0.4	0.26–0.46	62

[a]Adapted from Wilson D. A., Basic principles of ultrasound. In: Kraus R., editor. The Practice of Echocardiography, New York: John Wiley & Sons; 1985, p 15. This material is used by permission of John Wiley & Sons, Inc.

[b]1 rayl = 1 kg·m^{-2}·s^{-1}.

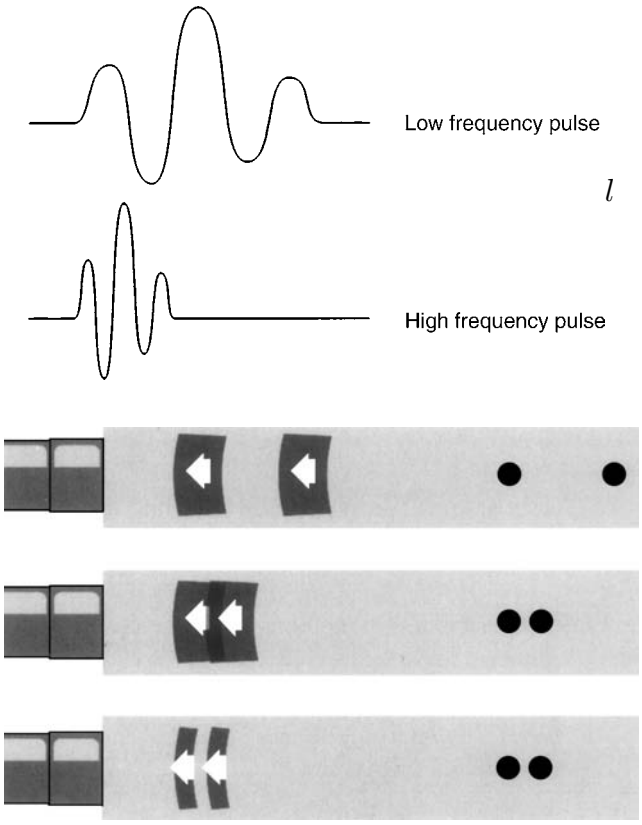


Figure 4. Upper panel Two depictions of an emitted pulse of ultrasound. Time moves horizontally and amplitude moves vertically. Note the high frequency pulse of three waves takes less time. Lower panel Effect of pulse duration on resolution. One echo pulse is delivered toward two reflectors and reflections are shown. In the top panel, the reflectors are well separated from each other and distinguished as two separate structures. In the middle panel, pulse frequency and duration are unchanged, but the reflectors are close together. The two returning reflections overlap, the instrument will register the two reflectors as one structure. In the lower panel the pulse frequency is increased, thus pulse duration is shortened. The two objects are again separately resolved. (Reprinted from Zagzebski JA Essentials of Ultrasound Physics. St. Louis: Mosby-Year Book; copyright © 1996, p 28, with permission from Elsevier.)

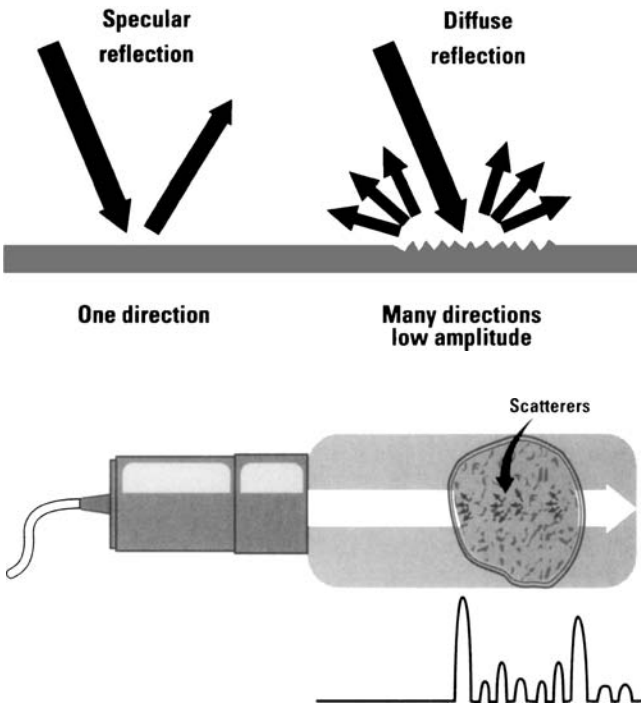


Figure 5. Upper panel Comparison between specular and diffuse reflectors. Note the diffuse reflector is less angle dependent than the specular reflector. Lower panel Example of combined reflections (shown at bottom of figure) returning from a structure, typical of reflections coming back from heart muscle. The large amplitude specular echo corresponds to the border of the structure. The interior of the structure produces low amplitude scattered reflections. (Reprinted from Zagzebski JA, Essentials of Ultrasound Physics. St. Louis, Mosby-Year Book; copyright © 1996 p 12 with permission from Elsevier.)

reflected at an interface between two different types of tissue (Table 1). When a short burst of ultrasound is directed at the heart, portions of this energy are reflected back to the receiver. It is these reflected waves that produce the image of the heart. A very dense structure such as calcified tissue has high impedance and is a strong reflector.

There are two types of reflected waves: *specular reflections* and *diffuse reflections* (Fig. 5). Specular reflections occur at the interface between two types of tissue. The greater the difference in acoustic impedance between two tissues, the greater the amount of specular reflection and the lower the amount of energy that penetrates beyond the interface. The interface between heart muscle and blood produces a specular echo, as does the interface between a heart valve and blood. Specular echoes are the primary echoes that are imaged by M-mode, 2D echo, and 3D echo and thus primarily form an outline of the heart. Diffuse reflected echoes are much weaker in energy. They are produced by small irregular more weakly reflective objects such as the myocardium itself. Scattered echoes "fill in the details" between the specular echoes. With modern equipment scattered echoes are processed and analyzed providing much more detail to tissue being examined.

Doppler echocardiography uses *scattered echoes* from red blood cells for detecting blood flow. Blood cells are Rayleigh scatterers since the diameter of the blood cells are much smaller than the typical wavelength of sound used to interrogate tissue. Since these reflected signals are even fainter than myocardial echoes, Doppler must operate at a higher energy level than M-mode or 2D imaging.

Harmonic imaging is a recent addition to image display made possible by advances in transducer design and signal processing. It was first developed to improve the display of contrast agents injected intravenously as they passed through the heart. These agents, gas filled microbubbles 2–5 μm in diameter, are highly reflective Rayleigh scatterers. At certain frequencies within the broadband transducer range the contrast bubbles resonate, producing a relatively strong signal at multiples of the fundamental interrogation frequency called harmonics. By using a high

pass filter (a system that blanks out all frequency below a certain level), to eliminate the fundamental frequency reflectors, selective reflections from the second harmonic are displayed. In second harmonic mode, reflections from the resonating bubbles are a much stronger than reflections from soft tissue and thus bubbles are preferentially displayed. This allows selective analysis of the contrast agent as it passes through the heart muscle or in the LV cavity (3).

Harmonic imaging has recently been applied to conventional 2D images without contrast. As the ultrasound wave passes through tissue, the waveform is modified by nonlinear propagation through tissue causing a shape change in the ultrasound beam. This progressively increases as the beam travels deeper into the heart. Electronic canceling of much of the image by filtering out the fundamental frequency allows selective display of the harmonic image, improving overall image quality by elimination of some artifacts. The spatial resolution of the signal is also improved since the reflected signal analyzed and displayed is double that of the frequency produced by the transducer (4).

ECHOCARDIOGRAPHIC INSTRUMENTATION

The transducer is a piezoelectric (pressure electric) device. When subjected to an alternating electrical current, the ceramic crystal (usually barium titanate, lead zirconate titanate, or a composite ceramic) expands and contracts producing compressions and rarefactions in its environment, which become waves. Various transducers produce ultrasonic waves within a frequency range of 1.0–40 MHz. The waves are emitted as brief pulses lasting ~ 1 μs out of every 100–200 μs. During the remaining 99–199 μs of each interval the transducer functions as a receiver that detects specular and diffuse reflections as they return from the heart. The same crystal, when excited by a reflected sound wave, produces an electrical signal and sends it back to the echocardiograph for analysis and display. Since one heartbeat lasts somewhere from 0.3 to 1.2 s, the echocardiographic device sends out a minimum of several hundred impulses per beat allowing precise tracking of cardiac motion throughout the beat.

After the specular and diffuse echoes are received they must be displayed in a usable format. The original ultrasound devices used an A-mode format (Fig. 6) that displayed depth on the *y* axis and amplitude of the signal on the *x*-axis. The specular echoes from boundaries between cardiac chambers register as the strongest echoes. No more than 1D spatial information is obtained from this format.

In a second format, B-mode, the amplitudes of the returning echoes are displayed as dots of varying intensity on a video monitor in what has come to be called a gray scale (Fig. 6). If the background is black (zero intensity), then progressively stronger echoes are displayed as progressively brighter shades of gray with white representing the highest intensity. Most echocardiographic equipment today uses between 64 and 512 shades of gray in its output display. The B-mode format, by itself, is not adequate for cardiac imaging and must be modified to image a continuously moving structure.

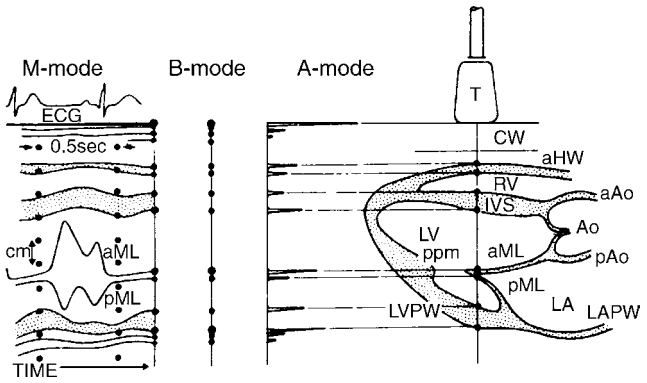
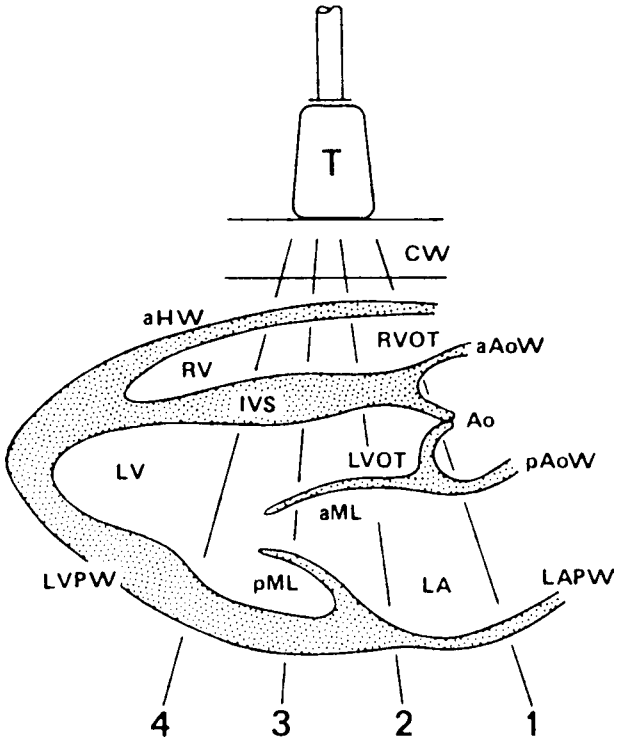


Figure 6. Composite drawing showing the three different modes of display for a one-dimensional (1D) ultrasound signal. In the right half of the figure is a schematic drawing of a cross-section through the heart. The transducer (T) sits on the chest wall (CW) and directs a thin beam of ultrasonic pulses into the heart. This beam traverses the anterior wall (aHW) of the right ventricle (RV), the interventricular septum (IVS), the anterior (aML), and posterior (pML) leaflets of the mitral valve, and the posterior wall of the left ventricle (LVPW). Each dot along the path of the beam represents production of a specular echo. These are displayed in the corresponding A-mode format, where vertical direction is depth and horizontal direction is amplitude of the reflected echo and B-mode format where again vertical direction is depth but amplitude is intensity of the dot. If time is added to the B-mode format, an M-mode echo is produced, which is shown in the left panel. This allows simultaneous presentation of motion of the cardiac structures in the path of the echo beam throughout the entire cardiac cycle; measurement of vertical depth, thickness of various structures, and timing of events within the cardiac cycle. If the transducer is angled in a different direction, a distinctly different configuration of echoes will be obtained. In the figure, the M-mode displayed is at the same beam location as noted in the right-side panel. Typical movement of the AML and PML is shown. ECG = electrocardiogram signal. (From Pierand L., Meltzer RS., Roelandt J, Examination techniques in M-mode and two-dimensional echocardiography. In: Kraus R editor, The Practice of Echocardiography, New York: John Wiley & Sons; copyright © 1985, p 69. This material is used by permission of John Wiley & Sons, Inc.)

To image the heart, the M-mode format (M for motion) was devised (Fig. 6). With this technique, the transducer is pointed into the chest at the heart and returning echoes are displayed in B-mode. A strip chart recorder (or scrolling video display) constantly records the B-mode signal with depth of penetration on the y-axis and time the parameter displayed on the *x*-axis. By adding an electrocardiographic signal to monitor cardiac electrical activity and to mark the beginning of each cardiac cycle, the size, thickness, and movement of various cardiac structures throughout a cardiac cycle are displayed with high resolution. By variation of transducer position, the ultrasound beam is directed toward several cardiac structures (Fig. 7).

The M-mode echo was the first practical ultrasound device for cardiac imaging and has produced a considerable amount of important data. Its major limitation is its limited field of view. Few spatial relationships between cardiac structures can be displayed that severely limits diagnostic capability. The angle of interrogation of the heart is also



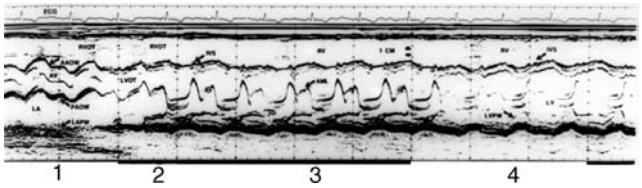


Figure 7. Upper panel Schematic diagram of the heart as in Fig. 6. The principal M-mode views are labeled 1–4. The corresponding M-mode image from these four views is shown in lower panel Abbreviations as in Fig. 6**a** Additional abbreviations: AV = Aortic valve, AAOW = anterior aortic wall, LA = left atrial posterior wall, LVOT = left ventricular outflow tract, RVOT = right ventricular outflow tract. (From Pierand L, Meltzer RS, Roelandt J, Examination techniques in M-mode and 2D echocardiography. In: Kraus R editor, The Practice of Echocardiography, New York, John Wiley & Sons; copyright © 1985, p 71. This material is used with permission of John Wiley & Sons, Inc.)

difficult to control. This can distort the image and render size and dimension measurements unreliable.

Since the speed of sound is rapid enough to allow up to 5000 short pulses of ultrasound to be emitted and received each second at depths typical for cardiac imaging, it was recognized that multiple B-mode scans in several directions could be processed rapidly enough to display a "real-time" image. Sector scanning in two dimensions was originally performed by mechanically moving a single element piezoelectric crystal through a plane. Typically, 128 B-mode scan lines were swept through a 60–90° arc 30 times $\cdot s^{-1}$ to form a video composite B-mode sector (Fig. 8). These mechanical devices have been replaced by transducer arrays that place a group of closely spaced piezoelectric elements, each with its own electrical connection to the ultrasound system, into a transducer. The type of array used depends on the structure

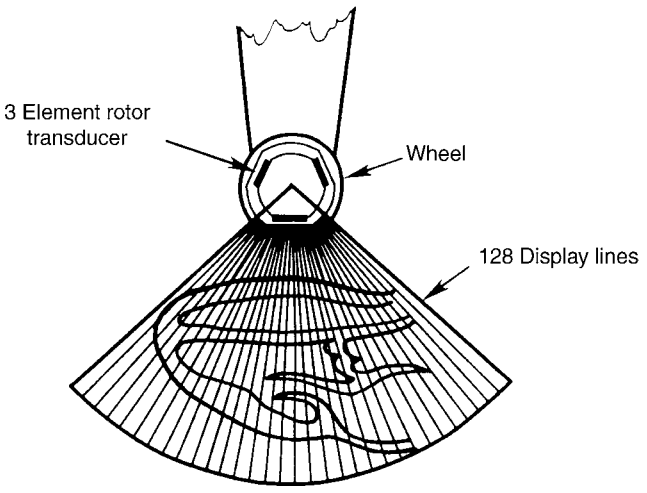


Figure 8. Diagram of a 2D echo mechanical transducer with three crystals. As each segment sweeps through the 90° arc, an element fires a total of 128 times. The composite of the 128 B-mode lines form a 2D echo frame. Typically there are 30 frames/s of video information, a rate rapid enough to show contractile motion of the heart smoothly. (From Graham PL, Instrumentation. In: Krause R. editor. The Practice of Echocardiography, New York: John Wiley & Sons; copyright © 1985, p 41. This material is used by permission of John Wiley & Sons, Inc.)

being imaged. For cardiac ultrasound, a phased array configuration is used, typically consisting of 128–1024 individual elements. In a phased array transducer, a portion of the elements are fired to produce each sector scan line. The sound beams are electronically steered through the sector by changing the time delay sequence of the array elements (Fig. 9). In a similar fashion, all elements are electronically sequenced to receive reflected sound from selected parts of the sector being scanned (5).

Use of a phased array device allows many other modifications of the sound wave in addition to steering. Further sophisticated electronic manipulations of the time of sound transmission and delays in reception allow selective focusing of the beam to concentrate transmit energy that enhance image quality of selected structures displaced in the sector. Recent design advances have markedly increased the sector frame rate (number of displayed sectors/second) to levels beyond 220 Hz, markedly increasing the time resolution of the imaging system. While the human visual processing system cannot resolve time at such a rapid rate, high frame rates allow for sophisticated quantitation based on the 2D image, such as high-resolution graphs of time based indexes.

DOPPLER ECHOCARDIOGRAPHY

Application of the Doppler effect allows analysis of blood flow within the heart and great vessels. The Doppler effect, named for its discoverer Christian Doppler, describes the change in frequency and wavelength that occurs with relative motion between the source of the waves and the receiver. If a source of sound remains stationary with respect to its listener, then the frequency and wavelength of the sound will also remain constant. However, if the

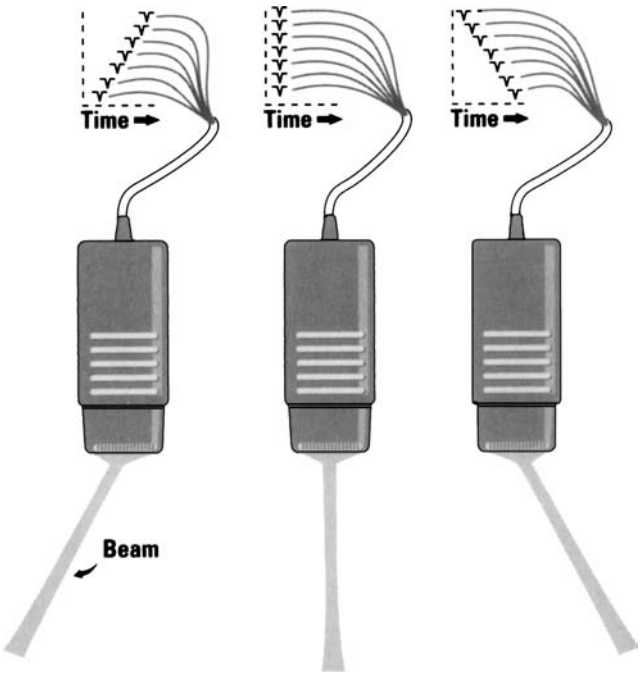


Figure 9. Diagram of a phased array transducer. Beam direction is varied by changing the delay sequence among the transmitted pulses produced by each individual element. (From Zagzebski JA. Essentials of Ultrasound Physics. St. Louis: Mosby-Year Book Inc.; copyright © 1996, with permission from Elsevier.)

sound source is moving away from the listener wavelength increases and frequency decreases. The opposite will occur if the sound source is moving toward the listener (Fig. 10).

The Doppler principle is applied to cardiac ultrasound in the following way: A beam of continuous wave ultrasound is transmitted into the heart and reflected off red blood cells as they travel through the heart. The reflected impulses are then detected by a receiver. If the red blood cells are moving toward the receiver, the frequency of the reflected echoes will be higher than the frequency of the transmitted echoes and vice versa (Fig. 10).

The difference between the frequency of the transmitted and received echoes (usually called the Doppler shift) can be related to the velocity of blood flow by the following equation:

$$V = \frac{c(f_r - f_t)}{2(f_t)(\cos\theta)30}$$

where V is the blood flow velocity, c is the speed of sound in soft tissue (1540 $m \cdot s^{-1}$), f_r is the frequency of the reflected echoes, f_t is the frequency of the transmitted echoes, and θ is the intercept angle between the direction of blood flow and the ultrasound beam. Thus, flow toward the transducer will produce a positive Doppler shift ($f_r > f_t$), while flow away from the transducer will produce a negative Doppler shift ($f_r < f_t$). The only variable that cannot be directly measured is θ. Since $\cos 0^\circ = 1$, it follows that maximal flow will be detected when the Doppler beam is parallel to blood flow. Since blood flow cannot be seen with 2D echo, at best, θ can only be estimated. Fortunately, if θ is within 20° of the direction of blood flow, the error introduced by angulation is small. Therefore,

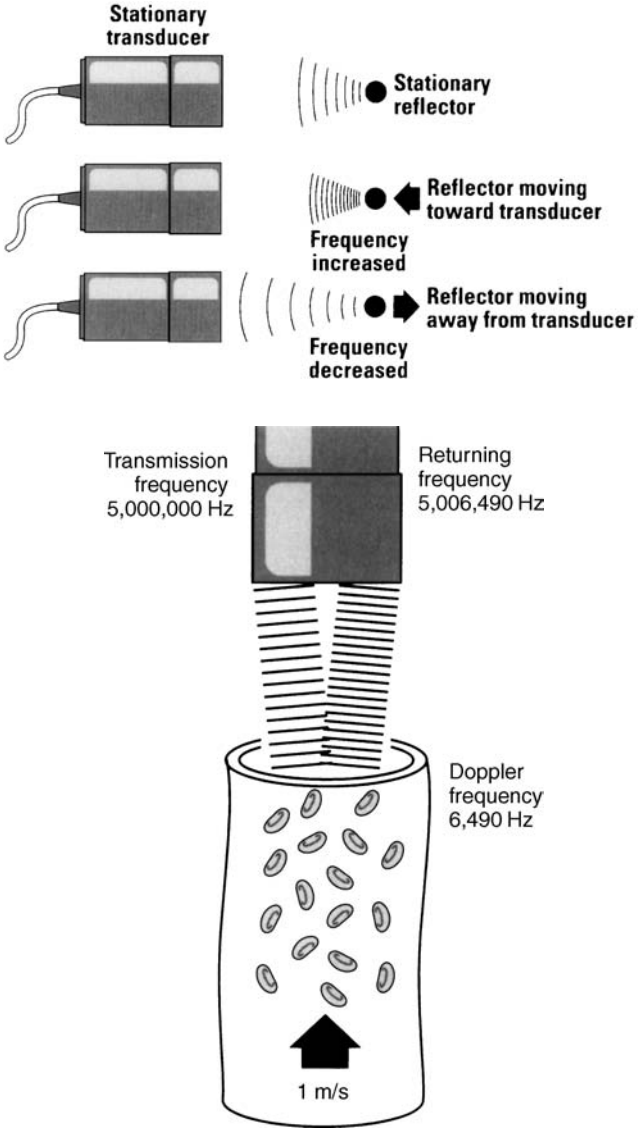


Figure 10. Upper panel The Doppler effect as applied to ultrasound. The frequency increases slightly when the reflector is moving toward the transducer and decreases when the reflector is moving away from the transducer. Lower panel Relative magnitude of the Doppler shift caused by red blood cells moving between cardiac chambers. In this example the frequency shift corresponds to a movement rate of 1 $m \cdot s^{-1}$. (Reprinted from Zagzebski JA. Essentials of Ultrasound Physics. St. Louis: Mosby-Year Book Inc.; copyright © 1996, with permission from Elsevier.)

most investigators do not formally correct for θ. Instead, the Doppler beam is aligned as closely as possible in the presumed direction of maximal flow and then adjusted until maximal flow is detected (6).

Doppler echo operates in two basic formats. Figure 10 depicts the CW method. An ultrasound signal is continuously transmitted into the heart while a second crystal (or array of crystals) in the transducer continually receives reflected signals. All red blood cells in the overlap region between the beam patterns of the transmit and receive crystals contribute to the calculated signal. The frequency

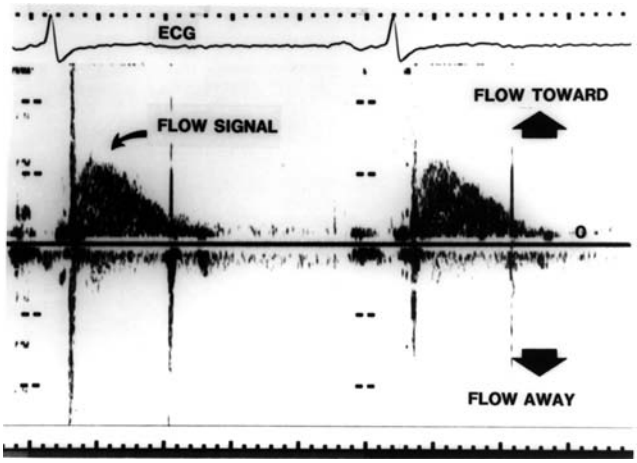


Figure 11. Example of a continuous wave Doppler signal taken from a patient. Flow toward the transducer is a positive (upward) deflection from the baseline and flow away from the transducer is downward. ECG = electrocardiogram signal.

content of this signal, combined with an electrocardiographic monitor lead, is then displayed on a strip chart similar to an M-mode echo (Fig. 11).

The advantage of the CW method is that it can detect a wide range of flow velocities encompassing every possible physiologic or pathologic flow state. The main disadvantage of the CW format is that the site of flow cannot be localized. To overcome the lack of localization of CW Doppler, a second format was developed called pulsed Doppler (PW). In this format, similar to B-mode echocardiographic imaging, brief bursts of ultrasound are transmitted at a given frequency followed by a silent interval (Fig. 12). Since the time it takes for a reflected burst of sound waves to return to the receiving crystal is directly related to the distance the reflecting structure is from the receiver, the position in the heart from which blood flow is sampled can be precisely controlled by limiting the time interval during which reflected ultrasound is received. This is known as range gating and allows the investigator to limit the area sampled to small portions of the heart or great vessel. There is a price to be paid for sample selectivity, however. The maximal detectable velocity PW Doppler is able to display is equal to one-half the pulse repetition frequency (frequently called the Nyquist limit). This reduces the number of situations in which flow velocity samples unambiguously display the flow phenomenon. A typical PW Doppler display shows flow both toward and away from the transducer (Fig. 13).

As one interrogates deeper structures progressively further from the transducer, the pulse repetition frequency must, of necessity, be decreased. As a result, the highest detectable velocity of the PW Doppler mode becomes progressively smaller. Due to attenuation, the sensitivity for detecting flow becomes progressively lower at greater distances from the transducer. Despite these limitations, selective sampling of blood flow allows interrogation of a wide array of cardiac structures in the heart.

When a measured flow has a velocity in a particular direction greater than the Nyquist limit, not all of the

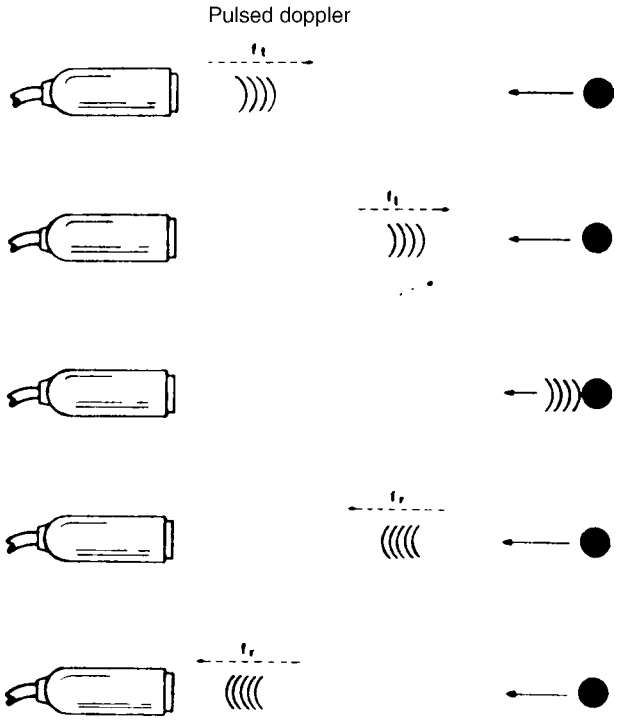


Figure 12. Pulsed Doppler echocardiography. In place of a continuous stream of ultrasound, brief pulses are emitted similar to M-mode or 2D imaging. By acquiring reflected signal data over a limited time window following each pulse, reflections emanating only from a certain depth may be received. (From Feigenbaum H, Echocardiography (5th ed), Philadelphia, PA: Lea & Febiger; 1994, p 29, with permission from Lippincott Williams & Wilkins ©.)

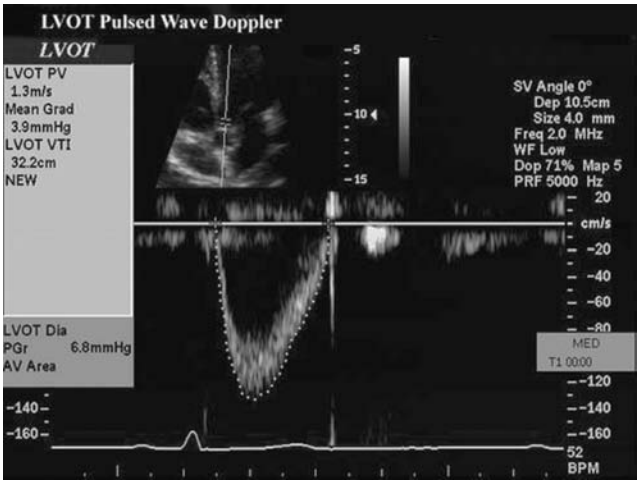


Figure 13. Example of a pulsed wave Doppler tracing. The study was recorded in duplex mode from the LVOT. The small upper insert shows the 2D image which guides positioning of the sample volume (arrow). The Doppler signal is shown below and has been traced by the sonographer using electronic analysis system. Data automatically detected from tracing the signal are shown on the left and include the peak velocity (PV) of flow and the integral of flow velocity (VTI) that can be used for calculations of cardiac output.

spectral envelope of the signal is visible. Indeed, the velocity estimates "wrap-around" to the other side of the velocity map and the flow appears to be going in the opposite direction. This phenomenon where large positives velocities are displayed as negative velocities is called aliasing. There are two strategies to mitigate or eliminate aliasing. The zero shift line (no velocity) may be moved upward or downward, effectively doubling the display range in the desired direction. This may be sufficient to "un-wrap" the aliased velocity and display it entirely in the appropriate direction. Some velocities may still be too high for this strategy to work. To display these higher velocities an alternative method called high pulse repetition frequency (high PRF) mode is employed. In this mode, sample volumes at multiples of the main interrogation sample volume are also interrogated. This is accomplished by sending out bursts of pulse packets at multiples of the burst rate necessary to sample at the desired depth. The system displays multiple sites from which the displayed signal might originate. While this creates some ambiguity in the exam, the anatomy displayed by the 2D exam usually allows a correct delineation as to which range gate is creating the signal (Fig. 14).

By itself, Doppler echo is a nonimaging technique that only produces flow patterns and audible tone patterns (since all Doppler shifts fall within the audible range). Phased array transducers, however, allow simultaneous display of both 2D images and Doppler in a mode called duplex Doppler echocardiography. By using this combination, the PW Doppler sample volume is displayed as an overlay on the 2D image and is moved to a precise area in the heart where the flow velocity is measured (Fig. 13). This combination provides both anatomic and physiologic information about the interrogated cardiac structure. Most commonly, Duplex mode is used with the PW wave format of Doppler echo. However, it is also possible to position the continuous wave beam anywhere in the 2D sector by superimposing the Doppler line of interrogation on top of the 2D image.

Just as changing from an M-mode echo to a 2D sector scan markedly increases the amount of spacial data simultaneously available to the clinician, Doppler information can be expanded from a single a PW wave sample volume or CW line to a full sector array. Color Doppler echocardiography displays blood flow within the heart or blood vessel as a sector plane of velocity information. By itself, a color flow scan imparts little information so the display is always combined with the 2D image as an overlay so blood flow may be instantly correlated with anatomic structures within the heart or vessel.

Color Doppler uses the same transmission principles as B-mode 2D imaging and PW Doppler. Brief transmit pulses of sound are steered along interrogation lines in a sector simultaneously with usual B-mode imaging pulses (Fig. 15). In place of just one pulse of sound, multiple pulses are transmitted. The multiple bursts of sound, typically 4–8 in number, are referred to as packets or ensembles of pulses. The first signal stores all reflected echoes along each scan line. Reflectors from subsequent pulses in the packet are received, stored, and rapidly compared to the previous packets. Reflected waves that are identical during each burst in the packet are canceled

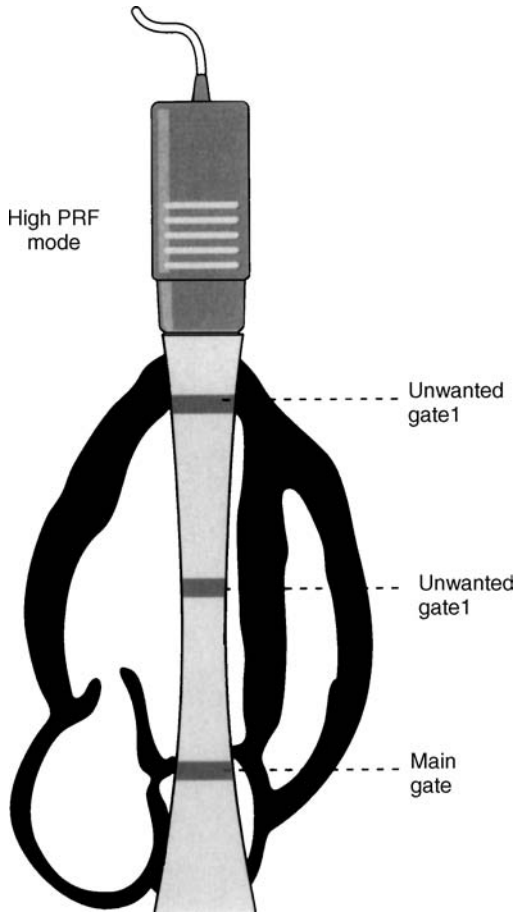


Figure 14. Example of high PFR mode of pulsed Doppler. The signal is used for velocity detected at the main gate because pulse packets are sent out more frequently at multiples of the frequency needed to sample at the main gate. Information can also be acquired at other gates that are multiples of the main gate. While the Nyquist limit is higher due to a higher sampling rate some signal ambiguity may occur due to information acquired from the unwanted gates. (Reprinted from Zagzebski JA. Essentials of Ultrasound Physics. St. Louis: Mosby-Year Book Inc. copyright © 1996 with permission from Elsevier.)

out and designated as stationary. Reflected waves that progressively change from burst to burst are acquired and processed rapidly for calculation of the phase shift in the ultrasound carrier. Both direction and velocity of movement are proportional to this phase change. Estimates for the average velocity are assigned to a pixel location on the video display. The velocity is estimated by an auto correlator system. On the output display, velocity is typically displayed as brightness of a given color similar to B-mode gray scale with black meaning no motion and maximum brightness indicating the highest velocity detected. Two contrasting colors are used to display direction of flow, typically a red-orange group for flow toward the transducer and a blue group away from transducer. Since the amount of data processed is markedly greater than a typical B-Mode, maximum frame rates of sector scan displays tend to be much lower. This limitation is due both to the speed of sound and the multiple packets of ultrasound evaluated in each interrogation line. To maximize

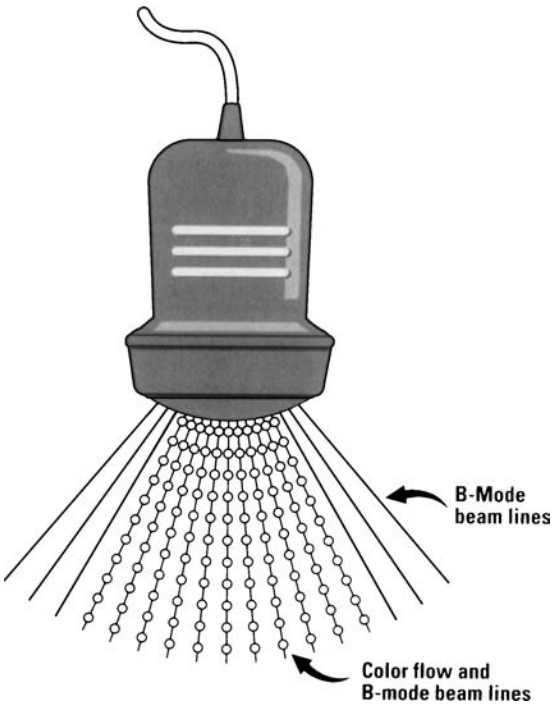


Figure 15. Diagram showing method of transmission of color flow Doppler ultrasound signals. Packets or ensembles of pulses represented by the open circles in the figure are sent out along some of the scan lines of the image. The reflected waves are analyzed in an autocorrelation circuit to allow display of the color flow image. (Reprinted from Zagzebski JA. Essentials of Ultrasound Physics. St. Louis: Mosby-Year Book Inc. copyright © 1996, with permission from Elsevier.)

time resolution, the color flow sector used to display data during a 2D echo may be considerably reduced in size compared to a usual 90° 2D sector scan. Some systems can only display color flow data at relatively slow frame rates of 6–10 Hz. Recent innovations, in which there are multiple receive lines for each transmit line allow much higher frame rates giving excellent time resolution on some systems (7).

Clinically, precise measurement of flow velocity is usually obtained with PW or CW Doppler. Color is used to rapidly interrogate a sector for the presence or absence of blood flow during a given part of the cardiac cycle. Another important part of the color exam is a display of the type of flow present. Normal blood flow is laminar; abnormal blood flow caused by valve or blood vessel pathology is turbulent. The difference between laminar and turbulent flow is easily displayed by color Doppler. With laminar flow, the direction of flow is uniform and variation in velocity of adjacent pixels of interrogation is small. With turbulent flow, both parameters are highly variable. The auto correlation system analyzing color Doppler compares the variance in blood flow between different pixels. The display can be set to register a third color for variance such as green, or the clinician may look for a "mosaic" pattern of flow in which non uniform color velocities and directions are scattered through the sector areas of color interrogation (Figs. 16 and 17).

As with pulsed Doppler there is a Nyquist limit restriction on maximal velocity than can be displayed. The zero flow position line may be adjusted as with PW Doppler to maximize the velocity limit in a given direction. High PRF is not possible with color Doppler. The nature of the color display is such that aliasing results in a shift from one color sequence to the next. Thus, in some situations a high velocity shift can be detected due to a clear shift in color (e.g., from a red-orange sequence to a blue sequence). This phenomenon has been put to use clinically by purposely manipulating the velocity range of color display to force aliasing to occur. By doing this, isovelocity lines are displayed outlining a velocity border of flow in a particular direction and a particular velocity (8).

Thus far, all discussion of Doppler has been confined to interrogation and display of flow velocity. Alternate modes of interrogation are possible. One mode, called power Doppler (or energy display Doppler) assesses the amplitude of the Doppler signal rather than velocity. By evaluating amplitude in place of velocity, this display becomes proportional to the number of moving blood cells present in the interrogation field rather than the velocity. This application is particularly valuable for perfusion imaging when the amount of blood present in a given area is of primary interest (9).

ULTRASOUND SIGNAL PROCESSING, DISPLAY, AND MANAGEMENT

Once each set of the reflected ultrasound data returns to the transducer, it is processed and then transmitted to video display. The information is first processed by a scan converter, which assigns video data to a matrix array of picture elements, "pixels." Several manipulations of the image are possible to reduce artifacts, enhance information in the display, and analyze the display quantitatively.

The concept of attenuation has been introduced earlier. In order to achieve a usable signal, the returning reflections must be amplified. The amplification can be done in multiple ways. Similar to a sound system, overall gain may be adjusted to increase or decrease the sensitivity of the received signal. More important, however, is the progressive loss of signal strength that occurs with reflections from deeper structures due to attenuation. To overcome this issue, ultrasound systems employ a variable gain circuit that selectively allows gain control at different depths. The applied gain is changed as a function of time (range) in the gain circuit, hence the term time gain compensation (TGC) is used to describe the process. This powerful tool can "normalize" the overall appearance of the image helping make much weaker returning echoes from great depth appear equal to near-field information (Fig. 18). The user also has slide pot control of gain as a function of depth. Some of this user-defined adjustment is applied as part of the TGC function or later as part of digital signal processing.

Manipulation of data prior to writing into the scan converter is called preprocessing. An important part of preprocessing is data compression. The raw data received by the transducer encompasses such a broad energy range

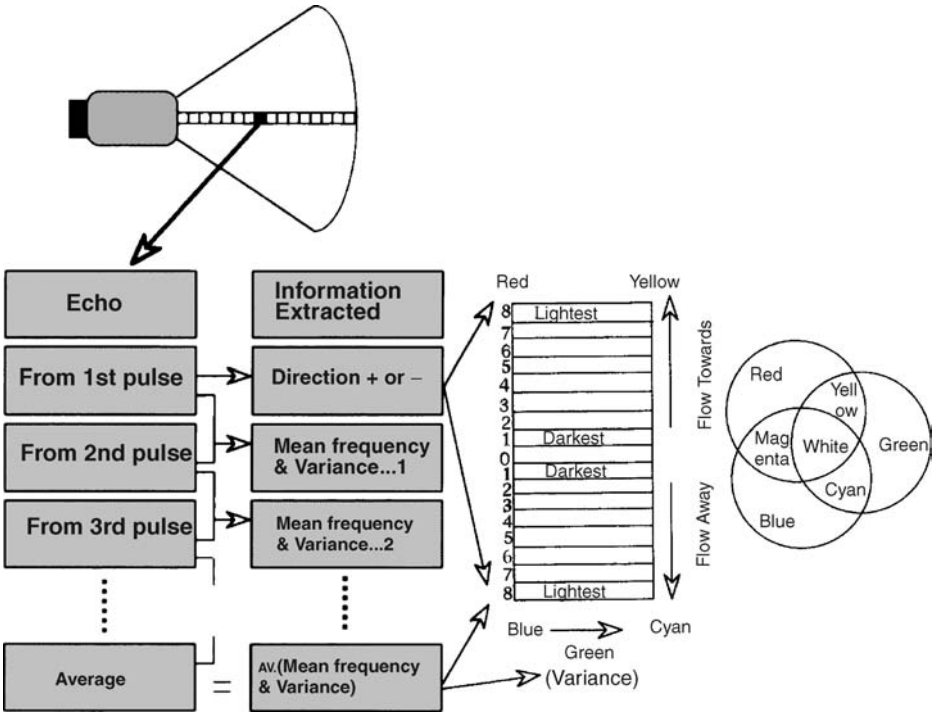


Figure 16. Diagram of color Doppler signal processing. At top, each box in the insonification line indicates a pulse of sound. A packet is made up of 4–8 pulses. The steps in the analysis cycle of a packet are shown in the "Echo" column and represent the comparison function of the autocorrelation system. From the first pulse the analysis determines the appropriate color representing direction. Comparisons between subsequent packets detect the velocity of blood flow. The brightness of the color selected corresponds to velocity. Comparisons of the variability of velocity are also done. Green is added proportional to the amount of variance. The final pixel color represents an average of the packet data for direction, velocity and variance. The right-hand column is the "color bar" that summarizes the type of map used, displays the range of color and brightness selected, and depicts how variance is shown. (From Sehgal CM, Principles of Ultrasound and Imaging and Doppler Ultrasound. In: Sutton, MG et al. editors: Textbook of Echocardiography and Doppler Ultrasound in Adults and Children (2nd ed). Oxford: Blackwell Science Publishing; 1996. p 29. Reprinted with permission of Blackwell Publishing, LTD. p 29)

that it cannot be adequately shown on a video display. Therefore, the dynamic range of the signal is reduced to better fit visual display characteristics. Selective enhancement of certain parts of the data is possible to better display borders between structures. Finally, persistence may be added to the image. With this enhancement, a fraction of data from the previous video frames at each pixel location may be added and averaged together. This helps define weaker diffuse echo scatterers and may work well in a static organ. However, with the heart in constant motion, only modest amounts of persistence add value to the image. Too much persistence reduces motion resolution of the image.

Once the digital signal is registered in the scan converter, further image manipulation is possible. This is called postprocessing of the image. Information in digital memory has a 1:1 ratio between ultrasound signal amplitude and video brightness. Depending on the structure imaged, considerable information may not be discernible in the image. With postprocessing, the relationship of video brightness to signal strength can be altered, frequently to enhance weaker echos and suppress high amplitude echoes. This may result in a better appreciation of less echogenic structures such as myocardium and the edges of the myocardium. The gray scale image may be transformed to a pseudo-color display that adds color to video amplitude data. In certain circumstances this may allow differentiation of pathologic changes from normal. Selective magnification of the image is also possible (Fig. 19).

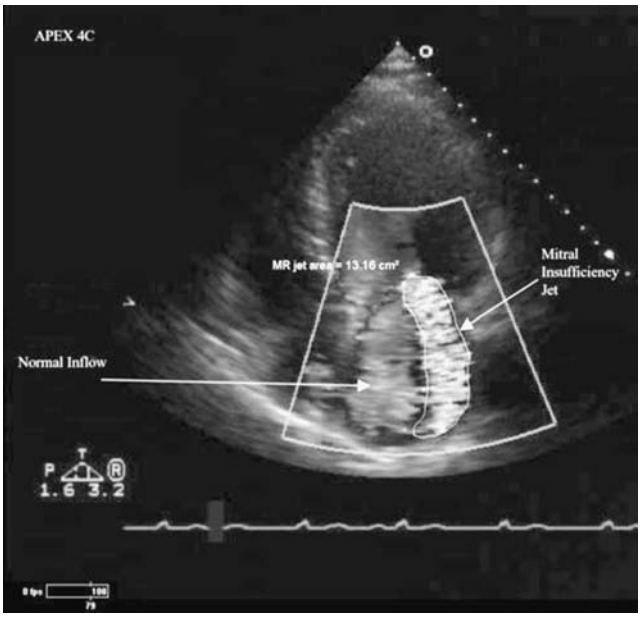


Figure 17. Color Doppler (here shown in gray scale only) depicting mitral valve regurgitation. The large mosaic mitral insufficiency jet is caused by turbulent flow coming through the mitral valve. The turbulent flow area has been traced to measure its area. The more uniform flow in the left atrium is caused by normal flow in the chamber. It is of uniform color and of much lower velocity.

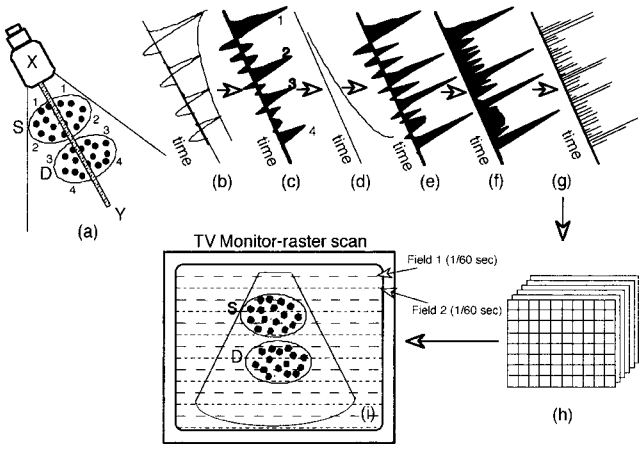


Figure 18. Processing steps during transmission of sound data to the scan converter. Raw data emerges from the transducer (a) that is imaging two objects that produce specular border echoes at points 1, 2, 3, and 4 and scattered echoes in between. (b) Specular echo raw data. (c) Low noise amplification is applied. Both specular and scattered echoes are now shown. Relative signal strength declines with depth due to attenuation of the signal, making signal 4 weaker than signal 1 even though the material border interface at borders 1 and 4 are identical. (d) Time gain compensation is applied proportionally by depth to electronically amplify signals received progressively later after the pulse leaves the transducer. This helps equalize perceived signal strength (e). The signal is then rectified (f) and smoothed (g) before entering the scan converter. (h) The process is repeated several times per second, in this case all new data appears every 1/30 of a second. The end result is a "real-time" video display of the two structures. (From Sehgal SH, Principles of Ultrasound Imaging and Doppler Ultrasound. In: Sutton MG et al. editors. Textbook of Echocardio graphy and Doppler Ultrasound in Adults and Children. Oxford: Blackwell Science; 1996. p 11. Reprinted with permission of Blackwell Publishing, LTD.)

Signal processing of PW and CW Doppler data includes filtering and averaging, but the most important component of the analysis is the computation of the velocity estimates. The most commonly used method of analysis is the fast Fourier transform analyzer, which estimates the relative amplitude of various frequency components of the input signal. This system (Fig. 20) divides data up into discreet time segments of very short duration (1–5 ms). For each segment, amplitude estimates are made for each frequency that corresponds to different velocity components in the flow and the relative amplitude of each frequency is recorded on gray scale display. Laminar flow typically has a narrow, discrete velocity range on PW Doppler while turbulent flow may be composed of the entire velocity range. Differentiation of laminar from turbulent flow may help define normal from abnormal flow states. Similar analysis is used to display the CW Doppler signal. Color Doppler displays can be adjusted by using multiple types of display maps. Each manufacturer has basic and special proprietary maps available to enhance color flow data.

All Doppler data can be subjected to selective band pass filters. For conventional Doppler imaging, signals coming from immobile structures or very slow moving structures such as chamber walls and the pericardium, are effectively blanked out. The range of velocity filtered can be changed

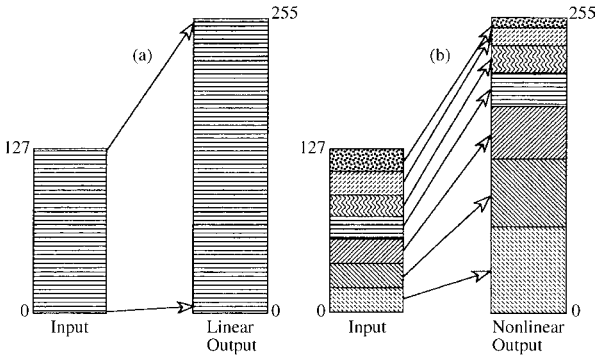


Figure 19. B-mode post processing occurs in which signal input intensity varies from 0 to 127 units. On the left side of the figure a linear output is equally amplified, the signal intensity range is now 0–255. On the right, nonlinear amplification is applied. The output is manipulated to enhance or compress the relative video intensity of data. In this example high energy specular reflection video data is relatively compressed (high numbers) while low energy data (from scattered echoes and weak specular echoes) is enhanced. Thus relatively more of the video range is used for relatively weaker signals in the final product. This postprocessing is in addition to time gain compensation done during preprocessing. (From Sehgal SH. Principles of Ultrasound Imaging and Doppler Ultrasound, In: Sutton MG et al. editors: Textbook of Echocardiography and Doppler Ultrasound in Adults and Children. Oxford: Blackwell Science; 1996. p 12. Reprinted with permission of Blackwell Publishing, LTD.)

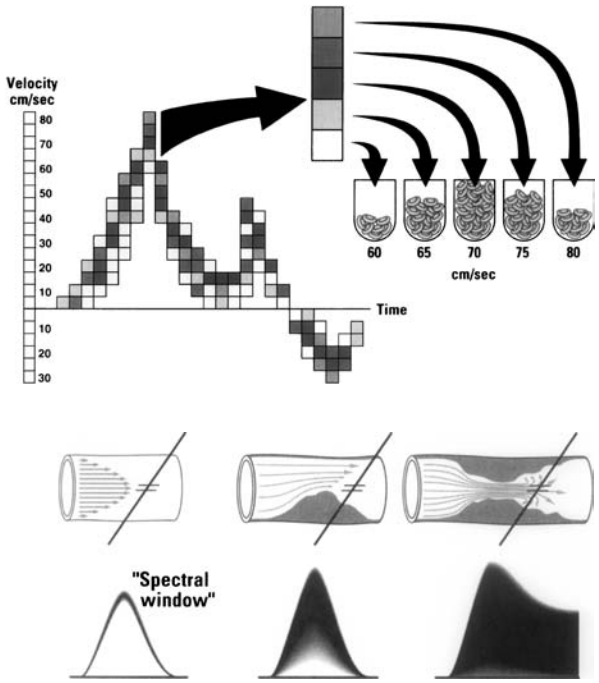


Figure 20. Upper panel Build-up of a spectral Doppler signal by Fast-Fourier analysis. The relative amplitude of the signal at each pixel location is assigned a level of gray. With laminar flow, the range of velocities is narrow resulting in a narrow window of displayed velocity. Lower panel As flow becomes more turbulent the range of velocities detected increases to the point that very turbulent signals may display all velocities. (Reprinted from Zagzebski JA. Essentials of Ultrasound Physics. St. Louis: Mosby-Year Book Inc. copyright © 1996, p 100,101, with permission from Elsevier.)

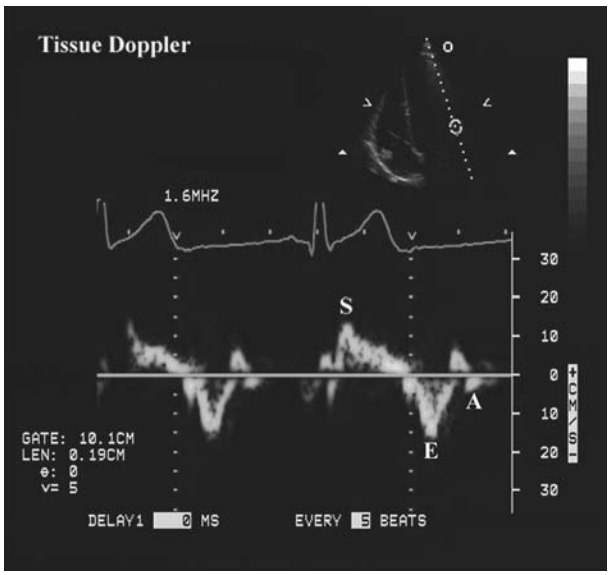


Figure 21. Example of tissue Doppler imaging. Duplex mode is used; the 2D image is shown in the upper insert. The sample volume has been placed outside the left ventricular cavity over the mitral valve annulus and is being used to detect movement of that structure. The three most commonly detected waveforms are shown: (S) systolic contraction wave, (E) early diastolic relaxation wave, and (A) atrial contraction wave.

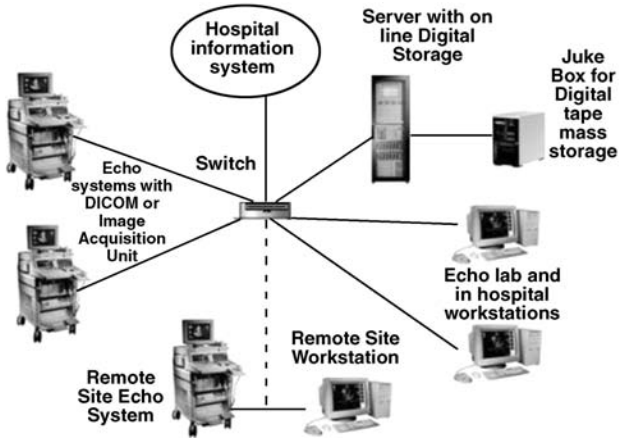


Figure 22. Organization chart showing a digital echo image storage system. Images generated by the ultrasound systems in the hospital and remote sites are input via high speed links through a switch to the server. The server has on line digital storage and is linked to a jukebox mass storage device. From the server, data may be viewed at any workstation either in the hospital or remotely. The hospital information system is linked to the system. This can allow electronic ordering of studies, downloading of demographic data on patients and interface with the workstations. Typically, images are reviewed at a workstation; a report is composed and is electronically signed. The report is stored in the server but may also be sent to the Hospital Information System electronic record or put out on paper as a final report.

for different clinical circumstances. In some situations, not all of the slower moving structures can be fully eliminated without losing valuable Doppler data, resulting in various types of artifact.

New techniques of Doppler analysis focus on analyzing wall motion with Doppler and displaying the data either in color or with Fourier transform analysis. In this setting, the band pass filters are set to eliminate all high velocity data from blood flow and only analyze very low velocity movement coming from the wall of the ventricles or other structures such as valve tissue. Analysis of tissue motion measures the velocity of wall movement at selected locations in the heart using the typical PW sample volume. Conventional strip chart display is used and velocity can be displayed to represent both the velocity of contraction and the velocity of relaxation (Fig. 21). Color Doppler uses the auto correlator system to calculate velocity of large segments of myocardium at once. Systems that record data at high color Doppler frame rates store sufficient information to allow selective time—velocity plots to be made at various locations of the cardiac chamber walls. Color Doppler may also be utilized to calculate strain and strain rate, another alternate display mode being investigated as a parameter of ventricular function (10,11).

Once the video signal processing is complete, it is displayed on a monitor. Until recently, most monitors were analogue and used standard NTSC video display modes. The composite (or RGB) signal was sent to a videocassette recorder for long-term storage and playback. The study could then be played on a standard video playback system for review and interpretation. Many laboratories continue to use this method of recording and storage of images.

Recently, many echo laboratories have changed to digital image display and storage. In some labs, the ultrasound system directly records the study in digital format, storing data on a large hard disk in the ultrasound system. The data is then output, typically using a DICOM standard signal to a digital storage and display system. Other laboratories may attach an acquisition box to the ultrasound system that digitally records and transmits the RGB signal to a central image server.

A digital laboratory set up is shown in Fig. 22. It has several advantages over videotape. Image data is sent to a server and on to a digital mass storage device. Depending on the amount of digital storage and volume of studies in the lab, days to months of image data may be instantly available for review. More remote data is stored in a mass storage system (PACS system) on more inexpensive media such as digital tape. Data in this remote storage may be held online in a second "juke box" server or be fully off line in storage media that must be put back on line manually.

Digital systems link the entire lab together and are typically integrated with the hospital information system. Common studies may be organized by patient and displayed on multiple workstations within the laboratory, hospital, and at remote clinics. This is a marked improvement compared to having each study isolated to one videotape. The quality of the image is superior. High-speed fiber optic links allow virtually simultaneous image retrieval at remote sites. Lower speed links may need several minutes to transmit a full study. To reduce the amount of storage, studies are typically compressed. Compression ratios of 30:1 can be used without evidence of any significant image degradation.

THE ECHOCARDIOGRAPHIC EXAMINATION

The Transthoracic Exam

A full-featured cardiac ultrasound system is designed to allow the operator to perform an M-mode echo, 2D echo, CW Doppler, PW Doppler, color Doppler, and tissue Doppler examination. Except for the specialized CW Doppler transducer, the entire exam is usually performed with a broadband multipurpose transducer. Occasionally, a specialized transducer of a different frequency or beam focus must be utilized to interrogate certain types of suspected pathology. The examination is performed in a quiet, darkened room with the patient supine or lying in a left lateral position on a specialized exam bed. The transducer is covered with a coupling medium (gel-like substance) that allows a direct interface between the transducer head and the patient's skin. If this coupling is broken the signal will be lost since ultrasound reflects strongly from tissue–air interfaces (Table 1). The transducer is then placed between the ribs, angled in the direction of the heart (ultrasound penetrates bone poorly and bone causes image artifacts), and adjusted until a satisfactory image is obtained (Fig. 2a). Electrodes for a single channel electrocardiogram are applied to the patient. The electrocardiogram signal is displayed continuously during the exam.

A standard examination begins in the left parasternal position (i.e., in the fourth and fifth rib interspaces just left of the sternum) (Fig. 2b). By orienting the 2D plane parallel to the long axis of the heart, an image of the proximal aorta, aortic valve, left atrium, mitral valve, and left ventricle can be obtained (Fig. 23). The standard M-mode echocardiographic views for dimension measurement are also obtained from this view (Fig. 6). Color Doppler is activated to examine flow near the mitral and aortic valves. By angulation from this position the tricuspid valve and portions of the right atrium and right ventricle are brought into view (Fig. 24). By rotation of the transducer $\sim 90^\circ$ from the long axis, the parasternal short-axis view is obtained.

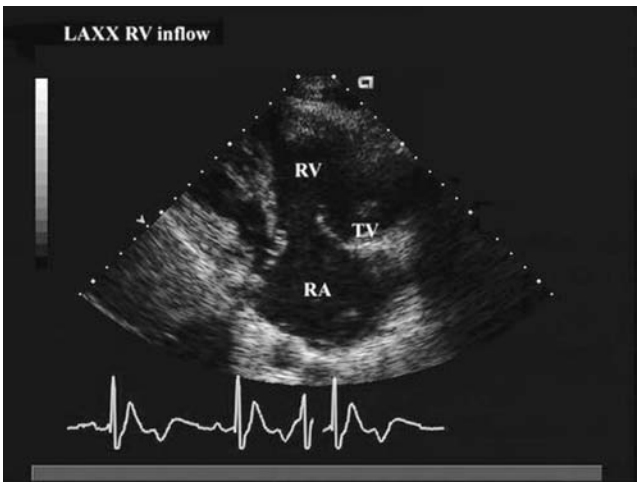


Figure 24. Parasternal long axis view angled toward the right ventricle. Right-sided structures are shown: right atrium (RA), right ventricle (RV) and tricuspid valve (TV).

By progressively changing the angle of the transducer it is possible to obtain a series of cross-sectional views through the left ventricle from near the apex to the base of the heart at the level of the origin of the great vessels (Figs. 25 and 26). In an optimal study, several cross-sections through the left ventricle, mitral valve, aortic valve, and to a lesser degree the tricuspid valve, pulmonic valve, and right ventricular outflow tract can be obtained. Since conventional 1D arrays yield only 2D images, only small slices of the heart are examined at any one time.

The transducer is then moved to the mid-left chest slightly below the left breast where the apex of the heart touches the chest wall (Fig. 2). The 2D transducer is then angled along the long axis toward the base of the heart. The result is a simultaneous display of all four chambers of the heart and the mitral and tricuspid valves (Fig. 27). No M-mode views are taken from this position. Using duplex

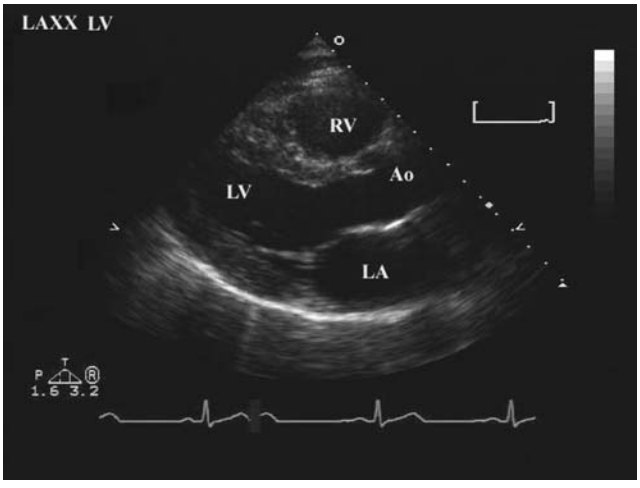


Figure 23. Parasternal long axis view of the heart (LAXX) similar to orientation shown in Fig. 6. Abbreviations are as follows: aorta (Ao), left ventricle (LV), left atrium (LA), and right ventricle (RV). This is a 2D image.

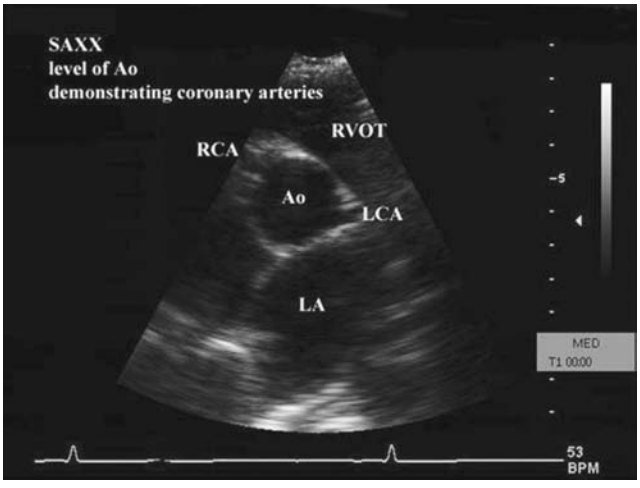


Figure 25. Parasternal short axis view at the level of the great vessels. Shown is the aorta (Ao), a portion of the left atrium (LA), the origin of the left (LCA), and right (RCA) coronary arteries, and a part of the right ventricular outflow tract.

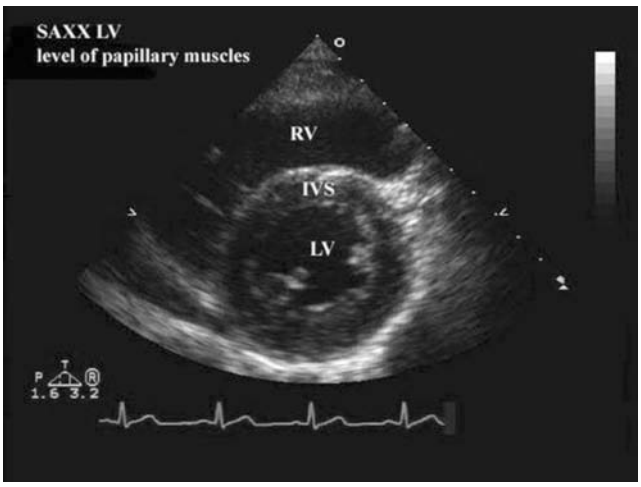


Figure 26. Short axis orientation at the transducer (SAXX) showing a cross-sectional view of the left ventricle (LV). The muscle appears as a ring. The septum (IVS) separates the LV from the right ventricle (RV).

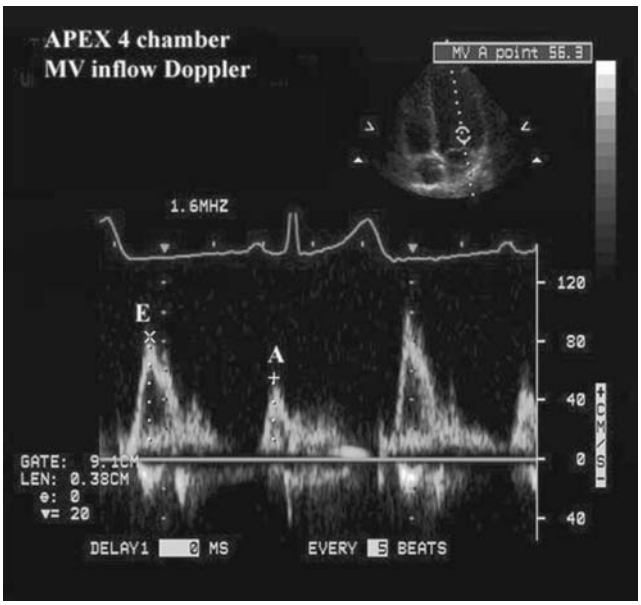


Figure 28. By using the same transducer orientation as Fig. 27, several other types of data are obtained. In this case, the system has been switched to Duplex mode. The 2D echo is used as a guide to position the mitral sample volume (small upper image). Flow is obtained across the mitral valve (MV) with waveforms for early (E), and atrial (A) diastolic flow shown.

mode, PW Doppler samples of blood flow are taken to show the forward flow signal across the mitral and tricuspid valves (Fig. 28). By further anterior angulation, the left ventricular outflow tract and aortic valve are imaged (Fig. 29) and forward flow velocities are sampled at each site. The 2D image allows precise positioning of the Doppler sample volume. By rotation of the transducer, further selective portions of the walls of the left ventricle are obtained (Fig. 30). Color Doppler is then used to sample blood flow across each heart valve, screening for abnormal turbulent flow, particularly related to valve insufficiency (leakage during valve closure). The transducer is then moved to a location just below the sternum (Fig. 2b) and aimed up toward the heart where the right atrium and atrial septum can be further interrogated, along with partial views of the other chambers (Fig. 31). In some patients, cross-sectional views equivalent to the short-axis view may be obtained.

However, this view in most patients is less useful because the heart is further from the transducer causing reduced resolution and increased attenuation of the echo signals. Finally, the heart may be imaged from the suprasternal approach (Fig. 2b), which generally will allow a view of the ascending aorta and aortic arch (Fig. 32).

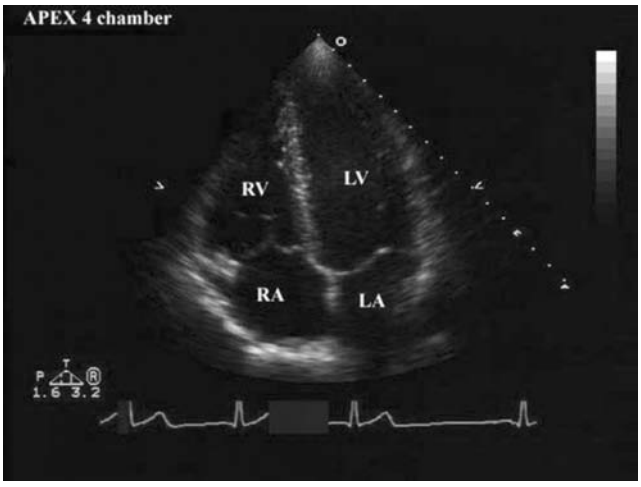


Figure 27. The transducer is placed in the apical position. The view shown is the apical "four chamber" view since all four chambers of the heart can be imaged simultaneously. Abbreviations are as follows: left ventricle (LV), left atrium (LA), right atrium (RA), and right ventricle (RV).

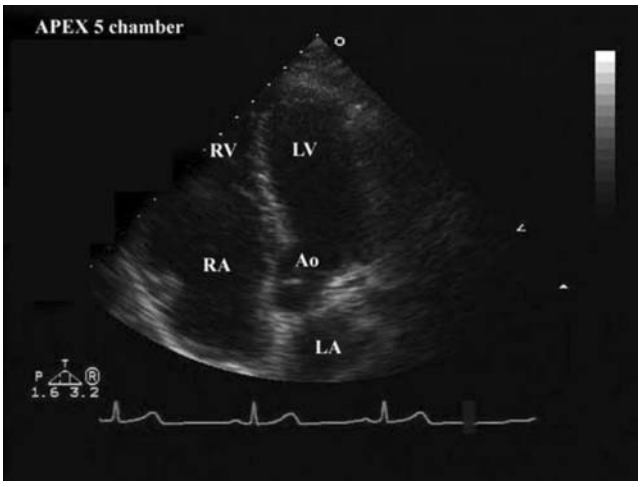


Figure 29. From the view shown in Fig. 27, the transducer has been tilted slightly to show the outflow region of the left ventricle (LV) where blood is ejected toward the aorta (Ao). Abbreviations are as follows: left atrium (LA), right atrium (RA), and right ventricle (RV).

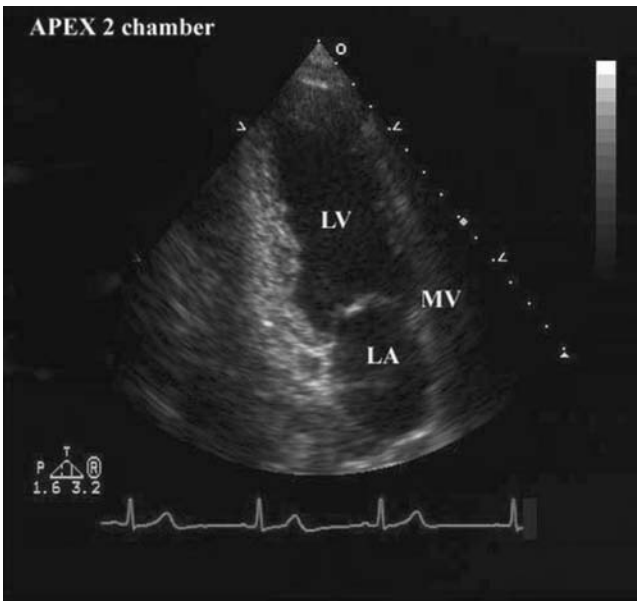


Figure 30. From the position shown in Fig. 27, the transducer is rotated to show a different perspective from the apical view. The "two-chamber" view shows only the left-sided chambers and the mitral valve (MV) that controls flow between these two chambers. Abbreviations are as follows: left ventricle (LV) and left atrium (LA).

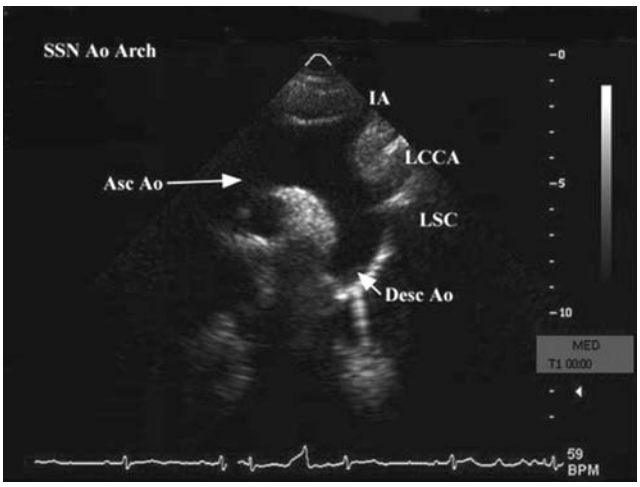


Figure 32. The transducer is positioned at the suprasternal notch (see Fig. 2**b**). A portion of the aorta (Asc Ao and Dsc Ao) is visible along with the origins of three branches coming off of this structure: innominate artery (IA), left common carotid artery (LCCA) and left subclavian (LSC).

The Transesophageal Exam

About 3–10% of hospital-based echocardiograms are performed using a specialized transesophageal (TEE) device. The ultrasound transducer, smaller but otherwise of virtually equal capability to the transthoracic device, is attached to the end of an endoscope. The patient is prepared using a topical anesthetic agent in the mouth and pharynx to eliminate the gag reflex and given conscious sedation to increase patient comfort. Patient status is monitored by a nurse, the ultrasound system operated by a sonographer and the transducer inserted by a physician who personally performs the exam. As the transducer is passed down in the esophagus, multiple imaging planes are obtained. These planes are obtained by a combination of movement of the transducer to different levels of the esophagus, changing the angle of the transducer and controlling the transducer head. Originally, the transducer was fixed in one location on the endoscope. Virtually all devices now made allow the operator to rotate the transducer through a 180° arc markedly increasing the number of imaging planes in the exam (Figs. 33–35). Using this method multiple views of structures in both a long and short axis configuration are possible. The transducer may also be passed into the stomach where additional views are possible.

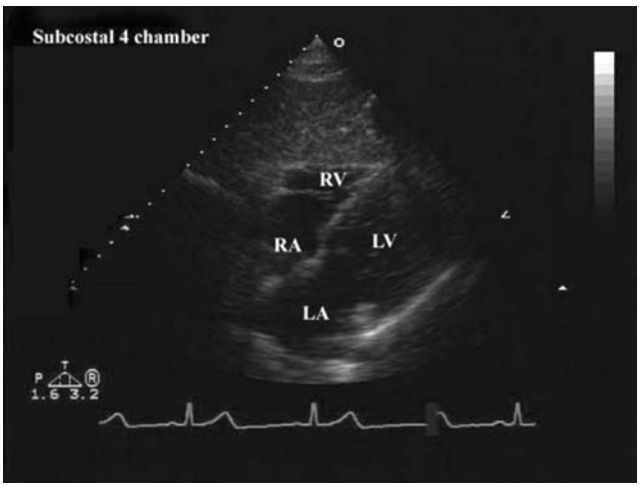


Figure 31. The transducer is moved to the subcostal position (see Fig. 2**b**). Portions of all four chambers are visible. Abbreviations are as follows: left atrium (LA), left ventricle (LV), right atrium (RA), and right ventricle (RV).

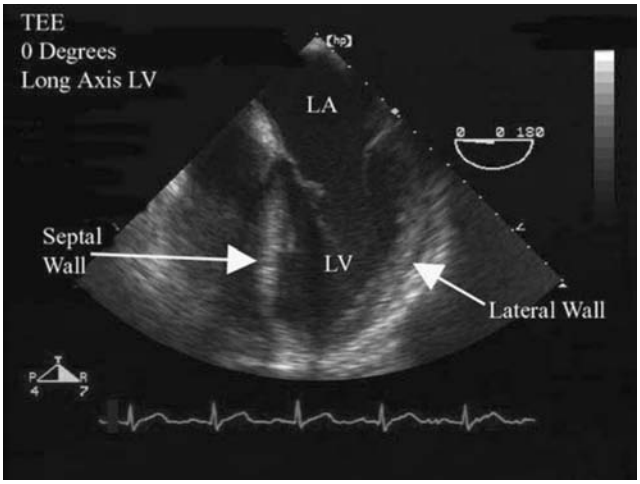


Figure 33. Image generated by a transducer placed in the esophagus. Abbreviations are as follows: left ventricle (LV) and left atrium (LA).

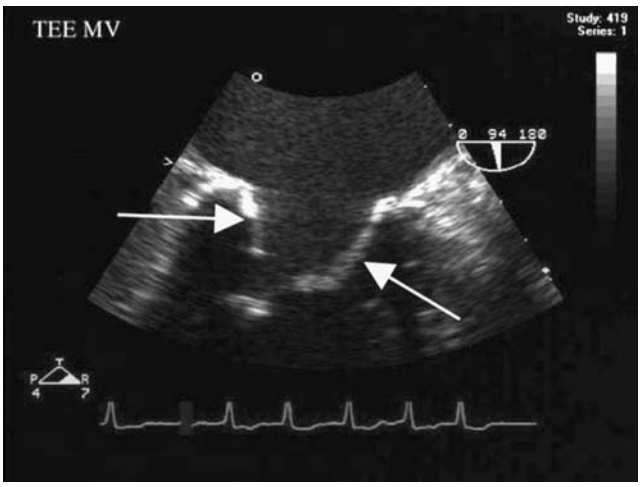


Figure 34. Image from a transesophageal transducer. This is a magnified view of the mitral valve leaflets (arrows).

The exam is performed in a similar fashion to the transthoracic exam in that multiple views of cardiac chambers and valves are taken in an organized series of views to achieve a complete examination of the heart. Many planes are similar to the transthoracic exam except for location of the transducer. Other views are unique to the TEE exam and may be the clinical reason the exam was performed. Once the exam is complete, the transducer is removed, and the patient is monitored until recovered from the sedation and topical anesthesia.

The Intracardiac Exam

During specialized procedures it may be valuable to examine the heart from inside. This is usually done during specialized sterile cardiac procedures when a transthoracic transducer would be cumbersome or impossible to use and a TEE also would be impractical.

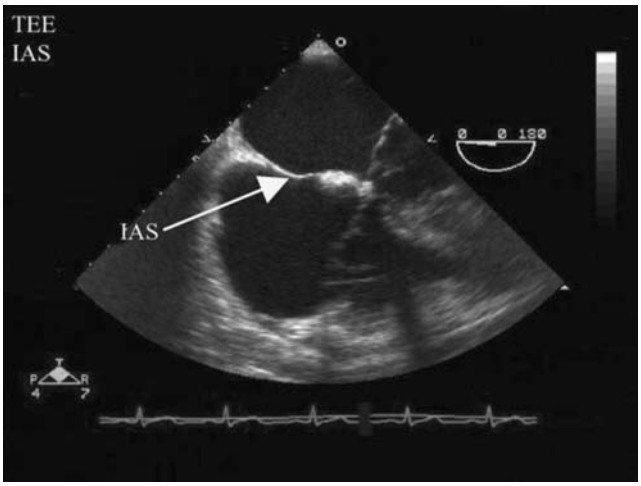


Figure 35. Image from a transesophageal transducer. The septal structure (IAS) that separates the two atrial chambers is shown (arrow).

A catheter with a miniaturized transducer is passed under fluoroscopic guidance up large central veins from the groin to the right side of the heart. For most applications, the device may be placed in the right atrium, or infrequently, in the right ventricle. The transducer is a miniature phased array device that can provide 2D and Doppler information. In most cases, the clinical diagnosis is already known and the patient is undergoing a specialized cardiac procedure in the catheterization laboratory to close a congenital defect or monitor radio frequency ablation during a complex electrophysiologic procedure, attempting to eliminate a cardiac rhythm disorder. The device contains controls that allow change of angle in four directions. This, combined with positional placement of the catheter in different parts of the cardiac chambers, allows several different anatomic views of cardiac structures (Fig. 36).

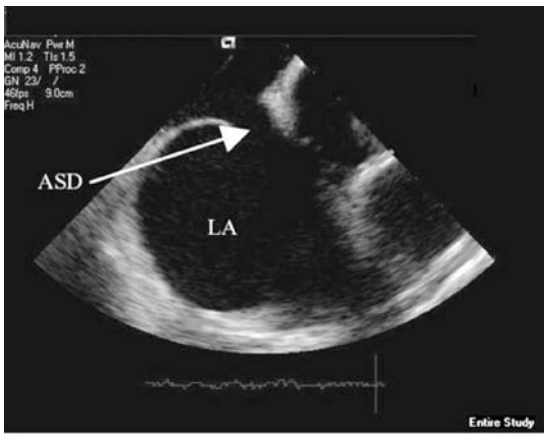
The Stress Echo Exam

Combining a transthoracic echo with a stress test was first attempted in the 1980s, became widely available in the mid-1990s, and is now a widely utilized test for diagnosis of coronary artery disease. The echo exam itself is purposely brief, usually confined to 4–6 2D views.

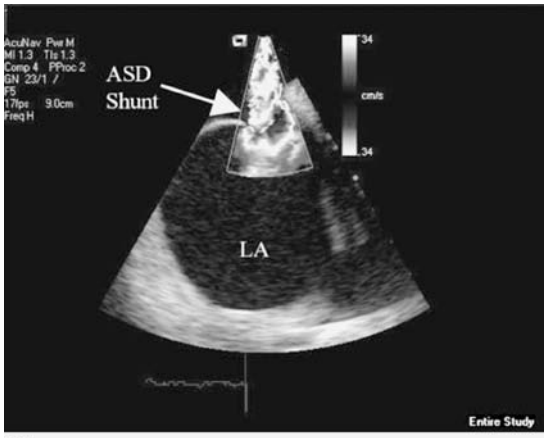
A patient having a stress echo is brought to a specialized stress lab that contains an ultrasound system set up to acquire stress images, an exam table, and a stress system that most commonly consists of a computerized electrocardiography system that runs a motorized treadmill (Fig. 37). The patient is connected to the 12-lead electrocardiogram stress system and a baseline electrocardiogram is obtained. The baseline echo is next performed, recording 2D views (Fig. 38). Then the stress test is begun. It can be performed in two different formats:

1. Exercise: If the patient is able to walk on a treadmill (or in some labs, pedal a bike), a standard maximal exercise test is performed until maximum effort has been achieved. Immediately upon completion of exercise, the patient is moved back to the echo exam table and repeat images are obtained within 1–2 min of completion of exercise.
2. Pharmacologic stimulation: If the patient is unable to exercise, an alternative is stimulation of the heart with an intravenous drug that simulates exercise. Most commonly, this is dobutamine, which is given, in progressively higher doses in 3–5 min stages. The patient remains on the exam table the entire time, connected to the electrocardiographic stress system. 2D echo images are obtained; at baseline, low dose, intermediate dose, and peak dose of drug infusion during the exam.

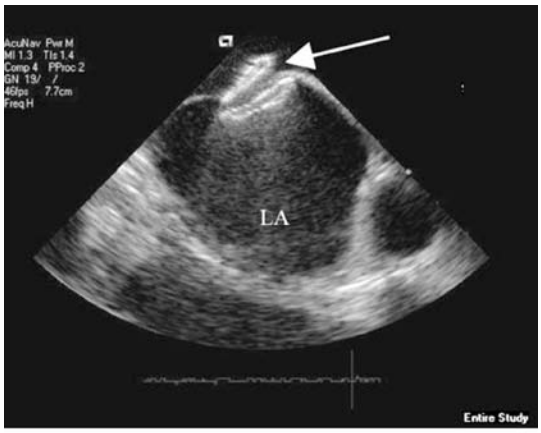
In both types of tests, clinical images are recorded and then displayed so that pretest and posttest data can be examined side by side. Comparisons of cardiac function are carefully made between the prestress and poststress images. The test relies on perceived changes in mechanical motion of different anatomic segments of the heart. Both inward movement of heart muscle and thickening of the walls of the ventricles are carefully evaluated. The normal



(a)



(b)



(c)

Figure 36. (a) Image of the atrial septum in a patient with an atrial septal defect (arrow). This image was taken with an intracardiac transducer placed in the right atrium (RA). The image is of similar quality to that seen in Fig. 35. (b) Image of same structure as (a) demonstrating color flow capabilities at the intracardiac transducer. Color Doppler (seen in gray scale) confirms the structure is a hole with blood passing through the hole (ASD shunt). (c) Image taken from same position as (a) and (b). The atrial septal defect has been closed with an occluder device (arrow). Doppler was used (not shown) to confirm that no blood could pass through the previously documented hole.

Figure 37. Typical set-up of a laboratory to perform stress echocardiograms. Shown are the echocardiographic examination table (ET), the treadmill (T), the treadmill computer control system (TC), and the ultrasound system (US).

heart responds by augmenting inward motion and wall thickening in all regions. If blood supply to a given segment is not adequate, mechanical motion and wall thickening either fails to improve or deteriorates in that segment, but improves in other segments. This change defines an abnormal response on a stress echo (Fig. 38). The location and extent of abnormal changes in wall motion and thickening are reported in a semiquantitative manner. In addition, the electrocardiogram response to stress is also compared with the baseline exam and reported along with patient symptoms and exercise capacity.

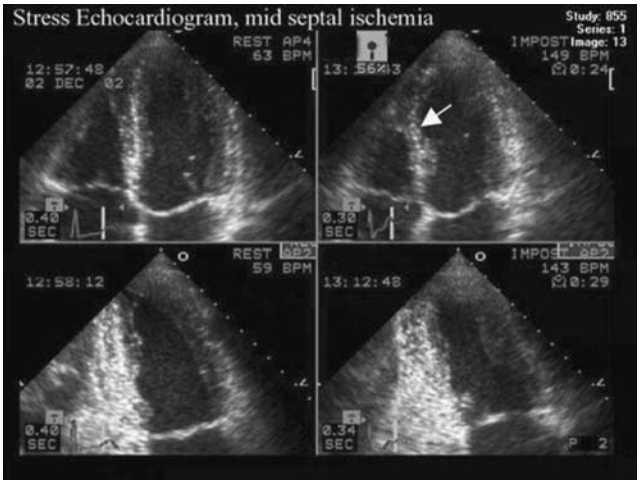


Figure 38. Quad screen format of view obtained from a stress echocardiography examination. Four views are displayed simultaneously, and the particular views shown can be changed to accommodate direct comparison of pre- (images on the left) and posttest images on the right views of the same part of the heart as shown in the example. The arrow indicates an area of the heart that failed to respond normally.

CLINICAL USES OF ECHOCARDIOGRAPHY

M-Mode Echocardiography

The M-mode echo was the original cardiac exam and for years was the dominant cardiac ultrasound study performed. The 2D echo has superseded M-mode echo as the primary examination technique and in most circumstances is superior. However, many laboratories continue to perform at least a limited M-mode exam because dimension measurements are well standardized. In addition, due to its high sampling rate, M-mode echo is superior to 2D echo for timing of events within the cardiac cycle and for recording simultaneously, with other physiologic measurements such as phonocardiograms and pulse tracings. Certain movement patterns of heart valves and chamber walls are only detected by M-mode, thus providing unique diagnostic information (Fig. 39).

Most M-mode echo data is obtained from multiple different "ice pick" views, all obtained from the parasternal position (Fig. 7). The 2D image is used to direct the cursor position of these views. The first view angles through a small portion of the right ventricle, the ventricular septum, the left ventricular chamber, the posterior wall of the left ventricle, and the pericardium. From this view, left ventricular wall thickness and the short-axis dimensions of this chamber can be measured. By calculating the change in dimension between end diastole (at the beginning of ejection) and end systole (at the end of ejection), the fractional shortening can be measured using the equation:

$$\frac{\text{LVEDD} - \text{LVESD}}{\text{LVEDD}} = \text{fractional shortening}$$

where LVEDD is the left ventricular end diastolic dimension and LVESD is the left ventricular end systolic dimension. This measurement estimates left ventricular function (Fig. 40). Dimension measurements give a relatively precise estimate of left ventricular chamber size to calculate whether the ventricle is inappropriately enlarged. In a similar manner wall thickness measurements can be utilized to determine if the chamber walls are inappropriately thick (left ventricular hypertrophy), asymmetrically thickened (hypertrophic cardiomyopathy), or inappropriately thin (following a myocardial infarction).

The second ice pick view passes through the right ventricle, septum, mitral valve leaflets, and posterior wall. This view is used primarily to evaluate motion of the mitral valve. Certain types of mitral valve disease alter the pattern of motion of this valve (Fig. 39). Other abnormal patterns of leaflet motion may indicate dysfunction elsewhere in the left ventricle.

The third ice pick view passes the echo beam through the right ventricle, aortic valve, and left atrium. From this view, analogous to the mitral valve, the pattern of aortic valve motion will change in characteristic ways allowing the diagnosis of primary aortic valve disease or diseases that cause secondary aortic valve motion changes. Also, from this view the diameter of the left atrium is measured and whether this structure is of normal size or enlarged can be of considerable importance in several circumstances. Other views are possible, but rarely used.

Measurements taken from M-mode may be performed during the exam by the sonographer. Usually, the scrolling M-mode video display is saved in digital memory on the system. The saved data is then reviewed until the best depiction of the various views discussed above is displayed. The sonographer then uses electronic calipers, automatically

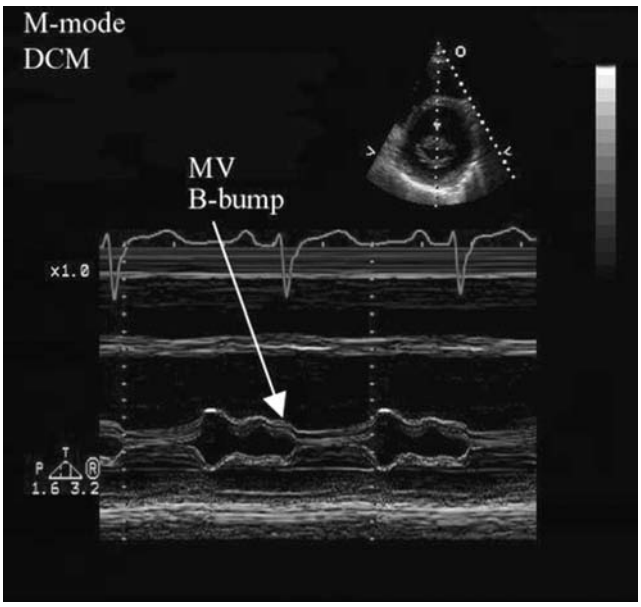


Figure 39. Example of an M-mode echocardiogram of the mitral vale showing abnormal motion, in this case a "B-bump" indicating delayed closure of the valve. DCM = dilated cardiomyopathy.

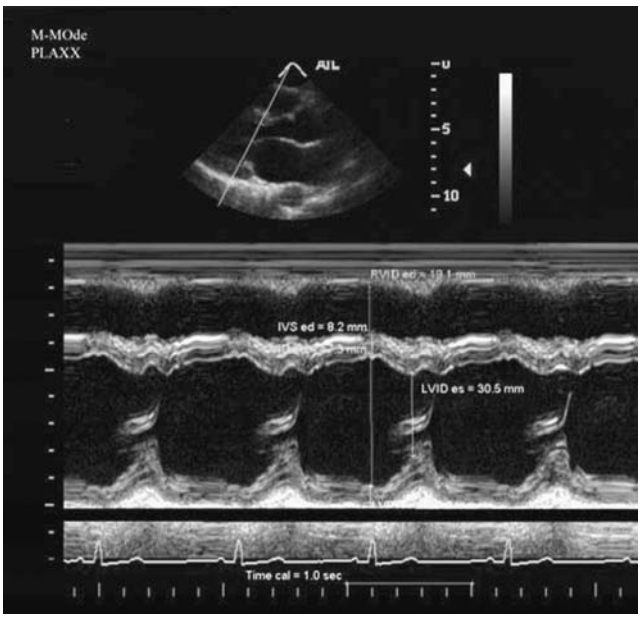


Figure 40. Example of an M-mode echocardiogram of the left ventricle. The typical measurements of wall thickness (IVSed and LVPWed) and chamber dimensions (RVIDed, LVIDed and LVIDes) are shown. Several calculated parameters are possible from these values.

calibrated by the system, to record dimension measurements (Fig. 40). Basic measures are made and formulas for derived data such as the % FS are automatically calculated. Measurements can also be made "off line" using special workstations that display the video information at the time the study is interpreted by the physician.

Two-Dimensional Echocardiography

The 2D exam gives information about all four cardiac chambers and all four cardiac valves. It also serves as the reference point for positioning all Doppler sample volumes and the color Doppler exam. The heart is imaged from multiple positions, which not only improves the chance of useful information being obtained, but also allows better characterization of a given structure because the structure is seen in several perspectives. The primary role of a 2D echo is to characterize the size of the left and right ventricles as normal or enlarged, and if enlarged, estimate the severity of the problem. Second, the 2D exam evaluates pump function of the two ventricles. Function is characterized globally (i.e., total ventricular performance) or regionally (i.e., performance of individual parts of each ventricle). Some types of disease affect muscle function relatively equally throughout the chambers. Other forms of disease, most notably coronary artery atherosclerosis, which selectively changes blood supply to various parts of the heart, cause regional changes in function. In this disease, some portions of the heart may function normally while other areas change to fibrous scar and decrease or stop moving entirely. Global function of the left ventricle can be characterized quantitatively. The most common measurements of function use calculations of volume during the cardiac cycle. This is quantified when the heart is filled maximally just before a beat begins and minimally just after ejection of blood has been completed. The volume calculations are used to determine of ejection fraction. The equation is

$$\text{Ejection fraction} = \frac{\text{LVEDV} - \text{LVESV}}{\text{LVEDV}} \times 100$$

where LVEDV = left ventricular end diastolic volume and LVESV = left ventricular end systolic volume.

The 2D echo is sensitive for detecting abnormalities within the chambers, such as blood clots or vegetations (infectious material attached to valves). Abnormalities surrounding the heart, such as pericardial effusions (fluid surrounding the heart), metastatic spread of tumors to the heart and pericardium, and abnormalities contiguous to the heart in the mediastinum or great vessels can be readily imaged. Most of this information is descriptive in nature (Figs. 41 and 42).

The ability to directly and precisely make measurements from 2D echo views for quantitative measurements of dimensions, areas, and volumes is built into the ultrasound system and is typically done by the sonographer during the exam in a similar fashion to M-mode. Further measurements may be performed off line on dedicated analysis computers. Many parts of the interpretation of the exam, however, remain primarily descriptive and are usually estimated by expert readers.

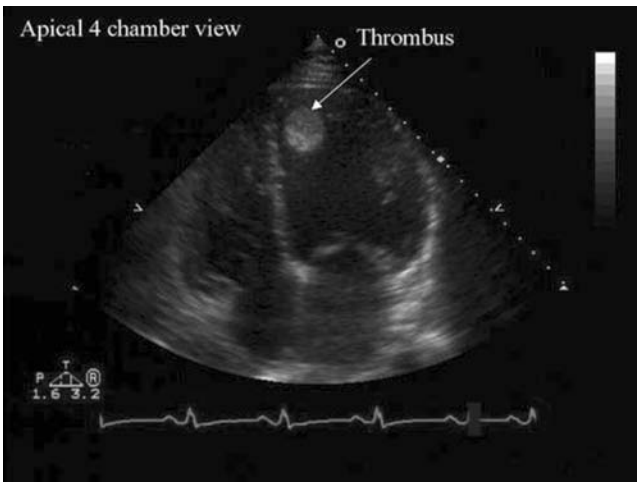


Figure 41. Example of detection of a blood clot (thrombus) in the left ventricular chamber.

Doppler Echocardiography

While the 2D echo has considerably expanded the ability to characterize abnormalities of the four heart valves, it has not been possible to obtain direct hemodynamic information about valve abnormalities by imaging alone. Doppler imaging provides direct measurement of hemodynamic information.

By using Doppler, six basic types of information can be obtained about blood flow across a particular region:

1. The direction of the blood flow.
2. The time during the cardiac cycle during which blood flow occurs.
3. The velocity of the blood flow.
4. The time the peak velocity occurs.
5. The rate at which velocity changes.

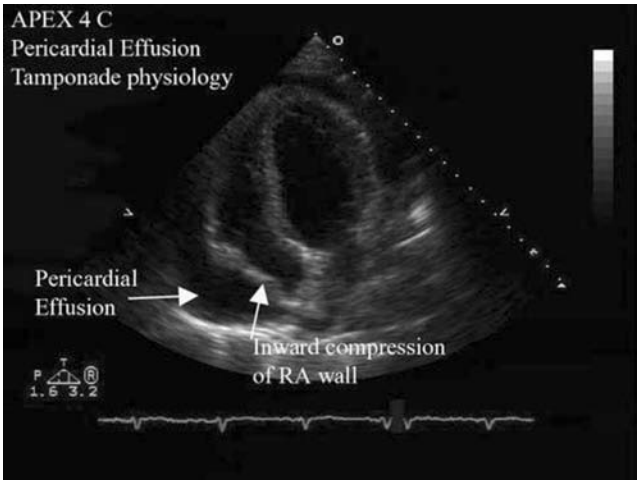


Figure 42. Example of a patient with fluid surrounding the heart. The dark area surrounding the heart (pericardial effusion) is shown. In this case, increased pressure caused by the effusion compresses part of the right atrium (RA).

6. The pressure drop or gradient across a particular valve or anatomic structure.

Data about pressure gradients is derived from the velocity measurement using the Bernoulli equation:

$$P_1 - P_2 = \underbrace{1/2\rho(V_2^2 - V_1^2)}_{\text{Convective acceleration}} + \underbrace{\rho \int_1^2 \frac{dv}{dt} ds}_{\text{Flow acceleration}} + \underbrace{R\ (v)}_{\text{Viscous friction}}$$

where $P_l - P_2$ is the pressure drop across the structure V_2 and V_1 being blood flow velocity on either side of the structure, and ρ is the mass density of blood (1.06×10^3 kg/m^3). For applications in the heart, the contributions by the flow acceleration and viscous friction terms can be ignored. In addition, V_1 is generally much less than V_2 (thus, V_1 can usually be ignored), and ρ is a constant for the mass density of blood (6). Combining all these changes together results in the final "simplified" form of the Bernoulli equation:

$$P_1 - P_2 = 4\,V^2$$

Cardiac Output. When the heart rate, blood flow velocity integral, and cross-sectional area of the region across which the blood flow is measured are known, cardiac output can be estimated using the following equation:

$$\text{CO} = A \times V \times \text{HR}$$

where CO is the cardiac output, A is the cross-sectional area, V is the integrated blood flow velocity, and HR is the heart rate.

The Character of the Blood Flow. The differentiation between laminar and turbulent blood flow can be made by observation of the spectral pattern. In general, laminar flow (all recorded velocities similar) occurs across normal cardiac structures of the heart, while disturbed or turbulent flow (multiple velocities detected) occurs across diseased or congenitally abnormal cardiac structures (Fig. 20).

Doppler is most valuable in patients with valvular heart disease and congenital heart disease. In the case of valve stenosis (abnormal obstruction to flow), use of Doppler echocardiography allows quantification of the pressure gradient across the valve (Fig. 43). Using the continuity principle, which states that the product of cross-sectional area and flow velocity must be constant at multiple locations in the heart, it is possible to solve for the severity of valve stenosis. The equation may be written as noted and then manipulated to solve for the area at the stenotic valve (A_2).

$$A_1 V_1 = A_2 V_2$$
$$\frac{A_1 V_1}{V_2} = A_2$$

For valvular insufficiency, Doppler echocardiography is most useful when the color Doppler format is used in conjunction with 2D echo. Since an insufficient valve (i.e., a valve that allows backward leakage of blood when closed) produces turbulent flow in the chamber behind the valve, color Doppler immediately detects its presence. The extent to which turbulent flow can be detected is then graded on a semiquantitative basis to characterize the amount of valve insufficiency (Fig. 17). Since only 2D are interrogated at any given time, the best results are obtained when Doppler sampling is done from more than one view.

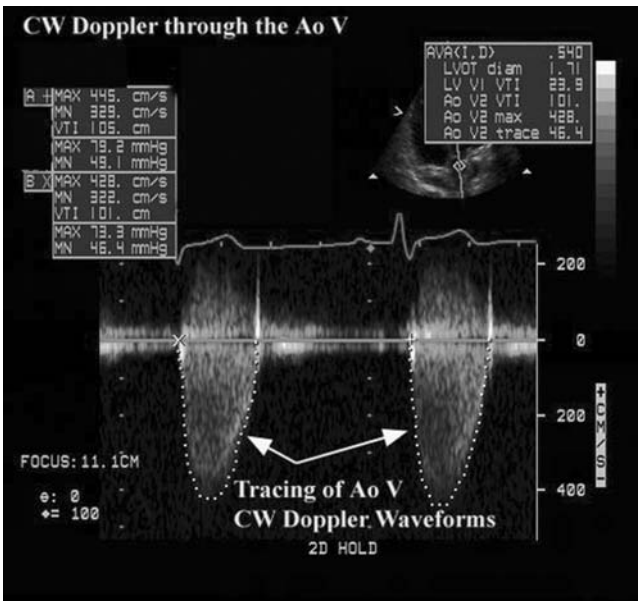


Figure 43. Example of a continuous wave Doppler signal through the aortic valve (AoV). Two tracings are shown along with measurement technique also demonstrated.

In patients with congenital heart disease, Doppler echocardiography allows the tracing of flow direction and velocity across anatomic abnormalities, such as holes between various cardiac chambers (i.e., atrial or ventricular septal defects). It can also display gradients across congenitally malformed valves and great vessels and also determine the direction and rate of flow through anatomically mal positioned chambers and great vessels.

In addition to direct interrogation of heart valves for detection of primary valve disease, Doppler flow sampling is used to evaluate changes in flow across normal valves that may indicate additional pathology. For example, the systolic blood pressure in the lungs (pulmonary artery pressure) may be estimated by quantifying the velocity of flow of an insufficiency jet across the tricuspid valve (another application of the Bernoulli equation). This calculated value, when added to the estimated central venous pressure (obtained in a different part of the exam) gives an excellent estimate of pulmonary artery pressure.

A second important measurement involves characterizing the way the left ventricle fills itself after ejecting blood into the aorta. There is a well-described set of changes in the pattern of flow across the mitral valve, changes in flow into the left atrium from the pulmonary veins and changes in the outward movement in the muscle itself characterized

by tissue Doppler that can help categorize the severity of changes in filling of the left ventricle.

SPECIALIZED CLINICAL DATA

Transesophageal Echocardiography

The TEE exam is used when a transthoracic exam either cannot be performed or gives inadequate information. One limitation of echo is the great degree of variability in image quality from patient to patient. In some circumstances, particularly in intensive care units, the TEE exam may provide superior image quality since its image quality is not dependent on patient position or interfered with by the presence of bandages, rib interfaces, air or other patient dependent changes. Similarly, during open heart surgery a TEE is routinely used to assess cardiac function pre- and postintervention and pre- and postheart valve replacement or repair. Since the TEE probe is in the esophagus, outside of the surgeon's sterile field images are obtained even when the patient's chest is open. This capability allows the surgeon to evaluate the consequences of, for example, a surgical repair of a heart valve, when the heart has been restarted, but before the chest is sutured closed. The TEE exam also visualizes parts of the heart not seen by any transthoracic view. A particular example of this is the left atrial appendage, a part of the left atrium. This structure sometimes develops blood clots that can only be visualized by TEE.

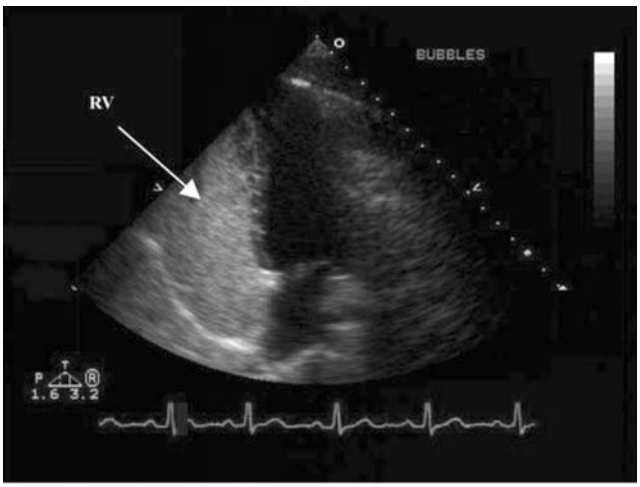
Three-Dimensional Reconstruction

While the 2D exam displays considerable data about spatial relationships between structures and quantification of volume, there is still considerable ambiguity in many circumstances. One way to further enhance the exam is to use 3D reconstruction.

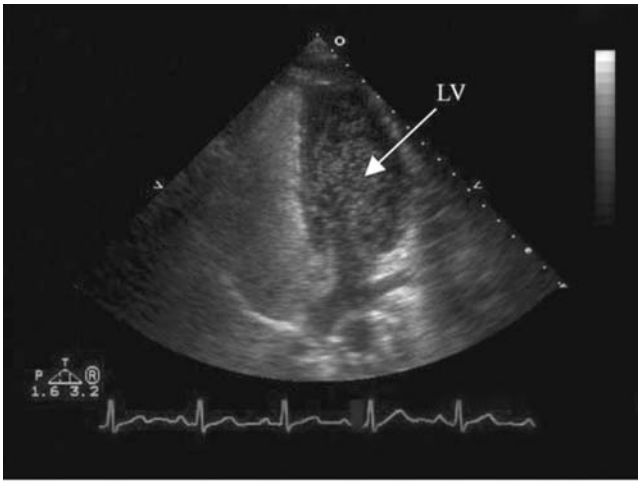
Figure 44. Example of on-line 3D reconstruction. The heart is imaged at the level of the aortic valve where all three leaflets are shown.

Its use has been a significant challenge. All early methods developed computerized routines that characterized the movement of the transthoracic transducer in space. Images were acquired sequentially and then reconstructed first using geometric formulae and later using more flexible algorithms without geometric assumptions. The data, while shown to be useful for both adding new insight into several cardiac diseases and improving quantitation, did not achieve practical acceptance due to the considerable operator time and effort required to obtain just one image (12).

Recently innovations in image processing and transducer design have produced 3D renditions of relatively small sections of the heart in real time. Use remains limited at present but further development is expected to make 3D imaging a practical reality on standard ultrasound systems (Fig. 44).



(a)



(b)

Figure 45. (**a**) Example of injection of agitated saline into systemic veins. The contrast moves into the right atrium (RA) and right ventricle (RV), causing transient full opacification. The left-sided chambers are free of contrast. (**b**) Similar to (**a**). However, some of the contrast bubbles have crossed to the left ventricle (LV), proving a communication exists between the right and left sides of the heart.

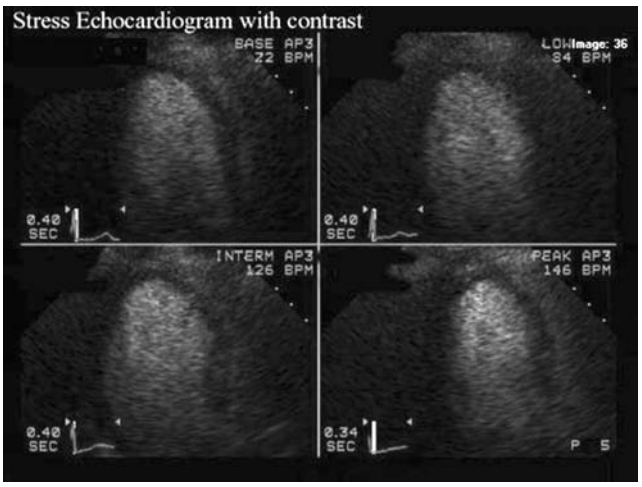


Figure 46. Example of contrast use to improve visualization of the walls of the left ventricle. The contrast agent used in this case is a gas-filled microbubble that fills the left ventricle during a stress test. The entire chamber is white with full contrast enhancement.

Contrast Imaging

Contrast agents are combined with certain 2D echo exams to enhance the amount of diagnostic information available on the exam, or improve the quality of the exam.

There are two types of contrast agents. One type is made from normal saline solution. An assistant generates the contrast for injection during regular 2D imaging. Typically 0.5 mL of air is added to a 10 mL syringe of saline and vigorously hand agitated between 2 syringes for about 15 s. This agitation causes production of several million small bubbles. The bubbles are too large to pass through capillaries, thus when injected into a vein, they are all filtered out by the passage of venous blood through the lungs. Thus when injected into a normal heart, saline contrast passes into the right atrium and right ventricle, produces a transient intense response and disappears, not making it to the left ventricle. This property makes saline injection ideal for detecting an abnormal hole, or shunt passage across the atrial septum or ventricular septum. When a communication of this type is present bubbles cross directly from the right side to the left side of the heart. This is an extremely sensitive method for making this diagnosis (Fig. 45).

A second type of contrast agent, developed and now commercially marketed, is a gas filled microbubble smaller than a red blood cell. This type of agent, made from perfluorocarbons covered by an albumen or lipid shell, when intravenously injected passes through the lungs and opacifies the left side of the heart as well as the right side. The bubbles are gradually destroyed by ultrasound and blood pressure, lasting for several minutes in ideal circumstances. When combined with harmonic imaging these contrast agents markedly improve the quality of the 2D image, particularly for the evaluation of left ventricular wall motion (Fig. 46). The contrast agent markedly enhances the border between blood and the chamber wall. Use of these agents is variable among laboratories, but its use substantially decreases the number of nondiagnostic studies. Still under investigation is whether these same contrast agents can be used for evaluation of blood flow within the heart muscle. This information could be of particular value for patients with coronary artery disease either during acute episodes of ischemia when a new coronary stenosis is suspected or as an enhancement to the stress echocardiogram (13).

BIBLIOGRAPHY

Cited References

1. Edler I, Hertz CH. The use of ultrasonic reflectoscope for the continuous recording of the movements of heart walls. Kungl Fysiografiska sällskapets I lund forhandlingar 1954;24(5):1–19.
2. Zagzebski JA. Physics of Diagnostic Ultrasound. In: Rowland J, editor. Essentials of Ultrasound Physics. St. Louis: Mosby; 1996. p 1–18.
3. Hancock J, Dittrich H, Jewitt DE, Monaghan MJ. Evaluation of myocardial, hepatic and renal perfusion in a variety of clinical conditions using an intravenous ultrasound contrast agent (Optison) and second harmonic imaging. Heart 1999;81:636–641.
4. Spencer KT, Bednarz J, Rafter PG, Korcarz C, Lang RM. Use of harmonic imaging without echocardiographic contrast to improve two-dimensional image quality. Am J Cardiol 1998; 82:794–799.
5. Kisslo J, VonRamm OT, Thurstone FL. Cardiac imaging using a phased array ultrasound system II: clinical technique and application. Circulation 1976;53:262–267.
6. Nishimura RA, Miller FA, Callahan MI, Benassl RC, Seward JB, Tajik AJ. Doppler echocardiography: Theory, instrumentation, technique and application. Mayo Clinic Proc 1985;60:321–343.
7. Omoto R, Kasai C. Physics and instrumentation of Doppler color mapping. Echocardiography 1987;4:467–483.
8. Vandervoort PM, Rivera JM, Mele D, Palacios IF, Dinsmore RE, Weyman AE, Levine RA, Thomas JD. Application of color Doppler flow mapping to calculate effective regurgitant orifice area. Circulation 1993;88:1150–1156.
9. Burns PN. Instrumentation for contrast echocardiography. Echocardiography 2002;19:241–258.
10. Sade LE, Severyn DA, Kanzaki H, Dohi K, Gorcsan J. Second generation tissue Doppler with angle corrected color codes wall displacement for quantitative assessment of regional left ventricular function. Am J Cardiol 2003;92:554–560.
11. Urhelm S, Edvardson T, Torpi H, Angelsen B, Smiseth O. Myocardial strain by Doppler echocardiography: Validation of a new method to quantify regional myocardial function. Circulation 2000;102:1158–1164.
12. Mele D, Levine RA. Quantitation of ventricular size and function: Principles and accuracy of transthoracic rotational scanning. Echocardiography 2000;17:749–755.
13. Porter TR, Cwajg J. Myocardial contrast imaging: A new gold standard for perfusion imaging. Echocardiography 2001; 18:79–87.

Reading List

Anderson B. Echocardiography: The Normal Examination and Echocardiographic Measurements. Brisbane: Fergies; 2000.
Goldberg BB. Ultrasound Contrast Agents. St. Louis: Mosby Year Book; 1997.
Feigenbaum H. Echocardiography. 5th ed. Philadelphia: Lea & Febiger; 1994.

McB Hodgson J, Sheehan HM. Atlas of Intravascular Ultrasound. New York: Raven Press; 1994.
Oh JK, Seward JB, Tajik AJ. The Echo Manual. Philadelphia: Lippincott-Raven; 1999.
Otto CM. Textbook of Clinical Echocardiography. 2nd ed. Philadelphia: W.B. Saunders; 2000.
Weyman AE. Principles and Practice of Echocardiography. 2nd ed. Philadelphia: Lea & Febiger; 1994.
Zagzebski JA. Essentials of Ultrasound Physics. St. Louis, Mosby; 1994.
Hatle L, Angelsen B. Doppler Ultrasound in Cardiology. Philadelphia: Lea and Febiger; 1985.
Marwick TH. Stress Echocardiography: Its role in the diagnosis and evaluation of coronary artery disease. Dordrecht: Kluwer Academic Publishers; 1994.
Freeman WK, Seward JB, Khandheria BK, Tajik AJ. Transesophageal Echocardiography. Boston: Little Brown; 1994.

See also BIOIMPEDANCE IN CARDIOVASCULAR MEDICINE; ULTRASONIC IMAGING.

ECT. See ELECTROCONVULSIVE THERAPY.

EDUCATION, BIOMEDICAL ENGINEERING. See BIOMEDICAL ENGINEERING EDUCATION.

EDUCATION, COMPUTERS IN. See MEDICAL EDUCATION, COMPUTERS IN.

EEG. See ELECTROENCEPHALOGRAPHY.

EGG. See ELECTROGASTROGRAM.

ELECTRICAL TREATMENT OF BONE NONUNION. See BONE UNUNITED FRACTURE, ELECTRICAL TREATMENT OF.

ELECTROANALGESIA, SYSTEMIC

AIME LIMOGE
The René Descartes
University of Paris
Paris, France

TED STANLEY
Salt Lake City, Utah

INTRODUCTION

Electroanalgesia, electroanesthesia, neurostimulation, neuromodulation, and other physical methods of producing analgesia, anesthesia, and/or decreased sensitivity to painful stimuli are old concepts that are beginning to be revitalized in the recent past. For > 40 years, there has been a revival of electrotherapy in the treatment of pain. Analgesia by electrical current is now based on transcutaneous or percutaneous nerve stimulation, deep stimulation, posterior spinal cords stimulation, and transcutaneous cranial electrical stimulation (1–8). One reason for this has been the increased awareness of spinal and supraspinal opioid analgesic mechanisms, including the precise pathways, receptors, and neurotransmitters involved in pain perception, recognition, modulation, and blockade. Another reason is the renewed belief that nonpharmacological manipulation of these receptors and transmitters should be possible with electricity since numerous progress have been made in the development of electric current waveforms that result in significant potentiation of the analgesic and hypnotics action of many intravenous and inhaled anesthetics without producing significant side effects (9–20). Finally, recent successes of transcutaneous electrical nerve stimulation (TENS) in the treatment of pain and transcutaneous cranial electrical stimulation (TCES) as a supplement during anesthesia to obtain postoperative analgesia by potentiating the anesthetic agents used during the intra- and postoperative phases. The popularity of electroacupuncture in a variety of pain and pain related areas have focused the attention of investigators and the public on electricity as a beneficial medical therapy. In this article, some of the most recent developments in nerve and brain stimulatory techniques using electrical stimulation to produce analgesia are addressed.

HISTORY

Alteration of pain perception utilizing forms of electrical stimulation dates back to the Greco-Roman period. Electrostimulation to decrease the pain started with the "electric fish" (torpedo marmorata), as 46 years after Jesus Christ, Scribonius Largus, physician to emperor Claudius, recommended the analgesic shock of the Torpille in the treatment of the pain (21,22). Unfortunately, in those days attempts were crude and success was limited for many reasons none-the-least of which was a poor understanding of the fundamentals of electricity. Interest in electroanalgesia was renewed in the seventeenth century when Von Guericke built the first electrostatic generator to apply locally to relieve pain; however, results were still marginal. At the beginning of the twentieth century, Leduc reawakened interest in the idea of producing sleep and local and general anesthesia with low frequency impulsional electrical current. He used unidirectional rectangular intermittent current of 100 Hz with an ON-time of 1 ms and OFF-time of 9 ms with a moderate amperage (0.5–10 mA) on a variety of animals and on himself to evaluate the effects of electricity on the central nervous system (23,24). Electrodes were placed on the forehead and kidney areas and electrostimulation resulted in apnea, cardiac arrhythmias, cardiac arrest, and convulsions in dogs and a "nightmarelike state" in which the subject was aware of pain. Despite these inauspicious beginnings, studies continued. In 1903, Zimmern and Dimier produced postepilectic coma with transcerebral currents, and in 1907, Jardy reported the first cases of surgery in animals with electroanesthesia. Between 1907 and 1910, Leduc and other performed a number of surgical operations on patients with electricity as an anesthetic supplement (1–5).

In the early decades of the twentieth century, electroanalgesia was always associated with intense side effects including muscle contractures, prolonged coma (cerebral shock), cerebral hemorrhage, hyperthermia, cardiac

arrhythmias, and convulsions. Because of these difficulties, interest waned. In 1944, Frostig and Van Harreveld began experimenting with an alternating current from 50 to 60 mA and a variable voltage bitemporally. The advantage of this more complex method of stimulation was less muscular spasm and contraction (4,5). Unfortunately, these approaches still resulted in transient periods of apnea, cardiac arrhythmias, and standstill as well as fecal and urinary soillage. These problems could be reduced, but not eliminated, by decreasing amperage. Numerous other investigators began using many diverse currents without much success. The most interesting results were obtained in 1951 by Denier (25,26) and in 1952 by Du Cailar (4). Denier began experimenting with high frequency (90 kHz) rectified sinusoidal current with a pulse duration of 3 ms (on time) and a resting time of 13 ms (OFF time), knowing that the effects of modulation at a high frequency current are those of the envelope of its waves. Du Cailar introduced the idea of electropharmaceutical anesthesia by utilizing a premedication of morphin–lobelin in association with a barbituric induction, along with the electrical current. This idea of electropharmaceutical anesthesia that was taken up again in the Soviet Union in 1957 by Ananev et al. using the current of Leduc combined with a direct current (1–3), and in the United States by Hardy et al. using an alternating sinusoidal current of 700 Hz current of Knutson (27–29), and during the same period by Smith using the current of Ananev (1). But with these currents the experimenters were always bothered by side effects (muscle contractions of the face with trismus, of the body with apnea, etc.) that required the use of curare and for all practical purposes made this approach to anesthesia more complicated that conventional anesthesia.

Other investigators began studying mixtures of pharmaceutical agents, including opioids, barbiturates, and later benzodiazepines and butyrophenones in combination with electric currents to reduce and hopefully eliminate these problems, which were often attributed to "the initial shock of the electric current". Others began studying the shape of the current waveform and its frequency. Sances Jr., in United States, used the current of Ananev associated with white noise (5,30) while Shimoji et al. in Japan, used a medium frequency (10 kHz) monophasic or biphasic current with sinusoidal or rectangular waves (31,32). Many were able to produce impressive analgesia and anesthesia in animals, but significant problems (apnea, hypersialorrhea, muscular contractures, convulsions, cardiac arrhythmias) continued to occur in humans. As a result, from the 1950s until the present time, many investigators focused on appropriate electrode placement. It was Djourno who thought that the principal problem to resolve was to find the ideal position for the electrodes to determine the trajectory of the electric current so as to touch precise zones of the brain. This is why he advocated electrovector anesthesia applied with three electrode pairs (vertex-palate, temporal-temporal, fronto-occipital) (4,33). During this time the Soviets Satchov et al. preferred interferential currents of middle frequencies (4000–4200 Hz) associated with barbiturates transmitted by two pairs of crossed electrodes (left temporal–right retromastoid and right temporal–left retromastoid). Others suggested that the problems can be minimized by using mixtures of sedative, hypnotic, and analgesic drugs, plus low amperage electrical currents to produce the ideal effect.

The result of all this activity is that there is still no general agreement on the importance of electrode placement (although frontal and occipital are probably most popular), waveform, wave frequency, current strength, interference currents, or the role of supplemental pharmacotherapy (4,34–38). What was agreed was that it appeared impossible to reliably produce problem-free "complete anesthesia" in humans using any available electrical generators and associated apparatus. Instead, the most successful approaches to electroanesthesia have used waveforms, frequencies, and currents that produce few, if any, side effects (and result in significant analgesia), but must be supplemented with pharmacological therapies to be a "complete anesthetic". While some may scoff at these modest gains, others remain optimistic because using a variety of neurostimulatory approaches, reproducible and quantifiable analgesia was now possible without pharmaceutical supplementation.

Analgesia and Electroneurostimulation

The advancement of the spinal gate control theory of pain by Melzach and Wall (39,40), the discovery of central nervous system opiate receptors, and the popularity and apparent effectiveness of acupuncture in some forms of pain management have given support to the basis that neurostimulatory techniques can produce analgesia via readily understandable neurophysiological changes rather than mysterious semimetaphysical flows of mysterious energy forces (41,42). It is now clear that electrical stimulation of the brain and peripheral nerves can markedly increase the concentration of some endogenous opiates (β-endorphin, δ-sleep producing factor, etc.) in certain areas of the brain and produce various degrees of analgesia. It is proposed that pain relief from electrical stimulation also results from a variety of other mechanisms including alteration in central nervous system concentrations of other neurotransmitters (serotonin, substance P), direct depolarization of peripheral nerves, peripheral nerve fatigue, and more complex nervous interactions (43–46).

Whatever the mechanisms producing analgesia with electrical stimulation, many clinicians are beginning to realize the advantages of these techniques. Neurostimulatory techniques are relatively simple, devices are often portable, their parameters (controls) are easy to understand and manipulate, and application usually requires minimal skills. Moreover, there are few, if any, side effects, addiction is unheard of, if a trial proves unsuccessful little harm is done, the techniques reduce requirements for other analgesics, and usually the stimulation itself is pleasant.

Transcutaneous Electrical Nerve Stimulators

Transcutaneous electrical nerve stimulation, currently called TENS, is the most frequently used device for treatment of acute postoperative and chronic pain of most etiologies. The first portable transcutaneous electrical stimulators were produced in the 1970s with controllable

wave forms and modulable patterns of stimulation. The goal was to produce a compact, lightweight, portable miniaturized current generator to provide stimulation by means of skin contacting electrodes, and able to be used as the patient went about normal daily activities. To that end, as well as safety reasons, the devices were battery powered. A plethora of electrical nerve stimulators can be found on the market. Dimensions are approximately the size of a pack of cigarettes and can be worn by the patient by use of straps or belts. These stimulators, that have one or more adjustable electric parameters that provide no ease of operation, deliver biphasic waves of low frequency of 1–250 Hz with current intensity from 50 to 100 mA. These electrical stimulations result in a tingling or vibrating sensation. Patients are able to adjust the dial settings with respect to frequency and intensity of the stimulus.

The stimulation electrodes must permit uniform current density and have a stimulation surface > 4 cm^2 in order to avoid cutaneous irritation caused by elevated current densities. The material must be hypoallergenic, soft, and flexible to allow maximal reduction of any discomfort while providing for lengthy stimulation in diverse situations. The impedance at the biologic electrode–skin interface can be minimized by the choice of material as well as the use of a conducting gel. Materials used to make the electrodes can be carbon-based elastomeres as well as malleable metals. Most recent developments use adhesive-type ribbons impregnated with silver and are activated by a solvent and provide improved conductibility. For a clinician who is inexperienced in electronics or electroneurophysiology, it is difficult to choose wisely as parameters available for use are created by inventors or producers with absolutely no scientific basis. Analysis of results obtained with the majority of these devices is based on subjectivity of the physician or the patient. The domain is merely empiric. It is a pity that the parameters chosen in the production and use of these devices is by researchers that have not taken advantage of the available scientific works in electrophysiology, notably those of Willer (47,48) on the nociceptive reflex of exercise in humans. A neurostimulator must be selected that will provide proper nerve excitation that is reproducible and durable and that does not cause lesions from burns or electrolysis. Consequently, all those stimulators that deliver direct or polarized current should be used carefully as well as those that deliver a radio frequency (RF) in excess of 800 kHz. One must chose stimulators that deliver a constant biphasic asymmetric current, that is, one that delivers a positive charge that is equal to the negative charge providing an average intensity of zero. To guide the clinician, it must be recalled that current always takes the path of least resistance, and therefore a current of low frequency can only be peripheral the more one increases the frequency. Otherwise, undesirable effects will be produced under electrodes. It is know that a sensation of numbness appears from 70 to 100 Hz and that a motor action appears from 1 to 5 Hz.

Implantable Electrical Nerve Stimulators

Other forms of stimulation consist of implanted neurostimulators, spinal cord stimulation (SCS) (dorsal column stimulators), and deep brain stimulation (DBS). Peripheral nerve neurostimulation implants are also often used for chronic pain but may be employed for acute ulnar, brachial plexus, or sciatic pain in critically ill patients (8).

There are two types of implantable electrical stimulators: Passive-type stimulator with RF made up of as totally implantable element (receptor) and an external element (transmitter) that supplies the subcutaneous receiver through the skin using an RF modulated wave (500 kHz–2 MHz). Active-type totally implantable stimulator, supplied by two mercury batteries (which lasts for 2–4 years) or a lithium battery, which lasts for 5 or 10 years. These devices enable several parameters to be controlled (amplitude peak, wave width, frequency gradient). The variation of these parameters obviously depends on the patient, the region stimulated and the symptom which it is desired to modify.

ACTUAL CLINICAL NEUROSTIMULATORY TECHNIQUES

Certain precautions must be taken and the patient must be well advised as to the technique, the principles of stimulation, and all desired effects. These techniques should not be used on patients wearing a cardiac pacemaker, pregnant women, or in the vicinity of the carotid sinus. The methods demand the utmost in patience, attention to detail, and perseverance. It must be regularly practiced by medical or paramedical personnel.

The most important application of neurostimulatory techniques in clinical use today is in management of acute postoperative and chronic pain, however, since 1980 numerous terms are used in the articles to describe the diverse techniques for electrical stimulation of nervous system. Certain words do not harmonize with reality, such as TransCranial Electrostimulation Treatment (TCET) or Transcranial Electrostimulation (TE). In reality, the microamperage and low frequency used do not enable penetration of the current into the brain, they correspond to a peripheral electrostimulation, which is a bad variant of Transcutaneous Electrical Nerve Stimulation, now being used for certain painful conditions.

Transcutaneous Electrical Methods

Transcutaneous Electrical Nerve Stimulation (TENS). The purpose of this method is to achieve sensitive stimulation, by a transcutaneous pathway, of the tactile proprioceptive fibers of rapid conduction with minimal response of nociceptive fibers of slow conduction and of efferent motor fibers. Numerous studies have documented that TENS in the early postoperative period reduces pain, and thus the need for narcotic analgesics, and improves pulmonary function as measured by functional residual capacity. TENS is also frequently applied in chronic unremitting pain when other approaches are less effective or ineffective. This method is the simplest technique, and appears to be effective by alleviating the appreciation of pain (6,49).

The points of stimulation and the stimulation adjustments must be multiple and carefully determined before concluding that the effect is negative. Different stimulation points are used by the various authors: One can stimulate

either locally by placing the electrodes in the patient at the level of the painful cutaneous area and more particularly on the trigger point that may be at times some distance from the painful zone (50), or along a nerve pathway "upstream" away from the painful zone to cause parasthesia in the painful area, or an acupuncture point corresponding to the points depicted an acupuncture charts (41,42). The stimulation time is usually 20–30 min and repeated at fixed hourly intervals, and discontinued when the pain is relieved. Whatever method is used, one must avoid the production of harmful stimulations or muscle contractions and the stimulation must be conducted with the patient at rest.

In acute injury states where pain is localized, TENS can produce analgesia in up to 80% of patients (51), but this percentage decreases to $\sim$ 20% effectiveness at the end of a year. In order to obtain this result, this stimulation has the sensation of "pins and needles" in the area of the cutaneous stimulation. This phenomenon appears to be in part similar to a placebo effect estimated at 33% regardless of the type of current employed or the location of applied current. As the affected area increases in size, TENS is less likely to be sufficient and is also less effective in chronic pain, especially if the cause of the pain itself is diffuse.

The mechanism by which TENS suppresses pain is probably related to spinal and/or brain modulation of neurotransmitter and/or opiate or other γ-aminobutyric acid (GABA) receptor function. This method works best with peripheral nerve injuries and phantom and stump pains. Transcutaneous nerve stimulators are usually less effective in low back pain or in patients who have had multiple operations. It is often totally unsatisfactory for pain (particularly chronic pain) that does not have a peripheral nerve cause such as pain with a central nervous system etiology or an important psychological component (depression and anxiety) (52–55).

Transcutaneous Acupoint Electrical Stimulation (TAES). Acupuncture, in its traditional form, depends on the insertion of needles into specific acupuncture points in the body as determined by historical charts. Electrical Acupuncture (EA) or TAES employs Low Frequency (LF) stimuli of 5–200 Hz in the needles inserted at the classical acupuncture points. Occasionally, nontraditional acupuncturists use the needles at or near the painful area. Usually these types of treatments produce mild degrees of analgesia. Electrical acupuncture is essentially as benign as TENS and produces its effects by similar mechanisms (42,53,56). Unfortunaly, EA is more expensive toperform than TENS because it necessitates the presence of an acupuncturist clinician. Thus, it is likely that EA will not become as popular as TENS for treatment of most pain problems.

Transcutaneous Cranial Electrical Stimulation (TCES). This method is a special form of electrical stimulation that employs a stimulator that gives a complex current (specific waveforms and high frequency). It was developed by a French group headed by Limoge (35–38,57–59). The TCES method has been used for analgesia during labor pain and before, during, and after surgery, and has recently been shown to be effective to potentiate the analgesic drugs for major surgery, and cancer pain (60). With TCES two electrodes are placed in back of the ear lobe and behind the mastoid bone and one electrode at intersection of the line of the eyebrowns and the sagittal plane. The resulting analgesia is systemic rather than regional (see the section Electrical Anesthesia for a more complete description of the current).

Neurosurgical Methods

Percutaneous Electrical Nerve Stimulation (PENS). This method consists of an electric stimulation by means of a surgically implanted electrode (subcutaneous) coupled by RF induction to an external stimulator nerve. This surgical technique produces long-term positive results of $\sim$ 70% (61,62). It is possible to carryout this procedure quite simply by temporarily implanting needle electrodes at the acupuncture points or auriculotherapy points. This technique produces results similar to those of classic TENS (63,64).

Spinal Cord Stimulation (SCS). This is a neurosurgical method utilized in cases of failure of simple pharmacological or physical treatment where the percutaneous test was positive. As with PENS an RF stimulator is implanted, which this time is connected to electrodes at a level with the posterior spinal cord. The electrodes are actually placed in the epidural space, to provide a percutaneous pathway, under local anesthesia and radiological control. It is often difficult to obtain good electrode position and electrodes can easily become displaced. This technique is reserved for desperate cases as the results are of long term. Approximately 30% are discouraging results (8).

Deep Brain Stimulation (DBS). This method is a complicated and awkward procedure bringing to mind stereotaxis (8). It consists of implanting electrodes at the level of the Ventral Postero-Lateral (VPL) nucleus of the thalamus, which is in relation to afferent posterior cords at the level of PeriAcqueductal Grey Matter (PAGM) or at the level of the PeriVentricular Grey Matter (PVGM), where endorphin and serotonin neurons are found at the motor cortex, which is the start of the pyramidal fascia (10–13). Results obtained are encouraging in cases of consecutive pains at the deafferentation (72%), but of no value in case of pains of nociception. Deep brain stimulation is employed in patients when pain is severe, when other approaches have failed, and when there is a desire to avoid a "drugged existence" and life expectancy is at best a few months. It is often an approach to patients with metastatic cancer. This method is less successful when pain originates from the central nervous system (secondary to stroke, trauma, quadriplegia). The DBS-stimulating probes are usually targeted for the periaqueductal gray matter when pain is deep seated, or for the sensory thalamus or medial lemniscus when is superficial.

Electrical Anesthesia

As mentioned previously, it has never been nor is not now possible to produce, "complete anesthesia" with electricity alone in humans without producing serious side effects. On

the other hand, work by numerous investigators has demonstrated that one or more methods of electropharmaceutical anesthesia (anesthesia consisting of a combination of an electric current with anesthetic agents) is not only possible, but also desirable because of the lack of side effects and reduced requirements for neurodepressants. During past years, progress in chemical anesthesia has been so successful that the objective was not to replace classical anesthesia, but to more precisely confirm studies performed on animals, potentiation of anesthetic drugs by Transcutaneous Cranial Electrical Stimulation (TCES) to obtain postoperative electromedicinal analgesia, to the end that toxicity induced by chemical drugs could be dramatically reduced. The use of TCES is not without considerable supporting data, as from 1972 to 2000, many clinical trials involving TCES had been carried out on patients under electromedicinal anesthesia and provide >20 specific references (7). During those clinical trials, anesthetists noticed that complaints of patients operated under TCES were less numerous in the recovery room than complaints of patients operated with chemical anesthesia. It seems that a state of indifference and reduction of painful sensation persisted in TCES-treated patients. These observations were scientifically confirmed in a study (59) in which 100 patients operated under electroanesthesia (EA) was compared to another 100 patients submitted to narco-neurolept-analgesia, a classical anesthesia (CA): the head nurses were ordered to administered 15 mg (i.m.) of pentazocine in case of patient complaints. It is worth noting that the first 16 postoperative hours, the average intake of pentazocine for the patients of the EA group was 8.1 mg/patient, whereas it was 29.7 mg/patient (3.67 time higher) for the patients of the CA group. This difference between groups is highly statistically significant ($p < 0.001$) (Fig. 1).

This residual and prolonged analgesia is surely one of the most important advantages of TCES, but few clinicians benefit from its advantages at the present time. The most likely reason that few clinicians benefit from TCES is that it is not yet approved for use in the United States, Canada, and many countries in Europe by the respective regulatory agencies. Recent research carried on in numerous laboratories has increased our knowledge of the neurobiological effects of these currents, and allowed the establishment of serious protocols dedicated to new clinical applications (7).

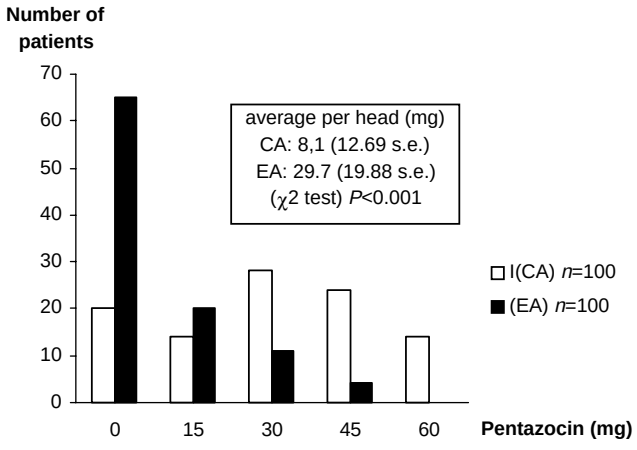


Figure 1. Comparison of two groups receiving pentazocin during the first 16 h after surgery.

Nature of the Limoge's Current. Limoge et al. demonstrated that complex currents of their design are capable of producing profound analgesia without provoking initial shock, pain, or unpleasant sensations, burns, other cutaneous damage, muscular contractures, cerebral damage or convulsions, and respiratory or circulatory depression (58). The Limoge current consists of high frequency (HF) biphasic asymmetrical wave trains composed of modulated high frequency (166 kHz) pulse trains, regularly interrupted with a repetition cycle of 100 Hz (7,57). These wave trains are composed of successive impulsional waves of a particular shape: one positive impulse of high intensity and short duration (2 μs), followed by a negative impulse of weak intensity and long duration (4 μs) adjusted in such a way that the positive surface is equal to the negative surface. The average intensity of this current equals 0 mA. The use of such a negative phase makes it possible to eliminate all risk of burns. The "on-time" of the low frequency (LF) wave trains is 4 ms, followed by a 6 ms "OFF-time" (Fig. 2).

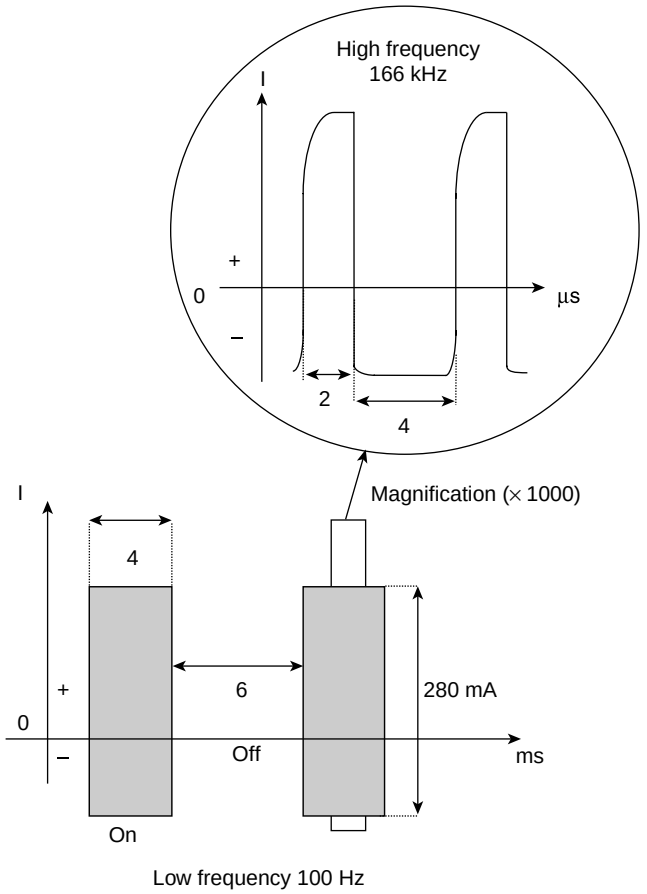


Figure 2. The Limoge waveform pattern: a modulated HF (166 kHz) pulse trains (top) regularly interrupted with a repetition cycle of 100 Hz. Concerning the high frequency, note the exponential ascent and acute fall of the waveform, and also note the area of the positive deflection is equal to that of negative deflection.

Figure 3. Application of the device on a patient during post operative period. See the placement of the frontal electrode.

This type of current was gradually developed over ~ 20 years through numerous human clinical studies. The shape and cyclic ration of the HF waves are felt to be of utmost importance in the production of analgesia. Various shapes of waves have been tested (triangular, rectangular, exponential). Clinical impressions suggest that the most profound analgesia occurs with HF waveforms having an exponential ascent and acute fall. The most effective cyclic ratios are 2:5 with LF waves and 1:3 with HF waves with peak-to-peak intensity between 250 and 300 mA and peak-to-peak voltage between 30 and 40 V.

Electrodes. Three electrodes are used. One frontal electrode is placed between the eyebrows and two posterior electrodes are placed behind the mastoid process on each side of the occiput (Fig. 3). It is hoped that the intracerebral electric field thus obtained spreads on each side of the median line and it thus successful in stimulating opioid receptors surrounding the third and fourth ventricles and the paraventricular areas of the brain. In addition, some of the electric current spreads over the scalp, thus provoking peripheral electrostimulation (Fig. 4). The use of HF biphasic current permits employment of self-sticking electrodes made of silver (active diameter 30 mm), without risk of burns and without unpleasant sensations under electrodes.

Transcutaneous Cranial Electrical Stimulators Using Limoge Currents. Until now three types of devices only give Limoge currents: two American devices called Foster Biotechnology Neurostimulation Device (FBND) and Electro-Analgesia Stimulation Equipment (EASE) and one French device called Anesthelec (Fig. 5). The electrical stimulator must abide by general safety rules. The use of electrosurgical units during TCES requires excellent electrical isolation of the generator to avoid any risk of return of the current or skin burns under electrodes, a fault in the electrosurgical unit. These portable devices of type LF with isolated output are composed of one HF oscillator, one oscillator with LF relaxation with internal power supply generating HF (166 kHz), and LF (100 Hz) currents for

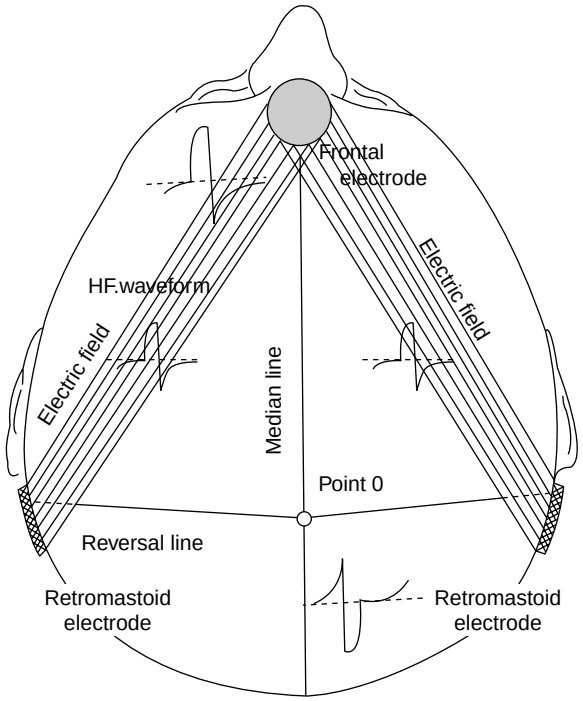


Figure 4. Location of the electrodes and shape of wave on the scalp. The center of the frontal electrode is situated at the intersection of the line of the eyebrowns and the sagittal plane. The center of the two retromastoid electrodes is localized in the retromastoid fossa. On the scalp the amplitude of HF waves diminish in measurement as the point O (occipital line) is approached, and behind that point there is an inversion of the wave form. The lines joining the frontal electrode to retromastoid electrodes represent the projected distribution of Limoge currents through the brain with action at the level of the periacqueductal gray matter and the limbic system.

therapeutic use and one delay circuit, to stop output when its level is too high. The battery pack must be protected against short circuit as well as polarity inversion and detachable as it is rechargeable. The patient cable must

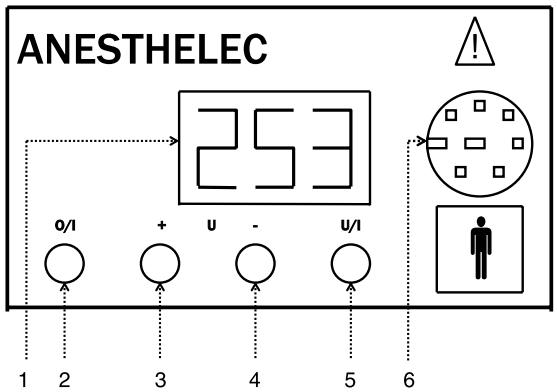


Figure 5. Front view of the Anesthelec generator box $1 \rightarrow 3$ digits displayer: Current intensity (from 000 to 300 mA); $2 \rightarrow$ Pushing button On/Off; $3 \rightarrow$ Pushing button for increment of the output current intensity; $4 \rightarrow$ Pushing button for decrement of the output current intensity; $5 \rightarrow$ Pushing button to select display (intensity of voltage); $6 \rightarrow$ Connector for the three electrodes.

be an interlocking type preventing accidental disconnection and the functioning of the device must be simplified with detectors to measure the intensity of the current and the voltage applied to the patient to confirm proper contact between skin and electrodes. Concerning electromagnetic compatibility, the device must be autonomous with no possible direct or indirect link mains supplies. Mini box and manipulation components must be made in isolated material and the patient cables must be shrouded and the applied elements must be protected against overvoltage.

Clinical Usage of TCES

The TCES method being used with increasing frequency in France, and many other european countries, in Russia, in Mexico, and in Venezuela. It is not yet approved for use in the United States, but is being evaluated both in patients and in volunteers. Numerous studies have demonstrated that TCES is particularly effective in urologic, thoracic, and gastrointestinal surgery, but is not limited to these types of operative procedures. Patients receiving TCES require less nitrous oxide (N_2O) (30–40% less) to prevent movement in response to a pain stimulus. This method potentiates both the amnestic and analgesic effects of N_2O and prolongs residual postanesthetic analgesia at sites of trauma. The mechanism of analgesia resulting from TCES during administration of N_2O is unknown. Volunteers getting TCES without N_2O for 1 h are not sleepy or amnesic, but do report a warm and tingling sensation all over their body, and are objectively analgesic to many forms of painful stimulation (14–16). Similar results have been obtained with some TENS units in patients with chronic pain and after operation in patients with acute postoperative pain (54,55). As mentioned previously, some have suggested that receptor sites situated in the central gray area of the brain, the spinal cord, and other areas in the central nervous system regulate the effects of painful stimulation, analgesia, and the perception of somatic pain. Electrical stimulation of these receptor sites has been shown to result in relief from pain and can be antagonized by narcotic antagonists. Furthermore, the analgesic actions of TENS can be reversed with antagonists like naloxone (9,10). This suggests that TENS and TCES may be producing analgesia by stimulating increased production and/or release of the body's endogenous analgesics, the endorphins, enkephalins, serotonin and/or other neurotransmitters and neuromodulators (5,11–13,43–46,63–65).

To separate facts from empiricism and anecdotal information for several years, teams of researchers and clinicians attempted to show in animals and in humans what are the neurobiological mechanisms brought into pay by the TCES with currents of Limoge. For that reason, a study was conducted in France on rats on TCES potentiation of halothane-induced anesthesia and the role of endogenous opioid peptides was addressed (19). Carried out in double blind for 10 h prior to tracheotomy and the inhalation of halothane, the TCES provoked in the stimulated rats (TCES group, $n = 10$), a significant decrease ($p < 0.001$) in the Minimum Alveolar Concentration of Halothane (MACH) in comparison with the nonstimulated rats (control group, $n = 10$). This effect was completely inhibited by a subcutaneous injection of 2 mg/kg of naloxone (antagonist of morphine), which restored the MACH to its initial value in the TCES group without affecting the control group (Fig. 6). Moreover, TCES potentiation of halothane-induced anesthesia was dramatically increased by inhibition of enkephalin degradation. Thus the decrease of the MACH is associated with the potentiation the analgesic action of enkephalins released in the cellular space by TCES. These results demonstrate the direct involvement of endogenous opioid peptides on therapeutic effects of TCES.

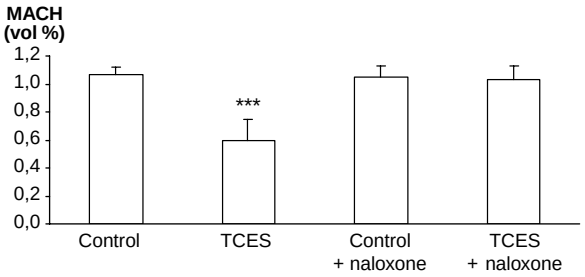


Figure 6. Effects of TCES on halothane requirements in rats. *** indicate significant difference between TCES and CONTROL groups (ANOVA, $p < 0.001$).

In addition, a double-blind study carried out during labor and delivery on parturients to provide evidence of a mode of action of TCES on maternal plasma secretion of β-endorphins (66). To evaluate the rate of β-endorphins, blood samples were drawn from two groups of voluntary women in parturition (a TCES group, $n = 23$, and a control group, $n = 17$) at four precise stages: at the moment the electric generator was attached, after 1 h of the current application, at the time of complete dilatation, and finally after the delivery. The dosages were achieved by the radioimmuno enzymatic method. The plasmatic rate of β-endorphins was identical in the beginning for the two groups as those described in the literature, but this rate was progressively augmented in a significant fashion during the course of the labor from the first hour ($p < 0.05$) for the TCES group (Table 1).

It is more interesting to know the rate of endorphins produced in the cerebral structures known for their abundance of opiate receptors, more so than in the plasma. The exploration of the effects of TCES on brain opioid peptides was conducted at the Vishnevski Institute in Moscow by dosing endorphins in the cerebral spinal fluid (CSF) before cardiac surgery and after 30 min of TCES. The dosage showed that TCES augmented significantly ($p < 0.01$) the rate of β-endorphins in the CSF when compared to the control group and the effects of TCES reversed by naloxone (49,67,68).

These studies can partially explain the mode of action of TCES with currents of Limoge in the brain and permit not only rectification of protocols for clinical trials already carried out, but also provide better indication for utilization of TCES.

For all clinical applications, it must be keep in mind that the currents of Limoge provoke endogenous neurosecretions (7), which are not immediate, they require a certain

Table 1. Evaluation of β-Endorphin During Labor and Delivery on Parturients[a]

			β-Endorphin Plasmatic Rate, $pg \cdot mL^{-1}$			
	Medication	Labor Time	Installation	After 1 h	Dilatation	Delivery
Control ($n = 17$)	Peridural: 1 patient Morphine: 11 patients None: 5 patients	$<$ 1 h 30 min: 4 patients $>$ 1 h 30 min: 13 patients	123 (±12)	127 (±10)	124	160
TCES ($n = 23$)	Peridural: 2 patient Morphine: 3 patients None: 18 patients	$<$ 1 h 30 min: 11 patients $>$ 1 h 30 min: 12 patients	133 (±11) N.S.[b]	167 (±12)[c]	186	182

[a]Results of β-endorphin rates are expressed as mean ± s.e.m. (when available).
[b]N.S. indicates no difference between β-endorphin rates of control and TCES groups when measured at the installation of labor.
[c]Indicates significant difference between β-endorphin rates of control and TCES groups (t-test, $p < 0.05$) when measured 1 h after the installation of labor.

amount of time for their induction, and then are maintained all along the stimulation application. In consequence, the utilization of this technique in classical anesthesia is not the best indication except for major interventions of long duration. One must also remember that during $>10{,}000$ major surgical interventions carried out under classical anesthesia combined with TCES it has been proven that the Limoge currents has a potentiation effect on opioid and non-opioid analgesics, morphinomimetics, psychotropes, and psycholeptics, (14–20) and this potentiation allows a decrease in drug doses, and therefore a decrease in toxicity. But one must admit objectively that, during past years, progress in chemical anesthesia has been so successful that the TCES will not replace classical anesthesia. The potentiation of drugs nevertheless by TCES can open new perspectives in the treatment of pain whether postoperative or chronic. To be precise the potentiation of opioid analgesia by TCES under specific conditions, was demonstrated by Stinus et al. (17). The authors showed that potentiation was a function of (a) the intensity of the stimulation, (b) the opioid dose administered, (c) the duration of TCES applied preceding opioid administration, and (d) the position and the polarity of the electrodes. This experimental approach was of prime importance as it allowed determination of the most efficient parameters, studied the therapeutic effects of TCES in humans, and increased our knowledge of the effects of TCES on neurobiological substrates.

Taking account of animal experimentation and clinical trials, one must know that to be successful in clinical applications, a correct basal protocol for TCES use should be followed. The main parameters are, the correct placement of the electrodes, starting electrostimulation no $<$ 2 h prior to the introduction of drugs and continuation of TCES delivery during the pharmacokinetic action of drugs.

Abolition of Postoperative Pain (Fig. 3). Patients operated under TCES associated with pharmaceutical anesthesia complain strikingly less often about pain than those operated with a classical anesthesia. The TCES method induces a postoperative analgesia for an average of 16 h. A double-blind study has been made during per and postoperative period on 39 patients (TCES group $n = 20$ and control group $n = 19$) undergoing an abdominal surgery (20). Upon arrival in the recovery room, patients were given a computerized, patient-controlled analgesia (PCA) device to deliver IV buprenorphine (50 μg boluses, 30 min lock-out) during the first four postoperative hours. The recorded variables included postoperative requirements, pain scores with pain visual analogue scale (VAS) (from 0 = no pain to 10 = worst), sedation, (from 0 = not arousable to 4 = awake) awake) and were collected hourly from the first to the sixth postoperative hour by a blinded investigator. There was a highly significant reduction of cumulative buprenorphine requirements in the TCES group compared with the control group (2.36 ± 0.19 vs. 3.43 ± 0.29 $\mu g \cdot kg^{-1}\,h^{-1}$; $p < 0.01$) (Table 2). At each postoperative hour, patients required less buprenorphine in the TCES group. These results indicate that TCES reduces narcotic requirements for postoperative analgesia. TCES may have potential to facilitate early postoperative analgesia in patients undergoing major surgery.Therefore this technique allows a maximal restriction of pharmaceutical contribution.

Obstetric Electroanalgesia (66,69). In order to test the analgesic efficacy of TCES with Limoge currents during

Table 2. Buprenorphine Consumption[a]

Postoperative hours (H)	TCES	Control
H1	1.35 ± 0.15	1.57 ± 0.13
H2	0.90 ± 0.16	1.21 ± 0.18
H3	0.60 ± 0.15	1.10 ± 0.16[b]
H4	0.60 ± 0.18	1.00 ± 0.15[b]
Total dose ($\mu g \cdot kg^{-1} \cdot h^{-1}$)	*2.36 ± 0.19*	*3.43 ± 0.29*[c]

[a]Data are expressed as mean ± SEM.
[b]$p < 0.05$.
[c]$p < 0.01$.

labor and delivery, a double-blind study was performed with "anesthelec" on 20 cases for whom analgesia was necessary (TCES group I, current "on", $n = 10$, and control group II, current "off ", $n = 10$). Labor and delivery were carried out by a medical team different from those using the anesthelec. The results showed that TCES, with or without nitrous oxide inhalation, decreases by 80% the number of epidural analgesia or general anesthesia that would otherwise have been unavoidable. To define the effects of TCES, maternal and fetal parameters of 50 deliveries carried out under TCES were compared with 50 deliveries carried out under epidural analgesia (70).

TCES was used only if analgesia was required. These clinical trials were a retrospective comparison between two similar nonpaired series. Despite the fact that analgesia obtained with TCES was less powerful than with epidural analgesia, this method showed many advantages: total safety for the child and the mother, easy utilization, shorter labor time, decreased number of instrumental extractions and potentially reduced costs. Good acceptance and satisfaction for the mother should stimulate a rapid evolution and acceptance of this new method.

The TCES method should be applied following the first contractions. Analgesia is established after 40 min of stimulation. A diminution of pain is achieved that is comparable to that obtained after an injection (IV) of morphine (but it is less profound than with epidural analgesia), a decrease in vigilance with euphoria is obtained without inducing sleep, but allowing compensatory rest between contractions. The pupils are enlarged. Stimulation is applied throughout the birthing procedure and residual analgesia persists for several hours following delivery. Results are best if the expectant mother participates in a preparatory course for the birthing experience or if she uses musicotherapy in conjunction with TCES. If analgesia is insufficient it is possible to have patients breath nitrous oxide and oxygen (50:50) or to administer an epidural analgesia for the remainder of procedure. Thus obstetrical analgesia utilizing the currents of Limoge allows a reduction of labor time in all primapares ($p < 0.001$) and is without risk to the mother or child. Mothers in labor appreciate this simple, nonmedicinal, nonpainful technique that allows them to actively participate in the delivery.

Electropharmaceutical Anesthesia in Long Duration Microsurgery. For major operations and those of long duration the results are most encouraging as TCES permits a reduction of anxiolytics and neuroleptics by 45% and reduction of morphinomimetics by 90% and demonstrates the possibilities of drug potentiation to prolong analgesia while at the same time providing a less depressive general anesthetic (7,58,59,68). Early results have improved thanks to animal research and revision of protocols more particularly (17–19). In 1972, it was not know to begin electrostimulation three hours prior to medicinal induction (4).

Potentiation of Morphine Analgesia for Patients with Chronic Pain and Associate Problems (71). For all neurophysiological applications, a basic protocol must be followed. This protocol is as follows: If the patient is being treated pharmacologically, for the first time, never stop the chemical medication but diminish the dosage each day until a threshold dose is obtained according to the particular pathology and the patient. Begin TCES at least 1 h before medication whether it be on awakening in the morning or 2 h prior to going to bed. (There is no contraindication in maintenance of stimulation all-night long.)

If the patient is not being treated chemically, the effect of the current is best if there is a "starter dosage" of medicine. It is therefore recommended that a weak medicinal dose be prescribed according to the pathology and begin the TCES 1 h before the patient takes the dose, and continue stimulation during the time of pharmacocinetic action of the medicine.

This protocol will permit treatment of cancer patients at home whenever possible under medical supervision: This is a less traumatizing course of action than having the patients come into hospital every day. In the beginning, one must maintain the standard pain medication therapy and the patient should be connected to the Limoge Current generator for 12 h (during the night, if possible); the potentiometer is turned clockwise to a reading of 35 V and 250–300 mA, peak to peak. After this first treatment phase, the patient can use the machine for 3 h whenever they feels the need. The analgesic effect of TCES may not appear until the third day of treatment. Then TCES is initiated upon awakening. After 1 h, standard pain medication is given and TCES therapy is continued for another hour. Three hours before bedtime, TCES is again administered for 2 h, then standard pain medication is given and TCES therapy continued for another hour. The patient should enjoy restful sleep. After 8 days, the standard pain medication therapeutic dose should be decreased gradually, but not totally terminated. After this status has been achieved, patients may use the machine whenever they feel the need, for 3 h preferably with the reduced dose of the standard pain medication. The minimal therapeutic dose of the pain medication, however, may have to be adjusted upward somewhat due to individual differences in some patients.

CONCLUSION

All numerous and previous clinical trials have demonstrated that TCES reduces narcotic (fentanyl) requirements in patients undergoing urologic operations with pure neuroleptanesthesia (droperidol, diazepam, fentanyl, and air-oxygen) (20,36–38). Use of TCES in a randomized double-blind trial of these patients resulted in a 40% decrease in fentanyl requirements for the entire operation. Unfortunenately, while available TCES units (using currents of 250–300 mA peak to peak, with an average intensity of zero) provide analgesia and amnesia, they do not produce complete anesthesia. Whether any form of TCES or the use of very high frequency (VHF) will provide more analgesia and amnesia, that is, amounts sufficient to result in complete anesthesia without need for pharmaceutical supplementation, without problems has yet to be carefully evaluated but obviously needs to be studied. Considerable research must continue in this area.

Theoretically, lower doses of narcotics or lower concentrations of inhalation anesthetics should result in fewer

alterations in major organ system function during anesthesia. This could mean that anesthesia with TCES produces less physiological insult than more standard anesthetic techniques and results in a shorter postoperative recovery period. It has been observed that TCES plus N_2O results in analgesia that persists after stimulation is terminated and N_2O is exhaled (7,14,15,20,58–60). This suggests that intraoperative use of TCES might reduce postanesthetic analgesic requirements, and that future clinical trials must be initiated to confirm this suggestion.

The 30.000 plus major interventions realized under TCES in France and in Russia since 1972 and the >5000 drug withdrawals undertaken in opioid addicted patients at the Medical Center of the University of Bordeaux since 1979 without even the most minor incident permits us to conclude that the currents of LIMOGE are absolutely innocuous and cause no side effects. This simple technique reduced the use of sedative medicaments such as psychotropes or psycholeptics that often lead to "legal" addiction. The TCES is atoxic, reproductible, causes no personality change and is without habituation. Briefly, this technique fits perfectly into the domaine of all aspects of classical and alternative medicine as well as human ecology.

BIBLIOGRAPHY

Cited References

1. Smith RH. Electrical Anesthesia. Springfield (IL): CC Thomas Publ.; 1963.
2. Smith RH, Tatsuno J, Zouhar RL. Electroanesthesia: a review-1966. Anesth Analg 1967;40:109–125.
3. Smith RH. Electroanesthesia Review article. Anesthesiology 1971;34:61–72.
4. Limoge A. An Introduction to Electroanesthesia. Baltimore: University Park Press; 1975. p 1–121.
5. Sances Jr A, Larson SJ. Electroanesthesia. New York: Academic Press; 1975. p 1–367.
6. Shealy CN, Maurer D. Transcutaneous nerve stimulation for control pain. Surg Neurol 1974;2:45–47.
7. Limoge A, Robert C, Stanley TH. Transcutaneous cranial electrical stimulation (TCES): a review 1998. Neurosci Biobehav Rev 1999;23:529–538.
8. White PF, Li S, Chiu JW. Electroanalgesia: Its role in acute and chronic pain management. Anesth Analg 2001;92:505–513.
9. Adams JE. Naloxone reversal of analgesia produced by brain stimulation in the human. Pain 1976;2:161–166.
10. Hosofrichi Y, Adams JE, Linchitz R. Pain relief by electrical stimulation of the central gray matter and its reversal by naloxone. Science 1977;197:183–186.
11. Snyder SH, Goodman RR. Multiple neurotransmitter receptors. Neurochemistry 1980;35:5–15.
12. Snyder SH. Brain peptides as neurotransmitters. Science 1980;209:976–983.
13. Pasternack GW. Opiate enkephalin and endorphin analgesia: Relations to a single subpopulation of opiate receptors. Neurology 1981;31:1311–1315.
14. Stanley TH, et al. Transcutaneous cranial electrical stimulation increases the potency of nitrous oxide in humans. Anesthesiology 1982;57:293–297.
15. Stanley TH, et al. Transcutaneous cranial electrical stimulation decreases narcotic requirements during neuroleptanesthesia and operation in man. Anesthol Analg 1982;61: 863–866.
16. Bourke DL, et al. TENS reduces halothane requirements during hand surgery. Anesthesiology 1982;61:769–772.
17. Stinus L, et al. Transcranial electrical stimulation with high frequency intermittent current (Limoge's) potentiates opiate-induced analgesia: blind studies. Pain 1990;42:351–363.
18. Auriacombe M, et al. Transcutaneous electrical stimulation with Limoge current potentiates morphine analgesia and attenuates opiate abstinence syndrome. Biol Psychiat 1990;28:650–656.
19. Mantz J, et al. Transcutaneous cranial electrical stimulation with Limoge's currents decreases halothane requirements in rats: evidence for involvement of endogenous opioids. Anesthesiology 1992;76:253–260.
20. Mignon A, et al. Transcutaneous cranial electrical stimulation (Limoge's currents) decreases bupremorphine analgesic requirements after abdominal surgery. Anesth Analg 1996;83:771–775.
21. Scribonius L. Compositiones medicae. Padua: Frambottus; 1655. Chapts. 11 and 162.
22. Kane K, Taub A. A history of local electrical anesthesia. Pain 1975;1:125–138.
23. Leduc S. Production du sommeil et de l'anesthésie générale et locale par les courants électriques. C R Acad Sci Paris 1902;135:199–200.
24. Leduc S. L'électrisation cérébrale. Arch Electr Med 1903; 11:403–410.
25. Denier A. Electro-anesthésie. Anesth Analg Réan 1951;8(1): 47–48.
26. Denier A. Anesthésie électrique. EMC 36550 A 10 1958;4:1–8.
27. Knutson RC. Experiments in electronarcosis. A preliminary study, Anesthesiology 1954;15:551–558.
28. Knutson RC, et al. The use of electric current as an anesthetic agent. Anesthesiology 1956;17:815–825.
29. Knutson RC, Tichy FY, Reitman J. The use of electrical current as an anesthetic agent. Anesthesiology 1966;17: 815–825.
30. Cara M, Cara-Beurton M, Debras C, Limoge A, Sances Jr A, Reigel DH. Essai d'anesthésie électrique chez l'Homme. Ann Anesth Franç 1972;13:521–528.
31. Shimoji K, et al. Clinical electroanesthesia with several methods of current application. Anesth Analg 1971;50:409–416.
32. Shimoji K, Higashi H, Kano T. Clinical application of electroanesthesia. In: Limoge A, Cara M, Debras Ch, editors. Electrotherapeutic Sleep and Electroanesthesia. Volume IV, Paris: Masson; 1978. p 96–102.
33. Djourno A, Kayser AD. Anesthésie et sommeil électriques. Paris: P.U.F.; 1968.
34. Sachkov VI, Liventsev NM, Kuzin MI, Zhukovsky VD. Experiences with interference currents in clinical surgery. In: Wageneder FM, Schuy S, editors. Electrotherapeutic Sleep and Electroanesthesia. Excerpta Medica Foundation; 1967. p 321–326.
35. Limoge A. The use of rectified high frequency current in electrical anaesthesia. In: Wageneder FM, Schuy St, editors. Electrotherapeutic sleep and electroanaesthesia. Volume I, Amsterdam: Excerpta Medica Foundation; 1967. p 231–236.
36. Cara M, Debras Ch, Dufour B, Limoge A. Essais d'anesthésie électromédicamenteuse en chirurgie urologique majeure. Bull Acad Méd 1972;156:352–359.
37. Debras C, Coeytaux R, Limoge A, Cara M. Electromedicamentous anesthesia in Man. Preliminary results Rev I E S A 1974; 18–19, 57–68.
38. Limoge A, Cara M, Debras C. Electrotherapeutic Sleep and Electroanesthesia. Paris: Masson; 1978.

39. Melzack R, Wall PD. Pain mechanism: a new theory. Science 1965;150:971–79.
40. Wall PD. The gate control theory of pain mechanisms. A re-examination and re-statement. Brain 1978;101:1–18.
41. Sjölund B, Ericson R. Electropuncture and endogenous morphines. Lancet 1975;2:1085.
42. Fox EJ, Melzach. Transcutaneous nerve stimulation and acupuncture. Comparison of treatment for low back pain. Pain 1976;2:141–149.
43. Akil H, et al. Encephaline-like material elevated in ventricular cerebrospinal fluid of pain patient after analgesic focal stimulation. Science 1978;201:463.
44. Henry JL. Substance P and pain: An updating. Trends Neurosc 1980;3:95–97.
45. Le Bars D. Serotonin and pain. In: Osbone NN, Hamon M, editors. Neuronal Serotonin. New York: John Wiley & Sons Ltd.; 1988. Chapt. 7. p 171–229.
46. Bailey PL, et al. Transcutaneous cranial electrical stimulation, experimental pain and plasma β-endorphin in man. Pain 1984;2:S66.
47. Willer JC, Boureau F, Albe-Fessard D. Role of large diameter cutaneous afferents in transmission of nociceptive messages: electrical study in man. Brain Res 1978;132: 358–364.
48. Willer JC, Boureau F, Albe-Fessard D. Human nociceptive reactors; effects of spacial sommation of afferent input from relatively large diameter fibers. Brain Res 1980;201: 465–70.
49. Pederson M, McDonald S, Long DM. An investigation determining the efficacy of TENS and the use of analgesia during labor in groups of women. Pain 1984;2:S69.
50. Melzack R, Stillwell D, Fox E. Trigger points and acupuncture points for pain: Correlations and implications. Pain 1977;3:23–28.
51. Hanai F. Effect of electrical stimulation of peripheral nerves on neuropathic pain. Spine 2000;25:1886–1892.
52. Campbell JN, Long DM. Peripheral nerve stimulation in the treatment of intractable pain. J Neurosurg 1976;45:692–699.
53. Woolf CJ. Transcutaneous electrical nerve stimulation and the reaction to experimental pain in human subjects. Pain 1979;7:115–127.
54. Kim WS. Clinical study of the management of postoperative pain with transcutaneous electrical nerve stimulation. Pain 1984;2:S73.
55. Park SP, et al. Transcutaneous electrical nerve stimulation. (TENS) for postoperative pain control. Pain 1984;2: S68.
56. Wilson OB, et al. The influence of electrical variables on analgesia produced by low current transcranial electrostimulation of rats. Anesth Analg 1989;68:673–681.
57. Limoge A, Boisgontier MT. Characteristics of electric currents used in human anesthesiology. NATO-ASI Series. In: Rybak B, editor. Advanced Technobiology. Germantown (MD): Sijthoff & Noordhoff; 1979. p 437–446.
58. Debras C, et al. Use of Limoge's current in human anesthesiology. NATO-ASI Series. In: Rybak B, editor. Advanced technobiology. Germantown (MD): Sijthoff & Noordhoff; 1979. p 447–465.
59. Limoge A, et al. Electrical anesthesia. In: Spiedijk J, Feldman SA, Mattie H, Stanley TH, editors. Developments in Drugs Used in Anesthesia. The Boerhave Series. Leiden: University Press; 1981. p 121–134.
60. Limoge A, Dixmerias-Iskandar F. A personal experience using Limoge's current during a major surgery. Anesth Analg 2004;99:309.
61. Shealy CN, Mortimer JT, Reswick JB. Electrical inhibition of pain by stimulation of the dorsal columns. Anesth Analg 1967;46:489–91.
62. Burton CV. Safety and clinical efficacy of implanted neuroaugmentive spinal devices for the relief of pain. Appl Neurophysiol 1977;40:175–183.
63. Prieur G, et al. Approche mathématique de l'action biologique des courants de Limoge. J Biophys Biomec 1985;9(2):67–74.
64. Malin DH, et al. Auricular microelectrostimulation: naloxone reversible attenuation of opiate abstinence syndrome. Biol Psychiat 1988;24:886–890.
65. Malin DH, et al. Augmented analgesic effects of enkephalinase inhibitors combined with transcranial electrostimulation. Life Sci 1989;44(19):1371–1376.
66. Stimmesse B, et al. β-endorphines plasmatiques et analgésie électrique durant l'accouchement. Cahiers d'Anesth 1986;34: 641–642.
67. Schloznikov BM, Kuzin MI, Avroustsky M. Influence de l'E.S.C.T. sur le contenu de β-endorphines dans le liquide céphalo-rachidien et dans le plasma sanguin. Nouvell Biol Expér 1984;5:515–516.
68. Kuzin MI, Limoge A. Electrischer strom und schmerzausschaltung. Berlin: Wissenschaft und Menschheit; 1985. p 50–57.
69. Kucera H, et al. The effects of electroanalgesia in obstetrics. In: Limoge A, Cara M, Debras Ch, editors. Electrotherapeutic Sleep and Electroanesthesia. Volume IV, Paris: Masson; 1978. p 73–77.
70. Champagne C, et al. Electrostimulation cérébrale transcutanée par les courants de Limoge au cours de l'accouchement. Ann Fr Anesth Réanim 1984;3:405–413.
71. Lakdja F. (personal communication).

See also ELECTROPHYSIOLOGY; SPINAL CORD STIMULATION; TRANSCUTANEOUS ELECTRICAL NERVE STIMULATION (TENS).

ELECTROCARDIOGRAPHY, COMPUTERS IN

HOMER NAZERAN
The University of Texas
El Paso, Texas

INTRODUCTION

Digital computers have the ability to store tremendous amount of data and retrieve them with amazing speed for further processing and display. These attributes of computers make them extremely useful in a modern clinic or hospital environment. Computers play a central role in medical diagnosis and treatment as well as management of all processes and information in the hospital including patient data. Computers greatly facilitate the recording and retrieval of patient data in a simple way. All areas of hospital including patient admittance and discharge, the wards, all specialty areas, clinical and research laboratories are now interconnected through computer intranet and even nationally or globally connected through computer internet networks. This arrangement provides for better coordination of patient management throughout various hospital departments, and reduces patient waiting

times. Obviously, computers also play a very important role in all aspects of administration and management of the hospital like other sophisticated institutions. Therefore, they greatly facilitate the overall planning and operation of the hospital resulting in improved healthcare services.

As the electrocardiographic (ECG) signal is one of the most, if not the most, measured and monitored vital signs, computers have had a tremendous impact in electrocardiography. One of the most well known areas of application of computers in medical diagnosis is their use in recording, monitoring, analysis and interpretation of ECG signals. Computers reduce interpretation time and ensure improved reliability and consistency of interpretation. Computers assist cardiologists by providing cost-effective and efficient means in ECG interpretation and relieve them from the tedious task of reviewing large numbers of ECG recordings. Computers also increase the diagnostic potential of ECG signals. Of course, cardiologists consider patient history, the details of the morphology of the ECG signals and other pertinent patient data backed by their clinical knowledge and experience to make a complete diagnosis. It should be clearly stated and emphasized that computers do in no way relieve the physicians of forming a complete clinical decision. However, computers provide them with information on the ECG examination in a clearer fashion and save them from the burden of routine, repetitive, and subjective calculations. An additional advantage of using computers in electrocardiography is the ease of storage and retrieval of ECG data for further study and analysis.

Computer processing of ECG signals, which are analogue in nature with amplitudes in low millivolt range and low frequency content (0.05–150 Hz), involves digitization with a typical sampling frequency of 500 Hz and 12 bit resolution in analog-to-digital conversion process. Further processing involves digital filtering, removal of powerline and biological artifacts (like EMG interference), averaging and automatic measurement of amplitudes and durations of different parts of the ECG signal. Computer analysis programs developed based upon the interpretative experience of thousands of experts performed on millions of ECG records, provide important ECG waveform components, such as amplitude, duration, slope, intervals, transform domain features, and interrelationships between individual waves of the ECG signal relative to one another. These parameters are then compared with those derived from normal ECGs to decide whether there is an abnormality in the recordings. There are a wealth of algorithms and methods developed over several decades to assist with computer-aided analysis and interpretation of ECG signals.

Computer-assisted ECG analysis is widely performed in ECG monitoring and interpretation. This method is effectively deployed in routine intensive care monitoring of cardiac patients, where the ECGs of several patients are continuously monitored to detect life-threatening abnormalities. The purpose of monitoring is not only to detect and treat ventricular fibrillation or cardiac arrests, but also to detect the occurrence of less threatening abnormalities like heart blocks and arrhythmias. The occurrence of such episodes and their timely detection helps clinicians to make early diagnostic decisions and take appropriate therapeutic measures. The accurate detection of trends resulting in dangerous cardiac abnormalities by visual inspection of ECG displays or chart recorders is a difficult task. Computer-assisted analysis of ECG signals to extract baselines, amplitudes, slopes, and other important parameters to establish minimum and maximum values for these parameters provides an efficient and accurate means to track the nature of the ECG rhythms.

The ECG acquisition and analysis systems like many other biomedical devices are designed to measure physiological signals of clinical relevance and significance. To design and effectively use appropriate instrumentation to measure and process such signals, an understanding of the origin and properties of these signals are of prime importance. This article starts off with an attempt to present a distilled overview of the origin of bioelectric signals in general, and the electorcardiographic signal in particular. This overview sets the basis to briefly describe the cardiac vector and its projection along specific directions (leads) providing the surface ECG recordings and review the basic instrumentation necessary to record these signals. A detailed description of a high end computer-based 12 lead clinical ECG acquisition and analysis system provides an example to appreciate the role of computers in diagnostic electrocardiography. An overview of the role of computers in high resolution electrocardiography (body surface potential mapping) and other ECG-based diagnostic devices then follows. A brief introduction to computer-based ECG monitoring systems and a block diagram description of a QRS detection algorithm illustrate how computer programming serves as a basis to detect cardiac arrhythmias. The article ends with a web-based ECG telemonitoring system as an example.

REVIEW OF BASIC CONCEPTS

From cellular physiology, we recall that biopotentials are produced as a consequence of chemical activity of excitable or irritable cells. Excitable cells are components of the neural, muscular, glandular as well as many plant tissues. More specifically, biopotentials are generated as a consequence of ionic concentration difference of electrolytes (mainly Na^+, K^+, Cl^- ions) across the cellular membrane of excitable cells. Ionic differences are maintained by membrane permeability properties and active transport mechanisms across the cellular membrane (ionic pumps.)

The cellular membrane is a semipermeable lipid bilayer that separates the extracellular and intracellular fluids having different ionic concentrations. As a consequence of semipermeability and differences in concentration of ions, electrochemical gradients are set up across the membrane. Ionic transfer across the membrane by diffusion and active transport mechanisms results in the generation of a voltage difference (membrane potential), which is negative inside. The resting membrane potential is mainly established by the efflux of K^+ ions due to diffusion and is balanced by the consequent inward electric field due to charge displacement. This equilibrium voltage can be estimated by the Nernst equation (1), which results from application of electric field theory and diffusion theory. If we consider the effects of the three main ions, potassium

(K^+), sodium (Na^+), and chloride (Cl^-), the Goldman–Hodgkin and Katz equation can be used to calculate the resting membrane potential (1).

The smallest sources of bioelectric signals (biosources) are single excitable cells. These cells exhibit a quiescent or resting membrane potential across the cellular membrane of several millivolts (mV) (−90 mV with several hundred milliseconds in duration for ventricular myocytes). When adequately stimulated, the transmembrane potential in excitable cells becomes positive inside with respect to outside (depolarization) and action potentials are generated (at the peak of the action potential in ventricular myocytes, the membrane potential reaches about +20 mV). Action potentials are produced by sudden permeability changes of cellular membrane to ions: primarily sodium and potassium ions. Action potentials are all-or-none monophasic waves of depolarization that travel unattenuated with a constant amplitude and speed along the cellular membrane (Fig. 1).

The excitable cells function in large groups as a single unit and the net effect of all stimulated (active) cells produces a time-varying electric field in the tissue surrounding the biosource. The surrounding tissue is called a volume conductor. The electric field spreads in the volume

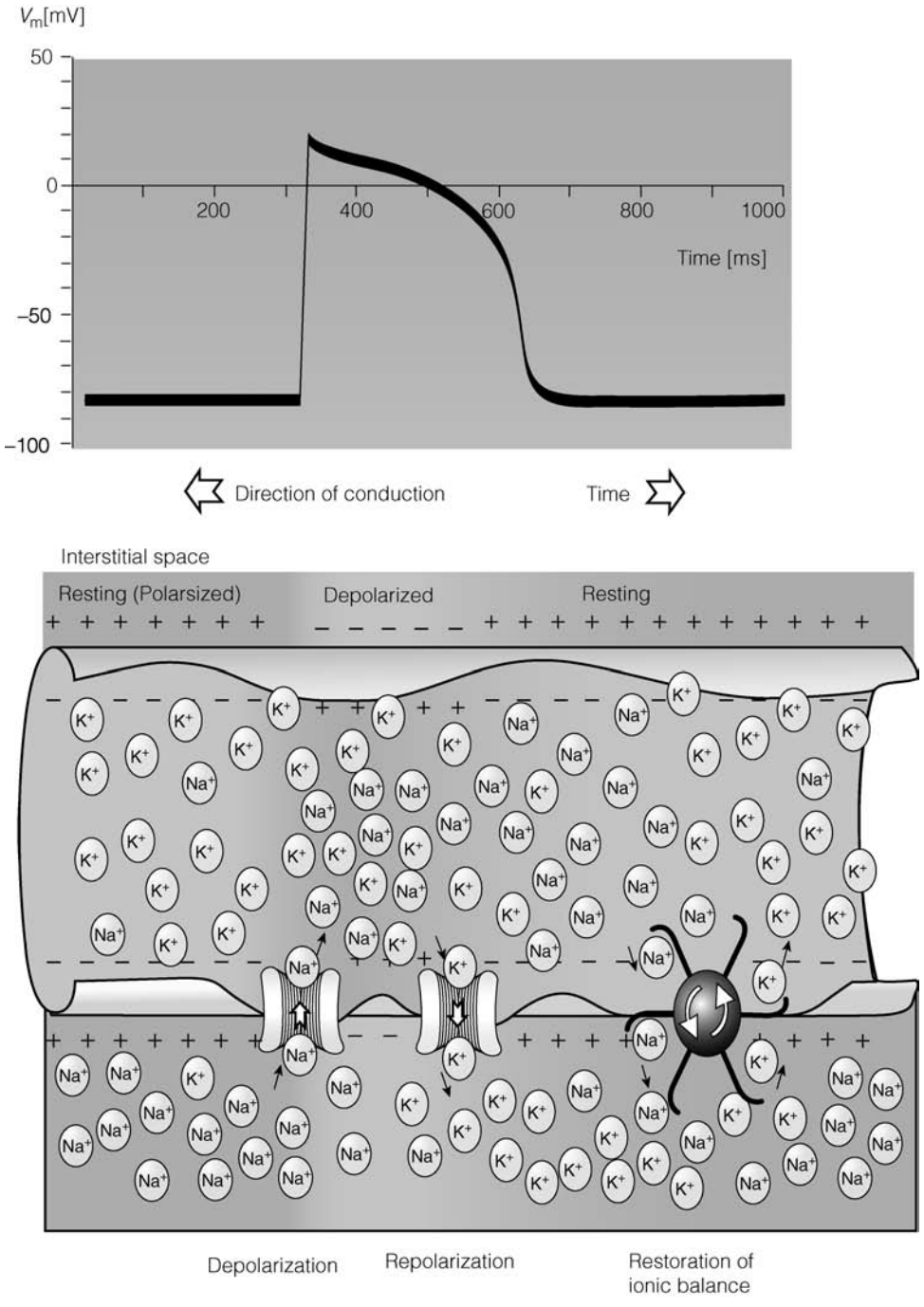


Figure 1. The monophasic action potential, direction of conduction of the action potential, and movement of ions across the cellular membrane. (Courtesy of Ref. 2.)

conductor and can be detected as small voltages by means of bioelectrodes or simply electrodes placed in the tissue or on the skin. Electrodes are sensors, which convert ionic current flow in the living tissue to electronic current flow in the electromedical instrument.

To understand the origin (electrogenesis) of biopotential signals like ECG, we should consider the following:

1. Electrical activity (bioelectric phenomena) at the cardiac cellular level and the extracellular potentials generated as the result of the electrical activity of single cardiac cells placed in a large homogeneous bathing (conducting) medium with the same composition as body fluids (volume conductor fields of simple bioelectric sources).
2. Extracellular potentials generated as the result of the electrical activity of a large number of myocardial cells (tissues) placed in a large conducting medium with the ionic composition of body fluids (volume conductor fields of complex bioelectric sources).
3. The relationship between these extracellular potentials and the gross electrical activity recorded on the body surface as ECG signals.

A simplified version of the volume conductor problem at the cellular level can be considered as follows. If a single excitable cell is placed in a bathing conductive medium, it acts like a constant current source. When the biosource becomes adequately depolarized, an action potential is generated across its membrane and it injects a current to the surrounding medium. The conductive medium presents as a load with a long range of loading conditions depending on its geometry, temperature, and so on. The lines of current flowing out of the excitable cell into the volume conductor with a specific resistance r, gives rise to an extracellular field potential proportional to the transmembrane current (i_m) and the medium resistance (r) according to Ohm's law. Obviously, the extracellular field potential increases with higher values of membrane current or tissue resistance.

There has been considerable debate about the exact relationship between the action potential across the cellular membrane and the shape of the extracellular field potential. However, the work of many researchers with different types of excitable cells has confirmed that the extracellular field potential resembles the second derivative of the transmembrane action potential. This means that a monophasic action potential creates a *triphasic* extracellular field potential (1). It has also been shown that the extracellular field potential is shorter in duration and much smaller in magnitude (μV compared to mV). Of course, this relationship has been established for cases when the geometry of the biosource and its surrounding environment is simple and the volume conductor is isotropic (uniform in all directions).

More realistically, when a piece of excitable tissue in the living organism becomes electrically active it becomes depolarized, acts like a point current source and injects a current into the anisotropic volume conductor comprised of tissues and body fluids surrounding it with different conductances (resistances). Consequently, the spatial distribution of current will not resemble that of a simple dipole placed in an isotropic volume conductor. However, it has been shown that an active nerve trunk (comprised of thousands of sensory and motor nerve fibers simultaneously stimulated) placed in a large homogeneous volume conductor generates an extracellular field potential which is quite similar in shape to that of a single nerve fiber (1). It is concluded that, the extracellular field potential is formed from the contributions of superimposed electric fields of the component biosources in the nerve trunk. The general form of the extracellular field potential of a nerve trunk in response to electrical stimulation is triphasic, it has amplitude in the microvolt range and it loses both amplitude and high frequency content at large radial distances from the nerve trunk. It is observed that the major contribution to the triphasic extracellular field potential is from the motor nerves in the trunk. It has also been shown that with a change in the volume conductor load (e.g., an increase in the specific resistance of the volume conductor or a decrease in the radial distance from the complex biosource) the amplitude of the recorded extracellular field potential increases (1).

The concepts discussed above are directly applicable to explain the relationship between the extracellular field potentials generated by complex and distributed biosources (current generators) like the cardiac tissue, the muscles, and the brain and their electrical activities recorded on the body surface as ECG, electromyogram (EMG) and electroencephalogram (EEG) signals.

In summary, excitable cells and tissues (biosources), when adequately stimulated, generate monophasic action potentials. These action potentials cause the injection of constant currents into a large bathing medium surrounding the biosource (considered as a point current source). As a result of the current flow in the volume conductor with specific resistance, extracellular field potentials are generated in the medium. These field potentials are triphasic in shape, of shorter duration and smaller amplitude compared to the transmembrane action potential. As the resistivity of the medium increases and the radial distance from the biosource decreases, the field potential increases. These field potentials are recorded as clinically useful signals on the body surface.

The biopotentials most frequently measured and monitored in modern clinics and hospitals are electrocardiogram (ECG: a recording of the electrical activity of the heart), electromyogram (EMG: a recording of the electrical activity of the muscle), electroencephalogram (EEG: a recording of the electrical activity of the brain), and others. Based on the physiological concepts reviewed above, now we present a brief overview of the electrogenesis of the ECG signals that carry a wealth of information about the state of health and disease of the heart. Having done this, we look at basics of electrocardiography.

BASICS OF ELECTROCARDIOGRAPHY

The conduction system of the heart consists of the sinoatrial (SA) node, the internodal tracts, the atrioventricular

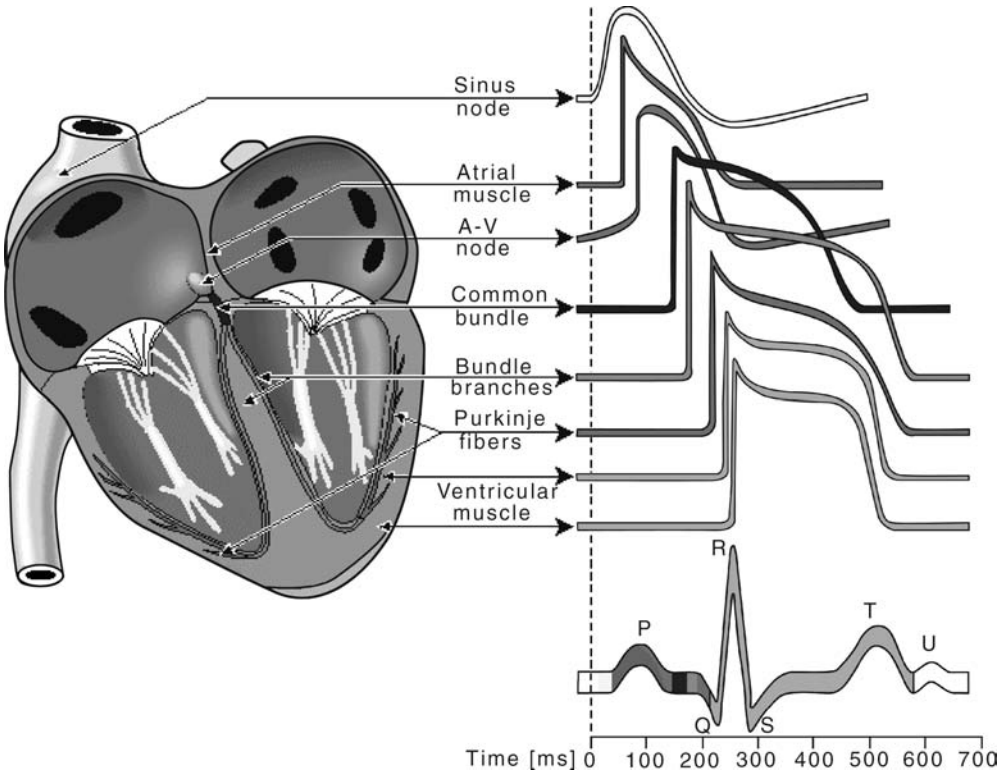


Figure 2. Wavesforms of action potentials in different specialized cells in the conductive pathway of a normal heart and their contribution with color coding to the surface ECG. (Courtesy of Ref. 2.)

(AV) node, the bundle of histidene (His), the right bundle branch (RBB), the left bundle branch (LBB), and the Purkinjie network. The rhythmic electrical activity of the heart (cardiac impulse) originates in the SA node. This node is known as the natural pacemaker of the heart, approximately the size of the tip of a pencil, located at the junction of the superior vena cava and the right atrium. The impulse then propagates through internodal and interatrial (Buchmans's bundle) tracts. As a consequence, the pacemaker activity reaches the AV node by cell-to-cell atrial conduction and activates the right and left atrium in an organised manner. The pacemaker action potential has a fast activation phase, a very short steady recovery phase, followed by a fairly rapid recovery phase and a characteristic slow depolarization phase leading to self-excitation (Fig. 2). The pacemaker cells of the SA node act as a biological oscillator.

As atria and ventricles are separated by fibrous tissue, direct conduction of cardiac impulse from the atria to the ventricles can not occur and activation must follow a path that starts in the atrium at the AV node. The cardiac impulse is delayed in the AV node for $\sim$ 100 ms. It then proceeds through the bundle of His, the RBB, the LBB, and finally to the terminal Purkinjie fibers that arborize and invaginate the endocardial ventricular tissue. The delay in the AV node is beneficial since electrical activation of cardiac muscle initiates its successive mechanical contraction. This delay allows enough time for completion of atrial contraction and pumping of blood into the ventricles. Once the cardiac impulse reaches the bundle of His, conduction is very rapid, resulting in the initiation of ventricular activation over a wide range. The subsequent cell-to-cell propagation of electrical activity is highly sequenced and coordinated resulting in a highly synchronous and efficient pumping action by the ventricles.

Essentially, an overall understanding of the genesis of the ECG waveform (cardiac field potentials recorded on the body surface) can be based on a cardiac current dipole model placed in an infinite (extensive) volume conductor. In this model, an active (depolarizing) region of the tissue is considered electronegative with respect to an inactive (repolarizing) region. Therefore, a boundary or separation exists between negative and positive charges. This is regarded as a current dipole: a current source and sink separated by a distance. According to the dipole concept, a traveling excitation region can be considered as a dipole moving with its positive pole facing the direction of propagation. Thus a nearby recording electrode placed in the surrounding volume conductor (referenced to an indifferent electrode placed in a region of zero potential) will detect a positive-going field potential as excitation approaches and a negative-going field potential as it passes away. Repolarization (recovery) is considered as a dipole with its negative pole facing the direction of propagation. Therefore, the propagation of excitation can be considered as the advance of an array of positive charges with negative charges trailing and the recovery could be considered as the approach of negative charges with positive ones trailing. Consequently, an upward deflection in the biopotential recording indicates the approaching of excitation

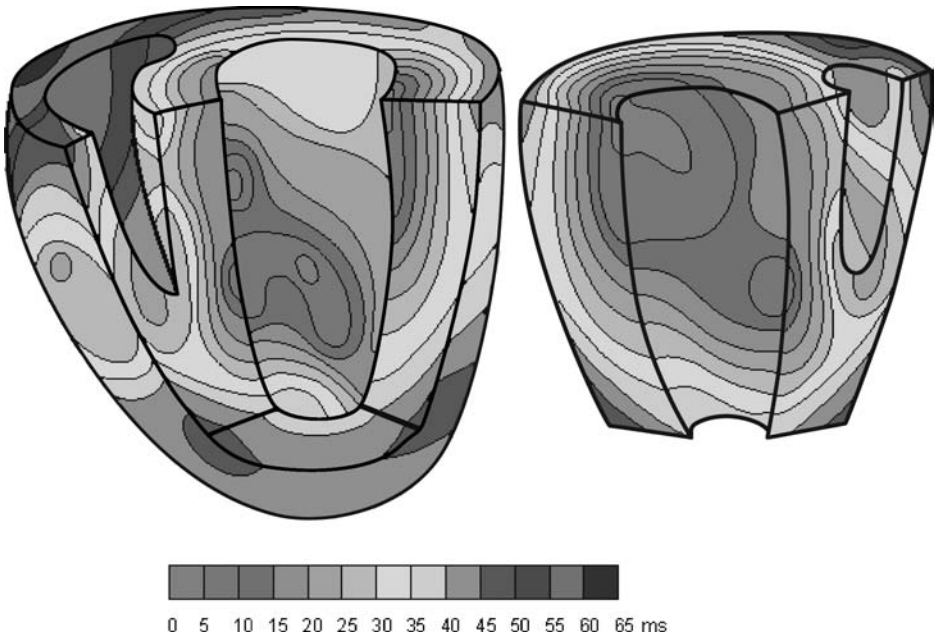


Figure 3. Isochrone surfaces in ventricular activation color coded to show spatiotemporal propagation. (Courtesy of Ref. 2.)

(depolarization) toward the positive (recording) electrode and a downward deflection indicates a recovery (depolarization) in the recorded signal.

As the wave of excitation (depolarization) spreads throughout the conductive pathways and tissues, specific excitation regions called isochrones are synchronously excited. In the ventricles, these synchronous activation regions propagate in a temporally and spatially orderly fashion from the endocardial to the epicardial direction (Fig. 3.)

In a localized region of the heart many cells are simultaneously activated because of the high electrical and mechanical coupling (functional syncytium) between the myocardial cells. Each activation region can be viewed as an elementary dipole, and all elementary dipoles could be vectorially added to all others to form a single net dipole. (For more details on simple and multiple dipole models see the section Electrocardiography.) Therefore, at each instant of time, the total cardiac activity can be represented by a net equivalent dipole current source. The electric field produced by this dipole source represents the total electrical activity of the heart and is recorded at the body surface as the ECG signal (Fig. 4). (For quantitative details see chapter 6 in Ref. 2.)

In the early 1900, Einthoven postulated that the cardiac excitation could be viewed as a vector. He drew an equilateral triangle with two vertices at two shoulders and one at the navel (representing the left leg). With the cardiac vector representing the spread of cardiac excitation inside the triangle, the potential difference measured between two vertices of the triangle (known as the limb leads) with respect to right leg, is proportional to the projection of the vector on each side of the triangle (Fig. 5).

In summary, based on the aforementioned concepts, electrocardiographers have developed an oversimplified model to explain the electrical activity of the heart. In this model, the heart is considered as an electric dipole (points of equal positive and negative charges separated from one another by a distance), denoted by a spatiotemporally changing dipole moment vector **M**. This dipole moment (amount of charge times distance between positive and negative charges) is called the cardiac vector. As the wave of depolarization spreads throughout the cardiac cycle, the magnitude and orientation of the cardiac vector changes and the resulting bioelectric potentials appear throughout the body and on its surface. The potential differences (ECG signals) are measured by placing electrodes on the body surface and connected to biopotential amplifier. (For details, see the section Bioelectrodes, and Electrocardiographic Monitors.) In making these potential measurements, the amplifier has a very high input impedance to minimally disturb the cardiac electric field that produces the ECG signal. As it was discussed before, when the depolarization wavefront points toward the recording positive electrode (connected to the + input terminal of the bioamplifier), the output ECG signal will be positive going, and when it points toward the negative electrode, the ECG signal will be negative going. The time varying cardiac vector produces the surface ECG signal with its characteristics P wave, QRS complex, and T wave during the cardiac cycle. These field potentials are measured by using bioelectrodes and biopotential amplifiers to record the ECG tracings.

BASIC INSTRUMENTATION TO RECORD ECG SIGNALS

As the electrical events in the normal heart precede its mechanical function, ECG signals are of great clinical value in diagnosis and monitoring of a wide variety of cardiac abnormalities including myocardial infarction and chamber enlargements. Therefore, ECG signal acquisition systems are widely used in cardiology, cardiac catheterization laboratories, intensive and cardiac care units, and at patient's bedside, among other areas.

As it was described earlier, the electrical activity of the heart can be best modeled and characterized by vector quantities. However, it is easier to measure scalar

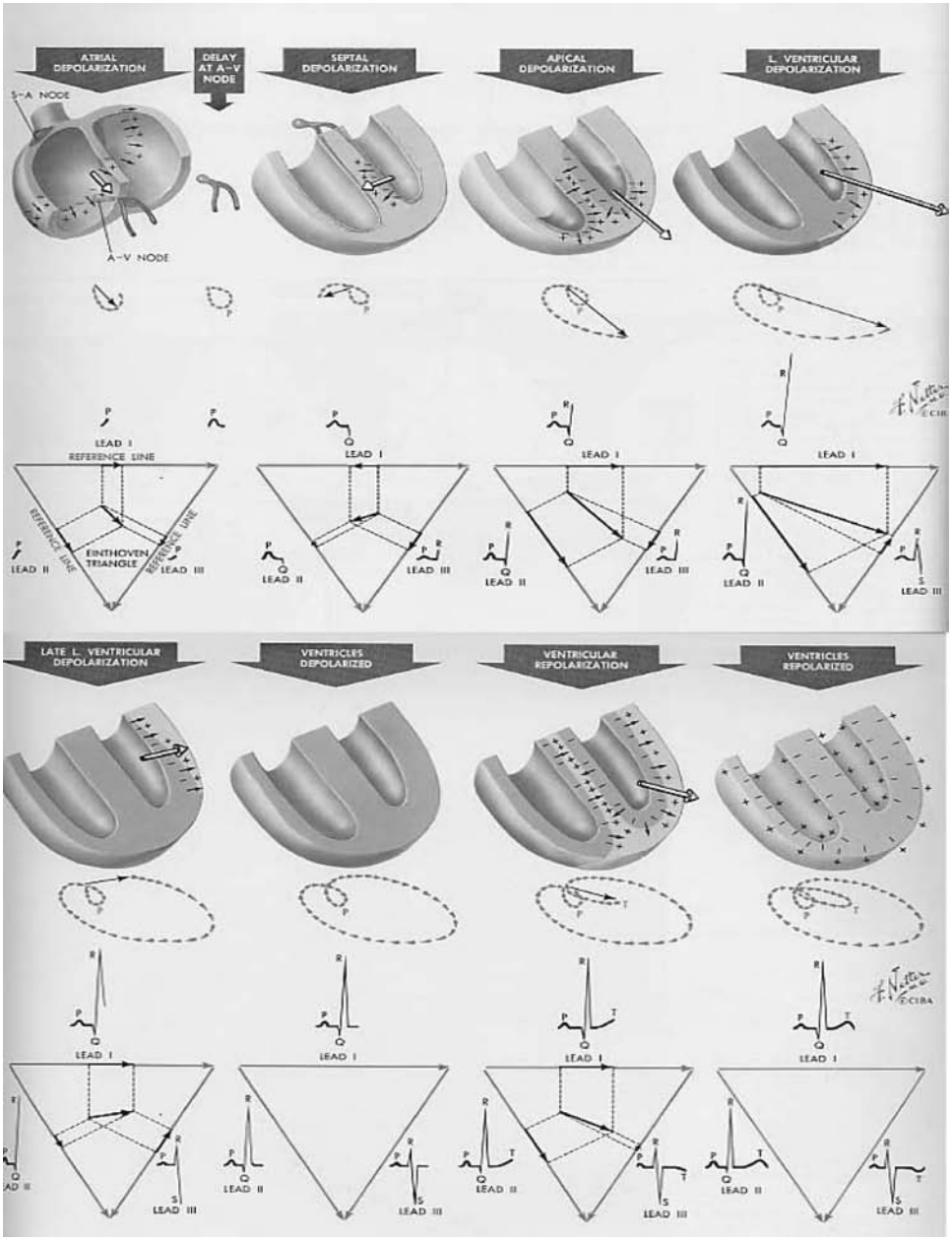


Figure 4. The total cardiac electrical activity represented by the net dipole (cardiac vector) during different phases of the cardiac cycle and its projection along the frontal plane electrocardiographic leads (I, II, and III). (Courtesy of Ref. 3.)

quantities, such as potential differences between specified points on the torso known as surface ECGs. These ECG signals have a diagnostically significant frequency content between 0.05 and 150 Hz. To ensure stability of the baseline, a good low frequency response is required. The instabilities in the baseline recordings originate from changes in the electrode–electrolyte interface at the point of contact of the bioelectrode with the skin. (For details see the section Bioelectrodes.) To faithfully record fast changes in the ECG signals and distinguish between other interfering signals of biological origin, adequate high frequency response is necessary. This upper frequency value is a compromise between several factors including limitation of mechanical recording parts of ECG machines using direct writing chart recorders.

To amplify the ECG signals and reject nonbiological (e.g., powerline noise) as well as biological interferences (e.g., EMG), differential amplifiers (DAs) with high gains (typically 1000 or 60 dB) and excellent common mode rejection capabilities must be used. Typically, common mode rejection ratios (CMMRs) in the range of 80–120 dB with 5 KΩ imbalance between differential amplifier input leads provide a desirable level of environmental and biological noise and artifact rejection in ECG acquisition systems. In addition to this, in very noisy environments it becomes necessary to engage a notch (band-reject) filter centered at 60 Hz or 50 Hz (in some countries) to reduce powerline noise further. A good review of adaptive filtering method applied to ECG powerline noise removal is given in Adaptive Filter Theory by Simon Haykin (4).

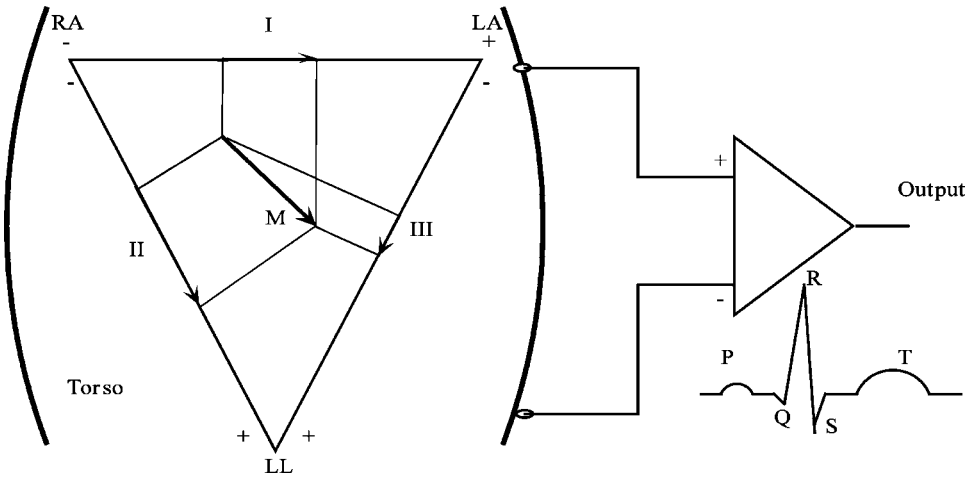


Figure 5. Einthoven equilateral triangle. The vertices are LA (left arm), RA (right arm), and LL (left leg). The *RL* (right leg) is used as a reference for potential difference measurements and is not shown. I, II and III represent electrocardiographic frontal limb leads. The + and – represent connection to the terminals of an ECG (biopotential) amplifier. Lead I is the potential difference between LA and RA. Lead II is the potential difference between LL and RA. Lead III is the potential difference between LL and LA. (The input polarity of the amplifier shown is for recording: III limb.)

Many modern biomedical instruments use instrumentation amplifiers (IAs). These amplifiers are advanced versions of differential amplifiers enhanced with many additional desirable characteristics such as very high input impedance > 100 MΩ) to prevent loading the small ECG signals to be picked up from the skin. The final stages of the ECG amplifier module limit the system's response (band pass filtering) to the desirable range of frequencies for diagnostic (i.e., 0.05–150 Hz) or monitoring (i.e., 0.5–40 Hz) purposes. A more limited bandwidth in the monitoring mode provides improved signal to noise ratio and removes ECG baseline drift due to half-cell potentials generated at the electrode/electrolyte interface and motion artifacts. Driven right-leg amplifiers improve the CMRR. The amplified and adequately filtered ECG signals are then applied to display, recording or digitization modules of a computer-based ECG acquisition system. Detailed specifications for diagnostic ECGs have been developed by the American National Standards Institute (5). For more detailed specification and design as well as other aspects of ECG instrumentation and measurements, see the Appendix, Bioelectrodes, Electrocardiography, and Electrocardiographic Monitors.

COMPUTER SYSTEMS IN ELECTROCARDIOGRAPHY

Diagnostic Computer-Based ECG Systems

Heart diseases cause a significant mortality rate in the world. Accurate diagnosis of heart abnormalities at an early stage could be the best way to save patients from disability or death. The ECG signal has considerable diagnostic value and ECG monitoring is a very well established and commonly used clinical method. Diagnostic ECG testing is performed in a doctor's office, in a clinic or hospital as a routine check up. In this test, a full 12-lead ECG (to be described later) is acquired from a resting subject and displayed on a chart recorder or a computer screen to diagnose cardiac diseases. In cardiac care units (CCUs), a patient's single lead ECG may be continuously acquired and displayed on a cathode ray tube (CRT) or a computer screen for signs of cardiac beat abnormalities. The ECG monitoring capabilities are now an integral part of a number of other medical devises, such as cardiotachometers, Holter monitors, cardiac pacemakers, and automatic defibrillators.

Most of the traditional clinical ECG machines used a single channel amplifier and recording system. The recoding was achieved using a direct writing chart recorder. In modern systems, however, computer memory, display screen and printing capabilities are deployed to record, display and report the ECG data. Traditional single channel systems used a multiposition switch to select the desired lead connection (I, II, III, aV_R, aV_L, aV_F, V_1, V_2, V_3, V_4, V_5, and V_6) and apply it to the biopotential amplifier and chart recorder. Only one ECG lead at a time could be selected and recorded with these machines. The block diagram of a modern single channel computer-based ECG acquisition system is shown in Fig. 6.

Most of the modern ECG machines are multichannel systems. They include several amplifier channels and record several ECG leads simultaneously. This feature enables them to considerably reduce the time required to complete a set of standard clinical ECG recordings. As the ECG leads are recorded simultaneously, they can be shown in their proper temporal relationship with respect to each other. These systems use microprocessors to acquire the cardiac signals from the standard 12-lead configuration by sequencing the lead selector to capture four groups of three lead signals, and switch groups every few seconds. The high end computer-based systems capture all 12-leads simultaneously and are capable of real-time acquisition and display of the standard 12-lead clinical ECG signals (see Example below.)

Modern ECG acquisition systems achieve certain desirable features, such as removal of artifacts, baseline wander, and centering of the ECG tracings by using specialized computer algorithms. These systems perform automatic self-testing on power up and check for lead continuity and polarity and indicate lead fall-off or reversal. They deploy digital filters implemented in software to considerably improve the ECG signal quality and automatically remove baseline drift and reduce excessive powerline and biological noise. Powerful software programs not only minimize

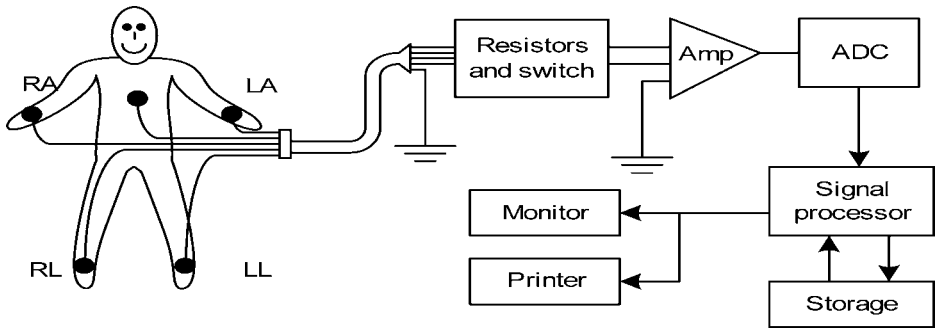


Figure 6. Block diagram of a three-lead electrocardiograph. The normal locations for surface electrodes are right arm (RA), right leg (RL = ground or reference), left arm (LA), and left leg (LL). In 12-lead electrocardiographs, six electrodes are attached on the chest of the patient as well. (Courtesy of Ref. 6.)

baseline drift without signal distortion during rest, they also produce high quality ECG tracings during patient monitoring, exercise and ambulation.

12-Lead Clinical Electrocardiography. The most commonly used or standard clinical ECG system is the 12-lead system. This system is comprised of three bipolar limb leads (I, II, III) connected to the arms and legs; three augmented leads (aV_R, aV_L, aV_F); and six unipolar chest or precordial leads (V_1, V_2, V_3, V_4, V_5, V_6).

The connections to measurement points in the 12-lead system are shown in Fig. 7. Six of these 12 leads are frontal leads (the bipolar and the augmented leads) and six of them are transverse leads (the precordial leads). The frontal leads are derived from three measurement points RA, LA, LL with reference to RL. Therefore, any two of these six leads contain exactly the same information as the other four. The dipole source model can be used to explain the total electrical behaviour of the heart (See the section Electrocardiography and Ref. 2.)

Basically, any two of the three I, II, and III leads could represent the cardiac activity in the frontal plane and only one chest lead could be used to represent the transverse activity. Chest lead V_2 is a good representative of electrical activity in the transverse plane as it is approximately orthogonal (perpendicular) to the frontal plane. Overall, as the cardiac electrical activity could be modeled as a dipole, the 12-lead system could be considered to have three independent leads and nine redundant leads. However, since the chest leads also detect unipolar elements of the cardiac activity directed toward the anterior region of the heart, they carry significant diagnostic value in the transverse plane. As such, the 12-lead ECG system can be considered to have eight independent and four redundant leads. Now the question arises why all the 12-leads are recorded then. The main reason is that the 12-lead system enhances pattern recognition and allows cardiologists to compare the projections of the resultant vectors in the two orthogonal planes and at different angles. This facilitates and validates the diagnostic process. Figure 8 shows the surface anatomical positions for the placement of the bioelectrodes in 12-lead electrocardiography.

A Detailed Example: The Philips PageWriter Touch 12-Lead ECG System. Currently, there are a number of highly advanced 12-Lead ECG data acquisition, analysis and interpretative systems on the market. A detailed description and comparison of all these systems are beyond the scope of this chapter (for information on these systems see manufacturers web sites). Due to space limitation, only one representative system will be described in detail as an example. The Philips PageWriter Touch (Philips Medical, MA) is an advanced 12-lead ECG acquisition and analysis system with many user-friendly features. It has a compact design and is best suited for busy hospitals and fast-paced clinical environments. It has an intuitive touch screen which is fully configurable and it has a powerful build-in interpretative 12-lead ECG signal analysis algorithm developed for rapid and accurate interpretation of ECG signals (Fig. 9).

The PageWriter supports a variety of ECG data acquisition modes, display settings and reporting schemes. It provides real-time color-coded ECG signals enabling the user to perform quality control checks on the acquired ECG data. Quality control features include surface anatomical diagrams that alert the user to the location of loose or inoperable electrodes (Fig. 8). It is equipped with a full-screen preview display of the ECG report before print out. It has an ergonomic design that facilitates ECG data collection from the patient bedside (Fig. 10).

The PageWriter Touch has an alphanumeric keyboard, an optional barcode scanner, and an optional magnetic card reader to increase the speed and accuracy of patient data entry and management. Data handling features include indexed ECG storage and quick transmission of stored ECG data to an ECG data management system. By using indexed thumbnail images of each ECG report the user can instantly view and print the stored reports. Multiple reporting formats may be applied to any saved ECG records.

In brief, the PageWriter Touch features accurate and fast 12-lead ECG signal acquisition and analysis capabilities for both adults and children patients. It has real-time color-coded display facilities and provides instantaneous snapshots of stored ECG data on the touch screen. It has a variety of ECG data review and printing capabilities. The touch screen provides easy and flexible configurability of features (Fig. 11). (For more details on Technical Specifications of PageWriter Touch see the Appendix.)

Overview of System Description. The PageWriter Touch electrocardiograph performs acquisition, analysis, presentation, printing, storage, and transfer of ECG signals as well as other patient data. A generalized block diagram of the system is shown in Fig. 12.

The system is comprised of three major subsystems equipped with an LCD display and touch screen module. A brief and high level description of these subsystems is as follows.

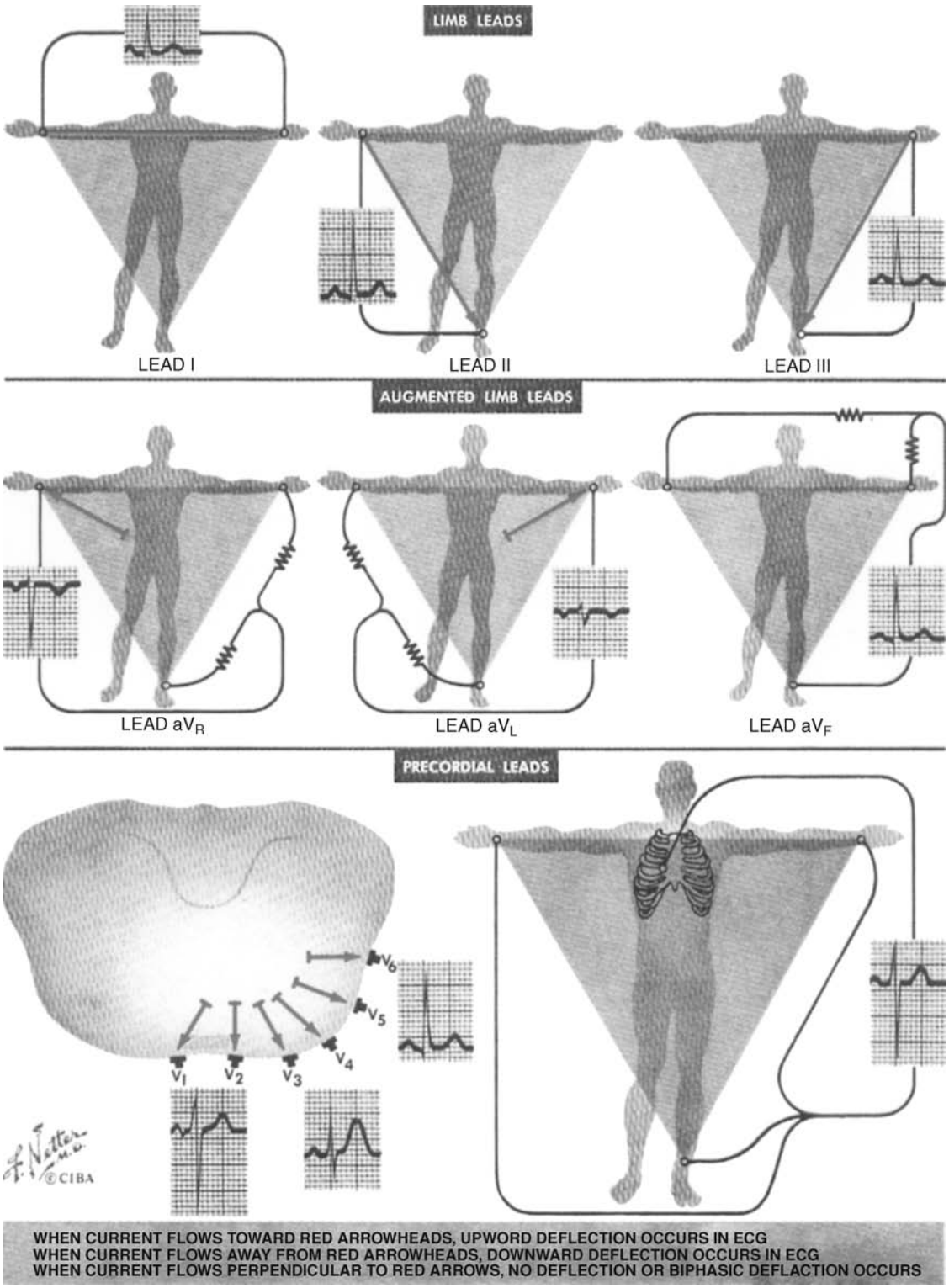


Figure 7. Lead connections in standard 12-lead electrocardiography. (Courtesy of Ref. 3.)

1. Main Controller Board. This is a single board computer (SBC) with extensive Input–Output facilities, running Windows CE 3.0. The PageWriter application software controlling the overall operation of the electrocardiograph runs on the Main Controller Board, which includes the display and user-input subsystems. The application software interacts with numerous hardware and software subsystems. The Main Controller Board SBC contains loader software and Windows CE kernel image in its

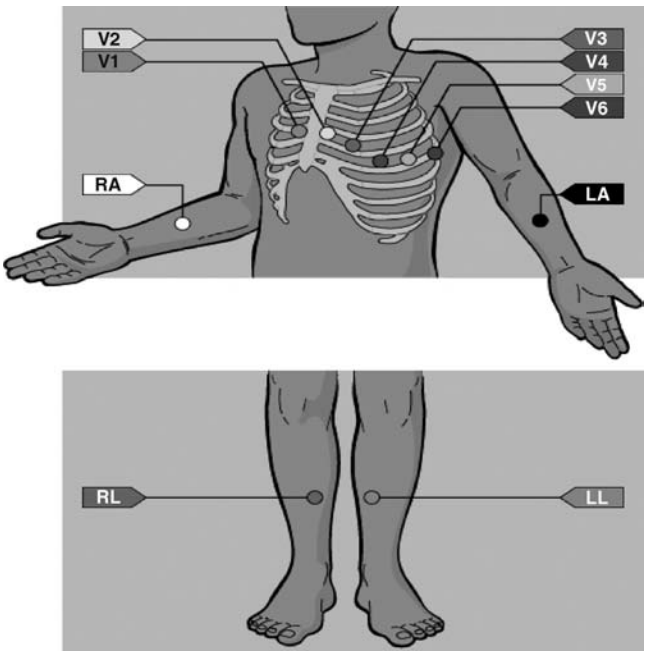


Figure 8. Surface anatomical positions for application of bioelectrodes in the 12-lead ECG system. (Courtesy of Philips Medical Systems, USA.) (Illustration courtesy of Mark Wing, Illustrator.)

internal flash memory (32 MB). At system boot, a system RAM test is performed by the loader (64 MB RAM onboard). The Windows CE kernel loads next. After loading of the CE, the application launcher runs, verifying system and executable images before loading the SierraGUI (Graphical User Interface) application. All interactions with the system operator (user) are implemented through the SierraGUI application. The application software and all ECG archives are stored on a separate 128 MB CompactFlash (CF) card installed on the Main Controller Board (Fig. 12.)

Figure 9. PageWriter Touch, a high-end 12-lead electrocardiograph with built-in 12-lead ECG signal interpretation capabilities. (Courtesy of Philips Medical Systems, USA.)

Figure 10. Easy data collection set up at bedside. (Courtesy of Philips Medical Systems, USA.)

2. Printer Controller Board. This is a controller board that provides all the real-time management of the printer. The Printer Controller Board communicates with the Main Controller Board through a USB port. The Printer Controller Board is a microprocessor-based control board for the electrocardiograph thermal printer mechanism. This board is connected by a USB port to the Main Controller Board and is powered by the power circuit of the Main Controller Board. It provides ECG waveform rendering and basic bitmap imaging operations, and uses a PCL-like control language API for page description and feed control. It controls the print head, motor, and detects drawer-open and top-of-form.
3. Patient Input Module (PIM). This is a controller running Windows CE 3.0, coupled with a signal acquisition board employing mixed-signal Application Specific Integrated Circuit (ASIC) technology developed for ECG data acquisition. The PIM communicates with the Main Controller Board through a USB port (Fig. 12.)

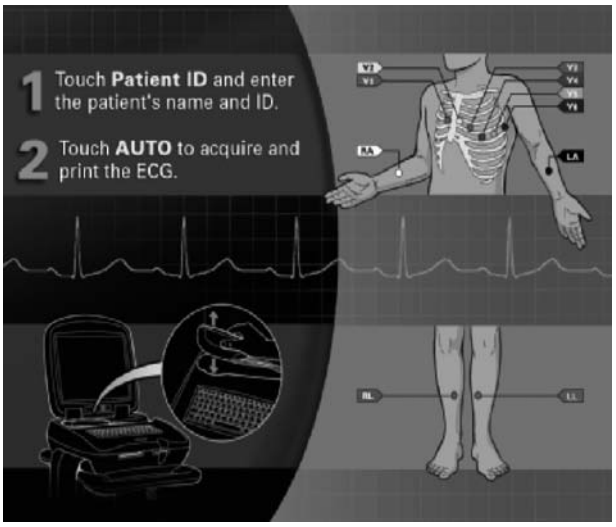


Figure 11. PageWriter Touch electrocardiograph has a user-friendly interface. (Courtesy of Philips Medical Systems, USA.) (Illustration courtesy of Mark Wing, Illustrator.)

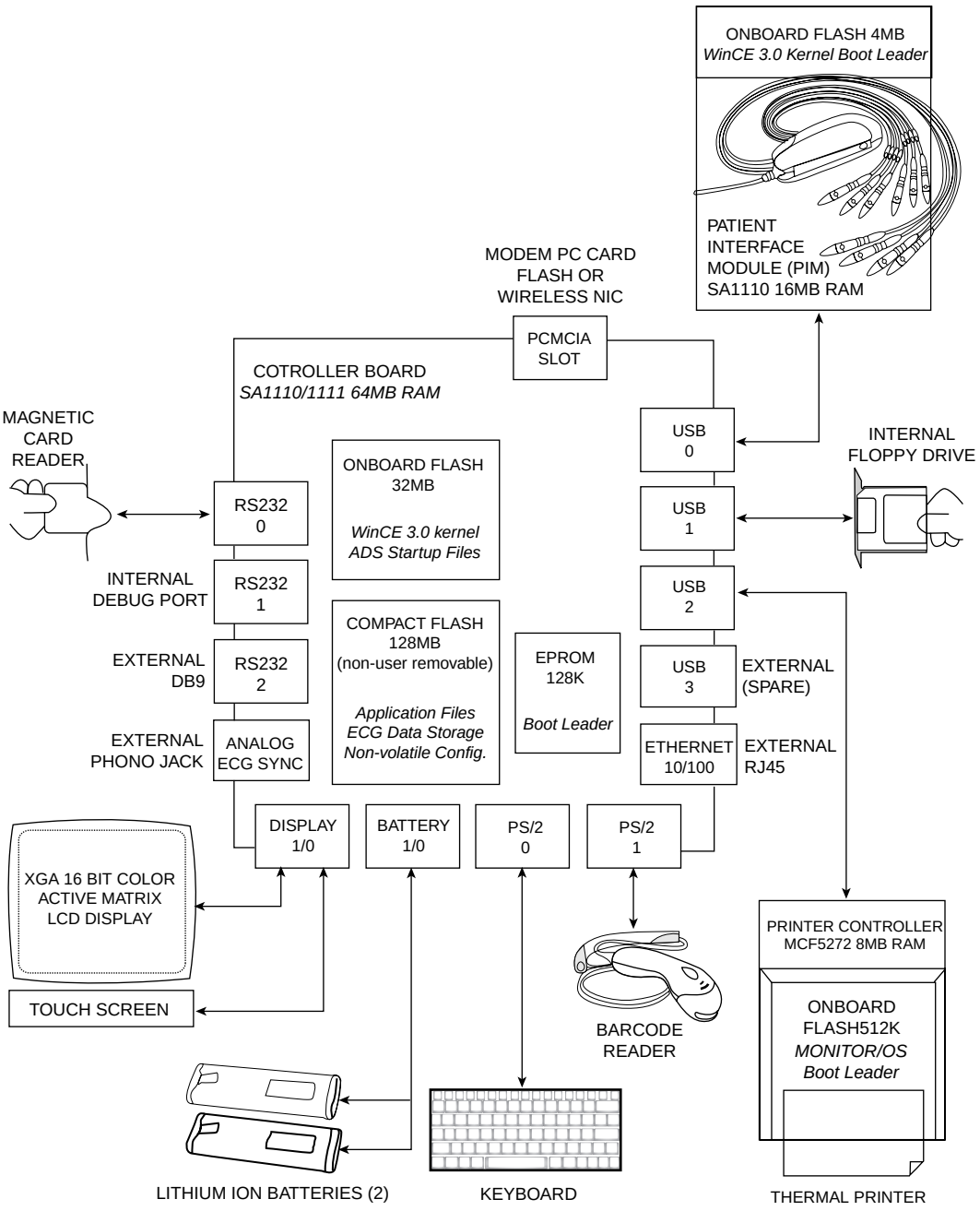


Figure 12. A generalized block diagram for PageWriter Touch. (Courtesy Philips Medical Systems, USA.)

The LCD display module for the PageWriter Touch electrocardiograph is an e**x**tended **g**raphic **a**rray or XGA-compatible, full-color with backlight and overlaid touch screen (Fig. 9.) It is driven by the Main Controller Board using a graphics accelerator chip and dedicated touch screen support hardware. The touch screen provides finger-tap input substituting for the normal Win32 mouse-click input.

ECG Data Flow and Storage in PageWriter Touch. The ECG data flow and storage in the PageWriter Touch electrocardiograph is shown in Fig. 13.

The ECG signals are picked up at the body surface using electrodes attached to properly prepared skin at specified anatomical sites. The ECG data stream is then routed in real-time to the Main Controller Board, where it is written into the application buffers in RAM. These buffers are used to present the ECG data stream on the LCD screen in real-time. When the user initiates an AUTO report print, presses the ACTION button on the Patient Input Module or the Snapshot button on the display, or uses the Timed ECG acquisition, the corresponding 10 s segments of the ECG data are then copied to the temporary ECG storage in RAM. These 10 s segments are named ECG reports that

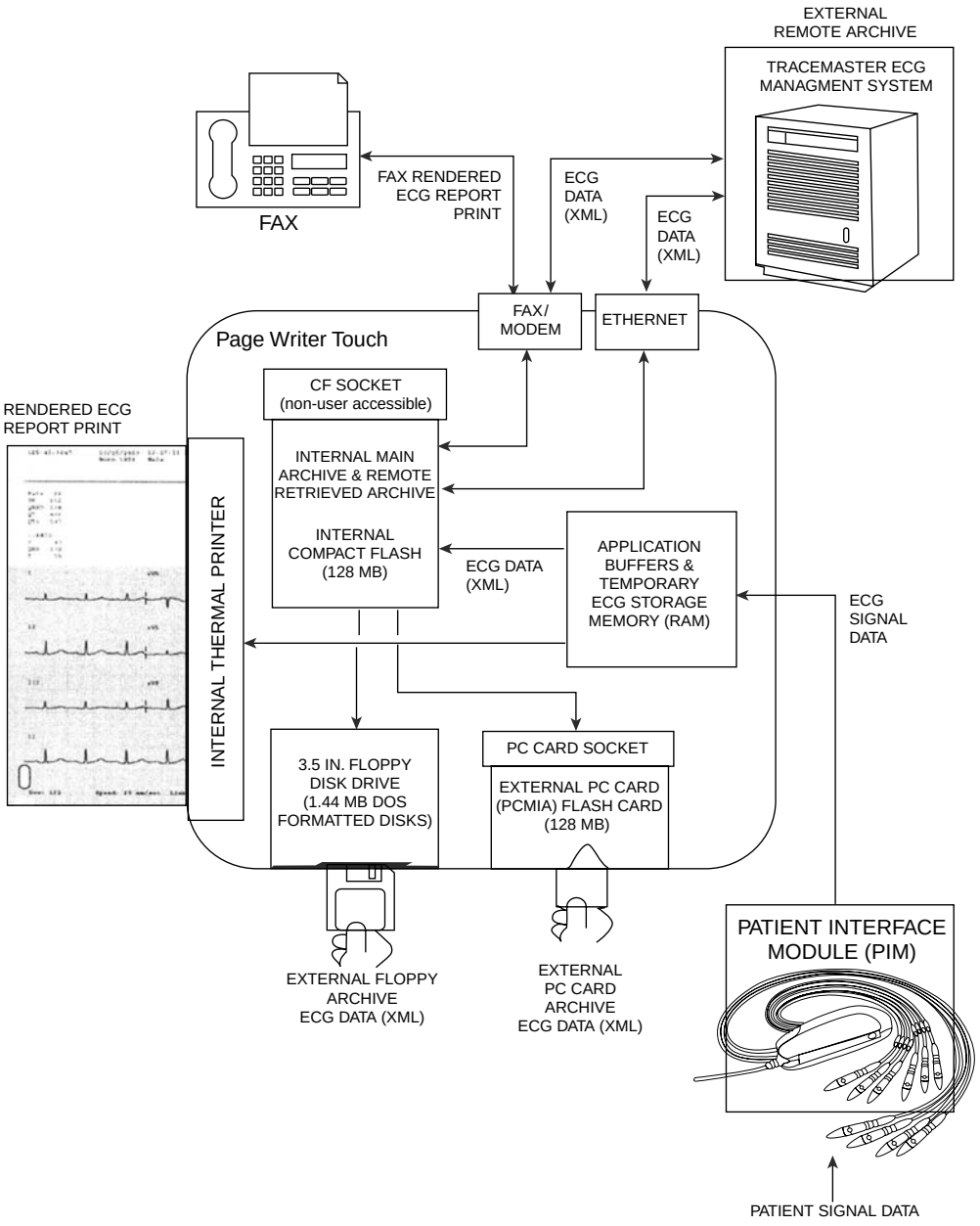


Figure 13. The ECG data flow and storage in PageWriter Touch electrocardiograph. (Courtesy of Philips Medical systems, USA.)

can be previewed and printed. In AUTO mode, the ECG report may be automatically printed after copying to storage. An ECG report contains waveforms, analysis information, patient demographics, and acquisition information, along with operator and device information. Figure 14 shows a rendered ECG report print sample.

The PageWriter Touch 12-Lead ECG Data Analysis Program. This program analyzes up to 12 simultaneously acquired ECG waveforms over a 10 s period using interpretive criteria by patient-specific information. For examples of ECG diagnostic criteria see Ref. (7). The analysis algorithm analyzes ECG data and produces an interpretive report to assist the clinician make more informed patient assessment in a broad range of clinical settings. The analysis program is developed to allow clinicians read and interpret ECG findings more quickly and efficiently, provide accurate, validated ECG measurements to facilitate physician overreading and improve treatment decision making, generate detailed findings that highlight areas of interest for physicians review, provide high levels of reproducibility for more consistent ECG interpretation.

The algorithm monitors the quality of the ECG waveforms acquired from 10 electrodes (12 leads), recognizes patterns, and performs basic rhythm analysis. Using advanced digital signal processing techniques, the algorithm removes biological and non-biological noise and artifacts while minimizing distortion of the ECG waveforms

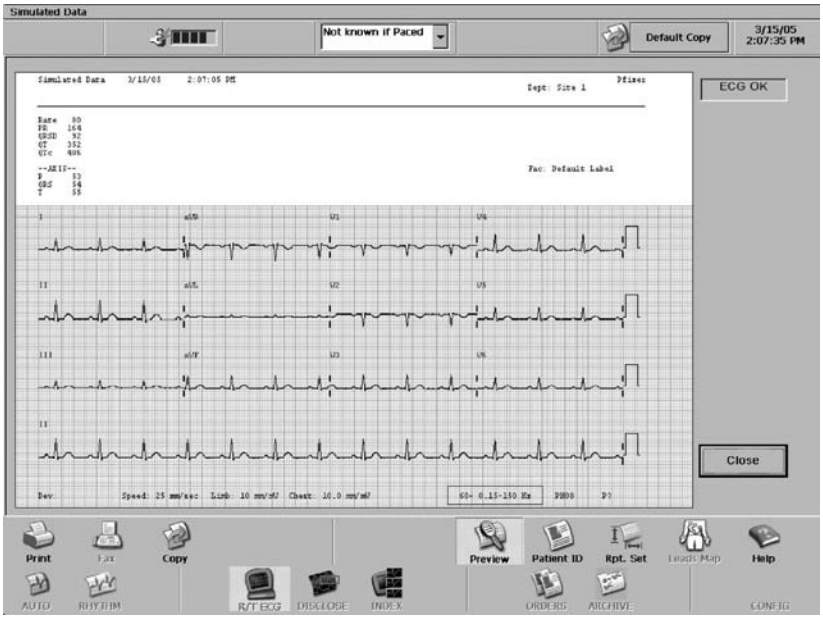


Figure 14. A 12-lead ECG report sample for printing or copying. (Courtesy of Philips Medical systems, USA.)

and preserving their diagnostic quality. By using a set of diagnostic criteria, the 12-lead ECG analysis program generates interpretative statements, summarizes the findings for the ECG signals, and highlights key areas of concern for physician review. The analysis program has a built-in pacemaker pulse detector and paced rhythm classification routine. The program distinguishes a variety of atrial, ventricular and atrioventricular (AV) sequential pacing modes to recognize asynchronous pacing typically seen with a magnet in place. In addition to automated detection capabilities, the algorithm provides user selected configuration of "pacemaker patient" or "non-pacemaker patient" for more accurate analysis.

The 12-lead analysis program also incorporates factors, such as age and gender that impact a patient's risk for developing specific forms of cardiac disease. It uses gender-specific interpretation criteria to take into account key physiological differences between males and females. For example, it applies gender-specific evaluation of Q waves for improved detection of Acute Myocardial Infarction (AMI), which is often missed in female patients. It also uses more gender-specific criteria for detection of axis deviation, Left Ventricular Hypertrophy (LVH), and prolonged QT segment.

The 12-lead algorithm also includes an advanced Pediatric Criteria Program, which uses age to select clinically relevant interpretative statements related to cardiac rhythm and morphology. If a patient's age is < 16 years, the algorithm automatically uses pediatric ECG interpretation criteria, which accounts for higher rates and narrower QRS complexes in this patient population. The Pediatric Criteria Program recognizes 12 distinct age groups to ensure most relevant-age interpretation criteria are applied for analyzing the ECG data. In fact, patient age is used to define normal limits in heart rate, axis deviation, ECG segment time intervals, voltage values for interpretation accuracy in tachycardia, bradycardia, prolongation or shortening of PR and QT intervals, hypertrophy, early repolarization, myocardial ischemia and infarct, as well as other cardiac conditions. The pacemaker detection and classification algorithm built into the Pediatric Analysis Program has the ability to reliably distinguish pacemaker pulses from the very narrow QRS complexes often produced by neonatal and pediatric patients. It also reduces the likelihood of false diagnosis in non-paced patients, while enabling more accurate paced rhythm analysis for the pediatric age group.

To improve on the classification accuracy of a diagnostic dilemma in pediatrics to distinguish between mild Right Ventricular Hypertrophy (RVH) and Incomplete Right Bundle Branch Block (IRBBB), the algorithm combines 12-lead synthesized vectorcardiogram transverse plane measurements with scalar ECG measurements. In addition, the pediatric analysis program provides improved QT measurements in pediatric patients.

The ST segment elevation in ECG tracings is an effective indicator of acute myocardial infraction. The PageWriter Touch 12-Lead ECG Analysis Program provides advanced software for detection of ST Elevation Acute Myocardial Infarction (STEMI). This software feature provides clinicians with a valuable decision support tool when working with patients presenting symptoms that suggest accurate coronary syndromes.

Other cardiac conditions, such as benign early repolarization and acute pericarditis, tend to mimic the ECG diagnosis of STEMI, and degrade algorithm detection accuracy. To address this difficulty, the ECG Analysis Program separates the confounders by examining the patterns of ST elevation. Improved measurements in ST deviation enables the algorithm to achieve both high sensitivity and specificity in more accurate detection of STEMI condition.

Exercise (Stress) Electrocardiography. Exercise (Stress or Treadmill) electrocardiography is a valuable diagnostic and screening procedure primarily used to diagnose

coronary artery disease (CAD). Exercise on treadmill may induce ischemia that is not present at rest. Exercise electrocardiography is usually performed to screen for the presence of undiagnosed CAD, to evaluate an individual with chest pain, to clarify abnormalities found on a resting ECG test, and to assess the severity of known CAD.

Accurate interpretation of stress ECG is dependent on a number of factors including a detailed knowledge of any prior medical condition, presence of chest pain and other coronary risk factors. A modern exercise ECG system includes a 12-lead ECG diagnostic system with accurate and advanced ECG signal processing capabilities. It also includes appropriate interfaces for treadmills, ergometers as well as motion-tolerant noninvasive blood pressure (NIBP) monitors.

Vectorcardiography. In 12-lead clinical electrocardiography described above, the scalar ECG signals are recorded from 10 electrodes placed on specific sites on the body surface. These tracings show detailed projections of the cardiac vector as a function of time. However, they do not show the underlying cardiac vector. Vectorcardiography is a method in which the cardiac electrical activity along three orthogonal axes (x, y, z) in the principal planes (frontal, transverse, sagittal) are recorded and the activity of any two of the three is displayed on a CRT. This display is a closed loop showing the locus of the tip of the cardiac vector during the evolution of atrioventricular depolarization–repolarization phases in one cardiac cycle (Fig. 4.) These closed loops are known as vectorcardiogram (VCG). Figure 15 shows the 3D representation of the cardiac electrical activity and its projections onto the principal planes recorded as vectorcardiogram. Figure 16 shows the basic principles of vectorcardiography and the VCG.

In contrast to ECG tracings that show the detailed morphology of the electrical activity of the heart in any one single lead direction, the VCG is the simultaneous plot of the same electrical events in two perpendicular (orthogonal) lead directions. This gives a detailed representation of the cardiac vector and produces loop-type patterns on the CRT. The magnitude and orientation of the P, QRS, and T loops are then determined from the VCG. The VCG plots provide a clear indication of the cardiac axis and its deviation from normal. Each VCG exhibits three loops, showing the vector orientation of the P waves, the QRS complex and the T wave during the cardiac cycle. The high amplitude of the QRS complex loop predominates the VCG recordings and sensitivity adjustments need to be made to adequately display the loops resulting from the P and T waves.

The VCG is superior to the clinical 12-lead scalar ECG in recognition of undetected atrial and ventricular hypertrophy and some cases of myocardial infarction and the capability to diagnose multiple infarctions in the presence of bundle branch blocks. For example, in most infarcts the cardiac vector orientation moves away from the area of infarction. The advantage of VCG is that it only requires three orthogonal leads to provide full information on cardiac electrical activity. This translates into simpler algorithms and less computation. However, the full 12-lead

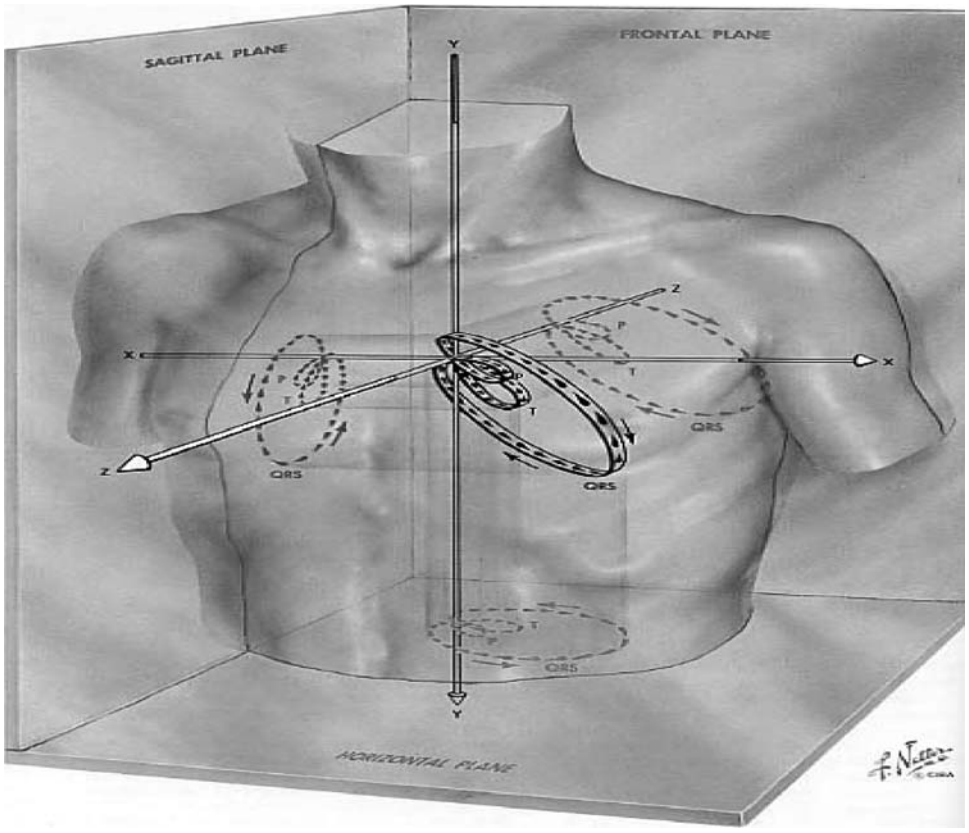


Figure 15. The three-dimensional (3D) representation of the evolution of the cardiac vector for P, QRS, and T loops during one cardiac cycle and their projections onto the three principal planes. (Courtesy of Ref. 3.)

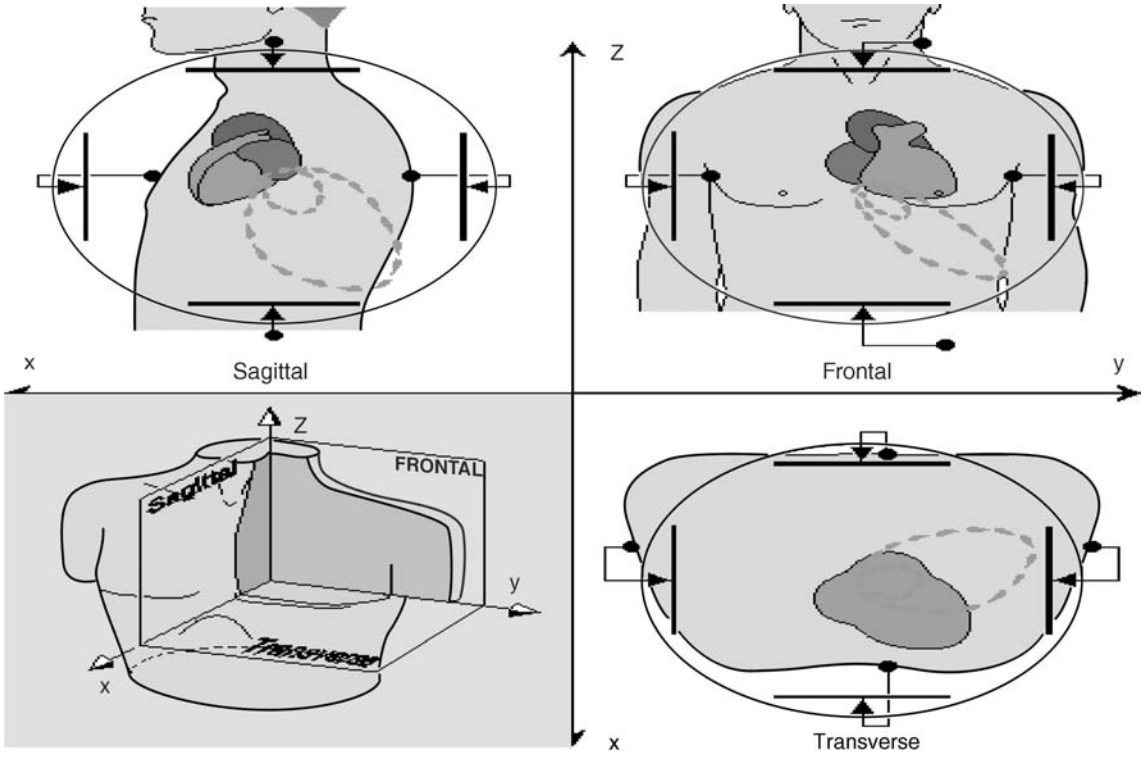


Figure 16. Vectorcardiography based on mutually orthogonal uniform lead fields set up by parallel electrode arrangements on opposite sides of the torso in three principal planes. (Courtesy of Ref. 2.)

scalar ECG provides detailed time plots of the ECG signals and highlights the changes in waveform morphology. Decades of clinical experience has made the 12-lead electrocardiography the standard method of practice in diagnosis of heart disease and favors that method over vecorcardiography. For a comprehensive and detailed description of Vectorcardiography see Refs. 2 and 8.

High Resolution Electrocardiography (Body Surface Potential Mapping). The electrical activity of the heart triggering its pumping action is a spatiotemporally distributed process. As discussed above, electrocardiography provides temporal information on timing of cardiac cycle. It can provide information on chamber enlargement and is an established tool for diagnosis of cardiac arrhythmias. Exercise testing using 12-lead clinical electrocardiography is the most common means of clinical investigation in evaluation of patients with chest pain. Even though clinical 12-lead ECG provides excellent temporal resolution in assisting clinical diagnosis of heart disease, it is limited in providing spatial information on the evolution of the electrical activity during the cardiac cycle.

The 12-lead ECG is now a standard diagnostic tool in clinical practice, but it is limited in detecting some cardiac abnormalities due to it sparse spatial sampling of the cardiac vector. Cardiac electrical imaging or enhanced electrocardiography alleviates this limitation by acquiring more information through deployment of a larger array (hundreds) of electrodes installed in a jacket to record more spatial samples of cardiac electrical activity.

As discussed in the sections above, the heart is a spatially and temporally distributed biosource embedded in an irregularly shaped volume conductor with conductivity inhomogenities. In standard electrocardiography one can measure heart rate and detect arrhythmias, determine the location of cardiac ischemia and infarction, localize some sites of conduction disorders between atria and ventricles or areas of transient ischemia or even detect chamber hypertrophy or congenital heart defect. However, standard ECG is specifically limited in the following: accurate detection of events in the non-anterior regions of the heart, underestimation of the deflections of the ST segment during ischemia if they are weak or nonexistent on the regions of the torso sampled by the ECG leads, inaccuracy of ECG amplitudes in reflecting the extend of physiological alterations due to the spatial integration effects of many simultaneous events, some of which partially cancel so that small defects can result in large changes in the ECG signal or vice versa.

Simply stated, standard ECG is limited in spatial resolution as it is used to describe the complex cardiac fields generated by the electrical activity of a spatiotemporally distributed biosource like the heart. Therefore, the need for enhanced resolution electrocardiography exists. Recent developments in cardiac imaging techniques have been largely driven by this need. A number of active research centers around the world are dedicated to the development of this technology (9–12). As with other imaging methods used in medicine, recent developments in cardiac electrical imaging have greatly benefited from better data

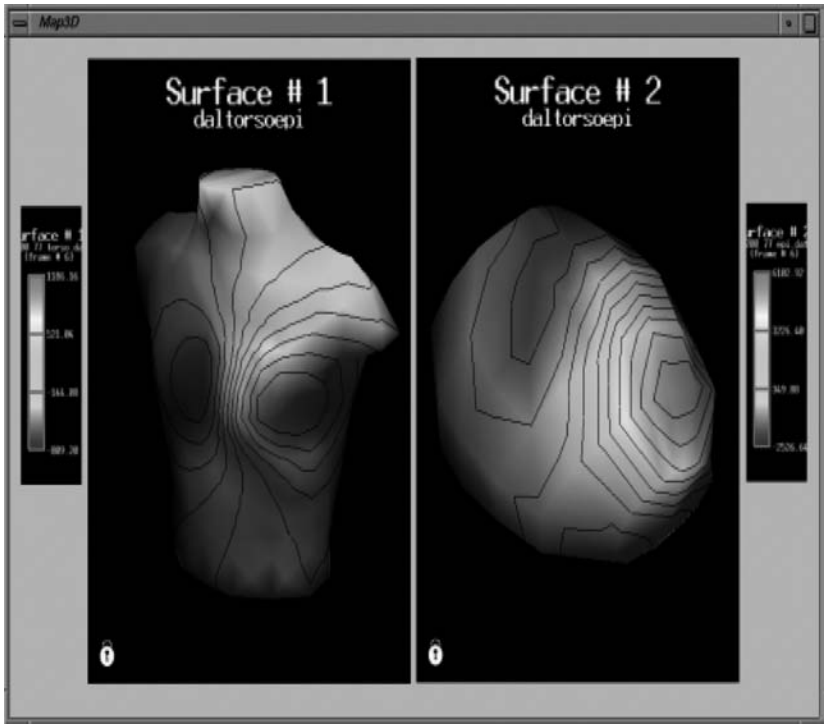


Figure 17. Color-coded BSPMs and computed potential distributions at the surface of the heart during the peak of the QRS complex using 120 ECG electrodes. The surface on the right shows the computed potential distribution on the surface of the heart at the same instant. The relative size of the two surfaces is distorted but their orientations are aligned as in the geometric model used for the inverse calculation. The scaling and isopotential contours are local to teach surface. Scale bars on both sides indicate the voltage (in microvolt) to color mapping. (Courtesy of Dr. Robert MacLeod, Scientific and Imaging Institute, University of Utah, Ref. 9.)

acquisition capabilities and advancements in electronics and computer hardware and software capabilities. Figure 17 shows color-coded BSPMs and computed potential distributions at the surface of the heart during the peak of the QRS complex using 120 ECG electrodes as an example of how advanced computing methods can improve the spatial resolution of electrocardiography.

In summary, a popular modern approach to extend the capabilities of the standard 12-lead ECG system in mapping the electrical activity of the heart is high resolution electrocardiography or body surface potential mapping (BSPM). This approach attempts to provide spatial information by including a larger number of recording electrodes covering the entire torso. The BSPM facilitates the spatial visualization of the cardiac electrical activity. It has been shown that BSPM is superior to 12-lead electrocardiography in detection and localization of acute and remote myocardial infarction (MI). However, the full capability of BSPM to localize acute MI is as yet, not established.

Computers play a central role in fast acquisition and real-time processing of a large number of ECG signal channels, as well as in modeling and visualization of the electrical images of the heart and eventual construction of 3D animated representations of body surface potential maps. These topics are active areas of research and deserve separate chapters in their own right.

Computer-Based Monitoring ECG Systems

All modern diagnostic and monitoring ECG acquisition and analysis systems are equipped with state-of-the-art computer technology. ECG monitoring devices constitute an essential part of the patient monitoring systems. Patient monitoring systems are used to continuously or intermittently measure the vital signs (generally ECG, heart rate, pulse rate, blood pressure, and temperature) and perform automatic analysis on them. Monitoring is carried out at the bedside, or a central station. Bedside monitors are now widely used in intensive care, cardiac care and operating rooms. Computer-based bedside monitors perform ECG analysis and generate alarms if life-threatening arrhythmias occur. The ECG monitoring systems can be networked to share common computing and analysis resources. Telecommunication technologies are now used to transmit ECG signals back and forth between computing facilities. For a detailed discussion of ECG-based monitoring systems see Ambulatory (Holter) Monitoring and Electrocardiographic Monitors.

Computer-Based Heart Beat Detection. Computer-based arrhythmia detectors that provide accurate detection of prevalent heart diseases could greatly help cardiologists in early and efficient diagnosis of cardiac diseases. Such devices now constitute an integral part of computer-based cardiac monitoring systems. The morphologic and rhythmic character of the ECG signal can be interpreted from its patterns, which normally have periodic waveforms such as PQRSTU-waves. Rhythmical patterns deviating from normal ECG (cardiac arrhythmias) have correlation with heart injuries or its malfunction in pumping the blood, such as premature beats, flutter and fibrillation patterns. The most serious pattern is ventricular fibrillation, which may be life threatening due to deficient supply of oxygen by the blood to the vital organs, including the heart itself. Considering other factors influencing the ECG patterns such as medications taken by patients, their physiological and psychological condition as well as their health records, an accurate diagnosis can be made based on cardiac arrhythmia classification.

Accurate determination of the QRS complex and more specifically, reliable detection of the R-wave peak plays a central role in computer-based ECG signal analysis, Holter data analysis and robust calculation of beat-to-beat heart rate variability (13). Fluctuations in the ECG signal baseline drift, motion artifact due to electrode movement and electromyographic (EMG) interference due to muscular activity frequently contaminate the ECG signal. In addition, morphological variations in the ECG waveform and the high degree of heterogeneity in the QRS complexes make it difficult to distinguish them from tall peaked P and T waves.

Many techniques have therefore been developed to improve the performance of QRS detection algorithms. These techniques are mainly based on band pass filtering, differentiation techniques, template matching and others. Recently, it has been shown that the first differential of the ECG signal and its Hilbert Transform can be used to effectively distinguish the R-waves from large, peaked T and P waves with a high degree of accuracy (14). Even though this method provides excellent R peak detection performance, it lacks automatic missing-R-peak correction capability and has not been tested on pediatric ECG data. Manual correction of missing R peaks in the ECG signal is time consuming and tedious and could play a critical role in error-free derivation of HRV signal. Figure 18 shows the block diagram for an Enhanced Hilbert Transform-based (EHT) method with automatic missing-R-peak correction capability for error-free detection of QRS complexes in the ECG signals.

In this algorithm, filtered ECG signals are differentiated first to reduce baseline drift and motion artifacts. This step sets the ECG peaks to zero. Hilbert Transform of the differentiated signal is then calculated and the conjugate of the Hilbert Transform is obtained. With this operation, the zero crossings in the differentiated signal becomes prominent peaks in the Hilbert transformed conjugate of the differentiated signal. A threshold is then selected based upon the normalized RMS value (as it gives a measure of the noise content in the signal) of the Hilbert transformed ECG signal. The time instants for which the signal amplitude is greater than the threshold value are stored in an array. Peak detection is performed on the original signal. The peak values and the time at which they occur are stored in separate arrays. The R–R intervals are then calculated. If an R–R interval is $>130\%$ of the previous R–R interval, then the algorithm is considered to have missed an R-peak from the ECG signal. Hence a correction for the missing beat is made based upon an updated moving average of the previous R–R intervals. The amplitude of the missing peak is estimated as the average of the two previous R-peaks adjacent to it (13).

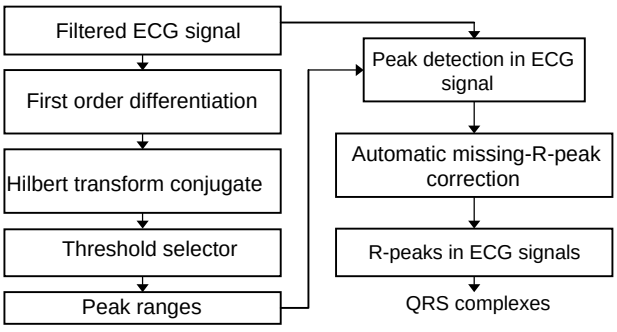


Figure 18. Block diagram for the Enhanced Hilbert Transform-based (EHT) QRS detection algorithm with automatic built-in missing R-peak correction capability. (Courtesy of Ref. 13.)

ECG Signal Telemonitoring

With advancements in bioinstrumentation, computer, and telecommunications technologies now it is feasible to design vital sign telemonitoring systems to acquire, record, display, and transmit physiological signals (e.g., ECG) from the human body to any location. At the same time, it has become more practical and convenient for medical and paramedical personnel to monitor vital signs from any computer connected to the Internet.

Telemonitoring can improve the quality, increase the efficiency, and expand access of the healthcare delivery system to the under-staffed, remote, hard-to-access, or under-privileged areas where there is a paucity of medical practitioners and facilities. It seems reasonable to envision that a telemonitoring facility could significantly impact areas where there are needs for uniform healthcare access such as under-served populations of rural areas, developing countries, space flights, remote military bases, combat zones, and security healthcare facilities. Mobile patient telemonitoring (i.e., emergency medicine), posthospital patient monitoring, home care monitoring, patient education and continuing medical education all will benefit from telemonitoring. Figure 19 shows the graphical user interface of a Web-based vital sign telemonitor accessed from a client site (15).

APPENDIX

The PageWriter Touch Technical Specifications. (Courtesy of Philips Medical Systems, USA.)

ECG Acquisition

R/T (real-time) ECG (12 leads)
AUTO (12 leads)
RHYTHM (up to 12 leads)
DISCLOSE (1–3 leads)

Keyboard

Full alphanumeric

Touch Screen Display

1024 × 768 pixel resolution
30.4 × 22.8 cm (15 in. diagonal) color liquid crystal touch screen display with backlight

Patient Module

Action button allows user to take ECG snapshots from the bedside

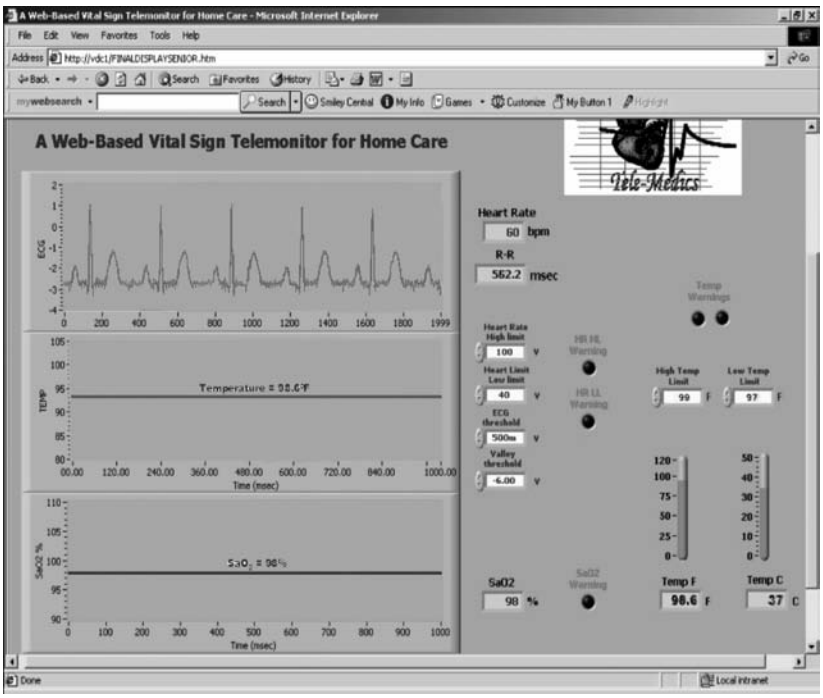


Figure 19. The graphical user interface of a web-based telemonitor. (Courtesy of Ref. 15.)

Signal Processing/Acquisition

Sampling Rate

1000 samples per second per electrode/lead
12 bit A/D conversion provides 5 μV resolution

Auto Frequency Response

0.05–150 Hz, 0.15–150 Hz, 0.5–150 Hz,
0.05–100 Hz, 0.15–100 Hz, 0.5–100 Hz,
0.05–40 Hz, 0.15–40 Hz, 0.5–40 Hz

Rhythm Frequency Response

0.05–150 Hz, 0.15–150 Hz, 0.05–100 Hz,
0.15–100 Hz, 0.05–40 Hz, 0.15–40 Hz

Filters

AC noise
Baseline wander
Artifact

Printer

Printer Resolution

High resolution, digital-array printer using thermal-sensitive paper
200 dpi (voltage axis) by 500 dpi (time axis)

Report Formats

3 × 4 (1R, 3R)
6 × 2
Panoramic 12 (Cabrera)
12 × 1
Rhythm (up to 12 selected leads)
Extended Measurements
One Minute Disclose (1 lead)
Full disclosure (5 min, 1 to s selected leads)

Battery Operation

Capacity

Typically 50 ECGs and copies on a single charge or 40 min of continuous rhythm recording

Recharge

Seven hours in standby mode to >90% capacity (typical)

Network Connections

10/100 Base-T IEEE 802.3 Ethernet Via RJ45 connector (standard)
Optional software required for Proxim Range LAN 27410 CE PC card wireless LAN connection

FAX Capability (optional)

Group 3, Class 1 or 2 fax

Modem (optional for USA & Canada)

V.90, K56 flex, enhanced *V*.34, *V*32 bits, *V*.32, *V*.22 bits and below

Barcode Reader (optional)

Reads Code 39 (standard and full ASCII)

Magnetic Card Stripe Reader (optional)

Reads cards adhering to ISO 7810, 7811-1, -2, -3, -4, and JIS X6301 and X6302

ECG Storage

XML File Format
150 ECGs to internal flash memory
2–3 ECGs typical per 1.4 MB floppy disk
150 ECGs per 128 MB PCMCIA card (optional)

ECG File Formats

XML and XML SVG

Power and Environment

Line Power 100–240 Vac, 50/60 Hz, 150 VA max

Environmental Operating Conditions

15–35 °C (50–104 °F)
15–70% relative humidity (noncondensing)
Up to 4550 m (15,000 ft.) altitude

Environmental Storage Conditions

0–40°C (32–122 °F)
15–80% relative humidity (noncondensing)
Up to 4550 m (15,000 ft) altitude

Cardiograph Dimensions

~45 × 45.8 × 16 cm (17.7 × 18.0 × 6.34 in.)

Cardiograph Weight

~13 kg (28 lb.) including accessories

Patient Interface Module

Remote, microprocessor-controlled module

Safety and Performance

Meets the following requirements for safety and performance:

IEC 60601-1:1988 + A1: 1991 + A2: 1995 General Requirements for Safety including all National Deviations
IEC 60601-1-2: 1993 General Requirements for Safety Electromagnetic Compatibility
IEC 60601-2-25: 1993 + A1: 1999 Safety of Electrocardiographs
IEC 55011: 1998 Radio Frequency disturbance, Limits and Methods of Test
AAMI EC11: 1991 Diagnostic Electrocardiographic Devices
JIST 1202: 1998 Japanese Industrial Standard for Electrocardiographs

ACKNOWLEDGMENT

The excellent support of Ms Mary-Lou Lufkin with the Diagnostic ECG Division at Philips Medical Systems, USA is greatly appreciated.

BIBLIOGRAPHY

Cited References

1. Clark JW. The origin of biopotentials. In: Webster JG, editor. Medical Instrumentation. 3rd ed. New York: John Wiley & Sons; 1998, p 121–182.
2. Malmivuo J, Plonsey R. Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields. New York: Oxford University Press; 1995.
3. Netter FH. The Heart, Vol. 5. The Ciba Collection of Medical Illustrations. Ciba Pharmaceutical Company; 1971.
4. Haykin S. Adaptive Filter Theory. 4th ed. New York: Prentice Hall; 2001.
5. Anonymous. American National Standard for Diagnostic Electrocardiographic Devices. ANSI/AAMI EC 13, New York: American National Standards Institute; 1983.
6. Webster JG, editor. Bioinstrumentation. New York: Wiley; 2004.
7. Available at http://medstat.med.utah.edu/kw/ecg/ACC_AHA.html.
8. Macfarlane PW, Lawrie TDV, editors. Comprehensive Electrocardiology: Theory and Practice in Health and Disease. 1st ed. Vols. 1–3. New York: Pergamon Press; 1989. 1785 p.
9. Available at http://www.sci.utah.edu.
10. Available at http://butler.cc.tut.fi/~malmivuo/bem/index.htm.
11. Available at http://rudylab.wustl.edu.
12. Available at http://www.bioeng.auckland.ac.nz/projects/ei/eimaging.php.
13. Chatlapalli S, et al. Accurate Derivation of Heart Rate Variability Signal for Detection of Sleep Disordered Breathing in Children. Proc 26th Annu Int Conf IEEE EMBS San Francisco, (CA) Sept., 2004.
14. Benitez D, Gaydecki PA, Zaidi A, Fitzpatric AP. The use of Hilbert Transform in ECG signal analysis. Comput Biol Med Sept.2001;31(5):399–406.
15. Mendoza P, et al. A Web-based Vital Sign Telemonitor and Recorder for Telemedicine Applications. Proc IEEE/EMBS, 26th Annu Int Conf, San Francisco (CA). Sept. 1–4, 2004.

See also ARRHYTHMIA ANALYSIS, AUTOMATED; GRAPHIC RECORDERS; PHONOCARDIOGRAPHY.

ELECTROCONVULSIVE THERAPHY

MILTON J. FOUST, JR
MARK S. GEORGE
Medical University of South Carolina
Charleston, South Carolina

INTRODUCTION

Electroconvulsive therapy (ECT) is a technique for the treatment of severe psychiatric disorders, which consists of the deliberate induction of a generalized tonic-clonic seizure by electrical means. Contemporary ECT devices typically deliver bidirectional (alternating current) brief-pulse square-wave stimulation through a pair of electrodes that are applied externally to the patient's scalp. The procedure is now almost always performed under general anesthesia, although, in some unusual situations, such as in developing countries with limited medical resources, it

may be occasionally done without anesthesia (1). As with other convulsive therapies that historically preceded ECT, the goal is to produce a seizure. The presence of seizure activity appears to be essential; stimuli that are below the seizure threshold appear to be clinically ineffective. However, although the production of a seizure appears to be necessary, a seizure alone is not sufficient. Some forms of seizure induction are, in fact, clinically ineffective. A variety of psychiatric and neurological conditions exist that respond favorably to ECT, although the majority of patients treated with ECT have mood disorders, such as unipolar or bipolar depression, particularly when severe or accompanied by psychotic symptoms. Certain other conditions, such as mania, schizoaffective disorder, catatonia, neuroleptic malignant syndrome, Parkinson's disease, and intractable seizures, may respond to ECT as well. Schizophrenia has also been treated with ECT, although the results tend to be less favorable than those obtained in patients with mood disorders. Those patients with schizophrenia who also have a prominent disturbance of mood probably respond best to ECT (2,3). Typically, a series or a course of treatments is prescribed. By convention, ECT treatments are usually given two to three times per week. A course usually consists of around six to eight treatments, which may then be followed by maintenance treatment in the form of either medication, additional ECT given at less frequent intervals, or both. A number of questions still remain regarding the most effective methods for performing ECT, the mechanism of action of ECT, and what role there may be in the future for ECT and other forms of brain stimulation, such as repeated transcranial magnetic stimulation (rTMS), magnetic seizure therapy (MST), and vagus nerve stimulation (VNS).

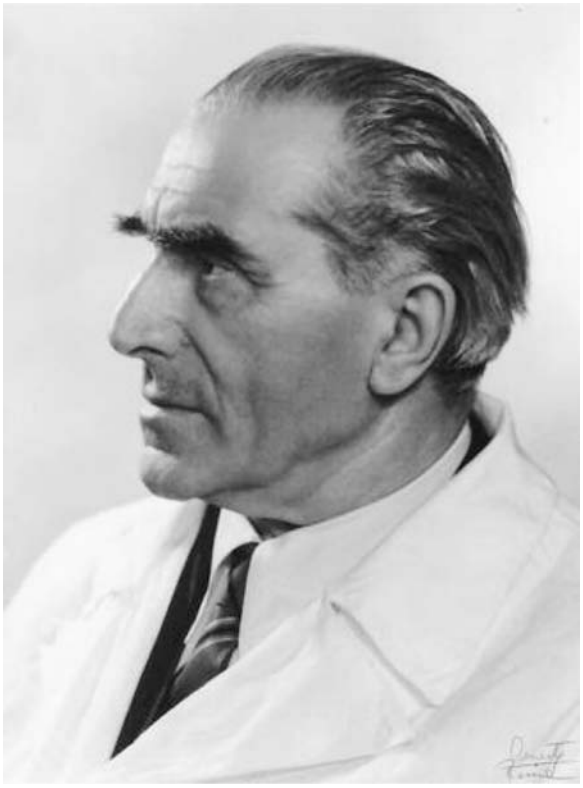


Figure 1. Ugo Cerletti 1877–1963. Reprinted with permission from the American Journal of Psychiatry, (Copyright 1999). American Psychiatric Association.

HISTORY

ECT was first used by the Italian psychiatrists Ugo Cerletti (Fig. 1) and Lucio Bini to treat a disorganized, psychotic man found wandering the streets of Rome in 1938. The results were dramatic, with complete recovery reported (4). The treatment was developed as an alternative to other, higher risk forms of artificial seizure induction, specifically Ladislas von Meduna's convulsive therapy involving the use of stimulants such as camphor, strychnine, and Metrazol (pentylenetetrazol) (4–6). ECT was welcomed due to its effectiveness with otherwise treatment-resistant and profoundly disabled patients. However, the procedure was at risk of being abandoned due to the incidence of fractures (up to 40%) caused by uncontrolled seizure activity. This problem was resolved by the introduction of muscle relaxation (originally in the form of curare, and later with depolarizing muscle relaxants such as succinylcholine) and general anesthesia (7). The use of anesthesia and muscle relaxation was one of the most important innovations in ECT treatment, another being the use of brief-pulse square-wave stimulation in place of sine-wave alternating current. Brief-pulse ECT was found to cause less cognitive impairment compared with sine-wave ECT, as well as less disruption of the EEG (8,9). The routine use of oxygen and monitoring of vital signs, cardiac rhythm, pulse oximetry, electromyography (EMG), and electroencephalography (EEG) have also helped to reduce the risks associated with the procedure.

PRE-ECT EVALUATION

In order to receive ECT, a patient must first be evaluated by a physician (typically a psychiatrist) who is trained and credentialed in the procedure and who agrees that the patient is a suitable candidate, based on psychiatric history and examination, physical condition, and capacity to consent. If a patient is unable to consent, a suitable substitute decision-maker must be identified, usually a family member (in some states, a court-order may be required). The process of evaluation typically consists, at a minimum, of a psychiatric interview and mental status examination, a medical history (including a past medical history, family and social history, and review of systems) and physical examination (including a screening neurological examination and fundoscopic examination to exclude papilledema) (1,2,10). It is necessary to review the patient's current medications, including those that are prescribed for concurrent medical conditions as well as psychiatric medications, and to obtain some information about previous medication trials for the psychiatric condition under consideration. Usually, it is desirable to obtain some basic laboratory studies (such as complete blood count, serum electrolytes, BUN, glucose and creatinine, urinalysis, liver

function tests, and thyroid function tests) both to screen for medical conditions that may cause depressive or psychotic symptoms and to identify conditions of increased ECT or anesthesia risk. Most patients, and especially older patients or patients with a history of cardiovascular disease, should have an electrocardiogram (ECG). The use of brain imaging is controversial; many practitioners prefer some form of pre-ECT brain imaging to identify or exclude the possibility of an intracranial mass, one of the few conditions that may be associated with a high risk of mortality with ECT (11). However, it has been argued that a neurologic examination should be sufficient to screen for this particular risk factor (12).

CONSENT

Prior to undergoing ECT or any other procedure, the patient (or patient's surrogate decision-maker) must demonstrate a satisfactory understanding of the nature of the procedure, its risks, benefits, and alternatives. In some states, the physician may need to petition the court for permission to perform ECT on patients who lack the capacity to consent. The issue of consent is complex, as those patients who are most in need of ECT are generally those who are the most ill, and are often the most vulnerable and impaired. It is possible to have a severe psychiatric illness and yet still retain the capacity to rationally evaluate the necessary issues involved in making a decision as to whether to have ECT, but one can be seriously incapacitated as well. Evaluating the patient's capacity to give informed consent is one of the most important parts of the pre-ECT consultation process, along with establishing the presence or absence of an appropriate indication for treatment and identifying concurrent medical conditions and medications that may increase the risk of treatment. It is part of the consultant's responsibility to educate the patient about what is, for many, an unfamiliar or frightening treatment. Often, much of what a patient or family may understand of ECT consists of disturbing images, such as those presented in films like "One Flew over the Cuckoo's Nest" (13). In most cases, the patient will be able to understand the information presented and engage in a rational decision-making process despite their illness. However, some patients may be able to express a superficial understanding of the facts at hand and yet be impaired in the ability to make rational decisions (14), which may be demonstrated through self-destructive behavior, lack of self-care, or irrational refusal of necessary treatment. In these cases, it becomes necessary to seek a substitute, usually a family member, who can make medical decisions on the patient's behalf. State laws differ regarding the details and circumstances under which another person can make these types of treatment decisions, and it is necessary to become familiar with the particular local laws governing consent.

INDICATIONS

The most common indication for ECT is severe, treatment-resistant depression, either of the unipolar or bipolar type (Table 1). The syndrome of depression is characterized by a sad or depressed mood, as well as disturbances in energy level, sleep, and the capacity to experience pleasure (anhedonia) (15). It may include psychotic symptoms such as delusions (fixed, false beliefs that are held despite evidence to the contrary) or hallucinations. Patients with so-called "unipolar" depression exhibit one or more episodes of depression without episodes of mania. Such patients will usually be formally diagnosed as having a major depressive disorder (MDD). Patients with so-called "bipolar" depression have also suffered from one or more episodes of mania, frequently in a cyclical pattern of alternating mania followed by depression. Those patients who are most severely ill, particularly those who are delusional or catatonic, will typically respond best to ECT. Although usually reserved for those patients who have not had a successful response to one or more medication trials, ECT is an appropriate first treatment when the patient's life is threatened by severe illness in the form of aggressively self-destructive behavior, refusal or inability to eat or drink, or extreme agitation. Mania, a condition characterized by an abnormally elevated or irritable mood, hyperactivity, agitation, impulsivity, and grandiosity, also responds well to ECT. Schizophrenia is often treated with ECT, particularly in some European and Asian countries, but may respond less well, unless mood disturbance is a prominent component of the patient's illness. Neuroleptic malignant syndrome (an antipsychotic drug-induced syndrome that shares many of the characteristics of catatonia) is a less common indication for ECT and may be used when the syndrome persists despite the usual interventions such as discontinuing neuroleptics and treatment with dopamine agonists. Catatonia (which may be an expression of either a mood disorder or schizophrenia) is characterized by mutism and immobility, sometimes with alternating periods of agitation. These patients may respond to treatment with benzodiazepines, but if they do not, ECT is indicated and frequently effective. Recurrent, treatment-refractory seizures may respond to ECT as well, suggesting an anticonvulsant mechanism of action for ECT (16). Patients with Parkinson's disease may improve with ECT, possibly due to the dopaminergic effect of ECT.

Table 1. Indications for ECT

Unipolar or bipolar depression
Mania
Schizoaffective disorder
Schizophreniform disorder
Schizophrenia
Catatonia
Neuroleptic malignant syndrome

Certain conditions exist, such as personality disorders, that do not respond well to ECT or may even reduce the likelihood of successful treatment when they coexist with a more suitable indication, such as a mood disorder (17–19). In some cases, the burden of disability and suffering may be so great (and the risk of serious complications so low) that ECT may reasonably be offered even if the patient's diagnosis is not one of those generally considered a standard indication for the treatment (20).

Table 2. Conditions of Increased Risk with ECT

Increased intra-cranial pressure
Recent myocardial infraction or stroke
Unstable angina
Severe cardiac valvular disease
Severe congestive heart failure
Unstable aneurysms
Severe pulmonary disease
Pheochromocytoma
Retinal detachment
Glaucoma

Table 3. Side Effects and Complications with ECT

Common:
Cognitive side effects
Transient postictal confusion or delirium
Retrograde and anterograde amnesia
Headaches
Muscle soreness
Nausea
Less common or rare:
Death (1/10,000 patients)
Aspiration
Brochospasm or laryngospasm
Cardiovascular
Arrhythmias
Severe hypertension or hypotension
Cardiac ischemia
Myocardial infarction
Cardiac arrest
Neurological
Prolonged or tardive seizures
Nonconvulsive status epilepticus
Stroke
Prolonged apnea due to pseudocholinesterase deficiency
Malignant hyperthermia of anesthesia

CONDITIONS OF INCREASED RISK

As will be discussed later in this article, ECT as currently practiced is a relatively low risk procedure. It can be safely used to treat all patient groups including children and adolescents (21–25), pregnant women (26,27), and the elderly (28). In particular, age alone should not be considered a barrier to treatment; elderly patients are among those who often have the most dramatic and favorable responses to ECT (29). However, certain medical conditions exist that may, to a greater or lesser degree, contribute to an increase in risk of morbidity or mortality with the procedure (Table 2). Most significant among these would be severe or unstable cardiac disease or the presence of a space-occupying lesion (such as a tumor) within the cranial cavity, resulting in increased intracranial pressure. Very small tumors without a visible mass effect on computerized tomography (CT) or magnetic resonance imaging (MRI) do not appear to pose a high risk with ECT (30). Detecting such masses and making the distinction between low and high risk lesions may help to support a rationale for pre-ECT brain imaging as a screening tool (11).

COMPLICATIONS

Serious complications with ECT are rare (Table 3). The risk of death has been estimated at 4 per 100,000 treatments (31). The risk of other potentially life-threatening complications, such as myocardial infraction and stroke, is also very low, although the risk of cardiac arrhythmias appears to be higher in persons with pre-existing cardiac disease (32,33). The introduction of general anesthesia and muscle relaxation has almost eliminated the risk of fractures with ECT. Both retrograde amnesia and anterograde amnesia are common, but it is unusual for cognitive impairment to be severe or prolonged. Minor side effects such as headaches, muscle aches, and nausea frequently occur; these side effects are usually transient and easily managed with symptomatic treatment.

The amnesia or memory loss that occurs with ECT typically takes two forms: loss of memory for past or previously learned information (retrograde amnesia) as well as difficulty in learning new information (anterograde amnesia). The retrograde amnesia associated with ECT tends to be greater for "public" or "impersonal" knowledge about the world than for autobiographical or personal memories. Memories for remote events also tend to be better preserved than that for more recent events. Bilateral ECT appears to produce greater and more persistent memory deficits than right-unilateral ECT (34). ECT-induced anterograde amnesia is typically greatest immediately following treatment and tends to rapidly resolve in the weeks following the last treatment of a series. Recovery also typically occurs over time with retrograde amnesia, although patients may notice some persistent gaps in memory for past events. Although anecdotal, it may be reassuring to patients to be made aware of the stories of psychologists and physicians who have had ECT, benefited, and resumed their professional activities (35). Uncommon exceptions to these rules exist, however. A few patients complain of severe and persistent problems with memory that cause them much distress (36). No satisfactory explanation exists for this phenomenon of severe ECT-induced memory loss (37). No reliable evidence exists that ECT causes damage or injury to the nervous system.

MEDICATIONS AND ECT

Experts in the past have recommended stopping antidepressants prior to ECT (38), although a recent study now suggests that tricyclic antidepressants (TCAs) and selective serotonin reuptake inhibitors (SSRIs) may be safe in combination with ECT (39). Evidence exists that certain antipsychotic medications (such as haloperidol, risperidone, and clozapine) may have beneficial effects in combination with ECT (40–44). Many drugs that were originally developed and used as anticonvulsants (such as carbamazepine, valproic acid, and lamotrigine) are frequently used as either mood stabilizers or as adjunctive agents in antidepressant regimens. As these drugs, by definition, can be expected to inhibit seizure activity, they are generally tapered and discontinued prior to beginning

ECT. An exception would be those patients for whom these drugs are prescribed for a concurrent seizure disorder. In these cases, anticonvulsant drugs should usually be continued in order to minimize the risk of uncontrolled seizure activity between treatments. Lithium, a commonly used mood stabilizer, has been associated with prolonged delirium following ECT. In general, it should be avoided, but, in special circumstances and with careful monitoring, its use in combination with ECT may be justified (45,46). Both chlorpromazine and reserpine, an antipsychotic and antihypertensive agent that acts through the depletion of neuronal dopamine, have been associated with severe hypotension and death when combined with ECT (47).

Certain drugs that are prescribed for concurrent medical conditions (and not primarily for psychiatric conditions) help to reduce the risks associated with ECT and anesthesia. Patients who have gastro-esophageal reflux disease (GERD) should be treated with a suitable medication (such as a histamine-2 receptor blocker or proton pump inhibitor) in the morning prior to ECT to reduce the risk of reflux and aspiration. Patients who may be especially prone to aspiration can be treated with intravenous metoclopramide to accelerate gastric emptying. Pregnant women may be given sodium citrate/citric acid solution by mouth. Patients with known hypertension should receive their usual antihypertensive regimen prior to ECT; an exception would be made for diuretics, which should be delayed until after ECT to avoid bladder filling before or during the procedure. Some patients may have an exaggerated hypertensive response to the outpouring of catecholamines that occurs with seizure response to ECT, which can usually be controlled with intravenous beta-adrenergic antagonists such as esmolol or labetalol. A patient with a pheochromocytoma (catecholamine-secreting tumor) is at especially high risk of severe hypertension (48). Such a patient may require more aggressive measures (such as arterial blood-pressure monitoring and intravenous sodium nitroprusside) in order to maintain satisfactory blood pressure control (49). Patients with pulmonary disease should have their pulmonary status assessed prior to treatment and should receive their usual medications, including inhaled beta-agonists or steroids. As ECT and the resulting seizure produces a transient increase in intraocular pressure (50), patients with retinal detachment may be at risk of further eye injury and should have this condition treated prior to ECT. Similarly, patients with glaucoma should have their intraocular pressures controlled with suitable medications prior to ECT (51).

ANESTHESIA

Methohexital was previously a very popular anesthetic drug for ECT, but manufacturing problems made it essentially unavailable for several years, forcing changes in ECT anesthesia practice (52). Anesthesia for ECT may be induced with an intravenous injection of other short-acting anesthetic agents such as propofol and etomidate. Etomidate may be substituted for methohexital or propofol in an effort to produce seizures of longer duration (53) or to stimulate seizures in those uncommon situations of no seizure activity despite maximum stimulus and bilateral electrode placement. Once anesthesia is induced (typically within seconds following the injection), a muscle relaxant is injected, typically succinylcholine, but a nondepolarizing muscle relaxant such as mivacurium may be used when there are coexisting medical conditions that increase the risk of exaggerated hyperkalemic response with succinylcholine, such as burns, renal failure, or neurologic disease, or if a history of malignant hyperthermia exists (54,55). The ECT stimulus may elicit a vagal (parasympathetic) response, which can lead to bradycadia, transient heart block, or even asystole, which has been explained as the result of forced expiration against a closed glottis during the stimulus, or it may be a direct effect of the stimulus on the central nervous system. Bradycardia and asystole have been observed in the postictal period as well (56). An anticholinergic compound such as glycopyrrolate or atropine may be injected proximate to the anesthetic to reduce the bradyarrhythmias that may occur with ECT (57). If the stimulus is successful in producing a seizure, it results in an outpouring of catecholamines and a sympathetic response with resulting tachycardia and hypertension as noted above, which is usually transient and without clinical significance, but in some patients, especially those with pre-existing hypertension or cardiovascular disease, it may be desirable to limit this response using an antihypertensive agent (54,55,58).

Patients are unable to breathe without assistance when anesthetized and fully relaxed; ventilation and oxygenation are provided by means of positive-pressure ventilation with 100% oxygen through a bag and mask. Endotracheal intubation is rarely required. In some patients (such as pregnant women), the risk of reflux may be higher and intubation may be the preferred option for airway management (27). Some patients may have abnormalities of the face and upper airway that interfere with mask ventilation. Other options for airway management including intubation, laryngeal mask airway (59,60), or even tracheostomy may be considered.

MONITORING

The type of medical monitoring that is used during ECT includes ECG blood pressure, pulse oximetry, EMG, and EEG. Modern ECT devices typically include the capacity to monitor and record ECG, EMG, and EEG; commonly, a paired right and left fronto-mastoid EEG placement is used. Two-channel EEG monitoring is helpful both to ensure that the seizure generalizes to both hemispheres and to evaluate the degree of inter-hemispheric EEG symmetry or coherence, as well as postictal EEG suppression, all factors that may predict the therapeutic efficacy of the seizure (61). Two EMG electrodes are placed on a foot, usually the right, which is isolated from the rest of the patient's circulation with a blood pressure cuff acting as a tourniquet, which minimizes the effect of the muscle relaxant and permits the observation and recording of motor seizure activity in the foot, even with complete paralysis otherwise.

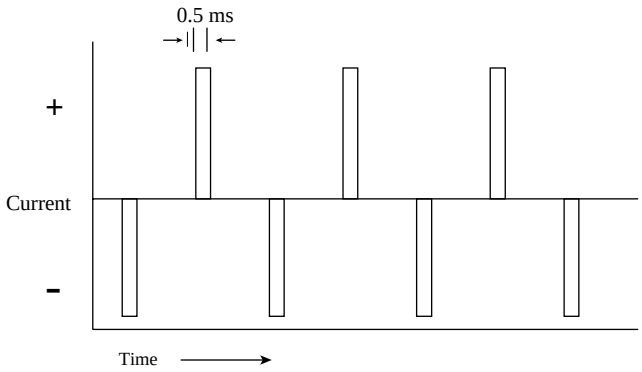


Figure 2. Brief-pulse square-wave ECT stimulus.

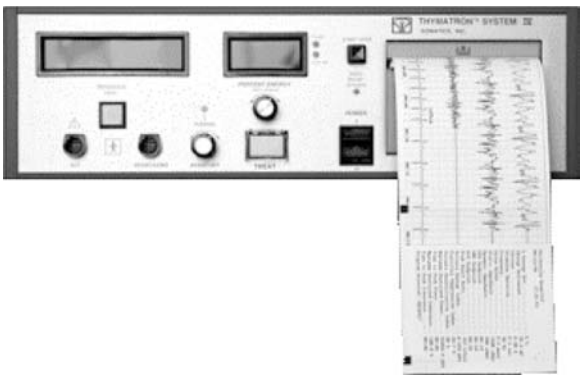


Figure 3. Somatics Thymatron system IV ECT device (Image provided courtesy of Somatics, LLC).

ECT STIMULUS

As noted previously, the ECT stimulus itself typically consists of a series of brief electrical pulses. Early ECT devices provided stimulation with sine-wave alternating current; essentially the same type of current that is distributed through public utilities for use in household appliances and lighting. Most modern ECT devices now provide a stimulus that consists of brief pulses with a bidirectional square-wave pattern (Fig. 2). (One exception is the Ectron series 5A, manufactured by Ectron, Ltd, Letchworth, Hertfordshire, England. This device delivers unidirectional pulses.) The American Psychiatric Association has recommended the use of "constant-current" devices; so-called because a relatively constant (rather than continuously varying) current is maintained for the duration of the pulse. The advantage of using a constant-current rather than constant-voltage or constant-energy device is that the clinician is able to deliver a predetermined quantity of charge by varying the time interval of exposure to the current (62).

At 100% of available output, a typical ECT device such as the Thymatron System IV (Somatics, LLC, Lake Bluff, IL) can provide approximately 504 mC of charge (Table 4) (64), which will be delivered as brief pulses of 0.5 ms each at a frequency of 70 Hz for a total stimulus duration of 8 s. Members of the Thymatron (Fig. 3) series (DX, DGX, and System IV) and similar devices such as those of the Spectrum series (4000M, 4000Q, 5000M, and 5000Q manufactured by MECTA, Lake Oswego, OR) are calibrated in such a way that the user can select incremental quantities of charge. The device labels and manuals will usually refer to these as increments of "percent energy," although the energy required to deliver the charge is actually dependent on the particular impedance of the patient being treated. The user (clinician) may have the option of individually adjusting such parameters as "percent energy" (charge), pulse frequency, pulsewidth, and total duration of pulse train. Both voltage and energy will be altered by the device to adjust for variations in impedance. The average impedance for a human receiving ECT is approximately 220 Ω, and the energy required to deliver 504 mC is 99.4 J (63).

Table 4. ECT Device Specifications (Thymatron System IV)

Current:	0.9 A (fixed)
Frequency:	10 to 140 Hz
Pulsewidth:	0.25 to 1.5 ms
Duration:	0.14 to 7.99 s
Output:	approx. 25 to 504 mC (5 to 99.4 J at 220 Ω)
	1008 mC (188.8 J at 220 Ω) with double-dose option
Input:	120 V AC, 60 Hz
Recording:	4 channel (2 EEG, EMG, ECG)

Over an 8 s total stimulus duration, the average power for that interval is 99.4 J/8 s = 12.4 W, although the power during each discrete electrical pulse is obviously much greater. Assuming a pulse frequency of 70 Hz (140 pulses/s with bidirectional current), delivered over a total stimulus duration (stimulus train) of 8 s, with a pulsewidth of 0.5 ms (or 0.0005 s), then the power during each discrete electrical pulse of this stimulus would be:

$$99.4\,\text{J}/(140\,\text{pulses/s}^{*}8\,\text{s}^{*}0.0005\,\text{s/pulse})$$
$$= 99.4\,\text{J}/0.56\,\text{s} = 177.5\,\text{W}$$

The stimulus is transmitted to the head of the patient (and, indirectly, to the brain) through externally applied scalp electrodes. These electrodes may either be stainless-steel discs held in place with an elastic band or self-adhesive flexible conductive pads. The stimulus electrodes can be applied in several configurations, including bilateral (also called bifronto-temporal), right unilateral, and bifrontal. In bilateral ECT, the electrodes are placed approximately 3 cm above the midpoint of a line between the canthus of the eye and tragus of the ear. With right unilateral ECT, the right electrode is placed in the conventional position for bilateral ECT, with the second electrode just to the right of the vertex of the skull. Bifrontal electrode placement is the newest of these various methods of performing ECT. Bifrontal placement is 5 cm above the angle of the orbit of the eye, which is usually found near the outer edge of the eyebrow. Comparative studies suggest that bilateral ECT may be more effective at lower energy levels, although right unilateral and bifrontal ECT may result in less cognitive impairment (64–68).

We consider the patient and device (including electrodes and cables) to be a circuit, and, for purposes of simplification, impedance in this circuit is usually treated as more or less equivalent to resistance. However, it should be recognized that impedance in the circuits being discussed has both resistive and reactive components (62). These reactive components are capacitance and inductance. Both inductance effects and capacitance in tissue are assumed to be low, but the treatment cables and electrode–skin interface may make a more significant contribution, altering both the amplitude and frequency of the stimulus. Complicating matters further, it should also be noted that the circuit impedance varies with the intensity of the stimulus. ECT devices typically provide a "static impedance" reading prior to treatment, which can vary widely, but is usually around 2000 Ω. The "dynamic impedance" can be measured during the stimulus and is substantially less, an average of about 200 Ω. The static impedance is measured with a very small test stimulus (insufficient to cause a seizure or even be detected by the patient). This static impedance is used primarily to test the integrity of the circuit. If the impedance is greater than 3000 Ω, it is assumed that either a break in the circuit occured or inadequate contact exists at the skin-electrode interface. Similarly, an excessively low impedance suggests shunting through a very low impedance channel (short-circuit) such as saline, conductive gel, or skin between closely applied electrodes. Interestingly, the seizure threshold (quantity of charge required to stimulate a seizure) is typically lower for unilateral than for bilateral ECT, despite the lower interelectrode distance with potential for lower impedance and greater shunting though scalp tissues with unilateral placement, which is thought to be a result of differing patterns of charge density, with maximum charge density in the frontal region for bilateral electrode placement and maximum charge density for unilateral placement in the region of the motor strip, an area of lower intrinsic seizure threshold (62).

SEIZURE RESPONSE

A seizure is characterized neurophysiologically by the paroxysmal synchronous discharge of populations of neurons. It is recognized clinically as abnormal behavior associated with abnormal neuronal activity and may present in a variety of different forms such as simple partial seizures, complex partial seizures, or generalized tonic-clonic seizures (69). During ECT, the patient is anesthetized and relaxed, so many of the clinical characteristics are altered. However, the observable characteristics of the ECT seizure are consistent with the generalized tonic-clonic type of seizure. The ECT stimulus predictably results in a pattern of spike and slow-wave activity that is distinctive and recognizable on EEG (70). The seizure activity rapidly spreads throughout both hemispheres and usually lasts for less than 60 s. The immediate goal of ECT is to produce such a seizure; stimuli that do not result in seizure activity appear to lack therapeutic benefit. However, it is possible to produce seizures, specifically with right unilateral ECT at or only slightly above the seizure threshold, that have a relatively weak therapeutic effect (64). Most experts recommend that right unilateral ECT be given with a stimulus that is at least five times the seizure threshold. Threshold may either be determined by titration (i.e., by giving a series of increasing stimuli), or the stimulus may be given at maximum energy. If bilateral or bifrontal electrode placement is used, energies just above the seizure threshold may be sufficient, and seizure threshold can be estimated based on the patient's age (71).

Certain characteristics of the seizure activity may be associated with therapeutic efficacy, including coherence or symmetry between hemispheres of high amplitude slow waves on EEG and marked postictal suppression following the end of the seizure (61). Adjustments can be made in technique to try to improve response based on these characteristics and other seizure characteristics, including overall seizure duration.

MECHANISM OF ACTION OF ECT

The precise mechanisms of action of ECT are not well understood. A number of biochemical and physiological changes exist that have been detected following both ECT in humans and electroconvulsive shock (ECS) in animals. Some of these parallel those of antidepressant drugs, but others do not. In vivo studies of long-term ECS (and repeated tricyclic antidepressant drug administration) show increases in postsynaptic 5-hydroxytryptamine type 1a (5-HT-1a) receptor sensitivity. Long-term ECS increases 5-HT-2a receptors, but antidepressant administration in animals decreases 5-HT-2a receptors. Both antidepressant treatment and ECS reduce beta-adrenergic receptors (72). Evidence for increased dopaminergic activity with ECS and ECT exists, probably as a result of increased dopamine synthesis (73). Proton magnetic resonance spectroscopy measurements following ECT have demonstrated increases in concentrations of cortical gamma-amino butyric acid (GABA), an inhibitory neurotransmitter (74). This observation, as well as the finding that increasing seizure threshold during a course of ECT is associated with clinical response, has led to the hypothesis that a linked anticonvulsant and antidepressant response to ECT exists (61).

Magnetic resonance imaging (MRI) scans obtained before and after ECT have failed to demonstrate any structural changes in the brain. Positron emission tomography (PET) and single-photon emission computerized tomography (SPECT) scans of the brain during the ECT seizure show marked increases in global and regional metabolic activity and blood flow (75). However, PET scans taken several hours or days following ECT have shown a marked reduction in absolute and prefrontal activity (76). The majority of functional imaging studies with depressed patients have shown pretreatment deficits in similar regions (77). Paradoxically, those subjects with the greatest post-ECT reduction in the left prefrontal cortex show the greatest clinical response (78,79). One possible explanation for this may be that specific areas of the brain exist that are hypermetabolic in the depressed state and are suppressed by the effects of ECT, although, as mentioned

above, most scans show resting prefrontal hypoactivity in depression.

At the level of gene expression, ECS appears to induce brain-derived neurotrophic factor (BDNF) and its receptor, protein tyrosine kinase B (TrkB). Using our current understanding of the effects of ECS and neurotransmitter signal transduction, a model can be constructed that traces the effects of ECS/ECT-induced neuronal depolarization through pathways of norepinephrine (NE), 5-HT and glutamate release, monoamine receptors and ionotropic glutamate receptor binding by these neurotransmitters, cyclic adenosine monophosphate (cAMP) and other second-messengers coupled to these receptors, and protein kinases stimulated by these second-messengers. Protein kinase A (PKA), protein kinase C (PKC), and calcium/calmodulin-dependent protein kinase (CaMK) phosphorylate and activate cAMP response element binding protein (CREB). CREB is a transcription factor for BDNF. BDNF then induces neuronal sprouting and may have other beneficial effects that reverse stress-induced neuronal atrophy (80) and, presumably, depression.

FUTURE DIRECTIONS FOR ELECTROMAGNETIC BRAIN STIMULATION

A number of unanswered questions regarding ECT still exist, including such questions as:

- What are the most effective methods of maintenance therapy (medications, ECT, or both) following an acute course of treatment?
- Which medications (including antidepressants, antipsychotics, and mood stabilizers) can be successfully and safely combined with ECT?
- What is the best electrode placement from a treatment efficacy and side-effect point of view?
- How can the ECT stimulus be modified to make it more therapeutically effective and diminish side effects (especially cognitive effects)?
- What is the actual mechanism (or mechanisms) by which ECT exerts its antidepressant, antipsychotic, and other beneficial effects?

ECT itself is no longer the only form of electromagnetic brain stimulation. Both rTMS (often referred to simply as TMS) and VNS have shown promise for the treatment of mood disorders (79,81). As it does not involve the production of seizures and therefore requires no anesthesia, TMS is a particularly attractive alternative to ECT, especially for mood disorders resistant to treatment with medication. MST involves magnetic stimulation at frequencies designed to provoke seizures. This technique obviously requires general anesthesia and muscle relaxation just as ECT does. The hope is that MST may be as effective as ECT, but with less cognitive impairment due to the more focal nature of the stimulus.

ECT will remain an important option for severe and treatment-refractory psychiatric illness for the foreseeable future. An improved understanding of the pathophysiology of psychiatric illness as well as the mechanism of action of electromagnetic brain stimulation will lead to further refinements in technique that will make ECT and related therapies safer, more effective, and more acceptable to patients and their families. The availability of alternative methods of brain stimulation will provide a wider range of choices and will create opportunities for combining therapies in ways that will be more compatible with individual patients' needs. Although changing social and political conditions may affect its image and public acceptance, it is unlikely that ECT will disappear or be replaced for some time to come.

ACKNOWLEDGMENTS

The authors would like to express their appreciation to Carol Burns, RN, ECT Program Coordinator of the Medical University of South Carolina Department of Psychiatry and Behavioral Sciences and to Drs. Melinda Bailey and Gary Haynes of the Medical University of South Carolina Department of Anesthesiology for their helpful advice during the preparation of this article.

BIBLIOGRAPHY

Cited References

1. Mudur G. Indian group seeks ban on use of electroconvulsive therapy without anesthesia. Br Med J 2002;324(6):806.
2. American Psychiatric Association. The Practice of Electroconvulsive Therapy: Recommendations for Treatment, Training and Privileging, A Task Force Report of the American Psychiatric Association. 2nd ed. Washington (DC): American Psychiatric Press, Inc.; 2000.
3. Kellner CH, Pritchett JT, Beale MD, Coffey CE. Handbook of ECT. Washington (DC): American Psychiatric Press, Inc.; 2001.
4. Fink M. Meduna and the origins of convulsive therapy. Am J Psychiatry 1984;141(9):1034–1041.
5. Fink M. Convulsive therapy: A review of the first 55 years. J Affect Disord 2001;63:1–15.
6. Fink M. Induced seizures as psychiatric therapy: Ladislas Meduna's contributions in modern neuroscience. J ECT 2004; 20:133–136.
7. Bennett AE. Curare: A preventive of traumatic complications in convulsive shock therapy. Convulsive Ther 1997;13:93–107. Originally published in the Am J Psychiatry 1941; 97:1040–1060.
8. Weiner RD, Rogers HJ, Davidson JR, Kahn EM. Effects of electroconvulsive therapy upon brain electrical activity. Ann NY Acad Sci 1986;462:270–281.
9. Weiner RD, Rogers HJ, Davidson JR, Squire LR. Effects of stimulus parameters on cognitive side effects. Ann NY Acad Sci 1986;462:315–325.
10. Klapheke MM. Electroconvulsive therapy consultation: An update. Convulsive Ther 1997;13:227–241.
11. Coffey CE. The role of structural brain imaging in ECT. Psychopharmacol Bull 1994;30:477–483.
12. Kellner CH. The CT scan (or MRI) before ECT: A wonderful test has been overused. Convulsive Ther 1996;12:79–80.
13. McDonald A, Walter G. The portrayal of ECT in American movies. J ECT 2001;17(4):264–274.
14. Appelbaum PS, Grisso T. Assessing patients' capacities to consent to treatment. N Engl J Med 1988;319(25):1635–1638.

15. Diagnostic and Statistical Manual of Mental Disorders, 4th ed. Text Revision (DSM-IV-TR®), Washington (DC): American Psychiatric Press, Inc.; 2000.
16. Griesemer DA, Kellner CA, Beale MD, Smith GM. Electroconvulsive therapy for treatment of intractable seizures: Initial findings in two children. Neurology 1997;49.
17. DeBattista C, Mueller K. Is electroconvulsive therapy effective for the depressed patient with comorbid borderline personality disorder? J ECT 2001;17(2):91–98.
18. Feske U, Mulsant BH, Pilkonis PA, Soloff P, Dolata D, Sackeim HA, Haskett RF. Clinical outcome of ECT in patients with major depression and comorbid borderline personality disorder. Am J Psychiatry 2004;161(11):2073–2080.
19. Sareen J, Enns MW, Guertin J. The impact of clinically diagnoses personality disorders on acute and one-year outcomes of electroconvulsive therapy. J ECT 2000;16(1):43–51.
20. Fink M. The broad clinical activity of ECT should not be ignored. J ECT 2001;17(4):233–235.
21. Ghaziuddin N, Kutcher SP, Knapp P, Bernet W, Arnold V, Beitchman J, Benson RS, Bukstein O, Kinlan J, McClellan J, Rue D, Shaw JA, Stock S, Kroeger K. Practice parameter for use of electroconvulsive therapy with adolescents. J Am Acad Child Adolesc Psychiatry 2004;43(12):1521–1539.
22. Cohen D, Paillere-Martinot M-L, Basquin M. Use of electroconvulsive therapy in adolescents. Convulsive Ther 1997;13: 25–31.
23. Moise FN, Petrides G. Case study: Electroconvulsive therapy in adolescents. J Am Acad Child Adolesc Psychiatry 1996;35: 312–318.
24. Schneekloth TD, Rummans TA, Logan K. Electroconvulsive therapy in adolescents. Convulsive Ther 1993;9:158–166.
25. Bertagnoli MW, Borchardt CM. A review of ECT for children and adolescents. J Am Acad Child Adolesc Psychiatry 1990; 29:302–307.
26. Yonkers KA, Wisner KL, Stowe Z, Leibenluft E, Cohen L, Miller L, Manber R, Viguera A, Suppes T, Altshuler L. Management of bipolar disorder during pregnancy and the postpartum period. Am J Psychiatry 2004;161:608–620.
27. Walker R, Swartz CM. Electroconvulsive therapy during high-risk pregnancy. Gen Hosp Psychiatry 1994;16:348–353.
28. Kelly KG, Zisselman M. Update on electroconvulsive therapy (ECT) in older adults. J Am Geriatr Soc 2000;48(5):560–566.
29. O'Connor MK, Knapp R, Husain M, Rummans TA, Petrides G, Snyder GK, Bernstein H, Rush J, Fink M, Kellner C. The influence of age on the response of major depression to electroconvulsive therapy: A C.O.R.E. report. Am J Geriatr Psychiatry 2001;9:382–390.
30. McKinney PA, Beale MD, Kellner CH. Electroconvulsive therapy in a patient with cerebellar meningioma. J ECT 1998; 14(1):49–52.
31. Abrams R. The mortality rate with ECT. Convulsive Ther 1997;13:125–127.
32. Nuttall GA, Bowersox MR, Douglass SB, McDonald J, Rasmussen LJ, Decker PA, Oliver WC, Rasmussen KG. Morbidity and mortality in the use of electroconvulsive therapy. J ECT 2004;20(4):237–241.
33. Zielinski RJ, Roose SP, Devanand DP, Woodring S, Sackeim HA. Cardiovascular complications in depressed patients with cardiac disease. Am J Psychiatry 1993;150:904–909.
34. Lisanby SH, Maddox JH, Prudic J, Davanand DP, Sackeim HA. The effects of electroconvulsive therapy on memory of autobiographical and public events. Arch Gen Psychiatry 2000;57:581–590.
35. Fink M. A new appreciation of ECT. Psychiatric Times 2004; 21(4):
36. Donahue AB. Electroconvulsive therapy and memory loss. J ECT 2000;16(2):133–143.
37. Abrams R. Does brief-pulse ECT cause persistent or permanent memory impairment? J ECT 2002;18:71–73.
38. American Psychiatric Association. The Practice of Electroconvulsive Therapy: Recommendations for Treatment, Training and Privileging, A Task Force Report of the American Psychiatric Association. Washington, DC: American Psychiatric Press, Inc.; 1990.
39. Lauritzen L, Odgaard K, Clemmesen L, Lunde M, Öhrström J, Black C, Bech P. Relapse prevention by means of paroxetine in ECT-treated patients with major depression: A comparison with imipramine and placebo in medium-term continuation therapy. Acta Psychiatrica Scandanavica 1996;94:241–251.
40. Hirose S, Ashby CR, Mills MJ. Effectiveness of ECT combined with risperidone against aggression in schizophrenia. J ECT 2001;17:22–26.
41. Sjatovic M, Meltzer HY. The effect of short-term electroconvulsive treatment plus neuroleptics in treatment-resistant schizophrenia and schizoaffective disorder. J ECT 1993;9: 167–175.
42. Chanpattana W, Chakrabhand MLS, Kongsakon R, Techakasem P, Buppanharun W. Short-term effect of combined ECT and neuroleptic therapy in treatment-resistant schizophrenia. J ECT 1999;15:129–139.
43. Tang WK, Ungvari GS. Efficacy of electroconvulsive therapy combined with antipsychotic medication in treatment-resistant schizophrenia: A prospective, open trial. J ECT 2002; 18:90–94.
44. Frankenburg FR, Suppes T, McLean P. Combined clozapine and electroconvulsive therapy. Convulsive Ther 1993;9: 176–180.
45. Kellner CH, Nixon DW, Bernstein HJ. ECT-drug interactions: A review. Psychopharmacol Bull 1991;27(4):
46. Mukherjee S. Combined ECT and lithium therapy. Convulsive Ther 1993;9(4):274–284.
47. Klapheke MM. Combining ECT and antipsychotic agents: Benefits and risks. Convulsive Ther 1993;9:241–255.
48. Carr ME, Woods JW. Electroconvulsive therapy in a patient with unsuspected pheochromocytoma. Southern Med J 1985; 78(5):613–615.
49. Weiner R. Electroconvulsive therapy in the medical and neurologic patient. In: Stoudemire A, Fogel BS, editors. Psychiatric Care of the Medical Patient. New York: Oxford University Press; 1993.
50. Edwards RM, Stoudemire A, Vela MA, Morris R. Intraocular changes in nonglaucomatous patients undergoing electroconvulsive therapy. Convulsive Ther 1990;6(3):209–213.
51. Good MS, Dolenc TJ, Rasmussen KG. Electroconvulsive therapy in a patient with glaucoma. J ECT 2004;20(1):48–49.
52. Kellner C. Lessons from the methohexital shortage. J ECT 2003;19(3):127–128.
53. Stadtland C, Erhurth A, Ruta U, Michael N. A switch from propofol to etomidate during an ECT course increases EEG and motor seizure duration. J ECT 2002;18(1):22–25.
54. Folk JW, Kellner CH, Beale MD, Conroy JM, Duc TA. Anesthesia for electroconvulsive therapy: A review. J ECT 2000;16:157–170.
55. Ding Z, White PF. Anesthesia for electroconvulsive therapy. Anesthesia Analgesia 2002;94:1351–1364.
56. Bhat SK, Acosta D, Swartz C. Postictal systole during ECT. J ECT 2002;18(2):103–106.

57. Rasmussen KG, Jarvis MR, Zorumski CF, Ruwitch J, Best AM. Low-dose atropine in electroconvulsive therapy. J ECT 1999;15(3):213–221.
58. McCall WV. Antihypertensive medicines and ECT. Convulsive Ther 1993;9:317–325.
59. Nishihara F, Ohkawa M, Hiraoka H, Yuki N, Saito S. Benefits of the laryngeal mask for airway management during electroconvulsive therapy. J ECT 2003;19:211–216.
60. Brown NI, Mack PF, Mitera DM, Dhar P. Use of the ProSeal™ laryngeal mask airway in a pregnant patient with a difficult airway during electroconvulsive therapy. Br J Anaesthesia 2003;91:752–754.
61. Sackeim HA. The anticonvulsant hypothesis of the mechanism of action of ECT: Current status. J ECT 1999;15(1):5–26.
62. Sackeim HA, Long J, Luber B, Moeller J, Prohovnik I, Davanand DP, Nobler MS. Physical properties and quantification of the ECT stimulus: I. Basic principles. Convulsive Ther 1994;10(2):93–123.
63. Abrams R, Swartz CM. Thymatron® System IV Instruction Manual. 8th ed. Somatics, LLC; 2003.
64. Sackeim HA, Prudic J, Devanand DP, Kiersky JE, Fitzsimons L, Moody BJ, McElhiney MC, Coleman EA, Settembrino JM. Effects of stimulus intensity and electrode placement on the efficacy and cognitive effects of electroconvulsive therapy. N Engl J Med 1993;328:839–846.
65. Delva NJ, Brunet D, Hawken ER, Kesteven RM, Lawson JS, Lywood DW, Rodenburg M, Waldron JJ. Electrical dose and seizure threshold: Relations to clinical outcome and cognitive effects in bifrontal, bitemporal, and right unilateral ECT. J ECT 2000;16:361–369.
66. Bailine SH, Rifkin A, Kayne E, Selzer JA, Vital-Herne J, Blieka M, Pollack S. Comparison of bifrontal and bitemporal ECT for major depression. Am J Psychiatry 2000;157:121–123.
67. Letemendia FJJ, Delva NJ, Rodeburg M, Lawson JS, Inglis J, Waldron JJ, Lywood DW. Therapeutic advantage of bifrontal electrode placement in ECT. Psycholog Med 1993;23:349–360.
68. Lawson JS, Inglis J, Delva NJ, Rodenburg M, Waldron JJ, Letemendia FJJ. Electrode placement in ECT: Cognitive effects. Psycholog Med 1990;20:335–344.
69. Benbadis S. Epileptic seizures and epileptic syndromes. Neurolog Clin 2001;19(2):251–270.
70. Beyer JL, Weiner RD, Glenn MD. Electroconvulsive Therapy: A Programmed Text. 2nd ed. Washington (DC): American Psychiatric Press; 1985.
71. Petrides G, Fink M. The "half-age" stimulation strategy for ECT dosing. Convulsive Ther 1996;12:138–146.
72. Newman ME, Gur E, Shapira B, Lerer B. Neurochemical mechanisms of action of ECT: Evidence from in vivo studies. J ECT 1998;14(3):153–171.
73. Mann JJ. Neurobiological correlates of the antidepressant action of electroconvulsive therapy. J ECT 1998;14(3):172–180.
74. Sanacora G, Mason GF, Rothman DL, Hyder F, Ciarcia JJ, Ostroff RB, Berman RM, Krystal JH. Increased cortical GABA concentrations in depressed patients receiving ECT. Am J Psychiatry 2003;160(3):577–579.
75. Nobler MS, Teneback CC, Nahas Z, Bohning DE, Shastri A, Kozel FA, Goerge MS. Structural and functional neuroimaging of electroconvulsive therapy and transcranial magnetic stimulation. Depression Anxiety 2000;12(3):144–156.
76. Nobler MS, Oquendo MA, Kegeles LS, Malone KM, Campbell CC, Sackeim HA, Mann JJ. Decreased regional brain metabolism after ECT. Am J Psychiatry 2001;158(2):305–308.
77. Drevets WC. Functional neuroimaging studies of depression: The anatomy of melancholia. Annu Rev Med 1998;49:341–361.
78. Nobler MS, Sackeim HA, Prohovnik I, Moeller JR, Mukherjee S, Scnur DB, Prudic J, Devanand DP. Regional cerebral blood flow in mood disorders, III. Treatment and clinical response. Arch Gen Psychiatry 1994;51:884–897.
79. George MS, Nahas Z, Li X, Kozel FA, Anderson B, Yamanaka K, Chae J-H, Foust MJ. Novel treatments of mood disorders based on brain circuitry (ECT, MST, TMS, VNS, DBS). Sem Clin Neuropsychiatry 2002;7:293–304.
80. Duman RS, Vaidya VA. Molecular and cellular actions of chronic electroconvulsive seizures. J ECT 1998;14(3):181–193.
81. George MS, Nahas Z, Kozel FA, Li X, Denslow S, Yamanaka K, Mishory A, Foust MJ, Bohning DE. Mechanisms and state of the art of transcranial magnetic stimulation. J ECT 2002;18: 170–181.

Reading List

Abrams R. Electroconvulsive Therapy. 4th ed. New York: Oxford University Press; 2002.
Sackeim HA, Devanand DP, Nobler MS. Electroconvulsive therapy. In: Bloom FE, Kupfer DJ, editors. Psychopharmacology: The Fourth Generation of Progress. 4th ed. New York: Raven Press; 1995.
Rasmussen KG, Sampson SM, Rummans TA. Electroconvulsive therapy and newer modalities for the treatment of medication-refractory mental illness. Mayo Clin Proc 2002;77:552–556.
Dolenc TJ, Barnes RD, Hayes DL, Ramussen KG. Electroconvulsive therapy in patients with cardiac pacemakers and implantable cardioverter defibrillators. PACE 2004;27:1257–1263.
Electroconvulsive Therapy. NIH Consensus Statement, June 10–12, 1985, National Institutes of Health, Bethesda, MD Online. Available at http://odp.od.nih.gov/consensus/cons/051/051_statement.htm.
Information about ECT. New York State Office of Mental Health. Online. Available at http://www.omh.state.ny.us/omhweb/ect/index.htm. and http://www.omh.state.ny.us/omhweb/spansite/ect_sp.htm (Spanish version).

See also ELECTROENCEPHALOGRAPHY; REHABILITATION, COMPUTERS IN COGNITIVE.

ELECTRODES.

See BIOELECTRODES; CO_2 ELECTRODES.

ELECTROENCEPHALOGRAPHY

DAVID SHERMAN
DIRK WALTERSPACHER
The Johns Hopkins University
Baltimore, Maryland

INTRODUCTION

The following section will give an overview of the electroencephalogram (EEG), its origin, and its validity for diagnosis in clinical use. Since Berger (1) demonstrated in 1929 that the activity of the brain can be measured using external electrodes placed directly on the intact skull, the EEG has been used to study functional states of the brain. Although the EEG signal is the most common indicator for brain injuries and functional brain disturbances, the complicated underlying process, creating the signal, is still not well understood.

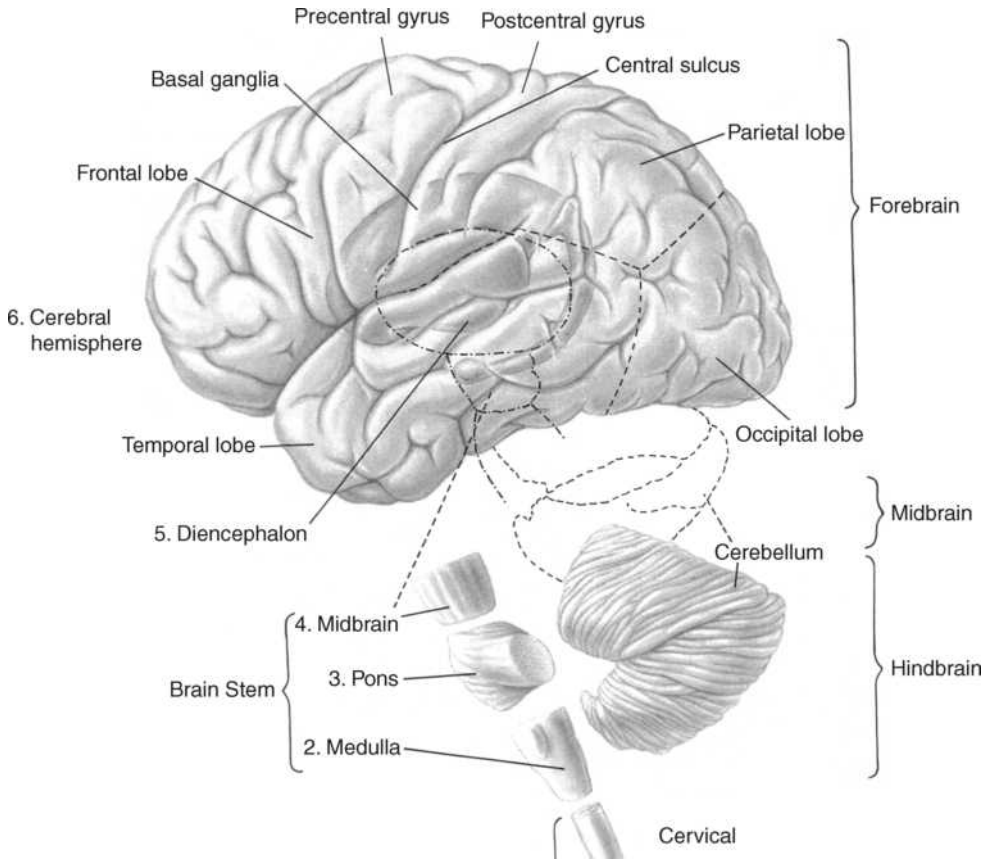


Figure 1. The major portions of the human cerebrum called lobes. Areas external to the cerebrum include the midbrain areas such as the diencephalon and the hindbrain areas such as the cerebellum, medulla, and pons. (Adapted from Ref. 5.)

The section is organized as follows. The biophysical basis of the origin of the EEG signal is described first, followed by EEG recordings and classification. Afterward, the validity and scientific basis for using the EEG signal as a tool for studying brain function and dysfunction is presented. Finally, logistical and technical considerations as they have to be made in measuring and analyzing biomedical signals are mentioned.

ORIGIN OF EEG

Valid clinical interpretation of the electroencephalogram ultimately rests on an understanding of the basic electrochemical and electrophysical processes through which the patterns are generated and the intimate nature of the brain's functional organization, at rest and in action. Therefore, the succeeding parts of the following discussion deal with the gross organization of the cortex, which is generally assumed to be the origin of brain electrical activity that is recorded from the surface of the head (2–4), the different kinds of electrical activity and resulting potential fields developed by cortical cells. Figure 1 shows the general organization of the human brain.

Organization of the Cerebral Cortex

Even though different regions of the cortex have different cytoarchitectures and each region has its own morphological patterns, aspects of intrinsic organization of the cortex are general (6,7). Most of the cortical cells are arranged in the form of columns, in which the neurons are distributed with the main axes of the dendritic trees parallel to each other and perpendicular to the cortical surface. This radial orientation is an important condition for the appearance of powerful dipoles. Figure 2 shows the schematic architecture of a cortical column. It can be observed that the cortex, and within any given column, consist of different layers. These layers are places of specialized cell structures and within places of different functions and different behaviors in electrical response. An area of very high activity is, for example, layer IV, which neurons function to distribute information locally to neurons located in the more superficial (or deeper) layers. Neurons in the superficial layers receive information from other regions of the cortex. Neurons in layers II, III, V, and VI serve to output the information from the cortex to deeper structures of the brain.

Activity of a Single Pyramidal Neuron

Pyramidal neurons constitute the largest and the most prevalent cells in the cerebral cortex. Large populations of these pyramidal neurons can be found in layers IV and V of the cortex. EEG potentials recorded from electrodes placed on the scalp represent the collective summation of changes in the extracellular potentials of pyramidal cells (2–4).

The pyramidal cell membrane is never completely at rest because it is continually influenced by activity arising in other neurons with which it has synaptic connections (8).

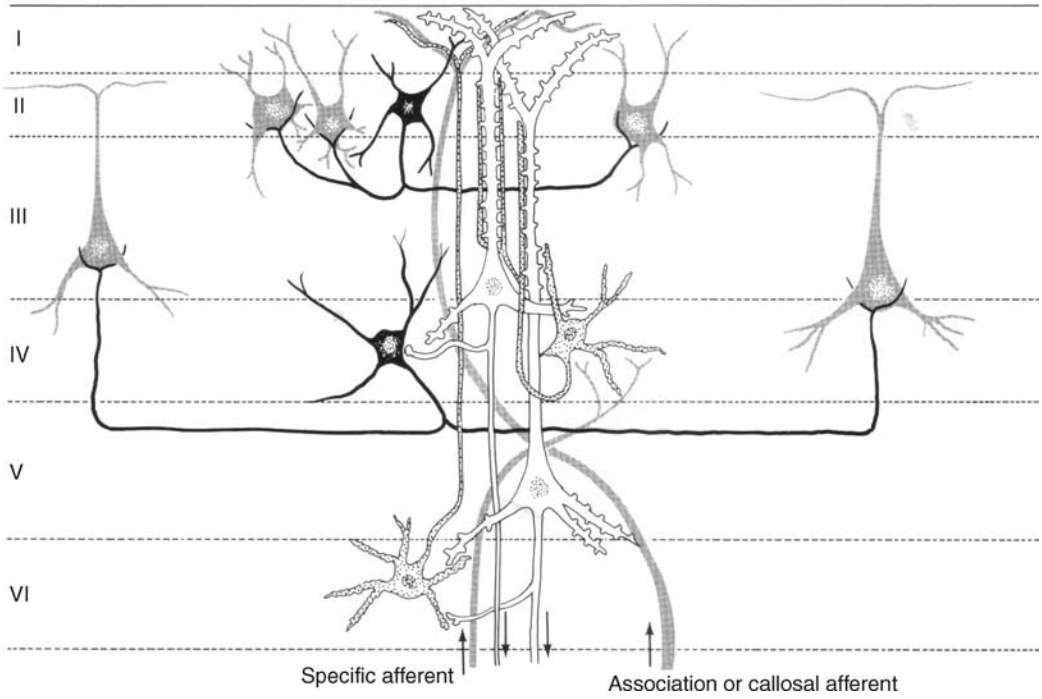


Figure 2. Schematic of the different layers of the cerebral cortex. Pyramidal cells that are mainly in layers III and V are mainly responsible for the generation of the EEG. (Adapted from Ref. 5.)

Such synaptic connections may be excitatory or inhibitory and the respective transmitter changes the permeability of the membrane for K^+ (and/or Cl^-), which results in a flow of current (for details, please see Ref. 8).

The flow of current in response to an excitatory postsynaptic potential at the site on the apical dendrite of a cortical pyramidal neuron is shown in Fig. 3. The excitatory postsynaptic potential (EPSP) is associated with an inward current at the postsynaptic membrane carried by positive ions and an outward current along the large expanse of the extra-synaptic membrane. For simplicity, only one path of outward current is illustrated through the soma membrane. This current flow in the extracellular space causes the generation of a small potential due to extracellular resistance (shown by R in Fig. 3).

As an approximation it is possible to estimate the extracellular field potential as a function of the transmembrane potential (9)

$$\phi_{\mathrm{e}} = \frac{a^2 \sigma_{\mathrm{i}}}{4\sigma_{\mathrm{e}}} \int \frac{\partial^2 v_{\mathrm{m}} / \partial x^2}{r} dx \qquad (1)$$

where ϕ_e is the extracellular potential, a is the radius of axon or dendrites, v_{m} is the transmembrane potential, σ_{i}, is the intracellular conductance, and σ_{e} is the extracellular conductance. For a derivation of the above mentioned equation, please see Ref. 9.

Although these extracellular potentials individually are small, their sum becomes significant when added over many of cells. This is because the pyramidal neurons are more or less simultaneously activated by this way of synaptic connections and the longitudinal components of their extracellular currents will add, whereas their transversal components will tend to cancel out.

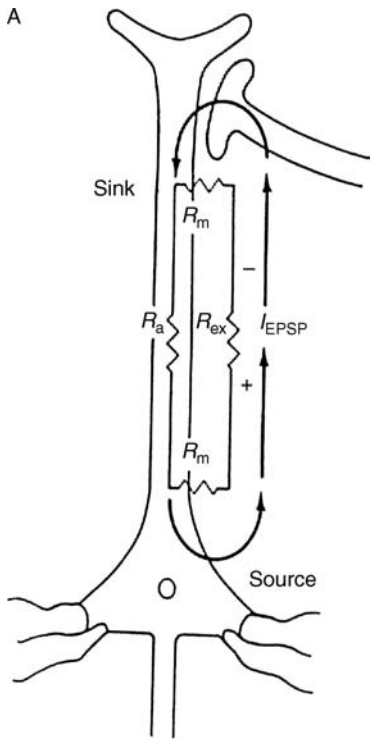


Figure 3. The current flow within a large pyramidal cell. Ionic flow is established to enable charge balance. (Adapted from Ref. 2.)

Generation of EEG Potentials

The bulk of the gross potentials recorded from the surface of the scalp results from the extracellular current flow associated with summated postsynaptic potentials in synchronously activated vertically oriented cortical pyramidal neurons. The exact configuration of the gross potential is related in a complex fashion to the site and the polarity of the postsynaptic potentials. Considering two cortical pyramidal neurons (shown in Fig. 3), the potential measured by a microelectrode at the location P is given by (4)

$$\Phi_p(t) = \frac{1}{4\pi\sigma}\left[\frac{I_a}{R_1}\cos(2\pi f_a t + \alpha_a) - \frac{I_a}{R_2}\cos(2\pi f_a t + \alpha_a) + \frac{I_b}{R_3}\cos(2\pi f_b t + \alpha_b) - \frac{I_b}{R_4}\cos(2\pi f_b t + \alpha_b)\right] + \text{Similar contributions from other dipoles} \quad (2)$$

where I_a and I_b are the peak magnitudes of current for each dipole with phases α_a and α_b, respectively, and is the conductivity of the medium. "Similar contribution from other dipoles" refers to the contribution of dipoles other than the two shown, most significantly from the dipoles physically close to the electrode. Dipoles that are farther from the recording electrode but in synchrony ($\alpha_a = \alpha_b = \ldots = \alpha$), and aligned in parallel, contribute an average potential proportional to the number of synchronous dipoles $|\overline{\Phi}_p| \approx m$. Dipoles that are either randomly oriented or with a random phase distribution contribute an average potential proportional to the square root of their number $\overline{\Phi}_p| \cong \sqrt{m}$. Thus, the potential measured by a microelectrode can be expressed by the following approximate relation (4):

$$\overline{\Phi}_p| \sim \frac{1}{8\pi\sigma}\left[\sum_{i=1}^{l}\frac{I_i}{R_i} + m\frac{I_s d}{R_s^2} + \sqrt{n}\frac{I_a d}{R_a^2}\right] \quad (3)$$

Here the subscripts i s, and a refer to local, remote synchronous, and remote asynchronous dipoles. I_s And I_a are the effective currents for remote synchronous and remote asynchronous sources, which may be less than the total current, depending on the orientation of the sources with respect to the electrode. Also l, m, and n are the numbers of local, remote synchronous, and remote asynchronous sources, which are located at average distances R_i, R_s, and R_a respectively. Note that a microelectrode like the scalp electrode used for EEG recordings measures a field averaged over a volume large enough to contain perhaps 10^7–10^9 neurons.

Contribution of Other Sources

A decrease in the membrane potential to a critical level of approximately 10 mV less than its resting state (depolarization) initiates a process that is manifested by the action potential (10). Although it might seem that action potentials traveling in the cortical neurons are a source of EEG, they contribute little to surface cortical records, because they usually occur asynchronously in time in large numbers of axons, which run in many directions relative to the surface. Other reasons are that the piece of membrane that is depolarized by an action potential at any instant of time is small in comparison with the portion of membrane activated by an EPSP and that action potentials are of short durations (1–2 ms) in comparison with the duration of EPSPs or IPSPs (10–250 ms). Thus, their net influence on potential at the surface is negligible. An exception occurs in the case of a response evoked by the simultaneous stimulation of a cortical input (11), which is generally called a compound action potential.

Other cells in the cortex like the glial cells are unlikely to contribute substantially to surface records because of their irregular geometric organization, such that produced fields of current flow sum to a small value when viewed from a relatively great distance on the surface. There is also activity in deep subcortical areas, but the resulting potentials are too much attenuated at the surface to be recordable.

The discussed principles show that the surface-recorded EEG can be observed as the result of many active elements, where the postsynaptic potentials from cortical pyramidal cells are the dominant source.

Volume Conduction of EEG

Recording from the scalp has the disadvantage that there are various layers with different conductivities between the electrode and the area of cortical potential under the electrode. Therefore, potentials recorded at the scalp are not only influenced by the above mentioned patterns, but also by regions with different conductivities. Layers lying around the brain are such regions. These include the cerebrospinal fluid (CSF), the skull, and the scalp. These layers account, at least in part, for the attenuation of EEG signals measured at the surface of the scalp, as compared with those recorded with a microelectrode at the underlying cortical surface or with a grid of electrodes directly attached to the cortex (ECoG). These shells surrounding the brain account for an attenuation factor of 10 to 20 (12). This attenuation mainly affects the high-frequency–low-voltage component (the frequency range above 40 Hz), which has been shown to carry important information about the functional state of the brain, but it is almost totally suppressed at the surface.

EEG SIGNAL

The EEG signal consists of spontaneous potential fluctuations that also appear without a sensory input. It seems to be a stochastic signal, but it is also composed of quasi-sinusoidal rhythms. The synchrony of cerebral rhythms may occur from pacemaker centers in deeper cortical layers like the thalamus or in subcortical regions, acting through diffuse synaptic linkages, reverberatory circuits incorporating axonal pathways with extensive ramifications, or electrical coupling of neuronal elements (13). The range of amplitudes is normally from 10 to 150 μV, when recorded from electrodes attached to the scalp. The EEG signal consists of a clinical relevant frequency range of 0.5–50 Hz (10).

EEG Categories

Categorizing EEG signals into waves of a certain frequency range has been used since the discovery of the electrical

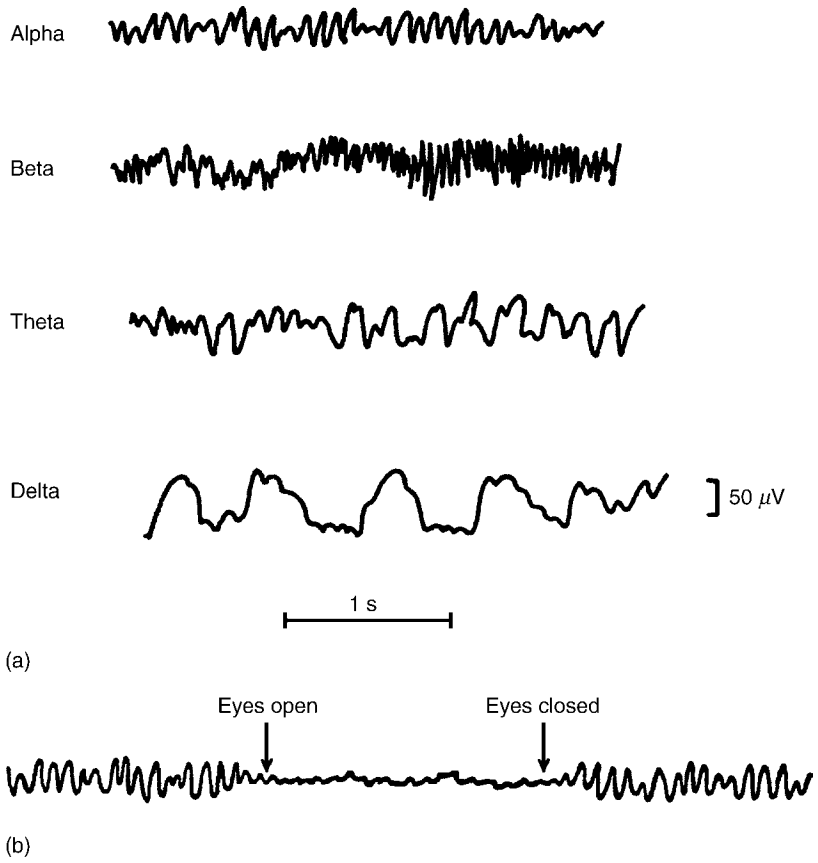


Figure 4. (a) Examples of EEG from different frequency bands. (b) The phenomenon of alpha desynchronization. (From Ref. 14).

activity of the brain. Therefore, these different frequency bands have been the most common feature in EEG analysis. Although this feature contains a lot of useful information as presented below, its use is not with criticism. It can be observed that there is some physiological and statistical evidence for the independence of these bands, but the exact boundaries vary between people and change with the behavioral state of each person (age, mental state, etc.). In particular, between human EEG and EEG signals recorded from different species of animals one can find different EEG patterns and frequency ranges. Nevertheless, the different frequency bands for human EEG are described below, because of their importance as an EEG feature. Most of the patterns observed in human EEG could be classified into one of the following bands or ranges:

Delta	below 3.5 Hz (usually 0.1–3.5 Hz)
Theta	4–7.5 Hz
Alpha	8–13 Hz
Beta	usually 14–22 Hz

A typical plot of these frequency bands is shown in Fig. 4, and the different bands are described below.

Delta Waves. The appearance of delta waves is normal in neonatal and infants' EEGs and during sleep stages in adult EEGs. When slow activity such as the delta band appears by itself, it indicates cerebral injury in the waking adult EEG. Dominance of delta waves in animals that have had subcortical transections producing a functional separation of cerebral cortex from deeper brain regions suggests that these waves originate solely within the cortex, independent of any activity in the deeper brain regions.

Theta Waves. This frequency band was included in the delta range until Walter and Dovey (15) felt that an intermediate band should be established. The term "theta" was chosen to allude to its presumed thalamic origin.

Theta frequencies play a dominant role in infancy and childhood. The normal EEG of a waking adult contains only a small amount of theta frequencies, mostly ovserved in states of drowsiness and sleep. Larger contingents of theta activity in the waking adult are abnormal and are caused by various forms of pathology.

Alpha Waves. These rhythmic waves are clearly a manifestation of the posterior half of the head and are usually found over occipital and parietal regions. These waves are best observed under conditions of awakeness but during physical relaxation and relative mental inactivity. The posterior alpha rhythm can be temporarily blocked by mental activities, or afferent stimuli such as influx of light while eye opening (Fig. 4b). This alpha blocking response was discovered by Berger in 1929 (1). Mainly thalamocortical feedback loops are believed to play a significant role in the generation of the alpha rhythm (16).

Beta Waves. Beta activity is found in almost every healthy adult and is encountered chiefly over the frontal and central regions of the cortex. The voltage is much lower than in alpha activity (seldom exceeds 30 μV). Beta

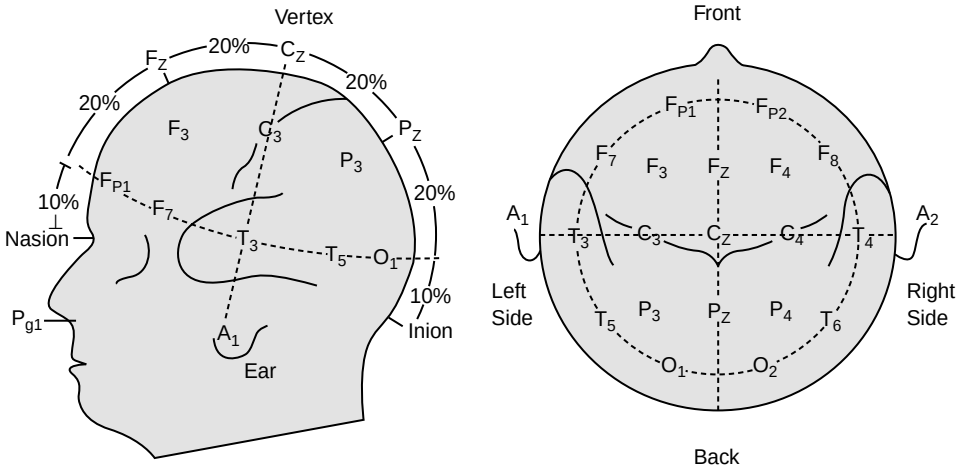


Figure 5. (a)and (b) The side and top view of the layout of a standardized 10–20 electrode system. Adapted from Ref. 14.

activity shows considerable increase in quantity and voltage after the administration of barbiturates, some nonbarbituric sedatives, and minor tranquilizers. It also appears during intense mental activity and tension.

Clinical EEG

To obtain EEG recordings, there are several standardized systems for electrode placement on the skull. The most common are those of the standard 10–20 system of the International EEG Federation, which uses 30 electrodes placed on four landmarks of the skull as observed in Fig. 5.

It is now possible to obtain unipolar (or monopolar) and bipolar derivations from these electrodes. Using a bipolar derivation, one channel is connected between a pair of electrodes and the resultant difference in the potential between these two electrodes is recorded. Therefore, bipolar derivations give an indication of the potential gradient between two cerebral areas. Unipolar (monopolar) derivations can either be obtained by recording the potential-difference between the "active" electrodes and one "indifferent" electrode, placed elsewhere on the head (ear, nose), or with respect to an average reference, by connecting all other leads through equal-valued resistances (e.g., 1 MΩ) to a common point (17). The advantages of unipolar derivations are that the amplitude of each deflection is proportional to the magnitude of the potential change that causes it and the demonstration of small time differences between the occurrence of a widespread discharge at several electrodes. Small, nonpolarizable, disk Ag-AgCl electrodes are used together with an electrode paste. Recorded potentials are amplified using a high gain, differential, capacitivly coupled amplifier. The output signals are displayed on a chart recorder or a monitor screen. For more details about unipolar or bipolar derivations and EEG-amplifiers, please see (Ref. 11).

SCIENTIFIC BASIS FOR EEG MONITORING

The scientific basis for using EEG as a tool for studying brain function and dysfunction rests on the following four neurobiologic qualities of EEG.

Link With Cerebral Metabolism

The above presented discussion on the origin of potential-differences, recorded from the brain, shows that the EEG can be observed as a result of the synaptic and cellular activity of cortical pyramidal neurons. These neuronal and synaptic activities are directly linked to cerebral metabolism. Cerebral metabolic activity in turn depends on multiple factors including enzyme synthesis, substrate phosphorylation, axonal transport, and adenosine triphosphate (ATP) production from mitochondrial and glytolytic pathways (18). Thus, the EEG is a composite phenomenon reflecting complicated intracellular, intraneuronal, and neuro-glial influences. Although this multifaceted system makes it obvious that a selection of any single mechanism underlying the electrocortical manifestations may not be possible, the EEG is still a highly sensitive indicator of cerebral function (19).

Sensitivity to Most Common Causes of Cerebral Injury

The most common causes of cerebral injury are hypoxia and ischemia. It can be observed that hypoxia-ischemia causes a severe neuronal dropout in the cortical layers 3 and 5, leading to well-known hisyophatologic patterns of laminar necrosis. As pyramidal neurons that occupy the cortical layers are the main source for EEG, this loss of neuronal activity changes the cortical potentials and therefore makes EEG very sensitive to these common insults.

Correlation With Cerebral Topography

The standardized systems for electrode placement (Jung, international 10–20 system, etc.) establish a consistent relationship between electrode placement and underlying cerebral topography (20). Therefore, changes in EEG recorded from these electrodes of different areas of the skull reflect a consistent topographical relationship with underlying cerebral structures and allows useful inferences about disease localization from abnormalities in EEG detected at the scalp.

Ability to Detect Dysfunctions at a Reversible Stage

Heuser and Guggenberger (21) showed in 1985 that EEG deteriorates before the disruption of neuronal membrane and before significant reduction of cellular ATP levels. Siesjo and Wieloch (22) demonstrated in 1985 that during cerebral ischemia, changes in EEG correlate with elevated tissue lactate levels while ATP levels remain normal. Astrup (23) showed that a reduction in cerebral blood flow (CBF) affects EEG much before it causes neuronal death. These and several other reports make it clear that EEG offers the ability to detect injury at a reversible stage.

EEG also allows a prediction of recovery after brain dysfunctions like cerebral ischemia after cardiac arrest (24). Various studies in this report show that EEG recordings at several stages during recovery allow the prediction of whether the patient has a favorable outcome. Goel (25) showed that parameters obtained from EEG recordings may serve as an indicator for the outcome after hypoxic-asphyxic encephalophaty (HAE). These attributes make the EEG a very attractive neurologic observational tool.

LOGISTICAL AND TECHNICAL CONSIDERATIONS

Although the last sections have shown that EEG signal detection may serve as an important indicator for detection of neurological status and disorders, its clinical use and acceptance under a neurological care regime is limited. This is due to the complicated nature of the EEG signal and because of the difficulties regarding the interpretation of the signals. Some challenges of EEG analysis are as follows.

Artifacts

Recordings of physiological signals, especially from the surface of the body, have the problem that they are superimposed or distorted by artifacts. EEG signals are especially prone to artifact distortions due to their weak character. Therefore, a knowledge about the possible sources of distortion is necessary to estimate the signal-to-noise ratio (SNR). These artifacts are mostly generated from various kinds of sources. The sources of artifacts can be devided into two mayor groups: the subject-generated artifacts and the artifacts generated by the equipment. Subject-generated artifacts include EMG artifacts like body movement, muscle contraction of the neck, chewing, swallowing, coughing, involuntary movements (like myoclonic jerks, palotal myoclonus, nystagmus, asymmetric oculomotor paralysis, and decerebrate or decorticate posturing), and eye movements. Scalp edema can produce artifactual reductions in amplitude regionally or over an entire hemisphere. Pulse and EKG could also contribute as artifacts in EEG.

Artifacts generated by the equipment include ventilator artifacts that typically appear as slow wave like activity, vibrations of electrical circuitry around the subject, and power line interference. By taking the appropriate methods, a lot of these artifacts can be prevented [for more details, see Mayer-Waarden (10)].

Another method, to eliminate both artifacts generated by the equipment and the subject, is to use a differential amplifier for the recording between two electrodes. The assumption here is that the transmission time of artifacts between two electrodes can be neglected, and therefore, the artifacts at both electrodes are in-phase. The signal to be recorded is assumed to have a time delay from one to another electrode, and taking the difference therefore eliminates the artifact but keeps the signals' nature.

Inter-User Variability

Interpreting physiological signals is difficult, even for specialists, because of their subject-specific nature. The complicated nature of EEG signals makes it even more difficult, and data generated by different EEG analysis methods (especially techniques like feature analysis, power spectral analysis, etc.) may be interpreted in different ways by different analysts. An analysis of inter-user-variability of clinical EEG interpretation was presented by Williams et al. (26), which showed that even EEG data interpretation by EEG analysts could be different. Therefore, more standardized and objective methods of EEG analysis are extremely desirable.

Inter-Individual Variability

As mentioned, the consistency of the human EEG is influenced by many parameters and makes EEG unique for a certain person and for a specific point-in-time. Intrinsic parameters are the age and the mental state of the subject (degree of wakefulness, level of vigilance), the region of the brain, hereditary factors, and influences on the brain (injuries, functional disturbances, diseases, stimuli, chemical influences, drugs, etc.). To detect deviations from "normal" EEG, it would be necessary to compare this "abnormal" EEG with the "normal" EEG as a reference. Therefore, attempts have been made to obtain normative EEG for each of the classes discussed above (i.e, normative EEG for various age groups, normative EEG under the influence of varying amounts of different drugs, etc.), but these databases of normative data are still not sufficient to cover the variety of situations possible in real-life recordings. On the other hand, these normative data vary too much for considering them as one person's "normal" EEG.

Labor-Intensive and Storage Problems

For patient monitoring in the operating room or for chronic monitoring tasks that are necessary for cases of gradual insults and injuries, the EEG recordings can be extremely labor intensive. This makes it necessary to have either efficient means of collecting, storing, and displaying the long-term recordings or to come up with new techniques of compressing the EEG data, while preserving its characteristic features. Better methods to overcome these problems have been developed in both directions, although primarily toward efficient storage and display techniques. Methods of compressed spectral array representation overcome the limitation of compressed display to a great extent.

DISCUSSION

The review presented above emphasizes that the EEG is sensitive to different states of the brain and therefore may

serve as a useful tool for neurological monitoring of brain function and dysfunction. In various clinical cases, the EEG has been used to observe patients and to make critical decisions. Nevertheless, its complicated nature and difficulty of interpretation has limited its clinical use. The following section should give an overview of the common techniques in analyzing EEG signals.

TECHNIQUES OF EEG ANALYSIS

Introduction

The cases presented above show that EEG has significant clinical relevance in detecting several diseases as well as in different stages of recovery. Still the complex nature of EEG has so far restricted its use in many clinical situations. The following discussion should give an overview of the state-of-the-art EEG monitoring and analysis techniques. The presented EEG analysis methods are divided into two basic categories, parametric and nonparametric, respectively, assuming that such a division is conceptually more correct than the more common differentiation between frequency and time-domain methods because they represent two different ways of describing the same phenomena.

NonParametric Methods

In most of these analysis methods, the statistical properties of EEG signals are considered realizations of a Gaussian random process. Thus, the statistics of an EEG signal can be described by the first-and second-order moments.

These nonparametric time-domain and frequency-domain methods have been the most common way in analyzing EEG signals. In the following description of the different methods, it is also mentioned whether the technique is still being used in clinical settings.

Clinical Inspection. The most prevalent method of clinically analyzing the EEG is the visual inspection of chart records obtained from EEG machines. It uses the features observed in real-time EEG (like low frequency-high amplitude activity, burst suppression activity, etc.) for diagnostic and prognostic purposes. Several typical deviations from the normal EEG are related to different stages of the brain. This method suffers from the limitations like inter-user variability, labor intensiveness, and storage problems. A detailed description of logistical and technical considerations faced with EEG analysis is given later in this section.

Amplitude Distribution. Amplitude distribution is based on the fact that a random signal can be characterized by the distribution of its amplitude and accompanying mean, variance, and higher order moments. It can be observed that the amplitude distribution of an EEG signal most of the time can be considered as Gaussian (27) and the deviations from Gaussianity and its time-varying properties have been clinically used to detect and analyze different sleep stages (28,29). This method is now less popular because of more powerful and sophisticated EEG analysis techniques.

Interval Distribution. This is one of the earliest methods of quantitating the EEG (30). The method is based on measuring the distribution of intervals between either zero or other level crossings, or between maxima and minima. Often, the level crossings of the EEG's first and second derivatives are also computed to obtain more information about the spectral properties of the signal. Due to its ease of computation, the method has been shown to be useful in monitoring long-term EEG changes during anesthesia or sleep stages. Although simple, some theoretical problems are associated with this technique: It is extremely sensitive to high-frequency noise in the estimation of zero crossings and to minor changes in EEG. Also the zero crossing frequency (ZXF), the number of times the EEG signal crosses the zero voltage line, is not unique to a given waveform. Very different waveforms could give rise to the same ZXF. Despite the limitations, modified versions of period analysis are still used for clinical applications (30).

Interval-Amplitude Analysis. Interval-amplitude analysis is the method by which the EEG decomposed in waves or half-waves, both defined in time, by the interval between zero crossings, and in amplitude by the peak-to-through amplitudes. The amplitude and the interval duration of a half-wave are defined by the peak through differences in amplitude and time; the amplitude and the interval duration of a wave are defined by the mean amplitude and the sum of the interval durations of two consecutive half-waves (31,32). This method has been used clinically for sleep monitoring and depth of anesthesia studies (33).

Correlation Analysis. The computation of correlation functions constituted the forerunner of contemporary spectral analysis of EEG signals (34,35). The correlation function for random data describes the general dependence of the values of the data at one time on the values of the same data in the case of autocorrelation analysis (or of different data in the case of cross-correlation analysis) at another time. The cross-correlation between two signals x and y is defined as

$$\Phi_{xy}(\tau) := E\{x(t)y(t+\tau)\} \tag{4}$$

where τ is the lag time (note that $\Phi_{xy}(\tau)$ becomes the autocorrelation function for $x = y$ and it can be estimated for discrete data by

$$\hat{\Phi}_{xy}(m) = \frac{1}{N-|m|}\sum_{n=0}^{N-|m|-1} x(n)y(n+m), \quad m\,\varepsilon\{0, 1, 2, \ldots, M << N\} \tag{5}$$

and m is the lag number, M is the maximum lag number and $\hat{\Phi}_{xy}(m)$ is the estimate of the correlation function at lag number m. This estimation is unbiased but not consistent (36). The following modifications of this method have been used clinically:

1. Polarity coincidence correlation function: In this method, the signals are replaced by their signum equivalents, where $\text{sign}[x(t)] = +1$ for $x(t) > 0$ and $\text{sign}[x(t)] = -1$ for $x(t) < 0$. This modification achieves

computational simplification and has been shown (37) to be useful for EEG analysis.

2. Auto- or cross-averaging: This method consists of making pulses at a certain phase of the EEG (e.g., zero-crossing, peak, or through) that are then used to trigger a device that averages the same signal (auto-averaging) or another signal (cross-averaging). In this way, rhythmic EEG phenomena can be detected (38,39).
3. Complex demodulation: This method is related to correlation functions and allows one to detect a particular frequency component and to follow it over time. This is done by multiplying EEG signals with a sine wave of a desired frequency to give a product at 0 Hz. The 0 Hz component is then retained using a low-pass filter, obtaining the frequency component of interest. This method has been used to analyze visual potentials (40) and sleep spindles (41). However, correlation analysis has lost much of its attractiveness for EEG analysis since the advent of the Fourier transformation (FT) computation of power spectra.

Power Spectra Analysis. The principal application for a power spectral density function measurement of physical data is to establish the frequency composition of the data, which in turn bears an important relationship to the basic characteristics of the physical or biological system involved. The power spectrum provides a statement of the average distribution of power of a signal with respect to frequency. The FT serves as a bridge between the time domain and the frequency domain by identifying the frequency components that make up a continous waveform. An equivalent of the FT for discrete time signals is the discrete Fourier transform (DFT), which is given by

$$X(\omega) = \sum_{n=-\infty}^{+\infty} x(n)\exp(-j\omega n) \tag{6}$$

An approximation of this DFT can be easily computed using an algorithm, developed in 1965 by Cooley and Tukey (42) and known as the fast Fourier transform (FFT).

An estimation of the power spectrum can now be obtained by Fourier-transforming (using FFT/DFT) either the estimation of the autocorrelation function, as developed in the previous section (Correlogram), or the signal and calculating the square of the magnitude of the result (Periodogram). Many modifications of these methods have been developed to obtain unbiased and consistent estimates (for details, please see Ref. 43). One estimator for the power spectrum, developed by Welch (44), will be used in the last section of this work.

Based on the frequency contents, human EEG has been classified into different frequency bands, as described. Correlations of normal function as well as dysfunctions of the brain have been made with the properties (frequency content, powers) of these bands. Time-varying power spectra have also been used to analyze time variations in EEG frequency properties (45). One of the main advantages of this kind of analysis is that it retains almost all the information content of EEG, while separating out the low-frequency artifacts into a small band of frequencies. On the other hand, it suffers from some of the limitations of feature analysis, namely, inter-user variability, labor intensiveness, and storage problems. There have been attempts to reduce the labor intensiveness by creating displays like linear display of spectral analysis and grayscale display of spectral analysis (30), which compromises the amount of information presented.

Cross-Spectral Analysis. This kind of analysis allows quantification of the relationship between different EEG signals. The cross-power spectrum $\{P_{xy}(f)\}$ is the product of the smoothed DFT of one signal and the complex conjugate of the other [see for details, Jenkins and Watts (46)]. As $P_{xy}(f)$ is a complex quantity, it has a magnitude and phase and can be written as

$$P_{xy}(f) = |P_{xy}(f)|\exp[j\phi_{xy}(f)] \tag{7}$$

where $j = \mathrm{sqrt}(-1)$, and $\phi_{xy}(f)$ is the phase spectrum. With the cross-power spectrum, a normalized quantity, the coherence function, can be defined as follows:

$$\mathrm{coh}_{xy}(f) = \frac{|P_{xy}(f)|^2}{P_{xy}(f)P_{yy}(f)} \tag{8}$$

where $P_{xx}(f)$ and $P_{yy}(f)$ are the autospectral densities of $x(t)$ and $y(t)$. The spectral coherence can be observed as a measurement of the degree of the "phase synchrony" or "shared activity" between spatially separated generators. Therefore, unity in this quantity indicates a complete linear relationship between two electrode sites, whereas a low value for the coherence function may indicate that the two EEG locations are connected via a nonlinear pathway and that they are statistically mostly independent.

Coherence functions have been used in several investigations of the EEG signal generation and their relation to brain functions, including studies of hippocampal theta rhythms (47), on limbic structures in humans (48), on thalamic and cortical alpha rhythms (49), on sleep stages in humans (29), and in EEG development in babies (50).

A more generalized form of coherence is the so called "spectral regression-amount of information analysis" [introduced and first applied to EEG analysis by Gersch and Goddard (51)] which expresses the linear relationship that remains between two time series after the influence of a third time series has been removed by a partial regression analysis. If the initial coherence decreases significantly, one can conclude that the coherence between the two initially chosen signals is due to the effect of the third one. The partial coherence between the signals x and z, when the influence of y is eliminated, can be derived from

$$P_{zz,y}(f) = P_{zz}(f)(1 - \mathrm{coh}_{zy}(f)) \tag{8}$$

and

$$P_{xx,y}(f) = P_{xx}(f)(1 - \mathrm{coh}_{xy}(f)) \tag{9}$$

$P_{xz,y}(f)$ is the conditioned cross-spectral density and can be calculated as

$$P_{xz,y}(f) = P_{xz}(f)\left(1 - \frac{P_{xy}(f)P_{yz}(f)}{P_{yy}(f)P_{xz}(f)}\right) \tag{10}$$

This method has been mainly used to identify the source of EEG seizure activity (51,52).

Bispectrum Analysis. The power spectrum essentially contains the same information as autocorrelation and hence provides a complete statistical description of a process only if it is Gaussian. In cases where the process is non-Gaussian or is generated by nonlinear mechanisms, higher order spectra defined in terms of higher order moments or cumulants provide additional information that cannot be obtained from the power spectrum (e.g., phase relations between frequency components). There are situations, due to quadratic nonlinearity, in which phase coupling between two frequency components of a process results in a contribution to the power at a frequency equal to their sum. Such coupling affects the third moment sequence, and hence, the bispectrum is used in detecting such nonlinear effects. Although used in experimental settings (53,54), bispectral analysis techniques have not yet been used in clinical settings, probably due to both the complexity of the analysis and the difficulty in interpreting results.

The bispectrum of a third-order stationary process can be estimated by smothing the triple product

$$B(f_1, f_2) = E\{F_{xx}(f_1)F_{xx}(f_2)^*F_{xx}(f_1 + f_2)\} \qquad (11)$$

where $F_{xx}(f)$ represents the complex FT of the signal and $F_{xx}(f)^*$ is the complex conjugate of $F_{xx}(f)$ [for details, please see Huber et al. (55) and Dumermuth et al. (56)].

Hjorth Slope Descriptors. Hjorth (57) developed the following parameters, also called descriptors, to quantify the statistical properties of a time series:

$$\text{activity}, \quad A = a_0$$

$$\text{mobility}, \quad M = \left[\left(\frac{a_2}{a_0}\right)\right]^{\frac{1}{2}}$$

$$\text{complexity}, \quad C = \left[\left(\frac{a_4}{a_2}\right) - \left(\frac{a_2}{a_0}\right)\right]^{\frac{1}{2}}$$

where

$$a_n = \int_{-\infty}^{+\infty} (2\pi f)^n S_{xx}\,(df)$$

Note here that a_0 is the variance of the signal ($a_0 = \sigma^2$), a_2 is the variance of the first derivative of the signal, and a_4 is the variance of the signal's second derivative. Hjorth also developed a special hardware for real-time computation of these three spectral moments, which allows the spectral moments to vary as a function of time. Therefore, this form of analysis can be applied to nonstationary signals, and it has been used in sleep monitoring (58) and in quantifying multichannel EEG recordings (59). It should be noted that Hjorth's descriptors give a valid description of an EEG signal only if the signals have a symmetric probability density function with only one maximum. As this assumption cannot be made in general practice, the use of the descriptors is limited.

Parametric Methods

The motivation for parametric models of random processes is the ability to achieve better power spectrum density (PSD) estimators based on the model, than produced by classical spectral estimators. In the last section, the PSD was defined as the FT of an infinite autocorrelation sequence (ACS). This relationship may be considered as a nonparametric description of the second-order statistics of a random process. A parametric description of the second-order statistic may also be devised by assuming a time-series model of the random process. The PSD of the time-series model will then be a function of the model parameters (and not of the ACS). A special class of models, driven by white noise processes and processing rational system functions, is the autoregressive (AR), the moving average (MA), and the autoregressive moving average (ARMA) model.

One advantage of using parametric estimators is, for example, better spectral resolution. Periodogram and correlogram methods construct an estimate from a windowed set of data or ACS estimates. The unavailable data or unestimated ACS values outside the window are implicitly zero, which is an unrealistic assumption, that leads to distortions in the spectral estimate. Some knowledge about the process from which the data samples are taken is often available. This information may be used to construct a model that approximates the process that generated the observed time sequence. Such models will make more realistic assumptions about the data outside the window instead of the null data assumption. Thus, the need for window function can be eliminated. Therefore, a parametric PSD estimation method is useful in real-time estimation because a short data sequence is sufficient to determine the model. The following parametric approaches have been used to analyze EEG signals.

ARMA Model. The ARMA model is the generalized form of the AR and MA model, which represents the time series $x(n)$ in the following form:

$$\begin{aligned} x(n) + a(1)x(n-1) + a(2)x(n-2)\ldots + a(p)x(n-p) \\ = w(n) + b(1)w(n-1) + b(2)w(n-2)\ldots \\ + b(q)w(n-q) \end{aligned} \qquad (12)$$

where $a(n)$ are the AR parameters, $b(n)$ are the MA parameters, $w(n)$ is the error in prediction, and p,q are the model orders for the AR and MA model, respectively.

The power spectrum $P_{xx}(z)$ of this time series $x(n)$ can be obtained by using the ARMA parameters in the following fashion:

$$P_{xx}(z) = \left|\frac{\sum_{i=0}^{q} 1 + b(1)z^{-1} + b(2)z^{-2} + \ldots b(q)z^{-q}}{\sum_{i=0}^{p} 1 + a(1)z^{-1} + a(2)z^{-2} + \ldots a(p)z^{-p}}\right|^2 W(z) \qquad (13)$$

where $W(z)$ is the z-transform of $w(n)$. Note here that if we set all $b(q)$ equal to zero, we obtain an AR model, represented by poles close to the unit circle only and therefore an all-pole-system, and if we set all $a(p)$ equal to zero, we obtain an MA model. The ARMA spectrum can model both sharp peaks as they are obtained from an AR spectrum and

deep nulls as they are typical for an MA spectrum (60). Although ARMA is a more generalized form of the AR model, in most EEG applications, it is sufficient to compute the AR model becuase EEG signals have been found to be represented effectively by such a model (45). The AR model will be described in more detail in the following section.

Inverse AR Filtering. Assuming that an EEG signal results from a stationary process, it is possible to approximate it as a filtered noise with a normal distribution. Consequently, passing such an EEG signal through the inverse of its estimated autoregressive filter could be performed to obtain the generator noise (also called the residues) of the signals, which is normally distributed with mean zero and variance σ^2. The deviation from a noise with a normal distribution can be used as an important tool to detect nonstationarity and nonlinearities in the original signal. This method has been used to detect transient nonstationarities present in epileptiform EEG (45).

Kalman Filtering. A method of analyzing time-varying signals consists of applying the so-called Kalman estimation method of tracking the parameters describing the signal (61,62). The Kalman filter recursively obtains estimates of the parametric model coefficients (such as those of an AR model) using earlier as well as current data. These data are weighted by the Kalman filter, depending on the signal-to-noise ratio (SNR) of the respective data. For the estimation of the parametric model, coefficients data with a high SNR are weighted higher than data with a lower SNR (37).

This method is not easy to implement due to its sensitivity to model order and initial conditions; it also tends to be computationally extensive. Despite these limitations, recursive Kalman filtering has been used in EEG analysis for deriving a measure of how stationary the signal is and for EEG segmentation signal (61,62). This segmentation of the EEG signal into quasi-stationary segments of variable length is necessary and useful in reducing data for the analysis of long EEG recordings under variable behavioral conditions. Adaptive segmentation based on Kalman filtering has been used to analyze a series of clinical EEGs to show a variety of normal and abnormal patterns (63).

BURST AND ANALYZING METHODS

Introduction

This article has shown that EEG signals are sensitive to various kinds of diseases and reflect different stages of the brain. Specific EEG patterns can be observed after ischemic brain damage and during deep levels of anesthesia with volatile anesthetics like enflurane, isoflurane, or babiturate anesthesia (64). The patterns are recognized as periods of electrical silence disrupted by bursts of high-voltage activity. This phenomenon has been known since Derbyshire et al. (65) showed that wave bursts separated by periods of electrical silence may appear under different anesthetics. The term "burst suppression" was introduced to describe the occurrence of alternating wave bursts and blackout sequences in narcotized animals (66), in the isolated cerebral cortex (67), during coma with dissolution of cerebral functions (68), after drama associated with cerebral anoxia (69), and in the presence of a cortical tumor (70). Other bursting-like patterns in the EEG are seizures as they occur during epilepsy. Although also episodes of high voltage, the background EEG is not suppressed in the presence of seizures.

The knowledge about occurrence of these bursts and perods of electrical silence in the EEG is of important clinical value. Although burst suppression during anesthesia with modern anesthetics is reversible and harmless, it often is an ominous sign after brain damage (71). Frequently occurring seizures may indicate a severe injury state. Thus, it is of great interest to detect these burst and burst suppression sequences during surgery or in other clinical settings. We have already presented several methods to analyze EEG signals, their advantages and disadvantages. In the case of short episodes of burst suppression or spikes, however, methods that maintain the time-varying character of the raw EEG signal are necessary.

In this section, we want to present the mechanisms of the underlying processes, which cause burst-suppression or spiking. Methods that show the loss of EEG signal power during the occurrence of burst suppression and methods that can follow the time-varying character of the raw input signal are presented. Finally, we will present some methods that have been used to detect bursts and seizures based on detection of changes in the power of the signal.

Mechanisms of Bursts and Seizures

Bursts can be observed as abrupt changes in the activity of the entire cortex. These abrupt changes led to the assumption that a nonlinear (ON-OFF or bang-bang control system) inhibiting mechanism exists in the central nervous system (CNS) that inhibits the burst activity in the EEG. Recent studies confirm this theory and have shown that during burst-suppression, the heart rate also is decreased (72,73). At the end of the suppression, this inhibition is released abruptly, permitting burst activity in EEG and increase in heart rate. The task of such a control system in the CNS may be to decrease the chaotic activity in a possibly injured or intoxicated brain. As cortical energy consumption is correlated with the EEG, decreased cortical activity also avoids excessive, purposeless energy consumption (74). Studies on humans under isoflurane anesthesia have shown that increased burst-suppression after increased anesthesia concentration does correlate with cerebral oxygen consumption (75).

Another interesting observation is the quasi-sinusoidal character of the EEG signal during bursting. This has been shown by Gurvitch et al. (76) for the case of hypoxic and posthypoxic EEG signals in dogs. In contrast to anesthesia evoked bursts, which also contain higher frequency components up to 30 Hz, these hypoxic and posthypoxic bursts are high-voltage slow-wave signals, with frequency components in the delta range (77). Figure 6 shows two typical cortical EEG recordings from an isoflurane-anesthetized dog and a piglet after hypoxic insult. The power spectrum of the first burst in each recording is shown, respectively. Bispectral analysis as described has shown that there is

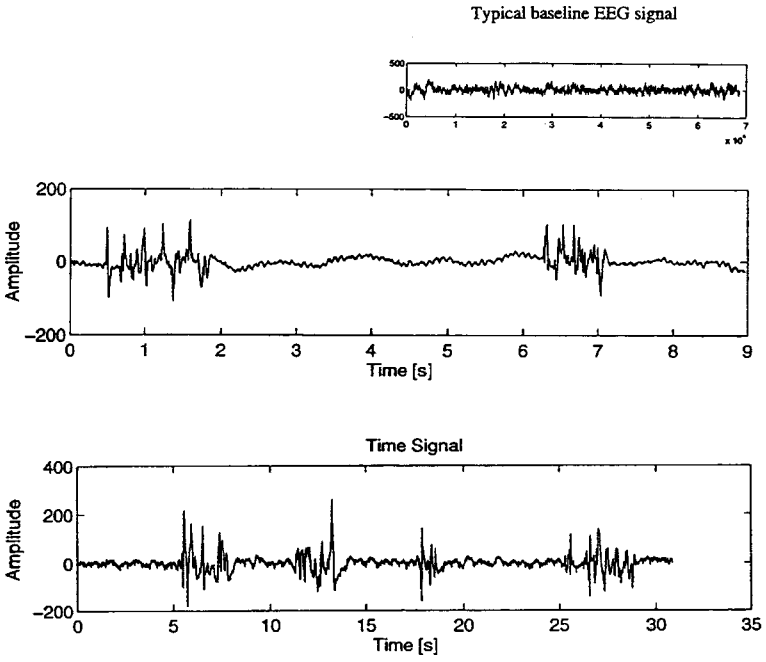


Figure 6. Burst suppresion under the inluence of isoflurane and after hypoxic insult. (a) Cortical EEG from a dog anesthetized with isoflurane/DEX. (b) Cortical EEG from a piglet during recovery from a hypoxic insult. Note the similarities in electrically silent EEG interrupted by high voltage EEG activity.

significant phase coupling during bursting (78). Due to these observations, we assume the EEG signal to be quasiperiodic during bursting and seizure sequences. Note that this observation is an important characteristic and will be used in the next section as a basic assumption for the use of an energy estimation algorithm.

The first cellular data on EEG burst suppression patterns were presented by Steriade et al. in 1994 (79). This study examined the electrical activity in cells in the thalamus, the brain stem, and the cortex during burst suppression in anesthetized cats. They showed that although the activity of intracellularly recorded cortical neurons matches the cortical EEG recording, the recording from thalamic neurons displays signs of activity during the periods of electrical silence in the cortex and the brain stem. But it has also been observed that the cortical neurons are not unresponsive during periods of electrical silence. Thalamic volleys delivered during the epochs of electrical silence were able to elicit neuronal firing or subthresholding depolarizing potentials as well as the revival of EEG activity. This observation led to the assumption that full-blown burst suppression is achieved through complete disconnection within the prethalamic, thalamocortical, and corticothalamic brain circuits and indicates that, in some instances, a few repetitive stimuli or even a single volley may be enough to produce recovery from the blackout during burst suppression. Sites of disconnection throughout thalamocortical systems are mainly inhibited synaptic transmissions due to an increase in GABAergic inhibitory processes at both thalamic and cortical synapses. Note that we showed that postsynaptic extracellular potentials at cortical neurons are the origin of the EEG signal. Therefore, this failure of synaptic transmission explains the flatness in the EEG during burst suppression. The spontaneous recurrence of cyclic EEG wave bursts may be observed as triggered by remnant activities in different parts of the affected circuitry, mainly in the dorsothalamic-RE thalamic network in which a significant proportion of neurons remains active during burst suppression. However, it is still unclear why this recovery is transient and whether there is a real periodicity in the reappearance of electrical activity. According to the state of the whole system, the wave bursts may fade and be replaced by electrical silence or may recover toward a normal pettern.

Seizures are sudden disturbances of cerebral function. The underlying causes of these disorders are heterogenous and include head trauma, lesions, infections, and genetic predisposition. The most common injury that causes seizures is epilepsy. Epileptic seizures are short, discrete episodes of abnormal neuronal activity involving either a localized area or the entire cerebrum. The abnormal time series may demonstrate abrupt decreases in amplitude, simple and complex periodic discharges, and transient patterns such as spikes (80) and large amplitude bursts.

Generalized seizures can be experimentally induced by either skull shocks to the animal or through numerous chemical compounds like pentylenetetrazol (PTZ). Several studies have shown that there are specific pathways through which the seizure activity is mediated from deeper cortical areas to the superficial cortex (81). Figure 7 shows a cortical EEG recording from a PTZ-treated rat. Nonconvulsive seizures are not severe or dangerous . In contrast, convulsive seizures like the seizures caused by epilepsy might be life threatening, and a detection of these abnormalities in the EEG at an early stage of the insult is desirable.

Reasons for Burst and Seizure Detection

We have seen in the previous section that there are various possible sources that can cause EEG abnormalities, like seizures or burst suppression interrupted by spontaneous activity outbreaks. In this section, now we want to describe why it is of importance to detect these events. Reasons for detecting bursts or seizures are as follows.

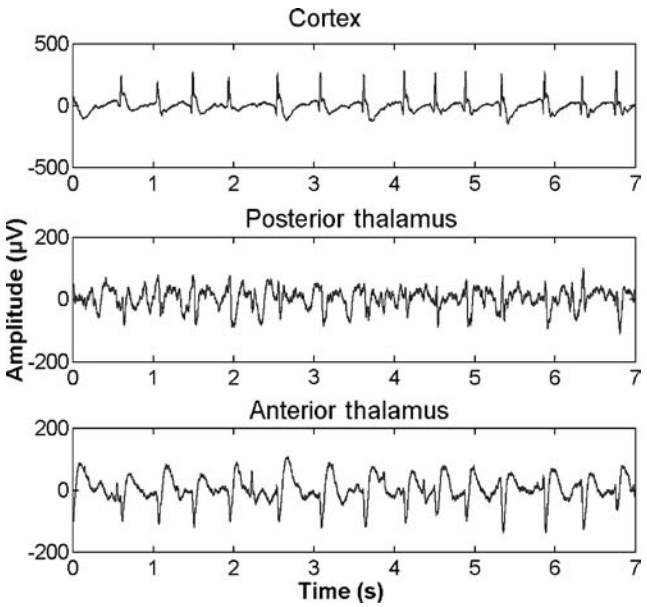


Figure 7. PTZ-induced generalized clonic seizure activity in the cortex and the thalamus of a rat. The figures from the top to the bottom show seizure activity recorded from the trans-cortex, the hippocampus, the posterior thalamus, and the anterior thalamus. Note the occurrence of spikes in the cortical recording before the onset of the seizure. At time point 40, one can see the onset of the seizure in the cortical recording, whereas the hippocampus shows increased activity already at time point 30. Such recordings can be used to study the origin and the pathways of seizures.

Confirmation of the Occurrence. In the case of seizures, it is obvious that it is desirable to detect these seizures as an indicator of possible brain injuries like epilepsy. Epileptic or convulsive seizures might be life-threatening, and detection at an early stage of the injury is necessary for medication. The frequency with which seizures occur in a patient is the basis on which the diagnosis of epilepsy is made. No diagnosis for epilepsy will be made based only on the occurrence of occasional seizures. Burst suppression under anesthesia is an indicator for the depth of anesthesia (82), and the relationship between the duration of burst suppression parts and bursting episodes is therefore desirable. In the case of hypoxia, however, burst suppression indicates a severe stage of oxygen deficiency and a possible risk of permanent brain damage or brain death. In the recovery period, in post-hypoxic analysis, the occurance of bursts might be of predictive value whether or not the patient has a good outcome (83–85).

To Further Analyze Seizure or Burst-Episodes. Not only the presence of bursts or seizures can serve as an physiological indicator, furthermore special features or characteristics of these EEG abnormalities are of importance. Intracranial EEG patterns at seizure onset have been found to correlate with specific pathology (86), and it has been suggested that the different morphologies of intracranial EEG seizure onset have different degrees of localizing value (87).

In the case of anesthesia or hypoxia-induced bursts, the duration of the bursts and the burst suppression parts may indicate the depth of anesthesia or the level of injury, respectively. Frequency or power analysis of these bursts may help discriminating these two kinds of bursts from one another (77). This is important, for example, in open heart surgery to detect reduced blood flow to the brain at a reversible stage.

Localization. Localization of the source of an injury or an unknown phenomenon is always desirable. This is valid especially in the case of epilepsy, where the injured, seizure-causing part of the brain can be operatively removed. Detecting the onsets of bursts or seizures in different channels from different regions of the brain may help us to localize the source of these events. In particular, recordings from different regions of the thalamus and the cortex have been used to study pathways of epileptic seizures.

Ability to Present Signal-Power Changes During Bursting

We have already mentioned that bursts can be observed as a sequence in the EEG signal with increased electrical activity and within sequences of increased power or energy. Therefore, looking at the power in the EEG signal can give us an idea about the presence of bursts and burst suppression episodes in the signal. Looking at the power in different frequences of a signal is classically done by estimating the PSD. We will present three methods here that have already been used in EEG signal analysis. First is a method to estimate the PSD by averaging over a certain number of periodograms, which is known as the Welch method. After obtaining the power spectrum over the entire frequency range, the total power in some certain frequency bands can then be obtained by summing together the powers in the discrete frequencies that fall in this frequency band. For this method, the desired frequency bands have to be known in advance. One method to obtain the knowledge where the dominant frequencies may be found in the power spectrum is to model the EEG signal with an AR model. Beside the fact that this method calculates the dominant frequencies, we also obtain the power in these dominant frequencies and can use this method directly to follow the power in the dominant frequencies. The third method will be a method to perform time-frequency analysis as a method to obtain the energy of a signal as a function of time as well as a function of frequency. We will present the short-time Fourier transform (STFT) as such a time-frequency distribution. As mentioned, these methods have been already used in EEG signal processing.

Feature Extraction. One major problem with these methods is the large amount of data that become available. Therefore, attempts have been made to extract spectral parameters out of the power spectrum that for themselves contain enough necessary information about the nature of the original EEG signal. The classic division of the frequency domain in four major subbands (called alpha, beta, theta, and delta waves) as described in the first section, has been one possibility of feature extraction and data reduction. However, we have also observed that these subbands may vary among the population and the major frequency components of a human EEG might be

different from the predominant frequency components of an EEG recorded from animals. Furthermore, some specific EEG changes typically involve an alteration or loss of power in specific frequency components of the EEG (88) or a shift in power over the frequency domain from one frequency range to another. This observation led to the assumption that the pre-devision of the frequency domain into four fixed subbands may not give features that are sensitive to such kinds of signal changes. We therefore propose the use of "dominant frequencies" as parameters; these are frequencies at which one can find a peak in the power spectrum, and therefore, these dominant frequencies can be observed as the frequencies with an increased activity. Recent studies (25) have shown that following the power in these dominant frequencies over time has a predictive value after certain brain injuries, whether or not the patient has a good outcome. In fact, detecting changes in power in dominant frequencies may be used as a method to visualize changes in the activity in certain frequency ranges. Another method to reduce some of the information of the power spectrum to one single value is to calculate the mean frequency of the spectrum at an instant point of time. This value can be used as a general indicator for changes in the power spectrum from one instant time point to another. Other spectral parameters that will not be described here are, for example, peak frequency or various different defined edge frequencies like the medium frequency. However, the effectiveness of these EEG parameters in detecting changes in the EEG, especially in detecting injury and the level at which they become sensitive to injury, has not been well defined. After the description of each method, we will present how we can obtain the power in the desired frequency bands and the mean frequency.

Power Spectrum Estimation Using the Welch-Method. We have already observed the use of the FT and its discrete performance in the DFT in the first section. In this section, we now want to show how we can use the DFT to obtain an estimator for the PSD.

To estimate the PSD there are two classic possibilities. The first and most direct method is the periodogram built by using the discrete-time data sequence and transforming it with DFT/FFT. We describe the algorithm in detail:

$$I(N) = \frac{1}{N}\left|\sum_{n=1}^{n=N} x(n)\exp(-j\omega n)\right| \tag{14}$$

where $x(n)$ is the discrete time signal and N is the number of FFT points. It can be observed that this basic estimator is not statistically stable, which means that the estimation has a bias and is not consistent because the variance does not tend to be zero for large values of N. The second method to achieve the PSD estimation is more indirect, in which the autocorrelation function of the signal is estimated and transformed via DFT/FFT. This estimation is called a correlogram:

$$I_N(\omega) = \sum_{m=-(N-1)}^{N-1} \hat{\Phi}_{xx}\exp(-j\omega m) \tag{15a}$$

where $\hat{\Phi}_{xx}(m)$ is the estimated autocorrelation function of a time signal $x(n)$:

$$\hat{\Phi}_{xx}(m) = \frac{1}{N}\sum_{n=0}^{N-|m|-1} x(n)x(n+|m|) \tag{15b}$$

To avoid these disadvantages of the periodogram as an estimator for PSD, many variations of this estimator were developed, reducing the bias and variance of the estimation. The most popular method among these estimators is the method of Welch (44). The given time sequence is devided into k overlapping segments of L points each, and the segments are windowed and transformed via DFT/FFT. The estimator of the PSD is then obtained by the mean of these spectra. It can be observed that as more spectral samples are used to build this estimator, the more the variance is reduced. Assuming a given sequence length of N points, the variance of the estimate will decrease if the number of points in each segment decreases. Note that a decrease in number of points results in a loss of good spectral resolution. Therefore, a compromise has to be found to achieve a small variance and a sufficient spectral resolution. To increase the number of segments, which are used to build the mean, an overlap of the segments of 50% is used.

Use of a finite segment length, $n = 0,\ldots,N-1$, of the signal $x(n)$ for computation of the DFT is equivalent to multiplying the signal $x(n)$ by a rectangular window $w(n)$. Therefore, due to the filtering effects of the window function, sidelobe energy is generated where the spectrum is actually zero. The window function also causes some smoothing of the spectrum when N is sufficiently large. To reduce the amount of sidelobe leakage caused by windowing, a nonrectangular window that has smaller sidelobes may be used. Examples of such windows include the Blackman, Hamming, and Hanning windows. However, use of these windows for reduction of sidelobe leakage also causes an increase in smoothing of the spectrum. Figure 8 shows the difference of sidelobe leakage effects between a rectangular and a Blackman window. In our case, a Tukey window is used in respect to a sufficient suppression of sidelobes and to obtain sharp mainlobes at the containing frequencies (90):

$$u(x) = \begin{cases} 0.5(1-\cos(\pi x/d)) & 0 \le x \le d \\ 1 & d \le x \le 1-d \\ 0.5(1-\cos(\pi(1-x)/d)) & 1-d \le x \le 1 \end{cases}$$

The resultant estimator of PSD is obtained by using the equation:

$$\hat{S}_{xx}(\exp(j\Omega)) = \frac{1}{kA}\sum_{i=1}^{K}\frac{1}{L}\left|\sum_{k=0}^{L-1} x_i(k)\, f_l(k)\exp(-j\Omega k)\right|^2 \tag{16}$$

where k is the number of segments, L is the number of points in each segment, $f_L(k)$ is the data window function, and

$$A = \frac{1}{L}\sum_{k=0}^{L-1} f_L^2(k) \tag{17}$$

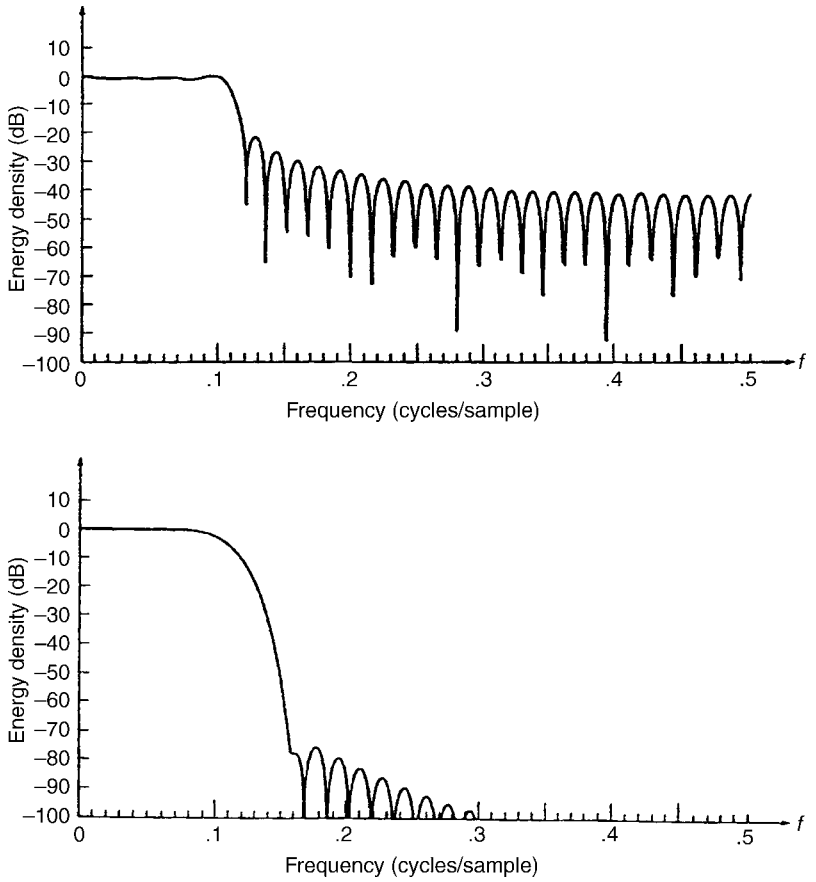


Figure 8. Comparison of sidelobe leakage effects in the spectrum using (a) rectangular window versus (b) Blackman window. Spectra are computed for a signal with (voltage spectrum $X(f) = 1$, $\text{abs}(f) < 1$; $X(f) = 0$ otherwise. The Blackman window reduces sidelobe effects and increases smoothing of the spectrum. (Adapted from Ref. 89.)

which is a factor to obtain an asymptotically unbiased estimation. Even if the PSD estimator (using the Welch method) is a consistent estimator, we have to note that it is only an approximation of the real PSD. Beside the above-mentioned limitations, unwanted effects also result from using DFT/FFT. These include aliasing, leakage, and the picket fence effect. Most of these effects may be avoided by using a window of appropriate characteristic and by fulfilling the Nyquist criterion, which is that the highest signal frequency component has to be less than one half the sampling frequency.

The FT and autocorrelation method can compute the power spectrum for a given segment of the EEG signal. The spectrum over this segment must therefore be assumed to be stationary. Loss of information will occur if the spectrum is changing over this segment, because temporal localization of spectral variations within the segment is not possible. Because burst suppression violates the assumptions (91) underlying power spectrum analysis and may cause misleading interpretations (92,93), it is necessary to increase time resolution. To track changes in the EEG spectrum over time, spectral anlysis can be performed on succesive short segments (or epochs) of data. Note that we used the spectral analysis of such epochs for our method above. We therefore may expect that the spectral analyses for the short segments are less consistent and that the effects of signal windowing will play a more important role.

Parameter Extraction. For selected sequences of EEG at each stage during a recording, spectral analysis might be performed using the Welch method. To obtain the power in the dominant frequencies, the powers in the average power spectrum are summed together over the frequency range of interest. This summation is made because the dominant frequency may vary in a small frequency range:

$$P(f_{\text{d}}) = \sum_{k=n}^{n+1} S(k) n + \ell < N/2 \tag{18}$$

where $S(k)$ is the average power spectrum, N is the FFT length, f_{d} is the dominant frequency, and $tN/2f_s$ is the bandwidth of the frequency band. Following the power in these specific frequency bands over time, we obtain a trend-plot of the power in different dominant frequency bands. This is shown in Fig. 9, where three sequences of 30 s are presented, which are recorded from a dog during different stages of isoflurane anesthesia. The dominant frequencies have been found using an AR model and are in the range of 0.5–5 Hz, 10–14.5 Hz, and 18–22.5 Hz. The recorded data are sampled with $f_s = 250$, and the sampled sequence is divided into segments of 128 points each with an overlap of 50%. The PSD estimator is obtained as described in Eq. 16.

The mean frequency (MF) of the power spectrum is computed from the following formula:

$$\text{MF} = \frac{\sum_{K=1}^{N/2} S(K)(KF_{\text{s}}/N)}{\sum_{K=1}^{N/2} S(K)} \tag{19}$$

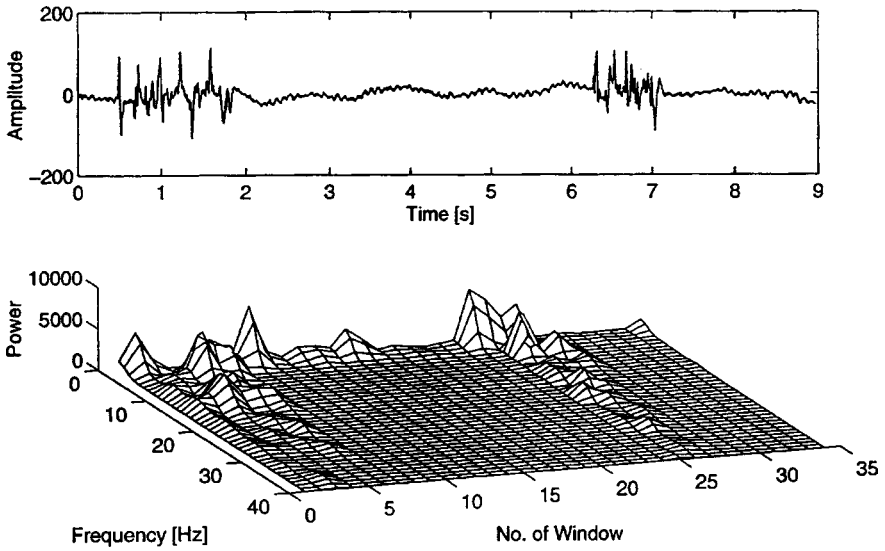


Figure 9. Trend in power in dominant frequency bands. The top row shows three input signals as they are obtained at different stages of anesthesia in a dog. The recordings show from left to right a baseline EEG followed by an EEG epoch obtained after administration of a sedative drug and finally the EEG after reversing the process. The three epochs were recorded in a distance of 1000 s to one another. The second row shows the averaged power spectra obtained with the Welch method, respectively. The data were sampled with $f = 250\ \pi$ and the FFT length is 128 points. The bottom row shows the trend in the three dominant frequency bands 0.5–5, 10–14.5, and 18–22.5 Hz. Note that this method does not provide the possibility to visualize the bursts in the EEG recording, but it can give a trend in power changes in dominant frequency bands.

where $S(K)$ is the average power spectrum, N is the FFT length, and F_s is the sampling frequency. The mean frequency can be observed as the frequency instant at which one can find the "center of mass" in the power spectrum.

Short-Time Spectral Analysis

The Algorithm. The STFT is one of the most used time-frequency methods. Time-frequency analysis is performed by computing a time-frequency distribution (TFD), also called a time-frequency representation (TFR), for a given signal. The main idea of a TFD is to allow determination of signal energy at a particular time as well as frequency. Therefore, these TFDs are functions of two dimensions, time and frequency, which have an inherent tradeoff between time and frequency resolution that can be obtained. This tradeoff between time and frequency resolution arises due to the required windowing of the signal to compute the time-frequency distribution. For good time resolution, a short time window is necessary; meanwhile a good frequency resolution requires a narrowband filter, which corresponds to a long time window. But these two conditions, a window with arbitrarily small duration and arbitrarily small bandwidth cannot be fulfilled at the same time. Thus, a compromise has to be found to achieve sufficiently good time and frequency resolution.

The STFT as one possible realization of a TFD is performed by sliding an analysis window across the signal time series and computing the FT for the current time point. The STFT for a continous-time signal $x(t)$ is defined as follows:

$$\mathrm{STFT}_x^{(y)}(t, f) = \int_{t'} [x(t')\gamma^*(t' - t)]' e^{-j2\pi f t'} dt' \qquad (20)$$

where $\gamma(t')$ is the analysis window and * denotes the complex conjugate. As discussed the analysis window chosen for the STFT greatly influences the result. Looking at the two extremes shows this influence best. Consider the case of the delta function as analysis window: $\gamma(t) = \delta(t)$. In this case, the $\mathrm{STFT} = \sum_{t=0}^{N-1} x(t)\exp(-j2\pi f t)$, which is essentially $x(t)$ and yields perfect time resolution, but no frequency resolution. On the other hand, if the analysis window is chosen to be a constant value $\gamma(t) = 1$ for all time, then the STFT becomes the Fourier transform $X(f)$, with perfect frequency resolution but no time resolution. Therefore, an appropriate window to provide both time and frequency resolution lies somewhere between these two extremes. In our case, a Hanning window is chosen as the analysis window. Figure 10 shows a plot of the STFT as it is obtained by transforming segments of the data, sampled with 250 points. The segments of 128 points each and an overlap of 50% are transformed via a 256 point FFT.

Feature Extraction. Calculating the power of the three dominant frequency bands in each segment, as described in equation 18, we obtain a contour plot of the power in these bands as shown in Fig 11. Also the mean frequency can be calculated in each segment, as described in Eq. 19.

Power Spectrum Estimation using AR-Model

The Algorithm. In the last section, the PSD was defined as the FT of an infinite ACS. This relationship may be considered a nonparametric description of the second-order statistics of a random process. A parametric description of the second-order statistic may also be devised by assuming a time-series model of the random process. The PSD of the time-series model will then be a function of the model parameters (and not of the ACS). A special class of models, driven by white noise processes and processing rational system functions, is AR, ARMA, and MA. One advantage of using parametric estimators is, for example, better spectral resolution. Periodogram and correlogram methods construct an estimate from a windowed set of data or ACS estimates. The unavailable data or unestimated ACS values outside the window are implicitly zero, which is an unrealistic assumption that leads to distortion in the spectral estimate. Some knowledge about the process from which the data samples are taken is often available. This information may be used to construct a model that approximates the process that generated the observed time sequence. Such models will make more realistic

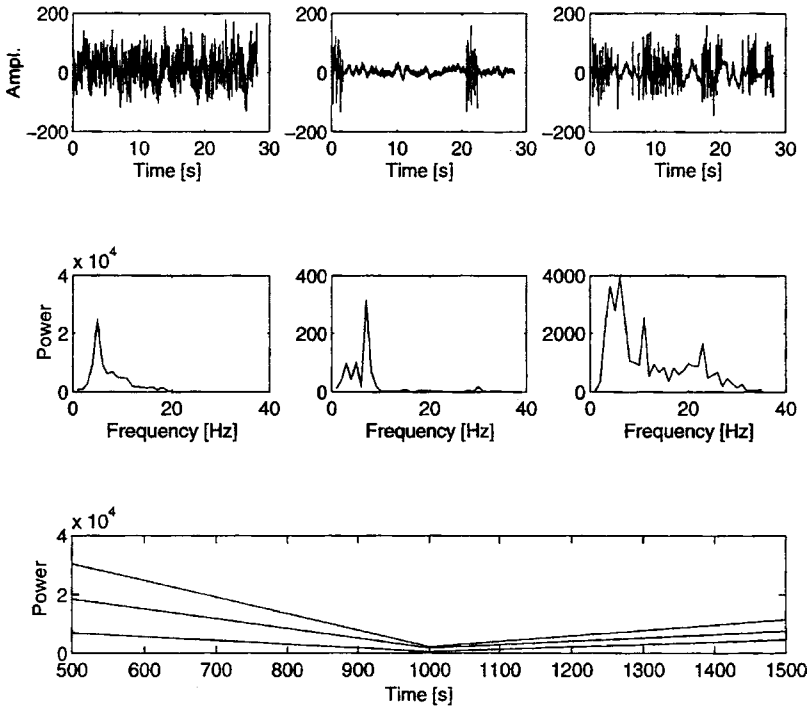


Figure 10. STFT. (a) Epoch of an EEG signal recorded from an isoflurane/DEX-treated dog. The time-varying character of this signal can be presented in the power spectrum using the STFT as shown in (b).STFT was performed with a 256 point FFT and a window length (Hamming window) of 128 points with an overlap of 50%. Data were sampled at $f = 250$ samples.

assumptions about the data outside the window instead of the null data assumption. In our case, the AR-modeling is used instead of AM or ARMA because of the advantage that the model parameter can be obtained by solving linear equations. The assumption is that if the model order p is chosen correctly, and the model parameters are calculated correctly, we obtain a PSD estimation with $p/2$ or $(p+1)/2$ sharp peaks in the spectrum at the so-called dominant frequencies.

The AR parameters have been shown to follow a recursive relation (94):

$$a_p(n) = a_{p-1}(n) + K_p a^*_{p-1}(p-n) \quad \text{for } n = 1 \ldots (p-1) \quad (21)$$

where $a_p(n)$ are the parameters for model order p, and $a_{p-1}(n)$ for model order $(p-1)$. K_p is the reflection coefficient for order p, and in Burg's maximum entropy method, K_p is determined by minimizing the arithmetic mean of the forward and backward linear prediction error power, i.e.,

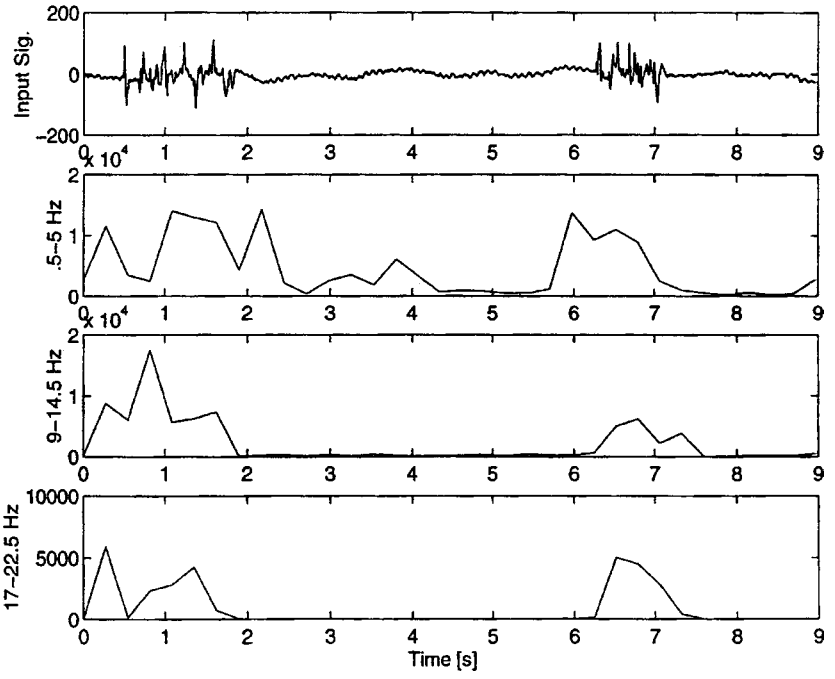


Figure 11. Energy profile in dominat frequency bands using the STFT. (a) EEG epoch recorded from isoflurane/DEX-treated dog. Summing together the energy in the frequency bands 0.5–5, 9–14.5, and 17–22.5 Hz, we obtain the energy profiles in the dominant frequency bands as shown in (b). Note that burst sequences and sequences of electrical silence in the EEG signal can be clearly distinguished with this method.

minimizing

$$\rho_p^{\mathrm{fb}} = \frac{1}{2N}\left(\sum_{n=p+1}^{N}|e_p^f(n)|^2 + \sum_{n=p+1}^{N}|e_p^b(n)|^2\right) \quad (22)$$

where e^{f} and e^{b} are the forward and backward linear prediction errors. Minimizing this equation gives us the reflection coefficients K_p for model order p:

$$K_p = \frac{-2\sum_{n=p+1}^{N} e_{p-1}^f(n)e_{p-1}^{b*}(n-1)}{\sum_{n=p+1}^{N}|e_{p-1}^f(n)|^2 + \sum_{n=p+1}^{N}|e_{p-1}^b(n)|^2} \quad (23)$$

To choose the right model order, we use the Akaike criterion (95), which is based on the maximum likelihood approach and is termed the Akaike Information Criterion (AIC) (95):

$$\mathrm{AIC}(p) = N^*\ln(\rho_p) + 2p \quad (24)$$

where ρ_p is the error variance for model order p. The error variance follows the relation:

$$\rho_p = \rho_{p-1}(1 - |K_p^2|) \quad (25)$$

The optimum model order p has to minimize AIC. With the obtained model parameter, it is now possible to present the data sequence $x(n)$ in the following way (44):

$$x(n) = w(n) - a(1)x(n-1) - a(2)x(n-2) - \ldots - a(p)x(n-p) \quad (26)$$

where the $a(i)$ are the model parameters, p is the model order, and $w(n)$ is the error in prediction. If we now choose the model order and the model parameters correctly for our estimation, $w(n)$ turns out to be zero. Taking the z-transform of this equation, we obtain

$$X(z) = \frac{W(z)}{1 + a(1)z^{-1} + a(2)z^{-2} + \ldots + a(p)z^{-p}} \quad (27)$$

where $W(z)$ is the z-transform of $w(n)$. Squaring the absolute value of $X(z)$, we obtain the estimated power spectrum:

$$P(z) = |\frac{W(z)}{1 + a(1)z^{-1} + a(2)z^{-2} + \ldots + a(p)z^{-p}}|^2 \quad (28)$$

Parameter Extraction. From equation 27, we can now obtain the dominant frequencies in the estimated power spectrum. The poles of $X(z)$ are obtained from the equation:

$$z^p + a(1)z^{p-1} + \ldots + a(p) \quad (29)$$

Evaluating this expression at the unit circle, we get frequencies ω at which there is a peak in the frequency spectrum of the data sequence and the analog frequencies of the spectral peaks are

$$F_{\text{do min ant}} = \frac{F_{\text{sampling}}}{2\pi}\omega_{\text{do min ant}} \quad (30)$$

It is now possible to evaluate the power in these frequencies either by integrating the power spectrum between desired frequencies or by the method of Johnsen and Anderson (96), which uses the residues to find the power in the peaks.

Thus, AR modeling provides the possibility to estimate the power spectrum of a signal, to calculate the frequencies at which we find a peak in the spectrum, and to obtain the power in these dominant frequencies. Note that an important assumption for a correct use of AR-modeling is a stationary signal. As this assumption cannot be made for EEG signals in long data sequences, the sample data have to be divided into small overlapping segments, in which the signal can be observed as quasi-stationary. The right segment length can be found using AIC criterion and that segment length is taken that minimizes the variance for the calculated model order. In our case for anesthetized dogs, a model order of eight was found to be appropriate. This leads to four dominant frequencies in the following frequency ranges: 0.5–5, 10–14.5, 18–22.5 Hz, and 27–31 Hz. However, it has been observed that the power in the highest dominant frequency is very small in comparison with the power in the lowest three dominant frequencies and this band is therefore ignored in our study.

Another problem with using AR models for single-channel EEG analysis during nonstationary events like burst suppression or seizures is the possibly change in model order (97). Therefore, AR models can be observed to be more appropriate for multichannel analysis.

Feature Extraction. The dominant frequencies and the power in the dominant frequencies are calculated in each segment, respectively. Therefore, a summation of power in a certain frequency band is not necessary. Figure 12 shows the power in the dominant frequencies over time. The input signal is sampled with 125 Hz and subdivided into segments of 50 points each. The segments are allowed to have an overlap of 50%.

Burst Detection Methods

The computerized detection of seizures and bursts requires differentiation of bursts from episodes of electrical silence and ictal (signal during seizure) from interictal (normal signal activity) parts. We already mentioned the characteristics of bursts and seizures, like high amplitude, high energy, and an increase in phase coupling. The change in energy at the onset and offset of bursts and seizures has been used to detect bursts in many applications. Babb et al. (98) constructed a circuit that signaled seizure detection when high amplitude–high frequency activity was observed for at least 5 s. Ives et al. (99) employed amplitude discrimination after summing and band pass filtering the EEG from 16 channels. Gotman (100) employs a detection paradigm based on measures of amplitude and time period obtained after a "half-wave decomposition" of the signals. Gotman has tested this method on numerous cases; many false positive are generated by this method. Murro et al. (101) used a discriminant function based on signal spectra. The advantage of this method is that it does not rely on visual determination of normal and abnormal signal characteristics. Recent studies have used wavelet analysis to detect the onset of bursts or seizures (102). The detection scheme is based on monitoring the variance structure of the wavelet coefficients over a selected scale range and power fluctuations in these scales individually. Webber et al. (103)

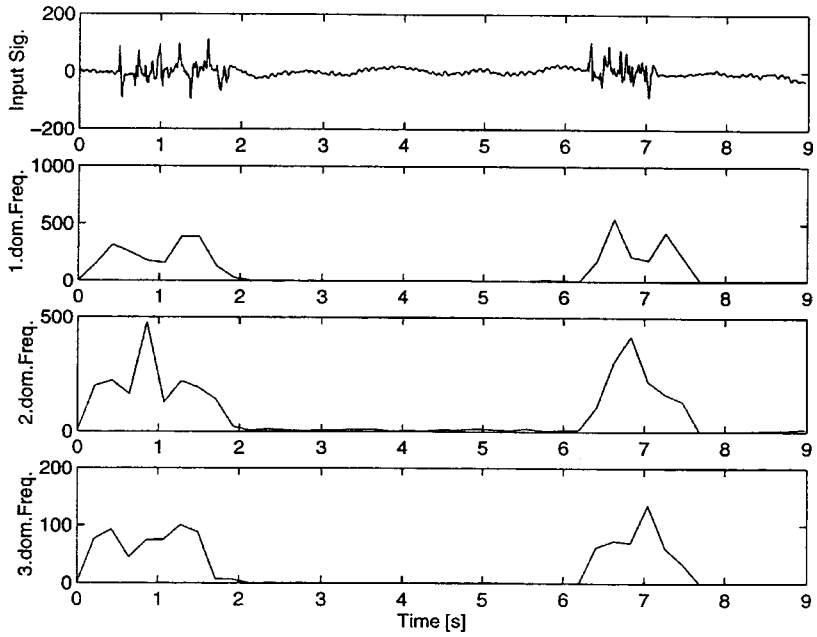


Figure 12. Energy profiles in dominant frequencies using an AR model. (a) EEG epoch recorded from an isoflurane/DEX-treated dog. An AR model (model order 8) was fitted to the EEG sample using segments of 50 points each and an overlap of 50%. The power in the three lowest dominant frequencies (ranges from 0.5–5, 10–14.5, and 18–22.5) was calculated with the residues. Note that this method provides the possibility to distinguish bursts from periods of electrical silence in the EEG signal.

presented in 1994 a detector for epileptiform discharges using an artificial neural network. The detector can detect seizure onsets and offsets in real time and was tested for raw EEG signals as well as for parametrized signals. Similar detectors have been developed by Oezdamer et al. (104) and Yaylali et al. (105). Bullmore et al. (106) presented in 1994 a detection method of ictal events based on fractal analysis of the EEG signal. The basic idea of this detector is that the EEG signal during bursts or seizures tends to have relatively low values for fractal dimensions. However, displaying the onsets and offsets of bursts and seizures very well, this method requires a visual inspection of the obtained image. Lehnerts and Elger (107) presented a detection algorithm in 1995 based on the neuronal complexity loss during ictal signal events. Alarkon et al. (108) presented a detector of seizure onsets in partial epilepsy based on feature changes extracted from the power spectrum. Features used were the total amplitude within specific frequency bands and the parameters activity, mobility, and complexity as developed by Hjorth (58) and presented earlier. The time course of these parameters was then displayed, and the onsets of ictal events were assessed when the parameter changes were above a preset threshold for more than four consecutive epochs. Franaszczuk et al. (109) presented a method in 1994, that allows not only the visualization of seizure onsets but also the flow of activity through different regions of the brain caused by epileptic discharges. The DFT method, a multichannel parametric method of analysis based on an AR model, was used for the analysis. Note that for localization of seizure onsets, this method requires recordings from different regions of the brain, especially recordings made with subdural grid and depth electrode arrays. A study that also used energy measurements in different frequency bands is that of Darcey and Williamson (110). The energy in different frequency bands is obtained from sequentially calculated power spectra, using short analysis windows to provide good time resolution. This method has been described in the previous section as STFT. The energy ratio of the energy in the ictal EEG with respect to the energy of the preictal baseline EEG was calculated, and detection was performed by comparing this quantity with a threshold. However, it has also been observed in their study that the time resolution obtained with the STFT cannot be increased upon a certain limit, due to simultaneous decrease in frequency resolution.

Agarwal and Gotman present a method of segmentation and clustering based on the nonlinear Teager Energy algorithm or TEA (111). The Teager Energy algorithm is a critical advance in energy measurement methods; based on the nonlinear operators or the second order Volterra kernel (112), it measures the frequency-dependent energy in a signal. The method uses multipliers beyond the simple square law detector ($x^2(t)$) to capture the energy in a filtered portion of the signal. It is based on the energy in a spring concept. The TEA is a two-term time-domain Agarwal, and Gotman uses a TEA variant that does not depend on a square law term. This enables studying the energy function without the evident emphasis of the equivalent zero-lagged autocorrelation term. This would cause undo emphasis of apparent white noise contaminant. For sleep staging and burst detection studies, the TEA turns out to be an essential preliminary analysis component. Sherman et al. (113) use it for burst detection, burst counting, and burst duration calculation in a cardiac arrest (CA) and recovery model in fetal pigs. This study paralleled earlier results highlighting the fact that early bursting in the recovery EEG after CA is indicative of potentially favorable outcomes. High burst counts and time spent in bursting was shown to afford group separability based on a neuro-deficit score (NDS).

CONCLUSIONS

We have presented that bursts during burst-suppression and seizures are signals of short duration with high

amplitude and therefore episodes of high energy. This led to the assumption that a discrimination of bursting episodes from burst-suppression parts could be made based on detecting a change in the energy in dominant frequency bands. However, representing the energy of a time-varying signal has its difficulties. The abrupt change from bursting parts to burst-suppression parts and the possibly short duration of these sequences make it impossible to visualize the change in energy at a certain point of time in the average power spectrum. Other methods like the STFT or parametric methods like the AR model provide the possibility to obtain both a time and a frequency resolution. Nevertheless, both methods still require a certain window length to estimate the power spectrum, and the estimated energy in each window represents the averaged energy over a certain number of sampling points.

BIBLIOGRAPHY

Cited References

1. Berger H. Ueber das Elektroenzephalogramm des Menschen. Arch Psychiat Nervenkrankheiten 1929;87:527–570.
2. Speckmann EJ, Elger CE. Introduction in the neurophysiological basis of the EEG and DC potentials. In: Niedermeyer E, daSilva FL, editors. Electroencephalography, Basic Principles Clinical Applications and Related Fields. Urban & Schwarzenberg.
3. Martin JH. Cortical neurons, the EEG, and the mechanisms of epilepsy. In: Kandel ER, Schwartz JH, editors. Principles of Neural Science. New York: Elsevier.
4. Nunez PL. Electric Fields of the Brain: The Neurophysics of EEG. New York: University Press; 1981.
5. Kandel ER, Schwartz JH, editors. Principles of Neural Science. New York: Elsevier; 1985.
6. Mountcastle VB. An organizing principle for cerebral function: The unit module and the distributed system. In: Edelman GM, Mountcastle VB, editors. The Mindful Brain. Cambridge (MA): MIT Press.
7. Hubel DH, Weisel TN. Functional architecture of macaque monkey visual cortex. Proc R Soc London Biol Sci 1978;198:1–59.
8. Kiloh, McComas, Osselton,. Clinical Electroencephalography. New York: Appleton-Century-Crofts; 1972.
9. Plonsey R, Barr RC. Bioelectricity. A Quantitative Approach. New York: Plenum Press.
10. Meyer-Waarden K. Biomedizinische Messtechnik I. Karlsruhe: Institut fuer Biomedizinische Technik; 1994.
11. Webster JG. ed. Encyclopedia of Medical Devices and Instrumentation. Vol. 1. New York: John Wiley & Sons; 1988.
12. daSilva FL, Rotterdam AV. Biophysical aspects of EEG and magnetoencephalogram generation. In: Niedermeyer E, daSilva FL, editors. Basic Principles, Clinical Applications and Related Fields. Urban & Schwarzenberg.
13. Kooi KA, Tucker RP, Marshall RE. Fundamentals of Electroencephalography. 2nd ed. New York; 1978.
14. Clark JW. Origin of Biopotentials. In: Webster JW, editor. Medical Instrumentation: Application and Design. 3rd ed., New York: John Wiley & Sons; 1998.
15. Walter WG, Dovey VJ. Electroencephalography in cases of sub-cortical tumour. J Neurol Neurosurg Psychiat 1944;7: 57–65.
16. Speckmann E.-J, Caspers H. Origin of Cerebral Field Potentials. Stuttgart: Thieme; 1979.
17. Goldman D. The Clinical Use of the "Average" Reference Electrode in Monopolar Recording. Electroenceph Clin Neurophysiol 1950;2:209.
18. Siegel G, et al. Basic Neurochemistry. 4th ed. New York: Raven Press; 1989.
19. Jordan KG. Continuous EEG and evoked potential monitoring in the neuroscience intensive care unit. J Clin Neurophysiol 1993;10(4):445–475.
20. Homan RW, Herman J, Purdy P. Cerebral location of international 10-20 system electrode placement. Electroencephalogr Clin Neurophysiol 1987;66:376–382.
21. Heuser D, Guggenberger H. Ionic changes in brain ischemia and alterations produced by drugs. A symposium on brain ischemia. Br J Anesthesia 1985;57:23–33.
22. Siesjo BK, Wieloch T. Cerebral metabolism in ischemia: Neurochemical basis for therapy. A symposium on brain ischemia. Br J Anesthesia 1985;57:47–62.
23. Astrup J, Simon L, Siesjo BK. Tresholds in cerebral ischemias—the ischemia penumbra. Stroke 1981;12:723–725.
24. Eleff SM, Hanley DH, Ward K. Post-resuscitation prognostication, declaration of brain death and mechanisms of organ salvage. In: Paradis NA, Halperin HR, Nowak RM, editors. Cardiac Arrest: The Pathophysiology & Therapy of Sudden Death Philadelphia (PA): Williams and Wilkens; 1995.
25. Goel V. A Novel Technique for EEG Analysis:Application to Neonatal Hypoxia-Asphyxia. In: Biomedical Engineering. Baltimore: Johns Hopkins University; 1995.
26. Williams GW, et al. Inter-observer variability in EEG interpretation. Neurology 1985;35:1714–1719.
27. Elul R. Gaussian behaviour of the electroencephalogram: changes during performance of mental task. Science 1969;164:328–331.
28. Dumermuth G, Gasser T, Lange B. Aspects of EEG analysis in the frequency domain. In: Dolce G, Kunkle H, editors. Computerized EEG Analysis Stuttgart Fischer: 1975.
29. Dumermuth G, et al. Spectral analysis of EEG activity during sleep stages in normal adults. Eur Neurol 1972;7:265–296.
30. Levy WJ, et al. Automated EEG processing for intraoperative monitoring. Anesthesiology 1980;53:223–236.
31. Leader HS, et al. Pattern reading of the electroencephalogram with a digital computer. Electroenceph Clin Neurophysiol 1967;23:566–570.
32. Legewie H, Probst W. On line analysis of EEG with a small computer. Electroenceph Clin Neurophysiol 1969;27:533–535.
33. Harner RN, Ostergren KA. Sequential analysis of quasistable and paroxysmal activity. In: Kellaway P, Petersen I, editors. Quantitative Analytic Studies in Epilepsy New York: Raven Press; 1975.
34. Brazier MAB, Barlow JS. Some applications of correlation analysis to clinical problems in electroencephalography. Electroenceph Clin Neurophysiol 1956;8:325–331.
35. Barlow JS, Brown RM. An analog correlator system for brain potentials. Cambridge (MA): Res Lab Electronics at MIT; 1955.
36. Kiencke U, Dostert K. Praktikum: Digitale Signalverarbeitung in der Messtechnik. Karlsruhe: Institut fuer Industrielle Informationstechnik; 1994.
37. Kaiser JF, Angell RK. New techniques and equipment for correlation computation. Cambridge (MA): Servomechanisms Lab; 1957.
38. Kamp A, Storm VLM, Tielen AM. A method for auto- and cross- relation analysis of the EEG. Electroenceph Clin Neurophysiol 1965;19:91–95.

39. Remond A, et al. The alpha average. I. Methodology and description. Electroenceph Clin Neurophysiol 1969;26:245–265.
40. Regan D. Evoked potentials in basic and clinical research. In: Remond A, editor. EEG Informatics. Didactic Review of Methods and Applications of EEG Data Processing. Amsterdam: Elsevier; 1977.
41. Campbell K, Kumar A, Hofman W. Human and automatic validation of a phase locked loop spindle detection system. Electroenceph Clin Neurophysiol 1980;48:602–605.
42. Cooley JW, Tukey JW. An algorithm for the Mchine Calculation of Complex Fourier Series. Math Computation 1965;19:297–301.
43. Kammeyer KD, Kroschel K. Digitale Signalverarbeitung. 2nd ed. Stuttgart: B.G. Teubner; 1992.
44. Welch PD. The use of Fast Fourier Transform for the estimation of power spectra. IEEE Trans Audio Electroacoust; 1970. AU-15:70–73.
45. daSilva FL. EEG analysis: Theory and practice. In: Niedermeyer E, daSilva FL, editors. Electroencephalography, Basic Principles Clinical Applications and Related Fields. Urban & Schwarzenberg.
46. Jenkins GM, Watts DG. Spectral analysis and its applications. San Fransisco: HoldenDay; 1968.
47. Walter DO, Adey WR. Analysis of brain wave generators as multiple statistical time series. IEEE Trans Biomed Eng 1965;12:309–318.
48. Brazier MAB. Studies of the EEG activity of the limbic structures in man. Electroenceph Clin Neurophysiol 1968; 25:309–318.
49. daSilva FL, et al. Organization of thalamic and cortical alpha rhythms: Spectra and coherence. Electroenceph Clin Neurophysiol 1973;35:627–639.
50. Vos JE. Representation in the frequency domain of nonstationary EEGs. In: Dolce G, Kunkel H, editors. CEAN-Computerized EEG Analysis. Stuttgart: Fischer; 1975.
51. Gersch W, Goddard G. Locating the site of epileptic focus by spectral analysis methods. Science 1970;169:701–702.
52. Tharp BR, Gersch W. Spectral analysis of seizures in humans. Comput Biomed Res 1975;8:503–521.
53. Sherman DL. Novel techniques for the detection and estimation of three wave coupling with applications to brain waves. Purdue University; 1993.
54. Braun JC. Neurological monitoring using time frequency and bispectral analysis of evoked potential and electroencephalogram signals. The Johns Hopkins University; 1995.
55. Huber PJ, et al. Statistical methods for investigating the phase relations in stationary stochastic processes. IEEE Trans Audio-Electroacoustics AU 1971;19:78–86.
56. Dumermuth G, et al. Analysis of interrelations between frequency bands of the EEG by means of the bispectrum. Electroenceph Clin Neurophysiol 1971;31:137–148.
57. Hjorth B. EEG analysis based on time domain properties. Electroenceph Clin Neurophysiol 1970;29:306–310.
58. Caille EJ, Bassano JL. Value and limits of sleep statistical analysis. Objective parameters and subjective evaluations. In: Dolce G, Kuekel H, editors. CEAN—Computerized EEG Analysis. Stuttgart: Fischer; 1975.
59. Luetcke A, Mertins L, Masuch A. Die Darstellung von Grundaktivitaet, Herd, und Verlaufsbefund sowie von paroxysmalen Ereignissen mit Hilfe der von Hjorth ausgegebenen normierten Steiheitsparameter. In: Matejcek M, Schenk GK, editors. Quantitative Analysis of the EEG. Konstanz: AEG-Telefunken; 1973.
60. Kay SM. Modern Spectral Estimation. Englewood Cliffs (NJ): Prentice-Hall, Inc.; 1987.
61. Duquesnoy AJ. Segmentation of EEGs by Means of Kalman Filtering. Utrecht: Inst. of Medical Physics TNO; 1976.
62. Isaksson A. On time variable properties of EEG signals examined by means of a Kalman filter method. Stockholm: Telecommunication Theory; 1975.
63. Creutzfeld OD, Bodenstein G, Barlow JS. Computerized EEG pattern classification by adaptive segmentation and probability density function classification. Clinical evaluation. Electroenceph Clin Neurophysiol 1985;60:373–393.
64. Clark DL, Rosner BS. Neurophysiologic effects of general anesthetics. Anesthesiology 1973;38:564.
65. Derbyshire AJ, et al. The effect of anesthetics on action potentials in the cerebral cortex of the cat. Am J Physiol 1936;116:577–596.
66. Swank RL. Synchronization of spontaneous electrical activity of cerebrum by barbiturate narcosis. J Neurophysiol 1949;12:137–160.
67. Swank RL, Watson CW. Effects of barbiturates and ether on spantaneous electrical activity of dog brain. J Neurophysiol 1949;12:137–160.
68. Bauer G, Niedermeyer E. Acute Convulsions. Clin Neurophysiol 1979;10:127–144.
69. Stockard JJ, Bickford RG, Aung MH. The electroencephalogram in traumatic brain injury. In: Bruyn PJVaGW, editor. Handbook of Clinical Neurology. Amsterdam: North Holland; 1975. 217–367.
70. Fischer-Williams M. Burst suppression activity as indication of undercut cortex. Electroenceph clin Neurophysiol 1963;15:723–724.
71. Yli-Hankala A, et al. Vibration stimulus induced EEG bursts in isoflurane anaesthesia. Electroencephal Clin Neurophysiol 1993;87:215–220.
72. Jaentti V, Yli-Hankala A. Correlation of instantaneous heart rate and EEG suppression during enflurane anaesthesia: Synchronous inhibition of hart rate and cortical electrical activity? Electroencephal Clin Neurophysiol 1990;76:476–479.
73. Jaentti V, et al. Slow potentials of EEG burst suppression pattern during anaesthesia. Acta Anaesthesiol Scand 1993; 37:121–123.
74. Ingvar DH, Rosen I, Johannesson G. EEG related to cerebral metabolism and blood flow. Pharmakopsychiatrie 1979;12: 200–209.
75. Schwartz AE, Tuttle RH, Poppers P. Electroencephalografic Burst Suppression in Elderly and Young Patients Anesthetized with Isoflurane. Anesth Analg 1989;68:9–12.
76. Gurvitch AM, Ginsburg DA. Types of Hypoxic and Posthypoxic Delta Activity in Animals and Man. Electroencephal Clin Neurophysiol 1977;42:297–308.
77. Ginsburg DA, Pasternak EB, Gurvitch AM. Correlation analysis of delta activity generated in cerebral hypoxia. Electroencephal Clin Neurophysiol 1977;42:445–455.
78. Muthuswamy J, Sherman D, Thakor NV. Higher order spectral analysis of burst EEG. IEEE Trans Biomed Eng 1999;46:92–99.
79. Steriade M, Amzica F, Contraras D. Cortical and thalamic cellular correlates of electroencephalographic burst-suppression. Electroenceph Clin Neurophysiol 1994;90:1–16.
80. Spehlmann R. EEG Primer. Amsterdam: Elsevier; 1985.
81. Mirski M. Functional Anatomy of Pentylenetetrazol Seizures. Washington University; 1986.
82. Schwilden H, Stoeckel H. Quantitative EEG analysis during anesthesia with isoflurane in nitrous oxide at 1.3 and 1.5 MAC. Br J Anaesth 1987;59:738.
83. Holmes GL, Rowe J, Hafford J. Significance of reactive burst suppression following asphyxia in full term infants. Clin Electroencephal 1983;14(3).

84. Watanabe K, et al. Behavioral state cycles, background EEGs and prognosis of newborns with perinatal hypoxia. Electroenceph Clin Neurophysiol 1980;49:618–625.
85. Holmes G, et al. Prognostic value of the electroencephalogram in neonatal asphyxia. Electroenceph Clin Neurophysiol 1982;53:60–72.
86. Lieb JP, et al. Neurophatological findings following temporal lobectomy related to surface and deep EEG patterns. Epilepsia 1981;22:539–549.
87. Ojemann GA, Engel JJ. Acute and chronic intracranial recording and stimulation. In: Engel JJ, editor. Surgical Treatment of the Epilepsy. New York: Raven Press; 1987. 263–288.
88. Vaz CA, Thakor NV. Monitoring brain electrical magnetic activity. IEEE Eng Med Biol Mag 1986;Sept.:9–15.
89. Proakis JG, Manolakis DG. Introduction to Digital Signal Processing. New York: Macmillan; 1988.
90. Dumermuth G, Molinari L. Spectral analysis of EEG background activity. In: Remond A, editor. Methods of Analysis of Brain Electrical and Magnetic Signals. Amsterdam: Elsevier; 1987.
91. Schwilden H, Stoeckel H. The derivation of EEG parameters for modelling and control of anesthetic drug effect. In: Stoeckel H, editor. Quantitation, Modelling and Control in Anaesthesia. Stuttgart: Thieme; 1985.
92. Schwilden H, Stoeckel H. Untersuchungen ueber verschiedene EEG parameter als Indikatoren des Narkosezustandes. Anaesth Intesivther Notfallmed 1980;15:279.
93. Levy WJ. Intraoperative EEG Patterns: implications for EEG monitoring. Anesthesiology 1984;60:430.
94. Isaakson A, Wennberg A, Zetterberg LH. Computer analysis of EEG signals with parametric models. Proc IEEE 1981;69: 451–461.
95. Akaike H. Recent Development of Statistical Methods for Spectrum Estimation. In: Yamaguchi NF, editor. Recent Advances in EEG and EMG Data Processing. Amsterdam: North-Holland Biomedical Press; 1981.
96. Johnsen SJ, Andersen N. On power estimation in maximum entropy spectral analysis. Geophysics 1978;43:681–690.
97. Hilfiker P, Egli M. Detection and evolution of rhythmic components in ictal EEG using short segment spectra and discriminant analysis. Electroenceph Clin Neurophysiol 1992;82:255–265.
98. Babb TL, Mariani E, Crandall PH. An electronic circuit for detection of EEG seizures recorded with implanted electrodes. Electroenceph Clin Neurophysiol 1974;37:305–308.
99. Ives JR, et al. The on-line computer detection and recording of spontaneous temporal lobe epileptic seizures from patients with implanted depth electrodes via a radio telemetry link. Electroenceph Clin Neurophysiol 1974;73:205.
100. Gotman J. Automatic seizure detection. Electroenceph Clin Neurophysiol 1990;76:317–324.
101. Murro AM, et al. Computerized seizure detection of complex partial seizures. Electroenceph Clin Neurophysiol 1991;79: 330–333.
102. Mehtu S, Onaral B, Koser R. Detection of seizure onset using wavelet analysis. In: Proc 16th Annual Int. Conf. IEEE-EMBS. Maryland, 1994.
103. Webber WRS, et al. Practical detection of epileptiform discharges (EDs) in the EEG using an artificial neural network: A comparison of raw and parameterized EEG data. Electroenceph Clin Neurophysiol 1994;91:194–204.
104. Oezdamer O, et al. Multilevel neural network system for EEG spike detection. In: Tsitlik JE, editor. Computer-Based Medical Systems. Washington: IEEE Computer Society Press; 1991. p 272–279.
105. Yaylali I, Jayakar P, Oezdamer O. Detection of epileptic spikes using artificial multilevel neural networks. Electroenceph clin Neurophysiol 1992;82.
106. Bullmore ET, et al. Fractal analysis of electroencephalographic signals intracerebrally recorded during 35 epileptic seizures: Evaluation of a new method for synoptic visualisation of ictal events. Electroenceph Clin Neurophysiol 1994;91:337–345.
107. Lehnerts K, Elger CE. Spatio-temporal dynamics of the primary epileptogenic area in temporal lobe epilepsy characterized by neuronal complexity loss. Electroenceph Clin Neurophysiol 1995;95:108–117.
108. Alarkon G, et al. Power spectrum and intracranial EEG patterns at seizure onset in partial epilepsy. Electroenceph Clin Neurophysiol 1995;94:326–337.
109. Franaszczuk PJ, Bergey GK, Kaminski MJ. Analysis of mesial temporal seizure onset and propagation using the directed transfer function method. Electroenceph Clin Neurophysiol 1994;91:413–427.
110. Darcey TM, Williamson PD. Spatio-Temporal EEG Measures and their Application to Human Intracranially Recporded Epileptic Seizures. Electroenceph Clin Neurophysiol 1985; 61:573–587.
111. Agarwal R, et al. Automatic EEG Analysis during long-term monitoring in the ICU. Electroencephal Clin Neurol 1998;107:44–58.
112. Bovik AC, Maragos P, Quatieri TF. AM-FM Energy Detection and Separation in Noise Using Multiband Energy Operators. IEEE Trans Signal Processing 1993;41(12): 3245–3265..
113. Sherman DL, et al. Diagnostic Instrumentation for Neural Injury. IEEE Instrum Measure 2002;28–35.

See also EVOKED POTENTIALS; MONITORING IN ANESTHESIA; REHABILITATION, COMPUTERS IN COGNITIVE; SLEEP STUDIES, COMPUTER ANALYSIS OF.

ELECTROGASTROGRAM

DZ CHEN
University of Kansas Medical Center
Kansas City, Kansas

ZHIYUE LIN
University of Texas Medical Branch
Galveston, Texas

INTRODUCTION

Electrogastrography, a term similar to electrocardiography (ECG), is usually referred to as the noninvasive technique of recording electrical activity of the stomach using surface electrodes positioned on the abdominal skin (1–3). The cutaneous recording obtained using the electrogastrographic technique is called electrogastrogram (EGG). In this article, both electrogastrography and electrogastrogram are abbreviated to EGG. Due to the noninvasive nature and recent advances in techniques of EGG recording and computerized analysis, EGG has become an attractive tool to study the electrophysiology of the stomach and pathophysiology of gastric motility disorders and is currently utilized in both research and clinical settings (4–7).

Although there are now several commercially available hardware–software packages making recording and analysis of EGG relatively easy to perform, many centers still use home-built equipment because the interpretations of specific frequency and EGG amplitude parameters are still debated and the clinical utility of EGG is still under investigation (6–10). Therefore, there are definite needs for better definition of the normal frequency range of EGG and dysrhythmias as well as standardization of EGG recording and advanced analysis methods for extraction and interpretation of quantitative EGG parameters. More outcome studies of EGG are also needed to determine the usefulness of EGG in the clinical settings. This article covers the following topics: a brief historic review of EGG, basics of gastric myoelectrical activity, measurement and analysis of EGG including multichannel EGG, interpretation of EGG parameters, clinical applications of EGG and future development of EGG.

HISTORIC REVIEW OF EGG

Electrogastrography was first performed and reported by Walter Alvarez back in the early 1920s (1,11). On October 14, 1921, Walter Alvarez, a gastroenterologist recorded the first human EGG by placing two electrodes on the abdominal surface of "a little old woman" and connected them to a sensitive string galvanometer. A sinusoid-like EGG with a frequency of 3 cycles/min (cpm) was then recorded. As Alvarez described in his paper, "the abdominal wall was so thin that her gastric peristalsis was easily visible" (1). Alvarez did not publish any other paper on EGG probably because of a lack of appropriate recording equipment.

The second investigator to discover the EGG is I. Harrison Tumpeer, a pediatrician who probably performed the first EGG in children (12). In a note in 1926 (12) and in a subsequent publication (13), Tumpeer reported the use of limb leads to record the EGG from a 5 week old child who was suffering from pyloric stenosis and observed the EGG as looking like an ECG (electrocardiogram) with a slowly changing baseline (11).

However, it took $\sim$30 years for EGG to be recovered by R.C. Davis, a psychophysiologist in the mid-1950s (14). Davis published two papers on the validation of the EGG using simultaneous recordings from needle electrodes and a swallowed balloon (14,15). Although Davis made only slow progress in EGG research, his two papers had stimulated several other investigators to begin doing EGG research, such as Dr. Stern who started working in Davis' lab in 1960 (11).

Stevens and Worrall (1974) were probably the first ones who applied the spectral analysis technique to EGG (16). They obtained simultaneous recordings from a strain gauge on the wall of the stomach and EGG in cats to validate the EGG. They not only compared frequencies recorded from the two sites visually in the old fashion way, but also used a fast paper speed in their polygraph and digitized their records by hand once per second, and then analyzed EGG data using Fourier transform (11).

Beginning in 1975, investigators in England published a number of studies on frequency analysis of the EGG signal and made numerous advances in techniques for analysis of the EGG, including fast Fourier transform (FFT) (17), phase-lock filtering (18), and autoregressive modeling (19). In some of their studies, they compared the EGG with intragastric pressure recordings and reported their findings similar to those of Nelson and Kohatsu (20). They found that there was no 1:1 correlation between the EGG and the contractions. The EGG could be used to determine the frequency of the contractions, but could not be used to determine when contractions were occurring (21).

During this same time, Smout and co-workers at Erasmus University in Rotterdam, The Netherlands, conducted several validation studies of the EGG and made major contributions in the area of signal analysis. In their landmark 1980 paper (22), they were the first ones who showed that the amplitude of the EGG increases when contractions occur. In 1985, Dr. Koch and Dr. Stern reported their study on simultaneous recordings of the EGG and fluoroscopy (23). They repeatedly observed the correspondence between EGG waves and antral contractions during simultaneous EGG-fluoroscopy recordings.

To extract information about both the frequency of EGG and time variations of the frequency, a running spectral analysis method using FFT was introduced by van der Schee and Grashus in 1987 (24), later used by some others (2,25,26) and now still used in most laboratories (5). To avoid the averaging effect introduced by the block processing of the FT, Chen et al. (27,28) developed a modern spectral analysis technique based on an adaptive autoregressive moving average model. This method yields higher frequency resolution and more precise information about the frequency variations of the gastric electrical activity. It is especially useful in detecting dysrhythmic events of the gastric electrical activity with short durations (29).

In 1962, Sobakin et al. (30) performed the EGG in 164 patients and 61 healthy controls and reported that ulcers caused no change in the EGG, but that pyloric stenosis produced a doubling of amplitude, and stomach cancer caused a breakup of the normal 3 cpm rhythm. This was probably the first large-scale clinical use of the EGG. In the past two decades, numerous studies have been reported on the clinical use of the EGG including understanding the relationship between the EGG and gastric motility (22,23,31–37), gastric myoelectrical activity in pregnant women (38–40), gastric myoelectrical activity in diabetics or gastroparetic patients (41–44), gastric myoelectrical activity in patients with dyspepsia (45–50), and prediction of delayed gastric emptying using the EGG (42,46,47,49,51).

As Dr. Stern wrote in 2000, "the history of EGG can be described as three beginnings, a length period of incubation, and a recent explosion" (11). It is beyond the scope of this article to cover every aspect of EGG studies. For more information about the EGG, readers are referred to some excellent articles, reviews, dissertations, and chapters (2,5,6,21,27,52–56).

ELECTROPHYSIOLOGY OF THE STOMACH

Normal Gastric Myoelectrical Activity

Along the gastrointestinal tract, there is myoelectrical activity. *In vitro* studies using smooth muscle strips of

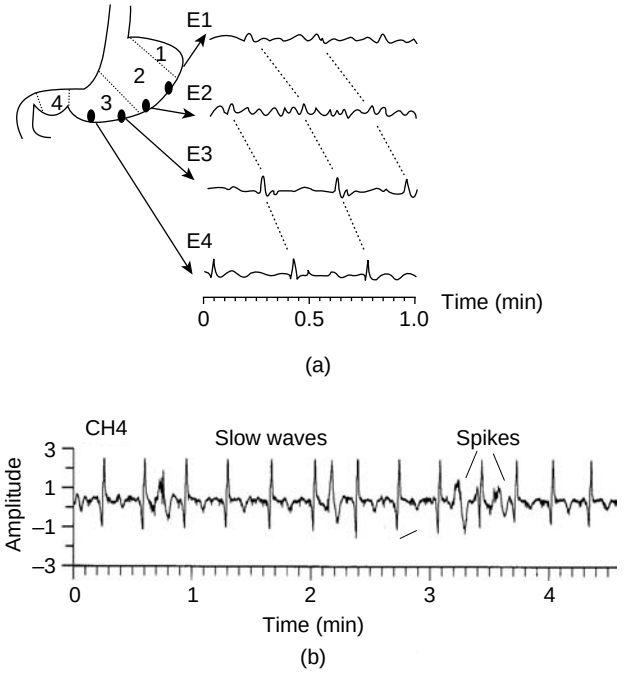


Figure 1. (a) Anatomy of the stomach (1–fundus, 2–body (corpus), 3–antrum, 4–pylorus) and the origin and propagation of the electrogastric signal from proximal (E1) to distal (E4) electrodes measured from the serosa of the human stomach. The dotted lines show the detection of a gastric slow wave traveling distally along the stomach wall. (b) Serosal recording obtained from the distal stomach in a patient. The trace shows slow waves of ~ 3 cpm with and without superimposed spike potentials.

the stomach have revealed independent gastric myoelectrical activity (GMA) from different regions of the stomach. The highest frequency of the gastric myoelectrical activity was recorded in the corpus and the lowest frequency in distal antrum. However, *in vivo* studies demonstrated a uniform frequency in the entire stomach under healthy conditions, because the myoelectrical activity in the corpus with the highest frequency drives or paces the rest of stomach into the same higher frequency (see Fig. 1).

Gastric myoelectrical activity is composed of slow waves and spike potentials. The slow wave is also called the pacesetter potential, or electrical control activity (57–59). The spike potentials are also called action potentials or electrical response activity (57–59). While slow waves are believed originated from the smooth muscles, recent *in vitro* electrophysiological studies suggest that interstitial cells of Cajal (ICC) are responsible for the generation and propagation of the slow wave (60). The frequency of normal slow waves is species-dependent, being ~ 3 cpm in humans and 5 cpm in dogs, with little day-to-day variations. The slow wave is known to determine the maximum frequency and propagation of gastric contractions. Figure 1 presents an example of normal gastric slow waves measured from a patient. Normal 3 cpm distally propagated slow waves are clearly noted.

Spike potentials are known to be directly associated with gastric contractions, that is, gastric contractions occur when the slow wave is superimposed with spike potentials. Note, however, that *in vivo* gastric studies have failed to reveal a 1:1 correlation between spike potentials measured from the electrodes and gastric contractions measured from strain gauges although such a relationship does exist in the small intestine. In the stomach, it is not uncommon to record gastric contractions with an absence of spike potentials in the electrical recording. Some other forms of superimposed activity are also seen in the electrical recording in the presence of gastric contractions.

Abnormal GMA: Gastric Dysrhythmia and Uncoupling

Gastric myoelectrical activity may become abnormal in diseased states or upon provocative stimulations or even spontaneously. Abnormal gastric myoelectrical activity includes gastric dysrhythmia and electromechanical uncoupling. Gastric dysrhythmia includes bradygastria, tachygastria, and arrhythmia. Numerous studies have shown that gastric dysrhythmia is associated with gastric motor disorders and/or gastrointestinal symptoms (4–7,20,61,62).

A recent study has revealed that tachygastria is ectopic and of an antral origin (63). In $> 80\%$ of cases, tachygastria is located in the antrum and propagates retrogradely toward the pacemaker area of the proximal stomach. It may partially or completely override the normal distally propagating slow waves. However, most commonly it does not completely override the normal gastric slow waves. In this case, there are two different slow wave activities: normal slow waves in the proximal stomach and tachygastrial slow waves in the distal stomach. A typical example is presented in Fig. 2.

Unlike tachygastria, bradygastria is not ectopic and reflects purely a reduction in frequency of the normal pacemaking activity. That is, the entire stomach has one single frequency when bradygastria occur (63). Bradygastria is originated in the corpus and propagates distally toward the pylorus. The statistical results showing the origins of tachygastria and bradygastria are presented Fig. 3. The data was obtained in dogs and gastric dysrhythmias were recorded postsurgically or induced with various drugs including vasopressin, atropine and glucagon (63).

MEASUREMENT OF THE EGG

Gastric myoelectrical activity can be measured serosally, intraluminally, or coutaneously. The serosal recording can be obtained by placing electrodes on the serosal surface of the stomach surgically. The intraluminal recording can be acquired by intubating a catheter with recording electrodes into the stomach. Suction is usually applied to assure a good contact between the electrodes and the stomach mucosal wall. The serosal and intraluminal electrodes can record both slow waves and spikes, since these recordings represent myoelectrical activity of a small number of smooth muscle cells. These methods are invasive and their applications are limited in animals and laboratory settings.

The EGG, a cutaneous measurement of GMA using surface electrodes, is widely used in humans and clinical settings since it is noninvasive and does not disturb ongoing activity of the stomach. A number of validation studies have documented the accuracy of the EGG by

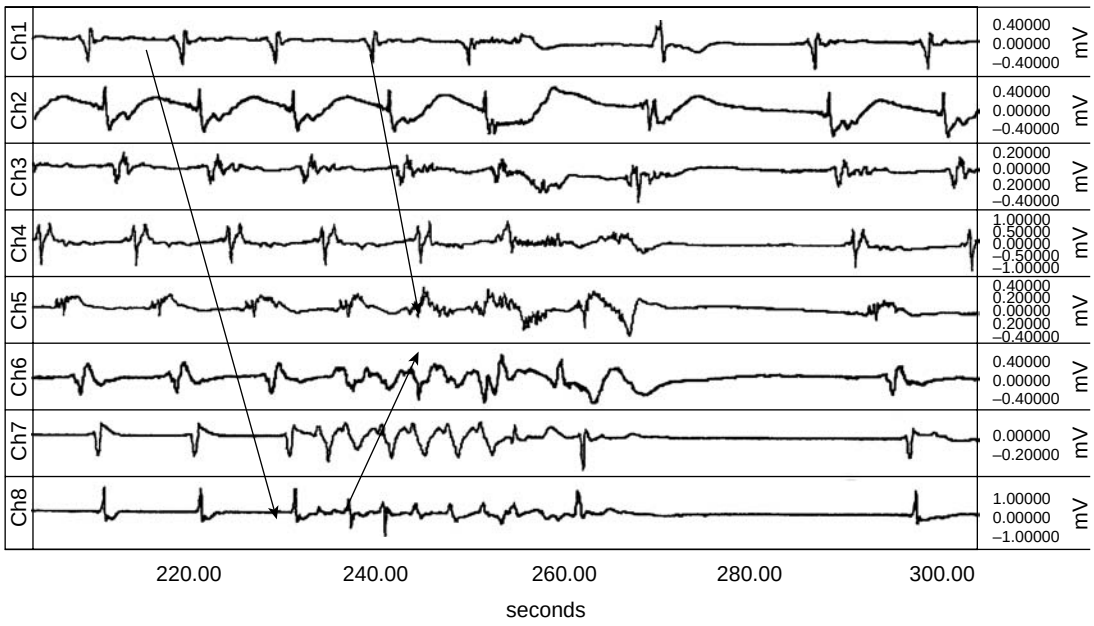


Figure 2. Origin and backward propagation of tachygastria and arrhythmia measured from the sorosa of the canine stomach. An ectopic pacemaker was located at the antrum (channel 8), which propagated both forward and backward. It changed the normal rhythm of original pacemaker that propagated down from the proximal part of the stomach (channel 1) to the distal part of the stomach (channel 8) and almost affected the entire rhythm of the stomach for a minute. A transient bradygastria was observed in all channels before normal rhythm changed back.

comparing it with the recording obtained from mucosal and serosal electrodes (19,22,31,61,64–66). Reproducibility of the EGG recording has been demonstrated, with no significant day-to-day variations (67). In adults, age and gender do not seem to have any influences on the EGG (68–71).

EGG Recording Equipment

The equipment required to record the EGG includes amplifiers, an analog-to-digital (A/D) converter and a personal computer (PC) (Figs. 4 and 5). The EGG signal must be amplified because it is of relatively low amplitude (50–500 μV). An ideal EGG amplifier should be able to enhance the gastric signal and effectively reduce interferences and noise. Abnormal frequencies of gastric slow waves may be as low as 0.5 and as high as 9 cpm. To effectively record the gastric slow wave, an appropriate recording frequency range is 0.5–18 cpm (5,6,72). It is recommended that a good choice of the sampling frequency should be three to four times of the highest signal frequency of interest (73,74). Therefore, a sampling rate for digitization of the EGG signal ≥ 1 Hz (60 cpm) is a proper choice.

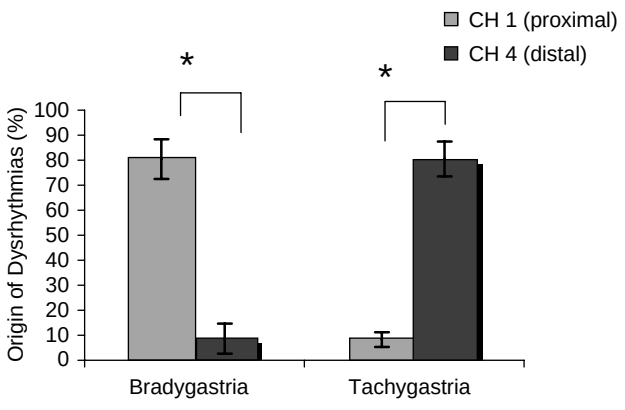


Figure 3. Statistical results showed that bradygastria originated primarily from the proximal body of the stomach, while tachygastria originated primarily from the distal stomach.

A typical EGG recording system is shown in Fig. 4. It is composed of two parts: data acquisition and data analysis. Venders who currently offer or have offered EGG equipment in the past included 3CPM Company, Medtronic/Synectics, Sandhill Scientific, Inc., RedTech, and MMS (The Netherlands), and so on (2,6). To date, there are two U.S. Food and Drug Administration (FDA)-approval EGG systems: one from Medtronic Inc (Minneapolis, MN) and the other from 3CPM Company (Crystal Bay, NV). The 3CPM Company's EGG device is a work station that consists of an amplifier with custom filters, strip chart recorder, and computer with proprietary software—the EGG Software Analysis System (EGGSAS). However, this device is only to record and analyze single-channel EGG.

The newly FDA-approved Medtronic's ElectroGastroGraphy system provides multichannel EGG recordings and analysis (75–78). It can be either running on Medtronic's Gastro Diagnostic Workstation or consisting of the Medtronic's Polygraf ID with a laptop to make a portable system (see Fig. 5). This system provides an Automatic Impedance Check function and optional Motion Sensor. With the Automatic Impedance Check, all EGG electrodes

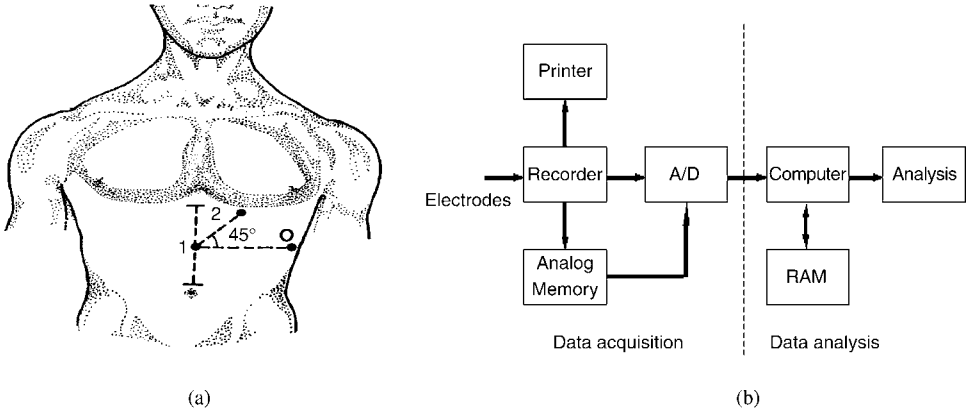


Figure 4. (a) Position of abdominal electrodes for the measurement of one-channel EGG. (b) Block diagram of an EGG recording system.

are verified to be in good electrical contact with the skin within 10 s. The optional Motion Sensor can record respiration and patient movements during data capture. This assists physicians to more easily identify motion artifacts, which can then be excluded from subsequent analysis. Currently, this system has been configured to make four-channel EGG recordings with placement of six surface electrodes on the subject's abdomen (75–78) (see Fig. 5a).

An ambulatory recording device is also available and has been used frequently in various research centers (42,46,48,71,76). For example, the ambulatory EGG recorder (Digitrapper EGG) developed by Synectics Medical Inc. (Shoreview, MN) is of the size and shape of a "walkman" (79). It contains one channel amplifier, an A/D conversion unit, and memories. It can be used to record up to 24-h one-channel EGG with a sampling frequency of 1 Hz. Information colleted during recording can be downloaded into a desktop computer for data storage and analysis (42,79).

Procedures for Recording EGG

Due to the nature of cutaneous measurement, the EGG is vulnerable to motion artifacts. Accordingly, a careful and proper preparation before the recording is crucial in obtaining reliable data.

Skin Preparation. Since the EGG signals are very weak, it is very important to minimize the impedance between

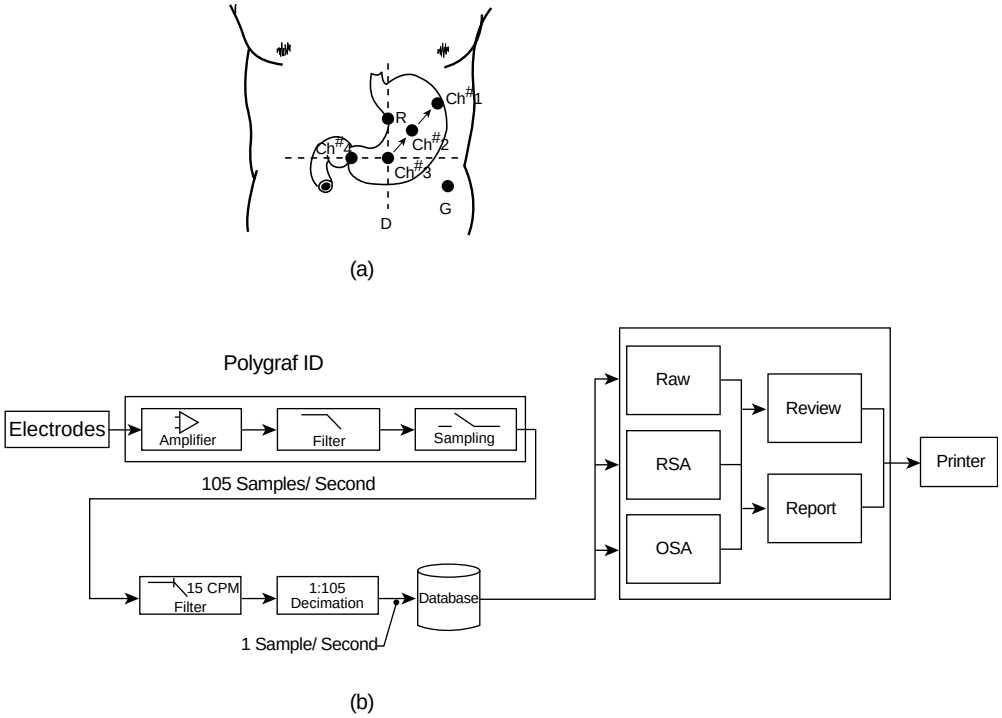


Figure 5. (a) Electrogastrogram electrodes placement for making four-channel EGG recordings consists of placing four active electrodes along the antral axis of the stomach, a reference electrode over the xyphoid process, and a ground electrode on the patient's left side. (b) Block diagram of Medtronic POLYGRAM NETTM ElectroGastroGraphy System (RSA: Running Spectral Analysis; OSA: Overall Spectral Analysis). (Reprinted with permission of Medtronic, Inc.).

Table 1. List of Systems and Procedures for Recording EGG Used by Different Groups

References	Hardware System	Analysis Method	Procedure (Duration and Meal)
Dutch research groups (8,83)	Custom-built four-channels (band-pass filter: 0.01–0.5 Hz, sampling rate: 1 Hz)	Running spectral analysis by short-time Fourier transform (STFT)	2 h before and 3 h after meal (a pancake, 276 kcal)
McCallum and Chen (42,79), Parkman et al. (46,71) Chen et al. (75,76)	MicroDigitrapper (Synectics Medical, Inc.): single channel (cut-off frequency: 1–16 cpm, sampling rate: 1 or 4 Hz). Commercial four-channel EGG recording device (cut-off frequencies: 1.8–16 cpm) (Medtronic-Synectics, Shoreview, MN)	Running spectral analysis by STFT (Gastrosoft Inc., Synetics Medical) or adaptive analysis method)	(1) 30 min before and 2 h after meal (turkey sandwich, 500 kcal) (2) 60 min before and 60 min after meal (two scrambled egg with two pieces of toasted bread 200 mL water, 282 kcal)
Penn State groups (2,84)	An amplifier with custom filters, a strip chart recorder (cut-off frequency: 1–18 cpm, sampling frequency: 4.27 Hz)	Running spectral analysis by STFT and a data sheet with percentage distribution of EGG power in four frequency ranges	Water load test (45 min)

the skin and electrodes. The abdominal surface where electrodes are to be positioned should be shaved if necessary, cleaned and abraded with some sandy skin-preparation jelly (e.g., Ominiprep, Weaver, Aurora, CO) in order to reduce the impedance between the bipolar electrodes to > 10 kΩ. The EGG may contain severe motion artifacts if the skin is not well prepared.

Position of the Electrodes. Standard electrocardiographic-type electrodes are commonly used for EGG recordings. Although there is no established standard, it is generally accepted that the active recording electrodes should be placed as close to the antrum as possible to yield a high signal-to-noise ratio (80). The EGG signals can be recorded with either unipolar or bipolar electrodes, but bipolar recording yields signals with a higher signal-to-noise ratio. One commonly used configuration for recording one-channel EGG is to place one of two active electrodes on the midline halfway between the xiphoid and umbilicus and the other active electrode 5 cm to the left of the first active electrode, 30 cephalad, at least 2 cm below the rib cage, in the midclavicular line. The reference electrode is placed on the left costal margin horizontal to the first active electrode (42,81) (Fig. 4a). One commonly used configuration of electrodes for making four-channel EGG recordings is shown in Fig. 4a, including four active electrodes along the antral axis of the stomach, a reference EGG electrode over the xyphoid process, and a ground EGG electrode on patient's left side (75–78).

Positioning the Patient. The subject needs to be in a comfortable supine position or sit in a reclining chair in a quiet room throughout the study. Whenever possible, the supine position is recommended, because the subject is more relaxed in this position, and thus introduces fewer motion artifacts. The subject should not be engaged in any conversations and should remain as still as possible to prevent motion artifacts (7,8,79).

Appropriate Length of Recording and Test Meal

The EGG recording is usually performed after a fast of 6 h or more. Medications that might modify GMA (prokinetic and antiemetic agents, narcotic analgesics, anticholinergic drugs, non-steroidal anti-inflammatory agents) should be stopped at least 48 h prior to the test (6,7). The EGG should be recorded for 30 min or more (no < 15 min in any case) in the fasting state and for 30 min or more in the fed state. A recording < 30 min may not provide reliable data and may not be reproducible attributed to different phases of migrating motor complex (82) in the fasting state.

The test meal should contain at least 250 kcal with no $> 35\%$ of fat (82). Solid meals are usually recommended although a few investigators have used water as the test "meal" (see Table 1).

EGG DATA ANALYSIS

In general, there are two ways to analyze EGG signals. One is time-domain analysis or waveform analysis, and the other is frequency-domain analysis. Numerous EGG data analysis methods have been proposed (18,19,24,27,53,55,56,84–97).

Time-Domain Data Analysis

Like other surface electrophysiological recordings, the EGG recording contains gastric signal and noise.

Table 2. Composition of the EGG

	Components	Frequency (cpm)
Signal	Normal slow wave	2.0–4.0
	Bradygastria	0.5–2.0
	Tachygastria	4.0–9.0
	Arrhythmia	NA[a]
Noise	Respiratory	12–24
	Small bowel	9–12
	ECG	60–80
	Motion artifacts	Whole range

[a]Not available

Compared with other surface recordings, such as ECG, the quality of EGG is usually poor. The gastric signal in the EGG is disturbed or may even be completely obscured by noise (see Table 2). The frequency of gastric signals is from 0.5 to 9.0 cpm, including normal (regular frequency of 2–4 cpm) and abnormal frequencies. The gastric signals with abnormal frequencies may be divided further into bradygastria (regular frequency of 0.5–2.0 cpm), tachygastria (regular frequency of 4–9 cpm) and arrhythmia (irregular rhythmic activities) (62).

The noise consists of respiratory artifacts, interferences from the small bowel, ECG, and motion artifacts (see Table 2). The respiratory artifact has a frequency from 12 to 24 cpm. It is a common and thorny problem. It is superimposed upon almost every EGG recording if not appropriate processed. Occasionally, the slow wave of the small intestine may be recorded in the EGG. The frequency of intestinal slow waves is 12 cpm in duodenum and 9 cpm in the ileum. The intestinal slow wave is usually weaker than the gastric slow wave. One can avoid recording intestinal slow waves by placing electrodes in the epigastric area. The frequency of ECG is between 60 and 80 cpm. It can be eliminated using conventional low pass filtering because its frequency is much higher than that of the gastric signal component. The frequency of motion artifacts is in the whole recording frequency range. To minimize motion artifacts, the subject must not talk and should remain still during recording.

The time-domain analysis methods with the aid of computers that were introduced to facilitate the EGG data analysis include (*1*) adaptive filtering. It is used to reduce noise such as respiratory artifacts with minimal distortion of the gastric signal component of interest (27,90,91), (*2*) coherent averaging. It is applied to filter out random noise by averaging a large number of EGG waves (85), (*3*) use of feature analysis and artificial neural networks to automatically detect and delete motion artifacts (93,94), and (*4*) use of independent component analysis to separate gastric signals from multichannel EGGs (97).

When noise level is low, it is possible to visually analyze the raw EGG tracing (3,6,84,98) to identify periods of artifact and provide a qualitative determination of recording segments with normal frequencies of ~ 3 cpm and those of abnormally high (tachygastria) or low (bradygastria) and the presence or absence of a signal power increase after eating a test meal. Artifacts usually are readily recognized visually as sudden, high amplitude off-scale deflections of the EGG signal (see Fig. 6). Artifactual periods must be excluded before analysis. This is because (a) they are usually strong in amplitude and may completely obscure the gastric signal; (b) they have a broad-band spectrum and their frequencies overlap with that of the gastric signal; therefore they are not separable using even spectral analysis method, and jeopardize any kind of quantitative analyses of the EGG data (79).

EGG Parameters (99)

Although a noise-free EGG signal is attainable by means of advanced signal processing techniques (27,86), the waveform analysis of the EGG has rarely been used, because the waveform of the EGG is related to many factors, including the thickness of the abdominal wall of the subject, skin preparation, position of the electrodes, and characteristics of the recording equipment (100). Furthermore, the number of specific characteristics of the EGG is limited. With single-channel EGG recording, only frequency and

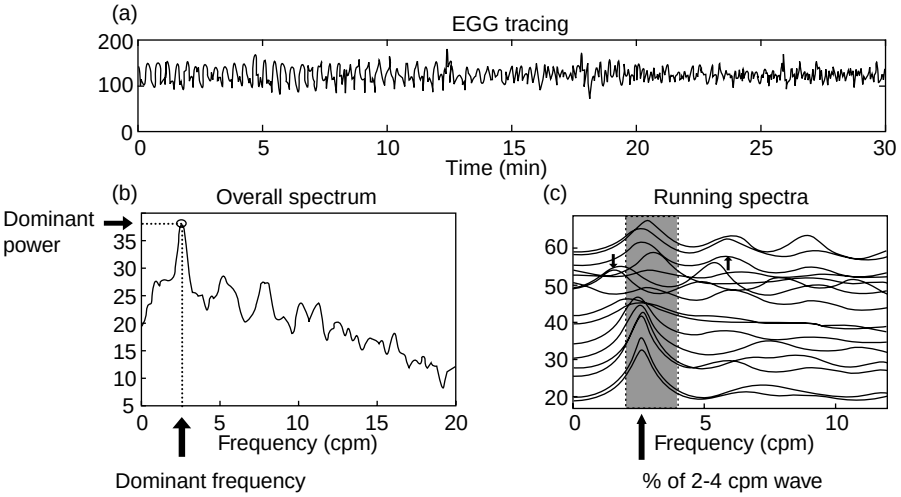


Figure 6. The EGG data analysis. (a) A 30 min EGG recording. (b) The power spectra of the 30 min EGG data. The EGG dominant frequency (DF) and power at DF can be determined from the spectrum, (c) adaptive running spectra. Each curve (from bottom to top) is the spectrum of 2 min EGG data. The percentage of normal slow waves (or dysrhythmias) can be determined from the spectra by counting spectral peaks in each frequency band.

amplitude can be measured. Recent computer simulations and experiments have shown that the propagation of the gastric slow wave can be identified from the multichannel EGG recordings (27,100), it is, however, difficult to get this information from waveform analysis (27). Accordingly, quantitative data analyses of the EGG are mostly based on spectral analysis methods. Some important EGG parameters obtained by the spectral analysis methods are described as in the following sections.

EGG Dominant Frequency and Dominant Power. The frequency believed to be of gastric origin and at which the power in the EGG power spectrum has a peak value in the range of 0.5–9.0 cpm is called the EGG dominant frequency. The dominant power is the power at the dominant frequency. The EGG power can be presented in a linear or decibel (dB) unit. The dominant frequency and power of the EGG are often simplified as EGG frequency and EGG power. Figure 6 shows the definition of the dominant frequency and power of the EGG. Simultaneous cutaneous and serosal (13,17–19) or mucosal (11,16) recordings of GMA have shown that the dominant frequency of the EGG accurately represents the frequency of the gastric slow wave. The dominant power of the EGG reflects the amplitude and regularity of gastric slow waves. The gastric slow wave is regarded as abnormal if the EGG dominant frequency is not within a certain frequency range (e.g., 2–4 cpm). Although there is no established definition for the normal range of the gastric slow wave, it is generally accepted that the dominant frequency of the EGG in asymptomatic normal subjects is between 2.0 and 4.0 cpm (5,6,41,72). The EGG, or a segment of the EGG, is defined as tachygastria if its frequency is > 4.0 cpm, but < 9.0 cpm, bradygastria if its frequency is < 2.0 cpm and arrhythmia if there is a lack of a dominant frequency (see Table 2).

Power Ratio or Relative Power Change. As the absolute value of the EGG dominant power is related to many factors, such as the position of the electrodes, the preparation of skin and the thickness of the abdominal wall, it may not provide much useful information. One of the commonly used EGG parameters associated with the EGG dominant power is the power ratio (PR) or the relative power change after an intervention such as meal, water, or medication. Note that the power of the EGG dominant frequency is related to both the amplitude and regularity of the EGG. The power of the EGG dominant frequency increase when EGG amplitude increases. It also increases when the EGG becomes more regular. Previous studies have shown that relative EGG power (or amplitude) change reflects the contractile strength of the gastric contractions (22,33,37).

Percentage of Normal Gastric Slow Waves. The percentage of normal slow waves is a quantitative assessment of the regularity of the gastric slow wave measured from the EGG. It is defined as the percentage of time during which normal gastric slow waves are observed in the EGG.

Percentage of Gastric Dysrhythmias Including Bradygastria, Tachygastria, and Arrhythmia. The percentage of gastric dysrhythmia is defined as the percentage of time during which gastric dysrhythmia is observed in the EGG. In contrast to the percentage of normal gastric slow waves in an EGG recording, this parameter represents the abnormality of the EGG or gastric slow waves.

Instability Coefficients. The instability coefficients are introduced to specify the stability of the dominant frequency and power of the EGG (99). The instability coefficient (IC) is defined as the ratio between the standard deviation (SD) and the mean:

$$\mathrm{IC} = \mathrm{SD}/\mathrm{mean} \times 100\%$$

The clinical significance of the instability coefficient has been demonstrated in a number of previous studies (37,40,99). The instability coefficients defined by Geldof et al. is slightly different from the one defined above. More information can be found in Refs. 83,101.

Percentage of EGG Power Distribution. The percentage of EGG power distribution was introduced by Koch and Stern (102) and is defined as the percentage of total power in a specific frequency range in comparison with the power in the total frequency range from 1 to 15 cpm. For example;

$$\%\ \text{of (2.4–3.6 cpm)} = \text{the power within 2.4–3.6 cpm/(the total power from 1 to 15 cpm)} \times 100\%$$

Using this parameter, Koch et al. (102) found that the percentage of power in the 3-cpm range was significantly lower in patients with idiopathic gastroparesis compared to patients with obstructive gastroparesis. They also found that the percentage of power in the tachygastria range (3.6–9.9 cpm) correlated significantly with the intensity of nausea reported during vector-induced motion sickness (103). The advantage of this method is that it is easy for computation. We should be aware, however, that only relative values of this parameters in comparison with the control data should be used. Even in normal subjects, the percentage of normal EGG activity computed in this way will never be 100%. Note that this parameter is sensitive to noise, since any noise component in the frequency band of 1–15 cpm affects the computation of this parameter. Harmonics of the fundamental 3 cpm slow wave may be computed as tachygastria (8,74).

Percentage of Slow Wave Coupling. Slow wave coupling is a measure of the coupling between two EGG channels. The percentage of slow wave coupling is defined as the percentage of time during which the slow wave is determined to be coupled. The slow waves between two EGG channels are defined as coupled if the difference in their dominant frequencies is < 0.2 (77,78,95) or 0.5 cpm (76).

Methods to Obtain EGG Parameters

Spectral analysis methods are commonly used for calculation of the EGG parameters, including power spectral analysis and running spectral analysis (RSA) or time-frequency analysis. The frequency and power of the EGG can be derived from the power spectral density. The periodogram is one of the commonly used methods for the calculation of

the power spectrum density (73). In this method, EGG data samples are divided into consequent segments with certain overlap. A FT is performed on each data segment, and the resultant functions of all segments are averaged. The periodogram method is more appropriate for the computation of the power spectrum of a prolonged EGG recording. Whenever there are enough data samples, the periodogram method instead of the sample spectrum should be used for the calculation of the dominant frequency and power of the EGG (21). Another method to estimate the frequency and power of EGG is to use a parametric method such as autoregressive modeling (AR) parameters (19). These AR parameters are initially set at zeros and are iteratively adjusted using the EGG samples. After a certain number of iterations, the EGG signal can be represented by these AR parameters. That is, the power spectrum of the EGG signals can be calculated from these parameters (19,27). The problem is the appropriate selection of the model order or the number of parameters. Too few parameters reduce the accuracy, and too many increasing the computing time (98). Although this method is somewhat time consuming, the advantage is that compared to FFT-based methods, the period over which the signal is analyzed can be much shorter (19).

To extract not only information about the frequency of the EGG, but also information about time variations of the frequency, a running spectral analysis method using FFT was first introduced by a Dutch group (24) and later used by others (25,26,39). This method consists of a series of sequential sample spectra. It is calculated as follows: For a given data set of EGG, a time window (e.g., Hanning window) with a length of D samples is applied to the first D samples, a FFT with the same length is calculated, and a sample spectrum is obtained for the first block of data. The sample spectrum of the next time step is obtained in the same way by shifting the windows of some samples forward. The advantage of this method is easy for implantation. Its drawback is that it may not be able to provide accurate time frequency estimations when the characteristics of the EGG signal change rapidly (72,104).

To avoid the averaging effect introduced by the FFT, Chen et al.(28) developed an adaptive spectral analysis based on the adaptive autoregressive moving average model (27). The main advantage of this method is that it is able to provide the instantaneous frequency of an EGG signal with short duration (29). Thus it is very useful for the detection of gastric dysrhythmias with brief duration, but may not be a good choice for the estimation of the EGG power (104). Recently, an exponential distribution (ED) method was also introduced for representation of EGG (92). The performance of the ED method is in between the RSA and the adaptive method. The cross-terms may deteriorate the performance of the ED method if the EGG signal contains several different frequency components (104). Time–frequency analysis methods other than those mentioned above have also been used, such as wavelet transform and fast Hartley transform (89). The description of these methods is mathematically complex and beyond the scope of this article. The detailed information can be found in (5) and (89).

An example of the calculation of the EGG parameters is shown in Fig. 6. The upper panel presents an EGG recording obtained in a human subject. The power spectrum of this 30-min EGG is shown in the lower left panel. The lower right panel shows the power spectra of the 30-min EGG calculated by the adaptive spectral analysis method (28). Each line in Fig. 6c represents the power spectrum of 2-min nonoverlap data (from bottom to top). The percentage of normal slow wave and dysrhythmias can be calculated from these spectra. Of 15 spectra, 12 have peaks in the 2–4 cpm range, that is, 80% of the EGG recording has normal slow waves. Three spectra (two in the bradygastria range and one in the tachygastria range) have peaks outside the normal slow wave range. The percentage of dysrhythmias is then 20%.

EGG IN ADULTS

EGG in Healthy Subjects

Definitions of what constitutes a normal EGG have been provided by careful analysis of EGG recordings from normal volunteers (6).

Normal EGG Frequency Range. Several studies (4,41,49,68,70,71) in healthy adults have shown that EGG in the fasting state is characterized by a stable slow wave dominant frequency (DF) (median: 3.0 cpm; range: 2–4 cpm) with a relatively small amplitude. Immediately after the test meal, the EGG frequency decreases from the baseline for a short period [$\sim$5 min. (4)] and then gradually increases to above the baseline level (media: 3.2 cpm; range: 2–4 cpm). It has been shown that the postprandial EGG DF is also dependent on the type and specific qualities of the ingested test meal (99). Solid meals slightly, but significantly, increase EGG DF, whereas liquid meals temporarily reduce the EGG DF.

Based on the normal EGG frequency range of 2–4 cpm, the overall results of four studies in 189 normal subjects suggest that 70% is an appropriate lower limit of normal for the percentage of the recording time for the EGG rhythm to be in the 2.0–4.0-cpm range (4,41,49,68,70,71).

Note that the definition of normal EGG frequency range reported in the literature varies considerately (see Table 1). The Penn State group defined the percentage of normal EGG activity as the percentage of the power in the frequency range of 2.4–3.6 cpm compared to the total power from 1 to 15 cpm (84). Accordingly, dysrhythms are considered present when too much power is in the low frequency range (bradygastria) or in the high frequency range (tachygastria). This approach is debatable due to the following reasons: (*1*) The EGG is not sinusoid and thus its power spectrum contains harmonics that are related to the waveform, but not at all associated with gastric rhythmicity. In this method, however, the harmonics are considered as tachgastria (8). (*2*) The method is very sensitive to motion artifacts that can result in abnormal frequency spectra with significant power in the low frequency range (8,9). Apparently, the differences in the definitions of the normal frequency range of EGG or dysrhthmias are at least related the following two factors: (1) Relative small numbers of subjects were included in the above EGG studies; (2) different analysis methods were used to analyze the EGG

data. To establish better definitions of normal frequency rang and dysrhythmias, an international multicenter EGG study with a relative large sample size is needed and the currently used analysis methods should be applied to compare the results.

EGG Power in the Fasting and EGG Power Ratio. Absolute values of EGG power during fasting and the postprandial period are affected by a number of variables including body habitus, electrodes placement, and body position (6,105). However, these factors do not influence the relative value in EGG power, that is, the power ratio between the pre- and postprandial powers. Depending on meals consumed, 90–95% of healthy volunteers exhibit increased postprandial power at DF (6,41,71). Note that different meals may have different effects on the EGG. The main results for the effects of different meals on the EGG power are summarized as follows (99):

Water: Water induces an increase in EGG dominant power and a decrease in EGG dominant frequency. In a study with 10 normal subjects drinking 140 mL water (106), it was found that the EGG dominant frequency was slightly, but significantly, lower than the baseline in the fasting state during the first 10 min after the drink (2.95 vs. 2.73 cpm, $p < 0.05$). The power of the EGG at the dominant frequency was significantly higher after the drink than the baseline. A 3 dB increase in EGG dominant power (equivalent to 41% increase in 3 cpm amplitude) was observed. Similar observations were reported by several other investigators (66,107). In a recently performed study, simultaneous EGG and serosal recordings of gastric myoelectrical activity were made in patients before and after a drink of water (66). Statistical analysis demonstrated that the EGG dominant power change after a drink of water was correlated with that observed in the serosal recording (Spearman's correlation coefficient: $r = 0.757$, $p = 0.007$) and the change of EGG dominant frequency was the same as that from serosal recordings.

Milk: To investigate whether there is a different effect between non-nutritive (water) and a nutritive liquid meal, Chen et al.repeated the study procedure mentioned above in Ref. 106 by asking volunteers to drink 140 mL of 2% milk. The results showed that milk decreases the amplitude of EGG. The average decrease in EGG power in the 10 subjects within the first 10 min was 3.8 dB (equivalent to $\sim 50\%$ decrease in amplitude (108).

Solid meal: The effects of solid meal on the EGG have been studied by numerous investigators (5,16,33,101,106). More significant increase in EGG dominant power was observed after the solid meal than after a drink of water. For example, the average EGG dominant power after a solid meal over 10 subjects was 6 dB higher than the preprandial value, equivalent to a 100% increase in amplitude (106) (see Fig. 7a and b). The actual amount of increase in EGG dominant power is believed to be associated with the volume and content of the meal (see next section). The dominant frequency of the EGG seems to increase as well after a test meal. Similar to the change in EGG dominant frequency after a drink of water, this increase is often small, but significant (106).

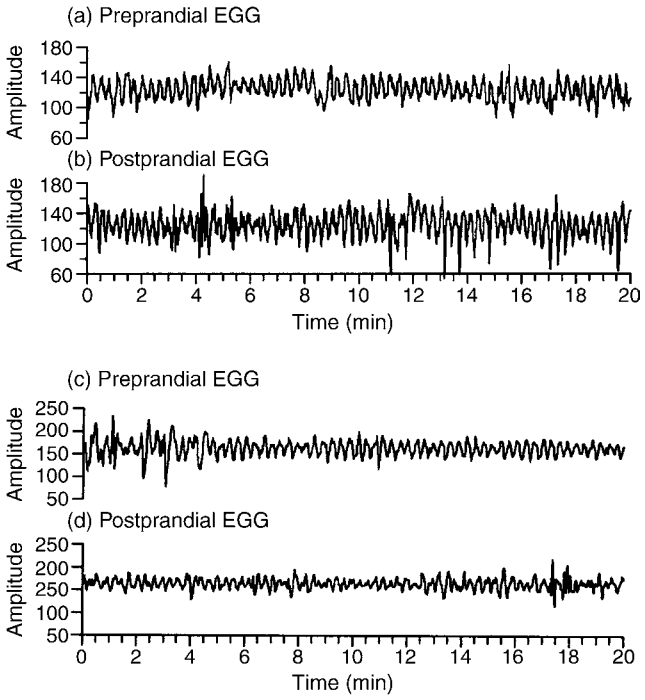


Figure 7. Pre- and postprandial EGG recordings in humans. (a) and (b): normal EGG patterns that show an increase in EGG amplitude postprandialy; (c) and (d): dysryhthmic EGG pattern and a substantial decrease in EGG amplitude after the meal.

EGG IN PATIENTS

Abnormal EGG

A variety of abnormalities have been described on EGG recordings from patients with motility disorders. For example, abnormal EGGs are noted with nausea, vomiting, early satiety, anorexia, and dyspepsia including gastroparesis (41–44,102), nonulcer or functional dyspepsia (46–50,76,109), motion sickness (5,25,110), pregnancy (38–40), and eating disorders (35). Typical EGG abnormalities in patients with motility disorders or symptoms include (*1*) absence of normal slow waves, which is shown in the EGG power spectra as a lack of peaks in the 2–4 cpm frequency range; (*2*) gastric dysrhythmias, including bradygastria, tachygastria, and arrhythmia (see Fig. 8); (*3*) deterioration of the EGG after a test meal, which is shown as a decrease in EGG power in the 2–4 cpm frequency range (see Fig. 7c and d), (*4*) slow wave uncoupling between different gastric segments detected from a multichannel EGG recording (76–78,95,96).

Definition of an abnormal EGG is mainly determined by comparison of EGG findings in healthy volunteers and symptomatic patients (41). At present, it is widely accepted that an EGG is considered as abnormal if the DF is in the tachy- and/or bradygastric frequency ranges for $> 30\%$ of the time. This number takes into account the observation that healthy volunteers exhibit periods of time representing up to 30% of recording time in which recognizable EGG rhythms are not distinguishable from background electrical noise either on visual inspection or computer analysis.

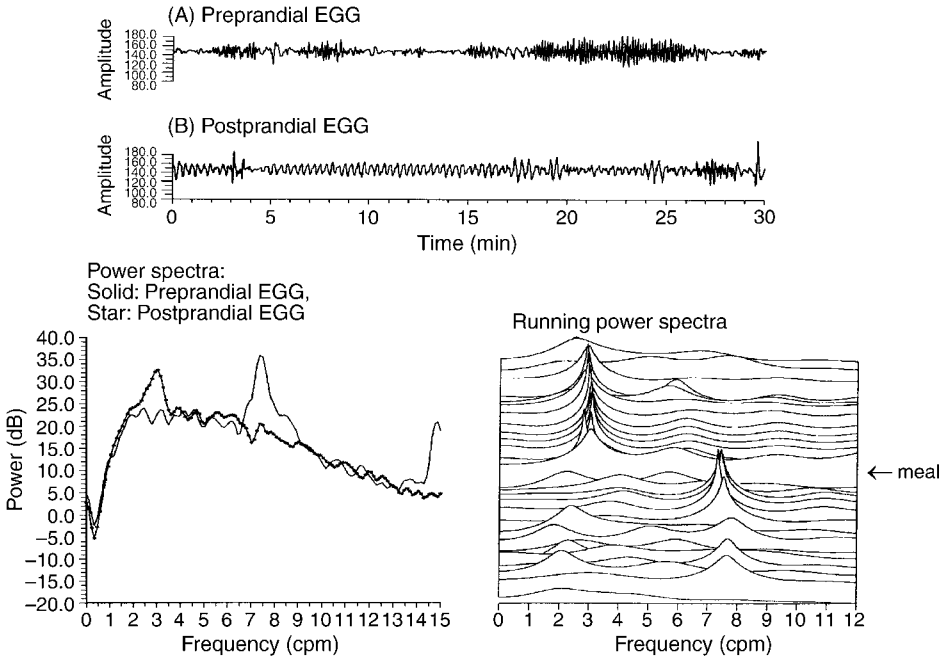


Figure 8. The EGG recordings in a patient with gastroparesis: (a) 30-min preprandial EGG recording, (b) 30-min postprandial EGG. Lower left panel: Power spectra of preprandial EGG (solid line) and postprandial EGG (line with star) shows abnormal response to a meal (decrease in postprandial EGG power) and tachgastria after meal (EGG dominant frequency: 7.4 cpm). Lower right panel shows the running spectra of 30-min preprandial EGG and 30-min postprandial EGG demonstrating the presence of 7–8 cpm tachygastrial peaks before meal and normal 3-cpm peaks after meal.

In addition, a decreased power ratio after a solid meal is also an indication of an abnormal EGG (6,7,41,68–71).

Some institutions have advocated the use of percentage distribution of EGG power in the three major frequency bands to summarize the absolute signal amplitude in the bradygastric, normal rhythm, and tachygastria ranges (2,84,111). For this parameter, an EGG is considered abnormal if the percentage distribution of total EGG power in the tachygastric range is $> 20\%$ (6,112). Power distributions in the bradygastric frequency range are highly variable and may be affected by minor variations in the signal baseline or subtle motion artifacts. Thus the calculation of the percentage of the total EGG power in the bradygastric frequency range may not be reliable for the determination of bradygastria (6).

Clinical Role of EGG. The FDA approved EGG as a test for patient evaluation in 2000. The FDA statement on EGG concluded that EGG is a noninvasive test for detecting gastric slow waves and is able to differentiate adult patients with normal myoelectrical activity from those with bradygastrias and techygastrias. The EGG can be considered as part of a comprehensive evaluation of adult patients with symptoms consistent with gastrointestinal motility disorders (6).

The members of the American Motility Society Clinical GI Motility Testing Task Force proposed the following indications for EGG as a diagnostic study to noninvasively record gastric myoelectrical activity in patients with unexplained persistent or episodic symptoms that may be related to a gastric motility and/or myoelectrical disorder (6). The EGG can be obtained: (*1*) to define gastric myoelectric disturbances in patients with nausea and vomiting unexplained by other diagnostic testing or associated with functional dyspepsia and (*2*) to characterize gastric myoelectric disturbances associated with documented gastroparesis.

The future clinical applications of EGG are in three main areas: (*1*) To assist in the clinical evaluation and diagnosis of patients with gastric motility disorders. (*2*) To determine the gastric response to either caloric stimuli or exogenous stimuli, such as pharmacologic and prokinetic agents or gastrointestinal hormones or gastric electrical stimulation or for patients before and after kidney–pancreas (KP) transplant (113–115). (*3*) To further evaluate the role of EGG in research and clinical work in infants and children (6,7,116).

EGG IN INFANTS AND CHILDREN

Although the majority of EGG studies are being performed in adults, there is an increased interest for the clinical application of EGG to pediatric patients. In infants, current diagnostic methods for the assessment of gastric motility, such as intraluminal manometry and radionuclide isotope study, are very much limited. Consequently, little is known on gastric myoelectrical in infants since mucosal–serosal recordings are not feasible, and much less information is available in infants than adults on gastric motility. The EGG is therefore an attractive noninvasive alternative to study gastric myoelectrical and motor activities in infants and children. In recent years, a small number of pediatric gastroenterologists and researchers, including Peter Milla, Alberto Ravelli, Salvatore Cucchiara, Giuseppe Riezzo, and Jiande Chen, et al. have began to use the EGG to study the pathophysiology of gastric motility in infants and children (11).

Patterns of GMA in Healthy Pediatric Subjects with Different Ages

To investigate whether EGG patterns are associated with ages, Chen et al. (117) performed EGG studies in five groups of healthy subjects including 10 preterm newborns,

8 full-term newborns, 8 full-term infants (ages 2–6 months), 9 children (ages 4–11 years), and 9 adults. The Digitrpper EGG recorder was used to record EGG signals for 30 min before and 30 min after a test meal in each subject. Spectral analysis methods were applied to computer EGG parameters. The results showed that the percentage of 2–4 cpm slow waves was 26.6±3.9% in the preterm newborns, 30.0±4.0% in full-term newborns, 70±6.1% in 2–6-months old infants ($P < 0.001$ compared with newborns), 84.6±3.2% in 4–11-year old children ($P < 0.03$ compared with infants), and 88.9±2.2% in the adults ($P > 0.05$ compared with children). This study has shown that regular gastric slow waves (2–4 cpm) are absent at birth, present at age of 2–4 months, and well developed at the age of 4–11 years. The EGG in healthy children is similar to that in healthy adults.

Using the percentage of total EGG power in the frequency range 2.5–3.6 cpm as a measure of normal gastric slow waves, Koch et al. reported similar findings in preterm and full-term infants with ages from 3 to 50 days: a low percentage of normal gastric slow waves, no difference between preterm and fullterm infants, and no difference between fasting EGG and fed EGG (111). These studies suggest that gastric slow waves are largely absent at birth, and there is a maturing process after birth.

To study the development or maturation of gastric slow waves in preterm infants, Liang et al. (118) performed a follow-up EGG study in 19 healthy preterm infants at postnatal ages of 1 and 2 weeks and 1, 2, 4, and 6 months (gestational age at birth: 33.5 ± 2.6 week). The results showed that the percentage of normal slow waves was low at birth and there was a progressive increase with age during the first 6 months of life (see Fig. 9). These results suggest that normative EGG data should be established for different age groups and age-matched controls

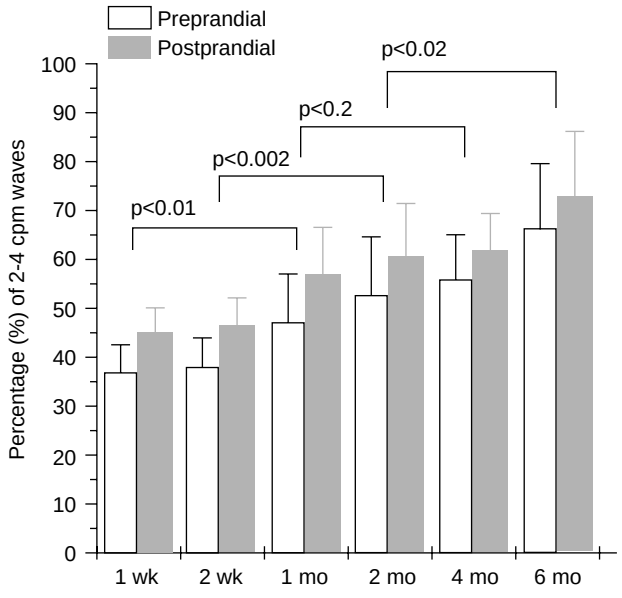


Figure 9. Preprandial and postprandial percentages of the normal 2–4-cpm slow waves in preterm infants at the ages of 1 and 2 weeks and 1, 2, 4, and 6 months. The *P* values resulted from comparison between paired fasting data or paired fed data.

are necessary for the interpretation of EGG from diseased neonates and infants.

EGG Norms in Healthy Children and Effects of Age, Gender, and BMI

As with any novel technique, the establishment of normal values is a prerequisite for reliable application across populations. Studies on EGG performed in healthy adults have found no major differences in EGG characteristics among age group (68,69,71). Using the same EGG recording and analysis system as utilized in healthy adults, Riezzo et al. performed EGG studies before and after a meal in 114 healthy children (age range: 6–12 years) (69) and Levy et al.conducted EGG studies in 55 healthy children (age range: 6–18 years) for a 1 h baseline preprandial period and a 1 h postprandial period after consummation of a standard 448 kcal meal (119). These studies have shown that the EGG patterns in the healthy children ages from 4 to 18 years are very similar to those in the healthy adults and the key normative values are not influenced by age and gender.

Applications in Pediatric Patients

Electrogastrogram has also been applied to evaluate pediatric patients with various diseases that are associated with gastric motility disorders. Functional dyspepsia presents as a challenge to clinicians with its debilitating features and no organic findings. Invasive diagnostic tests are limited in pediatric practice. Whereas, noninvisive EGG has been used and will be increasingly used to identify possible malfunctioning of the stomach, Cucchiara et al. (120) detected abnormal patterns of the EGG, encompassing all range of dysrhythmia, in 12 out of 14 patients with functional dyspepsia. Abnormalities in gastric myoelectrical activity were also observed from the EGG in pediatric patients with central nervous system disorders, chronic renal failure, and intestinal pseudoobstruction (98,121,122).

FUTURE PROSPECTS

The recording and analysis of the EGG are well established although not yet completely standardized. It is clear that the EGG is a reliable measurement of gastric slow waves and reflects relative contractile activity of the stomach. Clinical applications of the EGG are indicated in numerous studies. However, cautions should be made during recording to minimize motion artifact and in the interpretation of the EGG parameters. Future development in EGG methodology should be focused to reveal more information regarding spatial abnormalities of gastric slow waves and relevant information directly related to gastric contractility. In this regard, multichannel EGG may play a more important role in both electrophysiological studies of the stomach and clinical applications. Recently, several studies have been performed using a four-channel EGG system to derive spatial information from the multichannel EGG. These include spatial distribution of slow-wave frequency and amplitude, slow-wave coupling and propagation (75–78). The gastric slow wave propagation was measured by analyzing the phase shifts or time lags of

the EGG signal among different recording channels using cross-covariance analysis (75). Slow-wave coupling was defined as similar frequencies of the EGG signals in different channels and a cross-spectral analysis method has been established to compute the percentage of slow-wave coupling (93). Two single-center studies have suggested that patients with functional dyspepsia have impaired slow-wave propagation and coupling (76,78). Further multicenter studies are needed to determine how the expanded utility of the multichannel EGG.

The EGG is not only of great clinical significance in diagnosing gastric motility disorders, but may also play an important role in predicting and validating gastric electrical stimulation (GES) (116,123). Previous studies have established the efficacy of GES in reducing symptoms of gastroparesis. However, approximately one-third of the patients were shown to have poor response to GES therapy and the mechanism by which GES acts to improve gastroparetic symptoms is currently unclear (124,125). Physiological studies have demonstrated that gastric slow wave activity depends on the function of interstitial cells of Cajal (ICC) in the stomach (126–128). The ICC networks in the myenteric region of the gastric corpus and antrum (IC–MY) have been identified as the source of electrical slow waves that control the maximum frequency and propagation direction of gastric contractions. Ordog et al. (127) demonstrated that the loss of gastric IC–MY resulted in the loss of electrical slow waves and eliminated the ability the musculature to generate slow waves in response to depolarizing stimuli. Thus functional ICC is critical for the production and maintenance of normal slow wave activity and gastric motility. Degeneration of ICC is often found in some GI tract conditions characterized by ineffective motility of the affected segment. Recently, preliminary experiments have demonstrated (a) a normal baseline EGG is significantly correlated with an improved symptomatic outcome after chronic high frequency GES for refractory gastroparesis compared to an abnormal baseline EGG with which $\sim 50\%$ of patients were nonresponders to the high frequency GES therapy; (b) the absence of ICC was correlated with abnormal EGG and patients with an absence of ICC showed less symptom improvement (123,129). These data suggest that baseline EGG status could be used to predict long-term symptom improvement in gastroparetic patients treated with GES. Specifically, marked gastric dysrhythmias as determined by the baseline EGG study is a predictor to a poor response to high frequency GES. Further studies are required to determine the role of EGG in the management of GI motility disorders.

BIBLIOGRAPHY

Cited References

1. Alvarez WC. The electrogastrogram and what it shows. JAMA 1922;78:1116–1118.
2. Stern RM, Koch KL, Stewart WR, Vasey MW. Electrogastrography: Current issues in validation and methodology. Psychophysiology 1987;24:55–64.
3. Abell TL, Malagelada J-R. Electrogastrography: Current assessment and future perspectives. Dig Dis Sci 1988;33: 982–992.
4. Geldof H, Van Der Schee EJ. Electrogastrography: Clinical applications. Scand J Gastroenterol 1989;24(Suppl. 171): 75–82.
5. Chen JZ, McCallum RW. Electrogastrography. Principles and Applications. New York: Raven Press; 1994.
6. Parkman HP, Hasler WL, Barnett JL, Eaker EY. Electrogastrography: a document prepared by the gastric section of the American Motility Society Clinical GI Motility Testing Task Force. Neurogastroenterol Motil 2003;15:89–102.
7. Camilleri M, Hasler W, Parkman HP, Quigley EMM, Soffer E. Measurement of gastroduodenal motility in the GI laboratory. Gastroenterology 1998;115:747–762.
8. Verhagen MAM, VanSchelven L J, Samsom M, Smout AJPM. Pitfalls in the analysis of electrogastrographic recordings. Gastroenterology 1999;117:453–460.
9. Mintchev MP, Kingma YJ, Bowes KL. Accuracy of coutaneous recordings of gastricelectrical activity. Gastroenterology 1993;104:1273–1280.
10. Lorenzo CD, Reddy SN, Flores AF, Hyman PE. Is electrogastrography a substitute for manometric studies in chidren with functional gastrointestinal disirders? Dig Dis Sci 1997;42(1):2310–2316.
11. Stern RM. The history of EGG. Neurogastroenterologia 2000;1:20–26.
12. Tumpeer IH, Blitzsten PW. Registration of peristalsis by the Einthoven galvanometer. Am J Dis Child 1926;11:454–455.
13. Tumpeer IH, Phillips B. Hyperperistaltic electrographic effects. Am J Med Sci 1932;184:831–836.
14. Davis RC, Galafolo L, Gault FP. An exploration of abdominal potentials. J Com Psysiol Psychol 1957;50:519–523.
15. Davis RC, Galafolo L, Kveim K. Conditions associated with gastrointestinal activity. J Com Physiol Psychol 1959;52: 466–475.
16. Stevens LK, Worrall N. External recording of gastric activity: the electrogastrogram. Physiol Psychol 1974;2:175–180.
17. Brown BH, Smallwood RH, Duthie HL, Stoddard CJ. Intestinal smooth muscle electrical potentials recorded from surface electrodes. Med Biol Eng Comput 1975;13:97–102.
18. Smallwood RH. Analysis of gastric electrical signals from surface electrodes using phase-lock techniques. Med Biol Eng Comput 1978;16:507–518.
19. Linkens DA, Datardina SP. Estimations of frequencies of gastrointestinal electrical rhythms using autoregressive modeling. Med Biol Eng Comput 1978;16:262–268.
20. Nelsen TS, Kohatsu S. Clinical electrogastrography and its relationship to gastric surgery. Am J Surg 1968;116:215–222.
21. Chen J, McCallum RW. Electrogastrography: measurement, analysis and prospective applications. Med Biol Eng Comput 1991;29:339–350.
22. Smout JPM, van der Schee EJ, Grashuis JL. What is measured in electrogastrography? Dig Dis Sci 1980;25:179–187.
23. Koch KL, Stern RM. The relationship between the cutaneously recorded electrogastrogram and antral contractions in men. In: Stern RM, Koch KL, editors. Electrogastrography. New York: Praeger; 1985.
24. van der Schee EJ, Grashuis JT. Running spectral analysis as an aid in the representation and interpretation of electrogastrographical signal. Med Biol Eng Comput 1987;25:57–62.
25. Stern RM, Koch KL, Stewart WR, Lindblad IM. Spectral analysis of tachygastria recorded during motion sickness. Gastroenterology 1987;92:92–97.
26. Pfister CJ, Hamilton JW, Nagel N, Bass P, Webster JG, Thompkins WJ. Use of spectral analysis in the detection of frequency differences in the electrogastrograms of normal and diabetic subjects. IEEE Trans Biomed Eng 1988;BME-35:935–941.

27. Chen J. Adaptive filtering and its application in adaptive echo cancellation and in biomedical signal processing, Ph.D. dissertation, Department of Electrical Engineering. Katholieke Universiteit Leuven, Belgium; 1989.
28. Chen J, Vandewalle J, Sansen W, Vatrappen G, Jannssens J. Adaptive spectral analysis of cutaneous electrogastric signals using autoregressive moving average modelling. Med Biol Eng Comput 1990;28:531–536.
29. Chen JDZ, Stewart WR, McCallum RW. Spectral analysis of episodic rhythmic variations in the cutaneous electrogastrogram. IEEE Trans Biomed Eng 1993;BME-40:128–135.
30. Sobakin MA, Smirnov IP, Mishin LN. Electrogastrography. IRE Trans Biomed Electron 1962;BME-9:129–132.
31. Hamilton JW, Bellahsene B, Reichelderfer M, Webster JG, Bass P. Human electrogastrograms: Comparison of surface and mucosal recordings. Dig Dis Sci 1986;31:33–39.
32. Bortolotti M, Sarti P, Barbara L, Brunelli F. Gastric myoelectrical activity in patients with chronic idiopathic gastroparesis. J Gastrointest Motility 1990;2:104–108.
33. Koch KL, Stewart WR, Stern RM. Effect of barium meals on gastric electromechanical activity in man. Dig Dis Sci 1987;32:1217–1222.
34. Chen J, Richards R, McCallum RW. Frequency components of the electrogastrogram and their correlations with gastrointestinal motility. Med Biol Eng Comput 1993;31:60–67.
35. Abell TL, Malagelada J-R, Lucas AR, Brown ML, Camilleri M, Go VL, Azpiroz F, Callaway CW, Kao PC, Zinsmeister AR, et al. Gastric electromechanical and nerohorminal function in anorexia nervosa. Gastroenterology 1987;93:958–965.
36. Geldof H, van der Schee EJ, Grashuis JL. Electrogastrographic characteristics of interdigestive migrating complex in humans. Am J Physiol 1986;250:G165–171.
37. Chen JDZ, Richards RD, McCallum RW. Identification of gastric contractions from the cutaneous electrogastrogram. Am J Gastroenterol 1994;89:79–85.
38. Abell TL. Nausea and vomiting of pregnancy and the electrogastrogram: Old disease, new technology. Am J Gastroenterol 1992;87:689–681.
39. Koch KL, Stern RM, Vasey M, Botti JJ, Creasy GW, Dwyer A. Gastric dysrhythmias and nausea of pregnancy. Dig Dis Sci 1990;35:961–968.
40. Riezzo G, Pezzolla F, Darconza G, Giorgio I. Gastric myoelectrical activity in the first trimester of pregnancy: A cutaneous electrogastrographic study. Am J Gastroenterol 1992;87:702–707.
41. Chen J, McCallum RW. Gastric slow wave abnormalities in patients with gastroparesis. Am J Gastroenterol 1992;97: 477–482.
42. Chen JZ, Lin Z, Pan J, McCallum RW. Abnormal gastric myoelectrical activity and delayed gastric emptying in patients with symptoms suggestive of gastroparesis. Dig Dis Sci 1996;41(8):1538–1545.
43. Abell TL, Camilleri M, Hench VS, et al.Gastric electromechanical function and gastric emptying in diabetic gastroparesis. Eur J Gastroenterol Hepatol 1991;3:163–167.
44. Koch KL, Stern RM, Stewart WR, Dwyer AE. Gastric emptying and gastric myoelectrical activity in patients with diabetic gastroparesis: Effect of long-term domperidone treatment. Am J Gastroenterol 1989;84:1069–1075.
45. Cucchiara S, Riezzo G, Minella R, et al. Electrogastrography in non-ulcer dyspepsia. Arch Disease Childhood 1992;67(5): 613–617.
46. Parkman HP, Miller MA, Trate D, Knight LC, Urbain J-L, Maurer AH, Fisher RS. Electrogastrography and gastric emptying scintigraphy are complementary for assessment of dyspepsia. J Clin Gastroenterol 1997;24:214–219.
47. Pfaffenbach B, Adamek RJ, Bartholomaus C, Wegener M. Gastric dysrhythmias and delayed gastric emptying in patients with functional dyspepsia. Dig Dis Sci 1997;42 (10):2094–2099.
48. Lin X, Levanon D, Chen JDZ. Impaired postprandial gastric slow waves in patients with functional dyspepsia. Dig Dis Sci 1998;43(8):1678–1684.
49. Lin Z, Eaker EY, Sarosiek I, McCallum RW. Gastric myoelectrical activity and gastric emptying in patients with functional dyspepsia. Am J Gastroenterol 1999;94(9): 2384–2389.
50. Leahy A, Besherdas K, Clayman C, Mason I, Epstein O. Abnormalities of the electrogastrogram in functional dyspepsia. Am J Gastroenterol 1999;94(4):1023–1028.
51. Hongo M, Okuno Y, Nishimura N, Toyota T, Okuyama S. Electrogastrography for prediction of gastric emptying state. In: Chen JZ, McCallum RW, editors. Electrogastrography: Principles and Applications. New York: Raven Press; 1994.
52. Smallwood RH. Gastrointestinal electrical activity from surface electrodes. Ph.D. dissertation Sheffield (UK).
53. Smout AJPM. Myoelectric activity of the stomach: gastroelectromyography and electrogastrography. Ph.D. Dissertation, Erasmus Universiteit Rotterdam, Delft University Press, 1980.
54. Kingma YJ. The electrogastrogram and its analysis. Crit Rev Biomed Eng 1989;17:105–132.
55. van der Schee EJ. Electrogastrography: signal analysis aspects and interpretation. Ph.D. Dissertation, Erasmus Universiteit Rotterdam, Delft University Press, 1984.
56. Stern RM, Koch KL, editors. Electrogastrography: Methodology, Validation and Application. New York: Praeger; 1985.
57. Sarna SK, Daniel EE. Gastrointestinal electrical activity: Terminology. Gastroenterology 1975;68:1631–1635.
58. Hinder RA, Kelly KA. Human gastric pacemaker potential: Site of orgin, spread, and response to gastric transsaction and proximal gastric vagotomy. Am J Surg 1977;133:29–33.
59. Kelly KA. Motiltiy of the stomach and gastroduodenal junction. In: Johnson IA, editor. Physiology of the Gastrointestinal Tract. New York: Raven; 1981.
60. Sanders KM. A case for interstitial cells of Cajal as pacemakers and mediators of neurotransmission in the gastrointestinal tract. Gastroenterology 1996;112(2):492–445.
61. Abell TL, Malagelada J-R. Glucagon-evoked gastric dysrhythmias in humans shown by an improved electrogastrographic technique. Gastroenterology 1985;88:1932–1940.
62. Chen JDZ, Pan J, McCallum RW. Clinical significance of gastric myoelectrical dysrhythmas. Dig Dis 1995;13: 275–90.
63. Qian LW, Pasricha PJ, Chen JDZ. Origin and patterns of spontaneous and drug-induced Canine Gastric Myoelectrical Dysrhythmias. Dig Dis Sci 2003;48:508–515.
64. Familoni BO, Bowes KL, Kingma YJ, Cote KR. Can transcutaneous recordings detect gastric electrical abnormalities? Gut 1991;32:141–146.
65. Chen J, Schirmer BD, McCallum RW. Serosal and cutaneous recordings of gastric myoelectrical activity in patients with gastroparesis. Am J Physiol 1994;266:G90–G98.
66. Lin Z, Chen JDZ, Schirmer BD, McCallum RW. Postprandial response of gastric slow waves: correlation of serosal recordings with the electrogastrogram. Dig Dis Sci 2000; 45(4):645–651.
67. Riezzo G, Pezzolla F, Thouvenot J, et al. Reproducibility of cutaneous recordings of electrogasography in the fasting state in man. Pathol Biol 1992;40:889–894.
68. Pfaffenbach B, Adamek RJ, Kuhn K, Wegneer M. Electrogastrography in health subjects: Evaluation of normal values, influence of age and gender. Dig Dis Sci 1995;40: 1445–1450.

69. Riezzo G, Chiloiro M, Guerra V. Electrogastrography in health children: Evaluation of normal values, influence of age, gender and obesity. Dig Dis Sci 1998;43:1646–1651.
70. Riezzo G, Pezzolla F, Giorgio I. Effects of age and obesity on fasting gastric electrical activity in man: a cutaneous electrogastrographic study. Digestion 1991;50:176–181.
71. Parkman HP, Harris AD, Miller MA, Fisher RS. Influence of age, gender, and menstrual cycle on the normal electrogastrogram. Am J Gastroenterol 1996;91:127–133.
72. Chen JDZ, McCallum RW. Clinical application of electrogastrography. Am J Gastroenterol 1993;88:1324–1336.
73. Oppenheim AV, Schafer RW. Digital Signal Processing. New Jersey: Prentice Hall; 1975.
74. Mintchev MP, Rashev PZ, Bowes KL. Misinterpretation of human electrogastrograms related to inappropriate data conditioning and acquisition using digital computers. Dig Dis Sci 2000;45(11):2137–2144.
75. Chen JDZ, Zhou X, Lin XM, Ouyang S, Liang J. Detection of gastric slow wave propagation from the cutaneous electrogastrogram. Am J Physiol 1999;227:G424–G430.
76. Lin XM, Chen JDZ. Abnormal gastric slow waves in patients with functional dyspepsia assessed by multichannel electrogastrography. Am J Physiol 2001;280:G1370–G1375.
77. Simonian HP, Panganamamula K, Parkman HP, Xu X, Chen JZ, Lindberg G, Xu H, Shao C, Ke M-Y, Lykke M, Hansen P, Barner B, Buhl H. Multichannel electrogastrography (EGG) in normal subjects: A multicenter study. Dig Dis Sci 2004;49:594–601.
78. Simonian HP, Panganamamula K, Chen JDZ, Fisher RS, Parkman HP. Multichannel electrogastrography (EGG) in symptomatic patients: A single center study. Am J Gastroenterol 2004;99:478–485.
79. Chen JZ, Lin Z, McCallum RW. Toward ambulatory recording of electrogastrogram. In: Chen JZ, McCallum RW, editors. Electrogastrography: Principles and Applications. New York: Raven Press; 1994.
80. Mirizzi N, Scafoglieri U. Optimal direction of the electrogastrographic signal in man. Med Biol Eng Comput 1983; 21:385–389.
81. Patterson M, Rintala R, Lloyd D, Abernethy L, Houghton D, Williams J. Validation of electrode placement in neonatal electrogastrography. Dig Dis Sci 2001;40(10):2245–2249.
82. Levanon D, Zhang M, Chen JDZ. Efficiency and efficacy of the electrogastrogram. Dig Dis Sci 1998;43(5):1023–1030.
83. Smout AJPM, Jebbink HJA, Samsom M. Acquisition and analysis of electrogastrographic data. In: Chen JDZ, McCallum RW, editors. Electrogastrography: principles and applications. New York: Raven Press; 1994. p 3–30.
84. Koch KL, Stern RM. Electrogastrographic data and analysis. In: Chen JDZ, McCallum RW, editors. Electrogastrography: Principles and Applications. New York: Raven Press; 1994. p 31–44.
85. Volkers ACW, van der Schee EJ, Grashuis JL. Electrogastrography in the dog: Waveform analysis by a coherent averaging technique. Med Biol Eng Comput 1983;21:51–64.
86. Chen J. A computerized data analysis system for electrogastrogram. Comput Biol Med 1992;22:45–57.
87. Familoni BO, Kingma YJ, Bowes KL. Study of transcutaneous and intraluminal measurement of gastric electrical activity in human. Med Biol Eng Comput 1987;25:397–402.
88. Familoni BO, Kingma YJ, Bowes KL. Noninvasive assessment of huma gastric motor function. IEEE Trans Biomed Eng 1987;BME-34:30–36.
89. Mintchev MP, Bowes KL. Extracting quantitative information from digital electrogastrograms. Med Biol Eng Comput 1996;34:244–248.
90. Chen J, Vandewalle J, Sansen W, et al. Adaptive method for cancellation of respiratory artifact in electrogastric measurements. Med Biol Eng Comput 1989;27:57–63.
91. Chen JZ, Lin Z. Comparison of adaptive filtering in time-, transforem- and frequency-domain: A electrogastrographic study. Ann Biomed Eng 1994;22:423–431.
92. Lin Z, Chen JDZ. Time-frequency representation of the electrogastrogram–application of the exponential distribution. IEEE Trans Biomed Eng 1994;41(3):267–275.
93. Lin Z, Chen JDZ. Applications of feed-forward neural networks in the electrogastrograms. In: Akay M, editor. Nonlinear Biomedical Signal Processing. Piscataway (NJ): IEEE Press; 2000. p 233–255.
94. Wang ZS, He Z, Chen JDZ. Filter banks and neural network-based future extraction and automatic classification of electrogastrogram. Ann Biomed Eng 1999;27:88–95.
95. Wang ZS, Elsenbruch S, Orr WC, Chen JDZ. Detection of gastric slow wave uncoupling from multi-channel electrogastrogram: validations and applications. Neurogastroenterol Motil 2003;15:457–465.
96. Liang J, Cheung JY, Chen JDZ. Detection and deletion of motion artifacts in electrogastrogram using feature analysis and neural networks. Ann Biomed Eng 1997;25:850–857.
97. Wang ZS, Cheung JY, Chen JDZ. Blind separation of multichannel electrogastrograms using independent component analysis. Med Biol Eng Comput 1999;37:80–86.
98. Ravelli AM, Milla PJ. Electrogastrography in vomiting chidren with disorders of the central nervous system. In: Chen JDZ, McCallum RW, editors. Electrogastrography: Principles and Applications. New York: Raven Press; 1994. p 403–410.
99. Chen JDZ, McCallum RW. Electrogastrographic parameters and their clinical significance. In: Chen JDZ, McCallum RW, editors. Electrogastrography: Principles and Applications. New York: Raven Press; 1994. p 45–73.
100. Liang J, Chen JDZ. What can be measured from surface electrogastrography? Dig Dis Sci 1997;42(7):1331–1343.
101. Geldof H, van der Schee EJ, Smout AJPM. Myoelectrical activity of the stomach in gastric ulcer patients: An electrogastrographic study. J Gastrointest Motil 1989;1:122–130.
102. Koch KL, Bingaman S, Sperry N, Stern RM. Electrogastrography differentiates mechanical vs. idiopathic gastroparesis in patients with nausea and vomiting. Gastroenterology 1991;100:A99, (Abstract).
103. Xu L, Koch KL, Summy-Long J, Stern RM, Demers L, Bingaman S. Hypothalamic and gastric myoelectrical responses to vection in healthy Chinese subjects. Am J Physiol 1993;265:E578–E584.
104. Lin Z, Chen JDZ. Time-frequency analysis of the electrogastrogram. In: Akay M, editor. Time-Frequency and Wavelets in Biomedical Engineering. Piscataway, NJ: IEEE Press; 1996. 147–181.
105. Sanaka MR, Xing JH, Soffer EE. The effect of body posture on electrogastrogram. Am J Gastroenterol 2001;96:S73. (Abstract).
106. Chen J, McCallum RW. The response of electrical activity in normal human stomach to water and solid meal. Med Biol Eng Comput 1991;29:351–357.
107. Watanabe M, Shimada Y, Sakai S, Shibahara N, Matsumi H, Umeno K, Asanoi H, Terasawa K. Effects of water ingestion on gastric electrical activity and heart-rate variability in healthy human subjects. J Autonomic Nervous System 1996;58:44–50.
108. Chen J, McCallum RW. Effect of milk on myoelectrical activity in normal human stomach: An electrogastrographic study. Med Biol Eng Comput 1992;30:564–567.

109. Lin ZY, Chen JDZ, McCallum RW, Parolisi S, Shifflett J, Peura D. The prevalence of electrogastrogram abnormalities in patients with non-ulcer and H. pylori infection: results of H. pylori eradication. Dig Dis Sci 2001;46(4):739–745.
110. Stern RM, Koch KL, Leibowitz HW, Lindblad I, Shupert C, Stewart WR. Tachygastria and motion sickness. Aviat Space Environ Med 1985;56:1074–1077.
111. Koch KL, Tran TN, Stern RM, Bringaman S, et al.Gastric myoelectrical activity in premature and term infants. J Gastrointest Motil 1993;5:41–47.
112. Koch KL, Bringaman S, Tran TN, Stern RM. Visceral perceptions and gastric myoelectrical activity in healthy women and in patients with bulimia nervosa. Neurogastronetrol Motil 1998;10:3–10.
113. Hathaway DK, Abell T, Cardoso S, Heartwig MS, Gebely S, Gaber AO. Improvement in autonomic and gastric function following pancreas-kidney versus kidney-alone transplantation and the correlation with quality of life. Transplantation 1994;57:816.
114. Gaber AO, Hathaway DK, Abell T, Cardoso S, Heartwig MS, Gebely S. Improved autonomic and gastric function in pancreas-kidney vs. kidney-alone transplantation contributes to quality of life. Transplant Proc 1994;26:515.
115. Cashion AK, Holmes SL, Hathaway DK, Gaber AO. Gastroparesis following kidney/pancreas transplant. Clin Transplant 2004;18:306–311.
116. Levanon D, Chen JDZ. Electrogastrography: its role in managing gastric disorders. (Invited Review). J Pediatr Gastroenterol Nutr 1998;27:431–443.
117. Chen JDZ, Co E, Liang J, Pan J, Sutphen J, Torres-Pinedo RB, Orr WC. Patterns of gastric myoelectrical activity in human subjects of different ages. Am J Physiol 1997;272: G1022–G1027.
118. Liang J, Co E, Zhang M, Pineda J, Chen JDZ. Development of gastric slow waves in preterm infants measured by electrogastrography. Am J Physiol (Gastrointest Liver Physiol) 1998;37:G503–G508.
119. Levy J, Harris J, Chen J, Sapoznikov D, Riley B, De La Nuez W, Khaskeberg A. Electrogastrographic norms in children: toward the development of standard methods, reproducible results, and reliable normative data. J Pediatr Gastroenterol Nutr 2001;33(4):455–461.
120. Cucchiara S, Riezzo G, Minella R, et al. Electrogastrography in non-ulcer dyspepsia. Arch Disease Childhood 1992;67(5): 613–617.
121. Ravelli AM, Ledermann SE, Bisset WM, Trompeter RS, Barratt TM, Milla PJ. Gastric antral myoelectrical activity in children wit chronic renal failure. In: Chen JDZ, McCallum RW, editors. Electrogastrography: Principles and Applications. New York: Raven Press; 1994. p 411–418.
122. Devane SP, Ravelli AM, Bisset WM, Smith VV, Lake BD, Milla PJ. Gastric antral dysrhythmias in children wit chronic idiopathic intestinal pseudoobstruction. Gut 1992; 33:1477–1481.
123. Forster J, Damjanov I, Lin ZY, Sarosiek I, Wetzel P, McCallum RW. Absence of the interstitial cells of Cajal in patients with gastroparesis and correlation with clinical findings. J Gastrointest Sur 2005;9:102–108.
124. Lin ZY, Chen JDZ. Advances in electrical stimulation of the gastrointestinal tract. Crit Rev Biomed Eng 2002;30(4–6): 419–457.
125. Abell T, McCallum RW, Hocking M, Koch K, Abrahamssion H, LeBlang I, Lindberg G, Konturek J, Nowak T, Quigley EMM, Tougas G, Starkebaum W. Gastric electrical stimulation for medically refractory gastroparesis. Gastroenterology 2003;125:421–428.
126. Dickens EJ, Hirst GD, Tomita T. Identification of rhythmically active cells in guinespig stomach. J Physiol (London) 1999;514:515–531.
127. Ordog T, Ward SM, Sanders KM. Interstitial cells of Cajal generate slow waves in the murine stomach. J Physiol 1999; 518:257–269.
128. Hanani M, Freund HR. Interstitial cells of Cajal—their role in pacing and signal transmission in digestive system. Acta Physiol Scand 170;177–190.
129. Lin ZY, Sarosiek I, Forster J, McCallum RW. Association between baseline parameters of the electrogastrogram and long-term symptom improvement in gastroparetic patients treated with gastric electrical stimulation. Neurogastroenterol Motil 2003;15:345–346.

Reading List

Stern RM, Koch KL. Using electrogastrography to study motion sickness. In: Chen JDZ, McCallum RW, editors. Electrogastrography: Principles and Applications. New York: Raven Press; 1994. p 199–218.

See also GASTROINTESTINAL HEMORRHAGE; GRAPHIC RECORDERS.

ELECTROMAGNETIC FLOWMETER.

See FLOWMETERS, ELECTROMAGNETIC.

ELECTROMYOGRAPHY

CARLO DE LUCA
Boston University
Boston, Massachusetts

INTRODUCTION

Electromyography is the discipline that deals with the detection, analysis, and use of the electrical signal that emanates from contracting muscles.

This signal is referred to as the electromyographic (EMG) signal, a term that was more appropriate in the past than in the present. In days past, the only way to capture the signal for subsequent study was to obtain a "graphic" representation. Today, of course, it is possible to store the signal on magnetic tape, disks, and electronics components. Even more means will become available in the near future. This evolution has made the graphics aspect of the nomenclature a limited descriptor. Although a growing number of practitioners choose to use the term "myoelectric (ME) signal", the term "EMG" still commands dominant usage, especially in clinical environments.

An example of the EMG signal can be seen in Fig. 1. Here the signal begins with a low amplitude, which when expanded reveals the individual action potentials associated with the contractile activity of individual (or a small group) of muscle fibers. As the force output of the muscle contraction increases, more muscle fivers are activated and the firing rate of the fibers increases. Correspondingly, the amplitude of the signal increases taking on the appearance and characteristics of a Gaussian distributed variable.

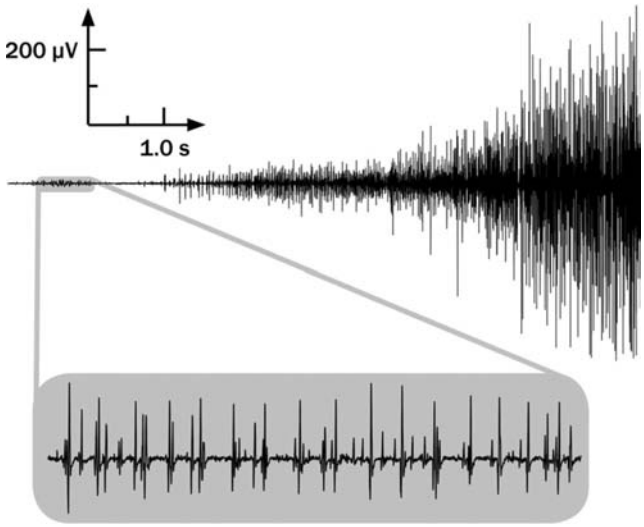


Figure 1. The EMG signal recorded with surface electrodes located on the skin above the first dorsal interosseous muscle in the hand. The signal increases in amplitude as the force produced by the muscle increases.

The novice in this field may well ask, why study electromyography? Why bother understanding the EMG signal? There are many and varied reasons for doing so. Even a superficial acquaintance with the scientific literature will uncover various current applications in fields such as neurophysiology, kinesiology, motor control, psychology, rehabilitation medicine, and biomedical engineering. Although the state of the art provides a sound and rich complement of applications, it is the potential of future applications that generates genuine enthusiasm.

HISTORICAL PERSPECTIVE

Electromyography had its earliest roots in the custom practiced by the Greeks of using electric eels to "shock" ailments out of the body. The origin of the shock that accompanied this earliest detection and application of the EMG signal was not appreciated until 1666 when an Italian, Francesco Redi, realized that it originated from muscle tissue (1). This relationship was later proved by Luigi Galvani (2) in 1791 who staunchly defended the notion. During the ensuing six decades, a few investigators dabbled with this newly discovered phenomenon, but it remained for DuBois Reymond (3) in 1849 to prove that the EMG signal could be detected from human muscle during a voluntary contraction. This pivotal discovery remained untapped for eight decades awaiting the development of technological implements to exploit its prospects. This interval brought forth new instruments such as the cathode ray tube, vacuum tube amplifiers, metal electrodes, and the revolutionary needle electrode which provided means for conveniently detecting the EMG signal. This simple implement introduced by Adrian and Bronk (4) in 1929 fired the imagination of many clinical researchers who embraced electromyography as an essential resource for diagnostic procedures. Noteworthy among these was the contribution of Buchthal and his associates.

Guided by the work of Inman et al. (5), in the mid-1940s to the mid-1950s several investigations revealed a monotonic relationship between the amplitude of the EMG signal and the force and velocity of a muscle contraction. This significant finding had a considerable impact: It dramatically popularized the use of electromyographic studies concerned with muscle function, motor control, and kinesiology. Kinesiological investigations received yet another impetus in the early 1960s with the introduction of wire electrodes. The properties of the wire electrode were diligently exploited by Basmajian and his associates during the next two decades.

In the early 1960s, another dramatic evolution occurred in the field: myoelectric control of externally powered prostheses. During this period, engineers from several countries developed externally powered upper limb prostheses that were made possible by the miniaturization of electronics components and the development of lighter, more compact batteries that could be carried by amputees. Noteworthy among the developments of externally powered prostheses was the work of the Yugoslavian engineer Tomovic and the Russian engineer Kobrinski, who in the late 1950s and early 1960s provided the first examples of such devices.

In the following decade, a formal theoretical basis for electromyography began to evolve. Up to this time, all knowledge in the field had evolved from empirical and often anecdotal observations. De Luca (6,7) described a mathematical model that explained many properties of the time domain parameters of the EMG signal, and Lindstrom (8) described a mathematical model that explained many properties of the frequency domain parameters of the EMG signal. With the introduction of analytical and simulation techniques, new approaches to the processing of the EMG signal surfaced. Of particular importance was the work of Graupe and Cline (9), who employed the autoregressive moving average technique for extracting information from the signal.

The late 1970s and early 1980s saw the use of sophisticated computer algorithms and communication theory to decompose the EMG signal into the individual electrical activities of the muscle fibers (10–12). Today, the decomposition approach promises to revolutionize clinical electromyography and to provide a powerful tool for investigating the detailed control schemes used by the nervous system to produce muscle contractions. In the same vein, the use of a thin tungsten wire electrode for detecting the action potential from single fibers was popularized for clinical applications (13,14). Other techniques using the surface EMG signal, such as the use of median and mean frequencies of the EMG signal to describe the functional state of a muscle and the use of the conduction velocity of the EMG signal to provide information on the morphology of the muscle fibers began to take hold. For a review, see De Luca (15).

The 1990s saw the effective application of modern signal processing techniques for the analysis and use of the EMG signal. Some examples are the use of time and frequency analysis of the surface EMG signal for measuring the relative contribution of low back muscles during the presence and absence of low back pain (16); the use of

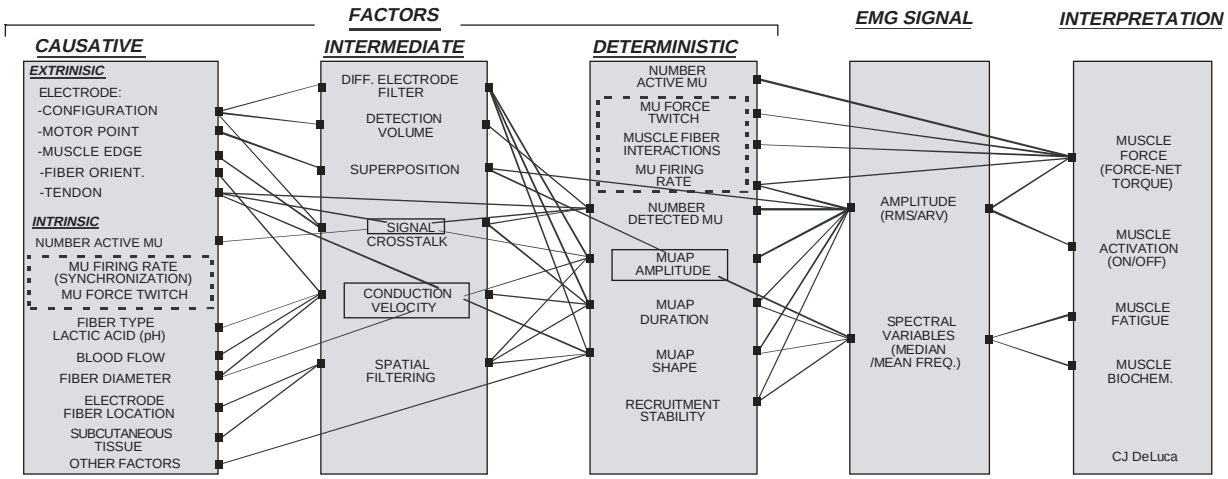


Figure 2. Relationship among the various factors that affect the EMG signal. [Reproduced with Permission from C. J. De Luca, "The Use of Surface Electromyography in Biomechanics." In the Journal of Applied Biomechanics, Vol. 13(No 2): p 139, Fig. 1.]

systematic measurements of the muscle fiber conduction velocity for measuring the severity of the Duchenne Dystrophy (17); the analysis of motor unit action potential delay for locating the origin, the ending and the innervation zone of muscle fibers (18); and the application of time–frequency analysis of the EMG signal to the field of laryngology (19).

New and exciting developments are on the horizon. For example, the use of large-scale multichannel detection of EMG signals for locating sources of muscle fiber abnormality (20); application of neural networks to provide greater degrees of freedom for the control of myoelectric prostheses (21), and for the analysis of EMG sensors data for assessing the motor activities and performance of sound subjects (22) and Stroke patients (23). Yet another interesting development is the emerging use of sophisticated Artificial Intelligence techniques for the decomposing the EMG signal (24). The reader who is interested in more historical and factual details is referred to the book *Muscles Alive* (25).

DESCRIPTION OF THE EMG SIGNAL

The EMG signal is the electrical manifestation of the neuromuscular activation associated with a contracting muscle. The signal represents the current generated by the ionic flow across the membrane of the muscle fibers that propagates through the intervening tissues to reach the detection surface of an electrode located in the environment. It is a complicated signal that is affected by the anatomical and physiological properties of muscles and the control scheme of the nervous system, as well as the characteristics of the instrumentation used to detect and observe it. Some of the complexity is presented in Fig. 2 that depicts a schematic diagram of the main physiological, anatomical and biochemical factors that affect the EMG signal. The connecting lines in the diagram show the interaction among three classes of factors that influence the EMG signal. The causative factors have a basic or elemental effect on the signal. The intermediate factors represent physical and physiological phenomena that are influenced by one or more of the causative factors and in turn influence the deterministic factors that represent physical characteristics of the action potentials. For further details see De Luca (26).

In order to understand the EMG signal, it is necessary to appreciate some fundamental aspects of physiology. Muscle fibers are innervated in groups called motor units, which when activated generate a motor unit action potential. The activation from the central nervous system is repeated continuously for as long as the muscle is required to generate force. This continued activation generates motor unit action potential trains. These trains from the concurrently active motor units superimpose to form the EMG signal. As the excitation from the Central Nervous System increases to generate greater force in the muscle, a greater number of motor units are activated (or recruited) and the firing rates of all the active motor units increases.

Motor Unit Action Potential

The most fundamental functional unit of a muscle is called the motor unit. It consists of an α-motoneuron and all the muscle fibers that are innervated by the motoneuron's axonal branches. The electrical signal that emanates from the activation of the muscle fibers of a motor unit that are in the detectable vicinity of an electrode is called the motor unit action potential (MUAP). This constitutes the fundamental unit of the EMG signal. A schematic representation of the genesis of a MUAP is presented in Fig. 3. Note the many factors that influence the shape of the MUAP. Some of these are (*1*) the relative geometrical relationship of the detection surfaces of the electrode and the muscles fibers of the motor unit in its vicinity; (*2*) the relative position of the detection surfaces to the innervation zone, that is, the region where the nerve branches contact the muscle fibers; (*3*) the size of the muscle fibers (because the amplitude of the individual action potential is proportional to the diameter of the fiber); and (*4*) the number of muscle fibers of an individual motor unit in the detectable vicinity of the electrode.

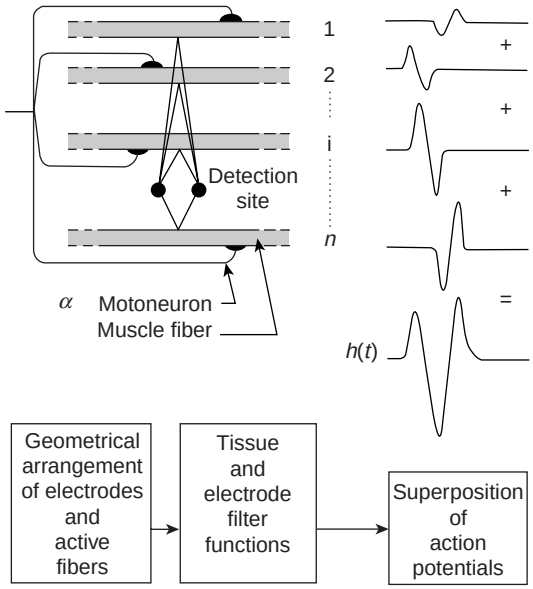


Figure 3. Schematic representation of the generation of the motor unit action potential.

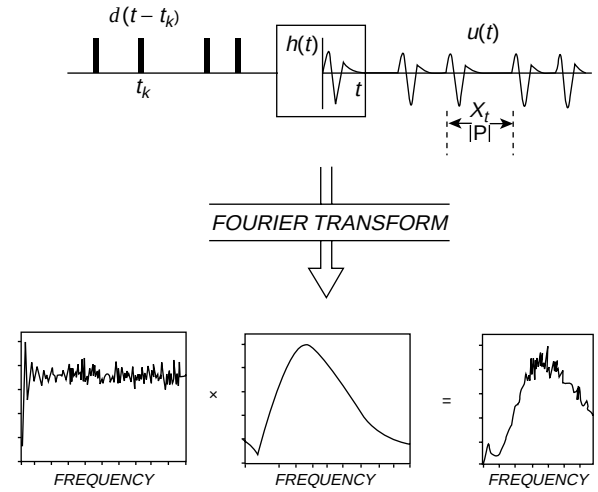


Figure 4. Model for a motor unit action potential train (MUAPT) and the corresponding Fourier transform of the interpulse intervals (IPIs), the motor unit actions potentials (MUAP), and the MUAPT.

The last two factors have particular importance in clinical applications. Considerable work has been performed to identify morphological modifications of the MUAP shape resulting from modifications in the morphology of the muscle fibers (e.g., hypertrophy and atrophy) or the motor unit (e.g., loss of muscle fibers and regeneration of axons). Although usage of MUAP shape analysis is common practice among neurologists, interpretation of the results is not always straightforward and relies heavily on the experience and disposition of the observer.

Motor Unit Action Potential Train

The electrical manifestation of a MUAP is accompanied by a contractile twitch of the muscle fibers. To sustain a muscle contraction, the motor units must be activated repeatedly. The resulting sequence of MUAPs is called a motor unit action potential train (MUAPT). The waveform of the MUAPs within a MUAPT will remain constant if the geometric relationship between the electrode and the active muscle fibers remains constant, if the properties of the recording electrode do not change, and if there are no significant biochemical changes in the muscle tissue. Biochemical changes within the muscle can affect the conduction velocity of the muscle fiber and the filtering properties of the muscle tissue.

The MUAPT may be completely described by its interpulse intervals (the time between adjacent MUAPs) and the waveform of the MUAP. Mathematically, the interpulse intervals may be expressed as a sequence of Dirac delta impulses $\delta i(t)$ convoluted with a filter $h(t)$ that represents the shape of the MUAP. Figure 4 presents a graphic representation of a model for the MUAPT. It follows that the MUAPT, $u_i(t)$ can be expressed as

$$u_i(t) = \sum_{k=1}^{n} h_i(t - t_k)$$

where

$$t_k = \sum_{l=1}^{k} x_l \quad \text{for} \quad k, l = 1, 2, 3, \ldots, n$$

In the above expression, t_k represents the time locations of the MUAPs, x represents the interpulse intervals, n is the total number of interpulse intervals in a MUAPT, and i, k, and l are integers that denote specific events.

By representing the interpulse intervals as a renewal process and restricting the MUAP shape so that it is invariant throughout the train, it is possible to derive the approximations

$$\text{Mean rectified value} = E\{|u_i(t,F)|\} \cong \lambda_i(t,F) \int_0^\infty |h_i(t)|dt$$

$$\text{Mean squared value} = MS\{|u_i(t,F)|\} \cong \lambda_i(t,F) \int_0^\infty h_i^2(t)dt$$

where F is the force generated by the muscle and is the firing rate of the motor unit.

The power density spectrum of a MUAPT was derived from the above formulation by LeFever and De Luca [(27) and independently by Lago and Jones (28)]. It can be expressed as

$$S_{u_i}(\omega,t,F) = S_{\delta_i}(\omega,t,F)|H_i(j\omega)|^2$$
$$\frac{=\lambda_i(t,F)\cdot\{1-|M(j\omega,t,F)|^2\}}{1-2\cdot \text{Real}\{M(j\omega,t,F)\}+|M(j\omega,t,F)|^2}\{|H_i(j\omega)|^2\}$$

for $\neq 0$

where is the frequency in radians per second, $H_i(\ j\omega)$ is the Fourier transform of $h_i(t)$, and $M(\ j\omega, t, F)$ is the Fourier transform of the probability distribution function, $px(x, t, F)$ of the interpulse intervals.

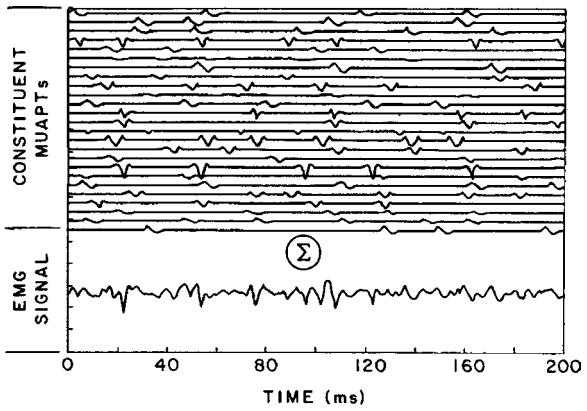


Figure 5. An EMG signal formed by adding (superimposing) 25 mathematically generated MUAPTs.

The EMG Signal

The EMG signal may be synthesized by linearly summing the MUAPTs. This approach is expressed in the equation

$$m(t,F) = \sum_{i=1}^{p} u_i(t,F)$$

and is displayed in Fig. 5, where 25 mathematically generated MUAPTs were added to yield the signal at the bottom. This composite signal bears striking similarity to the real EMG signal.

From this concept, it is possible to derive expressions for commonly used parameters: mean rectified value, root-mean-squared (rms) value, and variance of the rectified EMG signal. The interested reader is referred to *Muscles Alive* (25).

Continuing with the evolution of the model, it is possible to derive an expression for the power density spectrum of the EMG signal:

$$S_m(\omega,t,F) = R(\omega,d)\left[\sum_{i=1}^{p(F)} S_{u_i}(\omega,t) + \sum_{\substack{i,j=1 \\ i\neq j}}^{q(F)} S_{u_i u_j}(\omega,t)\right]$$

where $R(\omega,d) = K\sin^2(\omega d/2\upsilon)$ is the bipolar electrode filter function; d is the distance between detection surfaces of the electrode; is the angular frequency; υ is the conduction velocity along the muscle fibers; $S_{u_i}(\omega)$ is the power density of the MUAPT, $u_i(t)$; $S_{u_i u_j}(\omega)$ is the cross-power density spectrum MUAPTs $u_i(t)$ and $u_j(t)$; p is the total number of MUAPTs that constitute the signal; and q is the number of MUAPTs with correlated discharges.

Lindstrom (8), using a dipole model, arrived at another expression for the power density spectrum:

$$S_m(\omega,t,F) = R(\omega,d)\left[1\upsilon^2(t,F)G(\omega d2\upsilon(t,F))\right]$$

This representation explicitly denotes the interconnection between the spectrum of the EMG signal and the conduction velocity of the muscle fibers. Such a relationship is implicit in the previously presented modeling approach because any change in the conduction velocity would directly manifest itself in a change in the time duration of $h(t)$ as seen by the two detection surfaces of a stationary bipolar electrode.

ELECTRODES

Two main types of electrodes are used to detect the EMG signal: one is the surface (or skin) electrode and the other is the inserted (wire or needle) electrode. Electrodes are typically used singularly or in pairs. These configurations are referred to as monopolar and bipolar, respectively.

Surface Electrodes

There are two categories of surface electrode: passive and active. Passive electrode consists of conductive (usually metal) detection surface that senses the current on the skin through its skin electrode interface. Active electrodes contain a high input impedance electronics amplifier in the same housing as the detection surfaces. This arrangement renders it less sensitive to the impedance (and therefore quality) of the electrode–skin interface. The current trend is towards active electrodes.

The simplest form of passive electrode consists of silver disks that adhere to the skin. Electrical contact is greatly improved by introducing a conductive gel or paste between the electrode and skin. The impedance can be further reduced by removing the dead surface layer of the skin along with its protective oils; this is best done by light abrasion of the skin.

The lack of chemical equilibrium at the metal electrolyte junction sets up a polarization potential that may vary with temperature fluctuations, sweat accumulation, changes in electrolyte concentration of the paste or gel, relative movement of the metal and skin, as well as the amount of current flowing into the electrode. It is important to note that the polarization potential has both a direct current (dc) and an alternating current (ac) component. The ac component is greatly reduced by providing a reversible chloride exchange interface with the metal of the electrode. Such an arrangement is found in the silver–silver chloride electrodes. This type of electrode has become highly popular in electromyography because of its light mass (0.25 g), small size (<10 mm diameter), and high reliability and durability. The dc component of the polarization potential is nullified by ac amplification when the electrodes are used in pairs. This point is elaborated upon in later sections of this article.

The active surface electrodes have been developed to eliminate the need for skin preparation and conducting medium. They are often referred to as "dry" or "pasteless" electrodes. These electrodes may be either resistively or capacitively coupled to the skin. Although the capacitively coupled electrodes have the advantage of not requiring a conductive medium, they have a higher inherent noise level. Also, these electrodes do not have long term reliability because their dielectric properties are susceptible to change with the presence of perspiration and the erosion of the dielectric substance. For these reasons, they have not yet found a place in electromyography.

An adequately large input impedance is achieved when resistance is on the order of 10 TΩ and capacitance is small

Figure 6. Examples of active surface electrode in bipolar configurations from Delsys Inc. The spacing between the bars is 10 mm, the length of the bars is 10 mm and the thickness is 1 mm. These electrodes do not require any skin preparation or conductive paste or gels.

(typically, 3 or 4 pF). The advent of modern microelectronics has made possible the construction of amplifiers housed in integrated circuitry which have the required input impedance and associated necessary characteristics. An example of such an electrode is presented in Fig. 6. This genre of electrodes was conceptualized and first constructed at the NeuroMuscular Research Laboratory at Children's Hospital Medical Center, Boston, MA in the late 1970s. They each have two detection surfaces and associated electronic circuitry within their housing.

The chief disadvantages of surface electrodes are that they can be used effectively only with superficial muscles and that they cannot be used to detect signals selectively from small muscles. In the latter case, the detection of "cross-talk" signals from other adjacent muscles becomes a concern. These limitations are often outweighed by their advantages in the following circumstances:

1. When representation of the EMG signal corresponding to a substantial part of the muscle is required.
2. In motor behavior studies, when the time of activation and the magnitude of the signal contain the required information.
3. In psychophysiological studies of general gross relaxation of tenseness, such as in biofeedback research and therapy.
4. In the detection of EMG signals for the purpose of controlling external devices such as myoelectrically controlled prostheses and other like aids for the physically disabled population.
5. In clinical environments, where a relatively simple assessment of the muscle involvement is required, for example, in physical therapy evaluations and sports medicine evaluations.
6. Where the simultaneous activity or interplay of activity is being studied in a fairly large group of muscles under conditions where palpation is impractical, for example, in the muscles of the lower limb during walking.
7. In studies of children or other individuals who object to needle insertions.

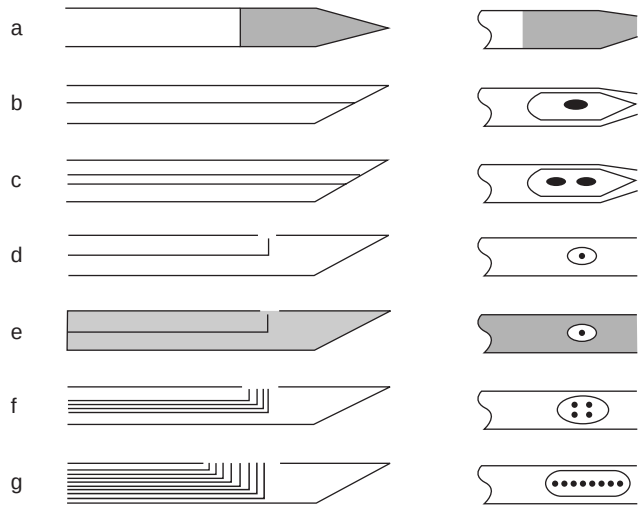


Figure 7. Examples of various needle electrodes: (a) A solid tip single-fiber electrode. If it is sufficiently thin, it can be inserted into a nerve bundle and detect neuroelectrical signals. (b) Concentric needle with one monopolar detection surface formed by the beveled cross-section of centrally located wire typically 200 μm in diameter. Commonly used in clinical practice. (c) Bipolar needle electrode with two wires exposed in cross-section, typically 100 μm in diameter. Used in clinical practice. (d) Single-fiber electrode with 25 μm diameter wire. Used to detect the activity of individual muscle fibers. (e) Macroelectrode with 25 μm diameter wire and with the cannula of the needle used as a detection surface. Used to detect the motor unit action potential from a large portion of the motor unit territory. (f) Quadrifilar planar electrode with four 50 μm wires located on the corners of a square 150 μm apart (center to center). Used for multiple channel recordings and in EMG signal decomposition technique. (g) Multifilar electrode consisting of a row of wires, generally used to study the motor unit territory.

Needle Electrodes

By far, the most common indwelling electrode is the needle electrode. A wide variety is commercially available. (see Fig. 7). The most common needle electrode is the "concentric" electrode used by clinicians. This monopolar configuration contains one insulated wire in the cannula. The tip of the wire is bare and acts as a detection surface. The bipolar configuration contains a second wire in the cannula and provides a second detection surface. The needle electrode has two main advantages. One is that its relatively small pickup area enables the electrode to detect individual MUAPs during relatively low force contractions. The other is that the electrodes may be conveniently repositioned within the muscle (after insertion) so that new tissue territories may be explored or the signal quality may be improved. These amenities have naturally led to the development of various specialized versions such as the multifilar electrode developed by Buchthal et al. (29), the planar quadrifilar electrode of De Luca and Forrest (30), the single fiber electrode of Ekstedt and Stålberg (13), and the macroelectrode of Stålberg (14). The single-fiber electrode consists of a thin, stiff, sharpened metal filament, usually made of tungsten. When inserted into a muscle it detects the action potentials of individual fibers. This electrode has

proven to be useful for neurological examinations of deinnervated muscles. Examples of these electrodes may be seen in Fig. 7.

Wire Electrodes

Since the early 1960s, this type of electrode has been popularized by Basmajian and Stecko (31). Similar electrodes that differ only in minor details of construction were developed independently at about the same time by other researchers. Wire electrodes have proved a boon to kinesiological studies because they are extremely fine, they are easily implanted and withdrawn from skeletal muscles, and they are generally less painful than needle electrodes whose cannula remains inserted in the muscle throughout the duration of the test.

Wire electrodes may be made from any small diameter, highly nonoxidizing, stiff wire with insulation. Alloys of platinum, silver, nickel, and chromium are typically used. Insulations, such as nylon, polyurethane, and Teflon, are conveniently available. The preferable alloy is 90% platinum, 10% iridium; it offers the appropriate combination of chemical inertness, mechanical strength, stiffness and economy. The Teflon and nylon insulations are preferred because they add some mechanical rigidity to the wires, making them easier to handle. The electrode is constructed by inserting two insulated fine (25–100 μm in diameter) wires through the cannula of a hypodermic needle. Approximately 1–2 mm of the distal tips of the wire is deinsulated and bent to form two staggered hooks (see Fig. 8 for completed version). The electrode is introduced into the muscle by inserting the hypodermic needle and then withdrawing it. The wires remain lodged in the muscle tissues. They may be removed by gently pulling them out: They are so pliable that the hooks straighten out on retraction.

In kinesiological studies, where the main purpose of using wire electrodes is to record a signal that is proportional to the contraction level of muscle, repositioning of the electrode is not important. But for other applications, such as recording distinguishable MUAPTs, this limitation is counterproductive. Some have used the phrase "poke and hope" to describe the standard wire electrode technique for this particular application. Another limitation of the wire electrode is its tendency to migrate after it has been inserted, especially during the first few contractions of the muscle. The migration usually stops after a few contractions. Consequently, it is recommended to perform a half dozen or so short duration contraction before the actual recording session begins.

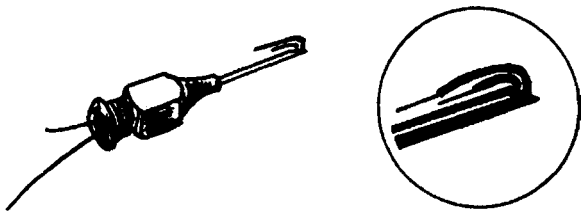


Figure 8. A bipolar wire electrode with its carrier needle used for insertion.

Electrode Maintenance

Proper usage of wire and needle electrodes requires constant surveillance of the physical and electrical characteristics of the electrode detection surfaces. Particular attention should be given to keeping the tips free of debris and oxidation. The reader is referred to the book *Muscles Alive* (25) for details on these procedures as well as suggestions for sterilization.

How to Choose the Proper Electrode

The specific type of electrode chosen to detect the EMG signal depends on the particular application and the convenience of use. The application refers to the information that is expected to be obtained from the signal; for example, obtaining individual MUAPs or the gross EMG signal reflecting the activity of many muscle fibers. The convenience aspect refers to the time and effort the investigator wishes to devote to the disposition of the subject or patient. Children, for example, are generally resistant to having needles inserted in their muscles.

The following electrode usage is recommended. The reader, however, should keep in the mind that crossover applications are always possible for specific circumstances.

Surface Electrodes

Time force relationship of EMG signals.

Kinesiological studies of surface muscles.

Neurophysiological studies of surface muscles.

Psychophysiological studies.

Interfacing an individual with external electromechanical devices.

Needle Electrode

MUAP characteristics.

Control properties of motor units (firing rate, recruitment, etc.).

Exploratory clinical electromyography.

Wire Electrodes

Kinesiological studies of deep muscles.

Neurophysiological studies of deep muscles.

Limited studies of motor unit properties.

Comfortable recording procedure from deep muscles.

Where to Locate the Electrode

The location of the electrode should be determined by three important considerations: (*1*) signal/noise ratio, (*2*) signal stability (reliability), and (*3*) cross-talk from adjacent muscles. The stability consideration addresses the issue of the modulation of the signal amplitude due to relative movement of the active fibers with respect to the detection surfaces of the electrode. The issue of cross-talk concerns the detection by the electrode of signals emanating from adjacent muscles.

For most configurations of needle electrodes, the question of cross-talk is of minor concern because the electrode

is so selective that it detects only signals from nearby muscle fibers. Because the muscle fibers of different motor units are scattered in a semirandom fashion throughout the muscle, the location of the electrode becomes irrelevant from the point of view of signal quality and information content. The stability of the signal will not necessarily be improved in any one location. Nonetheless, it is wise to steer clear of the innervation zone so as to reduce the probability of irritating a nerve ending.

All the considerations that have been discussed for needle electrodes also apply to wire electrodes. In this case, any complication will be unforgiving in that the electrode may not be relocated. Since the wire electrodes have a larger pickup area, a concern arises with respect to how the location of the insertion affects the stability of the signal. This question is even more dramatic in the case of surface electrodes.

For surface electrodes, the issue of cross-talk must be considered. Obviously, it is not wise to optimize the signal detected, only to have the detected signal unacceptably contaminated by an unwanted source. A second consideration concerns the susceptibility of the signal to the architecture of the muscle. Both the innervation zone and the tendon muscle tissue interface have been found to alter the characteristics of the signal. *It is suggested that the preferred location of an electrode is in the region halfway between the center of the innervation zone and the further tendon. See the review article by De Luca* (12) *for additional details*.

SIGNAL DETECTION: PRACTICAL CONSIDERATIONS

When attempting to collect an EMG signal, both the novice and the expert should remember that the characteristics of the observed EMG signal are a function of the apparatus used to acquire the signal as well as the electrical current that is generated by the membrane of the muscle fibers. The "distortion" of the signal as it progresses from the source to the electrode may be viewed as a filtering sequence. An overview of the major filtering effects is presented in Fig. 9. A brief summary of the pertinent facts follows. The reader interested in additional details is referred to *Muscles Alive* (25).

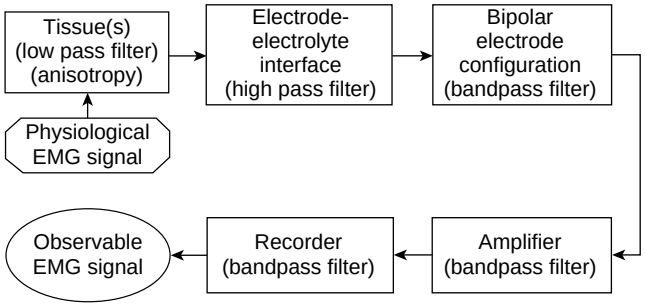


Figure 9. Block diagram of all the major aspects of the signal acquisition procedure. Note the variety of physical properties that act as filters to the EMG signal before it can be observed. The term "physiological EMG signal" refers to the collection of signals that emanate from the surface of the muscle fibers. These are not observable.

Electrode Configuration

The electrical activity inside a muscle or on the surface of the skin outside a muscle may be easily acquired by placing an electrode with only one detection surface in either environment and detecting the electrical potential at this point with respect to a "reference" electrode located in an environment that either is electrically quiet or contains electrical signals unrelated to those being detected. ("Unrelated" means that the two signals have minimal physiological and anatomical associations.) A surface electrode is commonly used as the reference electrode. Such an arrangement is called monopolar and is at times used in clinical environments because of its relative technical simplicity. A schematic arrangement of the monopolar detection configuration may be seen in Fig. 10. The monopolar configuration has the drawback that it will detect all the electrical signals in the vicinity of the detection surface; this includes unwanted signals from sources other than the muscle of interest.

The bipolar detection configuration overcomes this limitation (see Fig. 10). In this case, two surfaces are used to detect two potentials in the muscle tissue of interest each with respect to the reference electrode. The two signals are then fed to a differential amplifier which amplifies the

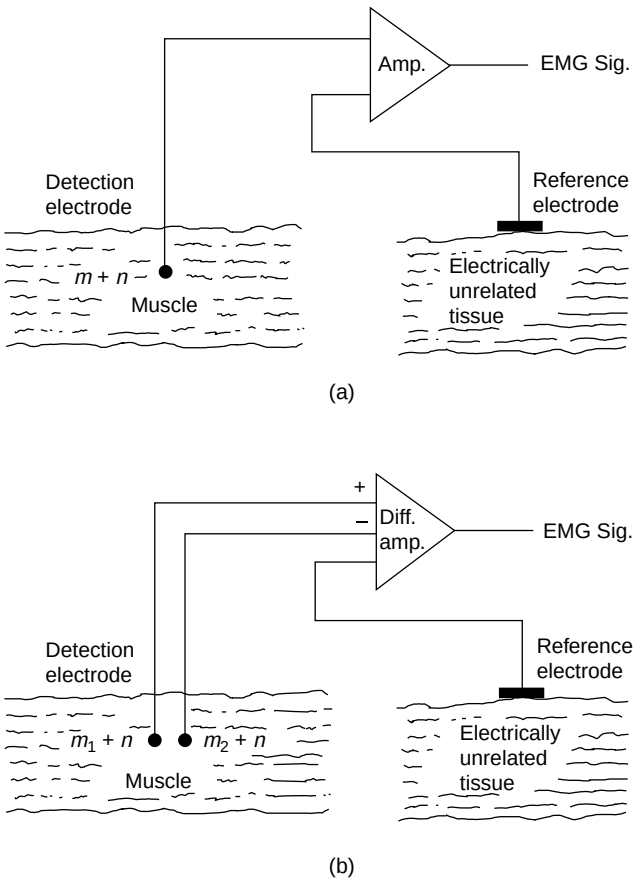


Figure 10. (a) Monopolar detection arrangement. (b) Bipolar detection arrangement. Note that in the bipolar detection arrangement, the EMG signals are considered to be different, whereas the noise is similar.

difference of the two signals, thus eliminating any "common mode" components in the two signals. Signals emanating from the muscle tissue of interest near the detection surface will be dissimilar at each detection surface because of the localized electrochemical events occurring in the contracting muscle fibers, whereas "ac noise" signals originating from a more distant source (e.g., 50 or 60 Hz electromagnetic signals radiating from power cords, outlets, and electrical devices) and "dc noise" signals (e.g., polarization potentials in the metal electrolyte junction) will be detected with an essentially similar amplitude at both detection surfaces. Therefore, they will be subtracted, but not necessarily nullified prior to being amplified. The measure bf the ability of the differential amplifier to eliminate the common mode signal is called the common mode rejection ratio.

Spatial Filtering

1. As the signal propagates through the tissues, the amplitude decreases as a function of distance. The amplitude of the EMG signal decreases to approximately 25% within 100 μm. Thus, an indwelling electrode will detect only signals from nearby muscle fibers.
2. The filtering characteristic of the muscle tissues is a function of the distance between the active muscle fibers and the detection surface(s) of the electrode. In the case of surface electrodes, the thickness of the fatty and skin tissues must also be considered. The tissues behaves as a low pass filter whose bandwidth and gain decrease as the distance increases.
3. The muscle tissue is anisotropic. Therefore, the orientation of the detection surfaces of the electrode with respect to the length of the muscle fibers is critical.

Electrode Electrolyte Interface

1. The contact layer between the metallic detection surface of the electrode and the conductive tissue forms an electrochemical junction that behaves as a high pass filter.
2. The gain and bandwidth will be a function of the area of the detection surfaces and any chemical electrical alteration of the junction.

Bipolar Electrode Configuration

1. This configuration ideally behaves as a bandpass filter; however, this is true only if the inputs to the amplifier are balanced and the filtering aspects of the electrode electrolyte junctions are equivalent.
2. A larger interdetection surface spacing will render a lower bandwidth. This aspect is particularly significant for surface electrodes.
3. The greater the interdetection surface spacing, the greater the susceptibility of the electrode to detecting measurable amplitudes of EMG signals from adjacent and deep muscles. Again, this aspect is particularly significant for surface electrodes.
4. An interdetection surface spacing of 1.0 cm is recommended for surface electrodes.

Amplifier Characteristics

1. These should be designed and/or set for values that will minimally distort the EMG signal detected by the electrodes.
2. The leads to the input of the amplifier (actually, the first stage of the amplification) should be as short as possible and should not be susceptible to movement. This may be accomplished by building the first stage of the amplifier (the preamplifier) in a small configuration which should be located near (within 10 cm) the electrode. For surface EMG amplifiers the first stage is often located in the housing of the electrodes.
3. The following are typical specifications that can be attained by modern day electronics. It is worth noting that the values below will improve as more advanced electronics components become available in the future.

(a) Common-mode input impedance: As large as possible (typically $> 10^{15}$ Ω in parallel with < 7 pF).
(b) Common mode rejection ratio: > 85 dB.
(c) Input bias current: as low as possible (typically < 5 fA).
(d) Noise (shorted inputs) < 1.5 μV rms for 20–500 Hz bandwidth.
(e) Bandwidth in hertz (3 dB points for 12 dB/octave or more rolloff):

Surface electrodes	20–500
Wire electrodes	20–2,000
Monopolar and bipolar needle electrodes for general use	20–5,000
Needle electrodes for signal decomposition	1,000–10,000
Single fiber electrode	1,000–10,000
Macroelectrode	20–5,000

An example of an eight-channel modern surface EMG amplifier is presented in Fig. 11. Such systems are

Figure 11. An eight-channel surface EMG system from Delsys Inc. The dimensions of this device (205 × 108 × 57 mm) are typical for current day units. Note that the active electrodes connect to an input unit that is separate from the body of the amplifier and can be conveniently attached to the body of the subject.

available in configurations of various channels up to 32, but 8 and 16 channel versions are most common.

Recording Characteristics

The effective or actual bandwidth of the device or algorithm that is used to record or store the signal must be greater than that of the amplifiers.

Other Considerations

1. It is preferable to have the subject, the electrode, and the recording equipment in an electromagnetically quiet environment. If all the procedures and cautions discussed in this article are followed and heeded, high quality recordings will be obtained in the electromagnetic environments found in most institutions, including hospitals.
2. In the use of indwelling electrodes, great caution should be taken to minimize (eliminate, if possible) any relative movement between the detection surfaces of the electrodes and the muscle fibers. Relative movements of 0.1 mm may dramatically alter the characteristics of the detected EMG signal and may possibly cause the electrode to detect a different motor unit population.

SIGNAL ANALYSIS TECHNIQUES

The EMG signal is a time and force (and possibly other parameters) dependent signal whose amplitude varies in a random nature above and below the zero value. Thus, simple average aging of the signal will not provide any useful information.

Rectification

A simple method that is commonly used to overcome the above restriction is to rectify the signal before performing mode pertinent analysis. The process of rectification involves the concept of rendering only positive deflections of the signal. This may be accomplished either by eliminating the negative values (half-wave rectification) or by inverting the negative values (full-wave rectification). The latter is the preferred procedure because it retains all the energy of the signal.

Averages or Means of Rectified Signals

The equivalent operation to smoothing in a digital sense is averaging. By taking the average of randomly varying values of a signal, the larger fluctuations are removed, thus achieving the same results as the analog smoothing operation. The mathematical expression for the average or mean of the rectified EMG signal is

$$|\overline{m(t)}|_{t_j - t_i} = 1t_j - t_i \int_{t_i}^{t_j} |m(t)|dt$$

where t_i and t_j are the points in time over which the integration and, hence, the averaging is performed. The shorter the time interval, the less smooth the averaged value will be.

The preceding expression will provide only one value over the time window $T = t_j - t_i$. To obtain the time varying average of a complete record of a signal, it is necessary to move the time window T duration along the record. This operation is referred to as moving average.

$$|\overline{m(t)}| = 1T \int_{t}^{t+T} |m(t)|dt$$

Like the equivalent operation in the analogue sense, this operation introduces a lag; that is, T time must pass before the value of the average of the T time interval can be obtained. In most cases, this outcome does not present a serious restriction, especially if the value of T is chosen wisely. For typical applications, values ranging from 100 to 200 ms are suggested. It should be noted that shorter time windows, T, yield less smooth time dependent average (mean) of the rectified signal.

Integration

The most commonly used and abused data reduction procedure in electromyography is integration. The literature of the past three decades is swamped with improper usage of this term, although happily within the past decade it is possible to find increasing numbers of proper usage. When applied to a procedure for processing a signal, the temp integration has a well-defined meaning that is expressed in a mathematical sense. It applies to a calculation that obtains the area under a signal or a curve. The units of this parameter are volt seconds (V·s). It is apparent that an observed EMG signal with an average value of zero will also have a total area (integrated value) of zero. Therefore, the concept of integration may be applied only to the rectified value of the EMG signal.

$$I\{|m(t)|\} = \int_{t}^{t+T} |m(t)|dt$$

Note that the operation is a subset of the procedure of obtaining the average rectified value. Since the rectified value is always positive, the integrated rectified value will increase continuously as a function of time. The only difference between the integrated rectified value and the average rectified value is that in the latter case the value is divided by T, the time over which the average is calculated. If a sufficiently long integration time T is chosen, the integrated rectified value will provide a smoothly varying measure of the signal as a function of time. There is no additional information in the integrated rectified value.

Root-Mean-Square (rms) Value

Mathematical derivations of the time and force dependent parameters indicate that the rms value provides more a more rigorous measure of the information content of the signal because it measures the energy of the signal. Its use in electromyography, however, has been sparse in the past. The recent increase is due possibly to the availability of analog chips that perform the rms operation and to the increased technical competence in electromyography. The time-varying rms value is obtained by performing the

operations described by the term in reverse order; that is,

$$\text{rms}\,\{m(t)\} = \left(1T\int_t^{t+T} m^2(t)dt\right)^{1/2}$$

This parameter is recommended above the others.

Zero Crossings and Turns Counting

This method consists of counting the number of times per unit time that the amplitude of the signal contains either a peak or crosses a zero value of the signal. It was popularized in electromyography by Williston (32). The relative ease with which these measurements could be obtained quickly made this technique popular among clinicians. Extensive clinical applications have been reported, some indicating that discrimination may be made between myopathic and normal muscle; however, such distinctions are usually drawn on a statistical basis.

This technique is not recommended for measuring the behavior of the signal as a function of force (when recruitment or derecruitment of motor units occurs) or as a function of time during a sustained contraction. Lindström et al. (33) showed that the relationship between the turns or zeros and the number of MUAPTs is linear for low level contractions. But as the contraction level increases, the additionally recruited motor units contribute MUAPTs to the EMG signal. When the signal amplitude attains the character of Gaussian random noise, the linear proportionality no longer holds.

Frequency Domain Analysis

Analysis of the EMG signal in the frequency domain involves measurements and parameters that describe specific aspects of the frequency spectrum of the signal. Fast Fourier transform techniques are commonly available and are convenient for obtaining the power density spectrum of the signal.

Three parameters of the power density spectrum may be conveniently used to provide useful measures of the spectrum. They are the median frequency, the mean frequency, and the bandwidth of the spectrum. Other parameters, such as the mode frequency and ratios of segments of the power density spectrum, have been used by some investigators, but are not considered reliable measures given the inevitably noisy nature of the spectrum. The median frequency and the mean frequency are defined by the equations:

$$\int_0^{f_{\text{med}}} S_m(f)df = \int_{f_{\text{med}}}^{\infty} S_m(f)df$$

$$f_{\text{mean}} = \int_0^f fS_m(f)df \int_0^f S_m(f)df$$

where $S_m(f)$ is the power density spectrum of the EMG signal. Stulen and De Luca (34) performed a mathematical analysis to investigate the restrictions in estimating various parameters of the power density spectrum. The median and mean frequency parameters were found to be the most reliable, and of these two the median frequency was found to be less sensitive to noise. This quality is particularly useful when a signal is obtained during low level contractions where the signal to-noise ratio may be < 6.

The above discussion on frequency spectrum parameters removes temporal information from the calculated parameters. This approach is appropriate for analyzing signals that are stationary or nearly stationary, such as those emanating from isometric, constant-force contractions. Measurement of frequency parameters during dynamic contractions requires techniques that retain the temporal information. During the past decade time–frequency analyses techniques have evolved in the field of Electromyography, as they have in the realm of other biosignals such as ECG and EEG. Early among the researchers to apply these techniques to the EMG signal were Contable et al. (35) who investigated the change in the frequency content of EMG signals during high jumps, and Roark et al. (19) who investigated the movement of the thyroarytenoid muscles during vocalization. In both these applications, the time–frequency techniques were essential because they investigated muscles that contracted dynamically and briefly.

Much of the work presented here is adapted, with permission, from Refs. 25, pp. 38, 58, 68, 74, and 81. The author thanks Williams & Wilkens for permission to extract this material.

BIBLIOGRAPHY

Cited References

1. Biederman W. Electrophysiology. 1898.
2. Galvani L. De Viribus Electricitatis. (R. Green, Transl.) London and New York: Cambridge University Press; 1953.
3. Du Bois RE. Untersuchungen uber theirische electricität. 2, 2nd P. Berlin: Verlag von G. Reimer; 1849.
4. Adrian ED, Bronk DW. J Physiol (London) 1929;67:19.
5. Inman VT, Sauders JBCM, Abbott LC J. Bone Jt Surg 1944;26:1.
6. De Luca CJ. MS [dissertation]. University of New Brunswick; 1968.
7. De Luca CJ. Biol Cybernet 1975;19:159.
8. Lindstrom LR. On the Frequency Spectrum of EMG Signals. Technical Report, Research Laboratory of Medical Electronics. Gothenburg, Sweden: Chalmers University of Technology; 1970.
9. Graupe D, Cline WK. IEEE Trans Syst Man Cybernet SMC 1975;5:252.
10. LeFever RS, De Luca CJ. Proceedings of the 8th Annual Meeting of Social Neuroscience; 1985. p 299.
11. LeFever RS, De Luca CJ. IEEE Trans Biomed Eng BME 1982;29:149.
12. McGill KC, Cummins KL, Dorfman LJ. IEEE Trans Biomed Eng 1985;32:470–477.
13. Ekstedt J, Stålberg E. In: Desmedt JE, editor. New Development EMG Clinical Neurophysiology 1. S. Karger; 1973. p 84.
14. Stålberg EJ. Neurol Neurosurg Psychiat 1980;43:475.
15. De Luca CJ. CRC Crit Rev Biomed Eng 1984;11:251–279.
16. Roy SH, De Luca CJ, Emley MC. J Rehab Res Dev 1997;34(4): 405–414.
17. Knaflitz M, Balestra G, Angelini C, Cadaldini M. Basic App Myol 1996;6(2):70,115.
18. Masuda T, Miyano H, Sadoyama T. EEG Clin Neurophysiol 1983;55(5):594–600.
19. Roark RM, Dowling EM, DeGroat RD, Watson BC, Schaefer SD. J Speech Hear Res 1995;38(2):289–303.

20. Zwarts MJ, Stegeman DF. Muscle Nerve 2003;28(1):1–17.
21. Light CM, Chappell PH, Hudgins B, Engelhart K. Med Eng Technol 2002;26(4):139–146.
22. Nawab SH, Roy SH, De Luca CJ. The 26th Int Conf IEEE Eng Med Biol Soc; San Francisco; 2004. p 979–982.
23. Roy SH, Cheng MS, De Luca CJ. Boston: ISEK Congress; 2004.
24. Nawab SH, Wotiz R, Hochstein L, De Luca CJ. Proceedings of the Second Joint Meeting of the IEEE Eng Med and Biol Soc and the Biomed Eng Soc; Houston: 2002. p 36–36.
25. Basmajian JV, De Luca CJ. Muscles Alive. 5th ed. Baltimore: Williams & Wilkins; 1985.
26. De Luca CJ. Muscle Nerve 1993;16:210–216.
27. LeFever RS, De Luca CJ. Proc Annu Conf Eng Med Biol 1976;18:56.
28. Lago P, Jones NB. Med Biol Eng Comput 1977;15:648.
29. Buchthal F, Guld C, Rosenfalck P. Acta Physiol Scand 1957;39:83.
30. De Luca CJ, Forrest WJ. IEEE Trans Biomed Eng BME 1972;19:367.
31. Basmajian JV, Stecko GA. J Appl Physiol 1962;17:849.
32. Willison RG. J Physiol (London) 1963;168:35.
33. Lindstrom LR, Broman H, Magnusson R, Petersen I. Neurophysiology 1973;7:801.
34. Stulen FB, De Luca CJ. IEEE Trans Biomed Eng BME 1981;28:515.
35. Constable R, Thornhill RJ, Carpenter RR. Biomed Sci Instrum 1994;30:69.

See also ELECTROPHYSIOLOGY; REHABILITATION AND MUSCLE TESTING.

ELECTRON MICROSCOPY.

See MICROSCOPY, ELECTRON.

ELECTRONEUROGRAPHY

THOMAS SINKJÆR
KEN YOSHIDA
WINNIE JENSEN
VEIT SCHNABEL
Aalborg University
Aalborg, Denmark

INTRODUCTION

Recording techniques developed over the past three decades have made it possible to study the peripheral and central nervous system (CNS) in detail and often during unconstrained conditions. There have been many studies using the ElectroNeuroGram (ENG) to investigate the physiology of the neuromuscular system, in particular, chronic studies in freely moving animals (1,2). Other studies relate to monitoring the state of the nerve (e.g., in relation to axotomized nerves and regeneration of nerve fibers). Clinically, the ENG is used to measure the conduction velocities and latencies in peripheral nerves by stimulating a nerve at different points along the nerve. Extracellular potentials can be recorded by either concentric needle electrodes or surface electrodes. The potentials can be derived from purely sensory nerve, from sensory components of mixed nerve, or from motor nerves (3). The study of extracellular potentials from sensory nerves in general has been shown to be of considerable value in diagnosing peripheral nerve disorders. For an indebt description of the ENG in clinical neurophysiology see Ref. 4. Several studies pertain to the use of sensory signals as feedback information to control neuroprosthetic devices. Studies (5) have shown that the application of *closed-loop* control techniques can improve the regulation of the muscle activation. Techniques using an electrical interface to nerves innervating natural sensors (6–12), such as those found in the skin, muscles, tendons, and joints are an attractive alternative to artificial sensors. Since these natural sensors are present throughout the body, remain functional after injury, and are optimally placed through evolution to provide information for natural feedback control, a rich source of information that could be used to control future FES devices exists as long as a method can be found to access them.

Interestingly, much of the peripheral sensory apparatus in spinal and brain-injured human individuals is viable. This means that the natural sensors are transmitting relevant nerve signals through the peripheral nervous system. Therefore, if the body's natural sensors are to provide a suitable feedback signal for the control of FES systems in paralyzed subjects, the challenge is to be able to extract reliable and relevant information from the nerve innervating the sensors over extended periods.

The nerve cuff electrode still has an unrivaled position as a tool for recording ENG signals from peripheral nerves in chronic experiments (13,14) and as means to provide information to be used in neural prosthesis systems (9,15,16). Other kinds of electrodes are to challenge the cuff electrode, such as intra-fascicular electrodes (6,10,18) or multisite electrodes with hundreds of contacts within a few cubic millimeters (2,10,11,17–19). These types of nerve-interface provide advantages with respect to selectivity and number of sensors.

This article describes the characteristics of the peripheral nerve signals, principals of neural recordings, and the signals obtained with different electrodes in long-term implants. An essential part in the success of recording peripheral neural signals or activating peripheral neural tissue is the neural interface.

THE PERIPHERAL NERVOUS SYSTEM

A peripheral nerve contains thousands of nerve fibers, each of them transmitting information, either from the periphery to the CNS or from the CNS to the periphery. The efferent fibers transmit information to actuators; mainly muscles, whereas afferent fibers transmit sensory information about the state of organs and events (e.g., muscle length, touch, skin temperature, joint angles, nociception, and several other modalities of sensory information). Most of the peripheral nerves contain both afferent and efferent fibers, and the peripheral nerve can thus be seen as a bidirectional information channel.

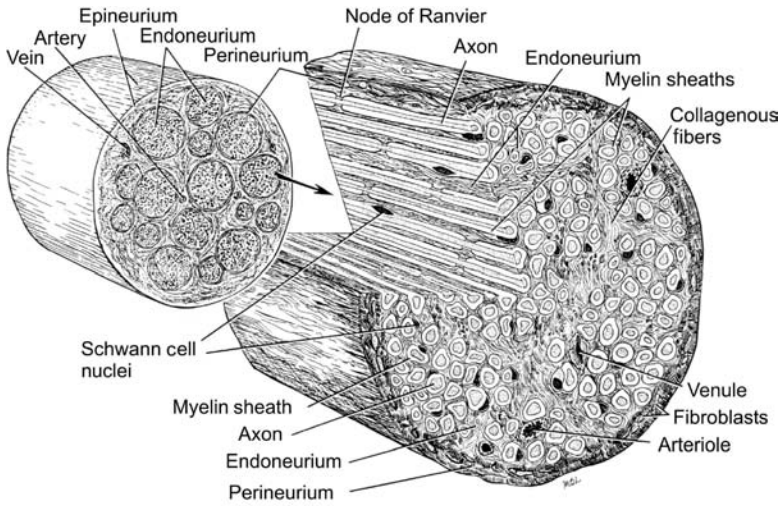


Figure 1. Figure showing the structure and components of the peripheral nerve trunk. (Adapted from Crouche, 1969, with permission.)

Anatomical Definitions and Structures

If the CNS is defined as the tissues and neural circuitry that comprise the brain and spinal cord, the peripheral nervous system can be defined as the part of the nervous system that lies outside of the CNS. Given some generalization of this definition, the peripheral nervous system consists of the spinal roots, dorsal root ganglions, axons/dendrites, support cells, and sensory end organs. The structure of the peripheral nerve trunk is illustrated below showing the following: nerve trunk, nerve fascicle, axon, schwann cells, epineurium, perineurium, and endoneurium. Though not generally considered part of the peripheral nerve, the nerve can also contain resident macrophages, leucocytes, and other cell types involved in the inflammatory response (Fig. 1).

THE NEURO-ELECTRONIC INTERFACE

The ENG can be described as the extracellular potential of an active peripheral nerve recorded at some distance from the nerve or from within the nerve trunk. The extracellular potential is formed from the contributions of the superimposed electrical fields of the active sources within the nerve. The general form of the extracellular response of a whole nerve to electrical stimulation is triphasic, it is in the lower end of the microvolt scale in amplitude, and looses both amplitude and high frequency content at larger radial distances from the nerve trunk (20).

Principles of Nerve Recordings

All cells of the body have a membrane potential. The membrane potential is a consequence of the cell membrane being selectively permeable to different ion species, resulting in different ion concentrations at the intracellular and extracellular space. The difference in ion concentrations causes an electrochemical voltage difference across the membrane, called the membrane potential. Under normal physiological conditions, the interior of the cell is negative with respect to the extracellular space without a net electric current flowing through the membrane. Therefore, if a macroscopic view of the cell membrane is justified (i.e., at distances that are large with respect to the thickness of the membrane), the existence of the resting membrane potential cannot be detected at the extracellular space.

Action Potentials. The change in membrane permeability is achieved by opening and closing of ion channels. The kinetics of the opening and closing of the channels determines the shape of the membrane current, which is characteristic for different types of nerve and muscle cells. The membrane potential changes as current flows through the membrane. The course of the membrane potential, from its resting state over an entire cycle of opening and closing of ion channels back to the initial state, is called the action potential. The associated current flowing through the membrane is the action current (Fig. 2).

A widely used model of peripheral myelinated nerve fibers is based on the model of single action potentials of fibers from rabbit sciatic nerve by Chiu et al. (22), comprising only sodium current and a leakage current. This model was adapted to 37 °C and to human peripheral nerve conduction velocity by Sweeney et al. (23).

An extensive review of different membrane models can be found in Varghese (24). More recent modeling work of human fibers includes Wesselink et al. (25) and McIntyre et al. (25).

Extracellular Currents and Potential Distribution. During the action potential, current is entering and leaving the cell through the membrane. Since there are no sources or sinks of electricity in the nerve fibers, the current flows in closed loops, and the current entering the cell at one site has to leave at a remote location. Likewise, the current leaving the cell has to reenter at a distant location, resulting in the redistribution and conduction of the current through the extracellular space. The extracellular conduction current (or, more precisely, its associated electric or magnetic field) can be detected and measured. The electroneurogram measures the electric potential field generated by the ensemble of extracellular conduction currents originating from active sites at many different fibers.

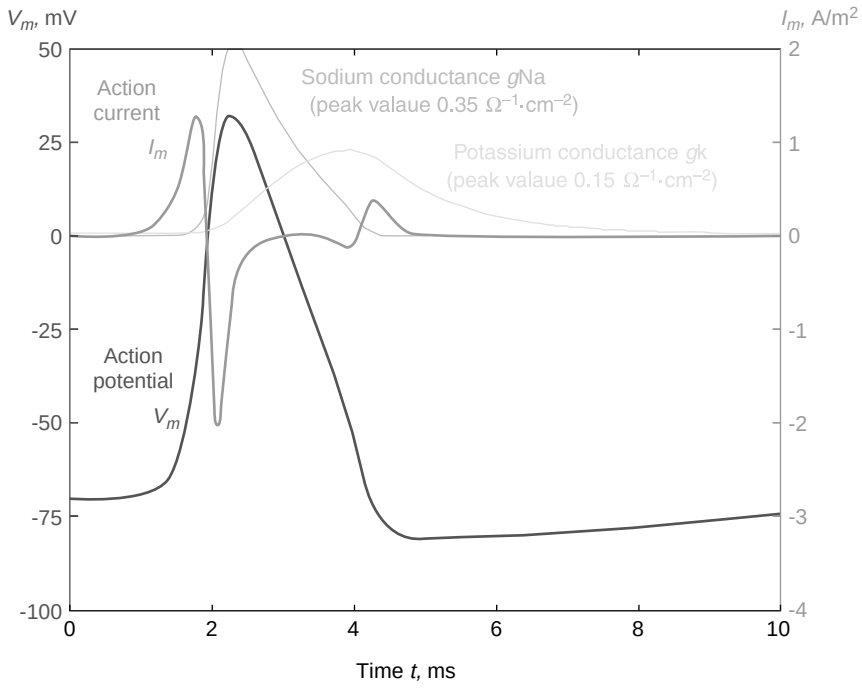


Figure 2. Action potential and action current of the squid giant axon (unmyelinated nerve fiber). Time course of the propagating action potential Vm and action current Im according to the Hodgkin and Huxley model (21). After initial depolarization, the sodium channels open, resulting in a 100-fold increase of sodium conductance, and thus inflow of sodium ions, which depolarizes the membrane potential even further. The delayed opening of the potassium channels (potassium outflow) and closing of sodium channels repolarizes the membrane potential, with a phase of hyperpolarization (below resting potential).

For the case of a single active myelinated fiber in an unbound homogeneous medium, the fiber can simply be modeled as a series of point current sources, representing the nodes of Ranvier, where current is entering and leaving the fiber (if myelin is assumed to be a perfect insulator). The extracellular potential φ of an active fiber can then easily be calculated by superposition of the electric potential fields of point current sources as follows (26):

$$\varphi(x,y,t) = \frac{1}{4\pi\sigma}\sum_n \frac{I_m(t - x_n/v)}{\sqrt{y^2 + (x - x_n)^2}}$$

where σ is the conductivity, I_m the action current (Fig. 2), v the propagation velocity, and $(x_n, 0)$ the position of the node of Ranvier n.

Figure 2 shows the extracellular potential $\varphi(0,y,t)$ of a single active myelinated fiber (10 μm) in unbound homogeneous medium for different distances y from the fiber. For very short distances from the fiber ($y = 1$ mm, corresponding to one internodal distance), the extracellular potential follows mostly the shape of the action current from the closest node of Ranvier. With increasing distance from the fiber, not only does the amplitude of the extracellular potential drop rapidly down, but also its shape changes as the potential is averaged over increasingly numbers of active nodes. The influence of fiber diameter and distance from the fiber on the amplitude and duration of the extracellular potential has been discussed in detail in Struijk (26).

The amplitude and shape of the extracellular potential are also strongly influenced by inhomogeneities in the surrounding medium, called the volume conductor (e.g., the highly resistive perineurium, surrounding tissue, or cuff electrodes). For a more realistic (complex) volume conductor than the above infinitely large, homogeneous case, numerical methods are required to calculate the distribution of the extracellular potential, such as harmonic series expansion, finite difference, or finite element methods.

Electrodes measure the extracellular potential by physically sampling the electric potential in the extracellular space. Depending on the electrode type, this yields either point-wise information (intrafascicular electrodes, needle electrodes, electrode arrays) or measurements averaged over a larger volume (cuff electrodes, surface electrodes). The configuration (monopolar, bipolar, tripolar) and spacing of the electrodes further affect the recording characteristics (e.g., noise rejection, spatial selectivity, diameter selectivity). Computer models are helpful in understanding these recording characteristics and for the development of electrodes with improved selectivity (27–29).

Neural Signals

Neural Signal Characteristics. The previous section, gives the theoretical basis of how the action potential is generated, and how this, in turn, results in a change in the electric field in the extracellular space. These potentials can be detected using electrodes in various recording configurations to maximize the neural signal detected and minimize the noise pickup. These configurations will be discussed in the next section. Here, the ENG signal is characterized. A starting point is to take recordings made by electrodes placed within the nerve fascicle, such as a longitudinal intra–fascicular electrode (LIFE).

Intrafascicular electrodes place their recording sites within the nerve fascicle, but outside of the nerve fiber. The site can be potentially adjacent to an active nerve fiber. An ENG record from an intrafascicular electrode placed in a branch of the tibial nerve in the rabbit model is shown Fig. 3. The record shows the activity from many active nerve fibers firing asynchronously from one another and superimposed upon one another.

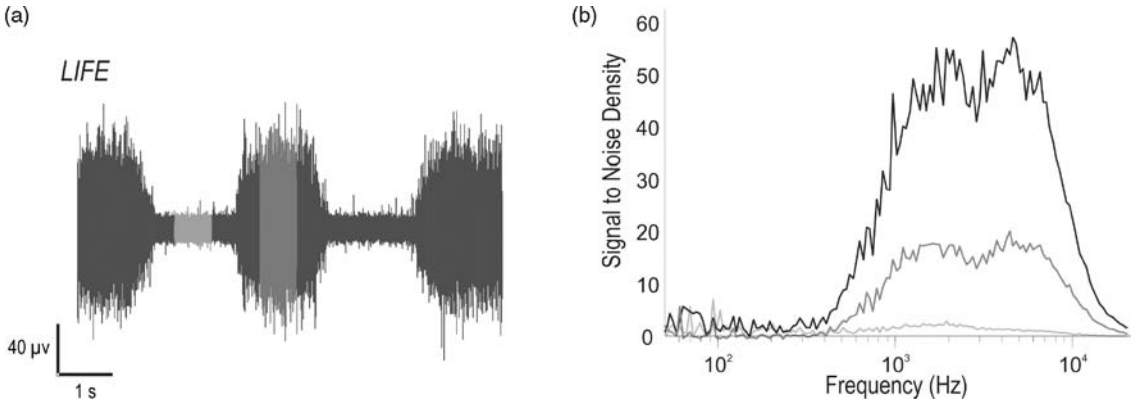


Figure 3. (a) Shows the raw record from a LIFE implanted in a medial gastrocnemius nerve branch of the tibial nerve in response to ankle flexion and extension. (b) Shows the spectral composition of the modulated LIFE activity for three different levels of activity starting from baseline, the lowest, to maximum, the highest. It illustrates that the amplitude of the spectral composition varies with the amount of activity, but the distribution remains constant.

If a closer look is taken of this multiunit activity recording, it is possible to see the spiking activity of single nerve fibers, or single unit activity. Analysis on the largest of these single spikes reveals that the amplitude of the activity is on the order of $\sim$10–100 μV_{pp}. The spectral components of intrafascicular ENG records show that, similar to extrafascicular ENG records, the activity has energy starting at ca. $\sim$100 Hz. However, unlike the unimodal extrafascicular ENG spectrum, the intrafascicular ENG spectrum is bimodal, with higher frequency components in the 5–$\sim$10 kHz range (30) (Fig. 4).

Whether the ENG signals are recorded from extrafascicular or intrafascicular electrodes, it can be seen that ENGs are low amplitude signals and require amplification of between 1000$\times$ and 100000$\times$ to bring the amplitude into the $\sim$1 V range where they can be adequately digitally sampled and stored, or recorded on analog tape. Approximate orders of magnitudes of various signals and their approximate spectral frequencies that might appear in an ENG record are shown in the Table 1.

Given the amount of amplification required and the relative amplitudes of the ambient noise that could be picked-up by the recording electrodes, recording configurations and filtering must be considered to minimize the noise pick-up and maximize the neural signals.

Recording Configurations

Components of a Recording Setup. The previous section gave the theoretical basis of how the action potential is generated and how this, in turn, results in a change in the electric field in the extracellular space. These potentials can, in turn, be detected using a combination of electrodes placed in or around the nerve, and recording instrumentation. A typical recording setup consists of the following components starting (from the input biological input end and working toward the data storage or output):

Electrodes: The active electrode(s), the ground electrode.
Differential preamplifier.
Main-amplifier and signal conditioning.
Digitization, Interpretation, and storage.
Visualization and audio monitoring.

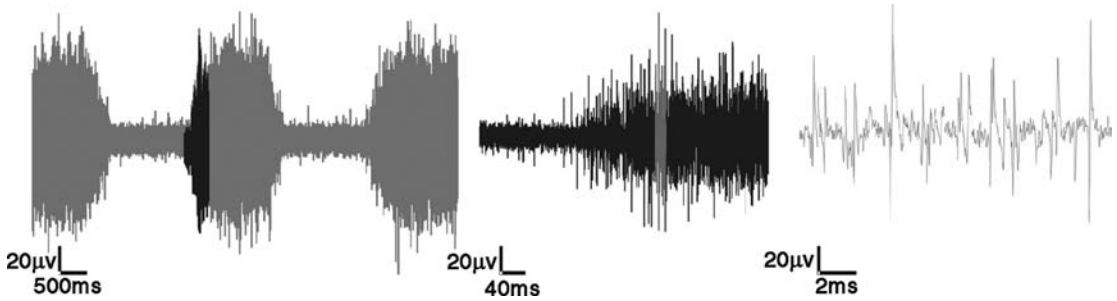


Figure 4. The figure shows an ENG recording in response to repeated mechanical activation of muscle afferents by stretching the muscle at various time scales. The amplitudes in all three traces are the same. (Left panel) Trace is the lowest time resolution and the right trace has the highest time resolution. The darker portion of the left trace is shown in the middle trace, which is zoomed in time by a factor of 12.5. The highlighted portion of the middle trace is represented in the right trace, which is zoomed in time by a factor of 20. (Right panel) Trace single spikes are now resolved, while in the other two traces only the gross mass activity can be resolved.

Table 1. Relative Amplitudes and Frequency Ranges of Components of Signals and Noise that Could Be Present in ENG Records

Source	Amplitude Order	Frequency Range
Electrochemical potential	1 V	dc[a]
Motion artefact	100 mV	Broad
Line noise	10 mV	50–60, 100–120
Thermal (electronics)	10 μV	Broad
Electrocardiogram (ECG)	1 mV	0.1–100 Hz
Electromyogram (EMG)	1 mV	1–1 kHz
Electroneurogram (ENG)	10 μV	100–10 kHz
Electroencephologram (EEG)	1 μV	dc–1 Hz

[a]Direct current = dc.

This instrumentation chain is represented schematically below in Fig. 5.

The components of the instrumentation chain can be blocked into four blocks: Preamplifier, Universal Amplifier/Filter, Computer, and Monitor. The preamplifier accepts signals from the electrode. Its function and configuration will be discussed in detail below. The Universal Amplifier/Filter block performs the analog filtering and main amplification, to reduce out of band noise, antialias and further amplify the signal and prepare the signal for the next block. The Computer block performs the digitization interpretation and depending on the application, storage of the neural data. It extracts and stores the information in the neural signal. Parallel to this is a monitoring block, which provides real-time feedback to the experimenter on the quality of the data being collected. In some cases, the monitoring block could be part of the Computer block. Neural signals are conveniently in the audio range, and the quality of the recording can be very effectively evaluated by simply listening to the signal through an audio monitor.

Differential Preamplifier. The differential preamplifier is a key component used in all measurement schemes. This component can also be called the headstage amplifier, preamplifier, or bioamplifier, and so on. In the context of biopotential recordings, the preamplifier is a low noise amplification stage used to amplify biopotentials to voltage levels in which signal conditioning can be performed using standard filters, amplifiers, discriminators, and so on. It also serves the purpose of impedance matching between the relatively high impedance electrodes at its input and the relatively low impedance input of standard instrumentation.

Differential preamplifiers can be described as an active (powered), three input (± reference), one output device. Its main function is to measure a potential at each of its two active input relative to a reference potential measured at the ground input terminal, take the difference of the measured potentials and multiply by a fixed gain. In equation form, it performs the following function:

$$v_{\text{out}} = A_2 \times [A_1(v_+ - v_{\text{ref}}) - A_1(v_- - v_{\text{ref}})]$$

A typical differential preamplifier fitted with a standard connector is shown in Fig. 6 (b), though in practice packaging, connector and form factor for different preamplifiers vary widely. Common to differential preamplifiers are the three input terminals, and one output terminal. Although they can be realized using discrete components, most differential preamplifiers are realized using low noise instrumentation amplifier chips. Since the electrodes attached to the v_+, v_-, and v_{ref} terminals of the instrumentation amplifier are physically located in different places; this equation implies that a spatial derivative is performed. It can also be seen that the voltage at the reference terminal is common to both active input terminals and should ideally be rejected along with the common mode rejection capabilities of the amplifier, measured by the common mode rejection ratio, CMRR.

One factor, which has a strong influence on the recording performance, is the impedance match of the electrodes at the active input terminals. In voltage controlled voltage source amplifiers the input impedance is much larger than the impedance of components (the electrode + tissue) preceding it so that most of the voltage drops across the input impedance of the preamplifier, and the voltage is adequately measured without attenuation. At the same time, there must be a finite dc current path from the input of the amplifier to ground to supply the required bias current to the input of the amplifier. Since there are no dc components in the extracellular nerve action potential, a commonly used scheme is to capacitively decouple the input to the amplifier. However, mismatches in the impedance of the electrode seen at each terminal of the amplifier can result in degradation of the CMR capacity of the amplifier.

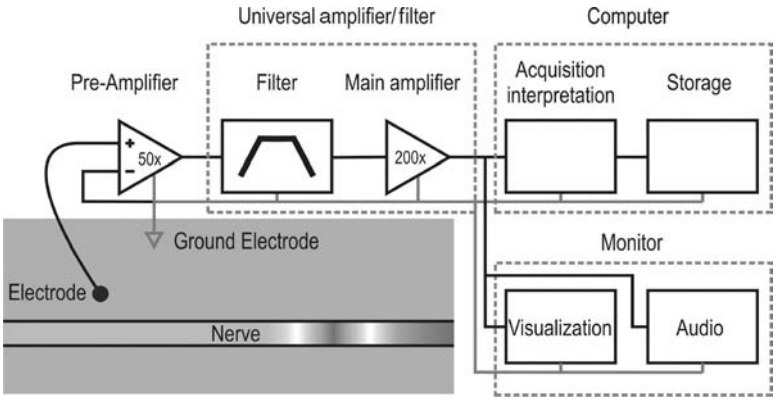


Figure 5. A generic instrumentation chain consisting of a preamplifier that impedance matches the recording electrode to the relatively low impedance filter–amplifier chain, a bandpass filter, main amplifier, data storage and monitoring. Shown also is the grounding configuration of the instrumentation attached to a ground electrode in the tissue.

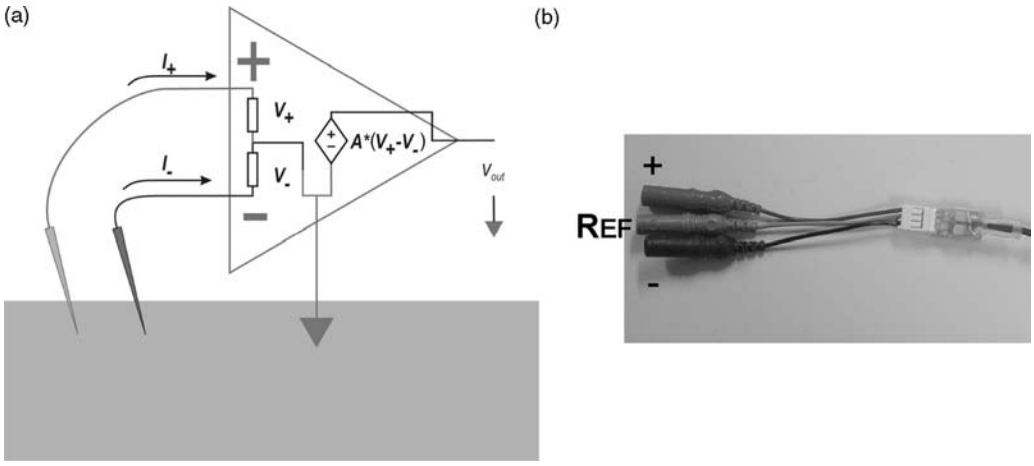


Figure 6. (a) Shows a schematic representation of the differential preamplifier and its electrode connections when used as a bioamplifier on tissues. (b) Shows a low noise differential preamplifier that could be used to record neural activity with a low impedance electrode.

Though the voltage, the controlled voltage source preamplifier scheme is the most commonly used type in current systems, current mode amplifiers, such as the application of microphone transformers have also been successfully used (31). The main limitations to the transformer amplifier scheme are the physical size of the transformer, which precludes their use in multichannel recording systems, and the necessity to carefully match the electrode impedance and transformer bandwidth.

Monopolar. Figure 7 shows the simplest recording configuration, the monopolar recording configuration where the activity recorded from the + terminal of the preamplifier is amplified relative to a common reference electrode. Though the recording, configuration is commonly used for recording EEG, *in vitro* electrophysiological preparations and implanted cortical signals, the monopolar recording configuration would be sufficient to record activity from the peripheral nerve only in the ideal noise-free case.

In practice, peripheral nerve electrodes, unlike the surface EEG electrodes on the scalp, or intracortical electrodes, are implanted and used on or in nerves that are generally surrounded by large biological noise sources, such as nearby skeletal muscles or the heart. As seen in the previous section, these bioelectrical sources have amplitudes that are orders of magnitude larger than the nerve signal, whose spectral domain can overlap that of the nerve activity making it difficult to filter the unwanted EMG activity out of the recording. Altering the recording configuration can reduce the pick-up of these unwanted activities. The monopolar recording configuration could be modified so that the reference electrode is place very close to the recording electrode, but far enough away from the nerve so that it only samples the unwanted EMG activity in the vicinity of the recording electrode.

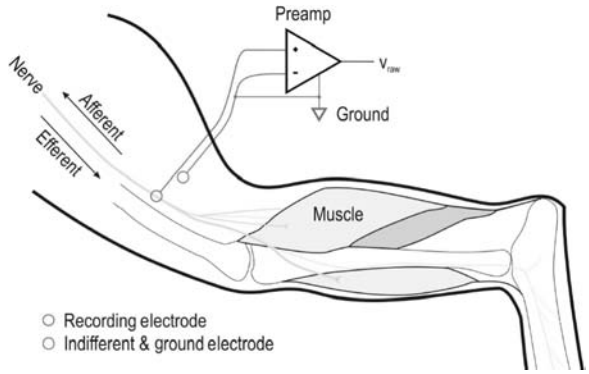


Figure 7. The monopolar recording configuration consisting of a recording electrode on or in the peripheral nerve and a common indifferent/ground electrode.

The pick-up of unwanted radio-frequency (RF) noise, such as line noise, through the lead wires between the electrode and the amplifier, is a second problem. To adequately ground the recording system, the reference must have a low impedance dc path between the tissue and the electronic ground points. By necessity, the reference electrode should have a large area. The recording electrode, on the other hand, must sample the electric potential from a small area of nerve in order not to short out the nerve activity. The relative impedance of the recording and reference electrodes is not balanced, making the RF pick-up in each of these two electrodes leads different, and not cancelled by the CMR of the differential amplifier.

Another problem with the monopolar recording configuration arises when one wants to record from more than one channel. To minimize the pick-up of local EMG activity, one might use multiple recording and ground electrodes. However, in general, it is not good practice to have multiple ground electrodes in a single recording set-up (see Fig. 12). Each ground electrode represents a low impedance electrical pathway, through which large stray ground loop currents might pass. Such currents could exist during electrical stimulation or accidental electrocution.

Bipolar. The bipolar recording configuration overcomes many of the problems of the monopolar recording configuration and is the most commonly used recording configuration for most neural electrodes excluding cuff

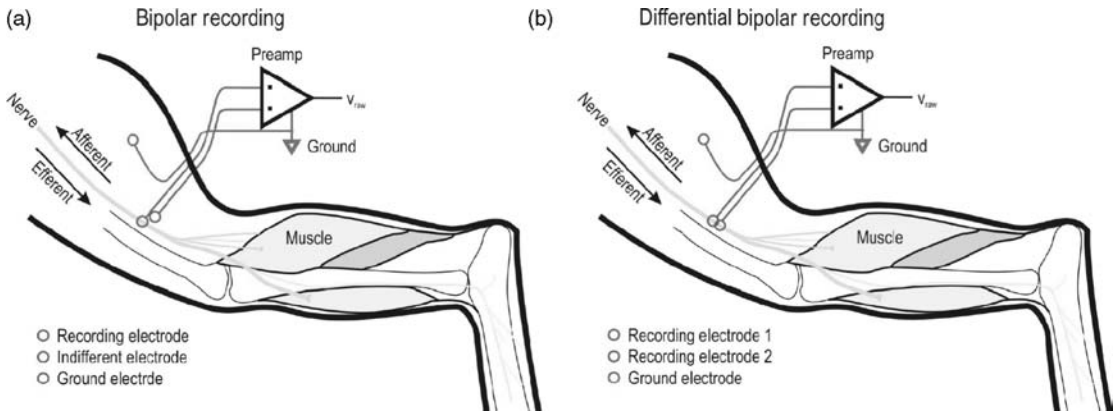


Figure 8. Bipolar recording configurations. (a) Shows a standard bipolar configuration where the signals from one recording electrode are amplified with respect to a second electrode that samples the ambient noise, the indifferent electrode. A separate ground electrode is placed somewhere in the tissue. The differential bipolar recording configuration (b) places a second recording electrode in or on the nerve and the difference of the signals from two active electrodes is amplified.

electrodes. The bipolar recording configuration requires two recording electrodes and a low impedance reference electrode. The second recording electrode is matched to the first recording electrode and is introduced to sample the local noise in the vicinity of the first recording electrode. The two recording electrodes are attached to a differential preamplifier, which amplifies the relative difference of the potentials seen by the first electrode and those seen by that of the second electrode. Thus, the local EMG activity is seen by the differential preamplifier as common mode and is rejected. Similarly, since the two recording electrodes are matched, the RF noise induced on their lead wires is similar and is rejected by the preamplifier. The second electrode effectively samples the noise, but not the neural signal, and is referred to as the indifferent electrode (Fig. 8).

A variation on the bipolar recording configuration is the differential bipolar recording configuration. In this configuration, a second active recording electrode is introduced in or on the nerve instead of an indifferent electrode reducing the total number of electrodes that needs to be implanted. Additional channels can be introduced with the addition of matched recording electrode pairs and their associated preamplifier. Each additional preamplifier shares a common reference electrode to maintain only one low impedance path to ground, and minimizing the hazards of ground loops (see below).

The neural component of the potentials recorded by the electrodes is generated by a traveling waveform, which appears at the different recording sites along the nerve at different times. Thus, there are some consequences in using the differential bipolar recording configuration to the shape of the recorded action potential. Because the action potential is a traveling waveform, measuring the action potential at two different locations along the nerve and taking the difference, $v(t_1,\boldsymbol{s_1}) - v(t_1,\boldsymbol{s_2})$, can be related to measuring the action potential measured at one location but at two different times, $v(\boldsymbol{t_1},s_1) - v(\boldsymbol{t}_2,s_1)$. This relationship assumes that the two electrodes record the action potential equally, and the distance between the recording sites are appropriate. $v(t_1,\boldsymbol{s_1}) - v(t_1,\boldsymbol{s_2})$ is $\Delta v/\Delta s$, while $v(\boldsymbol{t_1},s_1) - v(\boldsymbol{t}_2,s_2)$ is $\Delta v/\Delta t$. The relationship between the space interval, Δs, and the time interval, Δt, is the conduction velocity of the action potential, $\Delta s/\Delta t$. Therefore, the shapes of action potentials recorded using the differential bipolar configuration are roughly the derivative of the action potential recorded using the monopolar configuration. This alters the generally monophasic action potentials recorded by monopolar recordings to biphasic action potentials.

There is a dependence on the amplitude of the recorded action potential with the spacing between electrodes. The amplitude increases with electrode spacing until the spacing approaches the wavelength of the action potential. The wavelength of the action potential is dependent on the conduction velocity, which is related to the caliber of the nerve fiber. In general, distances should be kept > 1 cm for cuff electrode recordings (32), and > 2 mm in the case of differential bipolar recordings with LIFE (33).

Tripolar. The tripolar configuration is the most commonly used recording configuration with the nerve cuff electrode. The nerve cuff electrode is an extrafascicular recording configuration which places the active recording sites outside of the nerve fascicle. Since the perineurium has a relatively high impedance, most of the current generated by the nerve fiber during an action potential is restricted to within the fascicle. To maximize the signals, the nerve cuff design surrounds the nerve bundle with an insulating silicone tube that restricts whatever current leakage from the nerve bundle to the space between the nerve bundle and the insulating cuff. Let us first consider a cross sectional schematic of the electrode structure with a bipolar recording scheme as shown in Fig. 9.

One particular advantage of the cuff electrode is that the electrode is generally immune to noise sources that are lateral to the nerve and cuff electrode since they produce electrical fields that are shielded by the nerve cuff and shorted by the circumferential recording site of the electrode. However, noise sources producing electrical fields

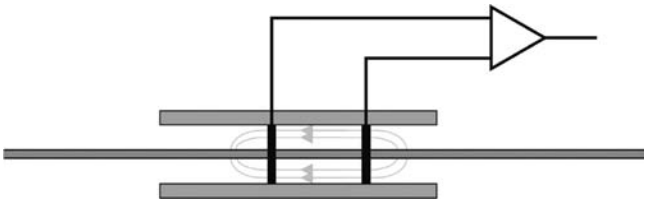


Figure 9. Bipolar differential recording scheme used with the cuff electrode. The insulating cuff restricts the current from the nerve bundle to within the cuff to maximize the voltage generated from the nerve.

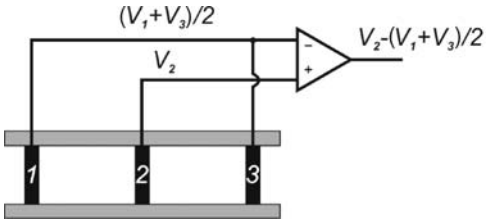


Figure 11. Tripolar recording configuration in a cuff electrode.

that are perpendicular produce a gradient through the nerve cuff, which is amplified by the bipolar recording configuration.

The tripolar recording configuration is a means to reduce or eliminate the longitudinal potential gradients of noise sources outside of the cuff. Two variants of the tripolar configuration are shown below in Fig. 10.

Longitudinal potentials from sources outside of the nerve cuff are influenced by the insulating cuff and linearlized within the cuff. The tripolar recording configuration takes advantage of this property and averages the potentials seen by the outer two electrodes to predict the extracuff noise potential seen at the central electrode, either by directly measuring and averaging the signals, as in the True Tripolar configuration, or by shorting the contacts of the outer two electrodes, as in the Pseudo-Tripolar configuration (shown below in Fig. 11).

The tripolar recording configuration introduces another differentiation operation on the shape of the action potentials recorded, and action potentials recorded with the tripolar recording configuration are triphasic.

Cabling. The cabling between the electrode active site and preamplifier is a major source of noise that can be minimized by taking care during cabling. Line noise and RF noise is induced in the leads between the electrode and the preamplifier since they act as antennae. The effect is similar to that with high impedance wired microphones. Similar strategies could be used to reduce noise pickup. Lead lengths should be kept as short as possible, though, in the case of implanted electrodes, this may not always be practical or possible since cables from multiple locations are often bundled together and routed to a common percutaneous connector to minimize the number of holes through the skin. Another strategy is to twist the differential lead pairs together so that the noise induced in each pair is nearly the same and can be reduced using the CMR of the differential pre-amplifier. When appropriate, shielded cabling could be further be used, though care should be taken to minimize ground loops when multiple shielded cables are used.

Grounding. A critical factor influencing the quality of the ENG recording is how the system is grounded. A low impedance path to ground must be established between the preamplifier and tissue to keep the input stage of the preamplifier stable and out of saturation. In acute animal preparations, the experiment is commonly performed on a conductive metal surface, which is tied to ground to a common internal ground electrode to reduce line noise. In chronic animal experiments as well as in human implanted systems, this means to reduce noise is not an option since the subject is autonomous and freely moving. In this case, a ground electrode with large surface area, such as braided stainless steel shields, can be used to establish a single ground point to the tissue. Because the ground electrode is intended to provide a low impedance pathway to a ground potential, the ground electrode

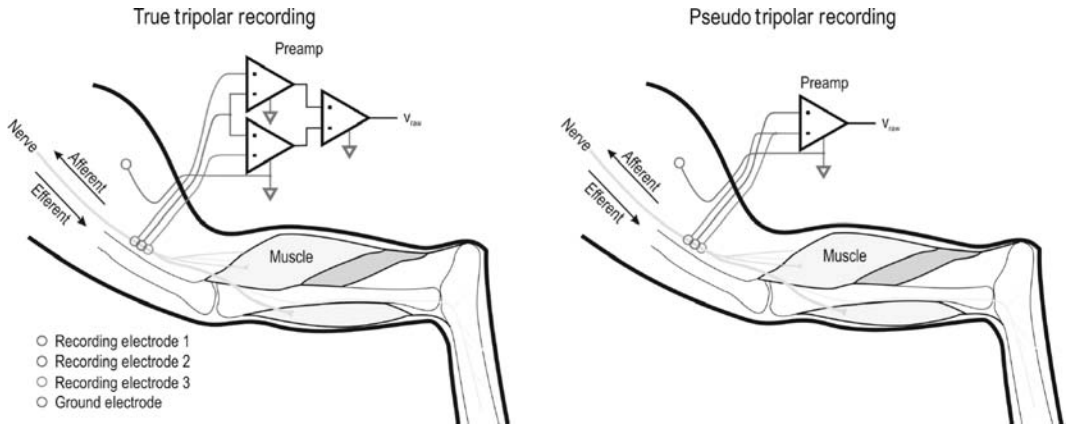


Figure 10. Tripolar recording configurations. (Left panel) Shows the true tripolar recording configuration consisting of a double differential amplification circuit. (Right panel) Shows a pseudo–tripolar recording technique, which approximates the true tripolar recording configuration by shorting the outer two electrodes 1 and 3.

provides a reference 0 potential in the tissue, and the potential of the tissues immediately surrounding the ground electrode forms a 0 isopotential. Multiple ground points to the tissue are typically not a good idea since each ground electrode would have a different chemoelectric potential even if they are made of the same material, and currents would flow from ground electrode to ground electrode to maintain a 0 isopotential at each ground point.

The issue of ground loop currents is particularly a problem when stimulation is performed at the same time as recording and if the stimulator shares the same ground with the recording circuit. Since each ground electrode is at the same potential, the current from a stimulator can flow into one ground electrode and out of another ground electrode, as shown in Fig. 11, making it nearly impossible to predict the current flow and location of stimulation. If the magnitude of stimulation is large, such as with surface stimulation, a significant portion of the current could flow through the ground electrodes since the impedance of the tissue and stimulating electrodes can be considerably higher than the impedance of the ground electrode and ground line. Current flow through the ground line can result in large artifacts and potential damage to high gain amplifiers.

One strategy to reduce ground loop currents especially during stimulation is to electrically isolate different circuits from one another. Circuits, such as the analog amplifier chain that demand low noise, are kept physically in different subcircuits from circuits that are noise-producing, such as stimulator circuits or digital circuits. If these different types of circuits must reside on a common PCB or ASIC, Parallel or Mecca grounding schemes with ground planes surrounding each circuit type can be established to prevent noise from spilling over and interfering with noise sensitive parts of the circuit (34). The strategy applied to Functional Electrical Stimulation (FES) systems where stimulation and recording sites are in close proximity to one another is to electrically isolate the stimulating circuit from the recording circuit. This makes each circuit independent from the other so that currents originating from the stimulation circuit see a high impedance pathway with no return current path in the recording system (Fig. 12).

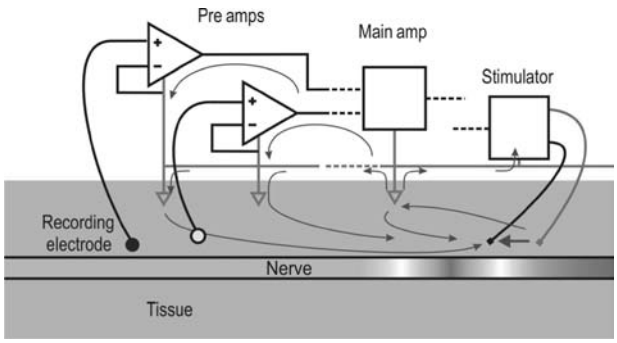


Figure 12. Illustrates potential ground loop currents that can be induced during stimulation if multiple ground electrodes (shown as triangles) are used. Since each ground electrode represents the same potential, the nearest ground electrode can draw current, which passes through the ground plane and out of the other ground electrodes, increasing the uncertainty of the point of stimulation.

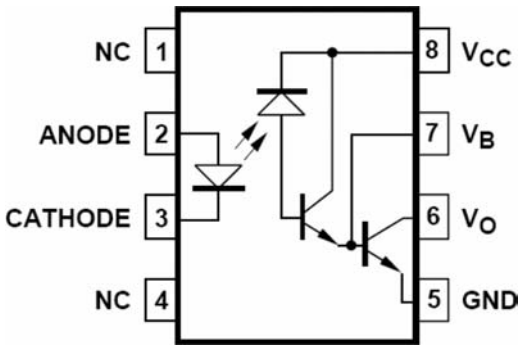


Figure 13. Schematic of a typical optoisolator using diode–diode coupling.

Optoisolators. An effective means to isolate circuits is with optoisolators. Optoisolators consist of two electrically independent halves: a transmitter, and a receiver. The basic analog optoisolator transmitter translates electrical currents into light intensity, typically through an LED. The transmitted light is detected with a phototransistor or photodiode receiver, which gates a current in the receiver that is mapped to the transmitter light intensity (Fig. 13).

If the transmitter and receiver are independently grounded and powered, this combination of optical transmitter/receiver electrically isolates the circuits at its input and output. This, of course, implies that when using optocouplers, each isolated circuit must have its own power supply and ground. The unisolated receivers can share a single power supply. A point should be made here about the number of ground electrodes in a recording set-up using multiple isolated recording/stimulation circuits. Since each isolated circuit must have its own ground, if several optoisolated recording/stimulation circuits are used, then each independent circuit must have its own single ground electrode. Ground loops between the circuits cannot form because the grounds are independent from one another, and there is no low impedance return path for ground loop currents to follow.

The optocoupler shown in the above Fig. 13 is a nonlinear device that is dependent on the coupling of the diode characteristics of the transmitter and receiver. Linear analog optocouplers also exist that linearizes the diode characteristics using feedback with a diode-transistor coupling instead of the diode–diode coupling. The diode–diode coupling, though nonlinear, has a relatively fast response time, with rise and fall times on the order of $2\,\mu s$. The diode-transistor coupling scheme, however, is bandwidth limited to considerably $< 10\,kHz$. Newer linear optocouplers attempt to overcome the bandwidth limitation by sampling and flash converting the input voltage to a digital value, transmitting the digital value and reconstructing the analogue value with a DAC (Fig. 14).

Electrode Impedance

Frequency Dependence of the Impedance. The impedance of the electrode is a function of the electrochemical properties of the electrochemical interface. Further discussion on the electrochemical interface and impedance can be

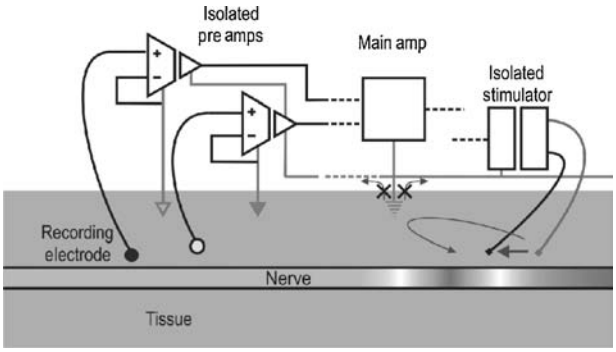


Figure 14. Use of optoisolated amplifiers and an optoisolated stimulator eliminates the ground loops shown in Fig. 8 by eliminating low impedance return paths to the tissue or stimulator circuit. Note that each isolated subcircuit has its own independent ground. Moreover, since there is no current flowing through the main ground from the tissue (shown as a dashed triangle), the main ground does not need to be grounded to the tissue.

found in other sections of this volume and in Grimnes and Martinsen (35).

A typical impedance spectrum for a platinum electrode is shown in Fig. 13. Though the absolute magnitudes of the impedance will vary depending on the total area of the electrode active site and the material used, this figure shows the general shape and characteristics of the impedance spectrum. Depending on the type of electrode used, the upper corner frequency and the lower corner frequency can be shifted. Similarly, the low frequency and high frequency segments (the segment $< 1\,\text{Hz}$, and $> 1\,\text{kHz}$, respectively, in Fig. 15) can be higher or lower than shown.

As seen earlier, most recording configurations rely on the common mode rejection ratio, CMRR, of the differential preamplifier to reduce noise pick-up and maintain stability of their neural recording. One consequence of the frequency dependence on the electrode impedance is that it influences the common mode rejection ratio. The differential preamplifier and its input can be represented by the following, where V_{cm} represents the common mode noise, which is to be rejected, and V_{dm} represents the differential nerve signal, which is to be amplified (Fig. 16).

Assuming an ideal preamplifier where the two input resistances $R_{\text{in}+}$ and $R_{\text{in}-}$ are equal, and applying the differential preamplifier equation given earlier, the common mode rejection ratio can be represented by

$$\text{CMRR} = \left|\frac{A_{\text{dm}}}{A_{\text{cm}}}\right| = \left|\frac{Z_+ + Z_- + 2R}{Z_+ - Z_-}\right|$$

In the ideal case, where the electrode and tissue impedances are matched, then the CMRR for an ideal differential preamplifier becomes infinite. However, real electrodes are relatively difficult to match, especially when dealing with small, high selectivity active site electrodes. A 10% error in electrode impedance at 1 kHz is common even with matched electrodes. Based on the CMRR equation and typical impedance shown earlier, the CMRR dependence versus, frequency can be calculated (Fig. 17).

It shows that because of the relatively high R_0 impedance at low frequency, there can be degradation in the CMRR, which results in inclusion of some of the relatively

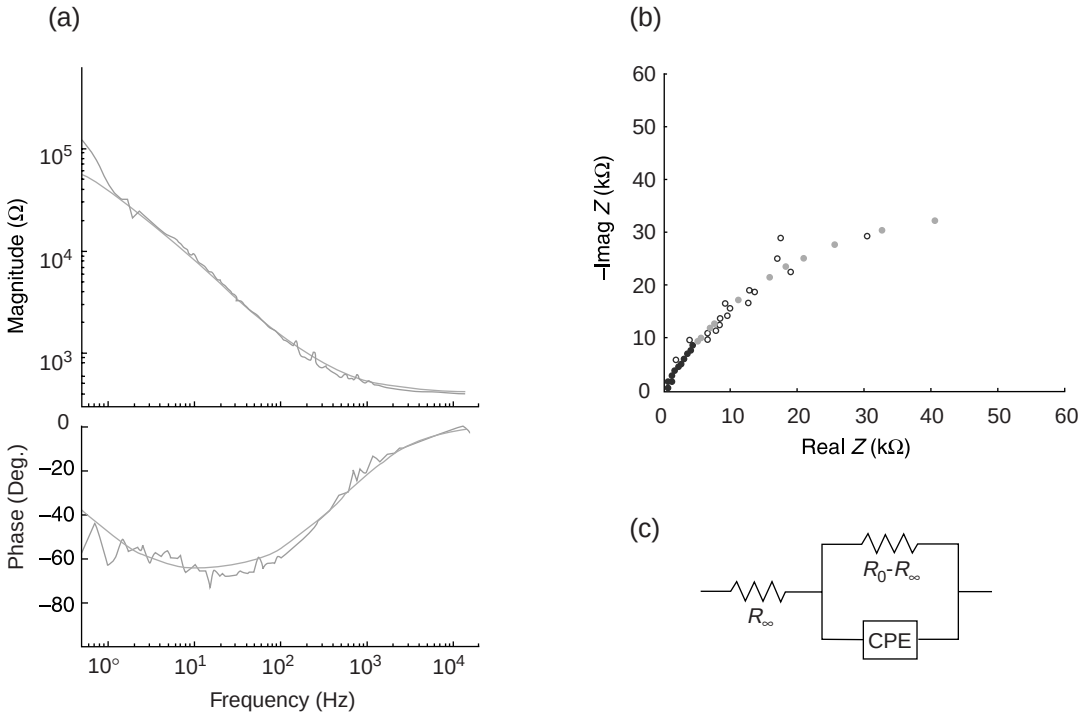


Figure 15. (a) Shows a typical metal electrode impedance spectrum and its fit to the Cole model shown in (c). (b) is the Warburg representation of the impedance spectrum where the real Z is plotted with respect to the negative of the imaginary Z.

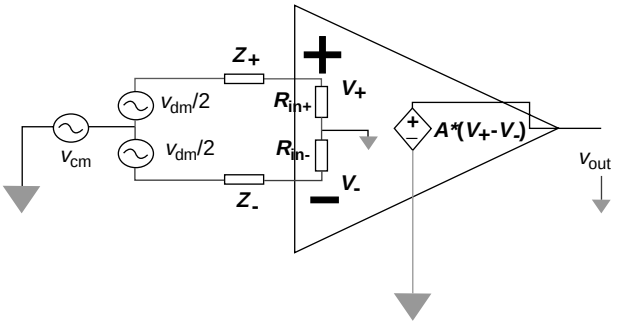


Figure 16. Shows a model of the differential preamplifier and its connections to input impedance loads, common mode, and differential mode inputs. The common mode input can represent noise that we hope to exclude from the ENG while the differential mode input represents the neural signal. The input Z represents the tissue, electrode, and whatever input network is used between the amplifier input and the signal sources.

high amplitude noise that exists at these frequencies. Similarly, not all of the ECG and EMG will be rejected. Therefore, inclusion of noise components from these sources should be expected in the raw ENG record and must be filtered before analysis and use of the ENG can be considered. High pass filtering the output of the differential preamplifier can remove much of the out of band noise not rejected by the CMR of the differential amplifier, but care should be taken to ensure that the differential preamplifier is operating within its linear dynamic range.

EXAMPLES OF LONG-TERM PERIPHERAL NERVE INTERFACES

Peripheral neural electrodes can be grouped according to the way that they interface with the peripheral neural tissue. They can be grouped into three main types: *regenerating*, *intraneural*, and *extraneural electrodes*. Common for all electrode interfaces presented here are that they

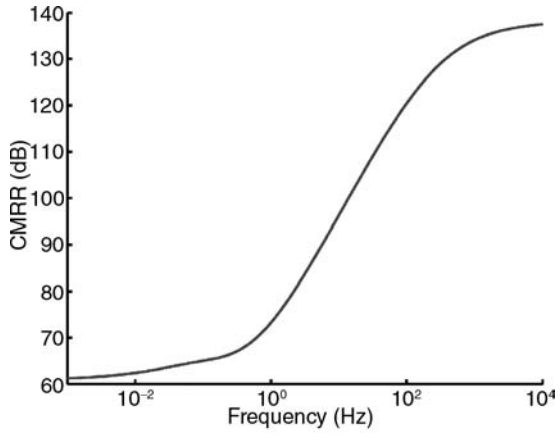


Figure 17. The frequency dependence of the CMRR is shown for an ideal preamplifier using electrodes that are mismatched by 10%. The frequency dependence of the electrodes has a large influence in the CMRR, which resembles the inverse of the electrode impedance.

record the extracellular potential from one or more peripheral nerve axons simultaneously (also referred to as extracellular electrodes). A fourth electrode group aims instead to record the intracellular potential of individual cells (also referred to as intracellular electrodes). In this case, the active recording site is placed inside a cell, so the electrode must therefore be carefully positioned and held in place during the entire duration of the recording. With their use, there is the risk of permanently breaking the membrane and injuring or killing the cell. As such, intracellular electrodes can be used acutely *in vitro*, *in situ*, or *in vivo*, however, they are not suitable for chronic animal, human or clinical use because of mechanical stability issues and the unavoidable damage to the cells that they record from. These electrodes are therefore not considered further here.

The choice of an optimal peripheral neural electrode largely depends on the application or the experimental design. One electrode might prove useful for one purpose, but not for suitable for another. A vast number of parameters have been used in the literature to describe, characterize, and compare neural interfaces. Most electrode interfaces can not only be used for recording from peripheral nerves, but also for applying electrical stimulation to the nervous tissue. However, the focus here will be on peripheral neural interfaces for recording. To provide a base for comparison between electrode types, the discussion will focus on the following issues:

Design parameters, including: The rationale for design, the ease of use and the cost of manufacturing and ease of implantation.

Application of the electrode, including: Recording selectivity and recording stability, expected lifespan, biocompatibility and robustness, and application examples.

Regenerating Electrodes

Introduction. Peripheral nerve axons spontaneously regrow after dissection or traumatic injury. The working principle of a *regenerating electrode* is to fit an electrode in the path between a proximal and distal nerve stump after a deliberate cut or traumatic nerve injury, and to guide the neural regrowth through perforations in the electrode structure. Ideally, the axons will grow through the perforations and allow very high resolution (single axon) recordings.

Sieve Electrodes

Design. Regenerating electrodes are often referred to as *Sieve electrodes*, inspired from their appearance. The structures are typically based on a sieve-shaped or perforated diaphragm, which contains an array of holes. The active sites are placed inside the sidewalls of the hole or in their immediate surroundings (similar to a printed circuit board via) to make physical contact with an axon growing through the hole. They are often designed incorporating a used small flexible tube made of a biocompatibile material (e.g., polyimide or silicone) that is used to hold the two ends of the transected nerve stumps in place with the sieve

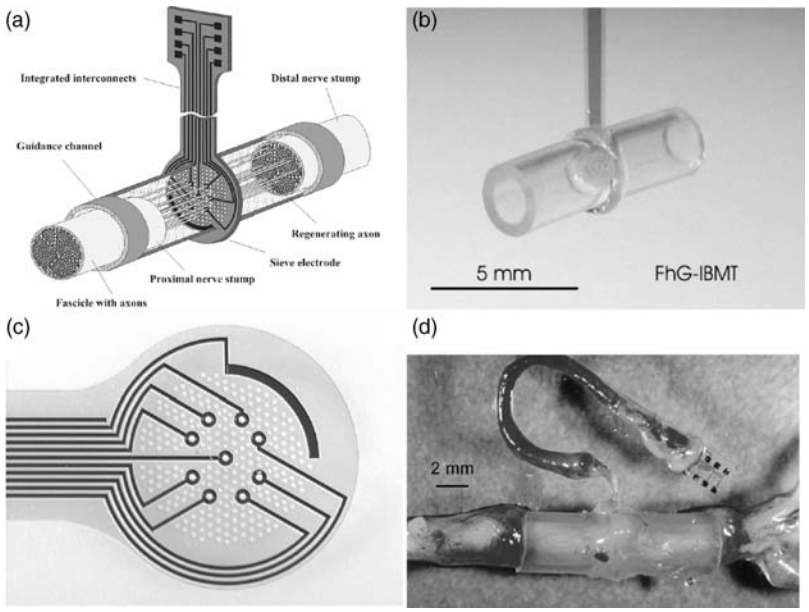


Figure 18. (a) Sketch of a Sieve electrode. The proximal and distal nerve stumps are fed into a guide tube. Neural axons must grow through holes in the Sieve electrode to reconnect with the neural axons on the other side. (b,c) Examples of Sieve electrodes. (d) shows the Sieve electrode *in situ*. (Reproduced with permission from T. Stieglitz, Fraunhofer Inst., Germany.)

electrode structure fixed in place between them. The tubes work to help guide the growing axons and to hold the electrode structure in place (36) (see Fig. 18). The first generations of Sieve electrodes were developed more than two decades ago (37). With the development of micromachining techniques, the devices have become considerably smaller and further enabled more suitable connections between the electrode and the surrounding world (38). These newer generations of sieve devices are made of silicon (39–41), polyimide or silicone substrates (38,42). Sieve electrodes have been designed with various shapes of the holes [round (40), squares (43), or longitudinal slots (39)]. The shape of the hole does not seem to have a major effect on the regrowth of the axons (36), but other design parameters, (such as the hole size, the thickness of the sieve diaphragm, and the relative amount of open versus closed space (also referred to as the *transparency factor*) have been found to be important. The thickness of the sieve diaphragm is typically in the order of 4–10 μm (36,40,41). Several findings suggest that 40–65 μm diameter Sieve holes work well (41,43,44), however, Bradley et al. (41) found that the holes could be as small as 2 μm. This may be explained by the difference in transparency factors of the electrodes. A high transparency factor is needed if the holes are small to secure regrowth (45). Major effort has also been put in to researching the use of neural growth factors (NGF) and how these may assist in promoting the spontaneous regeneration, and this may become an important factor in the future success of the *Sieve* electrodes.

Application. Bradley et al. (46) was among the first that obtained long-term recordings using sieve electrodes in mammals. They recorded from the rat glossophraryngeal nerve, which is mainly sensory, mediating somatosensory and gustatory information from the posterior oral cavity. Spontaneous activity was observed up to 7 weeks after implant. Evoked compound sensory responses were recorded up to 17 weeks after implantation. Other long-term evaluation studies have been reported by, for example, Mesninger et al. (42,47,48). Despite the growing number of reports on long-term usage of Sieve electrodes, they have not yet been tested in humans. An ideal axon-to-axon reconnection through a *Sieve electrode* could provide the ultimate interface to monitoring peripheral neural activity, but to date, this concept has not been proven in humans. Deliberate transection of a peripheral nerve to allow implant of a sieve electrode is a highly invasive procedure, and the success of the implant can only be evaluated several weeks after implantation (36). The current designs are not optimal since some axons may stop grow when the meet a wall and not a hole (36). Furthermore, it is currently not possible to control what fiber types regenerate. It has been shown that myelinated fibers are more likely to grow through than unmyelinated fibers, and large fibers are more likely to grow through than small fibers (49). It will be necessary to obtain a detailed map of what kind of sensory or motor axons have reconnected in order to interpret the peripheral neural signals properly (Fig. 19).

Intraneural Electrodes

Introduction. Intraneural electrodes are defined as electrodes that penetrate the endoneurium and perineurium of a peripheral nerve or the dorsal root ganglion (DRG) neuropil. The active sites are placed in the extracelluar space in close proximity with the nerve fibers, and therefore the electrodes aim to record from one single axon or from a small population of axons. The intraneural, penetrating electrodes are dependent on the insulating properties of the epineurium/perineurium to record.

Microneurography. Microneurography is a procedure in which a small needle electrode is placed directly inside a nerve fascicle. The aim is to record from one axon or a small a population of axons to study basic neural coding in the animal or human subjects, or as a diagnostic tool in clinical practice. Sketches of two types of microneurography electrodes are depicted in Fig. 19.

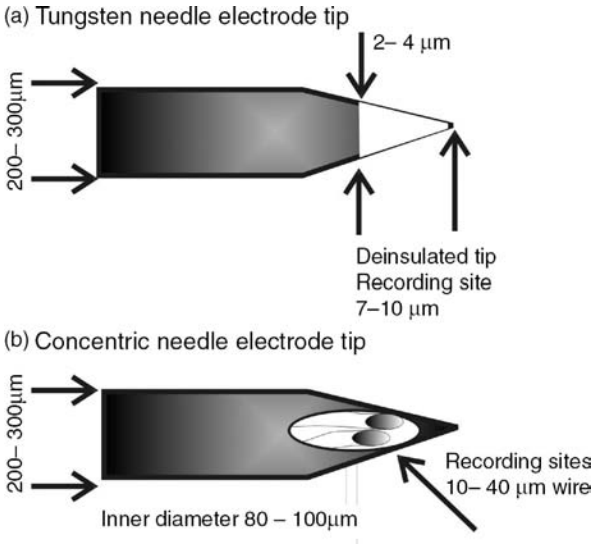


Figure 19. Microneurography electrodes. (a) A drawing of the geometry of a typical, commercially available tungsten needle electrode. A separate ground or indifferent electrode must be place in the close surroundings (Drawings appear courtesy of Dr. K. Yoshida.) (b) A drawing of a concentric needle electrode. One or two fine wires are threaded through a hypodermic needle and attached

Design. A classic needle electrode is manufactured from a stiff core material (e.g., tungsten, stainless steel, or carbon fiber) surrounded by a layer of insulating material. The Tungsten needle electrode typically has a 200–300 μm base diameter and a 2–4 μm tip diameter to ensure that the material is stiff enough to be pushed through the skin and muscle tissue and enter the perineurium without damaging the tip. The needle tip is uninsulated to create the active site. A separate ground wire must be placed subcutaneously in the close surroundings of the recording electrode (50,51).

A variation of the classic needle electrode was developed to allow more recording sites to be placed inside the nerve. The concentric needle electrode consists of one or more fine recordings wires threaded through a hypodermic needle. The fine wires are glued in place with an epoxy adhesive. The wires are typically made from tungsten, platinum, or platinum–iridium, and have a 20–25 μm diameter that provides a smaller, and thereby more selective recording, site than with the Tungsten needle electrode. An active site is made at the end of the wire by simply cutting the wire at a straight or slightly oblique angle. The hypodermic needle has a typical outer diameter of 200–250 μm and an inner diameter of 100 μm. The hypodermic needle tip provides a cutting edge for easy insertion of the electrode. Further, the needle shaft works as the ground during recording (52).

The microneurography electrodes are implanted perpendicular to the nerve trunk, and they are typically inserted by hand or using a micromanipulator. These electrodes are inexpensive, and they are typically and both types are easily manufactured by hand.

Application. Since Vallbo and colleagues developed the microneurography technique for more than three decades ago, the technique has proved to be a powerful tool in clinical experiments on conscious humans in hundreds of studies (51). However, the technique has a number of clear limitations. Placement of one single recording electrode may be extremely time consuming, and the delicate placement may easily be jeopardized if the animal or human subject moves. The microneurography needle electrodes are considered safe for short-term experiments, but they are not applicable for long-term implant and recording in animals or humans.

Longitudinal Intrafascicular Electrodes. The longitudinal intrafascicular electrodes (LIFE) are implanted parallel to the main axis of the peripheral nerve, and the active site of the electrode is placed along recording wires and not at the tip. The active sites typically record from a small population of axons.

Design. One of the first LIFE designs consisted of a simple 300 μm coiled stainless steel wire that was implanted into the nerve using a 30G hypodermic needle (53). The size of the hypodermic needle and implanted electrode may cause unnecessary injury to the nerve, and it also made it unsuitable for implantation in small nerves. Pioneering work was later done at University of Utah, as an attempt to solve these problems. A metal-wire LIFE was developed, see Fig. 21b (54). Here the 1–2 mm recording sites are created on 25 μm insulated, platinum–iridium wire. The recording wires are soldered to larger stainless steel lead-out wires, and the connections are protected by sliding a silicone tube over the solder joint. A polyaramid strand is threaded through this loop and glued to a 50–100 μm electrosharpened Tungsten needle (much similar to the tungsten needle electrode shown in Fig. 19). The needle is used to penetrate the perineurium/epineurium and pull the recording wires into place. The large lead-out wires remain outside the nerve. The electrode is sutured to the epineurium to hold it in place. The lead-out wires are then tunneled subcutaneously through the body to a chosen exit point. These metal-based LIFEs are relatively low cost, and they can be hand built. One of the disadvantages is the limited number of fine wires that can be fitted into one nerve. Further, the hand-building procedure does induce some variability in the size of the recording site, which influences the impedance and signal/noise ratio.

Later generations include polymer-wire LIFEs. Instead of using platinum–iridium wire, the recording wires are here based on 12 μm kevlar fibers. A metal is first deposited onto the kevlar to make the wire conductive, and a layer of silicone insulation is then added only leaving the recording sites exposed (55,56). The purpose of decreasing the intrafascicular wire size was to make the wires more flexible.

Application. The development of the LIFE has mainly been driven by the need for alternative and safe neural interfaces for clinical neuroprosthetic systems. Goodall et al. implanted the 25 μm platinum–iridium wire LIFEs in cat radial nerves and demonstrated that it was possible to obtain selective recordings of afferent activity from up to 6 months (59). Yoshida et al. used the same types of LIFES for recording neural activity in chronically

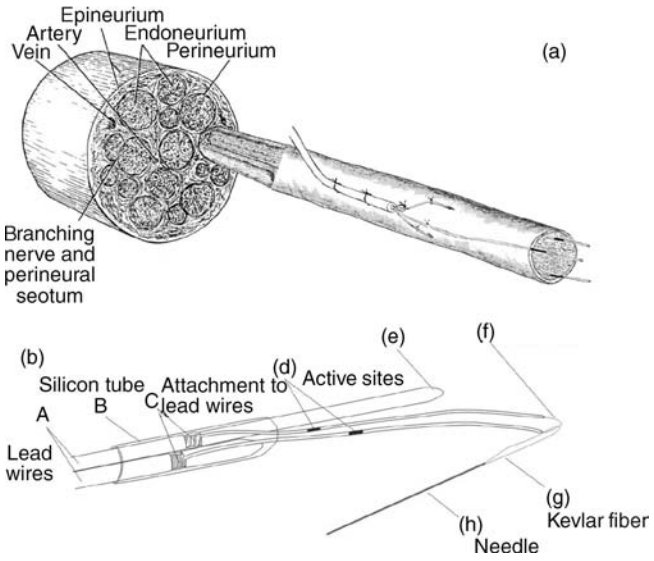


Figure 20. (a) Placement of a metal-wire LIFE electrode inside a whole nerve fascicle. Only the fine wires with the active sites are threaded through the nerve whereas the remaining part of the LIFE is located outside the epineurium. (b) is a line drawing of a two-channel metal-wire LIFE electrode. (a) Lead-out wires are located outside the body and tunneled through the body to a chosen exit point. (b,c) A silicone-tube is places over the connection point for protection. (d) Active sites on the recording wires. (f–h) A polyaramid fiber links the recording wire to a tungsten needle. The tungsten needle is used to thread the fine wires through the nerve and is removed after implantation. (Courtesy of Dr. K. Yoshida, Aalborg University, Denmark.)

implanted cats. These studies characterized the effect of electrode spacing for rejecting EMG and stimulus artifacts and explored the possibility of using multichannel recordings for noise reduction. Yoshida and Stein (33) also explored the possibility of extracting joint angle information from muscle afferent signals recorded with LIFEs (6). The information was used as feedback in closed-loop control of the ankle joint in the cats. The chronic recording stability and biocompatibility of the polymer-wire LIFES have been demonstrated by Malstrom et al. (58) in dorsal rootlets and in peripheral nerve (59) (Fig. 20).

Silicon-Based electrodes. The design and manufacturing of silicon-based electrodes take advantage of often sophisticated microfabrication techniques. These electrodes may be inserted transversely into the peripheral nerve as the microneurography needle electrodes, or they may be inserted longitudinally as the LIFE electrodes. Many groups around the world have designed, developed, and tested different silicon-based electrodes, however, the vast majority of those are developed for interfacing with the cerebral cortex tissue. Many of these designs are currently considered unsuitable as chronic peripheral nerve implants because of the risk of mechanical failure of the rigid base structure or small, thin electrode shafts. A notable exception to this view is the Utah array. One electrode that has been used for both peripheral nerve and cerebral cortex implantation is presented here.

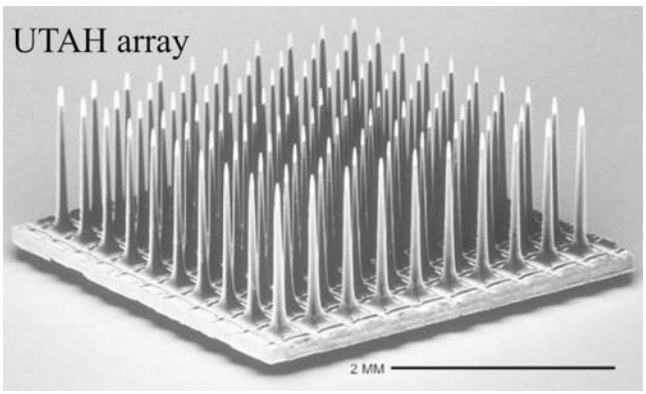


Figure 21. The 100-channel UTAH array. The electrode was originally developed at University of Utah, by Dr. R. A. Normann and colleagues. The electrode is commercially available through Cyberkinetics Inc. (Courtesy of Dr. R. A. Normann, University of Utah.)

Design. The UTAH array was originally developed by Normann and colleagues at University of Utah, and this electrode is now commercially available through Cyberkinetics Inc. The UTAH array is a three-dimensional (3D) silicon structure usually consisting of 100 penetrating electrodes arranged in a 10×10 array (see Fig. 21). Each needle is a long, tapered structure that is 1.5 mm long and has an 80 μm diameter at the base. The distance between the needles is 400 μm. The base substrate for this electrode is silicon, and glass is used as insulation material. Only the very tip of each needle is deposited with gold, platinum, or iridium to form the active site (59). Fig. 22 shows an array with equal height needles, which was originally designed for recording from the cerebral cortex where the neurons of interest are often located at the same depth. A slanted UTAH array was designed with different lengths of the needles to better target different layers in the cerebral cortex or to accommodate for the fact that fascicles in a peripheral nerve are located at different depths. Implantation of the UTAH array requires a special insertion tool that shoots the electrode into the neural tissue in a controlled manner. This is necessary because the high density of the needles increases the amount of force that is necessary to penetrate the dura or the epineurium. Further, the insertion tool avoids the elastic compression of the neural tissue and possibly damage of the neural tissue that was found during manual insertion (60).

Application. The recording properties of the Slanted UTAH array was first tested by Branner and colleagues in acute cat model (61,62) and later in similar chronic cat implants (2). The 100-electrode array was inserted in the cat sciatic nerve using the pneumatic insertion tool (described above). The acute work demonstrated that it was possible to achieve highly selective recordings with the array. In the chronic experiments, the array was held in place using a silicone cuff (self-spiraling or simply oval shaped) to protect the implant and surrounding tissue, to hold the implant in place, and to electrically shield the electrode. This containment was found to be important for the long-term survival of the electrode. In the chronic experiments, electrodes with lead wires were implanted

up to 7 months, however, it was only possible to obtain stable recordings for a few days after implant.

The slanted UTAH array has also been implanted in dorsal root ganglions (through the dura) in acute cat model to evaluate the neural coding of the lower leg position (10). In this work, it was possible to obtain selective recordings from the UTAH arrays over the duration of the experiment. The UTAH array has recently been implanted in a human volunteer (in a peripheral nerve in the lower arm), however, the results are pending (63). The UTAH array is also currently under clinical evaluation as a cortical interface in humans by Cyberkinetic, Inc.

Extraneural Electrodes

Introduction. The extraneural electrodes are the least invasive interfacing technique presented here. They are defined as electrodes that encircle the whole nerve or are placed adjacent to the nerve without penetrating the epineurium or dorsal root ganglion neuropil. The electrodes record from a large number of nerve fibers at the same time.

Circumferential Cuff Electrodes. The cuff electrode is probably the peripheral nerve interface that appears with the largest number of variations in designs and configurations. Only the main designs will be presented here. It is the most mature and widely used chronically implanted electrode for use in neuroprosthetics and has been used in a large number of animal studies and implants in human subjects (see section below).

Design. A circumferential cuff electrode consists of a tube or cuff that is placed around the whole nerve. The active sites are attached to the inside of the cuff wall to make contact with the epineurium (see Fig. 22). The tube wall works as an insulating barrier keeping extracellular currents emerging from the nerve within the cuff and blocking electric currents from other surrounding tissue (predominantly muscle activity, but also activity from other nerves) outside of the cuff. The tube wall has found to be essential to obtain measurable voltages from the nerve trunk and to achieve a good signal/noise ratio. The cuff tube or wall is made of a flexible, biocompatible material, such as silicone, (1,64,65) or polyimide (66,67). The cuffs may be handmade or microfabricated. The microfabrication technique has the advantage that thinner walls can be obtained.

The shape and number of active sites placed inside the cuff wall vary. The active sites are typically made of inert metals, such as platinum or platium–iridium. In the classic configuration, a cuff electrode consists of a silicone tube lined with two or more thin conductive rings that are displaced from one another along their common axis (69). In this design, the number of active sites is usually three, but more may be fitted into the cuff depending on the application. The large conductive rings record from a large number of fibers within the nerve they encircle. Later designs placed several, smaller active sites around the inner cuff wall in an attempt to interface individual fascicles within the nerve trunk, and thereby increase the selectivity of the cuff recording (68,69). The number and size of the active sites have a large impact on the recording and stimulation selectivity, which will be discussed in the next section.

The cuff electrode is fitted onto a nerve by sliding the nerve into a longitudinal slit in the cuff wall while traction is placed on the cuff wall to keep the slit open. The procedure may be made easier for the experimenter/surgeon by gluing a number of sutures to the outside of the cuff wall (see Fig. 22a). The experimenter/surgeon can grab onto these sutures, pull the cuff open, and slide the cuff in place around the nerve. Knots are tied to close the cuff and secure that extracellular current is contained inside the cuff, which is important to achieve good recordings. Other closure mechanisms have been attempted.

1. A hinge or zipper-like closure has been suggested by Dr. M. Haugland, Aalborg University, Denmark [see Fig. 22c and patented by (70)]. Here the cuff is cut open and integrated with a piano hinge structure, that can be zipped up and threaded with a suture for closure. The disadvantage of this

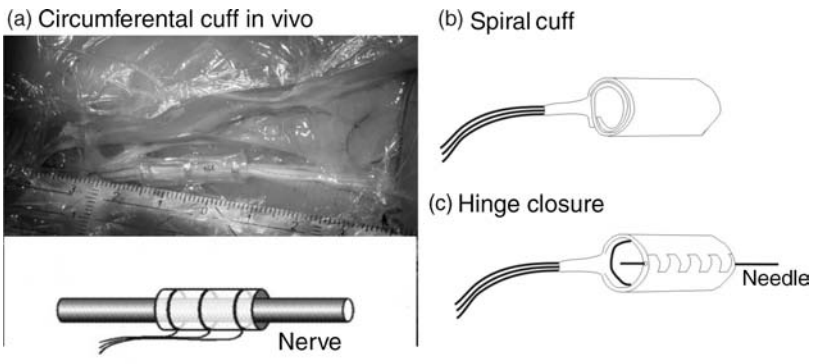


Figure 22. The sketch representation of the cuff (around the nerve) electrodes (left) and the newly designed 12-contact circular cuff electrode (with a longitudinal slit) for selective stimulation with a four-channel stimulator. In order to minimize contamination of the neural signal from the surrounding biological signals, the electrode must fit tightly around the nerve and close well. An electrode can be closed with the flap covering the longitudinal slit, and kept in position by means of the surgical threads around the cuff (left side). It can be closed by using a "zipper" method where the ends of the cuff have little holes, through which a plastic baton can be pulled (right side). A self-wrapping polyimide multipolar electrode for selective recording/stimulation of the whole nerve is shown in (c). (Part a is courtesy of Dr. W. Jensen, Aalborg University, Denmark.) (Parts b and c are courtesy of Dr. T. Sinkjær and Dr. M. Haugland, Aalborg University, Denmark.)

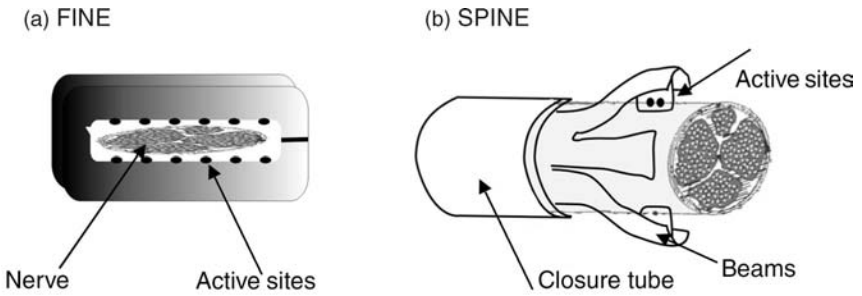


Figure 23. Sketches of two cuff-electrode types that reshape the nerve. (a) The flat-interface electrode (FINE) has a rectangular shape and flattens the peripheral nerve in order to make selective contact with individual fascicles in the nerve. The active sites are placed on the cuff-wall. (b) The slowly penetrating interfascicular electrode (SPINE) is designed with a number of beams holding the active sites. A closure tube is moved over the beams to force the beams into position. The nerve will reshape around the beams and thereby make selective contact with the fascicles in the nerve. (Redrawn with permission from Dr. Durand, Case Western Reserve University.)

closure mechanism is that it adds several steps to the fabrication process. Furthermore, greater skill is required from the experimenter/surgeon for correct implantation and closure.

2. A second cuff is placed over the first cuff. This is an easy way to securely cover the longitudinal slit, and it is easy to implant. The disadvantage of this method is that overall implant diameter and stiffness increases.

The diameter of a fixed-sized cuff is an important design parameter. The inner diameter cannot be too large since the cuff wall works an electrical shield, and the active sites must make contact with the epineurium. It has been suggested that the fixed-sized cuff should have an inner diameter of $\sim$20–50% larger than the nerve to avoid post-implant nerve damage caused by edema or swelling of the nerve (1,32).

A spiral cuff or a self-wrapping cuff design was originally suggested by Naples et al. (71) to circumvent some of the problems with the fixed-sized cuff. The spiral cuff electrode consists of a planar silicone or polyimide sheeting containing the active sites. The planar sheeting is flexible and is designed to automatically coil around a peripheral nerve (see Fig. 22b). These electrodes are generally handmade, but also have been manufactured by microfabrication techniques. The electrode will make a tight connection with the epineurium, leaving no space between the nerve and the contact sites. The self-sizing property of this design provides several advantages; (*1*) It is not necessary to determine the cuff diameter in advance as it is with the fixed-sized cuff electrodes. (*2*) The coiling will automatically and easily close the cuff and hold the cuff into place, and it is therefore very easy to implant. (*3*) The coiling will provide the necessary constriction of the extracellular current emerging from the nerve fibers without any possible leakages. (*4*) The coiling will allow the electrode to change size to accommodate for possible edema in the days after implantation.

An example of a cuff-electrode design that incorporates a number of the already discussed principles is a design referred to as the polyimide-hybrid cuff electrode (69) (see Fig. 22d). The electrode is based on a polyimide spiral cuff design containing multiple, platinum contacts on the inner wall. The polyimide-hybrid cuff is manufactured using microfabrication techniques that makes it possible to place a relatively high number of contacts inside the cuff and at the same time allow space for lead-out wires, which was a problem with the first multicontact cuffs. The microfabrication technique further secures high repeatability in the manufacturing process. Several patents on this type of hybrid fabrication has been issued (72,73).

Application. The first circumferential cuff electrode used for chronic recording was described by Hoffer et al. (74) and later by Stein et al. (75). It is to date the most successful and most widely used chronic peripheral interface in both animals and humans. An overview of current neuroprosthetic applications based on long-term recording from peripheral nerves is given elsewhere in this chapter. Many of these systems use cuff electrodes as their neural interface, and at the time of this writing, is the only implanted nerve interface being used in humans with spinal cord injury or stroke to provide FNS systems with natural sensory feedback information.

Reshaping Cuff Electrodes. The fixed-diameter, circumferential cuffs or the spiral cuffs have their active recording sites placed around the epineurium, and these electrodes therefore mainly record from the most superficial axons or fascicles. The reshaping cuff electrode attempt to accommodate for that by remodeling the shape of the nerve to access individual fascicles.

Design. Two types of reshaping cuff electrodes were developed by Durand and colleagues at Case Western Reserve University. In the slowly penetrating interfascicular electrode (SPINE) the active sites are placed inside the nerve, however, without penetrating the epineurium (76) (see Fig. 23). This is accomplished by a set of blunt penetrating beams that is slowly pushed into the nerve trunk by a closure tube. The active sites are placed on the side of the beams and may therefore access different fascicles of the nerve trunk. A refinement on this concept, the flat interface nerve electrode (FINE) was later developed (77). This electrode has a fixed shape like the fixed diameter, circumferential cuffs, however, the shape is rectangular instead of round (see Fig. 23). Ideally, the individual fascicles will then make contact with different contact sites along the cuff wall.

Application. The reshaping cuff electrodes were mainly developed as a stimulation platform; however,

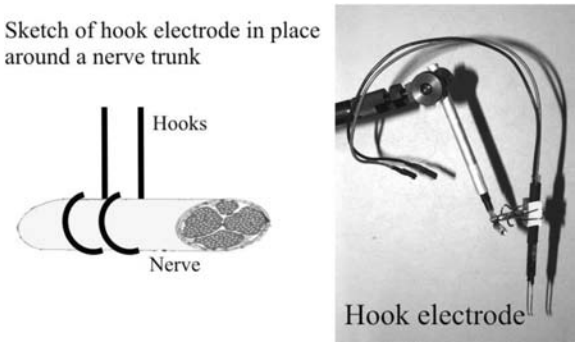


Figure 24. A hook electrode consists of two simple wires/hooks that are held in close contact with the nerve surface. The picture shows an example of a custom made hook electrode. (Courtesy of Dr. K. Yoshida, Aalborg University, Denmark.)

both types could potentially be used for recording. Acute studies with the SPINE electrode show that the slowly penetrating beams can safely place active sites within the nerve trunk, and it is possible to selectively activate different fascicles of the nerve (77). One of the main design issues with the FINE electrode has been to choose the right size of the electrode without causing nerve damage. Animal studies have shown that the FINE electrode is suitable for both recording and stimulation in chronic animal implants (78,79).

The Hook Electrode. The hook electrode was among the first extraneural electrodes used in experimental neurophysiology and neuroscience. The hook electrode is not useful for long-term nerve implants, however, it is still widely used in acute animal experiments or as an intraoperative tool for nerve identification, and therefore a brief introduction is given here.

Design. The hook electrode is constructed of usually two or more hook-shaped wires that form the recording/stimulation sites (typical materials used are platinum, stainless steel, or tungsten) (see Fig. 24). The nerve trunk or fascicle is placed so the epineurium makes close contact with the hooks. The hook-wires are soldered to insulated lead-out wires that are used to connect with the recording/stimulation equipment. The lead-out wires are usually threaded inside a tube or glued onto a stiff rod to form a base for holding the hook electrode in place. The hook wires must be stiff enough to support the whole nerve or nerve fascicle without yielding. One of the main advantages of the hook electrode is that it is easily handmade at a very low cost.

Application. The distance between the hooks may be varied to change the electrical recording/stimulation properties, however, the relatively large distance between the active sites means that the electrode has a poor selectivity. To use the hook electrode, the whole nerve must first be made accessible in the body, and surrounding fascia must be dissected away to place the hooks underneath the nerve. To avoid electrical shunting between the hooks caused by the extracellular fluid, a paraffin oil or mineral oil pool is often created around the hook electrode and nerve in acute animal studies. For human experiment, the hook electrode and the nerve may alternatively be suspended into the air to obtain a similar, insulating effect between the hook wires.

USE OF PERIPHERAL NERVE SIGNALS IN NEUROPROSTHETIC APPLICATIONS

Traumatic or nontraumatic injury to nerves at peripheral, spinal, or cortical level may cause permanent loss of sensory and motor functions. The goal of neuroprosthesis applications is to replace, restore, or augment the lost neural function in order to regain the lost mobility or sensation. Popovic and Sinkjær (80) discussed clinical areas where FES is already important or where it holds great potential if the adequate technology will be developed. In the present section, the state-of-the-art of using information provided by peripheral nerve signals in preclinical evaluated and experimental neuroprosthetic applications will be discussed.

Neurorprostheses systems traditionally work by using functional electrical stimulation (FES) to elicit controlled muscle contraction. The functional electrical stimulation can be applied using an open- (feed forward) or a closed-loop (feedback) controller scheme. The open-loop systems have proven to work, however, since no regulation is provided, excessive stimulation and muscle fatigue are commonly observed. Once the stimulation has been determined in the open-loop system, it is not changed, and it is not possible to detect or account for unexpected events or external disturbances. Several studies have shown that the application of closed-loop control techniques can improve the accuracy and stability of FES activated muscles. However, the closed-loop systems are dependent on feedback on the current state of the part of the body under control. The availability of sensors to provide reliable feedback information is, therefore, essential. An illustration of the basic elements in a closed-loop FES control system is given in Fig. 25.

Artificial sensors placed outside the body have been used widely within closed-loop FES applications to provide the necessary feedback signals [e.g., including goniometers

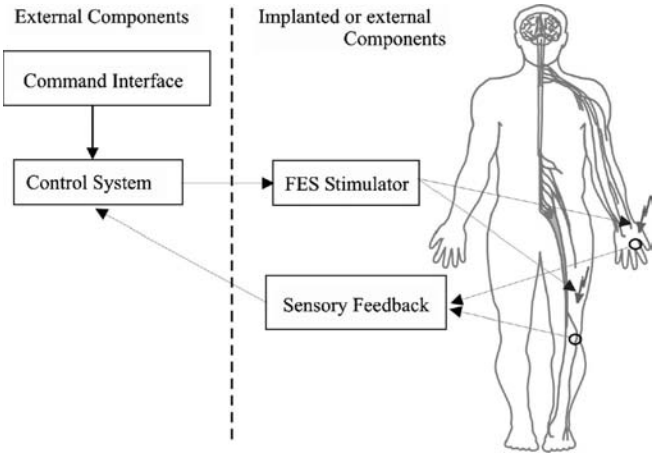


Figure 25. Illustration of the basic elements in a closed-loop neuroprosthesis application. The instrumentation typically includes an external (i.e., outside of the body) command interface and control system and external/implanted interfaces for sensing or activating the biological tissue.

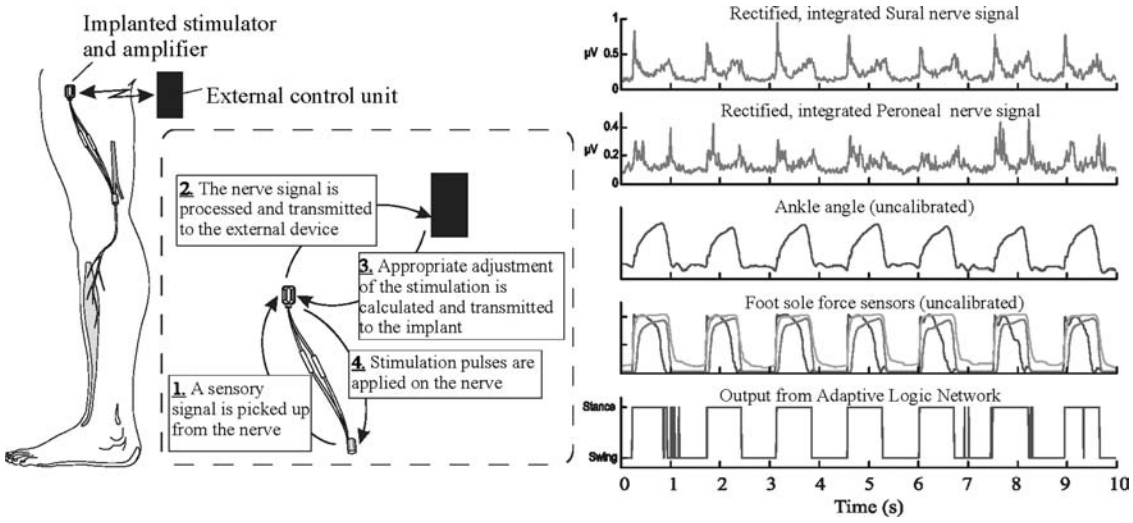


Figure 26. Footdrop correction system, using natural sensory information from the Sural or Peroneal nerve. The nerve signals were processed to determine heel contact with the floor and to decide the timing of the ankle dorsi-flexion stimulation. The stimulation may be applied through a peroneal nerve cuff electrode innervating the Tibials Anterior muscle or by surface electrodes. (Reproduced with permission from Dr. T. Sinkjær, Dr. M. Haugland, and Dr. M. Hansen.)

and accelerometers (80)]. However, the excess time consumed donning and doffing the artificial sensors, and the poor cosmetic visibility of bulky sensors has proven to be a major disadvantage for end user acceptance (1). Artificial sensors have also been placed inside the body; however, they require power to work and can be a problem to replace in case of failure.

The natural sensors, such as those found in, the skin (cutaneous mechanoreceptors, nociceptors), muscles (proprioceptors), or internal organs (visceral neurons) normally providing the CNS with feedback information on the state of the body. They are presently an attractive alternative to the artificial sensors in the cases where the sensors are still viable below the level of brain or spinal cord injury. Unlike artificial sensors, natural sensors are distributed throughout most of the body, and they do not require power or maintenance. Also, efferent motor signals are of interest to monitor the command signals from the CNS to the motor end organs. In many cases, they can be accessed through the same nerve based electrodes used for FES.

Therefore, the peripheral nerve signals of interest here are any neural tissue that is accessible outside the vertebral canal, including afferent and efferent somatic and autonomic pathways (the peripheral nervous system was defined in section The peripheral nervous system and specific peripheral nerve interfaces are described above).

Preclinical Neuroprosthetic Systems

Cutaneous Afferent Feedback for the Correction of Dropfoot. Foot-drop following upper or lower motor neuron deficits is defined as the inability to dorsiflex the foot during the swing phase of the gait (81). Electrical stimulation of the peroneal nerve has proven to be a potentially useful mean for enhancing foot dorsi-flexion in the swing phase of walking and thereby make the patient walk faster and more securely. The stimulator is often located just distal to the knee and can be either externally mounted or partly implantable. A key element in the system is the external heel-switch placed under the foot, which provides sensory information on heel-to-ground events necessary to switch on the stimulation at the right time. This system has proved clinically functional in large groups of patients. If the external heel switch is replaced with an implantable sensor, it can provide not only the advantages of foot drop corrections systems without the need for footwear, but also eliminates daily problems of mounting the heel switch or swapping them in different pairs of shoes.

The sural nerve is purely sensory and innervates the skin underneath the foot. It has been shown that whole nerve cuff electrode recordings from this nerve can provide information about foot contact events during walking, including heel-to-ground and toe-off-ground events (82). An example of recorded sural nerve activity during walking is shown in Fig. 26. The sural nerve activity has been processed (i.e., filtered, rectified, and bin-integrated). During walking, the nerve signal modulates, and two large spikes are repeatedly observed that correlate with the transitions between swing-to-stance phase and stance-to-swing phase. One of the most significant problem in correlating the nerve signal activity to the actual heel contact during walking is the noise originating from nearby muscle activity. The inclusion of a decision-assistive method like adaptive logic networks as a part of the postprocessing of the neural signal has been used improved the consistency of detecting reliable events (12,83).

The different elements of the drop foot correction system based on natural sensory feedback are shown in Fig. 26, including the external control unit, the implanted sural nerve cuff electrode and amplifier, and the peroneal nerve cuff electrode and stimulator. The sural nerve cuff electrode is connected to an external amplifier, and the signal output is fed to a microprocessor-controlled stimulator that activates the ankle dorsiflexor muscles.

The use of cutaneous afferent feedback in correction of drop-foot has to date been evaluated in several patients. The first evaluation included a 35 year-old subject with a hemiplegic dropfoot. In this case, the cuff electrode was connected to the control system through percutaneous wires, and an external stimulator was placed over the Tibialis Anterior muscle (84). A later version of the system was tested in a 32-year old subject, and the external stimulator was here replaced with an implanted cuff electrode around the peroneal nerve, which innervates the tibialis anterior muscle. An implantable system is currently under development and clinical testing by the Danish company Neurodan A/S.

Cutaneous Afferent Feedback for the Restoration of Hand Grasp. The fingers of the human hand contain an estimated 200–300 touch sensing receptors per square centimeters within the skin. It is among the highest densities of sensory receptors in the body (85). The cutaneous receptors and their responses have been studied extensively using microneurography techniques in normal subjects (85–87). The sensors mainly signal information about indentation and stretch of the skin, and the sensors play an important role in the control of precise finger movements and hand grasp while manipulating objects. It was found that slips across the skin were shown to elicit automatic motor reflex responses that increased the grasp force.

A FES system was developed to evaluate the use of cutaneous sensory information in restoration of hand grasp functions. The system has been evaluated in three spinal cord injured subjects (88–92). Results from a 27 years-old tetraplegic male with a complete C5 spinal cord injury (2 years postinjury) are presented in detail here. Before system was implanted, the patient had no voluntary elbow extension, no wrist function, and no finger function. Furthermore, he had only partial sensation in the thumb, but no sensation in 2nd to 5th fingers. He was implanted with a tripolar nerve cuff electrode on the palmar interdigital nerve (branching of the median nerve). Eight intramuscular stimulation wire electrodes were simultaneously placed in the following muscles: Extensor Pollicis Brevis, Flexor Pollicis Brevis, Adductor Pollicis, and Flexor Digitorum Longus to provide the grip control.

The cuff electrode was implanted around the palmar interdigital nerve, which is a pure cutaneous nerve innervating the radial aspect of the index finger (branching off the median nerve). The cuff electrode recordings were used to detect the occurrence of fast slips (increased neural activity was recorded when objects were slipping through the subject's fingers, and the neural activity showed similar characteristics as the cutaneous sural nerve activity shown in Fig. 26).

The stimulator was controlled by a computer, which also sampled the nerve signals and performed the signal analysis. When stimulation was turned on, the object could be held in a lateral grasp (key grip) by the stimulated thumb. If the object slipped, either because of decreasing muscle force or increasing load, the computer detected this from the processed nerve signal and increased the stimulation intensity with a fixed amount so that the slip was arrested, and the object again held in a firm grip. When extra weight was added, the slipped distance was also comparable to the performance of healthy subjects.

Today this system is developed to an extent where the subject can use it during functional tasks. During an eating session where the subject has a fork in his instrumented hand (grasped between the thumb and index finger), the control system is designed to decrease the stimulation of the finger muscles until the feedback signal from the skin sensors detects a slip between the index finger and the fork. When a slip is detected, the stimulation to the thumb increases automatically proportional to the strength of the sensory feedback, and if no further slips are detected, the controller starts to decrease the stimulation. This continuous tracking of the minimally necessarily needed stimulation means that the hand muscles are only loosely stimulated when the subject decides to rest his arm (which typically happens when they have a conversation with one or more of the persons at a dinner table). The stimulation will automatically increase when they start to eat again by placing the fork in the food. A typically eating session will last 20–30 min. A large fraction of this time is dedicated to "noneating" activities. During such times, the stimulation is at a minimum (keeping the fork in the hand with a loose grasp) and thereby preventing the hand muscles to be fatigued. When the feedback is taken away, the subject will typically leave the stimulation on at high stimulation intensity for the full eating session. This will fatigue the stimulated muscles, and the subject will try to eat their dinner faster, or they will rest their muscles at intervals by manually decreasing the stimulation. It is an effort that requires more attention from the subject than the automatic adjustment of the stimulation intensity.

The natural sensory feedback system was developed at Aalborg University, Denmark, and based on a telemetric system and on the Freehand system (NeuroControl Corp.). The Freehand system consists of a number of epimysial stimulation electrodes to provide muscle activation. More than 200 patients have to date received this system (Fig. 27).

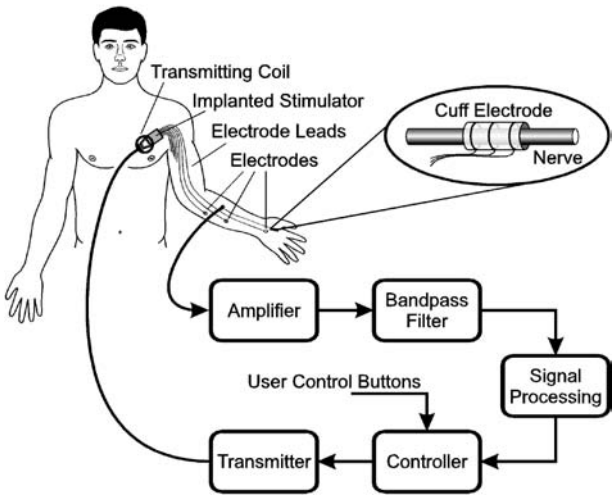


Figure 27. A neural prosthetic system for lateral-key grip restoration after spinal cord injury including sensory feedback. The FreeHand stimulation system (developed at Case Western Reserve University, Ohio) was modified to take advantage of natural sensory feedback signals obtained from a cutaneous, digital nerve in the hand. (Adapted from Ref. 92 with permission.)

Experimental Neural Prosthetic Systems

Somatic Afferent Feedback in Control of Joint Movement. The natural sensors that carry information about movement and orientation of the body in space (proprioception) are primarily located in the muscles, tendons, joints, and ligaments. The sensory signals are essential for the central nervous system in voluntary control of movement. In contrast to the more on–off type of information provided by the natural sensors located in the skin (e.g., heel contract and toe off events as used in correction of drop foot), the muscle spindle afferents are believed to carry information on muscle displacement (length, velocity, and acceleration). The proprioceptive sensors may therefore be useful in FES systems for control of joint movement. A majority of the research within this area has so far focused on restoration of standing and walking after spinal cord injury, however, the same principles explained here may be applied in control of upper extremity joints. Muscle afferent signals have been studied in animal models using intrafascicular (6,93) and whole nerve cuff electrodes (94,95), and these studies have revealed some characteristic features in muscle afferent responses that must be taken into consideration in a future FES system design.

Muscle afferent signals were obtained from passively stretched agonist-antagonist muscle groups around the ankle joint innervated by the tibial and the peroneal nerve branches in an animal model of spinal cord injury (6,94). The muscle afferent signals were recorded during simple flexion-extension rotations of the ankle joint, and it was found that the muscle afferent activity from the two nerves responded in a complementary manner, see Fig. 28. Thus, the muscle afferent activity increased during muscle stretch, but the activity stopped as soon as the movement was reversed, and the muscle was shortened. In the animal models used here, the muscles are not under normal efferent control from the central nervous system, and the intrafusal muscle fibers (where the muscle spindles are located) therefore become slack during muscle shortening, and the muscle spindles stop responding. The muscle afferents can therefore only provide information about muscle length during muscle stretch. In order to obtain a continuous and reliable muscle afferent feedback signal, it is necessary to use information from more muscles or muscle groups acting around a joint.

In spite of the nonlinear nature of the muscle afferent signals, it has shown to be possible to use the signals as feedback in a closed-loop control system. Yoshida and Horch (6) obtained control of ankle extension (ramp-and-hold movement) against a load and ankle flexion-extension (sinosoidal movement) in a cat model. A look-up table was first established by creating a map between the neural activity and the ankle joint angle. The tibial and peroneal nerve activity was simultaneously recorded using intrafascicular electrodes and decoded according to the look-up table.

To better incorporate the nonlinear behavior, more sophisticated information extraction algorithms have been employed to decode the muscle afferent signals, including a Neuro-Fuzzy network (96), and neural networks (97,98). In all cases, continuous information from both an agonist and antagonist muscle group was used as input to the networks.

The next step is to explore if muscle afferent signals on selected peripheral nerves can be recorded in humans and applied to improve FES standing.

Visceral Neural Signals in Control of Bladder Function. Normal bladder function (storage and emptying of urine) is dependent on mechanoreceptive sensors detecting bladder volume and efferent control that trigger a bladder contraction. Normal bladder function can be affected by a number of pathological diseases or neural injuries, for example when spinal cord injury occurs at the thoracic or cervial level, the neural pathways normally controlling the bladder are affected. Bladder dysfunction can result in an either overactive bladder (small volume capacity and incomplete emptying, referred to as neurogenic detrusor overactivity) or an underactive bladder (high volume capacity), however, the result of both states is incontinence. Neurogenic detrusor overactivity is the most common form of detrusor dysfunction following spinal cord injury. An overfilling of the bladder can also lead to a possible life-threatening condition of autonomic dysreflexia (overactivity of the autonomic nervous system).

Conventional treatments include suppression of reflex contraction by drugs, manual catheterization, or inhibition of the reflex contraction by surgical intervention, which is also referred to as dorsal rhizotomy. Dorsal rhizotomy

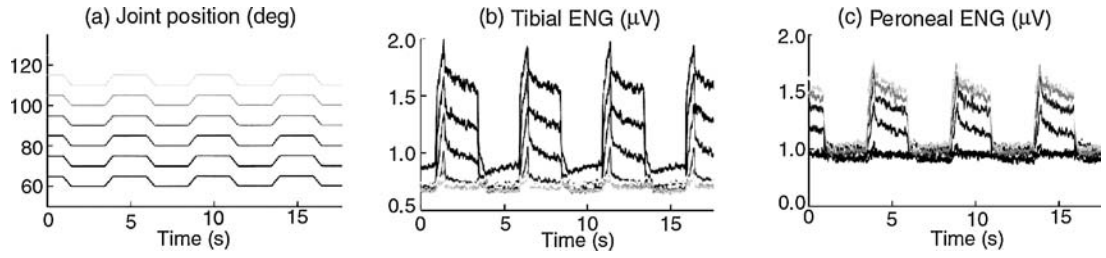


Figure 28. Muscle afferent signals recorded from the Tibial and Peroneal nerve branches during passive joint rotation in a rabbit model. The Tibial and Peroneal nerve activity modulated in a push–pull manner (i.e. the nerves responded to muscle stretch of the Gastrocnemius/Soleus and the Tibialis Anterior muscles, respectively). The magnitude of the response modulated with the initial position of the joint (the amount of prestretch in the muscles). (Courtesy of Dr. Jensen, Aalborg University.)

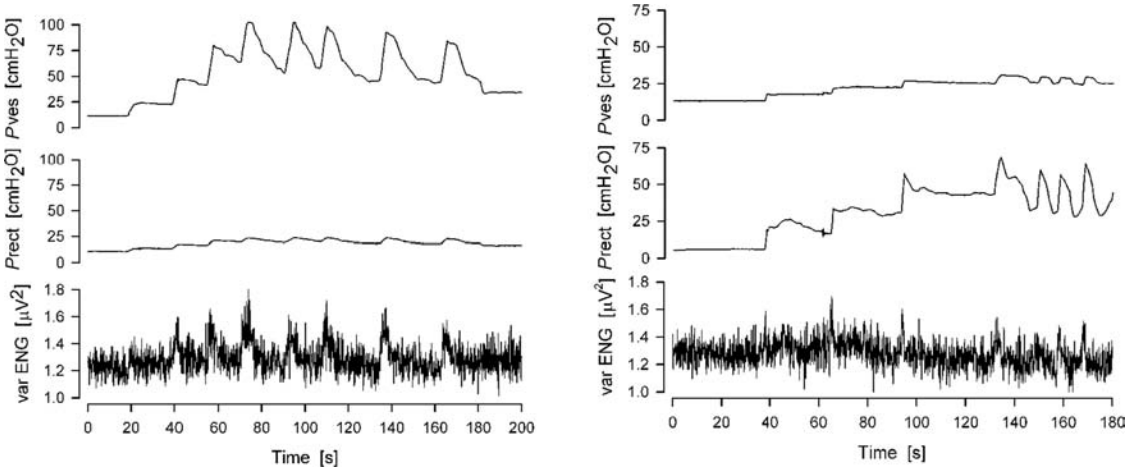


Figure 29. Two examples of recorded neural activity from the sacral roots in an anaesthetized human subject in preparation for surgery for implantation of a FineTech-Brindley Bladder system. The traces show the bladder pressure (top), rectal pressure (middle) and extradural sacral root activity (bottom) over time. (Courtesy of M. Kurstjens.)

increases the bladder capacity, however, reflex erection in male patients is lost, and this is in most cases not an attractive solution. To reestablish continence a closed-loop FES system has been suggested including sensory feedback on bladder volume and/or detrusor activation from bladder nerves.

Work in anaesthetized cats showed correlation between recorded activity from the S2 sacral root and the intravesical pressure associated with fullness of the bladder. Recently, similar information on bladder volume information has also been demonstrated in nerve cuff recordings from the S3 sacral root and the pelvic nerve in anaesthetized pigs (99) and extradural cuff recordings from sacral roots in humans (100) (see Fig. 29). The findings reveal, however, some problems in finding a reliable sensory feedback signal. It proved difficult to detect slow increases in bladder volume (99). Further, recordings from the pelvic nerve or sacral roots, however, are frequently contaminated with activity from other afferents, such as cutaneous and muscle afferents from the pelvic floor, rectum, and anus (100).

A closed-loop FES system for the management of neurogenic detrusor activity would work by using the sensory information from the bladder nerves to inhibit detrusor contractions by stimulating pudendal or penile nerves, or block efferent or afferent pelvic nerve transmission to prevent reflex detrusor contractions. The principle was tested in an anesthetized cat prepration where nerve cuff recordings from the S1 in cats could detect bladder hyperreflexive like contractions. However, the contractions were detected with a time delay in the order of seconds (using a CUMSUM alogrithm), which decreases the feasibility of the system (101).

Visceral Neural Signals in Respiratory Control. Obstructive sleep apnea is characterized by occlusions of upper airways during sleep. Electrical stimulation of the genioglossus muscle directly or via the hypoglossal nerve can improve the obstructions, however, a signal for detecting the occurrence of the obstructions will be useful in a FES closed-loop application (102).

This principle has been tested in a closed-loop FES system where the response of hypoglossal nerve was recorded during external loading of the pharynx during sleep to simulate upper airway obstructions in a dog model. It was shown that the hypoglossal nerve activity modulated with the pressure both during REM and NREM sleep. Any change in the cyclical rhythm of the recorded activity was used as an indication of the onset of apnea to the controller and triggered the onset of the stimulation of the same hypoglossal nerve to prevent the airway obstruction (103). Stimulation and recording from the same nerve is feasible in this case because the only information needed from the hypoglossal nerve is an indication of the onset of airway obstruction (i.e., no continuous information is needed).

CONCLUSIONS

The goal of the article is to provide the reader with an overview of currently available neural electrodes for long-term interfacing peripheral neural tissue. This is done to evaluate their suitability to monitor the neural traffic in peripheral nerves and their suitability to perform as sensors in neural prosthetic devices.

The role of neural prosthetic systems for increasing the quality of life of disabled individuals is becoming more evident each day. As the demand to develop systems capable of providing expanded functionality increases, so too does the need to develop adequate sensors for control. The use of natural sensors represents an innovation. Future research will show whether nerve-cuff electrodes and other types of peripheral nerve electrodes can be used to reliably extract signals from the large number of other receptors in the body to improve and expand on the use of natural sensors in neural prosthetic systems.

ACKNOWLEDGMENTS

The authors gratefully acknowledge The Danish National Research Foundation and The Danish Research Councils for their financial support.

BIBLIOGRAPHY

Cited References

1. Hoffer JA. Techniques to study spinal-cord, peripheral nerve and muscle activity in freely moving animals. In Boulton AA, Baker GB, Vanderwolf CH, editors. Neural Prostheses. Replacing Motor Function After Disease or Disability. New York Oxford: Oxford University Press; 2005.
2. Branner A, et al. Long-term stimulation and recording with a penetrating microelectrode array in cat sciatic nerve. IEEE Trans Biomed Eng 2004;51(1):146–157.
3. Lee DH, Claussen GC, Oh S. Clinical nerve conduction and needle electromyography studies. J Am Acad Orthop Surg 2004 July–Aug; 12(4):276–287. Review.
4. Fuller G. How to get the most out of nerve conduction studies and electromyography. J Neurol Neurosurg Psychiatry. 2005; 76(Suppl. 2):ii41–46. Review.
5. Crago PE, Nakai RJ, Chizeck HJ. Feedback regulation of hand grasp opening and contact force during stimulation of paralyzed muscle. IEEE Trans Biomed Eng 1991;38(1): 17–28.
6. Yoshida K, Horch K. Closed-loop control of ankle position using muscle afferent feedback with functional neuromuscular stimulation. IEEE Trans Biomed Eng 1996;43(2):167–176.
7. Haugland M, Hoffer JA. Slip Information Provided by Nerve Cuff Signals: Application in Closed-Loop Control of Functional Electrical Stimulation. IEEE Trans Rehab Eng 1994;2:29–36.
8. Sinkjaer T, Haugland M, Haase J. Natural neural sensing and artificial muscle control in man. Exp Brain Res 1994;98(3):542–545.
9. Sinkjær T, Haugland M, Struijk J, Riso R. Long-term cuff electrode recordings from peripheral nerves in animals and humans. In: Windhorst U, Johansson H, editors. Modern Techniques in Neuroscience. New York: Springer; 1999. p 787–802.
10. Stein RB, et al. Encoding mechanisms for sensory neurons studied with a multielectrode array in the dorsal root ganglion. Can J Physiol Phramacol 2004;82:757–768.
11. Stein RB, et al. Coding of position by simultaneously recorded sensory neurones in the cat dorsal root ganglion. J Physiol 2004;560(3):883–896.
12. Hansen M, Haugland M, Sinkjær T. Evaluating Robustness of gait event detection based on machine learning and natural sensors. IEEE Trans Neural Syst Rehab Eng 2004;12: 1:81–87.
13. Hoffer JA, Kallesoe K. Nerve cuffs for nerve repair and regeneration. Prog Brain Res 2000;128:121–134.
14. Struijk JJ, Thomsen M, Larsen JO, Sinkjaer T. Cuff electrodes for long-term recording of natural sensory information. IEEE Eng Med Biol Mag 1999;18(3):91–98.
15. Haugland M, Sinkjær T. 2000.
16. Riso RR. Perspectives on the role of natural sensors for cognitive feedback in neuromotor prostheses. Automedica 1998;16:329–353.
17. McDonnall D, Clark GA, Normann RA. Selective motor unit recruitment via intrafascicular multielectrode stimulation. Can J Physiol Pharmacol 2004;82(8–9):599–609.
18. McDonnall D, Clark GA, Normann . Interleaved, multisite electrical stimulation of cat sciatic nerve produces fatigue-resistant, ripple-free motor responses. IEEE Trans Neural Syst Rehabil Eng 2004;12(2):208–215.
19. Aoyagi Y, et al. Capabilities of a penetrating microelectrode array for recording single units in dorsal root ganglia of the cat. J Neurosci Methods 2003;30: 128(1–2):9–20
20. Webster. 1998.
21. Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol (London) 1952;117:500–544.
22. Chiu SY, Ritchie JM, Rogart RB, Stagg D. A quantitative description of membrane currents in rabbit myelinated nerve. J Physiol 1979;292:149–166.
23. Sweeney JD, Mortimer JT, Durand D. Modeling of mammalian myelinated nerve for functional neuromuscular stimulation. Proceedings of the 9th Annual International Conference of the IEEE EMBS 1987; 1577–1578.
24. Varghese A. Membrane Models. In: Bronzino JD, editor. The Biomedical Engineering Handbook: 2nd ed. Boca Raton (FL): CRC Press LLC; 2000.
25. McIntyre CC, Richardson AG, Grill WM. Modeling the excitability of mammalian nerve fibers: influence of afterpotentials on the recovery cycle. J Neurophysiol 2002;87(2):995–1006.
26. Struijk JJ. The extracellular potential of a myelinated nerve fiber in an unbounded medium and in nerve cuff models. Biophys J 1997;72:2457–2469.
27. Perez-Orive , Durand . Modeling study of peripheral nerve recording selectivity. IEEE Trans Rehabil Eng 2000;8(3):320–329.
28. Taylor J, Donaldson N, Winter J. Multiple-electrode nerve cuffs for low-velocity and velocity-selective neural recording. Med Biol Eng Comput 2004;42(5):634–643.
29. Chemineau ET, Schnabel V, Yoshida K. A modeling study of the recording selectivity of longitudinal intrafascicular electrodes. In: Wood D, Taylor J, editors. Getting FES into clinical practice. Proceedings of IFESS-FESnet 2004, 9th Annual Conference of the International Functional Electrical Stimulation Society and the 2nd Conference of FESnet 6–9 September, 2004Bournemouth, UK. 2004 p 378–380.
30. Yoshida K, Struijk JJ. The theory of peripheral nerve recording. In: Horch KW, Dhillon GS, editors. Neuroprosthetics: Theory and Practice World Scientific Publishing Co.; 2004.
31. Nikolic ZM, Popovic DB, Stein RB, Kenwell Z. Instrumentation for ENG and EMG recordings in FES systems. IEEE Trans Biomed Eng 1994;41(7):703–706.
32. Stein RB, et al. Stable long-term recordings from cat peripheral nerves. Brain Res 1977;128(1):21–38.
33. Yoshida K, Stein RB. Characterization of signals and noise rejection with bipolar longitudinal intrafascicular electrodes. IEEE Trans Biomed Eng 1999;46(2):226–234.
34. Ott HW. Noise Reduction Techniques in Electronic Systems. 2nd ed, New York: John Wiley & Sons, Inc.; 1988.
35. Grimnes S, Martinsen ø G. Bioimpedance and Bioelectricity Basics. London: Academic Press; 2000 p 221.
36. Heiduschka P, Thanos S. Implantable Bioelectronic Interfaces for Lost Nerve Functions. Prog Neurobiol 1998;55:433–461.
37. Mannard A, Stein RB, Charles D. Regeneration electrode units: implants for recording from single peripheral nerve fibers in freely moving animals. Science 1974;183(124):547–549.
38. Stieglitz T, et al. A biohybrid system to interface peripheral nerves after traumatic lesions: design of a high channel sieve electrode. Biosensors Bioelectronics 2002;17: 685–696.

39. Edell DJ. A peripheral nerve information transducer for amputees: long-term multichannel recordings from rabbit peripheral nerves. IEEE Trans Biomed Eng 1986;33(2):203–214.
40. Akin T, Najafi K, Smoke RH, Bradley RM. A micromachined silicon sieve electrode for nerve regeneration studies. IEEE Trans Biomed Eng 1994;41:305–313.
41. Bradley RM, Smoke RH, Akin T, Najafi K. Functional regeneration of glossopharyngeal nerve through michromachined sieve electrode arrays. Brain Res 1992;594:84–90.
42. Lago N, Ceballos D, RodrÚguez FJ, Stieglitz T, Navarro X. Long-term assessment of axonal regeneration through polyimide regnerative electrodes to interface the peripheral nerve. Biomaterials 2005;26:2021–2031.
43. Kovacs GTA, et al. Silicon-substrate microelectrode arrays for parallel recording of neural activity in peripheral and cranial nerves. IEEE Trans Biomed Eng 1994;41(6):567–577.
44. Zhao Q, Dahlin LB, Kanje M, Lundborg G. Specificity of muscle reinnervation following repair of the transected sciatic nerve. A comparative study of different repair techniques in the rat. J Hand Surg 1992;17:257–261.
45. Yoshida K, Riso R. Peripheral nerve recording electrodes and techniques. In Horch K, Dhillon GS, editors. Neuroprosthetics. Theory and Practice. 1st ed. New York: World Scientific; 2004. p 683–744.
46. Bradley RM, Cao X, Akin T, Najafi K. Long term chronic recordings from peripheral sensory fibers using a sieve electrode array. J Neurosc Methods 1997;73:177–186.
47. Mesinger AF, et al. Chronic recording of regenerating VIIIth nerve axons with a sieve electrode. J Neurophysiol 2000;83:611–615.
48. Shimantani Y, Nikles SA, Najafi K, Bradley RM. Long-term recordings from afferent taste fibers. Physiol Beh 2003;80: 309–315.
49. Berthold CH, Lugnegard H, Rydmark M. Ultrastructural morphometric studies on regeneration of the lateral sural cutaneous nerve in the white rat after transection of the sciatic nerve. Scan J Plast Reconstr Surg Supp l 1985;30:1–126.
50. Vallbo Å, Hagbart KE. Activity from skin mechanoreceptors ecorded percutaneously in awake human subjects. Exper Neurol 1968;21:270–289.
51. Vallbo Å, Hagbart KE, Wallin BG. Microneurography: how the technique developed and its role in the investigation of the sympathetic nervous system. J Appl Physiol 2004;96: 1262–1269.
52. Hallin RG, Wiesenfeld-Hallin Z, Duranti R. Percutaneous microneurography in man does not cause pressure block of almost all axons in the impaled nerve fascicle. Neurosc Lett 1986;68:356–361.
53. Bowmann BR, Erickson RC. Acute and chronic implanation of coiled wire intraneural electrodes during cyclical electrical stimulation. Ann Biomed Eng 1985;13:75–93.
54. Malagodi MS, Horch K, Schoenberg A. An intrafascicular electrode for recording action potentials in peripheral nerves. Ann Biomed Eng 1989;17:397–410.
55. McNaughton TG, Horch K. Metalized polymer fibers as leadwires and intrafacicular microelectrodes. J Neurosc Methods 1996;8:391–397.
56. Gonzalez C, Rodriguez M. A flexible perforated microelectrode arry probe for action potential recording in nerve and muscle tissue. J Neurosc Methods 1997;72:189–195.
57. Lefurge T, et al. Chronically implanted intrafascicular recording electrodes. Ann Biomed Eng 1991;19:197–207.
58. Malmstrom M, McNaughton TG, Horch K. Recording properties and biocompatibility of chronically implanted polymer-based intrafascicular electrodes. Ann Biomed Eng 1998;26: 1055–1064.
59. (a) Jones KE, Campebll PK, Normann RA. A Glass/Silicon Composite Intracortical Electrode Array. Ann Biomed Eng 1992;20:423–437. (b) Lawrence SM, et al. Long-term biocompatibility of implanted polymer-based intrafascicular electrodes. J Biomed Mater Res 2002; 63(5):501–506.
60. Rousche PJ, Normann RA. A method for pneumatically inserting an array of penetrating electrodes into cortical tissue. Ann Biomed Eng 1992;20:413–422.
61. Branner A, Normann RA. A multielectrode array for intrafasciclar recording and stimulation in sciatic nerve of cats. Brain Res Bull 1999;51(4):293–306.
62. Branner A, Stein RB, Normann RA. Selective Stimulation of Cat Sciatic Nerve Using an Array of Varying-Length Microelectrodes. J Neurophysiol 2001;85:1585–1594.
63. Gasson MN, et al. Bi-directional human machine interface via direct neural connection. Proceedings of the IEEE Conference on Robot and Human Interactive Communication, Berlin, Germany; 2002, p 26–270.
64. Stein RB, et al. Principles underlying new methods for chronic neural recording. Can J Neurol Sci 1975;2(3):235–244.
65. Haugland M. A flexible method for fabrication of nerve cuff electrodes. In: Proceedings of the 18th Annual Conference IEEE Engineering in Medicine and Biology. Amsterdam The Netherlands: 1996p 964–965.
66. Stieglitz T, Meyer JU. Implantable microsystems: polyimide-based neuroprostheses for interfacing nerves. Med Device Technol 1999;10(6):28–30.
67. Rodriguez FJ, et al. Polyimide cuff electrodes for peripheral nerve stimulation. J Neurosci Methods 2000;98(2):105–118.
68. Veraat C, Grill WM, Mortimer JT. Selective control of muscle activation with a multipolar nerve cuff electrode. IEEE Trans Biomed Eng 1993;40:640–653.
69. Schuettler M, Stieglitz T. 18 polar hybrid cuff electrodes for stimulation of peripheral nerves. Proceedings of the IFESS. Aalborg, Denmark: 2000. p 265–268.
70. Kallesoe K, Hoffer JA, Strange K, Valenzuela I. Simon Fraser University, Implantable cuff having improved closure. US patent 5,487,756, 1994.
71. Naples GG, Mortimer JT, Schiner A, Sweeney JD. A spiral nerve cuff electrode for peripheral nerve stimulation. IEEE Trans Biomed Eng 1988;35(11):905–916.
72. Grill WM, et al. Thin film implantable electrode and method of manufacture. Case Western Reserve University, US patent 5,324,322. 1994 June 28.
73. Grill WM, Tarler MD, Mortimer JT. Case western Reserve University. Implantable helical spiral cuff electrode. US patent 5,505,201. 1996 April 9.
74. Hoffer JA, Marks WB, Rymer Z. Nerve fiber activity during normal movements. Soc Neurosc Abs 1974: 258.
75. Stein RB, et al. Impedance properties of metal electrodes for chronic recording from mammalian nerves. IEEE Trans Biomed Eng 1978;25:532–537.
76. Tyler DJ, Durand DD. A slowly penetrating interfascicular nerve electrode for selective activation of peripheral nerves. IEEE Trans Rehab Eng 1997;5(1):51–61.
77. Tyler DJ, Durand DD. Functionally Selective Peripheral Nerve Stimulation With a Flat Interface Nerve Electrode. IEEE Trans Neural Syst Rehabil Eng 2002;10(4):294–303.
78. Yoo PB, Sahin M, Durand DD. Selective stimulation of the canine hypoglossal nerve using a multi-contact cuff electrode. Ann Biomed Eng 2004;32(4):511–519.

79. Leventhal DK, Durand DD. Chronic measurement of the stimulation selectivity of the flat interface nerve electrode. IEEE Trans Biomed Eng 2004;51(9):1649–1658.
80. Popovic D, Sinkjær T. Control of movement for the physically disabled: control for rehabilitation technology. 2nd ed. Aalborg: Center for Sensory Motor Interaction, Aalborg University; 2003.
81. Lyons GM, Sinkjær T, Burridge JH, Wilcox DJ. A review of portable FES-based neural orthoses for the correction of drop foot. IEEE Trans Neural Syst Rehabil Eng 2002;10(2): 260–279.
82. Haugland M, Sinkjaer T. Interfacing the body's own sensing receptors into neural prosthesis devices. Technol Health Care 1999;7(6):393–399.
83. Kostov A, Hansen M, Haugland M, Sinkjær T. Adaptive restrictive rules provide functional and safe stimulation patterns for foot drop correction. Art Organs 1999;23(5): 443–447.
84. Haugland M, Sinkjær T. Cutaneous Whole Nerve Recordings Used for Correction of Footdrop in Hemiplegic Man. IEEE Trans Rehab Eng 1995;3:307–317.
85. Johansson RS, Westling G. Tactile sensibility in the human hand: Relative and absolute densities of four types of mechanireceptive units in glabrous skin. J Physiol 1979; 21:270–289.
86. Vallbo Å, Johannson RS. Properties of cutaneous mechanoreceptors in the human hand related to touch sensation. Human Neurobiol 1984;3:3–14.
87. Westling G, Johansson RS. Responses in glabrous skin mechanoreceptors during precision grip in humans. Exper Brain Res 1987;66:128–140.
88. Haugland M, Lickel A, Haase J, Sinkjær T. Control of FES Thumb Force Using Slip Information Obtained from the Cutaneous Electroneurogram in Quadriplegic Man. IEEE Trans Rehab Eng 1999;7:215–227.
89. Haugland M, et al. Restoration of lateral hand grasp using natural sensors. Art Organs 1997;21(3):250–253.
90. Inmann A, Haugland M. Implementation of natural sensory feedback in a portable control system for a hand grasp neuroprosthesis. Med Eng Phy 2004;26:449–458.
91. Inmann A, Haugland M. Functional evaluation of natural sensory feedback incorporated in hand grasp neurprosthesis. Med Eng Phys 2004; 26(6):439–447.
92. Inmann A, et al. Signals from skin mechanoreceptors used in control of hand grasp neurosthesis. Neuroreport 2001;12(13): 2817–2820.
93. Jensen W, Yoshida K. Long-term recording properties of longitudinal intra-fascicular electrodes. 7th Annual Conference of the International Functional Electrical Stimulation Society, IFESS 2002, June 25–29, Ljubljana: Slovenia; 2002 p 138–140.
94. Riso R, Mossallaie FK, Jensen W, Sinkjær T. Nerve Cuff Recordings of Muscle Afferent Activity from Tibial and Peroneal Nerves in Rabbit During Passive Ankle Motion. IEEE Trans Rehab Eng 2000;8(2):244–258.
95. Jensen W, Riso R, Sinkjær T. Position Information in Whole Nerve Cuff Recordings of Muscle Afferents in a Rabbit Model of Normal and Paraplegic Standing. Proceedings of the 20th Annual International Conferences of the IEEE/EMBS Society. 1998.
96. Micera S, et al. Neuro-fuzzy extraction of angular information from muscle afferents for ankle control during standing in paraplegic subjects: an animal model. IEEE Trans Biomed Eng 2001;48(7):787–789.
97. Jensen W, Sinkjær T, Sepulveda F. Improving signal reliability for on-line joint angle estimation from nerve cuff recordings of muscle afferents. IEEE Trans Neural Syst Rehabil Eng 2002; 10(3):133–139.
98. Sepulveda F, Jensen W, Sinkjær T. Using Nerve Signals from Muscle Afferent Electrodes to Control FES-based Ankle Motion in a Rabbit. Proceedings 23rd Annual International Conference of the IEEE-EMBS. 2001.
99. Jezernik S, et al. Analysis of Bladder Related Nerve Cuff Electrode Recordings from Preganglion Pelvic Nerve and Sacrral Roots in Pigs. J Urol 2000;163:1309–1314.
100. Kurstjens M, et al. Interoperative recording of sacral root nerve signals in humans. Art Organs 2005;29(3):242–245.
101. Jezernik S, Grill WM, Sinkjær T. Detection and inhibition of hyperreflexia-like bladder contractions in the cat by sacral nerve root recording and stimulation. Neurouol Urodyn 2001;20:215–230.
102. Sahin M, Durand DD, Haxhiu MA. Closed-loop stimulation of hypoglossan nerve in a dog model of upper airway obstruction. IEEE Trans Biomed Eng 2000;47(7):919–925.
103. Sahin M, Durand DD, Haxhiu MA. Chronic recordings of hypoglossal nerve activity in a dog model of upper airway obstruction. J App Physi 1999;87(6):2197–2206.

See also ELECTROMYOGRAPHY; EVOKED POTENTIALS; NEUROLOGICAL MONITORS.

ELECTROPHORESIS

JOHN M. BREWER
University of Georgia
Athens, Georgia

INTRODUCTION

The "electrostatic effect" causes particles with the same sign of charge to repel each other, while two particles of opposite charge attract: Charged particles produce electric fields and a charged particle in an electric field experiences a force. Electrophoretic techniques are based on the movement or flow of ions in a solvent produced by an electric field, causing separation of different ions.

THEORY

An external electric field V produces a force on an ion with a charge Q equal to VQ(1). An ion in a solvent will be accelerated, but movement in the solvent will be slowed by an opposite force that comes from the viscosity of the solvent. For small velocities, the opposite (viscous) force is proportional to the velocity of electrophoresis. We call the proportionality constant the "frictional coefficient". The forces from the electric field and viscosity are opposed, so the net acceleration becomes zero and the velocity of electrophoresis (v) constant:

$$v = VQ/\text{frictional coefficient}$$

Mobilities of Ions

The velocity an ion reaches for a given applied field is a specific property of that ion. It is called the "mobility" of

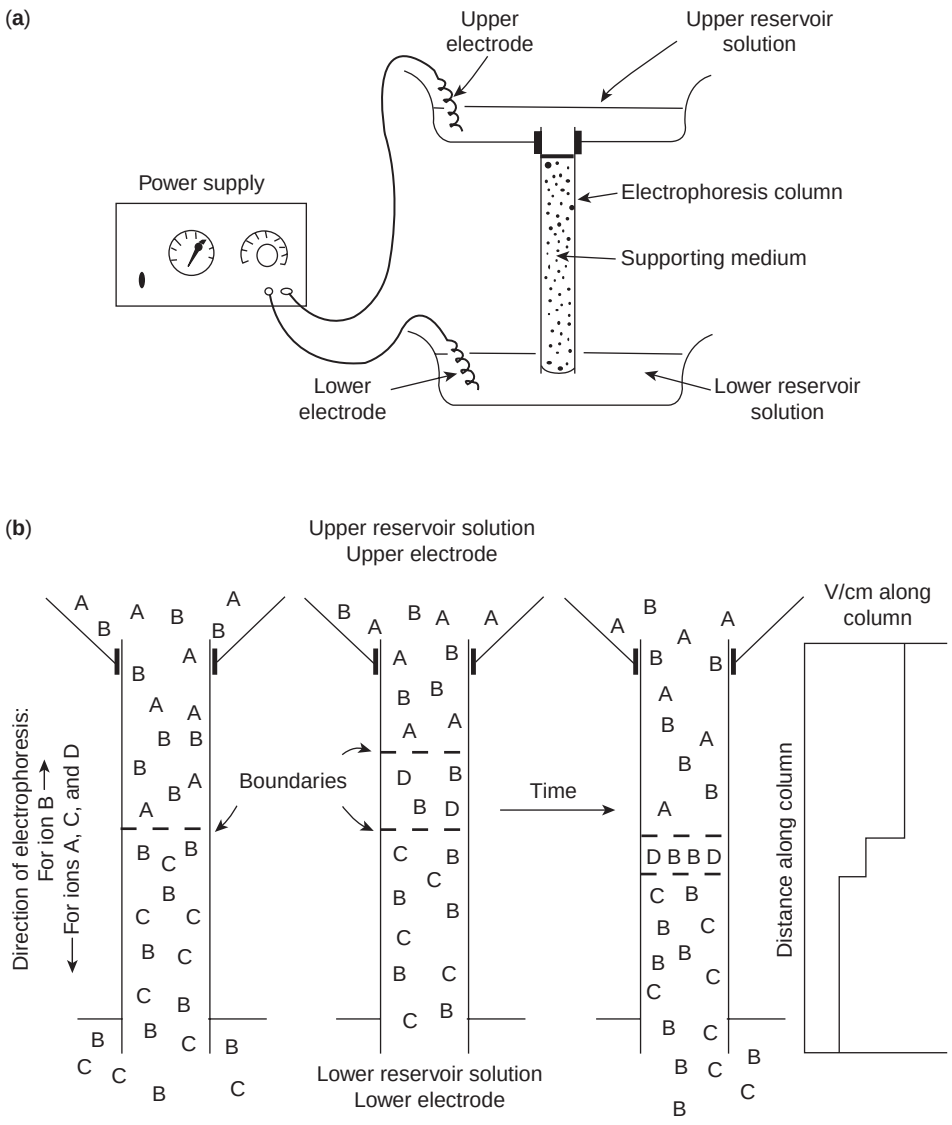


Figure 1. (a) Representation of a vertical electrophoresis experiment. (b) The lower drawing represents the stacking phase of disk electrophoresis or isotachophoresis of a single component, ion D. Charge signs are omitted as cationic and anionic isotachophoretic or disk systems exist. (Reprinted from Clinical Chemistry: Theory, Analysis, Correlation, 4th ed. Kaplan LA, Pesce AJ, Kazmierczak SC, editors, Electrophoresis, p 204, copyright © 2003, with permission from Elsevier.)

that ion:

$$v/V = \text{mobility of ion} = Q/\text{frictional coefficient}$$

For ions such as proteins that are very large compared with solvent molecules, the frictional coefficient is 6π times the product of the solvent viscosity and the effective radius (size) of the ion (1).

The direction the ion moves is determined by the sign of its charge Q and the direction of the electric field. The electric field is produced by two oppositely charged substances (electrodes) placed at opposite ends of the solution the ion is in, and their polarity (charge) determines the field's direction (Fig. 1) (2).

How fast ions in solution move is also affected by some other factors, but these are still hard to calculate (1,3). In aqueous solvents, ions are "hydrated", bound to a number of water molecules (4,5). A protein has many charged groups, both positive and negative, and those in contact with the solvent also will be hydrated (6) and will in addition have hydrated ions of opposite charge, called "counterions", nearby. In an electrophoresis apparatus (Fig. 1), the protein will move toward the electrode of opposite charge in an irregular fashion because random diffusional motion adds to the uniform electrophoretic movement. After each change of direction, the smaller ions of opposite charge are out of their equilibrium positions and need time(s) to readjust or move back into position. During the readjustment period(s), the effective voltage that the protein is exposed to is lower than the applied voltage, because the smaller counterions, while out of position, produce a field in the opposite direction.

Positively charged ions (cations) move in the opposite direction to negative ions (anions). They all are hydrated and the movement of the protein or other large ion of interest is slowed by a flow of hydrated ions of opposite charge (counterions) in the opposite direction. This is called an "electrophoretic effect" and, together with the "relaxation" or "asymmetry" effect described before, makes it difficult to use the measured velocities of large ions to calculate their real net charge or effective radius (1,3). Electrophoresis is consequently used largely as an empirical technique.

The mobilities of many small ions are known (7). These values are for fully charged ions. The actual (average) charge of ions like acetate depends on the pH, so the observed or actual (effective) mobility of some ions will also be pH dependent.

$$\text{Effective mobility} = (\text{mobility})(\text{average charge})$$

Proteins or polynucleotides have charged groups that determine their ionic properties. These charged groups in proteins are anionic and cationic, with pK values from 3 to 13, so the effective mobilities of proteins are very pH dependent. For proteins, overall (net) charges vary from positive at low pH to zero (isoelectric) to negative at more alkaline pH values. Substances whose charge can vary from positive to negative are called "amphoteric". Routine electrophoretic separations are done at constant pH values, in a "buffered" solution, to minimize effects of pH on velocity of movement.

Electrical Effects

These techniques are electrical in character. An applied voltage V produces movement of oppositely charged ions in opposite directions. These flows of ions are the current i. Solvents resist movement of ions through them and this resistance is R, so that $V = iR$, just as in a wire or any other electrical device. People using electrophoresis are interested in ion movement, so the reciprocal of the resistance, called the "conductivity" (σ), is used instead: $\sigma V = i$.

Several factors determine the conductivity of a solution. The conductivity is proportional to the concentrations of the ions present (actually to their thermodynamic activities). It is also proportional to their mobilities or, if the pH is a factor, to their effective mobilities. Since all ions contribute,

$$\sigma \propto \sum \sigma_i \propto \sum (\text{concentration})_i (\text{mobility})_i (\text{average charge})_i$$

The conductivity is also proportional to the cross-sectional area through which the ions move, but the area is usually constant.

The current must be constant in any tube or channel through which electrophoresis is occurring. There can be no gaps or accumulations of current. Since the conductivity in different parts of the tube may be different (see below), the voltage along a tube or channel may also differ in different parts of the tube or channel.

The purpose of electrophoresis is to separate ions and this may be measured from the movements of a large number of molecules using band broadening of scattered laser light owing to Doppler effects (8), by following the movement of the interface or "boundary" between a solution of a particular ion and the solvent, or by following the movement of a region or "zone" containing an ion of interest through a solvent (Fig. 1). Currently, most workers follow the position(s) of relatively thin zone(s) of material, since complete separation of a mixture takes less time the thinner the original zone of the mixture is. However, even the thinnest zone contains a leading boundary at the front and a trailing boundary at the back, so the following discussion applies to zone and boundary methods. We assume the ion of interest is a large one, say a protein.

The large ion is usually restricted at first to one part of the tube or channel through which electrophoresis occurs. The solvent in other parts must also contain ions, usually small ones, to carry the current where the large ion is not present. In other words, the large ion is part of the electrical system and so carries part of the current and contributes to the conductivity. It will move in the same direction as small ions with the same charge, called "co-ions". So the large ion is competing with the co-ions to carry the current.

The large ions must replace equivalent amounts (in terms of charge) of co-ions so that an excess of ions of one charge never accumulates anywhere. In other words, total positive charges must always equal total negative charges (electroneutrality principle). Large ions often have lower mobilities than co-ions. The co-ions are diluted by the large ions with a lower mobility. The moving zone with the large ions has then a lower average conductivity.

Since gaps in the current cannot occur, the zone with the large ions must keep up with the co-ions it is partially replacing. To do this, the voltage in the large ion-containing zone must increase to keep the current constant. If a single large ion diffuses across the leading large ion–solvent boundary, it has moved into a region of higher conductivity and lower voltage. Consequently, the single large ion slows down until the large ion zone catches up (2,7).

At the trailing boundary, co-ions with higher mobilities are replacing lower mobility large ions. This increases the conductivity in the region the large ions just left. The voltage right behind the trailing boundary is consequently lower. If a single large ion diffuses back into that region, it will be in a lower voltage and will move more slowly than the large ions in the large ion zone.

The effect of this competition is to cause zones of large ions with mobilities lower than those of the co-ions to be sharp at the leading boundary and diffuse at the trailing boundary. In other words, zone broadening due to diffusion of the large ions is either reduced or increased at the front (leading) and rear (trailing) edges, respectively, of the zone, depending on the relative mobilities of large ions and co-ions.

There may be effects of pH as well. The small co-ions and counterions are often also used to maintain the pH. The competition between large ions and co-ions, by reducing co-ion movement, leads to changes in the ratio of co-ion to counterion where the large ions have moved in or out of a region. The pH is affected by the ratio of small conjugate acid and base ions, either of which may be the co-ion, so movement of the large ion is accompanied by changes in pH. Changes in pH can change the net charge on the large ion, especially if it is a protein, and so change its effective mobility. Changes in the pH where electrophoresis is occurring can add to or reduce the competition effects described above.

Enhancing Resolution

Separations of substances produced by any procedure will be counteracted by the resulting diffusion of the separated substances, so diffusion should be reduced as much as possible. The competition between ions with the same

charge during electrophoresis can add to or counteract diffusion, and this is the basis of "mobility-based" enhanced resolution techniques.

Mobility-Based Methods. Instead of a zone of solution containing large and small ions set between zones of solvent containing the same small ions, we put different small ions on either side of the large ion zone. This is a "discontinuous" solvent system (2,3,7).

Figure 1b shows an upright cylinder whose lower half is filled with co-ion C and counterion B. The upper half of the cylinder has co-ion A and counterion B. An electric field is applied so that the A ions follow the C ions toward one electrode. If the mobility of the A ion and conductivity of the A solution are less than those of the C solution, the boundary between the solutions will move but diffusion across the boundary by either ion will be restricted. Since the current must be the same all along the tube, the lower conductivity A solution must have a higher voltage making the A ions keep up with the C ions. If A ions diffused ahead into the C solution, the A ions would be in a lower voltage region and would slow until the rest of the A solution caught up. If C ions diffused back into the A solution, the higher voltage there would drive them back to the C solution.

Suppose a solution of ions B and D was inserted between the AB and CB solutions. If ion D has a mobility intermediate between those of A and C, it will remain "sandwiched" between the AB and CB solutions and diffusion will be restricted across the AD and DC boundaries. Suppose further that a mixture of ions DEFG ... of the same charge as A and C were originally present in the intermediate solution and these had mobilities intermediate to those of A and C. Even if DEFG ... were originally mixed together, they would separate and travel, each in separate but adjacent subzones, in order of decreasing mobility going from the C zone to the A zone, all at the same velocity and all with sharply maintained boundaries.

Another effect that occurs in discontinuous systems involves concentrations. If the DEFG ... ions were originally very dilute in their original zone, their zone would compress and each ion's subzone would become relatively thin upon starting the electrophoresis. This is because each ion subzone must carry as much current as the A and C zones, and the higher the conductivities of the AB and CB solutions, the thinner the DEFG ... subzones will get. If DEFG ... are different proteins, which generally have very high molecular weights and relatively low net charges, the subzones will become very thin indeed. The "running" concentrations of DEFG ... will become very high.

If the pattern produced by DEFG ... as they move down the column is determined or if they are made to elute from the column as they emerge from it, the process is called "isotachophoresis", since DEFG ... all move at the same velocity (7). Having DEFG ... unseparated from each other makes analysis or preparation difficult, and isotachophoresis is currently relatively little used.

The most frequently employed procedure to separate DEFG ... is to increase the effective mobility of the A ion so that it runs through the DEFG ... subzones (2,7). These are then in a uniform electric field, that of the AB solution, and so electrophorese independently. While the separating subzones also begin to diffuse, they were "stacked" in very concentrated, and hence thin subzones. The thinner the subzone is before independent electrophoresis, the thinner it will be at the end of electrophoresis and the better the resolution. This approach is employed in "disk electrophoresis".

The A ion is of lower mobility than the C ion and the pH used lowers its effective mobility further. To increase the effective mobility of the A ion, the original CB solution contains a high concentration of the uncharged conjugate acid or base of the B ion. The ions (and conjugate acid or base) are chosen so that when A replaces C its average net charge increases because of the new or "running" pH, which is controlled by the ratio of B ion to its conjugate acid or base.

It must be emphasized that the running concentrations of the ions and uncharged acids or bases are not arbitrary but are controlled by the physical (electrical) constraints mentioned before and alter if they do not conform initially. The constraints are summarized mathematically in Ref. 7.

The importance of the B ion is emphasized by examination of another electrophoretic technique. Suppose the concentration of the B ion is reduced to zero. Since transport of ions must occur in both directions, reduction of the B ion concentration to zero will reduce migration of A, DEFG ..., and C to zero (or nearly zero: in aqueous solvents, there will always be some movement due to ionization of water) (7,9). A, DEFG ..., and C will be stacked. Since B must be replaced, because of electrical neutrality, with protons or hydroxyl ions, amphoteric ions such as proteins will be "isoionic". The pH of each zone or subzone is that at which each amphoteric ion has zero net charge. This technique is termed "isoelectric focusing".

To separate DEFG ... a mixture of relatively low molecular weight amphoteric substances such as "carrier ampholyte" is added before voltage is applied. These substances have a range of isoelectric points, but the pK values of their acidic and basic groups are close to each other, so they stack to produce a buffered pH gradient (7,10). Indeed, they are prepared and sold on the basis of the pH range they cover.

The proteins "band" or collect at their characteristic isoionic pH values. This state does not last indefinitely, however, since the stacked ampholytes and proteins behave like an isotachophoretic system unless the pH gradient is physically immobilized (7) (see below). Isoelectric focusing is frequently used preparatively, since large amounts of protein may be processed.

Supporting Media-Based Methods. A mechanical problem results from any zone separation procedure. Higher resolution results in thinner separated zones. These are more concentrated, and so are significantly denser than the surrounding solvent, which leads to "convection": the zones fall through the solvent and mix with it. This must be prevented. A preparative isoelectric focusing apparatus uses a density gradient of sucrose or other nonionic substance that contains ions DEFG ... and the carrier ampholytes. The separated zones of D, E, and other ions are buoyed by the increasingly dense solvent below them.

Sucrose density gradients have no mechanical strength, so the electrophoresis pathway must be vertical, and in any case the gradient only reduces convection. Generally, a solid or semisolid supporting medium is used. This can be fibrous (paper, cellulose acetate), particulate (cellulose powder, glass beads), or a gel (starch, agar, polyacrylamide). A gel is a network of interacting or tangled fibers or polymers that traps large quantities of solvent in the network. To increase mechanical strength, some polymers can be covalently crosslinked as is routinely done with polyacrylamide supports. Gels usually have a uniform percentage of gel material, but can be prepared with a gradient of gel material or other substances such as urea (10).

Support media should allow as free a passage as possible to electrophoresing ions while restricting convection. Convection (bulk flow) in a capillary is proportional to the fourth power of the capillary radius (1), but the area available for electrophoretic movement is proportional to the square of the radius. Reducing the radius reduces convection much more than the carrying capacity (area) of the capillary.

Capillary tubes can be used without supporting material inside (capillary electrophoresis) (11), but the other materials operate by offering a restricted effective pore size for electrophoretic transport. The effective pore size varies with the medium: 1–2% agar gels have larger average pore sizes than polyacrylamide gels made from 5 to 10% acrylamide solutions.

The importance of the supporting medium is illustrated by an immunoelectrophoretic technique. Electrophoresis is used to separate antigens before antibody diffuses into the separated antigens (immunodiffusion) (12,13). Immunodiffusion requires supports with average pore sizes large enough to allow antibodies, which are large and asymmetric immunoglobins, to diffuse through them, while still stabilizing the zones of precipitate that form. Agar gels, at 1–2% concentrations, are used.

A support medium is in very extensive contact with the solution. Even a capillary tube has a great deal of surface relative to the volume of solution within. Interaction with the support or capillary surface is often not wanted: chemical interactions with the ions being separated or with the other constituents of the solvent interfere with the ion movement. Adsorption of some substances by the fused silica surfaces of capillaries used in electrophoresis can occur (11). The surface must then be coated with inert material. Support media are often chosen for having as little effect on electrophoresis as possible. Such effects are called "electroosmosis": the solvent flows in an electric field (4). If the medium or capillary inner surface has charged groups, these must have counterions. The charged groups fixed in the matrix of the medium or capillary tube and their counterions are hydrated, but only the counterions can move in an electric field. The counterion movement causes solvent flow next to the capillary wall or support medium.

This case is extreme, but some electroosmosis occurs with any medium. Interaction of any two different substances results in a chemical potential gradient across the interface (4). This is equivalent to an electrical potential difference. Electroosmosis can distort zones of separated ions, reducing separation, but as capillaries become thinner or average pore sizes smaller, sideways (to electrophoretic movement) diffusion tends to overcome effects of electroosmosis (7).

Electrophoretic procedures are to produce separations, and if interactions of any kind improve separations, they are desirable. Separation of serum proteins by capillary electrophoresis in a commercial apparatus (Beckman-Coulter Paragon CZE 2000) depends partly on electroosmotic flow stemming from charged groups on the silica surface of the capillary.

Electrophoresis in solutions containing ionic detergents can separate uncharged substances. Detergents form clusters in water solutions called "micelles" with hydrophilic surfaces and hydrophobic interiors. Micelles of ionic detergents move in an electric field, and uncharged substances, if they partition into the micelles, will be moved along in proportion to the extent of their partitioning. This is "micellar electrokinetic chromatography" (11).

Electrophoresis in support media that reversibly adsorb substances being separated is called "electrochromatography". Unlike conventional chromatography, which uses hydrostatic pressure to move solutions of substances to be separated through the medium, electrochromatography uses an electric field.

The interactions with support media used to increase resolution of electrophoretic separations range from indirect to actual adsorption.

Acrylamide derivatives that will act as buffers over a desired pH range are incorporated into polyacrylamide gels and used in isoelectric focusing (7,9,10). These are called Immobilines. They provide stable pH gradients (7) and control movement of amphoteric ions by controlling the pH. They also stabilize zones of protein that form.

Electrophoretic or diffusional movement of large ions is reduced by supports, since they increase the tortuousness of the path the ions must follow. If the average pore size is made comparable to the effective sizes of the ions undergoing electrophoresis, the ions are literally filtered though the medium so that smaller ions would be least retarded and larger ions retarded more. This "molecular sieving" effect can improve the resolution achievable by electrophoretic techniques: the leading edges of zones of large ions are slowed by the gel matrix allowing the rest of the zones to catch up. This makes zones of larger ions thinner and restricts subsequent diffusion.

This effect can also provide information about the physical characteristics of the ions. The effective mobilities of large ions are determined by their net charges and frictional coefficients. The latter is affected by the shape of the large ion: A very asymmetric molecule has a larger effective size than a more compact one (1,3). Two molecules of identical charge and molecular weight can be separated even in the absence of a support if they have sufficiently different shapes. Effective sizes can be determined from electrophoresis on several gels with different average pore sizes. Effective sizes can be used to calculate diffusion constants (1,3), though they are normally used to obtain molecular weights. Large ions of known molecular weights may be electrophoresed on the same gels to calibrate the pore sizes of the gels (12,13) (see below).

If a mixture of large ions of differing shapes and sizes is converted to a mixture with different sizes but the same shape, then they will migrate according to their net charges and molecular weights. If the charges are made a constant function of the molecular weight then the ions will separate according to their molecular weights alone (3,10,12,13). The degree of separation will also be affected by the average pore size of the medium. Proteins that consist entirely or nearly so of amino acids and that have been treated to reduce any disulfide bonds are usually converted to rod-like structures with a nearly constant charge-to-mass ratio by adding the detergent, sodium dodecyl sulfate (SDS). The detergent binds to such proteins in large and highly uniform amounts (3,12,13). The protein–SDS complexes then migrate in order of decreasing polypeptide molecular weight (see below). It is important to recognize that some proteins may not behave "normally" in these respects, for example, if they contain large amounts of carbohydrate. The SDS–PAGE method, as it is called, is currently probably the single most widely used electrophoresis technique in research laboratories.

Polynucleotide fragments have a relatively constant charge-to-mass ratio at physiological pH values, since each nucleotide is of similar molecular weight and has only a single charged group: the phosphate. If they have the same shape, separation of such fragments can be effected on the basis of molecular weight (i.e., number of nucleotides) by electrophoresis (10,13). This is the basis of the technique for determining polynucleotide sequences (see below).

It is possible to separate and measure molecular weights of native (double-stranded) DNA molecules that are larger than the average pore size of the support medium (14). The electric field pulls on all segments of the DNA equally, but eventually, because of random thermal movement, one end is farther along the field direction and the entire molecule is pulled after that end along a path through the gel. If the field is turned off, the stretched molecule contracts or relaxes into a more compact, unstretched condition. If the electric field is turned on again, the process starts over. The rates of these processes are slower with longer DNAs, so, if the rate of change of field direction is appropriate, the DNAs will separate according to size. If not, they will tend to run together. The change in direction of the applied field and the length of time the field is applied in any direction can be adjusted in commercial instruments. The gel used is agarose (agar without the agaropectin). This is "pulsed-field" gel electrophoresis, and can separate DNA molecules as large as small chromosomes.

Sometimes adsorption strong enough to terminate electrophoretic movement is desired. Substances are separated on a sheet or thin slab of ordinary support medium, then an electric field is applied perpendicular to the face of the sheet and the separated substances are electrophoresed onto a facing sheet of adsorbent material such as nitrocellulose paper. This is "electroblotting" (13,15,16). It concentrates the electroeluted substances on the adsorbent sheet. This makes immunological detection (immunoblotting) or detection by any other method more sensitive (15,16). If the separated substances blotted are DNA, this is "Southern blotting" (17); RNA, "Northern blotting"; protein, "Western blotting" (10,13,15,16). These procedures are to identify particular proteins or polynucleotides with particular sequences.

Cells in living organisms produce tens of thousands of proteins and polynucleotides, and in very different amounts: for example, the concentrations of serum proteins differ by up to 10^7-fold, complicating analysis. Increased resolution has been also achieved through use of two-dimensional (2D) electrophoresis. Some 30–60 proteins can be resolved on a column or sheet of polyacrylamide. If electrophoresis is done in two dimensions, with separation on the basis of different properties of the proteins, 900–3600 proteins could be resolved (18,19). The most popular form of 2D electrophoresis involves isoelectric focusing on a gel strip for one dimension, attaching the gel strip to a gel sheet, and performing SDS–gel electrophoresis for the second dimension. Densitometers to scan the gels and software for computerized mapping and indexing of isolated proteins to help analyze the data are available commercially, for example, from BioRad Corp. and Amersham Biosciences Corp. The procedure is laborious and exacting; many factors affect the patterns obtained (19). Use of Immobiline gel strips for the first dimension separation is very important in improving reproducibility (19). The realized or potential information from 2D electrophoresis is so great that it is a very popular technique in research laboratories.

Voltage-Based Methods. Since diffusion is the enemy of separation and increases with the square root of the time (1), it is obviously better to carry out the electrophoresis faster. Since the velocity of electrophoresis increases with the voltage applied, using a higher voltage would improve resolution. However, increasing the voltage also increases the current and this increases electrophoretic heating, called "Joule heating". Joule heating is proportional to the wattage (volts times current) and can cause convective disturbances in solutions, even if a support medium is present. Joule heating is part of the reason electrophoresis is done in solutions with moderate conductivities, equivalent to 0.1 *M* NaCl or thereabouts (13): Ion concentrations must be high enough to minimize electrostatic interactions between large molecules (1), but high ion concentrations mean high conductivities, which means high current flow at a given voltage and therefore high heat generation. Joule heating is in fact the ultimate limiting factor in any electrophoretic procedure.

Use of thinner supports or capillaries minimizes the effects of heating. Thin (< 1 mm thick) sheets of polyacrylamide allow more efficient heat dissipation (from both sides) and are also useful when comparing samples. They are necessary for 2D electrophoresis (18,19). Thinner supports use less material but require more sensitivity of analysis. Capillaries are even more easily cooled (from all sides) and very high voltages, of the order of several thousand volts, can be used (11). The Beckman Coulter Paragon CZE 2000 has seven capillaries so can accommodate seven samples at once, and serum protein electrophoresis turnaround times are 10 min. Very small sample volumes, often a few nanoliters, are used. On the other hand, electroosmosis can be a major problem and the detection sensitivity is very low: the separated substances

pass a detector, and the capillaries, usually < 0.1 mm in diameter, provide a very short optical pathlength. Detection at very short wavelengths such as 214 nm, where extinction coefficients are higher or use of other measuring methods such as fluorescence or mass spectrometry are alternatives that are employed. Still, the ability to use higher voltages for separation, thus improving resolution, has directed much attention to capillary electrophoresis.

EQUIPMENT AND PROCEDURES

The equipment for electrophoresis includes a power supply, the apparatus on which electrophoresis is actually performed, and the reagents through which electromigration occurs.

Power Supply

This is sometimes combined with the apparatus. The power supply provides a dc voltage that produces the electrophoretic movement. Maximum power outputs should be matched with requirements.

Some power supplies may provide for applying a constant voltage, constant power or constant current. The conductivity of a setup changes with time during electrophoresis because of ion movement, so constant voltage is used for ordinary electrophoresis, constant current for isotachophoresis or disc electrophoresis (to make the zone migration velocity constant) and constant wattage for isoelectric focusing (10). Many power supplies also feature timers to prevent running samples too long. Most have an automatic shutdown of power if the apparatus is accidentally opened. Because even low voltages can be dangerous, safety issues must be addressed.

Apparatus

Many types are available, from simple partitioned plastic boxes (e.g., from Beckman-Coulter) to large and elaborate ones (e.g., from Helena). All basically hold the material through which electromigration occurs, horizontally or vertically, either as a plug or cylindrical rod or, most commonly for routine clinical work, as a sheet. Sizes and shapes of electrophoresis chambers are usually determined by the cooling efficiency of the apparatus. Some have special chambers for circulating cold water and some instruments have incorporated cooling units. These may be able to electrophorese a few samples at a time or many.

Horizontal electrophoresis is preferred when the supporting medium cannot stand the mechanical stress of standing upright. Evaporation from the top surface is often a problem; DNA fragments are separated in "submarine" systems, where the support (an agarose gel) is covered in buffer. The supports used when electrophoresis is done vertically are sometimes encased in glass sheets.

The apparatus also provides for the connection between the electrophoresis chamber and the electrodes. Typically, the connection is made through two relatively large volumes of solution, or "reservoirs". These reservoirs minimize the effects of electrolysis. To reduce contamination from extraneous metal ions, the electrodes are normally made of platinum.

Accessories include drying ovens and containers for staining and washing supports, though these may be integrated with the rest of the apparatus.

Reagents

The reagents constitute the supporting medium and provide the ions for electrical conduction throughout much of the apparatus. The ionic substances should be of good purity, but the major requirement is that the ions should not interact strongly with the substances to be separated (like borate with glycoproteins) unless a specific interaction (e.g., SDS) is desired. This means low heavy metal content and avoidance of large polyanions. Generally, buffers and salts with low ionic charges (no greater than phosphate at pH 7 or so) are preferred.

The reagents for supporting media prepared in the laboratory are usually "electrophoresis grade" for the reasons just given; that is, ionic contaminants such as acrylic acid in acrylamide are preferred to be at low levels. Some workers claim that cheaper grades of many reagents are satisfactory. In our experience, the quality of SDS seems to be important, however.

For research work, gel supports are prepared by heating slurries of agar or agarose in buffer or polymerizing acrylamide and derivatives in buffer and pouring them into a support to set. A "comb" with wide teeth is placed in the cooling or polymerizing liquid to form a series of identical rectangular holes or depressions called "wells" in the support. The samples are mixed with sucrose or glycerol (to make them dense enough to fall into the wells) and also with a low molecular weight dye called "tracking dye" (to visibly mark the progress of electrophoresis), then pipetted into the wells before electrophoresis. For routine clinical use, samples of a few microliters of samples not mixed with sucrose or tracking dye are applied to the surface of the support through a template or mask and allowed to soak into the support a few minutes before electrophoresis.

Commercially prepared supports are convenient and tend to be more uniform, so are employed for routine clinical analysis. These include agarose, cellulose acetate and polyacrylamide. They have plastic sheets as backing for strength. Washing or preelectrophoresing these before use is usually unnecessary.

Measurements

This means determining velocity(ies) or location(s) of the substance(s) electrophoresed. In the cases of isotachophoresis or capillary electrophoresis, measurements of absorbence, fluorescence, heat generation, conductivity, and so on are made while the separation is occurring. If the substances are allowed to elute and are collected in separate fractions, analysis may be done at leisure using any appropriate technique.

Usually, the (hopefully) separated substances remain on the supporting medium; however, diffusion continues after electrophoresis and measurements must be made quickly or diffusion slowed or stopped after electrophoresis ends. Fluorescent labeled single-stranded polynucleotides are measured on DNA sequencing gels (polyacrylamide) using a scanning fluorometer. Double-stranded (native)

DNA is detected using dyes that become fluorescent when the dye molecules move between (intercalate) stacked nucleotide bases, which provides great sensitivity and selectivity. Photographs are made of the patterns. Direct measurement of proteins on the support is difficult to do as the support itself often absorbs in the ultraviolet (UV) or scatters. Transparency of cellulose acetate supports can be increased (clearing) by soaking the support in methanol-acetic acid before measuring absorbence. A scanning spectrophotometer must be used. Otherwise, the support is dried or the separated substances, usually proteins, precipitated in the support.

With autoradiography, radioactivity is measured, most often using photographic film (13) after the support is dried. Autoradiography is potentially the most sensitive method of measurement of distributions, since isotopes of very high specific activity can be used.

Precipitation of proteins in the support matrix is usually done by soaking the matrix in 7–10% acetic acid, though trichloroacetic acid is sometimes used. Any SDS must be removed or precipitation does not occur. Adding methanol (30–40%) to the acetic acid and soaking is usually done to remove SDS. The next step is staining using a dye: This is to enhance visibility, for purposes of inspection, photography, or, if a quantitative record is needed (generally the case in routine clinical work), measurement of the absorbence using a scanning spectrophotometer (densitometry). Some commercial clinical electrophoresis apparatuses (e.g., Helena) have automated staining; otherwise, it is done manually.

The support is soaked in a solution of a dye that strongly and specifically adsorbs to the denatured protein. Excess dye is removed by electrophoresis or by diffusion: subsequent soaking in the same solution but without the dye. This leaves a support with colored bands, dots, or stripes where protein can be found.

The important parameters for detection are specificity and sensitivity. One may be interested in staining specifically for phosphoproteins or lipoproteins, for example. Otherwise, the benchmark parameter is sensitivity. Tables showing frequently used stains are given in Refs. 2 and 13. The sensitivity of soluble protein stains increases with the molar extinction coefficient of the dye (20). Substances with very high effective extinction coefficients are more sensitive, such as metal stains (e.g., silver stain) or colloidal stains such as India ink or colloidal gold. Using silver staining, as little protein as 0.1 ng/band can be detected (13).

Sometimes some specific property must be measured, to identify a particular band or spot. With enzymes, the activity is the most specific property available, and two means of identification can be used.

If a substrate or product of the enzyme has a distinctive color, the support is soaked in an assay medium containing the reactant until a color change due to product (or substrate) appears. Naturally, the support and electrophoresis conditions must be such that the activity was not lost beforehand. If substrate or product is of low molecular weight, some way of restricting its diffusion must be used. Enzyme activity-specific stains are called "zymograms" (21).

Alternatively, the support can be cut into successive sections, incubated in assay medium, and absorbence or other changes resulting from enzyme activity measured. Generally, a duplicate support or half of a single support is stained for protein for comparison.

Recovery of material from supports is sometimes desired, for example, for mass spectrometry. The ease of recovery of large molecules such as proteins by diffusion is proportional to the pore size of the supporting medium. Recovery (especially from acrylamide gels) is sometimes done using electrophoresis.

The times required for electrophoresis, including staining and densitometry, are of the order of hours. This has limited clinical applications of electrophoretic techniques, though capillary electrophoresis techniques may change this.

EVALUATION

The result is a pattern or distribution profile given as a function of distance migrated. Often this is expressed as R_f values, relative to the distance migrated by a known ion, usually tracking dye. Electrophoresis of proteins without denaturants (native gel electrophoresis) on several gels of different percentages (pore sizes) is followed by plotting logarithms of their R_f values versus gel concentration, a "Ferguson plot" (12,13). This plot enables separation of the contributions of size and charge: isozymes, for example, have the same size (slopes of the lines) but different charges (ordinate intercepts). In the presence of SDS, protein complexes are dissociated (3,10,12,13). A protein of unknown subunit molecular weight is electrophoresed alongside of a mixture of proteins of known subunit molecular weight in the presence of SDS. The logarithms of the known molecular weights are plotted versus their R_f values, yielding an approximate straight line, and the unknown molecular weight obtained from its R_f value.

However, the pattern is often presented without further analysis. Electrophoresis experiments are often used to assess the purity or "homogeneity" of a preparation of a macromolecule. Electrophoresis is still the single most widely used criterion of purity. It is best if purity is tested under widely different electrophoresis conditions: different pH values, or the presence of absence of denaturants, such as urea or SDS, in the case of proteins.

Isoelectric points can be obtained most readily by isoelectric focusing, which can also be used as a criterion of purity.

A major use of electrophoresis currently is for DNA sequencing, owing to ongoing determinations of entire genomes of organisms. Capillary electrophoresis using linear (uncross-linked) polyacrylamide has replaced use of polyacrylamide sheets as the method of choice for this purpose, since throughput is increased several-fold because of the higher voltages that can be applied. Applied Biosystems has an instrument with 96 capillaries and Amersham Biosciences has models with up to 384 capillaries. Analysis of the profiles is done by the instruments: each fluorescent peak observed is assigned to adenine, cytosine, guanine, or thymine on the basis of the emission

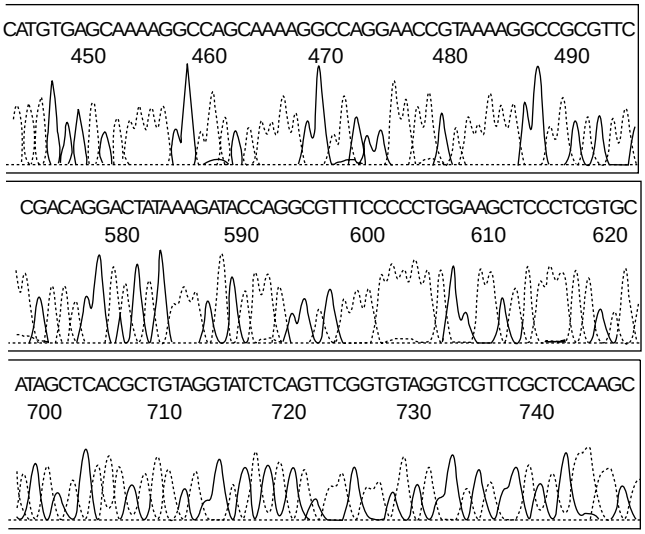


Figure 2. Section of data printout from the Applied Biosystems Prism capillary electrophoresis apparatus. The profile shows signals (fluorescence) given by polynucleotides with a fluorescent nucleotide analogue at the end, obtained as these migrate past the detector. Each peak is produced by a polynucleotide that is one residue (nucleotide) longer than that producing the signal to its left and one residue shorter than the polynucleotide producing the peak to its right. The separation by molecular length is produced by electrophoresis through linear (uncrosslinked) polyacrylamide. Different nucleotide analogues give different emission maxima, so the output identifies these as A, C, G, or T. Note the increasing peak widths due to diffusion. The numbers are the lengths in residues of the polynucleotides. (Courtesy of Dr. John Wunderlich of the Molecular Genetics Instrumentation Facility of the University of Georgia.)

spectra observed as polynucleotides terminated by dideoxynucleotide derivatives (22) with different emission spectra electrophorese past the detector. Data from each capillary are printed separately (Fig. 2).

Routine clinical electrophoresis separations are mostly serum proteins, hemoglobins, or isozymes: creatine kinase, lactic dehydrogenase, and alkaline phosphatase. Then quantitation of the separated proteins is done.

Usually, the proteins are separated on agarose or cellulose acetate sheets, stained, and the resulting pattern scanned using a densitometer. The Beckman-Coulter CZE 2000 instrument with seven capillaries is approved by the FDA for serum protein analysis, and directly measures absorbencies as different proteins move past the detector. In either situation, the electrophoretic patterns from samples (blood serum, urine, or cerebrospinal fluid) obtained from patients are compared with those from samples taken from healthy people (Fig. 3). Higher immunoglobin levels are seen with a number of conditions such as liver disease and chronic infections for example. The profiles may be analyzed visually, but some densitometers have analytical capabilities. Sometimes fluorescence [usually of reduced nicotinamide adenine denucleotide (NADH) in creatine kinase or lactic dehydrogenase isozyme assays] is scanned instead.

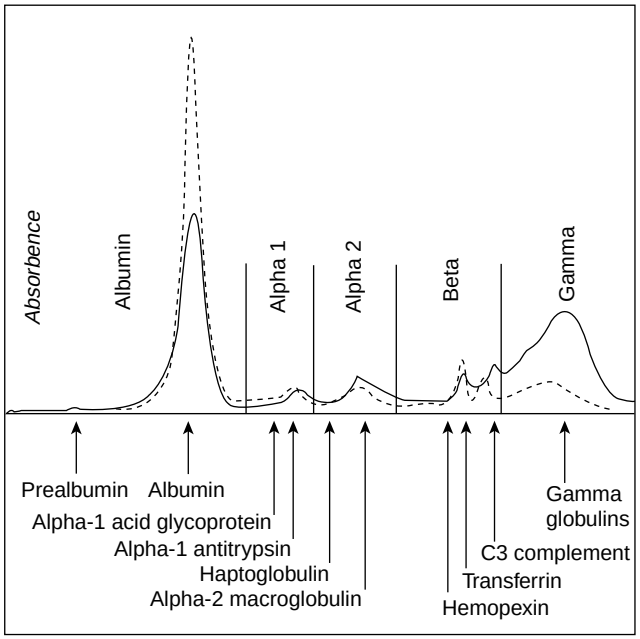


Figure 3. Recordings of absorbence at 214 nm versus time of serum protein electrophoresis on the Beckman-Coulter Paragon CZE 2000. The solid line is the electrophoretogram of a serum protein sample from a person with chronic hepatitis C, relative to a sample from a healthy person (broken line). Characteristic increase in the polyclonal gamma and decreases in albumin and transferrin zones occur. (Courtesy of Beckman-Coulter Inc.)

Reliability

Resolution of proteins whose isoelectric points differ by as little as 0.001 pH unit is claimed for isoelectric focusing using Immobiline gels (13).

The resolving power of electrophoresis on polyacrylamide is sufficiently high that hundreds of polynucleotides, each differing in length from one another by one nucleotide, are separated in DNA sequencing (Fig. 2). Estimates of molecular weights of proteins in SDS–gel electrophoresis is also very widely done. Small differences, of the order of 1000 Da, in molecular weights can be detected (13); absolute molecular weight estimates are less reliable, since the SDS binding depends on the composition of the protein (13). In the author's experience, most protein molecular weights estimated by this technique are not reliable beyond average limits of $\pm 10\%$. It is best to remember that "molecular sieving" techniques in general measure effective sizes, and estimates of molecular weight involves comparisons with proteins of known mass whose shapes and hydrations are assumed to be similar.

While electrophoresis results are generally fairly reproducible from experiment to experiment, it is best to use internal standards whenever appropriate. This is for location and identification of separated substances and for quantitation.

Electrophoresis is widely used, especially in the biological sciences for investigations of macromolecules. It is an excellent analytical method characterized by high resolution and sensitivity and moderately good reproducibility, a method capable of yielding considerable information about

size, shape, composition, and interactions of specific molecules and about distribution of large numbers of molecules at one time.

For clinical purposes, the resolution of isozymes such as lactic dehydrogenase is reliable enough that the technique has been used in critical care diagnoses, that is, life or death decisions.

BIBLIOGRAPHY

Cited References

Note: Developments in this field are covered particularly in the journal Electrophoresis. The catalogues of companies that sell electrophoresis apparatus and supplies, such as Amersham Biosciences, BioRad, Invitrogen, and National Diagnostics are another sources of information.

1. Tanford C. Physical Chemistry of Macromolecules. New York: John Wiley & Sons; 1961.
2. Brewer JM. Electrophoresis. In: Kaplan LA, Pesce AJ, Kazmierczak SC, editors. Clinical Chemistry: Theory, Analysis, Correlation. 4th ed. St. Louis (MO): Mosby; 2003. Chapt. 10, p 201–215.
3. Cantor CR, Schimmel PR. Biophysical Chemistry. San Francisco: W.H. Freeman and Company; 1980.
4. Moore WJ. Physical Chemistry. 4th ed. Englewood Cliffs (NJ): Prentice-Hall; 1972.
5. Kiriukhin MY, Collins KD. Dynamic hydration numbers for biologically important ions. Biophys Chem 2002; 99:155–168.
6. Rupley JA, Gratton E, Careri G. Water and globular proteins. Trends Biochem Sci 1983;8:18–22.
7. Mosher RA, Saville DA, Thormann W. The Dynamics of Electrophoresis. Weinheim, Germany: VCH; 1992.
8. Marshall AG. Biophysical Chemistry: Principles, Techniques and Applications. New York: John Wiley & Sons; 1978.
9. Righetti PG. Immobilized pH gradients, theory and methodology. Amsterdam, The Netherlands: Elsevier; 1990.
10. Hawcroft DM. Electrophoresis: the basics. Oxford (UK): Oxford University Press; 1997.
11. Whatley H. Basic Principles and Modes of Capillary Electrophoresis. In: Petersen JR, Mohammad AA, editors. Clinical and Forensic Applications of Capillary Electrophoresis. Totowa, NJ: Humana Press; 2001. Chapt. 2, p 21–58.
12. Van Holde KE. Physical Biochemistry. 2nd ed. Engelwood Cliffs (NJ): Prentice-Hall; 1985.
13. Dunn MJ. Gel Electrophoresis: Proteins. Oxford (UK): Bios Scientific; 1993.
14. Noolandi J. Theory of DNA gel electrophoresis. In: Chrambach A, Dunn MJ, Radola BJ, editors. Volume 5, Advances in Electrophoresis. New York: VCH; 1992. p 1–57.
15. Gallagher SR, Winston SE, Fuller SA, Harrell JGR. Immunoblotting and immunodetection. In: Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K, editors. Current protocols in molecular biology. Unit 10.8, New York: Greene Publishing and Wiley Interscience; 2000.
16. Baldo BA. Protein blotting: Research, Applications and its place in protein separation methodology. In: Chrambach A, Dunn MJ, Radola BJ, editors. Advances in Electrophoresis. Volume 7, New York: VCH; 1994. p 409–478.
17. Highsmith WE, Jr., Constantine NT, Friedman KJ. Molecular Diagnostics. In: Kaplan LA, Pesce AJ, Kazmierczak SC, editors. Clinical Chemistry: Theory, Analysis, Correlation. 4th ed. St. Louis (MO): Mosby; 2003. Chapt. 48, p 937–959.
18. Hochstrasser DF, Tissot JD. Clinical application of high-resolution two-dimensional gel electrophoresis. In: Chrambach A, Dunn MJ, Radola BJ, editors. Advances in Electrophoresis. Volume 6, New York: VCH; 1993. p 270–375.
19. Gorg A, Obermaier C, Boguth G, Harder A, Scheibe B, Wildgruber R, Weiss W. The current state of two- dimensional electrophoresis with immobilized pH gradients. Electrophoresis 2000;21:1037–1053.
20. Merril CR. Gel-staining techniques. Volume 182, Methods Enzymology. New York: Academic Press; 1990. p 477–488.
21. Heeb MJ, Gabriel O. Enzyme localization in gels. Volume 104, Methods Enzymology. New York: Academic Press; 1984. p 416–439.
22. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977;74:5463–5467.

See also CHROMATOGRAPHY; DNA SEQUENCE.

ELECTROPHYSIOLOGY

SEMAHAT S. DEMIR
The University of Memphis and
The University of Tennessee
Memphis, Tennessee

INTRODUCTION

The Resting Membrane Potential

The cell membrane or sarcolemma, composed of lipid bilayer, is hydrophobic and is highly impermeable to most water-soluble molecules and ions. A potential is developed across the lipid bilayer of the cell membrane due to unequal distribution of charges on the two sides of the membrane, thus the membrane acts as a capacitor. Membrane proteins that span across cell membranes form ion channels allowing transport of small water-soluble ions across the membrane. These channels are highly selective and their selectivity depends on diameter, shape of the ion channel, on the distribution of charged amino acids in its lining (1). The movements of the ions through these channels across the membrane govern its potential.

The transport of the ions across the membrane is either passive or active. The passive mechanism of transport of any ion is governed by its electrochemical gradient, a combination of chemical force exerted by diffusion of ions due to concentration gradient and an electrical force exerted by the electric field developed due the charges accumulated on either side of the membrane (capacitor) (2). Physiologically, cells have a high intracellular potassium concentration, $[K^+]_i$, and a low sodium concentration, $[Na^+]_i$. Conversely, the extracellular medium is high in Na^+ and low in K^+. In cardiac cells at rest, the membrane is mostly permeable to K^+ ions through K^+ leak channels. As K^+ flows out down its concentration gradient, a negative potential is built up inside the cell. This increases as long as it counterbalances the chemical driving force generated by concentration gradient. This potential at which the net ion flux is zero is called Nernst equilibrium potential of that ion. The equilibrium potential for K^+ is given by its

Nernst equation:

$$E_K = \frac{RT}{ZF} \ln \frac{[K^+]_o}{[K^+]_i}$$

where R is the universal gas constant, T is temperature in kelvin, F is Faraday constant, z is the valency of the ion.

Typical resting membrane potentials of excitable cells vary from -90 to -50 mV, depending on the type of the cell. Epithelial cell and erythrocytes have smaller, but still negative membrane potentials. It may tend toward the excitatory threshold for an action potential as in a cardiac pacemaker cell or remain stable with approximately no net ion flux observed in nonpaced cardiac ventricular cells. As the ventricular cell at rest is more permeable to K^+ ions than to any other ion, the resting membrane potential (ca. -84 mV) is close to E_K at 37 °C. Due to its permeability to other ions and also due to other transport mechanisms the resting membrane potential does not reach exactly E_K.

The active ionic transport mechanisms maintain the homeostasis of ionic concentrations in both the intra- and extracellular media. These membrane proteins are called carrier (pump) proteins and they utilize energy from hydrolysis of adenosine triphosphate (ATP) to transport ions against their concentration gradient.

EXPERIMENTAL TECHNIQUES TO QUANTIFY IONIC MECHANISMS IN CELLS

The advent of patch clamp technique (3–7) has made it possible to record the current from a single ion channel. The technique involves clamping a patch of the cell membrane and recording either voltage (current–clamp) or current (voltage–clamp or patch–clamp) across the membrane. Using this technique current of order as low as 10^{-12} A can be measured. This could be done using different configurations of patch clamping (7).

1. A freshly made glass pipette with a tip diameter of only a few micrometers is pressed gently on the cell membrane to form a gigohm seal. This is called as cell-attached patch configuration. The pipette solutions form the extracellular solution and the currents across the channel within the patch can be recorded.
2. When gentle suction is applied to the pipette in cell-attached configuration, the membrane ruptures while maintaining the tight seal and the cytoplasm and pipette solution start to mix. After a short time, this mixing is complete and the ionic environment in the cell is similar to the filling solution used in the pipette. This configuration is called whole-cell patch configuration. A recording obtained using this configuration is from whole cell and not from a patch. The advantage of this technique is that the intracellular environment is accessible through the pipette. Current–clamp technique is used in this configuration to measure the action potentials (APs) of excitable cells.
3. Sudden pulling out of the pipette from cell-attached configuration holds the patch that formed the gigohm seal giving raise to the inside–out configuration (inside of the cell membrane is exposed to external bath).
4. Slow pulling out of the pipette from whole cell configuration holds the patch that formed the gig-ohm seal giving rise to outside–out configuration. Both inside–out and outside–out configurations allow single channel recordings. Both the intracellular and extracellular baths are accessible in these cases.
5. The fifth configuration is obtained by creating artificial channels (permealizing membrane) on the cell-attached patch by administering antibiotics, like amphotericin. The voltage and current–clamping recordings obtained in this configuration recordings are similar to whole-cell recordings, the advantage being the intracellular medium is not dialyzed.

VOLTAGE-CLAMP TECHNIQUE

The method of voltage clamping has been the primary experimental tool used to reconstruct models of cellular electrical activity. As the behavior of the ion channels is highly nonlinear under changing action potential, this method enables us to quantify their properties by holding the transmembrane potential (membrane potential) at a particular voltage. The basic principle relies on providing current to balance those currents through the ionic channels that are open and thus the transmembrane voltage is clamped at a chosen constant level (clamp voltage) For example, if the Na^+ channel is studied, the membrane potential is initially held at rest. When this potential is changed instantaneously to a depolarized (more positive) potential, sodium channels open and Na^+ ions tends to move in. The voltage amplifier senses these small changes in voltage and a feedback current of equivalent amount is applied in opposite direction of the ion flow. This measurable current changes for different clamp potentials as the driving force (V_{Clamp}–E_{Na}), and the gating parameters at that V_{Clamp} changes enabling us to quantify the channel properties. Ion channels conduct ions at a rate sufficiently high that the flux through a single channel can be detected electrically using patch clamp technique. The basic circuit of voltage clamp setup is shown in Fig. 1.

CURRENT–CLAMP TECHNIQUE

The current–clamp technique is used to record action potentials. This technique involves clamping the cell in whole cell configuration and applying a suprathreshold current pulse for a short duration until the Na^+ channels start to activate. The transmembrane change in voltage gives the action potential recording.

A key concept to modeling of excitable cells is the idea of ion channel selectivity of the cell membrane. As the molecular behavior of channels is not known, modeling of

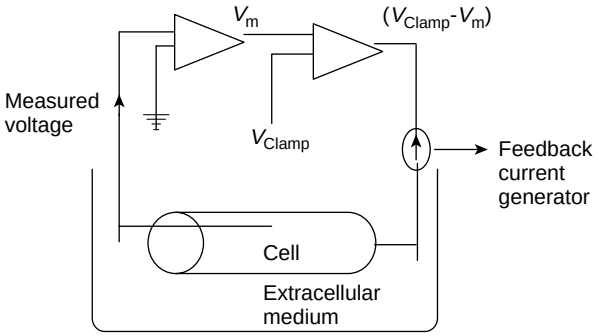


Figure 1. Simplified circuit representation of a voltage–clamp setup (8).

nonlinear empirical models of membrane processes help us to understand the role of various currents in the depolarization and repolarization phases and the phenomena that involve the interaction between these processes.

EXAMPLE OF ACTION POTENTIAL AND UNDERLYING IONIC BASIS: THE CARDIAC ACTION POTENTIAL

Membrane excitability is the fundamental property of the nerve and muscle cells, that is, in response to certain environmental stimuli, generates an all-or-none electrical signal or AP. Many different types of cardiac ion channels and ion pumps altogether form a complex process that results in cardiac AP (9). For example, the normal cardiac action potentials can be classified into two broad categories; those that are self-oscillatory in nature, such as pacemaker cells (sinoatrial and atrioventrciular cells) and those that need an external stimulus above a threshold, also called supra threshold, in order to be evoked, such as atrial, Purkinjee fiber, or ventricular cells. An extraordinary diversity in the action potential configurations can be seen in different regions of the heart. The ventricular tissue in particular displays a wide variety of action potential waveforms. These APs include pacemaker potentials in purkinje cells, and disparate action potential durations (APD) and morphologies in cells from the epicardial, mid-myocardial, and the endocardial layers of the ventricle. The ventricular action potential has been studied more frequently than other representative cardiac membrane potentials because ventricular arrhythmias are believed to constitute the majority of reportedly fatal incidences of cardiac arrhythmias (10).

Phases of Cardiac Action Potential

A typical ventricular action potential in higher mammals such as canine and human consists of four distinct phases (Fig. 1a). Phase 0 corresponds to a rapid depolarization or upstroke of the membrane action potential. Phase 1 is the initial rapid repolarization, and is followed by phase 2, which constitutes the action potential plateau. Phase 3 represents the final repolarization, which allows the ventricular cell to return to its resting state in phase 4. In addition to its morphological features, ventricular APs are commonly measured experimentally to determine its characteristics. These include the resting membrane potential (V_{rest}), the peak overshoot (PO) value that is the maximum positive value achieved during the initial phase 0 depolarization, the maximum upstroke velocity (dV/dt_{max}), which occurs during phase 0, and the APDs measured when the APs have repolarized to 50 and 90% of their final repolarization value, also called APD_{50} and APD_{90}, respectively.

One or more of these characteristics is usually altered in the setting of a pathophysiological condition, and helps to quantify the differences between the normal and the abnormal action potentials.

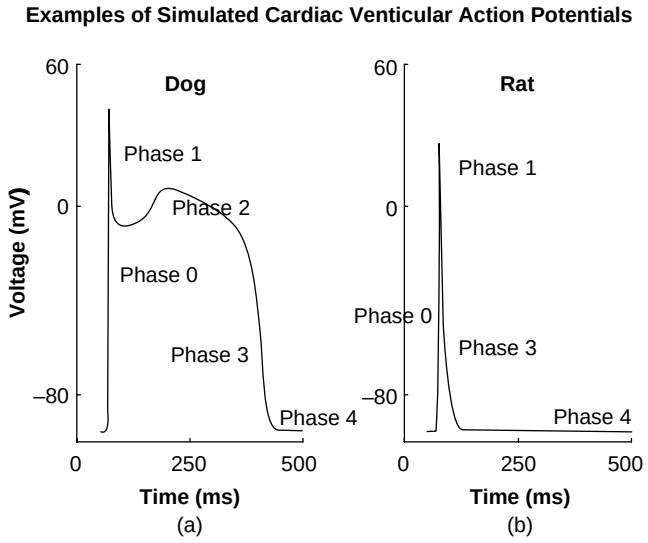


Figure 2. Examples of typical cardiac action potential waveforms of the dog (11) and rat (12) ventricular cell models. The action potential of the human ventricular cell would be similar to that of the dog (panel a). Please consult Tables 1 and 2 for the underlying ionic currents.

Contributions of Ion Channels to Cardiac Action Potential

The temporal changes in a typical ventricular action potential configuration, that is, depolarization followed by a repolarization (Fig. 2) are governed by the movement of different ions, such as Na^+, K^+, and Ca^{2+} ions across the sarcolemma. These ions are usually transported between the intracellular and the extracellular spaces by means of carrier proteins and channel proteins embedded in the cardiac membrane. These proteins form the passive (ion channel-mediated and carrier-mediated) and active (carrier proteins such as pumps and exchanger) transporters of the cell membrane. The ion channels, the pumps and exchanger are the major ionic currents that form an action potential in mathematical representations. A summary of these currents that are present in a typical cardiac ventricular cell and the role of the currents in action potential generation are summarized in Table 1. The major ionic currents contributing to different phases of the typical action potential are presented in Table 2. The ventricular action potential is the result of a delicate balance of the inward and outward ionic currents and the active transporters (pumps and exchangers).

Table 1. Major Ionic Membrane Mechanisms Underlying a Typical Cardiac Ventricular Action Potential

Membrane Mechanism	Description	Gene (α-subunit)	Role in Action Potential
	Inward Ionic Currents		
I_{Na}	Na^+current	SCN5A	Initial depolarization of action potential
I_{CaL}	L-type Ca^{2+} current	α_{1C}, α_{1D}	Maintains plateau phase of action potential
I_{CaT}	T-type Ca^{2+} current	α_{1G}, α_{1H}	Present in the late plateau phase
	Outward Ionic Currents		
I_t	Ca^{2+}-independent transient outward K^+ current	Kv4.2 Kv4.3 Kv1.4	Responsible for early repolarization
I_{Kr},I_{Ks}	Rapid and slow delayed K^+ rectifier currents	HERG KvLQT1	Aids repolarization during plateau
I_{ss},I_{Kslow}	Slow inactivating K^+ currents	Kv2.1 Kv1.5	Aids late repolarization
I_{K1}	Inward rectifier K^+ current	Kir2.1 Kir2.2	Late repolarization, helps establish V_{rest}
	Other Ionic Currents		
I_{NaCa}	Na^+–Ca^{2+} exchanger current	NCX1 NCX2	Late depolarization
I_{NaK}	Na^+–K^+ pump current	Na^+-K^+-ATPase (α)	Late repolarization

To restore the intracellular and extracellular ionic concentrations so that homeostasis is maintained, the ions that cross the cell membrane during an action potential are brought back by active mechanisms like Na^+–K^+ pump, Ca^{2+} pump and coupled transporters like Na^+–Ca^{2+} exchanger, Na^+–H^+ exchanger. All the active mechanisms utilize hydrolysis of adenosine triphosphate (ATP), the cellular source of energy to achieve this. Of these, Na^+–K^+ pump, which brings in 2 K^+ ions for 3 Na^+ ions out per ATP consumed, results in a net positive current (I_{NaK}) in outward direction and Na^+–Ca^{2+} exchanger, which exchanges 3 Na^+ ions for one Ca^{2+} ion, results in a net positive current (I_{NaCa}) contributing a little to the action potential. In most species, the exchanger current is in its Ca^{2+} influx mode (reverse mode) during depolarization resulting in outward current and is inward during Ca^{2+} efflux mode during repolarization. This is because the equilibrium potential of the current given by ($3E_{Na}$–$2E_{Ca}$) is around −40 mV (13). The current contributed by Ca^{2+} pump is negligible as it pumps few ions across membrane. The Na^+–H^+ exchanger transports one Na^+ for H^+ thereby causing no net flux across the membrane.

By convention, any current inward is considered negative and contributes to depolarization and any current outward is considered positive and contributes to repolarization. An inward current of a cation adds positive charge to the intracellular content, thereby making the transmembrane potential more positive; that is, depolarizes the membrane away from the resting potential.

Table 2. The Action Potential Phases and the Major Ionic Current Contributors

Phases of Action Potentials	Description	Major Contributing Ionic Currents
Phase 0	Initial depolarization	I_{Na}
Phase 1	Early repolarization	I_t
Phase 2	Plateau phase	I_{CaL}, I_{Kr}, I_{Ks}, I_{CaT}
Phase 3	Late repolarization	I_{K1}, I_{NaK}, I_{NaCa},I_{ss}, I_{Kslow}, I_{Kr}, I_{Ks},
Phase 4	Resting potential	I_{K1}

MATHEMATICAL MODELING OF CARDIAC CELLS

The Hodgkin–Huxley (HH) paradigm (14) type formalism is approached for numerical reconstruction of ventricular AP. At any particular moment, the movement of any particular ion across the membrane depends on the relative density of the channels, the probability of the channel that is selective to the ion being open, the conductance of the ion, and the net driving force of the ion given by the difference (V_m– E_{ion}), where V_m is the transmembrane voltage and E_{ion} is the Nernst potential of the ion (2). Also, it is assumed that ion fluxes are independent of each other, that is, the probability of an ion crossing the membrane does not depend on the probability of a different ion crossing the membrane. Based on this, the cell membrane is modeled as a capacitance in parallel to the resistances that represent the flow of ions through their respective channels along with their driving force. The resistive components are characterized in the original HH model as conductance's (g), the reciprocals of the resistances. The resistive currents can therefore be written

$$I_{ion} = g_{ion} * (V_m - E_{ion})$$

The experiments suggested that these conductances could be voltage and time dependent resulting in a gating mechanism. In HH type models, this gating mechanism is explained by considering the conductance g_{ion} as the product of maximum conductance $g_{ion\text{-}max}$ the channel can achieve and gating variables whose value lie between 0 and 1. The behavior of a gating variable x is given by first order differential equation:

$$dx/dt = (x_\infty - x)\tau_x$$

where x_∞ is the steady-state value the variable reaches at a particular voltage and τ_x is the time constant at that voltage that determines the rate at which steady state is reached. These variables are voltage dependent. All of these parameters are constrained by experimental data. These models represent the lumped behavior of the channels.

Increased understanding of the behavior of ion channels at a single channel level due to improved patch–clamp techniques lead to the development of state specific Markov models. Based on single channel recordings, it is observed that the channel opening or closing is random. Hence, the conductance $g_{\text{ion-max}}$ is multiplied by the total open channel probability of the ion (P_{ion}). These models represent the channel behavior based on their conformational changes and are capable of reproducing single channel behavior (15).

The rate of change of membrane potential to a stimulus current (I_{st}) is given by

$$(dV/dt) = (-1/C_{\text{m}}) * \left(\sum I_{\text{i}} + I_{\text{St}}\right)$$

where C_{m} is the membrane capacitance, and I_{i} values are different ionic currents.

THEORETICAL RESEARCH AND EXPERIMENTAL RESEARCH IN MURINE CARDIAC VENTRICULAR CELLS

After the first models of the mammalian ventricular cells by Beeler and Reuter (16) and Drouhard and Roberge (17), sophisticated mathematical models that simulate the cardiac action potentials in ventricular cells from different species such as canine (18,19), guinea pig (20–24), human (25,26), frog (27), and rabbit (28) have been published during the past decade. The model equations have usually based on the Hodgkin–Huxley (14) paradigm, wherein an ionic current is described by a set of nonlinear differential equations, and the parameters within these equations are constrained by experimental data obtained via voltage–clamp experiments in ventricular myocytes. There is a growing recognition that it is important to understand the complex, nonlinear interactions between the ionic milieu of the cardiac cell, that ultimately influence the action potential (29).

The mathematical models have demonstrated to be useful didactic tools in research, and have also quantified the important functional differences in the action potential properties between different species. Additionally, the computational models have also provided valuable, semi-quantitative insights into the diverse ionic mechanisms underlying the normal/abnormal action potential behavior in different animal models. It is not always possible to make precise experimental measurements regarding the contribution of a particular ionic mechanism to an aberrant action potential. The simulation results from these cardiac models have helped in planning for future experimental studies, and also in making predictions in cases where suitable technology is unavailable (or not developed) to make direct experimental measurements (e.g., visualizing the transmural activity within the ventricular wall). These models will play increasingly important roles in addition to experimental studies in the design and development of future drugs and devices (30). An additional and important feature of these ventricular models has been their ability to simulate intracellular Ca^{2+} transient ($[Ca^{2+}]_i$). Thus these models incorporate the feedback mechanism between the APD and the intracellular calcium $[Ca^{2+}]_i$. The APD is known to influence the amplitude of the $[Ca^{2+}]_i$ in ventricular cells (31), and $[Ca^{2+}]_i$ in turn influences the action potential waveform by Ca^{2+}-induced Ca^{2+} inactivation of I_{CaL}, and by determining the peak magnitude of I_{NaCa} (13).

The previously developed mathematical models of human, dog, guinea pig, rabbit and frog provide a good basis for the understanding of the ionic mechanisms responsible for the generation of the cardiac action potential. However, there are significant differences in the action potential waveforms and their corresponding properties between different species. The unique nature of the rat cardiac action potential, coupled with the recent available experimental data for the ionic mechanisms involved in the genesis of the action potential in isolated rat myocytes, provided the motivation for us to develop the first detailed mathematical model of the rat ventricular action potential. An adult male rat ventricular myocyte model was constructed (32) and utilized this model to study the ionic basis underlying the action potential heterogeneity in the adult rat left ventricle. Important insights into the role of long lasting Ca^{2+} current (I_{CaL}), the Ca^{2+}-independent transient outward K^+ current (I_{t}), and the steady-state outward K^+ current (I_{ss}) in determining the electrophysiological differences between epicardial and endocardial cells were obtained. This ventricular cell model has been used to investigate the ionic mechanisms that underlie altered electrophysiological characteristics associated with the short-term model of streptozotocin induced, type-I diabetic rats (33) and spontaneously hypertensive rats (34).Our rat ventrcicular myocyte model was further utilized to develop models for the mouse apex and septal left ventricular cells (35–37). Thus these model simulations reproduce a variety of experimental results, and provide quantitative insights into the functioning of ionic mechanisms underlying the regional heterogeneity in the adult rat and mouse ventricle.

The ventricular cell models of dog, guinea pig, human, and rabbit described in the previous section have been mainly used to simulate the so-called spike and dome configurations for action potentials (Fig. 2A) commonly observed in ventricular cells from larger mammalian species (38). However, no mathematical model has been published to represent the murine (rat or mouse) cardiac action potential (Fig. 2B) until our rat ventricular cell (12). The murine ventricular action potentials have a much shorter APD (typically the APD at 90% repolarization (APD_{90}) is < 100 ms), and lack a well-defined plateau phase (triangular in shape) (39–41). A comparison of the experimentally recorded ionic currents underlying action potentials in rat–mouse and other mammalian ventricular cells shows that they display markedly different amplitudes and time-dependent behavior. In fact, despite the similarity of action potential waveforms in rat and mouse, the underlying nature of the repolarizing K^+ currents are

different (41–43). Thus the unique action potential characteristics, and the lack of models to quantify these membrane properties provided the motivation to develop the rat and mouse ventricular cell models. The other motivation in this case was the widespread use of the murine cardiovascular system for the investigation of the cellular and molecular physiology of the compromised cardiovascular function (44).

Experimental studies indicate that the patterns of action potential waveforms are somewhat similar in rodents (rat or mouse), although the APD is shorter in mouse, and the complement of the K^+ currents underlying the cardiac repolarization in mouse are also different than those in rat (45,46). The cardiac repolarization in rat is controlled by two distinct depolarization activated K^+ currents, the Ca^{2+}-independent transient outward K^+ current (I_t) and the steady-state outward K^+ current (I_{ss}), (40,47). In mouse ventricular myocytes, an additional current, the 4-AP sensitive (at concentrations less than 100 μM), slowly inactivating, delayed rectifier K^+ current (I_{Kslow}) has been deemed to play an important role (41,48). The properties of the depolarization-activated K^+ currents have now been well characterized in rats (40,49) and mouse (43,48), and appear to be significantly different. It is therefore interesting to investigate in computational modeling whether the reported differences in the properties of the depolarization-activated K^+ currents can account for the dissimilar nature of the action potential configurations observed in rats and mice.

COMPUTATIONAL MODELING OF THE MURINE VENTRICULAR ACTION POTENTIALS

The goal of my computational modeling laboratory has been to unify different experimental data and to develop biophysically detailed models for the rat and mouse ventricular cells and to determine the underlying ionic channels responsible for differences in cardiac action potential variations in rats and mice under normal and diseased conditions. A computational model has been developed for the rat cardiac ventricular cell based on electrophysiology data. Our control model (12) represents the bioelectric activity in the left ventricular cells in adult male rats. The differences in the membrane properties within the left ventricle to simulate the action potential variations of the endocardial and epicardial cells have been formulated. Also, a right ventricular cell model from our control model was built (the left ventricular cell model) to investigate ionic mechanisms in diabetic rats (33). Our right ventricular cell model was also the template for us to develop a mouse ventricular cell model by utilizing experimental data (8,32).

The left (LV) and right (RV) ventricular cell models for the rat consist of a Hodgkin–Huxley type membrane model that is described by the membrane capacitance, various ionic channels; the fast Na^+ current (I_{Na}), long-lasting Ca^{2+} current (I_{CaL}), the 4AP sensitive, Ca^{2+} independent transient outward K^+ current (I_t), steady-state outward K^+ current (I_{ss}), inward rectifier K^+ current (I_{K1}), hyperpolarization activated current (I_f), linear background current (I_B); the Na^+/Ca^{2+} ion exchanger (I_{NaCa}), and the Na^+/K^+ (I_{NaK}) and Ca^{2+} membrane (I_{CaP}) pumps, that are experimentally observed in rat ventricular cells.

The mouse ventricular cell model was constructed by using the rat right ventricular cell model as the template. The mouse LV apex cell was developed by adding the 4AP sensitive slowly inactivating, delayed rectifier K^+ current (I_{Kslow}) based on the data of Fiset et al. (41) and Zhou et al. (48), and by reformulating I_t and I_{ss} based on experiments performed by Agus et al. (35–37) and Xu et al. (43) in mice. Further, a mouse LV septum cell model was developed by formulating a new current I_{tos} based on the data of (Xu et al. (43)), and by reducing the densities of I_{tof}, I_{Kslow}, and I_{ss} by 70, 23, and 30%, respectively, based on data of Gussak et al. (45).

The important results of our simulation studies are

1. The action potential heterogeneity (Fig. 3) in the adult rat LV is mainly due to the changes in the density and recovery kinetics of I_t and due to the altered density of I_{Na} (12).
2. The RV cell model can be developed from the LV cell model by changing the densities of I_t, I_{ss}, I_{CaL}, and I_{NaK} based on experimental data.
3. The changes in the density and the reactivation kinetics of I_t can account for the action potential prolongation differences in RV myocytes of diabetic (type-I, short term) rats (33) and LV myocytes of spontaneously hypertensive rats (35) (Fig. 4).
4. The presence of I_{Kslow} in mouse is one of the main factors contributing to the faster rate of repolarization seen in mouse compared to rats (Fig. 5) (8).
5. The LV septum cell model had more prolonged action potentials than the apex cells and these simulation results (Fig. 6a) are qualitatively similar to the experimental data of (52) (Fig. 6b).

Simulated Action Potentials of the Epicardial and Endocardial Cardiac Ventricular Cells

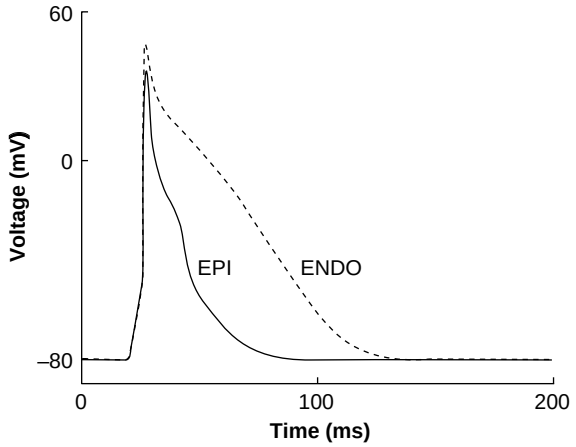


Figure 3. Simulated action potentials of the rat left ventricular (LV) epicardial (EPI) (solid line) and endocardial (ENDO) (dashed line) cells (35–37).

Simulated Action Potentials of the Epicardial Cardiac Ventricular Cells for Normal and Spontaneously Hypertensive Rats

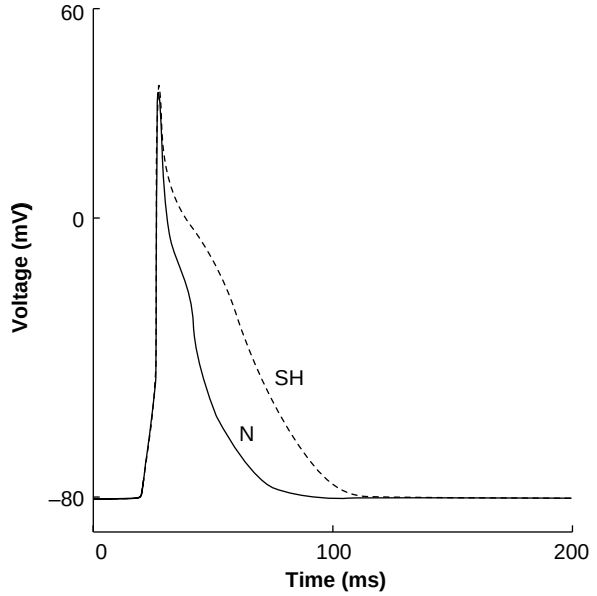


Figure 4. Model generated ventricular action potentials of the epicardial cells for the normal rat (N) (solid line) and spontaneously hypertensive (SH) rat (dashed line) (35–37).

6. The rat epicardial and endocardial ventricular cell models were more rate-sensitive than the mouse ventricular cell model and these simulation data match the experimental data well.

In conclusion, the mathematical modeling study of murine ventricular myocytes complements our knowledge of the biophysical data with simulation data and provide us with quantitative descriptions to understand the ionic currents underlying the cardiac action potential variations in different species. This kind of computational work will enhance our understanding of the ionic mechanisms that contribute to the cardiac action potential variation in normal and diseased animals, and will provide us with better treatments for diseases in humans.

Simulated Cardiac Action Potentials of Rat and Mouse Ventricular Cells

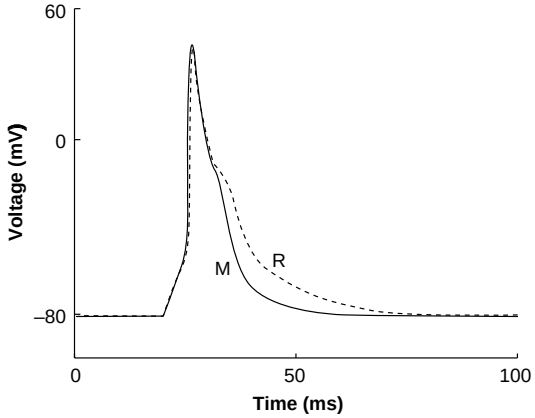


Figure 5. Simulated action potentials of the mouse left ventricular apex cell (M) (solid line) and the rat right ventricular cell (R) (dashed line) (35–37).

Simulated Ventricular Action Potentials for Mouse LV Apex and Septum Cells

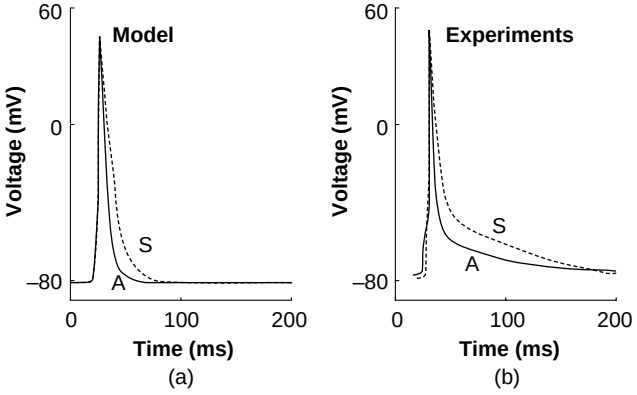


Figure 6. Simulated and experimentally recorded action potentials of the mouse left ventricular apex (A) (solid line) and the septum (S) (dashed line) cells (35–37).

DISSEMINATION OF COMPUTATIONAL MODELS

My computational modeling laboratory has developed an interactive cell modeling web site, iCell (http://ssd1.bme.memphis.edu/icell/) since 1998. iCell, that integrates research and education, was specifically developed as a simulation-based teaching and learning resource for electrophysiology (35–37,51,52). The main objectives for the development of iCell were (*1*) to use technology in electrophysiology education, (*2*) to provide a learning and teaching resource via internet for cellular electrophysiology, (*3*) to allow the user to understand the cellular physiology mechanisms, membrane transport and variations of action potentials and ion channels by running simulations, and (*4*) to provide a computer platform independent resource for cellular models to be used for teaching, learning and collaboration. The site consists of JAVA applets representing models of various cardiac cells and neurons, and provides simulation data of their bioelectric transport activities at cellular level. Each JAVA-based model allows the user to go through menu options to change model parameters, run and view simulation results. The site also has a glossary section for the scientific terms. iCell has been used as a teaching and learning tool for seven graduate courses at the Joint Biomedical Engineering Program of University of Memphis and University of Tennessee. This modeling tool was also used as a collaboration site among our colleagues interested in simulations of cell membrane activities. Scientists from the fields of biosciences, engineering, life sciences and medical sciences in 17 countries, Argentina, Belgium, Brazil, Canada, China, England, Germany, Greece, Ireland, Japan, Korea, the Netherlands, New Zealand, Spain, Taiwan, Turkey and the United States, have tested and utilized iCell as a

simulation-based teaching, learning and collaboration environment. The platform-independent software, iCell, provides us with an interactive and user-friendly teaching and learning resource, and also a collaboration environment for electrophysiology to be shared over the Internet. The usage of simulations for teaching and learning will continue advancing simulation-based engineering and sciences for research and development.

The simulations provided by iCell and other resources, such as CellML (http://www.cellml.org/public/news/index.html), Virtual Cell (http://www.nrcam.uchc.edu/), JSIM (http://nsr.bioeng.washington.edu/PLN/Software/), simBio Cell/Biodynamics Simulation Project of Kyoto University (http://www.biosim.med.kyoto-u.ac.jp/e/index.html), and E-Cell (http://ecell.sourceforge.net/), will continue signifying the important verification and prediction capabilities of the computer models to represent, analyze and complement the physiological data and knowledge. The model development demonstrates that computational models have to be constructed from experimental electrophysiology data, not only to explain and to verify the data that they are based on, but also to predict results for experiments that have not been performed and to guide future experiments.

Overall, computational modeling and simulation results continue to advance our understanding of living systems at the cellular level in cardiac electrophysiology while promoting collaborations and training in interdisciplinary field of bioengineering between scientists in life scientists and engineers. The presentation of computational models in user friendly, interactive and menu driven software are important in bringing collaborators of different disciplines and training scientists and students in cross-disciplinary projects.

IMPACT OF COMPUTATIONAL MODELING IN VENTRICULAR CELLS

The following summarizes impacts of the computational model development of ventricular bioelectric activity and the model-generated data in different disciplines of life sciences. I. *Biophysics and Physiology*: The results of the computational studies expand our knowledge of the living systems at the cellular level in electrophysiology. II. *Clinical Physiology and Medicine*: The insights gained and conclusions derived from the computational studies enhance our understanding of the biocomplexity of the heart, and provide us with better knowledge to be used in the future in treatments for diseases in humans. The cardiac cells' responses to various pathophysiological states with simulation data will also be better understood. III. *Pharmacology*: The differences in ventricular membrane ionic currents, especially outward K^+ currents in different species have very important practical implications. Different drugs are known to affect different ionic currents and to change action potential waveforms in different mammalian heart preparations under various conditions of development, aging and gender. A better understanding of the role of the ionic currents that control repolarization in the ventricular myocytes obtained from various species including rat and mouse, as presented in this paper, will provide motivation and explanations for species differences in treatment and drug actions, and also promote pharmacological research that may lead to the development of more specific drugs to be used in children and adults.

ACKNOWLEDGMENTS

These computational research projects presented here were funded by the Whitaker Foundation (PI: Dr. S. S. Demir). The author acknowledges the contributions of her former students S. Pandit, S. Padmala, and E. Damaraju to these research projects.

The author thanks her former students, Joe E. McManis, Yiming Liu, Dong Zhang, Srikanth Padmala and Eswar Damaraju for coding JAVA applets in the iCell project, her students Chris Oehmen and Jing Zheng for posting the html pages and Dr. Emre Velipasaoglu, Siddika Demir and Asim Demir for valuable collaborations and discussions. This research was also funded by the Whitaker Foundation (PI, Dr. S. S. Demir).

BIBLIOGRAPHY

Cited References

1. Alberts B, et al. Membrane Transport. Essential Cell Biology: An Introduction to the Molecular Biology of the Cell. New York: Garland Publishing; 1997 p 371–407. Chapt. 12
2. Plonsey R, Barr R. Bioelectricity: A Quantitaive Approach. New York: Kluwer Academic Publications; 2000.
3. Hamill O, et al. Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches. Pflugers Archiv 1981;391:85–100.
4. Neher E, Sakmann B. Noise analysis of drug induced voltage clamp currents in denervated frog muscle fibres. J Physiol 1976;258(3):705–729.
5. Neher E, Sakmann B. Single-channel currents recorded from membrane of denervated frog muscle fibres. Nature (London) 1976;260(5554):799–802.
6. Neher E, Sakmann B, Steinbach JH. The extracellular patch clamp: a method for resolving currents through individual open channels in biological membranes. Pflugers Arch 1978;375(2):219–228..
7. Neher E, Sakmann B. The patch clamp technique. Sci Am 1992;266(3):44–51.
8. Damaraju E. A Computational Model of Action Potential Heterogeneity in Adult Mouse Left Ventricular Myocytes. M.S. dissertation, University of Memphis, 2003.
9. Fozzard H. Cardiac Electrogenesis and the Sodium Channel. In: Spooner P, Brown A, Catterall W, Kaczorowski G, Strauss H, editors. Ion Channels in the Cardiovascular system:Function and dysfunction. Futura Publishing Company, Inc.; 1994. Chapt. 5.
10. Spooner PM, Rosen MR, editors. Foundations of Cardiac Arrhythmias. 1st ed. New York: Marcel Dekker; 2000.
11. Demir SS, et al. Action Potential Variation in Canine Ventricle: A Modeling Study. Comput Cardiol 1996; 221–224.
12. Pandit SV, Clark RB, Giles WR, Demir SS. A Mathematical Model of Action Potential Heterogeneity in Adult Rat Left Ventricular Myocytes. Biophy J 2001;81:3029–3051.
13. Bers DM. Excitation-Contraction Coupling and Cardiac Contractile Force 2nd ed. The Netherlands: Kluwer Academic Publications; 2001.

14. Hodgkin L, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol 1952;117:500–544.
15. Irvine L, Jafri M, Winslow R. Cardiac sodium channel Markov model with temperature dependence and recovery from inactivation. Biophysic J 1999;76(4):1868–1885.
16. Beeler GW, Reuter H. Reconstruction of the action potential of ventricular myocardial fibres. J Physiol 1997;268: 177–210.
17. Drouhard J, Roberge FA, Revised formulation of the Hodgkin-Huxley representation of the sodium current in cardiac cells. Comput Biomed Res 1987;20:333–350.
18. Winslow RL, et al. Mechanisms of altered excitation-contraction coupling in canine tachycardia-induced heart failure, II: model studies. Circ Res 1999;84:571–586.
19. Fox JJ, McHarg JL, Gilmour RF. Ionic mechanism of electrical alternans. Am J Physiol Heart Circ Physiol 2002;282:H516–H530.
20. Nordin C, Computer model of membrane current and intracellular Ca^{2+} flux in the isolated guinea pig ventricular myocyte. Amer J Physiolo 1993;265:H2117–H2136.
21. Luo C-H, Rudy Y. A model of the ventricular cardiac action potential. Circulation Res 1991;68:1501–1526.
22. Luo C-H, Rudy Y. A dynamic model of the cardiac ventricular action potential. I Simluation of ionic currents and concentration changes. Circulation Res 1994;74:1071–1096.
23. Zeng J, Laurita K, Rosenbaum DS, Rudy Y. Two components of the delayed rectifier K^+ current in ventricular myocytes of the guinea pig type: theoretical formulation and their role in repolarization. Circulation Res 1995;77:140–152.
24. Noble D, Varghese A, Kohl P, Noble P. Improved guinea-pig ventricular cell model incorporating a diadic space, I_{Kr} and I_{Ks}, and length- and tension-dependent processess. Can J Cardiol 1998;14:123–134.
25. Priebe L, Beuckelmann D. Simulation study of cellular electric properties in heart failure. Circ Res 1998;82: 1206–1223.
26. ten Tusscher KHWJ, Noble D, Noble PJ, Panfilov AV. A model for human ventricular tissue.Am J Physiol Heart Circ Physiol 2004;286:H1573–H1589.
27. Riemer TL, Sobie A, Tung L. Stretch-induced changes in arrhythmogenesis and excitability in experimentally based heart cell models. Am J Physiol 1998;275:H431–H442.
28. Puglisi JL, Bers DM. LabHEART: an interactive computer model of rabbit ventricular myocyte ion channels and Ca transport. Am J Physiol Cell Physiol 2001;281:C2049–C2060.
29. Winslow RL, et al. Electrophysiological modeling of cardiac ventricular function: from cell to organ. Annu Rev Biomed Eng 2000;2:119–155.
30. Members of the Sicilian Gambit. New approaches to antiarrhythmic therapy, Part I: Emerging therapeutic applications of the cell biology of cardiac arrhythmias. Circulation 2001;104:2865–2873.
31. Bouchard RA, Clark RB, Giles WR. Effects of action potential duration on excitation-contraction coupling in rat ventricular myocytes. Action potential voltage-clamp measurements. Circ Res 1995;76:790–801.
32. Pandit SV. Electrical Activity in Murine Ventricular Myocytes: Simulation Studies. Ph.D. dissertation, University of Memphis; 2002.
33. Pandit SV, Giles WR, Demir SS. A Mathematical Model of the Electrophysiological Alterations in Rat Ventricular Myocytes in Type-I Diabetes. Biophys J 2003;84(2):832–841.
34. Padmala S, Demir SS. A computational model of the ventricular action potential in adult spontaneously hypertensive rats. J Cardiovasc Electrophysiol 2003;14:990–995.
35. Demir SS. Computational Modeling of Cardiac Ventricular Action Potentials in Rat and Mouse. Rev Jne J Physiol 2004;54(6):523–530.
36. Demir SS. An Interactive Electrophysiology Training Resource for Simulation-Based Teaching and Learning. Institute of Electrical and Electronics Engineers (IEEE) Engineering in Medicine & Biology Society Proceedings; 2004. p 5169–5171.
37. Demir SS. The Significance of Computational Modelling in Murine Cardiac Ventricular Cells. J Appl Bionics Biomech 2004;1(2):107–114.
38. Antzelevitch C, Yan G-X, Shimuzu W, Burashnikov A. Electrical Heterogeneity, the ECG, and Cardiac Arrhythmias. In: Zipes DP, Jalife J, editors. Cardiac Electrophysiology: From Cell to Bedside. 3rd ed. Philadelphia: WB Saunders; 1999. p 222–238.
39. Watanabe T, Delbridge LM, Bustamante JO, McDonald TF. Heterogeneity of the action potential in isolated rat ventricular myocytes and tissue. Circ Res 1983;52:280–290.
40. Clark RB, et al. Heterogeneity of action potential waveforms and potassium currents in rat ventricle. Cardiovas Res 1993;27:1795–1799.
41. Fiset C, Clark RB, Larsen TS, Giles WR. A rapidly activating sustained K^+ current modulates repolarization and excitation-contraction coupling in adult mouse ventricle. J Physiol 1997;504:557–563.
42. Nerbonne JM, Nichols CG, Schwarz TL, Escande D. Genetic manipulation of cardiac K^+ channel function in mice: what have we learned, and where do we go from here? Circ Res 2001;89:944–956.
43. Xu H, Guo W, Nerbonne JM. Four kinetically distinct depolarization-activated K^+ currents in adult mouse ventricular myocytes. J Gen Physiol 1999;113:661–678.
44. Chien KR. To Cre or not to Cre: the next generation of mouse models of human cardiac diseases. Circ Res 2001;88:546–549.
45. Gussak I, Chaitman BR, Kopecky SL, Nerbonne JM. Rapid ventricular repolarization in rodents: electrocardiographic manifestations, molecular mechanisms, and clinical insights. J Electrocardiol 2000;33:159–170.
46. Nerbonne JM. Molecular analysis of voltage-gated K^+ channel diversity and functioning in the mammalian heart. In: Page E, Fozzard HA, Solaro RJ, editors. Handbook of Physiology: The Cardiovascular System. New York: Oxford University Press; 2001. 568–594.
47. Shimoni Y, Severson D, Giles WR. Thyroid status and diabetes modulate regional differences in potassium currents in rat ventricle. J Physiol 1995;488:673–688.
48. Zhou J, et al. Characterization of a slowly inactivating outward current in adult mouse ventricular myocytes. Circ Res 1998;83:806–814.
49. Shimoni Y, Light PE, French RJ. Altered ATP sensitivity of ATP-dependent K^+ channels in diabetic rat hearts. Am J Physiol 1998;275:E568–E576.
50. Demir SS. iCell: an Interactive Web Resource for Simulation-Based Teaching and Learning in Electrophysiology Training. Institute of Electrical and Electronics Engineers (IEEE), Engineering in Medicine & Biology Society Proceedings; 2003; p 3501–3504.
51. Demir SS. Simulation-based Training in Electrophysiology by iCell Institute of Electrical and Electronics Engineers, Engineering in Medicine and Biology Engineering in Medicine & Biology Society Proceedings, 4 pages; 2005.

See also BLADDER DYSFUNCTION, NEUROSTIMULATION OF; ELECTROCONVULSIVE THERAPY; PACEMAKERS; SPINAL CORD STIMULATION; TRANSCUTANEOUS ELECTRICAL NERVE STIMULATION (TENS).

ELECTRORETINOGRAPHY

GRAHAM E. HOLDER
Moorfields Eye Hospital
London, United Kingdom

INTRODUCTION

The retina, situated at the back of the eye, is highly complex, consisting of different layers and containing many different cell types. It serves to encode images of the outside world into a suitable form for transmission to the brain for interpretation and the process of "seeing". It is possible to view the retina *in situ* using ophthalmoscopic techniques, and although this may reveal anatomical abnormalities, it may not reveal either the extent or nature of retinal dysfunction. The principal challenge for electroretinography is to provide information regarding retinal function to facilitate patient care.

In essence, a controlled light stimulus is used to stimulate the retina, which responds by generating very small electrical signals that can be recorded, with suitable amplification, using electrodes situated in relation to the eye, usually contacting the cornea. These electrical signals, the electroretinogram (ERG), have defined parameters (timing, shape, size) in normal individuals, and are altered in a predictable manner in disease. In general, the brighter the stimulus, the higher is the amplitude and the shorter the peak time of the ERG. Modification of the adaptive state of the eye (dark adapted or scotopic; light adapted or photopic) facilitate the separation of different cell types and layers within the retina. The objective information provided by electrophysiological examination has a significant effect both on diagnosis and patient management (1).

TECHNIQUES

The main tests of retinal function are the ERG, the massed retinal responses to full-field luminance stimulation, which reflects the function of the photoreceptor and inner nuclear layers of the retina, and the pattern electroretinogram (PERG), which, in addition to being "driven" by the macular photoreceptors, largely arises in relation to retinal ganglion cell function. Knowledge of this latter response can also be particularly useful in improved interpretation of an abnormal cortical visual evoked potential (VEP), but that topic is beyond the remit of this contribution, and the reader is referred elsewhere for a full discussion of the interrelationships between PERG and ERG, and PERG and VEP (2). Brief reference will also be made to the electrooculogram (EOG), which examines the function of the retinal pigment epithelium (RPE) and the interaction between the RPE and the (rod) photoreceptors, and is often used in conjunction with the ERG.

Electrophysiological recordings are affected not only by stimulus and recording parameters, but also by the adaptive state of the eye, and standardization is mandatory for meaningful scientific and clinical communication between laboratories. The International Society for Clinical Electrophysiology of Vision (ISCEV) has published Standards for EOG (3), ERG (4), PERG (5), and the VEP (6). Readers are strongly encouraged not only to adhere to the recommendations of those documents, but also to consider that the Standards are intended as minimum data sets, and that recording protocols in excess of the Standards may be necessary to accurately establish the diagnosis in some disorders. Typical normal traces appear in Fig. 1.

A brief description of each test follows, with emphasis on response generation. Referencing has been restricted; the reader is referred to standard texts for further details (7,8). The multifocal ERG (mfERG) is a relatively recent addition to the diagnostic armamentarium, and although currently more of a research application than a mainstream clinical tool, this is likely to change in the future as more knowledge is gained of the clinical applications and underlying mechanisms. The ISCEV has published guidelines for mfERG, to which the reader is referred (9).

THE ELECTROOCULOGRAM

The EOG enables assessment of the function of the RPE, and the interaction between the RPE and the retinal photoreceptors. The patient makes fixed 30° lateral eye movements during a period of 20 min progressive dark adaptation, followed by a 12–15 min period of progressive light adaptation. The eye movements are made every 1–2 s for ~10 s each minute. The amplitude of the signal recorded between electrodes positioned at medial and lateral canthi reaches a minimum during dark adaptation, known as the dark trough, and a maximum during light adaptation, the light peak. The development of a normal light peak requires normally functioning photoreceptors in contact with a normally functioning RPE, and reflects progressive depolarization of the basal membrane of the RPE. The EOG is quantified by calculating the size of the light peak in relation to the dark trough as a percentage, the Arden index. A normal EOG light rise is >175% for most laboratories.

THE ELECTRORETINOGRAM

The functional properties of the retinal photoreceptors underpin the principles of ERG recording. The retinal rod system, with ~120,000,000 rod photoreceptors, is sensitive under dim lighting conditions, has coarse spatial and poor temporal resolution. The rods adapt slowly to changes in lighting conditions. They do not enable color vision. They have a peak spectral sensitivity in the region of 500 nm. There are three types of retinal cone. (1) Short wavelength (S-cone), (2) medium (M-cone), and (3) long wavelength (L-cone). In the past, they have been referred to as blue, green, and red, respectively. There are perhaps 7,000,000 M- and L-cones and 800,000 S-cones. The relative proportion of L- versus M-cones varies from individual to individual, but approximates to 50% over a population. They are sensitive under bright lighting conditions; their high spatial resolution enables fine visual acuity; they adapt rapidly to changes in lighting conditions and can follow a fast flicker (L- and M-cones). The overall maximum spectral sensitivity is ~550 nm, in the green-yellow

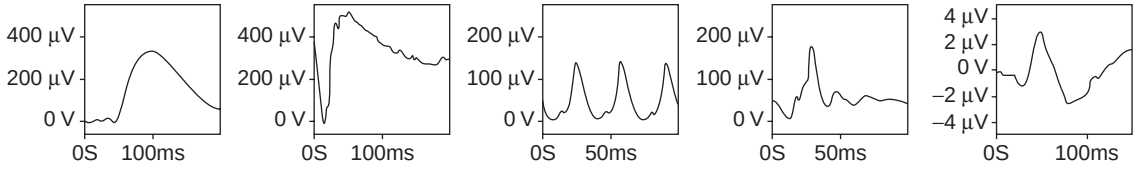


Figure 1. Typical normal ERG recordings. The rod specific ERG consists of the inner-nuclear layer generated b-wave. With a bright flash the waveform now contains an additional a-wave, the first 10–12 ms of which arise in relation to photoreceptor hyperpolarization. The rod specific and bright flash responses are recorded with full scotopic adaptation. After restoration to photopic adaptation the cone flicker and single flash ERGs are recorded. The former consists of a sinusoidal type waveform, the latter containing clear a- and b-waves. The pattern ERG (PERG) is the response of the macula to a reversing black and white checkerboard. See text for further details.

region of the color spectrum. Normal color vision requires all three cone types, but providing at least two cone types are present (there are some disorders in which that is not the case), at least some color vision is enabled. There are no S-cones in the foveola, the very central part of the macula responsible for very fine acuity.

The ERG is recorded using corneal electrodes and is the mass electrical response of the retina using a brief flash of light as a stimulus. The stimuli are delivered using a Ganzfeld bowl, an integrating sphere that enables uniform whole field illumination (Fig. 2a, b). In addition to flash stimulation, the Ganzfeld also allows a diffuse background for photopic adaptation. Some corneal electrodes are bipolar contact lenses with a built-in reference electrode. If such an electrode is not used, the reference electrodes should be sited at the ipsilateral outer canthi. A standard flash is defined by ISCEV as 1.5–3.0 $cd{\cdot}s{\cdot}m^{-2}$. The response to this flash under scotopic conditions, with a fully dilated pupil, is the Standard or mixed response (Fig. 1). It is probably this response that may be regarded as the "typical" ERG, but although there is a cone contribution, the standard response is dominated by rod driven activity. The "maximal" ERGs that appear in this article were recorded to $\sim$11.0 $cd{\cdot}s{\cdot}m^{-2}$ flash better to view the a-wave. The use of a brighter flash of such intensity is "suggested" in the most recent ISCEV ERG Standard (4). The initial $\sim$10 ms of the a-wave arises in relation to hyperpolarisation of the (rod) photoreceptors and the slope of the a-wave can be related to the kinetics of phototransduction (10). The larger positive b-wave is generated postreceptorally in the inner-nuclear layer of the retina in relation to depolarization of the ON-bipolar cells (11). The oscillatory potentials, the small wavelets on the ascending limb of the b-wave, are probably generated in relation to amacrine cell activity. When the standard flash is attenuated by 2.5 log units, the stimulus intensity falls below the cone threshold, and a rod-specific b-wave is obtained. At this relatively low luminance there is insufficient photoactivation to record an a-wave (Fig. 1, column A, top).

The ERGs that reflect cone system activity are obtained using a rod-saturating photopic background (17–34 $cd{\cdot}m^{-2}$) using superimposed single flash and 30 Hz flicker stimulation. The rod system has low temporal resolution and use of a 30 Hz stimulus, combined with a rod-suppressing background, allows a cone-system specific waveform to be recorded. This response is probably the more sensitive measure of cone dysfunction, but is generated at an inner-retinal level (12) and thus does not allow the distinction between cone photoreceptor and cone inner-nuclear layer dysfunction. Although there is a demonstrated contribution from hyperpolarizing (OFF-) bipolar cells to shaping the photopic a-wave (13), this component nonetheless has some contribution from cone photoreceptor function, and some localization within the retina may be obtained with the single flash cone response. The cone b-wave reflects postphototransduction activity, and to a

Figure 2. (a) A conventional Ganzfeld used for ERG recording (front view). (b) The subject in position at the Ganzfeld. (c) Photograph taken using an infrared (IR) camera at the back of the Ganzfeld, which is used to monitor eye position and eye opening during both dark adapted and light adapted conditions. The two gold-foil corneal recording electrodes are well seen. The central forehead ground electrode is easily seen; the outer canthus reference electrodes are just visible. (Courtesy of Chris Hogg.)

Figure 3. A "mini-Ganzfeld" based on light emitting diode technology. The device shown has four independent color channels, blue, green, orange and red, each of which can be used as stimulus or background alone or in combination. (Courtesy of Chris Hogg, CH electronics, Bromley, Kent, UK; www.ch-electronics.net.)

short flash stimulus ON and OFF activity within the photopic system is effectively synchronized.

Separation of the cone ON (depolarizing bipolar cells, DBCs) and OFF (hyperpolarizing bipolar cells, HBCs) responses can be achieved using a long duration stimulus with a photopic background (14,15). The stimulus can be generated either via a shutter system or by using light emitting diodes (Fig. 3). Stimulators based on light emitting diodes (LEDs) offer several advantages over standard stimulators. They are of low cost, have a stable output intensity over time (reducing the need for calibration), enable variable and highly accurate stimulus duration, and a have a well-defined narrow band spectral output. Further, being driven by relatively low voltage and current, they are intrinsically safe, and generate low electrical noise. Their use in ERG systems can be expected to increase.

It is also possible to elicit the activity of the S-cone population. In the author's laboratories this is achieved using blue stimuli superimposed upon a bright orange photopic background, again delivered using a LED based device. The background thus serves to suppress activity from rod and L-/M-cone systems. The response under appropriate recording conditions consists of an early component at ~30 ms arising in relation to L-/M-cone systems (there is overlap of the spectral sensitivities of the different cone systems and a small response arises from L-/M-cones with a bright blue stimulus), followed by a component specific for S-cone function at 45–50 ms (16).

The retinal ganglion cells do not significantly contribute to the clinical (flash) ERG. Also, as a mass response, the ERG is normal when dysfunction is confined to small retinal areas, and, despite the high photoreceptor density, this also applies to macular dysfunction; the full-field ERG is normal if dysfunction is confined to the macula (e.g., Fig. 4, column B).

THE PATTERN ELECTRORETINOGRAM

The response of central retina to a structured isoluminant stimulus can be measured, and is known as the pattern ERG. The stimulus is usually a reversing black and white checkerboard. The PERG has largely inner retinal origins, but is "driven" by the macular photoreceptors, and PERG measurement thus provides both a measure of central retinal function and, in relation to its origins, of retinal ganglion cell function. It is thus of clinical importance not only in the objective assessment of macular function, but also in the electrophysiological differentiation between optic nerve and macular dysfunction by providing a measure of the retinal response to a similar stimulus to that used to evoke the VEP (see Ref. 2 for a comprehensive review). It is a much smaller signal than the (full-field) ERG and computerized signal averaging is used to extract the PERG signal.

The PERG is recorded using noncontact lens electrodes in contact with the cornea or bulbar conjunctiva to preserve the optics of the eye. Suitable electrodes are the gold foil (17), the DTL (18), and the H–K loop (19). Ipsilateral outercanthus reference electrodes are essential to avoid contamination from the cortically generated VEP, such as occurs if forehead or ear "reference" electrodes are used (20). Pupillary dilation is not used.

There are two main components of PERG to a reversing checkerboard with a relatively slow reversal rate (<6 reversals s^{-1}). There is a prominent positive component, P50, at ~50 ms followed by a larger negative component, N95, at ~95 ms (21). Clinical measurement of the PERG usually comprises the amplitude of P50, measured from the trough of the early negative N35 component; the peak latency of P50; and the amplitude of N95, measured to trough from the peak of P50 (Fig. 1). Approximately 70% of P50 is likely to be related to retinal ganglion cell function, but the remainder is not related to spiking cell function and may be generated more distally in the retina (22). The exact origins have yet to be ascertained at the time of writing. The N95 is a contrast-related component generated in the retinal ganglion cells.

An analysis time of 150 ms or greater is usually used for recording the PERG, with ~150 averages per trial needed to obtain a reasonable signal-to-noise ratio. As it is a small response, stringent technical controls are important during recording and are fully discussed elsewhere (8). Binocular stimulation and recording is preferred so the better eye can maintain fixation and accommodation, but it is necessary to use monocular recording if there is a history of squint. P50 is sensitive to optical blur, and accurate refraction is needed. At low stimulus frequencies the amplitude of the PERG is related almost linearly to stimulus contrast. A high contrast black and white reversing checkerboard with 0.8° checks in a 10–16° field is recommended by ISCEV.

MULTIFOCAL ERG

The mfERG attempts to provide spatial information regarding cone system function in central retina. The stimulus usually consists of multiple hexagons displayed on a screen (Fig. 5a) each of which flashes on with its own pseudo-random binary sequence (an M-sequence). A cross-correlation of the local flash sequence with the mass

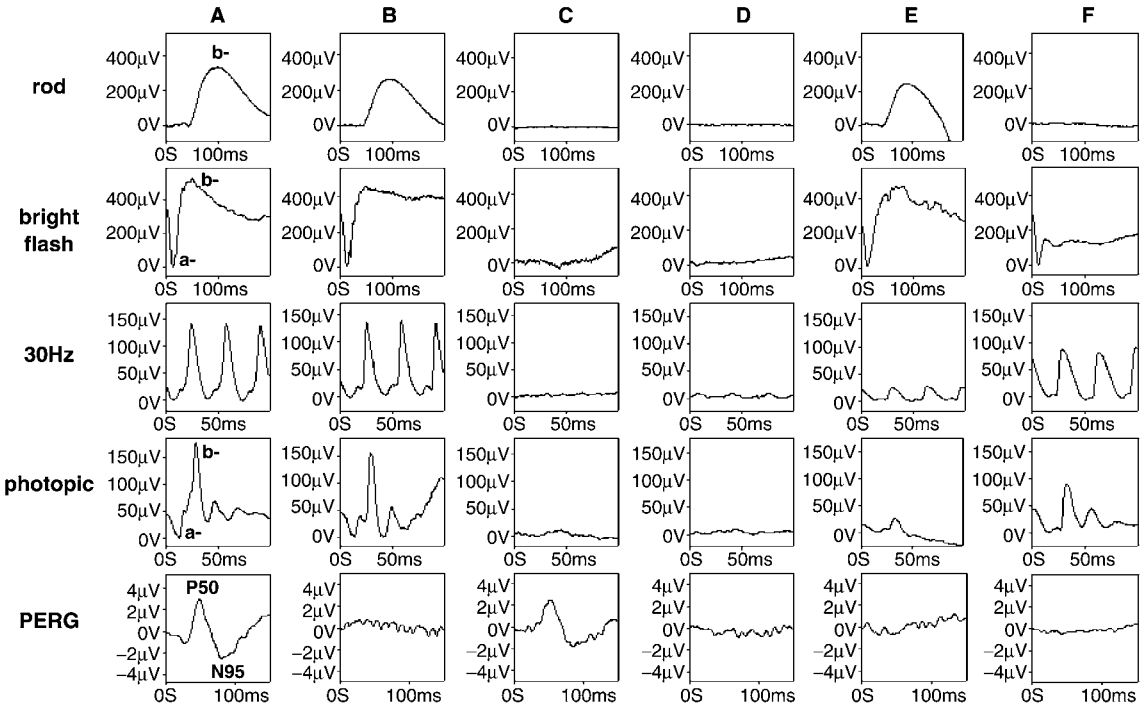


Figure 4. Typical electroretinographic abnormalities in selected diseases compared to those in a normal subject. Column A: Normal subject. Column B: A patient with macular dysfunction; the PERG is undetectable, but full-field ERGs are normal. Column C: "Classical" retinitis pigmentosa; all full-field ERGs are virtually extinguished, but the PERG is normal reflecting sparing of central retinal function. Column D: Rod-cone dystrophy (retinitis pigmentosa); the rod and cone ERGs are markedly abnormal (with the rod ERG being more affected). Note the delayed and reduced cone ERGs, typical abnormalities present when there is generalized cone involvement in the context of photoreceptor degeneration. The abnormal PERG reflects involvement of the macula. Column E: Cone dystrophy; the rod and bright flash ERGs are normal, but the cone single flash and flicker ERGs are delayed and reduced in keeping with generalized cone system dysfunction. The abnormal PERG reflects involvement of the macula. Column F: X-linked congenital stationary night blindness (complete type). The rod specific ERG is undetectable, but the normal a-wave of the bright flash dark adapted ERG confirms the dysfunction to be postphototransduction. There are subtle but significant changes in cone-system derived ERGs (note particularly the broadened trough and reduced amplitude sharply rising peak of the b-wave), and reduction in the PERG.

response derives the responses relating to each individual hexagon thus giving multiple cone system ERG waveforms from a single recording electrode. The mfERG can be of use in disturbances of macular function and to assess the degree of central retinal involvement in generalized retinal disease, but is highly susceptible to poor fixation, and the ability of a patient accurately to maintain good fixation throughout the recording session is a pre-requisite to obtaining clinically meaningful data. Increasing use and development of systems that can control stimulus delivery in relation to eye position can be anticipated. Possibilities include the use of "eye-tracking" devices and direct fundus visualization during stimulation.

CLINICAL APPLICATIONS

EOG

Disorders of rod photoreceptor function can affect the EOG, and the light rise is typically reduced in generalized photoreceptor degenerations such as retinitis pigmentosa (RP, rod–cone dystrophy), a genetically determined group of disorders. Usually, the reduction in EOG light rise parallels the degree of rod photoreceptor dysfunction, but generalized RPE dysfunction can also manifest a reduced EOG light rise. Indeed, it is the latter property that leads to the main clinical use of the EOG, the diagnosis of Best disease. Best disease, or vitelliform macular dystrophy, is a dominantly inherited macular degeneration related to mutation in the gene VMD2. At presentation there are often distinctive vitelliform lesions at the maculae on funduscopy, but other appearances may occur. The diagnostic findings are of a severely reduced or absent EOG light rise accompanied by normal ERGs. Best disease may present in childhood, but a child may find the repetitive eye movements needed for EOG recording difficult or impossible to maintain for the required 30–40 min. Under such circumstances it is appropriate to test both parents; due to the dominant inheritance pattern one of the disorder, one of the parents will have carry the mutant gene and will manifest a reduced EOG. Adult vitelliform macular dystrophy (pattern dystrophy) may sometimes clinically

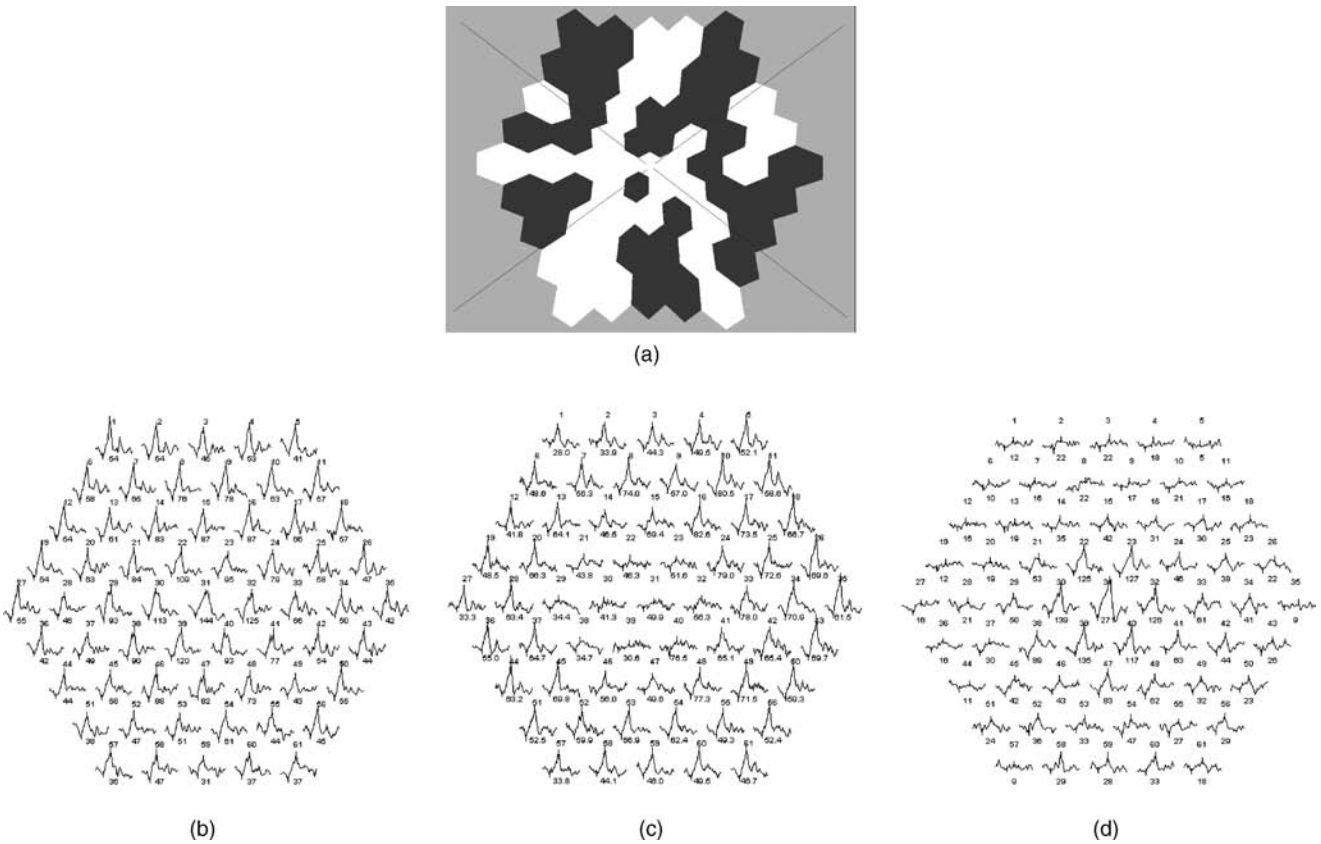


Figure 5. (a) the typical multifocal ERG stimulus; (b) a normal subject; (c) a macular dystrophy. There is loss of the responses to central hexagons but preservation of more peripheral responses. (d) A retinal dystrophy with sparing of central macular function but loss of the responses in the periphery.

be mistaken for Best disease, but although the EOG light rise may be mildly subnormal, it is not reduced to the same extent as in Best disease. The electrophysiological recordings will usually resolve any clinical dilemma in differential diagnosis.

ERG

Although the rod specific ERG b-wave is a sensitive indicator of retinal rod system dysfunction, the fact that it is generated in the inner-nuclear layer of the retina means that reduction in this response does not allow localization of the defect either to those structures or the upstream rod photoreceptors. It is the a-wave of the responses to brighter flashes that directly reflects activity of the photoreceptors and enables the distinction between photoreceptor dysfunction and a primary disorder of inner-retinal function. Genetically determined photoreceptor degenerations, such as the rod–cone (retinitis pigmentosa, RP) or cone–rod dystrophies, thus give overall ERG reduction (Fig. 4, columns C, D). The cone-derived ERGs in generalized photoreceptor degeneration characteristically show abnormalities of both amplitude and timing, particularly evident in the flicker ERG, but RP may occasionally only affect the rod-derived ERGs in the early stages of disease. Truly restricted disease, such as sector RP, is associated with amplitude reduction, but no implicit time change, whereas diffuse or generalized disease is usually also associated with an abnormally delayed implicit time. Retinitis pigmentosa is associated with pigmentary migration from RPE into retina consequent upon photoreceptor cell death, but the clinical appearance of the ocular fundus may not reflect the severity or nature of the disorder. Electroretinography not only enables accurate diagnosis, when interpreted in clinical context, but may also provide useful prognostic information. There is no rod system involvement in a pure cone dystrophy; such disorders have normal rod responses, but abnormal cone responses, with the 30 Hz flicker response usually showing both amplitude reduction and delay (Fig. 4, column E).

A waveform in which the bright flash a-wave is spared, but there is selective b-wave reduction, is known as a "negative" or electronegative ERG (e.g., Fig. 4, column F, row 2), and is associated with dysfunction postphototransduction, often postreceptoral. For example, in central retinal artery occlusion (CRAO) the finding of a "negative" ERG reflects the duality of the retinal blood supply, with the photoreceptors supplied via choroidal circulation, and the inner-nuclear layer supplied via the central retinal artery. Other causes of negative ERG include X-linked congenital stationary night blindness (CSNB, Fig. 4, column F), X-linked retinoschisis, quinine toxicity, melanoma

associated retinopathy (MAR, an autoimmune mediated disorder that can occur in patients with a history of cutaneous malignant melanoma), Batten disease (one of the ceroid neuronal lipofuscinoses), and occasionally cone–rod dystrophy. Carcinoma associated retinopathy (CAR), unlike MAR, usually give profound global ERG reduction in keeping with dysfunction at the level of the photoreceptor rather than a "negative" ERG. A negative ERG is also a relatively common occurrence in Birdshot chorioretinopathy (BCR), an inflammatory disease, but in that disorder such an appearance may normalize following successful treatment, usually with steroids and/or immunosuppressive agents.

The most common ERG abnormality in BCR, or other forms of inflammatory retinal disease, such as uveitis, is a delayed 30 Hz flicker ERG, but there may be much less marked amplitude change than occurs in photoreceptor degeneration. The ERG abnormalities may occur prior to the development of symptoms, and can normalize following treatment. Electrophysiology can thus play an important role not only in the characterization of the disease, but also in the initiation and monitoring of treatment (23). This relatively recent role of the ERG in the management of inflammatory disease can be expected to receive increasing clinical interest in the future.

PERG

Primary Evaluation of Macular Function. Disorders of macular function result in an abnormality of the P50 component of the PERG, often with preservation of the N95/P50 ratio. It is usually P50 amplitude that is affected; latency changes are only occasionally present, particularly in association with macular oedema or serous detachment at the macula. In clinical practice, the PERG and the (full-field) ERG provide complementary information regarding retinal function; the ERG assesses peripheral retinal function, and the PERG the degree of central retinal involvement. For example, dysfunction confined to the macula will have a normal ERG and an abnormal PERG (Fig. 4, column B), a common combination in macular dystrophies, such as Stargardt-fundus flavimaculatus (S-FFM), whereas generalized retinal dysfunction with macular involvement will have both an abnormal ERG and an abnormal PERG. This facilitates the distinction between macular dystrophy, cone dystrophy, and cone–rod dystrophy in a patient with an abnormal macular appearance and a history suggestive of a genetically determined disorder, important to the prognosis and accurate counseling of the patient. In relation to S-FFM, note that some patients have additional full-field abnormalities that may be of prognostic value (24).

The PERG may be normal even when the ERG is almost extinguished in patients with rod–cone dystrophy but normal central retinal function. Further, the objective assessment of macular function provided by the PERG can sometimes demonstrate early central retinal abnormalities prior to the appearance of symptoms or signs of macular involvement.

Ganglion Cell Dysfunction. The PERG will often be normal in disturbance of optic nerve function. However, there may be retrograde degeneration to the retinal ganglion cells in optic nerve disease and this may selectively affect the ganglion cell derived N95 component. It is N95 loss that is the common abnormality if the PERG is abnormal in optic nerve disease. That is unlike macular dysfunction, where it is the P50 component that is primarily affected. Shortening of P50 latency may also occur in more severe disease, but, again, is not a feature of macular dysfunction. Primary disorders of retinal ganglion cell, such as Leber hereditary optic neuropathy (LHON) and dominantly inherited optic atrophy (DOA), are associated with N95 component loss, marked at presentation in LHON, but often occurring later in the disease process in DOA. There may be additional P50 amplitude reduction in advanced retinal ganglion cell dysfunction, and the associated shortening of P50 latency then becomes an important diagnostic factor. Further, providing there is sufficient vision remaining in at least one eye to maintain fixation for binocular PERG recording, total extinction of the PERG probably does not occur in optic nerve disease. Even in an eye blind from optic nerve disease (no perception of light), a PERG may still readily be detectable (2).

THE PERG IN RELATION TO VEP INTERPRETATION

Although detailed discussion of the VEP is beyond the scope of this article, a short discussion of the use of the PERG in the improved interpretation of VEP abnormality is warranted. The cortically generated VEP to pattern reversal stimulation is a powerful clinical tool in the detection and assessment of optic nerve dysfunction, and pattern VEP latency delay or loss is frequently associated with optic nerve disease. However, the VEP is generated in the occipital cortex, and a delayed PVEP must never be assumed necessarily to indicate optic nerve dysfunction in a visually symptomatic patient. Similar abnormalities can occur either in macular disease or optic nerve disease. The appearance of the macula may be a poor indicator of function, and remember that a normal macular appearance does not necessarily equate to normal macular function. The different types of abnormality present in the PERG in optic nerve and macular diseases usually allow the differentiation between delayed VEP due to retinal macular disease and that due to optic nerve disease. An abnormal VEP with a normal PERG (or a normal P50 component with an abnormality confined to N95) is consistent with optic nerve/ganglion cell dysfunction, whereas pronounced P50 reduction suggests a disturbance of macular function (e.g., Fig. 4, columns B, D, E, F).

MULTIFOCAL ERG

The multifocal ERG can be used to assess the spatial extent of central retinal cone involvement in disease. Normal traces appear in Fig. 5b. Two clinical examples are shown; Fig. 5c shows a patient with a retinal dystrophy in whom there is sparing of central macular function; Fig. 5d shows a patient with a macular dystrophy with loss of the responses to central hexagons, but preservation of more peripheral responses. As a restricted test of central retinal

cone function, in clinical circumstances the mfERG should always be taken in conjunction with conventional full-field ERG.

CONCLUSIONS AND FUTURE DIRECTIONS

Diagnosis and management of the patient with visual pathway disease is greatly assisted by the objective functional information provided by electrophysiological examination. Separation of the function of different retinal cell types and layers enables characterization both of acquired and inherited retinal disorders, of great importance when counseling families affected by or at risk of a genetically determined disease. The PERG is complementary to the ERG by providing a measure of central retinal function; the mfERG may play a similar role.

BIBLIOGRAPHY

Cited References

1. Corbett MC, Shilling JS, Holder GE. The assessment of clinical investigations: the Greenwich grading system and its application to electrodiagnostic testing in ophthalmology. Eye 1995;9 (Suppl.): 59–64.
2. Holder GE. The pattern electroretinogram and an integrated approach to visual pathway diagnosis. Prog Retin Eye Res 2001;20:531–561.
3. Marmor MF. Standardization notice: EOG standard reapproved. Doc Ophthalmol 1998;95:91–92.
4. Marmor MF, Holder GE, Seeliger MW, Yamamoto S. Standard for clinical electroretinography (2004 update). Doc Ophthalmol 2004;108:107–114.
5. Bach M et al. Standard for Pattern Electroretinography. Doc Ophthalmol 2000;101:11–18.
6. Odom JV et al. Visual Evoked Potentials Standard (2004). Doc Ophthalmol 2004;108:115–123.
7. Heckenlively JR, Arden GB, editors. Principles and Practice of Clinical Electrophysiology of Vision. St. Louis: Mosby Year Book; 1991.
8. Fishman GA, Birch DG, Holder GE, Brigell MG. Electrophysiologic Testing in Disorders of the Retina, Optic Nerve, and Visual Pathway. 2nd ed. Ophthalmology Monograph 2. San Francisco: The Foundation of the American Academy of Ophthalmology; 2001.
9. Marmor MF, et al. Guidelines for basic multifocal electroretinography (mfERG). Doc Ophthalmol 2003;106:105–115.
10. Hood DC, Birch DG. Rod phototransduction in retinitis pigmentosa: Estimation of parameters from the rod a-wave. Invest Ophthalmol Vis Sci 1994;35:2948–2961.
11. Shiells RA, Falk G. Contribution of rod, on-bipolar, and horizontal cell light responses to the ERG of dogfish retina. Vis Neurosci 1999;16:503–511.
12. Bush RA, Sieving PA. Inner retinal contributions to the primate photopic fast flicker electroretinogram. J Opt Soc Am A 1996;13:557–565.
13. Bush RA, Sieving PA. A proximal retinal component in the primate photopic ERG a-wave. Invest Ophthalmol Vision Sci 1994;35:635–645.
14. Sieving PA. Photopic ON- and OFF-pathway abnormalities in retinal dystrophies. Trans Am Ophthalmol Soc 1993; 91:701–773.
15. Koh AHC, Hogg CR, Holder GE. The Incidence of Negative ERG in Clinical Practice. Doc Ophthalmol 2001;102: 19–30.
16. Arden GB et al. S-cone ERGs elicited by a simple technique in normals and in tritanopes. Vision Res 1999;39:641–650.
17. Arden GB et al. A gold foil electrode: extending the horizons for clinical electroretinography. Invest Ophthalmol Vision Sci 1979;18:421–426.
18. Dawson WW, Trick GL, Litzkow CA. Improved electrode for electroretinography. Invest Ophthalmol Vision Sci 1979;18: 988–991.
19. Hawlina M, Konec B. New noncorneal HK-loop electrode for clinical electroretinography. Doc Ophthalmol 1992;81:253–259.
20. Berninger TA. The pattern electroretinogram and its contamination. Clin Vision Sci 1986;1:185–190.
21. Holder GE. Significance of abnormal pattern electroretinography in anterior visual pathway dysfunction. Br J Ophthalmol 1987;71:166–171.
22. Viswanathan S, Frishman LJ, Robson JG. The uniform field and pattern ERG in macaques with experimental glaucoma: removal of spiking activity. Invest Ophthalmol Vis Sci 2000;41:2797–2810.
23. Holder GE, Robson AG, Pavesio CP, Graham EM. Electrophysiological characterisation and monitoring in the management of birdshot chorioretinopathy. Br J Ophthalmol 2005; in press.
24. Lois N, Holder GE, Bunce C, Fitzke FW, Bird AC. Stargardt macular dystrophy—Fundus flavimaculatus: Phenotypic subtypes. Arch Ophthalmol 2001;119:359–369.

See also BLIND AND VISUALLY IMPAIRED, ASSISTIVE TECHNOLOGY FOR; CONTACT LENSES; VISUAL PROSTHESES.

ELECTROSHOCK THERAPY.

See ELECTROCONVULSIVE THERAPY.

ELECTROSTIMULATION OF SPINAL CORD.

See SPINAL CORD STIMULATION.

ELECTROSURGICAL UNIT (ESU)

JOHN PEARCE
The University of Texas
Austin, Texas

INTRODUCTION

Electrosurgery means the application of radio frequency (RF) current at frequencies between $\sim$300 kHz and 5 MHz to achieve a desired surgical result; typically the fusion of tissues or surgical cutting in which the tissue structure is disrupted. In either case, the effect is achieved by heat dissipated in the tissues from the RF current by resistive, or joule, heating. This method has the ability to cut and coagulate tissues simultaneously; and, as a consequence, has made substantial contributions to several branches of clinical medicine since its introduction in the late 1920s.

The tissue effects applied in electrosurgery are typically described as (a) white coagulation, named for its appearance, in which the tissue proteins are degraded at lower temperatures, typically 50–90 °C; (b) black coagulation or carbonization in which tissues are completely dried out (desiccated) and reduced to charred carbonaceous

remnants at higher temperatures; and (c) cutting in which tissue structures are separated by the rapid boiling of small volumes of tissue water. These three results usually occur in some combination depending on the applied current and voltage at the so-called active or surgical electrode.

Electrosurgery accomplishes many surgical jobs better than any other device or technique while drastically reducing the morbidity and mortality associated with surgery. It does this by reducing the time under anesthesia and complications due to operative and postoperative hemorrhage. Many of the delicate techniques in neurosurgery would be impossible without electrosurgery—and it is likely that open heart surgery and much of urologic surgery would likewise not be done.

Historical Background

The application of heat for the treatment of wounds dates back to antiquity. According to Major (1), Neolithic skulls unearthed in France show clear evidence of thermal cauterization. The Edwin Smith papyrus ($\sim$3000 BC) (2) describes the use of thermal cautery for ulcers and tumors of the breast. Licht (3) reports that according to Jee (4) the ancient Hindu god Susruta, the highest authority in surgery, said that "caustic is better than the knife, and the cautery is better than either." Cautery in ancient Hindu medicine included heated metallic bars, boiling liquids, and burning substances.

In cauterization the essential physical mechanism behind the treatment is conduction heat transfer from a hot object placed on the surface to raise the temperature high enough to denature the tissue proteins. Cutting and coagulation by means of electrosurgery is also accomplished by heating tissue to high temperatures, but the essential difference is that the primary mechanism is electrical power dissipation directly in the affected tissues themselves, rather than heat transfer from an external hot object on the tissue surface. It is rather like the difference between heating food in a conventional oven and a microwave oven, in a loose sense. Electrosurgery is sometimes erroneously referred to as electrocautery. Since the physical mechanisms are different it is important to keep these two techniques separate by precise terminology. Electrosurgery is also referred to as surgical diathermy, particularly in Europe. While this term is generally understood, it is a bit of a misnomer since diathermy literally means through-heating, such as might be applied in physical medicine for the relief of pain or in hyperthermia therapy for tumor treatment. Electrosurgical devices in operating rooms are designed and built for surgical use only, and the terminology in this section is standardized on that basis.

Early Experiments with High Frequency Current. The origin of the application of rf current to achieve surgical results is difficult to establish owing to the rapid pace of development in the electrical arts during the late nineenth and early twentieth centuries. Lee De Forrest, the inventor of the vacuum tube (the audion, 1907 and 1908), filed a patent for a spark gap rf generator to be used for electrosurgery in February of 1907 (it was granted in December of 1907) (5). Also, during the same year, Doyen noted that the effect of a surgical arc on the tissue was not a function of the length of the arc, and that the temperatures of carbonized tissue were as high as 500–600 °C (6). He also found that final temperatures in the range of 65–70 °C resulted in white coagulation while there was no damage to tissues for temperatures < 60 °C (7).

By far, the most effective promoters of electrosurgery were Cushing and Bovie. W. T. Bovie was a physicist attached to the Harvard Cancer Commission. He had developed two electrosurgical units, one for coagulating and one for cutting. Harvey Cushing, the father of neurosurgery, had been concerned for some time with the problem of uncontrolled hemorrhage and *diabetes insipidus*, the often fatal complications of hypophysectomy, among other concerns (8). In 1926, working with Bovie, he applied high frequency current in cerebral surgery with excellent results. They published their work in 1928, emphasizing the three distinct effects of electrosurgery: desiccation, cutting and coagulation (9).

Early Electrosurgical Generators. Cameron-Miller offered the Cauterodyne in 1926, similar to the later model of 1930 that featured both vacuum tube cutting and spark gap coagulation. It came in a burled walnut bakelite case for $200. It is not known if the original 1926 device included tube cutting. The device designed and manufactured by W. T. Bovie was of higher fundamental frequency than the other early devices. The Bovie device used a 2.3 MHz vacuum tube oscillator for pure cutting (i.e., Cut 1) and a 500 kHz fundamental frequency spark gap oscillator for fulguration and desiccation (Cut 2–4 and Coag). The overall size, circuit and configuration of the Bovie device remained essentially constant through the 1970s. Bovie's name is so closely associated with electrosurgery that it is frequently referred to as the Bovie knife in spite of the extensive work which preceded his device and the numerous devices of different manufacture available then and now. In essence, the available electrosurgical generators between $\sim$1930 and the 1960s were of the similar design to that used by Bovie, and consisted of a spark gap generator for coagulating and a vacuum tube generator for cutting.

The introduction of solid-state electrosurgical generators in the early 1970s by Valleylab and EMS heralded the modern era of isolated outputs, complex waveforms, more extensive safety features and hand-activated controls. Interestingly, hand-activated controls are actually a recent rediscovery and improvement on those used by Kelly and Ward (10) and those available on the Cauterodyne device. In some ways, higher technology devices of all designs are made inevitable by the recent proliferation of delicate measurement instrumentation that is also attached to a patient in a typical surgical procedure.

Clinical Applications

The topics chosen in this introductory survey are by no means comprehensive. Those readers interested in specific surgical techniques should consult the texts by Kelly and Ward (10) and Mitchell et al. (11) for general, gynecologic, urologic and neurosurgical procedures; Otto (12) for minor

electrosurgical procedures; Harris (13), Malone (14), and Oringer (15,16) for dental electrosurgery; and Epstein (17) and Burdick (18), for dermatologic procedures.

When high frequency currents are used for cutting and coagulating, the tissue at the surgical site experiences controlled damage due either to disruptive mechanical forces or distributed thermal damage. Cutting is accomplished by disrupting or ablating the tissue in immediate apposition to the scalpel electrode. Continuous sine waveforms (e.g., those obtained from vacuum tube or transistor oscillators) have proven most effective for cutting. Coagulating is accomplished by denaturation of tissue proteins due to thermal damage. Interrupted waveforms, such as exponentially damped sinusoids (obtained from spark gap or other relaxation-type oscillators) are effective for coagulation techniques requiring fulguration, or intense sparking (*fulgur* is Latin for lightning). However, when no sparks are generated, coagulation is created by tissue heating alone, and the specific waveform is immaterial: only its effective heating power, or root-mean-square (rms) voltage or current determine the extent of the effect. Suffice it to say that the difference between cutting and coagulation is due to combined differences in heat-transfer mechanisms and spatial distribution of mechanical forces. In general, the tissue damage from cutting current is confined to a very small region under the scalpel electrode and is quite shallow in depth. Cells adjacent to the scalpel are vaporized and cells only a few cellular layers deep are essentially undamaged. Though dependent on surgical technique, generally only the arterioles and smaller vessels are sealed when a cut is made using pure sine wave currents. Coagulation currents are used to close larger vessels opened by the incision, to fuse tissue volumes by denaturing the proteins (chiefly the collagen) and to destroy regions of tissue. The tissue damage when coagulating is deeper than when cutting. In the majority of applications for coagulating current, the damage in the tissue is cumulative thermal damage rather than tissue disruption or ablation.

Cutting is a monopolar procedure, although some experiments have been performed with bipolar cutting. That is, the scalpel electrode represents essentially a point current source and the surgical current is collected at a remote site by a large area dispersive, or return electrode. Coagulation may be accomplished using either monopolar or bipolar electrodes. In bipolar applications, both the current source and current sink electrodes are located at the surgical site. A typical bipolar electrode might consist of forceps with the opposing tongs connected the two active terminals of the generator. In both cutting and coagulating processes, whether monopolar or bipolar electrodes are used, a layer of charred tissue often condenses on the cool electrode that must periodically be removed (19).

The histologic effects of cutting current are varied and apparently depend on technique. Knecht et al. (20) found that the healing process in an electrosurgical incision was a bit slower than for incisions made by a cold scalpel. In their study, the wound strength of an electrosurgical cut was less than that of a cold scalpel cut until ~ 21 days after surgery. After 21 days, no difference in wound strength was measurable. Ward found that an electrosurgical cut generally formed slightly more scar tissue on healing than a cold scalpel cut if the closure of the two wounds was identical (21). The cellular layers within ~ 0.1 mm of the scalpel electrode showed electrodesiccation effects when sine wave cutting was used (21). In a later series of studies on tissues of the oral cavity, Oringer observed that when the cutting current was carefully controlled, the damage was confined to the cut cellular layer, and the layer of cells adjacent to the cut was undamaged (14,15). The cell destruction was apparently self-limiting to the extent that no damage to the cytoplasm or cell nucleus of the cut layer was visible in light or electron micrographs (14). Oringer (16) describes the margin of an excised squamous cell carcinoma that had been removed with electrosurgery. Under the electron microscope at a magnification of 47,400 the margin was seen to contain several clear examples of cells sheared in half with no damage to the remainder. Oringer, and others, observed faster healing with less scar tissue in the electrosurgical incision. The variety of results obtained is likely due to differences in waveform, surgical technique, tissue characteristics and scalpel electrodes used in the studies.

When combined sine wave and interrupted (coagulating) waveforms are used, or when spark gap sources are used for cutting, a coagulum layer extends deeper into the tissues under the desiccated layer (21). Coagulation techniques include (*1*) fulguration (also called spray coagulation or black coagulation) in which the tissue is carbonized by arc strikes, (*2*) desiccation, in which the cells are dehydrated resulting in considerable shrinking, and (*3*) white coagulation, in which the tissue is more slowly cooked to a coagulum. In fulguration techniques, the active electrode is held a few millimeters from the surface of the tissue and arcs randomly strike from the electrode to the tissue. The cell structure is destroyed at a very high temperature resulting in charring of the tissue. In desiccation, the water is evaporated from the cell relatively slowly leaving a white dry flake of powder, the cells appear shrunken and drawn out with elongated nuclei. The overall cellular anatomical characteristics are preserved (21). Desiccation techniques normally take longer to accomplish than fulguration for the same volume of tissue. In white coagulation, the electrode is in intimate contact with the tissue. No arcs strike so the electrode voltage is low by comparison. The total electrode current may be high, but the tissue current density (current per unit area) at all points on the electrode is moderate and the duration of the activation is therefore relatively long. The cellular effects are varied, ranging from a tough coagulum when connective tissue predominates to granular debris easily removed by a curet when the majority of the tissue is epithelial (21). Often the goal of coagulation is to shrink and fuse or thermally damage tissue collagen.

Minor Surgery. Minor surgery may be described as surgery applied to tissues on exterior surfaces of the body under local anesthetic, which includes dermatology. Other external surfaces, which include oral, vaginal and cervical tissues, are also routinely treated as minor surgery cases on an outpatient basis or in the office. Typical minor surgery procedures include the removal of warts, moles

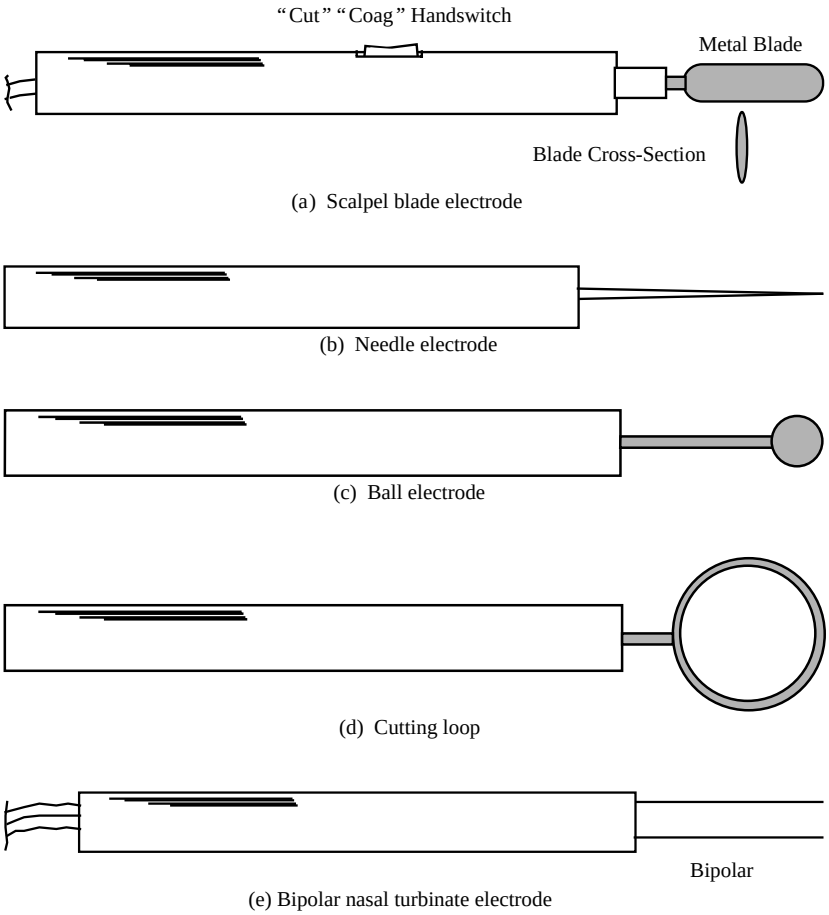


Figure 1. Electrodes typically used for general electrosurgery.

and vascular nevi, the surgical reduction of sebaceous cysts, epiliation, cervical conization, relief of chronic nasal obstruction, and the removal of polyps and tumors. The techniques used in minor surgery are very similar to those of general surgery, although the cutting electrodes come in a wider variety of shapes and sizes.

Figure 1 illustrates some of the electrodes used (19). The standard scalpel blade electrode used for incision is elliptical in cross-section with a cutting edge which is of small radius (not sharp) in order to yield the very high current densities required for tissue ablation. The scalpel electrode is sometimes angled for special techniques. Scalpel electrodes are usually used for incisions, but may also be used for coagulation. The flat side of the blade can be applied to a tissue mass to obtain coagulation. The tip of the electrode may be suspended above the tissue for fulguration. Other electrode shapes accumulate less carbonized tissue residue than the scalpel, and are often preferred for coagulation. The needle, standard coagulation and ball electrodes are used either for desiccation or fulguration. The ball electrode may be used with coagulating current to treat persistent nose bleed that does not respond to other methods.

Bipolar forceps and the turbinate electrode are used in bipolar procedures. The forceps electrode has one electrical connection on each side of the forceps, and is used, for example, to grasp a seeping vessel; current between the forceps electrodes then seals off the vessel end by fusing the vessel walls together. The turbinate electrodes are used to obtain submucous coagulation (desiccation) in the nasal turbinates for relief of chronic vasomotor rhinitis: swelling of the soft tissue in the nasal cavity caused by, for example, allergies or irritants.

Neurosurgery and General Surgery. Blood is toxic to neural tissue. Consequently, blood loss during neurosurgery is to be avoided at all costs. At its inception, electrosurgery presented the first technique that accomplished cutting with negligible blood loss. This was this feature of electrosurgery that so strongly attracted Cushing. Many of the now commonplace neurosurgical procedures would be impossible without some method for obtaining hemostasis while cutting.

White coagulation is generally used in neurosurgery since it does not cause charring. White coagulation takes a relatively long time to obtain compared to fulguration or cutting. Holding the scalpel electrode to a location for long activation times allows deep penetration of the high temperature zone. This effect is used to advantage in rf lesion generation for selective disabling of neural tissue. Bipolar applicators restrict the current to a smaller region and are used extensively for microneurosurgical techniques, since they are more precise and safer. Many of the techniques are properly classed as microsurgery. Often the tissue being cut or coagulated is stabilized with a suction probe while current is applied. The suction probe is usually nonconductive glass or plastic; however, on occasion a metal

suction probe is used with monopolar coagulation to localize the current to the immediate area (11). Incisions of the cerebral cortex are usually made with monopolar needle electrodes. Surface tumors are removed by applying a suction probe to the tumor and excising it at the base with a cutting loop.

Dental Electrosurgery. The tissues of the oral cavity are particularly highly vascularized. Also, the mouth and alimentary canal contain high concentrations of bacteria. Since the ability to ingest food is critical to survival these tissues are among the fastest healing of the body. Electrosurgery plays an important role in oral surgery in that it drastically reduces bleeding that would obscure the operative field and the undesirable postoperative effects of pain, edema and swelling in the submaxillary triangle (15). It can be quite difficult to accurately resect redundant tissue masses by cold scalpel techniques in the oral cavity. Electrosurgery allows accurate resection with minimal elapsed time and complications, an important feature when fitting prosthodontic devices. Electrosurgery reduces the hazard of transient bacteremia and secondary infection (14), and the danger of surgical or mechanical metastasis of malignant tumor emboli during biopsy (15). In short, all of the advantages obtained in other types of surgery are experienced in dental surgery as well as some additional beneficial aspects.

The active electrodes used in dental electrosurgery are for the most part similar to those shown in Fig. 2 (19). Several shapes specific to dental procedures are: the open hook electrodes in Fig. 2a are used along with other needle electrodes to create subgingival troughs, those in Fig. 2b are used with ball electrodes, and loop electrodes for seating orthodontic appliances and denture prostheses and exposing roots. Other interesting applications include the removal of premalignant neoplastic lesions from the surface of the mucosa using a planing loop and the removal of tissue masses from the tongue. Carefully applied electrosurgery can be used to advantage on teeth as well as the soft tissues of the oral cavity.

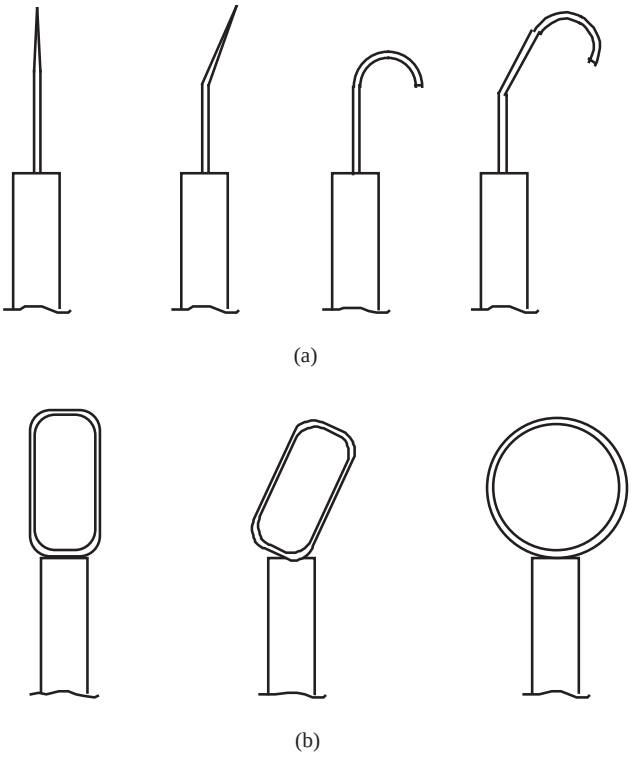


Figure 2. Electrodes typically used for dental electrosurgery. (a) Open hooks and needles, straight and angulated to reach difficult locations. (b) Straight and angulated cutting loops.

Urologic Surgery. Electrosurgery is used extensively in urologic procedures. Urologic procedures, specifically transurethral resections of the prostate, utilize by far the highest currents at high voltage for the longest durations and largest number of activations per procedure of any electrosurgical technique. Other urologic applications of electrosurgery include: the resection of bladder tumors, polyp removal by means of a snare electrode or desiccating needle, kidney resection to remove a stone-bearing pocket, and enlarging the urethral orifice in order to pass stones. These other procedures utilize power levels similar to those of general surgery.

A transurethral resection is intended to increase the caliber of a urethra that has been partially closed by an enlarged prostate gland. This procedure is one of the earliest applications of electrosurgery in urology, having been described in considerable detail by Kelly and Ward in 1932 (21). During transurethral resection, the urethra is irrigated with a nonconductive sterile solution, either dextrose or glycerine, while a cutting loop is advanced to remove the encroaching tissue. Surgical cutting accomplished in a liquid medium requires higher currents since the liquid carries heat away and disperses the current (even though it is nonconductive) more than a gaseous environment would. A typical resectoscope used in this procedure is shown diagrammatically in Fig. 3. A long outer metal sheath, which is plastic coated, contains a cutting loop that can be extended from the sheath, a fiber optic cable bundle for viewing the cutting action, an illumination source (either a bulb at the end or an illumination fiber optic bundle), and spaces for influx and efflux of irrigating fluid. The cutting loop (of thin diameter to yield the high current densities required) is moved in and out of the sheath by the surgeon as required to accomplish the resection.

Gynecologic Surgery. One of the common gynecologic applications of electrosurgery is cervical conization. Other common gynecologic applications of electrosurgery include the removal of tumors, cysts, and polyps. The use of electrosurgery in laparoscopic tubal ligations will be treated in some detail since it is also a very common procedure. The laparoscopic tubal ligation is a minimally invasive sterilization procedure typically accomplished by advancing an electrode through a small incision in order to coagulate the Fallopian tube. The position of the uterus is fixed by a cannula inserted through the cervix under the surgeon's control. The abdominal cavity is insufflated with CO_2 gas in order to separate the tissue structures. The coagulating electrode is then advanced through a small incision in the abdominal wall to the Fallopian tube by the surgeon,

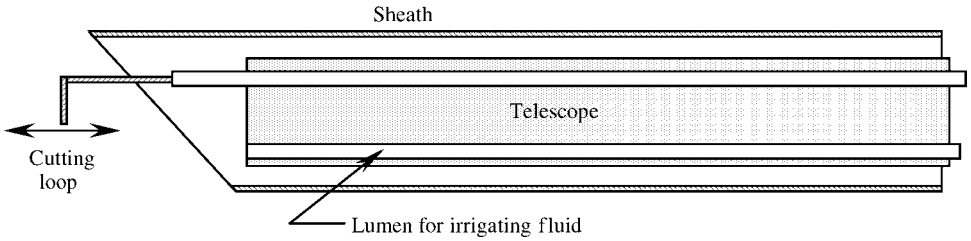


Figure 3. Resectoscope typical for transurethral resections of the prostate (TURP).

observing through an endoscope (laparoscope), which has been inserted through a separate small incision. The isolated Fallopian tube is then coagulated at high current but relatively low voltage.

Both monopolar and bipolar electrosurgical electrodes have been used for this procedure; however, monopolar tubal ligation methods are to be avoided as there have been many instances of bowel wall perforations and other complications following monopolar tubal ligation procedures owing to surgical current flow in the bowel wall. Bipolar techniques confine the current to a small region of tissue and the risk of bowel perforation is minimal in comparison. A bowel wall perforation can still result due to heat transfer from the coagulated tissue, so the coagulating forceps and Fallopian tube must be held away from the bowel wall and allowed to cool before being released. Note that a surrounding CO_2 gas environment will increase the time required for tissue cooling.

FUNDAMENTAL ENGINEERING PRINCIPLES OF ELECTROSURGERY

RF Generators for Electrosurgery

In general, the RF frequencies used for electrosurgery fall between 350 kHz and 4 MHz, depending on manufacturer and intended use. The available output power ranges from $\sim$30–300 W. Peak open circuit output voltages vary from $<$200 V to 10 kV. Higher open circuit voltages are used to strike longer arcs for fulguration, while the lower voltages are used for bipolar coagulation. Most devices are capable of generating several different waveforms, said to be appropriate for differing surgical procedures.

Vacuum Tube and Spark Gap Oscillators. The original rf generators used in electrosurgery, diathermy, radiotelegraphy, and radar circuits were spark gap oscillators (Fig. 4). The exponentially damped sine wavefrom (Fig. 4b) results from a breakdown of the spark gap (SG in Fig. 4a) that initializes the oscillations. The waveform is often called a Oudin waveform, though it is typical of all spark gap units, Oudin output circuit or not. The RFC is a RF choke to prevent the rf signal from coupling to the power line. Later generator designs utilized vacuum tube oscillator circuits that were typically Tuned-Plate, Tuned-Grid Oscillators (22), as shown in Fig. 5a (Birtcher Electrosectilis), or Hartley oscillators, Fig. 5b (Bovie AG). The output of vacuum tube electrosurgical units is available as either partially rectified [meaning that the RF oscillator is active only on one of the half-cycles of the mains power (i.e., one vacuum tube)] or fully rectified [meaning that the RF oscillator is active on both half-cycles of the mains power (Fig. 5c)]. In both circuits of Fig. 5 each vacuum tube, V1 and V2, oscillates on opposite half cycles of the mains power, period T. Electrosurgery generator designs varied little from the standard units built by Bovie and Cameron-Miller in the 1920s until $\sim$1970 when solid-state generators became available. Solid-state generators made possible much more sophisticated waveforms and safety devices in a smaller overall package. Though not specific to solid-state technology, isolated outputs became common when solid-state electrosurgery units were introduced. Until $\sim$1995 all electrosurgery generators acted essentially as voltage sources with a typical output resistance in the

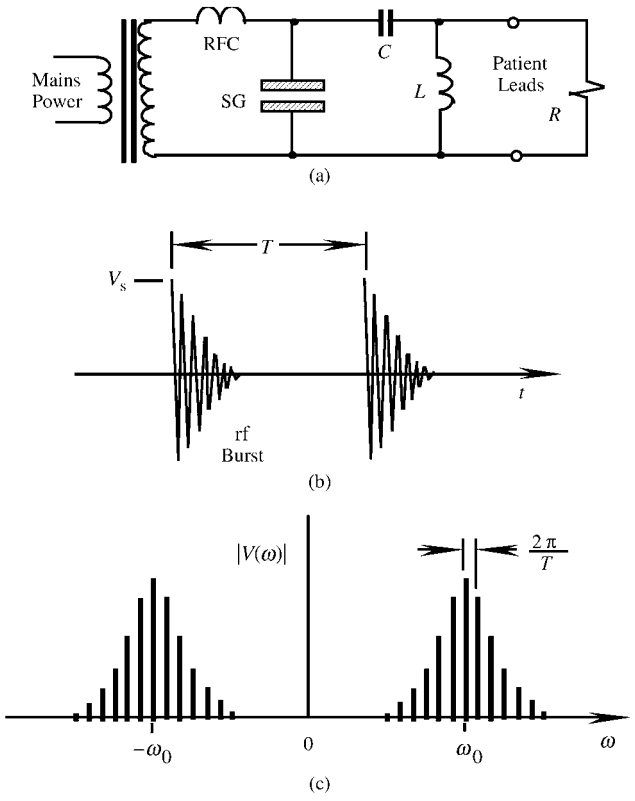


Figure 4. Spark gap oscillator rf generator. (a) Spark gap oscillator circuit. SG = spark gap, which acts as a switch, RFC = radio frequency choke, L and C determine the fundamental rf frequency, f_0 and R the damping (R = the patient). (b) Oudin waveform with amplitude determineed by the supply voltage peak, V_s. (c) Frequency spectrum of the output Oudin waveform has energy centered at $\pm_0$, the fundamental RF oscillation frequency with energy concentrated at harmonics of the repeat frequency to both high and low frequencies.

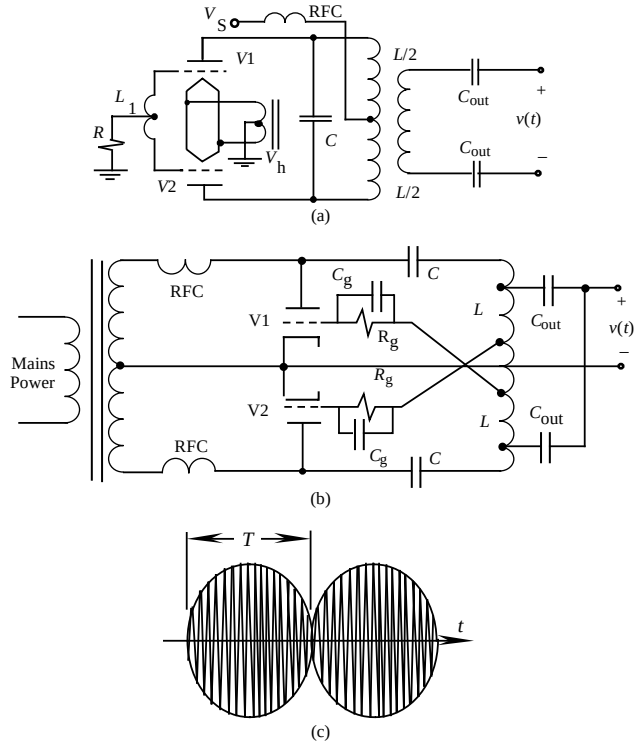


Figure 5. Vacuum tube electrosurgical circuits. (a) Tuned plate, tuned grid oscillator, as used in the Birtcher Electrosectilis. (b) Modified Hartley oscillator, as used in the Bovie AG. (c) Fully rectified output waveform.

neighborhood of 300–500 Ω. Recent generator designs incorporate embedded microprocessors and have constant power delivery modes. Interestingly, both Bovie and Cameron-Miller electrosurgical devices are available in the present day, though the designs are considerably different.

Solid-State Generators. Solid-state electrosurgical generators generally operate on a different design principle than the vacuum tube devices. Rather than combine the oscillator and power amplifier into one stage, solid-state generators utilize wave synthesis networks that drive a power amplifier output stage. This approach has the advantage that quite complex waveforms may be employed for cutting and coagulating; although to date the waveforms used vary little, if at all, from those of vacuum tube and spark gap oscillators. Many solid-state generators (chiefly those with bipolar transistors in the output amplifier) do not have as high open circuit output voltages as vacuum tube and spark gap devices. It sometimes appears to the user of those devices that there is less power available for cutting and coagulating since the lower open circuit voltages will not strike arcs as far away from the tissue. This turns out to be a limitation of concern in the high voltage procedures, namely in TURPs, but not in high current procedures, such as laparoscopic tubal ligations. In general, solid-state generators that use bipolar transistors in the high voltage output amplifier stage are vulnerable to transistor failure. The more recently introduced solid-state generators (after $\sim$1985) employ high voltage VMOS or HEXFET field effect transistors (23) in the output stage to give higher open circuit voltages and/or to reduce the stress on the bipolar output transistors.

A general block diagram of a typical solid-state electrosurgical generator is shown in Fig. 6 (24). The fundamental frequency, most often $\sim$500 kHz, is generated by a master oscillator circuit, typically an astable multivibrator. The primary oscillator acts as the clock or timing reference for the rest of the generator. The unmodified master oscillator signal is amplified and used for cutting. An interrupted waveform is formed by gating the continuous oscillator output through an external timing circuit, as shown in the figure. The repeat frequency of the timer is typically on the order of 20 kHz (24), much higher than that of spark gap oscillators. The duty cycle of a waveform is the ratio of duration of the output burst to the time between initiation of bursts. Duty cycles for solid state coagulating waveforms vary, but a duty cycle of 10–20% would be considered typical. This is in sharp contrast to the spark gap devices that have duty cycles often $< 1\%$. The higher duty cycle of solid-state units compensates in part for their lower peak output voltages so the actual maximum available power is similar in both families of devices.

Constant power output is obtained by measuring the output voltage and current and adjusting the drive signal to compensate for changes in the equivalent load impedance (25), as in Fig. 7a. The sampling rate for this adjustment is on the order of hundreds of hertz ($\sim$200 Hz for the device depicted). In Fig. 7b, the performance of the example system is compared to a standard voltage source generator

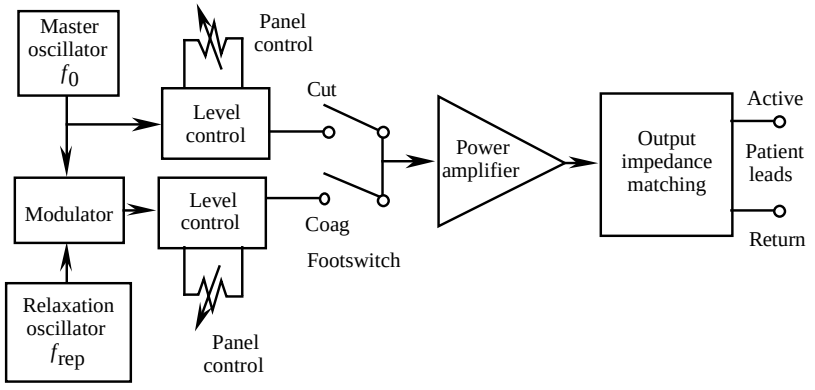


Figure 6. Block diagram for typical solid state ESU. Footswitch (and hand switch) controls simplified by omitting the interlock circuitry that prevents simultaneous activation. Master oscillator sets fundamental RF frequency, f_0. Interrupted waveform repeat frequency, f_{rep}. Provisions for blending cut and coag modes often provided. Power amplifier either bipolar junction transistors or HEXFET or VMOS transistors.

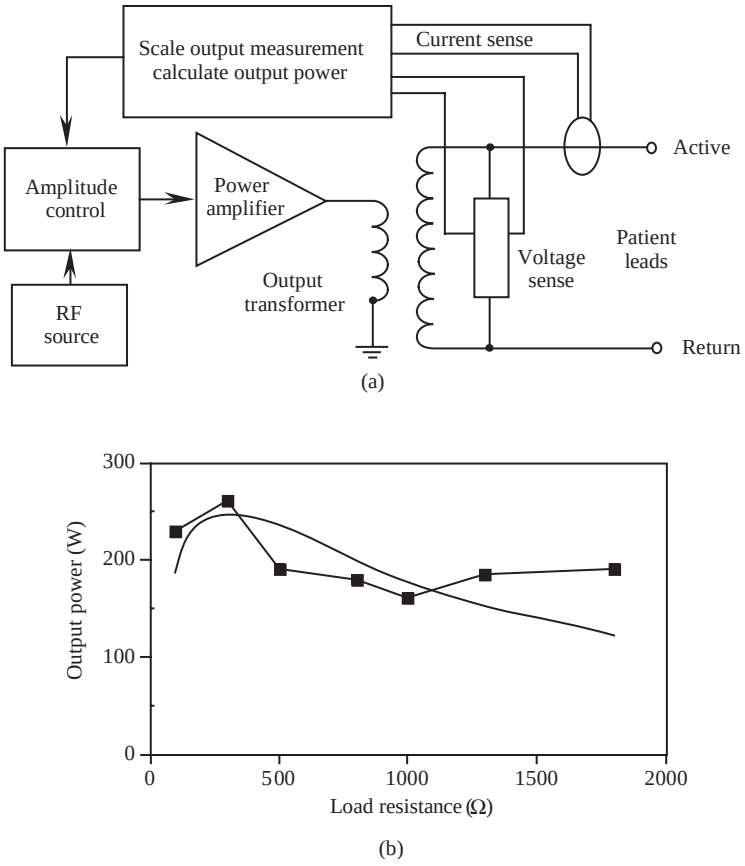


Figure 7. Constant power generator system. (a) Descriptive block diagram. (b) Typical performance (squares) compared to standard voltage source with fixed source voltage and equivalent output impedance (solid line), Nominal 250 W into 300 Ω (24).

with fixed output impedance 300 Ω and maximum output power of 250 W at 300 Ω load impedance. The differences at high load impedances typical of electrosurgical cutting (up to ∼2 kΩ) are easily seen in the figure.

Safety Appliances. Since the electrosurgical unit (ESU) represents a high power source in the operating room, the potential for adverse tissue effects and accidental results must be managed. One of the more common types of accident has historically been one or more skin burns at the dispersive, or return, electrode site (during monopolar procedures). Often these have resulted from the return electrode losing its attachment to the skin. Earlier ESU designs incorporated a so-called circuit sentry, or its equivalent, which ensured that the return electrode cable made contact with the return electrode by means of monitoring current continuity through the cable and electrode connection at some low frequency. This ensured that the electrode was connected to the ESU, but did not ensure that it was connected to the patient.

Modern ESU devices monitor the patient return electrode to ensure its continued connection to the skin. In these devices, the patient return electrode is divided approximately into halves and a small current (typically on the order of 5 mA or less) at high frequency but below the surgical frequency (typically in the neighborhood of 100 kHz) is applied between the electrode halves, the connection being through the skin, only, as in Fig. 8 (25). The impedance between them is monitored, and if it exceeds maximum or minimum limits, or a set fraction of the baseline value (typically 40% or so over the baseline) an alarm sounds and the generator is deactivated. The patient return monitor system resets the baseline impedance if it falls below the initial value.

The high frequency of operation of electrosurgery units also contributes to the hazards associated with its use. At very high frequencies, current may flow in two different ways. First, in good conductors, such as skin and other wet tissues, the current flows by conduction and Ohm's law applies. Second, in insulating dielectric substances, such as air, surgical rubber or plastic, the current flows as so-called displacement current. The value of the displacement current density is linearly related to the frequency, so for the same materials, voltages and geometric arrangement of conductors, higher frequencies conduct more displacement current in dielectric substances than do lower frequencies. A consequence of this relationship is that at electrosurgical frequencies and voltages the capacitance between the wire conductor in a scalpel electrode and tissue over which the wire passes may have low impedance if the insulation is thin. Scalpel electrode wires are covered with thick insulation of high dielectric constant in order to prevent tissue damage at high scalpel electrode voltages. If the wire is inadequately designed or the insulation is damaged, the displacement current in the wire insulation may be dense enough to cause damage to the underlying tissue.

Electrosurgical unit outputs may be isolated, referred to ground, or grounded terminals (Fig. 9). Grounded patient return leads (Fig. 9a) have a hard wired connection to the ESU chassis ground wire. Referred to ground means that

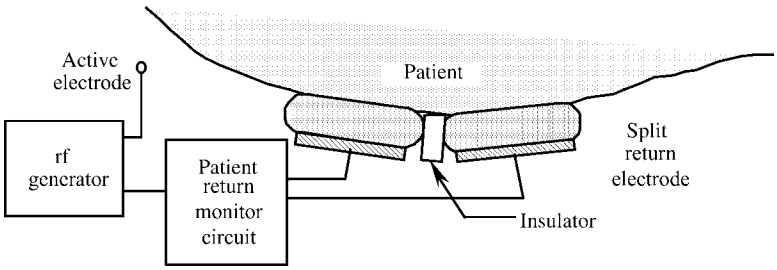


Figure 8. Patient return monitor circuit ensures that the return electrode remains attached to the patient. Split return electrode conforms to patient contours; both halves of return electrode carry surgical rf current. Insulator prevents lower frequency interrogation current from finding a shorter pathway. Current pathway is through the patient's skin, ensuring contact.

there is a fixed impedance, typically a capacitor, between the patient return lead and ground (Fig. 9b). The capacitance is chosen to represent a low impedance at the ESU rf frequency and a high impedance at power mains frequency and frequencies associated with stimulation of electrically excitable tissues. Isolated outputs (Fig. 9c) have high impedance to earth ground at both output terminals, usually by means of an output transformer; these are almost uniformly used for bipolar modes, and are common in monopolar modes as well. No isolation system is ideal,

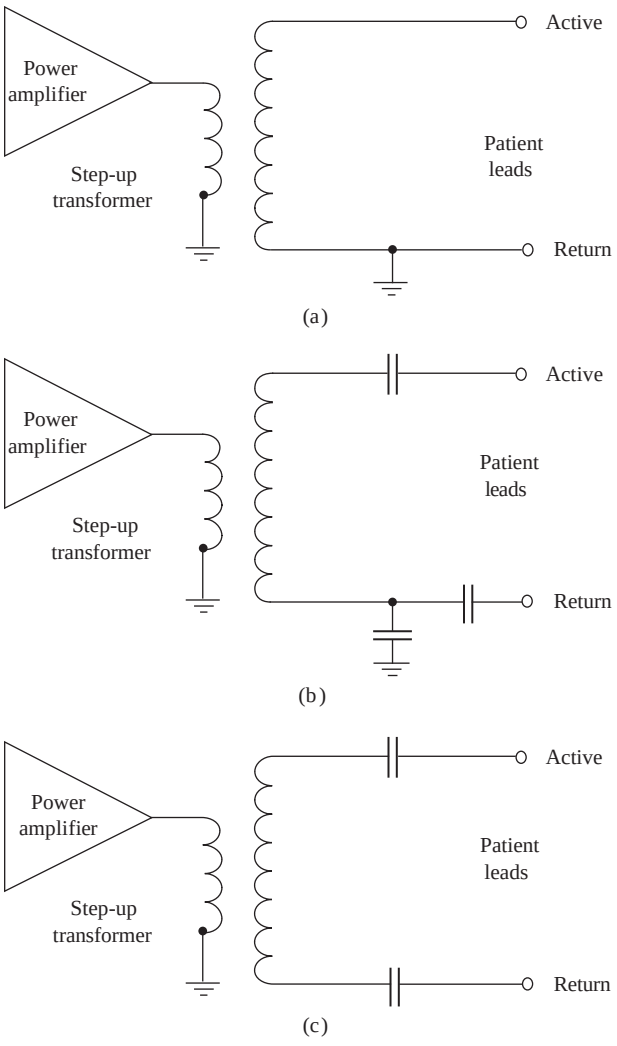


Figure 9. Generator output impedance matching circuits. The step up transformer is used to achieve higher open circuit output voltages. (a) grounded, (b) referred-to-ground, (c) isolated.

however, and all rf currents should be thought of as ground seeking in some sense. In an isolated system, either patient lead can become a source of RF current.

A less obvious consequence of the high frequency is that every object in the room (the surgeon, the patient, the operating table and associated fixtures, the electrosurgical generator, other instrumentation) all have a small but finite parasitic or distributed capacitance to earth. This makes all of the RF currents in the operating room ground seeking, to some extent. Even the best isolated generator will have some ground leakage current at RF, and this value may be much more than the mains frequency leakage current (depending on the design of the output circuitry). Certainly, all of the grounded output generators can have significant differences between the active cable current and the return or dispersive electrode cable current. The difference current flows in all of the parallel pathways, conductive and/or capacitive. If any of these pathways carry too much current in too small an area, a thermal burn could develop. To the extent reasonable, direct conductive pathways through clamps, foot stirrups, and other metallic fixtures should be eliminated. This can be accomplished by using devices that have insulating covers over the pieces likely to contact the tissue, or insulated couplings, connectors or other barriers at some location between the tissue and the clamp which connects the device to the surgical table. Additional safety can be obtained by using monitors and other instruments that have RF isolation built into their electrode leads. These precautions and others will greatly reduce the hazards associated with alternative current pathways. There are International Electrotechnical Commission standards that cover the requirements for electrosurgical applicators, generator output connections, and isolation schemes (26,27). It is important to note that safe operation of electrosurgical devices can be obtained by more than one design strategy. The ESU, patient, and surroundings should be thought of as a system, in order to ensure a safe environment.

Representative Surgical Procedures

The output voltages and currents that are required of an electrosurgical generator depend on the particular procedure for which it is to be used. Fulguration requires high voltages to initiate the arcs, but not large currents, so a generator of high output impedance works quite well. Spray coagulation uses higher currents at the high voltages, the difference between spray coagulation and fulguration being one of degree rather than principle. White coagulation requires relatively high current at low voltage

since no arc is formed at the scalpel electrode. The highest power, that is voltages and currents together, is required for transurethral resections of the prostate (TURP) procedures. The cutting characteristics of the various generator designs make them more or less optimal for certain procedures. Given the variety of designs available and variations in surgical technique among the users, it is no surprise to find that generator selection often boils down to personal preference.

The two modes used in electrosurgical procedures, bipolar, and monopolar, have quite different current field distributions. The bipolar mode is very effective in coagulating small vessels and tissue masses. Bipolar techniques are especially useful in microsurgical procedures. In bipolar electrosurgery, a two-wire coagulating electrode is clamped around a tissue lump or vessel and electrosurgical current is applied between the electrodes until the desired surgical effect is obtained. In monopolar electrosurgery, there are two separate electrodes, an active or cutting electrode and a large area dispersive or ground or return electrode located at some convenient remote site. Current is concentrated at the active electrode to accomplish the surgery while the dispersive or return electrode is designed to distribute the current in order to prevent tissue damage. The majority of applications of electrosurgery utilize monopolar methods, since that is the traditional technique and is most effective for cutting and excision. There are, however, many situations in which bipolar methods are preferred.

The power needed for a particular cutting or coagulating action depends on whether or not an arc is established at the scalpel electrode, on the volume of tissue, and on the type of electrosurgical action desired. Bipolar actions, which are typified by forceps electrodes grasping a volume of tissue to be coagulated, engage only a very small volume of tissue, since the current field is confined to be near the forceps, and usually white coagulation is desired. Consequently, the current is moderate to high and the voltage is low: tens to hundreds of milliamps (rms) at 40–100 V (rms), typically. In bipolar activations the current is determined primarily by the size of forceps electrodes used, and to a lesser extent by the volume of tissue. The volume of tissue more directly affects the time required to obtain adequate coagulation.

Monopolar actions, which may be for cutting or coagulation, are more variable and difficult to classify. In a study reported in 1973 in *Health Devices* (28), the voltages, currents, resistances, powers, and durations of activation during various monopolar electrosurgical procedures were measured. The resistance measured in this study was the total resistance between the active electrode cable and the dispersive electrode cable. Two types of procedure have significantly different electrical statistics compared to all other surgical cases: laparoscopic tubal ligations and transurethral resections of the prostate. The data in Table 1 have been grouped into general surgery (hernia repair, laparotomies, cholesystectomies, craniotomies etc.), laparoscopic tubal ligations, and TURPs. The table has been assembled from data collected at several different hospitals during procedures performed by different surgeons. For each separate surgical case, the minimum, average and maximum of each variable was recorded. The data in the table represent the means and standard deviations of the recorded minimum, average and maximum value calculated over the cases as originally presented in Ref. 19. While the total

Table 1. Typical Statistics for Representative Surgical Procedures in Which Electrosurgery Is Used[a]

	General Surgery (8 cases)	Laparoscopic Tubal Ligation (19 cases)	Transurethral Resection (8 cases)
Number of activations	22 (s.d = 24)	9 (7.5)	168 (151)
Voltage, V, rms			
Min	118 (58)	82 (20)	212 (43)
Avg	179 (57)	140 (65)	340 (12)
Max	267 (143)	207 (108)	399 (17)
Current, mA, rms			
Min	128 (29)	311 (108)	304 (109)
Avg	243 (116)	423 (73)	600 (30)
Max	423 (240)	615 (202)	786 (47)
Power, W			
Min	18 (7.5)	28 (9.3)	86 (41)
Avg	43 (25)	57 (19)	208 (15)
Max	103 (94)	99 (44)	290 (17)
Resistance, Ω			
Min	620 (720)	200 (50)	400 (40)
Avg	1070 (760)	410 (430)	580 (14)
Max	1960 (890)	660(670)	1110 (360)
Duration of Activation, s			
Min	1	1	1
Avg	1.6 (1.0)	4.8 (3.3)	1.1 (0.12)
Max	3.6 (3.0)	13.6 (13.1)	2.9 (1.9)

[a]Data collected by ECRI and reported in *Health Devices*(28). Data given include the means of each variable and its associated standard deviation in parentheses. The raw data were given as the minimum, average and maximum value during a particular procedure. The mean is the mean of the recorded minimum, average, or maximum value over all procedures in the study. This table originally appeared in Ref. 19.

number of cases studied under each category is not large, the data do give an overall indication of the range expected in surgical cases.

On the average, laparoscopic tubal ligations required the fewest activations (the range was 5–29), and TURPs by far the most activations (the range was 70–469). The resistances presented by the series combination of the scalpel electrode and tissue were similar for all procedures at 410 Ω for laparoscopic tubal ligations (range 130–1080 Ω), 580 Ω for TURPs (range 340–1800 Ω) and 1070 Ω for general surgery (range 180–2650 Ω). Higher equivalent resistances correlate with arc formation during the cutting process, so the values associated with general surgery might reasonably be expected to be higher.

Ablation, Coagulation and Tissue Fusion

The electrosurgical unit is designed to create irreversible thermal alteration of tissues; that is, controlled thermal damage. The objective is to heat target tissues to temperatures for times sufficient to yield the desired result. All of the physical effects of rf current are the result of elevated temperatures. The key observation is that the degree of alteration depends on both the temperature and the time of exposure. This section describes tissue effects resulting from the rf current from lower to higher temperature ranges.

Kinetic models of thermal damage processes based on an Arrhenius formulation have been used for many years to describe and quantify thermal damage (29):

$$\Omega(\tau) = \int_0^\tau A e^{-[E/RT]} dt \tag{1}$$

The dimensionless parameter, Ω is an indicator of the relative severity of the thermal damage. Thermal damage is a unimolecular reaction in which tissue proteins change irreversibly from their native ordered state to an altered damage state. In the kinetic model, A is a measure of the molecular collision frequency (s^{-1}), E is an energy barrier the molecules surmount in order to transform from native state to denatured state ($J \cdot mol^{-1}$), R is the universal gas constant ($J \cdot mol^{-1}$-K), T is the absolute temperature (K), and t is the time (s). The damage process coefficients, A and E, must be experimentally determined. This model assumes that only a single first-order process is active: The model can be used on multiple-process damage accumulation if each process is thermodynamically independent with its own set of process coefficients. A damage process may be described by its critical temperature, T_{crit}, defined as the temperature at which $d\Omega/dt = 1.0$.

The physical significance of Ω is that it is the logarithm of the ratio of the initial concentration of undamaged material, $C(0)$, to the remaining undamaged material at the conclusion, $C(\tau)$:

$$\Omega(\tau) = \ln\left\{\frac{C(0)}{C(\tau)}\right\} \tag{2}$$

However, typical damage end points have been qualitative tissue indicators such as edema formation or hyalinization of collagen (i.e., amorphous collagen as opposed to the normal regular fibrous array). One exception that has proved useful is the birefringent properties of some tissues, primarily muscle and collagenous structures. Birefringent tissue acts similarly to a quarter wave transformer in that polarized light has its polarization rotated as it passes through a regular array. Consequently, when observed through an analyzer filter rotated 90° with respect to the polarizer, birefringent tissue appears bright and nonbirefringent tissue dark (Fig. 10). In muscle, the birefringent properties are due to the regularity and spacing of the actin-myosin array. In collagenous connective tissues, it is the array of collagen fibers that determines birefringence. In both cases, elevated temperatures destroy the regularity of the array and birefringence is lost. In numerical models the transient temperature history of each point may be used along with equations 1 and 2 to predict the extent of thermal damage for comparison to histologic sections.

Ablation. Electrosurgical (RF) current, lasers, ultrasound, and microwaves have been used to obtain coagulative necrosis in myocardium *In vivo* for the elimination of ectopic contractile foci. The first meaning for ablation is to remove, as by surgery. The second meaning, to wear away, melt or vaporize is more familiar in the engineering sense (as in ablative heat transfer). In cardiac ablation, the goal

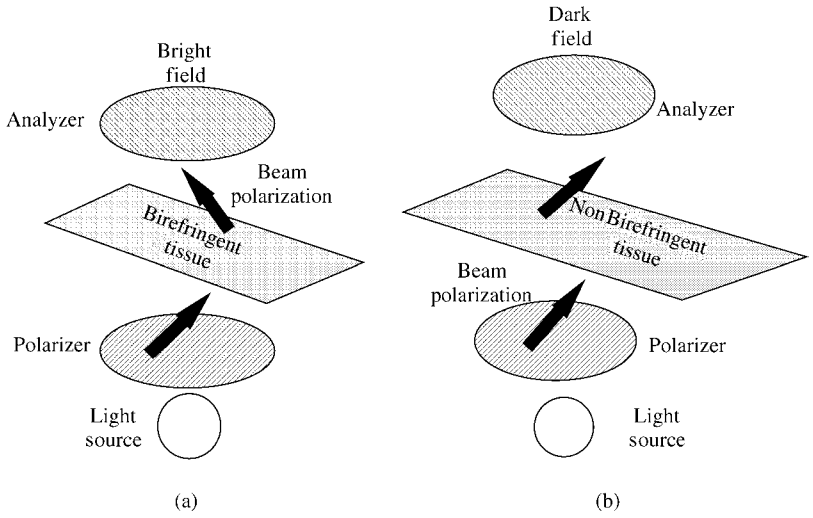


Figure 10. Principle of tissue birefringence. (a) Birefringent tissue is able to rotate the polarization (large arrow) of polarized light. (b) Thermally damaged tissue loses this property.

is to deactivate badly behaved cardiac muscle (reentrant pathways). Though no actual mass defects are created in the tissue structure, the process is termed "ablation" in the sense that the affected tissue is removed from the inventory of electrophysiologically active cardiac muscle. Results are evaluated in clinical use by monitoring electrophysiologic activity during treatment. Additional feedback may be applied to improve the repeatability of results.

Figure 11 illustrates the loss of birefringence at elevated temperatures in cardiac muscle. Figure 11a shows a circular disk active electrode applied to the epicardial surface of excised myocardium. The ground plane is applied to the endocardial surface. Figure 11b is a light microscopic (LM) view of the histologic section at an original magnification of 10× stained with hematoxylin and eosin. Figure 11c is the corresponding Transmission Polarizing Microscopy (TPM) view of the same section. The views are taken from the region shown in dashes at the outer edge of the disk electrode. While the boundary of thermal damage can just be identified by a skilled observer in the LM image (Fig. 11b), a clear line of demarcation is visible in the TPM image (Fig. 11c). For heating times in the range of 1–2 min useful estimates of the kinetic coefficients are $E = 1.45 \times 10^5$ ($J \cdot mol^{-1}$) and $A = 12.8 \times 10^{21}$ (s^{-1}). These coefficients give erroneous predictions for heating times outside of this range, however.

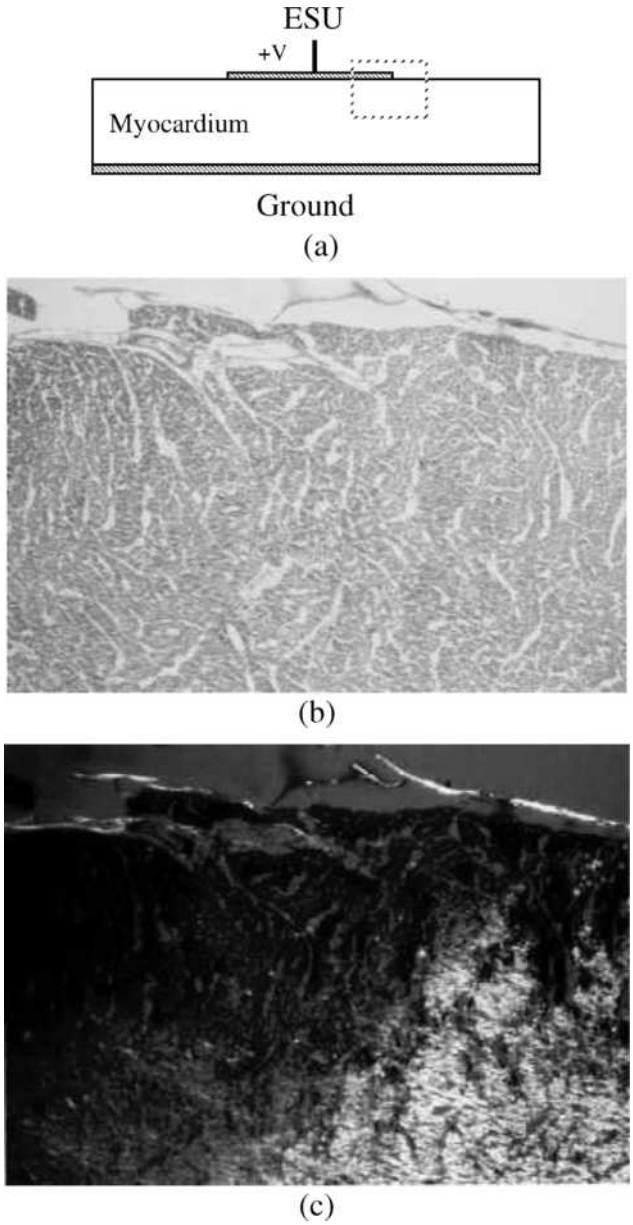


Figure 11. Disk electrode applied to myocardium. (a) Geometry. (b) Light microscopic view of region in dashed rectangle at the edge of the electrode. Original magnification 40×. (c) Transmission Polarizing Microscopy (TPM) view of the same section showing the clear delineation of the zone of birefringence loss. Original magnification 40×.

Birfringence loss identifies irreversible major structural damage in the cardiac myocyte. It is certainly arguable that electrophysiologic function is probably lost at lower temperatures than the 50+ °C required to achieve the result shown in the figure. However, clinically, the 50 °C isotherm seems to correspond with the desired result (30), and also with the birefringence loss boundary for heating times in this range (31).

Tissue Fusion and Vessel Sealing. Irreversible thermal alteration of collagen is apparently the dominant process in coagulation and successful tissue fusion (32–34). Electron microscopic (EM) studies suggest that the end-to-end fusion process in blood vessels is dominated by random reentwinement of thermally dissociated adventitial collagen fibrils (Type I) during the end stage heating and early cooling phases (34). Successful vessel seals depend on the fusion of intimal surface tissues and the support provided by thermal alteration of medial and adventitial tissues.

Collagen is a ubiquitous tissue structural protein consisting of three left-hand α-helices wound in a rope-like molecular form (Fig. 12) with a periodicity of ∼68 nm (35). In this figure a small segment of a typical 300 nm long molecule is shown. The 300 nm molecules spontaneously organize into quarter-staggered microfibrils 20–40 nm in diameter and these coalesce into larger diameter collagen fibers *In situ*. There are at least 13 different types of collagen fibers that form the structural scaffolding for all tissues, the most common are Types I, II, and III collagen. Collagen *In situ* is birefringent. One useful measure of irreversible thermal alteration in collagen is that when heated for sufficient time to temperatures in excess of ∼60 °C the regularity of the fiber array is disrupted and collagen loses its birefringent property (see Fig. 13). When viewed through the analyzer, native state collagen shows up as a bright field due to its birefringent properties. Thermally damaged collagen loses this property and is dark in the field. The kinetic coefficients for collagen birefringence loss in rat skin are (36): $A = 1.606 \times 10^{45}$ (s^{-1}) and $E = 3.06 \times 10^5$ ($J \cdot mol^{-1}$). The birefringence-loss damage process in collagen has a critical temperature of 80 °C. These coefficients have proven useful over a wide range of heating times from milliseconds to hours.

Collagen shrinks in length as well as losing its organized regular rope-like arrangement of fibers. A model for

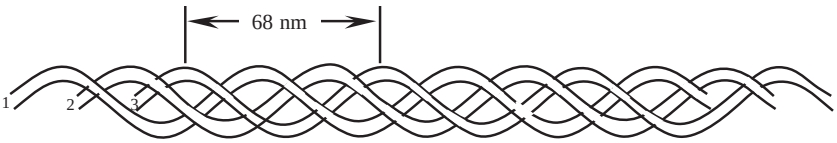


Figure 12. Sketch of periodic structure of the basic collagen molecule.

collagen shrinkage obtained by Chen et al. (37,38) is a bit different in style from the first-order kinetic model. They measured shrinkage in rat *chordae tendonae* over several orders of magnitude in time for temperatures between 65 and 90 °C and under different applied stresses. They were able to collapse all of their measured data into a single curve, sketched approximately in Fig. 14. In their experiments an initial "slow" shrinkage process (for equivalent exposure time $< \tau_1$) is followed by a rapid shrinkage rate ($\tau_1 < t < \tau_2$) and a final slow shrinkage process. The practical maximum for shrinkage in length, ξ (%), is 60%. Longer equivalent exposures result in gellification of the collagen and complete loss of structural properties. After initial shrinkage, the collagen partially relaxes during cooling, indicated by the shrinkage decay region in Fig. 14. The curve fit functions utilize a nondimensional time axis, t/τ_2, where the fit parameters are expressed in the form of the logarithm of the time ratio:

$$\nu = \ln\left\{\frac{t}{\tau_2}\right\} \tag{3}$$

The shrinkage is obtained by interpolation between the two slow region curves (through the fast region):

$$\xi = (1 - f(\nu))[a_0 + a_1\nu] + f(\nu)[b_0 + b_1\nu] \tag{4}$$

where $a_0 = 1.80 \pm 2.25$; $a_1 = 0.983 \pm 0.937$; $b_0 = 42.4 \pm 2.94$; and $b_1 = 3.17 \pm 0.47$ (all in %). The best-fit interpolation function, *f(ν)*, is given by

$$f(\nu) = \frac{e^{\alpha(\nu - \nu_m)}}{1 + e^{\alpha(\nu - \nu_m)}} \tag{5}$$

where $\alpha = 2.48 \pm 0.438$, and $\nu_m = \ln\{\tau_1/\tau_2\} = -0.77 \pm 0.26$. Finally, at any temperature τ_2 is given by

$$\tau_2 = e^{[\alpha + \beta P + M/T]} \tag{6}$$

where $\alpha = -152.35$; $\beta = 0.0109$ (kPa^{-1}); $P =$ applied stress (kPa); and $M = 53{,}256$ (K).

The functional form of τ_2 contains the kinetic nature of the process, but is in the form of an exposure time rather than a rate of formation, as was used in Eq. 2, and so the coefficient, M, is positive. To use the collagen shrinkage model, the shrinkage is referred to an equivalent τ_2. That is, at each point in space and time an equivalent value for the increment in t/τ_2 is calculated and accumulated until shrinkage is calculated.

A representative clinical application of collagen shrinkage is the correction of hyperopia using rf current, termed Conductive Keratoplasty. In this procedure a small needle electrode is inserted into the cornea to a depth of just over one-half of the thickness (see Fig. 15a). A teflon shoulder controls the depth of insertion. The speculum used to retract the eyelids comprises the return electrode in this procedure (Fig. 15b). The RF lesions are placed at 45° increments on circumferences with diameters of 6, 7, or 8 mm—8, 16, or 24 lesions depending on the degree of curvature correction (i.e., diopter change) required (Fig. 15c). Pulsed RF current heats and shrinks the corneal collagen to decrease the circumference and thus the radius of curvature of the cornea. Figure 16 is a histologic cross-section of a typical lesion seen a few minutes after its creation. Figure 16a is a light microscopic section (hematoxylin and eosin stain) and Fig. 16b is the corresponding transmission polarizing microscopy section showing the loss of collagenous birefringence near the electrode path. The effect of shrinkage on the collagen fibers is clearly visible as the normal fibrous wave is stretched in the vicinity of the electrode site (Fig. 16b).

A representative example of tissue fusion processes is the sealing of blood vessels by fusion of the intimal surfaces. In this application, forceps electrodes grasp the vessel and a bipolar rf current is used to heat and fuse the tissue (Fig. 17). An example experimental result is shown in Fig. 18, where a successful seal was obtained *In vivo* in a femoral artery. In that experiment a thermocouple was advanced through the vessel to the site of the electrodes, accounting for the hole in the cross-section.

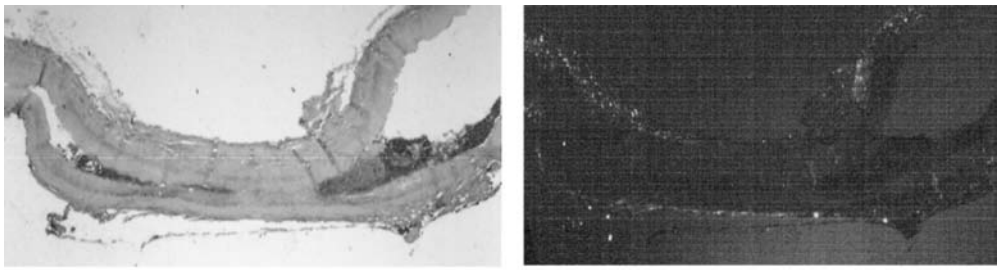


Figure 13. Vessel collagen birefringence loss. (a) Thermally fused canine carotid artery, H&E stain (Original magnification 40×). (b) TPM view of same section showing loss of birefringence in adventitial collagen under bipolar plate electrodes.

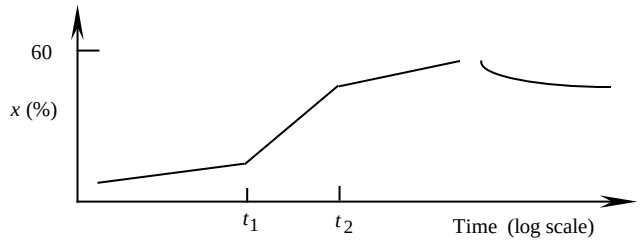


Figure 14. Collagen shrinkage model curve (38). [Reproduced with permission from "Corneal reshaping by radio frequency current: numerical model studies" Proc. SPIE v4247 (2001) pp 109–118.]

Sealing and fusion of vessels by electrosurgical current is strongly influenced by the inhomogeneous architecture of the tissue constituents, particularly in the large arteries. Inhomogeneities in electrical properties of the constituents, specifically smooth muscle, collagen and elastin, lead to sharp spatial gradients in volumetric power deposition that result in uneven heating. The mechanical properties of the various tissue constituents are also of considerable importance. Vascular collagen and elastin distribution varies from vessel to vessel, species to species in the same vessel, and point to point in the same vessel of the same species.

Cutting Processes

The essential mechanism of cutting is the vaporization of water. Water is by far the most thermodynamically active tissue constituent. Its phase transition temperature near 100 °C (depending on local pressure) is low enough that vaporization in tissue exerts tremendous forces on the structure and underlying scaffolding. The ratio of liquid to vapor density at atmospheric pressure is such that the volume is multiplied by a factor of $\sim$1300 when liquid water vaporizes. Expansion of captured water is capable of disrupting tissue structure and creating an incision. The goal in electrosurgical cutting is to vaporize the water in a very small tissue volume so quickly that the tissue structure bursts before there is sufficient time for heat transfer to thermally damage (coagulate) surrounding tissues (39). The same strategy is used in laser cutting.

Cutting electrodes have high rates of curvature to obtain the high electric fields necessary to accomplish cutting. The electrodes are not generally sharp, in a tactile sense, but the edge electrical boundary conditions are such that extremely strong electric fields result. With the possible exception of needle electrodes, the scalpel electrodes of Figs. 1 and 2 are not capable of mechanically damaging tissues. The intense electric fields vaporize tissue water

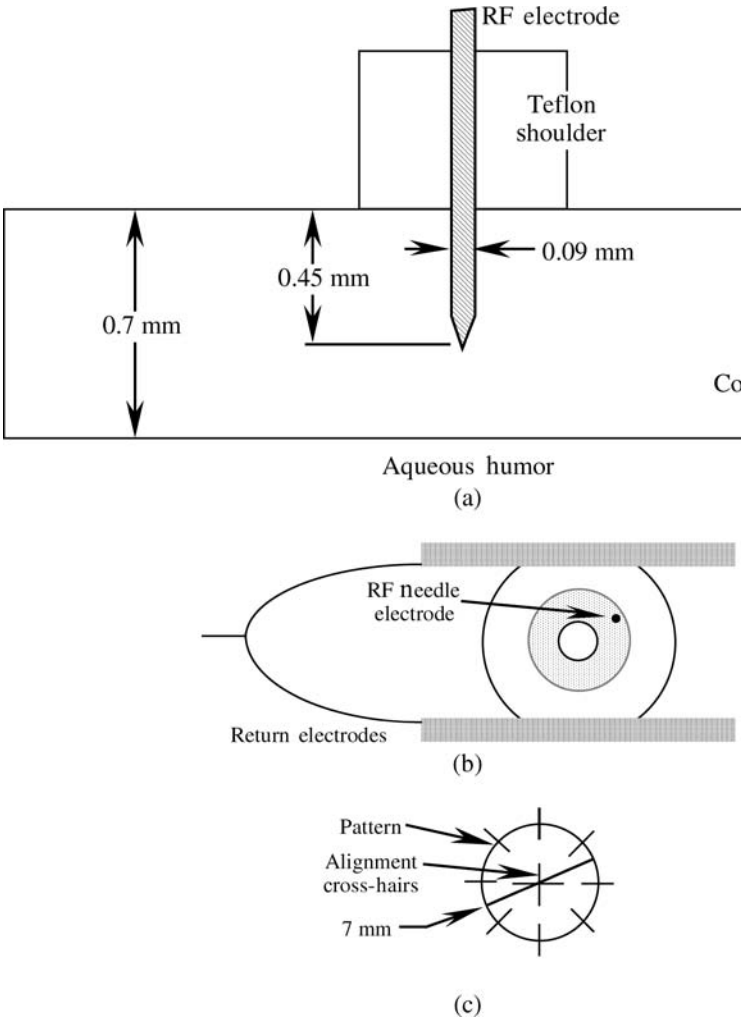


Figure 15. Needle electrode used to shrink corneal collagen and change curvature for correction of hyperopia. (a) Cross-section of electrode in place, (b) View of speculum return electrodes, (c) spot pattern for shrinkage lesion location. [Reproduced with permission from "Corneal reshaping by radio frequency current: numerical model studies" Proc. SPIE v4247 (2001) pp109–118.]

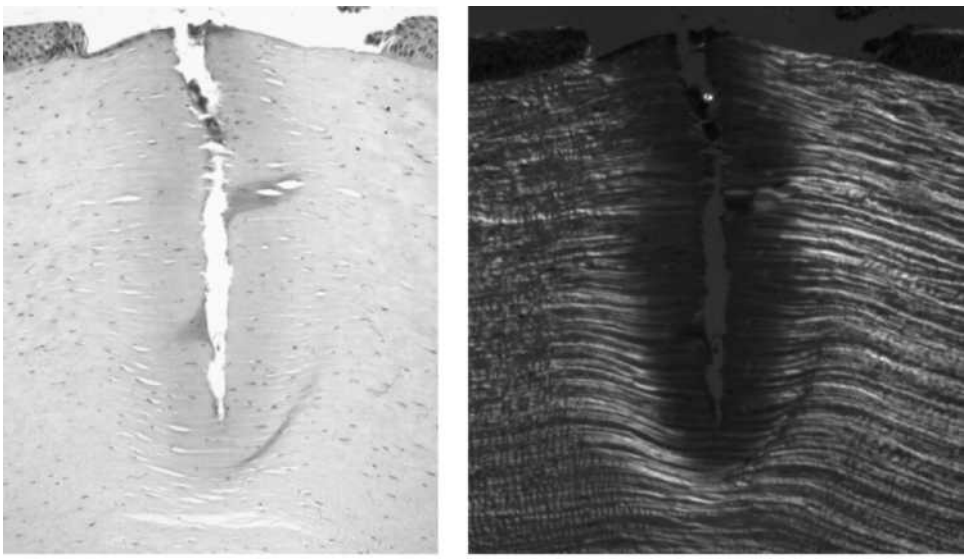


Figure 16. Histologic view of collagen shrinkage for Conductive Keratoplasty. (a) Light microscopic view, hematoxylin, and eosin stain. (a) The TPM view, same section. Original magnification 40×.

ahead of the electrode, and the tissue structure parts due to internal tension to allow the electrode to pass essentially without tissue contact. Certainly, if the surgeon advances the electrode too quickly there is a dragging sensation due to perceptible friction, and much more radiating thermal damage around the incision site.

Good cutting technique requires matching the generator output characteristics (primarily the source impedance, but also to some extent the fundamental frequency), open circuit voltage setting, electrode shape and cutting speed to optimize the surgical result. Continuous sine waves are more effective for cutting than interrupted waveforms since the time between vaporization episodes is so much shorter that radiating heat transfer is virtually eliminated. Well-hydrated compartmentalized tissues, like skeletal muscle or gingiva, cut more cleanly than drier tissues, such as skin. Higher fundamental frequencies seem to give cleaner cuts than lower frequencies, but the reason for this is not known.

ADVANCED PRINCIPLES OF ELECTROSURGERY

Hazards of Use and Remedies

Any energy source, including electrosurgery, has hazards associated with its use. The goal of the user is to achieve safe use of the device by minimizing the risk. While it is not possible to completely eliminate hazards, it is possible by careful technique to reduce them to acceptable levels. Several of the common hazards associated with electrosurgery are alternate site burns, explosions, stimulation of excitable tissues, and interference with monitoring devices and pacemakers. The RF currents are designed to thermally damage tissues, so the possibility of alternate site burns (at sites other than the surgical site) is always present. Explosions of combustible gases were, and still are, a hazard of electrosurgery. Combustible gases include at least two relatively little-used anesthetic agents (ether and cyclopropane) and bowel gas, which has both hydrogen and methane, as a result of bacterial metabolism. Arcs and/or sparks are routinely generated when cutting tissue so there is always an ignition source for a combustible mixture. While the fundamental frequency of electrosurgery is above the threshold for stimulation, arcs or sparks generate low frequency components that can stimulate electrically excitable tissues, that is, muscle and nerve. Radiated rf in the operating room and rf currents in the tissues posed few problems in early days, unless a faulty ground or other machine fault led to a burn at some alternate site. In the present day, electrosurgery raises considerable havoc with instrumentation amplifiers, pacemakers, and other electronic measurement

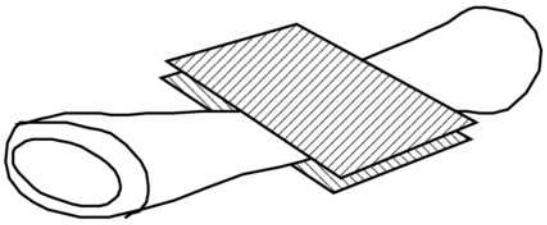


Figure 17. Bipolar forceps electrodes for vessel sealing.

Figure 18. Histologic cross-section of a vessel sealing experiment in the canine femoral artery. Successful seal obtained with a center temperature of 89 °C for 2 s. Hole in center of section is location of thermocouple used to monitor temperature. Original magnification 40×.

devices. Because so many potentially current-carrying objects and devices are now routinely attached to the patient, the attendant hazards have increased.

The remarkable feature of electrosurgical accidents is that, although they are rare (one estimate puts the probability at $\sim 0.0013\%$, or 200 out of 15 million procedures using electrosurgery, on an annual basis) they are usually a profound and traumatic event for the patient and the surgical team, and often cause for expensive litigation. The non-surgeon might reasonably wonder, in the light of the problems, why electrosurgery is used. The answer is that the tremendous advantages achieved by the use of electrosurgery (the remarkable reduction in morbidity and mortality and the low cost compared to competing technologies) make it a very effective technology for use in clinical medicine. It is important to note that hazards attend any energy source (and electrosurgery is one among several in the operating room) and they can be reduced and managed, but not eliminated. Clever application of appropriate technology can make considerable contributions in this area by reducing device interactions.

Alternate Site Burns

Electrosurgical units in general have very high open circuit output voltages that may appear at the scalpel electrode when it is not in contact with the tissue. When current is flowing the scalpel voltage is much reduced. Standard vacuum tube electrosurgery units may have open circuit voltages approaching 10,000 V peak-to-peak. This high open circuit output voltage is very effective in the initiation of fulguration techniques and in the spray coagulation of large tissue segments, which makes these units popular for certain procedures, especially in urologic surgery. However, the high voltages at the scalpel electrode also can initiate arcs to other objects, and must be handled with caution. This can be an especially difficult problem in minimally invasive surgery through an endoscope since the surgeon's field of view may be limited. The solid-state electrosurgery units usually have lower maximum output voltages (somewhere in the range of 1000–5000 V peak-to-peak depending on design) the exception being recent designs based on HEXFET or VMOS technology, which approach the open circuit voltage of vacuum tube devices. All electrosurgery units have output voltages that are potentially hazardous.

It is prudent to inspect all surgical cables periodically for damage to the insulation: especially those that are routinely steam sterilized. While in use, it is not a good idea to wrap the active cable around a towel clamp to stabilize it while accomplishing cutting or coagulation: The leakage current to the towel clamp will be concentrated at the tips of the towel clamp and may cause a burn. One should not energize the electrosurgical unit when the scalpel is not being used for cutting or coagulating. This is because when not using electrosurgery, the full open circuit voltage is applied between the active and return electrode cables, and one runs the risk of inadvertent tissue damage. Care should be taken to ensure that an unused scalpel electrode is not placed in a wet environment during surgery. The leakage current of any scalpel electrode may be increased by dampness. Additionally, some of the hand control designs can be activated by moisture at the handle. In one recent case, a hand control scalpel was placed on the drape in between activations. The scalpel was in a valley in the drape sheet in which irrigating fluid collected. The pooled fluid activated the electrosurgery machine and a fairly severe steam or hot water burn resulted on the patient's skin.

The high voltages and frequencies typical of electrosurgery make it essential to use apparatus with insulation of good integrity. It should also be kept in mind that the electric fields extend for some distance around cables in use at high voltage. An effort should be made not to have active and return electrode cables running parallel to each other for any distance in order to reduce the capacitive coupling between them, unless the apparatus is specifically deigned to be used this way, as in bipolar device cables.

There is an additional problem in monopolar endoscopic surgery that must be addressed. It arises when a metal laparoscope with an operating channel is isolated from a voltage reference plane, as depicted in Fig. 19. In this case, the insulating skin anchor isolates the metal laparoscope (or metallic trochar sleeve) so that its potential is determined by parasitic capacitances between the metallic elements. The resulting circuit creates a capacitive voltage divider, diagrammed in Fig. 20. The two parasitic capacitances are unavoidable. The voltage divide ratio for this

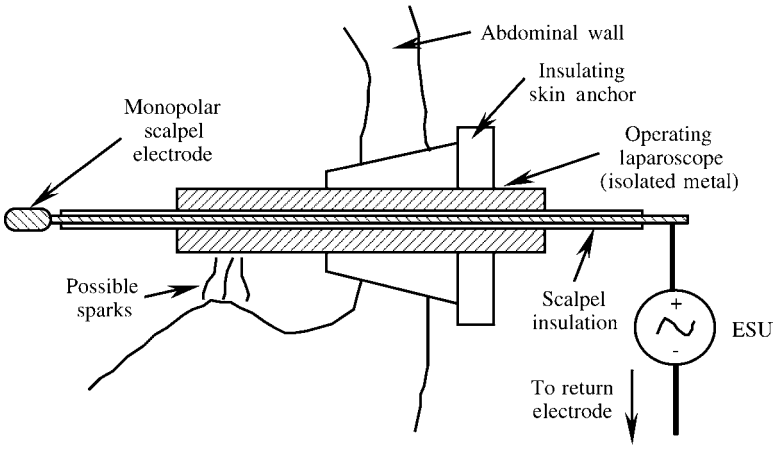


Figure 19. Diagram of isolated laparoscope with active operating channel.

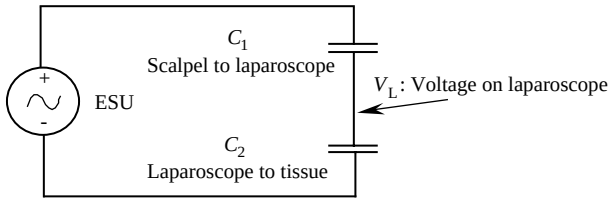


Figure 20. Equivalent capacitive voltage divider circuit for isolated laparoscope with active operating channel.

equivalent circuit is

$$V_L = \frac{C_1}{C_1 + C_2} \tag{7}$$

The parameter C_1 is always quite a bit larger than C_2, and the surface potential on the laparoscope, V_L, may be as much as 80–90% of the surgical voltage. This induced voltage is capable of generating sparks to the surrounding tissues out of the field of view of the surgeon. Several cases of distributed severe burns have resulted from situations similar to this. There are two remedies. First, the operating laparoscope may be grounded. This does not eliminate the C_1 capacitance, so there will be capacitively coupled currents between the scalpel electrode and the laparoscope. However, the laparoscope will be at zero potential. The second approach involves monitoring the scalpel electrode and return electrode currents. If they are substantially different then a hazard situation exists. There is at least one commercial device designed to do this.

Explosions. The original explosion hazard in the operating room arose from the use of explosive anesthetic agents, such as ether and cyclopropane. Combustible mixtures of these gases can be ignited by any arc discharge of sufficient energy—by static discharges between personnel and grounded objects, by arcs due to power switch openings in mains-powered devices, and by electrosurgical arcs at the cutting site. A few early instances of explosions and ensuing fires in operating rooms stimulated the establishment of requirements for operating room construction and use to lessen this hazard. In particular, operating rooms in the United States were required to have semiconducting flooring that was to be checked periodically for conductivity; operating room personnel were required to wear conductive shoe covers that would drain off accumulated static charge to the conductive flooring; and explosion-proof switches and plugs were required below the five foot level where the relatively heavy combustible gases might collect. Also, surgical table coverings and fixtures are connected to ground through moderate impedance in order to prevent static accumulation. These requirements greatly reduced the explosion risk in the operating room, but lead to some rather amusing *non sequiturs*. For example, it makes little sense to attach an explosion-proof foot switch to an electrosurgical generator (particularly to a spark gap oscillator) since arcing is inherent to the use of electrosurgery. Of course, since the wall plugs for mains power were required to be explosion proof, the power line plug on the generator was also often explosion proof. Because of the inordinate expense of explosion proof construction and the rare use of combustible gases, the majority of new operating room construction is designed to satisfy the requirements for oxygen use only and specifically designated not for use of explosive gases.

There are, however, other explosion–combustion sources. Two constituents of bowel gas are combustible, hydrogen and methane, and there is almost always sufficient swallowed air to comprise an explosive mixture. Approximately 33% of the adult population are significant methane producers due to the indigenous gut flora and fauna responsible for certain aspects of digestion. One preventive measure that has been used is insufflation of the bowel with carbon dioxide, or some other inert gas, during electrosurgery. A large volume of CO_2 will dilute and displace the combustible mixture, reducing the chance of explosion.

Additionally, care must be taken in pure oxygen environments not to ignite organic materials. Body hair has been ignited in a pure oxygen environment by an electrosurgical arc. In surgery on the larynx, both electrosurgery and CO_2 lasers have ignited plastic endotracheal tubes and surgical gloves in the oxygen-rich environment.

Stimulation of Electrically Excitable Tissues. The stimulation of motor nerves is a well-known phenomenon accompanying electrosurgery. Stimulation of the abdominal muscles is often observed, and the obturator nerve bundle (connecting the spinal column to the leg) is also quite often stimulated. Obturator nerve stimulation is inherent to transurethral resections, and the resulting movement is always undesirable. Part of the stimulation problem arises from the low frequency components of the surgical arc, and part from the confinement of current. For example, the obturator nerve is clustered with the major blood vessels supplying the leg (iliac/femoral artery and vein) and emerges from the pelvis under the inguinal ligament into the upper thigh area. The pelvic bone structures are essentially nonconductive and impede current flow. The nerve–artery–vein cluster represents a relatively high conductivity pathway of small cross-section between the abdomen and the upper thigh, a typical location for the return electrode. As such, the surgical current in TURP procedures is somewhat concentrated in that tissue cluster, and stimulation is often observed. In the case of general abdominal surgery, the majority of surgical current flows in the surface tissue structures, reducing the likelihood of obturator stimulation. In the case of transurethral resections, the current source is deep in the floor of the abdomen, and the surface tissues carry a smaller fraction of the total surgical current. The very high currents in TURPs and the frequent use of spark gap oscillators means that an abundant supply of intense low and high frequency signals is contained in the surgical current. It is a personal observation that motor unit stimulation is closely correlated with relatively intense arc strikes while accomplishing cutting or coagulating. This is to be expected from the frequency spectrum of the surgical signals.

There are, in addition, several reported instances of ventricular fibrillation induced by electrosurgery (40–43). The low frequency arc components and high intensity high frequency signals are capable of stimulating

excitable tissues, with ventricular fibrillation as a possible outcome.

Interference with Instrumentation. The cutting current generated by standard oscillators, such as the vacuum tube and solid-state devices, has a well defined, narrow bandwidth. When an arc is struck at the surgical site, extensive signal energy is added at the high frequency end and at the low frequency end of the spectrum. Signals from spark gap oscillators have these components whether or not an arc is established at the tissue. Over the range of physiological frequencies the arc components are of very low amplitude compared to the generator fundamental frequency; but since the surgical currents and voltages are very high, the arc components may be many times larger than physiological signals such as the ECG. This creates a considerable problem for all measuring instruments including demand, or noncompetitive, pacemakers. There are many reported instances of inhibition of demand pacemakers by electrosurgery (44–47).

The population most at risk for pacemaker inhibition is patients undergoing either open heart surgery or transurethral resections of the prostate. Open heart operations put the electrosurgical scalpel electrode in close proximity to the pacemaker electrodes. The TURPs use the highest currents and voltages, and the interference signals are potentially large at the pacemaker electrodes. There is some indication that a higher incidence of prostate problems is associated with long-term pacemaker implantation, so it is to be expected that this problem will continue. The present use of shielded pacemakers virtually eliminates microwave sensitivity, but not electrosurgical effects.

All pacemakers may be damaged to failure by electrosurgery signals, and some caution should be observed when performing electrosurgery in the presence of pacemakers. Bipolar electrosurgery is much less likely than monopolar to cause interference with pacemakers because the surgical current field is confined and the voltages are low. There are many techniques for which bipolar electrosurgery is unacceptable, however, so monopolar methods must be used. In those cases, care should be exercised to encourage the surgical current field to avoid the right ventricle. To the extent that is practical, elevate the pacing electrode site (either the right ventricular wall or the external surface) away from the conductive fluids and tissues when cutting near the heart. Place the dispersive electrode on the left side of the body (this will only help a little bit). Anticipate some form of pacemaker malfunction.

Representative Electric Field Distributions

The effect of rf current on tissues is to heat them according to the first law of thermodynamics:

$$\rho_t c_t \frac{\partial T}{\partial t} = \nabla \bullet (k \nabla T) + q_{gen} + q_m + w_t c_b (T_a - T) - h_{fg} \frac{\partial m}{\partial t} \tag{8}$$

where ρ = density ($kg \cdot m^{-3}$), c = specific heat ($J \cdot kg^{-1} \cdot K^{-1}$), T = temperature (K), t = time (s), k = thermal conductivity ($W \cdot m^{-1} \cdot K^{-1}$), q_{gen} = externally applied rf power ($W \cdot m^{-3}$), q_m = metabolic heat ($W \cdot m^{-3}$), w = perfusion ($kg_b \cdot kg_t^{-1} \cdot s^{-1}$), h_{fg} = vaporization enthalpy ($J \cdot kg^{-1}$) and dm/dt is the mass rate of vaporization ($kg \cdot s^{-1}$). Each term in the relation has units of ($W \cdot m^{-3}$) and the t subscript refers to tissue, while the b subscript refers to blood properties. This form does not include vaporization processes typical in cutting. Typically, both metabolic heat and perfusion heat transfer are negligible in electrosurgery. It is important to note that all of the tissue properties are dependent on the water content. As tissue dries out the density, electrical, and thermal conductivity decrease. As temperature rises the electrical conductivity increases $\sim$1–2%/°C, until drying at higher temperatures causes a decrease.

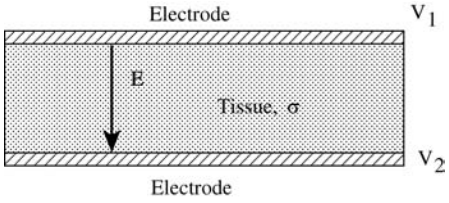


Figure 21. Geometry for infinite flat plate electrode calculation.

At electrosurgical frequencies the quasistatic assumption applies (simplifying calculations) and heat dissipation is dominated by joule, or resistive, heating ($W \cdot m^{-3}$):

$$q_{gen} = \mathbf{E} \cdot \mathbf{J} = \sigma |\boldsymbol{E}|^2 = \frac{|\boldsymbol{J}|^2}{\sigma} \tag{9}$$

where $\mathbf{E}$ = electric field vector ($V \cdot m^{-1}$), $\mathbf{J}$ = current density vector ($A \cdot m^{-2}$), and σ = electrical conductivity ($S \cdot m^{-1}$). This section reviews several representative electric field distributions.

Simple Field Geometries: Analytical Solutions. The simplest case is that of the electric field between parallel plates (Fig. 21). For a uniform medium between the plates the electric field is uniform (i.e., V increases linearly from V_2 to V_1) and is equal to the voltage difference divided by the distance between the plates: $(V_1 - V_2)/d = \Delta V/d$. It points from the higher potential (V_1) to the lower potential (V_2). In this case the heating term, q_{gen}, is also simple: $q_{gen} = \sigma \Delta V^2 / d^2$. For an operating voltage difference of 50 V(rms) and plate separation distance of 2 cm, the electric field is 2500 ($V \cdot m^{-1}$), and if the electrical conductivity is 0.3 ($S \cdot m^{-1}$) the power density would be 1.88 $W \cdot cm^{-3}$.

Another simple analytical solution is for the case of coaxial cylinders (Fig. 22). In this geometry the potential increases as the natural log of the radius from V_2 to V_1, and the electric field (pointing in the radial direction) is given by

$$\boldsymbol{E} = \frac{I_L}{2\pi\sigma r} \quad a_r \tag{10}$$

where I_L = the current per unit length of coaxial structure (i.e., length into the page, $A \cdot m^{-1}$). For this geometry the rf power generation decreases as $1/r^2$:

$$q_{gen} = \mathbf{E} \bullet \mathbf{J} = \sigma \left[\frac{I_L}{2\pi\sigma} \right]^2 \quad \frac{1}{r^2} \tag{11}$$

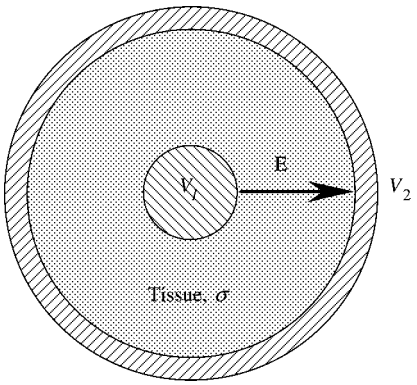


Figure 22. Geometry for coaxial cylinder electrode example. The outer cylinder typically comprises the return electrode, $V_2 = 0$.

For a center cylinder electrode 2 mm in diameter and outer cylinder 2 cm in diameter and tissue electrical conductivity 0.3 ($S \cdot m^{-1}$), the overall conductance per unit length would be $G = 0.816$ ($S \cdot m^{-1}$)— that is, $G = 2\sigma/(\ln\{b/a\})$. At a voltage difference of 50 V(rms), the current per unit length would be $I_L = 40.9$ ($A \cdot m^{-1}$) and maximum electric field 21.7 ($kV \cdot m^{-1}$) at $r = 1$ mm. The maximum volume power density would be 142 $W \cdot cm^{-3}$ at $r = 1$ mm in this example.

These examples are useful in understanding simple surgical fields. The uniform electric field case is not very different from that obtained near the center of a bipolar forceps electrode used to coagulate a tissue volume. The coaxial case is close to the electric field around a needle electrode as long as one is not too close to the tip. In both cases the analytical field expressions are simple because the boundaries are parallel to Cartesian and cylindrical coordinates, respectively. Also, the boundaries are chiefly isopotential, or Dirichlet, boundaries.

Disk Electrode Field. The electric field around a disk electrode applied to a large volume of tissue is a much more complex geometry (Fig. 23). This geometry was solved analytically by Wiley and Webster (48) for the $V = 0$ reference electrode located at $r = +\infty$ and $z = -\infty$. The electrode is an isopotential surface ($\partial V/\partial r = 0$) and the air is insulating (zero flux, $\partial V/\partial z = 0$); and at the edge of the disk ($r = a$) the field must satisfy two mutually exclusive boundary conditions. The consequence is that the electric field and current density both increase without bound at

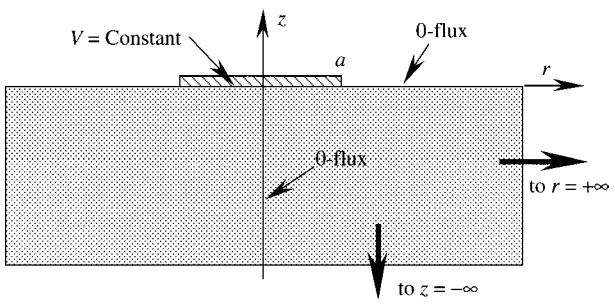


Figure 23. Solution geometry for disk electrode applied to infinite medium. Return electrode, $V = 0$, located at $r = \infty$ and $z = -\infty$.

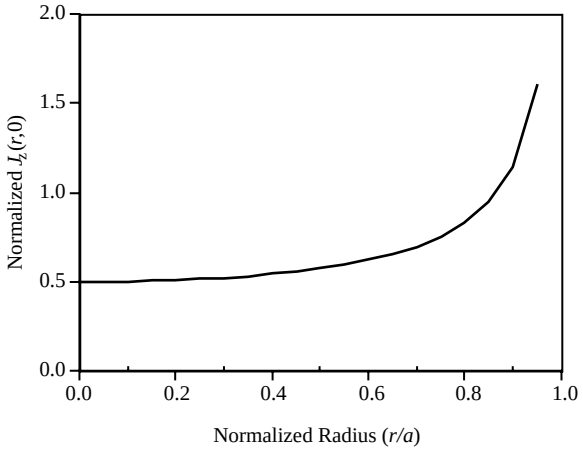


Figure 24. Normalized surface current density under the electrode from $r = 0$–$0.95a$.

the electrode edge, $r = a$ (48):

$$J_z(r,0) = \frac{J_0}{2\left[1 - (r/a)^2\right]^{1/2}} \tag{12}$$

where $J_0 = I/\pi r^2$, the average current density. Figure 24 plots the normalized surface current ($J_0 = 1$) density versus normalized radius under the electrode. For radii less than $\sim 0.89a$ the current density is less than the average value. This is at first glance an alarming result: It is not obvious why burns do not result at the edge of every electrode. There are three mitigating factors. First, the singularity at $r = a$ is an integrable singularity—meaning that one may integrate $J_z(r,0)$ from $r = 0$ to $r = a$ and get a finite result (i.e., integral $\{J_z\, dA\} = I$, the total current). Second, actual electrodes do not have mathematically sharp edges, but rather are filleted. Third, heat transfer dissipates the temperature rise at the edge and even though q_{gen} increases without bound (in a microscopic sense) the temperature rise at the edge reaches a finite maximum—still much higher than that at the center of the electrode, however.

The electric fields near a disk electrode, such as that used to ablate myocardium (Fig. 11) is very similar to the analytical solution. Figure 25 is the result of a quasi static Finite Element Method (FEM) calculation for the geometry used to create Fig. 11. The assumed electrode potential is 50 V(rms), and the isopotential lines (25a) are not significantly different in shape from those of the analytical solution (48). The electric field (current density lines) lines (Fig. 25b) are very crowded at the edge of the electrode, as expected from the above discussion. The result is that the volume power density, q_{gen}, (Fig. 25c) is higher by a factor of 1000 or more (see the 12 $W \cdot cm^{-3}$ contour) than about one electrode radius away (cf. to the 0.062 $W \cdot cm^{-3}$ contour).

Needle Electrode Field. Needle electrodes are often used to desiccate or coagulate small lesions. They are also used to shrink corneal collagen to change its curvature, as was discussed above (in the section Ablation, Coagulation and Tissue Fusion). The electric field around the tip of such an electrode is extremely high compared to the average value.

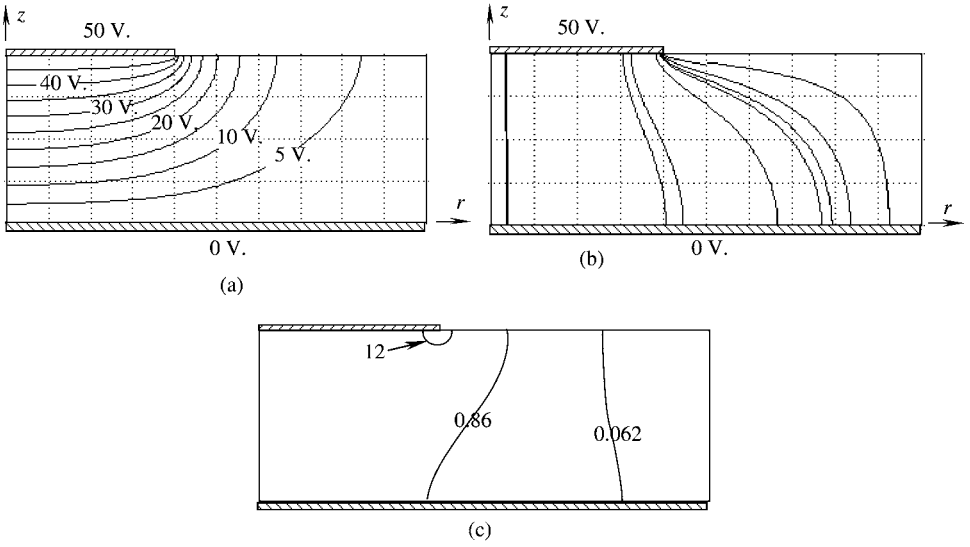


Figure 25. Circular disk electrode FEM model results. Muscle conductivity $90.3\,S{\cdot}m^{-1}$. (a) Potential field, *V(r,z)*. (b) Electric field streamlines. (c) Volume power density with contours at 0.062, 0.86, and $12\,W{\cdot}cm^{-3}$.

For comparison, the electric field and corresponding volume power density field were calculated around a needle electrode in cornea: Cornea electrical conductivity, at $1.125\,S{\cdot}m^{-1}$, is quite a bit higher that the $0.3\,S{\cdot}m^{-1}$ assumed for muscle in previous calculations. An electrical conductivity of $1.5\,S{\cdot}m^{-1}$ was assumed for the aqueous humor, accounting for the discontinuity in the constant power density contours at the interface. The results are shown in Fig. 26 for an assumed electrode potential of 50 V(rms).

Note that the volume power densities are extremely high compared to the disk electrode calculation due to the small dimensions of the needle electrode. Power densities of 100 $kW{\cdot}cm^{-3}$ have adiabatic (no heat transfer) heating rates near 27,000 $^{\circ}C{\cdot}s^{-1}$ and will vaporize significant amounts of water in microseconds. In practice, the applied voltage is in the neighborhood of 50–80 V (rms), however, pulsed rf is used with pulse times in the neighborhood of 15 μs. Vaporization begins during the first rf pulses in a very small volume around the electrode tip. Subsequent drying near the tip causes the dominant heating field to advance from the tip toward the surface while the overall impedance decreases due to an average increase in temperature around the electrode. At the higher voltage settings the heating may be sufficient to cause the impedance to increase near the end of rf activation due to drying in the cornea (49).

Bipolar Forceps Electrode Field. The bipolar forceps electrodes typically used in vessel sealing are essentially flat plate electrodes. The electric field between them is nearly uniform, with some small electrode edge effect. Figure 27 is a representative FEM calculation assuming an applied potential of 50 V(rms) and electrode separation of 2 mm. Here the power densities are in the neighborhood of 150–200 $W{\cdot}cm^{-3}$. The calculations have again assumed a uniform electrical conductivity of $0.3\,S{\cdot}m^{-1}$ while the actual value is a complex combination of collagen ($\sigma = 0.26$–$0.42\,S{\cdot}m^{-1}$ depending on orientation), elastin ($\sigma = 0.67$–$1.0\,S{\cdot}m^{-1}$ depending on orientation) and smooth muscle ($\sigma = 0.3\,S{\cdot}m^{-1}$).

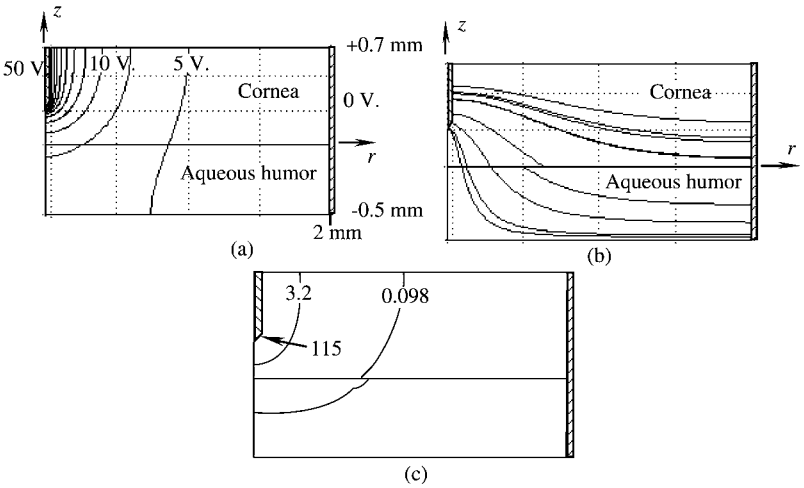


Figure 26. Needle electrode in cornea, FEM model results. Cornea conductivity $1.125\,S{\cdot}m^{-1}$, aqueous humor $1.5\,S{\cdot}m^{-1}$. (a) Potential field *V(r,z)* for 50 V (rms) electrode potential. (b) Electric field streamlines. (c) Volume power density with contours at 115, 3.2, and $0.098\,kW{\cdot}cm^{-3}$.

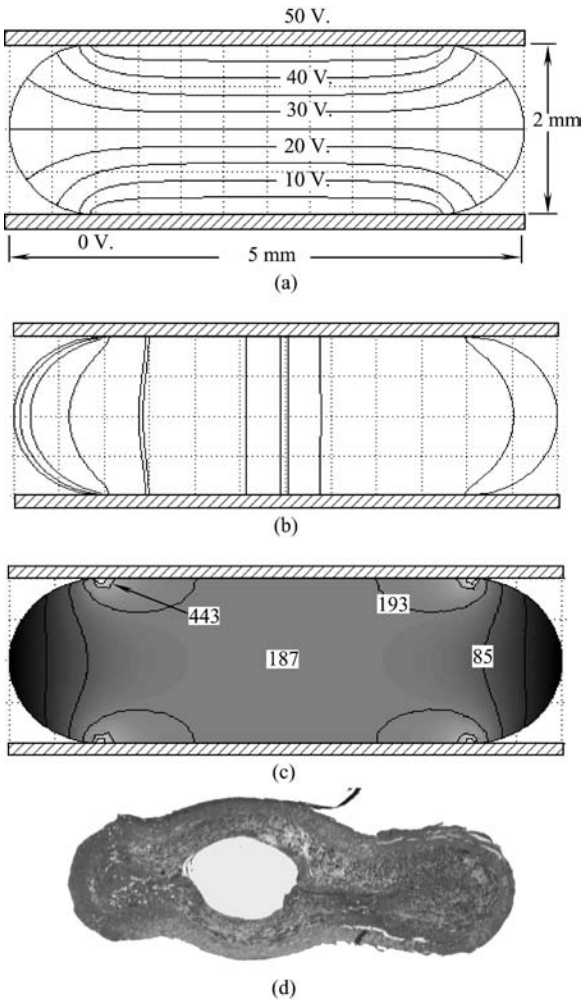


Figure 27. Bipolar forceps applied to a vessel, FEM model results. (a) Potential field, *V(x,y)* for 50 V(rms) electrode potential. (b) Electric field streamlines. (c) Volume power density, q_{gen}, with contours at 85, 193, and 443 ($W \cdot cm^{-3}$). (d) Typical result: canine femoral artery temperature 89 °C for 2 s (thermocouple in center of artery; original magnification 40×, Mallory's trichrome stain).

SUMMARY

Radio frequency current has been used for many years to reduce time under anesthesia and improve the clinical result. The ability to simultaneously cut tissue and coagulate blood vessels greatly reduces morbidity and mortality, and enables a large number of procedures that are otherwise not possible, especially in neurosurgery. Electrosurgery is one of several energy sources making substantial contributions in surgery, sources that include: lasers, microwaves, ultrasound, micromechanical disruption, and cryogenics. The major advantage of electrosurgery over the other modes originates from its ability to shape the treated tissue volume by design of the applicator electrode geometry alone. Consequently, a single electrosurgical generator is capable of a wide range of applications from various forms of coagulation and tissue fusion to cutting, either through an endoscope or in open procedures, either microsurgery or, as it were, macrosurgery. Since the electrodes are in contact with tissue, apposition pressure may be applied to improve the fusion probability. Electrode contact also gives the surgeon tactile feedback not characteristic of many of the other energy sources.

Useful predictions of clinical outcomes may be made with quasistatic electric field models coupled to transient thermal models including kinetic formulations of damage accumulation. Numerical model work can be used (*1*) to study trends due to uncertainty in tissue properties or changes in electrode voltage, current, power, or time of application; (*2*) to reveal discontinuities in the electric field due to boundary, conditions; and (*3*) to study the relative importance of each of the governing physical phenomena individually. When trends in the model match experimental results, one knows that all of the relevant physical phenomena have been included and the surgical process is well understood.

As with any such source of energy there are hazards in the use of electrosurgery. The hazards may be minimized, but not eliminated. Nevertheless, the advantages of use far outweigh the disadvantages, particularly in several very important procedures. The best safeguard is an informed and vigilant user population.

BIBLIOGRAPHY

Cited References

1. Major RA. History of Medicine Vols. I and II. Springfield (IL): Charles C. Thomas; 1954.
2. Breasted JH. The Edwin Smith Surgical Papyrus. Chicago: 1930.
3. Licht S. The history of therapeutic heat. Therapeutic Heat and Cold. 2nd ed. Elizabeth Licht; 1965. Chapt. VI.
4. Jee BS. A Short History of Aryan Medical Sciences. London: 1896.
5. Geddes LA, Roeder RA. DeForest and the first electrosurgical unit. IEEE Eng Med Biol Mag Jan/Feb 2003; 84–87.
6. Doyen E. Sur la destruction des tumeurs cancereuses accessibles; par la methode de la voltaisation bipolaire et de l'electro-coagulation thermique. Archg Elec Med Physiother Cancer 1909;17:791–795.
7. Doyen E. Surgical Therapeutics and Operative Techniques. Vol. 1. New York: Wm. Wood; ~ 1917; p 439–452.
8. Moyer CA, Rhoads JE, Allen JG, Harkins HN. Surgery Principles and Practices. 3rd ed. Philadelphia: Lippincott; 1965.
9. Cushing H, Bovie WT. Electrosurgery as an aid to the removal of intracranial tumors. Surg Gyn Obstet 1928;47: 751–784.
10. Kelly HA, Ward GE. Electrosurgery. Philadelphia: Saunders; 1932.
11. Mitchell JP, Lumb GN, Dobbie AK. A Handbook of Surgical Diathermy. Bristol: John Wright & Sons; 1978.
12. Otto JF, editor. Principles of Minor Electrosurgery. Liebel-Flarsheim Co.; 1957.
13. Harris HS. Electrosurgery in Dental Practice. Philadelphia: Lippincott; 1976.
14. Malone WF. Electrosurgery in Dentistry; Theory and Application in Clinical Practice. Springfield (IL): Chas C Thomas; 1974.
15. Oringer MJ. Electrosurgery in Dentistry. 2nd ed. Philadelphia: Saunders; 1975.
16. Oringer MJ. Color Atlas of Oral Electrosurgery. Chicago: Quintessence; 1984.

17. Epstein E, Epstein E, Jr, editors. Techniques in Skin Surgery. Philadelphia: Lea & Febiger; 1979.
18. Burdick KH. Electrosurgical Apparatus and Their Application in Dermatology. Springfield (IL): Chas C Thomas; 1966.
19. Pearce JA. Electrosurgery. Chichester: Chapman & Hall; 1986.
20. Knecht C, Clark RL, Fletcher OJ. Healing of sharp incisions and electroincisions in dogs. JAVMA 1971;159:1447–1452.
21. Kelly HA, Ward GE. Electrosurgery. Philadelphia: Saunders; 1932.
22. Operating and service instructions for model 770 Electrosectilis, The Birtcher Corporation, El Monte (CA).
23. Operator and service manual, MF-380, Aspen Laboratories Inc., Littleton (CO).
24. Valleylab SSE2-K service manual. Boulder (CO): The Valleylab Corporation;
25. User's guide, Force 2 electrosurgical generator. The Valleylab Corporation, Boulder (CO).
26. Medical electrical equipment part 1: General requirements for safety. IEC/EN 60601-1, International Electrotechnical Commission; 1988.
27. Medical electrical equipment part 2-2: Particular requirements for safety—High frequency surgical equipment. IEC/EN 60601-2-2, International Electrotechnical Commission; 1999.
28. Electrosurgical units. Health Devices. Jun–Jul 1973;2:n8–9.
29. Henriques FC. Studies of thermal injury V: The predictability and the significance of thermally induced rate processes leading to irreversible epidermal injury. Arch Pathol 1947;43:489–502.
30. Panescu D, Fleischman SD, Whayne JG, Swanson DK. Contiguous Lesions by Radiofrequency Multielectrode Ablation. Proc IEEE-Eng Med Biol Soc 17th Annu Meet. 1995; 17, n1.
31. Pearce JA, Thomsen S. Numerical Models of RF Ablation in Myocardium. Proc IEEE-Eng Med Biol Soc 17th Annu Meet; vol. 17, n1, 1995; p 269–270.
32. Lemole GM, Anderson RR, DeCoste S. Preliminary evaluation of collagen as a component in the thermally-induced 'weld'. Proc SPIE 1991;1422:116–122.
33. Kopchock GE, et al. CO_2 and argon laser vascular welding: acute histologic an thermodynamic comparison. Lasers Surg. Med. 1988;8(6):584–8.
34. Schober R, et al. Laser-induced alteration of collagen substructure allows microsurgical tissue welding. Science 1986;232:1421–11.
35. Collagen Volume I Biochemistry. Nimni ME, editors. Boca Raton (FL): CRC Press; 1988.
36. Pearce JA, Thomsen S, Vijverberg H, McMurray T. Kinetic rate coefficients of birefringence changes in rat skin heated in vitro. Proc SPIE 1993;1876:180–186.
37. Chen SS, Wright NT, Humphrey JD. Heat-induced changes in the mechanics of a collagenous tissue: isothermal, isotonic shrinkage. Trans ASME J Biomech Eng 1998;120:382–388.
38. Chen SS, Wright NT, Humphrey JD. Phenomenological evolution equations for heat-induced shrinkage of a collagenous tissue. IEEE Trans Biomed Eng 1998;BME-45:1234–1240.
39. Honig WM. The mechanism of cutting in electrosurgery. IEEE Trans Bio-Med Eng 1975;BME-22:58–62.
40. Geddes LA, Tacker WA, Cabler PA. A new electrical hazard associated with the electrocautery. Biophys Bioengr Med Instrum 1975;9n2:112–113.
41. Hungerbuhler RF, et al. Ventricular fibrillation associated with the use of electrocautery: a case report. JAMA 21 Oct 1974;230n3:432–435.
42. Orland HJ. Cardiac pacemaker induced ventricular fibrillation during surgical diathermy. Anesth Analg Nov 1975;3n4; 321–326.
43. Titel JH, et al. Fibrillation resulting from pacemaker electrodes and electrocautery during surgery. Anesthesiol Jul–Aug 1968;29:845–846.
44. Batra YK, et al. Effect of coagulating and cutting current on a demand pacemaker during transurethral resection of the prostate: a case report. Can Anesthes Soc J Jan 1978;25n1: 65–66.
45. Fein RL. Transurethral electrocautery procedures in patients with cardiac pacemakers. JAMA 2 Oct 1967;202: 101–103.
46. Greene LF. Transurethral operations employing high frequency electrical currents in patients with demand cardiac pacemakers. J Urol Sept 1972;108:446–448.
47. Krull EA, et al. Effects of electrosurgery on cardiac pacemakers. J Derm Surg Oct 1975;1n3:43–45.
48. Wiley JD, Webster JG. Analysis and control of the current distribution under circular dispersive electrodes. IEEE Trans BioMed Eng May 1982;BME-29n5:381–384.
49. Choi BJ, Kim J, Welch AJ, Pearce JA. Dynamic impedance measurements during radio-frequency heating of cornea. IEEE Trans Biomed Eng 2002;49n12:1610–1616.

See also CRYOSURGERY; ION-SENSITIVE FIELD EFFECT TRANSISTORS.

EMERGENCY MEDICAL CARE.

See CARDIOPULMONARY RESUSCITATION.

EMG.

See ELECTROMYOGRAPHY.

ENDOSCOPES

BRETT A. HOOPER
Areté Associates
Arlington, Virginia

INTRODUCTION

The word endoscope is derived from two Greek words, endo meaning "inside" and scope meaning "to view". The term endoscopy is defined as, "using an instrument (endoscope) to visually examine the interior of a hollow organ or cavity of the body." In this second edition of the Encyclopedia of Medical Devices and Instrumentation, we will revisit the excellent background provided in the first edition, and then move on from the conventional "view" of endoscopes and endoscopy to a more global "view" of how light can be delivered inside the body and the myriad light-tissue interactions that can be used for both diagnostic and therapeutic procedures using endoscopes. We will update the medical uses of endoscopy presented in the first edition, and then look at the medical specialties that have new capabilities in endoscopy since the first edition; cardiology and neurosurgery. This will be by no means an exhaustive list, but instead a sampling of the many capabilities now available using endoscopes. We will also present new delivery devices (fibers, waveguides, etc.) that have been introduced since the optical fiber to allow a broader range of wavelengths and more applications that deliver light inside the body.

HISTORY

Pre-1800s

Endoscopy had its beginnings in antiquity with the inspection of bodily openings using the speculum, a spoon-shaped instrument for spreading open the mouth, anus, and vagina. Another instrument of fundamental importance to endoscopy is the catheter, as it has been used to evacuate overfilled bladders for more than 3000 years. Catheters and a rectoscope were used by Hippocrates II (460–377 BC), where inspection of the oral cavity and of the pharynx was routine, including operations on tonsils, uvula, and nasal polyps. This trend continued for the better part of two millennia.

1800–1900

In 1804, Phillip Bozzini came upon the scene with the *Lichtleiter*, and attempts were made to view into the living body through the endoscope's narrow opening (1). It took almost 100 years to achieve this goal. These prototype endoscopes consisted of hollow tubes through which light of a candle, and even sunlight, was projected in order to visualize the inside of a bodily opening. The *Lichtleiter*, the light conductor, consists of two parts: (*1*) the light container with the optical part; and (*2*) the mechanical part, which consists of the viewing tubes fitted to accommodate the anatomical accesses of the organs to be examined. The apparatus is shaped like a vase, is $\sim$35 cm tall, made of hollow lead, and covered with paper and leather. On its front is a large, round opening divided into two parts by a vertical partition. In one half, a wax candle is placed and held by springs such that the flame is always in the same position. Concave mirrors are placed behind the candle and reflect the candle light through the one-half of the tube onto the object to be examined. The image of the object is then reflected back through the other half of the tube to the eye of the observer. Depending on the width of the cavity to be examined, different specula were used. These specula consisted of leaves that could be spread open by use of a screw device in order to expand the channels. An example is shown in Fig. 1.

Figure 1. Phillip Bozzini's *Lichtleiter* endoscope.

In 1828, the physicist C. H. Pfaff mentioned that platinum wires could be made incandescent through electric current. The glowing platinum wire has a desirable side effect; its white-hot heat illuminates the cavity of a surgical field so brightly that it provided the first internal light source. In 1845, the Viennese dentist Moritz Heider was the first to combine the illumination and tissue heating capabilities of the platinum wires when he cauterized a dental pulp with this method. However, the simultaneous heat that is produced by the wire is so intense that it was difficult to attach to the tip of small endoscopes. The term endoscopy was first used in the medical literature in 1865 by two French physicians, Segals and Desmoreaux. In 1873, Kussmaul, successfully passed a hollow tube into the stomach of a sword-swallower. Observation was limited because these long stiff hollow tubes were poorly lit by sun, candlelight, and mirrors. It took the advent of X rays to prove the swallowing of a sword was not a trick (2). Later that year, Gustave Trouvé, an instrument maker from Paris, introduced the polyscope at the World Exhibition in Vienna. His polyscope took many forms; rectoscope, gastroscope, laryngoscope, and cystoscope for looking into the rectum, stomach, larynx, and urinary tract, respectively. Trouvé was responsible for the idea of using electric current successfully for endoscopic illumination by placing the light source at the tip of the instrument, and the first to accomplish internal illumination. He was also the first to utilize a double prism for two observers by splitting the field with a Lutz prism by 90°, and incorporated a Galilean lens system, which magnified 2.5-fold, but did not enlarge the visual field. In 1876, Maximillian Nitze started his work on the urethroscope, and by the fall of 1877 he had instruments for illumination of the urethra, bladder, and larynx that used the platinum wire illumination method. Nitze also reported the first cystoscopy of the urinary bladder in 1877. As early as 1879, Nitze had his instrument patented for bladder, urethra, stomach, and esophagus in both Europe and the United States. The use of fiber optics was foreshadowed by Dr. Roth and Professor Reuss of Vienna in 1888 when they used bent glass rods to illuminate body cavities. With the invention of the light bulb by Thomas Edison, small light sources could be attached to the tip of the endoscopes without the need for cooling. The mignon lamp with a relatively small battery to operate it, led the next resurgence in endoscopy.

1900–Present

It was not until the early 1900s, however, that endoscopy became more than a curiosity. With the advancement of more sophisticated equipment and the discoveries of electrically related technologies, illumination capabilities were greatly improved. New and better quality lenses, prisms, and mirrors dramatically enhanced the potential applications of endoscopy and opened up new avenues of application. During this period, Elner is credited in 1910 with the first report of a technically advanced gastroscope (3) to view the stomach, and only 2 years later, Sussmann reported the first partially flexible gastroscope, which utilized screw systems and levers to contort the scope's shaft (4). The diagnostic merit of endoscopy was beginning to be

realized, as an atlas of endoscopic pathologies was being developed (5). Methods for saving images of the scene under investigation ranged from the early metal halide photographic process to 35 mm film cameras. Cameras were attached on view ports to allow simultaneous documentation of the view that the endoscopist had. Only within the last few decades has the use of digital photography been applied to the field of endoscopy, with an enhancement in diagnostic potential because of digital signal processing capabilities. In spite of the technical advances, problems with heat produced at the tip, blind spots in the field of view, and organ perforation limited the widespread use of endoscopes.

The most significant technological development began in the 1950s with the introduction of fiber optics. These thin glass or plastic optical fibers (on the order of 1 mm) allowed for "cold light" illumination at the tip, controlled flexibility to remove blind spots, and with quality image transmission opened up photographic capabilities, thus improving previous limitations. The use of fiber optics also allowed for the incorporation of ancillary channels for the passage of air, water, suction, and implementation of biopsy forceps or instruments for therapeutic procedures. Obviously, the most significant capabilities fiber optics has brought to endoscopes are their small size, flexibility, and ability to deliver laser light. Lasers are now used in both diagnostic and therapeutic procedures. In fact, the overwhelming advantage of using laser light is that it can act as both a diagnostic and therapeutic light source, and often in the same endoscope. Let us now look briefly at the theory behind efficient propagation of light in an optical fiber.

LIGHT DELIVERY WITH OPTICAL FIBERS

Efficient light delivery is made possible in solid-core optical fibers by use of total internal reflection. Total internal reflection (TIR) is achieved at the interface between two media when the incident medium has a refractive index larger than the transmitting medium, $n_i > n_t$, and the angle of incidence is greater than the critical angle. The critical angle is defined as the angle of incidence $_i$ where the transmitted angle $_t$ goes to 90°, and with no transmittance there is total internal reflection. For a given interface where n_i and n_t are known, the critical angle can be calculated from Snell's law as (6)

$$\theta_{\text{critical}} = \sin^{-1}(n_t \sin \theta_t / n_i) \tag{1}$$

For the optical fiber, TIR is achieved by designing the core refractive index n_i to be greater than the cladding refractive index n_t, and by focusing the input light so that it propagates along the axis of the fiber thereby maintaining a large angle with respect to the interface between core and cladding (see Fig. 2). Apart from choosing a fiber material that has close to lossless transmission at the wavelength of choice, one needs to focus the light beam so that the diameter of the fiber core is at least three times the radius of the input light beam and the cone of convergent focused light of angle 2 is less than the acceptance cone of the fiber core. The acceptance half-angle of the fiber is related to the

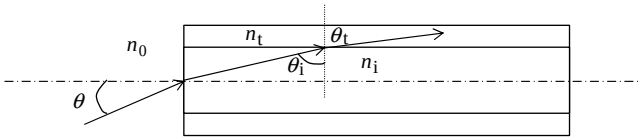


Figure 2. Light coupling into an optical fiber with core refractive index n_i and cladding refractive index n_t, and the associated incident angles (θ and θ_i), refracted angles, and their ray trajectories. The numerical aperture (NA) is a measure of the acceptance angle θ of a fiber.

NA and the refractive indexes by

$$\text{NA} = n_0 \sin\theta = (n_i^2 - n_t^2)^{1/2} \tag{2}$$

Numerical apertures for solid-core fibers range from 0.2 to 1 for fibers in air, where $n_0 = 1.0$, and are a measure of the light gathering capability of a fiber. An NA of 1 indicates that essentially all light incident at the fiber, even light near a 90° angle of incidence, can be launched into the fiber. In practicality, the NA approaches 1 and focusing and transverse mode distribution then play an important role in efficient light coupling. This is the case for unclad single-crystal sapphire fibers where the NA approaches one. For hollow-core waveguides (HWGs) the same rules apply for efficient coupling, that is, core diameter at least three times the radius of the input light beam and the convergence angle of the focused light θ less than the acceptance angle of the waveguide. Waveguiding is accomplished by coating the inside of the HWG with a metallic or dielectric coating depending on the desired wavelength. Fibers are available in diameters from several micrometers, for single-mode transmission, to ~1 mm, for multimode fibers. A propagating mode in a fiber is a defined path in which the light travels. Light propagates on a single path in a single-mode fiber or on many paths in a multimode fiber. The mode depends on geometry, the refractive index profile of the fiber, and the wavelength of the light. In the next section, we will look at the many different kinds of "optical fibers" that have become available since the previous edition and the myriad wavelengths that are now available for diagnostic and therapeutic procedures through endoscopes.

NEW "OPTICAL FIBER" DEVICES

There are now available to the endoscopist an array of different light delivery devices that can be incorporated into an endoscope. Conventional silica glass fibers transmit wavelengths from the visible (400 nm) to the mid-infrared (IR) (2.5 μm). New solid-core fibers are available for the transmission of IR wavelengths; these include germanium-oxide glass, fluoride glass, sapphire (out to 5 μm), and silver halide fibers (out to 30 μm) (7–9). In addition to the conventional solid-core optical fiber, there are now fibers available that have a hollow core and fibers that are intentionally "holey", that is to say they have a solid core with an array of periodic holes that run the length of the fiber. The hollow waveguides come in two flavors, but each efficiently transmits IR wavelengths.

The first hollow waveguide is a small silica tube that is metal coated on the inside and efficiently transmits IR light

from ~ 2 to 10 μm. These HWGs can be optimized for transmission of 2.94 μm Er:YAG laser light or 10.6 μm CO_2 laser light, and can therefore handle high laser power for therapeutic procedures (Polymicro Technologies, LLC). The second HWG is known as a photonic bandgap fiber (PBG) and is made from a multilayer thin-film process, rolled in a tube, and then heat drawn into the shape of a fiber. These PBG fibers are designed to transmit IR wavelengths, but only over a narrow band of 1–2 μm, for example, 9–11 μm for delivery of CO_2 laser light (Omniguide Communications, Inc.). The "holey" fibers are silica fibers that have been very carefully etched to create holes in a periodic pattern about the center of the core that span the length of the fiber. The pattern of these holes is chosen to propagate light very efficiently at a particular band of wavelengths, hence these fibers can be designed for a particular application; to-date largely in the visible and near-IR.

With each of these new fibers there are trade-offs in flexibility, transmission bandwidth, and ability to handle high power laser light. There are also differing fiber needs for diagnostic and therapeutic application. Diagnostic imaging fibers, for example, require a large NA for high light gathering capability. Comparatively, a therapeutic fiber endoscope may require a small NA to confine the delivery of the high power laser light to a specific region. This can also be achieved by use of evanescent waves launched at the interface between a high refractive-index optic and the tissue. Evanescent waves are different from the freely propagating light that typically exits fibers in that it is a surface wave. This surface wave only penetrates into the tissue on the order of the wavelength of the light. For typical wavelengths in the visible and IR, this amounts to light penetration depths on the order of microns, appropriate for very precise delivery of light into tissue. Both diagnostic and therapeutic applications have been demonstrated by coupling a high refractive-index conic tip (sapphire and zinc sulfide) to a HWG (see Fig. 3) (10). The diagnostic capability allows Fourier transform infrared (FTIR) spectroscopy to be performed on living (*In vivo*) tissue, where, for example, fatty tissues can be distinguished from normal intimal aorta tissue. Through the same HWG catheter, tissue ablation of atherosclerotic plaque has proven the therapeutic capabilities. These diagnostic and therapeutic applications can potentially take advantage of evanescent waves, HWGs, and mid-infrared FT spectroscopy in the 2–10 μm wavelength range (10).

A hybrid optical fiber consisting of a germanium trunk fiber and a low OH silica tip has shown the ability to transmit up to 180 mJ of Er:YAG power for applications requiring contact tissue ablation through a flexible endoscope (11). This pulse energy is more than sufficient for ablation of a variety of hard and soft tissues.

Next, we will look at the potential influence these new light delivery systems may have in endoscopy.

MEDICAL APPLICATIONS USING ENDOSCOPY

Endoscopy has had a major impact on the fields of medicine and surgery. It is largely responsible for the field of minimally invasive surgery. The ability to send diagnostic and therapeutic light into the body via minimally invasive procedures has reduced patient discomfort, pain, and length of hospital stay; and in some cases has dramatically changed the length of stay after a procedure from weeks to days. Optical fibers have had a profound effect on endoscopy, and in doing so, dramatically changed medicine and surgery. These small, flexible light pipes allowed physicians to direct light into the body where it was not thought possible, and even to direct laser light to perform microsurgeries in regions previously too delicate or intricate to access.

In this section, we will examine the state of endoscopy in arthroscopy, bronchoscopy, cardiology, cystoscopy, fetoscopy,

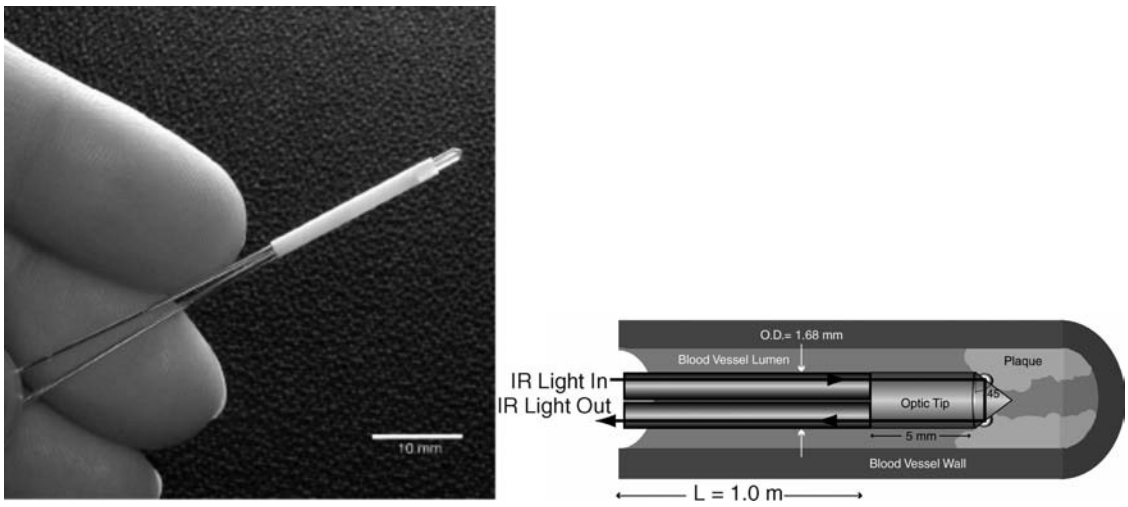


Figure 3. An example of a HWG endoscope. Both diagnostic and therapeutic applications have been demonstrated by coupling a high refractive-index conic tip (sapphire and zinc sulfide) to a HWG. In this geometry with two waveguides, diagnostic spectrometer light can be coupled into the tip in contact with tissue via one waveguide and sent back to a detector via the other waveguide.

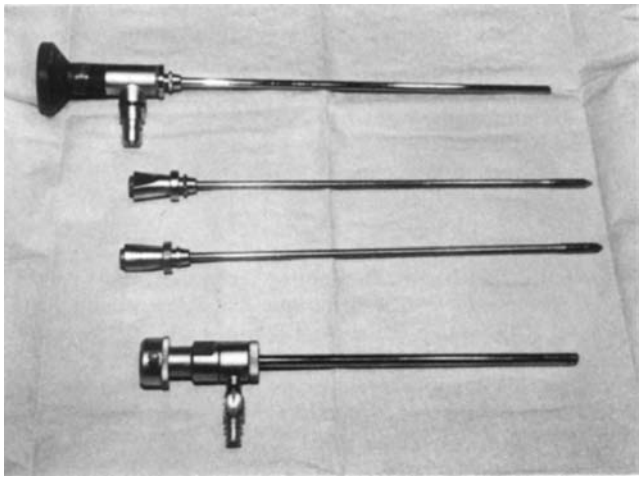


Figure 4. Arthroscope and internal components used to view the interior of knee joint space.

gastrointestinal endoscopy, laparoscopy, neurosurgery, and otolaryngology.

Arthroscopy

Arthroscopy had its birth in 1918 when Takagi modified a pediatric cystoscope and viewed the interior of the knee of a cadaver (12). Today its use is widespread in orthopedic surgery with major emphasis on the knee joint. Arthroscopy has had a major impact on the way knee surgery is performed and with positive outcomes. The surgery is minimally invasive often with only two small incisions; one for an endoscope to visualize the interior of the knee, and another to pass microsurgical instruments for treatment. Presently, endoscopes have the ability to view off-axis at a range of angles from 0 to 180°, with varying field-of-view (see Fig. 4). The larger the field-of-view the more distorted the image, as in a fish-eye lens. A 45° off-axis endoscope, for example, with a 90° field-of-view can be rotated to visualize the entire forward-looking hemisphere.

The most common procedure is arthrotomy, which is simply the surgical exploration of a joint. Possible indications include inspection of the interior of the knee or to perform a synovial biopsy. Synovia are clear viscous fluids that lubricate the linings of joints and the sheaths of tendons. Other indications include drainage of a hematoma or abscess, removal of a loose body or repair of a damaged structure, such as a meniscus or a torn anterior cruciate ligament, and the excision of an inflamed synovium (13,14).

Future directions in orthopedic surgery will see endoscopes put to use in smaller joints as endoscopes miniaturize and become more flexible. The ultimate limit on the diameter of these devices will likely be 10s of micrometers, as single-mode fibers are typically 5–10 μm in diameter. These dimensions will allow for endoscopy in virtually every joint in the body, including very small, delicate joints. Work in the shoulder and metacarpal-phalanges (hand–finger) joints is already increasing because of these new small flexible fiber optic endoscopes.

Bronchoscopy

Bronchoscopy is used to visualize the bronchial tree and lungs. Since its inception in the early 1900s bronchoscopy has been performed with rigid bronchoscopes, with much wider application and acceptance following the introduction of flexible fiber optics in 1968. An advantage of the equipment available is its portability, allowing procedures to be done at a patient's bedside, if necessary. With the introduction of fiber optics, in 1966, the first fiber optic bronchoscope was constructed, based on specifications and characteristics that were proposed by Ikeda. He demonstrated the instrument's use and application and named it the bronchofiberscope. Development over the last several decades has seen the use of fiberoptic endoscopes in the application of fiberoptic airway endoscopy in anesthesia and critical care. These endoscopes have improved the safe management of the airway and conduct of tracheal and bronchial intubation. Fiber optic endoscopy has been particularly helpful in the conduct of tracheal and bronchial intubation in the pediatric population.

Bronchoscopy is an integral part in diagnosis and treatment of pulmonary disease (15,16). Bronchoscopic biopsy of lung masses has a diagnostic yield of 70–95%, saving the patient the higher risk associated with a thoracotomy. Pulmonary infection is a major cause of morbidity and mortality, especially in immuno-compromised patients, and bronchoscopy allows quick access to secretions and tissue for diagnosis. Those patients with airway and/or mucous plugs can quickly be relieved of them using the bronchoscope. Another diagnostic use of the bronchoscope is pulmonary alveolar lavage, where sterile saline is instilled into the lung then aspirated out and the cells in the lavage fluid inspected for evidence of sarcoidosis, allergic alveolitis, for example. Lavage is also of therapeutic value in pulmonary alveolar proteinosis.

Bronchoscopy is usually well tolerated by the patient with complications much less than 1% for even minor complications. Laser use has also allowed for significant relief of symptoms in cancer patients.

Cardiology

Early developments in minimally invasive cardiac surgery included the cardiac catheter (1929), the intra-aortic balloon pump (1961), and balloon angioplasty (1968). Endoscopy in cardiology has largely focused on using intravascular catheters to inspect the inside of blood vessels and more recently the inside of the heart itself. Catheters are thin flexible tubes inserted into a part of the body to inject or drain away fluid, or to keep a passage open. Catheters are similar to endoscopes, and they also have many diagnostic and surgical applications.

For diagnostic purposes, angiography uses X rays in concert with radio-opaque dyes (fluoroscopy) to look for blockages in vessels, usually the coronary arteries that supply oxygenated blood to the heart. A catheter is introduced into the femoral artery and sent up the aorta and into the coronary arteries to assess blood flow to the heart. The catheter releases the dye and real-time X-ray fluoroscopy tracks the dye as it is pumped through the coronary artery. Angiography in concert with intravascular

Figure 5. Image of a blood vessel with a stent using IVUS. (Courtesy of LightLab Imaging.)

ultrasound (IVUS) is the currently accepted diagnostic in cardiology (17–21). IVUS emits acoustic energy out the tip of a catheter and listens for echoes to image the inside of coronary arteries (see Fig. 5). An IVUS image is a cross-sectional view of the blood vessel, and complements the X-ray fluoroscopy image. The IVUS has been used to assess atherosclerotic plaques in coronary arteries and has been very successful at guiding the placement of stents. Stents are expandable metal mesh cages in the shape of cylinders that act as scaffolding to open obstructions in vessels caused by atherosclerosis.

Another technique showing promise in endovascular imaging uses light. The light-driven technology, optical coherence tomography (OCT), is successful at detecting fatty plaques, including those that are vulnerable to rupture. Figure 6 compares OCT images of coronary artery to

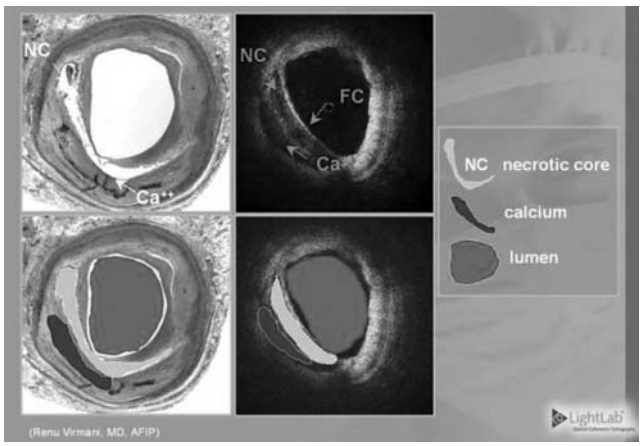


Figure 6. Images of coronary artery comparing OCT with conventional histology. (Courtesy of LightLab Imaging.)

conventional histology (LightLab Imaging, Inc.). Visible are the necrotic core of an atherosclerotic plaque and the thin cap at the intimal surface of the vessel. Near-IR light has been used diagnostically for differentiating between oxygenated and deoxygenated blood in myocardium (22–24). The NIR light in the 2–2.5 μm wavelength range has also been used to characterize myocardial tissue (25). In this wavelength range, the absorption due to water is declining through a local minimum, while there are absorption peaks in fat at 2.3 and 2.4 μm that show distinct spectral signatures. It has also been suggested that IR light might be used to identify atherosclerotic plaques (26–29), including plaques that are at high risk of rupture or thrombosis (30). Arai et al. showed that fibrofatty plaques have characteristic absorbances at mid-IR wavelengths of 3.4 and 5.75 μm that are significantly greater than normal aorta tissue (29). Peaks at these wavelengths, and potentially other subtleties in the absorption spectra [derived from multivariate statistical analysis (MSA)], can be targeted for developing a diagnostic profile similar to that described by Lewis and co-workers for IR and Raman spectroscopic imaging (31,32). These mid-IR frequencies may also be optimized for selective therapy via precise laser surgery (10,33,34).

Surgical laser techniques have become routine over the past two decades in a number of medical specialties such as ophthalmology and dermatology. However, in cardiology initial enthusiasm for fiber optic catheter ablation of atherosclerotic plaque (laser angioplasty) waned in the face of unpredictable vascular perforations, restenosis, and thrombosis (35,36). Therapeutically, IR light has found application in transmyocardial revascularization (TMR) through several partial myocardial perforations (37). Ideally, lasers can ablate tissue with exquisite precision and nearly no damage to the surrounding tissue. This precision requires extremely shallow optical penetration, such that only a microscopic zone near the tissue surface is affected. This has been accomplished by using short laser pulses at wavelengths that are strongly absorbed by proteins (193 nm in the UV) (38) or water (3 μm in the IR) (39).

Therapeutic techniques using catheters recently have targeted the atherosclerotic plaque that is deposited in the vessel wall as we age. The deposition of atherosclerotic plaque in coronary arteries is important for two reasons; the arteries are small (1–3 mm), and they supply blood to the heart muscle itself, so any reduction or blockage of oxygen-supplying blood to the heart shows up symptomatically as angina or worse, a heart attack. Catheters equipped with lasers and inflatable balloons have been used to open these blockages by ablating the plaque or compressing the plaque back into the vessel wall, respectively. Problems with balloon angioplasty have been vessel wall perforation, thrombosis (blood clot formation), and restenosis (reobstruction of the vessel due to an immune response). Recently, balloon angioplasty was augmented by the use of stents, as mentioned above. These expandable cages are opened via a balloon catheter in the vessel lumen and act as a mechanical scaffold to hold the plaque obstruction in place on the vessel wall. Stenting, however, still suffered from restenosis, because of the mechanical injury induced by the stent. Stents coated with small amounts of

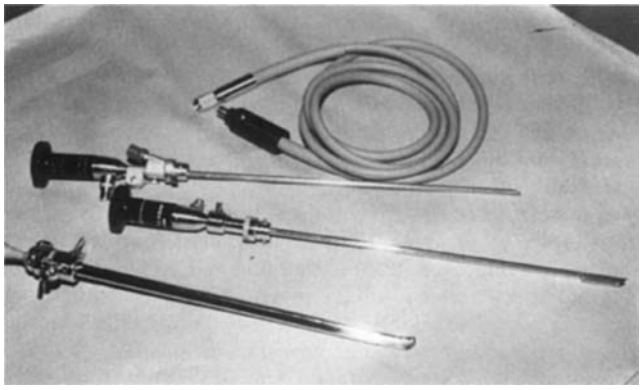


Figure 7. Cystoscope and accessories used for viewing the urinary tract in the medical specialty urology.

drugs (drug-eluding stents) have met with some success, in that they are able to retard the regrowth of intimal hyperplasia responsible for the restenosis.

Cystoscopy

Endoscopy is an integral part of the practice of urology. There are many procedures that look inside the urinary tract using a cystoscope, with typical access through the ureter (see Fig. 7). Common cystoscopy procedures include ureteral catheterization, fulguration and or resection of bladder tumor(s), direct vision internal urethrotomy, insertion of stent, removal of bladder stone, and kidney stone fragmentation (40,41).

A wide variety of therapeutic accessories are available for endoscopic treatment, including snares and baskets for stone removal, and electrohydraulic lithotripsy and laser delivery for fragmenting large stones (42,43). These procedures are minimally invasive and, therefore, can be accomplished with a much lower rate of morbidity and mortality than would be achieved in an open surgical procedure.

Laser incision of urethral, bladder neck, and urethral strictures and fragmentation of kidney stones is being investigated using flexible endoscopes made from germanium fibers to transport Er:YAG laser light. Bladder neck strictures are defined as a narrowing or stenosis of the bladder neck that may result in recalcitrant scarring and urinary incontinence. A significant number of patients undergoing surgery for benign or malignant prostate cancer suffer from bladder neck strictures, and there is no simple and effective minimally invasive treatment. The Er:YAG laser can ablate soft tissue $\sim$20–30 times better than a Ho:YAG laser (2.12 μm), which is the laser of choice in urology. The absorption coefficient is many orders of magnitude different for these two lasers, $\sim$10,000 cm^{-1} for Er:YAG versus 400 cm^{-1} for the Ho:YAG. This translates to a 1/*e* depth of optical penetration of 1 versus 25 μm for these two lasers. Water is the dominant absorptive chromophore in the tissue in this mid-IR region of the spectrum. Hence, the Er:YAG is better suited to procedures where precision is required in laser ablation of soft tissue.

Fetoscopy

Fetoscopy allows for the direct visualization of a fetus in the womb. It also allows for the collection of fetal blood samples and fetal skin sampling, for diagnosis of certain hemoglobinopathies and congenital skin diseases, respectively (44–46).

The instrument is a small-diameter (1–2 mm) needlescope with typical entry through the abdominal wall under local anesthesia and guided by ultrasound. Optimal viewing is from 18 to 22-weeks gestation when the amniotic fluid clarity is greatest. Once the instrument is introduced, a small-gauge needle is typically used to obtain blood samples or small biopsy forceps can be used for skin biopsy.

Complete fetal visualization is not usually achieved. The procedure has some risk with fetal loss $\sim$10% and prematurity another 10%. The future of this procedure is not clear, as it is not currently in general use, despite the safe and relatively simple prenatal diagnosis it offers. It has been replaced in many circumstances by ultrasound for the fetal visualization aspect, but is still valuable for blood and skin sampling.

Gastrointestinal Endoscopy

The techniques of fiberoptic gastrointestinal (GI) endoscopy were developed in the 1960s, but surgeons were slow to embrace the techniques. These procedures were developed by gastroenterologists who became skilled practitioners and teachers of the art (see Fig. 8). Gradually, GI surgeons adopted these procedures, and in 1980 the American Board of Surgery mandated that endoscopic training be a part of the curriculum in general surgical training. Endoscopy has since become an integral part of surgical education and practice (47). The GI endoscopy is used in the esophagus, stomach, small bowel and liver, biliary, colon, and in pediatric endoscopy. There are numerous accessories available for the GI endoscope. They include biopsy forceps, graspers, cautery tools, and wire snares (see Figs. 9 and 10).

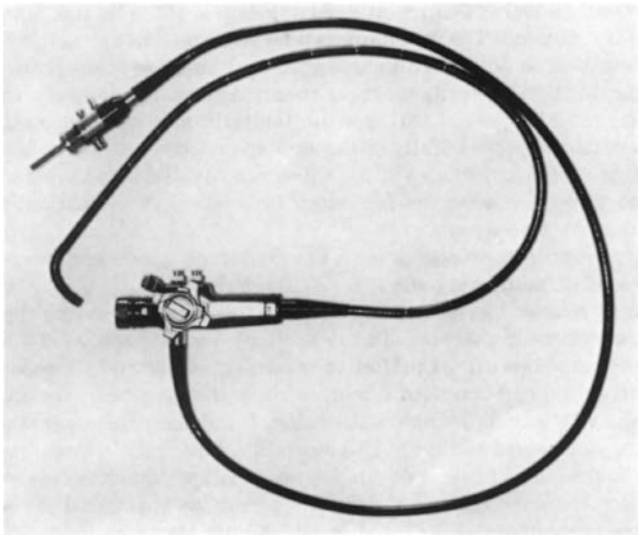


Figure 8. Upper intestinal panendoscope for the adult patient.

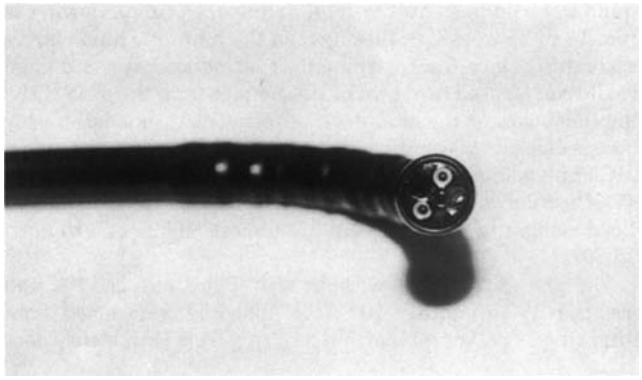


Figure 9. An end view of the distal tip of a panendoscope illustrating the accessory channels and illumination ports.

In the esophagus, endoscopy is used in the dilation of benign esophageal strictures, balloon dilatation of Achalasia, management of foreign bodies and bezoars of the upper GI tract, and endoscopic laser therapy of GI neoplasms (48–50). Endoscopy of benign esophageal strictures is a common procedure performed by gastroenterologists to treat esophageal narrowing and relieve dysphagia. Achalasia is a motility disorder of the esophagus characterized by symptoms of progressive dysphagia for both solids and liquids with (1) aperistalsis in the body of the esophagus, (2) high lower esophageal sphincter (LES) pressure, and (3) failure of the LES to relax. Examples of foreign bodies are coins, meat impaction, frequently in the elderly population, sharp and pointed objects, such as, a toothpick, a chicken or fish bone, needles, and hatpins. Bezoars are a hard mass of material often found in the stomach and are divided into three main types: phytobezoars, trichobezoars, and miscellaneous. These bezoars can be managed by endoscopy dependent on size and location, typically by capture and removal rather than endoscopic fragmentation. Since the 1970s, the incidence of esophageal adenocarcinoma has increased more rapidly than any other form of cancer and now represents the majority of esophageal neoplasms in the West (51). Esophagectomy is considered the gold standard for the treatment of high grade dysplasia in Barrett's esophagus (BE) and for noninvasive adenocarcinoma (ACA) of the distal esophagus (52). Barrett's esophagus is the replacement of native squamous mucosa by specialized intestinal metaplasia and is known to be the major risk factor for the development of adenocarcinoma. A recent study of 45 patients supports the use of endoscopic surveillance in patients who have undergone "curative" esophagectomy for Barrett's dysplasia or localized cancer (53–55). OCT imaging has recently shown the ability to image Barrett's esophagus through a small fiber endoscope. The GI neoplasm treatment is also particularly amenable using neodymium-YAG laser palliative treatment of malignancies of the esophagus and gastroesophageal junction as first described by Fleischer in 1982.

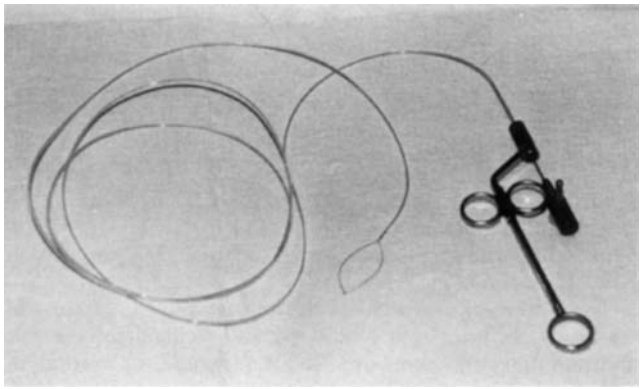


Figure 10. A polyp wire snare device with retraction handle.

For the stomach, endoscopic therapy for benign strictures of the gastric outlet is but one procedure. These strictures of the gastric outlet are similar to the esophageal stricture mentioned above. They are most frequently caused by peptic ulcer disease in the region of the pylorus, although chronic ingestion of nonsteroidal anti-inflammatory drugs is a frequent cause as well. Other endoscopic procedures in the stomach include sclerotherapy of esophageal varices (abnormally swollen or knotted vessels, especially veins), percutaneous endoscopic gastrostomy and jejunostomy, injection therapy for upper GI hemorrhage, and thermal coagulation therapy for upper GI bleeding.

In the small bowel and liver, enteroscopy is an endoscopic procedure to directly inspect the small bowel mucosa and take biopsies by enteroscopy from selected sites in the jejunum and proximal ileum. Until the recent development of the enteroscope, these parts of the small bowel were not possible to endoscopically evaluate. Enteroscopy is in its technological infancy. It presently allows for access to the most proximal aspects of the small bowel to obtain tissue at bleeding lesions. However, it cannot be utilized to adequately visualize the entire mucosa.

Endoscopy in the biliary system is used to perform sphincterotomies, papillotomies and stent insertions, to manage large bile duct stones and malignant biliary strictures, biliary and pancreatic manometry, and endoscopic retrograde cholangiopancreatography. The use of endoscopes all but changed the surgical technique for gallbladder removal and moved its outcome from a major inpatient surgery to a minimally invasive surgery with only small incisions for the endoscope.

In the colon, nonoperative and interventional management of hemorrhoids is performed with endoscopes. Other procedures include dilatation of colonic strictures, approaches to the difficult polyp and difficult colonic intubation, and clinical approaches of anorectal manometry. Additionally, colonoscopy is used for investigating irritable bowel syndrome, Crohn's disease, and ulcerative colitis (47,56).

Pediatric endoscopy has been used to perform gastroscopy, colonoscopy, and endoscopic retrograde cholangiopancreatography. Advances have largely involved the application of techniques used in adults to the pediatric patient; this was made possible with the introduction of smaller fiberoptic endoscopes.

And finally general endoscopy has been used as a-surveillance program for premalignant lesions, to assess outpatient endoscopy, and endosonography and echo probes.

Laparoscopy

Laparoscopy or peritoneoscopy is an important diagnostic procedure that allows direct visualization of the surface of many intra-abdominal organs, as well as allowing the performance of guided biopsies and minimally invasive therapy. Landmarks in laparoscopic surgery include the first laparoscopic appendectomy (1983) and the first laparoscopic cholescystectomy (1987). The first laparoscope was an ordinary proctoscope with illumination coming from an electric headlight worn by the endoscopist. Since then, with the advent of newer technologies, a fiber optic cable was added to the rigid telescope making flexible laparoscopes available. Today, despite the advent of various noninvasive scanning technologies, laparoscopy is still clinically useful for visualizing and biopsying intra-abdominal tumors, particularly those that involve the liver, peritoneum, and pelvic organs (57,58).

The laparoscopic procedure begins with the introduction of a trocar into the abdomen at the inferior umbilical crease for insufflation of carbon dioxide gas. A trocar is a sharply pointed steel rod sheathed with a tight-fitting cylindrical tube (cannula), used together to drain or extract fluid from a body cavity. The whole instrument is inserted then the trocar is removed, leaving the cannula in place. The gas acts as a cushion to permit the safe entry of sharp instruments into the peritoneal cavity and enable a better view. Common procedures include laparoscopic cholecystectomy and laparoscopic appendectomy, laparoscopic removal of the gallbladder and appendix, respectively. Laparoscopic cholecystectomy has all but replaced the conventional surgery for removal of the gallbladder. What used to involve opening the abdominal cavity for gallbladder removal and a 5–7 day stay at the hospital, has been transformed to a minimally invasive procedure with only a few days in the hospital. Lastly, laparoscopic sterilization and abortion are also performed with endoscopes.

Neuroendoscopy

Fiber optics is particularly well suited for the field of neuroendoscopy for both diagnostic and therapeutic procedures in the inner brain, because of size, flexibility, visualization, and laser delivery. We briefly review a case study that highlights some of the advantages of fiber optic endoscopes for minimally invasive surgery in the inner brain.

A recent case report on laser-assisted endoscopic third ventriculostomy (ETV) for obstructive hydrocephalus shows the use of diagnostic and therapeutic endoscopy in neurosurgery (59). Under stereotactic and endoscopic guidance, multiple perforations in the ventricular floor using a 1.32 μm neodymium–yttrium–aluminum–garnet (Nd:YAG) or an aluminum–gallium–arsenide (AlGaAs) 0.805 μm diode laser and removal of intervening coagulated tissue ensued with a 4–6 mm opening between third ventricle and basilar cisterns. These perforations allow for the cerebrospinal fluid (CSF) to be diverted so that a permanent communication can be made between the third cerebral ventricle and arachnoid cisterns of the cranial base. In a series of 40 consecutive cases, 79% of the patients had a favorable outcome. This compares well with a recent series summarizing >100 patients and long-term follow-up with success rates ranging from 50 to 84%.

When the 1.32 μm Nd:YAG laser is used, the high absorption in water requires that the fiber be placed in contact with the ventricular floor. Conversely, the high power diode laser's dispersion-dominant properties can lead to damage to neural structures around the ventricular cavity. Therefore, the 0.805 μm diode laser was used in a contact mode, but only after carbonization of the fiber tip so that thermal increase of the fiber tip allowing ventricular floor perforation was due to absorption of the laser energy by the carbon layer only and not by direct laser–tissue interaction. The 1.32 μm Nd:YAG laser was found to have higher efficiency for coagulation and perforation than the 0.805 μm diode laser, and would appear to be the better choice for neuroendoscopic use in this procedure. The endoscope allows for visualization and treatment of a very difficult part of the brain to access, and the use of lasers in endoscopes is advantageous in cases of distorted anatomy and microstructures and may reduce technical failures.

In another case of endoscopic-delivered laser light for therapeutic purpose, an IR free-electron laser (FEL) was used to ablate (cut) a suspected meningioma brain tumor at Vanderbilt's Keck FEL Center (60). A HWG catheter was used to deliver 6.45 μm IR FEL light to a benign neural tumor in the first human surgery to use a FEL. The 6.45 μm FEL light is a candidate for soft tissue surgery because of its ability to ablate (cut) soft tissue with a minimum of thermal damage to the surrounding tissue; on the order of micrometers of damage. This is obviously very important for surgery in the brain where viable, eloquent tissue may be in contact with the tumorous tissue that is being removed.

The FEL is a research center device that generates laser light over a vast portion of the electromagnetic spectrum; to date FELs have generated laser light from the UV (190 nm) to millimeter (61). The FEL is beneficial in identifying wavelengths, particularly in the IR where there are no other laser sources, for selective laser-tissue interaction.

Otolaryngology

Early use of endoscopes for ear, nose, and throat often focused on the interior of the ear. The benefit of endoscopes for diagnosis and therapy had been recognized early on with the advent of the laser and fiber optics (62–65). Recent investigations at the University of Ulm on the use of an Er:YAG laser with a germanium-oxide fiber delivery system has focused on tympanoplasty and stapedotomy (middle ear surgery) (66). The Er:YAG laser was found to be optimum for operating on the eardrum along the ossicles as far as the footplate without carbonization, and with sharp-edged, 0.2-mm-diameter canals "drilled" through the bone. Using this technique, children with mucotympanon could have their eardrums reopened in the doctor's office without the need for drain tubes.

An endoscope suitable for quantitatively examining the larynx (vocal chords) uses a green laser and a double reflecting mirror (67). The device can be clipped onto the shaft of a commercial rigid laryngoscope. The double reflecting mirror sends out two beams parallel to one

another that allows for quantitative morphometry of laryngeal structures such as, vocal cords, glottis, lesions, and polyps.

The miniaturization and flexibility of fiber optics has allowed endoscopes to be applied in the small and delicate organ of the ear with much success. A case in point for the unique capabilities that the fiber optic endoscope has that can be applied to many fields of medicine in a very productive manner.

Future Directions

One pressing issue for "reusable" endoscopes is the ability to guarantee a clean, sterile device for more than one procedure. With the advent of the World Wide Web (WWW), many web sites are available to gain information on endoscopes, as well as the procedures they are used in. The U.S. Food and Drug Administration (FDA) has a site at their Center for Devices and Radiological Health (CDRH) that monitors medical devices and their performance in approved procedures. The FDA has also created guidelines for cleaning these devices.

The results of an FDA-sponsored survey, Future Trends in Medical Device Technology: Results of an Expert Survey in 1998, expressed a strong view that endoscopy and minimally invasive procedures would experience significant new developments during the next 5 and 10 year periods leading to new clinical products (68). The 15 participants included physicians, engineers, healthcare policymakers and payers, manufacturers, futurists and technology analysts. In interviews and group discussions, survey participants expressed an expectation of continuing advancements in endoscopic procedures including fiber optic laser surgery and optical diagnosis, and a range of miniaturized devices. Clinically, most participants expected an emphasis on minimally invasive cardiovascular surgery and minimally invasive neurosurgery; two new fields we introduced in this edition. Also predicted were continuing advances in noninvasive medical imaging, including a trend to image-guided procedures. Most profound expectations were for developments in functional and multimodality imaging. Finally, participants observed that longer term trends might ultimately lead to noninvasive technologies. These technologies would direct electromagnetic or ultrasonic energy, not material devices, transdermally to internal body structures and organs for therapeutic interventions.

Perhaps a stepping-stone on this path is the PillCam (Given Imaging), a swallowable 11×26-mm capsule with cameras on both ends and a flashing light source, used to image the entire GI tract from the esophagus to the colon. The PillCam capsule has a field of view of 140°, and enables detection of objects as small as 0.1 mm in the esophagus and 0.5 mm in the small bowel. Figure 11 shows the PillCams used for small bowel (SB) and esophagus (ESO) procedures and samples of the images obtained. Shown are examples of active bleeding, Crohn's disease, and tumor in the small bowel; and normal Z-line, esophagitis, and suspected Barrett's in the esophagus. Patient exam time is 20 min for an esophageal procedure and 8 h for a small bowel procedure. As the PillCam passes through the GI tract images are acquired at 2 Hz and the information is transmitted via an array of sensors secured to the patient's chest and abdomen and passed to a data recorder worn around the patient's waist. The PillCam generates $\sim$57,000 images in a normal 8 h procedure, while the patient is allowed to carry on their normal activity. An obvious enhancement of patient comfort.

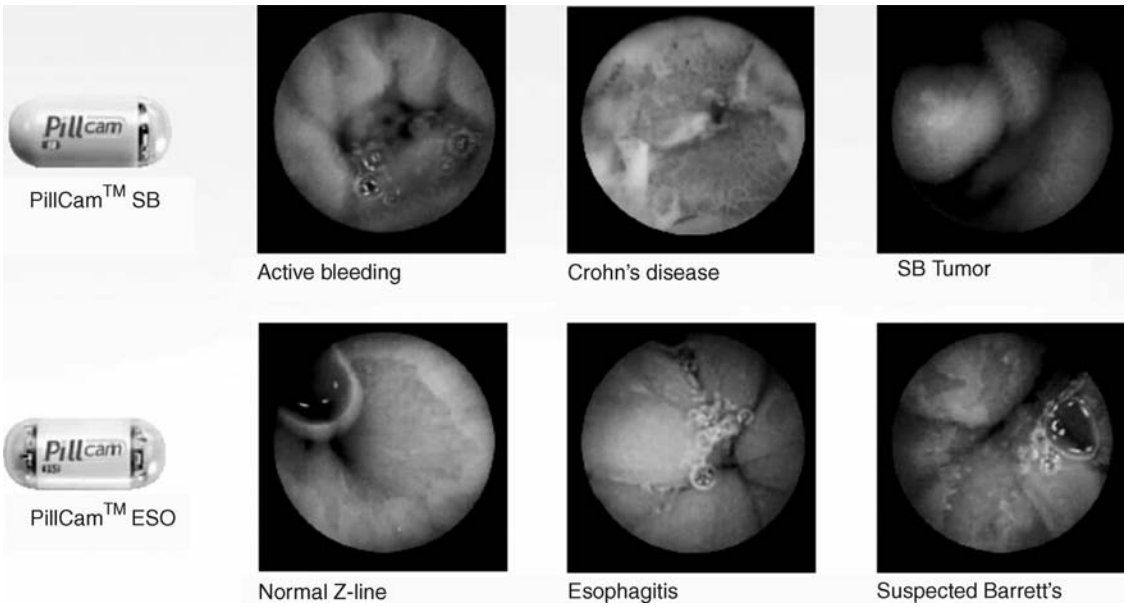


Figure 11. The PillCam from Given Imaging and sample images from the small bowel and esophagus. Shown are examples of active bleeding, Crohn's disease, and tumor in the small bowel; and normal Z-line, esophagitis, and suspected Barrett's in the esophagus. The PillCam is a swallowable 11×26 mm capsule with cameras on both ends and a flashing light source, used to image the entire GI tract from the esophagus to the colon.

Figure 12. The da Vinci Surgical System consists of (a) a surgeon console, patient-side cart, instruments, and image processing equipment. The surgeon operates while seated at a console viewing a 3D image of the surgical field. The surgeon's fingers (b) grasp master controls below the display that translate the surgeon's hand, wrist, and finger movements into precise, real-time movements of the surgical instruments (c) inside the patient. The patient-side cart provides the robotic arms, with two or three instrument arms and one endoscope arm, that execute the surgeon's commands. The laparoscopic arms pivot at the 1 cm operating ports eliminating the use of the patient's body wall for leverage and minimizing tissue damage. The instruments are designed with seven degrees of motion that mimic the dexterity of the human hand and wrist. Operating images are enhanced, refined, and optimized for 3D viewing through the (d) 3D endoscopic camera system. (Courtesy of Intuitive Surgical, Inc.)

AESOP (formerly Computer Motion and now Intuitive Surgical) was the first robot, FDA approved in 1994, to maneuver a tiny endoscopic video camera inside a patient according to voice commands provided by the surgeon. This advance marked a major development in closed chest and port-access bypass techniques allowing surgeons direct and precise control of their operative field of view. In 1999, one-third of all minimally invasive procedures used AESOP to control endoscopes.

The da Vinci Surgical System (Intuitive Surgical) provides surgeons the flexibility of traditional open surgery while operating through tiny ports by integrating robotic technology with surgeon skill. The da Vinci consists of a surgeon console, a patient-side cart, instruments and image processing equipment (see Fig. 12). The surgeon operates while seated at a console viewing a 3D image of the surgical field. The surgeon's fingers grasp the master controls below the display with hands and wrists naturally positioned relative to their eyes, and the surgeon's hand, wrist, and finger movements are translated into precise, real-time movements of endoscopic surgical instruments inside the patient. The patient-side cart provides up to four robotic arms, three instrument arms, and one endoscope arm that execute the surgeon's commands. The laparoscopic arms pivot at the 1 cm operating ports and are designed with seven degrees of motion that mimic the dexterity of the human hand and wrist. Operating images are enhanced, refined, and optimized using image synchronizers, high intensity illuminators, and camera control units to provide enhanced 3D images of the operative field via a dual-lens three-chip digital camera endoscope. The FDA has cleared da Vinci for use in general laparoscopic surgery, thoracoscopic (chest) surgery, laparoscopic radical prostatectomies, and thoracoscopically assisted cardiotomy procedures. Additionally, the da Vinci System is also presently involved in a cardiac clinical trial in the United States for totally endoscopic coronary artery bypass graft surgery. This technology will likely find application in vascular, orthopedic, spinal, neurologic, and other surgical disciplines that will certainly enhance minimally invasive surgery.

Minimally invasive technologies that enhance the present state of endoscopy will continue. The expectation that microelectromechanical systems (MEMS) technology will add a plethora of miniaturized devices to the armament of the endoscopist is well founded, as it is an extension of the impact that fiber optics had on the field of endoscopy. The MEMS will likely add the ability to have light source and detector at the tip of the endoscope, instead of piping the light into and out of the endoscope. Many other functions including "lab on a chip" MEMS technology may allow for tissue biopsies to be performed *in situ*. This miniaturization will likely lead to more capabilities for endoscopes, as well as, the ability to access previous inaccessible venues in the body. Endoscopy is poised to continue its substantial contribution to minimally invasive procedures in medicine and surgery. This will pave the way for the likely future of noninvasive procedures in surgery and medicine.

BIBLIOGRAPHY

Cited References

1. Matthias Reuter, Rainer Engel, Hans Reuter. History of Endoscopy. Stuttgart: Max Nitze Museum Publications; 1999.

2. Wolf RFE, Krikke AP. The X-files of sword swallowing. Available at www.ecr.org/Conferences/ECR1999/sciprg/abs/p010189.htm.
3. Elner HD. Ein gastroskop. Klin Wochenschr 1910;3:593.
4. Sussmann M. Zur Diptrik des gastroskop. Ther Gegenw 1912;53:115.
5. Schindler R, Lehrbuch U. Atlas D Gastroskop, Munich: Lehmann 1923.
6. Hecht E. Optics. 4th ed. New York: Addison-Wesley; 2002.
7. Harrington JA. A review of IR transmitting, hollow waveguides. Fiber Integr Opt 2000;19:211–227.
8. Rave E, Ephrat P, Goldberg M, Kedmi E, Katzir A. Silver halide photonic crystal fibers for the middle infrared. Appl Opt 2004;43(11):2236–2241.
9. Mackanos MA, Jansen ED, Shaw BL, Sanghera JS, Aggarwal I, Katzir A. Delivery of midinfrared (6 to 7-μm) laser radiation in a liquid environment using infrared-transmitting optical fibers. J Biomed Opt 2003;8(4):583–593.
10. Hooper BA, Maheshwari A, Curry AC, Alter TM. A Catheter for Diagnosis and Therapy using IR Evanescent Waves. Appl Opt 2003;42:3205–3214.
11. Chaney CA, Yang Y, Fried NM. Hybrid germanium/silica optical fibers for endoscopic delivery of erbium:YAG laser radiation. Lasers Surg Med 2004;34:5–11.
12. Altman RD, Kates J. Arthroscopy of the knee. Semin Arthritis Rheum 1983;13:188–199.
13. Drez D, Jr. Arthroscopic evaluation of the injured athelete's knee. Clin Sports Med 1985;4:275–278.
14. Altman RD, Gray R. Diagnostic and Therapeutic Uses of the Arthroscope in Rheumatoid Arthritis and Osteoarthritis. New York: American Journal of Medicine; 1983. p 50–55.
15. Phillon DP Collins, JV. Current status of fiberoptic bronchoscopy. Postgrad Med J 1984;60:213–217
16. Mitchell DM, Emerson CJ, Collyer J, Collins JV. Fiberoptic bronchoscopy: Ten years on. Br Med J 1980;2:360–363.
17. Nissen S. Coronary angiography and intravascular ultrasound. Am J Cardiol 2001;87(4A):15A–20A.
18. Nissen SE. Who is at risk for atherosclerotic disease? Lessons from intravascular ultrasound. Am J Med 2002;(Suppl 8A):27S–33S.
19. Tuzcu EM, De Franco AC, Goormastic M, Hobbs RE, Rincon G, Bott-Silverman C, McCarthy P, Stewart R, Mayer E, Nissen SE. Dichotomous pattern of coronary atherosclerosis 1 to 9 years after transplantation: Insights from systematic intravascular ultrasound imaging. JACC 1996;27(4):839–846.
20. Nadkarni SK, Boughner D, Fenster A. Image-based cardiac gating for three-dimensional intravascular ultrasound imaging. Ultrasound Med Biol 2005;1:53–63.
21. Bourantas CV, Plissiti ME, Fotiadis DI, Protopappas VC, Mpozios GV, Katsouras CS, Kourtis IC, Rees MR, Michalis LK. *In vivo* validation of a novel semi-automated method for border detection in intravascular ultrasound images. Br J Radiol 2005;78(926):122–129.
22. Jobsis FF. Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. Science 1977;198:1264–1267.
23. Parsons WJ, Rembert JC, Bauman RP, Greenfield Jr JC, Piantadosi CA. Dynamic mechanisms of cardiac oxygenation during brief ischemia and reperfusion. Am J Physiol 1990;259 (Heart Circ. Physiol. 28):H1477–H1485.
24. Parsons WJ, Rembert JC, Bauman RP, Greenfield Jr JC, Duhaylongsod FG, Piantadosi CA. Myocardial oxygenation in dogs during partial and complete coronary artery occlusion. Circ Res 1993;73(3):458–464.
25. Nilsson M, Heinrich D, Olajos J, AnderssonEngels S. Near infrared diffuse reflection and laser-induced fluorescence spectroscopy for myocardial tissue characterisation. Spectrochim Acta Part A-Mol Biomol Spectrosc 1997;51(11):1901–1912.
26. Manoharan R, Baraga JJ, Rava RP, Dasari RR, Fitzmaurice M, Feld MS. Biochemical-analysis and mapping of atherosclerotic human artery using FT-IR microspectroscopy. Atherosclerosis 1993;103(2):181–193.
27. Baraga JJ, Feld MS, Rava RP. Detection of atherosclerosis in human artery by midinfrared attenuated total reflectance. Appl Spectrosc 1991;45(4):709–710.
28. Rava RP, Baraga JJ, Feld MS. Near-infrared Fourier-Transform Raman spectroscopy of human artery. Spectrochim Acta Part A-Mol Biomol Spectrosc 1991;47(3–4):509–512.
29. Arai T, Mizuno K, Fujikawa A, Nakagawa M, Kikuchi M. Infrared absorption spectra ranging from 2.5 to 10 μm at various layers of human normal abdominal aorta and fibrofatty atheroma in vitro. Laser Surg Med 1990;10:357–362.
30. Casscells W, Hathorn B, David M, Krabach T, Vaughn WK, McAllister HA, Bearman G, Willerson JT. Thermal detection of cellular infiltrates in living atherosclerotic plaques: possible implications for plaque rupture and thrombosis. Lancet 1996;347(9013):1447–1449.
31. Colarusso P, Kidder L, Levin I, Fraser J, Arens J, Lewis EN. Infrared spectroscopic imaging: from planetary to cellular systems. Appl Spectrosc 1998;52(3):106A–119A.
32. Kodali DR, Small DM, Powell J, Krishna K. Infrared microimaging of atherosclerotic arteries. Appl Spectosc 1991; 45:1310–1317.
33. Edwards G, Logan R, Copeland M, Reinisch L, Davidson J, Johnson J, Maciunas R, Mendenhall M, Ossoff R, Tribble J, Werkhaven J, O'Day D. Tissue ablation by a Free-Electron Laser tuned to the Amide II band. Nature (London) 1994;371:416–419.
34. Awazu K, Nagai A, Aizawa K. Selective removal of Cholesterol Esters in an Arteriosclerotic region of blood vessels with a free-Electron Laser. Laser Surg Med 1998;23:233–237.
35. Holmes DR, Bresnahan JF. Interventional cardiology. Cardiol Clin 1991;9:115–134.
36. Linsker R, Srinivasan R, Wynne JJ, Alonso DR. Far-ultraviolet laser ablation of atherosclerotic lesions. Laser Surg Med 1984;4:201–206.
37. Hughes GC, Kypson AP, Yin B, St Louis JD, Biswas SS, Coleman RE, DeGrado TR, Annex BH, Donovan CL, Lanolfo KP, Lowe JE. Induction of angiogenesis following transmyocardial laser revascularization in a model of hibernating myocardium: a comparison of holmium:YAG, carbon dioxide, and excimer lasers. Surgical Forum L 1999;115–117.
38. Puliafito CA, Steinert RF, Deutsch TF, Hillenkamp F, Dehm EJ, Alder CM. Excimer laser ablation of cornea and lens: experimental studies. Ophthalmology 1985;92:741–748.
39. Cummings JP, Walsh Jr. JT Erbium laser ablation—the effect of dynamic optical properties. Appl Phys Lett 1993; 62:1988–1990.
40. Walsh PC, Gittes RF, Perlmutter AD, Stamey TS, editors. Campbell's Urology. Philadelphia: W. B. Saunders; 1986. p 510–540.
41. Segura JW. Endourology. J Urol 1984;132:1079–1084.
42. Powell PH, Manohar V, Ramsden PD, Hall RR. A flexible cytoscope. Br J Urol 1984;56:622–624.
43. Hoffman JL, Clayman RV. Endoscopic visualization of the supravesical urinary tract: Transurethral ureteropuleloscopy and percutaneous nephroscopy. Semin Urol 1985;3:60–75.
44. Benzie RJ. Amniocentesis, amnioscopy, and fetoscopy. Clin Obstet Gynecol 1980;7:439–460.
45. Rodeck CH, Nicolaides KH. Fetoscopy and fetal tissue sampling. Br Med Bull 1983;39:332–337.
46. Rauskolt R. Fetoscopy. J Perinat Med 1983;11:223–231.

47. Sleisenger MH, Fordtran JS, editors. Gastrointestinal Disease. Philadelphia: Saunders; 1983. p 1599–1616.
48. Sleisenger MH, Fordtran JS, editors. Gastrointestinal Disease. Philadelphia: Saunders; 1983. p 1617–1626.
49. Silvis SE, editor. Therapeutic Gastrointestinal Endoscopy. New York: Igaku-Shoin; 1985. p 241–268.
50. Huizinga E. On esophagoscopy and sword swallowing. Ann Otol Rhinol Laryngol 1969;78:32–34.
51. Shaheen N, Ransohoff DF. Gastroesophageal reflux, Barrett esophagus, and esophageal cancer: scientific review. JAMA 2002;287:1972–1981.
52. Spechler SJ. Clinical practice. Barrett's Esophagus. N Engl J Med 2002;346:836–842.
53. Sampliner RE. Updated guidelines for the diagnosis, surveillance, and therapy of Barrett's esophagus. Am J Gastroenterol 2002;97:1888–1895.
54. Wolfsen HC, Hemminger LL, DeVault KR. Recurrent Barrett's esophagus and adenocarcinoma after esophagectomy. BMC Gastroenterology 2004;4:18.
55. Jean M, Dua K. Barrett's Esophagus: Best of Digestive Disease Week 2003. Curr Gastroenterol Rep 2004;6:202–205.
56. Hunt RH, Waye JD, editors. Colonoscopy. London: Chapman & Hall; 1981. p 11–18.
57. Ohligisser M, Sorokin Y, Hiefetz M. Gynecologic laparoscopy, a review article. Obstet Gynecol Surv 1985;40:385–396.
58. Robinson HB, Smith GW. Application for laparoscopy in general surgery. Surg Gynecol Obstet 1976;143:829–834.
59. Devaux BC, Joly L, Page P, Nataf F, Turak B, Beuvon F, Trystram D, Roux F. Laser-assisted endoscopic third ventriculostomy for obstructive hydrocephalus: Technique and results in a series of 40 consecutive cases. Lasers Surg Med 2004;34:368–378.
60. Cram GP, Copeland ML. Nucl Instrum Methods Phys Rev B 1998;144:256.
61. Edwards G et al. Free-electron-laser-based biophysical and biomedical instrumentation. Rev Sci Instrum 2003;74:3207–3245.
62. Paparella MM, Shumrick DA, editors. Otolaryngology Philadelphia: W.B. Saunders; 1980. p 2410–2430.
63. Ballenger JJ, editor. Diseases of the Nose, Throat, Ear, Head, and Neck. Philadelphia: Lea & Febiger; 1985. p 1293–1330.
64. Steiner W. Techniques of diagnostic and operative endoscopy of the head and neck. Endoscopy 1979;1:51–59.
65. Vaughan CW. Use of the carbon dioxide laser in the endoscopic management of organic laryngeal disease. Otolaryngol Clin North Am 1983;16:849–864.
66. Pfalz R, Hibst R, Bald N. Suitability of different lasers for operations ranging from the tympanic membrane to the base of the stapes. Adv in Oto-Rhino-Laryngol 1995;49:87–94.
67. Schade G, Leuwer R, Kraas M, Rassow B, Hess M. Laryngeal morphometry with a new laser 'clip on' device. Lasers Surg Med 2004;34:363–367.
68. Future Trends in Medical Device Technology: Results of an Expert Survey. Available at http://www.fda.gov/cdrh/ost/trends/TOC.html.

Further Reading

Ponsky JL. Atlas of Surgical Endoscopy. St. Louis (MO): Mosby-Year Book; 1992.
Barkin J, O'Phelan C, editors. Advanced Therapeutic Endoscopy. New York: Raven Press; 1990.
Ovassapian A. Fiberoptic Airway Endoscopy in Anesthesia and Critical Care. New York: Raven Press; 1990.
Niemz M. Laser-Tissue Interactions: Fundamentals and Applications. 2nd ed. Berlin Heidelberg: Springer-Verlag; 2002.
Infrared Fiber Systems. Available at http://www.infraredfibersystems.com.
Polymicro Technologies, LLC. Available at http://www.polymicro.com.
Omniguide Communications, Inc. Available at http://www.omniguide.com.
LightLab Imaging, Inc. Available at http://www.lightlabimaging.com.
Given Imaging. Available at www.givenimaging.com.
Intuitive Surgical, Inc. Available at www.intuitivesurgical.com.

See also ESOPHAGEAL MANOMETRY; FIBER OPTICS IN MEDICINE; MINIMALLY INVASIVE SURGERY.

ENGINEERED TISSUE

GREGORY E. RUTKOWSKI
University of Minnesota-Duluth
Duluth, Minnesota

INTRODUCTION

History of Tissue Engineering

In 1988, researchers gathered at the Granlibakken Resort in Lake Tahoe, CA under the sponsor ship of the National Science Foundation (NSF) to develop the fundamental principles of tissue engineering as an emerging technology. Based on an earlier proposal by Dr. Y.C. Fung to develop an Engineering Research Center focused on the engineering of living tissues, NSF held several meetings that led to the decision to designate tissue engineering as a new emerging field. A formal definition was finally agreed upon at the Granlibakken workshop. Based on this meeting, tissue engineering is defined as "the application of the principles and methods of engineering and the life sciences toward the fundamental understanding of structure–function relationships in normal and pathological mammalian tissues and the development of biological substitutes to restore, maintain, and improve function" (1). This was further refined in 1992 by Eugene Bell who developed a list of more specific goals:

1. Providing cellular prostheses or replacement parts for the human body.
2. Providing formed acellular replacement parts capable of inducing regeneration.
3. Providing tissue or organ-like model systems populated with cells for basic research and for many applied uses such as the study of disease states using aberrant cells.
4. Providing vehicles for delivering engineered cells to the organism.
5. Surfacing nonbiological devices (2).

These discussions eventually culminated in the pioneering review article by Langer and Vacanti in 1993 (3). The general strategies to create engineered tissue would include the isolation of cells or cell substitutes, the use of tissue-inducing substances, and development of three-dimensional (3D) matrices on which to grow tissue.

Much of tissue engineering owes its beginnings to reconstructive surgery and internal medicine. In the sixteenth century, Gaspare Tagliacozzi developed a method for nasal reconstruction using flaps of skin taken from the arm and grafted onto the injury site (4). With the scientific development of the germ theory of disease and the sterile techniques that were introduced, modern surgery became established as a means to treat patients with internal injuries. In the late nineteenth century, surgeons used veins and arteries as conduits to enhance the nerve regeneration (5). World War I saw improvements in reconstructive surgery as doctors were able to hone their skills due to the number of soldiers injured in battle. Reconstructive surgery had it limitations in terms of the availability of biological material. By the 1940s, much progress had been made in understanding the function of the immune system is accepting tissue from a donor. This eventually led to the first successful organ transplant (kidney) in 1954. The next 50 years, would see tremendous advances in organ transplants as well as in immunosuppressive drug therapy.

An alternative to organ transplant has been the development of artificial devices to mimic biological function. The mid-nineteenth century also saw the rise in the use of prosthetic limbs that would initiate the use of artificial devices to replace biological functions. Artificial limbs were used as far back as the Dark Ages to assist knights heading off to battle. The intervening centuries saw improvements over such devices through the use of stronger, lighter materials, and a better understanding of biomechanics (6). Besides limbs, artificial devices have also been invented to replace the function of certain internal organs. The dialysis machine was created in the 1940s to assist patients with acute renal failure. In 2001, an implantable artificial heart was first used in a human. While many advances have been made in artificial devices, some of the drawbacks include the breakdown of the artificial materials, a lack of interaction with the human body, and the inability to self-renew.

The modern era of tissue engineers seeks to overcome the limitations of reconstructive surgery, organ transplantations, and prosthetic devices by creating functional, completely biocompatible tissues and organs. Since the late 1980s, the field of tissue engineering has grown exponentially and continues to draw scientists from diverse fields, from the biological and medical sciences to engineering and materials science.

THEORY

Tissue engineering adds to the modern health care system by providing the tools to assist in the repair of tissue and organs damaged by injury and disease. An exact replica of the tissue could potentially be grown in the lab and later inserted into the patient. Alternatively, precursors may be placed in the body with the expectation that it will develop into fully formed functional tissue. Also, devices may be implanted into the body to encourage the regeneration of already existing tissue in the body.

Engineered tissue is created by combining relevant cells and chemical factors within a 3D matrix that serves as a scaffold (Fig. 1). Sterile cell culture techniques allows for the expansion of cells *in vitro* to obtain sufficient quantities for use in engineered tissue. Cells can originate from the patient, another donor, or even animal tissue. As our understanding of biochemical cues during development expands, tissue formation can be better controlled through the delivery of pertinent growth factors or through the interaction of complex multiple cell cultures. Scaffold materials may be biological (collagen, extra cellular matrix) or synthetic. Synthetic materials can be designed to mimic the extra cellular matrix and may be infused with chemical factors to support tissue regeneration. Since the discovery and availability of synthetic polymers in the 1940s, scientific advances have made these materials biodegradable and biocompatible.

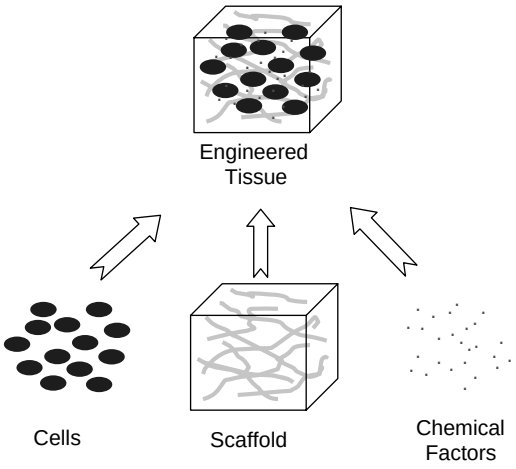


Figure 1. Basic components of engineered tissue.

Cellular Processes

During the formation of tissue, either during development or in the lab, cells undertake many different processes. In order to maintain the integrity of the tissue, the cells tissue must adhere to others as well as to the surrounding matrix material. Certain cells must migrate through the 3D space in order to properly position themselves within the tissue. Once in position, they must also continue to multiply to provide adequate tissue growth. Vascularization and innervation of the tissue is required to complete integration of the engineered tissue with its surroundings. Tissue derived from donor material also has immune concerns.

Development. Engineered tissue is best created by mimicking the processes that occur during development (7). Following fertilization, the zygote divides until the blastocyst is formed. The early embryonic tissue is comprised of two cell phenotypes: epithelial cells and migratory mesanchymal cells. The transformation between these cells is regulated by growth factors, the extra cellular matrix, and intracellular signaling. As the embryo develops, these cells will become diversified as they commit to various tissue forms. These cells will eventually differentiate into the various cell types found in the body.

Morphogenesis describes the cascade of events that leads to the spatial pattern formation of the developing

embryo. Regeneration of tissue mimics this process. Specific chemical factors, called morphogens, provide information on the pattern via diffusion. The combination of various morphogens due to diffusion leads to complex pattern formation. One well-studied example is the development of bone. Several bone morphogenic factors have been isolated and observed to affect the bone and cartilage morphogenesis and formation. Also, BMPs have been implicated in the development of tissue as diverse as skin, eye, heart, and kidney. These morphogens have the potential to assist in the engineering of bone tissue.

Adhesion. Tissue is held together by the adhesion processes between different cells and the cells and extracellular matrix. The cell–cell interactions consist of tight, anchoring, and communication junctions. Tight junctions are composed of transmembrane proteins that form connections between cells. More contact points will decrease the permeability between the cells so that the tissue can become essentially impermeable. This is typically found with epithelial cells, as in the intestinal, reproductive, and respiratory tract, where the barrier that is formed is essential to function.

Anchoring junctions are loosely held connections that take advantage of the cytoskeleton. With adherens junctions, actin filaments of the cytoskeleton are connected and linked to integrins on the cell exterior. These integrins form focal contacts that interact with cadherins found on the surface of other cells. The focal contacts can also interact with extracellular domains. With the involvement of the cytoskeleton, these junctions can also affect cell function by providing a mechanism to signal changes in cell growth, survival, morphology, migration, and differentiation.

Desmosomes are similar to adherens junctions, but they involve intermediate filament protein, such as vimentin, desmin, and keratin. These junctions connect with other cells via cadherins. A similar junction called the hemidesmosome behaves in the same manner, but connects with basal lamina proteins via integrin.

Communication junctions are those that provide direct communication between cells. These are large proteins that form pore structure that connects the two cells. These large pores (connexons) allow the transport of molecules between cells. These junctions are commonly found between neurons for rapid signal conduction.

In order to connect with other cells or the extracellular matrix, the junction proteins must interact with some receptor molecule. Integrins will bind with the amino acid sequence arginine-glycine-aspartic acid (RGD). This sequence is found in several proteins, such as collagen and fibronectin and these peptides can be incorporated onto other surfaces in order to improve cell adhesion. Cadherin–cadherin binding is mediated via Ca^{2+}. Without the presence of Ca^{2+}, the connection is subject to proteolysis. The Ig-like receptors contain motifs found in immunoglobulins. These receptors can interact with neural cell adhesion molecule (N-CAM) and are present during development. Finally, selectins are specific receptor types that are expressed during the inflammatory response. The lectin domain found in oligosaccharides on neutrophils allows these cells to interact with endothelial cells along the surface of blood vessels.

Migration. During embryogenesis, diffusible factors and ECM composition are important factors in the pattern formation of tissue development. The migration of cells is necessary for the formation of tissue not just for development, but also during regeneration. Cell migration is also observed in other body functions. Angiogenesis, or blood vessel formation, involves the migration of endothelial cells into new tissues. For immune responses, B and T cells patrol the body ready to attack invaders. Tumor invasion and, more importantly, metastasis relies on the migration of cancer cells into other parts of the body. Controlling migration can help to control the spread of cancer.

Signals from various factors can lead to the release of cells from their contact with other cells or with the extracellular matrix. Once released, different mechanisms can affect the migration pattern of the cell depending on the cell type. Cells may move in random directions. The movement can be modeled by Brownian motion. In this case, motion is due to collisions with other particles. This motion is characterized by the time between collisions, the mean free path (average distance before hitting another particle) and the average speed. The characteristics are dependent on the density of the particles.

Of more relevance is the issue of directed migration. Direct migration depends on some sort of gradient. In response to some kind of stimulus, cells may move toward or away from the source of the stimulus. The stimulus may affect speed, direction, or both. Chemotaxis is the general term describing the response of cells to a chemical gradient. The strength of the effect is dependent on the absolute concentration as well as the steepness of the gradient.

Growth. Mitosis is a tightly controlled process to regulate cell growth and depends on signals from the cell's environment. During mitosis, the deoxyribonucleic acid (DNA) is replicated and copies are separated as the cells divide into two exact copies. This process repeats itself depending on the extracellular signaling and the properties of the cell itself. Certain cells, such as pancreatic beta cells, are locked in the cell cycle and mitosis is arrested. In order to overcome this barrier, stem cells can be used to generate new differentiated cells. Also, cells may be converted into a precursor cell form that can undergo proliferation before switching back to the differentiated form.

Cell growth relies on the availability of essential nutrients. As the need for nutrients increases during cells proliferation, the availability becomes reduced barring some mechanism for distributing the nutrients to the growing tissue. Distribution of nutrients occurs naturally in the human body in the form of the network of blood vessels. The lack of such a vasculature for engineered tissue limits the effective size that tissue can grow.

Vascularization. For most engineered tissue to become integrated into the human body, it must become vascularized by the body's own blood vessel network. The body provides a natural mechanism for neovascularization by the process of wound healing. When tissue has been

damaged, a cascade of events is initiated to block the bleeding wound and encourage healing. Platelets, fibrin, and fibronectin first form a mesh plug. Mast cells release chemotactic agents to recruit other cells as part of the inflammatory response. Keratinocytes migrate to the site of the injury and begin to proliferate. Vascular endothelial growth factor (VEGF) and fibroblast growth factor (FGF) are released to encourage blood vessel formation. Once the new tissue has been vascularized, remodeling of the tissue occurs to complete the healing process.

For engineered tissue, two options are used to ensure vascularization and promote tissue growth and integration. The engineered tissue can be designed to contain a vascular network *in vitro* that would then be connected to the body's own network when it is implanted. This presents a significant engineering challenge in trying to create several different types of engineered tissue in one construct simultaneously. An alternative method is to engineer the tissue to recruit blood vessels from the body's existing framework. This has been accomplished though the controlled release of VEGF from the biodegradable support matrix of the engineered tissue (8).

Innervation. To become fully integrated into the body, certain tissues and organs must also reconnect with the body's own nervous system. Several organs of the body form connections with the sympathetic nervous system. These connections are important to regulation of the organ. Skeletal muscle tissue makes sensory and motor connections via the peripheral nervous system. In any case, the cells of these engineered tissues need to make synaptic connection with the axons of the relevant neurons. Because the new tissue is being engineered to replace that which has been lost to injury or disease, the neural framework may not be available for integration. If it is present, the neurons may be encouraged to regenerate toward the tissue and form new connections. Another complex, but theoretically possible, option may be to engineer the tissue with neural connections that can be later integrated into the existing nervous system. Also, artificial devices may be integrated into the system to provide the appropriate control of the tissue function.

Immune Concerns. As with organ transplants, engineered tissue also has concerns of rejection by the immune system. The major histocompatibility complex I (MHC I) present on the cells of the engineered tissue are recognized by the T cells of the immune system as a foreign body. This would eventually lead to cell lysis. The primary means of preventing rejection, though, is the use of immune suppressant drug therapy. While this may be acceptable, for a patient receiving a life saving organ transplant, it is not for those receiving engineered tissue. Certain tissues, such as cartilage, may not interact with the immune system. Metabolic tissues that only interact chemically can be physically separated from the environment. Some engineered tissues may contain cells that are only temporary until the body's own cells can take over. In these cases, immune suppressant drug may be a viable option. For most other tissues, methods are being developed for side stepping the immune system.

Because the cells of the immune system are formed within the bone marrow, one method is to transplant the bone marrow from the cell and tissue donor along with the organ. The donated bone marrow can form a chimera with the patient's existing bone marrow to allow the adaptation of the immune system to the new tissue (9).

At the molecular level, the antigen of the foreign cell can be blocked or completely eliminated. In order to block the antigen, a fragment of antibody to the antigen can be added to the tissue to mask the foreign cells from the patient's own immune system (10). This effect is temporary as the fragments will eventually separate from the antigen. Another drawback is that is may not counter all the mechanisms of immune response. The antigen can also be removed by placing an inactive form of the antigen gene into the cells (11). A gene can also be added to the cells to produce a protein that will inhibit rejection (12). Finally, oligonucleotides can be added to hybridize with either ribonucleic acid (RNA) or DNA in order to inhibit the transcription or translation of the antigen molecule (13).

Microencapsulation is a means of physically separating the cells from the environment. For this method, cells are surrounded by a porous synthetic material. As long as the membrane coating remains intact, the cells are isolated from the immune system. The pore size can be adjusted to allow chemical interaction with the environment while preventing cellular interaction. The encapsulation should allow nutrients to permeate through the membrane and reach the cells. This was first used in clinical studies for the encapsulation of xenogenic pancreatic islets (14).

Cell Types. While the response of the immune system is of great importance to the success of the engineered implant, the source of cells and their application will determine the best method for immunomodulation. Cells may come from the patients themselves (autologous), from a human donor (allogeneic), or from another species (xenogeneic). Each type has its own advantages and disadvantages.

Autologous cells are derived from the patient, expanded in culture, and then placed back into the patient. These cells are completely biocompatible with the patient and this eliminates the need for any immune modulation. Genzyme, for example, has developed a successful protocol for the repairing articular cartilage. One drawback to using autologous cells is that it requires a period of time to expand the cells. This would not be acceptable for patients needing an immediate tissue replacement. As a result, the engineered tissue does not have off-the-shelf availability. Depending on the tissue and the amount of damage, the amount of cells that may be harvested from the patient may be insufficient to form tissue.

Allogeneic cells can help to overcome some of the drawbacks to autologous cells because the cells come from donor sources that may be pooled together. This can provide off-the-shelf availability, but at the expense of immune rejection. The cells can be engineered to become immune acceptable or immunosuppressive drug therapy may be used. These cells are also well suited for tissues that do not interact with the patient's vasculature system or may only be used until the native tissue regenerate.

Advanced Tissue Science developed a skin replacement tissue (Dermagraft) with cells derived from circumcision surgeries.

Xenogeneic cells are derived from nonhuman species. An animal source can provide an unlimited amount of cells, but they provoke an acute immune response within minutes of implantation into the body. One useful application of such cells is the encapsulation of pancreatic islets (14,15). The membrane can be adjusted to allow the islets to regulate blood sugar and insulin levels while protecting the cells from the immune system. Another drawback to xenogeneic cells is the threat of transmission of animal viruses and well as endogenous retroviruses that may interact with the patient.

Cells that have been isolated may be modified to alter their characteristics before being incorporated into engineered tissue. Primary cells that are subcultured eventually become cell lines. These cell lines may be finite or continuous. The cells may also be normal or transformed compared to the primary cell from which it is derived. Transformation is associated with genetic instabilities that may lead to a change in the phenotype. These changes may impart different growth characteristics for the cell, such as immortalization, anchorage independence, and overall growth rate. The genetic material may be altered to effect gene expression. Changes in protein expression may effect the secretion of chemical factors or even the formation of extracellular matrix. In the worst case, cells may become cancerous and prove to be detrimental to the patient. Examples of stable, well-characterized cell lines used in tissue engineering include HEK-293 for nerve growth factor secretion (16), and 3T3-L1 for adipose tissue formation (17).

Primary cells as well as cell lines can be artificially transformed by introducing new genes by the process of transfection. In this process, a new gene is carried by some vector, such as a liposome or virus, to the host cells. The gene is transferred into the cells and delivered into the nucleus where it can be transcribed. This method allows the creation of cells with desirable characteristics for engineered tissue. Such cells may secrete growth factor that will enhance tissue formation. Examples of engineered tissue utilizing transfected cells include insertion of hBMP-4 gene in bone marrow stromal cells for bone tissue engineering (18) and aquaporin gene transfection in the LLC-PK1 cell line for a bioartificial renal device (19), secretion of neuronal growth factors from fibroblasts to enhance axonal regeneration (20).

Regardless of the source, primary cells have limitations depending on their age and genetic makeup. Adult cells may only have a limited number of cell divisions before they reach senescence or even become cancerous. Transformed cell lines may allow for continuous growth, but without proper control can become a problem for the patient. These cells are typically used to enhance regeneration of existing tissue, but do not become integrated into the body. Because of these issues, a universal cell that can develop into multiple tissue types has an infinite capacity to proliferate without loss of function, and if immune acceptable would be ideal for engineered tissue. Stem cells come closest to meeting these criteria.

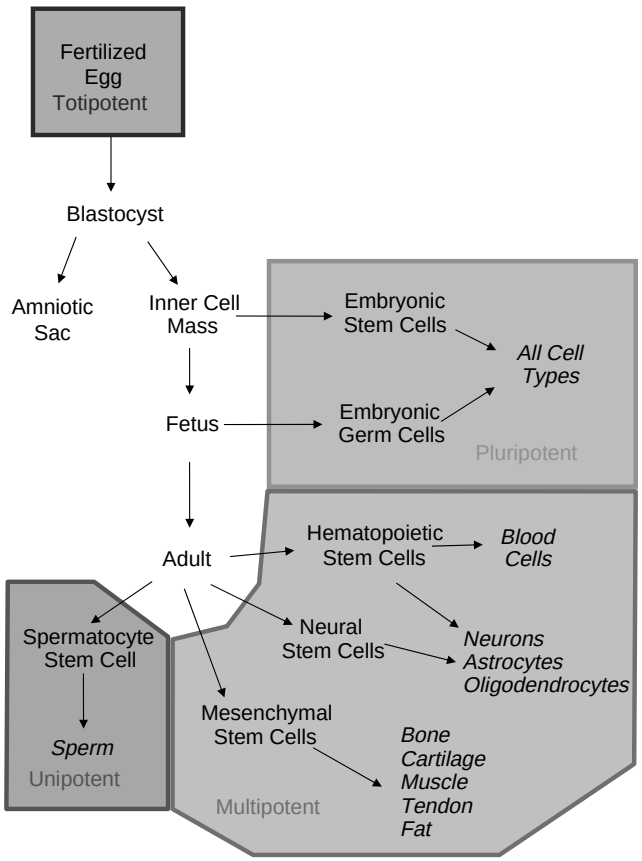


Figure 2. Stem cell potency. Relationship of stem cells to various stages of development. Cells and tissue derived from stem cells are shown in italics.

Stem Cells. The potency of stem cells is defined by their ability to differentiate into one or more cell genotypes (see Fig. 2). Unipotent stem cells give rise to only one cell type (*ex. Spermatogonia*). Multipotent stem cells are more functional and can differentiate into multiple cell types. Embryonic stem cells are considered pluripotent in that they can form all the cell types found in an embryo and adult. The fertilized egg develops into an embryo as well as the amniotic sac and is considered totipotent. Embryonic stem cells hold the most promise for engineering tissue, but more work must be done on the basic principles of stem and progenitor cell biology as well as on the control of differentiation.

Multipotent stem cells obtained from embryonic or adult sources are considered the optimal choice for tissue engineering applications because of their ability to form multiple tissue type. Hematopoietic stem cells (HSCs) are the most well characterized of stem cells. While isolated from bone marrow, HSCs can differentiate to form skeletal muscle, cardiac muscle, hepatocytes, endothelial cells, and epithelial cells. Multipotent stem cells can be found in several other tissues, such as brain, heart, pancreas, retina, liver, and lung, and skin.

To take full advantage of the functionality of stem cells, more work needs to be done to elucidate the mechanisms for differentiation. Once established, stem cells can be encouraged to differentiate into the required cell type.

The mechanism may rely on an environmental cue as to whether they are physical contact, chemical, or chemotactic in nature. The ability of stem cell to proliferate, integrate, and differentiate also depends on the methods of identifying, isolating, and expanding. Protocols for stem cell culture need to be developed and optimized to ensure the cells achieve their full potential.

Since some stem cells are harder to obtain than others, getting stem cells to transdifferentiate from one form to another would allow for further flexibility. Evidence suggests that bone marrow stromal cells may convert to neural stem cells (21) and neural stem cells to hematopoietic stem cells (22). Recent evidence, though, points to the fusion of stem cells with other cells instead of true transdifferentiation (23).

Another source for stem cells is from embryos discarded from *in vitro* fertilization clinics. Human embryonic stem cells have an even greater capacity to form tissues than the multipotent stem cells. These cells have been isolated from humans and protocols for long-term cultures have successfully been developed (24). The use of embryonic stem cells has raised many ethical concerns because of the destruction of the fetus from which they are derived. This has led to legislation tightly regulating their use. To avoid these concerns, several proposals have been suggested to obtain embryonic stem cells without compromising the potential for life (25).

Stem cells have been used successfully for tissue engineering applications. Stem cells have been seeded on scaffolds to form cartilage, small intestine, and bone (26). Neural stem cells have also been used for repair of spinal cord injuries (27). Embryonic stem cells have also been seeded onto scaffolds where the native vasculature integrated into the engineered tissue (28).

Instead of preseeding cells on scaffolds, stems cells may also be injected directly to the site of injury to promote tissue regeneration. Clinical studies in nonhuman primates have shown that neural stem cells can enhance repair of spinal cord injuries (29). Stem cells have also been used to regenerate damaged cardiac muscle tissue (30). Embryonic stem cells may also be used to regenerate pancreatic beta cells in order to alleviate diabetes (31). Mesanchymal stem cells also show some promise for the repair of articular cartilage for those suffering from osteoarthritis (32).

Scaffolds

Primary cells will form monolayer cultures when dissociated from tissue. In order to encourage the cells to form engineered tissue, they must be placed in a 3D matrix that acts as a support scaffold. Ideally, these scaffolds should be compatible with the cells as well as being biodegrade. The scaffold should have a high porosity to ensure the diffusion of nutrients and other chemical factors as well as provide room for cell migration and proliferation. The material should have a higher surface area to ensure adequate room for cell attachment. The matrix should maintain its structural integrity until tissue integration has been completed. For certain applications, the final construct may need to be formed into a specific 3D shape (see Fig. 3).

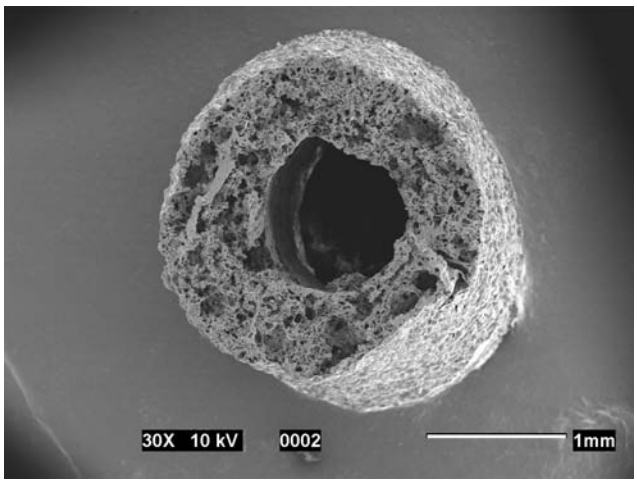


Figure 3. Example of a scaffold for engineered tissue. This scaffold is designed for nerve regeneration. The cylindrical conduit was manufactured by using the solvent casting method (see below) with particle leaching. Poly(vinyl alcohol) rods were coated with a suspension of salt crystals in a poly(lactic acid) (PLA)–chloroform solution. Dissolution of the rods permitted the formation of the large central pore through which nerve tissue could regenerate. A porosity of 85% was obtained by controlling the volume fraction of salt crystals.

For tissue engineering, scaffolds may come in different forms. An acellular matrix can be used to recruit cells from the host tissue (33). With this form, the immune response is not a concern though inflammation may occur. For most applications, cells must be seeded onto the scaffold to ensure successful tissue formation. Cells can be seeded onto collagen gels that mimic the naturally occurring extracellular matrix (34,35). Cells can also be encouraged to self-assemble in culture. These systems are ideal for tissue that forms simple shapes, such as sheets and cylinders (36,37). Biodegradable polymers provide a flexible means of creating scaffolds with properties desirable for tissue formation.

Natural Polymers. Macromolecules found with the extracellular matrix provide chemical and structural support for living tissue. Because of their function in tissue formation, these natural polymers provide an excellent scaffold for engineering new tissue. These natural polymers have some limitations in their application. The mechanical properties and degradation rates cannot be controlled as well as synthetic polymers (38). Also, material derived from donors may elicit an immune response (39). Many of these limitations can be overcome by chemical modification of the materials as well as creating composite materials of natural polymers or natural–synthetic polymers.

Collagen is a large family of proteins that make up much of the extracellular matrix. At least 19 different types have been isolated. Fibrillar collagen is comprised of several collagen types that combine to form extended structures that assist is maintaining tissue integrity. The relative amounts of collagen are characteristic of the tissue type. Collagen has been used for cartilage (40), nerve regeneration (41), gallbladder engineering (42), corneal tissue (43),

and skin (44). It can be cross-linked to enhance mechanical properties (45). Magnetically aligned collagen fibers can also be used to enhance nerve regeneration (46).

Matrigel is comprised of proteins found in the basal lamina. The basal lamina, or basement membrane, is a specialized area of the extracellular matrix found at epithelial–stromal boundaries. It is composed of collagen, laminin, nidogen, and perlecan, (a heparin sulfate proteoglycan). It has been used for spinal cord repair (47) vascular network formation (48), and cardiac tissue (49).

Alginates are synthesized naturally by brown seaweeds as well as some bacteria. They are comprised of units of mannuronic and guluronic acid. The proportion and repeat unit sequence is dependent on the organism of origin. Alginates will form a gel in the presence of calcium ions. Alginates have been used for cartilage (50), cardiac tissue (51), and liver engineering (52).

Fibrin is formed from the polymerization of fibrinogen that occurs during the wound healing process. The process is controlled by the presence of thrombin. Fibrin has been used for cardiovascular repair (53), bone (54), cartilage (55), and skin (56). Like collagen, the fibrin chains can be cross-linked to alter the characteristics of the matrix (57).

Chitosan is derived from chitin, a polysaccharide that comprises the exoskeleton of crustaceans. The amide group on chitin is replaced by an amine group. Since chitosan is not protein based, it does not elicit an immune response. Chitosan has been used for the engineering of cartilage (58) and bone (59). It has also been combined with other materials to create novel composite matrices (60,61).

Besides isolating single components as scaffolding material, intact decellularized tissue can also be used. In this method, tissue is obtained from a donor and the cells are destroyed chemically using compounds, such as sodium dodecyl sulfate. This methods is currently used to engineer blood vessels (62) and heart valves (63). Decellularized tissue may also be applied to the engineering of other types of tissues, such as utilizing porcine skin for urinary tract reconstruction (64).

Synthetic Polymers. While natural polymers are an excellent choice for scaffolds because of their biocompatibility, synthetic polymers offer more flexibility. Composites with natural polymers can also be made to take advantage of their properties as well. Biodegradable polymers also have the additional benefit of decomposing into nontoxic metabolites that can be removed from the body as waste. The molecules can be modified to obtain mechanical properties and appropriate degradation rates that are well suited to the specific application.

The most common polymers in use are PLA and poly(glycolic acid) (PGA). These are the first two polymers to obtain approval by the Food and Drug Administration (FDA) for use in medical implants. During degradation, the polymer is hydrolyzed to form lactic and glycolic acids. Since PLA is more hydrophobic than PGA, it has a slower degradation rate. As such, the degradation rate of the polymers can be easily adjusted by changing the ratio of lactic acid to glycolic acid within copolymers of PLA and PGA, poly(lactic-*co*-glycolic acid) (PLGA). The degradation rate is also affected by the overall molecular weight and extent of crystallinity of the polymer. During degradation, the polymer swells with water. The nonspecific interactions between the water and polymer lead to hydrolysis at random locations along the polymer chain. This leads to bulk degradation of the matrix that can be detrimental to the release profile of drugs. The hydrophobicity of PLA also makes the matrix less amenable for cell adhesion. Adsorption of proteins, such as laminin (65), to the matrix surface encourages cell adhesion. Amines can also be incorporated into the polymer chain to control protein and cell attachment (66), because many types of tissue have been engineered using PLA, PGA, as well as combinations of their stereoisomers and copolymers (see Ref. 67 for a review).

The limitations due to bulk degradation can be overcome using matrices based on polyanhydrides. These polymers degrade only at the surface. When used for drug delivery, the release profiles can be affected just by altering the shape of the matrix. For tissue engineering, these polymers provide a mechanism for renewing the surface to allow new areas for cells to grow. As the surface degrades, cells will slough off providing a fresh surface for cell growth. One example of such a polymer, recently approved by the FDA for use in the treatment of glioblastoma multiformae, is derived from bis(*p*-carboxyphenoxy propane) and sebacic acid.

While PGA, PLA, and their copolymers provide much flexibility for developing scaffolds that have particular degradation properties, they are fairly brittle. This makes them unsuitable for use in certain engineered tissues, such as tendon and ligament, where mechanical stresses may be present. Polycaprolactone (PCL) is also a polyester, but its longer monomer unit provides greater elasticity. The polymer and degradation product is nontoxic, which has led to its FDA approval for use as a long-term contraceptive implant (68). Other elastomeric polymers are poly-4-hydroxybutyrate and polyhydroxyalkanoate, which have been used in various engineered tissues (69).

Certain polyesters have been developed to degrade into biocompatible products that make them suitable for drug delivery and tissue engineering applications. Poly(propylene fumarate) (PPF), for example, will degrade into fumaric acid, a natural component of the Kreb's cycle, and 1,2-propandiol, a common drug diluent. Mechanical properties can be improved through the use of chemical cross-linkers or certain ceramic composites (70). A tyrosine-based polycarbonate is another polyester that has been used in bone engineering (71). This osteoconductive polymer has excellent structural properties and has been shown to promote bone growth (72).

Like polyanhydrides, poly(ortho esters) can be designed for surface degradation. Their use in drug delivery (73) also make them suitable candidates for tissue engineering. Some poly(ortho esters) may degrade into acidic byproducts that can autocatalyze the degradation process, which may be useful for systems where fast degradation of the scaffold is desired.

In contrast to most hydrocarbon-based polymers, poly(phosphazenes) consist of a phosphorous and nitrogen chain. These polymers undergo hydrolysis to form

phophate and ammonium salts. Ethyl glycinate substituted poly(phosphazenes) have shown promise for osteoblast attachment (74).

Polyurethanes have long been used for medical implants because of their good biocompatibility and mechanical properties. They have been used in long-term implants, such as cardiac pacemakers and vascular grafts. Though not degradable, these polymers can be cross-linked with degradable compounds to create scaffolds for engineered tissue (75). The main disadvantage is the toxicity of the degradation byproducts especially for cross-linkers based on diisocyanates.

Poly(amino acids) have good potential as a scaffold material because they are biocompatible and release amino acids as their degradation product. Poly-L-lysine is a common example that is adsorbed to surfaces in order to improve cell adhesion. Enzymatic degradation makes the breakdown of the polymer difficult to control. These polymers also have high reactivity and moisture sensitivity. Along with being expensive to produce, these materials have limited use for engineered tissue. An alternative is to create a pseudo-poly(amino acid), where the amino group in the polymer backbone is replaced with another nonamide linkage. This can improve the stability and mechanical properties of the polymer (76).

Hydrogels. Hydrogels are a subcategory of biomaterials defined by their ability to retain water within their polymeric matrix. Because of the high water content, hydrogels have mechanical properties similar to that of soft tissue. This may limit the application of hydrogels to certain tissues, but the environment closely simulates the environment of native tissue. Hydrogels are created by the cross-linking of water soluble polymers while in an aqueous environment. The cross-linking can be initiated chemically or via exposure to light of a particular wavelength. Natural polymers, such as fibrin and collagen, can be used to create hydrogels. For synthetic hydrogels, polyethylene glycol is a popular choice because of its biocompatibility, hydrophilicity and customizable transport properties (77).

Cells can also be entrapped within the matrix during the gelling process that permits a more uniform distribution. Cell entrapped in hydrogels include chondrocytes (78), fibroblasts (79), and smooth muscle (80). Photoinitiated cross-linking can be used to create cell–matrix composites *in situ* (81). In such systems, cells still maintain their viability (82). This process can be used to create a cell–polymer matrix that fits exactly into the shape of the tissue defect.

Scaffold Fabrication. Once a biomaterial has been chosen that has the properties crucial to the particular tissue to be engineered, it must be fabricated into an appropriate matrix in which the cells can survive and form the tissue. Several factors, such as porosity and pore structure, surface area, structural strength, and shape, are relevant to the design of a suitable scaffold. In general, a high porosity is important to the formation of tissue because it provides space for the cells to grow, as well as allows nutrients to diffuse into the matrix and promote cell survival (3). For certain situations, though, the porosity should be optimized to ensure that growth factors important to tissue regeneration are retained within the matrix (83).

The strength of the scaffold is important to ensure that the scaffold will protect the regenerating tissue until it becomes integrated into the body. Also, some engineered tissue, such as bone and cartilage, may require a strong scaffold in order to retain integrity while functioning under physiological loading conditions. The structural strength is dependent on the mechanical properties of the polymer as well as the processing of the scaffold. A high surface area will ensure adequate contact for the cell adhesion. Depending on the polymer, the surface may have to be modified to improve adhesion. Depending on the application, the scaffold may need to be processed to form a particular 3D shape. For example, cylinders can be used as vascular grafts and for nerve regeneration.

Several processing techniques are available for the fabrication of scaffolds (Table 1). These techniques allow for the control of porosity and pore structure as well as the contact area for cell adhesion. In fiber bonding, a polymer is dissolved in a suitable solvent and fibers from another polymer are suspended within the solution (84). The solvent is then removed by vacuum drying. Heat is slowly applied to the composite material to cause the fibers to bond to one another. The other polymer is again dissolved leaving behind the bonded fiber matrix. To ensure success in this process, solvents must be chosen that do not dissolve the fibers and the fibers must have a melt temperature lower than that of the matrix polymer.

A similar method, called particle leaching, involves the incorporation of particles suspended within the polymer solution (85). As with fiber bonding, the particles are locked within the polymer matrix after vacuum drying is done to remove the solvent. In this method, though, the particles

Table 1. Summary of Various Scaffold Fabrication Methods

Fabrication Method	Characteristics
Solvent casting	No thermal degradation
	Residual solvent may harm cells
Fiber bonding	High porosity
	Thermal degradation
	Limited solvent–polymer combination
Particle leaching	Porosity easy to control
	Entrapment of particle with matrix
	Brittle foams
Gas foaming	No organic solvents
	Noncontinuous pores
Freeze drying	Small pore sizes
	Low temperature
Phase separation	For entrapment of small bioactive molecules
	Low temperature
Extrusion	Long fibers
	Thermal degradation
Membrane lamination	Three dimensional shapes
3D printing	Slow processing
	Control of shapes
In situ polymerization	Limited polymer–cell combinations
	Injectable, shape to fit defect

are removed via dissolution leaving behind a porous matrix. Typically, salt or sugar crystals, which are soluble in water, are suspended within a relatively hydrophobic polymer. The porosity can be controlled by altering the amount of particles suspended in the matrix. If the amount of particles is too low, they may become trapped within the polymer and remain undissolved. The foams that are formed tend to be brittle and also may require prewetting with ethanol to promote fluid infiltration. Instead of suspending the particles in a solution, they can be placed into a polymer melt that is subsequently cooled and the particles later dissolved away. A major drawback to this process is the thermal degradation of the polymer that can greatly affect its mechanical properties.

With gas foaming, carbon dioxide is added to the solid polymer at a very high pressure so that it infiltrates the matrix (86). When the pressure is rapidly dropped, the gas will expand within the polymer creating a porous foam. This method eliminates the need for an organic solvent whose presence may be detrimental to cell survival and tissue formation. One disadvantage is that the pores may not connect. This can be overcome by using this method in combination with particulate leaching.

Freeze drying can also be used to create porous polymer matrices (87). A polymer dissolved in an organic solvent can be mixed with water to form an emulsion. The emulsion is then quenched in liquid nitrogen and the water is removed by freeze drying. This method can create a matrix with 90% porosity and very high surface area, but the pore size may be too small.

Bioactive molecules may be added to the polymer matrix so that their release can enhance tissue formation. In many cases, though, the bioactive molecule may be incompatible with the organic solvent or thermal conditions used in the processing of the polymer matrix. Phase separation may be used to overcome this problem (88). In this method, the bioactive molecule is dispersed within the polymer solution and the solution is cooled until two liquid phases are formed. The two immiscible liquids are quickly frozen and the solvent is removed by sublimation. Removal of the solvent-rich phase leads to the highly porous structure while the bioactive molecule remains trapped within the solid polymer phase. The cold temperatures used in this process ensure that the bioactive molecule is not adversely affected. This method works well with small molecules, but can be difficult with large proteins.

For certain structural applications, such as bone, the mechanical strength of the polymer matrix is important to the success of the engineered tissue. In these cases, the matrix should have a high compressive strength to help the tissue maintain its integrity until the formation of the engineered tissue. To improve the mechanical properties of the polymer, hydroxyapatite fibers may be included to reinforce the polymer matrix (89).

Specific 3D shapes may be required for certain engineered tissue. Simple shapes like rods and cylinders can be formed via thermal extrusion (90). Particulates can be added to create porous structures as with the particulate leaching methods described earlier (91). The extrusion process may also lead to thermal degradation of the polymer matrix. Membranes formed by other techniques may be cut into particular patterns and stacked to form specific 3D shapes (92). The membranes can be bonded through the use of a suitable solvent. The 3D printing is a method akin to rapid prototyping that can also be used to form specific shapes (93). In this method, a polymer powder is spread into an even layer, and solvent is sprayed in a specific pattern to bind the powder together. Another powder layer is added and the solvent is sprayed on to bind the powder to the previous layer. The process is repeated until the entire structure has been formed. Using this method, features as small as 300 μm can be formed.

When specific shapes are not required, *in situ* polymerization can be used to deliver cells and matrix molecules to the site of the of tissue defect. This process can be used to create a cell–polymer matrix to precisely fill the site of the defect. Poly(propylene fumarate) has been used as bone cement (94). Hydrogels containing chondrocytes have been used for cartilage engineering (95).

To ensure proper cell adhesion and growth, the surface of the polymer must be suitable for cell attachment. The wettability and surface free energy greatly influence the extent of cell adhesion. A more hydrophilic surface is necessary to ensure cells will attach to the surface. This must be balanced with the cells needed to interact with surrounding cells in order to form tissue. Surface eroding polymers, such as polyanhydrides, can also renew their surface providing additional area for cell adhesion, which helps to promote cell growth. The surface of the polymer may need to be modified. Amphipathic molecules may be adsorbed to the surface. For example, the polar side chains of poly-L-lysine provide a suitable surface for cell attachment. Extracellular matrix molecules, such as laminin, fibronectin, and collagen, can also be adsorbed to the surface. These molecules contain the amino acid sequence (RGD), which binds to integrins present on the cell surface. The RDG along with IKVAV and YIGSR (found on laminin) can also be covalently bonded to the polymer to create sites for cell adhesion.

The surface of the scaffold can also be modified using microfabrication techniques, these methods can be used to alter the surface chemistry for improved cell adhesion or to create micron scale structural features to affect cell function (96). Besides altering the surface chemistry, microfabrication techniques can also be used to create microscale morphological features on the surface. Microscale features can effect the attachment, motility, and proliferation of fibroblasts in culture (97). The texture of the surface can also influence cardiac myocytes at the molecular level, such as protein localization and gene expression (98). Grooved surfaces can also be used to physically align cells and direct nerve regeneration (99).

Photolithography is a technique commonly used in the production of computer chips. With this method, a special type of photopolymer called a photoresist is coated onto the surface of the scaffold material. Separately, metal is coated onto a glass substrate and laser etched to form a mask in the pattern of interest. The mask is placed over the photoresist and ultraviolet (UV) light is shown through to polymerize the photoresist into the pattern of interest. A solvent is used to remove the unpolymerized photoresist and expose the scaffold. The surface can now be modified

by several methods. Compounds, such as collagen and laminin, can be adsorbed to the surface. Small molecules, such as the RGD tripeptide, can be covalently bonded to the surface. The surface can also be exposed to oxygen plasma to make the area more hydrophilic. Once modified, the remaining photoresist can be removed using another solvent. The success of this method depends on the choice of appropriate solvents and photoresist to use in conjunction with the scaffold and surface modifiers. The major drawback of this technique is that it can only be applied to planar surfaces and relies on the use of expensive specialized equipment. Lithography can also be used to create elastomeric stamps. These stamps can be used to place molecules onto the substrate in a defined pattern. The stamp can be used repeatedly, which helps to reduce the cost compared to traditional photolithography.

Reactive ion etching builds on the premise of photolithography to create microstructural features on the surface (see Fig. 4). Using glass or quartz for the substrate, metal can be deposited onto the patterned photoresist. When the photoresist is removed, the surface is exposed to ion plasma that etches into the substrate leaving deep grooves in the surface. The substrate can now be used as a mold. The polymer matrix can be dissolved and cast onto the mold and dried. When the matrix is lifted from the mold, it will contain the micropatterned structure. As with photolithography the matrix can only be planar. Despite this disadvantage, this method has been used to improve nerve regeneration in bioartificial nerve grafts (100).

Surface of the biodegradable matrix can be etched directly using the process of laser ablation. In this method, a laser beam is used to etch grooves into the substrate. The laser is pulsed rapidly to allow the dispersion of thermal energy and thus reduce the degradation of the polymer matrix.

Another method for depositing molecules in specific patterns is through the use of microfluidics. In this method, a microscale channel system is placed on top of the substrate. Various solutions are passed through the channel and molecules are allowed to interact with the substrate surface. The channel system is removed leaving behind the patterned substrate.

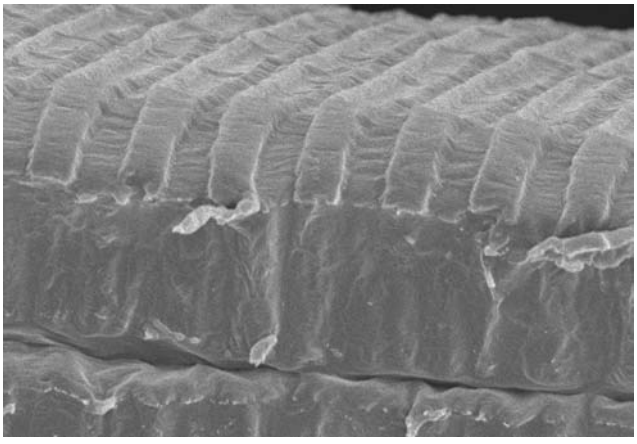


Figure 4. Microfabrication of polymer surface using a combination of reactive ion etching and compression molding. Grooves are 10 μm wide and 3 μm deep.

Signals

Once cells have been seeded onto a biodegradable matrix, signals must be provided to ensure the cells will continue to grow and form fully functional tissue (see Fig. 5). These signals must mimic the environment in which tissue naturally regenerates or develops. Signals may come from contact with the extracellular matrix or other cells. Diffusible factors may be delivered to the cells spatially and transiently. Mechanical and electrical stimuli may also promote tissue formation for certain cell types. Cues may also be provided based on the spatial arrangement of the cells.

Extracellular Matrix. When forming tissue, cells will interact with various proteins that make up the ECM. The ECM is comprised of collagens, proteoglycans, hyaluronic acid, fibronectin, vitronectin, and laminin. Integrins that are found on the surface of cells can interact with short amino acid sequences located on various ECM proteins. For example, cells will adhere to the arginine-glycine-aspartic acid-serine sequence (RGDS) found in fibronectin (101). Integrins are transmembrane proteins that also interact with cytoskeletal proteins like actin. In response to changes in the ECM, they modulate signals to the cells that result in a change in cell function. Certain amino acid sequences have also been implicated in the interactions between integrins and the ECM for specific cell types. These include REDV for endothelial cell adhesion (102), IKVAV for neurite outgrowth (103), and LRE for synaptic development (104).

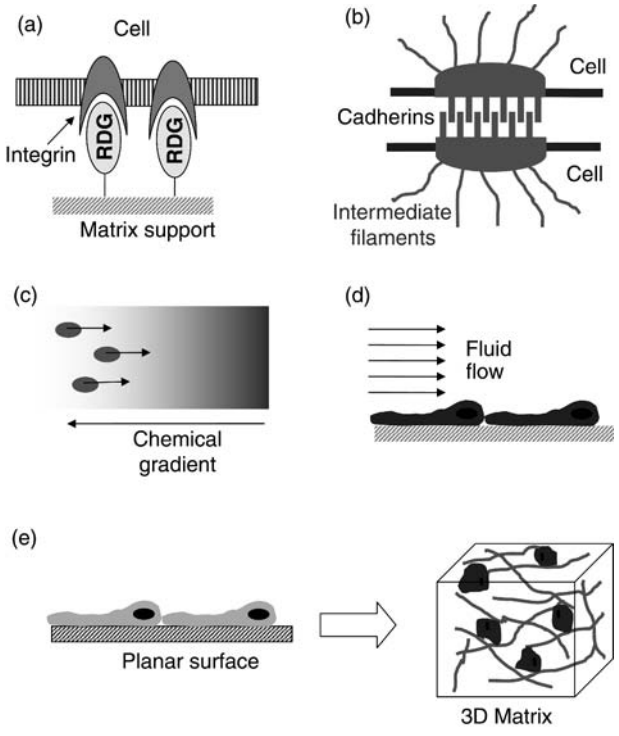


Figure 5. Examples of signals used to control tissue formation. (a) Cell–matrix interaction (RGD–integrin binding). (b) Cell–cell interaction (desmosome). (c) Diffusible factors (chemotaxis). (d) Mechanical forces (shear). (e) Spatial organization.

The primary purpose of the ECM is to anchor the cells so that tissue can form. The connection can ensure that polarized cells are oriented in the correct direction. When cells die, the ECM may retain its integrity until new cells can grow and form tissue. The composition of the ECM varies among tissues. The differences help to define the boundaries between tissue types as well as promote a microenvironment that is best suited for that particular tissue. Soluble regulatory factors may also be stored within the ECM. Disruption of ECM, and thus the tissue, lead to the release of these regulatory factors in order to affect cell function. The ECM can affect functions, such as cell growth, differentiation, and apoptosis. During development, the ECM plays a role in the repatterning of the epithelial-mesanchymal transformation. The physical connect between the cells and the ECM also helps with the modulation of signals due to mechanical stresses.

Diffusible Factors. During development or tissue regeneration, soluble diffusible factor help to control various cell functions, such as proliferation, adhesion, migration, and differentiation. Examples include the various families of growth factors and morphogenic proteins as well as nutrient to ensure cell survival. These factors may be expressed by the cell itself (autocrine) or it may come from a nearby cell (paracrine) or a remote site (endocrine). These factors can elicit changes in the cell cycle, promote differentiation, encourage cell motility and regulate the synthesis of DNA and protein.

During normal development or regeneration, these factors are secreted from various cells and delivered to the target tissue. In order to mimic this effect in engineered tissue, various methods may be used. Prior to implantation, the engineered tissue may be subject to a medium that is supplemented with factor to promote tissue growth. The scaffold for the engineered tissue may have soluble factors incorporated into the matrix. As the substrate degrades, the factors are released in a controlled manner. Molecules may also be immobilized onto the surface of the scaffold in order to alter cell function. Immobilized growth factors can still impart their effect on the cell without being internalized by endocytosis. This provides a mechanism to potentiate the effect of the growth factor.

For controlled release, the scaffold has a dual purpose to provide structural support for the engineered tissue as well as deliver signaling factors at the appropriate time and dose. Other particles that provide controlled release, such as microspheres, may also be incorporated into the engineered tissue. Using a combination of chemical factors with different release profiles can mimic the developmental processes.

Spatial Organization. The spatial arrangement of the cells in culture can have a direct impact on their function. Three-dimensional scaffolds provide support, but also encourage cells to display the appropriate phenotype for tissue formation. Hepatocytes are prone to loss of phenotype when cultured as a monolayer (105). Chondrocytes exhibit fibroblast behavior in monolayers compared to 3D systems. The 3D cultures secrete a great amount of type II collagen that is essential for cartilage formation (106).

The shape of the scaffold is an important consideration for the regeneration of certain tissue types. Nerve tissue engineering uses tubular conduits to help guide the regenerating axons. A luminal diameter that is 2.5 times the diameter of the nerve bundle is optimal for axon extension (107).

Mechanical Modulation. When grown in static cultures, cells tend to settle and form monolayers. This can lead to a loss of phenotype that would be detrimental to tissue formation. To prevent this disaggregation, cells can be cultured in microgravity (108). Microgravity is achieved through the used of specialized bioreactor systems that keeps the cell-scaffold matrix suspended within the media. The net mechanical forces on the tissue are essentially zero.

Mechanical stress may need to be imparted on the engineered tissue in order to ensure the proper development. Shear stresses can encourage the alignment of fibrillar extracellular matrix proteins, such as collagen. This effect will encourage the cell to align with the matrix (109). Alignment of collagen can also be achieved through the use of magnetic forces (46). Pulsatile shear stresses can also effect the orientation of endothelial cells in engineered blood vessels (110). This cyclic stress lead to the orientation of the cells along the circumference of the blood vessel as compared to cells aligned in the axial direction for constant shear loading.

Physical loading of tissue can also impact the mechanical properties of tissues. Signals are modulated via the interaction between the cells and the ECM. Cyclic loading can improve the tensile strength of tissues, such as arteries, heart valves, ligament, muscle, and bone (111). Hydrostatic pressure can also effect the formation of tissue. Cyclic pressure can alter the metabolic function of chondrocytes to produce greater amounts of type II collagen for cartilage formation (112).

Electrical Signals. Besides chemical and mechanical signals, cells will also respond to electrical stimuli, such as electric fields and direct current. Electric fields describe the amount of force exhibited by a charged particle. These forces can have an impact on materials that can be induced to carry a charge, such as cells, or individual ionic molecules. Electric fields can be used to move polarizable materials (dielectrophoresis) or encourage the motion of ions (iontophoresis). Application of direct current (dc) can also be used to mimic the conduction of electrical signals. The effect depends on the tissue type and its conductivity as well as the type of current (alternating or direct), the frequency and magnitude of voltage, and the uniformity of the electric field.

Because the neuromuscular system relies on electrical signaling for its functions, such stimulation can affect tissue formation. Muscle will atrophy when not exposed to an electrical stimulus. Providing an electrical stimulus to muscles until new neural connection can be made will greatly improve the likelihood of success for the nerve regenerative process. Neurons themselves may also respond to electrical stimuli by altering the effect of biochemical cure on growth cone dynamics (113). Myocytes will

develop into functional heart tissue under electrical stimulation (114). Muscle progenitor cells also show improved contractility when exposed to electrical stimulation (115).

Other tissues that do not typically exhibit electrical behavior may still respond to such stimuli. Corneal epithelium has been found to emit dc electrical fields with injured tissues. These cells will migrate toward the cathode in the presence of artificial electrical fields (116). Direct and alternating current (ac) electrical fields can also be used to reduce wound area and wound volume comparatively (117,118).

Cell-to-Cell Interactions. Cells can communicate directly with each other through physical contact via tight junctions, gap junctions, cadherins, and desmosomes or through chemical interactions using soluble factors. Such contact is important for architectural support and for inducing a favorable phenotype. In complex tissue containing multiple cell types, one cell type can act as a physical support for other cells. For example, smooth muscle cells help to support various tubular tissues, such as bladder ducts and blood vessels. Other cells may provide molecular signals to control the metabolic processes in other cells. Schwann cells secrete growth factors that stimulate neurons to regenerate. Without this chemical support, some cells may cease to proliferate and eventually die.

Bioreactor Technology

Once the appropriate cell source, scaffold, and signaling molecules have been selected, the engineered tissue construct may require a specialized bioreactor in which to grow. The bioreactor can provide the vehicle for supplying chemical factors, as well as place mechanical stresses on the tissue if necessary. Liquid media optimized for tissue growth supplies nutrients to the growing tissue and waste products are removed.

A perfusion bioreactor describes the general type of system where cells are retained. The cell-scaffold construct is held in place and media is continuously fed into the bioreactor. The tissue may be physically pinned in place or fixed due to a balance of gravitational and shear forces. Spinner flasks provide mixing through the use of stirrers (119). The engineered tissue is pinned in place while the stirrer keeps the fluid well mixed. The mixing ensured that adequate nutrients are delivered to the cells and wastes are quickly removed to prevent toxic buildup.

Several tissues may require mechanical stimuli in order to achieve full functionality. Novel bioreactor systems have been developed to provide these stresses. Pulsatile shear forces have been used to create engineered arteries with burst strengths comparable to native vessels (110). Placing chondrocytes under cyclic hydrodynamic pressure can enhance the formation of articular cartilage (120). Mechanical cyclic stretching can also improve the strength of engineered ligaments (121).

Tissue Properties

Biomechanics. Engineered tissue should have the same mechanical properties of native tissue in order to achieve full functionality. This issue is especially important with tissue, such as bone and cartilage. Mechanical properties can also influence the tissue architecture by altering the structure of the extracellular matrix. The cytoskeleton interactions with the extracellular matrix can also be affected by the mechanical properties of the tissue. This can alter cell growth, movement, and metabolic function.

Certain tissues also interact with the nervous system based on mechanical signals. The alveoli and passageways in lungs respond to inhalation (122). Stretch receptors in the urinary bladder can detect when the organ must be emptied. Sensory receptors in the skin can detect touch. Engineered tissue must account for physical interaction with other tissue as well as the outside environment.

The influence of biomechanics can be seen at multiple spatial dimension. At the lowest level, biomechanics is influenced by the forces of individual molecules. The individual proteins of the extracellular matrix determine its mechanical properties. The formation and dissolution of chemical bonds can also alter the strength of the tissue. At the cellular level, the plasma membrane and cytoskeleton influence the physical properties of the cell. Cells can also interact with their environment via mechanotransduction pathways. The adhesion, aggregation, and migration of cells will affect the overall tissue and its mechanical properties. At the tissue level, cells interact with the extracellular matrix to create a material with specific physical properties that are essential to its function. An example of this hierarchy can be seen in tendon tissue (123). Individual collagen molecules are bundled together to form microfibers and fibrils. These fibrils combine with fibroblasts to create fascicles that in turn comprise the tendon tissue.

The mechanical properties of structural tissue can be defined by the relationship between stress and strain. As a certain level of stress is placed on the tissue, it will deform. This deformation relative to the initial length defines the strain. In many cases, stress is proportional to strain. The constant of proportionality is called the Young's modulus. For some tissue-like bone, the Young's modulus may vary from location to location (124).

Tissue exhibits characteristics of viscoelastic behavior. As with most biological materials, tissue is slow to deform in response to stress. This slow deformation is known as creep. When stress is quickly placed on tissue and the tissue deforms, the amount of force needed to maintain the deformation decreases over time. This phenomenon is known as stress relaxation. This viscoelastic behavior may be due to the presence of interstitial fluid or the movement of collagen fiber as with cartilage. Subtissue components, like lamellae in bone, may slip relative to one another. Also, some tissue may undergo internal friction due to repeated motions as is seen in tendons and ligaments.

Besides normal forces due to stress, shear forces can also influence the physical properties of tissue. Shear forces are present in tissue where fluid flow is involved. For example, blood flow will influence the properties of endothelial cells in veins and arteries. Compressive forces cause the flow of interstitial fluid in cartilage.

Structural tissue must be engineered to respond to the shear and normal forces it will encounter when placed into the body. In some cases, such as bone, the material placed

into the site of injury must meet the mechanical requirements of the surrounding tissue. If not, the engineered tissue will become damaged thus reducing the chances of recovery.

Biocompatibility. Ideally, the cells and scaffolds that are used in engineered tissue would not provoke an immune or inflammatory response. Unless cells are obtained from the patients themselves, the possibility of rejection remains. The possibility of rejection is dependent on the access that the immune system has with the implanted engineered tissue. For tissue that lacks an established vasculature, for example, cartilage and epidermis, the immune response is limited. Also, certain engineered tissues may utilize cells for a short period of time until the patient's own cells recover. For example, nerve regeneration may be promoted through the use of donor Schwann cells that may become unnecessary once the nerve has recovered. Barring these special cases, methods must be developed to protect the tissue from the immune system.

As with organ transplants, the primary method of controlling the immune response to engineered tissue is through the use of immunosuppressant therapy. These drugs are required for the life of the patient. To avoid side effects and the long-term implications of having a suppressed immune system, methods must be developed to mask the cells. Bone marrow from the donor can be transplanted with the engineered tissue to create a chimera. The combined bone marrow can reduce the likelihood of rejection.

At the molecular level, surface antigens can be altered to prevent rejection. Fragments of antibodies that do not elicit cytotoxic T cell attack may be used to make the antigen from the patients own immune system. This is only temporary and may not prevent all immune mechanisms. Gene ablation can be used to create cells where the antigen coding gene has been inactivated. Alternatively, a gene may be added that is used to synthesize a protein that inhibits rejection. The DNA or RNA that codes for the antigen can be blocked with oligonucleotides that hybridize to prevent transcription or translation.

For cells whose function is primarily metabolic, the cells can be encapsulated to isolate them from the immune system. Encapsulation is accomplished by surrounding the cells with an artificial membrane. The pores in the membrane must be small enough to prevent interaction with the immune system, but large enough to allow nutrients and therapeutic proteins to pass through. The membrane must also be able to withstand any mechanical stress that may disrupt the membrane and allow interaction with the immune system.

Besides the immune system, the implantation of engineered tissue will, to some extent, elicit an inflammatory response. The response initiates a cascade of chemical reactions that alters the gene expression of circulating blood cells (granulocytes, platelets, monocytes, and lymphocytes), resident inflammatory cells (e.g., macrophages, mast cells) and endothelial cells, In conjunction with the cellular activation, many compounds (e.g., growth factors, cytokines, chemokines) are released. The overall effect is characterized by increased blood flow to the tissue that increases temperature and causes redness, swelling, and pain. This creates an environment that isolates the inflamed tissue from the body and promotes wound healing.

During the wound healing, damaged tissue is cleared by the circulating blood cells, A fibrin and fibronectin meshwork is created to act as a substrate for migrating and proliferating cells. Granular tissue containing fibroblasts, myofibroblasts, monocytes, lymphocytes, and endothelial cells forms at the site of the injury. The endothelial cells lead to the formation of new blood vessels (angiogenesis). The tissue undergoes remodeling as tissue regenerates to replace what had been lost. At the end of the process, fibroblast, myofiborblast, and endothelial cells will undergo programmed cells to reduce scarring and return the tissue to its original form.

Engineered tissue and biomaterials implanted in the body can be affected by the inflammatory and wound healing processes. During the granulation process, fibroblasts may migrate to the implant and encase it. This may isolate the engineered tissue from the surrounding tissue and prevent the integration of the engineered tissue with the native tissue. The additional layer of cells will reduce the diffusion of nutrients to the implant. Also, vascularization may be reduced further, hampering regeneration. Though, the layer of fibroblasts may be problematic, the complete lack of fibroblasts may also indicate some level of toxicity from the implant. The ideal implant will enhance the wound healing process while ensuring the engineered tissue does not become isolated.

Cryopreservation

To engineer tissue that can be delivered on demand, methods must be developed for long-term storage. Cryopreservation can be used not only for the storage of the final tissue product, but also for the cellular components. When new cells are isolated for use in engineered tissue, they may be preserved for latter expansion to prevent senescence or DNA mutation. The original tissue from which cells are derived may also be stored for later use. The isolated cells can be preserved while a sample is screened for infectious agents as well as the ability to form functional tissue. Cells must be banked in accordance with FDA regulations to ensure their genetic stability. During the production stage, the engineered tissue may be preserved at various steps for later quality control testing. The engineered tissue can be stored at hospitals to ensure availability

The protocol for cryopreservation depends on the cooling rate as well as the addition of compounds to reduce cell damage. If the cooling rate is too low, osmosis will drive water from the cell leading to severe dehydration. A high cooling rate will promote intracellular ice formation. The expansion of the ice forming inside the cell may lead to damage of the plasma membrane. Additives may be used to prevent cell damage caused by the increased ionic concentration in the unfrozen part of the tissue. Additives may be permeating [dimethyl sulfoxide (dmss) and ethylene glycol] or nonpermeating [poly(vinyl pyrrolidone) and starch]. The use of permeating additives would necessitate additional

steps to remove the additive in preparation of the engineered tissue for implantation. One drawback to nonpermeating additives is that the osmotic stress due to high extracellular concentrations may lead to dehydration of the cells. The viability of the cells that have been preserved will depend not just on the method, but also on the cell concentrations and the type of cell.

Regulations

During the 1990s, the FDA recognized the need to develop regulations for products derived from tissue engineering principles. Engineered tissue is renamed by the FDA as tissue engineered medical products (TEMPs), which would include products used to replace damage tissue. Organs for transplantation and blood were excluded from this designation. Examples include artificial skin, bone graft, vascular grafts, nerve grafts and metabolic assist devices.

The primary concerns of TEMPs were disease transmission, controls of product manufacturing to ensure product integrity, clinical safety and efficacy, promotional claims, and monitoring the industry. The characteristics that determine the level risk include the cells source, the viability of cells and tissue, homologous function, manipulation of the implant chemically or genetically, systemic versus local effects, the long-term storage of the device, and the combination of the cells with other cells or drugs.

The first rule (63 FR 26744 Establishment Registration and listing of manufacturers of Human Cellular and Tissue-based Products) proposed the registration of the various established companies involved in the manufacturing of TEMPs in order to provide better communication between the FDA and industry. The second rule (Suitability Determination for Donors of Human Cellular and Tissue-Based Products, Final Rule 5/24/04 changes to 21 CFR 210, 211, 820) is designed to make changes to the regulations for Good Manufacturing Practices to require industry to test cells and tissue in order to prevent the transmission of disease and the unwitting use of contaminated tissue. The thirds rule (Current Good Tissue Practice for Manufacturers of Human Cellular and Tissue-based Products; Inspection and Enforcement, final rule 5/25/05 changes to 21 CFR 820) defines the methods, facilities, and controls used in the manufacturing of TEMPs.

EXAMPLES OF ENGINEERED TISSUE

Engineered tissue implants are designed from a combination of cells, a biodegradable scaffold, and chemical signal. Once a need has been established for an implant, the system must be designed to be compatible with the patient. The implant relies on an adequate source of cells that can form functional tissue. Alternatively, the implant may be designed to recruit cells from the patient. Since the human body is not a static system, the engineered tissue must be able to interact with the patients existing tissue. Cells must continue to survive. The structure of the scaffold must maintain its integrity until the tissue is fully integrated. The tissue structure should be achieved very quickly and be fully accepted by the host. The implant should foster cell migration into and out of the new tissue. The implant should seamlessly integrate with surrounding tissue.

Tissue can be categorized into three main types: structural, metabolic, or combination. Structural tissue provides physical strength and a stable structure. The strength is determined by the extracellular matrix supported by the cells. Examples of structural tissue include bone, ligament, and vascular grafts. Metabolic-type tissues are defined as having functions based on the secretion or absorption of chemicals. These can be subdivided into tissues that respond to some stimulus (pancreas) and those that function independent of stimuli (liver). Combined tissue exhibit characteristics of both structural and metabolic tissue (skin).

Metabolic

Pancreas. The pancreas controls the blood sugar level through the regulation of insulin. The loss of this function leads to diabetes mellitus and requires daily injection of insulin. Over the long term, diabetes can damage other tissue such as eyes, kidneys, and nerve. Islets of Langerhans comprise only 1–2% of pancreatic tissue, but are the cells that regulate insulin levels. The islets are comprised of different cell types that lack the capacity to expand. As a result, islets must come from donor material.

The number of organ donors cannot supply sufficient islets for fill the demand of diabetes patients. Xenografts offer a viable alternative. Pig pancreatic tissue is currently used for insulin production and so is considered a good candidate for islet transplantation. To prevent an acute immune response, the islets must be encapsulated. Glucose and insulin along with other small molecules would diffuse across the membrane to allow chemical interaction while preventing antibodies from initiating an immune response.

Pancreatic islets have been encapsulated in various polymers. A biodegradable hydrogel-based coating has been developed for timed degradation of the capsule (125). The polymer can be modified to degrade at the same time that the islets reach the end of their functional life. The cells would then be removed by the immune system. Additional islets can be injected into the patient as necessary. A bioartificial device incorporating microencapsulated pancreatic islets can be connected to the vasculature to allow chemical interaction between the islets and the blood stream (126).

An alternative to the bioartificial pancreas is to inject islets directly into the existing pancreatic tissue. Islets have successfully been transplanted into diabetic patients (127). For this method to be successful, the islets must be completely biocompatible with the patient. Progenitor cells have been isolated from pancreatic tissue that exhibits the capacity to differentiate into cells similar to beta cells (128). Evidence indicates that adult stem cells may also develop into pancreatic cells without showing any indication of transdifferentiation (129). More work, though, needs to be done to create a universally acceptable stem cell derived pancreatic islet.

Liver. The primary cause of liver damage is cirrhosis due to alcohol consumption and hepatitis. This damage will

prevent the proper metabolism of nutrients as well as toxic substance that must be removed from the body. When damage is severe enough, the only treatment available is an organ transplant, but due to the shortage of liver donors, many patients die waiting. Partial liver transplants use a single donor liver that is divided and transplanted into multiple patients. This technique demonstrates the ability of implanted tissue to assist the native liver (130). Individual cultured hepatocytes may also be transplanted to assist in liver function (131). Cells are placed in organs with a high level of vascularization (liver, pancreas, spleen). This enhances the chemical interaction between the hepatocytes and the blood stream. Like islets, the hepatocytes may also be microencapsulated to eliminate the need for immune suppressant therapy.

Until an organ becomes available, a patient can be sustained through the use of a liver assist device. Several different liver assist devices have been developed (132). In these systems, hepatocytes are placed within a bioartificial device and allowed to contact blood for plasma. The cells are retained behind some form of membrane and small molecules are allowed to diffuse through and become metabolized. The technology behind liver assist devices may eventually lead to the development of a whole engineered organ.

Future research is geared toward creating an engineered organ. Techniques will need to be developed to ensure a high level of vascularization through the engineered organ. Another method is to create a totally implantable liver assist device to act as a permanent liver replacement. The system will need to be engineered to be self-contained and small enough to fit into the abdomen of the patient.

Structural Tissue

Bone. The main purpose of bone is to provide structural support for the human body. Bone also helps to protect the internal organs of the body. It provides attachment sites for muscles to allow locomotion.

The primary cells found in bone are osteoclasts, osteoblasts, and osteocytes. Osteoblasts are responsible for the deposition of bone matrix while osteoclasts erode it. The cell dynamics permits the continuous turnover of bone matrix material that helps the tissue quickly respond to damage. Osteocytes can respond to mechanical stimuli and also promote blood–calcium homeostasis. The bone matrix is comprised of 65–70% hydroxyapatite. The remainder is composed of organic molecules, such as collagen 1, fibronectin, and various glycoproteins, sialoproteins and, proteoglycans (133).

Autologous bone grafts are the primary method for bone repair. This method has been very successful, but is restricted by the amount of tissue that can be obtained from the patient. Allogenic tissue may be used, but its rate of graft incorporation is much lower. Also, rejection of the tissue may be problematic and may introduce pathogens. Metals and ceramics offer an alternative to the grafts, but cannot become fully integrated into the existing tissue. This may lead to fatigue failure at the metal–bone interface or breakup of the ceramic material.

For engineered bone tissue, the graft combines scaffold with bone cells and compounds that promote osteoinduction. The most common scaffold is based on natural or synthetic hydroxyapatite (134). These materials have major drawbacks in terms of their brittleness and rapid dissolution rate. Many synthetic and natural polymers have been considered for use as a scaffold (133). Osteoblasts are the primary cell component of engineered tissue because of their ability to synthesize bone matrix. These cells may be isolated from the patient or donor source. Another option is to isolate mesanchymal stem cells from the bone marrow. The cells have been show to differentiate into the various bone cells when exposed to dexamethasone (135). The primary growth factors that have been found to affect bone tissue formation are bone morphogenetic proteins (BMPs), transforming growth factor beta (TGF-), fibroblast growth factors (FGFs), insulin growth factor I and II (IGF I/II), and platelet derived growth factor (PDGF). The VEGF may also be incorporated into the matrix to promote vascularization of the bone tissue.

The bioreactors used for bone tissue engineering have mostly been confined to spinner flask and rotating wall vessels. The rotating wall vessels help to promote cell interactions, but the microgravity imposed by the bioreactor may actually lead to bone loss (136).

Currently, most engineered bone tissue utilizes a combination of mesanchymal stem cells with a highly porous biodegradable matrix (133). Alternatively, multipotent cells isolated from the periosteum seeded on PLGA scaffolds created bone tissue that was well integrated with the native tissue (137). Transfecting bone progenitor cells with BMP-2 can also enhance their ability to create new tissue (138).

Because of the large number of growth factor that can influence bone formation, more work must be done to understand the underlying mechanisms of how these growth factors influence the various bone cell types. New materials with surface modifications are also being evaluated for their ability to control cell differentiation and migration. Rapid prototyping is also emerging as a processing technique to create scaffolds with control spatial arrangement, porosity, and bulk shape (139).

Cartilage. Cartilage is divided into two types: fibrous and articular. Fibrocartilage is used to provide shape, but flexibility to body parts, such as the ear and nose, while articular cartilage is found in joints where bone meets bone. Damage to articular cartilage can lead to painful arthritis and loss of motion. Traditional surgical repair techniques involve removal of damaged cartilage and reshaping the tissue to prevent further damage. Grafts from autologous tissue have not been successful in repair tissue. One reason for the difficulty in repair is the lack of a vasculature. The wound healing process available to other tissue types will not effect cartilage regeneration.

Genzyme has developed a procedure that involves the use of autologous cells that are expanded in culture and reintroduced into the site of injury. This method has had a fairly high success rate and can be used for defects up to 15 cm^2. Larger defects require the use of a scaffold to hold the engineered tissue in place. Photomonomers can be

combined with chondrocytes and other growth factors to engineer tissue *in vivo* (140). Ultraviolet light is then used to initiate polymerization. This method is ideal for irregularly shaped defects and is less invasive. The hydrogel structure of the cross-linked polymer is also ideal for cartilage growth because the chondrocytes have a spherical morphology like that of native cells.

One of the main characteristics of cartilage that gives it strength is the trizonal arrangement of collagen fiber. This characteristic is difficult to mimic *in vitro*. Polyethylene oxide based photopolymers have successfully been used to encapsulate chondrocytes into discrete layers (141). Chondrocytes also respond to mechanical stresses that will affect its production of extracellular matrix. Cyclic hydrostatic pressure that mimics the forces present in knee joints can increase the synthesis of type II collagen (112). Type II collagen is an important protein in the extracellular matrix of articular cartilage. Fibrocartilage ECM consists primarily of type I cartilage.

Cell culture technology has allowed for the rapid expansion of autologous chondrocytes. Problems associated with immune rejection can be avoided, but two surgeries are required to create the engineered cartilage. Inroads have been to engineer cartilage that is structurally equivalent to native tissue. With proper control of differentiation, stems cells may be injected directly into the injured site to form new cartilage. Advances in immunomodulation technology may render stem cells resistant to the immune system that will eliminate the need for acquiring autologous cells. Also, injectable biomaterials may eventually be created that can be spatially arranged *in situ* to mimic the trizonal arrangement of collagen.

Blood Vessels. Blood vessels play a major role in delivering nutrient and removing wastes from all parts of the body. Disruptions to the flow of blood can lead to tissue loss downstream from the site of injury. In response to transection of the blood vessel, a cascade of events leads to the clotting of the blood and wound repair. Beside physical damage to the blood vessel, arthrosclerosis can lead to partial or complete blockage of the blood vessel. Within the heart, this can lead to myocardial infarction and even death unless bypass surgery is performed. Artificial grafts are used to bypass the flow of blood around the site of the blockage. This method has only been successful for grafts >6 mm (142). Small diameter grafts are being engineered to fill this gap (143). The major problem with grafting is restenosis, which can lead to failure.

The artery consists of three layers: intimal, medial, and adventitia. The intimal layer is comprised of the endothelial cells that line the inside of the blood vessel. For engineered arteries, these cells are typically obtained from donor material. Advances still need to be made in stem cell research to create new endothelial cells. The smooth muscle cells of the medial layer control the flow of blood via constriction or dilation of the artery. The adventitia layer is made up of fibroblasts and extra cellular matrix. While this layer is commonly left out of engineered arteries, its presence provides additional mechanical strength (37).

For the scaffold, PLA, PGA and their copolymers as well as poly-4-hydroxybutyrate are most commonly used. Engineered arteries have also been created without the use of synthetic materials (37). In these cases, donor arteries have their cells removed. The remaining matrix structure is reseeded with endothelial cells on the inner-lumen and smooth muscle cells on the exterior. Once the cell–scaffold implant has been created, a mechanical stimulus can be used to confer increased strength. Specialized bioreactors that provide cyclic shear flow that mimic the pulsatile flow of blood have been developed for engineering arteries (144).

The ideal engineered artery would have the mechanical strength to withstand the pulsatile shear stresses and blood pressure, would be biocompatible, and would integrate seamlessly with the existing blood vessel. Since autologous vessels may not be practical for patients suffering from arthrosclerosis, an acellular implant would be better suited. Acellular implants have displayed excellent results in animal models (145). These systems have a greater capacity to remodel their structure to better mesh with the native tissue. Cells are recruited from the existing tissue. Better understanding of the remodeling mechanism in humans would help to improve these implants.

Combined

Skin. Skin is considered to be the largest organ of the body. Its primary function is to protect the body from the environment. It helps to regulate fluid content and body temperature. Skin acts as the first line of defense for immune surveillance. Sensory receptor found in the skin help the body examine the environment. When injured, skin has the capacity to self-repair. The common form of injury to skin is a burn. Ulcerations may also form due to venal stasis, diabetes, and pressure sores. Skin continuity may be disrupted due to accidental physical trauma or the removal of skin cancers. Though the skin has the capacity to regenerate, engineered skin may be required to provide physical protection and metabolic regulation until the native tissue covers the wound.

Skin is comprised of two layers: the dermis and epidermis. The epidermis is the outer layer comprised of keratinocytes. This layer protects the dermis layer and helps to regulate heat and water loss. The dermis contains the vasculature, nerve bundles, and lymphatic systems that connect the skin to the rest of the body.

The traditional method to large area skin repair was to obtain skin from intact portions of the body and spread it around to increase the rate of regeneration. This was problematic for patients with severe injuries. Also, the area for harvesting may become damaged as well. An alternative was to use donor cadaver tissue that usually provoked an immune response. To overcome these obstacles, skin may be engineered to be biocompatible with the patient and help to promote the body's ability to self-repair.

Engineered skin should contain a belayed structure that allows for rapid vascularization and innervation from the body. The dermis layer should promote rapid wound repair. The epidermal layer should have the capacity to protect the body from the environment. The system should become fully integrated into the wound (146).

Initially, engineered skin was designed to act as a wound dressing and not become integrated with the native

tissue. Alloderm, from Life Cell Technologies and approved in 1992, was an acellular dermal matrix derived from cadaver tissue. Integra was approved in 1996 and consisted of a silicone sheet coated with collagen and glycoaminoglycans. The silicone sheet helped to prevent fluid loss, but needed to be removed when the tissue eventually healed. Autologous cultured skin substitutes have been used in combination with Integra for engraftment into burn wounds of pediatric patients. The elastic quality of this engineer skin allowed the new tissue to grow with the patient (147). In 1998, Organogenesis received FDA approval for the first tissue engineer product (Apligraf). The product utilized a collagen gel sheet seeded with fibroblast. Dermagraft (Advanced Tissue Sciences) was the first skin substitute to utilize biodegradable polyglactin. The fiber mesh was coated with fibroblasts that secrete growth promoting factors. More advanced cultured skin substitutes, next generation engineered skin, should have a full thickness bilayered structure that can become integrated into the wound for faster recovery.

For full thickness engineered skin to become integrated into the patient's own tissue, the dermis layer must promote immediate vascularization. Without a supply of nutrients to the tissue, the epidermal layer would begin to die and slough off (148). The inclusion of fibroblasts and endothelial cells in the dermis layer promoted the formation of a capillary bed that improved neovascularization (149).

Engineered skin can also be used for other functions besides wound repair. Stem cells that are normally present in the epidermis may be transfected and included in engineered skin to produce therapeutic proteins for patient suffering from chronic disorders (150). Engineered skin may contain hair follicle cells to create tissue for hair implants (151). Future engineered skin should include melanocytes to match the skin to the patient's native tones as well as sweat glands to regulate sweating and sensory receptors that integrate with the existing nervous system.

Nerves. Nerve tissue engineering presents a unique problem not encountered with other tissue. Nerve tissue is comprised of neurons and satellite cells. In the peripheral nervous system, the satellite cells are Schwann cells. These cells secrete growth factors to stimulate nerve regeneration when they lose contact with neurons. In contact with neurons, Schwann cells ensheath or myelinate the axon in order to enhance the conduction of the electrical signal. The central nervous system (brain and spinal column) contains oligodendrocytes that ensheath axons and astrocytes forming the blood–brain barrier. Instead of promoting regeneration, oligodendrocytes are actually inhibitory. Damage to these neurons can lead to the loss of sensory and motor function, paralysis, and even death.

When a nerve becomes severed, the proximal segment of the axon closest to the cell body will extend and enter the degenerated distal portion and continue growing until new synaptic connections are made. If connections are not made in a timely manner, the neuron may lose function and die making the loss of function permanent. The traditional method of nerve repair is to surgically reconnect the two severed ends. When the damage is too extensive to reconnect the tissue, donor tissue from a less critical nerve is used to bridge the gap. As an alternative, an engineered bioartificial nerve graft can be used to bridge the gap to eliminate the need for the donor material. Several nerve grafts have been developed that utilize different scaffolds materials, various cells, and combinations of growth factor in order to enhance nerve regeneration (152). Clinical trials for peripheral nerve repair have shown success using poly(tetrafluoroethylene) to regenerate nerve tissue with gaps up to 4 cm long (153). The FDA has also approved a PGA conduit (Neurotube, Neuroregen LLC, Bel Air, MD) and a collagen-based nerve tube (NeuraGen, Integra Neurosciences, Plainsboro, NJ) for peripheral nerve repair. For the material that is chosen, the physical parameters of the conduit, such as porosity, can be optimized to ensure adequate nutrient reach the regenerating tissue while retaining growth factor with the implant (83).

Various chemical matrices, such as ECM components, can be added to the conduit to further enhance nerve regeneration (152). While the inclusion of these factors may provide incremental improvements in the overall design, they have not surpassed traditional surgical techniques to regenerate nerves (154). In the 1940s, Weiss, a pioneer in artificial nerve graft research, claimed that the ideal substrate for nerve regeneration is degenerated nerve (155). Degenerated nerve contains the matrix molecules and cellular components that have been naturally optimized to promote nerve regeneration.

Non-neuronal cells may be added to the artificial nerve graft in order to create an environment that mimics the natural nerve regeneration process. Schwann cells can be used to enhance the regeneration of both peripheral and central nerve tissue (156). Though they are not found in the central system, Schwann cells can overcome the inhibitory effect of the oligodendrocytes. Alternatively, fibroblasts transfected to produce nerve growth factor and seeded into an artificial nerve graft have been used to regenerate nerve tissue (157). Neural stem cells used in conjunction with a structured polymer scaffold have led to functional recovery from spinal cord injuries in rat models (27).

The next generation of nerve grafts should regenerate nerve tissue over much longer distances than currently achieved. Grafts containing a matrix comparable to degenerated nerve should make this possible. Microfabrication techniques also show promise for control of the growth of axons at the cellular level (100). Eventually, techniques must be developed to transplant new neural tissue into the existing system. This will become necessary for the innervation of engineered tissue and organs that relay on communication with the nervous system. Additional research is needed to understand the impact of additional neural contact on the overall system. These techniques may make possible a cure for paralysis where nerve tissue has degenerated beyond repair.

FUTURE PROSPECTS

The advances made in tissue engineering since the 1980s are set to transform the standard methods for medical care. The previous section is just a small sampling of the tissues currently under investigation. Just about every tissue in

the body is currently being studied for repair using engineered tissue or cell therapy. Though much has been accomplished, more work still needs to be done.

Stem cells have shown great promise as a universal cell for developing engineered tissue implants. Research adult stem cells, such as mesanchymal and hematopoietic cells, hints at their capacity to act as a pluripotent cell type. Additional work may reveal more cell types that these can differentiate into. Strict regulations may hinder investigation of embryonic stem cells, but additional research in cultivation methods can alleviate ethical concerns. More work still needs to be done to control the differentiation of these cells to prevent the formation of teratomas.

The variety of tissue types being engineered makes development of a single universal scaffold difficult. The mechanical characteristics of individual tissues require scaffolds with specialized properties. Although a few polymers have been accepted by the FDA, more will need to be evaluated to ensure enough options are available for the engineered tissue. New biomaterials should also be investigated to ensure a wide variety of options. For a given class of biomaterials, though, the polymer should be customizable to ensure the appropriate mechanical and degradation properties. Also, the material should be capable of manipulation to enhance adhesion and control tissue formation.

As stem cells increase in importance, more signaling factors to control differentiation will be needed. Such control should not be limited to diffusible factors, but should also include surface bound molecules that can mimic the cell–cell contact interactions that occur during development. Such molecules may be bound to the surface using microfabrication techniques to encourage differentiation of the stem cells into spatially arranged multiple cell types. With advances in gene transfer and other immunomodulation techniques, stem cells may be rendered fully biocompatible with the patient regardless of the donor source.

Several specific examples of engineered tissues were presented earlier. Each tissue type has its own issues that must be overcome for the engineered tissue to be successful. In general, though, many tissues must interact with the existing vascular and nervous systems. As engineered tissue turns to engineered organs, protocols must be developed to ensure adequate nutrient will reach the cells. Vascular endothelial growth factor provides a means of recruiting blood vessels to the new tissue, but would not be useful in growing whole organs *in vitro*. Bioreactors will need to be designed to accommodate multiple tissue types, including endothelial cells for neovasculature formation. As the mechanisms for stem cell differentiation become better understood, the possibility of growing entire organs may become a reality.

Another general concern for engineered tissues is the connection with the existing nervous system. Muscle and skin interact with the peripheral nervous system. Several internal organs interact with the brain via the autonomic nervous system. For such organs and tissue to be fully integrated and functional, methods must be developed to attract existing neurons. In some cases, nerve tissue may not be present so the engineered tissue may need to include a neural component that can be integrated with the nervous system.

Engineered tissue will ultimately be used to repair practically any tissue in the body. With the almost infinite combinations of biomaterials, growth factors and cells, the only limit to creating new tissue is one's imagination.

BIBLIOGRAPHY

Cited References

1. Skalak R, Fox CF. Tissue Engineering: Proceedings of a Workshop, Held at Granlibakken, Lake Tahoe, California; 1988 Feb. 26–29. New York: Liss; 1988.
2. Bell E. Tissue Engineering: Current Perspectives. Boston: Birkhäuser; 1993.
3. Langer R, Vacanti JP. Tissue engineering. Science 1993;260 (5110):920–926.
4. Zimbler MS. Gaspare tagliacozzi (1545–1599):Renaissance surgeon. Arch Facial Plast Surg 2001;3(4):283–284.
5. Huber GC. A study of the operative treatment for loss of nerve substance inperipheral nerve. J Morph 1895;11:629–740.
6. NUPOC. Prosthetics History. 9/19 2004;2004(12/29).
7. Birchmeier C, Birchmeier W. Molecular aspects of mesenchymal-epithelial interactions. Annu Rev Cell Biol 1993;9:511–540.
8. Ajioka I, Akaike T, Watanabe Y. Expression of vascular endothelial growth factor promotes colonization, vascularization, and growth of transplanted hepatic tissues in the mouse. Hepatology 1999;29(2):396–402.
9. Charlton B, Auchincloss Jr H, Fathman CG. Mechanisms of transplantation tolerance. Annu Rev Immunol 1994;12:707–734.
10. Faustman D, Coe C. Prevention of xenograft rejection by masking donor HLA class I antigens. Science 1991;252(5013): 1700–1702.
11. Markmann JF, et al. Indefinite survival of MHC class I-deficient murine pancreatic islet allografts. Transplantation 1992;54(6):1085–1089.
12. Platt JL. A perspective on xenograft rejection and accommodation. Immunol Rev 1994;141:127–149.
13. Ramanathan M, et al. Characterization of the oligodeoxynucleotide-mediated inhibition of interferon-gamma-induced major histocompatibility complex class I and intercellular adhesion molecule-1. J Biol Chem 1994;269(40):24564–24574.
14. Soon-Shiong P, et al. Insulin independence in a type 1 diabetic patient after encapsulated islet transplantation. Lancet 1994;343(8903):950–951.
15. Soon-Shiong P, et al. Long-term reversal of diabetes by the injection of immunoprotected islets. Proc Natl Acad Sci USA 1993;90(12):5843–5847.
16. McConnell MP, et al. In vivo induction and delivery of nerve growth factor, using HEK-293 cells. Tissue Eng 2004;10(9–10):1492–1501.
17. Fischbach C, et al. Generation of mature fat pads in vitro and in vivo utilizing 3-D long-term culture of 3T3-L1 preadipocytes. Exp Cell Res 2004;300(1):54–64.
18. Jiang XQ, et al. The ectopic study of tissue-engineered bone with hBMP-4 gene modified bone marrow stromal cells in rabbits. Chin Med J (Engl) 2005;118(4):281–288.
19. Fujita Y, et al. Transcellular water transport and stability of expression in aquaporin 1-transfected LLC-PK1 cells in the development of a portable bioartificial renal tubule device. Tissue Eng 2004;10(5–6):711–722.
20. Jin Y, Fischer I, Tessler A, Houle JD. Transplants of fibroblasts genetically modified to express BDNF promote axonal regeneration from supraspinal neurons following chronic spinal cord injury. Exp Neurol 2002;177(1):265–275.

21. Mezey E, et al. Turning blood into brain: Cells bearing neuronal antigens generated in vivo from bone marrow. Science 2000;290(5497):1779–1782.
22. Bjornson CR, et al. Turning brain into blood: A hematopoietic fate adopted by adult neural stem cells in vivo. Science 1999; 283(5401):534–537.
23. Ying QL, Nichols J, Evans EP, Smith AG. Changing potency by spontaneous fusion. Nature London 2002;416(6880): 545–548.
24. Thomson JA, et al. Embryonic stem cell lines derived from human blastocysts. Science 1998;282(5391):1145–1147.
25. Holden C, Vogel G. Cell biology. A technical fix for an ethical bind? Science 2004;306(5705):2174–2176.
26. Levenberg S, Langer R. Advances in tissue engineering. Curr Top Dev Biol 2004;61:113–134.
27. Teng YD, et al. Functional recovery following traumatic spinal cord injury mediated by a unique polymer scaffold seeded with neural stem cells. Proc Natl Acad Sci USA 2002;99(5):3024–3029.
28. Levenberg S, et al. Endothelial cells derived from human embryonic stem cells. Proc Natl Acad Sci USA 2002;99(7): 4391–4396.
29. Iwanami A, et al. Transplantation of human neural stem cells for spinal cord injury in primates. J Neurosci Res 2005;80(2):182–190.
30. Smits AM, et al. The role of stem cells in cardiac regeneration. J Cell Mol Med 2005;9(1):25–36.
31. Trucco M. Regeneration of the pancreatic beta cell. J Clin Invest 2005;115(1):5–12.
32. Luyten FP. Mesenchymal stem cells in osteoarthritis. Curr Opin Rheumatol 2004;16(5):599–603.
33. Ellis DL, Yannas IV. Recent advances in tissue synthesis in vivo by use of collagen-glycosaminoglycan copolymers. Biomaterials 1996;17(3):291–299.
34. Bell E, Ivarsson B, Merrill C. Production of a tissue-like structure by contraction of collagen lattices by human fibroblasts of different proliferative potential in vitro. Proc Natl Acad Sci USA 1979;76(3):1274–1278.
35. Weinberg CB, Bell E. A blood vessel model constructed from collagen and cultured vascular cells. Science 1986;231(4736): 397–400.
36. Auger FA, et al. Skin equivalent produced with human collagen. In Vitro Cell Dev Biol Anim 1995;31(6):432–439.
37. L'Heureux N, et al. A completely biological tissue-engineered human blood vessel. FASEB J 1998;12(1):47–56.
38. Lee CH, Singla A, Lee Y. Biomedical applications of collagen. Int J Pharm 2001;221(1–2):1–22.
39. Schmidt CE, Baier JM. Acellular vascular tissues: Natural biomaterials for tissue repair and tissue engineering. Biomaterials 2000;21(22):2215–2231.
40. Wakitani S, et al. Repair of large full-thickness articular cartilage defects with allograft articular chondrocytes embedded in a collagen gel. Tissue Eng 1998;4(4):429–444.
41. Liu S, et al. Axonal regrowth through collagen tubes bridging the spinal cord to nerve roots. J Neurosci Res 1997;49(4):425–432.
42. Atala A. Tissue engineering for bladder substitution. World J Urol 2000;18(5):364–370.
43. Orwin EJ, Hubel A. In vitro culture characteristics of corneal epithelial, endothelial, and keratocyte cells in a native collagen matrix. Tissue Eng 2000;6(4):307–319.
44. Pomahac B, et al. Tissue engineering of skin. Crit Rev Oral Biol Med 1998;9(3):333–344.
45. Elbjeirami WM, Yonter EO, Starcher BC, West JL. Enhancing mechanical properties of tissue-engineered constructs via lysyl oxidase crosslinking activity. J Biomed Mater Res A 2003;66(3):513–521.
46. Ceballos D, et al. Magnetically aligned collagen gel filling a collagen nerve guide improves peripheral nerve regeneration. Exp Neurol 1999;158(2):290–300.
47. Xu XM, Zhang SX, Li H, Aebischer P, Bunge MB. Regrowth of axons into the distal spinal cord through a schwann-cell-seeded mini-channel implanted into hemisected adult rat spinal cord. Eur J Neurosci 1999;11(5):1723–1740.
48. Sieminski AL, Padera RF, Blunk T, Gooch KJ. Systemic delivery of human growth hormone using genetically modified tissue-engineered microvascular networks: Prolonged delivery and endothelial survival with inclusion of nonendothelial cells. Tissue Eng 2002;8(6):1057–1069.
49. Zimmermann WH, Melnychenko I, Eschenhagen T. Engineered heart tissue for regeneration of diseased hearts. Biomaterials 2004;25(9):1639–1647.
50. Stevens MM, Qanadilo HF, Langer R, Prasad Shastri V. A rapid-curing alginate gel system: Utility in periosteum-derived cartilage tissue engineering. Biomaterials 2004;25 (5):887–894.
51. Dar A, Shachar M, Leor J, Cohen S. Optimization of cardiac cell seeding and distribution in 3D porous alginate scaffolds. Biotechnol Bioeng 2002;80(3):305–312.
52. Dvir-Ginzberg M, Gamlieli-Bonshtein I, Agbaria R, Cohen S. Liver tissue engineering within alginate scaffolds: Effects of cell-seeding density on hepatocyte viability, morphology, and function. Tissue Eng 2003;9(4):757–766.
53. Mol A, et al. Fibrin as a cell carrier in cardiovascular tissue engineering applications. Biomaterials 2005;26(16):3113–3121.
54. Karp JM, Sarraf F, Shoichet MS, Davies JE. Fibrin-filled scaffolds for bone-tissue engineering: An in vivo study. J Biomed Mater Res 2004;71A(1):162–171.
55. Fussenegger M, et al. Stabilized autologous fibrin-chondrocyte constructs for cartilage repair in vivo. Ann Plast Surg 2003;51(5):493–498.
56. Bannasch H, et al. Skin tissue engineering. Clin Plast Surg 2003;30(4):573–579.
57. Pittier R, Sauthier F, Hubbell JA, Hall H. Neurite extension and in vitro myelination within three-dimensional modified fibrin matrices. J Neurobiol 2004.
58. Suh JK, Matthew HW. Application of chitosan-based polysaccharide biomaterials in cartilage tissue engineering: A review. Biomaterials 2000;21(24):2589–2598.
59. Seol YJ, et al. Chitosan sponges as tissue engineering scaffolds for bone formation. Biotechnol Lett 2004;26(13):1037–1041.
60. Li K, et al. Chitosan/gelatin composite microcarrier for hepatocyte culture. Biotechnol Lett 2004;26(11):879–883.
61. Li Z, et al. Chitosan-alginate hybrid scaffolds for bone tissue engineering. Biomaterials 2005;26(18):3919–3928.
62. Schaner PJ, et al. Decellularized vein as a potential scaffold for vascular tissue engineering. J Vasc Surg 2004;40(1):146–153.
63. Grabow N, et al. Mechanical and structural properties of a novel hybrid heart valve scaffold for tissue engineering. Artif Organs 2004;28(11):971–979.
64. Kimuli M, Eardley I, Southgate J. In vitro assessment of decellularized porcine dermis as a matrix for urinary tract reconstruction. BJU Int 2004;94(6):859–866.
65. Miller C, et al. Oriented schwann cell growth on micropatterned biodegradable polymer substrates. Biomaterials 2001; 22(11):1263–1269.
66. Cook AD, Hrkach JS, Gao NN, Johnson IM, Pajvani UB, Cannizzaro SM, Langer R. Characterization and development of RGD-peptide-modified poly(lactic acid-co-lysine) as an interactive, resorbable biomaterial. J Biomed Mater Res 1997;35(4):513–523.

67. Webb AR, Yang J, Ameer GA. Biodegradable polyester elastomers in tissue engineering. Expert Opin Biol Ther 2004;4(6):801–812.
68. Pitt CG. Biodegradable polymers as drug delivery systems. In: Chasin M, Langer RS, editors. Poly-e-Caprolactone and its Copolym, Ers. New York: Marcel Dekker; 1990.
69. Martin DP, Williams SF. Medical applications of poly-4-hydroxybutyrate: A strong felxible absorbable biomaterial. Biochem Eng J 2003;16:97–105.
70. Temenoff JS, Mikos AG. Injectable biodegradable materials for orthopedic tissue engineering. Biomaterials 2000;21(23): 2405–2412.
71. Tangpasuthadol V, Pendharkar SM, Kohn J. Hydrolytic degradation of tyrosine-derived polycarbonates, a class of new biomaterials. part I: Study of model compounds. Biomaterials 2000;21(23):2371–2378.
72. Muggli DS, Burkoth AK, Anseth KS. Crosslinked polyanhydrides for use in orthopedic applications: Degradation behavior and mechanics. J Biomed Mater Res 1999;46(2):271–278.
73. Daniels AU, et al. Evaluation of absorbable poly(ortho esters) for use in surgical implants. J Appl Biomater 1994;5(1): 51–64.
74. Laurencin CT, et al. Use of polyphosphazenes for skeletal tissue regeneration. J Biomed Mater Res 1993;27(7): 963–973.
75. Zdrahala RJ, Zdrahala IJ. Biomedical applications of polyurethanes: A review of past promises, present realities, and a vibrant future. J Biomater Appl 1999;14(1):67–90.
76. James K, Kohn J. In: Park K, editor. Pseudo-Poly (Amino Acids):Examples for Synthetic Materials Derived from Natural Metabolites. Washington (DC): American Chemical Society; 1997.
77. Peppas NA, Bures P, Leobandung W, Ichikawa H. Hydrogels in pharmaceutical formulations. Eur J Pharm Biopharm 2000;50(1):27–46.
78. Elisseeff J, et al. Photoencapsulation of chondrocytes in poly(ethylene oxide)-based semi-interpenetrating networks. J Biomed Mater Res 2000;51(2):164–171.
79. Gobin AS, West JL. Cell migration through defined, synthetic ECM analogs. FASEB J 2002;16(7):751–753.
80. Mann BK, et al. Smooth muscle cell growth in photopolymerized hydrogels with cell adhesive and proteolytically degradable domains: Synthetic ECM analogs for tissue engineering. Biomaterials 2001;22(22):3045–3051.
81. Anseth KS, et al. In situ forming degradable networks and their application in tissue engineering and drug delivery. J Control Release 2002;78(1–3):199–209.
82. Elisseeff J, et al. Transdermal photopolymerization for minimally invasive implantation. Proc Natl Acad Sci USA 1999;96(6):3104–3107.
83. Rutkowski GE, Heath CA. Development of a bioartificial nerve graft. II. nerve regeneration in vitro. Biotechnol Prog 2002; 18(2):373–379.
84. Mikos AG, et al. Preparation of poly(glycolic acid) bonded fiber structures for cell attachment and transplantation. J Biomed Mater Res 1993;27(2):183–189.
85. Mikos AG, Lyman MD, Freed LE, Langer R. Wetting of poly(L-lactic acid) and poly(DL-lactic-co-glycolic acid) foams for tissue culture. Biomaterials 1994;15(1):55–58.
86. Mooney DJ, et al. Novel approach to fabricate porous sponges of poly(D,L-lactic-*co*-glycolic acid) without the use of organic solvents. Biomaterials 1996;17(14):1417–1422.
87. Whang K, Thomas H, Healy KE. A novel method to fabricate bioabsorbable scaffolds. Polymer 1995;36:837–841.
88. Lo H, Ponticiello MS, Leong KW. Fabrication of controlled release biodegradable foams by phase separation. Tissue Eng 1995;1:15–27.
89. Thomson RC, Yaszemski MJ, Powers JM, Mikos AG. Hydroxyapatite fiber reinforced poly(alpha-hydroxy ester) foams for bone regeneration. Biomaterials 1998;19(21):1935–1943.
90. Widmer MS, et al. Manufacture of porous biodegradable polymer conduits by an extrusion process for guided tissue regeneration. Biomaterials 1998;19(21):1945–1955.
91. Thomson RC, Yaszemski MJ, Powers JM, Mikos AG. Fabrication of biodegradable polymer scaffolds to engineer trabecular bone. J Biomater Sci Polym Ed 1995;7(1):23–38.
92. Mikos AG, Sarakinos G, Leite SM, Vacanti JP, Langer R. Laminated three-dimensional biodegradable foams for use in tissue engineering. Biomaterials 1993;14(5):323–330.
93. Park A, Wu B, Griffith LG. Integration of surface modification and 3D fabrication techniques to prepare patterned poly(L-lactide) substrates allowing regionally selective cell adhesion. J Biomater Sci Polym Ed 1998;9(2):89–110.
94. Peter SJ, Kim P, Yasko AW, Yaszemski MJ, Mikos AG. Crosslinking characteristics of an injectable poly(propylene fumarate)/beta-tricalcium phosphate paste and mechanical properties of the crosslinked composite for use as a biodegradable bone cement. J Biomed Mater Res 1999;44(3):314–321.
95. Bryant SJ, Durand KL, Anseth KS. Manipulations in hydrogel chemistry control photoencapsulated chondrocyte behavior and their extracellular matrix production. J Biomed Mater Res A 2003;67(4):1430–1436.
96. Andersson H, van den Berg A. Microfabrication and microfluidics for tissue engineering: State of the art and future opportunities. Lab Chip 2004;4(2):98–103.
97. Berry CC, Campbell G, Spadiccino A, Robertson M, Curtis AS. The influence of microscale topography on fibroblast attachment and motility. Biomaterials 2004;25(26): 5781–5788.
98. Motlagh D, Senyo SE, Desai TA, Russell B. Microtextured substrata alter gene expression, protein localization and the shape of cardiac myocytes. Biomaterials 2003;24(14): 2463–2476.
99. Miller C, Jeftinija S, Mallapragada S. Micropatterned schwann cell-seeded biodegradable polymer substrates significantly enhance neurite alignment and outgrowth. Tissue Eng 2001;7(6):705–715.
100. Rutkowski GE, Miller CA, Jeftinija S, Mallapragada SK. Synergistic effects of micropatterned biodegradable conduits and schwann cells on sciatic nerve regeneration. J Neural Eng 2004;1:151–157.
101. Pierschbacher MD, Ruoslahti E. Variants of the cell recognition site of fibronectin that retain attachment-promoting activity. Proc Natl Acad Sci USA 1984;81(19):5985–5988.
102. Hubbell JA, Massia SP, Desai NP, Drumheller PD. Endothelial cell-selective materials for tissue engineering in the vascular graft via a new receptor. Biotechnology (NY) 1991;9(6):568–572.
103. Tashiro K, et al. A synthetic peptide containing the IKVAV sequence from the A chain of laminin mediates cell attachment, migration, and neurite outgrowth. J Biol Chem 1989;264(27):16174–16182.
104. Hunter DD, et al. An LRE (leucine-arginine-glutamate)-dependent mechanism for adhesion of neurons to S-laminin. J Neurosci 1991;11(12):3960–3971.
105. Hsiao CC, et al. Receding cytochrome P450 activity in disassembling hepatocyte spheroids. Tissue Eng 1999;5(3):207–221.
106. Ronziere MC, et al. Ascorbate modulation of bovine chondrocyte growth, matrix protein gene expression and synthesis in three-dimensional collagen sponges. Biomaterials 2003;24(5):851–861.

107. Buti M, et al. Influence of physical parameters of nerve chambers on peripheral nerve regeneration and reinnervation. Exp Neurol 1996;137(1):26–33.
108. Yoffe B, et al. Cultures of human liver cells in simulated microgravity environment. Adv Space Res 1999;24(6):829–836.
109. Brown RA, et al. Tensional homeostasis in dermal fibroblasts: Mechanical responses to mechanical loading in three-dimensional substrates. J Cell Physiol 1998;175(3): 323–332.
110. Niklason LE, et al. Functional arteries grown in vitro. Science 1999;284(5413):489–493.
111. Fink C, et al. Chronic stretch of engineered heart tissue induces hypertrophy and functional improvement. FASEB J 2000;14(5):669–679.
112. Carver SE, Heath CA. Increasing extracellular matrix production in regenerating cartilage with intermittent physiological pressure. Biotechnol Bioeng 1999;62(2):166–174.
113. Ming G, et al. Electrical activity modulates growth cone guidance by diffusible factors. Neuron 2001;29(2):441–452.
114. Radisic M, et al. Functional assembly of engineered myocardium by electrical stimulation of cardiac myocytes cultured on scaffolds. Proc Natl Acad Sci USA 2004;101(52): 18129–18134.
115. Cannon TW, et al. Improved sphincter contractility after allogenic muscle-derived progenitor cell injection into the denervated rat urethra. Urology 2003;62(5):958–963.
116. Zhao M, Dick A, Forrester JV, McCaig CD. Electric field-directed cell motility involves up-regulated expression and asymmetric redistribution of the epidermal growth factor receptors and is enhanced by fibronectin and laminin. Mol Biol Cell 1999;10(4):1259–1276.
117. Bogie KM, Reger SI, Levine SP, Sahgal V. Electrical stimulation for pressure sore prevention and wound healing. Assist Technol 2000;12(1):50–66.
118. Spadaro JA. Mechanical and electrical interactions in bone remodeling. Bioelectromagnetics 1997;18(3):193–202.
119. Sikavitsas VI, Bancroft GN, Mikos AG. Formation of three-dimensional cell/polymer constructs for bone tissue engineering in a spinner flask and a rotating wall vessel bioreactor. J Biomed Mater Res 2002;62(1):136–148.
120. Carver SE, Heath CA. Influence of intermittent pressure, fluid flow, and mixing on the regenerative properties of articular chondrocytes. Biotechnol Bioeng 1999;65(3):274–281.
121. Matsuda N, Yokoyama K, Takeshita S, Watanabe M. Role of epidermal growth factor and its receptor in mechanical stress-induced differentiation of human periodontal ligament cells in vitro. Arch Oral Biol 1998;43(12):987–997.
122. Schumacker PT. Straining to understand mechanotransduction in the lung. Am J Physiol Lung Cell Mol Physiol 2002;282(5):L881–2.
123. Kastelic J, Galeski A, Baer E. The multicomposite structure of tendon. Connect Tissue Res 1978;6(1):11–23.
124. Augat P, et al. Anisotropy of the elastic modulus of trabecular bone specimens from different anatomical locations. Med Eng Phys 1998;20(2):124–131.
125. Lanza RP, et al. Xenotransplantation of cells using biodegradable microcapsules. Transplantation 1999;67(8):1105–1111.
126. Iwata H, et al. Bioartificial pancreas research in japan. Artif Organs 2004;28(1):45–52.
127. Shapiro AM, et al. Islet transplantation in seven patients with type 1 diabetes mellitus using a glucocorticoid-free immunosuppressive regimen. N Engl J Med 2000;343(4): 230–238.
128. Seaberg RM, et al. Clonal identification of multipotent precursors from adult mouse pancreas that generate neural and pancreatic lineages. Nat Biotechnol 2004;22(9):1115–1124.
129. Ianus A, Holz GG, Theise ND, Hussain MA. In vivo derivation of glucose-competent pancreatic endocrine cells from bone marrow without evidence of cell fusion. J Clin Invest 2003;111(6):843–850.
130. Chan C, et al. Hepatic tissue engineering for adjunct and temporary liver support: Critical technologies. Liver Transpl 2004;10(11):1331–1342.
131. Selden C, Hodgson H. Cellular therapies for liver replacement. Transpl Immunol 2004;12(3–4):273–288.
132. Kulig KM, Vacanti JP. Hepatic tissue engineering. Transpl Immunol 2004;12(3–4):303–310.
133. Salgado AJ, Coutinho OP, Reis RL. Bone tissue engineering: State of the art and future trends. Macromol Biosci 2004;4(8):743–765.
134. LeGeros RZ. Properties of osteoconductive biomaterials: Calcium phosphates. Clin Orthop 2002;(395)(395):81–98.
135. Pittenger MF, et al. Multilineage potential of adult human mesenchymal stem cells. Science 1999;284(5411):143–147.
136. Droppert PM. The effects of microgravity on the skeletal system–a review. J Br Interplanet Soc 1990;43(1):19–24.
137. Perka C, et al. Segmental bone repair by tissue-engineered periosteal cell transplants with bioresorbable fleece and fibrin scaffolds in rabbits. Biomaterials 2000;21(11):1145–1153.
138. Lee JY, et al. Effect of bone morphogenetic protein-2-expressing muscle-derived cells on healing of critical-sized bone defects in mice. J Bone Joint Surg Am 2001;83-A(7):1032–1039.
139. He C, Xia L, Luo Y, Wang Y. The application and advancement of rapid prototyping technology in bone tissue engineering. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi 2004;21 (5):871–875.
140. Elisseeff J. Injectable cartilage tissue engineering. Expert Opin Biol Ther 2004;4(12):1849–1859.
141. Kim TK, et al. Experimental model for cartilage tissue engineering to regenerate the zonal organization of articular cartilage. Osteoarth Cart 2003;11(9):653–664.
142. Quinones-Baldrich WJ, et al. Is the preferential use of polytetrafluoroethylene grafts for femoropopliteal bypass justified? J Vasc Surg 1988;8(3):219–228.
143. Schmedlen RH, Elbjeirami WM, Gobin AS, West JL. Tissue engineered small-diameter vascular grafts. Clin Plast Surg 2003;30(4):507–517.
144. Barron V, et al. Bioreactors for cardiovascular cell and tissue growth: A review. Ann Biomed Eng 2003;31(9): 1017–1030.
145. Daly CD, Campbell GR, Walker PJ, Campbell JH. In vivo engineering of blood vessels. Front Biosci 2004;9:1915–1924.
146. Auger FA, et al. Tissue-engineered skin substitutes: From in vitro constructs to in vivo applications. Biotechnol Appl Biochem 2004;39(Pt 3):263–275.
147. Boyce ST, et al. The 1999 clinical research award. cultured skin substitutes combined with integra artificial skin to replace native skin autograft and allograft for the closure of excised full-thickness burns. J Burn Care Rehabil 1999;20(6):453–461.
148. Supp DM, Wilson-Landy K, Boyce ST. Human dermal microvascular endothelial cells form vascular analogs in cultured skin substitutes after grafting to athymic mice. FASEB J 2002;16(8):797–804.
149. Black AF, et al. In vitro reconstruction of a human capillary-like network in a tissue-engineered skin equivalent. FASEB J 1998;12(13):1331–1340.

150. Andreadis ST. Gene transfer to epidermal stem cells: Implications for tissue engineering. Expert Opin Biol Ther 2004;4(6):783–800.
151. Cooley J. Follicular cell implantation: An update on hair follicle cloning. Facial Plast Surg Clin North Am 2004; 12(2):219–224.
152. Belkas JS, Shoichet MS, Midha R. Peripheral nerve regeneration through guidance tubes. Neurol Res 2004;26(2):151–160.
153. Stanec S, Stanec Z. Reconstruction of upper-extremity peripheral-nerve injuries with ePTFE conduits. J Reconstr Microsurg 1998;14(4):227–232.
154. Hentz VR, et al. The nerve gap dilemma: A comparison of nerves repaired end to end under tension with nerve grafts in a primate model. J Hand Surg (Am) 1993;18(3):417–425.
155. Fields RD, Le Beau JM, Longo FM, Ellisman MH. Nerve regeneration through artificial tubular implants. Prog Neurobiol 1989;33(2):87–134.
156. Heath CA, Rutkowski GE. The development of bioartificial nerve grafts for peripheral-nerve regeneration. Trends Biotechnol 1998;16(4):163–168.
157. Patrick Jr CW. et al. Muristerone A-induced nerve growth factor release from genetically engineered human dermal fibroblasts for peripheral nerve tissue engineering. Tissue Eng 2001;7(3):303–311.

Reading List

Atala A, Lanza RP. Methods of tissue engineering. San Diego: Academic Press; 2001. p 1285.
Lanza RP, Langer RS, Vacanti J. Principles of tissue engineering. San Diego: Academic Press; 2000. p 995.
Saltzman WM. Tissue engineering: Engineering principles for the design of replacement organs and tissues. Oxford, New York: Oxford University Press; 2004. p 523.

See also BIOMATERIALS: TISSUE ENGINEERING AND SCAFFOLDS; SKIN TISSUE ENGINEERING FOR REGENERATION.

ENVIRONMENTAL CONTROL

DENIS ANSON
College Misericordia
Dallas, Pennsylvania

INTRODUCTION

Electronic aids to daily living (EADLs) are devices that can be used to control electrical devices in the client's environment (1). Before 1998 (2), these devices were generally known by the shorter term, "Environmental Control Unit" (ECU). Technically, this term should be reserved for furnace thermostats and similar controls. The more generic EADL applies to control of lighting and temperature, but also applies to control of radios, televisions, telephones, and other electrical and electronic devices in the environment of the client (3,4). See Fig. 1.

These systems all contain some method for the user to provide input to the EADL, some means of determining the current state of the device to be controlled (although this is often visual inspection of the device itself, since EADLs are generally thought of as being applied to the immediate environment), and a means of exerting control over the targeted device. The degree and generalization of control differs among various EADL systems. These systems may provide a means of switching power to the target device, of controlling the features of an external device, or may subsume an external device to provide enhanced control internally.

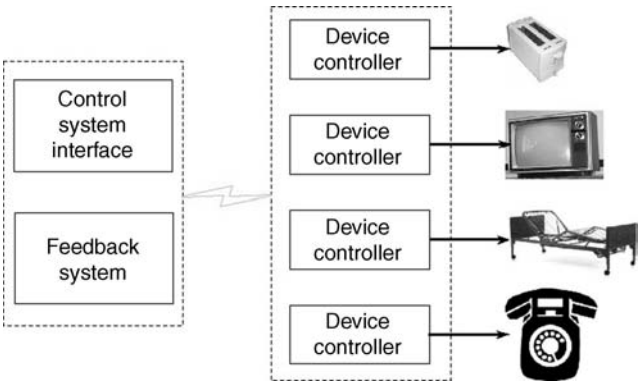


Figure 1. Components of an EADL system.

POWER SWITCHING

The simplest EADLs only provide switching of the electrical supply for devices in a room. Although not typically considered as EADLs, the switch-adapted toys provided to severely disabled children would, formally, be included in this category of EADLs. To adapt a conventional battery powered toy, the therapist inserts a "battery interrupter" to allow an external switch to control the flow of power from the batteries to the workings of the toy. Power switching EADLs operate in precisely the same manner. A switch is placed in series with the device to be controlled, so that the inaccessible power switch of the device can be left in the "ON" position, and the device can be activated by a more accessible external switch. To provide control over appliances and lights in the immediate environment, primitive EADL systems consisted of little more than a set of electrical switches and outlets in a box that connected to devices within a room via extension cords. Control 1(5), for example, allowed the connection of eight devices to the receptacles of the control unit. Such devices are limited in their utility and safety, since extension cords pose safety hazards to people in the environment through risks of falls (tripping over the extension cords) and fires (overheated or worn cords). Because of the limitations posed by extension cords, EADL technology was driven to use remote switching technologies (Fig. 2).

Second generation EADL systems used various remote control technologies to activate power to electrical devices in the environment. These strategies include the use of ultrasonic pulses [e.g., TASH Ultra 4 (6)] (Fig. 4) infrared (IR) light [e.g., Infrared Remote Control (7)], and electrical signals propagated through the electrical circuitry of the home [e.g., X-10 (8) Fig. 3]. All of these switching

Figure 2. Electralink power switching module.

Figure 3. X-10 switching modules.

technologies remain in use, and some are used for much more elaborate control systems. Here we are only considering power switching, however (Fig. 4).

The most prevalent power-switching EADL control system is that produced by the X-10 Corporation. The X-10 system uses electrical signals sent over the wiring of a home to control power modules that are plugged into wall sockets in series with the device to be controlled. (In a series connection, the power module is plugged into the wall, and the remotely controlled device is plugged into the power module.) The X-10 supports up to 16 channels of control, with up to 16 modules on each, for a total of up to 256 devices controlled by a single system. The signals used to control X-10 modules will not travel through the home's power transformer so, in single family dwellings, there is no risk of interfering with devices in a neighbor's home. This is not necessarily true, however, in an apartment setting, where it is possible for two X-10 users to inadvertently control each other's devices. The general set-up of early X-10 devices was to control up to 16 devices on an available channel so that such interference would not occur. In some apartments, the power within a single unit may be on different "phases" of the power supplied to the building. (These phases are required to provide 220-V power for some appliances.) If this is the case, the X-10 signals from a controller plugged into one phase will not cross to the second phase of the electrical wiring. A special "phase cross-over" is available from X-10 to overcome this problem. The X-10 modules, in addition to switching power on and off, can be used, via special lighting modules, to dim and brighten room lighting. These modules work only with incandescent lighting, but add a degree of control beyond simple switching. For permanent installations, the wall switches and receptacles of the home may be replaced with X-10 controlled units. Because X-10 modules do not prevent local control, these receptacles and switches will work like standard units, with the added advantage of remote control.

When they were introduced in the late 1970s, X-10 modules revolutionized the field of EADLs. Prior to X-10, remote switching was a difficult and expensive endeavor, restricted largely to applications for people with disabilities and to industrial applications. The X-10 system, however, was intended as a convenience for able-bodied people who did not want to walk across a room to turn on a light. Because the target audience was able to perform the task without remote switching, the technology had to be inexpensive enough that it was easier to pay the cost than get out of a chair. The X-10 made it possible for an able-bodied

Figure 4. Tash Ultra-4 power switching modules.

person to remotely control electrical devices for under $100, where most disability-related devices cost several thousand dollars.

Interestingly, the almost universal adoption of X-10 protocols by disability related EADLs did not result in sudden price drops in the disability field. In cases where simple power switching provided adequate control, many clinicians continue to adapt mass-market devices for individuals with disabilities. While this may not be a good use of clinician time, it does allow those with limited funding to gain a degree of control over their environments.

FEATURE CONTROL

As electronic systems became more pervasive in the home, simply switching of lights and coffee pots failed to meet the needs of individuals with disabilities who wanted to control the immediate environment. With wall current control, a person with a disability might be able to turn a light or television on and off, but would have no control beyond that. A person with a disability might want to be able to surf cable channels as much as an able-bodied person with a television remote control (9). When advertisements are blaring from the speakers, a person with a disability might want to be able to turn down the sound, or tune to another radio station. Because the ability to control a radio from across the room became a sales advantage when marketing to sedentary, able-bodied adults, nearly all home electronic devices are now delivered with a remote control, generally using IR signals. Most of these remote controls are not usable by a person with a disability, however, due to the small buttons and labels that require fine motor control and good sensory discrimination (Fig. 5).

The EADL systems designed to provide access to the home environment of a person with a disability must provide more than on/off control of home electronics. They must also provide control of the features of home electronic devices. Because of this need, EADL systems frequently have hybrid capabilities. They will incorporate a means of directly switching power to remote devices, often using X-10 technology. This allows control of devices such as lights, fans, and coffee pots, as well as electrical door openers and other specialty devices (10). They will also typically incorporate some form of IR remote control, which will allow them to mimic the signals of standard remote control devices. This control will be provided either by programming in the standard sequences for all commercially available VCRs, televisions, and satellite decoders, or through teaching systems, where the EADL learns the codes beamed at it by the conventional remote. Preprogrammed control codes allow a simple set-up process to enable the EADL to control various devices, but only those devices whose codes existed prior to the manufacture of the EADL. The advantage of the learning approach is that it can learn any codes, even those that have not yet been invented. The disadvantage is that the controls must be taught, requiring more set-up and configuration time for the user and caregivers. In addition, there are cases where the internal IR switching speed of the EADL differs enough from that of the device to be controlled that some signals cannot be reproduced reliably.

Figure 5. Imperium 200H provides infrared remote control of entertainment systems, power switching, and hospital bed control.

Infrared remote control, as adopted by most entertainment systems controllers, is limited to approximate line of sight control. Unless the controller is aimed in the general direction of the device to be controlled (most have wide dispersion patterns), the signals will not be received. This means that an EADL cannot directly control, via IR, any device not located in the same room. However, IR repeaters, such as the X-10 Powermid (11) can overcome this limitation by using radio signals to send the control signals received in one room to a transmitter in the room of the device to be controlled. With a collection of repeaters, a person would be able to control any infrared device in the home from anywhere else in the home.

One problem that is shared by EADL users and able-bodied consumers is the proliferation of remote control devices. Many homes now are plagued with a remote control for the television, the cable–satellite receiver, the DVD player/VHS recorder (either one or two devices), the home stereo/multimedia center, and other devices, all in the same room. Universal remote controls allow switching from controlling one device to another, but are often cumbersome to control. Some hope is on the horizon for improved control of home audiovisual devices with less difficulty. In November of 1999, a consortium of eight home electronics manufacturers released a set of guidelines for home electronics called HAVi (12). The HAVi specification will allow compliant home electronics to communicate so that any HAVi remote control can operate the features of all of the HAVi devices sharing the standard. A single remote control can control all of the audiovisual devices in the home, through a single interface. Such standards are effective to the extent that they are actually implemented. As of the summer of 2004, the HAVi site lists six products, from two manufacturers, that actually use the HAVi standard.

The Infrared Data Association (13) (IrDA) is performing similar specifications work focusing purely on IR controls. The IrDA standard will allow an IR remote control to operate features of computers, home audiovisual equipment, and appliances equipped with IR controls through a

single standard protocol. In addition to allowing a single remote control to control a wide range of devices, IrDA standards will allow other IrDA devices, such as PDAs, personal computers, and augmentative communications systems to control home electronics. Having a single standard for home electronics will simplify the design of EADL systems for people with disabilities.

A more recent standard, V2 (14), offers a much greater level of control. If fully implemented, V2 would allow a single EADL device to control the all of the features of all electronic devices in its vicinity, from the volume of the radio through the setting of the thermostat in the hall, to the "Push to Walk" button on the cross-walk. Using a V2 EADL, a person with a disability could move from environment to environment, and be able to control the V2 enabled devices in any location.

One interesting aspect of feature control by EADLs is the relationship between EADLs and computers. Some EADL systems, such as the Quartet Simplicity (15), include features to allow the user to control a personal computer through the EADL. In general, this is little more than a passthrough of the control system of the EADL to a computer access system. Other EADLs, such as the PROXi (16), are designed to accept control inputs from a personal computer. The goal in both cases is to use the same input method to control a personal computer as to control the EADL. In general, the control demands of an EADL system are much less stringent than those for a computer. An input method that is adequate for EADL control may be very tedious for general computer controls. On the other hand, system that allows fluid control of a computer will not be strained by the need to also control an EADL. The "proper" source of control will probably have to be decided on a case-by-case basis.

One of the most important features of the environment to be controlled by an EADL is not generally thought of as an electronic device. In a study of EADL users conducted in Finland (17), users reported that the feature that provided the most gain in independence was the ability to open doors independently. While doors are not electronic devices, powered door openers may be, and can be controlled through the EADL.

SUBSUMED DEVICES

Finally, modern EADLs frequently incorporate some common devices that are more easily replicated than controlled remotely. Some devices, such as the telephone, are so pervasive that an EADL system can assume that a telephone will be required. Incorporating telephone electronics into the EADL is actually less expensive, due to telephone standardization, than inventing special systems to control a conventional telephone. Other systems designed for individuals with disabilities are so difficult to control remotely that the EADL must include the entire control system. Hospital bed controls, for example, have no provisions for remote control, but should be usable by a person with a disability. Hence, some EADLs include hospital bed controls internally, so that a person who is in the bed can control its features (Fig. 6).

A telephone conversation may be considered as having several components. The user must "pick up" to connect to the telephone system (able-bodied individuals do this by

Figure 6. Relax II scanning EADL with IR remote control and X-10 switching.

picking up the handset). If the user is responding to an incoming call, the act of picking up initiates the connection. If the user is "originating" a call, the act of picking up will be followed by "dialing", which describes the person to whom the user wishes to speak. When a connection is made (both parties have picked up), a conversation may ensue. At the end of the conversation, the call is "terminated" by breaking the connection.

Many EADL systems include a built-in speakerphone, which will allow the user to originate and answer telephone calls using the electronics of the EADL as the telephone. Because of the existing standards, these systems are generally analogue, single-line telephones, electronically similar to those found in the typical home. Many business settings now use multiline sets, which are not compatible with home telephones. Other settings use digital interchanges, which are also not compatible with conventional telephones. Finally, there is a move currently to use "Voice Over Internet Protocol" (VOIP) to bypass telephone billing and use the internet to carry telephone conversations. Because of this, the telephone built in to a standard EADL may not meet the needs of a disabled client in an office or other setting. Before recommending an EADL as an access solution for a client, therapists should check that the EADL communication system is compatible with the telecommunications systems in that location.

Because the target consumer for an EADL will have severe restrictions in mobility, the manufacturers of many of these systems consider that a significant portion of the customer's day will be spent in bed, and so include some sort of control system for standard hospital beds. These systems commonly allow the user to adjust head and foot height independently, extending the time the individual can be independent of assistance for positioning. As with telephone systems, different brands of hospital bed use different styles of control. It is essential that the clinician match the controls provided by the EADL with the inputs required by the bed to be controlled.

CONTROLLING EADLs

For an EADL to provide improved function to the individual with a disability, it must provide a control interface that is more usable than that of the devices it controls. The common strategies for control found in EADLs are scanning and voice command.

Scanning Control

While scanning control is not particularly useful for computer control (17), it may be quite satisfactory for control of an EADL. When controlling a computer, the user must select between hundreds of options, and perform thousands of selections per day. When controlling the environment, the user may select among dozens of options, but will probably not be making more than a hundred selections during the day. While the frequent waits for the desired action to be offered in a computer scanning input system can have a crippling effect on productivity, the difference between turning on a light *now* versus a *few seconds from now* is of little consequence. Because of this, many EADL systems provide a scanning control system as the primary means of controlling the immediate environment.

As with computer scanning, EADL scanning may be arranged in a hierarchical pattern. At the topmost level, the system may scan between lights, telephone, bed, entertainment, and appliances. When "lights" are selected, the system might scan between the various lights that are under EADL control. When a single light is selected, the system may scan between "Off", "Dimmer", "Brighter", and "On". As the number of devices to be controlled increases, the number of selections required to control a particular feature will increase, but the overall complexity of the device need not.

Voice Command

Like scanning, voice command may be significantly more functional in controlling an EADL than a computer. If the user is willing to learn specific voice commands for control that are selected to be highly differentiable, a voice command EADL system can be highly reliable. As the potential vocabulary increases, the likelihood of misrecognitions increases and the quality of control goes down.

As with all voice-command systems, background noise can impair the accuracy of recognition. An EADL may easily recognize the commands to turn on a radio in a quiet room, for example, but may be unable to recognize the command to turn off the radio when it is playing in the background. This problem will be exacerbated if the voice commands are to be received by a remote microphone, and when the user is not looking directly at the microphone. On the other hand, the use of a headset microphone can improve control accuracy, but at the cost of encumbering the user.

In addition to microphone type and quality, voice quality can affect the reliability of the system. For individuals with high level spinal cord injuries or other neurological deficits, voice quality can change significantly during the course of a day. A command that is easily recognized in the morning when the person is well rested may be ignored later in the day, after the user has fatigued. At this later time, the user's tolerance for frustration is also likely to be lessened. The combined effect of vocal changes and frustration may result in the user abandoning the device if an alternative control is not provided. While voice command of EADLs has a "magical" quality of control, for the person whose disability affects vocal quality or endurance, it can be a sporadic and limiting magic.

Other Control Strategies

For those systems that are controlled by a computer, any access method that can be used to control the computer can provide control over the environment as well. This includes mouse emulators and expanded keyboards as well as scanning and voice input. A recent study (18) has also demonstrated the utility of switch-encoding as a means of controlling the environment for children as young as 4, or, by extension, for individuals with significant cognitive limitations. In this strategy, the user closes one or two switches in coded patterns (similar to Morse code) to

operate the features of devices in the environment. The devices can be labeled with the control patterns for those users who cannot remember the controls, or during the training phase.

The technology now exists, though it has not been applied to EADL systems, to combine head and eye tracking technologies so that a person with a disability could control devices in the environment by simply looking at them. Head-tracking technology could determine, from the user's location, what objects were in the field of view. Eye tracking could determine the distance and fine direction of the intended object. A "heads-up" display might show the features of the device available for control, and provide the control interface. With such a system, a user wearing a headset might be able to turn on a light simply by looking at it and blinking. This type of control strategy, combined with the V2 control protocol, would allow a person with a disability truly "magical" control of the environment.

THE FUTURE OF EADLs

Recognizing that the EADL provides a bridge between the individual with a disability and the environment in which they live, EADL development is likely to occur in two directions. The interface between the individual and the EADL will be enhanced to provide better input to the EADL, and remote controls will become more pervasive so that more of the world can be controlled by the EADL.

The user's ability to control the EADL is a function of the ability of the person to emit controlled behavior. Scanning requires the presence or absence of an action, and ignores any grading. As assistive technologists develops sensing technologies to identify gradients of movement or neural activity, the number of selections that can be made directly increases. For example, current EADLs may be controlled by eye-blink for severely disabled individuals. But in the future, EADLs could use eye-tracking technology to allow a person to look at the device in the environment to be controlled, and then blink to activate it. Electroencephalography might, eventually, allow the control of devices in the environment by simply *thinking at* them (19,20).

The continued development of remote control technologies will, in the future, allow EADLs to control more of the environment that is currently available. Currently, EADLs can switch power to any device that plugs into the wall or that uses batteries. Feature control, however, is very limited. Cross-walk controls, elevators, and ATM machines do not have the means of remote control today, and are often beyond the reach of a person with a significant disability. Microwave ranges, ovens, and air conditioners also have feature controls that might be, but are not, remotely controllable. If interoperability standards like V2 are accepted, the controls that are provided for sedentary, able-bodied users may provide control for the EADLs of the future. It is generally recognized that the inclusion of remote control for EADL access is not cost-effective, but if the cost of providing remote control for able-bodied users becomes low enough, such options might be made available, which will allow EADLs of the future to control them as well.

BIBLIOGRAPHY

Cited References

1. Assistive Technology Partners. 2002. Direct Access Electronic Aids. Available at http://www.uchsc.edu/atp/library/fastfacts/Direct%20Access%20Electronic%20Aids.htm. Accessed July 26 2004.
2. MacNeil V. 1998. Electronic Aids to Daily Living. Team Rehabilitation Report, 53–56.
3. Center for Assistive Technology. Environmental Control Units. 2004. Available at http://cat.buffalo.edu/newsletters/ecu.php. Accessed 26 July 2004.
4. Cook AM, Hussey SM. Assistive technologies: Principles and practice. 2nd ed. Philadelphia: Mosby International; 2002.
5. Prentke Romich Company, 1022 Heyl Road, Wooster, OH 44691. Phone: (800) 262-1984. Available at http://www.prentrom.com/index.html.
6. Tash Inc., 3512 Mayland Ct., Richmond VA 23233 Phone: 1-(800) 463-5685 or (804) 747-5020. Available at http://www.tashinc.com/index.html.
7. DU-IT Control Systems Group, Inc., 8765 Township Road #513, Shreve, OH 44676, Phone: (216) 567-2906.
8. SmarthHome, Inc., 16542 Millikan Avenue, Irvine, CA 92606-5027, Phone: (949) 221-9200 x109.
9. Butterfield T. 2004. Environmental Control Units. Available at http://www.birf.info/artman/publish/article_418.shtml. Accessed 26 July 2004.
10. Quartet Technology. 2004. Quartet Technology, Inc.-News. Available at http://www.qtiusa.com/ProdOverview.asp?ProdTypeID =1. Accessed at 29 July 2004.
11. asiHome, 36 Gumbletown Rd., CS1, Paupack, PA 18451, Phone: 800-263-8608. Available at http://www.asihome.com/cgi-bin/ASIstore.pl?user_action=detail&catalogno= X10-PEX01.
12. HAVi, Inc., 40994 Encyclopedia Circle, Fremont, CA 94538 USA. Phone: (510) 979-1394. Available at http://www.havi.org/.
13. IrDA Corporate Office, P.O. Box 3883, Walnut Creek, CA 94598. Phone: (925) 943-6546. Available at http://www.irda.org/index.cfm.
14. InterNational Committee for Information Technology Standards. What exactly is V2 - and How Does It Work? Available at http://www.myurc.com/whatis.htm. Accessed at 30 July 2004.
15. Quartet Technology, Inc., 87 Progress Avenue, Tyngsboro, Massachusetts 01879, phone: 1-(978) 649-4328.
16. Madentec, Ltd., 4664 99 St., Edmonton, Alberta, Canada T6E 5H5, phone: (877) 623-3682. Available at http://madentec.com.
17. Anson DK. Alternative Computer Access: A Guide to Selection. Philadelphia: F. A. Davis; 1997.
18. Anson D, Ames C, Fulton L, Margolis M, Miller M. 2004. Patterns For Life: A Study of Young Children's Ability to Use Patterned Switch Closures for Environmental Control. Available at http://atri.misericordia.edu/Papers/Patterns.php. Accessed 4 Oct 2004.
19. Howard T. Beyond the Big Barrier. Available at http://www.cs.man.ac.uk/aig/staff/toby/writing/PCW/bci.html. Accessed 11 Oct 2004.
20. Wolpaw JR. 2004. Brain-Computer Interfaces For Communication And Control. Available at http://www.nichd.nih.gov/about/ncmrr/symposium/wolpaw_abstract.htm. Accessed 11 Oct 2004.

See also COMMUNICATIVE DISORDERS, COMPUTER APPLICATIONS FOR; MOBILITY AIDS.

EQUIPMENT ACQUISITION

ROBERT STIEFEL
University of Maryland
Baltimore, Maryland

INTRODUCTION

Equipment acquisition is the process by which a hospital introduces new technology into its operations. The process involves determining the hospital's needs and goals with respect to new technology and equipment, how best to meet those needs, and instituting the decisions. The process involves virtually every clinical and support department of the hospital. This is consistent with the Joint Commission on Accreditation of Healthcare Organizations (JCAHO) (1) medical equipment management standards which require hospitals to have a process for medical equipment acquisition.

Unfortunately, in many hospitals it is a ritual, the control and details of which are jealously guarded. In fact, the needs of the hospital's operation would be much better served if all departments knew how the process worked. If the rules of the process were known and based upon the stated goals and priorities of the institution, then the people who must attempt to justify requests for new equipment would be able to do their jobs better, and with subsequently better results. If a new technology can improve the hospital's finances and/or clinical or support functions, then the methods by which this can be used to justify requests should be clearly explained. If the hospital's reimbursement structure is such that the introduction of new technology is difficult, but the improvement of current functions is more easily funded, then this should be explained.

In short, there should be a policy and procedure for the method by which the hospital acquires new equipment. It should define what the hospital means by a capital expenditure, reflect the hospital's overall goals and objectives, clearly state how to prepare a justification, and explain, at least in general terms, how a decision is to be made about the funding of a proposal.

The scope of this article is limited to the justification, selection, and implementation of new medical technology and equipment (Table 1). It is not intended to provide cookbook methods to be followed exactly, but instead to explain principles that can be applied to different hospitals and their differing needs. This seems to be particularly appropriate because needs vary not only between hospitals, but also with time.

Table 1. Equipment Acquisition: Outline of Process

- **Justification**
  - Needs assessment
  - Proposal
    - Clinical
    - Financial
    - Environmental
  - Budget request
- **Selection**
  - Literature review
    - Library
    - Subscriptions
    - Standards
    - Manufacturer's literature
  - Vendor list
    - Request for proposal
    - Preliminary review
  - Engineering testing
    - Safety
    - Performance
  - Contact other users
  - Clinical testing
    - Use in expected application
    - Questionnaire or interview
  - Assessment
    - Ranking
    - Requests for quotations
  - Final choice
    - Negotiate
    - Contract
- **Implementation**
  - Purchase order
  - Installation
  - Acceptance testing
  - Training
    - Operator
    - Service
- **Conclusion**
  - Report
  - Follow-up

JUSTIFICATION PROCESS

The justification process lays the groundwork for the acquisition of medical equipment. The better the justification, the more dependable the results will be. If the rules of the justification process are carefully planned, and just as carefully followed, the equipment acquisition function of the hospital will be respected and adhered to by other components of the hospital system. It is via the justification process that the hospital's needs are recognized, proposals are created to meet these needs, and sufficient funds are budgeted to fulfill the most acceptable proposals.

Needs assessment is the first step in the justification process. The requirement for new equipment can be based upon a variety of disparate requirements. There can be a need for new technology or expansion of an existing service. Need for equipment can be based upon cost effectiveness of new technology, safety, maintenance costs, or simply a need to replace old equipment. The justification for acquiring equipment based upon any of these reasons must be supported by facts.

The age of equipment by itself is not sufficient justification for replacing equipment. If the age of equipment exceeds accepted guidelines, and maintenance costs, for example, are also exceeding accepted guidelines, then the replacement of equipment is adequately justified. What this points out, however, is that there must be guidelines for replacement of equipment based upon both age and maintenance costs.

A general guideline is that when equipment is over 7 years old, it is time to start considering its replacement. The age when equipment should be considered for replacement can also be based on its depreciation life, which varies depending on the type of device. Wear and tear, along with the advancement of the state of the art in equipment design, will start catching up with the equipment at about this time. When total maintenance costs exceed approximately one and one-half times replacement cost or when an individual repair will cost more than one-half the replacement cost, it is probably more appropriate to consider replacing the equipment. Age and/or maintenance costs by themselves do not necessarily require equipment replacement. There can be mitigating circumstances, such as the fact that the older equipment might be impossible to replace. If an item is one of a standardized group, or part of a larger system, it might be unrealistic to consider replacing the entire group or system.

Safety considerations should be relatively easy to document. If the performance of equipment puts it out of

compliance with accepted standards, the standards and the performance of the equipment can be documented. Again, however, judgment can mitigate these standards. If a piece predates current safety or performance standards, it is perfectly acceptable to continue using it if clinical and engineering personnel believe it still performs safely and as intended. This judgment should also be documented.

Cost effectiveness is, by itself, adequate justification for equipment acquisition. If it can be shown that within 3–5 years, the acquisition of new equipment will save more than the cost of the acquisition, it will be money well spent. The justification must be done very carefully, however, to recognize all costs that will be incurred by the acquisition of new equipment, and any additional operating costs.

If an existing clinical service is covering its costs and it can be shown that there is a need to expand the service, the equipment necessary to support the expansion can be cost justified. Again, the justification must be done carefully to ascertain that there are sufficient trained people, or people who can be trained to use the equipment, as well as a sufficient population of patients requiring the service.

The acquisition of new technology requires the most difficult and demanding justification. The technology itself must be assessed. Determine whether the technology is viable, provides a needed service, and will be accepted. There must be clinical professionals capable of utilizing the technology or in a position to be trained to do so. Cost justification of new technology is equally difficult. Very careful thought will have to be given to identifying all costs and revenues. The cost of the equipment, the cost of the supplies necessary to operate it, and the cost of the personnel to operate it must all be determined.

In addition, the space requirements and special installation requirements, such as electrical power, computer networking, air conditioning, plumbing, or medical gases, will all have to be determined. Maintenance costs will have to be identified and provisions made to cover them. It must be determined that there is an adequate population base that will provide patients to take advantage of the new technology. State and local approval (certificate of need) may also have to be met. The entire process is very time consuming, and should be carefully planned to meet the scheduling of the hospital.

Once needs have been justified, primarily by the people intending to apply the new equipment, it will be necessary to create a formal proposal. The formal proposal should be prepared by a select group of people (medical, nursing, purchasing, administration, and clinical engineering) most closely associated with the management and use of the equipment. Physicians are concerned with the function performed by the equipment, nursing with the equipment's operations, and clinical engineering with the design, safety and performance, and dependability of the equipment. Administration and purchasing are involved with managing costs. Information technology staff may need to be involved if equipment is networked, or facilities staff if there are special installation requirements.

The proposal must provide a precise definition of the clinical needs, the intended usage of the equipment, any restrictions of a clinical nature, and a thorough financial plan. The financial planning, in particular, should include accepted planning techniques. For example, life-cycle cost analysis is perhaps the most thorough method for determining the financial viability of a new project. (Life-cycle analysis is the method of calculating the total cost of a project by including the initial capital cost, the operating costs for the expected lifetime of the equipment, and the time cost of money.)

At this stage, it is appropriate to consider a variety of physical or environmental factors that affect or are affected by the proposed new equipment. The equipment will require space for installation, use, and maintenance. Its power requirements might call for a special electrical source, medical gases or vacuum, or a water supply and drain. Its size might prevent if from passing through doors. The equipment might require special air-handling consideration if it generates heat or must be operated in a temperature- and/or humidity-controlled environment. Its weight might preclude transport in elevators or require modification of floors, or its sensitivity to vibration might require a special installation. If the equipment is either susceptible to or generates electrical or magnetic fields, special shielding may be required. There might be special standards that restrict the selection or installation of the type of equipment.

After the staff intending to use and manage the equipment have completed their proposal, it will be necessary to present this to the committee responsible for budgeting new equipment or new projects. This "capital equipment budget committee" will typically be chaired by an individual from the hospital's budget office and should include representatives from central administration, nursing, and clinical engineering. It is their responsibility to review proposals for completeness and accuracy as well as feasibility with respect to the hospital's long-range plans and patient population. Their judgment to approve or disapprove will be the necessary step before providing funds.

SELECTION PROCESS

Once the acquisition of new equipment has been justified, planned, and budgeted, the next step is to select the equipment that will actually be purchased. Again, this is a formal process, with sequential steps that are necessary to achieve the most appropriate equipment selection. There are commercial software products available that help hospitals establish priorities, develop proposals, and select capital equipment (e.g., Strata Decision Technology).

For most equipment, a standing capital equipment committee can oversee the selection and acquisition process. This committee should at least include representatives from clinical engineering, finance, and purchasing. It might also include representatives from nursing and information technology.

For major equipment or new technology, a selection committee should be formed specifically for each acquisition. Such a committee should include physician, nursing, and administrative personnel from the area for which the equipment is intended, and a representative from clinical engineering. Since the representative from clinical engineering will serve on virtually all of these ad hoc selection committees, this person's experience would make him/her

the best choice for chairperson. Realistically, however, it might be more politically expedient to allow one of the representatives from the area to chair the committee.

The first step involves a literature review. A library search should be conducted, and the clinical engineering department's professional subscriptions reviewed, for example, *Health Devices* and *Biomedical Instrumentation and Technology*. Look for applicable standards from AAMI or the American National Standards Institute (ANSI). Obtain product literature from the manufacturers being considered. Research the information maintained by the FDA Center for Devices and Radiological Health (CDRH) at http://www.fda.gov/cdrh. The FDA-CDRH Manufacturer and User Facility Device Experience Database (MAUDE) contains reports of adverse events involving medical devices, and their "Safety Alerts, Public Health Advisories, and Notices" page contains safety-related information on medical devices.

A list of proposed vendors to be contacted can be created by consulting the *Health Devices Sourcebook* (ECRI) (2). These vendors should be contacted and their literature on the type of equipment being evaluated requested. Part of the evaluation includes evaluating the manufacturers and vendors. Commencing with this initial contact, notes should be kept on the responsiveness and usefulness of the representatives contacted.

A request for proposal (RFP) should be written on the basis of the needs determined during the justification process and the information acquired from the literature review. When the selection criteria are straightforward, the RFP and the request for quotation (RFQ) can be combined. For more complex selections, such as equipment systems or new technology, the RFP and RFQ should be separate processes.

The RFP should be carefully written, well organized, and thoroughly detailed. It should contain useful background information about the institution and the specific user area. Applicable documents, such as drawings, should be either included or explicitly made available for review by the vendors. The equipment requirements should include a statement regarding the major objectives to be fulfilled by the equipment. The description of the specific requirements must include *what* is to be done, but should avoid as much as possible restrictions on *how* it is to be done. It is likely that, given reasonable latitude in addressing the equipment requirements, manufacturers can make useful suggestions based upon their experience and the unique characteristics of their equipment.

The RFP should contain a description of the acceptance testing that will be conducted before payment is approved. It should also contain a request for a variety of ancillary information: available operator and service documentation; training materials and programs; warranties; and maintenance options and facilities. There should also be a request for the names of at least three users of comparable equipment, located as close as possible to the hospital so that site visits can be conveniently arranged. The RFP should be reviewed by the entire in-house evaluation committee.

A cover letter should accompany the RFP to explain the general instructions for submitting proposals: who to call for answers to questions, deadline for submission, format, and so on. When appropriate, there can also be such information as to how the proposals will be evaluated, how much latitude the vendors have in making their proposals, and the conditions under which proposals can be rejected. It is not necessary to send the RFP to all known vendors of the type of equipment being evaluated. If there is any reason why it would be undesirable to purchase from particular vendors, for example, poor reputation in the area or unsatisfactory dealings in the past, it would be best to eliminate them from consideration before the RFP process.

The response of the vendors to the RFP will allow the selection committee to narrow the field to equipment that is likely to meet the defined needs. The full evaluation committee should review the proposals. There will have been a deadline for submission of proposals, but it might not be appropriate to strictly enforce it. It is more important to consider the long-term advantages to the hospital and not to discount an otherwise acceptable proposal for being a few days late. The proposals should be reviewed for completeness. Has all of the requested information been provided? Are there any misinterpretations? Are there exceptions? The vendors should be contacted and additional information requested or clarifications discussed as necessary. It will now also be possible to determine the type of acquisition required, that is, whether the equipment can be purchased from a single vendor, whether it will have to be purchased from more than one vendor and assembled, or whether the equipment will require some special development effort to meet the clinical needs.

Evaluate the quality of each proposal. A simple form can be used to record the results. The form should include the elements of the review. Score the responses according to the relative importance of each element, and whether there was a complete, partial, or no response (Table 2). Include comments to explain the reason for each score. Based upon a review of the proposals submitted, the evaluation

Table 2. Proposal Evaluation Form

Evaluation of Response to Request for Proposal

Date:
Reviewer:

Manufacturer:
Equipment included in proposal (models, options, quantity):
Response submitted on time: Score (0,1):__
Response followed specified format: Score (0,1,2):__
Response included all information requested: Score (0,2,4):__
Response includes exceptions: Score (−2,−1,0):__
Response included additional, useful suggestions: Score (0,1,2):__
Response included reasonable alternatives: Score (0,1,2):__
Total Score:__
Percent Score {[*total* score/*total* possible score (11)] × 100}:_____

committee should agree on which vendors and equipment to consider further (i.e., the proposals that offer equipment that meets the established criteria).

The next step will be to request equipment for an in-house evaluation. The equipment requested for testing should be the exact type that would be ordered. If preproduction prototypes or engineering models are accepted for evaluation, it should be realized that there are likely to be changes in the operation and performance of the final product. The dependability of prototype equipment cannot be judged. In short, the evaluation will not be complete, and, to some extent, the ability to judge the equipment or compare it with other equipment will be hampered.

The most important aspect of the entire equipment acquisition process, especially for equipment types not already in use, is the in-house, comparative evaluation. This evaluation has two phases: engineering and clinical. The equipment requested for comparative evaluation has been selected from the proposals submitted. The field should be narrowed to include only vendors and equipment worthy of serious consideration. Therefore, it will be worthwhile to commit significant effort to this evaluation.

The engineering phase of the evaluation will need test procedures, test equipment, and forms for documenting the results. The clinical phase will need representative areas, users, protocols, a schedule for training as well as use, and a method for collecting results: an interview, a review meeting, or a questionnaire. All of these details should be settled before the arrival of the equipment. For example, it might be determined that sequential testing would be preferable to simultaneous testing. Neither the hospital nor the vendor would want to store equipment while waiting for its evaluation.

The first step in testing equipment is the engineering evaluation, done in the laboratory. The tests include safety and performance aspects. Mechanical safety criteria include consideration of the ruggedness or structural integrity of the equipment as well as the potential for causing injury to patients or personnel. Electrical safety tests are conducted per the requirements in the AAMI/ANSI *Electrical* Safety *Standard* and NFPA electrical standards as appropriate. Performance testing is done as described by any published standards, according to the manufacturer's own service literature, and per the needs determined in the justification process. The results of the engineering tests should be summarized in a chart to facilitate comparison (Table 3). Differences and especially flaws should be highlighted.

Table 3. Engineering Evaluation Form

Engineering Evaluation

Date:
Evaluator:

Manufacturer:
Equipment included in evaluation (models, options):
Safety
  Mechanical: Score (0,1,2):__
  Electrical: Score (0,1,2):__
Safety Subtotal (weight 0.2 × average score):_____

Performance
  Controls: Score (0,1,2):__
  (List performance features; score according to test results; weight according to importance): Score (0,1,2):__
or Score (0,2,4):__
Performance Subtotal (weight 0.2 × average score):_____

Manufacturer's Specifications
  (List important specifications; score according to test results; weight according to importance): Score (0,1,2):__
or Score (0,2,4):__
Manufacturer's Specifications Subtotal (weight 0.1 × average score):_____

Technical Standards
  (List applicable technical standards; score according to test results; weight according to importance): Score (0,1,2):__
or Score (0,2,4):__
Technical Standards Subtotal (weight 0.2 × average score):_____

Human Engineering
  Design: Score (0,1,2):__
  Size: Score (0,1,2):__
  Weight: Score (0,1,2):__
  Ease of use: Score (0,1,2):__
  Reliability: Score (0,1,2):__
  Serviceability: Score (0,1,2):__
  Operator's manual: Score (0,1,2):__
  Service manual: Score (0,1,2):__
Human Engineering Subtotal (weight 0.1 × average score):_____
Total Score:_____
Percent Score [(*total* score/*total* possible score) × 100]:_____

Table 4. User Evaluation Form

User Interview

Date:
Interviewer:

Institution:
Name and title of person(s) interviewed:
Manufacturer:
Equipment at site (models, options):
Years equipment in use:
Safety (any incidents involving patients or personnel)

Score (0,1,2):__
Safety Score (weight 0.2 × score):_____

Performance (does equipment meet user needs)

Score (0,1,2):__
Performance Subtotal (weight 0.2 × average score):_____

Reliability (frequency of equipment failures)

Score (0,1,2):__
Reliability Subtotal (weight 0.1 × average score):_____

Ease of Use (satisfaction with ease of use)

Score (0,1,2):)__
Ease of Use Subtotal (weight 0.1 × average score):_____

Ease of Service (satisfaction with ability to inspect and repair)

Score (0,1,2):__
Ease of Service Subtotal (weight 0.1 × average score):_____

Manufacturer's Support (quality of training and service)

Score (0,1,2):__
Manufacturer's Support Subtotal (weight 0.1 × average score):_____

Overall Satisfaction (would user buy equipment again)

Score (0,1,2):__
Overall Satisfaction Subtotal (weight 0.2 × average score):_____
*Total Score:*_____
Percent Score {[*total* score/*total* possible score (2.0)] × 100}:_____

In addition to the straightforward safety and performance tests, a number of characteristics should be evaluated based upon engineering judgment. The physical construction should be judged, especially if there are constraints imposed by the intended application or installation. A study of the construction and assembly can also allow a judgment regarding reliability. This judgment should be based upon the quality of hardware and components, the method of heat dissipation (fans suggest an exceptional requirement for heat dissipation and require periodic cleaning), the method of construction (the more wiring and connectors, the more likelihood of related failure), and whether the design has had to be modified by such means as alterations on circuit boards or "piggybacked" components.

The maintainability of the equipment is reflected both in the methods of assembly and in the maintenance instructions in the operator and service manuals. The manuals should explain how and how often preventive maintenance or inspection or calibration should be performed. Finally, a clinical engineer should be able to judge human engineering factors. For example, ease of use, how logical and self-explanatory is the front panel, self-test features, and the chances or likelihood of misuse all affect the safety and efficacy of the equipment.

Design a form to record the engineering evaluation results. The form will vary in the specific features and tests that are included, according to the type of equipment that is being evaluated, but the basic format will remain the same. Include comments to explain the reason for the score of each item. Table 3 is an example of an engineering evaluation form.

The physicians, nurses, and clinical engineers on the evaluation committee should contact their counterparts at the institutions named as users of their equipment by the vendors. Site visits can be particularly useful in cases where the equipment or technology being considered is new to the hospital. The committee can see the application firsthand, and talk to the actual users face to face. Record and score the results of interviews (Table 4).

Clinical testing is performed after engineering testing. Equipment must satisfactorily pass the engineering testing to be included in the clinical testing; there is no point in wasting people's time or taking a chance on injuring a patient. Clinical testing should not be taken lightly; it is usually more important than the technical testing. The clinical testing is done only on equipment that has survived all of the previous tests and by the people who will actually be using it.

The clinical testing should be designed so that the equipment is used for the intended application. The equipment included in the clinical testing should be production equipment unless special arrangements are likely to be made with the manufacturer. Users should be trained just as they would be for the actual equipment to be purchased. In fact, the quality of the training should be included in the evaluation. Users should also be given questionnaires or be interviewed after the equipment has been used (Table 5).

Table 5. Clinical Testing Questionnaire

Clinical Trial

Date:
Clinician:

Manufacturer:
Equipment used (models, options):
Number of patients on whom equipment used:
Safety (any incidents involving patients or personnel)

Score (0,1,2):__
Safety Score (weight 0.2 × score):_____

Performance (does equipment meet user needs)

Score (0,1,2):__
Performance Subtotal (weight 0.2 × average score):_____

Reliability (number of equipment failures)

Score (0,1,2):__
Reliability Subtotal (weight 0.1 × average score):_____

Ease of Use (satisfaction with ease of use)

Score (0,1,2):__
Ease of Use Subtotal (weight 0.1 × average score):_____

Manufacturer's Support (quality of training and support)

Score (0,1,2):__
Manufacturer's Support Subtotal (weight 0.1 × average score):_____

Overall Satisfaction (would user recommend equipment)

Score (0,1,2):__
Overall Satisfaction Subtotal (weight 0.2 × average score):__
Total Score:_____
Percent Score {[*total* score/*total* possible score (1.8)] × 100}:_____

One further consideration to be judged by the full evaluation committee is that of standardization. There are numerous valid reasons for standardizing on equipment. Repairs are easier to accomplish because of technicians' familiarity with the equipment. Repair parts are easier and less expensive to keep in inventory. Standardization allows exchange of equipment between clinical areas to help meet varying demands. Training is also more easily provided.

There are also valid drawbacks, however. Standardization makes the hospital dependent upon a limited number of vendors. It can interfere with user acceptance if users feel they have little or no say in the selection process. It can also delay the introduction of new technology. It is important that the evaluation committee consider the relative importance of the pros and cons of standardization in each case.

Assessment of the equipment should result in a ranking of the equipment that has successfully completed the engineering and clinical testing. The advantages and disadvantages of each item should be determined and ranked in order of importance. Alternatively, the criteria can be listed in order of importance. If possible, the list should be divided between criteria that are mandatory and those that are desirable. Then a judgment on whether the equipment does or does not satisfy the criteria can be made. From this charting, it is likely that clear preferences will become obvious.

Ideally, there will be two or more finalists who are close to equal in overall performance at this point. These finalists should be sent a request for quotation (RFQ). The responses to the RFQ will allow a cost comparison. A life-cycle cost analysis can be accomplished with the aid of the hospital's financial officer. This will give a more accurate depiction of the total cost of the equipment for its useful lifetime. While it is beyond the scope of this article to describe life-cycle cost analysis, it takes into account the initial costs (capital cost of equipment plus installation), operating costs over the anticipated life of the equipment (supplies, service, fees), and the time cost of money (or present value). It may or may not take personnel costs into account, since these will likely be the same or similar for different models. To calculate a percent score for purposes of evaluation, divide the cost of the least expensive equipment by the cost of the equipment being evaluated, and multiply by 100.

With the completion of the evaluation, a comparison of the results should lead to a final selection. This should be as objective as possible. If additional information is needed, there should be no hesitation in contacting vendors. In fact, it may be useful to have a final presentation by the vendors. Develop a point system for the individual elements of the proposal reviews, engineering evaluations, user interviews, clinical testing, and cost comparison. Weight these elements according to their relative importance (Table 6).

Once the hospital has made its final choice of vendor and equipment, it is not necessary to end negotiations. Negotiations should be conducted before a vendor knows that they are the preferred provider. Before a vendor is certain of an order, they are much more likely to make concessions. Requests for extra features, such as spare equipment, spare parts, special tools, and test equipment, may be included in the final quote. Trade-in of old equipment and compromises on special features or warranties, for example, can help reduce the cost of equipment. Consider negotiating hardware and software upgrades for a specified period.

For large orders (e.g., dozens of infusion pumps), or for installations (e.g., monitoring systems), request that the vendor unpack, set up, inspect, distribute or install, and document (using the hospital's system) the equipment at no additional cost. In most cases, the hospital will want

Table 6. Evaluation Scoring Form

Overall Evaluation Score

Date:
Reviewer:

Manufacturer:
Equipment (models, options):

Proposal Percent Score:__
Proposal Score (weight 0.1 × score):_____

Engineering Evaluation Percent Score:__
Engineering Evaluation Score (weight 0.2 × score):_____

User Evaluation Percent Score:__
User Evaluation Score (weight 0.1 × score):_____

Clinical Testing Percent Score:__
Clinical Testing Score (weight 0.3 × score):_____

Cost Comparison Percent Score:__
Cost Comparison Score (weight 0.1 × score):_____

Total Score:_____

the vendor to provide user training. This training should meet the hospital's specified needs (e.g., all users, all shifts, train the trainer, videotapes or computer programs, etc.). The hospital may also need service training for its clinical engineering department. The hospital can negotiate not only the service school tuition, but also room, board, and travel. The negotiated quote should be reviewed by the end-users, clinical engineering, finance, purchasing, and a contracts lawyer.

IMPLEMENTATION PROCESS

The successful conclusion of the equipment acquisition process cannot be realized until the equipment is satisfactorily put into use. As with all the previous steps in the equipment acquisition process, the implementation process requires planning and monitoring.

The purchase order with which the equipment is ordered is a legal document: a contract. As such, it can protect the rights of the hospital. Therefore, it is important to include not only the equipment being ordered, but also all agreed-upon terms and conditions. These terms and conditions should include delivery schedules, the work to be performed by the vendor, warranty conditions, service agreements, operator and service manuals, acceptance criteria and testing, and operator and service training.

Before the equipment is delivered, arrangements should be made for installation, for personnel to transport the equipment, for storage, and for personnel to test the equipment. If any of these are to be the responsibility of the vendor, they should be included in the purchase order. If the hospital has to perform any modifications or fabrications, all necessary parts and preparations should be in place.

Acceptance testing should be done upon the completion of installation. The test procedures for acceptance testing should come from the initial specifications and from the manufacturer's data sheets and service manuals. Acceptance testing should be thorough, because this is the ideal time to obtain satisfaction from the manufacturer. It is also appropriate to initiate the documentation per the standard system of the hospital at this time. The invoice for the equipment should not be approved until the equipment has been satisfactorily tested.

After the equipment is in place and working properly, training should be scheduled. The manufacturer or their representative, or the selected hospital personnel, should have the training program completely prepared. If ongoing training will be necessary over the lifetime of the equipment, in-service instructors should be involved in the initial training as well. The clinical engineering personnel responsible for inspection and maintenance should also receive operator and service training.

An often overlooked, but useful, adjunct to the implementation process is follow-up on the installation. Users should be polled for the acceptance of the equipment and their perception of its usefulness. Engineering personnel should review the dependability of the equipment from their service records. All of the people involved in the equipment acquisition process should learn from every acquisition, and what they have learned should be reviewed during this follow-up.

CONCLUSION

Table 1 can be used as a list of tasks that should be accomplished (or at least considered) in the equipment acquisition process. A schedule and checklist can be generated from this list.

The equipment acquisition should be fully documented, ideally by a write-up of the entire process. Table 1 can be used to create an outline for the final report. The results of the equipment acquisition process should be shared with manufacturers and other interested parties. It should always be the intention of the hospital personnel to improve the situation with respect to the manufacturer. Perhaps the most dependable way for medical equipment manufacturers to learn what is important to hospitals is to review what hospitals have said about the evaluation of new equipment during their acquisition process.

BIBLIOGRAPHY

Cited References

1. Comprehensive Accreditation Manual for Hospitals, 2005. Oakbrook Terrace (IL): Joint Commission on Accreditation of Healthcare Organizations; 2004.

Reference List

Health Devices Sourcebook, 2005. Plymouth Meeting (PA): ECRI; 2004.

Larson E, Maciorowski L. Rational Product Evaluation. JONA 16(7, 8):31–36.

Stiefel R, Rizkalla E. The Elements of a Complete Product Evaluation. Biomed Instrum Technol. Nov/Dec 1995. p 482–488.

Stiefel RH. Medical Equipment Management Manual, 2004 Edition. Arlington (VA): Association for the Advancement of Medical Instrumentation; 2004.

Staewen WS. The Clinical Engineer's Role in Selecting Equipment. Med Instrum 18(1):81–82.

See also EQUIPMENT MAINTENANCE, BIOMEDICAL; MEDICAL RECORDS, COMPUTERS IN; OFFICE AUTOMATION SYSTEMS; RADIOLOGY INFORMATION SYSTEMS.

EQUIPMENT MAINTENANCE, BIOMEDICAL

ARIF SUBHAN
Masterplan Technology Management
Chatsworth, California

INTRODUCTION

Preventive maintenance (PM) is one of the many functions of a clinical engineering department. The other functions include incoming inspection/testing, prepurchase evaluation, coordination of outside equipment service, hazard and recall notification, equipment installation, equipment repair and upgrade, purchase request review, equipment replacement planning, device incident review, and regulatory compliance maintainance. The primary objective of a PM program for biomedical equipment is to prevent failure, which is achieved through (*1*) detecting the degradation of any non-durable parts of the device and restoring them to like-new condition; (*2*) identifying any significant degradation of the performance of the device and restoring it to its proper functional level; and (*3*) detecting and repairing any partial degradation that might create a direct threat to the safety of the patient or operator.

The term preventive maintenance has its origins with mechanical equipment that has parts that are subject to wear and need to be restored or replaced sometime during the useful lifetime of the device. PM refers to the work performed on equipment on a periodic basis and should be distinguished from "repair" work that is performed in response to a complaint from the equipment user that the device has failed completely or is not working properly. The term "corrective maintenance" is often used instead of the term "repair."

Today, medical equipment uses electronic components that fail in an unpredictable manner, and their failure cannot be anticipated through measurements and checks performed during a PM (1). Also, equipment failures that are attributable to incorrect set up or improper use of the device or the use of wrong or defective disposable accessory cannot be prevented by doing PM (2). It has been argued that current medical devices are virtually error-free in terms of engineering performance. Device reliability is an intrinsic function of the design of the device and cannot be improved by PM. In modern equipment, the need for performance and safety testing is greatly reduced because of the design and self-testing capability of the equipment (3,4). For example, the Philips HeartStart FR2+ defibrillator performs many maintenance activities itself. These activities include daily and weekly self-tests to verify readiness for use and more elaborate monthly self-tests that verify the shock waveform delivery system, battery capacity, and internal circuitry. The manufacturer also states that the FR2+ requires no calibration or verification of energy delivery. If the unit detects a problem during one of the periodic self-tests, the unit beeps and displays a flashing red or a solid red warning signal on the status indicator (5).

The term "scheduled (planned) maintenance" was introduced in 1999 by a joint AAMI/Industry Task Force on Servicing and Remarketing while developing a document for submission to the FDA entitled "Joint Medical Device Industry Proposed Alternative to the Regulation of Servicers, Refurbishers, and Remarketers" (6). However, many still use the traditional term preventive maintenance rather than this new, more carefully defined term. According to the document developed by the AAMI/Industry Task Force, the term scheduled (planned) maintenance "consists of some or all of the following activities: cleaning; decontamination; preventive maintenance; calibration; performance verification; and safety testing." Among these activities, the three key activities are (*1*) preventive maintenance (PM); (*2*) performance verification (PV) or calibration; and (*3*) safety testing (ST).

These three terms are defined by the AAMI/Industry Task Force as follows. "Preventive maintenance is the inspection, cleaning, lubricating, adjustment or replacement of a device's nondurable parts. Nondurable parts are those components of the device that have been identified either by the device manufacturer or by general industry experience as needing periodic attention, or being subject to functional deterioration and having a useful lifetime less than that of the complete device. Examples include filters, batteries, cables, bearings, gaskets and flexible tubing." PM performed on a medical device is similar to the oil, filter, and spark plug changes for automobiles.

"Performance Verification is testing conducted to verify that the device functions properly and meets the performance specifications; such testing is normally conducted during the device's initial acceptance testing." This testing is important for detecting performance deterioration that could cause a patient injury. For example, if the output of the device (such as temperature, volume, or some form of energy) is not within specifications, it could result in an adverse patient outcome (7). If the performance deterioration can be detected by visual inspection or by a simple user test, then periodic performance verification becomes less critical. The data obtained during maintenance will also assist in determining the level and intensity of performance verification.

"Safety testing is testing conducted to verify that the device meets the safety specifications, such testing is normally conducted during the device's initial acceptance testing."

HISTORICAL BACKGROUND

The concept of a facility-wide medical equipment management program first emerged in the early 1970s. At that time, there was little management of a facility's medical equipment on a centralized basis. Examples of the lack of proper management that affected the quality of care are described in detail in the literature (8) and include poor frequency response of ECG monitors, heavy accumulation of dirt inside equipment, delay in equipment repair, little attention to the replacement of parts that wear over time, and the consequential high cost of repairs.

The Joint Commission on Accreditation of Healthcare Organizations (JCAHO) has played an important role in

promoting equipment maintenance programs in hospitals in the United States. Almost all hospitals in the United States are accredited by the JCAHO and are required to comply with their standards.

Poor or nonexistent equipment maintenance programs (8) coupled with the "great electrical safety scare" of the early 1970s (9) created the impetus that has popularized electrical safety programs in hospitals. Since then, equipment maintenance and electrical safety testing has been an important part of clinical and biomedical engineering programs in hospitals. The JCAHO standards in the mid-1970s required that all electrically powered equipment be tested for leakage current four times a year (10).

In 1983, the Medicare reimbursement program for hospitals changed to a much tighter fixed-cost approach based on so-called Diagnostic Related Groups (DRGs). In 1979, JCAHO increased the maximum testing interval for patient care equipment to six months. This change was made in response to pressure from hospitals to reduce the cost of complying with their standards. It has been reported that there was no other rationale for changing the testing intervals (10).

In 2001, JCAHO removed the annual PM requirement for medical equipment (11). The reason cited for the change was that the safety and reliability of medical equipment had improved significantly during the prior decade. It was also stated that some equipment could benefit from maintenance strategies other than the traditional "interval-based" PM. Other PM strategies suggested included predictive maintenance, metered maintenance (e.g., hours of usage for ventilators), and data-driven PM intervals.

INVENTORY

A critical component of an effective PM program is an accurate inventory, which has been a key requirement in the JCAHO equipment standards since they were introduced in the early 1970s. The 2005 JCAHO standard (EC.6.20) requires a current, accurate, and separate inventory of medical equipment regardless of ownership (12). The inventory should include all medical equipment used in the hospital (owned, leased, rented, physician-owned, patient-owned, etc.). A complete and accurate inventory is also helpful for other equipment management functions including tracking of manufacturer's recalls, documenting the cost of maintenance, replacement planning, and tracking model-specific and device-specific issues. The equipment list should not be limited to those devices that are included in the PM program (13).

INCLUSION CRITERIA

In 1989, JCAHO recognized that not all medical devices are equally critical with respect to patient safety and introduced the concept of risk-based inclusion criteria for the equipment management program. Fennigkoh and Smith developed a method for implementing a risk-based equipment management program by attributing numerical values to three variables; "equipment function," "physical risks associated with clinical application," and "maintenance requirements" (14). They compounded these values into a single variable; the equipment management (EM) number. The EM was calculated as follows: EM = Function rating + Risk rating + Required Maintenance rating. The three variables were not weighted equally. Equipment function constituted 50% of the EM number whereas risk and maintenance each constituted 25%. The authors acknowledge the somewhat arbitrary nature of weighting the equipment function rating at 50% of the EM number; however, they viewed this variable as the device's most significant attribute. The authors also arbitrarily set a level for the EM number at greater than or equal to 12 to determine whether the device will be included in the EM program. Devices included in the EM program are assigned a unique control number. Devices with an EM number less than 12 were excluded from the EM program and are not assigned a control number.

Although risk-based calculators using the original Fennigkoh and Smith model are still widely used, criticism of this method exists on the basis that this particular risk-based approach is arbitrary. The factors used in the calculation, their weighting, the calculated value above which one decides to include the device in the management program, and what PM interval to use are criticized as all being somewhat subjective or arbitrary (13).

SELECTING THE PM INTERVAL

Selecting an appropriate PM interval is a very important element of an effective maintenance program. Several organizations have offered guidelines, requirements, and standards for the selection of an appropriate PM interval for the various devices. This list includes accreditation organizations (e.g., JCAHO), state authorities (e.g., California Code of Regulations, Title 22), device manufacturers, and national safety organizations (e.g., NFPA). JCAHO is the primary authority that most people look to for minimum requirements for the PM intervals for medical devices. However, the JCAHO 2005 Standards [EC.6.10 (4) and EC.6.20 (5)] put responsibility back on the hospital to define PM intervals for devices based on manufacturer recommendations, risk levels, and hospital experience. JCAHO also requires that an appropriate maintenance strategy be selected for all of the devices in the inventory (12).

According to one source (7) two approaches exist to determining the proper PM interval. One is fixed and the other is evidence-based. Other sources present some contradictory views on how to select a PM interval for a device. Some (13,15) argue that PMs should not necessarily follow the manufacturer's recommended intervals but should be adjusted according to the actual PM failure rate. Ridgway and Dinsmore (16,17) argue that if a device is involved in an incident where a patient or staff member is injured, a maintainer who has not followed the manufacturer's recommendations will be deemed to have some legal liability for the injury, irrespective of how the device contributed to the injury. Dinsmore further argues that the manufacturer recommendations should be followed

Table 1. Conflicting Recommendations for the PM Intervals of Some Devices

Device Type	ASHE PM Interval (months)	ECRI PM Interval (months)
Defibrillator	3	6
Apnea monitor	6	12
Pulse oximeter	6	12

because the manufacturer developed the device and proved its safety to the FDA. He states that if no recommendations exist from the manufacturer, then it is up to the clinical engineering department to decide on the maintenance regimen. National organizations such as the American Society of Hospital Engineering (18) and ECRI (2) have published recommendations on PM intervals for a wide range of devices. The recommendations by ASHE were published in 1996 and ECRI in 1995. However, conflicting recommendations exist for certain devices in the list (see Table 1).

The ANSI/AAMI standard EQ56 (19) "Recommended Practice for a Medical Equipment Management Program" does not attempt to make specific recommendations for PM intervals for devices. And, in many cases, PM interval recommendations from the manufacturers are not readily available. Those that have been made available are often vague with little or no rationale provided. In some cases, their recommendations simply state that testing should be conducted periodically "per hospital procedures" or "per JCAHO requirements."

Intervals for the electrical safety testing of patient care equipment are discussed in the latest edition of NFPA 99-2005, Health Care Facilities, Section 8.5.2.1.2.2. Although this document is a voluntary consensus standard and not a regulation, many private and government agencies reference and enforce NFPA standards. The following electrical safety testing intervals are recommended. The actual interval is determined by the device's normal location or area in which the device is used. For general care areas, the recommended interval is 12 months; for critical care areas and wet locations, the recommended interval is 6 months. The standard does allow facilities to use either longer or shorter intervals if they have a documented justification from previous safety testing records or evidence of unusually light or heavy use (20).

In a similar fashion, the evidence-based method allows adjustment of the interval up or down depending on the documented finding of prior testing (7). This method is based on the concepts of reliability-centered maintenance (21). Ridgway proposes an evidence-based method using a PM optimization program that is based on periodic analyses of the results of the PM inspections. His approach takes into consideration the severity of the problems found during the PMs. It classifies the problems found into one of the four PM problem severity levels, level 1 through 4. This method incorporates the concepts found in Failure Modes and Effects Analysis (FMEA)—a discipline that has been recently embraced by the JCAHO. It requires the classification of the problems discovered during the PM testing into four levels of criticality. Level 1 problems are potentially life-threatening, whereas the other levels are progressively less severe. Examples of level 1 problems include defibrillator output energy being significantly out of specification or an infusion pump significantly under- or over-infusing. For a more detailed explanation of this method, see Ref. (22). Others (23,24) have used maintenance data in a similar way to rationalize longer maintenance intervals.

It should be noted that some regulations, codes, or guidelines and some manufacturer's recommendations may advance more stringent requirements than other standards for the same type of device or an interval that the facility has used successfully in the past. In these instances, the facility needs to balance the risks of noncompliance, which may include injury to the patient, liability, and penalties for the organization, against the cost of compliance. Many factors exist that need to be evaluated before modifying the generally recommended intervals, including equipment maintenance history and experience, component wear, manufacturer recommendations, device condition, and the level of technology used in the device (2).

Note also that the manufacturer's recommendations for their newer models of medical devices are generally less stringent than those for their older models. In many instances, the recommended PM intervals are greater than 12 months.

PM PROCEDURES

Selecting an appropriate PM procedure is another important element of an effective maintenance program. Both ECRI (2) and ASHE (18) have published PM procedures for a broad range of devices. It should be noted that the manufacturers usually provide more detailed PM procedures for their specific devices than those published by ECRI and ASHE. It is recommended that manufacturers' recommendations that appear to be too complex be evaluated to see if they can be simplified (2). See Fig. 1 for a sample PM procedure for an infusion pump.

EVALUATION OF A PM PROGRAM

The facility's PM completion or completed on-time rates are parameters that have been favored by both facility managers and external accreditation agencies (7). However, the value of PM completion rate as an indicator of program quality has been debated for many years (25). Until 2003, to pass the maintenance portion of the medical equipment program, the JCAHO required a consistent 95% PM completion rate. The shortcoming of this standard is that the 5% incomplete PMs could, and sometimes did, include critical devices such as life-support equipment. To address this undesirable loophole, the JCAHO, in 2004, discontinued the 95% requirement. The 2005 standard EC.6.20 now segregates medical equipment into two categories: life-support and nonlife-support devices. Life-support equipment is defined by JCAHO as those devices that are intended to sustain life and whose failure to perform their primary function is expected to result in death. Examples include ventilators, anesthesia

For: **INFUSION PUMP** PM Procedure No. **IP001**

Interval: **xx months** Estimated annual m-hrs: **x.0**

Check the **AC** box if **Action Completed** and the **AMR** box if **Adjustment or Minor Repair** was required to bring the device into conformance with the performance or safety specifications. In this case provide a **note*** on the nature of the problem found, in sufficient detail to identify the ***level of potential severity*** of the problem.

AC/AMR **Preventive Maintenance**

SM1. Inspect/clean the chassis/housing especially any moving parts including any user-accessible areas under covers (if applicable). Examine all cable connections. Replace any damaged parts, as required.

SM2. Confirm that all markings and labeling are legible. Clean or replace, as required.

SM3. Replace or recondition the battery, as required.

AC/AMR **Performance Verification**

PV1. Verify that the flow rate or drip rate is within specification. Perform self-test, if applicable.

PV2. Verify that all alarms (including occlusion and maximum pressure) and interlocks operate correctly.

PV3. Verify that the battery charging system is operating within specification.

PV4. Verify the functional performance of all controls, switches, latches, clamps, soft touch keys, etc.

PV5. Verify the functional performance of all indicators and displays, in all modes.

PV6. Verify that the time/date indication is correct, if applicable.

AC/AMR **Safety Testing**

ST1. Check that the physical condition of the power cord and plug, including the strain relief, is OK.

ST2. Check the ground wire resistance. (< 0.5 ohm).

ST3. Check chassis leakage to ground. (< 300 micro amps).

*** Notes**:

Figure 1. Sample PM procedure for an infusion pump.

machines, and heart–lung bypass machines (12). ECRI suggests that the following additional devices be included in the life-support category: anesthesia ventilators, external pacemakers, intra-aortic balloon pumps, and ventricular assist devices (26). The completion of PMs for life-support equipment is scored more stringently than the completion rate performance for nonlife-support equipment. It is theoretically possible that if the PM of a single life-support device is missed, an adverse, noncompliance "finding" could be generated by the survey team. However, it has been reported that the surveyors will probably be more focused on investigating gaps in the process that led to the missed PM (27).

The 2005 JCAHO standards do not contain an explicit PM completion requirement. However, many organizations appear to have set the goal for on-time PM completion rate for life-support equipment at 100% and the goal for the nonlife-support equipment at better than 90%. An important aspect of PM completion is calculating the on-time PM completion rate. The JCAHO does not state how the PM completion rate should be calculated. Hospitals are free to specify in their management plan that they have allowed themselves either 1 or 2 months of extra time to complete the scheduled maintenance. It is important for those devices that have maintenance intervals of three months or less, or are identified as life-support equipment, that the PMs be completed within the month they are scheduled for PM. Sometimes, additional time may be needed for devices that are unavailable because they are continuously in use. In this case, the appropriate department and the safety/environment of care committee should be informed about the delay.

The real issue to be addressed is the effectiveness of the program. However, "effectiveness" of a PM program is difficult to measure. If the purpose of the restoration of any nondurable parts is to reduce device failures, then the effectiveness of this element can be measured by the resulting reduction in the device failure rate. In principle, repair rates and incident analyses can provide a relationship between PM and equipment failures (7).

Two challenges associated with achieving an acceptable PM completion rate exist. The first is to find the missing equipment listed on the monthly PM schedule. Multiple documented attempts should be made to locate any devices that are not found, which should be followed with a written communication to the clinical users seeking their assistance in locating the device. It is important to get a written acknowledgment from the users that they have attempted to locate the devices in question. This acknowledgment will help when explaining to the surveyors that efforts were made to locate the device. However, if the devices cannot be located after this extended search it must be assumed that they are no longer in use in the facility. The question now is, are they still on site but deliberately or accidentally hidden away? Or have they really been removed from the facility? Until these questions can be answered, no satisfactory way to deal with accounting for the failure to complete the overdue PM exists. One strategy to reduce the potential administrative paperwork is to classify the missing equipment as "unable to locate." This action (classifying the devices as unable to locate) should be communicated to the appropriate departments including the user departments, the safety/environment of care committee, materials management, and security. Consistent loss or unable to locate device(s) should be viewed negatively by the facility and should serve as an indicator that the facility needs to re-evaluate its security plan. The users should be reminded not to use the missing device(s) if they reappear. These device(s) should be readily identifiable with the "out-of-date" PM sticker. The service provider should be notified so that they can give the re-emerging device(s) an incoming inspection and restart its PM sequence.

The second challenge is to gain access to equipment that is continually in use for patient care. It is difficult to complete the PM if the device is in constant use for monitoring or treatment of patients. Examples of devices that often fall into this category include ventilators, patient monitors, and certain types of laboratory or imaging equipment. Ideally, to alleviate this problem, a back-up unit should be available to substitute while maintenance is performed on the primary unit. This spare unit can also serve as a back-up for emergency failures as well. A good working relationship with the users is very helpful in these circumstances.

DOCUMENTATION

Documentation of maintenance work is another important element of an effective maintenance program. The 2005 JCAHO standards require the hospital to document performance, safety testing, and maintenance of medical equipment (12).

Complete scheduled maintenance and repair documentation is required when a device is involved in an incident, or when the reliability of the device is questioned, and also to determine the cost of maintenance. Most accrediting and regulatory organizations accept a system of exception reporting, which is recording only the results of steps failed during performance verification as part of scheduled maintenance or after repair (18). It has been generally accepted that it is reasonable to record only what went wrong during the PM procedure (13). Having extensive documentation is considered a safe practice. ECRI supports the use of exception reporting and recommends that before deciding on using exception reporting the hospital take into account that exception reporting, with its lack of affirmative detail, may be less than convincing evidence in the event of a liability suit (2).

Two ways to document scheduled (planned) maintenance exist. One is to record the maintenance by hand on a standard form or preprinted PM procedure and file the records manually. The other is to use a computerized maintenance management system (CMMS). Computerized records should help in reducing the time and space required for maintaining manual documentation. Even with CMMS, the department may need to have a way to store or transpose manual service reports from other service providers. Other technologies like laptops and personal data assistants (PDAs) can further assist in implementing the maintenance program. A comprehensive review of CMMSs, that are currently in use can be found in

the literature (28). The CMMS helps the clinical engineering staff to manage the medical equipment program effectively. The core of the CMMS consists of equipment inventory, repair and maintenance record, work order subsystem, parts management subsystem, reporting capabilities, and utilities. A CMMS can be classified broadly as internally developed (typically using commercial off-the-shelf personal computer hardware and database software) and commercially available applications (desktop and web-based) (29).

STAFFING REQUIREMENTS

The number of full-time equivalents (FTEs) required to do scheduled maintenance varies based on the experience of the staff, the inventory mix, and the inventory inclusion criteria. See Ref. 30 for the method to determine the number of FTEs required for maintenance. Based on a clinical engineering practice survey, an average of one FTE is required to support 590 medical devices (31). The generally cited relationship between the support required for scheduled maintenance and repair for a typical inventory mix of biomedical devices is 750–1250 devices per FTE. Cost of the maintenance program includes salaries, benefits, overtime and on-call pay, cost of test equipment and tools, and training and education expenses. Salaries for the clinical engineering staff can be obtained from the annual survey published by the Journal of Clinical Engineering and the 24×7 magazine.

PM STICKERS

PM stickers or tags are placed on a device to indicate that maintenance has been completed and the device has been found safe to use. They also convey to the user a warning not to use the device when the sticker is out-of-date. Color-coded stickers (see Fig. 2) or tags are not required by the JCAHO, but they can be very helpful in identifying devices that need maintenance (2,32).

ENVIRONMENTAL ROUNDS

Conducting environmental rounds is another important aspect of the medical equipment maintenance program. The intent of these rounds is to discover situations, which are, or could lead to, an equipment-related hazard or safety problem. The clinical engineering staff should conduct regular inspections of all clinical areas that are considered to be "medical equipment-intensive" areas (e.g., ICU, CCU, NICU, ER, and surgery). The interval between these equipment-related environmental safety rounds should be adjusted according to the frequency with which hazards are found. Other areas that are considered less equipment-intensive such as medical/surgical and other patient care floors, laboratory areas, and outpatient clinics should also be surveyed for electrical safety hazards but less frequently. The equipment items in these areas are usually line-powered items that do not have nondurable parts requiring periodic attention and they usually do not require periodic calibration or performance checking. These items need periodic visual inspections to ensure that they are still electrically safe (i.e., that no obvious defects exist such as a damaged power cord or evidence of physical damage that might electrify the case or controls). The search should be focused on the physical integrity of the equipment, particularly the power cord and the associated connectors and other equipment defects such as dim displays, torn membrane switches, taped lead wires, and so on. Items with power cords that are exposed to possible abuse from foot traffic or the wheels of other equipment should receive closer attention than those items with less exposed cords. Damaged wall outlets should be noted, as should the use of extension cords or similar "jury-rigged" arrangements. Other indicators that should be noted and investigated further are equipment enclosures (cases) that are obviously distorted or damaged from being dropped. In addition to these potential electrical safety hazards, other targets of this survey are devices with past-due PM stickers or with no sticker at all. When a safety hazard is identified, or a past due or unstickered item is found, appropriate corrective actions should be taken immediately and documented as required (33).

Figure 2. Color coded sticker.

BIBLIOGRAPHY

Cited References

1. Wang B, Levenson A. Are you ready? 24×7 2004; Jan: 41.
2. ECRI. Health Devices Inspection and Preventive Maintenance System, 3rd ed. Plymouth Meeting (PA): ECRI; 1995.
3. Keil OR. Evolution of the clinical engineer. Biomed Technol Manag 1994; July–Aug: 34.
4. Keil OR. The buggy whip of PM. Biomed Technol Manag 1995; Mar–Apr: 38.
5. Philips. HeartStart FR2+ defibrillator. Instructions for use. M3860A, M3861A, Edition 8, 4-1, 4-9. Seattle (WA): Philips; 2002.
6. Hatem MB. From regulation to registration. Biomed Instrument Technol 1999;33:393–398.
7. Hyman WA. The theory and practice of preventive maintenance. J Clin Eng 2003;28:31–36.
8. JCAHO. Chapter 1. The Case for Clinical Engineering. The Development of Clinical Engineering. Plant, Technology & Safety Management Series Update Number 3. Chicago (IL): JCAHO; 1986. pp 9–11.
9. Nader R. Ralph Nader's most shocking expose. Ladies Home J 1971;88: 98 176–178.
10. Keil OR. Is preventive maintenance still a core element of clinical engineering? Biomed Instrument Technol 1997;31: 408–409.

11. JCAHO. EC Revisions approved, annual equipment PM dropped. Environ Care News 2001;4:1, 3, 9.
12. JCAHO. Hospital Accreditation Standards. Oakbrook Terrace, IL: Joint Commission Resources; 2005.
13. Stiefel RO. Developing an effective inspection and preventive maintenance program. Biomed Instrument Technol 2002;36: 405–408.
14. Fennigkoh L, Smith B. Clinical equipment management. Plant, Technology & Safety Management Series Number 2, Chicago (IL): JCAHO; 1989.
15. Maxwell J. Prioritizing verification checks and preventive maintenance. Biomed Instrument Technol 2005;39:275–277.
16. Ridgway MG. Personal communication. 2005.
17. Baker T. Journal of Clinical Engineering Roundtable: Debating the medical device preventive maintenance dilemma and what is the safest and most cost-effective remedy. J Clin Eng 2003;28:183–190.
18. ASHE. Maintenance Management for Medical Equipment. Chicago, IL: American Hospital Association; 1996.
19. AAMI. Recommended Practice for Medical Equipment Management Program. ANSI/AAMI EQ 56:1999. Arlington, VA: AAMI; 1999.
20. NFPA. NFPA 99, Standard for Health Care Facilities. Quincy (MA): NFPA; 2005.
21. Moubray J. Reliability-Centered Maintenance. New York: Industrial Press; 1997.
22. Ridgway MG. Analyzing planned maintenance (PM) inspection data by failure mode and effect analysis methodology. Biomed Instrument Technol 2003;37:167–179.
23. Acosta J. Data-driven PM Intervals. Biomed Instrument Technol 2000;34:439–441.
24. JCAHO. Data-driven medical equipment maintenance. Environ Care News 2000;3:6–7.
25. Ridgway MG. Making peace with the PM police. Healthcare Technol Manag 1997; Apr: 44.
26. ECRI. Maintaining life support and non-life support equipment. Health Devices 2004;33(7):244–250.
27. AAMI. JCAHO connection. Questions on life support equipment and NPSG # 6. Biomed Instrument Technol 2005;39: 284.
28. Cohen T. Computerized Maintenance Management Systems for Clinical Engineering. Arlington (VA): AAMI; 2003.
29. Cohen T, Cram N. Chapter 36. Computerized maintenance management systems. In: Dyro JF, editor. Clinical Engineering Handbook Burlington (MA): Elsevier Academic Press; 2004;.
30. AHA. Maintenance Management for Medical Equipment. Chicago, IL: American Hospital Association; 1988.
31. Glouhove M, Kolitsi Z, Pallikarakis N. International survey on the practice of clinical engineering: Mission, structure, personnel, and resources. J Clin Eng 2000;25:269–276.
32. Guerrant S. A sticker is worth a thousand words. 24 × 7 44, March 2003.
33. Ridgway MG. Masterplan's Medical Equipment Management Program. Chatsworth (CA): Masterplan; 2005.

See also CODES AND REGULATIONS: MEDICAL DEVICES; SAFETY PROGRAM, HOSPITAL.

ERG. See ELECTRORETINOGRAPHY.

ERGONOMICS. See HUMAN FACTORS IN MEDICAL DEVICES.

ESOPHAGEAL MANOMETRY

VIC VELANOVICH
Henry Ford Hospital
Detroit, Michigan

INTRODUCTION

The purpose of the esophagus is to act as a conduit for food from the mouth to the stomach. This is an active process that is initially a conscious act with chewing and swallowing, then becomes unconscious when the bolus of food enters the esophagus. As part of this process, the esophageal musculature acts to propagate the food bolus to the stomach and to prevent reflux of gastric contents to the esophagus to protect the esophagus itself. There are several disease processes of the esophagus for which the assessment the esophageal musculature function would contribute to the diagnosis and management. This assessment is done indirectly through the measurement of intraesophageal pressures. This pressure measurement is accomplished with esophageal manometry (1).

ESOPHAGEAL ANATOMY AND PHYSIOLOGY

The esophagus is, in essence, a muscular tube that connects the mouth to the stomach. The esophagus anatomically starts in the neck, traverses the thorax, and enters into the abdomen to join the stomach. It is ~20–25 cm in length, with the start of the esophagus being the upper esophageal sphincter and the end being the gastroesophageal junction. The upper esophageal sphincter is primarily made up of the cricopharyngeus muscle attached to the cricoid cartilage anteriorly. The musculature of the upper esophagus is made of striated muscle, and this transitions to smooth muscle in the lower esophagus. The esophagus is made up of three primarily layers: the inner-mucosa, the middle-submucosa, and the outer-muscle layer. The muscle layer is made up of an inner-circular muscle layer and an outer-longitudinal layer. At the inferior end of the esophagus is the lower esophageal sphincter. This is not a true anatomical sphincter, but rather a physiological high pressure zone that is the cumulation of thickening of the distal 5 cm of the esophageal musculature, the interaction of the esophagus with the diaphragmatic hiatus, and at 2 cm of intraabdominal esophagus. The lower esophageal sphincter should not be confused with the gastroesophageal junction, which is the entrance of the esophagus to the stomach, nor with the Z line, which is the transition of the esophageal squamous mucosa to the glandular gastric mucosa.

The lower esophageal sphincter is tonically contracted during rest to prevent reflux of gastric contents into the esophagus. It relaxes with a swallow. Swallowing is divided into an oral stage, pharyngeal stage, and esophageal stage. The oral and pharyngeal stage prepare food by chewing, the tongue pushing the food bolus to the pharynx, the soft palate is pulled upward, the vocal cords are closed and the epiglottis covers the larynx, the upper esophageal sphincter relaxes, and the contraction of the pharyngeal

muscles initiate the primary peristaltic wave. The esophageal stage consists of a peristaltic wave that pushes the food bolus to the stomach. There are primary peristaltic waves, which are initiated by the pharynx with a swallow. Secondary peristaltic waves are initiated within the esophagus due to distention of the esophagus with food. As the bolus reaches the lower esophagus, the lower esophageal sphincter relaxes.

INDICATIONS FOR ESOPHAGEAL MANOMETRY

Indications for esophageal manometry include the following five problems. (*1*) Dysphagia, to assess for an esophageal motility disorder, such as achalasia or nutcracker esophagus. It is important that mechanical causes of dysphagia, such as cancer, have been excluded. (*2*) Gastroesophageal reflux disease, not for primary diagnosis, but for possible surgical planning. (*3*) Noncardiac chest pain, which may be of esophageal origin. (*4*) Exclusion of generalized gastrointestinal disease (e.g., scleroderma or chronic idiopathic pseudoobstruction), and exclusion of an esophageal etiology for anorexia nervosa. (*5*) Determination of lower esophageal sphincter location for proper placement of an esophageal pH probe (2).

EQUIPMENT

The basic equipment used for manometry is the manometry catheter, infusion system (for water perfused systems), transducers, Polygraf or A/D converter, and computer with appropriate software (Fig. 1).

Esophageal manometry catheters come in two basic types: water perfused and solid state. The water perfused catheter contains several hollow tubes. Each tube has one side opening at a specific site on the catheter. The openings are at various points around the circumference of the catheter. There is a 4 cm catheter with side holes placed 5 cm apart, and a 8 cm catheter with 4 lumens placed 1 cm apart in the most distal end of the catheter, then 5 cm apart for the next proximal 4 lumens. The radial spacing helps to accurately measure upper and lower esophageal sphincter pressures that are asymmetrical. The water perfused catheters require an infusion system. The manometric pump uses regulated compressed nitrogen gas to deliver distilled water through the channels of the catheter. The pressurized water from each channel is connected to a single opening in the catheter within the patient's esophagus. The pressure changes are transmitted to the transducer, and these pressure changes are recorded and charted by a computer with the appropriate software.

The solid-state catheter has internal microtransducers. These directly measure intraesophageal pressures and contractions. Some advantages are that the sensors respond faster, do not require that the patient lie down, and record pressures circumferentially. An additional advantage is that as these catheters do not require fluid containers, potential pitfalls with fluid disinfection are avoided. Some drawbacks are that they are delicate, more expensive to purchase and repair, and are prone to baseline drift. As with the water-perfused system, the solid-state system requires that the catheter be connected to a computer console for recording and storing pressure data.

Figure 1. The esophageal manometry console with computer. (Courtesy of Medtronic Corp., Minneapolis, MN.)

Additional items are needed. These include a manometer calibration tube for calibrating the solid-state catheter, a viscous lidocaine, a lubricating jelly, tissues, tape, an emesis basin, a syringe, a cup of water, a penlight, and tongue blades.

CONDUCT OF THE TESTING

Prior to the procedure, the pressure channels require calibration and connection to the computer console. The patient sits upright and one of the nares is anesthetized. Lubricating jelly is applied to the catheter and inserted through one the nares into the pharynx. The patient is asked to swallow and the catheter advance into the esophagus to the stomach. The catheter is advanced for $\sim$65 cm to insure all the sensor ports are within the stomach. If using water-perfused catheters, the patient is moved to the supine position. If using solid-state catheters, the patient is kept in the semi-Fowler position, which is the head and torso of the patient at a 45° angle bent at the waist.

The esophageal motility study consists of (*1*) the lower esophageal sphincter study, (*2*) esophageal body study, and (*3*) the upper esophageal sphincter study. Before beginning the study, insure that all sensor ports are within the stomach by having the patient take a deep breath and

watching the motility recording. It should show a smooth tracing with a pressure increase during inspiration. An alternative method of determining that all side holes are within the stomach is to apply gentle pressure to the epigastrium to confirm that there is a simultaneous increase in the recorded pressure. When all channels are within the stomach, a gastric baseline is set to establish a reference for pressure changes.

The lower esophageal sphincter study measures sphincter pressure, length, location, and relaxation. This is done using either the "station pull-through" or "slow continuous pull through" methods. With the station pull-through technique, the catheter is slowly withdrawn through the sphincter at 1 cm increments while looking for changes in pressure. When the first channel enters the sphincter, pressure will increase from baseline by at least 2 mmHg (0.266 kPa). This identifies the lower border of the sphincter. As the catheter is continued to be pulled back, the "respiratory inversion point" is reached. This is the transition from the intraabdominal to the intrathoracic esophagus. With inspiration, the catheter will record positive pressures within the abdomen, but negative pressures within the thorax. Inferior to the respiratory inversion point is the lower esophageal sphincter high pressure zone. This is the physiologic landmark used to perform pressure measurements and relaxation studies. When the distal channel passes through the upper border of the lower esophageal sphincter, the pressure should decrease from baseline. The length traveled from the lower to the upper

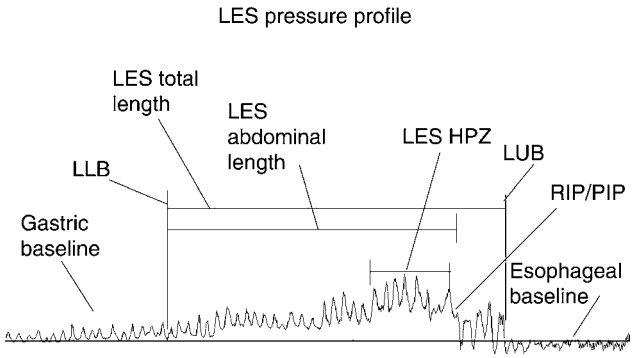


Figure 2. Lower esophageal sphincter profile as determined by esophageal manometry. (Courtesy of Medtronic Corp., Minneapolis, MN.)

border of the sphincter measures the sphincter length. The slow continuous pull-through method is done while the catheter is pulled back continuously, while pressures are being recorded. The catheter is pulled back 1 cm every 10 s. These methods lead to identifying the distal and proximal borders of the sphincter, overall sphincter length, abdom inal sphincter length, and resting sphincter pressure (Fig. 2). Sphincter relaxation involves observing the response of the lower esophageal sphincter to swallowing (Fig. 3). This is done by asking the patient to swallow 5 mL of water with the catheter positioned in the high

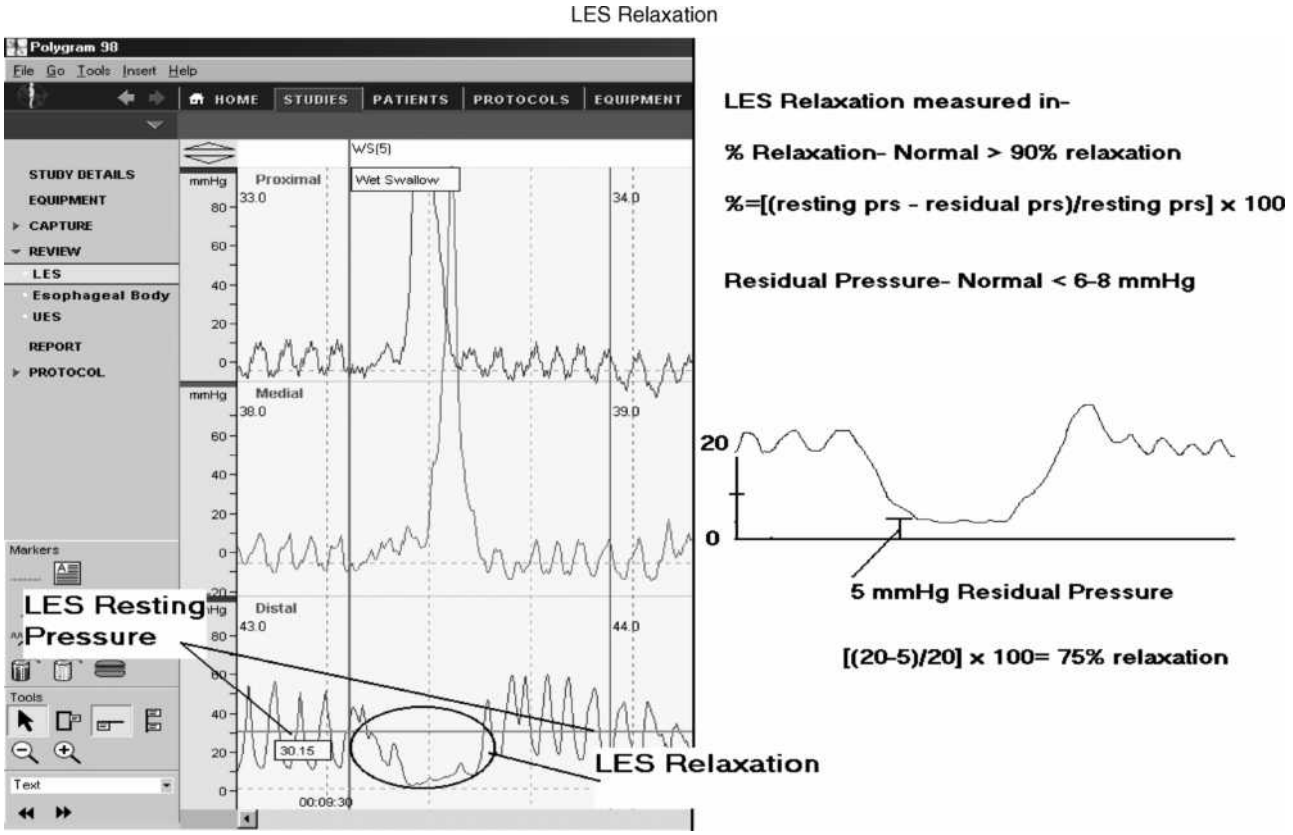


Figure 3. Manometric tracings of lower esophageal sphincter relaxation. (Courtesy of Medtronic Corp., Minneapolis, MN.)

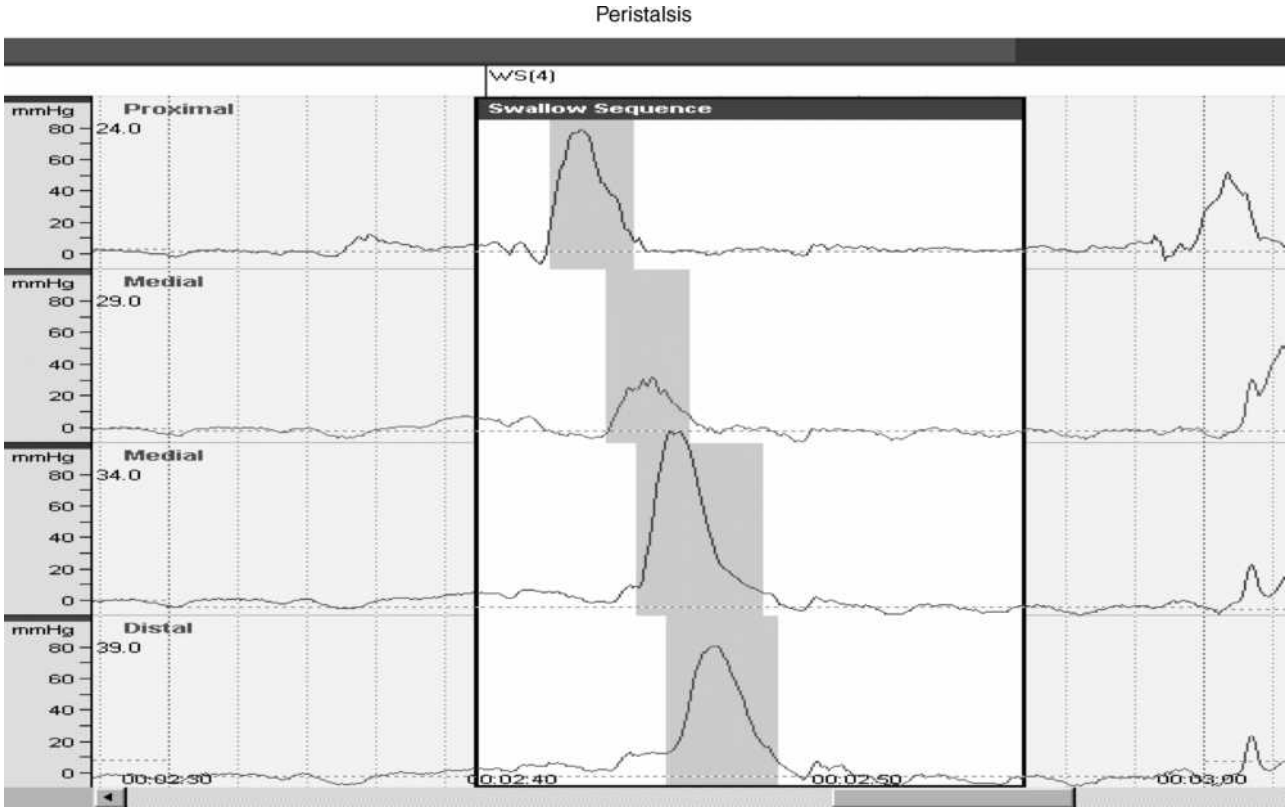


Figure 4. Manometric tracings of esophageal body peristalsis. (Courtesy of Medtronic Corp., Minneapolis, MN.)

pressure zone. The pressures are recorded with the swallow. Accurate determination of lower esophageal relaxation requires a Dent sleeve. Although this can be measured using side holes, artifact can be created.

The study of the esophageal body determines the muscular activity of the esophagus during swallowing. There are four components of the study: (*1*) peristalsis, (*2*) amplitude of contractions, (*3*) duration of contraction, and (*4*) contraction morphology (Fig. 4). These measurements are made with the distal pressure sensor positioned 3 cm superior to the upper border of the lower esophageal sphincter. The patient takes 10 wet swallows with 5 mL of room temperature water. Esophageal body amplitude is the force with which the esophageal musculature contracts. The amplitude is measured from baseline to the peak of the contraction wave. Duration is the length of time that the esophageal muscle remains contracted. It is measured from the point at which the major upstroke of the contraction begins to the point at which it ends. Velocity is a measurement of the time it takes for a contraction to migrate down the esophagus (unit of measure is cm s^{-1}). These measurements are used to determine esophageal body motility function.

The study of the upper esophageal sphincter includes (*1*) resting pressure, (*2*) relaxation, and (*3*) cricopharyngeal coordination (Fig. 5). The study is done by withdrawing the catheter in 1 cm increments until the upper esophageal sphincter is reached. This is determined when the pressure measured rises above the esophageal baseline. The catheter is positioned so that the first sensor is just superior to the sphincter and the second sensor is at the proximal border of the sphincter. The remaining channels are in the body of the esophagus. The patient is given 5 mL of water for wet swallows. The catheter is withdrawn during this process. However, it should be emphasized that the upper esophageal sphincter is quite asymmetric; therefore, pressure readings are meaningless unless the exact position of the side holes are known.

This concludes the study and the catheter is removed from the patient.

INTERPRETATION OF THE TEST

Esophageal motility disorders are categorized into primary, secondary, and nonspecific (3). Primary esophageal disorders are those in which the dysfunction is limited only to the esophagus. Examples of these include the hypotensive lower esophageal sphincter associated with gastroesophageal reflux disease, achalasia, diffuse esophageal spasm, hypertensive lower esophageal sphincter, and nutcracker esophagus. Secondary esophageal motility disorders are those in which the swallowing occurs as a result of a generalized disease. Examples of these include collagen-vascular disease (e.g., scleroderma), endocrine and metabolic disorders (diabetes mellitus), neuromuscular diseases (myasthenia gravis, multiple sclerosis, and Parkinson's disease), chronic idiopathic intestinal pseudo-obstruction,

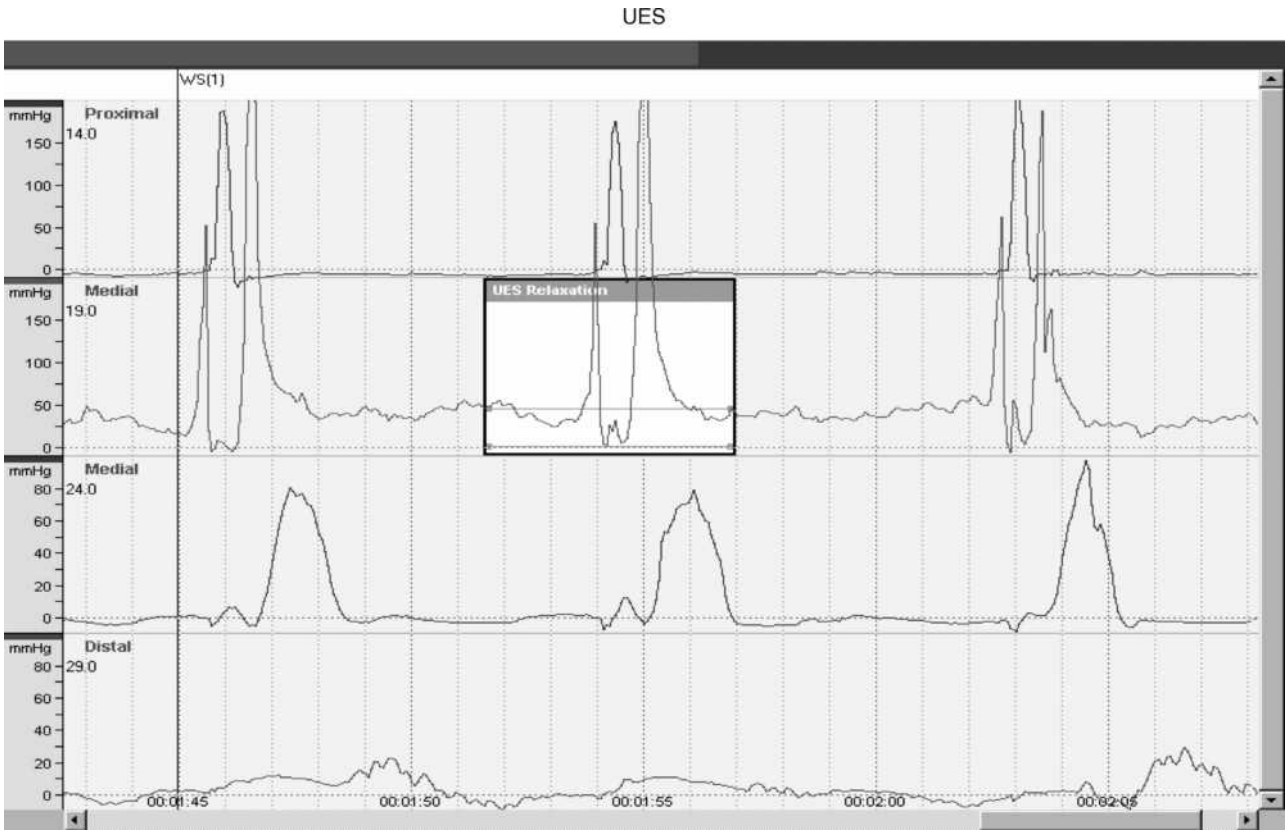


Figure 5. Manometric tracings of the upper esophageal sphincter. (Courtesy of Medtronic Corp., Minneapolis, MN.)

and Chagas' disease. Nonspecific esophageal motility disorders are those that are associated with the patient's symptoms, but the pattern of manometric dysmotility does not fit into the primary or secondary categories. Another category that is becoming better recognized and characterized is ineffective esophageal motility, which occurs when esophageal motility appears normal, but does not result in effective propagation of the food bolus down the esophagus.

BIBLIOGRAPHY

Cited References

1. Castell DO et al. Esophageal Motility Testing, 2nd ed. New York: Elsevier; 1993.
2. Castell JD, Dalton CB. Esophageal manometry. In: Castell DO, editor. The Esophagus. Boston:Little, Brown and Co.; 1992.
3. Stein HJ, Demeester TR, Hinder RA. Outpatient physiologic testing and surgical management of foregut motility disorders. Curr Prob Surg 1992;29:415–555.

See also ENDOSCOPES; GASTROINTESTINAL HEMORRHAGE.

ESU. See ELECTROSURGICAL UNIT (ESU).

EVENT-RELATED POTENTIALS. See EVOKED POTENTIALS.

EVOKED POTENTIALS

RODRIGO QUIAN QUIROGA
University of Leicester
Leicester, United Kingdom

INTRODUCTION

Our knowledge about the brain has increased dramatically in the last decades due to the incorporation of new and extraordinary techniques. In particular, fast computers enable more realistic and complex simulations and boosted the emergence of computational neuroscience. With modern acquisition systems we can record simultaneously up to few hundred neurons and deal with issues like population coding and neural synchrony. Imaging techniques such as magnetic resonance imaging (MRI) allow an incredible visualization of the locus of different brain functions. On the other extreme of the spectrum, molecular neurobiology has been striking the field with extraordinary achievements. In contrast to the progress and excitement generated by these fields of neuroscience, electroencephalography (EEG) and evoked potentials (EPs) have clearly decreased in popularity. What can we learn from scalp electrodes recordings, when one can use sophisticated devices like MRI, or record from dozens of intracranial electrodes? Still a lot.

There are mainly three advantages of the EEG: (*1*) it is relatively inexpensive; (*2*) it is noninvasive, and therefore it can be used in humans, (*3*) it has a very high temporal resolution, thus enabling the study of the dynamics of brain processes. These features make the EEG a very accessible and useful tool. It is particularly interesting for the analysis of high level brain processes that arise from the activity of large cell assemblies and may be poorly reflected by single neuron properties. Moreover, such processes can be well localized in time and even be reflected in time varying patterns (e.g., brain oscillations) that are faster than the time resolution of imaging techniques. The caveat of noninvasive EEGs is the fact that they reflect the average activity of sources far from the recording sites, and therefore do not have an optimal spatial resolution. Moreover, they are largely contaminated by noise and artifacts.

Although the way of recording EEG and EP signals did not change as much as multiunit recordings or imaging techniques, there have been significant advances in the methodology for analyzing the data. In fact, due to their high complexity, low signal/noise ratio, nonlinearity, and nonstationarity, they have been an ultimate challenge for most methods of signal analysis. The development and implementation of new algorithms that are specifically designed for such complex signals allow us to get information beyond the one accessible with previous approaches. These methods open a new gateway to the study of high level cognitive processes in humans with noninvasive techniques and at no great expense. Here, we review some of the most common paradigms to elicit evoked potentials and describe basic and more advanced methods of analysis with special emphasis on the information that can be gained from their use. Although we focus on EEG recordings, these ideas also apply to magnetoencephalograpic (MEG) recordings.

RECORDING

The electroencephalogram measures the average electrical activity of the brain at different sites of the head. Typical recordings are done at the scalp with high conductance electrodes placed at specific locations according to the so-called 10–20 system (1). The activity of each electrode can be referenced to a common passive electrode (or to a pair of linked electrodes placed at the earlobes)—monopolar recordings—or can be recorded differentially between pairs of contiguous electrodes—bipolar recordings—. In the latter case, there are several ways of choosing the electrode pairs. Furthermore, there are specific montages of bipolar recordings designed to visualize the propagation of activity across different directions (1). Intracranial recordings are common in animal studies and are very rare in humans. Intracranial electrodes are mainly implanted in epileptic patients refractory to medication in order to localize the epileptic focus, and then evaluate the feasibility of a surgical resection.

Figure 1 shows the 10–20 electrode distribution (a) and a typical monopolar recording of a normal subject with eyes

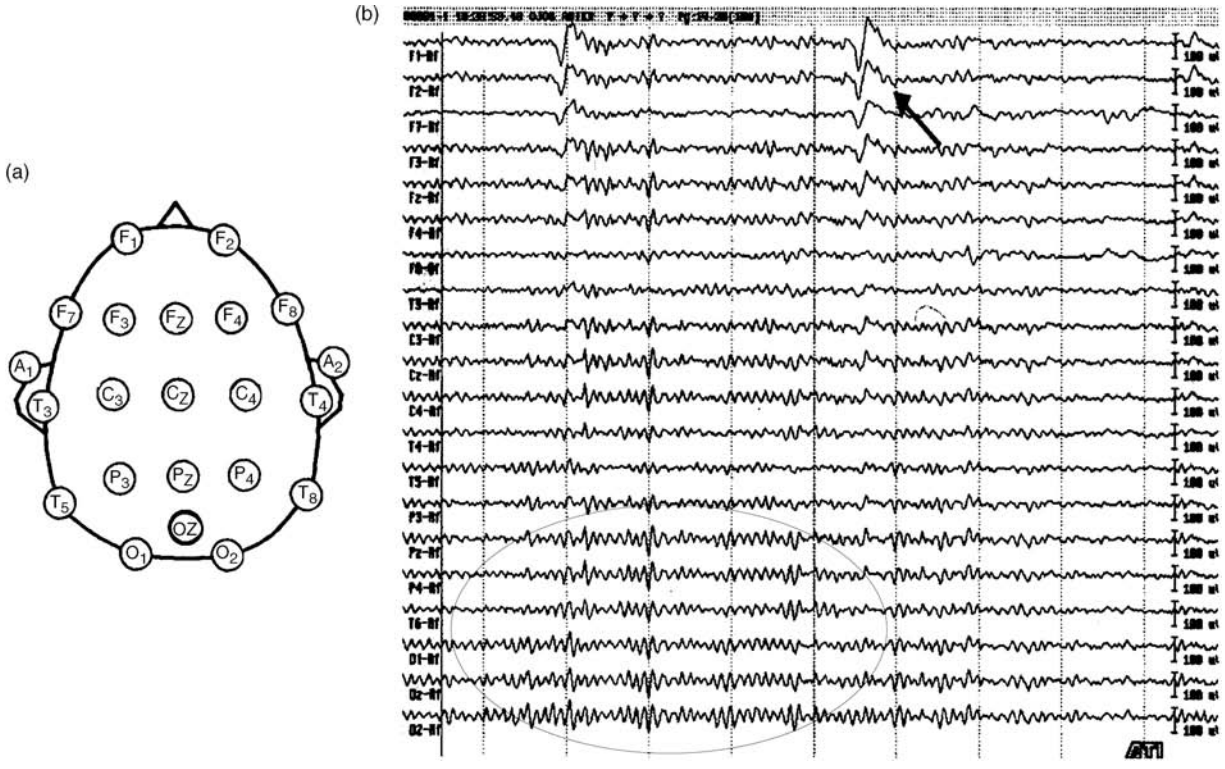


Figure 1. (a) Electrode montage of the 10–20 system and (b) an exemplary EEG recording with this montage. All electrodes are referenced to a linked earlobes reference (A1 and A2). F = frontal, C = central, P = parietal, T = temporal, and O = occipital. Note the presence of blinking artifacts (marked with an arrow) and of posterior alpha oscillations (marked with an oval).

open (b). Note the high amplitude deflections in the anterior recordings due to blinking artifacts. In fact, one of the main problems in EEG analysis is the very low signal/noise ratio. Note also the presence of ongoing oscillations in the posterior sites. These oscillations are $\sim$10 Hz and are known as the alpha rhythm. The EEG brain oscillations of different frequencies and localizations have been correlated with functions, stages and pathologies of the brain (2–4).

In many scientific fields, especially in physics, one very useful way to learn about a system is by studying its reactions to perturbations. In brain research, it is also a common strategy to see how single neurons or large neuronal assemblies, as measured by the EEG, react to different types of stimuli. Evoked potentials are the changes in the ongoing EEG activity due to stimulation. They are time locked to the stimulus and they have a characteristic pattern of response that is more or less reproducible under similar experimental conditions. They are characterized by their polarity and latency, for example, P100 meaning a positive deflection (P for positive) occurring 100 ms after stimulation. The recording of evoked potentials is done in the same way as the EEGs. The stimulus delivery system sends triggers to identify the stimuli onsets and offsets.

GENERATION OF EVOKED POTENTIALS

Evoked potentials are usually considered as the time-locked and synchronized activity of a group of neurons that add to the background EEG. A different approach explains the evoked responses as a reorganization of the ongoing EEG (3,5). According to this view, evoked potentials can be generated by a selective and time-locked enhancement of a particular frequency band or by a phase resetting of ongoing frequencies. In particular, the study of the EPs in the frequency domain attracted the attention of several researchers (see section on Event-Related Oscillations). A few of these works focus on correlations between prestimulus EEG and the evoked responses (4).

SENSORY EVOKED POTENTIALS

There are mainly three modalities of stimulation: visual, auditory, and somatosensory. Visual evoked potentials are usually evoked by light flashes or visual patterns such as a checkerboard or a patch. Figure 2 shows the grand average visual evoked potentials of 10 subjects. Scalp electrodes were placed according to the 10–20 system, with linked earlobes reference. The stimuli were a color reversal of the (black/white) checks in a checkerboard pattern (sidelength of the checks: 50'). There is a positive deflection at $\sim$ 100 ms after stimulus presentation (P100) followed by a negative rebound at 200 ms (N200). These peaks are best defined at the occipital electrodes, which are the closest to the primary visual area. The P100 is also observed in the central and frontal electrodes, but not as well defined and appearing later than in the posterior sites. The P100–N200 complex can be seen as part of an $\sim$10 Hz event-related oscillation as it will be described in the following sections. Visual EPs can be used clinically to identify lesions in the visual pathway, such as the ones caused by optic neuritis and multiple sclerosis (6–9).

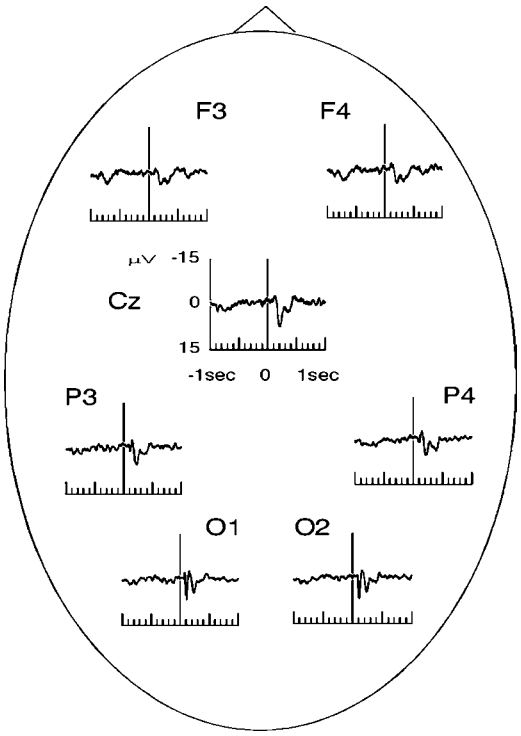


Figure 2. Grand average visual evoked potential. There is mainly one positive response at 100 ms after stimulation (P100) followed by a negative one at 200 ms (N200). These responses are best localized in the posterior electrodes.

Auditory evoked potentials are usually elicited by tones or clicks. According to their latency they are further subdivided into early, middle, and late latency EPs. Early EPs comprise: (*1*) the electrocochleogram, which reflects responses in the first 2.5 ms from the cochlea and the auditory nerve, and (*2*) brain stem auditory evoked potentials (BSAEP), which reflect responses from the brain stem in the first 12 ms after stimulation and are recorded from the vertex. The BSAEP are seen at the scalp due to volume conduction. Early auditory EPs are mainly used clinically to study the integrity of the auditory pathway (10–12). They are also useful for detecting hearing impairments in children and in subjects that cannot cooperate in behavioral audiometry studies. Moreover, the presence of early auditory EPs may be a sign of recovery from coma.

Middle latency auditory EPs are a series of positive and negative waves occurring between 12 and 50 ms after stimulation. Clinical applications of these EPs are very limited due to the fact that the location of their sources is still controversial (10,11). Late auditory EPs occur between 50 and 250 ms after stimulation and consist of four main peaks labeled P50, N100, P150, and N200 according to their polarity and latency. They are of cortical origin and have a maximum amplitude at vertex locations. Auditory stimulation can also elicit potentials with latencies of $>$ 200 ms. These are, however, responses to the context of the

stimulus rather than to its physical characteristics and will be further described in the next section.

Somatosensory EPs are obtained by applying short lasting currents to sensory and motor peripheral nerves and are mainly used to identify lesions in the somatosensory pathway (13). In particular, they are used for the diagnosis of diseases affecting the white matter like multiple sclerosis, for noninvasive studies of spinal cord traumas and for peripheral nerve disorders (13). They are also used for monitoring the spinal cord during surgery, giving an early warning of a potential neurological damage in anesthetized patients (13).

Evoked potentials can be further classified as exogenous and endogenous. Exogenous EPs are elicited by the physical characteristics of the external stimulus, such as intensity, duration, frequency, and so on. In contrast, endogenous EPs are elicited by internal brain processes and respond to the significance of the stimulus. Endogenous EPs can be used to study cognitive processes as discussed in the next section.

EVOKED POTENTIALS AND COGNITION

Usually, the term evoked potentials refers to EEG responses to sensory stimulation. Sequences of stimuli can be organized in paradigms and subjects can be asked to perform different tasks. Event-related potentials (ERPs) constitute a broader category of responses that are elicited by "events", such as the recognition of a "target" stimulus or the lack of a stimulus in a sequence.

Oddball Paradigm and P300

The most common method to elicit ERPs is by using the oddball paradigm. Two different stimuli are distributed pseudorandomly in a sequence; one of them appearing frequently (standard stimulus), the other one being a target stimulus appearing less often and unexpectedly. Standard and target stimuli can be tones of different frequencies, figures of different colors, shapes, and so on. Subjects are usually asked to count the number of target appearances in a session, or to press a button whenever a target stimulus appears.

Figure 3 shows grand-average (10 subjects) visual evoked potentials elicited with an oddball paradigm. Figure 3 a shows the average responses to the frequent (non target) stimuli and (b) shows one to the targets. The experiment was the same as the one described in Fig. 2, but in this case target stimuli were pseudorandomly distributed within the frequent ones. Frequent stimuli (75%) were color reversals of the checks, as in the previous experiment, and target stimuli (25%) were also color reversals but with a small displacement of the checkerboard pattern (see Ref. (14) for details). Subjects had to pay attention to the appearance of the target stimuli.

The responses to the nontarget stimuli are qualitatively similar to the responses to visual EPs (without a task) shown in Fig. 2. As in the case of pattern visual EPs, the P100–N200 complex can be observed upon nontarget and target stimulation. These peaks are mainly related with primary sensory processing due to the fact that they do not depend on the task, they have a relatively short latency

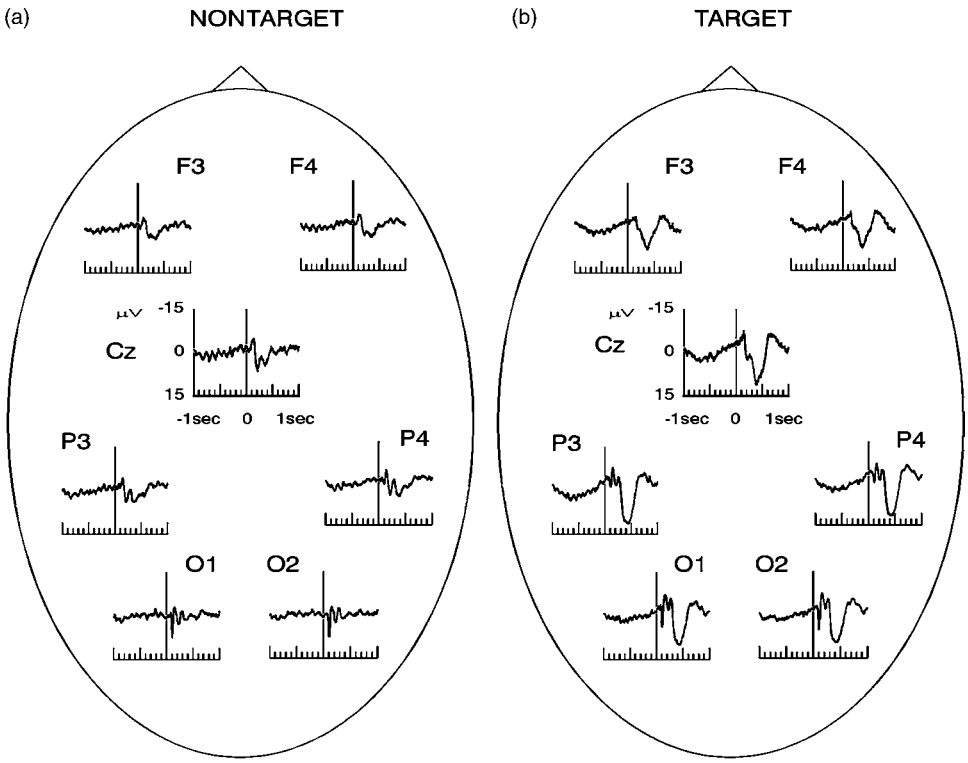


Figure 3. Grand-average pattern visual evoked potentials with an oddball paradigm. (a) Average responses for the nontarget stimuli and (b) average responses for the target stimuli. Note the appearance of a positive deflection at $\sim$ 400 ms after stimulation (P300) only upon the target stimuli.

(100 ms) and they are best defined in the primary visual area (occipital lobe). Note, however, that these components can also modulate their amplitude in tasks with different attention loads (15,16). Target stimulation led to a marked positive component, the P300, occurring between 400 and 500 ms. The P300 is larger in the central and posterior locations.

While the localization of the P300 in the scalp is well known, the localization of the sources of the P300 in the brain are still controversial (for a review see Ref. (17). Since the P300 is task dependent and since it has a relatively long latency, it is traditionally related to cognitive processes such as signal matching, recognition, decision making, attention and memory updating (6,18,19). There have been many works using the P300 to study cognitive processes [for reviews see Refs. (18–20)]. In various pathologies cognition is impaired and this is reflected in abnormal P300 responses, as shown in depression, schizophrenia, dementia and others [for reviews see (18,21)].

The P300 can be also elicited by a passive oddball paradigm (i.e., an oddball sequence without any task). In this case, a P300 like response appears upon target stimulation, reflecting the novelty of the stimulus rather than the execution of a certain task. This response has been named P3a. It is earlier than the classic P300 (also named P3b), it is largest in frontal and central areas and it habituates quickly (22,23).

Mismatch Negativity

Mismatch negativity (MMN) is a negative potential elicited by auditory stimulation. It appears along with any change in some repetitive pattern and peaks between 100 and 200 ms after stimulation (24). It is generally elicited by the passive (i.e., no task) auditory oddball paradigm and it is visualized by subtracting the frequent stimuli from the deviant one. The MMN is generated in the auditory cortex. It is known to reflect auditory memory (i.e., the memory trace of preceding stimuli) and can be elicited even in the absence of attention (25). It provides an index of sound discrimination and has therefore being used to study dyslexia (25). Since MMN reflects a preattentive state, it can be also elicited during sleep (26). Moreover, it has been proposed as an index for coma prognosis (27,28).

Omitted Evoked Potentials

Omitted evoked potentials (OEPs) are similar in nature to the P300 and MMN, but they are evoked by the omission of a stimulus in a sequence (29–31). The nice feature of these potentials is that they are elicited without external stimulation, thus being purely endogenous components. Omitted evoked potentials mainly reflect expectancy (32) and are modulated by attention (31,33). The main problem in recording OEPs is the lack of a stimulus trigger. This results in large latency variations from trial to trial, and therefore OEPs may not be visible after ensemble averaging. Note that trained musicians were shown to have less variability in the latency of the OEP responses (latency jitter) in comparison to non-musicians due to their better time-accuracy (34).

Contingent Negative Variation

Contingent negative variation (CNV) is a slowly rising negative shift appearing before stimulus onset during periods of expectancy and response preparation (35). It is usually elicited by tasks resembling conditioned learning experiments. A first stimulus gives a preparatory signal for a motor response to be carried out at the time of a second stimulus. The CNV reflects the contingency or association between the two stimuli. It has been useful for the study of aging and different psychopathologies, such as depression and schizophrenia (for reviews see Refs. (36,37)). Similar in nature to the CNVs are the "Bereitschaft" or "Readiness" potentials (38), which are negative potential shifts preceeding voluntary movements [for a review see Ref. (36)].

N400

Of particular interest are ERPs showing signs of language processing. Kutas and Hillyard (39,40) described a negative deflection between 300 and 500 ms after stimulation (N400), correlated with the appearance of semantically anomalous words in otherwise meaningful sentences. It reflects "semantic memory"; that is, the predictability of a word based on the semantic content of the preceding sentence (16).

Error Related Negativity

The error related negativity (ERN) is a negative component that appears after negative feedback (41,42). It can be elicited with a wide variety of reaction time tasks and it peaks within 100 ms of an error response. It reaches its maximum over frontal and central areas and convergent evidence from source localization analyses and imaging studies point toward a generation in the anterior cingulated cortex (41).

BASIC ANALYSIS

Figure 4a shows 16 single-trial visual ERPs from the left occipital electrode of a typical subject. These are responses to target stimuli using the oddball paradigm described in the previous section. Note that it is very difficult to distinguish the single-trial ERPs due to their low amplitude and due to their similarity to spontaneous fluctuations in the EEG. The usual way to improve the visualization of the ERPs is by averaging the responses of several trials. Since evoked potentials are locked to the stimulus onset, their contribution will add, whereas one of the ongoing EEG will cancel. Figure 4b shows the average evoked potential. Here it is possible to identify the P100, N200, and P300 responses described in the previous section.

The main quantification of the average ERPs is by means of their amplitudes and latencies. Most research using ERPs compare statistically the distribution of peak amplitudes and latencies of a certain group (e.g., subjects in some particular state or doing some task) with a matched control group. Such comparisons also can be used clinically and, in general, pathological cases show peaks with long latencies and small amplitudes (2,6).

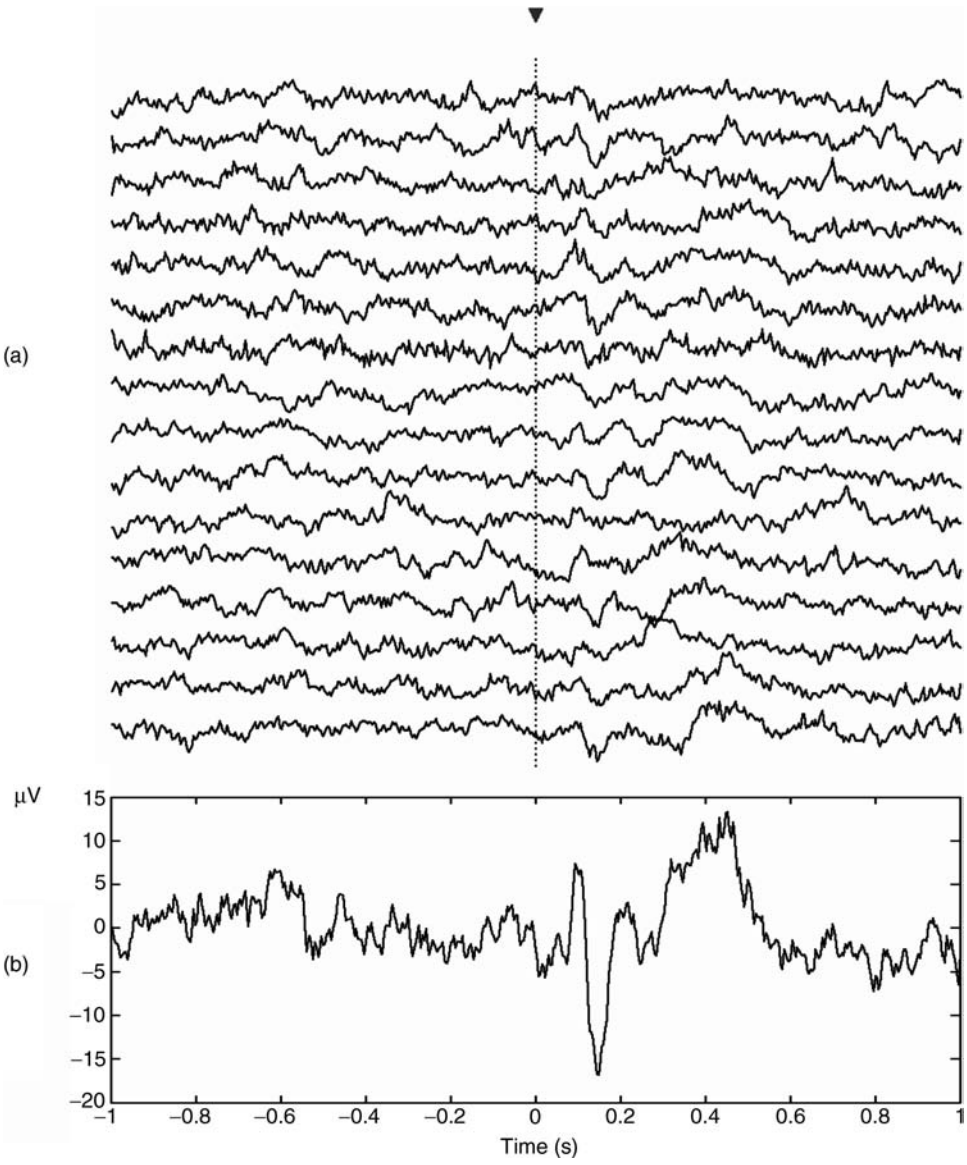


Figure 4. (a) Sixteen single-trial responses of a typical subject and (b) the average response. The triangle marks the time of stimulation. Note that the evoked responses are clearly seen after averaging, but are hardly identified in the single trials.

Another important aspect of ERPs is their topography. In fact, the abnormal localization of evoked responses can have clinical relevance. The usual way to visualize the topography of the EPs is via contour plots (43–47). These are obtained from the interpolation of the EP amplitudes at fixed times. There are several issues to consider when analyzing topographic plots: (1) the way the 3D head is projected into two dimensions, (2) the choice of the reference, (3) the type of interpolation used, and (4) the number of electrodes and their separation (46). These choices can indeed bias the topographic maps obtained.

SOURCE LOCALIZATION

In the previous section, we briefly discussed the use of topographic representations of the EEG and EPs. Besides the merit of the topographic representation given by these maps, the final goal is to get a hint on the sources of the activity seen at the scalp. In other words, given a certain distribution of voltages at the scalp one would like to estimate the location and magnitude of their sources of generation. This is known as the inverse problem and it has no unique solution. The generating sources are usually assumed to be dipoles, each one having six parameters to be estimated, three for its position and three for its magnitude. Clearly, the complexity of the calculation increases rapidly with the number of dipoles and, in practice, no more than two or three dipoles are considered. Dipole sources are usually estimated using spherical head models. These models consider the fact that the electromagnetic signal has to cross layers of different impedances, such as the dura mater and the scull. A drawback of spherical head models is the fact that different subjects have different head shapes. This led to the introduction of realistic head models, which are obtained by modeling the head shape using MRI scans and computer simulations. Besides all these issues, there are already some impressive results in the literature [see Refs. (49,86)] and references cited therein describing the use and applications of the LORETA

software; Ref. (50) and an extensive list of publications using the BESA software at http://www.besa.de). Since a reasonable estimation of the EEG and EP sources critically depends on the number of electrodes, dipole location has been quite popular for the analysis of magnetoencephalograms, which have more recording sites.

EVENT-RELATED OSCILLATIONS

Evoked responses appear as single peaks or as oscillations generated by the synchronous activation of a large network. The presence of oscillatory activity induced by different type of stimuli has been largely reported in animal studies. Bullock (51) gives an excellent review of the subject going from earlier studies by Adrian (52) to more recent results in the 1990s (some of the later studies are included in Ref. (53). Examples are event-related oscillations of 15–25 Hz in the retina of fishes in response to flashes (54), gamma oscillations in the olfactory bulb of cats and rabbits after odor presentation (55,56) and beta oscillations in the olfactory system of insects (57,58). Moreover, it has been proposed that these brain oscillations play a role in information processing (55). This idea became very popular after the report of gamma activity correlated to the binding of perceptual information in anesthetized cats (59).

Event-related oscillations in animals are quite robust and in many cases visible by the naked eye. In humans, this activity is more noisy and localized in time. Consequently, more sophisticated time–frequency representations, like the one given by the wavelet transform, are needed in order to precisely localize event-related oscillations both in time and frequency. We finish this section with a cautionary note about event-related oscillations, particularly important for human studies. Since oscillations are usually not clear in the raw data, digital filters are used in order to visualize them. However, one should be aware that digital filters can introduce "ringing effects" and single peaks in the original signal can look like oscillations after filtering. In Fig. 5, we exemplify this effect by showing a delta function (a) filtered with a broad and a narrow band elliptic filter (b,c, respectively). Note that the original delta function can be mistaken for an oscillation after filtering, especially with the narrow band filter [see also Ref. (51)].

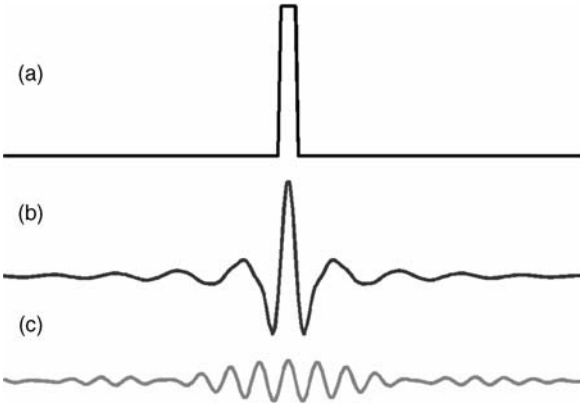


Figure 5. A delta function (a) after broad (b) and narrow (c) band pass filtering. Note that single peaks can look like oscillations due to filtering.

WAVELET TRANSFORM AND EVENT-RELATED OSCILLATIONS

Signals are usually represented either in the time or in the frequency domain. The best time representation is given by the signal itself and the best frequency representation is given by its Fourier transform (FT). With the FT it is possible to estimate the power spectrum of the signal, which quantifies the amount of activity for each frequency. The power spectrum has been the most successful method for the analysis of EEGs (2), but it lacks time resolution. Since event-related oscillations appear in a short time range, a simultaneous representation in time and frequency is more appropriate.

The Wavelet transform (WT) gives a time–frequency representation that has two main advantages: (*1*) optimal resolution in the time and frequency domains; (*2*) no requirement of stationarity. It is defined as the correlation between the signal *x(t)* and the wavelet functions $\psi_{a,b}(t)$

$$W_\psi X(a,b) = \langle x(t)|\psi_{\alpha,\beta}(\tau)\rangle \tag{1}$$

where $\psi_{a,b}(t)$ are dilated (contracted) and shifted versions of a unique *wavelet function* $\psi(t)$

$$\psi_{a,b} = |a|^{-1/2}\psi\left(\frac{t-b}{a}\right) \tag{2}$$

(a, b are the scale and translation parameters, respectively). The WT gives a decomposition of *x(t)* in different scales, tending to be maximum at those scales and time locations where the wavelet best resembles *x(t)*. Moreover, Eq. 1 can be inverted, thus giving the reconstruction of *x(t)*.

The WT maps a signal of one independent variable t onto a function of two independent variables a, b. This procedure is redundant and not efficient for algorithm implementations. In consequence, it is more practical to define the WT only at discrete scales a and discrete times b by choosing the set of parameters $\{a_j = 2^{-j};\ b_j,\ k = 2^{-j}k\}$, with integers j, k.

Contracted versions of the wavelet function match the high frequency components of the original signal and the dilated versions match low frequency oscillations. Then, by correlating the original signal with wavelet functions of different sizes we can obtain the details of the signal at different scales. The correlations with the different wavelet functions can be arranged in a hierarchical scheme called multiresolution decomposition (60). The multiresolution decomposition separates the signal into "details" at different scales and the remaining part is a coarser representation named "approximation".

Figure 6 shows the multiresolution decomposition of the average ERP shown in Fig. 4. The left part of the figure shows the wavelet coefficients and the right part shows the corresponding reconstructed waveforms. After a five octave wavelet decomposition using B-Spline wavelets (see Refs. 14,61 for details) the coefficients in the following bands were obtained (in brackets the EEG frequency bands that approximately correspond to these values): D1: 63–125 Hz, D2: 31–62 Hz (gamma), D3: 16–30 Hz (beta), D4:

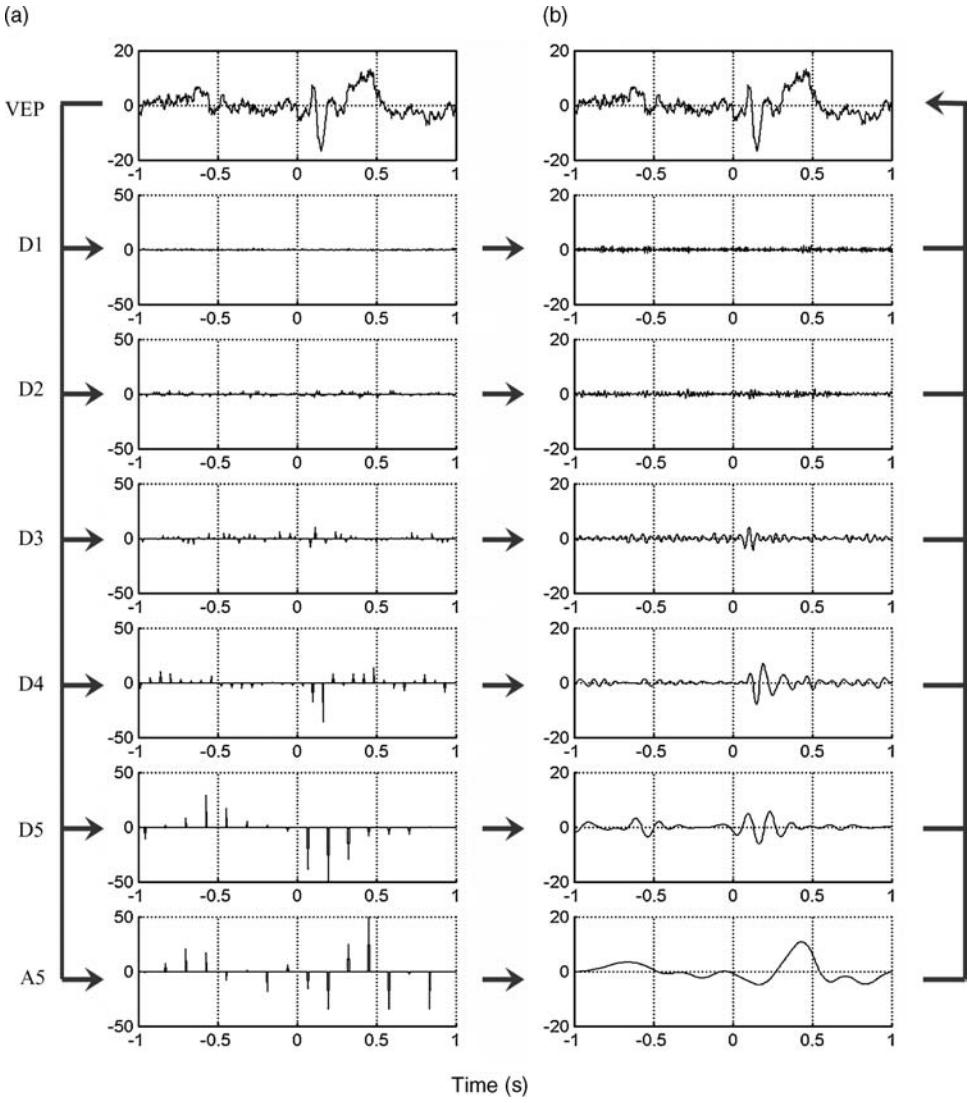


Figure 6. Multiresolution decomposition (a) and reconstruction (b) of an average evoked potential. D1–D5 and A5 are the different scales in which the signal is decomposed.

8–15 Hz (alpha), D5: 4–7 Hz (theta), and A5: 0.5–4 Hz (delta). Note that the addition of the reconstructed waveforms of all frequency bands returns the original signal. In the first 0.5 s after stimulation there is an increase in the alpha and theta bands (D4, D5) correlated with P100–N200 complex and later there is an increase in the delta band (A5) correlated with the P300. As an example of the use of wavelets for the analysis of event related oscillations, in the following we focus on the responses in the alpha band.

The grand average (across subjects) ERP is shown on left side of Fig. 7. Upper plots correspond to the responses to NT stimuli and lower plots to *T* stimuli. Only left electrodes and Cz are shown, the responses of the right electrodes being qualitatively similar. For both stimulus types we observe the P100–N200 complex and the P300 appears only upon target stimulation. Center and right plots of Fig. 7 show the alpha band wavelet coefficients and the filtered ERPs reconstructed from these coefficients, respectively. Amplitude increases are distributed over the entire scalp for the two stimulus types, best defined in the occipital electrodes. They appear first in the occipital electrodes, with an increasing delay in the parietal, central, and frontal locations. The fact that alpha responses are not modulated by the task and the fact that their maximal and earliest appearance is in occipital locations (the primary visual sensory area) point toward a distributed generation and a correlation with sensory processing (14,61). Note that these responses are localized in time, thus stressing the use of wavelets.

In recent years, there have been an increasing number of works applying the WT to the study of event-related oscillations. Several of these studies dealt with gamma oscillations, encouraged by the first results by Gray and coworkers (59). In particular, induced gamma activity has been correlated to face perception (62), coherent visual perception (63), visual search tasks (64) cross-modal integration (64,65), and so on.

Another interesting approach to study event-related oscillations is the one given by the concepts of event-related synchronization (ERS) and event-related desynchronization (ERD), which characterize increases and decreases of the power in a given frequency band (66,67). Briefly, the band limited power is calculated for each single trial and then averaged across trials. Since ERS and ERD are defined as an

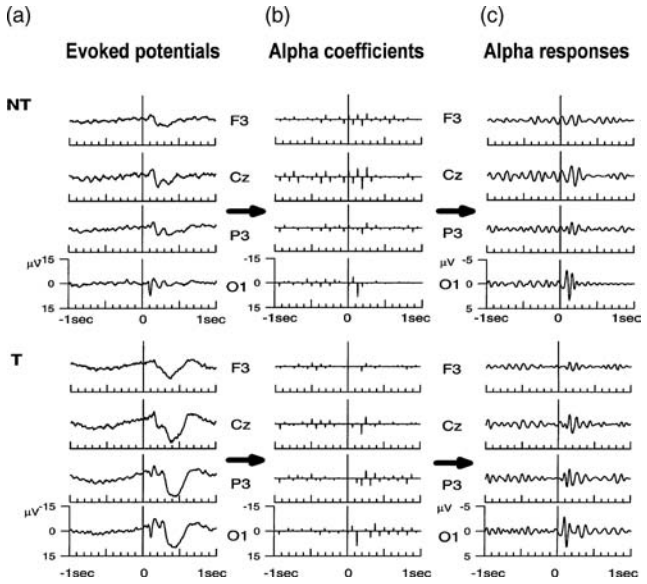


Figure 7. Grand average visual EPs to nontarget (NT) and target (T) stimuli in an oddball paradigm (a). Both b+ and c plots shows the wavelet coefficients in the alpha band and the corresponding reconstruction of the signal from them, respectively.

average of the power of the signal, they are sensitive to phase locked as well as nonphase locked oscillations. Interestingly, similar concepts and a simple measure of phase locking can be also defined using wavelets (68).

SINGLE–TRIAL ANALYSIS

As shown in Fig. 4, averaging several trials increases the signal/noise ratio of the EPs. However, it relies on the basic assumption that EPs are an invariant pattern perfectly locked to the stimulus that lays on an independent stationary and stochastic background EEG signal. This assumption is in a strict sense not valid. In fact, averaging implies a loss of information related to systematic or unsystematic variations between the single trials. Furthermore, these variations (e.g., latency jitters) can affect the validity of the average EP as a representation of the single trial responses.

Several techniques have been proposed to improve the visualization of the single-trial EPs. Some of these approaches involve the filtering of single-trial traces by using techniques that are based on the Wiener formalism. This provides an optimal filtering in the mean-square error sense (69,70). However, these approaches assume that the signal is a stationary process and, since the EPs are compositions of transient responses with different time and frequency localizations, they are not likely to give optimal results. A obvious advantage is to implement time-varying strategies. In the following, we describe a recently proposed denoising implementation based on the WT to obtain the EPs at the single trial level (71,72). Other works also reported the use of wavelets for filtering average EPs or for visualizing the EPs in the single trials (73–77) see a brief discussion of these methods in Ref. 72.

In Fig. 6, we already showed the wavelet decomposition and reconstruction of an average visual EP. Note that the P100–N200 response is mainly correlated with the first poststimulus coefficient in the details D4–D5. The P300 is mainly correlated with the coefficients at ~400–500 ms in A5. This correspondence is easily identified because: (1) the coefficients appear in the same time (and frequency) range as the EPs and (2) they are relatively larger than the rest due to phase-locking between trials (coefficients related with background oscillations are diminished in the average). A straightforward way to avoid the fluctuations related with the ongoing EEG is by equaling to zero those coefficients that are not correlated with the EPs. However, the choice of these coefficients should not be solely based on the average EP and it should also consider the time ranges in which the single-trial EPs are expected to occur (i.e., some neighbor coefficients should be included in order to allow for latency jitters).

Figure 8a shows the coefficients kept for the reconstruction of the P100–N200 and P300 responses. Figure 8b shows the contributions of each level obtained by eliminating all the other coefficients. Note that in the final reconstruction of the average response (uppermost right plot) background EEG oscillations are filtered. We should remark that this is usually difficult to be achieved with a Fourier filtering approach due to the different time and frequency localizations of the P100–N200 and P300 responses, and also due to the overlapping frequency components of these peaks and the ongoing EEG. In this context, the main advantage of Wavelet denoising over conventional filtering is that one can select different time windows for the different scales. Once the coefficients of interest are identified from the average ERP, we can apply the same procedure to each single trial, thus filtering the contribution of background EEG activity.

Figure 9 shows the first 15 single trials and the average ERP for the recording shown in the previous figure. The raw single trials have been already shown in Fig. 4. Note that with denoising (red curves) we can distinguish the P100–N200 and the P300 in most of the trials. Note also that these responses are not easily identified in the original signal (gray traces) due to their similarity with the ongoing EEG. We can also observe some variability between trials. For an easier visualization Fig. 10 shows a contour plot of the single trial ERPs after denoising. This figure is the output of a software package for denoising EPs (EP_den) available at www.vis.caltech.edu/~rodri. In the denoised plot, we observe a gray pattern followed by a black one between 100 and 200 ms, corresponding to the P100–N200 peaks. The more unstable and wider gray pattern at ~ 400–600 ms corresponds to the P300. In particular, it has been shown that wavelet denoising improves the visualization of the single trial EPs (and the estimation of their amplitudes and latencies) in comparison with the original data and in comparison with previous approaches, such as Wiener filtering (72).

APPLICATIONS OF SINGLE-TRIAL ANALYSIS

The single-trial analysis of EPs has a wide variety of applications. By using correlations between the average EP and the single-trial responses, it is possible to calculate

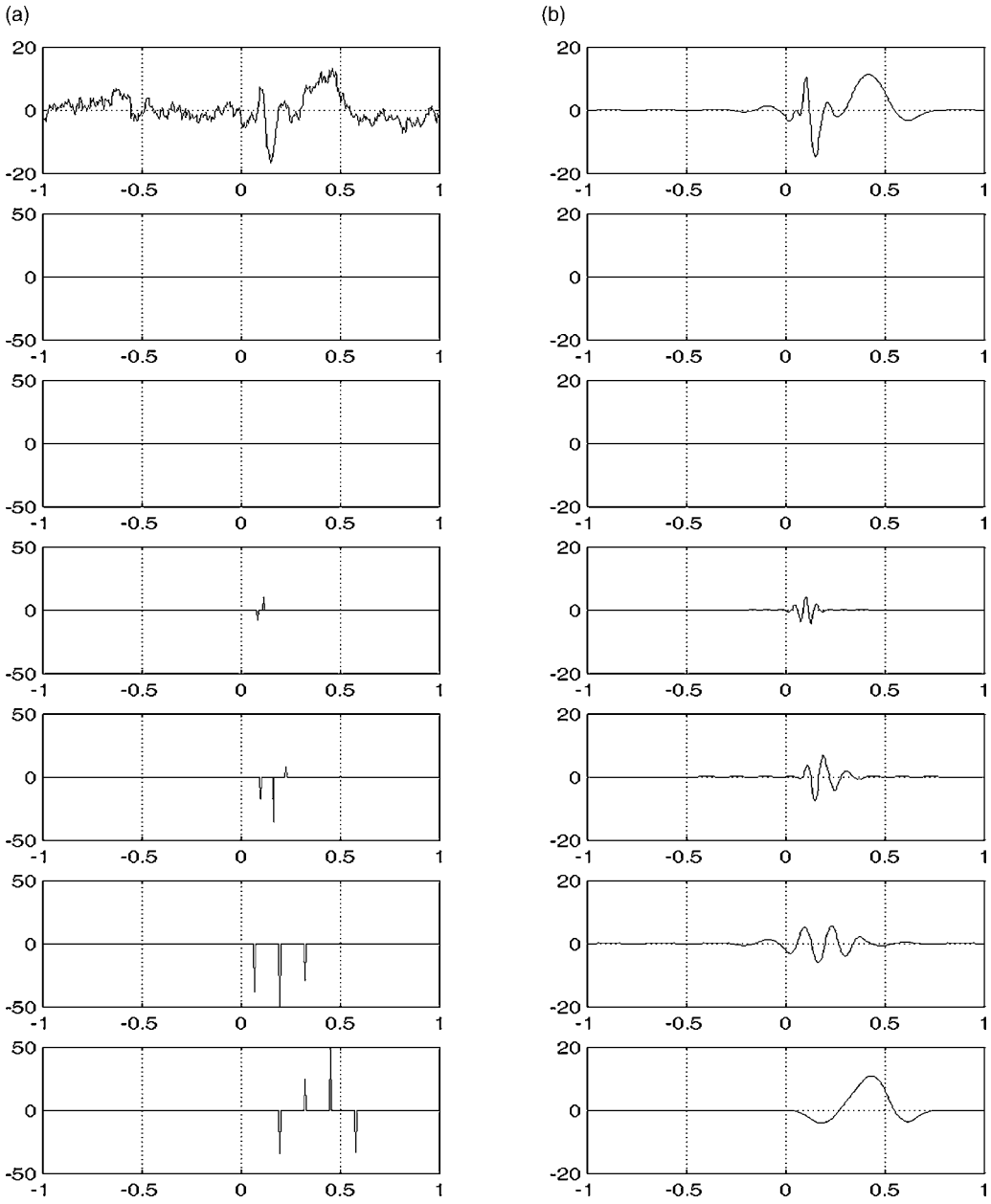


Figure 8. Principle of wavelet denoising. The reconstruction of the signal is done using only those coefficients correlated with the EPs. See text for details.

selective averages including only trials with good responses (71,72). Moreover, it is possible to eliminate effects of latency jitters by aligning trials according to the latency of the single-trial peaks (71,72). The use of selective averages as well as jitter corrected averages had been proposed long ago (78,79). Wavelet denoising improves the identification of the single-trials responses, thus facilitating the construction of these averages.

Some of the most interesting features to study in single-trial EPs are the changes in amplitude and latency of the peaks from trial to trial. It is possible to calculate amplitude and latency jitters: information that is not available in the average EPs. For example, trained musicians showed smaller latency jitters of omitted evoked potentials in comparison with nonmusicians (34). Variations in amplitude and latency can be also systematic. Exponential decreases in different EP components have been related to habituation processes both in humans and in rats (80–82). Furthermore, the appearance of a P3-like component in the rat entorhinal cortex has been correlated to the learning of a go/no-go task (83). In humans, it has recently been shown that precise timing of the single-trial evoked responses accounts for a sleep-dependent automatization of perceptual learning (84).

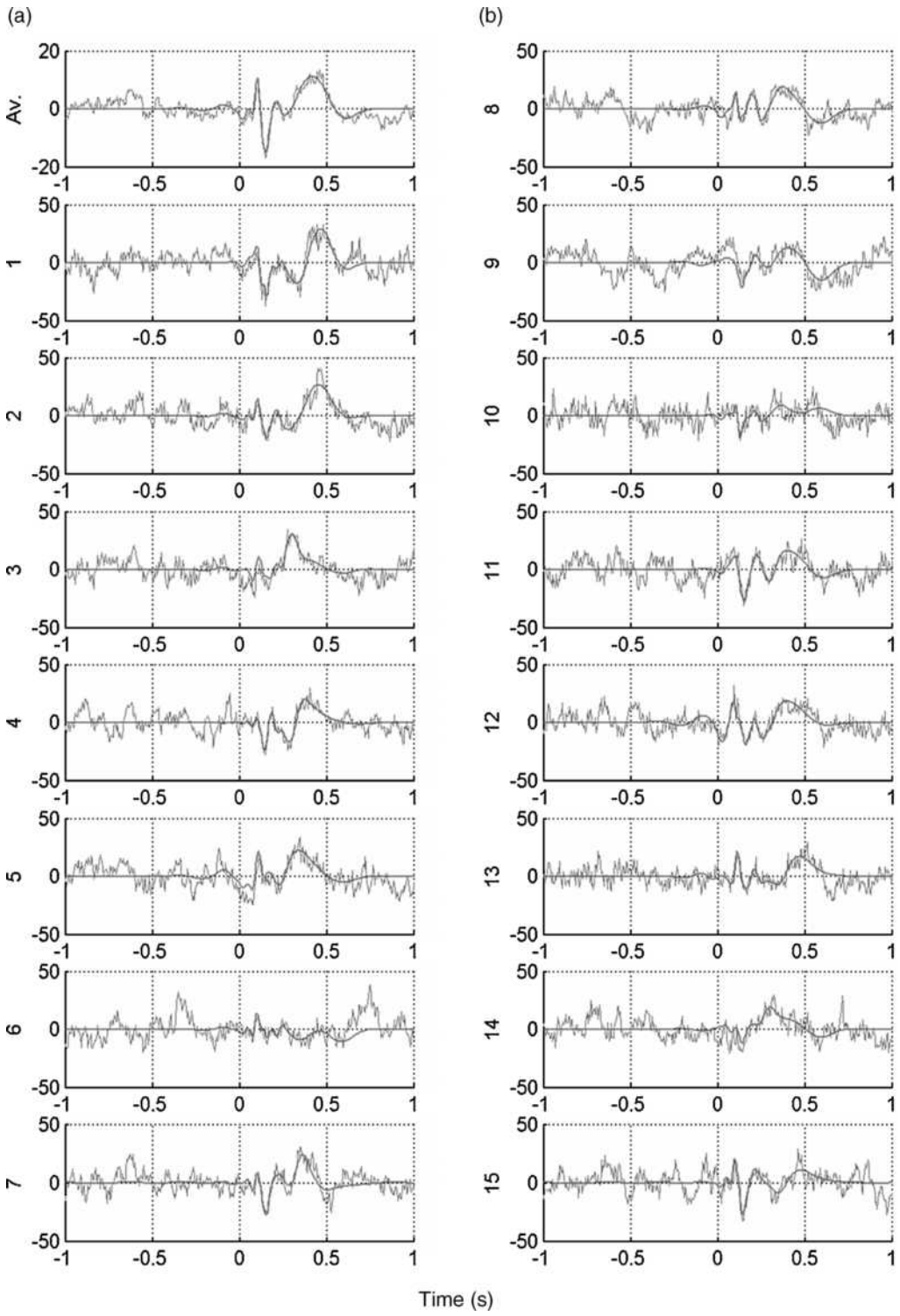


Figure 9. Average EP and the single-trial responses corresponding to the data shown in the previous figure, with (black) and without (gray) denoising. Note that after denoising it is possible to identify the single-trial responses.

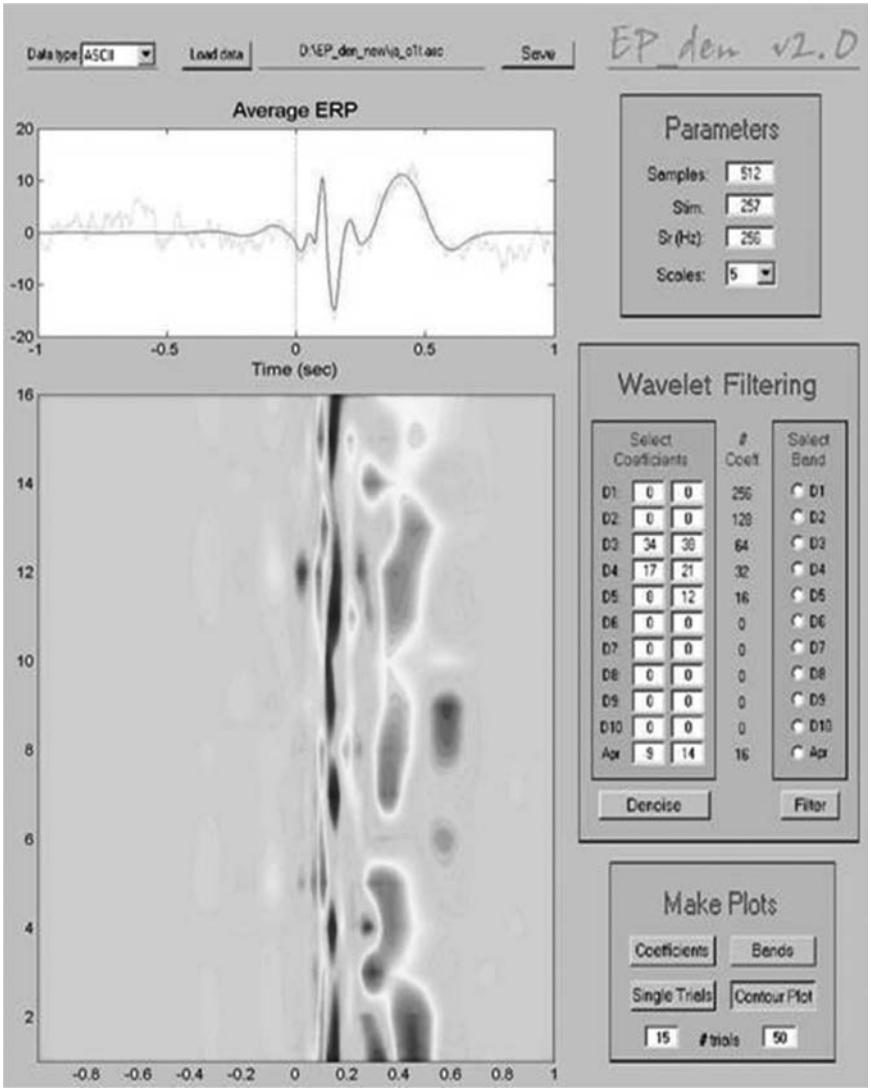


Figure 10. Contour plot of the single-trial responses shown in the previous figure. The graphic user interface used to generate this plot is available at www.vis.caltech.edu/ ~ rodri.

CONCLUDING COMMENT

In addition to clinical applications, EPs are very useful to study high level cognitive processes. Their main advantages over other techniques are their low cost, their non-invasiveness and their good temporal resolution. Of particular interest is the study of trial-to-trial variations during recording sessions. Supported by the use of new and powerful methods of signal analysis, the study of single trial EPs and their correlation to different behavioral processes seems one of the most interesting directions of future research. In conclusion, the good and old EEG and its cousin, the EP, have a lot to offer, especially when new and powerful methods of analysis are applied.

BIBLIOGRAPHY

Cited References

1. Reilly EL. EEG recording and operation of the apparatus. In: Niedermeyer E, Lopes da Silva F, editors. Electroencephalography: Basic principles, clinical applications and related fields. Baltimore: Williams and Wilkins; 1993.
2. Niedermeyer E, Lopes da Silva F, editors. Electroencephalography: Basic principles, clinical applications and related fields. Baltimore: Williams and Wilkins; 1993, p 1097–1123.
3. Başar E. EEG-Brain dynamics. Relation between EEG and brain evoked potentials. Amsterdam: Elsevier; 1980.
4. Başar E. Brain function and oscillations. Vol. I: Brain oscillations, principles and approaches. Vol. II: Integrative brain function: Neurophysiology and cognitive processes. Berlin-Heidelberg-New York: Springer; 1999.
5. Sayers B, Beagley HA, Hanshall WR. The mechanisms of auditory evoked EEG responses. Nature (London) 1974;247: 481–483.
6. Regan D. Human brain electrophysiology. Evoked potentials and evoked magnetic fields in science and medicine. Amsterdam: Elsevier; 1989.
7. Celesia GG. Visual evoked potentials and electroretinograms. In: Niedermeyer E, Lopes da Silva F, editors. Electroencephalography: Basic principles, clinical applications and related fields. Baltimore: Williams and Wilkins; 1993.
8. Desmetdt JE, editor. Visual evoked potentials in man: new developments. Oxford: Clarendon Press; 1977a.
9. Epstein CM. Visual evoked potentials. In: Daly DD, Pedley TA, editor. Current practice of clinical electroencephalography. New York: Raven Press; 1990.

10. Celesia GG, Grigg MM. Auditory evoked potentials. In: Niedermeyer E, Lopes da Silva F, editors. Electroencephalography: Basic principles, clinical applications and related fields. Baltimore: Williams and Wilkins; 1993.
11. Picton TW. Auditory evoked potentials. In: Daly DD, Pedley TA, editors. Current practice of clinical electroencephalography. New York: Raven Press; 1990.
12. Desmetdt JE, editor. Auditory evoked potentials in man. Basel: S. Karger; 1977b.
13. Erwin CW, Rozear MP, Radtke RA, Erwin AC. Somatosensory evoked potentials. In: Niedermeyer E, Lopes da Silva F, editors. Electroencephalography: Basic principles, clinical applications and related fields. Baltimore: Williams and Wilkins; 1993.
14. Quian Quiroga R, Schürmann M. Functions and sources of event-related EEG alpha oscillations studied with the Wavelet Transform. Clin Neurophysiol 1999;110:643–655.
15. Hillyard SA, Hink RF, Schwent VL, Picton TW. Electrical signs of selective attention in the human brain. Science 1973;182:177–179.
16. Hillyard SA, Kutas M. Electrophysiology of cognitive processing. Ann Rev Psychol 1983;34:33–61.
17. Molnar M. On the origin of the P3 event-related potential component. Int J Psychophysiol 1994;17:129–144.
18. Picton TW. The P300 wave of the human event-related potential. J Clin Neurophysiol 1992;9:456–479.
19. Polich J, Kok A. Cognitive and biological determinants of P300: an integrative review. Biol Psychol 1995;41:103–146.
20. Pritchard WS. Psychophysiology of P300. Psychol Bull 1984;89:506–540.
21. Polich J. P300 in clinical applications: Meaning, method and measurement. Am J EEG Technol 1991;31:201–231.
22. Polich J. Neuropsychology of P3a and P3b: A theoretical overvier. In: Arikan K, Moore N, editors. Advances in Electrophysiology in Clinical Practice and Research. Wheaton (IL): Kjellberg; 2002.
23. Polich J, Comerchero MD. P3a from visual stimuli: Typicality, task, and topography. Brain Topogr 2003;15:141–152.
24. Naatanen R, Tervaniemi M, Sussman E, Paavilainen P, Winkler I. 'Primitive intelligence' in the auditory cortex. Trends Neurosci 2001;24:283–288.
25. Naatanen R. Mismatch negativity: clinical research and possible applications. Int J Psychophysiol 2003;48:179–188.
26. Atienza M, Cantero JL. On-line processing of complex sounds during human REM sleep by recovering information from long-term memory as revealed by the mismatch negativity. Brain Res 2001;901:151–160.
27. Kane NM, Curry SH, Butler SR, Gummins BH. Electrophysiological indicator of awakening from coma. Lancet 1993; 341:688.
28. Fischer C, Morlet D, Bouchet P, Luante J, Jourdan C, Salford F. Mismatch negativity and late auditory evoked potentials in comatose patients. Clin Neurophysiol 1999;11:1601–1610.
29. Simson R, Vaughan HG Jr, Ritter W. The scalp topography of potentials associated with missing visual or auditory stimuli. Electr Clin Neurophysiol 1976;40:33–42.
30. Ruchkin DS, Sutton S, Munson R, Silver K, Macar F. P300 and feedback provided by absence of the stimulus. Psychophysiology 1981;18:271–282.
31. Bullock TH, Karamursel S, Achimowics JZ, McClune MC, Basar-Eroglu C. Dynamic properties of human visual evoked and omitted stimulus potentials. Electr Clin Neurophysiol 1994;91:42–53.
32. Jongsma MLA, Eichele T, Quian Quiroga R, Jenks KM, Desain P, Honing H, VanRijn CM. The effect of expectancy on omission evoked potentials (OEPs) in musicians and nonmusicians. Psychophysiology 2005;42:191–201.
33. Besson M, Faita F. An event-related potential (ERP) study of musical expectancy:Comparisons of musicians with nonmusicians. J Exp Psychol, Human Perception and Performance 1995;21:1278–1296.
34. Jongsma MLA, Quian Quiroga R, VanRijn CM. Rhythmic training decreases latency-jitter of omission evoked potentials (OEPs). Neurosci Lett 2004;355:189–192.
35. Walter WG, Cooper R, Aldridge VJ, McCallum WC, Winter AL. Contingent negative variation. An electric sign of sensorimotor association and expectancy in the human brain. Nature (London) 1964;203:380–384.
36. Birbaumer N, Elbert T, Canavan AGM, Rockstroh B. Slow potentials of the cerebral cortex and behavior. Physiol Rev 1990;70:1–41.
37. Tecce JJ, Cattanach L. Contingent negative variation (CNV). In: Niedermeyer E, Lopes da Silva F, editors. Electroencephalography: Basic principles, clinical applications and related fields. Baltimore: Williams and Wilkins; 1993. p 1097–1123.
38. Kornhuber HH, Deeke L. Hirnpotentialanderungen bei Willkurbewegungen und passiven Bewegungen des Menschen. Bereitschaftspotential und reafferente Potentiale. Pfluegers Arch 1965;248:1–17.
39. Kutas M, Hillyard SA. Event-related brain potentials to semantically inappropriate and surprisingly large words. Biol Psychol 1980a;11:99–116.
40. Kutas M, Hillyard SA. Reading senseless sentences: Brain potentials reflect semantic incongruity. Science 1980b;207: 203–205.
41. Holroyd CB, Coles GH. The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. Psychol Rev 2002;109:679–709.
42. Nieuwenhuis S, Holroyd CB, Mol N, Coles MGH. Neurosci Biobehav Rev 2004;28:441–448.
43. Vaughan HG Jr, Costa D, Ritter W. Topography of the human motor potential. Electr Clin Neurophysiol 1968;27:(Suppl.) 61–70.
44. Duffy FH, Burchfiel JL, Lombroso CT. Brain electrical activity mapping (BEAM): a method for extending the clinical utility of EEG and evoked potential data. Ann Neurol 1979;5:309–321.
45. Lehman D. Principles of spatial analysis. In: Gevins AS, Remond A, editors. Methods of analysis of brain electrical and magnetic signals. Amsterdam: Elsevier; 1987.
46. Gevins AS. Overview of computer analysis. In: Gevins AS, Remond A, editors. Methods of analysis of brain electrical and magnetic signals. Amsterdam: Elsevier; 1982.
47. Lopes da Silva F. EEG Analysis: Theory and Practice. In: Niedermeyer E, Lopes da Silva F, editors. Electroencephalography: Basic principles, clinical applications and related fields. Baltimore: Williams and Wilkins; 1993.
48. Fender DH. Source localization of brain electrical activity. In: Gevins AS, Remond A, editors. Methods of analysis of brain electrical and magnetic signals. Amsterdam: Elsevier; 1987.
49. Pascual-Marqui RD, Esslen M, Kochi K, Lehmann D. Functional imaging with low resolution brain electromagnetic tomography (LORETA): a review. Methods Findings Exp Clin Pharmacol 2002;24:91–95.
50. Scherg M, Berg P. New concepts of brain source imaging and localization. Electr Clin Neurophysiol 1996;46: (Suppl.) 127–137.
51. Bullock TH. Introduction to induced rhythms: A widespread, hererogeneous class of oscillations. In: Başar E, Bullock T, editors. Induced rhythms in the brain. Boston: Birkhauser; 1992.
52. Adrian ED. Olfactory reactions in the brain of the hedgehog. J Physiol 1942;100:459–473.
53. Başar E, Bullock T, editors. Induced rhythms in the brain. Boston: Birkhauser; 1992.

54. Bullock TH, Hofmann MH, New JG, Nahm FK. Dynamics properties of visual evoked potentials in the tectum of cartilaginous and bony fishes, with neuroethological implications. J Exp Zool 1991;5: (Suppl.) 142–155.
55. Freeman WJ. Mass action in the nervous system. New York: Academic Press; 1975.
56. Freeman WJ, Skarda CA. Spatial EEG-patterns, non-linear dynamics and perception: the neo-Sherringtonian view. Brain Res Rev 1981;10:147–175.
57. Laurent G, Naraghi M. Odorant-induced oscillations in the mushroom bodies of the locust. J Neurosci 1994;14:2993–3004.
58. Laurent G, Wehr M, Davidowitz H. Odour encoding by temporal sequences of firing in oscillating neural assemblies. J Neurosci 1996;16:3837–3847.
59. Gray CM, Koenig P, Engel AK, Singer W. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. Nature (London) 1989;338:334–337.
60. Mallat S. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Machine Intell 1989;2:674–693.
61. Quian Quiroga R, Sakowicz O, Basar E, Schurmann M. Wavelet transform in the analysis of the frequency composition of evoked potentials. Brain Res Protocols 2001;8:16–24.
62. Rodriguez E, George N, Lachaux JP, Martinerie J, Renault B, Varela FJ. Perception's shadow: Long-distance synchronization of human brain activity. Nature (London) 1999;397:430–433.
63. Tallon-Baudry C, Bertrand O, Delpuech C, Pernier J. Simulus specificity of phase-locked and non-phase-locked 40 Hz visual responses in human. J Neurosc 1996;16:4240–4249.
64. Sakowicz O, Quian Quiroga R, Başar E, Schürmann M. Bisensory stimulation increases gamma-responses over multiple cortical regions. Cogn Brain Res 2001;11:267–279. Tallon-Baudry C, Bertrand O, Delpuech C, Pernier J. Activity induced by a visual search task in humans. J Neurosc 1997;15:722–734.
65. Sakowicz O, Quian Quiroga R, Schürmann M, Basar E. Spatio-temporal frequency characteristics of intersensory components in audio-visually evoked Potentials. Cog Brain Res 2005;23:87–99.
66. Pfurtscheller G, Lopes da Silva RH. Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin Neurophysiol 1999a;110:1842–1857.
67. Pfurtscheller G, Lopes da Silva RH, editors. Event-related desynchronization. Amsterdam: Elsevier; 1999b.
68. Quian Quiroga R, Basar E, Schürmann M. Phase locking of event-related alpha oscillations. In: Lehnertz K, Elger CE, Arnhold J, Grassberger P, editors. Chaos in Brain? World Scientific; 2000.
69. Walter DO. A posteriori "Wiener filtering" of average evoked responses. Electr Clin Neurophysiol 1969;27: (Suppl.) 61–70.
70. Doyle DJ. Some comments on the use of Wiener filtering for the estimation of evoked potentials. Electr Clin Neurophysiol 1975;38:533–534.
71. Quian Quiroga R. Obtaining single stimulus evoked potentials with wavelet denoising. Phys D 2000;145:278–192.
72. Quian Quiroga R, Garcia H. Single-trial event-related potentials with wavelet denoising. Clin Neurophysiol 2003;114: 376–390.
73. Bartnik EA, Blinowska KJ, Durka PJ. Single evoked potential reconstruction by means of wavelet transform. Biol Cybern 1992;67:175–181.
74. Bertrand O, Bohorquez J, Pernier J. Time-frequency digital filtering based on an invertible wavelet transform: an application to evoked potentials. IEEE Trans Biomed Eng 1994;41:77–88.
75. Thakor N, Xin-rong G, Yi-Chun S, Hanley D. Multiresolution wavelet analysis of evoked potentials. IEEE Trans Biomed Eng 1993;40:1085–1094.
76. Effern A, Lehnertz K, Fernandez G, Grunwald T, David P, Elger CE. Single trial analysis of event related potentials: non-linear de-noising with wavelets. Clin Neurophysiol 2000a;111:2255–2263.
77. Effern A, Lehnertz K, Schreiber T, David P, Elger CE. Nonlinear denoising of transient signal with application to event related potentials. Physica D 2000b;140:257–266.
78. Pfurtscheller G, Cooper R. Selective averaging of the intracerebral click evoked responses in man: An improved method of measuring latencies and amplitudes. Electr Clin Neurophysiol 1975;38:187–190.
79. Woody CD. Characterization of an adaptive filter for the analysis of variable latency neuroelectric signals. Med Biol Eng 1967;5:539–553.
80. Quian Quiroga R, van Luijtelaar ELJM. Habituation and sensitization in rat auditory evoked potentials: a single-trial analysis with wavleet denoising. Int J Psychophysiol 2002;43:141–153.
81. Sambeth A, Maes JHR, Quian Quiroga R, Coenen AML. Effects of stimulus repetitions on the event-related potentials of humans and rats. Int J Psychophysiol 2004a;53:197–205.
82. Sambeth A, Maes JHR, Quian Quiroga R, van Rijn CM, Coenen AML. Enhanced re-habituation of the orienting response of the human event related potential. Neurosci Lett 2004b;356:103–106.
83. Talnov A, Quian Quiroga R, Meier M, Matsumoto G, Brankack J. Enthorinal imputs to dentate gyrus are activated mainly by conditioned events with long time intervals. Hippocampus 2003;13:755–765.
84. Atienza M, Cantero JL, Quian Quiroga R. Precise timing accounts for posttraining sleepdependent enhancements of the auditory mismatch negativity. Neuroimage 2005;26:628–634.

See also ELECTROENCEPHALOGRAPHY; ELECTROPHYSIOLOGY; SLEEP STUDIES, COMPUTER ANALYSIS OF.

EXERCISE FITNESS, BIOMECHANICS OF.

See BIOMECHANICS OF EXERCISE FITNESS.

EXERCISE, THERAPEUTIC.

See REHABILITATION AND MUSCLE TESTING.

EXERCISE STRESS TESTING

HENRY CHEN
Stanford University
Palo Alto, California
VICTOR FROEHLICHER
VA Medical Center
Palo Alto, California

INTRODUCTION

Exercise is the body's most common physiologic stress, and while it affects several systems, it places major demands on the cardiopulmonary system. Because of this interaction, exercise can be considered the most practical test of cardiac perfusion and function. Exercise testing is a noninvasive tool to evaluate the cardiovascular system's response to exercise under carefully controlled conditions.

THEORY

The major cardiopulmonary adaptations that are required during acute exercise make exercise testing a practical evaluation of cardiac perfusion and function. Exercise testing is not only useful in clinical evaluation of coronary status, but also serves as a valuable research tool in the study of cardiovascular disease, evaluating physical performance in athletes, and studying the normal and abnormal physiology of other organ systems.

A major increase and redistribution of cardiac output underlies a series of adjustments that allow the body to increase its resting metabolic rate as much as 10–20 times with exercise. The capacity of the body to deliver and utilize oxygen is expressed as the maximal oxygen uptake, which is defined as the product of maximal cardiac output and maximal arteriovenous oxygen difference. Thus, the cardiopulmonary limits are defined by (*1*) a central component (cardiac output) that describes the capacity of the heart to function as a pump, and (*2*) peripheral factors (arteriovenous oxygen difference) that describe the capacity of the lung to oxygenate the blood delivered to it as well as the capacity of the working muscle to extract this oxygen from the blood. Hemodynamic responses to exercise are greatly affected by several parameters, such as presence or absence of disease and type of exercise being performed, as well as age, gender, and fitness of the individual.

Coronary artery disease is characterized by reduced myocardial oxygen supply, which, in the presence of an increased myocardial oxygen demand, can result in myocardial ischemia and reduced cardiac performance. Despite years of study, a number of challenges remain regarding the response to exercise clinically. Although myocardial perfusion and function are intuitively linked, it is often difficult to separate the impact of ischemia from that of left ventricular dysfunction on exercise responses. Indexes of ventricular function and exercise capacity are poorly related. Cardiac output is considered the most important determinant of exercise capacity in normal subjects and in most patients with cardiovascular or pulmonary disease. However, among patients with disease, abnormalities in one or several of the links in the chain that defines oxygen uptake contribute to the determination of exercise capacity.

The transport of oxygen from the air to the mitochondria of the working muscle cell requires the coupling of blood flow and ventilation to cellular metabolism. Energy for muscular contraction is provided by three sources: stored phosphates [adenosine triphosphate (ATP) and creatine phosphate]; oxygen-independent glycolysis, and oxidative metabolism. Oxidative metabolism provides the greatest source of ATP for muscular contraction. Muscular contraction is accomplished by three fiber types that differ in their contraction speed, color, and mitochondrial content. The duration and intensity of activity determine the extent to which these fuel sources and fiber types are called upon.

All of the physiologic responses to exercise are mediated by the autonomic nervous system. As such, the exercise test and the response in recovery after exercise are increasingly recognized as important surrogate measures of autonomic function. The balance between sympathetic and parasympathetic influences to the cardiovascular sytem is critical, as they determine heart rate, blood pressure, cardiac output redistribution, and vascular resistance during exercise. Furthermore, indirect measures of autonomic function, such as heart rate variability and the rate in which heart rate recovers after exercise, are important prognostic markers in patients with cardiovascular disease (1).

There are several advantages to using the exercise test. These include the test's widespread availability, multiple uses, and high yield of clinically useful information. These factors make it an important "gatekeeper" for more expensive and invasive procedures. However, a major drawback has been the non-uniform application in clinical practices. To approach this problem, excellent guidelines have been developed based on expert consensus and research performed over the last 20 years. These include an update of the AHA/ACC guidelines on exercise testing, the American Thoracic Society/American College of Chest Physicians Statement on Cardiopulmonary Exercise Testing, an AHA Scientific Statement on Exercise and Heart Failure, new editions of the American Association of Cardiovascular and Pulmonary Rehabilitation Guidelines, and American College of Sports Medicine Guidelines on Exercise Testing and Prescription (cardiologyonline.com/guidelines.htm, cardiology.org). These have made substantial contributions to the understanding and greater uniformity of application of exercise testing.

EQUIPMENT

Current technology, although adding both convenience and sophistication, has raised new questions about methodology. For example, all commercially available systems today use computers. Do computer-averaged exercise electrocardiograms (ECGs) improve test performance, and what should the practitioner be cautious of when considering computer measurements? What about the many computer-generated exercise scores? When should ventilatory gas exchange responses be measured during testing and what special considerations are important when using them? The following text is intended to help answer such questions.

Blood Pressure Measurement

While there have been a number of automated devices developed for blood pressure measurement during exercise, our clinical impression is that they are not superior to manual testing. The time-proven method of the physician holding the patient's arm with a stethoscope placed over the brachial artery remains most reliable. An anesthesiologist's auscultatory piece or an electronic microphone can be fastened to the arm. A device that inflates and deflates the cuff with the push of a button can be helpful.

A dropping or a flat systolic response during exercise are ominous and can be the most important indicator of adverse events occurring during testing. If systolic blood pressure fails to increase or appears to be decreasing, it should be taken again immediately. The exercise test should be stopped if the systolic blood pressure drops by

10 mmHg (1.33 kPa) or more or falls below the value of the blood pressure taken in a standing position before testing. Additionally, if the systolic blood pressure exceeds 250 mmHg (33.33 kPa) or the diastolic blood pressure reaches 115 mmHg (15.33 kPa), the test should be stopped.

ECG Recording

Electrodes and Cables. Several disposable electrodes perform adequately for exercise testing. Disposable electrodes can be quickly applied and do not have to be cleaned for reuse. A disposable electrode that has an abrasive center spun by an applicator after the electrode is attached to the skin (Quickprep) is available from Quinton Instrument Co. This approach does not require skin preparation. The applicator of this system has a built-in impedance meter that stops the spinning when the skin impedance has been appropriately lowered.

Previously used buffer amplifiers and digitizers carried by the patient are no longer advantageous. Cables develop continuity problems with use and require replacement rather than repair. It is often found that replacement is necessary after roughly 500 tests. Some systems have used analog-to-digital converters in the electrode junction box carried by the patient. Because digital signals are relatively impervious to noise, the patient cable can be unshielded and is therefore very light.

Careful skin preparation and attention to the electrode–cable interface are important for a safe and successful exercise test and are necessary no matter how elaborate or expensive the ECG recording device used.

Lead Systems

Bipolar Lead Systems. Bipolar leads have been traditionally used owing to the relatively short time required for placement, the relative freedom from motion artifact, and the ease with which noise problems can be located. The usual positive reference is an electrode placed the same as the positive reference for V_5 (2). The negative reference for V_5 is Wilson's central terminal, which consists of connecting the limb electrodes to the right arm, left arm, and left leg. Virtually all current ECG systems, however, use the modified 12-lead system first described by Mason and Likar (3).

Mason–Likar Electrode Placement. Because a 12-lead ECG cannot be obtained accurately during exercise with electrodes placed on the wrists and ankles, the electrodes are placed at the base of the limbs for exercise testing. In addition to lessening noise for exercise testing, the Mason–LIkar modified placement has been demonstrated to exhibit no differences in ECG configuration when compared to the standard limb lead placement (3). However, this finding has been disputed by others who have found that the Mason–Likar placement causes amplitude changes and axis shifts when compared with standard placement. Because these could lead to diagnostic changes, it has been recommended that the modified exercise electrode placement not be used for recording a resting ECG. The preexercise ECG has been further complicated by the recommendation that it should be obtained standing, since that is the same position maintained during exercise. This situation is worsened by the common practice of moving the limb electrodes onto the chest to minimize motion artifact.

In the Mason–Likar torso-mounted limb lead system, the conventional ankle and wrist electrodes are replaced by electrodes mounted on the torso at the base of the limbs. This placement avoids artifact caused by movement of the limbs. The standard precordial leads use Wilson's central terminal as their negative reference, which is formed by connecting the right arm, left arm, and left leg. This triangular configuration around the heart results in a zero voltage reference through the cardiac cycle. The use of Wilson's central terminal for the precordial leads (V leads) requires the negative reference to be a combination of three additional electrodes rather than the single electrode used as the negative reference for bipolar leads.

The modified exercise electrode placement should not be used for routine resting ECGs. However, the changes caused by the exercise electrode placement can be kept to a minimum by keeping the arm electrodes off the chest and putting them on the anterior deltoid and by having the patient supine. In this situation, the modified exercise limb lead placement of Mason–Likar can serve well as the resting ECG reference before an exercise test.

For exercise testing, limb electrodes should be placed as far from the heart as possible, but not on the limbs; the ground electrode (right leg) can be on the back out of the cardiac field and the left leg electrode should be below the umbilicus. The precordial electrodes should be placed in their respective interspaces.

Inferior Lead ST-Segment Depression

One potential area of confusion in interpretation lies in ST-segment depression in the inferior leads. Miranda et al. (4) studied 178 men who had undergone exercise testing and coronary angiography to evaluate the diagnostic value of ST-segment depression occurring in the inferior leads. The area under the curve in lead II was not significantly > 0.50, suggesting that for the identification of coronary artery disease, isolated ST-segment depression in lead II appears unreliable. The ST depression occurring in the inferior leads alone (II, AVF) can sometimes represent a false positive response. ST elevation in these leads, however, suggests transmural ischemia in the area of RCA blood distribution.

Number of Leads to Record

In patients with normal resting ECGs, a V_5 or similar bipolar lead along the long axis of the heart is usually adequate. The ECG evidence of myocardial infarction or history suggestive of coronary spasm necessitate use of additional leads. As a minimal overall approach, it is advisable to record three leads: a V_5-type of lead, an anterior V_2-type of lead, and an inferior lead, such as a V_F. Alternatively, Frank *X*, *Y*, and *Z* leads may be used. Either of these approaches can be helpful additionally for the detection and identification of dysrhythmias. It is also advisable to record a second three-lead grouping consisting of V_4, V_5, and V_6. Occassionally abnormalities seen as borderline in V_5 can be better defined in neighboring V_4 or V_6.

Because most meaningful ST-segment depression occurs in the lateral leads (V_4, V_5, and V_6) when the resting ECG is normal, other leads are only necessary in patients who had a myocardial infarction, those with a history consistent with coronary spasm or variant angina, or those who have exercise- induced dysrhythmias of an uncertain type.

Although as much as 90% of abnormal ST depression occurs in V_5 (or the two adjacent precordial leads), other leads should also be used, particularly in patients with history of known myocardial infarction or variant angina, and especially since ST elevation can localize ischemia to the area beneath the electrodes, and multiple-vector leads can be useful for studying arrhythmias (best diagnosed with inferior and anterior leads where the P waves are best seen).

ECG Recording Instruments

There have been several advances in ECG recorder technology. The medical instrumentation industry has promptly complied with specifications set forth by various professional groups. Machines with a high input impedance ensure that the voltage recorded graphically is equivalent to that on the surface of the body despite the high natural impedance of the skin. Optically isolated buffer amplifiers have ensured patient safety, and machines with a frequency response up to 100 Hz are commercially available. The lower end is possible because direct current (dc) coupling is technically feasible.

Waveform Processing

Analog and digital averaging techniques have made it possible to filter and average ECG signals to remove noise. There is a need for consumer awareness in these areas because most manufacturers do not specify how the use of such procedures modifies the ECG. Both filtering and signal averaging can, in fact, distort the ECG signal. Averaging techniques are nevertheless attractive because they can produce a clean tracing when the raw data is noisy. However, the clean-looking ECG signal produced may not be a true representation of the waveform and in fact may be dangerously misleading. Also, the instruments that make computer ST-segment measurements are not entirely reliable because they are based on imperfect algorithms. While useful in reducing noise, filtering and averaging can cause false ST depression due to distortion of the raw data.

Computerization

There are a host of advantages of digital over analog data processing. These include more accurate measurements, less distortion in recording, and direct accessibility to digital computer analysis and storage techniques. Other advantages are rapid mathematical manipulation (averaging), avoidance of the drift inherent in analog components, digital algorithm control permitting changes in analysis schema with ease (software rather than hardware changes), and no degradation with repetitive playback. When outputting data, digital processing also offers higher plotting resolution and easy repetitive manipulation.

Computerization also helps meet the two critical needs of exercise ECG testing: the reduction of the amount of ECG data collected during testing and the elimination of electrical noise and movement artifact associated with exercise. Because an exercise test can exceed 30 min (including data acquisition during rest and recovery) and many physicians want to analyze all 12 leads during and after testing, the resulting quantity of ECG data and measurements can quickly become excessive. The three-lead vectorcardiographic [or three-dimensional (3D) (i.e., aV_F, V_2, V_5)] approach would reduce the amount of data; however, clinicians favor the l2-lead ECG. The exercise ECG often includes random and periodic noise of high and low frequency that can be caused by respiration, muscle artifact, electrical interference, wire continuity, and electrode–skin contact problems. In addition to reducing noise and facilitating data collection, computer processing techniques may also make precise and accurate measurements, separate and capture dysrhythmic beats, perform spatial analysis, and apply optimal diagnostic criteria for ischemia.

Although most clinicians agree that computerized analysis simplifies the evaluation of the exercise ECG, there has been disagreement about whether accuracy is enhanced (5). A comparison of computerized resting ECG analysis programs with each other and with the analyses of expert readers led to the conclusion that physician review of any reading is necessary (6). Although computers can record very clean representative ECG complexes and neatly print a wide variety of measurements, the algorithms they use are far from perfect and can result in serious differences from the raw signal. The physician who uses commercially available computer-aided systems to analyze the results of exercise tests should be aware of the problems and always review the raw analog recordings to see whether they are consistent with the processed output.

Even if computerization of the original raw analog ECG data could be accomplished without distortion, the problem of interpretation still remains. Numerous algorithms have been recommended for obtaining the optimal diagnostic value from the exercise ECG. These algorithms have been shown to provide improved sensitivity and specificity compared with standard visual interpretation. Often, however, this improvement has been demonstrated only by the investigator who proposed the new measurement. Furthermore, the ST measurements made by a computer can be erroneous. It is advisable to have the devices mark both the isoelectric level and the point of ST0. Even if the latter is chosen correctly, misplacement of the isoelectric line outside of the PR segment can result in incorrect ST level measurements. Computerized ST measurements require physician over reading; errors can be made both in the choice of the isoelectric line and the beginning of the ST segment.

Causes of Noise

Many of the causes of noise in the exercise ECG signal cannot be corrected, even by meticulous skin preparation. Noise is defined in this context as any electrical signal that

is foreign to or distorts the true ECG waveform. Based on this definition, the types of noise that may be present can be caused by any combination of line-frequency (60 Hz), muscle, motion, respiration, contact, or continuity artifact. Line-frequency noise is generated by the interference of the 60 Hz electrical energy with the ECG. This noise can be reduced by using shielded patient cables. If in spite of these precautions this noise is still present, the simplest way to remove it is to design a 60-Hz notch filter and apply it in series with the ECG amplifier. A notch filter removes only the line frequency; that is, it attenuates all frequencies in a narrow band $\sim$ 60 Hz. This noise can also be removed by attenuating all frequencies $>$ 59 Hz; however, this method of removing line-frequency noise is not recommended because it causes waveform distortion and results in a system that does not meet AHA specifications. The most obvious manifestation of distortion caused by such filters is a decrease in R-wave amplitude; therefore a true notch filter is advisable.

Muscle noise is generated by the activation of muscle groups and is usually of high frequency. This noise, with other types of high frequency noise, can be reduced by signal averaging. Motion noise, another form of high frequency noise, is caused by the movement of skin and the electrodes, which causes a change in the contact resistance. Respiration causes an undulation of the waveform amplitude, so the baseline varies with the respiratory cycle. Baseline wander can be reduced by low-frequency filtering; however, this results in distortion of the ST segment and can cause artifactual ST-segment depression and slope changes. Other baseline removal approaches have been used, including linear interpolation between isoelectric regions, high order polynomial estimates, and cubic-spline techniques, which can each smooth the baseline to various degrees.

Contact noise appears as low frequency noise or sometimes as step discontinuity baseline drift. It can be caused by poor skin preparation resulting in high skin impedance or by air bubble entrapment in the electrode gel. It is reduced by meticulous skin preparation and by rejecting beats that show large baseline drift. Also, by using the median rather than the mean for signal averaging, this type of drift can be reduced. Continuity noise caused by intermittent breaks in the cables is rarely a problem because of technological advances in cable construction, except, of course, when cables are abused or overused.

Most of the sources of noise can be effectively reduced by beat averaging. However, two types of artifact can actually be caused in the signal-averaging process by the introduction of beats that are morphologically different from others in the average and the misalignment of beats during averaging. As the number of beats included in the average increases, the level of noise reduction is greater. The averaging time and the number of beats to be included in the average have to be compromised, though, because the morphology of ECG waveforms changes over time.

For exercise testing, the raw ECG data should be considered first, and then the averages and filtered data may be used to aid interpretation if no distortion is obvious.

ECG Paper Recording

For some patients, it is advantageous to have a recorder with a slow paper speed option of $5\,\mathrm{mm}\cdot\mathrm{s}^{-1}$. This speed is optimal for recording an entire exercise test and reduces the likelihood of missing any dysrhythmias when specifically evaluating such patients. Some exercise systems allow for a total disclosure print out option similar to that provided and many holter systems. In rare instances, a faster paper speed of $50\,\mathrm{mm}\cdot\mathrm{s}^{-1}$ can be helpful for making particular evaluations, such as accurate ST-segment slope measurements.

Thermal head printers have effectively become the industry standard. These recorders are remarkable in that they can use blank thermal paper and write out the grid and ECG, vector loops, and alpha-numerics. They can record graphs, figures, tables, and typed reports. They are digitally driven and can produce very high resolution records. The paper price is comparable with that of other paper, and these devices have a reasonable cost and are very durable, particularly because a stylus is not needed.

Z-fold paper has the advantage over roll paper in that it is easily folded, and the study can be read in a manner similar to paging through a book. Exercise ECGs can be microfilmed on rolls, cartridges, or fiche cards for storage. They can also be stored in digital or analog format on magnetic media or optical disks. The latest technology involves magnetic optical disks that are erasable and have fast access and transfer times. These devices can be easily interfaced with microcomputers and can store megabytes of digital information. Lasers or ink jet printers have a delay making them unsuitable for medical emergencies but offer the advantage of the inexpensiveness of standard paper and long lived images.

Many available recording systems have both thermal head and laser or inkjet printers and use the cheaper, slower printers for final reports and summaries while the thermal head printers are used for live ECG tracings (i.e., real time). The standard 3 lead $\times$ 4 lead groups print out leaves only 2.5 s to assess ST changes or arrhythmias.

Exercise Test Modalities

Types of Exercise. Two types of exercise can be used to stress the cardiovascular system: isometric and dynamic, though most activities are a combination of the two. Isometric exercise, which involves constant muscular contraction with minimal movement (e.g., a handgrip), imposes a disproportionate pressure load on the left ventricle relative to the body's ability to supply oxygen. Dynamic exercise, or rhythmic muscular activity resulting in movement, initiates a more appropriate balance between cardiac output, blood supply, and oxygen exchange. Because a delivered workload can be accurately calibrated and the physiological response easily measured, dynamic exercise is preferred for clinical testing. Dynamic exercise is also superior because it can be more easily graduated and controlled. Using gradual, progressive workloads, patients with coronary artery disease can be protected from rapidly increasing myocardial oxygen demand. Although bicycling is also a dynamic exercise, most individuals perform more work on a treadmill because of greater muscle mass

involved and generally more familiarity with walking than cycling.

Numerous modalities have been used to provide dynamic exercise for exercise testing, including steps, escalators, ladder mills, and arm ergometers. Today, however, the bicycle ergometer and the treadmill are the most commonly used dynamic exercise devices. The bicycle ergometer is usually cheaper, takes up less space, and makes less noise. Upper body motion is usually reduced, but care must be taken so that isometric exercise is not performed by the arms. The workload administered by the simple, manually braked cycle ergometers is not well calibrated and depends on pedaling speed. It can be easy for a patient to slow pedaling speed during exercise testing and decrease the administered workload, making the estimation of exercise capacity unreliable. More expensive electronically braked bicycle ergometers keep the workload at a specified level over a wide range of pedaling speeds, and have become the standard for cycle ergometer testing today.

Dynamic exercise, using a treadmill or a cycle ergometer, is a better measure of cardiovascular function and better method of testing than isometric exercise.

Arm Ergometry. Alternative methods of exercise testing are needed for patients with vascular, orthopedic, or neurological conditions who cannot perform leg exercise. Arm ergometry can be used in such patients (7). However, nonexercise techniques (such as pharmacologic stress testing) are currently more popular for these patients.

Bicycle Ergometer. The bicycle ergometer is usually cheaper, takes up less space, and makes less noise than a treadmill. Upper body motion is usually reduced, but care must be taken so that the arms do not perform isometric exercise. The workload administered by the simple bicycle ergometers is not well calibrated and is dependent on pedaling speed. It can be easy for a patient to slow pedaling speed during exercise testing and decrease the administered workload. More modern electronically braked bicycle ergometers keep the workload at a specified level over a wide range of pedaling speeds and are recommended. Since supine bicycle exercise is so rarely used, we will not address it here except to say that maximal responses are usually lower than with the erect position.

Treadmill. Treadmills should have front and side rails to allow patients to steady themselves. Some patients may benefit from the help of the person administering the test. Patients should not grasp the front or side rails because this decreases the work performed and the oxygen uptake and, in turn, increases exercise time, resulting in an overestimation of exercise capacity. Gripping the handrails also increases ECG muscle artifact. When patients have difficulty maintaining balance while walking, it helps to have them take their hands off the rails, close their fists, and extend one finger to touch the rails after they are accustomed to the treadmill. Some patients may require a few moments to feel comfortable enough to let go of the handrails, but grasping the handrails after the first minute of exercise should be strongly discouraged.

Table 1. Two Basic Principles of Exercise Physiology

Myocardial oxygen consumption	$\approx$ Heart rate $\times$ systolic blood pressure (determinants include wall tension $\cong$ left ventricular pressure $\times$ volume; contractility; and heart rate)
Ventilatory oxygen consumption (VO_2)	$\approx$ External work performed, or cardiac output $\times$ a-VO_2 difference[a]

[a]The arteriovenous O_2 difference is $\sim$ 15–17 vol% at maximal exercise in most individuals; therefore, the VO_2 max generally reflects the extent to which cardiac output increases.

A small platform or stepping area at the level of the belt is advisable so that the patient can start the test by pedaling the belt with one foot before stepping on. The treadmill should be calibrated at least monthly. Some models can be greatly affected by the weight of the patient and will not deliver the appropriate workload to heavy patients. An emergency stop button should be readily available to the staff only.

Bicycle Ergometer versus Treadmill

Bicycle ergometry has been found to elicit similar maximum heart rate values to treadmill exercise in most studies comparing the methods. However, maximum oxygen uptake has been 6–25% greater during treadmill exercise (8). Some studies have reported similar ECG changes with treadmill testing as compared with bicycle testing (9), whereas others have reported more significant ischemic changes with treadmill testing (10). Nonetheless, the treadmill is the most commonly used dynamic testing modality in the United States, and the treadmill may be advantageous because patients are more familiar with walking than they are with bicycling. Patients are more likely to give the muscular effort necessary to adequately increase myocardial oxygen demand by walking than by bicycling.

Treadmills usually result in a higher MET values, but maximal heart rate is usually the same as with a bike. Thus, bike testing can result in a lower prognosis estimate but has similar ability to predict ischemic disease.

Exercise Protocols

The many different exercise protocols in use have led to some confusion regarding how physicians compare tests between patients and serial tests in the same patient. A recent survey performed among VA exercise laboratories confirmed that the Bruce protocol remains the most commonly used; 83% of laboratories reported using the Bruce test for routine testing. This protocol uses relatively large and unequal increments in work (2–3 MET) every 3 min. Large and uneven work increments, such as these have been shown to result in a tendency to overestimate exercise capacity and the lack of uniform increases in work rate, can complicate the interpretation of some ST segment measurements and ventilatory gas exchange responses (11,12). (Table 1). Thus, exercise testing guidelines have recommended protocols with smaller and more equal increments. It is also important to individualize the test to target duration in the range of 8–12 min.

Ramp Testing

An approach to exercise testing that has gained interest in recent years is the ramp protocol, in which work increases constantly and continuously.

Questionnaires

The key to appropriately targeting a ramp is accurately predicting the individual's maximal work capacity. If a previous test is not available, a pretest estimation of an individual's exercise capacity is quite helpful to set the appropriate ramp rate. Functional classifications are too limited and poorly reproducible. One problem is that usual activities can decrease, so an individual can become greatly limited without having a change in functional class. A better approach is to use the specific activity scale of Goldman et al. (13) (Table 2), the DASI (Table 3), or the VSAQ (Table 4). Alternatively, the patient may be questioned regarding usual activities that have a known MET cost (Table 5) (14).

Borg Scale

Instead of simply tracking heart rate to clinically determine the intensity of exercise, it is preferable to use the 6–20 Borg scale or the nonlinear 1–10 scale of perceived exertion (Table 6) (15). The Borg scale is a simple, valuable way of assessing the relative effort a patient exerts during exercise.

Table 2. Specific Activity Scale of Goldman

Class I ($\geq$ 7 METs)	A patient can perform any of the following activities: Carrying 24 lb (10.88 kg) up eight steps Carrying an 80 lb (16.28 kg) object Shoveling snow Skiing Playing basketball, touch football, squash, or handball Jogging/walking 5 mph
Class II ($\geq$ 5 METs)	A patient does not meet Class I criteria, but can perform any of the following activities to completion without stopping: Carrying anything up eight steps Having sexual intercourse Gardening, raking, weeding Walking 4 mph
Class III ($\geq$ 2 METs)	A patient does not meet Class I or Class II criteria but can perform any of the following activities to completion without stopping: Walking down eight steps Taking a shower Changing bedsheets Mopping floors, cleaning windows Walking 2.5 mph Pushing a power lawnmower Bowling Dressing without stopping
Class IV ($\leq$ 2 METs)	None of the above

Table 3. Duke Activity Scale Index[a]

Activity	Weight
Can you?	
1. Take care of yourself, that is, eating, dressing, bathing, and using the toilet?	2.75
2. Walk indoors, such as around your house?	1.75
3. Walk a block or two on level ground?	2.75
4. Climb a flight of stairs or walk up a hill?	5.50
5. Run a short distance?	8.00
6. Do light work around the house like dusting or washing dishes?	2.7
7. Do moderate work around the house like vacuuming, sweeping floors, or carrying in groceries?	3.50
8. Do heavy work around the house like scrubbing floors or lifting and moving heavy furniture?	8.00
9. Do yard work like raking leaves, weeding, or pushing a power mower?	4.5
10. Have sexual relations?	5.25
11. Participate in moderate recreational activities like golf, bowling, dancing, doubles tennis, or throwing a basketball or football?	6.00
12. Participate in strenuous sports like swimming, singles tennis, football, basketball, or skiing?	7.50

[a]Duke activity scale index = DASI = sum of weights for "yes" replies. $VO_2 = 0.43 \times \text{DASI} + 9.6$.

Postexercise Period

The patient should assume a supine position in the postexercise period to achieve greatest sensitivity in exercise testing. It is advisable to record $\sim$ 10 s of ECG data while the patient is motionless, but still experiencing near-maximal heart rate before having the patient lie down. Some patients must be allowed to lie down immediately to avoid hypotension. Letting the patient have a cool-down walk after the test is discouraged, as it can delay or eliminate the appearance of ST-segment depression (16). According to the law of La Place, the increase in venous return and thus ventricular volume in the supine position increases myocardial oxygen demand. Data from our laboratory (17) suggests that having patients lie down may enhance ST-segment abnormalities in recovery. However, a cool-down walk has been suggested to minimize the postexercise chances of dysrrhythmic events in this high risk time when catecholamine levels are high. The supine position after exercise is not as important when the test is not being performed for diagnostic purposes, for example, fitness testing. When testing is not performed for diagnostic purposes, it may be preferable to walk slowly (1.0–1.5 mph) or continue cycling against zero or minimal resistance (up to 25 W when testing with a cycle ergometer) for several minutes after the test.

Monitoring should continue for at least 6–8 min after exercise or until changes stabilize. An abnormal response occurring only in the recovery period is not unusual. Such responses are not false positives. Experiments confirm mechanical dysfunction and electrophysiological abnormalities in the ischemic ventricle after exercise. A cool down walk can be helpful when testing patients with an

Table 4. Veterans Specific Activity Questionnaire[a,b]

1 MET:	Eating; getting dressed; working at a desk
2 METs:	Taking a shower; shopping; cooking; walking down eight steps
3 METs:	Walking slowly on a flat surface for one or two blocks; doing moderate amounts of work around the house like vacuuming, sweeping the floors, or carrying in groceries
4 METs:	Doing light yard work, i.e., raking leaves, weeding, sweeping, or pushing a power mower; painting; light carpentry
5 METs:	Walking briskly; social dancing; washing the car
6 METs:	Playing nine holes of golf, carrying your own clubs; heavy carpentry; mowing lawn with a push mower
7 METs:	Carrying 60 lb; performing heavy outdoor work, that is, digging, spading soil; walking uphill
8 METs:	Carrying groceries upstairs; moving heavy furniture; jogging slowly on flat surface; climbing stairs quickly
9 METs:	Bicycling at a moderate pace; sawing wood; jumping rope (slowly)
10 METs:	Briskly swimming; bicycling up a hill; jogging 6 mph
11 METs:	Carrying a heavy load (i.e., a child or firewood) up two flights of stairs; cross-country skiing; bicycling briskly and continuously
12 METs:	Running briskly and continuously (level ground, 8 mph)
13 METs:	Performing any competitive activity, including those that involve intermittent sprinting; running competitively; rowing competitively; bicycle racing.

[a]Veterans Specific Activity Questionnaire = VSAQ

[b]Before beginning your treadmill test today, we need to estimate what your usual limits are during daily activities. Following is a list of activities that increase in difficulty as you read down the page. Think carefully, then underline the first activity that, if you performed it for a period of time, would typically cause fatigue, shortness of breath, chest discomfort, or otherwise cause you to want to stop. If you do not normally perform a particular activity, try to imagine what it would be like if you did.

established diagnosis undergoing testing for other than diagnostic reasons, when testing athletes, or when testing patients with dangerous dysrhythmias.

The recovery period is extremely important for observing ST shifts and should not be interrupted by a cool down walk or failure to monitor for at least 5 min. Changes isolated to the recovery period are not more likely to be false positives.

The recovery period, particularly between the second and fourth minute, are critical for ST analysis. Noise should not be a problem and ST depression at that time has important implications regarding the presence and severity of coronary artery disease (CAD).

Additional Techniques

Several ancillary imaging techniques have been shown to provide a valuable complement to exercise electrocardiography for the evaluation of patients with known or suspected CAD. They can localize ischemia and thus guide interventions. These techniques are particularly helpful among patients with equivocal exercise electrocardiograms or those likely to exhibit false-positive or false-negative responses. The guidelines call for their use when testing patients with more than 1.0 mm of ST depression at rest, LBBB, WPW, and paced rhythms. They are frequently used to clarify abnormal ST segment responses in asymptomatic people or those in whom the cause of chest discomfort remains uncertain, often avoiding angiography. When exercise electrocardiography and an imaging technique are combined, the diagnostic and prognostic accuracy is enhanced. The major imaging procedures are myocardial perfusion and ventricular function studies using radionuclide techniques, and exercise echocardiography. Some of the newer add-ons or substitutes for the exercise test have the advantage of being able to localize ischemia as well as diagnose coronary disease when the baseline ECG exhibits the above-mentioned abnormalities, which negate the usefulness of ST analysis. While the newer technologies are often suggested to have better diagnostic characteristics, this is not always the case particularly when more than the ST segments from the exercise test are used in scores. Pharmacologic stress testing is used in place of the standard exercise test for patients unable to walk or cycle or unable to give a good effort. These nonexercise stress techniques (persantine or adenosine with nuclear perfusion, dobutamine or arbutamine with echocardiography) permit diagnostic assessment of patients unable to exercise.

The ancillary imaging techniques are indicated when the ECG exhibits more than a millimeter of ST depression at rest, LBBB, WPW, and paced rhythms or when localization of ischemia is important.

Ventilatory Gas Exchange Responses

Because of the inaccuracies associated with estimating METs (ventilatory oxygen uptake) from workload (i.e., treadmill speed and grade), it can be important for many patients to measure physiologic work directly using ventilatory gas exchange responses, commonly referred to as cardiopulmonary exercise testing. Although this requires metabolic equipment, a facemask or mouthpiece and other equipment, advances in technology have made these measurements widely available. Cardiopulmonary exercise testing adds precision to the measurement of work and also permits the assessment of other parameters, including the respiratory exchange ratio, efficiency of ventilation, and the ventilatory anaerobic threshold. The latter measurement is helpful because it usually represents a comfortable sub maximal exercise limit and can be used for setting an optimal exercise prescription or an upper limit for daily activities. Clinically, this technology is often used to more precisely evaluate therapy, for the assessment of disability, and to help determine whether the heart or lungs limit exercise. Computerization of equipment has also led to the widespread use of cardiopulmonary exercise testing in sports medicine. Gas exchange measurements

Table 5. MET Demands for Common Daily Activities[a]

Activity	METs
Mild	
Baking	2.0
Billiards	2.4
Bookbinding	2.2
Canoeing (leisurely)	2.5
Conducting an orchestra	2.2
Dancing, ballroom (slow)	2.9
Golfing (with cart)	2.5
Horseback riding (walking)	2.3
Playing a musical instrument	2.0
Volleyball (noncompetitive)	2.9
Walking (2 mph)	2.5
Writing	1.7
Moderate	
Calisthenics (no weights)	4.0
Croquet	3.0
Cycling (leisurely)	3.5
Gardening (no lifting)	4.4
Golfing (without cart)	4.9
Mowing lawn (power mower)	3.0
Playing drums	3.8
Sailing	3.0
Swimming (slowly)	4.5
Walking (3 mph)	3.3
Walking (4 mph)	4.5
Vigorous	
Badminton	5.5
Chopping wood	4.9
Climbing hills	7.0
Cycling (moderate)	5.7
Dancing	6.0
Field hockey	7.7
Ice skating	5.5
Jogging (10 min mile)	10.0
Karate or judo	6.5
Roller skating	6.5
Rope skipping	12.0
Skiing (water or downhill)	6.8
Squash	12.0
Surfing	6.0
Swimming (fast)	7.0
Tennis (doubles)	6.0

[a]These activities can often be done at variable intensities if one assumes that the intensity is not excessive and that the courses are flat (no hills) unless so specified.

Table 6. Borg Scales of Perceived Exertion[a]

Borg 20-Point Scale of Perceived Exertion

6	
7	Very, very light
8	
9	Very light
10	
11	Fairly light
12	
13	Somewhat hard
14	
15	Hard
16	
17	Very hard
18	
19	Very, very hard
20	

Borg Nonlinear 10-Point Scale of Perceived Exertion

0	Nothing at all
0.5	Extremely light (just noticeable)
1	Very light
2	Light (Weak)
3	Moderate
4	Somewhat heavy
5	Heavy (Strong)
6	
7	Very heavy
8	
9	
10	Extremely heavy (almost maximal)
•	Maximal

[a]Top: Borg 20-point scale; bottom: Borg nonlinear 10-point scale.

can supplement the exercise test by increasing precision and providing additional information concerning cardiopulmonary function during exercise. It is particularly needed to evaluate therapies using serial tests, since workload changes and estimated METs can be misleading. Because of their application for assessing prognosis in patients with heart failure, their use has become a standard part of the work-up for these patients.

Nuclear Techniques

Nuclear Ventricular Function Assessment. One of the first imaging modalities added to exercise testing was radionuclear ventriculography (RNV). This involved the intravenous injection of technetium tagged red blood cells. Using ECG gating of images obtained from a scintillation camera, images of the blood circulating within the LV chamber could be obtained. While regurgitant blood flow from valvular lesions could not be identified, ejection fraction and ventricular volumes could be estimated. The resting values could be compared to those obtained during supine exercise and criteria were established for abnormal. The most common criteria involved a drop in ejection fraction. This procedure is now rarely performed because its test characteristics have not fulfilled their promise.

Nuclear Perfusion Imaging. After initial popularity, the blood volume techniques have become surpassed by nuclear perfusion techniques. The first agent used was thallium, an isotopic analog of potassium that is taken up at variable rates by metabolically active tissue. When taken up at rest, images of metabolically active muscle such as the heart are possible. With the nuclear camera placed over the heart after intravenous injection of this isotope, images were initially viewed using X-ray film. The normal complete donut shaped images gathered in multiple views would be broken by cold spots where scar was present. Defects viewed after exercise could be due to either scar or ischemia. Follow up imaging confirmed that the cold spots were due to ischemia if they filled in later. As computer imaging techniques were developed, 3D imaging (SPECT) and subtle differences could be plotted and scored. In recent years, ventriculograms based on the imaged wall as apposed to the blood in the chambers (as with RNV) could be constructed.

Because of the technical limitations of thallium (i.e., source and lifespan), it has largely been replaced by chemical compounds called isonitriles which could be tagged with technetium, which has many practical advantages over thallium as an imaging agent. The isonitriles are trapped in the microcirculation permitting imaging of the heart with a scintillation camera. Rather than a single injection as for thallium, these compounds require an injection at maximal exercise then later in recovery. The differences in technology over the years and the differences in expertise and software at different facilities can complicate the comparisons of the results and actual application of this technology. The ventriculograms obtained with gated perfusion scans do not permit the assessment of valvular lesions, or as accurate an assessment of wall motion abnormalities or ejection fraction as echocardiography.

Nuclear perfusion scans can now permit an estimation of ventricular function and wall motion abnormalities.

Echocardiography

Echocardiography has made a significant and impressive impact on the field of cardiology. This imaging technique comes second only to contrast ventriculography via cardiac catheterization for measuring ventricular volumes, wall motion, and ejection fraction. With Doppler added, regurgitant flows can be estimated as well. With such information available, this imaging modality was quickly added by echocardiographers to exercise testing. Most studies showed that supine, posttreadmill assessments were adequate and the more difficult imaging during exercise was not necessary. The patient must be placed supine as soon as possible after treadmill or bicycle exercise and imaging begun. A problem can occur when the imaging requires removal or displacement of the important V_5 electrode where as much as 90% of the important ST changes are observed.

Biomarkers

The latest ancillary measures added to exercise testing in an attempt to improve diagnostic accuracy are biomarkers. The first and most logical biomarker evaluated to detect ischemia brought out by exercise was troponin. Unfortunately, it has been shown that even in patients who develop ischemia during exercise testing, serum elevations in cardiac specific troponin do not occur, demonstrating that myocardial damage does not occur (18,19). B-type natriuretic peptide (BNP) is a hormone produced by the heart that is released by both myocardial stretching and by myocardial hypoxia. Armed with this knowledge, investigators have reported several studies suggesting improvement in exercise test characteristics with BNP and its isomers (20,21). BNP is also used to assess the presence severity of (CHF) coronary heart failure, and has been shown to be a powerful prognostic marker (22,23). The point of contact analysis techniques available for these assays involves a hand held battery powered unit that uses a replaceable cartridge. Finger stick blood samples are adequate for these analyses and the results are available immediately. If validated using appropriate study design (similar to QUEXTA), biomarker measurements could greatly improve the diagnostic characteristics of the standard office–clinic exercise test.

In summary, use of proper methodology is critical for patient safety and accurate results. Preparing the patient physically and emotionally for testing is necessary. Good skin preparation will cause some discomfort but is necessary for providing good conductance and for avoiding artifact. The use of specific criteria for exclusion and termination, physician interaction with the patient, and appropriate emergency equipment are essential. A brief physical examination is always necessary to rule out important obstructive cardiomyopathy and aortic valve disease. Pretest standard 12-lead ECGs are needed in the supine and standing positions. The changes caused by exercise electrode placement can be kept to a minimum by keeping the arm electrodes off the chest, placing them on the shoulders, placing the leg electrodes below the umbilicus, and recording the baseline ECG supine. In this situation, the Mason–Likar modified exercise limb lead placement, if recorded supine, can serve as the resting ECG reference before an exercise test.

Few studies have correctly evaluated the relative yield or sensitivity and specificity of different electrode placements for exercise-induced ST-segment shifts. Using other leads in addition to V_5 will increase the sensitivity; however, the specificity is decreased. The ST-segment changes isolated to the inferior leads can on occasion be false-positive responses. For clinical purposes, vectorcardiographic and body surface mapping lead systems do not appear to offer any advantage over simpler approaches.

The exercise protocol should be progressive with even increments in speed and grade whenever possible. Smaller, even, and more frequent work increments are preferable to larger, uneven, and less frequent increases, because the former yield a more accurate estimation of exercise capacity. The value of individualizing the exercise protocol rather than using the same protocol for every patient has been emphasized by many investigators. The optimum test duration is from 8 to 12 min; therefore the protocol workloads should be adjusted to permit this duration. Because ramp testing uses small and even increments, it permits a more accurate estimation of exercise capacity and can be individualized to yield targeted test duration. An increasing number of equipment companies manufacture a controller that performs such tests using a treadmill.

Target heart rates based on age is inferior because the relationship between maximum heart rate and age is poor and scatters widely around many different recommended regression lines. Such heart rate targets result in a submaximal test for some individuals, a maximal test for others, and an unrealistic goal for some patients. Blood pressure should be measured with a standard stethoscope and sphygmomanometer; the available automated devices cannot be relied on, particularly for detection of exertional hypotension. Borg scales are an excellent means of quantifying an individual's effort. Exercise capacity should not be reported in total time but rather as the oxygen uptake or MET equivalent of the workload achieved. This method permits the comparison of the results of many different exercise testing protocols. Hyperventilation should be

avoided before testing. Subjects with and without disease may exhibit ST-segment changes with hyperventilation; thus, hyperventilation to identify false-positive responders is no longer considered useful by most researchers. The postexercise period is a critical period diagnostically; therefore the patient should be placed in the supine position immediately after testing.

EVALUATION

Hemodynamics involves studying the body's adaptations to exercise, commonly evaluated in heart rate and blood pressure. However, it also includes changes in cardiac output and its determinants, as well as the influence of cardiovascular disease on cardiac output. Exercise capacity is an important clinical measurement and is also influenced strongly by exercise hemodynamics. There are several factors that affect exercise capacity, and there is an important issue of how normal standards for exercise capacity are expressed.

When interpreting the exercise test, it is important to consider each of its responses separately. Each type of response has a different impact on making a diagnostic or prognostic decision and must be considered along with an individual patient's clinical information. A test should not be called abnormal (or positive) or normal (or negative), but rather the interpretation should specify which responses were abnormal or normal, and each particular response should be recorded. The final report should be directed to the physician who ordered the test and who will receive the report. It should contain clear information that helps in patient management rather than vague "medspeak". Interpretation of the test is highly dependent upon the application for which the test is used and on the population tested.

Predicting Severe Angiographic Disease

Exercise Test Responses. Studies have long attempted to assess for disease in the left main coronary artery using exercise testing (24–26). Different criteria have been used with varying results. Predictive value here refers to the percentage of those with the abnormal criteria that actually had left main disease. Naturally, most of the false positives actually had coronary artery disease, but less severe forms. Sensitivity here refers to the percentage of those with left main disease only that are detected. These criteria have been refined over time and the last study by Weiner using the CASS data deserves further mention (27). A markedly positive exercise test (Table 7) defined as 0.2 mV or more of downsloping ST-segment depression beginning at 4 METs, persisting for at least six minutes into recovery, and involving at least five ECG leads had the greatest sensitivity (74%) and predictive value (32%) for left main coronary disease. This abnormal pattern identified either left main or three-vessel disease with a sensitivity of 49%, a specificity of 92% and a predictive value of 74%.

Table 7. Markedly Positive Treadmill Test Responses

Markedly Positive Treadmill Test Responses
More than 0.2 mV downsloping ST-segment depression
Involving five or more leads
Occurring at < 5 METs
Prolonged ST depression late into recovery

It appears that individual clinical or exercise test variables are unable to detect left main coronary disease because of their low sensitivity or predictive value. However, a combination of the amount, pattern, and duration of ST-segment response was highly predictive and reasonably sensitive for left main or three-vessel coronary disease. The question still remains of how to identify those with abnormal resting ejection fractions, those that will benefit the most with prolonged survival after coronary artery bypass surgery. Perhaps those with a normal resting ECG will not need surgery for increased longevity because of the associated high probability of normal ventricular function.

Blumenthal et al. (28) validated the ability of a strongly positive exercise test to predict left main coronary disease even in patients with minimal or no angina. The criteria for a markedly positive test included (*1*) early ST-segment depression, (*2*) 0.2 mV or more of depression, (*3*) downsloping ST depression, (*4*) exercise-induced hypotension, (*5*) prolonged ST changes after the test, and (*6*) multiple areas of ST depression. While Lee et al. (29) included many clinical and exercise test variables, only three variables were found to help predict left main disease: angina type, age, and the amount of exercise-induced ST segment depression.

Meta Analysis of Studies Predicting Angiographic Severity

To evaluate the variability in the reported accuracy of the exercise ECG for predicting severe coronary disease, Detrano et al. (30) applied meta analysis to 60 consecutively published reports comparing exercise-induced ST depression with coronary angiographic findings. The 60 reports included 62 distinct study groups comprising 12,030 patients who underwent both tests. Both technical and methodologic factors were analyzed. Wide variability in sensitivity and specificity was found [mean sensitivity 86% (range 40–100%); mean specificity 53% (range 17–100%)] for left main or triple vessel disease. All three variables found to be significantly and independently related to test performance were methodological. Exclusion of patients with right bundle branch block and receiving digoxin improved the prediction of triple vessel or left main coronary artery disease and comparison with a better exercise test decreased test performance.

Hartz et al. (31) compiled results from the literature on the use of the exercise test to identify patients with severe coronary artery disease. Pooled estimates of sensitivity and specificity were derived for the ability of the exercise test to identify three-vessel or left main coronary artery disease. One millimeter criteria averaged a sensitivity of 75% and a specificity of 66% while two millimeters criteria averaged a sensitivity of 52% and a specificity of 86%. There was great variability among the studies examined in the estimated sensitivity and specificity for severe coronary artery disease that could not be explained by their analysis.

Table 8. Summary of Studies Assessing Maximal Heart Rate

Investigator	No. Subjects	Population[a] Studied	Mean Age ±SD (Range)	Mean HR Max (SD)	Regression Line	Correlation[a] Coefficient	Standard Error of the Estimate, beats/min[a]
Astrand[b]	100	Asymptomatic men	50 (20–69)	166 ± 22	$y = 211 - 0.922$ (age)	NA	NA
Bruce	2091	Asymptomatic men	44 ± 8	181 ± 12	$y = 210 - 0.662$ (age)	−0.44	14
Cooper	2535	Asymptomatic men	43 (11–79)	181 ± 16	$y = 217 - 0.845$ (age)	NA	NA
Ellestad[c]	2583	Asymptomatic men	42 ± 7 (10–60)	173 ± 11	$y = 197 - 0.556$ (age)	NA	NA
Froelicher	1317	Asymptomatic men	38 ± 8 (28–54)	183	$y = 207 - 0.64$ (age)	−0.43	10
Lester	148	Asymptomatic men	43 (15–75)	187	$y = 205 - 0.411$ (age)	−0.58	NA
Robinson	92	Asymptomatic men	30 (6–76)	189	$y = 212 - 0.775$ (age)	NA	NA
Sheffield	95	Men with CHD	39 (19–69)	176 ± 14	$y \pm 216 - 0.88$ (age)	−0.58	11[d]
Bruce	1295	Men with CHD	52 ± 8	148 ± 23	$y = 204 - 1.07$ (age)	−0.36	25[d]
Hammond	156	Men with CHD	53 ± 9	157 ± 20	$y = 209 - 1.0$ (age)	−0.30	19
Morris	244	Asymptomatic men	45 (20–72)	167 ± 19	$y = 200 - 0.72$ (age)	−0.55	15
Graettinger	114	Asymptomatic men	46 ± 13 (19–73)	168 ± 18	$y = 199 - 0.63$ (age)	−0.47	NA
Morris	1388	Men referred for evaluation for CHD, normals only	57 (21–89)	144 ± 20	$y = 196 - 0.9$ (age)	−0.43	21

[a]CHD = coronary heart disease; HR max = maximal heart rate; NA = not able to calculate from available data.
[b]Astrand used bicycle ergometry; all other studies were performed on a treadmill.
[c]Data compiled from graphs in reference cited.
[d]Calculated from available data.

Studies Using Multivariate Techniques to Predict Severe Angiographic CAD

Multiple studies have reported combining the patient's medical history, symptoms of chest pain, hemodynamic data, exercise capacity and exercise test responses to calculate the probability of severe angiographic coronary artery disease (32–43). The results are summarized in Table 8. Of the 13 studies, 9 excluded patients with previous coronary artery bypass surgery or prior percutaneous coronary intervention (PCI) and in the remaining 4 studies, exclusions were unclear. The percentage of patients with one vessel, two vessels and three vessels was described in 10 of the 13 studies. The definition of severe disease or disease extent also differed. In 5 of the 13 studies disease extent was defined as multivessel disease. In the remaining 8 studies, it was defined as three-vessel or left main disease and in one of them as only left main artery disease and in another the impact of disease in the right coronary artery disease on left main disease was considered. The prevalence of severe disease ranged from 16 to 48% in the studies defining disease extent as multivessel disease and from 10 to 28% in the studies using the more strict criterion of three-vessel or left main disease.

Chosen Predictors

Interestingly, some of the variables chosen for predicting disease severity are different than those for predicting disease presence. While gender and chest pain were chosen to be significant in more than half of the severity studies, age was less important and resting ECG abnormalities and diabetes were the only other variables chosen in more than half the studies. In contrast, the most consistent clinical variables chosen for diagnosis were age, gender, chest pain type, and hypercholesterolemia. ST depression and slope were frequently chosen for severity, but METs and heart rate were less consistently chosen than for diagnosis. Double product and delta SBP were chosen as independent predictors in more than half of the studies predicting severity.

Consensus to Improve Prediction

So far, only two studies [Detrano et al. (41) and Morise et al. (38)] have published equations that have been validated in large patient samples. Even though validated, however, the equations from these studies must be calibrated before they can be applied clinically. For example, a score can be discriminating but provide an estimated probability that is higher or lower than the actual probability. The scores can be calibrated by adjusting them according to disease prevalence; most clinical sites however, do not know their disease prevalence and even so, it could change from month to month.

In NASA trajectories of spacecraft are often determined by agreement between three or more equations calculating the vehicle path. With this in mind, we developed an agreement method to classify patients into high, no agreement, or low risk groups for probability of severe disease by requiring agreement in all three equations [Detrano, Morise, and ours (LB–PA)] (44). This approach adjusts the calibration and makes the equations applicable in clinical populations with varying prevalence of coronary artery disease.

It was demonstrated that using simple clinical and exercise test variables could improve the standard application of

ECG criteria for predicting severe coronary artery disease. By setting probability thresholds for severe disease of $< 20\%$ and $> 40\%$ for the three prediction equations, the agreement approach divided the test set into populations with low, no agreement, and high risk for severe coronary artery disease. Since the patients in the no agreement group would be sent for further testing and would eventually be correctly classified, the sensitivity of the agreement approach is 89% and the specificity is 96%. The agreement approach appeared to be unaffected by disease prevalence, missing data, variable definitions, or even by angiographic criterion. Cost analysis of the competing strategies revealed that the agreement approach compares favorably with other tests of equivalent predictive value, such as nuclear perfusion imaging, reducing costs by 28% or $504 per patient in the test set.

Requiring agreement of these three equations to make diagnosis of severe coronary disease has made them widely applicable. Excellent predictive characteristics can be obtained using simple clinical data entered into a computer. Cost analysis suggests that the agreement approach is an efficient method for the evaluation of populations with varying prevalence of coronary artery disease, limiting the use of more expensive noninvasive and invasive testing to patients with a higher probability of left main or three vessel coronary artery disease. This approach provides a strategy for assisting the practitioner in deciding when further evaluation is appropriate or interventions indicated.

Predicting Improved Survival with Coronary Artery Bypass Surgery

Which exercise test variables indicate those patients who would have an improved prognosis if they underwent coronary artery bypass surgery (CABS)? Research in this area is limited by the fact the available studies did not randomize patients to surgery based on the results of their exercise test results and the retrospective nature of the studies.

Bruce et al. (45) demonstrated noninvasive screening criteria for patients who had improved 4 year survival after coronary artery bypass surgery. Their data have come from 2000 men with coronary heart disease enrolled in the Seattle Heart Watch who had a symptom-limited maximal treadmill test; these subjects received usual community care, which resulted in 16% of them having coronary artery bypass surgery in nonrandomized fashion. Patients with cardiomegaly, < 5 MET exercise capacity and/or a maximal systolic blood pressure of < 130 would have a better outcome if treated with surgery. Two or more of the above parameters present the highest risk and the greater differential for improved survival with bypass. Four year survival in this group would be 94% for those that had surgery versus 67% for those who received medical management (in those who had two or more of the above factors). In the European surgery trial (46), patients who had an exercise test response of 1.5 mm of ST segment depression had improved survival with surgery. This also extended to those with baseline ST segment depression and those with claudication.

From the CASS study group (47), in > 5000 nonrandomized patients, though there were definite differences between the surgical and nonsurgical groups, this could be accounted for by stratification in subsets. The surgical benefit regarding mortality was greatest in the 789 patients with 1 mm ST segment depression at < 5 METs. Among the 398 patients with triple vessel disease with this exercise test response, the 7 year survival was 50% in those medically managed versus 81% in those who underwent coronary artery bypass surgery. There was no difference in mortality in randomized patients able to exceed 10 METs. In the VA surgery randomized trial (48), there was a 79% survival rate with CABS versus 42% for medical management in patients with two or more of the following: 2 mm or more of ST depression, heart rate of 140 or greater at 6 METs, and/or exercise-induced PVCs.

Evaluation of Percutaneous Coronary Interventions

One important application of the exercise test is to assess the effects of percutaneous coronary intervention (PCI) on physical function, ischemic responses, and symptoms in the immediate and longer period following the various interventions that now fall under the general term PCI. The exercise test has been used for this purpose in numerous trials of PCI, and a few notable examples are described in the following. Berger et al. (49) reported follow-up data in 183 patients who had undergone PCI at least 1 year earlier. PCI was initially successful in 141 patients (79%). Of the 42 patients in whom PCI was unsuccessful, 26 underwent CABG, while 16 were maintained on medical therapy. When compared to the medical patients at time of follow-up, successful PCI patients experienced less angina (13 vs. 47%), used less nitroglycerin (25 vs. 73%), were hospitalized less often for chest pain (8 vs. 31%), and subjectively felt their condition had improved (96 vs. 20%).

Vandormael and colleagues reported the safety and short-term benefit of multi-lesion PCI in 135 patients (50). Primary success, defined as successful dilation of the most critical lesion or all lesions attempted, occurred in 87% of the 135 patients. Exercise-induced angina occurred in 11 (12%) and an abnormal exercise ECG in 30 (32%) of the 95 patients with post-PCI exercise test data. Of 57 patients who had paired exercise test data before and after angioplasty, exercise-induced angina occurred in 56% of patients before the procedure, compared with only 11% of patients after angioplasty. Exercise-induced ST-segment depression of > 0.1 mV occurred in 75% of patients before PCI versus 32% after the procedure.

Rosing et al. (51) reported that exercise testing after successful PCI exhibited improved ECG and symptomatic responses, as well as improved myocardial perfusion and global and regional left ventricular function.

Prediction of Restenosis with the Exercise Test

To determine whether a treadmill test could predict restenosis after angioplasty, Honan et al. (52) studied 289 patients six months after a successful emergency angioplasty of the infarct-related artery for acute myocardial infarction (MI). After excluding those with interim interventions, medical events, or medical contraindications

to follow-up testing, both a treadmill test and a cardiac catheterization were completed in 144 patients; 88% of those eligible for this assessment. Of six clinical and treadmill variables examined by multivariable logistic analysis, only exercise ST deviation was independently correlated with restenosis. The sensitivity of ST deviation of 0.10 mV or greater for detecting restenosis was only 24% (13 of 55 patients), and the specificity was 88% (75 of 85 patients). Extent or severity of wall motion abnormalities at follow-up did not affect the sensitivity of exercise-induced ST deviation for detection of restenosis, by the timing of thrombolytic therapy or of angioplasty, or by the presence of collateral blood flow at the time of acute angiography. A second multivariable analysis evaluating the association of the same variables with number of vessels with significant coronary disease at the 6 month catheterization found an association with both exercise ST deviation and exercise duration. Angina symptoms and exercise test results in this population had limited value for predicting anatomic restenosis six months after emergency angioplasty for acute myocardial infarction.

Bengtson et al. (53) studied 303 consecutive patients with successful PCI and without a recent myocardial infarction. Among the 228 patients without interval cardiac events, early repeat revascularization or contraindications to treadmill testing, 209 (92%) underwent follow-up angiography, and 200 also had a follow-up treadmill test and formed the study population. Restenosis occurred in 50 patients (25%). Five variables were individually associated with a higher risk of restenosis: recurrent angina, exercise-induced angina, a positive treadmill test, greater exercise ST deviation, and a lower maximum exercise heart rate. However, only exercise-induced angina, recurrent angina, and a positive treadmill test were independent predictors of restenosis. Using these three variables, patient subsets could be identified with restenosis rates ranging from 11 to 83%. The exercise test added independent information to symptom status regarding the risk of restenosis after elective PCI. Nevertheless, 20% of patients with restenosis had neither recurrent angina nor exercise-induced ischemia at follow-up.

The ROSETTA registry was studied to demonstrate the effects of routine post-PCI functional testing on the use of follow-up cardiac procedures and clinical events (54). The ROSETTA (Routine versus Selective Exercise Treadmill Testing after Angioplasty) registry is a prospective multicenter observational study examining the use of functional testing after PCI. A total of 788 patients were enrolled in the registry at 13 clinical centers in 5 countries. The frequencies of exercise testing, cardiac procedures and clinical events were examined during the first 6 months following a successful PCI. Patients were predominantly elderly men (mean age, 61 ± 11 years; 76% male) who underwent single-vessel PCI (85%) with stent implantation (58%). During the 6 month follow-up, a total of 237 patients underwent a routine exercise testing strategy (100% having exercise testing for routine follow-up), while 551 pts underwent a selective (or clinically driven) strategy (73% having no exercise testing and 27% having exercise testing for a clinical indication). Patients in the routine testing group underwent a total of 344 exercise tests compared with 165 tests performed in the selective testing group (mean, 1.45 tests/patient vs. 0.3 tests/patient). However, clinical events were less common among those who underwent routine exercise testing, for example, unstable angina (6% vs. 14%), myocardial infarction (0.4% vs. 1.6%), death (0% vs. 2%) and composite clinical events (6% vs. 16%). After controlling for baseline clinical and procedural differences, routine exercise testing had a persistent independent association with a reduction in the composite clinical event rate. This association may be attributable to the early identification and treatment of patients at risk for follow-up events, or it may be due to clinical differences between patients who are referred for routine and selective exercise testing.

The ACC/AHA Guidelines for the Prognostic Use of the Standard Exercise Test

The task force to establish guidelines for the use of exercise testing has met and produced guidelines in 1986, 1997, and 2002. The following is a synopsis of these evidence-based guidelines.

Indications for Exercise Testing to Assess Risk and prognosis in patients with symptoms or a prior history of CAD:

Class I (Definitely Appropriate). Conditions for which there is evidence and/or general agreement that the standard exercise test is useful and helpful to assess risk and prognosis in patients with symptoms or a prior history of CAD.

- Patients undergoing initial evaluation with suspected or known CAD. Specific exceptions are noted below in Class IIb.
- Patients with suspected or known CAD previously evaluated with significant change in clinical status.

Class IIb (Maybe Appropriate). Conditions for which there is conflicting evidence and/or a divergence of opinion that the standard exercise test is useful and helpful to assess risk and prognosis in patients with symptoms or a prior history of coronary artery disease but the usefulness/efficacy is less well established.

- Patients who demonstrate the following ECG abnormalities: pre excitation (Wolff–Parkinson White) syndrome; electronically paced ventricular rhythm; > 1 mm of resting ST depression; and complete left bundle branch block.
- Patients with a stable clinical course who undergo periodic monitoring to guide management

Class III (Not Appropriate). Conditions for which there is evidence and/or general agreement that the standard exercise test is not useful and helpful to assess risk and prognosis in patients with symptoms or a prior history of CAD and in some cases may be harmful.

- Patients with severe comorbidity likely to limit life expectancy and/or candidacy for revascularization.

In summary, the two principal reasons for estimating prognosis are to provide accurate answers to patient's questions regarding the probable outcome of their illness and to identify those patients in whom interventions might improve outcome. There is a lack of consistency in the available studies because patients die along a pathophysiological spectrum ranging from those that die due to CHF with little myocardium remaining to those that die from an ischemic related event with ample myocardium remaining. Clinical and exercise test variables most likely associated with CHF deaths (CHF markers) include a history or symptoms of CHF, prior MI, Q waves, and other indicators of LV dysfunction. Variables most likely associated with ischemic deaths (ischemic markers) are angina, and rest and exercise ST depression. Some variables can be associated with either extremes of the type of CV death; these include exercise capacity, maximal heart rate, and maximal systolic blood pressure that may explain why they are reported most consistently in the available studies. A problem exists that ischemic deaths occur later in follow up and are more likely to occur in those lost to follow up whereas CHF deaths are more likely to occur early (within 2 years) and are more likely to be classified. Work-up bias probably explains why exercise-induced ST depression fails to be a predictor in most of the angiographic studies. Ischemic markers are associated with a later and lesser risk, whereas CHF or left ventricular dysfunction markers are associated with a sooner and greater risk of death.

Recent studies of prognosis have actually not been superior to the earlier studies that considered CV endpoints and removed patients from observation who had interventions. This is because death data is now relatively easy to obtain while previously investigators had to follow the patients and contact them or review their records. CV mortality can be determined by death certificates. While death certificates have their limitations, in general they classify those with accidental, GI, Pulmonary and cancer deaths so that those remaining are most likely to have died of CV causes. This endpoint is more appropriate for a test for CV disease. While all-cause mortality is a more important endpoint for intervention studies, CV mortality is more appropriate for evaluating a CV test (i.e., the exercise test). Identifying those at risk of death of any cause does not make it possible to identify those who might benefit from CV interventions, one of the main goals of prognostication.

The consistencies actually overshadow the differences. Considering simple clinical variables can assess risk. A good exercise capacity, no evidence or history of CHF or ventricular damage (Q waves, history of CHF), no ST depression or only one of these clinical findings are associated with a very low risk. These patients are low risk in exercise programs and need not be considered for interventions to prolong their life. High risk patients can be identified by groupings of the clinical markers; that is, two or more. Exertional hypotension is particularly ominous. Identification of high risk implies that such patients in exercise training programs should have lower goals and should be monitored. Such patients should also be considered for coronary interventions to improve their longevity. Furthermore, with each drop in METs there is a 10–20% increase in mortality so simple exercise capacity has consistent importance in all patient groups.

The mathematical models for determining prognosis are usually more complex than those used for identifying severe angiographic disease. Diagnostic testing can utilize multivariate discriminant function analysis to determine the probability of severe angiographic disease being present or not. Prognostic testing must utilize survival analysis which includes censoring for patients with uneven follow-up due to "lost to follow up" or other cardiac events (i.e., CABS, PCI) and must account for time-person units of exposure. Survival curves must be developed and the Cox proportional hazards model is often preferred.

From this perspective, it is obvious that there is substantial support for the use of the exercise test as a first noninvasive step after the history, physical exam, and resting ECG in the prognostic evaluation of coronary artery disease patients. It accomplishes both of the purposes of prognostic testing: to provide information regarding the patient's status and to help make recommendations for optimal management. The exercise test results help us make reasonable decisions for selection of patients who should undergo coronary angiography. Since the exercise test can be performed in the doctor's office and provides valuable information for clinical management in regard to activity levels, response to therapy, and disability, the exercise test is the reasonable first choice for prognostic assessment. This assessment should always include calculation of the estimated annual mortality using the Duke treadmill score though its ischemic elements have less power in the elderly.

There has been considerable debate in screening asymptomatic patients. Screening has become a controversial topic because of the incredible efficacy of the statins (drugs that lower cholesterol) even in asymptomatic individuals (55). There are now agents that can cut the risk of cardiac events almost in half. The first step in screening asymptomatic individuals for preclinical coronary disease should be using global risk factor equations, such as the Framingham score. This is available as nomograms that are easily applied by healthcare professionals or it can be calculated as part of a computerized patient record. Additional testing procedures with promise include the simple ankle-brachial index (particularly in the elderly), CRP, carotid ultrasound measurements of intimal thickening, and the resting ECG (particularly spatial QRS-T wave angle). Despite the promotional concept of atherosclerotic burden, EBCT does not have test characteristics superior to the standard exercise test. If any screening test could be used to decide regarding statin therapy and not affect insurance or occupational status, this would be helpful. However, the screening test should not lead to more procedures.

True demonstration of the effectiveness of a screening technique requires randomizing the target population, applying the screening technique to half, taking standardized action in response to the screening test results, and then assessing outcomes. Efficacy of the screening test necessitates that the screened group has lower mortality and/or morbidity. Such a study has been completed for

mammography, but not for any cardiac testing modalities. The next best validation of efficacy is to demonstrate that the technique improves the discrimination of those asymptomatic individuals with higher risk for events over that possible with the available risk factors. Mathematical modeling makes it possible to determine how well a population will be classified if the characteristics of the testing method are known.

Additional follow-up studies and one angiographic study from the CASS population (where 195 individuals with abnormal exercise-induced ST depression and normal coronary angiograms were followed for 7 years) improve our understanding of the application of exercise testing as a screening tool. No increased incidence of cardiac events was found, and so the concerns raised by Erikssen's findings in 36 subjects that they were still at increased risk have not been substantiated.

The later follow-up studies (MRFIT, Seattle Heart Watch, Lipid Research Clinics, and Indiana State Police) have shown different results compared to prior studies, mainly because hard cardiac end points and not angina were required. The first ten prospective studies of exercise testing in asymptomatic individuals included angina as a cardiac disease end point. This led to a bias for individuals with abnormal tests to subsequently report angina or to be diagnosed as having angina. When only hard end points (death or MI) were used, as in the MRFIT, Lipid Research Clinics, Indiana State Police or the Seattle Heart Watch studies, the results were less encouraging. The test could only identify one-third of the patients with hard events and 95% of abnormal responders were false positives; that is, they did not die or have a MI. The predictive value of the abnormal maximal exercise electrocardiogram ranged from 5 to 46% in the studies reviewed. However, in the studies using appropriate endpoints (other than angina pectoris) only 5% of the abnormal responders developed coronary heart disease over the follow-up period. Thus, $>90\%$ of the abnormal responders were false positives. However, the exercise test's characteristics as a screening test probably lie in between the results with hard or soft endpoints since some of the subjects who develop chest pain really have angina and coronary disease. The sensitivity is probably between 30 and 50% (at a specificity of 90%) but the critical limitation is the predictive value (and risk ratio) which depends upon the prevalence of disease (which is low in the asymptomatic population).

Although some of these individuals have indolent coronary disease yet to be manifest, angiographic studies have supported this high false positive rate when using the exercise test in asymptomatic populations. Moreover, the CASS study indicates that such individuals have a good prognosis. In a second Lipid Research Clinics study, only patients with elevated cholesterol's were considered, and yet only a 6% positive prediction value was found. If the test is to be used to screen it should be done in groups with a higher estimated prevalence of disease using the Framingham score and not just one risk factor. The iatrogenic problems resulting from screening must be considered. Hopefully, using a threshold from the Framingham score would be more successful in identifying asymptomatic individuals that should be tested.

Some individuals who eventually develop coronary disease will change on retesting from a normal to an abnormal response. However, McHenry and Fleg have reported that a change from a negative to a positive test is no more predictive than is an initially abnormal test. One individual has even been reported who changed from a normal to an abnormal test, but was free of angiographically significant disease (56). In most circumstances an added imaging modality (echocardiographic or nuclear) should be the first choice in evaluating asymptomatic individuals with an abnormal exercise test.

The motivational impact of screening for CAD is not evidence based with one positive study for exercise testing and one negative study for EBCT. Further research in this area certainly is needed.

While the risk of an abnormal exercise test is apparent from these studies, the iatrogenic problems resulting from screening must be considered (i.e., employment, insurance). The recent U.S. Preventive Services Task Force statement states that "false positive tests are common among asymptomatic adults, especially women, and may lead to unnecessary diagnostic testing, over treatment and labeling". This statement summarizes the current U.S. Preventive Services Task Force (USPSTF) recommendations on screening for coronary heart disease and the supporting scientific evidence and updates the 1996 recommendations on this topic. The complete information on which this statement is based, including evidence tables and references, is available in the background article and the systematic evidence review, available through the USPSTF Web site (http://www.preventiveservices.ahrq.gov) and through the National Guideline Clearinghouse (http://www.guideline.gov) (57). In the majority of asymptomatic people, screening with any test or test add-on, is more likely to yield false positives than true positives. This is the mathematical reality associated with all of the available tests.

There are reasons to include exercise testing in the preventative health recommendations for screening healthy, asymptomatic individuals along with risk factor assessment. The additional risk classification power documented by the data from Norway (2000 men, 26 year follow up), the Cooper Clinic (26,000 men, 8 year follow up), and Framingham (3000 men, 18 year follow up) provide convincing evidence that the exercise test should be added to the screening process. Furthermore, exercise capacity itself has substantial prognostic predictive power. Given the emerging epidemic of physical inactivity, including the exercise test in the screening process sends a strong message to our patients that we consider their exercise status as important.

However, if screening could be performed in a logical way with test results helping to decide on therapies rather than leading to invasive interventions, insurance or occupational problems, then the recent results summarized above should be applied to preventive medicine policy.

Because of the inherent difficulties, few preventive medicine recommendations are based on randomized trials demonstrating improved outcomes but rely on reasonable assumptions from available evidence. There is now enough evidence to consider recommending a routine exercise test every five years for men >40 and women >50 years of age,

especially if one of the potential benefits is the adoption of an active lifestyle (58).

CONCLUSION

While there are important technological considerations and the need to be very knowledgeable of the guidelines to insure its proper application, exercise testing remains one of the most widely used and valuable noninvasive tools to assess cardiovascular status.

The following precepts regarding methodology are important to follow:

- The treadmill protocol should be adjusted to the patient and one protocol is not appropriate for all patients.
- Exercise capacity should be reported in METs not minutes of exercise.
- Hyperventilation prior to testing is not indicated, but can be utilized at another time, if a false positive is suspected
- ST measurements should be made at ST0 (J-junction) and ST depression should only be considered abnormal if horizontal or downsloping; 95% of the clinically important ST depression occurs in V_5 particularly in patients with a normal resting ECG.
- Patients should be placed supine as soon as possible postexercise with a cool down walk avoided in order for the test to have its greatest diagnostic value.
- The 2–4 min recovery period is critical to include in analysis of the ST response.
- Measurement of systolic blood pressure during exercise is extremely important and exertional hypotension is ominous; at this point, only manual blood pressure measurement techniques are valid.
- Age-predicted heart rate targets are largely useless because of the wide scatter for any age; a relatively low heart rate can be maximal for a given patient and submaximal for another.
- The Duke Treadmill Score should be calculated automatically on every test except for the elderly.
- Other predictive equations and heart rate recovery should be considered a standard part of the treadmill report.

BIBLIOGRAPHY

Cited References

1. Freeman J, Dewey R, Hadley D, Froelicher V. Evaluation of the Autonomic Nervous System with Exercise Testing. Prog Cardiovasc Dis 2005 Jan-Feb;47(4):285–305.
2. Froelicher VF, et al. A comparison of two-bipolar electrocardiographic leads to lead V_5. Chest 1976;70:611.
3. Gamble P, et al. A comparison of the standard 12-lead electrocardiogram to exercise electrode placements. Chest 1984;85: 616–622.
4. Miranda CP, et al. Usefulness of exercise-induced ST-segment depression in the inferior leads during exercise testing as a marker for coronary artery disease. Am J Cardiol 1992;69: 303–307.
5. Milliken JA, Abdollah H, Burggraf GW. False-positive treadmill exercise tests due to computer signal averaging. Am J Cardiol 1990;65:946–948.
6. Willems J, et al. The diagnostic performance of computer programs for the interpretation of ECGs. N Engl J Med 1991;325:1767–1773.
7. Balady GJ, et al. Value of arm exercise testing in detecting coronary artery disease. Am J Cardiol 1985;55:37–39.
8. Myers J, et al. Comparison of the ramp versus standard exercise protocols. J Am Coll Cardiol 1991;17:1334–1342.
9. Wickes JR, et al. Comparison of the Electrocardiographic changes induced by maximum exercise testing with treadmill and cycle ergometer. Circulation 1978;57:1066–1069.
10. Hambrecht RP, et al. Greater diagnostic sensitivity of treadmill versus cycle exercise testing of asymptomatic men with coronary artery disease. Am J Cardiol 1992 Jul 15;70(2): 141–146.
11. Sullivan M, McKirnan MD. Errors in predicting functional capacity for post myocardial infarction patients using a modified Bruce protocol. Am Heart J 1984;107:486–491.
12. Webster MWI, Sharpe DN. Exercise testing in angina pectoris: the importance of protocol design in clinical trials. Am Heart J 1989;117:505–508.
13. Goldman L, et al. Comparative reproducibility and validity of systems for assessing cardiovascular function class: advantages of a new specific activity scale. Circulation 1981;64: 1227–1234.
14. Fletcher GF, et al. Exercise standards for testing and training: a statement for healthcare professionals from the American Heart Association. Circulation 2001;104:1694–1740.
15. Borg G, Holmgren A, Lindblad I. Quantitative evaluation of chest pain. Acta Med Scand 1981;644:43–45.
16. Gutman RA, et al. Delay of ST depression after maximal exercise by walking for two minutes. Circulation 1970;42: 229–233.
17. Lachterman B, et al. "Recovery only" ST segment depression and the predictive accuracy of the exercise test. Ann Intern Med 1990;112:11–16.
18. Ashmaig ME, et al. Changes in serum concentrations of markers of myocardial injury following treadmill exercise testing in patients with suspected ischaemic heart disease. Med Sci Monit 2001;7:54–57.
19. Akdemir I, et al. Does exercise-induced severe ischaemia result in elevation of plasma troponin-T level in patients with chronic coronary artery disease? Acta Cardiol 2002;57: 13–18.
20. Sabatine MS, et al. TIMI Study Group. Acute changes in circulating natriuretic peptide levels in relation to myocardial ischemia. J Am Coll Cardiol 2004;44(10):1988–1995.
21. Foote RS, Pearlman JD, Siegel AH, Yeo KT. Detection of exercise-induced ischemia by changes in B-type natriuretic peptides. J Am Coll Cardiol 2004;44(10):1980–1987.
22. Wang TJ, et al. Plasma natriuretic peptide levels and the risk of cardiovascular events and death. N Engl J Med 2004 Feb 12; 350(7):655–663.
23. Kragelund C, Gronning B, Kober L, Hildebrandt P, Steffensen R. N-terminal pro-B-type natriuretic peptide and long-term mortality in stable coronary heart disease. N Engl J Med 2005 Feb 17;352(7):666–75.
24. Cheitlin MD, et al. Correlation of "critical" left coronary artery lesions with positive submaximal exercise tests in patients with chest pain. Am Heart J 1975;89(3):305–310.
25. Goldschlager N, Selzer A, Cohn K. Treadmill stress tests as indicators of presence and severity of coronary artery disease. Ann Int Med1976;85:277–286.
26. NcNeer JF, et al. The role of the exercise test in the evaluation of patients for ischemic heart disease. Circulation 1978; 57:64–70.

27. Weiner DA, McCabe CH, Ryan TJ. Identification of Patients with left main and three vessel coronary disease with clinical and exercise test variables. Am J Cardiol 1980;46:21–27.
28. Blumenthal DS, Weiss JL, Mellits ED, Gerstenblith G. The predictive value of a strongly positive stress test in patients with minimal symptoms. Am J Med 1981;70:1005–1010.
29. Lee TH, Cook EF, Goldman L. Prospective evaluation of a clinical and exercise-test model for the prediction of left main coronary artery disease. Med Decis Making 1986;6: 136–144.
30. Detrano R, et al. Exercise-induced ST segment depression in the diagnosis of multivessel coronary disease: A meta analysis. J Am Coll Cardiol 1989;14:1501–1508.
31. Hartz A, Gammaitoni C, Young M. Quantitative analysis of the exercise tolerance test for determining the severity of coronary artery disease. Int J Cardiol 1989;24:63–71.
32. Cohn K, et al. Use of treadmill score to quantify ischemic response and predict extent of coronary disease. Circulation 1979;59:286–296.
33. Fisher L, et al. Diagnostic quantification of CASS (Coronary artery surgery study) clinical and exercise test results in determining presence and extent of coronary artery disease. Circulation 1981;63:987–1000.
34. McCarthy D, Sciacca R, Blood D, Cannon P. Discriminant function analysis using thallium 201 scintiscans and exercise stress test variables to predict the presence and extent of coronary artery disease. Am J Cardiol 1982;49:1917–1926.
35. Lee T, Cook E, Goldman L. Prospective evaluation of a clinical and exercise test model for the prediction of left main coronary artery disease. Med Decis Making 1986;6:136–144.
36. Hung J, et al. A logistic regression analysis of multiple noninvasive tests for the prediction of the presence and extent of coronary artery disease in men. Am Heart J 1985;110:460–469.
37. Christian T, Miller T, Bailey K, Gibbons R. Exercise tomographic thallium-201 imaging in patients with severe coronary artery disease and normal electrocardiograms. Ann Intern Med 1994;121:825–832.
38. Morise A, Bobbio M, Detrano R, Duval R. Incremental evaluation of exercise capacity as an independent predictor of coronary artery disease presence and extent. Am Heart J 1994;127:32–38.
39. Morise A, Diamond G, Detrano R, Bobbio M. Incremental value of exercise electrocardiography and thallium-201 testing in men and women for the presence and extent of coronary artery disease. Am Heart J 1995;130:267–276.
40. Moussa I, Rodriguez M, Froning J, Froelicher VF. Prediction of severe coronary artery disease using computerized ECG measurements and discriminant function analysis. J Electrocardiol 1992;25:49–58.
41. Detrano R, et al. Algorithm to predict triple-vessel/left main coronary artery disease in patients without myocardial infarction. Circulation 1991;83(3):89–96.
42. Christian TF, Miller TD, Bailley KR, Gibbons RJ. Noninvasive identification of severe coronary artery disease using exercise tomographic thallium-201 imaging. Am J Cardiol 1992;70:14–20.
43. Hung J, et al. Noninvasive diagnostic test choices for the evaluation of coronary artery disease in women:a multivariate comparison of cardiac fluoroscopy, exercise electrocardiography and exercise thallium myocardial perfusion scintigraphy. J Am Coll Cardiol 1984;4:8–16.
44. Do D, West JA, Morise A, Froelicher VF. Agreement Predicting Severe Angiographic Coronary Artery Disease Using Clinical and Exercise Test Data. Am Heart J 1997;134: 672–679.
45. Bruce RA, Hossack KF, DeRouen TA, Hofer V. Enhanced risk assessment for primary coronary heart disease events by maximal exercise testing: 10 years' experience of Seattle Heart Watch. J Am Coll Cardiol 1983;2:565–73.
46. European Cooperative Group. Long-term results of prospective randomized study of coronary artery bypass surgery in stable angina pectoris. Lancet 1982; 1173–1180.
47. Weiner DA, et al. The role of exercise testing in identifying patients with improved survival after coronary artery bypass surgery. J Am Coll Cardiol 1986;8(4):741–748.
48. Hultgren HN, Peduzzi P, Detre K, Takaro T. The 5 year effect of bypass surgery on relief of angina and exercise performance. Circulation 1985;72:V79–V83.
49. Berger E, Williams DO, Reinert S, Most AS. Sustained efficacy of percutaneous transluminal coronary angioplasty. Am Heart J 1986;111:233–236.
50. Vandormael MG, et al. Immediate and short-term benefit of multilesion coronary angioplasty: Influence of degree of revascularization. J Am Coll Cardiol 1985;6:983–991.
51. Rosing DR, et al. Exercise, electrocardiographic and functional responses after percutaneous transluminal coronary angioplasty. Am J Cardiol 1984;53:36C–41C.
52. Honan MB, et al. Exercise treadmill testing is a poor predictor of anatomic restenosis after angioplasty for acute myocardial infarction. Circulation 1989;80:1585–1594.
53. Bengtson JR, et al. Detection of restenosis after elective percutaneous transluminal coronary angioplasty using the exercise treadmill test. Am J Cardiol 1990;65:28–34.
54. Eisenberg MJ, et al. ROSETTA Investigators. Utility of routine functional testing after percutaneous transluminal coronary angioplasty: results from the ROSETTA registry. J Invasive Cardiol 2004;16:318–322.
55. Downs JR, et al. Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: Results of AFCAPS/TexCAPS. Air Force/Texas Coronary Atherosclerosis Prevention Study. JAMA 1998;279: 1615–1622.
56. Thompson AJ, Froelicher VF. Normal coronary angiography in an aircrewman with serial test changes. Aviat Space Environ Med 1975;46:69–73.
57. U.S. Preventive Services Task Force. Screening for coronary heart disease: recommendation statement. Ann Intern Med 2004 Apr 6; 140(7):569–572.
58. DiPietro L, Kohl HW 3rd, Barlow CE, Blair SN. Improvements in cardiorespiratory fitness attenuate age-related weight gain in healthy men and women: the Aerobics Center Longitudinal Study. Int J Obes Relat Metab Disord 1998 Jan; 22(1):55–62.

See also BIOMECHANICS OF EXERCISE FITNESS; BLOOD PRESSURE MEASUREMENT; ELECTROCARDIOGRAPHY, COMPUTERS IN.

EXTRACORPOREAL SHOCK WAVE LITHOTRIPSY.

See LITHOTRIPSY.

EYE MOVEMENT, MEASUREMENT TECHNIQUES FOR

JOSHUA BORAH
Applied Science Laboratories
Bedford, Massachusetts

INTRODUCTION

The terms eye movement measurement, eye tracking, and oculogragphy refer to measurement of the orientation and motion of the eye, either with respect to the head, or with respect to the visual environment. This may include not

only rotations of the eye that cause changes in gaze direction, but also rotations of the eyeball about the line of sight, called ocular torsion. Point-of-gaze is the point in the visual environment whose image forms on the small, high acuity area of the retina, called the fovea. Line-of-gaze is the imaginary line connecting the eye to the point-of-gaze. Sometimes the term gaze tracker is used to describe a system whose primary function is to determine a subject's fixation point or line of gaze with respect to the visual environment, rather than the dynamics of eyeball motion with respect to the head.

Eye movement measurement devices have long been used for research in reading, various aspects of visual perception and cognition, neurology, instrument panel layout, and advertising. Technological advances, especially in the areas of digital processing and solid-state sensor technology, have made eye tracking possible under progressively less and less restrictive conditions. In recent years, uses have expanded to include computer application usability research, communication devices for the disabled, sports and gait research, Lasik surgery instrumentation, and research requiring simultaneous fMRI (functional magnetic resonance imaging) measurement. In the past decade it has also become practical to measure ocular torsion with optical, noncontacting methods.

Figure 1 shows some of the structures and dimensions of the eye that are important in eye movement measurement (1,2). In an idealized model, the optical axis of the eye is the line that passes through the centers of curvature of the cornea, and lens, and the center of rotation of the eyeball. The visual axis (or line of sight) is the ray that passes from the fovea, through the nodal points of the lens and intersects the point-of-gaze. It is important to note that the fovea is not centered on the retina, but rather is located 5–7° toward the temporal side. The visual and optical axes are therefore, not identical. An idealized model of the eye usually assumes the pupil and iris to be centered on the optical axis, and assumes that the eyeball and eye socket operate as a perfect ball and socket joint, with the eyeball rotating about a single point within the eye socket.

The idealized model is often perfectly adequate for making good measurements of gaze direction and eye movement dynamics, but is not precisely accurate. For example, the eye does not rotate about a single center of rotation within the eye socket (3). The pupil is not precisely centered with respect to the optical axis, visual axis, or iris, and its center moves as the iris opens and closes (4).

Eye position with respect to the head can be described by a three element rotation vector, by a four element rotation specification called a quarternion, or by a set of three angles that describe the positions of an imaginary set of nested gimbals, with the outer gimbal fastened to the head, and the eye attached to the inner most gimbal. In the latter case, the three angles are usually referred to as Fick or Euler angles, and consist of an azimuth (or horizontal) angle, an elevation (or vertical) angle, and a roll (or torsion) angle. In all cases rotations are measured from a somewhat arbitrary reference position that loosely corresponds to looking straight ahead when the head is upright. A complete description of the methods and underlying mathematics for specifying eye rotation is available in an article by Haslwanter (5).

Gaze tracking devices usually report point-of-gaze in terms of a coordinate system defined on a surface in the

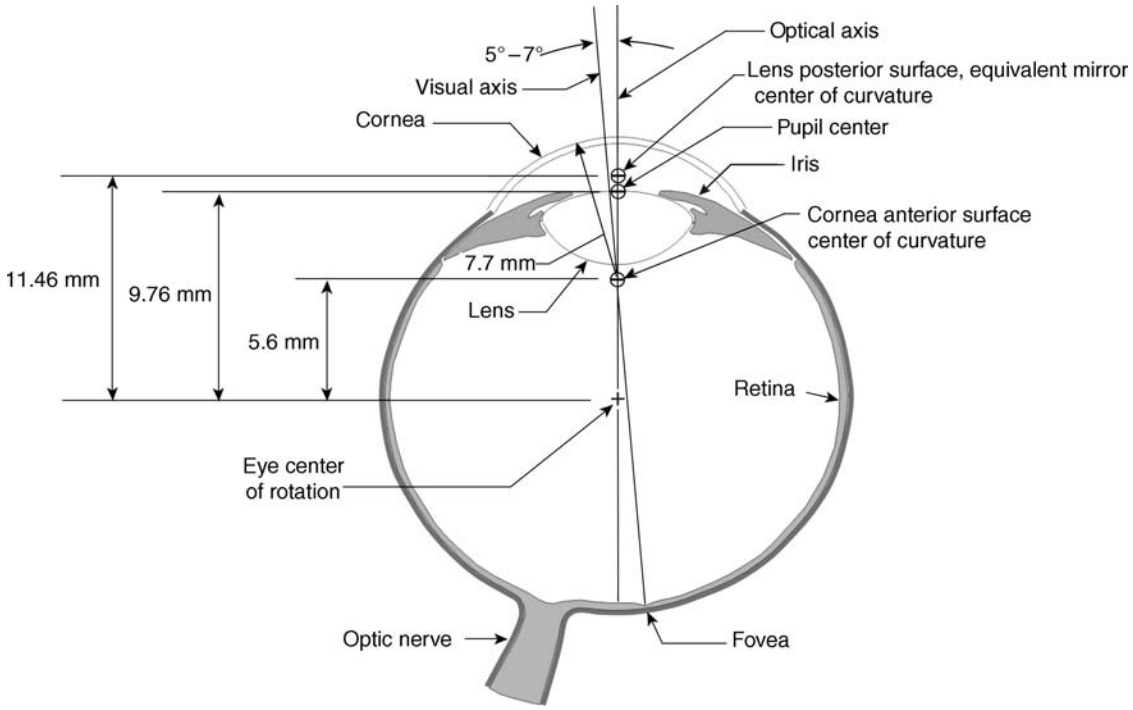


Figure 1. Schematic diagram of the eye showing typical values of dimensions that are important in eye movement measurement. The dimension values, which do vary between individuals, are derived from Refs. 1 and 2.

environment; or an eye location in space plus a gaze direction vector, with respect to an environment coordinate frame.

Normal human eye movements fall into the broad categories of conjugate and nonconjugate movements. Conjugate movements, in which both eyes move together, include the rapid, ballistic jumps between fixation points, called saccades; smooth compensatory movements to hold the gaze steady in the presence of head motion; smooth movements to track objects that are moving across the visual field; and the saw tooth pattern of movement called nystagmus that occurs in response to an inertial rotation of the body or a rotation of the visual field. There are also miniature motions during the relatively stationary fixation periods, which are $< 1^\circ$ and are not perceived. Vergence is a nonconjugate motion used to keep a visual target at the same position on both retinas. As a visual target moves closer, the visual axes of the two eyes rotate toward each other. Ocular torsion occurs in response to inertial rotation or lateral acceleration, or rotation of the visual field about a horizontal axis. It is associated with perceptions of tilt. A thorough review of eye movement behavior can be found in Hallett (6).

The eye movement measurement techniques currently in most frequent use fall into the major categories of magnetic search coil, a technique that measures magnetically induced current in a tiny wire coil fastened to the eye; electrooculography, which uses surface electrodes to measure the direction of an electrical potential between the cornea and retina; and optical techniques that rely on optical sensors to detect the position or motion of features on the eye. Their optical technique category includes many subcategories, and has the largest variety of different systems in current use. Background theory and system descriptions for eye movement measurement devices, in all of these categories, are presented in the following sections.

Eye tracker performance is usually described by some subset of the following parameters. *Accuracy* is the expected difference between the measured value and the true value. *Resolution* is the smallest change that can be reported by the device. *Precision* is the expected difference in repeated measurements of the same true value. *Range* describes the span of values that can be measured by the device. *Linearity* is the degree to which a given change in the real quantity results in a proportional change in the measured value, usually expressed as percent of the measurement range. *Update rate* is the frequency with which data is output (samples per second). *Bandwidth* is the range of sinusoidal input frequencies that can be measured without significant distortion or attenuation. *Transport delay* is the time required for data to pass through the system and become available for use.

SCLERAL SEARCH COIL

The scleral search coil technique, first described by Robinson (7), requires that a sensing element be placed on the eye. The technique is based on the principle that a changing electric field can induce a current in a coil of wire. If the coil lies in a plane parallel to a uniform, alternating current (ac) magnetic field, no current is induced. If the plane of the coil is not parallel to the field lines, an ac current *will* be induced in the coil. Current amplitude will be proportional to the coil area projected onto the plane that is perpendicular to the magnetic field lines. For example, if the plane of the coil is tilted about an axis perpendicular to the magnetic field lines, the induced current will be proportional to the sine of the tilt angle. A tilt in the opposite direction results in an induced current with the opposite phase (180° phase shift). The sign of the tilt angle can, therefore, be deduced from the phase of the induced current.

As shown in Fig. 2, a pair of Helmholz coils, which set up uniform ac magnetic fields in both vertical and horizontal axes, surrounds the subject's head. The driving circuitry ensures that the two fields are exactly 90° out of phase. An induction coil, made of very fine wire, is held on the eye so

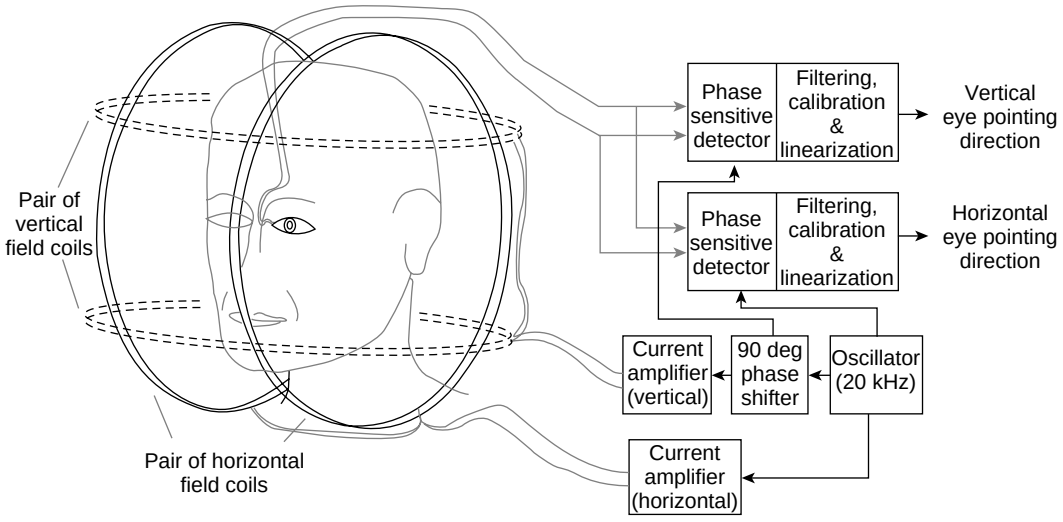


Figure 2. Schematic illustrating the scleral search coil method of measuring eye movement. (Adapted from a diagram in Ref. 8.)

that it forms a circle about the limbus (iris–sclera boundary). Robinson embedded the coil in a scleral contact lens, which was held to the limbus by the action of a tiny suction tube. Collewijn et al. (9) developed a technique for using an annular ring made of silicone rubber and having a slightly hollow inner surface. The ring adheres to the limbic area by capillary action and does not require the suction tube. Fine wire leads, from the induction coil, extend out of the eye at the canthus (the corner of the eye). A drop of anesthetic is generally administered prior to insertion, but the devise is usually tolerated well once the anesthetic wears off (9).

The induction coil encloses an area that is approximately in the plane of the pupil. Horizontal eye movement varies the current induced by the horizontal ac magnetic field, and vertical motion varies the current induced by the vertical field. By detecting the phase, as well as the amplitude of the induced current, it is possible to obtain separate analog signals proportional to the sine of vertical and horizontal eye rotations. A simple calibration is required to find the initial reference orientation of the eye.

It is possible to embed, in the annular ring, a second induction coil that encloses an area having a component parallel to the optical axis of the eye. The second coil is shown in Fig. 3. Torsional eye movement can be computed from the current induced in this second coil.

When used with nonhuman primates, a magnetic induction coil is often implanted surgically, using a method described by Judge et al. (10).

Scleral search coil systems can be expected to measure with a resolution > 1 arc·min over a range of ± 15–20°, and accuracy of $\sim$ 1–2% of the range. Slippage of the annular ring on the eyeball is possible, and can produces significant additional error. Temporal bandwidth is a function of the coil excitation frequency and filtering, and depends on the specific implementation, but 0–200 Hz or better is probably achievable. The accuracy of the torsional measurement is generally assumed to be a fraction of a degree, but may be affected by slippage of the annulus as well as variation in eyeball curvature, and is not well documented.

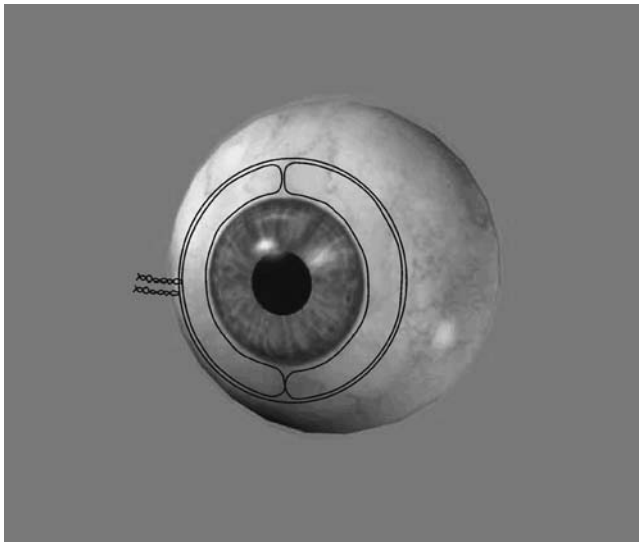


Figure 3. Schematic showing the configuration of the second coil in a dual induction coil system. The second coil forms a loop whose enclosed area has a component parallel to the optical axis of the eye (perpendicular to the plane of the pupil), and is used to measure torsional movement. Although the diagram shows only one winding for each coil, the actual device uses multiple windings.

The method is distinctly invasive, and requires the head to be confined within the Helmholz coil assembly, but does not require the head to be rigidly fixed. It offers good measurement performance, measures all three axes of rotation simultaneously, and is not affected by eyelid closure, or ambient illumination. In the past, complete systems for use with humans have been commercially available from C-N-C Engineering, Seattle, WA; and Skalar Medical BV (11), The Netherlands. Current commercial availability is uncertain. Systems for use with animals are available from Riverbend Instruments, Inc., Birmingham, AL (12).

ELECTRO-OCULOGRAPHY

Electro-oculography (EOG) has a relatively long history, dating from the 1920s and 1930s (13–17). The retina of the eye carries a slightly negative electrical charge, varying from $\sim$ 0.4 to 1.0 mV, with respect to the cornea, probably because the retina has a higher metabolic rate. This charge difference constitutes an electrical dipole, which is approximately, although not exactly, aligned with the optical axis of the eye. Electro oculography refers to the use of surface skin electrodes to measure the position of the cornea-retinal dipole. When used with ac recording techniques to measure nystagmus, or when used in a neurological test setting to measure any type of eye movement, it is often called electronystagmography (ENG).

Ideally, when the electrical dipole is midway between two electrodes that have been placed near the eye, the differential voltage between the electrodes would be zero, and as the eye rotates from this position, the differential voltage would increase with the sine of the angle. Although this is indeed the qualitative result, in practice there is a great deal of direct current (dc) drift. Skin conductance varies over time, and the corneo-retinal potential changes with light adaptation, alertness, and the diurnal cycle. In fact, EOG is sometimes used explicitly to measure the changes in corneo-retinal potential as a function of light stimuli, rather than eye movement (18,19). Electromyographic activity from facial muscles can also interfere with EOG measurement.

After cleaning the skin with an alcohol swab, electrodes are often placed as shown in Fig. 4. In this case, the differential voltage between the electrodes placed near the outer canthi (junction of upper and lower eyelid) of each eye is used to measure horizontal motion of the two eyes together. It is also possible to use additional electrodes near the nose, or on the bridge of the nose, to measure the two horizontal eye positions independently. The vertically positioned electrode pairs measure both eyes together when wired as shown in the diagram, but can also be used to measure vertical position of each eye, independently. The electrode at the center of the forehead is used as a reference. Other placement patterns can be used as well.

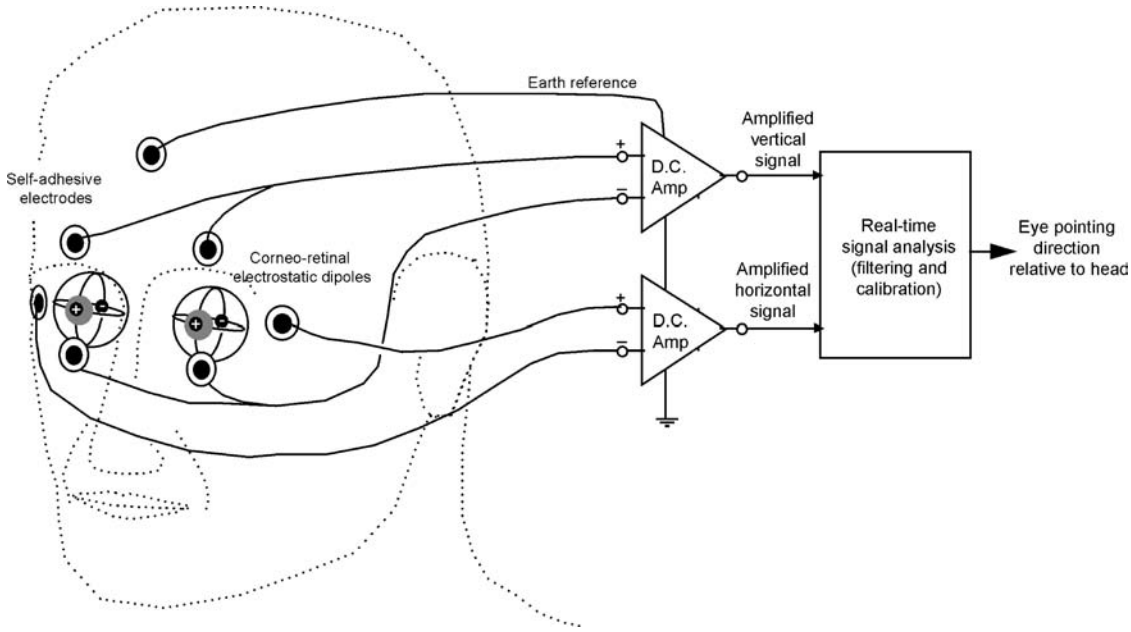


Figure 4. Schematic illustrating the EOG method of measuring eye movement. (Adapted from a diagram in Ref. 8.)

The electrodes most often used are silver–silver chloride, self-adhesive models, designed as infant heart monitor electrodes. They are connected to high gain, low impedance, low noise, differential amplifiers. The output from the differential amplifiers is now most commonly digitized and input to digital processors for linearization, scaling, and other processing. The ac coupling or frequent rezeroing is required to keep the analog signal in range of the analog to digital converters.

The high and unpredictable rate of dc drift makes this technique less suitable than others for point of gaze measurement. However, the drift is usually slower than eye movement velocities, and EOG provides an excellent means for measuring eye velocity profiles, slow and fast phases of nystagmus, and patterns of fixations and saccades, as long as the precise point of regard is unimportant. Research laboratories often assemble their own EOG devices from commercially available amplifiers, electrodes, and digital data processing software packages. Both EOG and ENG devices are sometimes used for neurological testing in clinical settings, and commercially available EOG or ENG devices are most often packaged as part of neurological testing suites.

The EOG and ENG systems are commercially available from: Cambridge Research Systems Ltd., UK (20); GN Otometrics, Denmark (21); Guymark UK Ltd., UK (22); Metrovision, Pérenchies, France (23), and Neuro Kinetics, Inc., Pittsburgh, PA (24).

OPTICAL TECHNIQUES

Noncontacting optical sensors can be used to deduce the orientation of the eyeball from the position of optical features, optically detectable geometry, or the pattern of reflectivity on the eye and the facial area surrounding the eye. The sensors may range from a small number of individual photodiodes to CCD or CMOS arrays, which provide two-dimensional (2D) gray scale image data. In most cases the eye area is illuminated with a near infrared (IR) light that is within the sensitive spectral region for solid-state light sensors, but minimally visible to the human eye. Optics may be mounted to head gear and move with the head, the head may be restrained to prevent or limit motion with respect to optics that are not head mounted, or movement with respect to non-head-mounted optics may be allowed. In the latter case the sensor field of view must either be large enough to accommodate the expected head movement, or some component of the optical assembly must automatically move to keep the sensor aimed at the eye being measured.

OPTICAL SENSORS

Sensors used by optical eye trackers include the following: *Quadrant and bicell photodetectors*: The disk shaped detector surface is divided into two (bicell) or four (quadrant) discrete photosensitive areas. The devices are configured to produce analog signals (one for bicell detectors, and two for quadrant detectors) proportional to the difference in light intensity sensed on adjacent areas. The signals are monotonically related to small displacements of a light spot from the center of the detector either in one (bicell) or two (quadrant) dimensions. The light spot must remain completely on the detector surface, and displacements must be smaller than the diameter of the spot. *Lateral effect photo diodes (position sensitive detectors):* A solid-state detector provides analog information proportional to the one (1D) or two dimensional location of the incident light center of gravity. *Small arrays of discreet solid state photosensors:* The array provides a small number of

analog light intensity signals. *Large, linear, photo-sensor arrays*: The array provides gray scale image data in a single dimension. *Large, 2D, solid-state, photosensor arrays (CCD and CMOS)*: The array provides two dimensional gray scale image data. Commercially available video cameras, based on CCD and CMOS sensors provide analog and digital signals in standard formats, usually at 50 or 60 fields/second. Using a CCD chip that supports double the normal pixel output rate, a small number of cameras are available that output 120 fields/second. By using a subset of the video lines for each field, these devices can also deliver field update rates that are higher than 120 Hz. Some CMOS sensor chips allow even more flexibility to receive data from a dynamically determined subset of the pixels and to vary the update rate. Higher update rates always mean that each pixel has less time to accumulate charge, resulting in less effective sensitivity and lower signal to noise ratios.

Quadrant detectors, lateral effect photo diodes, and small arrays of discrete photo sensors provide information with low spatial bandwidth content. The information can usually be processed at high temporal bandwidth with relatively little digital processing requirement. Large linear and 2D arrays offer much richer spatial information, but require more processing power to interpret the information, often leading to reduced temporal bandwidth.

FEATURES OF THE EYE

The eye image features most often used for eye movement measurement are *Limbus:* Boundary between the colored iris and white sclera. *Iris*: "Colored" ring that opens and closes to adjust pupil size. *Pupil*: Circular opening defined by inner boundary of iris. *First Purkinje image (corneal reflection)*: Reflection of a light source from the outer surface of the cornea. *Fourth Purkinje imag*e: Reflection of a light source from the inner surface of the lens.

These features are shown schematically in Fig. 5.

The pattern of blood vessels on the retina, if imaged with an opthalmoscope or fundus camera, also constitute markings that can be used to track eye movement, but this is a less commonly used technique.

In some cases facial landmarks, such as the canthus (corner of the eye, where the upper and lower eyelid meet), another facial feature, or a dot placed on the skin, may be used to compare with the location of features on the eye. As discussed later on, facial features can also be used to guide remote camera based trackers.

The iris has a distinctive pattern of radial markings that can be used to track ocular torsion. Its inner boundary, defining the pupil, is nominally circular and centered with respect to the limbus. Detailed examination, however, reveals a slightly noncircular shape that is off center (usually toward the nasal side) with respect to the limbus. Furthermore, both the shape and position of the pupil change slightly with pupil diameter and vary across the population. The characteristics of the pupil form are described and quantified in considerable detail by Wyatt (4). He found that pupil position tends to shift in the nasal and superior directions (with respect to the limbus) as the pupil contracts, and that pupil diameter and circularity tend to decrease with age.

Light rays that enter the eye through the pupil are reflected by the retina and directed back toward their

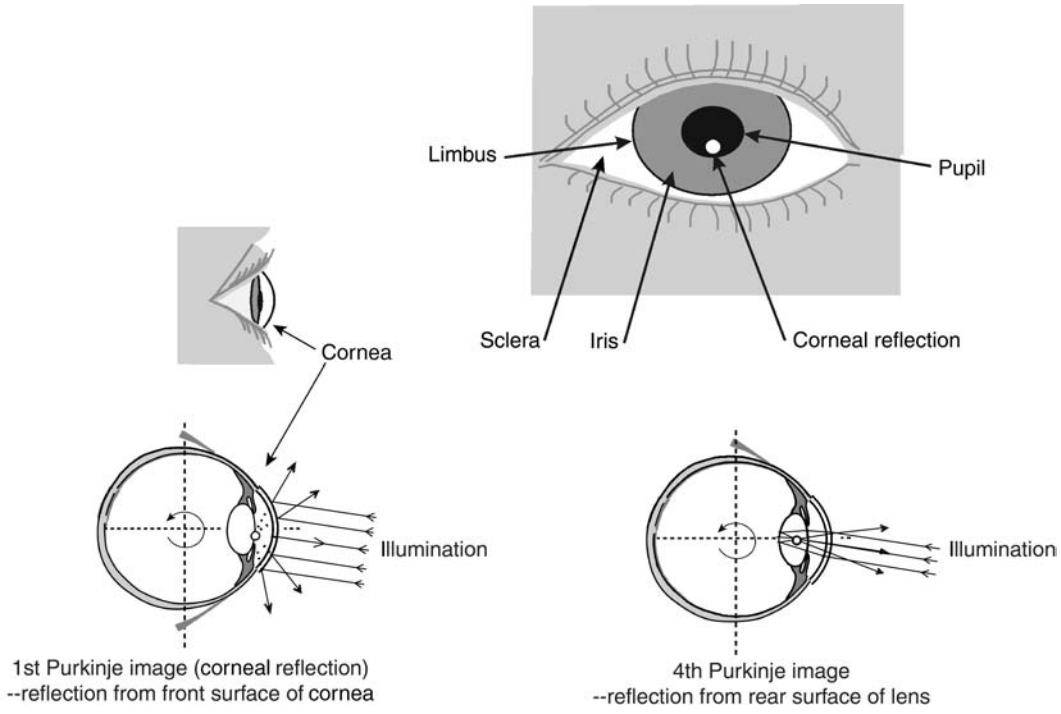


Figure 5. Schematic illustrating the various features of the eye often used by optical eye movement measurement techniques. (Adapted from a diagram in Ref. 8.)

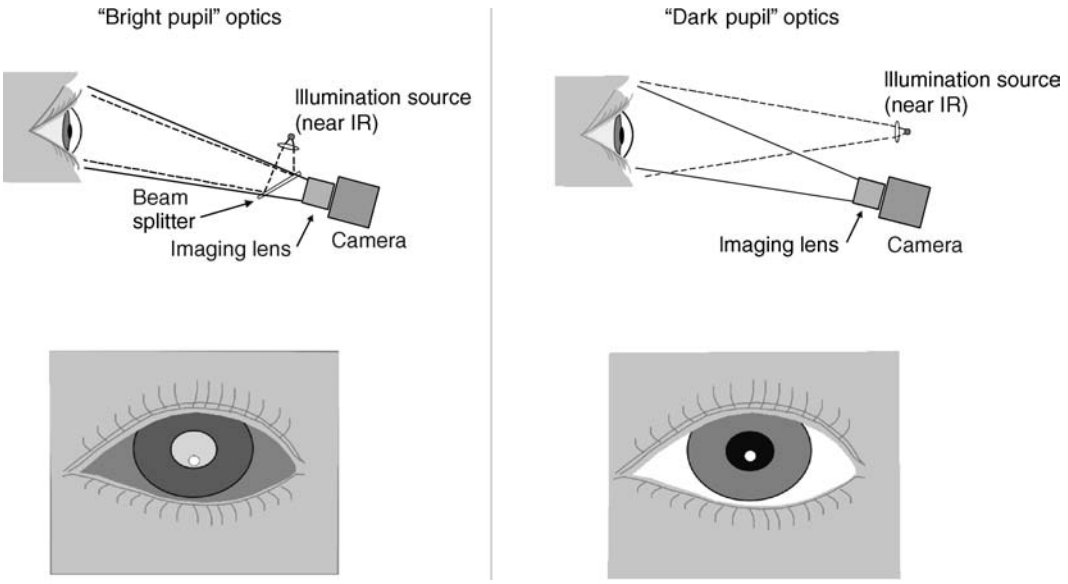


Figure 6. Illustration of bright and dark pupil optics. In the bright pupil example, retroreflected light from the retina will beam back into the camera lens, resulting in a bright, back lit, pupil image.

source. The eye therefore acts as a retroreflector. If the eye is viewed by a detector that is coaxial with an illumination beam, as shown in Fig. 6, the retroreflected light from the retina makes the pupil appear to be a bright, back lit circle. Some of the retinal reflection can be received slightly off axis from the illumination beam (25). Although the bright pupil effect falls off sharply as the detector moves off axis, this accounts for the "red eye" effect in flash photography. In an idealized case, the apparent brightness of the pupil retroreflection will vary inversely with the square of pupil diameter (26). As shown by Nguyen et al. (25), brightness of the retinal reflection also varies between individuals and as a function of gaze direction with respect to the illumination source, but these variations are small compared to the effect of pupil diameter.

If the detector is off axis from the illumination beam, the retroreflected light does not enter the detector and the pupil appears as the familiar dark circle.

The corneal reflection is a virtual image that appears to be just behind the plane of the pupil. If the cornea is assumed to be spherical, the corneal reflection image will form at point half way between the surface of the cornea and its center of curvature, along a ray that is parallel to the illumination beam and passing through the corneal center of curvature. If the light source is far away compared to the radius of the eyeball, then as the eye rotates, the corneal reflection always appears to move the same amount as the corneal center of curvature. A similar analysis will show that the fourth Purkinje image forms in almost the same plane as the corneal reflection, but appears to move the same amount as the posterior lens surface center of curvature.

FEATURE RECOGNITION

Systems that use 2D detector arrays typically perform a pattern recognition task to identify features in the 2D image. Digital processing power has always been a critical limitation. When video cameras were first used as sensors, analog preprocessing was often used to find edge points, or other pixel subgroups, and thereby reduce the amount of data that needed to be processed digitally. Algorithms requiring relatively few computational steps were used to process the reduced digital information in order to recognize features and find their centers.

Increased digital processing capability has very significantly eased, although by no means eliminated, this limitation. It is now practical to digitally process 2D, gray scale image buffers while maintaining reasonable, real-time update rates. It has become possible to use, in real-time, elements of classical digital image processing such as Sobel or other convolution based edge detection algorithms, and circle or ellipse best-fit algorithms. Less computationally intensive algorithms are still often used, however, to maximize update rates.

Digital processing components in current use range from commercially available PCs and frame grabbers to custom processing boards that include field programmable gate arrays (FPGAs) and digital signal processors (DSPs), and microcontrollers.

When using an array sensor, the location of individual boundary points on an image feature (e. g., the pupil) are often identified with only single pixel resolution. If the object covers multiple pixels, knowledge of the object shape allows its position to be computed with subpixel resolution. For example, if a group of pixels are thought to define the edge of circular object, a least mean squared error circle fit will define the circle center location with sub pixel resolution. If sufficient computation time is available, it is also possible to use gray scale information to define edge points with subpixel accuracy.

Pupil Recognition

The pupil is generally recognized as a circular or elliptical area that is darker (in the case of a dark pupil image) or

brighter (in the case of a bright pupil image) than surrounding features. The pupil is often partially occluded by eyelids and corneal reflections, and algorithms must therefore recognize the circular or elliptical shape even when occluded to some degree. Furthermore, it is important that only the real pupil boundaries be used to determine the center of the object, rather than the occlusion boundaries. Examples of pupil detection algorithms can be found in Zhu (27), Mulligan (28), Ohno et al. (29), Charlier et al. (30), and Sheena (31).

The retroreflective property of the eye makes possible a signal enhancement technique that can be exploited to help identify the pupil (32–35). If a camera is equipped with both a coaxial and off-axis illuminator, a bright pupil image will be produced when only the coaxial illuminator is on, and a dark pupil image will be produced when only the off-axis source is on. If the illuminators are alternately activated for sequential camera images, the result will be alternating bright and dark pupil images. Assuming elements in the camera field of view have not moved between images, all other image features, which are not retroreflectors, will remain essentially unchanged.

If two such images are subtracted, one from the other, the result should leave only the pupil image. In practice, there will still be small differences in all parts of the image due to the different illumination angles, but contrast between the retroreflective pupil and the rest of the image is still greatly enhanced. There are some drawbacks to the technique. If the eye moves significantly between the sequential images, subtraction results in a distorted pupil image. The need to digitize and subtract two images increases memory and processing requirements, and the fact that two sequential images are required to create one data sample limits the temporal bandwidth of the measurement. Note that a similar result can be obtained with images from two cameras that are carefully aligned with respect to the same image plane, and an illumination source that is coaxial with only one of them.

Corneal Reflection Recognition

The corneal reflection (first Purkinje image) is usually recognized, within a larger 2D image, by its intense brightness, predictable size and shape, and proximity to the pupil. It can, however, be confused with small reflections from tear ducts or eyeglass frames, and reflections from external sources emitting light in same spectral band as the intended source. Its center can be identified with subpixel accuracy only if it is large enough to cover multiple pixels. Eizenman et al. (36) describe a technique for using knowledge of brightness pattern across the corneal reflection (first Purkinjie image) to find its position, on a linear array, with subpixel resolution. To accomplish this they used a precisely designed illumination source to insure a known pattern of luminance.

Fourth Purkinje Image Recognition

The fourth Purkinje image is very dim and very difficult to reliably identify in a larger image. The one system in common use that requires the fourth Purkinje image relies on careful initial alignment to focus the image on a quadrant detector. The detector is not much larger than the Purkinje image; and, in this case, automatic recognition in a larger field is not necessary.

Face Recognition

In some cases wide-angle images are now used to recognize the presence of a face, and to find the location of one or both eyes. If the eye is identified in a wide angle image, and assuming the camera is well calibrated to account for lens or sensor distortions, it is reasonably straight forward to compute the direction of the line extending from the camera to the eye. Finding the distance to the eye is more difficult. If the eye is identified on the images from two cameras, it is possible to triangulate. If both eyes are identified on a single camera image, knowledge of the true interpupillary distance can be used to compute distance to the face. However, head rotation with respect to the camera will cause some error if not taken into account. Some face recognition systems are able to determine head orientation. For example, Xiao et al. (37) and Matthew and Baker (38) describe a method, based on a technique known as active appearance modeling (39), to make real-time measurements of the position and 3D orientation of a face.

Information about eye location can be used to direct a separate sensor to obtain a more magnified view of the eye, or the wide-angle image itself may also be used to find the position of features within the eye. In the latter case, there is a clear trade off between spatial resolution and wide-angle coverage. Recognition of facial features can also be exploited in order to use a facial landmark, such as the canthus or the center of the eyeball as one of the elements in a dual feature-tracking algorithm (40,41). However, the plasticity of facial features makes it difficult to determine their position with the same precision as the pupil and Purkinjie images.

EYE ORIENTATION AS A FUNCTION OF SINGLE OR DUAL FEATURE POSITION

If we have a sphere of radius r, with a mark on its outer surface as shown in Fig. 7, and if we assume the center of the sphere is fixed with respect to an observer, then the observer can compute the rotation (θ) of the sphere about its center by noting the displacement (d) of the surface mark.

$$\theta = \arcsin(d/r)$$

This is the principle behind single feature eye tracking. However, if the center of the sphere moves with respect to the observer, observation of a single mark provides no way to distinguish such motion from rotation.

If there are two visible marks on the sphere, which are fixed to the sphere at different distances from its center, then observing the relative position of these marks does allow rotation of the sphere to be distinguished from translations. So long as the distance between the sphere and observer remains the same or is independently known, translation with respect to the observer can be unambiguously distinguished from rotation.

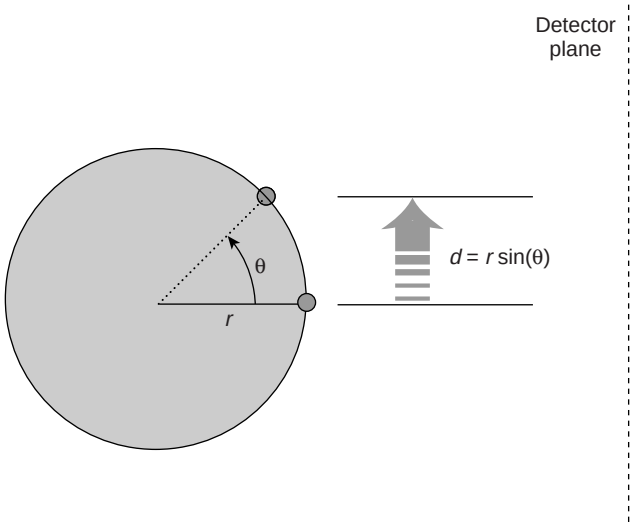


Figure 7. Schematic illustrating the principle behind single feature tracking. Observation of a single land mark on the sphere allows determination of its orientation, so long as the center of the sphere does not move.

If the two marks are located along the same radius line, at distances r_1 and r_2 from the center of the sphere, as shown in Fig. 8, then the rotation angle (θ) of this radius line, with respect to the line connecting the sphere and observer, is

$$\theta = \arcsin(\Delta d/(r_1 r_2))$$

where Δd is the observed separation between the marks. This is the basic principle behind dual feature tracking techniques. The underlying sine function has a steep slope at small angles, maximizing the sensitivity of the technique. Note that if the distance between the observer and sphere changes, the dual feature relation still leaves some ambiguity since the apparent separation between the marks will change as a function of distance.

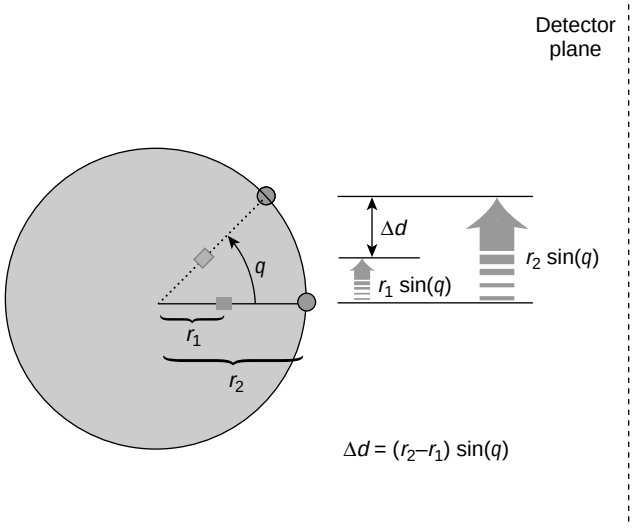


Figure 8. Schematic illustrating the principle behind dual feature tracking. The relative position of the two landmarks defines the orientation of the sphere, even if its center moves in a direction parallel to the plane of the detector.

In the case of the eye, referring to Fig. 1, the center of the pupil or center of the iris is a mark that is ~ 9.8 mm from the center of the eyeball. Although both images appear to be just behind the plane of the pupil, the first Purkinje image moves the same amount, with eye rotation, as would a mark ~ 5.6 mm from the center of the eyeball (the position of the anterior corneal surface center of curvature); and the fourth Purkinje image moves the same amount as would a mark ~ 11.5 mm from the center of the eyeball (the position of the posterior lens surface center of curvature).

Eye Orientation as a Function of Just Pupil or Corneal Reflection Position

The pupil, and corneal reflection (first Purkinje image) are the features most commonly used for single-feature tracking. If a sensor and light source are mounted so that they do not move with respect to a person's head, either of these features can be used as a marker to compute eye rotation in the eye socket. Either the head must be stabilized with some type of head restraint mechanism, or the sensor must be mounted to head gear that moves with the subject.

When measuring only a single feature, any translation of the head with respect to the sensor will be erroneously interpreted as eye rotation. For example, if using pupil position, a 1 mm motion of the head parallel to the detector image plane may be mistaken for about a 5° eye rotation. If the corneal reflection is the feature being tracked, a 1 mm slippage will be indistinguishable from an $\sim 12^\circ$ eye rotation.

Eye Orientation as a Function of Relative Pupil and Corneal Reflection Positions (CR/Pupil)

The pupil and first Purkinje image (corneal reflection), or the first and fourth Purkinje image are the most commonly used feature pairs for dual feature tracking. The pupil to corneal reflection technique (CR/Pupil) was first described by Merchant et al. (42). As shown in Fig. 9, if the sensor is close to the light source that produces the corneal reflection, the angle (θ) of the eye optical axis with respect to the sensor is described by

$$\theta = \arcsin(d/k)$$

where d is the apparent distance between the pupil and corneal reflection (from the point of view of the sensor), and k is the distance from the pupil to the corneal center of curvature. If the sensor is not close to the illumination source, the relation changes only slightly so long as the sensor does not move with respect to the illumination source.

$$\theta = \arcsin((d - k_d)/k)$$

$$k_d = k_{\mathrm{cr}} \sin(\gamma)$$

where k_{cr} is half the cornea radius of curvature and γ is the angle between the illumination beam and the sensor line of sight (see Fig. 10). If the sensor and illumination source are very far away, compared to the radius of the eyeball, then γ,

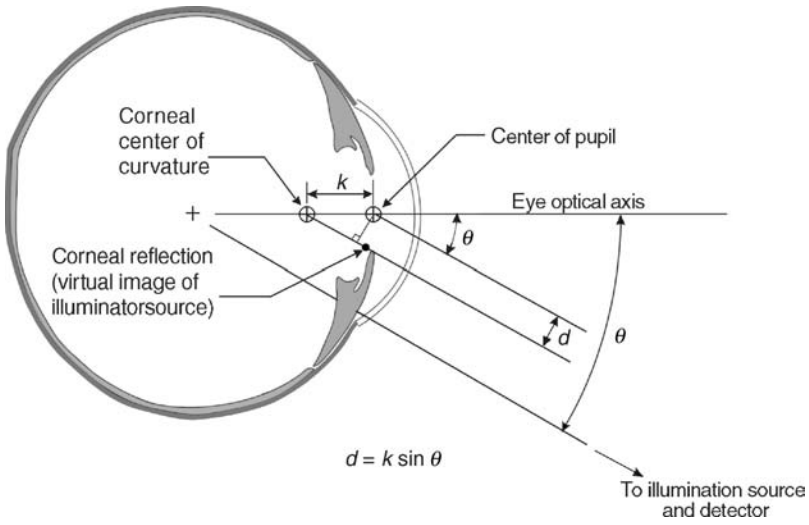


Figure 9. Schematic illustrating the basic relationship behind the pupil-to-corneal-reflection technique for measurement eye movement. The diagram assumes that the detector and illumination source are coaxial, or very close together, and are very far from the eye, compared to the radius of the eyeball. The optical axis of the eye is rotated away from the detector by an angle θ. From the vantage point of the detector, the pupil and corneal reflection appear to be separated by distance d.

and hence k_d, are constants. One drawback to the pupil to corneal reflection technique is that the pupil center is not completely stable with respect to the optical or visual axis, but moves slightly as the pupil size changes (4). Theoretically, this effect can be measured in an individual and accounted for (43,44), but would add to the time and effort required to calibrate an individual. Note also that the pupil-to-corneal-reflection vector is less sensitive to eye rotation than either of the individual features. A 5° eye rotation, from an initial position in which the pupil and corneal reflections are aligned, causes the pupil and corneal reflection images to separate by only ~ 0.4 mm, whereas the pupil center moves ~ 1 mm.

The equations given above describe the major effect, but not all secondary effects, and are not precise. For example, the cornea is not perfectly spherical, the eyeball does not rotate about a perfectly stable central point, and the pupil image is slightly magnified by the refractive power of the cornea. To the extent that secondary effects are large enough to be detected, they are often accounted for by the results of an empirical calibration procedure, as discussed later.

The range of gaze angles that can be measured by the pupil to corneal reflection technique is limited by the range over which the corneal reflection remains visible to the detector. In the horizontal axis, this range is usually $\sim 70^\circ$ visual angle for a given illumination source and sensor pair. It is usually less in the vertical axis due to occlusion of either the pupil or corneal reflection by the eyelids. The range can be extended by using multiple illumination sources, at different positions, to create multiple corneal reflections, but the system must be able to uniquely recognize each reflection even when not all are visible.

Eye Orientation as a Function of Relative First and Fourth Purkinje Image Positions (CR/4PI)

As described by Cornsweet and Crane (45), the first and fourth Purkinje images can also be used for dual feature tracking. The same type of arcsine relation applies, but with d the apparent distance between the two Purkinje images, and with k equal to the distance between the corneal center of curvature and the posterior lens surface center of curvature. The technique has the advantage that the Purkinje image positions can be more precisely defined than the pupil center. In addition, the separation of the posterior lens surface and corneal centers of curvature (~ 6 mm) is greater than that between the pupil and corneal centers of curvature (~ 4.2 mm), yielding greater

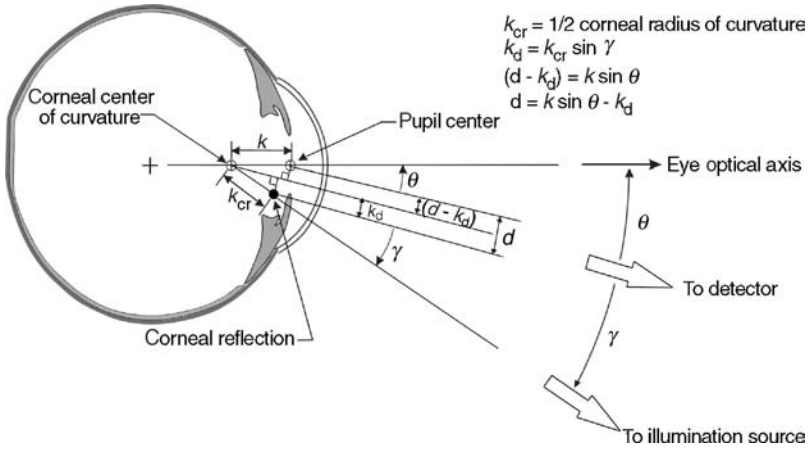


Figure 10. Schematic illustrating the detected pupil to corneal reflection separation (d) when the detector and illuminator optical paths are not coaxial. The diagram assumes that the detector and illumination source are far from the eye, so that lines from the detector to various points on the eye are essentially parallel, as are rays from the illumination source to various points on the eye. The angle θ is between the eye and detector optical axes, and γ is the angle between the detector optical axis and the illumination beam.

sensitivity to eye rotation than the pupil to corneal reflection technique. Drawbacks are that the fourth Purkinje image is relatively dim and difficult to find, and is only visible within the iris opening.

Eye Orientation as a Function of Relative Pupil or Iris and Facial Landmark Positions

The relative position of the pupil or iris and a facial landmark can also be used to measure eye orientation. This is sometimes done when image magnification and resolution is such that the smaller Purkinje images cannot be reliably detected, or the relative motion of the pupil and corneal reflection cannot be sufficiently resolved. Since the facial landmark does not move with eye rotation, the governing relationship is described by

$$\theta = \arcsin(\ (d - d_i)/r)$$

where d is the distance (in the camera image plane) from the facial landmark to the pupil or iris center, d_i is the distance (also in the camera image plane) from the facial landmark to the eye center of rotation, r is the distance from the pupil or iris center to the eyeball center of rotation, and θ is the angle of the eye optical axis with respect to the detector. If the head rotates with respect to the detector, d_i will appear to shorten, in the direction perpendicular to the rotation axis, by an amount proportional to the cosine of the rotation angle. Advantages over other dual feature techniques are greater sensitivity (larger change of the measured quantity for a given eye rotation), and the possibility of identifying the features in wider, less magnified images. Disadvantages are that facial image features are not stable, but usually move at least slightly as a function of facial muscle activity, and head orientation must also be measured or must remain stable with respect to the detector. Zhu and Yang (40) describe a dual feature technique using the relative position of the canthus and iris; and Tomono et al. (41) describe an algorithm for using the relative position of the pupil, and a computed position of the center of the eyeball.

EYE ORIENTATION AS A FUNCTION OF FEATURE SHAPE

It is also possible to extract eye orientation information from feature shape. A circle (e. g., the pupil outline or outer boundary of the iris) appears elliptical if viewed from an angle. As the circle tilts, the minor axis of the ellipse, which is perpendicular to the axis of rotation, appears shortened by the cosine of the rotation angle. A major drawback to using pupil or iris ellipticity as an orientation measure is that, due to symmetry considerations, the direction of rotation about the rotation axis remains ambiguous. A second major limitation is that the underlying cosine function has a shallow slope at small angles resulting in poor sensitivity.

EYE ORIENTATION AS A FUNCTION OF REFLECTIVITY PATTERN MOVEMENT

Different structures on the eye, primarily the pupil, iris, and sclera have different reflectivity properties. As the eye rotates this reflectivity pattern moves, and that property can be exploited to measure eye movement. Movement of the reflectivity pattern can be detected with small numbers of individual photodetectors, and this, in turn, makes it possible to achieve relatively high temporal bandwidth. The technique was pioneered by Torok et al. (46) and Smith and Warter (47), using a photomultiplier as the detector, and further developed by Stark and Sandberg (48), Wheeless et al. (49), Young (50), Findlay (51) and Reulen et al. (52). The most prominent contrast feature in the pattern is generally produced by the boundary between the iris and sclera. Therefore, devices that use small arrays of photodetectors to measure motion of this pattern are often called limbus trackers.

The iris sclera boundary is easily visible along the horizontal axis, but along the vertical axis it is usually obscured by the eyelids. In fact the boundary between the eyelids and iris are often the most prominent reflectivity boundaries along the vertical axis. The eyelids do tend to move in proportion to vertical eye motion, and are useful as measures of vertical eye position; but motion of the reflectivity pattern remains a much less dependable function of vertical (as opposed to horizontal) eye rotation.

In principle, reflectivity pattern tracking is similar to single feature tracking. As with single feature tracking, any movement of the sensors with respect to the eye produces erroneous measurements.

MEASURING POINT-OF-GAZE IN THE PRESENCE OF HEAD MOTION

Dual feature techniques, such as the pupil to corneal reflection method, permit computation of gaze direction with respect to a detector. However, head motion may still need to be measured in order to accurately determine the point of gaze on other objects in the environment.

First, consider an example in which the head is free to move with respect to a stationary detector, and the task is to measure point of gaze on other stationary objects in the environment. This is illustrated by the two dimensional example in Fig. 11. The point of gaze, defined by x, is dependent not only on θ, which can be measured by one of the dual feature tracking methods previously described, but also on ϕ and d_1, which define head position. If head motion is small compared to distance to the detector and scene surface, then changes in head position will have little effect and can be ignored.

If the location of the eye in the environment space can be independently measured, and if the detector and scene surface positions are known with respect to the same environment space, the following general algorithm can be used to determine point of gaze on a surface. Use a dual feature technique to determine direction of the gaze vector with respect to the detector, and knowledge of the detector orientation to express this as a direction in the environment space. Use the gaze direction and known start point (the location of the eye in space) to write the parametric equation for a line in the environment coordinate space. Use knowledge of the scene surface position to solve for the intersection of a line and a plane. Ohno et al. (29) describe a version of this strategy.

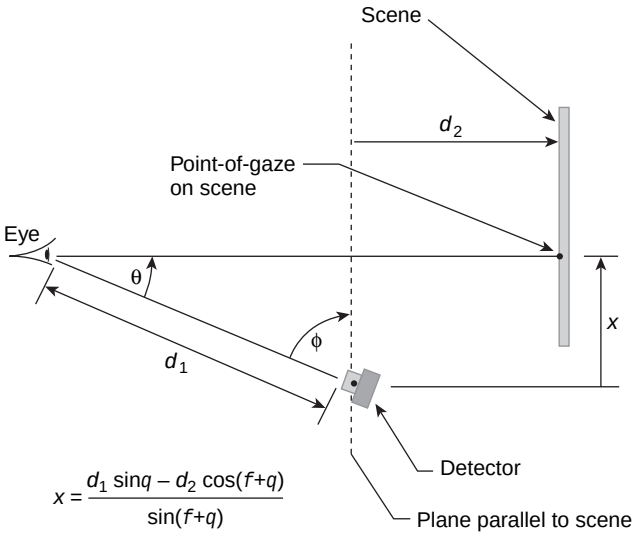


Figure 11. Relationship, in one plane, between point-of-gaze on a flat scene and relative eye, detector, and scene positions. (From Ref. 26.)

Next, consider an example in which the detector is fastened to headgear worn by the subject. Single as well as dual feature techniques may be adequate to measure gaze direction with respect to the head (although single feature methods will have larger errors if there is any slippage of the headgear). In order to find point-of-gaze on objects in the environment it is clearly necessary to know the position and orientation of the head with respect to those objects. The following strategies can be used.

1. A second camera can be mounted to the head gear so that it shares the same reference frame as the eye sensor, but points toward the subject's field of view. Point-of-gaze can be indicated as a cursor superimposed on the image from this scene camera.
2. Light-emitting diodes (LEDs) or other special emitters can be fastened to an object in the environment, such as the bezel of a computer monitor, and detected by head mounted sensors to locate the object in the head reference frame. Measurement of gaze in the head reference frame can then be related to position on that object.
3. A separate head tracking system can be used to measure head position and orientation with respect to the environment. This information can then be used to compute the location and direction of the gaze vector in the environment coordinate frame. If the locations of surfaces are also known in the same environment reference frame, it is possible to solve for the intersection of a line (line-of-gaze) with a surface, to find point-of-gaze on various surfaces in the environment. A general method for doing this computation is described in Appendix E of Leger et al. (8). Duchowski (53) also describes specific algorithms for handling this type of task.

In the case of the head mounted scene camera described above, it is important to be aware of possible parallax error. The scene camera is usually not viewing the scene from exactly the same vantage point as the eye being tracked. As shown in Fig. 12, eye rotation angle data can be mapped to the scene camera image plane at a particular image plane distance, but the relation changes as the image plane distance changes. The resulting parallax error can be easily corrected if there is knowledge of the distance, from he subject to the gaze point. The parallax error may be negligible if the distance to the gaze point is large compared to the distance of the scene camera from the eye. It is also possible, as shown in Fig. 13, to minimize parallax by bending the scene camera optical path with a beam splitter, such that the scene camera has the same vantage point as the eye being measured (54).

If a person with normal ocular function is fixating a point not infinitely far away, the lines of gaze from the two eyes should converge. If the gaze angles of both eyes are measured with head mounted optics, the intersection of the two lines-of-gaze theoretically indicates the three-dimensional (3D) point-of-gaze in space, with respect to a head fixed coordinate system. If head position and orientation are known, this can be transformed to a position in environment space. Duchowski (53) describes an algorithm for this computation. It should be noted that, in practice, the measured lines-of-gaze from the two eyes will almost

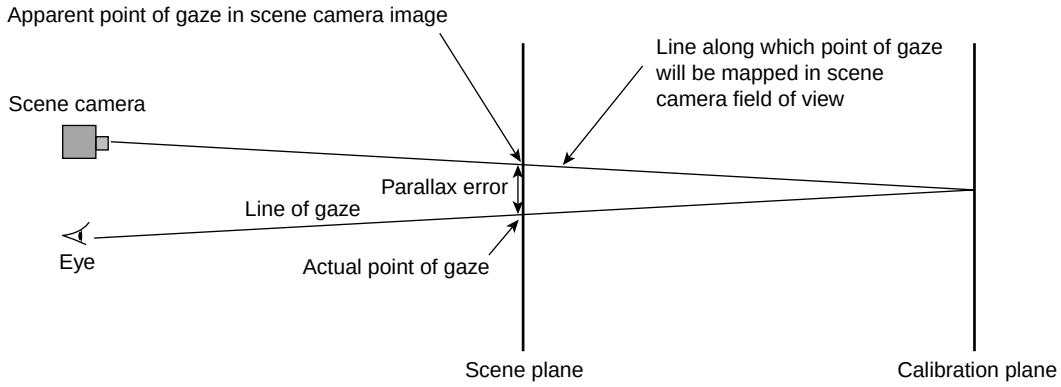


Figure 12. Parallax error when gaze direction is mapped (calibrated) to a scene camera image plane that is different from the plane being fixated.

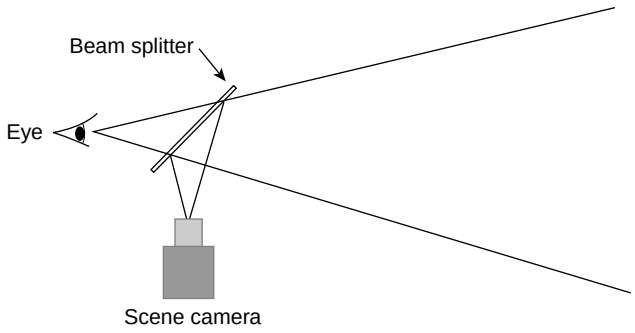


Figure 13. Use of a beam splitter to minimize scene camera parallax.

never intersect. Neither the measurement nor the system being measured is infinitely precise. The discrepancy can be resolved as follows. Consider the two planes that are parallel to the "vertical" axis of the head and which contain the gaze vector from each of the two eyes. If the two lines-of-gaze converge, these two planes will intersect along a line, which is also intersected by the line-of-gaze from each eye. The 3D point-of-gaze can be chosen to be either of these intersection points or a point half way between the two.

As the point-of-gaze moves farther from the head, the vergence angle (the angle formed by the gaze vector from each eye) diminishes. The relation is reasonably sensitive at close distances. As distances become longer, a very small change in vergence corresponds to an increasingly large change in distance, and moderate measurement noise or error in the eye tracker may result in a very noisy or inaccurate point of gaze computation.

MEASUREMENT OF TORSIONAL EYE MOVEMENT

Optical measurement of torsional eye movement was first accomplished by offline photographic analysis techniques and later by tracking artificial marks placed on the eye. For example, Edelmann (55) created a visible mark by sandwiching a human hair between two contact lenses. Use of standard surgical markers, applied just outside the limbus, has also been reported (56), although this requires that a local anesthetic be applied to the cornea.

Over the last decade or so it has become practical to make real time measurements using the patterns that are naturally visible on the iris, as captured by CCD or CMOS cameras. This has generally been done either by using a cross-correlation technique, or a template-matching scheme to compare the iris over sequential video frames. The first method, first described by Hatamian and Anderson (57), and further developed and automated by Clarke et al. (58) and Bucher et al. (59), cross-correlates the pixel sequence from a 1 pixel wide path around the pupil, sampled from each video frame, with that from an initial reference frame.

It is important that the same strip of iris be sampled each time, so unless the eye is stationary in the two nontorsional degrees of freedom, the pupil center must be accurately tracked. Even so, at eccentric eye positions, geometric image distortions can cause errors. The points on the iris that form a circle on the camera image plane when the subject looks directly at the camera, begin to form a more elliptical shape on the camera image plane as the subject looks farther away from the camera. Moore et al. (60), and Peterka et al. (61) describe a method for avoiding this type of error by correctly computing the projection of the eye onto the camera image plane and empirically solving for parameters that correspond to physical characteristics of the eye or eye-to-camera geometry. A template-matching scheme described by Groen (62) and further developed by Zhu et al. (63) is also designed to minimize geometrical perspective errors by tracking distinctive landmarks on the iris. Changes in pupil diameter can affect the pattern of radial markings on the iris and lead to some torsion measurement error. Guillemant et al. (64) describes a technique using neural network software to identify the pupil and iral patterns for torsion measurement.

CALIBRATION

Most eye tracking systems require a practical method to relate a measured quantity, such as the relative position of the pupil and corneal reflection, to a desired quantity, such as point of gaze on a particular scene space. The underlying relationships behind several techniques for measuring eye orientation have been presented in preceding sections. In practice, however, eye tracking systems often rely on completely empirical techniques to map the measured quantity to gaze points on a scene space. The measured quantity is recorded as a subject looks at several known points in the scene space and either a polynomial curve fit, an interpolation scheme, or some combination is used to map (transform) one to the other.

The process of gathering data to compute the transform is referred to as the calibration. In this way, the precise physical dimensions, such as the corneal radius of curvature or angle between the optical and visual axis of the eye, and precise geometrical relationships between detectors, illumination sources and scene surfaces do not have to be explicitly determined. Rather, these relations are automatically incorporated in the implicitly determined function.

Theoretically, the calibration transformation can remove any systematic error that is a function of the measured variables. More calibration data points allow higher order polynomial transforms or more interpolation points, and usually improve the result, but with diminishing returns. Too many calibration points also result in a time consuming and onerous procedure. Systems often require subjects to look at either five or nine target points, and rarely > 20. In some cases, precise knowledge of the geometrical relation between components, along with knowledge of the underlying mechanism, can be used to reduce the number or calibration points required while preserving accuracy. Ohno et al. (29) describes a scheme using this type of strategy.

A cascaded polynomial curve fit scheme, used with pupil to corneal reflection method eye trackers, is described by Sheena and Borah (43). The same paper describes a method to account for changes in pupil position associated with change in pupil diameter. A 2D interpolation scheme is

described by McConkie (65), and Kliegle and Olson (66). Possible variations are unlimited, and available systems employ a wide variety of calibration schemes. Sometimes, in order to further reduce systematic error, the users of commercially produced eye trackers add their own calibration and transform process onto data that has already been processed by the manufacturer's calibration and transform scheme. Jacob (67) and Duchowsky (53) describe specific examples.

Amir et al. (68) describe a method for using two cameras and illumination sources to compute the location of the eye optical axis with no calibration requirement. The pupil and corneal reflection images, on each camera, can be used to determine a plane that must contain the eye optical axis. The intersection of the two planes, one computed from each camera image, defines the optical axis. The orientation of the visual axis relative to the optical axis of the eye varies across the population, however, and this uncertainty cannot be removed without requiring at least one calibration point.

If eye tracker optics are head mounted, the calibration transform is often designed to map gaze to the image plane of a head mounted scene camera or a similar imaginary plane that travels with the head. Head position and orientation measurements can then be combined with this result, as described in the previous section, to derive the line-of-gaze in space, and to compute its intersection with known surfaces.

COMPATIBILITY WITH EYEGLASSES AND CONTACT LENSES

Eye glasses may present mechanical problems for systems that require sensors to be very close to the eye. Systems having sensors that are farther away and that view the eye through the spectacle lens must contend with mirror reflections from the spectacle lens and frame, or obstruction by elements of the frame. Distortion of the image by the spectacle lens usually does not present a significant problem since such effects are removed by the calibration scheme. The biggest reflection problem is often posed by the illumination source that is part of the eye tracking system, especially since the sensor must be sensitive in the spectral region of this light. Antireflective coatings are usually not good enough to eliminate the problem.

The position of the specular reflection (mirror image of the illumination source) is determined by the incidence angle of the illumination beam with the spectacle lens surface, and it is often possible to position the source so that the specular reflection does not cover a feature of interest, although this can become more difficult in the case of very high power (high curvature) spectacle lenses. If the specular reflection is not occluding an important feature, but is still in the sensor field of view, the system must be able to distinguish the features of interest from the specular reflection without confusion. This ability varies among systems, but video based systems can often be used successfully with eyeglasses.

Contact lenses also are often tolerated well by optical eyetracking devices, but may sometimes present the following problems. An edge of the contact lens may sometimes be visible or generate a bright reflection, and may confuse feature recognition, especially if the edge intersects a feature of interest.

When a light source reflection from the outer surface of the contact lens is visible to the detector, it usually appears to replace the first Purkinje image (corneal reflection). This is not a problem for systems that use the corneal reflection, so long as the visible corneal reflection is always the reflection from the contact lens. Although the contact lens surface will have a slightly different position and curvature than the cornea, the difference in the motion of this reflection from that of the real corneal reflection is easily accounted for by whatever calibration scheme is used. However, if the contact lens moves so that the reflection appears to fall off the edge of the contact lens and onto the cornea, there will be a shift in the computed gaze position. Hard contact lenses, which tend to be relatively small and float about on the tear film, are more likely to cause this problem than the larger soft lenses.

The contact lens surface may be less reflective than the cornea, resulting in a dimmer first Purkinje image, and making detection more difficult.

ILLUMINATION SAFETY

Most eye trackers that use optical sensors also include a means to illuminate the eye, usually with nonlaser light at the lower end of the near-infrared (IR-A) spectral region. The IR-A region spans the wavelengths between 770 and 1400 nm. To prevent harming the eye, it is important to avoid excessively heating the cornea and lens, and to avoid focusing too much energy on too small a spot on the retina.

The American Conference of Governmental Industrial Hygienists (ACGIH) suggests the following safety criteria for extended exposure to nonlaser, near-IR light (69). To protect the cornea and lens from thermal injury, irradiance at the eye should be no $> 10\,\mathrm{mW \cdot cm^{-2}}$. To protect the retina, near-IR radiance, expressed in units of $\mathrm{W \cdot (cm^2 \cdot sr)^{-1}}$, should be limited to no $> 0.6/\alpha$, where α is the angular subtense, in radians, of the source as seen by the subject.

Various safety standards are specified by many other organizations, including the American National Standards Institute (ANSI), The U. S. Food and Drug Administration (FDA), the International Electrotechnical Commission (IEC), and others, although some of these are intended specifically for laser sources. A comprehensive review of light safety issues can be found in a book by Sliney and Wolbarsht (70).

SPECIFIC IMPLEMENTATIONS OF OPTICAL TECHNIQUES

Photo Electric, Reflectivity Pattern (Limbus) Trackers

There are a small number of commercially available systems that use a photo electric reflectivity pattern (limbus) tracking technique to measure eye movements. Figure 14 shows a schematic for a basic system measuring horizontal eye position. The difference in the signal received by the two photodetectors is roughly proportional to the

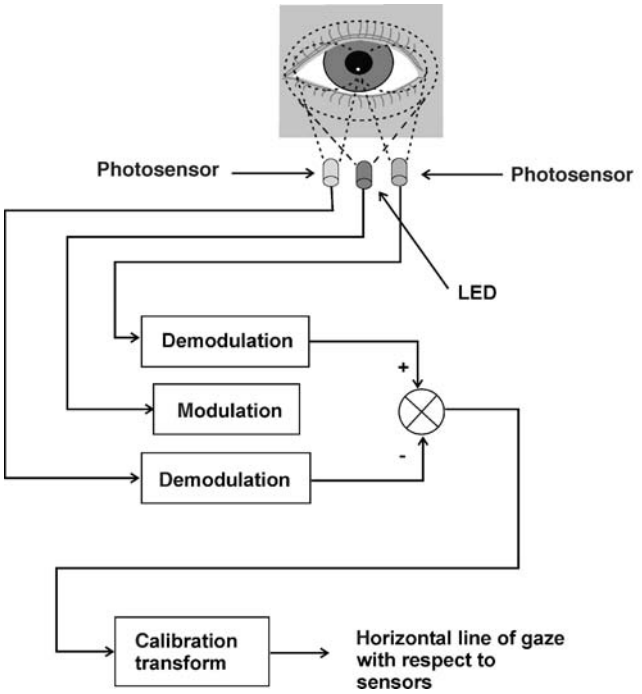


Figure 14. Schematic showing simple reflectivity pattern (limbus) tracker for horizontal measurement. (Adapted from a diagram in Ref. 8.)

horizontal position of the reflectivity pattern across the eye, with the most prominent feature of the pattern being the contrast between the white sclera and darker iris (limbus). Modulation of the LED, typically at 2 kHz or higher, and corresponding demodulation of photosensor signals diminishes the effect of ambient light. The signal is low pass filtered to remove modulation artifacts; and is either scaled and linearized with analog controls, or sampled and processed digitally, in order to scale and linearize the values.

Vertical position can be measured by orienting a similar LED and sensor array vertically, instead of horizontally. The results are less dependable because the high contrast iris to sclera boundary is often obscured by the eyelids. Alternately, vertical position is measured by aiming the horizontally oriented LED and sensor array at the boundary between the eye and the lower eyelid, and summing (instead of differencing) the photodetector signals. The result is really a measure of lower eyelid position, and takes advantage of the fact that the lower eyelid moves roughly in proportion to vertical eye motion. In either case, the vertical measurement is less accurate and less repeatable than the horizontal measure.

The LED and photosensors are positioned within ~ 2 cm of the eye, and are mounted to a head band, goggles, or spectacle frames. The detector assembly inevitably obscures some of the visual field. Since the reflectivity pattern moves more or less as would a landmark on the eyeball surface, a 1 mm shift of the optics on the head is expected to produce an error of $\sim 5^\circ$. There is often a distinct cross-talk effect. The horizontal measurement values are affected by vertical eye position, and visa versa. If vertical eye position is also measured, the calibration transform can attempt to correct cross-talk.

System bandwidth is limited, primarily, by the low pass filtering needed due to the modulation scheme, and is typically 50–100 Hz. If the signal is processed digitally, sample rates are often at least 1000 Hz. It is possible to achieve resolutions of $> 0.05^\circ$ visual angle. While bandwidth and resolution are very good, accuracy is somewhat undependable because of headgear slippage affects, cross-talk effects, and, especially in the vertical axis, eye lid effects. Accuracy of 1° along a horizontal axis and 2° along a vertical axis may be achievable over a short period, but errors of several degrees would not be unusual, especially over longer periods or in the presence of vigorous head motion. Devices that use one LED and two photo sensors for a given axis tend to become very nonlinear, and difficult to calibrate over > 30 or 40° in either axis. The range can be extended somewhat by using multiple LEDs and sensors for each axis. Neither torsional eye movements, nor pupil diameter is measured.

Photoelectric, reflectivity pattern (limbus) trackers are best suited to measure dynamics of horizontal eye movements as opposed to point of regard measurement, although they are sometimes used for point of regard measurement as well. Systems in this category are commercially available from Applied Science Laboratories Bedford, MA (71), Cambridge Research Systems Ltd, UK (20), and Optomotor Laboratory, Freiburg, Germany (72).

Cambridge Research Systems offers a version of their device designed for use in fMRI environments. In this case, the LEDs and photodiodes are located outside of the magnet bore, and connected to the eye piece, within the magnet bore, via fiber optic cables (20).

Dual Purkinje Image Measurement (CR/4PI)

A technique was described by Cornsweet and Crane (45) and further developed by Crane and Steele (73,74) in which the eye is illuminated with a modulated IR source, and servo controlled mirrors are used to image the first and fourth Purkinje images onto solid state quadrant detectors. Demodulated, analog signals from the quadrant detectors are used, in separate feed back loops, to move the mirrors and keep the images centered on the detectors. The resulting mirror positions constitute measures of the feature positions, and can be used to compute eye rotation with respect to the optics.

A schematic representation of the system is shown in Fig. 15. The entire optics platform is mounted to a servo controlled *XYZ* stage, which automatically moves to optimize overall system alignment. A hot mirror beam splitter is used to direct light to and from the optical unit, which is positioned off to one side of the subject. By looking through the beam splitter, the subject has an unimpeded view of the forward visual field. Not shown in the simplified schematic is an autofocus mechanism, implemented with a beam splitter, off axis aperture, and bicell detector, in the first Purkinje image optical path.

The subject's motion is usually restricted with a head rest or bite bar assembly. Because the fourth Purkinje image is visible only through the iris opening, measurement range is

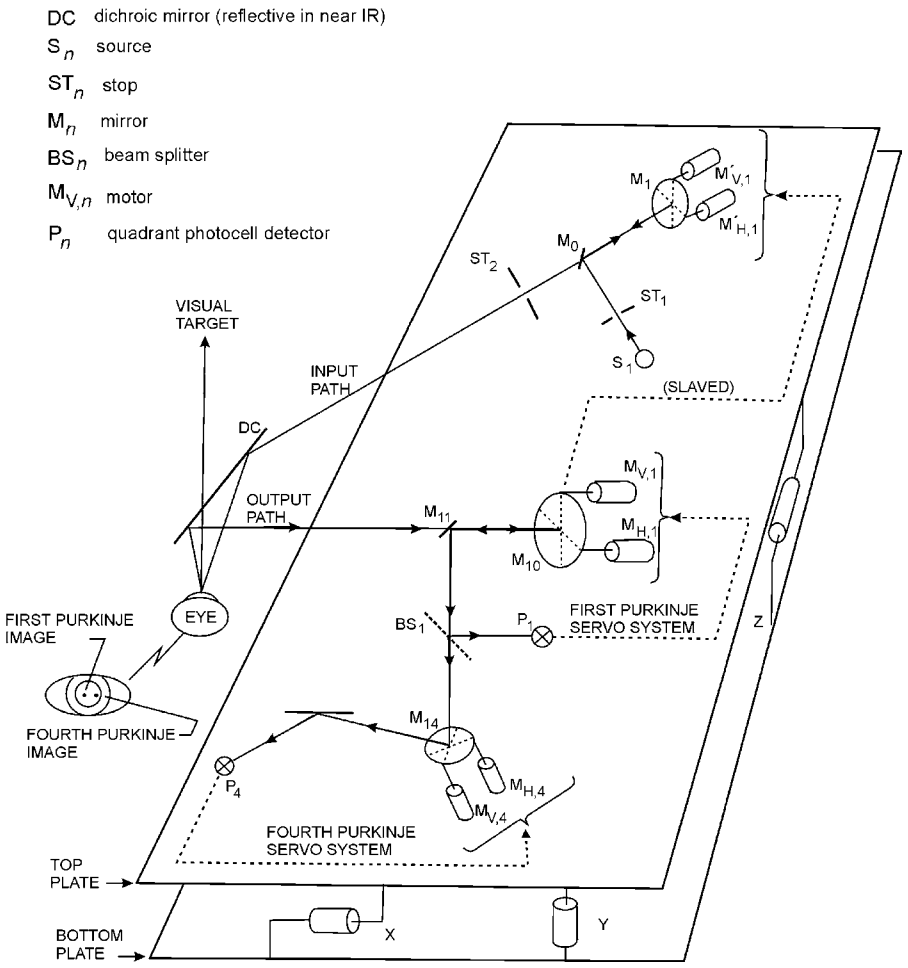


Figure 15. Schematic representation of Double Purkinje image eye tracker. (Redrawn from Ref. 73.)

restricted to about $\pm$ 10° visual angle, although this can be extended to about $\pm$ 15° by using drops to dilate the pupil. Accuracy is $\sim$ 1 min of arc, and the sampling rate is 1000 Hz. Using a model eye, Crane and Steele (74) empirically measured a bandwidth of $\sim$ 500 Hz (for small eye movements of several degrees), noise levels of 20 arc·s rms, response delay of 0.25 ms, and linear tracking of slew rates up to $\sim$2000°·s. The manufacturer of a current version of the device specifies a bandwidth of 400 Hz and a 1 ms response time (75).

Although the subject's motion is restricted and the measurement range is small, the accuracy, precision and temporal bandwidth are exceptionally good. The Dual Purkinje image eye-tracking device can provide a precise enough and fast enough measurement, for example, to allow another device to effectively stabilize an image on the retina (74). Neither torsional eye movement nor pupil diameter is measured, but it is possible to attach an infrared optometer, which provides a real-time, analog measure of changes in eye accommodation (75). An updated version of the device described by Crane and Steele (74) is offered commercially by Fourward Technologies, Inc, Buena Vista, VA. (75).

Systems Using Two-Dimensional Video Sensor Arrays

Eye tracking devices that use 2D CCD or CMOS sensor arrays exist in wide variety, and are commercially available from numerous sources. The term video oculography (VOG) is often used to describe this type of system. Most current systems in this category use the pupil to corneal reflection technique (CR/Pupil) to measure gaze direction, but there are exceptions. Some systems designed primarily to measure ocular torsion might rely on a fixed position of the head with respect to the sensor, and use only pupil position to correct for gaze direction changes. Some systems that normally use the pupil-to-corneal reflection technique have options to measure with only pupil position or only corneal reflection position under certain circumstances. Other systems are designed to measure the position of the head in space, and use a camera (or cameras) with a wide-angle view on which the corneal reflection can not easily be resolved. These systems may use the pupil position relative to a facial feature to compute gaze direction.

Note that a pupil-only or corneal-reflection-only measurement offers some advantages if it can be assured that the head does not move with respect to the camera. Pupil and corneal reflection motion have greater sensitivity to eye rotation than the pupil-to-corneal reflection vector, and the measurement is therefore more precise (less noisy).

Figure 16 is a functional diagram consistent with most current video-based eye tracking systems. The functions may be distributed among physical components in various ways, and a separate head tracking system may or may not be included.

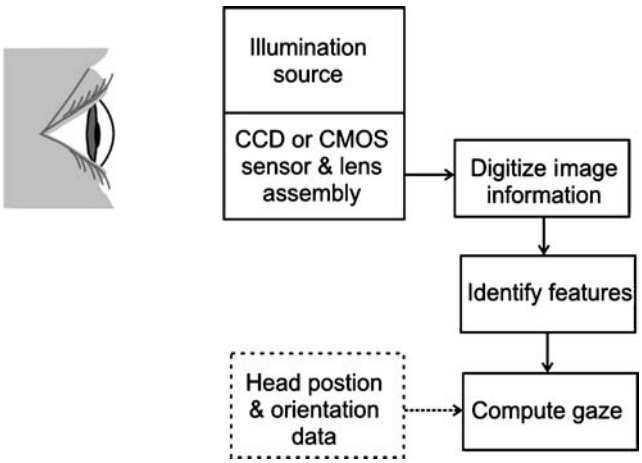


Figure 16. Schematic showing the basic functional architecture of most video-based eye tracking systems.

Some systems input video from the eye camera to PC memory via a USB or firewire connection, either directly from a camera with compatible output, or via a video–firewire converter. The PC CPU then does all the processing. Alternately, a frame grabber installed on the PC may be used to digitize a composite video signal. In some cases, the resulting digitized signal is processed by the PC, or in other cases, a frame grabber that also has image processing capabilities may do some portion of the processing before making a reduced information set available to the PC. In still other cases, the eye video is input to a completely custom processing board or board set, often populated with some combination of field programmable gate array (FPGA) and digital signal processing (DSP) components. The custom processor may do virtually all of the processing or may send a reduced data set to a PC. Other workstations may be used in place of the PCs referred to above. In short, there is enormous variety possible. Systems that support > 60 Hz update rates often make some use of custom processing components to help achieve the high speed. Although not yet capable of the same temporal bandwidth as several other types of system, camera based systems are now available with high enough update rates to make them useful for study of eye movement dynamics. Video-based systems almost always measure pupil diameter as a byproduct of computations for eye rotation measurement, and these systems are sometimes used as pupillometers.

The two main subcategories within this group are head mounted systems, having the sensor and small optics packages mounted to headgear; and remote systems, relying on optics mounted to the environment (table, instrument panel, etc.). In both cases these are often used as real time point of regard systems. The following sections describe video based, head mounted and remote eye tracking systems, respectively; followed by a section describing systems designed to handle the special problem of operation in MRI devices.

Head Mounted, Video Based Systems. Head mounted sensors are often designed to view a reflection of the eye from a beam splitter, as shown in Fig. 17, so that the

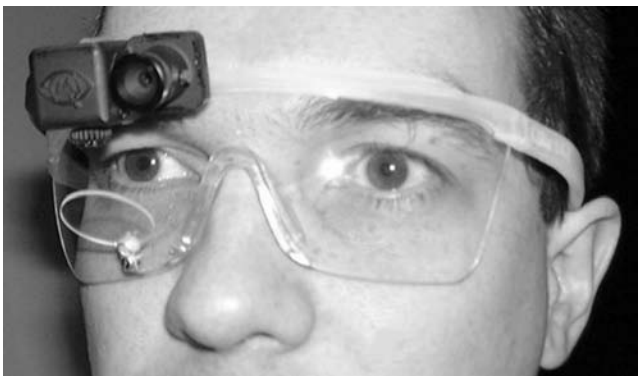


Figure 17. The head mounted module directly over the subject's right eye contains both a camera and illuminator, which are aimed down at the tilted, hot mirror beam splitter. The camera views a reflection of the eye from the hot mirror. The module just to the nasal side of the eye camera module is a scene camera, aimed toward the subject's field of view. (Courtesy of Applied Science Laboratories, Bedford, MA.)

sensor, lens, and illumination source need not be in the subject's field of view. The beam splitter is coated to be a hot mirror, very reflective in the near-IR, but not in the visible spectrum. In some cases, however, the module containing the sensor is positioned on a stalk (Fig. 18), so that it can obtain a direct view of the eye. Illumination is most often provided by one or more near infrared LEDs. Both dark-pupil and bright-pupil type optics are represented among available systems.

The image of the eye produced by a head mounted sensor yields information about eye orientation with respect to the head. In other words, line of gaze is measured in the head reference frame. In some cases, this is the

Figure 18. Head mounted camera and illuminator module is mounted on a stalk, and is aimed directly at the eye. (Courtesy of SR Research, Inc. Ontario.)

desired quantity, but in other cases the desired measurement is point of gaze in the environment. In order to find point of gaze on objects in the environment, it is necessary either to detect the positions of those objects with respect to the head, or to detect the position and orientation of the head with respect to the environment containing the objects. Techniques for accomplishing this were discussed in the previous section titled "Measuring point of gaze in the presence of head motion". All of these methods have been used by commercially available, pupil–corneal reflection-type systems.

Head tracking devices commonly used with head mounted eye trackers include magnetic systems that measure the position and orientation of a small, head gear mounted, set of coils with respect to a larger set of stationary coils; optical systems that use head mounted sensors to detect moving laser beams from a stationary source; and a system which uses a head mounted inertial package for high frequency measurements, corrected by lower frequency measurements from ultrasonic emitter–receiver pairs (8).

Head tracking systems capable of sufficient 6 degree of freedom accuracy usually restrict subject to 1 or 2 m of movement. Applications requiring a subject to walk or run a significant distance usually rely on a head mounted scene camera to capture information about point of gaze on the environment. There are available systems that allow subjects to move about with no physical connection to stationary components while either recording data on a recording device attached to the subject, or using wireless transmission to send data to a base station.

Systems intended for neurological testing are usually concerned with eye movement with respect to the head rather than point of gaze in the environment. Some of these systems measure ocular torsion as well as measuring gaze direction with either pupil-to-corneal-reflection or pupil position techniques.

Measurement accuracy, of head mounted, video-based, eye trackers, tends to be from ~ 0.5–1.0° visual angle, with resolutions varying from 0.01 to 0.1° visual angle. Measurement range is 60–70° visual angle in the horizontal axis and usually $\sim 20^\circ$ less in the vertical axis due to eye lid interference. These figures refer to measurement of gaze direction with respect to the head. The measurable field of view with respect to the environment is unrestricted because subjects are free to turn their heads and bodies. Measurement update rates range from 25 to 360 Hz, with at least one commercially available system offering 500 Hz with pupil-only (as opposed to pupil-to-corneal-reflection) measurements.

When a head mounted eye tracker is used in combination with a head tracking device, in order to measure point of gaze in the environment, error in the final measurement is increased due to noise and error in the head position and orientation measurement. The result can vary widely depending on the particular combination of equipment being used and the conditions under which measurements are being made. It is probably not unreasonable to expect additional errors corresponding to ~ 0.2–1.0° visual when using a head mounted eye tracker to measure point of gaze on elements in the environment. Conditions that create difficulties for either the eye or head tracking components, for example a heavily metallic environment when using a magnetic head tracker, or occlusion of the pupil image by reflection from some external light source, may cause errors to increase dramatically or make the measurement impossible.

Manufacturers of systems that include ocular torsion measurement usually specify torsion measurement range of 18–20°, resolution of $\sim 0.1^\circ$, and linearity between 1 and 4% of full scale. Video-based, head mounted systems, using the pupil to corneal reflection measurement technique, currently allow point of gaze measurement in the most diverse environments, and with the greatest freedom of subject motion and measurable field of view. They do not have as high a temporal bandwidth as scleral search coil, double Purkinje image, EOG, or reflectivity (limbus) tracking systems, and are not as accurate or precise as scleral search coil, or double Purkinje image systems. They do require that the subject wear head gear, but the size and obtrusiveness of the headgear has been steadily reduced over the past decade. It is becoming practical to use this type of system in ever more flexible and varied settings. The amount and severity of subject motion, the amount of variation in ambient light environment, and the length of time over which continuous measurements must be made, all tend to trade off somewhat against the accuracy and dependability of the measurements.

Video-based eye trackers, with head mounted optics, are offered commercially by: Alphabio, France (76); Applied Science Laboratories, Bedford, MA (71); Arrington Research, Inc., Scottsdale, AZ (77); Chronos Vision GmbH, Berlin, Germany (78); Guymark UK Ltd, UK (22); EL-MAR Inc, Downsview, ON (79); ISCAN, Inc., Burlington, MA (80); Neuro Kinetics Inc. Pittsburgh, PA (24); SensoMotoric Insturments GmbH, Berlin, Germany (81); and SR Research Ltd, Osgoode, ON (82).

Head mounted systems that measure ocular torsion are available from: Alphabio, France (72); Arrington Research, Inc, Scottsdale, AZ (77); Chronos Vision GmbH, Berlin, Germany (78); Neuro Kinetics Inc. Pittsburgh, PA (24); and SensoMotoric Insturments GmbH, Berlin, Germany (81).

Remote, Video-Based Systems. Remote (nonhead mounted) systems have one or more videosensors and illumination sources that are fixed to the environment and aimed toward the subject. The optics may sit on a table next to or underneath a monitor being viewed by the subject, may be mounted on a vehicle (or vehicle simulator) instrument panel, or mounted to the environment in some other way. Both bright and dark pupil optics are used. In the case of bright pupil optics, the illuminator may be a ring of LEDs placed close enough to the lens to produce the bright pupil effect, an LED actually mounted over the front of the lens so that it blocks only a small portion of the lens surface, or a beam splitter arrangement like that shown in Fig. 6.

Some system configurations require the subject's head to be stabilized. This maximizes achievable accuracy and dependability. A chin and forehead rest, or cheek and forehead rest are commonly used, although these do allow

a small amount of head motion. A bite bar with a dental impression can be used to completely stabilize the head. In the case of nonhuman primate research the head is often stabilized by other means appropriate for neural recording. The camera, lens, and illumination sources can all be stationary (after manual positioning) and may safely be designed with a field of view just large enough to comfortably accommodate pupil and corneal reflection motion due to eye rotation. If the subject's head is sufficiently stabilized there is no need to differentiate between eye rotation and head translation, and, as previously discussed, there is a precision advantage to using a pupil-only (rather than pupil-to-corneal-reflection) measurement. Some systems allow a choice between pupil-to-corneal-reflection and pupil-only or corneal-reflection-only measurement.

Alternately, the eye camera field of view may be wide enough to accommodate the motions of a seated subject who is not restrained, but is voluntarily remaining still. If the subjects are provided with feedback, so that they know when they are moving out of the legal space, it is only necessary to allow ~ 5 cm of head motion.

Many remote systems are designed to allow enough head motion to accommodate normal motions of a person working at a computer terminal or driving a car, and so on. This is accomplished either by dynamically rotating the camera optical axis so that it always points toward the eye being tracked, by using a camera field of view wide enough to accommodate the desired head motion, using multiple cameras, or some combination of these. Head motion toward and away from the eye camera must either be small enough to remain within the lens system depth-of-field, or must be accommodated by an autofocus mechanism. Figure 19 shows an example of an eye camera that automatically moves in azimuth (pan) and elevation (tilt) to follow the eye as the head moves about. Figure 20 is an example of a camera with a moving mirror used to direct the optical path, and Fig. 21 shows a system with stationary sensor and illumination components in an enclosure.

Systems with a moving camera or moving mirrors often use a closed loop control, based on the detected position of the pupil in the camera field of view. The moving element is driven to move the image toward center. If the pupil is completely lost from the camera field of view, the system may execute a search pattern, use information from a separate head tracking system to reacquire the eye, or require that a human operator intervene to reacquire the eye image. Such systems can have a relatively narrow eye camera field of view, thus maximizing image resolution for the features of interest. The disadvantage is the need for moving parts, and the possibility of failing to maintain the eye image in the camera field of view.

Systems that do not require moving parts have a significant advantage in terms of system simplicity and dependability. Furthermore, there is the possibility of using the same image to measure head position and to track the movement of both eyes. The trade-off is that as the field of view is increased, resolution is reduced (the pupil and corneal reflection are imaged onto fewer pixels) and the features of interest must be identified within a larger, more complex image.

The pupil-to-corneal-reflection method alone can be used to determine gaze angle with respect to the eye camera. However, in the presence of significant head motion, this is not sufficient to accurately determine point of gaze on other surfaces. Some remote systems use head trackers to find the position of the eye in space, and use the information to more accurately compute point of gaze on other stationary surfaces. Head position information can also be used to help aim moving cameras or mirrors. The same types of head tracker mentioned in the previous section, as being appropriate for use with head mounted eye trackers, are sometimes used in conjunction with remote eye trackers.

Figure 19. Example of a remote eye tracker optics module that moves in azimuth (pan) and elevation (tilt) to keep a telephoto eye image within the camera field of view. (Courtesy of Applied Science Laboratories, Bedford, MA.)

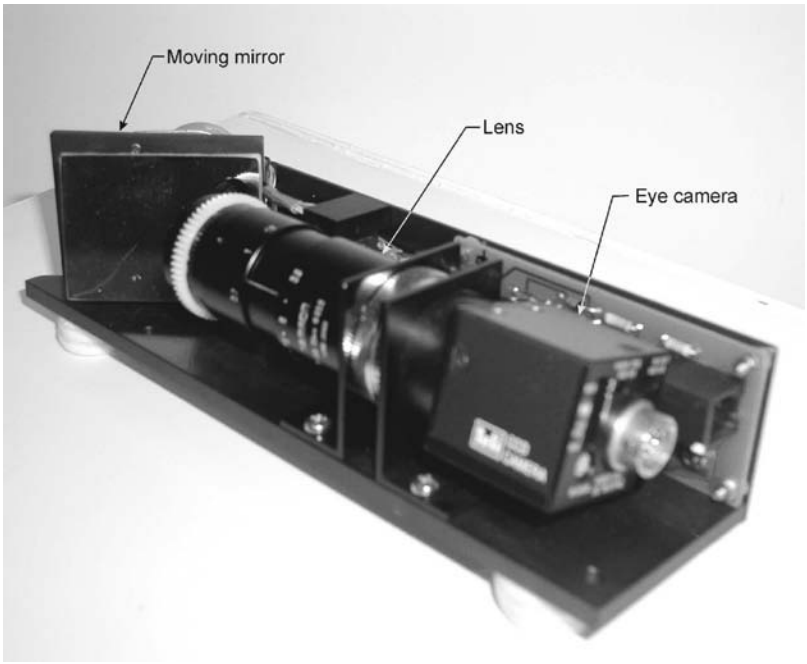


Figure 20. Example of a remote eye tracker optics module that uses a moving mirror to keep a telephoto eye image within the camera field of view. The optics module is shown with its cover removed. (Courtesy of Applied Science Laboratories, Bedford, MA.)

In the case of remote eye trackers, it is also possible to use one or more wide angle cameras to locate the head and eyes. The head tracking camera (or cameras) may be separate from the eye tracking camera, or alternately, the same cameras (or cameras) may be used for both.

At present, remote systems that offer >60 Hz update rates either require the head to be stabilized, or use a narrow field image that is dynamically directed to follow head motions. Systems that use stationary wide-angle optics typically update at 60 Hz or less, probably because of the extra processing required for feature recognition in the more complex wide-angle image. Systems that either require the head to be stabilized or direct a narrow field of camera to follow head motion are available with update rates of up to 360 Hz. As with head mounted systems, higher update rates result in lower signal/noise ratios for the sensor.

Figure 21. Example of a remote eye tracker optics module using stationary wide angle optics. Sensor, lens, and illumination components are within the enclosure. (Courtesy of Tobii Tehnology AB, Stockholm, Sweden.)

For configurations that stabilize the subject's head, accuracy of remote, video-based eye tracking systems tends to be ~ 0.5–1.0° visual angle, with resolutions varying from 0.01 to 0.1° visual angle. When a remote eye tracker is used to measure point of gaze on a stationary surface, and head motion is allowed, some additional error can be expected. If no attempt is made to account for head motion, the equation in Fig. 11 can be used to estimate the amount of error expected from a given change in head position. If an attempt is made to correct for head movement, the amount of additional error depends on the accuracy of the head position measurements, and the way the information is used in the point-of-gain computation, as well as the range of head motion. It is probably not unreasonable to expect an additional 0.5–1.5° of error when there is significant head motion. Data may also have additional error, or even be briefly lost, during fast head motion, due to instability of the eye image on the eye-camera field of view.

Remote systems using one eye-camera and one illumination source, and the pupil-to-corneal-reflection method, can generally measure gaze directions that are within $\sim$ 25–35° visual angle from the eye-camera lens. At more eccentric angles the corneal reflection is either not visible to the eye camera, or easily confused with multiple reflections from the sclera. The exception is that when line of gaze is below the eye-camera, the upper eyelid often begins to occlude part of the pupil. The amount of eye lid occlusion under this condition varies from subject to subject. Furthermore, different eye tracking systems, using different recognition and center computation algorithms, are tolerant of different amounts of occlusion. This often limits measurement range to 5–10° visual angle below the eye-camera, and for this reason, the environment is usually arranged with the eye-camera at the bottom of the scene space that is of interest. The result is a horizontal gaze measurement range of ~ 50–70° visual angle, and a vertical range of ~ 35–45°.

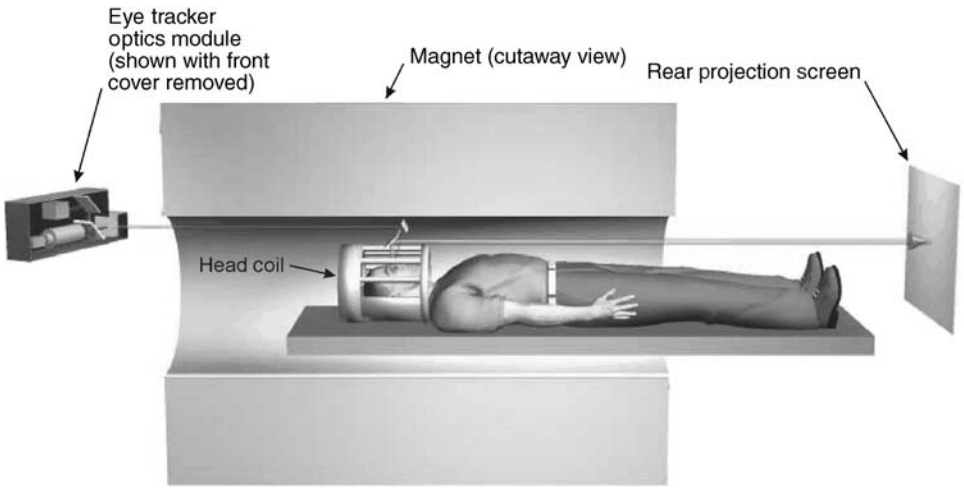


Figure 22. Example of eye tracker optical path in an fMRI system. The drawing shows the magnet in cutaway view. One mirror, just above the subject's face, reflects the eye image to the eye tracker camera, while another allows the subject to view a stimulus on a rear projection screen. Many variations to this arrangement are used, including configurations in which both the projection screen and eye tracker optics module are at the same end of the magnet. (Courtesy of Applied Science Laboratories, Bedford, MA.)

The range of measurable gaze directions, using the pupil-to-corneal-reflection method, can be increased by using multiple illumination sources (and thus multiple corneal reflections). The horizontal range can be increased to as much as 90° by widely spacing illumination sources. Only a more modest range expansion is possible in the vertical axis because of eyelid interference.

Remote, video-based, eye trackers, offer the possibility of measuring eye movement very unobtrusively. Unlike head mounted systems, nothing need be worn by the subject (unless using a type of head tracker that requires a head mounted sensor), and the optics can be hidden from obvious view by a filter that is transmissive in the near-IR. The range of subject motion and measurable field of view is significantly more restricted than for systems with head mounted optics, although less so than for most other techniques. Measurement accuracy, resolution, and temporal bandwidth are not as good as that available with the much more restrictive and obtrusive double Purkinje image method, or scleral search coil method.

Remote, video-based, eye movement measurement systems are commercially available from: Applied Science Laboratories, Bedford, MA (71); Arrington Research, Inc., Scottsdale, AZ (77); Cambridge Research Systems Ltd, UK (20), ISCAN, Inc., Burlington, MA (80); LC Technologies Inc., Fairfax, VA (83); Metrovision, Pérenchies, France (23), Seeing Machines, Canberra, Australia (84); SensoMotoric Insturments GmbH, Berlin, Germany (81); SR Research Ltd, Osgoode, ON (82); and Tobii Technology AB, Stokholm, Sweeden (85).

Video-Based Systems for use in Conjunction with fMRI. There are commercially available, video-based eye tracker systems, which use the pupil-to-corneal-reflection method, and are specifically configured to be used in conjunction with fMRI measurements of brain activity.

During fMRI measurements, the subject's head and torso are inside the main MRI magnet bore, and a smaller head coil is placed fairly closely around the subject's head. The head coil always has an opening over the subject's eyes, and there is usually some provision for the subject to look at a visual display, often with the help of a mirror or set of mirrors. The environment is challenging because there is not much room in which to arrange an optical path from the subject's eye to a camera, and also because any electronic equipment used must operate in a high magnetic field, without creating even small amounts of magnetic field noise that would interfere with the MRI measurement.

One approach is to use a camera equipped with a telephoto lens, and placed outside of the magnet. The camera is aimed down the magnet bore, and views a reflection of the eye on a small mirror placed inside of the magnet, near the subject's head. The optical path length from the camera lens to the eye is typically at least 2.5–3 m, necessitating a fairly large telephoto lens, and powerful illumination source. An example is shown in Fig. 22, and is similar to a setup described by Gitleman et al. (86). The system shown is using a bright pupil technique, so there is only one coaxial path for both the camera and illumination beam. In this case, a second mirror allows the subject to look at a display screen. Many variations of both the eye camera and display screen optical paths are possible.

Other systems bring the eye image out of the magnet, to the sensor, with a coherent, fiber optic bundle. This has the advantage of not requiring large, telephoto optics, but the resolution of fiber bundles are limited and light is attenuated somewhat unevenly by the individual fibers that comprise the bundle. Fiber optics may also be used to relay the output from an illumination source into the magnet. It is possible to place a small camera inside the magnet, but it can be difficult to avoid some interference with the magnetic field and resulting degradation of the fMRI measurement.

Video-based, eye tracking systems designed for use during fMRI measurement, are commercially available from Applied Science Laboratories, Bedford, MA (71); Arrington Research, Inc., Scottsdale, AZ (77); ISCAN, Inc., Burlington, MA (80); and SensoMotoric Insturments GmbH, Berlin, Germany (81).

COMPARISON OF EYE MOVEMENT MEASUREMENT TECHNIQUES

Table 1 is an abbreviated comparison of the various techniques discussed.

Table 1. Summary of Most Prevalent Eye Tracking Techniques[a]

Method	Typical Applications	Typical Attributes	Typical Reference Frame(s)	Typical Performance
EOG	Dynamics of saccades smooth pursuit nystagmus	High bandwidth Eyes can be closed In expensive Drift problem (poor position accuracy) Requires skin electrodes - otherwise unobtrusive	head	static accuracy: ~3-7° resolution: with low pass filtering and periodic rezero, virtually infinite bandwidth: ~100 Hz
Scleral Coil	Dynamics of saccades, smooth pursuit, nystagmus Miniature eye movements Point of gaze Scan path Torsion	Very high accuracy and precision Invasive Very obtrusive High bandwidth	Room	accuracy:~0.2° resolution: ~1 arc min. range: ~ 30° bandwidth: ~200 Hz
Limbus (using small number of photo sensors)	Dynamics of saccades, smooth pursuit, nystagmus Point of gaze Scan path	High bandwidth Inexpensive Poor vertical accuracy Obtrusive (sensors close to eye) Head gear slip errors	head gear	accuracy: varies resolution: 0.1° (much better res. possible) range: ~30° update rate: 1000 samples/s
CR/Pupil	Point of gaze Line of gaze Scan path Dynamics of eye movements (only with subset of systems offering high update rates) Torsion (available only on some systems)	Minimal head gear slip error Unobtrusive Low to medium bandwidth Problems with sunlight Wide variety of configurations and environments	Head gear Room Cockpit	accuracy: ~1° resolution: ~ 0.2° hor. range:~50° vert. range: ~40° update rate: 50 or 60 samples/s. Update rates up to 360 samples/s available on some systems (higher for pupil only measurement);
CR/4PI	Dynamics of saccades, smooth pursuit, nystagmus Miniature eye movements Point of gaze Scan path Image stabilization on retina Accommodation	Very high accuracy and precision High bandwidth Obtrusive (large optics package, restricted head motion) Limited range	Room	accuracy: ~1 arc min. resolution: < 1 arc min range: ~20° bandwidth: ~400 Hz

[a]Adapted from Ref. 8.

An earlier, but comprehensive and still pertinent review of eye movement measurement techniques, can be found in Young and Sheena (87). A book by Duchowski (53) presents a detailed treatment of many aspects of eye tracking methodology. The journal *Computer Vision and Image Understanding* has devoted an entire special edition to developments in optical eye tracking, with an emphasis on real-time, nonintrusive techniques (88). A database of commercially available eye movement measurement equipment is available at a web site hosted by the University of Derby (89).

BIBLIOGRAPHY

Cited References

1. Roth EM, Finkelstein S. Light environment. In: Roth EM, editor. Compendium of Human responses to the Aerospace Environment. NASA CR-1205(1); 1968.
2. Westheimer G. The eye, In: Mountcastle VB, editor, Medical Physiology, Volume One. Saint Louis: The C. V. Mosby Co.; 1974.
3. Fry GA, Hill WW. The center of rotation of the eye. Am J Optom Arch Am Acad Optom 1962;390:581–595.
4. Wyatt HJ. The form of the human pupil. Vision Res 1995;35(14):2021–2036.
5. Haslwanter T. Mathematics of three-dimensional eye rotations. Vision Res 1995;35:1727–1739.
6. Hallett P. Eye movements. In: Boff KR, Kaufman L, Thomas JP, editors. Handbook of Perception and Human Performance. New York: John Wiley & Sons, Inc.; 1986.
7. Robinson DA. A method of measuring eye movement using a scleral search coil in a magnetic field, IEEE Transactions on Bio-Medical Electronics 1963;BMI-10:137–145.
8. Leger A et al. Alternative Control Technologies, North Atlantic Treaty Organization, RTO-TR-7, 1998.
9. Collewijn F, van der Marx S, Jansen TC. Precise recording of human eye movements. Vision Res 1975;15:447–450.

10. Judge SJ, Richmond BJ, Chu FC. Implantation of magnetic search coils for measurement of eye position: an improved method. Vision Res 1980;20:535–538.
11. Skalar Medical BV. (2002, December 20). Home. (Online). Available at http://www.skalar.nl, Accessed Feb. 9, 2005.
12. Riverbend Instruments. (No date). Home. (Online). Available at http://www.riverbendinst.com/index.htm. Accessed 13 April 2005.
13. Mowrer OH, Ruch RC, Miller NE. The corneo-retinal potential difference as the basis of the galvanometric method of recording eye movements. Am J Physiol 1936;114:423.
14. Marg E. Development of electro-oculography—standing potential of the eye in registration of eye movement. AMA Arch Ophthalmol 1951;45:169.
15. Kris C. Vision: electro-oculography. Glasser O, editor. Medical Physics: Chicago: Chicago Yearbook Publishers; 1960.
16. Shackel B. Review of the past and present in oculography. Medical Electronics Proceedings of the Second International Conference. London: Hiffe; 57, 1960.
17. Shackel B. Eye movement recording by electro-oculography. In: Venables PH, Martion I, editors, A manual of Psychophysiological Methods. Amsterdam: North-Holland Publishing Co.; pp 300–334, 1967.
18. Arden GB, Barrada A, Kelsey JH. New clinical test of retinal function based upon the standing potential of the eye. Br J Ophthalmol 1962;46:467.
19. De Rouck A, Kayembe D. A clinical procedure of the simultaneous recording of fast and slow EOG oscillations. Int Ophthalmol 1981;3:179–189.
20. Cambridge Research Systems. (2005, February 9). Home. (Online). Available at http://www.crsltd.com Accessed 2005, Feb. 10.
21. GN Otometrics. (2004). Products. (Online). Available at http://www.gnotometrics.com/products.htm. Accessed 2005 Feb. 9.
22. Guymark UK Ltd. (No date). Vestibular Assessment. (Online). Available at http://www.guymark.com/vestibul.html. Accessed 2005, Feb. 9.
23. Metrovision. (No date). Homepage (Online). Available at http://www.metrovision.fr. Accessed 2005, Feb. 9.
24. Neuro Kinetics Inc. (No date). Products and Services. (Online). Available at http://neuro-kinetics.com/products-research.htm. Accessed 2005, Feb. 9.
25. Nguyen K, Wagner C, Koons D, Flickner M. Differences in the infrared bright pupil response of human eyes. Proceedings ETRA 2002 Eye Tracking Research & Applications Symposium. New Orleans: March 25–27, 2002.
26. Borah J. Helmet Mounted EyeTtracking for Virtual Panoramic Display Systems–Volume I: Review of Current Eye Movement Measurement Technology. US Air Force report AAMRL-TR-89019, 1989.
27. Zhu S, Moore T, Raphan T. Robust pupil center detection using a curvature algorithm. Computer Methods Programs Biomed 1999;59:145–157.
28. Mulligan JB. A software-based eye tracking system for the study of air-traffic displays. Proceedings ETRA 2002 Eye Tracking Research & Applications Symposium. New Orleans: March 25–27, 2002.
29. Ohno T, Mukawa N, Yoshikawa A. FreeGaze: a gaze tracking system for everyday gaze interaction. Proceedings ETRA 2002 Eye Tracking Research & Applications Symposium. New Orleans: March 25–27, 2002.
30. Charlier J et al. Real time pattern recognition and feature analysis from video signals applied to eye movement and pupillary reflex monitoring. Proceedings of the 6th Int Visual Field Symposium. In: Heijl A, Greve EL, editors. Dordrecht, The Netheralands: Dr. W Junk Publisheres; 1985.
31. Sheena D. Pattern-recognition techniques for extraction of features of the eye from a conventional television scan. In: Monty RA, Senders JW, editors. Eye Movements and Psychological Processes. Hillsdale (NJ): Lawrence Erlbaum; 1976.
32. U.S. Pat. 5,302,819 1994. Kassies M. Method of and Apparatus for, Detecting an Object.
33. Zhai S, Morimoto C, Ihde S. Manual and Ggaze input cascaded (MAGIC) pointing. Proceedings CHI'99: ACM Conference on Human Factors in Computing Systems. Pittsburgh: ACM: 246–253, 1999.
34. Tomono I, Muneo X, Lobayashi Y. A TV camera system that extracts feature points for non-contact eye movement detection. iSPIE vol. 1194 Optics, Illumination, and Image Sensing for Machine Vision IV; 1989.
35. Ebisawa Y, Satoh S. Effectiveness of pupil area detection technique using two light sources and image difference method. 15th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. San Diego: 1993.
36. Eizenman M, Frecker RC, Hallett PE. Precise non-contacting measurement of eye movements using the corneal reflex. Vision Res 1984;24:167–174.
37. Xiao J, Baker S, Matthews I, Kanade T. Real-Time Combined 2D+3D Active Appearance Models. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. June, 2004.
38. Matthew X, Baker S. (none) Real-Time AAM fitting algorithms. [Online]. Carnegie Mellon University. Available at http:// www.ri.cmu.edu/projects/project_448.html. Accessed 9 Feb. 2005.
39. Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence 2001;23:681–685.
40. Zhu J, Yang J. Subpixel Eye-Gaze Tracking. Proceedings of the fifth IEEE International Conference on Face and Gesture Recognition 2002; 131–136.
41. U.S. Pat. 5,818,954 1998. Tomono M, Iida K, Ohmura X. Method of Detecting Eye Fixation using Image Processing.
42. Merchant J, Morrissette R, Perterfield JI. Remote measurement of eye direction allowing subject motion over one cubic foot of space. IEEE Transactions on Biomedical Engineering 1974; BME-21:309–317.
43. Sheena D, Borah J. Compensation for some second order effects to improve eye position measurements. In: Fisher DF, Monty RA, Senders JW, editors. Eye Movements: Cognition and Visual Perception. Hillsdale (NJ): Lawrence Erlbaum Associates; 1981.
44. U.S. Pat. 6,5987,971, 2003. Cleveland D. Method and System for Accommodating Pupil Non-concentricity in Eyetracker Systems.
45. Cornsweet TN, Crane HD. Accurate two-dimensional eye tracker using first and fourth Purkinje images. J Opt Soc Am 1973;63:921.
46. Torok N, Guillemin B, Barnothy JM. Photoelectric nystagmography. Ann Otol Rhinol Laryngol 1951;60:917.
47. Smith WM, Warter PJ. Eye movement and stimulus movement: New Photoelectric electromechanical system for recording and measuring tracking motions of the eye. J Opt Soc Am 1960;50:245.
48. Stark L, Sandberg A. A simple instrument for measuring eye movements, Quarterly Progresss Report 62, Research Laboratory of Electronics, Massachusetts Insstitute of Technology; 1961. p. 268.
49. Wheeless LL, Boynton RM, Cohen GH. Eye movement responses to step and pulse-step stimuli. J Opt Soc Am 1966;56:956–960.

50. Young LR. Recording eye position. In: Clynes M, Milsum JH, editors. Biomedical Engineering Systems. New York: McGraw-Hill; 1970.
51. Findlay JM. A simple apparatus for recording micorsaccades during visual fixation. Quart J Exper Psychol 1974;26:167–170.
52. Reulen JP et al. Precise recording of eye movement: the IRIS technique. part 1. Med Biol Eng Comput 1988;26:20–26.
53. Duchowski AT. Eye Tracking Methodology Theory and Practice. London: Springer-Verlag; 2003.
54. U.S. Pat. 4,852,988, 1989. Velez J, Borah J. Visor and camera providing a parallax-free field-of-view image for a head-mounted eye movement measurement system.
55. Edelman ER. Video based monitoring of torsional eye movements. S.M. dissertation, Massachusetts Institute of Technology, Cambridge (MA). 1979.
56. Proceedings of the 3rd VOG workshop, Tubingen, Germany, Nov. 30–Dec. 2, 1999.
57. Hatamian M, Anderson DJ. Design considerations for a real-time ocular counterroll instrument. IEEE Trans Biomed Eng 1983;30:278–288.
58. Clarke AH, Teiwes W, Scherer H. Videooculography—an alternative method for measurement of three-dimensional eye movements. In: Schmid R, Zambarbieri D, editors. Oculomotor Control and Cognitive Processes, Amsterdam: Elsevier; 1991.
59. Bucher U, Heitger F, Mast F. A novel automatic procedure for measuring ocular counterrolling. Behav Res Methods Instr Comp 1990;22:433–439.
60. Moore ST, Haslwanter T. Curthoys IS, Smith ST. A geometric basis for measurement of three-dimensional eye poisition using image processing. Vision Res 1996;36: 445–459.
61. Petrka RJ, Merfeld DM. Calibration techniques for video-oculography, Poster Presentation at Barany Society Meeting. Sydney Australia; 1996.
62. Groen E, Nacken PFM, Bos JE, De Graaf B. Determination of ocular torsion by means of automatic pattern recognition. IEEE Trans Biomed Eng 1996;43(5): 471–479.
63. Zhu D, Moore ST, Raphan T. Teal-time torsional eye position calculation from video images. Soc Neurosci Abstr 1999;25:1650.
64. Guillemant P, Ulmer E, Freyss G. 3-D eye movement measurements on four Comex's drivers using video CCD cameras, during high pressure diving. Acta Otolaryngol (Suppl) (Stockh) 1995;520:288–292.
65. McConkie GW. Evaluating and reporting data quality in eye movement research. Behav Res Methods Instrum 1981;13: 97–106.
66. Kliegle R, Olson RK. reduction and calibration of eye monitor data. Behav Res Methods Instrum 1981;13:107–111.
67. Jacob RK. Eye tracking in advanced interface design. In: Barfield W, Furness T, editors. Virtual Environments and Advanced Interface Design. New York: Oxford University Press; 1995.
68. U.S. Pat. 6,578,962, 2003. Amir MD, Flickner DB, Koons X, Russell GR. (to Calibration-Free Eye Gaze Tracking).
69. American Conference of Governmental Industrial Hygienists, 2001 TLVs and BEIs, ACGIH, 1330 Kemper Meadow Drive, Cincinnati, OH; 2001.
70. Sliney DH, Wolbarsht M. Safety with Lasers and Other Optical Sources: A Comprehensive Handbook. New York: Plenum Press; 1980.
71. U. S. Pat. (to Applied Science Laboratories). (2005, January 5) Home. (Online). Available at http:http://www.a-s-l.com. Accessed 9 Feb. 2005.
72. U. S. Pat. (to Applied Science Laboratories). (No date). Express Eye. (Online), Available at http://www.optom.de/english/exe. htm. Accessed 9 Feb. 2005.
73. Crane HD, Steele CM. Accurate three-dimensional eyetracker. Appl Opt 1978;17:691.
74. Crane HD, Steele CM. Generation-V dual-Purkinje image eyetracker. Appl Opt 1985;24:527–537.
75. Fourward Technologies Inc. (2004, November 11). Home (Online). Available at http://www.fourward.com. Accessed 9 Feb. 2005.
76. Alphabio. (No date). Home. (Online). Available at http://www.electronica.fr/alphabio. Accessed 9 Feb. 2005.
77. Arrington Research, Inc. (2004, September 9). Home. (Online). Available at http://www.arringtonresearch.com. Accessed 9 Feb. 2005.
78. Chronos Vision. (no date). Eye Tracking. (Online). Available at http://www.chronos-vision.de/eyetracking/default_start_eyetracking.htm. Accessed 9 Feb. 2005.
79. EL-MAR, Inc. (No date). Menu. (Online). Available at http://www.interlog.com/~elmarinc/menu.htm. Accessed 9 Feb. 2005.
80. ISCAN Inc., (no date). Home. (Online). Available at http://iscaninc.com. Accessed 9 Feb. 2005.
81. SensoMotoric Instruments. (No date). Home. (Online). Available at http://www.smi.de. Accessed 9 Feb. 2005.
82. SR Research Ltd. (2004, Nov. 24). Home. (Online). Available at http://www.eyelinkinfo.com/. Accessed 9 Feb. 2005.
83. LC Technologies Inc. (2003)., Home. (Online). Available at http://www.eyegaze.com. Accessed 9 Feb. 2005.
84. Seeing Machines. (2004). Home. (Online). Available at http://www.seeingmachines.com. Accessed 9 Feb. 2005.
85. Tobii Technology. (2005). Home. (Online). Available at http://www.tobii.se. Accessed 9 Feb. 2005.
86. Gitelman DR, Parrixh TB, LaBar KS, Mesulam MM. Real-time monitoring of eye movements using infrared video-oculography during functional magnetic resonance imaging of the frontal eye fields. NeuroImage 2000;11:58–65.
87. Young LR, Sheena D. Survey of eye movement recording methods. Behav Res Methods Instrumen 1975;7:397–429.
88. Wechsler H, Duchowski A, Flickner M. Editorial, special issue: eye detection and tracking. Computer Vision Image Understanding 2005;98:1–3.
89. Eye Movement Equipment Data Base (EMED). (2002, May 2). Home. (Online). Available at http://ibs.derby.ac.uk/emed. Accessed 9 Feb. 2005.

See also ELECTRORETINOGRAPHY; FIBER OPTICS IN MEDICINE; OCULAR MOTILITY RECORDING AND NYSTAGMUS.

FES. See FUNCTIONAL ELECTRICAL STIMULATION.

FETAL MONITORING

MICHAEL R. NEUMAN
Michigan Technological University
Houghton, Michigan

INTRODUCTION

Fetal monitoring is a special type of electronic patient monitoring aimed at obtaining a record of vital physiologic functions during pregnancy and birth. Such monitoring is applied in assessing the progress of pregnancy and labor, and it can identify conditions that concern the clinician caring for the patient. These nonreassuring recordings can lead to special considerations in caring for the pregnant patient and in managing her labor. Although these recordings are no longer considered to be definitive in identifying most forms of fetal distress, they can help to reassure patient and clinician that the fetus is able to withstand the physiologic stress of labor and delivery. The technology is also useful in assessing high risk pregnancies, which, in most cases, is only reassuring as opposed to giving a definitive diagnosis. Although this technology is now recognized to have diagnostic limitations, it is still frequently used in the hospital and clinics as an adjunct to other diagnostic evaluations.

PHYSIOLOGIC VARIABLES MONITORED

The goal of fetal monitoring is to ensure that vital fetus organs receive adequate perfusion and oxygen so that metabolic processes can proceed without compromise and these organs can carry out their functions. Thus, an ideal situation for monitoring from the physiologic standpoint would be to monitor the perfusion and oxygen tension in the fetal central nervous system, heart, kidneys, and brain, with the brain being by far the most important. It is also important to know that the fetus is receiving adequate oxygen and nutrients from the mother through the placenta. Unfortunately, it is not possible to directly or even indirectly measure these variables in the fetus *in utero* using currently available technology. One, therefore, must look for related secondary variables that are practical for monitoring and are related to these critical variables. In the following paragraphs, some of these variables and the methods used to obtain them are described.

METHODS OF MONITORING BY A HUMAN OBSERVER

Any discussion of fetal monitoring must begin by pointing out an obvious, but often overlooked, fact that fetal monitoring does not always require expensive electronic equipment. Basic fetal monitoring can be carried out by a trained clinician using his or her hands, ears, and brain. A fetoscope is a stethoscope especially designed for listening to the fetal heart sounds through the maternal abdomen, which can be used to follow the fetal heart rate (FHR), and a hand placed on the abdomen over the uterus can be used to detect the relative strength, frequency, and duration of uterine contractions during the third trimester of pregnancy and labor. Any woman who has experienced labor will point out that the patient is also able to detect the occurrence of uterine contractions during labor. Although these techniques are only qualitative, they can be quite effective in providing information on the patient in labor and frequently represent the only fetal monitoring that is necessary in following a patient.

The main problems with this type of fetal monitoring are associated with convenience, fatigue, data storage and retrieval, and the difficulty of simultaneously processing multiple inputs. Electronic instrumentation can help to overcome these types of problems. Although electronic devices are less flexible and, at the present time, unable to interpret data as well as their human counterparts, the electronic devices can provide quantitative data, continuously monitor patients with minimal interruption of hospital routines, monitor for extended periods of time without fatigue, store data in forms that can be reevaluated at a later time, and, in some circumstances, make elementary logical decisions and calculations based on the data. Thus, the electronic monitor can serve as an extension of the clinician's data-gathering senses and provide a convenient method of recording and summarizing these data. Such a monitoring apparatus has the potential of allowing the clinician to optimize his or her limited available time.

FETAL HEART RATE MONITORING

The widespread use of electronic fetal monitoring was the result of the development of a practical method of sensing the fetal electrocardiogram and determining the instantaneous fetal heart rate from it. Much of the early work in this area was carried out by Dr. Edward Hon and associates who demonstrated a practical technique for directly obtaining the fetal electrocardiogram during labor (1). Techniques for obtaining the fetal heart rate can be classified as direct or indirect. The former involves invasive procedures in which a sensor must come into contact with the fetus to pick up the fetal electrocardiogram; the latter techniques are relatively noninvasive procedures where the mother's body serves as an intermediary between the fetus and the electronic instrumentation. In this case, the maternal tissue conducts a signal (electrical or mechanical) between the fetus and the surface of the mother's abdomen.

Direct Determination of Fetal Heart Rate

Direct FHR determinations are made from the fetal electrocardiogram (FECG). This signal is obtained by placing an electrode on the fetus and a second electrode in the maternal vaginal fluids as a reference point. These electrodes are connected to a high input impedance, high common-mode rejection ratio bioelectric amplifier. Such a direct connection to the fetus can be made only when the mother is committed to labor, the uterine cervix has dilated at least 2 cm, and the chorioamniotic membranes have been ruptured. In principle, it is possible to pass a wire through the maternal abdominal wall into the uterine cavity and beneath the fetal skin to obtain the FECG, and this technique was experimentally reported in the past (2). The method, however, places the mother and fetus at risk and is not used or suitable for routine clinical application. Indirect methods of determining the FHR that are available today make the application of such a technique unnecessary. Thus, the method that is directly used to obtain the FECG is to attach an electrode to the fetal presenting part through the cervix once the mother is committed to labor and the fetal membranes can be ruptured.

Although many different types of electrodes for obtaining the FECG have been described, best results are obtained when the electrode actually penetrates the fetal skin. The reason is illustrated in Fig. 1. The fetus lines in a bath of a amniotic fluid that is electrically conductive due to its electrolyte content. This amniotic fluid tends to short-out the fetal electrocardiogram on the skin surface, therefore, those potentials that are seen on the surface are relatively weak and affected by noise. Even if it were physically possible to place conventional chest electrodes on the fetus for picking up the electrocardiogram, a poor-quality signal would be obtained because of this shunting effect. The amniotic fluid does, however, provide a good central terminal voltage for the fetal electrocardiogram because it contacts most of the fetal body surface.

The fetal head is normally positioned against the dilating cervix when the mother is in labor, but it is possible for the fetal buttocks or other parts to present first. As the cervix dilates, the skin on the presenting part can be observed through the vagina, and it is possible to place an electrode on or within this skin. If this electrode penetrates the fetal scalp (or other exposed skin surface), it contacts the subcutaneous tissue. As an electrical resistance associated with the surface layers of the fetal skin exists, as indicated in Fig. 1, placing the electrode subcutaneously bypasses this resistance and gives a stronger, more reilable signal. Penetrating the skin also helps to physically keep the electrode in place on the fetus during movement associated with labor.

Various types of penetrating fetal electrodes ranging from fish hooks (3) to wound chips (4) have been developed over the years. Today, the most frequently applied electrode in the helical electrode originally described by Hon et al. (5). This electrode, as illustrated in Fig. 2, consists of a section of a helix of stainless-steel wire on an electrically insulating support. The tip of the wire is sharpened to a point that can penetrate the fetal skin when pressed against it and rotated to advance the helix. Typical dimensions of the wire helix are 5 mm in diameter with 1.25 turn of the wire exposed so that the tip of the helix is 2 mm from the surface of the insulator. A second stainless-steel electrode consisting of a metal strip is located on the opposite end of the insulator from the helix and is used to establish contact with the amniotic fluid through the fluid in the vagina. Lead wires connect the two electrodes to the external monitor.

The electrode is attached to the fetal presenting part by means of a special applicator device, which allows the electrode helix to be pressed against the fetal head to penetrate the skin and be twisted so that the entire wire is advanced beneath the surface of the skin until the insulating portion of the electrode contacts the skin. The flexible lead wires then exit through the vagina and can be connected to the monitoring electronics.

Signal Processing

In fetal heart monitoring, it is desired to have a continuous recording of the instantaneous heart rate. A fetal monitor must, therefore, process the electrocardiogram sensed by

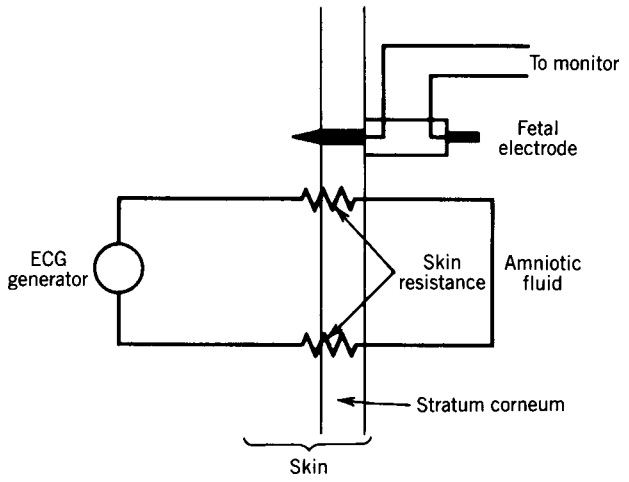


Figure 1. Schematic view and equivalent circuit of a direct fetal ECG electrode penetrating the fetal skin.

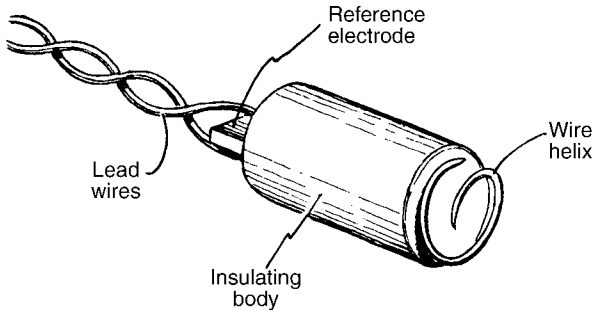


Figure 2. A helical direct fetal ECG scalp electrode of the type described by Hon et al. (5).

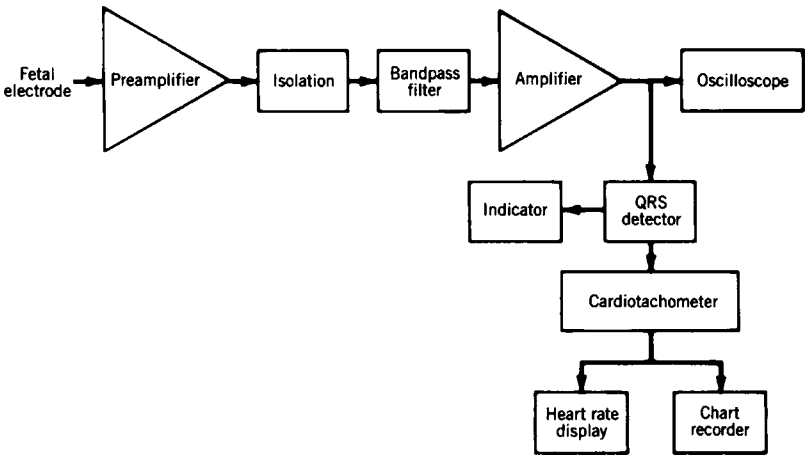


Figure 3. Block diagram of the electronic circuit of a direct fetal heart rate monitor.

the electrode and present the results on a computer monitor or paper printout. A typical electronic system for doing this recording is illustrated in Fig. 3. The signal from the fetal electrode has an amplitude ranging from 50 μV to 1.2 mV, which is amplified to a more suitable level for processing by an amplifier stage. The input of this aimplifier is electrically isolated and must have a very high input impedance and low lekage current because of the polarizable nature of most fetal electrodes. A high common-mode rejection ratio is also important, because a relatively strong maternal electrocardiogram signal is present on both electrodes. Another characteristic of the amplifier system is that it includes filtering to minimize the amplification of noise and motion artifact from the fetal electrode. As the purpose of the electronics is primarily to display the instantaneous heart rate, the filtering can distort the configuration of the fetal electrocardiogram as long as it does not affect the time at which the QRS complex appears, as it is used to determine the heart rate. For this reason, a relatively narrow band-pass filter is often used. The QRS complex contains higher frequencies than the rest of the electrocardiogram, and noise frequently has a predominance of the lower frequencies. For this reason, the bandpass filter can be centered at frequencies as high as 40 Hz.

Many ways exist for the QRS complex can be detected. The simplest of these is a threshold detector that indicates whenever the output voltage from the amplifier exceeds a preset threshold. The level of this threshold is adjusted such that it is usually greater than the noise level but less than the minimum amplitude of a typical QRS complex. The majro limitation of this method lies in the fact that wide variation exists in fetal QRS complex amplitudes. If the threshold level were fixed such that the minimum fetal QRS complex would cross it, this would mean that, for stronger signals, the threshold would not be optimal and interference from noise exceeding the threshold level would be quite possible. One way to get around this problem is to use some type of adaptive threshold. In this case, the threshold level is adjusted based on the amplitude of the electrocardiogram. A simple example of how this can be done is illustrated in Fig. 3. An automatic gain control circuit determines the amplitude of the fetal electrocardiogram at the output of the amplifier, and uses this amplitude to set the gain of that amplifier. This closed-loop control system, therefore, results in a constant-amplitude electrocardiogram appearing at the output of the amplifier even though the actual signal from the fetal electrode at the input might vary in amplitude from one patient to the next. Using a simple threshold detector with this automatic gain control will greatly improve the reliability of the fetal monitor in detecting true fetal heartbeats. Often, instead of using a simple threshold detector, a detector with hysteresis is used to minimize multiple triggers in the presence of noise. One can also use matched filters in the amplifier to recognize only true QRS complexes. A peak detector may be used to locate the true peak of the QRS complex (the R wave) for better timing, and pattern-recognition algorithms can be used to confirm that the detected pulse is most likely to be a fetal heartbeat. Of course, the best consideration for an accurate determination of the instantaneous fetal heart rate is to have a good signal at the input to the electronic instrumentation. Thus, care should always be taken to have the fetal electrode well positioned on the fetal presenting part so that one has the best possible input to the electronic system.

The cardiotachometer block of the fetal monitor determines the time interval between successive fetal QRS complexes and calculates the heart rate for that interval by taking the reciprocal of that time. Although it is obvious that such a cardiotachometer can introduce errors when it erroneously detects a noise pulse rather than a fetal QRS complex, other errors resulting from the method of heartbeat detection can exist. For a cardiotachometer to accurately determine the heart rate, it must measure the time interval over one complete cycle of the electrocardiogram. In other words, it must detect each QRS complex at the same point on the complex to ensure that the complete cycle period has been recorded. If one beat is detected near the peak of the R wave and the next beat is detected lower on the QR segment, the beat-to-beat interval measured in that case will be too short and the heart rate determined from it will be slightly greater than it should be. Normally, such a concern would be of only minimal significance, because the Q-R interval of the fetal electrocardiogram is short. However, because the variability in fetal heart rate from one interval to the next may be important in interpreting the fetal heart rate pattern, detection problems of this type can affect the apparent variability of the signal

and, perhaps, influence the interpretation of the pattern. The output from the cardiotachometer is recorded on one channel of a strip chart recorder and is also often indicated on a digital display. In both cases, the output is presented in the units of beats per minute, and standard chart speeds of 1 and 3 $cm \cdot mm^{-1}$ are used.

Indirect Sensors of Fetal Heart Rate

Indirect methods of sensing the fetal heart rate involve measurement of a physiologic variable related to the fetal heartbeat from the surface of the maternal abdomen. Unlike the fetal scalp ECG electrode, these methods are noninvasive and can be used prior to committing the patient to labor. The most frequently applied method is transabdominal Doppler ultrasound. Lesser used techniques involve transabdominal phonocardiography and electrodiography. Each of these techniques will be described in the paragraphs that follow.

Transabdominal Doppler Ultrasound. Ultrasonic energy propagates relatively easily through soft tissue, and a portion of it is reflected at surfaces where the acoustic impedance of the tissue changes such as at interfaces between different tissues. If such an interface is in motion relative to the source of the ultrasound, the frequency of the reflected signal radiation will be shifted from that of the incident signal according to the Doppler effect. This principle can be used to detect the fetal heartbeat from the maternal abdominal surface. A beam of ultrasound is passed through the abdomen from a transducer acoustically coupled to the abdominal surface. Frequencies around 2 MHz are generally used, because ultrasound of moderate source energy at this frequency can penetrate deep enough into the abdomen to sufficiently illuminate the fetus. Wherever this ultrasound beam encounters an abrupt change in tissue acoustical impedance, some of it is reflected back toward the transducer. If the incident ultrasound illuminates the fetal heart, some of it will be reflected from the various heart-blood interfaces in this organ. Many of these interfaces, such as the valve leaflets, experience periodic movement at rhe rate of the cardiac cycle. In the case of the valve leaflets, relatively high velocities can be obtained during portions of the cardiac cycle. Ultrasound reflected from these interfaces can, therefore, be significantly shifted in frequency so that the reflected wave can be identified at the maternal abdominal surface because of its frequency shift. This frequency shift will be related to the velocity of the reflecting surface and, hence, will be able to indicate each fetal heartbeat. Thus, by detecting and processing this reflected Doppler-shifted ultrasonic wave, it is possible to determine each heartbeat and, hence, the fetal heart rate.

A block diagram of a typical indirect fetal heart rate monitoring system using Doppler ultrasound is shown in Fig. 4(b). As continuous wave ultrasound is used, separate adjacent transducers are employed to establish the ultrasonic beam and detect the Doppler-shifted reflected waves. The reflected ultrasound signal is mixed with the transmitted wave, and beat frequencies are produced when a Doppler shift in frequency occurs for the reflected wave.

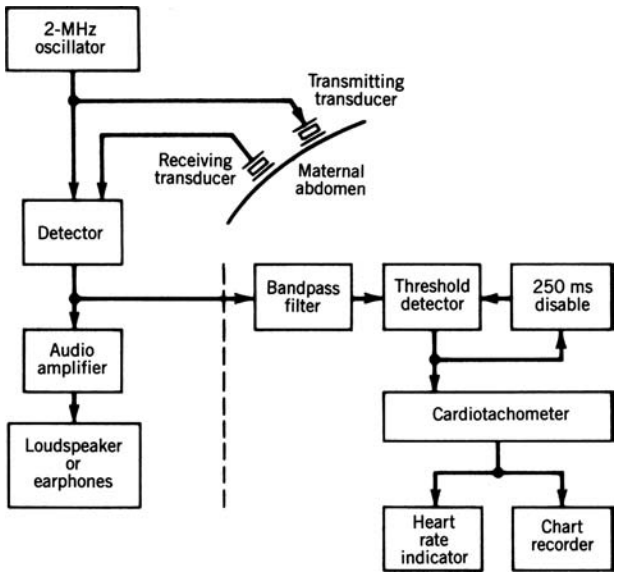


Figure 4. Block diagram of a Doppler ultrasound indirect fetal heart rate monitor. [Reprinted with permission from CRC Press (6).]

This beat frequency is amplified and used to indicate the occurrence of a heartbeat to a cardiotachometer. Many monitors also provide this signal to a loudspeaker to assist the clinical personnel in positioning the transducers for optimal signal pickup or for auditory monitoring.

The reflected ultrasound signal is different from an electrocardiogram, although it can also be used to identify various events in the cardiac cycle. A typical signal is illustrated in Fig. 5. Here one sees two principal peaks per heartbeat, one corresponding to valve opening and the other to valve closing. Actually, such signals can be quite useful in measuring fetal systolic time intervals, but from the standpoint of the cardiotachometer for determining heart rate, they can create problems. If the cardiotachometer is set to trigger at the peak of each wave it sees, as it is for the electrocardiogram, it could measure two beats per cardiac cycle and would give an erroneously high fetal heart rate. One way to avoid this problem is to detect only the first peak of the signal, and, once it is detected, to disable the detection circuit for a period of time that is less than the shortest expected beat-to-beat interval but longer than the time necessary for the second Doppler-shifted signal to occur. In this way, only one peak per cardiac cycle will be registered.

A second, more sophisticated method for detecting the fetal heartbeat involves the use of short-range autocorrelation techniques. The monitor recognizes the beat signal from the reflected wave for a given cardiac cycle and looks for a signal that most closely correlates with this signal over the period of time in which the next heartbeat is likely to occur. The time interval between that time when the initial wave was measured and the point of best correlation corresponds to a beat-to-beat interval of the fetal heart. Thus, instead of relying only on the peaks of the ultrasound signal, this method looks at the entire signal and, therefore, is more accurate. Some manufacturers of commercial fetal monitors claim their ultrasonic systems using this

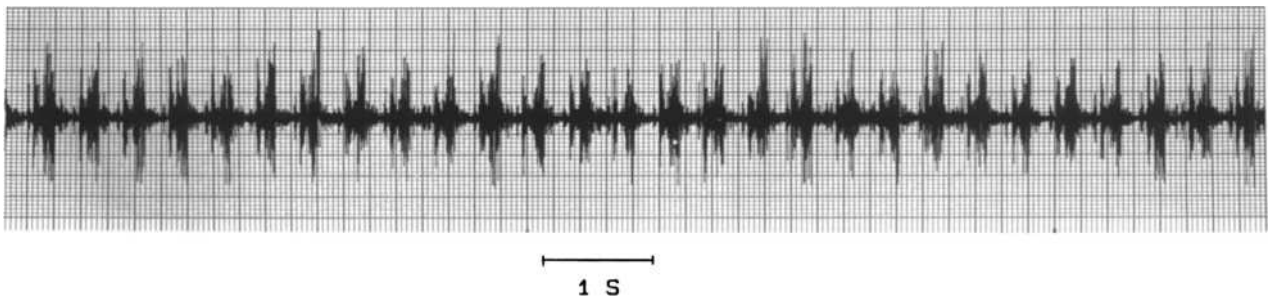


Figure 5. Illustration of the raw reflected ultrasound signal from the beating fetal heart *in utero*. Note that two ultrasonic bursts occur per cardiac cycle.

type of autocorrelation technique can detect fetal heartbeats as well as can be done by the direct electrocardiographic technique.

The major limitations of the Doppler ultrasound technique are related to its sensitivity to movement. As the Doppler effect will respond to movements of any tissue interfaces illuminated by the ultrasound beam with respect to the signal source, movement of the mother or fetus can result in Doppler-shifted reflected waves that are stronger than the cardiac signal, which, with artifact, can completely obliterate the signal of interest. Thus, this technique is really only reliable when the patient is resting quietly, and it often fails to provide reliable information in the active phase of labor. The other movement-related problem is that the fetus can move *in utero* so that the heart is no longer illuminated by the ultrasound beam or the orientation of the heart with respect to the ultrasonic beam is such that it produces only a minimum. Doppler shift in the reflected ultrasonic wave. Thus, while monitoring a patient, it sometimes necessary to reposition the ultrasonic sensors on the maternal abdomen from time to time because of the movement of the fetus.

Acoustic Pickup of the Fetal Heart. Until the advent of electronic fetal monitoring, the standard method of detecting the fetal heartbeat to measure the fetal heart rate was to use a fetoscope. When the bell of this instrument was firmly pressed against the maternal abdomen, fetal heart sounds could be heard and the heart rate could be determined from them. The acoustic method of indirect fetal heart monitoring follows the fetal heartbeat by a similar technique (7). A sensitive contact microphone is placed on the maternal abdomen over the point where the loudest fetal heart sounds are heard with a fetoscope. The signal picked up by this microphone is filtered to improve the signal-to-noise ratio, and the resulting signal drives a cardiotachometer to give the instantaneous fetal heart rate. The acoustic signal from the fetal heart is similar to the Doppler ultrasound signal in that it generally has two components per heartbeat. The cardiotachometer is set to trigger when the peak signal comes from the acoustic transducer, so it is possible that two apparent fetal heartbeats can exist for each cardiac cycle. Thus, as was the case for the Doppler ultrasound, it is wise to have a processing circuit that selects only the first of the two heart sounds to trigger the cardiotachometer. Unlike the ultrasonic Doppler signal, the fetal heart sounds produce sharp pulses that are narrower so that the detection of the time of the peak can be more precise. Thus, it is generally more accurate to measure the beat-to-beat interval using a peak detector with the acoustic signal than it is with the Doppler ultrasound. The use of the electrocardiogram still represents the best way to measure beat-to-beat cardiac intervals.

The major limitation of the acoustic method of detecting fetal heart sounds is the poor selectivity of the acoustic transducer. It not only is sensitive to the fetal heart sounds, but it will also respond to any other intraabdominal sounds in its vicinity. Also, a finite sensitivity to environmental sounds, exists which is an especially severe limitation for patients in active labor on a busy, noisy delivery service. For this reason, the acoustic method is limited primarily to patients who can lie quietly in a quiet environment to be monitored. The advent and use of the home-like labor/delivery rooms has helped to create an atmosphere that is more conducive to acoustic fetal heart monitoring, yet it is still not a widely applied approach.

The acoustic technique also has the limitation that when used for antepartum (before labor and delivery) monitoring the fetus can move such that the microphone is no longer ideally positioned to pick up the fetal heart sounds. Thus, it is frequently necessary to relocate the microphone on the maternal abdomen with this monitoring approach.

The major advantages of the acoustic method lie in the fact that not only is there better accuracy in determining the instantaneous fetal heart rate, but unlike the ultrasound method, which must illuminate the fetus with ultrasonic energy, the acoustic method derives its energy entirely from the fetus, and no possibility exists of placing the fetus at risk due to exogenous energy. As a result, investigators have considered the possibility of using the acoustic method for monitoring the high-risk fetus at home (8).

Abdominal Electrocardiogram. Although the fetus is bathed in amniotic fluid located within the electrically conductive uterus and maternal abdomen, one can still see small potentials on the surface of the maternal abodomen that correspond to the fetal electrocardiogram. These signals are generally very weak, ranging in amplitude from 50 to 300 μV. Methods of obtaining the abdominal fetal electrocardiogram and clinical application of the information have been known for many years as described by

Larks (9), yet signal quality remains a major problem. Nevertheless, some methods of improving the quality of the signal have been developed. These methods can allow a much more detailed fetal electrocardiogram to be obtained from the maternal abdomen under ideal conditions, and such electrocardiograms can be used in some cases for more detailed diagnosis than from just looking at heart rate. One of these methods involves applying signal-averaging techniques to several subsequent fetal heartbeats using the fetal R wave as the time reference (10). In this way, the full P-QRS-T wave configuration can be shown, but heart rate information and its variability will be lost.

As the fetal electrocardiogram at the maternal abdominal surface is very weak, it is easy for other signals and noise to provide sufficient interference to completely obliterate the fetal signal. Having the subject rest quietly during the examination and removing the stratum corneum of the skin at the electrode sites can reduce noise due to motion artifact and electromyograms from the abdominal muscles. Nevertheless, one major interference source exists that requires other types of signal processing to eliminate. This source is the component of the maternal electrocardiogram seen on the abdominal leads. This signal is generally considerably higher in amplitude than the fetal signal. Thus, observation of the fetal electrocardiogram could be greatly improved by the elimination or at least the reduction of the maternal signal. One method of reducing this signal involves simultaneously recording the maternal electrocardiogram from chest electrodes and subtracting an appropriate component of this signal from the abdominal lead so that only the fetal signal remains. Under idealized circumstances, this process can give a greatly improved abdominal fetal electrocardiogram, but the conditions for subtraction of the maternal signal are likely to vary during a recording session so that frequent adjustments may be necessary to maintain the absence of the maternal signal (11).

The abdominal fetal electrocardiogram can be used for antepartum fetal heart monitoring. In this case, the goal of the instrumentation is to collect fetal R-R intervals as done with the direct monitoring of the fetal electrocardiogram and to determine the instantaneous heart rate from these intervals, which strong maternal component in the abdominal fetal electrocardiogram can make a very difficult task electronically, and so most abdominal fetal electrocardiogram fetal heart rate monitors need to eliminate the maternal component of the abdominal signal. The substraction method described in the previous paragraph would be ideal for this purpose because if fetal and maternal heartbeats occur in approximately the same time, subtracting the maternal component should leave the fetal component unaffected. Unfortunately, because the conditions under which the maternal component is added to the fetal signal change from one minute to the next, it is not always practical to use this subtraction technique. Thus, a simpler technique that loses more information is used.

A typical abdominal fetal electrocardiogram is shown in Fig. 6 (lower panel) along with a direct fetal electrocardiogram taken from a scalp electrode and the maternal electrocardiogram taken from a chest lead. In the abdominal

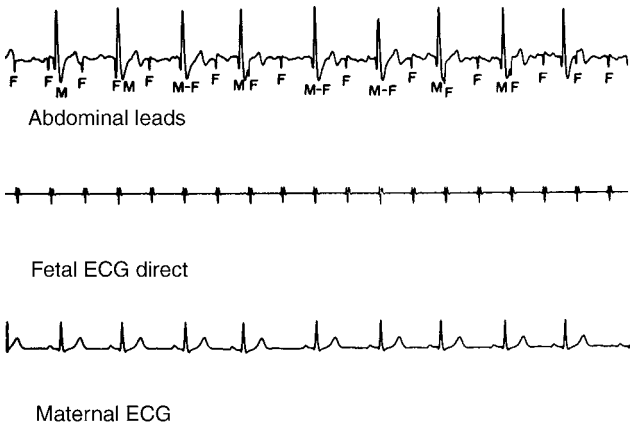


Figure 6. An example of a fetal electrocardiogram as obtained from the maternal abdomen. F, fetal QRS complexes; M, maternal QRS complexes. The direct fetal electrocardiogram and maternal electrocardiogram are recorded simultaneously for comparison. [Reprinted with permission from CRC Press (6).]

fetal electrocardiogram, fetal heartbeats are indicated by F and maternal heartbeats by M. Note that some beats exist where the fetal and maternal heartbeats occur at the same time. The strategy of the abdominal fetal electrocardiogram/fetal heart rate monitor is to monitor two signals, the maternal electrocardiogram from a chest lead and the fetal and maternal electrocardiograms from an abdominal lead. As shown in the block diagram in Fig. 7, the maternal electrocardiogram triggers a gate such that the input from the abdominal lead is interrupted every time a maternal beat occurs. Thus, this process eliminates the maternal component from the abdominal signal, but it can also eliminate a fetal QRS complex if it occurs at a time close to or during the maternal QRS complex. Thus, the cardiotachometer estimates the intervals where one or more fetal beats is missing. Due to the random relationship between maternal and fetal heartbeats, it is most likely that only one fetal beat would be missing at a time because of this mechanism, and so when maternal and fetal beats coincide, the fetal R-R interval should be approximately double the previous interval. Some monitors look for this condition and imply that it is the result of simultaneous maternal and fetal beats. The monitor, therefore, artificially introduces a fetal beat at the time of the maternal beat so that an abrupt (and presumably incorrect) change in the fetal heart rate will not occur.

Although such processing of the fetal signal makes the resulting heart rate recordings appear to have less artifact, this technique can loose some important information. For example, if the fetus suffers from a cardiac arrhythmia such as second-degree heart block, in which the fetal heart can miss a beat every so often, the monitor would reintroduce the missing beat, and this arrhythmia would not be detected. The principal advantage of the abdominal electrocardiogram method of fetal heart rate monitoring is that it can, under optimal conditions, provide the closest indirect observation of the fetal heart rate as compared with direct observations. No risk to the patient exists from this procedure, and inexpensive disposable electrodes can be used as the sensors.

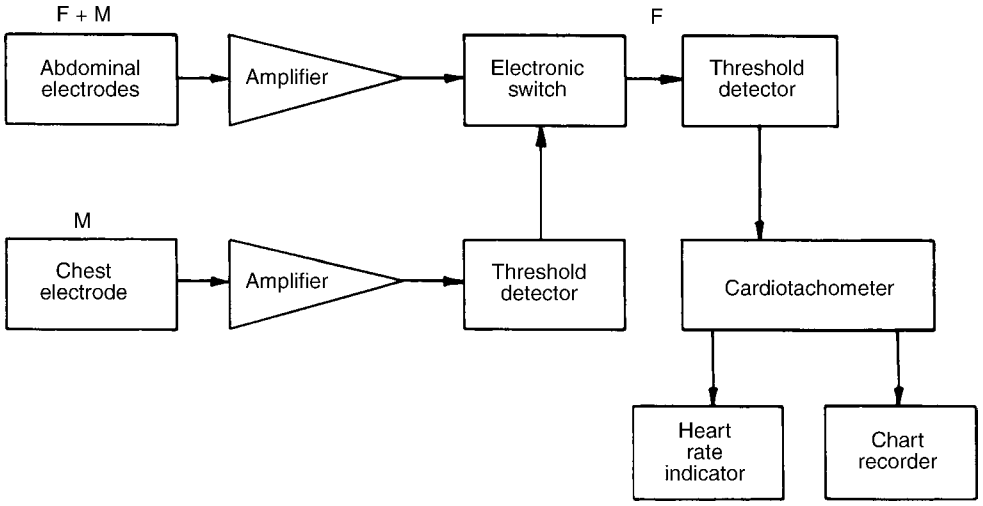


Figure 7. Block diagram of a monitor for processing the abdominal electrocardiogram shown in Fig. 6 using the anticoincidence detector method. [Reprinted with permission from CRC Press (6).]

The limitations of this method include its being based on a very weak signal in an environment that can contain a great amount of artifact. Thus, low signal-to-noise ratios are frequently encountered. Patients must be resting quietly for the method to work. Furthermore, electrodes must be optimally placed for good results, which requires some experimentation with different electrode sites, and skill is required on the part of the user in finding optimal electrode positions for a particular patient. Various signal processing techniques have been used over the years to get a more reliable fetal signal from the abdominal surface, but most of these techniques only improve signal quality under very special circumstances (12,13).

UTERINE CONTRACTIONS

Although the fetal heart rate is an important fetal variable for clinical monitoring, an equally important maternal variable is uterine activity. In fetal monitoring, one must detect the occurrence of uterine contractions, their frequency, their duration, and their intensity. As was the case with the fetal heart rate, it is possible to monitor uterine contractions by both direct methods and indirect methods.

Direct Monitoring of Uterine Contractions

Uterine contractions are periodic coordinated contractions of the myometrium, the muscle of the uterine wall. In an ideal method of direct measurement of uterine contractions, the tension and displacement of the myometrium would be measured, but this measurement cannot be done for routine fetal monitoring as only invasive methods of making this measurement exist. Uterine contractions, however, are reflected in increases in hydrostatic pressure of the amniotic fluid within the pregnant uterus. If this fluid is continuous and in a closed system, pressure increases resulting from uterine contractions should be seen throughout the amniotic fluid and should be related to the overall strength of the contraction but not necessarily to the tension at any one particular location in the myometrium.

The pressure change in the amniotic fluid during a contraction can be measured directly by coupling the amniotic fluid to a manometer, which consists of an electrical pressure sensor and the appropriate electronic circuitry for processing and indicating or recording the measured pressure. Intrauterine pressure can be measured by placing the pressure sensor directly in the amniotic fluid or by using a fluid-filled catheter to couple the amniotic fluid to an external pressure sensor. This latter method is the method most frequently employed in clinical fetal monitoring. The catheter used for coupling the amniotic fluid to an external pressure sensor can be placed only when the membranes surrounding the fetus have been ruptured, which should only be done if the patient is in labor. Unfortunately, rupture of the fetal membranes sometimes occurs spontaneously before the patient goes into labor or when the patient is in premature labor. It is unwise to place a catheter under these circumstances unless labor will be induced and the patient will deliver within 24 h. The reason is that the catheter can serve as a conduit for introducing infectious agents into the uterus or such agents can be introduced during the process of placing the catheter. When the distal tip of the catheter is within the amniotic fluid and its proximal end is connected to a pressure sensor at the same elevation as the distal end, the pressure seen at the sensor will, according to Pascal's law, be the same as that in the amniotic fluid. Thus, when a contraction occurs, the pressure increase will be transmitted along the catheter to the external pressure sensor.

Although the fluid-filled catheter provides a direct conduit from the amniotic fluid to the externally located pressure sensor, it can also be responsible for measurement errors. As was pointed out earlier, the proximal and distal ends of the catheter must be at the same level if one is to avoid the gravitational hydrostatic errors that give incorrect baseline pressure readings. Pascal's law applies only to the static solution where no fluid movement exists in the system. Once fluid movement occurs in the catheter,

pressure drops along the length of the catheter can result. Such fluid movement can occur when a small leak in the plumbing system exists at the sensor end of the catheter. Movement of the catheter itself or of the patient with respect to the pressure sensor can also produce alterations in the observed dynamic pressure. The most serious violation of Pascal's law is that once the fetal membranes have been ruptured, a truly closed system no longer exists. The fetal head or other presenting part approximated against the cervix does, indeed, isolate the intrauterine amniotic fluid from the outside world, but amniotic fluid can leak through the cervix, thus, no longer providing a static situation. Furthermore, after membranes have been ruptured, the total amount of amniotic fluid in the uterine cavity is reduced. It is possible that there might be local non-communicating pools of amniotic fluid delineated by fetal parts on one side and by the uterine wall on the other. The pressure in one of these isolated pools possibly can be different from that of another. The measured pressure will, therefore, be dependent on which pool contains the distal tip of the catheter. A statistical study by Knoke et al. has shown that when three identical catheters are placed in the pregnant uterus, the pressure measured by each can be considerably different, and differences of more than 10 mmHg (1.3 kPa) can be seen between different sensors (14), which is probably due to the fact that the distal tip of each catheter is located in a different part of the uterus and is coupled to a pocket of amniotic fluid at a different pressure.

Other problems exist that can affect the quality of intrauterine pressure measurement with the catheter-external sensor method. Poor recordings are obtained when catheter when catheter placement is not optimal and when limited communication exists between the fluid in the catheter and the intrauterine amniotic fluid. Many catheters in use today have a single hole either in the end or at the side of the catheter that communicates with the amniotic fluid. Mucus or vernix caseosa (a substance of a consistency similar to soft cheese that is found on the fetus) can obstruct or partially obstruct this opening resulting in poor quality recordings. When the distal tip of an open-ended intrauterine catheter becomes obstructed, the obstruction can frequently be "blown" off by forcing fluid through the catheter. In practice, this procedure is done by attaching a syringe filled with normal physiologic saline at the proximal end of the catheter near the pressure sensor and introducing fluid into the catheter when an obstruction is suspected.

It is possible to minimize these obstruction problems by modifying the catheter (15). Increasing the number of openings at the catheter tip is one of the simplest ways of minimizing obstructions. By placing an open-celled sponge on the catheter tip, it is possible to obtain a greater surface area in contact with the amniotic fluid because of the multiple openings of the sponge, which also tends to keep the tip of the catheter away from fetal or uterine structures, minimizing the possibility of complete obstruction or injury to the fetus. A small balloon placed at the distal tip of the catheter will prevent the fluid in the catheter from making actual contact with the amniotic fluid, and so this interface cannot be obstructed. As the wall of the balloon is flexible, the pressure in the fluid within the balloon will be equal to the pressure in the fluid surrounding the balloon plus the pressure resulting from the tension in the balloon wall itself. This system, however, has the disadvantage that it will respond to a direct force on the balloon as well as to hydrostatic pressure in the fluid outside of the balloon; thus, fetal movements or the entrapment of the balloon between the fetus and the uterine wall during a contraction can lead to erroneous pressure measurements.

Interuterine pressure can be directly measured using a miniature pressure sensor that can be placed in the intrauterine cavity (16). These devices are based on a miniature silicon pressure sensor that can be placed on the tip of a probe that has a similar appearance to an intrauterine catheter. In some cases, the probe at the catheter tip is no longer than the catheter itself, so the method of placement is the same as that used for the catheter. The advantage of the intrauterine pressure sensor is its location within the uterine cavity, which aviods artifact introduced by the fluid-filled catheter, and the problem of zeroing the pressure measurement system due to elevation differences is avoided because the sensor is at the pressure source. Investigators have compared the performance of intrauterine sensors with that of intrauterine catheters and have found the newer devices to provide equivalent data to the previously accepted technology (17).

Indirect Monitoring of Uterine Contractions

The clinician is able to sense uterine contractions by palpating (feeling) the maternal abdomen. Indirect uterine contraction sensors known as tocodynamometers are electrical sensors for doing the same thing. The basic principle of operation of these sensors is to press against the abdomen to measure the firmness of the underlying tissues. A contracting muscle will feel much more firm than a relaxed one. Most tocodynamometers carry out this function by pressing a probe against the abdomen and measuring its displacement.

The construction of a typical tocodynamometer is shown in Fig. 8. The sensor is held in place against the surface of

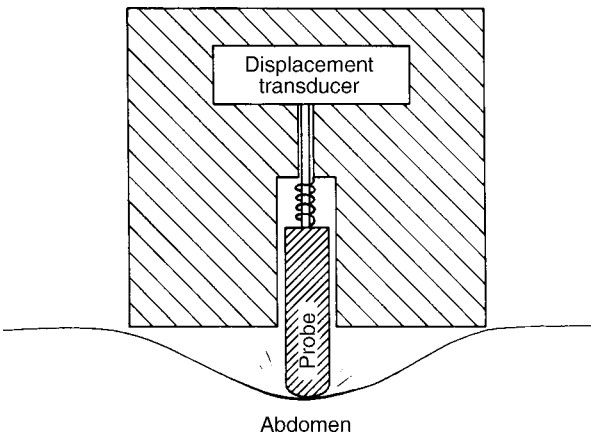


Figure 8. Schematic cross-sectional view of a tocodynamometer. [Reprinted with permission from CRC Press (6).]

the abdomen with an elastic strap. A movable probe protrudes beyond the surface of the sensor so that it causes a slight indentation in the abdominal wall. It is loaded by a spring, which makes it somewhat compliant; it can be either extended further into the abdominal wall or retracted into the body of the sensor depending on the firmness of the tissue of the abdominal wall under the probe. In some tocodynamometers, the spring tension, and hence the force that the probe exerts on the abdomen wall, can be adjusted by means of a small knob so that optimal operation of the sensor can be achieved. What a uterine contraction occurs, the abdominal wall will become tense, and it tends to push the probe back into the housing of the tocodynamometer. Following the contraction, the spring is again able to push the probe deeper into the abdomen. In some tocodynamometers this actual movement is very slight, whereas in others it can be as great as a few millimeters.

A displacement sensor inside the tocodynamometer provides an electrical signal proportional to the position of the probe. This displacement reflects myometrial activity. Different types of displacement sensors can be used in tocodynamometers. Including a strain gage on a cantilever arm, mutual inductance coils, a linear variable differential transformer, or a piezoelectric crystal.

The principal advantage of the tocodynamometer is the noninvasive way in which it measures uterine contractions. It is the only method that can be safely used before the patient is in active labor. It has serious limitations, however, in the quantitative assessment of labor. The method can be used only to quantitatively determine the frequency and duration of uterine contractions. Its output is only qualitative with respect to the strength of the contractions. Signal levels seen are a function of the position of the sensor on the maternal abdomen and the tension of the belt holding it in place. Signal amplitudes are also strongly related to maternal anatomy, and the method is virtually useless in obese patients. Many patients in active labor complain that the use of the tocodynamometer with a tight belt is uncomfortable and irritating.

Electronic Signal Processing

A block diagram for the uterine contraction channel of an electronic fetal monitor is illustrated in Fig. 9. The sensor can be either an internal or external pressure transducer or a tocodynamometer. Signals are sometimes filtered in the amplifier stages of the monitor because the uterine contraction information includes only dc and very low ac frequencies. Nevertheless, filtering is generally not necessary for high quality signals, and often the presence of artifact due to breathing movements of the patient is useful in demonstrating that the pressure measuring system is functional.

In some cases, it is necessary to adjust the baseline pressure to establish a zero reference pressure when using the monitor. In the case of direct uterine contraction monitoring when a single pressure sensor is always used with the same monitor, this adjustment should be made by the manufacturer, and additional adjustment should not be necessary. As a matter of fact, making such a zero-level adjustment control available to the operator of the monitor runs the risk of having significantly altered baseline pressures that can affect the interpretation of uterine basal tone. The adjustment of a zero-level control should not replace the requirement of having the proximal and distal end of the fluid-filled catheter at the same level. It is far better to adjust zero levels in uterine pressure monitoring by raising or lowering the external pressure transducer than by adjusting the electrical zero. On the other hand, when the tocodynamometer is used, no physiologically significant zero level exists. It is not possible to establish uterine basal tone with a tocodynamometer. Baseline levels are frequently dependent on how the tocodynamometer is attached to the patient and the structure of the sensor itself. In this case, it is reasonable to adjust the baseline level between uterine contractions so that the tracing conveniently fits on the chart. When doing so, it is important that the chart indicates that the uterine contractions were measured using a tocodynamometer so that the individual reading the chart does not ascribe inappropriate information to the baseline.

Uterine Electromyogram

The uterus is a muscle, and electrical signals are associated with its contraction as they are for any kind of muscle. These signals can be detected from electrodes on the maternal abdomen or the uterine cervix during uterine contractions. Garfield and others have studied these signals and suggested that they might be useful in managing

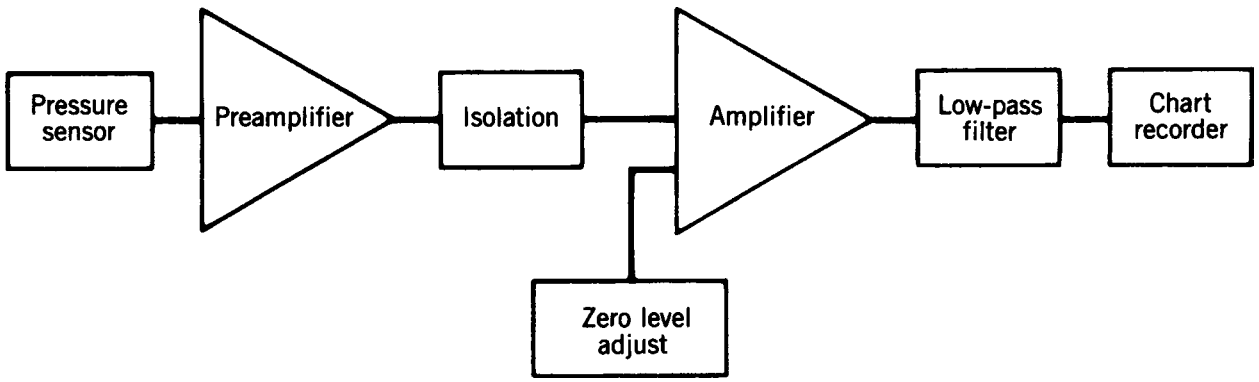


Figure 9. Block diagram of the signal processing electronics for intrauterine pressure measurement.

patients during pregnancy and, perhaps, even during labor (18–20). These techniques are still experimental and not yet ready for clinical application. Nevertheless, they offer a new approach to assessing uterine activity and the possibility of differentiating contractions leading to cervical dilatation from those that are nonprogressive.

THE FETAL CARDIOTOCOGRAPH

Electronic fetal monitoring is accomplished using a fetal cardiotocograph, such as illustrated in Fig. 10, which is basically a two-channel instrument with a two-channel chart recorder as the output indicator. One of the channels records the fetal heart rate, whereas the second channel records the uterine contractions. Most cardiotocographs are capable of accepting direct or indirect signals as inputs for each channel, although specialized monitors for antepartum assessment have only the indirect signal input capability. To aid clinicians in interpreting monitored patterns, most instruments use chart paper that is 70 mm wide for the fetal heart rate channel and calibrated from 30 to 240 beats·min^{-1}. The uterine contraction channel is 40 mm wide and calibrated with a scale from 0 to 100. The scale is only qualitative when a tocodynamometer is the input source, but corresponds to the pressure in millimeters of mercury when direct sensors of uterine contractions are used. The standard speeds for the chart paper are 1 or 3 cm/min. The use of a standardized chart and chart speed results in fetal heart rate—uterine contraction patterns that appear the same no matter what monitoring device is used—which is important because cardiotocograms are read by visually recognizing patterns on the chart. Changing the chart speed and scale significantly changes the appearance of the patterns even though the data remain unchanged. Thus, a clinician would have to learn to interpret patterns from each of the different types of monitors used if they each had different chart speeds and scales, because the same pattern can appear quite different when the chart speed or signal amplitude is changed.

Information Obtained from Fetal Cardiotocography

In interpreting a cardiotocogram, a clinician considers the heart rate and uterine contraction information separately as well as the interaction between the two signals. The frequency, duraction, and, in the case of direct monitoring, amplitude and baseline information have already been discussed. Similar types of information can be obtained from the directly and indirectly monitored fetal heart rate recordings. Specifically in the fetal heart rate channel, one looks for the average baseline value of the fetal heart rate, which should generally be in the range 120–160 beats·min^{-1} and when outside of this range can be cause for concern. The beat-to-beat variability of the fetal heart rate can also be an important indicator of fetal condition, and so the use of an instantaneous cardiotachometer in a cardiotocograph is mandatory. Certain recurring patterns in the fetal heart rate recording can also be important indicators of fetal condition. Sinusoidally varying fetal heart rate has been described as an ominous sign (21), and sometimes fetal cardiac arrhythmias can be detected by observing the heart rate pattern.

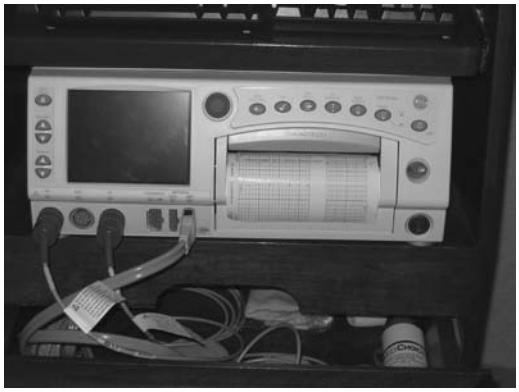


Figure 10. A commercially available fetal cardiotocograph. (Courtesy of Portage Health System, Hancock, Michigan.)

The information that is most frequently obtained from the cardiotocogram and applied clinically comes from both the heart rate and uterine contraction channels and is concerned with the relationship between these two signals. One can consider a uterine contraction as a stress applied to the fetus and the resulting changes in the fetal heart rate as the response to this stress. When the changes occur in direct relationship to the uterine contractions, they are referred to as periodic changes in the fetal heart rate. Several possibilities exist for fetal heart rate changes during and following a uterine contraction. One can see no change, an acceleration, or a deceleration in the fetal heart rate. In the case of decelerations, three basic patterns are seen, and representative examples of these are shown in Fig. 11. The different patterns are characterized by the shape of the deceleration curve and the temporal relationship of its onset and conclusion with the uterine contraction.

Early decelerations begin during the rising phase of the uterine contraction and return to baseline during the falling phase. They frequently appear to be almost the inverse of the uterine contraction waveform. Periodic decelerations of this type are thought to not represent a serious clinical problems.

Late decelerations refer to fetal heart rate decelerations that begin during a utterine contraction but late in the duration of that contraction. The rate of heart rate descent is not rapid, and the deceleration lasts beyond the end of the contraction and then slowly returns to baseline. Such patterns sometimes can be associated with fetal distress, although they should not be considered definitive of fetal distress.

The third type of periodic deceleration of the fetal heart rate is known as a variable deceleration. In this pattern, the deceleration of heart rate is sharp and can occur either early or late in the duration of the uterine contraction. Following the contraction, a rapid return to baseline values occurs. Sometimes one sees rapid return to baseline while the uterus is still contracting and then a rpaid fall back to the reduced heart rate. Variable decelerations have a flat "U" shape, whereas early and late decelerations represent a more smooth curve that could be characterized as shaped as the letter "V" with the negative peak rounded. Variable

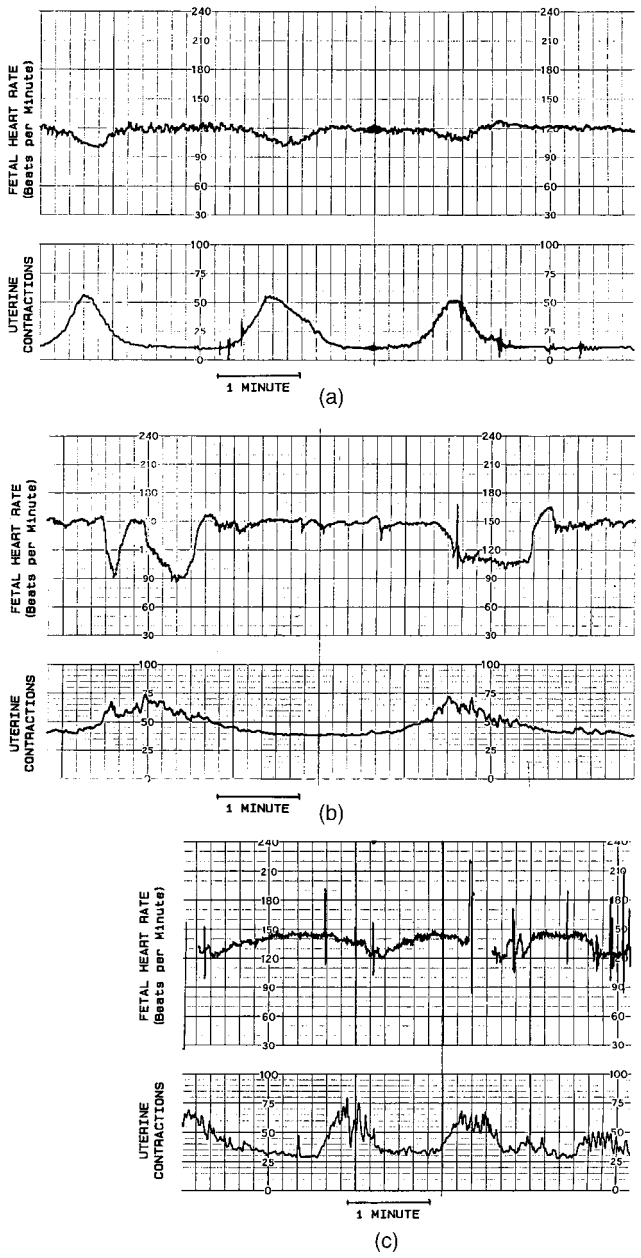


Figure 11. Examples of fetal cardiotocograms showing the three basic patterns: (a) early deceleration, (b) late deceleration, and (c) variable deceleration.

decelerations can be sometimes associated with involvement of the umbilical cord, and, in some cases, they can indicate the presence of fetal distress. As more recent clinical studies have shown that a simple relationship between late and variable decelerations and fetal compromise, does not exist these patterns are not considered to indicate fetal distress as they once were. Now clinicians refer to them as being "nonreassuring," and their presence should encourage the application of other clinical measures to evaluate the fetus.

These basic thoughts for interpreting the fetal cardiotocogram are very elementary and should not be used for diagnostic purposes. The reader is referred to the current obstetrical literature for more detailed descriptions of fetal cardiotocogram and their clinical significance.

Clinical Applications of Fetal Cardiotocography

Electronic fetal monitoring can be applied during the antepartum (before labor and delivery) and intraparturn (during labor and delivery) periods of pregnancy. In the anterpartum period, only indirect methods of fetal monitoring can be used. A primary application of fetal cardiotocography in this period is in nonstress testing. In this test, a cardiotocograph is applied to a patient who is resting quietly. In the United States, the ultrasonic Doppler method of detecting the fetal heart rate and the tocodynamometer are the sensors of choice. The patient is monitored for 1–2 h, and the cardiotocogram is examined for spontaneously occurring uterine contractions of fetal movements, which can also be indicated by the tocodynamometer. In some cases, the mother is asked to activate an event marker on the chart when she feels a fetal movement. The response of the fetal heart rate to these stimuli is noted in interpreting the cardiotocogram. In a reactive nonstress test, a response to these stimuli occurs, which is usually in the form of a brief fetal heart rate acceleration following the uterine contraction or fetal movement. Although nonstress testing is not routinely applied to apparently normal pregnancies, it is indicated for complications of pregnancy such as maternal diabetes, Rh sensitization, intrauterine growth retardation , decreased fetal movement, known fetal anomalies, oligohydrainnios or poly-hydramnios (too little or too much amniotic fluid), pregnancy-induced hypertension, pregnancy lasting beyond the normal 40 weeks, and other maternal and fetal complications.

A second antepartum test involving fetal cardiotocography is the oxytocin challenge test, which is usually applied when the nonstress test yields positive results, such as when fetal heart rate decelerations follow spontaneous uterine contractions or fetal movements. In this test, the patient is given intravenous oxytocin, a hormone that stimulates uterine contractions. The response of the fetal heart rate to the induced contractions is then examined, looking for the periodic changes described before.

Intrapartum monitoring of the fetal heart and uterine contractions can be carried out using the indirect techniques in early labor with the direct techniques applied during active labor. The indications for intrapartum fetal monitoring are controversial. Some obstetricians feel that all labors should be monitored whether they are complicated or not, whereas others feel that only those patients considered being at risk should have monitors. As monitoring is no longer considered to give a definitive diagnosis of fetal distress, some clinicians find it of little value and to not make use of the technology. As internal monitoring gives the most efficacious results, this modality is recommended in cases when it can be applied and the indirect methods do not give satisfactory results. Otherwise, indirect methods can be used as long as they give readable results.

The preceding paragraphs describe fetal cardiotocography as clinically applied in most major medical centers. Although this technology has the advantage of providing

continuous surveillance of the mother and fetus, it also has some limitations that prevent if from providing optimal information to obtain the earliest indications of fetal or maternal problems. The major limitation is in the data. Although uterine contractions provide a good indication of the intensity of labor, they do not necessarily indicate its effectiveness in dilating the cervix and expelling the fetus. If, in addition to uterine contractions, one should monitor whether labor is progressing, better information about some maternal aspects of labor and delivery could be obtained.

A similar argument can be made for the use of the fetal heart rate as the primary variable for evaluating the status of the fetus. Heart rate is a very non specific variable, and, in same cases, the fetus must be seriously compromised before any problem is detected by the heart rate. The goal of fetal monitoring as mentioned at the beginning of this article is to make certain that vital organs such as the fetal brain are adequately perfused so as to receive necessary nutrients and oxygen. Although the heart rate is related, it is not the principal variable for determining this perfusion.

Accepting these principal limitations for the variables measured, limitations still exist to the practical application of the cardiotocograph. Sensor placement, especially for indirect monitoring, is important for optimal recordings. The operator of the instrumentation, therefore, must be skilled in determining the best placement for the sensors. Most cardiotocographs are connected to the sensors on the patients by wires and catheters. Although this method is quite adequate while the patient is in bed, it can become quite inconvenient when it is necessary to transfer the patient to another location or to have the patient stand up and walk around. Many of these problems have been overcome by the use of biotelemetry for fetal monitoring (see Biotelemetry).

A final limitation of fetal cardiotocography is associated with the fact that some of the monitored patterns are not easily recognized and interpreted, which means that different clinicians looking at the data can see different things, lead to uncertain diagnoses. Periodic decelerations are usually not as clear, as illustrated in Fig. 11. Again, experience is an important factor here. Even when patterns can be readily determined, the relationship between certain patterns and pathology is not completely clear. As is so often the case medicine, one can only suggest from monitored tracets that certain problems might be present, and other tests need to be performed for confirmation.

OTHER METHODS OF FETAL MONITORING

Although the cardiotocogram is the usual method used to monitor the fetus, other techniques have been developed and experimentally employed to more accurately assess fetal status during the antepartum and intrapartum periods. One of these techniques, fetal microblood analysis is routinely used at major medical centers that care for patients deemed to have high risk pregnancies; the other techniques are still experimental or relatively new and have not enjoyed routine application at the time of this writing.

Fetal Microblood Analysis

About the time when electronic fetal monitoring was developed, Saling (22) was working on a new technique for taking a small sample of fetal capillary blood during active labor and measuring its hydrogen ion activity. This technique, known as fetal microblood analysis, made it possible to determine whether acidosis that could be associated with fetal distress was present during the labor. The technique involves observing a portion of the fetal presenting part (usually the scalp) through the cervix using a vaginal endoscope. By cleaning this portion of fetal skin and even, in some cases, shaving a small amount of hair from the scalp, the obstetrician is able to make a small superifical incision in the skin using a scalpel blade. A droplet of capillary blood will form at this site, and it can be collected in a miniature heparinized glass pipet. Generally, 100–300 μL of blood can be collected in this way. The blood sample is transferred to a special instrument designed to measure the pH of very small blood specimens. This instrument can be a part of a more extensive blood gas analysis instruments in a blood gas laboratory or it can be a relatively simple bedside device that uses disposable pH sensor cartrnidges. In either case, it is possible to measure the pH of this small sample and get the results back to the clinician within a few minutes of collecting the sample.

Chronic hypoxia can cause tissue and, hence, blood pH to drop as a result of the formation of acidic products of anaerobic metabolism such as lactic acid. Thus, if a blood sample is found to have a low pH (most clinical guidelines say lower than 7.2 or in some cases 7.15), it is possible that the fetus is experiencing some form of distress. Often, this technique is used in conjunction with fetal cardiotocography. When the cardiotocograph indicates possible fetal distress, such as when late decelerations are seen, the clinician can get a better idea as to whether distress is indeed present by performing a fetal microblood analysis. If the results indicate acidosis, the probability of actual fetal distress is higher, and appropriate actions can be taken.

A major limitation of the Saling technique is that it gives only an indication of the fetal acid-base status at the time the blood sample was taken. It would be far better to have a continuous or quasi-continuous measure of fetal tissue pH. Stamm et al. (23) have described a technique in which a miniature glass pH sensor is placed in the fetal scalp during active labor. This sensor can continuously record the pH of the fetal scalp. Clinical studies of this technique have shown that a drop in fetal tissue pH can occur along with a cardiotocographic indication of fetal distress (24). The major limitation of this as yet experimental technique is technical. The sensor is fragile, and it is not always possible to obtain efficacious recordings from it. Other sensor are under development in an attempt to overcome some of these limitations (25), yet this technique remains experimental due to the lack of practical devices.

Monitoring of Fetal Blood Gases

Many investigators have been interested in developing technology to continuously monitor fetal oxygenation during active labor and delivery. A review of some of the earlier techniques showed different types of oxygen sensors that

could be placed in the fetal scalp using structures similar to electrodes for directly obtaining the fetal electrocardiogram. Investigators also have used transcutaneous oxygen sensors on the fetus (26), and the most recent approach has been the uses of fetal pulse oximetry (27–29). In the transcutaneous oxygen case (see Blood Gas Measurement, Transcutaneous), a miniature sensor is attached to the fetal scalp once the cervix has dilated enough to make this physically possible, and fetal membranes have been ruptured. The technique is considerably more difficult than that for neonates, and it is important to have a preparation where the sensor surface is well approximated to the feal skin so no chance exists for environmental air to enter the electrode, as fetal Po_2 is much lower than that of the air. Most investigators who use this technique experimentally find that gluing the sensor to a shaved region of fetal scalp is the best technique to maintain contact (26).

Fetal pulse oximetry is performed in a similar way, but the sensor probe does not have to be physically fixed to the fetus as was the case for the transcutaneous oxygen tension measurement described above (27–29). Instead, the probe is a flat, flexible structure that contains ligh-emitting diodes at two different wavelengths and photodetector for sensing the reflected light. It is slid between the fetal head and the cervix once the head is engaged and membranes have been ruptured and is oriented so that the light sources and detector are pressed against the fetal skin by the uterine wall. The reflected light at each wavelength will vary in intensity as the blood volume in the fetal tissue changes over the cardiac cycle. As with the routine clinical pulse oximeter, the ratio of amplitudes of the reflected light at the different wavelengths is used to determine the oxygen saturation of the fetal arterial blood.

Recent improvements in the technology of making transcutaneous carbon dioxide sensors have allowed miniature transcutaneous sensors to be built in the laboratory. These have been applied to the fetus during active labor to continuously measure carbon dioxide tensions (30). All of these transcutaneous methods of measuring fetal blood gases are experimental at the time of this writing and have limitations regarding the technique of application and the quality of recorded information. Nevertheless, they present an interesting new approach to monitoring the fetus using variables more closely related to fetal metabolism and, hence, with greater potential for accurately detecting fetal distress.

Fetal Activity and Movements

The amount of time that the fetus spends in different activity states may be an important indicator of fetal condition. The fetus, as does the neonate, spends time in different activity states. Part of the time it may be awake and active, moving around in the uterus; at other times, it may be quiet and resting or sleeping. By establishing norms for the percentage of time that the fetus spends in these states, one can measure the activity of a particular fetus over a period of time and determine whether it falls within the normal classifications as a means of evaluating fetal condition.

One of the simplest ways to measure fetal activity is to have the mother indicate whether she feels fetal movements over a period of time, which can be done and recorded for several days as an assessment of fetal well-being. Fetal movements can also be detected by tocodynamometers. If the fetus is located under the probe of a tocodynamometer and moves or kicks, it can be detected as a short-duration pulse of activity on the chart recording from the sensor. Maternal movements can appear on this sensor as well, and so it is not easy to differentiate between the two. Timor-Trich et al. have developed a technique using two tocodynamometers to minimize this problem (31). By placing one over the fundus of the uterus and the second at a lower level, and recording the signals on adjacent channels of a chart recorder, fetal movements very often either are seen only on one sensor or produce pulses of opposite sign on the two sensors. Maternal movements, on the other hand, are usually seen on both sensors and are similar in shape and sign.

One of the most elegant methods of measuring fetal movements is to directly observe these movements using real-time ultrasonic imaging (see Ultrasonic Imaging). The main limitation of this technique is that an ultrasonographer must continuously operate the apparatus and reposition the ultrasonic transducer to maintain the best image. It also requires the subject to rest quietly during the examination. Although not believed to be a problem, no definite evidence currently exists that long-term exposure of the fetus to ultrasonic energy is completely safe.

One special type of fetal movement that is of interest to obstetricians is fetal breathing movement. The fetus goes through periods of *in utero* movement that are very similar to breathing movements. The relative percentage of these movements during a period of time may be indicative of fetal condition (32). Such movements can be observed using real-time ultrasound as described above. One can also select specific points on the chest and abdomen and use the ultrasonic instrument to record movements of these points as a function of time as one does for echocardiography (see Echocardiography). Measurement of fetal breathing movements by this technique also requires an experienced ultrasonographer to operate and position the instrumentation during examinations. For this reason, it is not a very practical technique for routine clinical application.

Fetal Electroencephalography

As one of the principal objectives of fetal monitoring is to determine if conditions are adequate to maintain fetal brain function, it is logical to consider a measure of this function as an appropriate measurement variable. The electroencephalogram (EEG) is one such measure that is routinely used in the neurological evaluation of patients. The EEG from the fetus during labor has been measured and shown to undergo changes commensurate with other indicators of fetal distress during labor and delivery (33,34). The monitoring of fetal EEG involves placement of two electrodes on the fetal scalp and measurement of the differential signal between them. These electrodes can be similar to the electrodes used for detecting the fetal electrocardiogram, or they can be electrodes especially designed for EEG. Of course, when either of these electrodes is used in the unipolar mode, the fetal electrocardiogram can be obtained.

Rosen et al. obtained good-quality fetal EEG recordings using a specially designed suction electrode (34). By observation of configurational or power spectrum changes in the EEG, it may be possible to indicate conditions of fetal distress.

Continuous Monitoring of Cervical Dilatation

In the routine method used to assess the progress of labor, the examiner places his or her fingers in the vagina and feels the uterine cervix to determine its length, position, and dilatation. Although this technique is simple and quick and requires no special apparatus, it has some limitations as well. It is an infrequent sampling method, and each time a measurement is made there can be discomfort for the patients as well as risk of intrauterine infection. The technique is also very subjective and depends on the experience of the examiner. A more reliable and reproducible technique that is capable of giving continuous records could be useful in the care of high-risk patients and patients with increased risk of intrauterine infection. Mechanical, caliper-like devices attached to opposite sides of the cervix have been described by Friedman (35) and others. These devices measure a cervical diameter with an electrical angular displacement transducer attached to the calipers. These devices are somewhat big and awkward, and Richardson et al. have optimised the mechanical structure by reducing its size (36). Other investigators have eliminated the mechanical calipers and used a magnetic field to measure the distance between two points on diametrically opposed sides of the cervix (37). In another technique for continuously monitoring cervical dilatation reported by Zador et al., ultrasound is used to measure the cervical diameter (38). A brief pulse of ultrasound is generated at a transducer on one side of the cervix and is detected, after propagating across the cervical canal, by a similar transducer on the opposite side. By measuring the transit time of the ultrasonic pulse between the two transducers, one can determine the distance between them, because ultrasound propagates through soft tissue a nearly constant known velocity. By generating an ultrasonic pulse once a second, a continuous recording of cervical dilatation as a function of time can be produced, which can be recorded either on an adjacent channel with the fetal cardiotocogram or on a separate display that generates a curve of cervical dilatation as a function of time known as a labor graph. Many clinicians plot such a curve as a result of their digital examinations of the cervix.

SUMMARY

As seen from this article, the use of biomedical instrumentation in obstetrical monitoring is fairly extensive, but the variables measured are not optimal in achieving the goals of fetal monitoring. Some of the newer and yet experimental techniques offer promise of getting closer to the question of whether vital structures in the fetus are being adequately perfused, but at the present time, none of these techniques are ready for general widespread application. Fetal monitoring is important if it can detect correctable fetal distress, as the results of such distress can remain with the newborn for life. It is important that the fetal monitoring techniques used will eventually benefit this patient. Some critics of currently applied fetal cardiotocography claim that the only result of fetal monitoring has been increase in the number of cesarean sections performed, and this might have a negative rather than positive effect on patient care. It is important that as this area of biomedical instrumentation progresses, biomedical engineers, clinicians, and device manufacturers are not only concerned with the technology. Instead , true progress will be seen when measured variables and their analysis are more closely and more specifically related to fetal status, and measurements can be made in a less invasive way without disturbance or discomfort. The application of this technology must be a benefit to the patients and to society.

BIBLIOGRAPHY

Cited References

1. Hon EH. Apparatus for continuous monitoring of the fetal heart rate. Yale J,Biol Med 1960;32:397.
2. Shenker L. Fetal electrocardiography. Obstet,Gynecol Surv 1966;21:367.
3. LaCroix GE. Fetal electrocardiography in labor: A new scalp electrode. Mich Med 1968;67:976.
4. Hon EH. Instrumentation of fetal heart rate and fetal electrocardiography. II. A vaginal electrode. Am J Obstet Gynecol 1963;86:772.
5. Hon EH, Paul RH, Hon RW. Electrode evaluation of the fetal heart rate. XI. Description of a spiral electrode. Obstet Gynecol 1972;40:362.
6. Roux JF, Neuman MR, Goodlin RC. Monitoring intrapartum phenomena. CRC Crit Rev Bioeng 1975;2:119.
7. Hammacher K. Neue methode zur selectiven registrierung der fetalen herzschlagfrequenz. Geburteh Frauenkeilk 1962;22:1542.
8. Talbert DO, Davies WL, Johnson F, Abraham N, Colley N, Southall DP. Wide bandwidth fetal phonography using a sensor matched to the compliance of the mother's abdominal wall. IEEE Trans Biomed Eng 1986;BME-33:175.
9. Larks SD. Normal fetal electrocardiogram, statistical data and representative waveforms. Am J Obstet Gynecol 1964;90:1350.
10. Cox JR. An algorithmic approach to signal estimation useful in electrocardiography. IEEE Trans Biomed Eng 1969;BME-16:3.
11. Nagel J, Schaldach M. Processing the abdominal fetal ECG using a new method. In: Rolfe P, editor. Fetal and Neonatal Physiological Measurements. London: Pitman; 1980. p. 9.
12. Tal Y, Akselrod S. A new method for fetal ECG detection. Comput Biomed Res 1991;24(3):296–306.
13. Assaleh K, Al-Nashash H. A novel technique for the extraction of fetal ECG using polynomial networks. IEEE Trans Biomed Eng 2005;52(6):1148–1152.
14. Knoke JD, Tsao LL, Neuman MR, Roux JF. The accuracy of measurements of intrauterine pressure during labor: A statistical analysis. Comput Biomed Res 1976;9:177.
15. Csapo A. The diagnostic significance of the intrauterine pressure. Obstet Gynecol Surv 1970;25:403–515.
16. Neuman MR, Picconnatto J, Roux JF. A wireless radiotelemetry system for monitoring fetal heart rate and intrauterine pressure during labor and delivery. Gynecol Invest 1970;1(2):92–104.
17. Devoe LD, Gardner P, Dear C, Searle N. Monitoring intrauterine pressure during active labor. A prospective comparison of two methods. J Reprod Med 1989;34(10):811–814.

18. Garfield RE, Maner WL, Maul H, Saade GR. Use of uterine EMG and cervical LIF in monitoring pregnant patients. Brit J Obstet Gynecol 2005;112(Suppl 1):103–108.
19. Garfield RE, Maul H, Shi L, Maner W, Fittkow C, Olsen G, Saade GR. Methods and devices for the management of term and preterm labor. Ann N Y Acad Sci 2001;943:203–24.
20. Devedeux D, Marque C, Mansour S, Germain G, Duchêne J. Uterine electromyography: A critical review. Am J Obstet Gynecol 1993;169(6):1636–1653.
21. Egley C. The clinical significance of intermittent sinusoidal fetal heart rate. Am J Obstet Gynecol 1999;181(4):1041–1047.
22. Saling E. A new method of safeguarding the life of the fetus before and during labor. J Int Fed Gynaecol Obstet 1965; 3:100.
23. Stamm O, Latscha U, Janecek P, Campana A. Development of a special electrode for continuous subcutaneous pH measurement in the infant scalp. Am J Obstet Gynecol 1976;24:193.
24. Lauersen NH, Hochberg HM. Clinical Perinatal Biochemical Monitoring. Baltimore: Williams & Wilkins; 1981.
25. Hochberg HM, Hetzel FW, Green B. Tissue pH monitoring in obstetrics. Proc 19th Annual Meeting Association Advanced Medical Instrumentation. Boston: 1984. 36.
26. Huch A, Huch R, Schneider H, Rooth G. Continuous transcutaneous monitoring of fetal oxygen tension during labour. Br J Obstet Gynaecol 1977;84(Suppl 1):1.
27. Johnson N, Johnson VA, Fisher J, Jobbings B, Bannister J, Lilford RJ. Fetal monitoring with pulse oximetry. Br J Obstet Gynaecol 1991;98(1):36–41.
28. König V, Huch R, Huch A. Reflectance pulse oximetry–Principles and obstetric application in the Zurich system. J Clin Monit Comput 1998;14(6):403–412.
29. Yam J, Chua S, Arulkumaran S. Intrapartum fetal pulse oximetry. Part I: Principles and technical issues. Obstet Gynecol Surv 2000;55(3):163–172.
30. Lysikiewicz A, Vetter K, Huch R, Huch A. Fetal transcutaneous Pco_2 during labor. In: Huch R, Huch A, editors. Conenuous Transcutaneous Blood Gas Monitoring. New York: Dekker; 1983. p. 641.
31. Timor-Trich I, Zador I, Hertz RH, Roser MG. Classification of human fetal movement. Am J Obstet Gynecol 1976;126: 70.
32. Boddy K, Dawes GS. Fetal breathing. Br Med Bull 1975;31:3.
33. Mann LI, Prichard J, Symms D. EEG, EKG and acid-base observations during acute fetal hypoxia. Am J Obstet Gynecol 1970;106:39.
34. Rose MO, Scibetta J, Chik L, Borgstedt AD. An approach to the study of brain damage. Am J Obstet Gynecol 1973;115:37.
35. Friedman EA, Micsky L. Electronic cervimeter: A research instrument for the study of cervical dilatation in labor. Am J Obstet Gynecol 1963;87:789.
36. Richardson JA, Sutherland IA, Allen DW. A cervimeter for continuous measurement of cervical dilatation in labor-preliminary results. Br J Obstet Gynaecol 1978;85:178.
37. Kriewall TJ, Work BA. Measuring cervical dilatation in human parturition using the Hall effect. Med Instrum 1977;11:26.
38. Zador I, Neuman MR, Wolfson RN. Continuous monitoring of cervical dilatation during labor by ultrasonic transit time measurement. Med Biol Eng 1976;14:299.

Reading List

Hon EH. An Atlas of Fetal Heart Rate Patterns. New Haven, (CT): Marty Press; 1968.

See also ELECTROCARDIOGRAPHY, COMPUTERS IN; ELECTROENCEPHALOGRAPHY; INTRAUTERINE SURGICAL TECHNIQUES.

FETAL SURGERY. See INTRAUTERINE SURGICAL TECHNIQUES.

FEVER THERAPY. See HYPERTHERMIA, SYSTEMIC.

FIBER OPTICS IN MEDICINE

MARK D. MODELL
LEV T. PERELMAN
Harvard Medical School
Beth Israel Deaconess Medical Center
Boston, Massachusetts

INTRODUCTION

In the first edition of the Wiley *Encyclopedia of Medical Devices and Instrumentation*, our friend and colleague Max Epstein, Professor Emeritus at Northwestern University, wrote an excellent article on Fiber Optics in Medicine. Now, almost 20 years later, applications of fiberoptics in medicine underwent dramatic changes and expansions. Thus on Max's recommendation, this article has been updated and rewritten for the application of fiber optics in medicine for the second edition while keeping, where it was appropriate, the original text.

For a long time optical fibers in medicine have been primarily used in endoscopy, where they have been employed for transmission of illumination to the distal end of the fiberoptic endoscope and for conveying images for the visualization of otherwise inaccessible organs and tissues. However, in the past 20 years, the science of biomedical optics of the light–tissue interaction has been dramatically advanced. The new methods of imaging, often based on substantial utilization of the optical fibers, for example, optical coherence tomography (OCT) and fiber-based confocal microscopy have been introduced. Also, the new methods of the diagnostics employing various spectroscopic techniques, for example reflectance spectroscopy, light scattering spectroscopy (LSS), fluorescence spectroscopy, and Raman spectroscopy have been developed. To be useful in the diagnosis of tissue in the lumens of the human body, these methods utilize fiber-based catheters. Phototherapy and diagnoses of internal organs also require optical fiber catheters.

The goal of this article is to give the reader basic tools necessary to understand principles of biomedical fiber optics and its applications. In addition to diagnostic, imaging, and therapeutic applications that are described in this article, optical fibers have been employed in a number of biomedical applications, for example, laser surgery and fiber-based transducers for monitoring physiologically important parameters (temperature, pressure, oxygen saturation, blood flow). All those subjects have been covered in detail in the dedicated articles of this encyclopedia.

The structure of this article is the following. The first section provides general physical and engineering principles of fiber optics needed to understand the rest of the article. It discusses the physics of total internal reflection and throughput, fiber propagation modes, optical fiber construction, and

types of fibers. The next section provides the reader with the review of illumination applications of fibers in medicine. The third section discusses the diagnostic applications of the biomedical fibers, including imaging and spectroscopy. The last section reviews therapeutic applications of fibers.

GENERAL PRINCIPLES OF FIBER OPTICS

The Physics of Fiber Optics: Total Internal Reflection

For almost 400 years, it has been well known from classical optics that when light enters a medium with a lower refractive index it bends away from the imaginary line perpendicular to the surface of the medium. However, if the angle of incidence is sufficiently large, the angle in the medium with lower refractive index can reach 90°. Since the maximum possible angle of refraction is 90°, the light with higher angles of incidence would not enter the second medium and will be reflected entirely back in the medium from which it was coming. This particular angle is called the critical angle and the effect is called total internal reflection.

The effect of total internal reflection that makes an optical fiber possible is depicted in Fig. 1. The light in an optical fiber propagates through the medium with high refractive index, which is called core (usually the silica glass). The core is surrounded by another medium with lower refractive index, which is called cladding (usually another type of the silica glass). If light reaches the core–cladding interface with an incident angle higher than the critical angle it will be entirely reflected back into the optical fiber. However, if light reaches the core–cladding in terface with an incident angle lower than the critical angle, it will leave the core and will be lost. Thus, optical fibers can propagate light only at a certain angular composition, which depends on the critical angle of the core–cladding interface and thus on the refractive indexes of the core and the cladding.

Light, being continuously reflected from the core–cladding interface, can propagate very far through an optical fiber, even if the fiber is bent or is placed in a highly absorptive medium. However, if the fiber is bent too much some of the light can escape the core of the fiber. Other sources of losses are impurities in the glass. Typical optical fiber has 50–60% losses per kilometer of its length.

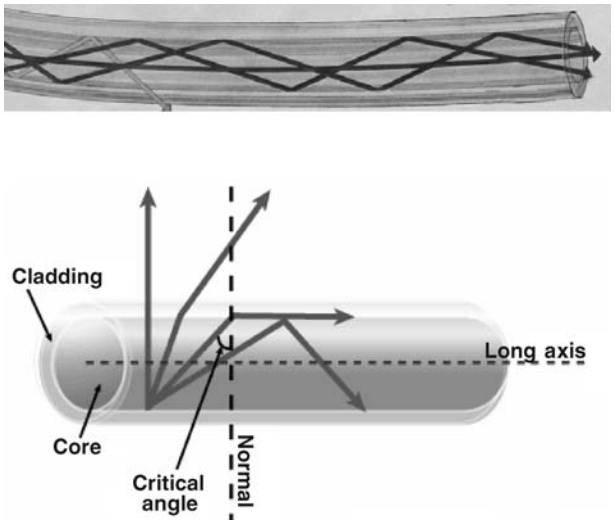


Figure 1. Total internal reflection confines light within optical fibers. (From The Basics Of Fiber Optic Cable, ARC Electronics.)

Throughput

There is a limit on how much light an optical fiber can transmit. Intuitively, it should be limited by an acceptance angle of a fiber and its area. This rule is often formulated as a conservation of throughput principle, which is a very general principle in optics.

For a given aperture in an optical system, the throughput T is defined by

$$T = S(\mathrm{NA})^2$$

where S is the area of the aperture, and NA is numerical aperture of the optical element equal to the sine of the maximum divergence angle of radiation passing through the aperture (1). Conservation of throughput says that it can be no greater than the lowest throughput of any aperture in the system (2). It is very important to take the throughput of the fiber into consideration when one calculates the power, which can pass through the fiber.

Propagation Modes

By solving Maxwell's equations for an optical fiber one can find various patterns of the electromagnetic field inside the fiber. Those patterns are modes of the fiber. There are two main types of an optical fiber. The fiber can be either single mode or multimode (see Fig. 2). The difference between these types is the number of modes that the fiber can propagate.

A single-mode fiber is a fiber through which only one mode can propagate. Usually, a single-mode fiber has a very small core, ~5–10 μm in diameter. Due to their size and also because of their small NA, these fibers have a very small throughput. In medicine, such fibers are used, for example, in confocal microscopy because of the requirement for the small core diameter of the fiber tip (see the section Fiber-Based Confocal Microscopy) and OCT. However, in OCT they are used not for their size, but because the coherence of the light pulse is critical for OCT to work (this is described in detail in the section Optical Coherence Tomography characterization of flexible imaging fiber Bundles).

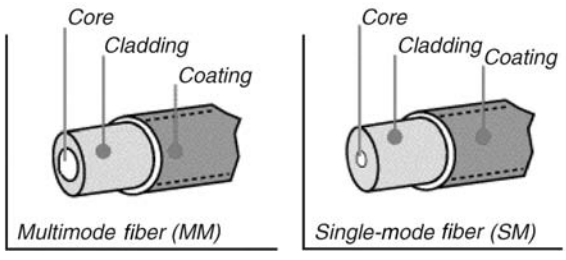


Figure 2. Types of fibers. (From Basic Principles of Fiber Optics, Corning Incorporated © 2005.)

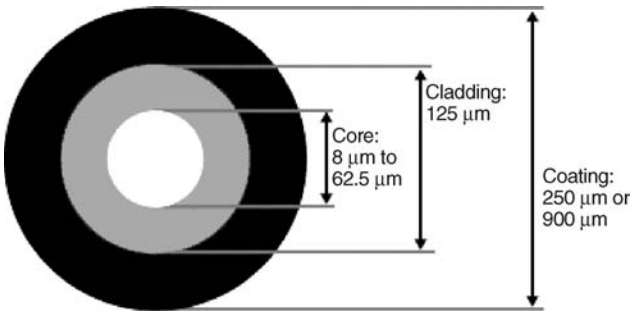


Figure 3. Fiber components. (From Basic Principles of Fiber Optics, Corning Incorporated © 2005.)

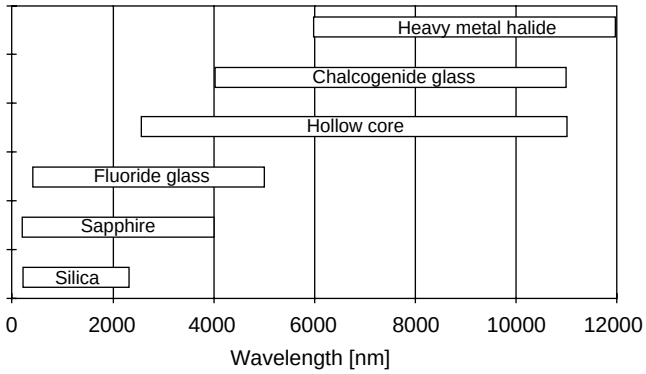


Figure 4. Transmission ranges of materials used for fibers (From Ref. 3)

The core of the multimode fiber can be much larger, somewhere between 50 and 1500 μm in diameter. These fibers are ideal for light delivery and collection, and are used in medicine when the throughput is important (see the Section Spectroscopy).

In addition to the single-mode and multimode fibers, there is another important type, that is, a polarization-maintaining fiber. The polarization-maintaining capability of a fiber is provided by the induced birefringence in the core of the fiber. This birefringence causes the polarization in these fibers to remain in the axis that it was launched into the fiber and is not changing randomly as in the regular fiber.

There are two main types of the polarization-maintaining fibers: one with the geometrical birefringence and another with the stress-induced birefringence. Geometrical birefringence is created by the elliptically shaped core. The stress-induced birefringence is created by using two stress rods as a core.

In the application where fibers are exposed to the physical stress and temperature changes, the former type is used mostly since it maintains its polarization.

Optical Fiber Construction

Optical fiber consists of three main components: core, cladding, and coating as shown in Fig. 3. The fiber is constructed by drawing a solid glass rod in a high purity graphite furnace. The rod consists of a core with high refractive index inside a low refractive index cladding. Thus both core and cladding are produced from a single piece of glass.

After core and cladding are formed, the protective coating is applied to the fiber. This protective coating is called a jacket and it guarantees that the fiber is protected from the outside environment.

Types of Fibers

Transmission ranges of materials used for fibers are shown in Fig. 4 (3). Most lasers operate in the range from 300 to 2500 nm, where silica fibers have the best overall properties and thus are commonly used.

Ultraviolet, Visible, and Near-Infrared Fibers. Ultraviolet(UV), visible, and near-infrared (NIR) light spans the range from 200 nm to 2.5 μm. Visible and near-IR light propagates in the silica-based fibers with practically no losses due to the low absorption of the order of tenths of percent per meter. In the IR range (wavelength >2.4 μm) and UV (wavelength <400 nm) absorption is higher. The silica-based fiber absorption is caused primarily by hydroxyl radicals (OH), and thus is determined by OH concentration resulting from the presence of free water during the fiber production. Low OH concentration determines excellent transmission of these fibers in the NIR range up to 2.4 μm. At wavelengths longer than 2.5 μm, the absorption of silica limits the use of silica fibers. In the UV range, most of the silica fibers are usable down to 300 nm, particular the fibers with a high OH concentration. For shorter wavelengths fibers with both core and cladding made of silica, silica–silica fibers are used.

For applications in wavelengths <230 nm, special attention should be paid to the solarization effect caused by the exposure to the deep UV light. The solarization effect is induced by the formation of "color centers" with an absorbance at the wavelength of 214 nm. These color centers are formed when impurities (like Cl) exist in the core and cladding fiber materials, and form unbound electron pairs in the Si atom, which are affected by the deep UV radiation. Recently, solarization resistant fibers have been developed. It consist of a silica core, surrounded by silica cladding that is coated in aluminum, which prevents the optical fiber from solarizing. The fiber preform (a high grade silica rod used to make the fiber) is hydrogen loaded in a hydrogen-rich environment that helps to heal the silicone–oxygen bonds broken down by UV radiation.

As far as power-handling capability is concerned, the typical glass optical fiber is quite adequate in applications where the laser beam energy is delivered continuously or in relatively long pulses such that the peak power in the optical fiber does not exceed power densities of several megawatts per square millimeter. When the laser energy is delivered in very short pulses, however, even a moderate energy per pulse may result in unacceptable levels of-peak power. Such may be the case of Nd-YAG lasers, operating in mode-locked or Q-switched configurations, which produce laser beam energy in the form of pulses of nanosecond duration or less. On the other hand, excimer lasers, which are attractive in a number of applications (4), generate energy in the UV range of the spectrum (200–400 nm) in

very short pulses; they, therefore, require solarization-resistant silica–silica fibers, which can transmit light en ergy at such short wavelengths and, at the same time, carry the high power densities.

The limitation on power-handling capability of a glass optical fiber is due to several nonlinear effects, for example, Raman and Brillouin scattering (5), avalanche breakdown (6), and self-focusing (7). Stimulated Raman scattering, which occurs when, because of molecular vibrations, a photon of one wavelength, say that of the laser, is absorbed and a photon of another wavelength, known as a Stoke's photon, is emitted, has been observed at power densities of $6\,MW{\cdot}mm^{-2}$. The time varying electric field of the laser beam generates, by electrostriction, an acoustic wave, which in turn modulates the refractive index of the medium and gives rise to Brillouin scattering. Thus, Brillouin scattering is analogous to stimulated Raman scattering wherein the acoustic waves play the same role as the molecular vibrations. Although the Brillouin gain is higher than the one measured for the stimulated Raman scattering, the latter is usually the dominant process in multimode fibers (8).

Under the influence of an intense electromagnetic field, free electrons, which may exist in the optical fiber as a result of ionized impurities, metallic inclusions, background radiation, or multiphoton ionization, are acceler ated to energies high enough to cause impact ionization within the medium. If the rate of electron production due to ionization exceeds the electron loss by diffusion out of the region, by trapping, or by recombination, then an avalanche breakdown may occur, resulting in material damage. If high enough power densities ($>100\,MW{\cdot}mm^{-2}$) are applied to the fiber core, avalanche breakdown is the main mechanism of permanent damage to the optical fiber. The fiber surface should be polished and chemically processed with great care to avoid reduction in the damage threshold level of the fiber surfaces. The latter is usually lower by two orders of magnitude than that of the bulk material as a result of the presence of foreign materials embedded during improper polishing or because of mechanical defects. The threshold of induced Raman and Brillouin scattering and avalanche breakdown can be further substantially reduced by self-focusing of the laser beam. Self-focusing may occur when the refractive index of the nonlinear medium increases with beam intensity. The possible physical mechanisms involved are vibration, reorientation, and redistribution of molecules, electrostrictive deformation of electronic clouds, heating, and so on. Thus, a laser beam with a transverse Gaussian profile causes an increase in the refraction index in the central portion of its path of propagation, and becomes focused toward the center. Self-focusing is counteracted by the diffraction of the beam and the balancing effects of the two determine the threshold of power that causes self-focusing; for glass it was found to be $\sim 4\,MW$. Damage to optical fibers can also occur if a pulsed-laser beam is not properly aligned with the entrance face of the fiber (8,9).

IR Fibers. For the IR region beyond 2500 nm, materials other than silica are being used. These IR fibers can be classified into three categories: IR glasses fibers, crystalline fibers, and hollow fibers (10,11). Fluorozirconate and fluoroaluminate glass fibers can be used in the 0.5–4.5 μm region. Commercially, they are produced in diameters from 100 to 400 μm. Because of their low melting point, they cannot be used $>150\,^{\circ}C$; however, they have a high damage threshold. The refractive index is similar to silica (~ 1.5) and the transmission is >95% for several meters. For longer wavelengths (4–11 μm) chalcogenide glass fibers are available in diameters of 150–500 μm. The disadvantage of these fibers is that they are mechanically inferior to silica fibers and are toxic. They have high Fresnel reflection because of the high refractive index (2.8) and relatively high absorption; as a result they have low transmission [losses are several tens of percent per meter (12)].

The crystalline fibers can be better candidates for the mid-IR range. Sapphire can be grown to a single crystal with a diameter of 200–400 μm, which is strong, hard, and flexible (13). It can be used up to a wave length of 4 μm. Sapphire, however, has a high index of refraction (1.75), which produces rather high reflection losses at each surface. Silver halide and thallium halide polycrystalline alloys (e.g., KRS-13), in contrast, can successfully transmit even high power CO_2 light (14). From these alloys, good quality fibers are manufactured with high transmission of a few $dB{\cdot}m^{-1}$ and that are insoluble in water, are nontoxic, and are fairly flexible.

Hollow fibers are built as flexible tubes, which are hollow inside, that is with air. They transmit light in the whole IR range with high efficiency (15–17). One type of these fibers comprises metallic, plastic, or glass tubing that is coated on the inside with a metallic or dielectric film with a refractive index $n>1$ (18). Another type has the tubing coated with a dielectric coating of $n<1$ for 10.6 μm on the inside of hollow glass (15) or crystalline tubes (19). The losses are 0.4–7 $dB{\cdot}m^{-1}$ depending on the core size. The losses due to bending are inversely proportional to the core radius. Power transmissions >100 W have been achieved. The damage threshold for high power densities is comparable with that of solid core fibers. It has been reported that the dielectric-coated metallic hollow fiber is the most promising candidate for IR laser light transmission. The standard hollow fiber is 2 m in length with an inner diameter of 0.7 mm and has transmission >75% of Er-YAG or CO_2 laser light under practical usage conditions (20).

ILLUMINATION APPLICATIONS

Introduction and Applications

Optical fibers are used for various illumination needs. The use of fiber optics bundles allows illuminating the desired area without the problem associated with the presence of the lamp-based light source.

For example, the fiber bundle brings the visible light to illuminate the area under examination with the colposcope and surgical microscope while the light source is placed in the area not interfering with the physician's activities.

In endoscopy, the fiber optic bundles are incorporated into the small diameter flexible cylindrical body of the endoscope, which is dictated by the necessity to pass through narrow (<2 cm at the most) pathway of lumens

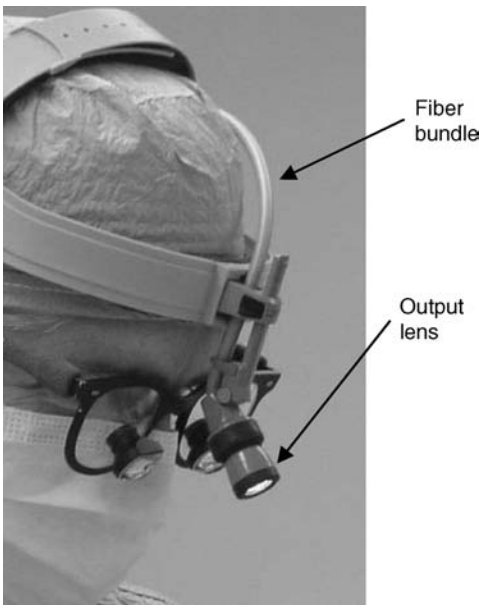


Figure 5. Headlight attached to the head bend on the surgeon head. The light from the lamp is transmitted through the fiber bundle to the output lens. (From www.luxtec.com.)

of the human body. The fiber optic bundles are used to bring the light down to the distal end and illuminate the target tissue, which the physician is examining through the imaging channel of the endoscopes.

In surgery, especially in microsurgery, dentistry, and so on fiber optic bundles are used to build a headlight, which creates bright illumination of the surgery area where the surgeon eyes are pointed. Here, the fiberoptic illumination allows mounting the output lens on the headband of the surgeon and leaving their hands free for the operation (see Fig. 5). Recent development in the fiber optics flat and flexible illumination panels (see below) allows bringing the visible light inside the deep cavities of the human body for illumination during surgery.

In ophthalmology, early application of fiber optic illumination has included its use as a light source in the indirect ophthalmoscope. The resulting small light spot at the distal end of the optical fibers allows for the use of variable apertures and provides sharply focused and uniformly illuminated circles of light upon the retina (21). Fiber optic illuminators are also used in conjunction with intraocular surgery. Thus a variety of miniature devices are available for the visualization of the interior of the eye to provide improved illumination in microsurgery of the retina and vitreous.

Currently, a number of companies are developing the "solid-state" light based on the light emitting diodes (LED). The advantages of this type of illumination is that it produces much less heat and emits in the visible spectral range thus making the illumination fiber bundle less useful for some a pplications. However, it will take another decade before this type of the light will become commercially viable.

In addition to transmitting the light through the fiber bundle to the target for illumination, the fiber optics often serves to shape the light in the form most advantages for the application. For example, in the form of the rigid or flexible flat panel that can be used during almost any deep cavity surgery. These fiberoptic panels are made of woven plastic optical fibers as shown in Fig. 6.

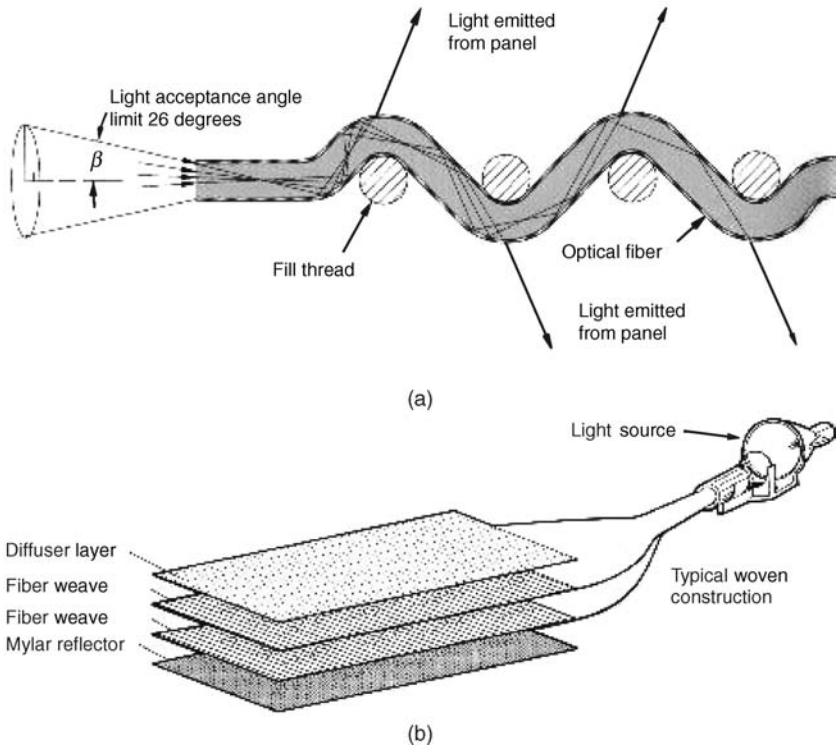


Figure 6. Principle of the fiber optics panels with the side emission. (From www.lumitex.com.) (a) All fiber optics illumination bundles, li ght enters the panel through each highly polished fiber end. Here, the computer controlled "macrobends" in the fibers cause the transmitted light to be emitted from the sides of the fibers through the cladding. Precisely engineered construction causes all light to be emitted uniformly along the length of the panel. (b) Layers of fiber optic weave are assembled together with double-sided adhesive into as many as eight layers. A mylar reflector is laminated to the back and a clear vinyl top layer is added for extra durability. For some applications (e.g., LCD backlighting), a semitransparent diffuser layer is placed between the top weave layer and the clear vinyl. The optical fibers extend from the panel in cable form and are bundled into a brass ferrule and highly polished. These ferrules are then connected to a remote light source.

Requirements for Illumination Fibers

A thin optical fiber is obtained by drawing in a furnace a glass rod inside glass tubing, which together form the high refractive index core and the low refractive index cladding, respectively. The core material usually used for the illumination fibers is silica, which, as was discussed earlier, is the best material for the visible spectral range. When drawn to a core diameter of 25 μm or less, the optical fiber is quite flexible, allowing for bending radii of <1 cm.

Light Sources Used with Illumination Fibers. The light sources for the fiberoptic illuminators are tungsten or quartz halogen projection lamps, mercury or xenon high pressure arc in quartz glass enclosures, and metal halide lamps. Color temperatures of tungsten sources vary between 2800 and 3500 K, which is a low color temperature with the illumination appearing yellow. Arc and halide lamps are used in the commercially available illuminators and provide light at the color temperature of $\sim$5400 K.

These light sources are a black body type radiator, thus they radiate into the sphere, that is, solid angle of 4π steradian and the amount of the output light they emit is almost proportional to the emitting body (filament, for the tungsten lamps, and arc, for other types). Thus they have the emitting body of $>1\,\text{mm}^2$ for the arc lamp and much greater for the tungsten and quartz halogen lamps.

Coupling Efficiency. The fact that the light source radiates in 4π steradian from the emitting body lead to requirements that the optical fibers used for the illumination have a high aperture and sufficient cross-section. To transmit the required illumination through the bundle and keep the fibers flexible, the conventional practice is to gather a lot of thin flexible fibers into a bundle, with the ends bound and polished. This bundle is almost as flexible as the single fiber. The fibers in the ends of the illumination fiber bundles are arranged at random and these bundles are called incoherent . This organization of the fibers in a bundle, in addition to being inexpensive to assemble, is also sometimes useful to provide uniform illumination at the output end.

Various schemes have been employed to maximize the light intensity delivered from the source to the fiber or fiber bundle and to obtain uniform illumination at the distal end of the fiber. For example, a special reflecting mirror, in the form of an ellipsoid of revolution, has the light source and the fiber input at the focal points with the major axis of the ellipsoid at an angle to the receiving fiber. However, note that light lamp sources, as opposed to lasers, could not be focused or concentrated onto areas smaller than their emitting body without considerable loss of the luminous flux. This follows from the fundamental relationship in radiometry, which states that the radiance of an image cannot exceed that of an object for the case when both lie in a medium with the same index of refraction (22). Consequently, the optimum illumination from an optical fiber is obtained when the image of the source completely fills the entry face of the fiber and the cone of light forming this image is equal to the numerical aperture of the fiber. In some cases, when the light source consists of a filament, the use of reflectors increases the effective area of the source by redirecting light rays through the voids in the filament.

Removal of Heat, UV, and IR Radiation. All lamps produce heat directly or by absorption of light in the visible to IR spectral ranges. Also, the UV radiation portion of the spectrum may be hazardous to the eyes and illuminated tissue. The high intensity light sources needed to deliver adequate illumination to and through the optical fiber cause the latter to become very hot. In the case of fiber bundles, the heat generated at the proximal end of the fibers requires that they not be bonded with epoxy, but instead be fused together. In some cases this may also apply to the distal end if enough heat is conducted through the fiber or is generated by absorption of light in the fiber. Of course, this is highly undesirable and should be avoided. After all, one of the main objectives of the use of illumination fibers is to keep the heat of the light source away from the area being illuminated. Indeed, the early applications of optical fibers were referred to as "cold-light" illumination. Special provisions must, therefore, be made to dissipate and divert large amounts of heat. Most light sources utilize dichroic reflectors and/or heat filters to block the heat from reaching the fiber.

Transmission. The light propagating through the optical fiber is attenuated by bulk absorption and scattering losses. These losses vary with the wavelength of light and are usually higher at short wavelengths, that is, the blue end of the visible spectrum of 400–700 nm. In most illumination applications, the typical length of the fibers does not exceed 2 m and thus the attenuation of light due to absorption is of no consequence. Antireflection coating can reduce Fresnel losses due to reflection at the end faces of the fiber.

The faithful rendition of color in viewing depends on the spectral content of the light illumination. This is of particular significance in medicine where the diagnosis of disease is often determined by the appearance and color of organs and tissue. Hence, optical fibers made of plastic materials, for example, polystyrene for the core and Lucite, polymethylene methacrylate (PMMA), for the cladding, despite their flexibility and low cost, do not find extensive use as illumination fibers.

DIAGNOSTIC APPLICATIONS

Introduction

One of the earliest applications of optical fiber in medicine were in imaging. The bundle of the optical fibers has been used to transmit image from the distal to the proximal end, where the physician could see the image of the target tissue in real time. The device using this technique is known as an endoscope. The endoscopes are used widely in current medical practice for imaging of lumens in the human body.

In the past 20 years, many new diagnostic applications of fiber optics have appeared as a result of the developments in biomedical optics (23). Most all of them utilize a single or a few fibers. Some of them, for example OCT (24)

and confocal imaging (CI) (25) use scanning to produce the images of the lateral or axial cross-sections of the body. Others are utilizing the spectroscopic differentiation of the tissue, using natural (23) or man-made markers (26). Among these techniques, fluorescence (23), reflectance (23), light scattering, and Raman spectroscopic methods (27) are most promising and thus most developed.

Another new area for the application of fiber optics in medicine is the fiberoptic biosensor (FOBS). This is a sensor consisting of optical fiber with a light source and detector and an integrated biological component that provides the selective reaction to the biochemical conditions of the tissue and body fluids. These sensors have been applied for glucose and blood gas measurements, catheter-based oximetry, billirubin, and so on. The detailed discussion about the principle and applications of these sensors is in "Optical Sensors" article of this encyclopedia.

Imaging

Endoscopy. The word endoscopy derives from two Greek words meaning "inside" and "viewing". Its use is limited to applications in medicine and is concerned with visualization of organs and tissue by passing the instrument through natural openings and cavities in the human body or through the skin, that is percutaneously. The endoscope has become an important and versatile tool in medicine. It provides a greater flexibility than it is possible with instruments that consist of a train of optical lenses, and transmits illumination to the distal end of the probe. The greater flexibility of the flexible endoscope enables the visualization around corners and the eliminati on of "blind areas" obtained with the rigid instrument. It should be noted that optical fibers are used to deliver illumination light in rigid endoscopes that in some cases may employ lenses for image transmission.

Endoscopes, which utilize optical fibers to transmit the image from the distal to the proximal end, are often called fiberscopes to differentiate them from the video or electronic endoscopes where a semiconductor imager is placed at the distal end and the image is transmitted electronically. Progress in the production of the relatively inexpensive high quality miniature semiconductor imagers based on Charge Coupled Device (CCD) technology and Complementary Metal Oxide Semiconductor (CMOS) led to the development of the high quality electronic (video) endoscopes. Currently, most of the endoscope vendors produce such endoscopes. It appears that these video endoscopes produce higher quality images with considerably higher magnification (28) and are replacing the fiberscopes where it is prac tical (29–31). However, a large number of fiberscopes are still used in the clinics and new fiberscopes are being sold (e.g., see www.olympusamerica.com). Moreover, in the areas that require thin and ultrathin endoscopes of $<$2–4 mm (32), fiberscopes are still the only practical solutions. The general discussion on endoscopy, its features, and applications is presented in the Endoscopy article. This article will primarily discuss fiberscopes.

In addition to the imaging and illumination channels, a typical endoscope includes channels to accommodate tools for biopsy to aspirate liquids from the region being inspected, and to inflate the cavity or to inject clear fluids to allow for better visualization. The overall dimensions of such instruments vary between 5 and 16 mm in diameter, the thinner versions being smaller and more versatile than the corresponding rigid systems. The fiberscope can be made as long as nece ssary, since the light losses in most fibers made of glass cores and cladding are tolerable over distances of up to several meters. These images can be recorded using film, analog and digital still, and video cameras.

Most of the fiberscopes use similar optical structures and vary mainly in length, total diameter, maneuverability, and accessories, for example, biopsy forceps. The diameter of the individual glass fibers in the image-conveying aligned bundle are made as small as possible limited only by the wavelength of the light to be transmitted. In practical applications, the diameter is ranging from 2 to 15 μm. Moreover, if the fibers are densely packed, cross-talk problems may arise due to the evanescent electromagnetic field (light waves) in each individual fiber (33).

A variety of fiberscopes have been developed, each with features that are best suited for specific applications.

Transmission of Images Through Optical Fibers. As shown in the section General Principles of Fiber Optics, an optical fiber cannot usually transmit images. However, a flexible bundle of thin optical fibers (obviously, with silica core) can be constructed in a manner that does allow for the transmission of images. If the individual fibers in the bundle are aligned with respect to each other, each optical fiber can transmit the intensity and color of one object point. This type of fiber bundles is usually called a "coherent" or "aligned" bundle. The resulting array of aligned fibers then conveys a halftone image of the viewed object, which is in contact with the entrance face of the fiber array. To obtain the image of objects that are at a distance from and larger than the imaging bundle, or imaging guide, it is necessary to use a distal lens or lens system that images the distal object onto the entrance face of the aligned fiberoptic bundle. The halftone screen-like image formed on the proximal or exit face of a bundle of aligned fibers can be viewed through magnifying systems or on the video monitor if this exit face is projected onto the video camera.

Fabrication of Flexible Imaging Bundles. The fabrication of flexible imaging bundles involves winding of the optical fibers on a highly polished and uniform cylinder or drum, with the circumference of the latter determining the length of the imaging structure. The aligned fibers can be wound directly from the fiber-drawing furnace or a separate spool of individual fibers. When the entire bundle is obtained in a single winding process, similar to a coil, an overhang of fibers wound on the outer layers develops after the structure is removed from the winding cylinder. Such overhang can be eliminated by winding single or a small number of layers and cutting them into strips to form the aligned fiber bundle. This process, although more laborious, usually renders better uniformity in the imaging bundles. Some users find the evenness of the fiber arrangement distracting and prefer the irregular structure. This may be compared

with the viewing of a television image wherein the horizontal line scan has accentuated by imperfect interlacing. In either method, the diameter of the fibers is on the order of 10 μm, which is thinner than hair or about the size of a strand in a cottonball, and must therefore be soaked in binding epoxy during the winding process. When the completed imaging bundle is cut and remove d from the drum, its ends are bound to secure the alignment and the epoxy is removed along the structure to let the individual fibers remain loose, and thus flexible. A flexible imaging bundle can also be obtained by first drawing a rigid preform made up of individual optical fibers, each with a double coating, where the outer cladding is chosen to be selectively etched afterwards. Such a structure has nearly perfect alignment, sin ce the preform can be constructed of fibers thick enough to allow parallel arrangement and the individual fibers in the drawn bundle are fused together without any voids be tween them. Moreover, such fabrication technique, as compared with the winding method, is far less expensive since it permits batch processing.

The size of the imaging bundle varies from 1 to 2 mm in diameter for miniature endoscopes, for example, angioscopes or pediatric bronchoscopes, where the fiberscopes are currently primarily utilized, to 6 mm in diameter for large colonoscopes, and lengths exceeding 2 m. Large imaging bundles may consist of as many as 100,000 individual fibers (28), each fiber providing one point of resolution or pixel. For such large numbers of fibers, it is often necessary to fabricate the final device by draw ing first fiber bundles containing a smaller number of fibers and then joining them together to form a larger structure.

Characterization of Flexible Imaging Fiber Bundles. The resolution attainable with a perfectly aligned fiberoptic imaging structure is determined by the fiber diameter. For a hexagonal configuration, the resolution in optical line pairs per millimeter is, at best, equal to, and usually slightly $<1/2d$, where d is the fiber diameter in millimeters. Figure 7 shows a cross-section of an imaging bundle of fibers aligned in a closely packed hexagonal pattern, i.e., each fiber has six closest neighbors. In Fig. 7a are shown two opaque stripes separated by a distance equal to their width, which are placed at the distal end of an entrance face of the imaging bundle. When this end is illuminated, the corresponding fibers at the output or proximal exit end of the bundle will be either partially or totally darkened, depending on the relative location of the opaque stripes with respect to the fiber pattern. In Fig. 7b, the partially illuminated fibers indicate that the smallest resolvable separation of the two stripes is equal to twice the fiber diameter. In practice, the packing of the fiber bundle is often not perfect, leaving spaces between the fibers; thus, the line resolution is further reduced and is usually $<1/2d$. For a typical imaging bundle in endoscopy, the diameter of the individual fiber is $\sim$10 μm and, therefore, the resolution is better than 50 line-pairs·mm^{-1}. This resolution is considerably poorer than in most lens systems.

The line resolution for the hexagonally aligned imaging bundle does depend on the orientation of the stripes (Fig. 7). Thus, for an orientation different from that shown in Fig. 7, the line resolution may be somewhat different from that

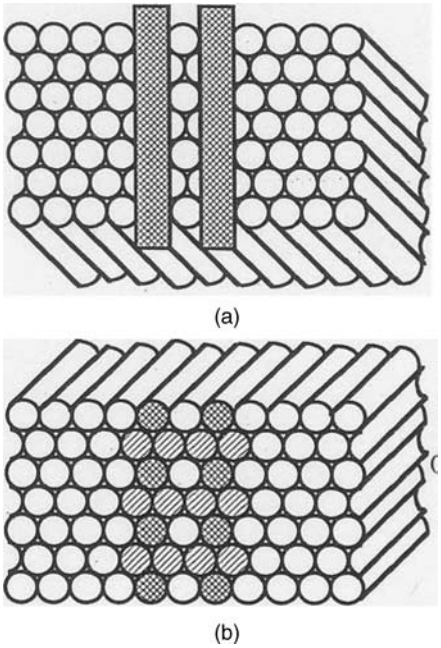


Figure 7. Image transmission of two opaque stripes from (a) distal end to (b) proximal face of an imaging fiber optics bundle.

obtained. The spatial variance of the image transfer properties of imaging bundle has led to the use of averaging techniques in the evaluation of their limits of image resolution (34). The modulation-transfer function (MTF) of an imaging bundle is the contrast response at its exit face to a sinusoidal test pattern of varying spatial periodicity imaged onto the input or entrance face. The fabrication of an imaging bundle may result in the misalignment or deviation from perfect alignment of the adjacent fibers. A method of evaluation of the image-conveying properties of practical imaging bundles has been developed (35), which takes into account the fact that the arrangement of the fibers may differ at the input and output faces of the imaging bundle. It also uses a statistical approach in determining an average modulation-transfer function. The aligned fiber bundle exhibits a mosaic pattern that represents the boundaries of the individual fibers and that appears superimposed on the viewed image. Hence, the viewer sees the image as if through a screen or mesh. Any broken fiber in the imaging bundle appears as a dark spot. In some applications, these two features can be very annoying and distracting, although most medical practitioners have become accustomed to these peculiarities and have learned to discount them. In some sense, it is comparable to viewing through a window with a wire screen, which in most cases is unnoticed.

Rigid Imaging Bundles. Fiberoptics bundles composed of individual optical fibers far smaller than those described earlier can be obtained by drawing the entire assembly in the furnace in a manner that preserves their alignment. This method is similar to that employed in the preparation of flexible imaging structures by etching a double-clad rigid imaging bundle. However, since the fibers become fused together during the drawing process, the structure

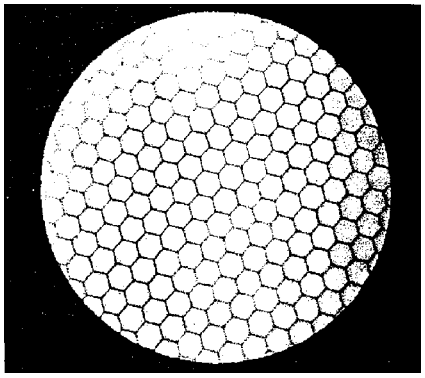


Figure 8. Cross-section of the imaging bundle.

remains a rigid solid-glass rod. A segment of a cross-section of such an imaging bundle, which is 0.5 mm in diameter and contains 11,000 fibers, each 4.2 μm in diameter, is shown in Fig. 8. The honeycomb pattern represents the boundaries of the individual fibers and is superimposed on the image, similar to the flexible fiberoptic imaging bundle.

Fiber-Based Confocal Microscopy. Confocal imaging (CI) is based on illuminating a single point on the sample and collecting scattered light from the same single point of the sample (Fig. 9a). The illumination point on the sample is the image of the illumination pinhole and is imaged into the detector pinhole, thus making both pinholes and the illuminated spot on the sample optically conjugated. In this way, stray light from points outside the collection volume is effectively filtered, improving contrast, particularly for scattering media, for examples, biological tissues. Scanning the position of this point on the sample and acquiring the signal from each position creates a lateral image of the sample. For axial cross-sectional imaging, the focal point on the sample is axially moved while acquiring the signal from each position. Combining both lateral and axial scans provides a perpendicular cross-sectional image of the tissue. The nature of the signal from the illuminated point depends on the specific embodiment. Either backscattered or fluorescent signals are most often used for CI; however, other signals have been collected and proven clinically useful.

Both pinholes contribute to depth sectioning ability. The source pinhole causes the illuminating irradiance to be strongly peaked at the image formed by the objective and to fall off rapidly both transversely and along the line of sight (depth). There is, therefore, very little light available to illuminate out-of-focus volumes. This implies that these volumes will not significantly mask the image of the more brightly illuminated focal region. The detector pinhole acts similarly. It is most efficient in gathering the radiation from the volume at the focal point, and its efficiency falls off rapidly with distance, both transversely and in depth. Together they assure that the totality of the out-of-focus signal is strongly rejected.

Confocal imaging provides enhanced lateral and axial resolutions and improved rejection of light from the out-of–focus tissue. As a result, the confocal imaging (CI) can achieve a resolution sufficient for imaging cells to depths of several hundreds of microns. At these depths, CI has been especially helpful in imaging the epithelium of many organs, including internal organs via endoscopic access. Images displayed by the CI system are similar to images of histopathology slides under high resolution conventional microscope, thus, are familiar to the physicians.

The fiberoptical confocal microscope has been introduced in late 1980s and early 1990s (see, e.g., Ref. 37). In this kind of the CI microscope, the light source is not a point source, but the tip of an optical fiber, and the signal is collected by another optical fiber that delivers the signal to a detector. This makes the CI system compact and thus convenient in numerous biomedical applications.

There are a variety of optical systems and scanning arrangements, which allows for optimization of the CI device for the required image and particular clinical application (25). For example, the scanning can be or ganized by moving the fiber tips, especially, when the common fiber is used for the illumination and detection as shown in Fig. 10b, or by moving the objective lens or by placing in

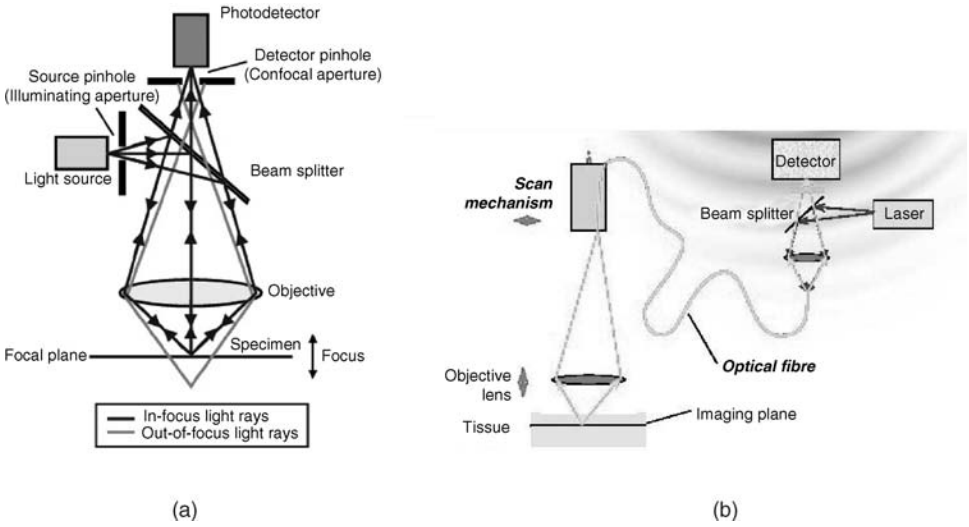


Figure 9. Confocal microscopy. (a) Principle of confocal arrangement. (From MPE Tutorial, Coherent Incorporated © 2000.) (b) Possible implementation of fiber-based confocal microscope (36).

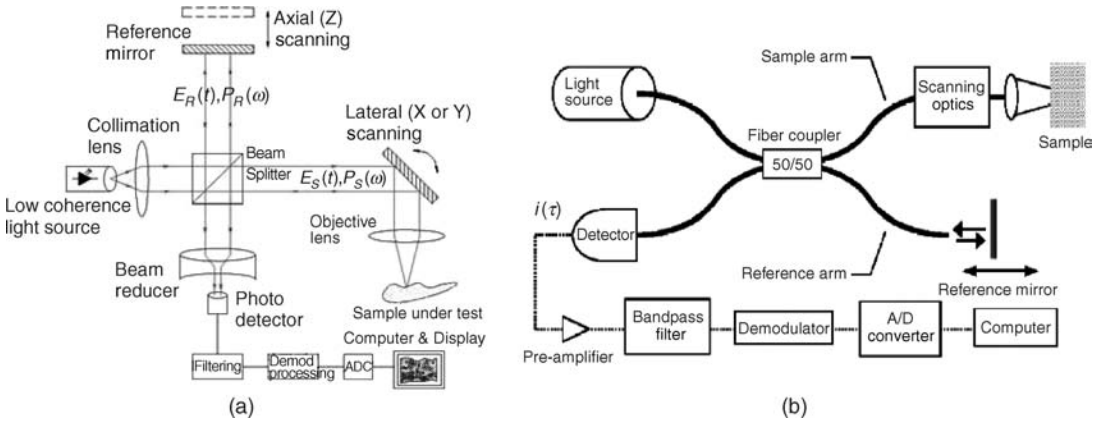


Figure 10. Optical coherence tomography (OCT). (a) Principle of OCT (From Wikipedia.org). (b) Possible implementation of fiber-based OCT (38).

the pinhole plane an output face of coherent fiber bundle and scanning an mage of a pinhole at the entrance end of it.

The CI system can provide a cellular level resolution: A lateral integrated image of tissue (similar to C-scan imaging), a lateral cross-sectional image at the desirable depth of tissue, a perpendicular-to-the-surface cross-sectional image of the tissue (similar to B-scan imaging), and a combination of the above images, thus a three-dimensional (3D) image of tissue. In each case, the CI system could be optimized to achieve the best overall performance while utilizing the same basic platform. The application of the CI imaging has been successfully demonstrated in diagnostics of intraepithelial neoplasia and cancer of the colon (39). It is reasonable to envision that the CI-based endoscope will be used for the screening of Barrett's esophagus and cancer and other areas of the upper GI tract. It appears that there is potential for application of the CI device combined with one of the scattering-based spectroscopies for the vulnerable plaque screening and triage.

The optical fibers used for the CI application are usually a single mode type because of the requirement for the small core diameter of the fiber tip as discussed above.

Optical Coherence Tomography. Optical coherence tomography is a relatively new technology and has been developed and investigated for the past 15 years (38). It uses the wave property of light called coherence to perform ranging and cross-sectional imaging. In OCT systems, a light beam from a light source is split into a reference light beam and a sample light beam (see Fig. 10a). The sample light beam is direct ed onto a point on the sample and the light scattered from the sample is combined w ith the reference light beam. The combined reference and sample light beams interfere if the difference of their optical paths is less than the coherence length. The reference and the collected sample beam are mixed in a photodetector, which detects the interference signal. The light source used in OCT has a low coherence length so that only the light s cattered from the sample within the close proximity around a certain depth will satisfy this condition. The strength of the interference signal corresponds to the scattering around this depth. To acquire the signal from another depth in the sample the optical pa th of one of the beams is changed so that the same condition is now satisfied by the li ght scattered from another depth of the sample. The component of the OCT system providing this change is called an optical delay line. By sequentially changing the optical path of one of the beams and processing the photodetector output, a cross sectional image of the sample is generated. By laterally moving the sample beam along the sample provides a perpendicular cross-sectional image of the sample. The OCT image is similar to high frequency ultrasound B-scan images.

Usually a moving mirror in the optical path of one of the beams performs the continuous scan of the optical path. The shortest coherence length of available light sources allows the OCT systems to achieve a depth resolution higher than in high frequency ultrasound imagers, but lower than in confocal imaging. As a direct correlate, the depth penetration of OCT systems is lower than for the high frequency ultrasound imagers and higher than the confocal imaging. These parameters make OCT a useful technology in the biological and medical examinations and procedures that require good resolutions to 2 mm depths.

The OTC systems utilizing the fiberoptic components are most often used (Fig. 10b). These systems are compact, portable, and modular in design. The sample arm of the OCT can contain a variety of beam-delivery options including fiberoptic radial- and linear-scanning catheter-probes for clinical endoscopic imaging. The aiming beam is used so that the clinicians could see the location on the tissue where the OCT catheter is acquiring image. The optical fiber based OCT systems require using single mode fibers in both reference and sample arms to keep the coherence of the light guided by the fiber. In some cases, when the OCT system is using the polarization properties of the light, it must utilize the polarization-maintaining fibers.

There is a variety of OCT optical systems and scanning arrangements, which provide room for optimization of the OCT device for the specific clinical application. Recently, several system modifications of the basic OCT have been reported; for example, Fourier transform OCT (40,41), spectroscopic OCT (42), and polarization OCT (43). Some of them appear to promise practical OCT systems with higher data acquisition rates, higher resolution (comparable with

the resolution of CI), better noise immunity, or capabilities to acquire additional information on tissue.

In the endoscopic applications, the sample arm fiber is designed as a probe that goes through the endoscope to the target tissue. There are a number of OCT probes developed and suggested designs (44–54) that provide room for optimization of the OCT device for the specific clinical application. A number of successful clinical studies have been carried out demonstrating the clinical applicability of the endoscopic fiber OCT technique for clinical imaging, for example, for imaging of Barrett's esophagus (54) and esophageal cancer (55), bile duct (56), and colon cancer (55).

Spectroscopy

Reflectance and Fluorescence Spectroscopy. Diffuse reflectance spectroscopy is one of the simplest spectroscopic techniques for studying biological tissue. Light delivered to the tissue surface undergoes multiple elastic scattering and absorption, and part of it returns as diffuse reflectance carrying quantitative information about tissue structure and composition.

This technique can serve as a valuable supplement to standard histological techniques. Histology entails the removal, fixation, sectioning, staining, and visual examination of a tissue sample under the microscope. Tissue removal is subject to sampling errors, particularly when the lesions are not visible to the eye. Also, the multiple-stage sample preparation process is time-consuming, labor intensive, and can introduce artifacts that are due to cutting, freezing, and staining of the tissue. Most importantly, the result is largely qualitative in nature, even though quantitative information is available through techniques, for example, morphometry and DNA multiploidy analysis.

Spectroscopy, in contrast, can provide information in real time, is not greatly affected by artifacts or sampling errors, and can provide quantitative information that is largely free of subjective interpretation. Because it does not require tissue removal, it can be conveniently used to examine extended tissue areas.

Usually, light is delivered and collected using an optical fiber probe that can be advanced through the accessory channel of the endoscope and brought into contact with the tissue. The probe can consist of several delivery and collection fibers. Probably the simplest geometry is a central optical fiber for light delivery and six fibers for light collection arranged in a circle around the central fiber. In Zonios et al. (57), all fibers had a 200 μm core diameter and a NA of 0.22, and were packed tightly with no gap between them. The probe tip was fitted with a quartz shield ~1.5 mm in length and in diameter, which provided a fixed delivery and collection geometry with uniform circular delivery and collection spots in the form of overlapping cones. The tip was beveled at an angle of 17° to eliminate unwanted specular reflections from the shield–tissue interface (see Fig. 11).

To extract quantitative properties of tissue collected with the above probe the model of light transport in tissue should take into account probe geometry. To find the total light collected by the probe, diffuse reflectance from a point source must be integrated over the spatial extent of the light delivery and collection areas, characterized by

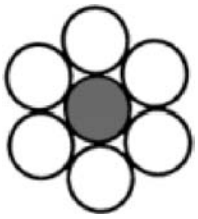


Figure 11. Configuration of fibers in a typical reflectance and fluorescence optical probe. Probe contains six fibers for light collection arranged in a circle around the central fiber.

radii r_d and r_c, respectively. Assuming the incident light intensity to be uniform over the entire delivery area, the diffuse reflectance $R_p(\lambda)$ collected by the probe is given by (57).

$$R_p(\lambda) = \frac{1}{{r_d}^2}\int_0^{r_c} r dr \int_0^{2\pi} d\phi \int_0^{r_d} R(\lambda, |\mathrm{r} - \mathrm{r}'|) r' dr'$$

where r_d and r_c are radii of the delivery and collection spots of the fiber optics diffuse reflectance probe and $R(\lambda,|\mathbf{r}-\mathbf{r}'|)$ is diffuse reflectance predicted by the physical model, which depends on tissue morphological and biochemical composition.

A similar probe (Fig. 11) can be used for fluorescence spectroscopy measurements. For fluorescence measurements, it is especially important that delivery and collection signals are delivered over the separate optical fibers. This is because intense illumination light can easily induce certain amount of fluorescence in the delivery fiber. This fluorescence of the fiber is likely to be weak, however, tissue fluorescence is also very weak. Thus if the delivery and collecti on fibers coincide, the fluorescence from the probe itself would significantly perturb the tissue fluorescence observed by the instrument. Hence, fluorescence fiber probes should always have separate delivery and collection fibers.

Light Scattering Spectroscopy. In addition to the multiply scattered light described by the diffuse reflectance, there is a single scattering component of the returned light that contains information about the structure of the uppermost epithelial cells (58). It has been shown that light scattering spectroscopy (LSS) (Fig. 12) enables

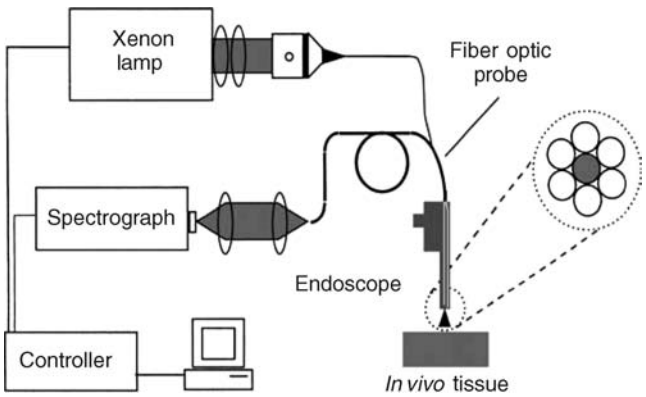


Figure 12. Schematic diagram of the LSS system (59).

quantitative characterization of some of the most important changes in tissues associated with precancerous and early cancerous transformations, namely, enlargement and crowding of epithelial cell nuclei (58,60). Typical nondysplastic epithelial cell nuclei range in size from 4 to 10 μm. In contrast, dysplastic and malignant cell nuclei can be as large as 20 μm. Single scattering events from such particles, which are large compared to the wavelength of visible light (0.5–1 μm), can be described by the Mie theory. This theory predicts that the scattered light undergoes small but significant spectral variations. In particular, the spectrum of scattered light contains a component that oscillates as a function of wavelength. The frequency of these oscillations is proportional to the particle size. Typically, normal nuclei undergo one such os cillation cycle as the wavelength varies from blue to red, whereas dysplastic/malignant nuclei exhibit up to two such oscillatory cycles. Such spectral features were observed in the white light directly backscattered from the uppermost epithelial cell nuclei in human mucosae (60,61).

When the epithelial nuclei are distributed in size, the resulting signal is a superposition of these single frequency oscillations, with amplitudes proportional to the number of particles of each size. Thus, the nuclear size distribution can be obtained from the amplitude of the inverse Fourier transform of the oscillatory component of light scattered from the nuclei. Once the nuclear size distribution is known, quantitative measures of nuclear enlargement (shift of the distribution toward larger sizes) and crowding (increase in area under the distribution) can be obtained. This information quantifies the key features used by pathologists in the histological diagnosis of dysplasia and Carcinoma *in situ* (CIS), and can be important in assessing premalignant and noninvasive malignant changes in biological tissue *in situ*.

However, single scattering events cannot be directly observed in tissue *in vivo*. Only a small portion of the light incident on the tissue is directly backscattered. The rest enters the tissue and undergoes multiple scattering from a variety of tissue constituents, where it becomes randomized in direction, producing a large background of diffusely scattered light. Light returned after a single scattering event must be distinguished from this diffuse background. This requires special techniques because the diffusive background itself exhibits prominent spectral features dominated by the characteristic absorption bands of hemoglobin and scattering of collagen fibers (there is abundance of them in the connective tissue laying below the epithelium). The first technique of diffusive background removal uses a standard reflectance probe described in the section References and Fluorescence Spectroscopy. This technique is based on observation that the diffuse background is typically responsible for >95–98% of the total reflectance signal. Therefore, the diffusive background is responsible for the coarse features of the reflectance spectra. The diffusion approximation-based model may account for th is component by fitting to its coarse features. After the model fit is subtracted, the single backscattering component becomes apparent and can be further analyzed to obtain nuclear size distribution (58).

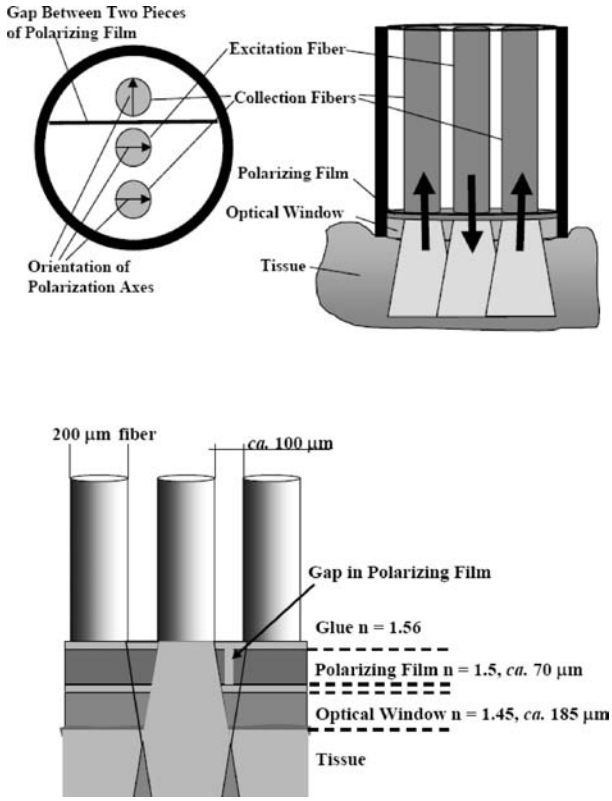


Figure 13. Design of the polarized fiber probe (62).

Another technique would use a special polarized probe. One possible implementation of the polarized probe is described by Utzinger and Richards-Kortum (62) (see Fig. 13). Recently, similar polarized fiber probe was developed by Optimum Technologies, Inc. The working principle of this probe is based on the fact that initially polarized light loses its polarization when traversing a turbid medium such as biological tissue. Consider a mucosal tissue illuminated by linearly polarized light. A small portion of the incident light will be backscattered by the epithelial cell nuclei. The rest of the signal diffuses into the underlying tissue and is depolarized by multiple scattering. In contrast, the polarization of the light scattered backward after a single scattering event is preserved. Thus, by subtracting the unpolarized component of the reflected light, the contribution due to the backscattering from epithelial cell nuclei can be readily distinguished. The residual spectrum can then be analyzed to extract the size distribution of the nuclei, their population density, and their refractive index.

Raman Spectroscopy. Raman spectroscopy is a very powerful technique, which should be capable of performing detailed analysis of tissue biochemical composition. However, in biological tissues the intensity of the Raman signal is only 10^{-9}–10^{-11} of the intensity of the incident light. What make it even worse is the fact that fluorescence signal excited in tissue by the same incident light is much higher, $\sim 10^{-3}$–10^{-5} of the incident light. And if the signal is detected by the regular optical fiber probe it will also overlap with the fluorescence signal originated in the fiber

Figure 14. Design of the fiber tip of a typical fiber optic probe for Raman spectroscopy (62).

itself. Hence, development of a reliable biomedical Raman fiber probe is still a challenge.

In Fig. 14 one can see the design of a Raman fiber optic probe developed by Visionex Inc. It consists of a central delivery fiber and seven collection fibers. A bandpass filter is located in front of the delivery fiber and a longpass filter in front of the collection fibers. Those filters ensure that no laser light and fluorescence light originated in the delivery fiber can enter the collection fiber.

THERAPEUTIC APPLICATIONS

Introduction

In therapeutic application, fibers are used primarily as a part of light delivery system. Utilization of fibers allows for flexible and convenient methods of delivering light from bulky high energy light sources to the diseased location. Often, using the fiber optics delivery system makes otherwise inaccessible locations accessible. Variety of applications, light sources and energy requirements require different fibers with different features.

Fiber Surgical Applications

The monochromatic nature of laser beam en ergy, temporal coherence, and the ability to focus it onto a very small spot because of its spatial coherence, allow for efficient surgical methods, for example, tissue cutting or ablation, as well for the transmission of large amounts of power through a single optical fiber. This localized energy can be used to cauterize tissue by evaporation while reducing bleeding by coagulation. The use of lasers in surgical and other therapeutic applications is most effective when employed in conjunction with the flexible delivery systems and, in particular, with laser-beam delivery which is compatible with the endoscopy. Lasers utilized in medical applications range in wavelength from vacuum UV (excimer laser at 193 nm) to infrared (IR) (CO_2 laser at 10.6/11.1 μm).

Laser-beam-delivery systems depend on the wavelength laser energy. At the present time there is no single delivery system that can be used over the entire range of medical lasers. The primary design considerations in such systems are efficiency, maximum power delivered, preservation of beam quality, and mechanical properties (flexibility, degrees of freedom, range, size, and weight).

Low efficiency (output power/input power) results in losses in the delivery system, which, in turn, requires higher power and thus more expensive lasers. Moreover, the power lost in the delivery system is generally dissipated as heat, resulting in a temperature rise that causes damage to the device, leading to a further deterioration of efficiency or catastrophic failure, Hence, efficiency, together with heat dissipation, can be considered to be the limiting factors in maximum power delivery.

A well-designed laser oscillator emits a highly collimated beam of radiation, which can then be focused to a spot size of just a few wavelengths, yielding power densities not achievable with conventional light sources. A useful delivery system should, therefore, preserve this quality of the beam as much as possible; otherwise the main advantages of using a laser are lost.

The most desirable flexible system should be easy to handle mechanically, perform a large variety of tasks, and, of course, still satisfy the above properties. From a mechanical point of view, the ideal laser-beam guide would be an optical fiber of small cross section and long enough to reach any site, at any orientation, over a wide range of curvatures, through openings and inside complex structures.

The choice of fibers for such system is mainly determined by the wavelength of the laser (see Fig. 4) and discussion in the section Types of Fibers.

Most lasers operate in the range of 300–2500 nm, where silica fibers have the best overall properties and, thus, are commonly used.

Outside this range, there are few wavelengths used for laser surgery, in IR, 2.94 μm of the erbium:YAG laser, 5–6 μm of the CO laser, 10.6/11.1 μm of the CO_2 laser and in the UV, the 193 and 248 nm of the ArF and KF excimer lasers, respectively. In the infrared range, silver halide and sapphire are being used for the fiber core. Also, hollow fibers have become available for efficient guidance of infrared light with minimal losses (20,63). In the UV range, the solarization resistant silica/silica fibers are available (64) and hollow fibers have been reported recently (16,17). For pulsed lasers, utilization of fibers can be limited by the high peak powers, which could damage the fiber; for example, the peak power of 106 $kW \cdot cm^{-2}$ is a threshold for silica. In this case, the flexible arms with reflective surfaces have to be used.

As far as power-handling capability is concerned, the typical glass optical fiber is quite adequate in applications where the laser beam energy is delivered continuously or in relatively long pulses such that the peak power in the optical fiber does not exceed power densities of several megawatts per square millimeter. When the laser energy is delivered in very short pulses, however, even a moderate energy per pulse may result in unacceptable levels of-peak power. Such may be the case of Nd-YAG lasers, operating in a mode-locked or Q-switched configuration, which produces laser beam energy in the form of pulses of nanosecond duration or less. On the other hand, excimer lasers, which are attractive in a number of applications (4), generate energy in the UV range of the spectrum (200–400 nm) in very short pulses; they, therefore, require

solarization-resistant silica/silica fibers, which can transmit light energy at such short wavelengths and, at the same time, carry the high power densities.

For lasers in the IR region beyond 2500 nm, fibers made of materials other than silica are being used as shown in the section Types of Fibers.

Photodynamic Therapy

Photodynamic therapy utilizes the unique properties of a substance known as photosensitizer (65). When administered systemically it is retained selectively by cancerous tissue. The substance is photosensitive and produces two effects: When excited by light at a photosensitizer-specific wavelengt hit fluoresces. This effect is used in photodynamic diagnostics. When irradiated with light, the photosensitizer undergoes a photochemical reaction, which results in the formation of a singlet oxygen and the subsequent destruction of the cell (usually malignant cell) that retained the substance.

In photodynamic diagnostics, since the fluorescence efficiency of the photosensitizer is low, high intensity illumination at the photosensitizer-specific wavelength and high gain imaging systems are required to detect very small tumors. The excitation is in the spectral range, where silica fibers have excellent properties thus glass (silica) fibers are used for delivery of the excitation light.

In order to obtain an effective cure rate in photodynamic therapy, it is essential that the optical fibers, which deliver the light energy to the tumor site, provide uniform light distribution. Since an optical fiber, even if uniformly irradiated, yields a nonuniform output light distribution, it is necessary to employ special beam shapers at the exit face of the fiber (66). The specifics of the fiber used for the delivery is determined by the absorption wavelength of the selected photosensitizer, the laser power and mode of operation, and the site being treated. These are the same consideration as in the fiber delivery for the laser surgery as discussed in the section Surgical Application of Fibers.

It is noteworthy that the illumination in photodynamic therapy does not require that it be obtained with coherent light. The only reason that lasers are used is that unlike conventional light sources, spatially coherent radiation can be efficiently focused onto and transmitted through a small diameter fiber. Light emitting diodes are often used as a compromise between efficient but expensive lasers and inexpensive and inefficient conventional light sources.

The application of photodynamic therapy in oncology has been investigated over the past 25 years. It appears that this modality is being used more often now (67) for variety of cutaneous and subcutaneous tumors.

BIBLIOGRAPHY

Cited References

1. Webb MJ. Practical considerations when using fiber optics with spectrometers. Spectroscopy 1989;4:26.
2. Basic Principles of Fiber Optics, Corning Incorporated.
3. Verdaasdonk RM, van Swol CFP. Laser light delivery systems for medical applications. Phys Med Biol 1997;42:869–894.
4. Parrish JA, Deutsch TF. Laser photomedicine. IEEE J Quantum Electron, 1984; QE- 20:1386.
5. Stolen RH. Nonlinearity in fiber transmission. Proc IEEE- 68: 1980; 1232.
6. Bass M, Barrett HH. Avalanche breakdown and the probabilistic nature of laser induced damage. IEEE J. Quantum Electron 1972; QE- 8:338.
7. Chiao RY, Garmire E, Townes CH. Self-trapping of optical beams. Phys Rev Lett 1964;13:479.
8. Smith RG. Optical power handling capacity of low loss optical fibers as determined by stimulated Raman and Brillouin scattering. Appl Opt 1972;11:2489.
9. Allison W, Gillies GT, Magnuson DW, Pagano TS. Pulsed laser damage to optical fibers. Appl Opt 1985;4:3140.
10. Harrington JA, ed. Selected Papers on Infrared Fiber Optics (SPIE Milestone Series MS-9). Bellingham (WA): SPIE; 1990.
11. Merberg GN. Current status of infrared fiber optics for medical laser power delivery. Lasers Surg Med 1993;13:572–576.
12. Harrington JA. Laser power delivery in infrared fiber optics. Proc SPIE Eng Comput 1992;1649:14–22.
13. Barnes AE, May RG, Gollapudi S, Claus RO. Sapphire fibers: optical attenuation and splicing techniques. Appl Opt 1995;34:6855–6858.
14. Shenfeld O, Ophir E, Goldwasser B, Katzir A. Silver halide fiber optic radiometric temperature measurement and control of CO_2 laser-irradiated tissues and application to tissue welding. Lasers Surg Med 1994;14:323–328.
15. Abel T, Harrington JA, Foy PR. Optical properties of hollow calcium aluminate glass waveguides. Appl Opt 1994;33: 3919–3922.
16. Matsuuga Y, Yamamoyo T, Miyagi M. Delevirey of F2-excimer laser light by aluminum hollow fibers. Opt Exp 2000;6:257–261.
17. Matsuuga Y, Miyagi M. Hollow Optical Fibers for Ultraviolet Light. IEEE J Quantum Electron 2004; QE- 10:1430–1439.
18. Cossmann PH, et al. Plastic hollow waveguides: properties and possibilities as a flexible radiation delivery system for CO_2-laser radiation. Lasers Surg Med 1995;16:66–75.
19. Matsuura Y, Abel T, Harrington JA. Optical properties of small-bore hollow glass waveguides. Appl Opt 1995;34:6842–6847.
20. Hongo A, Koike T, Suzuki T. Infrared hollow fibers FOR medical applications. Hitachi Cable Rev 2004;23:1–5.
21. Havener WH. The fiber optics indirect ophthalmoscope. Eye Ear Nose Throat Mon 1970;49:26.
22. Bass M, editor in chief, Handbook of Optics. Vol 1 New York: McGraw-Hill; 1995. Chapt. 1.
23. Tuchin VV, ed., Handbook of Optical Biomedical Diagnostics. Bellingham (WA): SPIE Press; 2002.
24. Bouma BE, Tearney GJ. Handbook of Optical Coherence Tomography. New York: Marcel Dekker; 2001.
25. MacAulay C, Lane P, Richards-Kortum R, In vivo pathology: microendoscopic imaging modality. Gastroint Endoscopy Clin N Am 2004;14:595–620.
26. Wagnieres GS, Star WM, Wilson BC. In vivo fluorescence spectroscopy and imaging for oncology applications. Photochem Photobiol 1998;68:603–632.
27. Perelman LT, Modell MD, Vitkin E, Hanlon EB. Scattering spectroscopy: from elastic to inelastic. In: Tuchin VV, ed., Coherent Domain Optical Method: Biomedical Diagnostics, Environmental and Material Science. Vol. 1. Boston: Kluwer Academic; 2004. pp 355–396.
28. Niederer P, et al. Image quality of endoscope. Proc SPIE 2001;4158:1–9.
29. Nelson DB. High resolution and high magnification endoscopy. Gastrointes Endosc 2000;52:864–866.
30. Kourambas J, Preminger GM. Advances in camera, video, and imaging technologies in laparoscopy. Urolog. Clinics N. Am 2001;28:5–14.

31. Korman LY. Digital imaging in endoscopy. Gastrointest Endosc 1998;48:318-326.
32. Nelson DB. Ultrathin endoscopes esophagogastroduodenoscopy. Gastrointest Endosc 2000;51:786–789.
33. Lipson SG, Lipson HS, Tannhauser DS. Optical Physics. New York: Cambridge University Press; 1995.
34. Sawatari T, Kapany NS. Statistical evaluation of transfer properties in fiber assemblies. SPIE Proc 1970;21:23.
35. Marhie ME, Schacham SE, Epstein M. Misalignment of imaging multifibers. Appl Opt 1978;17:3503.
36. OptiScan Pty, Ltd. [Online], Investor Presentation, October 2003. Available at http://www.optiscan.com.
37. Gu M, Sheppard CJR, Gan X. Image formation in a fiber-optical confocal scanning microscope. J Opt Soc Am A 8(11): 1755 (November 1991).
38. Huang D, et al. Optical coherence tomography. Science 1991;254:1178–1181.
39. Kiesslich R, et al. Confocal laser endoscopy for diagnosing intraepithelial neoplasias and colorectal cancer in vivo. Gastroenterology 2004;127:706–713.
40. Morgner U, et al. Spectroscopic optical coherence tomography. Opt Lett 2000;25:111–113.
41. Wax A, Yang C, Izatt JA. Fourier-domain low-coherence interferometry for light-scattering spectroscopy. Opt Lett 2003;28:1230–1232.
42. Vakhtin AB, Peterson KA, Wood WR, Kane DJ. Differential spectral interferometry: an imaging technique for biomedical applications. Opt Lett 2003;28:1332–1334.
43. Jiao S, Yu W, Stoica G, Wang LV. Optical -fiber-based Mueller optical coherence tomography. Opt Lett 2003;28:1206–1208.
44. Tearney GJ, et al. Scanning single- mode fiber optic catheter-endoscope for optical coherence tomography. Opt Lett 1996;21:543–545.
45. Gelikonov FI, Gelikonov VM. Design of OCT Scanners. Bouma BE, Tearney GJ, editors. Handbook of Optical Coherence Tomography. New York: Marcel Dekker; 2001. pp 125–142.
46. Liu X, Cobb MJ, Chen Y, Li X. Miniature lateral priority scanning endoscope for real-time forward-imaging optical coherence tomography. OSA Biomed Top Meeting Tech Dig SE6 2004.
47. Zara JM, et al. Electrostatic micromachine scanning mirror for optical coherence tomography. Opt Lett 2003;28:628–630.
48. Zara JM, Smith SW. Optical scanner using a MEMS actuator. Sens Actuators A 2002;102:176–184.
49. Piyawattanametha W, et al. Two-dimensional endoscopic MEMS scanner for high resolution optical coherence tomography. Tech Digest Ser Conf Lasers Electro-Optics (CLEO) CWS 2 2004.
50. Pan Y, Xie H, Fedder GK. Endoscopic optical coherence tomography based on a microelectromechanical mirror. Opt Lett 2001;26:1966–1968.
51. Xie H, Pan Y, Fedder GK. Endoscopic optical coherence tomographic imaging with a CMOS-MEMS micromirror. Sens Actuators A 2003;103:237–241.
52. Pan Y, Fedder GK, Xie H. Endoscopic imaging system. U.S. Pat. Appl. US2003/0142934, 2003.
53. Tran PH, Mukai DS, Brenner M, Chen Z. In vivo endoscopic optical coherence tomography by use of a rotational microelectromechanical system probe. Opt Lett 2004;29:1236–1238.
54. Qi B, et al. Dynamic focus control in high-speed optical coherence tomography based on a microelectromechanical mirror. Opt Commun 2004;232:123–128.
55. Brand S, et al. Optical coherence tomography in the gastrointestinal tract. Endoscopy 2000;32:796–803.
56. Seitz U, et al. First in vivo optical coherence tomography in the human bile duct. Endoscopy 2001;33:1018–1021.
57. Zonios G, et al. Diffuse reflectance spectroscopy of human adenomatous colon polyps in vivo. Appl Opt 1999;38:6628–6637.
58. Perelman LT, et al. Observation of Periodic Fine Structure in Reflectance from Biological Tissue: A New Technique for Measuring Nuclear Size Distribution. Phys Rev Lett 1998;80:627–630.
59. Wallace MB, et al. Endoscopic Detection of Dysplasia in Patients with Barrett's Esophagus using Light Scattering Spectroscopy. Gastroenterology 2000;119:677–682.
60. Perelman LT, Backman V. Light scattering spectroscopy of epithelial tissues: principles and applications. In: Tuchin VV, editor. Handbook on Optical Biomedical Diagnostics. Bellingham: SPIE Press; 2002.
61. Backman V, et al. Diagnosing cancers using spectroscopy. Nature (London) 2000;405:35–36.
62. Utzinger U, Richards-Kortum RR. Fiber optic probes for biomedical optical spectroscopy. J Biomed Opt 2003;8:121–147.
63. Matsuura Y, Miyagi M. Er:YAG, CO, and CO^2 laser delivery by ZnS-coated Ag hollow waveguides. Appl Opt 1993;32:6598–6601.
64. Solarization Resistant Optical Fiber, SolarGuide 193 [online], Fiberguide Industries, Inc, Stirling, N.J. Available at www.fiberguide.com.
65. Dougherty TJ, et al. Photodynamic Therapy. J Nat Cancer Inst 1998;90:889–905.
66. Panjehpour M, Overholt DF, Haydek JM. Light sources and delivery devices for photodynamic therapy in the gastrointestinal tract. Gastrointest Endosc Clin N Am 2000;10:513–532.
67. Brown SB, Brown EA, Walker I. The present and future role of photodynamic therapy in cancer treatment. The Lancet 2004;5:497–508.

See also ENDOSCOPES; OPTICAL SENSORS.

FICK TECHNIQUE. See CARDIAC OUTPUT, FICK TECHNIQUE FOR.

FIELD-EFFECT TRANSISTORS, ION-SENSITIVE. See ION-SENSITIVE FIELD-EFFECT TRANSISTORS.

FITNESS TECHNOLOGY. See BIOMECHANICS OF EXERCISE FITNESS.

FIXATION OF ORTHOPEDIC PROSTHESES. See ORTHOPEDICS, PROSTHESIS FIXATION FOR.

FLAME ATOMIC EMISSON SPECTROMETRY AND ATOMIC ABSORPTION SPECTROMETRY

ANDREW W. LYON
MARTHA E. LYON
University of Calgary
Calgary, Canada

INTRODUCTION

The observation that atoms of each element can emit and absorb light at specific wavelengths is a fundamental property of matter that fostered the study of chemistry and development of structural models of atoms during the past century. The interaction of light and matter can be traced to use of the lens of Aristophanes $\sim$ 423 BC, and to studies of mirrors by Euclid (300 BC) and Hero (100 BC). Seneca (40 AD) observed the ability of prisms to scatter

Table 1. Selected Wavelengths and Excitation Energies for Alkali and Alkaline Earth Elements

Element	Wavelength, nm Light Emission	Excitation Energy, eV
Lithium	670.7	1.85
Sodium	589.0/589.6	2.1
Magnesium	285.2	4.34
Potassium	766.5/769.9	1.61
Calcium	422.7	2.93
Manganese	403.1/403.3/403.4	3.07
Rubidium	780.0/794.8/ 420.2/421.6	1.56/1.59/ 2.95/2.94
Strontium	460.7	2.69
Cesium	894.3/852.1/ 455.5/459.3	1.39/1.45/ 2.72/2.69
Barium	553.6	2.24

light and Ptolemy (100 AD) studied angles of incidence and refraction. Sir Isaac Newton (1642–1727) performed many experiments to separate light into its component spectrum described in the 1718 volume Opticks: or, a treatise of the reflections, refractions, inflections and colours of light (sic).

The scientific basis of flame emission spectrometry (also called flame photometry) arose from studies of the light emitting and absorbing behaviors of matter when introduced into a flame. Observations can be traced to Thomas Melvill (1752), who observed the change of color and intensity of light when different materials were introduced into flames (1). Wollaston (1802) separated candlelight by a prism and noted that it contained a bright yellow line (2). The term line spectra is used to describe spectra composed of discontinuous narrow bands of light wavelengths or lines of different colors. Herschel (1822) observed different spectral diagrams for different salts and Talbot suggested that the yellow light indicated the presence of soda (various forms of sodium carbonate) (3,4). In 1860, Kirchhoff and Bunsen reported the correlation of spectral lines with specific elements by introducing solutions into a Bunsen burner using a platinum wire or hydrogen spray (5). This body of work is Table 1 is depicted in Fig. 1 and summarized by Kirchhoff's laws of spectroscopy that describe emission and absorption of light by matter. Kirchhoff's laws are a hot solid, liquid, or gas, under high pressure, gives off a continuous spectrum; a hot gas, under low pressure, produces a bright line or emission line spectrum; a dark line or absorption line spectrum is seen when a source of continuous spectrum is viewed behind a cool gas at low pressure.

In 1930, Lundegardh developed the first practical method for quantitative spectrochemical analysis by showing that the intensity of light emitted at specific wavelengths represented the concentration of an element being introduced into a flame (6,7). Flame emission methods were rapidly developed to determine alkali and alkaline earth metals. With the introduction of interference filters by Leyton and Ivanov in the 1950s to select narrow bands of wavelengths of light, multichannel flame emission photometers were developed that allow quantification of several elements simultaneously (potassium, sodium, lithium) from a single flame (8,9).

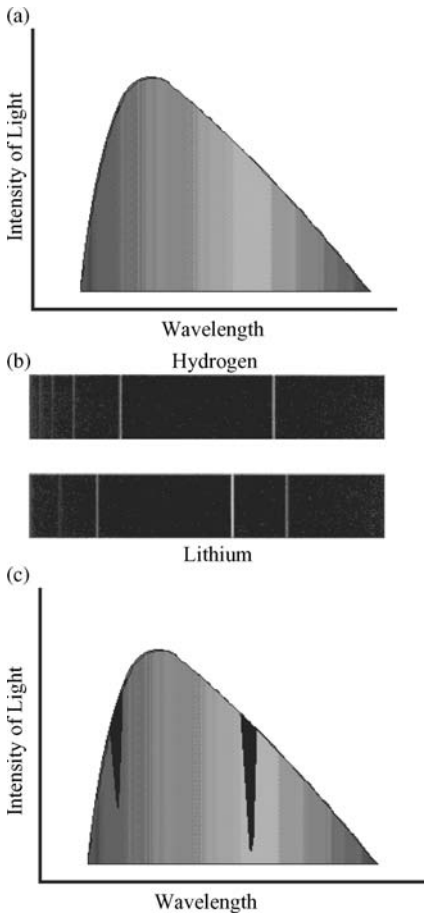


Figure 1. Emission and absorption spectra: (a) A continuous spectrum emitted from a hot solid, liquid, or gas under pressure. (b) A line spectrum emitted from a low pressure hot gas. (c) A dark line or absorption spectrum observed when a continuous spectrum is viewed behind a cool gas at low pressure.

Following the development of commercial flame emission photometers, clinical laboratories used the analyzers to measure sodium and potassium concentrations in body fluids from the mid-1950s until the mid-1970s, when potentiometric technologies largely replaced this technology. Flame emission photometers were initially retained in clinical laboratories to allow monitoring of serum lithium in patients treated with oral lithium salts as a therapy for manic depressive disorders. The rugged and reliable flame emission photometers are still used in specialized clinical laboratories to analyze fluid matrices that can render potentiometric analyses unstable (e.g., the analysis of stool electrolyte concentrations).

In addition to flame, other energy sources can be used to produce atomic emission spectra in the ultraviolet (UV) and visible regions including electric arc, electric spark, or inductively coupled plasma. The use of higher temperature energy sources to induce emission allows this technology to assess quantitative elemental composition of materials. These high energy instruments are used to support analysis of specimens for geology, manufacturing, and clinical toxicology.

Kirchhoff's third law describes the absorption of specific wavelengths of light by cool gases, the basis of atomic absorption spectrometry. In the 1950s, Walsh (10) and Alkemade and Milatz (11) independently reported the analytical potential of light absorption by atoms in a flame, extending the observation that Woodson (12) reported for determination of mercury in air. The temperature of the flame is sufficient to disperse atoms in a low density gaseous phase, enabling the atoms of individual elements to absorb light at specific wavelengths. An intermittent external light source is used to assess light absorption by atoms in the flame while the continuous light emission from the flame is measured and subtracted. Hollow cathode lamps were subsequently developed that enabled atomic absorption methods to achieve both greater sensitivity and selectivity than flame emission photometry and many more analytical applications. The development of electrothermal atomic absorption spectrometers and preanalytical derivitization of chemicals into compounds that minimize interference during analysis make this technology a valuable part of analytical laboratories. Clinical laboratories continue to use atomic absorption spectrometric methods to evaluate the concentration of lead, copper, zinc, and other trace metal elements in body fluids.

THEORY: FLAME ATOMIC EMISSION SPECTROMETRY

A theoretical basis for flame emission spectrometry can be described using a model of the atom, where the atom is composed of a nucleus surrounded by a cloud of electrons that fluctuate between different energy levels. The temperature of a solution of salt is rapidly elevated when the solution is introduced into a flame. The solvent evaporates and most metallic ions are reduced to the elemental neutral atom state by free electrons in the flame. A small proportion of monatomic ions are also formed. The temperature of the flame can confer sufficient quantum energy to excite electrons in a low energy level to a higher energy state, however, the higher energy state is inherently unstable. When the unstable high energy electron returns to the ground state, energy is released from the atom in the form of monochromatic light; a spectral line that is characteristic of that atom and energy transition. The analysis of the intensity and wavelengths of the line spectra emitted from materials in high temperature sources is the basis of flame atomic emission spectrometry, Fig. 2.

The intensity of light in the line spectra is related to the number of atoms with electrons in the higher energy level. The relationship between number of atoms in each energy state (N_0: ground state or N_1 elevated energy state) is dependent on the temperature: T, and the size of the energy interval between the ground state and elevated energy state: E_1–E_0. This relationship is depicted by the Boltzmann equation:

$$\frac{N_1}{N_0} = \frac{P_1}{P_0} e^{-(E_1 - E_0)/kT}$$

The important features of this equation are that the ratio of atoms in each energy state (N_1/N_0) is dependent on the magnitude of the quantity $-(E_1-E_0)/kT$. As temperature increases, the number of atoms in the high energy state increases in an exponential manner. The values P_1 and P_0 are the quantum mechanical properties called statistical weights and represent the number of ways in which an atom can be arranged in a given state.

The Boltzmann equation describes the proportion of atoms with electrons in an excited or high energy state at specific temperatures. As electrons decay from excited states to lower energy states, light energy is released in discrete energy intervals associated with the line spectrum for that element. The relationship between the light energy

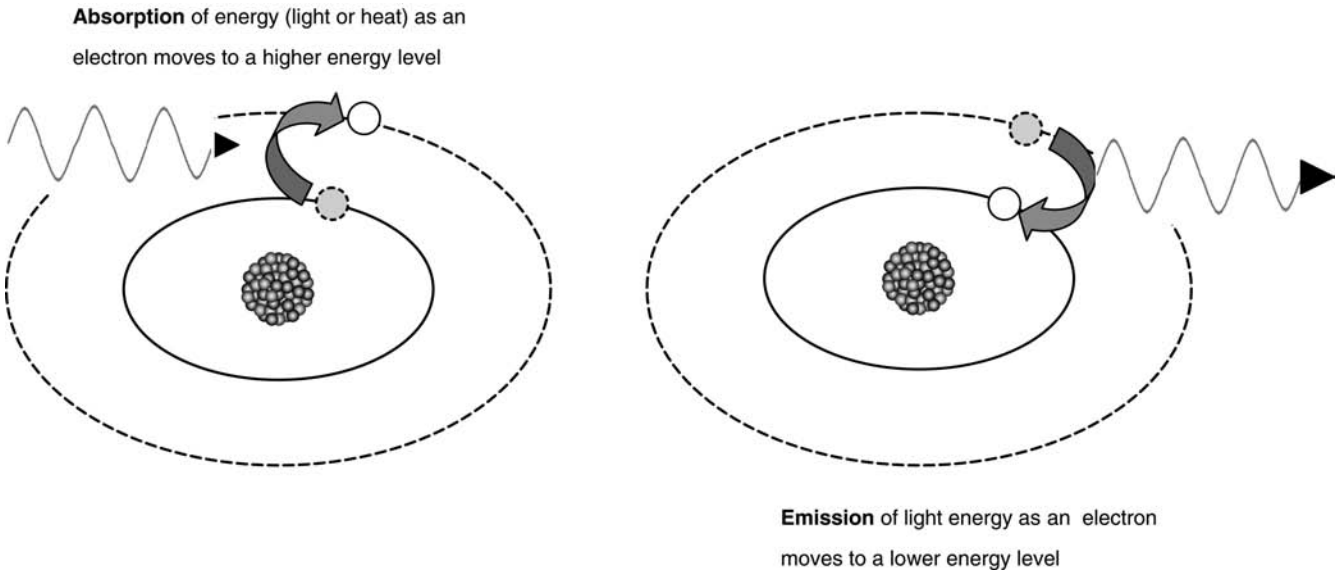


Figure 2. A model of the atom depicting (left panel) the absorption of energy associated with an electron moving to a higher energy state or (right panel) the emission of light energy associated with an electron moving to a lower energy state.

released and the wavelength or color of light emitted is described by Planck's law: E: Energy of electron transition (J), h: Planck's constant (J · s), c: speed of light (m/s), λ: wavelength of light (m^{-1}).

$$E = hc/\lambda$$

Atoms can have electrons at many different energy levels and this results in line spectra emissions at many different wavelengths. When a flame has limited energy and the atoms are at low density, only a small fraction of atoms are excited and only a simple line spectrum of low energy wavelengths is observed. Line spectra are characteristic of each element allowing analysts to measure emitted light to determine the chemical composition of solutions in a flame or the chemical composition of a star, based on starlight Table 1 (13). For example, the line spectrum of sodium is characterized by an intense doublet of yellow lines at 589 and 589.6 nm. The intensity of light at 588–590 nm can be measured in a flame photometer to measure the concentration of sodium in a fluid. Other alkali metals have low excitation energy that results in emission of light at longer wavelengths. Lithium and rubidium produce a red color and potassium a red-violet color when introduced into a flame. Higher energy associated with higher temperatures is required for many other elements to emit light in the visible or UV regions. To achieve the higher temperatures, electrical arc, spark, ionized gases or plasmas are used to induce light emission instead of a cooler conventional flame. The inductively coupled plasma atomic emission spectrometer (ICP–AES) (also called an inductively coupled plasma optical emission spectrometer, ICP–OES) devices are versatile instruments capable of measuring elemental composition with high sensitivity and accuracy and are currently used to detect trace elements and heavy metals in human urine or serum. The ICP–AES devices use temperatures in excess of 5000 °C and can achieve resolution of 0.0075–0.024 nm.

The line spectra of pure elements often contains doublets, triplets, or multiplets of emission lines. Planck's law implies there is very little difference in energy level associated with spectral lines that have similar wavelengths. Spectral line splitting is attributed to the quality of electron spin that can have two quantum values of slightly different energy. Doublets are observed from atoms with a single outer electron (with two possible electron spin values), triplets from atoms with two outer electrons and multiplets from atoms with several outer electrons. High resolution instruments are required to distinguish some of the multiplet lines. With use of high temperature flames or ICP source and high resolution instruments, the spectra of atoms, ions, and molecules can be observed. Alkaline earth metals can become ionized and the resulting cations can generate emission spectra. Oxides and hydroxides of alkaline earth metals can be present in samples or generated in the flame and can also generate emission spectra. In practice, an emission wavelength used to measure an element is selected when the line provides sufficient intensity to provide adequate sensitivity and the line is free from interference from other lines near the selected wavelength. The amount of light emitted at a specific emission wavelength may not be sufficient for analysis. For this reason, flame emission spectrometry methods may not be applicable and alternate methods, such as atomic absorption spectrometry or inductively coupled plasma mass spectrometry, may be preferred.

THEORY: ATOMIC ABSORPTION SPECTROMETRY

Atomic absorption spectrometry developed from the observations used to establish Kirchhoff's third law: 'That a dark line or absorption line spectrum is seen when a source of continuous spectrum is viewed behind a cool gas under pressure' (Fig. 1). When a sample is introduced into a flame, only a small fraction of the atoms are in an excited or high energy state (according to the Boltzmann equation) and most atoms of the element remain in the unexcited or ground state and are capable of absorbing energy. If light energy is provided as a continuous spectrum, ground-state atoms will selectively absorb wavelengths of light that correspond with the energy intervals required to excite electrons from low energy to higher energy levels. The wavelengths of light that are absorbed as electrons are excited from ground state to higher energy levels are analogous to the wavelengths of light emitted as high energy electrons return to the ground state. The advantage of atomic absorption spectrometry is that most of the atoms in a flame remain in the unexcited ground state and are capable of light absorption, allowing this method to be ~100-fold more sensitive than flame emission spectrometry.

Atomic absorption of light is described by an equation analogous to the Lambert–Beer law. The absorbance of light, A, is defined as the logarithm of the ratio of initial light beam intensity I_0 to the intensity of the beam after light absorption, I. The absorbance of light is directly proportional to the concentration of absorbing atoms c and the path length of the light beam in the atoms d. The value k is a constant referred to as the absorptivity constant.

$$A = \log\left(\frac{I_0}{I}\right) = kcd$$

The emission of characteristic wavelengths or colors of light from materials in a flame was initially observed by the unaided human eye and preceded the measurement of atomic absorption that required instrumentation. To observe atomic absorption of light, a beam of light was passed through atoms introduced to a flame and the ratio of initial and postabsorption light beams was determined. However, the light beam emitted from the flame also contains light derived from the combustion of materials in the flame that was not present in the initial light beam. To measure the atomic absorption of light, the intensity of background light emitted from the flame itself must be subtracted from the postabsorption light beam intensity prior to calculating the absorbance. Electrothermal atomic absorption (also called flameless atomic absorption or graphite furnace atomic absorption) uses an electrically heated furnace to atomize the sample. By design, the electrothermal instruments have greater sensitivity and lower detection limits because analyte atoms can achieve higher gaseous

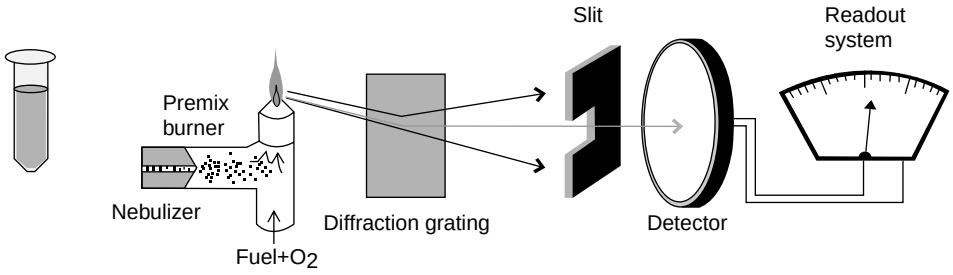


Figure 3. Components of a flame atomic emission spectrometer.

concentrations and residence time in the light beam and there is less background light emission.

The sensitivity of atomic absorption methods was improved by using light sources that generated wavelengths of light that exactly matched the wavelength of maximum absorption by the element undergoing analysis. The light sources are referred to as hollow cathode lamps. These lamps heat a small amount of the element of interest and generate line spectra emission for the light beam at wavelengths that exactly match the wavelengths maximally absorbed by atoms of that element in the ground state in the flame. Often hollow cathode lamps contain a single element, however, multielement lamp sources with six or more elements are commercially available.

EQUIPMENT

Flame Atomic Emission Spectrometry

The components of a flame photometer are illustrated schematically in Fig. 3 and consist of a nebulizer within a spray chamber, a premix burner, flame, monochromator, detector, and readout system. The monochromator, detector, and readout systems are similar to those used in a spectrophotometer: monochromatic light is obtained by means of an interference filter or diffraction grating, a detector consisting of a photomultiplier tube, result display (visual meter, digital display, or data capture a computer system), and the flame serves as a light source and sample compartment.

Each of the basic components of the instrument contributes to the analytical process. Various combinations of combustible gases and oxidants can be used to create flames with different temperatures (e.g., acetylene, propane, natural gas, oxygen, compressed air). The nebulizer is an important component responsible for mixing the analyte liquid with compressed air and converting it into a fine mist that enters the flame. For precise analytical measurements, the nebulizer must create a mist of consistent composition (10–15% of the aspirated sample) and mix the mist into the gases that combust to create a consistent flame. Within the spray chamber, large droplets settle out and are drained to waste, and only a fine mist enters the flame. A wetting agent (e.g., a nonionic detergent) may be added to standards and samples to minimize changes in nebulizer flow rates that result from differences in the viscosity or matrix of samples.

The intensity of emitted line spectra from atoms excited in the flame must be discriminated from other light in the flame. Less sophisticated instruments (e.g., flame photometers) rely on filters with narrow bandpass to eliminate background light emitted from the flame. Diffraction gratings are used as monochromators in more sophisticated instruments with a slit to limit the bandwidth of light being detected.

To improve the reliability, an internal standard may be used when designing a method. Usually an element not present in the sample is introduced into the calibration standards, quality control, and unknown solutions. In biological applications, lithium or cesium is frequently used for this purpose. By measuring the ratio of emission light intensity of the target element to the internal standard, there is compensation for small variation in nebulization rate, flame stability and solution viscosity, and methods perform with greater precision. Addition of either lithium or cesium to all solutions can also prevent interference by acting as an ionization buffer (previously referred to as a radiation buffer). A relatively high concentration of an alkali metal ion in the solutions creates a buffer mixture of its ionic and atomic species and free electrons in the flame. The elevated free-electron concentration in the flame represses and stabilizes the degree of ionization of analyte atoms and improves the atomic emission signal.

Inductively coupled plasma emission spectrometers have similar components. The sample is nebulized into argon gas that is heated to extreme temperatures ($>5000\,^{\circ}C$) that efficiently breaks molecular bonds to generate gaseous atoms. Because few molecules exist at these temperatures, ICP–AES has few chemical interferences compared to flame techniques, however, because of the high temperature and large proportion of atoms and ions that emit light, ICP–AES methods are more prone to spectral interferences associated with inadequate resolution of spectral lines.

Atomic Absorption Spectrometry

The components of a flame atomic absorption spectrometer are illustrated schematically in Fig. 4 and consist of a nebulizer, premix burner, hollow cathode lamp, modulation system or beam chopper, flame, monochromator, detector, and readout system. The monochromator, detector, and readout systems are similar to those used in a spectrophotometer. Monochromatic light is obtained by isolating a single spectral line emitted from a hollow cathode lamp using an interference filter. The detector is a photomultiplier tube and the results can be displayed by a visual meter, digital display, or captured by a computer system. The flame serves as a sample compartment and

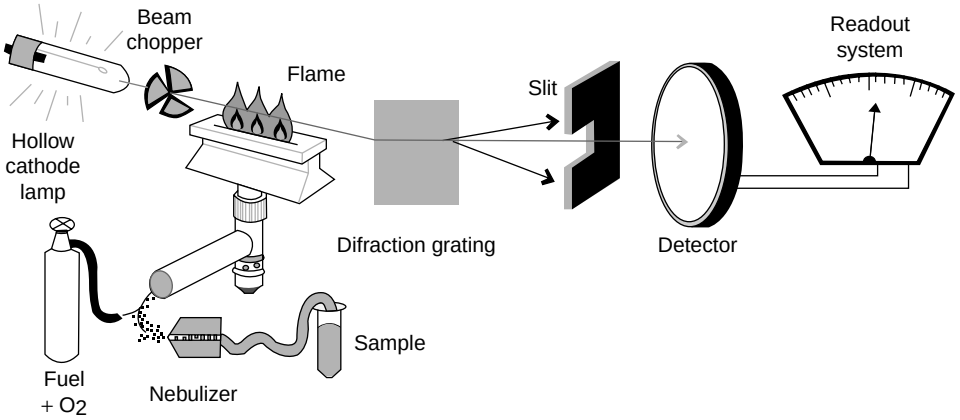


Figure 4. Components of a flame atomic absorption spectrometer.

light from the hollow cathode lamp is passed through the flame to measure absorption. Electronic modulation or a mechanical beam chopper is used to interrupt the light from the hollow cathode lamp into pulses so the extraneous light emitted from the flame alone can be measured and subtracted prior to measuring atomic absorption.

Electrothermal atomic absorption instruments typically use a graphite furnace to rapidly heat a sample in a uniform manner to ~2600 °C. Samples are typically introduced into the furnace on a carbon rod. As the temperature is raised, the samples are dried, charred, rendered to ash, and atomized (e.g., Table 2). Hollow cathode lamps are used as a source of monochromatic light to determine atomic absorption. In contrast to flame-based methods that rapidly dilute and disperse analyte atoms, electrothermal methods retain atoms at higher concentrations with longer residence time in the light beam. In addition, there is less extraneous background light. These advantages allow electrothermal atomic absorption instruments to measure many trace elements and heavy metals not capable of measurement with flame atomic absorption instruments (15).

There are four categories of interference that can occur with atomic absorption spectrometry: chemical, spectral, ionization, and matrix. The element undergoing analysis may exist or form stable molecules within the flame. The atoms in the stable molecules do not generate line spectra at the characteristic wavelengths of the free atoms and this chemical interference results in a negative analytic bias for the method. Spectral interference can occur when there is nonspecific absorption of light by material in the sample. Nonspecific absorption of light can be caused by solids or particles in the flame. Spectral interference can often be reduced by performing maintenance on the burner head to remove residues and assure a steady flame is maintained. Emission of light from other components in the flame can interfere with analysis and can often be removed with a background correction method. Ionization interference occurs when atoms in the flame are converted to ions that will not absorb the line spectra at the characteristic wavelengths for the atoms. Ionization interference can be reduced by an ionization buffer or by lowering flame temperature to reduce the extent of ionization. Matrix interferences occur when the calibration standards have a different composition than the samples (e.g., viscosity, salt concentration or composition, pH, or presence of colloids, e.g., protein). Matrix interferences can be avoided in the method design by assuring calibrator solutions have similar composition to the samples or that samples are highly diluted so they acquire a matrix similar to the calibrator solutions.

Table 2. Graphite Furnace Settings for Blood Lead Determination[a]

Stage	Temperature, °C	Ramp Time, s	Hold Time, s	Gas flow, $L \cdot min^{-1}$
Dry	120	5	15	300
Preash	260	1	5	300
Ash	800	5	27	300
Cool	20	1	4	300
Atomize	1600	0	5	0
Clean	2700	1	2	300

[a]See Ref. 14.

Several methods have been developed to improve the sensitivity of atomic absorption by reducing background radiation spectral interference. Electronically modulated light beams or mechanical beam choppers have been used to generate pulsed light so the background emission of light from the flame can be distinguished, measured and subtracted. Analogous to molecular absorption spectrophotometers, atomic absorption spectrometers can have a single beam design (Fig. 4) or a split double-beam design where the beam modulator acts as a beam splitter to create the light beam that passes through the sample and a second beam that acts as a reference. Split double-beam instrumentation has an advantage of compensating for variation in light intensity from the lamp or variation of detector sensitivity.

In electrothermal atomic absorption methods, the rapid heating of samples can cause the release of smoke or nonspecific vapor that blocks some light from reaching the detector. To correct for this nonspecific light scatter, in addition to the beam from a hollow cathode lamp, a beam from a deuterium continuum lamp can be direct through the sample chamber and light intensity measured at a wavelength other than that used for the assay to assess the extent of light scatter.

A third method of background correction is based on the Zeeman effect. When atoms are placed in a strong magnetic

field, the emission and absorption line spectra are dramatically altered as the single line spectra are split into triplets or more complex patterns. Splitting of a spectral line into two different wavelengths implies that the electrons involved have slightly different energy levels, according to Planck's law. The small difference in the energy levels is attributed to slightly different quantum numbers for the electrons involved and this difference in energy level can only be detected in the presence of the magnetic field. In the absence of the magnetic field, slight differences in energy levels are not apparent and a single wavelength of absorption or emission is observed. By turning the magnetic field off and on during atomic absorption spectrometry, the population of atoms being tested is modulated from a uniform population to discrete populations with different quantum numbers. By measuring the element-specific difference in atomic absorption as the magnetic field is modulated, nonspecific background absorption or emission can be accurately subtracted. Implementation of instruments with Zeeman correction is complex, expensive, and requires optimization of the magnetic field for each element. Consequently, Zeeman correction approaches are usually reserved for methods involving strong background interferences.

MEDICAL APPLICATIONS

Flame emission spectrometry and atomic absorption spectrometry are sophisticated methods of elemental analyses that are robust and flexible techniques that have been applied to the analysis of many different types of biological or clinical samples. In the 1950s and 1960s, use of flame photometers provided state-of-the-art accurate and precise determinations of serum, sodium, and potassium. Stable, reliable, and inexpensive potentiometric methods for determination of electrolyte concentrations were developed and replaced flame photometric methods for most medical applications. The hardy reliable nature of flame photometers is still used for some analysis (e.g., the determination of stool electrolytes). While flame emission photometric methods can be applied to determine the concentration of trace elements or heavy metals in biological specimens, many clinical laboratories use electrothermal atomic absorption spectrometry or inductively coupled plasma mass spectrometry that offer greater specificity, sensitivity, precision, and adaptability.

Flame emission spectrometric determinations are prone to negative bias when biological samples undergoing analysis have high concentrations of lipid or protein (16). Biological samples are mixed with a diluent to create a mist in the nebulizer and the mist is created by mixing constant proportions of each fluid. However, when a biological fluid has a high concentration of lipid or protein, the constant volume of sample fluid has a reduced amount of aqueous dissolved ions because lipid particles suspended in the fluid occupy volume and reduce the volume of water available to dissolve the ions. In a similar manner, proteins bind water molecules creating a hydration shell and at high concentration in an aqueous solution, the proteins occupy volume in the sample and reduce the volume of water available to dissolve ions. This negative bias is known as the volume exclusion effect. The term pseudohyponatremia is used to describe the misleading low serum sodium concentrations attributed to this bias. This volume exclusion effect was initially characterized with flame photometers and it remains a concern with potentiometric instruments that use indirect electrodes and rely on sample dilution (17).

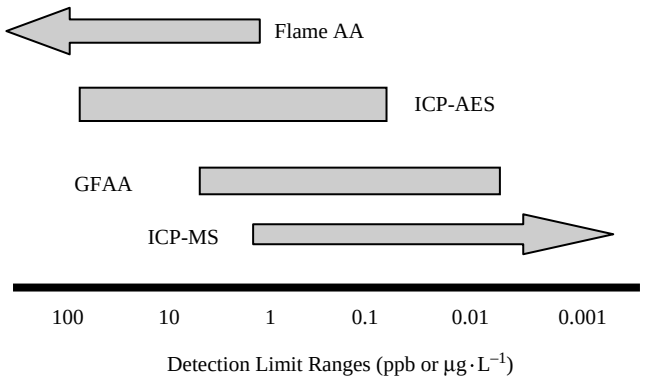


Figure 5. Typical detection limit ranges for major spectrometry techniques. Flame atomic absorption, inductively coupled plasma-atomic emission spectrometry, graphite furnace atomic absorption, inductively coupled plasma–mass spectrometry (19).

The use of atomic absorption spectrometers in clinical laboratories for the determination of calcium, magnesium, and other metal cations reached a peak in the 1970s. In the following decades, potentiometric methods became commercially available for the most common tests (e.g., calcium, magnesium, sodium, and potassium) and atomic absorption spectrometers were not commonly used in each hospital laboratory. A concise review of the history of atomic absorption spectrometry application in clinical chemistry was published in 1993 (18). The atomic absorption spectrometer remained an important platform for the determination of serum zinc, copper, and lead. Medical laboratory services have often been consolidated, and infrequent tests (e.g., serum zinc, copper, lead, trace elements, and heavy metals) are shipped to reference laboratories for analyses. While inductively coupled plasma mass spectrometry offers greater sensitivity and flexibility of analysis, flame and electrothermal atomic absorption spectroscopic methods remain cost-effective, precise analytic methods, widely used in clinical reference laboratories to support the evaluation of serum or blood concentrations of zinc, copper, aluminum, chromium, cadmium, mercury and lead (Fig. 5).

BIBLIOGRAPHY

Cited References

1. Melvill T. Observations on light and colors, Essays Observ. Phys Lit Edinburgh 1756;2:12–90.
2. Wollaston WH. A method of examining refractive and dispersive powers. Philos Trans R Soc London 1802;92:365–380.
3. Herschel JFW. On the absorption of light by coloured media. Trans R Soc Edinburgh 1823;9:445–460.
4. Talbot HF. Some experiments on colored flames. Edinburgh J Sci 1826;5:77–81.
5. Kirchhoff G, Bunsen R. Chemical analyses by means of spectral observations. Ann Phy (Leipzig) 1860;110(2):160–189.

6. Lundegardh H. New contributions to the technique of quantitative chemical spectral analysis. Z Phys 1930;66: 109–118.
7. Lundegardh H. Investigations into the quantitative emission spectral analysis of inorganic elements in solutions. Lantbrukshoegsk Ann 1936;3:49–97.
8. Leyton L. An improved flame photometer. Analyst 1951;76: 723–728.
9. Ivanov DN. The use of interference filters in the determination of sodium and potassium in soils. Pochvovedenie [N.S.] 1953;1: 61–66.
10. Walsh A. The application of atomic absorption spectra to chemical analysis. Spectrochim Acta 1955;7:108–117.
11. Alkemade CTJ, Milatz JMW. Double beam method of spectral selection and flames. J Opt Soc Am 1955;45:583–584.
12. Woodson TT. A new mercury vapor detector. Rev Sci Instrum 1939;10:308–311.
13. Lutz RA, Stojanov M. Flame Photometry. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation. New York: John Wiley & Sons Inc; 1988.
14. Bannon DI, et al. Graphite furnace atomic absorption spectroscopic measurement of blood lead in matrix-matched standards. Clin Chem 1994;40(9):1730–1734.
15. Butcher DJ. Joseph Sneddon. A Practical Guide to Graphite Furnace Atomic Absorption Spectrometry. New York: John Wiley & Sons Inc; 1998.
16. Waugh WH. Utility of expressing serum sodium per unit of water in assessing hyponatremia. Metabolism 1969;18: 706–712.
17. Lyon AW, Baskin LB. Pseudohyponatremia in a myeloma patient: Direct electrode potentiometry is a method worth its salt. Lab Med 2003;34(5):357–360.
18. Willis JB. The birth of the atomic absorption spectrometer and its early applications in clinical chemistry. Clin Chem 1993; 39(1):155–160.
19. Perkin Elmer Life Sciences (2004). Guide to Inorganic Analysis. Available at http://las.perkinelmer.com/Content/RelatedMaterials/005139C_01_Inorganic_Guide_web.pdf Accessed 1 Aug, 2005.

See also ANALYTICAL METHODS, AUTOMATED; FLUORESCENCE MEASUREMENTS.

FLAME PHOTOMETRY. See FLAME ATOMIC EMISSION SPECTROMETRY AND ATOMIC ABSORPTION SPECTROMETRY.

FLOWMETERS

ARNOLD A. FONTAINE
STEVEN DEUTSCH
KEEFE B. MANNING
Pennsylvania State University
University Park, Pennsylvania

INTRODUCTION

Fluid flow occurs throughout biomedical engineering in areas as different as air flow in the lungs (1,2) to diffusion of nutrients or wastes through membranes (3–5). Flow-related problems can involve fluid media in the form of gas, liquid, or multiphase flows of both liquids and gas together or in combination with solid matter. Biomedical flows occur in both *in vivo* (6) and *in vitro* (7,8). They can involve relatively benign flows like that of saline through an intravenous tube to a biochemically active flow of a non-Newtonian fluid such as blood. Many biomedical or bioengineering processes require the quantification of some flow field that may be directly or indirectly related to the process. Such quantification can involve the measurement of volume or mass flow, the static and dynamic pressures, the local velocity of the fluid, the motion (speed and direction) of particles such as cells, the flow-related shear, or the diffusion of a chemical species.

Interest in understanding fluid flows and attempts to measure flow-related phenomena has had a long history in the scientific and medical communities with early work by Newton, DaVinci, and others. Early studies often involved observations of flow-related phenomena that can be characterized as simple flow visualization or particle tracking (9) or the estimation of a pressure by the displacement of a fluid. Rouse and Ince (10) provide a historical review of these early works. Throughout the years, flow measurement techniques have advanced significantly in capability (what is measured and how), versatility, and in many ways, complexity. Some techniques, such as photographic flow visualization, have changed little in over 100 years, whereas others are only possible because of advances in electronics, optics, and physics. Improved capability and versatility are evidenced through the increased ease of use in some systems and the ability to measure more quantities with increased accuracy and resolution. However, this improved capability and versatility has, in some cases, come at the cost of increased complexity in system hardware, calibration requirements, and application complexity.

Measurement techniques can be characterized as invasive or noninvasive and direct or indirect. Invasive measurement techniques require the insertion of a sensing or a sampling element directly into the flow field. As a result of this direct insertion, invasive probes may alter the flow field characteristics or induce bias errors associated with the presence of the probe in the flow field or by the operation of the probe (11,12). Invasive probes are often designed to minimize flow disturbance by miniaturizing the sensing elements or by displacing the sensing elements some distance from the hardware holding the probe in the flow, as illustrated in Fig. 1. Invasive probes also require some type of closure at the penetration site through the boundary of the flow, which must be accounted for in the test design and can be particularly important in *in vivo* applications to prevent fluid loss or infection. White et al. (13) measured wall shear stress in the abdominal aorta of dogs using an invasive, flush-mounted hot-film probe and described how the probe tip is modified to provide an effective entry mechanism through the arterial wall with an adequate seal.

Noninvasive techniques do not involve direct insertion of a sensor into the flow but provide sensing capability through either access to the flow at the flow boundary or through the use of some form of electromagnetic radiation (EMR) transfer or propagation. Wall-mounted thermal sensors, static pressure taps and transducers, or surface

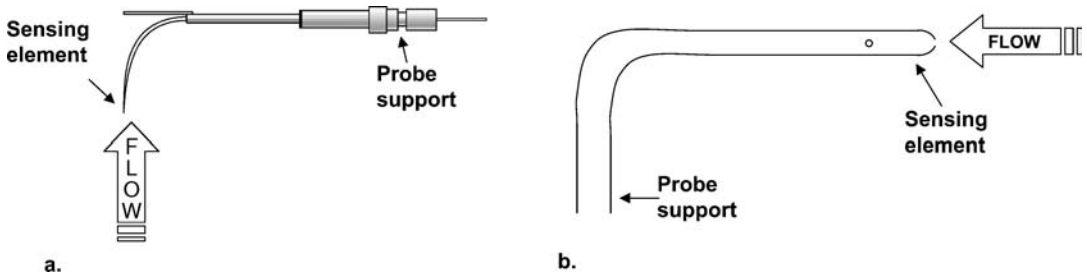


Figure 1. Examples of invasive velocity measurement sensors with displaced sensing elements relative to their probe supports. (a) a boundary layer style hot-wire probe for thermal anemometry. Picture from TSI Inc. catalog, probe catalog 1261A. (b) Pitot static probe for velocity measurement.

sampling probes are examples of noninvasive techniques that require access to the boundary of the flow field through a wall penetration (13). Ultrasound (14) magnetic resonance (MR), X-ray, and optical techniques are all examples of electromagnetic radiation that can be used to probe flow fields (15). Unlike the wall-mounted invasive probes describe above, EMR-based measurement systems do not require physical penetration into the flow field or access through the flow field boundary. They do, however, require a "window" into the flow through the boundary enclosing the flow field of interest. This "window" depends on the type of technique being used. Optical-based techniques require an optically clear window that may not be suitable for *in vivo* applications, whereas ultrasound and X-ray techniques require that the material properties of the flow boundaries are transparent to these forms of EMR waves. For example, lead will shield X-ray penetration and metal objects on a surface or in the flow may create local artifacts (noise or error) in MR measurements.

Direct and indirect measurements are defined by how quantities of interest are measured. The displacement of a particle or cell in a flow can be directly measured by photographing the movement of the particle over a finite time interval (16). Most flow-related measurement systems, however, are indirect. In general, velocity or flow is indirectly calculated from the direct measurement of a quantity and the application of a calibration that relates the magnitude of the measured quantity to the parameter of interest. This calibration may involve not only a conversion of a measured quantity like a voltage to a physical quantity such as a pressure, but may also involve the application of a functional relationship (i.e., Bernoulli's equation), which requires assumptions about the flow. For example, volume flow probes often assume a characteristic velocity profile at the location of the probe (17). Blood flow in the microcirculation can be estimated by indirectly measuring the cell velocity using a time-of-flight optical technique where the time a cell takes to move a known distance is measured and the cell velocity is calculated by the ratio of the distance divided by the transit time (18).

Indirect measurement can also impact the uncertainty in the estimated quantity. The measurement of velocity using a Pitot probe is one example of an indirect measurement (19). The Pitot probe measures the local dynamic and static pressures in the flow. These pressures are most often measured using a pressure transducer that provides a voltage output in response to an applied pressure. A calibration is then applied to the measured voltage to convert it to pressure. Velocity is indirectly calculated from the estimated pressures using the Bernoulli equation. The error in Pitot probe velocity measurements includes the pressure transducer calibration uncertainty, noise and statistical uncertainty during the pressure measurement, electronic noise in the acquisition of the transducer output voltage, and transducer drift and potential bias due to the physical size of the probe relative to the flow scales being measured. These errors are nonlinearly propagated into the estimate of the velocity uncertainty.

Measurement accuracy is also a function of the physical and operating characteristics of the probe itself. Many flows exhibit a range of spatial and temporal scales. The physical size and the frequency response of the sensing element must be taken into account when choosing a measurement system for a particular application. A large sensing element or an element with poor frequency response has the effect of low pass filtering the measured signal (20). This low pass filtering will cause a bias in the measured quantity. The total measurement uncertainty must also take into account statistical errors associated with random processes, cycle-to-cycle variability in pulsatile systems, and noise. The reader is referred to the texts by Coleman and Steele (21) and Montgomery (22) for a detailed approach to experimental uncertainty analysis. The focus of this chapter will be on measurement techniques, their fundamentals of operation, their advantages and disadvantages, and examples of their use.

The name flow measurement is a broad term that can encompass the measurement of many different flow-related parameters. In this chapter, the authors will focus on the measurement of those parameters that are most often desired in a biomedical/bioengineering application, volume flow rate, and velocity. Imaging, Doppler echocardiography, and MR techniques are addressed in other chapters within the encyclopedia and, thus, will only be briefly introduced in this article when applicable. This chapter is subdivided into sections that will address volume flow and velocity separately, with a detailed presentation of systems that are available for the measurement of each. Although ultrasound and MR techniques are often used to measure flow-related parameters, a detailed discussion of the principles of operation will not be presented here as these topics are covered in depth in other chapters of this Encyclopedia.

FLOW MEASUREMENT APPLICATIONS

Volume Flow Measurement

In both the clinical environment and the laboratory environment, the measurement of the volume flow rate of a fluid as a function of time can be an important parameter. In internal flow applications, which comprise most biomedical flows of interest, the volume flow of a fluid (Q) is related to the local fluid velocity (V) through the integration of the velocity over the cross sectional area of the duct or vessel (19,23).

$$Q = \int V\, dA \tag{1}$$

The flow rate Q, velocity V, and area A have dimensions of volume per time, length per time, and length squared, respectively. The SI units are typically used in the bioengineering field with mass units of grams or kilograms, length units of meters (millimeter and centimeters), and time units of seconds. The mass flow (M) is directly related to the flow volume through the fluid density, ρ, with units of mass/volume.

$$M = \rho \cdot Q \tag{2}$$

Fluid pressure and velocity are related through the Navier–Stokes equations, which govern the flow of fluids in internal and external flows [see White (23)].

The volume flow rate of blood is often measured in many cardiovascular applications (24). For example, cardiac output (CO) is the integrated average of the instantaneous volume flow rate of blood (Q_b) exiting the aortic valve over one cardiac cycle (T_c):

$$\mathrm{CO} = \left[\int Q_b dt\right] / T_c \tag{3}$$

The cardiac output has units of volume flow rate, volume per unit time. The following subsections will provide an overview of measurement techniques typically used for both volume flow and velocity measurement in *in vivo* and *in vitro* studies. This article will be limited to those flow measurement techniques most often used in the biomedical and bioengineering fields. Specialized measurement techniques, such as concentration or species measurement through laser-induced fluorescence (LIF) or mass spectrometry, will not be addressed.

Electromagnetic Flow Probes. Carolina Medical Inc. developed the first commercially available electromagnetic flow meter in 1955. The design provided scientists and clinicians with a noninvasive tool that could directly measure the volume flow rate (17). Clinical probes were developed that could be attached to a vessel for extravascular measurement of blood flow without the need for cannulation of a surgically exposed vessel.

Electromagnetic flow meters are volumetric flow measuring devices designed to measure the flow of an electrically conducting liquid in a closed vessel or pipe. Commercial meters, used in the biomedical and general engineering fields, come in a variety of sizes and designs that can measure flow rates from $\sim 1\ \mathrm{mL \cdot min^{-1}}$ to $>100{,}000\ \mathrm{L \cdot min^{-1}}$. Reported uncertainties are typically on the order of a few percent, but can be as low as 0.5% of reading in some specialized meter designs. Low uncertainties are dependent on proper use and installation of the meter. Improper use or installation (mounting the meter in a location with a complex flow profile) will increase measurement uncertainty. Most clinical quality meters that are mounted externally to a vessel exhibit uncertainties that can approach 15% in some applications (Carolina Medical Inc., product support literature). In cardiovascular applications, these meters may also be susceptible to electrical interference by the heart or by measurement anomalies due to motion of the probe.

The principle governing the operation of an electromagnetic flow meter is Faraday's law of electromagnetic induction. This law states that an induced electrode voltage is proportional to the velocity of flow of a conductor through a magnetic field of a known density. Mathematically, this law is represented as:

$$E_e = K[V \cdot B \cdot L_e] \tag{4}$$

Here, E_e is the induced voltage between two electrodes (with units of volts) separated by a known conductor length L_e (provided by the conducting fluid between the electrodes) in units of millimeters or centimeters, B is the magnetic field strength in units of Tesla's, and V is the conducting fluid average velocity in units of length per time. The parameter K is a dimensionless constant. Meter output is linear and proportional to flow velocity.

Fluid properties such as viscosity and density are absent from Eq. 4. Thus, the output of an electromagnetic flow meter is independent of these properties and its calibration is independent of the type of fluid, provided the fluid meets minimum conductivity levels. This meter can then be used for highly viscous fluids, Newtonian fluids, and nonNewtonian fluids such as blood. The requirement of an electrically conductive fluid can disqualify an electromagnetic meter in some applications. Typical meters require a minimum fluid conductivity of $\sim 1\,\mu\mathrm{S/cm^{-1}}$. However, low conductivity designs are capable of operating with fluid conductivities as low as $0.1\,\mu\mathrm{S/cm^{-1}}$. The presence of gas bubbles in the fluid can cause erratic behavior in the meter output.

A typical meter design has electromagnetic coils mounted on opposing sides of an electrically insulated duct with two opposing electrodes mounted 90° relative to the electromagnets. The two electrodes are mounted such that they are in contact with the conducting fluid. Figure 2 illustrates the typical configuration.

The meter is designed to generate a magnetic field that is perpendicular to the axis of motion of the flowing fluid. A voltage is generated when the conducting fluid flows through the magnetic field. This voltage is sensed by the two opposing electrodes. The supporting material around the meter is made of a nonconducting material to prevent leakage of the voltage generated in the moving fluid into the surrounding material. In practice, the conductor length, L_e, is not the simple path illustrated, but is rather the integral of all possible path lengths between the two electrodes across the cross section of the duct or vessel. The

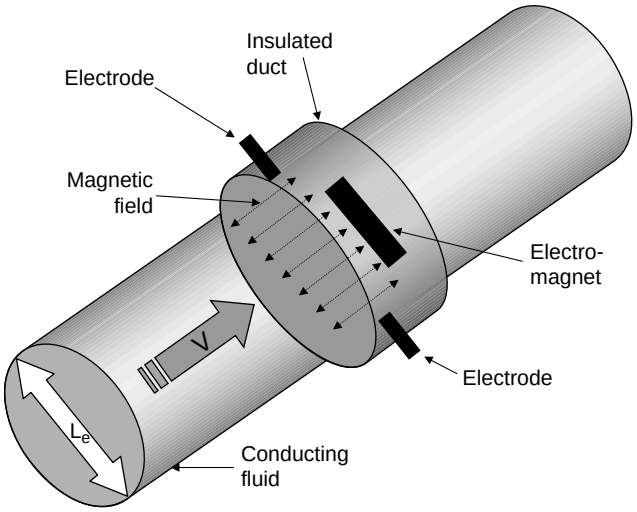


Figure 2. Illustration of the principle of operation of the electromagnetic flow meter. Note, the magnetic field, electrodes, and flow direction are all mutually perpendicular to one another.

signal generated along each path length is proportional to the fluid velocity across that path. Thus, the two electrodes measure the integrated sum of all velocities across every possible path in the vessel cross section. This signal is then directly proportional to the volume flow rate of the fluid passing through the magnetic field.

The magnetic field generated in commercial meters may be anything from a uniform field to a specifically designed field with prescribed characteristics. Meters with uniform magnetic fields can exhibit some sensitivity to the conducting liquid's velocity profile. Fluid flowing through a vessel does not have the same velocity at all locations across the vessel. The no-slip condition at the walls of the duct ensures that the fluid velocity at the wall is zero (19). Viscosity then generates a gradient between the flowing fluid in the vessel and the stationary fluid at the wall. In complex geometries, secondary flows may occur and velocity gradients in other directions may also develop (25,26). This variation in fluid velocity across the magnetic field coupled with variations in the conductor length generates a variation in the magnitude of the voltage measured across the duct. As a result, installation of these meters must be carefully performed to ensure that the velocity profile of the liquid in the tube is close to that used during calibration.

A number of commercial meters shape the magnetic field coils to generate a magnetic field that exhibits a field strength with a prescribed pattern across the duct. This field shaping compensates for some velocity variations in the duct and provides a meter with a reduced sensitivity to flow profile. As a result, this type of meter is better able to measure in vessels with upstream and downstream characteristics, such as curvature and nonuniformity in the vessel cross section, that generate asymmetric flow profiles with secondary flow velocity components.

Commercial meters generate magnetic fields using either an ac excitation or a pulsed dc excitation. The ac excitation generates a magnetic field with a strength that varies with the frequency of the applied ac voltage. This configuration produces a meter with a relatively high frequency response but with the disadvantage that the output signal not only varies with the flow velocity but also with the magnitude of the alternating excitation voltage. Thus, the output of the meter for a flow with constant velocity across the vessel will exhibit a sinusoidal pattern. In addition, zero flow will produce an offset output due to the presence of the nonmoving conductor in a moving magnetic field. Quadrature signal rejection techniques can be used to filter the unwanted signal generated by the ac excitation from the desired signal generated by the flowing liquid, but this correction requires careful compensation and zeroing of the meter in the flow field before data acquisition.

The pulsed dc excitation was developed to reduce or eliminate the zero shift encountered with ac excitation. This improvement has the cost of reduced frequency response and increased sensitivity to the presence of particulates in a fluid. Particles that impact the electrodes in a pulsed dc operated meter produce output fluctuations that can be characterized as noise. The accuracy in each system is comparable.

A low sensing voltage at the electrodes requires a signal conditioning unit to provide a measurable output with a good signal-to-noise ratio. Meter calibrations typically involve one of two calibration techniques. The meter and signal conditioning unit are calibrated separately or the meter and signal conditioning unit are calibrated as a system. The latter provides the most accurate calibration with accuracies that can approach 0.5% of reading in certain applications. The reader is referred to literature by various manufacturers of electromagnetic flow meters for a more comprehensive discussion of the techniques, operation, and use for specific applications.

Ultrasound Techniques – Transit Time Volume Flow Meters. Ultrasonic transit time flow meters provide a direct measure of volume flow by correlating the change in the transit time of sound waves across a pipe or vessel with the average velocity of the liquid flowing through the pipe (17,27,28). Transit time ultrasonic flow meters are widely used in clinical cardiovascular applications. In recent years, a number of studies have been performed to evaluate and compare transit time ultrasonic flow measurement techniques with other techniques used clinically (29). The typical configuration for an ultrasonic flow probe involves one or two ultrasonic transducers (transmitters/receivers) and possibly an ultrasonic reflector. Transducers and reflectors are positioned in opposing configurations across a tube or vessel as illustrated in Fig. 3. The time it takes an ultrasound wave to propagate across a fluid depends on the distance of propagation and the acoustic velocity in the fluid. If the fluid is moving, the motion of the fluid will positively or negatively add a phase shift to the time of propagation (transit time) through the fluid, which can be written mathematically as:

$$T_t = D_p/(c \pm V \cdot \cos\theta) \tag{5}$$

Here, T_t is the measured transit time (s), D_p is the total propagation distance of the wave (length units), c is the acoustic speed of the fluid (units of length per time), V is

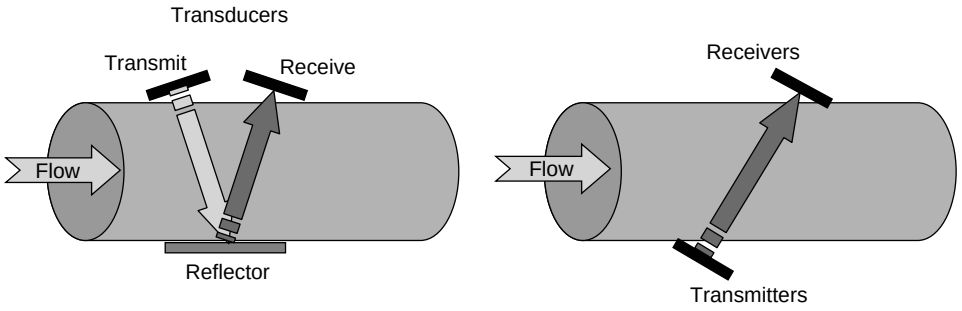


Figure 3. Illustration of principal of transit time ultrasonic flow probe operation.

the average velocity of the fluid (units of length per time) over the propagation length, and θ is the angle between the flow direction and the propagation direction of the wave. The configurations illustrated in Fig. 3 have an inherent dependency of the measured transit time on the coupling of the transducer with the vessel. Acoustic impedance characteristics of the vessel wall, and mismatches in impedance at the vessel wall–transducer and wall–fluid interfaces will affect the accuracy of the flow rate measurement.

The approach to using equation 5 in a metering device is to incorporate bidirectional wave propagation in opposing directions, as shown in Fig. 4, which will produce two independent transit time measurements (T_{t1} and T_{t2}), one from each direction of propagation. The forward direction transit time T_{t1} is defined by Eq. 5 with a plus sign before V, and T_{t2} is defined by the minus sign. The fluid velocity can then be obtained by taking the difference of the transit times ($T_{t1}-T_{t2}$). It can be shown that, for fluid velocities small relative to the acoustic velocity of the fluid ($V^2 << c^2$), this difference reduces to:

$$(T_{t1} - T_{t2}) = 2D_p(V \cdot \cos\theta)/c^2 \qquad (6)$$

With the probe geometry defined (the propagation distance and propagation angle relative to the vessel or flow direction) and with the fluid properties known (acoustic speed of the fluid), the average speed of the fluid along the propagation path of a narrow beam can be calculated from the transit times. Wide beam illumination, where the beam width is wider than the vessel diameter, effectively integrates the measured transit time phase shift over the cross section of the vessel. The wide beam transmission can be approximated by the summation of many infinitesimally narrow beams adjacent to one another. Thus, the measured wide beam phase shift is proportional to the sum of these narrow beams propagating through the vessel. As the phase shift encountered in a narrow beam transmission is proportional to the average fluid velocity times the beam path length, integrating or summing over all narrow beams across the vessel results in a measured total transit time phase shift that is proportional to the average fluid velocity times the area of the vessel sliced by the ultrasound beam, or the volume flow rate.

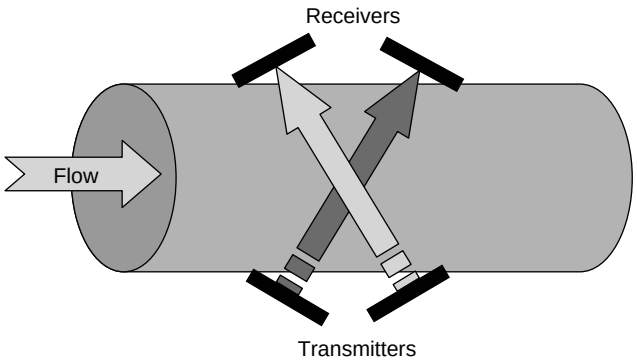


Figure 4. Bidirectional wave propagation.

The popular Transonic Systems Inc. flow meter uses bidirectional transmission with two transducers operating in transmit and receive modes, and a single reflector configured as illustrated in the left schematic of Fig. 4. This approach increases the propagation length while effectively reducing sensitivity to wall coupling or misalignment of the probe with the wall. The increased path length improves uncertainty and provides a probe body with a relatively small footprint, an advantage in *in vivo* or surgical applications.

The basic operation of a bidirectional transit time meter involves the transmission of an ultrasound plane wave at a specific frequency. This wave propagates through the vessel wall and fluid where it is either received at the opposite wall or is reflected to another transducer operating as an acoustic receiver. This received signal is recorded, processed, and digitized before the transducer is reconfigured to transmit a second pulse in the opposite direction. The overall frequency response of such a probe is dependent on the pulse time, the time delay between the forward and reverse pulses, the acoustic speed through the medium, the propagation distance, and the signal conditioning electronics, which can include analog signal acquisition and filtering. The meter size governs the propagation distance and, thus, the size of the vessel that the meter can be mounted on.

The frequency response of commercial probes varies from approximately 100 Hz to more the 1 kHz, where the highest frequency responses are obtained in the smaller probes. As a result, commercial probes have sufficient frequency response for most clinically or biomedically relevant flow regimes. Velocity and flow resolution is governed, in part, by the propagation length over which the flow is integrated and the resolution of the transit time measurement. The reader is referred to the meter manufacturers for detailed information about the operating specifications of particular meters. Reported uncertainties in transit time meters can be better than 15%. Actual uncertainties will depend on meter use, the experience of the operator, the meter calibration, and the acoustic properties of the

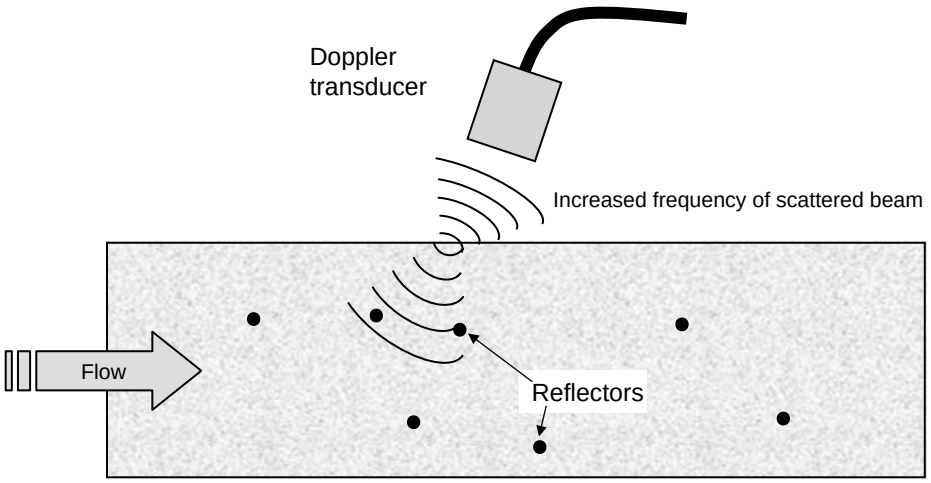


Figure 5. Schematic of the operation of a Doppler ultrasound probe.

fluid measured and how these properties differ from those of the calibration fluid.

Ultrasound Techniques – Doppler Volume Flow Meters. Flow can also be measured by ultrasound using the Doppler shift in a propagating sound wave generated by moving objects in a fluid flow (17,30). The primary difference is in the principal of operation. Devices using the Doppler approach measure the Doppler frequency shift of the transmitted beam due to the motion of particles encountered along the beam path, as illustrated in Fig. 5. The Doppler shift due to reflection of an incident wave by a moving particle is given by

$$F_D = 2F_o V \cos\theta / c \tag{7}$$

The shift frequency F_D (units of $L{\cdot}s^{-1}$) is linearly related to the component of the speed of a particle, V, in the direction of the wave propagation, the initial transmission frequency of the wave, F_o, and the speed of sound in the fluid, c. The reader is referred elsewhere in this Encyclopedic series and to the text by Weyman (30) for a detailed presentation of the Doppler technique.

The Doppler meter provides a direct measure of velocity that can be used to calculate the volume flow rate indirectly. Most biomedical applications involving volume flow measurement are performed on flow through a duct or vessel of given shape and size. Thus, the volume flow is the integral of the measured velocity profile across the vessel cross sectional area as defined in equation 1. The integral in equation 1 can be related to the average velocity $\bar{U}$ across the duct multiplied by the cross-sectional area of the duct (23). The Doppler technique then requires not only an estimate of the average velocity in the vessel but knowledge of the vessel area as well.

Commercially available Doppler volume flow meters, although not commonly used in biomedical applications, can be attached to a pipe or duct wall as with transit time meters. The commercial Doppler flow meters measure volume flow by integrating the measured Doppler shift frequency generated by particles throughout the flow. This integration is performed over a predefined path length in the flow, and is dependent on the number and type of particles, their size and distribution. The meter accuracy is also dependent on the velocity profile in the flow. Careful *in situ* calibrations are often needed to obtain accuracies of less than 10%. The Doppler meter has several disadvantages when compared with the transit time meter. It requires a fluid that contains a sufficient concentration of suspended particles to act as scattering sites on the incident ultrasound wave. These particles must be large enough to scatter the incident beam with a high intensity level but small enough to ensure that they follow the fluid flow (31). As a result of their dependence on flow profile, Doppler flow meters are not well-suited for measurement of flow in vessels with curvature or branching. Doppler flow meter measurements in blood rely on blood cells to act as scatterers. Clinical Doppler ultrasound machines, commonly used in echocardiography, can also be used to indirectly infer volume flow through the direct measure of the fluid velocity, and will be discussed later in the subsection on velocity measurements.

Invasive or Inline Volume Flow Measurement. Invasive or inline flow meters must be installed inline as a part of the piping or vessel network and involve hardware that is in contact with the fluid. These meters often have a nonnegligible pressure drop and may adversely interact with the flowing fluid. As a result, these meters are not often used in *in vivo* applications. Meters that fall in this category are variable area rotameters, turbine/paddle wheel meters, and vortex shedding meters. The primary advantage of these meters, is low cost and ease of use. However, these meters typically exhibit sensitivity both to fluid properties, which can be dependent on temperature and pressure, and to flow profile, White (23).

Variable area rotameters are simple measurement devices that can be used with a variety of liquids and gases. The flow of fluid through the meter raises a float in a tapered tube, as shown in Fig 6. The higher the float is raised, the larger the diameter of the tapered tube, increasing the cross-sectional area of the tube for passage of the fluid. As the flow rate increases, the float is elevated higher in the tube. The height of the float is directly proportional to the fluid flow rate. In liquid flow, the float is raised by the combination of the buoyancy of the liquid and the fluid drag on the float. Buoyancy is negligible in gaseous flows

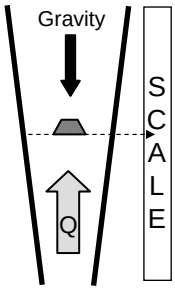


Figure 6. Schematic of a variable area flowmeter.

and the float moves in response to the drag by the gas flow. For constant flow, the float reaches a stable position in the tube when the upward force generated by the flowing fluid equals the downward force of gravity. The float will move to a new equilibrium position in response to a change in flow. These meters must be installed vertically to operate properly, although spring-loaded meters have been designed to eliminate the need for gravity and permit installation in other orientations.

Rotameters are designed and calibrated for the type of fluid (fluid properties such as viscosity and density) and flow range expected. They do not function properly in nonNewtonian fluids. Use of a meter with a fluid different from that the meter was calibrated for, or with, a fluid at a different temperature or pressure requires a correction to the meter reading. Meter uncertainty and repeatability will vary with operation of the meter, but can approach a few percent with proper operation.

Turbine and paddle wheel meters measure volume flow rate through the rotation of a vaned rotor in contact with the flowing fluid. These meters are intrusive flow measurement devices that produce higher pressure drops than others in the class of invasive flow probes. The turbine meter has a turbine mounted across the pipe or duct in full contact with the flow, whereas the paddle wheel meter has a vaned wheel mounted on the side of the duct with half of the wheel in contact with the flow. Accuracy and repeatability is best with the turbine meter, but the pressure drop is also higher. An ac voltage is induced in a magnetic pickup coil as the turbine or paddle wheel rotates. Each pulse in the ac signal represents the passage of one blade of the turbine. As the turbine fills the flow path, a pulse represents a distinct volume of fluid being displaced between two turbine blades. This design provides an accurate direct measure of the volume flow rate.

Flowmeter selection must take into account the type of fluid, the flow rate range under study, and the acceptablepressure drop for a given flow application. In general, these meters have a sensitivity to flow profile, and the pressure drop is dependent on the fluid properties. The meters incorporate moving parts within the flow and thus use bearings to ensure smooth operation. Bearing wear will affect the meter accuracy and must be monitored for the life of the meter. The paddle wheel meter operates in a similar manner as the turbine meter. The primary difference is that only part of the rotor is in contact with the fluid and, thus, as the paddle wheel meter is more sensitive to flow profile it has a smaller pressure drop. Installation of these

Figure 7. Vortex shedding from a circular cylinder. Picture from White (23), courtesy of the U.S. Naval Research Laboratory.

meters often involves a specified number of straight pipe sections upstream and downstream of the meter and may also require installation of a flow straightener inline upstream of the meter.

Vortex meters operate on the principal of Strouhal shedding. Separating flow over an obstruction such as a cylinder or sharp-edged bar results in a pulsatile or oscillatory pattern as shown in Fig. 7. The shedding frequency, ω (units of $\mathrm{L}\cdot\mathrm{s}^{-1}$), is related to the fluid velocity by

$$\omega = V \cdot S_\mathrm{t} / L \tag{8}$$

where S_t is the Strouhal number, which is a nondimensional number that is a function of the flow Reynolds number and geometry of the obstruction, and L is a characteristic length scale (23). For a cylinder, L is the diameter of the cylinder. The Reynolds number is a dimensionless number that is the ratio of inertial to viscous forces in the flow, and is defined as

$$\mathrm{Re} = V \cdot L / \upsilon \tag{9}$$

Here, V and L are defined as in Eq. 8 and υ is the kinematic viscosity of the fluid with units of length squared per time.

The vortex meter is an intrusive meter that has a "shedder bar" installed across the diameter of the duct. The flow separates off this bar and generates a shedding frequency that is transmitted through the bar to a piezoelectric sensor attached to the bar. The meter is sensitive to flow and fluid properties, and rated accuracy and pressure drop depend on application.

Volume flow rate can also be estimated through an indirect measure of the velocity profile in the flow and the use of Eq. 1. A number of instruments are available that measure fluid velocity in biomedical engineering applications. Doppler ultrasound and MR phase velocity encoding are standard clinical techniques used to measure velocity of a flowing fluid noninvasively. *In vitro* systems that are commonly used to measure fluid velocity, in addition to Doppler and MR, are laser Doppler velocimetry (LDV), particle image velocimetry (PIV), and thermal anemometry. Besides an estimate of volume flow rate, fluid velocity measurement can provide quantification of flow profiles, fluid wall shear, pressure gradient, and flow

mixing. The following section summarizes velocity measurement techniques commonly used in biomedical/bioengineering applications.

Velocity Measurements

Thermal Anemometry. Thermal anemometry is an invasive technique used to measure fluid velocity or wall shear. A heated element is inserted into the flow, and specialized electronic circuitry is used to measure the rate of change in heat input into the element in response to changes in the flow field (32). Thermal anemometry, when used properly, is characterized by high accuracy, low noise, and high spatial and temporal resolution. Its main disadvantages are sensitivity to changes in fluid temperature and properties, particulates and bubbles suspended in a fluid, nonlinear response to velocity, and its invasive characteristics (geometry, size, vibration, etc.).

Hot-film anemometry has been used to measure blood velocity and wall shear in biomedical and bioengineering applications both *in vitro* and *in vivo*. Arterial blood flow velocity measurements were performed by Nerem et al. (33,34) in horses and by Falsetti et al. (35) in dogs. Tarbell et al. (36) used flush-mounted hot films to measure wall shear in the abdominal aorta of a dog. *In vitro* applications of hot-film anemometry include the measurement of wall shear in an LVAD device (37), and the *in vitro* measurement of flow in rigid models of arterial bifurcations by Batten and Nerem (38). Although rarely used in biomedical applications now, we will briefly present hot-film anemometry here for completeness. The reader is referred to the text by Bruun (39) and the symposium proceedings by Stock (40) for a detailed presentation of thermal anemometry.

Thermal anemometry operates on the principal of convective cooling of a heated element. A thin wire or coated quartz element, mounted between supports, is heated and exposed to a fluid flow, as shown in Fig. 8. The element is heated by passing a current through the wire or element. The amount of heat generated is proportional to the current, I (A), and the resistance of the element, R (Ω), by I^2R. The element is convectively cooled by flow until an equilibrium state is reached between electrical heating of the element and flow-induced convective cooling, $DE/Dt = W + H$, where E is the thermal energy stored in the element, W is the heat added by joule heating, and H is the heat loss to the environment by cooling. At equilibrium, $DE/Dt = 0$ and $W = H$.

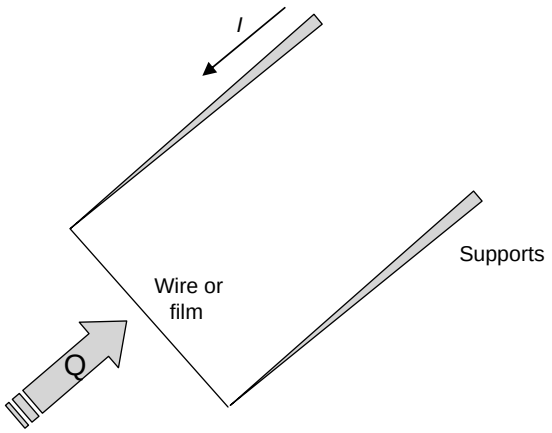


Figure 8. Illustration of a hot wire or film element.

Changes in flow velocity will increase or decrease the amount of convective cooling and produce changes in the equilibrium state of the element or the temperature of the element. Commercial anemometers employ a four-arm electronic bridge circuit to maintain constant element temperature, current, or voltage in response to changes in convective cooling. As convective cooling changes, the anemometer output, current, or voltage changes in response to maintaining the desired set condition. This equilibrium condition assumes that radiation losses are small, conduction to supports is small, temperature is uniform over length of sensor, velocity impinges normally on the sensor, velocity is uniform over sensor length and is small compared with the sonic speed, and finally, the fluid temperature and density are constant.

An energy balance between convective heat cooling and joule heating can be performed to derive a set of governing equations that relate input current, I, to convective velocity, V. The "King's law" is the classic result of this energy balance:

$$I^2R^2 = V_o^2 = (T_w - T_a)(A + B \cdot V^n) \tag{10}$$

where V_o is the measured voltage drop in response to a velocity, V, and T_w and T_a are the wire and ambient fluid temperatures (°C), respectively. The coefficients, A and B, and power, n, are determined through careful calibration over the velocity and temperature range that will be observed experimentally. In the event of a three-component flow, the probe must be calibrated for yaw and pitch angles between the probe and the flow velocity vector, and the velocity, V, in Eq. 10 must be replace by a term related to the velocity vector magnitude. Bridge-type circuits are also prone to stable and unstable performance under unsteady operation. Thus, the overall calibration of a hot-wire/film system must involve the element and electronics as a system and must also involve dynamic calibrations to characterize the frequency response of the system.

Hot-wire/film probes come in a variety of sizes, shapes, and configurations. Probes are manufactured from platinum, gold-plated tungsten, or nickel-plated quartz, and come in single-element or multielement configurations for measurement in a variety of flow conditions. The reader is referred to the hot-wire/film manufacturers for a complete summary of probe types and conditions for use. In general, wire probes are used when possible due to lower cost, improved frequency response, and ease of repair. However, wire probes are more fragile compared with film-type probes and are usually restricted to air flows. Film probes are used in rough environments, such as liquid flows.

The following considerations should be addressed to ensure accurate measurements when using thermal anemometry. The type of flow should be assessed for velocity range, flow scales, and fluid properties (clean gas or particle contaminated liquid, etc.). The flow characteristics will define the right probe, anemometer configuration, and

A/D setup to use. Perform appropriate calibrations with complete hardware setup. Perform the experiment and post calibrations to ensure that the anemometer/probe calibration has not changed.

Doppler Ultrasound And Magnetic Resonance Flow Mapping. The focus of this subsection is to introduce the concept of Doppler ultrasound and MR flow mapping for local velocity measurement. Flow measurement using clinical Doppler can suffer from the same limitations as the small Doppler meter, but has several advantages over these small meters. Most ultrasound machines can operate in continuous wave or pulsed Doppler modes of operation; see Weyman (30) for a more detailed discussion of the modes of operation.

Pulsed Doppler ultrasound offers the advantage of localizing a velocity measurement within a flow and can be used to measure the velocity profile across a vessel or lesion. This information coupled with echocardiographic imaging of the geometry can be used to calculate the flow rate from Eq. 1. Unfortunately, the implementation of this technique is not straightforward due to limitations in resolution, velocity aliasing, and the need to know the relative angle between the transmitted ultrasound beam and the local flow.

Velocity aliasing in pulsed-mode Doppler occurs because the signal can only be sampled once per pulse transmission (e.g., the pulse repetition frequency). Frequency aliasing, or the ambiguous characterization of a waveform, occurs in signal processing when a waveform is sampled at less than one-half of its fundamental frequency, referred to as the Nyquist condition in signal processing. In a pulsed Doppler system, velocity aliasing will occur when the Doppler shift of the moving particles exceeds half of the pulse repetition frequency. As the pulse repetition frequency is a function of the depth at which a sample is measured, the alias velocity will vary with imaging depth. Increasing the imaging depth will lower the velocity at which aliasing will occur. Continuous wave Doppler signals are typically digitized at higher sampling frequencies, limited by the Nyquist frequency associated with the frequency of the transmission wave. Velocities observed clinically produce Doppler shifts that are generally lower than sampling frequencies used in continuous wave Doppler. As a result, velocity alias is not usually observed with continuous mode Doppler in clinical applications.

Aliased signals that are processed by the measuring system are not directly proportional to the Doppler shift generated by the velocity of the particle but can be related to the Doppler shift. Undersampling a wave underestimates the frequency and produces a phase shift. Most clinical Doppler machines use the frequency and phase of the sampled wave to infer velocity magnitude and direction. Velocity alias will produce a lower velocity magnitude with increasing Doppler shift above the alias limit. As frequency is, by definition, positive and Doppler machines use the signal phase to determine direction, the measured frequency is usually reported as a negative velocity above the alias limit, which is often displayed as an increasing positive velocity magnitude with increasing Doppler shift up to the alias limit. Further increases in the Doppler shift (particle velocity) result in a sign change at the velocity magnitude of the alias velocity with a continued decrease in velocity with increasing Doppler shift. Velocity alias can be reduced or eliminated by frequency unwrapping and baseline shifting, or through the careful selection of machine settings during data acquisition.

Frequency unwrapping is simply correcting the reported aliased velocity by a factor that is related to the alias velocity limit and the magnitude of the reported negative velocity. This correction is, roughly speaking, adding the relative difference in magnitude between the measured aliased velocity and the velocity alias limit to the velocity alias limit. This method of addressing velocity alias is often accomplished by baseline shifting in commercial Doppler machines. In baseline shifting, the phase angle at which a negative velocity is defined is shifted with the effect of a relative shift in the reported alias velocity. Baseline shifting or frequency unwrapping does not eliminate velocity alias but provides a correction to extend the measurement to higher frequencies.

Velocity alias can be "eliminated" by reducing the Doppler frequency of moving particles and thereby shifting the measurable range below the alias limit, which can be accomplished by reducing the carrier frequency of the ultrasound wave that will, in turn, reduce the Doppler frequency shift induced by a moving particle and increase the maximum velocity that can be recorded before reaching the Nyquist limit. Alternatively, the Doppler shift frequency can be reduced by increasing the angle between the propagation of the ultrasound wave and the velocity vector, which reduces the magnitude of the Doppler shift frequency by the cosine of this included angle. Angle correction has limitations in that the flow direction must be known and the uncertainty in the correction increases with increasing included angle. As color flow mappers operate in the pulsed Doppler mode, they are subject to velocity alias. Color flow mappers indicate velocity direction by a color series (for example, shades of red or blue). Velocity alias is displayed as a change in a color series from red-to-blue or blue-to-red.

The clinical measurement of many velocity ensembles across a vessel and the integration across the vessel geometry can be time-consuming and problematic in pulsatile flow through a compliant duct. Furthermore, lesions are often complex in shape and cannot be adequately defined by echo. Doppler echocardiographers and scientists have exploited the physics of fluid flow to develop diagnostic tools that complement these capabilities of commercial Doppler systems. The text by Otto (41) provides an excellent review of these diagnostic tools. Techniques such as the PISA or proximal flow convergence use the capability of color Doppler flow mapping machines to estimate volume flow through an orifice, such as a heart valve. Figure 9 illustrates the concept of the proximal flow convergence method.

The flow accelerating toward a small circular orifice will increase in velocity V_a, until a maximum velocity at the orifice V_j is reached. This acceleration occurs in a symmetric pattern around the orifice and is characterized by hemispheres of constant velocity. As the orifice is approached, the velocity increases and the radius of the

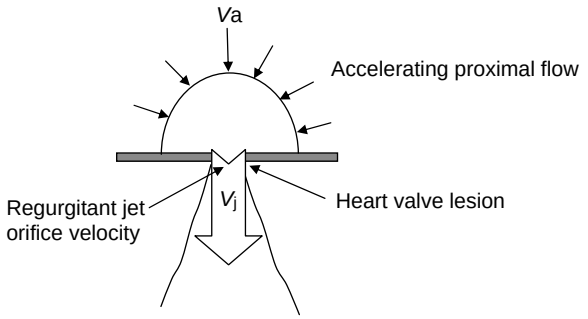


Figure 9. Illustration of the proximal isovelocity surface area (PISA) concept.

hemisphere decreases. The regurgitant flow through the orifice can then be calculated as:

$$Q = (2\pi r^2)V_a \tag{11}$$

The combined imaging and Doppler characteristics of color Doppler flow mapping are exploited in the PISA approach. The location of the alias velocity in the flow map provides a measure of V_a, which is then coupled with a measure of the radial location from the orifice using the imaging capability of color flow mapping. As flow is velocity times area, the hemispheric assumption provides a shell with a surface area of $2\pi r^2$ that velocity Va is passing through. Figure 10 illustrates the concept of PISA with a color Doppler flow image of valvular regurgitation in a patient.

The PISA approach assumes a round orifice with a hemispherical acceleration zone. In clinical applications, orifices are rarely round and the acceleration zone is not hemispherical with the result of under or over estimation of the flow rate depending on what radial velocity contour is used in Eq. 11. The semielliptic method is one approach at considering nonhemispheric geometries in an attempt to correct for errors associated with the PISA technique.

The combination of continuous wave and pulsed Doppler ultrasound is exploited in the turbulent jet decay method of measuring flow through a stenotic lesion or a regurgitant valve. While continuous wave Doppler does not suffer from velocity alias as does pulsed Doppler, it cannot provide spatial localization of the velocity. The turbulent

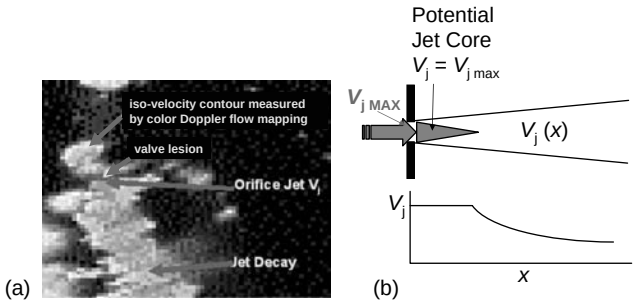


Figure 10. (a) Color Doppler flow map image of the proximal isovelocity surface area (PISA) in valvular regurgitation. (b) Illustration of the jet decay downstream of an orifice. The parameter V is the jet velocity and x is measured from the orifice.

jet decay method uses continuous wave Doppler to measure the high velocity at the lesion orifice and then uses pulsed Doppler to measure the velocity decay at specified location downstream of the orifice. Turbulent jet theory can be used to relate the flow rate of the turbulent jet to the decay of the jet velocity downstream of the orifice, as in Eq. 12

$$Q = (\pi V_m^2 x^2)/160\, V_j \tag{12}$$

The velocity V_m is measured by pulsed Doppler at location x measured from the jet orifice, whereas the orifice velocity, V_j, is measured by continuous wave Doppler; Fig. 10b illustrates this decay phenomenon. This equation is valid for round jets and has been extended to jets with other geometries by Cape et al. (42) with the resulting change to Eq. 12:

$$Q = (V_m^2 H_x)/5.78\, V_j \tag{13}$$

where H is the width of the jet measured by color Doppler.

Doppler velocity measurements are also used to estimate pressure gradients in various cardiovascular applications. The Bernoulli equation can be used to estimate the pressure drop across a stenotic lesion or through a valve by measuring the velocity both upstream and downstream of the lesion or valve. The Bernoulli equation is

$$\Delta P = (P_1 - P_2) = 1/2{\cdot}\rho(V_2^2 - V_1^2)$$

where position 1 is often measured upstream of the lesion and position 2 is at the lesion or downstream. In this equation, the pressure drop ΔP has units of Pascal's (Pa). A Pascal is a Newton (N) per square meter, where a Newton has units of mass (kg) times length (meter) per time squared. Bioengineering and biomedical applications often use the units of millimeters of mercury (mmHg) in defining a pressure value. A mmHg is related to a Pa by the conversion; 1 mmHg = 133.32 Pa.

Magnetic resonance flow mapping has the advantage over Doppler that it can measure the full three-component velocity field over a volume region (43–45), which eliminates the uncertainty in flow direction and enables the use of standard fluid dynamic control volume analysis. The advantages of MR flow mapping come at the cost of long imaging times and increased sensitivity to motion artifacts in *in vivo* applications, where phase locking to the heart rate or breathing cycle can increase complexity.

The velocity of moving tissue can be detected by a time-of-flight technique (46) and by phase velocity encoding (47,48). The time-of-flight method tracks a selected number of protons in a plane and measures the displacement of the protons over a time interval defined by the imaging rate. *In vivo* (49) and phantom (50) studies have shown that the time-of-flight technique is capable of accurate velocity measurement up to velocities at least as high as 0.5 $m{\cdot}s^{-1}$. However, the time-of-flight method requires a straight length of vessel on the order of several centimeters for accurate velocity estimation. This requirement reduces its usability in most *in vivo* applications. The phase velocity encoding method has become the preferred technique in most clinical applications.

Phase velocity encoding directly relates the local velocity of nuclei to the induced phase shift in an imaging voxel.

Properly defined bipolar magnetic field gradients are produced in the direction of interest for velocity measurement. The velocity of hydrogen nuclei are then encoded into the phase of the detected signal (51). Chatzimavroudis et al. (52) and Firmin et al. (53) provide a discussion of the phase encoding technique with an assessment of its accuracy and limitations for flow measurement.

The technique uses two image acquisitions to provide velocity compensation and velocity encoding. Velocity information in each voxel is obtained by a voxel-by-voxel subtraction of the two images with respect to the phase of the signal. Like Doppler ultrasound, phase velocity encoding can suffer from aliasing effects, alignment error, and limits in spatial and temporal resolution. Velocity estimation using phase shift measurement is limited to a maximum range of phase of 2π radians without ambiguity or aliasing. However, the estimation of the phase shift using phase subtraction between two images reduces that sensitivity to this problem. Studies have been conducted that show MR phase velocity encoding can measure velocities covering the complete physiologic range up to several meters per second (54). Misalignment of the flow direction with the encoding direction will produce a negative bias in the measured flow where the measured velocity will be lower than the true velocity. Like Doppler, this bias follows a cosine behavior where $V_{\text{meas}} = V_{\text{act}} \cos(\theta)$, where V_{meas} is the measured velocity, V_{act} is the actual velocity, and θ is the misalignment angle. This error is typically less than 1% in most applications.

The size of a voxel and the sampling capabilities of the hardware characterize the spatial and temporal resolution of the system. Spatial resolution affects the size of a flow structure that can be measured without spatially filtering or averaging the structure or velocity measurement. Spatial velocity gradients that are small relative to the voxel size will not be adequately resolved and will be averaged over the voxel volume (55). In addition, rapidly varying velocity fluctuations in time will produce a similar low pass frequency filtering effect if these fluctuations occur with a time scale that is much smaller than the imaging time scale of the measurements. Turbulent flow can produce spatial and temporal scales that could be small relative to the imaging characteristics and can result in what is referred to as signal loss in the image (56). Stenotic lesions and valvular regurgitation are clinical examples where turbulent flow can occur with spatial and temporal scales that could compromise measurement accuracy.

Phase velocity encoding has the drawback of fairly long imaging or magnet residence times, which is particularly true for three-component velocity mapping. Although this may be acceptable for *in vitro* testing with flow loop phantoms, it can present problems and concerns with clinical measurements. Patients can be exposed to long time intervals in the magnetic with the associated problems of patient comfort and care. *In vivo* velocity measurements are often phase-locked with cardiac cycle or breathing rhythm. Long imaging times can increase potential for measurement error due to patient movement and variability in the cardiac cycle or breathing rhythm, which can cause noise in the phase-averaged, three-component velocity measurements. Research, in recent years, has focused on hardware and software improvements to increase spatial resolution and reduce imaging time [see, e.g., Zhang et al. (57)].

Magnetic resonance phase velocity encoding provides coupled 3D geometric imaging using traditional MR imaging methods with three-component velocity information. This coupled database provides a powerful diagnostic tool for blood flow analysis and has been used extensively in *in vitro* and clinical applications. Jin et al. (6) used this coupled imaging flow mapping capability to study the effects of wall motion and compliance on flow patterns in the ascending aorta. Standard imaging was used to measure aortic wall movement and the range of lumen area and diameter change over the cardiac cycle. This aortic wall motion was phase-matched with phase velocity encoded axial velocity distributions in the ascending aorta. Similar to the PISA approach in Doppler ultrasound, a control volume approach using phase velocity encoded MR velocities can be applied to the assessment of valvular regurgitation (58,59). The control volume approach is illustrated in Fig. 11.

Laser Doppler Velocimetry. The Doppler shift of laser light scattered by particles or cells in a fluid is the basis of laser Doppler velocimetry (LDV). Detailed presentations of the LDV technique are provided in the works by Drain (60) and Durst et al. (61). The scattered radiation, from a laser beam directed at moving particles in a fluid, has a Doppler-shifted frequency defined as:

$$f_{\text{D}} \sim (1 - V/C_1)\, f' \tag{14}$$

where C_1 is the speed of light in a vacuum, V is the particle velocity, and f' is incident light frequency. The Doppler-shifted frequency is very small relative to the frequency of

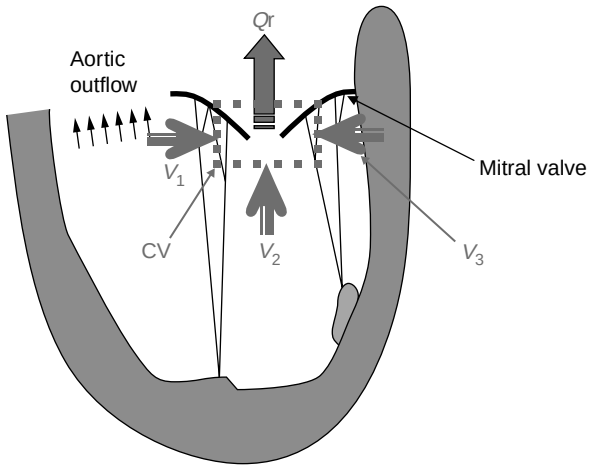


Figure 11. Illustration of the control volume method in MR phase velocity assessment of valvular regurgitation. The control volume (CV) is the heavy dotted line box around the mitral regurgitant orifice. The box edges are usually selected to correspond with rows and columns in the MR image. V_{i} represents the three-component velocities measured with MR through the i faces of the box. Faces 4 and 5 are in the plane of the image at $\pm Z$ offsets from the plane of the image. The regurgitant flow Q_{r} is the sum of the $V_{\text{i}}A_{\text{i}}$ on each face.

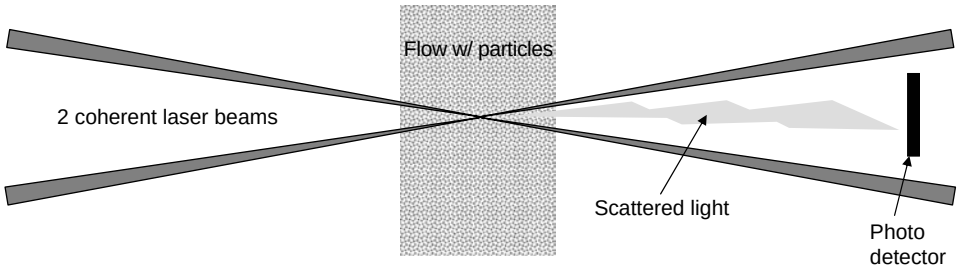


Figure 12. Illustration of the dual-beam or fringe mode LDV setup.

light and, thus, dual-beam or fringe mode LDV is the system configuration of choice. The dual-beam mode of operation is schematically shown in Fig. 12. In fringe mode LDV, two coherent laser beams of the same wavelength or frequency are focused to a common point (control volume) in the flow field. The scattered light from a particle moving through the control volume is received by a photodetector.

The crossing of two coherent, collimated laser beams forms interference fringes as the propagating light waves constructively and destructively interfere with one another. This interference creates a series of light and dark bands with spacing, d_f, of

$$d_f = \lambda/2\sin(\kappa) \tag{15}$$

The number of fringes, N_{FR}, in the measurement volume is given by

$$N_{FR} = 1.27d/D_e^{-2} \tag{16}$$

where d is the spacing between the two parallel laser beams before the focusing lens and D_e^{-2} is the beam diameter before the lens. Figure 13 illustrates the probe geometry generated by the intersection of two focused coherent laser beams with a common wavelength.

The spatial resolution of a dual-beam system is affected by the distribution of the light intensity at the intersection of the two focused beams, referred to as the probe or measurement volume. When the laser is operating in the TEM_{oo} mode, the laser cavity sustains a purely longitudinal standing wave oscillation along its axis with no transverse modes. The laser output has an axisymmetric intensity profile that is approximately a Gaussian function

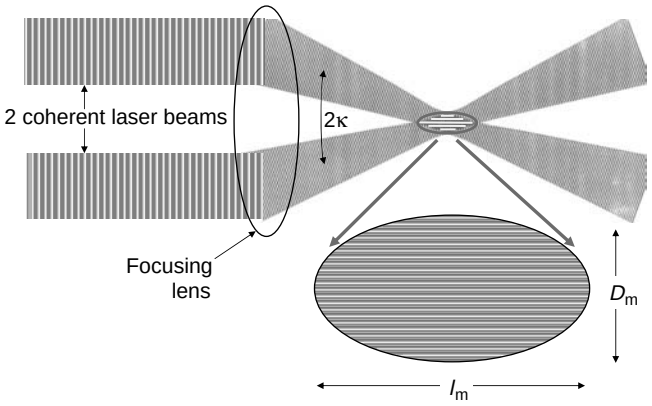


Figure 13. Illustration of the measurement volume generated in fringe mode LDV.

of radial distance from the axis. In the far field, the beam divergence is small enough to appear as a spherical wave from a point source located at the front of the lens. A lens is used to focus the beam into a converging spherical wave. The radius of this wave decreases until the focal point of the lens is reached. At the focal point, the beam has almost a constant radius and planar behavior. The beam is focused to its minimum diameter or focal waist, d_e^{-2}, and is defined as:

$$d_e^{-2} = (4\lambda f)/(\pi D_e^{-2})$$

where λ is the wavelength of the laser beam and f is the focal length of the lens. A single pair of laser beams generates an ellipsoidal geometry having dimensions of major axis l_m and minor axis d_m given by

$$l_m = d_e^{-2}/\sin(\kappa) \text{ and } d_m = d_e^{-2}/\cos(\kappa) \tag{17}$$

where κ is the half angle between the two laser beams, as illustrated in Fig. 13.

The particle velocity is calculated by the fluctuating light intensity collected by the receiver as the particle passes through the measurement volume and scatters light from the fringes. The intensity change of the scattered light from the light and dark fringes is converted into an electrical signal by a photomultiplier tube (PMT). The electrical signal represents an amplitude-modulated sine wave, with frequency proportional to the Doppler frequency shift (f_D) of the particle traveling through the measurement volume. The particle velocity is then equal to the Doppler frequency multiplied by the fringe spacing. In a two-beam LDV system, the measured velocity component is in the plane of the two laser beams and in the direction perpendicular to the long axis of the measurement volume.

Coherent laser beams with the same frequency produce stationary fringes. A particle moving in either direction across the fringes will produce the same frequency independent of sign, such that a stationary fringe system can only determine the magnitude of the velocity, not the direction. To avoid this directional ambiguity, one of the laser beams of a beam pair is shifted to a different frequency, using a Bragg cell, to provide a moving fringe pattern. One laser beam from each beam pair passes through a transparent medium such as glass, in which acoustic waves, generated by a piezoelectric transducer, are traveling. If the angle between the laser beam and the acoustic waves satisfies the Bragg condition, reflections from successive acoustic wave fronts reinforce the laser beam. The beam exits at a higher frequency and a prism

directs the beam to its original direction. The Bragg shift causes the fringes in the probe volume to move at a constant speed in either the positive or negative direction relative to the flow. The measured frequency by the PMT and processor is then the sum or difference of the Bragg cell frequency (typically 40 MHz) and the Doppler shift frequency. This measured frequency is then downmixed with a frequency that is a percent of the Bragg frequency (called the shift frequency) producing a frequency that has a zero shifted to a higher baseline frequency (usually on the order of several MHz). This zero shift eliminates directional ambiguity in LDV signal processing.

Laser Doppler velocimetry has excellent spatial and temporal frequency response compared with most other measurement systems. It is considered a gold standard measurement technique in biomedical applications and is the noninvasive measurement system of choice for turbulence measurements. Two disadvantages of LDV worth noting are (*1*) LDV noise and (*2*) velocity bias. The LDV is noisy when compared with other turbulence measurement systems, such as thermal anemometry, due to the use of photomultiplier tubes. These optical detectors, used for their sensitivity and high frequency response, suffer from higher noise floors than other photo detectors.

Velocity bias is a result of the random sampling characteristics of LDV. As a velocity ensemble is randomly recorded when a particle passes through a probe volume, the statistics of the measured velocity ensembles are not independent of the particle velocity. A greater number of higher speed particles will cross the measurement volume over a specified time than will slower speed particles. Standard ensemble averaging will produce mean velocity estimates that are biased toward higher velocities. This velocity bias can have a significant impact on the velocity statistics, particularly in turbulent flow. In addition to velocity bias, two other biases may occur, fringe bias and gradient bias.

Fringe bias is an error that is minimized by frequency shifting. This type of bias is created by not having enough fringe crossings to satisfy processor validation criteria when calculating a velocity, which occurs when a particle crosses the edge of the probe volume or if the particle velocity is nearly parallel to the fringes. Thus, velocity ensemble averages weight velocities from particles traveling near the center of the measurement volume or those particles that cross more fringes then others. By frequency shifting with a fringe velocity at least two times greater than the flow velocity, particles moving parallel to the fringes can cross the minimum number of fringes for validation by a processor.

Gradient bias results from a nonnegligible mean gradient across the probe volume. This bias depends on the fluid flow and the measurement volume dimensions. The mean velocity and the odd order moments are the only statistics affected by gradient bias. In general, LDV-transmitting optics are chosen to provide as small a measurement volume as possible to increase spatial resolution and reduce gradient bias. As the LDV measurement volume is longer than it is wide, experiments should be designed to ensure that the LDV optical setup is oriented to position the measurement volume diameter in the direction of the maximum gradients in the flow.

Several post processing techniques have been developed to reduce velocity bias. The recommended technique is to use a transit time weighting when computing the velocity statistics. This transit time weighting approximates the ensemble average as a time average. The reader is referred to Edwards (62) for a detailed discussion of the transit time technique and its implementation in LDV data processing.

Multiple pairs of laser beams with different wavelength (color) or polarization can be used to produce a multicomponent velocity measuring system. Two or three laser beam pairs can be focused to the same point in the flow. Each beam pair can then be used to independently measure a different component of the velocity vector. As more than one particle can pass through a measurement volume at one time, it is possible to get valid velocity component estimates from different particles. The ellipsoidal geometry of the measurement volumes exaggerates this problem. As a result, LDV data are often processed in one of two modes, random and coincident.

Random mode processing records every valid velocity ensemble as it arrives at the measurement volume, which can generate uneven sample distributions in the different velocity components or LDV channels being measured. Random mode processing has a negligible impact on mean velocity statistics but can be detrimental to turbulence estimates. Coincident mode processing uses hardware or software filters to ensure that each velocity component ensemble is measured from the same particle. Filters are used to ensure that the Doppler bursts measured on the different LDV channels overlap in time. Bursts generated by one particle should be measured on each channel with perfect overlap. Time window filters are used to reject bursts that do not occur within a window defined by a percentage of the burst length. The effect of coincident mode processing is usually a reduction in the overall data rate by a factor of at least two but provides the necessary data quality for turbulence measurements.

Laser Doppler velocimetry is primarily an *in vitro* tool, although systems have been developed for blood flow measurement (17,63). Blood is a multiphase fluid composed of a carrier liquid, plasma, and a variety of cells and other biochemical components. Plasma is optically clear to the wavelengths of light used in LDV. The optical opacity of blood is due to the high concentration of cells, in particular red cells. On the microscopic level, however, blood can transmit light over a short distance due to the plasma carrier fluid. Clinical-style probes have been developed to measure the velocity of blood cells in blood using catheter-type insertion into vessels of suitable size or through transcutaneous measurement of capillary flow below the skin. These *in vivo* systems are designed with very short focal length transmitting lenses providing a small measurement volume located a very short distance from the transmitting lens. Laser light is propagated through the plasma and focused a few millimeters from the probe tip. Blood cells act as particles in the fluid and scatter light that is collected by the transmitting lens and directed to a PMT system for recording of the Doppler bursts. Manning et al. (64) and Ellis et al. (65) have used LDV to measure the velocity fields around mechanical heart valves in *in vitro* studies. Figure 14 shows the measured velocity

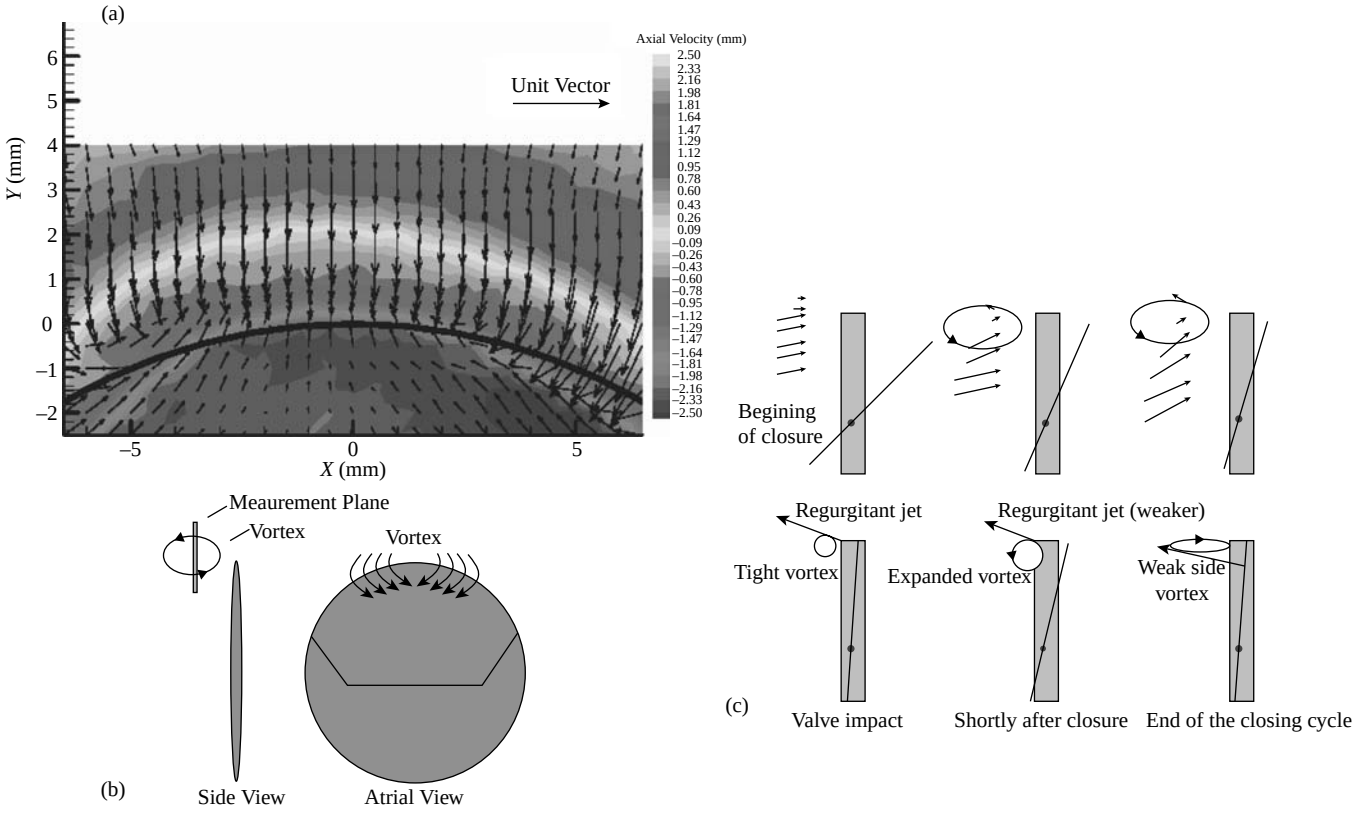


Figure 14. 3D phase-averaged velocity map of major orifice regurgitant flow in Bjork–Shiley monostrut valve. (a) 3 mm from valve housing, 4 ms after impact. (b) Illustration of measurement plane and vortex flow pattern. (c) Flow field schematic during valve closure determined from multicomponent LDV measurements (64).

distributions associated with impact of a Bjork–Shiley monostrut valve (64).

Particle Image Velocimetry and Particle Tracking Velocimetry. Particle image velocimetry (PIV) and particle tracking velocimetry (PTV) have been applied in fluid flow applications for over a decade. They are noninvasive experimental techniques that provides a quantitative, instantaneous velocity vector field with good spatial resolution, an appealing feature when studying complex, time-dependent fluid flows that can occur in biomedical applications. The method allows the quantitative visualization of instantaneous flow structures over a spatial region, as opposed to a point measurement like LDV, which can provide insight into the flow physics. The two techniques, PIV and PTV, differ in the way particle displacements are measured. Particle tracking follows and tracks the motion of individual particles whereas PIV measures the displacement of groups of particles within the flow. Particle tracking velocimetry, although not commonly used, is a subset of the more common PIV technique and is still used in specific applications of PIV. Raffel et al. (66) provide a comprehensive discussion of the PIV and PTV technique with a detailed presentation of the physics, processing tools, capabilities, and errors.

The instantaneous velocity field is computed by measuring an instantaneous particle displacement field over a specified, finite time interval. Laser-based PIV and PTV are noninvasive velocity measurement tools that require good optical access to the flow field. As a result, they are essentially *in vitro* tools (8,67) that are of limited use *in vivo*. Figure 15 shows an example of the use of PIV in a bioengineering application of flow through one chamber of an artificial heart (68). X-ray-based PTV systems are being developed and will be capable of *in vivo* use. In this section, the authors will focus on PIV; however, system concepts (seeding, acquisition, processing, noise, and errors) would be applicable to some degree to systems like X-ray PTV (69).

Particle image velocimetry uses a double-pulsed light source, usually a laser, to illuminate a thin sheet in the flow field. Particles suspended in the fluid scatter light during each pulse, and this scattered light is recorded on a digital camera. The optimal setup has the recording device located 90° to the illumination sheet. Figure 16 illustrates the typical PIV setup. (66)

Two lasers with coincident beam paths are used to illuminate a desired plane of the flow by incorporating optics to produce thin laser sheets. During image acquisition, the two lasers are pulsed with the specified time separation (typically between 1 and 1000 μs). A trigger system, referred to as a synchronizer, controls the firing of the two lasers relative to the shuttering of a CCD camera. The camera, usually placed orthogonal to the laser sheet, collects the light scattered by tracer particles in the flow and records an image. The synchronizer, used in cross-correlation-based PIV systems, delays the firing of the first

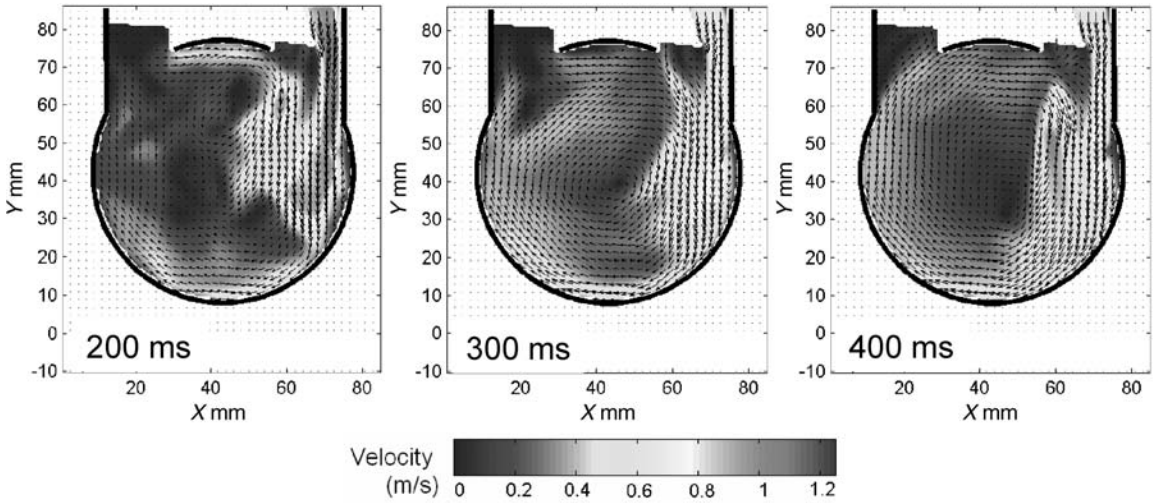


Figure 15. Phase-average velocity maps from mid to late diastole for a prototype artificial heart ventricular chamber (time reference is from the onset of diastole, 4.7 L·min^{-1} CO, 75 bpm) (68).

laser such that the camera shutter is centered between the firing of the two lasers. This synchronization technique is called frame straddling and produces two sequential images of each laser beam pulse. Although the time between successive camera frames may be much larger than the time duration between laser pulses, the two images of the particle field created are separated by the specified time interval between the two laser pulses.

A discussion of PIV must begin with a brief introduction of the terminology commonly used. Figure 17 provides a schematic representation of geometric imaging (66). The "light plane" is the "volume" of the fluid illuminated by the light sheet. The "image plane" is the image from the light plane captured on the CCD sensor. It is important to note that the light plane is a 3D space or volume, whereas the image plane is a 2D space or "surface." The subvolume selected from the light plane for cross correlation is called the interrogation subvolume. The corresponding location of this interrogation volume captured on the image plane is called the interrogation subregion. Please note that the displacement vectors in an interrogation volume are three-component vectors, whereas those in an interrogation area are two-component vectors. "Particle" is the physical particle suspended in the fluid volume. "Particle image" is the image of each particle in the image plane. Particle density, intensity, and pair refer to particle properties, whereas image density, intensity, and pair refer to particle image properties.

Most commercial systems use a cross-correlation-based image processing technique to compute the particle displacement. Images are subdivided into small interrogation regions, typically a fraction of the overall image size, and

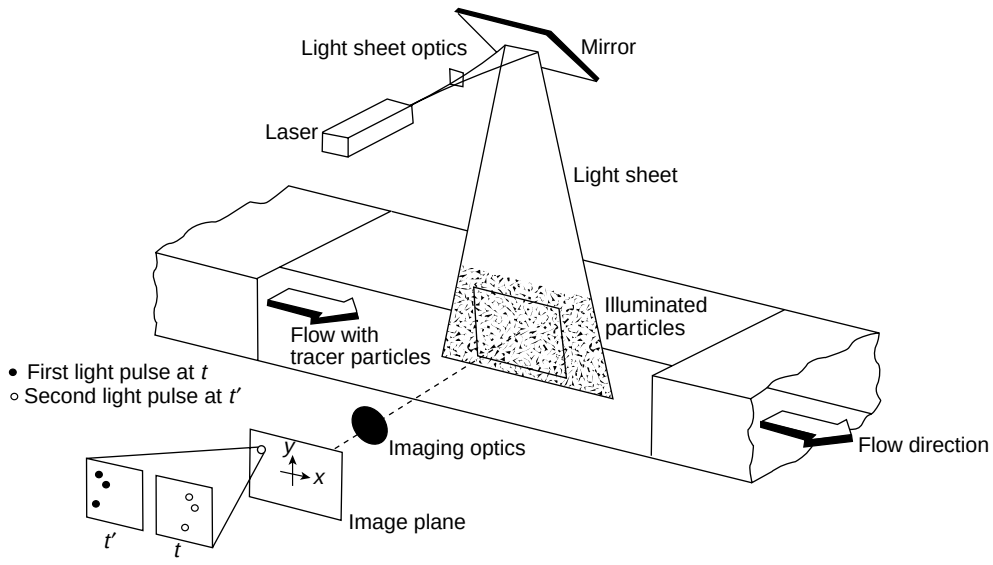


Figure 16. Schematic of a PIV setup (66). (With kind permission of Springer Science and Business Media.)

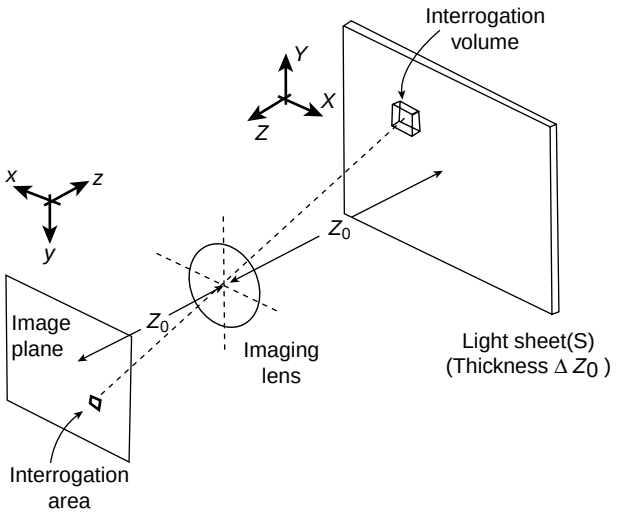


Figure 17. Schematic representation of geometric imaging (66). (With kind permission of Springer Science and Business Media.)

the same two subregions are cross-correlated between the two images to obtain the average displacement of the particles within that subregion. From this displacement and the known time delay, the velocities within the interrogation region are obtained.

Statistical PIV assumes all particles in an interrogation subregion move a similar distance and direction. The processing algorithm then computes the mean displacement vector for the particles in the interrogation volume. Therefore, the local particle distribution pattern captured on each exposure should be similar; but the group of local particles is displaced from image to image. Statistical PIV is then "pattern recognition" of the particle distribution within an interrogation subregion, instead of the averaged particle displacements. Sophisticated pattern recognition schemes have been developed by a number of researchers; however, the cross-correlation tends to be the algorithm of choice. The use of a cross-correlation as opposed to an autocorrelation eliminates directional ambiguity in the velocity measurement. Most commercial systems use advanced cross-correlation algorithms, such as the Hart-Correlation, developed to improve signal to noise in the correlation estimate and enhance resolution (70,71).

The cross-correlation function for two interrogation subregions of frames A and B is defined by:

$$R_{\mathrm{II}}(s,\Gamma,D) = \langle \mathrm{I}(x,\Gamma)\mathrm{I}'(x+s,\Gamma',D)\rangle \tag{18}$$

where s is the shifting vector of the second interrogation window, Γ is the series of location vectors for each particle in the interrogation volume, D is the displacement vector for each particle, x is the spatial domain vectors within the interrogation area, I is the intensity matrix for the interrogation area from frame A, and I' is the intensity matrix for the interrogation area from frame B. A detailed mathematical derivation of the cross-correlation for (group) particle tracking is beyond the scope of this presentation. The location of the maximum value of cross-correlation function represents the mean particle displacement for the interrogation image. Figure 18 is an example of a cross-correlation function between two images.

The location of the cross-correlation peak should be determined with subpixel accuracy. Several curve-fitting algorithms have been developed to identify the peak in the cross-correlation. Gaussian, parabolic, and centroid are three commonly used methods in commercial software, although others exist. A Gaussian peak fit is most commonly used because the cross-correlation function of two Gaussian functions also yields a Gaussian function, which

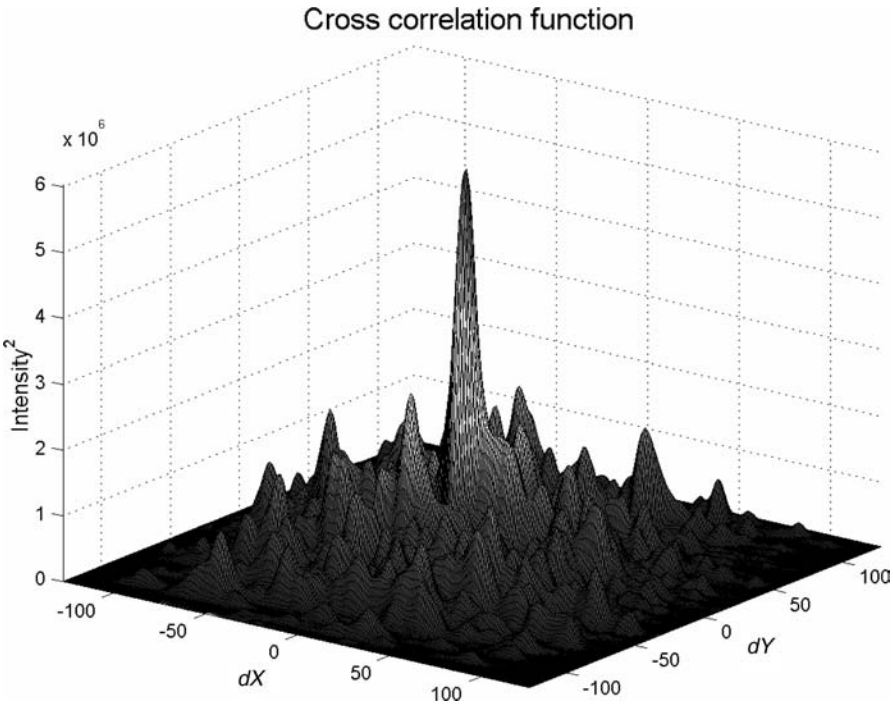


Figure 18. Representative cross-correlation map between frames A and B.

means that if the intensities of individual particle images in the interrogation area fit a Gaussian shape, the cross-correlation function can be accurately approximated by the Gaussian function, which occurs only under the condition of low displacement gradient, so that the particle distribution pattern is preserved in windows A and B. The cross-correlation function in distorted, particle image intensity distribution patterns are less accurately approximated by a Gaussian distribution. A three-point estimator for Gaussian fit works better with a narrow and symmetric peak. Centroid peak finding should be considered when the cross-correlation peak is expected to have asymmetric and irregular patterns. Such cases occur for particle images larger than 2–3 pixels in diameter, for a low intensity ratio of particle image to background noise, or for a high gradient displacement field. For "correlation-based correction," the centroid peak finding might be more suitable than Gaussian because the multiplications could distort the cross-correlation peak.

The use of a digital CCD camera presents an error source known as "peak locking." This error impacts the accuracy of the subpixel estimation in the correlation peak and thus impacts the velocity measurement. This error will be discussed later.

Like LDV and Doppler, PIV and PTV require that the fluid be seeded with tracer particles that follow the fluid motion. The particle density, the number of particles per unit volume of fluid, determines what technique should be used, PIV or PTV. Flows with a low particle density are more suited to PTV, wheras PIV works best in high particle density flows. It is assumed that the tracer particles follow the flow accurately to give an exact representation of the flow field at various times and locations. The particle density, however, should be sufficiently low to preserve the original flow dynamics. Such a dilute condition is expressed by the inequality:

$$(\rho_p \pi d_p^4 \upsilon_r)/(18\mu\,\delta_p^3) < 1 \tag{19}$$

where d_p and ρ_p are the particle diameter and density, respectively, μ is fluid viscosity, υ_r is the averaged particle velocity relative to neighboring particles, and δ_p is the average distance between particles.

Particles must be small enough to follow the fluid flow but large enough to be seen. The particle relaxation time, τ_s, should be small relative to the flow time scales.

$$\tau_s = d_p^2(\rho_p/18\mu) \tag{20}$$

In practice, τ_s is made small by using very small particles. The Stokes number for PIV experiments, St_{PIV}, can be defined as τ_s/τ_{PIV}, where τ_{PIV} is the small finite separation time between two observations (pulse separation time). St_{PIV} should be much less than 1 to assure negligible particle-fluid velocity differences over the pulse separation. However, particles must be large enough to scatter sufficient light energy to be visualized on the recording device (e.g., a CCD camera) with the goal that a particle image is at least several pixels in size. Increasing the light source energy can improve visibility, but a saturation point is reached where increasing light source energy does not help. Furthermore, high energy can damage windows and plastic test models.

The time separation of the two laser pulses must be small enough to minimize particle loss through too large a particle displacement between the first and second frames of the interrogation window. However, the time separation must be long enough both to permit adequate displacement of particles at the lowest measurable velocities in each velocity component and to minimize the impact of pixel peak locking (72). Complex and highly 3D flows can be biased in a 2D PIV system. PIV can provide very high spatial resolution, but suffers from low temporal resolution. Furthermore, high magnification imaging used in high resolution PIV introduces additional constraints and limitations that must be overcome to achieve high quality vector maps.

The challenge for PIV is to correctly track particle motion. Figure 19 shows an example of a PIV particle image. The statistical cross-correlation approach is used to track the displacement of a "group" of particles within an indicated small volume or subregion. The location of a velocity vector is at the center of the subregion spot. The spatial resolution for a "velocity vector" in PIV is the size of the interrogation subregion. Overlapping adjacent interrogation subregions is commonly used to reduce the distance between adjacent vectors and provide additional vectors in the overall vector map. However, this overlapping does not increase the spatial resolution of the velocity field, which is limited to the size of the interrogation subregion.

Commercial PIV systems use a multigrid recursive method to reduce interrogation subregion size. In the hierarchical approach, a PIV measurement with large interrogation subregions is first computed. Subsequently, the initial interrogation area is evenly divided into smaller areas for processing. Each smaller interrogation area is offset by the displacement obtained from its parent interrogation area. This process is repeated until the smallest possible interrogation size is reached (e.g., $128 \times 128 \rightarrow 64 \times 64 \rightarrow 32 \times 32 \rightarrow 16 \times 16$) for the given flow field. An iterative method can be applied at the final grid level.

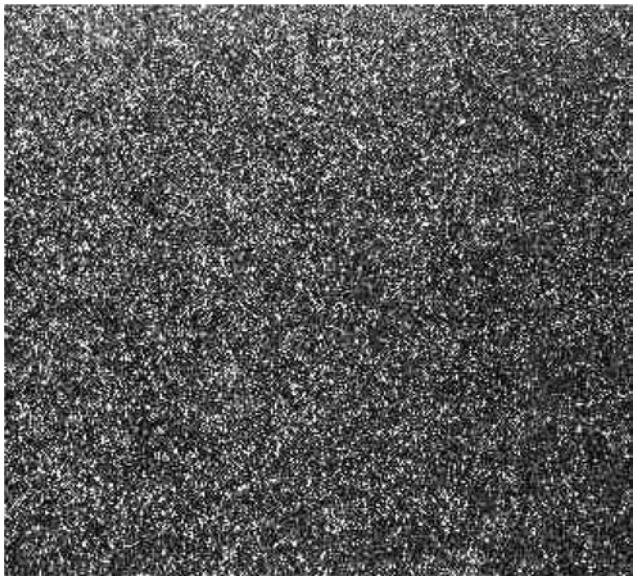


Figure 19. Example of a PIV particle image.

Similar to the multigrid method, the iterative method uses the obtained displacement from the first cross correlation to offset the second window for the next cross correlation. The difference is that it does not break down the interrogation areas into smaller windows, which is repeated until the difference between the displacements from successive cross correlations is less than one pixel. If the window B is shifted by the converged displacement, windows A and B should be virtually the same, as long as the gradient is sufficiently low. Iterative cross correlation is another way to increase accuracy.

A minimum number of particle pairs (on the order of 10) are required in PIV processing. The particle density in the flow will determine the minimum size subregion that can be used to obtain adequate vector maps. Thus, the spatial resolution is governed by the particle density. Near solid surfaces, the particle density is often lower in flows with strong wall gradients. Reducing the interrogation window size increases spatial resolution. However, an overly small window causes in-plane particle loss due to particles moving out of the interrogation spot. Several techniques exist to capture the particles moving out of the window without enlarging the interrogation spot. The first is simply to enlarge the second window to cover the expected displaced particle. The original interrogation window (frame A) is enlarged to the same size as window B and zero-padded at the extended part. The second technique is to offset the second window to the location of anticipated displacement.

Errors in PIV processing can occur from several sources. The spatial resolution for a velocity vector is the dimension of the interrogation volume. If the particles are evenly distributed, the center of an interrogation volume can be used as the vector location. The accuracy of the 'displacement depends on both the subpixel accuracy of the peak finding algorithm and the image quality. A one-tenth pixel is the most accurate (66). Time resolution for the velocity is the separation time between two pulses, as the information during this period is not recorded. The velocity error is composed of systematic and residual error. Systematic errors come from the algorithm and experiment setting or image quality, which can be minimized and uncertainty in the time separation. Residual errors are inherent in the processing, such as errors due to the peak finding algorithm. The residual errors are usually not a function of Δt. Therefore, a too small separation time increases the velocity error as this error is proportional to $1/\Delta t$.

The following discussion is relevant to the effect of image quality on PIV accuracy. Large particle images can result in wide cross-correlation peaks, which can reduce the accuracy of the peak finding algorithm. In addition, large particles require larger interrogation spots to contain an appropriate number of particles, which leads to a reduction in spatial resolution. Particle images smaller than 2 pixels in diameter, or particle displacements that are less than 2 pixels, can introduce a displacement bias, called "peak locking." The displacement peaks tend to be biased toward integer values. Peak locking presents itself as a "staircased" velocity pattern, in a region with a velocity gradient where the velocity distribution should be smooth. The calculation of spatial derivatives of this vector map then produces a mosaic pattern in the gradient map. Figure 20 illustrates these patterns. Techniques, such as multigrid or continuous window-shifting or iterative image deformation, have been proposed to overcome peak locking. Image preconditioning, such as filtering, or defocusing can optimize the image diameter. The resolution of the CCD sensor, therefore, also limits the use of smaller particles to increase the velocity resolution.

The methods developed to increase displacement accuracy rely on the assumption of low displacement gradient. High gradient tends to bias the displacement toward low values because the particles with smaller displacements are more likely to remain in the interrogation volume longer than those with higher displacements. This bias can be minimized by reducing the size of the interrogation volume and separation time. For high distortion of the particle pattern in a high gradient spot, the centroid peak finding algorithm is more suitable than the Gaussian. However, PTV, as it follows an individual particle, is not affected by high gradient. Several research groups use the displacement results from PIV to guide the local search areas for PTV in a coupled PIV/PTV processing algorithm. These coupled techniques relieve the gradient limit in PIV and increases resolution of both velocity vectors and velocity fields.

The motion across the light sheet in highly 3D flows can bias the local velocity estimation due to perspective projection, if the particle has not left the light sheet. The effect of perspective projection velocity bias, illustrated in Fig. 21 (68), is usually more severe at the edges of the image, where the projection angle increases. At high image magnification, the focal length becomes shorter and the projection angle increases, which worsens the perspective projection. Strong perspective projection could vary the magnification factor through the image plane resulting in an image distortion.

In general, the light sheet thickness is smaller than the depth of focus, δ_z. The light sheet thickness, therefore, determines the thickness of the effective interrogation volume: All illuminated particles are well focused. Most commercial systems using standard-grade Yag lasers with appropriate optics can generate light sheets that have a thickness on the order of 100–200 μm, although light sheets can be as thick as a 1 mm. In high magnification imaging, the depth of focus can become smaller than the light sheet thickness. The thickness of the effective interrogation volume is then constrained to the depth of focus. In this case, particles located beyond the focal plane but within the illuminated plane are out-of-focus and appear as highly nonuniform background image noise and can affect the cross correlation. In addition, the thickness of the effective interrogation volume determines the tolerance to out-of-plane motion. A smaller effective volume thickness increases the probability for out-of-plane particle loss. In general, the estimated maximum out-of-plane motion should be less than one-fourth of the effective volume thickness.

Data validation is another source of uncertainty in PIV. Bad velocity vectors will ultimately appear within a vector map due to noise in the displacement estimation. Several filtering algorithms have been developed to remove these bad vectors. These filter routines operate on thresholding velocity at a particular location and use either the

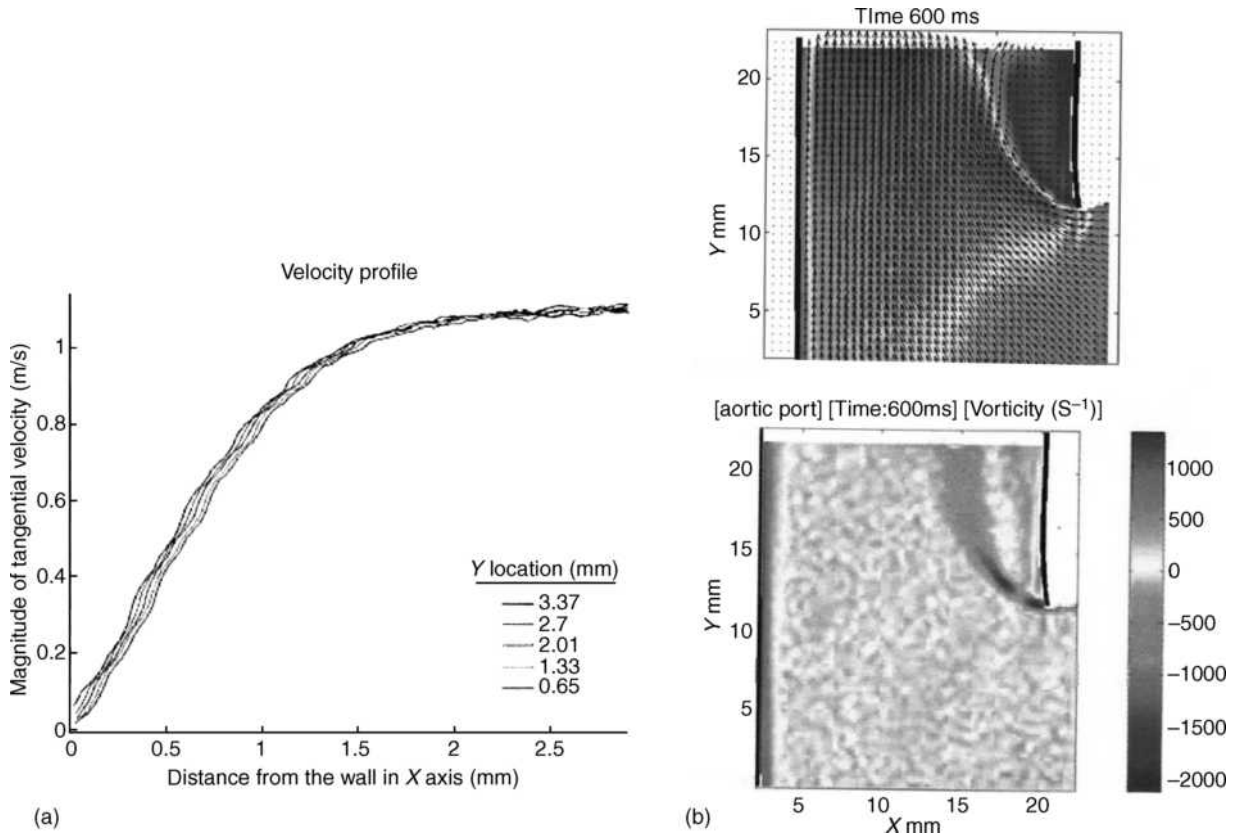


Figure 20. (a) Staircased pattern in a velocity profile at a wall. (b) Gradient field calculation (bottom image) of a measured velocity field (top image) showing mosaic pattern (68).

magnitude of the velocity estimate, the mean, the median or the rms within a predefined subregion, or other more complicated thresholds to low pass filter the estimates. Improper application of a validation scheme can overfilter the velocity map and throw away good data. For example, rms validation techniques should be carefully used in turbulent shear layers where high rms values are normally encountered. It is possible to inadvertently filter good instantaneous turbulent velocity ensembles with a tight rms filter setting. In general, some knowledge of the flow under study is needed to accurately perform vector validation.

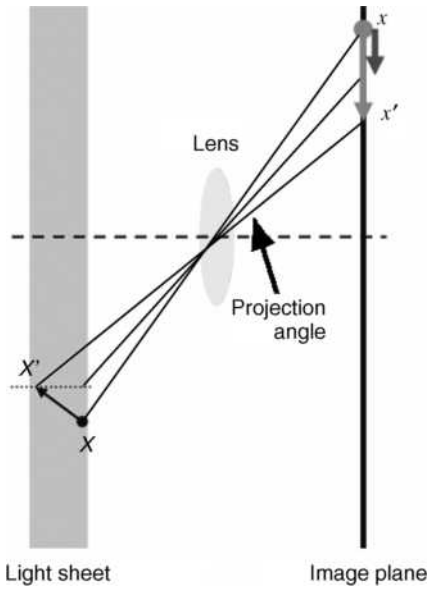


Figure 21. Illustration of the perspective projection; the red arrow is the vector obtained from perspective projection, the blue is the correct projection of displacement vector on the *XY* plane (68).

Commercial PIV systems can be two-component or three-component, planar or volume systems. A two-component, planar PIV system provides information on two components of the velocity vector. In two-component PIV, the measured displacement vector is the projection of the three-component velocity vector on the 2D plane of the light sheet. Flow information for highly three-component flows can be inaccurately represented by planar PIV images. Stereographic and holographic PIV systems have been developed for three-component measurement in a plane or volume, respectively. Although the instantaneous velocity field obtained by PIV has an advantage over LDV (or Doppler), two-exposure PIV only provides information on the particle motion during the two exposures and also suffers from poor temporal frequency response in the measurement of adjacent vector maps in time. Particle acceleration cannot be measured by direct two-exposure PIV. Four-exposure systems have been developed to permit calculation of the particle acceleration by Hassan and Phillip (73) and Lui and Katz (74), although the temporal resolution for the acceleration is not yet comparable with that of LDV.

BIBLIOGRAPHY

Cited References

1. Fung YC. Respiratory gas flow. In: Biomechanics: Motion, Flow, Stress and Growth. New York: Springer-Verlag; 1990.
2. Primiano FP. "Measurements of the respiratory system. In: Webster JG, editor. Medical Instrumentation: Application and Design. New York: John Wiley & Sons; 1998.
3. Hampers CL, Schuback E, Lowrie EG, Lazarus JM. Clinical engineering in hemodialysis and anatomy of an artificial kidney unit. In: Long Term Hemodialysis. New York: Grune and Stratten; 1973.
4. Neuman MR. Therapeutic and prosthetic devices. In: Webster JG, ed., Medical Instrumentation: Application and Design. 3rd ed. New York: John Wiley & Sons; 1998.
5. Tedgui A, Lever MJ. Filtration through damaged and undamaged rabbit thoracic aorta. Am J Physiol 1984;247:784.
6. Jin S, Oshinski J, Giddens DP. Effects of wall motion and compliance on flow patterns in the ascending aorta. J Biomech Eng 2003;125:347–354.
7. Hochareon P, Manning KB, Fontaine AA, Tarbell JM, Deutsch S. Wall shear-rate estimation within the 50cc Penn State artificial heart using particle image velocimetry. J Biomech Eng 2004;126:430–437.
8. Hochareon P, Manning KB, Fontaine AA, Tarbell JM, Deutsch S. Correlation of in vivo clot deposition with the flow characteristics in the 50cc Penn State artificial heart: A preliminary study J ASAIO 2004;50(6):537–542.
9. Merzkirch W. Flow Visualization. New York: Academic Press; 1974.
10. Rouse H, Ince S. History of hydraulics. Iowa Institute of Hydraulics Research Report. Ames, (IA); State University of Iowa; 1957.
11. Latto B, El Riedy O, Vlachopoulos J. Effect of sampling rate on concentration measurements in nonhomogeneous dilute polymer solution flow. J Rheol 1981;25:583–590.
12. Lauchle GC, Billet ML, Deutsch S. Hydrodynamic measurements in high speed liquid flow facilities. Gad-el-Hak M. In: Lecture Notes in Engineering, 46, Experimental Fluid Mechanics. New York: Springer Verlag, 1989.
13. White KC, Kavanaugh JF, Wang DM, Tarbell JM. Hemodynamics and wall shear rate in the abdominal aorta of dogs. Circ Res 1994;75(4):637–649.
14. Povey MJW. Ultrasonic Techniques for Fluids Characterization. New York: Academic Press; 1997.
15. Siedband MP. Medical imaging systems. In: Webster J, editor. Medical Instrumentation: Application and Design. New York: John Wiley & Sons; 1998.
16. Adrian RJ. Particle imaging techniques for experimental fluid mechanics. Ann Rev Fluid Mech 1991;23:261–304.
17. Webster JG. Measurement of flow and volume of blood. In: Webster JG, editor. Medical Instrumentation: Application and Design. 3rd ed. New York: John Wiley & Sons; 1998.
18. Lipowski HH, McKay CB, Seki J. Transit time distributions of blood flow in the microcirculation. In: Lee, Skalak, editors. Microvascular Mechanics. New York: Springer Verlag; 1989. p 13–27.
19. Shaughnessy EJ, Katz IM, Schaffer JP. Introduction to Fluid Mechanics. New York: Oxford University Press; 2005.
20. Bendat JS, Peirsol AG. Random Data: Analysis and Measurement Procedures. New York: John Wiley and Sons; 1986.
21. Coleman HW, Steele WG. Experimentation and Uncertainty Analysis for Engineers. New York: John Wiley and Sons; 1999.
22. Montgomery DC. Design and Analysis of Experiments, 3rd ed. New York: John Wiley and Sons; 1991.
23. White FM. Fluid Mechanics. New York: McGraw-Hill; 1979.
24. Caro CG, Pedley TJ, Schroter RC and Seed WA. The Mechanics of the Circulation. New York: Oxford University Press; 1978.
25. Pedley TJ, Schroter RC, Sudlow MF. Flow and pressure drop in systems of repeatedly branching tubes. J Fluid Mech 1971; 46(2):365–383.
26. Mori Y, Nakayama W. Study on forced convective heat transfer in curved pipes. Int J Heat Mass Transfer 1965;8:67–82.
27. Drost CJ. Vessel diameter-independent volume flow measurements using ultrasound. Proceedings of the San Diego Biomedical Symposium, 17, 1978: p 299–302.
28. Lynnworth LC. Ultrasonic flow meters. In: Mason WP, Thurston RN eds. Physical Acoustics. Academic Press; 1979.
29. Beldi G, Bosshard A, Hess OM, Althaus U and Walpoth BH. Transit time flow measurement: Experimental validation and comparison of three different systems. Ann Thorac Surg 2000;70:212–217.
30. Weyman AE. Principles and Practice of Echocardiography, 2nd ed., New York: Lea & Febiger; 1994.
31. Crowe CT, Sommerfeld M, Tsuji Y. Multiphase Flows with Droplets and Particles. Boca Raton (FL): CRC Press; 1998.
32. Comte-Bellot G. Hot-wire anemometry. Ann Rev Fluid Mech 1976;8:209–232.
33. Nerem RM, Rumberger JA, Gross DR, Muir WW, Geiger GL. Hot film coronary artery velocity measurements in horses. Cardiovasc Res 1976;10(3):301–313.
34. Nerem RM, Rumberger JA, Gross DR, Hamlin RL, Geiger GL. Hot film anemometer velocity measurements of arterial blood flow in horses. Circ Res 1974;34(2):193–203.
35. Falsetti HL, Carroll RJ, Swope RD, Chen CJ. Turbulent blood flow in the ascending aorta of dogs. Cardiovasc Res 1983;17(7): 427–436.
36. Baldwin, JT, Tarbell, JM, Deutsch S, Geselowitz DB, Rosenberg G, Hot-film wall shear probe measurements inside a ventricular assist device," J Biomech Eng, 1988;110(4):326–333.
37. Baldwin JT, Tarbell KM, Deutsch S, Geselowitz DB. Wall shear stress measurements in a ventricular assist device. J BioMech Eng 1988;V110:326–333.
38. Batten JR, Nerem RM. Model study of flow in curved and planar arterial bifurcations. Cardiovasc Res 1982;16(4):178–186.
39. Bruun HH. Hot Wire Anemometry: Principles and Signal Analysis. New York: Oxford University Press; 1995.
40. Stock DE Ed. Thermal anemometry 1993. Proceedings of the 3rd Int. Symposium on Thermal Anemometry – ASME Fluids Engineering Conference, Washington, (DC), FED 167 1993.
41. Otto, CM, The Practice of Clinical Echocardiography, Philadelphia, PA., WB Saunders Co., 1997.
42. Cape EG, Nanda NC, Yoganathan AP. Quantification of regurgitant flow through Bileaflet heart valves: theoretical and *in vitro* studies. Ultrasound Med Biol 1993;19:461–468.
43. Pettigrew RI. Magnetic resonance in cardiovascular imaging. In Zaret BL et al., editors. Frontiers in Cardiovascular Imaging. Baltimoc, MD: Raven Press; 1993.
44. Ku DN, Biancheri CL, Pettigrew RI, et al. Evaluation of magnetic resonance velocimetry for steady flow. J Biomech Eng 1990;112:464–472.
45. Hahn EL. Detection of sea-water motion by nuclear preemission. J Geophys Res 1960;65:776–777.
46. Singer JR, Crooks LE. Nuclear magnetic resonance blood flow measurements in the human brain. Science 1983;221:654–656.
47. Moran PR, Moran RA, Karstaedt N. "Verification and evaluation of internal flow and motion. True magnetic resonance imaging by the phase gradient modulation method." Radiology. 1985 Feb; 154(2):433–41.
48. Bryant DJ, Payne JA, Firmin DN, Longmore DB. Measurement of flow with NMR imaging using a gradient pulse and phase difference technique. J Comp Assist Tomogr 1984;8: 588–593.

49. Matsuda T, Shimizu K, Sakurai T, et al. Measurement of aortic blood flow with MR imaging: Comparative study with Doppler ultrasound. Radiology 1987;162:857–861.
50. Edelman RR, Heinrich PM, Kleefield J, Silver MS. Quantification of blood flow with dynamic MR imaging and presaturation bolus tracking. Radiology 1989;171:551–556.
51. Moran, PR, "A flow velocity zeugmatographic interlace for NMR imaging in humans," Magn. Res. Imaging, 1982, 1:197–203.
52. Chatzimavroudis GP, Oshinski JN, Franch RH, et al. Evaluation of the precision of magnetic resonance phase velocity mapping for blood flow measurements. J Card Mag Res 2001;3:11–19.
53. Firmin DN, Nayler GL, Kilner PJ, Longmore DB. The application of phase shifts in NMR for flow measurement. Mag Res Med 1990;14:230–241.
54. Zhang H, Halliburton SS, White RD, Chatzimavroudis GP. Fast measurements of flow through mitral regurgitant orifices with magnetic resonance phase velocity mapping. Ann Biomed Eng 2004;32(12):1618–1627.
55. Kraft KA, Fei DY, Fatouros PP. Quantitative phase-velocity MR imaging of in-plane laminar flow: Effect of fluid velocity, vessel diameter, and slice thickness. Med Phys 1992;19: 79–85.
56. Suzuki J, Caputo GR, Kondo C, Higgins CB. Cine MR imaging of valvular heart disease: Display and imaging parameters affect the size of the signal void caused by valvular regurgitation. Am J Roentgenol 1990;155:723–727.
57. Zhang H, Halliburton SS, Moore JR, Simonetti OP, et al. Ultrafast flow quantification with segmented k-space magnetic resonance phase velocity mapping. Ann Biomed Eng 2002;30:120–128.
58. Walker PG, Oyre S, Pedersen EM, Houlind K, Guenet FS, Yoganathan AP. A new control volume method for calculating valvular regurgitation. Circ 1995;92:579–586.
59. Chatzimavroudis GP, Oshinski JN, Pettigrew RI et al. Quantification of mitral regurgitation with magnetic resonance phase velocity mapping using a control volume method. J Mag Reson Imag 1998;8:577–582.
60. Drain LE. The Laser Doppler Technique. New York: John Wiley and Sons; 1980.
61. Durst F, Melling A, Whitelaw JH. Principles and Practice of Laser Doppler Anemometry. San Diago, (CA): Academic Press; 1976.
62. Edwards, Robert V, "Report of the special panel on statistical particle bias problems in laser anemometry," J Fluids Eng, Transactions of the ASME, V 109, n 2, Jun, 1987, p 89–93.
63. Tomonaga G, Mitake H, Hoki N, Kajiya F. Measurement of point velocity in the canine coronary artery by laser doppler velocimeter with optical fiber. Jap J Surg 1981;11(4):226–231.
64. Manning KB, Przybysz TM, Fontaine AA, Tarbell JM, Deutsch S. Near field flow characteristics of the Bjork-Shiley monostrut valve in a modified single shot valve chamber. ASAIO J 2005; 51(2):133–138.
65. Ellis JT, Healy TM, Fontaine AA, Westin MW, Jarret CA, Saxena R, Yoganathan AP. An *in vitro* investigation of the retrograde flow fields of two bileaflet mechanical heart valves. J Heart Valve Disease 1996;5:600–606.
66. Raffel M, Willert CE, Kompenhans J. Particle Image Velocimetry: A Practical Guide. New York: Springer; 1998.
67. Oley LA, Manning KB, Fontaine AA, Deutsch S. "Off design considerations of the 50cc Penn State ventricular assist device. Art Organs 2005;29(5):378–386.
68. Hochareon P. Development of particle image velocimetry (PIV) for wall shear stress estimation within a 50cc Penn State artificial heart ventricular chamber. Ph.D. dissertation Bioengineering Department, Penn State University, 2003.
69. Lee SJ, Kim GB. X-ray particle image velocimetry for measuring quantitative flow information inside opaque objects. J Appl Phys 2003;94:3620–3623.
70. Hart DP. Super-resolution PIV by recursive local correlation. J Visualization 1999;10:1–10.
71. Hart DP. PIV error correction. Exp Fluids 2000;29(1):13–22.
72. Christensen, KT, "The influence of peak-locking errors on turbulence statistics computed from PIV ensembles," Experiments in Fluids, Vol. 36, n 3, March, 2004, p 484–497.
73. Hassan YA, Phillip OG. A new artificial neural network tracking technique for particle image velocimetry. Exp Fluids 1997;23(2):145–154.
74. Lui X, Katz J. Measurements of pressure distribution in a cavity flow by integrating the material acceleration. Proceedings of 2004 Heat Transfer and Fluids Engineering Conference, ASME HT-FED04-56373, July, 2004.

Reading List

Adrian RJ. Laser velocimetry. In: Fluid Mechanics Measurements. *New York*: 1983.

See also BLOOD RHEOLOGY; HEMODYNAMICS; PERIPHERAL VASCULAR NONINVASIVE MEASUREMENTS.

FLOWMETERS, RESPIRATORY. See PNEUMOTACHOMETERS.

FLUORESCENCE MEASUREMENTS

ROBERT J. KLEBE
GUSTAVO ZARDENETA
PAUL M. HOROWITZ
University of Texas
San Antonio, Texas

INTRODUCTION

Following the absorption of light, fluorescent compounds release light at a less energetic wavelength. This phenomenon can be used to detect the presence of many compounds with exceedingly high sensitivity and selectivity.

The objective of this article is to acquaint the reader with the basic principles of fluorescence spectroscopy. For a more detailed analysis of this area, several reviews exist (1–5) that should be useful to readers with different levels of sophistication.

Most readers are probably familiar with the basic phenomenon involved in fluorescence measurements. For example, when certain minerals are irradiated with ultraviolet light, light in the visible region of the electromagnetic spectrum is released. In this case, absorption of high energy ultraviolet (UV) light excites electrons to leave their lowest energy state (the ground state); upon return to the ground state, energy is emitted as either light or heat. The energy emitted during fluorescence must be of a lower energy than the light that initially excited a compound and, hence, high energy ultraviolet light is emitted from a mineral as lower energy visible light.

When one deals with biological samples, one is always confronted with the problem of detecting extraordinarily small amounts of a compound mixed in an array of other compounds. As this article will show, fluorescence analysis provides a rather simple means to detect the existence of a compound within a complex mixture. For obvious

reasons, the ability of fluorescence spectroscopy to detect exceptionally small amounts of a compound in biological specimens has many applications of biomedical interest.

The analytical power of fluorescence measurements may be appreciated by consideration of the two following practical analogies. The sensitivity of fluorescence measurements is similar to the increased sensitivity with which one could measure the light output of a flashlight in a completely darkened room versus a sunlit room. In the case of the sunlit room, the light output of the flashlight would represent $< 1\%$ of the total light in the room; indeed, it would be difficult for an observer in a sunlit room to detect whether the flashlight was on or off. In contrast, in a darkened room, an observer would readily be able to sense whether a flashlight was turned on. In fluorescence measurements, light strikes a sensor only if the compound of interest is present and, thus, as in the case of the darkened room, the sensitivity of the fluorescence measurements is quite high.

The ability of fluorescence measurements to detect particular compounds in complex mixtures is due to the fact that several conditions must be met before a fluorescence signal is detected. Just as the location of a person would be precisely defined if one knew their street address, which consists of a city, state, street name, and house address, the selective ability of fluorescence measurements to detect the presence of only desired compounds arises from the fact that a compound will fluoresce only under a series of quite restrictive conditions. In the case of fluorescence spectroscopy, one observes light emanating from a compound if and only if the following conditions are met: (a) the compound must absorb light; (b) the compound must be capable of emitting light following excitation (the compound must be fluorescent); (c) the compound must be irradiated at a particular excitation wavelength; (d) a detector must be set to sense light at a different, less energetic emission wavelength; (e) a discrete time (in nanoseconds) must elapse between the time of excitation and emission; and (f) other conditions, such as type of solvent and pH, must be satisfactory. As a street address defines the location of a person, the parameters involved in the fluorescence of a compound can be used to determine the presence or absence of just a single compound. For example, the presence of the compound, fluorescein, can be accurately detected in a biological specimen because (a) fluorescein is a fluorescent compound; (b) fluorescein is excited to fluoresce only at wavelengths near 488 nm (nanometer) in aqueous solvents in the neutral to alkaline range; and (c) following excitation, fluorescein gives off light maximally at 514 nm (see Fig. 1). Hence, since only a very few compounds have fluorescence properties similar to that of fluorescein, selecting light with a wavelength of 488 nm to irradiate the sample and setting a detector to only respond to light with a 514 nm wavelength allows one to detect fluorescein in a complex mixture.

In addition to an overview of the theory involved in this area, a brief introduction into the instrumentation involved in fluorescence determinations will be presented. While there are numerous applications of fluorescence that are outside of the scope of this article (1–5), a few specific examples of the use of fluorescence measurements in biomedical studies will be presented and technical problems in interpretation of fluorescence data will be pointed out.

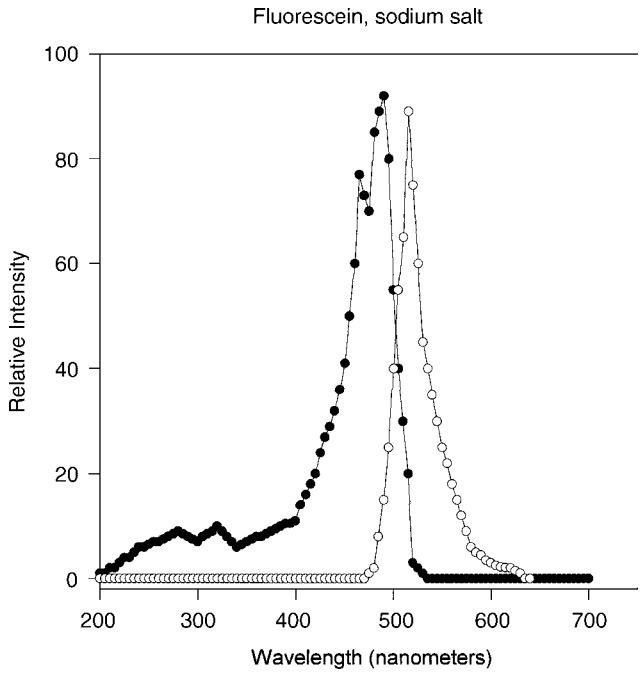


Figure 1. Excitation and emission spectra of the fluorescent compound, fluorescein. As described in the text, excitation and emission maxima are determined empirically and, then, spectra are obtained for the excitation and emission of fluorescein. Excitation of fluorescein at any wavelength in the excitation spectrum will produce emission at all wavelengths in the emission spectrum, with the exception of wavelengths of higher energy. Thus, fluorescence involves the conversion of high energy photons into photons of lower energy plus heat.

THEORY AND INSTRUMENTATION

After absorption of light energy, all compounds release some energy as heat; fluorescent compounds represent a special case in which some energy is also given off as light. Spectrofluorimeters are optical devices that measure the amount of light emitted by fluorescent compounds.

Spectrofluorimeters

While a complete description of the theory of fluorescence is quite involved (see Ref. 3, for a review), one can grasp the basic principles of the phenomenon from the following examples. If one were to aim a flashlight at a mirror, one would find that light would reflect off the surface of the mirror; in contrast, if one were to direct a flashlight beam at a black wall, one would see no light reflected from the wall. In this example, molecules in the black wall stop the transmission of light by absorbing the energy of the light and then converting the energy of light into heat energy. (Very precise measurements would reveal that the black wall was slightly warmer at the site at which the light beam struck it.) In the case of fluorescent compounds, light absorbed by a compound is converted into light (of a lower energy) as well as heat. The brighter the light that strikes a fluorescent compound, the stronger the fluorescent light emitted by the compound. The particular wavelength absorbed by a compound and the wavelength that is later emitted are characteristics of each fluorescent compound (Fig. 1).

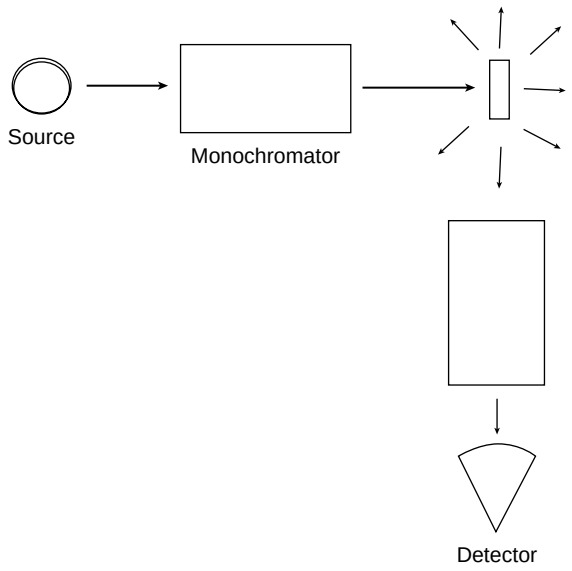


Figure 2. Design of a typical spectrofluorimeter. As described in the text, a desired wavelength is selected with the aid of a monochromator and is used to irradiate a sample (large arrows); light emitted by the fluorescent sample is released in all directions (small arrows). The detector is positioned at a right angle with respect to the excitation beam in order to avoid light from the exciting beam entering the detector.

While light reflection from a mirror occurs at an angle determined by the angle of the incident light, light emission by a fluorescent compound occurs in all directions. For example, if one observed a beam of light striking a mirror in a smoke filled room, one would see a beam of light reflected at a unique angle from the mirror. In contrast, if one observed a beam of light striking a solution of fluorescent compound, one would see the emitted light appear as if it originated from a glowing sphere (and, compared to the original beam of light, the emitted light would have a color that was shifted toward the red end of the spectrum, i.e., lower energy).

The above features of fluorescence dictate the design of instruments used in the measurement of fluorescence (see Fig. 2). First, an intense beam of light is passed through a filter (or monochromator) that permits only light of a desired band of wavelengths to proceed toward the sample. As indicated in Fig. 2, light not absorbed by the sample passes straight through the sample without change in angle while the light emitted by the fluorescent sample is released in all directions. By positioning the detector at a 90° angle with respect to the exciting beam, instrument designers can minimize the amount of stray light from the exciting beam that reaches the detector. As indicated above, the light emitted by the fluorescent sample is given off in all directions and, thus, the positioning of the detector with respect to the exciting beam does not affect the relative amount of the emitted fluorescent light received by the detector. Hence, as we have shown above, the design of a fluorimeter is predicated upon maximizing the signal/noise ratio.

The sensitivity of fluorescence measurements is inherent in the design of instruments used to make such measurements. Since only light that is emitted from a fluorescent compound reaches the detector, the sensitivity of the measurement is equivalent to our analogy of measuring the light from a flashlight in a darkened room. The ability to selectively measure the presence of a given compound is determined by the appropriate selection of filters (or settings on monochomators) used to make the measurement. In contrast, analytical measurements based on absorption are inherently less sensitive (due to signal/noise problems).

Fluorescence Microscopy

The design of a fluorescence microscope is based on the same principles used in the construction of a spectrofluorimeter. In the technique of reflected fluorescence microscopy, light of a desired wavelength is used to irradiate a sample and, then, a dichroic mirror is used to separate light emitted by the fluorescent compounds in the sample from light in the original exciting beam (Fig. 3). A dichroic mirror is an optical device that reflects light of certain

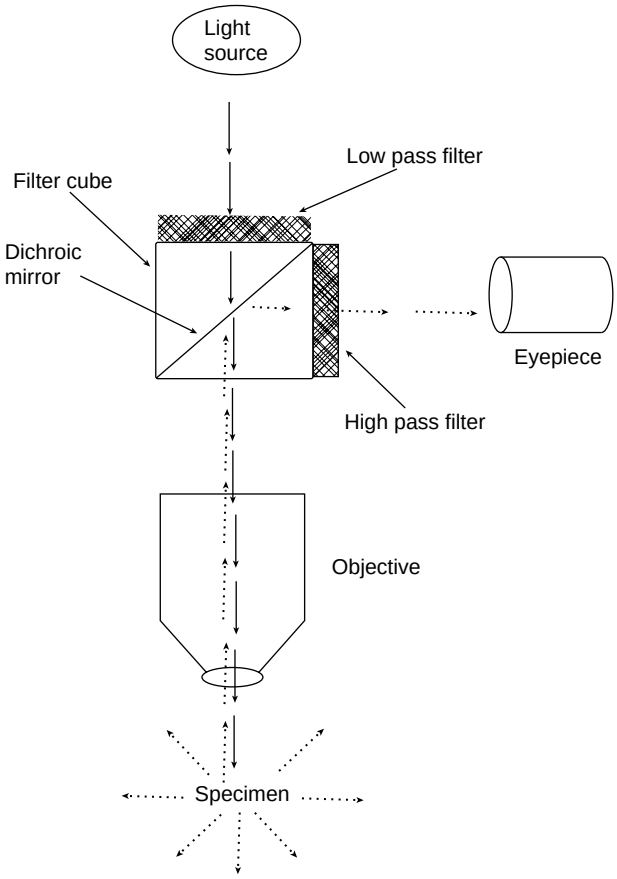


Figure 3. Fluorescence microscope design. A fluorescence microscope employes optical components that allow one to irradiate a specimen at one wave lenght and then observe the fluorescence of a microscopic object at a second wavelength. The excitation beam (dashed line) is directed toward the specimen by the dichroic mirror and objective lens. Due to the fact that light emitted by a fluorescent object is released in all directions, one can observe the fluorescence of the specimen with the same objective lens that was used to irradiate the specimen. The emitted light (dotted line) is collected by the objective, passes through the dichroic mirror, and is observed via the eyepiece.

desired wavelengths and transmits light of other wavelengths.

Fluorescence microscopes can be used to detect fluorescence compounds in exceedingly small specimens. Via the technique of immunofluorescence, one can visualize the distribution of virtually any biological compound in living or fixed tissues by using antibodies tagged with fluorescent compounds.

PRACTICAL APPLICATIONS

The high sensitivity and selectivity of fluorescence methods permits the detection of exceedingly low amounts of many compounds of biological and biomedical interest.

In the preceding section, the basic methods that are used to detect fluorescent compounds were described. In the following section, the criteria that are used in establishing the presence or absence of particular chemical species are presented.

Detection of Fluorescent Compounds

It should be pointed out initially that only a small percentage of the known compounds are fluorescent. Thus, while most compounds cannot be studied by fluorescence techniques, the mere fact that an unknown compound is fluorescent can be used as a diagnostic means to positively identify it. In a following section, literature references are provided that list many fluorescent compounds and their properties (6,7).

Once it is established that a sample contains a fluorescent compound, one can use several criteria to establish the identity of the unknown compound. First, by trial-and-error, one establishes the wavelength(s) at which the compound is maximally stimulated to fluoresce (the excitation maxima). Second, again by trial-and-error, one determines the wavelength(s) at which the emission of light is highest (the emission maxima). Fluorescence emission spectra are then generated by holding the excitation monochromator at the excitation maximum wavelength and recording the intensity of light released from the sample as the wavelength of the emission monochromator is varied (Fig. 2). In this manner, a spectrum is produced that (a) describes the wavelengths at which a compound emits light and (b) establishes the relative efficiency with which light of different wavelengths is emitted (Fig. 1). The shape of the emission spectrum and the number of major and minor peaks are characteristics of each compound and are important parameters in establishing the identity of an unknown compound. The excitation spectrum, when appropriately corrected, is often identical to the absorption spectrum of the fluorescent compound. (In a similar fashion, the fluorescence excitation spectra are established by holding the emission monochromator at the emission maximum and varying the settings of the excitation monochromator.)

Strong evidence for the identity of an unknown compound is provided by (a) the establishment that a compound is fluorescent and (b) the shapes of the excitation and emission spectra. In addition, one could use other parameters involved in fluorescence to characterize a compound, namely, (a) the fluorescent lifetime of a compound, (b) the quantum yield, and (c) the perturbation of fluorescence by various solvents. It is possible that two compounds could have identical excitation and emission spectra just as it is possible that two individuals could have the same blood groups. Other analytical methods may be required to establish the identity of a compound with certitude. If the sample under study is quite impure, it is possible to use high performance liquid chromatography (HPLC) in conjunction with a fluorescence detector to gain yet another parameter that is characteristic of a compound, namely, its elution time from a particular HPLC column. The nature of the sample and the reasons for interest in it will determine the rigor required in its characterization.

Biomedical Applications

As described above, fluorescence methods provide means to identify compounds in complex mixtures. The sensitivity of fluorescence methods coupled with their inherent selectivity provide valuable means for biomedical analysis; several examples are described below.

Identification of Compounds in Biological Samples. If a compound of interest proves to be fluorescent, several analytical methods for the detection and quantitation of this compound immediately become available. Many drugs and biological compounds can be detected by fluorescence analysis (6,7). Fluorescence can be used to identify these and many other biologically active compounds in complex pathological specimens.

Due to the sensitivity of fluorescence methods to detect compounds at the level of one part per billion, such methods can be used to determine the presence of environmental pollutants with great sensitivity. The presence of specific pesticides and potentially carcinogenic aromatic hydrocarbons from cigarette smoke can be easily detected. Leakage of industrial wastes into the environment can be proven by placing fluorescent compounds into waste containers and later detecting the fluorescent compound in streams and lakes.

Fluorescence spectra of many chemicals and pharmaceuticals have been obtained under well controlled conditions and are published as the Sadtler Standard Fluorescence Spectra (6). Passwater (7) published a three volume series that presents literature citations for the fluorescence properties of a wide variety of compounds.

Fluorescence Microscopy. Following staining with fluorescent dyes, the use of fluorescence at the microscopic level can be used to determine the sex of an individual from single hair follicle cells, teeth, or from blood smears (8). Sex determination is based upon the fact that male cells have a highly condensed Y chromosome that appears as an intense fluorescent spot in the nucleus of each male cell. Fluorescent dyes have found many applications in the biomedical sciences (1,5).

Immunofluorescence. Antibodies can be prepared that specifically recognize a wide variety of molecules and microbial organisms. By labeling such antibodies with fluorescein or other fluorescent probes, one can visualize

the presence of antigens at the subcellular level; this approach has been widely used in molecular biology (9). In addition, one can visualize specific molecules and organisms in pathological specimens. The identification of disease causing microorganisms in pathological specimens by immunofluorescence (10) can be used. Immunofluorescence microscopy can also be employed in the identification of bacteria in food products (11).

Fluorescence Imaging. In addition to localizing molecules in histological sections, fluorescence has found numerous novel applications in cell biology. The availability of highly sensitive optical detection systems has permitted the localization of specific molecules in living cells. By tagging recombinant proteins with the green fluorescent protein (GFP) of jellyfish origin, one can track the expression and translocation of specific proteins in living cells during hormonal responses (12). The rate at which molecules move within living cells can be determined by fluorescence recovery after photobleaching (FRAP), which involves the photobleaching of molecules in a small region of a cell and then monitoring of the recovery of fluorescence with time as unbleached molecules return to the bleached area (13). Fluorescence energy transfer and fluorescence polarization methods (14–16) can also be used to study the interaction of molecules within living cells (15,16).

Fluorescence Polarization. Fluorescence polarization is perhaps the only method in biology that is directly attributable to Albert Einstein. The principle of fluorescence polarization involves the fact that emission of a photon is delayed by a few nanoseconds following absorption of light. During the delay in emission of light, Brownian motion will result in the movement of a molecule and smaller molecules will move more than larger molecules. Thus, molecules excited by polarized light will emit progressively depolarized light as the size of the molecule increases. Hence, if a small fluorescent molecule binds to a large nonfluorescent molecule, the light emitted by the bound small fluorescent molecule will become more polarized. Thus, the method of fluorescence polarization permits one to measure the binding of ligands to large molecules in real time (14).

Forster Resonance Energy Transfer (FRET). This method involves the transfer of energy from a fluorophore, which absorbs light to a second molecule that emits light at a different wavelength. Since the efficiency of FRET depends on the distance between the absorbing and emitting molecules, FRET permits one to obtain information about the distance between two molecules. Thus, if the molecule that absorbs light binds to a molecule that emits light via FRET, one can measure the binding event via release of light.

Molecular Biology Applications of Fluorescence. The high sensitivity of Fluorescence has been used as the basis of several important methods in molecular biology. For example, real-time polymerase chain reaction (PCR) methods can be used to quantify the amount of a specific ribonucleic acid (RNA) species in a small sample of cells. In this method, fluorescence energy transfer is used to detect the increase in PCR product with time (16).

The jellyfish green fluorescent protein (GFP) has become an important tool in molecular biology because it is fluorescent without the necessity of binding or reacting with a second molecule. Mutant GFP and GFP-like molecules from various species have been described (17) that emit light at a variety of wavelengths. Thus, one can engineer the sequence of the GFP into a recombinant protein and have a fluorescently tagged protein (12). In contrast, the firefly luciferase protein must react with ATP and luciferin in order to release a photon (18).

The high sensitivity and specificity of fluorescence should find many new applications in the future.

GLOSSARY

Dichroic Mirror. An optical device that reflects light of a desired band of wavelengths yet permits the transmission of light of another band of wavelengths. Dichroic mirrors are used in fluorescence microscopes to separate the light that excites a sample from the light emitted from the sample.

Emission (Fluorescent). The release of light by a compound that follows the absorption of a photon.

Emission Wavelengths. Following absorption of light at a wavelength capable of inducing fluorescence, light is released by the compound at less energetic wavelengths, termed the emission wavelengths (Fig. 1).

Excitation (Fluorescent). Following the absorption of a photon, one or more electrons of a fluorescent compound are promoted to a more energetic,"excited" state of the compound.

Excitation Wavelengths. While light can be absorbed by fluorescent compounds at many wavelengths, only certain wavelengths, termed excitation wavelengths, are capable of inducing fluorescent emission of light. The wavelengths are often the same as those absorbed by the compound of interest.

Filter (Optical). Generally a colored piece of glass that transmits only certain wavelengths of light. Interference filters reflect light that is not transmitted while color glass filters absorb light and convert the absorbed energy into heat.

Fluorescence. The emission of light by certain compounds that is induced by the absorption of light at a more energetic wavelength.

Fluorescent Lifetime. The amount of time, generally in nanoseconds, that expires between the absorption of a photon and the emission of a photon.

Monochromator. An optical device that is used (a) to separate light into its component wavelengths and, then, (b) to isolate a desired group of wavelengths. Such devices employ either a prism or a diffraction grating.

Nanosecond. $0.000000001\ \text{s} = 10^{-9}$ s. (There are 1000 million ns in 1 s).

Quantum Yield. The percentage of excited compounds that release a photon of light during their return to the ground state. In most cases, the absorption of light by a compound is followed by the liberation of heat, rather than light.

Spectrofluorimeter. An optical device that is used to measure the amount of light that is emitted by fluorescent compounds.

BIBLIOGRAPHY

Cited References

1. Hof M, Hutterer R, Fidler V. Fluorescence Spectroscopy in Biology. New York: Springer; 2005.
2. Lakowicz JR. Principles of Fluorescence Spectroscopy. New York: Plenum; 1983.
3. Valeur B. Molecular Fluorescence: Principles and Applications. New York: Wiley-VCH; 2004.
4. Albani JR. Structure and Dynamics of Macromolecules: Absorption and Fluorescence Studies. New York: Elsevier; 2004.
5. Kohen E, Hirschberg JG, Santus R. Fluorescence Probes in Oncology. Imperial College Press; 2002.
6. Sadtler Standard Fluorescence Spectra. Philadelphia: Sadtler Research Laboratories;
7. Passwater RA. Guide to Fluorescence Literature. Vols. 1–3. New York: Plenum; 1967.
8. Kringsholm B, Thomsen JL, Henningsen K. Fluorescent Y-chromosomes in hairs and blood stains. Forensic Sci 1977;9:117.
9. Giambernardi TA, et al. Neutrophil collagenase (MMP-8) is expressed during early development in neural crest cells as well as in adult melanoma cells. Matrix Biol 2001;20:577–587.
10. Stenfors LE, Raisanen S. Quantification of bacteria in middle ear effusions. Acta Otolaryngol 1988;106:435–440.
11. Rodrigues UM, Kroll RG. Rapid and selective enumeration of bacteria in foods using a microcolony epifluorescence microscopy technique. J Appl Bacteriol 1988;64:65–78.
12. Rivera OJ, et al. Role of promyelocytic leukemia body in the dynamic interaction between the androgen receptor and steroid receptor coactivator-1 in living cells. Mol Endocrinol 2003;17:128–140.
13. Snickers YH, van Donkelaar CC. Determining diffusion coefficients in inhomogeneous tissues using fluorescence recovery after photobleaching. Biophys J 2005.
14. Bentley KL, Thompson L, Klebe RJ, Horowitz P. Florescence polarization: A general method for studing ligand interactions. Bio Techniques 1985;3:356–366.
15. Rizzo MA, Piston DW. Hight-contrast imaging of fluorescent protein FRET by fluorescence polarization microscopy. Biophys J 2005;88:14–16.
16. Peng XH, et al. Real-time detection of gene expression in cancer cells using molecular beacon imaging: New Strategies for cancer research. Cancer Res 2005;65: 1909–1917.
17. Shagin DA, et al. GFP-like proteins as ubiquitous metazoan superfamily: Evolution of functional features and structural complexity. Mol Biol Evol 2004;21:841–850.
18. Branchini BR, et al. An alternative mechanism of bioluminescence color determination in firefly luciferase. Biochemistry 2004;43:7255–7262.

See also COLORIMETRY; MICROSCOPY, FLUORESCENCE; ULTRAVIOLET RADIATION IN MEDICINE.

FLUORESCENCE MICROSCOPY. See MICROSCOPY, FLUORESCENCE.

FLUORESCENCE SPECTROSCOPY. See FLUORESCENCE MEASUREMENTS.

FLUORIMETRY. See FLUORESCENCE MEASUREMENTS.

FRACTURE, ELECTRICAL TREATMENT OF. See BONE UNUNITED FRACTURE AND SPINAL FUSION, ELECTRICAL TREATMENT OF.

FUNCTIONAL ELECTRICAL STIMULATION

GANAPRIYA VENKATASUBRAMANIAN
RANU JUNG
JAMES D. SWEENEY
Arizona State University
Tempe, Arizona

INTRODUCTION

Functional electrical stimulation (FES) is a rehabilitative technique where low level electrical voltages and currents are applied to an individual in order to improve or restore function lost to injury or disease. In its broadest definition, FES includes electrical stimulation technologies that, for example, are aimed at restoration of a sense of hearing for the deaf, vision for the blind, or suppression of seizures in epilepsy or tremors for people with Parkinson's disease. Most FES devices and systems are known then as "neuroprostheses" because through electrical stimulation they artificially modulate the excitability of neural tissue in order to restore function. While sometimes used synonymously with FES, the term functional neuromuscular stimulation (FNS) is most commonly used to describe only those FES technologies that are applied to the neuromuscular system in order to improve quality of life for people disabled by stroke, spinal cord injury, or other neurological conditions that result in impaired motor function (e.g., the abilities to move or breathe). Another technology closely related to FES is that of therapeutic electrical stimulation (TES), wherein electrical stimulation is applied to provide healing or recovery of tissues (e.g., muscle conditioning and strengthening, wound healing). As will be seen, some FES and FNS technologies concurrently provide or rely upon such therapeutic effects in order to successfully restore lost function. For illustrative purposes, much of this article is centered on FNS and related TES devices and technologies. For a wider exposure to additional FES approaches and neural prosthetic devices, the reader is referred to this article's *Reading List*, which contains references to a number of general books, journal articles, and on-line resources.

An important consideration in most all FNS technologies is that significant neural tissue remains intact and functional below the level of injury or disease so that electrical stimulation can be applied effectively. Individuals exhibiting hemiplegia (i.e., paralysis on one side of the body) due to stroke, for example, will exhibit paralysis in an impaired limb due to loss of control from the central nervous system (CNS), not because the peripheral nervous system (PNS) innervation of skeletal muscles in the limb has been lost. Similarly, while spinal cord injury (SCI) destroys motor neurons at the level of injury either partially or completely, many motor neurons below the level of injury may be spared and remain intact. Therefore, in stroke or SCI the axons of these intact motor neurons can be artificially excited by introducing an appropriate

electrical field into the body using electrodes located on the skin surface, or implanted within the body. Artificial excitation of motor nerves by electrical excitation can generate action potentials (propagating excitation waves) along axons that, when they arrive at synaptic motor-endplate connections to skeletal muscle fibers, act to generate muscle force much as the intact nervous system would. Thus, lower extremity FNS systems often have the objective of restoring or improving mobility for stroke or SCI individuals. Upper extremity FNS systems often are designed to restore or augment reaching and grasping movements for SCI subjects. Both FNS and TES technologies are of course not a cure for stroke, spinal cord injury or diseases (e.g., cerebral palsy or multiple sclerosis where FNS also has been used). They are also not universally beneficial, and must be carefully matched by a clinician to an individual and their medical condition (1). On the other hand, as will be seen in the remainder of this article, FES and TES systems can provide greatly improved quality of life for many people who use them.

THEORY AND APPLICATION

In 1961, Liberson and co-workers proposed the usage of electrical stimulation in what was called functional electrotherapy to restore or augment movement capability that has been lost or compromised due to injury or disease (2). Specifically, Liberson's group developed the first electrical stimulation system for correction of hemiplegic drop foot: a gait disability occurring in some stroke survivors (for an excellent review of the history of development of neural orthoses for the correction of drop foot see Ref. 3). Moe and Post subsequently coined the term functional electrical stimulation to describe such techniques (4).

Electrical stimulation devices and systems now have been developed to activate paralyzed muscles in human subjects for a variety of applications in both the research lab and the clinic. Both FES and FNS systems have seen their greatest use as a tool for long-term rehabilitation of persons with neurological disorders (e.g., spinal cord injury, head injury, stroke) (5–10). For example, implanted electrical stimulation devices have been developed that can restore hand-grasp function to people with tetraplegia (11). Stimulation devices that utilize percutaneous electrodes (thin wires that cross the skin) have been developed to provide individuals with thoracic-level spinal cord injury with the ability to stand and step (12–14). Other devices that utilize electrodes placed on the surface of the skin can restore standing and locomotor function to individuals with spinal cord injury or other neuromuscular disorders (6,8,15,16). One system that uses surface electrodes (Parastep, Sigmedics Inc.) is FDA approved for use by people with thoracic level spinal cord injury and has been used at several rehabilitation centers worldwide. These efforts have clearly demonstrated that neuromuscular stimulation can be effectively used to activate paralyzed muscles for performing motor activities of daily living.

The basis by which all neuromuscular stimulation systems function is artificial electrical activation of muscle force, usually through excitation of the nerve fibers that innervate the skeletal muscle(s) of interest.

Excitation, Recruitment, and Rate Modulation

The nerve fibers that innervate skeletal muscle fibers are myelinated in nature, which means that they are regularly along their lengths ensheathed within layers of Schwann-cell derived myelin separating exposed axonal membrane at nodes of Ranvier. Myelination enables increased propagation velocities via saltatory conduction in such nerve fibers. The cell bodies of these alpha motor neurons lie within the ventral horn of the spinal cord. The efferent axons of these cells ($\sim$ 9–20 μm in diameter) pass out from the spinal cord via the ventral roots and project then to muscle fibers within peripheral nerve trunks. When spared during damage or disease of the nervous system, alpha motor neurons and their axons usually form the substrate of electrical activation of skeletal muscle force in FNS applications. This may come as something of a surprise to the reader, in that skeletal muscle cells are themselves also excitable. Why then is indirect stimulation of the innervating nerve fiber generally the mechanism by which force is generated rather than direct stimulation of the muscle cells themselves? The reason is that large myelinated nerve fibers are usually excited at lower stimulus amplitudes (voltage or current) and with shorter stimulus pulse widths than are skeletal muscle cells (assuming similar spatial separations of electrodes to cells) (17). Electrical stimulation of myelinated nerves to threshold occurs when a critical extracellular potential distribution is created along or near the cell. At threshold, outward transmembrane currents are sufficient to depolarize the nerve cell membrane voltage to the level where an action potential is generated.

In normal physiology, there exist two natural control mechanisms to regulate the force a single muscle produces—recruitment and rate coding. Motor units are recruited naturally according to the Size Principle (18,19). Small alpha motor neurons innervating slow motor units have a low synaptic threshold for activation, and therefore are recruited first. As more force is demanded by an activity, progressively larger alpha motor neurons that innervate fast motor units are recruited. The second method of natural force regulation is called rate coding. Within a given motor unit there is a range of firing frequencies. Alpha motor neurons innervating fast-twitch motor units have firing rates that are higher than those that innervate slow-twitch units (20,21). Within that range, the force generated by a motor unit increases with increasing firing frequency. If an action potential reaches a muscle fiber before it has completely relaxed from a previous impulse, then force summation occurs. Twitches generated by the slow motor units have a fusion frequency of 5–10 Hz and reach a tetanic state at 25–30 Hz. The fast motor units may achieve fusion at 80–100 Hz (21,22).

The contractile properties of the muscle are largely dependent on the composition of the skeletal muscle (i.e., the muscle fiber types). The composition of muscle fibers varies across species. The composition of muscle fibers in the hindlimbs of the rat are predominantly fast fibers (23) whereas, human skeletal muscle is composed of a heterogenous collection of muscle fiber types (24). This is also indicated in the differences in fusion frequencies observed

Table 1. Skeletal Muscle Fiber Types and Their Characteristics

	Skeletal Muscle Fiber Types and Characteristics		
Fiber type	Type I	Type IIa	Type IIb
Other names	Slow red	Fast red	Fast white
	Slow oxidative (SO)	Fast oxidative (FOG)	Fast glycolytic (FG)
	Slow (S)	Fast resistant (FR)	Fast fatigable (FF)
Motor unit size	Smallest	Moderate	Largest
Firing order	1	2	3
Stimulation threshold	Lowest	Moderate	Highest
Force production	Lowest	Moderate	Highest
Resistance to fatigue	Highest	Moderate	Lowest
Contraction time	Slowest	Fast	Fastest
Mitochondrial density	High	High	Low
Capillary density	Highest	Moderate	Lowest

in the two species. The fusion frequency for muscles in the human is 25 Hz (25) and those for the muscles in the rat are higher ($\sim$ 75 Hz) (26). As summarized in Table 1, from various mammalian studies, skeletal muscle fibers have been grouped into many different types according to physiological, ultrastructural, and metabolic properties. Based on histochemical measurements of adenosinetriphosphatase (ATPase) reactivities, muscles were classified into type I, type IIA, and type IIB (27). A differentiation based on combination of physiological and metabolic properties categorized muscle fibers as SO-, FOG-, FG- (28). Based on twitch speed and fatigue resistance, muscle fiber types were identified as S, FR, and FF (29). There is also an intermediate type of fast muscle fiber in certain muscles denoted type IIAB or FI (Fast Intermediate resistance to fatigue). The different muscle fiber types vary in the amount of force generated, speed of contraction, and fatigability. The slow fiber types (SO, Type I, S) generate lower force, but for a prolonged duration. They are very fatigue resistant. The fast fiber types (FG, IIB, and FF) are on the other end of the spectrum with greater force generating capacity, but briefer intervals of time. Also, these fatigue very quickly compared to slow fibers. Therefore, there is a trade off between the ability to produce force quickly and powerfully or slowly and steadily. Though slow fibers are able to generate a steady force for long periods of time, their force output is less. Fast fibers on the other hand can generate quicker, greater forces, but they fatigue very fast. Some fibers are classified in between the two extremes of slow and fast and are termed intermediate fibers. These are fast fibers, but with fatigue resistant capability (FOG, IIA, FR, IIAB, FI). The properties of these intermediate fibers lie between those of slow fibers and fast fibers. The force generated by these fibers is less that those generated by fast fibers and greater than the force produced by slow fibers.

The heterogeneity of muscle fibers within the muscle is in part due to the hierarchy of motor unit recruitment order (the Size Principle, described above) (30) indicating the influence of motor neuron activity upon muscle fiber phenotypes. The fiber-type composition within a muscle can be altered by altering the excitation patterns delivered to the muscle (induced by various exercise regimes). The best documented effects of such transformations are those that occur after chronic, low frequency stimulation (CLFS) of a predominantly fast muscle using implanted electrode systems. The fast skeletal muscles of a number of mammalian species have been shown to change to the slower phenotype in response to chronic electrical stimulation (31–39). The muscle phenotype can be manipulated to enhance fatigue resistance at the expense of contractile power and speed (40–45). Changes in metabolic activity, and muscle mass have been documented too (38,46). These transformations are also dose dependent. A continuous stimulation of rabbit fast muscle at 10 Hz completely transform the muscle fibers to the slow phenotype, but lower frequencies of stimulation produce an intermediate state of conversion. However, stimulation at 2.5 Hz for 12 weeks (47,48) or 10 months (49) results in a whole muscle consisting mainly of the fast phenotype.

CLFS has been shown to affect human muscle in a manner similar to that in animals (50–57). Electrical stimulation has shown to increase strength–force and build fatigue resistance in muscles in both healthy and SCI individuals (56,58–63). An increase in passive range of motion has also been observed (64). Electrical stimulation has been shown to prevent the shift and loss of fibers in patients with paralyzed muscles thereby increasing fatigue resistance (60,65–67). A well-defined progression of changes is observed, whereby the muscle changes first its metabolic and then its contractile properties to become slow muscle (68). This has been documented in different species and muscles suggesting that probably the effects observed are not species or muscle specific. Following transformation, the new slow fibers are indistinguishable from normal slow skeletal muscle fibers. Also, from time series studies (69) and single fiber biochemistry (70,71) it is clear that the changes that occur result from transformation at the level of the single fiber and not from fast-fiber degeneration with subsequent slow-fiber regeneration.

From the above sections, it is clear that skeletal muscle is very adaptive, and therefore provides an opportunity for conditioning and therapy after an injury. Electrical stimulation based exercise has gained much significance in toning and conditioning muscles. Even though electrical stimulation techniques are being used increasingly for rehabilitation and therapy, note that in general electrical stimulation systems generate activation patterns and

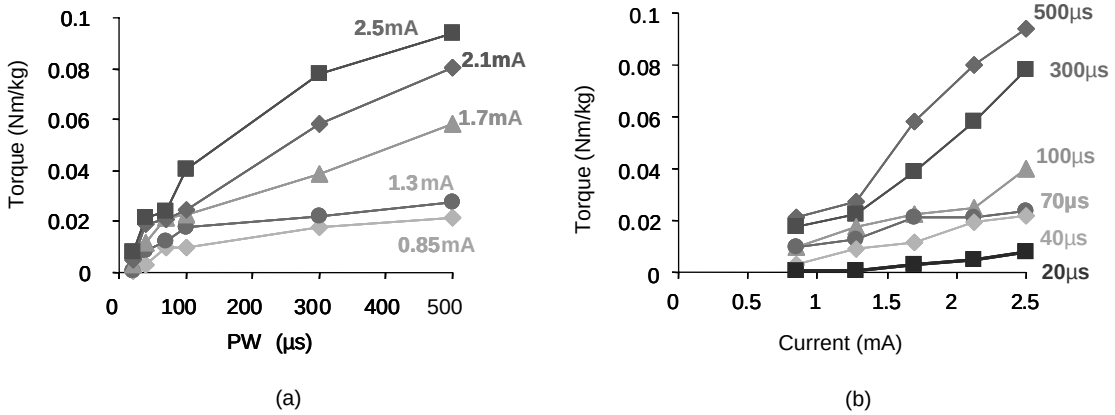


Figure 1. Typical force recruitment curves obtained from the ankle dorsiflexor muscle (Tibialis anterialis) of a rat through intramuscular stimulation. The recruitment curves indicate two techniques of force–torque modulation (a) pulse width modulation (PWM) and (b) pulse amplitude modulation (PAM). Single, symmetric, charge balanced, biphasic (cathodic first) pulses at an interval of 60 s were delivered. The currents were chosen as multiples of the twitch threshold current at 40 μs.

recruitment characteristics quite different from the normal physiological mechanisms. With electrical stimulation, physiological muscle force regulation is controlled either by spatial summation or by temporal summation (72). Spatial summation (or electrical recruitment) is achieved by increasing the pulse width (Fig. 1a) and/or the pulse amplitude (Fig. 1b) of the electrical stimulus—extending the excitatory extracellular potential distribution further out from the stimulating electrode(s) to greater numbers of nerve fibers, and/or longer in time. Force recruitment curves are in general quite nonlinear. The isometric recruitment curve (IRC) of a muscle can be defined as the static gain relation between stimulus level and output force/torque when the muscle is held at a fixed length. The features of a typical IRC are an initial deadzone region, a high slope, monotonically increasing region, and a saturation region (73,74). These features can be explained by recognizing that the slope of the IRC is primarily a function of the electrode–nerve interface. The shape is dictated by the location and size distributions of the individual motor unit axons within the nerve with large diameter axons having a lower stimulus activation threshold than small diameter axons. The IRC depends on the past history of muscle activation and location of the electrode relative to the motor point. The motor point functionally is defined as the location (on the skin surface, or for implanted electrodes on the muscle overlying its innervation) where stimulation thresholds are lowest for the desired motor response. There is a drop in the maximum magnitude and slope of the monotonic region of the IRC on muscle fatigue (73,75). The IRC is also influenced by the muscle length tension curve (76) and, if muscle force is estimated by measuring joint torque, by the muscle nonlinear moment arm as it crosses the joint. Because of these factors, the IRC shape will be different for each muscle and set of experimental configurations and will also vary between subjects.

Temporal summation (also called rate modulation) varies the stimulus frequency or the rate of action potential firing on the nerve fiber(s). When electrodes are located closer to the motor point for stimulation, enhanced spatial selectivity can be achieved because the electric field introduced can be focused closer to the α motor neuron fibers of interest. Another aspect of recruitment selectivity is fiber diameter, which relates to the tendency to stimulate subpopulations of nerve fibers based on their size. In electrical stimulation of myelinated fibers, there will be a tendency to recruit large axons at small stimulus magnitudes and then smaller axons with increased stimulus levels unlike during normal physiological recruitment—this is often dubbed reverse recruitment (77–79). Such reversed recruitment of motor units will inappropriately utilize fast, more readily fatigued muscle fibers for low force tasks. Slower fatigue resistant muscle fibers will only be recruited at higher stimulus levels. This also results in an undesirable steep relation between force output and stimulus magnitude. After injuries causing paralysis and disuse of muscle, many fatigue resistant muscle fibers tend to shift their metabolism toward less oxidative and more anaerobic, more readily fatigued mechanisms. Electrical stimulation therapy in such instances will recruit the faster muscle fibers first thereby inducing fatigue at a very early stage in the therapy.

FES DEVICES AND SYSTEMS

As illustrated in Fig. 2, all modern FES and FNS devices and systems incorporate (*1*) surface or implanted electrodes to generate an excitatory electric field within the body, (*2*) a regulated-current or regulated-voltage output stage that delivers stimulus pulses to the electrodes, (*3*) the stimulator pulse conditioning circuitry that creates the desired pulse shape, amplitude, timing, and pulse delivery (often within trains of pulses at set frequencies and for intended intervals), and (*4*) an open- or closed-loop stimulator controller unit. Systems may be completely or partially implanted and often incorporate a microcontroller or computer interface. Smith and colleagues at the Cleveland FES Center, for example, have developed an externally

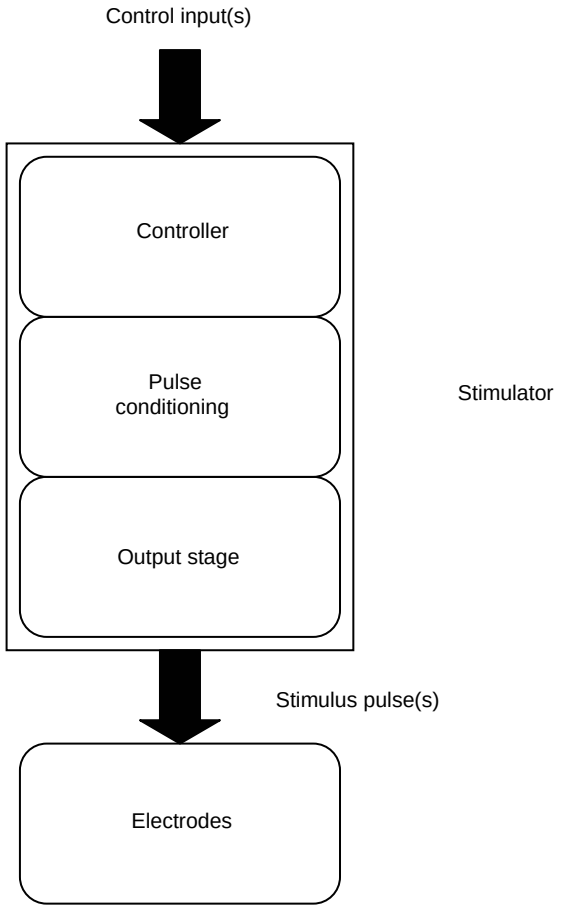


Figure 2. The FES systems typically incorporate control signals from the user that a Controller stage acts upon. Patterns of stimulation pulses are shaped with a pulse conditioning module that in turn feeds pulse information to an output stage that delivers regulated-current or regulated-voltage pulses of the desired amplitudes and timing to one or more channels of electrodes which are in contact with, or implanted within, the body.

powered, multichannel, implanted stimulator with telemetry for control of grasp and release functions in individuals with cervical level (C5 and C6) spinal cord injuries (80). Wu et al. designed a PC-based LabView controlled multichannel FES system with regulated-current or regulated-voltage arbitrary stimulation waveform pattern capability (81).

Commercialized FES systems include, for example, the Bioness, Inc. H200/Handmaster. This U.S. Food and Drug Administration (FDA) approved device incorporates microprocessor controlled surface stimulation into a portable, noninvasive hand–wrist orthosis for poststroke rehabilitation [see, e.g., (82)]. The FreeHand System, commercialized by NeuroControl Corporation in Cleveland, implements implanted receiver-stimulator, external controller, electrode, and sensor technologies (Fig. 3) developed through the Cleveland FES Center into a system for restoration of control of hand grasp and release for C5/C6 level spinal cord injured individuals. Compex Motion (Fig. 4), a programmable transcutaneous electrical stimulation product of Compex SA, is designed as a multipurpose FES system for incorporation into rehabilitation therapies (83). The Parastep System developed by Sigmedics, Inc. is designed to enable independent, unbraced standing and walking for spinal cord injured people. Parastep is a noninvasive system that incorporates a battery-powered, microcomputer controlled stimulator unit (Fig. 5), surface electrodes, and a control and stability walker with finger activated control switches.

Electrode Designs for Electrical Stimulation

In the implementation of FES and FNS techniques, surface or implanted electrodes are used to create an excitatory electric field distribution within the targeted tissues. Researchers over the years have identified a number of important criteria for stimulation electrode selection and have developed a variety of electrode designs in order to meet specific application requirements (for an excellent recent review see Ref. 84).

Criteria for Electrode Selection. A few of the important factors identified for long-term applications are anatomical and surgical factors, mechanical and electrochemical characteristics, biocompatibility, long-term stability, and economics. Anatomical and surgical factors include ease of identification of stimulation site, either on the skin surface or through implantation. In the event of damage to the electrode, any implanted region should be easily accessible for retrieval and replacement. The mechanical properties of electrodes are important particularly with respect to implants whose lifetime is measured in years. Electrodes that are flexible, and consequently smaller in diameter, induce less trauma to muscles during movement. Instead of straight wires, coiled electrode wires provide for greater tension, and reduce the stress. The use of multistranded wires reduces breakage or provides redundancy if some wires should fail.

The electrical stability of the electrode is usually judged based upon reproducibility of muscle force recruitment curves. These depict some stimulation parameter (e.g., pulse width or current) against muscle force or torque output. As we have seen, the normal order of recruitment is generally reversed (larger motor units are activated before smaller ones). The threshold and the steepness of the curve are important properties that vary with electrode design, fiber size, and strength duration relations.

Another important criterion of consideration for choice of electrodes that are chronically implanted and tested over time is biocompatibility. The charge carriers in the electrode material (metal) are electrons unlike in our body wherein the charge carriers are ions. This results in a change of charge carriers when currents cross the metal–body interface. A capacitive double layer of charge arises at the metal–electrolyte interface; the single layer in the metal arises because of its connection to the battery, whereas that in the electrolyte is due to the attraction of ions in the electric field (85,86). These layers are separated by the molecular dimensions of the water molecule so the effective capacitance (being inversely proportional to charge separation) is quite high. At sufficiently low levels, the current will be primarily capacitive. But for high currents that exceed the capabilities of the capacitance channel, irreversible chemical reactions will take place

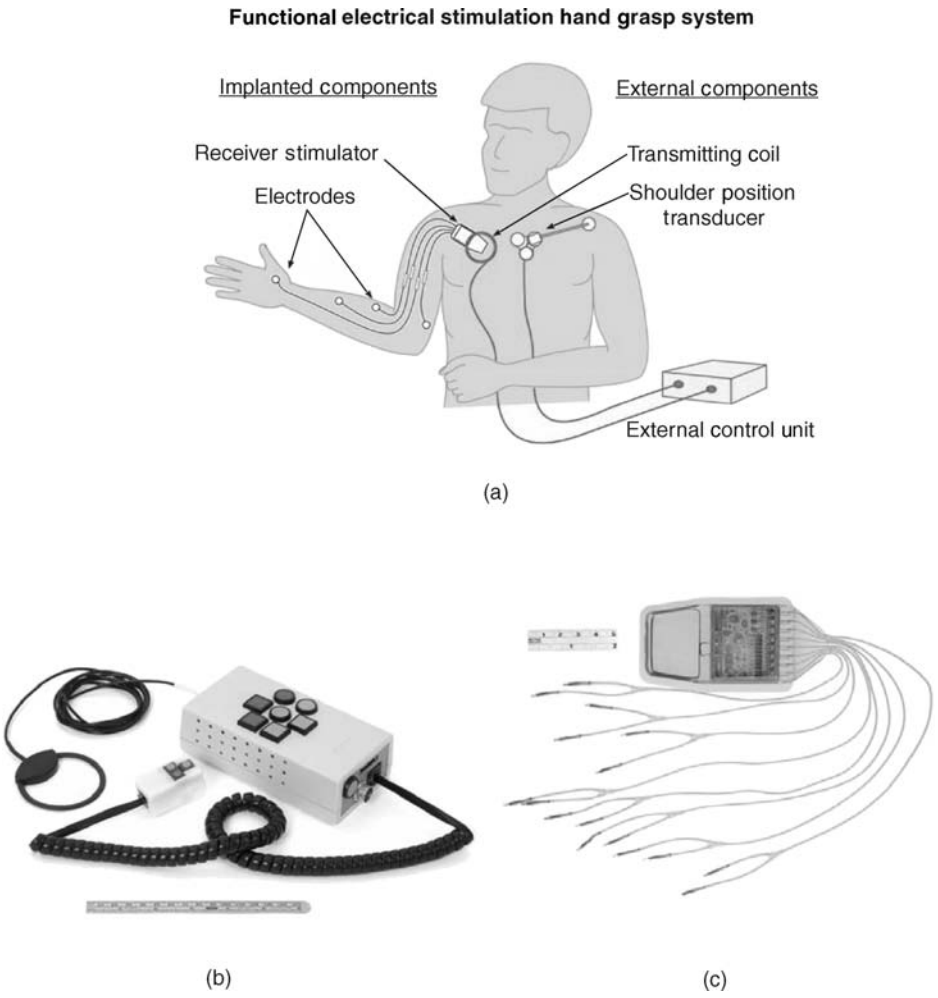


Figure 3. (a) Diagram of components for the implanted stimulation system developed at the Cleveland FES Center and commercialized as the Freehand neuroprosthesis by NeuroControl Corp. In the hand-grasp example shown, shoulder position is transduced for use as the command input. (b) The external control unit (ECU) provides the transducer interface, user control algorithm, multichannel stimulus coordination, and power for the implanted receiver-stimulator system. (c) The implanted receiver-stimulator provides multiple channels of stimulus output via the leads seen in the figure. It also transmits implantable sensor data to the ECU, and is powered through an inductive link that forms a coreless bidirectional transformer. Intramuscular or epimysial electrodes implanted in the forearm or hand are attached to the stimulator leads (not shown). (Courtesy of the Cleveland FES Center.)

that are undesirable since they are detrimental to the tissue or electrode or both. Therefore, the electrode material must have little impact on the electrochemistry at the electrode–tissue interface. For biocompatibility and to avoid local tissue damage induced by high current levels, the electrode materials used are essentially inert (e.g., platinum, platinum–iridium, and 316LVM stainless steel).

The above mentioned criteria for electrode selection are a general guideline for either skin surface or chronically implanted electrode systems. However, the choice of electrode is also application dependent. For example, during stimulation of the brain, of particular concern is prevention of breakdown of the blood–brain barrier. For nerve stimulation circular (82) electrodes can be placed within an insulating cuff; consequently, smaller amounts of current are required because the field is greatly confined. Also, lower current tends to minimize unwanted excitation of surrounding tissue. Finally, intramuscular electrodes, because of the implant flexing that must be withstood, are usually of the coiled-wire variety discussed above.

Electrode Classification. In general, electrodes designed to deliver electrical pulses to excitable tissue are classified based on the site of stimulation or placement of electrodes. Motor nerves can be stimulated through electrodes placed on the surface of the skin (surface electrodes) or implanted within the body. Implanted electrodes include those placed on or in the muscle (epimysial or intramuscular electrodes, respectively); as well as within or adjacent to a motor nerve (intraneural or extraneural electrodes). Electrodes that stimulate the spinal cord and BIONs (electrodes integrated with sensing and processing and packaged into a capsule) are recent additions to the family of implanted electrode technologies. The above classification of electrodes is further described below and summarized in Table 2.

Surface Electrodes. Surface electrodes as the name implies are placed on the surface of the skin and are the earliest of the electrodes to be used for applications in electrotherapy. These consist of conductive plates and are available in many types including conductive rubber patches coated with electrolyte gel, metal plates contacting the skin via thin, moist sponges and flexible, disposable, stainless steel mesh or rubber electrodes with self-adhesive conductive polymers (98–100). They do not need any implantation and are therefore noninvasive and relatively easy to apply and replace. An excellent description on the placements of these electrodes can be found in the Rancho Los Amigos Medical Center's practical guide to neuromuscular electrical stimulation (101). Surface electrodes

Figure 4. The Compex Motion FES system, manufactured by the Swiss based company Compex SA, is a general purpose programmable transcutaneous electrical stimulation device. Seen are the stimulator unit, three memory chip-cards that are inserted into the stimulator and used to store all pertinent information for a specific protocol, two EMG sensors, and two surface electrodes. (Reprinted from Ref. 83 with permission from the Institute of Physics and Engineering in Medicine.)

Figure 5. The neuromuscular stimulation unit for the Parastep system manufactured by Sigmedics, Inc. is battery-powered and microcomputer controlled. Cables connect the unit to surface electrodes, as well as to finger activated control switches on a walker. (Courtesy of Sigmedics, Inc.)

do have some disadvantages. They offer relatively poor selectivity for stimulation, have elevated threshold levels, may activate skin pain receptors, and do not have highly reproducible positioning capability. When higher currents are delivered to stimulate deeper muscles, spill over of charge to the nontargeted superficial muscles occurs. It is sometimes difficult to anchor surface electrodes in moving limbs and electrical properties at the skin–electrode interface can be variable.

Surface electrodes have been used for both lower limb and upper limb motor prosthesis, including the aforementioned Parastep system for ambulation (Fig. 6). WalkAid was designed for the management of foot drop to help toe clearance during the swing phase of walking (102). A single channel stimulator, the Odstock Dropped Foot Stimulator (ODFS) and later a two channel stimulator (O2CHS) designed for foot drop correction, used self-adhesive skin surface electrodes placed on the side of the leg (103,104). MikroFES was another orthotic stimulator for correction of foot drop in paralyzed patients (9). The Hybrid Assist System (HAS) (105) and the RGO system (106) use surface stimulation along with braces. Upper extremity applications include the Handmaster (107), the Belgrade Grasp System (BGS) (108), and the Bionic Glove (109) which focus on improving hand grasp.

Implanted Electrodes. Implanted electrodes can either be in direct contact with a muscle or peripheral nerve, within a muscle and only separated by muscle tissue from the motor nerves innervating the muscles, or within the spinal cord. Since peripheral electrodes are closer to the motor nerves than surface electrodes, they allow for better selectivity and more repeatable excitation. Their positioning and implantation is more permanent. Implanted electrodes have the advantage of place and forget by comparison to surface electrodes. That is, once the system is implanted, the user potentially can forget it is there. The chances of spill over are reduced since the electrodes can be placed close to the target muscle or nerve. The sensation to the user is usually much more comfortable as the implantation is away from the cutaneous pain receptors and the threshold current amplitude is lower. However, the implant procedure is invasive and in case of implant failure an invasive revision procedure can be required. Improper design and implantation can lead to tissue damage and infection. Insufficient tensile strength, high threshold levels, highly nonlinear recruitment curves, poor selectivity of activation and repeatability and adverse pain sensation (110–112) indicate failure. Excess encapsulation and infection (113); mechanical failures of electrode lead breakage and corrosion of electrodes and the insulator (114,115) can also impair the system.

Electrodes in or on the Muscle: Intramuscular and Epimysial Electrodes. Implanted electrodes that are placed on or in the muscle consist of intramuscular (87,88,116–121) and epimysial electrodes (89,122–125). Intramuscular electrodes (88,126) can, for example, be fabricated from multistranded Teflon coated stainless steel wires. This configuration provides good tensile strength and flexibility. They are implanted by injecting a hypodermic needle

Table 2. Electrical Stimulation Electrode Classifications and Types

Location/Type	Features and Advantages	Example	References
Surface	Metal plate with electrolyte gel, noninvasive	WalkAid, ODFS, MikroFES, HAS, RGO, Handmaster, BGS, Bionic Glove	
In/On Muscle	lower thresholds and better selectivity compared to surface electrodes		
Intramuscular	Implanted in the muscle, multistranded Teflon coated stainless steel wire, monopolar and bipolar configurations, good tensile strength, and flexibility		87,88
Epimysial	Implanted under the skin: on the muscle, monopolar and bipolar configurations, less prone to mechanical failure		89
BIONs	Injected into or near the muscle, hermetically sealed glass/ceramic capsule integrated with electronics		90
Near/On Nerve	Lower threshold levels and better selectivity than the above mentioned electrodes		
Nerve Cuffs	Monopolar, bipolar and tripolar configurations, good power efficiency, improved selectivity, comparatively stable		91,92
FINE	Reshape or maintain nerve geometry		93
Intrafascicular	Penetrate the epineurium and into the fascicle, selective stimulation, lower current and charge levels		
LIFE	Stable, suitable for stimulating and recording		94
SPINE	Reduced nerve damage		95
Intraspinal			
Microwires	Near to normal recruitment, reduced fatigue, highly selective stimulation		96,97

Figure 6. Examples of self-adhesive, reusable surface electrodes. The electrodes shown are used in the Parastep neuromuscular stimulation system. (Courtesy of Sigmedics, Inc.)

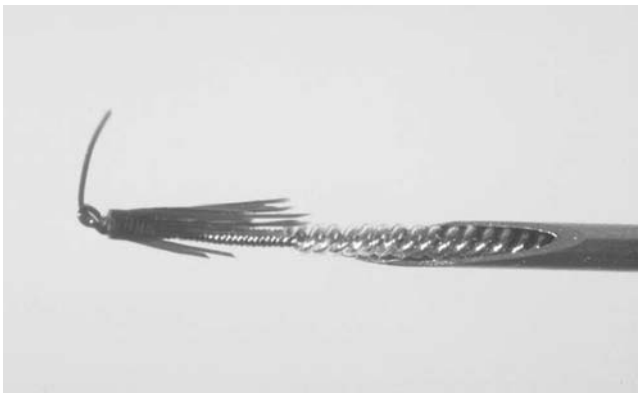


Figure 7. A "Peterson" type intramuscular electrode design. This is a helically wound PFS insulated multistranded 316LVM stainless steel wire design that is attached to a barb-like anchoring structure constructed of polypropylene suture material. The wound section of the electrode is $\sim 800\,\mu m$ in diameter and is partially loaded into a hypodermic needle. (Courtesy of J.T. Mortimer and reproduced by permission of World Scientific Publishing Co.)

either nonsurgically or through an open incision. A fine needle probe used by itself or in conjunction with a surface probe is used to detect the motor point; the motor point for an intramuscular electrode is usually just below the muscle surface beneath the motor point position as defined by surface electrode. These electrodes can elicit a maximal muscular contraction with only $\sim$ 10% of the stimulus charge required by equivalent surface electrodes (25). Figure 7 depicts a Peterson type intramuscular electrode developed at Case Western Reserve University (121).

Both monopolar and bipolar intramuscular electrodes have been used. Bipolar intramuscular electrodes that straddle the nerve entry point can be as effective at activating the muscles as a nerve cuff. If bipolar electrodes do not straddle the nerve entry point, full recruitment of the muscle can require large stimulation charge and stimulation cannot be achieved without activating the surrounding muscles. In contrast, monopolar stimulation is less position dependent, though it cannot match the selectivity obtained with good bipolar placement (127). The size of the

muscle will determine the limit of electrode size, although large electrodes are more efficacious.

A recent development in the intramuscular stimulating electrode world are BIONs (for BIOnic Neurons), that can potentially provide precise and inexpensive interfaces between electronic controllers and muscles (90). The BIONs consist of a hermetically sealed glass–ceramic capsule with integral capacitor electrodes for safety and reliability (128). The internal electronics include an antenna coil wrapped around a sandwich of hemicylindrical ferrites over a ceramic microprinted circuit board carrying a custom integrated circuit chip. In animal studies, these electrodes have demonstrated long-term biocompatibility (129) and ability to achieve selective muscle stimulation (130). The first generation of BIONs, BION1, generates stimulation pulses of 0.2–30 mA at 4–512 μs duration. This system is now in clinical trials to provide therapeutic electrical stimulation to patients with disabilities (131–135). The second generation BION, BION2, is under development. BION2s are expected to sense muscle length, limb acceleration and bioelectrical potentials for feedback control in FES (136–138).

Intramuscular electrodes have been used to activate paralyzed muscles that retain a functional motor neuron in the muscles of the upper extremity (139,140), lower extremity (118,140,141) and the diaphragm (142). Muscles also have been stimulated to correct spinal deformities in the treatment of scoliosis (143).

Epimysial electrodes (89,110) are positioned on the surface of a muscle below the skin but not within the muscle. They have a smooth circular disk on one side and a flat, insulating backing, reinforced with mesh. The motor point is usually identified by moving a stimulating electrode across the muscle surface to locate the surface position that requires the least amplitude to fully excite the muscle. Replacing this electrode in the event of failure is comparatively easier. The stimulation levels and impedance are also similar to that of intramuscular electrodes. A perceived advantage of epimysial electrodes over intramuscular electrodes is that they are less prone to mechanical failure and less likely to move in the hours and days immediately after implantation.

Epimysial electrodes also can be used either in the monopolar mode or the bipolar mode (89,108,119,120,123). Use of a monopolar epimysial electrode close to the motor nerves results in reduced threshold stimulus amplitude, higher gain and selectivity, and decrease in length dependent recruitment. When a bipolar epimysial electrode is used, the stimulus current is constrained to regions closer to the two electrodes. Compared to the results with monopolar electrodes, the threshold is increased, relative gain decreased, and though greater selectivity is found with stimulation current levels close to twitch threshold poorer selectivity is present in the stimulus range needed for maximum activation of the muscle (108).

Epimysial electrodes have been used for a number of years in the implementation of upper extremity assist devices for C5 or C6 adult subjects with tetraplegia (Fig. 8), including incorporation into the FDA approved FreeHand System (144) and more recently for providing the capability of standing after paraplegia (117).

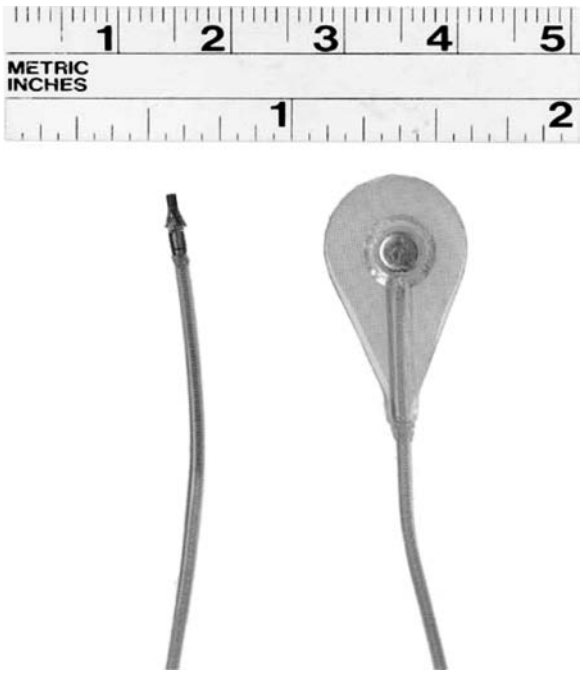


Figure 8. An example implantable epimysial electrode (right) with intramuscular electrode (left), typical of those used with the Cleveland FES Center's implanted hand-grasp system. (Courtesy of the Cleveland FES Center.)

Implanted Nerve Electrodes. Electrodes that are placed in contact with the nerve include extraneural and intraneural electrodes. Extraneural electrodes do not penetrate the epineurium and include varying designs of nerve cuffs (91,92,145–149) and the recently investigated flat interface nerve electrodes (FINE) (93,150,152). Intraneural electrodes penetrate the epineurium and include intrafascicular and interfascicular electrodes (94,95,153–157). Nerve electrodes have several potential advantages over intramuscular electrodes—including, lower power requirements, the ability to control several muscles with a single implant, and the ability to place the electrodes far from contracting muscles (158).

Electrodes placed on the surface of the nerve, and housed in an insulative carrier that encompasses the nerve trunk, are cuff electrodes (91,151,159,160). The cuff material is often silicone rubber and sometimes reinforced with Dacron. Cuff-type electrodes hold the stimulating contacts in close proximity to the nerve trunk. Holding the target tissues close to the stimulating contacts offers opportunities for power efficiency and improved selectivity. Less power is spent on electrical conduction through space between the electrode and target tissues. Improved selectivity is possible because the electric potential gradient is larger when the spacing between the stimulating contact and the target tissue is least. Further, these electrodes are less likely to move in relationship to the target tissues after implantation (161–164). However, while nerve cuffs stimulate effectively and selectively they require invasive surgery for implantation. They may also damage the nerves they enclose unless carefully designed, sized, and implanted.

Figure 9. A self-sizing cuff electrode design fabricated using PMP (polymer–metal–polymer) technology and laser machining. (Courtesy of J.T. Mortimer and M. Tarler.)

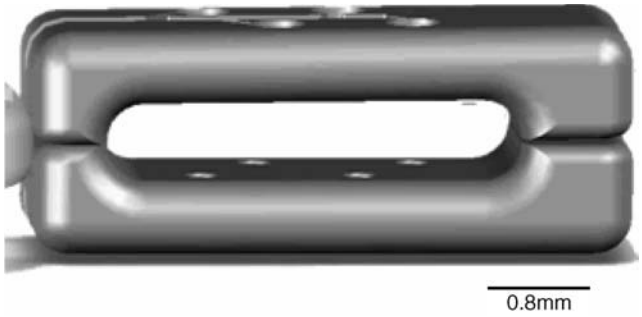


Figure 10. The FINE nerve cuff design, intended to flatten peripheral nerve trunks into a layering of nerve fascicles. Electrode contacts are seen as small dots within the overall structure. (Courtesy of D. Durand.)

To overcome potential problems with a fixed cuff-size, nerve cuff electrodes have been designed with different configuration. The Huntington nerve cuff (165), is a helix-type nerve electrode system that has exposed metal sections as stimulating contacts along the internal diameter of the helix. The open helix design can accommodate some swelling. Other self-sizing cuff electrode designs sometimes have a spiral configuration that enables opening or closing to accommodate a range of different diameter nerves (91). Figure 9, for example, is a photo of a self-sizing nerve cuff fabricated at Case Western Reserve University using PMP technology and laser machining. Both cuff and spiral electrode configurations can be used in various monopolar, bipolar or tripolar configurations (91,164). Cuff electrodes with multiple electrical contacts can produce selective activation of two antagonistic muscle groups innervated by that nerve trunk (166). Increased function and additional control of muscles with minimum number of electrodes can be achieved. Self-sizing nerve-cuff electrodes, with multiple contacts in a tripolar configuration, have been shown to produce controlled and selective recruitment of some motor nerves in a nerve trunk (145,158,167–170). A monopolar electrode with four radially placed contacts can work as well as a tripolar electrode with four radially placed tripoles (171,172). A four contact self-sizing spiral cuff electrode has been described as a tunable electrode that is capable of steering the excitation from an undesirable location to a preferred location (92).

The flat interface nerve electrode, or FINE system as seen in Fig. 10, has been introduced in an attempt to improve the stimulation selectivity of extraneural electrodes (151). The goal with the FINE is to create a geometry that optimizes stimulation selectivity. In contrast to cylindrical electrodes, the FINE either reshapes the nerve into, or maintains the nerve in, an ovoid geometry. Chronic studies in rats have demonstrated that nerves and fascicles can be safely reshaped (150,173). Also, acute experiments and finite element models have demonstrated that it is possible to selectively activate individual fascicles in the cat sciatic nerve using this electrode (151,152,174). This could be important in both reducing fatigue and selectively activating individual muscles (153,175). A potential disadvantage is that a fibrous capsule with electrical properties different from the surrounding tissues will envelope the electrode (176,177), potentially rendering the recruitment properties unstable, although a recent study has shown that both selectivity measurements and the recruitment curve characteristics can remain stable for a prolonged implant period (93).

Intraneural electrodes are positioned to penetrate the epineurium around the nerve trunks. Intraneural electrodes utilize a conductor that invades the epineurium. Maximal contraction is elicited at stimulation levels an order of magnitude lower than with nerve cuff electrodes (200 μA, pulse duration 300 μs). However, connectors, fixation, and neural damage are still not completely resolved to allow routine clinical usage. Intraneural multipolar sword type electrodes have been made out of solid silicon with golden contacts and can be very selective (178). Such electrodes could minimize the needs for using many electrodes for activation of different muscles that are innervated from a single nerve (179).

A subset of intraneural electrodes are meant to enter the perineurium around the fascicles and go between the nerve fibers: These are so-called intrafascicular electrodes. Intrafascicular electrodes place stimulating elements inside the fascicles, in close proximity to axons (126,153,160,175,178,180,181). They have been shown to produce axonal recruitment with almost no excitation of muscles that are not targeted (181). A variation of the intrafascicular electrode is the longitudinal intrafascicular electrode (LIFE) (94,153). Compared with extraneural electrodes, LIFEs have many advantages and can be implanted into any of the fascicles of peripheral nerves to selectively stimulate a single fascicle thereby offering highly selective stimulation. Also they serve as excellent recording electrodes. When LIFEs are used as recording electrodes, the amplitudes of motor evoked potentials (MEPs) recorded by LIFEs implanted in fascicles are much larger than those of EMGs recorded from the skin by surface electrodes and the signals recorded are not affected by external electrical fields (155,182). Therefore, the signals recorded by LIFE can be used to control a prosthetic limb more accurately than those controlled by EMGs (183). In addition, LIFEs have excellent biocompatibility with peripheral fascicles (156,184,185).

While intrafascicular electrodes can provide high degrees of selectivity, it remains unclear whether penetrating the perineurium will lead to long-term nerve injury (126,186). Interestingly, an intraneural electrode system dubbed the slowly penetrating interfascicular electrode (SPINE) has been developed, which has been reported to penetrate a peripheral nerve within 24 h without evidence of edema or damage of the perineurium and showed functional selectivity (95).

In general, compared to externally placed electrodes, the current and charge stimulation requirements for intraneural electrodes are low since they are positioned inside the nerve trunk to be excited. Also, the stimulation selectivity is high compared to extraneural electrodes where stimulation selectivity suffers from the relatively large amount of tissue interposed between the stimulating contacts and the target axons.

Micro wires: Electrodes for Intraspinal Stimulation

Spinal circuits that are shown to have the capacity of generating complex behaviors with coordinated muscle activity can be activated by intraspinal electrical stimulation (187–190). Microwires that are finer than a human hair have been used to stimulate the spinal cord neurons to control single muscles or small group of synergists (96,97,191–193). Stimulation through single wires in a few sites has been shown to have the ability to elicit whole-limb activation sufficient to support the animal's weight (191,192,194–196). The stimuli were not perceived but were able to produce strong coordinated movements. Near normal recruitment order, minimal changes in kinematics and little fatigue and functional, synergistic movements induced by stimulation in the lumbosacral cord (97,194,196) are some of the promising advantages of stimulating the spinal cord with microwires. However, the clinical and long-term feasibility of implanting many fine microwires into the spinal cord remains questionable. In addition, stimulating the spinal cord results in steep recruitment curves compared to muscle and nerve stimulation thereby limiting the degree of control achievable.

Controllers and Control Strategies

Besides stimulating the paralyzed muscles, it is also important to control and regulate the artificial movements produced. The control task refers to specification of the temporal patterns of muscle stimulation to produce the desired movements; and the regulation task is the modification of these patterns during use to correct for unanticipated changes (disturbances) in the stimulated muscles or in the environment. A major impediment to the development of satisfactory control systems for functional neuromuscular stimulation has been the nonlinear, time varying properties of electrically activated skeletal muscle that make control difficult to achieve (7,76,197). With FNS, the larger, fatigable muscle fibers are recruited at low levels of stimulation before the more fatigue-resistant fibers are activated thereby inducing rapid fatigue (56). It is important that the output of any FNS control system results in stable, repeatable, regulated muscle input–output properties over a wide range of conditions of muscle length, electrode movement, potentiation, and fatigue. To improve control strategies to provide near physiological control, inherent muscle characteristics (force-activation, force-length, and force-velocity), muscle modeling studies, studies on understanding how to model the patterns of neural prostheses and how neural prostheses respond to disturbances have been performed (197–200).

As depicted in Fig. 11 (201), FNS control methods include feedforward (open-loop), feedback, and adaptive control. Feedforward control requires a great deal of information about the biomechanical behavior of the limb. The control algorithms specify the stimulus parameters (musculoskeletal system inputs) that are expected to be needed to produce the desired movement (system outputs). In an open-loop control system these parameters are often identified by trial and error (6,13,202–205). The same stimulation pattern, which is often stored in the form of a lookup table, is delivered for each cycle of movement.

Three major problems exist with this form of fixed-parameters, open-loop control (204–206). First, the process of specifying the parameters for a single stimulation pattern for a single user often requires several extensive sessions involving the user, therapist, physician, and engineer. This process is often expensive, time consuming, and often only minimally successful in achieving adequate performance. Second, the fixed parameter stimulation pattern may not be suitable after muscles fatigue that is exacerbated by the stimulation paradigm itself. The third problem is that the open-loop stimulation pattern does not respond to changing environments (e.g., slope of walking surface) and external perturbations (e.g., muscle spasms).

To address the limitations of open-loop control systems feedback control was implemented (12,14,207,208). In a feedback control system, sensors monitor the output and corrections are made if the output does not behave as

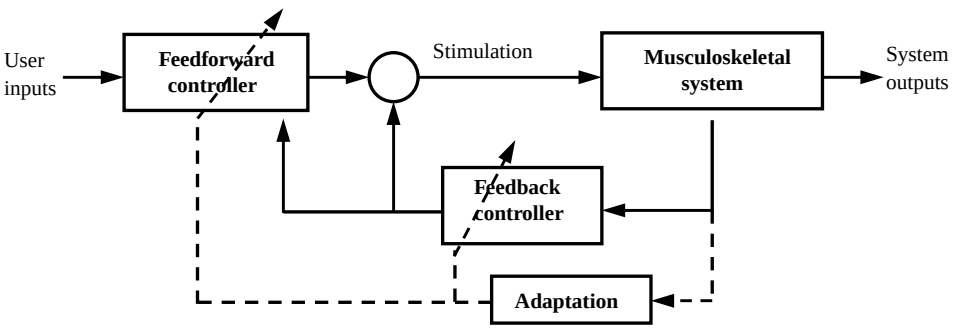


Figure 11. A representation of FNS control system components and strategies (feedforward, feedback and adaptive). (Reproduced by permission from Neuromodulation 2001;4: 187–195.)

desired. The corrections are made based on a control law, which is a mathematical prescription for how to change the input to reduce the difference (error) between the desired output and the actual output. Feedback control requires output sensors, and compensation is generally slower than in feedforward control since an output error must be present to generate a controller response. Thus feedback control might best be used for slow movements and for maintaining a steady posture. Since the output of the feedback controller is highly dependent on sensor signals, the quality of the control that is achieved will be compromised by the relatively low quality of sensors that are available. Feedback control has been successful in regulating hand grasp (209) and standing posture (12), but it appears that another strategy, adaptive feedforward control, is likely to be required for dynamic activities such as locomotion.

To improve performance of feedback control systems, adaptive control strategies were developed that automatically adjusted the overall system behavior (i.e., the combined response of the controller and the system) so that it is more linear, repeatable, and therefore predictable (75,210–213). These techniques adjust the parameters of the control system and attempt to self-fit the system to the user in order to make it easier to use and learn to use (206,212,214). The control system developed by Abbas and Chizeck has a pattern generator (PG) and a pattern shaper (PS) (211,215). The PG generates the basic rhythm for controlling a given movement. The PS adaptively filters those signals and sends its output to the muscles. The adaptive properties of the PS provide the control system with the ability to customize stimulation parameters for a particular individual and to adjust them on-line to account for fatigue. In some of the computer simulation experiments a proportional-derivative feedback controller was also active. Studies have shown that the pattern generator/pattern shaper (PG/PS) adaptive neural network controller is able to account for nonlinear and dynamic system properties and muscle fatigue (73,75,213). To summarize, adaptive control systems have replaced other developed control system strategies because this strategy can (*1*) provide the ability to automatically customize the stimulation pattern for a given user, (*2*) automatically adjust stimulation parameters to account for fatigue, and (*3*) automatically adjust to allow the voluntary motor commands to recover control of the movement pattern (in the case of partial recovery in a person with an incomplete spinal cord lesion).

Apart from the above other strategies, such as fuzzy logic (216) and proportional–integral–derivative (PID) controllers (217) have also been implemented to investigate automatic fatigue compensation. However, fatigue remains one of the major factors limiting utility of FES/FNS because such adaptive systems can adjust for fatigue only up to the contractile limits of the muscle.

Rather than initiating and modulating control of FES systems indirectly through residual motor function (e.g., as in the Freehand system for grasping, where paralyzed hand closure and opening were command controlled through sensing of opposite shoulder position), future FES devices might be controlled directly through thought—by tapping into the subject's remaining cortical intent to move via a brain–machine interface (BMI) [or sometimes brain–computer interface (BCI)]. So-called direct brain–machine interfaces utilize arrays of intracortical recording electrodes to sense action potentials from a host of individual neurons in regions of the brain where cells code for movement and its intent. A number of research teams have in recent years demonstrated the feasibility of recording and processing movement related signals from cortex (in both animals and in humans), and then enabling the subject to control computers or devices directly through such processed thought (218–220). Ultimately, BMI technologies hold promise that paralyzed individuals might one day be able to control FES devices for movement restoration with little or no effort or learning other than forming the simple intent to move (221).

THERAPEUTIC EFFECTS OF ELECTRICAL STIMULATION

While this article is focused mainly on electrical stimulation therapies for restoring lost function, it is important to recognize that electrical stimulation techniques are used also for therapeutic reasons. A recent review summarizes the current state of therapeutic and neuroprosthetic applications of electrical stimulation after spinal cord injury and identifies some future directions of research and clinical and commercial development (222). Functional electrical stimulation therapy individually and in combination with other rehabilitation therapies also is being utilized after incomplete spinal cord injury to influence the plasticity within the nervous system for improved recovery (9,223–228).

Therapeutic electric stimulation (TES) can affect the restoration of muscle strength (229). Therapeutic electric stimulation in humans has been shown to prevent muscle atrophy thereby increasing muscle cross-sectional area, torque, and force (230–234). Such electrical therapy has been effective in reversing the increased fatigability associated with the change in fiber type in both animals (31–37) and humans (56,59–61,65–67) after spinal cord injury. Electrical stimulation has also been able to reduce spasticity among patients with neurological disorders (reference).

While osteoporosis has been prevented in the limbs of paralyzed individuals, in menopausal women, and in the elderly and fracture patients through electrical stimulation therapy (235–240), certain other studies have shown little or no change in bone density (235,241–244). These contradictory results suggest the importance of other characteristics, such as the stimulation patterns, specifications for training (intensity, duration, loading), and the time postinjury. Enhancing fracture–wound healing is another therapeutic application of electrical stimulation (245–249). The theory here is to attract negatively or positively charged cells into the wound area, such as neutrophils, macrophages, epidermal cells, and fibroblasts that in turn will contribute to wound healing processes by way of their individual cellular activities (250). Electrical stimulation may also play a role in wound healing through improved blood flow (251,252), prevent occurrence of pressure sores thereby improving general tissue health (253). A recent

review details all the theories suggested and experimental studies and clinical trials performed on wound healing through electrical stimulation (254).

Recent applications of electrical stimulation have also been successful in altering neural function. For example, deep brain stimulation (DBS) is being used to treat a variety of disabling neurological symptoms, most commonly the debilitating symptoms of Parkinson's disease (PD), such as tremor, rigidity, stiffness, slowed movement, and walking problems [for a review, see (255,256)]. Deep brain stimulation uses a surgically implanted, neurostimulator approximately the size of a stopwatch. The implanted device delivers electrical stimulation to targeted areas in the brain that control movement, blocking the abnormal nerve signals that cause tremor and PD symptoms. Vagal nerve stimulator (VNS), approved by the FDA in 1997 are used to treat patients with intractable epilepsy. These devices controls seizures by sending electrical pulses to the vagus nerve (257,258). Transcutaneous electrical nerve stimulation (TENS), wherein electrical signals are sent to underlying nerves, can relieve a wide range of chronic and acute pain (259). The TENS devices are small battery-powered stimulators that produce low intensity electrical signals through electrodes on or near a painful area, producing a tingling sensation that reduces pain. Chronic electrical stimulation of the GI tract has been found to be a potential therapy for the treatment of obesity (260–262). It is clear that in future development of electrical stimulation technologies many devices will be designed to achieve both therapeutic and functional outcomes.

ACKNOWLEDGMENT

This work was in part supported by NIH (NCMRR)–HD-40335.

BIBLIOGRAPHY

Cited References

1. Kilgore KL, Kirsch RF. Upper and lower extremity motor neuroprostheses. In: Horch KW, Dhillon GS, editors. Neuroprosthetics: Theory and Practice, New Jersey: World Scientific; 2003. pp 844–877.
2. Liberson WT, Holmquest HJ, Scot D, Dow M. Functional electrotherapy: stimulation of the peroneal nerve synchronized with the swing phase of the gait of hemiplegic patients. Arch Phys Med Rehabil 1961;42:101–105.
3. Lyons GM, Sinkjaer T, Burridge JH, Wilcox DJ. A review of portable FES-based neural orthoses for the correction of drop foot. IEEE Trans Neural Syst Rehabil Eng 2002;10(4): 260–279.
4. Moe JH, Post HW. Functional electrical stimulation for ambulation in hemiplegia. J Lancet 1962;82:285–288.
5. Bajd T, Andrews BJ, Kralj A, Katakis J. Restoration of walking in patients with incomplete spinal cord injuries by use of surface electrical stimulation-preliminary results. Prosthet Orthot Int 1985;9(2):109–111.
6. Krajl A, Bajd T. Functional Electrical Stimulation: Standing and Walking After Spinal Cord Injury. Boca Raton (FL): CRC Press; 1989.
7. Yarkony GM, Roth EJ, Cybulski G, Jaeger RJ. Neuromuscular stimulation in spinal cord injury: I: Restoration of functional movement of the extremities. Arch Phys Med Rehabil 1992;73(1):78–86.
8. Stein RB, et al. Electrical systems for improving locomotion after incomplete spinal cord injury: an assessment. Arch Phys Med Rehabil 1993;74(9):954–959.
9. Bajd T, Kralj A, Stefancic M, Lavrac N. Use of functional electrical stimulation in the lower extremities of incomplete spinal cord injured patients. Artif Organs 1999;23(5):403–409.
10. Stein RB. Functional electrical stimulation after spinal cord injury. J Neurotrauma 1999;16(8):713–717.
11. Peckham PH, Keith MW. Motor prostheses for restoration of upper extremity function., in Neural prostheses: Replacing motor function after disease or disability. New York: Oxford University Press; 1992. pp 162–190.
12. Chizeck HJ, et al. Control of functional neuromscular stimulation systems for standing and locomotion in paraplegics. Proc IEEE 1988;1155–1165.
13. Marsolais EB, Kobetic R. Development of a practical electrical stimulation system for restoring gait in the paralyzed patient. Clin Orthop 1988;233:64–74.
14. Abbas JJ, Chizeck HJ. Feedback control of coronal plane hip angle in paraplegic subjects using functional neuromuscular stimulation. IEEE Trans Biomed Eng 1991;38(7):687–698.
15. Solomonow M. Biomechanics and physiology of a practical functional neuromuscular stimulation walking orthosis for paraplegics. In: Stein RB, Popovic DP, editors. Neural Prostheses: Replacing motor function after disease or disability. New York: Oxford University Press; pp 202–232.
16. Graupe D, Kohn KH. Functional electrical stimulation for ambulation by paraplegics, in Functional electrical stimulation for ambulation by paraplegics. Krieger; 1994. p 194.
17. Mortimer JT. Motor Prostheses. In: Brookhart JM, Mountcastle VB, Brooks VB, Geiger SR, editors. Handbook of Physiology, Section 1: The Nervous System, Vol. II Motor Control, Part I. Bethesda (MD): American Physiological Society; 1981.
18. Henneman E, Somjen G, Carpenter DO. Functional Significance of Cell Size in Spinal Motoneurons. J Neurophysiol 1965;28:560–580.
19. Henneman E, Somjen G, Carpenter DO. Excitability and inhibitability of motoneurons of different sizes. J Neurophysiol 1965;28(3):599–620.
20. Burke RE. Firing patterns of gastrocnemius motor units in the decerebrate cat. J Physiol 1968;196(3):631–654.
21. Burke RE. Motor units: Anatomy, physiology and functional organization. In: Brooks VB, editor. Handbook of Physiology Section 1: The Nervous System. Vol. III. Motor Systems. Bethesda (MD): American Physiology Society; 1981. pp 345–422.
22. McPhedran AM, Wuerker RB, Henneman E. Properties of Motor Units in a Heterogeneous Pale Muscle. J Neurophysiol 1965;28:85–99.
23. Armstrong RB, Phelps RO. Muscle Fiber Type Composition of the Rat Hindlimb. Am J Anat 1984;171:256–272.
24. Staron RS. Human skeletal muscle fiber types: delineation, development, and distribution. Can J Appl Physiol 1997; 22(4):307–327.
25. Popovic D, Sinkjaer T. Control of Movement for the Physically Disabled. London: Springer-Verlag; 2003.
26. Ichihara K, et al. Muscle stimulation in a rodent model: electrode design, implantation and assessment. 9th Annual Conference of the International FES Society. Bournemouth (UK): 2004.

27. Brooke MH, Kaiser KK. Muscle fiber types:How many and what kind? Arch Neurol 1970;23:369–379.
28. Peter JB, et al. Metabolic profiles of three fiber types of skeletal muscle in guinea pigs and rabbits. Biochemistry 1972;11:2627–2633.
29. Burke RE, Levine DN, Tsairis P, Zajac FE. Physiological types of histochemical profiles in motor units of the cat gastrocnemius. J Physiol 1973;234:723–748.
30. Pette D, Staron RS. Cellular and molecular diversities of mammalian skeletal muscle fibers. Rev Physiol Biochem Pharmacol 1990;116:1–76.
31. Brown WE, Salmons S, Whalen RG. The sequential replacement of myosin subunit isoforms during muscle type transformation induced by long term electrical stimulation. J Biol Chem 1983;258(23):14686–14692.
32. Brownson C, et al. Changes in skeletal muscle gene transcription induced by chronic stimulation. Muscle Nerve 1988; 11(11):1183–1189.
33. Brownson C, Little P, Jarvis JC, Salmons S. Reciprocal changes in myosin isoform mRNAs of rabbit skeletal muscle in response to the initiation and cessation of chronic electrical stimulation. Muscle Nerve 1992;15(6): 694–700.
34. Carraro U. Contractile proteins of fatigue-resistant muscle. Semin Thorac Cardiovasc Surg 1991;3(2):111–115.
35. Kirschbaum BJ, Heilig A, Hartner KT, Pette D. Electrostimulation-induced fast-to-slow transitions of myosin light and heavy chains in rabbit fast-twitch muscle at the mRNA level. FEBS Lett 1989;243(2):123–126.
36. Pette D, Muller W, Leisner E, Vrbova G. Time dependent effects on contractile properties, fibre population, myosin light chains and enzymes of energy metabolism in intermittently and continuously stimulated fast twitch muscles of the rabbit. Pflugers Arch 1976;364(2):103–112.
37. Sreter FA, Gergely J, Salmons S, Romanul F. Synthesis by fast muscle of myosin light chains characteristic of slow muscle in response to long-term stimulation. Nat New Biol 1973;241(105): 17–19.
38. Pette D, et al. Partial fast-to-slow conversion of regenerating rat fast-twitch muscle by chronic low-frequency stimulation. J Muscle Res Cell Motil 2002;23(3):215–221.
39. Putman CT, et al. Fiber-type transitions and satellite cell activation in low-frequency-stimulated muscles of young and aging rats. J Gerontol A Biol Sci Med Sci 2001;56(12):B510–B519.
40. Jarvis JC. Power production and working capacity of rabbit tibialis anterior muscles after chronic electrical stimulation at 10 Hz. J Physiol 1993;470:157–169.
41. Mannion JD, et al. Histochemical and fatigue characteristics of conditioned canine latissimus dorsi muscle. Circ Res 1986;58(2):298–304.
42. Trumble DR, LaFramboise WA, Duan C, Magovern JA. Functional properties of conditioned skeletal muscle: implications for muscle-powered cardiac assist. Am J Physiol 1997;273(2 Pt. 1):C588–C597.
43. Salmons S, Vrbova G. The influence of activity on some contractile characteristics of mammalian fast and slow muscles. J Physiol 1969;201(3):535–549.
44. al-Amood WS, Buller AJ, Pope R. Long-term stimulation of cat fast-twitch skeletal muscle. Nature (London) 1973; 244(5413): 225–257.
45. Glatz JF, et al. Differences in metabolic response of dog and goat latissimus dorsi muscle to chronic stimulation. J Appl Physiol 1992;73(3):806–811.
46. Ferguson AS, et al. Muscle plasticity: comparison of a 30-Hz burst with 10-Hz continuous stimulation. J Appl Physiol 1989;66(3):1143–1151.
47. Jarvis JC, et al. Fast-to-slow transformation in stimulated rat muscle. Muscle Nerve 1996;19(11):1469–1475.
48. Mayne CN, et al. Induction of a fast-oxidative phenotype by chronic muscle stimulation: histochemical and metabolic studies. Am J Physiol 1996;270(1 Pt 1):C313–C320.
49. Sutherland H, et al. The dose-related response of rabbit fast muscle to long-term low-frequency stimulation. Muscle Nerve 1998;21(12):1632–1646.
50. Andersen JL, et al. Myosin heavy chain isoform transformation in single fibres from m. vastus lateralis in spinal cord injured individuals: effects of long-term functional electrical stimulation (FES). Pflugers Arch 1996;431(4):513–518.
51. Theriault R, Theriault G, Simoneau JA. Human skeletal muscle adaptation in response to chronic low-frequency electrical stimulation. J Appl Physiol 1994;77(4):1885–1889.
52. Gordon T, Pattullo MC. Plasticity of muscle fiber and motor unit types. Exerc Sport Sci Rev 1993;21:331–362.
53. Lenman AJ, et al. Muscle fatigue in some neurological disorders. Muscle Nerve 1989;12(11):938–942.
54. Rutherford OM, Jones DA. Contractile properties and fatiguability of the human adductor pollicis and first dorsal interosseus: a comparison of the effects of two chronic stimulation patterns. J Neurol Sci 1988;85(3):319–331.
55. Scott OM, Vrbova G, Hyde SA, Dubowitz V. Effects of chronic low frequency electrical stimulation on normal human tibialis anterior muscle. J Neurol Neurosurg Psychiat 1985; 48(8): 774–781.
56. Stein RB, et al. Optimal stimulation of paralyzed muscle after human spinal cord injury. J Appl Physiol 1992;72(4): 1393–1400.
57. Theriault R, Boulay MR, Theriault G, Simoneau JA. Electrical stimulation-induced changes in performance and fiber type proportion of human knee extensor muscles. Eur J Appl Physiol Occup Physiol 1996;74(4):311–317.
58. Currier DP, Mann R. Muscular strength development by electrical stimulation in healthy individuals. Phys Ther 1983;63(6):915–921.
59. Hartkopp A, et al. Effect of training on contractile and metabolic properties of wrist extensors in spinal cord-injured individuals. Muscle Nerve 2003;27(1):72–80.
60. Mohr T, et al. Long-term adaptation to electrically induced cycle training in severe spinal cord injured individuals. Spinal Cord 1997;35(1):1–16.
61. Gerrits HL, et al. Variability in fibre properties in paralysed human quadriceps muscles and effects of training. Pflugers Arch 2003;445(6):734–740.
62. Ragnarsson KT, et al. Clinical evaluation of computerized functional electrical stimulation after spinal cord injury: a multicenter pilot study. Arch Phys Med Rehabil 1988;69(9): 672–677.
63. Sloan KE, et al. Musculoskeletal effects of an electrical stimulation induced cycling programme in the spinal injured. Paraplegia 1994;32(6):407–415.
64. Baker LL, Yeh C, Wilson D, Waters RL. Electrical stimulation of wrist and fingers for hemiplegic patients. Phys Ther 1979;59(12):1495–1499.
65. Martin TP, Stein RB, Hoeppner PH, Reid DC. Influence of electrical stimulation on the morphological and metabolic properties of paralyzed muscle. J Appl Physiol 1992;72(4): 1401–1406.
66. Crameri RM, et al. Effects of electrical stimulation leg training during the acute phase of spinal cord injury: a pilot study. Eur J Appl Physiol 2000;83(4–5):409–415.

67. Munsat TL, McNeal D, Waters R. Effects of nerve stimulation on human muscle. Arch Neurol 1976;33(9):608–617.
68. Salmons S, Henriksson J. The adaptive response of skeletal muscle to increased activity. Muscle Nerve 1981;4: 94–105.
69. Eisenberg BR, Salmons S. The reorganization of subcellular structure in muscle undergoing fast-to-slow type transformation. A stereological study. Cell Tissue Res 1981; 220(3):449–471.
70. Nemeth PM. Electrical stimulation of denervated muscle prevents decreases in oxidative enzymes. Muscle Nerve 1982;5(2): 134–139.
71. Sreter FA, Pinter K, Jolesz F, Mabuchi K. Fast to slow transformation of fast muscles in response to long-term phasic stimulation. Exp Neurol 1982;75(1):95–102.
72. Peckham PH. Principles of electrical stimulation. Top spinal cord injury rehabilitation 1999;5(1):1–5.
73. Abbas JJ, Triolo RJ. Experimental evaluation of an adaptive feedforward controller for use in functional neuromuscular stimulation systems. IEEE Trans Rehabil Eng 1997; 5(1):12–22.
74. Durfee WK, MacLean KE. Methods for estimating isometric recruitment curves of electrically stimulated muscle. IEEE Trans Biomed Eng 1989;36(7):654–667.
75. Riess J, Abbas JJ. Adaptive control of cyclic movements as muscles fatigue using functional neuromuscular stimulation. IEEE Trans Neural Syst Rehabil Eng 2001;9(3):326–330.
76. Crago PE, Peckham PH, Thrope GB. Modulation of muscle force by recruitment during intramuscular stimulation. IEEE Trans Biomed Eng 1980;27(12):679–684.
77. Fang ZP, AJTM. A method of attaining natural recruitment order in artificially activated muscles. Proceedings 9th IEEE-EMBS Conference; 1987. pp 657–658.
78. Blair EA, Erlanger J. A comparison of the characteristics of axons through their individual electrical responses. Am J Physiol 1933;106:565–570.
79. Petrofsky JS. Control of the recruitment and firing frequencies of motor units in electrically stimulated muscles in the cat. Med Biol Eng Comput 1978;16(3):302–308.
80. Smith B, et al. An externally powered, multichannel, implantable stimulator-telemeter for control of paralyzed muscle. IEEE Trans Biomed Eng 1998;45(4):463–475.
81. Han-Chang Wu, Young S-T, Kuo T-S. A versatile multichannel direct-synthesized electrical stimulator for FES applications. IEEE Trans Instrum Meas 2002;51(1):2–9.
82. Ring H, Rosenthal N. Controlled study of neuroprosthetic functional electrical stimulation in sub-acute post-stroke rehabilitation. J Rehabil Med 2005;37(1):32–36.
83. Popovic MR, Keller T. Modular transcutaneous functional electrical stimulation system. Med Eng Phys 2005;27(1):81–92.
84. Mortimer JT, Bhadra N. Peripheral Nerve and Muscle Stimulation. In: Horch KW, Dhillon GS, editors. Neuroprosthetics: Theory and Practice. New Jersey: World Scientific (Series on Bioengineering & Biomedical Engineering; 2004.
85. Conway B. Theory and Principles of Electrode Processes. New York: Ronald Press; 1965.
86. Dymond AM. Characteristics of the metal-tissue interface of stimulation electrodes. IEEE Trans Biomed Eng 1976;23(4): 274–280.
87. Scheiner A, Polando G, Marsolais EB. Design and clinical application of a double helix electrode for functional electrical stimulation. IEEE Trans Biomed Eng 1994;41(5):425–431.
88. Daly JJ, et al. Performance of an intramuscular electrode during functional neuromuscular stimulation for gait training post stroke. J Rehabil Res Dev 2001;38(5):513–526.
89. Uhlir JP, Triolo RJ, Davis JA Jr, Bieri C. Performance of epimysial stimulating electrodes in the lower extremities of individuals with spinal cord injury. IEEE Trans Neural Syst Rehabil Eng 2004;12(2):279–287.
90. Loeb GE, Peck RA, Moore WH, Hood K. BION system for distributed neural prosthetic interfaces. Med Eng Phys 2001;23(1):9–18.
91. Naples GG, Mortimer JT, Scheiner A, Sweeney JD. A spiral nerve cuff electrode for peripheral nerve stimulation. IEEE Trans Biomed Eng 1988;35(11):905–916.
92. Tarler MD, Mortimer JT. Selective and independent activation of four motor fascicles using a four contact nerve-cuff electrode. IEEE Trans Neural Syst Rehabil Eng 2004; 12(2):251–257.
93. Leventhal DK, Durand DM. Chronic measurement of the stimulation selectivity of the flat interface nerve electrode. IEEE Trans Biomed Eng 2004;51(9):1649–1658.
94. Lawrence SM, Dhillon GS, Horch KW. Fabrication and characteristics of an implantable, polymer-based, intrafascicular electrode. J Neurosci Methods 2003;131(1–2): 9–26.
95. Tyler DJ, Durand DM. A slowly penetrating interfascicular nerve electrode for selective activation of peripheral nerves. IEEE Trans Rehabil Eng 1997;5(1):51–61.
96. Mushahwar VK, Gillard DM, Gauthier MJ, Prochazka A. Intraspinal micro stimulation generates locomotor-like and feedback-controlled movements. IEEE Trans Neural Syst Rehabil Eng 2002;10(1):68–81.
97. Saigal R, Renzi C, Mushahwar VK. Intraspinal microstimulation generates functional movements after spinal-cord injury. IEEE Trans Neural Syst Rehabil Eng 2004;12(4): 430–440.
98. McNeal DR, Baker LL. Effects of joint angle, electrodes and waveform on electrical stimulation of the quadriceps and hamstrings. Ann Biomed Eng 1988;16(3):299–310.
99. Bowman BR, Baker LL. Effects of waveform parameters on comfort during transcutaneous neuromuscular electrical stimulation. Ann Biomed Eng 1985;13(1):59–74.
100. Bajd T, Kralj A, Turk R. Standing-up of a healthy subject and a paraplegic patient. J Biomech 1982;15(1):1–10.
101. Baker L, et al. NeuroMuscular Electrical Stimulation: A Practical Guide. 4th ed. Los Amigos Research & Education Institute; 2000.
102. Wieler M, SN, Stein RB. WalkAid: An improved functional electrical stimulator for correcting foot-drop. Proceeding of the 1st Anuual Conference IFES; Cleveland (OH): 1996.
103. Burridge J, Taylor P, Hagan S, Swain I. Experience of clinical use of the Odstock dropped foot stimulator. Artif Organs 1997;21(3):254–260.
104. Taylor PN, et al. Clinical use of the Odstock dropped foot stimulator: its effect on the speed and effort of walking. Arch Phys Med Rehabil 1999;80(12):1577–1583.
105. Popovic D, Tomovic R, Schwirtlich L. Hybrid assistive system–the motor neuroprosthesis. IEEE Trans Biomed Eng 1989;36(7):729–737.
106. Solomonow M, et al. Reciprocating gait orthosis powered with electrical muscle stimulation (RGO II). Part II: Medical evaluation of 70 paraplegic patients. Orthopedics 1997; 20(5):411–418.
107. Snoek GJ, et al. Use of the NESS handmaster to restore handfunction in tetraplegia: clinical experiences in ten patients. Spinal Cord 2000;38(4):244–249.

108. Popovic MR, Popovic DB, Keller T. Neuroprostheses for grasping. Neurol Res 2002;24(5):443–452.
109. Popovic D, et al. Clinical evaluation of the bionic glove. Arch Phys Med Rehabil 1999;80(3):299–304.
110. Grandjean PA, Mortimer JT. Recruitment properties of monopolar and bipolar epimysial electrodes. Ann Biomed Eng 1986;14(1):53–66.
111. Gruner JA, Mason CP. Nonlinear muscle recruitment during intramuscular and nerve stimulation. J Rehabil Res Dev 1989;26(2):1–16.
112. Crago PE, Peckham PH, Mortimer JT, Van der Meulen JP. The choice of pulse duration for chronic electrical stimulation via surface, nerve, and intramuscular electrodes. Ann Biomed Eng 1974;2(3):252–264.
113. Mortimer T. Motor prosthesis. In: B VB, editor. Handbook of Physiology. Bethesda (MD): American Physiologer Society; 1981.
114. Smith BT, Betz RR, Mulcahey MJ, Triolo RJ. Reliability of percutaneous intramuscular electrodes for upper extremity functional neuromuscular stimulation in adolescents with C5 tetraplegia. Arch Phys Med Rehabil 1994;75(9):939–945.
115. Scheiner A, Mortimer JT, Roessmann U. Imbalanced biphasic electrical stimulation: muscle tissue damage. Ann Biomed Eng 1990;18(4):407–425.
116. Daly JJ, Ruff RL. Feasibility of combining multi-channel functional neuromuscular stimulation with weight-supported treadmill training. J Neurol Sci 2004;225(1-2):105–115.
117. Uhlir JP, Triolo RJ, Kobetic R. The use of selective electrical stimulation of the quadriceps to improve standing function in paraplegia. IEEE Trans Rehabil Eng 2000;8(4): 514–522.
118. Prochazka A, Davis LA. Clinical experience with reinforced, anchored intramuscular electrodes for functional neuromuscular stimulation. J Neurosci Methods 1992;42(3): 175–184.
119. Marsolais EB, Kobetic R. Functional walking in paralyzed patients by means of electrical stimulation. Clin Orthop Relat Res 1983;175:30–36.
120. Handa Y, Hoshimiya N, Iguchi Y, Oda T. Development of percutaneous intramuscular electrode for multichannel FES system. IEEE Trans Biomed Eng 1989;36(7):705–710.
121. Peterson DK, et al. Electrical activation of respiratory muscles by methods other than phrenic nerve cuff electrodes. Pacing Clin Electrophysiol 1989;12(5):854–860.
122. Degnan GG, Wind TC, Jones EV, Edlich RF. Functional electrical stimulation in tetraplegic patients to restore hand function. J Long Term Eff Med Implants 2002;12(3):175–188.
123. von Wild K, et al. Computer added locomotion by implanted electrical stimulation in paraplegic patients (SUAW). Acta Neurochir Suppl 2002;79:99–104.
124. Davis JA Jr, et al. Preliminary performance of a surgically implanted neuroprosthesis for standing and transfers—where do we stand? J Rehabil Res Dev 2001; 38(6):609–617.
125. Sharma M, et al. Implantation of a 16-channel functional electrical stimulation walking system. Clin Orthop Relat Res 1998;347:236–242.
126. Bowman BR, Erickson RC. 2nd, Acute and chronic implantation of coiled wire intraneural electrodes during cyclical electrical stimulation. Ann Biomed Eng 1985;13(1):75–93.
127. Popovic D, Gordon T, Rafuse VF, Prochazka A. Properties of implanted electrodes for functional electrical stimulation. Ann Biomed Eng 1991;19(3):303–316.
128. Singh J, Peck RA, Loeb GE. Development of BION Technology for functional electrical stimulation: Hermetic Packaging. Proceedings of the 23rd Annual EMBS International Conference; Istanbul, Turkey: 2001. pp 1313–1316.
129. Cameron T, Liinamaa TL, Loeb GE, Richmond FJ. Long-term biocompatibility of a miniature stimulator implanted in feline hind limb muscles. IEEE Trans Biomed Eng 1998; 45(8):1024–1035.
130. Cameron T, et al. Micromodular implants to provide electrical stimulation of paralyzed muscles and limbs. IEEE Trans Biomed Eng 1997;44(9):781–790.
131. Richmond FJ, et al. Therapeutic electrical stimulation with BIONs to rehabilitate shoulder and knee dysfunction. 2002. Ljubljana, Slovenia:IFESS.
132. Dupont A, et al. Therapeutical electrical stimulation with BIONS: Clinical trial report. in 2nd Joint Conference of the IEEE Engineering in Medicine and Biology Society and the Biomedical Engineering Society; Huston (TX): 2002.
133. Dupont AC, et al. Clinical Trials of BION Injectable Neuromuscular Stimulators. Reno (NV): RESNA; 2001.
134. Baker L. Rehabilitation of the Arm and Hand Following Stroke - A Clinical Trial with BIONsTM. Proceeding of the 26th Annual International Conference IEEE Engineering in Medicine and Biology Society; San Francisco: 2004.
135. Dupont AC, et al. First patients with BION implants for therapeutic electrical stimulation Neuromodulation. Neuromodulation 2004;7:38–47.
136. Arcos I, et al. Second-generation microstimulator. Artif Organs 2002;26(3):228–231.
137. Troyk PR, Brown IE, Moore WH, Loeb GE. Development of BION Technology for functional electrical stimulation: Bidirectional Telemetry. Istanbul, Turkey: IEEE-EMBS; 2001.
138. Zou Q, Kim ES, Loeb GE. Implantable Bimorph Piezoelectric Accelerometer for Feedback Control of Functional Neuromuscular Stimulation. The 12th International Conference on Solid State Sensors, Actuators and Microsystems; Boston: 2003.
139. Peckham PH, Mortimer JT, Marsolais EB. Controlled prehension and release in the C5 quadriplegic elicited by functional electrical stimulation of the paralyzed forearm musculature. Ann Biomed Eng 1980;8(4–6):369–388.
140. Triolo RJ, et al. Implanted Functional Neuromuscular Stimulation systems for individuals with cervical spinal cord injuries: clinical case reports. Arch Phys Med Rehabil 1996;77(11):1119–1128.
141. Marsolais B, RK. Experience with a helical percutaneous electrode in the human lower extremity. Proceedings of the RESNA 8th Annual Conference; 1985. pp 243–245.
142. Peterson DK, Nochomovitz ML, Stellato TA, Mortimer JT. Long-term intramuscular electrical activation of the phrenic nerve: efficacy as a ventilatory prosthesis. IEEE Trans Biomed Eng 1994;41(12):1127–1135.
143. Herbert MA, Bobechko WP. Paraspinal muscle stimulation for the treatment of idiopathic scoliosis in children. Orthopedics 1987;10(8):1125–1132.
144. Peckham PH, et al. Efficacy of an implanted neuroprosthesis for restoring hand grasp in tetraplegia: a multicenter study. Arch Phys Med Rehabil 2001;82(10):1380–1388.
145. Veraart C, Grill WM, Mortimer JT. Selective control of muscle activation with a multipolar nerve cuff electrode. IEEE Trans Biomed Eng 1993;40(7):640–653.
146. Walter JS, et al. Multielectrode nerve cuff stimulation of the median nerve produces selective movements in a raccoon animal model. J Spinal Cord Med 1997;20(2):233–243.

147. Crampon MA, Brailovski V, Sawan M, Trochu F. Nerve cuff electrode with shape memory alloy armature: design and fabrication. Biomed Mater Eng 2002;12(4):397–410.
148. Navarro X, Valderrama E, Stieglitz T, Schuttler M. Selective fascicular stimulation of the rat sciatic nerve with multipolar polyimide cuff electrodes. Restor Neurol Neurosci 2001; 18(1):9–21.
149. Loeb GE, Peck RA. Cuff electrodes for chronic stimulation and recording of peripheral nerve activity. J Neurosci Methods 1996;64(1):95–103.
150. Tyler DJ, Durand DM. Chronic response of the rat sciatic nerve to the flat interface nerve electrode. Ann Biomed Eng 2003;31(6):633–642.
151. Tyler DJ, Durand DM. Functionally selective peripheral nerve stimulation with a flat interface nerve electrode. IEEE Trans Neural Syst Rehabil Eng 2002;10(4):294–303.
152. Leventhal DK, Durand DM. Subfascicle stimulation selectivity with the flat interface nerve electrode. Ann Biomed Eng 2003;31(6):643–652.
153. Yoshida K, Horch K. Selective stimulation of peripheral nerve fibers using dual intrafascicular electrodes. IEEE Trans Biomed Eng 1993;40(5):492–494.
154. McDonnall D, Clark GA, Normann RA. Selective motor unit recruitment via intrafascicular multielectrode stimulation. Can J Physiol Pharmacol 2004;82(8–9):599–609.
155. Zheng X, Zhang J, Chen T, Chen Z. Longitudinally implanted intrafascicular electrodes for stimulating and recording fascicular physioelectrical signals in the sciatic nerve of rabbits. Microsurgery 2003;23(3):268–273.
156. Lawrence SM, et al. Long-term biocompatibility of implanted polymer-based intrafascicular electrodes. J Biomed Mater Res 2002;63(5):501–506.
157. Yoshida K, Jovanovic K, Stein RB. Intrafascicular electrodes for stimulation and recording from mudpuppy spinal roots. J Neurosci Methods 2000;96(1):47–55.
158. Grill WM Jr, Mortimer JT. Quantification of recruitment properties of multiple contact cuff electrodes. IEEE Trans Rehabil Eng 1996;4(2):49–62.
159. Goodall EV, de Breij JF, Holsheimer J. Position-selective activation of peripheral nerve fibers with a cuff electrode. IEEE Trans Biomed Eng 1996;43(8):851–856.
160. Veltink PH, van Alste JA, Boom HB. Multielectrode intrafascicular and extraneural stimulation. Med Biol Eng Comput 1989;27(1):19–24.
161. Hoffer JA, Loeb GE. Implantable electrical and mechanical interfaces with nerve and muscle. Ann Biomed Eng 1980; 8(4–6):351–360.
162. Juch PJ, Minkels RF. The strap-electrode: a stimulating and recording electrode for small nerves. Brain Res Bull 1989; 22(5):917–918.
163. Stein RB, et al. Stable long-term recordings from cat peripheral nerves. Brain Res 1977;128(1):21–38.
164. Sweeney JD, Mortimer JT. An asymmetric two electrode cuff for generation of unidirectionally propagated action potentials. IEEE Trans Biomed Eng 1986;33(6): 541–549.
165. Agnew WF, McCreery DB, Yuen TG, Bullara LA. Histologic and physiologic evaluation of electrically stimulated peripheral nerve: considerations for the selection of parameters. Ann Biomed Eng 1989;17(1):39–60.
166. McNeal DR, Bowman BR. Selective activation of muscles using peripheral nerve electrodes. Med Biol Eng Comput 1985;23(3):249–253.
167. Sweeney JD, Ksienski DA, Mortimer JT. A nerve cuff technique for selective excitation of peripheral nerve trunk regions. IEEE Trans Biomed Eng 1990;37(7):706–715.
168. Sweeney JD, Crawford NR, Brandon TA. Neuromuscular stimulation selectivity of multiple-contact nerve cuff electrode arrays. Med Biol Eng Comput 1995;33(3 Spec No): 418–425.
169. Rozman J, Sovinec B, Trlep M, Zorko B. Multielectrode spiral cuff for ordered and reversed activation of nerve fibres. J Biomed Eng 1993;15(2):113–120.
170. Rozman J, Trlep M. Multielectrode spiral cuff for selective stimulation of nerve fibres. J Med Eng Technol 1992;16(5): 194–203.
171. Deurloo KE, Holsheimer J, Boom HB. Transverse tripolar stimulation of peripheral nerve: a modelling study of spatial selectivity. Med Biol Eng Comput 1998;36(1):66–74.
172. Tarler MD, Mortimer JT. Comparison of joint torque evoked with monopolar and tripolar-cuff electrodes. IEEE Trans Neural Syst Rehabil Eng 2003;11(3):227–235.
173. Tyler DJ. Functionally Selective Stimulation of Peripheral Nerves: Electrodes That Alter Nerve Geometry. Cleveland (OH): Case Western Reserve University; 1999.
174. Choi AQ, Cavanaugh JK, Durand DM. Selectivity of multiple-contact nerve cuff electrodes: a simulation analysis. IEEE Trans Biomed Eng 2001;48(2):165–172.
175. Branner A, Stein RB, Normann RA. Selective stimulation of cat sciatic nerve using an array of varying-length microelectrodes. J Neurophysiol 2001;85(4):1585–1594.
176. Anderson JM. Inflammatory response to implants. ASAIO Trans 1988;34(2):101–107.
177. Grill WM, Mortimer JT. Neural and connective tissue response to long-term implantation of multiple contact nerve cuff electrodes. J Biomed Mater Res 2000;50(2):215–226.
178. Rutten WL, van Wier HJ, Put JH. Sensitivity and selectivity of intraneural stimulation using a silicon electrode array. IEEE Trans Biomed Eng 1991;38(2):192–198.
179. Rutten WL, Meier JH. Selectivity of intraneural prosthetic interfaces for muscular control. Med Biol Eng Comput 1991;29(6):3–7.
180. Meier JH, Rutten WL, Boom HB. Force recruitment during electrical nerve stimulation with multipolar intrafascicular electrodes. Med Biol Eng Comput 1995;33(3 Spec No): 409–417.
181. Nannini N, Horch K. Muscle recruitment with intrafascicular electrodes. IEEE Trans Biomed Eng 1991;38(8):769–776.
182. Lawrence SM, et al. Acute peripheral nerve recording characteristics of polymer-based longitudinal intrafascicular electrodes. IEEE Trans Neural Syst Rehabil Eng 2004; 12(3):345–348.
183. Dhillon GS, Lawrence SM, Hutchinson DT, Horch KW. Residual function in peripheral nerve stumps of amputees: implications for neural control of artificial limbs. J Hand Surg [Am] 2004;29(4):605–615; discussion 616–618.
184. Zheng XJ, et al. [Experimental study of biocompatibility of LIFEs in peripheral fascicles]. Zhonghua Yi Xue Za Zhi 2003;83(24):2152–2157.
185. Malmstrom JA, McNaughton TG, Horch KW. Recording properties and biocompatibility of chronically implanted polymer-based intrafascicular electrodes. Ann Biomed Eng 1998;26(6):1055–1064.
186. Lundborg G, Richard P. Bunge memorial lecture. Nerve injury and repair—a challenge to the plastic brain. J Peripher Nerv Syst 2003;8(4):209–226.
187. Herman R, He J, D'Luzansky S, Willis W, Dilli S. Spinal cord stimulation facilitates functional walking in a chronic, incomplete spinal cord injured. Spinal Cord 2002;40(2): 65–68.

188. Pinter MM, Dimitrijevic MR. Gait after spinal cord injury and the central pattern generator for locomotion. Spinal Cord 1999;37(8):531–537.
189. Field-Fote E. Spinal cord stimulation facilitates functional walking in a chronic, incomplete spinal cord injured subject. Spinal Cord 2002;40(8):428.
190. Prochazka A, Mushahwar V, Yakovenko S. Activation and coordination of spinal motoneuron pools after spinal cord injury. Prog Brain Res 2002;137:109–124.
191. Mushahwar VK, Horch KW. Selective activation of muscle groups in the feline hindlimb through electrical microstimulation of the ventral lumbo-sacral spinal cord. IEEE Trans Rehabil Eng 2000;8(1):11–21.
192. Mushahwar VK, Horch KW. Proposed specifications for a lumbar spinal cord electrode array for control of lower extremities in paraplegia. IEEE Trans Rehabil Eng 1997; 5(3):237–243.
193. Tai C, et al. Multi-joint movement of the cat hindlimb evoked by microstimulation of the lumbosacral spinal cord. Exp Neurol 2003;183(2):620–627.
194. Mushahwar VK, Horch KW. Selective activation and graded recruitment of functional muscle groups through spinal cord stimulation. Ann NY Acad Sci 1998;860:531–535.
195. Mushahwar VK, Collins DF, Prochazka A. Spinal cord microstimulation generates functional limb movements in chronically implanted cats. Exp Neurol 2000;163(2):422–429.
196. Mushahwar VK, Horch KW. Muscle recruitment through electrical stimulation of the lumbo-sacral spinal cord. IEEE Trans Rehabil Eng 2000;8(1):22–29.
197. Durfee WK, Palmer KI. Estimation of force-activation, force-length, and force-velocity properties in isolated, electrically stimulated muscle. IEEE Trans Biomed Eng 1994;41(3):205–216.
198. Durfee W. Muscle model identification in neural prosthesis systems. In: Stein R, Peckham H, editors. Neural Prostheses: Replacing Motor Function After Disease or Disability. Oxford University Press; 1992.
199. Durfee WK. Control of standing and gait using electrical stimulation: influence of muscle model complexity on control strategy. Prog Brain Res 1993;97:369–381.
200. Veltink PH, Chizeck HJ, Crago PE, el-Bialy A. Nonlinear joint angle control for artificially stimulated muscle. IEEE Trans Biomed Eng 1992;39(4):368–380.
201. Wame X. Newromoclulation 2001;4:187–195.
202. Marsolais EB, Kobetic R. Functional electrical stimulation for walking in paraplegia. J Bone Joint Surg Am 1987;69(5): 728–733.
203. Yamaguchi GT, Zajac FE. Restoring unassisted natural gait to paraplegics via functional neuromuscular stimulation: a computer simulation study. IEEE Trans Biomed Eng 1990; 37(9):886–902.
204. Quintern J, Minwegen P, Mauritz KH. Control mechanisms for restoring posture and movements in paraplegics. Prog Brain Res 1989;80:489–502; discussion 479–480.
205. Nathan RH. Control strategies in FNS systems for the upper extremities. Crit Rev Biomed Eng 1993;21(6):485–568.
206. Crago PE, et al. New control strategies for neuroprosthetic systems. J Rehabil Res Dev 1996;33(2):158–172.
207. Bajzek TJ, Jaeger RJ. Characterization and control of muscle response to electrical stimulation. Ann Biomed Eng 1987; 15(5):485–501.
208. Lan N, Crago PE, Chizeck HJ. Feedback control methods for task regulation by electrical stimulation of muscles. IEEE Trans Biomed Eng 1991;38(12):1213–1223.
209. Crago PE, Nakai RJ, Chizeck HJ. Feedback regulation of hand grasp opening and contact force during stimulation of paralyzed muscle. IEEE Trans Biomed Eng 1991;38(1): 17–28.
210. Kataria P, Abass J. Adaptive user-specific control of movements with functional neuromuscular stimulation. Proceedings of the IEEE/BMES Conference; Atlanta (GA): 1999. p. 604.
211. Abbas JJ, Chizeck HJ. Neural network control of functional neuromuscular stimulation systems: computer simulation studies. IEEE Trans Biomed Eng 1995;42(11):1117–1127.
212. Chang GC, et al. A neuro-control system for the knee joint position control with quadriceps stimulation. IEEE Trans Rehabil Eng 1997;5(1):2–11.
213. Riess J, Abbas JJ. Adaptive neural network control of cyclic movements using functional neuromuscular stimulation. IEEE Trans Rehabil Eng 2000;8(1):42–52.
214. Chizeck HJ. Adaptive and nonlinear control methods for neural prostheses. In: Stein PP, RB, Popovic DB, editor. Neural prostheses: replacing motor function after disease or disability. New York: Oxford University Press; 1992. pp 298–328.
215. Abbas J, Chizeck H. A neural network controller for functional neuromuscular stimulation systems. Proceedings IEEE/EMBS Conference. Orlando (FL): 1991. p 1456–1457.
216. Chen JJ, et al. Applying fuzzy logic to control cycling movement induced by functional electrical stimulation. IEEE Trans Rehabil Eng 1997;5(2):158–169.
217. Veltink PH. Control of FES-induced cyclical movements of the lower leg. Med Biol Eng Comput 1991;29(6):NS8–NS12.
218. Friehs GM, et al. Brain-machine and brain-computer interfaces. Stroke 2004;35(11 Suppl. 1):2702–2705.
219. Patil P, Carmena J, Nicolelis MA, DA T. Ensemble recordings of human subcortical neurons as a source of motor control signals for a brain-machine interface. Neurosurgery 2004; 55(1):27–35.
220. Schwartz AB. Cortical neural prosthetics. Annu Rev Neurosci 2004;27:487–507.
221. Donoghue JP. Connecting cortex to machines: recent advances in brain interfaces. Nat Neurosci 2002;5(Suppl.): 1085–1088.
222. Creasey GH, et al. Clinical applications of electrical stimulation after spinal cord injury. J Spinal Cord Med 2004;27(4): 365–375.
223. Postans NJ, Hasler JP, Granat MH, Maxwell DJ. Functional electric stimulation to augment partial weight-bearing supported treadmill training for patients with acute incomplete spinal cord injury: A pilot study. Arch Phys Med Rehabil 2004;85(4):604–610.
224. Field-Fote EC. Combined use of body weight support, functional electric stimulation, and treadmill training to improve walking ability in individuals with chronic incomplete spinal cord injury. Arch Phys Med Rehabil 2001;82(6): 818–824.
225. Field-Fote EC, Tepavac D. Improved intralimb coordination in people with incomplete spinal cord injury following training with body weight support and electrical stimulation. Phys Ther 2002;82(7):707–715.
226. Barbeau H, Ladouceur M, Mirbagheri MM, Kearney RE. The effect of locomotor training combined with functional electrical stimulation in chronic spinal cord injured subjects: walking and reflex studies. Brain Res Brain Res Rev 2002; 40(1–3):274–291.
227. Barbeau H, et al. Tapping into spinal circuits to restore motor function. Brain Res Brain Res Rev 1999;30(1):27–51.

228. Field-Fote EC. Electrical stimulation modifies spinal and cortical neural circuitry. Exerc Sport Sci Rev 2004;32(4): 155–160.
229. Waters RL, Campbell JM, Nakai R. Therapeutic electrical stimulation of the lower limb by epimysial electrodes. Clin Orthop 1988;233:44–52.
230. Kagaya H, Shimada Y, Sato K, Sato M. Changes in muscle force following therapeutic electrical stimulation in patients with complete paraplegia. Paraplegia 1996;34(1): 24–29.
231. Baldi J, Jackson RD, Moraille R, Mysiw WJ. Muscle atrophy is prevented in patients with acute spinal cord injury using functional electrical stimulation. Spinal Cord 1998;36(7): 463–469.
232. Scremin AM, et al. Increasing muscle mass in spinal cord injured persons with a functional electrical stimulation exercise program. Arch Phys Med Rehabil 1999;80(12): 1531–1536.
233. Crameri RM, et al. Effects of electrical stimulation-induced leg training on skeletal muscle adaptability in spinal cord injury. Scand J Med Sci Sports 2002;12(5):316–322.
234. Kern H, et al. Long-term denervation in humans causes degeneration of both contractile and excitation-contraction coupling apparatus, which is reversible by functional electrical stimulation (FES): a role for myofiber regeneration? J Neuropathol Exp Neurol 2004;63(9):919–931.
235. Hangartner TN, Rodgers MM, Glaser RM, Barre PS. Tibial bone density loss in spinal cord injured patients: effects of FES exercise. J Rehabil Res Dev 1994;31(1):50–61.
236. Mohr T, et al. Increased bone mineral density after prolonged electrically induced cycle training of paralyzed limbs in spinal cord injured man. Calcif Tissue Int 1997;61(1):22–25.
237. Bloomfield SA, Mysiw WJ, Jackson RD. Bone mass and endocrine adaptations to training in spinal cord injured individuals. Bone 1996;19(1):61–68.
238. Lee YH, Rah JH, Park RW, Park CI. The effect of early therapeutic electrical stimulation on bone mineral density in the paralyzed limbs of the rabbit. Yonsei Med J 2001;42(2):194–198.
239. Pettersson U, Nordstrom P, Lorentzon R. A comparison of bone mineral density and muscle strength in young male adults with different exercise level. Calcif Tissue Int 1999;64(6):490–498.
240. Belanger M, et al. Electrical stimulation: can it increase muscle strength and reverse osteopenia in spinal cord injured individuals? Arch Phys Med Rehabil 2000;81(8):1090–1098.
241. Pacy PJ, et al. Muscle and bone in paraplegic patients, and the effect of functional electrical stimulation. Clin Sci (London) 1988;75(5):481–487.
242. Leeds EM, et al. Bone mineral density after bicycle ergometry training. Arch Phys Med Rehabil 1990;71(3): 207–209.
243. Eser P, et al. Effect of electrical stimulation-induced cycling on bone mineral density in spinal cord-injured patients. Eur J Clin Invest 2003;33(5):412–419.
244. BeDell KK, Scremin AM, Perell KL, Kunkel CF. Effects of functional electrical stimulation-induced lower extremity cycling on bone density of spinal cord-injured patients. Am J Phys Med Rehabil 1996;75(1):29–34.
245. Weiss DS, Kirsner R, Eaglstein WH. Electrical stimulation and wound healing. Arch Dermatol 1990;126(2):222–225.
246. Castillo E, Sumano H, Fortoul TI, Zepeda A. The influence of pulsed electrical stimulation on the wound healing of burned rat skin. Arch Med Res 1995;26(2):185–189.
247. Reich JD, Tarjan PP. Electrical stimulation of skin. Int J Dermatol 1990;29(6):395–400.
248. Thawer HA, Houghton PE. Effects of electrical stimulation on the histological properties of wounds in diabetic mice. Wound Repair Regen 2001;9(2):107–115.
249. Reger SI, et al. Experimental wound healing with electrical stimulation. Artif Organs 1999;23(5):460–462.
250. Kloth LC. Physical modalities in wound management: UVC, therapeutic heating and electrical stimulation. Ostomy Wound Manage 1995;41(5):18–20, 22–24, 26–27.
251. Gentzkow GD. Electrical stimulation to heal dermal wounds. J Dermatol Surg Oncol 1993;19(8):753–758.
252. Kloth LC, McCulloch JM. Promotion of wound healing with electrical stimulation. Adv Wound Care 1996;9(5):42–45.
253. Agarwal S, et al. Long-term user perceptions of an implanted neuroprosthesis for exercise, standing, and transfers after spinal cord injury. J Rehabil Res Dev 2003;40(3):241–252.
254. Kloth LC. Electrical stimulation for wound healing: a review of evidence from in vitro studies, animal experiments, and clinical trials. Int J Low Extrem Wounds 2005;4(1): 23–44.
255. Stewart RM, Desaloms JM, Sanghera MK. Stimulation of the subthalamic nucleus for the treatment of Parkinson's disease: postoperative management, programming, and rehabilitation. J Neurosci Nurs 2005;37(2):108–114.
256. Garcia L, D'Alessandro G, Bioulac B, Hammond C. High-frequency stimulation in Parkinson's disease: more or less? Trends Neurosci 2005;28(4):209–216.
257. Uthman BM. Vagus nerve stimulation for seizures. Arch Med Res 2000;31(3):300–303.
258. McLachlan RS. Vagus nerve stimulation for intractable epilepsy: a review. J Clin Neurophysiol 1997;14(5):358–368.
259. Carroll D, et al. Transcutaneous electrical nerve stimulation (TENS) for chronic pain. Cochrane Database Syst Rev 2001;3:CD003222.
260. Chen JD, Lin HC. Electrical pacing accelerates intestinal transit slowed by fat-induced ileal brake. Dig Dis Sci 2003;48(2):251–256.
261. Sun Y, Chen J. Intestinal electric stimulation decreases fat absorption in rats: therapeutic potential for obesity. Obes Res 2004;12(8):1235–1242.
262. Xing J, et al. Gastric electrical-stimulation effects on canine gastric emptying, food intake, and body weight. Obes Res 2003;11(1):41–47.

Reading List

Baker L, et al. NeuroMuscular Electrical Stimulation: A Practical Guide 4th ed., California: Los Amigos Research & Education Institute, 2000, Available at www.ranchorep.org/Publications.htm, An excellent resource book on the basics of electrical stimulation, emphasizing the clinical uses and outcomes of neuromuscular stimulation.

Horch KW, Dhillon GS, editors. Neuroprosthetics: Theory and Practice. New Jersey: World Scientific (Series on Bioengineering & Biomedical Engineering—Vol. 2, J K-J Li, Series Editor); 2004. An extensive and comprehensive text covering a wide range of neuroprosthetic subjects.

McNeal DR. 2000 Years of Electrical Stimulation. In: Hambrecht FT, Reswick JB, editors. Functional Electrical Stimulation: Applications in Neural Prostheses. New York: Marcel Dekker; 1977. pp 3–35. A now classic historical perspective on the usage of electrical stimulation for medical purposes from the time of the ancient Greeks up until the modern

advent of functional electrical stimulation techniques in the 1970s.

Popovic D, Sinkjaer T. Control of Movement for the Physically Disabled. London: Springer Verlag; 2003. Reviews the state of the art of rehabilitation systems and methods used to restore movement, including the combined use of electrical stimulation systems and orthotics.

Reilly JP. Applied Bioelectricity: From Electrical Stimulation to Electropathology. New York: Springer-Verlag; 1998. A detailed text covering the fundamental principles of electrical stimulation, including sensory, motor and cardiac responses.

The web-site of the Cleveland FES Center at Case Western Reserve University contains an excellent "Resource Guide" as well as an extensive glossary of FES terms. Available at fescenter.case.edu/.

The International Functional Electrical Stimulation Society (IFESS) acts to promote the research, application and understanding of electrical stimulation as it is utilized in the field of medicine. The official journal of IFESS along with the International Neuromodulation Society (INS) is Neuromodulation. For a number of general resources on FES technologies see the IFESS web-site available at www.ifess.org/index.htm.

See also ELECTROPHYSIOLOGY; REHABILITATION AND MUSCLE TESTING; SPINAL CORD STIMULATION.

FUNCTIONAL NEUROMUSCULAR STIMULATION. See FUNCTIONAL ELECTRICAL STIMULATION.

GAMMA CAMERA. See ANGER CAMERA.

GAMMA KNIFE

STEVEN J. GOETSCH
San Diego Gamma Knife Center
La Jolla, California

INTRODUCTION

The Leksell Gamma Knife is one of the most massive and costliest medical products ever created (see Fig. 1). It is also one of the most clinically and commercially successful medical products in history, with > 180 units installed worldwide at this writing. The device is used exclusively for the treatment of brain tumors and other brain abnormalities. The Gamma Knife, also known as the Leksell Gamma Unit, contains 201 sources of radioactive cobalt-60, each of which emits an intense beam of highly penetrating gamma radiation (see Cobalt-60 units for radiotherapy). Due to the penetrating nature of the gamma rays emitted by these radiation sources, the device must be heavily shielded, and therefore it weighs ∼ 22 tons. The Gamma Knife must also be placed in a vault with concrete shielding walls 2–4-ft thick.

This remarkable device is used in the following way: A patient known from prior medical diagnosis to have a brain tumor or other treatable brain lesion, is brought to a Gamma Knife Center on the selected day of treatment. Gamma Knife treatment is thus intended for elective surgery and is never used for emergency purposes. The patient is prepared for treatment, which normally occurs with the patient alert and awake, by a nurse. Then, a neurosurgeon injects local anesthetic under the skin of the forehead and posterior of the skull. He/she then affixes a stereotactic head frame (see Fig. 2) with sharp pins to the patient's head (much like a halo fixation device for patients with a broken neck). The patient is transported by wheelchair or gurney to a nearby imaging center where a Computed Tomography (CT) X-ray scan or a Magnetic Resonance Imaging (MRI) scan of the brain (with the stereotactic head frame on) is obtained (see articles on Computed Tomography and Magnetic Resonance Imaging). Specially constructed boxes consisting of panels containing geometric localization markers are attached to the stereotactic frame and surround the patient's head during imaging. The markers contained in the localization boxes are visible on the brain scan, just outside the skull (see Fig. 3). All imaging studies are then exported via a PACS computer network or DAT tape (see the article on Picture Archiving and Communication Systems) into a powerful computer, where a treatment plan is created. A neurosurgeon, a radiation oncologist, and a medical physicist participate in the planning process. When the plan is satisfactorily completed, the patient (still wearing the stereotactic frame) is brought into the treatment room. The patient is then placed on the couch of the treatment unit and the stereotactic frame is docked with the trunnions affixed to the helmet (see Fig. 4). After the staff members leave the room and the room shielding doors are closed, the Gamma Knife vault door automatically opens and the patient couch is pushed forward into the body of the device, so that the holes in the collimating helmet line up with the radiation source pattern inside the central body of the device. The treatment begins at that point. Any given patient may be treated in this manner with a single "shot" (e.g., treatment) or with

Figure 1. The Leksell Gamma Unit Model U for treatment of patients with brain tumors and other brain abnormalities.

Figure 2. Patient with Leksell Model G stererotactic frame affixed to their head. This frame restricts patient motion during imaging and treatment and also allows placement of fiducial markers to localize the volume to be treated.

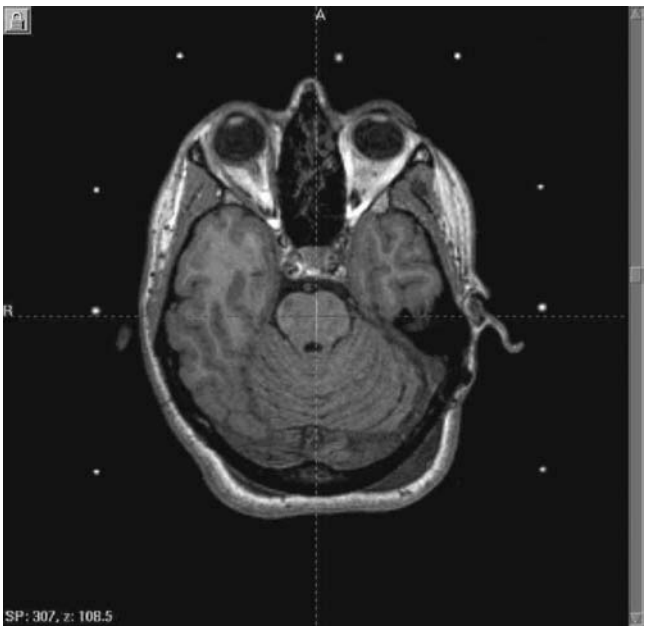


Figure 3. Axial MRI scan of patients brain, with external fiducial markers filled with copper sulfate solution to enable localization of target volumes.

many shots (30 or more in some cases). The collimating helmet may be changed to use one or more of the available helmet sizes, corresponding to a roughly spherical volume 4, 8, 14, or 18 mm in diameter. At the conclusion of treatment, the stereotactic frame is removed and most patients are then discharged. Thus Gamma Knife radiosurgery is most commonly performed on an outpatient basis.

Gamma Knife radiosurgery has shown rapidly increasing acceptance, since the first commercial unit was introduced at the University of Pittsburgh in 1987 (1). Despite the high purchase price (a little >$3 million) and single purpose, Gamma Knife units are widely available in the United States, Europe, Asia, and other parts of the world. All units are manufactured by Elekta Instruments, of Stockholm, Sweden. Use of this device can eliminate the need for open surgery of the brain. Modern surgical techniques and nursing follow-up have reduced the death rate due to brain surgery from as much as 50% in the 1930s to <1% in the United States in 2002. However, Gamma Knife patients most commonly do not have to remain overnight in the hospital at all (an important consideration in a very cost conscious healthcare environment), while craniotomy patients typically have a 2–5 day stay. Conventional brain surgery patients sometimes require 30 days or more of hospitalization if extremely adverse effects occur. Thus, the cost of the Gamma Knife outpatient procedure is typically far less than that for inpatient open brain surgery. Recovery of the patient is much more rapid for Gamma Knife patients, with most patients going home immediately and returning to work or other normal routines in 1–2 days.

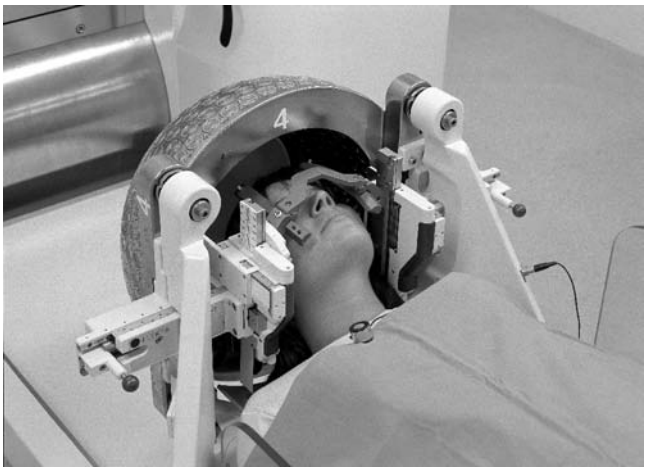


Figure 4. Supine patient in position for treatment in Gamma Knife. Stereotactic head frame is carefully docked with trunnions, which slide in precise channels in the secondary collimator helmet.

EARLY HISTORY OF THE DEVICE

Gamma Knife radiosurgery was an outgrowth of several prior inventions. Dr. Lars Leksell, a Swedish neurosurgeon, was one of the pioneers in the field of stereotaxis (see the article on Stereotactic Surgery). Dr. Leksell was motivated to find minimally invasive ways to treat brain abnormalities by the appalling death rate for early twentieth century brain surgery, which could be as high as 50% (2). Leksell was one of the first surgeons to create a workable stereotactic frame (in 1949) that could be affixed to a patients skull, together with a set of indexed external markers (called fiducials) that were visible on an X-ray of the patient's head. Only primitive brain imaging procedures were available in 1950 to the late 1970s, so stereotactic surgery patients had to undergo a painful procedure called pneumoencephalography. A lumbar puncture was used to introduce air into the spinal canal while the patient (strapped into a special harness) was manipulated upside down, back and forth, while air was injected under positive pressure to displace the cerebro-spinal fluid in the ventricles of the brain. A pair of plane orthogonal X-ray images (anterior–posterior and lateral) were then taken. Since the air-filled ventricles were well imaged by this technique, standard atlases of the human brain such as Schaltenbrand and Wahren (3) were then used to compute the location of the desired target relative to these landmarks. The imaging procedure alone was considered extremely painful and typically required hospitalization. The early stereotactic frames were applied by drilling into the patient's skull and driving screws into the calvarium (outer table of the skull), which was then topped with a Plaster of Paris cap that could be rigidly fixed. A twist drill could then be guided to create a small hole (a few millimeters in diameter) in the patient's skull, through which a catheter could be passed to a designated target, such as the globus pallidum for treatment of Parkinsons disease. A radio frequency transmitter was passed through the catheter and a small volume of tissue was heated to high

temperature, creating a deliberate, controlled brain lesion. This procedure, though rigorous, was far less invasive and dangerous than open brain surgery, called craniotomy.

Leksell then attached an X-ray treatment unit to his stereotactic frame and used it in 1951 to treat brain structures without opening of the skull. He called this procedure "radiosurgery", which he defined as "a single high dose of radiation stereotactically directed to an intracranial region of interest" (4). Leksell was successfully in treating previously intractable cases of trigeminal neuralgia, an extremely painful facial nerve disease, by stereotactically irradiating the very narrow ($\sim$ 2–4 mm diameter) nerve as it enters the brainstem. Only a few patients were treated with this X-ray unit.

Leksell then collaborated with physicist Borge Larsson in treating patients at a cyclotron located at Uppsala University near Stockholm beginning in 1957. Tobias and others had just begun treating patients with proton therapy (see article on Proton Beam Radiotherapy) at the University of California Berkeley in 1954. The proton is a positively charged subatomic particle, a basic building block of matter, which has extremely useful properties for treatment of human patients. The charged particles, accelerated to very high energies by a massive cyclotron (typically located at a high energy physics research laboratory) are directed at a patient, where they begin to interact through the Coulomb force while passing through tissue. At the end of the proton range, however, the particles give off a large burst of energy (the Bragg peak) and then stop abruptly. Leksell and Larsson utilized these properties with well-collimated beams of protons directed at intracranial targets. A few other centers in the United States and Russia also began proton therapy in the 1950s and 1960s.

The Gamma Knife was invented in Stockholm, Sweden by Leksell and Larsson and was manufactured (as a prototype) by the Swedish shipbuilding firm Mottola. The first unit had 179 radioactive cobalt-60 sources and was installed in 1968 at Sophiahemmet Hospital in Stockholm, Sweden (5). This original unit had three interchangeable helmets with elliptically shaped collimators of maximum diameter 4, 8, or 14 mm. Despite the lack of good brain imaging techniques at that time, the Gamma Knife was used to successfully treat Parkinson's disease (a movement disorder), trigeminal neuralgia (extreme facial pain), and arteriovenous malformations (AVMs), which are a tangle of congenitally malformed arteries and veins inside the brain. Several years later a second nearly identical unit was manufactured for Leksell when he became a faculty member at Karolinska Hospital in Stockholm. The original unit lay idle for a number of years, until it was donated to UCLA Medical Center in Los Angeles, where it was moved in 1982 (Fig. 5). It was used in animal research and treated a limited number of human patients before it was retired permanently in 1988 (6). The two original, custom-made Gamma Knife units were unique in the world and did not immediately enjoy widespread acceptance or gain much notice. A large number of patients with AVMs began to be treated at the Gamma Knife Center in Karolinska, by Dr. Ladislau Steiner, a neurosurgical colleague of Lars Leksell. Arteriovenous malformations are prone to spontaneous hemorrhage that can cause sudden coma or death. Open surgical techniques for removal of these life-threatening vascular anomalies were extremely difficult and dangerous in the 1970s. Patients came from all over the world to be treated for these AVMs at the Gamma Knife Center in Stockholm.

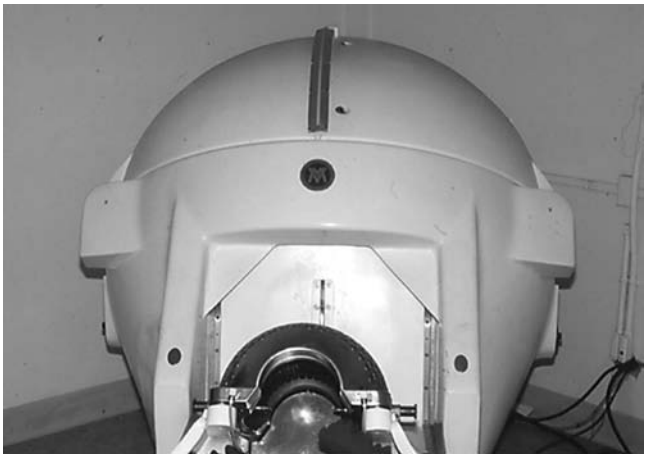


Figure 5. Original Gamma Knife, after being moved from Stockholm to UCLA Medical Center in Los Angeles.

In 1984 and 1985, two new Gamma Knife units were manufactured using Dr. Leksell's specifications by Nucletec SA of Switzerland (a subsidiary of Scanditronix AB, Sweden) for installation in hospitals in Buenos Aires, Argentina and Sheffield, England, respectively (7,8). These units also had three sets of collimators, which were now circular in shape, of 4, 8, and 14 mm diameter, but the number of cobalt-60 sources was increased to 201. The mechanical tolerance was exquisite: The convergence of all 201 beams at the focal point was specified as $\pm$ 0.1 mm. The total radioactivity was 209 TBq (5500 Ci) and the sources were distributed evenly over a $160 \times 60^{\circ}$ sector of the hemispherical secondary collimators (Fig. 6). An ionization chamber (a type of radiation detector) placed at the center of a spherical phantom 16 cm in diameter was used to measure an absorbed dose rate of $\sim$ 2.5 gray $\cdot$ min^{-1} for the Sheffield unit. This was adequate to treat patients to a large radiation dose in a reasonable period of time. Both Gamma Knife units were successfully used fo many years to treat patients in their respective countries.

A new corporation, called Elekta Instruments, AB, of Stockholm was created in 1972 by Laurent and Dan Leksell, sons of Lars Leksell, to manufacture neurosurgical products, including the Gamma Knife, which is now a trademark of this firm. Elekta created the first commercial Gamma Knife product, the Model U, and has manufactured all Gamma Knife units worldwide since that time. This new 201 source unit was installed at the University of Pittsburgh in 1987 and expanded the available beam diameters to include a fourth secondary collimator with a nominal diameter of 18 mm (1). The trunnions, which connect the secondary collimator helmets to the patient, were now configured to dock with connecting points located

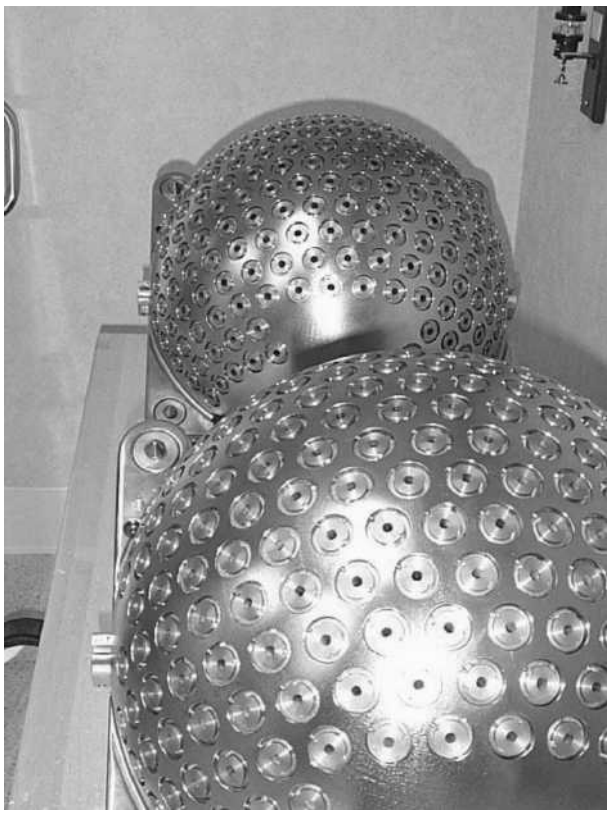


Figure 6. (Upper panel) Collimator helmets for Leksell Gamma Unit Model U. (Lower panel) Collimator helmet for Leksell Gamma Unit Models B and C.

on removable *Y–Z* positioning bars on the patients headframe. The earliest versions of the Gamma Knife had required implantation of screws into the patients skull and covering with Plaster of Paris to achieve this docking. The unit (like the two previous units) utilized a hydraulic drive to open the shielding door in the central body and propel the patient couch into treatment position.

Elekta also introduced a radiation therapy treatment plan called KULA to calculate the size and shape of the radiation volume to be treated for each patient and compute the necessary duration of the treatment. The Sophiahemmet and Karolinska Gamma Knife Centers had relied on manual calculations until this time. The KULA plan could calculate isodose lines (lines of equal radiation dose) in a two-dimensional (2D) plane, which could then be manually traced onto an axial brain image. The advent of stereotactic imaging with computed tomography also eliminated the need for the difficult and painful pneumoencephalograms and was capable of localizing brain tumors as well as anatomical targets. The University of Pittsburgh Gamma Knife Center enjoyed a relatively high degree of acceptance from the time of installation, and was soon joined by other Leksell Gamma Units in the United States, Europe, and Asia.

One drawback of the Leksell Gamma Unit Model U, which is no longer manufactured, is that it was shipped to the hospital unloaded and then loaded with cobalt-60 sources on site. This necessitated the shipment of many tons of shielding materials to create a temporary hot cell, complete with remote manipulating arms (Fig. 7). A further difficulty with Gamma Units is that the radioactive cobalt-60 is constantly being depleted by radioactive decay. The half-life is cobalt-60 is $\sim$ 5.26 years, which means that the radiation dose rate decreases $\sim$ 1%/month. The Sheffield Gamma Unit (manufactured by Nucletec) was reloaded after a number of years of use by British contractors who had not been involved in designing or building the unit and it therefore took $\sim$ 12 months to complete the task. The University of Virginia Leksell Model U Gamma Unit was the first to be reloaded (in 1995) and it was out of service for only 3 weeks. Nevertheless, the necessity of having the treatment unit down for a period of weeks after 5–6 years of operation, at a cost approaching $500,000 with a very elaborate construction scenario inhibited the early acceptance of these units. A compensating advantage of the Gamma Unit Model U was the extremely high reliability of these devices. Routine maintenance is required once every 6 months and mechanical or electrical breakdowns preventing use of the device are very rare.

Leksell introduced the Gamma Unit Model B in Europe in 1988, although it was not licensed in the United States until 5 years later. The new unit, which strongly resembles the later Model C (Fig. 8), departed from the unique spherical shape of the earlier unit and more closely resembled the appearance of a CT scanner. The source configuration was changed to five concentric rings (Fig. 6b), although the number and activity of the cobalt-60 sources remained the same as in the Model U. The new Gamma Unit Model B was designed so that the radioactive cobalt-60 sources could be loaded and unloaded by means of a special

Figure 7. Loading cobalt-60 sources into Gamma Unit with remote manipulating arms.

11 ton loading device (Fig. 9), without the necessity for creating a large and costly hot cell. This significantly reduced installation and source replenishment costs and speeded up these operations as well. The hydraulic operating system used to open the shielding doors and to move the patient treatment couch was replaced with a very quiet electrically powered system.

Extensive innovations were introduced with the Leksell Gamma Unit Model C with optional Automatic Positioning System (APS) in calendar year 2000. This unit was

Figure 8. Leksell Gamma Unit Model C, which strongly resembles the previous Leksell Gamma Unit Model B.

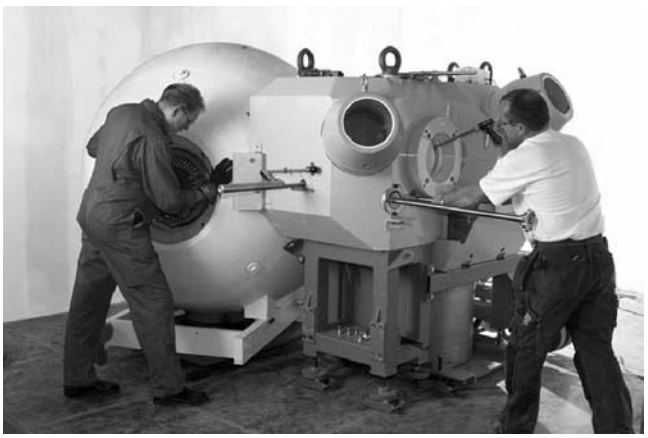


Figure 9. Special loading device for insertion and removal of cobalt-60 sources with the Leksell Gamma Units Models B and C.

awarded three American patents and one Swedish patent. The new unit provided several upgrades at once: a new computer control system operates the shielding door and patient transport mechanism and is networked via RS232C protocol with the Leksell GammaPlan treatment planning computer. All previous models required manual setting (and verification) of helmet size, stereotactic coordinates and treatment time for each shot. An optional APS system (Fig. 10) has motorized trunnions that permit the patient to be moved from one treatment isocenter to another without human intervention. This automated system can only be utilized if the secondary helmet, gamma angle (angle of patients stereotactic frame Z axis with respect to the longitudinal axis of symmetry of the Gamma Unit helmet) and patient position (prone or supine) are identical to the values for these respective treatment parameters as provided in the final approved treatment plan. In addition, the isocenters (or shots) are grouped into "runs" having stereotactic coordinates within a predefined distance of each other (typically ± 2 cm) so as not to introduce unacceptable strain on the patient's neck while their head is being moved

Figure 10. Close-up view of trunnions and helmet of Leksell Gamma Unit model C with Automatic Positioning System in place.

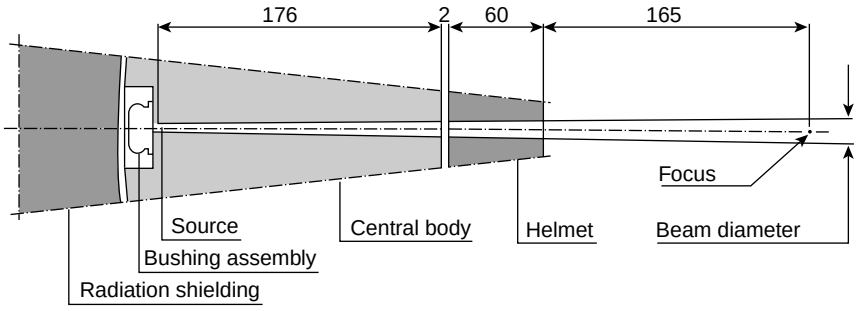


Figure 11. Geometry of sources installed in Leksell Gamma Units.

to a new position relative to the immobile body. Within these limitations the efficiency of a complex treatment plan can be greatly increased. Additionally, two independent electromechanical devices verify the positioning of the patient's stereotactic coordinates to within 50 μm (below the resolution of the unaided human eye).

THEORY

The invention of the Gamma Knife built on seven decades of previous experience with radiation therapy (see related articles). Early external beam radiation treatments used X-ray sets with low energies, in the range of 100–300 kV, which have the disadvantage of depositing a maximum dose at the surface of the skin. This physical characteristic makes it difficult to treat deep seated tumors without causing unacceptable damage to the overlying skin and tissue. Lars Leksell used a 200 kV X-ray set to treat his first radiosurgery patient in 1951, but abandoned that technique after only a few patients to work with far more penetrating proton beam radiotherapy (see article Proton beam radiotherapy). The disadvantage of proton beam therapy was that the patient had to be brought to a high energy physics laboratory, which was not otherwise equipped to treat sick or infirm patients and was often located at a great distance from the surgeon's hospital. This made treatments somewhat difficult and awkward, and the cyclotron was not always available. An important breakthrough came when two Canadian cancer centers introduced cobalt-60 teletherapy (see article Cobalt-60 units for radiotherapy) in the early 1950s. Leksell and Larsson realized that this new, more powerful radiation source could be utilized in a hospital setting. They also realized that rotational therapy, where a radiation source is revolved around a patient's tumor to spread out the surface dose, could be mimicked in this new device by creating a static hemispherical array of smaller radiation sources. Since Leksell was interested only in treating intracranial disease, where the maximum patient dimension is only $\sim$ 16 cm, the device could place radiation sources relatively close to the center of focus. The Leksell Gamma Knife uses a 40 cm Source to Surface Distance (SSD), far shorter than modern linear accelerators (see article Medical linear accelerator), which typically rotate around an isocenter at a distance of 100 cm (see Fig. 11). This short SSD allows the manufacturer to take advantage of the inverse square principle, which implies that a nominal 30-curie source at 40 cm achieves the same dose rate at the focus as a 187.5 curie source would achieve at 100 cm. This makes loading and shielding a Gamma Knife practical.

The Gamma Knife treats intracranial tumors or other targets by linear superposition of 201 radiation beams. The convergence accuracy of these sources is remarkable: The radiation focus of the beams converge at the center point of stereotactic space (e.g., 100, 100, 100 in Leksell coordinates) to within < 0.3 mm. Thus the targeting accuracy of treatment of brain tumors is essentially not limited at all by mechanical factors, and is primarily limited by the inaccuracy of imaging techniques and by target definition. Each cobalt-60 beam interacts by ionization and excitation (primarily by Compton scattering) as it passes through the skull of the patient. The intensity of each beam is diminished by $\sim$ 65% while passing through 16 cm of tissue (a typical skull width, approximated as water for purposes of calculation). At the mechanical intersection of all 201 radiation sources, which are collimated to be 18, 14, 8, or 4 mm in diameter, the useful treatment volume is formed (see Fig. 12). Outside this volume the radiation dose rate drops off precipitously (90% of Full Maximum to 50% in 1 mm for the 4 mm beam) thereby mimicking the behavior of protons at the end of their range. The mathematics of this 3D convergent therapeutic beam irradiation

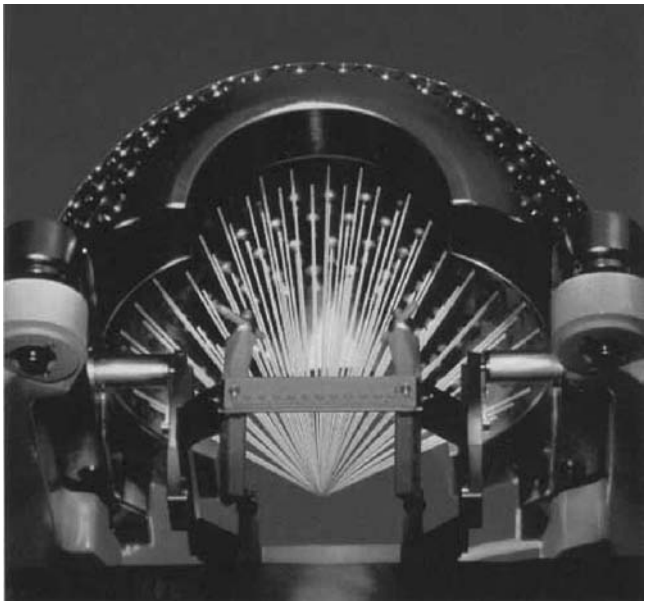


Figure 12. Illustration of convergence of 201 radiation beams to create treatment volume.

is relatively simple: The radiation absorbed dose adds up in linear superposition.

$$D(P) = D_{\mathrm{fi}} / \sum D_{\mathrm{fi}} \times [d_{\mathrm{fs}}/(d_{\mathrm{fs}} - dz)]^2 \times \mu dz$$

where $D(P)$ is the total dose at arbitrary point P, D_{fi} is the relative contribution from source i to the total dose at the point of focus, d_{fs} is the distance from the source to the focus (40 cm), dz is the distance along the beam axis from the focal point to intersection with the perpendicular from point P, and μ is the linear attenuation coefficient for Co-60 gamma radiation in tissue.

A radiation therapy treatment planning code (see article Radiation Therapy Treatment Planning) called Leksell GammaPlan is provided by the manufacturer for the purpose of preparing treatment plans for the Leksell Gamma Unit for use with human patients. An early treatment plan called KULA calculated treatment time and created a 2D plot of lines of equal radiation dose (or isodose lines), but with severe limitations. The early code could only calculate plans with a single center of irradiation (called an isocenter in general radiosurgery applications, or a "shot" in Gamma Knife usage). Calculated isodose lines had to be transferred by hand from a plot to a single CT slice in the axial plane. In 1991 the Leksell GammaPlan software was introduced (and premarket clearance by the FDA was obtained), which permitted on-screen visualization of isodose lines in multiple CT slices. The improved code could calculate and display the results of multiple shots and could model the effect of "plugging" some of the 201 source channels with thick steel plugs to "turn off" certain radiation sources. The software was written for UNIX workstations and has rapidly become increasingly powerful and much more rapid as processing speed and computer memory increased in the last decade. Leksell GammaPlan utilizes a matrix of $> 27{,}000$ equally spaced points (in the shape of a cube), which can be varied from 2.5 cm on a side to 7.5 cm on a side. Within this cube a maximum radiation dose is computed from the linear superposition of all 201 radiation beams (some of which may be plugged), from collimator helmets of 18, 14, 8, or 4 mm diameter, and this calculation is integrated over each "shot". More than 30 different shots (each with a discrete X, Y, and Z stereotactic coordinate, in 0.1 mm increments) can be computed and integrated, with user selectable relative weighting of each shot. Whereas the original KULA plan required $\sim$ 15 min for one single shot calculation, modern workstations with Leksell GammaPlan can now compute 30 shot plans in up to 36 axial slices in < 1 min. Leksell GammaPlan can now utilize stereotactic CT, MRI, and Angiographic studies in the planning process. Each study must be acquired with the stereotactic frame in place and registered separately. Image fusion is now available. Figures 13 and 14 give two of the many possible screen presentations possible with a very sophisticated Graphical User Interface.

CLINICAL USE OF GAMMA KNIFE

The Gamma Knife has gained widespread acceptance in the neurosurgical and radiation oncology community as an effective treatment for many different pathologies of brain tumors, neurovascular abnormalities and functional disorders. Gamma Knife radiosurgery may in some cases be used as an alternative to open craniotomy while for other patients it may be used after previous surgeries have been attempted. Since Gamma Knife radiosurgery infrequently requires overnight hospitalization, and generally has a very low probability of adverse side effects, it may in many cases be much less costly, with lower chance of complication and much less arduous recovery.

A typical Gamma Knife procedure is performed after a patient has been carefully screened by a neurosurgeon, a radiation oncologist, and a neuroradiologist. The procedure is scheduled as an elective outpatient procedure and typically lasts from 3 to 8 h. The first critical step is placement of a stereotactic frame (see article Stereotactic Surgery) to provide rigid patient fixation and to allow stereotactic imaging to be performed with special fiducial attachments. The exact position of the frame (offset to left or right, anterior or posterior) is crucial, since the tumor must be well centered in stereotactic space to permit

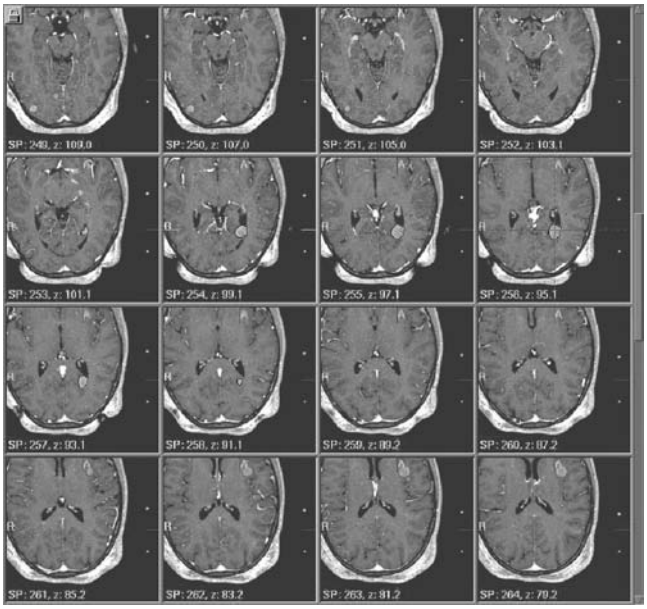


Figure 13. Screen from Leksell GammaPlan illustrating multiple MRI slices with superimposed isodose display.

Figure 14. Screen from Leksell GammaPlan illustrating angiographic study and three-dimensional display of AVM nidus.

accommodation of the patient's skull (with the frame attached) inside the small volume of the secondary collimator helmet. The Leksell Model G frame encloses ~ 2900 cm^3 of accessible stereotactic space, significantly less than other stereotactic frames which do not have to fit inside the Gamma Knife helmet. A stereotactic localization study is immediately performed, using one or more modalities such as computed tomography, and magnetic resonance imaging. Patients with vascular abnormalities may also undergo an angiographic study: A radiologist inserts a thin catheter (wire) into a vein in the patient's groin and then carefully advances the wire up through one of the major arteries leading to the brain, then into the area of interest. The catheter is then used to inject X-ray opaque dye, which reveals the extent of the vascular lesion (see Fig. 14). These imaging studies must then be promptly networked (via a hospital PACS system) to the planning computer. There the images are registered from couch coordinates (left-right, in-out, up-down) to stereotactic space (X, Y, and Z). At that point, each individual point in the brain corresponds to a specific stereotactic coordinate, which can be identified from outside the brain.

Gamma Knife radiosurgery, both in the United States and worldwide, has enjoyed a very rapid acceptance since the first commercial unit was produced in 1987. The number of procedures performed annually, subdivided by indication, is compiled by the nonprofit Leksell Society. Only results voluntarily reported by Gamma Knife centers are tallied, with no allowance for nonreporting centers, so their statistics are conservative. The growth in use of this device has been explosive, with < 7000 patients treated worldwide by 1991 and $> 297{,}000$ patients reported treated through December, 2004 (see Fig. 15). This parallels the increase in number of installed Leksell Gamma units, going from 17 in 1994 in the United States to 96 centers by the end of 2004. The number of Gamma Knife cases reported performed in the United States has increased by an average of 17%/year for the last 10 years, a remarkable increase. Table 1 indicates the cumulative number of patients treated with Gamma Knife radiosurgery in the western hemisphere and worldwide through December, 2004, subdivided by diagnosis.

Treatment objectives for Gamma Knife patients vary with the diagnosis. The most common indication for treatment is metastatic cancer to the brain. An estimated

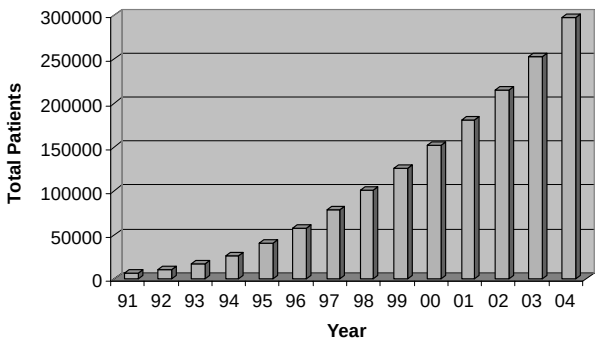


Figure 15. Cumulative number of Gamma Knife patients treated worldwide from 1991 through December, 2004.

Table 1. Cumulative Reported Gamma Knife Radiosurgery Procedures through December, 2004

Indication	Western Hemisphere Procedures	Worldwide Procedures
AVM and other vascular	9,793	43,789
Acoustic neuroma	7,719	28,306
Meningioma	11,016	36,602
Pituitary adenoma	3,577	24,604
Other benign tumors	3,137	14,884
Metastatic brain tumors	29,285	100,098
Glial tumors	7,727	20,614
Other malignant tumors	1,501	6,492
Trigeminal neuralgia	11,609	17,799
Other functional disease	1,135	4,441
TOTAL INDICATIONS:	***67,336***	***297,529***

1,334,000 cancers (not including skin cancers) were diagnosed in the United States in calendar year 2004. Approximately 20–30% of those patients will ultimately develop metastatic tumors in the brain, which spread from the primary site. These patients have a survival time (if not treated) of 6–8 weeks. The treatment objective with such patients is to palliate their symptoms and stop the growth of known brain tumors, thereby extending lifespan. A recent analysis (9) reported a median survival of patients treated with radiosurgery of 10.7 months, a substantial improvement. Approximately 18,000 patients were diagnosed with primary malignant brain tumors in the United States in calendar year 2004, with 13,000 deaths from this cause. Patients with primary malignant brain tumors (i.e., those originating in the brain) have a lifespan prognosis varying from 6 months to many years, depending on the grade of the pathology. Many glioma patients are offered cytoreductive brain surgery to debulk the tumor and may have an extended period of recovery and significant loss of quality of life afterwards. At time of tumor recurrence for these patients, the noninvasive Gamma Knife procedure may accomplish as much as a second surgery, while sparing the patient the debilitation of such a procedure. Recent reports in the clinical literature indicate that Gamma Knife radiosurgery is effective in improving survival for glioma patients.

Many patients with nonmalignant brain tumors are also treated with Gamma Knife radiosurgery. Meningiomas are the most common nonmalignant brain tumor, arising from the meninges (lining of the brain) as pathologically altered cells and causing neurological impairment or even death. Approximately 7000 new meningiomas are diagnosed in the United States each year. Most grow very slowly (~ 1 mm $\cdot$ year^{-1}) while the most aggressive tumors may grow rapidly to as much as 12–15 cm in length and may even invade the bone. Gamma Knife radiosurgery has been reported for treatment of meningioma as far back as 1987 and is considered a well-established treatment for this extremely persistent disease, with > 1000 Gamma Knife treatments reported for meningioma in the United States during calendar year 2001. Another common nonmalignant tumor is the acoustic neuroma (also called vestibular schwannoma), which arises from the auditory nerve

(cranial nerve VIII). It can cause deafness and imbalance, and in severe cases motor impairment as it compresses the brainstem. The incidence of newly diagnosed acoustic neuromas is 2500–3000/year in the United States. Craniotomy for acoustic neuroma is among the most challenging brain operations, typically requiring 8–24 h on the operating table. Potential complications range from loss of residual hearing to devastating facial palsy to cerebrospinal fluid leak requiring as much as 30 days of hospitalization. Extremely high control rates of up to 97% (no additional tumor growth or moderate shrinkage) have been reported for Gamma Knife radiosurgery of these tumors with extremely low complication rates (10). This may explain why >1000 acoustic neuromas were treated with Gamma Knife radiosurgery in the United States during Calendar Year 2003, nearly one-third of all such tumors diagnosed that year.

Arteriovenous malformations are a rare genetic disorder of the vascular system of the brain and spinal cord. Estimates of incidence ranges from 5 to > 600/100,000 people. The lesion consists of a tangle of abnormal arteries and veins that may not be detected until late in life. The AVMs can cause debilitating headaches, epileptic seizures, coma, and even sudden death due to cerebral hemorrhage. Arteriovenous malformations were first described in the 1800s and the first surgical resection was credited to Olivecrona in Stockholm in 1932. The AVMs were categorized by Spetzler and Martin into five distinct surgical categories in order of increasing surgical risk and one additional category for inoperable lesions (11). Surgery for these lesions remained extremely challenging until late in the twentieth Century. Therefore, when angiography became available in the 1960s, Ladislau Steiner (a neurosurgical colleague of Lars Leksell at Karolinska Hospital) began to treat AVMs with the Gamma Knife as early as 1970 (12). A large number of AVMs were treated in the early days of Gamma Knife radiosurgery both because of the extreme risk of open surgery and the early success with this technique in obliterating these lesions. Recent clinical studies report an obliteration rate for these lesions of 75–85% within 3 years of Gamma Knife radiosurgery, with similar obliteration rates if a second Gamma Knife treatment is necessary. Over 33,000 AVMs have been treated with Gamma Knife radiosurgery worldwide, making it the second most common indication after metastatic brain tumors.

Trigeminal neuralgia is a neurological condition marked by excruciating pain of the fifth cranial nerve that enervates the face in three branches between the eyebrows and the jawline. The pain may be caused by a blood vessel pressing on a nerve, by a tumor, by multiple sclerosis, or for unknown reasons. This is the first condition ever treated by Lars Leksell, using a 200 kVp X-ray unit affixed to a stereotactic frame in a treatment performed in 1951. The root entry zone of the nerve as it enters the brainstem is the target volume. The nerve diameter at that point is only 2–4 mm and the consequences of a geometric miss with the radiosurgery treatment volume accidentally being directed to the brainstem could be quite severe. Alternative treatments include injection of glycerol into the cistern under radiographic guidance, radio frequency "burn" of the nerve under radiographic guidance and microvascular decompression which is a fairly major brain surgery. Physicians at the University of Pittsburgh recently reviewed their first 10 years of treatments on 202 trigeminal neuralgia patients and found that > 85% had complete or partial relief of pain at 12 months after Gamma Knife radiosurgery (13). Over 12,500 patients with trigeminal neuralgia have been treated with Gamma Knife radiosurgery at this writing.

QUALITY CONTROL/QUALITY ASSURANCE

Quality Control and Quality Assurance for Gamma Knife radiosurgery is of critical importance. Unlike fractionated radiation therapy, Gamma Knife treatments are administered at one time, with the full therapeutic effect expected to occur in weeks, months, or years. Errors in any part of the radiosurgery process, from imaging to planning to the treatment itself could potentially have severe or even fatal consequences to the patient. An international group of medical physicists published a special task group report on Quality Assurance in stereotactic radiosurgery in 1995 (14) and the American Association of Physicists in Medicine discussed Quality Assurance for Gamma Knife radiosurgery in a task group report in that same year (15). Each group stressed the need for both routine Quality Control on a monthly basis, examining all physical aspects of the device, and calibration of radiation absorbed dose measurements with traceability to national standards. Both groups also emphasized detailed documentation and independent verification of all treatment parameters for each proposed isocenter before the patient is treated. An Information Notice was published by the U.S. Nuclear Regulatory Commission (NRC) on December 18, 2000 that documented 16 misadministrations in Leksell Gamma Knife radiosurgery cases in the United States over a 10-year period (16). The Nuclear Regulatory Commission defines a misadministration as "A gamma stereotactic radiosurgery radiation dose: (*1*) Involving the wrong individual, or wrong treatment site; or (*2*) When the calculated total administered dose differs from the total prescribed dose by > 10% of the total prescribed dose." Fifteen of the 16 incidences were ascribed to human error while utilizing the Leksell Gamma Knife models U, B, and B2. Six of the reported errors involved setting incorrect stereotactic coordinates (often interchanging Y and Z coordinates). Two errors occurred when the same shot was inadvertently treated twice. One error involved interchanging left and right side of the brain. One error involved using the wrong helmet. No consequences to patients were reported, but would be expected to be minor in most of the reported cases.

It is important to note in this respect that the new Leksell Gamma Unit Model C with (optional) Automatic Positioning System has the potential to eliminate many of the reported misadministrations. The older Leksell Gamma Unit Models U and B are manual systems in which the treatment plan is printed out and hand carried to the treatment unit. Stereotactic coordinates for each of the isocenters (shots) are set manually by one clinician and checked by a second person. It is thus possible to treat the

patient with the wrong helmet, prone instead of supine, wrong gamma angle, incorrect plugged shot pattern, wrong time, or to repeat or omit shots. The operation of the new Model C Gamma Unit is computer controlled. The Leksell GammaPlan workstation is networked via an RS232C protocol with full error checking, thus transferring the treatment plan electronically. The shots may be treated in any order, but no shot may be repeated and the screen will indicate shots remaining to be treated. The helmet size is remotely sensed and treatment cannot commence if an incorrect helmet is selected. If the optional Automatic Positioning System is used, the X, Y, and Z stereotactic coordinates are automatically sensed to within 0.05 mm. The device will not permit treatment until the X, Y, and Z coordinates sensed by the APS system match those indicated on the treatment plan. Thus, it appears that all of the 15 misadministrations due to human error as reported by the Nuclear Regulatory Commission would have been prevented by use of the Model C with APS.

RISK ANALYSIS

The concept of misadministration should be placed in the larger concept of risk analysis. All medical procedures have potential adverse effects and, under state laws, patients must be counseled about potential consequences and sign an informed consent before a medical procedure (even a very minor procedure) may be performed. The relative risk of misadministration of Gamma Knife misadministration may be computed, utilizing the NRC report and data from the Leksell society on number of patients treated per year in the United States. Since $\sim$ 28,000 patients received Gamma Knife radiosurgery between 1987 and 1999, while 16 misadministrations were reported during the same interval, a relative risk of misadministration of 0.00057 per treatment may be computed for that period. Using the most recent year (1999) for which both NRC and patient treatment data are available, the relative risk drops to 0.0001/patient treatment.

These risks may be compared with other risks for patients undergoing an alternative procedure to Gamma Knife radiosurgery, namely, open craniotomy with hospital stay (17). A report by the National Institute of Medicine estimates that medical errors kill between 44,000 and 98,000 patients/year in the United States (18). These deaths reportedly occur in hospitals, day-surgery centers, outpatient clinics, retail pharmacies, nursing homes, and home care settings. The committee report states that the majority of medical errors do not result from individual recklessness, but from basic flaws in the way the health system is organized. A total of 33.6 million hospital admissions occur in the United States each year, which yields a crude risk estimate range of 0.0013–0.0029 death per admission to hospital or outpatient facility.

A second source of inadvertent risk of injury or death must also be considered. The National Center for Infectious Diseases reported in December, 2000 that an estimated 2.1 million nosocomial (hospital based) infections occur in the United States annually (19). These infections are often drug resistant and require extremely powerful antibiotics with additional adverse effects. Given that there are 31 million acute care hospital admissions annually in the United States, the relative risk of a hospital based infection may be computed as 0.063/patient admission, or roughly one chance in 16. The risk of infection from craniotomy was given by the same report as 0.82/100 operations for the time period January, 1992–April, 2000.

The Leksell Gamma Knife Model C system is one example of a computer-controlled irradiation device. The rapidly developing field of Intensity Modulated Radiation Therapy (IMRT) is the subject of a separate article in this work. These complex treatments require extraordinary care on the part of treatment personnel to minimize the possibility of misadministration. Only rigorous Quality Assurance and Continuing Quality Improvement in radiation oncology can make such treatments safe, reliable and effective. Leveson has studied the use of computers to control machinery which could potentially cause human death or injury, such as linear accelerators, nuclear reactors, modern jet aircraft and the space shuttle (20).

EVALUATION

The Leksell Gamma Knife, after a long period as a unique invention of limited applicability, has enjoyed explosive growth in medical application in the last 10 years. It is one of a number of medical instruments specifically created to promote minimally invasive surgery. Such instruments subject human patients to less invasive, painful, and risky procedures, while often enhancing the probability of success or in fact treating surgically inoperable patients. Over 297,000 Gamma Knife treatments have been performed worldwide as of the last tally. Most treatments are successful in achieving treatment objectives in 85–90% of patients treated. Although the Gamma Knife is the most massive and probably the costliest single medical product ever introduced, it has enjoyed widespread commercial and clinical success in 31 countries. The simplicity and reliability of operation of the unit make its use an effective treatment strategy in lesser developed countries where difficult craniotomies may not be as successful as in more developed countries. The newest version of the unit addresses the issues of automation, efficiency, treatment verification, and increased accuracy. The instrument appears to be well established as an important neurosurgical and radiological tool.

BIBLIOGRAPHY

Cited References

1. Wu A, et al. Physics of Gamma Knife Approach on Convergent Beams in Stereotactic Radiosurgery. Int J Rad Oncol Biol Phys 1990;18:941–949.
2. Greenblatt SH. The crucial decade: modern neurosurgery's definitive development in Harvey Cushing's early research and practice. 1900 to 1910, J Neurosurg 1997;87:964–971.
3. Schaltenbrand G, Wahren W. Atlas for stereotaxy of the human brain. 2nd ed. Stuttgart: Thieme; 1977.
4. Leksell L. The Stereotactic Method and Radiosurgery of the Brain. Acta Chir Scand 1951;102:316–319.

5. Leksell L. Stereotaxis and radiosurgery: an operative system. Springfield: Thomas Publishers; 1971.
6. Rand RW, Khonsary A, Brown WJ. Leksell stereotactic radiosurgery in the treatment of eye melanoma. Neurol Res 1987; 9:142–146.
7. Walton L, Bomford CK, Ramsden D. The Sheffield stereotactic radiosurgery unit: physical characteristics and principles of operation. Br J Rad 1987;60:897–906.
8. Bunge HJ, Guevara JA, Chinela AB. Stereotactic brain radiosurgery with Gamma Unit III RBS 5000. Barcelona Proceeding 8th European Congress Neurology Surgery; 1987.
9. Sanghavi SN, Miranpuri BS, Chapell R. Multi-Institutional Analysis of Survival Outcome for Radiosurgery Treated Brain Metastases, Stratified by RTOG RPA Classification. Int J Rad Oncol Biol Phys October, 2001;51(2):426–434.
10. Flickinger JC, et al. Results of acoustic neuroma radiosurgery: an analysis of 5 years' experience using current methods. J Neurosurg Jan, 2001;94(1):141–142.
11. Spetzler RF, Martin NA. A proposed grading system for arteriovenous malformations. J Neurosurg 1986;65(4):476–83.
12. Steiner L, Lindquist C, Adler JR, Torner JC, Alves W, Steiner M. Clinical outcome of radiosurgery for cerebral arteriovenous malformations. J Neurosurg 1992;77(1):1–8.
13. Kondziolka D, Lunsford LD, Flickinger JC. Stereotactic radiosurgery for the treatment of trigeminal neuralgia. Clin J Pain 2002;18(1):42–7.
14. Lutz W, Arndt J, Ermakov I, Podgorsak EB, Schad L, Serago C, Vatnitsky SM. Quality Assurance Program on Stereotactic Radiosurgery. Berlin: Springer; 1995.
15. Schell MC, Bova FJ, Larson DA, Leavitt DD, Lutz WR, Podgorsak EB, Wu A. Stereotactic Radiosurgery. AAPM Report No. 54, College Park: American Association of Physicists in Medicine; 1995.
16. Medical Misadministrations Caused by Human Errors Involving Gamma Stereotactic Radiosurgery, Bethesda. U.S.: Nuclear Regulatory Commission; 2000.
17. Goetsch SJ. Risk Analysis of Leksell Gamma Knife Model C with Automatic Positioning System. Int J Rad Oncol Biol Phys March, 2002;52(3):869–877.
18. To Err is Human: Building a Safer Health System. Bethesda: National Academy Press; 2000.
19. Geberding JL. Semiannual Report. National Nosocomial Infections Surveillance (NNIS) System. Rockville; U.S.: Public Health Service; June, 2000.
20. Leveson N. Safeware: System Safety and Computers. Reading: Addison-Wesley; 1995.

Reading List

De Salles AAF, Lufkin R, Minimally Invasive Therapy of the Brain. New York: Thieme Medical Publishers, Inc.; 1997.
Pollock B. Contemporary Stereotactic Radiosurgery: Technique and Evaluation. Oxford: Futura Publishing Company; 2002.
Coffey R, Nichols D. A Neuroimaging Atlas for Surgery of the Brain: Including Radiosurgery and Stereotaxis. Philadelphia: Lippincott-Raven; 1998.
Ganz J. Gamma Knife Radiosurgery. 2nd ed. New York: Springer-Verlag; 1997.
Webb S. Physics of Three-Dimensional Radiation Therapy: Conformal Radiotherapy, Radiosurgery and Treatment Planning. London: Institute of Physics Publishing; 1993.
Lunsford LD, Kondziolka D, Flickinger JC. Gamma Knife Brain Surgery. S. Karger Publishing; March, 1998.

See also COBALT 60 UNITS FOR RADIOTHERAPY; RADIOSURGERY, STEREOTACTIC.

GAS AND VACUUM SYSTEMS, CENTRALLY PIPED MEDICAL

BURTON KLEIN
Burton Klein Associates
Newton, Massachusetts

INTRODUCTION

Terminal gas outlets or vacuum inlets are as common a fixture today in hospital rooms as stethoscopes. Even clinics, outpatient surgery facilities, and some nursing homes utilize them. But how did they get there? And have they helped medical and nursing staff give better patient care?

This article is intended to give readers a brief look at how and why these systems were developed, how they operate, what hazards they pose, what standards have been developed to mitigate hazards as well as to standardize operation, and why maintenance of these systems is very important. In a sense, medical gas and vacuum systems are a reflection, in part, of how the practice of medicine has changed over the past 60–70 years: Both systems have become more complex and sophisticated in order to meet and treat more serious illnesses.

The systems discussed below are those involving the distribution of pressurized gases (or suctioning of air)or the creation of a vacuum via rigid metal pipes, with the source of gas or suction *not* in the same room as the end-use terminals of the system. Further, the description of these systems is a generalized one; specific systems may have different operating characteristics to meet a particular need. The authority(ies) having jurisdiction (AHJ) should be consulted for specific locations (e.g., hospital, clinic, nursing home) and application purpose (medical surgical, dental, laboratory, veterinary).

Finally, the limited bibliography provided at the end of this article has been included (*1*) for readers who wish to pursue this subject further, and (*2*) to show the various standards organizations involved in setting standards that are used in designing, installing and using these systems.

GAS SYSTEMS (PRESSURIZED)

To understand how and why the piping of medical gases to operating rooms and other patient care areas came into practice, it is necessary to briefly review how the practice of medicine, and in particular the practice of anesthesiology, changed from the mid-1800s to the early 1900s, for it was advances in administering anesthesia that led to the piping of gases into operating rooms, and from there to many other patient care areas.

Some History

The first public demonstration of inhalation anesthetics took place on October 16, 1846 at the Massachusetts General Hospital in Boston. There had been some experimentation prior to this date, but this publicized demonstration by Dr. John W. Collins clearly showed that patients could

be kept unconscious as long as necessary and have surgery performed without their sensing pain. A giant step forward in the practice of medicine had been achieved.

These first years of anesthesiology were relatively simple in that only a sponge soaked in ether and placed over the nose and mouth of patients was used to induce anesthesia. In 1868, Andrews introduced oxygen mixed with nitrous oxide as an adjunct to inhalation anesthesia. In 1871, cylinders of nitrous oxide became available. In 1882, cyclopropane was discovered, though it was not until the 1930s that it was found useful for anesthesia. And in 1887, Hewitt developed the first gas anesthesia machine that used compressed gases in cylinders.

This controlled unconsciousness sparked a dramatic increase and change in medical practice and hospital activities. No longer did patients enter a hospital only for terminal care or to feel the cutting edge of a scalpel. Problems occurring inside the body could now be exposed for examination and possible correction. And, as anesthesia systems became more available and sophisticated, the volume and type of operations increased dramatically. Finally, the discovery that oxygen enrichment helped patients during anesthesia and operations increased the use of oxygen in operating rooms tremendously.

By the 1920s, cylinders of oxygen and nitrous oxide were constantly in motion about hospitals, from loading docks to storage rooms to operating rooms and back again. But, occasionally, cylinders did not make the entire circuit in one piece. Thus, the question occurred to healthcare staff: Was there another, better way to provide gas in operating rooms?

Some sources credit the late Albert E. McKee, who was working with a Dr. Waters at the University of Wisconsin Medical Center in the 1920s, with installing the first medical piped gas system that used high pressure cylinders of oxygen connected by pipes to outlets in nearby operating rooms. He (and his counterparts) saw this as a better method of providing gases to operating rooms. Their installation had some very positive effects (it also had some negative ones that will be discussed shortly):

1. There was an immediate reduction in operating costs. Instead of many small cylinders, fewer but larger cylinders could be utilized, with concurrent reduction in the unit cost per cubic foot of gas. (It has been reported to this author that the amount saved after McKee installed his system was sufficient to pay the salaries of the University of Wisconsin anesthesiology departmental staff.) Fewer cylinders also meant less loss of residual gas that remained in empty cylinders. When individual cylinders were used, they would be replaced when the pressure inside the cylinder dropped down to $\sim$500 psi ($lb \cdot in^{-2}$ or 3448 kPa); when two or more cylinders were manifolded together as a source, however, individual cylinders could be allowed to go down to $\sim$ 40 psi (276 kPa), since there were other cylinders in the system from which gas could be drawn.
2. This method provided immediate access to gases. Operating room staff only needed to connect hoses to gas outlets to obtain gas. The large supply at the central dispersion point could be monitored by one person (instead of each anesthesiologist worrying about their own individual small cylinders). Since several large cylinders were grouped together, when one became empty, or nearly empty, others could be switched on line and the empty one replaced. Thus, operating room staff were assured of a constant supply of gas.
3. Safety was improved. No longer were cylinders, with their inherent hazards, inside the operating room. Cylinder movement around the hospital was dramatically reduced.

Industry had been using gas under pressure in pipes since the late 1800s (e.g., street lamps). Piping gases around a hospital was, thus, a natural extension of this methodology, though components had to be modified to meet medical needs. These new installations were not without problems, however. The system had to be leak-free, since an escape and buildup of gases (flammable or oxidizing) within a building was dangerous. Also, having these gases carried in pipes around a healthcare facility meant that an incident in one place now had a means of becoming an incident in another place. Finally, if more than one gas were piped, the possibility of cross-connection and mixing of gases existed (and cross-connecting of some gases can create explosive possibilities).

This last problem was of particular concern since initially there was no restriction on the piping of flammable anesthetic gases. Several institutions, including the University of Wisconsin, installed systems to pipe ethylene gas. Even though the standardization of terminal connectors began in the late 1940s, explosions in operating rooms continued to occur. While the number of such incidents was not large, the occurrence was always devastating, almost always killing the patient, and sometimes maiming medical–surgical–nursing staff. In 1950, the National Fire Protection Association (NFPA) Committee on Hospital Operating Rooms proposed a number of changes, including prohibiting the piping of flammable anesthetic gases. The proposal, adopted by the NFPA membership, eliminated one possible source of explosions and fire.

The relatively recent introduction (late-1940s) of storing a large volume of oxygen on-site in a liquid state presented a new host of concerns. While large-volume storage replaced the use of many cylinders, it vastly increased the amount of oxygen in one location and introduced the hazard associated with gas in a cryogenic state (i.e., gas at an extremely low temperature).

System Components

The following is a general description of components used in piped gas systems today (Fig. 1). An actual system may not utilize all these components. However, all systems have certain minimum safety features, as discussed below. In addition, standardization of some components (e.g., threaded station outlet connections) and practices (e.g., operating pressures) has evolved over the years, which will be discussed later as well.

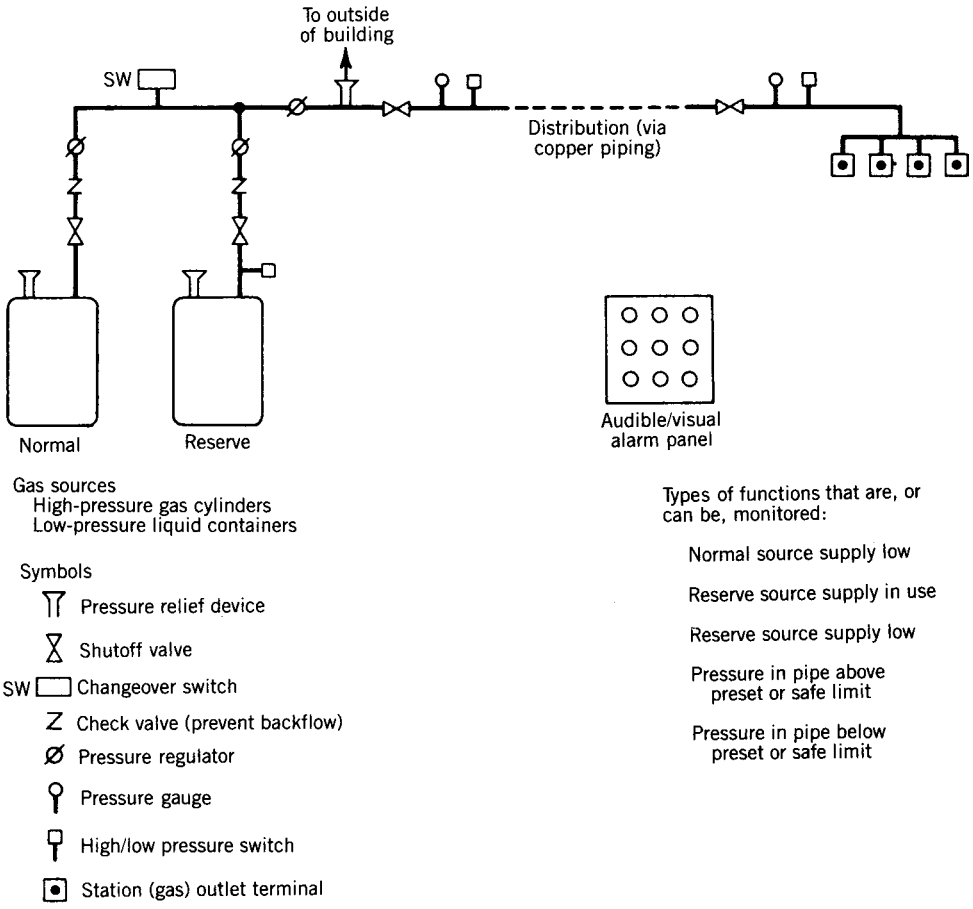


Figure 1. Components of a medical gas (pressurized) central piping system (simplified). Standards have been developed for component and total-system performance and safety.

Gases. The most common nonflammable medical gases piped under pressure today include medical air, oxygen, nitrogen, and nitrous oxide. These gases are available from manufacturers in cylinders into which a large volume of the gas has been compressed. The pressure of the gas in these cylinders can be $>$ 2000 psig (13.8 GPa). Some of these gases are also available in a liquefied state, through a refrigeration process, and are supplied in portable containers or in large stationary bulk units (tanks). (When the gas is used, it is allowed to evaporate and return to its gaseous state.) The gas in the liquefied state is placed under relatively low pressure [$\sim$ 75 psig (520 kPa)]. One gas (air) can also be obtained on-site using compressors. Whichever method is used to obtain a specific gas, it must interface with the piping portion of the system; that is, the mechanical parts must interconnect. It also means that the pressure of the source gas needs to be regulated to pressure at which the system is operating. Gas in the liquid state must be transformed to the gaseous state. For all gases, except nitrogen, a pressure between 50 and 55 psig (344 and 379 kPa) at station outlets has become the standard. For nitrogen, which is used to power nonelectric surgical tools, such as drills, bone saws, and dermatomes, a pressure between 160 and 185 psig (1103 and 1379 kPa) is used. This regulation can be likened to electricity and the use of transformers that are installed between the power generators of utility companies (where voltages upward of 10,000 V are generated) and buildings where the voltage is regulated down to 208 or 110 V. In gas systems, these transformers are called pressure regulators.

In the last few years, other nonpatient medical gases (called support gases in NFPA 99, *Standard for Health Care Facilities*) have begun to be piped. These gases are used for powering equipment that use pressurized gas in order to function (e.g., pneumatically operated utility columns). Gases in this category include nitrogen and instrument air.

Source Equipment

Other devices used at the source portion of the piped gas system include (*1*) shutoff valves at prescribed locations so that a complete or partial shutdown of a source can be accomplished; (*2*) check valves to control the direction of gas flow (i.e., one direction only); (*3*) pressure-relief valves, which are preset to vent gas to the atmosphere if the pressure in a cylinder, container, or pipeline becomes excessive enough to cause a rupture or explosion if allowed to continue to increase; and (*4*) signals to alarm panels to indicate such conditions as low and high pressure.

A separate reserve supply of the gas is also included in some piped systems. This reserve serves as a backup if the main (normal) source is interrupted or requires repair. This reserve can be adjacent to, or remote from, the main source. Its remote location precludes both sources from damage should an accident occur to one source. Such separation, however, may not always be possible.

A requirement added in NFPA 99 in the early 1990s called for a piped bulk-oxygen system that has its source sources located outside the building to have a separate connection to the piping system, also located outside of the building and accessible to a bulk oxygen delivery truck (ODT). Thus, if both the main and reserve supplies of oxygen were to fail or become damaged or depleted, the ODT could be used as the source. This emergency connection is required only for the oxygen supply because it is a life-support gas.

Finally, if extra cylinders or containers of gases are kept stored within a facility or within close proximity to a healthcare facility, a safe means of storing the gas must be provided. These storage requirements are intended to provide safety for occupants should an incident occur outside the storage room (i.e., in order to protect the cylinders from adding to the incident), or should an incident occur inside the storage room (i.e., in order to protect occupants in the building from the incident in the storage room).

Piping (Distribution) System. From the source piping is installed to distribute the gas to patient care areas. (Standards require gases piped to laboratory areas to be supplied from a separate system from gases piped to patient care areas. This is to prevent any backfeeding of gas from laboratory systems into patient care systems, and to allow for different pressures where required or desired for laboratory purposes.) Sizes and locations of main, riser, and lateral pipes should take into consideration both present and future needs or plans. As with the source, shutoff valves and flow-control devices (check valves) are required by standards at certain locations in the piping (distribution) system.

Terminal Units (Station Outlets). The endpoints (called outlets) of the piped gas system are very important since it must be very clear what gas is flowing to each outlet. To eliminate any chance of mix-up, noninterchangeable mechanical connectors have been designed for each type of gas. These different connectors are similar to the different configurations of electrical outlets for 110, 220–208 (single-phase), 220–208 V (three-phase), and so on. Labeling of gas outlets and piping is also required. Color coding of new piping became a requirement in NFPA 99 in 2005. However, it requires staff to remember the color coding scheme. It also poses problems for persons who are colorblind.

Alarm Panels/Monitoring. Because gases are relied upon for life support, system monitoring is essential and has become standard practice. Sensors and alarms are required to be installed in all critical care areas to detect if the pressure decreases or increases beyond specified limits (e.g., should someone inadvertently or deliberately close a shutoff valve). Other sensors are required to detect when the normal source and/or reserve supply are low and when the reserve supply has been switched in.

All signals are fed to one or more master alarm panels, one of which is required to be constantly monitored by facility staff. The electrical power for these alarms is to be connected to the facility's emergency power system so that alarms will continue to function if normal electrical power is interrupted. This constant surveillance is required because of fire hazards that could develop should something in the system malfunction, and for patient safety should gas delivery be interrupted. Immediate action (corrective, responsive) is necessary in either situation.

Installation of Systems. In the early 1990s, concern about the quality of the installation of medical piped gas (and vacuum) systems resulted in the technical committee responsible for piping system requirements listed in NFPA 99, *Standard for Health Care Facilities*, to revise and expand requirements for their installation. To assure the system has been installed according to the design drawings, extensive requirements were included not only for the installer, but also for a verifier who is to be totally independent of the installer, and who tested the system after everything was connected and readied for operation (i.e., for patient use).

Performance Criteria and Standards

When first installed, medical piped gas systems generally followed the practices then in use for the piping of nonmedical gases. These practices were considered adequate at the time. In 1932, the subject came to the attention of the NFPA Committee on Gases, which noted the following hazards that the installation of these systems posed for hospitals:

1. Pipes, running through an extensive portion of a building into operating rooms, carried gases that were of the flammable type (those that burn or explode if ignited) or of the oxidizing type (those that support and intensify the burning of combustibles that have been ignited).
2. A large quantity of gas in cylinders was being concentrated and stored in one area.
3. The possible buildup of potentially hazardous gas concentrations existed should the pipes leak.
4. A possible explosion in an operating room was possible if a hose on an anesthesia machine were connected to the wrong gas.
5. A compromising of patient safety existed in that a mix-up of gases could be injurious or even fatal.

This notification came in the form of identification of hazards resulted in a request by the National Board of Fire Underwriters that this Committee develop a set of guidelines on the subject. The Committee studied the subject and, in 1933, proposed "Recommended Good Practice Requirements for the Construction and Installation of Piping Systems for the Distribution of Anesthetic Gases and Oxygen in Hospitals and Similar Occupancies, and for the Construction and Operation of Oxygen Chambers." The proposed document contained guidance on acceptable types of piping, the length of pipe runs, the identification of piping, the kind of manifolds that were acceptable, and the number and location of shutoff valves. As noted

in the title, it permitted the distribution of anesthetic gases, which were flammable. The NFPA did not formally adopt the proposed Recommended Good Practice until 1934. Over the years, as more knowledge was gained from the hazards, installation, and use of piped gas systems, the NFPA standard also changed. In addition, other organizations prepared standards addressing other aspects of piped gas systems (1–10). A brief summary of these efforts follows.

National Fire Protection Association Standards. The original NFPA document, first designated NFPA 565 and later, NFPA 56F, which in turn was incorporated into NFPA 99 (11) remained unchanged until 1950 when the NFPA Hospital Operating Room Committee, working with the NFPA Gas Committee, recommended that piping of flammable anesthetic gases be prohibited. Later, specific safety requirements were added, such as for the storage of gases, shutoff valve locations, check valves, line-pressure gages, pressure switches and alarm panels, and installation and testing criteria. Performance criteria, in the form of operating pressure limits for different gases, were added because no other organization had included them in their documents, and uniformity of systems operations was helpful to both medical staff, designers, and the industry producing the equipment for these piped systems.

Other NFPA documents have been developed over the years that impact on medical piped gas systems. These include documents on the subjects of emergency electric power, bulk oxygen supplies, and building construction (11–16).

Compressed Gas Association (CGA) Standards. The CGA, an organization of manufacturers of gases and gas equipment, publishes many documents on the subject of gases. Some of these apply directly to medical gas piping systems; others are generic and affect any closed gas system. Topics addressed include gas cylinder criteria; noninterchangeable connectors for cylinders and terminal outlets; liquefied gas transfer connections; compressed gas-transfer connections; and commodity specifications for nitrogen, air, nitrous oxide, and oxygen (1–6).

Other Organizations. (7–10).

VACUUM SYSTEMS

Some History

The development of vacuum central piped vacuum systems, in place of portable suction machines, occurred over a period of time from the late-1940s to the early-1950s. These systems did not have to face the same unknowns and problems that the development of piped gases faced 20 years earlier. While they did not pose as great a threat as piped gases [i.e., they were not carrying oxidizing gases at 50 psig (344.8 kPa)], they did have their own hazards (e.g., they carried flammable and/or nonflammable oxidizing gases around a facility; created patient risks should the system stop; created possible restrictive contamination of orifices and low vacuum levels if orifices became clogged or contaminated; and created excecssive loading on emergency electrical power systems if not properly provided for in the system). Since many vacuum systems were and still are installed simultaneously with central piping systems for gases, they added to the problems associated with installing two or more systems simultaneously (e.g., cross-connections, incorrect labeling).

Until vacuum central piping systems were installed, medicine utilized small portable suction pumps that created a vacuum much the same way an ordinary vacuum cleaner creates suction (Fig. 2). A major difference, however, is the type of collection container used. For a home vacuum cleaner, a semiporous bag collects dirt; for a medical vacuum machine, a nonporous "trap" is necessary because of the products collected (e.g., body fluids of all kinds, semiliquid bulk material). In addition, a major problem with portable suction machines is the airborne bacteria it can spread as it operates. Since vacuum pumps operate on the principle of moving air from one place to another, this movement can be unhealthy in a healthcare setting where airborne bacteria can be infectious. Another problem with individual suction pumps was their need to be safe when flammable anesthetics were in use. (This ceased to be a problem as flammable anesthetics were replaced by nonflammable anesthetics in the 1960s and 1970s.) A central vacuum system eliminated these two problems, since it exhausted contaminated air outdoors and no electric motor was needed in the patient area in order to provide the vacuum. (It should *not* be concluded that portable suction pumps are no longer used. With bacteria filters now available and used on suction pumps,

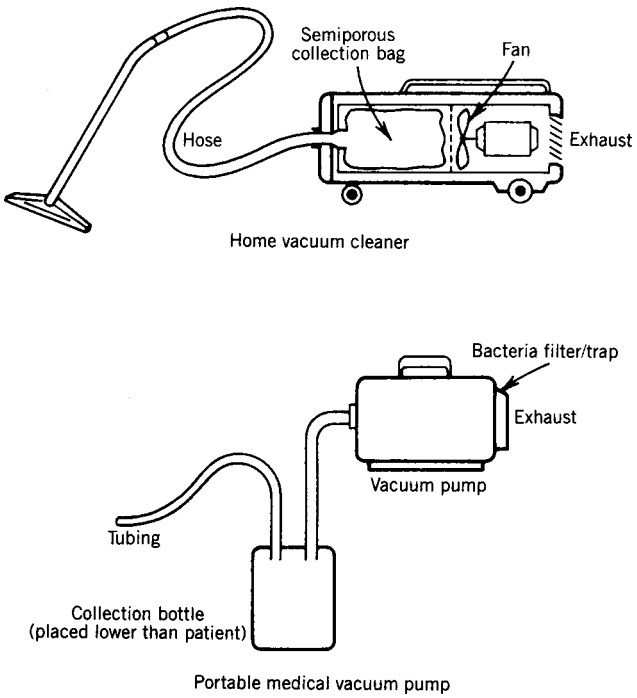


Figure 2. Home vacuum cleaner (high volume, low degree of vacuum) versus portable medical vacuum pump [low volume, low-to-high (adjustable) level of vacuum].

these devices are still quite suitable, in the same fashion that individual gas cylinders are still used.) It is necessary, however, that a trap unit be used between the patient and the vacuum control regulator and station inlet, so that nothing but air is drawn into the piping system.

The other reason central (add) vacuum systems began to be installed was the result of studies by the late David A. McWhinnie, Jr., in the early 1950s that showed the economic viability of these systems. Initially, vacuum central piped vacuum systems served only operating rooms and specialty areas, such as postanesthesia recovery rooms and emergency rooms. General patient care areas were added as demand for suction increased and economics made their installation viable. The reduction in the spread of airborne bacteria that central piped vacuum systems provided also contributed to their installation in general patient care areas as hospitals become more aware and concerned about this hazard. Pediatric and neonatal areas were last to install piped vacuum systems because of concern over what high degrees of vacuum and flow rates might do to babies (e.g., damage to very delicate tissues, possible collapse of newly functioning lungs). With improvements in the regulation of the degree of vacuum and more staff education, this concern abated, and piped vacuum systems were installed in these areas as well.

System Components

A medical piped vacuum system can be diagrammed, as shown in Fig. 3, in a fashion similar to the piped gas system described above. However, remember that the flow of subatmospheric air is opposite to the flow of pressurized gases in centrally piped gas systems. Note that piped vacuum systems require much larger orifices at inlet terminals than those at outlet terminals for gas systems because of (*1*) the pressures involved [i.e., 12 in. (30.5 cm) of Hg (40.6 kPa) (negative pressure) as opposed to 50 psi (344.8 kPa)]; and (*2*) the need for high flow. As noted previously for piped gas systems, the following description for piped vacuum systems includes the major components of a large system. Of course, individual systems will vary.

Sources for Vacuum. Pumps provide the means by which suction is created. They draw in the air that exists within the piped system, and exhaust it via a vent discharge located on the outside of the building (generally on a roof) and away from any intake vents. This configuration allows exhausted air, which may be infectious, to dissipate into the atmosphere.

At least two pumps are required to be installed, each with either one capable of providing adequate vacuum to the entire system. This redundancy is necessary to keep the vacuum system functioning in case one pump fails or needs maintenance. To smooth out pump impulses and provide a constant vacuum, a receiver (surge tank) is required to be installed at the source site between the pumps and the rest of the system. Shutoff valves and check valves are to be installed for maintenance, and efficiency, and to shut down the system (or portions of the system) in the event of an emergency.

Piping (Distribution) System. Like piped gas systems, the first standard on piped vacuum systems required metal pipes to be used to connect the various patient care areas to the receiver. And like gas systems, there were and still are prescribed locations for shutoff valves, check valves, vacuum switches, and vacuum-level gages.

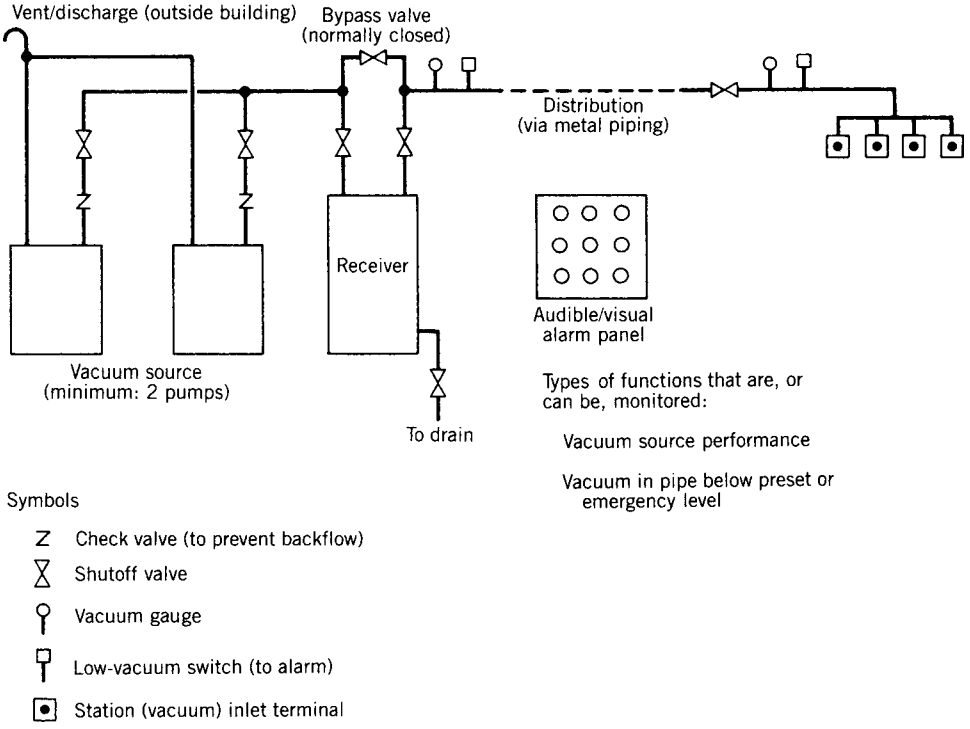


Figure 3. Components of a medical piped vacuum central piping system (simplified). Standards have been developed for component and total-system performance and safety. A collection unit and a trap are required between the inlet terminal and the patient.

However, because of the subatmospheric operating pressures and lower concentration of oxidizing gases in a piped vacuum system as opposed to a piped gas system, more types of metal pipes are allowed in the first standard on these vacuum systems. Piping for vacuum systems may have to be larger than piping for gas systems because of the level of airflow (vacuum) required by medical staff. Also, originally, the melting point allowed for joints can be lower for piped vacuum systems was permitted to be lower than the 1000° withstand-temperature required for piped gas systems. However, it is recognized that vacuum systems are sometimes installed at the same time as gas systems; as such, it may be prudent in those situations to use one type of piping throughout in order to reduce the chance of using the wrong piping and/or brazing on the piped gas system. In recent years, the committee responsible for piped vacuum system requirements has gradually required the type of piping for vacuum systems to be closer to that required for piped gas systems.

A significant difference of piped vacuum systems from piped gas systems permits connection of medical laboratories into patient care vacuum systems, though with the stipulation that the connection be made directly into the receiver and not via the pipes serving patient areas, so that a fluid trap and manual shut off valve are included. Separate systems, however, are encouraged.

Terminal Units (Station Inlets). The terminals for vacuum systems (called inlets), resemble the outlets of gas systems. Thus, it is required that they be clearly labeled vacuum or suction. To preclude problems (since piped vacuum systems sometimes are installed along at the same time with piped gas systems), the connector used for vacuum inlets is to be mechanically different from all gas outlet connectors, thereby reducing the chance of interconnection of gas and vacuum equipment.

Alarm Panels/Monitoring. Because vacuum is now a critical tool in the practice of medicine, it, too, requires constant monitoring. An audible/visual alarm panel (integrated with one for a piped gas system if also installed) alerts staff to problems similar to those of gas systems (e.g., pump malfunction, a drop in vacuum below a prescribed level).

Performance Criteria and Standards

With no vacuum standards in existence, the first piped vacuum systems installed were based on prevailing engineering expertise. While vacuum systems may seem similar to gas systems (e.g., piping, the movement of gas, although in the opposite direction), the design criteria for them are very different technically. With a piped gas system, after the source gas has been connected, the whole system reaches and stabilizes at a narrow range of pressure. In a piped vacuum system, a pump is trying to evacuate a space and provide a degree of vacuum [measured in inches of Hg (negative) *and* in volume displacement (flow)] at each inlet. In the former, the gas itself within the system provides a positive pressure and flow; in the latter, a pump is required to create a subatmospheric pressure and flow.

In the early 1950s, ineffective performance plagued many piped vacuum systems. Staff techniques, the lack of appropriate check valves, and widely divergent pump sizing contributed to the problems. One city known to have been investigating the problem was Detroit. During the 1950s, the city attempted to establish a municipal standard for the piped vacuum systems in city hospitals. Several of the major manufacturers of vacuum pumps became involved in the effort. Because general agreement could not be reached, the manufacturers suggested that industry try to develop a standard. This led to Compressed Gas Association (CGA) involvement, since many of its members were by the late 1950s supplying vacuum pumps and inlet connectors. In 1961, the CGA released a document (designated P-2.1) that included recommendations on pumps, warning systems, piping, installation, and labeling. It also included recommendations on pump sizing.

During the 1960s, staff practices were improved or standardized. This included the location of collection bottles (below patient level) and the use of regulator bypasses. This helped system performance as well. Because there continued to be differences of opinion in the engineering world regarding piped vacuum system design, the CGA approached the NFPA in the early-1970s about the NFPA developing a medical–surgical vacuum system standard. The NFPA agreed to the idea and a subcommittee of the then Committee on Hospitals was established. After tests of various pumps and suction-therapy equipment, and surveys of actual systems in hospitals, a recommended practice (designated NFPA 56K) was adopted by NFPA in 1980. After 3 years, it was revised and changed to a standard (being incorporated into NFPA 99, Standard for Health Care Facilities, at the same time) (11). The NFPA recommended practice (and then standard) generally contained the same topics as the CGA document. Other standards that impact piped vacuum systems have been developed. Most have already been mentioned or listed for piped gas systems and cover such subjects as cleaning and purging, pressure testing, and connection for emergency electrical power.

As noted, the initial criteria for installing vacuum central piped vacuum systems differed from piped gas systems. Of late, the major document on the subject (NFPA 99) has gradually revised piped vacuum system requirements, particularly on piping material, to that required for piped gas systems. But if they a piped vacuum system is are installed alongside a piped gas system at the same time the gas system is installed, the installation standards of the gas system should be considered to avoid possible degradation of the gas system, which requires more stringent standards.

Requirements on piped vacuum system design have also been deleted from the NFPA 99 document as it was seen to be outside the scope of the document (NFPA 99 is a minimum performance safety standard), as well as not changed in the document for $>$ 20 years.

MAINTENANCE OF SYSTEMS

A separate note on maintenance is deemed warranted because of the inherent hazards posed by piped gas and

vacuum systems, as well as the high reliance now placed on these systems by medical–surgical–nursing staff. In the former, everyone is affected by the hazards; in the latter, failure of these systems can place patients at considerable medical risk.

Like any system, periodic maintenance is necessary in order to assure continuous, and optimum and safe level of operation. For piped gas or vacuum systems, this includes visual inspection of exposed pipes and outlets–inlets, sampling of gases (gas systems), measurement of pressures (gas systems), measurement of flow rates (vacuum systems), and testing of alarms. Guidance on this subject is included in such documents as NFPA 99, *Standard for Health Care Facilities* (11).

While today's standards assure a high probability of a safe and reliable system, mechanical failures can and do occur, and human error or abuse still remain. Thus, should a fault of some kind occur, or a wrong connection be made, periodic maintenance should detect the condition so that corrective action can be taken before a serious incident occurs. This maintenance is particularly necessary whenever either system is breached for upgrading, component maintenance occurs, or system expansion purposesis made. The value of these systems in the treatment of patients demands no less.

Originial manuscript for this article was reviewed for technical accuracy by John M.R. Bruner, M.D., W.E. Doering, William H.L. Dornette, M.D., James F. Ferguson, Edwin P. Knox, (the late) David A. McWhinnie, Jr., Ralph Milliken, M.D., and (the late) Carl Walter, M.D.

BIBLIOGRAPHY

Cited References

1. Compressed Gas Association. Commodity Specification of Air. G-7.1, Arlington Chantilly (VA): CGA; 1973–2004.
2. Compressed Gas Association. Standard for the Installation of Nitrous Oxide System at Consumer Sites. G-8.1. Chantilly Arlington (VA): CGA; 1979–1990.
3. Compressed Gas Association. Standard Color-Marking of Compressed Gas Cylinders Intended for Medical Use in the OR, C-9. ChantillyArlington (VA): CGA; 1982–2004.
4. Compressed Gas Association. Commodity Specification for Nitrogen, G-10.1. ChantillyArlington (VA): CGA; 1985–2004.
5. Compressed Gas Association. Compressed Gas Cylinder Valve Outlet and Inlet Connections. V-1. ChantillyArlington (VA): CGA; 1977–2003.
6. Compressed Gas Association. Diameter Index Safety System, V-5. ChantillyArlington (VA): CGA; 1978–2005.
7. American Society of Sanitary Engineering. Professional Qualifications Standard for Medical Gas Systems, Installer, Inspectors, Verifiers, Maintenance Personnel and Instructors. ASSE, Series 6000: Westlake (OH); 2004.
8. American Society for Testing and Materials. Specification for Seamless Copper Water Tube. B-88, Philadelphia: ASTM; 1986–2003.
9. American Society for Testing and Materials. Specifications for Seamless Copper tube for Medical Gas Systems. B-819, Philadelphia: ASTM; 1992.
10. American Society for Testing and Materials. Standard Test Method for Behavior of Materials in a Vertical Tube Furnace at 750°C, E-136, Philadelphia: ASTM; 1982–2004.
11. National Fire Protection Association. Standard for Health Care Facilities (which includes criteria on piped medical gas systems, piped medical–surgical vacuum systems, and emergency electrical power), NFPA 99. Quincy (MA): NFPA; 1987–2005.
12. National Fire Protection Association. Standard for Bulk Oxygen System at Consumer Sites, NFPA 50. Quincy (MA): NFPA; 1985 (now included in NFPA 55, Standard for the Storage, Use, and Handling of Compressed Gases and Cryogenic Fluids in Portable and Stationary Containers, Cylinders, and Tanks; 2005).
13. National Fire Protection Association. National Electrical Code. NFPA 70, Quincy (MA): NFPA; 1987–2005.
14. National Fire Protection Association. Life Safety Code. NFPA 101, Quincy (MA): NFPA; 1985–2003.
15. National Fire Protection Association. Standard on Types of Building Construction. NFPA 220, Quincy (MA): NFPA; 1985–1999.
16. National Fire Protection Association. Standard Test Method for Potential Heat of Building Materials. NFPA 259, Quincy (MA): NFPA; 1987–2003.

Reading List

National Fire Protection Association. Historical Proceedings. Annual Meeting, Quincy (MA): NFPA; 1933.
National Fire Protection Association. Historical Proceedings. Annual Meeting, Quincy (MA): NFPA; 1934.
National Fire Protection Association. Historical Proceedings. Annual Meeting, Quincy (MA): NFPA; 1950.
National Fire Protection Association. Historical Proceedings. Annual Meeting, Quincy (MA): NFPA; 1951.
American Welding Society. Specification for Brazing Metal. A5.8, Miami (FL): AWS; 1981–2003.
American Society of Mechanical Engineers. Boiler and Pressure Vessel Code. New York: ASME; 1986–2001.

See also CODES AND REGULATIONS: MEDICAL DEVICES; EQUIPMENT MAINTENANCE, BIOMEDICAL; SAFETY PROGRAM, HOSPITAL.

GAS EXCHANGE.

See RESPIRATORY MECHANICS AND GAS EXCHANGE.

GASTROINTESTINAL HEMORRHAGE

R.C. BRITT
L.D. BRITT
Eastern Virginia Medical School
Norfolk, Virginia

INTRODUCTION

Gastrointestinal (GI) hemorrhage is a common medical problem, with significant morbidity and mortality. Traditionally, GI hemorrhage was managed by medically supporting the patient until the bleeding stopped or surgical intervention was undertaken. The modern day management of GI hemorrhage involves a multidisciplinary approach, including gastroenterologists, surgeons, interventional radiologists, primary care physicians, and intensivists. Despite the evolution in management of GI hemorrhage, the mortality rate has remained fairly constant, concentrated in the elderly with greater comorbidity (1). Additionally,

medical advances, e.g., proton pump inhibitors, H_2 blockers, antimicrobial treatment of *Helicobacter pylori*, and endoscopic management have led to a decrease in the number of operations for hemorrhage, but not in the actual number of hemorrhages (2). The incidence of upper GI bleeding has remained relatively constant at 100–150/100,000 people (3), with an estimated 300,000–350,000 admissions annually and a mortality rate of 7–10% (4). Lower GI hemorrhage accounts for 20–30% of GI hemorrhage, and typically has a lower mortality rate than upper GI bleeding.

There are three major categories of GI hemorrhage, including esophageal variceal bleeding, nonvariceal upper GI bleeding, and lower GI bleeding. Typically, upper GI bleeding is classified as that bleeding occurring from a source proximal to the ligament of Treitz, with lower GI bleeding occurring distally. When bleeding occurs in the upper GI tract, it can be vomited as bright red blood, referred to as hematemesis. Slower bleeding from the upper GI tract is often referred to as "coffee-ground emesis", which refers to the vomiting of partially digested blood. Black, tarry stool is referred to as melena, and usually originates from an upper GI source, with the black color due to the action of acid on hemoglobin. Visible blood in the stool, or bright red blood per rectum, is referred to as hematochezia. Hematochezia is usually indicative of lower GI bleeding, although brisk upper GI bleeding may also present as hematochezia. The stool may also be maroon, suggesting the blood has mixed with liquid feces, usually in the right colon.

INITIAL EVALUATION AND RESUSCITATION

Upon presentation with GI hemorrhage, two large-bore (16 gauge or larger) peripheral IVs should be placed and intravascular volume resuscitation initiated with an isotonic solution. Lactated Ringers is frequently preferred to 0.9% Normal Saline because the sodium and chloride concentrations more closely approximate whole blood. The ABCs of resuscitation are a priority in the initial evaluation of the massive GI bleed, with careful attention given to the airway because of the high incidence of aspiration. The patient must be carefully monitored to ensure the adequacy of resuscitation. In the presence of continued rapid bleeding or failure of the vital signs to improve following 2 L of crystalloid solution, the patient should also begin receiving blood. If type-specific blood is not yet available, the patient may receive O negative blood.

On presentation, blood is drawn for hematocrit, platelets, coagulation profile, electrolytes, liver function tests, and a type and cross. Caution must be used when evaluating the initial hematocrit, as this does not accurately reflect the true blood volume with ongoing hemorrhage. A foley catheter should be inserted to monitor for adequate urine output as a marker for adequate resuscitation. An NG tube should be inserted to evaluate for the presence of upper GI bleeding, as bright red blood per NG tube indicates recent or active bleeding. While clear, bilious aspirate usually indicates that the source of bleeding is not upper GI, this is not a definite as absence of blood on nasogastric aspirate is associated with a 16% rate of actively bleeding lesions found on upper endoscopy (5).

A thorough history is paramount when evaluating a patient presenting with GI hemorrhage. The clinical history may suggest the etiology of hemorrhage, as well as offer prognostic indicators. Important features in the history include a history of previous bleeding, history of peptic ulcer disease, history of cirrhosis or hepatitis, and a history of alcohol abuse. Also important is a history of medication use, particularly aspirin, nonsteroidals, and anticoagulants. Symptoms the patient experiences prior to the onset of bleeding, such as, the presence or absence of abdominal pain, can also be useful in the diagnosis.

A comprehensive physical exam must be done to evaluate the severity of the hemorrhage, as well as to assess for potential etiology. Massive hemorrhage is associated with signs and symptoms of shock, including tachycardia, narrow pulse pressure, hypotension, and cool, clammy extremities. The rectal exam may reveal the presence of bright red blood or melena, as well as evidence of bleeding hemorrhoids in a patient with bright red blood per rectum. Physical exam is also useful to evaluate for stigmata of liver failure and portal hypertension, such as, jaundice, ascites, telangiectasia, hepatosplenomegaly, dilated abdominal wall veins, and large hemorrhoidal veins.

When faced with a patient presenting with GI hemorrhage, the complete history and physical exam will help direct further management by assessing the likely source of bleed. The initial questions that must be answered to determine management priorities include whether the likely source of hemorrhage is from the upper or lower GI tract, and if the bleeding is from an upper source, whether the bleed is likely variceal or nonvariceal (Table 1).

UPPER GASTROINTESTINAL BLEEDING

Upper gastrointestinal bleeding is shown in Table 2.

Gastroesophageal Varices

Portal hypertension, defined as an increase in pressure > 5 mmHg (0.666 kPa) in the portal venous system (6), can lead to acute variceal hemorrhage. Cirrhosis, related to either chronic alcohol abuse or hepatitis, is the most common cause of portal hypertension, and leads to an increased outflow resistance, which results in the formation of a collateral portosystemic circulation. Collaterals form most commonly in the gastroesophageal junction and form submucosal variceal veins. Patients with isolated splenic vein thrombosis often form submucosal varices in the fundus of the stomach. Some 30–60% of all cirrhotic patients will have varices at the time of diagnosis, and 5–8% develop new varices each year (7). One-third of patients with varices will experience variceal hemorrhage, with mortality from the first variceal bleed as high as 50% (8). Rebleeding occurs frequently, especially in the first 6 weeks. Risk factors for early rebleeding, within the first 6 weeks, include renal failure, large varices, and severe initial bleeding with hemoglobin $< 8\,g{\cdot}dL^{-1}$ at admission (6). The risk of late rebleeding is related to the severity of liver

Table 1. Localization of Gastrointestinal Hemorrhage

Diagnosis	History	Physical Examination
	Esophagus	
Nasopharyngeal bleeding	Epistaxis	Blood in nares, blood dripping down pharynx, evidence for telangiectasias
Esophagogastric varices	Alcoholism, lived in area where schistosomiasis is endemic, history of blood transfusions or hepatitis B	Stigmata of chronic liver disease, (e.g., gynecomastia, testicular atrophy, parotid enlargement Cachexia, Kaposi's sarcoma, oral candidiasis)
Esophagitis	Dysphagia, odynophagia; immunosuppressed, (e.g., AIDS); diabetes mellitus, lymphoma, elderly	
Esophageal neoplasm	Progressive dysphagia for solids	Cachexia
Mallory–Weiss tear	Retching or vomiting prior to hematemesis	Not specific
	Stomach	
Acute gastric ulcer	Intensive care unit setting	Comatose, multiple burns, on respirator
Chronic gastric ulcer	Peak between 55 and 65 years old	Not specific
Acute hemorrhagic gastritis	History of aspirin use, intensive care unit setting	Similar to acute gastric ulcer
Gastric neoplasm	Weight loss, early satiety; obstructive symptoms	Cachexia, Virchow's node; abdominal mass
Gastric angiodysplasia	Elderly	Aortic stenosis murmur
Gastric telangiectasia	Epistaxis, family history of Osler-Weber-Rendu disease or history of renal failure	Telangiectasias on lips, buccal mucosa, palate
	Duodenum	
Duodenal ulcer	Epigastric pain	Not specific
Aortoenteric fistula	History of abdominal aortic aneurysm repair	Laparotomy scar
	Colon	
Colonic neoplasm	Often occult; if located in rectosigmoid then may have obstructive symptoms	Mass on rectal examination
Cecal angiodysplasia	Elderly, recurrent bleeding, low grade	Aortic stenosis murmur
Colonic diverticuloses	Severe, single episode of bright red blood per rectum	Not specific

failure, ongoing alcohol abuse, variceal size, and renal failure (6).

Variceal bleeding classically presents as massive, painless upper GI hemorrhage in a patient with known cirrhosis. The management of acute variceal bleeding requires attention to adequate resuscitation as well as control of the active bleeding and minimization of complications related to the bleed (Fig. 1). Early endoscopy is imperative for the successful management of variceal bleeding. Frequently, endoscopy is performed in conjunction with pharmacologic therapy. Endoscopy is essential to confirm the diagnosis of bleeding varices, as many patients with cirrhosis bleed from a source other than varices. Endoscopic sclerotherapy is effective in managing active variceal hemorrhage 70–90% of the time, and is superior to balloon tamponade or vasopressin (6). Intravariceal and paravariceal injections are equally efficacious. Sclerotherapy should be repeated at 1 week, and then at 1–3 week intervals until the varices are obliterated. Endoscopic variceal band ligation achieves hemostasis 90% of the time, and is felt to have a lower rebleeding and complication rate than sclerotherapy (9).

Table 2. Upper Gastrointestinal Bleeding

Differential Diagnosis of Upper GI Hemorrhage
Gastroesophageal varices
Mallory–Weiss tear
Esophagitis
Neoplasm esophagus, stomach, small bowel
Gastritis: stress, alcoholic, drug-induced
Angiodysplasia of stomach, small bowel
Peptic ulcer disease: stomach, duodenum
Dieulafoy ulcer
Aortoenteric fistula
Hemobilia

Pharmacologic therapy is used in conjunction with early endoscopy in massive variceal bleeding. Vasopressin, which works to cause splanchnic and systemic vasoconstriction and thus decrease portal venous flow, was traditionally used to control hemorrhage, but its use is limited by systemic side effects in 20–30% of patients (10). Vasopressin causes systemic vasoconstriction, which is particularly problematic in patients with coronary artery disease, in which vasoconstriction may induce myocardial infarction. Simultaneous administration of intravenous nitroglycerine will minimize the cardiac complications

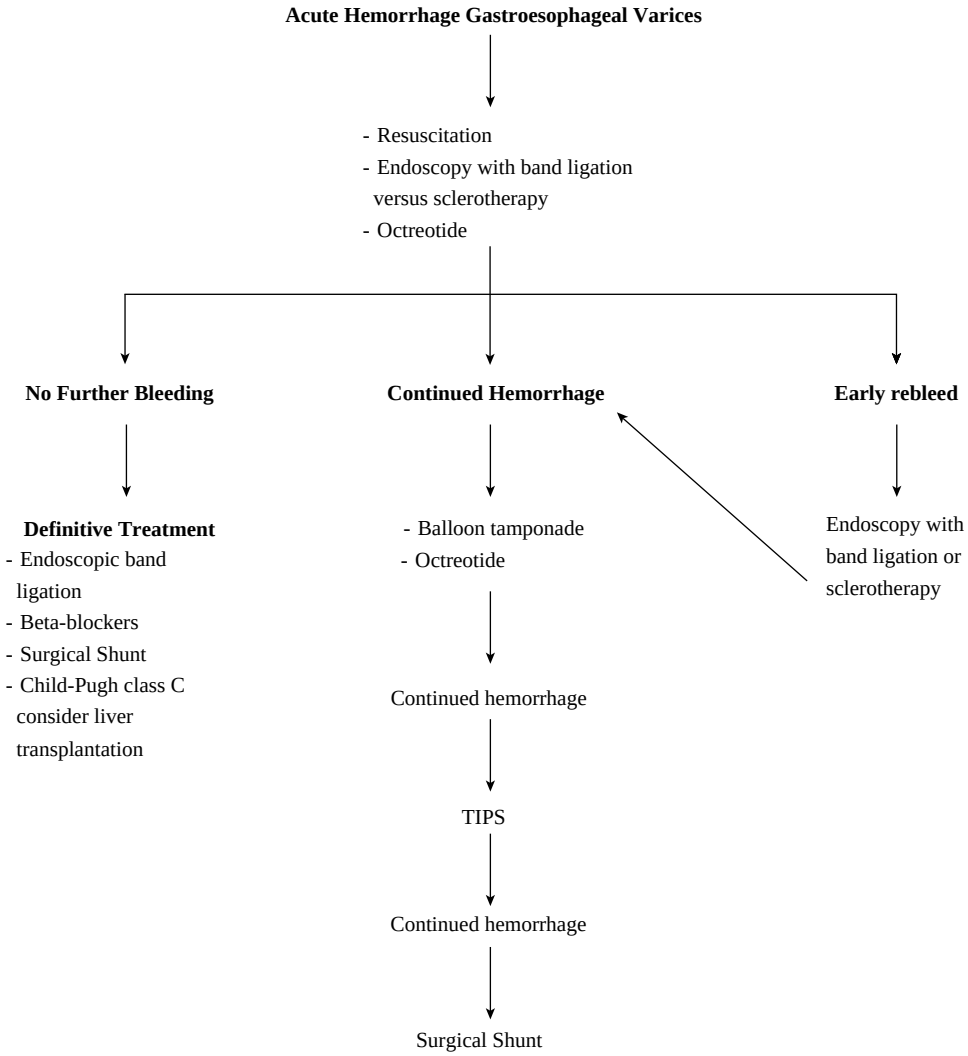


Figure 1. Management of acute variceal bleeding.

associated with vasopressin. Somatostatin, and its long-acting analog Octreotide, work via inhibition of various vasodilatory hormones, and therefore inhibit splanchnic vasodilatation and decrease portal pressure. Somatostatin is as effective as vasopressin, but without the systemic side effects (11), and is currently the medication of choice to reduce portal pressure.

Occasionally, a patient presents with massive variceal hemorrhage not amenable to endoscopic or pharmacologic therapy. Balloon tamponade, generally done with the Sengstaken–Blakemore tube, can be used to achieve short-term hemostasis, which is successful 60–90% of the time (12). Caution must be taken to secure the airway with endotracheal intubation prior to placement of the tamponade balloon because of the high risk of aspiration. Care must be used to ensure that the gastric balloon is in the stomach prior to full insufflation, as migration or inflation of the gastric balloon in the esophagus can lead to esophageal rupture. The balloon can be left in place for 24 h, at which time endoscopic band ligation or sclerotherapy can be performed.

Bleeding that cannot be controlled by endoscopic therapy or that recurs should be managed with portal pressure reduction. The initial approach currently used is the transjugular intrahepatic portosystemic shunt (TIPS), which can be done with or without general anesthesia. Potential benefits to the use of general anesthesia include advanced management of fluid dynamics by the anesthesiologist and pain management for the patient. The TIPS method works by creating a channel between the hepatic and portal veins, which is kept patent by a metal stent, which achieves

Table 3. Lower Gastrointestinal Bleeding

Differential Diagnosis of Lower GI Hemorrhage
Colonic diverticular disease
Colonic arteriovenous malformations
Neoplasm: colon, small bowel
Meckel's diverticulum
Ulcerative colitis
Crohn's disease
Colitis: infectious, ischemic, radiation-induced
Internal hemorrhoidal disease

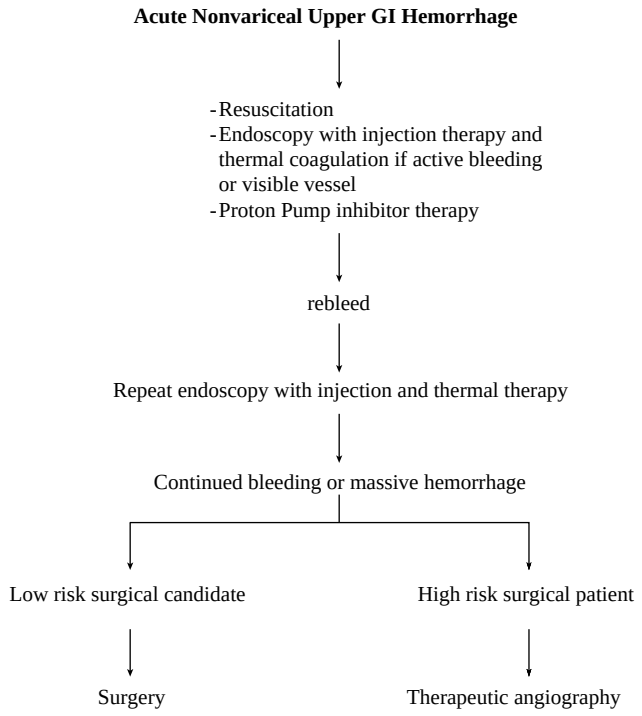


Figure 2. Management of acute nonvaricea bleeding.

hemostasis in 90% of patients (13). The downside of TIPS is related to shunt thrombosis, which can occur early or late and may result in recurrent variceal bleeding. Approximately 20% of patients by 1 year and 30% by 2 years experience recurrent bleeding related to shunt thrombosis (14,15). Complications related to TIPS procedures include a 30% rate of encephalopathy, procedure related complications including inadvertent puncture of the portal vein leading to massive hemorrhage, stent stenosis and malfunction, and TIPS-associated hemolysis.

Traditionally, reduction of portal pressure was achieved by surgical shunt procedures or devascularization. Surgical shunt procedures include nonselective shunts, which divert all the portal flow away from the liver, and selective shunts. Nonselective shunts include the portacaval end-to-side and side-to-side shunts, the central spleno-renal shunt, and the interposition portacaval shunt. Nonselective shunts are successful in achieving hemostasis in the actively bleeding patient, but frequently lead to hepatic encephalopathy as well as acceleration of liver failure. Selective shunts include the distal splenorenal shunt and the small-diameter mesocaval shunt. The selective shunts have a lower rate of encephalopathy, but are frequently complicated by uncontrollable ascites given the continued portal flow. Nonshunt operations, including esophageal transection and devascularization of the gastroesophageal junction are rarely used today. In the setting of emergent operation for ongoing hemorrhage, a nonselective portacaval shunt is most frequently employed. The distal splenorenal shunt is the most common shunt used for elective control.

Once control of active bleeding is achieved, the focus shifts to prevention of future bleeding. Endoscopic band ligation is the treatment of choice for long-term management of variceal hemorrhage (16). Beta-blockers in combination with nitrates have been shown to synergistically lower portal pressures and thus decrease the risk of rebleeding. Surgical shunting is an option in patients refractory to endoscopic or pharmacologic therapy, with the distal splenorenal shunt the most frequently used for this indication. For patients with liver failure, liver transplantation is effective for both longterm prevention of bleeding as well as hepatic decompensation and death.

NONVARICEAL UPPER GI BLEEDS

Peptic Ulcer Disease

Peptic ulcer disease is the most common cause of upper GI hemorrhage; accounting for between 40 and 50% of all acute upper GI bleeds. Major complications related to peptic ulcer disease include perforation, obstruction, and hemorrhage and occur in $\sim 25\%$ of patients, with hemorrhage the most common complication. Risk factors for peptic ulcer disease include infection with *H. pylori*, nonsteroidal antiinflammatory use, and physiologic stress related to critical illness. Medical advances including proton pump inhibitors and H2 blockers have led to a decreased need for operation for hemorrhage, but no decrease in the actual number of hemorrhages (17).

Hemorrhage related to peptic ulcer will classically present as hematemesis. In the setting of massive bleeding, the patient may also present with hematochezia. The patient may give a history of midepigastric abdominal pain preceding the bleeding. Important elements in the history include a history of peptic ulcer disease and recent usage of aspirin or nonsteroidal medications. Adverse clinical prognostic factors include age >60 years, comorbid medical conditions, hemodynamic instability, hematemesis, or hematochezia, the need for emergency surgical interventions, and continued or recurrent bleeding (18).

The initial diagnostic test on all patients presenting with an upper GI bleed should be endoscopy. Endoscopy is the best test for determining the location and nature of the bleeding lesion, provides information regarding the risk of further bleeding, and allows for therapeutic interventions. Endoscopy should be performed urgently in all high-risk patients, and within 12–24 h for patients with acute, self-limited episodes of bleeding. The goal of endoscopy is to stop the active hemorrhage and reduce the risk of recurrent bleeding. Stigmata of recent hemorrhage (SRH) are endoscopically identified features that help determine which patients should receive endoscopic therapy. The SRH include active bleeding visible on endoscopy, visualization of a nonbleeding visible vessel, adherent clot, and a flat, pigmented spot (18). Certainly, patients with the major SRH including active bleeding or a visible vessel should undergo endoscopic therapy, as meta-analysis has shown a significant reduction in rates of continued or recurrent bleeding, emergency surgery, and mortality in those who received endoscopic therapy versus those who did not (19).

A variety of modalities exist for endoscopic therapy, including injection, thermal, and mechanical therapy. Epinephrine diluted 1:10,000 is the most frequently used

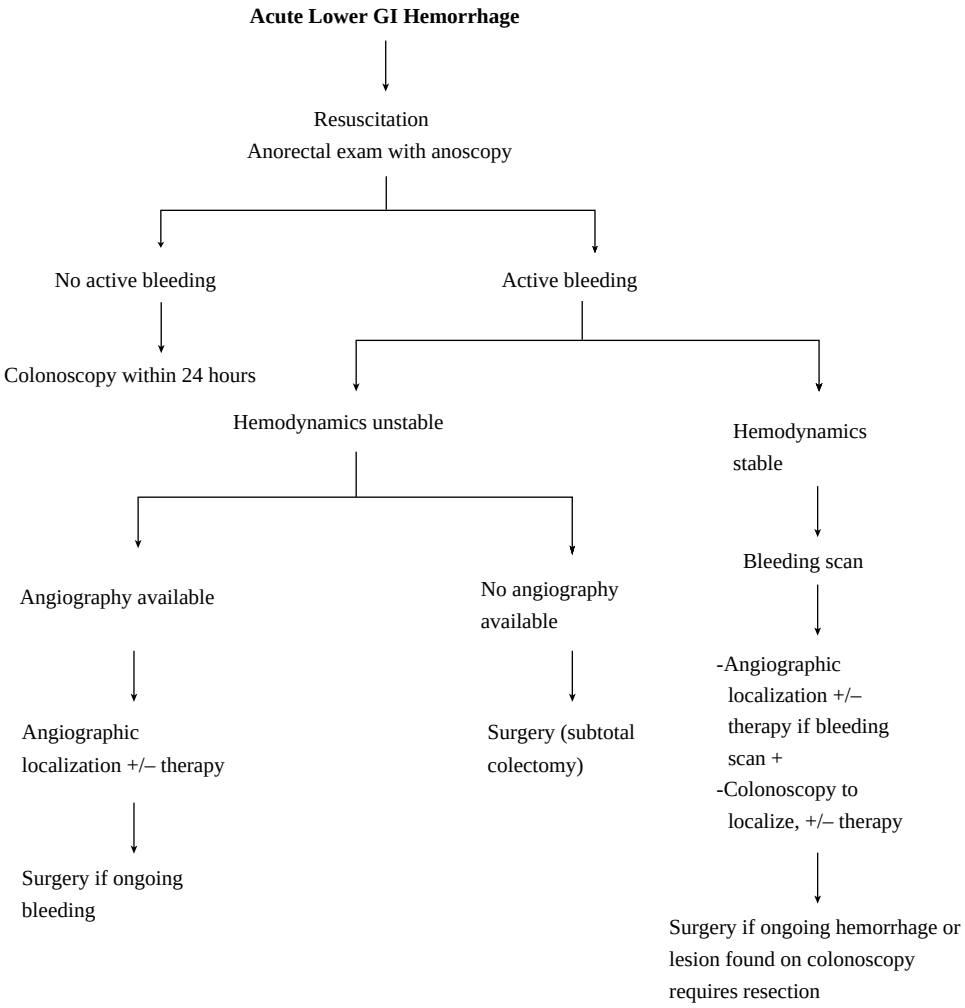


Figure 3. Management of acute lower gastrointestinal bleeding.

injection therapy, with injection into and adjacent to the bleeding point until hemostasis is achieved. Other agents used for injection therapy include ethanol, ethanolamine, thrombin, and fibrin. Thermal therapy is generally delivered by coaptive techniques, including bipolar electrocoagulation or heater probe. With coaptive coagulation, the probe is used to physically compress the vessel prior to delivery of heat to seal the vessel. Laser photocoagulation and argon beam plasma coagulation are noncoaptive techniques that are used less frequently. Mechanical therapy with hemoclips can also be used in bleeding, although the efficacy may be limited by ulcer location or a firm, scarred ulcer base preventing adequate application of the clips. A combination of injection therapy with epinephrine and bipolar thermal therapy is the most common endoscopic management of an acute bleed.

Despite initial success with endoscopic therapy, 15–20% of patients will rebleed, generally within the initial 72 h (18). While this was traditionally considered a surgical indication, endoscopic retreatment is now recommended for most patients. Repeat endoscopy rather than surgery was found in a prospective, randomized study to be associated with less complications and similar mortality (20). Surgical indications include massive hemorrhage unresponsive to resuscitation and continued bleeding unresponsive to nonoperative management. For bleeding gastric ulcers, the operation of choice is a wedge resection to include the portion of the stomach containing the ulcer with or without vagotomy. For duodenal ulcers, truncal vagotomy, pyloroplasty, and direct oversewing of the bleeding ulcer via duodenotomy is the most common operation. Care must be taken to incorporate the proximal and distal gastroduodenal artery as well as the transverse pancreatic artery.

Therapeutic angiography is an option when therapeutic endoscopy is unsuccessful and may be performed prior to surgical intervention, as is effective, less invasive than surgery, and does not impact on the ability to surgically manage the bleeding if the intervention is unsuccessful. Angiographic options include selective intra-arterial vasopressin infusion or embolotherapy with microcoils, poly-(vinyl alcohol) (PVA) particles, or gelatin sponge. Embolization is considered the first line angiographic therapy, with success rates as high as 88% (21). Vasopressin is selectively infused after bleeding has been identified by contrast extravasation at an initial rate of 0.2 units per minute, with an increase to 0.4 units per minute then 0.6 units per minute if hemostasis is not achieved. The infusion

is continued for 12–24 h, and then gradually tapered. Efficacy of vasopressin infusion ranges from 60 to 90% (22). Side effects related to selective infusion of vasopressin include abdominal cramping, fluid retention, hyponatremia, cardiac arrhythmias, and systemic hypertension. Vasopressin should not be used in patients with coronary artery disease because of the risk for myocardial ischemia.

Pharmacologic therapy to reduce gastric acidity is generally started as an adjunct to endoscopic therapy. The H2 blockers were found in meta-analysis to reduce the rate of continued bleeding, surgery, and death (23); however, a subsequent multicenter randomized trial found no difference in rebleeding rates in patients randomized to famotidine infusion versus placebo (24). Intravenous proton pump inhibitors have been shown to reduce rebleeding rates, length of hospital stay, and need for blood transfusion (25). Treatment with a proton pump inhibitor is generally started on admission for upper GI bleed, and continued as an adjunct to endoscopic therapy.

Mallory–Weiss Tear

The Mallory–Weiss syndrome describes acute upper GI bleeding that occurs after retching or vomiting, and was first described by Kenneth Mallory and Soma Weiss in 1929 (26). The increased intragastric pressure caused by vomiting causes mucosal lacerations, which are usually longitudinal. The typical presentation is a patient who initially vomits gastric material, followed by hematemesis and melena. Mallory–Weiss tears can also occur after anything that raises intragastric pressure, such as blunt abdominal trauma, severe coughing, childbirth, seizures, and closed chest cardiac compression. Mallory–Weiss tears classically occur at the gastroesophageal junction, but can occur in the distal esophagus. The lesion is common, occurring in 5–15% of patients presenting with upper GI bleeding. The majority of Mallory–Weiss tears will stop bleeding spontaneously, although some patients will require emergency treatment for ongoing hemorrhage. Endoscopic options for Mallory–Weiss bleeding include band ligation, epinephrine injection, and hemoclip application. In cases not amenable to endoscopic management, operative therapy involves oversewing the laceration via a gastrotomy.

Gastritis

Stress gastritis is associated with multiple superficial gastric ulcerations and is typically seen in the critically ill patient. Mechanical ventilation and coagulopathy increase the risk for hemorrhage in the critically ill. Prophylaxis with H2 blockers and proton pump inhibitors has led to a decrease in the incidence of stress gastritis in the critically ill. Bleeding from gastritis usually is self-limited, not requiring intervention. Early endoscopy is essential to establish the diagnosis and rule out other sources of upper GI bleeding. The patient should be started on pharmacologic therapy with either proton pump inhibitors or H2 blockers at a therapeutic dose. Endoscopic laser anticoagulation has been used for bleeding gastritis. Intraarterial infusion of vasopressin or selective embolization may also be used to arrest hemorrhage in gastritis. Ongoing hemorrhage not amenable to nonsurgical management is operatively managed with vagotomy, pyloroplasty, and oversewing of the bleeding sites versus total gastrectomy. The mortality for a patient requiring surgical management of bleeding gastritis remains quite high.

Esophagitis

Esophagitis is an unusual cause of acute gastrointestinal bleeding, and only rarely occurs in association with severe reflux esophagitis. The history would be suggestive of gastroesophageal reflux, with symptoms, such as, heartburn, cough, and hoarseness. Rare causes of esophagitis associated with bleeding in the immunocompromised patient include infection with candida, herpes, or cytomegalovirus (27).

Neoplasm

Acute upper GI bleeding is a rare manifestation of esophageal neoplasms, with $< 5\%$ of esophageal malignancies presenting with an acute bleed. Occult, slow GI bleeding is much more common with esophageal neoplasms. Benign tumors of the esophagus include leiomyomas and polyps, and are very unlikely to present with GI bleeding. Esophageal hemangiomas, which constitute only 2–3% of benign esophageal tumors, may present with potentially massive GI hemorrhage. Leiomyosarcomas are more likely to bleed than benign leiomyomas. When brisk bleeding occurs in the setting of esophageal cancer, one also must consider the erosion of the tumor into a major thoracic vessel.

Dieulafoy Vascular Malformation

Dieulafoy lesions are the result of arterioles of large diameter (1–3 mm) running through the submucosa, with erosion of the overlying mucosa resulting in bleeding. The mucosal defect is usually small, without evidence of chronic inflammation. Dieulafoy lesions generally present with brisk hemorrhage, due to their arterial nature. Diagnosis is made by upper endoscopy, with visualization of a small mucosal defect with brisk bleeding. Management is initially endoscopic with epinephrine injection and bipolar thermal therapy. Catheter embolization is generally successful in patients who fail endoscopic management. For patients requiring surgical management, the operation involves a wedge resection of the lesser curve of the stomach at the site of the lesion.

AORTOENTERIC FISTULA

Aortoenteric fistula classically occurs as a communication between a prosthetic aortic graft and the distal duodenum, and the diagnosis should be entertained in any patient presenting with an upper GI bleed who has undergone aortic reconstruction. The time period from aortic surgery to presentation is varied, and many patients present years down the road. The patient will frequently present initially with a sentinel bleed, which may be followed by massive upper GI hemorrhage. Upper endoscopy is paramount to making the diagnosis, as well as ruling out other sources of upper GI bleeding. Upon making the diagnosis of aortoenteric fistula, the optimal management is surgical, with removal of the aortic prosthesis, extra-anatomic bypass, and repair of the duodenum.

HEMOBILIA

Hemobilia classically presents as upper GI bleeding, melena, and biliary colic. Diagnosis is established by upper endoscopy, with visualization of blood from the ampulla. Endoscopic retrograde cholangiopancreatography can more clearly delineate the source of hemobilia. A variety of disease processes can lead to hemobilia, including hepatobiliary trauma, chronic cholelithiasis, pancreatic cancer, cholangiocarcinoma, and manipulation of the hepatobiliary tree. While hemobilia remains a rare cause of upper GI bleeding, the frequency is increasing related to increased manipulation of the hepatobiliary system and improved diagnostic modalities. Many cases of hemobilia will resolve without intervention. In the setting of ongoing hemorrhage, angiography with selective embolization of the bleeding vessel is the primary treatment modality. Surgery is reserved for failure of angiographic management.

LOWER GI BLEEDING

The passage of bright red or maroon blood via the rectum suggests a bleeding source distal to the ligament of Treitz, although blood can originate from any portion of the GI tract, depending on the rate of bleeding. Some 80–90% of lower GI bleeding will stop spontaneously. Initial resuscitation is similar to the patient presenting with upper GI bleeding, with hemodynamic assessment, establishment of appropriate access, and thorough history and physical exam. Visual inspection of the anorectal region, followed by anoscopy is essential to rule out a local anorectal condition such as hemorrhoids as the source of bleeding. A variety of modalities are available to further define the etiology of the bleeding, including endoscopy, nuclear medicine, angiography, and intraoperative localization.

The timing of colonoscopy for acute lower GI bleeding is controversial, with early (within 24 h of admission) colonoscopy increasingly advocated. Certainly, visualization may be difficult in a massively bleeding patient. Aggressive bowel prep can be given for 6–12 h prior to endoscopy, with the benefit of improved visualization. The benefit of early colonoscopy, similar to early upper endoscopy for upper GI bleed, is the opportunity for endoscopic diagnosis and therapy, using injection therapy and thermal modalities. Colonoscopy can directly visualize the bleeding source, which is beneficial in directing the surgeon in resection if the patient has continued or recurrent hemorrhage. Additionally, early colonoscopy may shorten hospital length of stay (28).

Nuclear scans can localize the site of lower GI bleeding and confirm active bleeding, with sensitivity to a rate of bleeding as low as 0.05–0.1 $mL \cdot min^{-1}$. Bleeding scans use either ^{99m}Tc sulfur colloid or ^{99m}Tc-labeled erythrocytes, with radioactivity detected by a gamma camera, analyzed by computer, and recorded onto photographic film. The ^{99m}Tc sulfur colloid has the advantage of detection of bleeding as slow as 0.05 $mL \cdot min^{-1}$, is inexpensive, and easy to prepare, but only will detect bleeding within 10 min of injection as it disappears quickly from the bloodstream (21). ^{99m}Tc-labeled erythrocytes detect bleeding as slow as 0.1 $mL \cdot min^{-1}$, and circulate within the bloodstream for 24 h. The ^{99m}Tc-labeled erythrocyte technique is generally considered the test of choice because of an increased sensitivity and specificity when compared with the ^{99m}Tc sulfur colloid (21). When a bleeding scan is positive, angiography or endoscopy is recommended to confirm the location of bleeding, to diagnose the specific cause, and to possible apply either endoscopic or angiographic therapy.

Angiography is advantageous because of the potential for both localization and treatment. Angiographic control can permit elective rather than emergent surgery in patients who are good surgical candidates, and can provide definitive treatment for poor surgical candidates. Bleeding can be detected at a rate as low as 0.5 $mL \cdot min^{-1}$. The SMA is cannulated initially, with cannulation of the IMA if the SMA study is nondiagnostic. When bleeding is localized, either vasopressin infusion or superselective catheter embolization may be used. Vasopressin is used in a method similar to upper GI bleeding, with an infusion rate of 0.2–0.4 units $\cdot min^{-1}$. Efficacy varies from 47–92%, with rebleeding in up to 40% of patients (21). Vasopressin is particularly effective for bleeding from diverticula.

Angiographic embolization may be done with a variety of agents, including coil springs, gelatin sponge, cellulose, and (PVA). There is less collateral blood supply in the lower G tract than in the upper, so embolization was initially thought to be a salvage therapy for those patients who would not tolerate an operation. Recent innovations in catheter and guidewire design, however, have enabled the interventional radiologist to superselectively embolize the bleeding vasa recta, sparing the collateral vessels and thus minimizing ischemia. Several small studies have reported successful embolization without intestinal infarction (21), with combined results showing successful hemostasis in 34 of 37 patients. Superselective embolization with coaxial microcatheters is currently considered the optimal angiographic therapy.

Traditionally, emergency operations for lower GI bleeding were required in 10–25% of patients presenting with bleeding (29). Surgical indications traditionally include hemodynamic instability, ongoing transfusion requirements, and persistent or recurrence of hemorrhage. If the bleeding has not been localized, a total abdominal colectomy with ileorectal anastomosis or end ileostomy is performed. If the lesion has been localized to either the right or left side of the colon, a hemicolectomy may be performed. With the advances in angiography available, the surgical indications are evolving. If an angiographer is readily available, angiographic localization and therapy is a viable option even for the hemodynamically unstable or persistently bleeding patient, thus avoiding the high morbidity and mortality associated with emergent total colectomy in this patient population.

Diverticulosis

The most common cause of lower GI bleeding is diverticular disease. Diverticular disease increases with age and is present in 50% of people $>$ 80 years. Less than 5% of these patients, however, will hemorrhage. While most diverticula are found distal to the splenic flexure, bleeding

diverticula more frequently occur proximal to the splenic flexure. Classically, the patient will present with sudden onset of mild lower abdominal pain and the passage of maroon or bright red bloody stool per rectum. The majority of diverticular bleeds will stop spontaneously, with a recent study showing spontaneous resolution in 76% of patients (1). About 20–30% of patients will have a recurrent bleeding episode, of which the majority will again stop without intervention. Patients that have persistent or recurrent bleeding should be considered for surgical therapy, particularly if the site of bleeding has been localized. High risk surgical patients can be treated with angiographic or endoscopic therapy.

Angiodysplasia

Angiodysplasia arises from age-related degeneration of submucosal veins and overlying mucosal capillaries, with frequency increasing with age. The bleeding tends to be less severe than with diverticular bleeds, and frequently resolves spontaneously, although recurrence is common. Diagnosis can be made by colonoscopy, with electrocoagulation as definitive therapy. Angiography may also be used for diagnosis, with the angiographic hallmarks a vascular tuft from an irregular vessel, an early and intensely filling vein resulting from arteriovenous communication, and persistent venous filling (30). Angiographic therapy with vasopressin can be used for treatment.

Neoplasm

While polyps and cancers frequently present with blood per rectum, they rarely cause massive hemorrhage as the presenting symptom. Diagnosis is made with colonoscopy. Management of a polyp is via colonoscopic polypectomy, while cancer requires surgical resection. Occasionally, a patient will present up to 1 month following a polypectomy with lower GI bleeding, which should prompt colonoscopy and thermal or injection therapy to the bleeding polypectomy site.

Meckel's Diverticulum

Meckel's diverticulum is an unusual cause of GI bleeding, and usually occurs in the first decade of life. The etiology of the bleeding is ectopic gastric mucosa in the diverticulum with resultant ulceration of adjacent bowel. Diagnosis is usually demonstrated by nuclear scanning demonstrating the ectopic gastric mucosa. Management is with surgical resection of the diverticulum as well as the adjacent bowel.

Ischemic Colitis

Ischemic colitis generally presents with bloody diarrhea, and massive lower GI bleeding is rare in this population. The bloody diarrhea is due to mucosal sloughing. Ischemic colitis should be suspected in patients with a history of vascular disease and in critically ill, hypotensive patients with a low flow state. Diagnosis is made by flexible endoscopy showing evidence of ischemia. Management for early ischemia is resuscitation and improvement of blood flow. Advanced ischemia requires surgical resection of the necrotic portion of the bowel.

Inflammatory Bowel Disease

GI bleeding characterizes both Crohn's disease and ulcerative colitis; however, massive bleeding is quite uncommon. The bleeding from inflammatory bowel disease is usually self-limited, and rarely acutely requires surgical attention. Diagnosis is made by colonoscopy, with identification of features unique to these entities and biopsy for pathology. Occasionally, ulcerative colitis will present fulminantly with massive hemorrhage and require surgical resection, consisting of total colectomy, end ileostomy, and Hartman's pouch, leaving the possibility for future conversion to an ileo-pouch anal anastomosis. Both entities are managed with immunosuppressive medications.

BIBLIOGRAPHY

Cited References

1. Hamoui N, Docherty DO, Crookes PH. Gastrointestinal hemorrhage: is the surgeon obsolete? Emerg Med Clin N Am 2003;21:1017–1056.
2. Bardhan KD, et al. Changing patterns of admission and operations for duodenal ulcer. Br J Surg 1989;76:230–236.
3. Longstreth GF. Epidemiology of hospitalization for acute upper gastrointestinal hemorrhage: a population based study. Am J Gastroenterol 1995;90:206–210.
4. Yavorski R, et al. Analysis of 3294 cases of upper gastrointestinal bleeding in military medical facilities. Am J Gastroenterol 1995;90:568–573.
5. Gilbert DA, et al. The national ASGE survey on upper gastrointestinal bleeding III. Endoscopy in upper gastrointestinal bleeding. Gastrointest Endosc 1982;27:94
6. Comar KM, Sanyal AJ. Portal hypertensive bleeding. Gastroenterol Clin N Am 2003;32:1079–1105.
7. Lebrec D, et al. Portal hypertension, size of esophageal varices, and risk of gastrointestinal bleeding in alcoholic cirrhosis. Gastroenterology 1980;79:1139–1144.
8. Pagliaro L, et al. Prevention of the first bleed in cirrhosis. A metaanalysis of randomized trials of non-surgical treatment. Ann Intern Med 1992;117:59–70.
9. Stiegmann GV, Goff GS, Sun JH, Wilborn S. Endoscopic elastic band ligation for active variceal hemorrhage. Am Surg 1989;55:124–128.
10. Conn HO. Vasopressin and nitroglycerine in the treatment of bleeding varices: the bottom line. Hepatology 1986;6:523–525.
11. Inperiale TF, Teran JC, McCullough AJ. A meta-analysis of somatostatin versus vasopressin in the management of acute esophageal variceal hemorrhage. Gastroenterology 1995;109:1289–1294.
12. Feneyrou B, Hanana J, Daures JP, Prioton JB. Initial control of bleeding from esophageal varices with the Sengstaken-Blakemore tube: experience in 82 patients. Am J Surg 1988;155:509–511.
13. LaBerge JM, et al. Creation of transjugular intrahepatic portosystemic shunts with the wallstent endoprothesis: results in 100 patients. Radiology 1993;187:413–420.
14. Sanyal AJ, et al. Transjugular intrahepatic portosystemic shunts compared with endoscopic sclerotherapy for the prevention of recurrent variceal hemorrhage: a randomized, controlled trial. Ann Intern Med 1997;126:849–857.
15. LaBerge JM, et al. Two-year outcome following transjugular intrahepatic portosystemic shunt for variceal bleeding: results in 90 patients. Gastroenterology 1995;108:1143–1151.

16. Grace ND. Diagnosis and treatment of gastrointestinal bleeding secondary to portal hypertension. American College of Gastroenterology Practice Parameters Committee. Am J Gastroenterol 1997;92:1081–1091.
17. Bardhan KD, et al. Changing patterns of admissions and operations for duodenal ulcer. Br J Surg 1989;76:230–236.
18. Huang CS, Lichtenstein DR. Nonvariceal upper gastrointestinal bleeding. Gastroenterol Clin N Am 2003;32: 1053–1078.
19. Cook DJ, Guyatt GH, Salena BJ, Laine LA. Endoscopic therapy for acute nonvariceal upper gastrointestinal hemorrhage:a meta-analysis. Gastroenterology 1992;102:139–148.
20. Lau JY, et al. Endoscopic retreatment compared with surgery in patients with recurrent bleeding after initial endoscopic control of bleeding ulcers. N Engl J Med 1999;340: 751–756.
21. Gomes AS, Lois JF, McCoy RD. Angiographic treatment of gastrointestinal hemorrhage: comparison of vasopressin infusion and embolization. Am J Roentgenol 1986;146:1031–1037.
22. Lefkovitz Z, et al. Radiologic diagnosis and treatment of gastrointestinal hemorrhage and ischemia. Med Clin N Am 2002;86:1357–1399.
23. Collins R, Langman M. Treatment with histamine H2 antagonists I acute upper gastrointestinal hemorrhage: implications of randomized trials. N Engl J Med 1985;313: 660–666.
24. Walt RP, et al. Continuous infusion of famotidine for hemorrhage from peptic ulcer. Lancet 1992;340:1058–1062.
25. Javid G, et al. Omeprazole as adjuvant therapy to endoscopic combination injection sclerotherapy for treating bleeding peptic ulcer. Am J Med 2001;111:280–284.
26. Mallory GK, Weiss S. Hemorrhage from lacerations of the cardiac orifice of the stomach due to vomiting. Am J Med Sci 1929;178:506.
27. Onge GS, Bezahler GH. Giant esophageal ulcer associated with cytomegalovirus. Gastroenterology 1982;83:127–130.
28. Schmulewitz N, Fisher DA, Rockey DC. Early colonoscopy for acute lower GI bleeding predicts shorter hospital stay: A retrospective study of experience in a single center. Gastrointest Endosc 2003;58:841–846.
29. Colacchio TA, Forde KA, Patsos TJ, Nunez D. Impact of modern diagnostic methods on the management of rectal bleeding. Am J Surg 1982;143:607–610.
30. Boley SJ, et al. The pathophysiologic basis for the angiographic signs of vascular ectasias of the colon. Radiology 1977;125:615–621.

See also ELECTROGASTROGRAM; ENDOSCOPES.

GEL FILTRATION CHROMATOGRAPHY.

See CHROMATOGRAPHY.

GLUCOSE SENSORS

FARBOD N. RAHAGHI
DAVID A. GOUGH
University of California,
La Jolla, California

INTRODUCTION

Glucose assay is arguably the most common of all medical measurements. Billions of glucose determinations are performed each year by laypeople with diabetes based on "fingersticking" and by healthcare professionals based on blood samples. In fingersticking, sample collection involves the use of a lancet to puncture the skin of the fingertip or forearm to produce a small volume of blood and tissue fluid, followed by collection of the fluid on a reagent-containing strip and analysis by a handheld meter. Glucose measurements coupled to discrete sample collection continue to be the most common method of glucose monitoring. However, new types of sensors capable of continuous glucose monitoring are nearing clinical introduction. Continuous or near-continuous glucose sensors may make possible new and fundamentally different approaches to the therapy of the disease. This article reviews recent progress in the development of new glucose sensors and describes the potential roles for these sensors in the improved treatment of diabetes.

THE CASE FOR NEW GLUCOSE SENSORS

The objective of all forms of therapy for diabetes is the maintenance of blood glucose near normal levels (1). The Diabetes Control and Complications Trial (or DCCT) and counterpart studies such as the United Kingdom Prevention of Diabetes Study (UKPDS) have clearly demonstrated (Fig. 1) that lower mean blood glucose levels resulting from aggressive treatment can lead to a reduced incidence and progression of retinopathy, nephropathy, and other complications of the disease (2,3). These prospective studies showed definitively that there exists a cause-and-effect relationship between poor blood glucose control and the complications of diabetes. As convenient means for frequent glucose assay were not available at the time, glucose control was assessed in these trials by glycosylated hemoglobin levels (Hb_{A1c}), which indicate blood glucose concentrations averaged over the previous 3 month period. Although Hb_{A1c} levels are useful for assessment of longitudinal blood glucose control, the values indicate only *averaged* blood glucose, rather than blood glucose *dynamics* (i.e., how blood

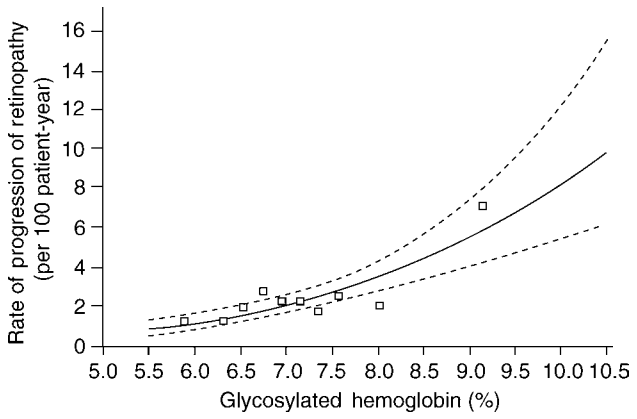


Figure 1. The results of the DCCT (2). Results show that improved glucose control, measured by a reduction in the fraction of glycosylated hemoglobin, leads to reduced long-term complications of diabetes. (Copyright © 1993, Massachusetts Medical Society.)

glucose changes with time), and cannot be used for immediate adjustment of therapy (4). There is general agreement that frequent determination of glucose by a sensing method that is convenient and widely acceptable to people with diabetes would allow a finer degree of control. Normalization of blood glucose dynamics may be of equal or greater importance than normalization of average blood glucose. The results of the DCCT and related studies point to the need for practical new approaches to achieve control.

The primary need for a new type of glucose sensor is to facilitate improved treatment of type 1 diabetes. In this case, the insulin producing ability of the pancreas has been partially or fully destroyed due to a misdirected autoimmune process, making insulin replacement essential. The sensor would help avoid the long-term complications associated with hyperglycemia (i.e., above-normal blood glucose) by providing information to specify more timely and appropriate insulin administration. It is now becoming widely appreciated that a new sensor could also be beneficial for people with the more common type 2 diabetes, where a progressive resistance of peripheral tissues to insulin develops, leading to glucose imbalances that can eventually produce long-term clinical consequences similar to type 1 diabetes. Type 2 diabetes is related to obesity, lifestyle, and inherited traits. In recent years, the incidence of type 2 diabetes has increased at extraordinary rates in many populations, to the point of becoming a worldwide epidemic (5). It is estimated that within 10 years, the prevalence of diabetes may approach 210 million cases worldwide (6). This places increased urgency on developing new approaches to managing or preventing the disease where possible, and a meliorating its consequences.

In addition, an automatic or continuous sensor may also have an important role in preventing hypoglycemia (i.e., below-normal blood glucose). Hypoglycemia is caused primarily by a mismatch between the insulin dosage used and the amount of insulin actually needed to return the blood glucose level to normal. Many people with diabetes can reduce the mean blood glucose by adjustment of diet, insulin, and exercise, but when aggressively attempted, this has led to a documented increase in the incidence of hypoglycemia (7). Below-normal glucose values can rapidly lead to cognitive lapses, loss of consiousness, and life-threatening metabolic crises. In children, there is concern that severe hypoglycemic events may lead to neurologic sequelea (8). A significant percentage of deaths of people under 40 with type 1 diabetes is due to the "dead-in-bed" syndrome (9), which may be linked to nocturnal hypoglycemia. Some experts claim that "... the threat of severe hypoglycemia remains the single most important barrier to maintaining normal mean blood glucose" (10). A continuous glucose sensor that does not depend on user initiative could be part of an automatic alarm system to warn of hypoglycemia and provide more confidence to the user to lower mean blood glucose, in addition to preventing hypoglycemia by providing improved insulin dosages. Hypoglycemia detection may be the most important application of a continuous glucose sensor. Ultimately, a glucose sensor may also be useful in the prediabetic state to indicate behavior modification for reduction of metabolic stress on the pancreas.

Beyond applications in diabetes, it has recently been shown that stricter glycemic control during surgery and intensive care can reduce mortality in non-diabetic patients and significantly shorten the hospital stay (11). The exact mechanism of this effect has not been elucidated, but the benefit is closely tied to the extent of glucose control and not simply insulin dosage (12). This is another important application for new glucose sensors.

Alternatives to sensor-based therapies for diabetes are more distant. Several biological approaches to diabetes treatment have been proposed, including pancreatic transplantation, islet transplantation, genetic therapies, stem cell-based therapies, new pharmaceutical strategies, islet preservation, and others. Whole or partial organ and islet transplantation requires discovery of methods for assuring immuno-tolerance that do not rely on anti-rejection drugs and approaches for overcoming the shortage of transplantable pancreatic tissue. Potential therapies based on stem cells, if feasible, require basic research on growth, regulation, and implementation of the cells, and share the immuno-intolerance problem. Genetic therapies are limited by incomplete understanding of the complex genetic basis of diabetes, as well as progress in developing site-specific gene delivery, activation, and inactivation. Although transplantation, stem cell, and genetic approaches are based wholly on biological materials, it is not certain that the glucose and insulin dynamics resulting from their use will necessarily be near-normal or readily adjustable. Immunotherapeutic approaches for *in situ* preservation of islets are also being studied but, if eventually feasible, are far off and may require lifetime immune system modulation. The possibility of prevention of type 1 diabetes relies on development of timely methods for early detection of the disease and discovery of an acceptable approach to avoid or interrupt the islet destruction process. Furthermore, prevention will have little value for people who already have diabetes. These alternatives require substantial basic research and discovery, and while often highly publicized, are not likely to be available until far into the future, if eventually feasible.

Although new glucose sensors have the advantage of being closer to clinical introduction, there are certain other advantages as well. First, no anti-rejection medication will be needed. Second, the sensor will provide real-time information about blood glucose dynamics that is not available from other technologies. Third, in addition to real-time monitoring, continuous sensor information may be useful to *predict* blood glucose ahead of the present (13), a capability not feasible with the other approaches. Real-time monitoring and predictive capabilities may lead to entirely new applications of present therapies. Fourth, the sensor could operate in parallel with various other therapies, should they become available. The glucose sensor will likely have broad application, regardless of whether or when other technologies are introduced.

The sensor is also key to the implementation of the mechanical artificial beta cell. In the ideal configuration, this device would have an automatic glucose sensor, a refillable insulin pump, and a controller containing an

algorithm to direct automatic pumping of insulin based on information provided by the sensor. There has been progress on development of several of the components of this system, including: (1) external insulin pumps, which operate in a substantially preprogrammed mode with minor adjustments by the user based on fingerstick glucose information; (2) long-term implantable insulin pumps that operate in a similar way; (3) models of glucose and insulin distribution in the body that may eventually be useful in conjunction with control systems; and (4) controllers to direct insulin pumping based on sensor information. In contrast to other approaches to insulin delivery, the mechanical artificial beta cell has the advantage that the insulin response can be reprogrammed to meet the changing needs of the user. Development of an acceptable glucose sensor has thus far been the most difficult obstacle to implementation of the mechanical artificial beta cell.

THE IDEAL GLUCOSE SENSOR

The likelihood that glucose monitoring will reach its full potential as a tool for the therapy of diabetes depends on the technical capabilities of candidate sensors and the general acceptance of sensors by people with diabetes. Technical requirements of the sensor system include: specificity for glucose in the presence of interfering biochemicals or physiological phenomena that may affect the signal; sensitivity to glucose and adequate concentration resolution over the relevant range; accuracy as compared to a "gold standard" blood glucose assay; a sufficiently short response lag to follow the full dynamic range of blood glucose variations; reliability to detect mild hypoglycemia without false positives or negatives; and sufficient stability that recalibration is rarely needed. The specific criteria for sensor performance remain a matter of consensus and may become better defined as sensors are introduced. The general acceptance of new sensors by people with diabetes will be based on such factors as safety, convenience, reliability, automatic or initiative-independent operation, infrequent need for recalibration, and independence from fingersticking.

For the glucose sensor to be a widely accepted innovation, the user must have full confidence in its accuracy and reliability, yet remain uninvolved in its operation and maintenance. Sensor systems under development have yet to reach this ideal, but some promising aspirants are described below. Short of the ideal, several intermediate sensing technologies with limited capabilities may find some degree of clinical application and, if used effectively, may lead to substantial improvements in blood glucose control. Nevertheless, the most complete capabilities will lead to the broadest adoption by users.

GLUCOSE SENSORS AND SENSING METHODOLOGIES

Several hundred physical principles for monitoring glucose have been proposed since the 1960s. Many are capable of glucose measurement in simple solutions, but have encountered limitations when used with blood, employed as implants, or tested in clinically relevant applications. Certain others have progressed toward clinical application. A brief summary of the history of events related to glucose sensor development is shown in Figure 2.

Present Home Glucose Monitoring

A major innovation leading to improved blood glucose management was the widespread use of home glucose

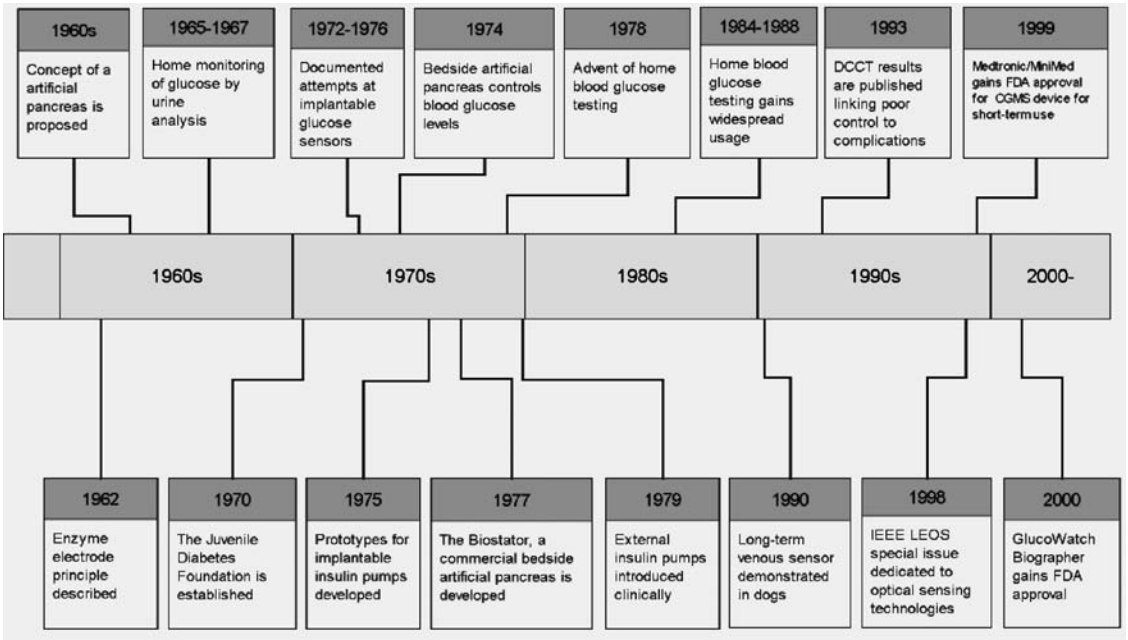


Figure 2. A time-line of some important developments relating to glucose sensors (2,14–24).

Figure 3. A small collection of home glucose monitoring equipment developed over the past decade. At either end (above) are devices used to puncture the skin for sample collection. Examples of commercial glucose meter (above, center) are also shown. Strips (below) contain immobilized glucose oxidase and are discarded after a single measurement.

monitoring in the 1980s (22). Present commercial versions of this technology are available with respective methods for glucose assay, data presentation and storage, sample volume requirements, and various convenience features (Fig. 3). These devices employ single-use strips based on enzyme methods discussed below. The widespread application of home glucose monitoring has permitted laypeople with diabetes to assume a newfound role in the management of their disease. The present standard-of-care recommends glucose measurement three or more times a day for insulin-dependent individuals (25), but a small number of individuals samples 10 or more times daily. It is generally suspected that the average sampling rate is inadequate and a recent publication noted that only 56% of diabetic individuals sampled their blood glucose once or more daily (26). The general resistance to more frequent sampling may be related to several factors, including: the pain associated with finger puncture, the requirement for user initiative, the general inconvenience of the assay, and unwillingness to carry out nocturnal testing (27).

When sampling is not sufficiently frequent, undetected blood glucose excursions can occur between samples. It has been shown that blood glucose measurements must be obtained every 10 min to detect all blood glucose excursions in the most severe diabetic subjects (28), although slower blood glucose excursions in the majority of people with diabetes may not require sampling at this frequency. The fact that the sample frequency required to detect all glycemic excursions is not clinically feasible with present technology indicates that the dynamic control of blood glucose is currently not practiced in diabetes management.

To compensate for infrequent monitoring, users typically adopt various strategies to estimate blood glucose concentration using subjective *ad hoc* models. These strategies rely on the most recent reported values of glucose, in conjunction with the timing and content of recent or upcoming meals, insulin therapy, and exercise. The effectiveness of these strategies is limited and the constant attention required to make such estimates represent a substantial intrusion in lifestyle. Although glucose monitoring by fingersticking is likely to become more acceptable as the sample volume and the pain associated with sample collection are reduced, the problem of infrequent sampling and the requirement for user initiative will continue to be the major obstacles to the improvement of glucose control based on this technology.

Noninvasive Optical Sensing Concepts

Noninvasive optical methods are based on directing a beam of light onto the skin or through superficial tissues, and recording the reflected, transmitted, polarized, or absorbed components of the light (29). A key requirement for success of these methods is a specific spectral region that is sufficiently sensitive to glucose, but insensitive to other similar optically active interfering molecules and tissue structures. Several optical methods allow straightforward glucose measurement in simple aqueous solutions, but are ineffective at detecting glucose in tissue fluid, plasma, or blood. If an optical approach can be validated, a noninvasive sensor might be possible. For this reason, an intensive research effort and substantial industrial investment over the past two decades have gone into investigation of these concepts.

Infrared (IR) absorption spectroscopy is based on excitation of molecular motions that are characteristic of the molecular structure. The near-infrared (NIR) region of the spectrum (750–2500 nm) is relatively insensitive to water content so that the beam penetration depth in tissues can be substantial (30). Trials to identify a clinical correlation between NIR signals and blood glucose have employed various computational methods for analyzing the absorption spectrum. Basic studies have focused on identifying the absorbing species and tissue structures responsible for optical signals. However, after much effort the operating conditions that provide selectivity for glucose have yet to be established, leading one investigator to conclude that "... signals can be attributed to chance" (31).

Raman spectroscopy relies on detecting scattered emissions associated with vibrational molecular energy of the chemical species (as opposed to transmitted, rotational, or translational energy). Early studies compared the measurement in water of three different analytes (urea, glucose, lactic acid) and found that glucose levels could be determined with limited accuracy (32). Raman spectroscopy has been applied in the aqueous humor of the eye (33), which is thought to reflect delayed blood glucose levels over certain ranges (34). As with other optical methods, adequate specificity for glucose in the presence of other molecules remains to be demonstrated.

Measurement of the concentration using *polarimetry* is based on ability of asymmetric molecules such as glucose to rotate the plane of polarized light (35). This method is limited by the presence of other interfering asymmetric molecules, as well as the thickness and light scattering by tissues in the region of interest (30). Considerable development of polarimetry has centered on measurements in the anterior chamber of the eye (36), but there is yet to be a demonstration of sufficient selectivity under biological conditions.

Attempts at validation of optical sensor concepts have involved two general approaches. One approach endeavors to establish selectivity for glucose by identification of the components of tissues besides glucose that contribute to the optical signal, and determine if the effects of these interfering substances can be eliminated or the observed signals can be reconstructed based on all contributions. This has not yet been successful, in spite of intensive efforts. The impediment is the large number of optically active components in tissues, many of which produce much stronger effects than glucose. A second approach to validation involves identifying an empirical relationship between the observed optical signal *in vivo* and simultaneously recorded blood glucose concentration. Noninvasive optical approaches have been the premise of several human clinical trials, all of which have been unsuccessful. The prospects for a non-invasive optical glucose sensor are distant.

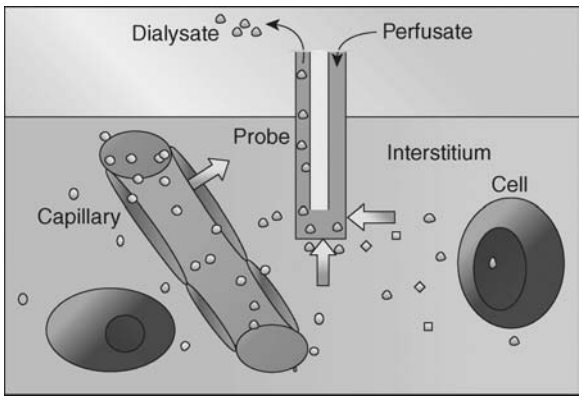


Figure 4. Diagram of a microdialysis probe (39). The semipermeable membrane at the probe tip allows exchange of soluble molecules between the probe and surrounding tissue. Samples are continuously collected and analyzed. (Used with permission of BMJ Publishing Group.)

Implantable Optical Sensor Concepts

Implanted optical sensors offer the prospect of a less congested optical path, at the expense of requiring a more complicated device and confronting the foreign body response. One promising optical concept is based on chemical interaction between glucose and an optically active chemical species that is immobilized in an implanted, glucose-permeable chamber. The interaction creates a change in the optical signal which, under ideal conditions, may indicate glucose concentration. An example is the "affinity sensor" (37), in which glucose competes with a fluorescent substrate for binding with a macromolecule, Con A, resulting in a change in the optical signature. A similar detection strategy has been proposed as part of an implantable intraocular lens (38). There are difficulties with biocompatibility of the implant, design of the chamber, specificity of the optical detector, as well as toxicity and photobleaching of the indicator molecules (38). These systems have yet to be extensively tested.

Tissue Fluid Extraction Techniques

The interstitial fluid that irrigates tissues contains glucose derived from the blood in local capillaries. Several strategies have been devised to extract this fluid for glucose assay.

Microdialysis is based on a probe consisting of a fine hairpin loop of glucose-permeable dialysis tubing in a probe that is inserted into subcutaneous tissues (39). A fluid perfusate is continuously circulated through this tubing by a pump contained in an external apparatus (40), collected, and assayed for glucose concentration using an enzyme electrode sensor (Fig. 4). This methodology relies on the exchange of glucose between the microvascular circulation and the local interstitial compartment, transfer into the dialysis tube, and appropriate adjustment of the pumping pressure and perfusion rate (41). The advantage of the system is that a foreign body response of the tissue and local mass transfer resistance are slow to develop due to the sustained tissue irrigation, but drawbacks include the requirement for percutaneous access, the need for frequent relocation of the probe to minimize the chance of infection, and management of the external apparatus by the user. This device may find clinical applications for short-term monitoring.

Reverse iontophoresis employs passage of electrica current between two electrodes placed on the surface of the body to extract tissue fluid directly through the intact skin (42). Glucose in the fluid has been measured by an enzyme electrode-type sensor as part of a wristwatch-like apparatus (43) (Fig. 5). With a 2 h equilibration process after placing the device and a fingerstick calibration, the sensor can take measurements as often as every 10 min for 12 h, at which time sensor components must be replaced and the sensor recalibrated (44). This sensor was approved by the Food and Drug Administration (FDA) for indicating glucose trends, but users are instructed to revert to more reliable conventional assays for insulin dosing decisions. Minor skin irritation has been reported as a side effect (45). Although this sensor was briefly available commercially, it

Figure 5. The Glucowatch Biographer (43). An integrated system for sample extraction by reverse iontophoresis and glucose sensing. (Used with permission from John Wiley & Sons, Inc.)

was not successful as a product due to its limited capabilities and inconvenience.

Implantable Enzyme Electrode Sensors

The most promising approaches have been various configurations of the enzyme electrode sensor based on immobilized glucose oxidase coupled to electrochemical detectors. The enzyme catalyzes the reaction:

$$\text{glucose} + O_2 + H_2O \rightarrow \text{gluconic acid} + H_2O_2 \qquad (1)$$

Monitoring of glucose can be based on detection of hydrogen peroxide production, oxygen depletion, or electron transfer via a conductive polymer link, as described below. Enzyme electrode sensors must contact the sample fluid to be assayed, and therefore require either sensor implantation or sample extraction (as in the case of reverse iontophoresis, microdialysis sensors and fingerstick devices). By employing the enzyme, sensors can have a significant advantage over non-enzymatic sensors of being *specific* for glucose rather than just selective. However, the benefits of enzyme specificity may not be fully realized unless the sensor is properly designed. To achieve the best performance, enzyme electrode sensors must include design features to address enzyme inactivation, biological oxygen variability, mass transfer dependance, generation of peroxide, electrochemical interference, and other effects.

From the perspective of biocompatibility, sensors can be implanted either in direct contact with blood or with tissues. Biocompatibility in contact with blood depends on the surface properties of the sensor as well flow characteristics at the implant site. Implantation in an arterial site, where the pressure and fluid shear rates are high, poses the threat of blood clotting and embolization, and is rarely justified. Central venous implantation is considerably safer, and there are several examples of successfull long-term implants in this site.

Implantation of the sensor in a tissue site is safer, but involves other challenges. The sensing objective is to infer blood glucose concentration from the tissue sensor signal, and factors that affect glucose mass transfer from nearby capillaries to the implanted sensor must be taken into account. These factors include: the pattern and extent of perfusion of the local microvasculature; regional perfusion of the implant site, the heterogeneous distribution of substrates within tissues, and the availability of oxygen. There are also substantial differences in performance between short- and long-term implant applications. In the short term, a dominant wound healing response prevails, whereas in the long term, encapsulation may occur. Definitive studies are needed to establish the real-time accuracy of implanted sensors and determine when recalibration is necessary. Studies should be designed to ascertain whether signal decay is due to enzyme inactivation, electrochemical interference, or tissue encapsulation. More information is needed about the effect of these processes on the sensor signals.

There are $>10^4$ technical publications and several thousand patents related to glucose measurement by glucose oxidase-based enzyme electrodes, although only a fraction of these address implant applications. Rather than an attempt to be comprehensive, comments here are limited to examples of the most advanced approaches intended for implant applications.

Enzyme Electrode Sensors Based on Peroxide Detection.

Detection of hydrogen peroxide, the enzyme reaction product, is achieved by electrochemical oxidation of peroxide at a metal anode resulting in a signal current that passes between the anode and a counterelectrode (46). A membrane containing immobilized glucose oxidase is attached to the anode and, in the presence of glucose and oxygen under certain conditions, the current can reflect glucose concentration.

The peroxide-based sensor design is used in several home glucose monitoring devices and has been highly successful for glucose assay on an individual sample basis. However, it is not easily adapted as an implant, especially for long-term applications. The peroxide-based sensor is subject to electrochemical interference by oxidation of small molecules due to its requirement of a porous membrane and an aqueous pathway to the electrode surface for transport of the peroxide molecule. This factor partially accounts for a documented decay in sensitivity to glucose during sensor use. In addition, this sensor design can incorporate only a limited excess of immobilized glucose oxidase to counter enzyme inactivation, as high enzyme loading reduces peroxide transport to the electrode (47). Coimmobilization of catalase to avoid peroxide-mediated enzyme inactivation is not an option because it would prevent peroxide from reacting with the anode. There are also no means to account for the effects of physiologic variation in oxygen concentration and local tissue perfusion on the sensor response.

There have, nevertheless, been proposals to address some of these challenges. Composite membranes with reduced pore size have markedly reduced electrochemical interference from a variety of species over the short-term (48). A "rechargeable" enzyme system has been devised for periodically replenishing enzyme activity (49), in which a slurry of carbon particles with immobilized glucose oxidase is pumped between membrane layers of a peroxide electrode from a refillable reservoir. A gas-containing chamber has been proposed (50) to address the "oxygen deficit" (51), or stoichiometric limitation of the enzyme reaction by the relativley low tissue oxygen concnetration. Certain other challenges of the peroxide sensor principle remain to be addressed. As a result of the inherent features of this sensor principle, the peroxide-based sensor may be best suited to short-term implant applications and where frequent sensor recalibration is acceptable.

Small, needle-like *short-term peroxide-based sensors* connected by wire to a belt-mounted monitor have been developed for percutaneous implantation (52) (Fig. 6). The sensor was ultimately intended for insertion by the user for operation up to 3 days at a given tissue site before relocation. Sensors based on peroxide detection have been tested extensively in animals and humans (52–55) and, in some cases have functioned remarkably well, although frequent recalibration was required. In

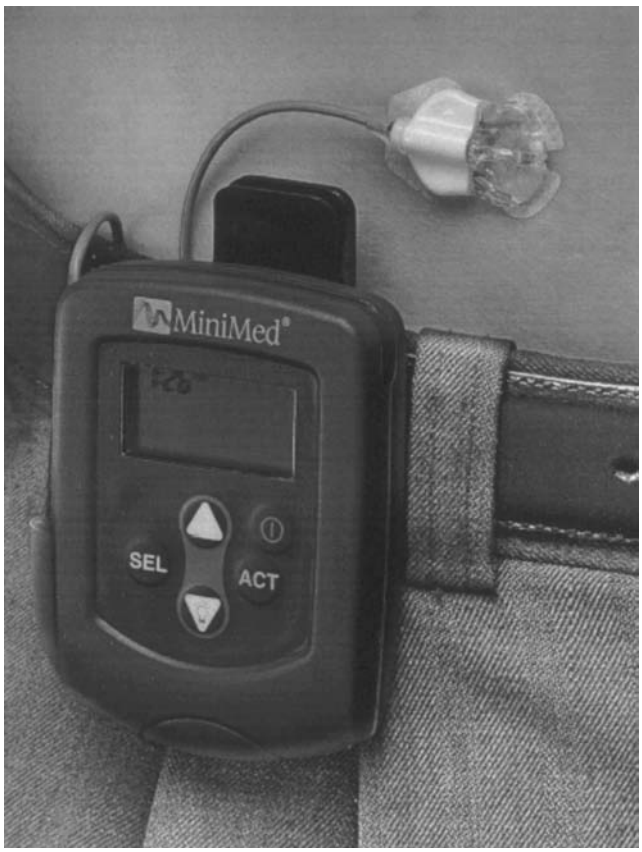


Figure 6. Components of the MiniMed CGMS system (62). This sensor, based on the peroxide-based sensing principle, is FDA approved for short-term monitoring. (Copyright © 1999, Elsevier.)

some cases, difficulties of two types have been identified (56–61). First, the sensors themselves have not been specific for glucose, sufficiently sensitive, or stable. In some cases where the sensor response in simple buffer solutions was acceptable, ambiguous signals have sometimes resulted when used as an implant. Examples of such responses are: sensor signals that decay over the short term while blood concentrations remain constant, signals that apparently follow blood concentrations during some periods, but not at other times; and identical sensors implanted in different tissue sites in a given subject that sometimes produce opposite signals (55). The peroxide-based subcutaneous sensor was the first commercially available near continuous sensor. In small controlled studies, use of the sensor was shown to lower Hb_{A1c} levels (62). Latter versions of this sensor have been approved by the FDA to be used as monitors to alarm for hyper- and hypoglycemia in real time. Although there are reservations about its accuracy, the needle sensor has been used in clinical research settings (Figs. 7, 8). A recent study found substantial error in values produced by a prominent commercial needle sensor and concluded that this sensor "... cannot be recommended in the workup of hypoglycemia in nondiabetic youth" (66) and, by extension, to other diabetic subjects. Reasons for these signal deviations are not fully understood.

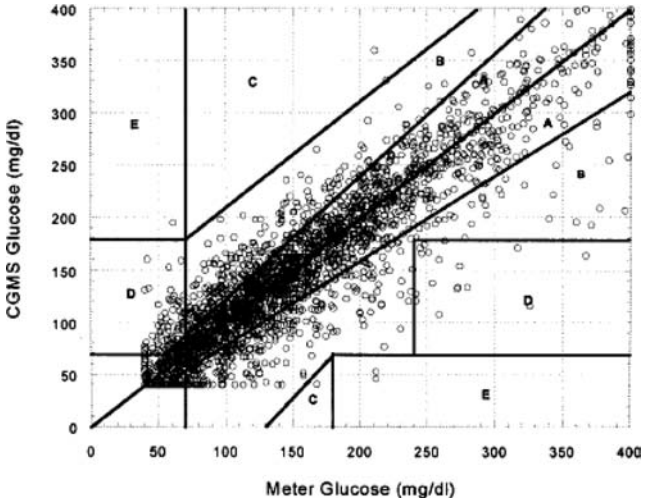


Figure 7. An approach to sensor validation. Comparison of 2477 glucose values determined by a CGMS sensor system and a standard meter (63). Data pairs were collected during home use from 135 patients. The plot, known as the Clarke Error Grid, has zones with differing clinical implications. This type of plot is widely used to describe glucose sensor performance, but has limited ability to discriminate ineffective sensors (64). (Copyright © 1999, Elsevier.)

Although the needle sensor may be acceptable only to a relatively small group of the most motivated individuals, it represents an advance in glucose sensor technology. Perhaps the most important contribution of the short-term needle sensor has been the revelation to users and clinicians that blood glucose excursions generally occur much more frequently and in a greater number of people than previously thought. This heightened awareness of blood glucose dynamics may lead to a greater appreciation of the need for dynamic control in improved metabolic management.

Long-term peroxide-based sensors have been implanted in the peritoneal cavity of dogs and in humans in conjunction with battery-operated telemetry units (67,68). Although the sensors remained implanted in humans for up to 160 days, the sensitivity to glucose decayed during the study and frequent recalibration was required.

Peroxide-based sensors with telemetry systems have also been implanted in the subcutaneous tissues of human type 1 diabetic subjects to determine if the mere presence of a nonreporting sensor can improve metabolic control (69). Glucose was monitored in parallel by fingerstick throughout the study as a basis for insulin dosage. Study subjects were able to reduce the time spent in hyperglycemia, increase the time spent in normoglycemia and modest hypoglycemia, and markedly reduce the time spent in severe hypoglycemia, but reductions in Hb_{A1c} values were not observed. The study was not specifically designed to validate sensor function and a more straightforward and informative study design is needed.

Short-Term Enzyme Electrodes Based on Conductive Polymers. Another principle for glucose monitoring is based on

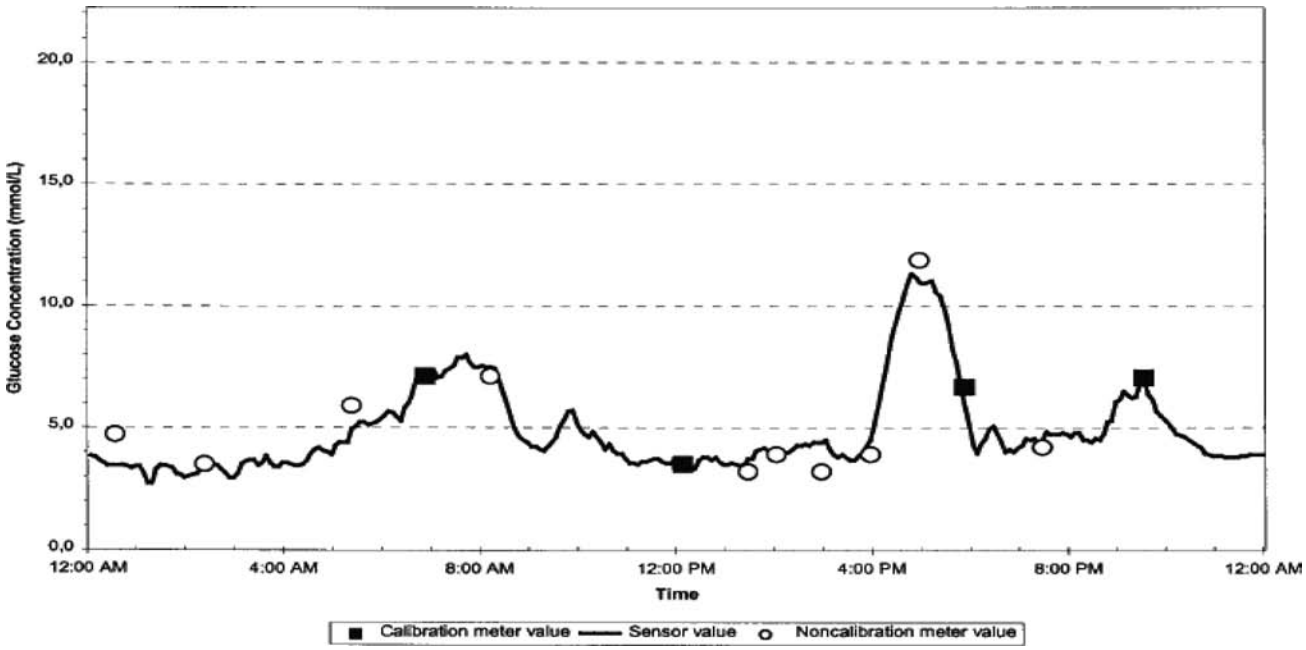


Figure 8. An example of the CGMS sensor response (65). Squares are reference values utilized in sensor calibration. Circles are additional reference blood glucose values. Reference values were obtained from a standard glucose meter. Values are in mmol $\cdot$ L^{-1}. (Used with permission.)

immobilization of glucose oxidase to electron-conducting polymers that can act as "chemical wires", providing a means for direct electron transport between glucose oxidase and the electrode (70). This priniciple eliminates the need for oxygen as a coreactant and, although a porous membrane that can allow passage of ionic current and interferants is still required, the electrode can be operated at lower anodic potentials to reduce electrochemical interference (71). A short-term needle-like version of this sensor for 3 day operation is under development.

Long-Term Enzyme Electrode Sensors Based on Oxygen Detection. Glucose can also be monitored by detecting differential oxygen consumption from the glucose oxidase reaction. In this case, the process is based either on glucose oxidase alone (reaction 1), or a two-enzyme reaction including catalase in excess, which produces the following overall process:

$$\text{Glucose} + 0.5\ O_2 \rightarrow \text{gluconic acid} \qquad (2)$$

The enzymes are immobilized within a gel membrane in contact with the electrochemical oxygen sensor. Excess oxygen not consumed by the enzymatic process is detected by an oxygen sensor and, after comparison with a similar background oxygen sensor without enzymes, produces a differential signal current that is related to glucose concentration.

This approach has several unique features (23). Electrochemical interference and electrode poisoning from endogenous biochemicals are prevented by a pore-free silicone rubber membrane between the electrode and the enzyme layer. This material is permeable to oxygen but completely impermeable to polar molecules that cause electrochemical interference. Appropriate design of the sensor results in sufficient supply of oxygen to the enzyme region to avoid a stoichiometric oxygen deficit (51), a problem that has not been addressed in the peroxide-based sensor system. The differential oxygen measurement system can also readily account for variations in oxygen concentration and local perfusion, which may be particularly important for accurate function of the implant in tissues. Vast excesses of immobilized glucose oxidase can be incorporated to extend the effective enzyme lifetime of this sensor, a feature not feasible with peroxide- and conductive polymer-based sensors. Co-immobilization of catalase can further prolong the lifetime of glucose oxidase by preventing peroxide-mediated enzyme inactivation, the main cause of reduced enzyme lifetime (72). This sensor design also avoids current passage through the body and hydrogen peroxide release into the tissues.

A long-term oxygen-based sensor has been developed as a central venous implant (23) (Fig. 9). The sensor functioned with implanted telemetry (73) in dogs for >100 days and did not require recalibration during this period (Fig. 10). The central venous site permitted direct exposure of the sensor to blood, which allowed simple verification of the sensor function without mass transfer complications. This was particularly beneficial for assessing sensor stability. In clinical trials, this system has been reported (74) to function continuously for >500 days in humans with <25% change in sensitivity to glucose over that period. This achievement represents a world record for long-term, stable, implanted glucose sensor operation, although there may still exist hurdles to commercialization.

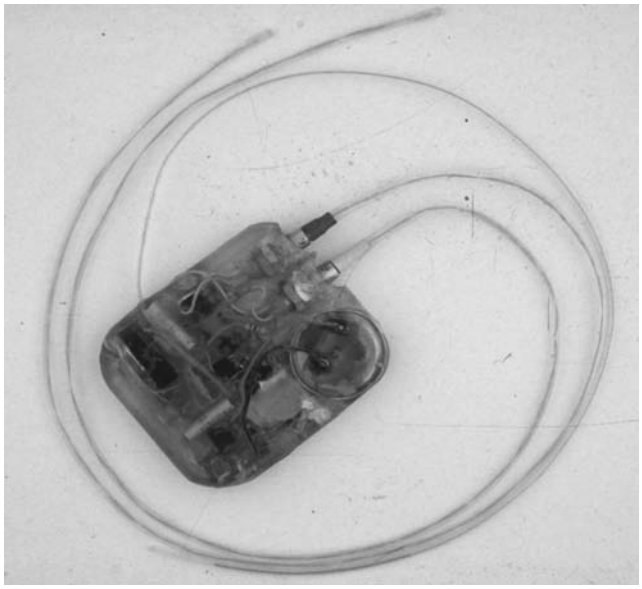


Figure 9. Animal prototype long-term central venous glucose sensor with implanted telemetry (73). Glucose and oxygen sensors are at the end of the catheters. Telemetry antenna emerges from the top, left. The telemetry body is 2×2.5 in.

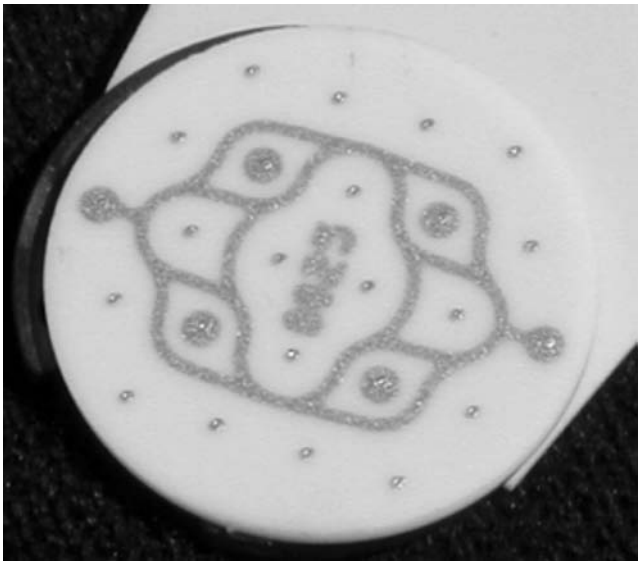


Figure 11. Close-up view of tissue glucose and oxygen sensor array (77). Sensor array with small (125 μm diameter) independent platinum working electrodes, large (875 μm diameter) common platinum counterelectrodes, and a curved common Ag/AgCl reference electrode. The membrane is not present. (Copyright 2003, American Physiological Society.)

These results have lead to several unanticipated conclusions. Although native glucose oxidase is intrinsically unstable, with appropriate sensor design the apparent catalytic lifetime of the immobilized enzyme can be substantially extended (75). The potentiostatic oxygen sensor is remarkably stable (76) and the oxygen deficit, once thought to be an insurmountable barrier, can be easily overcome (51). The central venous implant site, which is uniquely characterized by slow, steady flow of blood, allows for sufficient long-term biocompatibility with blood that the sensor stability can be documented (23). Nevertheless, the potential for thromoembolic events, anticipated to be rare but potentially significant over many years of sensor use, suggests reservations that may limit clinical acceptance and provides motivation for development of a potentially safer long-term sensor implant in tissues.

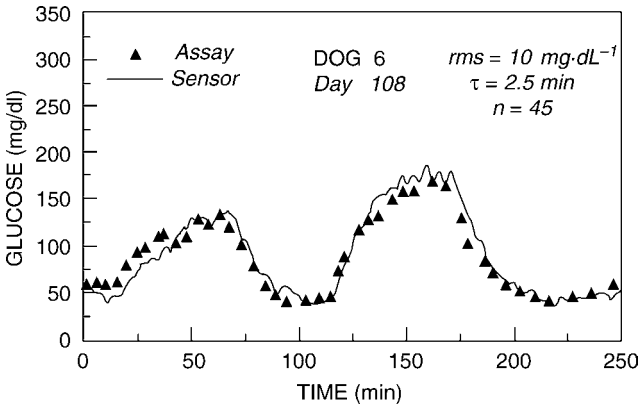


Figure 10. Response of an implanted intravenous sensor to glucose challenges on day 108 after implantation in a dog (23). The solid line is the sensor signal and triangles are venous blood glucose assays. Blood glucose excursions with initial rates of 0.2-8 m$M \cdot \text{min}^{-1}$ were produced by infusions of sterile glucose solutions through an intravenous catheter in a foreleg vein. (*Note*: 90 mg $\cdot$ dL^{-1} glucose = 5.0 m*M*.) (Copyright © 1990, American Diabetes Association.)

Long-term oxygen-based sensors have also been implanted in tissues. The successful central venous sensor cannot be simply adopted for use in the safer tissue site, but certain design features of that sensor which promote long-term function, such as immobilized enzyme design, the stable potentiostatic oxygen sensor, and membrane design to eliminate the oxygen deficit, can be incorporated (Fig. 11).

A systematic approach is required to validate sensor function, based on quantitative experimentation, mass transfer analysis, and accounting for properties of tissues that modulate glucose signals. Several new tools and methods have been developed. A tissue window chamber has been developed that allows direct optical visualization of implanted sensors in rodents, with surrounding tissue and microvasculature, while recording sensor signals (77) (Fig. 12). This facilitates determination of the effects of microvascular architecture and perfusion on the sensor signal. A method has been devised for sensor characterization in the absence of mass transfer boundary layers (78) that can be carried out before implantation and after explantation to infer stability of the implanted sensor. This allows quantitative assessment of mass transfer resistance within the tissue and the effects of long-term tissue changes. A sensor array having multiple glucose and oxygen sensors has also been developed that shows the range of variation of sensor responses within a given tissue (77). This provides a basis for averaging

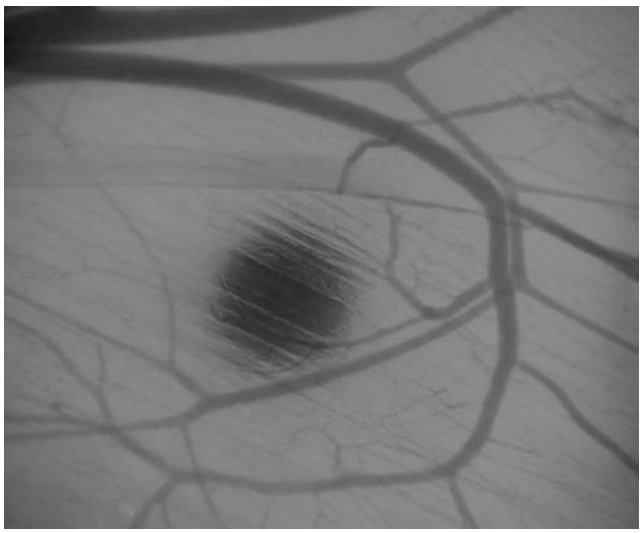


Figure 12. An implanted glucose sensor and nearby microvasculature (79). Optical image taken in a hamster window chamber. Sensor diameter is 125 μm.

sensor signals for quantitative correlation to blood glucose concentration.

REMAINING CHALLENGES FOR SENSOR DEVELOPMENT

Although there has been recent progress in sensor development and implementation, certain challenges remain. In many cases, there is need for improvement in data presentation and sensor validation. Standard glucose measurements for validation of sensor signals are often either not reported or are not obtained frequently enough to validate sensor responses. Requirements for sensor recalibration are not always given. Published results are often selected to show what may be ideally possible for a particular sensor rather than what is typical, sometimes conveying the impression that sensors are more accurate than may be the case.

There is a need to understand the effects of physiologic phenomena such as local perfusion, tissue variability, temperature and movement, that modulate sensor responses to glucose and affect measurement accuracy. A detailed understanding of these effects and their dynamics is needed for a full description of the glucose sensing mechanism. Robust sensor designs and modes of operation are required that assure reliable determination of glucose during exercise, sleeping and other daily life conditions.

A complete explanation for the response of every sensor should be sought, whether it is producing "good" or "bad" results, as more can often be learned for sensor improvement from sensors that produce equivocal results than from those that produce highly predictable signals (56). Present definitions as to what constitutes an acceptable sensor are based on narrow technical criteria proposed by sponsors of individual sensors that apply under specific conditions and lead to limited-use approvals by the FDA. There is a need to establish rational criteria for sensor validation and performance specific for the intended use (80). As sensors must be useful for hypoglycemia detection, sensor function must be validated in the hypoglycemic state. Correlation with Hb_{A1c} levels may not be useful for sensor validation, as the detailed information from sensors is likely to supplant Hb_{A1c} as a more comprehensive index of control.

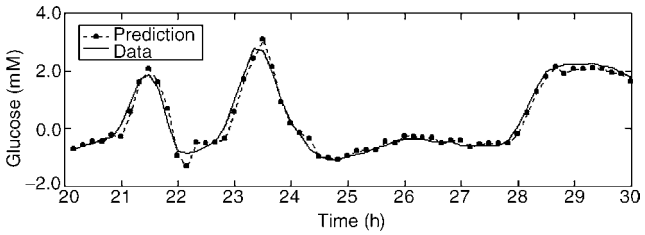


Figure 13. Blood glucose prediction based on recently sampled values (13). A 10 min prediction in a non-diabetic, average rms error = 0.2 m*M*. (Copyright © 1999, American Diabetes Association.)

BLOOD GLUCOSE PREDICTION

The ability to monitor blood glucose in real-time has major advantages over present methods based on sample collection that provide only sparse, historical information. There exists, however an additional possibility of using sensor information to *predict* future blood glucose values. It been demonstrated that blood glucose dynamics are not random and that blood glucose values can be predicted using autoregressive moving average (ARMA) methods, at least for the near future, from frequently sampled previous values (13) (Fig. 13). Prediction based only on recent blood glucose history is particularly advantageous because there is no need to involve models of glucose and insulin distribution, with their inherent requirements for detailed accounting of glucose loads and vascular insulin availability. This capability may be especially beneficial to children. Glucose prediction can potentially amplify the considerable benefits of continuous glucose sensing, and may represent an even further substantial advance in blood glucose management.

CLOSING THE LOOP

Glucose control is an example of a classical control system (Fig. 14). To fully implement this system, there is a need to establish a programmable controller based on continuous glucose sensing, having control laws or algorithms to counter hyper- and hypoglycemic excursions, identify performance targets for optimal insulin administration, and employ insulin pumps. The objective is restore optimal blood glucose control while avoiding over-insulinization by adjusting the program, a goal that may not be possible to achieve with alternative cell- or tissue-based insulin replacement strategies.

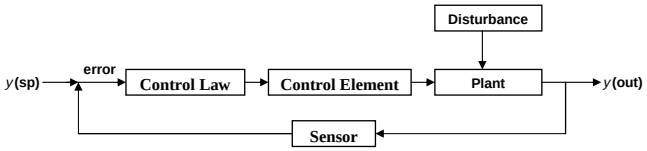


Figure 14. A simple control system for blood glucose. y(out) is the blood glucose concentration, y(sp) is the desired blood glucose, the natural sensor is in the pancreatic beta cell, the plant is the body over which glucose is distributed, and the disturbance is absorption of glucose from the gut via digestion. The control element can be an insulin pump. The control law is an algorithm that directs the pump in response to the difference between measured and target blood glucose.

Programmable external pumps that deliver insulin to the subcutaneous tissue are now widely used and implanted insulin pumps may soon become similarly available. At present, these devices operate mainly in a preprogrammed or *open-loop* mode, with occasional adjustment of the delivery rate based on fingerstick glucose information. However, experimental studies in humans have been reported utilizing *closed-loop* systems based on implanted central venous sensors and intra-abdominal insulin pumps in which automatic control strategies were employed over periods of several hundred days (81) (Fig. 15). Initial inpatient trials using subcutaneous peroxide sensors to close the loop with an external insulin pump are also underway. There is a need to expand development of such systems for broad acceptance. Extensive reviews of pump development can be found elsewhere (82–84).

These results demonstrate that an implantable artificial beta cell is potentially feasible, but more effort is required to incorporate a generally acceptable glucose sensor, validate the system extensively, and demonstrate its robust response.

CONCLUSIONS

The need for new glucose sensors in diabetes is now greater than ever. Although development of an acceptable, continuous and automatic glucose sensor has proven to be a substantial challenge, progress over the past several decades has defined sensor performance requirements and has focused development efforts on a limited group of promising candidates. The advent of new glucose sensing technologies could facilitate fundamentally new approaches to the therapy of diabetes.

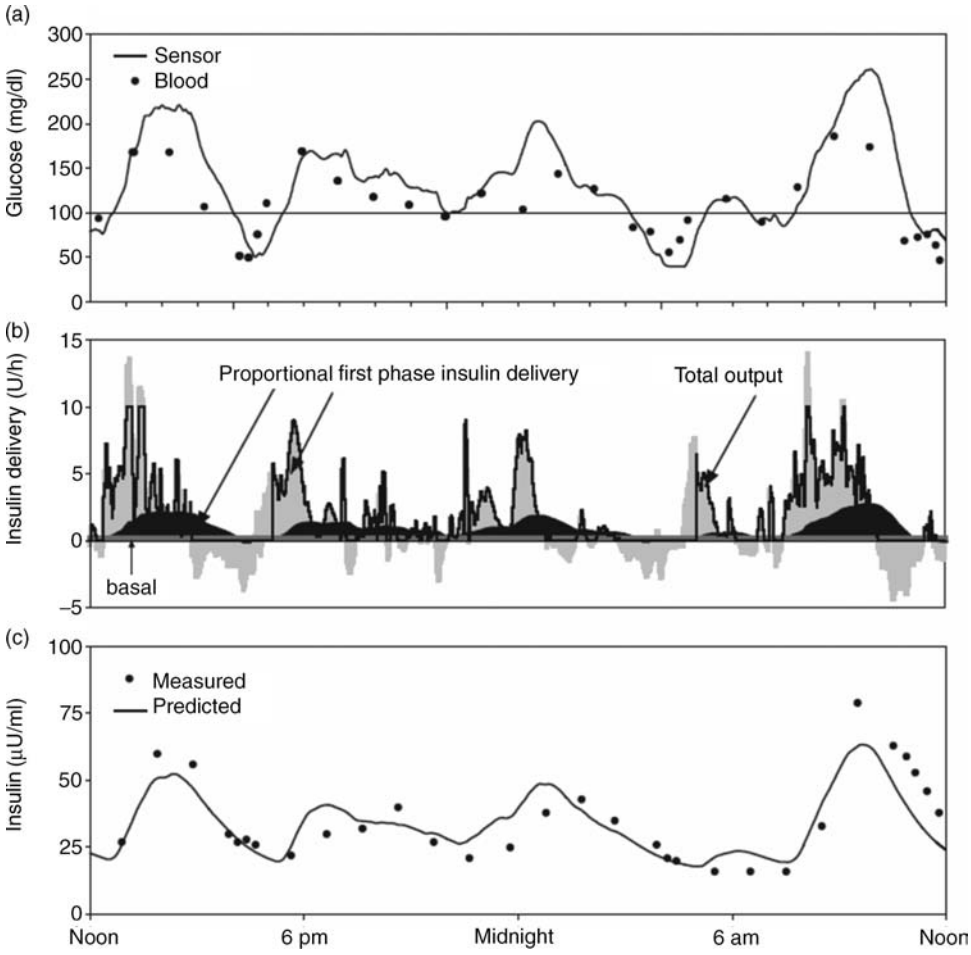


Figure 15. Blood glucose control in humans by an implanted artificial beta cell. A chronic, central venous blood glucose sensor and implanted insulin pump (Medtronic/MiniMed) implanted in a human subject. (a) Plasma (solid circles) and sensor (line) glucose following initiation of closed-loop control (noon). Solid line at 100 $mg \cdot dL^{-1}$ indicates setpoint. (b) Proportional (medium shading), basal (light shading), and derivative (dark shading) insulin delivery during the closed-loop (solid line indicates total, which is not allowed to go below zero). (c) Plasma (circles) and predicted insulin (solid line) concentrations. (Study performed by Medical Research Group. Copyright 2004, Elsevier.)

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health and the Technology for Metabolic Monitoring Initiative of the Department of Defense. D.G. holds equity interest in GlySens, Inc., a company dedicated to the development of a new glucose sensor, and is a scientific advisor. This arrangement that has been approved by the University of California, San Diego in accordance with its conflict of interest policies.

BIBLIOGRAPHY

Cited References

1. Cahill Jr GF, Etzwiler LD, Freinkel N. Editorial: "Control" and diabetes. N Engl J Med 1976;294(18): 1004–1005.
2. The effect of intensive treatment of diabetes on the development and progression of long–term complications in insulin-dependent diabetes mellitus. The Diabetes Control and Complications Trial Research Group. N Engl J Med 1993;329(14): 977–986.
3. Stratton IM, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. Bmj 2000; 321(7258):405–412.
4. Nathan DM, et al. The clinical information value of the glycosylated hemoglobin assay. N Engl J Med 1984;310(6): 341–346.
5. Skyler JS, Oddo C. Diabetes trends in the USA. Diabetes Metab Res Rev 2002;18(3 Suppl):S21–S26.
6. Zimmet P, Alberti KG, Shaw J. Global and societal implications of the diabetes epidemic. Nature (London) 2001;414 (6865):782–787.
7. Egger M, et al. Risk of adverse effects of intensified treatment in insulin-dependent diabetes mellitus: a meta-analysis. Diabet Med 1997;14(11):919–928.
8. Rovet JF, Ehrlich RM. The effect of hypoglycemic seizures on cognitive function in children with diabetes: a 7-year prospective study. J Pediatr 1999;134(4):503–506.
9. Sovik O, Thordarson H. Dead-in-bed syndrome in young diabetic patients. Diabetes Care 1999;22(2 Suppl):B40–B42.
10. Santiago JV. Lessons from the Diabetes Control and Complications Trial. Diabetes 1993;42(11):1549–1554.
11. van den Berghe G, et al. Intensive insulin therapy in the critically ill patients. N Engl J Med 2001;345(19):1359–1367.
12. Van den Berghe G, et al. Outcome benefit of intensive insulin therapy in the critically ill: Insulin dose versus glycemic control. Crit Care Med 2003;31(2):359–366.
13. Bremer T, Gough DA. Is blood glucose predictable from previous values? A solicitation for data. Diabetes 1999;48 (3):445–451.
14. Clark Jr LC, Lyons C. Electrode systems for continuous monitoring in cardiovascular surgery. Ann NY Acad Sci 1962; 102:29–45.
15. Chang KW, et al. Validation and bioengineering aspects of an implantable glucose sensor. Trans Am Soc Artif Intern Org 1973;19:352–360.
16. Slama G, Bessman SP. [Results of *in vitro* and *in vivo* use of a prototype implantable glucose sensor]. J Annu Diabetol Hotel Dieu 1976; 297–302.
17. Albisser AM, et al. An artificial endocrine pancreas. Diabetes 1974;23(5):389–396.
18. Thomas LJ, Bessman SP. Prototype for an implantable insulin delivery pump. Proc West Pharmacol Soc 1975;18:393–398.
19. Clemens AH, Chang PH, Myers RW. The development of Biostator, a Glucose Controlled Insulin Infusion System (GCIIS). Horm Metab Res 1977; (7 Suppl):23–33.
20. Skyler JS, et al. Home blood glucose monitoring as an aid in diabetes management. Diabetes Care 1978;1(3):150–157.
21. Pickup JC, et al. Clinical application of pre-programmed insulin infusion: continuous subcutaneous insulin therapy with a portable infusion system. Horm Metab Res 1979; (8) (Suppl):202–204.
22. McCall AL, Mullin CJ. Home monitoring of diabetes mellitus–a quiet revolution. Clin Lab Med 1986;6(2):215–239.
23. Armour JC, et al. Application of chronic intravascular blood glucose sensor in dogs. Diabetes 1990;39(12):1519–1526.
24. Sage Jr BH. FDA panel approves Cygnus's noninvasive GlucoWatch. Diabetes Technol Ther 2000;2(1): 115–116.
25. Clinical Practice Recommendations 2005. Diabetes Care 2005;28(1 Suppl):S1–S79.
26. Lojo J, et al. Prevantive Care Practices Among Persons with Diabetes, United States, 1995 and 2001. Morbidity Mortality Weekly Rep, Center Disease Control Prevention 2002;51(43): 965–967.
27. Bennion N, Christensen NK, McGarraugh G. Alternate site glucose testing: a crossover design. Diabetes Technol Ther 2002;4(1):25–33; discussion 45–47.
28. Gough DA, Kreutz-Delgado K, Bremer TM. Frequency characterization of blood glucose dynamics. Ann Biomed Eng 2003;31(1):91–97.
29. McNichols RJ, Cote GL. Optical glucose sensing in biological fluids: an overview. J Biomed Opt 2000;5(1):5–16.
30. Cote GL. Noninvasive and minimally-invasive optical monitoring technologies. J Nutr 2001;131(5):1596S–604S.
31. Arnold MA, Burmeister JJ, Small GW. Phantom glucose calibration models from simulated noninvasive human near-infrared spectra. Anal Chem 1998;70(9):1773–1781.
32. Goetz Jr MJ, et al. Application of a multivariate technique to Raman spectra for quantification of body chemicals. IEEE Trans Biomed Eng 1995;42(7):728–731.
33. Steffes PG. Laser-based measurement of glucose in the ocular aqueous humor: an efficacious portal for determination of serum glucose levels. Diabetes Technol Ther 1999;1(2):129–133.
34. Cameron BD, Baba JS, Cote GL. Measurement of the glucose transport time delay between the blood and aqueous humor of the eye for the eventual development of a noninvasive glucose sensor. Diabetes Technol Ther 2001;3(2):201–207.
35. Blass DA, Adams E. Polarimetry as a general method for enzyme assays. Anal Biochem 1976;71(2):405–414.
36. Baba JS, et al. Effect of temperature, pH, and corneal birefringence on polarimetric glucose monitoring in the eye. J Biomed Opt 2002;7(3):321–328.
37. Schultz JS, Mansouri S, Goldstein IJ. Affinity sensor: a new technique for developing implantable sensors for glucose and other metabolites. Diabetes Care 1982;5(3): 245–253.
38. March WF, Ochsner K, Horna J. Intraocular lens glucose sensor. Diabetes Technol Ther 2000;2(1):27–30.
39. Muller M. Science, medicine, and the future: Microdialysis. Bmj 2002;324(7337):588–591.

40. Meyerhoff C, et al. On line continuous monitoring of subcutaneous tissue glucose in men by combining portable glucosensor with microdialysis. Diabetologia 1992;35(11): 1087–1092.
41. Hoss U, et al. A novel method for continuous online glucose monitoring in humans: the comparative microdialysis technique. Diabetes Technol Ther 2001;3(2):237–243.
42. Tamada JA, Bohannon NJ, Potts RO. Measurement of glucose in diabetic subjects using noninvasive transdermal extraction. Nat Med 1995;1(11):1198–1201.
43. Potts RO, Tamada JA, Tierney MJ. Glucose monitoring by reverse iontophoresis. Diabetes Metab Res Rev 2002;18 (1 Suppl):S49–S53.
44. Lenzen H, et al. A non-invasive frequent home blood glucose monitor. Practical Diabetes Int 2002;19(4):101–103.
45. Tierney MJ, et al. The GlucoWatch biographer: a frequent automatic and noninvasive glucose monitor. Ann Med 2000; 32(9):632–641.
46. Bindra DS, et al. Design and *In vitro* studies of a needle-type glucose sensor for subcutaneous monitoring. Anal Chem 1991;63(17):1692–1696.
47. Jablecki M, Gough DA. Simulations of the frequency response of implantable glucose sensors. Anal Chem 2000;72(8):1853–1859.
48. Ward WK, et al. A new amperometric glucose microsensor: *in vitro* and short-term *In vivo* evaluation. Biosens Bioelectron 2002;17(3):181–189.
49. Xie SL, Wilkins E. Rechargeable glucose electrodes for long-term implantation. J Biomed Eng 1991;13(5):375–378.
50. Clark Jr LC. Membrane Polarographic Electrode System and Method with Electrochemical Compensation. US pat 3,539,455. 1970.
51. Gough DA, Lucisano JY, Tse PH. Two-dimensional enzyme electrode sensor for glucose. Anal Chem 1985;57(12):2351–2357.
52. Mastrototaro JJ. The MiniMed continuous glucose monitoring system. Diabetes Technol Ther 2000;2(1 Suppl): S13–S18.
53. Rebrin K, et al. Subcutaneous glucose predicts plasma glucose independent of insulin: implications for continuous monitoring. Am J Physiol 1999;277(3 Pt. 1): E561–E571.
54. Gross TM, Mastrototaro JJ. Efficacy and reliability of the continuous glucose monitoring system. Diabetes Technol Ther 2000;2(1 Suppl):S19–S26.
55. Metzger M, et al. Reproducibility of glucose measurements using the glucose sensor. Diabetes Care 2002;25(7):1185–1191.
56. Gough DA, Armour JC. Development of the implantable glucose sensor. What are the prospects and why is it taking so long? Diabetes 1995;44(9):1005–1009.
57. Shichiri M, et al. Telemetry glucose monitoring device with needle-type glucose sensor: a useful tool for blood glucose monitoring in diabetic individuals. Diabetes Care 1986;9 (3):298–301.
58. Abel P, Muller A, Fischer U. Experience with an implantable glucose sensor as a prerequisite of an artificial beta cell. Biomed Biochim Acta 1984;43(5):577–584.
59. Moatti-Sirat D, et al. Towards continuous glucose monitoring: *In vivo* evaluation of a miniaturized glucose sensor implanted for several days in rat subcutaneous tissue. Diabetologia 1992;35(3):224–230.
60. Johnson KW, et al. *In vivo* evaluation of an electroenzymatic glucose sensor implanted in subcutaneous tissue. Biosens Bioelectron 1992;7(10):709–714.
61. Kerner W, et al. The function of a hydrogen peroxide-detecting electroenzymatic glucose electrode is markedly impaired in human sub-cutaneous tissue and plasma. Biosens Bioelectron 1993;8(9–10):473–482.
62. Bode BW, et al. Continuous glucose monitoring used to adjust diabetes therapy improves glycosylated hemoglobin: a pilot study. Diabetes Res Clin Pract 1999;46(3):183–190.
63. Gross TM, et al. Performance evaluation of the MiniMed continuous glucose monitoring system during patient home use. Diabetes Technol Ther 2000;2(1): p. 49–56.
64. Gough DA, Botvinick EL. Reservations on the use of error grid analysis for the validation of blood glucose assays. Diabetes Care 1997;20(6): p. 1034–1036.
65. Kerssen A, de Valk HW, Visser GH. The Continuous Glucose Monitoring System during pregnancy of women with type 1 diabetes mellitus: accuracy assessment. Diabetes Technol Ther 2004;6(5): p. 645–51.
66. Mauras N, et al. Lack of accuracy of continuous glucose sensors in healthy, nondiabetic children: results of the Diabetes Research in Children Network (DirecNet) accuracy study. J Pediatr 2004;144(6):770–775.
67. Gilligan BJ, et al. Evaluation of a subcutaneous glucose sensor out to 3 months in a dog model. Diabetes Care 1994;17(8): 882–887.
68. Updike SJ, et al. A subcutaneous glucose sensor with improved longevity, dynamic range, and stability of calibration. Diabetes Care 2000;23(2):208–214.
69. Garg SK, Schwartz S, Edelman SV. Improved glucose excursions using an implantable real-time continuous glucose sensor in adults with type 1 diabetes. Diabetes Care 2004; 27(3):734–738.
70. Csoregi E, Schmidtke DW, Heller A. Design and optimization of a selective subcutaneously implantable glucose electrode based on "wired" glucose oxidase. Anal Chem 1995;67(7): 1240–1244.
71. Heller A. Implanted electrochemical glucose sensors for the management of diabetes. Annu Rev Biomed Eng 1999;1:153–175.
72. Tse PHS, Leypoldt JK, Gough DA. "Determination of the Intrinsic Kinetic Constants of Immobilized Glucose Oxidase and Catalase". Biotechnol Bioeng 1987;29:696–704.
73. McKean BD, Gough DA. A telemetry-instrumentation system for chronically implanted glucose and oxygen sensors. IEEE Trans Biomed Eng 1988;35(7):526–532.
74. Medtronic Minimed talk at the Diabetes Technology and Therapeutics Conference, S.F., CA. 2003.
75. Gough DA, Bremer T. Immobilized glucose oxidase in implantable glucose sensor technology. Diabetes Technol Ther 2000; 2(3):377–380.
76. Lucisano JY, Armour JC, Gough DA. *In vitro* stability of an oxygen sensor. Anal Chem 1987;59(5):736–739.
77. Makale MT, et al. Tissue window chamber system for validation of implanted oxygen sensors. Am J Physiol Heart Circ Physiol 2003;284(6):H2288–H2294.
78. Makale MT, Jablecki MC, Gough DA. Mass transfer and gas-phase calibration of implanted oxygen sensors. Anal Chem 2004;76(6):1773–1777.
79. Makale MT, Chen PC, Gough DA. Variants of the tissue/sensor array chamber. Am J Physiol Heart Circ Physiol 2005;286, in press.
80. Bremer TM, Edelman SV, Gough DA. Benchmark data from the literature for evaluation of new glucose sensing technologies. Diabetes Technol Ther 2001;3(3):409–418.
81. Steil GM, Panteleon AE, Rebrin K. Closed-loop insulin delivery-the path to physiological glucose control. Adv Drug Deliv Rev 2004;56(2):125–144.

82. Saudek CD. Implantable Pumps. 3rd ed. International Textbook of Diabetes Mellitus; 2004.
83. Selam JL. External and implantable insulin pumps: current place in the treatment of diabetes. Exp Clin Endocrinol Diabetes 2001;109(2 Suppl):S333–S340.
84. Vague P, et al. The implantable insulin pump in the treatment of diabetes. Hopes and reality?. Bull Acad Natl Med 1996; 180(4):831–41. discussion 841–843.

See also FIBER OPTICS IN MEDICINE; OPTICAL SENSORS; OXYGEN SENSORS; PANCREAS, ARTIFICIAL.

H

HBO THERAPY. See HYPERBARIC OXYGENATION.

HEARING IMPAIRMENT. See AUDIOMETRY.

HEART RATE, FETAL, MONITORING OF. See FETAL MONITORING.

HEART VALVE PROSTHESES

K. B. CHANDRAN
University of Iowa
Iowa City, Iowa

INTRODUCTION

The human circulatory system provides adequate blood flow without interruption to the various organs and tissues and regulates blood supply to the demands of the body. The contracting heart supplies the energy required to maintain the blood flow through the vessels. The human heart consists of two pumps in series. The right side of the heart, a low pressure pump, consisting of the right atrium and the right ventricle supplies blood to the pulmonary circulation. The left side consisting of the left atrium and the left ventricle is the high pressure pump circulating blood through the systemic circulation. Figure 1 is a schematic representation of the four chambers of the heart and the arrows indicate the direction of blood flow. The pressure gradients developed between the main arteries supplying blood to the systemic and pulmonary circulation and the respective venous ends are the driving forces causing the blood flow and the energy is dissipated in the form of heat due to frictional resistance.

The four valves in the heart ensure that the blood flows only in one direction. The blood from the systemic circulation supplies nutrients and oxygen to the cells for the various tissues and organs and removes carbon dioxide at the level of capillaries. The oxygen depleted blood returns through the systemic veins to the right atrium. During the ventricular relaxation or diastole, the blood passes through the tricuspid valve into the right ventricle. In the ventricular contraction phase of the cardiac cycle or systole, the tricuspid valve closes and the pulmonic valve opens to pump the blood to the lungs through the pulmonary arteries. Carbon dioxide is removed and oxygen is absorbed by the blood in the capillaries of the lungs that is surrounded by the alveolar sac with the air we breathe. The oxygen-rich blood returns to the left atrium via the pulmonary veins and passes through the mitral (bicuspid) valve into the left ventricle during the ventricular diastole. During the ventricular contraction, the mitral valve closes and the aortic valve opens to pump the blood through the systemic circulation. The main function of the heart valves is to control the direction of blood flow permitting flow in the forward direction and preventing regurgitation or back flow through the closed valves.

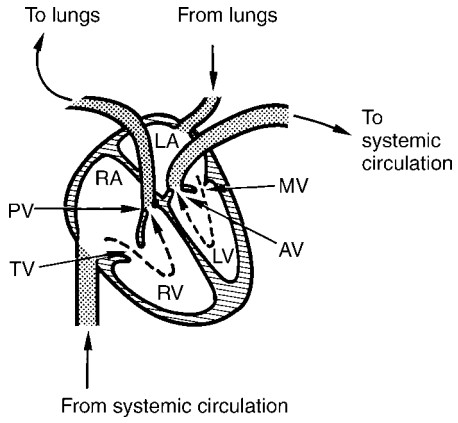


Figure 1. Schematic of blood flow in the human heart. LA-Left atrium; RA-Right atrium; LV-Left ventricle; RV-Right ventricle; PV-pulmonary valve; TV-Tricuspid valve; AV-Aortic valve; and MV-Mitral valve.

Anatomy of the Native Valves

The aortic valve (Fig. 2) consists of three semicircular (semilunar) leaflets or cusps within a connective tissue sleeve (1) attached to a fibrous ring. The cusps meet at three commissures that are equally spaced along the circumference at the supraaortic ridge. This ridge is thickening of the aorta at which the cusps insert and there is no continuity of tissue from one cusp to the other across the commissure. The leaflet consists of three layers as shown in Fig. 3: the aortic side layer is termed the fibrosa and is the

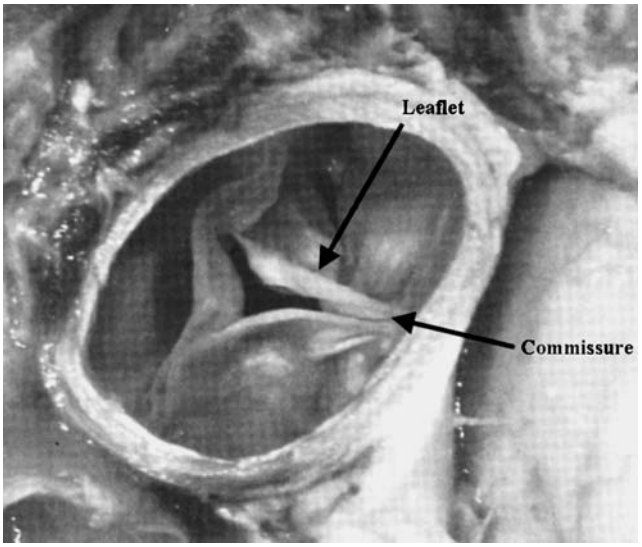


Figure 2. Human aortic valve viewed from the aorta. (Adapted with permission from Otto, C. M. Valvular Heart Disease, Second Edition, 2004, Saunders/Elsevier, Inc., Philadelphia, PA.)

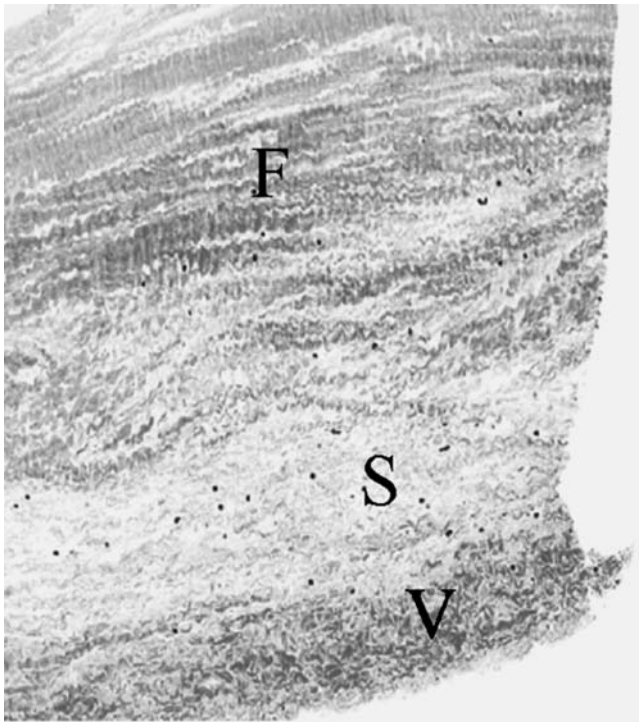


Figure 3. A histologic section of an aortic valve leaflet depicting the three layers along the thickness of the leaflet: F = Fibrosa; S = Spongiosa; and V = Ventricularis. (Courtesy of Prof. Michael Sacks of the University of Pittsburgh, Pittsburgh, PA.)

major fibrous layer within the body of the leaflet; layer on the ventricular side termed ventricularis is composed of both collagen and elastin; and the central portion of the valve termed the spongiosa consisting of loose connective tissue, proteins, and glycosaminoglycans (GAG). The leaflet length is larger than the radius of the annulus, and hence a small overlap of the tissue from each leaflet protrudes and forms a coaptation surface when the valve is closed to ensure that the valve is sealed in the closed position. The sinus of Valsalva is attached to the fibrous annular ring on the aortic side and is comprised of three bulges at the root of the aorta. Each bulge is aligned with the belly or the central part of the valve leaflet. The left and the right sinuses contain the coronary ostia (openings) giving rise to the left and right coronary arteries, respectively, providing blood flow and nutrients to the cardiac muscles. When the valve is fully open, the leaflets extend to the upper edges of the sinuses. The anatomy of the pulmonic valve is similar to that of the aortic valve, but the sinuses in the pulmonary artery are smaller than the aortic sinuses, and the pulmonic valve orifice is slightly larger. The average aortic valve orifice area is $\sim$4.6 cm^2 and is $\sim$4.7 cm^2 for the pulmonic valve (2). In the closed position, the pulmonic valve is subject to a pressure of $\sim$30 mmHg (3.99 kPa) while the load on the aortic valve is $\sim$100 mmHg (13.30 kPa).

The mitral and tricuspid valves are also anatomically similar with the mitral valve consisting of two main leaflets (cusps) compared to three for the valve in the right side of the heart. The valves consist of the annulus, leaflets, papillary muscles, and the chordae tendinae (Fig. 4). The average mitral and tricuspid valve orifice areas are 7.8 and 10.6 cm^2, respectively (2). The atrial and ventricular walls are attached to the mitral annulus, consisting of dense collagenous tissue surrounded by muscle, at the base of the leaflets. The chordae tendinae are attached to the free edge of the leaflets at multiple locations and extend to the tip of the papillary muscles. Anterior and posterior leaflets of the mitral valve are actually one continuous tissue with two regularly spaced indentations called the commissures. The combined surface area of both the leaflets is approximately twice the area of the valve orifice and thus the leaflets coaptate during the valve closure. The posterior leaflet encircles two-thirds of the annulus and is quadrangular shaped, while the anterior leaflet is semilunar shaped. The left ventricle has two papillary muscles that attach to the ventricular free wall and tether the mitral valve in place via the chordae tendinae. This tethering prevents the leaflets from prolapsing into the left atrium during ventricular ejection. Improper tethering will result in the leaflets extending into the atrium and incomplete apposition of the leaflets will permit blood to

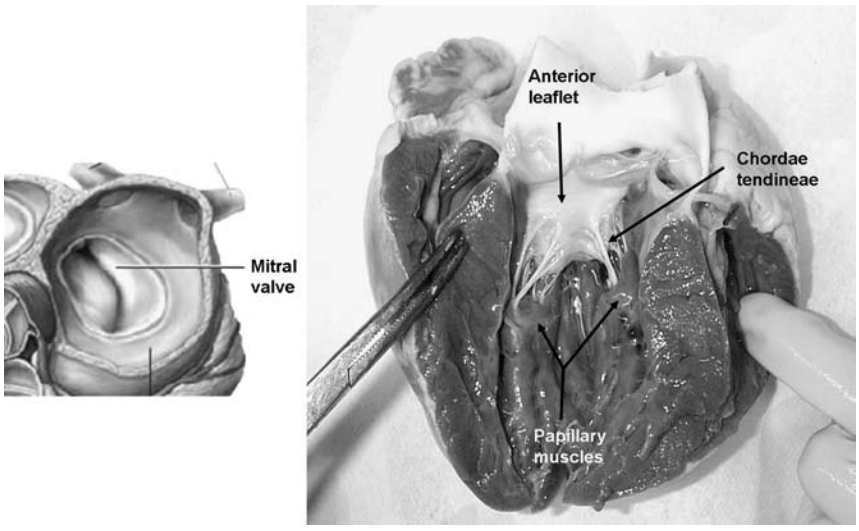


Figure 4. Schematic of the human mitral (bicuspid) valve and a photograph showing the anterior leaflet with the chordae tendinae attachment with papillary muscles. (Courtesy of Prof. Ajit Yoganathan from Georgia Institute of Technology.)

regurgitate back to the atrium. The tricuspid valve has three leaflets, a septal leaflet along with the anterior and posterior leaflets, and is larger and structurally more complicated than the mitral valve.

Valve Dynamics

At the beginning of systole, the left ventricle starts to contract and with the increase in pressure, the mitral valve closes preventing regurgitation of blood to the left atrium. During the isovolumic contraction with both the mitral and aortic valves closed, the ventricular pressure rises rapidly. The aortic valve opens when the left ventricular pressure exceeds the aortic pressure. The blood accelerates rapidly through the open valve and peak velocity of flow occurs during the first third of systole. The pressure difference between the left ventricle and the aorta required to open the valve is of the order of 1–2 mmHg (0.13–0.26 kPa). During the forward flow phase, vortices develop in the three sinuses behind the open leaflets and the formation of such vortices was first described by Leonardo da Vinci in 1513. Several studies have suggested that the vortices and the difference in pressure between the sinuses and the center of the aortic orifice pushes the leaflets toward closure even during the second third of systole when forward flow of blood continues. With the ventricular relaxation and rapid drop in the ventricular pressure, an adverse pressure gradient between the ventricle and the aorta moves the leaflets toward full closure with negligible regurgitation of blood from the aorta to the ventricle. Systole lasts for about one-third of the cardiac cycle and the peak pressure reached in the aorta during systole in healthy humans is ~120 mmHg (15.96 kPa) and the diastolic pressure in the aorta with the aortic valve closed is ~80 mmHg (10.64 kPa).

At the beginning of diastole, the aortic valve closes and the ventricular pressure decreases rapidly during the isovolumic relaxation. As the ventricular pressure falls below the atrial pressure, the mitral valve opens and the blood flows from the atrium to the ventricle. The pressure difference between the left atrium and the ventricle required to open the mitral valve and drive the blood to fill the ventricle is smaller than that required with the aortic valve (< 1 mmHg or 0.13 kPa). As the blood fills the ventricle, vortical flow is established in the ventricle, and it has been suggested that the leaflets move toward closure due to the same. The atrial contraction induces additional flow of blood from the atrium into the ventricle during the second half of diastole and the adverse pressure gradient at the beginning of ventricular contraction forces the mitral valve to close and isovolumic contraction takes place. The chordae tendinae prevents the prolapse of the leaflets into the left atrium when the mitral valve is in the closed position. The dynamics of the mitral valve opening and closing is a complex process involving the leaflets, mitral orifice, chordae tendinae, and the papillary muscles. During systole, the closed mitral valve is subjected to pressures of ~120 mmHg (15.96 kPa).

The dynamics of opening and closing of the pulmonic and the tricuspid valves are similar to the aortic and mitral valve, respectively, even though the pressures generated in the right ventricle and the pulmonary artery are generally about a one-third of the corresponding magnitudes in the left side of the heart. From the description of the valve dynamics, one can observe several important features on the normal functioning of the heart valves. These include opening efficiently with minimal difference in pressure between the upstream and downstream sides of the valve, and efficient closure to ensure minimal regurgitation. In addition, the flow past the valves are laminar with minimal disturbances in flow and the fluid induced stresses do not activate or destroy the formed elements in blood such as the platelets and red blood cells. As the valves open and close, the leaflets undergo complex motion that includes large deformation, as well as bending. The leaflet material is also subjected to relatively high normal and shear stresses during these motions. The valves open and close at about once every second, and hence functions over several million cycles during the normal life of a human. These are some of the important considerations in the design and functional evaluation of heart valve prostheses that we consider in detail below.

Valvular Diseases and Replacement of the Diseased Valves

Valvular diseases are more common on the left heart due to the high pressure environment for the aortic and mitral valves and also with the tricuspid valve on the right side of the heart. Valvular diseases include stenosis and incompetence. Stenosis of the leaflets is due to calcification resulting in stiffer leaflets that will require higher pressures to open the valves. Rheumatic fever is known to affect the leaflets resulting in stenosed valves (3). Premature calcification of the bicuspid valve, as well as significant obstruction of the left ventricular outflow in congenital aortic valve stenosis, also affects the valves of the left heart (3). Aortic sclerosis due to aging can also advance to valvular stenosis in some patients. Mitral stenosis may be the result of commissural fusion in younger patients and may also be due to cusp fibrosis. In the case of valvular stenosis, higher pressure needs to be generated to force the stiffer leaflets to open and the valvular orifice area in the fully open position may be significantly reduced. Effective orifice area (EOA) can be computed by the application of the Gorlin equation (4) based on the fluid mechanic principles and is given by the following relationship:

$$\mathrm{EOA}(\mathrm{cm}^2) = \frac{Q_{\mathrm{rms}}}{C\sqrt{\Delta\overline{p}}} \tag{1}$$

In this equation, Q_{rms} is the root-mean-square (rms) flow rate (mL/s) during the forward flow through the valve and $\Delta\overline{p}$ is the mean pressure drop (mmHg) across the open valve. The measurement of mean pressure drop *in vivo* is described later, and the flow rate across the valve during the forward flow phase is computed from the measurement of cardiac output and the heart rate. The parameter C represents a constant that is based on the discharge coefficient used for the aortic or mitral valve, and the unit conversion factors to result in the computed area in terms of square centimeter. A more direct technique to estimate the effective orifice area is the application of conservation of mass principle. The systolic volume flow through the left

ventricular outflow tract is determined as the product of the outflow tract cross-sectional area and the flow velocity–time integral. Since the same blood volume must also pass through the valve orifice, the valve orifice area is computed by dividing the volume flow with the measured aortic valve velocity–time integral (5). Replacement of the aortic valve is generally considered when the measured valvular orifice area is $< 0.4\ cm^2 \cdot m^{-2}$ of body surface area (6). The corresponding value for the mitral stenosis is $1.0\ cm^2 \cdot m^{-2}$.

Valvular incompetence results from incomplete closure of the leaflets resulting in significant regurgitation of blood. Incompetence could be the result of decrease in leaflet area due to rheumatic disease or perforations in the leaflets due to bacterial endocarditis. Structural alterations due to loss of commissural support or aortic root dilatation can result in aortic valve incompetence. Rupture of chordae tendinae, leaflet perforation, and papillary muscle abnormality may also affect the mitral valve closure and increase in regurgitation. Optimal timing for valvular replacement in the case of native valve incompetence is not clearly defined.

Various methods for valvular reconstruction or repair are also being developed instead of replacement with prostheses since these techniques are associated with lower risk of mortality and lower risk of recurrence (7). Valvular repair rather than replacement is generally preferred for regurgitation due to segmental prolapse of the posterior mitral leaflet. Implantation of a prosthetic ring to reduce the size of the mitral orifice and improve leaflet coaptation is performed in the case of mitral regurgitation due to ring dilatation. Mitral endocarditis with valvular or chordal lesions is also repaired rather than replacing the whole valve. Dilatation of the root and prolapse of the cusps are also the most important causes for regurgitation of the aortic valves and several techniques have also been developed to correct these pathologies (7).

The cardiopulmonary bypass technique to reroute the blood from the vena cava to the ascending aorta, and the introduction of cold potassium cardioplegia to arrest the heart to perform open heart surgery introduced in the 1950s enabled the replacement of diseased valves. Replacement of severely stenosed and/or incompetent valves with prostheses is a common treatment modality today and patients with prosthetic valves lead a relatively normal life. Yet, significant problems are also encountered with implanted prosthetic valves and efforts are continuing to improve the design of the valves for enhanced functionality and minimizing the problems encountered with implantation.

PROSTHETIC HEART VALVES

Ideal Valve Design

An ideal valve to be used as replacement for a diseased valve should mimic the functional characteristics of the native human heart valves with the following characteristics (adapted from Ref. 8). The prosthetic valve should open efficiently with a minimal transvalvular pressure drop. The blood should flow through the orifice with central and undisturbed flow as is observed with healthy native heart valves. The valve should close efficiently with minimal amount of regurgitation. The material used for the valve should be biocompatible, durable, nontoxic, and nonthrombogenic. The valve will be anticipated to open and close for > 40 million cycles/year for many years and must maintain the structural integrity throughout the lifetime of the implant. Blood is a corrosive and chemically unstable fluid that tends to form thrombus in the presence of foreign bodies. To avoid thrombus formation, the valve surfaces must be smooth, and the flow past the valve should avoid regions of stagnation and recirculation as well as minimize flow-induced stresses that are factors related to thrombus initiation. The prosthetic valve should be surgically implantable with ease and should not interfere with the normal anatomy and function of the cardiac structures and the aorta. The valve should be easily manufactured in a range of sizes, inexpensive, and sterilizable.

Transplanting freshly explanted human heart valves from donors who died of noncardiovascular diseases is probably the most ideal replacement and such homograft valves have been successfully used as replacements. These are the only valves entirely consisting of fresh biological tissue and sewn into place resulting in an unobstructed central flow. The transplanted homograft valves are no longer living tissue, and hence lack the cellular regeneration capability of the normal valve leaflets. Thus, the transplanted valves are vulnerable to deterioration on a long-term use. Moreover, homograft valves are difficult to obtain except in trauma centers in large population areas, and hence not a viable option generally.

Numerous prosthetic valves have been developed over the past 40 years and most of the design and development of valvular prostheses have been empirical. The currently available heart valve prostheses can be broadly classified into two categories: mechanical heart valves (MHV) and bioprosthetic heart valves (BHV). Even though valves from both categories are routinely implanted in patients with valvular disease and the patients with prosthetic implants lead a relatively normal life, several major problems are encountered with the mechanical and biological prosthetic valves (9). These problems include: (*1*) thromboembolic complications; (*2*) mechanical failure due to fatigue or chemical changes; (*3*) mechanical damage to the formed elements in blood including hemolysis, activation, and destruction of platelets and protein denaturation; (*4*) perivalvular leak due to healing defects; (*5*) infection; and (*6*) tissue overgrowth. The first three problems with implanted valves can be directly attributed to the design of the mechanical and biological prostheses and the fluid and solid mechanics during valve function. Thrombus deposition on the valvular surface and subsequent breakage of the thrombus to form emboli that can result in stroke or heart failure is still a major problem with MHV implants. Hence, patients with MHV implants need a long-term anticoagulant therapy that can lead to complications with bleeding. On the other hand, patients implanted with bioprostheses do not generally require anticoagulant therapy except immediately after surgery. Yet, leaflet tearing with structural disintegration results in the need for BHV implants to be replaced at about 10–12 years after implantation on the average. Due to the necessity of multiple

surgeries during a lifetime, the tissue valves are generally not implanted in younger patients. Patients who cannot tolerate or cannot be on long-term anticoagulant therapy are also candidates for the BHV.

Mortality is higher among patients after prosthetic valve replacement than among age-matched controls. Mortality rate is not significantly different between MHV and BHV implantation. In addition to the mortality rate and valve related morbidity, quality of life for the patient must be an important consideration in the choice of valve implantation and the quality of life is difficult to quantify. Individual patients may place different emphasis on mortality, freedom from reoperation, risk of thromboembolism and stroke, risk of anticoagulation-related hemorrhage, and lifestyle modification required with chronic anticoagulation. Patients may choose to accept the high probability of reoperation within 10–12 years with BHV in order to avoid long-term anticoagulation with MHV, whereas others may want to avoid the likelihood of reoperation (10).

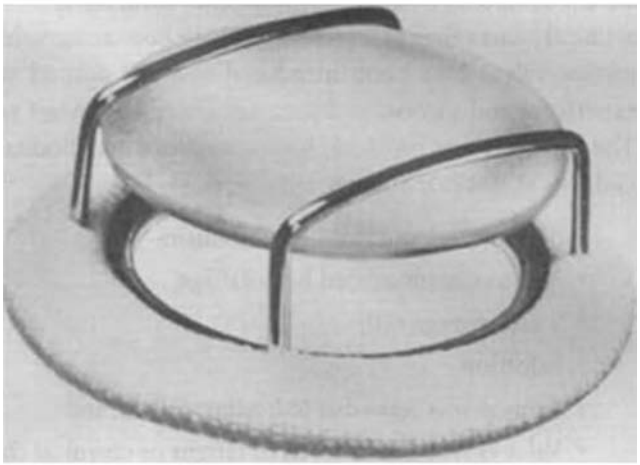
We will review the historical development of heart valve prostheses, functional evaluation of these valves in order to understand the relationship between the dynamics of the valve and the problems associated with the valve implants and our continuing efforts on the understanding of the problem and improvements in design.

Mechanical Heart Valves

Mechanical valves are made of blood compatible, nonbiological material, such as metals, ceramics, or polymers. The initial designs of mechanical valve prostheses were of the centrally occluding type with either a ball or disk employed as the moving occluder. The occluder passively responds to the difference in pressure in opening and closing of the valves. Starr-Edwards caged ball valve (Fig. 5a) was the first mechanical valve to be implanted in 1960 in the correct anatomical position as replacement for the diseased native mitral valve (11–14). The caged disk prostheses (Fig. 5b), in which a flat disk was employed as the occluder, were of a lower profile than the ball valves, and hence were thought to be advantageous especially as a replacement in the mitral position in order to minimize interference with the cardiac structures. However, increased flow separation and turbulence in the flow past the flat disk compared to that for the spherical ball occluder in the caged ball prostheses resulted in larger pressure drop across the valve with the caged disk design. An increased propensity for thrombus deposition in the recirculation region was also observed, and hence this design was not successful in spite of the low profile. The cage in the caged ball prostheses is made of a polished cobalt–chromium alloy and the ball is made of a silicone rubber that contains 2% by weight barium sulfate for radiopacity. The sewing ring contains silicone rubber insert under knitted composite polytetrafluoroethylene (PTFE: Teflon) and polypropylene cloth. With the centrally occluding design, the flow dynamics past the caged ball prostheses is vastly different from that of flow past native aortic or mitral valves.

Within the next two decades of 1970s and 1980s, valve prostheses with a tilting disk or bileaflet designs were introduced with significantly improved flow characteristics. The first tilting disk valve that was clinically used was a notched Teflon occluder that engaged in another pair of notches in the housing (15). The stepped occluder with the notches was not free to rotate. Clinical data soon indicated severe thrombus formation around the notches and wear of the Teflon disk leading to severe valvular regurgitation or disk embolization (16). A major improvement to this design was the introduction of hinge-less free-floating tilting disk valves in the Bjork–Shiley (17) and the Lillehei–Kaster valves. The Bjork–Shiley valve had a depression in the disk and two welded wire struts in the valve housing to retain the disk. The occluder tilted to the open and closed position and it was free to rotate around its center. The Bjork–Shiley valve housing was made from Stellite-21 with a Teflon sewing ring and a Delrin disk. Compared to the

(a)



(b)

Figure 5. Photographs of (a) Starr-Edwards caged ball valve (Edwards Lifesciences, LLC, Irvine, CA); and (b) a caged disk valve (Courtesy of Prof. Ajit Yoganathan of Georgia Institute of Technology, Atlanta, GA) as examples of early mechanical valve designs.

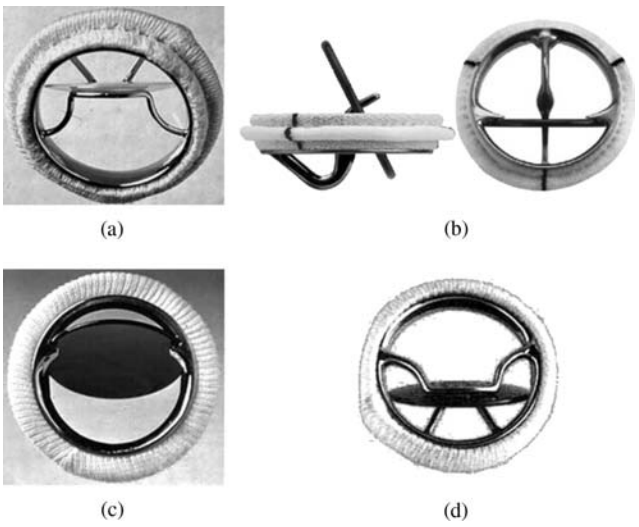


Figure 6. Tilting disk mechanical valve prostheses: (a) Bjork–Shiley valve; (b) Medtronic Hall valve (Medtronic, Inc., Minneapolis, MN; (c) Omni Carbon valve (MedicalCV Inc., Minneapolis, MN); (d) Sorin valve (Sorin Biomedica Cardio S.p.A., Via Crescentino, Italy).

centrally occluding caged ball valves, a large annulus diameter compared to the tissue annulus diameter in the tilting disk valve resulted in a very low pressure drop and thus energy loss in the flow across the valve in the open position. The disk opened to an angle of 60° or more, and hence the flow was more central. The free floating disk that rotated during the opening and closing phases prevented any build up of thrombus. However, the Delrin disk had a propensity for swelling during autoclaving that may compromise the proper functioning of the leaflets (18,19). The disk for the Lillehei–Kaster valve consisted of a graphite substrate coated with a 250 μm thick layer of a carbon–silicon alloy (Pyrolite). The pyrolytic carbon has proven to be a very durable and blood-compatible material for use in prosthetic heart valves and is the preferred material for the MHVs currently available for implants. The Bjork–Shiley valve also had the Delrin disk replaced with pyrolytic carbon disk shortly thereafter (Fig. 6a). The Medtronic Hall tilting disk valve (Fig. 6b) has a central, disk control strut. An aperture in the flat pyrolytic carbon disk affixes it to this central guiding strut and allows it to move downstream by ~2.0 mm. This translation improves the flow velocity between the orifice ring and the rim of the disk. The ring and strut combination is machined from a single piece of titanium for durability and the housing can be rotated within the knitted Teflon sewing ring for optimal orientation of the valve within the tissue annulus. The Omniscience valve was an evolution of the Lillehei–Kaster valve and the Omnicarbon valve (Fig. 6c) is the only tilting disk valve with the occluder and housing made of pyrolytic carbon. Sorin Carbocast tilting disk valve (Fig. 6d), made in Italy and available in countries outside United States, has the struts and the housing made in a single piece by a microcast process and thus eliminates the need for welding the struts to the housing. The cage for this valve is made of a chrome–cobalt alloy and coated with a carbon film. The tilting disk valves of the various manufacturers open to a range of angles varying from 60 to 85° and in the fully open position, the flow passes through the major and minor orifices. Some of the valve designs, such as the Bjork–Shiley valve, encountered unforeseen problems with structural failure due to further design modifications, and hence are currently not used for replacement of diseased native heart valves. However, some of these designs are still being used in the development of artificial heart and positive displacement left ventricular assist devices.

Another major change in the MHV design was the introduction of a bileaflet valve in the late 1970s. The St. Jude Medical bileaflet valve (Fig. 7a) incorporates two semicircular hinged pyrolytic carbon leaflets that open to an angle of 85° and the design is intended to provide minimal disturbance to flow. The housing and the leaflets of the bileaflet valve is made of pyrolytic carbon. Numerous other bileaflet designs have since been introduced into the market. Several design improvements have also been incorporated in the bileaflet valve models in order to improve their hemodynamic performance. The design improvements have included a decrease in thickness of the sewing cuff that allows the placement of a larger

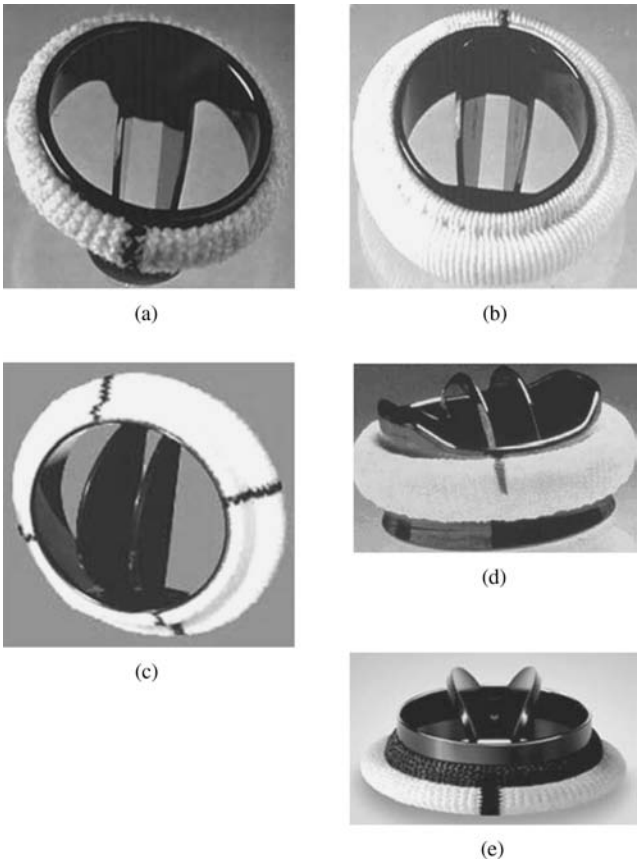


Figure 7. Bileaflet mechanical valve prostheses: (a) St. Jude valve (St. Jude Medical, Inc., St. Paul, MN); (b) Carbomedics valve (Carbomedics Inc., Austin Texas); (c) ATS valve (ATS Medical Inc., Minneapolis, MN); (d) On-X valve (Medical Carbon Research Institute, LLC, Austin, Texas); and (e) Sorin valve (Sorin Biomedica Cardio S.p.A., Via Crescentino, Italy).

housing within the cuff for a given tissue annulus diameter with the resulting hemodynamic improvement. Structural reinforcement of the housing has also allowed reducing its thickness that increases the internal orifice area for improved hemodynamics. The Carbomedics bileaflet valve (Fig. 7b) was introduced into the market in the 1990s with a recessed pivot design. The aortic version of this valve is designed to be implanted in the supraannular position enabling a larger size valve to be implanted with respect to the aortic annulus. More recent bileaflet valve designs available in the market include the ATS valve (Fig. 7c) with an open leaflet stop rather than the recessed hinges and the On-X bileaflet valve (Fig. 7d) that has a length to diameter ratio close to the native heart valves, a smoothed pivot recess allowing for the leaflet to open to 90°, a flared inlet for reducing flow disturbances, and a two point landing mechanism for smoother closing of the leaflets. The Sorin Bicarbon valve (Fig. 7e), marketed outside the United States, has curved leaflets, and hence increases the area of the central orifice. The pivots of this valve with two spherical surfaces enable the leaflet projections to roll against the surfaces rather than with the sliding action between the leaflet and the housing at the hinges.

Studies have shown that the bileaflet valves generally have a smaller pressure drop compared to the tilting disk valves, especially in the smaller sizes. However, there are several characteristic differences between the bileaflet and tilting disk valve designs that must be noted. The bileaflet valve designs include a hinge mechanism generally by introducing a recess in the housing in which a protrusion from the leaflets interacts during the opening and closing of the leaflets, or open pivots for the retention of the leaflets. On the other hand, the tilting disk valve designs do not have a hinge mechanism for retaining the occluder and the occluder is freefloating. The free-floating disk rotates as the valve opens and closes, and hence the stresses are distributed around the leaflets as opposed to the bileaflet designs. In spite of the advances in the MHV valves by the introduction of the tilting disk and bileaflet designs, design improvements aimed at enhancing the flow dynamics, and material selection, problems with thromboembolic complications and associated problems with bleeding (20) are still significant with the implanted valves and the possible relationship between the flow dynamics and initiation of thrombus will be discussed in detail later. An example of thrombus deposition and tissue ingrowth with an explanted MHV is shown in Fig. 8.

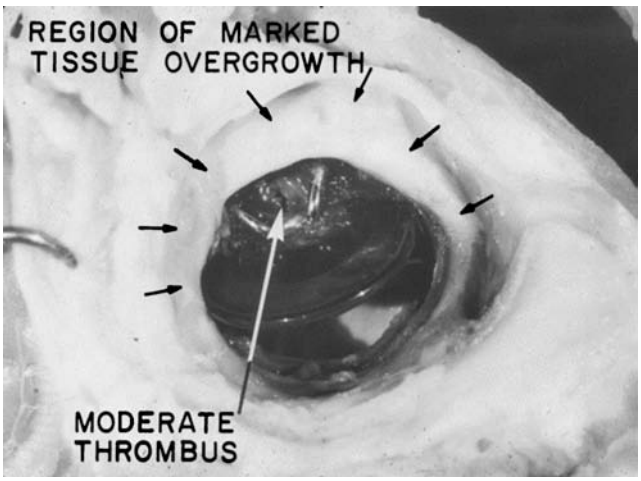


Figure 8. Photograph of an explanted mechanical heart valve prosthesis with thrombus deposition on the leaflet surface and tissue ingrowth (Courtesy of Prof. Ajit Yoganathan of Georgia Institute of Technology, Atlanta, GA.)

Bioprosthetic Valves

With the lack of availability of homograft valves as replacement of diseased valves, and as alternative to MHV that required long-term anticoagulant therapy, numerous attempts have been made in the use of various biological tissues as valvular replacement material. BHV made out of fascia lata (a layer of membrane that encases the thigh muscles) as well as human duramater tissue has been attempted. Fascia lata tissue was prone to deterioration, and hence unsuitable while the duramater tissue suffered from lack of availability for commercial manufacture in sufficient quantities. Harvested and preserved porcine aortic valve as well as BHV made of bovine pericardial tissue have been employed as replacement and have been available commercially for >30 years. The early clinical use of a xenograft (valve made from animal tissue) employed treatment of the leaflets with organic mercurial salts (21) or formaldehyde (22) to overcome the problems of rejection of foreign tissue by the body. Formaldehyde is used to fix and preserve the tissue in the excised state by histologists and results in shrinkage as well as stiffening of the tissue. Formaldehyde treated valves suffered from durability problems with 60% failure rates at 2 years after implantation. Subsequently, it was determined that the durability of the tissue cross-links was important in maintaining the structural integrity of the leaflets and gluteraldehyde was employed as the preservation fluid (23). Glutaraldehyde also reduces the antigenicity of the foreign tissue, and hence can be implanted without significant immunological reaction.

Porcine aortic valves are excised from pigs with the aortic root and shipped to the manufacturer in chilled saline solution. Support stents, configured as three upright wide posts with a circular base, is manufactured out of metal or plastic material in various sizes and covered in a fabric. A sewing flange or ring is attached to the base of the covered stent and used to suture the prostheses in place during implantation. The valve is cleaned, trimmed, fitted, and sewn to the appropriate size cloth covered stent. The stented valve is fixed in gluteraldehyde with the valve in the closed position. Glutaraldehyde solution with concentrations ranging from 0.2 to 0.625% is used in the fixation process at pressures of <4 mmHg (0.53 kPa) to maintain the valve in the closed position. The low pressure fixation maintains the microstructure of collagen. The first glutaraldehyde-treated porcine aortic valve prosthesis mounted on metallic stent was implanted in 1969 (24). The metallic frame was soon replaced by a flexible stent on a rigid base ring; the Hancock Porcine Xenograft was commercially introduced in 1970. The stent in this valve is made of polypropylene with stainless steel radiopaque marker,

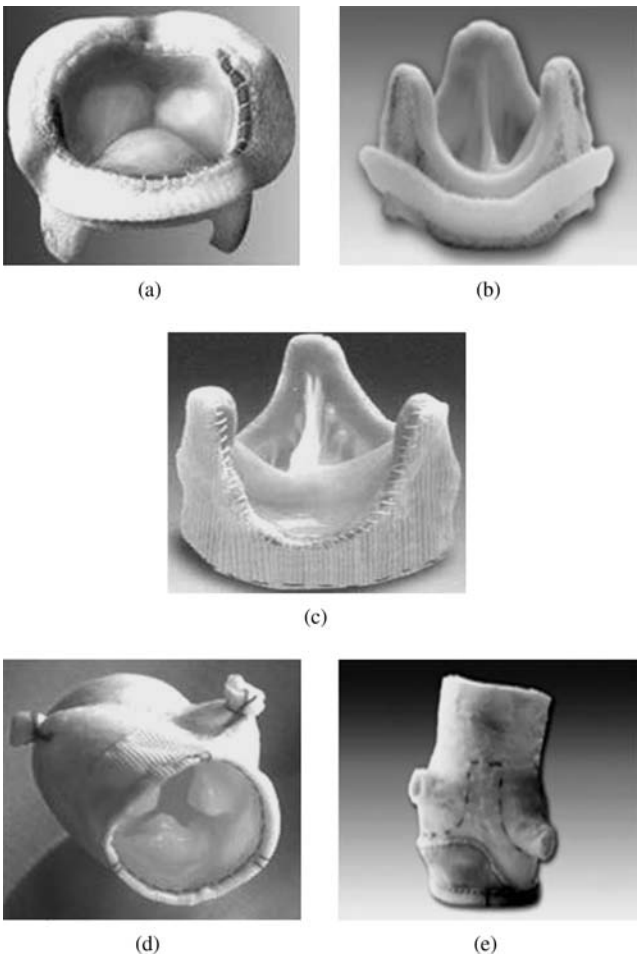


Figure 9. Porcine bioprosthetic valve prostheses: (a) Hancock II valve (Medtronic Inc., Minneapolis, MN); (b) Carpentier–Edwards valve (Edwards Lifesciences, LLC, Irvine, CA); (c) Toronto stentless valve (St. Jude Medical, Inc., St. Paul, MN); (d) Medtronic Freestyle stentless valve (Medtronic Inc., Minneapolis, MN); and (e) Edwards Prima Plus stentless valve (Edwards Lifesciences, LLC, Irvine, CA).

sewing ring made of silicone rubber foam fiber and polyester used as the cloth covering. Hancock Modified Orifice valve (Fig. 9a) was introduced in 1977 as a refinement of the earlier valve. The right coronary cusp of the pig's aortic valve is a continuation of the septal muscle, and hence stiffer. In the modified orifice valve, the right coronary cusp is replaced with a non-coronary cusp of comparable size from another valve. The Carpentier–Edwards Porcine valve (Fig. 9b) also employs a totally flexible support frame. The Hancock II and the Carpentier–Edwards supra-annular porcine bioprostheses employed modified preservation techniques in which the porcine tissue is initially fixed at 1.5 mmHg (0.2 kPa), and then at high pressure in order to improve the preservation of the valve geometry. The supra-annular valve is designed to be implanted on top of the aortic annulus while aligning the internal diameter of the valve to the patient's annulus and this technique allows implantation of a larger valve for any given annular size. These valves are also treated with antimineralization solution such as sodium dodecyl sulfate (SDS) in order to reduce calcification.

Another major innovation in the bioprosthetic valve design was the introduction of stentless bioprostheses. In the stented bioprostheses, the regions of stress concentration are observed at the leaflet–stent junction and the stentless valve design is intended to avoid such regions prone to failure. The absence of the supporting stents also results in less obstruction to flow, and hence should improve the hemodynamics across the valves. Due to the lack of stents, larger size valve can be implanted for a given aortic orifice to improve the hemodynamics. Stentless porcine bioprostheses are only approved for aortic valve replacement in the United States. *In vitro* studies have shown improved hemodynamic performance with the stentless designs in the mitral position, but questions remain about the durability of these valves, and the implantation techniques are also complex in the mitral position. Examples of stentless bioprostheses currently available in the United States include: St. Jude Medical Toronto SPV (Fig. 9c); Medtronic Freestyle (Fig. 9d); and Edwards Prima (Fig. 9e) valves. The Edwards Prima prosthesis is the pig's aortic valve with a preserved aortic root, with a woven polyester cloth sewn around the inflow opening to provide additional support and with features that make it easier to implant. The other stentless valve designs are also porcine aortic valves with the intact aortic root and specific preservation techniques in order to improve the hemodynamic characteristics and durability after implantation. In the Toronto SPV valve, a polyester cloth covering around the prosthesis separates the xenograft from the aortic wall of the host, making it easier for handling and suturing, and also promotes tissue ingrowth.

Both stented and stentless porcine tissue valves are from the pig's native aortic valve, and hence individual leaflets need not be manufactured. In order to have sufficient quantities of these valves in various sizes available for implant, a facility to harvest adequate quantities of these valves become necessary. As an alternative, pericardial valves are made by forming the three leaflets from the bovine pericardial tissue, and hence valves of various sizes can be made. In the pericardial valves, bovine pericardial sac is harvested and shipped in chilled saline solution. At the manufacturing site, the tissue is debrided of fatty deposits and trimmed to remove nonusable areas before the tissue is fixed in glutaraldehyde. After fixation, leaflets are cut out from the selected areas of the pericardial tissue and sewn to the cloth-covered stent in such a fashion to obtain coapting and fully sealing cusps. Since the valve is made in the shape of the native human aortic valve, the hemodynamic characteristics were also superior to the porcine valves in comparable sizes. The Ionescu-Shiley pericardial valve introduced into the market in the 1970s was discontinued within a decade due to problems associated with calcification and decreased durability. For this reason, pericardial valve were not marketed by the valve companies for several years. With advances in tissue processing and valve manufacturing technology, pericardial valves were reintroduced into the commercial market in the 1990s. The Edwards Lifesciences introduced the Carpentier-Edwards PERIMOUNT Bioprostheses

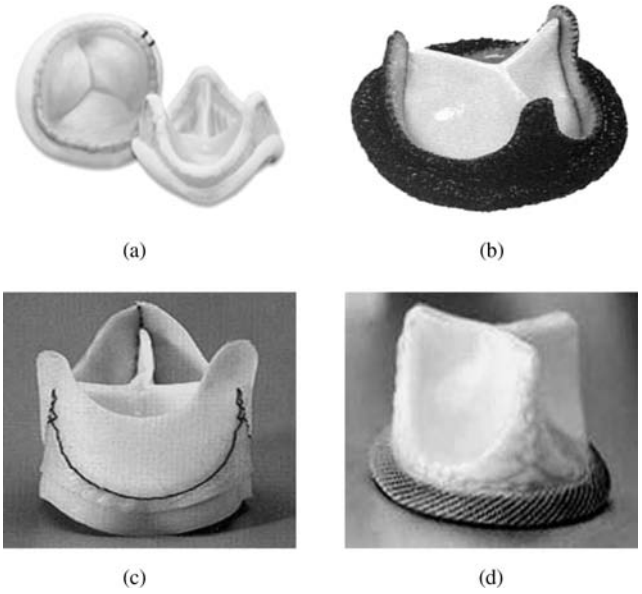


Figure 10. Pericardial bioprosthetic valve prostheses: (a) Carpentier-Edwards valve (Edwards Lifesciences, LLC, Irvine, CA), (b) Sorin Pericarbon valve (Sorin Biomedica Cardio S.p.A., Via Crescentino, Italy); (c) Sorin Pericarbon stentless valve (Sorin Biomedica Cardio S.p.A., Via Crescentino, Italy); and (d) Mitroflow Aortic pericardial valve (Sorin Group Canada, Inc., Mitroflow Division, Burnaby, B.C., Canada.)

(Fig. 10a) in the early 1990s. The pericardial tissue in this valve is mounted on a lightweight frame that is covered with porous, knitted PTFE material. A sewing ring made of molded silicone rubber covered with PTFE cloth is incorporated to suture the valve in place. Sorin pericardial valve (Fig. 10b), Sorin stentless pericardial valve (Fig. 10c) and the Mitroflow aortic pericardial valve (Fig. 10d) are available in markets outside the United States.

FUNCTIONAL CHARACTERISTICS

From the functional point of view, the implanted valve prostheses should open with minimal transvalvular pressure drop and have minimal energy loss in blood flow across the valve, and the valve must close efficiently with minimal regurgitation. The valve should mimic the central flow characteristics that are observed with native human heart valves with minimally induced fluid dynamic stresses on the formed elements in blood. The flow characteristics across the implants should also avoid regions of stasis or flow stagnation where thrombus deposition and growth can be enhanced. *In vivo* measurements, *in vitro* experimental studies, and computational simulations have been used over the past 40 years in order to assess the functional characteristics of the mechanical and bioprosthetic heart valves. The information gained from such studies have been exploited to improve the design of the valves in order to improve the performance characteristics of the valves and also increase the durability in order to provide a "normal" life style for the patient with prosthetic valve implants.

Pressure Drop and Effective Orifice Area

In vivo measurement of pressure drop requires placing pressure transducers inserted via a catheter both on the inflow and outflow side of the valve, and computing the pressure drop during the phase when the valve is open. For the aortic valve, the pressure transducers are placed in the left ventricular outflow tract and in the ascending aorta. The peak or the average pressure drop during the forward flow phase is computed from the recorded data. To avoid the invasive technique of catheterization, the fluid mechanics principle can also be applied to estimate the pressure drop across the valve in the aortic position. The average velocity, V in $\mathrm{m \cdot s^{-1}}$, in the ascending aortic cross-section is measured noninvasively and the pressure drop, Δp expressed in millimeters of mercury can be computed using the equation

$$\Delta p \cong 4V^2 \tag{2}$$

Using this simplified equation and noninvasive measurement of the aortic root velocity using the Doppler technique, the pressure drop across the aortic valve can be computed. With the availability of several designs of MHV and BHV, it is desirable to compare the pressure drop and regurgitation for the various valve designs. In the case of development of a new valve design, United States Federal Drug Administration (FDA) requires that these quantities measured *in vitro* for the new designs are compared with currently approved valves in the market. *In vitro* comparisons are performed in pulse duplicators that mimic the physiological pulsatile flow in the human circulation. One of the initial designs of a pulse duplicator for valve testing was that of Wieting (25) that consisted of a closed-loop flow system that is actuated by a pneumatic pump to initiate pulsatile flow through the mitral and aortic valves in their respective flow chambers. Pressure transducers were inserted through taps in the flow chambers on the inflow and outflow sides to measure the pressure drop across the valves. The fluid used in such *in vitro* experimental studies, referred to as the blood-analogue fluid, is designed to replicate the density ($1060\ \mathrm{kg \cdot m^{-3}}$) and viscosity coefficient (0.035 P or 35×10^{-4} Pa·s) of whole human blood. A glycerol solution (35–40% glycerin in water) has been generally used as the blood analog fluid in these studies. Prosthetic valves are made in various sizes (specified in sewing ring diameter magnitude). In comparing the pressure drop data for the various valve designs, proper comparison can be made on data only with comparable valve sizes. Since the flow across the valve during the forward flow phase is not laminar, the pressure drop has a nonlinear relationship with flow rate. Hence, pressure drop is measured for a range of flow rates and the pressure drop data is presented as a function of flow rate. Numerous studies comparing the pressure drop data for the various mechanical and bioprosthetic valves have been reported in the literature. Typical pressure drop comparisons for MHV and BHV (of nominal size of 25 mm) are shown in Fig. 11 (14). As can be observed, stented porcine tissue valves have the higher pressure drop, and hence are considered to be stenotic, especially in smaller valve sizes. The advantage of the stentless bioprostheses design is

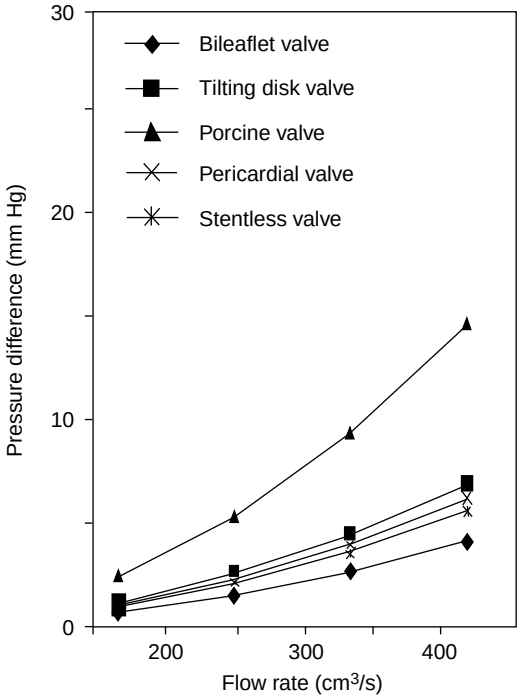


Figure 11. Typical plots for the pressure drop as a function of flow rate for the various mechanical and bioprosthetic valve prostheses (Courtesy of Prof. Ajit Yoganathan of Georgia Institute of Technology, Atlanta, GA.)

obvious especially in smaller sizes since the flow orifice will be larger due to the absence of supporting stents. Supra-annular design also permits the implantation of a larger sized valve for a given annulus orifice, thus providing a smaller pressure drop and energy loss. Smaller pressure drops across the valve prostheses will result in a reduced workload of the left ventricle as the pump. Gorlin equation, described in Eq. 1, has also been employed to compute the effective orifice area for the various valve designs (14). Generally, the pericardial and bileaflet valve designs have the largest effective orifice area, followed by the tilting disk, and porcine valves, with the caged ball valve exhibiting the smallest effective orifice area for a given sewing ring diameter. Valves with the larger EOA correspond to a smaller pressure drop and energy loss in flow across the valve. Performance index (PI), computed as the ratio of effective orifice area to the sewing ring area is also used for comparison of the various valve designs. Table 1 includes data on the EOA and PI for the various MHV and BHV with a 27 mm tissue annulus diameter from *in vitro* experimental studies. It can be observed that the centrally occluding caged ball valve and stented porcine valves have lower values of PI where as the tilting disk, bileaflet, pericardial valves, and stentless tissue valves have higher values indicating improved hemodynamics for the same values of the tissue annulus diameter.

Regurgitation

In discussing the flow dynamics with native heart valves, it was observed that the anatomy and the fluid dynamics enable the leaflets to close efficiently with minimal amount

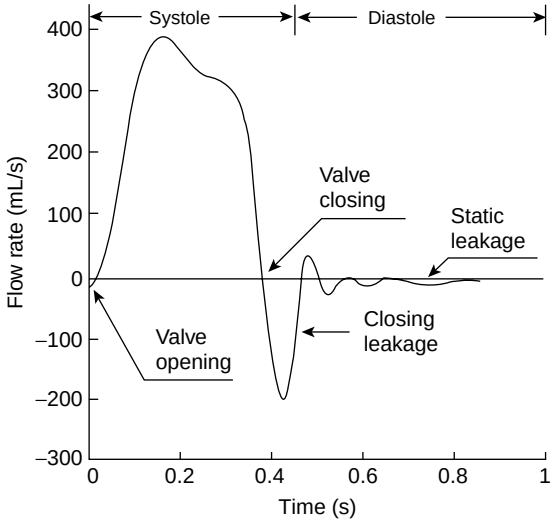


Figure 12. Flow rate across mechanical valve prosthesis in the aortic position obtained in a pulse duplicator *in vitro* measured with an electromagnetic flow meter.

of regurgitation. On the other hand, the adverse pressure gradient at the end of the forward flow motion induces the occluders to move toward closure and all prosthetic valves exhibit a finite amount of regurgitation. Figure 12 shows a typical flow rate versus time curve obtained from an electromagnetic flow meter recording obtained *in vitro* in a pulse duplicator with a mechanical valve in the aortic position. As can be observed, a certain volume of reverse flow is observed as the valve closes and is termed as closing leakage. The closing leakage is related to the geometry of the valve and the closing dynamics. The rigid occluders in the mechanical valves also prevent the formation of a tight seal between the occluder and the seating ring when the valve is closed. With the tilting disk and bileaflet valves, a small gap between the leaflet edge and the valve housing is also introduced in order to provide a continuous wash-out of blood in the hope of preventing any thrombus deposition. Hence, even when the valve is fully closed, a small volume of blood is continuously leaking and is termed the static leakage. Percent regurgitation is defined as the ratio of the leakage volume over the net forward flow volume expressed as a percentage. Percent regurgitation can be computed by recording the flow rate versus time curve in an *in vitro* experimental set up and measuring the area under the forward and reverse flow phases from the data. These can be compared for the various size valves of the same model and also for comparison across the various valve models. Table 1 shows typical data of regurgitant volumes measured *in vitro* under physiological pulsatile flow in a pulse duplicator. The BHV designs result in more efficient leaflet closure with relatively small regurgitant volumes followed by the caged ball valve design. The magnitudes of the regurgitant volumes for the tilting disk and bileaflet valves are relatively larger and comparable to each other.

Quantitative measurement of percent regurgitation *in vivo* with both incompetent native valves or with prostheses has not been successful, even though attempts have

Table 1. Comparison of Effective Orifice Area (EOA)[a], Performance Index (PI), Regurgitation Volume, and Peak Turbulent Shear Stresses for the Various Models of Commercially Available Valve Prostheses

Valve Type	EOA[b] cm^2	PI	Reg. Vol., cm^3/beat	Peak Turb. SS, Pa[c]
Caged ball	1.75	0.30	5.5	185
Tilting Disk[d]	3.49	0.61	9.4	180
Bileaflet[d]	3.92	0.68	9.15	194
Porcine (Stented)[d]	2.30	0.40	< 2	298
Pericardial (Stented)[d]	3.70	0.64	< 3	100
Stentless BHV	3.75	0.65	< 4	NA[e]

[a]EOA = Effective orifice area computed by the application of Gorlin's equation (4).
[b]Values compiled from Yoganathan (14).
[c]Turbulent stresses were measured at variable distances from the valve seat.
[d]Values reported are mean values from several valve models of the same type. Data reported are for 27-mm tissue annulus diameter size of the valves with measurements obtained *in vitro* in a pulse duplicator with the heart rate of 70 bpm and a cardiac output of 5.0 $L \cdot min^{-1}$.
[e]Not available = NA.

been made to employ fluid mechanical theories for regurgitant flow in order to estimate the leakage volume. Turbulent jet theory and proximal flow convergence theory have been employed in an attempt to measure the regurgitant glow volume quantitatively (26,27). However, *in vivo* application has not been successful due to the restrictive assumptions of steady flow and alterations due to impingement of the jet on the ventricular wall in the theoretical considerations as well as lack of *in vivo* validation.

Dynamics of Valve Function

As discussed earlier, significant problems still exist with the implantation of heart valve prostheses in patients with disease of native heart valves. These include thrombus initiation and subsequent embolic complications with MHV implantation. The thromboembolic rates with MHV have been estimated at 2%/patient year (28). Structural disintegration and tearing of leaflets are the major complications with BHV requiring reoperation in ~10–12 years after implantation. Flow past healthy native valves are central with minimal flow disturbances and fluid induced stresses and it can be anticipated that the fluid dynamics past the mechanical valve prostheses will be drastically different from those of native heart valves. Flow induced stresses with MHV function have long been implicated with hemolysis and activation of platelets that may trigger thrombus initiation. Regions of stress concentration on the leaflets during the opening and closing phases have been implicated on structural alterations of collagen fibers resulting in leaflet tears with BHV. Detailed functional analysis of implanted valve prostheses *in vivo* is impractical. Limited attempts have been made in the measurement of velocity profiles distal to the valve prostheses in the aortic position with hot film anemometry (29). Doppler and MR phased velocity mapping techniques have also been used to measure the velocity profiles distal to heart valves (30–32). However, detailed velocity measurements very close to the leaflets and housing of prosthetic valves are not possible *in vivo*, and hence *in vitro* experimental studies and computational fluid dynamic simulation are necessary for the same. Limited *in vivo* studies in animal models have also been employed to describe the complex leaflet motion with native aortic valves (33–37). *In vitro* studies and computer simulations are also necessary for a detailed analysis of stress distribution in the leaflets of native valves and bioprostheses during the opening and closing phases of the valve function and to determine its relationship with failure of the leaflets.

Flow Dynamics Past Mechanical Valves

With the assumption that the deposition of thrombi in MHV implants is related to the flow induced stresses, studies are continuing to date on the deterministic relationship between fluid induced stresses and damage to formed elements in blood. Subjecting blood cells to precise flow fields and assessing the destruction or activation (of platelets), magnitudes of turbulent stresses beyond which damage can be expected has been established. In addition to the magnitude of the flow induced stresses, the time for which the blood elements are exposed to the stresses also need to be considered in assessing the destruction or activation of the platelets. Nevaril et al. (38) reported that blood cells can be hemolyzed with shear stresses of the order of 150–400 Pa. In the presence of foreign surfaces, the threshold for red blood cell damage reduces to ~1–10 Pa (39). Sublethal damage to red blood cells have also been reported at turbulent shear stress levels of about 50 Pa (40). Shear induced platelet activation and aggregation is observed to be a function of both magnitude and duration of shear stresses. The larger the magnitude of the shear stress, the shorter is the duration to which platelets are subjected to the shear before they get activated. Platelets have been shown to be activated with 10–50 Pa of shear stresses with a duration of the order of 300 ms (41). Platelet damage also increases linearly with time of exposure when subjected to constant magnitudes of shear (42).

Hence, it is of interest to determine the level of wall shear and turbulent shear stresses in flow past valve prostheses as factors causing initiation and deposition of thrombus.

For the first two to three decades after the implantation of the first mechanical valve, investigations concentrated on the flow dynamics past the valves during the forward flow phase and measurements of velocity profiles, regions of stasis and recirculation, high wall shear and bulk turbulent shear stresses. Wieting (25) employed a pulse duplicator and flow visualization studies using illuminated neutrally buoyant particles in order to qualitatively describe the nature of flow past the prosthetic valves.

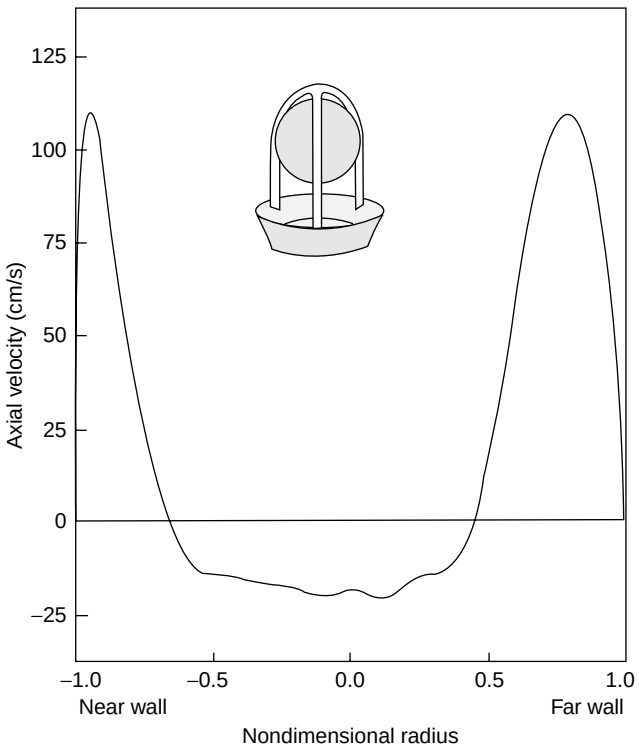


Figure 13. Velocity profile measured distal to a caged ball valve in the aortic position *in vitro* in a pulse duplicator using laser Doppler anemometry technique.

Yoganathan (43) employed laser Doppler velocimetry (LDV) technique to measure the velocity profiles and turbulent shear stresses under steady flow past the valve prostheses. Since then numerous detailed studies have been reported in the literature on the detailed measurement of velocity profiles and turbulent shear stresses distal to the prostheses under physiological pulsatile flow (13,14).

Figure 13 shows the velocity profile distal to a caged ball valve during the peak forward flow phase measured under physiological pulsatile flow *in vitro* (44). Jet-like flow is observed around the circumference that is separated by the ball and high turbulent stresses were measured at the edge of the jet. A wake is observed behind the ball with slow moving fluid. With the caged ball valve, higher incidences of thrombus deposition have been observed at the top of the cage and correspond to the slow moving fluid in this region behind the wake of the ball. With the tilting disk valves in the fully open position, the blood flows through the major and minor orifices as shown in Fig. 14 where the velocity profile during peak forward flow phase is once again depicted (44). Two jets are formed corresponding to the two orifices with the major orifice jet having larger velocity magnitudes. The amount of blood flow through the major and minor orifices will depend on the angle of opening of the occluder as well as the geometry. A region of reverse flow is also observed adjacent to the valve housing in the minor orifice. The velocity profile measured distal to the leaflet along the major flow orifice in the perpendicular orientation is also included in the figure.

Velocity profiles with three jets corresponding to the central orifice and two peripheral orifices are observed with the bileaflet valve as shown in Fig. 15 (45). The velocity profile along the central orifice in the perpendicular orientation is also included in this figure. Regions of flow reversals near the valve housing are also observed in the figure. Typical magnitudes of turbulent shear stresses measured in the various positions distal to the MHV under pulsatile flow conditions are also included in Table 1. It can be

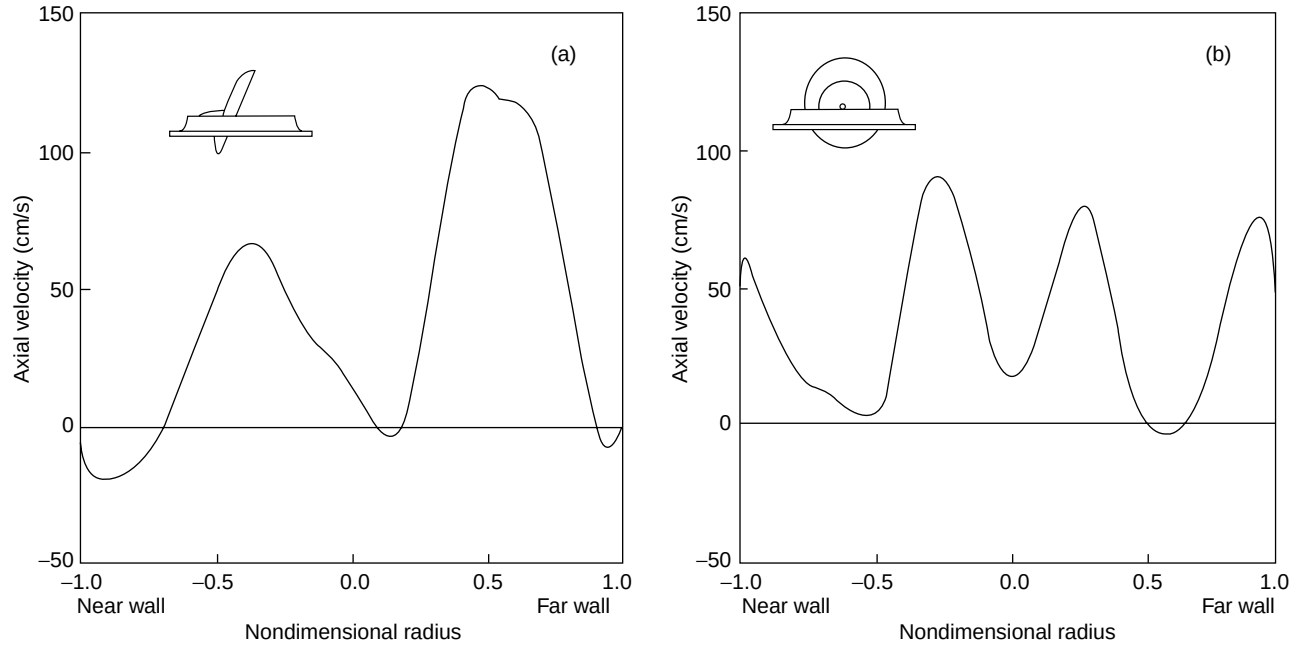


Figure 14. Velocity profile measured distal to a tilting disk valve in the aortic position *in vitro* in a pulse duplicator using laser Doppler anemometry technique.

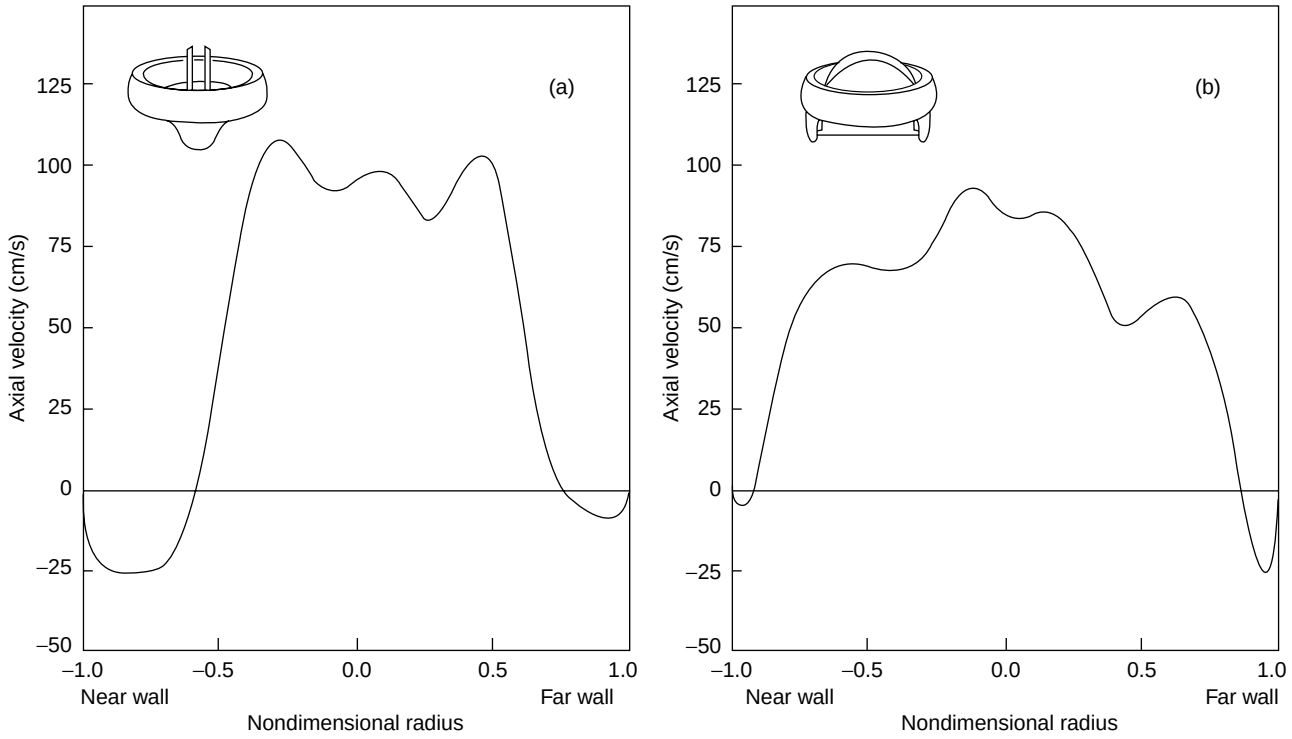


Figure 15. Velocity profile measured distal to a bileaflet valve in the aortic position *in vitro* in a pulse duplicator using laser Doppler anemometry technique.

observed that the measured bulk turbulent stresses are large enough to cause hemolysis and platelet activation that can be related to thrombus deposition with MHV. Thrombus deposition is generally observed on the leaflets and the valve housing with the tilting disk valves and also in the hinge region in the case of bileaflet valves.

More recently, it has been suggested that the relatively high turbulent stresses observed during the forward flow phase may not necessarily be the only reason for problems associated with MHV implantation. High turbulent stresses that may damage the formed elements occur in bulk flow distal to and moving away from the valve during the forward flow phase. The activated platelets will need to go through the systemic and pulmonic circulation before it will get deposited once again in the vicinity of the housing in the case of tilting disk and bileaflet valves. Several other experiences with mechanical valve prostheses designs have also indicated the importance of the valve dynamics during the closing phase to be more important for structural integrity and also in the initiation of thrombus. Medtronic Parallel valve design was introduced in the European market in the 1990s with the two leaflets opening to 90° in the fully open position. *In vitro* studies suggested that the fluid dynamics past this valve in the forward flow phase is superior or at least comparable to the currently available bileaflet valves. However, soon after human implantation trials in Europe, increased incidences of thrombus deposition was observed with this valve model, and hence it was withdrawn from clinical trials.

Another example is a design change in the tilting disk valve resulting in major changes in valve dynamics that resulted in structural failure in a small percentage of implanted valves. In an effort to increase the flow through the minor orifice with an aim of preventing thrombus deposition, the flat disk geometry of the original Bjork–Shiley valve was changed to a curved geometry in the Bjork–Shiley convexo-concave valve. Even though this design change resulted in improved forward flow hemodynamics, this change resulted in alterations in the dynamics of valve closure with the leaflet overrotating and subjecting the outlet strut to additional loading (46). In a small percentage of valves particularly in the mitral position, single leg separation followed by outlet strut fracture resulted in leaflet escape, and hence this valve was withdrawn from the market. These developments also suggest the importance of understanding the mechanics of valve function throughout the cardiac cycle with any mechanical valve designs. In addition, structural failure and leaflet escape was reported with the implantation of a newly introduced bileaflet valve (Edwards-Duromedics) that resulted in the withdrawal of the valve from the market (47,48). The structural failure was thought to be due to pitting and erosion of the valve structures due to cavitation damage on the pyrolytic carbon (49). These reports also spurred a number of investigations on the closing dynamics and the potential for the mechanical valves to cavitate during the closing phase.

The occluders in the mechanical valves move toward closure with the onset of adverse pressure gradients, and the time taken to move from the fully open to the fully closed position is ∼30 ms. Toward the end of the leaflet closure, the leaflet edge moves with a velocity of ∼3–4 $m \cdot s^{-1}$

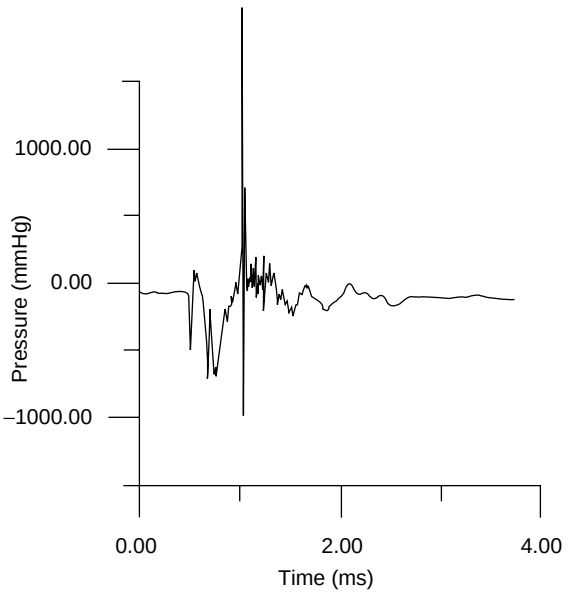


Figure 16. Typical negative pressure transients recorded at the instant of mechanical heart valve closure (in the mitral position) with the pressure transducer placed very close to the leaflet on the inflow side (atrial side) from *in vitro* experiments.

(50–53) and comes to a sudden stop as it impacts the seating lip. This produces a water hammer effect with large positive pressure transient on the outflow side (left ventricle in the mitral position and aorta in the aortic position) and a large negative pressure transient on the inflow side (left atrium in the mitral position and the left ventricular outflow tract in the aortic position). Several *in vitro* studies have recorded negative pressure transients (54) with magnitudes below that of the vapor pressure for blood (ca. −713 mmHg or −94.8 kPa) and cavitation bubbles have also been visualized at the edge of the leaflets (53–58) in the region corresponding to large negative pressure transients and where the linear velocity of the leaflet edge will be the largest. Figure 16 depicts measured negative pressure transients with the pressure transducer placed near the leaflet edge on the inflow side (atrial side) of the leaflet of a tilting disk valve at the instant of valve closure from *in vitro* experiments. Note that structural failure due to cavitation type of damage has been reported with only one model of the bileaflet valve and there are no other reports of pitting and erosion on the valve material reported with implanted mechanical valves. It is also not possible to visualize cavitation bubbles *in vivo* with implanted mechanical valves. However, *potential* for mechanical valves to cavitate has been demonstrated in an animal model with the recording of negative pressure transients, similar to those measured *in vitro*, in the left atrium in the vicinity of the implanted mechanical valves in the mitral position (59,60). The actual mechanism of cavitation bubble formation, whether due to the negative pressure magnitudes below the vapor pressure for blood or due to strong vortices forming in the atrial chamber providing additional pressure reductions, is still being debated. It has also been suggested that vortex cavitation bubbles forming away from the valve surfaces, can trap the dissolved gas from blood and form stable gas bubbles that travel with blood to the circulation and induce neurological deficit due to the gas emboli (61). Number of attempts has also been reported on the detection of the induced cavitation *in vivo* with implanted mechanical valves from acoustic signals (62–64). Another aspect of MHV cavitation that has been neither fully understood nor fully investigated is the development of stable bubbles, found by microembolic signals (MES) or high intensity transient signals (HITS) during and post-MHV implantation. *In vitro* studies have shown the development of stable bubbles (HITS) in an artificial heart and closing dynamics experimental models, and are affected by the concentration of CO_2(65–67). *In vivo*, HITS have been visualized during and post-MHV implantation through transcranial Doppler ultrasound (68,69). These events have been implicated as a cause of strokes and neurological deficits. Further evidence has shown that these HITS are in fact, gaseous, and not solid. Patients placed on pure O_2 after MHV implantation showed a large decrease in the number of HITS recorded, when compared to patients on normal air (70). These stable bubbles are believed to develop when gaseous nuclei that are present in blood, flow into low pressure regions associated with valve closure. As the valve closes and rebounds inducing vaporous cavitation, gas diffuses into the nuclei enlarging the bubble. When the pressure recovers and the vapor collapses, the bubble dynamics and local fluid mechanics prevent the gas from diffusing back into solution causing the bubble to stabilize and allowing it to flow freely in the vasculature. There is some discussion as to which gas stabilizes the nuclei. Both N_2 and CO_2 have been suggested as the link to MES/HITS/stable bubble formation (71), but there has yet to be concrete proof indicating which one does.

Large negative pressure transients occur due to the rigidity of the occluder and negative pressures do not occur at the instant of valve closure in the case of bioprosthetic valves (59). *In vitro* measurements with a tilting disk valve design employing a flexible occluder being implanted in India has also demonstrated that large negative pressures do not develop in such designs because the leaflets deform at the instant of valve closure and absorb part of the energy (12,59).

Irrespective of the formation of cavitation bubbles and subsequent collapse with implanted mechanical valves, the flow induced stresses during the valve closing phase has been suggested as of sufficient magnitude to induce platelet activation and initiation of thrombus. Even if the negative pressure transients do not reach magnitudes below the vapor pressure for blood, the large positive and negative pressure transients on the outflow and inflow sides of the valve at the instant of valve closure can induce high velocity flows through the gap between the leaflet and the housing, in the central gap between the two leaflets in the bileaflet valve, and also through the hinge region. The wall shear stress in the clearance region have been computed to be relatively high, even though present only for a fraction of a second. They induce platelet activation in the region where thrombus deposition is observed with mechanical valves (72). Relatively

high turbulent shear stresses have also been reported from *in vitro* studies distal to the hinge region of bileaflet valves during the valve closing phase (73,74) indicating the presence of high fluid induced regions near the leaflet edges during the valve closing phase that may be a significant contributor for thrombus initiation.

Most of the studies described above for the measurement of velocity profiles and turbulent stresses employed the laser Doppler velocimetry (LDV) technique. This is a point-velocity measurement technique, and hence measurement of the complete three-dimensional (3D) velocity profile distal to the valve under unsteady flows is tedious and time consuming. On the other hand, particle image velocimetry (PIV) technique has the ability to capture the whole flow field information in a relatively shorter time. Lim et al.employed the PIV technique to study the flow field distal to prosthetic heart valves in steady (75) and pulsatile (76) flow conditions. Along with the description of the flow field at the aortic root that was employed to identify the regions of flow disturbances, they also presented the results of turbulent Reynolds stress and turbulent intensity distributions. Under pulsatile flow conditions distal to a BHV, the velocity vector fields and Reynolds stress mappings at different time steps were used to estimate the damage of shear induced damage to formed elements of blood (76). Browne et al. (77) and Castellini et al. (78) compared the LDV and PIV techniques for the measurement of flow past MHV. Both these works conclude that PIV has the advantage of describing the detailed flow field distal to the valve prostheses in a relatively short time, but LDV technique affords more accurate results in the measurement of turbulent stresses. Regurgitant flow fields and the details of the vortical flow, a potential low pressure field for cavitation initiation have also been measured employing PIV techniques (79,80) and with a combination of PIV and LDV techniques (81).

Bioprosthetic Valve Dynamics

Measurement of velocity profiles and turbulent stresses from *in vitro* tests in pulse duplicators past BHV have also been reported in the literature (82). Figure 17 depicts typical velocity profiles past a porcine bioprosthesis under normal physiological flow simulation (82). With the porcine valve, a jet-like flow is observed during the peak forward flow phase with high turbulent shear stresses at the edge of the jet. With the pericardial valves (82), the peak velocity magnitudes in the jet-like flow during the peak forward flow phase were smaller than those for the porcine valves in comparable sizes (Fig. 18). It can be observed from Table 1 that the peak turbulent stresses are also smaller in the pericardial valves with geometry closer to the native aortic valves compared to that of the porcine prostheses. It should be noted that the magnitudes of turbulent stresses with the BHV also exceed those values suggested for activation of platelets. However, the leaflets of the BHV are treated biological tissue rather than artificial surfaces. Long-term anticoagulant therapy is generally not required with the implantation of bioprostheses since thrombus initiation is not a significant problem with these valves.

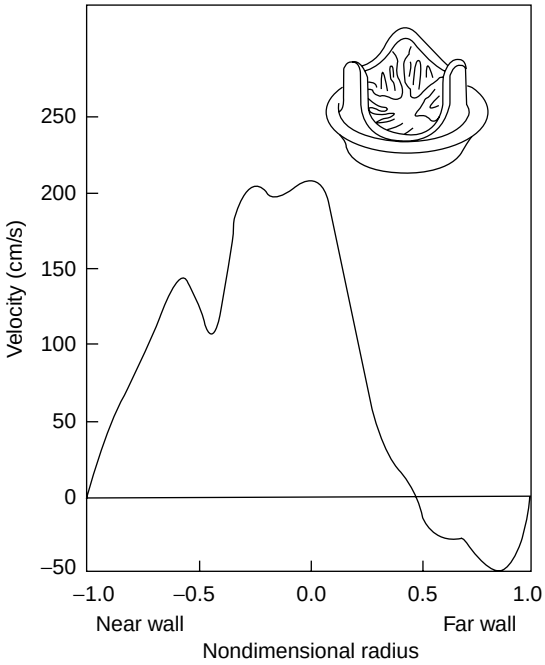


Figure 17. Velocity profile measured distal to a porcine bioprosthetic valve in the aortic position *in vitro* in a pulse duplicator using laser Doppler anemometry technique.

On the other hand, these valves fail after an average of 10–12 years of implantation and replacement surgery is required with leaflet failure. A number of studies have been reported on the analysis of the complex leaflet motion during the valve function in order to determine a causative relationship between regions of high stress concentration

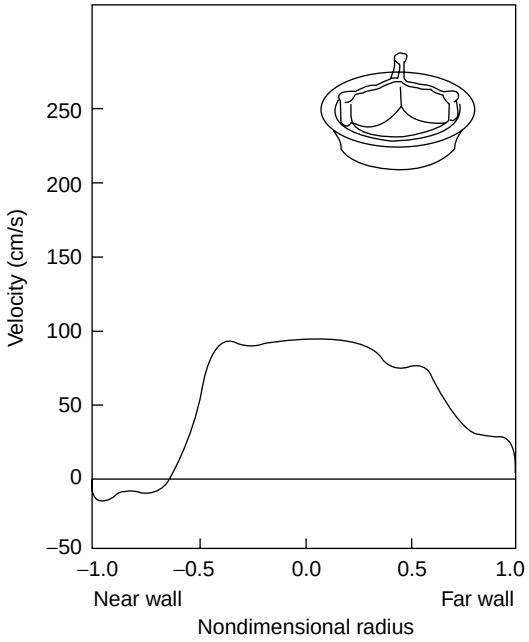


Figure 18. Velocity profile measured distal to a pericardial bioprosthetic valve in the aortic position *in vitro* in a pulse duplicator using laser Doppler anemometry technique.

on the leaflets and its attachment sites and structural failure. Based on the analysis of native aortic leaflet motion *in vivo* with the use of radiopaque markets, it has been reported that the design of the native aortic leaflets affords minimal stresses on the leaflets (33,35). On the other hand, stress analysis on the BHV valve leaflets *in vivo* suggests that mechanical stresses play a role in calcification of the leaflets (33,34,83). A number of studies has been reported on a finite element stress analysis on the BHV valve leaflets in the closed position in order to correlate regions of high stresses with calcification and tissue failure (84,85). Recent studies have indicated that damage to the valvular structural matrix occurs as the result of mechanical stresses and that the structural damage occurs in spatially distinct sites from those of cuspal mineralization (86). Hence, there is a renewed interest in analyzing the stresses on the leaflets during the opening and closing phases of the leaflet function since loss in flexural rigidity has been demonstrated after the valves undergo 50 million cycles of loading *in vitro*(87). Detailed analysis on the mechanism of structural failure under cyclic loading requires experimental quantification of the complex motion and also detailed description of the nonlinear anisotropic material property description of the BHV leaflets (88). Leaflet motion quantification by the application of ink markers on the leaflet surface (89), laser profiling technique (90), and noncontacting structured laser light projection technique (91) have been employed for the BHV leaflets. Several studies have been reported on the material characterization of chemically treated BHV leaflet tissue (92–94). Even though these studies have yielded valuable information about the nonlinear material characterization of the leaflets, true physiological characterization requires biaxial loading tests. Recently, constitutive models for the BHV leaflets under biaxial loading have been reported (95,96), which takes into consideration the local architecture of the valve fibers. Since the treated aortic valve leaflet consists of three layers, it can also be anticipated that the behavior of each layer will be different under the physiological loading during the valve opening and closing phases. Recent studies have included separating the fibrosa and ventricularis layers by dissecting microscope and each layer being subjected to biaxial testing (97). Incorporating such detailed material description in the computational analysis in order to determine the flexural and shear stresses on the multilayered tissue leaflets may yield valuable information on the nature of the effect of mechanical stresses on structural disintegration and limited durability with implanted BHV.

Computational Simulation of Heart Valve Function

With the advent of high speed computing capabilities, advances in computational fluid dynamic and finite element numerical analyses algorithms, simulations to determine the mechanical stresses on the blood elements and leaflets during the valve function, is being increasingly employed to understand the mechanics of valve function. In the case of MHV, recent studies have focused on the mechanical stresses developed during the closing phase of valve function. Wall shear stresses of the order of 1000 Pa have been reported through a numerical simulation in the central clearance of a bileaflet valve with the leaflets in the fully closed position (98). Since the simulation was with the leaflets in the closed position, this simulation did not incorporate the effects of a large pressure gradient across the leaflet at the instant of valve closure (54). A quasistatic simulation in which the leaflets were in the fully closed position and with the application of the measured transient pressures at the instant of valve closure indicated the presence of shear stresses of ~ 2200 Pa (99,100). Employing moving boundaries for the mechanical valve leaflet very near the time of valve closure, simulations of the flow dynamics in the clearance region between the edge of the leaflet and the seat stop have demonstrated fluid velocities of the order of 28 $\mathrm{m \cdot s^{-1}}$ with large negative pressure regions on the inflow side of the occluder (101,102). The computed resulting wall shear stress magnitudes at the edge of the leaflet exceeds 17 kPa for a fraction of a second at the instant of valve closure and first rebound after impact with the seat stop (72).

In the case of BHV, finite element analysis has been the popular technique employed to determine the stress distribution on the leaflets with the blood pressure applied as the load on the leaflets. Such analyses have also been employed to perform stress analysis with native aortic and mitral valves (103,104). Stress analysis with BHV geometry have incorporated the nonlinear material property of the leaflets and employed both the rigid and flexible stent support (84,85). These results have generally suggested a correlation between regions of high stresses with the leaflets in the fully closed position and calcification observed with implanted valves.

Numerical simulation of native and prosthetic heart valve function is quite challenging with the necessity of addressing several difficult issues. In the case of MHV simulation, the mesh generation for the detailed 3D geometry of the tilting disk and bileaflet valves, including the complex geometry of the hinge mechanism, is required. The simulation must also have the ability to deal with the moving leaflets. In the case of BHV simulation, it is necessary to incorporate the nonlinear anisotropic material property of the leaflets and compute the complex motion of the leaflets due to the external load imparted on the same by the surrounding fluid. In modeling the valve function, the fluid domain is most conveniently described using the Eulerian reference frame in which the fluid moves through a fixed mesh. A Lagrangian formulation is more appropriate for the leaflet motion in which the mesh moves together with the leaflet. The complete simulation of the heart valve function requires a fluid-structure interaction simulation and the two formulations are incompatible for such an analysis. Two methods have generally been employed to circumvent this problem. An Eulerian method used in the simulation of heart valve function is the fictitious domain method employed by de Hart et al. (105,106) and Baaijens (107). This simulation requires careful attention to accurate mesh representation for flows near the leaflet due to the use of fixed grid and numerical stability. It is also computationally very intensive. The second approach is the arbitrary Lagrangian–Eulerian (ALE) method in which the computational mesh is allowed

to deform (108) and move in space arbitrarily to conform to the moving boundaries. It has also been successfully employed in the heart valve simulation. The disadvantage with this method is the need for a mechanism to adapt or deform the mesh to conform to the boundary motion at each new time step. Large and complex deformation of the BHV leaflets within the computational domain makes the mesh adaptation very difficult when structured mesh is used and the mesh topology has to be maintained. Mesh regeneration has been employed to avoid the problems with the maintenance of mesh topology, and requires reinterpolation of the flow variables that can be expensive to perform and may result in large artificial errors. The ALE method recently has been employed in the simulation of prosthetic valve function by several investigators (102,109–111). More recently, a fluid-structure interaction simulation for the mechanical valve closing dynamics employing the ALE method has been presented for both two-dimensional (2D) (112) and 3D (72) geometry of a bileaflet valve. In this analysis in which the details of the hinge geometry was not included, the leaflet was specified to rotate around an axis and the motion of the leaflets was computed by solving the governing equation of motion for the leaflet with fluid-induced stresses specified as the external load. The pressure and velocity field is calculated employing the CFD solver employing the ALE method. The grid points on the leaflets are rotated, based on the solution of the equation of motion of the leaflet. An elliptic solver is employed to the entire mesh using the initial methods and the computed displacement of the leaflet grid points. This simulation has also clearly demonstrated the presence of abnormally high wall shear stresses in the clearance region of the leaflet and the valve housing at the instant of valve closure and leaflet rebound. These studies indicate that the shear stress-time magnitude present in this region far exceeds magnitudes suggested for platelet activation, and hence may be the critical factor for thrombus initiation with MHV implants.

SUMMARY

Since the introduction of the first prosthetic valve replacement for a diseased native heart valve >40 years ago, numerous developments in design and manufacturing process have resulted in improved performance of these implants and patients are leading a relatively normal life. Continued efforts are underway to minimize the effect of thrombus deposition with MHV and to improve the durability of implanted BHV. State-of-the-art experimental techniques and computational simulations are being applied to improve our understanding on the relationship between the complex solid and fluid mechanics during the prosthetic valve function and its relationship with the problems still continuing to be observed with the implants. With the advent of high performance computers and advances in computational flow dynamics algorithms, more detailed 3D unsteady laminar and disturbed flow simulations are becoming a reality today. Development of fluid–structure interaction simulations, inclusion of detailed structural analysis of biological leaflet valves during the valve function, and the behavior of platelets and red blood cells in the unsteady 3D flow field to simulate the platelet and red blood cell motion in the crevices are crucial for our further understanding of the mechanical valve function. However, complementary experimental studies to validate the simulations are also essential to gain confidence in the results of the complicated numerical simulations. A deterministic study of the effect of such stresses on the leaflet structures as well as formed elements in blood will require an analysis that includes computational algorithms that span multiple length and time scales. Efforts are underway to develop such simulations. These studies will also provide valuable information toward the development of tissue engineered valve replacements.

BIBLIOGRAPHY

Cited References

1. Yoganathan A, Lemmon JD, Ellis JT. Heart Valve Dynamics. In: Bronzino JD, editor. The Biomedical Engineering Handbook. 2nd ed. Boca Raton: CRC Press and IEEE Press; 2000. p 29.1–29.15.
2. Westaby S, Karp RB, Blackstone EH, Bishop SP. Adult human valve dimensions and their surgical significance. Am J Cardiol 1984;53(4):552–556.
3. Davies MJ. Pathology of Cardiac Valves. London: Butterworth; 1980.
4. Gorlin R, Gorlin SG. Hydraulic formula for calculation of the area of the stenotic mitral valve, other cardiac valves, and central circulatory shunts. I Am Heart J 1951;41(1):1–29.
5. Schlant RC, Alexander RW, editors. Technique of Doppler and color flow Doppler in the evaluation of cardiac disorders and function. in The Heart: Arteries and Veins. The Heart: Arteries and Veins. 6th ed. New York: McGraw-Hill; 1994.
6. Braunwald E. Heart Disease: A Textbook of Cardiovascular Medicine. 2nd ed. Philadelphia: W. B. Saunders; 1984.
7. Aazami M, Schafers HJ. Advances in heart valve surgery. J Interv Cardiol 2003;16(6):535–541.
8. Harken DE, Taylor WJ, Lefemine AA, Lunzer S, Low HB, Cohen ML, Jacobey JA. Aortic valve replacement with a caged ball valve. Am J Cardiol 1962;9:292–299.
9. Hammermeister K, Sethi GK, Henderson WG, Grover FL, Oprian C, Rahimtoola SH. Outcomes 15 years after valve replacement with a mechanical versus a bioprosthetic valve: final report of the Veterans Affairs randomized trial. J Am Coll Cardiol 2000;36(4):1152–1158.
10. Bach DS. Choice of prosthetic heart valves: update for the next generation. J Am Coll Cardiol 2003;42(10):1717–1719.
11. Starr A, Edwards ML. Mitral replacement: Clinical experience with a ball valve prosthesis. Ann Surg 1961;154:726.
12. Bhuvaneshwar G, Ramani AV, Chandran KB. Polymeric occluders in tilting disk heart valve prostheses. In: Dumitriu S, editor. Polymeric Biomaterials. 2nd ed. New York: Marcel Dekker; 2002. p 589–610.
13. Chandran KB. Dynamic behavior analysis of mechanical heart valve prostheses. In: Leondes C, editor. Cardiovascular Techniques. Boca Raton: CRC Press; 2001. p 3.1–3.31.
14. Yoganathan A. Cardiac valve prostheses. In: Bronzino JD, editor. Biomedical Engineering Handbook. Boca Raton: CRC Press and IEEE Press; 2000. p 127.1–127.23.
15. Wada J. Knotless suture method and Wada hingeless valve. Jpn J Thor Sur 1967;15:88.
16. Bjork VO. Experience with the Wada-Cutter valve prostheses in the aortic area- One year follow-up. J Thorac Cardiovasc Surg 1970;60:26.

17. Bjork VO. A new tilting disk valve prostheses. Scand J Cardiovasc Surg 1969;3:1.
18. Bjork VO. Delrin as implant material for valve occluders. Scand J Thorac Cardiovasc Surg 1972;6:103.
19. Bjork VO. The pyrolytic carbon occluder for the Bjork–Shiley tilting disk valve prostheses. Scand J Thorac Cardiovasc Surg 1972;6:109–113.
20. Butchart EG, Bodnar E. Thrombosis, Embolism, and Bleeding. Marlow, Bucks (UK): ICR Publishing; 1992.
21. Carpentier A, Lamaigre RL, Carpentier S, Dubost C. Biological factors affecting long-term results of valvular heterografts. J Thorac Cardiovasc Surg 1969;58:467.
22. O'Brien M, Clarebrough JK. Heterograft aortic valve transplant for human valve disease. Aust Med J 1966;2(228).
23. Carpentier A, Dubost C. From xenograft to bioprosthesis: Evolution of concepts and techniques of valvular xenografts. In: Ionescu MI, Ross DN, Wooler GH, editors. Biological Tissue in Heart Valve Replacement. London: Butterworth; 1971. p 515–541.
24. Yoganathan AP. Cardiac Valve Prostheses. In: Bronzino JD, editor. The Biomedical Engineering Handbook. 2nd ed. Boca Raton (FL): CRC Press; 2000. p 127.1–127.23.
25. Wieting DW. *Dynamic flow characteristics of heart valves*, Ph.D., dissertation. University of Texas, Austin, TX.
26. Cape EG, Nanda NC, Yoganathan AP. Quantification of regurgitant flow through bileaflet heart valve prostheses: theoretical and *in vitro* studies. Ultrasound Med Biol 1993;19(6):461–468.
27. Cape EG, Kim YH, Heinrich RS, Grimes RY, Muralidharan E, Broder JD, Schwammenthal E, Yoganathan AP, Levine RA. Cardiac motion can alter proximal isovelocity surface area calculations of regurgitant flow. J Am Coll Cardiol 1993;22(6):1730–1737.
28. Edmunds LH Jr. Thrombotic and bleeding complications of prosthetic heart valves. Ann Thorac Surg 1987;44(4):430–445.
29. Paulsen PK, Nygaard H, Hasenkam JM, Gormsen J, Stodkilde-Jorgensen H, Albrechtsen O. Analysis of velocity in the ascending aorta in humans. A comparative study among normal aortic valves, St. Jude Medical and Starr-Edwards Silastic Ball valves. Int J Artif Organs 1988;11(4): 293–302.
30. Farthing S, Peronneau P. Flow in the thoracic aorta. Cardiovasc Res 1979;13(11):607–620.
31. Rossvoll O, Samstad S, Torp HG, Linker DT, Skjaerpe T, Angelsen BA, Hatle L. The velocity distribution in the aortic anulus in normal subjects: a quantitative analysis of two-dimensional Doppler flow maps. J Am Soc Echocardiogr 1991;4(4):367–378.
32. Kilner PJ, Yang GZ, Mohiaddin RH, Firmin DN, Longmore DB. Helical and retrograde secondary flow patterns in the aortic arch studied by three-directional magnetic resonance velocity mapping. Circulation 1993;88(5 Pt 1):2235–2247.
33. Thubrikar M, Piepgrass WC, Shaner TW, Nolan SP. The design of the normal aortic valve. Am J Physiol 1981; 241(6):H795–801.
34. Thubrikar MJ, Skinner JR, Eppink RT, Nolan SP. Stress analysis of porcine bioprosthetic heart valves *in vivo*. J Biomed Mater Res 1982;16(6):811–826.
35. Thubrikar M, Skinner JR, Aouad J, Finkelmeier BA, Nolan SP. Analysis of the design and dynamics of aortic bioprostheses in vivo. J Thorac Cardiovasc Surg 1982;84(2):282–290.
36. Thubrikar M, Eppink RT. A method for analysis of bending and shearing deformations in biological tissue. J Biomech 1982;15(7):529–535.
37. Thubrikar M, Carabello BA, Aouad J, Nolan SP. Interpretation of aortic root angiography in dogs and in humans. Cardiovasc Res 1982;16(1):16–21.
38. Nevaril C, Hellums J, Alfrey C Jr. Physical effects in red blood cell trauma. J Am Inst Chem Eng 1969;15:707.
39. Mohandas N, Hochmuth RM, Spaeth EE. Adhesion of red cells to foreign surfaces in the presence of flow. J Biomed Mater Res 1974;8(2):119–136.
40. Sutera SP, Mehrjardi MH. Deformation and fragmentation of human red blood cells in turbulent shear flow. Biophys J 1975;15(1):1–10.
41. Ramstack JM, Zuckerman L, Mockros LF. Shear-induced activation of platelets. J Biomech 1979;12(2):113–125.
42. Anderson GH, Hellums JD, Moake JL, Alfrey CP Jr. Platelet lysis and aggregation in shear fields. Blood Cells 1978;4(3): 499–511.
43. Yoganathan AP, Corcoran WH, Harrison EC. In vitro velocity measurements in the vicinity of aortic prostheses. J Biomech 1979;12(2):135–152.
44. Chandran KB, Cabell GN, Khalighi B, Chen CJ. Laser anemometry measurements of pulsatile flow past aortic valve prostheses. J Biomech 1983;16(10):865–873.
45. Chandran KB. Pulsatile flow past St. Jude Medical bileaflet valve. An in vitro study. J Thorac Cardiovasc Surg 1985; 89(5):743–749.
46. Chandran KB, Lee CS, Aluri S, Dellsperger KC, Schreck S, Wieting DW. Pressure distribution near the occluders and impact forces on the outlet struts of Bjork–Shiley convexo-concave valves during closing. J Heart Valve Dis 1996; 5(2):199–206.
47. Dimitri W, Williams BT. Fracture of a Duromedics mitral valve housing with leaflet escape. J Cardiovas Sur 1990; 31:41–46.
48. Deuvaert FE, Devriendt J, Massaut J. Leaflet escape of a mitral Duromedics prosthesis (case report). Acta Chirur 1989;89:15–18.
49. Kafesjian R, Howanec M, Ward GD, Diep L, Wagstaff LS, Rhee R. Cavitation damage of pyrolytic carbon in mechanical heart valves. J Heart Valve Dis 1994;3:S2–S7.
50. Guo GX, Chiang TH, Quijano RC, Hwang NH. The closing velocity of mechanical heart valve leaflets. Med Eng Phys 1994;16(6):458–464.
51. Wu ZJ, Shu MC, Scott DR, Hwang NH. The closing behavior of Medtronic Hall mechanical heart valves. ASAIO J 1994;40(3):M702–6.
52. Wu ZJ, Wang Y, Hwang NH. Occluder closing behavior: a key factor in mechanical heart valve cavitation. J Heart Valve Dis 1994;3(Suppl 1):S25–33; discussion S33–4.
53. Chandran KB, Aluri S. Mechanical valve closing dynamics: relationship between velocity of closing, pressure transients, and cavitation initiation. Ann Biomed Eng 1997;25 (6):926–938.
54. Chandran KB, Lee CS, Chen LD. Pressure field in the vicinity of mechanical valve occluders at the instant of valve closure: correlation with cavitation initiation. J Heart Valve Dis 1994;3(Suppl 1):S65–75; discussion S75–6.
55. Graf T, Reul H, Dietz W, Wilmes R, Rau G. Cavitation of mechanical heart valves under physiologic conditions. J Heart Valve Dis 1992;1(1):131–141.
56. Graf T, Fischer H, Reul H, Rau G. Cavitation potential of mechanical heart valve prostheses. Int J Artif Organs 1991;14(3):169–174.
57. Zapanta CM, Liszka EG Jr., Lamson TC, Stinebring DR, Deutsch S, Geselowitz DB, Tarbell JM. A method for real-time in vitro observation of cavitation on prosthetic heart valves. J Biomech Eng 1994;116(4):460–468.
58. Garrison LA, Lamson TC, Deutsch S, Geselowitz DB, Gaumond RP, Tarbell JM. An *in vitro* investigation of prosthetic heart valve cavitation in blood. J Heart Valve Dis 1994; 3(Suppl 1):S8–22; discussion S22–4.

59. Chandran KB, Dexter EU, Aluri S, Richenbacher WE. Negative pressure transients with mechanical heart-valve closure: correlation between in vitro and in vivo results. Ann Biomed Eng 1998;26(4):546–556.
60. Dexter EU, Aluri S, Radcliffe RR, Zhu H, Carlson DD, Heilman TE, Chandran KB, Richenbacher WE. *In vivo* demonstration of cavitation potential of a mechanical heart valve. ASAIO J 1999;45(5):436–441.
61. Kleine P, Perthel M, Hasenkam JM, Nygaard H, Hansen SB, Laas J. High-intensity transient signals (HITS) as a parameter for optimum orientation of mechanical aortic valves. Thorac Cardiovasc Surg 2000;48(6):360–363.
62. Johansen P, Andersen TS, Hasenkam JM, Nygaard H. In-vivo prediction of cavitation near a Medtronic Hall valve. J Heart Valve Dis 2004;13(4):651–658.
63. Johansen P, Manning KB, Tarbell JM, Fontaine AA, Deutsch S, Nygaard H. A new method for evaluation of cavitation near mechanical heart valves. J Biomech Eng 2003;125(5):663–670.
64. Johansen P, Travis BR, Paulsen PK, Nygaard H, Hasenkam JM. Cavitation caused by mechanical heart valve prostheses—a review. APMIS 2003;(Suppl)(109):108–112.
65. Biancucci BA, Deutsch S, Geselowitz DB, Tarbell JM. *In vitro* studies of gas bubble formation by mechanical heart valves. J Heart Valve Dis 1999;8(2):186–196.
66. Lin HY, Bianccucci BA, Deutsch S, Fontaine AA, Tarbell JM. Observation and quantification of gas bubble formation on a mechanical heart valve. J Biomech Eng 2000;122 (4):304–309.
67. Bachmann C, Kini V, Deutsch S, Fontaine AA, Tarbell JM. Mechanisms of cavitation and the formation of stable bubbles on the Bjork–Shiley Monostrut prosthetic heart valve. J Heart Valve Dis 2002;11(1):105–113.
68. Dauzat M, Deklunder G, Aldis A, Rabinovitch M, Burte F, Bret PM. Gas bubble emboli detected by transcranial Doppler sonography in patients with prosthetic heart valves: a preliminary report. J Ultrasound Med 1994;13(2):129–35.
69. Reisner SA, Rinkevich D, Markiewicz W, Adler Z, Milo S. Spontaneous echocardiographic contrast with the carbomedics mitral valve prosthesis. Am J Cardiol 1992;70(18): 1497–1500.
70. Droste DW, Hansberg T, Kemeny V, Hammel D, Schulte-Altedorneburg G, Nabavi DG, Kaps M, Scheld HH, Ringelstein EB. Oxygen inhalation can differentiate gaseous from nongaseous microemboli detected by transcranial Doppler ultrasound. Stroke 1997;28(12):2453–2456.
71. Georgiadis D, Wenzel A, Lehmann D, Lindner A, Zerkowski HR, Zierz S, Spencer MP. Influence of oxygen ventilation on Doppler microemboli signals in patients with artificial heart valves. Stroke 1997;28(11):2189–2194.
72. Cheng R, Lai YG, Chandran KB. Three-dimensional fluid-structure interaction simulation of bileaflet mechanical heart valve flow dynamics. Ann Biomed Eng 2004;32(11): 1469–1481.
73. Ellis JT, Yoganathan AP. A comparison of the hinge and near-hinge flow fields of the St Jude medical hemodynamic plus and regent bileaflet mechanical heart valves. J Thorac Cardiovasc Surg 2000;119(1):83–93.
74. Saxena R, Lemmon J, Ellis J, Yoganathan A. An in vitro assessment by means of laser Doppler velocimetry of the Medtronic advantage bileaflet mechanical heart valve hinge flow. J Thorac Cardiovasc Surg 2003;126(1):90–98.
75. Lim WL, Chew YT, Chew TC, Low HT. Steady flow dynamics of prosthetic aortic heart valves: a comparative evaluation with PIV techniques. J Biomech 1998;31(5):411–421.
76. Lim WL, Chew YT, Chew TC, Low HT. Pulsatile flow studies of a porcine bioprosthetic aortic valve in vitro: PIV measurements and shear-induced blood damage. J Biomech 2001; 34(11):1417–1427.
77. Browne P, Ramuzat A, Saxena R, Yoganathan AP. Experimental investigation of the steady flow downstream of the St. Jude bileaflet heart valve: a comparison between laser Doppler velocimetry and particle image velocimetry techniques. Ann Biomed Eng 2000;28(1):39–47.
78. Castellini P, Pinotti M, Scalise L. Particle image velocimetry for flow analysis in longitudinal planes across a mechanical artificial heart valve. Artif Organs 2004;28(5):507–513.
79. Manning KB, Kini V, Fontaine AA, Deutsch S, Tarbell JM. Regurgitant flow field characteristics of the St. Jude bileaflet mechanical heart valve under physiologic pulsatile flow using particle image velocimetry. Artif Organs 2003;27(9):840–846.
80. Kini V, Bachmann C, Fontaine A, Deutsch S, Tarbell JM. Flow visualization in mechanical heart valves: occluder rebound and cavitation potential. Ann Biomed Eng 2000; 28(4):431–441.
81. Kini V, Bachmann C, Fontaine A, Deutsch S, Tarbell JM. Integrating particle image velocimetry and laser Doppler velocimetry measurements of the regurgitant flow field past mechanical heart valves. Artif Organs 2001;25(2):136–145.
82. Chandran KB, Cabell GN, Khalighi B, Chen CJ. Pulsatile flow past aortic valve bioprostheses in a model human aorta. J Biomech 1984;17(8):609–619.
83. Thubrikar MJ, Deck JD, Aouad J, Nolan SP. Role of mechanical stress in calcification of aortic bioprosthetic valves. J Thorac Cardiovasc Surg 1983;86(1):115–125.
84. Hamid MS, Sabbah HN, Stein PD. Influence of stent height upon stresses on the cusps of closed bioprosthetic valves. J Biomech 1986;19(9):759–769.
85. Rousseau EP, van Steenhoven AA, Janssen JD. A mechanical analysis of the closed Hancock heart valve prosthesis. J Biomech 1988;21(7):545–562.
86. Sacks MS, Schoen FJ. Collagen fiber disruption occurs independent of calcification in clinically explanted bioprosthetic heart valves. J Biomed Mat Res 2002;62(3):359–371.
87. Gloeckner DC, Billiar KL, Sacks MS. Effects of mechanical fatigue on the bending properties of the porcine bioprosthetic heart valve. ASAIO J 1999;45(1):59–63.
88. Sacks MS. The biomechanical effects of fatigue on the porcine bioprosthetic heart valve. J Long-Term Effects Med Implants 2001;11(3–4):231–247.
89. Lo D, Vesely I. Biaxial strain analysis of the porcine aortic valve. Ann Thorac Surg 1995;60(2 Suppl):S374–8.
90. Donn AW, Bernacca GM, Mackay TG, Gulbransen MJ, Wheatley DJ. Laser profiling of bovine pericardial heart valves. Int J Artif Organs 1997;20(8):436–439.
91. Iyengar AKS, Sugimoto H, Smith DB, Sacks MS. Dynamic in vitro quantification of bioprosthetic heart valve leaflet motion using structured light projection. Ann Biomed Eng 2001;29(11):963–973.
92. Lee JM, Boughner DR, Courtman DW. The gluteraldehyde-stabilized porcine aortic valve xenograft. II. Effect of fixation with or without pressure on the tensile viscoelastic properties of the leaflet material. J Biomed Mater Res 1984;18 (1):79–98.
93. Lee JM, Courtman DW, Boughner DR. The gluteraldehyde-stabilized porcine aortic valve xenograft. I. Tensile viscoelastic properties of the fresh leaflet material. J Biomed Mater Res 1984;18(1):61–77.
94. Vesely I, Noseworthy R. Micromechanics of the fibrosa and the ventricularis in aortic valve leaflets. J Biomech 1992;25(1): 101–113.
95. Billiar KL, Sacks MS. Biaxial mechanical properties of the native and gluteraldehyde-treated aortic valve cusp: Part II—A structural constitutive model. J Biomech Eng 2000;122(4):327–335.
96. Billiar KL, Sacks MS. Biaxial mechanical properties of the natural and gluteraldehyde treated aortic valve cusp—Part I: Experimental results. J Biomech Eng 2000;122(1):23–30.
97. Sacks M, Stella J, Chandran KB. *A bi-layer structural constitutive model for the aortic valve leaflet*. Philadelphia: BMES Annu Conf, Abstract No. 1164, 2004.

98. Reif TH. A numerical analysis of the backflow between the leaflets of a St. Jude Medical cardiac valve prosthesis. J Biomech 1991;24(8):733–741.
99. Lee CS, Chandran KB. Instantaneous back flow through peripheral clearance of Medtronic Hall tilting disk valve at the moment of closure. Ann Biomed Eng 1994;22(4):371–380.
100. Lee CS, Chandran KB. Numerical simulation of instantaneous backflow through central clearance of bileaflet mechanical heart valves at closure: shear stress and pressure fields within clearance. Med Biol Eng Comput 1995; 33(3):257–263.
101. Bluestein D, Einav S, Hwang NH. A squeeze flow phenomenon at the closing of a bileaflet mechanical heart valve prosthesis. J Biomech 1994;27(11):1369–1378.
102. Makhijani VB, Yang HQ, Singhal AK, Hwang NH. An experimental-computational analysis of MHV cavitation: effects of leaflet squeezing and rebound. J Heart Valve Dis 1994;3(Suppl 1):S35–44; discussion S44–8.
103. Grande KJ, Cochran RP, Reinhall PG, Kunzelman KS. Stress variations in the human aortic root and valve: the role of anatomic asymmetry. Ann Biomed Eng 1998;26(4): 534–545.
104. Kunzelman KS, Quick DW, Cochran RP. Altered collagen concentration in mitral valve leaflets: biochemical and finite element analysis. Ann Thorac Surg 1998;66(6 Suppl):S198–205.
105. De Hart J, Peters GW, Schreurs PJ, Baaijens FP. A two-dimensional fluid-structure interaction model of the aortic valve [correction of value]. J Biomech 2000;33(9):1079–1088.
106. De Hart J, Peters GW, Schreurs PJ, Baaijens FP. A three-dimensional computational analysis of fluid-structure interaction in the aortic valve. J Biomech 2003;36(1):103–112.
107. Baaijens FPT. A fictitious domain/mortar element method for fluid-structure interaction. Int J Numerical Met Fluids 2001;35(7):743–761.
108. Lai YG. Unstructured grid arbitrarily shaped element method for fluid flow simulation. AIAA J 2000;38 (12):2246–2252.
109. Makhijani VB, Yang HQ, Dionne PJ, Thubrikar MJ. Three-dimensional coupled fluid-structure simulation of pericardial bioprosthetic aortic valve function. ASAIO J 1997;43(5):M387–M392.
110. Aluri S, Chandran KB. Numerical simulation of mechanical mitral heart valve closure. Ann Biomed Eng 2001; 29(8): 665–676.
111. Lai YG, Chandran KB, Lemmon J. A numerical simulation of mechanical heart valve closure fluid dynamics. J Biomech 2002;35(7):881–892.
112. Cheng R, Lai YG, Chandran KB. Two-dimensional fluid-structure interaction simulation of bi-leaflet mechanical heart valve flow dynamics. Heart Valve Dis 2003;12: 772–780.

See also BIOCOMPATIBILITY OF MATERIALS; BIOMATERIALS, CORROSION AND WEAR OF; TISSUE ENGINEERING.

HEART VALVE PROSTHESES, IN VITRO FLOW DYNAMICS OF

STEVEN CECCIO
University of Michigan
Ann Arbor, Michigan

INTRODUCTION

The cardiac cycle begins with venous blood passing through the tricupsid valve in response to relaxation of the right ventricle. Then, during ventricular contraction, the tricuspid valve closes and blood flows through the open pulmonary valve to the lungs. Similarly, oxygenated blood leaving the lungs crosses the mitral valve during filling of the left ventricle and is then ejected through the aortic valve when the left ventricle contracts. A normally functioning heart valve must open and close approximately once every second without posing significant resistance to flow during opening and without allowing significant leakage during closure.

Due to the critical role that heart valves play in controlling pressures and flows in the heart and throughout the body, valve disease is a serious health risk and can be fatal if not treated. In the United States almost 20,000 people die annually as a result of heart valve disease (1). Although the causes and mechanisms of heart valve diseases are varied, their effect can usually be reduced to either failure of the valve to open fully (stenosis) or failure to prevent leakage of blood (regurgitation). Patients with stenosis or regurgitation may experience chest pain, labored breathing, lightheadedness, and a reduced tolerance for exercise.

Because of the mechanical nature of valve dysfunction, treatments for severe valve disease usually involve surgical intervention to restore the flow control function of the valve. Early surgical treatments consisted of a surgeon using a tool, or his fingers, to reach into the beating heart and forcefully open a stenotic mitral valve. With the advent of cardiopulmonary bypass in the 1950s, the notion of fabricating and implanting a prosthetic valve became more feasible and by the early 1960s the first successful and repeatable prosthetic valve implants were performed by Starr (2,3). The valve he developed with Edwards, an engineer, consisted of a ball trapped in a rigid, dome-shaped cage. At the inflow edge of the valve was a cloth flange, the sewing ring, which enabled the surgeon to sew the valve into the patient's heart (see Fig. 1). Although crude in comparison to the native valve structure, this

Figure 1. Starr–Edwards ball–cage valve. (Courtesy of Edwards Lifesciences, Irvine, CA).

Starr–Edwards valve and subsequent models had been used successfully in >175,000 people by 1991 (4).

Over the past four decades, numerous valve designs have been developed and used clinically. There were ~80,000 heart valve-related surgeries in the United States in 1999, ~50,000 of which were implants of prosthetic aortic valves (5). But despite the success of prosthetic valve technology, currently no valve is optimal, so surgeons and engineers continue their collaborative efforts in pursuit of improved designs.

Due to the lower pressures in the right ventricle, the tricuspid and pulmonary valves are implicated far less in heart disease than the valves of the left heart (1). Consequently, prosthetic heart valve technology is focused almost entirely on mitral and aortic valves. The following discussion will focus on the primary tools, techniques, and data that are of interest when evaluating these valves *in vitro*.

NATIVE VALVE STRUCTURE AND HEMODYNAMICS

Although it has been shown that a prosthetic heart valve need not look like a native heart valve to have adequate *in vivo* function, it must have geometry suitable for the intended implantation site and must function without impeding other aspects of cardiac function. It is therefore important to understand the anatomy and physiology of native heart valves as well as the process of surgical implantation of prosthetic valves. These will be reviewed briefly here; more detailed reviews can be found elsewhere (6–10). Figure 2 shows schematic drawings of a cross-section of the left ventricle in systole and diastole.

The base, or inflow perimeter, of the aortic valve is contiguous with the left ventricular outflow tract and the anterior leaflet of the mitral valve. The valve is comprised of three flexible, triangular leaflets, each of which attaches along ~120° of the circumference of the aorta. The inflow attachment line curves upward from the annulus at both ends, giving the leaflet its curved, three-dimensional (3D) geometry. At the outflow aspect of the valve, each adjacent leaflet pair meet at a commissure, a junction point on the aortic wall, and the aorta surrounding the valve leaflets is comprised of three bulbous regions, the sinuses of Valsalva.

Contraction of the left ventricle causes ventricular pressure to increase until it exceeds that in the aorta at which point the aortic valve opens rapidly and allows flow into the aorta. The flow reaches its peak amplitude about one-third of the way through the flow cycle. As the left ventricle relaxes, ventricular pressure falls, which reduces the pressure gradient and causes flow in the aorta to decelerate. Eventually, the pressure in the aorta exceeds that in the left ventricle and the aortic valve closes. During forward flow, a vortex forms in each sinus, which may play a role in the subsequent closure of the leaflets (11).

Aortic flow continues forward for a short period of time after the reversal of the pressure gradient due to the momentum of the blood (12), and a small volume of blood, the closing volume, is pushed back into the ventricle at closure due to the motion of the closing leaflets. During diastole, the closed aortic valve is under a back pressure of 80–100 mmHg (10.66–13.33 kPa). Representative pressure and flow waveforms of the aortic valve are shown in Fig. 3. At a heart rate of 70 beats · min^{-1}, systole will typically last ~35% of the whole cardiac cycle, or ~300 ms. At an exercise heart rate of 120 beats · min^{-1}, the systolic ratio will increase to near 50%.

Although there are many factors involved in deciding whether a patient needs surgery, general guidelines suggest that a diseased aortic valve may be considered for replacement with a prosthesis when the area of the valve has been reduced to 0.5–1.0 cm^2 (compared to a normal range of 3–4 cm^2), or when regurgitation has caused an ejection fraction of <50% (13). When a diseased aortic valve is replaced with a prosthetic valve, the aorta is cut open, all three leaflets are cut out, and any calcium is removed from the valve annulus. The diameter of the annulus is then measured to determine the appropriate sized prosthetic valve to use. The prosthetic valve is then implanted by stitching the sewing ring to the tissue of the native annulus, although the exact implantation process as well as the positioning of the valve will vary based on valve type.

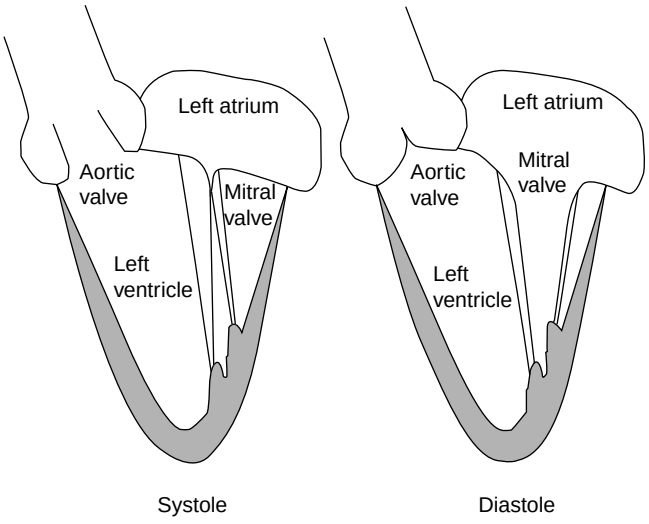


Figure 2. Schematic of the left ventricle and valves during systole and diastole.

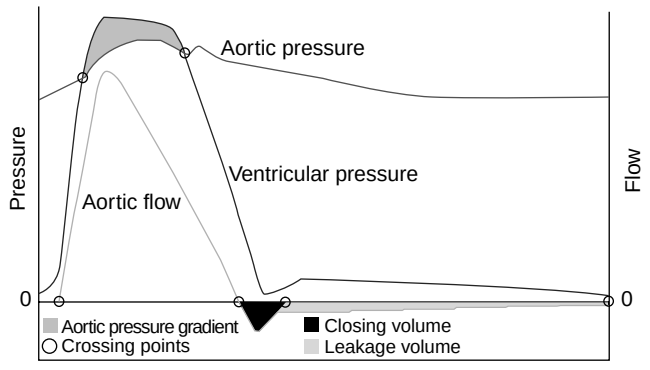


Figure 3. Idealized waveforms of pressure and flow through the aortic valve.

The mitral valve is structurally and functionally distinct from the aortic valve. It has two primary, unequally sized leaflets, each of which is made from several segments. The longer anterior leaflet, adjacent to the aortic valve, is attached along one-third of the annulus, while the shorter posterior leaflet connects to about two-thirds of the annulus. The leaflets contain many furrows or wrinkles that allows the valve to occupy minimal space in the heart while still having a large surface area for preventing leakage and supporting the stress of closure. The annulus itself is not circular, but D-shaped, and has a 3D, saddle-shaped curvature (14).

Many tendonous chords emanate from the underside and edge of the leaflets and attach to the papillary muscles, which are part of the wall of the left ventricle. The chords and papillary muscles comprise the tensioning component of the mitral apparatus, helping the valve to support and balance the stresses on the leaflets during closure. The entire mitral apparatus is a dynamic and active structure: the annulus, leaflets, and papillary muscles all move in coordination throughout the cardiac cycle in support of proper valve function (15,16). Currently, no prosthetic mitral valve can replicate or synchronize with the complicated force and motion dynamics of the native mitral apparatus.

During left ventricular filling, diastolic flow through the mitral valve is equal to the subsequent flow out of the aortic valve, assuming there are no leaks through either valve. Although the same volume passes through the mitral valve, the flow profile is very different than that of aortic flow. First, diastolic flow occupies ~65% of the cardiac cycle, lasting ~557 ms at a heart rate of 70 beats $\cdot$ min^{-1}. Due to the longer flow period, peak flow rates and pressure gradients are usually lower through the mitral valve than the aortic valve. Second, diastolic flow occurs in two phases. As the left ventricle relaxes, pressure falls to near zero and the mitral valve opens to allow ventricular filling from the left atrium, which acts as a compliant filling chamber and maintains a fairly constant blood pressure of ~15 mmHg (1.99 kPa). The pressure gradient between the atrium and the ventricle lessens as the ventricle fills, causing the flow to approach zero and the leaflets to nearly close. The left atrium then contracts, opening the leaflets again and sending a second bolus of blood, less than the first, into the ventricle. The two waves of the biphasic diastolic flow pattern are referred to as the E and A waves. The valve closes fully at the end of the A wave and is under a back pressure of >100 mmHg (13.33 kPa) during systole. Figure 4 shows a schematic representation of pressure and flow waveforms through the mitral valve.

The decision to surgically replace the mitral valve with a prosthesis, as with the aortic valve, is based on many factors. But general functional criteria include an effective orifice area <1.0 cm^2 (depending on body size) or regurgitation causing an ejection fraction <50% (13). During surgical replacement, the left atrium is opened, the native leaflets are cut out, and any calcium is removed. As with aortic replacement, the diameter of the annulus is measured to determine the prosthetic valve size needed. For valves with stent posts, care must be taken that the posts do not impinge on the wall of the left ventricle. The chords

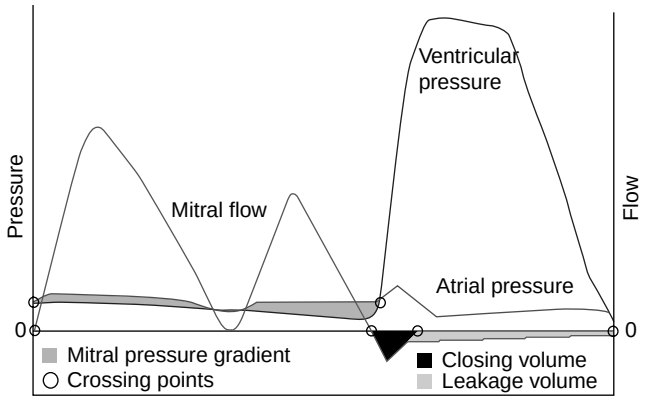


Figure 4. Idealized waveforms of pressure and flow through the mitral valve.

are generally removed, but may be left in the heart in some cases to provide structural support to the left ventricle. For some patients, the mechanical functionality of the mitral valve can be restored with surgical repair techniques, obviating the need for a prosthesis.

PROSTHETIC HEART VALVE TECHNOLOGY

A prosthetic heart valve must meet several basic functional requirements. In addition to having adequate opening and closing characteristics, it must also be durable, and biocompatible. All prosthetic valves compromise at least one of these features in favor of one of the others. The primary distinction between valve types is materials: Most valve designs use either synthetic, rigid components or flexible, biologically derived tissue. Based on these material features, a prosthetic valve is generally categorized as either a mechanical valve or a tissue valve.

Most mechanical heart valves (MHVs) are made from a rigid ring of metal or pyrolytic carbon, the outer perimeter of which is covered with a cloth sewing ring. The ring, or housing, contains one or more rigid occluders that are free to swivel on struts or hinges in response to a pressure gradient. The occluders are constrained within the housing, but are not mechanically coupled to it, allowing blood to flow completely around them, which helps to avoid flow stagnation.

Although they can adequately prevent backflow, MHVs do not create a seal during closure, and thus allow some regurgitation. The volume of blood regurgitated is tolerable for the patient, but the squeezing of blood through the closed valve creates high velocity, high shear jets that may be responsible for blood damage that leads to thrombosis (17). Whatever the mechanism, all MHV patients must take a daily dose of anticoagulant to counteract the thrombogenic effects of the valve. Without anticoagulation MHVs will develop clots that can impede valve function or become embolized into the bloodstream. One advantage of MHVs is that they are highly durable and can usually function for the duration of the recipient's life.

Although widely used, the Starr–Edwards valve was eventually surpassed by MHVs with better flow characteristics. Bileaflet valves and tilting disk valves are the two

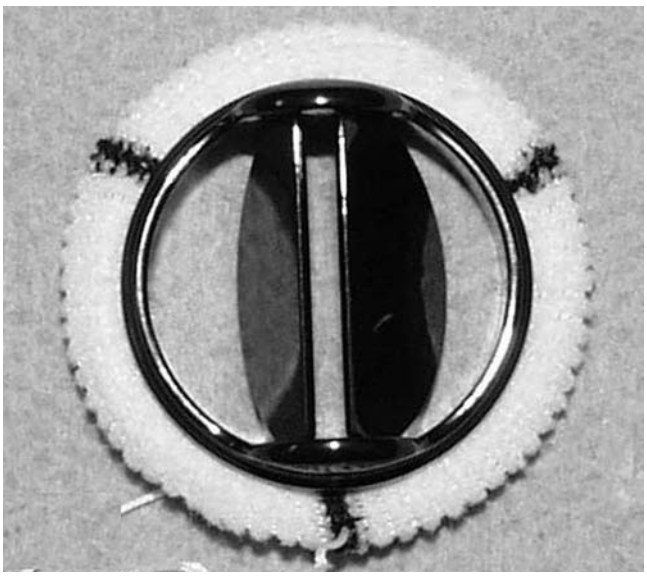
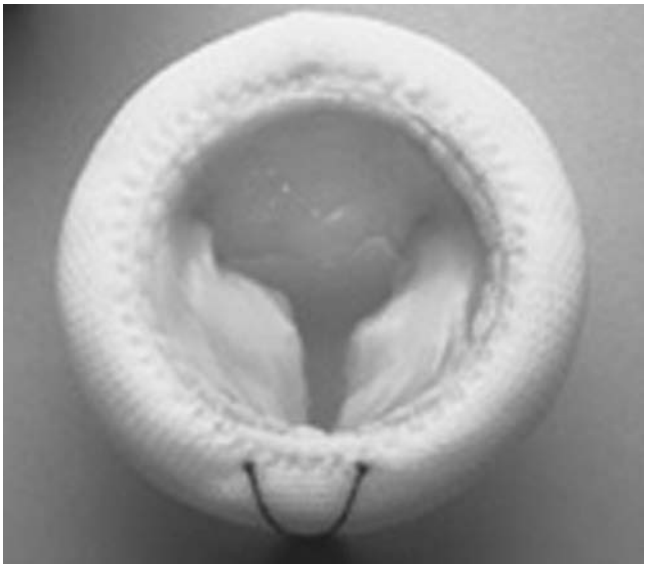


Figure 5. Examples of two mechanical heart valves: A bileaflet St. Jude Mechanical valve, and a Medtronic tilting disc valve. (Courtesy of Medtronic, Inc.)

mostly popular types in use clinically (see Fig. 5). The St. Jude bileaflet mechanical valve is by far the most widely used mechanical valve. It accounted for >70% of all mechanical valves sold in the United States in 2003 (18). Most innovations in MHVs in the last 10 years have focused on optimizing this type of bileaflet design, either through improved materials, geometries, or sewing rings. Other manufacturers of mechanical heart valves include Carbomedics, Medtronic, Sorin, and Medical Carbon Research Institute.

Tissue heart valves (THVs), also called bioprosthetic valves, were developed based on the idea that valves made of biologic tissue and with a structure similar to the native heart valve would function better *in vivo* than rigid mechanical valves. The leaflets of these valves are made from animal tissue that has been treated with a dilute solution of glutaraldehyde. Glutaraldehyde cross-links the collagen in the tissue, which prevents its breakdown and reduces its antigenicity *in vivo*. The cross-linked tissue is slightly stiffer than fresh tissue, but it still retains a functional amount of flexibility.

Porcine THVs are made from the aortic valve of a pig. To construct the prosthesis, the aortic valve is first excised from the pig heart, trimmed to remove excess tissue, incubated in glutaraldehyde, and then mounted on a cloth-covered frame with three commissure posts and a sewing ring. The frame, made of metal wire or plastic, provides structural support and allows ease of handling and implantation. The other main type of THV is the pericardial valve. Pericardium, the tissue that surrounds the heart, is separated from the heart of a cow or horse and then flattened and incubated in glutaraldehyde, producing a sheet of material that may be cut and assembled as desired to form a valve. Commercial pericardial valves typically incorporate three separate leaflets onto a three-pronged support structure similar to those used for porcine valves. Although structurally similar to the aortic valve, both porcine and pericardial valves are also implanted in the mitral position. Examples of a porcine valve and a pericardial valve are shown in Fig. 6. The Edwards Perimount pericardial valve is currently the most widely used THV, comprising >70% of all tissue valves used in the



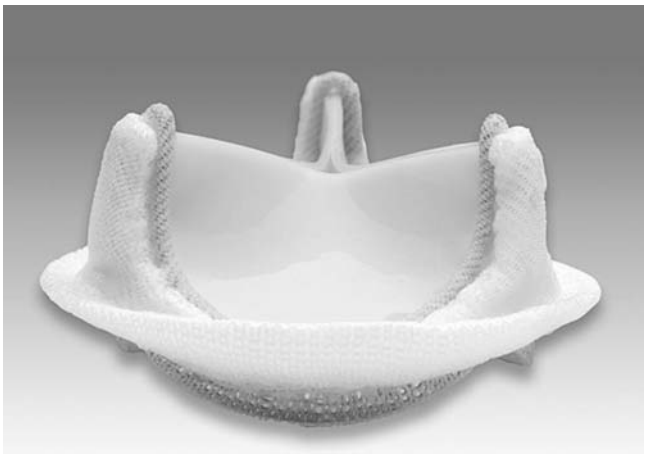


Figure 6. Examples of two tissue heart valves: A Mosaic porcine tissue valve. (Courtesy of Medtronic, Inc.) A Perimount pericardial tissue valve. (Courtesy of Edwards Lifesciences, Irvine, California.)

United States (18). Other manufacturers of tissue valves include Medtronic, St. Jude, Sorin, and 3F Therapeutics.

In contrast to MHVs, THVs generally have an unobstructed flow orifice, seal during closure, and do not typically require long-term anticoagulation. However, THVs have limited structural durability compared to MHVs, which usually last the duration of the recipient's lifetime. Although some pericardial valves have performed well for as long as 17 years in patients (19), the THVs in general are expected to degenerate or calcify within 10–15 years, at which time they must be surgically replaced with a new prosthesis. Due to this limitation, THVs are typically only implanted in patients older than 65 years of age (13). Over the past several years there has been an increased use of tissue valves in the United States, while mechanical valve usage has declined, a trend that is expected to continue (18).

Another type of THV that was widely pursued in the 1990s is the stentless valve. This valve type, intended for aortic valve replacement only, is made from porcine aortic roots that are fixed in glutaraldehyde, but do not have any rigid support materials added. The surgeon must attach both ends of the device into the patient's aorta without the aid of a sewing ring or support structures. The intended benefit of these valves is better hemodynamics because of their flexibility and lack of a sewing ring, and greater durability due to lower stresses in the tissue. However, they are not used as extensively as other THVs because they are more difficult to implant, which results in extended surgery time. Examples of two stentless valves used clinically are shown in Fig. 7. Similarly, human aortic roots can be removed from cadavers and processed with preservative techniques to make an implantable replacement. These valves, called homografts or allografts, have all the perceived benefits of stentless valves, but are just as difficult to implant, and long-term clinical results have been mixed (5). The primary commercial source of homografts is Cryolife, which cryogenically preserves the aortic roots.

There have been numerous attempts at fabricating prosthetic valves from polymers but, to date, none have achieved clinical success. Polymer valves are conceptually attractive because they would be flexible, with a reproducible geometry and relatively economical and straightforward to manufacture, and ideally would be more durable than tissue valves, while not requiring chronic anticoagulation like mechanical valves. But design difficulties and calcification have prevented these valves from realizing their full potential (20). Like polymer valves, tissue-engineered valves have many theoretical advantages over current mechanical and tissue valves. As a result, processes for growing a valve *in vitro* from human cells seeded on a scaffold has been an area of active research (21–23), but has also not yet produced a clinically viable product.

Figure 7. Examples of two stentless porcine heart valves: A Prima stentless valve. (Courtesy of Edwards Lifesciences, Irvine, Ca.) A Freestyle stentless valve. (Courtesy of Medtronic, Inc.)

IN VITRO TESTING

In vitro evaluations of prosthetic heart valves are performed to understand the detailed flow characteristics of a given design. The flow area, the amount of leakage, the ultrasound compatibility, the presence of deleterious flow patterns, the velocity magnitude of regurgitant jets, the motion of the leaflets during forward flow, and the position of the leaflets during closure can all be assessed *in vitro* and be used to improve valve designs and assess the appropriateness for human implantation.

Equipment

To evaluate the hemodynamic performance of prosthetic valves, a pulse duplicator or mock flow loop is implemented to act as an artificial left ventricle, and therefore must be able to simulate the pressure and flow waveforms shown in

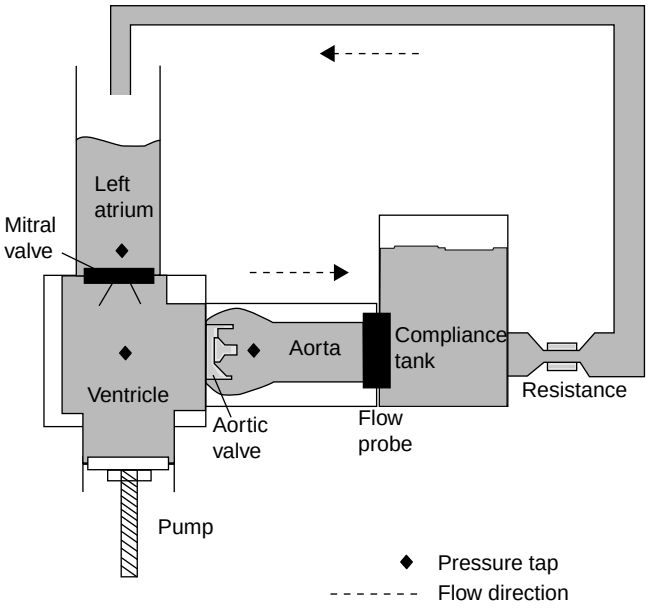


Figure 8. Schematic of a pulse duplicator for *in vitro* heart valve testing.

Figs. 2 and 3 over a range of physiologic hemodynamic conditions.

Figure 8 is a schematic representation of a pulse duplicator for *in vitro* heart valve testing. Generally, the left atrium is simulated with an open or compliant reservoir that maintains a static fluid height above the mitral valve so as to provide a diastolic filling pressure of ~10–15 mmHg (1.33–1.99 kPa). The mitral valve should open directly into the left ventricle as it does anatomically. The left ventricle can be simulated with a rigid chamber with a volume similar to the human left ventricle (~70–100 mL), although some pulse duplicators utilize a flexible, ventricular-shaped sac that can be hydraulically compressed to simulate the squeezing action of the heart (24,25). The aortic valve should be mounted at the outflow of the ventricular chamber so that flow enters the valve directly from the chamber. Tubing between the ventricle and the valve should be avoided as it will cause higher than physiologic velocities upstream of the valve. The flow exiting the aortic valve enters a tubular chamber with dimensions similar to the native aorta, including three sinuses and a diameter appropriate for the aortic valve being tested. Although not necessary for valve closure, the presence of sinuses allows for more realistic flow patterns behind the leaflets during systole.

The system should be instrumented to allow instantaneous pressure measurements in the left atrium, left ventricle, and aorta and flow measurements through both valves. The flowmeter must be able to accurately measure both forward flow, which can reach peak valves of $30\,L \cdot min^{-1}$, and leakage flows, which may be on the order of $1\,mL \cdot s$. Test fluids can be either saline or a blood analog fluid with a density and viscosity similar to that of blood, $\sim 1.1\,g \cdot mL^{-1}$ and 3.5 cp, respectively. A mixture of water and glycerin is the most common blood analog fluid. Because blood is a non-Newtonian fluid (i.e., its viscosity varies with shear rate), polymeric solutions, which more closely mimic this property, have also been used for *in vitro* valve testing (26).

In order to produce physiologic waveforms, compliant and resistive elements must be present downstream of the aorta. Compliant elements are expansive chambers that may utilize air, springs, flexible beams, or compliant tubing to absorb fluid volume during the systolic stroke. The compliant element should be located as close as possible to the aortic valve to minimize "ringing" in the pressure and flow waveforms. The impedance of the arterial system can be simulated with simple pinch clamp on flexible tubing, or a variably restrictive orifice.

The flow is driven with a pulsatile pumping system that interfaces with the left ventricular chamber. The pump must be able to create physiologic beat rates, flow rates, and systolic/diastolic ratios. One type of pumping system utilizes positive and negative air pressure, cycled with solenoid valves and timers, to drive a flexible diaphragm (27). Alternatively, a large, cam-driven syringe (piston–cylinder) pump can be employed with controlled piston displacement designed to produce the desired systolic and diastolic timing. Syringe pumps driven by computer-controlled servo motors are the optimal pumping systems because they allow the greatest amount of control over the motion of the piston and can best simulate complicated physiologic flow patterns (e.g., biphasic diastolic flow).

Valve mounting in a pulse duplicator can be accomplished by mechanical compression of the sewing ring, or by suturing the ring to a gasket. However, any valve that is flexible and mechanically coupled to the native anatomy during implantation (e.g., a stentless aortic valve) must be tested in a valve chamber that mimics the dynamic motion of the valve *in vivo*, because this may affect hemodynamic performance. Standardized testing guidelines suggest that stentless valves be tested in flexible chambers that undergo a 4 and 16% change in diameter per 40 mmHg (5.33 kPa) pressure change (28).

Test Conditions and Data Analysis

Hemodynamic conditions considered typical for an adult would be the following: cardiac output: $5\,L \cdot min^{-1}$; left atrial pressure: 10–15 mmHg (1.33–1.99 kPa); left ventricular pressure: 120/0 mmHg (systolic/diastolic) (15.99/0 g); aortic pressure: 120/80 mmHg (systolic/diastolic) (15.99/10.66 kPa); heart rate: 70 $beats \cdot min^{-1}$.

Although flow and pressure conditions are often quantified according to mean, peak, and minimum values this way, the shape of the pressure and flow waveforms are also important in determining the appropriateness and quality of the test condition. Excess "ringing", large pressure spikes, and square waveforms are examples of unphysiologic waveform features that may be seen *in vitro*.

A thorough *in vitro* investigation of heart valve performance should also include different flow, heart rate, and pressure conditions that correspond to the range of physiologic hemodynamic conditions that the valve will experience *in vivo*. At rest, the heart rate may decrease to $45\,beats \cdot min^{-1}$ or lower, while during exercise it may increase to >120 $beats \cdot min^{-1}$; cardiac output will vary

accordingly at these conditions, typically ranging from 2 to $7\,L \cdot min^{-1}$. It is at these extremes that leaflet dynamics, for example, may change and reveal performance limitations of the valve.

Typically, a minimum of 10 cycles, or heart beats of flow, inflow pressure and outflow pressure are collected for analysis at each test condition. The waveforms are analyzed by first identifying the key crossing points that define the start and end of forward flow, regurgitant flow, and positive (forward flow) pressure gradient (see Figs. 3 and 4). These pressure and flow data are used to calculate the key variables that define the hemodynamic performance of a valve: the pressure gradient (ΔP), the effective orifice area (EOA), and regurgitant volume.

The pressure gradient is the most basic measure of valve resistance to flow. *In vivo*, a prosthetic aortic valve with a high pressure gradient will force the left ventricle to expend more energy than one with a low pressure gradient. The regions of positive pressure gradient for aortic and mitral valves are shown in Figs. 3 and 4. These pressures can be measured in a pulse duplicator through wall taps in the testing chambers or through a catheter inserted into the flow. The pressure gradient will increase with decreasing valve size and, for a given valve size, increasing flow rate. It has also been shown that aortic pressure gradient measurements can be dependent on left ventricular ejection time (29) and aortic pressure (30).

The EOA is not a physical valve dimension, but a calculation of the minimal flow area allowed by the open valve. Under pulsatile flow conditions, it is calculated as $Q_{rms}/[51.6(\Delta P)^{1/2}]$, where Q_{rms} is the root-mean square of the average flow rate through the valve. The EOA is a more useful measure than geometric orifice dimensions because mechanical valves, with occluders that occupy the orifice, restrict the orifice flow in variable ways based on the degree of opening and flow rate. Similarly, the extent of tissue valve leaflet opening will vary based on leaflet stiffness and flow rate. A method of dynamically measuring the actual valve area in an *in vitro* tester by measuring the amount of light that passes through it has also been proposed and tested (31).

All valves move some volume of fluid back into the inflow chamber during closure. The closing volume of a prosthetic valve is calculated from the area under the flow curve immediately after valve closure as shown in Figs. 3 and 4. This volume is a function of the valve area and the travel distance of the leaflets during closure. Although it will vary based on valve design, the closing volume is relatively small and typically does not have a significant hemodynamic effect. Leakage volume, by contrast, is the volume of blood that passes through the valve during closure, and is a critical measure of valve performance. Excessive leakage reduces the efficiency of the heart and can cause progressive deterioration of ventricular function. Ideally, the leakage volume through a valve is zero, although most mechanical valves allow some leakage of blood in order to provide "washout" of mechanical junctions. Tissue valves do not allow any leakage as long as there is adequate apposition, or coaptation, of the leaflets.

The total regurgitant volume is the sum of the closing and leakage volumes and represents the total fluid loss during one valve closure. Regurgition can also be expressed as a percentage of forward volume. Clinically, the ejection fraction, the ratio of the ejected blood volume to the left ventricular volume, is used to assess the hemodynamic severity of regurgitation. As with other hemodynamic variables, regurgitant volumes will vary with valve design, valve size, flow rate, pressure, and heart rate.

When assessing prosthetic valve regurgitation it is important to discriminate between transvalvular regurgitation and paravalvular regurgitation. Transvalvular regurgitation occurs through the valve mechanism, such as past the hinges or occluder gaps in mechanical valves or through regions of inadequate leaflet coaptation in tissue valves. Because prosthetic valves are sewn in place with a porous cloth sewing ring, leakage can also occur through the sewing ring or spaces created by inadequate fixation to the annulus. This is paravalvular leakage, occurring around or adjacent to the valve, and should be differentiated from transvalvular leakage *In vitro*, test valves should be sealed to eliminate paravalvular leakage so that transvalvular leakage can be measured independently.

In vitro performance comparisons of different prosthetic valve designs are complicated by variations between labeled valve size and actual valve size (32). Most commercial prosthetic valves are sized according to diameter in 2 mm increments and typically range from 19 to 29 mm for aortic valves, and 25 to 31 mm for mitral valves. However, the diameter measurement is made at different locations based on manufacturer and valve type. The performance index (PI), which is the ratio of EOA to actual sewing ring area (33), is calculated in order to control for these variations and allow for more reliable valve-to-valve comparisons. Clinically, an EOA Index (EOAI) is calculated by dividing the EOA by the patient's body surface area, in order to normalize EOA values based on patient size and, indirectly, hemodynamic requirements.

Due to the variability of pulse duplicator systems and the sensitivity of test results on hemodynamic conditions, comparing valves evaluated in different test systems can also be problematic (34). Differences in pressure and flow waveform shapes, drive types, location of pressure taps, chamber geometries, and other system configuration characteristics can cause variations in measured valve performance. Comparative *in vitro* studies of different mechanical valves (35,36) and tissue valves (37) tested in the same pulse duplicator under the same conditions can be found in the literature. Figures 9–10 show *in vitro* pressure drop results for various tissue valves from one of these studies.

Doppler Ultrasound

Doppler echocardiography is the most widely used modality for assessing prosthetic valve performance *in vivo* and can be useful for *in vitro* studies as well. *In vitro* Doppler and imaging assessments are an important part of valve evaluations prior to human implants because they may reveal signature ultrasound features (e.g., acoustic shadowing, eccentric flow profiles, jet morphology), which are unique to the valve design and different from a native valve. It is important for clinical sonographers to understand these

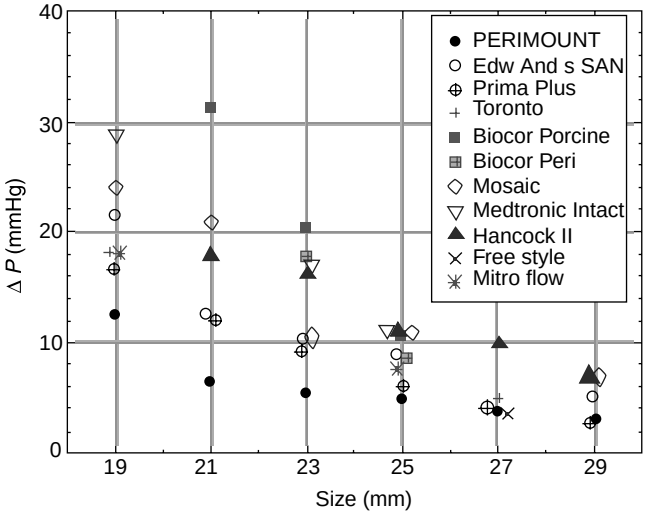


Figure 9. *In Vitro* pressure drop data of various prosthetic aortic valves. (Reprinted with permission from *The Journal of Heart Valve Disease*.)

features so they can distinguish between normal and abnormal valve function. In addition, two-dimensional (2D) echo images of the valve can be used to obtain an axial cross-sectional view of the valve during cycling that can be useful in assessing the motion and coaptation morphology of tissue leaflets. And although a flowmeter can be used to quantify the amount of regurgitation through a valve, Color Doppler ultrasound allows for visualization of the size and shape of any jets and detecting their point of origin.

Doppler measurements and 2D imaging can be performed *in vitro* providing there is an acoustic window for the ultrasound beam that allows the transducer to be aligned roughly perpendicular to the axis of valve flow and does not attenuate the acoustic signal. It is also

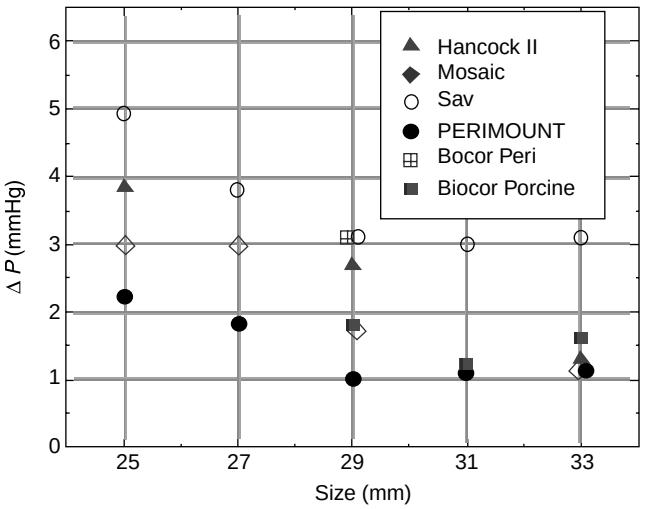


Figure 10. *In Vitro* pressure drop data of various prosthetic mitral valves. (Reprinted with permission from *The Journal of Heart Valve Disease*.)

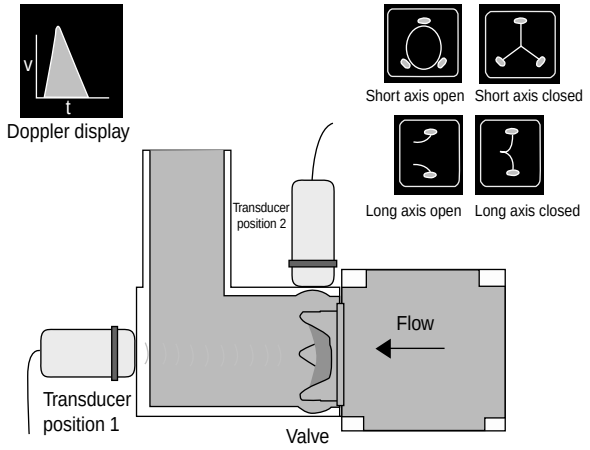


Figure 11. Ultrasound measurements in a pulse duplicator. In transducer position 1, the transducer is aligned with the flow and emits a sound wave that reflects off particles in the fluid; the return signal is measured by the transducer and the Doppler equation is used to calculate velocity. Transducer position 2 is used to obtain short-axis (cross-sectional) and long axis images of the valve as it opens and closes.

necessary to add small, insoluble particles to the test fluid to provide acoustic scatter for Doppler measurements and Color Doppler imaging. Figure 11 shows how an ultrasound transducer would be positioned on a pulse duplicator and the types of images that can be obtained.

Doppler measurements of pressure gradients can be performed *in vitro* to assess the hemodynamic performance of a valve as it would be done in a patient. During forward flow, continuous wave Doppler measures a real-time spectrum of velocities across the open valve. Standard echocardiographic software allows the user to acquire and analyze the spectrum and determine an average velocity.

Clinically, the average velocity value is used to compute pressure gradient using a simplified version of the Bernoulli equation. Ignoring viscous losses, acceleration effects, and changes in height, the Bernoulli relationship between two points, 1 and 2, on a streamline is:

$$P_1 + \frac{1}{2}\rho v_1^2 = P_2 + \frac{1}{2}\rho v_2^2 \qquad (1)$$

where P is the pressure, is the fluid density, and v is the velocity.

Rearranging Eq. 1 and expressing the upstream, or proximal valve velocity as v_p, and the downstream, or distal velocity as v_d, the pressure gradient ($P_1 - P_2 = \Delta P$) across a valve can be expressed as:

$$\Delta P = \frac{1}{2}\rho(v_d^2 - v_p^2) \qquad (2)$$

Doppler ultrasound machines typically report velocity in units of meter per second ($\mathrm{m \cdot s^{-1}}$) and pressure in units of millimeter of mercury (mmHg). Assuming a blood density of $\sim 1.1\,\mathrm{g \cdot mL^{-1}}$ and applying the appropriate unit conversions, the 1/2 term is approximated as $4\,\mathrm{mmHg \cdot m^{-2} \cdot s^{-2}}$. A further simplification can be made if the proximal velocity term (the velocity at the left ventricular outflow tract for aortic valves) is small compared to the distal velocity.

Neglecting the proximal velocity term yields the equation used by ultrasound software programs to compute pressure gradient from continuous wave velocity measurements:

$$\Delta P = 4(v^2) \quad (3)$$

If the proximal velocity term is needed, it can be measured with pulsed wave Doppler ultrasound, which, unlike continuous wave Doppler, provides velocity spectra at a user-defined location in space (e.g., the left ventricular outflow tract).

Prior to the use of Doppler, hemodynamic assessment of prosthetic valves *in vivo* was performed by placing a pressure-measuring catheter into the heart. Comparisons between pressure gradients measured by catheter and by Doppler typically show some discrepancy between the two methods, causing some concern as to which method is more accurate. Several groups have performed *in vitro* and *in vivo* studies to compare and contrast catheter-based measurements and Doppler measurements (38–41). These comparisons are complicated by the fact that the two techniques use fundamentally different methods to arrive at the pressure gradient and they each have individual sources of approximation and error.

The primary valve-related reason for differences in Doppler and catheter measurements of pressure gradient is pressure recovery. Pressure recovery occurs when some portion of the downstream kinetic energy is converted back to potential energy (i.e., pressure). Like flow through a Venturi nozzle, the lowest pressure and highest velocity through a heart valve will occur at the narrowest point, which is typically the region immediately downstream of the valve. If the flow out of the valve is allowed to gradually expand and remains attached to the walls, there will be an increase, or recovery, of pressure further downstream relative to the pressure at the valve exit. Due to this local variation in pressure, the location of pressure measurement becomes critical. Because continuous wave Doppler measures velocities all along the beam path, it will detect the highest velocity (regardless of its location), which will then be used to calculate the pressure gradient. In contrast, catheters take direct pressure measurement at a specific location; if that location is not at the point of lowest pressure, catheter measurements of pressure gradient will tend to be lower than those derived from Doppler.

Pressure measurements across bileaflet MHVs (e.g., the St. Jude Mechanical) are further complicated because, during opening, the two side orifices are larger than the central orifice, resulting in a nonuniform velocity profile. This may be one reason why better agreement between *in vitro* Doppler and catheter measurements has been found for a tissue valve, which has an unobstructed central orifice, than for a bileaflet mechanical valve (39).

Flow Visualization

Clear chambers that provide visual access to the valve as well as the flow fields in its vicinity are an important feature of *in vitro* test systems. The use of video to asses leaflet dynamics can provide important information about valve performance that cannot be obtained from hemodynamic measurements or *in vivo* studies. Tissue contact with a stent, poor coaptation, asymmetric or asynchronous leaflet closure, and leaflet rebound are all visual clues that either the valve design is inadequate or the test conditions are inappropriate.

Flow field visualization and measurements are equally important because thrombotic complications *in vivo* can be caused by unphysiologic, valve-induced flow patterns. Both flow stasis and high velocity leakage jets (and associated turbulent stresses) can trigger thrombosis and subsequent thromboemboli. *In vitro* pulsatile flow studies using blood are not a practical means of determining the thrombogenicity of a valve design because of the complexity and sensitivity of the coagulation process. Although the use of milk as an enzymatically coagulable test fluid has been reported (42–44), most *in vitro* assessments of a valve's thrombogenicity are made indirectly based on its flow characteristics.

Illuminating the flow field of interest with a sheet of laser light and seeding the fluid with neutrally buoyant particles will allow for qualitative visual assessment of flow patterns (i.e, uniformity and direction of forward flow, presence of vortices or recirculation zones, areas of washout during closure.) The motion of the particles can be analyzed quantitatively with digital particle image velocimetry (DPIV), which uses the translocation of seeded particles in consecutive, laser-illuminated video frames to compute the velocity of many particles at one point in time. With these numerous velocity measurements, a velocity map of the flow field can be created from any point in the cardiac cycle.

Velocity measurements can also be performed with laser Doppler velocimetry (LDV), which allows good temporal resolution of velocity, but only at a single point. Multiple point measurements at different locations can be made sequentially and then compiled in order to construct a phase averaged velocity profile of the entire flow field. Simultaneous LDV and DPIV measurements have been performed on a MHV in an attempt to integrate the relative benefits of each method (45). Thorough flow field measurements of several prosthetic heart valves, both tissue and mechanical, have been published by Yoganathan and coworkers (46,47). During forward flow, MHVs are seen to have disrupted or eccentric velocity profiles reflective of the open occluder position, while THVs tend to have more uniform, unobstructed flow profiles.

Since blood cell damage will likely occur above some critical shear stress threshold, *in vitro* velocity data is used to calculate shear stresses created throughout the flow cycle, and indirectly assess the potential for hemolysis and platelet activation *in vivo*. The MHV leakage jets, in particular, have the potential to create high shear stresses in blood (48). It is difficult, however, for *in vitro* flow characterization studies to be conclusive with regard to blood damage because of the myriad of variables that interact to trigger coagulation. In addition to the magnitude of shear stresses, the exposure time to those stresses, as well as flow stagnation, material surface interactions, and patient factors can contribute to prosthetic valve thrombosis.

Design Specific Testing

Guidelines for *in vitro* testing heart valves prior to human use were originally introduced by the U.S. Food and Drug

Administration (FDA) in 1982. A revised document, that included guidelines for testing stentless valves, was introduced in 1994 (28). The International Standards Organization publishes a similar set of guidelines for heart valve evaluation, including the equipment and data requirements for *in vitro* studies (49). Many of the testing techniques in use today are motivated by and in response to these guidelines. However, standardized guidelines are often insufficient for testing new and innovative designs, since each new valve design will have unique features that may require special testing methods to evaluate. A risk analysis or failure mode analysis can be used to assess the need for *in vitro* testing beyond that described in the standards.

In addition to predictive studies of *in vivo* performance, *in vitro* studies may be conducted retrospectively, in order to elucidate a particular failure mode seen after a valve has been used clinically. These types of studies typically require the development of new or improved testing methodologies to investigate a particular phenomenon, as shown in the following examples.

In 1988, structural failures of the Baxter Duromedics MHV were reported in several patients (50). Surface pitting was observed on the pyrolytic carbon leaflets of the explanted valves, suggestive of cavitation-induced erosion. Cavitation occurs in a fluid when the pressure drops rapidly below the vapor pressure of the fluid, causing the formation of small vaporous cavities, which collapse violently when the pressure increases. Many *in vitro* studies were conducted, employing novel measurement techniques, to assess the propensity for cavitation to occur during leaflet closure of mechanical heart valves (51,52) and cavitation testing is now part of U.S. Food and Drug Administration (FDA) required preclinical testing for all new MHVs.

In the mid-1990s, the Medtronic parallel bileaflet valve experienced an unanticipated number of thrombosed valves in early clinical trials (53). Explanted valves were observed to have clot formation in the hinge region of the valve, indicating a flow-related problem in this vicinity (54). These failures occurred despite the full set of required tests having been conducted with apparently satisfactory results. Prototype valves with clear housings were constructed to allow flow visualization studies in the region of the hinge mechanism (55). Results of these studies suggested the geometry of the hinge created stagnant flow regions which may have been responsible for the clinical failures.

Clinical studies of patients with stentless valves revealed that some developed aortic regurgitation several years after implantation (56,57). Because these valves lack a support structure, it was believed that age-related dilation of the aorta strained the outflow edge of the valve, which lead to insufficient leaflet coaptation. Although FDA testing requirements for stentless valves included flow testing at elevated aortic pressures in compliant chambers, testing of mechanical dilation without an increase in pressure was not required. An *in vitro* study using canine hearts showed that simply pulling the aorta outward at the commisssures with sutures prevented the aortic leaflets from closing in the center (58). This study helped confirm the need to band the aorta during implantation of some stentless valves in order to prevent later dilation.

In some cases, *in vitro* testing may also be employed to test the actual valve that was explanted from a patient. Depending on the age and condition of the valve, it may be mounted in a pulse duplicator and studied for signs of performance, structural or manufacturing abnormalities responsible for an adverse event in a patient.

Finally, *in vitro* flow studies may be conducted using valves excised from animal hearts in order to study the flow or mechanical dynamics of the native valve that in turn may be used to design improved prosthetic devices. Fresh porcine aortic roots have been installed in pulse duplicators to study the motion of the aortic valve as well to serve as a testing chamber for prosthetic valves (59–61). The porcine mitral apparatus, including chords and papillary muscles, has also been mounted and tested *in vitro* (62,63).

FUTURE DIRECTIONS

After more than four decades of prosthetic heart valve development, surgeons have come to rely on just a few valve designs, both mechanical and tissue, for the large majority of implants. These valves have achieved widespread use because their performance is reliable and their limitations and failure modes are known and reasonably predictable. None of these valves are ideal, however, and new designs that try to improve upon the state of the art will continue to emerge.

Although polymer valves and tissue-engineered valves still hold promise, the greatest change to be expected in the near future is not the valve itself but the way it is implanted in the heart. The success of catheter-based technologies in treating other heart diseases (e.g., coronary stents, septal defect closure devices) has inspired the pursuit and clinical evaluation of a prosthetic heart valve that can be placed in the beating heart with a catheter (64–66). This technology is attractive because it would not require opening the chest and stopping the heart, nor the use of cardiopulmonary bypass equipment.

Catheter-delivered valves will still require all the testing and analysis described above, but will also necessitate new equipment and methodologies that take into account their unique delivery and implantation method. For example, the actual delivery and deployment of these valves may first need to be simulated under *in vitro* pulsatile conditions, requiring testers that simulate the entry point and anatomy of the delivery path. Once deployed, the ability of the valve to seal around its perimeter and remain in place without migrating must be evaluated, which will require deployment chambers with the appropriate surface properties, mechanical properties, and dynamic motion. Also, current *in vitro* valve tests rarely simulate coronary flow out of the sinuses. But when testing valves that are placed in the aortic annulus under image guidance, rather than direct visualization, it will be important to evaluate valve designs in terms of their propensity to block coronary flow.

With aging populations in the United States and Europe, heart valve disease is likely to remain a significant and prevalent problem. Better, more efficient prosthetic heart

valve technology will surely emerge to address this need and *in vitro* testing technologies will need to keep pace to continue helping to improve these devices and ensure their safety and efficacy for human use.

BIBLIOGRAPHY

Cited References

1. Heart Disease and Stroke Statistics—2004 Update. American Heart Association, 2004.
2. Starr A, Edwards ML. Mitral replacement: clinical experience with a ball-cage prosthesis. Ann Surg 1961;154:726–740.
3. Starr A, Edwards ML. Mitral replacement: late results with a ball valve prosthesis. J Cardiovasc Surg 1963;45:435–447.
4. Pluth JR. The Starr valve revisited. Ann Thor Surg 1991;51: 333–334.
5. Otto CM. Valvular Heart Disease, 2nd ed. Philadelphia: Saunders; 2004.
6. Anderson RH. Clinical anatomy of the aortic root. Heart 2000;84:670–673.
7. Thubrikar M. The Aortic Valve. Boca Raton (FL): CRC Press; 1990.
8. Antunes MJ. Functional Anatomy of the Mitral Valve. In: Barlow JB, editor. Perspectives on the Mitral Valve. Philadelphia: FA Davis; 1987.
9. Kalmanson D. The mitral valve: a pluridisciplinary approach. London: Arnold; 1976.
10. Zipes DP. Braunwald's Heart Disease. 7th ed. Philadelphia: WB Saunders; 2005.
11. Thubrikar MJ, Heckman JL, Nolan SP. High speed cineradiographic study of aortic valve leaflet motion. J Heart Valve Dis 1993;2(6):653–661.
12. Nichols WW, O'Rourke MF. McDonald's Blood Flow in Arteries, 4th ed. London: Arnold; 1998.
13. Bonow RO, et al. ACC/AHA guidelines for the management of patients with valvular heart disease: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am College Cardiol 1998;32: 1486–1588.
14. Levine RA, et al. Three-Dimensional echocardiographic reconstruction of the mitral valve, with implications for the diagnosis of mitral valve prolapse. Circulation 1989;80: 589–598.
15. Yellin EL, et al. Mechanisms of mitral valve motion during diastole. Am J Physiol 1981;241:H389–H400.
16. Komeda M, et al. Papillary muscle-left ventricular wall "complex". J Thor Cardiovasc Surg, 1997;113:292–301.
17. Ellis JT, et al. An in vitro investigation of the retrograde flow fields of two bileaflet mechanical heart valves. J Heart Valve Dis 1996;5(6):600–606.
18. Anonymous Focus on Heart Valves, Medical Device and Diagnostic Industry, vol. 126, March 2004.
19. Banbury MK, et al., Hemodynamic stability during 17 years of the Carpentier-Edwards aortic pericardial bioprosthesis. Ann Thor Surg 2002;73(5):1460–1465.
20. Hyde JA, Chinn JA, Phillips RE., Jr., Polymer Heart Valves. J Heart Valve Dis 1999;8(3):331–339.
21. Hoerstup SP, Kadner A, Melnitchouk S. Tissue engineering of a functional trileaflet heart valve from human marrow stromal cells. Circulation 2002;106 (Suppl. I):143–150.
22. Cebatori S, Mertsching H, Kallenbach K. Construction of autologous human heart valves based on an acellular allograft matrix. Circulation 2002;106 (Suppl. I):63–68.
23. Bertipaglia B, Ortolani F, Petrelli L. Cell cellularization of porcine aortic valve and decellularized leaflets repopulated with aortic valve interstitial cells: the VESALIO project. Ann Thor Surg 2003;75:1274–1282.
24. Reul H, Minamitani H, Runge J. A hydraulic analog of the systemic and pulmonary circulation for testing artificial hearts. Proc ESAO 1975;2:120.
25. Scotten LN, et al. New tilting disc cardiac valve prostheses. J Thor Cardiovasc Surg 1986;82:136–146.
26. Pohl M, et al. *In vitro* testing of artificial heart valves:comparison between Newtonian and non-Newtonian fluids. Arti Organs 1996;20(1):37–46.
27. Weiting DW. Dynamic flow characteristics of heart valves dissertation. University of Texas, Austin, 1969.
28. U.S. Department of Health and Human Services, Public Health Service, Food and Drug Administration, Center for Devices and Radiological Health, Replacement Heart Valve Guidance, 1994.
29. Kadem L, et al. Independent contribution of left ventricular ejection time to the mean gradient in aortic stenosis. J. Heart Valve Dis 2002;11(5):615–623.
30. Razzolini R, et al. Transaortic gradient is pressure-dependent in a pulsatile model of the circulation. J Heart Valve Dis 1999;8(3):279–283.
31. Scotten LN, Walker DK. New laboratory technique measures projected dynamic area of prosthetic heart valves. J Heart Valve Dis 2004;13(1):120–132.
32. Cochran RP, Kunzelman KS. Discrepancies between labeled and actual dimensions of prosthetic valves and sizers. J Cardiovasc Surg 1996;11:318–324.
33. Stewart SFC, Bushar HF. Improved statistical characterization of prosthetic heart valve hemodynamics using a performance index and regression analysis. J Heart Valve Dis 2002;11:270–274.
34. Van Auker MD, Strom JA. Inter-Laboratory comparisons: approaching a new standard for prosthetic heart valve testing *in vitro*. J Heart Valve Dis 1999;8(4):384–391.
35. Walker DK, Scotten LN. A database obtained from *in vitro* function testing of mechanical heart valves. J Heart Valve Dis 1994;3(5):561–570.
36. Fisher J. Comparative study of the hydrodynamic function of six size 19 mm bileaflet heart valves. Eur J Cardiothora Surg 1995;9(12):692–695.
37. Marquez S, Hon RT, Yoganathan AP. Comparative hemodynamic evaluation of bioprosthetic heart valves. J Heart Valve Dis 2001;10(6):802–811.
38. Vandervoort PM, et al. Pressure recovery in bileaflet heart valve prostheses. Circulation 1995;92:3464–3472.
39. Baumgartner H, et al. Discrepancies between Doppler and catheter gradients in aortic prosthetic valves *in vitro*. A manifestation of localized gradients and pressure recovery. Circulation 1990;82:1467–1475.
40. Burstow DJ, et al. Continuous wave Doppler echocardiographic measurement of prosthetic valve gradients. A simultaneous Doppler-catheter correlative study. Circulation 1989; 80(3):504–514.
41. Bech-Hanssen O, et al. Assessment of effective orifice area of prosthetic valves with Doppler echocardiography: and *in vivo* and *in vitro* study. J Thorac Cardiovasc Surg 2001;122(2): 287–295.
42. Lewis JM, Macleod N. A blood analogue for the experimental study of flow-related thrombosis at prosthetic heart valves. Cardiovasc Res 1983;17(8):466–475.
43. Keggen LA, et al. The use of enzyme activated milk for *in vitro* simulation of prosthetic valve thrombosis. J Heart Valve Dis 1996;5(1):74–83.
44. Martin AJ, Christy JR. An *in vitro* technique for assessment of thrombogenicity in mechanical prosthetic cardiac valves: evaluation with a range of valve types. J Heart Valve Dis 2004;13(3):509–520.

45. Kini V, et al. Integrating Particle Image Velocimetry and Laser Doppler Velocimetry Measurements of the Regurgitant Flow Field Past Mechanical Heart Valves. Art Organs 2001;25(2):136.
46. Woo YR, Yoganathan AP. *In vitro* pulsatile flow velocity and shear stress measurements in the vicinity of mechanical mitral heart valve prostheses. J Biomechan 1986;19(1): 39–51.
47. Yoganathan AP, Woo YR, Sung HW. Turbulent shear stress measurements in the vicinity of aortic heart valve prostheses. J Biomechan 1986;19(6):433–442.
48. Ellis JT, et al. An in vitro investigation of the retrograde flow fields of two bileaflet mechanical heart valves. J Heart Valves Dis 1996;5:600–606.
49. International Standard 5840:2004(E), Cardiovascular Implants—Cardiac Valve Prostheses, International Standards Organization, June 22, 2004.
50. Quijano RC. Edwards-Duromedic dysfunctional analysis, Proceedings of Cardiostimulation; 1988.
51. Hwang NH. Cavitation potential of pyrolytic carbon heart valve prostheses: a review and current status. J Heart Valve Dis 1998;7:140–150.
52. Hwang NHC, editor. Cavitation in mechanical heart valves, Proceedings of the First International Symposium. J Heart Valve Dis 3(Suppl. I), 1994.
53. Bodnar E. The Medtronic Parallel valve and the lessons learned. J Heart Valve Dis 1996;5:572–673.
54. Ellis JT, et al. Velocity measurements and flow patterns within the hinge region of a Medtronic Parallel bileaflet mechanical valve with clear housing. J Heart Valve Dis 1996;5:591–599.
55. Gross JM, et al. A microstructural flow analysis within a bileaflet mechanical heart valve hinge. J Heart Valve Dis 1996;5:581–590.
56. Jin XY, Westaby S. Aortic root geometry and stentless porcine valve competence. Seminars Thorac Cardiovasc Surg 1999;11(Suppl I):145–150.
57. David TE, et al. Dilation of the sinotubular junction causes aortic insufficiency after aortic valve replacement with the Tornoto SPV bioprosthesis. J Thorac Cardiovascu Surg 2001;122(5):929–934.
58. Furukawa K, et al. Does dilation of the sinotubular junction cause aortic regurgitation. Ann Thorac Surg 1999;68:949–954.
59. Thubrikar MJ, et al. The influence of sizing on the dynamic function of the free-hand implanted porcine aortic homograft: an in vitro study. J Heart Valve Dis 1999;8(3): 242–253.
60. Revanna P, Fisher J, Watterson KG. The influence of free hand suturing technique and zero pressure fixation on the hydrodynamic function of aortic root and aortic valve leaflets. Eur J Cardiothor Surg 1997;11(2):280–286.
61. Jennings LM, et al. Hydrodynamic function of the second-generation mitroflow pericardial bioprosthesis. Ann Thorac Surg 2002;74:63–68.
62. Jensen MO, et al. Harvested porcine mitral xenograft fixation: impact on fluid dynamic performance. J Heart Valve Dis 2001;10(1):111–124.
63. Jensen MO, Fontaine AA, Yoganathan AP. Improved *in vitro* quantification of the force exerted by the papillary muscle on the left ventricular wall: three-dimensional force vector measurement system. Ann Biomed Eng 2001;29(5):406–413.
64. Cribier A, et al. Early experience with percutaneous transcatheter implantation of heart valve prosthesis for the treatment of end-stage inoperable patients with calcific aortic stenosis. J Am College Cardiol 2004;43(4):698–703.
65. Boudjemline Y, Bonhoffer P. Steps toward percutaneous aortic valve replacement. Circulation 2002;105(6):775–778.
66. Lutter G, et al. Percutaneous aortic valve replacement: an experimental study. J Thorac Cardiovasc Surg 2002; 123(4): 768–776.

Reading List

Thubrikar M. The Aortic Valve. Boca Raton (FL): CRC Press; 1990.
Antunes MJ. Functional Anatomy of the Mitral Valve. In: Barlow JB, editor. Perspectives on the Mitral Valve, Philadelphia: FA Davis; 1987.
Kalmanson D. The mitral valve: a pluridisciplinary approach. London: Arnold; 1976.
Zipes DP. Braunwald's Heart Disease. 7th ed. Philadelphia: WB Saunders; 2005.
Otto CM, Valvular Heart Disease. 2nd ed. Philadelphia: WB Saunders; 2004.
Nichols WW, O'Rourke MF. McDonald's Blood Flow in Arteries. 4th ed. London: Arnold; 1998.
Reul H, Talukder N. Heart Valve Mechanics. In: Hwang NHC, et al., editors. Quantitative Cardiovascular Studies: Clinical and Research Applications of Engineering Principles. Vol. 12. Baltimore: University Park Press; 1979. p 527–564.
Nanda N. Doppler Echocardiography. 2nd ed. Philadelphia: Lee & Febiger; 1993.

See also BLOOD PRESSURE MEASUREMENT; MONITORING, HEMODYNAMIC.

HEART VALVES, PROSTHETIC

ROBERT MORE
Austin, Texas

INTRODUCTION

Blood flow through the four chambers of the normal human heart is controlled by four one-way valves. These valves open and close in response to local pressure gradients and flow during the cardiac cycle. The atrioventricular mitral and tricuspid valves open to admit blood flow from the atria into the ventricles during diastole and then close during systolic ventricular contraction to prevent backflow into the atria. The semilunar aortic and pulmonary valves open during systole to eject blood from the ventricles and then close during diastole to prevent backflow. Various pathological states, either congenital or acquired, may result in the failure of one or more of these valves. A valve may become stenotic, in which case forward flow through the valve is impaired, or a valve may become incompetent or regurgitant, closing improperly, which allows excessive backflow losses. Loss of valve function has a profound degenerative effect on quality of life and is ultimately life-threatening.

THE CHALLENGE

The goal for valve prosthesis design is to provide a functional substitute for a dysfunctional native valve that will endure for a patients' lifetime while requiring minimal chronic management. The development of open-heart surgery opened the technical possibility of valve replacement in the late 1950s and early 1960s. However, the practicality of an effective, durable artificial heart valve was to prove

elusive in the early 1960s because of problems with valve construction materials causing blood clotting, material fatigue, and degradation. Materials, related problems were greatly diminished when Carpentier (1) introduced the use of aldehyde preserved biological valves in the early 1960s. In 1963, Bokros discovered isotropic pyrolytic carbon (PyC), and shortly thereafter Gott and Bokros discovered the remarkable blood compatibility of PyC (2). These events led to two separate avenues of approach to the development of successful heart valve replacements, namely *biological valves*, derived from biological valves themselves or biological tissues, and *mechanical valves*, manufactured from synthetic materials. Significant process has occurred toward the goal of a lifelong valve substitute, last 30 years to the extent that valve replacement surgery is now a commonplace procedure worldwide. However, todays valve prostheses are nonideal and a tradeoff remains between long-term durability and the need for chronic anticoagulant management.

Although the history of early development of replacement heart valves is fascinating in it's own right, it is well chronicled and, thus, will not be addressed here (3–5). See http://members.evansville.net/ict/prostheticvalveimagegallery.htm for a chronology of valve designs. Rather, the focus here is to contemporary replacement heart valve designs, which is a story of interplay between advances in materials technology, design concepts, methods for evaluation, and regulatory issues.

EVOLUTION OF REGULATION

Prior to 1976, heart valve prostheses were unregulated devices—clinical studies, use, and introduction to the marketplace could be initiated without a formalized performance evaluation conducted under regulatory oversight. Regulation in the United States by the Food and Drug Administration (FDA), and by similar agencies in other countries, became necessary because not all valve prostheses were clinically successful. With the passage of the Medical Device Amendments in 1976, heart manufacturers were required to register with the FDA and to follow specific control procedures. Heart valve replacements were required to have premarket approval (PMA) by the FDA before they could be sold. To quote from the FDA website, http://www.FDA.gov, (the) "FDA's mission is to promote and protect the public health by helping safe and effective products reach the market in a timely way, and monitoring products for continued safety after they are in use".

The establishment of FDA oversight was a watershed event because it formalized the approval process and led to the development of criteria for evaluating new valve prosthesis designs. The preclinical performance tests and clinical performance evaluation criteria that have been developed are given in the FDA Document *Replacement Heart Valve Guidance - Draft Document, October 14, 1994* (6). This document may be downloaded from the FDA website at: http://www.fda.gov/cdrh/ode/3751.html. The document is deliberately labeled "Draft" because it is a "living" document that is periodically updated as heart valve technology evolves. Furthermore, requirements for postmarket monitoring exist to continually verify valve safety and effectiveness.

A parallel international effort gave rise to the International Standards Organization (ISO) Standard 5840, *Cardiovascular Implants-Cardiac Valve* (7), which is currently being harmonized with the European Committee for Standardization (CEN) to produce a European Standard EN prEN12006. A major benefit of the European Union (EU) is that member countries agree to recognize a single set of standards, regulatory processes and acceptance criteria. Prior to the establishment of the European Union, each individual country imposed their own standards and regulatory requirements. The ISO committee hopes to ultimately merge with or at least to philosophically harmonize the ISO, EN requirements with the FDA Guidance Document requirements in order to produce a single international standard.

The regulatory approval process for a new heart valve design consists of two general stages, a preclinical study and a subsequent clinical study. The preclinical study consists of *in vitro* tests, orthotopic animal implants, and a survey of manufacturing quality assurance systems. When the preclinical requirements are satisfied, the FDA or EU agency may issue an Investigational Device Exemption (IDE) that allows human clinical trials. Clinical trials are conducted according to specified study designs, patient numbers, and evaluation methods. Upon the successful completion of clinical trials, the regulatory agency such as the FDA may issue a PMA, or equivalent EU approval, that allows the open marketing of the valve design. As a result of the importance and complexity of regulatory requirements in heart valve design, evaluation, manufacture, marketing, and monitoring, major heart valve companies now have dedicated "Regulatory Affairs" departments.

DEMOGRAPHICS AND ETIOLOGY

Patients require heart valve replacement for a number of reasons, which vary by sex, valve position, and the prevailing levels of socioeconomic development. Valves become dysfunctional because of stenosis, inadequate forward flow when open or insufficient and excessive backflow when closed. A summary of typical patient data by sex, lesion, and disease state that led to dysfunction is given in Table 1. This data is taken from PMA studies of the On-X mechanical valve conducted in North America and Western Europe (8,9).

Patients in developing countries tend to require valves at a younger age (mean 33 years) and have a significantly higher incidence of rheumatic disease and endocarditis. The proportion of aortic implants relative to mitral implants is higher in North America and Western Europe, whereas the proportion of mitral implants and double-valve implants is higher in the developing countries.

CLINICAL PERFORMANCE

Regulatory agencies (6,7) gage the results of observational clinical (IDE) trials in terms of Objective Performance

Table 1. Summary of Typical Patient Data

Category	Aortic	Mitral
Mean age years	60	59
Sex	%	%
Male	66	38
Female	34	62
Lesion	%	%
Stenosis	47	13
Insufficiency	21	48
Mixed/other	32	39
Etiology (can be > one per patient)	%	%
Calcific	50	16
Degenerative	28	27
Rheumatic	13	38
Congenital	10	2
Endocarditis	4	7
Previous prosthetic valve dysfunction	3	3
Other	3	16

Criteria (OPC). Ultimately, the performance of a new investigational device is compared with existing devices. The OPCs are linearized complication rate levels. An IDE study uses statistical methods to demonstrate that observed complication rates for the investigational device are less than two times the OPC rate. The linearized rates are given in units of percent per patient year (6,7) in Table 2. A patient year is simply the total post implant follow-up duration in years for patients enrolled in the study. For example, 100 patients followed one year each post implant is 100 patient years (pt-yr). For the occurrence of complication events, the linearized rate is 100% $\times$ (number of events/number of patient years). Currently, a followup of 400 patient years is the minimum FDA regulatory requirement for each valve position, aortic and mitral each.

Thromboembolism (blood clotting-related strokes) and hemmohrage (excessive bleeding) are measures of the stability of the valve design relative to hemostasis and chronic anticoagulation therapy. Perivalvular leaks occur at the interface between the valve sewing cuff attachment mechanism and the cardiac tissue and are a measure of how well the valve design seals. Endocarditis is any infection involving the valve. A number of other complications, or morbid events, including hemolysis, unacceptable hemodynamics, explant, reoperation, and death are also monitored and compared on the basis of percent per patient year. Other presentations of data often used are Kaplan Meier "Survival or Freedom from complications" plots or tables in which the percentage of the study population that has survived, or has not experienced a given complication, is plotted vs. implant duration in years (10).

Table 2. Linearized Rates in Units of Percent Patient Year

Complication	Mechanical Valve (%/pt-yr)	Biological Valve (%/pt-yr)
Thromboembolism	3.0	2.5
Valve thrombosis	0.8	0.2
All hemmhorage	3.5	1.4
Major hemmhorage	1.5	0.9
All perivalvular leak	1.2	1.2
Major perivalvular leak	0.6	0.6
Endocarditis	1.2	1.2

HEMODYNAMIC EVALUATION

Methods of evaluating heart valve hemodynamic function, both natural and prosthetic, have advanced considerably in the past 30 years. With today's technology, most evaluations of valve performance can be performed noninvasively, and more information regarding valve design performance is available than ever before, both *in vivo* and *in vitro*. Early methods, such as ascultation, listening to valve sounds with a stethescope, have been greatly improved on, and invasive catheter techniques are no longer needed except in special cases.

Cine fluoroscopy involves recording a fluoroscopic examination with film or digital recording media. It provides a direct dynamic visualization of radio-opaque prosthetic valve features and, thus, can be used to identify a valve type and to assess function. Cine fluoroscopy used in combination with a catheter delivered radio-opaque contrast media, cineangiography, provides a means for basic flow visualization. Noninvasive echocardiography, however, has become the most common and important method for assessing valve function.

The physics of echocardiography are simple; a beam of ultrasound (1–5 MHz) is aimed into the patient's chest and, as the propagating sound wave encounters the different tissue layers, some of the energy is reflected back. The existence and depth of the reflecting tissue layers are then inferred from the time history of the reflections relative to the incident beam and the strength of the reflection. The reflection time history and strength from an angular sweep are then used to construct a 2D tomographic image of the underlying tissue structures. Thus, cardiac structures and their motions are readily visualized.

Echocardiography also produces an angiographic image for visualizing flow by using the moving red blood cells as a contrast medium. A Doppler shift occurs between the frequency of the incident sound and the sound reflected by the moving red blood cells, which is in proportion to the cells' velocity. This information is used to map flow direction and velocity throughout the cardiac cycle. Furthermore, the Doppler frequency shifts can be rendered audible to produce sound signatures. Thus, echocardiography and Doppler echocardiography provide powerful visual and audible diagnostic tools to interrogate cardiac dynamic structure and function.

Although the above description of echocardiography is admittedly brief, the real complexity of echocardiography lies in transducer mechanics, sophisticated electronics, and computational techniques required to produce near real-time dynamic 2D images and flow maps. Further details are readily available; for example, a very accessible tutorial can be found at the http://www.echoincontext.com/basicEcho.asp "Introduction to Echocardiography," prepared by Duke University, 2000.

For the purposes of heart valve evaluation, echocardiography and Doppler echocardiography give dynamic visualizations of valve component motion and reasonable measurements of flow velocities, flow distributions valve effective orifice areas, and transvalvular pressure differences. Definitions and technical details of valve evaluations can be found in Appendix M of the FDA Document *Replacement Heart Valve Guidance - Draft Document, October 14, 1994* (6,7). A summary of the important echocardiographic parameters for valve evaluation are listed below.

1. Visualizations of valve motion and structure: Provides a diagnostic tool for both the native valve and valve prostheses and can detect the presence of thrombosis.
2. Doppler measurements of transvalvular pressure difference (or gradient): A stenotic valve, with impaired forward flow, either native or prosthetic, will have an elevated transvalvular pressure difference. The transvalvular gradient (peak or mean) is determined by measuring the forward flow velocity through the valve. If the proximal velocity is greater than 1 $\mathrm{m \cdot s^{-1}}$, then it is also measured. Velocity values are diagnostic and can be used to derive an estimate of the transvalvular pressure difference using the complete (long) Bernoulli equation,

$$\Delta P = 4(V_2^2 - V_1^2),$$

where ΔP is the peak pressure difference mmHg, V_2 is the continuous-wave Doppler peak transvalvular velocity ($\mathrm{m \cdot s^{-1}}$), and V_1 is the proximal pulse-wave Doppler peak velocity ($\mathrm{m \cdot s^{-1}}$). If the proximal velocity is less than $1\,\mathrm{m \cdot s^{-1}}$, as is usual in the mitral position, the V_1 term can be dropped and, as such, the single-term equation is called the short Bernoulli equation. Forward flow velocity-time envelopes may be integrated to give mean velocity values or mean pressure values depending on the echo system's processing capabilities. The mean velocity values are substituted into the Bernoulli equation, or the mean transvalvular and proximal pressures are subtracted to give the mean transvalvular pressure difference.
3. Effective orifice area: A stenotic valve, either native or prosthetic, will have decreased effective orifice area. Effective orifice area is determined using the classical continuity equation:

$$\mathrm{EOV(cm^2)} = \mathrm{CSA}\, vti_1/vti_2$$

CSA is the cross-sectional area of the left ventricular outflow tract (LVOT) for the aortic position. Variables vti_1 and vti_2 are the velocity-time integrals at the pulse-wave proximal LVOT vti_1 and transvalvular continuous wave for the aortic position vti_2 . For the mitral position, CSA is the aortic cross-sectional area; vti_1 and vti_2 are the velocity-time integrals for the continuous-wave transvalvular velocities for the aortic valve, vti_1, and vti_2 for the mitral valve.
4. Regurgitation: Regurgitation or incompetence is graded as trivial, mild, moderate, or severe according to the regurgitant flow jet size expressed as a height or area relative to the LVOT area in the aortic position or left atrium in the mitral position. An incompetent, leaky valve will have severe regurgitation.

As a result of variations in patient body size and because the valve must fit the native valve annulus, prosthetic valve designs are prepared in annulus sizes ranging from an approximate 19 mm annulus diameter up to 33 mm diameter. However, valve sizes are not scalar in the sense that a size 19 valve or its sizer will often not be 19 mm. The valve and its associated sizer are supposed to fit a 19 mm annulus. For this reason, it is critical to use the sizers and associated instrumentation provided by the manufacturer for a specific valve design.

Typically, a design will have six sizes in 2 mm diameter increments. The size differences can consist of larger valve components or larger sewing cuffs. In general, it is desired to present the largest internal diameter possible for flow. Thus, it is important to recognize that hemodynamic quantities vary strongly with valve size in the aortic position and, for this reason, aortic hemodynamics comparisons must be interpreted according to size. One can find studies in the literature that report single values for a valve design that lump together and ignore the effects of valve size, the utility of such information is limited to the engineer, but meaningful to the cardiologist (11).

Patient body surface area (BSA) is often used to normalize valve effective orifice area, to give an orifice area $\mathrm{cm^2 \cdot m^{-2}}$. EOA is measured using echocardiography and BSA is calculated using the patient height and weight. The ratio of valve EOA to patient BSA is known as the indexed orifice area (IEOA)

Patient prosthesis mismatch has recently been recognized as a potential error in selecting valves for patients. It is possible to replace a diseased aortic valve with a valve prosthesis that is too small for the patient body size. A generally accepted criteria for sizing is that the indexed effective orifice area should not fall below a value of 0.85 (12).

Left ventricular mass regression is another contemporary measure of valve effectiveness. Ventricular dimensions are measured from echocardiography for patients preoperatively and postoperatively. When overloaded by a dysfunctional valve, the heart responds with an increase in mass. An effective prosthetic valve lowers the work load on the heart and thus causes a regression back toward normal size (13).

The information presented above provides a basic framework for evaluating and comparing valve design clinical performance: complication rates and hemodynamics. Complication rates are assessed using statistics presented as linearized rates, percent per patient year. However, the linearized rates tend to assume that the risk of a complication is constant with time (14), which may not hold for all complications. Furthermore, it is necessary to compare linearized rates determined for comparable population demographics, implant experience, and complication definition. Short-term clinical experience study rates do not

compare well with long-term clinical experience. For this reason, it is convenient to compare results from PMA preclinical studies. Such studies have comparable patient numbers, demographics, duration, statistics, and study methods.

Results from PMA preclinical studies can be found by searching a database for "Summary of Safety and Effectiveness" documents on the FDA website (http://www.fda.gov/cdrh/). A Summary of Safety and Effectiveness is prepared for each device that has received a PMA and contains a summary of the clinical study, the preclinical studies, and the manufacturing facilities. Furthermore, most valves in the United States are packaged with "Instructions for Use" that include the clinical complication and hemodynamics data. Most valve manufacturers have websites that either contain the pertinent data or provide relevant references to the literature because of the importance of this data for marketing purposes. In general, high survival, freedom from complications or low complication rates, low pressure gradients, high effective orifice areas, and low regurgitation are essential.

Much of the same type of information is also available in the open literature; however, because of differences in study design, methods, and size that occur, it is more difficult to compare results from different studies and sources. An excellent review of heart valve complication comparison problems can be found in a paper by G. Grunkmeier and Y. Wu, "Our complication rates are lower than theirs: Statistical critique of heart valve comparisons" (14). An excellent review of hemodynamics for a number of valves may also be found in a paper by Zang and Chambers (15). As with complication rates, hemodynamics comparisons based on studies in the open literature can be difficult because of differences in technique and quantity definitions.

Examples of valve comparisons with the above information will be presented in subsequent sections of this text.

VALVE DESIGN ELEMENTS

The challenge is to provide valve prostheses that can endure the aggressive biological and mechanical environment for a patients' lifetime without structural or functional degradation. A myriad of disciplines, including medicine, biology, engineering design, materials engineering, structural engineering, and fluid mechanics, are required to meet this challenge. Advances in each of these disciplines over the past 30 years have made possible significant improvements in the safety and efficacy of valve prostheses. At the current state of development, common structural features exist among the various valve designs required to meet the goals of function and durability. Conceptually, a valve prosthesis is a simple one-way valve that opens easily to allow forward flow and closes adequately to prevent regurgitant losses during the cardiac cycle. Each design has:

- An orifice body that defines the open forward flow channel.
- Leaflets or occluders that reversibly open to permit forward flow and close (or occlude) to prevent reverse flow.
- Struts or pivots that capture and guide leaflet motion during opening and closing.
- A sewing cuff or sewing ring that is used to attach the valve to the cardiac anatomy.

In addition, bioprosthetic valves may have stents, which are metallic or polymeric structures that define and support the soft tissue valve shape.

Many of the design elements of valves are radio-opaque and present a unique appearance upon examination by X-ray or cine fluoroscopy. Guides to identifying valves by radiographic appearance have recently been published by Butany et al. (16,17).

The size, shape, strength, durability, and biocompatibility of these common design elements define the performance of the valve. Each element is, in turn, defined by the materials used in its construction, hence, the importance of materials technology.

MATERIALS TECHNOLOGY

The materials available in many ways dictate design possibilities. Our repertoire of manmade materials for mechanical valves includes certain polymers, metals, and pyrolytic carbons in a class generally known as biomaterials. For heart valve applications, the materials used must be blood- and tissue-compatible. The presence of the material itself, wear particles, or the material in combination with others cannot provoke adverse biological reactions. Some of the more important adverse biological reactions include intense inflammatory response, carcinogenicity, mutagenicity, or the inception of thrombosis.

The materials must be biodurable: The mechanical integrity must resist degradation by the *in situ* physiochemical environment. Assuming that the human heart beats approximately 40 million times a year, valve prostheses must endure on the order of 600 million cycles over a 15 year period without functional degradation due to exposure to the aggressive biological reactions and repeated cyclic stresses in the cardiac environment. Early valve designs failed within a short period because of thrombosis (blood clotting), and damage due to exposure to the rigorous biological environment manifest as distortion, wear, and fatigue. Additionally, the materials must tolerate sterilization without adverse affects. A more current requirement is that a material cannot interact adversely with high strength electromagnetic radiation, as may be encountered with modern diagnostic techniques such MRI (magnetic resonance imaging).

The short list of materials that meet these requirements for heart valve structural components includes isotropic pyrolytic carbons, certain cobalt-chrome alloys, and certain titanium alloys. The polymers most often used today in the fabric sewing cuffs are polyethylene terephthalate (PET, tradename Dacron) and polytetrafluoroethylene (PTFE, tradename Dacron). Polyacetyl (polyformaldehyde, tradename Delrin) and poly dimethyl siloxane (Silicone rubber) are used as occluders in some of the older designs. Some important physical and mechanical characteristics of these materials are given in Table 3.

Table 3. Representative Material Properties

Property	Unit	PyC	CoCr	Delrin
Density	$g \cdot cm^{-3}$	1.93	8.52	1.41
Bend strength	MPa	494	690 (uts)	90
Young's modulus	GPa	29.4	226	3.1
Hardness	HV	236[a]	496	86 Shore D
Fracture toughness K_{1c}	$MN \cdot m^{3/2}$	1.68		
Elongation at failure	%	2		30
Poisson's ratio		0.28	0.3	0.35

[a]The hardness value for PyC is a hybrid definition that represents the indentation length at a 500 g load with a diamond penetrant indentor.

As the mechanical properties of biological materials are more complex than for manmade materials, they will not be addressed here, rather references are suggested in the recommended reading section.

Each material used in a prosthetic heart valve must meet a battery of tests for biocompatibility as currently defined in the FDA Guidance document (8) ISO 5840 and ISO 10993. The interested reader can find the details of these tests in the reference documents.

DESIGN CONCEPTS

Two broad categories of valve prostheses have evolved, mechanical and biological, each with advantages and disadvantages. The selection of valve type for a given recipient is made by balancing the individual's needs relative to the valve type advantages and disadvantages. Factors for consideration in valve type selection include:

- Thromboembolism/Stroke,
- Bleeding Risk,
- Reoperation,
- Survival.
- Guidelines from The American Heart Association/ American College of Cardiology (AHA/ACC) are listed below in Table 4 (18).

Table 4. AHA/ACC Guidelines[a]

AHA/ACC Indications for Valve types	
Mechanical Valve	Tissue Valve
Expected long lifespan	AVR >65 years age, No risk factors for TE
Mechanical valve in another position	MVR >70 years age
AVR $\leq$ 65 years age	Cannot take warfarin
MVR $\leq$70 years age	
Requires warfarin due to risk factors	
Atrial fibrillation	
LV Dysfunction	
Prior thromboembolism	
Hypercoagulable condition	

[a]AVR –aortic valve replacement MVR – mitral valve replacement.

In the absence of concomitant disease, aortic mechanical valves are recommended for patients less than 65 years of age and mitral mechanical valves for patients less than 70 years of age.

Biological valves offer the advantage that chronic anticoagulation therapy (warfarin) may not be required beyond an aspirin taken daily. Many patients benefit from biological valves because of an inability to tolerate or to comply with chronic anticoagulation. However, many patients with biological valves may require anticoagulation for other concomitant conditions, such as atrial fibrillation that pose a coagulation risk, and some physicians may recommend anticoagulation for mitral replacements.

The disadvantage is that biological valves have limited lifetimes for reasons inherent to host responses that degenerate the implant tissue and the relatively low fatigue resistance of the devitalized biological tissues. Expected patient survival (median) in the United States equals expected biological valve survival at 18 years for aortic valves and 69 years for mitral valves (19).

The risk of biological valve dysfunction increases inversely with age, with failure being more rapid in younger patients. Degenerative mechanisms such as calcification are accelerated in young growing patients, and an active lifestyle may more severely stress the valve. As a mitigating factor, biological valve degeneration tends to be gradual, hence readily detected so that a re-operation to replace the valve can be scheduled before the patients' health becomes compromised. Here, the tradeoff is the risk of re-operation vs. the risk of anticoagulant-related complications. Therefore, the ideal biological valve candidate patient tends to be an elderly person with a relatively sedentary lifestyle and an expected lifetime that is not likely to exceed the useful lifetime of the valve.

A mechanical valve has the advantage of unlimited durability, but has the disadvantage of requiring chronic anticoagulant therapy.

Many excellent resources are available on the Internet providing illustrations, some design, details, and performance details for valve prostheses. A general overview of valves and other devices approved for marketing in the United States can be found at the Cardiovascular and Thoracic Surgery website (http://www.ctsnet.org/sections/industryforum/products/index.cfm). Other very accessible resources include the Evansville Heart Center prosthetic heart valve gallery (http://members.evansville.net/ict/prostheticvalveimagegallery.htm) and the Cedar Sinai Medical Center (http://www.csmc.edu/pdf/Heart Valves.pdf). Heart valve manufacturers also have websites that provide information regarding valves and performance. A partial list follows:

- ATS Medical, Inc. http://www.atsmedical.com/
- Carbomedics, A Sorin Group Company, http://www.carbomedics.com/, and Sorin Biomedica, http://www.sorin-cid.com/intro_valves.htm
- Cryolife, Inc., http://www.cryolife.com/products/cryovalvenew.htm.
- Edwards Lifesciences, http://www.edwards.com/Products/HeartValves

- Medical Carbon Research Institute, LLC., http://www.onxvalves.com/
- Medical CV, http://www.vitalmed.com/products/medicalcv/omnicarbon_heart_valve.htm
- Medtronic, Inc., http://www.edwards.com/Products/HeartValves
- St. Jude Medical, Inc., http://www.sjm.com/conditions/condition.aspx?name=Heart+Valve+Disease&section=RelatedFeaturesProducts

Many other resources are available, other manufacturers, a number of professional journals and archives of news stories, and press releases available on the web. A web search for "artificial heart valves" will yield more than 300,000 sites. The sites listed above are given as a simple quick start for the reader, without implied endorsement.

As a point of interest, the valve manufacturers' websites are highly colored by competitive marketing and tend to present a rather focused view. However, they often provide copies of studies in the literature along with supplemental interpretative information. As an example, aspects of the selection of type, biological and mechanical, for a given patient are controversial. To address this controversy, the Edwards Lifesciences and St. Jude Medical sites have detailed, authoritative panel discussions of issues related to valve selection that are well worth reviewing.

BIOLOGICAL VALVES

Bioprosthetic valve replacements tend to mimic the semilunar aortic valve directly by using a natural aortic valve from a donor (allograft) or an aortic valve from another species (xenograft). Alternatively, the natural aortic valve may be imitated by design using a combination of nonvalvular natural tissues and synthetic materials (heterograft). Typically, the semilunar aortic-type valve is used in all positions for valve replacement.

The native aortic valve can be visualized as a short flexible tube within a longer tube. The outer tube, the aorta, has three symmetric scalloped bulbous protuberances (sinus of Valsalva) at its root. Within the sinus of Valsalva, there is the short inner tubular aortic valve, with a length approximately the same as the diameter. The aortic valve tube is constrained at three equidistant points along its upper circumference at the commissures. When pressurized from above, the aortic valve tube wall collapses onto itself into three symmetric crescent-shaped leaflet cusps. Along the lower circumference, the leaflets attach to the aorta wall following the scalloped curve of the aortic root. When pressurized from below, the valve opens, the cusps rise together and separate to form the short open internal tube, which is the flow region of the valve. When closed, the internal tube walls, cusps, collapse, inflate, and form a seal (coaptation) at the center of the valve (nodus Aranti) along lines 120° apart (see Fig. 1). A good dynamic, multimedia visualization of this collapsing tube concept of the aortic valve can be found at the 3F Technologies website (http://www.3ftherapeutics.com/us/products.html).

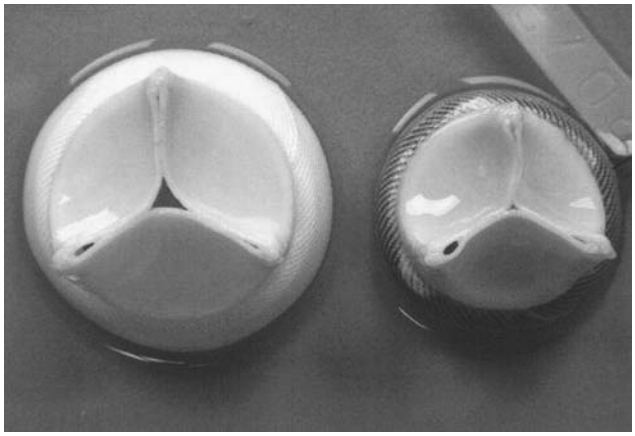


Figure 1. Mitroflow bovine pericardial prostheses heterograft shown in two sizes. This figure illustrates the collapsed tube concept of aortic valve structure.

Valves constructed from biological materials have several categories, including:

- Homograft (also Allograft), native valves taken from members of the same species,
- Xenograft (also Heterograft), valves taken from other species,
- Autograft, valves transposed from one position to another in the same patient. Valves may also be constructed from nonvalvular tissues such as pericardium, dura mater, or facia lata and the tissues may be autogenic, allogenic, or xenogenic. Valves may be stented (with added rigid support structures) or unstented.

Valves transplanted from one individual to another do not remain vital. Although the valve may be initially coated with endothelium, the valve eventually becomes acellular with only the extracellular matrix collagenous structures remaining. A biological tissue originating from another species (xenograft) must be killed and stabilized by aldehyde cross-linking or some other method. Fixation masks autoimmune rejection processes and adds stability and durability. However, the tissue again consists of an acellular, but now cross-linked, extracellular matrix collagenous structure. Biological valves lifetimes are limited due to degradation of the relatively fragile nonvital, acellular material, which results in structural and functional failure. Degradation processes are exeraberated by calcification and elevated applied stress levels in the cardiovascular environment.

Valve Homografts (Also Allograft)

Use of an orthotopic implanted homologous cardiac valve (e.g., a valve transplant from the same species) as a valve substitute was first successfully accomplished in humans by Donald Ross in 1962 (20). Harvest must occur as rapidly as possible following the donor's death, if within 48 h, the valve may still be vital. However, with human or animal origin, a risk of disease transmission exists, particularly if the valve is not sterilized. Furthermore, competition exists with the need for intact donor hearts for

heart transplantation, so the supply of homografts is limited.

During the period between 1962 and 1972, homografts were widely used because other valve substitutes were not yet satisfactory. As other valve substitutes became effective and more readily available during the 1970s, homograft use decreased. With the development of cryopreservation technology in the mid-1980s and commercialization by Cryolife, homograft use has increased again. However, unless the homograft is an autograft (from the same patient), it becomes a devitalized acellular, collagenous structure within the first year of implantation (21). Homograft lifetime in the aortic position tends to be limited to approximately 10 years due to primary tissue failure because of the inherent fragility of the devitalized, acellular homograft.

The most successful use of homografts is for a pulmonary valve replacement in the Ross procedure in which the patient's own viable pulmonary valve (autograft) is transplanted into the aortic position as a replacement for diseased aortic valve and a heterograft used to replace the patient's pulmonary valve (22). In principal, the autograft pulmonary valve in the aortic position remains vital and grows with the patient. The heterograft in the pulmonary position endures well because the pressure loads and the consequences of regurgitation are much less severe on the right side of the heart.

Xenografts

Aldehyde-fixed porcine xenograft valves were initially developed in 1969 by Carpentier et al. (1), which consisted of porcine aortic valves mounted on a stent frame to maintain shape and then fixed with dilute gluteraldehyde (see Fig. 2). Aldehyde fixation masks autogenic host reactions, cross-links the extracellular matrix collagen, and kills the native valve cells. Infectous cells and viri are also cross-linked and killed, so fixation provides a means of disinfection. During the same time period, Hancock et al. independently developed another porcine prosthesis (23). Yet another design, constructed from bovine pericardium, was developed by Ionescu in 1971 (24). These valves provided readily available alternatives to homografts and were widely used during the early 1970s. Although these valves offered freedom from chronic anticoagulation, lifetimes were limited to around 10 years. The valves failed because of calcification, leaflet tears, and perforations. Freedom from reoperation 15 years post implant was about 50% for the porcine valves and slightly worse for the bovine pericardial valves. Other important lessons were learned the Hancock valve polymeric Delrin stents would creep, leading to valve dysfunction. With the Ionescu valve, a suture located at the valve commissures initiated tears in the leaflets. The attachment of leaflets at the commisures tends to be the most highly stressed point, because of pressure and reversed bending stresses, which leads to commissural tearing.

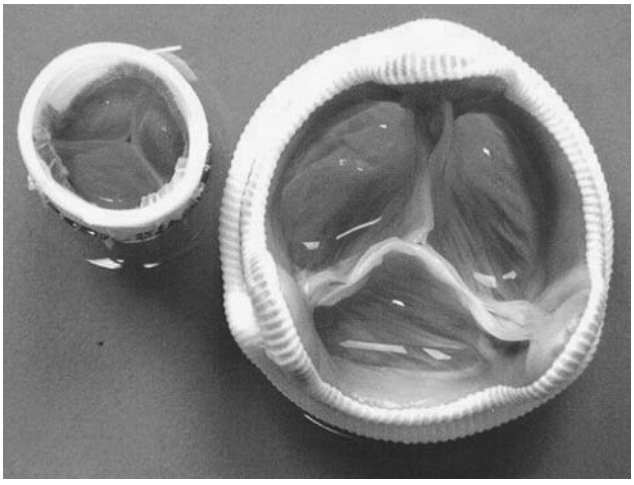


Figure 2. A porcine valve xenograft design (Xenomedica) shown in two sizes. Here, a porcine valve is mounted in a stented structure.

Xenograft biological valve replacements have gone through three generations of evolution. The first generation and some of its problems are described above by Hancock and Ionescu. Tissues used in these valves were fixed with aldehydes at high pressure in the closed position. As the collagen fibers within the porcine leaflets tend to have a wavy or crimped structure, the high pressure fixation tended to straighten the fibers out, removing the crimp, which was found to have an adverse effect on fatigue endurance.

For the second generation valves, low and zero pressure aldehyde fixation were used, flexible stent materials were employed, and assembly strategies improved. These improvements retained the natural collagen crimp and lowered stresses at the commissural attachment. Also, additional fixation treatments and chemical posttreatments were employed in hopes of reducing calcification.

Contemporary third-generation-valves tend to use fixation at physiological pressures, flexible stents materials, and include stentless designs. Some of these newer aortic valve designs are challenging the age limits given above for valve-type selection (25). A number of stentless designs are available that typically consist of a xenograft or homograft valve mounted in fabric or tissue to enable sewing (see Fig. 3). Some designs include the portions of the aortic arch along with the valve. Stenting and annular support structures tend to decrease the annular flow area. Removal of the stents should enhance hemodynamic performance.

A unique design using equine pericardium and an external expandable stent for use with a transcatheter delivery system is under development by 3F Theraputics. Furthermore, tissue engineering concepts are being explored as a means of developing valve scaffolding structures in expectation that the host cells will infiltrate, populate, and ultimately render the scaffolding into an autologous tissue structure. One such example is a chemically decellularized valve in development by Cryogenics. Carbomedics has also developed a method for nonaldehyde cross-linking and decellularization of collageneous cardiac tissues (26).

Hemodynamics for some of the contemporary biological valves cited from FDA PMA Summaries of Safety and

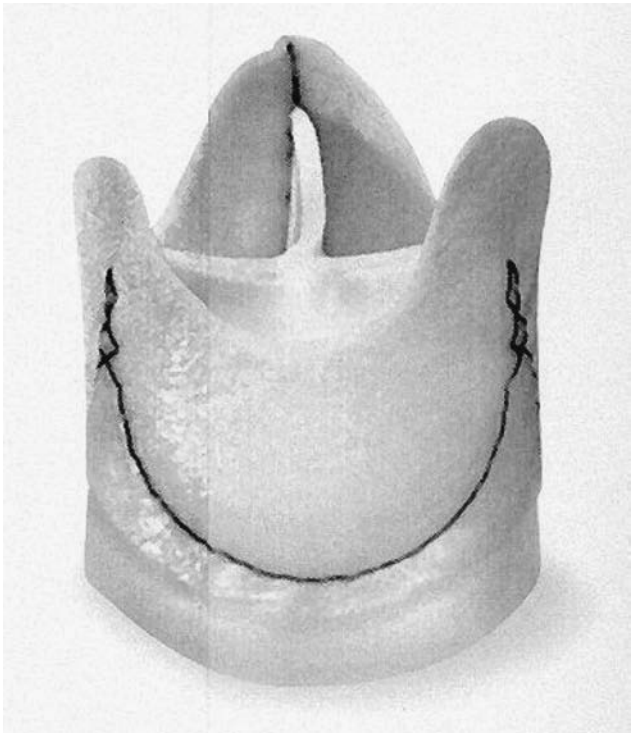


Figure 3. A stentless pericardial aortic valve. Sorin Biomedica Pericarbon.

Effectiveness and the literature for uniform echocardiographic protocols and comparable patient experience are listed in Tables 5 and 6. All of the contemporary biological valves exhibit satisfactory hemodynamics and acceptably low risk of valve-related complications.

MECHANICAL VALVES

Contemporary mechanical valves are primarily bileaflet valve designs constructed from pyrolytic carbon (PyC). Some monoleaflet valves and ball in cage valves, such as the Medtronic Hall valve and the Starr-Edwards valve, are still used. Some examples of bileaflet and monoleaflet designs are shown in Fig. 4. The bileaflet designs consist of two flat half circular leaflets inserted into a cylindrical annulus. The orifice may be stiffened by external metallic rings and is encircled by a fabric sewing cuff. Small depressions, or extrusions, within the orifice accept mating features on the leaflet and provide a pivot mechanism to capture and guide the leaflets during opening and closing excusions. Figure 5 provides a cut-away illustration of general bileaflet valve features. Figure 6 provides a high detail view of a hinge pivot.

PyC is used almost exclusively in bileaflet design structural components because of excellent blood compatibility, biostability, strength, and fatigue endurance. This biomaterial is an isotropic, fine-grained turbostratic carbon that is typically prepared as a coating on an inner graphite substrate. Most PyC valve components are on the order of slightly less than 1 mm in thickness and have a low enough density so as to be easily moved by cardiovascular flow and pressure. Two general types are available, silicon-alloyed and pure carbon; both have roughly equivalent and adequate properties for heart valve application. The history of PyC applications in heart valves and its material science is another story in itself and will not be addressed here. Rather, sources about PyC can be found in the recommended reference list.

Supraannular placement has been the most prominent design evolution in contemporary mechanical valves. Originally, the sewing cuff girdled the mechanical valve orifice at approximately mid-height and was intended to fit entirely within the native valve annulus. However, inserting the bulk of the sewing cuff and valve into the native annulus reduces the available flow area. Supraannular designs are modifications to the sewing cuff and valve orifice exterior wall that allow the sewing cuff to be positioned above the annulus, which removes the sewing cuff bulk from within the native annulus and, thus, increases the available flow area.

Table 5. Bioprosthetic Aortic Valve Pressure Gradients (mmHg)

Valve/Size		19	21	23	25	27	29	PMA
Toronto	Stentless	577	10 ± 9.0	7.3 ± 4.8	6.4 ± 5.1	5.1 ± 3.1	3.8 ± 2.3	P970030
Freestyle	Stentless	631	11.7 ± 4.7	9.8 ± 7.4	8.8 ± 6.8	5.1 ± 3.3	4.4 ± 2.9	P970031
Prima Plus	Stentless	366	15.9 ± 7.0	9.9 ± 5.7	8.8 ± 4.9	6.5 ± 3.9		P000007
SAV	Stented	337	12.7 ± 4.2	10.5 ± 4.3	11.3 ± 5.5	8.3 ± 3.3		P010041
Mosaic	Stented	1252	14.5 ± 5.3	12.8 ± 5.0	11.8 ± 5.2	10.0 ± 4.0	10.3 ± 2.6	P990064
Hancock II	Stented	205	12.9 ± 4.2	13.2 ± 4.6	11.3 ± 4.4	11.7 ± 4.8	10.5 ± 3.6	P980043

Table 6. Bioprosthetic Aortic Valve Effective Orifice Areas (cm^2)

Valve/Size		19	21	23	25	27	29	PMA
Toronto	Stentless	577	1.3 ± 0.7	1.5 ± 0.6	1.7 ± 05	2.0 ± 0.6	2.5 ± 0.8	P970030
Freestyle	Stentless	631	1.1 ± 0.3	1.4 ± 0.4	1.7 ± 0.5	2.0 ± 0.5	2.5 ± 0.7	P970031
Prima Plus	Stentless	366	1.1 ± 0.4	1.6 ± 0.5	1.9 ± 0.4	2.1 ± 0.6		P000007
SAV	Stented	337	1.3 ± 0.4	1.5 ± 0.4	1.7 ± 0.5	1.8 ± 0.6		P010041
Mosaic	Stented	1252	1.3 ± 0.4	1.5 ± 0.4	1.8 ± 0.5	1.9 ± 0.6	2.2 ± 0.7	P990064
Hancock II	Stented	205	1.4 ± 0.5	1.3 ± 0.2	1.4 ± 0.4	1.6 ± 0.4	1.4 ± 0.3	P980043

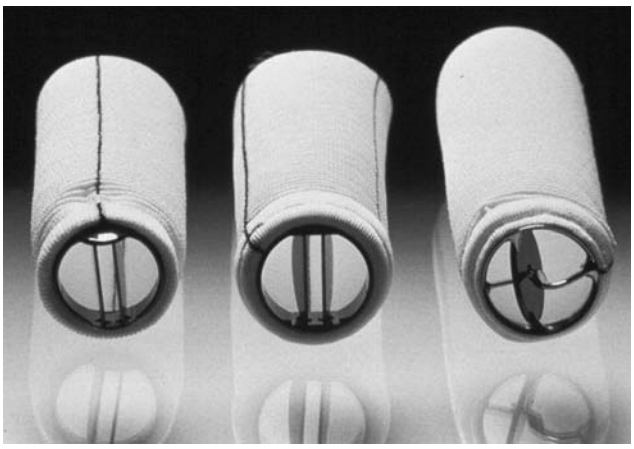


Figure 4. Three mechanical valve designs shown attached to aortic grafts. Left to right: St. Jude Medical bileaflet, Carbomedics bileaflet, and Medtronic-Hall monodisk.

Effectively, a supraannular sewing cuff allows the placement of an upsized valve relative to the intraannular cuff. Most often, the valve orifice and leaflet components are unchanged, only the sewing cuff is modified. Thus, supraannular cuffs are a subtle modification in that the renamed supraannular design is not really a new design, rather it is just an upsized version of the original valve. Virtually every bileaflet valve manufacturer now has supraannular cuffs. Renaming the supraannular versions has caused a bit of confusion because it is not always obvious that improved hemodynamic performance is due solely to the placement of an upsized valve of the same design. However, two recent designs, the Medical Carbon Research Institute On-X valve and the SJM Regent, actually incorporate significant design features as supraannular valves beyond a modified sewing cuff.

Hemodynamics for some contempory mechanical valves are given in Tables 7 and 8 (27–38). All of the contemporary mechanical valves exhibit satisfactory hemodynamics and acceptably low risk of valve-related complications. As a

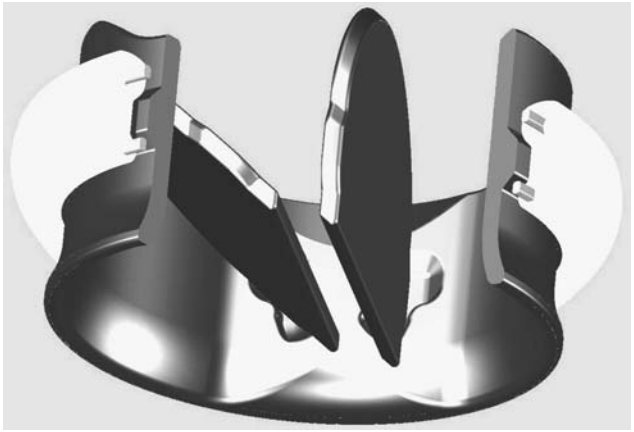


Figure 5. A cut-away view of the On-X valve showing leaflets in the closed (left) and open (right) positions. The valve orifice or annulus ring has pivot depressions in the inside diameter wall and is encased in a sewing cuff. The orifice wall has a bulge and two external metal wires for stiffening.

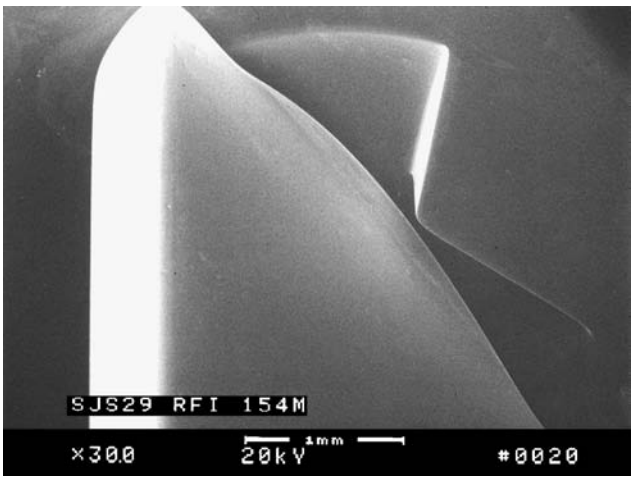


Figure 6. Detail of the St. Jude Medical bileaflet valve pivot mechanism. A tab on the leaflet inserts into a butterfly shaped depression in the valve annulus ring. The pivot depression retains and guides the leaflet motion during opening and closing excusions.

point of interest, only aortic hemodynamics are presented for both the biological and mechanical valve designs, because clinical data for aortic valves demonstrates a strong dependence on valve size. Effective orifice areas increase and pressure gradients decrease with increasing valve annulus size. Clinical hemodynamic data for mitral valves typically does not depend as strongly on annulus size (15). Mitral sizes range from 25 to 33 mm annulus diameters. In many designs, the larger annulus diameter sizes, above 27–29 mm, are attained by increasing the sewing cuff bulk. However, the difference between effective orifice area and pressure gradients with size is minimal. For mechanical mitral valves, mean pressure gradients are on the order of 4 mmHg and effective orifice areas of 2.5–3 cm^2.

COMPARISON OF BIOLOGICAL AND MECHANICAL

Detailed comparisons of biological and mechanical valve experience in large multicenter randomized studies have been recently published (39–41). Freedom from complications were found to be roughly comparable. Rates for thromboembolism and stroke have been found to be the same, and strongly related to nonvalve factors. Bleeding is more common for mechanical valves, but depends on the anticoagulant control and the use of aspirin. Freedom from reoperation is significantly better for mechanical valves. At 15 years, the freedom from reoperation for mechanical valve patients exceeded 90%, whereas while for biological valves, the percent free from reoperation was 67% for the aortic position and 50% for the mitral position. Survival at 15 years was slightly better with aortic mechanical valves, mortality of 66 versus 79% for biological valves. No difference exizted in survival for mitral valves. Overall, the survival of biological and mechanical valve patients was similar over extended periods, with the tradeoff being an increased risk of hemorrhage with mechanical valves vs. and increased risk of reoperation for biological valves. This risk of reoperation for biological valves increases in time. Thus, valve type selection is a balance

Table 7. Mechanical Aortic Valve Pressure Gradients (mmHg)

Valve/Size		19	21	23	25	27	29	PMA
On-X	SA	8.9 ± 3.0	7.7 ± 2.9	6.7 ± 3.1	4.3 ± 2.4	5.6 ± 0.3	5.6 ± 0.3	P000037
SJM Regent	SA	9.7 ± 5.3	7.6 ± 5.2	6.3 ± 3.7	5.8 ± 3.4	4.0 ± 2.6		[a]
SJM HP	SA	13.6 ± 5	12.6 ± 6.5	13 ± 6				[a]
SJM		17 ± 7	15.1 ± 3.2	16 ± 6	13 ± 6	11.5 ± 5	7 ± 1	[a]
ATS		20.2 ± 2.8	18.0 ± 1.6	13.1 ± 0.8	11.1 ± 0.8	8.0 ± 0.8	7.8 ± 1.1	P990046
CMI		21.7 ± 9.1	16.2 ± 7.9	9.9 ± 4.2	10.5 ± 2.8	7.2 ± 3.9	5.1 ± 2.8	P00060
Sorin Bicarbon		14.1 ± 2.9	10.1 ± 3.3	7.7 ± 3.3	5.6 ± 1.6			[b]
Omnicarbon				19 ± 8	16 ± 8	12 ± 4		P8300039

[a]See Refs. 27–36.
[b]See Refs. 37,38.

Table 8. Mechanical Aortic Valve Effective Orifice Areas (cm^2)

Valve/Size		19	21	23	25	27	29	PMA
On-X	SA	1.5 ± 0.3	1.9 ± 0.5	2.4 ± 0.7	2.7 ± 0.7	2.9 ± 0.7	2.9 ± 0.7	P000037
SJM Regent	SA	1.6 ± 0.4	2.0 ± 0.7	2.2 ± 0.9	2.5 ± 0.9	3.6 ± 1.3		[a]
SJM HP	SA	1.25 ± 0.2	1.3 ± 0.3	1.8 ± 0.4				[a]
SJM		0.99 ± 0.2	1.3 ± 0.2	1.3 ± 0.3	1.8 ± 0.4	2.4 ± 0.6	2.7 ± 0.3	[a]
ATS		1.2 ± 0.3	1.5 ± 0.1	1.7 ± 0.1	2.1 ± 0.1	2.5 ± 0.2	3.1 ± 0.4	P990046
CMI		0.9 ± 0.3	1.3 ± 0.4	1.4 ± 0.4	1.5 ± 0.3	2.2 ± 0.7	3.2 ± 1.5	P00060
Sorin Bicarbon		0.8 ± 0.2	1.1 ± 0.2	1.6 ± 0.2	2.4 ± 0.3			[b]
Omnicarbon				1.8 ± 0.9	1.9 ± 0.8	2.5 ± 1.4		P8300039

[a]See Refs. 27–36.
[b]See Refs. 37,38.

between the patient ability to tolerate chronic anticoagulation therapy and the risk of re-operation. To reiterate the AHA/ACC criteria, in the absence of concomitant disease, aortic mechanical valves are recommended for patients less than 65 years of age and mitral mechanical valves for patients less than 70 years of age.

IMPROVEMENTS IN ANTICOAGULATION AND DURABILITY

As mentioned earlier, although contemporary PyC mechanical valves have virtually unlimited durability and extremely low incidences of structural failure, the biocompatibility is not perfect. Hence, because of imperfect biocompatibility, chronic anticoagulation therapy is required, which is the disadvantage of mechanical valves. Anticoagulation therapy carries with it the risk of hemorrhage, thromboembolism, and thrombosis.

Valve thrombogenicity can be thought of in terms of an extrapolation of Virchow's triad from veins to valves. Here, a predisposition to thrombogenicity is attributed to three factors, (*1*) the blood compatibility of the artificial valve material, (*2*) the hemodynamics of blood flow through the valve, and (*3*) patient specific hemostasis. In effect, resistance to valve thrombogenicity occurs through a balance of all three factors. Conversely, valve thrombosis could be provoked by a poor material, poor hemodynamics, or a patient-specific hypercoagulable condition.

Improvements are being made in the need for and control of anticoagulant therapy for mechanical valve patients. Anticoagulant-related complications have been shown to be reduced by the use of INR (international normalized ratio) self-monitor units (42). The success of INR self-monitoring leads to the possibility of reduced anticoagulant levels, which affects the patient-specific hypercoagulable condition of Virchow's triad. Valve manufacturers have also made improvements in material quality and hemodynamics to the extent that aspirin-only and reduced warfarin level studies are planned and in place for certain low risk aortic mechanical valve patients (43).

On the other hand, concurrent improvements in biological valve design and tissue treatments can lead to extended durability, which reduces the risk of re-operation. However, comparison of Tables 5 and 8 for effective orifice areas shows that biological valves tend to be more stenotic than mechanical valves. In the interest of avoiding patient-prosthesis mismatch, smaller patients tend to be better candidates for biological valves. Valve selection should ultimately be a matter of patient and surgeon preference.

FUTURE DIRECTIONS

Advances in tissue engineering probably hold the brightest promise, because a viable, durable, autologous valve is the best possible valve replacement. However, until realized, improvements in materials technology, anticoagulant therapy, and valve hemodynamics can all be expected to improve the outcome of valve replacements. Developments in techniques for valve annuloplasty and valve repair (44) have been highly successful in allowing the surgical restoration of the native valve, which often eliminates the need for a valve replacement. In fact, mitral valve implant rates have decreased with the popularity of valve repair. Strides are also being made in minimally invasive techniques, robotic surgery, and transcatheter surgery so as to minimize the risk and trauma of valve replacement surgery.

CONCLUSION

Valve replacements are a commonplace occurrence worldwide, and many advances in design and patient management have been made over the past 30 years. Current replacements are certainly safe and effective. Surgical outcomes and patient quality of life can only be expected to improve.

BIBLIOGRAPHY

Cited References

1. Carpentier A, et al. Biological factors affecting long-term results of valvular heterografts. J Thorac Cardiovas Surg 1969;58:467–482.
2. LaGrange LD, Gott VL, Bokros JC, Ramos MD. In: Johnson Hegyeli R, editor. Compatibility of Carbon and Blood, Chapter 5. Artificial Heart Program Conference, Washington, DC: National Heart Institute Artificial Heart Program; June 9–13,1969; 47–58.
3. Roe B. Chapter 13. Extinct. In: Bodnar E, Frater R, editors. Cardiac Valve Prostheses, Replacement Cardiac Valves, New York: Pergamon Press; 1969; 307–332.
4. Dewall RA, Qasim N, Carr L. Evolution of mechanical heart valves. Ann Thorac Surg 2000;69:1612–1621.
5. Brewer L, Prosthetic Heart Valves, Springfield (IL): C.C. Thomas; 1969.
6. Replacement Heart Valve Guidance—Draft Document, October 14, 1994. Available at http://www.fda.gov/cdrh/ode/3751.html.
7. ISO Standard 5840, Cardiovascular Implants—Cardiac Valve.
8. Summary of Safety and Effectiveness Data, On-X Prosthetic Heart Valve Model ONXA, P000037, May 30, 2001.
9. Summary of Safety and Effectiveness Data, On-X Prosthetic Heart Valve Model ONXM and ONXMC, P000037/S1, Marxh 6, 2002.
10. Hammermeister K, Sethi GK, Henderson WG, Grover FL, Oprian C, Rahimtoola SH. Outcomes 15 years after valve replacement with a mechanical versus a bioprosthetic valve: Final report of the Veterans Affairs randomized trial. J Am Coll Cardiol 2000 Oct;36(4):1152–1158.
11. Walther T, Lehmann S, Falk V, Kohler J, Doll N, Bucerius J , Gummert J, Mohr FW. Experimental evaluation and early clinical results of a new low-profile bileaflet aortic valve. Artif Organs. 2002 May;26(5):416–419.
12. Pibarot P, et al. Impact of prosthesis-patient mismatch on hemodynamics and asymptomatic status, morbidity after aortic valve replacement with a bioprosthetic heart valve. J Heart Valve Dis 1998;7:211–218.
13. Walther T, et al. Left ventricular remodeling after stentless aortic valve replacement. In: Huysmans H, David T, Westaby S, editors. Stentless Bioprosthesis. Oxford (UK): Isis Medical Medica Ltd; 1999; p 161–165.
14. Grunkemeir GL, Wu Y. Our complication rates are lower than theirs: Statistical critique of heart valve comparisons. J Thorac Cardiovas Surg 2003;125(2):290–300.
15. Wang Z, Grainger N, Chambers J. Doppler echocardiography in normally functioning replacement heart valves: A literature review. J Heart Valve Dis 1995;4:591–614.
16. Butany J, Fayet C, Ahluwalia MS, Blit P, Ahn C, Munroe C, Israel N, Cusimano RJ, Leask RL. Biological replacement heart valves. Identification and evaluation. Cardiovasc Pathol 2003 May–Jun; 12(3):119–39.
17. Butany J, Ahluwalia MS, Munroe C, Fayet C, Ahn C, Blit P, Kepron C, Cusimano RJ, Leask RL. Mechanical heart valve prostheses: Identification and evaluation. Cardiovasc Pathol 2003 Jan–Feb;12(1):1–22.
18. Bonow R, et al. ACC/AHA Guidelines for the Management of Patients With Valvular Heart Disease. Executive Summary. A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee on Management of Patients With Valvular Heart Disease). J Heart Valve Dis 1998 Nov;7(6):672–707.
19. Khan S. Long-term outcomes with mechanical and tissue valves. J Heart Valve Dis 2002 Jan; 11 (Suppl 1): S8–S14.
20. Ross DN. Homograft replacement of the aortic valve. Lancet 1962 2:447.
21. Koolbergen DR, Hazenkamp MG, de Heer E, Bruggemans EF, Huysmans HA, Dion RA, Bruijn JA. The pathology of fresh and cryopreserved homograft heart valves: An analysis of forty explanted homograft valves. J Thoracic Cardiovas Surg; 124;4:689–697.
22. Ross DN. Replacement of the aortic and mitral valves with a pulmonary autograft. Lancet 1967;2:956–958.
23. Reis RL, Hancock WD, Yarbrough JW, Glancy DL, Morrow AG. The flexible stent: A new concept in the fabrication of tissue heart valve prostheses. J Thoracic Cardiovas Surg 1971;62:683–689.
24. Masters RG, Walley VM, Pipe AL, Keon WJ. Long-term experience with the Ionescu-Shiley pericardial valve. Ann Thorac Surg 1995;60(Suppl):S288–S291.
25. Glower DD, et al. Determinants of 15-year outcome with 1,119 standard Carpentier-Edwards porcine valves. Ann Thorac Surg 1998;66(Suppl):S44–S48.
26. Moore M. PhotoFix: Unraveling the mystery. J Long Term Eff Med Implants 2001;11(3–4):185–197.
27. Bach DS, Goldbach M, Sakwa MP, Petracek M, Errett L, Mohr F. Hemodynamic and early performance of the St. Jude Medical regent aortic valve prosthesis. J Heart Valve Dis 2001;10:436–442.
28. Lercher AJ, Mehlhorn U, Muller-Riemschneider, Rainer de Vivie E. In vivo evaluation of the SJM regent valve at one-year follow up after aortic valve replacement. The Society For Heart Valve Disease FIRST BIENNIAL MEETING Queen Elizabeth II Conference Centre, London June 15–18; 2001; Abstract 101.
29. Prasad SU, Prendergast B, Codispoti M, Mankad PS. Evaluation of a new generation mechanical prosthesis: Preliminary results of the St. Jude regent aortic prosthesis. The Society For Heart Valve Disease FIRST BIENNIAL MEETING Queen Elizabeth II Conference Centre, London June 15–18;2001; Abstract 104.
30. Chafizadeh E, Zoghbi W. Doppler echocardiographic assessment of the St. Jude Medical prosthetic valve in the aortic position using the continuity equation. Circulation 1991; 83:213–223.
31. Flameng W, Vandeplas A, Narine K, Daenen W, Herijgers P, Herregods M. Postoperative hemodynamics of two bileaflet valves in the aortic position. J Heart Valve Dis 1997;6:269–273.
32. Kadir I, Izzat M, Birdi I, Wilde P, Reeves B, Bryan A, Angelini G. Hemodynamics of St. Jude Medical prostheses in the small aortic root: In vivo studies using dobutamine Doppler echocardiography. J Heart Valve Dis 1997;6:123–129.
33. Zingg U, Aeschbacher B, Seiler C, Althaus U, Carrel T. Early experience with the new masters series of St Jude Medical heart valve: In vivo hemodynamics and clinical results in patients with narrowed aortic annulus. J Heart Valve Dis 1997(6):535–541.
34. De Paulis R, Sommariva L, De Matteis G, Polisca P, Tomai F, Bassano C, Penta de Peppo A, Chiariello L. Hemodynamic performance of small diameter carbomedics and St. Jude valves. J Heart Valve Dis 1996;(5: SIII):S339–S343.

35. Carrel T, Zingg U, Jenni R, Aeschbacher B, Turina M. Early in vivo experience with the hemodynamic plus St. Jude Medical valve in patients with narrowed aortic annulus. Ann Thorac Surg 1996;61:1418–1422.
36. Vitale N, et al. Clinical evaluation of St. Jude Medical hemodynamic plus versus standard aortic valve prostheses: The Italian multicenter, prospective, randomized study. J Thorac Cardiovas Surg 2001;122(4):691–698.
37. Flameng W, et al. Postoperative hemodynamics of two bileaflet heart valves in the aortic position. J Heart Valve Dis 1997; 6:269–273.
38. Kadir I, Wan IY, Walsh C, Dip-Rad, Wilde P, Byran AJ, Angelini GD. Hemodynamic performance of the 21-mm sorin bicarbon mechanical aortic prosthesis using dobutamine Doppler echocardiography. Ann Thorac Surg 2001; 72:49–53.
39. Hammermeister K, Sethi GK, Henderson WG, Grover FL, Oprian C, Rahimtoola SH. Outcomes 15 years after valve replacement with a mechanical versus a bioprosthetic valve: Final report of the Veterans Affairs randomized trial. J Am Coll Cardiol 2000 Oct; 36(4):1152–1158.
40. Khan S, et al. Twenty-year comparison of tissue and mechanical valve replacement. J Throrac Cardiovas Surg 2001; 122:257–269.
41. Bloomfield P. Choice of prosthetic heart valves: 20-year results of the Edinburgh Heart Valve Trial. J Am Coll Cardiol 2004 Aug 4;44(3):667.
42. Koertke H, Minami K, Boethig D, Breymann T, Seifert D, Wagner O, Atmacha N, Krian A, Ennker J, Taborski U, Klovekorn WP, Moosdorf R, Saggau W, Koerfer R. INR self-management permits lower anticoagulation levels after mechanical heart valve replacement. Circulation 2003; 108(Suppl 1):I175–178
43. On-X Aspirin-Only Study with Selected Isolated Aortic Valve Replacements. Available at http://www.onxvalves.com/Med_Aspirin_Study.asp.
44. Cohn L, Soltesz E. The evolution of mitral valve surgery: 1902-2002. Am Heart Hosp J 2003 Winter; 1(1):40–46.

Reading List

Standards

FDA Document *Replacement Heart Valve Guidance-Draft Document, October* 14, 1994.
ISO 5840, *Cardiovascular Implants—Cardiac valve*.
EN prEN12006 *Cardiovascular Implants—Cardiac valve*.

Pathology

Schoen FJ. Cardiovascular Pathology: Pathology of Heart Valve Substitution With Mechanical and Tissue Prostheses. New York: Churchill Livingston; 2001.

Biological Tissue Properties

Sacks MS, Schoen FJ. Mechanical damage to collagen independent of calcification limits bioprosthetic heart valve durability. Biomed Mater Res 2002;62:359–371.
Billiar KL, Sacks MS. Biaxial mechanical properties of the fresh and glutaraldehyde treated porcine aortic valve: Part I - Experimental results. J Biomechan Eng 2000;122:23–30.
Wells SM, Sacks MS. Effects of fixation pressure on the biaxial mechanical behavior of porcine bioprosthetic heart valves with long-term cyclic loading. Biomaterials 2002;23(11): 2389–2399.
Sacks MS. The biomechanical effects of fatigue on the porcine bioprosthetic heart valve. Long-term Effects Med Implants 2001;11(3&4):231–247.

Pyrolytic Carbon

More R, Bokros J. Carbon Biomaterials, Encyclopedia of Medical Devices and Instrumentation EMD 023. Heart Valve Prostheses In Vitro Flow Dynamics, Encyclopedia of Medical Devices and Instrumentation.

HEART VIBRATION. See PHONOCARDIOGRAPHY.

HEART, ARTIFICIAL

CONRAD M. ZAPANTA
Penn State College of Medicine
Hershey, Pennsylvania

INTRODUCTION

Artificial hearts are broadly defined as devices that either supplement or replace the native (natural) heart. These devices can be classified into two groups: ventricular assist devices and total artificial hearts. This article will define the clinical need, review the native heart anatomy and function, describe design considerations for ventricular assist devices and total artificial hearts, review selected designs, and recommend areas for future development.

CLINICAL NEED

Cardiovascular disease accounted for 38% of all deaths (almost 1.4 million people) in the United States in 2002 (1). Coronary heart disease (53%) represented the majority of these deaths, followed by stroke (18%), and congestive heart failure (6%). Almost 4.9 million Americans suffer from congestive heart failure, with $\sim$550,000 new cases diagnosed each year. Over 80% of men and 70% of women with congestive heart failure under the age of 65 will die within 8 years. In people diagnosed with congestive heart failure, sudden cardiac death occurs at six to nine times the rate of the general population.

One treatment for congestive heart failure is heart transplantation. It is estimated that 40,000 Americans could benefit from a heart transplant each year (1,2). However, only $\sim$2100 donor hearts were available each year from 1999 to 2004. The number of donor hearts dropped during this period, from a high of 2316 in 1999 to 1939 in 2004. Over 3300 patients were on the waiting list for a donor heart at any time during this period, with $>$65% of these patients on the waiting list for $>$1 year. From 1998 to 2004, $\sim$630 patients died each year waiting for transplant.

These numbers clearly demonstrate the clinical need for ventricular assist devices and total artificial hearts that support the patient until transplant (bridge to transplant) or permanently assist or replace the natural heart (destination therapy).

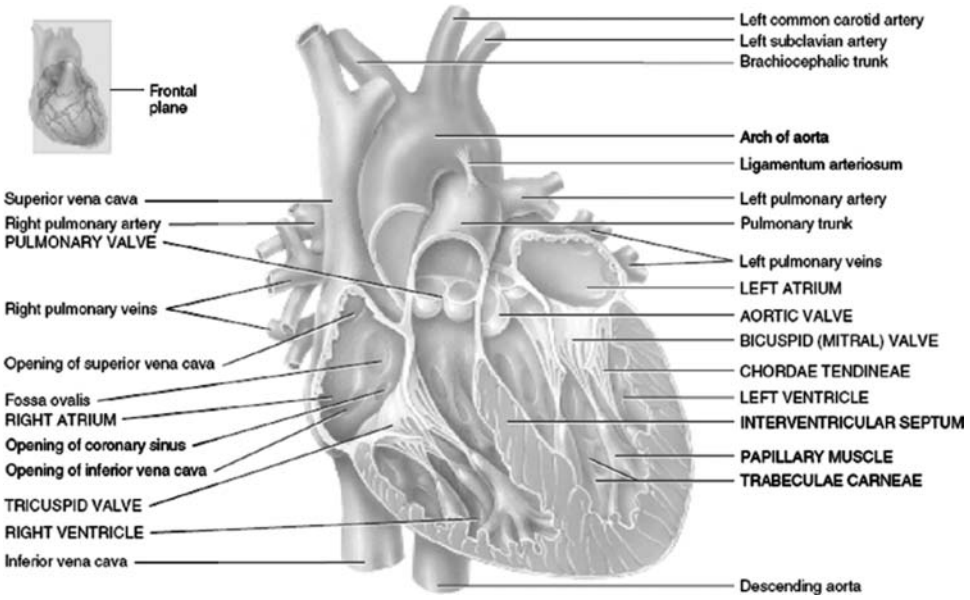


Figure 1. Anatomy of the Native Heart. The heart is composed of two pumps (right and left side) that work simultaneously. The right pump delivers blood to the pulmonary circulation (the lungs), while the left pump delivers blood to the systemic circulation (the body). Each pump consists of an atrium and ventricle. [Reprinted with permission from Gerard J. Tortora and Bryan H. Derrickson, *Principles of Anatomy and Physiology, 11th ed.*, Hoboken (NJ): John Wiley & Sons; 2006.]

NATIVE HEART ANATOMY AND FUNCTION

The anatomy of the native (or natural) heart is shown in Fig. 1. The heart is composed of two pumps (right and left side) that work simultaneously. The right pump delivers blood to the pulmonary circulation (the lungs) while the left pump delivers blood to the systemic circulation (the body). Each pump consists of an atrium and ventricle that make up the heart's four distinct chambers: right atrium, right ventricle, left atrium, and left ventricle. The atria act as priming chambers for the ventricles. The ventricles pump blood out of the heart to either the pulmonary or systemic circulation. Heart valves located between each atrium and ventricle and at the outlet of each ventricle maintain flow direction during pulsatile flow.

Blood from the systemic circulation enters the right atrium through the superior vena cava (from the head and upper extremities) and inferior vena cava (from the trunk and lower extremities). The blood is then pumped to the right ventricle, which pumps blood to the pulmonary circulation via the pulmonary arteries. Oxygenated blood returns to the left atrium heart from the lungs via the pulmonary vein and is then pumped to the left ventricle. The left ventricle pumps blood to the systemic circulation via the aorta.

Table 1 lists the nominal pressures and flows in the native heart (3). A ventricular assist device or total artificial heart must be able to generate these pressures and flows in order to meet the needs of the recipient.

DESIGN CONSDIRATIONS FOR VENTRICULAR ASSIST DEVICES AND TOTAL ARTIFICIAL HEARTS

Several design considerations must be taken into account when developing a ventricular assist device or total artificial heart. These considerations are detailed below:

1. *Size of the Intended Patient*: The size of the patient will determine the amount of blood flow required to adequately support the patient. This then determines the size of the ventricular assist device or total artificial heart. For example, a total artificial heart designed for adults would most likely be too large to be implanted within small children. A larger ventricular assist device may be placed externally, while a smaller ventricular assist device could be placed within the native heart. In addition, the size of the patient may dictate the location of some of the components. For example, the power sources may be located either internally (in the abdominal cavity) or externally depending on the size and type of the power source.
2. *Pump Performance:* A ventricular assist device or total artificial heart can be used to support or replace the native heart. Each of these support modes requires a different cardiac output. For example, a ventricular assist device can provide

Table 1. Nominal Pressures and Flows in the Native (Natural) Heart

Pressures	
Left ventricle	120 mmHg (16.0 kPa) peak systolic normal (into aorta)
	10 mmHg (1.33 kPa) mean diastolic (from left atrium)
Right ventricle	25 mmHg (3.33 kPa) peak systolic (into pulmonary artery)
	5 mmHg (0.667 kPa) mean diastolic (from right atrium)
Flows	
Normal healthy adult at rest: 5 $L \cdot min^{-1}$	
Maximum flow: 25 $L \cdot min^{-1}$	

either a portion of the blood flow required by the patient (partial support) or the entire blood flow (total support). In addition, the decision must be made whether to include a controller that will either passively or actively vary the cardiac output of the ventricular assist device or total artificial heart based on the patient demand.

3. *Reliability:* The National Institutes of Health (NIH) proposed a reliability goal for ventricular assist devices and total artificial hearts of 80% for a 2 year operation with an 80% confidence level before an artificial heart can begin clinical trials. However, the desired reliability may need to be more demanding for long-term clinical use, such as 95% reliability with 95% confidence for a 5 year operation. The design and components of ventricular assist devices and total artificial hearts must be carefully selected to achieve this reliability goal.
4. *Quality of Life:* The patient's quality of life can have a significant impact on the design of a ventricular assist device or total artificial heart. It is important to clearly define what constitutes an acceptable quality of life. For example, if a patient desires to be ambulatory following the implantation of a ventricular assist device or total artificial heart, the power supply must be easily transportable. The environment of the patient (home, work, car, etc.) should also be considered to insure proper function in these different environments. The patient should be able to monitor basic pump operation without the need for a physician or other trained medical personnel. The ventricular assist device or total artificial heart should be designed to clearly provide information through displays and provide alarms to warn the patient of potentially dangerous situations, such as a battery that is running low on power.

VENTRICULAR ASSIST DEVICES

A ventricular assist device (VAD) is designed to assist or replace the function of either the left or right ventricle. These devices are intended to provide either temporary support until a donor heart has been located or the native heart has recovered function, or as a permanent device.

As shown in Table 1, the left ventricle pumps against a higher pressure system than the right ventricle. Therefore, the left ventricle is typically more in need of assistance. Consequently, left ventricular assist devices (LVADs) are more prevalent than right ventricular assist devices (RVADs).

Ventricular assist devices can generate either pulsatile or continuous (nonpulsatile) flow.

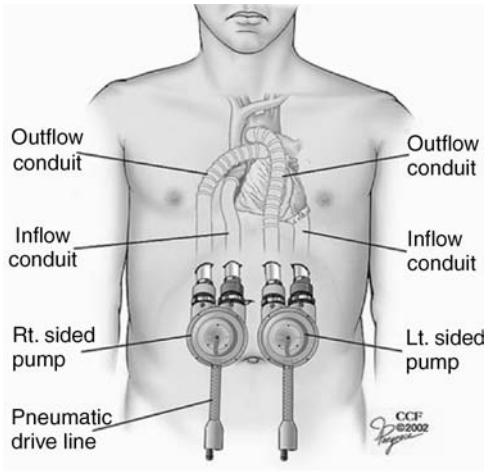
Pulsatile

Pulsatile flow ventricular assist devices are composed of a single ventricle that mimics the native ventricle. The ventricle is placed either outside the patient's body or within the abdominal cavity. There are two types of pulsatile flow ventricular assist devices: pneumatic and electric.

Figure 2. Penn State/Thoratec Pneumatic Ventricular Assist Device. The ventricle contains a flexible blood sac made of segmented polyurethane that is housed within a rigid polysulfone case. Mechanical heart valves are located in the inlet and outlet positions of the ventricle to control flow direction. Air pulses that are generated by an external drive unit are used to periodically compress the flexible blood sac. (Reprinted with permission from Thoratec Corporation.)

Pneumatic Ventricular Assist Devices. Figure 2 shows a pneumatic ventricular assist device that was originally developed by the Pennsylvania State University and later purchased by Thoratec Corporation (Pleasanton, CA) (4). The ventricle contains a flexible blood sac made of segmented polyurethane that is housed within a rigid polysulfone case. This blood sac is extremely smooth to prevent the formation of clots (or thrombi). Mechanical heart valves are located in the inlet and outlet positions of the ventricle to control flow direction. Air pulses that are generated by an external drive unit are used to periodically compress the flexible blood sac. An automatic control system varies the cardiac output by adjusting the heart rate and the time for ventricular filling in response to an increase in filling pressure.

The device is placed outside the patient on the patient's abdomen (paracorporeal). The device can be used to assist a single ventricle, or simultaneously with an additional device that assists the both ventricles, as shown in Fig. 3. For the LVAD configuration (right pump in Fig. 3), the inlet cannula is inserted into the apex of the left ventricle and connected to the inlet port of the ventricular assist device. The outflow cannula is attached between the outflow port of the ventricular assist device and the ascending aorta. For the RVAD configuration (left pump in Fig. 3), the inlet cannula is connected to the right atrium and the outlet cannula to the main pulmonary artery. For both types of configurations, the inflow and outflow cannulae pass through the skin below the rib cage. Over 2850 implants have occurred worldwide with the longest duration of 566 days (5). An implantable version of this pump (with a titanium pump casing) was approved by the FDA in August of 2004 (6).



Permission Pending

Figure 3. Implant Location of Penn State/Thoratec Pneumatic Ventricular Assist Device in the RVAD (left) and LVAD (right) Configuration. For the LVAD configuration, the inlet cannula is inserted into the apex of the left ventricle and connected to the inlet port of the ventricular assist device. The outflow cannula is attached between the outflow port of the ventricular assist device and the ascending aorta. For the RVAD configuration, the inlet cannula is connected to the right atrium and the outlet cannula to the main pulmonary artery. For both types of configurations, the inflow and outflow cannulae pass through the skin below the rib cage. (Reprinted with permission from The Cleveland Clinic Foundation.)

Another type of pneumatic ventricular assist device is the Thoratec HeartMate IP (intraperitoneal). The HeartMate IP is an implantable blood pump that is connected to an external drive unit via a percutaenous air drive line (7). The interior of this device is shown in Fig. 4. A unique feature of this device is the use of a textured blood surface in the ventricle that promotes the formation of a cell layer. The cell layer is believed to decrease thrombus formation because the layer mimics the blood contacting surface of a blood vessel. Bioprosthetic heart valves are used to regulate the direction of flow. The HeartMate IP has been implanted in >1300 patients worldwide with the longest duration of 805 days.

Two other types of pneumatic devices include the BVS5000 and AB5000 (both made by ABIOMED, Danvers, MA). Both devices are intended to provide cardiac support as a bridge to transplant or until the native heart recovers. The BVS5000, illustrated in Fig. 5, is an external dual-chamber device that can provide support to one or both ventricles as a bridge to transplant (8). The chambers utilize polyurethane valves to regulate the flow direction. More than 6000 implants have been performed worldwide (9). The AB5000 is a pneumatically driven, paracorporeal device that is similar to a single ventricle of the AbioCor total artificial heart (described in a later section) (10). This device was approved by the FDA in October of 2003 as has been used in >88 patients. The average duration of support is 15 days with the longest duration of 149 days.

Additional types of pneumatic devices include the Berlin Heart Excor (Berlin Heart AG, Berlin, Germany), the

Figure 4. Interior of Thoratec HeartMate IP Pneumatic Ventricular Assist Device. A unique feature of this device is the use of a textured blood surface in the ventricle that promotes the formation of a cell layer. The cell layer is believed to decrease thrombus formation because the layer mimics the blood contacting surface of a blood vessel. (Reprinted with permission from Thoratec Corporation.)

MEDOS/HIA (Aachen, Germany), and the Toyobo Heart (National Cardiovascular Center, Osaka, Japan). The Berlin Heart Excor is available in a range of sizes (10–80 mL stroke volume) with either tilting disk or polyurethane valves, and has been implanted in >500 patients (11). The MEDOS/HIA system is also available in a range of sizes and has been implanted in >200 patients (12). The Toyobo LVAS has been implanted in >120 patients (13).

ELECTRIC VENTRICULAR ASSIST DEVICES

Electric ventricular assist devices mainly differ from their pneumatic counterparts in their source of power. Electric ventricular assist devices are typically placed within the

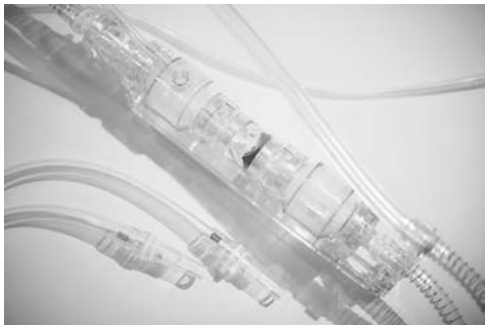


Figure 5. ABIOMED BVS 5000 Pneumatic Ventricular Assist Device. The BVS5000 is an external dual-chamber device that can provide support to one or both ventricles as a bridge to transplant (8). The chambers utilize polyurethane valves to regulate the flow direction. (Reprinted with permission from ABIOMED, Inc.)

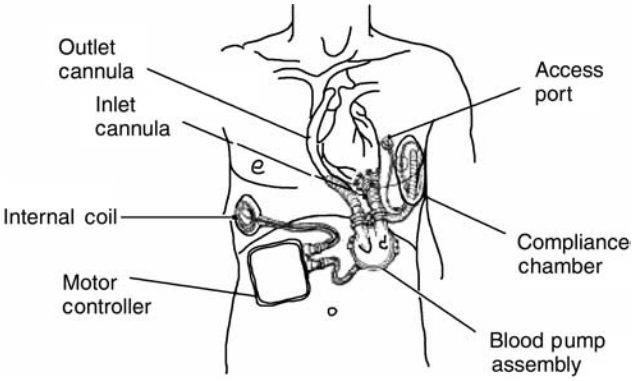


Figure 6. Penn State/Arrow Electric Ventricular Assist Device (LionHeart). The blood pump assembly utilizes a rollerscrew energy converter with a pusher plate. The motion of the pusher plate compresses the blood sac and ejects blood from the ventricle. Mechanical heart valves are used to control the direction of flow into and out of the pump. Energy passes from the external power coil to the subcutaneous (internal) coil by inductive coupling via the transcutaneous energy transmission system (TETS). (Reprinted with permission from Arrow International.)

abdominal cavity. The inlet cannula is inserted into the apex of the native left ventricle and connected to the inlet port of the device. The outlet cannula is attached between the outflow port of the device and the ascending aorta via an end-to-side anastomosis. These types of devices can be used as either bridge-to-transplant or as permanent implants (destination therapy).

Figure 6 illustrates an electric ventricular assist device (LionHeart) developed by the Pennsylvania State University in conjunction with Arrow International (Reading, PA) (14). The blood pump assembly utilizes a rollerscrew energy converter with a pusher plate. The motion of the pusher plate compresses the blood sac and ejects blood from the ventricle. Mechanical heart valves are used to control the direction of flow into and out of the pump. Energy passes from the external power coil to the subcutaneous (internal) coil by inductive coupling via the transcutaneous energy transmission system (TETS). The controller and internal battery supply are also implanted in the abdomen. The internal battery permits operation without the external power coil for ~ 20 min. Air displaced by the blood pump housing enters the polyurethane compliance chamber. Because the air in the compliance chamber can slowly diffuse across the wall of the compliance chamber, the air in the chamber is periodically replenished via the subcutaneous access port. The LionHeart is intended to be used as destination therapy. This device was approved for use in Europe in 2003.

Another type of electric ventricular assist devices is the Novacor LVAS (left ventricular assist system), produced by WorldHeart Corporation (Ottawa, ON). The Novacor LVAS, illustrated in Fig. 7, contains a polyurethane blood sac that is compressed between dual pusher plates (15). The pusher plates are actuated by a solenoid that is coupled to the plates via springs. Bioprosthetic heart valves are utilized to control the direction of flow. A percutaneous power line connects the pump to an external battery pack and controller. The Novacor LVAS has been implanted in

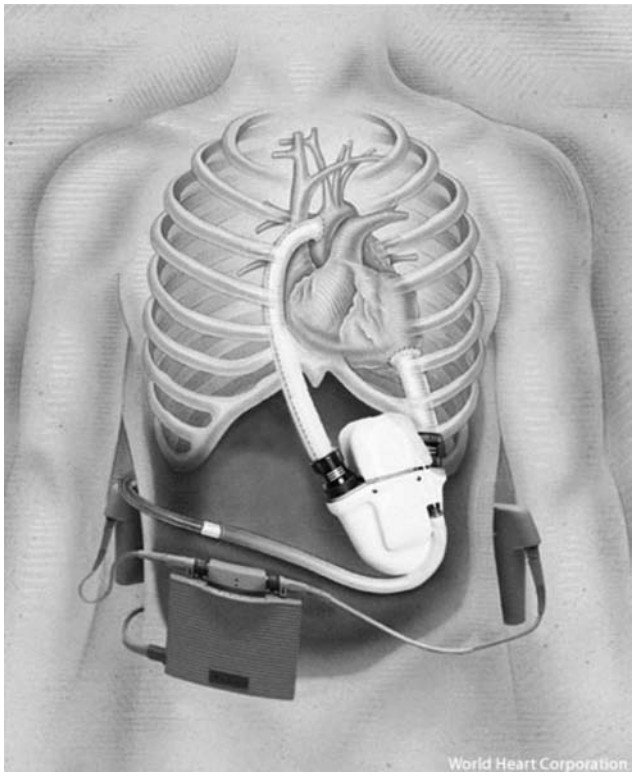


Figure 7. WorldHeart Novacor Electric Ventricular Assist Device. The Novacor contains a polyurethane blood sac that is compressed between dual pusher plates. The pusher plates are actuated by a solenoid that is coupled to the plates via springs. A percutaneous power line connects the pump to an external battery pack and controller. (Reprinted with permission from World Health Corporation, Inc.)

over 1500 patients worldwide. The longest implant duration is > 6 years. No deaths have been attributed to device failure with only 1.4% of the devices needing replacement. The Novacor LVAS is approved as a bridge-to-transplant in the United States and Europe and is in clinical trials for destination therapy in the United States.

The HeartMate XVE (illustrated in Fig. 8) is a derivative to the HeartMate IP (7). The HeartMate XVE uses an electric motor and pusher plate system to pump blood. A percutaneous power line is used to connect the pump to an external battery pack and controller. The HeartMate VE (an earlier version of the XVE) and XVE have been implanted in > 2800 patients worldwide with the longest duration of 1854 days. The HeartMate SNAP-VE was recently approved by the FDA as destination therapy.

The Randomized Evaluation of Mechanical Assistance for the Treatment of Congestive Heart Failure (REMATCH) study examined the clinical utility of ventricular assist devices (16). Patients with end-stage heart failure who were ineligible for cardiac transplantation were split into two groups. The first group ($n = 68$) received the HeartMate VE LVAS while the second group ($n = 61$) received optimal medical management. The results showed a reduction of 48% in the risk of death from any cause in the LVAD group versus the medical-therapy group ($p = 0.001$). The 1 year survival was 52% for the VAD group and 25% for the medical-therapy group ($p = 0.002$). The 2 year survival

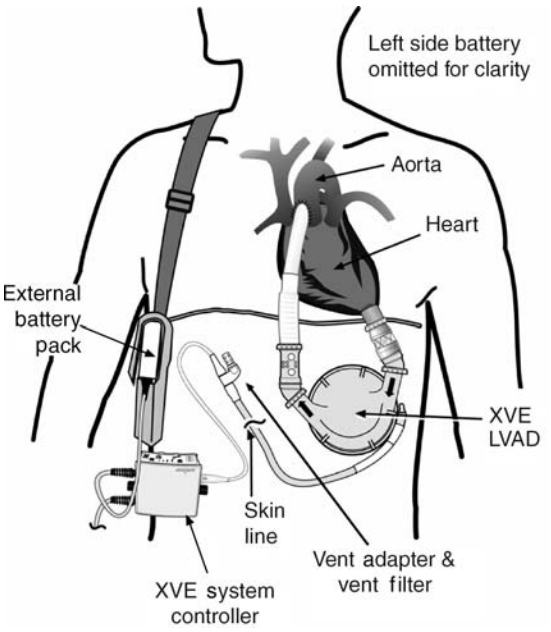


Figure 8. Thoratec HeartMate XVE Electric Ventricular Assist Device. The HeartMate XVE uses an electric motor and pusher plate system to pump blood. A percutaneous power line is used to connect the pump to an external battery pack and controller. (Reprinted with permission from Thoratec Corporation.)

was 23% for the VAD group and 8% for the medical-therapy group ($p = 0.09$). Finally, the median survival was 408 days for the VAD group and 150 days for the medical-therapy group. This study clearly showed the clinical utility of ventricular assist devices.

Continuous Flow

Continuous flow ventricular assist devices deliver nonpulsatile flow. Consequently, they do not require heart valves to regulate the direction of blood flow. Continuous flow ventricular assist devices are classified as either centrifugal flow or axial flow pumps based on the direction of the flow as it passes through the pump. These types of pumps are implanted in a similar fashion as their pulsatile counterparts.

Continuous flow assist devices have several potential advantages over pulsatile systems. First, these devices are typically smaller than their pulsatile counterparts and can be used in smaller patients (such as small adults and children). In addition, these pumps have fewer moving parts and are simpler devices than pulsatile systems. These types of devices typically require less energy to operate than the pulsatile pumps.

However, continuous flow pumps have several potential disadvantages. The main disadvantage is that the long-term effects of continuous flow in patients are unknown. Some studies suggest that continuous flow results in lower tissue perfusion (17,18). In addition, these types of devices typically have higher fluid stresses than their pulsatile counterparts, potentially exposing blood components to stress levels that may damage or destroy the cells. However, due to the short residence time of the blood components within these pumps, the potential for damage or destruction is reduced (19). Finally, feedback control mechanisms for altering pump speed and flow in response to patient demand are complex and unproven.

Centrifugal Flow Ventricular Assist Device. In a centrifugal flow ventricular assist device, the direction of the outlet port is orthogonal (at a right angle) to the direction of the inlet port. Blood flowing into a centrifugal pump moves onto a spinning impeller. This causes the blood to be propelled away from the impeller due to centrifugal forces. The blood is then channeled to the outlet port by a circular casing (known as the volute) around the impeller. Finally, the blood is discharged through the outlet at a higher pressure than the inlet pressure.

The impeller typically consists of various numbers and geometric configurations of blades, cones, or disks. Typical motor speeds (or rotation rates) for centrifugal flow pumps range vary from 1500 to 5000 rpm (revolutions per minute). This results in flow rates of 2–10 $L \cdot min^{-1}$. Many centrifugal flow pumps utilize electromagnetic impellers that do not make any contact with the interior of the pump when the impeller is spinning. The inlet and outlet ports are connected to the native ventricle and the aorta, respectively, as described previously for pulsatile electric ventricular assist devices.

A major drawback with centrifugal flow pumps is that they are outlet pressure sensitive and may not produce flow if the outflow pressure (the pressure that the pump is working against) becomes greater than the outlet pressure. When this happens, the impeller will continue to spin without producing any flow. In order for the pump to produce flow, either the outflow pressure must be reduced or the impeller speed must be increased (to increase the outlet pressure).

The Bio-Pump (Medtronic BioMedicus, Inc., Minneapolis, MN), shown in Fig. 9, is an extracorporeal, centrifugal flow pump that was originally developed for cardiopulmonary bypass (20). It has been used to provide support for one or both ventricles as a bridge to transplant for short periods

Figure 9. Medtronic BioMedicus Bio-Pump Centrifugal Flow Ventricular Assist Device. The Bio-Pump is an extracorporeal, centrifugal flow pump that was originally developed for cardiopulmonary bypass. It has been used to provide support for one or both ventricles as a bridge to transplant for short periods. (Reprinted with permission from Medtronic, Inc.)

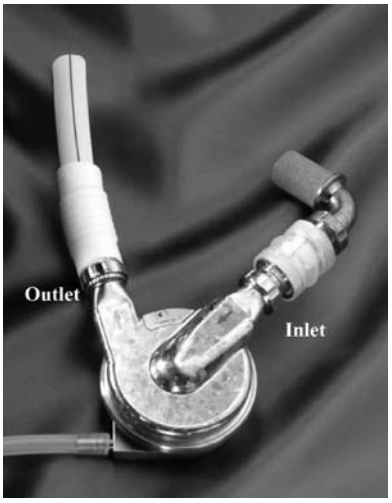


Figure 10. Thoratec HeartMate III Centrifugal Flow Ventricular Assist Device. The HeartMate III is a centrifugal pump that features a magnetically levitated impeller. (Reprinted with permission from Thoratec Corporation.)

(5 days or less). The pump consists of an acrylic pump head with inlet port and outlet ports placed at right angles to each other. The impeller consists of a stack of parallel cones within a conical acrylic housing. A single magnetic drive unit is coupled with a magnet in the impeller. The pump is driven by an external motor and power console. Two different sizes are available to provide support for both adults and children. Recipients of the Bio-Pump have had mixed results (21). The Sarns/3M Centrifugal system (Terumo, Ann Arbor, MI) is another centrifugal pump that is used primarily for cardiopulmonary bypass (22).

The HeartMate III (Thoratec), shown in Fig. 10, is a centrifugal pump that features a magnetically levitated impeller (23,24). The entire pump is fabricated from titanium. The interior of the pump uses the same type of textured blood contacting surfaces utilized in the HeartMate VE. In addition, the HeartMate III incorporates a TETS that permits it to be fully implantable as a permanent device for destination therapy. The controller is designed to respond automatically to the patient's needs and to permit both pulsatile and continuous flow. This pump is currently under development. Other centrifugal pumps that utilize a magnetically levitated impeller include the HeartQuest (MedQuest, Salt Lake City, UT) (25) and the Duraheart (Terumo) (26). The Duraheart was first implanted in 2004.

Two centrifugal flow pumps utilize hydrodynamic forces, rather than magnetic levitation, to suspend the impeller: the CorAide (Arrow International) (27) and the VentrAssist (Ventrcor Limited, Chatswood, Australia) (28). The CorAide (shown in Fig. 11) began clinical trials in Europe in 2003 (29), while the VentrAssist (shown in Fig. 12) began clinical trials in Europe in 2004 (30).

Axial Flow Ventricular Assist Devices. An axial flow ventricular assist device is also composed of an impeller spinning in a stationary housing. However, the blood that flows into and out of the device travels in the same direction as the axis of rotation of the impeller. The impeller transfers energy to the blood by the propelling, or lifting, action of the vanes on the blood. Stators (stationary flow straighteners) stabilize the blood flow as it enters and exits the impeller. Magnets are embedded within the impeller and are coupled with a rotating magnetic field on the housing. The pumps are typically constructed of titanium.

Figure 11. Arrow CorAide Centrifugal Flow Ventricular Assist Device. The CorAide utilizes a hydrodynamic bearing, rather than magnetic levitation, to suspend the impeller. (Reprinted with permission from Arrow International, Inc.)

Axial flow pumps run at speeds of 10,000–20,000 rpm, generating flow rates of up to 10 $L \cdot min^{-1}$. These high motor speeds are not expected to cause excessive hemolysis (damage to blood components) because of the limited exposure of blood within the axial flow pump (19). Like centrifugal pumps, axial flow pumps are also outlet pressure sensitive and may not produce flow in cases when the outflow pressure exceeds the outlet pressure. Mechanical bearings are typically used to support the impeller within the stator.

Figure 12. Ventracor VentrAssist Centrifugal Flow Ventricular Assist Device. The VentrAssist utilizes a hydrodynamic bearing, rather than magnetic levitation, to suspend the impeller. (Reprinted with permission from Ventracor, Inc.)

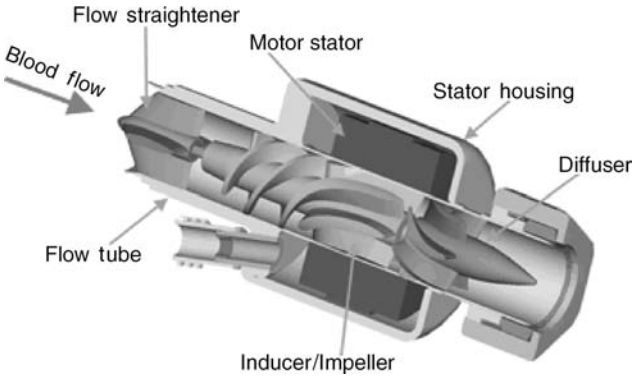


Figure 13. MicroMed Debakey Axial Flow Ventricular Assist Device. The MicroMed is connected to an external controller and power unit. The pump speed is varied manually to meet the needs of the patient. (Reprinted with permission from Micromed Technology, Inc.)

Figure 13 shows the MicroMed Debakey VAD (MicroMed Technology, Houston, TX) axial flow pump. This device operates from 7500 to 12,000 rpm and can provide flows up to 10 $L \cdot min^{-1}$ (31). The flow curves, speed, current and power are displayed in a bedside monitor unit. A pump motor cable along with the flow probe wire exit transcutaneously from the implanted device and connect to the external controller and power unit. The pump speed is varied manually to meet the needs of the patient. The pump can be actuated by two 12 V dc batteries for 4–6 h. This device was approved in Europe in 2001 (32). Clinical trials in the United States began in 2000. Over 280 patients have received the MicroMed Debakey VAD as of January 2005 worldwide. Although this device was originally approved as a bridge to transplant, clinical trials are underway to use the device for destination therapy.

Figure 14 shows the HeartMate II (Thoratec) axial flow ventricular assist device. The rotating impeller is surrounded by a static pump housing with an integral motor (33). The pump's speed can be controlled either manually or by an automatic controller that relies on an algorithm based on pump speed, the pulsatility of the native heart, and motor current. The HeartMate II is designed to operate between 6000 and 15,000 rpm and deliver as much as 10 $L \cdot min^{-1}$. The initial version of this device is powered through a percutaneous small-diameter electrical cable connected to the system's external electrical controller. A fully implantable system utilizing a TETS is under development. The first implant HeartMate II implant occurred in 2000 (34). Clinical trials in Europe and the United States are ongoing. This device is intended for both bridge to transplant and destination therapy.

Figure 15 illustrates the Jarvik 2000 (Jarvik Heart, New York). The Jarvik 2000 is intraventricular axial flow pump. The impeller is a neodymium–iron–boron magnet, which is housed inside a welded titanium shell and supported by ceramic bearings (35). A small, percutaneous cable delivers power to the impeller. All of the blood-contacting surfaces are made of highly polished titanium. The normal operating range for the control system is 8000–12,000 rpm, which generates an average pump flow rate of 5 $L \cdot min^{-1}$. The pump is placed within the left ventricle with a sewing cuff sutured to the ventricle, eliminating the need for an inflow cannula. Over 100 patients have received the Jarvik 2000 as a bridge to transplant or destination therapy, with the longest implant duration of >4 years (36).

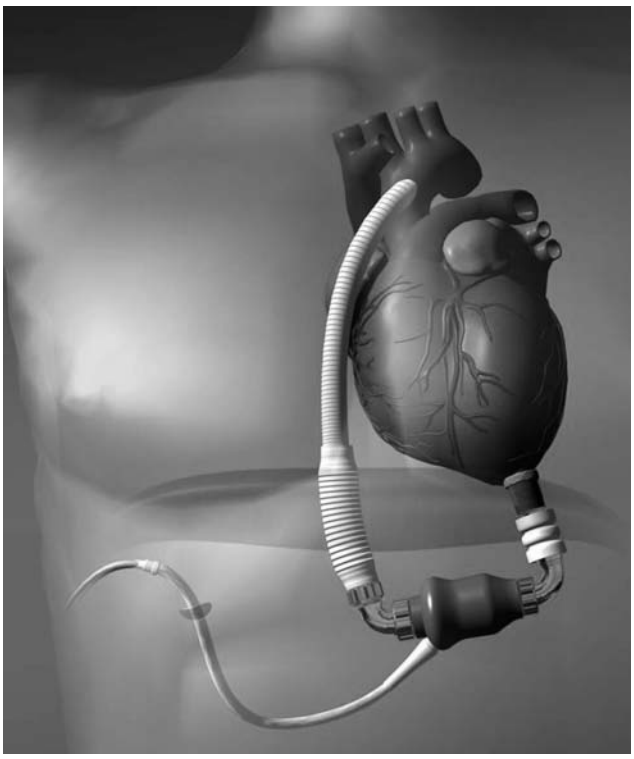


Figure 14. Thoratec HeartMate II Axial Flow Ventricular Assist Device. The rotating impeller is surrounded by a static pump housing with an integral motor. The pump's speed can be controlled either manually or by an automatic controller that relies on an algorithm based on pump speed, the pulsatility of the native heart, and motor current. (Reprinted with permission from Thoratec Corporation.)

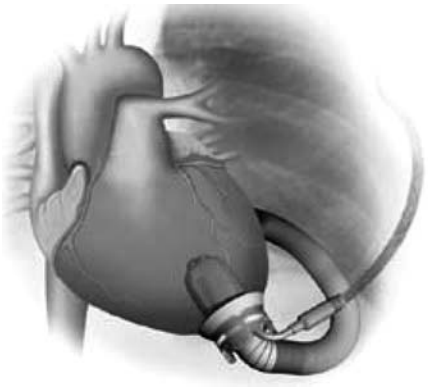


Figure 15. Jarvik 2000 Axial Flow Ventricular Assist Device. Unlike most other axial flow devices, the Jarvik 2000 is intraventricular axial flow pump. The impeller is a neodymium-iron-boron magnet, which is housed inside a welded titanium shell and supported by ceramic bearings. (Reprinted with permission from Jarvik Heart, Inc.)

TOTAL ARTIFICIAL HEARTS

The total artificial heart (TAH) is designed to support both the pulmonary and systemic circulatory systems by replacing the native heart. Two types of artificial hearts have been developed: pneumatic and electric.

Pneumatic Total Artificial Heart

A pneumatic total artificial heart is composed of two ventricles that replace the native left and right ventricle. Each ventricle is of similar design to the Penn State/Thoratec pneumatic ventricular assist device (as described in a previous section) (4). Both ventricles are implanted within the chest. The air pulses are delivered to the ventricles via percutaneous drivelines. An automatic control system varies cardiac output by adjusting the heart rate and the time for ventricular filling in response to an increase in filling pressure.

Pneumatic total artificial hearts are currently used as a bridge to transplant. Several different pneumatic artificial hearts have been used clinically around the world. The only pneumatic TAH approved as a bridge to transplant in the United States is the CardioWest (SynCardia, Tucson, AZ) TAH, illustrated in Fig. 16 (37). The CardioWest (with a stroke volume of 70 mL) is based on the Jarvik-7, which has a stroke volume of 100 mL. A study of 81 recipients of the CardioWest revealed a survival rate to transplant of 79% and a 1 year survival rate of 70%.

Pneumatic total artificial hearts have also been used as a permanent replacement device. The Jarvik-7 pneumatic TAH was permanently implanted in five patients (38). Although the longest survivor lived for 620 days, all five patients had hematologic, thromboembolic, and infectious complications. The pneumatic artificial heart is no longer considered for permanent use because of infections associated with the percutaneous pneumatic drive lines and quality of life issues related to the bulky external pneumatic drive units.

Electric Total Artificial Heart

The electric TAH is completely implantable and is designed for permanent use. The Penn State/3M Electric TAH is shown in Fig. 17. The artificial heart is composed of two blood pumps that are of similar design to the Penn State/Arrow electric ventricular assist device (39). However, the electric TAH uses a single implantable energy converter that alternately drives each ventricle. The implantable controller adjusts the heart rate in response to ventricular filling and maintains left–right balance. The design for this system was completed in 1990 and was the first to incorporate the controller, transcutaneous energy transmission system (TETS), telemetry, and internal power (via rechargeable batteries) into a completely implantable system. The Penn State electric TAH has been successfully implanted in animals for >1 year without thromboembolic complications. In 2000, ABIOMED acquired the rights to the Penn State/3M Electric TAH.

The ABIOMED AbioCor TAH, illustrated in Fig. 18, uses an electrohydraulic energy converter to alternately compress each blood sac (40,41). In addition, the AbioCor uses polymer valves to control flow into and out of each ventricle. The AbioCor is currently undergoing clinical

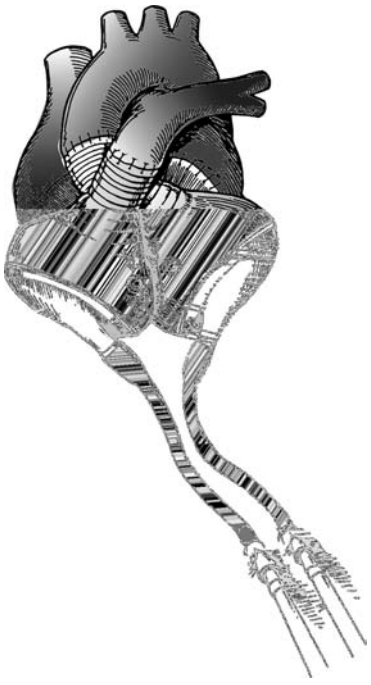


Figure 16. SynCardia CardioWest Pneumatic Total Artificial Heart. The CardioWest is based on the Jarvik-7 and is the only pneumatic TAH approved as a bridge to transplant in the United States. (Reprinted with permission from SynCardia Systems, Inc.)

Figure 17. Penn State/3M Electric Total Artificial Heart. This artificial heart is composed of two blood pumps that are of similar design to the Penn State/Arrow electric ventricular assist device. However, the electric TAH uses a single implantable energy converter that alternately drives each ventricle.

Figure 18. ABIOMED AbioCor Electric Total Artificial Heart. The AbioCor TAH uses an electro hydraulic energy converter to alternately compress each blood sac. The AbioCor is currently undergoing clinical trials in the United States. Two patients were discharged from the hospital, with one patient surviving for >1 year. (Reprinted with permission from ABIOMED, Inc.)

trials in the United States. Fourteen critically ill patients (with an 80% chance of surviving <30 days) have been implanted. Two patients were discharged from the hospital (one to home), with one patient surviving for >1 year. The causes of death were typically end organ failure and strokes. One pump membrane wore out at 512 days. Smaller, improved totally implantable artificial hearts are currently under development.

FUTURE DIRECTIONS OF RESEARCH

The ventricular assist devices and total artificial hearts presented in this article successfully provide viable cardiac support by either assisting or replacing the native heart. However, there are several areas for future research on artificial hearts. These include the following: Power sources to permit longer intervals between battery changes; Improved control schemes for both pulsatile and nonpulsatile devices that enhance the response of the cardiac assist device to meet physiologic demands; Decrease thromboembolic events associated by modifying the device geometry and/or blood-contacting materials; Determine the long-term effects of continuous, nonpulsatile flow; Decrease incidence of infection by the elimination of all percutaneous lines and creating smaller implantable electronic components; Reduced pump sizes to fit smaller adults, children, and infants; Increased reliability for 5 or more years to 95% (with a 95% confidence level).

Significant progress has been made in the last 20 years. One can only imagine what the next 20 years will bring!

ACKNOWLEDGMENTS

The author would like to acknowledge the support of William S. Pierce, M.D., Gerson Rosenberg, Ph.D., David B. Geselowitz, Ph.D., and the past and present faculty, staff, and graduate students at the Division of Artificial Organs at the Penn State College of Medicine and the Department of Bioengineering at the Pennsylvania State University. The financial support from the National Institutes of Health is also recognized.

BIBLIOGRAPHY

Cited References

1. American Heart Association. Heart Disease and Stroke Statistics—2005 Update. Dallas (TX): American Heart Association; 2005.
2. Organ Procurement and Transplantation Network Data as of May 29, 2005. Available at http://www.optn.org. 2005.
3. Guyton A, Hall J. Textbook of Medical Physiology. Philadelphia: Saunders; 2000.
4. Richenbacher WE, Pierce WS. In: Braunwald HE, editor. Assisted Circulation and the Mechanical Heart Disease: A Textbook of Cardiovascular Medicine, 6th ed. Philadelphia: Saunders; 2001. p 534–547.
5. Thoratec VAD Clinical Results, Available at http://www.thoratec.com. Accessed Nov 2004.
6. Thoratec Corporation Press Release, Aug. 5, 2004 [Online]. Thoratec. Available at http://www.thoratec.com/index.htm. [5/19/2005]. Accessed 2005.
7. HeartMate LVAS Clinical Results, Nov. 2004. Available at http://www.thoratec.com/index.htm. Accessed 2004.
8. Berger EE. ABIOMED's BVS 5000 biventricular support system. J Heart Lung Transplant 2004;23(5):653.
9. Clinical Information, BVS Clinical Update 2004 [online] ABIOMED. Available at http://www.abiomed.com/clinical information/BVS5000Update.cfm. [5/19/2005]. Accessed 2004.
10. Clinical Information, AB5000 Clinical Update 2004 [online] ABIOMED. Available at http://www.abiomed.com/clinical information/AB5000Update.cfm. [5/19/2005]. Accessed 2004.
11. Berlin Heart AG- The VAD System [online] Berlin Heart. http://www.berlinheart.com/download/system.pdf. [5/17/2005]. 2003.
12. Reul H. The MEDOS/HIA system: development, results, perspectives. Thorac Cardiovasc Surg 1999;47(Suppl 2):311–315.
13. Takano H, Nakatani T. Ventricular assist systems: experience in Japan with Toyobo pump and Zeon pump. Ann Thorac Surg 1996;61(1):317–322.
14. El-Banayosy A, et al. Preliminary experience with the LionHeart left ventricular assist device in patients with end-stage heart failure. Ann Thorac Surg 2003;75(5):1469–1475.
15. Novacor LVAS-Products-WorldHeart [Online] WorldHeart. Available at http://www.worldheart.com/products/novacor lvas.cfm. [5/17/2005]. Accessed 2005.
16. Rose EA, et al. Long-term mechanical left ventricular assistance for end-stage heart failure. N Engl J Med 2001;345(20): 1435–1443.
17. Baba A, et al. Microcirculation of the bulbar conjunctiva in the goat implanted with a total artificial heart: effects of pulsatile and nonpulsatile flow. ASAIO J 2004;50(4):321–327.
18. Undar A. Myths and truths of pulsatile and nonpulsatile perfusion during acute and chronic cardiac support. Artif Organs 2004;28(5):439–443.
19. Arora D, Behr M, Pasquali M. A tensor-based measure for estimating blood damage. Artif Organs 2004;28(11):1002–1015.

20. Noon GP, BallJr JW, Papaconstantinou HT. Clinical experience with BioMedicus centrifugal ventricular support in 172 patients. Artif Organs 1995;19(7):756–760.
21. Noon GP, Lafuente JA, Irwin S. Acute and temporary ventricular support with BioMedicus centrifugal pump. Ann Thorac Surg 1999;68(2):650–654.
22. Williams M, Oz M, Mancini D. Cardiac assist devices for end-stage heart failure. Heart Dis 2001;3(2):109–115.
23. Bourque K, et al. HeartMate III: pump design for a centrifugal LVAD with a magnetically levitated rotor. ASAIO J 2001;47(4):401–405.
24. Bourque K, et al. Incorporation of electronics within a compact, fully implanted left ventricular assist device. Artif Organs 2002;26(11):939–942.
25. Chen C, et al. A magnetic suspension theory and its application to the HeartQuest ventricular assist device. Artif Organs 2002;26(11):947–951.
26. Terumo Heart, Inc. Press Release, Jan. 19, 2004 [Online] : Available at http://www.terumocvs. com/newsandevents/rendernews.asp?newsId=5. Terumo [5/17/2005]. Accessed 2005.
27. Doi K, et al. Preclinical readiness testing of the Arrow International CorAide left ventricular assist system. Ann Thorac Surg 2004;77(6):2103–2110.
28. James NL, et al. Implantation of the VentrAssist Implantable Rotary Blood Pump in sheep. ASAIO J 2003;49(4):454–458.
29. Arrow International-Cardiac Assist-CorAide [online] Arrow International. Available at http://www.arrowintl.com/products/cardassist/. [2/28/05]. Accessed 2003.
30. Ventracor Half-Year Report Summary, Feb 16 2005 [online] Ventracor. Available at http://www.ventracor.com/default.asp?cp=/news/newsitem.asp%3FnewsID%3D316. [5/19/2005]. Accessed 2005.
31. Goldstein DJ. Worldwide experience with the MicroMed DeBakey Ventricular Assist Device as a bridge to transplantation. Circulation 2003;108(Suppl 1):II272–277.
32. MicroMed Technology Press Release, Jan. 20 2005 [Online] MicroMed. Available at http://www.micromedtech.com/news/01-20-05.htm. [5/19/2005]. Accessed 2005.
33. Burke DJ, et al. The Heartmate II: design and development of a fully sealed axial flow left ventricular assist system. Artif Organs 2001;25(5):380–385.
34. Frazier OH, et al. First clinical use of the redesigned HeartMate II left ventricular assist system in the United States: a case report. Tex Heart Inst J 2004;31(2):157–159.
35. Westaby S, et al. The Jarvik 2000 Heart. Clinical validation of the intraventricular position. Eur J Cardiothorac Surg 2002;22(2):228–232.
36. Frazier OH, et al. Use of the Flowmaker (Jarvik 2000) left ventricular assist device for destination therapy and bridging to transplantation. Cardiology 2004;101(1–3):111–116.
37. Copeland JG, et al. Cardiac replacement with a total artificial heart as a bridge to transplantation. N Engl J Med 2004; 351(9):859–867.
38. DeVries WC. The permanent artificial heart. Four case reports. JAMA 1988;259(6):849–859.
39. Weiss WJ, et al. Steady state hemodynamic and energetic characterization of the Penn State/3M Health Care Total Artificial Heart. ASAIO J 1999;45(3):189–193.
40. Dowling RD, et al. The AbioCor implantable replacement heart. Ann Thorac Surg 2003;75(6 Suppl):S93–S99.
41. ABIOMED AbioCor Press Release, Nov. 4, 2004 [Online] ABIOMED. Available at http://www.abiomed.com/news/Fourteenth-AbioCor-Patient.cfm. [5/19/2005]. Accessed 2004.

Reading List

Lee J, et al. Reliability model from the in vitro durability tests of a left ventricular assist system. ASAIO J 45(6):595–601 1999.
Raman J, Jeevanadam V. Destination therapy with ventricular assist devices. Cardiology 101(1–3):104–110 2004.
Rosenberg G. Artificial Heart and Circulatory Assist Devices. In: Bronzino J, editor. The Biomedical Engineering Handbook, Boca Raton (FL): CRC Press; 1995: 1839–1846.
Song X, et al. Axial flow blood pumps. ASAIO J 49(4):355–364 2003.
Stevenson LW, Kormos RL. Mechanical Cardiac Support 2000: Current applications and future trial design. J Thorac Cardiovasc Surg 121(3):418–424 2001.
Zapanta CM, et al. Durability testing of a completely implantable electric total artificial heart. ASAIO J 51(3):214–223 2005.

See also HEART VALVE PROSTHESES; HEMODYNAMICS; QUALITY OF LIFE MEASURES, CLINICAL SIGNIFICANCE OF.

HEART–LUNG MACHINES

DAVID D'ALESSANDRO
ROBERT MICHLER
Montefiore Medical Center
Bronx, New York

INTRODUCTION

The heart–lung machine is perhaps the most important contribution to the advancement of surgery in the last century. This apparatus was designed to perform the functions of both the human heart and the lungs allowing surgeons to suspend normal circulation to repair defects in the heart. The development of a clinically safe and useful machine was the rate-limiting step to the development of modern cardiac surgery. Since its inception, the heart–lung machine has enabled the surgical treatment of congenital heart defects, coronary heart disease, valvular heart disease, and end-stage heart disease with heart transplantation and mechanical assist devices or artificial hearts.

The heart–lung machine consists of several components that together make up a circuit that diverts blood away from the heart and lungs and returns oxygenated blood to the body. Commercial investment and production of these components has resulted in wide variability in the design of each, but the overall concept is preserved. During an operation, a medical specialist known as a perfusionist operates the heart–lung machine. The role of the perfusionist is to maintain the circuit, adjust the flow as necessary, prevent air and particulate emboli from entering the circulation, and maintain the various components of the blood within physiologic parameters.

HISTORY OF CARDIAC SURGERY AND ASSISTED CIRCULATION

The emergence of modern heart surgery and the ability to correct congenital and acquired diseases of the heart were dependent on the work of innovative pioneers who developed a means to stop the heart while preserving blood flow to the remainder of the body. This new technology was aptly named *cardiopulmonary bypass* (CPB), which simply

means circulating blood around the heart and lungs. Since the inception of the heart–lung machine, continued refinements and widespread adoption of CPB led to the rapid growth of congenital and adult heart surgery.

Lessons learned during World War II and corresponding advances in critical care emboldened surgeons to consider surgical solutions for diseases of the heart, an organ long considered to be inoperable. One such surgeon was Dr. Dwight Harken who pioneered closed heart surgery, which he initially used to remove foreign bodies from the heart such as bullets and shrapnel. Having achieved success with this approach, he and others modified the techniques, developing a blinded technique for performing closed valvuloplasty primarily used to ameliorate rheumatic valvular disease. Although this method proved safe and reproducible, its applicability to other diseases of the heart was limited.

The next leap came when surgeons attempted open-heart procedures using brief periods of total circulatory arrest. As the brain is highly sensitive to hypoxic injury, very few patients were successfully treated with his approach. The safety of circulatory arrest was greatly increased when Dr. Bill Bigelow at the University of Minnesota introduced the concept of induced hypothermia. This method of lowering body temperature to reduce metabolic demand provided protection for the oxygen-starved organs allowing for modestly longer periods of circulatory arrest. Inspired by Bigelow's work, Lewis and Taufic first used this approach clinically on September 2, 1952, at the University of Minnesota (1). Under moderate total body hypothermia, Lewis and Taufic used a short period of circulatory arrest to repair a congenital defect in a 5 year-old girl. This was a landmark achievement in surgery and marks the true beginning of open-heart surgery. For the first time, surgeons had the ability to open the heart to repair defects under direct vision. Despite this great achievement, however, the relatively brief periods of circulatory arrest that this technique provided were sufficient only for the repair of simple defects and did little to broaden the scope of surgically treatable cardiac diseases.

The development of assisted circulation was a quantum leap in the field of cardiac surgery. Although investigation into mechanical means of circulation began during the early part of the twentieth century, an effective and safe design would not emerge for several years. An alternative approach named *cross-circulation* was used for several years in the interim. Dr. C. Walt Lillehei, again at the University of Minnesota, was the first to use this technique clinically when on March 26, 1954, when he repaired a VSD in a 12 month-old infant. During this operation, the child's mother acted as a blood reservoir, a pump, and an oxygenator, allowing a safe period of extended circulatory arrest to repair a complicated congenital defect. The child's circulation was connected in series to her mother's diverting oxygen poor blood away from the patient's heart and lungs and returning oxygen-saturated blood to her arterial system. Although many were amazed by and congratulatory of Dr. Lillehei's efforts, critics were outspoken of their disapproval of a technique that risked the death of two patients for the benefit of one. Nevertheless, these early successes provided proof of the concept and a mechanical substitute for cross-circulation soon followed.

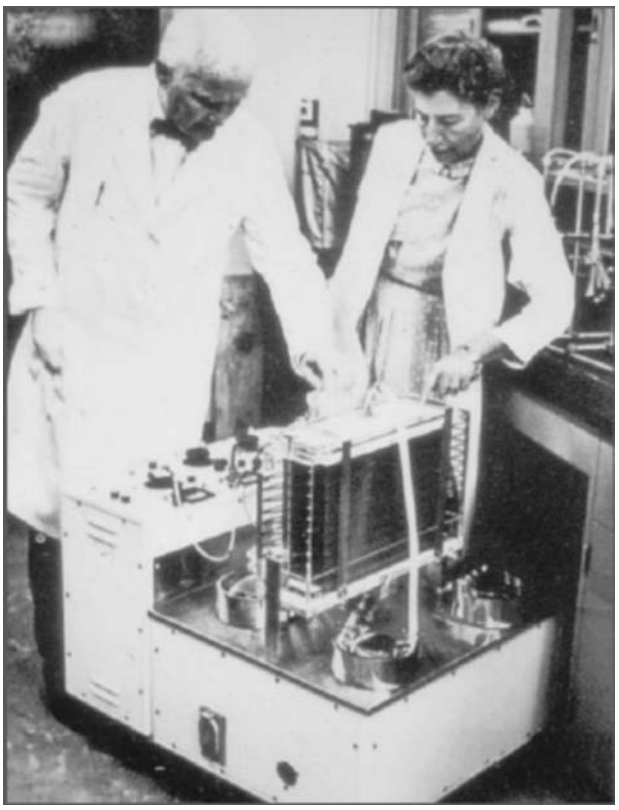


Figure 1. John and Mary Gibbon with their heart–lung machine. (Reprinted with permission from Thomas Jefferson University Archives.)

John and Mary Gibbon are credited with building the first usable CPB machine, a prototype of which John Gibbon employed in 1935 to support a cat for 26 min (2) (Fig. 1). This pump, which used two roller pumps and other similar early designs, were traumatic to blood cells and platelets and allowed for easy air entry into the circulation, which often proved catastrophic. Later, in 1946, Dr. Gibbon in collaboration with Thomas Watson, then chairman of IBM, made further refinements. Together they were able to successfully perform open-heart surgery in dogs, supporting the animals for period exceeding 1 h (3). Finally on May 6, 1953, Dr. Gibbon used his heart–lung machine to successfully repair an atrial septal defect in an 18 year-old girl, marking the first successful clinical use of a heart–lung machine to perform open-heart surgery. Gibbon's first attempt in 1952 followed two unsuccessful attempts by Clarence Dennis et al. in 1951 (4), and it too ended in failure. Sadly, subsequent failures led Gibbon to finally abandon this new technology, finally giving up heart surgery completely.

Kirklin's group at the Mayo Clinic modified the Gibbon pump oxygenator and reported a series of eight patients in 1955 with a 50% survival (5). Around the same time, Frederick Cross and Earl Kay developed a rotating disk oxygenator that had a similar effectiveness (6,7). This apparatus was similar in design to that used by Dennis et al. several years earlier (8). While this Kay-Cross unit became commercially available, it shared many of the

Figure 2. Typical cardiopulmonary bypass circuit demonstrating the various components. This circuit contains four accessory pumps for suction/venting and the administration of cardioplegia. (Reprinted with permission from Ref. 13.)

limitations of the Gibbon system. These pumps were cumbersome, difficult to clean and sterilize, and were inefficient and thus wasteful of blood.

Collaboration between Richard DeWall and Lillehei provided the next advance in pump oxygenators, and in 1956, they reported their first clinical experience using a bubble oxygenator in seven patients, five of whom survived (9). Further refinements by Gott et al. (10) resulted in a reproducible, disposable plastic sheet oxygenator, a system that was readily accepted by pioneering centers around the world. The bubble oxygenator became the predominant pump oxygenator for the next three decades. More recently, the membrane oxygenator, an early design of which was described by Kolff et al. in 1956 (11), has replaced it. Lande et al. described the first commercially available disposable membrane oxygenator (12), but its advantages to the bubble oxygenator were not readily apparent. By the mid-to-late 1980s, advanced microporous membrane oxygenators began to supplant the bubble oxygenator in the clinical arena and they remain the predominant design in use today (Fig. 2).

THE CIRCUIT

The first requirement of a CPB circuit is a means to evacuate and return blood to the circulation. Special drainage and return catheters known as cannulae have been devised for this purpose. One or more venous cannulae are placed in a central vein or in the heart to empty the heart. The venous cannulae are connected to a series of poly(vinyl chloride) tubing that deliver the blood to the remainder of the circuit.

The next component is a reservoir that collects and holds the blood. The reservoir is often used to collect shed blood from the operative field and to store excess blood volume while the heart and lungs are stopped. Drainage of the blood is accomplished by gravitational forces and is sometimes assisted with vacuum. The movement of blood through the remainder of the circuit, however, requires the use of a pump. Many pump designs have been used over the years, but the most common in current use are the roller pump and the centrifugal pump. A roller pump is a positive displacement pump and has been the most prolific design to date for this purpose. These pumps use rollers mounted on the ends of a rotating arm that displace blood within the compressed tubing, propelling it forward with faint pulsatility. Both the rate and direction of rotation and the occlusion pressure can be adjusted to adjust the flow. In contrast, centrifugal pumps use a magnetically driven impeller design that create a pressure differential between the inlet and the outlet portion of the pump housing propelling blood. Less traumatic to the blood elements, the centrifugal pump has largely replaced the roller pump for central circulation at most large centers.

The blood is actively pumped through a series of components before it is returned to the body. A heat exchange system is used to first lower the blood and body temperature during heart surgery. Controlled hypothermia reduces the body's energy demands, safely allowing reduced or even complete cessation of blood flow, which is necessary during certain times in the course of an operation. Reversing the process later rewarms the blood and body.

The pump oxygenator is by far the most sophisticated component of the circuit. As discussed, the membrane oxygenator is the predominant design in use today. Membrane oxygenators employ a microporous membrane that facilitates gas exchange in the circulating blood. Pores less than 1 μm prevent the leakage of serum across the membrane yet allow efficient gas exchange. An integrated gas mixer or blender enables the perfusionist to adjust the oxygen and carbon dioxide content in the blood and thus regulate the acid–base balance. In addition, an inhalational gas vaporizer provides anesthetic to the patient for the period during CPB. In line oxygen saturation monitors supplemented with frequent blood gas analysis ensure physiologic requirements are met during CPB.

Finally, an in line filter (typically 40 μm) prevents air and thromboembolic particles from returning to the arterial circulation. In parallel with this circuit, several additional roller pumps are used to provide suction that returns shed blood form the operative field. These are critical in open-heart procedures where a dry operative field is imperative for good visualization. An additional pump is often needed to deliver a cardioplegia solution, a mixture of blood and hyperkalemic solution used to stop the heart. Various monitors, hemofiltration units, blood sampling manifolds, gauges, and safety valves are usually present.

The main arterio-venous circuit is generally primed with a balanced electrolyte solution, which obviates the need for the blood prime used in the earlier, higher volume circuits. Other prime constituents include buffers such as sodium bicarbonate, oncotics such as albumin and mannitol, and anticoagulation in the form of heparin. Some protocols also call for the addition of an antifibrinolytic such as aminocaproic acid or aprotinin to the prime.

The overall construction of the CPB circuit is highly variable among institutions depending on the preferences of the surgeons and perfusionists. Although the principals of perfusion are universal, individual and group practices

are not. Perfusionists are a highly trained group of specialists dedicated to the safe operation of the heart–lung machine. Perfusion education and training has evolved concurrently with improvements in the bypass circuit, and accredited perfusion programs are included in many allied health curriculums at select universities. Perfusionists are required to pass a board examination, and many government organizations are now enacting licensure legislation. These persons must be able to adapt to the ever-changing technology present in cardiac surgery as well as to protocols and circuitry that vary among surgeons and institutions.

FUTURE DIRECTIONS

There are many commercial designs currently available that incorporate several of these components into a single disposable unit. Some units, for example, combine a reservoir with a membrane oxygenator and a heat exchanger. Continued innovation has resulted in more efficient, compact designs that limit the blood contacting surface area and decrease the need for large priming volumes and thus the need for patient transfusions. Improving biocompatibility of materials has lessened the inflammatory response that is associated with extracorporeal (outside the body) circulation. Sophisticated cannula designs are less traumatic to the vessels and have improved flow dynamics, causing less sheer stress on circulating red blood cells. New safety mechanisms such as alarm systems, pop-off valves, and automatic air evacuation devices will further increase the efficacy of these lifesaving machines.

The field of cardiac surgery is similarly evolving. Minimally invasive approaches to many heart operations are developing driven by patient demand as well as efforts to reduce postoperative hospital stay and patient morbidity. Recent debate has also focused on the possible adverse neurologic sequelae associated with the use of CPB. A growing number of coronary bypass procedures are now performed without the use of a heart–lung machine. Some centers have demonstrated success with this approach, decreasing the need for postoperative blood transfusions and end-organ dysfunction. To date, however, there is no conclusive evidence that avoidance of the heart lung machine results in improved neurologic outcomes or patient survival.

The creation of the heart–lung machine was the rate-limiting step to the development of the field of cardiac surgery. The ability to stop the heart while preserving flow the remainder of the heart has given surgeons the ability to repair defects in the heart and great vessels in patients of all ages. The pioneers in this field demonstrated remarkable courage and conviction in persevering in the face of overwhelming odds. Their collaborative efforts during one of the most prolific periods in the history of medicine have had a remarkable impact on human health. The future of cardiac surgery is largely dependent on continued advances in perfusion techniques and in the components of the heart–lung machine. In this endeavor, industry plays a pivotal role in developing and manufacturing improved products tailored to meet the needs of an everchanging field. Ultimately, evidence-based outcomes research will help ensure these innovations result in improved outcome for patients.

BIBLIOGRAPHY

Cited References

1. Lewis FJ, Taufic M. Closure of atrial septal defects with the aid of hypothermia; experimental accomplishments and the report of one successful case. Surgery 1953;33:52–59.
2. Gibbon JH Jr. Artificial maintainance of circulation during experimental occlusion of the pulmonary artery. Arch Surg 1937;34:1105–1131.
3. Gibbon JH Jr., et al. the closure of interventricular septal defects in dogs during open cardiotomy with the maintainance of the cardiorespiratory functions by a pump oxygenator. J Thor Surg 1954;28:235–240.
4. Dennis C, et al. Development of a pump oxygenator to replace the heart and lungs: an apparatus applicable to human patients and application to one case. Ann Surg 1951;134: 709–721.
5. Kirklin JW, et al. Intracardiac surgery with the aid of a mechanical pump oxygenator system (Gibbon type): Report of eight cases. Proc Mayo Clin 1955;30:201–206.
6. Kay EB, et al. Certain clinical aspects in the use of the pump oxygentor. JAMA 1956;162:639–641.
7. Cross FS, et al. Description and evaluation of a rotating disc type reservoir oxygenator. Surg Forum 1956;7:274–278.
8. Dennis C, Karleson KE, Nelson WP, Eddy FD, Sanderson D. Pump-oxygenator to supplant the heart and lunf for brief periods. Surgery 1951;29:697–713.
9. Lillehei CW, et al. Direct vision intracardiac surgery in man using a simple, disposable artificial oxygenator. Dis Chest 1956;29:1–8.
10. Gott VL, et al. A self-contained, disposable oxygenator of plastic sheet for intracardiac surgery. Thorax 1957;12:1–9.
11. Kolff WJ, et al. Disposable membrane oxygenator (heart-lung machine) and its use in experimental surgery. Clev Clin Q 1956;23:69–97.
12. Lande AJ, et al. A new membrane oxygenator-dialyzer. Surg Clin North Am 1967;47:1461–1470.
13. Gravlee GP, Davis RF, Kurusz M, Utley JR, editors. Cardiopulmonary Bypass; Principles and Practice, Philadelphia: Lippincott Williams and Wilkins; 2000.

Reading List

Edmunds LH. Cardiopulmonary bypass after 50 years. New Engl J Med 2004;351:1603–1606.
Iwahashi H, Yuri K, Nosé Y. Development of the oxygenator: Past, present and future. J Artif Organs 2004;7:111–120.

See also BLOOD GAS MEASUREMENTS; PULMONARY PHYSIOLOGY; RESPIRATORY MECHANICS AND GAS EXCHANGE.

HEAT AND COLD, THERAPEUTIC

MARY DYSON
United Kingdom

INTRODUCTION

Therapeutic levels of heating accelerate the resolution of inflammation, relieve pain, and promote tissue repair and

regeneration. Therapeutic levels of cooling reduce inflammation, relieve pain, and can reduce the damage caused by some injurious agents. To use therapeutic heating (thermotherapy) and therapeutic cooling (cryotherapy) effectively, the clinician should know:

- The physiological effects of heat and cold
- What devices are available to produce temperature changes
  - How they work
  - When and when not to use them
  - Their advantages and disadvantages.

This information assists the clinician in selecting the most suitable therapy and device for each patient.

The aim of this article is to provide this information, following a brief description of the history of these therapies, (this being included because history informs current and future usage).

HISTORY OF THERMOTHERAPY AND CRYOTHERAPY

Heat and cold have been used to treat diseases, aid healing, and reduce pain for many millennia. Exposure to the warmth of sunlight and thermal mineral springs continues to be used to this day. The ancient Greeks and Romans also advocated heated sand and oils for the treatment of injuries (1). Heated aromatic oils were rubbed on the body before massage, a form of therapy still in use today. In contrast, lesions accompanied by burning sensations (e.g., abscesses) were commonly treated by the application of cooling substances such as cold water and ice packs, another form of therapy still in use today. Fever was treated with cold drinks and baths; eating snow alleviated heartburn.

Thermotherapy

In the seventeenth century, one treatment for arthritis and obesity was to bury patients up to their necks in the sun-warmed sand of beaches. Warm poultices made from vegetable products such as leaves and cereals were used to treat musculoskeletal and dermal ailments. Molten wax was used to treat bruises around the eyes and infected eyelids, a technique not recommended because of the danger of getting hot wax into the orbit and damaging the eye.

In the eighteenth century, hot air produced by ovens and furnaces was used to induce perspiration and improve the circulation (2).

In the nineteenth century, Guynot found that wounds healed faster when the ambient temperature was 30 °C than when the ambient temperature was lower than this. After the invention of electric light bulbs, these were used to produce light baths and heating cabinets that were used to treat neuralgia, arthritis, and other conditions. Different wavelengths of light were also produced, initially by the use of prisms and passing light through media of different colors, and most recently by the use of lasers. There is evidence that different wavelengths of the electromagnetic spectrum including light and infrared radiation have different biomedical effects (3). When light is used to produce heat, both phototherapeutic and thermotherapeutic effects are produced that may reinforce each other.

In the late nineteenth and early twentieth century, the availability of electricity as a power supply led to the development of a host of novel thermotherapeutic devices including whirlpool baths, paraffin baths, and diathermy machines. In diathermy, high frequency electric currents are used to heat deeply located muscles. This effect was discovered in the 1890s when d'Arsonval passed a 1 A current at high frequency through himself and an assistant. He experienced a sensation of warmth (4). d'Arsonval worked on the medical application of high-frequency currents throughout the 1890s, work that led to the design of a prototype medical device by Nagelschmidt in 1906 and the coining of the term "diathermy" for what had previously been known as "darsonvalization." The first diathermy machines were long wave, and in the second decade of the twentieth century, these were used to treat a wide range of diseases, including arthritis, poliomyelitis, and unspecified pelvic disorders. In 1928, short-wave diathermy was invented, superceding long-wave diathermy in Europe after the Second World War.

Therapeutic ultrasound also became popular in the post-war period, initially as a method of heating tissues deep within the body to relieve pain and assist in tissue repair. It was demonstrated experimentally that ultrasonic heating was accompanied by primarily nonthermal events such as micromassage, acoustic streaming, and stable cavitation. These events, which also occur at intensities less than are required to produce physiologically significant heating of soft tissues, produce cellular events that accelerate the healing process (3,5). Therapeutic ultrasound is generally used at frequencies in the 0.75–3 MHz range. The higher the frequency, the shorter the wavelength and the lower the intensity needed to produce physiologically significant heating of soft tissues. However, higher frequencies, because they are more readily absorbed than lower frequencies, are less penetrative and are therefore less suitable for heating deep soft tissues. Since the 1990s, low kilohertz ultrasound, also known as long-wave ultrasound, has been used to initiate and stimulate the healing of chronic wounds (6). Although long-wave ultrasound is not primarily a heating modality (7), as with all therapies in which energy is absorbed by the body, it is inevitably transduced into heat, although in this instance insufficient to be clinically significant (8).

Cryotherapy

Until the invention in 1755 of artificial snow, which was made by placing a container of water over nitrous ether as the latter vaporized, only cold water and naturally occurring ice and snow were available as means of cooling the body. In the nineteenth century, ice packs were used over the abdomen to reduce the pain of appendicitis and over the thorax to reduce the pain of angina. Ice was also used as a local anesthetic. In 1850, evaporative cooling methods were introduced; for example, ether applied to the warm forehead had a cooling effect as it evaporated. Since the early days of physical therapy, the local application of ice has

been used to treat acute musculoskeletal injuries; usually a combination of rest, ice, compression, and elevation (RICE) are recommended.

Contrast Bath Hydrotherapy

Hydrotherapy is the use of water as a therapeutic agent. Hot or cold water, usually administered externally, has been used for many centuries to treat a wide range of conditions, including stress, pain, and infection. Submersion in hot water is soothing and relaxing, whereas cold water may be anti-inflammatory and limit the extent of tissue damage. In contrast bath hydrotherapy, the patient is exposed to hot and cold water alternately to stimulate the circulation. The blood vessels dilate in the heat and constrict in the cold. Pain is also reduced. Contrast baths have been used to treat overuse injuries such as carpal tunnel syndrome and tendonitis of the hand and forearm (http://www.ithaca.edu/faculty/nquarrie/contrast.html). Although there has been little research on the efficacy of contrast baths, they remain in use and may be of clinical value (9).

THE PHYSIOLOGICAL EFFECTS OF HEAT AND COLD

When healthy we keep a fairly constant body temperature by means of a very efficient thermoregulatory system. We are homoeothermic organisms. Homoeothermic is a pattern of temperature regulation in which cyclic variation of the deep body (core) temperature is maintained within arbitrary limits of $\pm 2\,^{\circ}C$ despite much larger variations in ambient temperature. In health and at rest, our core temperature can be maintained by the thermoregulatory system within $\pm 0.3\,^{\circ}C$ of $37\,^{\circ}C$ in accordance with the body's intrinsic diurnal temperature cycle. Superimposed on the diurnal temperature cycles are monthly and seasonal temperature cycles. Hyperthermia is a core temperature greater than $39\,^{\circ}C$. Hypothermia is a core temperature less than $35\,^{\circ}C$.

The physiological effects of heat and cold are generally independent of the agent producing the temperature change, although in some cases, for example, ultrasound therapy, primarily nonthermally induced events accompany those induced thermally (3).

Physiological Effects of Heat

Heating of tissues occurs when these tissues absorb energy. The different effects of heating are due to many factors, including the following (10):

- The volume of tissue absorbing the energy
- The composition of the absorbing tissue
- The capacity of the tissue to dissipate heat, a factor largely dependent on its blood supply
- The rate of temperature rise
- The amount by which the temperature is raised

Cell Metabolism. Cell metabolism increases by about 13% for each $1\,^{\circ}C$ increase in temperature up to the temperature at which the proteins of the cell, many of which are vital enzymes, coagulate. Enzyme activity first increases as the temperature increases, then peaks, then declines as the enzymes denature, and is finally abolished when the temperature reaches about $45\,^{\circ}C$ when heat kills the cells. Only temperature increases less than those producing enzyme denaturation are therapeutic. The cells' membranes are particularly sensitive to heat, which increases the fluidity of their lipoproteinaceous components, producing changes in permeability (11). At sublethal temperatures, heat-shock proteins, which give some protection to cells reexposed to heat, accumulate in the cells.

Abnormal and normal cells are affected differently by mild hyperthermia ($\sim 40\,^{\circ}C$). Synthesis of deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and proteins can all be inhibited by mild hyperthermia in abnormal cells, irreversibly damaging their membranes and killing the cells; this does not occur in normal cells subjected to mild hyperthermia (12). The technique can therefore be used to kill abnormal cells selectively.

Collagen. The extensibility of collagen is increased by thermotherapy. Raising the temperature of a highly collagenous structure such as a tendon increases the extent of elongation produced by stretch of a given intensity (13). Joint capsules are also highly collagenous, so thermotherapy decreases the resistance of a joint to movement. Cryotherapy, however, increases this resistance. Changes in resistance to movement are also due to changes in the viscosity of the synovial fluid.

Laboratory experiments suggest that heat should be applied to tendons with caution because *in vitro* exposure of excised tendons to temperatures in the range of $41–45\,^{\circ}C$ is accompanied by a reduction in their tensile strength (13). It is unlikely, however, that the stresses produced by passive stretch during physical therapy and normal exercise will reach the levels at which rupture occurs, although overvigorous exercise could be damaging, particularly in those who ignore protective pain.

Blood Flow. Thermotherapy causes blood vessels to widen (vasodilatation), increasing local blood flow and causing the skin to redden (erythema). Heat-induced vasodiltation has several causes, including the following:

- The direct effect of heat on the smooth muscle cells of the arterioles and venules.
- If there is local damage, then the damage-induced release of vasodilators such as bradykinin will cause further vasodiltation.

Bradykinin, histamine, and other chemical mediators released in response to injury and to heating increase capillary and venule permeability that, together with an increase in capillary hydrostatic pressure, can produce local swelling of the tissues (edema). The application of local heat immediately after injury should therefore be avoided (14).

Heating also induces changes in blood flow in subcutaneous tissues and organs. These changes depend on the amount of heating. First blood flow increases in these structures, but then it decreases if heating is sufficient for the core temperature to rise, as blood is diverted to the

skin where heat is lost from it to the external environment as part of the thermoregulatory process.

Different local heating techniques can have different effects on blood flow, due to differences in their depth of penetration. Methods producing superficial heating include infrared radiation and contact with a heated material; those producing deep heating include short-wave and microwave diathermy. Infrared (IR) radiation increases cutaneous blood flow (15) but not in the underlying skeletal muscle (16). In contrast, diathermy increases blood flow and temperature in both skin and skeletal muscle in humans (17), hyperemia being sustained for at least 20 min after the cessation of treatment, probably because of an increase in the metabolic rate of the heated tissues.

Neurological Effects. Therapeutic heat produces changes in muscle tone and pain relief.

Muscle Tone. Increased muscle tone can sometimes be reduced by the application of either superficial or deep heat. In 1990, Lehmann and de Lateur (18) showed that heating skin and skeletal muscle of the neck relieved muscle spasm secondary to underlying pathology. The Ia afferents of muscle spindles increase their firing rate on receipt of heat within the therapeutic range, as do tendon organs (19). Most secondary afferents decrease their firing rate (20). Collectively these neurological changes reduce muscle tone locally.

Pain Relief. People in pain generally consider heat to be beneficial, even on the intense pain experienced by patients with cancer (21).

Therapeutic heat produces vasodiltation and may therefore relieve pain related to ischemia. Pain related to muscle spasm may also be relieved by therapeutic heat; this reduces muscle spasm secondary to underlying pathology (18). Heat may also act as a counterirritant and relieve pain via the pain gate mechanism (22), in that the thermal sensations take precedence in the central nervous system over nociceptive sensations.

Increasing the temperature within the therapeutic range increases the velocity of nerve conduction. An increase in sensory conduction increases the secretion of pain-relieving endorphins.

Tissue Injury and Repair. The initial response of vascularised tissues to injury is acute inflammation. This is characterized by:

- Heat
- Redness
- Swelling (edema)
- Pain
- Loss of function

During it a host of chemical mediators and growth factors are secreted; these collectively limit the extent of tissue damage and lead to healing. The application of therapeutic levels of heat can accelerate the resolution of acute inflammation leading to faster healing. Although uncomfortable, acute inflammation is not a disease but a vital component of tissue repair and regeneration. The pain associated with it is generally of a short duration and is of survival value in eliciting protective actions.

The management of acute trauma by thermotherapy and cryotherapy is based on a pathophysiological model often referred to as the secondary injury model (23). According to this model, secondary injury is the damage that occurs as a consequence of the primary injury in previously unaffected cells. The mechanisms initially hypothesized as producing this damage were classified as being either enzymatic or hypoxic (24). Since then, knowledge of the mechanisms involved in cell death from trauma has increased dramatically and a third mechanism is now postulated, namely the delayed death of primary injured cells. A review and update by Merrick in 2002 (25) attempts to reconcile the secondary injury model with current knowledge of pathophysiology. Secondary hypoxic injury has been reclassified as secondary ischemic injury, and specific mechanisms for ischemic injury have been identified. In addition to changes of vascular origin, there is now evidence that apparently secondary injury may be due, in part, to the delayed death of cells subjected, for example, to mitochondrial damage during the primary trauma. A better understanding of secondary injury should inform methods of treatment, some of which, although traditional, may need to be altered to improve their effectiveness. For example, the rationale that short-term cryotherapy of acute injuries was effective because it limited edema through vasoconstriction has been replaced by the currently accepted theory that it also retards secondary injury, regardless of the cellular mechanisms by which this occurs, be they lysosomal mechanisms, protein denaturation mechanisms, membrane permeability mechanisms, mitochondrial mechanisms, and/or apoptotic mechanisms (25). By reducing metabolic demand, the application of cryotherapy as soon as possible after an acute injury should reduce the speed and, possibly the extent, of secondary injury. There is evidence that continuous cryotherapy for 5 hours after a crush injury inhibited the loss of mitochondrial oxidative function that follows such injuries (25). The effect of continuous and intermittent cryotherapy for other durations on this and other pathophysiological events remains to be examined. This must be done if clinical treatments are to be improved.

Systemic and local therapeutic heating can reduce postoperative wound infection (26). The use of a warm-up dressing (Augustine Medical), which radiates heat and provides a moist wound environment, eradicated methicillin-resistant *Staphylococcus aureus* (MRSA) infection from pressure ulcers within 2 weeks (27). It should be appreciated that patients with MRSA-infected pressure ulcers, also known as pressure sores and bed sores, have increased morbidity and mortality; these ulcers can kill. Warming the tissues induces vasodiltation, giving rise to the high oxygen tension required for the production of oxygen-free radicals; these are an important part of the body's defense against bacterial infection (28) and initiate collagen synthesis and re-epithelialization. Vasodiltation also aids healing by increasing the efficiency with which the cells and growth factors needed for this are transported to the injured region and the efficiency with which waste products are removed from it.

Warming has been shown to increase the proliferation of fibroblasts (29), the cells that synthesize collagen, and of endothelial cells (30), the cells lining blood vessels, *in vitro*. If this also occurs *in vivo*, then warming will accelerate the formation of granulation tissue. More research is required to confirm whether therapeutic heating reduces the risk of wound infection and accelerates healing in a clinically significant fashion. The data gathered to date suggest that it may. If so, then "there is a real prospect of reducing complications, improving patient outcomes, shortening hospital stays and minimizing costs" (31).

Physiological Effects of Cold

The physiological effects of therapeutic cold (9) depend on a range of factors, including

- The volume of tissue to which cooling is applied
- The composition of the tissues cooled
- The capacity of the tissues to moderate the effects of cooling
- The rate of temperature fall
- The amount by which the temperature of the tissues is lowered

Cell Metabolism. The vital chemical processes occurring in cells generally slow as the temperature is lowered. These processes are catalyzed by enzymes, many of which are associated with the cells' membranes. Cell viability relies on passive and active membrane transport systems, the latter involving ionic pumps activated by enzymes. These transport systems are necessary to maintain the intracellular ionic composition required for cell viability. Below a threshold temperature, the pumps fail, and consequently, the membranes lose their selective permeability; the intracellular concentration of Na^+ and Ca^{2+} increases whereas that of K^+ decreases. Between normal body temperature and this threshold, cooling is therapeutic. The application of therapeutic levels of cold can reduce cell degeneration and therefore limit the extent of tissue damage (31). The induction of mild hypothermia in the brain of a baby starved of oxygen at birth can interrupt the cascade of chemical processes that cause the neurons of the brain to die after being deprived of oxygen (32), reducing disability.

Collagen. Collagen becomes stiffer when cooled. People with rheumatoid arthritis generally experience a loss of mobility of their affected joints at low temperatures, due in part to increased stiffness of the collagen of their joint capsules (9).

Blood Flow. Lowering the skin temperature is detected by dermal thermoreceptors that initiate an autonomic reflex narrowing of the blood vessels of the skin (vasoconstriction). This results in reduction of the flow of blood to the dermis. Cold also has a direct constrictor effect on the smooth muscle of the blood vessels, and the arteriovenous anastomoses that shunt blood to the skin close. The resulting reduction in the flow of blood to the dermis diminishes heat loss through it. Countercurrent heat exchange between adjacent arteries and veins reduces heat loss. These processes collectively reduce the rate at which the core temperature of the body falls.

Dermal vasoconstriction induced by lowering the temperature of the skin to approximately 10 °C is followed after a few minutes by cold-induced vasodiltation (CIVD) followed by cycles of alternating vasoconstriction and vasodiltation, resulting in alternating decrease and increase on dermal blood flow. Originally thought to be either a local neurogenic axon reflex or due to the local release of vasodilator materials into the tissues, CIVD is now believed to be due to paralysis of vascular smooth muscle contraction in direct response to cold (33). This reaction may provide some protection to the skin from damage caused by prolonged cooling and ischemia. However, in CIVD of the skin, the erythema produced is a brighter red than in erythema produced by heating because at low temperatures, the bright red oxyhemoglobin dissociates less readily than at higher temperatures. Therefore, although in CIVD the skin receives oxygen-rich blood, it is still starved of oxygen, suggesting that cryotherapy may not aid wound healing (9).

In contrast to skin, the blood flow in the skeletal muscles is determined more by local muscle metabolic rate than by temperature changes at the skin because muscles are insulated from these by subcutaneous fat.

Neurological Effects. Cold can be used therapeutically to affect muscle tone and pain.

Muscle Tone. Although cooling generally decreases muscle tone, this can be preceded by a temporary increase in tone (34), possibly related to tactile stimulation accompanying the application of cryotherapy by means of ice massage. Decrease in muscle tone in response to cryotherapy is likely to be due to a decrease in muscle spindle sensitivity as the temperature falls, together with slowing of conduction in motor nerves and skeletal muscle fibers.

Pain Relief. The immediate response of the skin to cold is a stimulation of the sensations of cold and pain. However, if the cold is sufficiently intense, nerve conduction is inhibited, causing both sensations to be suppressed. Less intense cold can be used as a counterirritant, its effects being explicably by the pain gate theory of Wall and Melzack (35). Enkephalins and endorphins may also be involved (36).

Tissue Injury and Repair. Cryotherapy has beneficial effects on the acute inflammatory phase of healing in that it reduces bleeding, decreases edema at the injury site, gives pain relief, and reduces local muscle spasm, as described above. Once these events have occurred, it should be replaced by thermotherapy because, as previously described, this accelerates the resolution of acute inflammation leading to faster healing.

DEVICES AVAILABLE FOR PRODUCING TEMPERATURE CHANGE

Thermotherapy Devices

These can be classified into those producing superficial heating and those producing deep heating. All should be

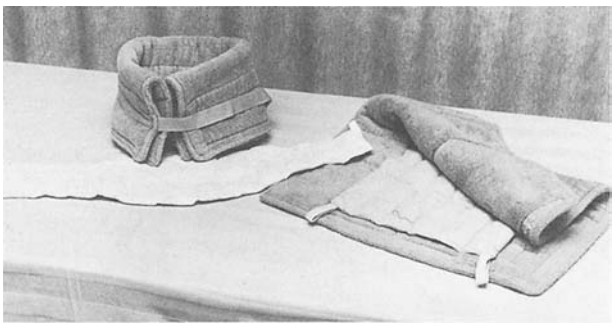


Figure 1. Hydrocollator pack for moist heating by conduction. Each pack consists of silica gel covered by cotton fabric. After submersion in hot water, the pack is applied to the area to be heated over layers of terry cloth.

used with caution in patients unable to detect heat-induced pain or over tissues with a poor blood supply because these cannot dissipate heat efficiently.

Superficial Heating. This is provided by devices that transfer energy to the body by conduction, convection, or radiation (37).

Conduction. Devices that heat by conduction include hydrocollator packs, hot water bottles, electric heating pads, and baths of heated paraffin wax.

Hydrocollator packs Figs. 1 and 2 consist of cotton fabric bags containing a silica gel paste that absorbs water equal to 10 times its weight. They are placed in thermostatically controlled tanks of hot water, and after they have absorbed this, one or more packs are placed on toweling-covered skin, into which heat is conducted until the pack has cooled, usually or 20–30 min, depending on the ambient temperature. Heat is also conducted into the subcutaneous tissues to a depth of about 5 mm. The transportation of blood warmed in the superficial tissues to deeper tissues may result in the latter also becoming warmer. Hot water bottles and thermostatically controlled electric heating pads act in a similar fashion.

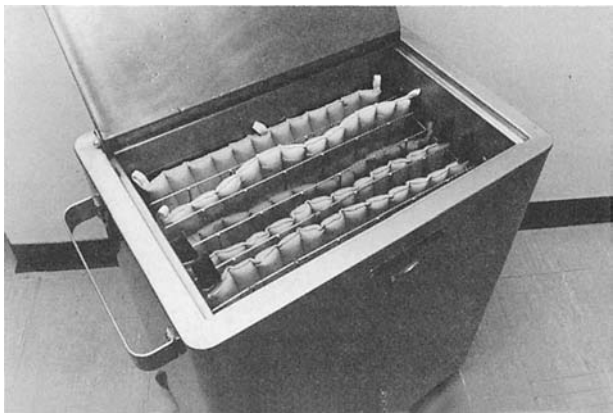


Figure 2. Hydrocollator packs being heated in a thermostatically controlled tank containing hot water.

When and When Not to Use These Devices. They are useful in the relief of pain and muscle spasm. Abdominal cramps can be treated effectively through the induction of smooth muscle relaxation. Superficial thrombophlebitis and localized skin infections such as boils or furuncles may also be treated by heat conduction with these devices. It has been advised that these and other primarily heating devices should not be used over injured tissue within 36 h of the injury. Nor should they be used where there is decreased circulation, decreased sensation, deep vein thrombophlebitis, impaired cognitive function, malignant tumors, a tendency toward hemorrhage or swelling, an allergic rash, an open cut, skin infection, or a skin graft (http://www.colonialpt.com/CPTFree.pdf). A search of English–language textbook and peer–reviewed sources and computerized databases from January 1992 to July 2002 (38) has revealed "generally good agreement among contraindication sources for superficial heating devices"; however, agreement ranged from 11% to 95% and was lower for pregnancy, metal implants, edema, skin integrity, and cognitive/communicative concerns.

Advantages and Disadvantages. The moist heat produced by the hydrocollator packs has a greater sedative effect than the dry heat produced by the other conduction devices listed above. The hydrocollator packs are, however, heavy because of the amount of water they absorb and should not be applied to very frail patients, for example those with advanced osteoporosis. Hot water bottles occasionally leak, producing scalding and should therefore be inspected after filling and before each use. Electric pads should be thermostatically controlled and used with an automatic timer to reduce the risk of burning the skin.

Paraffin Wax Baths. Molten paraffin wax is mixed with liquid paraffin wax and kept molten in a thermostatically controlled bath (Fig. 3) at between 51.7 and 59.9 °C. Typically used for heating hands or feet, these are either dipped into the bath several times so that a thick layer of wax forms on them, or immersed in the bath for 20–30 min. After dipping, the hand or foot is wrapped in a cotton towel to retain the heat; after 20–30 min, the towel and wax are peeled off.

When and When Not to Use These Devices. Paraffin wax baths are suitable for treating small areas with irregular surfaces. The heat provides temporary relief of joint stiffness and pain in patients with chronic osteoarthritis and rheumatoid arthritis. By increasing the elasticity of collagen, it helps to increase soft-tissue mobilization in the early stages of Dupuytren's contracture and after traumatic hand and foot injuries provided that the lesions have closed. Paraffin wax baths should not be used to open or infected wounds.

Advantages and Disadvantages. Molten paraffin wax is suitable for applying heat to small irregular surfaces such as those of the hands and feet. The baths can be used several times without changing the wax, and several patients can be treated at the same time if the baths are sufficiently large.

Figure 3. Paraffin wax bath used for heating hands and feet by the dip method.

A disadvantage is that the wax collects debris from the surface of the skin and other particles that settle at the bottom of the bath and are difficult to remove without emptying the bath. Treatments must be supervised and the temperature of the wax monitored to avoid the risk of burning.

Convection. This involves immersion of either the whole of the body or part of the body in heated water. Hubbard tanks (Fig. 4), hydrotherapy pools (Fig. 5), akin to heated swimming pools, and baths can be used for either total or partial immersion. The Hubbard tanks can be fitted with agitation devices. Some baths incorporate low-frequency ultrasound transducers. It is recommended that the water temperature should not exceed 40.6 °C when used for total immersion and 46.1 °C when used for partial immersion (1). Treatments typically last for 20–30 min.

Figure 4. Hubbard tank used for partial or total body submersion.

Hydrotherapy is useful for treating multiple arthritic joints simultaneously and for treating extensive soft-tissue injuries. Although total immersion can raise the core body temperature, the main use of hydrotherapy is for the relief of pain, muscle spasm, and joint stiffness. Exercise regimes can be incorporated into hydrotherapy treatment to take advantage of the increase in joint movement made possible by the decrease in joint stiffness and the pain associated with movement. The addition of an agitation device to a Hubbard tank allows it to be used for the gentle debridement of burns, pressure ulcers, and other skin conditions, provided that thorough cleaning and disinfection of the tank are carried out between treatments. Baths incorporating low frequency (kHz) ultrasound transducers have the additional advantage that they can stimulate the repair process as well as assist in debridement.

The greater the surface of the body that is immersed, the greater the number of joints and the greater the area of skin and subcutaneous tissue that can be heated. Immersion also provides buoyancy, helping the patient to exercise more easily.

Disadvantages are that

- Each patient needs one-to-one supervision to ensure that the mouth and nose are kept free from water.
- The equipment is relatively expensive and is time-consuming to maintain.

Radiation. Heat transmission by radiation is generally provided by infrared lamps (Fig. 6). These are photon producers, emitting wavelengths from the red end of the visible part of the electromagnetic spectrum, together with IR. The production of visible light provides a visual indication when the lamps are active. The IR component provides heating.

Infrared emitters used for heating are either nonlaser or laser devices, the former being the most commonly used.

Nonlaser.

- Luminous
- Nonluminous

Luminous emitters are effectively light bulbs, each mounted in the center of a parabolic reflector that projects a bright beam like a floodlight. About 70% of the energy emitted consists of IR rays in the wavelength range of 700–4000 nm.

Luminous emitters can be classified into those with large, high-wattage (600–1500 W) bulbs, and those with smaller, lower wattage (250–500 W) bulbs.

1. The large luminous emitters are used to treat large regions of the body such as the lumbar spine. They

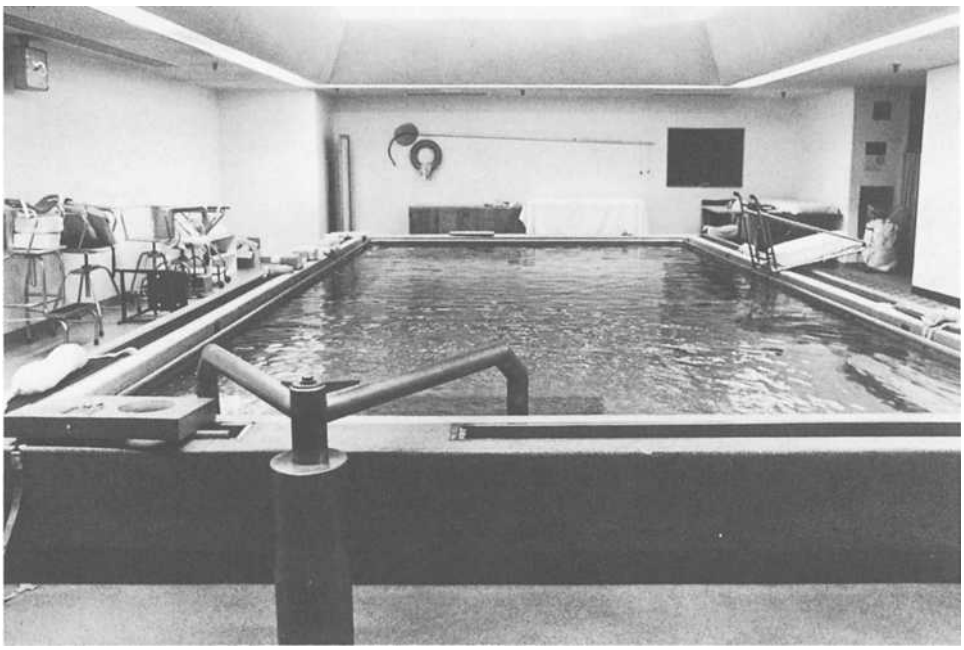


Figure 5. Hydrotherapy pool used for partial or total body submersion.

are positioned about 0.6 m from the patient and used for 20 min.

2. The small luminous emitters are used to treat smaller regions of the body such as the hands and feet. They are positioned closer to the patient, about 0.45–0.5 m away, and they are typically used for 20 min.

Figure 6. Luminous generator infrared lamp used for superficial heating by radiation. The height and angle of the lamp can be adjusted.

The former produce a greater sensation of warmth than the latter. The skin becomes temporarily erythematous, resembling blushing. If the heat is too great because, for example, the lamp is too close to the skin, *erythema ab igne*, a mottled appearance indicative of damage, may be produced.

Nonluminous emitters are usually cylinders of either metal or an insulating material, the radiation source, around which a resistance wire, the heat source, is coiled. They emit longer wavelengths of IR that penetrate deeper than the shorter IR wavelengths of the luminous emitters. About 90% of the emission of the nonluminous IR device is in the wavelength range of 3500–4000 nm.

When and When Not to Use These Devices. IR radiation when used as a heating modality can alleviate pain and muscle spasm. It also accelerates the resolution of acute inflammation. It should not be applied to skin that is photosensitive. It should only be applied to skin free from creams that, if heated, could burn the skin. Goggles should be worn to protect the eyes from radiation damage to either the lens or retina. Patients with abnormally low blood pressure should not receive any form of treatment that increases the temperature of large areas of skin and subcutaneous tissue because of the redistribution of blood to these areas and away from vital organs such as the brain, heart, and lungs.

Advantages and Disadvantages. IR heating is suitable for use in the home as well as in clinics and hospitals. Because the devices described in this section are noncontact, they can be used to treat open and infected wounds, but only for short periods because they cause fluid evaporation and wounds need to be kept moist if they are to heal efficiently.

The main disadvantage is that of burning if the IR heater is placed too close to the patient. Only mechanically stable devices should be used to ensure that they do not fall on the patient.

Laser. Lasers emitting IR energy have been used to treat injured tissue in Japan and Europe for over 25 years (39). The term "LASER" is an acronym for light amplification by the stimulated emission of radiation. Laser devices are a very efficient means of delivering energy of specific wavelengths to injured tissue. Laser radiation is coherent; that is the waves are in phase, their peaks and troughs coinciding in time and space. Some semiconductor diodes produce coherent and others noncoherent IR radiation. Both can accelerate repair and relieve pain, as does red light (39). The technique of using low intensity lasers therapeutically is usually referred to as LILT, an acronym for low intensity laser therapy, or low level laser therapy (LLLT) to distinguish it from the surgical use of high intensity lasers. Surgical CO_2, ruby, and Nd-YAG surgical lasers can be used for thermal biostimulation provided that the amount of energy absorbed by the tissues is reduced sufficiently. Manufacturers achieve this either by defocusing and spreading the beam over an area large enough to reduce the power density below the threshold for burning or by scanning rapidly over the area to be treated with a narrow beam. The patient will then feel only a mild sensation of heat (39).

Nonsurgical lasers are typically used at power densities below those producing a sensation of heating; although their bioeffects are primarily nonthermal, absorption of light inevitably involves some heating. LILT can be used in either continuous or pulsed mode, the latter further reducing heating.

Single probes, each containing a single diode, are used to treat small areas of skin and small joints. Some are suitable for home use; their output is generally too low to produce the sensation of heat although the radiation they emit is transduced into heat after absorption. The effects of these devices on pain and tissue repair are essentially nonthermal. Clusters of diodes are used in clinics and hospitals to treat large areas of skin, skeletal muscle, and larger joints. These cluster probes generally include, in the interest of cost-effectiveness, nonlaser diodes. Treatment times vary with the type of probe and the area to be treated. Typically they range from about 3 to 10 min. They are calculated in terms of the time taken to deliver an effective amount of energy, generally 4 or more $J \cdot cm^{-2}$, joules being calculated by multiplying the power density of the probe (in $W \cdot cm^{-2}$) by the irradiation time in seconds. The larger cluster probes have sufficient output to produce a sensation of warmth. They also have essentially nonthermal effects that assist in the relief of pain and the stimulation of tissue repair (39).

When and When Not to Use These Devices. LILT devices are effective on acute and chronic soft-tissue injuries of patients of all ages. Most are designed for use in clinics and hospitals, but small, handheld devices are also available for home use as part of a first aid kit. The radiation they produce is absorbed by all living cells and transduced into chemical and thermal energy within these cells, which are activated by it. Temporary diltation of superficial blood vessels occurs, aiding oxygenation.

Contraindications for LILT have been described in some detail by Tuner and Hode (39). They include the following:

- As a matter of prudence, LILT should not be applied over the abdomen during pregnancy to ensure that no harm comes to the fetus although it can be applied to other parts of the body.
- In patients with epilepsy who may be sensitive to pulses of light in the 5–10 Hz range, it is advisable to use either unpulsed (continuous) LILT or much higher pulsing frequencies.
- Treatment over the thyroid gland is contraindicated because this gland may be sensitive to absorbed electromagnetic radiation in the red and IR parts of the spectrum.
- Patients with cancer or suspected cancer should only be treated by a cancer specialist.

Advantages and Disadvantages. LILT is easy to apply directly to the skin. The treatment is rapid and painless, the only sensations being of those of contact and, if the power is sufficiently high, of warmth. If used over an open wound, this should first be covered with a transparent dressing through which the radiation can pass unimpeded. The use of goggles is recommended as a precaution. The larger devices require clinical supervision, but some of the smaller predominantly nonthermal devices are suitable for home use.

Deep Heating. Deep heating devices produce heat within the tissues via electrostatic, electromagnetic, and acoustic fields. The term diathermy is used to describe the conversion or transduction of any form of energy into heat within the tissues. The devices used for deep heating include short-wave, microwave, and ultrasound generators. They are relatively expensive and should only be used by trained operators.

Short-Wave diathermy (SWD). SWD equipment Figs. 7 and 8 produces nonionizing radiation in the form of radio waves in the frequency range of 10–100 MHz. The most commonly used frequency is 27.12 MHz; this has a wavelength of 11 m.

SWD machines consist of:

- An oscillating circuit that produces the high frequency current
- A patient circuit connected to the oscillating circuit through which electrical energy is transferred to the patient
- A power supply

The patient's electrical impedance is a component of the patient circuit. Because the patient's electrical impedance is variable, the patient circuit must be tuned to be in resonance with the oscillating circuit to ensure maximum flow of current through the patient.

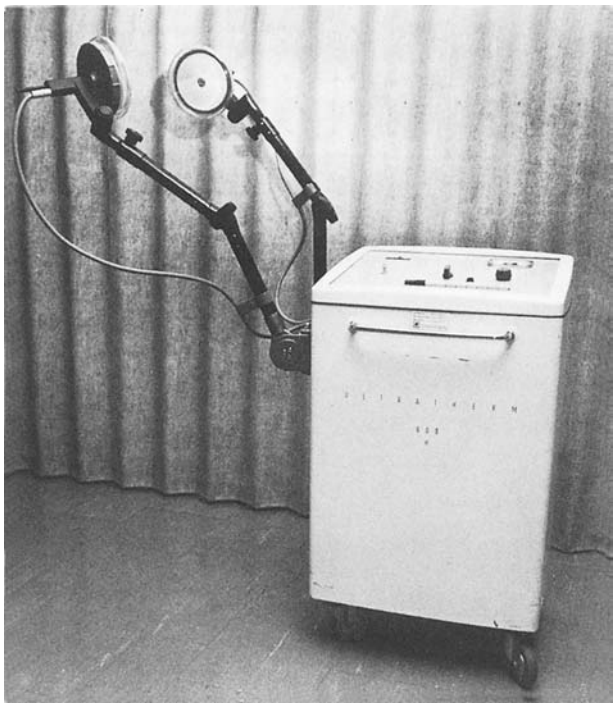


Figure 7. SWD equipment. The SW energy leaving the insulated disk-shaped applicators produces small eddy currents in the tissues, thus heating them.

The most commonly used means of application of SWD are as follows:

- The capacitor plate technique
- The inductive method

The capacitor plate technique entails the placement of two capacitor plates near to the part of the patient's body that is to be heated, with a spacer between each plate and the skin. The spacer allows the electric radiation to diverge just before entering, preventing thermal damage to the surface of the skin. The current density depends on the resistance of the tissues and on capacitance factors. Both subcutaneous fat and superficial skeletal muscle can be heated with this technique. The patient feels a sensation of warmth. Treatment is usually for $20\,\text{min} \cdot \text{day}^{-1}$.

In the inductive method, a high-frequency alternating current is applied to a coiled cable generally incorporated into an insulated drum that is placed close to the part of the body requiring treatment. Alternatively, but these days rarely, an insulated cable is wrapped around the limb to be treated. The passing of an electric current through the cable sets up a magnetic field producing small eddy currents within the tissues, increasing tissue temperature. The patient feels a sensation of warmth. Treatment is usually for $20\,\text{min} \cdot \text{day}^{-1}$.

Pulsed Shortwave. Some SWD devices allow the energy to be applied to the patient in short pulses. This form of application is termed pulsed short–wave diathermy (PSWD). The only physical difference between SWD and PSWD is that in the latter the electromagnetic field is interrupted at regular intervals. Pulsing reduces the amount of energy available for absorption by the tissues and therefore reduces the thermal load, allowing its nonthermal effects to be exploited without the risk of a potentially damaging thermal overload. It has been suggested that the applied energy produces ionic and molecular vibration affecting cellular metabolism (40).

With some PSWD machines, the therapist can vary the:

- Pulse repetition rate (PRR)
- Pulse duration (PD)
- Peak pulse power (PPP).

The mean power applied is the product of these three variables.

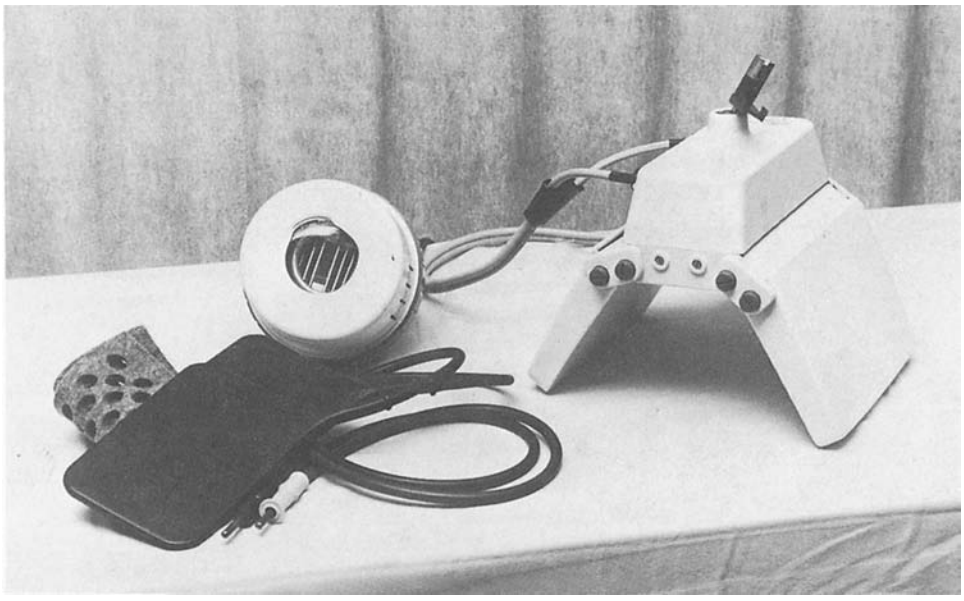


Figure 8. Applicator heads for administration of SWD to produce deep heating of tissues.

Figure 9. Microwave diathermy equipment. The microwave generator contains a magnetron that produces high frequency alternating current. This current is carried by a coaxial cable shown on the left to a transducer (upper left) containing an antenna and a reflector. The antenna converts the electric current into microwaves that are collected by the reflector. This focuses them and beams them into the body, where they are transduced into heat.

Tissues with good conductivity, such as muscle and blood, i.e., with a high proportion of ions, should absorb energy preferentially from the SWD field. However, there is still debate about which tissues are heated the most during SWD and PSWD treatments.

Microwave Diathermy. Microwaves are that part of the electromagnetic spectrum within the frequency range of 300 MHz–300 GHz and, therefore, with wavelengths between 1 m and 1 mm. Microwave diathermy, although deeper than superficial (surface) heating, is not as deep as capacitative short-wave or ultrasonic heating (40).

Microwave generators (Fig. 9) contain a magnetron, which produces a high frequency alternating current that is carried to a transducer by a coaxial cable. The transducer consists of an antenna and a reflector. The electric current is transduced into electromagnetic energy on passing through the antenna. The reflector focuses this energy and beams it into the tissues to be heated (1).

On entering the body the microwave energy is absorbed, reflected, refracted, or transmitted according to the physical properties of the tissues in its path. When microwaves are absorbed, their energy is transduced into heat. Tissues with a low water content (e.g., superficial adipose tissue) absorb little microwave energy, transmitting it into those with a high water content (e.g., skeletal muscle) that are very absorptive and therefore readily heated, warming adjacent tissues by conduction. As with SWD, there is no objective dosimetry, the intensity of treatment being judged by the patient's sensation of warmth (41).

When and When Not to Use These Devices. As with other forms of heating, SWD and microwave diathermy are used to relieve pain and muscle spasm. They are also used to stimulate repair.

According to the Medical Devices Agency of the United Kingdom, there are several groups of patients on whom these devices must not be used:

1. Those with implanted drug infusion pumps because the energy provided during therapy may affect the pump's electronic control mechanisms causing temporary alteration in drug delivery.
2. Women in the first trimester of pregnancy should not be treated with SWD because heating may be teratogenic.
3. Patients with metallic implants because metals are heated preferentially and may burn the surrounding tissue.
4. Patients fitted with active (i.e., powered) implants such as neurostimulators and cardiac pacemakers/defibrillators because there have been reports of tissue damage, including nerve damage adjacent to stimulation electrodes on implanted lead systems.
5. Patients with pacemakers are also unsuitable for SWD because the frequency of the short-wave may interfere with cardiac pacing.

Patients with rheumatoid arthritis have had their joint pain reduced and walking time increased after treatment with microwave diathermy, and it has been suggested that the heat produced in the joints may have potentiated the effects of concurrent anti-inflammatory medication (42).

In the interests of safety, microwave diathermy should be restricted to patients in whom skin pain and temperature sensation are normal. It should not be used near the eyes, sinuses, and moist open wounds, all of which could be heated excessively because of their high water contact. Nor should it be used on any of the groups of patients in whom SWD is contraindicated.

Advantages and Disadvantages. The devices do not need to be in direct contact with the body; however, the body part being treated must be immobile during treatment because movement interferes with the field, resulting in too little heating in some regions and too much in other. Its disadvantages include its potential for burning tissue exposed to it due, for example, to treatment over regions in contact with metal. As with SWD, the therapist must be careful to avoid being exposed to the radiation, bearing in mind that some of the microwaves will be reflected from the patient.

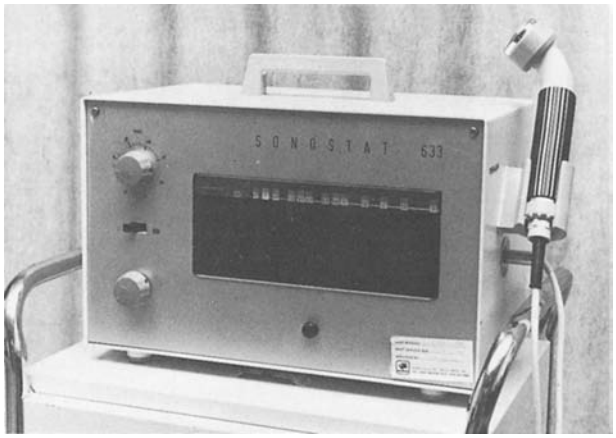


Figure 10. Portable ultrasound therapy equipment. The ultrasound transducer is housed in the head of the applicator shown on the right.

Ultrasound. Ultrasound therapy devices (Fig. 10 designed to produce clinically significant tissue heating operate at 0.75–3.0 MHz, the lower frequencies mechanical waves being more penetrative and affecting deeper tissues than the higher frequency waves. The ultrasound (US) is produced by the reverse piezoelectric effect when a high frequency alternating current is applied to a ceramic crystal causing it to vibrate at the same frequency. The crystal is located inside an applicator fitted with a metal cap into which these vibrations pass. A coupling medium such as either water or a gel with a high water content transmits the ultrasonic vibration into the tissues. US is transmitted readily by water and by tissue fluids. As US is more readily absorbed by protein than by fat, it can be used to heat muscle and collagenous tissue selectively without heating subcutaneous fat to a clinically significant level. Heating occurs where the US energy is absorbed. If the US enters the tissue at an angle other than 90°, it is refracted; the transducer should therefore be held perpendicular to the surface being treated.

Absorption, and therefore heating, is greatest at higher frequencies. As a consequence, it is generally accepted that higher frequency US (e.g., 3 MHz) penetrates less deeply into the body than lower frequency US (e.g., 1 MHz). However, in some circumstances, 3 MHz may heat deeper tissues than originally theorized, when for example, it is transmitted into muscle over a bony prominence via a gel pad coated on both sides with an ultrasound transmitting gel (43). In this clinical investigation, it was found that 3 MHz ultrasound applied over the lateral malleolus of the ankle heated the musculotendinous tissue deep to the peroneal groove 0.5 cm deeper than suggested by others. The 3 MHz ultrasound produced heating deeper into muscle in healthy volunteers than did the 1 MHz ultrasound. This interesting but surprising observation may be an artifact due to variations in coupling of the US transducers to the skin. The authors point out that air trapped at the gel/pad interfaces might result not only in nonuniform heating but also in hot spots on the US transducer faceplate and possibly the skin.

In recent years, US therapy devices producing kilohertz US have been developed (6); these are more penetrative than the MHz devices but produce little heat, using the nonthermal effects of US to produce their effects (7).

When US enters tissue, its intensity gradually decreases as energy is lost from the US beam by absorption, scattering, and reflection. This lessening of the force of the US is termed attenuation. US is scattered by structures smaller than its wavelength. It is reflected at the interfaces between materials with different acoustic impedances, e.g., air and skin, collagen and ground substance, periosteum, and bone. Muscle and bone are preferentially heated at their interfaces with other tissues, e.g., tendon and periosteum. This is because mode conversion occurs when US waves are reflected from these interfaces, the longitudinal incident waves being changed into reflected transverse waves, creating additional heating. Therapists can take advantage of this enhanced thermal effect to target inflamed joints. The thickness of tissue through which the US must pass for its intensity to be reduced to half the level applied to its surface is termed the half-value thickness. The half-value thickness of 1 MHz US is theoretically three times more than that of 3 MHz US although artifacts during clinical use may alter this (43).

The intensity range of US necessary to elevate tissue temperature to between 40 and 45 °C is 1.0–2.0 $W \cdot cm^{-2}$ applied in continuous mode for 5–10 min (44). This increase is in the generally accepted thermal therapeutic range. Temperatures exceeding this can cause thermal necrosis and must be avoided. In poorly vascularized tissues, even the upper end of the therapeutic range can produce thermal damage. When US is absorbed or when mode conversion occurs, heat is generated in the tissues. This produces local increases in blood flow and softens collagenous deposits such as those in scars.

If the temperature increase is less than 1 °C, this is not considered to be clinically significant. Therapeutic effects produced by US used in a manner that produces a temperature increase of less than 1 °C are considered to be predominantly nonthermal. These effects, which include stable cavitation and acoustic streaming (3,44), occur at intensities lower than those necessary to produce therapeutic increases in temperature. The amount of acoustic energy entering the tissues can be reduced by pulsing the US. A commonly used pulsing sequence is 2 ms ON, 8 ms OFF. By reducing the total amount of energy supplied to the tissues, the thermal load on these tissues is reduced. The intensity during the pulse is sufficient to permit the predominantly nonthermal therapeutic events to occur. This is of particular value when ultrasound is used to treat tissue with a blood supply too poor to ensure adequate heat dissipation via the circulation.

The nonthermal therapeutic effects of US include stable cavitation and acoustic streaming. Cavitation is the production of bubbles of gas a few microns in diameter in fluid media such as tissue fluid and blood. The bubbles increase in size during the negative pressure or rarefaction part of the US wave and decrease during the positive pressure part. If the intensity is sufficiently great, the bubbles collapse, damaging the tissues. At lower intensities, within the therapeutic range, the

cavities are stable and acoustic streaming is increased around them. Acoustic streaming has been shown to cause reversible changes in membrane permeability to calcium and other ions, stimulating cell division, collagen synthesis, growth factor secretion, myofibroblast contraction, and other cellular events that collectively accelerate the resolution of acute inflammation leading to more rapid tissue repair (3).

When and When Not to Use These Devices. US therapy is used as a thermotherapeutic modality to treat acute and chronic injuries of the

- Skin, e.g., pressure ulcers and venous ulcers
- Musculoskeletal system, e.g., arthritis, bursitis, muscle spasms, and traumatic injuries to both soft tissue and bone.

In addition to superficial injuries, US can be used to heat tissues at a considerable depth from the surface of the skin. Sites of pathology such as damaged skeletal muscle, tendon, and ligaments within 5 cm of the surface can be heated preferentially; any adipose tissue superficial to these lesions is heated less because it absorbs less US than highly proteinaceous muscle, tendons, and ligaments. It is the most suitable form of thermotherapy to use on deeply located areas of damage in obese patients. At least 3 treatments per week are generally recommended, daily treatment being preferable. The treatment head should be moved throughout treatment to reduce the possibility of thermal tissue damage due to local hot spots and mechanical damage due to standing wave formation.

In the interests of safety, continuous US should only be used to heat tissues in patients sensitive to heat-induced pain. In patients lacking this sensitivity, pulsed ultrasound can be used. Although this may not heat the tissues significantly, nonthermal events will occur that can accelerate healing (3,44).

It should not be used either over tumors because it can increase cell division or over the eye because of the risk of collapse cavitation. Nor should it be used on any of the groups of patients in whom SWD is contraindicated.

Advantages and Disadvantages. Its differential absorption makes it the therapy of choice for treating muscle, tendons, ligaments, and bone without heating superficial adipose tissue.

The effects of ultrasound are, however, local and only small regions can be treated because of the small size of the treatment heads, rarely greater than 5 cm^2. Another disadvantage is that a coupling medium such as water or a gel with a high water content must be used to transit the US from the applicator to the tissues. If there is an open wound, this must be covered with an ultrasound transmitting dressing such as Opsite to the surface of which the coupling medium can be applied. Alternatively the intact tissue adjacent to the wound and from which reparative tissue grows into the wound can be treated (3). Care must be taken to ensure that there are no bubbles in the coupling medium because reflection from these can reduce the efficiency to energy transfer from the treatment head to the tissue. If the US energy is reflected back into the treatment head because of inadequate coupling, this will increase in temperature and could burn a patient, lacking adequate sensitivity to heat-induced pain.

Cryotherapy

Cryotherapy produces a rapid fall in the temperature of the skin and a slower fall in the temperature of the subcutaneous tissues, skeletal muscle, and the core temperature of the body. The rate of fall in skeletal muscle and the core temperature is dependent, in part, on the amount and distribution of adipose tissue. In a slender person with less than 1 cm of adipose tissue in the hypodermis, cooling extends almost 2.5 cm into the skeletal muscle after 10 min of applying cryotherapy to the skin surface. In an obese person with more than 2.5 cm of adipose tissue in the hypodermis, cooling extends only 1 cm into the muscle in the same time (1). To get deeper cooling in an obese person, the cooling agent must be applied for longer than in a slender person, for example, 30 min compared with 10 min to get adequate cooling extending 2.5 cm into the muscle. This concept has been confirmed recently by Otte et al. (45) who found that although 25 min of treatment may be adequate for a slender patient with a skinfold of 20 mm or less, 40 min is needed if the skinfold is between 21 and 30 mm, and 60 min if the skinfold is between 30 and 40 mm. The subcutaneous adipose tissue thickness is an important determinant of the time required for cooling in cryotherapy.

Cryotherapy is used to relieve pain, retard the progression of secondary injury, and hasten return to participation in sport and work. In a literature review by Hubbard et al. in 2004 (46), the conclusion drawn was that cryotherapy may have a positive effect on these aims, but attention was drawn to the relatively poor quality of the studies reviewed. There is a clear need for randomized, controlled clinical studies of the effect of cryotherapy on acute injury and return to participation in sport or work.

The main equipment used in cryotherapy is a refrigerator/freezer necessary for cooling gels and for producing ice. The ice is mixed with water, reducing the temperature of the water to just above its freezing point. The temperature of an injured limb can be reduced by immersing it in this ice/water mixture, an effective but initially uncomfortable experience for the patient. Alternatively, a cold compress containing the mixture can be applied to the region to be cooled. Also, a terry cloth can be soaked in the mixture, wrung out, and then applied. Blocks of ice can be used to massage an injured area if the skin over the injury is intact, an initially uncomfortable experience for both patient and therapist.

Another technique is to spray a vapor coolant on the skin. As the coolant evaporates, the temperature of the skin is reduced, but there is no clinically significant cooling of subcutaneous tissues. Ethylene chloride, chlorofluoromethane, or preferably a non-ozone-depleting vapor coolant, is sprayed over the area to be treated in a stroking fashion at a rate of about 1 $cm \cdot s^{-1}$ (1). Concern over the ozone-depleting properties of chlorofluorocarbons has led

to the development of vapor coolant sprays that are not ozone-depleting (http://www.gebauerco.com/Default.asp) and that may be substituted for those noted above.

When and When Not to Use Cooling. The main use of cryotherapy is after acute skin and musculoskeletal injury to reduce swelling, bleeding, and pain. It helps give temporary relief to painful joints. It also reduces muscle spasms and spasticity, again temporarily. Trigger points, myofascial pain, and fibrositis may be treated with vapor coolant sprays.

Cooling is an excellent way of minimizing the effect of burns and scalds by reducing the temperature at the injury site provided that it is done as soon as possible after the incident. It can also slow brain damage after the deprival of oxygen in, for example, babies for whom delivery has been prolonged and difficult. A cool cap filled with chilled water is placed on the baby's head within a few hours of birth and kept there for several days while the baby is in intensive care. A computerized controller circulates cold water through the cap, reducing the temperature of the brain by several degrees, minimizing cell death within the brain. The results of a clinical trial showed a significantly lower disability and death rate in children at risk of post-natal brain damage due to oxygen deprival if they were given this treatment than in those whose brains had not been cooled in this way (47). It has been suggested that a similar technique may help patients with hemorrhagic strokes where bleeding has occurred on to the surface of the brain.

Mild systemic hypothermia has also been reported to reduce brain damage after severe head trauma. Patients in whom the core temperature was reduced to and maintained at 33–35 °C with a cooling blanket for 4 days had reduced extradural pressure, an increase in the free radical scavenger superoxide dismutase, and consequently, less neuron loss and improved neurological outcomes (48).

Retarding secondary injury is an important benefit of cryotherapy. Secondary tissue death occurring after the initial trauma has been attributed to subsequent enzymatic injury and hypoxic injury (23). Cryotherapy reduces tissue temperature, slowing the rate of chemical reactions and therefore the demand for adenosine triphosphate (ATP), which in turn decreases the demand for oxygen, leading to longer tissue survival during hypoxia. By decreasing the amount of damaged and necrotic tissue, the time taken to heal may be reduced. In an extensive review of the secondary injury model and the role of cryotherapy, Merrick et al. (24) addressed the following question: "Is the efficacy of short-term cryotherapy explained by rescuing or delaying the death of the cells that were primarily injured but not initially destroyed?" He recommended the replacement of the term "secondary hypoxic injury" by "secondary ischemic injury" because hypoxia presents tissue with the challenge of reduced oxygen only, whereas ischemia presents inadequacies in not only oxygen but also fuel and anabolic substrates and in waste removal, all of which may contribute to secondary injury. Short-term cryotherapy may lessen the demand for these necessities.

Cooling should not be used:

- In people who are hypersensitive to cold-induced pain
- Over infected open wounds
- In people with a poor circulation

Advantages and Disadvantages. Disadvantages are that many people find prolonged exposure to cold therapy uncomfortable. Frostbite can occur if the skin freezes, as is possible if vapor coolant sprays are overused. Furthermore, its effects on chronic inflammatory conditions are generally temporary.

SUMMARY

Changing the temperature of tissues can reduce pain, muscle spasms, spasticity, stiffness, and inflammation. Heating the tissues by a few degrees centigrade increases tissue metabolism and accelerates the healing process. Cooling tissue by a few degrees centigrade can limit secondary damage to soft tissues, nerves, and the brain after trauma.

The correct selection of either thermotherapy or cryotherapy depends on an understanding of the physiological effects of heat and cold, and on knowledge of the patient's medical condition. It is suggested that acute skin lesions and musculoskeletal injuries be treated:

- First by the application of cold as soon as possible after the injury to limit its extent
- Followed a few hours later by the application of moderate heat to accelerate the resolution of inflammation enabling healing to progress more rapidly.

It is recommended that the progress of healing be monitored noninvasively so that treatment can be matched to the response of the injury. In recent years this has become possible by means of high resolution diagnostic ultrasound or ultrasound biomicroscopy (49).

Having decided which is required, heat or cold, the next question is which modality to use. The following should be considered:

- Size and location of the injury
- Type of injury
- Depth of penetration of the modality
- Ease of application
- Duration of application
- Affordability
- Medical condition of the patient
- Contraindications

Patients heal their own injuries if they can to. This can be facilitated by the timely and correct application of therapeutic heat or cold. Their pain can also be relieved, improving their quality of life.

BIBLIOGRAPHY

Cited References

1. Tepperman PS, Kerosz V. In: Webster JG, ed., Encyclopedia of Medical Devices and Instrumentation. 1st ed. New York: John Wiley & Sons; 1988. p 14811–1493.

2. Lehmann JF. Therapeutic Heat and Cold. 3rd ed. Baltimore: Williams & Wilkins; 1982.
3. Dyson M. Adjuvant therapies: ultrasound, laser therapy, electrical stimulation, hyperbaric oxygen and negative pressure therapy. In: Morison MJ, Ovington LG, Wilkie LK, editors. Chronic Wound Care: A Problem-Based Learning Approach. Edinburgh: Mosby; 2004. 129–159.
4. Guy AW, Chou CK, Neuhaus B. Average SAR and distribution in man exposed to 450 mMHz radiofrequency radiation. IEEE Trans Microw Theory Tech 1984;MTT–32:752–762.
5. Dyson M, Suckling J. Stimulation of tissue repair by ultrasound. A survey of the mechanisms involved. Physiotherapy 1978;64:105–108.
6. Peschen M, Weichenthal MM, Schopf E, Vanscheidt W. Low-frequency ultrasound treatment of chronic venous ulcers in an outpatient therapy. Acta Dermatovenereol 1997;77:311–314.
7. Ward AR, Robertson VJ. Comparison of heating of nonliving soft tissue produced by 45 kHz and 1 MHz frequency ultrasound machines. J Org Sports Physiotherapists 1996;23:258–266.
8. Kitchen S, Dyson M. Low-energy treatments: non-thermal or thermal? In: Kitchen S, editor. Electrotherapy. Evidence-Based Practice. Edinburgh: Churchill Livingstone; 2002. 107–112.
9. Stanton DB, Bear-Lehman J, Graziano M, Ryan C. Contrast baths: What do we know about their use? J Hand Ther 2003;16:343–346.
10. Kitchen S. Thermal effects. In: Kitchen S, editor. Electrotherapy. Evidence-Based Practice. Edinburgh: Churchill Livingstone; 2002. 89–105.
11. Bowler K. Cellular heat injury: Are membranes involved? In: Bowler K, Fuller BJ, editors. Temperature and animal cells. Cambridge: Company of Biologists; 1987. 157–185.
12. Westerhof W, Siddiqui AH, Cormane RH, Scholten A. Infrared hyperthermia and psoriasis. Arch Dermatol Res 1987;279:209–210.
13. Lehmann JF, Masock AJ, Warren CG, Koblanski JN. Effects of therapeutic temperatures on tendon extensibility. Arch Phys Med Rehab 1970;51:481–487.
14. Feebel H, Fast H. Deep heating of joints: A reconsideration. Arch Phys Med Rehab 1976;57:513–514.
15. Millard JB. Effects of high frequency currents and infrared rays on the circulation of the lower limb in man. Ann Phys Med 1961;6:45–65.
16. Wyper DJ, McNiven DR. Effects of some physiotherapeutic agents on skeletal muscle blood flow. Physiotherapy 1976;62: 83–85.
17. McMeeken JM, Bell C. Microwave irradiation of the human forearm and hand. Physiother Theory Practice 1990;6:171–177.
18. Lehmann JF, de Lateur BJ. Therapeutic heat. In: Lehmann JF, editor. Therapeutic Heat and Cold. 4th ed. Baltimore: Williams & Watkins; 1990. 444.
19. Mense S. Effects of temperature on the discharges of muscle spindles and tendon organs. Pflug Arch 1978;374:159–166.
20. Lehmann JF, de Lateur BJ. Ultrasound, shortwave, microwave, laser, superficial heat and cold in the treatment of pain. In: Wall PD, Melzack R, editors. Textbook of Pain. 4th ed. New York: Churchill Livingstone; 1999. 1383–1397.
21. Barbour LA, McGuire DS, Kirchott KT. Nonanalgesic methods of pain control used by cancer outpatients. Oncol Nurs For 1986;13:56–60.
22. Doubell P, Mannon J, Woolf CJ. The dorsal horn: state dependent sensory processing, plasticity and the generation of pain. In: Wall PD, Melzack R, editors. Textbook of Pain. 4th ed. New York: Churchill Livingstone; 1999. 165–182.
23. Knight KL. Cryotherapy in Sports Injury Management. Champaign (IL): Human Kinetics; 1995.
24. Merrick MA, Rankin JM, Andrea FA, Hinman CL. A preliminary examination of cryotherapy and secondary injury in skeletal muscle. Med Sci Sports Exerc 1999;31:1516–1521.
25. Merrick MA. Secondary injury after musculoskeletal trauma: A review and update. J Athl Train 2002;37:209–217.
26. Melling DAC, Baqar A, Scott EM, Leaper DJ. Effects of preoperative warming on the incidence of wound infection after clean surgery: a randomized controlled trial. Lancet 2001;358:876–880.
27. Ellis SL, Finn P, Noone M, Leaper DJ. Eradication of methicillin-resistant Staphylococcus aureus from pressure sores using warming therapy. Surg Infect 2003;4:53–55.
28. MacFie CC, Melling AC, Leaper DJ. Effects of Warming on Healing. J Wound Care 2005;14:133–136.
29. Xia Z, Sato A, Hughes MA, Cherry GW. Stimulation of fibroblast growth in vitro by intermittent radiant warming. Wound Rep Regen 2000;8:138–144.
30. Hughes MA, Tang C, Cherry GW. Effect of intermittent radiant warming on proliferation of human dermal endothelial cells in vitro. J Wound Care 2003;12:135–137.
31. Zarro V. Mechanisms of inflammation and repair. In: Michlovitz SL, editor. Thermal Agents in Rehabilitation. Philadelphia: Davis; 1986. 3–17.
32. Gluckman PD, Wyatt JS, Azzopardi D, Ballard R, Edwards AD, Ferriero DM, Polin RA, Robertson CM, Thoresen M, Whitelaw A, Gunn AJ. Selective head cooling with mild systemic hypothermia after neonatal encephalopathy: multicentre randomized trial. Lancet 2005;365:632–634.
33. Keatinge WR. Survival in Cold Water. Oxford: Blackwell; 1978. 39–50.
34. Price R, Lehmann JF, Boswell-Bessett S, Burling S, de Lateur R. Influence of cyrotherapy on spasticity at the human ankle. Arch Phys Med Rehab 1993;74:300–304.
35. Wall PD, Melzack R, editors. Textbook of Pain. 4th ed. New York: Churchill Livingstone; 1999.
36. Fields HL, Basbaum AI. Central nervous system mechanisms of pain. In: Wall PD, Melzack R, editors. Textbook of Pain. 4th ed. New York: Churchill Livingstone; 1999. 309–330.
37. Tepperman PS, Devlin M. Therapeutic heat and cold. Postgrad Med 1983;73:69–76.
38. Batavia M. Contraindications for superficial heat and therapeutic ultrasound: Do sources agree? Arch Phys Med Rehabil 2004;85:1006–1012.
39. Tuner J, Hode L. The Laser Therapy Handbook. Grangesberg: Prima Books AB; 2004.
40. Scott S, McMeeken J, Stillman B. Diathermy. In: Kitchen S, editor. Electrotherapy. Evidence-Based Practice. Edinburgh: Churchill Livingstone; 2002. 145–170.
41. Low J. Dosage of some pulsed shortwave clinical trials. Physiotherapy 1995;81:611–616.
42. Weinberger A, Fadilah R, Lev A, et al. Treatment or articular effusions with local deep hyperthermia. Clin Rheumatol 1989;8:461–466.
43. Hayes BT, Merrick MA, Sandrey MA, Cordova ML. Three-MHz ultrasound heats deeper into tissue than originally theorized. J Athl Train 2004;39:230–243.
44. Sussman C, Dyson M. Therapeutic and diagnostic ultrasound. In: Sussman C, Bates-Jensen BM, editors. Wound Care. A Collaborative Practice Manual for Physical Therapists and Nurses. 2nd ed. Gaithersburg: Aspen Publishers; 2001. 596–620.
45. Otte JW, Merick MA, Ingersoll CD, Cordova ML. Subcutaneous adipose tissue thickness alters cooling time during cryotherapy. Arch Phys Med Rehabil 2002;83:1501–1505.
46. Hubbard TJ, Aronson SL, Denegar CR. Does cryotherapy hasten return to participation? A systematic review. J Athl Train 2004;39:88–94.

47. Gluckman PD, Wyatt JS, Azzopardi D, Ballard R, Edwards AD, et al. Selective head cooling with mild systemic hypothermia after neonatal encephalopathy: A multicentre randomised trial. Lancet 2005;365:663–670.
48. Qui WS, Liu WG, Shen H, Wang WM, Hang ZL, Jiang SJ, Yang XZ. Therapeutic effect of mild hypothermia on severe traumatic head injury. Chin J Traumatol 2005;8:27–32.
49. Dyson M, Moodley S, Verjee L, Verling W, Weinman J, Wilson P. Wound healing assessment using 20 MHz ultrasound and photography. Skin Res Technol 2003;9: 116–121.

HEAVY ION RADIOTHERAPY.

See RADIOTHERAPY, HEAVY ION.

HEMODYNAMICS

PATRICK SEGERS
PASCAL VERDONCK
Ghent University
Belgium

INTRODUCTION

"Arterial blood pressure and flow result from the interaction of the heart and the arterial system. Both subsystems should be considered for a complete hemodynamic profile and a better diagnosis of the patient's disease". This statement seems common sense, and a natural engineering approach of the cardiovascular system, but is hardly applied in clinical practice, where clinicians have to deal with limitations imposed by the clinical environment and ethical and economical considerations. The result is that the interpretation of arterial blood pressure is (too) often restricted to the interpretation of systolic and diastolic blood pressure measured using the traditional cuff around the upper arm (cuff sphygmomanometry). Blood flow, if even measured, is usually limited to an estimate of cardiac output.

The purpose of this article is to provide the reader with an overview of both established and newer methods and techniques that allow us to gain more insight into the dynamics of blood flow in the cardiovascular system (the hemodynamics), based on both invasive and noninvasive measurements. The emphasis is that hemodynamics results from the interaction between the action of the heart and the arterial system, and can be analyzed as the interplay between a (complex) pump and a (complex) tube network. This article, has been divided into three main sections. First the (mechanical function of the) heart is considered, followed by a major section on arterial function analysis. The final section deals with cardiovascular interaction.

THE HEART AS A PUMP...

The heart is a hollow muscle, consisting of four chambers, whose function is to maintain blood flow in two circulations: the systemic (or large) and the pulmonary circulation. The left atrium receives oxygenized blood from the lungs via the pulmonary veins. Blood flows (through the mitral valve) into the left ventricle, where it is pumped into the aorta (through the aortic valve) and distributed toward the organs, tissue, and muscle for exchange of O_2 and CO_2, nutrients, and waste products. Deoxygenated blood is collected via the systemic veins (ultimately the inferior and superior vena cava) into the right atrium and flows, via the tricuspid valve, into the right ventricle, where it is pumped (through the pulmonary valve) into the pulmonary artery toward the lungs. Functionally, the pulmonary and systemic circulation are placed in series, and there is a "serial interaction" between the left and right heart. Anatomically, however, the left and right heart are embedded within the pericardium (the thin membrane surrounding the whole heart) and are located next to each other. The part of the cardiac muscle (myocardium) that they have in common is called the myocardial septum. Due to these constraints, the pumping action of one chamber has an effect on the other, a form of "parallel interaction". In steady-state conditions, the left and right heart generate the same flow (cardiac output), on average ~ 6 L/min in an adult at rest.

The Cardiac Cycle and Pressure–Volume Loops

The most heavily loaded chamber is the left ventricle (LV), pumping ~ 80 mL of blood with each contraction (70 beats $\cdot$ min^{-1}), with intraventricular pressure increasing from ~ 5–10 mmHg (1 mmHg $= 133.3$ Pa) at the onset of contraction (i.e., at the end of the filling period or diastole) to 120 mmHg (6.0 kPa) in systole (ejection period) (Fig. 1). In heart physiology research, it is common to study the function of the ventricle using pressure–volume loops (PV loops; Fig. 1), with volume on the x axis and pressure on the y axis. Considering the heart at the end of diastole, it has reached its maximal volume (EDV; end-diastolic volume). Specialized pacemaker cells within the heart generate the (electrical) stimulus for the contraction, initiating the depolarization of cardiac muscle cells (myocytes). Electrical depolarization causes the muscle to contract and ventricular pressure increases. With this, the mitral valve closes, and the ventricle contracts at closed volume, with a rapidly increasing pressure (isovolumic contraction). When LV pressure becomes higher than the pressure in the aorta, the aortic valve opens, and the ventricle ejects blood. The ventricle then starts its relaxation, slowing down the ejection, with a decrease in LV pressure. At the end of ejection, the LV has reached its end-systolic volume (ESV), and LV pressure drops below the pressure in the aorta, closing the aortic valve. Relaxation then (rapidly) takes place at closed volume (isovolumic relaxation), until LV pressure drops below LA pressure and LV early filling begins (E-wave). After complete relaxation of the ventricle, contraction of the LA is responsible for an extra (late) filling wave (A-wave). The difference between EDV and ESV is the stroke volume, SV. Multiplied with heart rate, one obtains cardiac output (CO), the flow generated by the heart, commonly expressed in $L \cdot min^{-1}$. The time course of cardiac and arterial pressure and flow is shown in Fig. 1.

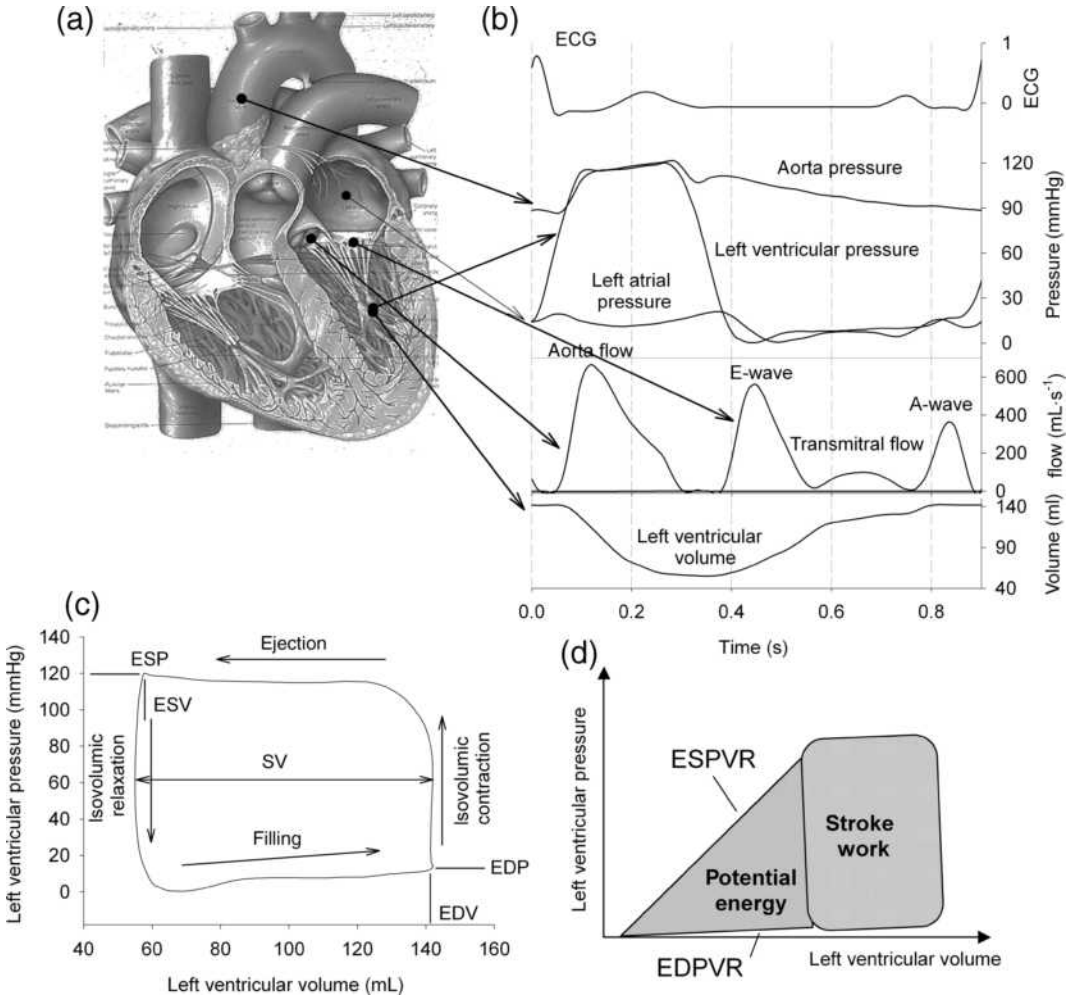


Figure 1. The heart (a), and the variation in time of pressure, flow, and volume within the left ventricle (b). Plotting left ventricular pressure as a function of volume (c), a pressure–volume loop is obtained. (d) Illustrates the association between area's defined within the pressure–volume plane and the mechanical energy.

Time Varying Elastance and Cardiac Contractility

The intrinsic properties of cardiac muscle are responsible for making the functional pumping performance of the ventricle determined by different factors (1): the degree to which the cardiac muscle is prestretched prior to contraction (preload), the intrinsic properties of the muscle (the contractility or inotropy), the load against which the heart ejects (the afterload), and the speed with which the contraction takes place (reflected by the heart rate; chronotropy). In *muscle physiology*, preload is muscle length and is related to the overlap distance of the contractile proteins (actin and myosin) of the sarcomere (the basic contractile unit of a muscle cell), while afterload is the load against which a muscle strip or fiber contracts. In *pump physiology*, ventricular end-diastolic volume is often considered as the best approximation of preload (when unavailable, ventricular end-diastolic pressure can be used as a surrogate). To characterize afterload, one can estimate maximal ventricular wall stress (e.g., using Laplace formula), but most often, mean or systolic arterial blood pressure is taken as a measure of afterload.

Most difficult to characterize is the intrinsic contractility of the heart, which is important to know in diagnosing the severity of cardiac disease. At present, the gold standard is still considered to be the slope of the end-systolic pressure–volume relation (2,3). To fully comprehend this measure, the time varying elastance concept has to be introduced.

Throughout a cardiac cycle, cardiac muscle contracts and relaxes. The functional properties of fully relaxed muscle-at the end of diastole-can be studied in the pressure–volume plane. This relation, sometimes called the (end-) diastolic pressure–volume relation (EDPVR), is nonlinear, that is, for higher volumes, a higher increase in pressure (ΔP) is required to realize a given increase in volume (ΔV). With the volume/pressure ratio defined as compliance, the ventricle behaves less compliant (stiffer) at high volume. The EDPVR represents the passive properties of the ventricular chamber.

Similarly, if one could "freeze" the ventricle at its maximal contraction (as is reached at the end of systole), and measure the pressure–volume relation of the chamber in

this maximally contracted state, one could assess the (end-) systolic pressure–volume relation (ESPVR) and the maximal stiffness of the ventricle. The ESPVR represents the active, contractile function of the ventricle. Throughout the cycle, the stiffness (or elastance) of the ventricle varies in between its diastolic and end-systolic value, hence the conceptual model of the time-varying elastance (Fig. 2).

The basis of this time-varying elastance model was laid by Suga et al. in the 1970s. They performed experiments in isolated hearts, and found that by increasing the initial volume of the heart (increasing preload), the maximally developed pressure in an isovolumic beat increased linearly with preload (2,3) (Frank–Starling mechanism). Obviously, the minimal pressure, determined by the passive properties, also increased. When they allowed these ventricles to eject against a quasiconstant afterload pressure, PV loops were obtained, with wider loops (higher stroke volume) being obtained for the more filled ventricles. When connecting all end-systolic points of the PV loops, it was found that the slope of this line, the end-systolic ventricular stiffness, was not different from the line obtained with the isovolumic experiments, demonstrating that it is independent of the load against which the ventricle ejects. Moreover, connecting data points on the PV loops occurring at the same instant in the cardiac cycle (isochrones), these points were also found to line up (Fig. 2). The slopes of these lines have the dimension of stiffness ($\Delta P/\Delta V$; $mmHg \cdot mL^{-1}$ or $Pa \cdot mL^{-1}$) or elastance (E). In addition, it is often assumed that these isochrones all have the same intercept with the volume axis (which is, however, most often not the case). This volume is called V_0 and represents the volume for which the ventricle no longer develops any pressure.

The slope of the isochronic lines, E, is given by $E = P/(V - V_0)$ and can be plotted as a function of time, yielding the time varying elastance curve, $E(t)$ (Fig. 2). The experiments of Suga and co-workers further pointed out that the maximal slope of the ESPVR, also called end-systolic (E_{es}) elastance, is sensitive to inotropic stimulation. The parameter E_{es} is, at present, still considered as the gold standard measurement of ventricular contractility.

Since these experiments, it has been shown that the ESPVR is not truly linear (4,5), especially not in conditions of high contractility or in small mammals. Since V_0 is a value derived from linear extrapolation of the ESPVR, one often finds negative values, which clearly have no physiological meaning at all. Nevertheless, the time varying elastance remains an attractive concept to concisely describe ventricular function. In practice, PV loops with altered loading conditions are obtained via inflation of a balloon in one of the caval veins, reducing venous return, or with a Valsalva maneuver. The PV loops can be measured invasively with a conductance catheter (6), or by combining intraventricular pressure (measured with a catheter) with volumes measured with a medical imaging technique that is fast enough to measure instantaneous volumes during the load manipulating operations (e.g., echocardiography).

The area enclosed within the PV loop is the work performed by the heart per stroke (stroke work, SW). Furthermore, when the heart contracts, it pressurizes the volume within the ventricle, giving it a potential energy (PE). In the PV plane, PE is represented by the area enclosed within the triangle formed by V_0 on the volume axis, the end-systolic point, and the left bottom corner of the PV loop (Fig. 1). The sum of SW and PE is also called the total pressure–volume area (PVA) and it has been shown

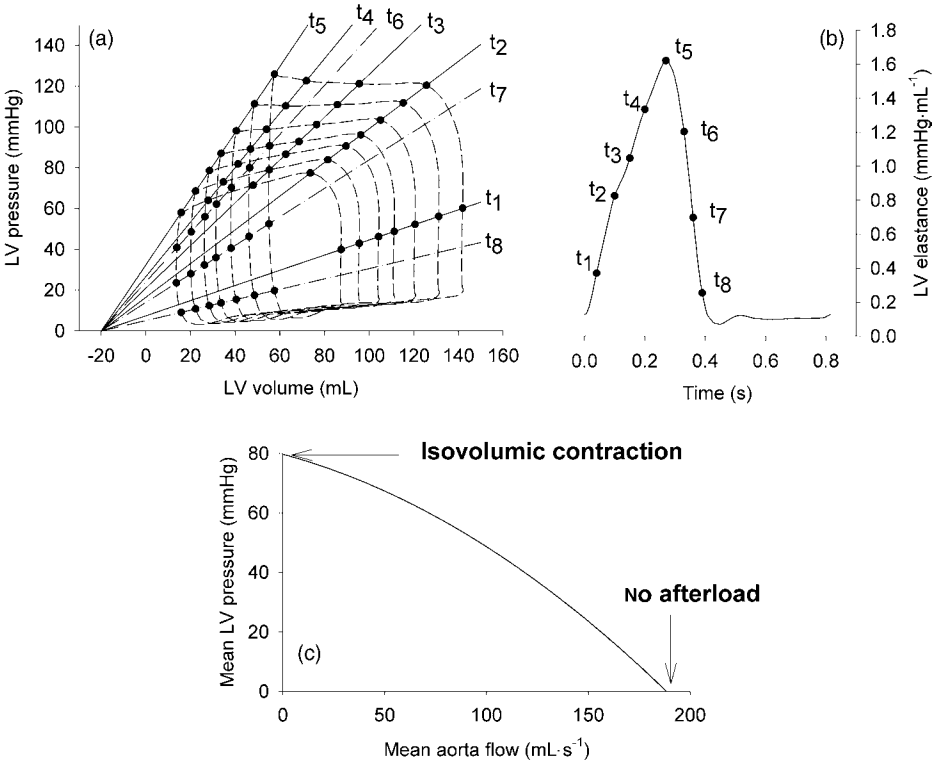


Figure 2. Pressure–volume loops recorded in the left ventricle during transient loading conditions (a), and the concept of the time-varying elastance (b). (c) Illustrates an alternative representation of ventricular function through the pump function graph.

that the consumption of oxygen by the myocardium, $V\mathrm{O}_2$, is proportional to PVA: $V\mathrm{O}_2 = c_1\mathrm{PVA} + c_2E_{\mathrm{es}} + c_3$, with c_{1-3} constants to be determined from experiments. The constant c_1 represents the O_2 cost of contraction, c_2 is the O_2 cost of Ca handling related to the inotropic state, and c_3 is the O_2 cost of basal metabolism. Mechanical efficiency can then be expressed as the ratio of SW and $V\mathrm{O}_2$. Recent reviews of the relation between pressure–volume area and ventricular energetics can be found in Refs. 7,8.

Another measure of ventricular function, also derived from PV loop analysis, is the so-called preload recruitable stroke work (PRSW) (9). Due to the Frank–Starling effect (10), a ventricle filled up to a higher EDV will generate a higher pressure and/or stroke volume, and hence a higher SW. Plotting SW as a function of EDV yields a quasilinear relation, of which the slope is sensitive to the contractile state of the ventricle (9).

Alternative Ways of Characterizing LV Systolic Function

Pump function of the ventricle may also be approached in a way similar to hydraulic pumps through its pump function graph (11,12), where the pressure generated by the pump (e.g., mean LV pressure) is plotted as a function of its generated flow (cardiac output). With no outflow, the ventricle contracts isovolumically, and the highest possible pressure is generated. Pumping against zero load, no pressure is built up, but outflow is maximal. The ventricle operates at some intermediate stage, in between these two extreme cases (Fig. 2). One such pump function curve is obtained by keeping heart rate, inotropy, and preload constant, while changing afterload. Although the principle is attractive, it appears to be difficult to measure pump function curves *in vivo*, even in experimental conditions.

Assessing Cardiac Function in Real Life

Although pressure–volume loop-based cardiac analysis still has the gold standard status in experimental work, the applicability in clinical conditions is rather limited. First, the method requires intraventricular pressure and volume. While volumes can, more and more, be measured noninvasively with magnetic resonance imaging (MRI) and even real-time three-dimensional (3D) echocardiography, the pressure requirement still implies invasive measurements. Combined pressure–volume conductance catheters are available (6), but these require calibration to convert conductance into volume data. This requires knowledge of the conductance of the cardiac structures and blood outside the cardiac chamber under study (offset correction), and an independent measurement of stroke volume for scaling of the amplitude of the conductance signal. Second, and perhaps even more important, measuring preload-recruitable stroke work or the end-systolic pressure–volume relation requires that PV loops are recorded during transient loading conditions, which is experimentally obtained via inflation of a balloon in the caval vein to reduce the venous return and cardiac preload. It is difficult to (ethically) justify an extra puncture and the insertion of an additional catheter in patients, knowing also that these maneuvers induce secondary changes in the overall autonomic state of the patient and the release of catecholamines, making this method limited to assess the pump function in a "steady state". To avoid the necessity of the caval vein balloon, so-called "single beat" methods have been proposed, where cardiac contractility is estimated from data measured at baseline steady-state conditions (13,14). The accuracy, sensitivity, and specificity of these methods, however, remains a matter of debate (15,16).

Since it is easier to measure aortic flow than ventricular volume, indicies based on the concept of "hydraulic power" have been proposed. As the ventricle ejects blood, it generates hydraulic power (Pwr), which is calculated as the instantaneous product of aorta pressure and flow. The peak value of power ($\mathrm{Pwr}_{\mathrm{max}}$) has been proposed as a measure of ventricular performance. However, due to the Frank–Starling mechanism, ventricular performance in general, and hydraulic power in particular, is highly dependent on the filling state of the ventricle, and correction of the index for EDV is mandatory. Preload-adjusted maximal power, defined as $\mathrm{Pwr}_{\mathrm{max}}/\mathrm{EDV}^2$, has been proposed as a "single beat" (i.e., measurable during steady-state conditions) index of ventricular performance (17). It has, however, been suggested that the factor 2 used in the denominator is not a constant, but depends on ventricular size (18). It has been demonstrated that the most correct approach is to correct $\mathrm{Pwr}_{\mathrm{max}}$ for $(\mathrm{EDV}-V_0)^2$, V_0 being the intercept of the end-systolic pressure–volume relation (19,20). Obviously, the index then loses its main feature, that is, the fact that it can be deduced from steady-state measurements.

In clinical practice, cardiac function is commonly assessed with ultrasound echocardiography in its different modalities. Imaging the heart in two-dimensional (2D) planar views (2 and 4 chamber long axis views, short axis views), allows us to visually inspect ventricular wall motion and to identify noncontracting zones. With (onboard) image processing software, parameters such as ejection fraction or the velocity of circumferential fiber shortening can be derived. Echocardiography has played a major role in quantifying "diastolic function", that is, the filling of the heart (21,22). Traditionally, this was based on the interpretation of flow velocity patterns at the mitral valve (23) and the pulmonary veins (24). With the advent of more recent ultrasound processing tools, the arsenal has been extended. Color M-mode Doppler, where velocities are measured along a base-to-apex directed scanline, allows us to measure the propagation velocity of the mitral filling wave (25,26). More recent, much attention has been and is being paid to the velocity of the myocardial tissue (in particular the motion of the mitral annulus) (27). Further processing of tissue velocity permits us to estimate the local strain and strain rate within sample volumes positioned within the tissue. Strain and strain rate imaging are new promising tools to quantify local cardiac contractile performance (28,29). Further advances are directed toward real time 3D imaging with ultrasound and quantification of function. Some of the aforementioned ultrasound modalities are illustrated in Fig. 3.

An important domain where assessment of cardiac function is important, is in the catheterization laboratory (cathlab), where patients are "catheterized" to diagnose and/or treat cardiovascular disease. Access to the vasculature is gained via a large vein (the catheter then ends in the right

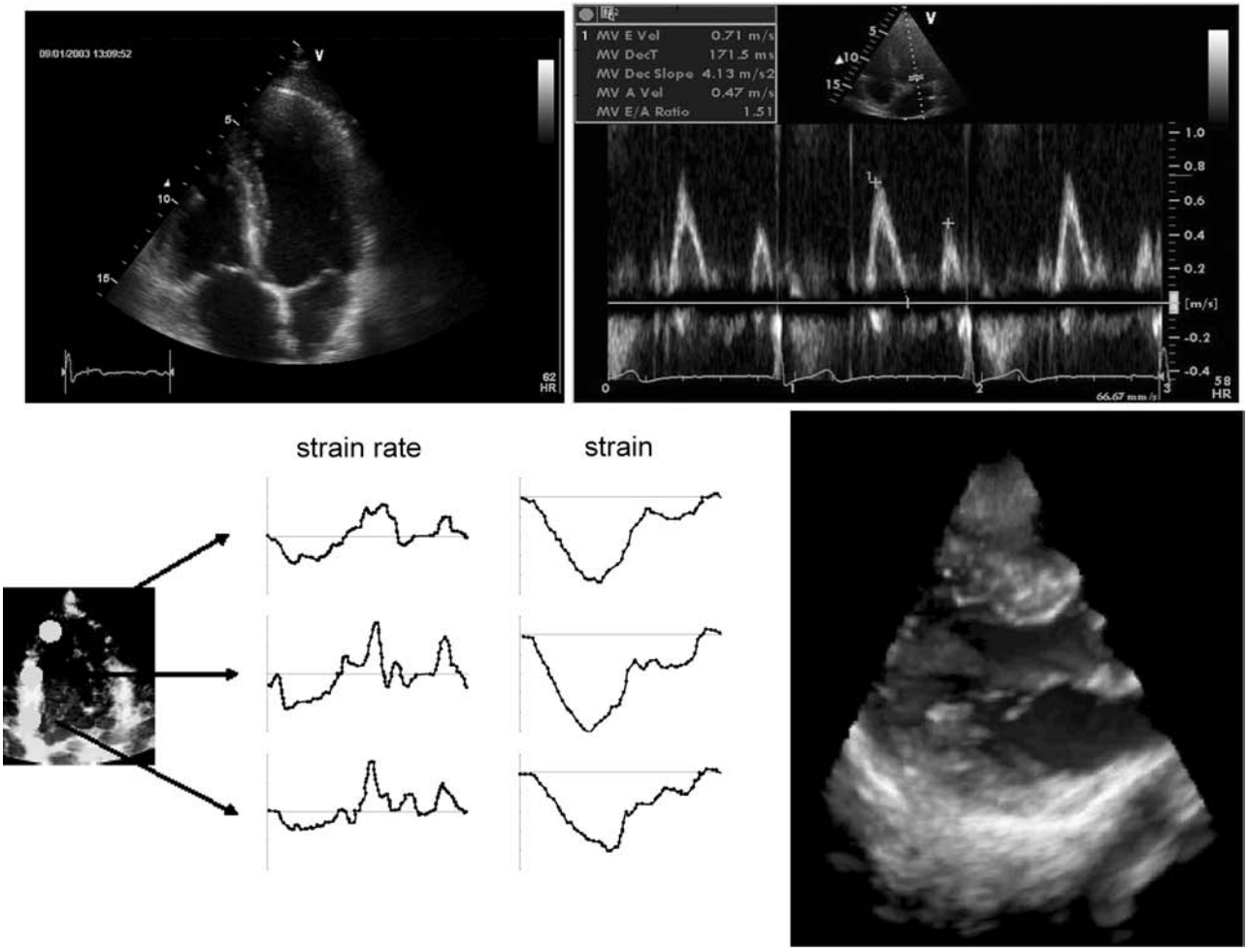


Figure 3. Echocardiography of the heart. (a) Classical four-chamber view of the heart, with visualization of the four cardiac chamber; (b) transmitral flow velocity pattern; (c) strain and strain rate imaging in an apical, mid and basal segment of the left ventricle [adapted from D'hooge et al. (29)]; (d) real-time 3D visualization of the mitral valve with the GE Dimension. (Courtesy of GE Vingmed ultrasound, Horten, Norway.)

atrium/ventricle/pulmonary artery: right heart catheterization) or large artery, typically the femoral artery in the groin (the catheter then resides in the aorta/left ventricle/left atrium: left heart catheterization). Cath-labs are equipped with X-ray scanners, allowing us to visualize cardiovascular structures in one or more planes. With injection of contrast medium, vessel structures (e.g., the coronary arteries) can be visualized (angiography) as well ventricular cavity (ventriculography). The technique, through medical image processing, allows us to estimate ventricular volumes at high temporal resolution. At the same time, arterial pressures can be monitored. Although one cannot deduce intrinsic ventricular function from pressure measurements alone, it is common to use the peak positive (dp/dt_{max}) and peak negative value (dp/dt_{min}) of the time derivative of ventricular pressure as surrogate markers of ventricular contractility and relaxation, respectively. These values are, however, highly dependent on ventricular preload, afterload, and heart rate, and only give a rough estimate of ventricular function. Measurement of dp/dt_{max} or dp/dt_{min} at different loading states, and assessing the relation between changes in volume and changes in dp/dt, may compensate for the preload dependency of these parameters.

There are also clinical conditions where the physician is only interested in a global picture of cardiac function, for example, in intensive care units, where the principal question is whether the heart is able to deliver a sufficient cardiac output. In these conditions, indicator-dilution methods are still frequently applied to assess cardiac output. In this method, a known bolus of indicator substance (tracer) is injected into the venous system, and the concentration of the substance is measured on the arterial side (the dilution curve), after the blood has passed through the heart/cardiac output = [amount of injected indicator]/[area under the dilution curve]. A commonly used indicator is cold saline, injected into the systemic veins, and the change in blood temperature is then measured with a catheter equipped with a thermistor, positioned in the pulmonary artery. This methodology is commonly referred to as "thermodilution". With valvular pathologies, as tricuspid or pulmonary valve regurgitation, the method becomes less

accurate. The variability of the method is quite high, and the method should be repeated (three times or more) so that results can be averaged to provide a reliable estimate. Obviously, the method only yields intermittent estimates of cardiac output. Cardiac output monitors based on other measuring principles are in use, but their accuracy and/or responsiveness is still not optimal and they often require a catheter in the circulation (30,31).

Finally, it is also worth mentioning that cardiac MRI is an emerging technology, able to provide full 3D cardiac morphology and function data (e.g., with MRI tagging) (32,33). Especially for volume estimation, MRI is considered the gold standard method, but new modalities also allow us to measure intracardiac flow velocities. The high cost, the longer procedure, the fact that some materials are still banned from the scanner (e.g., metal containing pacemakers) or cause image artifacts and the limited availability of MRI scanner time in hospitals, however, make that ultrasound is still the first method of choice. For reasons of completeness, computed tomography (CT) and nuclear imaging are mentioned as medical imaging techniques that provide morphological and/or functional information on cardiac function.

The Coronary Circulation

The coronary arteries branch off the aorta immediately distal to the aortic valve (Fig. 4), and supply blood to the heart muscle itself. With their specific anatomical position, they are often considered as part of the heart, although they could as well be considered as being part of the arterial circulation. This ambiguity is also reflected in the medical specialism: coronary artery disease is the territory of the (interventional) cardiologist, and not of the angiologist. The right coronary artery mainly supplies the right heart, while the left coronary artery, which bifurcates into the left anterior descending (LAD) and left circumflex (LCX) branch, mainly supplies the left heart.

As they have to perfuse the cardiac muscle, the coronaries protrude the ventricular wall, which has a profound effect on coronary hemodynamics (34,35). Upon ventricular contraction, blood flow in the coronaries is impeded, leading to a typical biphasic flow pattern, with systolic flow impediment, and predominantly flow during the diastolic phase (Fig. 4). This pattern is most obvious in the LAD, which supplies oxygenized blood to the left ventricle. The resistance of the coronary arteries is thus not constant in time, and contains an active component. When coronary flow is plotted as a function of coronary pressure, other typical features for the coronary circulation are observed. Under normal conditions, coronary flow is highly regulated, so that blood flow is constant for a wide range of coronary perfusion pressures (35,36). The level up to which the flow is regulated is a function of the activity of the ventricle, and hence of the metabolic demands. This seems to suggest that at least two different mechanisms are involved: metabolic autoregulation (flow is determined by the metabolic demand) and myogenic autoregulation. Myogenic autoregulation is the response of a muscular vessel on an increase in pressure: the vessel contracts, reducing its diameter and increasing its wall thickness, which tends to normalize the wall stress. It is only after maximal dilatation of the coronary vessels [e.g., through infusion of vasodilating pharmacological substances such as adenosine or papaverine, or immediately following a period of oxygen deficiency (ischemia)] that autoregulation

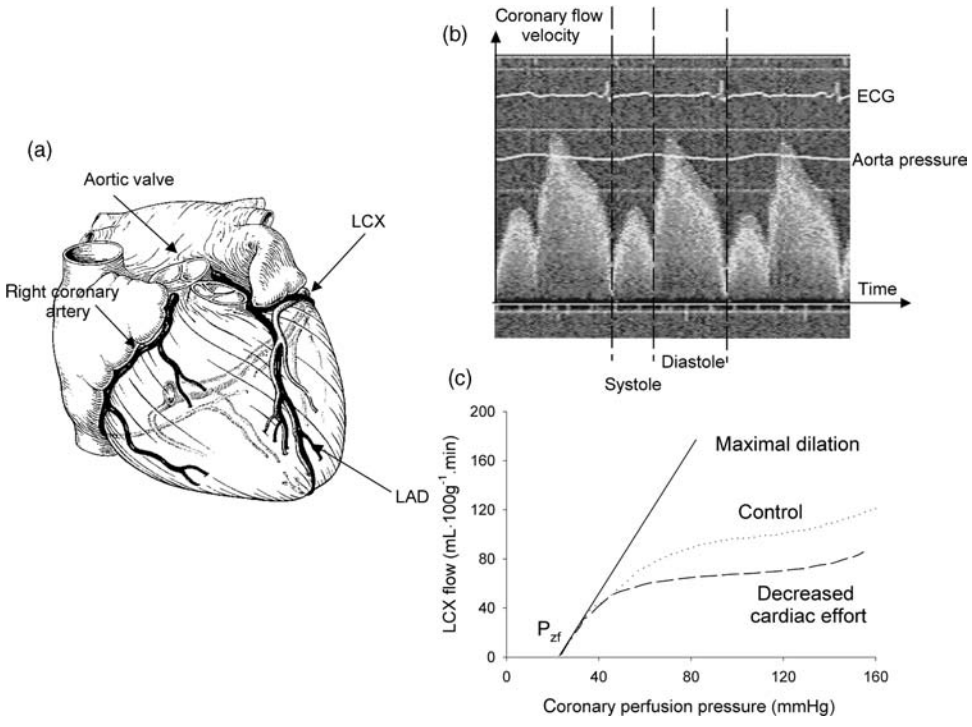


Figure 4. (a) Anatomical situation of the coronary arteries. (b) Demonstration of flow impediment during systole. (c) Coronary flow as a function of perfusion pressure, demonstrating the aspect of autoregulation and the non-zero intercept with the pressure axis (P_{zf}, zero-flow pressure). (Reconstructed after Ref. 36.)

can be "switched off", and that the pressure-flow relation becomes linear (Fig. 4). Note, however, that the pressure-flow relation does not pass through the origin: it requires a certain pressure value (zero-flow pressure P_{zf}) to generate any flow, an observation first reported by Bellamy et al. (37). The origin of P_{zf} is still not fully understood and has been attributed to (partial) vessel collapse by vascular tone and extravascular compression. This is a conceptual model, also known as the "waterfall" model, as flow is determined by the pressure difference between inflow and surrounding pressure (instead of outflow pressure) and pressure changes distal to the point of collapse (the waterfall) have no influence on flow (38,39). Note, however, that P_{zf} is often an extrapolation of pressure-flow data measured in a higher pressure range. It is plausible that the relation becomes nonlinear in the lower perfusion pressure range, due to the distensibility of the coronary vessels and compression of vessels due to intramyocardial volume shifts (35). Another model, based on the concept of an intramyocardial pump and capable of explaining phasic patterns of pressure and flow waves, was developed by Spaan et al. (35,40), and is mentioned here for reasons of completeness.

An important clinical aspect of the coronary circulation, which we will only briefly touch here, is the assessment of the severity of coronary artery stenosis (34). The stenosis forms an obstruction to flow, causes an extra pressure drop, and may result in coronary perfusion pressures too low to provide oxygen to the myocardial tissues perfused by that coronary artery. Imaging of the coronary vessels and the stenosis with angiography (in the cath-lab) is still an important tool for clinical decision making, but more and more attention has been attributed to quantification of functional severity of the stenosis. One of the most known indicies in use are the coronary flow reserve (CFR), that is, the ratio of maximal blood flow (velocity) through a coronary artery (after induction of maximal vasodilation) and baseline blood flow (velocity), with values >2 indicating sufficient reserve and thus a subcritical stenosis. Coronary flow reserve has the drawback that flow or velocity measurements are required, which are not common in the cath-lab. From that perspective, the fractional flow reserve (FFR) is more attractive, as it requires only pressure measurements. It can be shown that $\mathrm{FFR} = P_d/P_a$ (with P_d the pressure distal to the stenosis and P_a the aorta pressure) is the ratio of actual flow through the coronary, and the hypothetical flow that would pass through the coronary in the absence of the stenosis (34,41). A FFR value >0.75 indicates nonsignificant stenosis (34). Measurement of FFR is done in conditions of maximal vasodilation, and requires ultrathin pressure catheters (pressure wires) that can pass through the stenosis without causing (too much) extra pressure drop. Other indicies combine pressure and flow velocity (42) and are, from fluid dynamic perspective, probably the best characterization of the extra resistance created by the stenosis. It deserves to be mentioned that intravascular ultrasound (IVUS) is, currently, increasingly being applied, especially when treating complex coronary lesions. In addition to quantifying stenosis severity, much research is also focused on assessing the histology of the lesions and their vulnerability and risk of rupture via IVUS or other techniques.

THE ARTERIAL SYSTEM: DAMPING RESERVOIR AND/OR A BLOOD DISTRIBUTING NETWORK...

Basically, there are two ways of approaching the arterial system (43): (1) one can look at it in a "lumped" way, where abstraction is being made of the fact that the vasculature is a network system with properties distributed in space; or (2) one can take into account the network topology, and analyze the system in terms of pressure and flow waves propagating along the arteries in a forward and backward direction.

Irrespective of the conceptual framework within which one works, the analysis of the arterial system requires (simultaneously measured) pressure and flow, preferably measured at the upstream end of the arterial tree (i.e., immediately distal to the aortic valve). The analysis of the arterial system is largely analogous to the analysis of electrical network systems, where pressure is equivalent to voltage and flow to current. Also steming from this analogy is the fact that the arterial system is often analyzed in terms of impedance. Therefore, before continuing, the concept of impedance is introduced.

Impedance Analysis

While electrical waves are sine waves, with a zero time-average value, arterial pressure and flow waves are (approximately) periodical, but certainly nonsinusoidal, and their average value is different from zero (in humans, mean systemic arterial blood pressure is ~ 100 mmHg (13.3 kPa), while mean flow is ~ 100 mL·s^{-1}). To bypass this limitation, one can use the Fourier theorem, which states that any periodic signal, such as arterial pressure and flow, can be decomposed into a constant (the mean value of the signal) and a series of sinusoidal waves (harmonics). The frequency of the first harmonic is cardiac frequency (the fundamental frequency), while the frequency of the nth harmonic is n times the fundamental frequency.

Fourier decomposition is applicable if two conditions are fulfilled: (1) the cardiovascular system operates in steady-state conditions (constant heart rate; no respiratory effects); (2) the mechanical properties of the arterial system are sufficiently linear so that the superposition principle applies, meaning that the individual sine waves do not interact and that the sum of the effects of individual harmonics (e.g., the flow generated by a pressure harmonic) is equal to the effect caused by the original wave that is the sum of all individual harmonics.

Harmonics can be represented using a complex formalism. For the nth harmonic, the pressure (P) and flow (Q) component can be written as

$$P_n = |P_n| e^{i(n\omega t + \Phi_{P_n})} \qquad Q_n = |Q_n| e^{i(n_{\omega t} + \Phi_{Q_n})}$$

where $|P_n|$ and $|Q_n|$ are the amplitudes (or modul) of the pressure and flow sine waves, having phase angles Φ_{P_n} and Φ_{Q_n} (to allow for a phase lag in the harmonic), respectively. Time is indicated by t, and ω is the fundamental angular frequency, given by $2\pi/T$ with T the duration of a heart cycle (RR-interval). For a heart rate of 75 beats · min^{-1}, T is 0.8 s. The fundamental frequency is 1.25 Hz, and ω becomes 7.85 rad · s^{-1}.

In general, 10–15 harmonics are sufficient to describe hemodynamic variables, such as pressure, flow, or volume

(44,45). Also, to avoid aliasing, the sampling frequency should be twice as high as the frequency of the signal one is measuring (Nyquist limit). Thus, when measuring hemodynamic data in humans, the frequency response of the equipment should be $> 2 \times 15 \times 1.25$ Hz, or > 37.5 Hz. These requirements are commonly met by Hi–Fi pressure tip catheters (e.g., Millar catheters) with a measuring sensor embedded within the tip of the sensor, but not by the fluid-filled measuring systems that are frequently used in the clinical setting. Here, the measuring sensor is outside the body (directly or via extra fluid lines) connected to a catheter. Although the frequency response of the sensor itself is often adequate, the pressure signal is being distorted by the transmission via the catheter (and fluid lines and eventual connector pieces and three-way valves). The use of short, rigid, and large bore catheters is recommended, but it is advised to assess the actual frequency response of the system (as it is applied *in vivo*) if the pressure data is being used for purposes other than patient monitoring. Note that in small rodents like the mouse, where heart rate is as high as 600 beats min^{-1}, the fundamental frequency is 10 Hz, posing much higher measuring equipment requirements with a frequency response flat up to 300 Hz.

Impedance Z is generally defined as the ratio of pressure and flow: $Z = P/Q$, and thus has the dimensions of $\text{mmHg}\cdot\text{mL}^{-1}$ s [$\text{kg}\cdot\text{m}^{-4}\cdot\text{s}^{-1}$ in SI units; $\text{dyn}\cdot\text{cm}^{-5}\cdot$ s in older units] if pressure is expressed in mmHg and flow in $\text{mL}\cdot\text{s}^{-1}$. The parameter Z is usually calculated for each individual harmonic, and displayed as a function of frequency. Since both P and Q are complex numbers, Z is complex as well, and also has a modulus and a phase angle, except for the steady (dc) component at 0 Hz, which is nothing but the ratio of mean pressure and mean flow (i.e., the value of vascular resistance). For all higher harmonics, the modulus of the nth harmonic is given as $|Z_n| = |P_n|/|Q_n|$, and its phase, Φ_z, is given as $\Phi_{Pn} - \Phi_{Q_n}$.

Arterial impedance requires simultaneous measurement of pressure and flow at the same location. Although it can be calculated all over the arterial tree, it is most commonly measured at the entrance of the systemic or pulmonary circulation, and is called 'input' impedance, often denoted as Z_{in}. The parameter Z_{in} fully captures the relation between pressure and flow, and is determined by all downstream factors influencing this relation (arterial network topology, branching patterns, stiffness of the vessel, vasomotor tone, ...). In a way, it is a powerful description of the arterial circulation, since it captures all effects, but this is at the same time its greatest weakness, as it is not very sensitive to local changes in arterial system properties, such as focal atherosclerotic lesions.

The interpretation of (input) impedance is facilitated by studying the impedance of basic electrical or mechanical "building blocks" (43–45): (1) When a system behaves strictly resistive, there is a linear relation between the pressure (difference) and flow. Pressure and flow are always in phase, Z_n is a real number and Φ_z is zero. (2) In the case where there is only inertia in a system, pressure is ahead of flow; for sine waves, the phase difference between both is a quarter of a wavelength, or $+90^\circ$ in terms of phase angle. (3) In the case where the system behaves like a capacitor, flow is leading pressure, again 90° out of phase, so that Φ_z is -90°.

Figure 5 displays the input impedance of these fundamental building blocks, as well as Z_{in} calculated from the

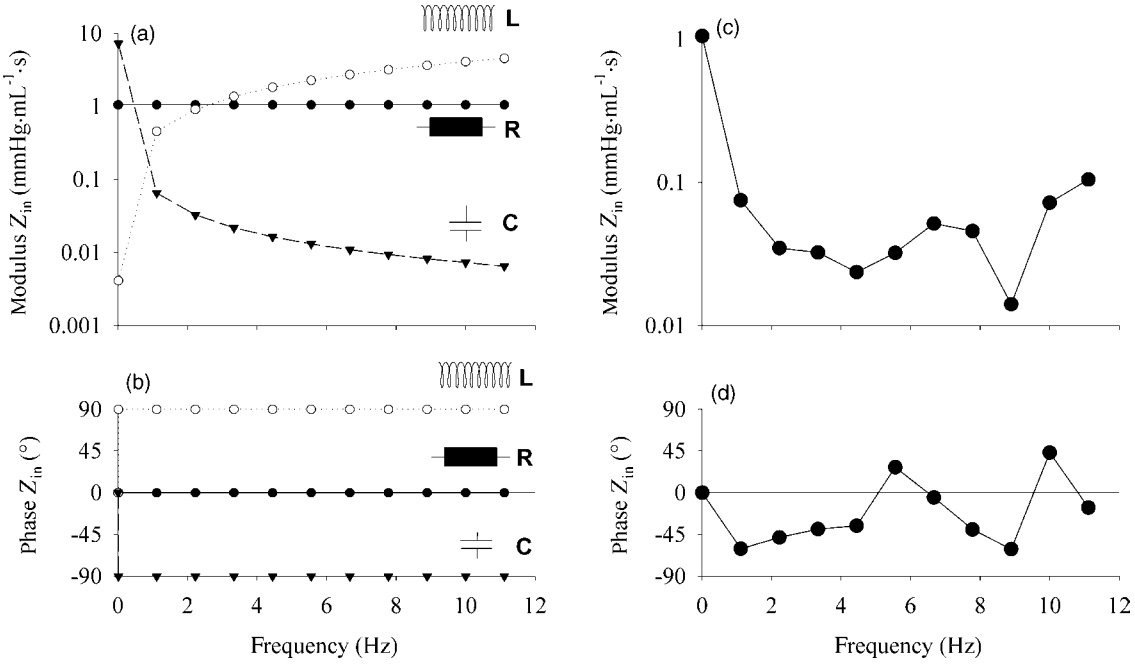


Figure 5. (a and b) Impedance modulus (a) and phase angle (b) of fundamental electrical/mechanical building blocks of the arterial system: resistance (R; 1.028 $\text{mmHg}\cdot\text{mL}^{-1}\cdot\text{s}$ (137.0×10^6 $\text{Pa}\cdot\text{m}^{-3}\cdot\text{s}$)), inertia [$L$; 0.065 $\text{mmHg}\cdot\text{mL}^{-1}\cdot\text{s}^2$ (8.7×10^6 $\text{Pa}\cdot\text{m}^{-3}\cdot\text{s}^2$)] and compliance [$C$; 2.25 $\text{mL}\cdot\text{mmHg}^{-1}$ (16.9×10^{-9} $\text{m}^3\cdot\text{Pa}^{-1}$)]. (c and d) input impedance modulus (c) and phase (d) calculated from aortic pressure and flow given in Fig. 1.

aorta pressure and flow shown in Fig. 1. The impedance modulus drops from the value of total systemic vascular resistance at 0 Hz to much lower values at higher frequencies. The phase angle is negative up until the fourth harmonic, showing that capacitive effects dominate at these low frequencies, although the phase angle never reaches -90°, indicating that inertial effects are present as well. For higher harmonics, the phase angle is close to zero, or at least oscillating around the zero value. At the same time, the modulus of Z_{in} is nearly constant. For these high frequencies, the system seems to act as a pure resistance, and the impedance value, averaged over the higher harmonics, has been termed the characteristic impedance (Z_0).

The Arterial System as a "Windkessel" Model

The most simple approximation of the arterial system is based on observations of reverend Stephen Hales (1733), who drew the parallel between the heart ejecting in the arterial tree, and the working principle of a fire hose (46). The pulsatile action of the pump is damped and transformed into a quasicontinuous outflow at the downstream end of the system, that is, the outflow nozzle for the fire hose and the capillaries for the cardiovascular system. In mechanical terms, this type of system behavior can be simulated with two mechanical components: a buffer reservoir (compliant system) and a downstream resistance. In 1899, Otto Frank translated this into a mathematical formulation (47,48), and it was Frank who introduced the terminology "windkessel models", windkessel being the German word for air chamber, as the buffer chamber used in the historical fire hose. The windkessel models are often described as their electrical analogue (Fig. 6).

The Two-Element Windkessel Model. While there are many different "windkessel" models in use (49,50), the two basic components contained within each model are a compliance element, C ($mL \cdot mmHg^{-1}$ or $m^3 \cdot Pa^{-1}$), and a resistor element, R ($mmHg \cdot mL^{-1} \cdot s$ or $Pa \cdot m^{-3} \cdot s$ in SI units). The compliance element represents the volume change associated with a unit change in pressure; R is the pressure drop over the resistor associated with a unit flow. In diastole, when there is no new inflow of blood into the compliance, the arterial pressure decays exponentially following $P(t) = P_0 e^{-t/RC}$. The parameter RC is the product

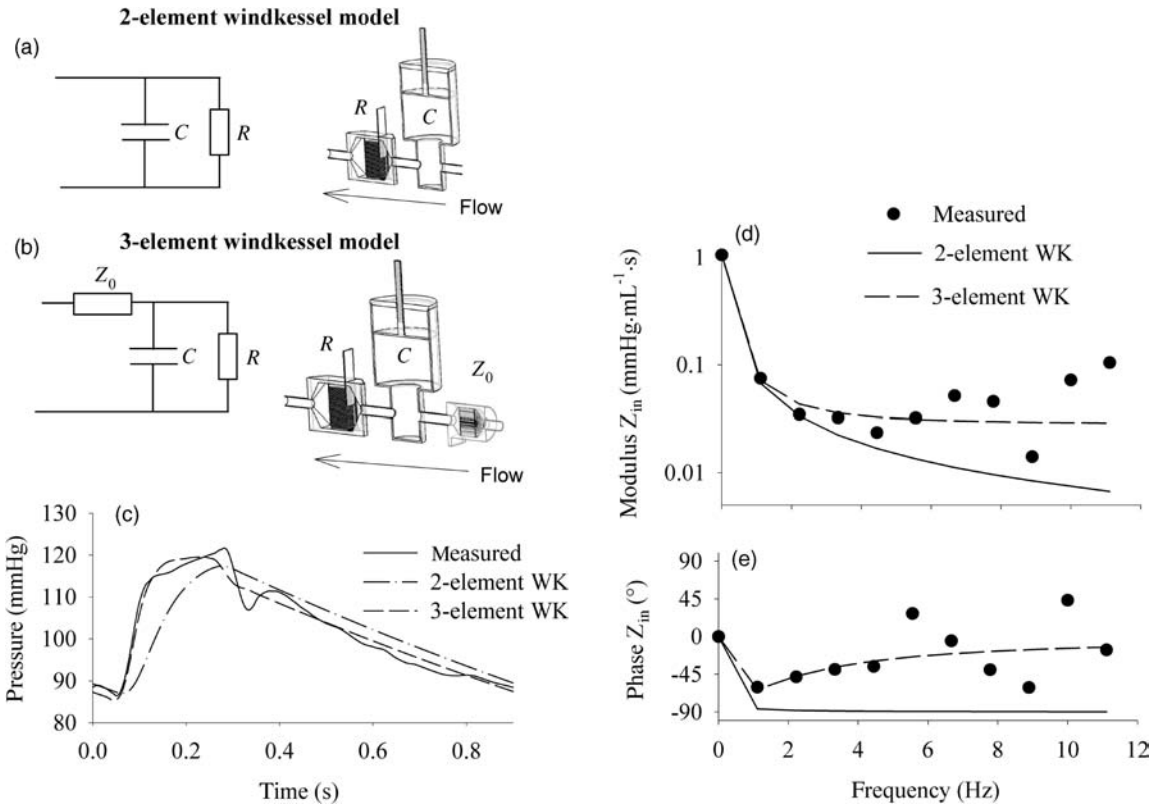


Figure 6. (a and b) Electrical analog and mechanical representation of a two- and three-element windkessel model. (c) Agreement between measured pressure, and the pressure obtained after fitting a two- and three-element windkessel model to the pressure (1 mmHg = 133.3 Pa) and flow data from Fig. 1. Model parameters are $R = 1.056$ $mmHg \cdot mL^{-1} \cdot s$ (140.8×10^6 $Pa \cdot m^{-3} \cdot s$) and $C = 2.13$ $mL \cdot mmHg^{-1}$ (16.0×10^{-9} $m^3 \cdot Pa^{-1}$) for the two-element windkessel model; $R = 1.028$ mm $Hg \cdot mL^{-1} \cdot s$ (137.0×10^6 $Pa \cdot m^{-3} \cdot s$). $C = 2.25$ $mL \cdot mmHg^{-1}$ (16.9×10^{-9} $m^3 \cdot Pa^{-1}$) and $Z_0 = 0.028$ $mmHg \cdot mL^{-1} \cdot s$ (3.7×10^6 $Pa \cdot m^{-3} \cdot s$) for the three-element windkessel model. (d and e) Input impedance modulus (d) and phase angle (e) of these lumped parameter models and their match to the *in vivo* measured input impedance (1 $mmHg \cdot mL^{-1} \cdot s = 133.3 \times 10^6$ $Pa \cdot m^{-3} \cdot s$).

of R and C and is called the arterial decay time. The higher the RC time, the slower the pressure decay. It is the time required to reduce P_0 to 37% of its initial value (note that the 37% is a theoretical value, usually not reached *in vivo* because the next beat impedes a full pressure decay). One can make use of this property to estimate the arterial compliance: By fitting an exponential curve to the diastolic decaying pressure, RC is obtained and thus, when R is known we also know C (49,51–53). This method is known as the decay time method.

For typical hemodynamic conditions in humans at rest (70 beats $\cdot$ min^{-1}, systolic/diastolic, and mean pressure of 120 (16.0 kPa), 80 (10.7 kPa), and 93 (12.4 kPa) mmHg respectively, stroke volume 80 mL), mean flow is 93 mL$\cdot$s^{-1} (0.93×10^{-4} m$^3\cdot$s^{-1}), and R is 1 mmHg (mL$\cdot$s^{-1}) (133.3×10^6 Pa$\cdot$m$^{-3}\cdot$s). Assuming that the whole stroke volume is buffered in systole, C can be estimated as the ratio of stroke volume and pulse pressure (systolic–diastolic pressure difference), ~ 2 mL$\cdot$mmHg^{-1} (1.50×10^{-8} m$^3\cdot$Pa^{-1}). This value is considered as an overestimation of the actual arterial compliance (49,54). In humans, RC time is thus of the order of 1.5–2 s.

The question, How well does a windkessel model represents the actual arterial system?, can be answered by studying the input impedance of both. In complex formulation, the input impedance of a two-element windkessel model is given as

$$Z_{i-\mathrm{WK2}} = \frac{R}{1 + i\omega RC}$$

with i the complex constant, and $\omega = 2\pi f$, f is the frequency. The dc value (0 Hz) of Z_{in} is thus R; at high frequencies, it becomes zero. The phase angle is 0 at 0 Hz, and -90° for all other frequencies. Compared to input impedance as measured in mammals, the behavior of a two-element windkessel model reasonably represents the behavior of the arterial system for the low frequencies (up to third harmonic), but not for higher frequencies (43,49,54) (Fig. 6). This means that it is justified to use the model for predicting the low frequency behavior of the arterial system, that is, the low frequency response to a flow input. This property is used in the so-called "pulse pressure method", an iterative method to estimate arterial compliance: with R assumed known, the pulse pressure response of the two-element windkessel model to a (measured) flow stimulus is calculated with varying values of C. The value of C yielding the pulse pressure response matching the one measured *in vivo*, is considered to be the correct one (55). Compared to the decay time method, the advantage is that the pulse pressure method is insensitive to deviations of the decaying pressure from the true exponential decay (53).

The Three-Element and Higher Order Windkessel Models. The major shortcoming of the two-element windkessel model is the inadequate high frequency behavior (43,49,56). Westerhof et al.resolved this problem by adding a third resistive element proximal to the windkessel, accounting for the resistive-*like* behavior of the arterial system in the high frequency range (56). The third element represents the characteristic impedance of the proximal part of the ascending aorta, and integrates the effects of inertia and compliance. Adding the element, the input impedance of the three-element windkessel model becomes

$$Z_{i-\mathrm{WK3}} = Z_0 + \frac{R}{1 + i\omega RC}$$

it making the phase angle negative for lower harmonics, but returns to zero for higher harmonics, where the impedance modulus asymptotically reaches the value of Z_0. For the systemic circulation, the ratio of Z_0 and R is 0.05–0.1 (57).

The major disadvantage of the three-element windkessel model is the fact that Z_0, which should represent the high frequency behavior of the arterial system, plays a role at all frequencies, including at 0 Hz. This has the effect that, when the three-element windkessel model is used to fit data measured in the arterial system, the compliance is systematically overestimated (53,58): the only way to "neutralize" the contribution of Z_0 at the low frequencies, is to artificially increase the compliance of the model. The "ideal" model would incorporate both the low frequency behavior of the two-element windkessel model, and the high frequency Z_0, though without interference of the latter at all frequencies. This can be achieved by adding an inertial element in parallel to the characteristic impedance, as demonstrated by Stergiopulos et al. (59), elaborating on a model first introduced by Burattini et al. (60). For the DC component and low frequencies, Z_0 is bypassed through the inertial element. For the high frequencies, Z_0 takes over. It has been demonstrated, fitting the four-element windkessel model to data generated using an extended arterial network model, that L effectively represents the total inertia present in the model (59).

Obviously, by adding more elements, it is possible to develop models that are able to further enhance the matching between model and arterial system behavior (49,50), but the uniqueness of the model may not be guaranteed, and the physiological interpretation of the model elements is not always clear.

Although lumped parameter models cannot explain all aspects of hemodynamics, they are very useful as a concise representation of the arterial system. Mechanical versions are frequently used in hydraulic bench experiments, or as highly controllable afterload systems for *in vivo* experiments. An important field of application of the mathematical version is for parameter identification purposes: fitting arterial pressure and flow data measured *in vivo* to these models, the arterial system can be characterized and quantified (e.g., the total arterial compliance) through the model parameter values (43,49).

It is important to stress that these lumped models represent the behavior of the arterial system as a whole, and that there is no relation between model components and anatomical parts of the arterial tree (43). For example, although Z_0 represents the properties of the proximal aorta, there is no drop in mean pressure along the aorta, which one would expect if the three-element windkessel model were to be interpreted in a strict anatomical way. The combination of elements simply yields a model that represents the behavior of the arterial system as it is seen by the heart.

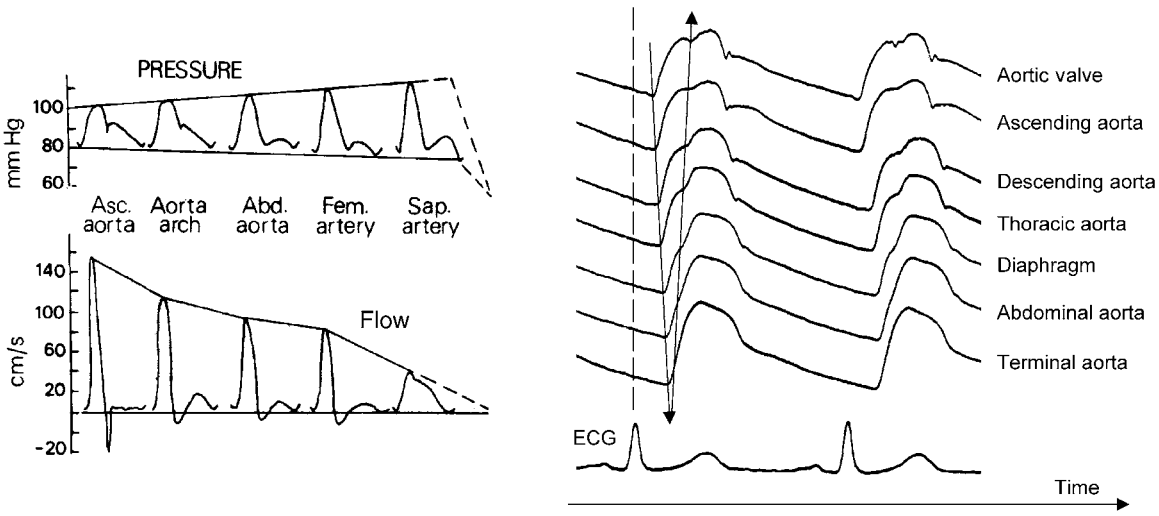


Figure 7. (a) Shows both pressure (1 mmHg = 133.3 Pa) and flow velocity measured along the arterial tree in a dog [after Ref. (44).] (b) Pressure wave forms measured between the aortic valve (AoV) and the terminal aorta (Term Ao) in one patient. [Modified from Ref. 61.]

Wave Propagation and Reflection in the Arterial Tree

As already stressed, lumped parameter models simply represent the behavior of the arterial system as a whole, and as seen by the heart. As soon as the model is subjected to flow or pressure, there is an instantaneous effect throughout the whole model. This is not the case in the arterial tree: when measuring pressure along the aorta, it can be observed that there is a finite time delay between the onset of pressure rise and flow in the ascending and distal aorta (Fig. 7). The pressure pulse travels with a given speed from the heart toward the periphery.

When carefully analyzing pressure wave profiles measured along the aorta, several observations can be made: (*1*) there is a gradual increase in the steepness of the wave front; (*2*) there is an increase in peak systolic pressure (at least in the large-to-middle sized arteries); (*3*) diastolic blood pressure is nearly constant, and the drop in mean blood pressure (due to viscous losses) is negligible in the large arteries. The flow (or velocity) wave profiles exhibit the same time delay in between measuring locations, but their amplitude decreases. Also, by comparing the pressure and flow wave morphology, one can observe that these are fundamentally different.

In the absence of wave reflection (assuming the aorta to be a uniform, infinitely long elastic tube), dissipation would only lead to less steep wave fronts, and damping of maximal pressure. Also, in these conditions, one would expect similarity of pressure and flow wave morphology. Thus, the above observations can only be explained by wave reflection. This is, of course, not surprising given the complex anatomical structure of the arterial tree, with geometric and elastic tapering (the further away from the heart, the stiffer the vessel), its numerous bifurcations, and the arterioles and capillaries making the distal terminations.

The Arterial Tree as a Tube. Despite the complexity described above, arterial wave reflection is often approached in a simple way, conceptually considering the arterial tree as a single tube [or T-tube (62)], with one (or 2) discrete reflection site(s) at some distance from the heart. Within the single tube concept, the arterial tree is seen as a uniform or tapered (visco-)elastic tube (63,64), with an "effective length" (65,66), and a single distal reflection site. The input impedance of such a system can be calculated, as demonstrated in Fig. 8 for a uniform tube of length 50 cm, diameter 1.5 cm, Z_0 of 0.275 mmHg·mL^{-1}·s (36.7×10^6 Pa·m^{-3}·s) and wave propagation speed of 6.2 m·s^{-1}. The tube is ended

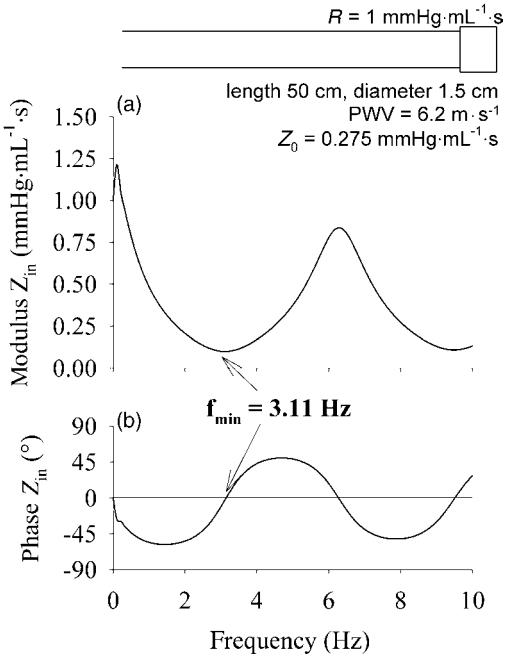


Figure 8. Input impedance modulus and phase of a uniform tube with length 50 cm, diameter 1.5 cm, characteristic impedance of 0.275 mmHg·mL^{-1}·s (36.7×10^6 Pa·m^{-3}·s) and wave propagation speed of 6.2 m·s^{-1}. The tube is ended by a (linear) resistance of 1 mmHg·mL^{-1}·s (133.3×10^6 Pa·m^{-3}·s).

by a (linear) resistance of 1 mmHg·mL^{-1}·s (133.3×10^6 Pa·m^{-3}·s). The impedance mismatch between the tube and its terminal resistance gives rise to wave reflections and oscillations in the input impedance pattern, and many features, observed *in vivo* (67), can be explained on the basis of this model.

Assume there is a sinusoidal wave running in the system with a wavelength λ being four times the length of the tube. This means that the phase angle of the reflected wave, when arriving back at the entrance at the tube, will be 180° out of phase with respect to the forward wave. The sum of the incident and reflected wave will be zero, since they interfere in a maximally destructive way. If this wave is a pressure wave, measured pressure at the entrance of the tube will be minimal for waves with this particular wave length. There is a relation between pulse wave velocity (PWV), λ, and frequency f (i.e., $\lambda = \mathrm{PWV}/f$). Thus, in an input impedance spectrum, the frequency $f_{\min}$, where input impedance is minimal, corresponds to a wave with a wavelength that is equal to four times the distance to the reflection site, L,: $4L = \mathrm{PWV}/f_{\min}$, or $L = \mathrm{PWV}/4f_{\min}$ and $f_{\min} = \mathrm{PWV}/4L$. Applied to the example of the tube, $f_{\min}$ is expected at $6.2/2 = 3.1$ Hz. This equation is known as the "quarter wavelength" formula, and is used to estimate the effective length of the arterial system.

Although the wave reflection pattern in the arterial system is complex (68), pressure (P) and flow (Q) are mostly considered to be composed of only one forward running component, P_f (Q_f) and one backward running component, P_b (Q_b), where the single forward and backward running components are the resultant of all forward and backward traveling waves, including the forward waves that result from rereflection at the aortic valve of backward running waves (69).

At all times,

$$P = P_f + P_b \quad \text{and} \quad Q = Q_f + Q_b$$

Furthermore, if the arterial tree is considered as a tube, defined by its characteristic impedance Z_0, the following relations also apply:

$$Z_0 = P_f/Q_f = -P_b/Q_b$$

since Z_0 is the ratio of pressure and flow in the absence of wave reflection (43–45), which is the case when only forward or backward running components are taken into consideration. The negative sign in the equation above appears because the flow is directional, considered positive in the direction away from the heart, and negative toward the heart, while the value of the pressure is insensitive to direction.

Combining these equations, $P = P_f - Z_0 Q_b = P_f - Z_0 (Q - Q_f) = 2P_f - Z_0 Q$, so that

$$P_f = (P + Z_0 Q)/2$$

Similarly, it can be deduced that

$$P_b = (P - Z_0 Q)/2$$

These equations were first derived by Westerhof et al., and are known as the linear wave separation equations (67). In principle, the separation should be calculated on individual harmonics, and the net P_f and P_b wave then follows from summation of all forward and backward harmonics. In practice, however, the equations are often used in the time domain, using measured pressure and flow as input. Note, however, that wave reflection only applies to the pulsatile part of pressure and flow, and mean pressure and flow should be subtracted from measured pressure and flow before applying the equations. An example of wave separation is given in Fig. 9.

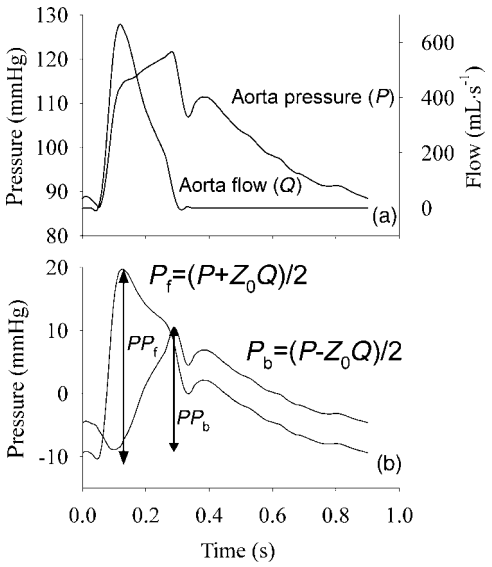


Figure 9. Application of linear wave separation analysis to the pressure (1 mmHg = 133.3 Pa) and flow data of Fig. 1. The ratio of PP_b and PP_f can be used as a measure of wave reflection magnitude.

Note also that wave separation requires knowledge of characteristic impedance, which can be estimated both in the frequency and time domain (70). When discussing input impedance, it was already noted that for the higher harmonics (> fifth harmonic), Z_{in} fluctuates around a constant value, Z_0, and with a phase angle approaching zero. Assuming wave speed to be $\sim 5\ \mathrm{m \cdot s^{-1}}$, the wave length λ for these higher harmonics (e.g., the fifth harmonic for a heart rate of 60 beats · min^{-1}), being the product of the wave speed and wave period (0.2 s) becomes shorter (1 m) than the average arterial pathlength. Waves reflect at distant locations throughout the arterial tree (with distal ends of vascular beds < 50 cm to ~2 m away from the heart in humans), return back to the heart with different phase angles and destructively interfere with each other, so that the net effect of the reflected waves appears inexistent. For these higher harmonics, the arterial system thus appears reflectionless, and under these conditions, the ratio of pressure and flow is, by definition, the characteristic impedance. Therefore, averaging the input impedance modulus of the higher harmonics, where the phase angle is about zero, yields an estimate of the characteristic impedance (43–45).

Characteristic impedance can also be estimated in the time domain. In early systole, the reflected waves did not yet reach the ascending aorta, and in the early systolic ejection period, the relation between pressure and flow is

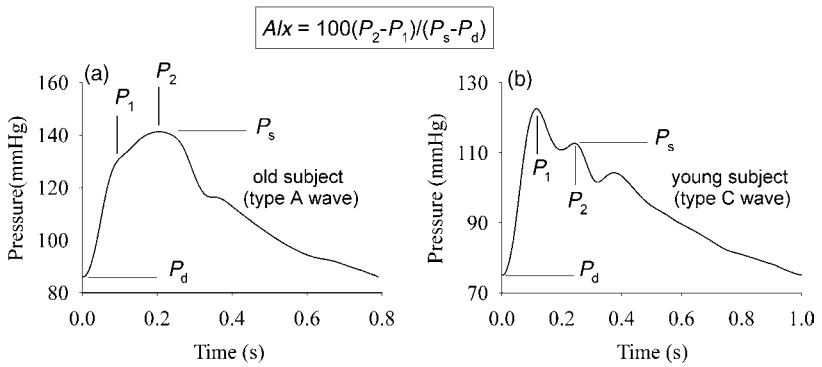


Figure 10. Typical pressure wave contours measured noninvasively at the common carotid artery with applanation tonometry in a young and old subject, and definition of the "augmentation index" (AIx) (1 mmHg = 133.3 Pa). See text for details.

linear, as can be observed when plotting P as a function of Q (70,71). The slope of the Q–P relationship during early systole is the time domain estimate of Z_0, and is in good agreement with the frequency domain estimate (70).

Both the time and frequency domain approach, however, are sensitive to subjective criteria, such as the selection of the early systolic ejection period, or the selection of the harmonic range that is used for averaging.

Pathophysiological Consequences of Arterial Wave Reflection. Figure 10 displays representative carotid artery pressure waves (∼ aorta pressure) for a young subject, (b), and for an older, healthy subject, (a). It is directly observed that the morphology of the pressure wave is fundamentally different. In the young subject, pressure first reaches its maximal systolic pressure, and an "inflection point" is visible in late systole, generating a late shoulder (P_2). In the older subject, this inflection point appears in early systole, generating an early shoulder (P_1).

Measurements of pressure along different locations in the aorta, conducted by Latham et al. (67), showed that, when the foot of the wave on one hand, and the inflection point on the other, are interconnected (Fig. 7): (1) these points seem to be aligned on two lines; (2) the lines connecting these characteristic marks intersect. This pattern is consistent with the concept of a pressure wave being generated by the heart, traveling down the aorta, reflect, and superimpose on the forward going wave. The inflection point is then a visual landmark, designating the moment in time where the backward wave becomes dominant over the forward wave (61).

In young subjects, the arteries are most elastic (deformable), and pulse wave velocity (see below) is much lower than in older subjects, or in patients with hypertension or diabetes. In the young, it takes more time for the forward wave to travel to the reflection site, and for the reflected wave to arrive at the ascending aorta. Both interact only in late systole (late systolic inflection point), causing little extra load on the heart. In older subjects, on the other hand, PWV is much higher, and the inflection point shifts to early systole. The early arrival of the reflected wave literally boosts systolic pressure, causing pressure augmentation, and augmenting the load on the heart (72). At the same time, the early return of the reflected wave impedes ventricular ejection, and thus may have a negative impact on stroke volume (72,73). An increased load on the heart increases the energetic cost to maintain stroke volume, and will initiate cardiac compensatory mechanisms (remodeling), which may progress into cardiac pathology (74–76).

Wave Intensity Analysis. With its origin in electrical network theory, much of the arterial function analysis, including wave reflection, is done in the frequency domain. Besides the fact that this analysis is, strictly speaking, only applicable in linear systems with periodic signal changes, the analysis is quite complex due to the necessity of Fourier decomposition, and it is not intuitively comprehensible. An alternative method of analysis, performed in the time domain and not requiring linearity and periodicity is the analysis of the wave intensity, elaborated by Parker and Jones in the late 1980s (77).

Disturbances to the flow lead to changes in pressure (dP) and flow velocity (dU), "wavelets", which propagate along the vessels with a wave speed (PWV), as defined above. By accounting for conservation of mass and momentum, it can be shown that

$$dP_{\pm} = \pm \mathrm{PWV} \rho dU_{\pm}$$

where the "+" denotes a forward traveling wave (for a defined positive direction), while "−" denotes a backward traveling wave. This equation is also known as the water-hammer equation. Waves characterized by a $dP > 0$, that is, a rise in pressure, are called compression waves, while waves with $dP < 0$ are expansion waves. Note that this terminology still reflects the origin of the theory in gas dynamics.

Basically, considering a tube, with left-to-right as the positive direction, there are four possible types of waves: (*1*) Blowing on the left side of the tube, pressure rises ($dP > 0$), and velocity increases ($dU > 0$). This is a *forward compression wave*. (*2*) Blowing on the right side of the tube, pressure increases ($dP > 0$), but velocity decreases ($dU < 0$) with our convention. This is a *backward compression wave*. (*3*) Sucking on the left side of the tube, pressure decreases ($dP < 0$), as well as the velocity ($dU < 0$), but the wavefront propagates from left to right. This wave type is a *forward expansion wave*. (*4*) Finally, one can also suck on the right side of the tube, causing a decrease in pressure ($dP < 0$) but an increase in velocity ($dU > 0$). This is a *backward expansion wave*.

The nature of a wave is most easily comprehended by analysing the wave intensity, dI, which is defined as the product of dU and dP, and is the energy flux carried by the

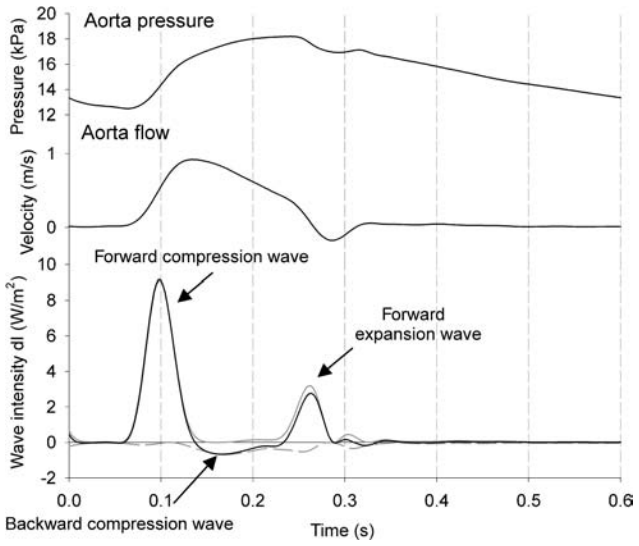


Figure 11. The concept of wave intensity analysis, applied to pressure and flow measured in the ascending aorta of a dog.

wavelet. It can be deduced from the above that dI is always positive for forward running waves, and always negative for a backward wave. When dI is positive, forward waves are dominant; otherwise, backward waves are dominating. Analysis of dP reveals whether the wave is a compression or an expansion wave. Figure 11 shows the wave intensity calculated from pressure and flow velocity measured in the ascending aorta of a dog (71). A typical wave intensity pattern is characterized by three major peaks. The first one is a forward compression wave, associated with the ejection of blood from the ventricle. The second positive peak is associated with a forward running wave, but $dP < 0$, and this second peak is thus a forward running expansion wave, due to ventricular relaxation, slowing down the ejection from the heart. During systole, reflected waves are dominant, resulting in a negative wave intensity, but with, in this case, positive dP. The negative peak is thus a backward compression wave, resulting from the peripheral wave reflections.

Further, note that similar to Westerhof's work (69), the wavelets dP and dU also can be decomposed in the forward and backward components (71). It can be derived that

$$dP_{\pm} = \frac{1}{2}(dP \pm \rho \mathrm{PWV}\, dU) \quad \text{and} \quad dU_{\pm} = \pm\frac{1}{2}\left(\frac{dP}{\rho \mathrm{PWV}} \pm dU\right)$$

The total forward and backward pressure and flow wave can be obtained as

$$P_{+} = P_{\mathrm{d}} + \sum_{t=0}^{t} dP_{+}$$

with P_{d} the diastolic blood pressure, which is added to the forward wave, and $P_{-} = \sum_{t=0}^{t} dP_{-}$ (71). Similarly, for the forward and backward velocity wave, it applies that

$$U_{+} = \sum_{t=0}^{t} dU_{+} \quad \text{and} \quad U_{-} = \sum_{t=0}^{t} dU_{-}$$

Wave Intensity in itself, dI, can also be separated in a net forward and backward wave intensity: $dI_{+} = dP_{+}dU_{+}$ and $dI_{-} = dP_{-}dU_{-}$ with $dI = dI_{+} + dI_{-}$.

Wave intensity is certainly an appealing method to gain insight into complex wave (reflection) patterns, as in the arterial system, and the use of the method is growing (78–80). The drawback of the method is the fact that dI is calculated as the product of two derivatives dP and dU, and thus is highly sensitive to noise in the signal. Adequate filtering of basic signals and derivatives is mandatory. As for the more "classic" impedance analysis, it is also required that pressure and flow be measured at the exact same location, and preferably at the same time.

Assessing Arterial Hemodynamics in Real Life

Analyzing Wave Propagation and Reflection. The easiest approach of analyzing wave reflection is to simply quantify the global effect of wave reflection on the pressure wave morphology. This can be done by formally quantifying the observation of Murgo et al. (61), and is now commonly known as the augmentation index (AIx). This index was first defined in the late 1980s by Kelly et al. (81). Although different formulations are used, AIx is nowadays most commonly defined as the ratio of the difference between the "secondary" and "primary peak", and pulse pressure: $\mathrm{AIx} = 100\ (P_2 - P_1)/(P_{\mathrm{s}} - P_{\mathrm{d}})$, and expressed as a percentage (Fig. 10). For A-type waves (Fig. 10) with a shoulder preceding systolic pressure, P_1 is a pressure value characteristic for the shoulder, while P_2 is systolic pressure. These values are positive. For C-type waves, the shoulder follows systolic pressure, and P_1 is systolic pressure, while P_2 is a pressure value characteristic for the shoulder, thus yielding negative values for AIx (Fig. 10). There are also alternative formulations in use, for example, $100(P_2 - P_{\mathrm{d}})/(P_1 - P_{\mathrm{d}})$, which always yields a positive value ($< 100\%$ for C-type waves, and $> 100\%$ for A-type waves). Both are, of course, mathematically related to each other. Since AIx is an index based on pressure differences and ratios, it can be calculated from noncalibrated pressure waveforms, which can be obtained noninvasively using techniques such as applanation tonometry (see below). In the past 5 years, numerous studies have been published using AIx as a quantification of the contribution of wave reflection to arterial load.

It is, however, important to stress that AIx is an integrated measure and that its value depends on all factors influencing the magnitude and timing of the reflected wave: pulse wave velocity, magnitude of wave reflection, and the distance to the reflection site. This AIx is thus strongly dependent on body size (82). In women, the earlier return of the reflected wave leads to higher AIx. The relative timing of arrival of the reflected wave is also important, so that AIx is also dependent on heart rate (83). For higher heart rates, systolic ejection time shortens, so that the reflected wave arrives relatively later in the cycle, leading to an inverse relation between heart rate and AIx.

Instead of studying the net effect of wave reflection, one can also measure pressure and flow (at the same location) and separate the forward and backward wave using the

aforementioned formulas. An example is given in Fig. 9, using aorta pressure and flow from Fig. 1. The ratio of the amplitude of P_b (PP_b) and P_f (PP_f) then yields an easy measure of wave reflection, which can be considered as a wave reflection coefficient, although it is to be emphasized that it is not a reflection coefficient in a strict sense. It is not the reflection coefficient at the site of reflection, but at the upstream end of the arterial tree and thus also incorporates the effects of wave dissipation, playing a role along the pathlength for the forward and backward component.

Another important determinant of the augmentation index is pulse wave velocity, which is an interesting measure of arterial stiffness in itself. This is easily demonstrated via the theoretical Moens–Korteweg equation for a uniform 1D tube, stating that PWV $= \sqrt{Eh/\rho D}$ with E the Young elasticity modulus (400–1200 kPa for arteries), ρ the density of blood ($\sim 1060\ \mathrm{kg \cdot m^{-3}}$), h the wall thickness, and D the diameter of the vessel. Another formula, sometimes used, is the so-called Bramwell–Hill (84) equation: PWV $= \sqrt{A \partial P/\rho \partial A} = \sqrt{A/\rho\, 1/C_A}$ with P intra-arterial pressure, A cross-sectional area and C_A the area compliance, $\partial A/\partial P$. For unaltered vessel dimensions, an increase in stiffness (increase in E, decrease in C_A) yields an increase in PWV. The most common technique to estimate PWV is to measure the time delay between the passage of the pressure, flow, or diameter distension pulse at two distinct locations, for example, between the ascending aorta (or carotid artery) and the femoral artery (85). These signals can be measured noninvasively, for example, with tonometry (86,87) (pressure pulse) or ultrasound [flow velocity, vessel diameter distension (88,89)]. Reference work was done by Avolio et al., who measured PWV is large cohorts in Chinese urban and rural communities (90,91).

When the input impedance is known, which implies that pressure and flow are known, the "effective length" of the arterial system, that is, the distance between the location where the input impedance is measured and an "apparent" reflection site at a distance L, can be estimated using the earlier mentioned quarter wavelength formula. Murgo et al. found the effective length to be ~ 44 cm in humans (61), which would suggest a reflection site near the diaphragm, in the vicinity of where the renal arteries branch off the abdominal aorta. Latham et al. demonstrated that local reflection is higher at this site (67), but one should keep in mind that the quarter wavelength formula is based on a conceptual model of the arterial system, and the apparent length does not correspond to a physical obstruction causing the reflection. Nevertheless, there is still no clear picture of the major reflection sites, which is complex due to the continuous branching, the dispersed distal reflection sites, and the continuous reflection caused by the geometric and elastic tapering of the vessels (45,64,67,68,92–96).

Central and Peripheral Blood Pressure: On the Use of Transfer Functions. Wave travel and reflection generally result in an amplification of the pressure pulse from the aorta (central) toward the periphery (brachial, radial, femoral artery) (44,45). Since clinicians usually measure blood pressure at the brachial artery (cuff sphygmomanometry), this implies that the pressure measured at this location is an overestimation of central pressure (97,98). It is the latter against which the heart ejects in systole, and which is therefore of primary interest.

In the past few years, much attention has been attributed to the relation between radial artery and central pressure (99–102). The reason for this is that radial artery pressure pulse is measurable with applanation tonometry, a noninvasive method (86,87,103). The relation between central and radial artery pressure can be expressed with a "transfer function", most commonly displayed in the frequency domain (Fig. 12). The transfer function expresses

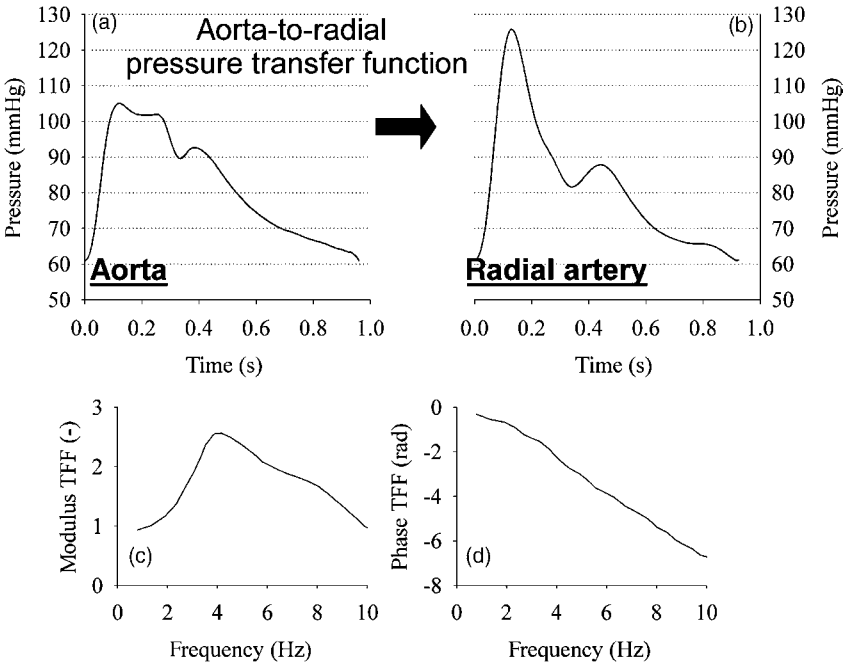


Figure 12. Demonstration of the aorta-to-radial pressure pulse amplification: (a and b) are carotid (as substitute for aorta pressure) and radial artery pressure measured in the same subject with applanation tonometry (1 mmHg = 133.3 Pa). (c and d) Display the modulus and phase angle of the radial-to-aorta transfer function as published by Ref. (99).

the relation between the individual harmonics at both locations, with a modulus (damping or amplification of the harmonic) and a phase angle (the time delay). In humans, the aorta-to-radial transfer function shows a peak ~ 4 Hz, so that amplification is maximal for a harmonic of that frequency (99,101). It has been demonstrated that the transfer function is surprisingly constant, with albeit little variation among individuals (99). It is this observation that led to the use of generalized transfer functions, embedded in commercial systems, that allow us to calculate central pressure from a peripheral pressure measurement (104). In these systems, the transfer is commonly done using a time domain approach [autoregressive exogenous (ARX) models (100)]. There have been attempts to "individualize" the transfer function, but until now these attempts were unsuccessful (105). The general consensus appears to be that the generalized transfer function can be used to estimate central systolic blood pressure and pulse pressure (104), but that one should be cautious when using synthesized central pressures for assessing parameters based on details in the pressure wave (e.g., the augmentation index) (106). The latter requires high frequency information that is more difficult to guarantee with a synthesized curve, both due to the increase in scatter in the generalized transfer function for higher frequencies, and the absence of high frequency information in the peripherally measured pressure pulse (99,107).

One more issue worth noting is the fact that the radial-to-aorta transfer function peaks at 4 Hz (in humans), which implies that the peripheral pulse amplification is frequency, and thus heart rate dependent (83,108,109). In a pressure signal, most power is embedded within the first two to three harmonics. When heart rate shifts from 60 to 120 beats/min (e.g., during exercise), the highest amplification thus occurs for these most powerful, predominant harmonics, leading to a more excessive pressure amplification. This means, conversely, that the overestimation of central pressure by a peripheral pressure measurement is a function of heart rate. As such, the effect of drugs that alter the heart rate (e.g., beta blockers, slowing down heart rate) may not be fully reflected by the traditional brachial sphygmomanometer pressure measurement (97,98).

Practical Considerations. The major difficulty, transferring experimental results into clinical practice, is the accurate measurement of the data necessary for the hemodynamic analysis. Nevertheless, there are noninvasive tools available that do permit "full" noninvasive hemodynamic assessment in clinical conditions, as possible with central pressure and flow.

Flow is at present rather easy to measure with ultrasound, using pulsed Doppler modalities. Flow velocities can be measured in the left ventricular outflow tract and, when multiplied with outflow tract cross-section, be converted into flow. Also, velocity-encoded MRI can provide aortic blood flow velocities.

As for measuring pressure, applanation tonometry is an appealing technique, since it allows us to measure pressure pulse tracings at superficial arteries such as the radial, brachial, femoral, and carotid artery (86,87,103,110,111). The carotid artery is located close to the heart, and is often used as a surrogate for the ascending aorta pulse contour (87,111). Others advocate the use of a transfer function to obtain central pressure from radial artery pressure measurement (104), but generalized transfer functions do not fully capture all details of a central pressure (106). Nevertheless, applanation tonometry only yields the morphology of the pressure wave and not absolute pressure values (although this is theoretically possible, but virtually impossible to achieve in practice). For the calibration of the tracings, one still relies on brachial cuff sphygmomanometry (yielding systolic and diastolic brachial blood pressure). Van Bortel et al. validated a calibration scheme in which first a brachial pressure waveform is calibrated with sphygmomanometer systolic and diastolic blood pressure (81,112). Averaging of this calibrated curve subsequently yields mean arterial blood pressure. Pulse waveforms at other peripheral locations are then calibrated using diastolic and mean blood pressure. Alternatively, one can use an oscillometric method to obtain mean arterial blood pressure, or estimate mean arterial blood pressure from systolic and diastolic blood pressure using the two-third/one-third rule of thumb (mean blood pressure $= \frac{2}{3}$ diastolic $+ \frac{1}{3}$ systolic pressure). With the oscillometric method, one makes use of the oscillations that can be measured within the cuff (and in the brachial artery) when the cuff pressure is lowered from a value above systolic pressure (full occlusion) to a value below diastolic pressure (no occlusion).

Nowadays, ultrasound vessel wall tracking techniques also allow us to measure the distension of the vessel upon the passage of the waveform (88,113). These waveforms are quasiidentical to the pressure waveforms [but not entirely, since the pressure–diameter relationship of blood vessels is not linear (114,115)] and can potentially be used as an alternative for tonometric waveforms (112,116).

When central pressure and flow are available, the arterial system can be characterized using all described techniques, from impedance analysis, parameter estimation by means of a lumped parameter windkessel or tube model, analysis of wave reflection, and so on. The augmentation index can be derived from pressure data alone. A parameter that is certainly relevant to measure in addition to pressure and flow is pulse wave velocity, as it provides both functional information, and helps to elucidate wave reflection. For measuring the transit time (e.g., from carotid to femoral) of the arterial pressure pulse, the flow velocity wave or the diameter distension wave is the most commonly applied technique (85).

HEART–ARTERIAL COUPLING

The main function of the heart is to maintain the circulation, that is, to provide a sufficient amount of blood at sufficiently high pressures to guarantee the perfusion of vital organs, including the heart itself. Arterial pressure and flow arise from the interaction between the ventricle and the arterial load. In the past few years, the coupling of the heart and the arterial system in cardiovascular pathophysiology has been recognized (74–76,117,118). With ageing or in hypertension, for example, arterial stiffening leads to an increase of the pressure pulse and an early

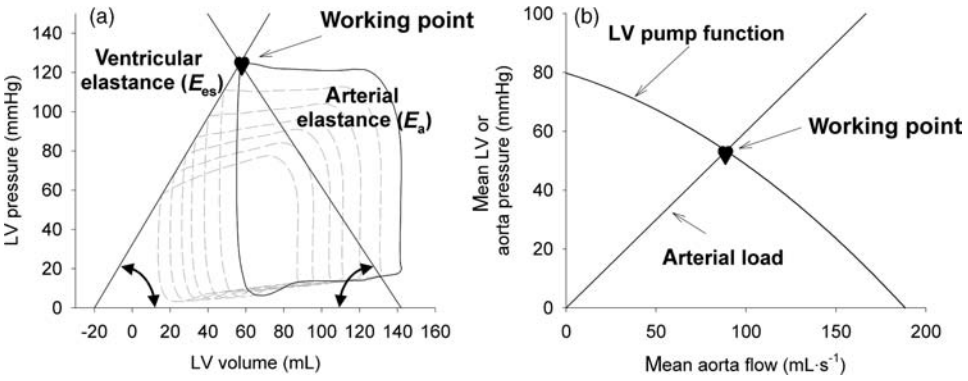


Figure 13. Heart–arterial coupling using the E_a–E_{es} (a) of the pump function graph (b) framework. In both cases, the intersection between the curves characterizing the ventricle and the arterial system determines the working point of the cardiovascular system.

return of reflected waves, increasing the load on the heart (73,119), which will react through adaptation (remodeling) to cope with the increased afterload.

Factors affecting ventricular pump function are preload (venous filling pressure and/or end diastolic volume EDV), heart rate, and the intrinsic contractile function of the heart muscle (here assumed to be best represented by the slope of the end-systolic PV relation, E_{es}). The two main mechanical afterload parameters determining systolic and pulse pressure are total peripheral resistance (R) and total arterial compliance (C) (120,121). Arterial blood pressure and flow are thus determined by a limited number of cardiac and arterial mechanical factors.

The most common framework, however, to study the heart–arterial (or ventricular–vascular) coupling, is the E_a–E_{es} framework (23,122–124), where E_a is effective arterial elastance (125), and where E_a/E_{es} is a parameter reflecting heart–arterial coupling. In the pressure–volume plane, E_a is the slope of the line connecting the end-systolic point (ESV, ESP) with the end-diastolic volume point on the volume axis (EDV, 0) (Fig. 13). As such, $E_a = \text{ESP/SV}$, with SV the stroke volume. The parameter E_a can be written as a function of three-element windkessel parameters describing the arterial system (125), but it has been shown that E_a can be approximated as R/T with T the duration of the cardiovc cycle (124,126).

The first major advantage of this framework is that it allows us to graphically study the interaction between the heart and the arterial system. Cardiac function is characterized by the ESPVR, which is a fully defined line in the PV plane. For a given preload (EDV) and arterial stiffness E_a, the line characterizing the arterial system can be drawn as well. The intersection of these two lines determines the end-systolic point (ESV, ESP), and thus the stroke volume and (end-)systolic pressure (Fig. 13).

Second, the analysis in the PV plane also allows us to calculate some energetic parameters, and to relate them to E_a/E_{es}. It can be shown that SW is maximal only when $E_a = E_{es}$. Thus when $E_a/E_{es} = 1$. Mechanical efficiency is maximal when $E_a/E_{es} = 0.5$, when arterial elastance is one-half of the end-systolic elastance (124). In normal hearts, the heart operates in conditions with an E_a/E_{es} ratio in the range 0.5:1, and it has been shown that this ratio is preserved in normal aging (74,75). Arterial stiffening is thus paralleled by an increase in ventricular end-systolic stiffness. This happens even in patients with heart failure, but with preserved ejection fraction (76). In this patient population, both E_a and E_{es} are higher than expected with normal ageing, but the ratio is more or less preserved.

Although the E_a/E_{es} parameter certainly has its value, it has been shown under experimental conditions that the E_a/E_{es} range, where the heart operates near optimal efficiency, is quite broad (122,127). Thus the parameter is not a very sensitive tool to detect "uncoupling" of the heart and the arterial system, which commonly implies a increase in E_a and a poorer ventricular function (depressed E_{es}) leading to increased E_a/E_{es} ratios (128,129). Also, one should not focus only on the coupling, but also on the absolute values of E_a and E_{es}. Increased stiffening leads to hypertensive response in exercise, where small changes in volume (filling volumes, stroke volumes) have an amplified effect on arterial pressure and workload, and lead to an increased energy cost to increase stroke volume (75,76).

By using dimensional analysis, Stergiopulos et al. (130) and later Segers et al. (131), demonstrated that blood pressure and stroke volume are mainly determined by ventricular preload, and by 2 dimensionless parameters: $E_{es}C$ and RC/T, where T is the heart period. The first is the product of ventricular (end-systolic) stiffness and arterial compliance; the latter is the ratio of the time constant of arterial blood pressure decay (RC) and the heart period.

It is also worth noting the link between the coupling parameters as proposed by Stergiopulos et al., and E_a/E_{es}: $(RC/T)/(E_{es}C) = (R/T)/E_{es} \approx E_a/E_{es}$. E_a/E_{es} thus combines the two coupling parameters following from the dimensional analysis into a single dimensionless parameter. This is, however, at the cost of eliminating the contribution of total arterial compliance from the parameter. Within this perspective, the terminology of E_a as arterial stiffness, is perhaps not entirely justified, as it is a parameter related to peripheral resistance and the heart period, rather than to arterial compliance or stiffness (126).

Finally, it should be mentioned that heart–arterial interaction can also be studied within the pump function framework of Westerhof and co-workers (Fig. 13). While the function of the heart can be described with the pump function graph, the function of the arterial system can be displayed in the same graph. In its simplest approximation, considering only mean pressure and flow, the arterial system is characterized by the total vascular resistance. In a pump function graph, this is represented by a straight line: mean flow is directly proportional to mean arterial pressure.

The intersection with the pump function graph yields the working point of the cardiovascular system. Note, however, that although both the pump and arterial function curve are displayed in the same figure, their y axis is not the same.

BIBLIOGRAPHY

Cited References

1. Strackee EJ, Westerhof N. Physics of Heart and Circulation. Bristol (UK): Institute of Physics; 1993.
2. Suga H, Sagawa K, Shoukas AA. Load independence of the instantaneous pressure–volume ratio of the canine left ventricle and effects of epinephrine and heart rate on the ratio. Circ Res 1973;32:314–322.
3. Suga H, Sagawa K. Instantaneous pressure–volume relationships and their ratio in the exercised, supported canine left ventricle. Circ Res 1974;35:117–126.
4. van der Velde ET, Burkhoff D, Steendijk P, Karsdon J, Sagawa K, Baan J. Nonlinearity and load sensitivity of end-systolic pressure–volume relation of canine left ventricle *in vivo*. Circulation 1991;83:315–327.
5. Kass DA, Beyar R, Lankford E, Heard M, Maughan WL, Sagawa K. Influence of contractile state on curvilinearity of in situ end-systolic pressure–volume relations. Circulation 1989;79:167–178.
6. Baan J, van der Velde ET, de Bruin HG, Smeenk GJ, Koops J, van Dijk AD, Temmerman D, Senden J, Buis B. Continuous measurement of left ventricular volume in animals and humans by conductance catheter. Circulation 1984;70:812–823.
7. Suga H. Cardiac energetics: from E(max) to pressure–volume area. Clin Exp Pharmacol Physiol 2003;30:580–585.
8. Suga H. Global cardiac function: mechano-energetico-informatics. J Biomech 2003;36:713–720.
9. Glower DD, Spratt JA, Snow ND, Kabas JS, Davis JW, Olsen CO, Tyson GS, Sabiston Jr DC, Rankin JS. Linearity of the Frank–Starling relationship in the intact heart: the concept of preload recruitable stroke work. Circulation 1985;71:994–1009.
10. Starling E, Vischer M. The regulation of the output of the heart. J Physiol Cambridge 1927;62:243–261.
11. Elzinga G, Westerhof N. How to quantify pump function of the heart. The value of variables derived from measurements on isolated muscle. Circ Res 1979;44:303–308.
12. Elzinga G, Westerhof N. Pump function of the feline left heart: changes with heart rate and its bearing on the energy balance. Cardiovasc Res 1980;14:81–92.
13. Takeuchi M, Igarashi Y, Tomimoto S, Odake M, Hayashi T, Tsukamoto T, Hata K, Takaoka H, Fukuzaki H. Single-beat estimation of the slope of the end-systolic pressure–volume relation in the human left ventricle. Circulation 1991;83: 202–212.
14. Chen CH, Fetics B, Nevo E, Rochitte CE, Chiou KR, Ding PA, Kawaguchi M, Kass DA. Noninvasive single-beat determination of left ventricular end-systolic elastance in humans. J Am Coll Cardiol 2001;38:2028–2034.
15. Lambermont B, Segers P, Ghuysen A, Tchana-Sato V, Morimont P, Dogne JM, Kolh P, Gerard P, D'Orio V. Comparison between single-beat and multiple-beat methods for estimation of right ventricular contractility. Crit Care Med 2004;32:1886–1890.
16. Kjorstad KE, Korvald C, Myrmel T. Pressure–volume-based single-beat estimations cannot predict left ventricular contractility *in vivo*. Am J Physiol Heart Circ Physiol 2002;282:H1739–H1750.
17. Kass DA, Beyar R. Evaluation of contractile state by maximal ventricular power divided by the square of end-diastolic volume. Circulation 1991;84:1698–1708.
18. Nakayama M, Chen CH, Nevo E, Fetics B, Wong E, Kass DA. Optimal preload adjustment of maximal ventricular power index varies with cardiac chamber size. Am Heart J 1998;136:281–288.
19. Segers P, Leather HA, Verdonck P, Sun Y-Y, Wouters PF. Preload-adjusted maximal power of right ventricle: contribution of end-systolic P-V relation intercept. Am J Physiol 2002;283:H1681–H1687.
20. Segers P, Tchana-Sato V, Leather HA, Lambermont B, Ghuysen A, Dogne JM, Benoit P, Morimont P, Wouters PF, Verdonck P, Kolh P. Determinants of left ventricular preload-adjusted maximal power. Am J Physiol Heart Circ Physiol 2003;284:H2295–H2301.
21. Ommen SR. Echocardiographic assessment of diastolic function. Curr Opin Cardiol 2001;16:240–245.
22. Garcia MJ, Thomas JD, Klein AL. New Doppler echocardiographic applications for the study of diastolic function. J Am Coll Cardiol 1998;32:865–875.
23. Appleton CP, Hatle LK, Popp RL. Relation of transmitral flow velocity patterns to left ventricular diastolic function: new insights from a combined hemodynamic and Doppler echocardiographic study. J Am Coll Cardiol 1988;12:426–440.
24. Thomas JD, Zhou J, Greenberg N, Bibawy G, McCarthy PM, Vandervoort PM. Physical and physiological determinants of pulmonary venous flow: numerical analysis. Am J Physiol 1997;272:H2453–H2465.
25. Garcia MJ, Ares MA, Asher C, Rodriguez L, Vandervoort P, Thomas JD. An index of early left ventricular filling that combined with pulsed Doppler peak E velocity may estimate capillary wedge pressure. J Am Coll Cardiol 1997;29:448–454.
26. De Mey S, De Sutter J, Vierendeels J, Verdonck P. Diastolic filling and pressure imaging: taking advantage of the information in a colour M-mode Doppler image. Eur J Echocardiog 2001;2:219–233.
27. Nagueh SF, Middleton KJ, Kopelen HA, Zoghbi WA, Quinones MA. Doppler tissue imaging: a noninvasive technique for evaluation of left ventricular relaxation and estimation of filling pressures. J Am Coll Cardiol 1997;30:1527–1533.
28. Heimdal A, Stoylen A, Torp H, Skjaerpe T. Real-time strain rate imaging of the left ventricle by ultrasound. J Am Soc Echocardiog 1998;11:1013–1019.
29. D'Hooge J, Heimdal A, Jamal F, Kukulski T, Bijnens B, Rademakers F, Hatle L, Suetens P, Sutherland GR. Regional strain and strain rate measurements by cardiac ultrasound: principles, implementation and limitations. Eur J Echocardiog 2000;1:154–170.
30. Chaney JC, Derdak S. Minimally invasive hemodynamic monitoring for the intensivist: current and emerging technology. Crit Care Med 2002;30:2338–2345.
31. Leather HA, Vuylsteke A, Bert C, M'Fam W, Segers P, Sergeant P, Vandermeersch E, Wouters PF. Evaluation of a new continuous cardiac output monitor in off-pump coronary artery surgery. Anaesthesia 2004;59:385–389.
32. O'Dell WG, McCulloch AD. Imaging three-dimensional cardiac function. Annu Rev Biomed Eng 2000;2:431–456.
33. Matter C, Nagel E, Stuber M, Boesiger P, Hess OM. Assessment of systolic and diastolic LV function by MR myocardial tagging. Basic Res Cardiol 1996;91:23–28.
34. Pijls NHJ, de Bruyne B. Coronary Pressure. Dordrecht: Kluwer Academic Publishers; 1997.
35. Spaan JAE. Coronary Blood Flow: Mechanics, Distribution, and Control. Dordrecht and Boston: Kluwer Academic Publishers; 1991.
36. Mosher P, Ross Jr J, McFate PA, Shaw RF. Control of Coronary Blood Flow by an Autoregulatory Mechanism. Circ Res 1964;14:250–259.

37. Bellamy RF. Diastolic coronary artery pressure-flow relations in the dog. Circ Res 1978;43:92–101.
38. Downey JM, Kirk ES. Inhibition of coronary blood flow by a vascular waterfall mechanism. Circ Res 1975;36:753–760.
39. Permutt S, Riley RL. Hemodynamics of Collapsible Vessels with Tone: The Vascular Waterfall. J Appl Physiol 1963;18:924–932.
40. Spaan JA, Breuls NP, Laird JD. Diastolic-systolic coronary flow differences are caused by intramyocardial pump action in the anesthetized dog. Circ Res 1981;49:584–593.
41. Matthys K, Carlier S, Segers P, Ligthart J, Sianos G, Serrano P, Verdonck PR, Serruys PW. In vitro study of FFR, QCA, and IVUS for the assessment of optimal stent deployment. Catheter Cardiovasc Interv 2001;54:363–375.
42. Siebes M, Verhoeff BJ, Meuwissen M, de Winter RJ, Spaan JA, Piek JJ. Single-wire pressure and flow velocity measurement to quantify coronary stenosis hemodynamics and effects of percutaneous interventions. Circulation 2004; 109:756–762.
43. Westerhof N, Stergiopulos N, Noble M. Snapshots of Hemodynamics. An aid for clinical research and graduate education. New York: Springer Science + Business Media; 2004.
44. Nichols WW, O'Rourke MF. McDonald's Blood Flow in Arteries. 3rd ed. London: Edward Arnold; 1990.
45. Milnor WR. Hemodynamics. 2nd ed. Baltimore (MA): Williams & Wilkins; 1989.
46. Hales S. Statical Essays: Containing Haemostatics (reprint 1964). New York: Hafner Publishing; 1733.
47. Frank O. Die Grundfurm des arteriellen Pulses. Erste Abhandlung. Mathematische Analyse. Z Biol 1899;37: 483–526.
48. Frank O. Der Puls in den Arterien. Z Biol 1905;46:441–553.
49. Segers P, Verdonck P. Principles of Vascular Physiology. In: Lanzer P, Topol EJ, editors. Pan Vascular Medicine. Integrated Clinical Management. Heidelberg: Springer-Verlag; 2002.
50. Toy SM, Melbin J, Noordergraaf A. Reduced models of arterial systems. IEEE Trans Biomed Eng 1985;32:174–176.
51. Liu Z, Brin K, Yin F. Estimation of total arterial compliance: an improved method and evaluation of current methods. Am J Physiol 1986;251:H588–H600.
52. Simon A, Safar L, London G, Levy B, Chau N. An evaluation of large arteries compliance in man. Am J Physiol 1979;237:H550–H554.
53. Stergiopulos N, Meister JJ, Westerhof N. Evaluation of methods for the estimation of total arterial compliance. Am J Physiol 1995;268:H1540–H1548.
54. Chemla D, Hébert J-L, Coirault C, Zamani K, Suard I, Colin P, Lecarpentier Y. Total arterial compliance estimated by stroke volume-to-aortic pulse pressure ratio in humans. Am J Physiol 1998;274:H500–H505.
55. Stergiopulos N, Segers P, Westerhof N. Use of pulse pressure method for estimating total arterial compliance *in vivo*. Am J Physiol 1999;276:H424–H428.
56. Westerhof N, Elzinga G, Sipkema P. An artificial arterial system for pumping hearts. J Appl Physiol 1971;31:776–781.
57. Westerhof N, Elzinga G. Normalized input impedance and arterial decay time over heart period are independent of animal size. Am J Physiol 1991;261:R126–R133.
58. Segers P, Brimioulle S, Stergiopulos N, Westerhof N, Naeije R, Maggiorini M, Verdonck P. Pulmonary arterial compliance in dogs and pigs: the three-element windkessel model revisited. Am J Physiol 1999;277:H725–H731.
59. Stergiopulos N, Westerhof B, Westerhof N. Total arterial inertance as the fourth element of the windkessel model. Am J Physiol 1999;276:H81–H88.
60. Burattini R, Gnudi G. Computer identification of models for the arterial tree input impedance: comparison between two new simple models and first experimental results. Med Biol Eng Comp 1982;20:134–144.
61. Murgo JP, Westerhof N, Giolma JP, Altobelli SA. Aortic input impedance in normal man: relationship to pressure wave forms. Circulation 1980;62:105–116.
62. Burattini R, Campbell KB. Modified asymmetric T-tube model to infer arterial wave reflection at the aortic root. IEEE Trans Biomed Eng 1989;36:805–814.
63. Chang KC, Tseng YZ, Kuo TS, Chen HI. Impedance and wave reflection in arterial system: simulation with geometrically tapered T-tubes. Med Biol Eng Comput 1995;33:652–660.
64. Segers P, Verdonck P. Role of tapering in aortic wave reflection: hydraulic and mathematical model study. J Biomech 2000;33:299–306.
65. Wang DM, Tarbell JM. Nonlinear analysis of oscillatory flow, with a nonzero mean, in an elastic tube (artery). J Biomech Eng 1995;117:127–135.
66. Campbell K, Lee CL, Frasch HF, Noordergraaf A. Pulse reflection sites and effective length of the arterial system. Am J Physiol 1989;256:H1684–H1689.
67. Latham R, Westerhof N, Sipkema P, Rubal B, Reuderink P, Murgo J. Regional wave travel and reflections along the human aorta: a study with six simultaneous micromanometric pressures. Circulation 1985;72:1257–1269.
68. Berger D, Li J, Laskey W, Noordergraaf A. Repeated reflection of waves in the systemic arterial system. Am J Physiol 1993;264:H269–H281.
69. Westerhof N, Sipkema P, Van Den Bos G, Elzinga G. Forward and backward waves in the arterial system. Cardiovasc Res 1972;6:648–656.
70. Dujardin J, Stone D. Characteristic impedance of the proximal aorta determined in the time and frequency domain: a comparison. Med Biol Eng Comp 1981;19:565–568.
71. Khir AW, O'Brien A, Gibbs JS, Parker KH. Determination of wave speed and wave separation in the arteries. J Biomech 2001;34:1145–1155.
72. O'Rourke MF. Mechanical principles. Arterial stiffness and wave reflection. Pathol Biol (Paris) 1999;47:623–633.
73. Westerhof N, O'Rourke MF. Haemodynamic basis for the development of left ventricular failure in systolic hypertension and for its logical therapy. J Hypertens 1995;13:943–952.
74. Chen C-H, Nakayama M, Nevo E, Fetics BJ, Maughan WL, Kass DA. Coupled systolic-ventricular and vascular stiffening with age. Implications for pressure regulation and cardiac reserve in the elderly. J Am Coll Cardiol 1998;32: 1221–1227.
75. Kass DA. Age-related changes in venticular-arterial coupling: pathophysiologic implications. Heart Fail Rev 2002;7:51–62.
76. Kawaguchi M, Hay I, Fetics B, Kass DA. Combined ventricular systolic and arterial stiffening in patients with heart failure and preserved ejection fraction: implications for systolic and diastolic reserve limitations. Circulation 2003;107: 714–720.
77. Parker KH, Jones CJ. Forward and backward running waves in the arteries: analysis using the method of characteristics. J Biomech Eng 1990;112:322–326.
78. Bleasdale RA, Mumford CE, Campbell RI, Fraser AG, Jones CJ, Frenneaux MP. Wave intensity analysis from the common carotid artery: a new noninvasive index of cerebral vasomotor tone. Heart Vessels 2003;18:202–206.
79. Wang Z, Jalali F, Sun YH, Wang JJ, Parker KH, Tyberg JV. Assessment of Left Ventricular Diastolic Suction in Dogs using Wave-intensity Analysis. Am J Physiol Heart Circ Physiol 2004.?
80. Sun YH, Anderson TJ, Parker KH, Tyberg JV. Effects of left ventricular contractility and coronary vascular resistance on

coronary dynamics. Am J Physiol Heart Circ Physiol 2004;286:H1590–H1595.
81. Kelly R, Hayward C, Avolio A, O'Rourke M. Noninvasive determination of age-related changes in the human arterial pulse. Circulation 1989;80:1652–1659.
82. Hayward CS, Kelly RP. Gender-related differences in the central arterial pressure waveform. J Am Coll Cardiol 1997;30:1863–1871.
83. Wilkinson IB, MacCallum H, Flint L, Cockcroft JR, Newby DE, Webb DJ. The influence of heart rate on augmentation index and central arterial pressure in humans. J Physiol 2000;525:263–270.
84. Bramwell CJ, Hill A. The velocity of the pulse wave in man. Proc R Soc London (Biol) 1922;93:298–306.
85. Lehmann ED. Noninvasive measurements of aortic stiffness: methodological considerations. Pathol Biol (Paris) 1999;47:716–730.
86. Drzewiecki GM, Melbin J, Noordergraaf A. Arterial tonometry: review and analysis. J Biomech 1983;16:141–152.
87. Chen CH, Ting CT, Nussbacher A, Nevo E, Kass D, Pak P, Wang SP, Chang MS, Yin FC. Validation of carotid artery tonometry as a means of estimating augmentation index of ascending aortic pressure. Circulation 1996;27:168–175.
88. Hoeks AP, Brands PJ, Smeets FA, Reneman RS. Assessment of the distensibility of superficial arteries. Ultrasound Med Biol 1990;16:121–128.
89. Segers P, Rabben SI, De Backer J, De Sutter J, Gillebert TC, Van Bortel L, Verdonck P. Functional analysis of the common carotid artery: relative distension differences over the vessel wall measured *in vivo*. J Hypertens 2004;22:973–981.
90. Avolio A, Chen S, Wang R, Zhang C, Li M, O'Rourke M. Effects of aging on changing arterial compliance and left ventricular load in a northern Chinese urban community. Circulation 1983;68:50–58.
91. Avolio A, Fa-Quan D, Wei-Qiang L, Yao-Fei L, Zhen-Dong H, Lian-Fen X, M. OR. Effects of aging on arterial distensibility in populations with high and low prevalence of hypertension: comparison between urban and rural communities in China. Circulation 1985;71:202–210.
92. Karamanoglu M, Gallagher D, Avolio A, O'Rourke M. Functional origin of reflected pressure waves in a multibranched model of the human arterial system. Am J Physiol 1994;267:H1681–H1688.
93. Karamanoglu M, Gallagher D, Avolio A, O'Rourke M. Pressure wave propagation in a multibranched model of the human upper limb. Am J Physiol 1995;269:H1363–H1369.
94. O'Rourke MF. Pressure and flow waves in the systemic arteries and the anatomical design of the arterial system. J Appl Physiol 1967;23:139–149.
95. Avolio A. Multi-branched model of the human arterial system. Med Biol Eng Comp 1980;18:709–718.
96. Stergiopulos N, Young DF, Rogge TR. Computer simulation of arterial flow with applications to arterial and aortic stenoses. J Biomech 1992;25:1477–1488.
97. Takazawa K, Tanaka N, Takeda K, Kurosu F, Ibukiyama C. Underestimation of Vasodilator Effects of Nitroglycerin by Upper Limb Blood Pressure. Hypertension 1995;26:520–523.
98. Vlachopoulos C, Hirata K, O'Rourke MF. Pressure-altering agents affect central aortic pressures more than is apparent from upper limb measurements in hypertensive patients: the role of arterial wave reflections. Hypertension 2001;38: 1456–1460.
99. Chen C-H, Nevo E, Fetics B, Pak PH, Yin FCP, Maughan L, Kass DA. Estimation of Central Aortic Pressure Waveform by Mathematical Transformation of Radial Tonometry Pressure: Validation of Generalized Transfer Function. Circulation 1997;95:1827–1836.
100. Fetics B, Nevo E, Chen CH, Kass DA. Parametric model derivation of transfer function for noninvasive estimation of aortic pressure by radial tonometry. IEEE Trans Biomed Eng 1999;46:698–706.
101. Karamanoglu M, O'Rourke M, Avolio A, Kelly R. An analysis of the relationship between central aortic and peripheral upper limb pressure waves in man. Eur Heart J 1993; 14:160–167.
102. Karamanoglu M, Fenely M. Derivation of the ascending aorta-carotid pressure transfer function with an arterial model. Am J Physiol 1996;271:H2399–H2404.
103. Matthys K, Verdonck P. Development and modelling of arterial applanation tonometry: a review. Technol Health Care 2002;10:65–76.
104. Pauca AL, O'Rourke MF, Kon ND. Prospective evaluation of a method for estimating ascending aortic pressure from the radial artery pressure waveform. Hypertension 2001;38: 932–937.
105. Hope SA, Tay DB, Meredith IT, Cameron JD. Comparison of generalized and gender-specific transfer functions for the derivation of aortic waveforms. Am J Physiol Heart Circ Physiol 2002;283:H1150–H1156.
106. Segers P, Qasem A, De Backer T, Carlier S, Verdonck P, Avolio A. Peripheral "Oscillatory" Compliance Is Associated With Aortic Augmentation Index. Hypertension 2001;37: 1434–1439.
107. O'Rourke MF, Nichols WW. Use of arterial transfer function for the derivation of aortic waveform characteristics. J Hypertens 2003;21:2195–2197; author reply 2197–2199.
108. Albaladejo P, Copie X, Boutouyrie P, Laloux B, Declere AD, Smulyan H, Benetos A. Heart rate, arterial stiffness, and wave reflections in paced patients. Hypertension 2001;38: 949–52.
109. Wilkinson IB, Mohammad NH, Tyrrell S, Hall IR, Webb DJ, Paul VE, Levy T, Cockcroft JR. Heart rate dependency of pulse pressure amplification and arterial stiffness. Am J Hypertens 2002;15:24–30.
110. Marcus R, Korcarz C, McCray G, Neumann A, Murphy M, Borow K, Weinert L, Bednarz J, Gretler D, Spencer K, Sareli P, Lang R. Noninvasive method for determination of arterial compliance using Doppler echocardiography and subclavian pulse tracings. Circulation 1994;89:2688–2699.
111. Kelly R, Karamanoglu M, Gibbs H, Avolio A, O'Rourke M. Noninvasive carotid pressure wave registration as an indicator of ascending aortic pressure. J Vas Med Biol 1989;1:241–247.
112. Van Bortel LM, Balkestein EJ, van der Heijden-Spek JJ, Vanmolkot FH, Staessen JA, Kragten JA, Vredeveld JW, Safar ME, Struijker Boudier HA, Hoeks AP. Noninvasive assessment of local arterial pulse pressure: comparison of applanation tonometry and echo-tracking. J Hypertens 2001;19:1037–1044.
113. Rabben SI, Baerum S, Sorhus V, Torp H. Ultrasound-based vessel wall tracking: an auto-correlation technique with RF center frequency estimation. Ultrasound Med Biol 2002;28: 507–517.
114. Langewouters G, Wesseling K, Goedhard W. The static elastic properties of 45 human thoracic and 20 abdominal aortas in vitro and the parameters of a new model. J Biomech 1984;17:425–435.
115. Hayashi K. Experimental approaches on measuring the mechanical properties and constitutive laws of arterial walls. J Biomech Eng 1993;115:481–488.
116. Meinders JM, Hoeks AP. Simultaneous assessment of diameter and pressure waveforms in the carotid artery. Ultrasound Med Biol 2004;30:147–154.
117. Lakatta EG, Levy D. Arterial and cardiac aging: major shareholders in cardiovascular disease enterprises:

Part II: the aging heart in health: links to heart disease. Circulation 2003;107:346–354.
118. Lakatta EG, Levy D. Arterial and cardiac aging: major shareholders in cardiovascular disease enterprises: Part I: aging arteries: a "set up" for vascular disease. Circulation 2003;107:139–146.
119. O'Rourke M. Arterial stiffness, systolic blood pressure, and logical treatment of arterial hypertension. Hypertension 1990;15:339–347.
120. Stergiopulos N, Westerhof N. Role of total arterial compliance and peripheral resistance in the determination of systolic and diastolic aortic pressure. Pathol Biol (Paris) 1999;47:641–647.
121. Stergiopulos N, Westerhof N. Determinants of pulse pressure. Hypertension 1998;32:556–559.
122. Sunagawa K, Maughan WL, Burkhoff D, Sagawa K. Left ventricular interaction with arterial load studied in isolated canine ventricle. Am J Physiol 1983;245:H773–H780.
123. Sunagawa K, Maughan WL, Sagawa K. Optimal arterial resistance for the maximal stroke work studied in isolate canine left ventricle. Circ Res 1985;56:586–595.
124. Burkhoff D, Sagawa K. Ventricular efficiency predicted by an analytical model. Am J Physiol 1986;250:R1021–R1027.
125. Kelly R, Ting C, Yang T, Liu C, Lowell W, Chang M, Kass D. Effective arterial elastance as index of arterial vascular load in humans. Circulation 1992;86:513–521.
126. Segers P, Stergiopulos N, Westerhof N. Relation of effective arterial elastance to arterial system properties. Am J Physiol Heart Circ Physiol 2002;282:H1041–H1046.
127. De Tombe PP, Jones S, Burkhoff D, Hunter WC, Kass DA. Ventricular stroke work and efficiency both remain nearly optimal despite altered vascular loading. Am J Physiol 1993;264:H1817–H1824.
128. Asanoi H, Sasayama S, Kameyama T. Ventriculoarterial coupling in normal and failing heart in humans. Circ Res 1989;65:483–493.
129. Sasayama S, Asanoi H. Coupling between the heart and arterial system in heart failure. Am J Med 1991;90: 14S–18S.
130. Stergiopulos N, Meister JJ, Westerhof N. Determinants of stroke volume and systolic and diastolic pressure. Am J Physiol 1996;270:H2050–H2059.
131. Segers P, Steendijk P, Stergiopulos N, Westerhof N. Predicting systolic and diastolic aortic blood pressure and stroke volume in the intact sheep. J Biomech 2001;34:41–50.

See also BIOIMPEDANCE IN CARDIOVASCULAR MEDICINE; BLOOD PRESSURE MEASUREMENT; FLOWMETERS, ELECTROMAGNETIC; MONITORING, HEMODYNAMIC.

HEMODYNAMIC MONITORING.

See MONITORING, HEMODYNAMIC.

HIGH FREQUENCY VENTILATION

J. BERT BUNNELL
Bunnell Inc.
Salt Lake City, Utah

INTRODUCTION

High frequency ventilators (HFVs) were designed to eliminate many of the problems that conventional ventilators create as they try to mimic normal breathing. When we breathe normally, we draw gas into the lungs by creating a negative pressure with our diaphragm. Iron lungs were created to replicate that activity, and they worked very well for thousands of polio patients. However, when patients are sealed off in airtight iron lungs numerous practical problems unrelated to breathing arise.

Positive pressure ventilators made assisting ventilation much easier. Attaching the ventilator to the patient's lungs via an endotracheal (ET) tube greatly simplified patient care. But, lungs, especially premature lungs, are not designed to tolerate much positive pressure.

The lungs of prematurely borne infants have yet to be fully formed, and they lack the surfactant that enables alveoli to expand with very little pressure gradient. Hence, a considerable pressure gradient must be applied to ventilate them. Applying that pressure from the outside in, as *conventional* ventilators (CVs) have been doing since the early 1970s, causes problems. Tiny infant's airways get distended, alveoli are ruptured, and inflammatory sensors are triggered. Even if an infant receives artificial surfactant to lessen the need for assisted ventilation, and they grow new lung fast enough to survive, they may well develop chronic lung disease at a time when most of us were just taking our first breaths. Most premature infants outgrow their chronic lung disease, but many struggle mightily with every virus they encounter in their first few years of life, and they have an increased incidence of neurodevelopmental problems, such as cerebral palsy. Some infants lose those struggles and die of pneumonia.

Acute respiratory distress syndrome (ARDS) is the primary problem for mechanical ventilation of adults. This disease affects ~50 people per 100,000 with a mortality of 30–50%, and there have been few improvements in this mortality rate over the past several decades.

High frequency ventilators were developed in response to problems associated with CVs, but HFVs do not try to replicate normal breathing. They assist ventilation using much smaller tidal volumes delivered at rates ~10 times higher than normal. Animal and clinical studies indicate that smaller tidal volumes cause less lung injury (1,2).

Swedish anesthesiologists in the 1970s turned up the rate of their anesthesia ventilators to enable them to use smaller breaths to assist patients during neurosurgery (3). Their regular ventilators caused pulsations in blood pressure, causing brain movement every time the ventilator pushed in a breath, which was an obvious problem during microsurgery.

Auto accident victims whose heads went through car windshields also pose problems during surgery when access to the lungs has to pass right through the area of the face and neck where major reconstruction is required. So, another anesthesiologist, Dr. Miroslav Klain, began sticking needles into patients' necks to gain access to their tracheas, and he made it work by delivering very tiny breaths at very rapid rates (4).

The HFVs have shown great promise in supporting premature infants where fragile, underdeveloped and surfactant deficient lungs need to be gently ventilated until growth and maturation allow the newborn to catch up both anatomically and physiologically. In newborn infants,

where various modes of positive pressure ventilation have produced lung injury, HFVs have been widely accepted. Early studies using HFV to treat adults with severe ARDS have also shown promise in lessening lung injury and improving survival (5,6).

This article will review our current understanding of how HFVs work, the types of HFV equipment that are available for treating infants and adults, the results of several key animal and clinical studies that indicate how to optimize applications of HFVs, and what controversies remain to be resolved before HFVs can be considered as a primary mode of ventilation.

THEORETICAL BASIS FOR HFV: HOW HFVs WORK

High frequency ventilators are different from all other types of mechanical ventilators. They do not mimic normal breathing; rather, they facilitate gas exchange in a manner similar to panting in animals.

There are two elements to explaining how HFVs work with the higher than normal frequencies and smaller than normal tidal volumes. We begin with the assumption that ventilation or CO_2 elimination is proportional to minute volume or frequency times tidal volume, as

$$\dot{V}_{CO_2} \propto \dot{V}_{min} = f \times V_T \quad (1)$$

where $\dot{V}_{CO_2}$ = rate of carbon dioxide elimination; $\dot{V}_{min}$ = minute volume; f = ventilator frequency; and V_T = tidal volume.

If the practical limits of the high frequency end of this relationship are considered, there must be a lower limit on tidal volume size that will effectively provide alveolar ventilation. That lower limit is related to the effective or physiologic dead space of the lungs by the following equation:

$$\dot{V}_A = f \times (V_T - V_D) \quad (2)$$

where $\dot{V}_A$ = alveolar ventilation, and V_D = effective or physiologic dead space.

Thus, as tidal volume approaches the size of the effective dead space of the lungs, ventilation of the alveoli becomes nil.

Physiologic Dead Space and the Lower Limit of Tidal Volume

Anatomic dead space of mammalian lungs is generally considered to be $2\,mL \cdot kg^{-1}$ body weight (7). When one breathes normally, effective or physiologic dead space must be at least as large as anatomic dead space, because one pushes the dead space gas back into the alveoli ahead of the fresh gas during inhalation. What happens in the panting animal is another matter, as Henderson and associates described in 1915 (8).

Henderson et al. (8) demonstrated that panting animals breathe very shallowly as well as rapidly. They hypothesized that physiologic dead space changes at rapid respiratory rates in mammals, and they measured these effects on themselves. They also performed a series of experiments using smoke to demonstrate how inhaled gas penetrates through the anatomic dead space with rapid inhalations in a manner that makes physiologic dead space become less than anatomic dead space (Fig. 1).

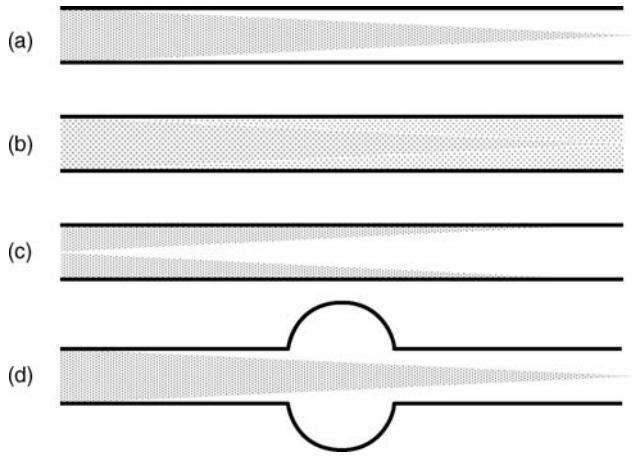


Figure 1. Henderson's smoke experiment. (a) A long thin spike or jet stream of smoke shoots downstream when suddenly blown into a glass tube. (b) The jet stream disappears when flow stops and diffusion takes place. (c) This effect can be duplicated in the opposite direction if fresh gas is drawn back into the smoke filled tube with a sudden inhalation. (d) Imperfections in the tube walls (such as a bulb) have little effect on the shape of the jet stream. (Adapted with permission from Ref. 8, p. 8. © 1915, American Physiology Society.)

Regardless of how much physiologic dead space can be reduced in the panting animal, there is still the matter of providing adequate alveolar ventilation as defined by Eq. 2. Thus, the extent to which smaller tidal volumes can be used to ventilate the alveoli has to be balanced by an increase in breathing frequency, as defined by Eq. 1. Panting animals breathe very rapidly, of course, but humans do not pant as a rule. So, how can the benefits of increasing ventilator frequency for humans be explained?

The Natural Frequency of the Lungs

There is a mechanical advantage to ventilating lungs at frequencies higher than normal breathing frequency. This phenomenon was revealed by a diagnostic technique for measuring airway resistance called forced oscillations (9).

Applying forced oscillations to measure airway resistance requires a person to hold a large bore tube in their mouth and allow small volumes of gas to be oscillated in and out of their lungs by a large loudspeaker. The frequency of oscillations produced by the speaker is varied through a spectrum from low (~1 Hz) to high (~60 Hz), and the pressure amplitudes of the oscillations are measured along with the flow rate of the gas that is passing in and out of the lungs (Fig. 2 depicts the test set up). Although the volume of gas moving in and out of the lungs is a constant, pressure amplitude varies with frequency and is minimized at the resonant or natural frequency of the lungs.

The concept that the lungs have a natural frequency is explained by consideration of lung mechanics. There are three elements to lung impedance (those things that impede the flow of gas in and out of the lungs): airway resistance, lung compliance, and inertance. We normally are not concerned about inertance, since it is concerned with the energy involved in moving the mass in the system,

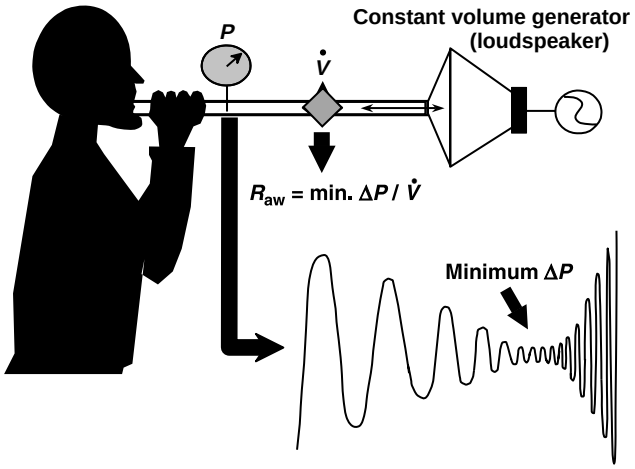


Figure 2. Measuring airway resistance at the resonant or natural frequency of the lungs using forced oscillations. (Used with permission. © 2003, Bunnell Inc.)

most of which is gas, and gas does not have much mass. Therefore, it does not take much energy to overcome inertance when one breathes: unless one is breathing very rapidly.

In the forced oscillations determination of airway resistance, the point of minimum pressure amplitude marks the frequency at which the energy necessary to overcome the elasticity of the lungs is supplied by the energy temporarily stored in the inertial elements of the system (i.e., the gas rushing in). (We normally measure lung elasticity inversely as lung compliance.) As the lungs recoil at the end of the gas-in phase, the elasticity of the lungs imparts its energy to turn the gas around and send it back out to the loudspeaker.

When the natural frequency or resonance is reached, the speaker and lungs exchange the gas being forced in and out of the lungs with ease. The lungs and the speaker accept the gas and recoil at just the right times to keep the gas oscillating back and forth with minimal energy required to keep the gas moving. At this point, the only element impeding gas flow is frictional airway resistance, which works against the gas coming in and going out. Its value can be calculated by dividing pressure amplitude by the gas flow rate when pressure amplitude is minimized.

The smaller the lungs are, the higher the natural frequency. The natural frequency of adult lungs is ~4 Hz, while that of premature infant lungs is closer to 40 Hz.

Putting Two and Two Together: How Can We HFV?

Combining the two concepts that describe the relationships of gas velocity, physiologic dead space, breathing frequency, and lung mechanics led us to HFV. We reported that one can then achieve adequate minute ventilation and compensate for very small tidal volumes in paralyzed animals by increasing ventilatory frequency to several hundred breaths per minute in 1978 (10). Pushing small volumes of gas into the lungs at high velocities reduced effective dead space volume and pushed the lower limit of effective tidal volume below anatomic dead space volume (~2 $mL \cdot kg^{-1}$). Increasing frequency to near resonant frequency also allowed us to minimize airway pressure.

As HFVs were developed and clinical use in newborn intensive care units (NICUs) became widespread in the 1980 and 1990s, numerous theories and experiments refined our concepts of how it all works. A number of prominent physiologists and bioengineers tackled the analysis and interpretation of gas exchange within the lungs during HFV while clinicians were seeking to identify appropriate applications of the new technique and all its intricacies. A few notable contributions will be discussed here.

Fredberg (11) and Slutsky et al. (12) analyzed mechanisms affecting gas transport during high frequency oscillation, expanding traditional concepts of convection and diffusion to include their combined effects, and termed the collection: augmented transport. Their analyses and those of Venegas et al. (13) and Permutt et al. (14) revealed that our traditional appreciation of the relationship between minute volume and CO_2 elimination must be modified during HFV to reflect the increased contribution of tidal volume, as

$$V_{CO_2} \propto f^a \times V_T^b \quad (3)$$

where the exponent b is greater than the exponent a. For practical purposes, most people now accept this relationship as

$$\dot{V}_{CO_2} \propto f \times V_T^2 \quad (4)$$

Slutsky also explored the limitations of HFV by measuring the effect of bronchial constriction on gas exchange. When the peripheral airways of dogs were constricted by administration of histamine, HFOV was no longer as effective at higher frequencies. (This issue is discussed later when the effectiveness of various types of HFVs for different pathophysiologies are explored.)

Venegas and Fredberg explored the importance of frequency during HFV in their classic paper of 1994, subtitled: "Why does high frequency ventilation work?" (15). They found that the resonant frequency of the smallest prematurely born infant with RDS (respiratory distress syndrome) is approximately 40 Hz. At that frequency, the minimum pressure amplitude is required to ventilate the lungs. However, the shape of theoretical curves of pressure amplitude measured at the carina versus frequency for infants with various lung conditions is most interesting as illustrated in Fig. 3, which was constructed using their concepts.

Figure 3 illustrates four essential considerations concerning the application of HFV for newborn infants.

1. Decreasing lung compliance moves the optimal frequency for HFV to the right (i.e., toward higher frequencies).
2. Increasing airway resistance moves the optimal frequency for HFV to the left (i.e., toward lower frequencies); and
3. There is diminishing value in applying HFV for infants at frequencies above ~ 10 Hz as far as airway pressure is concerned.
4. Choosing to operate at the "corner frequency" is an appropriate choice for HFV since there is little benefit

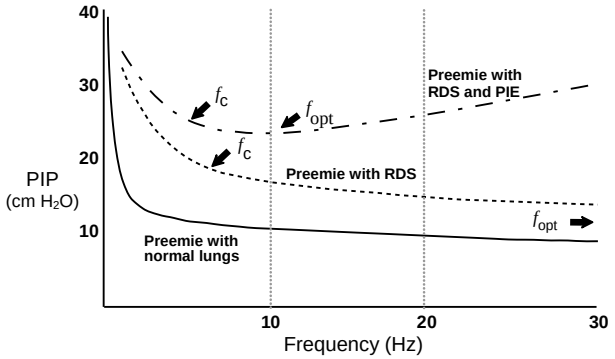


Figure 3. Theoretical peak carinal pressures for infants with normal lungs and lungs with poor compliance (RDS), poor airway resistance (asthma), and both conditions (RDS + PIE). Note how infants with RDS are well served using the corner frequency (f_c) of ~10 Hz (600 breaths per minute, bpm). Larger patients will exhibit curves with nearly identical shapes, but they will all be shifted to the left. (Adapted with permission from Ref. 15.)

above that frequency and more chance for gas trapping. Venegas and Fredberg define corner frequency as that frequency above which airway pressure required to provide adequate ventilation no longer rapidly decreases.

In other words, ventilating premature babies at 10 breaths $\cdot s^{-1}$ is *practically* as efficient as ventilating them at their theoretical resonant frequency of 40 "breaths" $\cdot s^{-1}$, where the danger of gas trapping is greatly increased. Patients with increased airway resistance require more careful consideration of the decreased benefits of exceeding the corner frequency and are more safely ventilated at lower frequencies.

One can calculate the resonant and corner frequencies if values of lung compliance (C_L), airway resistance (R_{aw}), and inertance (I) are known, but that is rarely the case with patients in intensive care. Venegas and Fredberg provided the following formulas:

$$f_0 = 1/(2\pi\sqrt{\tilde{C}_I}) \quad (5)$$

where f_0 = resonant frequency, and

$$f_c = 1/(2\pi CR) \quad (6)$$

where f_c = corner frequency. (Plug in typical values for lung compliance, inertance, and airway resistance of a premature infant, 0.5 mL$\cdot$cm^{-1} H_2O, 0.025 cm $H_2O\cdot L\cdot s^{-2}$, and 50 cm $H_2O\cdot L^{-1}\cdot s^{-1}$, respectively, and $f_0 = 45\cdot s^{-1}$ and $f_c = 6.4\cdot s^{-1}$.)

Finally, Venegas and Fredberg illustrated the value of using appropriate levels of positive end-expiratory pressure (PEEP). The PEEPs of 5–10 cm H_2O dramatically decrease the pressure amplitude necessary to ventilate premature infants at all frequencies when lung compliance is normal, and at all frequencies above ~6 Hz when lung compliance is reduced.

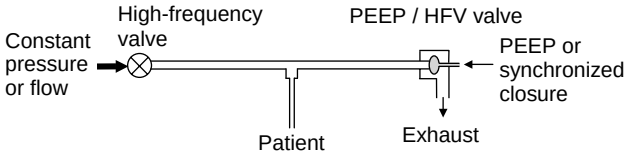


Figure 4. Basic design of HFPPVs. (Used with permission. © 2003, Bunnell Inc.)

HFV EQUIPMENT

Design Classifications

Figures 4–7 illustrate four different ways HFVs have been created. Figure 4 illustrates high frequency positive-pressure ventilation (HFPPV), which is basically a CV that operates at HFV rates. Early devices worked in this manner, but they seldom worked at the very high frequencies used with infants.

Figure 5 illustrates high frequency flow interruption (HFFI), where positive pressure oscillations are created by releasing gas under pressure into the breathing circuit via an HFV valve mechanism. The valve may be a solenoid valve or valves, a spinning ball with a hole in it, and so on. Early HFVs used long, small diameter exhaust tubes to increase impedance to gas flow oscillating at HFV frequencies in the expiratory limb so that the HFV oscillations would preferentially flow in and out of the patient.

High frequency oscillatory ventilators (HFOVs) work in a similar manner to HFFIs, as shown in Fig. 6, except that the pressure oscillations in the patient's breathing circuit are caused by an oscillating piston or diaphragm. Again, the impedance of the expiratory limb of the circuit tubing must be higher than the impedance of the patient and his ET tube when the gas flowing through the circuit is oscillating at HFV frequencies. The major difference between HFOV and HFFI is that pressure in the ventilator circuit during HFOV oscillates below atmospheric pressure in an effort to actively assist the patient's expiration. (This topic is discussed further below when gas trapping is addressed.)

Finally, Fig. 7 illustrates high frequency jet ventilation (HFJV), where inspiratory gas is injected into the patient's ET tube via a jet nozzle. Jet nozzles have been fashioned out of needles or built into special ET tubes or ET tube adapters as discussed below.

Each HFV approach introduces fresh gas into the patient's airway at about 10 times the patient's normal breathing frequency. The last three designs incorporate a separate constant flow of gas that passes by the patient's

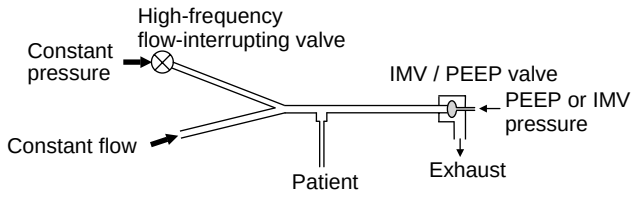


Figure 5. Basic design of HFFIs. (Used with permission. © 2003, Bunnell Inc.)

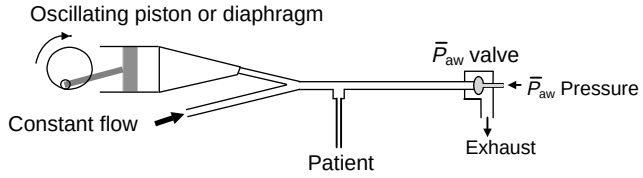


Figure 6. Basic design of HFOVs. (Used with permission. © 2003, Bunnell Inc.)

ET tube and out a large orifice valve to control baseline PEEP and mean airway pressure in the circuit. The patient may also breathe spontaneously from this gas stream, which may be provided by a built-in mechanism or by a separate conventional ventilator. Conventional IMV (intermittent mandatory ventilation) may be combined with HFV in this way. Additional hybrid devices that are more difficult to characterize have also been created, but the currently most common used HFVs are HFOVs, HFJVs, and conventional ventilators with built-in HFOV modules.

In the early 1980s, the FDA (U.S. Food and Drug Administration) decided that 150 breaths·min^{-1} would be the lower limit of what they would define as an HFV, and they placed rigorous Class III restrictions on any ventilator that operates above that frequency. As a result, there have been only six HFVs approved for use in the United States, three for infants and children, two for adults, and one HFV that was granted Class II approval (i.e., not needing proof of safety and efficacy since it was substantially equivalent to devices marketed before 1976, the date U.S. law was amended to require proof of safety and efficacy before new products can be marketed). At least four other HFVs are available outside the United States.

Of the FDA approved devices, two HFVs have been withdrawn from the market by the major corporation that acquired the smaller companies that developed them. Therefore, we are left with one HFJV, one HFOV for infants and children, one HFOV for children and adults, and a Class II HFV hybrid device designed for patients of all sizes. These four devices will be discussed in more detail below.

HFJV: High Frequency Jet Ventilators

The HFJVs inject inspired gas into the endotracheal tube via a jet nozzle. The Bunnell LifePulse High Frequency

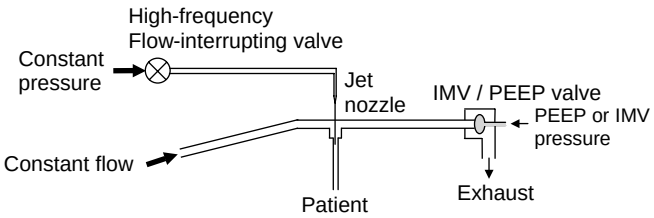


Figure 7. Basic design of HFJVs. (Used with permission. © 2003, Bunnell Inc.)

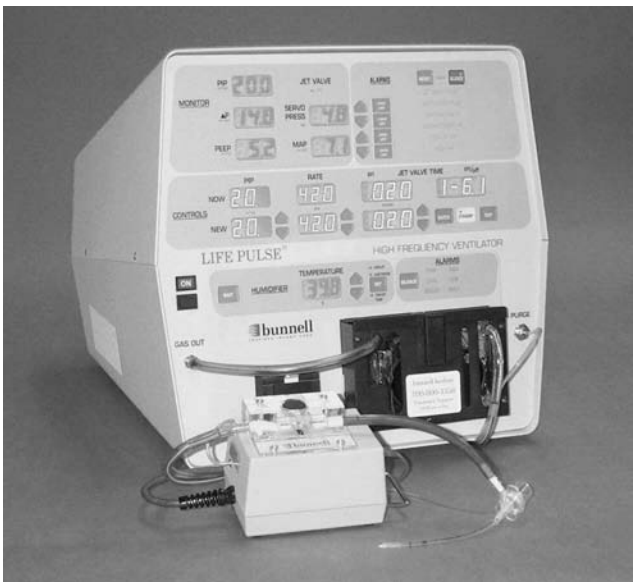


Figure 8. Bunnell life pulse HFV. (Used with permission. © 2003, Bunnell Inc.)

Ventilator is the only HFJV currently available for intensive care in the United States (Fig. 8). It was designed for infants and children up to ~10 years of age and operates at rates between 240 and 660 bpm. It is also used in tandem with a conventional ventilator, which provides for the patient's spontaneous breathing, delivery of occasional sigh breaths, and PEEP.

The LifePulse is a microprocessor controlled, pressure limited, time cycled ventilator that delivers heated and humidified breaths to the ET tube via a LifePort adapter (Fig. 9). A small Patient Box placed close to the patient's head contains an inhalation valve and pressure transducer for monitoring airway pressures in conjunction with the LifePort adapter. Peak inspiratory pressure (PIP) is feedback controlled by regulating the driving pressure (servo pressure) behind the jet nozzle. More detailed information on the device can be found on the manufacturer's website: www.bunl.com.

The theory of operation behind the LifePulse is that pulses of high velocity fresh gas stream down the center of the airways, penetrating through the dead-space gas, while exhaled gas almost simultaneously moves outward in the annular space along the airway walls. This countercurrent action facilitates mucociliary clearance while it minimizes effective dead space volume.

The pressure amplitude (ΔP) of LifePulse HFJV breaths is determined by the difference between PIP and the CV-controlled PEEP. Its value is displayed continuously on the LifePulse front panel along with mean airway pressure, PIP, PEEP, and Servo Pressure.

Servo Pressure on the LifePulse is a direct reflection of the gas flow needed to reach the set PIP, so it varies with the patient's changing lung mechanics. Alarm limits are automatically set around the Servo Pressure to alert the operator to significant changes in the patient's condition as well as the tubing connecting the patient to the HFJV.

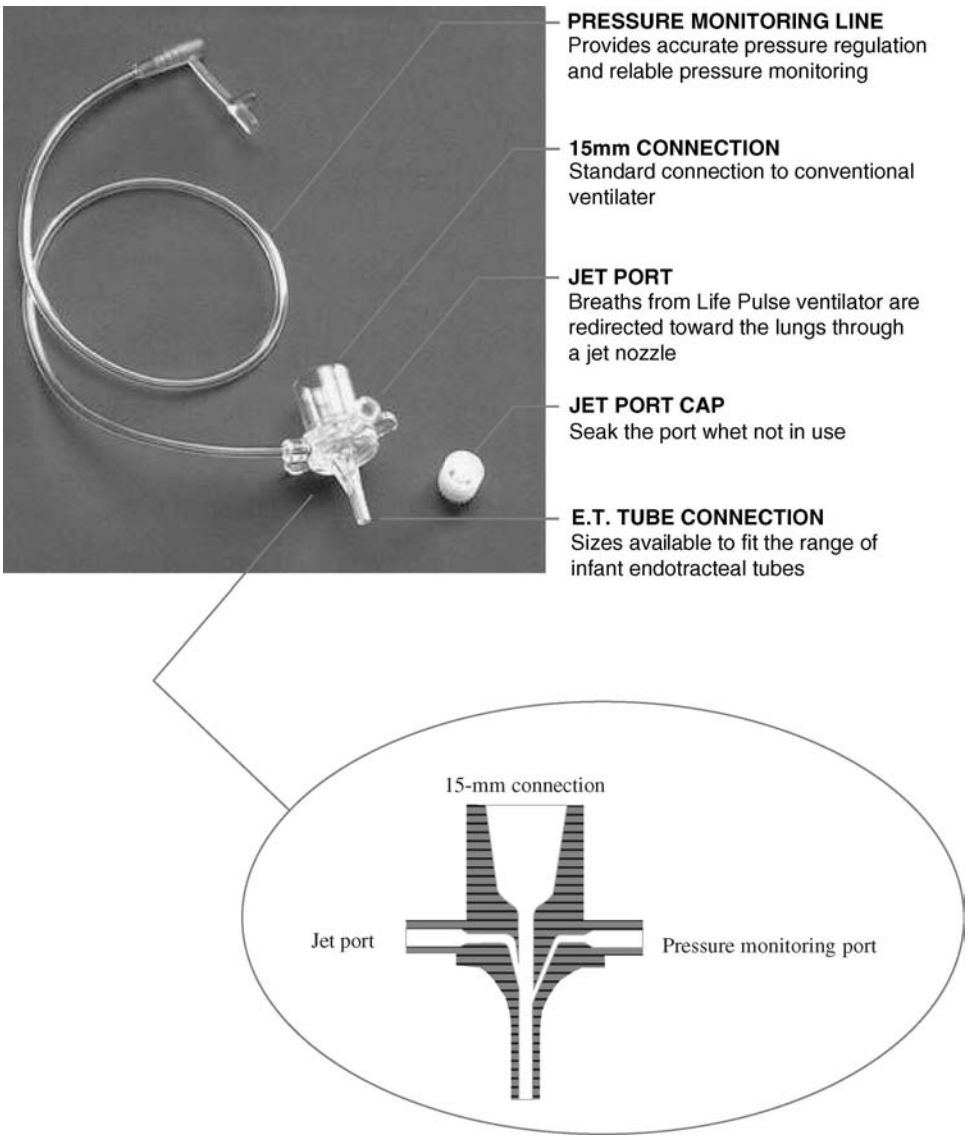


Figure 9. LifePort ET tube adapter for HFJV. (Used with permission. © 2003, Bunnell Inc.)

A low limit alarm would infer that the patient's lung compliance or airway resistance has worsened, and the LifePulse is using less gas (i.e., smaller tidal volumes) to reach the set PIP in that circumstance. A high limit alarm would infer that the patient's condition has improved, larger tidal volumes are being delivered, and the operator should consider weaning PIP in order to avoid hyperventilation.

Alarm limits are also automatically set around monitored mean airway pressure.

HFOV: High Frequency Oscillatory Ventilators

The SensorMedics 3100A HFOV for infants and children and its sister model, the 3100B for adults, are the only pure HFOVs currently available in the United States. Sinusoidal oscillatory ventilation is produced by an electromagnetically driven floating piston with adjustable frequency and amplitude. Inspiratory gas is supplied as bias flow, which escapes from the very large diameter (1.5 in. ID, 38 mm) patient breathing circuit via a traditional dome valve that controls mean airway pressure. All pressures are monitored at the connection to the ET tube. The SensorMedics 3100A HFOV is illustrated in Fig. 10 and more information is available at www.sensormedics.com.

The 3100 HFOVs operate on the principle that high frequency oscillations that are in tune with the natural frequency of the patients lungs will preferentially move in and out of the lungs, as opposed to the exhaust system of the patient circuit. Haselton and Scherer illustrated a new gas transport principle that applies to HFOV (16). Differences in the velocity profiles of inspiration and expiration during HFOV created by the branching architecture of the lungs enables inspiratory gas to advance down the center of the airways while exhaled gas moves up along the airway walls as the piston of the HFOV pulls the gas back. The net effect of many oscillations is similar to, but less pronounced than, the flow characteristics of HFJV flow in the airways. Fresh gas tends to flow down the center of the airways while exhaled gas recedes back along the airway walls.

Figure 10. SensorMedics 3100A high frequency oscillatory ventilator. (Used with permission. © 2003, SensorMedics Inc.)

The 3100 HFOVs have six control settings:

1. Frequency, which is adjusted to suit patient size and lung time constants.
2–4. Bias gas flow rate, Mean Pressure Adjust and Mean Pressure Limit, which together set mean airway pressure.
5. Power, which sets ΔP.
6. % Inspiratory Time, which sets I:E (inspiratory to expiratory time ratio; typically set at 33%).

Mean airway pressure is the primary determinant of oxygenation, and ΔP is the primary determinant of tidal volume and ventilation (CO_2 removal). However, all controls are open-loop: increasing frequency decreases tidal volume and visa versa, and changing bias gas flow rate or power may change mean airway pressure. Mean airway pressure is monitored in the HFOV circuit, so changes there are apparent, but changes in tidal volume due to setting changes or changes in a patient's lung mechanics are not apparent. Given the squared relationship between tidal volume and CO_2 removal noted above, changes in frequency move $PaCO_2$ in the opposite direction of what one would anticipate with conventional ventilation. (Increasing frequency increases $PaCO_2$; decreasing frequency decreases $PaCO_2$.) Thus, continuous or frequent monitoring of arterial $PaCO_2$ is recommended during HFOV, as it is with all HFVs and conventional ventilation of premature infants due to the potential for cerebral injury associated with hyperventilation (more on that topic later).

HFFIs and Other Hybrids

Conventional infant ventilators with built-in HFV modules, such as the Dräger Babylog 8000 *plus* (Dräger Medical AG & Co. KGaA) and the popular Infant Star Ventilator, which will no longer be supported by its manufacturer after May 2006, have been widely used in the United States, Canada, Europe, and Japan. In general, these hybrid HFVs are not as powerful as the stand-alone HFVs, so their use is limited to smaller premature infants (<2 kg). The VDR Servolator Percussionator (Percussionaire Corporation, Sand Point ID), however, was designed to ventilate adults as well as premature infants. (Detailed information on these devices can be viewed on the manufacturers' websites: www.draeger-medical.com and www.percussionaire.com.) The mechanical performances of these devices vary widely, as do their complexities of operation and versatilities as infant ventilators. (See the section Equipment Limitations.)

Design Philosophy for Clinical Applications

The philosophy for controlling arterial blood gases with HFVs is similar to that used with pressure-limited conventional ventilation, especially when HFVs are used to treat homogeneous lung disorders such as RDS (respiratory distress syndrome) in prematurely born infants. The alveoli of these surfactant-deficient lungs must be opened with some type of recruitment maneuver and kept open with appropriate mean airway pressure (*Paw*) or PEEP in order for the lungs to make oxygen available to the blood stream. Ventilation is accomplished at a frequency proportionate to the patient's size and lung mechanics using a peak airway pressure or pressure amplitude above PEEP that creates a tidal volume that produces an appropriate arterial PCO_2. Pulse oximeters, which report the oxygen percent saturation of arterial blood, are great indirect indicators of when lungs have opened up, because oxygenation is highly dependent on the number of alveoli that are open and participating in gas exchange with the blood stream. Chest wall motion is a good indirect indicator of ventilation since it reflects the amount of gas that is passing in and out of the lungs.

The usual approach to initiation of HFV is to choose a frequency that is appropriate for the size of the patient and his lung mechanics, starting with 10 Hz or 600 bpm for the smallest premature infant with RDS and working downward as the size of the patient increases and lung mechanics improve. A good rule of thumb is to choose a frequency 10 times greater that the patient's normal breathing frequency, which would put HFV for adults at rates <200 bpm. Higher rates may be used with HFOV since exhalation is nonpassive, but gas trapping can still result unless mean airway pressure is kept high enough to keep the airways open during the active exhalation phase. Operating an HFOV with a 33% inspiratory time ($I{:}E = 1{:}2$) lessens negative pressure during exhalation compared to longer I-times (e.g., $I{:}E = 1{:}1$) thereby decreasing the potential for causing airway collapse.

With HFJV, the shortest possible inspiratory time (~0.020 s) usually works best; it maximizes inspiratory velocity, which helps reduce effective dead space, and minimizes $I{:}E$, which allows more time for exhalation to avoid gas trapping. These characteristics also minimize mean airway pressure, which is very useful when treating airleaks and for ventilation during and after cardiac surgery. The high velocity inspirations also enable ventilation of patients with upper airway leaks and tracheal tears.

Treatment of obstructive lung disorders absolutely requires longer exhalation times, so HFV must be used at lower frequencies on these patients. HFJV *I*:*E* varies from 1:3.5 to 1:12 as frequency is reduced from 660 to 240 bpm when inspiratory time is held constant at its shortest value.

The HFV is not intended and may in fact be contraindicated for patients with asthma, unless helium–oxygen mixtures become part of the mix (17).

Once a frequency and duty cycle (% *I*-time or *I*:*E*) is chosen, airway pressure settings (PIP, PEEP, or ΔP) are set to provide HFV tidal volumes that noticeably move the chest. If chest wall movement is not apparent, ventilation is probably not adequate. Use of transcutaneous CO_2 monitoring is of great benefit here.

Finally, mean airway pressure (*Paw*) or PEEP must be optimized. Too little *Paw* or PEEP will lead to atelectasis and hypoxemia, and too much *Paw* or PEEP will interfere with cardiac output. One of the true benefits of HFV, however, is that higher *Paw* and PEEP can be used without increasing the risk of iatrogenic lung injury. (The small HFV tidal volumes do not create the same potential for creating alveolar "stretch" injury as larger CV tidal volumes do.) Pulse oximeters can be great indirect indicators of appropriate lung volume, but one must be vigilant in detecting signs of decreased cardiac output.

Conventional ventilation is sometimes required or available for tandem use with certain HFVs. The CV breaths are most useful with nonhomogeneous lung disorders and to facilitate alveolar recruitment with atelectatic lungs. The usual strategy is to reduce CV support when starting HFV (assuming the patient is on CV prior to HFV) to 5–10 bpm while optimal *Paw* and PEEP is being sought, and then reduce CV support further.

Now some of the performance differences in HFV equipment and how those differences may affect successful HFV implementation will be examined.

HFV Equipment Limitations

There have been few head-to-head comparisons of HFV equipment. The most recent comparison were by Hatcher et al. and Pillow et al. where they compared several neonatal HFOVs and found wide variations in performance, complexity, and versatility (18,19). Pillow et al. concluded that the clinical effects of manipulating ventilator settings may differ with each HFOV device. In particular, the pressure amplitude required to deliver a particular tidal volume varies with device, and the effect of altering frequency may result in very different effects on tidal volume and $P\text{a}CO_2$.

The first rigorous analysis of HFVs was undertaken by Fredberg et al. in preparation for the HiFi Study (20). They bench tested eight HFVs in an effort to provide the clinicians who were to participate in the study comparative data that they could use to select an HFV for use in their study. (They selected the Hummingbird HFOV, manufactured by MERA of Japan.) Despite the wide diversity of ventilator designs tested, certain common features emerged. In almost all devices, delivered tidal volume was sensitive to endotracheal tube size and airway resistance and invariant with respiratory system compliance. These results supported the theoretical basis for why high frequency ventilation may be a better treatment for RDS compared to pressure-limited CV (conventional ventilation), because low lung compliance is its paramount pathophysiologic feature.

These HFV bench tests also found that tidal volume decreased with increasing frequency with all HFOVs where I:E (inspiratory to expiratory time ratio) was held constant and was invariant with HFJV and HFFI devices where I-time was held constant. Peak inspiratory flow rates for a given tidal volume and frequency were significantly higher with the HFJV and HFFI as well. Proximal airway pressure was also a poor indicator of distal pressure with all devices.

Two other studies compared HFJV to HFOV. Boros and associates compared the pressure waveforms measured at the distal tip of the endotracheal tube of the Bunnell LifePulse HFJV and the Gould 4800 HFOV (precursor to the SensorMedics 3100A HFOV) in normal, paralyzed, and anesthetized cats (21). They found that the HFOV required higher PIP, ΔP, and *Paw* to get the same $P\text{a}CO_2$, $P\text{a}O_2$, and pH compared to HFJV. Likewise, $PaCO_2$ was higher and pH and $P\text{a}O_2$ were lower with HFOV when the same airway pressures were used. However, different frequencies were used with the two ventilators; 400 bpm with HFJV and 900 bpm (15 Hz) with HFOV.

Zobel and associates also found that HFJV was effective at lower airway pressure compared to HFOV (22). They used a piglet model of acute cardiac failure and respiratory failure and also measured airway pressure at the distal tip of the endotracheal tube. The HFJV used was an Acutronic AMS-1000® (Acutronic Medical Systems AG, Switzerland) operating at 150 bpm with an *I*:*E* of 1:2. The HFOV was a SensorMedics 3100A operating at 10 Hz and 1:2.

Why do HFOVs (presumably) operate at higher *Paw* compared to HFJV? The answer to this question may be related to gas trapping, HFV rates, and what happens during exhalation. In both of the animal studies just discussed, HFJV rate was considerably lower than HFOV rate. Exhalation is passive during HFJV, so lower rates must be employed to allow sufficient exhalation time to avoid gas trapping. The HFOVs suck the gas back out of the lungs during the expiratory phase, and the physiologic consequence can be, not surprisingly, airway collapse. However, the *Paw* employed during HFOV determines the importance of this effect.

Bryan and Slutsky set the tone for the future of HFVs when they noted that this mode of ventilation is ideally designed for treatment of patients with poor lung compliance (23). The higher *Paw* required to match the pathophysiology of such patients also serves to splint the airways open during HFOV so that the choking effect of active expiration is mitigated.

In conclusion, both modes of HFV can cause gas trapping; they just do it by different mechanisms. The HFOV can choke off airways when *Paw* is insufficient to mitigate the effect of active expiration, and HFJV will trap gas when expiratory time is insufficient to allow complete exhalation of inspired tidal volume. One cannot lower *Paw* during HFOV beyond the point where choking is made evident by

a rise in a patient's PCO_2. With HFJV, one should not increase frequency beyond the point where PEEP *monitored* in the endotracheal tube, as it is with the Bunnell LifePulse, begins to rise inadvertently. If the automatically set upper alarm limit on mean airway pressure with the LifePulse is activated, there is a good chance that this rise in *Paw* is due to inadvertent PEEP. The remedy for that circumstance is to decrease HFJV frequency, which lengthens exhalation time and allows the PEEP to fall back to *set* level.

Airway Pressure Monitoring During HFV

While airway pressures were monitored at the distal tip of the ET tube in the animal studies noted above, monitoring at this location is seldom done currently, because the Hi-Lo ET tubes (formerly manufactured by Mallinckrodt, Inc.) are no longer available. Thus, airway pressure monitoring is done either at the standard ET tube adapter connection during HFOV or at the distal tip of the special LifePort adapter during HFJV. In either case, the pressure waveform measured deep in the lungs at the alveolar level is greatly damped (Fig. 11). Gerstmann et al. reported that measurement of pressure amplitude in the alveoli of rabbits during HFOV at 15 Hz was only 10% of that measured proximal to the ET tube (24).

Meaningful monitoring of airway pressure during HFOV is limited to mean airway pressure, and that is only representative of mean alveolar pressure in the absence of gas trapping, as noted above. Relative values of pressure amplitude at the proximal end of the ET tube are indicative of tidal volume size, and they are typically expressed as such by the various HFOVs.

Peak inspiratory pressure (PIP) and PEEP as well as *Paw* are measured during HFJV at the distal tip of the LifePort ET adapter. The PEEP is representative of alveolar PEEP at this location in the absence, again, of gas trapping. However, the PIP at this location is a gross overestimate of peak pressure in the alveoli. Mean airway pressure may slightly overestimate mean alveolar pressure as shown by the study of Perez-Fontan et al. (25).

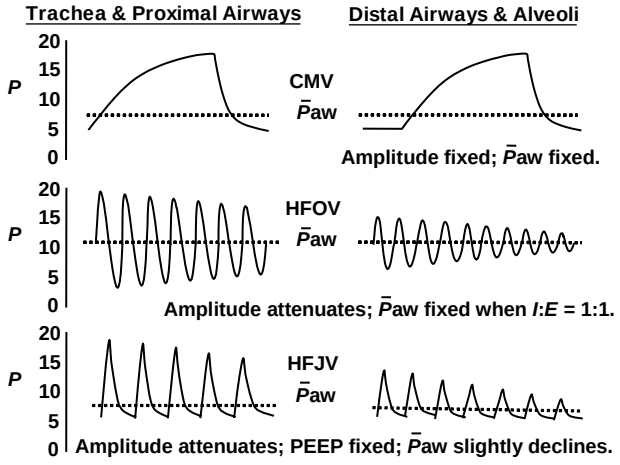


Figure 11. HFV Airway Pressure Waveform Dampening. (Used with permission. © 2003, Bunnell Inc.)

HFV APPLICATIONS IN NEONATES AND CLINICAL OUTCOMES

Homogeneous Atelectatic Lung Disease (e.g., RDS) and Prevention of Lung Injury

Ever since the completion of the first multicenter, randomized, controlled HFV trial was published in 1989, reporting no benefit for premature infants with RDS and an increased risk of severe cerebral injury (26), the choice of HFV to prevent lung injury in preterm infants has been hotly debated. Some recent trails have demonstrated that if HFVs are implemented within hours of a premature infant's birth with the proper strategy, results are positive. Other recent studies have not been positive.

The HiFi Trial, as the first multicenter, randomized, controlled trial was labeled, was criticized for the general lack of clinical experience of the investigators and failure to adhere to the most appropriate strategy for recruiting and maintaining appropriate lung volume (15). Later multicenter, randomized controlled trials conducted in the 1990s using both HFJV and HFOV demonstrated significant reductions in chronic lung disease (CLD) measured at 36 weeks postconceptional age (PCA) in this patient population with practically no difference in adverse effects (27,28). (There was a slightly higher incidence of PIE in the experimental group of the HFOV study.) The demographics and results of these two trials are illustrated in Tables 1 and 2.

The results of the HFJV study were criticized for a lack of well-defined ventilator protocols for the conventionally ventilated control group, whereas protocols for both the HFOV and SIMV control groups in the HFOV study, conducted several years later, were well conceived and monitored during the study. Therefore, it is interesting to note that the major outcome measures of CLD at 36 weeks PCA in the control groups of the two studies were almost identical.

Other HFV studies revealed an increase in severe cerebral injury that appears to be related to hyperventilation and hypocarbia during HFV (29–32). Other criticisms of recent trials with negative or equivocal results include the same strategy issues plus choice of HFV devices, limited time on HFV before weaning back to CV, and so on (33).

Because of these mixed results, HFVs have yet to be generally accepted for early treatment of premature infants with RDS and prevention of lung injury.

Table 1. Demographics of Two Multicenter, Randomized Controlled Trials with HFOV and HFJV

Design/Demographics	HFJV Study[a]		HFOV Study[10]	
Treatment Groups	HFJV	CV	HFOV	SIMV
Number of Patients	65	65	244	254
Mean Birth Weight, kg	1.02	1.02	0.86	0.85
Mean Gestational Age	27.3	27.4	26.0	26.1
Age at Randomization, h	8.1	8.3	2.7	2.7
1 min/5 min Apgar Scores	3.5/7	4/7	5/7	5/7
F_IO_2 at Entry	0.62	0.69	0.57	0.60
Mean Airway Pressure at Entry	10	10	8.2	8.3

[a]See Ref. 9.

Table 2. Significant Respiratory and Clinical Outcomes of HFOV and HFJV Early Application Trials on Premature Infants with RDS

Significant Respiratory and Clinical Outcomes	HFOV Study		HFJV Study	
	HFOV	SIMV	HFJV	CV
Alive w/o CLD at 36 weeks PCA	56%	47%	68%	48%
	$p = 0.046$		$p = 0.037$	
Age at extubation, days	13	21	-	
	$p < 0.001$			
Crossovers or Exits[a]	25/244 (10%)	49/254 (19%)	3/65 (5%)	21/65 (32%)
	$p = 0.07$		$p < 0.01$	
Success after Crossover			14/21 (67%)	0/3 (0%)
			$p = 0.06$	
Supplemental O_2	27%	31%	5.5%	23%
	$p = 0.37$		$p = 0.019$	
PIE	20%	13%		
	$p = 0.05$			
Pulmonary Hemorrhage	2%	7%	6.3%	10%
	$p = 0.02$		$p > 0.05$	

[a]Similar failure criteria were prospectively defined in both studies. Those who met the criteria in the HFJV study were crossed over to the other mode, while those who met the criteria in the HFOV study exited the study and were treated with whatever mode of ventilation the investigators deemed appropriate, including HFJV (personal communication, David Durand, MD). Data on all patients were retained in their originally assigned group in both studies.

Homogeneous Restrictive Lung Disease (e.g., Congenital Diaphragmatic Hernia)

While theories support use of HFV in cases where the lungs are uniformly restricted by acute intra-abdominal disease or postsurgically in infants with congenital diaphragmatic hernia, omphalocele, or gastroschisis, there are no randomized controlled trials due to the rarity of these disorders. Despite this lack of controlled trials, HFV has been widely accepted as an appropriate treatment for this category of lung disease due to the futility of CV treatment in severe cases.

Keszler et al. demonstrated improved gas exchange and better hemodynamics with HFJV in an animal model of chest wall restriction (34) and later reported improved ventilation and hemodynamics in a series of 20 patients with decreased chest wall compliance (35). Fok et al. reported improved gas exchange with HFOV in eight similar patients who were failing CV (36).

Nonhomogeneous Atelectatic and Restrictive Lung Disease (e.g., RDS with Tension PIE)

Pulmonary interstitial emphysema (PIE) in the premature infant creates a non-homogeneous lung disease: parts of the lungs are collapsed as a result of surfactant deficiency while other parts become overexpanded with gas trapped in interstitial areas. Air leaks like PIE originate most commonly in premature infants near the terminal bronchial (37). As gas dissects into interstitial spaces, it invades and dissects airway and vascular walls moving towards the larger airways and vessels and the pleural space where pneumothoraces are formed (38). While positive-pressure CV may successfully penetrate such restricted airways, the consequence may well be accumulation of trapped gas in the alveoli and subsequent alveolar disruption, which produces the classical picture of PIE on X ray.

The HFJV quickly gained a reputation for superior treatment of PIE in the early days of its clinical application. A multicenter randomized trial of HFJV compared to rapid rate (60–100 bpm), short *I*-time (0.20–0.35 s) CV for the treatment of PIE confirmed anecdotal findings of faster and more frequent resolution of PIE on HFJV. Survival in the stratified group of 1000–1500 g birth weight infants was most evident (79% with HFJV vs. 44% with CV; $p < 0.05$). There was no difference in the incidence of adverse side effects.

There is, as yet, no comparable randomized trial of HFOV treatment for PIE. While anecdotal success has been reported, attempts to show an advantage with HFOV in a randomized controlled trial have so far been unsuccessful. It may be that the physical characteristics of the two types of HFVs coupled with the pathophysiologic characteristics of PIE are the reasons for this lack of success. Recall that one difference between HFV devices reported in the pre-HiFi bench studies by Fredberg et al. was that HFJVs squirt gas into the lungs at much higher flow rates compared to HFOV. That fact may make HFJV more sensitive to airway patency compared to HFOV.

Since CV breath distribution may be more affected by lung compliance while HFV breaths may be more affected by airway resistance, especially HFJV breaths with their high velocity inspirations, the distribution of ventilation in the nonhomogeneous PIE lung may be markedly affected by mode of ventilation. While the path of least resistance for CV breaths may lead to more compliant, injured areas of the lungs, HFJV breaths may automatically avoid injured areas where airway and vascular resistances are increased. Therefore, HFJV breath distribution may favor relatively normal airways in the uninjured parts of the lungs where ventilation/perfusion matching is more favorable.

The CV tidal volumes delivered with higher PEEP and *Paw* may dilate airways enough to help gas get into

restricted areas in babies with PIE, but those larger tidal volumes take longer to get back out. Much smaller HFV tidal volumes are more easily expired, especially those that were unable to penetrate the restricted airways where the lungs are injured.

Upper Airway Fistulas and Pneumothoraces

Theoretically, the small tidal volumes, high inspiratory velocities, and short inspiratory times of HFJV are ideally suited for treating pneumothoraces and broncho-pleural and tracheal-esophageal fistulae. Gonzalez et al. found that gas flow in chest tubes, inserted in a series of infants with pneumothoraces, dropped an average of 54% when six infants were switched from CV to HFJV (39). Their mean $PaCO_2$ dropped from 43 to 34 Torr at the same time that their peak and mean airway pressures measured at the distal tip of the ET tube dropped from means of 41–28 and 15 to 9.7 cm H_2O, respectively.

Goldberg et al. (40) and Donn et al. (41) similarly reported improved gas exchange and reduced flow through tracheal–esophageal fistulas.

Homogeneous Obstructive Lung Disease (e.g., Reactive Airway Disease, Asthma)

The HFV should theoretically not be of much benefit in treating lung disorders such as asthma wherein airway resistance is uniformly increased. Low rates and long expiration times should be more effective. However, recent work with HFJV and helium-oxygen mixtures (heliox) demonstrated interesting potential for treating such disorders in patients requiring no more than 80% oxygen.

Tobias and Grueber improved ventilation in a one-year old infant with respiratory syncytial virus and progressive respiratory failure related to bronchospasm with HFJV by substituting a mixture of 80% helium/20% oxygen for compressed air at the air/oxygen blender (42). They hypothesized that the reduced density of helium compared to nitrogen enhanced distal gas exchange. Gupta and associates describe another case where HFJV and heliox rescued a 5 month old infant with acute respiratory failure associated with gas trapping, hypercarbia, respiratory acidosis, and air leak (43). The combination of HFJV with heliox led to rapid improvements in gas exchange, respiratory stabilization, and the ability to wean the patient from mechanical ventilation.

Nonhomogeneous Obstructive Lung Disease (e.g., MAS) and ECMO Candidates

Clinical studies of infants with meconium aspiration syndrome (MAS) provide support for the use of HFV with this type of lung disease. These patients are potential candidates for extracorporeal membrane oxygenation (ECMO), so ability to avoid ECMO is a typical outcome variable in such studies.

Clark et al. randomized 94 full-term infant ECMO candidates to HFOV or CV in a multicenter study (44). Prospectively defined failure criteria were met by 60% of those infants randomized to CV while only 44% of those randomized to HFOV failed. Cross-overs to the alternate mode by those who failed were allowed, and 63% of those who failed CV were rescued by HFOV, while only 23% of those who failed HFOV were rescued by CV. (The latter comparison was statistically significant.) Overall, 46% of the infants who met ECMO criteria required ECMO.

A similar single-center study of HFJV versus CV involved 24 ECMO candidates with respiratory failure and persistent pulmonary hypertension of the newborn (PPHN) (45). Most of the infants in the HFJV-treated group (8 of 11) and 5 of 13 of the conventionally treated infants had either MAS or sepsis pneumonia. Treatment failure within 12 h of study entry occurred in only two of the HFJV-treated infants versus seven of the conventionally treated infants. The ECMO was used to treat 4 of 11 HFJV infants versus 10 of 13 control infants. Zero of nine surviving HFJV-treated infants developed chronic lung disease compared to four of 10 surviving controls ($p = 0.08$). Survival without ECMO in the HFJV group was 5 of 11 (45%) versus 3 of 13 (23%) in the control group. There was no statistical significance in any of these comparisons due to the small number of patients.

The degree to which pathophysiology predicts positive outcomes with respect to the ability of HFVs to rescue infants that become ECMO candidates has been explored in two additional clinical studies. Baumgart et al. evaluated their success with HFJV prior to instituting an ECMO program in 73 infants with intractable respiratory failure who by age and weight criteria may have been ECMO candidates (46). They found survival after HFJV treatment to be much higher in infants with RDS and pneumonia (32/38, 84%) compared to MAS/PPHN (10/26, 38%) or congenital diaphragmatic hernia (3/9, 33%). All patients initially responded rapidly to HFJV as measured by oxygen index (O.I., calculated as mean airway pressure in cm H_2O multiplied by fraction of inhaled O_2 divided by P_{aO2} in Torr). However, that improvement in survivors was realized and sustained during the first 6 h of HFJV treatment.

Paranka et al. studied 190 potential ECMO candidates treated with HFOV during 1985–1992 (47). All patients were born at 35 weeks gestational age or more and developed severe respiratory failure, as defined by an arterial to alveolar oxygen ratio ($P_{(A-a)O_2}$) < 0.2 or the need for a peak pressure of >35 cm H_2O on CV. Fifty-eight percent (111 patients) responded to HFOV and 42% (79 patients) were placed on ECMO. Gas exchange improved in 88% of the infants with hyaline membrane disease (RDS), 79% of those with pneumonia, 51% with meconium aspiration, and 22% of those with congenital diaphragmatic hernia. They also found failure to demonstrate an improvement in $P_{(A-a)O_2}$ after six hours on HFOV to be predictive of failure.

During and After Cardiac Surgery

The ability of HFJV to hyperventilate while using lower mean airway pressure is a great asset when treating patients with cardiac problems. During surgery, the small tidal volumes and low mean airway pressure allow the surgeon to move the lungs out of the way, in order to visualize and work on the heart. After surgery, HFJV can gently hyperventilate the patient to encourage increased pulmonary blood flow while mean airway pressure is kept down (48–51).

PPHN and Nitric Oxide Therapy

Kinsella et al. demonstrated the potential of HFV to enhance delivery of nitric oxide (NO) for the treatment of PPHN in a large, multicenter, randomized controlled trial (52). Nitric oxide delivered with HFOV to infants with significant parenchymal lung disease was more effective than NO delivered by CV. NO has also been delivered successfully with HFJV (53). However, NO must be administered via the HFJV circuit in order for the patient to realize any beneficial effect from the gas (54). Inhaled NO does not work with HFJV when administered exclusively through the conventional ventilator circuit (55).

HFV APPLICATIONS IN CHILDREN AND ADULTS

While the bulk of the research and application of HFV has been aimed at the benefit of infants to date, the sheer number of potential applications for children and adults is far greater. Unfortunately, the number of HFVs available to treat adults is severely limited. There is only one instrument currently available in the United States specifically designed for ARDS in children and adults, the SensorMedics 3100B. (The Percussionaire VDR4-F00008 ventilator also provides HFV for adults. It was approved as a Class II device by the FDA.)

Acute respiratory distress syndrome is the obvious target for HFV treatment in adult intensive care. This syndrome affects ~50 per 100,000 population with a mortality of 30–50%. It is a clinical syndrome of noncardiogenic pulmonary edema associated with pulmonary infiltrates, stiff lungs, and severe hypoxemia (56). Although the pathology of ARDS involves a number of features similar to RDS in infants, such as hyaline membranes, endothelial and epithelial injury, loss of epithelial integrity, and increased alveolar-capillary permeability, it may have a much greater inflammatory component.

The only treatment shown to positively impact mortality over the past several decades came from the ARDSnet Trial where CVs were used with a low tidal volume ventilatory strategy designed to reduce iatrogenic lung injury (57). Comparative treatments in this multicenter study of 861 patients included an experimental group where mean tidal volumes for the first 3 days of their treatments were $6.2\,mL{\cdot}kg^{-1}$ body weight and a control group where tidal volumes were $11.8\,mL{\cdot}kg^{-1}$. The experimental group had lower mortality and fewer days on mechanical ventilators.

With ARDSnet trial pointing in the general direction of smaller tidal volumes, it is not surprising that recent HFV trials appear very promising, especially since HFV investigators focused on NICU patients and worked their way up the learning curve. The most important lesson learned, and one that took many years to learn in the treatment of infants, was the importance of recruiting and maintaining adequate lung volume during HFV. Adult trials of HFV for ARDS now begin with a *Paw* $5\,cm\,H_2O$ greater than that currently being used with CV. Just as was learned with infants, it is safe to use higher PEEPs and mean airway pressures with HFVs smaller tidal volumes.

HFV Clinical Trails with Children and Adults

The importance of starting early with HFV on adults and children with ARDS was highlighted in several anecdotal and pilot trials. Smith et al. treated 29 children with severe ARDS complicated by pulmonary barotrauma with HFJV (58). Twenty (69%) survived, and the only statistically significant difference between survivors and nonsurvivors was the mean time on CV before initiating HFJV (3.7 days in survivors vs. 9.6 days in nonsurvivors). Fort et al. similarly found that survivors in a pilot study of HFOV for adults with ARDS were on CV 2.5 days before initiation of HFOV, while nonsurvivors were on CV for 7.2 days (59). Expected survival in the pilot study was <20%, actual survival was 47%.

Arnold et al. compared HFOV to CV in children with respiratory failure (60). Optimizing lung volume was emphasized in both the experimental and control groups. The strategy for optimizing lung volume in the CV group was to lengthen inspiratory times and increase PEEP in order to decrease required PIPs. They found significant improvement in oxygenation in the HFOV group as well as a lower need for supplement oxygen at 30 days postenrollment.

A recent prospective trial of HFOV for ARDS had similar results. Mehta et al. treated a series of 24 adults with severe ARDS with HFOV (61). Five of the patients were burn victims. Within 8 h of HFOV initiation, F_IO_2 and $PaCO_2$ were lower and PaO_2/F_IO_2 was higher than baseline values during CV throughout the duration of the trial. An obvious focus was placed on recruiting and maintaining adequate lung volume while on HFOV, since *Paw* was also significantly higher than that applied during CV throughout the HFOV trial. Unfortunately, this increase in *Paw* was associated with significant changes in hemodynamic variables including an increase in pulmonary artery occlusion pressure (at 8 and 40 h) and central venous pressure (at 16 and 40 h), and a reduction in cardiac output throughout the study. Thus, *Paw* may not have been optimized. However, 10 patients were successfully weaned from HFOV and 7 survived. Again, there was a statistically significant difference in the time spent on CV prior to initiation of HFV: 1.6 days for survivors versus 5.8 days for the nonsurvivors.

Noting the importance of early intervention, Derdak et al. designed a multicenter, randomized, controlled trial comparing the safety and effectiveness of HFOV versus CV in adults with less severe ARDS (62). (The authors nicknamed their trial: the MOAT Study.) Inclusion criteria included $PaO_2/F_IO_2 \leq 200\,mmHg$ (26.66 kPa) on $\geq 10\,cm$ H_2O PEEP, and 148 adults were evenly randomized. Applied *Paw* was significantly higher in the HFOV group compared with the CV group throughout the first 72 h. The HFOV group showed improvement in PaO_2/F_IO_2 at <16 h, but this difference did not persist beyond 24 h. Thirty day mortality was 37% in the HFOV group and 52% in the CV group ($p = 0.102$). At 6 months, mortality was 47% in the HFOV group and 59% in the CV group ($p = 0.143$). There were no significant differences in hemodynamic variables, oxygenation failure, ventilation failure, barotraumas, or mucus plugging between treatment groups.

The MOAT Study indicates that HFOV is safe and effective for ARDS, and the FDA approved the SensorMedics 3100B for ARDS. Outcome data from this study are comparable to those of the ARDSnet Trial. The control group in the MOAT study was not ventilated with tidal volumes as small as those used in the experimental group of the ARDSnet trial (6–10 vs. 6.2 $mL \cdot kg^{-1}$), but they were generally smaller than the ARDSnet control group (11.8 $mL \cdot kg^{-1}$). Mortality at 30 days in the MOAT Study was not quite as good as that in the ARDSnet Trial (37 vs. 31%, respectively), but sepsis was much more prevalent in the MOAT Study compared to the ARDSnet Trial (47 vs. 27%, respectively).

STATUS OF HFV, RISKS, AND OUTLOOK FOR THE FUTURE

Are HFVs Safe and Effective?

Use of HFVs for newborn infants and adults began in the early 1980s. Fifteen randomized controlled trials with infants and about one-half that many randomized studies with children and adults were conducted over the next 20+ years. Over 1000 articles about HFV have been published. Yet, there are still questions about HFV safety and efficacy.

There are certainly adequate data to suggest that HFVs are effective in lessening chronic lung injury. The fact that not all studies have been successful in this regard is a reflection of differences in infant populations, ventilator strategies, and devices used. There is little argument that use of antenatal steroids, exogenous surfactant, and ventilator strategies using smaller tidal volumes have greatly improved mortality and morbidity of premature infants.

Not surprisingly, as clinicians have become more successful with HFV and other small tidal volume strategies, the age of viability of premature infants has gone down. Thus, the challenge of preventing chronic lung disease in NICU patients never gets easier, because the patients keep getting more premature.

What Are the Risks Associated with HFV in the NICU?

The greatest controversy in consideration of HFVs as a primary mode of ventilation of premature infants is safety, particularly whether HFV use increases the risk of cerebral injury. Clark et al. evaluated the probability of risk of premature infants suffering from intraventricular hemorrhage (IVH) or periventricular leukomalacia (PVL) by conducting a meta-analysis of all prospective randomized controlled trials of HFV published by 1996 (63). The meta-analysis showed that use of HFV was associated with an increased risk of PVL (odds ratio = 1.7 1.7 with a confidence interval of 1.06–2.74), but not IVH or severe (≥grade 3) IVH. In addition, since the largest study in the group by far was the HiFi Trial (14), where implementation strategy was reputed to be less than optimal, they repeated the analysis without that study. When the results of the HIFI study were excluded, there were no differences between HFV and conventional ventilation in the occurrence of IVH or PVL.

Since 1996, seven additional randomized controlled trials of early use of HFV have been conducted on 1726 patients. Only one of the newer studies demonstrated a possible increased risk of cerebral injury (64), and that study included 273 patients or 16% of the total in these 7 studies. Thus, a more current meta-analysis would be even more convincingly positive today, and one could even say that there is *little* evidence of increased risk of cerebral injury during HFV. Why then, is this matter still controversial?

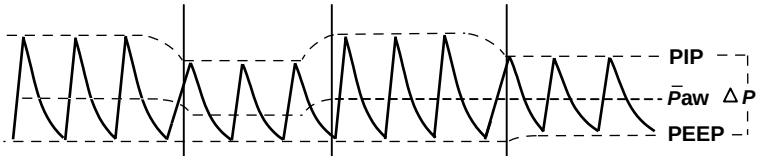
The risk of causing cerebral injury in premature infants is associated with hyperventilation and hypocarbia as noted earlier. There will never be a randomized controlled trial to prove cause and effect here, for obvious reasons. Therefore, all we can do is try to avoid hyperventilation and hypocarbia and see if outcomes get better over time.

Avoiding hyperventilation and hypoxemia first requires proper monitoring. Pulse oximetry, transcutaneous CO_2 monitoring, and continuous or frequent arterial blood gas monitoring are essential during HFV. Control of $PaCO_2$ during HFV often requires optimization of PEEP, *Paw*, and pressure amplitude (ΔP) as shown in Fig. 12. The HFVs are noted for their ease of blowing off CO_2 at lower airway pressures compared to CV, so PEEP and *Paw* must often be increased above those used during CV, if hypoxemia is to be avoided.

With HFOV, one often adjusts mean airway pressure without knowing the resulting baseline pressure or PEEP, whereas with HFJV, PEEP is adjusted to get an appropriate *Paw*. Therefore, one must not be fearful of higher PEEP when higher mean airway pressure is required. PEEP as high as 10 cm H_2O is not unusual when HFJV is being used to treat premature infants.

One must also recognize that raising PEEP will reduce ΔP when PIP is held constant, as shown in Fig. 12, which causes both PaO_2 and $PaCO_2$ to rise.

Other safety concerns with early use of HFVs for preventing lung injury are interference with cardiac output by using too much PEEP or *Paw*. Since interference with



Higher ΔP Higher $\bar{P}aw$	Lower ΔP Lower $\bar{P}aw$	Higher ΔP Higher $\bar{P}aw$	Lower ΔP Higher $\bar{P}aw$
Hypocapnia Good PaO_2	Good $PaCO_2$ Hypoxemia	Hypocapnia Good PaO_2	Good $PaCO_2$ Good PaO_2

Figure 12. Adjusting pressure waveforms to correct arterial blood gases. (Used with permission. © 2003, Bunnell Inc.)

venous return by elevated intrathoracic pressure raises intracranial pressure, there is associated fear of causing IVH by this mechanism as well.

A related issue, for those HFVs with that capability, is using too many CV breaths or using overly large CV tidal volumes during HFV. The latter use increases the risk of causing lung injury when HFV is implemented with higher PEEP.

Optimizing PEEP and minimizing the risk of using too many CV breaths during HFV can be achieved at the same time. The following flowchart for finding optimal PEEP during HFJV illustrates this point (Fig. 13).

The flowchart in Fig. 13 is based on the concept that CV breaths will be most effective in opening up collapsed alveoli, while PEEP or baseline pressure will prevent alveoli from collapsing during exhalation. The longer *I* times and tidal volumes of CV breaths provide a greater opportunity to reach the critical opening pressure of collapsed alveoli, and if PEEP is set above the critical closing pressure of those alveoli, they will remain open throughout the ventilatory cycle. Once PEEP is optimized, there is less value in using CV in tandem with HFV.

Although Fig. 13 was designed for use during HFJV, its principles are equally applicable to HFOV when CV may not be available. In this case, mean airway pressure is raised until an improvement in oxygenation makes it apparent that alveolar recruitment has occurred. At that point, it should be possible to decrease mean airway pressure somewhat without compromising oxygenation. However, the appropriate strategy here would be to set a goal for lowering the fraction of inhaled oxygen (F_IO_2) *before* attempting to lower *Paw*. In this way, one should avoid inadvertently weaning *Paw* too fast and risking catastrophic collapse of alveoli. An appropriate F_IO_2 goal in this circumstance might be 0.3–0.4 depending on the vulnerability of the patient to high airway pressures and the magnitude of the mean airway pressure present at the time.

The final risk to be mentioned here will be the greatest risk associated with HFV: inadequate humidification and airway damage. In unsuccessful applications of HFV in infants in the early 1980s, necrotizing tracheal bronchitis (NTB) was frequently noted at autopsy (65). First discovered during HFJV, it was subsequently discovered during HFOV as well (66). Fortunately, the development of better humidification systems coupled with earlier implementation of HFV seems to have eradicated this problem as an extraordinary adverse side effect. None of the 14 randomized controlled trials has found an increase in NTB associated with HFV treatment.

Humidification during HFV is challenging, especially for HFOVs that use high gas flow rates and HFJVs. Gas humidified under pressure will not hold as much water as unpressurized gas, so HFJVs must humidify their inspiratory gas at higher than normal temperatures in order to reach anything near 100% relative humidity at body temperature.

Bunnell Incorporated's HFJV addressed this inherent problem by minimizing the driving pressure behind their jet nozzle using an inspiratory pinch valve in a little box placed near the patient's head. The pinch valve reduces the pressure drop through that part of the system because of

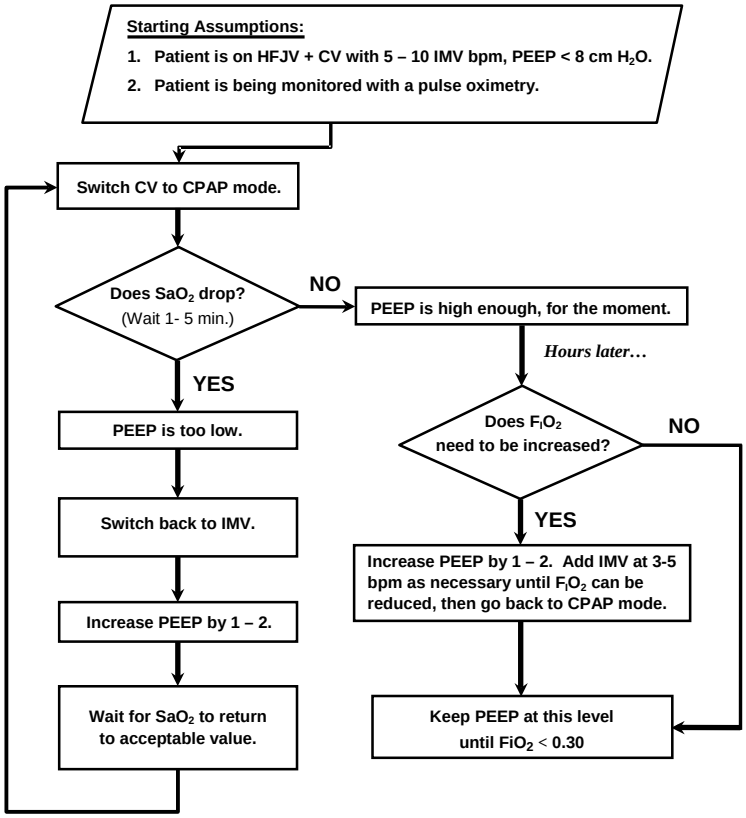


Figure 13. Optimal PEEP flowchart. (Used with permission. © 2005, Bunnell Inc.) (*Warnings*: Lowering PEEP may improve SaO_2 in some cases. Optimal PEEP may be lower in patients with active air leaks or hemodynamic problems. Do not be shocked if optimal PEEP = 8 – 12 cm H_2O. Using IMV PIP with high PEEP is hazardous. Do not assume high PEEP causes overexpansion.)

the relatively large internal diameter (ID) of the tubing in the valve (0.13 in., 3.2 mm). Placing the pinch valve within 35 cm of the patient where inspired gas is delivered via a jet nozzle embedded in a special ET tube adapter also enables the LifePulse Ventilator to deliver its tiny tidal volumes without much driving pressure. (A typical driving pressure needed for HFJV with the LifePulse on a premature infant is between 1.5 and 3.5 psi.) The HFVs that work at higher pressures and gas flow rates sometimes provide humidity with liquid water. (See Acutronic jet ventilation systems on their website: www.acutronic-medical.ch.)

Working Within the Limitations of HFVs in the NICU

The HFVs have been widely accepted for treating newborn infants with lung injury. Whether HFV will ever be widely accepted as a primary mode of ventilation is another matter. There have been many advances in conventional ventilator therapy over the past several years and, to a certain extent, techniques used to optimize HFVs have been applied to CVs (67).

Like most therapies, the skill with which HFV is implemented is probably the most critical determinant of clinical success. Starting HFV on patients sooner rather than waiting for worsening lung disease to become nearly hopeless is also extraordinarily important. Having written protocols defining the patient population and how HFV is optimized for that patient population can also be very helpful.

Early-to-moderately early stage treatment of homogeneously noncompliant lungs is the most obvious choice as an appropriate indication for HFV. If hyperventilation and gas trapping are avoided and appropriate resting lung volume is achieved, better outcomes should result. The ability of all HFVs to ventilate lungs using less pressure, independent of lung compliance, is nearly universal as long as the lungs are not too large for the ventilator's output. Many of the HFV modes built into CVs are not powerful enough to ventilate even a term infant, so operators need to know the relative output capacities of their HFVs.

Using HFVs as rescue ventilators usually means that the underlying lung disorder has become nonhomogeneous or even obstructive. Ironically, most clinicians will only use HFVs as rescue ventilators even though these disorders are much harder to treat with HFV. Gas trapping is a much greater risk with heterogeneous obstructive disorders, and it may be avoided via use of small HFV tidal volumes. However, even HFV tidal volumes can be trapped if expiratory times are not several times greater than inspiratory times. The HFJVs with their very short I:E ratios and very high velocity inspiratory flows have been demonstrated to be more effective with these types of lung disease. Combined CV and HFJV have also been shown to work even better than pure HFJV in severe nonhomogeneous lung disorders (68).

What Are the Risks and Limitations Associated with HFV in Treating Children and Adults?

The only risks unique to HFV for children and adults are those associated with the necessity for delivering larger tidal volumes at higher flow rates compared to HFV for infants. Humidification of HFV gases is crucial as was learned with the early trials in infants, and gas is more difficult to humidify under pressure, as discussed above. The most challenging mode of HFV in this respect is HFJV, since it takes elevated pressure to push gas through a jet nozzle. Perhaps this is one reason for the lack of success of the only HFJV for adults approved by the FDA for use with adults (the APT 1010 Ultrahigh Frequency Ventilator, developed by the Advanced Pulmonary Technologies, Inc., Glastonbury, CT). Humidification with this device was only provided via supersaturated entrained gas. However, the same corporation that pulled this product off the market is also planning to discontinue manufacturing the Infant Star HFV for infants, and that ventilator had no such humidification issues. Thus, it appears likely that these products were discontinued for other (i.e., business) reasons.

There is also danger of operator error when any machine is used by untrained or unskilled operators. Given the tendency for some hospitals to only use HFVs as last resort rescue ventilators, one must consider those types of risks. Since HFOV is most successful in homogeneous lung disorders, it only makes sense for it to be used relatively early in the course of ARDS before significant lung injury results from CV treatment. Once the patient's condition deteriorates into ARDS complicated by airleaks, chances for HFOV success may be significantly decreased.

Treating children and adults with HFVs also has to take optimal frequency into account. The primary determinant of optimal frequency is lung compliance, which is primarily determined by the patient's size. An adult's lung is larger than that of a child, which is larger than that of a term newborn, which is larger than that of a preemie. Thus, HFV frequency should be reduced as patient size increases. Optimal HFV for adults may occur at 150 bpm, whereas operation at that frequency with infants would not even be considered HFV.

Given the evidence that HFJVs have been more successful with nonhomogeneous lung disorders in infants, as described above, it would seem likely that HFJVs would find a role for treating adult patients with ARDS as well. Unfortunately, the application of HFJV for adults has been tarnished by a lack of success in very early studies.

Carlon et al. conducted a randomized controlled trial of HFJV versus volume-cycled CV on adults with acute respiratory failure (69). While they reported patients failing on CV improved more rapidly and in greater number when switched to HFJV compared to those who were failing on HFJV and crossed to CV, there were no advantages with respect to survival and total duration of stay in the ICU. Thus, they concluded that HFJV offered no obvious benefits over CV. The study was published in 1983, long before there was much appreciation of the need to optimize PEEP and maintain adequate lung volume during HFV. One wonders what results would come from a similar trial 20 years later, but such a trial will probably never happen now. The cost, time, and effort of seeking FDA approval for any Class III device is so high now, that HFJV may never find its way back into an adult ICU in the United States.

What Is the Outlook for HFV in the Future?

The HFVs evolved as people discovered problems with trying to replicate breathing in compromised patients. Unlike conventional ventilation, HFV is designed to facilitate gas exchange rather than mimic how people breathe normally. The differences between HFV and CV have led to creative solutions for many of the problems that investigators set out to solve, but they have also created their own problems.

It was discovered how HFVs can ventilate much more effectively than CV, and in the process, it was discovered how hypocapnia can lead to severe cerebral injury in premature infants.

The HFV still uses positive pressure where we create negative pressure to draw gas into the lungs. So, the problems related to use of ET tubes and its bypassing the normal humidification system of the body (i.e., the nose) are still there.

The HFV uses much smaller tidal volumes than CV, so the damage we have come to call volutrauma has lessened. In the process, it was discovered that more mean airway pressure or PEEP is required to keep sick lungs open while they are being ventilated. Then it was discovered that the new problems were associated with too much pressure in the thorax interfering with cardiac output.

So, HFVs do not solve all the problems, and they require increased vigilance to avoid creating new problems. However, the basic differences between HFV and CV provide a very reliable alternative to CV in circumstances where those differences are critical to survival.

The HFV tidal volumes are minimally affected by lung compliance and maximally affected by airway resistance when they are delivered via a jet nozzle. Therefore, in lung disorders where these conditions dictate treatment success or failure, wise users of HFVs have been very successful. When HFV users are not adequately trained or aware of the differences in HFV gas distribution caused by lung pathophysiology, success can be elusive.

Premature infants with RDS represent a large population with the potential of leading long and rich lives if they survive their first months of life with undamaged or even minimally damaged lungs and brains. Many randomized controlled trials have demonstrated the potential of HFVs to help these infants realize that potential.

The tens of thousands of adults who succumb to ARDS every year also have the potential of increased survival with less morbidity thanks to HFVs. These patients, when successfully treated with an HFV, will be considered rescued from a well-recognized disorder with a chronically high mortality rate.

Given the skill and training needed to master HFVs, their use may be considered risky indefinitely. As HFVs and associated monitoring equipment become better designed to help their users optimize assisted ventilation, HFV use should increase and evolve into earlier, more prophylactic applications to prevent lung injury. The HFVs are inherently a kinder, gentler form of mechanical ventilation. Hopefully, their true potential will someday be realized.

BIBLIOGRAPHY

Cited References

1. Lee PC, Helsmoortel CM, Cohn SM, Fink MP. Are low tidal volumes safe? Chest 1990;97:430–434.
2. Kacmarek RM, Chiche J.D. Lung protective ventilatory strategies for ARDS—the data are convincing! Resp Care 1998; 43:724–727.
3. Sjöstrand U. Review of the physiological rationale for and development of high-frequency positive-pressure ventilation—HFPPV. Acta Anaesthesiol Scand (Suppl) 1977;67: 7–27.
4. Klain M. Clinical applications of high-frequency jet ventilation, Part A: clinical use in the operating room. In: Carlon GC, Howland WS, editors. High-frequency ventilation in intensive care and during surgery. New York: Marcel Dekker; 1985. p 137–149.
5. Singh JM, Stewart TE. High-frequency mechanical ventilation principles and practices in the era of lung-protective ventilation strategies. Respir Care Clin N Am 2002;8:247–60.
6. MacIntyre NR. Setting the frequency-tidal volume pattern. Respir Care 2002;47:266–274.
7. Radford E. Ventilation standards for use in artificial respiration. J Appl Physiol 1955;7:451–460.
8. Henderson Y, Chillingworth FP, Whitney JL. The respiratory dead space. Am J Physiol 1915;38:1–19.
9. Dubois AB, Brody AW, Lewis DH, Burgess BF. Oscillation mechanics of lungs and chest in man. J Appl Physiol 1956;8: 587–594.
10. Bunnell JB, Karlson KH, Shannon DC. High-frequency positive pressure ventilation in dogs and rabbits. Am Rev Respir Dis 1978;117:289.
11. Fredberg JJ. Augmented diffusion in the airways can support pulmonary gas exchange. J Appl Physiol 1980;49:232–238.
12. Slutsky AS. Mechanisms affecting gas transport during high-frequency oscillation. Crit Care Med 1984;12:713–717.
13. Venegas JG, Hales CA, Strieder DJ. A general dimensionless equation of gas transport by high-frequency ventilation. J Appl Physiol 1986;60:1025–1030.
14. Permutt S, Mitzner W, Weinmann G. Model of gas transport during high-frequency ventilation. J Appl Physiol 1985;58: 1956–1970.
15. Venegas JG, Fredberg JJ. Understanding the pressure cost of high frequency ventilation: why does high-frequency ventilation work? Crit Care Med 1994;22:S49–S57.
16. Haselton FR, Scherer PW. Bronchial bifurcations and respiratory mass transport. Science 1980;208:69–71.
17. Tobias JD, Grueber RE. High-frequency jet ventilation using a helium-oxygen mixture. Paediatr Anaesth 1999;9:451–455.
18. Hatcher D, et al. Mechanical performance of clinically available, neonatal, high-frequency, oscillatory-type ventilators. Crit Care Med 1998;26:1081–1088.
19. Pillow JJ, Wilkinson MH, Neil HL, Ramsden CA. *In vitro* performance characteristics of high-frequency oscillatory ventilators. Resp Crit Care Med 2001;164:1019–1024.
20. Fredberg JJ, Glass GM, Boynton BR, Frantz 3rd ID. Factors influencing mechanical performance of neonatal high-frequency ventilators. J Appl Physiol 1987;62:2485–2490.
21. Boros SJ, et al. Comparison of high-frequency oscillatory ventilation and high-frequency jet ventilation in cats with normal lungs. Ped Pulmonol 1989;7:35–41.
22. Zobel G, Dacar D, Rodl S. Proximal and tracheal airway pressures during different modes of mechanical ventilation: An animal model study. Ped Pulmonol 1994;18:239–243.
23. Bryan AC, Slutsky AS. Lung volume during high frequency oscillation. Am Rev Resp Dis 1986;133:928–930.

24. Gerstmann DR, et al. Proximal, tracheal, and alveolar pressures during high-frequency oscillatory ventilation in a normal rabbit model. Pediatr Res 1990;28:367–373.
25. Perez Fontan JJ, Heldt GP, Gregory GA. Mean airway pressure and mean alveolar pressure during high-frequency jet ventilation in rabbits. J Appl Physiol 1986;61:456–463.
26. HIFI Study Group, HFOV compared with conventional mechanical ventilation in the treatment of respiratory failure in preterm infants. N Engl J Med 1989;320:88–93.
27. Keszler M, et al. Multicenter controlled clinical trial of high-frequency jet ventilation in preterm infants with uncomplicated respiratory distress syndrome. Pediatrics 1997;100:593–599.
28. Courtney SE, et al. Early high-frequency oscillatory ventilation versus conventional ventilation in very-low-birth-weight-infants. N Engl J Med 2002;347:643–653.
29. Johnson AH, et al. High-frequency oscillatory ventilation for the prevention of chronic lung disease of prematurity. N Engl J Med 2002;347:633–642.
30. Wiswell TE, et al. HFJV in the early management of RDS is associated with a greater risk for adverse outcomes. Pediatrics 1996;98:1035–1043.
31. Wiswell TE, et al. Effects of hypocarbia on the development of cystic periventricular leukomalacia in premature infants treated with HFJV. Pediatrics 1996;98:918–924.
32. Bryan AC, Froese AB. Reflections on the HiFi trial. Pediatrics 1991;87:565–567.
33. Stark AR. High-frequency oscillatory ventilation to prevent bronchopulmonary dysplasia—are we there yet? N Engl J Med 2002;347:682–683.
34. Keszler M, Goldberg LA, Wallace A. High frequency jet ventilation in subjects with low chest wall compliance. Pediatr Res 1993;33:331.
35. Keszler M, Jennings LL. High frequency jet ventilation in infants with decreased chest wall compliance. Pediatr Res 1997;41:257.
36. Fok TF, et al. High frequency oscillatory ventilation in infants with increased intra-abdominal pressure. Arch Dis Child 1997;76:F123–125.
37. Thibeault DW. Pulmonary barotrauma: Interstitial emphysema, pneumomediastinum, and pneumothorax. In: Thibeault DW, Gregory GA, editors. Neonatal Pulmonary Care. 2nd ed. New York : Appleton-Century-Crofts; 1986. p 499–517.
38. Macklin MT, Macklin CC. Malignant interstitial emphysema of the lung and mediastinum as an important occult complication in many respiratory diseases and other conditions: An interpretation of clinical literature in light of laboratory experiment. Medicine 1944;23:281.
39. Gonzalez F, Harris T, Richardson P. Decreased gas flow through pneumothoraces in neonates receiving high-frequency jet versus conventional ventilation. J Pediatr 1987;110:464–466.
40. Goldberg L, Marmon L, Keszler M. High-frequency jet ventilation decreases flow through tracheo-esophageal fistula. Crit Care Med 1992;20:547–550.
41. Donn SM, et al. Use of high-frequency jet ventilation in the management of congenital tracheoesophageal fistula associated with respiratory distress syndrome. J Pediatr Surg 1990;12:1219–1221.
42. Tobias JD, Grueber RE. High-frequency jet ventilation using a helium-oxygen mixture. Paediatr Anaesth 1999;9:451–455.
43. Gupta VK, Grayck EN, Cheifetz IM. Heliox administration during high-frequency jet ventilation augments carbon dioxide clearance. Respir Care 2004;49:1038–1044.
44. Clark RH, Yoder BA, Sell MS. Prospective, randomized comparison of high-frequency oscillation and conventional ventilation in candidates for extracorporeal membrane oxygenation. J Pediatr 1994;124:447–454.
45. Engle WA, et al. Controlled prospective randomized comparison of HFJV and CV in neonates with respiratory failure and persistent pulmonary hypertension. J Perinat 1997;17:3–9.
46. Baumgart S, et al. Diagnosis-related criteria in the consideration of ECMO in neonates previously treated with HFJV. Pediatrics 1992;89:491–494.
47. Paranka MS, Clark RH, Yoder BA, Null DM. Predictors of failure of high-frequency oscillatory ventilation in term infants with severe respiratory failure. Pediatrics 1995;95:400–404.
48. Meliones JN, et al. High-frequency jet ventilation improves cardiac function after the Fontan procedure. Circulation 1991;84(Suppl III):III-364–III-368.
49. Kocis KC, et al. High-frequency jet ventilation for respiratory failure after congenital heart surgery. Circulation 1992;86(Suppl II):II-127–II-132.
50. Greenspan JS, et al. HFJV: Intraoperative application in infants. Ped Pulmonol 1994;17:155–160.
51. Davis DA, et al. High-frequency jet versus conventional ventilation in infants undergoing Blalock-Taussig shunts. Ann Thorac Surg 1994;57:846–849.
52. Kinsella JP, et al. Randomized, multicenter trial of inhaled nitric oxide and high-frequency oscillatory ventilation in severe, persistent pulmonary hypertension of the newborn. J Pediatr 1997;131:55–62.
53. Day RW, Lynch JM, White KS, Ward RM. Acute response to inhaled nitric oxide in newborns with respiratory failure and pulmonary hypertension. Pediatrics 1996;98:698–705.
54. Platt D, Swanton D, Blackney D. Inhaled nitric oxide delivery with high-frequency jet ventilation. J Perinatol: (in press) 2003.
55. Mortimer TW, Math MCM, Rajardo CA. Inhaled nitric oxide delivery with high-frequency jet ventilation: A bench study. Respir Care 1996;41:895–902.
56. Ware LB, Matthay MA. The acute respiratory distress syndrome. N Engl J Med 2000;342:1334–1349.
57. The Acute Respiratory Distress Syndrome Network, Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. N Engl J Med 2000;342:1301–1308.
58. Smith DW, et al. High-frequency jet ventilation in children with the adult respiratory distress syndrome complicated by pulmonary barotrauma. Ped Pulmonol 1993;15:279–286.
59. Fort P, et al. High-frequency oscillatory ventilation for adult respiratory distress syndrome—a pilot study. Crit Care Med 1997;25:937–47.
60. Arnold JH, et al. Prospective, randomized comparison of high-frequency oscillatory ventilation and conventional mechanical ventilation in pediatric respiratory failure. Crit Care Med 1994;22:1530–1539.
61. Mehta S, et al. Prospective trial of high-frequency oscillation in adults with acute respiratory distress syndrome. Crit Care Med 2001;29:1360–1369.
62. Derdak S, et al. High-frequency oscillatory ventilation for acute respiratory distress syndrome in adults: A randomized, controlled trial. Am J Respir Crit Care Med 2002;166:801–808.
63. Clark RH, Dykes FD, Bachman TG, Ashurst JT. Intraventricular hemorrhage and high-frequency ventilation: A meta-analysis of prospective clinical trials. Pediatrics 1996;98:1058–1061.
64. Moriette G, et al. Prospective randomized multicenter comparison of high-frequency oscillatory ventilation and conventional ventilation in preterm infants of less than 30 weeks with respiratory distress syndrome. Pediatrics 2001;107:363–72.

65. Mammel MC, et al. Acute airway injury during high-frequency jet ventilation and high-frequency oscillatory ventilation. Crit Care Med 1991;19:394–398.
66. Kirpilani H, et al. Diagnosis and therapy of necrotizing tracheobronchitis in ventilated neonates. Crit Care Med 1985; 13:792–797.
67. Keszler M, Durand DJ. Neonatal high-frequency ventilation. Past, present, and future. Clin Perinatol 2001;28:579–607.
68. Spitzer AR, Butler S, Fox WW. Ventilatory response to combined HFJV and conventional mechanical ventilation for the rescue treatment of severe neonatal lung disease. Ped Pulmonol 1989;7:244–250.
69. Carlon GC, et al. High-frequency jet ventilation. A prospective randomized evaluation. Chest 1983;84:551–559.

See also CONTINUOUS POSITIVE AIRWAY PRESSURE; RESPIRATORY MECHANICS AND GAS EXCHANGE; VENTILATORY MONITORING.

HIP JOINTS, ARTIFICIAL

Z. M. JIN
J. L. TIPPER
M. H. STONE
E. INGHAM
J. FISHER
University of Leeds
Leeds, United Kingdom

INTRODUCTION

Natural synovial joints, such as hips, are remarkable bearings in engineering terms. They can transmit a large dynamic load of several times bodyweight during steady-state walking, yet with minimal friction and wear achieved through effective lubrication and with little maintenance. However, diseases such as osteoarthritis and rheumatoid arthritis or trauma sometimes necessitate the replacement of these natural bearings. Artificial hip joints replace the damaged natural bearing material, articular cartilage. As a result, pain in the joint is relieved and joint mobility and functions are restored. Total hip joint replacement has been considered as one of the greatest successes in orthopaedic surgery in the last century in improving the quality of life of the patients. Currently, >1 million hip joints are replaced worldwide each year, with ever increasing use of these devices in a wider range of patients.

The majority of current artificial hip joints consist of an ultrahigh molecular weight polyethylene (UHMWPE) acetabular cup against a metallic or ceramic femoral head as illustrated in Fig. 1. These devices can generally last 10–15 years in the body without too many problems. However, after this period of implantation, loosening of prosthetic components becomes the major clinical problem. It is now generally accepted that the loosening is caused by the osteolysis as a result of biological reactions to particulate wear debris mainly released from the articulating surfaces. Therefore, one of the main strategies to avoid the loosening problem and to extend the clinical life of the hip prosthesis is to minimize wear and wear particles. Application of *tribology,* defined as "the branch of science and technology concerned with interacting surfaces in relative motion and with associated matters (as friction, wear, lubrication, and the design of bearings" (Oxford English Dictionary), to biological systems (*biotribology*) such as artificial hip joints, can play an important role in this process. Coupled tribological studies of friction, wear and lubrication of the bearing surfaces, and biological studies of wear debris-induced adverse reactions become necessary.

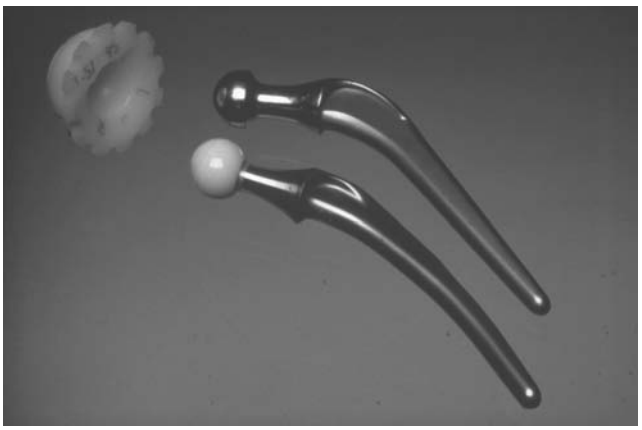


Figure 1. A typical Charnley hip prosthesis consisting of an UHMWPE acetabular cup against either a metallic (stainless steel) or a ceramic (alumina) femoral head.

HISTORICAL DEVELOPMENT

Early History: Hemiarthroplasty, Interposition Arthroplasty, and Total Hip Replacement

The first recognizable ball and socket joint was reported in Germany by Professor Gluck in 1890 in a dog with an ivory ball and socket hip joint. This did not gain popular support for use in humans until Hey Groves in Bristol reported his ivory hemiarthroplasty for fractured neck of femur in 1926. Attempts to use metal at this stage were unsuccessful. A significant breakthrough came in 1923. It began with a chance observation that a piece of glass left in an individual's back for 1 year stimulated a fibrous tissue and fluid producing reaction. It formed a fluid-filled synovial sac (Smith Peterson 1948). Smith Peterson went on to insert a glass cup-shaped mould between the surfaces of an ankylosed hip. Although the glass broke, at the time of its removal the acetabulum and the head of the femur were found to be covered with a smooth lining of fibrous tissue. Over the next few years a number of different materials were used including Viscaloid, Pyrex, Bakelite, and finally Vitallium (chromium–cobalt–molybdenum alloy) in 1938. This material worked well and was used for $\sim$1000 interposition arthroplasties at Massachusetts General Hospital alone over the next 10 years. It remained the standard treatment for hip arthritis until the advent of total hip replacement.

Charnley Era

Jean and Robert Judet reported their use of a replacement femoral head made of poly (methyl methacrylate) (PMMA). Although the prosthesis failed, it survived

long enough to squeak within the human body. It was this squeaking prosthesis that set Charnley on his quest for a low friction-bearing surface. He began with Teflon in 1958 and throughout the 1950s Charnley experimented with two thin cups of Teflon, one in the acetabulum and one over a reshaped femoral head. They failed within a year due to loosening of the cup and avascular necrosis of the femoral head. He abandoned this surface replacement for an endoprosthesis and as the acetabular cups wore through Charnley sought better wearing materials. He moved to high density PE and later to UHMWPE.

Low Friction Arthroplasty. Charnley began his cemented total hip replacement era with a large femoral head (Moore's hemiarthroplasties). He argued that distributing the load over a large area of contact would decrease wear. However, after loosening of these large head components, he began working on the low frictional torque prosthesis, which reduced the torque at the cement-bone and prosthesis interfaces. He achieved this by reducing the diameter of the femoral head from ~ 41 to 22 mm. In this way, Charnley developed his prosthesis of a 22 mm head on a metal stem, UHMWPE cup and PMMA cement. Charnley hips still have a survival today of $> 90\%$ at 10 years (1).

There continues to be debate as to the cause of up to 10% failures of Charnley hips. Early belief that it was due to the cement and cement particles led to the development of uncemented prostheses. Concern that the production of PE particles was producing bone lysis, led to the development of alternative bearing surfaces, for example, ceramics that wear less than metal against PE, and metal-on-metal prostheses, which produce lower volumes of wear debris. Impingement of the femoral prosthesis on the cup leading to loosening led to the narrowing of the neck of the femoral component and the use of cut away angle bore sockets. Concern about access of fluid to the femoral cement–prosthesis interface and subsequently to the cement–bone interface through cement deficiencies, with the production of osteolysis, have led some manufacturers to produce polished femoral stems. These self-locking tapers prevent fluid flowing between the cement and the femoral stem. The debate continues. It can be difficult to separate the improvements in surgical techniques that have improved the clinical results from the effect of modification and changes in materials used to produce joint replacements.

Current Developments

There is currently much interest in reducing the trauma of the surgery itself. These include surgical techniques of minimal incision surgery with the skin wound < 8 cm and a more conservative approach to femoral bone use at the time of surgery. Surface replacement has returned with "better" metal-on-metal surfaces, and a short-stemmed Judet type of metaphyseal fix femoral prosthesis is also available. Hydroxyapatite as a method of component fixation is also gaining popularity. The current short-term results of these techniques are interesting, and may lead to significant benefits especially for the younger patient in the future. However, the proof will only come from long-term clinical results that are not yet available.

JOINT ANATOMY AND ENVIRONMENT

The bearing material in the natural hip joint is articular cartilage, firmly attached to the underlying bone. Articular cartilage is an extremely complex material, consisting of both fluid (interstitial water) and solid (primarily collagen and proteoglycan) phases. Such a biphasic or poroelastic feature determines the time-dependent deformation of articular cartilage, and largely governs the lubrication mechanism of the natural hip joint. The lubricant present in the natural hip joint is synovial fluid, which is similar to blood plasma with hyaluronic acid added and becomes periprosthetic synovial fluid after total hip replacement. Rheological studies of these biological lubricants have shown shear-thinning characteristics, particularly at low shear rates and for the joint fluid taken from diseased or replaced joints (2).

The load experienced in the hip joint during steady-state walking varies both in direction and magnitude. The maximum load can reach five times bodyweight during the stance phase after heel-strike and is largely reduced in the swing phase after the toe-off. On the other hand, the speed is relatively low, particularly in the stance phase and during the motion reversal. However, the hip contact force can be substantially increased under other conditions. For example, the hip contact force has been reported to be 5.8 times bodyweight up a ramp, 6.6 times up and down stairs, and 7.6 times on fast level walking at a speed of 2.01 $\mathrm{m \cdot s^{-1}}$ (3).

CURRENT BEARING SURFACES

Biomaterials used for current artificial hip joints include UHMWPE, stainless steel, cobalt chromium alloy (CoCr), and ceramics (alumina and zirconia). A number of combinations for the bearing surfaces using these materials have been introduced since 1950s in order to minimize wear and wear particle generation. These can generally be classified as soft-on-hard and hard-on-hard as summarized in Table 1.

Two main parameters that govern the tribology of the articulating surfaces are geometrical and mechanical properties. The geometrical parameters of the bearing surfaces are the diameters of the acetabular cup (D_{cup}) and the femoral head (D_{head}). The size of the hip prosthesis is usually characterized by its diameter, which is important for both clinical and tribological considerations, such as stability, dislocation, and sliding distance. In addition to size, the diametral mismatch or clearance between the cup and the head ($d = D_{\mathrm{cup}} - D_{\mathrm{head}}$) is also important, particularly for hard-on-hard bearing surfaces. From a tribological point of view, these geometric parameters can often be approximated as a single equivalent diameter (D) defined as

$$D = \frac{(D_{\mathrm{head}} D_{\mathrm{cup}})}{d} \tag{1}$$

Typical values of equivalent diameter used in current hip prostheses are summarized in Table 2. In addition, geometric deviations from perfectly spherical surfaces, such as nonsphericity and surface toughness, are also very

Table 1. Typical Biomaterials and Combinations for the Bearing Surfaces of Current Artificial Hip Joint Replacements

	Acetabular Cup				
	Soft			Hard	
Femoral Head (Hard)	UHMWPE	Cross-linked UHMWPE	Polyurethane	CoCr	Alumina
Stainless Steel	√	√			
CoCr	√	√	√	√	
Alumina	√	√	√	√	√
Zirconia	√	√	√		√

Table 2. Typical Geometric Parameters of Various Bearing Couples for Artificial Hip Joints[a]

Bearing Couples	Femoral Head Diameter, mm	Diametral Clearance, μm	Equivalent Diameter, m
UHMWPE-on-metal	28 (22–40)	300 (160–1000)	2.6 (1.0–5.0)
Metal-on-metal	28 (28–60)	60 (60–300)	~10
Ceramic-on-ceramic	28 (28–36)	80 (20–80)	~10

[a]See Ref. 4.

important factors in determining the tribological performance of the prosthesis.

The mechanical properties of the bearing surfaces are also important tribological determinants. Typical values of elastic modulus and Poisson's ratio are given in Table 3 for the biomaterials used in artificial hip joints. Other parameters, such as hardness, particularly in the soft-on-hard combinations, are also important, in that the hard surface should be resistant to third-body abrasion to minimize the consequences of polymeric wear.

COUPLED TRIBOLOGICAL AND BIOLOGICAL METHODOLOGY

The vast majority of studies to evaluate the wear performance of hip prostheses have simply measured the volumetric wear rate (6–8). There are very few groups who have also investigated the characteristics of the wear particles generated in *in vitro* simulations (9–11). The cellular response to prosthetic wear particles, and thus the functional biological activity of implant materials, is complex and is dependent not only on the wear volume, but also the mass distribution of particles as a function of size, their concentration, morphology, and chemistry (see the section, Biological Response of Wear Debris).

During the latter 1990s, methods were developed for the isolation and characterization of UHMWPE particles from retrieved tissues and serum lubricants from simulators that allow discrimination between the particles generated in different patient samples and from different types of polyethylene tested *in vitro* (12–18). The basis of this method is to determine the mass distribution of the particles as a function of size. Determination of the number distribution as a function of size fails to discriminate between samples since the vast majority of the number of particles are invariably in the smallest size range detectable by the resolution of the imaging equipment.

Table 3. Typical Mechanical Properties in Terms of Elastic Modulus and Poisson's Ratio of the Bearing Materials for Artificial Hip Joints[a]

Bearing Materials	Elastic Modulus, GPa	Poisson's Ratio
UHMWPE	0.5–1.	0.4
Cross-Linked UHMWPE	0.2–1.2[a]	0.4
Stainless steel	210	0.3
CoCr	230	0.3
Zirconia	210	0.26
Alumina	380	0.26

[a]See Ref. 5.

In our laboratories we have pioneered cell culture studies with clinically relevant UHMWPE wear particles generated in experimental wear simulation systems operated under aseptic conditions (17–21). These studies have been extended to cell culture studies of clinically relevant metal (22), ceramic (23), and bone cement wear particles (24,25).

By combining volumetric wear determinations in hip joint simulations with experiments to determine the direct biological activity of the particles generated, we have developed novel methodologies to evaluate the functional biocompatibility of different materials used in prosthetic joint bearings. The functional biocompatibility can be used as a preclinical estimate of the *in vivo* performance of the material under test compared to historical materials. We have adopted two different approaches to determining functional biocompatibility. Our choice of method is dependent on the bearing material and the type of prosthesis.

The first approach is indirect, but can be applied to all materials and devices. It utilizes data obtained from the direct culture of UHMWPE wear particles in three different size ranges: 0.1–1, 1–10, and >10 μm at different volumetric concentrations with human peripheral blood macrophages. Measurements of the biological activity for unit volumes of particles in the different size ranges are generated (20). The use of TNF-α as a determinant is justified since, in our experience the major cytokines concerned in osteolysis (TNF-α, IL-1, IL-6, GM-csf) all show the same pattern of response to clinically relevant wear particles (19–21). By using our methods to determine the

volumetric concentration of particles generated in simulations as a function of size (see above), it is then possible to integrate the volume concentration and biological activity function to produce a relative index of specific biological activity (SBA) per unit volume of wear. The functional biological activity (FBA) has been defined as the product of volumetric wear and SBA (26). This has allowed us to compare the functional biological activity of different types of PE in hip joint simulators (27) and different types of bearing materials (23).

The second approach is to directly culture wear debris from wear simulators with primary macrophages. For metal and ceramic particles, we can directly culture wear particles from standard simulation systems after isolation, sterilisation, and removal of endotoxin by heat treatments (23). However, for PE this is not feasible since the heat treatment at elevated temperature required to remove endotoxin cannot be applied. For these materials, we have developed a sterile endotoxin free multidirectional wear simulator in which wear particles are generated in macrophage tissue culture medium. While this does not test whole joints, it allows the application of different kinematics to represent the hip and the knee. The advantage of this approach is that all the wear products are directly cultured with the cells, and there is no risk of modification during the isolation procedure. This method has recently been used to compare the biological reactivity of particles from PEs of different molecular weights and different levels of cross-linking. Higher molecular weight of GUR 1050 and higher levels of cross-linking of both GUR 1020 and 1050 produced particles that were more biologically reactive (18).

Tribology of Bearing Surfaces

Tribological studies of the bearing surfaces of artificial hip joints include friction, wear, and lubrication, which have been shown to mainly depend on the lubrication regimes involved. There are three lubrication regimes: boundary, fluid-film, and mixed. In the boundary lubrication regime, a significant asperity contact is experienced, and consequently both friction and wear are high. In the fluid film lubrication regime, where the two bearing surfaces are completely separated by a continuous lubricant, minimal friction and wear is expected. The mixed-lubrication regime consists of both fluid film lubricated and boundary contact regions. Friction and lubrication studies are usually performed to understand the wear mechanism involved in artificial hip joints. However, friction forces may be important in determining the stresses experienced at the interface between the implant and the cement bone (28) as well as temperature rise (29).

Friction in artificial hip joints is usually measured in a pendulum-like simulator with a dynamic load in the vertical direction and a reciprocating rotation in the horizontal direction. The coefficient of friction is usually expressed as a friction factor defined as

$$\mu = \frac{T}{w(d_{\text{head}}/2)} \tag{2}$$

where T is the measured friction torque and w is the load.

The measured coefficient of friction in a particular hip prosthesis itself can generally reveal the nature of the lubrication regime, since each mechanism is associated with broad ranges of the coefficient of friction. The variation in the coefficient of friction against a Sommerfeld number defined as, $S = (\eta u d_{\text{head}}/w)$, where is viscosity and u velocity, can further indicate the lubrication regime. If the measured friction factors remain constant, fall or increase as the Sommerfeld number is increased, the associated modes of lubrication are boundary, mixed, or fluid-film, respectively (30).

Lubrication studies of artificial hip joints are generally carried out using both experimental and theoretical approaches. The experimental measurement is usually involved with the detection of the separation between the two bearing surfaces using a simple resistivity technique. A large resistance would imply a thick lubricant film, while a small resistance is attributed to the direct surface contact. Such a technique is directly applicable to metal-on-metal bearings as well as UHMWPE-on-metal and ceramic-on-ceramic bearings if appropriate coatings are used (31,32). The theoretical analysis is generally involved with the solution to the Reynolds equation, together with the elasticity equation subjected to the dynamic load and speed experienced during walking. The predicted film thickness (h_{min}) is then compared with the average surface roughness (Ra) using the following simple criterion.

$$\lambda = \frac{h_{\text{min}}}{[Ra_{\text{head}}^2 + Ra_{\text{cup}}^2]^{1/2}} \tag{3}$$

The lubrication regime is then classified as fluid film, mixed, or boundary if the predicted ratio is > 3, between 1 and 3, or < 1, respectively.

Wear of artificial hip joints has been investigated extensively, due to its direct relevance to biological reactions and clinical problems of osteolysis and loosening. Volumetric wear and wear particles can be measured using the following machines, among others:

- Pin-on-disk machines.
- Pin-on-plate machines.
- Joint simulators.

A unidirectional sliding motion is usually used in the pin-on-disc machine, and the reciprocating motion is added to the pin-on-plate machine. Both of these machines are used to screen potential bearing materials under well controlled, and often simplified conditions. Generally, it is necessary to introduce additional motion in order to produce a multidirectional motion. The next stage of wear testing is usually carried out in joint simulators with a varied degree of complexity of the 3D loading and motion patterns experienced by hip joints, while immersing the test joints in a lubricant deemed to be physically and chemically similar to synovial fluid. Wear can be evaluated by either dimensional or gravimetric means.

Contact mechanics analysis is often performed to predict the contact stresses within the prosthetic components and to compare with the strength of the material. However, other predicted contact parameters such as the contact

area and the contact pressure at the bearing surfaces have been found to be particularly useful in providing insights into friction, wear, and lubrication mechanisms. Contact mechanics can be investigated either experimentally using pressure-sensitive film and sensors, or theoretically using the finite element method.

Biological Response of Wear Debris

Our current understanding of the mechanisms of wear particle-induced osteolysis has developed from >30 years experience with UHMWPE-on-metal. The major factor limiting the longevity of initially well-fixed UHMWPE total joint replacements is osteolysis resulting in late aseptic loosening (33). There is extremely strong evidence from *in vivo* and *in vitro* studies that osteolysis is a UHMWPE particle related phenomenon.

Following total hip arthroplasty, a pseudocapsule forms around the joint and this may have a pseudosynovial lining. A fibrous interfacial tissue may also form at the bone–cement or bone–prosthesis interface that is normally thin with few vessels or cells (34–36). At revision surgery for aseptic loosening, the fibrous membrane is thickened, highly vascularized, and contains a heavy infiltrate of UHMWPE-laden macrophages and multinucleated giant cells (37,38). There is a correlation between the number of macrophages and the volume of UHMWPE wear debris in the tissues adjacent to areas of aggressive osteolysis (39–45). Analyses of interfacial membranes have demonstrated the presence of a multitude of mediators of inflammation including cytokines that may directly influence osteoclastic bone resorption:-TNF-α (46), IL-1β (47), IL-6 (48), and M-CSF (49). There is a direct relationship between the particle concentration and the duration the implant, and there are billions of particles generated per gram of tissue (9,15,50,51). Osteolysis is likely to occur when the threshold of particles exceeds 1×10^{10}/g of tissue (45). Each milligram of PE wear has been estimated to generate 1.3×10^{10} particles (15).

The UHMWPE particles isolated from retrieved tissues vary in size and morphology, from large platelet-like particles, up to 250 μm in length, fibrils, shreds, and submicrometer globule-shaped spheroids 0.1–0.5 μm in diameter (15,52–54). The vast majority of the numbers of particles are the globular spheroids and the mode of the frequency distribution is invariably 0.1–0.5 μm, although the larger particles may account for a high proportion of the total volume of wear debris. Analysis of the mass distribution as a function of size is therefore necessary to discriminate between patient samples (15,55).

UHMWPE wear particles generated *in vitro* in hip joint simulators have a larger proportion of the mass of particles in the 0.01–1 μm sized range than those isolated from periprosthetic tissues (27,55). This may indicate that *in vivo*, the smaller particles are disseminated more widely away from the implant site. Recently, improvements to particle imaging techniques have revealed nanometer sized UHMWPE particles generated in hip joint simulators. These particles have yet to be identified *in vivo*. These nanometer size particles account for the greatest number of particles generated, but a negligible proportion of the total volume (18).

Studies of the response of macrophages to clinically relevant, endotoxin-free polyethylene particles *in vitro* have clearly demonstrated that particle stimulated macrophages elaborate a range of potentially osteolytic mediators (IL-1, IL-6, TNF-α, GM-CSF, PGE_2) and bone resorbing activity (19–21,56–58). Induction of bone resorbing activity in particle stimulated macrophage supernatants has been shown to be critically dependent on particle size and concentration with particles in the 0.1–1.0 μm size range at a volumetric concentration of 10–100 μm^3/cell being the most biologically reactive (19,56). The importance of UHMWPE particle size has also been demonstrated in animal studies (59). These findings have enabled the preclinical prediction of the functional biological activity of different polyethylenes by analysis of the wear rate and mass distribution of the particles as a function of particle size (26,27). For a review of the biology of osteolysis, the reader is referred to Ingham and Fisher (60).

In metal-on-metal bearings in the hip, an abundance of small nanometer size particles are generated (61,62). It is believed that the majority of metal debris is transported away from the periprosthetic tissue. While only isolated instances of local osteolysis have been found around metal-on-metal hips, this is most commonly associated with high concentrations of metal debris and tissue necrosis. *In vitro* cell culture studies have shown that these nanometer size metal particles are highly toxic to cells at relatively low concentrations (22). These particles have a very limited capacity to activate macrophages to produce osteolytic cytokines at the volumes likely to be generated *in vivo* (63), however, metal particles are not bioinert and concerns exist regarding their potential genotoxicity (22).

Ceramic-on-ceramic prostheses have been shown to have extremely low wear rates. Ceramic wear particles generated in hip joint simulations under clinically relevant conditions in the hip joint simulator (64) and *in vivo* (65) have a bimodal size distribution with nanometer sized (5–20 nm) and larger particles (0.2–$>$10 μm). Alumina ceramic particles have been shown to be capable of inducing osteolytic cytokine production by human mononuclear phagocytes *in vitro* (23). However, the volumetric concentration of the particles needed to generate this response was 100–500 μm^3/cell. Given the extremely low wear rates of modern ceramic-on-ceramic bearings, even under severe conditions, it is unlikely that this concentration will arise in the periprosthetic tissues *in vivo* (60).

APPLICATIONS

UHMWPE-on-Metal and UHMWPE-on-Ceramic

The friction in UHMWPE hip joints has been measured using a pendulum-type simulator with a flexionsol–extension motion and a dynamic vertical load. The friction factor has been found to be generally in the range 0.02–0.06 for 28 mm diameter metal heads and UHMWPE cups (66), broadly representative of mixed lubrication, and this has been confirmed from the variation in the friction factor with the Sommerfeld number. These experimental observations are broadly consistent with the theoretical prediction of typical lubricant film thicknesses between 0.1 and

Table 4. Volumetric Wear Rate, % wear volume < 1 μm, SBA, and FBA for Nonirradiated and Irradiated UHMWPEs and Alumina Ceramic-on-Ceramic Hip Joint Prostheses[a]

Material	Volumetric Wear rate, $mm^3/10^6$ cycles $\pm$ 95% CL	% Volume < 1 μm	SBA	FBA
Nonirradiated UHMWPE	50 ± 8	23	0.32	16
Gamma in air UHMWPE, 2.5 Mrad GUR1120	49 ± 9	46	0.55	55
Stabilized UHMWPE (2.5–4 Mrad) GUR1020	35 ± 9	43	0.5	17.5
Highly cross-linked UHMWPE, 10 Mrad GUR1050	8.6 ± 3.1	95	0.96	8
Alumina ceramic-on-ceramic (microseparation)	1.84 ± 0.38	100	0.19	0.35

[a]See Refs. 60,69.

0.2 μm and the average surface roughness of UHMWPE bearing surface between 0.1 and 1 μm. Therefore, wear of UHMWPE acetabular cups is largely governed by the boundary lubrication mechanism. An increase in the femoral head diameter can lead to an increase in sliding distance and consequently wear (41). As a result, 28 mm diameter femoral heads appear to be a better choice. Furthermore, reducing the surface roughness of the metallic femoral head or using harder alumina to resist third-body abrasion and to maintain the smoothness is also very important. For example, the wear factor in UHMWPE-on-ceramic implants is generally 50% of that in UHMPWE-on-metal (67). The introduction of cross-linked UHMWPE has been shown to reduce wear significantly in simulator studies. However, the degree of wear reduction appears to depend on cross-linking, kinematics, counterface roughness, and bovine serum concentration (68). It should be pointed out that volumetric changes are often accompanied by morphology changes, which may have different biological reactions as discussed below.

First, let us consider the effect of irradiation and cross-linking on the osteolytic potential of UHMWPE bearings. Historically, UHMWPE acetabular cups were gamma irradiated in air until it became clear that oxidative degeneration of the PE was occurring. This oxidative damage was caused by the release of free radicals, which produced strand scission of the long PE chains. Research has indicated that deterioration to important mechanical properties such as tensile strength, impact strength, toughness, fatigue strength, and Young's modulus occurs (12). These time-dependent changes have been shown to affect the volumetric wear of the UHMWPE and typical values are in the region of 100 mm^3/million cycles. In addition, UHMWPE that had been gamma irradiated in air produced a greater volumetric concentration of wear particles that were in the most biologically active size range, 0.1–1 μm (46% of the wear volume compared to 24% for nonirradiated UHMWPE). When the specific biological activity (biological activity per unit volume of wear; SBA) of the wear particles was calculated this gave an SBA that was 1.7-fold higher than the SBA of the nonirradiated material, which translated into a functional biological activity (FBA), which was 3.5-fold higher than the FBA of the nonirradiated material (Table 4).

Currently, UHMWPE is sterilized by gamma irradiation (2.5–4 Mrad) in an inert atmosphere. This material undergoes partial cross-linking as a result of this processing, and is often referred to as moderately cross-linked or stabilized PE. This material produces lower wear rates than the nonirradiated UHMWPE, but has a higher volumetric concentration of wear particles < 1 μm compared to the nonirradiated material as shown in Table 4 (69). Consequently, the specific biological activity of the wear particles is higher at 0.5 compared to 0.32 for the nonirradiated material. However, as the wear volume is substantially lower, the FBA value for the stabilized UHMWPE is very similar to the nonirradiated material.

As the level of cross-linking increases, the wear volume decreases (69). The highly cross-linked UHMWPE is GUR 1050, irradiated at 10 Mrad and remelted, and has very low wear volumes at 8.6 ± 3.1 mm^3/million cycles. However, as can be seen from Table 4, 95% of the wear volume is comprised of particles in the most biologically active size range, leading to an extremely high SBA. However, as the wear volume is significantly lower than the other UHMWPEs, the FBA is one-half of those of the nonirradiated and stabilized materials (Table 4).

In addition, the wear particles from the cross-linked materials have increased biological activity per unit volume of wear (Fig. 2). A recent study by Ingram et al. (18) has shown that when worn against a scratched counterface, PE irradiated with 5 and 10 Mrad of gamma irradiation produced higher volumetric concentrations of wear particles in the 0.01–1.0 μm size range compared to noncross-linked material. This increased volumetric concentration of wear particles in the 0.01–1.0 μm size range meant that both cross-linked materials were able to stimulate the release of elevated levels of TNF-α, an osteolytic cytokine, at a 10-fold lower volumetric dose than the

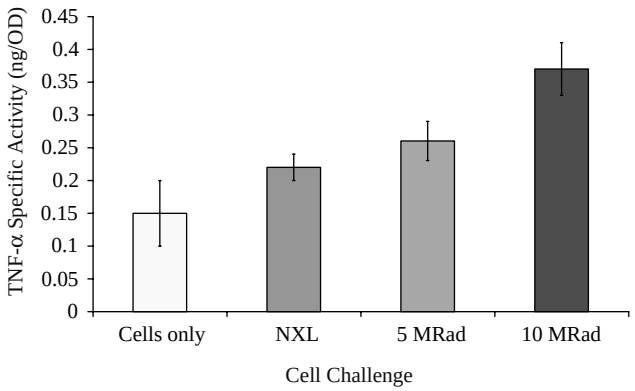


Figure 2. TNF-α release (specific activity $\pm$ 95% confidence limits) as a result of challenge with UHMWPE particles, which were noncross-linked (NXL), cross-linked with 5 Mrad of irradiation, cross-linked with 10 Mrad of irradiation compared to the cell only control.

noncross-linked polyethylene (0.1 μm^3 debris/cell compared to 1–10 μm^3 debris/cell). So, while the cross-linked materials produced lower wear volumes, the particles produced from these materials were more reactive compared to the noncross-linked PE.

However, when the same materials were worn against a smooth counterface, analysis of the wear particles showed that both cross-linked and noncross-linked PE produced very high numbers of nanometer-sized wear particles. In addition, the cross-linked and noncross-linked materials produced similar low volumes of particles in the 0.1–1.0 μm size range, which resulted in wear debris that was only stimulatory a the highest volumetric dose of 50 μm^3 debris/cell. This offers further explanation as to why the FBA or osteolytic potential of the highly cross-linked polyethylene's are lower than the moderately cross-linked and noncross-linked materials (Table 4).

Metal-on-Metal

The friction factor measured in metal-on-metal hip joints with different sizes and clearances in simple pendulum type machines is generally much higher than for UHMWPE-on-metal articulations, in the range between 0.1 and 0.2, indicating a mixed-boundary lubrication regime (66). However, the lubrication regime in metal-on-metal bearings has been shown to be sensitive to the surface roughness, loading and velocity, and design parameters (70–74). Consequently, different friction factors or wear factors are possible. Therefore, it is important to optimize the bearing system, in terms of the femoral head diameter, the clearance and the structural support (75,76). From a lubrication point of view, the femoral head diameter is the most important geometric parameter, since it is directly related to both the equivalent diameter defined in Eq. 1 and the sliding velocity (70). If the lubrication improvement is such that a fluid-film dominant lubrication regime is present, the increase in the sliding distance becomes unimportant. Such an advantage has been utilized in large-diameter metal-on-metal hip resurfacing prostheses (77). However, it should be pointed out that the lubrication improvement in large-diameter metal-on-metal hip resurfacing prostheses can only be realized with adequate clearances (78). A too large clearance can reduce the equivalent diameter, shifting the lubrication regime toward mixed–boundary regions. In addition, the increase in the sliding distance associated with the large diameter means that the bedding-in wear becomes important. The wear in optimised systems can be quite low, of the order of a few millimeters cubed.

The biological activity in terms of osteolytic potential of metal-on-metal hip prostheses is difficult to define. If macrophages and fibroblasts are challenged with clinically relevant cobalt chrome wear particles, there is some release of the osteolytic cytokine TNF-α (Fig. 3), however, this only takes place at very high levels of particulate load (50 μm^3 debris/cell), and the level of cytokine produced is at lower levels compared to UHMWPE particles [see Fig. 2(79)]. The predominant biological reaction is cytotoxicity or a reduction in cell viability (Fig. 4). Macrophage and fibroblast cell viability is significantly reduced when

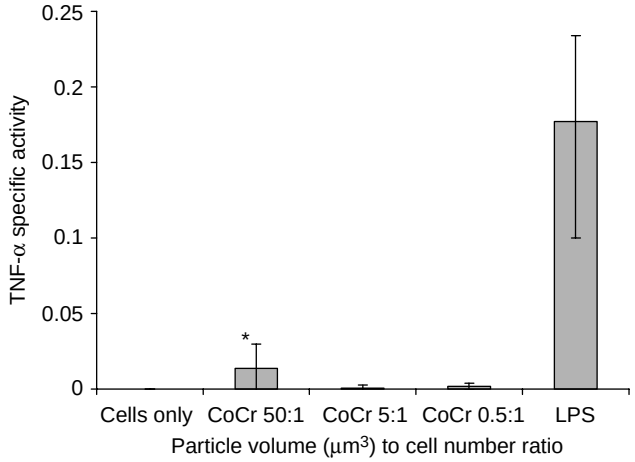


Figure 3. TNF-α production by human peripheral blood mononuclear cells stimulated with clinically relevant cobalt–chromium particles.

challenged with 50 or 5 μm^3 debris/cell (22). The specific biological activity of metal wear particles is difficult to assess as the cells may release cytokines, such as TNF-α as a consequence of cell death. In addition, the high levels of particulate load required to stimulate cytokine release

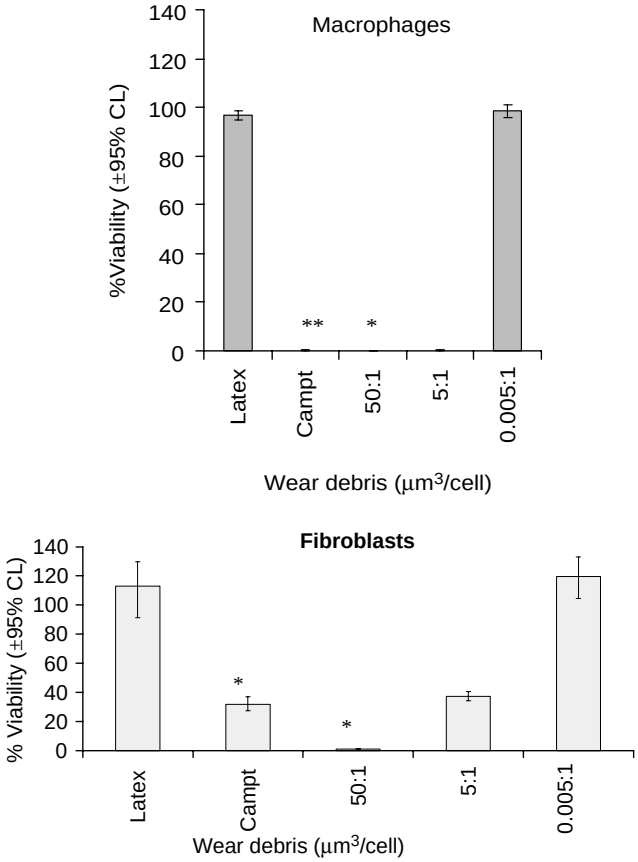


Figure 4. Effect of metal-on-metal cobalt–chromium wear particles on macrophage and fibroblast cell viability (*Significant ($p < 0.05$, ANOVA) reduction in cell viability).

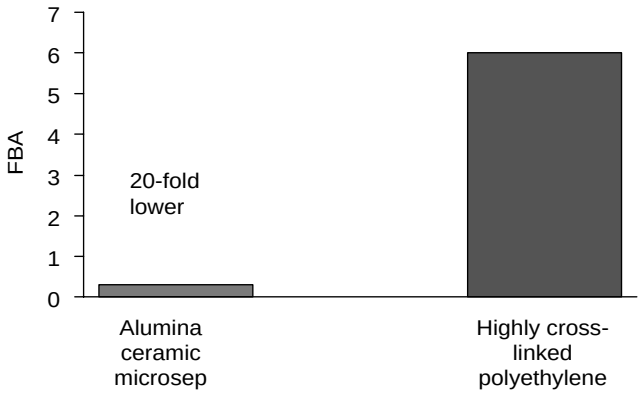


Figure 5. Predicted functional biological activity or osteolytic potential of alumina ceramic-on-ceramic and highly cross-linked UHMWPE-on-metal hip prostheses.

may only be achieved *in vivo* if a pairing is particularly high wearing.

Ceramic-on-Ceramic

The friction factor in ceramic-on-ceramic hip implants is quite low, particularly when the nonbiological type lubricant is used (66). However, when biological lubricants such as bovine serum are tested, the friction factor can be quite high due to the complex interactions with proteins. The wear under normal ideal conditions is low, but can be increased substantially under adverse conditions such as microseparation (80). Despite this, the wear in ceramic-on-ceramic hip implants is generally the lowest among current hip prostheses available clinically.

The introduction of microseparation conditions into the *in vitro* simulation model replicates clinically relevant wear rates, wear patterns, and wear particles. Alumina wear particles have a lower biological activity than UHMWPE particles. A 10-fold higher concentration of alumina wear particles is required to stimulate the release of osteolytic cytokine TNF-α from macrophages compared to UHMWPE wear particles (23). It is questionable whether the volume of alumina wear particles will reach this threshold *in vivo* given the extremely low wear rates of ceramic-on-ceramic prostheses even under severe microseparation conditions. Consequently, alumina wear particles have a lower specific biological activity than UHMWPE particles (Table 4). When this lower SBA is integrated with the comparatively small volumetric wear rates that are produced by ceramic-on-ceramic couples compared to metal-on-polyethylene, a substantially lower functional biological activity or osteolytic potential is produced (Table 4; Fig. 5). In fact, alumina ceramic-on-ceramic couples produce a 20-fold lower FBA than the currently used highly cross-linked UHMWPEs.

Summary

Typical values of friction factor, wear factor, and biological reactions are summarized in Tables 5, 6, and 7 for various hip implants with different bearing couples.

FUTURE DEVELOPMENTS

Cross-Linked Polyethylene

The introduction of highly cross-linked PE into clinical use in the last 5 years has been extensive. Standard UHMWPE is irradiated at high dose levels (5–10 Mrad), which produces chain scission and cross-linking between the molecular chains. Subsequent heating and remelting recombines the free radicals producing a more stable material (81). The additional cross-links provide improved wear resistance, particularly during kinematic conditions with high levels of cross-shear as found in the hip. A number of early simulator studies showed no wear for these materials (81), while other studies demonstrated measurable wear (82). Initial clinical studies, however, do show penetration and wear (83). Wear and surface cracking has been identified in a few isolated retrievals. The wear rates reported more recently in simulator studies have been found to be in the range 5–10 mm^3/million cycles, which is four to five times less than with conventional material (69). Recent work has also shown that cross-linked PE produces a greater proportion of smaller particles, per unit volume of wear debris and has been found to be up to three times more biologically active than conventional material (18). This leads to a functional reduction in osteolytic potential of about twofold compared to conventional PE. This improvement in functional osteolytic potential may not be sufficient for high demand patients, and in patients who require large head sizes. In these patients, larger diameter hard-on-hard, such as ceramic-on-ceramic or metal-on-metal, may be a more appropriate bearing choice.

Ceramic-on-Metal Bearing

Currently used hard-on-hard bearings are comprised of similar materials, such as alumina ceramic-on-ceramic or metal-on-metal. When lubrication conditions are depleted, like bearing materials can produce elevated adhesive friction and wear. The ceramic-on-metal hip was developed to produce a deferential hardness hard bearing (84). Laboratory simulation studies have shown a reduction in wear of

Table 5. Typical Friction Factors and Lubrication Regimes in Various Bearings for Hip Implants[a]

Bearing Couples	Friction Factor	Variation of Friction Factor Against Increasing Sommerfeld Number	Indicated Lubrication Regimes
UHMWPE-on-Metal	0.06–0.08	Constant/decreasing	Boundary/mixed
Metal-on-metal	0.22–0.27	Decreasing	Mixed
Ceramic-on-ceramic	0.002–0.2	Increasing	Fluid-film/mixed

[a]See Ref. 4.

Table 6. Typical Volumetric and Linear Wear Rates for Different Bearings for Hip Implants[a]

Bearing Couples	Volumetric Wear Rate, mm^3/million cycles	Linear Wear Rate, μm/million cycles
UHMWPE-on-metal	30–100	100–300
UHMWPE-on-ceramic	15–50	50–150
Metal-on-metal	0.1–1	2–20
Ceramic-on-ceramic	0.05–1	1–20

[a]See Ref. 4.

Table 7. Typical Particle Sizes and Biological Responses in Different Bearings for Hip Implants[a]

Bearing Couples	Dominant Particle Diameters, μm	Biological Responses
UHMWPE-on-metal/ceramic	UHMWPE, 0.01–1	Macrophages/osteoclasts/osteolysis
Metal-on-metal	Metallic, 0.02–0.1	Low osteolysis, cytotoxicity
Ceramic-on-ceramic	Ceramic, 0.01–0.02	Bioinert, low cytotoxicity
	Ceramic, 0.1–10	Macrophages/osteoclasts/osteolysis

[a]See Ref. 69.

up to 100-fold compared to metal on metal. The ceramic head does not wear and remains smooth, improving lubrication and reducing wear of the metallic cup. This new concept is currently entering clinical trials.

Surface Replacement Bearings

There is considerable interest in surface replacement solutions in the hip (85). In this approach, a large diameter metallic shell is placed over the reamed femoral head, preserving femoral bone stock, and this articulates against a large diameter acetabular cup. Both cobalt chrome cast and wrought alloys have been used in different bearing designs. The larger diameter head improves lubrication and reduces wear compared to smaller head sizes (70). However, it is important to maintain a low radical clearance between the components to ensure low bedding-in wear (78,86,87). Surface replacement metal on metal bearings are not suitable for all patients, due to the nature of the femoral bone, but are currently used in $\sim 10\%$ of patients receiving hip prostheses.

Surface Engineered Metal-on-Metal Bearings SUREHIP

Concerns still remain about wear particles in metal on metal bearings and elevated metal ion levels. Surface engineering solutions are an attractive option for reducing wear and metal ion levels, and can be readily applied to surface replacement hips. Recent research with thick AEPVD chromium nitride and chromium carbon nitride surface engineered coatings of thicknesses between 8 and 12 μm have shown a 30-fold reduction in wear and metal ion levels (88,89). These coatings are now undergoing product development in preparation for clinical studies.

Compliant Materials, Cushion Form Bearings

In recent years, the trend has been to move toward harder bearing materials that wear less, and away from the lower elastic modulus properties of articular cartilage. Compliant materials such as polyurethane have been investigated as bearing materials in acetabular cups. The cups have been formed as a composite structure with a higher modulus substrate to give structural support (90). The bearing has shown improved lubrication and reduced wear compared to conventional polyethylene bearings. However, concerns remain about the long-term stability of these low modulus materials. More recently an experimental polyurethane surface replacement cup has been investigated (91).

Hemiarthroplasty and Hemisurface Replacements

Interest in more conservative bone preserving, and minimally invasive surgery has generated renewed interest in surgical treatments that replace only one side of the diseased joint or potentially just the diseased surface itself. In these scenarios, consideration has not only to be given to the biomaterial replacing the degenerated tissue, but also the function of the apposing articulating surface.

In the hip hemiarthroplasty using compliant materials has just entered clinical trials, where the femoral head covered with a layer of polyurethane articulates against the natural cartilage in the acetabulum (http://www.impliant.com/home/index.html). Future designs will focus on full or partial replacement of the diseased cartilage on one side of the joint.

SUMMARY

This article summarizes the biotribology of artificial hip joints and development over the last four decades. Although adequate solutions exist for the elderly less active patients >65 years old with average life expectances <20 years, considerable technological advances are required to meet the demands and improved performance of younger patients. Recent moves toward large diameter heads to give greater function, stability and range of motion are placing greater demands on tribological performances and increasing use of the hard-on-hard bearings.

Nomenclature

d Diametral clearance

D Dearing diameter or equivalent diameter defined in Eq. 1

FBA Functional biological activity

h_{min} Minimum lubricant film thickness

PMMA Poly(methyl methacrylate)

Ra Average surface roughness

S Summerfeld number

SBA Specific biological activity

T Frictional torque

u Siding velocity

UHMWPE Ultrahigh molecular weight polyethylene

w Load

η Viscosity

λ Ratio defined in Eq.3

μ Frictional factor defined in Eq. 2

Subscripts:

Head Femoral head

Cup Acetabular cup

BIBLIOGRAPHY

Cited References

1. Malchau H, Herberts P, Ahnfelt L. Prognosis of total hip replacement in Sweden. Acta Orthop Scand 1993;64–65.
2. Yao JQ, Laurent MP, Johnson TS, Blanchard CR, Crowinshield RD. The influence of lubricant and material on polymer/CoCr sliding friction. Wear 2003;255:780–784.
3. Paul JP. Strength requirements for internal and external prostheses. J Biomech 1999;32(4):381–393.
4. Jin ZM, Medley JB, Dowson D. Fluid Film Lubrication In Artificial Hip Joints. Proc 29th Leeds-Lyon Symp Tribology; 2003. p 237–256.
5. Lewis G. Properties of crosslinked ultra-high-molecular-weight polyethylene. Biomaterials 2001;22(4):371–401.
6. Chiesa R, Tanzi MC, Alfonsi S, Paracchini L, Moscatelli M, Cigada A. Enhanced wear performance of highly cross-linked UHMWPE for artificial joints. J Biomed Mat Res 2000;50:381–387.
7. Bowsher JG, Shelton JC. A hip simulator study of the influence of patient activity level on the wear of cross-linked polyethylene under smooth and roughened counterface conditions. Wear 2001;250:167–179.
8. Hermida JC, Bergula A, Chen P, Colwell CW, D'Lima DD. Comparison of wear rates of twenty-eight and thirty-two millimeter femoral heads on cross-linked polyethylene acetabular cups in a wear simulator. J Bone Joint Surg 2003;85A:2325–2331.
9. McKellop HA, Campbell P, Park SH, Schmalzried TP, Grigoris P, Amstutz HC, Sarmiento A. The origin of submicron polyethylene wear debris in total hip arthroplasty. Clin Orthopaed Rel Res 1995;311:3–20.
10. Tipper JL, Ingham E, Fisher J. Characterisation of wear debris from UHMWPE, metal on metal and ceramic on ceramic hip prostheses. Wear 2001;250:120–128.
11. Saikko V, Calonius O, Keranen J. Wear of conventional and cross-linked ultra-high molecular -weight polyethylene acetabular cups against polished and roughened CoCr femoral heads in a biaxial hip simulator. J Biomed Res Appl Biomat 2002;63:848–853.
12. Besong AA, Tipper JL, Ingham E, Stone MH, Wroblewski BM, Fisher J. Quantitative comparison of wear debris from UHMWPE that has and has not been sterilized by gamma irradiation. J Bone Joint Sur 1998;80B:340–344.
13. Endo MM, Barbour PSM, Barton DC, Wroblewski BM, Fisher J, Tipper JL, Ingham E, Stone MH. A comparison of the wear and debris generation of GUR 1120 (compression moulded) and GUR 4150HP (ram extruded) ultra high molecular weight polyethylene. Bio Med Mat Eng 1999;9:113–124.
14. Endo MM, Barbour PS, Barton DC, Fisher J, Tipper JL, Ingham E, Stone MH. Comparative wear and wear debris under three different counterface conditions of cross-linked and non-cross-linked ultra high molecular weight polyethylene. Bio Med Mat Eng 2001;11:23–35.
15. Tipper JL, Ingham E, Hailey JL, Besong AA, Fisher J, Wroblewski BM, Stone MH. Quantitative analysis of polyethylene wear debris, wear rate and head damage in retrieved Charnley hip prostheses. J Mat Sci, Mat Med 2000;11:117–124.
16. Bell J, Besong AA, Tipper JL, Ingham E, Wroblewski BM, Stone MH, Fisher J. Influence of gelatin and bovine serum lubrication on ultra-high molecular weight polyethylene wear debris generated in *in vitro* simulations. Proc Inst Mech Eng J Eng Med 2000;214H:513–518.
17. Ingram J, Matthews JB, Tipper JL, Stone MH, Fisher J, Ingham E. Comparison of the biological activity of grade GUR 1120 and GUR 415 HP UHMWPE wear debris. Bio Med Mat Eng 2002;12:177–188.
18. Ingram JH, Stone MH, Fisher J, Ingham E. The influence of molecular weight, crosslinking and counterface roughness on TNF-alpha production by macrophages in response to ultra high molecular weight polyethylene particles. Biomaterials 2004;25:3511–3522.
19. Green TR, Fisher J, Matthews JB, Stone MH, Ingham E. Effect of size and dose on bone resorption activity of macrophages *in vitro* by clinically relevant ultra high molecular weight polyethylene particles. J Biomed Mat Res Appl Biomat 2000;53:490–497.
20. Matthews JB, Green TR, Stone MH, Wroblewski BM, Fisher J, Ingham E. Comparison of the response of primary human peripheral blood mononuclear phagocytes from different donors to challenge with polyethylene particles of known size and dose. Biomaterials 2000;21:2033–2044.
21. Matthews JB, Stone MH, Wroblewski BM, Fisher J, Ingham E. Evaluation of the response of primary human peripheral blood mononuclear phagocytes challenged with *in vitro* generated clinically relevant UHMWPE particles of known size and dose. J Biomed Mat Res Appl Biomat 2000;44:296–307.
22. Germain MA, Hatton A, Williams S, Matthews JB, Stone MH, Fisher J, Ingham E. Comparison of the cytotoxicity of clinically relevant cobalt-chromium and alumina ceramic wear particles *in vitro*. Biomaterials 2003;24:469–479.
23. Hatton A, Nevelos JE, Matthews JB, Fisher J, Ingham E. Effects of clinically relevant alumina ceramic wear particles on TNF-α production by human peripheral blood mononuclear phagocytes. Biomaterials 2003;24:1193–1204.
24. Ingham E, Green TR, Stone MH, Kowalski R, Watkins N, Fisher J. Production of TNF-α and bone resorbing activity by macrophages in response to different types of bone cement particles. Biomaterials 2000;21:1005–1013.
25. Mitchell W, Matthews JB, Stone MH, Fisher J, Ingham E. Comparison of the response of human peripheral blood mononuclear cells to challenge with particles of three bone cements *in vitro*. Biomaterials 2003;24:737–748.

26. Fisher J, Bell J, Barbour PSM, Tipper JL, Matthews JB, Besong AA, Stone MH, Ingham E. A novel method for the prediction of functional biological activity of polyethylene wear debris. J Eng Med Proc Inst Mech Eng 2001; 215H: 127–132.
27. Endo MM, Tipper JL, Barton DC, Stone MH, Ingham E, Fisher J. Comparison of wear, wear debris and functional biological activity of moderately crosslinked and non-cross-linked polyethylenes in hip prostheses. Proc Instn Mech Eng J Eng Med 2002;216:111–122.
28. Nassutt R, Wimmer MA, Schneider E, Morlock MM. The influence of resting periods on friction in the artificial hip. Clin Orthop 2003;407:127–38.
29. Bergmann G, Graichen F, Rohlmann A, Verdonschot N, van Lenthe GH. Frictional heating of total hip implants. Part 2: finite element study. J Biomech 2001;34(4):429–435.
30. Dowson D. New joints for the Millennium: wear control in total replacement hip joints. Proc Instn Mech Eng J Eng Med 2001;215(H4):335–358.
31. Smith SL, Dowson D, Goldsmith AJ, Valizadeh R, Colligon JS. Direct evidence of lubrication in ceramic-on-ceramic total hip replacements. Proc Inst Mech Eng J Mech Eng Sci 2001;215(3):265–268.
32. Murakami T, Sawae Y, Nakashima K, Sakai N, Doi S, Sawano T, Ono M, Yamamoto K, Takahara A. Roles of Materials and Lubricants in Joint Prostheses. Proc 4th Int Biotribol Forum 24th Biotribol Symp 2003;1–4.
33. Archibeck MJ, Jacobs JJ, Roebuck KA, Glant TT. The basic science of periprosthetic osteolysis. J Bone Joint Surg 2000;82A:1478–1497.
34. Goldring SR, Schiller AL, Roelke M, Rourke CM, O'Neil DA, Harris WH. The synovial-like membrane at the bone-cement interface in loose total hip replacements and its proposed role in bone lysis. J Bone Joint Surg 1983;65A:575–584.
35. Goodman SB, Chin RC, Chou SS, Schurman DJ, Woolson ST, Masada MP. A clinical-pathologic-biochemical study of the membrane surrounding loosened and non-loosened total hip arthroplasties. Clin Orthopaed 1989;244:182–187.
36. Bullough PG, DiCarlo EF, Hansraj KK, Neves MC. Pathologic studies of total joint replacement. Orthopaed Clin North Am 1988;19:611–625.
37. Mirra JM, Marder RA, Amstutz HC. The pathology of failed total joint arthroplasty. Clin Orthopaed Rel Res 1982; 170:175–183.
38. Willert HG, Buchhorn GH. Particle disease due to wear of ultrahigh molecular weight polyethylene. Findings from retrieval studies. In: Morrey BF, editor. Biological, Material, and Mechanical Considerations of Joint Replacement. New York: Raven Press; 1993.
39. Maloney WJ, Jasty M, Harris WH, Galante MD, Callaghan JJ. Endosteal erosion in association with stable uncemented femoral components. J Bone Joint Surg 1990;72A:1025–1034.
40. Santavirta S, Holkka V, Eskola A, Kontinen YT, Paavilainen T, Tallroth K. Aggressive granulomatous lesions in cementless total hip arthroplasty. J Bone Joint Surg 1990;72B:980–985.
41. Livermore J, Ilstrup D, Morrey B. Effect of femoral head size on the wear of the polyethylene acetabular component. J Bone Joint Surg 1990;72A:518–528.
42. Schmalzried TP, Jasty M, Harris WH. Periprosthetic bone loss in total hip arthroplasty: polyethylene wear debris and the concept of the effective joint space. J Bone Joint Surg 1992;74A:849–863.
43. Howie AW, Haynes DR, Rogers SD, McGee MA, Pearcey MJ. The response to particulate debris. Orthopaed Clinics North Am 1993;24:571–581.
44. Bobyn JD, Jacobs JJ, Tanzer M, Urban RM, Arbindi R, Summer DR, Turner TM, Brooks CE. The susceptibility of smooth implant surfaces to peri-implant fibrosis and migration of polyethylene wear debris. Clin Orthopaed Rel Res 1995;311:21–39.
45. Revell PA. Biological reaction to debris in relation to joint prostheses. Proc Inst Mech Eng J Eng Med 1997;211:187–197.
46. Xu JW, Kotinnen T, Lassus J, Natah S, Ceponis A, Solovieva S, Aspenberg P, Santavirta S. Tumor necrosis factor-alpha (TNF-α) in loosening of total hip replacement (THR). Clin Exp Rheumatol 1996;14:643–648.
47. Kim KJ, Rubash H, Wilson SC, D'Antonio JA, McClain EJ. A histologic and biochemical comparison of the interface tissues in cementless and cemented hip prostheses. Clin Orthopaed Rel Res 1993;287:142–152.
48. Sabokbar A, Rushton Role of inflammatory mediators and adhesion molecules in the pathogenesis of aseptic loosening in total hip arthroplasties. J Arthropl 1995;10:810–815.
49. Takei I, Takagi M, Ida H, Ogino S, Santavirta, Konttinen YT. High macrophage-colony stimulating factor levels in synovial fluid of loose artificial hip joints. J Rheumatol 2000;27:894–899.
50. Campbell P, Ma S, Yeom B, McKellop H, Schmalzried TP, Amstutz HC. Isolation of predominantly sub-micron sized UHMWPE wear particles from periprosthetic tissues. J Biomed Mat Res 1995;29:127–131.
51. Hirakawa K, Bauer TW, Stulberg BN, Wilde AH. Comparison and quantitation of wear debris of failed total hip and knee arthroplasty. J Biomed Mat Res 1998;31:257–263.
52. Maloney WJ, Smith RL, Hvene D, Schmalzried TP, Rubash H. Isolation and characterization of wear debris generated in patients who have had failure of a hip arthroplasty without cement. J Bone Joint Surg 1994;77A:1301–1310.
53. Margevicius KT, Bauer TW, McMahon JT, Brown SA, Merritt K. Isolation and characterization of debris from around total joint prostheses. J Bone Joint Surg 1994;76A: 1664–1675.
54. Shanbhag AS, Jacobs JJ, Glant T, Gilbert JL, Black J, Galante JO. Composition and morphology of wear debris in failed uncemented total hip replacement. J Bone Joint Surg 1994;76B:60–67.
55. Howling GI, Barnett PI, Tipper JL, Stone MH, Fisher J, Ingham E. Quantitative characterization of polyethylene debris isolated from periprosthetic tissue in early failure knee implants and early and late failure Charnley hip implants. J Biomed Mat Res Appl Biomater 2001;58:415–420.
56. Green TR, Fisher J, Stone MH, Wroblewski BM, Ingham E. Polyethylene particles of a critical size are necessary for the induction of cytokines by macrophages *in vitro*. Biomaterials 1998;19:2297–2302.
57. Matthews JB, Green TR, Stone MH, Wroblewski BM, Fisher J, Ingham E. Comparison of the response of primary murine peritoneal macrophages and the U937 human histiocytic cell line to challenge with *in vitro* generated clinically relevant UHMWPE particles. Biomed Mat Eng 2000;10:229–240.
58. Matthews JB, Green TR, Stone MH, Wroblewski BM, Fisher J, Ingham E. Comparison of the response of three human monocytic cell lines to challenge with polyethylene particles of known size and dose. J Mat Sci: Mat Med 2001;12:249–258.
59. Goodman SB, Fornasier VL, Lee J, Kei J. The histological effects of the implantation of different sizes of polyethylene particles in the rabbit tibia. J Biomed Mat Res 1990;24:517–524.
60. Ingham E, Fisher J. The role of macrophages in the osteolysis of total joint replacement. Biomaterials (In Press, 2005).
61. Doorn PF, Campbell PA, Worrall J, Benya PD, McKellop HA, Amstutz HC. Metal wear particle characterization from metal on metal total hip replacements: transmission electron microscopy study of periprosthetic tissues and isolated particles. J Biomed Mat Res 1998;42:103–111.
62. Firkins PJ, Tipper JL, Saadatzadeh MR, Ingham E, Stone MH, Farrar R, Fisher J. Quantitative analysis of wear and wear debris from metal-on-metal hip prostheses tested in a

physiological hip joint simulator. Biomed Mat Eng 2001; 11:143–157.
63. Germain MA. Biological reactions to cobalt chrome wear particles, Ph.D. dissertation, University of Leeds, 2002.
64. Tipper JL, Hatton A, Nevelos JE, Ingham E, Doyle C, Streicher R, Nevelos AA, Fisher J. Alumina-alumina artificial hip joints- Part II: Characterisation of the wear debris from *in vitro* hip joint simulations. Biomaterials 2002;23: 3441–3448.
65. Hatton A, Nevelos JE, Nevelos AA, Banks RE, Fisher J, Ingham E. Alumina-alumina artificial hip joints- Part I: a histological analysis and characterization of wear debris by laser capture microdissection of tissues retrieved at revision. Biomaterials 2002;23:3429–3440.
66. Scholes SC, Unsworth A. Comparison of friction and lubrication of different hip prostheses. Proc Inst Mech Eng J Eng Med 2000;214(1):49–57.
67. Ingham E, Fisher J. Biological reactions to wear debris in total joint replacement. Proc Inst Mech Eng J Eng Med 2000;214(H1):21–37.
68. Galvin AL, Tipper J, Stone M, Ingham E, Fisher J. Reduction in wear of crosslinked polyethylene under different tribological conditions. Proc Int Conf Eng Surg Joined Hip, IMechE 2002; C601/005.
69. Galvin AL, Endo MM, Tipper JL, Ingham E, Fisher J. Functional biological activity and osteolytic potential of non-cross-linked and cross-linked UHMWPE hip joint prostheses. Trans 7th World Biomat Cong 2004. p 145.
70. Jin ZM, Dowson D, Fisher J. Analysis of fluid film lubrication in artificial hip joint replacements with surfaces of high elastic modulus. Proc Inst Mech Eng J Eng Med 1997;211: 247–256.
71. Chan FW, Bobyn JD, Medley JB, Krygier JJ, Tanzer M. The Otto Aufranc Award-Wear and lubrication of metal-on-metal hip implants. Clin Orthopaed Rel Res 1999;369:10–24.
72. Firkins PJ, Tipper JL, Ingham E, Stone MH, Farrar R, Fisher J. Influence of simulator kinematics on the wear of metal-on-metal hip prostheses. Proc Inst Mech Eng J Eng Med 2001a;215(H1):119–121.
73. Scholes SC, Green SM, Unsworth A. The wear of metal-on-metal total hip prostheses measured in a hip simulator. Proc Inst Mech Eng J Eng Med 2001;215(H6):523–530.
74. Williams S, Stewart TD, Ingham E, Stone MH, Fisher J. Metal-on-metal bearing wear with different swing phase loads. J Biomed Mater Res 2004;15:70B(2):233–9.
75. Liu F, Jin ZM, Grigoris P, Hirt F, Rieker C. Contact Mechanics of Metal-on-Metal Hip Implants Employing a Metallic Cup With an UHMWPE Backing, Journal of Engineering in Medicine. Proc Inst Mech Eng 2003;217:207–213.
76. Liu F, Jin ZM, Grigoris P, Hirt F, Rieker C. Elastohydrodynamic Lubrication Analysis of a Metal-on-Metal Hip Implant Employing a Metallic Cup With an UHMWPE Backing Under Steady-State Conditions. J Eng Med Proc Inst Mech Eng 2004;218:261–270.
77. Smith SL, Dowson D, Goldsmith AAJ. The lubrication of metal-on-metal total hip joints: a slide down the Stribeck curve. Proc Inst Mech Eng J Eng Tribol 2001;215(J5):483–493.
78. Rieker CB, et al.In vitro tribology of large metal-on-metal implants. Proc 50th Trans Orthopaed Res Soc 2004; 0123.
79. Ingham E, Fisher J. Can metal particles (Theoretically) cause osteolysis? Proceedings of the Second International Conference on Metal-Metal Hip Prostheses: Past Performance and Future Directions, Montreal, Canada, 2003.
80. Nevelos JE, Ingham E, Doyle C, Streicher R, Nevelos AB, Walter W, Fisher J. Micro-separation of the centres of alumina–alumina artificial hip joints during simulator testing produces clinically relevant wear rates and patterns. J Arthroplasty 2000;15(6):793–795.
81. Muratoglu OK, Bragdon CR, O'Connor D, Jasty M, Harris WH, Gul R, McGarry F. Unified wear model for highly cross-linked ultra-high molecular weight polyethylene (UHMWPE). Biomaterials 1999;20:1463–1470.
82. McKellop H, Shen FW, Lu B, Campbell P, Salovey R. Development of an extremely wear-resistant ultra high molecular weight polyethylene for total hip replacements. J Orthop Res 1999;17:157–167.
83. Bradford L, Baker DA, Graham J, Chawan A, Ries MD, Pruitt LA. Wear and surface cracking in early retrieved highly cross-linked polyethylene acetabular liners. J Bone Joint Surg 2004;86A:1271–1282.
84. Firkins PJ, Tipper JL, Ingham E, Stone MH, Farrar R, Fisher J. A novel low wearing differential hardness, ceramic-on-metal hip joint prosthesis. J Biomech 2001;34(10):1291–1298.
85. McMinn D, Treacy R, Lin K, Pynsent P. Metal on metal surface replacement of the hip. Experience of the McMinn prosthesis. Clin Orthop 1996;329:S89–98.
86. Hu XQ, Isaac GH, Fisher J. Changes in the contact area during the bedding-in wear of different sizes of metal on metal hip prostheses. J Biomed Mater Eng 2004;14(2):145–149.
87. Dowson D, Hardaker C, Flett M, Isaac GH. A hip joint simulator study of the performance of metal-on-metal joints: Part II: Design. J Arthroplasty 2004;19(8 Suppl 1):124–130.
88. Fisher J, Hu XQ, Tipper JL, Stewart TD, Williams S, Stone MH, Davies C, Hatto P, Bolton J, Riley M, Hardaker C, Isaac GH, Berry G, Ingham E. An in vitro study of the reduction in wear of metal-on-metal hip prostheses using surface-engineered femoral heads. Proc Inst Mech Eng [H] 2002; 216(4):219–230.
89. Fisher J, Hu XQ, Stewart TD, Williams S, Tipper JL, Ingham E, Stone MH, Davies C, Hatto P, Bolton J, Riley M, Hardaker C, Isaac GH, Berry G. Wear of surface engineered metal-on-metal hip prostheses. J Mater Sci Mater Med 2004;15(3):225–235.
90. Bigsby RJ, Auger DD, Jin ZM, Dowson D, Hardaker CS, Fisher J. A comparative tribological study of the wear of composite cushion cups in a physiological hip joint simulator. J Biomech 1998;31(4):363–369.
91. Jennings LM, Fisher J. A biomechanical and tribological investigation of a novel compliant all polyurethane acetabular resurfacing system. Proceedings of the International Conference of Engineers and Surgeons Joined at the Hip, IMechE 2002; C601/032.

See also ALLOYS, SHAPE MEMORY; BIOMATERIALS, CORROSION AND WEAR OF; JOINTS, BIOMECHANICS OF; ORTHOPEDICS, PROSTHESIS FIXATION FOR.

HIP REPLACEMENT, TOTAL. See MATERIALS AND DESIGN FOR ORTHOPEDIC DEVICES.

HOLTER MONITORING. See AMBULATORY MONITORING.

HOME HEALTH CARE DEVICES

TOSHIYO TAMURA
Chiba University School of Engineering
Chiba, Japan

INTRODUCTION

The increase in the size of the elderly population and the importance of preventing life-related diseases, such as

cancer, hypertension, and diabetes, all emphasize the importance of home healthcare. Generally, the purpose of home healthcare is to reduce the distance the patient must travel to receive care and to reduce the number of hospital admissions.

The devices used in home healthcare must be simple to use, safe, inexpensive, and noninvasive or minimum invasive so that they excessively disturb normal daily activities. Recent developments in home healthcare devices meet most of these specifications, but some still pose a problem in that they disturb the activities of normal daily life and the effectiveness of some other devices in monitoring particular health-related parameters has been questioned.

The requirements for the devices used in home healthcare depend on both their purpose and the subject's condition. Monitoring vital signs, such as heart rate, blood pressure, and respiration, is routinely done for elderly individuals and for patients with chronic disease or who are under terminal care. These cases require simple, noninvasive monitors. When patients are discharged from the hospital and continue to need these parameters monitored at home, the devices used are essentially no different from those used in the hospital.

For health management and the prevention of disease, an automatic health monitoring system has been considered. The onset of lifestyle-related diseases, such as hypertension, arteriosclerosis, and diabetes, is highly correlated with daily activities, such as physical exercise, including walking, as well as other habits, such as sleep and smoking. To prevent such diseases, daily monitoring will be important for achieving healthy living and improving the quality of life. Although the monitoring of daily activities is not well established in evidenced-based health research, there have been many attempts at installing sensors and transducers and monitoring daily life at home.

Evidenced-based health research directed at finding correlations between daily activity monitoring and the onset of disease, and identifying risk factors, is a major subject of epidemiology. In general, large population studies of daily living activities, including daily food intake, based on the history using interviews or questionnaires are required. If an automatic monitoring system can be applied, more reliable and objective data can be obtained.

Recently, many new home healthcare devices have been developed because many individuals have become motivated to maintain their health. This article discusses recently developed homecare devices, as well as expected laboratory-based devices.

BLOOD PRESSURE

Blood pressure is one of the most important physiological parameters to monitor. Blood pressure varies considerably throughout the day and frequent blood pressure monitoring is required in many home healthcare situations. Usually, a blood pressure reading just before the patient wakes in the morning is required. The success of home blood pressure readings is highly dependent on the patient's motivation. Blood pressure readings at home are also recommended because many patients have elevated blood pressure readings in a clinical setting, the so-called "white-coat hypertension".

Figure 1. The standard blood pressure monitor. The device includes an inflatable cuff, a manometer, and a stethoscope. The bladder is inflated until the cuff compresses the artery in the arm; since no blood passes, the stethoscope detects no noise. Then, the cuff is deflated slowly, blood passes though the artery again, and the stethoscope perceives a noise, which is defined as the systolic pressure. The cuff continues to deflate and finally the stethoscope perceives no noise, defined as the diastolic pressure.

Medical doctors and nurses usually measure blood pressure using the auscultatory method as shown in Fig. 1, in which a pressure cuff is attached to the upper arm and inflated to compress the brachial artery to a value above the systolic pressure. Then, the cuff is gradually deflated while listening to the Korotkoff sounds though a stethoscope placed on the brachial artery distal to the cuff. The systolic and diastolic pressures are determined by reading the manometer when the sounds begin and end, respectively. However, this technique requires skill and it is difficult to measure blood pressure on some obese individuals using this method.

For home blood pressure monitoring, convenient automatic devices have been developed and are commercially available. The measurement sites are the upper arm, wrist, and finger.

The most common method is to attach the cuff to the upper arm, and the systolic and diastolic pressures are determined automatically (Fig. 2). The cuff is inflated by an electric pump and deflated by a pressure-released valve. To determine the pressures, two different methods are used: Korotkoff sounds and an oscillometric method.

A microphone installed beneath the cuff detects the Korotkoff sounds and when the systolic and diastolic pressures are detected, a pressure sensor measures the obtained sounds and pressure at the critical points. The advantage of this method is that this measurement principle follows the standard auscultatory method. When the cuff is attached correctly, a reliable reading can be obtained.

The size of the cuff is important. The cuff should accurately transmit pressure down to the tissue surrounding

Figure 2. Automatic blood pressure monitor using both auscultatory and oscillometric methods.

the brachial artery. A narrow cuff results in a larger error in pressure transmission. The effect of cuff size on blood pressure accuracy for the Korotkoff method has been studied experimentally (1).

The oscillometric method detects the pulsatile components of the cuff pressure as shown in Fig. 3. When the cuff pressure is reduced slowly, pulses appear in the systolic pressure and the amplitude of the pulses increases and then decreases again. The amplitude of these pulses is always maximal when the cuff pressure equals the mean arterial pressure. However, it is difficult to determine the diastolic pressure from the signal measured from the

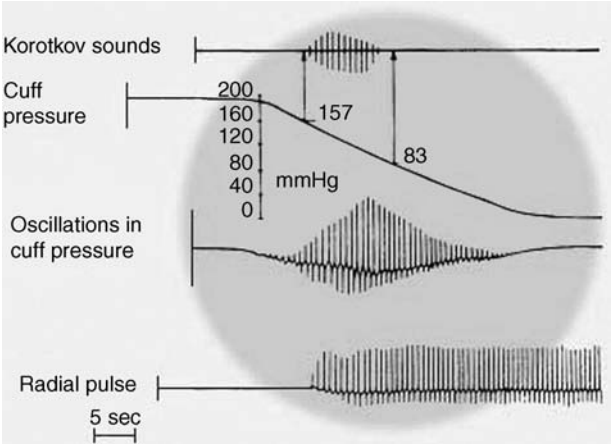


Figure 3. Principle of the oscillometric method. The pulsations induced by the artery differ when the artery is compressed. Initially, no pulsation occurs, and then the pulsation starts. As the pressure decreases in the cuff, the oscillation becomes more significant, until the maximum amplitude of the oscillations defines the average blood pressure. Then, the oscillations decrease with the cuff pressure until they disappear.

Figure 4. Wrist-type home blood pressure monitor. The measurement site is the wrist and during the measurement, the wrist must be at heart level.

cuff pressure. In general, the diastolic pressure is determined indirectly in commercial devices. One simple method that is often used is to calculate the diastolic pressure from the mean arterial pressure and systolic pressure (2). Several algorithms for this calculation have been used in commercial blood pressure monitors. The oscillometric method can only measure the cuff pressure.

Blood pressure measurement is not restricted to the upper arm. It is possible to measure blood pressure at the wrist and on a finger. However, if the measurement site is changed from the upper arm, the errors due to gravitational force and the peripheral condition increase. Wrist-type blood pressure monitors are now common in home use (Fig. 4).

Home blood pressure monitors are tested for accuracy against two protocols: the Association for the Advancement of Medical Instruments (AAMI) and the International Protocol of the European Society of Hypertension. Reports of the accuracy of home blood pressure monitors have been published (3). In addition, a 24 h home blood pressure monitor has been evaluated (4).

A home blood pressure monitor using the pulse wave transit time has also been studied. The principle used in this approach is that the arterial pulse wave transit time depends on the elasticity of the arterial vessel wall and the elasticity depends on the arterial pressure. Therefore, arterial pressure affects the pulse wave transit time. However, vascular elasticity is also affected by vasomotor activities, which depend on external circumstances; so this method is not reliable. Even so, with intermittent calibration we can estimate the blood pressure from the pulse wave transit time (5). The pulse wave transit time can be noninvasively determined from the arrival time of the arterial pulse at the beginning of cardiac contraction, which is determined from the QRS complex in an electrocardiogram.

ELECTROCARDIOGRAM

The electrocardiogram (ECG) gives important cardiac information. Recording the ECG at home can assist physicians to make a diagnosis. When monitoring the ECG at home

Figure 5. Holter ECG recorder. The Holter recorder is used for 24 h ECG monitoring and a digital Holter recorder is commonly used.

during either recovery from an acute disease or when the patient has a chronic disease, long-term recording is essential in order to detect rarely occurring abnormalities.

The Holter ECG recorder as shown in Fig. 5, has been widely used. It is a portable recorder that records the ECG on two or more channels for 24 or 48 h, on either an ordinary audiocassette tape or in a digital memory, such as solid-state flash memory. Most Holter recorders are lightweight, typically weighing 300 g or less, including the battery. The ECG must be recorded on the chest and electrodes need to be attached by clinical staff. A physician should also be available to monitor the ECG. Aside from these limitations, the Holter recorder can be used without obstructing a patient's daily life.

There are some special ECG recordings that can be taken in the home. The ECG can be recorded automatically during sleep and bathing.

In bed, the ECG can be recorded from a pillow and sheets or beneath the leg using electroconductive textiles (Fig. 6) (6). Since the contact between the textile electrodes and the skin is not always secure, large artifacts occur with body movements. In our estimation, 70–80% of ECGs during sleep can be monitored.

The ECG can also be recorded while bathing. If electrodes are installed on the inside wall of the bathtub as shown in Fig. 7, an ECG can be recorded through the water (7,8). The amplitude of the ECG signal depends on the conductivity of the tap water. If the conductivity is high, the water makes a short circuit with the body, which serves as the voltage source, and consequently the amplitude is reduced. If the water conductivity is low, however, the signal amplitude remains at levels similar to those taken on the skin surface. Fortunately, the electrical conductivity of ordinary tap water is on the order of $10^{-2}\,\mathrm{S{\cdot}m^{-1}}$, which is within the acceptable range for measurement using a conventional ECG amplifier. However, such an ECG signal cannot be used for diagnostic purposes because of the attenuation of the signal at lower frequencies.

HEART AND PULSE RATES

The heart rate (HR) is a simple indicator of cardiac function during daily life and exercise. The HR is the number of contractions of the heart per minute, and the pulse rate is defined as the number of arterial pulses per minute. Usually, both rates are the same. When there is an arrhythmia, some contractions of the heart do not produce effective ejection of blood into the arteries and this gives a lower pulse rate.

Figure 6. The ECG recorded from the bed. Electrodes are placed on the pillow and the lower part of the bed. An ECG signal can be obtained during sleep.

Figure 7. The ECG recorded from a bathtub. Silver–silver chloride electrodes are placed in the bathtub and the ECG signal can be obtained though the water.

The heart rate can be determined by counting the QRS complexes in an ECG or measuring the R–R interval when the cardiac rhythm is regular. In patients with an arrhythmia, the ECG waveforms are abnormal, and in this case detection algorithms with filtering are used. For accurate heart rate monitoring, electrodes are attached to the chest. Instead of surface electrodes, a chest strap is also available (Polar, Lake Success, NY)

The pulse rate can be obtained by detecting the arterial pulses using a photoplethysmograph, mechanical force measurements, vibration measurements, or an impedance plethysmograph. In photoplethysmograpy, the change in light absorption caused by the pulsatile change in the arterial volume in the tissue is detected. To monitor the pulse rate using photoplethysmography, the finger is commonly used. Light sources with wavelengths in the infrared (IR) region $\sim$ 800 nm are adequate for this purpose, because tissue absorbance is low and the absorbance of hemoglobin at this wavelength does not change with oxygen saturation.

The pulse oximeter, described below, can monitor both oxygen saturation and pulse rate. The pulsatile component of light absorption is detected and the pulse rate can be determined from the signal directly. The ring-type pulse oximeter is the most commonly used type of pulse oximeter.

A wristwatch type pulse rate meter is also available. It consists of a reflection-type photoplethysmograph. To measure pulse rate, the subject puts their fingertip on the sensor. A flashing icon on the display indicates a detected pulse, and the rate is displayed within 5 s.

The pulse rate can also be monitored in bed. In this case, the pulse rate is obtained directly from an electroconductive sheet. Vibration of the bed is detected by a thin flexible electric film (BioMatt, Deinze, Belgium) or an air mattress with a pneumatic sensor. In either case, the pulse rate is obtained through signal processing.

BODY TEMPERATURE

Body temperature has been checked at home for many years to detect fever. Frequent body temperature measurements are required for homecare in many chronic diseases. Basal body temperature measurement is also required when monitoring the menstrual cycle.

Stand-alone mercury-in-glass clinical thermometers have long been used both in clinical practice and at home, although they have recently been replaced by electronic thermometers because mercury contamination can occur if they are broken (Fig. 8). The ordinary electronic clinical thermometer uses a thermistor as a temperature sensor. The body temperature is displayed digitally. There are two types of clinical thermometer: the prediction and the real-time type. The real-time type waits until a stable temperature value is obtained. The prediction type attempts to predict the steady-state temperature using an algorithm involving exponential interpolation. The response time of a real-time electronic thermometer is 3 min and the response time of a prediction-type electronic thermometer is < 1 min, when both are placed in the mouth.

Figure 8. The electric thermometer contains a thermistor. Both predicting and real-time types are sold.

The tympanic thermometer as shown in Fig. 9, has become popular for monitoring the body temperature in children and the elderly because of its fast response. The device operates on the principle of IR radiation. The sensor is either a thermopile or pyroelectric sensor and is installed

Figure 9. The tympanic thermometer. Either a thermopile or a pyroelectric sensor is used as the temperature sensor. This device has a faster response than an electric thermometer.

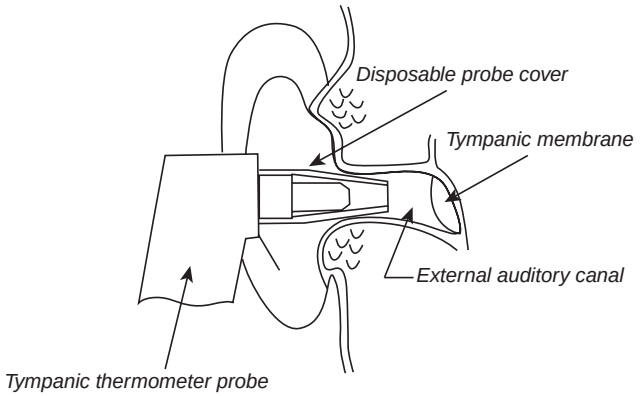


Figure 10. The principle of the tympanic thermometer. The sensor tip is inserted into the ear canal and the thermal distribution of tympanum is measured.

near the probe tip, as shown in Fig. 10. The probe tip is inserted into the auditory canal and the radiation from the tympanic membrane and surrounding tissue is detected. The tympanic temperature is close to the deep body temperature and the measurement can be made within a few seconds. Many studies have shown that when used properly, a tympanic thermometry is very accurate. However, IR tympanic thermometers produced measurements that were both less accurate and less reproducible when used by nurses who routinely used them in clinical practice (9,10).

A strip thermometer is sometimes used to monitor an acute fever. It is designed to be used once only and then discarded. It contains strips of thermosensitive liquid crystal that change color to indicate skin temperature, not body temperature. The color change is nonreversible. The strip is placed on the forehead and then read after 1 min. If a strip thermometer shows a high temperature, one should recheck the temperature with another type of thermometer.

BODY FAT

Body composition and body fat have both been proposed as indicators of the risk of chronic disease.

Body fat percentage is the proportion of fat in a person's body. Excess body fat was previously determined by measuring weight and comparing that value with height. Body fat is not always visible and cannot be measured on an ordinary scale. Obesity, which indicates a high degree of excess body fat, has been linked to high blood pressure, heart disease, diabetes, cancer, and other disabling conditions. To estimate the percentage of body fat, it is commonly derived from body density. The following equation gives an estimate of body density (D), which is then converted into the percent body fat (%BF) using the Siri equation:

$$\%\mathrm{BF} = (495/D) - 450$$

Body density, measured by weighting an individual while immersed in a tank of water, is based on Archimedes' principle and is a standard technique. However, this is not a convenient method for measurement in the home. Body volume can be determined from the air volume in an airtight chamber with the body inside by measuring the compliance of the air in the chamber (Bod Pod, Life Measurement Instruments, Concord, CA).

Figure 11. A scale with bioelectrical impedance analysis. This is a simple version of body impedance analysis using leg-to-leg bioimpedance analysis. The precision electronic scale has two footpad electrodes incorporated into its platform. The measurement is taken while the subject's bare feet are on the electrodes. The body fat percentage can be obtained from equations based on weight, height, and gender.

Body fat scales use the bioelectrical impedance analysis (BIA) technique. This method measures body composition using four electrodes, in which a constant alternating current (ac) of 50–100 kHz and 0.1–1 mA is applied between the outer electrode pair, and the alternating voltage developed between the inner electrode pair is detected (Fig. 11). Alternating current is applied between the toes of both feet, and the voltage developed between the electrodes at both feet is detected. The current passes freely through the fluids contained in muscle tissue, but encounters difficulty–resistance when it passes through fat tissue. This means that electrical impedance is different in different body tissues. This resistance of the fat tissue to the current is called bioelectrical impedance, and is accurately measured by body fat scales. Using a person's height and weight, the scales can then compute the body fat percentage. Recently, new commercial BIA instruments, such as the body segmental BIA analyzer, multifrequency BIA analyzer, lower body BIA analyzer, upper body BIA analyzer, and laboratory-designed BIA analyzers, have greatly expended the utility of this method (11). However, body composition differs by gender and race. Nevertheless, the impedance technique is highly reproducible for estimating the lean body mass (12).

The use of near-IR spectral data to determine body composition has also been studied (13). Basic data suggest that the absorption spectra of fat and lean tissues differ. The FUTREX-5000 (Zelcore, Hagerstown, MD) illuminates the body with near-IR light at very precise wavelengths (938 and 948 nm). Body fat absorbs the light, while lean body mass reflects the light. The intensity of back-scattered light is measured. This measurement provides an estimation of the distribution of body fat and lean body mass.

BLOOD COMPONENTS

In a typical clinical examination, the analysis of blood components is important. Medical laboratory generally use automatic blood analyzers for blood analysis. Usually, an invasive method is required to obtain a blood sample. Therefore, in a home healthcare setting, the monitoring and analysis of blood is uncommon, except for diabetic patients. In this section, we focus on the blood glucose monitor.

There are several commercial home blood glucose monitors. Self-monitoring of blood glucose (SMBG) is recommended for all people with diabetes, especially for those who take insulin. The role of SMBG has not been defined for people with stable type 2 diabetes treated with diet only. As a general rule, the American Diabetes Association (ADA) recommends that most patients with type 1 diabetes test glucose three or more times daily. Blood glucose is commonly measured at home using a glucose meter and a drop of blood taken from the finger (Fig. 12). A lancet device, which contains a steel needle that is pushed into the skin by a small spring, is used to obtain a blood sample. A small amount of blood is drawn into the lumen of the needle. The needle diameter is from 0.3 (30 G) to 0.8 (21 G) mm. In addition, laser lancing devices, which use a laser beam to produce a small hole by vaporizing the skin tissue, are available.

Once a small amount of blood is obtained, blood glucose can be analyzed using either a test strip or a glucose meter. The blood glucose level can be estimated approximately by matching the color of the strip to a color chart. In electrochemical glucose meters for homecare, a drop of blood of 10 μL or less is placed in the sensor chip. The blood glucose is measured by an enzyme-based biosensor. Most glucose meters can read glucose levels over a broad range of values, from as low as 0 to as high as 600 mg·dL. Since the range differs among meters, it is important to interpret very high or low values carefully. Glucose readings are not linear over their entire range.

Home blood glucose meters measure the glucose in *whole blood*, while most lab tests measure the glucose in *plasma*. Glucose levels in plasma are generally 10–15% higher than glucose measurements in whole blood (and this difference is even larger after a person has eaten). Many commercial meters now give results as the "plasma equivalent". This allows patients to compare their glucose measurements from lab tests with the values taken at home.

Minimally invasive and noninvasive blood glucose measurement devices are also sold. One of these uses near-IR spectroscopy to measure glucose. It is painless. There are increasing numbers of reports in the scientific literature on the challenges, strengths, and weaknesses of this and other new approaches to testing glucose without fingersticks (14,15).

The U.S. Food and Drug Administration (FDA) has approved minimally invasive meters and noninvasive glucose meters, but neither of these should replace standard glucose testing. They are used to obtain additional glucose values between fingerstick tests. Both devices require daily calibration using standard fingerstick glucose measurements.

The MiniMed system (Medtronic, Minneapolis, MN) consists of a small plastic catheter (a very small tube) inserted just under the skin. The catheter collects small amounts of liquid, which are passed through a biosensor to measure the amount of glucose present. The MiniMed is intended for occasional use and to discover trends in glucose levels during the day. Since it does not give readings for individual tests, it cannot be used for typical day-to-day

Figure 12. Glucose meter. This is used for self-monitoring blood glucose. The blood is taken from the fingertip and analyzed using a test strip.

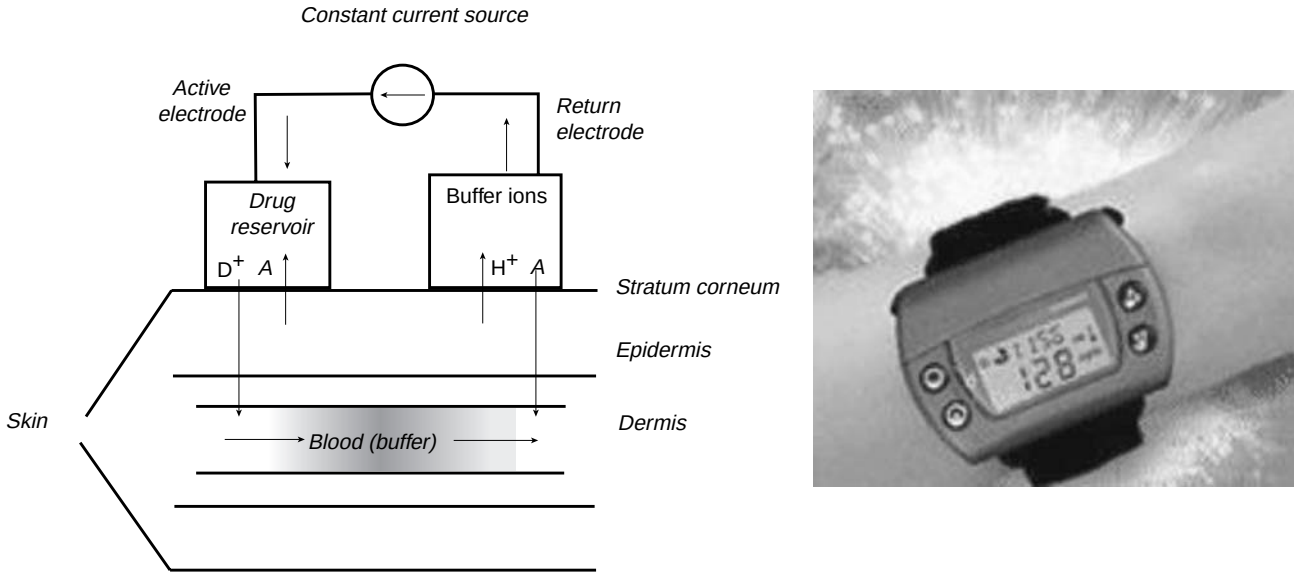


Figure 13. Glucowatch device (1) and principle of iontophoresis (2). This device provides noninvasive monitoring of glucose and uses reverse iontophoresis to extract glucose from the skin to monitor glucose. A low electric current is applied, which draws interstitial fluid through the skin. The glucose in this fluid is collected in a gel. A chemical process occurs, which generates an electrical signal that is converted into a glucose measurement.

monitoring. The device collects measurements over a 72 h period and then the stored values must be downloaded by the patient or healthcare provider.

GlucoWatch (Cygnus, Redwood City, CA) is worn on the arm like a wristwatch (Fig. 13). It pulls small amounts of interstitial fluid from the skin by iontophoresis and measures the glucose in the fluid without puncturing the skin. The device requires 3 h to warm up after it is put on the wrist. After this, it can measure glucose up to three times per hour for 12 h. The GlucoWatch displays results that can be read by the wearer, although like the MiniMed device, these readings are not meant to be used as replacements for fingerstick-based tests. The results are meant to show trends and patterns in glucose levels, rather than report any one result alone. It is useful for detecting and evaluating episodes of hyperglycemia and hypoglycemia. However, the values obtained must be confirmed by tests with a standard glucose meter before any corrective action is taken.

An elevated cholesterol level is one of the most important risk factors for coronary heart diseases. For home healthcare, a blood cholesterol test device is available. The test requires that a few drops of blood obtained from a finger stick sample be applied to the cholesterol strip, which contains cholesterol esterase and cholesterol oxidize. Total cholesterol, that is, the sum of free and esterified cholesterol, can be accurately and conveniently measured enzymatically using cholesterol oxidize and cholesterol esterase. The total amount of cholesterol is measured, and the results are obtained in 3–15 min.

URINE COMPONENTS

The analysis of urine components provides important diagnostic information for clinicians. Urine glucose and ketones indicate diabetes and urine protein indicates kidney disease. However, the only tool available for such testing is the urine test strip. A urine test can be done using a test strip without pain or discomfort. No fully automatic urine test system available, but there have been some attempts to monitor urine components at home with minimum disturbance. The instrument shown in Fig. 14 has been developed. It can be installed in the toilet and measures the urine glucose after a button is pushed (TOTO, Tokyo). The urine collector protrudes, collects urine automatically from the urine stream, and analyzes urine glucose within 1 min using an enzyme glucose sensor. The sensor must be replaced every 4 months and a calibration solution must be replenished every 3 months. This system is useful for monitoring the urine glucose level in diabetic patients.

BODY WEIGHT

Body weight monitored at home is an essential parameter for health management. To use body weight for health management, data must be taken regularly and stored. A digital scale connected to a laptop computer, together with temperature and blood pressure monitors, and a bed-sensor system has been developed (16).

A device to weigh the body automatically for health monitoring based on measurements on the toilet seat has been developed (17). A precision load cell system was installed in the floor of the toilet, and the seat was supported so that the weight on the seat was transferred to the load cell. This system also allows the measurement of urine and feces volume, urine flow rate, and the number and times of urination and evacuation.

For health management, the body mass index is commonly used. This is defined as the weight divided by the

Figure 14. Urine glucose monitor installed in the toilet. A small nozzle collects urine and then a biosensor analyzes urine glucose automatically.

square of the height. Excess body weight increases the risk of death from cardiovascular disease and other causes in adults between 30 and 74 years of age. The relative risk associated with greater body weight is higher among younger subjects (18).

NUTRITION

To prevent cardiac disease, diabetes, and some cancers, it is important to control body weight. The most accurate method that currently exists is to weigh foods before they are eaten. Like many other methods, however, this method can be inaccurate, time-consuming, and expensive. There are two basic ways to monitor nutrition. One is to monitor food intake.

Food consumed is photographed using a digital camera and the intake calories are calculated from the photographs (19,20). Digital photography and direct visual estimation methods, estimates of the portion sizes for food selection, plate waste, and food intake are all highly correlated with weighed foods.

The resting metabolism rate (RMR) is an important parameter for controlling body weight. The RMR represents the calories the body burns in order to maintain vital body functions (heart rate, brain function, and breathing). It equals the number of calories a person would burn if they were awake, but at rest all day. The RMR can represent up to 75% of a person's total metabolism if they are inactive or lead a sedentary lifestyle. Since the RMR accounts for up to 75% of the total calories we need each day, it is a critical piece of information for establishing appropriate daily calorie needs, whether one is trying to lose or maintain weight. Most healthcare and fitness professionals recognize that metabolism is affected by a variety of characteristics, such as fever, illness, high fitness, obesity, and active weight loss. When managing a subject's nutritional needs and calorie requirements, knowledge of their RMR is critical. Since metabolism differs individually, estimating the RMR value can lead to errors, and inaccurate calorie budgets. Consequently, individuals can be unsuccessful at reaching their personal goals, due to over- or undereating. As technology advances, professionals must reassess their practices. Caloric needs are assessed most accurately by measuring oxygen consumption and determining individual metabolism. Oxygen consumption estimates are obtained from the oxygen gas concentration and flow. Since it usually requires wearing a mask or mouthpiece, this measurement is difficult for some individuals. The BodyGem and MedGem (HealtheTech, Golden, CO) are devices that provide information vital for determining a personalized calorie budget, based on individual metabolism (Fig. 15). The BodyGem and MedGem consist of an ultrasound flow meter and fluorescence oxygen sensor with a blue LED excitation source, but the measurements are limited to an RMR monitor only. The RMR has been mentioned in the text.

We can also estimate the body's energy consumption from heat flow and acceleration measurements taken while an individual exercises (Body Media inc. Pittsburg, PA). For diabetes control, a pedometer with an accelerometer has been used and the energy consumption estimated using several algorithms.

Figure 15. A simple oxygen-uptake monitor. The subject wears the mask and a small ultrasonic flow meter measures the respiratory volume and a fluorescence oxygen monitor measures the oxygen concentration. This device is only used for measuring basal metabolism.

DAILY ACTIVITY

From the standpoint of health management, both the physical and mental health of an individual are reflected in their daily physical activities. The amount of daily physical activity can be estimated from the number of walking steps in a day, which are measured by a pedometer attached to the belt or waistband. To improve physical fitness 10,000 steps per day or more are recommended. For more precise measurement of physical activity, an accelerometer has been used. Behavior patterns, such as changes in posture and walking or running can be classified. The metabolic rate can be estimated from body acceleration patterns. The algorithms for calculating energy consumption differ for different pedometers. Each manufacturer has a different algorithm, and these have not been made public. However, the energy is likely evaluated using total body weight and walking time (21). This measurement requires attaching a device to the body, and requires continual motivation. An accelerometer equipped with a global positioning sensor has been developed and can monitor the distance and speed of daily activity (22).

There have been attempts to monitor daily activities at home without attaching any devices to the body. Infrared sensors can be installed in a house to detect the IR radiation from the body so that the presence or absence of a subject can be monitored, to estimate the daily activity at home, at least when the subject is living alone.

Other simple sensors, such as photointerrupters, electric touch sensors, and magnetic switches, can also be used to detect activities of daily living (23–25). The use of room lights, air conditioning, water taps, and electric appliances, such as a refrigerator, TV, or microwave oven, can be detected and used as information related to daily living. Habits and health conditions have correlated with these data to some extent, but further studies are required to give stronger evidence of correlations between sensor output and daily health conditions.

SLEEP

Sleep maintains the body's health. Unfortunately, in most modern industrial countries the process of sleep is disturbed by many factors, including psychological stress, noise, sleeping room temperature, and the general environment surrounding the bedroom. Insufficient sleep and poor sleeping habits can lead to insomnia. Another sleep problem is sleep apnea syndrome. In the laboratory, sleep studies aimed at the diagnosis of sleep apnea syndrome include polysomnography (PSG), electroencephalography (EEG), ECG, electromyography (EMG) pulse oximetry, and require chest and abdomen impedance belts. In the home, simple devices are required to evaluate sleep to determine if more detailed laboratory tests are needed.

A physical activity monitor actigraph (AMI, Ardsley, NY) can be used as a sleep detector. It is easy to wear and detects the acceleration of the wrist using a piezoelectric sensor. The wrist acceleration recorded by the actigraph accurately showed when the wearer was asleep (26).

Body movements during sleep can be measured without attaching sensors and transducers to the body using a pressure-sensitive sheet (BioMatt, VTT Electronics, Tampere, Finland). It consists of a 50 μm thick pressure- sensitive film, which can be installed under the mattress. This film is quite sensitive and not only detects body motions, but also respiration and heart rate. Therefore, it can be used as a sleep monitor for detecting insomnia and sleep disorders and as a patient monitor for detecting sleep apnea, heart dysfunctions, and even coughing and teeth grinding (27–29).

Body motion during sleep can also be monitored using a thermistor array installed on the bed surface at the waist or thigh level (30,31). The changes in temperatures show the body movement and sleep condition.

RESPIRATION THERAPY AND OXYGEN THERAPY

Respiration is the function of gas exchange between the air and blood in the body, and it consists of ventilation of the lung and gas transfer between the alveolar air and the blood in the pulmonary circulatory system. Lung ventilation can be monitored by either measuring the flow rate of the ventilated air or the volume change of the lung. Gas transfer is monitored by arterial blood oxygenation. Frequent respiratory monitoring is required for respiratory therapy at home.

Furthermore, many individuals have developed breathing difficulties as a consequence of increasing pollution, combined with an aging population.

For therapy, we use two types of respiration aid. One is for respiration related to cellular gas exchange. The other is for breathing difficulty, such as sleep apnea.

Reparatory therapy is included in the training of individuals involved in rehabilitation after thoracoabdominal surgery, in paraplegic or quadriplegic patients, and for patients requiring some form of mechanical ventilation. The fundamental parameters that must be monitored are the respiratory rate, respiratory amplitude, and respiratory resistance. Respiratory amplitude can be monitored using either airflow or lung movement.

For continuous monitoring of respiration in a home setting, it is inconvenient to use a mask or mouthpiece. Lung ventilation can be estimated by practice and from abdominal displacement. Inductance plethysmography has been used (32,33). This consists of two elastic bands placed at the rib cage and abdomen. Each band contains a zigzag coil and the inductance of this coil changes with its cross-sectional area. This system, Respitrace (Non-Invasive Monitoring Systems, North Bay Village, FL), gives the changes in volume of the rib cage and abdomen, tidal volume, and breathing rate. Respitrace was rated as the best noninvasive technology for the diagnosis of sleep-related breathing disorders by the American Academy of Sleep Medicine Task Force (1999).

Oxygen therapy, intermittent positive pressure breathing (IPPB) therapy, and respiratory assistance using a respirator can also be performed at home. In these situations, the arterial blood oxygenation must be monitored. Actually, there is a change in optical absorbance on the

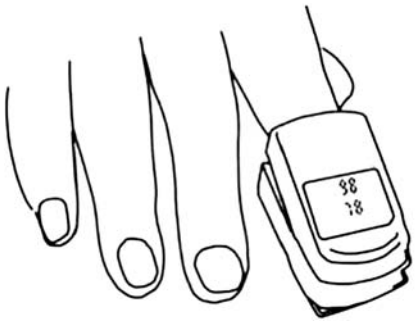


Figure 16. Pulse oximeter. A pulse oximeter is a simple noninvasive method of monitoring the percentage of hemoglobin (Hb) saturated with oxygen. It consists of a probe attached to the subject's finger. The device displays the percentage of Hb with oxygen together with an audible signal for each pulse beat and the calculated heart rate.

venous side that reflects changes in intrathoracic pressure due to breathing. Blood oxygenation is commonly monitored using a pulse oximeter, which can measure the oxygen saturation of arterial blood noninvasively from the light that is transmitted through a finger (Fig. 16).

The pulse oximeter is based on the principle that the pulsatile component in the transmitted light intensity is caused by the changes in the absorption of arterial blood in the light path while the absorption of the venous blood and tissue remains unchanged. The absorption spectrum of the blood changes with oxygen saturation, so the oxygen saturation of the arterial blood can be determined from the time-varying spectral components in the transmitted light. The oximeter contains two light-emitting diodes (LEDs), which emit light at two different wavelengths, and a photodiode to detect absorption changes at the two different wavelengths (Fig. 17). The measuring site is usually at a finger. However, a probe with a cable can sometimes disrupt the activities of daily life. A reflection-type probe that can be attached to any part of the body might be more convenient. Unfortunately, reflection-type probes are less reliable than transmission probes (34). A finger-clip probe without a cable (Onyx, Nonin Medical, Plymouth, MN) and a ring-type probe (35) are also available.

Recent advanced home healthcare devices are reviewed. These devices can be used effectively, not only for the elderly, but also for the middle-aged population and to establish home healthcare and telecare. Telecare and telemedicine are now popular for monitoring patients with chronic diseases and elderly people who live alone. The devices are placed in their homes and the data are transmitted to the hospital or a healthcare provider, who can check their clients' condition once every 12–24 h. Successful application has been reported for oxygen therapy and respiratory therapy.

We have solved several problems for more practical use. The major problems are the standardization of these devices and the agreement between medical use and home healthcare. Standardization of monitoring is important. For example, the principle of body impedance analysis differs for each manufacturer. Therefore, the values differ for different devices. This confuses customers, who then think that the devices are not reliable; hence, nobody uses such devices. There are similar problems with pedometers. Pedometers use either a mechanical pendulum or an accelerometer. The manufacturers should mention their limitations and reliability briefly, although most customers find this information difficult to understand.

The next problem is more serious. Some home healthcare devices have not been approved by health organizations, such as the FDA. For blood pressure monitors, a physician still needs to measure blood pressure during clinical practice even if the subject measures blood pressure at home. If the home healthcare device was sufficiently reliable, the physician would be able to trust the blood pressure values. Both researchers and members of industry must consider ways to solve this problem in the near future. There are additional social problems, such as insurance coverage of home healthcare devices, costs, handling, and interface design. The development of home heathcare devices must also consider the psychological and environmental factors that affect users. In the future, preventative medicine will play an important role in medical diagnosis. Hopefully, more sophisticated, high quality home healthcare devices will be developed. Technology must solve the remaining problems in order to provide people with good devices.

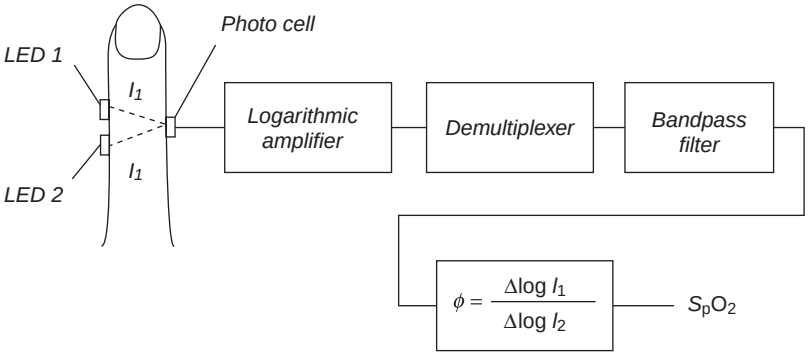


Figure 17. The principle of the oximeter. Hemoglobin absorbs light and the amount depends on whether it is saturated with oxygen. The absorption at two wavelengths (650 and 805 nm) is measured and used to calculate the proportion of hemoglobin that is oxygenated.

BIBLIOGRAPHY

Cited References

1. Geddes LA, Whistler SJ. The error in indirect blood pressure measurement with in correct size of cuff. Am Heart 1978; 96(July):4–8.
2. Sapinski A, Hetmanska ST. Standard algorithm of blood-pressure measurement by the oscillometric method. Med Biol Eng Comput 1992;30:671.
3. For example, Zsofia N, Katalin M, Gygorgy D. Evaluation of the Tensioday ambulatory blood pressure monitor according to the protocols of the British Hypertension Society and the Association for the Advancement of Medical Instrumentation. Blood Press Monit 2002;7:191–197.
4. O'Brien E, Atkins N, Staessen J. State of the market: a review of ambulatory blood pressure monitoring devices. Hypertension 1995;26:835–842.
5. Chen W, Kobayashi T, Ichikawa S, Takeuchi Y, Togawa T. Continuous estimation of systolic pressure using the pulse arrival time and intermittent calibration. Med Biol Eng Comput 2000;38:569–574.
6. Ishijima M. Monitoring of electrocardiograms in bed without utilizing body surface electrodes. IEEE Trans Biomed Eng BME 1993;40:593–594.
7. Ishijima M, Togawa T. Observation of electrocardiogram through tap water. Clin Phys Physiol Meas 1989;10:171–175.
8. Tamura T, et al. Unconstrained heart rate monito ring during bathing. Biomed Instrum Technol 1997;31:391–396.
9. Amoateng-Adjepong Y, Mundo JD, Manthous CA. Accuracy of an infrared tympanic thermometer. Chest 1999;115:1002–1005.
10. Robinson JL, Jou H, Spady DW. Accuracy of parents in measuring body temperature with a tympanic thermometer. BMC Family Pract. 2005;6:3.
11. Khan M, et al. Multi-dimension applications of bioelectrical impedance analysis. JEP Online 2005;8(1):56–71.
12. Segal KR, et al. Estimation of human body composition by electrical impedance methods. A comparison study. J Appl Physiol 1985;58:1565–1571.
13. Conway JM, Noms KH, Bodwell CE. A new approach for the estimation of body composition: infrared interactance. Am J Clin Nutr 1984;40:1123–1130.
14. Maruo K, Tsurugi M, Tamura M, Ozaki Y. *In vivo* non-invasive measurement of blood glucose by near-infrared diffuse-reflectance spectroscopy. Appl Spectrosc 2003;57(10): 1236–1244.
15. Malin SF, et al. Noninvasive prediction of glucose by near-infrared diffuse reflectance spectroscopy. Clin Chem 1999; 45:1651–1658.
16. Tuomisto T, Pentikäinen V. Personal health monitor for homes. ERCIM News 1997;29.
17. Yamakoshi K. Unconstrained physiological monitoring in daily living for healthcare. Frontiers Med Biol Eng 2000; 10:239–259.
18. Stevens J, et al. The effect of age on the association between body-mass index and mortality. N Engl J Med 1998;338:1–7.
19. Williamson D, et al. Comparison of digital photography to weighed and visual estimation of portion sizes. J Am Diet Assoc 2003;103:1139–1111.
20. Wang DH, Kogashiwa M, Ohta S, Kira S. Validity and reliability of a dietary assessment method: the application of a digital camera with a mobile phone card attachment. J Nutr Sci Vitaminol (Tokyo) 2002;48:498–504.
21. Tharion WJ, et al. Total energy expenditure estimated using a foot-contact pedometer. Med Sci Monit 2004;10(9):CR504–509.
22. Perrin O, Terrier P, Ladetto Q, Merminod B, Schutz Y. Improvement of walking speed prediction by accelerometry and altimetry, validated by satellite positioning. Med Biol Eng Comput 2000;38(2):164–168.
23. Celler BG, et al. Remote monitoring of health status of the elderly at home. A multi- disciplinary project on aging at the University of New South Wales. Intern J Bio-Medical Comput 1995;40:144–155.
24. Suzuki R, Ogawa M, Tobimatsu Y, Iwaya T. Time course action analysis of daily life investigation in the welfare techo house in Mizusawa. J Telemed Telecare 2001;7:249–259.
25. Ohta S, Nakamoto H, Shinagawa Y, Tanikawa T. A health monitoring system for elderly people living alone. J Telemed Telecare 2002;8:151–156.
26. Sadeh A, Hauri PJ, Kripke DF, Lavie P. The role of actigraphy in the evaluation of the sleep disorders. Sleep 1995;18: 288–302.
27. Salmi T, Partinen M, Hyyppa M, Kronholm E. Automatic analysis of static charge sensitive bed (SCSB) recordings in the evaluation of sleep-related apneas. Acta Neurol Scand 1986;74:360–364.
28. Salmi T, Sovijarvi AR, Brander P, Piirila P. Long-term recording and automatic analysis of cough using filtered acoustic signals and movements on static charge sensitive bed. Chest 1988;94:970–975.
29. Sjoholm TT, Polo OJ, Alihanka JM. Sleep movements in teethgrinders. J Craniomandib Disord 1992;6:184–191.
30. Tamura T, et al. Assessment of bed temperature monitoring for detecting body movement during sleep: comparison with simultaneous video image recording and actigraphy. Med Eng Phys 1999;21:1–8.
31. Lu L, Tamura T, Togawa T. Detection of body movements during sleep by monitoring of bed temperature. Physiol Meas 1999;20:137–148.
32. Milledge JS, Stott FD. Inductive plethysmography--a new respiratory transducer. J Physiol 1977;267:4P–5P.
33. Sackner JD, et al. Non-invasive measurement of ventilation during exercise using a respiratory inductive plethysmograph. I Am Rev Respir Dis 1980;122:867–871.
34. Mendelson Y, Ochs BD. Noninvasive pulse oximetry utilizing skin reflectance. IEEE Trans Biomed Eng 1988;35:798–805.
35. Rhee S, Yang BH, Asada HH. Artifact-resistant power-efficient design of fingerring plethysmographic sensors. IEEE Trans Biomed Eng 2001;48:795–805.

See also HUMAN FACTORS IN MEDICAL DEVICES; MOBILITY AIDS; NUTRITION, PARENTERAL; QUALITY-OF-LIFE MEASURES, CLINICAL SIGNIFICANCE OF; TEMPERATURE MONITORING; TRANSCUTANEOUS ELECTRICAL NERVE STIMULATION (TENS).

HOSPITAL SAFETY PROGRAM.

See SAFETY PROGRAM, HOSPITAL.

HUMAN FACTORS IN MEDICAL DEVICES

DANIEL W. REPPERGER
Wright-Patterson Air Force Base
Dayton, Ohio

INTRODUCTION

The human factors issues related to the use and design of medical devices has experienced significant paradigm shifts since this topic was addressed >20 years ago (1). Not only has the technology innovation of the Internet

vastly affected how medical professionals both gather and report information, but also standards are now more easily established. In addition, technology in the healthcare industry has concomitantly made significant advances. The evolving characteristics of legal liability with medical devices have also changed. Concurrently, the skill and sophistication of users with computer-aided systems has significantly improved with more tolerance and acceptance of automation. The computer and microprocessor-based medical devices are now the pervasive means of humans dealing with mechanical–electrical systems. First, it is important to define the term Human Factors within the context of medical devices and biomedical engineering. The phrase human factors can be broadly characterized as the application of the scientific knowledge of human capabilities and limitations to the design of systems and equipment to generate products with the most efficient, safe, effective, and reliable operation. A modern expression to describe a systematic procedure to evaluate risk when humans use medical devices is termed human factors engineering (2). The U.S Food and Drug Administration (FDA) is a strong proponent of the use of human factors engineering to manage risk, in particular with application to medical devices. The responsibility of the FDA is to guarantee the safety and efficacy of drugs and medical devices. Since, in the United States, the FDA is one of the leading authorities on medical standards, it is worthwhile to review (3,4) their interpretation on how human factors studies should be conducted with medical device use as perceived by this group. Other U.S. government organizations, such as the National Institute of Health (NIH) (5) and the Agency for Health Care Research and Quality (6), also offer their perspective on the application of human factors studies with respect to the manipulation of medical devices. Other sources of government information are found at (7–12). Also available online are a number of legal sources (13–15) related to injury issues and instrumentation affiliated with healthcare and how they perceive the relevance of human factors engineering. In these sources, there is a strong influence on how human factors procedures have some bearing on liability, abuse, misuse, and other troublesome issues associated with medical devices (16).

From an historical perspective, in the late 1980s, data collected by the FDA demonstrated that almost one-half of all medical device recalls resulted from design flaws. In 1990, the U.S. Congress passed the Safe Medical Devices Act, giving the FDA the ability to mandate good manufacturing practices. These practices involve design controls for manufacturers to use human factors engineering principles within medical device design.

In this article, we initially discuss human factors as defined by the FDA followed by three classic case studies. The ramifications of legal issues are then presented. Concurrent good human factors methods are then described, followed by some key topic areas including alarms, labeling, automation, and reporting. Future issues regarding human factors and medical devices are subsequently offered with conclusions and future directions of this field depicted.

First, it is instructive to review the present state of affairs on how the FDA defines human factors engineering within the context of when humans interact with medical devices. The term human factors engineering is a persuasive term in the literature describing present FDA standards.

A HUMAN FACTORS ENGINEERING PERSPECTIVE FROM THE FDA

The goal of the FDA is to promote medical device designers to develop highly reliable devices. Human factors engineering is a phrase used to help understand and optimize how people employ and interact with technology. A host of literature describes human factors engineering in many eclectic areas (17–30). When medical devices fail or malfunction, this impacts patients, family members, and professional healthcare providers. A common term used to characterize the potential source of harm is a hazard. A hazard may arise in the use of a medical device due to the inherent risk of medical treatment, from device failures (malfunctions) and also from device use. Figure 1, from the FDA, displays possible sources of device failure hazards that impact the human factors issues in medical devices. Figure 1 may be deceptive in the presumption that equal hazards exist between use related and device failure. More correctly (3,4) the use contribution to the total medical devices hazards may far exceed those from the device failures. In fact, from an Institute of Medicine report (31), as many as 98,000 people die in any given year from medical errors that occur in hospitals. This is more than the number who die from motor vehicle accidents, breast cancer, or acquired immune deficiency syndrome (AIDS). A proportion of these errors may not directly be attributed to the medical device itself; however, the importance of incorporating human factors engineering principles into the early design and use of these important interfaces is a key concern. It is instructive to examine the two major types of errors (hazards) in Fig. 1 and how they are delineated. Risk analysis will refer to managing the forms of risks to be described herein. After the hazards are first clarified, the goal is for the hazards to be mitigated or controlled by modifying the device user interface (e.g., control or display characteristics, logic of operation, labeling) or the background of the users employing the device (training, limiting the use to qualified users). The power in the human factors approach is to help identify, understand, and address use-related problems as well as the original design problem with the physical device itself prior to its acceptance in the workplace of the healthcare professional. Some institutions have now developed in-house usability laboratories, in order to rigorously test any medical device

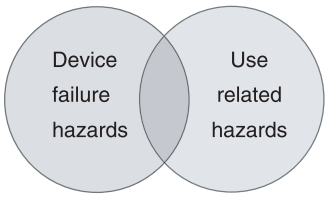


Figure 1. Hazards from device failure and use related.

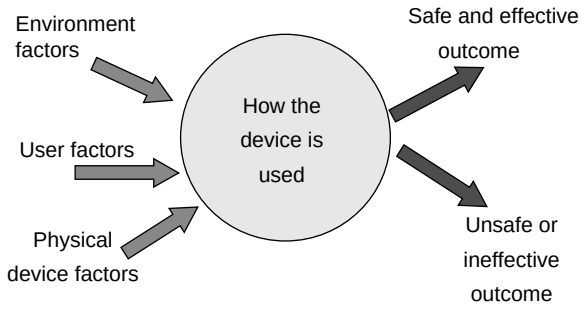


Figure 2. Human factors consideration in medical device use.

before utilization. It is worthwhile to first elaborate on use related hazards as they are relevant in this area.

Use-Related Hazards

Addressing the hazards related to device use, the essential components include (*1*) device users (patient, caregiver, physician, family member, etc.); (*2*) typical and atypical device use; (*3*) characteristics of the environment for the application of the medical device; and (*4*) the interaction between users, devices, and use environments.

Figure 2 portrays an abstraction on all possible uses for a medical device.

Device Failure Hazards

When understanding hazards from the perspective of risk analysis, it is common to consider the following classes of hazards as they pertain to device design: (*1*) chemical hazards (e.g., toxic chemicals); (*2*) mechanical hazards (e.g., kinetic or potential energy from a moving object); (*3*) thermal hazards (high temperature components); (*4*) electrical hazards (electric shock, electromagnetic interference); (*5*) radiation hazards (ionizing and nonionizing); and (*6*) biological hazards (allergic reactions, bioincompatibility, and infection).

In an effort to address good human factors design when dealing with medical devices, it is instructive to now discuss three classic misadventures in the medical device arena when human factors procedures could have been modified to preclude untoward events. One usually thinks of medical errors occurring, for example, in surgery (wrong site surgery), as the amputation of the wrong appendage (32,33) or from chemotherapy overdoses (34). In 2005, it is now a common practice, in holding areas before surgery, for the surgeon and medical team to discuss with the patient the impending surgery and to have the patient mark on his body precisely where the surgery will be performed with a magic marker. This procedure assures the patient that no confusion may occur as a function of a patient mix-up, after the patient is under anesthesia. Three examples are now presented of untoward events that could have been prevented with improved human factors use methodologies. In the final analysis, the enemy of safety is complexity, which appears in these case studies. Complex systems fail because of the contribution of multiple small failures, each individual failure may be insufficient to cause an accident, but in combination, the results may be tragic.

EXAMPLES OF CASE STUDIES WHERE HUMAN FACTORS ISSUES ARISE

It is worthwhile to examine a few classic case studies where human factors procedures interacting with medical devices needed to be reevaluated. It is emphasized that the errors described herein may not be attributed to any one person or system. Rather, the complex interaction of humans with poorly defined procedures involving certain medical devices has given rise to events, which were not planned or expected. However, with a more structured interaction of humans with these systems, improved results could be obtained. The reference by Geddes (35) provides many interesting examples where enhanced human factors planning would have prevented errors in medical treatment.

Case Study 1 from Ref. 35 and Reported in the South African *Cape Times* (1996)

"For several months, our nurses have been baffled to find a dead patient in the same bed every Friday morning" a spokeswoman for the Pelonomi Hospital (Free State, South Africa) told reporters. "There was no apparent cause for any of the deaths, and extensive checks on the air conditioning system, and a search for possible bacterial infection, failed to reveal any clues."

Although device failure could cause such deaths, why they occurred on Fridays is difficult to understand? The *Cape Times* later reported:

> It seems that every Friday morning a cleaner woman would enter the ward, remove the plug that powered the patient's life support system, plug her floor polisher into the vacant socket and then go about her business. When she had finished her chores, she would then replug the life support machine and leave, unaware that the patient was now dead. She could not, after all, hear the screams and eventual death rattle over the shirring of her polisher.
> "We are sorry, and have sent a strong letter to the cleaner in question. Further, the Free State Health and Welfare Department is arranging for an electrician to fit an extra socket, so there should be no repetition of this incident. The inquiry is now closed."

This example emphasizes that when unplanned influences or events interact with medical device operation, tragic results may occur. In a later section, we describe several human factors procedures and techniques that are now designed to help preclude these types of untoward events.

In Ref. 36, a second example shows how important (delicate) care is requisite to providing appropriate interaction of humans with medical devices.

Case Study 2 from Ref. 36

Besides putting patients at high risk for injury, clinicians who use a device they are not familiar with are placing themselves in legal jeopardy. The case of *Chin vs. St. Barnabos Medical Center* (160 NJ 454 [NJ 1999]) illustrates

this point. A patient died after gas inadvertently was pumped into her uterus during a diagnostic hysteroscopy. Evidence established that the two perioperative nurses implicated in the case had no experience with the new hysteroscope and used the wrong hook up for the procedure. They connected the exhaust line to the outflow port. The 45-year-old patient died from a massive air embolism. It was also discovered that the nurses never received any education regarding the device. The manufacturer was not found liable because records indicated that the device did not malfunction or break. Damages in the amount of $2 million were awarded to the plaintiff and apportioned among the surgeon, nurses, and hospital. As discussed in the sequel, new training methods have been developed in the human factors area to preclude the occurrence of events of this type.

The last case study deals with a poor interface design. Usability Testing, to be discussed later, provides methods to preclude some of these difficulties encountered.

Case Study 3 from Ref. 37

"A 67 year-old man has ventricular tachycardia and reaches the emergency room. Using a defibrillator, nothing happens. The doctor suggests the nurse to start a fluid bolus with normal saline. The nurse opens the IV tubing wide open, but within minutes the patient starts seizing. The nurse then realizes that the xylocaine drip instead of the saline had been inadvertently turned up. The patient is then stabilized and the nurse starts her paperwork. She then realizes that the defibrillator was not set on the cardio version, but rather on an unsynchronized defibrillation. This is because the defibrillator she uses, every day, automatically resets to the nonsynchronized mode after each shock. The second shock must have been delivered at the wrong time during the cardiac cycle, causing ventricular fibrillation." By performing a usability study as described later, with better training, this type of event could have been prevented.

The changing effect of legal influences also has had its impact on human factor interactions of caregivers with medical devices. It is worthwhile to briefly describe some of these recent influences and how they impact on human dealings with medical devices.

NEW LEGAL INFLUENCES THAT AFFECT HUMAN FACTORS

As mentioned previously in Refs. 13–15, there have been a number of modifications in the legal system that affect how humans now interact with medical devices. It is not unusual in the year 2005 to hear prescription drugs advertised on television or on the radio with a disclaimer near the end of each commercial. The disclaimer lists major possible side effects and warns the user to discuss the drug with their physician. As the aging "baby boomers" of the post-World War II era now require more and more medications, the drug companies make major efforts and studies to ensure that safe and effective drugs are delivered to an ever increasing public audience. Experience from prior mistakes has significantly modified how the drug industry must deal with a larger, and more highly informed, population base. Some major modifications that occur involving legal issues and medical device operation include

1. The legal profession (13) now recognizes the importance of human factors engineering [also known as usability engineering or ergonomics (a term used outside the United States)] in studying how humans interact with machines and complex systems. Ergonomics is a factor in the design of safe medical devices; A user-friendly device is usually a safe one (38).
2. Healthcare Institutions employ results of human factors engineering testing of devices in making key decisions in evaluation as well as major purchase judgments. If these components fail, but have been tested within a rigorous framework, the legal consequences are mitigated since standards were adhered to in the initial selection of the medical device and procedure of use.
3. Insurance premiums to Healthcare Institutions are correspondingly reduced if adherence to standards set forth by good human factors engineering principles are maintained. The insurance costs are directly related to the potential of legal expenses and thus sway the decisions on how medical devices are purchased and used.

A number of new ways of improving how human factor design with medical devices has evolved, which will now be discussed as pertinent to the prior discussion. These methods affect training, utilization procedures, design of the devices, testing, and overall interaction with caregivers in the workplace.

METHODS TO IMPROVE HUMAN FACTOR DESIGN AND MEDICAL DEVICES

Professionals working in the area of human factors have now devised a number of new means of improving how healthcare professionals can better deal with medical devices. We present some of the most popular methods in the year 2005, most of which now pervasively affect the development of a user's manual, training, and manner of use with respect to medical devices. Some of these techniques appear to have overlap, but the central theme of these approaches is to better assist the healthcare professional to mitigate untoward events. One of the most popular methods derived from human factors studies is cognitive task analysis (CTA). In short, the way that CTA works is that a primary task is subdivided up into smaller tasks that must be performed. It is necessary to specify the information needed to perform each subtask, and the decisions that direct the sequence of each subtask. Note, this type of task description is independent of the automation involved. For example, for the same tasks, information and decisions are required regardless of whether they are performed by a human or a machine. Also considered in this analysis are the mental demands that would be placed on the human operator while performing these selected subtasks.

Cognitive Task Analysis and How it Affects Human Factors and Medical Devices

The principle behind CTA is to take a large and complex task and divide it up into smaller subtasks (39). Each subtask should be attainable and reasonable within the scope of the user. If the subtask is not yet in the proper form, further subdivisions of that subtask are performed until the final subtask is in the proper form. The expression "proper form" implies the subtask is now attainable, sufficiently reasonable to perform, the proper information has been provided, and sufficient control is available to execute this subtask. An important aspect of CTA is to define if the user has the proper information set to complete his duties and also has the proper control means over the situation so that the task can be achieved adequately. Cognitive task analysis has great value in establishing a set of final subtasks that are both attainable and relevant to the overall mission. With this analysis, the caregiver can be provided better training and have an enhanced understanding of the role of each task within the overall mission. This procedure has been used in the nursing area for assessing risk of infants (40) and for patient-controlled analgesia machines (41). It has been noted (42) that 60% of the deaths and serious injuries communicated to the Medical Device Reporting system of the U.S. Food and Drug Administration (FDA) Center for Devices and Radiological Health have been attributed to operator error. Cognitive task analysis has evolved out of the original area of Critical Decision Methods (43) and is now an accepted procedure to analyze large and complex interactions of humans with machines.

A second popular method to shape procedures to interact with medical devices involves User Testing and Usability Engineering.

User Testing and How It Affects Human Factors and Medical Devices

User Centered Design has found popularity when humans have to interact with medical devices for ultrasound systems (44) and for people with functional limitations (45). Usability engineering methods are applied early in the system lifecycle to bridge the gap between users and technology. The ultimate goal is to design an easy to use system that meets the needs of its users. A basic principle of user-centered design is making design decisions based on the characteristics of users, their job requirements and their environments (46–50). It is often the complaint of human factors professionals that they are brought into the design process much too late to influence the construction of the overall system. It is all too common for the experimental psychologist to have to deal with an interface built without prior considerations of the human's limitations and preferences in the initial design construction. The usability engineering methods bring to light needs for users and tasks early on, and suggest specific performance testing prior to the recommendation of the final design of an interface.

A third method discussed here to influence how to work with medical devices involves Work Domain Analysis.

Work Domain Analysis and How It Affects Human Factors and Medical Devices

A third and popular method to better understand the interplay between human factors issues and medical devices is via an integrated method involving the technical world of physiological principles and the psychological world of clinical practice. This work domain analysis was originally proposed by Rasmussen et al. (51,52) and has found application in patient monitoring in the operating room (53). A variety of tables are constructed based on data to be utilized. Columns in the tables include a description of the task scenario and the relations between key variables of the work environment and the work domain. The tables portray interactions and different strategies that are elicited to help in monitoring and control.

As discussed earlier in Ref. 1, today many new advances have also been made in alarms. These inform the healthcare provider of troublesome events and many innovative changes and studies have been instituted in this area that warrant discussion. Alarm deficiencies compromise the ability to provide adequate healthcare. For alarms, some research has focused on the identification of alarm parameters that improve or optimize alarm accuracy (i.e., to improve the ratio of true positives to false positives: the signal/noise ratio).

ALARMS AND HUMAN FACTORS ISSUES

Alarms are key to the detection of untoward events when humans interact with medical devices. One does not want to generate designs that invite user error. We cannot deal with confusing or complex controls, labeling, or operation. Many medial devices have alarms or other safety devices. If, however, these features can be defeated without calling attention to the fact that something is amiss, they can be easily ignored and their value is diminished (54). The efficacy of alarms may be disregarded because it is not attention getting. For example, if a multifunction liquid-crystal display (LCD) has a low battery warning as its message, but is not blinking, it does not call attention to itself. Alternatively, an improved design occurs in the case of a low battery alarm design commonly found in household smoke detectors. In this case, a low battery will cause the unit to chirp once a minute for a week, during which the smoke detector is still functional. The chirp may be confusing at first, but it cannot be ignored for a week. A battery test button is still available for the testing when the battery power is satisfactory. Adequate alarm systems are important to design in a number of analogous medical scenarios. For example, use of auditory systems (55) is preferable to a visual display, since this reduces the visual workload associated with highly skilled tasks that may occur, for example, in the operating room. For anesthesia alarms (56), care must be exercised to not have too many low level alarms that indicate, for example, that limits are exceeded or that the equipment is not functioning properly. The danger of false positives (alarms sounding when not necessary) provides an opportunity for the user to ignore information, which may be critical in a slightly different setting. An example where information of this type cannot be

ignored is in applications of human factors training to the use of automated external defibrillators. It is known that without defibrillation, survival rates drop by 10% for every minute that passes after cardiac arrest (57). A great deal of work still continues in the area of management of alarm systems in terms of their efficacy and utility (58,59).

Another significant issue with human factors is proper labeling. Medication misadventures are a very serious problem. We briefly summarize changes that impact how humans will interact with their prescriptions as well as medical devices in general and how they are influenced by their labeling constraints.

LABELING AND HUMAN FACTORS ISSUES

By government regulation and industry practice, instructions accompanying distribution of medical devices to the public are termed "labeling". Medical device labeling comprises directions on how to use and care for such practices. It also includes supplementary information necessary for the understanding and safety, such as information about risks, precautions, warning, potential adverse reactions, and so on. From a humans factors perspective, the instructions must have the necessary efficacy, that is, they must provide the correct information to the user. There are a number of standards on how the instructions must be displayed and their utility in the healthcare industry (60). For example, for prescription medications (61–64), they represent the most important part of outpatient treatment in the United States. This provides >2 billion possible chances for patient error each year in the United States. To maximize the benefits and minimize the dangers of using these medications, users must comply with an often complex set of instructions and warnings. Studies show that seven specific system failures account for 78% of adverse drug events in hospitals. All seven of these failures could be corrected by better information systems that detect and correct for errors. The top seven system failures for prescription medications are (*1*) drug knowledge dissemination; (*2*) dose and identification checking; (*3*) patient information availability; (*4*) order transcription error; (*5*) allergy defense; (*6*) medication order tracking; (*7*) improved interservice communication.

As mentioned previously, as computers and the Internet become more persuasive, patients, caregivers, doctors, and others become more tolerant and dependent on automation. The goal is to make a task easier, which is true most of the time. There are a number of issues with regard to automation that need to be addressed.

AUTOMATION ISSUES AND HUMAN FACTORS

As computers and microprocessor-based devices have become more ubiquitous in our modern age, there is increased tendency to foster automation (65) as a means of improving the interaction of users with medical devices. The original goal of automation was to reduce the workload (physical and mental) and complexity of a task to the user. This is specific to the desired response from the device of interest. There is an obvious downside of this concept. The idea that the automation has taken over and has a mind of its own is ghastly within human thinking. Also, if the automated system is too complex in its operation and the user is not comfortable in understanding its causality (input–output response characteristics), the trust in the device will decrease accordingly and the human–machine interaction will degrade. The classical work by Sheridan (66) defines eight possible levels of automation, as portrayed in Fig. 3. One easily sees the relationship of loss of control to increased automation. For medical device usage, this may be problematic to trade off simplicity of use to loss of control and eventual efficacy. Automation studies continue to be of interest (67–73).

REPORTING AND HUMAN FACTORS ISSUES

Reporting of failures of proper medical device operation has now commonly advanced to Web-based systems (3,74). The healthcare provider must be increasingly skilled with the use of computer systems. Sometimes the terms digital

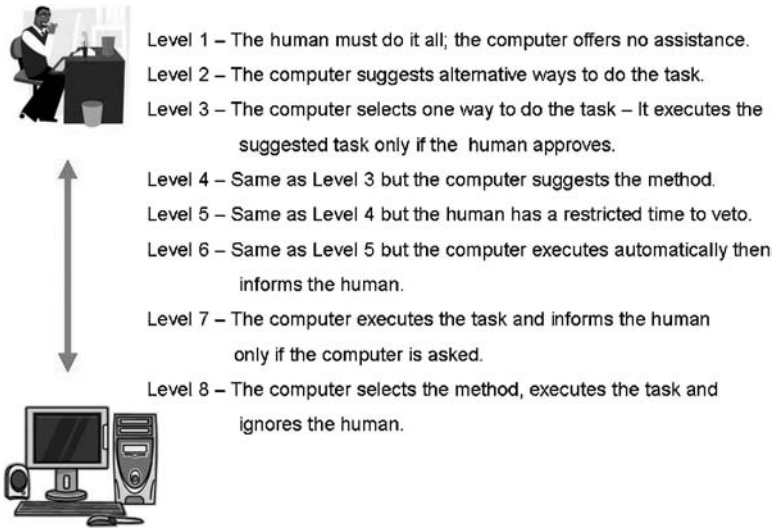


Figure 3. The eight levels of automation.

divide is employed to distinguish those that have the requisite computer skills from those that are not as competent in this area. This may be a consequence of humans factors procedures instituted to deal with a more elderly patient group (75–82) who may not be comfortable with computers. Education is the best means to deal with this problem (83). It is important that modern healthcare givers who use medical devices in their work setting have the obligatory skills to accurately report failures and have the suitable computer training to make relevant reports to the necessary sources.

A number of new areas are growing and influence how human factors have been evolving with the interaction of medical devices. It is important to spend some time mentioning these contemporary and emergent areas.

NEW ISSUES INVOLVING HUMAN FACTORS AND MEDICAL DEVICES

With advances in technology, several new areas should be mentioned that seem to have relevance to novel advances in medical devices. The concept of telemedicine is now a developing field that certainly addresses the future uses of medical devices.

The Growth of Telemedicine

The term telemedicine literally means "medicine at a distance" (84) and is now an increasing and popular means of providing healthcare. The advantages are obvious in rural settings and numerous other scenarios (cf. Fig. 4). For example, having an expert physician or surgeon located at a remote and safe location, but performing a medical procedure on a person in a hazardous or distant environment provides a distinct advantage. The human at the hazardous environment may be placed in a battlefield in a combat situation, they may be in a space shuttle, or simply be in another country or distant location from the expert medical practitioner. Another simple example of telemedicine occurs in a simple endoscopic (colonoscopy) procedure or in a laparoscopic operation, for example, for knee surgery. For these procedures, a small insertion is made into the patient and the process is carried out from a remote location with the physician observing the process on a television monitor. The advantages are obvious: (*1*) with smaller entrance incisions or openings into the patient, the trauma is significantly reduced; (*2*) recovery time is much quicker; and (*3*) the risk of infection is substantially mitigated. Some of the new human factors issues regarding medical devices used within this context include (*1*) Dealing with the time delay between observing and making actions and the associated instabilities that may occur in the closed loop interaction; (*2*) having the lack of a sense of presence about a remote environment; and (*3*) having a strong dependence on mechanical systems or automation at the end-effector of the device inside the patient. These methods have been advanced to the point where robots are now being used to perform heart surgery, and so on, and some operations have been conducted over the Internet. Recent applications of robots performing heart procedures on humans need only two very small incisions in the patient. This allows for much quicker recovery time to the patient (1 vs. 7 weeks for a typical open heart surgery). From the surgeon's perspective, a significant advantage for the small incisions is that: It is not necessary to insert my hands inside the chest cavity. Thus the size of the incisions can be substantially reduced. One sees the disadvantage of traditional surgery in this area because *it is the size of the surgeon's hands being required to be inside the chest cavity as the only reason for a large incision in the chest cavity*. When the surgeon's hands no longer have to be inside the chest cavity, then the correspondingly two small incisions give rise to reduced possible infection, less trauma to the patient, and a shorter recovery time before the patient can return to normal work and living activities. This area of healthcare will only continue to advance as medical practices move more and more to this form of undertaking. In recent times, a number of studies in teleoperation have shown the efficacy of improving the sense of presence of the operator about the remote environment through "haptic" feedback. Haptic refers to forces reflected back on the physician (by various robotic interface devices) to improve their sense of presence about the remote environment, so the operator can "feel" the task much as they see it on a television monitor. A number of studies have shown both an improved sense of presence and performance about these teleoperation scenarios using haptic feedback (85–88). Auditory systems have also found analogous use in surgery (89). The problems in anesthesia are also well studied (90–93).

Figure 4. The concept of telemedicine.

Another growth area includes more participation by the patient directly in their own healthcare.

The Growth of Increased Patient Participation in the Healthcare Process

As discussed previously, with the pervasive nature of automation, more and more of the healthcare responsibility and work will be performed by the patient, themself. Modern insurance procedures also encourage additional homecare scenarios and many times without a trained caregiver. This saves expensive care at the hospital, but transfers the burden onto the patient or their family members to become the primary caregiver. A paradigm

shift of this type is a consequence of present insurance reimbursement procedures requiring the patient to now spend much of their time away from the hospital. New human factors issues are consequently introduced when dealing with medical devices in this scenario. The design of the human computer interface (94–97) now becomes critical for the efficacy of the healthcare provided. Even in cancer treatment, the responsibility of the proper administration of radioisotopes may become the burden of the patient (98) or if they have to manipulate their own chemotherapy level. For pain treatment (99–101) the patient has to be proactive in the selection of the analgesia level of the device provided. Modern TENS (Transcutaneous Electrical Nerve Stimulator) units now have been constructed to be wireless and shown to have equivalent efficacy in terms of pain management as compared to long wired units that have been in existence for >40 years (102). The movement to more wireless medical devices is certainly in the future. For the TENS units, by eliminating the long and entangling wires, this provides more reliability, less chance of wires breaking or shorting, more efficient use of electric power, but different forms of control with these analgesia devices. For example, the physician or caregiver may use a remote control to program the voltage level of the TENS microprocessor in a wireless manner rather than making manual adjustments with the traditional, long wired, TENS devices.

A third area of modern concern is the impact of electromagnetic fields on the operation of other medical devices, especially if they are implanted.

The Growth of Electromagnetic Fields on the Operation of Other Medical Devices

The Geddes reference (35) describes numerous examples of documented problems when medical devices inadvertently interact with unexpected exposure to external electromagnetic fields. Electromagnetic interference (EMI) is used to describe the malfunction of a device exposed to electromagnetic waves of all types that propagate through space. The EMI can intervene with electrodes on a patient, it can bias results for EEG recording of generalized epileptiform activity, and can give false positives to alarm systems, Silbert et al. (103). Other cases where malfunctions can occur involve heart machines, apnea monitors, ventilator mishaps, and in drug infusion pumps. Pacemakers are known to be affected by low frequency EMI signals. There are many exogenous sources of EMI including, but not limited to, Electrostatic Discharge, arc welding, ambulance sirens, and other sources (104). The growth of EMI is only increasing and human factors professionals need to carefully consider sources of problems from EMI that may have to be dealt with.

A fourth area of potential problems of humans factors interaction with medical devices occurs when two or more medical devices are simultaneously in operation, but their concurrent action may interact with each other in a destructive way.

The Potential Interaction of Multiple Medical Devices

As medical devices become more and more sophisticated, they may concurrently be in operation on the same patient (105–109). The action of one medical device may produce an undesired response of another device, especially if it may be implanted. From Geddes (35): "Between 1979 and 1995, the Center for Devices and Radiological Health (CDRH) of the U.S. Food and Drug Administration (FDA) has received over one hundred reports alleging the electromagnetic interference (EMI) resulted in malfunction of electronic medical devices." The source of the EMI was from one medical device treating the patient. The malfunction occurred in a second medical device, which was also, simultaneously, being used to treat the same patient.

"For example, about 100,000 cardiac pacemakers are implanted annually. These stimulators can be interrogated and programmed by an external device that uses a radio frequency link. Electro surgery, which also uses radio frequency electric current may interact with the pacemaker causing serious arrhythmias. Even though the two technologies are safe, when used alone, their simultaneous combination has resulted in injury." More specifically, nowadays with the prevalence use of cellular telephones for both caregivers as well as the patients in care situations, there are documented cases of resulting injury to the patient. Cell phones have now been recognized to cause instances of malfunction of drug infusion pumps and patient monitors. These interactions have to be considered for new human factors interactions with medical devices in the future, Silbergerg (110) with increased interest in the errors created (111,112) and the specific technology used (113).

With the changing style of litigation with respect to the medical profession, there has been more public awareness of sources of human factor error induced by the medical professional working in a state of extreme fatigue.

Increased Public Awareness to Fatigue Issues and the Medical Professional

There has now been a substantial increased awareness of the general public to the fact that their medical professional may have compromised performance due to the fact that they are suffering from long hours of work. Fatigue studies continue to receive increased concern (114–120). This certainly has its influence on human factors procedures when dealing with medical devices and the overall success of the medical interaction. For example (115), it is known that physicians had demonstrated levels of daytime sleepiness worse then that of patients with narcolepsy or sleep apnea when required to perform long hours of duty.

Finally, since the publication of Ref. 1, the healthcare industry must now deal with a substantially larger population of acquired immune deficiency syndrome (AIDS) survivors who need medical, dental, and other types of interactions with healthcare professionals (121).

Changing Medical Procedures to Deal with Active Human Immunodeficiency Virus Patients

With the advent of advanced drug therapies, people with human immunodeficiency virus (HIV) are now living longer and longer. These same people need dental care, have medical attention requests, and require other types of consideration. The medical professional must exercise

forethought to not have exposure to body fluids and new procedures are in place to provide discrimination free care to these people. In the early days of public exposure to people with HIV, there were documented cases of health professionals refusing to give adequate care. For example, for resuscitation, fireman and others would avoid contact with individuals suspected of having HIV. New devices have now been constructed to keep body fluids and other contact more separated between the patient and the caregiver. There have been new laws passed to prevent discrimination to people suspected of having HIV in housing, in the workplace, and also in receiving adequate healthcare.

CONCLUSION

The modern human factors interactions with medical devices have been strongly influenced by the advent of new technologies including the Internet, microprocessors, and computers. People are becoming more accustomed to automation and dealing with other sophisticated means of delivering healthcare. One lesson that can be learned from the improvement of human factors interactions with medical devices is that we can create safety by anticipating and planning for unexpected events and future surprises. Another change in this new millennium is that the responsibility of the patient is now shifted more to the individual or their family to have a greater role and duty over their own therapies and venue, and perhaps work in their home setting. Telemedicine and wireless means of dealing with controls over the medical devices are certainly on the increase and will influence how the patient has to deal with their healthcare professional in the future.

BIBLIOGRAPHY

Cited References

1. Hyman WA. Human Factors in Medical Devices. In Webster JG editor Encyclopedia of Medical Devices and Instrumentation. 1st ed. New York: Wiley; 1988.
2. Fries RC. Reliable Design of Medical Devices. New York: Marcel Dekker; 1997.
3. U. S. Food and Drug Administration, Guidance for Industry and FDA Premarket and Design Control Reviewers–Medical Device Use-Safety: Incorporating Human Factors Engineering into Risk Management. Available at http://www.fda.gov/cdrh/humfac/1497.html.
4. Sawyer D. An Introduction to Human Factors in Medical Devices. U. S. Department of Health and Human Services, Public Health Service, FDA; December, 1996.
5. Murff HJ, Gosbee JW, Bates DW. Chapter 41. Human Factors and Medical Devices. Available at http://www.ncbi.nlm.-nih.gov.books/.
6. Human Factors and Medical Devices. Available at http://www.ahrq.gov/clinc/ptsafety/chap41b.htm.
7. Association for the Advancement of Medical Instrumentation. Human Factors Engineering Guidelines and Preferred Practices for the Design of Medical Devices. ANSI/AAMI HE48-1993, Arlington (VA); 1993.
8. Backinger CL, Kingsley P. Write it Right: Recommendations for Developing User Instructions for Medical Devices in Home Health Care. Rockville (MD): Department of Health and Human Services; 1993.
9. Burlington DB. Human factors and the FDA's goals: improved medical device design. Biomed Instrum Technol Mar.–Apr., 1996;30(2):107–109.
10. Carpenter PF. Responsibility, Risk, and Informed Consent. In: Ekelmen KB, editors. New Medical Devices: Invention, Development, and Use. Series on Technology and Social Priorities. Washington (DC): National Academy Press; 1988. p 138–145.
11. Policy statements adopted by the governing council of the American Public Health Association. Am J Public Health. Nov. 12, 1997;88(3):495–528.
12. Reese DW. The Problem of Unsafe Medical Devices for Industry, Government, Medicine and the Public. Dissertation Abs International: Sect B: Sci Eng 1994;54(11-B):5593.
13. Claris Law, Inc. Use of Human factors in Reducing Device-related Medical Errors, available at http://www.injuryboard.com/.
14. Green M. An Attorney's Guide to Perception and Human Factors. Available at http://www.expertlaw.com/library/attyarticles/perception.html.
15. Green M. Error and Injury in Computers and Medical Devices. Available at http://www.expertlaw.com/library/attyaricles/computer_negligence.html.
16. Sokol A, Jurevic M, Molzen CJ. The changing standard of care in medicine-e-health, medical errors, and technology add new obstacles. J Legal Med 2002;23(4):449–491.
17. Bruckart JE, Licina JR, Quattlebaum M. Laboratory and flight tests of medical equipment for use in U.S. army medevac helicopters. Air Med J 1993;1(3):51–56.
18. Budman S, Portnoy D, Villapiano AJ. How to Get Technological Innovation Used in Behavioral Health Care: Build It and They Still Might Not Come. Psychother Theory, Res Practice, Training 40 (1–2), Educational Publishing Foundation; 2003. p 45–54.
19. Burley D, Inman WH, editors. Therapeutic Risk: Perception, Measurement, Management. Chichester: Wiley; 1988.
20. Hasler RA. Human factors design-what is it and how can it affect you?. J Intravenous Nursing May–Jun, 1996;19(3)(Suppl.): S5–8.
21. McConnell EA. How and what staff nurses learn about the medical devices they use in direct patient care. Res Nursing Health 1995;18(2):165–172.
22. Obradovich JH, Woods DD. Users as designers: how people cope with poor HCI design in computer-based medical devices. Human Factors 1996;38(4):574–592.
23. Phillips CA. Human Factors Engineering. New York: Wiley; 2000.
24. Senders JW. Medical Devices, Medical Errors, and Medical Accidents. Human Error in Medicine. Hillsdale (NJ): Lawrence Erlbaum Associates, Inc.; 1994. p 159–177.
25. Ward JR, Clarkson PJ. An analysis of medical device-related errors: prevalence and possible solutions. J Med Eng Technol Jan–Feb, 2004;28(1):2–21.
26. Goldmann D, Kaushal R. Time to tackle the tough issues in patient safety. Pediatrics Oct. 2002;110(4):823–827.
27. Gosbee J. Who Left the Defibrillator On? Joint Comm J Quality Safety May, 2004;30(5):282–285.
28. Kaptchuk TJ, Goldman P, Stone DA, Stason WB. Do medical devices have enhanced placebo effects? J Clin Epidemiol 2000;53:786–792.
29. Lambert MJ, Bergin AE. The Effectiveness of Psychotherapy. In: Bergin AE, Garfield SL, editors. Handbook of Psychotherapy and Behavior Chance. 4th ed. New York: Wiley; 1994. p 143–189.
30. Perry SJ. An Overlooked Alliance: Using human factors engineering to reduce patient harm. J Quality Safety 2004; 30(8):455–459.

31. Leape L. Error in medicine. J Am Med Assoc 1994;21(3):272.
32. Leape LL. The preventability of medical injury. In: Bogner MS, editor. Human Error in Medicine Hillsdale (NJ): Lawrence Erbaum Associates; 1994. p 13–25.
33. Ganiats T. Error. J Am Med Asso 1995;273:1156.
34. Bogner MS. Medical human factors. Proc Human Factors Ergonomics Soc, 40th Annu Meet; 1996. p 752–756.
35. Geddes LA. Medical Device Accidents. 2nd ed. Lawyers and Judges, Publishers 2002.
36. Wagner D. How to use medical devices safely. AORN J Dec. 2002;76(6):1059–1061.
37. Fairbanks RJ, Caplan S. Poor interface design and lack of usability testing facilitate medical error. Human Factors Engineering Series. J Quality Safety Oct. 2004;30(10):579–584.
38. Phillips CA. Functional Electrical Rehabilitation, Springer-Verlag, 1991.
39. Militello LG. Learning to think like a user: using cognitive task analysis to meet today's health care design challenges. Biomed Instrum Technol 1998;32(5):535–540.
40. Militello L. A Cognitive Task Analysis of Nicu Nurses' Patient Assessment Skills. Proc Human Factors Ergonomics Soc, 39th Annu Meet; 1995. p 733–737.
41. Lin L, et al. Analysis, Redesign, and Evaluation of a Patient-Controlled Analgesia Machine Interface. Proc Human Factors Ergonomics Soc, 39th Annu Meet; 1995. p 738–741.
42. Bogner MS. Medical devices and human error. In: Mouloua M, Parasuraman R, editors. Human Performance in Automated Systems: Current Research and Trends. Hillsdale (NJ): Erlbaum; 1994. p 64–67.
43. Klein GA, Calderwood R, MacGregor D. Critical decision method for eliciting knowledge. IEEE Trans. Systems, Man, Cybernetics 1989;19(3):462–472.
44. Aucella AF, et al. Improving Ultrasound Systems by User-Centered Design. Proc Human Factors Ergonomics Society, 38th Annu Meet; 1994. p 705–709.
45. Law CM, Vanderheiden GC. Tests for Screening Product Design Prior to User Testing by People with Functional Limitations. Proc Human Factors Ergonomics Soc, 43rd. Annu Meet; 1999. p 868–872.
46. Neilsen J. Usability Engineering. Boston: Academic; 1993.
47. Whiteside BJ, Holtzblatt K. Usability engineering: our experience and evolution. In: Helander M, editor. The Handbook of Human Computer Interaction. New York: Elsevier Press; 1988.
48. Nielsen J. Heuristic evaluation. In: Nielson J, Mack R, editors. Usability Inspection Methods. New York: Wiley; 1994. p 54–88.
49. Welch DL. Human factors in the health care facility. Biomedical Instrum Technol. May–Jun, 1998;32(3):311–316.
50. Lathan BE, Bogner MS, Hamilton D, Blanarovich A. Human-centered design of home care technologies. NeuroRehabilitation 1999;12(1):3–10.
51. Rasmussen J, Pejtersen AM, Goodstein LP. Cognitive Systems Engineering. New York: Wiley; 1994.
52. Rasmussen J. Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering. New York: North-Holland; 1986.
53. Hajdukiewicz JR, et al. A Work Domain Analysis of Patient Monitoring in the Operating Room. Proc Human Factors Ergonomics Soc, 42th Annu Meet; 1998. p 1038–1042.
54. van Gruting CWD. Medical Devices–International Perspectives on Health and Safety. New York: Elsevier; 1994.
55. Simons D, Fredericks TK, Tappel J. The Evaluation of an Auditory Alarm for a New Medical Device. Proc Human Factors Ergonomics Soc, 41st Annu Meet; 1997. p 777–781.
56. Seagull FJ, Sanderson PM. Anesthesia Alarms in Surgical Context. Proc Human Factors Ergonomics Soc, 42nd. Annu Meet; 1998. p 1048–1051.
57. Aguirre R, McCreadie S, Grosbee J. Human Factors and Training Evaluation of Two Automated External Defibrillators. Proc Human Factors Ergonomics Soc, 43rd Annu Meet; 1999. p 840–844.
58. Woods DD. The alarm problem and directed attention in dynamic fault management. Ergonomics 1995;38(11):2371–2394.
59. Laughery KR, Wogalter MS. Warnings and risk perception. design for health and safety. In: Salvendy G, editor. Handbook of Human Factors and Ergonomics. 2nd ed. New York: Wiley; 1997. p 1174–1197.
60. Callan JR, Gwynee JW. Human Factors Principles for Medical Device Labeling. Available at http://www.fda.gov/cdrh/dsma/227.html.
61. Isaacson JJ, Klein HA, Muldoon RV. Prescription Medication Information: Improving Usability Through Human Factors Design. Proc Human Factors Ergonomics Soc, 43rd Annu Meet; 1999. p 873–877.
62. Collet JP, Bovin JF, Spitzer WO. Bias and confounding in pharmacoepidemiology. In: Strom BL, editor. Pharmacoepidemiology. New York: Wiley; 1994. p 741.
63. Senn S. Statistical Issues in Drug Development. New York: Wiley; 1997.
64. Twomey E. The usefulness and use of second-generation antipsychotic medications: review of evidence and recommendations by a task force of the World Psychiatric Association. Curr Opinion Psychiat 2002;15(Suppl 1):S1–S51.
65. O'Brien TG, Charlton SG. Handbook of Human Factors Testing and Evaluation. Lawrence Erlbaum Associates; 1996.
66. Sheridan TB. Humans and Automation: System Design and Research Issues. New York: Wiley; 2002.
67. Obradovich JH, Woods DD. Users as Designers: How People Cope with Poor HCI Design in Computer-based Medical Devices. Human Factors 1996;38(4):574–592.
68. Howard SK. Failure of an automated non-invasive blood pressure device: the contribution of human error and software design flaw. J Clin Monitoring 1993; 9.
69. Sarter NB, Woods DD, Billings CE. Automation surprises. In: Salvendy G, editor. Handbook of Human Factors/Ergonomics. 2nd ed. New York: Wiley; 1997. p 1926–1943.
70. Andre J. Home health care and high-tech medical equipment, caring. Nat Assoc Home Care Mag Sept. 1996; 9–12.
71. Dalby RN, Hickey AJ, Tiano SL. Medical devices for the delivery of therapeutic aerosols to the lungs. In: Hickey AJ, editor. Inhalation Aerosols: Physical and Biological Basis for Therapy. New York: Marcel Dekker; 1996.
72. Draper S, Nielsen GA, Noland M. Using 'No problem found' in infusion pump programming as a springboard for learning about human factors engineering. Joint Comm J Quality Safety Sept. 2004;30(9):515–520.
73. Weinger MB, Scanlon TS, Miller L. A widely unappreciated cause of failure of an automatic noninvasive blood pressure monitor. J Clin Monitoring Oct. 1992;8(4):291–294.
74. Walsh T, Beatty PCW. Human factors error and patient monitoring. Physiol Meas 2002;23:R111–132.
75. Agree EM, Freedman VA. Incorporating assistive devices into community- based long-term care: an analysis of the potential for substitution and supplementation. J Aging Health 2000;12:426–450.
76. Rogers WA, Fisk AD. Human Factors Interventions for the Health Care of Older Adults. Mahwah (NJ): Lawrence Erlbaum Associates, Publishers; 2001.
77. Vanderheiden GC. Design for people with functional limitations resulting from disability, aging, or circumstance. In: Salvendy G, editor. Handbook of Human Factors and Ergonomics. 2nd ed. New York: Wiley; 1997. p 2010–2052.

78. Billing J. The Incident Reporting and Analysis Loop. In Enhancing Patient Safety and Reducing Medical Errors in Health Care. Chicago: National Patient Safety Foundation; 1999.
79. Fisk D, Rogers WA. Psychology and aging: enhancing the lives of an aging population. Curr Directions Psychol Sci Jun. 2002;11(3):107–111.
80. Gardner-Bonneau D. Designing medical devices for older adults. In: Rogers WA, Fisk AD, editors. Human Factors Interventions for the Health Care of Older Adults Mahwah (NJ): Erlbaum; 2001. p 221–237.
81. Park C, Morrell RW, Shifren K. Processing of medical information in aging patients: cognitive and human factors perspectives. Lawrence Erlbaum Associates; 1999.
82. Sutton M, Gignac AM, Cott C. Medical and everyday assistive device use among older adults with arthritis. Can J Aging 2002;21(4):535–548.
83. Glavin RJ, Maran NJ. Practical and curriculum applications integrating human factors into the medical curriculum. Med Educ 2003;37(11):59–65.
84. Birkmier-Peters DP, Whitaker LA, Peters LJ. Usability Testing for Telemedicine Systems: A Methodology and Case Study. Proc Human Factors Ergonomics Soc, 41st Annu Meet; 1997. p 792–796.
85. Repperger DW. Active force reflection devices in teleoperation. IEEE Control Systems Jan. 1991;11(1) 52–56.
86. Repperger DW, et al. Effects of haptic feedback and turbulence on landing performance using an immersive cave automatic virtual environment (CAVE). Perceptual Motor Skills 2003;97:820–832.
87. Repperger DW. Adaptive displays and controllers using alternative feedback. CyberPschol Behavior 2004;7(6):645–652.
88. Repperger DW, Phillips CA. A haptics study involving physically challenged individuals. Encyclopedia of Biomedical Engineering. New York: Wiley; 2005.
89. Wegner CM, Karron DB, Surgical navigation using audio feedback. Studies in Health Technology and Informatics. 1997;39:450–458.
90. Cooper J. An analysis of major errors and equipment failures in anesthesia management: considerations for prevention and detection. Anesthesiology 1984;60:34–42.
91. Gaba DM, Howard SK, Jump B. Production pressure in the work environment: California anesthesiologists' attitudes and experiences. Anesthesiology 81: 1994; 488–500.
92. Howard SK, et al. Anesthesia crisis resource management training: teaching anesthesiologists to handle critical incidents. Aviation Space Environ Med 63:763–770.
93. Weinger MB. Anesthesia incidents and accidents. Misadventures in health care: Inside Stories. Mahwah, (NJ): Lawrence Erlbaum Associates, Publishers; 2004. p 89–103.
94. Obradovic JH, Woods DD. Users as Designers: How People Cope with Poor HCI Design in Computer-Based Medical Devices. Proc Human Factors Ergonomics Soc, 38th Ann Meet; 1994. p 710–714.
95. Kober, Mavor A, editors. Safe, Comfortable, Attractive, and Easy to Use: Improving the Usability of Home Medical Devices. Report of National Research Council to U.S. Congress 1996, Washington (DC): National Academy Press; 1996. p 5–8.
96. Baldwin GM. Experiences of Siblings of In-Home Technology-Dependent Children. Dissertation Abst Int: Sec B: Sci Eng 1997;58(5-B):2714.
97. Mykityshyn AL, Fisk AD, Rogers WA. Learning to use a home medical device: mediating age-related differences with training. Human Factors 2002;44(3):354–364.
98. Schoenfeld I. Risk Assessment and Approaches to Addressing Human Error in Medical Uses of Radioisotopes. Panel: Proc Human Factors Ergonomics Soc, 37th Ann Meet; 1993. p 859–862.
99. Lin L. Human Error n Patient-Controlled Analgesia: Incident Reports and Experimental Evaluation. Proc Human Factors Ergonomics Soc, 42th Ann Meet; 1998. p 1043–1047.
100. Lin L, et al. Applying human factors to the design of medical equipment: patient controlled analgesia. J Clin Monitoring 1998;14:253–263.
101. McLellan H, Lindsay D. The relative importance of factors affecting the choice of bathing devices. Pain B J Occup Therapy 2003;66(9):396–401.
102. Repperger DW, Ho CC, Phillips CA. Clinical short-wire TENS (transcutaneous electric nerve stimulator) study for mitigation of pain in the Dayton va medical center. J Clin Eng Sept./Oct. 1997; 290–297.
103. Silbert PL, Roth PA, Kanz BS. Interference from cellular telephones in the electroencephalogram. J Polysomnographic Technol Dec. 1994;10:20–22.
104. Radiofrequency interference with medical devices. A technical information statement. IEEE Eng Med Bio Mag 1998;17(3):111–114.
105. Roy G. Child-proofing of hearing aids to prevent hazards posed by battery swallowing. J Speech-Language Pathol Audiol 1992;16(3):243–246.
106. Beery TA, Sommers M, Sawyer M, Hall J. Focused life stories of women with cardiac pacemakers. Western J Nursing Res 2002;24(1):7–23.
107. Romano PS. Using administrative data to identify associations between implanted medical devices and chronic diseases. Ann Epidemiol 2000;10:197–199.
108. Wiederhold K, Wiederhold MD, Jang DP, Kim SI. Use of cellular telephone therapy for fear of driving. CyberPsychol Behavior 2000;3(6):1031–1039.
109. Zinn HK. A Retrospective Study of Anecdotal Reports of the Adverse Side Effects of Electroconvulsive Therapy. Dissertation Abs Int Sec B: Sci Eng 2000;60(9-B):4871.
110. Silbergerg J. What Can/Should We Learn From Reports of Medical Device Electromagnetic Interference? At Electromagnetic, Health Care and Health, EMBS 95, Sept. 19–20, 1995, Montréal, Canada: Standards Promulgating Organizations; 1995.
111. Leape LL, et al. Promoting patient safety by reducing medical errors. JAMA Oct. 1998;280:28 1444–1447.
112. Rasmussen J. The concept of human error: Is it useful for the design of safe systems in health care? In: Vincent C, DeMoll B, editors. Risk and Safety in Medicine. London: Elsevier; 1999.
113. Woods DD, Cook RI, Billings CE. The impact of technology on physician cognition and performance. J Clin Monitoring 1995;11:92–95.
114. Gaba DM, Howard SK. Fatigue among clinicians and the safety of patients. N Engl J Med 2002;347:1249–1255.
115. Howard SK, Gaba DM, Roseking MR, Zarcone VP. Excessive daytime sleepiness in resident physicians: risks, intervention, and implication. Acad Med 2002;77:1019–1025.
116. Cook RI, Render ML, Woods DD. Gaps in the continuity of care and progress on patient safety. Br Med J March 18, 2000;320:791–794.
117. Fennell PA. A Fourx-phase approach to understanding chronic fatigue syndrome. In: Jason LA, Fennell PA, Taylor RR., editors. The Chronic Fatigue Syndrome Handbook 2003. Hoboken (NJ): Wiley; 2003. p 155–175.
118. Fennell PA. Phase-based interventions. In: Jason LA, Fennell PA, Taylor RR, editors. The Chronic Fatigue Syndrome Handbook. Hoboken (NJ): Wiley; 2003. p 455–492.
119. Jason LA, Taylor RR. Community-based interventions. In: Jason LA, Fennell PA, Taylor RR, editors. The Chronic Fatigue Syndrome Handbook. Hoboken (NJ): Wiley; 2003. p 726–754.

120. Rogers SH. Work physiology—fatigue and recovery. The human factors fundamentals. In: Salvendy G, editor. Handbook of Human Factors and Ergonomics. 2nd ed. New York: Wiley; 1997. p 268–297.
121. Roy F, Robillard P. Effectiveness of and compliance to preventive measures against the occupational transmission of human immunodeficiency virus. Scand J Work Environ Health 1994;20(6):393–400.

See also CODES AND REGULATIONS: MEDICAL DEVICES; EQUIPMENT MAINTENANCE, BIOMEDICAL; HOME HEALTH CARE DEVICES; MONITORING IN ANESTHESIA; SAFETY PROGRAM, HOSPITAL.

HUMAN SPINE, BIOMECHANICS OF

VIJAY K. GOEL
ASHOK BIYANI
University of Toledo, and
Medical College of Ohio,
Toledo Ohio

LISA FERRARA
Cleveland Clinic Foundation
Cleveland, Ohio

SETTI S. RENGACHARY
Detroit, Michigan

DENNIS MCGOWAN
Kearney Notabene

INTRODUCTION

From a bioengineer's perspective, bio the spine involves an understanding of the interaction among spinal components to provide the desired function in a normal person. Thereafter, one needs to analyze the role of these elements in producing instability. Abnormal motion may be due to external environmental factors to which the spine is subjected to during activities of daily living (e.g., impact, repetitive loading, lifting) degeneration, infectious diseases, injury or trauma, disorders, and/or surgery. Furthermore, the field of spinal biomechanics encompasses a relationship between conservative treatments, surgical procedures, and spinal stabilization techniques. Obviously, the field of spinal biomechanics is very broad and it will not be practical to cover all aspects in one article. Consequently, this article describes several of these aspects, especially in the cervical and thoraco-lumbar regions of the human spine. A brief description of the spine anatomy follows since it is a prerequisite for the study of bio the human spine.

SPINE ANATOMY

The human spinal column consists of 33 vertebras interconnected by fibrocartilaginous intervertebral disks (except the upper most cervical region), articular facet capsules, ligaments, and muscles. Normally, there are 7 cervical vertebras, 12 thoracic vertebras, 5 lumbar vertebras, and 5 fused sacral vertebras, Fig. 1a (1). When viewed in the frontal plane, the spine generally appears straight and symmetric while revealing four curves in the sagittal plane. The curves are anteriorly convex or lordotic in the cervical and lumbar regions, and posteriorly convex or kyphotic in the thoracic and sacrococcygeal regions. The center of gravity of the spinal column generally passes from the dens of the axis (C2) through the vertebra to the promontory of the sacrum (2,3). The ligamentous spine anatomy can be best described through a functional spinal unit (FSU, Fig. 1b), comprising the two adjacent vertebras, the disk in between, and the other soft tissues structures. This segment can be divided into anterior and posterior columns. The anterior column consists of the posterior longitudinal ligament, intervertebral disk, vertebral body, and anterior longitudinal ligament. Additional stability is provided by the muscles that surround the ligamentous spine, Fig. 1c. The motion of this segment can be described as rotation about three axes and translation along the same axes, Fig. 2. In the following paragraphs, the anatomy of the cervical region is described in some detail followed by a descriptive section discussing the anatomy of the lumbar spine.

Cervical Spine Anatomy

The cervical spine usually is subdivided in two regions (upper and lower), based on the functional aspects and anatomical differences between the two regions. The lumbar region anatomy, in principle, is similar to the lower cervical region.

Upper Cervical Spine (C0-C1-C2)

The upper cervical spine has been commented to be the most complex combination of articulations in the human skeleton. This region is also commonly called the "craniocervical junction" or the "craniovertebral junction" (CVJ). It is composed of three bony structures: the occipital bone (C0), the atlas (C1), and the axis (C2, Fig. 3). The atlas (C1), serves to support the skull. The atlas is atypical of other cervical vertebras in that it possesses neither a vertebral body nor a spinous process. The lateral masses of the atlas have both superior and inferior articular facets. The superior facets are elongated, kidney-shaped, and concave, and serve to receive the occipital condyles. The inferior facets are flatter and more circular and permit axial rotation. Transverse processes extend laterally from each lateral mass. Within each transverse process is a foramen that is bisected by the vertebral artery. The second cervical vertebra, or axis (C2), is also atypical of other cervical vertebra due to its osseous geometry (5,6). The most noteworthy geometric anomaly is the odontoid process, or dens. The odontoid process articulates with the anterior arch of the atlas. Posterior and lateral to the odontoid process are the large, convex superior facets that articulate with the inferior facets of C1. The inferior facets of C2 articulate with the superior facets of C3. The axis contains a large bifid spinous process that is the attachment site delineating the craniovertebral and subaxial musculature and ligament anatomies.

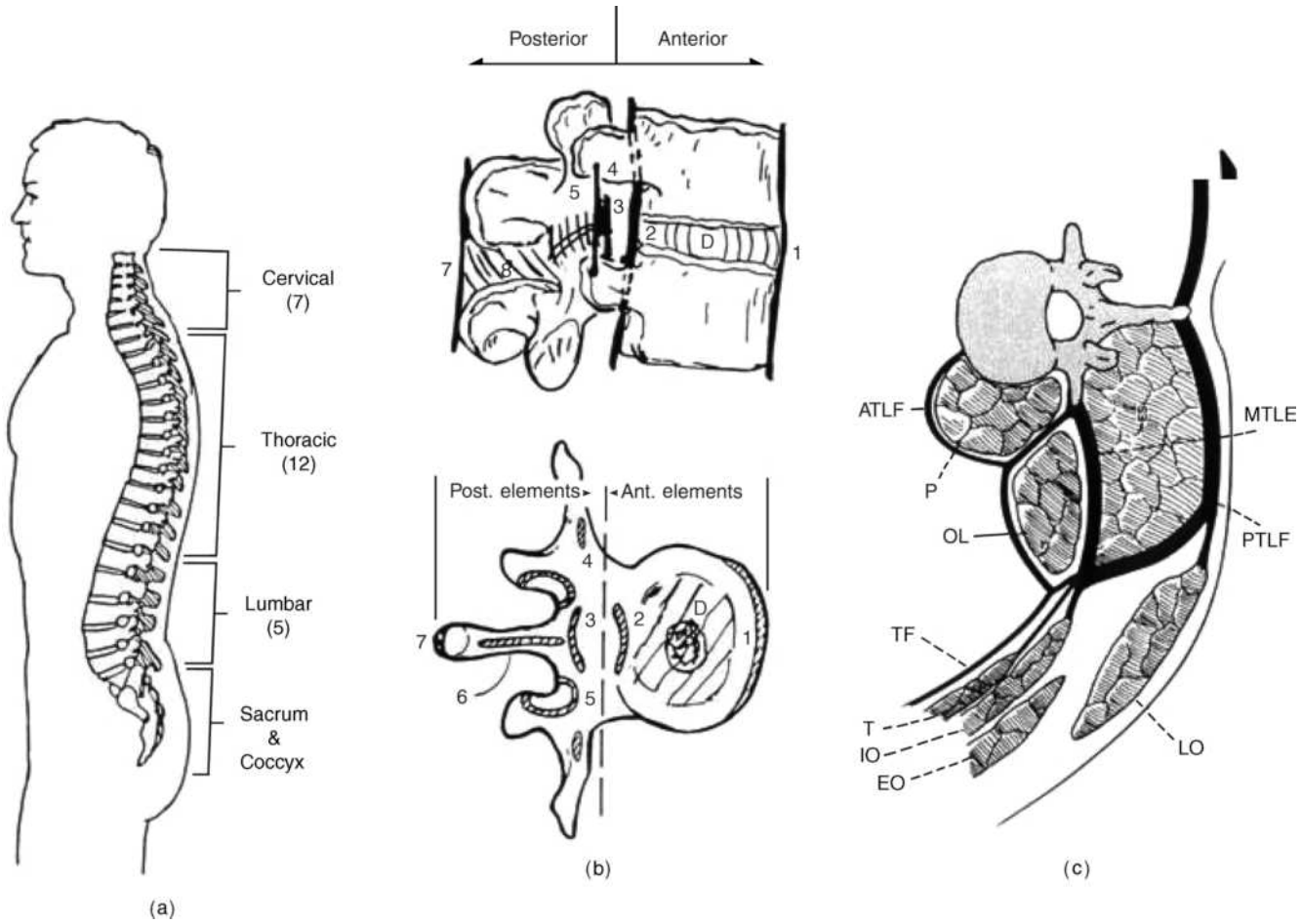


Figure 1. The ligamentous human spine. (a) The side view showing the three curvatures. (b) The functional spinal unit (FSU) depicts the spinal elements that contribute to its stability. (c) Additional stability is provided by the muscles that surround the spine. (Taken from Ref. 1.)

The trabecular anatomy of weight bearing bones provides information about the normal loading patterns of the bones, fracture mechanisms, and fixation capabilities. According to Heggeness and Doherty (6) the medial, anterior cortex of the odontoid process (1.77 mm at the anterior promontory) was found to be much thicker than the anterolateral (1.00 mm), lateral (1.08 mm), and posterior (0.84 mm) aspects of the axis. These authors feel that this is suggestive of bending and torsional load carrying capabilities. The same was found for the vertebral body, with thinner cortices were noted in the anterolateral and posterior directions. The trabecular bone in the tip of the odontoid process was found to be dense, maximizing in the anterior aspect of the medullary canal. One observation made by the authors was an area of cortical bone density at the center near the tip, which would seem to indicate that this area experiences elevated external forces, perhaps due to local ligamentous attachments. The lateral masses immediately inferior to the facets demonstrated dense regions of trabecular bone, with individual trabeculas spanning from this region to the inferior end plate, suggestive of a major axial load path.

The ligamentous structures of the upper cervical spine form a complex architecture (Fig. 3) that serves to join the

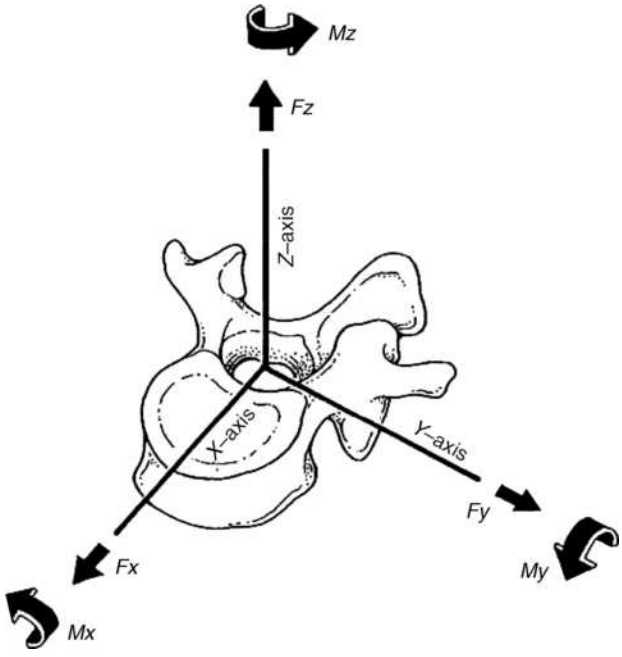


Figure 2. The spinal motion consists of six components (three translations and three rotations). (Adapted from Ref. 2.)

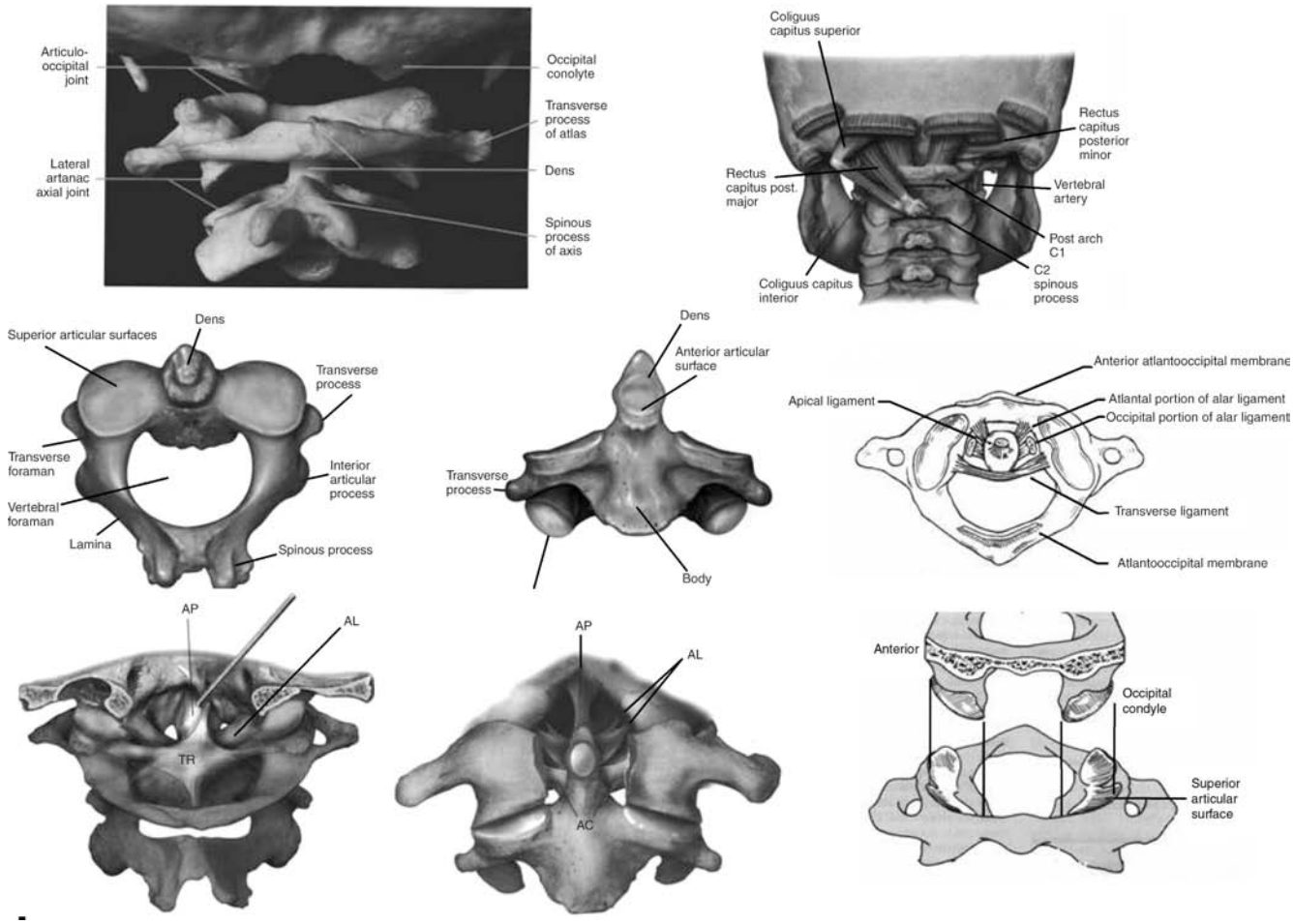


Figure 3. The anatomy of the upper region of the cervical spine C0 (occiput)-C1 (Atanlanto)-C2 (Axial). (Taken from Ref. 4c.)

vertebras, allow limited motion within and between levels, and provide stability. The cruciform ligament as a whole consists of two ligaments: the atlantal transverse ligament and the inferior–superior fascicles. The transverse ligament attaches between the medial tubercles of the lateral masses of the atlas, passing posterior to the odontoid process. Attachment of the cervical spine to the skull is also achieved by the paired alar ligaments. These ligaments run bilaterally from the occiptal condyles inferiolaterally to the tip of the odontoid process. The alar ligaments also contain fibers that run bilaterally from the odontoid process anterolaterally to the atlas. These ligaments have been identified as a check against overaxial rotation of the craniovertebral junction. Extending from the body of the axis to the inner surface of the occiput, the tectorial membrane is the most posterior ligament and actually represents the cephalad extension of the subaxial posterior longitudinal ligament. The tectorial membrane has been implicated as a check against extreme flexion motion. The apical dental ligament extends from the anterior portion of the magnum foramen to the tip of the odontoid process. The accessory atlantoaxial ligaments are bilateral structures that run between the base of the odontoid process and the lateral masses of the atlas. The most anterior of the major ligaments is the anterior longitudinal ligament. This ligament extends inferiorly from the anterior margin of the foramen magnum to the superior surface of the anterior arch of the atlas at the anterior tuberosity. The ligament continues inferiorly to the anterior aspect of the axial body. The nuchal ligament (ligamentum nuchae) extends from the occiput to the posterior tubercle of the axis, continuing inferiorly to the spinous process of the subaxial vertebras (7).

There are six synovial articulations in the occipitoatlantoaxial complex: the paired atlanto-occiptal joints, the paired atlantoaxial joints, the joint between the odontoid process and the anterior arch of the atlas, and the joint formed by the transverse ligament and the posterior aspect of the odontoid process, Fig. 3. The bilateral atlantooccipital joints are formed from the articulation of the occiptal condyles with the superior facets of the atlas. These joints are relatively stable due to the high degree of congruence between the opposing surfaces and the marked rounding that is displayed by both sides. They allow flexion and extension, limited lateral bending, and almost no rotation. The lack of allowed rotation is thought to be due to the ellipsoid form of the joint itself. Bilateral articulation of the inferior facets of the atlas with the superior facets of the axis form the atlantoaxial joints. Relatively small contact areas and opposed convexity

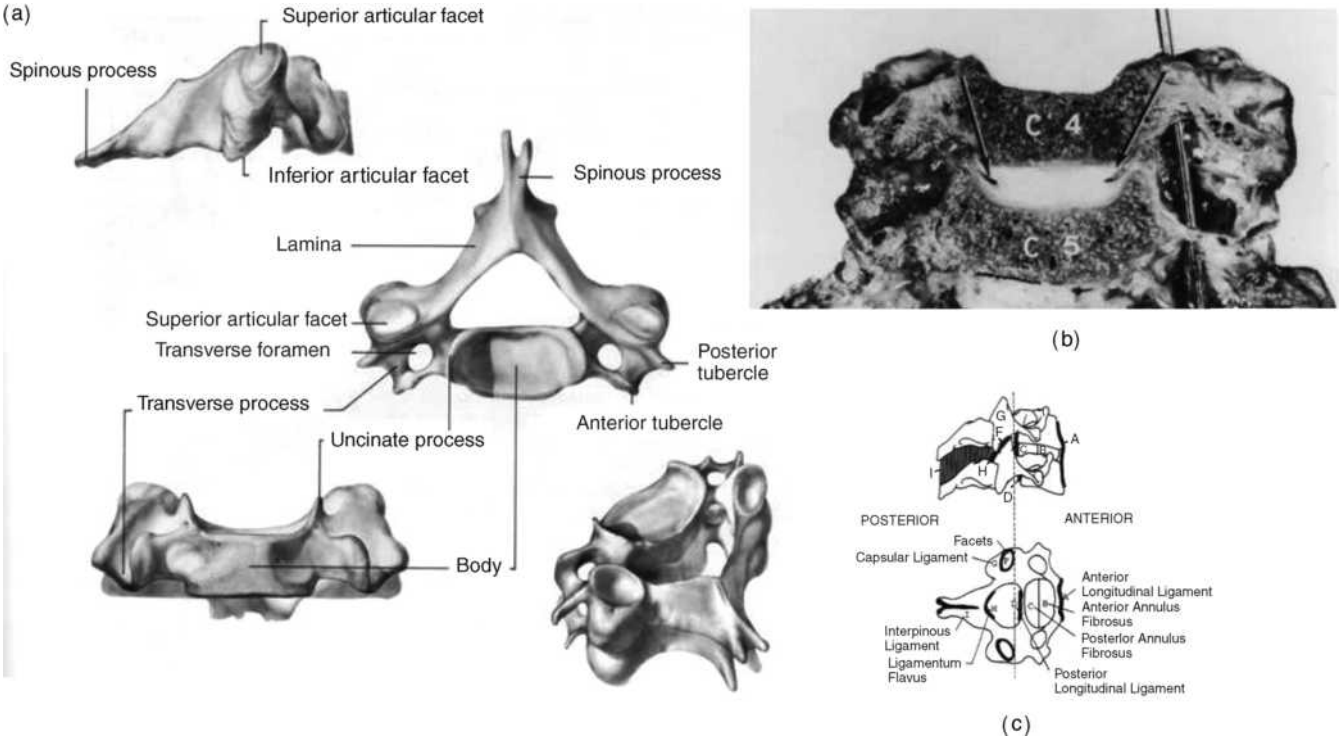


Figure 4. Anatomy of the lower cervical spine region. (Taken from Ref. 4d.)

result in a rather unstable joint. Movement is permitted in almost all six degrees of freedom: left and right axial rotation, flexion–extension, right and left lateral bending. Anteroposterior translation stability of this articulation is highly dependent on the transverse ligament. The odontoid process articulates anteriorly with the posterior aspect of the anterior atlantal ring. The joint is actually a bursal joint, with absence of specific capsular ligaments. The posterior aspect of the odontoid process and the transverse ligament form a joint via a bursa junction, creating the most unique articulation in the craniovertebral junction. This is necessitated by the large degree of axial rotation afforded at the atlantoaxial level.

Lower Cervical Spine (C3-C7). The lower cervical spinal vertebral column consists of osseous vertebras separated by fibrocartilaginous intervertebral disks anteriorly, facet joint structures posteriorly, and a multitude of ligamentous structures that provide stability and serve as motion control. Motion between adjacent vertebras is relatively limited due to these constraints, although overall motion of the lower cervical region is quite extensive. The lower cervical spine consists of five vertebras (C3-C7).

Cervical Vertebrae (Fig. 4a): The vertebral body is roughly in the shape of an elliptical cylinder and has a concave superior surface (due to the uncinate processes) and a convex inferior surface. A thin cortical shell (~ 0.3 mm thick anteriorly and 0.2 mm thick posteriorly) surrounds the cancellous bone of the inner vertebral body, while the superior and inferior surfaces of the vertebral body form the cartilaginous endplates, to which the intervertebral disks are attached. The superior aspect of each vertebra contains the uncinate process or uncus, a dorsolateral bilateral bony projection, which gives the body a concave shape superiorly in the coronal plane and allows for the vertebral body to fit around the convex inferior surface of the immediately superior vertebra. The height of these processes vary from level to level, but the highest uncinate processes are located at C5 and C6 (as high as 9 mm from the flat surface of the endplate) and the smallest are located at C3 and C7 (8–10). Vertebral bodies transmit the majority of load.

The transverse process of the vertebra contains the intervertebral foramen. The intervertebral foramen is elliptical or round in shape, and hides and protects the neurological and vascular structures of the cervical spine, specifically the vertebral artery. Also, the rostral side of each bilateral transverse process is grooved to allow space for the exiting spinal nerve root.

The bilateral diarthroidal facet (or zygapophyseal) joints are located posteriorly to the pedicles both superiorly and inferiorly. The average orientation for the C3-C7 facet joints is $\sim 45^\circ$ from the transverse plane, with steeper inclinations in the lower segments (11). This inclination allows far less axial rotation than occurs in the upper cervical spine. Together with the vertebral body (and intervertebral disks), the facets fulfill the primary role of load bearing in the spine. Typically a "three-column" aspect is applied to the cervical spine, consisting of bilateral facets and the anterior column (vertebral body plus intervertebral disk).

The pedicles, lamina, and spinous process of the cervical spine are made of relatively dense bone and, together with the posterior aspect of the vertebral body, form the spinal

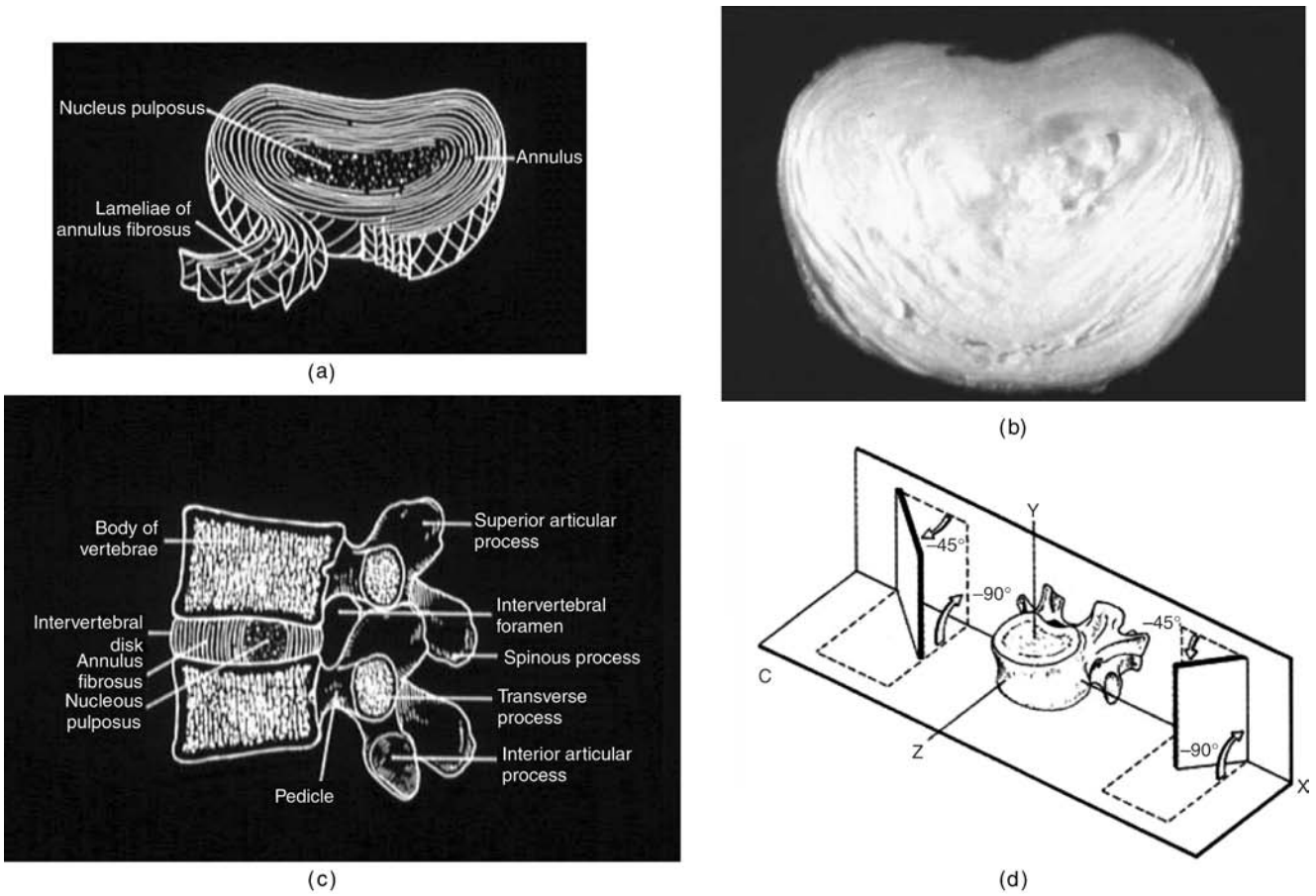


Figure 5. The anatomy of the lumbar spine. (a) and (b) show schematics of the disk and an actual disk (c) A FSU in the lumbar region. (d) The facet orientation in the lumbar region is more sagittal as compared to the other regions of the spine. (Adapted from Ref. 2.)

canal, within which lies the spinal cord. There are many ligament attachment points in this region and ligaments allow for resistance of flexion motion in the cervical spine.

The typical sagittal cervical spine alignment is thought to be a lordotic contour (11–15). The total average cervical lordosis was found to be $40 \pm 9.7^\circ$ for C0-C7, with the majority of this lordosis occurring at the C1-C2 level ($31.9 \pm 7.0^\circ$), and only 15% of the total lordosis occurring at the C4-C7 levels combined. The normal population seem to exhibit lordosis that ranges between 15 and 40°.

The Intervertebral Disk

Figure 5 forms the main articulation between adjacent vertebral bodies in the spine. It has the ability to transmit and distribute loads that pass between adjacent vertebral bodies. Its structure is a composite formation of outer layers of lamellas sheets called the annulus fibrosis, which surrounds the inner region of hydrophylic proteoglycan gel embedded in a collagen matrix called the nucleus pulposus. The material properties of the intervertebral disk appear to change markedly as a result of the aging process The matrix in which collagen and elastin fibers are embedded is composed of proteoglycan aggregates formed from proteoglycan subunits, hyaluronic acid, and link protein (16). In soft tissues, such as the intervertebral disk and cartilage, the proteoglycan aggregates are immobilized within a fibrous network and play a major biological role in the structure of collagen, in turn playing a major mechanical role in the intervertebral disk integrity. The viscoelastic properties of the intervertebral disk can be attributed to the interaction between the collagen fibrils and proteoglycan matrix composing the nucleus pulposus of the intervertebral disk. The proteoglycans function to attract fluids into the matrix, while the collagen fibers provide the tensile strength to the disk. As the human spine ages, the osmotic properties of the intervertebral disk decline, and the disks become dehydrated with age, causing a reduction in overall disk height.

The annulus fibrosis of the disk consists of a series of approximately twelve 1-mm thick lamellas sheets, each composed of collagen fibers. The anterior lamellas are generally thicker and more distinct than the posterior lamellas. According to a study by Pooni et al.(9), the collagen fibers running through a single laminar sheet are oriented at $\sim 65^\circ$ ($\pm 2.5^\circ$) with respect to the vertical axis. These fibers alternate direction in concentric lamellas to form a cross-pattern. The annulus fibrosus develops lesions as it ages.

The nucleus pulposus of the intervertebral disk consists of a hydrophylic proteoglycan gel embedded in a collagen matrix. The nucleus pulposus contains $\sim$80–88% water content in a young adult spine and occupies $\sim$30–50% of the total intervertebral disk volume (16,17). However, with aging the nucleus undergoes rapid fibrosis and loses its fluid properties such that, by the third decade of life, there is hardly any nuclear material distinguishable (18). In a normal healthy intervertebral disk, the nucleus pulposus is glossy in appearance.

Luschka's joints are something special in the cervical region. The human cervical intervertebral disk contains fissures, called Luschka's joints or uncovertebral joints that run along the uncinate process and radiate inward toward the nucleus (Fig. 4a and b). These fissures run through the annular lamellas and the adjacent annular fibers are oriented such that they run parallel to the fissure (19–21). These fissures appear within the latter part of the first decade of life and continue to grow in size as aging occurs (8). Although some argument exists as to the definition of the fissures as true joints or pseudojoints, the fissures have been shown to exist as a natural part of the aging process (19,20) and therefore are important aspects of biomechanical modeling of the human cervical intervertebral disks.

The ligaments of the cervical spine (Fig. 4c) provide stability and act to limit excessive motions of the vertebras, thereby preventing injury during physiologic movement of the spine. Ligaments can only transmit tensile forces, impeding excessive motion, but do follow the principles of Wolff's law, where the tissue will remodel and realign along lines of tensile stress. The ligaments that are biomechanically relevant include the anterior longitudinal ligament (ALL), posterior longitudinal ligament (PLL), ligamentum flavum (LF), interspinous ligament (ISL), and the capsular ligaments (CAP). The ALL and PLL each traverse the length of the spine. The ALL originates at an insertion point on the inferior occipital surface and ends at the first segment of the sacrum. It runs along the anterior vertebral bodies, attached to the osseous bodies and loosely attached to the intervertebral disks as well. The ALL is under tension when the cervical spine undergoes extension. The PLL also runs the length of the spine down the posterior aspect of the vertebral bodies, originating at the occiput and terminating at the coccyx. Similar to the ALL, it is firmly attached to the osseous vertebral bodies and to the intervertebral disks. The PLL is under tension when the spine undergoes flexion. The ligamentum flavum couples the laminas of adjacent vertebras. It is an extremely elastic ligament due to the higher percentage of elastin fibers (65–70%) as compared to other ligaments in the spine and any other structure in the human body. The LF resists flexion motion and lengthens during flexion and shortens during extension. The high elastin content minimizes the likelihood of buckling during extension. It is under slight tension when the spine is at rest and acts as a tension band in flexion. Torsion also places the ligamentum flavum under tension, and restraint of rotation may also be a significant function. The ISL insertion points lie between adjacent spinous processes. The ligament is typically slack when the head is in a neutral posture and only becomes tensile when enough flexion motion has occurred such that other ligaments have undergone significant tension, such as the capsular ligaments, PLL and LF. Additionally, the ISL insertion points are such that it is ideal for resisting the larger flexion rotations that can occur as a result of excessive flexion loading. The capsular ligaments (CAPs) enclose the cervical facet joints and serve to stabilize the articulations of these joints and limit excessive motions at these joints. Generally, the fibers are oriented such that they lie perpendicular to the plane of the facet joints. These ligaments potentially also serve to keep the facets aligned and allow for the coupled rotations.

Lumbar Spine Anatomy

The basic structural components of the lumbar spine are the same as that of the lower cervical spine with differences in size, shape, and orientation of the structures due to functional requirements being different from that of the cervical region, Fig. 5. For example, the lumbar vertebras are bigger in size, because of the higher axial loads they carry. With regard to the peripheral margin of the interverebral disk, annulus fibrosus is composed of 15–20 layers of collagenous fibrils obliquely running from one cartilage end plate to the other and crossing at 120° angles. As one progresses from the cervical into the thoracic region, the facet joints gradually orient themselves parallel with the frontal plane. The transition from the thoracic region into the lumbar region is indicated by a progressive change from the joints in the frontal plane to a more sagittal plane (4,22). This transition in facet orientation from the thoracic to the lumbar spine creates a different series of degenerative complications and disorders in the spine. Sagittal alignment of the facet joints increases the risk of subaxial and spondylolisthesis of the lumbar spine.

CLINICAL BIOMECHANICS OF THE NORMAL SPINE

The three basic functions of the spine are to protect the vital spinal cord, to transmit loads, and to provide the flexibility to accomplish activities of daily living. Components that provide stability to the spine are divided into four groups as follows:

1. Passive stabilizers: Passive stabilization is provided by the shape and size of vertebras and by the size, shape, and orientation of the facet joints that link them.
2. Dynamic stabilizers: Dynamic stabilization is provided by viscoelastic structures, such as the ligaments, capsules, and annulus fibrosus. The cartilage of the facet joints also acts as a damper.
3. Active stabilizers: Active voluntary or reflex stabilization is provided by the muscular system that governs the spine, Fig. 1c.
4. Hydrodynamic stabilizer: Hydrodynamic stabilization is due to the viscous nucleus pulposus.

The combination of these elements generates the characteristics of the entire spine. The diskussion of the kinematics will begin by further analyzing spinal elements as

either passive or active. It will then progress into the effect these stabilizers have on the different portions of the spine.

Passive Elements

The vertebral body acts to passively resist compressive force. The size, mineral content, and orientation of the cancellous bone of each vertebral body increase–change as one descends in the caudal direction, which is a morphologic response to the increasing weight it must bear (4). The cortical shell on the vertebral body serves as the chief load path. The shell also provides a rigid link in the FSU, and a platform for attachment of the intervertebral disk, muscles, and the anterior and posterior longitudinal ligaments. The transition area of the motion segment is the endplate. This serves to anchor the intervertebral disk to the vertebral body. Note that the endplate starts out as growth cartilage and transitions into bone as aging occurs (22). The disk acts as both a shock absorber and an intervertebral joint because the relative flexibility of the intervertebral disk is high when compared to the vertebral body. The intervertebral disk resists compression, tension, shear, bending, and torsion (4). It is relatively resistant to failure in axial compression while its annular portions fail in axial torsion first (23).

Dynamic Stabilizers

Although bone is viscoelastic in nature, it serves more as a structural component within the spine that passively resists axial forces and can transmit forces along the spinal column. The soft tissue spinal structures (ligamentous, capsules, annulus fibrosis) are far more elastic as compared to bone behavior and stabilize the spine in a dynamic manner, where rapid vamping of oscillatory motions occur. The main function of the facet joints is to pattern the motions of the spine so that during activities of daily living the neural elements are not strained beyond the physiological limits. Therefore, they play a major role in determining the range of motion across a joint and as a damper to any possible dynamic loading. The amount of stability provided by the facet joints depends on extent of the capsular ligaments, their shape, orientation, and level within the spine (2). For example, the thoracic facets have a limited capsular reinforcement and facilitate axial rotation, which is in contrast to the lumbar region where the facet ligaments are more substantial and the joint plane is configured to impede axial motion (24).

From a biomechanical perspective, the ligaments respond to tensile forces only (1). The effectiveness of a ligament depends on the morphology and the moment arm through which it acts. That is, not only the strength, but also the longer lever arm a ligament has, the more it participates in the stabilization of the spine (4). Ligaments also obey Wolff's law. The ligaments also undergo remodeling along the lines of applied tensile stresses in response to chronic loads, just like bones. The ligamentum flavum acts as a protective barrier for the entire spine.

Active Stabilizers

Muscles contribute significantly to maintain the stability of the spinal column under physiological conditions. Decreasing the muscle forces acting on a FSU, increases the motion and loading of the ligaments. A thoracolumbar (T1-sacrum) spinal column that is devoid of musculature is an unstable structure, with a load-carrying capacity of < 25 N (24). However, with properly coordinated muscle action, the spine can sustain large loads, which is exemplified by the action of weight lifters (24).

The internal force resisted by the muscle depends on factors such as cross-section and length at the initiation of contraction. The maximum force develops at approximately 125% of muscle resting length. In contrast, at approximately one-half of its resting length, the muscle develops very low force. The muscle stress (the maximum force per unit area) ranges from 30 to 90 $N \cdot cm^{-2}$ (25,26). Macintosh et al. (27) performed a modeling study based on radiographs from normal subjects to determine the effects of flexion on the forces exerted by the lumbar muscles. They found that the compressive forces and moments exerted by the back muscles in full flexion are not significantly different from those in the upright posture.

The remainder of this section is devoted to the biomechanics of the individual sections of the spinal column in a normal healthy person. Various methods for recording data with varying degrees of accuracy and repeatability are used ranging from the use of different types of goniometers, radiographs, *in vitro* cadaver based studies, magnetic resonance imaging (MRI) to visual estimation of motion. Although the value of assessing the ROM is not yet documented, the understanding and knowledge of normal age- and sex-related values of ROM is the basis for analysis of altered and possibly pathologic motion patterns as well as decreased or increased ROM (23,28). The issue of spinal instability (stability), although controversial in its definition, has immense clinical significance in the diagnosis and treatment of spinal disorders. Maintaining a normal range of motion in the spine is linked to spinal stability. The spine needs to maintain its normal range of motion to remain stable and distribute forces while bearing loads in several directions. The typical motion, for example, in response to the flexion–extension loads, as determined using cadaver testing protocols, is shown in Fig. 6. The two motion

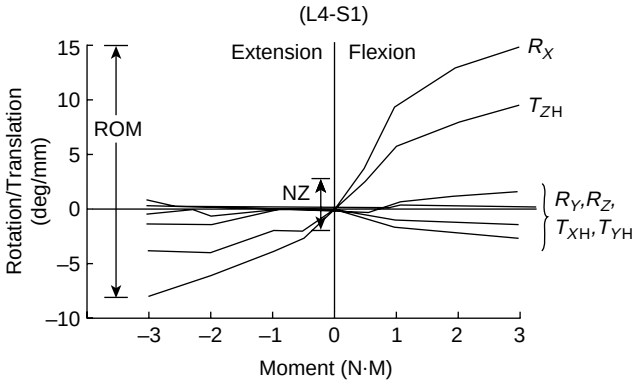


Figure 6. The load-displacement response of a FSU in flexion and extension. Two of the motion components are major and the other four are minor (or coupled). The range-of-motion and neutral zones, two of the terms used to describe the motion behavior of a segment, are also shown. (Adapted from Ref. 1.)

components (Flexion/extension rotation–R_x, and A-P translation-T_{zH}) are an order of magnitude higher than the other four. The two larger components are called the major–main motions and other four are the secondary–coupled motions. Range of motion, highlighted in the figure, will depend on the maximum load exerted on the specimen during testing. Likewise, the *in vivo* ranges of motion data will vary depending on the level of external force applied, that is, active or passive (2).

Biomechanics of the Occipital–Atlantoaxial Complex (C0-C1-C2)

As a unit, the craniovertebral junction accounts for 60% of the axial rotation and 40% of the flexion–extension behavior of the cervical spine (29,30).

Flexion–Extension. Large sagittal plane rotations have been attributed to the craniovertebral junction (Tables 1 and 2). Panjabi et al. (31) reported combined flexion–extension of C0-C1 and C1-C2 of 24.5 and 22.4°, respectively, confirming flexion–extension equivalence at the two levels. They also found that the occipitoatlantal joint level demonstrated a sixfold increase in extension as compared to flexion (21.0 vs. 3.5°), whereas the atlantoaxial level equally distributed its sagittal plane rotation between the two rotations, 11.5 (flexion) versus 10.9° (extension). Goel et al. (32) documented coupling rotations that occur with flexion and extension. They reported one-side lateral bending values of 1.2 and 1.4° for flexion and extension, respectively, at the C0-C1 level. In addition, they found C1-C2 coupled lateral bending, associated with flexion and extension movents, were lower than seen at the C0-C1 level. The largest axial rotation reported was 1.9°, which was an outcome of a C0-C1 extension of 16.5°. Note that the values reported by this study do not represent range of motion data, but rather intermediate rotation due to submaximal loading. Displacement coupling occurs between the translation of the head and flexion–extension of the occipitoatlanto–axial complex. Translation of the occiput with respect to the axis produces flexion–extension movements in the atlas. Anterior translation of the head extends the occipitoatlantal joints, with posterior motion resulting in converse flexion of the joint. This is postulated to occur due to the highly contoured articular surfaces of the atlanto-occiptal joint.

Table 1. Ranges of Motion Reported from *In Vivo* and *In Vitro* Studies for the Occipito-Atlantal Joint (C0-C1)[a]

Type of Study[b]	Total Flexion/ Extension	Unilateral Bending	Unilateral Axial Rotation
In vivo	50	34–40	0
In vivo	50	14–40	0
In vivo	13	8	0
In vivo	30	10	0
In vivo			5.2
In vivo			1.0
In vitro			4.0
In vitro	3.5/21.0	5.5	7.2

[a]In degrees.
[b]The *in vivo* studies represent passive range-of-motion, whereas the *in vitro* studies represent motion at 1.5 N·m occipital moment loading. (From Ref. 4c.)

Table 2. Ranges of Motion Reported from *In Vivo* and *In Vitro* Studies for the AtlantoAxial Joint (C1-C2)[a]

Type of Study[b]	Total Flexion/ Extension	Unilateral Bending	Unilateral Axial Rotation
In vivo	0	0	60
In vivo	11		30–80
In vivo	10	0	47
In vivo	30	10	70
In vivo			32.2
In vitro			43.1
In vivo			40.5
In vitro	11.5/10.9	6.7	38.9

[a]In degrees.
[b]The *in vivo* studies represent passive range-of-motion, whereas the *in vitro* studies represent motion at 1.5 N·m occipital moment loading. (From Ref. 4c.)

Lateral Bending. As is shown in Tables 1 and 2, early studies have shown that occipitoatlantal lateral bending dominates the overall contribution of this motion in the occipitoatlanto–axial complex. However, this is not the finding of the most recent study. Almost all other studies indicate a significantly greater contribution from the C0-C1 joint. Lateral flexion also plays an important role in rotation of the head. Rotation of the lower cervical spine (C2-T1) results in lateral flexion of this region.

Axial Rotation. Almost all of the occipitoatlanto–axial contribution to axial rotation occurs in the atlantoaxial region. Atlantoaxial rotation occurs about an axis that passes vertically through the center of the odontoid process. This axis remains halfway between the lateral masses of the atlas in both neutral and maximal rotation. In maximal rotation, there is minimal joint surface contact, and sudden overrotation of the head can lead to interlocking of the C1-C2 facets, making it impossible to rotate the head back to neutral. Table 2 lists the amount of rotation found in the atlantoaxial joint by various researchers. Although these studies have produced widely varying results, there seems to be a consensus among the more recent studies that one side axial rotation at the atlantoaxial level falls somewhere in the range of 35–45°. The findings in Table 1 demonstrate that there is a relatively small contribution from the C0-C1 joint, with researchers finding between 0 and 7.2° of rotation. One interesting anatomical note concerning axial rotation is the behavior of the vertebral artery during rotation. The vertebral artery possess a loop between the atlas and axis, thus affording it over-length. Upon atlantoaxial rotation, the slack is taken up in the loop and it straightens, thus preventing overstretching and possible rupture during maximal rotation.

The instantaneous axes of rotation (IARs) for the C0-C1 articulation pass through the center of the mastoid processes for flexion–extension and through a point 2–3 cm above the apex of the dens for lateral bending. There is a slight axial rotation at C0-C1. The IARs for the C1-C2 articulation are somewhere in the region of the middle third of the dens for flexion–extension and in the center of the dens for axial rotation. Lateral bending of C1-C2 is

Table 3. C3-C4 Ranges of Motion Compiled from Various *In Vivo* and *In Vitro* Studies[a,b]

Type of Study	Type of Loading	Total Flexion/Extension[c]	Unilateral Lateral Bending[c]	Unilateral Axial Rotation[c]
In vivo	Max. Rotation (active)	15.2 (3.8)	NA	NA
In vivo	Max. Rotation (active)	17.6 (1.5)	NA	NA
In vivo	Review	13.0 (range 7–38)	11.0 (range 9–16)	11.0 (range 10–28)
In vivo	Max. Rotation (active)	13.5 (3.4)	NA	NA
In vivo	Max. Rotation (active)	15.0 (3.0)	NA	NA
In vivo	Max. Rotation (active)	NA	NA	6.5 (range 3–10)
In vivo	Max. Rotation (active)	18.0 (range 13–26)	NA	NA
In vitro	1 N·m	8.5 (2.6)	NA	NA
In vitro	~3 N·m	NA	8.5 (1.8)	10.7 (1.3)

[a]In degrees.
[b]See Refs. 4b and d.
[c]Not available = NA.

controversial at the most 5–10° (4). During lateral bending, the alar ligament is responsible for the forced rotation of the second vertebra.

Middle and Lower Cervical Spine (C2-C7)

In the middle and lower cervical regions, stability and mobility must be provided; while, the vital spinal cord and the vertebral arteries must be protected. There is a good deal of flexion–extension and lateral bending in this area, Tables 3–6.

Flexion–Extension. Most of the flexion–extension motion in the lower cervical spine occurs in the central region, with the largest range of motion (ROM) generally occurring at the C5-C6 level. Except for extension, the orientation of the cervical facets (on average, ~45° in the sagittal plane) does not excessively limit spinal movements in any direction or rotation. Flexion–extension rotations are distributed throughout the entire lower cervical spine for total rotations typically in the range of 60–75° and sagittal A/P translation is usually in the range of ~2–3 mm at all cervical levels (1). There is relatively little coupling effect that occurs during flexion–extension due to the orientation of the facets. There have been many published *in vivo* and *in vitro* studies reporting "normal" rotations at the various cervical spinal levels. These studies are in general agreement, although there appears to be a wide variation within ROM at all levels of the cervical region.

An *in vitro* study by Moroney et al.(33) averaged rotations among 35 adult cervical motion segments and found that average rotations (± SD) in flexion and extension

Table 4. C4-C5 Ranges of Motion Compiled from Various *In Vivo* and *In Vitro* Studies[a,b]

Type of Study	Type of Loading	Total Flexion/Extension[c]	Unilateral Lateral Bending[c]	Unilateral Axial Rotation[c]
In vivo	Max. Rotation (active)	17.1 (4.5)	NA	NA
In vivo	Max. Rotation (active)	20.1 (1.6)	NA	NA
In vivo	Review	12 (range 8–39)	11.0 (range 0–16)	12.0 (range 10–26)
In vivo	Max. Rotation (active)	17.9 (3.1)	NA	NA
In vivo	Max. Rotation (active)	19 (3.0)	NA	NA
In vivo	Max. Rotation (active)	NA	NA	6.8 (range 1–12)
In vivo	Max. Rotation (active)	20 (range 16–29)	NA	NA
In vitro	1 N·m	9.7 (2.35)	NA	NA
In vitro	~3 N·m	NA	6.3 (0.6)	10.8 (0.7)

[a]In degrees.
[b]See Refs. 4b and d.
[c]Not available = NA.

Table 5. C5-C6 Ranges of Motion Compiled from Various *In Vivo* and *In Vitro* Studies[a,b]

Type of Study	Type of Loading	Total Flexion/Extension[c]	Unilateral Lateral Bending[c]	Unilateral Axial Rotation[c]
In vivo	Max. Rotation (active)	17.1 (3.9)	NA	NA
In vivo	Max. Rotation (active)	21.8 (1.6)	NA	NA
In vivo	Review	17.0 (range 4–34)	8.0 (range 8–16)	10.0 (range 10–34)
In vivo	Max. Rotation (active)	15.6 (4.9)	NA	NA
In vivo	Max. Rotation (active)	20.0 (3.0)	NA	NA
In vivo	Max. Rotation (active)	NA	NA	6.9 (range 2–12)
In vivo	Max. Rotation (active)	20.0 (range 16–29)	NA	NA
In vitro	1 N·m	10.8 (2.9)	NA	NA
In vitro	~3 N·m	NA	7.2 (0.5)	10.1 (0.9)

[a]In degrees.
[b]See Refs. 4b and d.
[c]Not available = NA.

Table 6. C6-C7 Ranges of Motion Compiled from Various *In Vivo* and *In Vitro* studies[a,b]

Type of Study	Type of Loading	Total Flexion/Extension[c]	Unilateral Lateral Bending[c]	Unilateral Axial Rotation[c]
In vivo	Max. Rotation (active)	18.1 (6.1)	NA	NA
In vivo	Max. Rotation (active)	20.7 (1.6)	NA	NA
In vivo	Review	16.0 (range 1–29)	7.0 (range 0–17)	9.0 (range 6–15)
In vivo	Max. Rotation (active)	12.5 (4.8)	NA	NA
In vivo	Max. Rotation (active)	19 (3)	NA	NA
In vivo	Max. Rotation (active)	NA	NA	5.4 (range 2–10)
In vivo	Max. Rotation (active)	15 (range 6–25)	NA	NA
In vitro	1 N·m	8.9 (2.4)	NA	NA
In vitro	~3 N·m	NA	6.4 (1.0)	8.8 (0.7)

[a]In degrees.
[b]See Refs. 4b and d.
[c]Not available = NA.

under an applied 1.8-N·m moment with 73.6-N preload (applied axially through the center of the vertebral bodies) were 5.55° (1.84) and 3.52° (1.94), respectively. These results demonstrate a total ROM in flexion–extension of ~9.02°. Although generally lower than the reported data in Tables 3–6, probably due to the effect of averaging across cervical levels, the measurements are within the range of motion for all levels diskussed above.

Lateral Bending. Lateral bending rotations are distributed throughout the entire lower cervical spine for total rotations typically in the range of 10–12° for C2-C5 and 4–8° for C7-T1 (1). Unlike flexion–extension motion, where coupling effects are minimal, lateral bending is a more complicated motion involving the cervical spine, mainly due to the increased coupling effects. The coupling effects, probably due to the spatial locations of the facet joints at each level, are such that the spinous processes are rotated in the opposite direction of the lateral bending direction. The degree of coupling that occurs at separate levels of the cervical region has been described (33). There is a gradual decrease in the amount of axial rotation coupled with lateral bending as one traverses from C2 to C7. At C2, for every 3° of lateral bending there is ~2° of coupled axial rotation, a ratio of 0.67. At C7, for every 7.5° of lateral bending there is ~1° of coupled axial rotation, a ratio of 0.13.

Axial Rotation. Most cervical rotation occurs about the C1-C2 level, in the range of 35–45° for unilateral axial rotation: ~40% of the total rotation observed in the spine (1). In the lower cervical spine, axial rotation is in the range of 5.4–11.0° per level. Again, as in the main motion of lateral bending, there exists a coupling effect with lateral bending when axial rotation is the main motion of the cervical spine. This coupling effect is in the range of 0.51–0.75° of lateral bending per degree of axial rotation (34). The effects of aging and gender on cervical spine motion have been investigated by numerous researchers. The average values for age decades for each motion, as well as average for the gender groups along with significant differences are shown in Table 7. Significantly less motion in the active tests was evident in comparison of lateral bending and axial rotation. Generally, for passive tests, the SD was lower. Women showed greater ROM in all these motions. In the age range of 40–49 years, women again showed significantly greater ROM in axial rotation and rotation at maximal flexion. There were no significant differences between gender groups for the group aged 60+ years. The well-established clinical observation that motion of the cervical spine decreases with age has been confirmed. An exception to this finding was the surprising observation that the rotation of the upper cervical spine, mainly at the atlantoaxial joint (tested by rotating the head at maximum flexion of the cervical spine that presumably locks the other levels)

Table 7. Average (SD) Head–Shoulder Rotations[a,b]

Age Decade	Flex/Ext		Lat Bending		Axial Rotation		Rot From Flex		Rot From Ext	
	M	F	M	F	M	F	M	F	M	F
20–29	152.7[c]	149.3	101.1	100.0	183.8	182.4	75.5[c]	72.6	161.8	171.5
	(20.0)	(11.7)	(13.3)	(8.6)	(11.8)	(10.0)	(12.4)	(12.7)	(15.9)	(10.0)
30–39	(141.1)	155.9[c]	94.7[c]	106.3[c]	175.1[c]	186.0[c]	66.0	74.6	158.4	165.8
	(11.4)[d]	(23.1)	(10.0)[d]	(18.1)	(9.9)[d]	(10.4)	(13.6)[d]	(10.5)	(16.4)	(16.0)
40–49	131.1	139.8	83.7	88.2[c]	157.4	168.2[b]	71.5	85.2	146.2	153.9[c]
	(18.5)	(13.0)	(13.9)	(16.1)	(19.5)[d]	(13.6)	(10.9)[c]	(14.8)	(33.3)	(22.9)
50–59	136.3[c]	126.9	88.3	76.1	166.2[c]	151.9	77.7	85.6	145.8	132.4[c]
	(15.7)	(14.8)	(29.1)[d]	(10.2)	(14.1)	(15.9)	(17.1)	(9.9)	(21.2)[d]	(28.8)
60+	116.3	133.2	74.2	79.6	145.6	154.2	79.4	81.3	130.9	154.5
	(18.7)	(7.6)	(14.3)	(18.0)	(13.1)	(14.6)	(8.1)	(21.2)	(24.1)	(14.7)

[a]In degrees.
[b]See Ref. 4h.
[c]Significant difference from cell directly adjacent to the right (i.e., gender within age group differences).
[d]Significant difference from cell directly adjacent below (i.e., age group within gender differentiation).

did not decrease with age. The measurement data for rotation out of maximum flexion suggests that the rotation of the atlantoaxial joint does not decrease with age, but rather remains constant or increases slightly perhaps to compensate for the reduced motion of the lower segments.

Lumbar Spine

The lumbar spine is anatomically designed to limit anterior translation and permit considerable flexion-extension and lateral bending, Tables 8A, B, and C. The unique characteristic of the spine is that it must support tremendous axial loads. The lumbar spine and the hips contribute to the considerable mobility of the trunk (34,35). The facets play a crucial role in the stability of the lumbar spine. The well-developed capsules of these joints play a major part in stabilizing the FSU against axial rotation and lateral bending. Lumbar facet joints are oriented in the sagittal plane, thereby allowing flexion–extension and lateral bending but limiting torsion (4).

In flexion–extension, there is usually a cephalocaudal increase in the range of motion in the lumbar spine. The L5-S1 joint offers more sagittal plane motion than the other joints, due to the unique anatomy of the FSU. The orientation of the facet becomes more parallel to the frontal plane as the spinal column descends toward S1. Both this facet orientation and the lordotic angle at this motion segment contribute to the differences in the motion at this level. For lateral bending, each level is about the same except for L5-S1, which shows a relatively small amount of motion.The situation is the same for axial rotation, except that there is more motion at the L5-S1 joint.

There are several coupling patterns that have been observed in the lumbar spine. Pearcy (36) observed coupling of 2° of axial rotation and 3° of lateral bending with flexion–extension. In addition, there is also a coupling pattern, in which axial rotation is combined with lateral bending, such that the spinous processes point in the same direction as the lateral bending (22). This pattern is the opposite of that in the cervical spine and the upper thoracic spine (34).

The rotation axes for the sagittal plane of the lumbar spine have been described in several reports. In 1930, Calve and Galland (37) suggested that the center of the intervertebral disk is the site of the axes for flexion–extension; however, Rolander (38) showed that when flexion is simulated starting from a neutral position, the axes are located in the region of the anterior potion of the disk. In lateral bending, the axes fall in the region of the right side of the disk with left lateral bending, and in the region of the left side of the disk with right lateral bending. For axial

Table 8. Ranges of Motion for Various Segments Based on *In Vivo* and *In Vitro* Data Collection Techniques Cited in the Literature[a,b]

(A) Flexion/Extension

	In vitro			*In vivo*/active			*In vivo*/active		
	Mean	Lower	Upper	Mean	Lower	Upper	Mean	Lower	Upper
L1/2	10.7	5.0	13.0	7.0	1.0	14.0	13.0	3.0	23.0
L2/3	10.8	8.0	13.0	9.0	2.0	16.0	14.0	10.0	18.0
L3/4	11.2	6.0	15.0	10.0	2.0	18.0	13.0	9.0	17.0
L4/5	14.5	9.0	20.0	13.0	2.0	20.0	16.0	8.0	24.0
L5/S1	17.8	10.0	24.0	14.0	2.0	27.0	14.0	4.0	24.0

(B) Lateral Bending

	In vitro			*In vivo*/active			*In vivo*/passive		
	Mean	Lower	Upper	Mean	Lower	Upper	Mean	Lower	Upper
L1/2	4.9	3.8	6.5	5.5	4.0	10.0	7.9		14.2
L2/3	7.0	4.6	9.5	5.5	2.0	10.0	10.4		16.9
L3/4	5.7	4.5	8.1	5.0	3.0	8.0	12.4		21.2
L4/5	5.7	3.2	8.2	2.5	3.0	6.0	12.4		19.8

(C) Axial Rotation

	In vitro			*In vivo*/active		
	Mean	Lower	Upper	Mean	Lower	Upper
L1/2	2.1	0.9	4.5	1.0	−1.0	2.0
L2/3	2.6	1.2	4.6	1.0	−1.0	2.0
L3/4	2.6	0.9	4.0	1.5	0.0	4.0
L4/5	2.2	0.8	4.7	1.5	0.0	3.0
L5/S1	1.3	0.6	2.1	0.5	−2.0	2.0

[a]In degrees.
[b]In general *in vitro* data differs from *in vivo* data and the magnitude of *in vivo* motions depend on the collection technique (active vs. passive). (Taken from Ref. 4h.)

rotation, the IARs are located in the region of the posterior nucleus and annulus (4,36).

BIOMECHANICS OF SPINAL INSTABILITY: ROLE OF VARIOUS FACTORS

The causes of spinal instability have been hypothesized to include environmental factors that contribute to spinal degeneration and host of other variables (39). For example, some diseases can lead to spinal instability without being the direct cause. Chronic spondylolisthesis can lead to permanent deformation of the annulus that increases the probability of instability, Fig. 7. Essentially, any damage to any of the components of the motion segment or neural elements can contribute to instability. Instability can result from ruptured ligaments, fractured facets, fractured endplates, torn disks, or many other causes. However, the elements within the spine that seem to contribute more to stability and can therefore be major sources of instability are the facet joints, the intervertebral disks, and the ligaments (40). Both *in vivo* investigations in humans and animals and *in vitro* investigations of ligamentous spinal segments have been undertaken to accumulate biomechanical data of clinical significance.

Role of Environmental Factors in Producing Instability–Injury

Upper Cervical Spine. High speed impact loads that may be imposed on the spine are one of the major causes of spinal instability in the cervical region, especially in the upper region. To quantify the likely injuries of the atlas, Oda et al. (39,40) subjected upper cervical spine specimens to high speed axial impact by dropping 3–6 kg weights from various heights. The load produced axial compression and flexion of the specimen. Both bony and soft tissue injuries, similar to Jefferson fractures, were observed. The bony fractures were six bursting fractures, one four-part fracture without a prominent bursting, and one posterior arch fracture. The major soft tissue injury involved the transverse ligament. There were five bony avulsions and three midsubstance tears. The study was extended to determine the three-dimensional (3D) load displacements of fresh ligamentous upper cervical spines (C0-C3) in flexion, extension, and lateral bending before and following the impact loading in the axial mode. The largest increase in flexibility due to the injury was in flexion–extension: $\sim 42\%$. In lateral bending, the increase was on the order of 24%; in axial rotation it was minimal: $\sim 5\%$. These increases in motion are in concordance with the actual instabilities observed clinically. In patients with burst fractures of the atlas, Jefferson noted that the patients could not flex their heads, but could easily rotate without pain (41).

Heller et al. (42) tested the transverse ligament attached to C1 vertebra by holding the C1 vertebra and pushing the ligament in the middle along the AP direction. The specimens were loaded with an MTS testing device at varying loading rates. Eleven specimens failed within the substance of the ligament, and two failed by bone avulsion.

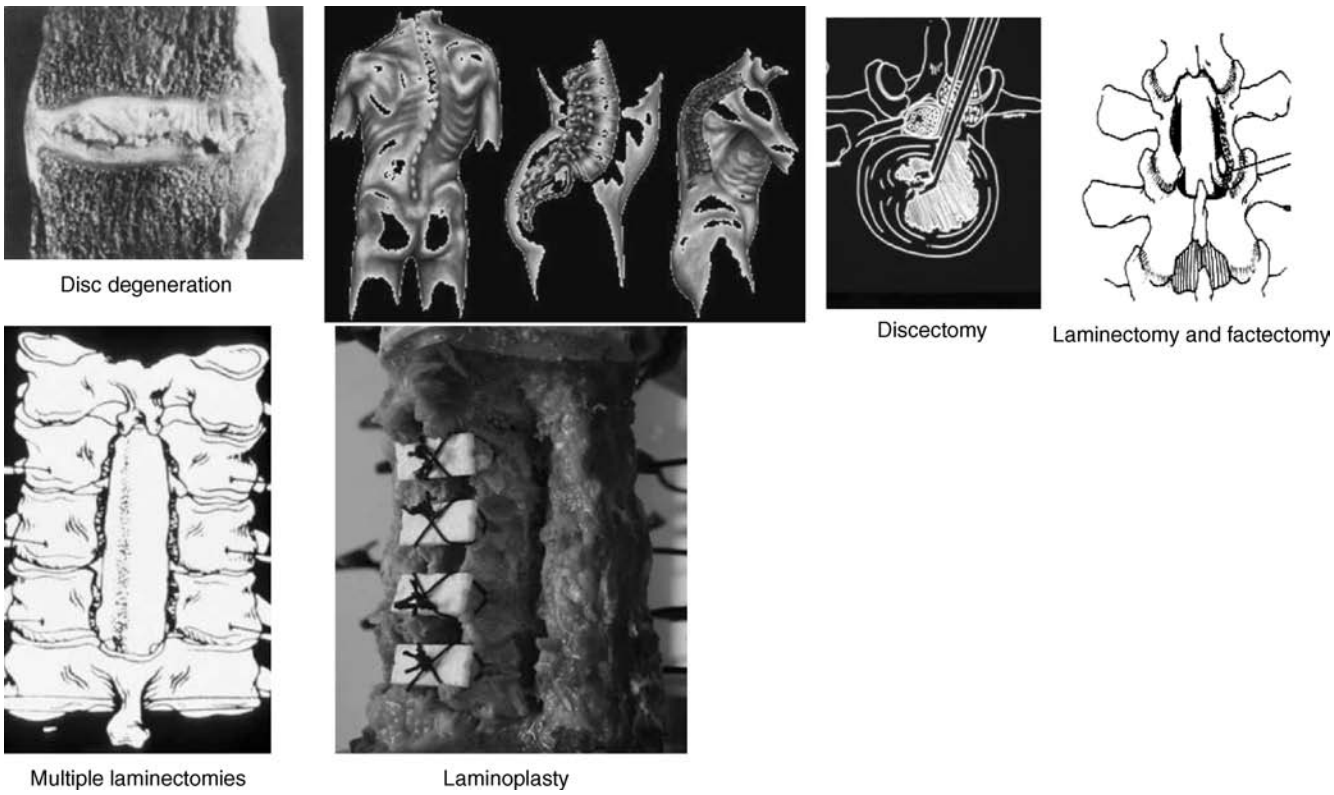


Figure 7. Various spinal disorders and surgical procedures that may lead to spinal instability. Such procedures are common for all of the spine regions.

The mean load to failure was 692 N (range 220–1590 N). The displacement to failure ranged from 2 to 14 mm (mean 6.7 mm). This study, when compared with the work of Oda et al. (39,40) suggests that (a) anteroposterior (AP) translation of the transverse ligament with respect to the dens is essential to produce its fracture; (b) rate of loading affects the type of fracture (bony versus ligamentous) but not the displacement at failure; and (c) even "axial" impact loads are capable of producing enough AP translation to produce a midsubstance tear of the ligament, as reported by Oda et al. (39).

The contribution to stabilization by the alar ligament of the upper cervical spine is of particular interest in evaluation of the effects of trauma, especially in the axial rotation mode. Goel and associates (43), in a study of occipitoatlantoaxial specimens, determined that the average values for axial rotation and torque at the point of maximum resistance were 68.1° and 13.6 N·m, respectively. They also observed that the value of axial rotation at which complete bilateral rotary dislocation occurred was approximately the point of maximal resistance. The types of injuries observed were related to the magnitude of axial rotation imposed on a specimen during testing. Soft tissue injuries (such as stretch–rupture of the capsular ligaments, subluxation of the C1-C2 facets) were confined to specimens rotated to or almost to the point of maximum resistance. Specimens that were rotated well beyond the point of maximum resistance also showed avulsion fractures of the bone at the points of attachment of the alar ligament or fractures of the odontoid process inferior to the level of alar ligament attachment. The alar ligament did not rupture in any of the specimens. Chang and associates (44) extended this study to determine the effects of rate of loading (dynamic loading) on the occipitoatlantoaxial complex. The specimens were divided into three groups and tested until failure at three different dynamic loading rates: 50°/s, 100°/s, and 400°/s as compared to the quasistatic (4°/s) rate of loading used by Goel et al.(43). The results showed that at the higher rates of loading, (a) the specimens became stiffer and the torque required to produce "failure" increased significantly (e.g., from 13.6 N·m at 4°/s to 27.9 N·m at 100°/s); (b) the corresponding right angular rotations (65–79°) did not change significantly; and (c) the rates of the alar ligament midsubstance rupture increased and that of "dens fracture" decreased. No fractures of the atlas were noted. This is another example of the rate of load application affecting the type of injury produced.

Fractures of the odontoid process of the second cervical vertebra comprise 7–13% of all cervical spine fractures (45). Most published reports involving odontoid fracture use the classification system detailed by Anderson and D'Alonzo (46). They described three types of odontoid process fracture (Fig. 8). Type I is an oblique fracture near the superior tip of the odontoid process and is thought to involve an avulsion defect associated with the alar–apical complex. Fracture of the odontoid process at the juncture of the process and vertebral body in the region of the accessory ligaments (Type II) is the most common osseous injury of the atlas. Fractures of this type lead to a highly unstable cervicovertebral region, commonly threatening the spinal

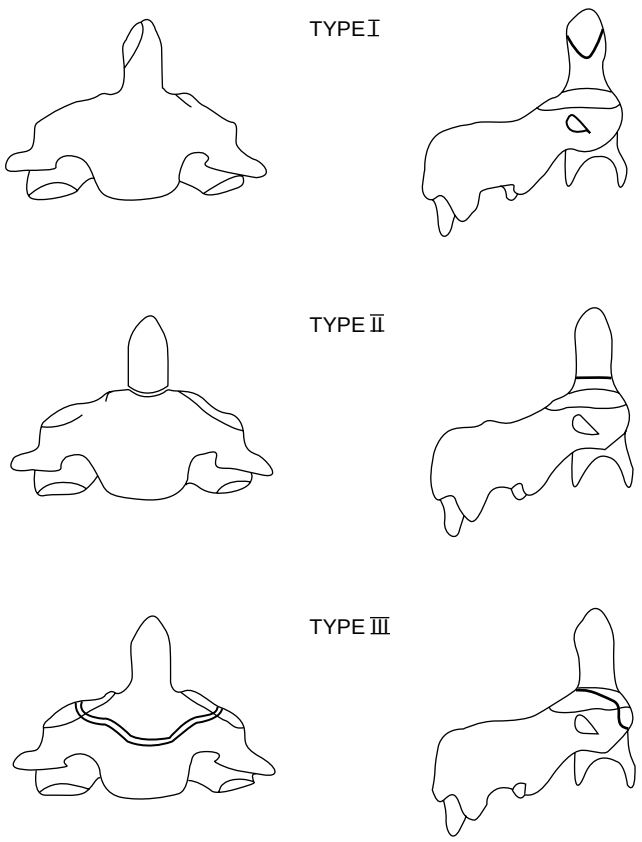


Figure 8. Fractures of the odontoid process. Taken from Ref. 46.

canal, and are often accompanied by ligamentous insult. Many of these fractures result in pseudoarthrosis if not properly treated. Type III fractures involve the junction of the odontoid process and the anterior portion of the vertebral body. These fractures are thought to be more stable than the Type I and Type II fractures. Type III fractures have high union rates owing to the cancellous bone involvement and the relatively high degree of vascularity (46,47).

Forces required to produce various types of dens fractures have been documented by Doherty et al. (45) who harvested the second cervical vertebra from fresh human spinal columns. Force was applied at the tip of the dens until failure occurred. The direction of the applied force was adjusted to exert extension bending or combined flexion and lateral bending on the tip of the dens. Extension resulted in type III fractures, and the combined load led to type II fractures of the dens. Furthermore, dynamic loading modes are essential to produce midsubstance ligament ruptures as opposed to dens fractures, especially in a normal specimen. Odontoid fractures have been implicated as being the result of high energy traumatic events. Indeed, there have been numerous accounts as to the events that lead to odontoid fracture. Schatzker et al. (47) reported that 16 of the 37 cases they reviewed were due to motor vehicle accidents and 15 cases were the result of high energy falls. Clark and White (48) report that all Type II (96 patients) and Type III (48 patients) fractures they reviewed were attributable to either motor vehicle accidents ($\sim$70%) or

falls. Alker et al. (19) examined postmortem radiographs of 312 victims of fatal motor vehicle accidents. The cohort exhibited 98 injuries of the cervical spine, of which 70 were seen in the craniovertebral junction. The authors, although not quantifying the degree of dens fractures, hypothesized that odontoid fractures were probably due to hyperextension because of the posterior displacement of the fracture pieces.

There is considerable controversy as to the major load path that causes odontoid fractures. A review of the clinical and laboratory research literature fails to designate a consensus on this issue. Schatzker et al. (47) reviewed clinical case presentations and concluded that odontoid fractures are not the result of simple tension and that there must exist a complex combination of forces needed to produce these failures. Althoff (49) performed a cadaver study, whereby he applied various combinations of compression and horizontal shear to the head via a pendulum. Before load onset the head was placed in neutral, extension or flexion. The position of the load and the angle of impact, determining the degree of compression with shear, was changed for each experiment. The results indicated that an impact in the sagittal plane (anterior or posterior) produced fractures that involved the C2 body (Type III). As the force vector moved from anterior to lateral, the location of the fracture moved superiorly, with lateral loading producing Type I fractures. This led the author to propose a new hypothesis: impact loading corresponding to combined horizontal shear and compression results in odontoid fractures. Althoff dismissed the contributions of sagittal rotation (flexion and extension) to the production of resultant odontoid fracture.

Mouradian et al. (50) reported on a cadaver and clinical model of odontoid fracture. In their opinion, "it seems reasonable to assume that shearing or bending forces are primarily involved." The cadaver experimentation involved anterior or lateral translation of the occiput as well as lateral translation of the atlantal ring. In forward loading, the odontoid was fractured in 9 of the 13 cases, with 8 Type III fractures and 1 Type II fracture. The lateral loading specimens evidenced similar patterns of odontoid fracture regardless of the point of load application (on the occiput or on the atlas). In 11 specimens, lateral loading resulted in 10 Type II fractures and 1 Type III fracture. The clinical model involved reviewing 25 cases of odontoid fracture. They reported that 80% of these cases resulted from flexion or flexion–rotation injuries. They pointed out that the clinical data does not reflect the lateral loading cadaver experimentation results. In fact, they state that "a pure lateral blow probably did not occur in any [clinical] case". However, their clinical data indicated that the remaining 20% of the odontoid injuries could be ascribed to extension injuries. The technical difficulties precluded cadaver experimentation of this possible mechanism. Experimental investigations dealing with the pathogenesis of odontoid fractures have failed to produce a consensus as to the etiology of these fractures. These findings may actually reflect the diversity of causal mechanisms, suggesting the various mechanical factors are coincident in producing these fractures. It is difficult to diskern if this is the case or if this is due to the inhomogeneity of cadaver experiment methodology. That is, some of the boundary and loading conditions used by the surveyed studies are vastly different and have produced divergent results. In addition, the anatomical variants of the craniovertebral osteo-ligamentous structures could also be integral to the cadaver study outcomes. The purpose of the study undertaken by Puttlitz et al. (51) was to utilize of the finite element method, in which the loading and kinematic constraints can be exactly designated, for elucidating the true fracture etiology of the upper cervical spine. Previous laboratory investigations of odontoid process failure have used cadaver models. However, shortcomings associated with this type of experimentation and the various loading and boundary conditions may have influenced the resulting data. Utilization of the FE method for the study of odontoid process failure has eliminated confounding factors often seen with cadaveric testing, such as interspecimen anatomic variability, age-dependent degeneration, and so on. This has allowed us to isolate changes in complex loading conditions as the lone experimental variable for determining odontoid process failure.

There are many scenarios, that are capable of producing fracture of the odontoid process. Force loading, in the absence of rotational components, can reach maximum von Mises stresses that far exceed 100 MPa. Most of these loads are lateral or compressive in nature. The maximum stress obtained was 177 MPa due to a force directed in the posteroinferior direction. The net effect of this load vector and its point of application, the posterior aspect of the occiput, is to produce a compression, posterior shear, and extension due to the load's offset from the center of rotation. This seems to suggest that extension and compression can play a significant role in the development of high stresses, and possibly failure, of the odontoid. The location of the maximum stress for this loading scenario was in the region of a Type I fracture. The same result, with respect to laterally loading, was obtained by Althoff (49). However, he dismissed the contribution of sagittal plane rotation to development of odontoid failures. The results of this study disagree with that finding. Posteroinferior loading with extension produced a maximum von Mises stress in the axis of 226 MPa. As stated above, the load vector for this case intensifies the degree of extension, probably producing hyperextension. The addition of the extension moment did not change the location of the maximum stress, still identifiable in the region of a Type I fracture. The clinical study by Moradian et al. (50) suggested that almost 20% of the odontoid fracture cases they reviewed involved some component of extension. The involvement of extension in producing odontoid process failures can be explained by its position with respect to the atlantal ring and the occiput. As extension proceeds, the contact force produced at the atlanto-dental articulation increases, putting high bending loads on the odontoid process. The result could be failure of the odontoid. Increasing tension of the alar ligaments as the occiput extends could magnify these bending stress via superposition of the loads, resulting in avulsion failure of the bone (Type I).

While the FE model predicted mostly higher stresses with the addition of an extension moment, the model showed that, in most cases, flexion actually mitigates

the osseous tissue stress response. This was especially true for compressive (inferior) force application. Flexion loading with posterior application of an inferior load vectorally decreases the overall effect of producing extension on the occiput. None of the studies surveyed for this investigation pinpointed flexion, per se, as a damage mechanism for odontoid failure. The findings of this study supported the lack of evidence in support of flexion as being a causal mechanism for failure. In addition, the data suggested that flexion can act as a preventative mechanism against odontoid fracture.

Once again, the lateral bending results support the hypothesis of extension being a major injury vector in odontoid process failure. Inferior and posteroinferior loads with lateral rotation resulted in the highest maximal von Mises stress in the axis. Lateral loading also intensified the maximal stress in compression, suggesting rotations that incorporate a component of both lateral and extension motion may cause odontoid failures. Many of the lateral bending scenarios resulted in the maximum von Mises stress being located in the Type II and Type III fracture regions. In fact, the only scenarios that lead to the maximum stress in the Type I area was when there was an inferior or posterior load applied with the lateral bending. This is, again, suggestive that the extension moment, produced by these vectors and their associated moment arms (measured from the center of rotation), can result in more superiorly-located fractures.

Overall, this investigation has indicated that extension and the application of extension via force vector application, causes the greatest risk of superior odontoid failure. The hypothesis of extension as a causal mechanism of odontoid fracture includes coupling of this motion to other rotations. Flexion seems to provide a protective mechanism against force application that would otherwise cause a higher risk of odontoid failure.

Middle and Lower Cervical Spine. In the C2-T1 region of the spine, as in the upper cervical region, instabilities in a laboratory setting have been produced in an effort to understand the dynamics of traumatic forces on the spine (19). In one study, fresh ligamentous porcine cervical spine segments were subjected to flexion-compression, extension-compression, and compression-alone loads at high speeds (dynamic–impact loading) (19). The resultant injuries were evaluated by anatomic dissection. The results that the severity of the injuries were related mostly to the addition of bending moments to high speed axial compression of the spine segment, since compression alone produced the least amount of injury and no definite pattern of injuries could be identified. Other investigators have reported similar results (19).

Lumbar Spine. The onset of low back pain is sometimes associated with a sudden injury. However, it is more often the result of cumulative damage to the spinal components induced by the presence of chronic loading on the spine. Under chronic loading, the rate of damage may exceed the rate of repair by the cellular mechanisms, thus weakening the structures to the point where failure occurs under midly abnormal loads. Chronic loading to structures may occur under a variety of conditions (52,53). One type of loading is heavy physical work prevalent among blue collar workers. Lifting not only induces large compressive loads across the segment, but tends to be associated with bending and twisting (54). Persons with jobs requiring the lifting of objects of > 11.3 kg > 25 times/day have over three times the risk for acute disk prolapse than people whose jobs do not require lifting (55). If the body is twisted during lifting, the risk is even higher with less frequent lifting. The other major class of loading associated with low back pain is posture related, for example, prolonged sitting–sedentary activities, and posture that involve bending over while sitting. Prolonged sitting may be compounded by vibration, such as observed in truck drivers (52,56,57).

The effects of various types of cyclic loads on the specimen behavior have been investigated (52,55). For example, Liu et al. subjected ligamentous motion segments to cyclic axial loads of varying magnitudes until failure or 10,000 cycles, which ever occurred first (53). Test results fell in to two categories, stable and unstable. In the unstable group, fracture occurred within the 6000 cycles of loading. The radiographs in the unstable group revealed generalized trabecular bony microfailure. Cracks were found to propagate from the periphery of the subcondral bone. After the removal of the organic phase, the unstable group specimens disintegrated into small pieces, as opposed to stable group specimens. This suggests that microcrack initiation occurs throughout the inorganic phase of the subchondral bone as a result of axial cyclic loading. In response to cyclic axial twisting of the specimens, Liu et al. noticed a discharge of synovial fluid from the articular joints (58). Specimens that exhibited an initial angular displacement of $< 1.5^{\circ}$, irrespective of the magnitude of the applied cyclic torque, did not show any failures. On the other hand, specimens, exhibiting initial rotations $< 1.5^{\circ}$, fractured before reaching 10,000 cycles. These fractures included bony failure of facets and/or tearing of the capsular ligaments.

Chronic vibration exposure and prolonged sitting are also known to lead to spinal degeneration. Spinal structures exhibit resonance between 5 and 8 Hz (56–59). *In vivo* and *in vitro* experimental and analytical studies have shown that the intradiscal pressure and motion increase when spinal structures experience vibration, such as during driving cars–trucks, at the natural resonant frequency (59). Prolonged sitting alone or in conjunction with chronic vibration exposure is also a contributing factor to spinal degeneration. A finite element-based study revealed that prolonged sitting led to an increase in disk bulge and the stresses in the annulus fibers located at the outer periphery (59,60).

Lee et al. (61) quantatively analyzed occlusion of the dural-sac in the lumbar spine was quantitatively analyzed by utilizing a finite element lumbar spine model. In the static analysis, it was found that < 2 kN of compressive load could not produce dural-sac occlusion, but the compression together with extension moment was more likely to produce the dural-sac occlusion. The 7.4% of occlusion was obtained when the 8 N·m of extension moment was added to 2 kN of compressive load that alone did not create any occlusion. The magnitude of occlusions was increased

to 10.5% as the extension moment increased to 10 N·m with the same 2 kN of compressive load. In creep analysis, 10 N·m extension, kept for 3600 s, induced 6.9% of occlusion, and 2.4% of volume reduction in the dural-sac. However, flexion moment did not produce any occlusion in the dural-sac, but increased the volume instead because it caused stretching of the dural-sac coupled with vertebral motion. As a conclusion, occlusions resulted mainly from the slackening of the ligamentum flavum and disk bulging. Furthermore, the amount of occlusion was strongly dependent with loading conditions and the viscoelastic behavior of materials as well.

Changes in Motion due to Degeneration–Trauma

The degenerative process can effect all of the spinal elements and trauma can lead to partial or full destruction of the spinal elements. As such the motion behavior of the segment will change.

Cervical Spine Region. The rotation-limiting ability of the alar ligament was investigated by Dvorak et al. (62,63). A mean increase of 10.8° or 30% (divided equally between the occipitoatlantal and atlantoaxial complexes) in axial rotation was observed in response to an alar lesion on the opposite side. Oda et al. (39,40) determined the effects of alar ligament transections on the stability of the joint in flexion, extension, and lateral bending modes. Their main conclusion was that the motion changes occurred subsequent to alar ligament transection. The increases, however, were directional-dependent. Crisco et al. (64) compared changes in 3D motion of C1 relative to C2 before and after the capsular ligament transections in axial rotation. Two groups of cadaveric specimens were used to study the effect of two different sequential ligamentous transections. In the first group ($n = 4$), transection of the left capsular ligament was followed by transection of the right capsular ligament. In the second group ($n = 10$), transection of the left capsular ligament preceded transection of left and right alar and transverse ligaments. The greatest changes in motion occurred in axial rotation to the side opposite the transection. In the first group, transection of left capsular ligaments resulted in a significant increase in axial rotation ROM to the right of 1°. After the right capsular ligament was transected, there was a further significant increase of 1.8° to the left and of 1.0° to the right. Lateral bending to the left also increased significantly by 1.5° after both ligaments were cut. In the second group, with the nonfunctional alar and transverse ligaments, transection of the capsular ligament resulted in greater increases in ROM: 3.3° to the right and 1.3° to the left. Lateral bending to the right also increased significantly by 4.2°. Although the issue is more complex than this, in general these studies show that the major function of the alar ligament is to prevent axial rotation to the contralateral side. Transection of the ligament increases the contralateral axial rotation by ~15%.

The dens and the intact transverse ligament provide the major stability at the C1-C2 articulation. The articular capsules between C1 and C2 are loose, to allow a large amount of rotation and provide a small amount of stability. Although the C1-C2 segment is clinically unstable after failure of the transverse ligament, resistance against gross dislocation is probably provided by the tectorial membrane, the ala, and the apical ligaments. With transection of the tectorial membrane and the ala ligaments, there is an increased flexion of the units of the occipital–atlantoaxial complex and a subluxation of the occiput (4h). It was also demonstrated that transection of the ala ligament on one side causes increased axial rotation to the opposite side by ~30%.

Fielding et al. (65) performed a biomechanical study investigating lesion development in rheumatoid arthritis. Their study tested 20 cadaveric occipitoatlanto–axial specimens for transverse ligament strength by application of a posterior force to the atlantal ring. They found atlantoaxial subluxation of 3–5 mm and increased atlas movement on the axis after rupture of the transverse ligament. From this study, Fielding et al. were able to conclude that the "transverse ligament represents a strong primary defense against anterior shift of the first cervical vertebra." Puttlitz et al. (66) developed an experimentally validated ligamentous, nonlinear, sliding contact 3D finite element (FE) model of the C0-C1-C2 complex generated from 0.5-mm thick serial computed tomography scans (Fig. 9). The model was used to determine specific structure involvement during the progression of RA and to evaluate these structures in terms of their effect on clinically observed erosive changes associated with the disease by assessing changes in loading patterns and degree of AAS (see Table 9 for terminology). The role of specific ligament involvement during the development and advancement of AAS was evaluated by calculating the AADI and PADI after reductions in transverse, ala, and capsular ligament stiffness. (The stiffness of transverse, alar, and capsular ligaments was sequentially reduced by 50, 75, and 100% of their intact values.) All models were subjected to flexion moments, replicating the clinical diagnosis of RA using full flexion lateral plane radiographs. Stress profiles at the transverse ligament-odontoid process junction were monitored. Changes in loading profiles through the C0-C1 and C1-C2 lateral articulations and their associated capsular ligaments were calculated. Posterior atlantodental interval (PADI) values were calculated to correlate ligamentous destruction to advancement of AAS. As an isolated entity, the model predicted that the transverse ligament had the greatest effect on AADI in the fully flexed posture. Without transverse ligament disruption, both ala and capsular ligament compromise did not contribute significantly to the development of AAS. Combinations of ala and capsular ligament disruptions were modeled with transverse ligament removal in an attempt to describe the interactive effect of ligament compromise, which may lead to advanced AAS. Ala ligament compromise with intact capsular ligaments markedly increased the level of AAS (Table 9). Subsequent capsular ligament stiffness loss (50%) with complete ala ligament removal led to an additional decrease in PADI of 0.92 mm. Simultaneous resection of the transverse, ala, and capsular ligaments resulted in a highly unstable situation. The model predicted stresses at the posterior base of the odontoid process greatly reduced, with transverse ligament compromise beyond 75%

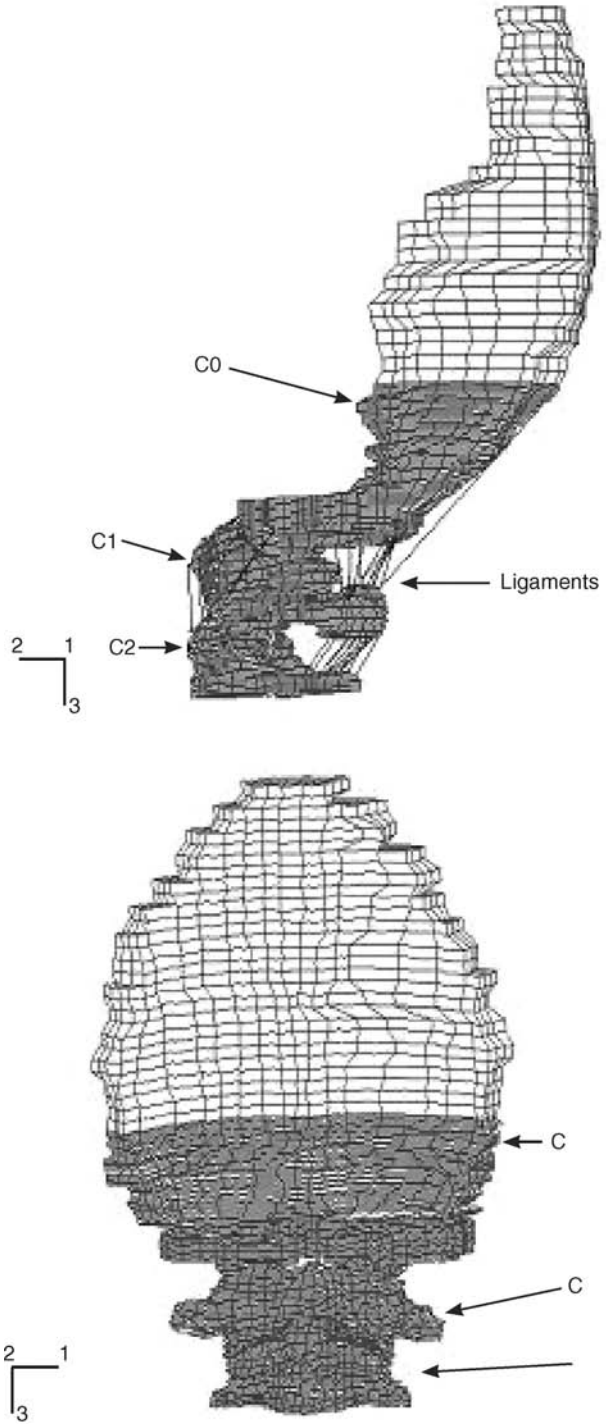


Figure 9. Finite element models of the upper cervical spine used to study the biomechanics of rheumatoid arthritis. (Taken from Ref. 66.)

(Fig. 10). Decreases through the lateral C0-C1 and C1-C2 articulations were compensated by their capsular ligaments. The data indicate that there may be a mechanical component (in addition to enzymatic degradation) associated with the osseous resorption seen during RA. Specifically, erosion of the base of the odontoid may involve Wolff's law unloading considerations. Changes through

Table 9. Combinations of Ligament Stiffness Reductions with the Resultant Degree of AAS, as Indicated by the AADI and PADI Values at Full Flexion (1.5 N · m moment)[a]

Reduction in Ligament Stiffness, %			Criteria, mm	
Transverse	Alar	Capsular	AADI	PADI
0	0	0	2.92	15.28
100	0	50	5.77	12.43
100	0	75	6.21	11.99
100	50	0	7.42	10.79
100	75	0	7.51	10.71
100	100	50	8.43	9.83

[a]Zero (0) ligament stiffness values represent completely intact ligament stiffness, "100" corresponds to total ligament destruction (via removal). (Taken from Ref. 66.) AAS = anterior atlantoaxial subluxation, AADI = anterior atlantodental interval, PADI = posterior atlantodental interval.

the lateral aspects of the atlas suggest that this same mechanism may be partially responsible for the erosive changes seen during progressive RA. The PADI values indicate that complete destruction of the transverse ligament coupled with alar and/or capsular ligament compromise exist if advanced levels of AAS are present.

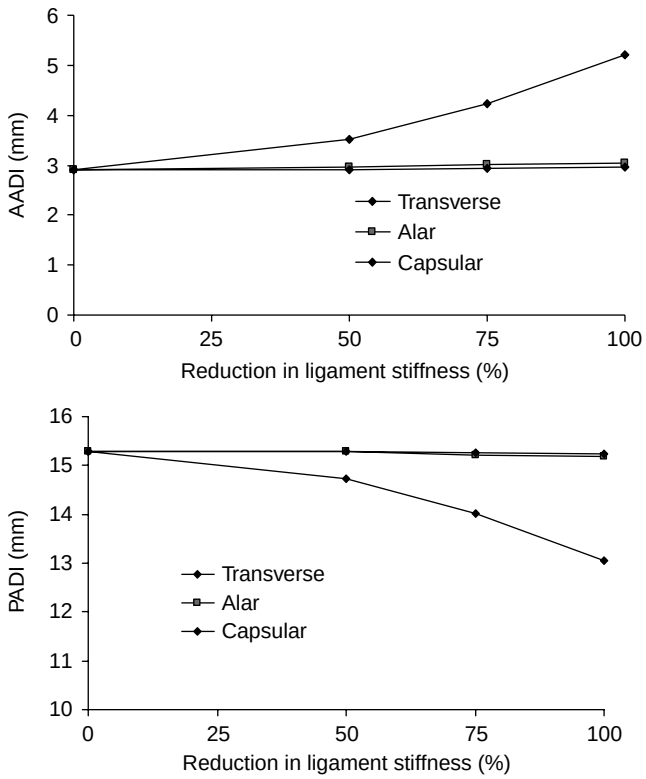


Figure 10. (a) Anterior atlantodental interval (AADI) (b) and posterior atlantodental interval (PADI) calculated for the intact model and models with stiffness reductions of the transverse, alar, and capsular ligaments at the fully flexed posture (1.5 N·m moment load). Each ligament's stiffness was altered while holding the other two at the baseline value (completely intact). (Taken from Ref. 66.)

In vitro studies to determine the feasibility of the "stretch test" in predicting instability of the spine in the cervical region were performed by Panjabi et al. (67). Four cervical spines (C1-T1; ages 25–29) were loaded in axial tension in increments of 5 kg to a maximum of one third of the specimen's body weight. The effects of sequential AP transection of soft tissues of a motion segment on the motion in one group and of posterior–anterior transections in another group were investigated. The intact cervical spine went into flexion under axial tension. Anterior transection produced extension. Posterior transection produced opposite results. Anterior injuries creating displacements of 3.3 mm at the disk space (with a force equal to one-third body weight) and rotation changes of $\sim 3.8^\circ$ were considered precursors to failure. Likewise, posterior injuries resulting in 27 mm separation at the tips of the spinous process and an angular increase of 30° with loading were considered unstable. This work supports the concept that spinal failure results from transection of either all the anterior elements or all the posterior plus at least two additional elements.

In a study by Goel et al. (68,69), the 3D load-displacement motion of C4-C5 and C5-C6 as a function of transection of C5-C6 ligaments was determined. Transection was performed posteriorly, starting with the supraspinous and interspinous ligaments, followed by the ligamentum flavum and the capsular ligaments. With the transection of the capsular ligaments, the C5-C6 motion segment (injured level) showed a significant increase in motion in extension, lateral bending, and axial rotation. A significant increase in flexion resulted when the ligamentum flavum was transected.

A major path of loading in the cervical spine is through the vertebral bodies, which are separated by the intervertebral disk. The role of the cervical intervertebral disk has received little attention. A finite element model of the ligamentous cervical spinal segment was used to compute loads in various structures in response to clinically relevant loading modes (70). The objective was to predict biomechanical parameters, including intradiskal pressure, tension in ligaments, and forces across facets that are not practical to quantify with an experimental approach. In axial compression, 88% of the applied load passed through the disk. The interspinal ligament experienced the most strain (29.5% in flexion, and the capsular ligaments were strained the most (15.5% in axial rotation). The maximum intradiskal pressure was 0.24 MPa in flexion with an axial compression mode (1.8 N·m of flexion moment + 73.6 N of compression). The anterior and posterior disk bulges increased with an increase in axial compression (up to 800 N). The results provide new insight into the role of various elements in transmiting loads.

This model was further used to investigate the biomechanical significance of uncinate processes and Luschka joints (71). The results indicate that the facet joints and luschka joints are the major contributors to coupled motion in the lower cervical spine and that the uncinate processes effectively reduce motion coupling and primary cervical motion (motion in the same direction as load application), especially in response to axial rotation and lateral bending loads. Luschka joints appear to increase primary cervical motion, showing an effect on cervical motion opposite to that of the uncinate processes. Surgeons should be aware of the increase in motion accompanied by resection of the uncinate processes.

Cervical spine disorders such as spondylotic radiculopathy and myelopathy are often related to osteophyte formation. Bone remodeling experimental–analytical studies have correlated biomechanical responses, such as stress and strain energy density, to the formation of bony outgrowth. Using these responses of the spinal components, a finite element study was conducted to investigate the basis for the occurrence of disk-related pathological conditions. An anatomically accurate and validated intact element model of the C4-C5-C6 cervical spine was used to simulate progressive disk degeneration at the C5-C6 level. Slight degeneration included an alteration of material properties of the nucleus pulposus representing the dehydration process. Moderate degeneration included an alteration of fiber content and material properties of the annulus fibrosus representing the disintegrated nature of the annulus in addition to dehydrated nucleus. Severe degeneration included decrease in the intervertebral disk height with dehydrated nucleus and disintegrated annulus. The intact and three degenerated models were exercised under compression, and the overall force-displacement response, local segmental stiffness, annulus fiber strain, disk bulge, annulus stress, load shared by the disk and facet joints, pressure in the disk, facet and uncovertebral joints, and strain energy density and stress in the vertebral cortex were determined. The overall stiffness (C4-C6) increased with the severity of degeneration. The segmental stiffness at the degenerated level (C5-C6) increased with the severity of degeneration. Intervertebral disk bulge and annulus stress and strain decreased at the degenerated level. The strain energy density and stress in vertebral cortex increased adjacent to the degenerated disk. Specifically, the anterior region of the cortex responded with a higher increase in these responses. The increased strain energy density and stress in the vertebral cortex over time may induce the remodeling process according to Wolff's law, leading to the formation of osteophytes.

Thoracolumbar Region. The most common vertebral levels involved with the thoracolumbar injuires are T12-L1 (62%) and L1-L2 (24%) (22,25). The injuires, depending on the severity of the trauma, have included disruption of the posterior ligaments, fracture and dislocation of the facets, and fracture of the vertebral bodies with and without neural lesions. Operative intervention is often suggested to restore spinal stability. These involve use of spinal instrumentation, vertebroplasty, and host of other procedures which have been described elsewhere in this article.

For ease in description of these injuries, conceptually the osteoligamentous structures of the spine have been grouped into three "columns"; anterior, middle, and posterior. The anterior column consists of the anterior longitudinal ligament, anterior annulus fibrosus, and the anterior part of the vertebral body. The middle column consists of the posterior longitudinal ligament, posterior annulus

fibrosus, and the posterior vertebral body wall. The posterior column contains the posterior bony complex or arch (including the facet joints), and the posterior ligamentous complex composed of the supraspinous ligament, interspinous ligament, facet joint capsules, and ligamentum flavum.

As per this classification, a compression fracture is a fracture of the anterior column with the middle and posterior columns remaining intact. In severe cases, there may also be a partial tensile failure of the posterior column, but the vertebral ring, consisting of the posterior wall, pedicles, and lamina, remains totally intact in a compression fracture. A burst fracture is a fracture of the anterior and middle columns under compression; the status of the posterior column can vary. In the burst fracture, there is fracture of the posterior vertebral wall cortex with marked retropulsion of bone into the spinal canal, obstructing, on average, 50% of the spinal canal cross-section. There may be a tilting and retropulsion of a bone fragment into the canal from one or both endplates. In contrast to the compression fracture, there is loss of posterior vertebral body height in a burst fracture. The seat-belt type injuries feature failure of the middle and posterior columns under tension, and either no failure or slight compression failure of the anterior column. In fracture dislocations, the anterior, middle, and posterior columns all fail, leading to subluxation or dislocation. There may be "jumped facets" or fracture of one articular process at its base or at the base of the pedicle. There is also disruption of the anterolateral periosteum and anterior longitudinal ligament. If the separation goes through the disk, there will be some degree of wedging in the vertebral body under the disk space. However, the fracture cleavage may pass through the vertebral body itself, resulting in a "slice fracture".

There are four mechanisms of fracture that have been hypothesized in the literature to explain why the thoracolumbar region experiences a higher frequency of injury than adjacent regions. The hypotheses state that a thoracolumbar fracture sequence can be put into motion by stress concentrations arising from (*1*) spinal loading conditions; (*2*) material imperfections in spine; (*3*) differences in spinal stiffness and physiological range of motion characteristics between the thoracic and lumbar regions; and (*4*) abrupt changes in spinal anatomy, especially facet orientations. As always, there is no conseus for these mechanisms.

A few of the experimental investigations that have attempted to reproduce the clinical fracture patterns are as follows: In one study, cadaver motion segments were subjected to loads of different magnitude and direction: compression, flexion, extension, lateral flexion, rotation, and horizontal shear to reproduce all varieties of spinal injury experimentally by accurately controlled forces. For a normal disk, increases of intradiskal pressure and bulging of the annulus occur under application of axial compressive load. With increased application of force, the end-plate bulges and finally cracks, allowing displacement of nuclear material into the vertebral body. Continued loading of the motion segment results in a vertical fracture of the vertebral body. If a forward shear component of force accompanies the compression force, the line of fracture of the vertebral body is not vertical but is oblique. Different forms of fracture could be produced by axial compressive loading if the specimens were from older subjects (i.e., the nucleus was no longer fluid), or if the compressive loading was asymmetrical. Under these conditions, the transmission of load mainly through the annulus is responsible for the (*1*) tearing of the annulus, (*2*) general collapse of the vertebra due to buckling of the sides (cortical wall), and (*3*) marginal plateau fracture.

Thoracolumbar burst fractures in cadaver specimens have also been produced by dropping a weight such that the prepared column is subjected to axial-compressive impact loads. The potential energies of the failing weights used by these researchers have been 200 and 300 N·m. Fracture in four of the seven specimens apparently started at the nutrient foramen. The nutrient foramen may perhaps be viewed as a local area of material imperfection where stresses may be concentrated during loading, leading to fracture. Other researchers are apparently unable to consistently produce burst fractures *in vitro* without first creating artificial "stress raisers" in the vertebral body by means of cuts or slices into the bone.

Panjabi et al. (3) conducted *in vitro* flexibility tests of 11 T11-L1 specimens to document the 3D mechanical behavior of the thoracolumbar junction region (see section on Construct Testing for an explanation). Pure moments up to 7.5 N·m were applied to the specimens in flexion–extension, left–right axial torque, and right–left lateral bending. The authors reported the average flexibility coefficients of the main motions (range of motion divided by the maximum applied load). For extension moment, the average flexibility coefficient of T11-T12. (0.32°/N·m) was significantly less than that of T12-L1 (0.52°/N·m). For axial torque, the average flexibility coefficient of T11-T12 (0.24°/N·m) was significantly greater than that of T12-L1 (0.16°/N·m). The authors attributed these biomechanical differences to the facet orientation. They speculated that thoracic-type facets would offer greater resistance to extension than the more vertically oriented lumbar-type facets while the lumbar-type facets would provide a more effective stop to axial rotation than thoracic-type facets. No other significant biomechanical differences were detected between T11-T12 and T12-L1. In addition to these observations, authors found that for flexion torque, the average flexibility coefficients of the lumbar spine (e.g., L1-L2, 0.58°/N·m; L5-S1, 1.00°/N·m) were much greater than those of both T11-T12 (0.36°/N·m) and T12-L1 (0.39°/N·m). They identified this change in flexion stiffness between the thoracolumbar and lumbar regions as a possible thoracolumbar injury risk factor.

Lumbar Spine Region. The porosity of the cancellous bone within the vertebral body increases with age, especially in women. The vertebral body strength is known to decrease with increase in porosity of the cancellous bone, a contributing factor to the kyphosis normally seen in an elderly person (4). As the trabeculae reduce in size and number, the cortical shell must withstand greater axial load, thus increasing the thickness of the shell obeying the principles of Wolff's law. Edwards et al. (72) demonstrated cortical shell thickening of osteoporotic vertebral bodies

compared to that of normal vertebral bodies. Furthermore, there was increased incidence of osteophytic development along the cortical shell in the regions of highest stress within the compromised osteoporotic vertebrae.

The normal disk consists of a gel-like nucleus encased in the annulus. In a normal healthy person, the disk acts like a fluid filled cavity. With age, the annulus develops radial, circumferential and rim lesions, and the nucleus becomes fibrous. Using a theoretical model in which cracks of varying lengths were simulated, Goel et al. found that the interlaminar shear stresses (and likewise displacements) were minimal until the crack length reached 70% of the annulus depth (73). Likewise, dehydration of the nucleus (extreme case totally ineffective like in a total nucleotomy) also was found to lead to separation of the lamina layers and an increase in motion (74). Thus, the results support the observation that the increase in motion really occurs in moderately degenerated disks.

Posner et al. investigated the effects of transection of the spinal ligaments on the stability of the lumbar spine (75). The ligaments were transected in a sequential manner, either anterior to posterior or posterior to anterior. While cutting structures from the anterior to posterior portion of the spine, extension loading caused a significant residual deformation after the anterior half of the disk was cut. Cutting from the posterior to anterior region, flexion loading caused significant residual motion upon facet joint transection. The role of ligaments becomes more prominent in subjects whose muscles are not fully functional. Using a finite element model in which the muscular forces during lifting were simulated, Kong et al. found that a 10% decrease in the muscle function increased loads borne by the ligaments and the disks (76). The forces across the facet joint decreased.

The orientation of facet becomes more parallel to the frontal plane as one goes down from L1 to S1 (77). Other factors can also contribute to changes in facet orientation in a person. The facet orientation, especially at L4-5 and L5-S1, plays a role in producing spondylolisthesis. Kong et al. using a finite element of the ligamentous lumbar segment (Fig. 11a) found that as the facet orientation becomes more sagittal, the A–P translation across the segment, increases in response to the load applied, Fig. 11b. The increase in flexion angle was marginal.

Changes in Motion Due to Surgical Procedures

Cervical Region. *In vivo* "injuries" result in disk degeneration and may produce osteophytes, ankylosed vertebras, and changes in the apophyseal joints (78). The effects of total diskectomy on cervical spine motions are of interest (79). Schulte and colleagues reported a significant increase in the motion after C5-C6 diskectomy (80). Motion between C5-C6 increased in flexion (66.6%), extension (69.5%), lateral bending (41.4%), and axial rotation (37.9%). In previous studies, Martins (81) and Wilson and Campbell (82) could not detect increases in motion roentgenographically and deemed the spines functionally stable.

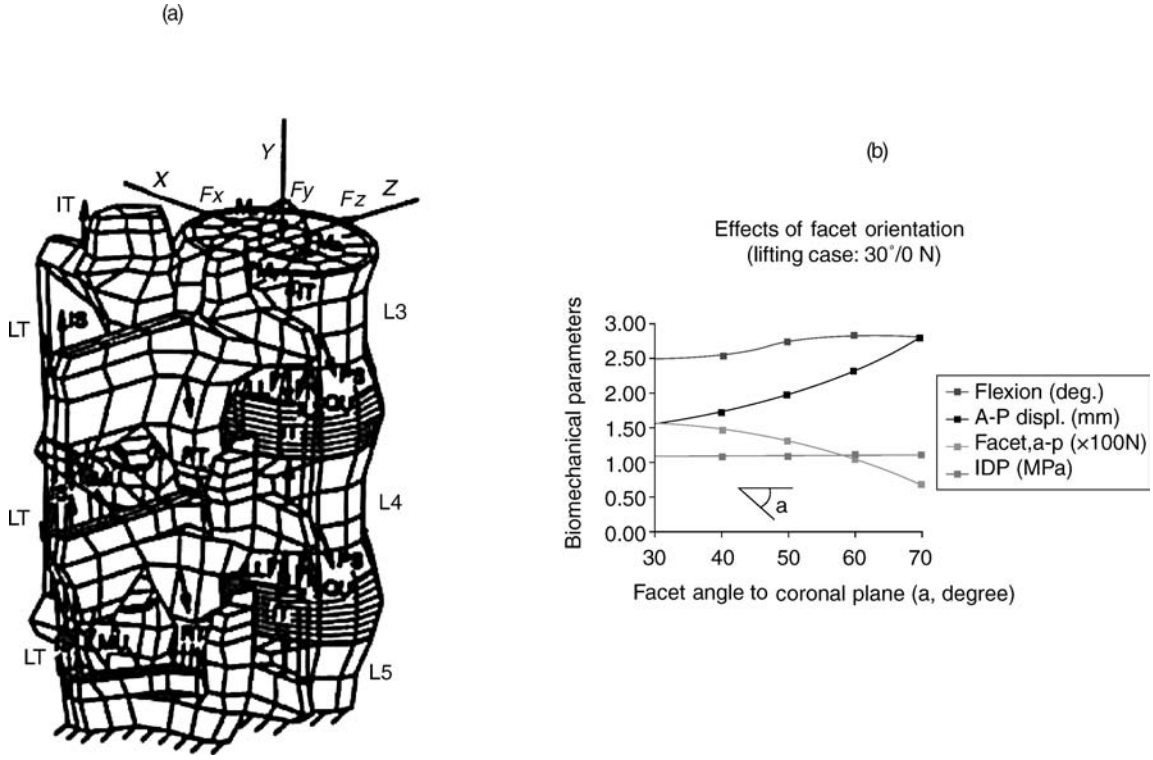


Figure 11. The finite element of a lumbar segment used to predict the effect of facet orientations on the motion and loads in various spinal components in a motion segment. (Taken from Ref. 77.)

The reasons for this diskrepancy in results are not apparent. The experimental designs were quite different as were the methods of motion measurement. However, the disk obviously is a major structural and functional component of the cervical spine.

The contribution of facet and its capsule to the stability of the cervical spine has been well documented using both *in vitro* laboratory models (83–85) and mathematical models (70,86,87,88). Facet joints play an integral part in the biomechanical stability of the cervical spine. Cusick et al. (89) found that total unilateral and bilateral facetectomies decreased compression-flexion strength by 31.6 and 53.1%, respectively. Facetectomy resulted in an anterior shift of the IAR, resulting in increased compression of the vertebral body and disk. This work confirmed the findings of Raynor et al. (63,64) who reported that bilateral facetectomy of as much as 50% did not significantly decrease shear strength; however, with a 75% bilateral facetectomy, a significant decrease in shear strength was noted. One should take great care when exposing an unfused segment to limit facet capsule resection to $< 50\%$. With resection of $> 50\%$ of the capsule, postoperative hypermobility can occur and may require stabilization.

In contrast, studies that focused on the effects of laminectomy alone have been few and still unclear. Goel et al. were the first to evaluate the effects of cervical laminectomy with *in vitro* spine models (83,90). They found 10% increase of motion in flexion-extension using 0.3 N·m after a two level laminectomy. Zdeblick et al. did not find motion changes in flexion–extension after one level laminectomy under 5 N·m (84,85). Cusick et al. successfully showed that three level cervical laminectomy (C4-C6) induces a significant increase in total column flexibility using physiologic compression-flexion forces (86,87). Nevertheless, it seems difficult to estimate the instantaneous combination of physiologic compression and flexion forces. Therefore, quantitative evaluation might be difficult with this model. Our results indicate significant increase of spinal column motion in flexion (24.5%), extension (19.1%), and axial rotation (23.7%) using 1.5 N·m after a four level (C3-C6) laminectomy. Cervical vertebral laminae may transmit loads. Laminectomies result in the removal of part of this loading path and the attachment points for the ligamentum flavum, interspinous ligament, and the supraspinous ligament. It is not surprising that total laminectomy results in significant modifications in the motion characteristics of the cervical spine, especially in children. For example, Bell et al. (91) reported that multiple-level cervical laminectomy can lead to increase in postoperative hyperlordosis or kyphosis in children. However, there was no correlation between diagnosis, sex, location, or number of levels decompressed and the subsequent development of deformity. Postlaminectomy spinal deformity in the cervical spine, however, is rare in adults, probably owing to stiffening of the spine with age and changes in facet morphology. Goel et al. (89) removed the laminae of multisegmental cervical spines (C2-T2) at the level of C5 and C6 (total laminectomy); in flexion–extension mode, demonstrating an increase in motion of $\sim 10\%$.

In another *in vitro* study, the effects of multilevel cervical laminaplasty (C3-C6) and laminectomy with increasing amounts of facetectomy (25% and more) on the mechanical stability of the cervical spine were investigated (88). Cervical laminaplasty was not significantly different from the intact control, except for producing a marginal increase in axial rotation. However, cervical laminectomy with facetectomy of 25% or more resulted in a highly significant increase in cervical motion as compared with that of the intact specimens in flexion, extension, axial rotation, and lateral bending. There was no significant change in the coupled motions after either laminaplasty or laminectomy. The researchers recommended that concurrent arthrodesis be performed in patients undergoing laminectomy accompanied by $> 25\%$ bilateral facetectomy. Alternatively, one may use laminaplasty to achieve decompression if feasible. More recently, the effect of laminaplasty on the spinal motion using *in vivo* testing protocols have also been investigated (92–94). Kubo et al. (95) undertook an *in vitro* 3D kinematic study to quantify changes after a double door laminoplasty. Using fresh cadaveric C2-T1 specimens, sequential injuries were created in the following order: intact, double door laminoplasty (C3-C6) with insertion of hydroxyapatite (HA) spacers, laminoplasty without spacer, and laminectomy. Motions of each vertebra in each injury status were measured in six loading modes: flexion, extension, right and left lateral bending, and right and left axial rotation. Cervical laminectomy showed significant increase in motion compared to intact control in flexion (25%: $P < 0.001$), extension (19%: $P < 0.05$), and axial rotation (24%: $P < 0.001$) at maximum load. Double door laminoplasty with HA spacer indicated no significant difference in motion in all loading modes compared to intact. Laminoplasty without spacer showed intermediate values between laminoplasty with spacer and laminectomy in all loading modes. Initial slack of each injury status showed similar trends that of maximum load although mean % changes of laminectomy and laminoplasty without spacer were greater than that of maximum load. Double door laminoplasty with HA spacer appears to restore the motion of the decompressed segment back to its intact state in all loading modes. The use of HA spacers well contribute to maintain the total stiffness of cervical spine. In contrast, laminectomy seems to have potential leading postoperative deformity or instability.

Kubo et al. (96) undertook another study with the aim to evaluate the biomechanical effects of multilevel foraminotomy and foraminotomy with double door laminoplasty as compared to foraminotomy with laminectomy. Using fresh human cadaveric specimens (C2-T1), sequential injuries were created in the following order: intact, bilateral foraminotomies (C3/4, C4/5, C5/6), laminoplasty (C3-C6) using hydroxyapatite spacer, removal of the spacers, and laminectomy. Changes in the rotations of each vertebra in each injury status were measured in six loading modes: flexion–extension, right–left lateral bending, and right–left axial rotation. Foraminotomy alone, and following laminoplasty showed no significant differences in motion compared to the intact with the exception of axial rotation. After removal of the spacers and following a laminectomy, the motion increased significantly in flexion and axial rotation. The ranges of initial slack showed similar trends when compared to the results at

maximum load. Clinical implications of these observations are presented.

Lumbar Region. The spine is naturally shaped to properly distribute and absorb loads, therefore, any surgical technique involving dissection of spinal components can disrupt the natural equilibrium of the spinal elements and lead to instability. The amount and origin of pain within the spine usually determines the type of surgical procedure for a patient. Such procedures include the removal of some or all of the laminae, facets, and/or disks. A certain increase in the range of motion within the spine can be attributed to each procedure. The increased range of motion can also lead to more pain, as noted by Panjabi and others, who used an external fixator to stabilize the spine (97,98). The fixator decreased the range of motion for flexion, extension, lateral bending, and axial rotation. The pain experienced by the patients who had the external fixator applied was significantly reduced. For these reasons, it is essential to learn the effects of various surgical procedures on the stability of the spine. In particular, we need to consider when procedures may lead to increase in motion to a point leading to instability.

Much of the debate surrounding laminectomy and instability involves the use of fusion after the laminectomy. The possibility that fusion will be necessary to stabilize the spine after a laminectomy is largely case specific and depends on the purpose of the surgery. In a study by Goel et al. , the results did not indicate the presence of instability after a partial laminectomy (99).

The facets are particularly important because they contribute to strength and resist axial rotation and extension. Subsequently, facetectomies can potentially be linked to instability. Abumi et al. developed some conclusions regarding partial and total facetectomies (2). They found that, although it significantly increased the range of motion, a partial facetectomy of one or both facets at a single level did not cause spinal instability. However, the loss of a complete facet joint on one or both sides was found to contribute to instability. Total facetectomy produced an increase of 65% in flexion, 78% in extension, 15% in lateral bending, and 126% in axial rotation compared with intact motion. Goel et al. also found similar results regarding partial facetectomy (99). Another study indicated that facetectomy performed within animals resulted in a large decrease in motion *in vivo* even though the increase in range of motion occurred acutely (2).

Goel et al. reported a significant increase in the range of motion for all loading modes except extension when a total diskectomy was performed across L4-5 level (99). A significant, but smaller increase in range of motion for subtotal disk removal was also observed, however, the postoperative instability was minimal. Both partial and total diskectomies produced a significant amount of intervertebral translational instability in response to left lateral bending at the L3-L4 and L4-L5 levels. They attributed the one-sided instability to the combination of injuries to the annulus and the right capsular ligament. Studies have also shown that more significant changes to the motion of the spine occur with removal of the nucleus pulposus as opposed to the removal of the annulus (4d). Discectomy by fenestration and minimal resection of the lamina did not produce instability either.

BIOMECHANICS OF STABILIZATION PROCEDURES

Stability (or instability) retains a central role in the diagnosis and treatment of patients with back pain. Several studies have been carried out that help to clarify the foundation for understanding stability in the spine, as summarized above. In recent years, to restore stability across an abnormal segment, surgeons have well-accepted surgical stabilization and fusion of the spine using instrumentation, Figs. 12 and 13. The types and complexity of procedures (e.g., posterior, anterior, interbody) (100–105) have produced novel design challenges, requiring sophisticated testing protocols. In addition, most contemporary implant issues of stabilization and fusion of the spine are mostly mechanical in nature. [Biologic factors related to the adaptive nature of living tissue further complicate mechanical characterization (103,105)] Accordingly, it becomes essential to understand the biomechanical aspects of various spinal instrumentation and their effectiveness in stabilizing the segment. Properly applied spinal instrumentation maintains alignment and shares spinal loads until a solid, consolidated fusion is achieved. With few exceptions, these hardware systems are used in combination with bone grafting procedures, and may be augmented by external bracing systems.

Spinal implants typically follow loosely standardized testing sequelae during the design and development stage and in preparation for clinical use. The design and development phase goal, from a biomechanical standpoint, seeks to characterize and define the geometric considerations and load-bearing environment to which the implant will be subjected. Various testing modalities exist that elucidate which components may need to be redesigned. Not including the testing protocols for individual components of a device, plastic vertebrae (corpectomy) models are one of the first-stage tests that involves placing the assembled device on plastic vertebral components in an attempt to pinpoint which component of the assembled device may be the weakest mechanical link in the worst-case scenario, vertebrectomy. The *in vivo* effectiveness of the device may be limited by its attachment to the vertebras (fixation). Thus, testing of the implant-bone interface is critical in determining the fixation of the device to biologic tissue. Construct testing on cadaveric specimens provides information about the effectiveness of the device in reducing intervertebral motion across the affected and adjacent segments during quasiphysiologic loading. Animal studies provide insight with respect to the long-term biologic effects of implantation. Analytic modeling, such as the finite element method, is an extremely valuable tool for determining how implants and osseous loading patterns change with varying parameters of the device design. This type of modeling may also provide information about temporal changes in the bone quality due to the changing loading patterns as bone adapts to the implant (e.g., stress shielding-induced bone remodeling). After a certain level of confidence in the implant's safety and effectiveness is

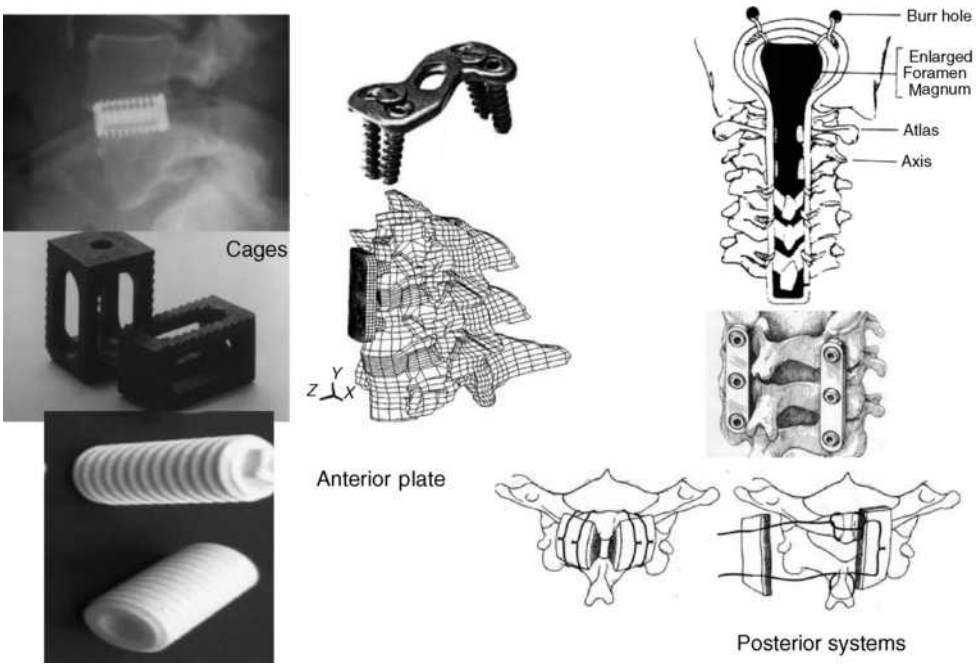


Figure 12. Devices used for stabilizing the cervical spine using the anterior and posterior approaches. Cages used both for the lumbar and cervical regions are also shown.

established via all or some of the aforementioned tests, controlled clinical trials allow for the determination of an implant's suitability for widespread clinical use. The following sections discuss each of these testing modalities, with specific examples used to illustrate the type of information that different tests can provide.

Implant–Bone Interface

Depending on the spinal instrumentation, the implant-bone interface may deal with the interface, where the spinal instrumentation abuts, encroaches, or invades the bone surface. It may include bony elements, such as the laminas, pedicles, the vertebral body itself, or the vertebral endplates.

Interlaminar Hooks. Interlaminar hooks are used as a means for fixing the device to the spine. Hook dislodgment, slippage, and incorrect placement have led to loss of fixation, however, resulting in nonfusion and pseudoarthrosis. Purcell et al. (106) investigated construct stiffness as a function of hook placement with respect to affected level in a thoracolumbar cadaver model. The failure moment was

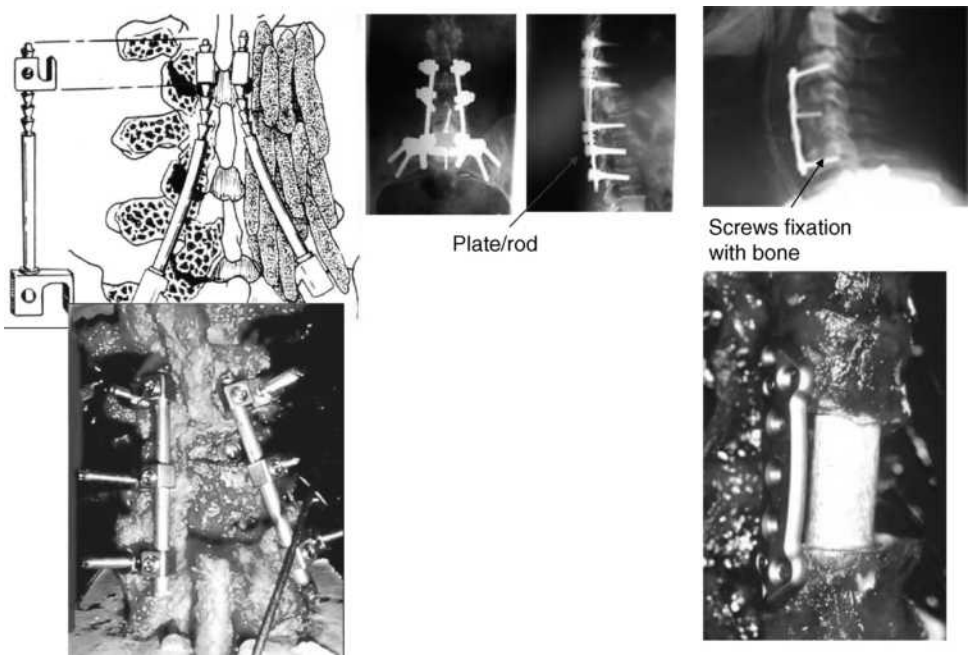


Figure 13. Examples of spinal instrumentation used in the lumbar region. Figure on the bottom right is an anterior plate.

found to be a function of the location of the hook placement with regard to the "injured" vertebra. The authors recommended hook placements three levels above and two levels below the affected area. This placement reduced vertebral tilting (analogous to intervertebral motion) across the stabilized segment, where fusion is to be promoted. Furthermore, the three-above, two-below surgical instrumentation strategy avoids the construct ending at the apex of a spinal deformity. Shortened fixation in this manner tends to augment a kyphotic deformity and cause continued progressive deformation. Overall, the use of hook fixation is a useful surgical stabilization procedure in patients with poor bone quality, where screw fixation is not an ideal choice for achieving adequate purchase into the bone.

Transpedicular Screws. Proper application of screw based anterior or posterior spinal devices requires an understanding of screw biomechanics, including screw characteristics and insertion techniques, as well as an understanding of bone quality, pedicle and vertebral body morphometries, and salvage options (107–109). This is best illustrated by the fact that the pedicle, rather than the vertebral body, contributes $\sim 80\%$ of the stiffness and $\sim 60\%$ of the pull out strength across the screw–bone interface (107).

Carlson et al. (110) evaluated the effects of screw orientation, instrumentation, and bone mineral density on screw translation, rotation at maximal load, and compliance of the screw–bone interface in human cadaveric bones. An inferiorly directed load was applied to each screw, inserted either anteromedially or anterolaterally, until failure of the fixation was perceived. Anteromedial screw placement with fully constrained loading linkages provided the stiffest fixation at low loads and sustained the highest maximal load. Larger rotation of the screws, an indication of screw pull out failure, was found with the semi-constrained screws at maximal load. Bone mineral density directly correlated with maximal load, indicating that bone quality is a major predictor of bone–screw interfacial strength. A significant correlation between BMD and insertional torque ($p < 0.0001$, $r = 0.42$), BMD and pullout force ($p < 0.0001$, $r = 0.54$), and torque and pullout force has been found (109–112).

Since the specimens used for pull-out strength studies primarily come from elderly subjects, Choi et al. used foams of varying densities to study the effect of bone mineral density on the pull out strength of several screws (112). Pedicle screws (6.0×40 mm, 2 mm pitch, Ti alloy) of several geometric variations used for the study included the buttress (B), square (S), and V-shape (V) screw tooth profiles. For each type of tooth profile, its core shape (i.e., minor diameter) also varied, either the straight (i.e., cylindrical, core diameter = 4.0 mm) or tapered (i.e., conical, core diameter = 4.0/2.0 mm). In addition, for the cylindrical screws the major diameter was kept straight or tapered. The conical screws had its major diameters tapered only. Therefore, screws with a total of nine different geometries were prepared and tested (Fig. 14a). The screws were implanted in the rigid polyurethane foams of three different grades. The pullout strengths for various screw designs are shown in Table 10. The highest purchasing power in any screw design was observed in foams with

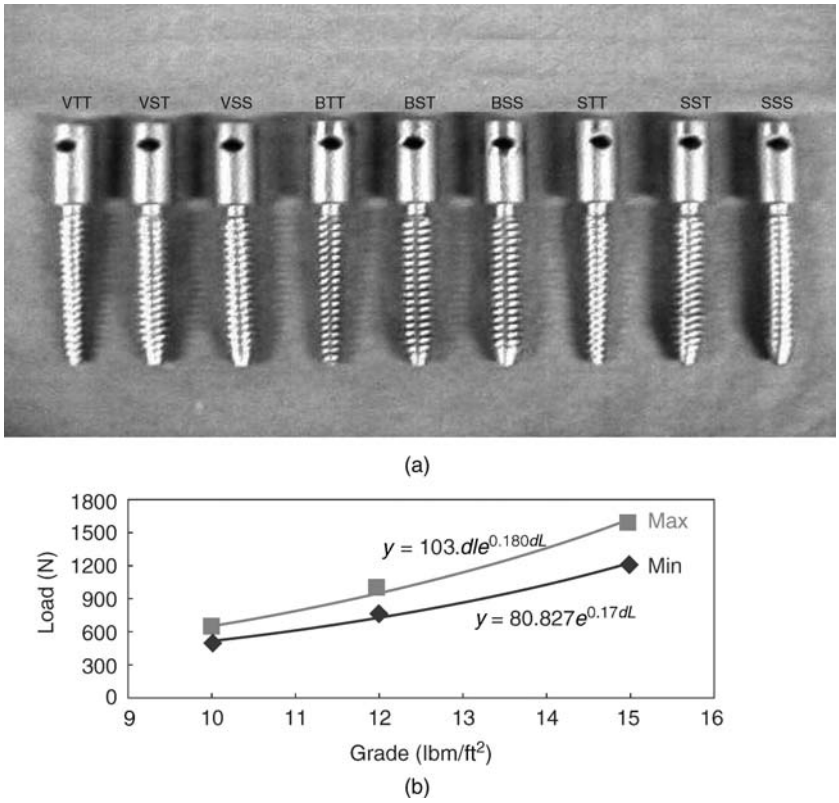


Figure 14. (a) Types of screws used in the foam model to determine the pull-out strength. The nomenclature used is as follows: Square = S, Buttress = B, V-shape = V. Screw diameters were SS = straight major diameter on straight core, ST = straight major diameter on tapered core, TT = tapered major diameter on tapered core. (b) Regression analysis. The maximum and minimum values from pull-out test for each foam grade were used regardless of tooth or core profiles. (Taken from Ref. 112.)

Table 10. Axial Pull Out Strength (N) Data for Different Types of Screws, Based on a Foam Model of Varying Densities

		Tooth Profile(Mean ± SD)		
Foam Grade	Body Profile	Square	Buttress	V-shape
10	SS[a]	591 ± 22	497 ± 80	615 ± 36
	ST[b]	622 ± 43	598 ± 25	634 ± 19
	TT[c]	525 ± 36	547 ± 30	568 ± 74
12	SS	864 ± 50	769 ± 56	987 ± 55
	ST	956 ± 30	825 ± 108	1005 ± 92
	TT	811 ± 41	808 ± 25	944 ± 32
15	SS	1397 ± 93	1303 ± 126	1516 ± 78
	ST	1582 ± 82	1438 ± 36	1569 ± 79
	TT	1197 ± 43	1352 ± 88	1396 ± 68

[a]Straight major diameter on Straight core.
[b]Straight major diameter on Tapered core.
[c]Tapered major diameter on Tapered core. (Taken from Ref. 112.)

the highest density (Grade 15). Exponential increase in pullout strength was seen when the foam density increased from Grade 10–15 (Fig. 14b). Overall, results demonstrated that the conical screws were consistently more effective against the pullout than the cylindrical designs. This was especially evident when the major diameter of the screw was kept straight. In this case, the contact area between the screw thread and surrounding foam was large. Although no consistent statistical superiority was found with the tooth profiles, results did suggest that the V-shape tooth screws ranked highest in many statistical comparisons and the buttress types showed comparatively lower pullout strength than the other types. This finding may be somewhat different from the literature. This can be due to the absence of the cortical purchase in the foam model used in this study. On the other hand, the square-tooth screws faired well in terms of pullout strength when the major diameter was kept straight but did not do so when tapered. Results also suggested that as the density of host site was decreased no clear choice of tooth profile could be found.

Lim et al. investigated the relationship between the bone mineral density of the vertebral body and the number of loading cycles to induce loosening of an anterior vertebral screw (113). (Screw loosening was defined as 1 mm displacement of the screw relative to bone). The average number of loading cycles to induce screw loosening was significantly less for specimens with bone mineral density $< 0.45\ \mathrm{g \cdot cm^{-2}}$, compared to those with bone mineral density $>$ or $= 0.45\ \mathrm{g \cdot cm^{-2}}$. These findings suggest that bone mineral density may be a good predictor of anterior vertebral screw loosening as well, just like the pedicle screws.

Since BMD seems to play a crucial role in the loosening of fixation screws, their use with osteoporotic bone is a contraindication. Alternatives have been proposed, including the use of bone cement to augment fixation and use of hooks along with pedicle screws (114,115).

The above findings related to increased pullout strength, number of cycles to failure, and tightening torque with BMD, are not fully corroborated with the corresponding *in vivo* work. For example, moments and forces during pedicle screw insertion were measured *in vivo* and *in vitro* and correlated to bone mineral density, pedicle size, and other screw parameters (material, diameter) (116). The mean *in vivo* insertion torque (1.29 N·m) was significantly greater than the *in vitro* value (0.67 N·m). The linear correlation between insertion torque and bone mineral density was significant for the *in vitro* data, but not for the *in vivo* data. No correlation was observed between insertion torque and pedicle diameter. However, another investigation that clinically evaluated 52 patients who underwent pedicle screw fixation augmenting posterior lumbar interbody fusion (PLIF) supports the *in vitro* findings. The BMD was measured using DEXA and radiographs were assessed for detecting loosening, and so on at the screw bone interface. Bone mineral density was found to have a close relation with the stability of pedicle screw *in vivo*, and BMD values $< 0.674 \pm 0.104\ \mathrm{g \cdot cm^{-2}}$ suggested a potential increased risk of "non-union".

Cages. Total disk removal alone or in combination with other surgical procedures invariably leads to a loss of disk height and an unstable segment. Both allo- and autologous bone grafts have been used as interbody spacers (103,117–120). Associated with the harvest and use of autogenous bone grafts are several complications: pain, dislodgment of the anterior bone graft, loss of alignment, and so on. Recently, the use of inserts, fabricated from synthetic materials (metal or bone-biologic), have gained popularity. These may be implanted through an anterior or posterior approach. Interbody devices promote fusion by imparting immediate postoperative stability, and by providing axial load-bearing characteristics, while allowing long-term fusion incorporation of the bone chips packed inside and around the cage (121,122). Many factors influence the performance of an interbody cage. The geometry, porosity, elastic modulus, and ultimate strength of the cage is crucial to achieving a successful fusion. An ideal fixation scenario should be to utilize the largest cross-sectional footprint of a cage in the interbody space so that the cortical margin can be captured by the fixation to decrease the risk of endplate subsidence. A modulus of elasticity close to bone is often an ideal choice to balance the mechanical integrity at the endplate–implant interface. A cage that has a large elastic modulus and high ultimate strength increases the risk to endplate subsidence and/or stress-shielding issues. Finally, cage design must possess a balance between an ideal porosity to augment bony fusion through the cage and mechanical strength to bear axial loads.

Steffen et al. undertook a human cadaveric study with the objectives to assess the axial compressive strength of an implant with peripheral endplate contact as opposed to full surface contact, and to assess whether removal of the central bony endplate affects the axial compressive strength (120). Neither endplate contact region nor its preparation technique affected yield strength or ultimate compressive strength. Age, bone mineral content, and the normalized endplate coverage were strong predictors of yield strength ($P < 0.0001$; $r^2 = 0.459$) and ultimate compressive strength ($P < 0.0001$; $r^2 = 0.510$). An implant with only peripheral support resting on the apophyseal ring offers axial mechanical strength similar to that of an implant with full support. Neither supplementary struts nor a solid implant face has any additional mechanical advantage, but reduces graft–host contact area. Removal of the central bony endplate is recommended because it does not affect the compressive strength and promotes graft incorporation. There are drawbacks to using threaded cylindrical cages (e.g., limited area for bone ingrowth, subsidence issues, and metal precluding radiographic visualization of bone healing). To somewhat offset these drawbacks, several modifications have been proposed, including changes in shape and material (123–125). For example, the central core of the barbell shaped cage can be wrapped with collagen sheets infiltrated with bone morphogenetic protein. The femoral ring allograft (FRA) and posterior lumbar interbody fusion (PLIF) spacers have been developed as biological cages that permit restoration of the anterior column with a machined allograft bone (123).

Wang et al. (126) looked at *in vitro* load transfer across standard tricortical grafts, reverse tricortical grafts, and fibula grafts, in the absence of additional stabilization. Using pressure sensitive film to record force levels on the graft, the authors found the greatest load on the graft occurred in flexion. As expected, the anterior portion of the graft bore increased load in flexion and the posterior portion of the graft bore the higher loads in extension. The authors did not supplement the anterior grafting with an anterior plate. Cheng et al. (127) performed an *in vitro* study to determine load sharing characteristics of two anterior cervical plate systems under axial compressive loads: the Aesculap system (Aesculap AGT, Tuttlingen, Germany) and the CerviLock system (SpineTech Inc., Minneapolis, MN). The percent loads carried by the plates at a 45 N applied axial load were as follows: Aesculap system $-6.2\% \pm 9.2\%$ and the CerviLock system $-23.8\% \pm 12.7\%$. Application of 90 N loads produced similar results to those of the 45 N loads. The authors stated that the primary factor in load transfer characteristics of the instrumented spine was a difference in plate designs. The study contained several limitations. Loading was performed solely in axial compression across a single functional spinal unit (FSU). The study did not simulate complex loading, such as flexion combined with compression. In the physiologic environment, load sharing in multisegmental cervical spine could be altered since the axial compressive load will produce additional flexion–extension moments, due to the lordosis. The upper and lower vertebrae of the FSU tested were constrained in the load frame, whereas in reality they are free to move, subject to anatomic constraints.

Rapoff et al. (128) recently observed load sharing in an anterior CSLP plate fixed to a three level bovine cadaveric spinal mid-thoracic segment under simple compression of 125 N. A Smith–Robinson diskectomy procedure was performed at the median disk space to a maximum distraction of 2 mm prior to plate insertion and loading. Results showed that at 55 N of load, mean graft load sharing was 53% ($\pm 23\%$) and the plate load sharing was 57% ($\pm 23\%$). This study was limited in several aspects, including the fact that no direct measurement of plate load was made, the spines were not human, and the loading mode was simplified and did not incorporate more complex physiologic motions, such as coupled rotation and bending or flexion/extension.

A recent study by An et al. (129) looking at the effect of endplate thickness, endplate holes, and BMD on the strength of the graft–endplate interphase of the cervical spine found that there existed a strong relationship between BMD and load to failure of the vertebrae, demonstrating implications for patient selection and choice of surgical technique. There was a significantly larger load to failure in the endplate intact group compared to the endplate resected group studied, suggesting that an intact endplate may be a significant factor in prevention of graft subsidence into the endplate. Results of an FE model observing hole patterns in the endplate indicated that the hole pattern only significantly affected the fraction of the upper endplate that was exposed to fracture stresses at 110 N loading. A large central hole was found to be best for minimization of fracture area and more effective at distribution of the compressive load across the endplate area.

Dietl et al. pulled out cylindrical threaded cages (Ray TFC Surgical Dynamics), bullet-shaped cages (Stryker), and newly designed rectangular titanium cages with an endplate anchorage device (Marquardt) used as posterior interbody implants (130). The Stryker cages required a median pullout force of 130 N (minimum, 100 N; maximum, 220 N), as compared with the higher pullout force of the Marquardt cages (median, 605 N; minimum, 450 N; maximum, 680 N), and the Ray cages (median, 945 N; minimum, 125 N; maximum, 2230 N). Differences in pullout resistance were noted depending on the cage design. A cage design with threads or a hook device provided superior stability, as compared with ridges. The pyramid shaped teeth on the surfaces and the geometry of the implant increased the resistance to expulsion at clinically relevant loads (1053 and 1236 N) (124,125).

Construct Testing

Spinal instrumentation needs to be applied to a spine specimen to evaluate its effectiveness. As a highly simplified model, two plastic vertebras serve as the spine model. Loads are applied to the plastic vertebras and their motions and applied loads to failure are measured. This gives some idea of the rigidity of the instrumentation. However, a truer picture is obtained by attaching the device to the cadaveric spine specimen.

Plastic Vertebra (Corpectomy) Models. Clinical reviews of failure modes of the devices indicate that most designs satisfactorily operate in the immediate postoperative period. Over time, however, these designs can fail because of the repeated loading environment to which they are subjected. Thus, fatigue testing of newer designs has become an extremely important indicator of long-term implant survivorship. Several authors have tested thoracolumbar instrumentation systems in static and fatigue modes using a plastic vertebral model (131–133). For example, Cunningham et al. compared 12 anterior instrumentation systems, consisting of 5 plate and 7 rod systems in terms of stiffness, bending strength, and cycles to failure (132). The stiffness ranged from 280.5 $kN \cdot m^{-1}$ in the Synthes plate (Synthes, Paoli, PA) to 67.9 $kN \cdot m^{-1}$ in the Z-plate (Sofamor-Danek, Memphis, TN). The Synthes plate and Kaneda SR titanium (AcroMed, Cleveland, OH) formed the highest subset in bending strength of 1516.1 and 1209.9 N, respectively, whereas the Z plate showed the lowest value of 407.3 N. There were no substantial differences between plate and rod devices. In fatigue, only three systems: Synthes plate, Kaneda SR titanium, and Olerud plate (Nord Opedic AB, Sweden) withstood 2 million cycles at 600 N. The failure mode analysis demonstrated plate or bolt fractures in plate systems and rod fractures in rod systems.

Clearly, studies, such as these involving missing vertebral (corpectomy) artificial models, reveal the weakest components or linkages of a given system. Results must be viewed with caution since these results do not shed light on the biomechanical performance of the device. Furthermore, we do not know the optimum strength of a fixation system. These protocols do not provide any information about the effects the device implantation may have on individual spinal components found *in vivo*. For these data, osteoligamentous cadaver models need to be incorporated in the testing sequelae and such studies are more clinically relevant.

Osteoligamentous Cadaver Models. For applications, such as fusion and stabilization, initial reductions in intervertebral motion are the primary determinants of instrumentation success, although the optimal values for such reductions are not known and probably not needed to determine relative effectiveness. Thus, describing changes in motion of the injured and stabilized segments in response to physiologic loads is the goal of most cadaver studies. Many times, these data are compared with the intact specimen, and the results are reported as the instrumentation's contribution to providing stability (134). To standardize, the flexibility testing protocol has been suggested (135). Here a load is applied and resulting unconstrained motions are measured. However, there are several issues pertaining to this type of testing, as described below.

More recently nonfusion devices have come on the market. These devices try to restore motion of the involved segment. With the paradigm shift from spinal fusion to spinal motion, there are dramatically different criteria to be considered in the evaluation of nonfusion devices. While fusion devices need to function for a short period and are differentiated primarily by their ability to provide rigid fixation, nonfusion devices must function for much longer time periods and need to provide spinal motion, functional stability, and tolerable facet loads. The classic flexibility testing protocol is not appropriate for the understanding of the biomechanics of the construct for the nonfusion devices, at the adjacent levels (136,137). However, constant pure moments are not appropriate for measuring effects of implants, like the total disk replacements, at adjacent levels. The pure moments distribute evenly down a column and are thus not effected by perturbation at a level(s) in a longer construct. Further, the net motion of a longer construct is not similar if only pure moments are applied: fusions will limit motion and other interventions may increase motion, a reflection of the change in stiffness of the segment. This may have shortcomings for clinical applications. For example, with forward flexion, there are clinical demands to get to ones shoes to tie them, to reach a piece of paper fallen to the floor, and so on. It would thus be advantageous to use a protocol that would achieve the same overall range of motion for the intact specimen and instrumented construct by applying pure moments that distribute evenly down the column.

Another issue is that the ligamentous specimens cannot tolerate axial compressive loads, specimens in the absence of the muscles will buckle. Thus, methods have been developed to apply preloads on the ligamentous spines during testing, since these indirectly simulate the effects of muscles on the specimens. A number of approaches have been proposed with one that stands out and is getting accepted by the research community. It is termed the follower-load concept (137).

It could be reasoned that coactivation of trunk muscles (e.g., the lumbar multifidus, longissimus pars lumborum, iliocostalis pars lumborum) could alter the direction of the internal compressive force vector such that its path followed the lordotic and kyphotic curves of the spine, passing through the instantaneous center of rotation of each segment. This would minimize the segmental bending moments and shear forces induced by the compressive load, thereby allowing the ligamentous spine to support loads that would otherwise cause buckling and providing a greater margin of safety against both instability and tissue injury. The load vector described above is called a "*follower load*".

Additionally, most of these studies involve quasistatic loading; however, short-term fatigue characteristics have also been investigated. Both posterior and anterior-instrumentation employed for the promotion of fusion and non fusioon have been evaluated. The following are examples of such devices, which are diskussed within the context of these testing modalities.

Cervical Spine Stabilization and Fusion Procedures

There are a variety of techniques that are utilized for spinal fusion in the lower cervical spine, among which are spinal wiring techniques (138–144), posterior plating (145–154), anterior plating, and (more recently) cervical interbody fusion devices. While fusions are effective in a majority of cases, they do have documented biomechanical shortcomings, particularly at the segments adjacent to the

fusion. Some of these problems include observations of excessive motion (sometimes due to pseudoarthrosis) (155–164), degenerative changes (165,166), fracture dislocation (167), screw breakage or plate pullout (160,168–170), and risks to neural structures. These problems are typically minimal when only one or two segments are involved in the injury. However, when the number of segments involved in the reconstruction increases to three or more, the incidence of failed fusion, screw breakage, and plate pullout increases dramatically.

Upper Cervical Spine Stabilization. Stabilization of the craniovertebral junction is not common; however, its importance for treating rheumatoid arthritis associated lesions, fractures and tumors cannot be underestimated. Currier et al. (171) studied the degree of stability provided by a rod-based instrumentation system. They compared this new device to the Ransford loop technique and a plate system using C2 pedicle screws. Transverse and alar ligament sectioning and odontoidectomy destabilized the specimen. All three-fixation systems significantly reduced motion as compared to intact and injured spines in axial rotation and extension. The new device did not significantly reduce motion at C1-C2 in flexion, and none of the devices were able to produce significant motion reductions in C1-C2 lateral bending. The authors claimed, based on these findings, that the new system is equivalent or superior to the other two systems for obtaining occipitocervical stability. Oda et al. (172) investigated the comparative stability afforded by five different fixation systems. Type II odontoid fractures were created to simulate instability. The results indicate that the imposed dens fracture decreased construct stiffness as compared to the intact case. Overall, the techniques that utilized screws for cervical anchors provided greater stiffness than the wiring techniques. Also, the system that utilized occipital screws with C2 pedicle screw fixation demonstrated the greatest construct stiffness for all rotations. Puttlitz et al. (1c) have used the finite element model of the C0-C1-C2 complex to investigate the biomechanics of a novel hardware system (Fig. 15). The FE models representing combinations of cervical anchor type (C1-C2 transarticular screws versus C2 pedicle screws) and unilateral versus bilateral instrumentation were evaluated. All models were subjected to compression with pure moments in flexion, extension, or lateral bending. Bilateral instrumentation provided greater motion reductions than the unilateral hardware. When used bilaterally, C2 pedicle screws approximate the kinematic reductions and hardware stresses (except in lateral bending) that are seen with C1-C2 transarticular screws. The FE model predicted that the maximum stress was always located in the region where the plate transformed into the rod. Thus, the authors felt that C2 pedicle screws should be considered as an alternative to C2-Cl transarticular screw usage when bilateral instrumentation is applied.

Other strategies to fix the atlantoaxial complex can be found in the literature. Commonly available fixation techniques to stabilize the atlantoaxial complex are posterior wiring procedures (Brooks fusion, Gallie fusion) (169), interlaminar clamps (Halifax) (170), and transarticular screw (Magerl technique), either alone or in combination.

Posterior wiring procedures and interlaminar clamps are obviously easier to accomplish. However, these do not provide sufficient immobilization across the atlantoaxial complex. In particular, posterior wiring procedures and place the patient at risk of spinal cord injury due to sublaminar passage of wires into the spinal canal (172). Interlaminar clamps offer the advantage of avoiding the sublaminar wire hazard and have more rigid biomechanical stiffness than posterior wiring procedures (173).

Transarticular screw fixation (TSF), on the other hand, affords a stiffer atlantoaxial arthrodesis than posterior wiring procedures and interlaminar clamps. The TSF does have some drawbacks including injury of vertebral artery, malposition, and screw breakage (174). Furthermore, body habitus (obesity or thoracic hyperkyphosis) may prohibit achieving the low angle needed for screw placement across C1 and C2. Recently, a new technique of screw and rod fixation (SRF) that minimizes the risk of injury to the vertebral artery and allows intraoperative reduction has been reported (175,176). The configuration of this technique, which achieves rigid fixation of the atlantoaxial

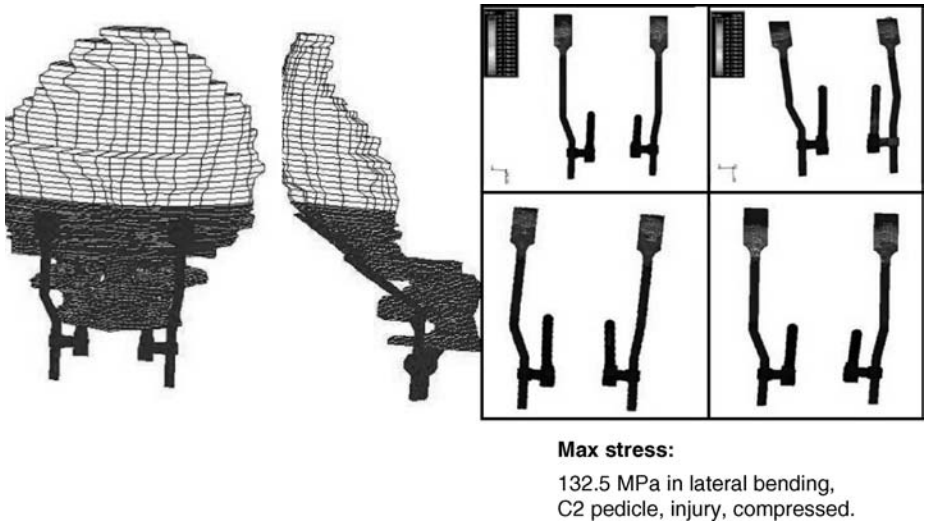


Figure 15. The finite elment model showing the posterior fixation system and the stress plots in the rods. (Taken from Ref. 4c.)

complex, consists of lateral mass screws at C1 and pedicle screws at C2 linked via longitudinal rods with constrained coupling devices.

One recent study compared the biomechanical stability impaired to the atlantoaxial complex by either the TSF or SRF technique and to assess how well these methods withstand fatigue in a cadaver model (177).

The results of this study suggested that in the unilateral fixations, the SRF group was stiffer than the TSF group in flexion loading, but there were no evident differences in other directions. In the bilateral fixations, SRF was more stable than TSF, especially in flexion and extension. These results were similar to those reported by Melcher et al. (178) and Richter et al. (179), yet different from Lynch et al. (180). The instrumentation procedure (screw length, type of constrained coupling device, etc.), the destabilization technique, and the condition of the specimens might have an influence on the results. In this study, when stabilizing the atlantoaxial segments, all screws were placed bicortically in both techniques in accordance with procedures by Harms and Melcher (181). Previous work has demonstrated that bicortical cervical vertical screws are superior to unicortical screws in terms of pullout strength and decreased wobble (182,183). Most surgeons, however, prefer unicortical screwing at C1 and C2 levels to reduce the risk of penetration during surgery. This could affect the outcome. They initially connected the screw to the rod using the oval shape constrained coupling device recommended for use in C1 and C2 vertebras. However, the stability was not judged adequate, So they altered the procedure to use the stiffer circle shape constrained coupling device. With regards to the destabilization procedure, there are three typical methods: sectioning of intact ligaments, odontoid fracture, and odontoidectomy. The atlantoaxial complex was destabilized by ligament transection to simulate ligamentous instability, while Lynch et al. (180) used odontoidectomy. Furthermore, the bone quality of specimens affects the screw-bone interface stability. These factors were possibly reflected in other results. However, both results were not statistically different between TSF and SRF, so they could be interpreted equivalent in terms of effective stabilization when compared with the intact specimen.

In unilateral TSF and SRF, the fixed left lateral atlantoaxial joint acted as a pivot in left axial rotation and as a fulcrum in left lateral bending, thus leading to an increase in motion. This motion could be observed with the naked eye.

Stability in flexion and extension of the bilateral TSF group was inferior to that of SRF group. Henriques et al. (184) and Naderi et al. (182) also reported similar tendency. Henriques et al. (184) felt that this was most likely due to the transarticular screws being placed near the center of motion between C1 and C2. This was judged as another reason that the trajectory of the screws is consistent with the motion direction of flexion and extension. So, if TSF is combined with some posterior wiring procedures, the stability in flexion and extension will increase.

Lower Cervical Spine

Anterior Plating Techniques for Fusion. The anterior approach in order to achieve arthrodesis of the cervical spine has become a widely utilized and accepted approach. However, many of these techniques rely on insertion of a bone graft only anteriorly and the use of an external immobilization device, such as a halo vest, or posterior fixation in order to allow for sufficient fixation. Problems encountered with these methods include dislodging of the bone graft (potentially causing neural compromise), loss of angular correction, and failure to maintain spinal reduction (185,186). The use of anterior plates has recently become popular partially because they address some of the complications stated above. The main reasons typically cited for the use of anterior plates are (*1*) advantage of simultaneous neural decompression via an anterior as opposed to posterior approach, (*2*) improved fusion rates associated with anterior cervical fusion (187,188), (*3*) help in reduction of spinal deformities, (*4*) provides for rigid segmental fixation, and (*5*) prevents bone graft migration. However, the efficacy of anterior plates alone is still debated by some authors, particularly in multilevel reconstruction techniques, due to the high rates of failure observed, up to 50% in some cases (189–192). Thus, more biomechanical research must be accomplished to delineate the contributions of anterior plates to load sharing mechanics in the anterior approach.

There have been several *in vitro* studies examining the efficacy of anterior plates for use in a multitude of procedures involving cervical spine stabilization. Grubb et al. (151) performed a study involving 45 porcine and 12 cadaveric specimens to study anterior plate fixation. Phase I of the study involved intact porcine specimens which were subjected to nondestructive testing in flexion, lateral bending, and axial rotation loading modes to determine structural stiffness. Maximum moments applied included 2.7 N·m for flexion and lateral bending and 3.0 N·m for axial rotation testing. After completion of the nondestructive testing, a flexion-compression injury was introduced by performing a C5 corpectomy and inserting an iliac strut bone graft in the resulting space. An anterior plate was then introduced across C4–C6. Three different anterior plates were tested, including a Synthes CSLP (cervical spine locking plate) with unicortical fixation, a Caspar plate with unicortical fixation, and a Caspar plate with bicortical fixation. Each instrumented specimen was then tested again nondestructively in flexion, lateral bending, and axial rotation. Finally, destructive testing in each loading mode was performed on particular specimens in each plated group. Phase II of the study involved intact cadaver specimens that were subjected to nondestructive testing in flexion, lateral bending, and axial rotation loading modes to determine structural stiffness. Maximum moments applied included 2.0 N·m for flexion, lateral bending, and axial rotation. After completion of the nondestructive testing, a flexion-compression injury was introduced by performing a C5 corpectomy and inserting an iliac strut bone graft in the resulting space. An anterior plate was then introduced across C4–C6. Two different anterior plates were tested: a Synthes CSLP (cervical spine locking plate) with unicortical fixation and a Caspar plate with bicortical fixation. Each instrumented specimen was then tested again nondestructively in flexion, lateral bending, and axial rotation. Finally, destructive testing in flexion

was performed on each specimen. Results of the study demonstrated that each of the stabilized specimens had stiffness characteristics greater than or equal to their paired intact test results. The CSLP was found to have a significantly higher stiffness ratio (plated: intact), higher failure moment, lower flexion neutral zone ratio, and higher energy to failure than the Caspar plates.

A study by Clausen et al. (193) reported the results of biomechanical testing of both the CSLP system with unicortical locking screws and Caspar plate system with unlocked bicortical screws. Fifteen cadaveric human spines were tested intact in flexion, extension, lateral bending, and axial rotation loading modes to determine stiffness characteristics. A C5-C6 instability was then introduced, consisting of a C5-C6 diskectomy with complete posterior longitudinal ligament (PLL) disruption. An iliac crest bone graft was then introduced into the C5-C6 disk space and the spine was instrumented with either the CSLP or Caspar system. Once instrumented, each of the spines were further destabilized through disruption of the interspinous and supraspinous ligaments, the *ligamentum flavum*, facet capsules, and lateral annulus. The specimens were then retested for stiffness. After initial postinstrumented testing was done, biomechanical stability of the specimens was reassessed following cyclic fatigue for 5000 cycles of flexion–extension. Finally, failure testing of each specimen was performed in flexion. Results of the study demonstrated that both devices stabilized the spine before but not after fatigue and that only the Caspar plate stabilized the spine significantly before and after fatigue. Failure moment did not differ between the two systems. Biomechanical stability discrepancy between the two devices was attributed to differences in bone–screw fixation. Kinematic testing of 10 cervical spines following single level (C5-C6) diskectomy and anterior plate insertion was studied by Schulte et al. (194). Results showed that the use of an anterior plate in addition to the bone graft provided significant stabilization in all loading modes. Traynelis et al. (195) performed biomechanical testing to compare anterior plating versus posterior wiring in an cadaver instability model involving a simulated C5 teardrop fracture with posterior disruption and fixation across C4-C6. Study results showed that bicortical anterior plating provided significantly more stability than posterior wiring in extension and lateral bending, and was slightly more stable than posterior wiring in flexion. Both provided equivalent stability in axial rotation. A variety of anterior constructs exist in the market today, typically using either bicortical screws or unicortical locking screws. Several studies have evaluated the purchase of unicortical versus bicortical screws in the cervical spine (195–197).

Wang et al. (198) looked at *in vitro* load transfer across standard tricortical grafts, reverse tricortical grafts, and fibula grafts, in the absence of additional stabilization. Using pressure sensitive film to record force levels on the graft, the authors found the greatest load on the graft occurred in 10° of flexion (~ 20.5 N) with a preload on the spine of 44 N. As expected, the anterior portion of the graft bore increased loading in flexion and the posterior portion of the graft bore the highest loads in 10° extension. The authors did not supplement the anterior grafting with an anterior plate. Cheng et al. (127) performed an *in vitro* study to determine load-sharing characteristics of two anterior cervical plate systems under axial compressive loads: the Aesculap system (Aesculap AGT, Tuttlingen, Germany) and the CerviLock system (SpineTech Inc., Minneapolis, MN). The percent loads carried by the plates at a 45 N applied axial load were as follows: Aesulap system $-6.2\% \pm 9.2\%$ and the CerviLock system $-23.8\% \pm 12.7\%$. Application of 90 N loads produced similar results to those of the 45 N loads. The authors stated that the primary factor in load transfer characteristics of the instrumented spine was a difference in plate designs. The study contained several limitations. Loading was performed solely in axial compression across a single FSU. The study did not simulate complex loading, such as flexion combined with compression. In the physiologic environment, load sharing in multisegmental cervical spine could be altered since the axial compressive load will produce additional flexion–extension moments, due to the lordosis. The upper and lower vertebras of the FSU tested were constrained in the load frame, whereas in reality they are free to move, subject to anatomic constraints. Foley et al. also performed *in vitro* experiments to examine the loading mechanics of multilevel strut grafts with anterior plate augmentation (199). The results of the study showed that application of an anterior plate in a cadaver corpectomy model unloads the graft in flexion and increases the loads borne by the graft under extension of the spine. The increase in load borne by the graft in the presence of the plate should increase the graft subsidence, a finding that is contrary to clinical follow-up studies, as stated earlier.

Finite element (FE) analysis has been used by our group on a C5-C6 motion segment model to determine load sharing in an intact spine under compressive loading and more clinically relevant combined loading of flexion–extension and compression (4b). Similarly, using the FE approach, stresses in various graft materials (titanium core, titanium cage, iliac crest, tantalum core, and tantalum cage), the adjacent disk space, and vertebra have been investigated by Kumareson et al. (200). These authors found that angular stiffness decreased with decreasing graft material stiffness in flexion, extension, and lateral bending. They also observed the stress levels in the disk and vertebral bodies as a whole due to the presence of a graft, but did not focus on the graft itself or the endplate regions, superior and inferior to the graft. The effects of anterior plates on load sharing were not investigated.

Scifert et al. (4d) developed an experimentally validated C4-C6 cervical spine finite element model was developed to examine stress levels and load sharing characteristics in an anterior plate and graft. Model predictions demonstrated good agreement with the *in vitro* data. The rotations across the stabilized segment significantly decreased in the presence of a plate as compared to graft alone case. Much like the *in vitro* studies, the model also predicted that the compressive load in the graft increased in extension in the presence of plate, as compared to graft alone case. Depending on the load type, stresses in graft were concentrated in its anterior or posterior region in the graft alone case and became more uniformly distributed in the presence of the plate. The predicted load-displacement data

and load sharing results reveal that plate is very effective in maintaining the alignment. Increase in load borne by the graft in the presence of a plate in the extension mode suggests that pistoning of the graft is a possible outcome. However, the stress data reported in the present study, and something that the *in vitro* studies are unable to quantify, show that pistoning of the graft is not likely to happen due to stresses being low, an observation in agreement with the clinical outcome data. For an optimal healing, the stress results suggest the placement of the tricortical bone graft with its cortical region towards the canal when a plate is used. For the graft case alone, this parameter does not seem to be that critical. A more uniform stress distribution in the graft in the presence of the plate would tend to promote bone fusion in a more uniform fashion, as compared to the graft alone case. In the later case fusion may initiate in a selective region.

Lower Cervical Spine

Posterior Plating Techniques for Fusion. The posterior approach in order to achieve cervical spine arthrodesis has been a widely utilized and accepted approach to dealing with cervical spine trauma, such as posterior trauma involving the spinous processes or facet dislocation or injury, and disease, such as degenerative spondylosis or ossification of the posterior longitudinal ligament. Recently, however, posterior fixation using cervical screw plates affixed to the lateral masses has gained acceptance due to a variety of factors, including the fact that they do not rely on the integrity of the lamina or spinous processes to allow for fixation, bone grafting is not always necessary to allow for long-term stability, greater rotational stability is achieved at the facets (201,202), and it eliminates the need for external immobilization such as halo vests. Problems encountered with these posterior methods include (*1*) risk to nerve roots, vertebral arteries, facets, and spinal cord (168); (*2*) screw loosening and avulsion (203); (*3*) plate breakage; (*4*) and loss of reduction. Additionally, contraindications exist where the patient has osteoporosis, metabolic bone disease, or conditions where the bone is soft (i.e., ankylosing spondylitis) (204). There also exists controversy as to the advantages of using posterior plating techniques when posterior cervical wiring techniques can be used (205). In theory, anterior stabilization of the spine in cases of vertebral body injury is superior to posterior plating. However, in practice, posterior plates are an effective means of stabilizing vertebral body injuries, and their application is easier than the anterior approach involving corpectomy, grafting, and anterior plating.

In addition to clinical *in vivo* studies, there have been several *in vitro* studies examining the efficacy of posterior plates for use in cervical spine stabilization. Roy-Camille et al. (202) utilized a cadaveric model to compare posterior lateral mass plating to spinous process wiring. They found that posterior plates increased stability by 92% in flexion and 60% in extension, while spinous process wiring enhanced flexion stability by only 33% and did not stabilize in extension at all Coe et al. (201) performed biomechanical testing of several fixation devices, including Roy-Camille posterior plates, on six human cadaveric spines. Complete disruption of the supraspinous and interspinous ligaments, *ligamentum flavum*, posterior longitudinal ligament, and facet joints was performed. They found no significant difference in static or cyclic loading results between the posterior wiring and posterior plates, although the posterior plating was stiffer in torsion. Overall, the authors recommended the Bohlmann triple wire technique for most flexion distraction injuries. In experimental studies performed in our lab on 12 cervical spines, Scifert et al. (202) found that posterior plates were superior to posterior facet wiring in almost every loading mode tested in both the stabilized and cyclic fatigue testing modes, excluding the cyclic extension case. Smith et al. (153) performed biomechanical tests on 22 spines to evaluate the efficacy of Roy-Camille plates in stabilization of the cervical spine following simulation of a severe fracture dislocation with three-column involvement caused by forced flexion-rotation of the head. Results of the study indicated that the posterior plating system decreased motion significantly compared to the intact spine, specifically by a factor of 17 in flexion–extension and a factor of 5 units in torsion. Raftopoulos et al. (203) found that both posterior wiring and posterior plating resulted in significant stability following severe spinal destabilization, although posterior plating provided superior stability compared to that of interfacet wiring. Similar to the results of anterior plates, Gill et al. (204) found that bicortical lateral posterior plate screw fixation provided greater stability than unicortical fixation. However, Grubb et al. (151) found that unicortical fixation of a destabilized spine using a cervical rod device provided equivalent stability in torsion and lateral bending as bicortical fixation using an AO lateral mass plate. Effectiveness of 360° plating techniques for fusion.

As stated previously, both anterior and posterior plating procedures contain inherent difficulties and drawbacks. Some authors have examined the utilization of both techniques concomitantly to ensure adequate stabilization. Lim et al. (205) examined both anterior only, posterior only, and combined techniques *in vitro* to determine efficacy of these techniques in stabilizing either a C4-C5 flexion-distraction injury or an injury simulating a C5 burst fracture involving a C5 corpectomy. The AXIS and Orion plates were used for posterior and anterior stabilization, respectively. In the C4-C5 flexion-distraction injury, both posterior and combined fixation reduced motion significantly from intact in flexion. Only the combined procedure was able to reduce motion effectively in extension. In lateral bending and axial rotation, posterior fixation alone and combined fixation were able to significantly reduce motion compared to intact. In the C5 corpectomy model, all constructs exhibited significantly less motion compared to intact in flexion, although the combined fixation was the most rigid. In extension, all constructs except the posterior fixation with bone graft were able to reduce motion significantly compared to intact. In lateral bending, only the posterior fixation and combined fixation were able to provide enhanced stability compared to intact. In axial rotation, only the combined fixation was able to significantly reduce motion compared to intact. Thus, the authors concluded that combined fixation provided the most rigid

stability for both surgical cases tested. In a clinical study of multilevel anterior cervical reconstruction surgical techniques, Doh et al. (190) found a 0% psuedoarthrosis rate for the combined fixation system only.

Although combined fixation almost certainly allows for the most rigid fixation in most unstable cervical spine injuries, there are other factors to consider, such as the necessity for an additional surgery, possibility of severely reduced range of motion, and neck pain, Jonsson et al. (206) found a propensity for 22 out of 26 patients with combined fixation to have pain related to the posterior surgery. Additionally, patients with the combined fixation were found to have considerably restricted motion compared to normal. These and other factors must be weighed with the additional advantages of almost assured stability with the combined fixations.

Interbody Fusion Cage Stabilization for Fusion

Interbody fusion in the cervical spine has traditionally been accomplished via the anterior and posterior methods, incorporating the use of anterior or posterior plates, usually with the concomitant use of bone grafts. However, recently, interbody fusion cages using titanium mesh cages packed with morselized bone have been reported for use in the cervical spine. Majid et al. (163) performed channeled corpectomy on 34 patients, followed by insertion of a titanium cage implant packed with autogenous bone graft obtained from the vertebral bodies removed in the corpectomy. The authors then performed additional anterior plating on 30 of the 34 patients that involved decompression of two or more levels. Results of the study indicated a 97% radiographic arthrodesis rate in the patient population, with a 12% complication rate including pseudoarthrosis, extruded cage, cage in kyphosis, and radiculopathy. The authors concluded that titanium cages provide immediate anterior column stability and offer a safe alternative to autogenous bone grafts.

Two recent studies examined the biomechanics of anterior cervical interbody cages. Hacker et al. (207) conducted a randomized multicenter clinical trial looking at three different study cohorts of anterior cervical diskectomy fusions: instrumented with HA-coated BAK-C, instrumented with noncoated BAK-C, and uninstrumented, bone graft only (ACDF) fusions. There were a total of 488 patients in the trial, with 288 included in the 1 year follow up and 140 in the 2 year follow up. There were 79.9% one-level fusions and 20.1% two-level fusions performed. Results showed no significant differences between the coated or noncoated BAK-C devices, leading the authors to combine these groups for analysis. Complication rate with the BAK-C group of 346 patients was 10.1% and the ACDF group of 142 patients demonstrated an overall complication rate of 16.2%. The fusion rates for the BAK-C and ACDF fusions at 12 months for one level were 98.7 and 86.4%, respectively; for two levels, 80.0 and 80.0%, respectively. The fusion rates for the BAK-C and ACDF fusions at 24 months for one level were 100 and 96.4%, respectively; for two levels, 91.7 and 77.8%, respectively. Overall, the authors found that the BAK-C cage performed comparably to conventional, uninstrumented, bone graft only anterior diskectomy and fusion. In an *in vitro* comparative study, Yang et al. (208) compared the initial stability and pullout strength of five different cervical cages and analyzed the effect of implant size, placement accuracy, and tightness of the implant on segmental stability. The cages analyzed included (*1*) SynCage-C Curved, (*2*) SynCage-C Wedged, (*3*) Brantigan I/F, (*4*) BAK-C, and (*5*) ACF Spacer. Overall, 35 cervical spines were used, with a total number of 59 segments selected for the study. Flexibility testing was performed under 50 N preload and up to 2 N·m in flexion, extension, lateral bending, and axial rotation. After quasistatic load tests were completed, the cages were subjected to an anterior pull-out test. Direct measurement on the specimen and biplanar radiographs allowed for quantification of distractive height, change in segmental lordosis, cage protrusion, and cage dimensions normalized to the endplate. Results from the study indicated that, in general, the cages were effective in reducing ROM in all directions by approximately one-third, but failed to reduce the neutral zone (NZ) in flexion/extension and axial rotation. Additionally, differences in implants were not significant and only existed between the threaded and nonthreaded designs. The threaded BAK-C was found to have the highest pullout force. Pullout force and lordotic change were both identified as significant predictors of segmental stability, a result the authors underscored as emphasizing the importance of a tight implant fit within the disk space.

RHAKOSS C synthetic bone spinal implant (Orthovita Inc., Malvern, PA) is trapezoidal in shape with an opening in the center for bone graft augmentation, and is fabricated from a bioactive glass/ceramic composite. *In vitro* testing conducted by Goel et al. (209) was conducted to evaluate the expulsion and stabilizing capabilities of the cervical cage in the lower cervical spine; *C6/7 and C4/5 motion segments. from five of the spinal donors were used for the expulsion testing. All specimens received the "Narrow Lordotic" version of the Rhakoss C design. The cages were implanted by orthopedic surgeons following manufacturer recommendations*. Specimens were tested in various modes; intact, destabilized with the cage in place, cage plus an anterior plate (Aline system, Surgical Dynamics Inc., Norwalk, CT), and again with the cage and plate after fatigue loading of 5000 flexion–extension cycles of 1.5 N·m. The results of the expulsion testing indicate that BMD and patient age are good predictors of implant migration resistance ($r = 0.8$). However, the high BMD/age correlation in the specimens makes it difficult to distinguish the relative importance of these two factors. The stability testing demonstrated the ability of a cage with a plate construct to sufficiently stabilize the cervical spine. However, BMD and specimen age play a major role in determining the overall performance of the cervical interbody cage.

Totribe (210) undertook a biomechanical comparison of a new cage made of a forged composite of unsintered-hydroxyapatite particles–poly-L-lactide (F-u-HA-PLLA) and the Ray threaded fusion cage. The objectiove was to compare the stability imparted to the human cadaveric spine by two different threaded cervical cages, and the effect of cyclic loading on construct stability. Threaded cages have been developed for use in anterior cervical

interbody fusions to provide initial stability during the fusion process. However, metallic instrumentation has several limitations. Recently, totally bioresorbable bone fixation devices made of F-u-HA/PLLA have been developed, including a cage for spinal interbody fusion. Twelve fresh ligamentous human cervical spines (C4-C7) were used. Following anterior diskectomy across C5-C6 level, stabilization was achieved with the F-u-HA/PLLA cage in six spines and the Ray threaded fusion cage in the remaining six. Biomechanical testing of the spines was performed with six degrees of freedom before and after stabilization, and after cyclic loading of the stabilized spines (5000 cycles of flexion–extension at 0.5 N·m). The stabilized specimens (with F-u-HA/PLLA cage or the Ray cage) were significantly more stable than the diskectomy case in all directions except in extension. In extension, both groups were stiffer, although not at a significant level ($P > 0.05$). Following fatigue, the stiffness, as compared to the prefatigue case, decreased in both groups, although not at a significant level. The Ray cage group exhibited better stability than the F-u-HA/PLLA cage group in all directions, although a significant difference was found only in right axial rotation.

Lumbar Spine

Anterior and Posterior Spinal Instrumentation. The stability analysis of devices with varying stiffness is best exemplified by a study of Gwon et al. (211) who tested three different transpedicular screw devices: spinal rod-transpedicular screw system (RTS), the Steffee System (VSP), and Crock device (CRK). All devices provided statistically significant ($P < 0.01$) motion reductions across the affected level (L4-L5). The differences among the three devices in reducing motion across L4-L5, however, were not significant. Also, the changes in motion patterns of segments adjacent to the stabilized level compared with the intact case were not statistically significant. These findings have been confirmed by Rohlmann and associates who used a finite element model to address several implant related issues, including this one (212).

In an *in vitro* study, Weinhoffer et al. (213) measured intradiskal pressure in lumbosacral cadaver specimens subjected to constant displacement before and after applying bilateral pedicle screw instrumentation across L4-S1. They noted that intradiskal pressure increased in the disk above the instrumented levels. Also, the adjacent level effect was confounded in two-level instrumentation compared with single-level instrumentation. Other investigators, in principle, have reported similar results. Completely opposite results, however, are presented by several others (212). Results based on *in vitro* studies must be interpreted with caution, being dependent on the testing mode chosen (displacement or load control) for experiments. In the displacement control-type studies, in which applied displacement is kept constant during testing of intact and stabilized specimens, higher displacements and related parameters (e.g., intradiskal pressure) at the adjacent segments are reported. This is not true for the results based on the load control-type studies, in which the applied loads are kept constant.

Lim et al., assessed the biomechanical advantages of diagonal transfixation compared to horizontal transfixation (214). Diagonal cross-members yielded more rigid fixation in flexion and extension, but less in lateral bending and axial rotational modes, as compared to horizontal cross-members. Furthermore, greater stresses in the pedicle screws were predicted for the system having diagonal cross members. The use of diagonal configuration of the transverse members in the posterior fixation systems did not offer any specific advantages, contrary to the common belief.

Biomechanical cadaver studies of anterior fusion promoting and stabilizing devices (214–217) have become increasingly more common in the literature, owing to this procedure's rising popularity (105). *In vitro* testing was performed using the T9-L3 segments of human cadaver spines (218). An L-1 corpectomy was performed, and stabilization was achieved using one of three anterior devices: the ATLP in nine spines, the SRK in 10, and the Z-plate in 10. Specimens were load tested. Testing was performed in the intact state, in spines stabilized with one of the three aforementioned devices after the devices had been fatigued to 5000 cycles at ± 3 N·m, and after bilateral facetectomy. There were no differences between the SRK- and Z-plate-instrumented spines in any state. In extension testing, the mean angular rotation ($\pm$ standard deviation) of spines instrumented with the SRK ($4.7 \pm 3.2^\circ$) and Z-plate devices ($3.3 \pm 2.3^\circ$) was more rigid than that observed in the ATLP-stabilized spines ($9 \pm 4.8^\circ$). In flexion testing after induction of fatigue, however, only the SRK ($4.2 \pm 3.2^\circ$) was stiffer than the ATLP ($8.9 \pm 4.9^\circ$). Also, in extension post-fatigue, only the SRK ($2.4 \pm 3.4^\circ$) provided more rigid fixation than the ATLP ($6.4 \pm 2.9^\circ$). All three devices were equally unstable after bilateral facetectomy. The SRK and Z-plate anterior thoracolumbar implants were both more rigid than the ATLP, and of the former two the SRK was stiffer. The results suggest that in cases in which profile and ease of application are not of paramount importance, the SRK has an advantage over the other two tested implants in achieving rigid fixation immediately postoperatively.

Vahldiek and Panjabi investigated the biomechanical characteristics of short-segment anterior, posterior, and combined instrumentations in lumbar spine tumor vertebral body replacement surgery (219). The L2 vertebral body was resected and replaced by a carbon-fiber cage. Different fixation methods were applied across the L1 and L3 vertebrae. One anterior, two posterior, and two combined instrumentations were tested. The anterior instrumentation, after vertebral body replacement, showed greater motion than the intact spine, especially in axial torsion (range of motion, 10.3 vs. 5.5°; neutral zone, 2.9 vs. 0.7°; $P < 0.05$). Posterior instrumentation provided greater rigidity than the anterior instrumentation, especially in flexion–extension (range of motion, 2.1 vs. 12.6°; neutral zone, 0.6 vs. 6.1°; $P < 0.05$). The combined instrumentation provided superior rigidity in all directions compared with all other instrumentations. Posterior and combined instrumentations provided greater rigidity than anterior instrumentation. Anterior instrumentation should not be used alone in vertebral body replacement.

Oda et al. nondestructively compared three types of anterior thoracolumbar multisegmental fixation with the objective to investigate the effects of rod diameter and rod number on construct stiffness and rod–screw strain (220). Three types of anterior fixation were then performed at L1-L4: (1) 4.75 mm diameter single rod, (2) 4.75 mm dual-rod, and (3) 6.35 mm single-rod systems. A carbon fiber cage was used for restoring intervertebral disk space. Single screws at each vertebra were used for single-rod and two screws for dual-rod fixation. The 6.35 mm single-rod fixation significantly improved construct stiffness compared with the 4.75 mm single rod fixation only under torsion ($P < 0.05$). The 4.75 mm dual rod construct resulted in significantly higher stiffness than did both single-rod fixations ($P < 0.05$), except under compression. For single-rod fixation, increased rod diameter neither markedly improved construct stiffness nor affected rod–screw strain, indicating the limitations of a single-rod system. In thoracolumbar anterior multisegmental instrumentation, the dual-rod fixation provided higher construct stiffness and less rod–screw strain compared with single-rod fixation.

Lumbar Interbody Cages. Cage related biomechanical studies range from evaluations of cages as stand alone devices to use of anterior or posterior instrumentation for additional stabilization.The changes in stiffness and disk height of porcine FSUs by installation of a threaded interbody fusion cage and those by gradual resection of the annulus fibrosus were quantified (117). Flexion, extension, bending, and torsion testing of the FSUs were performed in four sequential stages: stage I, intact FSU; stage II, the FSUs were fitted with a threaded fusion cage; stage III, the FSUs were fitted with a threaded fusion cage with the anterior one-third of the annulus fibrosus excised, including excision of the anterior longitudinal ligament; and stage IV, in addition to stage III, the bilateral annulus fibrosus was excised. Segmental stiffness in each loading in the four stages and a change of disk height induced by the instrumentation were measured. After instrumentation, stiffness in all loading modes ($p < 0.005$) and disk height ($p = 0.002$) increased significantly. The stiffness of FSUs fixed by the cage decreased with gradual excision of the annulus fibrosus in flexion, extension, and bending. These results suggest that distraction of the annulus fibrosus and posterior ligamentous structures by installation of the cage increases the soft-tissue tension, resulting in compression to the cage and a stiffer motion segment. This study explains the basic mechanism through which the cages may provide the stability in various loading modes.

Three posterior lumbar interbody fusion implant constructs (Ray Threaded Fusion Cage, Contact Fusion Cage, and PLIF Allograft Spacer) were tested for stability in a cadaver model (221). None of the standalone implant constructs reduced the neutral zone (amount of motion in response to minimal load application). The constructs decreased the range of motion in flexion and lateral bending. The data did not suggest any implant construct to behave superiorly. Specifically, the PLIF Allograft Spacer is biomechanically equivalent to titanium cages and is devoid of the deficiencies associated with metal cages. Therefore, the PLIF Allograft Spacer is a valid alternative to conventional cages.

The lateral, and other cage orientations within the disk have been increasingly used for fusion (222). In one study, 14 spines were randomized into the anterior group (anterior diskectomy and dual anterior cage—TFC placement) and the lateral group (lateral diskectomy and single transverse cage placement) for load-displacement evaluations. Segmental ranges of motion were similar between spines undergoing either anterior or lateral cage implantation. Combined with a decreased risk of adjacent structure injury through a lateral approach, these data support a lateral approach for lumbar interbody fusion. When used alone to restore stability, the orientation of the cage (oblique vs. posterior) effected the outcome (223). Likewise, in flexion, both the OBAK (Oblique placement of one cage) and CBAK (Conventional posterior placement of two cages) orientations provided significant stability. In lateral bending, CBAK orientation was found to be better then OBAK. In axial mode, CBAK orientation was significantly effective in both directions while OBAK was effective only in right axial rotation. Owing to the differences in the surgical approach and the amount of dissection, the stability for the cages when used alone as a function of cage orientation was different.

The high elastic modulus of the cages causes the structures to be very stiff and may lead to stress-shielded environments within the devices with potential adverse effect on growth of the cancellous bone within the cage itself (224). Using a calf spine model, a study was designed to compare the construct stiffness afforded by 11 differently designed anterior lumbar interbody fusion devices: four different threaded fusion cages: (BAK device, BAK Proximity, Ray TFC, and Danek TIBFD); five different nonthreaded fusion devices (oval and circular Harms cages, Brantigan PLIF and ALIF cages, and InFix device); two different types of allograft (femoral ring and bone dowel); and to quantify their stress-shielding effects by measuring pressure within the devices. Prior to testing, a silicon elastomer was injected into the cages and intra cage pressures were measured using pressure needle transducers. No statistical differences were observed in construct stiffness among the threaded cages and nonthreaded devices in most of the testing modalities. Threaded fusion cages demonstrated significantly lower intracage pressures compared with nonthreaded cages and structural allografts. Compared with nonthreaded cages and structural allografts, threaded fusion cages afforded equivalent reconstruction stiffness but provided more stress-shielded environment within the devices. (This stress shielding effect may further increase in the presence of supplementary fixation devices.)

It is known that micromotion at the cage–endplate interface can influence bone growth into its pores. Loading conditions, mechanical properties of the materials, friction coefficients at the interfaces, and geometry of spinal segments would affect relative micromotion and spinal stability. In particular, relative micromotion is related closely to friction at bone–implant interfaces after arthroplasty. A high rate of pseudarthrosis and a high overall rate of implant migration requiring surgical revision has been reported following posterior lumbar interbody fusion using BAK threaded cages (225). This may be due to poor fixation

of the implant, in addition to the stress shielding phenomena described above. Thus, Kim developed an experimentally validated finite element model of an intact FSU and the FSU implanted with two threaded cages to analyze the motion of threaded cages in posterior lumbar interbody fusion (226). Motion of the implants was not seen in compression. In torsion, a rolling motion was noted, with a range of motion of 10.6° around the central axis of the implant when left–right torsion (25 N·m) was applied. The way the implants move within the segment may be due to their special shape: the thread of the implants cannot prevent the BAK cages rolling within the disk space. However, note that the authors considered too high a value of torsional load; such values may not be clinically relevant. Relative micromotion (slip distance) at the interfaces was obvious at their edges under axial compression. The slip occurred primarily at the anterior edges under torsion with preload, whereas it occurred primarily at the edges of the left cage under lateral bending with preload. Relative micromotion at the interfaces increased significantly as the apparent density of cancellous bone or the friction coefficient of the interfaces decreased. A significant increase in slip distance at the anterior annulus occurred with an addition of torsion to the compressive preload. Relative micromotion was sensitive to the friction coefficient of the interfaces, the bone density, and the loading conditions. A reduction in age-related bone density was less likely to allow bone growth into surface pores of the cage. It was likely that the larger the disk area the more stable the interbody fusion of the spinal segments. However, the amount of micromotion may change in the presence of a posterior fixation technique, an issue that was not reported by the authors.

Almost every biomechanical study has shown that interbody cages alone, irrespective of their shapes, sizes, surface type, material, and approach used for implantation, does not stabilize the spine in all of the modes. It is suspected that this may be caused by the destruction of the appropriate spinal elements like the anterior longitudinal ligament, and anterior annulus fibrosus, or facets. Thus, use of additional instrumentation to augment cages seems to have become a standard procedure.

The 3D flexibility in ligamentous human lumbar spinal units have been investigated after the anterior, anterolateral, posterior, or oblique insertion of various types of interbody cages with supplemental fixtion using anterior or posterior spinal instrumentation (227). With the supplementary fixation using transfacet screws, the differences in stability due to the orientations were not noticeable at all, both before and after; underscoring the importance of using instrumentation when cages are used.

Patwardhan et al. (228) tested the hypothesis that the ability of the ALIF cages to reduce the segmental motions in flexion and extension will be significantly affected by the magnitude of the compressive preload. Fourteen human lumbar spine specimens (L1-sacrum) were tested intact, and after insertion of two threaded cylindrical cages at L5-S1. They were tested in flexion–extension with progressively increasing magnitude of compressive preload from 0 to 1200 N applied along the follower load path (described earlier). The stability of the stand-alone cage construct was significantly affected by the amount of compressive preload applied across the operated segment. In contrast to the extension instability reported in the literature, the two-cage construct exerted a stabilizing effect on the motion segment (reduction in segmental motion) in extension under physiologic compressive preloads. The cages provided substantially more stability, both in flexion and in extension, at larger preloads (800–1200 N) corresponding to standing and walking activities as compared to the smaller preloads (200–400 N) experienced during supine and recumbent postures. The compressive preload due to muscle activity likely plays a substantial role in stabilizing the segment with interbody cages.

The function of the interbody fusion cages is to stabilize the spinal segment primarily by distracting them as well as allowing bone ingrowth and fusion (122). An important condition for efficient formation of bone tissue is achieving adequate spinal stability. However, the initial stability may be reduced due to repeated movements of the spine during activities of daily living. Before and directly after implanation of a Zientek, Stryker, or Ray posterior lumbar interbody fusion cage, 24 lumbar spine segments were evaluated for stability analyses. The specimens were then loaded cyclically for 40,000 cycles at 5 Hz with an axial compression load ranging from 200 to 1000 N. The specimens were tested again in the spine tester. Generally, a decrease in motion in all loading modes was noted after insertion of the Zietek and Ray cages and an increase after implantation of a Stryker cage. In all three groups, greater stability was demonstrated in lateral bending and flexion then in extension and axial rotation. Reduced stability during cyclic loading was observed in all three groups; however, loss of stability was most pronounced in Ray cage group. Authors felt that this may be due to the damage of the cage: bone interface during cyclic loading that was not the case for the other two since they have a flat brick type interface. In order to reduce the incidence of stress risers at the bone–implant interface, it is essential that interbody fusion implants take advantage of the cortical periphery of the vertebral endplates. A larger cross-sectional footprint to the implant design will aid in dispersing the axial forces of spinal motion over a larger surface area and minimize the risk of stress risers, which may result in endplate fractures.

Animal Models

An approximation of the *in vivo* performance of spinal implants in humans can be attained by evaluation in animal models (229). Specifically, animal models provide a dynamic biologic and mechanical environment in which the implant can be evaluated. Temporal changes in both the host biologic tissue and instrumentation can be assessed with selective incremental sacrificing of the animals. Common limitations of animal studies include the method of loading (quadruped versus biped) and the size adjustment of devices needed such that proper fit is achieved in the animals.

Animal studies have revealed the fixation benefits of grouting materials in the preparation of the screw hole (230). The major findings were that the HA grouting of

the screw hole bed before insertion significantly increased fixation (pullout) of the screws. Scanning electron microscopy analysis revealed that HA plasma spraying had deleterious effects on the screw geometry, dulling the self-tapping portion of the screw and reducing available space for bony in-growth.

An animal model of anterior and posterior column instability was developed by McAfee et al. (231–233) to allow *in vivo* observation of bone remodeling and arthrodesis after spinal instrumentation. An initial anterior and posterior destabilizing lesion was created at the L5–6 vertebral levels in 63 adult Beagle dogs. Observations 6 months after surgery revealed a significantly improved probability of achieving a spinal fusion if spinal instrumentation had been used. Nondestructive mechanical testing after removal of all metal instrumentation in torsion, axial compression, and flexion revealed that the fusions performed in conjunction with spinal instrumentation were more rigid. Quantitative histomorphometry showed that the volumetric density of bone was significantly lower (i.e., device-related osteoporosis occurred) for fused versus unfused spines. In addition, a linear correlation occurred between decreasing volumetric density of bone and increasing rigidity of the spinal implant; device-related osteoporosis occurred secondary to Harrington, Cotrel-Dubousset, and Steffee pedicular instrumentation. However, the stress-induced changes in the bone quality found in the animal models is not likely to correlate well with the actual changes in the spinal segment of a patient. In fact, it is suggested that the degeneration in a patient may be determined more by individual characteristics than by the fusion itself (234).

In long bone fractures, internal fixation improves the union rate, but does not accelerate the healing process. Spinal instrumentation also improves the fusion rate in spinal arthrodesis. However, it remains unclear whether the use of spinal instrumentation expedites the healing process of spinal fusion (235,236). Accordingly, an *in vivo* sheep model was used to investigate the effect of spinal instrumentation on the healing process of posterolateral spinal fusion. Sixteen sheep underwent posterolateral spinal arthrodeses at L2-L3 and L4-L5 using equal amounts of autologous bone. One of those segments was selected randomly for further augmentation with transpedicular screw fixation (Texas Scottish Rite Hospital spinal system). The animals were killed at 8 or 16 weeks after surgery. Fusion status was evaluated through biomechanical testing, manual palpation, plain radiography, computed tomography, and histology. Instrumented fusion segments demonstrated significantly higher stiffness than did uninstrumented fusions at 8 weeks after surgery. Radiographic assessment and manual palpation showed that the use of spinal instrumentation improved the fusion rate at 8 weeks (47 vs. 38% in radiographs, 86 vs. 57% in manual palpation). Histologically, the instrumented fusions consisted of more woven bone than the uninstrumented fusions at 8 weeks after surgery. The 16-week-old fusion mass was diagnosed biomechanically, radiographically, and histologically as solid, regardless of pedicle screw augmentation. The results demonstrated that spinal instrumentation created a stable mechanical environment to enhance the early bone healing of spinal fusion.

Human Clinical Models

Loads in posterior implants were measured in 10 patients using telemeterized internal spinal fixation devices (237–239). Implant loads were determined in up to 20 measuring sessions for different activities, including walking, standing, sitting, lying in the supine position, and lifting an extended leg while in the supine position. Implant loads often increased shortly after anterior interbody fusion was performed. Several patients retained the same high level even after fusion had taken place. This explains the reason why screw breakage sometimes occurs more than half a year after implantation. The time of fusion could not be pinpointed from the loading curves. A flexion bending moment acted on the implant even when the body was in a relaxed lying position. This meant that already shortly after the anterior procedure, the shape of the spine was not neutral and unloaded, but slightly deformed, which loaded the fixators. In another study, the same authors used the telemeterized internal spinal fixation devices to study the influence of muscle forces on the implant loads in three patients before and after anterior interbody fusion. Contracting abdominal or back muscles in a lying position was found to significantly increase implant loads. Hanging by the hands from wall bars as well as balancing with the hands on parallel bars reduced the implant loads compared with standing; however, hanging by the feet with the head upside down did not reduce implant loads, compared with lying in a supine position. When lying on an operating table with only the foot end lowered so that the hips were bent, the patient had different load measurements in the conscious and anesthetized states before anterior interbody fusion. The anesthetized patient evidenced predominately extension moments in both fixators, whereas flexion moments were observed in the right fixator of the conscious patient. After anterior interbody fusion had occurred, the differences in implant loads resulting from anesthesia were small. The muscles greatly influence implant loads. They prevent an axial tensile load on the spine when part of the body weight is pulling, for example, when the patient is hanging by their hands or feet. The implant loads may be strongly altered when the patient is under anesthesia.

The above review clearly shows that a large number of fusion enhancement instrumentation are available to surgeons. However, none of the instrumentation is totally satisfactory in its performance and there is room to improve the rate of fusion success, if fusion is the goal. Naturally, alternative fusion approaches (mechanical, biological) are currently being pursued.

The rigidity of a spinal fixation device and its ability to share load with the fusion mass are considered essential for the fusion to occur. If the load transferred through the fusion mass, is increased without sacrificing the rigidity of the construct, a more favorable environment for fusion may be created. To achieve this objective, posterior as well as anterior "dynamized" systems have been designed (240–242). One such posterior system consists of rods and pedicle screws and has a *hinged* connection between the screw head and shaft compared with the rigid screws (Fig. 16a). Another example of the dynamized anterior system (ALC) is shown in Fig. 16b. Load-displacement

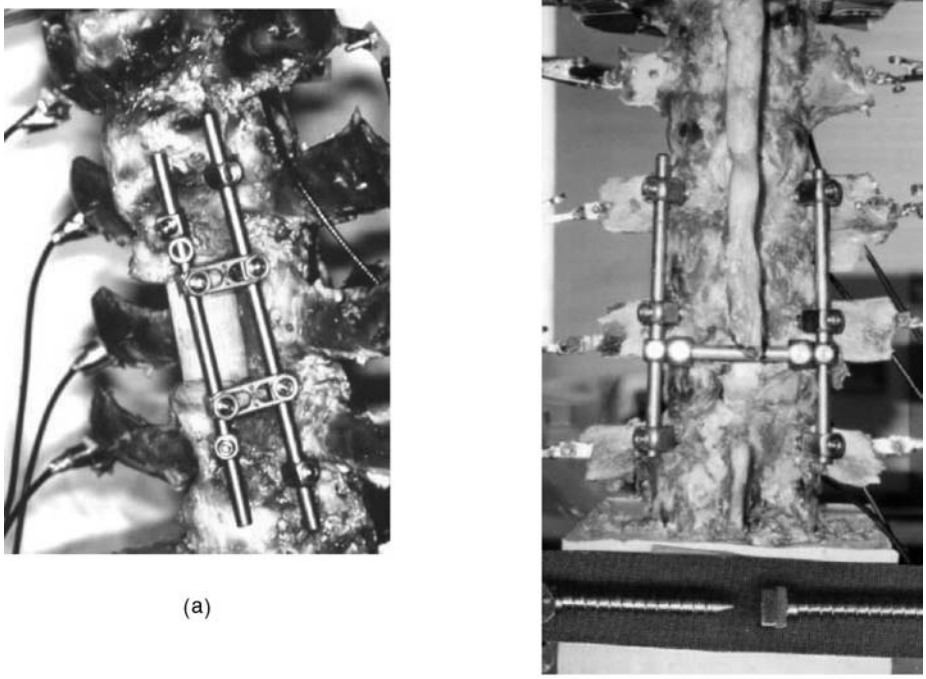


Figure 16. The two different types of dynamized systems used in a cadaver model to assess their stability characteristics. The data were compared with the corresponding "rigid" systems. (a) Posterior system and (b) anterior system. (Taken from Refs. 242 and 241.)

tests were performed to assess the efficacy of these devices in stabilizing a severally destabilized spinal segment. The hinged and rigid posterior systems provided significant stability across the L2-L4 segment in flexion, extension, and lateral bending as compared with the intact case ($P < 0.5$). The stabilities imparted by the hinged-type and its alternative rigid devices were of similar magnitudes. The ALC dynamized and rigid anterior systems also provided significant stability across the L3-L5 segment in flexion, extension, and lateral bending ($P < 05$). The stability imparted by the Dynamized ALC and its alternate rigid system did not differ significantly.

Dynamic stabilization may provide an alternative to fusion for patients suffering from early degenerative disk disease (DDD). The advantages of using a dynamic system are, preservation of the disk loading, allowing some physiologic load sharing in the motion segment. A finite element (FE) study was done to understand the effect of a commercially available dynamic system (DYNESYS, Zimmer Spine) compared to a rigid system on the ROM and disk stresses at the instrumented level (243). An experimentally validated 3-D FE model of intact L3-S1 spine was modified to simulate rigid and dynamic systems across L4-L5 level with the disk intact. The DYNESYS spacer and ligament were modeled with truss elements, with the "no tension" and "no compression" options, respectively. The ROM and disk stresses in response to a 400 N axial compression and 10.6-N·m flexion–extension moment were calculated. The ROM and disk stresses of the adjacent levels with rigid and DYNESYS systems had no significant change when compared to the intact. At the instrumented level in flexion–extension the decrease in motion when compared to the intact was 68/84% for rigid system and 50/56% for DYNESYS. The peak Von Mises disk stresses at the instrumented segment reduced by 41/80% for the rigid system, 27/45% for the DYNESYS system for flexion–extension loading condition. The predicted motion data for the dynamic system was in agreement with the experimental data. From the FE study it can be seen that the DYNESYS system allows more motion than the rigid screw-rod system, and hence allows for partial disk loading. This partial disk loading might be advantageous for a potential recovery of the degenerated disk, thus making dynamic stabilization systems a viable option for patients in early stages of DDD.

An anterior bone graft in combination with posterior instrumentation has been shown to provide superior support because the graft is in line with axial loads and the posterior elements are left intact. However, employing posterior instrumentation with anterior grafting requires execution of two surgical procedures. Furthermore, use of a posterior approach to place an interbody graft requires considerable compromise of the posterior elements, although it reduces the surgery time. It would be advantageous to minimize surgical labor and structural damage caused by graft insertion into the disk space via a posterior approach. Authors have addressed this issue by preparing an interbody bone graft using morselized bone (244–246). This device is a gauze bag of Dacron that is inserted into the disk space, filled with morselized bone, and tied shut, Fig. 17. *In vitro* testing measured the rotations of each vertebral level of mechanically loaded cadaver lumbar spines, both in intact and several experimental conditions. With the tension band alone, motion was restored to the

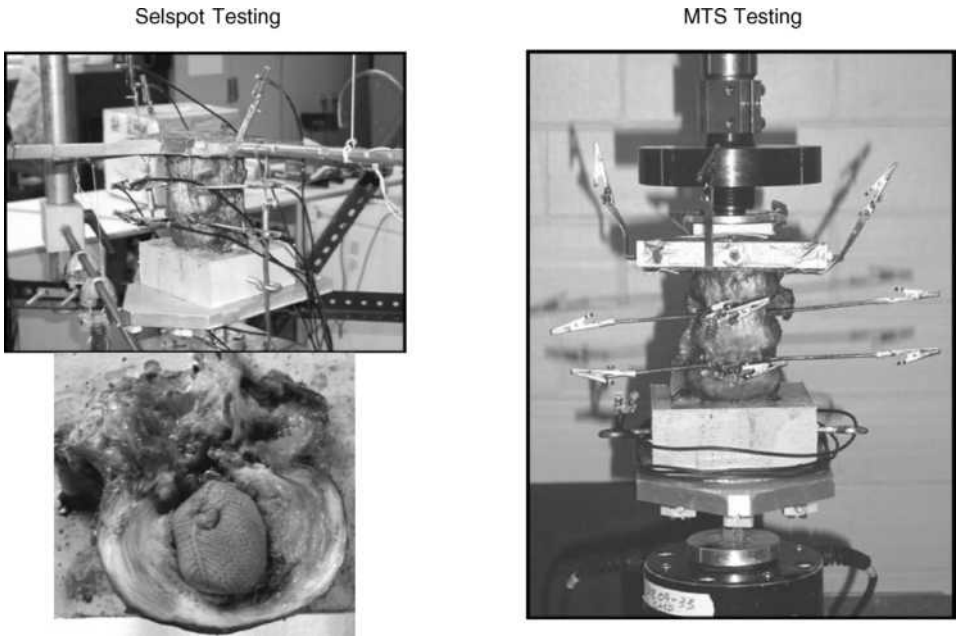


Figure 17. The Bag system developed by Spineology Inc. The increases and decreases in motion with respect to intact segment for bag alone and bag with a band are also shown. (Taken from Ref. 244.)

intact case, except in extension where it was reduced. With the graft implant, motion was restored to intact in all of the loading modes, except in flexion where it was reduced. With the tension band and graft, motion was again restored to intact except in flexion and extension where it was reduced. *In vitro* results suggest that a tension band increases stability in extension, while the bag device alone seems to provide increased stability in flexion. The implanted bag filled with morselized bone in combination with a posterior tension band, restores intact stiffness. Postcyclic results in axial compression suggest that the morselized bone in the bone-only specimens either consolidates or extrudes from the cavity despite confinement. Motion restoration or reduction as tested here is relevant both to graft incorporation and segment biomechanics. The posterior interbody grafting method using morselized bone is amenable to orthoscopy. It produces an interbody graft without an anterior surgical approach. In addition, this technique greatly reduces surgical exposure with minimal blood loss and no facet compromise. This technique would be a viable alternative to current 360° techniques pending animal tests and clinical trials.

Bone grafting is used to augment bone healing and provide stability after spinal surgery. Autologous bone graft is limited in quantity and unfortunately associated with increased surgical time and donor-site morbidity. Recent research has provided insight into methods that may modulate the bone healing process at the cellular level in addition to reversing the effects of symptomatic disk degeneration, which is a potentially disabling condition, managed frequently with various fusion procedures. Alternatives to autologous bone graft include allograft bone, demineralized bone matrix, recombinant growth factors, and synthetic implants (247,248). Each of these alternatives could possibly be combined with autologous bone marrow or various growth factors. Although none of the presently available substitutes provides all three of the fundamental properties of autograft bone (osteogeneticity, osteoconductivity, and osteoinductivity), there are a number of situations in which they have proven clinically useful. A literature review indicate that alternatives to autogenous bone grafting find their greatest appeal when autograft bone is limited in supply or when acceptable rates of fusion may be achieved with these substitutes. For example, bone morphogenetic proteins have been shown to induce bone formation and repair.

Relatively little research has been undertaken to investigate the efficacy of OP-1 in the above stated role (249,250). Grauer et al. performed single-level intertransverse process lumbar fusions at L5-L6 of 31 New Zealand White rabbits. These were divided into three study groups: autograft, carrier alone, and carrier with OP-1. The animals were killed 5 weeks after surgery. Five (63%) of the 8 in the autograft group had fusion detected by manual palpation, none (0%) of the 8 in the carrier-alone group had fusion, and all 8 (100%) in the OP-1 group had fusion. Biomechanical testing results correlated well with those of manual palpation. Histologically, autograft specimens were predominantly fibrocartilage, OP-1 specimens were predominantly maturing bone, and carrier-alone specimens did not show significant bone formation. OP-1 was found to reliably induce solid intertransverse process fusion in a rabbit model at 5 weeks. Smoking interferes with the success of posterolateral lumbar fusion and the above authors extended the investigation to study the effect of using OP-1 to enhance fusion process in patients who smoke. Osteoinductive protein-1 was able to overcome the inhibitory effects of nicotine in a rabbit posterolateral spine fusion model, and to induce bony fusion reliably at 5 weeks.

Finally, another study performed a systematic literature review on non-autologous interbody fusion materials in anterior cervical fusion, gathering data from 32 clinical- and ten laboratory studies. Ten alternatives to autologous bone were compared: autograft, allograft, xenograft, poly-(methyl methacrylate) (PMMA), biocompatible osteoconductive polymer (BOP), Hydroxyapatite compounds, bone

morphogenic protein (BMP), Carbon fiber, metallic devices and ceramics. The study revealed that autologous bone still provides the golden standard that other methods should be compared to. The team concluded that the results of the various alternative fusion options are mixed, and comparing the different methods proved difficult. Once a testing standard has been established, reliable comparisons could be conducted.

Finite Element Models

In vitro investigations and *in vivo* animal studies contain numerous limitations, including that these are both time consuming and monetarily expensive. The most important limitations of *in vitro* studies are that muscle contributions to loading are not usually incorporated and the highly variable quality of the cadaver specimens. As stated earlier, *in vivo* animal studies usually involve quadruped animals, and the implant sizes usually need to be scaled according to the animal size. In an attempt to compliment the above protocols, several FE models of the ligamentous spine have been developed (251–257).

Goel et al. (255) generated osteoligamentous FE models of intact lumbar one segment (L3-L4) and two segments (L3-L5). Using the L3-L4 model, they simulated fusion with numerous techniques in an attempt to describe the magnitude and position of internal stresses in both the biologic tissue (bone and ligament) and applied hardware. Specifically, the authors modeled bilateral fusion using unilateral and bilateral plating. Bilateral plating models showed that cancellous bone stresses were significantly reduced with the instrumentation simulated in the immediate postoperative period. Completely consolidated fusion mass case, load transmission led to unloading of the cancellous bone region, even after simulated removal of the device. Thus, this model predicted that removal of the device would not alleviate stress shielding-induced osteopenia of the bone and that this phenomenon may truly be a complication of the fusion itself. As would be expected, unilateral plating models revealed higher trabecular bone stresses than were seen in the bilateral plating cases. The degree of stability afforded to the affected segment, however, was less. Thus, a system that allows the bone to bear more load as fusion proceeds may be warranted. Several solutions have been proposed to address this question.

For example, a fixation system was developed that incorporated polymer washers in the load train (Steffee variable screw placement, VSP). The system afforded immediate postoperative stability and reduced stiffness with time as the washers underwent stress relaxation (a viscoelastic effect) (256). The FE modeling of this system immediately after implantation showed that internal bony stresses were increased by $\sim 20\%$ over the same system without the polymeric material. In addition, mechanical property manipulation of the washers simulating their *in vivo* stress relaxation revealed these stresses were continuously increasing, promoting the likelihood that decreased bone resorption would occur. The other solution is the use of dynamized fixation devices, as diskussed next.

The ability of a hinged pedicle screw-rod fixation (dynamized, see next section for details) device to transmit more loads across the stabilized segment compared with its rigid equivalent system was predicted using the FE models (240). In general, the hinged screw device allowed for slightly larger axial displacements of L3, while it maintained flexion rotational stability similar to the rigid screw device (Table 11). Slightly larger axial displacements may be sufficient enough to increase the load through the graft since the stiffness of the disk was increased by replacing it (shown as the "nucleus" in the tables) with a cancellous, cortical, or titanium interbody device to simulate the fusion mass in the model (Table 12).

Table 11. Axial Displacement and Angular Rotation of L3 with respect to L4 for the 800 N Axial Compression[a]

	Axial Displacement, mm		Rotation, deg	
Graft	Rigid	Hinged	Rigid	Hinged
Cancellous	−0.258	−0.274	0.407	0.335
Cortical	−0.134	−0.137	0.177	0.127
Titanium	−0.132	−0.135	0.174	0.126

[a]Taken from Ref. 240.

The FE modeling coupled with adaptive bone remodeling algorithms has been used to investigate temporal changes associated with interbody fusion devices. Grosland et al. predicted the change in bone density distribution after implantation of the BAK device (Fig. 18) (257). The major findings included hypertrophy of bone directly in the load train (directly overlying and underlying the implant) and lateral atrophy secondary to the relatively high stiffness of the implant. The model also predicted that bone growth into and around the larger holes in the implant, resulting in sound fixation of the device.

Nonfusion Treatment Alternatives

Various methods have been employed in the characterization of device effectiveness for which spinal fusion is indicated. Because of nonphysiological nature of fusing the spinal segments that are supposed to provide motion–flexibility, adjacent-level degeneration, and other complications associated with the fusion process, alternatives to fusion have been proposed.

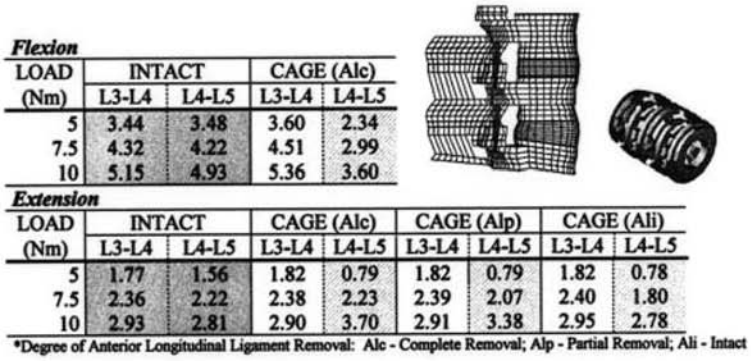
Ray Nucleus

In 1988, Ray presented a prosthetic nuclear replacement consisting of flexible woven filaments (Dacron) surrounding an internal semipermeable polyethylene membranous sac filled with hyaluronic acid and a thixotropic agent (i.e.,

Table 12. Loads Transferred Through the "Nucleus" and the Device for the 800 N Axial Compression in newtons[a]

	Rigid		Hinged	
Graft	"Nucleus"	Device	"Nucleus"	Device
Cancellous	712.4	87.6	767.9	32.1
Cortical	741.2	58.8	773.5	26.5
Titanium	742.5	57.5	774.3	25.7

[a]Taken from Ref. 37.



Flexion

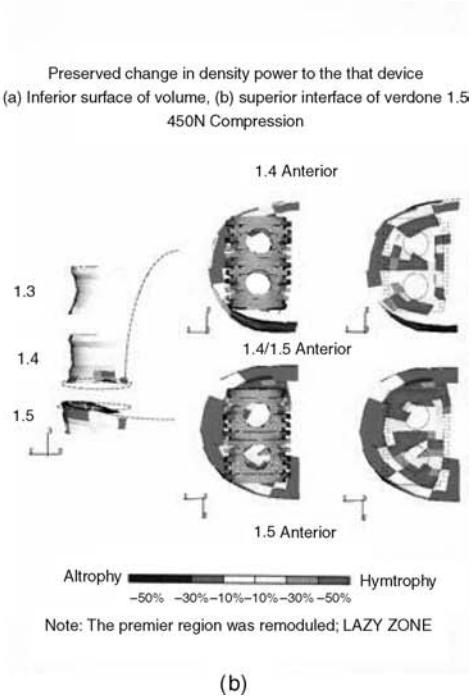
LOAD	INTACT		CAGE (Alc)	
(Nm)	L3-L4	L4-L5	L3-L4	L4-L5
5	3.44	3.48	3.60	2.34
7.5	4.32	4.22	4.51	2.99
10	5.15	4.93	5.36	3.60

Extension

LOAD	INTACT		CAGE (Alc)		CAGE (Alp)		CAGE (Ali)	
(Nm)	L3-L4	L4-L5	L3-L4	L4-L5	L3-L4	L4-L5	L3-L4	L4-L5
5	1.77	1.56	1.82	0.79	1.82	0.79	1.82	0.78
7.5	2.36	2.22	2.38	2.23	2.39	2.07	2.40	1.80
10	2.93	2.81	2.90	3.70	2.91	3.38	2.95	2.78

*Degree of Anterior Longitudinal Ligament Removal: Alc - Complete Removal; Alp - Partial Removal; Ali - Intact

(a)



(b)

Figure 18. (a) The FE model of a ligamentous motion segment was used to predict load-displacement behavior of the segment following cage placement. Alc = anterior longitudinal ligament completely removed/cut, Alp = partially cut, and Ali = intact; and (b) Percentage change in density of the bone surrounding the BAK cage. (Taken from Refs. 32,33, and 257.)

a hydrogel) (244,258,259). As a nucleus replacement, the implant can be inserted similar to a thoracolumbar interbody fusion device, either posteriorly or transversely. Two are inserted per disk level in a partly collapsed and dehydrated state, but would swell due to the strongly hygroscopic properties of the hyaluronic acid constituent. The designer expects the implant to swell enough to distract the segment while retain enough flexibility to allow a normal range of motion. An option is to include therapeutic agents in the gel that would be released by water flow in and out of the prosthesis according to external pressures.

Recent reports on biomechanical tests of the device show that it can produce some degree of stabilization and distraction. Loads of 7.5 N·m and 200 N axial were applied to six L4-L5 specimens. Nucleotomized spines increased rotations by 12–18% depending on load orientation, but implanted spines (implant placed transversely) showed a change of −12% to +2% from the intact with substantial reductions in neutral zone. Up to 2 mm of disk height was recovered by insertion. The implant, however, was implanted and tested in its no hydrated form. The biomechanics of the hydrated prosthesis may vary considerably from that of its desiccated form.

In Situ Curable Prosthetic Intervertebral Nucleus (PIN)

The device (Disc Dynamics, Inc, Minnetonka, MN) consists of a compliant balloon connected to a catheter (Fig. 19) (244,260). This is inserted and a liquid polymer injected into the balloon under controlled pressure inflating the balloon, filling the cavity, and distracting the intervertebral disk. Within 5 min the polymer is cured. Five fresh-frozen osteoligamentous three-segment human lumbar spines, screened for abnormal radiograph and low bone density, were used for the biomechanical study. The spines were tested under four conditions: intact, denucleated, implanted, and fatigued. Fatiguing was produced by cyclic loading from 250 to 750 N at 2 Hz for at least 100,000 cycles. Nuclectomy was performed through a 5.5 mm trephine hole in the right middle lateral side of the annulus. The device was placed in the nuclear cavity as described earlier. Following biomechanical tests, these specimens were radiographed and dissected to determine any structural damage inflicted during testing. Middle segment rotations generally increased with diskectomy, but were restored to the normal intact range with implantation. After fatiguing, rotations across the implanted segment increased. However, these were not more than, and often less than the intact adjacent segments. During polymer injection under compressive load the segment distracted as much as +1.8 mm (av) at the disk center as determined by the surrounding gauges. Over 1.6 mm was maintained during polymer cure with compression. The immediate goals of a disk replacement system are to restore disk height and provide segment mobility without causing instability. This study showed that PIN device could reverse the destabilizing effects of a nuclectomy and restore normal segment stiffness. Significant increases in disk height can also be achieved. Implanting the majority of disk replacement systems requires significant annulus removal, this device requires minimal surgical compromise and has the potential to be performed arthroscopically.

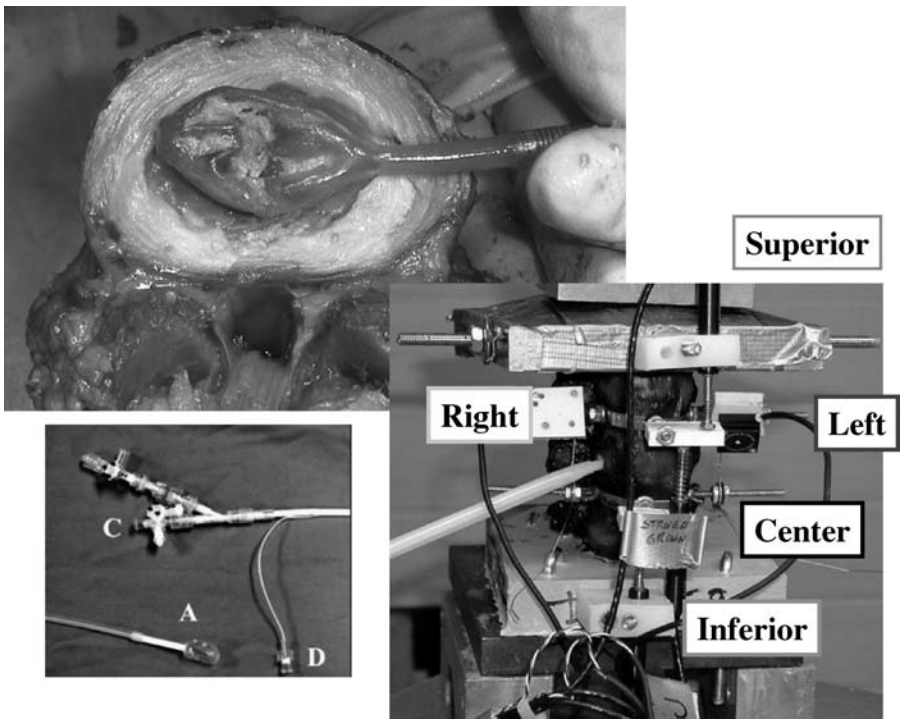


Figure 19. In situ curable prosthetic intervertebral nucleus (PIN) developed by Disc Dynamics, Inc. (Taken from Ref. 244.)

Artificial Disk

One of the most recent developments for nonfusion treatment alternatives is replacement of the intervertebral disk (244,261,262). The goal of this treatment alternative is to restore the original mechanical function of the resected disk. One of the stipulations of artificial disk replacement is that the remaining osseous spinal and paraspinal soft tissue components are not compromised by pathologic changes. Bao et al. (263) have classified the designs of total disk replacements into four categories: (*1*) low friction sliding surface; (*2*) spring and hinge systems; (*3*) contained fluid-filled chambers; and (*4*) disks of rubber and other elastomers. The former two designs seek to take advantage of the inherently high fatigue characteristics that all-metal designs afford. The latter two designs attempt to incorporate some of the viscoelastic and compliant properties that are exhibited by the normal, healthy intervertebral disk. Hedman et al. (264) outlined the major design criteria for intervertebral disk prosthesis: The disk must be able to maintain its mechanical integrity out to approximately 85 million cycles; consist of biocompatible materials; exist entirely within the normal disk space and maintain physiologic disk height; restore normal kinematic motion wherein the axes of each motion, especially sagittal plane motion, is correctly replicated; duplicate the intact disk stiffness in all three planes of rotation and compression; provide immediate and long-term fixation to bone; and, finally, provide *failsafe* mechanisms such that if an individual component of the design fails, catastrophic failure is not immediately imminent, and it does not lead to periimplant soft tissue damage. This is certainly one of the greatest design challenges that bioengineers have encountered to date. In the following, some of the methods are discussed that are being employed in an attempt to meet this rigorous challenge.

One of the available studies dealt iterative design of the artificial disk replacement based on measured biomechanical properties. Lee, Langrana and co-workers (265,266) looked at incorporating three different polymers into their prosthetic intervertebral disk design and tried to represent the separate components (annulus fibrosis and nucleus) of the normal disk in varying proportion. They loaded their designs under 800 N axial compression and in compression-torsion out to 5°. The results indicated that disks fabricated from homogeneous materials exhibited isotropy that could not replicate the anisotropic behavior of the normal human disk. Thus, 12 layers of fiber reinforcement were incorporated in an attempt to mimic the actual annulus fibrosis. This method did result in more closely approximating the mechanical properties of the normal disk. Through this method of redesign and testing, authors claim that eventually "a disk prosthesis that has mechanical properties comparable to the natural disk could be manufactured."

The FE analyses have also been recruited in an effort to perturbate design with an eye toward optimizing the mechanical behavior of artificial disks. Goel and associates modified a previously validated intact finite element model to create models implanted with a ball-and-cup and slip core-type artificial disk models via an anterior approach, Figs. 20 and 21 (244,245,261). To study surgical variables, small and large windows were cut into the annulus, and the implants were placed anteriorly and posteriorly within the disk space. The anterior longitudinal ligament was also restored. Models were subjected to either 800 N axial compression force alone or to a combination of 10 N·m flexion–extension moments and 400 N axial preload. Implanted model predictions were compared with those of the intact model. The predicted rotations for the two disk implanted models were in agreement with the experimental data.

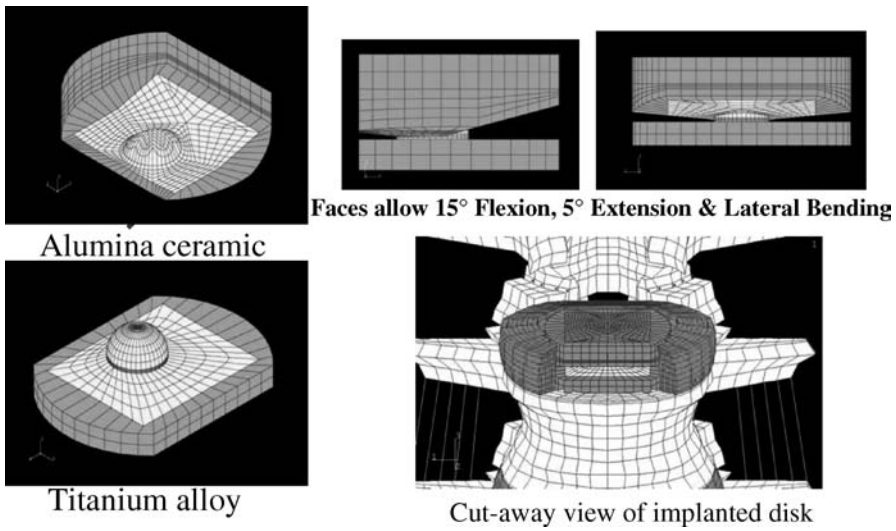


Figure 20. The intact finite element model of a ligamentous segment was modified to simulate the ball and socket type artificial disk implant. (Taken from Refs. 244,245.)

For the ball and socket design disk facet loads were more sensitive to the anteroposterior location of the artificial disk than to the amount of annulus removed. Under 800-N axial compression, implanted models with an anteriorly placed artificial disk exhibited facet loads 2.5 times greater than loads observed with the intact model, whereas posteriorly implanted models predicted no facet loads in compression. Implanted models with a posteriorly placed disk exhibited greater flexibility than the intact and implanted models with anteriorly placed disks. Restoration of the anterior longitudinal ligament reduced pedicle stresses, facet loads, and extension rotation to nearly intact levels. The models suggest that, by altering placement of the artificial disk in the anteroposterior direction, a surgeon can modulate motion-segment flexural stiffness and posterior load sharing, even though the specific disk replacement design has no inherent rotational stiffness.

The motion data, as expected, differed between the two disk designs (ball and socket, and slip core) and as compared to the intact as well, Fig. 22. Similar changes were observed for the loads on the facets, Fig. 23.

The experimentally validated finite element models of the intact and disk implanted L3-L5 segments revealed that both of these devices do not restore motion and loads across facets back to the intact case. (These design restore the intact biomechanics in a limited sense.) These differences are not only due to the size of the implants but the inherent design differences. Ball and socket design has a more "fixed" center of rotation as compared to the slip core design in which the COR undergoes a wider variation. Further complicating factor is the location of the disk within the annular space itself, a parameter under the control of the surgeon. Thus, it will be difficult to restore biomechanics of the segment back to normal using such designs. Only clinical follow up studies will provide the effects of such variations on the changes in spinal structures as a function of time.

More Recent and Future Initiatives

Although many of the well-accepted investigation techniques and devices have been discussed above, other

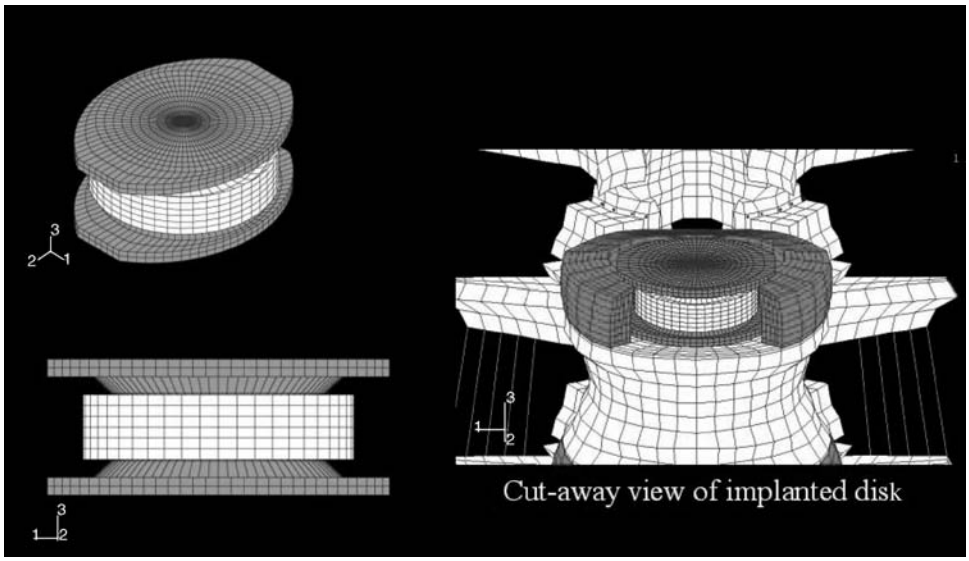


Figure 21. The intact finite element model of a ligamentous segment was modified to simulate the slip core type artificial disk implant. (Taken from Ref. 244.)

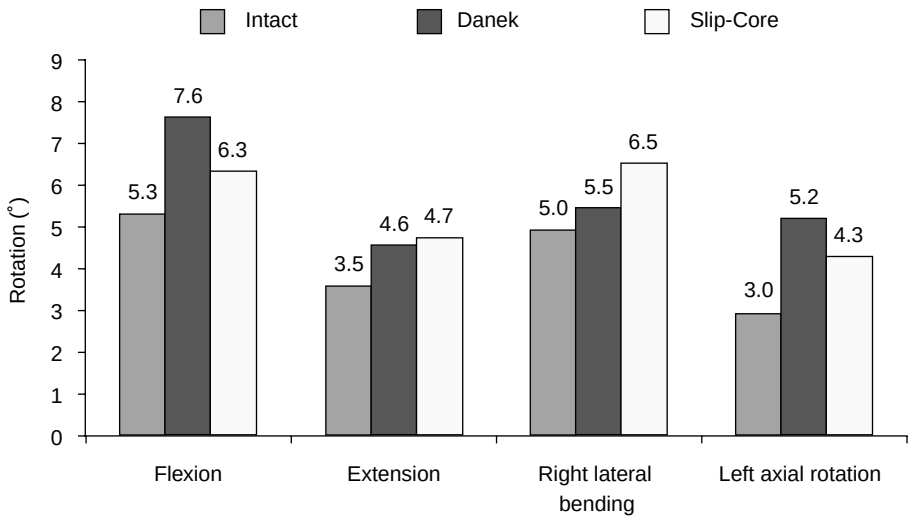


Figure 22. Predicted rotations for the two disk designs, shown in Figs. 20 and 21, as compared to the intact. (Taken from Ref. 244.)

techniques for the stabilization–fusion of the spine and nonfusion approaches are currently being investigated. These concepts are likely to play a significant role in future and are discussed. One such technique is vertebroplasty. Painful vertebral osteoporotic compression fractures leads to significant morbidity and mortality (263). Kyphoplasty and vertebroplasty are relatively new techniques that help decrease the pain and improve function in fractured vertebras.

Vertebroplasty is the percutaneous injection of PMMA cement into the vertebral body (263–269). While PMMA has high mechanical strength, it cures fast and thus allows only a short handling time. Other potential problems of using PMMA injection may include damage to surrounding tissues by a high polymerization temperature or by the unreacted toxic monomer, and the lack of long-term biocompatibility. Bone mineral cements, such as calcium carbonate and CaP, have longer working time and low thermal effect. They are also biodegradable while having a good mechanical strength. However, the viscosity of injectable mineral cements is high, and the infiltration of these cements into vertebral body has been questioned. Lim et al. evaluated the compression strength of human vertebral bodies injected with a new calcium phosphate (CaP) cement with improved infiltration properties before compression fracture and also for vertebroplasty in comparison with PMMA injection (268). The bone mineral densities of 30 vertebral bodies (T2-L1) were measured using dual-energy X-ray absorptiometry. Ten control specimens were compressed at a loading rate of 15 mm/min to 50% of their original height. The other specimens had 6 mL of PMMA ($n = 10$) or the new CaP ($n = 10$) cement injected through the bilateral pedicle approach before being loaded in compression. Additionally, after the control specimens had been compressed, they were injected with either CaP ($n = 5$) or PMMA ($n = 5$) cement using the same technique, to simulate vertebroplasty. Loading experiments were repeated with the displacement control of 50% vertebral

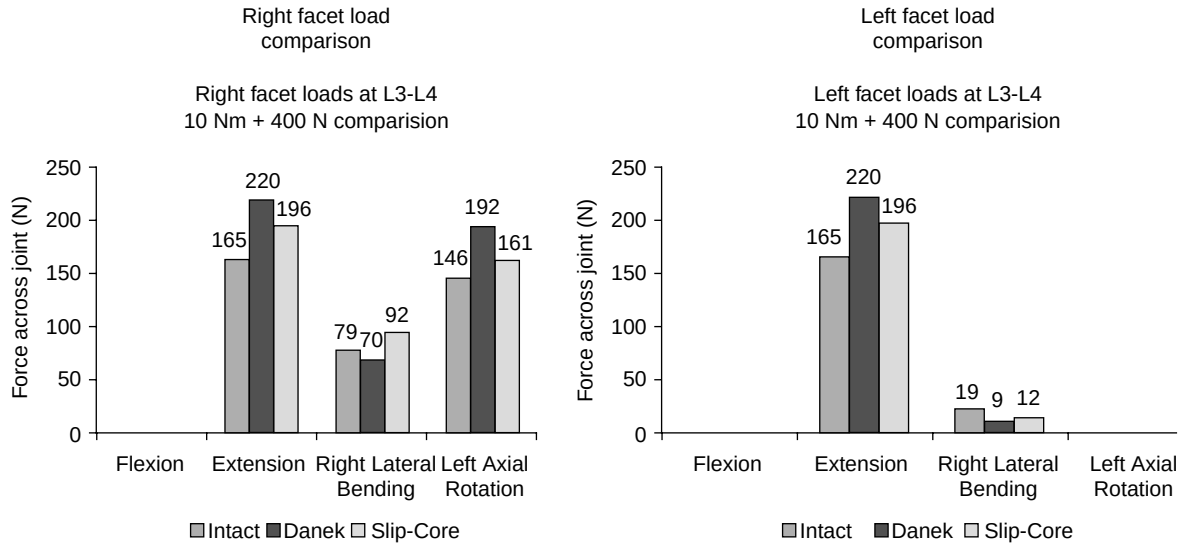


Figure 23. Predicted facet loads for the two disk designs, shown in Figs. 20 and 21, as compared to the intact. (Taken from Ref. 244.)

height. Load to failure was compared among groups and analyzed using analysis of variance. Mean bone mineral densities of all five groups were similar and ranged from 0.56 to 0.89 $g \cdot cm^{-2}$. The size of the vertebral body and the amount of cement injected were similar in all groups. Load to failure values for PMMA, the new CaP, and vertebroplasty PMMA were significantly greater than that of control. Load to failure of the vertebroplasty CaP group was higher than control but not statistically significant. The mean stiffness of the vertebroplasty CaP group was significantly smaller than control, PMMA, and the new CaP groups. The mean height gains after injection of the new CaP and PMMA cements for vertebroplasty were minimal (3.56 and 2.01%, respectively). Results of this study demonstrated that the new CaP cement can be injected and infiltrates easily into the vertebral body. It was also found that injection of the new CaP cement can improve the strength of a fractured vertebral body to at least the level of its intact strength. Thus, the new CaP cement may be a good alternative to PMMA cement for vertebroplasty, although further *in vitro*, *in vivo* animal and clinical studies should be done. Furthermore, the new CaP may be more effective in augmenting the strength of osteoporotic vertebral bodies, and for preventing compression fractures considering our biomechanical testing data and the known potential for biodegradability of the new CaP cement. Belkof et al. (266) found that the injection of either Orthocomp or Simplex P resulted in vertebral body strengths that were significantly greater than initial strength values. Vertebral bodies augmented with Orthocomp recovered their initial stiffness; and, vertebral bodies augmented with Simplex P were significantly less stiff than they were in their initial condition. However, these biomechanical results have yet to be substantiated in clinical studies.

Previous biomechanical studies have shown that injections of 8–10 mL of cement during vertebroplasty restore or increase vertebral body strength and stiffness; however, the dose-response association between cement volume and restoration of strength and stiffness is unknown. Belkof et al. (266) investigated the association between the volume of cement injected during percutaneous vertebroplasty and the restoration of strength and stiffness in osteoporotic vertebral bodies. Two investigational cements were studied: Orthocomp (Orthovita, Malvern, PA) and Simplex 20 (Simplex P with 20% by weight barium sulfate. Compression fractures were experimentally created in 144 vertebral bodies (T6-L5) obtained from 12 osteoporotic spines harvested from female cadavers. After initial strength and stiffness were determined, the vertebral bodies were stabilized using bipedicular injections of cement totaling 2, 4, 6, or 8 mL and recompressed, from which post-treatment strength and stiffness were measured. Strength and stiffness were considered restored when post-treatment values were not significantly different from initial values. Strength was restored for all regions when 2 mL of either cement was injected. To restore stiffness with Orthocomp, the thoracic and thoracolumbar regions required 4 mL, but the lumbar region required 6 mL. To restore stiffness with Simplex 20, the thoracic and lumbar regions required 4 mL, but the thoracolumbar region required 8 mL. These data provide guidance on the cement volumes needed to restore biomechanical integrity to compressed osteoporotic vertebral bodies.

Liebschner et al. undertook a finite element based biomechanical study to provide a theoretical framework for understanding and optimizing the biomechanics of vertebroplasty, especially the effects of volume and distribution of bone cement on stiffness recovery of the vertebral body, just like the preceding experimental study (269). An experimentally calibrated, anatomically accurate finite-element model of an elderly L1 vertebral body was developed. Damage was simulated in each element based on empirical measurements in response to a uniform compressive load. After virtual vertebroplasty (bone cement filling range of 1–7 cm^3) on the damaged model, the resulting compressive stiffness of the vertebral body was computed for various spatial distributions of the filling material and different loading conditions. Vertebral stiffness recovery after vertebroplasty was strongly influenced by the volume fraction of the implanted cement. Only a small amount of bone cement (14% fill or 3.5 cm^3) was necessary to restore stiffness of the damaged vertebral body to the predamaged value. Use of a 30% fill increased stiffness by $> 50\%$ compared with the predamaged value. Whereas the unipedicular distributions exhibited a comparative stiffness to the bipedicular or posterolateral cases, it showed a medial-lateral bending motion (toggle) toward the untreated side when a uniform compressive pressure load was applied. Only a small amount of bone cement (15% volume fraction) is needed to restore stiffness to predamage levels, and greater filling can result in substantial increase in stiffness well beyond the intact level. Such overfilling also renders the system more sensitive to the placement of the cement because asymmetric distributions with large fills can promote single-sided load transfer and thus toggle. These results suggest that large fill volumes may not be the most biomechanically optimal configuration, and an improvement might be achieved by use of lower cement volume with symmetric placement. These theoretical findings support the experimental observations described in the proceeding paragraph, except these authors did not analyze the relationship between cement type and volume needed to restore strength.

Hitchon et al. compared the stabilizing effects of the HA product, with PMMA in an experimental compression fracture of L1 (268). No significant difference between the HA and PMMA cemented-fixated spines was demonstrated in flexion, extension, left lateral bending, or right- and left-axial rotation. The only difference between the two cements was encountered before and after fatiguing in right lateral bending ($p \leqslant 0.05$). The results of this study suggest that the same angular rigidity can be achieved by using either HA or PMMA. This is of particular interest because HA is osteoconductive, undergoes remodeling, and is not exothermic.

Advances in the surgical treatment of spinal etiologies continues to evolve with the rapid progression of technology. The advent of robotics, Microelectromechanical systems (MEMS) (270), novel biomaterials, and genetic and tissue engineering are revolutionizing spinal medicine, Novel biomaterials, also termed "smart biomaterials" that are capable of conforming or changing their mechanical

properties in response to different loading paradigms are being investigated for their use in spinal implant design. The rapidly advancing field of tissue engineering opens new possibilities to solving spine problems. By seeding and growing intervertebral disk cells, it could be possible to grow a new bioartificial disk, to be implanted in to the spine. Studies are in progress at a number of centers, including our own (269).

CONCLUSION

The stability (or instability) of the human spine is integral to the diagnosis and treatment of patients with low back pain. The stability of the lumbar spine as portrayed by its motion characteristics can be determined through the use of clinical and radiographic criteria or other methods of determining the orientation of one spinal vertebra with respect to another. The instability can be the result of injury, disease, and many other factors, including surgery. Therefore, it is necessary to become familiar with recent findings and suggestions that deal with the instability that can result from such procedures. The prevalence of spinal fusion and stabilization procedures to restore spinal stability and host of other factors is continuously increasing. This article has presented many of the contemporary biomechanical issues germane to stabilization and fusion of the spine. Because of the wide variety of devices available, various testing protocols have been developed in an attempt to describe the mechanical aspects of these devices. These investigations reveal comparative advantages (and disadvantages) of the newer designs to existing hardware. Subsequent *in vivo* testing, specifically animal models, provides data on the performance of the device in a dynamic physiologic environment. All of the testing, *in vitro* and *in vivo*, helps to build confidence that the instrumentation is safe for clinical trial. Future biomechanical work is required to produce newer devices and optimize existing ones, with an eye toward reducing the rates of nonfusion and pseudoarthrosis. In addition, novel devices and treatments that seek to restore normal spinal function and loading patterns without fusion continue to necessitate advances in biomechanical methods. These are the primary challenges that need to be incorporated in future biomechanical investigations. Finally, one has to gain understanding of the effects of devices at the cellular level and one must undertake outcome assessment studies to see if the use of instrumentation is warranted for the enhancement of the fusion process.

ACKNOWLEDGMENTS

Manuscript is based on the work sponsored by various funding agencies over the last 20 years. Thanks also to a large number of coinvestigators who have contributed to the original work reported in this article.

BIBLIOGRAPHY

1. Goel VK, Weinstein JN. Clinical Biomechanics of the Lumbar Spine. Boca Raton (FL): CRC Press; 1990.
2. White AAA, Panjabi MM. Clinical Biomechanics of Spine. New York: Lippincot. 2nd ed.; 1990.
3. Doherty BJ, Heggeness MH. Quantitative anatomy of the second cervical vertebra. Spine 1995;20:513–517.
4. (a) Yoganandan N, Halliday A, Dickman C, Benzel E. Practical anatomy and fundamental biomechanics, in Spine Surgery: Techniques, Complication Avoidance, and Management. In: Benzel EC, editor. New York: Churchill Livingstone; 1999. p 93–118. (b) Clausen JD. Experimental & Theoretical Investigation of Cervical Spine Biomechanics—Effects of Injury and Stabilization. Ph.D. dissertation, University of Iowa, Iowa City (IA) 52242; 1996. (c) Puttlitz CM. A Biomechanical Investigation of the Craniovertebral Junction. Ph.D. dissertation, University of Iowa, Iowa City (IA) 52242; 1999. (d) Scifert JL. Biomechanics of the Cervical Spine. Ph.D. dissertation, University of Iowa, Iowa City (IA) 52242; 2000. (e) Goel VK, Panjabi MM, Kuroki H, Rengachary S, McGowan D, Ebraheim N. Biomechanical Considerations—Spinal Instrumentation. ISSLS Book. 2003. (f) Kuroki H, Holekamp S, Goel V, Panjabi M, Ebraheim N, Singer K. Biomechanics of Spinal Deformity in Inflammatory Disease. book chapter submitted; 2003. (g) Goel VK, Dooris AP, McGowan D, Rengachary S. Biomechanics of the Disk. In: Lewandrowski K-U, Yaszemski MJ, White III AA, Trantolo DJ, Wise DL, editors. Advances in Spinal Fusion—Molecular Science, Biomechanics, and Clinical Management. Somerville (CA): M&N Toscano; 2003. (h) Panjabi MM, Yue JJ, Dvorak J, Goel V, Fairchild T, White AA. Chapt. 4 Cervical Spine Kinematics and Clinical Instability. Cervical Spine Research Society; 2003 (submitted).
5. Xu R, Nadaud MC, Ebraheim NA, Yeasting RA. Morphology of the second cervical vertebra and the posterior projection of the C2 pedicle axis. Spine 1995;20:259–263.
6. Heggeness MH, Doherty BJ. The trabecular anatomy of the axis. Spine 1993;18:1945–1949.
7. Watkins RG. Cervical Spine Injuries in Athletes. In: Clark CR, editor. The cervical spine. Philadelphia: Lippincott-Raven; 1998. p 373–386.
8. Penning L. Differences in anatomy, motion, development, and aging of the upper and lower cervical disk segments. Clin Biomech 1988;3:337–347.
9. Pooni J, Hukins D, Harris P, Hilton R, Davies K. Comparison of the structure of human intervertebral disks in the cervical, thoracic and lumbar regions of the spine. Surg Rad Anat 1986;8:175–182.
10. Tulsi R, Perrett L. The anatomy and radiology of the cervical vertebrae and the tortuous vertebral artery. Aust Rad 1975; 19:258–264.
11. Frykholm R. Lower cervical vertebrae and intervertebral disks. Surgical anatomy and pathology. Acta Chir Scand 1951;101:345–359.
12. Hall M. Luschka's Joint. Springfield (IL): Charles C. Thomas; 1965.
13. Hardacker J, Shuford R, Capicotto P, Pryor P. Radiographic standing cervical segmental alignment in adult volunteers without neck symptoms. Spine 1997;22(13):1472–1480.
14. Harrison D, Janik T, Troyanovich S, Harrison D, Colloca C. Evaluation of the assumptions used to derive an ideal normal cervical spine model. JMPT 1997;20(4):246–256.
15. Harrison D, Janik T, Troyanovich S, Holland B. Comparisons of lordotic cervical spine curvatures to a theoretical ideal model of the static sagittal cervical spine. Spine 1996;21(6): 667–675.
16. Rosenberg LC. Proteoglycans. In: Owen R, Goodfellow J, Bullough P, editors. Scientific Foundations of Orthopaedics and Traumatology. Philadelphia: WB Saunders; 1980. p 36–42.

17. Humzah M, Soames R. Human intervertebral disk: structure and function. Anatomical Rec 1988;220:337–356.
18. Oda J, Tanaka H, Tsuzuki N. Intervertebral disk changes with aging of the human cervical vertebra: from neonate to the eighties. Spine 1988;13:1205–1211.
19. Alker GJ, Oh YS, Leslie EV. High cervical spine and craniocervical junction injuries in fatal traffic accidents: a radiological study. Orthop Clin NA 1978;9:1003–1010.
20. Freeman L, Wright T. Experimental observations of concussion and contusion of the spinal cord. Ann Surg 1953; 137(4): 433–443.
21. Hirsch C, Galante J. Laboratory conditions for tensile tests in annulus fibrosis from human intervertebral disks. Acta Orthop Scand 1967;38:148–162.
22. Singer KP, Breidahl PD, Day RE. Posterior element variation at the thoracolumbar transition. A morphometric study using computed tomography. Clin Biomech 1989;4:80–86.
23. Panjabi MM, Thibodeau LL, Crisco III JJ, et al. What constitutes spinal instability? Clin Neurosurg 1988; 34: 313–339.
24. Putz R. The detailed functional anatomy of the ligaments of the vertebral column. Anatomischer Anzeiger 1992;173:40–47.
25. Gilbertson LG. Mechnism of fracture and biomechanics of orthosis in thoracolumbar region. Ph.D. dissertation, University of Iowa (IA); 1993.
26. Ikegawa S, et al. The effect of joint angle on cross sectional area and muscle strength of human elbow flexors, Human Kinetics. Biomechanics SA: Champaign; 1979. p 35–44.
27. Macintosh JE, Bogduk N, Pearcy MJ. The effects of flexion on the geometry and actions of the lumbar erector spinae. Spine 1993;18:884–893.
28. White AA, et al. Spinal stability. Evaluation and treatment. instr course lect. AAOS 1981;30:457–484.
29. Penning L, Normal Kinematics of the Cervical Spine, Clinical Anatomy and Management of Cervical Spine Pain. In: Giles SK, editor. Oxford (UK): Butterworth Heinemann; 1998. p 53–70.
30. Panjabi MM, Dvorak J, Sandler AJ, Goel VK, White AA. Cervical Spine Kinematics and Clinical Instability. In: CR C, editor. The Cervical Spine. Philadelphia: Lippincott-Raven; 1998. p 53–77.
31. Panjabi MM, Dvorak J, Duranceau J, Yamamoto I, Gerber M, Rauschning W, Bueff HU. Three dimensional movements of the upper cervical spine. Spine 1988;13:726–730.
32. Goel VK, Clark CR, Gallaes K, Liu YK. Moment–rotation relationships of the ligamentous occipitoatlanto-axial complex. Spine 1988;21:673–680.
33. Moroney S, Schultz A, Miller J, Andersson G. Load-displacement properties of lower cervical spine motion segments. J Biomech 1988;21(9):769–779.
34. Lysell E. Motion in the cervical spine. Ph.D. dissertation. Acta Orthop Scand (123 Suppl); 1969.
35. Bernhardt M, White AA, Panjabi MM, McGowan DP. Biomechanical considerations of spinal stability. The Spine. 3rd ed. Philadelphia: WB Saunders Company; 1992. p 1167–1195.
36. Pearcy MJ. Stereoradiography of lumbar spine motion. Acta Orthop Scand 1985;56(212 Suppl).
37. Calve J, Galland M. Physiologie Pathologique Du Mal De Pott. Rev Orthop 1930;1:5.
38. Rolander SD. Motion of the lumbar spine with special reference to the stabilizing effect of the posterior fusion. Scta Ortho Scand 1966;90(Suppl):1–114.
39. Oda T, Panjabi MM, Crisco JJ, Oxland T, Katz L, Nolte L-P. Experimental study of atlas injuries II—Relevance to clinical diagnosis and treatment. Spine 1991;16:S466–473.
40. Oda T, Panjabi MM, Crisco JJ, Oxland TR. Multidirectional instabilities of experimental burst fractures of the atlas. Spine 1992;17:1285–1290.
41. Jefferson G. Fracture of the atlas vertebra: report of four cases, and a review of thos previously recorded. Br J Surg 1920;7:407–422.
42. Heller JG, Amrani J, Hutton WC. Transverse ligament failure—a biomechanical study. J Spinal Disord 1993;6: 162–165.
43. Goel VK, et al. Ligamentous laxity across CO-C1-C2 complex: axial torque-rotation characteristics until failure. Spine 1990;15:990–996.
44. Chang H, Gilbertson LG, Goel VK, Winterbottom JM, Clark CR, Patwardhan A. Dynamic response of the occipito-atlanto-axial (C0-C1-C2) complex in right axial rotation. J Orth Res 1992;10:446–453.
45. Doherty BJ, Heggeness MH, Esses SI. A biomechanical study of odontoid fractures and fracture fixation. Spine 1993;18:178–184.
46. Anderson LD, D'Alonzo RT. Fractures of the odontoid process of the axis. J Bone Joint Surg 1974;56-A:1663–1674.
47. Schatzker J, Rorabeck CH, Waddell JP. Fractures of the dens. J Bone Joint Surg 1971;53-B:392–405.
48. Clark CR, White AA. Fractures of the dens. JBJS 1985;67-A:1340–1348.
49. Althoff B. Fracture of the odontoid process. Acta Orthop Scand 1979;177(Suppl):1–95.
50. Mouradian WH, Fietti VG, Cochran GVB, Fielding JW, Young J. Fractures of the odontoid: a laboratory and clinical study of mechanisms. Orthop Clinics of NA 1978;9:985–1001.
51. Puttlitz CM, Goel VK, Clark CR, Traynelis VC. Pathomechanism of Failures of the Odontoid. Spine 2000;25:2868–2876.
52. Goel VK, Montgomery RE, Grosland NM, Pope MH, Kumar S. Ergonomic factors in the work place contribute to disk degeneration. In: Kumar S, editor. Biomechanics in Ergonomics. Taylor & Francis; 1999. p 243–267.
53. Liu YK, Njus G, Buckwalter J, Wakano K. Fatigue response of lumbar intervertebral joints under axial cyclic loading. Spine 1983;6:857.
54. Hooper DM. Consequences of asymmetric lifting on external and internal loads at the L3–5 lumbar levels. Ph.D. dissertation, Iowa City (IA): University of Iowa; 1996.
55. Kelsey JL. An epidemiological study of acute herniated lumbar intervertebral disk. Rehum Rehabil 1975;14:144.
56. Panjabi MM, Anderson GBJ, Jorneus L, Hult E, Matteson L. In vivo measurements of spinal column vibrations. JBJS 1986;68A:695.
57. Wilder DG, Woodworth BB, Frymoyer JW, Pope MH. Vibration and the human spine. Spine 1982;7:243.
58. Liu YK, Goel VK, DeJong A, Njus G, Nishiyama K, Buckwalter J. Torsional fatigue of the lumbar intervertebral joints. Spine 1985;10:894–900.
59. Kong W-Z, Goel VK. Ability of the finite element models to predict response of the human spine in sinusoidal vertical vibration. Spine 2003;28:1961–1967.
60. Furlong DR, Palazotto AN. A finite element analysis of the influence of surgical herniation on the viscoelastic properties of the intervertebral disk. J Biomech 1983;16:785.
61. Lee C-K, Kim Y-E, Jung J-M, Goel VK. Impact response of the vertebral segment using a finite element model. Spine 2000;25:2431–2439.
62. Dvorak J, Hayek J, Zehnder R. CT-functional diagnostics of the rotatory instability of the upper cervical spine: part 2. An evaluation on healthy adults and patients with suspected instability. Spine 1987;12:726–731.

63. Dvorak J, Penning L, Hayek J, Panjabi MM, Grob D, Zehnder R. Functional diagnostics of the cervical spine using computer tomography. Neuroradiology 1988;30:132–137.
64. Crisco JJ, Takenori O, Panjabi MM, Bueff HU, Dovrak J, Grob D. Transections of the C1-C2 joint capsular ligaments in the cadaveric spine. Spine 1991;16:S474–S479.
65. Fielding JW. Cineroentgenography of the normal cervical spine. J Bone Joint Surg 1957;39-a:1280–1288.
66. Puttlitz CM, Goel VK, Clark CR, Traynelis VC, Scifert JL, Grosland NM. Biomechanical rationale for the pathology of rheumatoid arthritis in the craniovertebral junction. Spine 2000;25:1607–1616.
67. Panjabi MM, White AA, Keller D, Southwick WO, Friedlaender G. Stability of the cervical spine under tension. J Biomech 1978;11:189–197.
68. Goel VK, Clark CR, Harris KG, Schulte KR. Kinematics of the cervical spine: effects of multiple total laminectomy and facet wiring. J Orthop Res 1988;6:611–619.
69. Goel VK, Clark CR, Harris KG, Kim YE, Schulte KR. Evaluation of effectiveness of a facet wiring technique: an in vitro biomechanical investigation. Ann Biomed Eng 1989;17:115–126.
70. Goel VK, Clausen JD. Prediction of load sharing among spinal components of a C5–C6 motion segment using the finite element approach. Spine 1998;23:684–691.
71. Clausen JD, Goel VK, Traynelis VC, Scifert JL. Uncinate processes and luschka's joints influence the biomechanics of the cervical spine-quantification using a finite element model of the C5-C6 segment. J Orth Res 1997;15:342–347.
72. Edwards WT, Zheng Y, Ferrara LA, Yuan HA. Structural features and thickness of the vertebral cortex in the thoracolumbar spine. Spine 2001;26(2):218–225.
73. Goel VK, Monroe BT, Gilbertson LG, Brinckmann P. Interlaminar shear stresses and laminae separation in a disk: finite element analysis of the L3-4 motion segment subjected to axial compressive loads. Spine 1995;20:689–698. (1994 Volvo Award Paper).
74. Goel VK, Kim YE. Effects of injury on the spinal motion segment mechanics in the axial compression mode. Clin Biomech 1989;4:161–167.
75. Posner I, White AA, Edwards WT, et al. A biomechanical analysis of clinical stability of the lumbar lumbrosacral spine. Spine 1982;7:374–389.
76. Kong WZ, Goel VK, Gilbertson LG, et al. Effects of muscle dysfunction on lumbar spine mechanics–a finite element study based on a two motion segments model. Spine 1996; 21:2197.
77. Kong W, Goel V, Weinstein J. Role of facet morphology in the etiology of degenerative spondylolisthesis in the presence of muscular activity. 41st Annual Meeting. Orlando, FL: Orthopaedic Research Society; Feb 13–16, 1995.
78. Lipson JL, Muir H. Proteoglycans in experimental intervertebral disk degeneration. Spine 1981;6:194–210.
79. Raynor RB, Moskovich R, Zidel P, Pugh J. Alterations in primary and coupled neck motions after facetectomy. J Neurosurg 1987;12:681–687.
80. Schulte KR, Clark CR, Goel VK. Kinematics of the cervical spine following discectomy and stabilization. Spine 1989;14: 1116–1121.
81. Martins A. Anterior cervical discectomy with and without interbody bone graft. J Neurosurg 1976;44:290–295.
82. Wilson D, Campbell D. Anterior cervical discectomy without bone graft. Neurosurgery 1977;47:551–555.
83. Goel VK, Nye TA, Clark CR, Nishiyama K, Weinstein JN. A technique to evaluate an internal spinal device by use of the Selspot system: an application to Luque closed loop. Spine 1987;12:150–159.
84. Zdeblick TA, Zou D, Warden KE, McCabe R, Kunz D, Vanderby R. Cervical stability after foraminotomy. A biomechanical in vitro analysis. J Bone Joint Surg [Am] 1992; 74:22–27.
85. Zdeblick TA, Abitbol JJ, Kunz DN, McCabe RP, Garfin S. Cervical stability after sequential capsule resection. Spine 1993;18:2005–2008.
86. Cusick JF, Yoganandan N, Pintar F, Myklebust J, Hussain H. Biomechanics of cervical spine facetectomy and fixation techniques. Spine 1988;13:808–812.
87. Voo LM, Kumaresan S, Yoganandan N, Pintar FA, Cusick JF. Finite element analysis of cervical facetectomy. Spine 1997;22:964–9.
88. Nowinski GP, Visarius H, Nolte LP, Herkowitz HN. A biomechanical comparison of cervical laminaplasty and cervical laminectomy with progressive facetectomy. Spine 1993;18: 1995–2004.
89. Goel VK, Clark CR, Harris KG, Schulte KR. Kinematics of the cervical spine: effects of multiple total laminectomy and facet wiring. J Orthop Res 1988;6:611–619.
90. Goel VK, Clark CR, McGowan D, Goyal S. An in-vitro study of the kinematics of the normal, injured and stabilized cervical spine. J Biomech 1984;17:363–376.
91. Bell DF, Walker JL, O'Connor G, Tibshirani R. Spinal deformity after multiple-level cervical laminectomy in children. Spine 1994;19:406–411.
92. Lee S-J, Harris KG, Nassif J, Goel VK, Clark CR. In vivo kinematics of the cervical spine; part I: development of a roentgen stereophotogrammetric technique using metallic markers and assessment of its accuracy. J Spinal Disord 1993;6:522–534.
93. Lee S-J. Three-dimensional analysis of post-operative cervical spine motion using simultaneous roentgen stereophotogrammetry with metallic markers. Ph.D. dissertation, Iowa City (IA): University of Iowa; 1993.
94. Kumaresan S, Yoganandan N, Pintar FA, Maiman DJ, Goel VK. Contribution of disk degeneration to osteophyte formation in the cervical spine—A biomechanical investigation. J Orth Res 2001;19(5):977–984.
95. Kubo S, Goel VK, Yang S-J, Tajima N. Biomechanical comparison of cervical double door laminoplasty using hydroxyapatite spacer. Spine 2003;28(3):227–234.
96. Kubo S, Goel VK, Tajima N. The biomechanical effects of multilevel posterior foraminotomy and foraminotomy with double door laminoplasty. J Spinal Disorders 2002;15:477–485.
97. Panjabi MM. Low back pain and spinal instability. In: Weinstein JN, Gordon SL, editors. Low back pain: a scientific and clinical overview. San Diego (CA): American Academy of Orthopaedic Surgeons; 1996. p 367–384.
98. Panjabi MM, Kaigle AM, Pope MH. Degeneration, injury, and spinal instability. In: Wiesel SW, et al., editors. The lumbar spine. 2nd ed. Volume 1, Philadelphia, PA: W.B. Saunders; 1996. p 203–211.
99. Goel VK, et al. Kinematics of the whole lumbar spine—effect of discectomy. Spine 1985;10:543–554.
100. Watters WC, Levinthal R. Anterior cervical discectomy with and without fusion–results, complications, and long-term follow-up. Spine 1994;19:2343.
101. Oxland TR, Teija L, Bernhard J, Peter C, Kurt L, Philippe J, Lutz-P N. The relative importance of vertebral bone density and disk degeneration in spinal flexibility and interbody implant performance An in vitro study. Spine 1996;21(22): 2558–2569.

102. Fraser RD. Interbody, posterior, and combined lumbar fusions. Spine 1995;20:167S.
103. Goel VK, Gilbertson LG. Basic science of spinal instrumentation. Clin Orthop 1987;335:10.
104. Penta M, Fraser RD. Anterior lumbar interbody fusion–a minimum 10 year follow-up. Spine 1997;22:2429.
105. Goel VK, Pope MH. Biomechanics of fusion and stabilization. Spine 1995;20:35S.
106. Purcell GA, Markolf KL, Dawson EG. Twelfth thoracic-first lumbar vertebral mechanical stability of fractures after Harrington-rod instrumentation. J Bone Joint Surg Am 1981;63:71.
107. Lehman RA, Kuklo TR, O-Brien MF. Biomechanics of thoracic pedicle screw fixation. Part I–Screw biomechanics. Seminars in Spine Surgery 2002;14(1):8–15.
108. Pfeiffer M, et al. Effect of specimen fixation method on pullout tests of pedicle screws. Spine 1996;21:1037.
109. Pfeiffer M, Hoffman H, Goel VK, et al. In vitro testing of a new transpedicular stabilization technique. Eur Spine J 1997;6:249.
110. Carlson GD, et al. Screw fixation in the human sacrum–an in vitro study of the biomechanics of fixation. Spine 1992;17:S196.
111. Ryken TC, Clausen John D, Traynelis Vincent C, Goel Vijay K. Biomechanical analysis of bone mineral density, insertion technique, screw torque, and holding strength of anterior cervical plate screws. J Neurosurg 1995;83:324–329.
112. Choi W, Lee S, Woo KJ, Koo KJ, Goel V. Assessment of pullout strengths of various pedicle screw designs in relation to the changes in the bone mineral density. 48th Annual Meeting, Dallas, TX: Orthopedic Research Society; Feb. 10–13, 2002.
113. Lim TH, et al. Prediction of fatigue screw loosening in anterior spinal fixation using dual energy x-ray absorptiometry. Spine 1995 Dec 1; 20(23):2565–2568; discussion 2569.
114. Hasagawa, et al. An experimental study of a combination method using a pedicle screw and laminar hook for the osteoporotic spine. Spine 1997; 22:958.
115. McLain RF, Sparling E, Benson DR. Early failure of short-segment pedicle instrumentation for thoracolumbar fractures. J Bone Joint Surg Am 1993;75:162.
116. Bühler DW, Berlemann U, Oxland TR, Nolte L-P. Moments and forces during pedicle screw insertion in vitro and in vivo measurements. Spine 23(11):1220–1227.
117. Goel VK, Grosland NM, Scifert JL. Biomechanics of the lumbar disk. J Musculoskeletal Res 1997;1:81.
118. Lund T, Oxland TR, Jost B, Cripton P, Grassmann S, Etter C, Nolte LP. Interbody cage stabilisation in the lumbar spine: biomechanical evaluation of cage design, posterior instrumentation and bone density. J Bone Joint Surg Br 1998 Mar; 80(2):351–359.
119. Rapoff AJ, Ghanayem AJ, Zdeblick TA. Biomechanical comparison of posterior lumbar interbody fusion cages. Spine 1997;22:2375.
120. Steffen T, Tsantrizos A, Aebi M. Effect of implant design and endplate preparation on the compressive strength of interbody fusion construct. Spine 2000;25(9):1077–1084.
121. Brooke NSR, Rorke AW, King AT, et al. Preliminary experience of carbon fibre cage prosthesis for treatment of cervical spine disorders. Br J Neurosurg 1997;11:221.
122. Ketller A, Wilke HJ, Dietl R, Krammer M, Lumenta C, Claes L. Stabilizing effect of posterior lumbar interbody fusion cages before and after cyclic loading. J Neurosurg 2000;92(1 Suppl):87–92.
123. Janssen ME, Nguyen, Beckham C, Larson R. A Biological cage. Eur Spine J 2000;9(1 Suppl):S102–S109.
124. Murakami H, Boden SD, Hutton WC. Anterior lumbar interbody fusion using a barbell-shaped cage: A biomechanical comparison. J Spinal Disord 2001;14(5):385–392.
125. Murakami H, Horton WC, Kawahara N, Tomita K, Hutton WC. Anterior lumbar interbody fusion using two standard cylindrical threaded cages, a single mega-cage, or dual nested cages: A biomechanical comparison. J Orthop Sci 2001;6(4):343–348.
126. Wang J, Zou D, Yuan H, Yoo J. A biomechanical evaluation of graft loading characteristics for anterior cervical discectomy and fusion. Spine 1998;23(22):2450–2454.
127. Cheng B, Moore D, Zdeblick T. Load sharing characteristics of two anterior cervical plate systems. in 25th Annual Meeting of the Cervical Spine Research Society;. 1997. Rancho Mirage (CA).
128. Rapoff A, O'Brien T, Ghanayem A, Heisey D, Zdeblick T. Anterior cervical graft and plate load sharing. J Spinal Disorders 1999;12(1):45–49.
129. An H, et al. Effect of endplate conditions and bone mineral density on the compressive strength of the graft-endplate interphase in the cervical spine. in 14th Annual Meeting of the North American Spine Society. Chicago Hilton and Towers; 1999.
130. Dietl R, Krammer HJ, Kettler M, Wilke A, Claes H-J, Lumenta L, Christianto B. Pullout test with three lumbar interbody fusion cages. Spine 2002;27(10):1029–1036.
131. Clausen JD, et al. A protocol to evaluate semi-rigid pedicle screw systems. J Biomech Eng 1997; 119:364.
132. Cunningham BW, et al. Static and cyclic biomechanical analysis of pedicle screw constructs. Spine 1993;18: 1677.
133. Goel VK, Winterbottom JM, Weinstein JN. A method for the fatigue testing of pedicle screw fixation devices. J Biomech 1994;27:1383.
134. Chang KW, et al. A comparative biomechanical study of spinal fixation using the combination spinal rod plate and transpedicular screw fixation system. J Spinal Disord 1989;1:257.
135. Panjabi MM. Biomechanical evaluation of spinal fixation devices: Part I. A conceptual framework. Spine 1988; 13(10):1129–1134.
136. Panjabi MM, Goel VK. Adjacent-Level Effects: Design of a new test protocol and finite element model simulations of disk replacement. In: Goel K, Panjabi MM, editors. Roundtables in Spine Surgery; Spine Biomechanics: Evaluation of Motion Preservation Devices and Relevant terminology, Chapt. 6. Vol 1, Issue 1, St. Louis: Quality Medical Publishing; 2005.
137. Patwardhan AG, et al. A follower load increases the load-carrying capacity of the lumbar spine in compression. Spine 1999;24:1003–1009.
138. Cusick J, Pintar F, Yoganandan N, Baisden J. Wire fixation techniques of the cervical facets. Spine 1997;22(9):970–975.
139. Fuji T, et al. Interspinous wiring without bone grafting for nonunion or delayed union following anterior spinal fusion of the cervical spine. Spine 1986;11(10):982–987.
140. Garfin S, Moore M, Marshall L. A modified technique for cervical facet fusions. Clin Orthoped Rel Res 1988;230:149–153.
141. Geisler F, Mirvis S, Zrebeet H, Joslyn J. Titanium wire internal fixation for stabilization of injury of the cervical spine: clinical results and postoperative magnetic resonance imaging of the spinal cord. Neurosurgery 1989;25(3):356–362.
142. Scuderi G, Greenberg S, Cohen D, Latta L, Eismont F. A biomechanical evaluation of magnetic resonance imaging–compatible wire in cervical spine fixation. Spine 1993; 18(14):1991–1994.

143. Stathoulis B, Govender S. The triple wire technique for bifacet dislocation of the cervical spine. Injury 1997; 28(2):123–125.
144. Weis J, Cunningham B, Kanayama M, Parker L, McAfee P. In vitro biomechanical comparison of multistrand cables with conventional cervical stabilization. Spine 1996; 21(18):2108–2114.
145. Abumi K, Panjabi M, Duranceu J. Biomechanical evaluation of spinal fixation devices. Part III. Stability provided by six spinal fixation devices and interbody bone graft. Spine 1989;14:1239–1255.
146. Anderson P, Henley M, Grady M, Montesano P, Winn R. Posterior cervical arthrodesis with AO reconstruction plates and bone graft. Spine 1991: 16(3S):S72–S79.
147. Ebraheim N, An H, Jackson W, Brown J. Internal fixation of the unstable cervical spine using posterior Roy-Camille plates: preliminary report. J Orthop Trauma 1989;3(1): 23–28.
148. Fehlings M, Cooper P, Errico T. Posterior plates in the management of cervical instability: longterm results in 44 patients. J Neurosurg 1994;81:341–349.
149. Bailey R. Fractures and dislocations of the cervical spine. Postgrad Med 1964;35:588–599.
150. Graham A, Swank M, Kinard R, Lowery G, Dials B. Posterior cervical arthrodesis and stabilization with a lateral mass plate. Spine 1996;21(3):323–329.
151. Grubb M, Currier B, Stone J, Warden K, An K-N. Biomechanical evaluation of posterior cervical stabilization after wide laminectomy. Spine 1997;22(17):1948–1954.
152. Nazarian S, Louis R. Posterior internal fixation with screw plates in traumatic lesions of the cervical spine. Spine 1991;16(3 Suppl):S64–S71.
153. Smith MMC, Langrana N, Lee C, Parsons J. A biomechanical study of a cervical spine stabilization device: Roy-Camille plates. Spine 1997;22(1):38–43.
154. Swank M, Sutterlin C, Bossons C, Dials B. Rigid internal fixation with lateral mass plates in multilevel anterior and posterior reconstruction of the cervical spine. Spine 1997; 22(3):274–282.
155. Aebi M, Zuber K, Marchesi D. Treatment of cervical spine injuries with anterior plating: indications, techniques, and results. Spine 1991;16(3 Suppl):S38–S45.
156. Bose B. Anterior cervical fusion using Caspar plating: analysis of results and review of the literature. Surg Neurol 1998;49:25–31.
157. Ebraheim N, et al. Osteosynthesis of the cervical spine with an anterior plate. Orthopedics 1995;18(2):141–147.
158. Grubb M, et al. Biomechanical evaluation of anterior cervical spine stabilization. Spine 1998;23(8):886–892.
159. Naito M, Kurose S, Sugioka Y. Anterior cervical fusion with the Caspar instrumentation system. Inter Orthop 1993; 17:73–76.
160. Paramore C, Dickman C, Sonntag V. Radiographic and clinical follow–up review of Caspar plates in 49 patients. J Neurosurg 1996;84:957–961.
161. Randle M, et al. The use of anterior Caspar plate fixation in acute cervical injury. Surg Neurol 1991;36:181–190.
162. Ripa D, Kowall M, Meyer P, Rusin J. Series of ninety–two traumatic cervical spine injuries stabilized with anterior ASIF plate fusion technique. Spine 1991;16(3 Suppl):S46–S55.
163. Majd M, Vadhva M, Holt R. Anterior cervical reconstruction using titanium cages with anterior plating. Spine 1999; 24(15):1604–1610.
164. Pearcy M, Burrough S. Assessment of bony union after interbody fusion of the lumbar spine using biplanar radiographic technique. J Bone Joint Surg 1982;64B:228.
165. Capen D, Garland D, Waters R. Surgical stabilization of the cervical spine. A comparative analysis of anterior and posterior spine fusions. Clin Orthop 1985;196:229.
166. Hunter L, Braunstein E, Bailey R. Radiographic changes following anterior cervical spine fusions. Spine 1980;5:399.
167. Drennen J, King E. Cervical dislocation following fusion of the upper thoracic spine for scoliosis. J Bone Joint Surg 1978;60A:1003.
168. Heller J, Silcox DH, Sutterlin C. Complications of posterior plating. Spine 1995;20(22):2442–2448.
169. Gallie WE. Fractures and dislocations of the cervical spine. Am J Surg 1939;46:495–499.
170. Cybulski GR, Stone JL, Crowell RM, Rifai MHS, Gandhi Y, Glick R. Use of Halifax interlaminar clamps for posterior C1-C2 arthrodesis. Neurosurgery 1988;22:429–431.
171. Currier BL, Neale PG, Berglund LJ, An KN. A biomechanical comparison of new posterior occipitalcervical instrumentation. New Orleans, LA: Orthopaedic Research Society; 1998.
172. Oda I, Abumi K, Sell LC, Haggerty CJ, Cunningham BW, Kaneda K, McAfee PC. An in vitro study evaluating the stability of occipitocervical reconstruction techniques. Anaheim, CA: Orthopaedic Research Society; 1999.
173. Dickman CA, Crawford NR, Paramore CG. Biomechanical characteristics of C1–2 cable fixations. J Neurosurgery 1996;85: 316–322
174. Jun B-Y. Anatomic study for the ideal and safe posterior C1-C2 transarticular screw fixation. Spine 1998;23:1703–1707.
175. Goel A, Desai KI, Muzumdar DP. Atlantoaxial fixation using plate and screw method: A report of 160 treated patients. Neurosurgery 2002;51:1351–1357.
176. Goel A, Laheri V. Plate and screw fixation for atlanto-axial subluxation. Acta Neurochir (Wien) 1994;129:47–53.
177. Kuroki H, Rengachary S, Goel V, Holekamp S, Pitkänen V, Ebraheim N. Biomechanical comparison of two stabilization techniques of the atlantoaxial joints – transarticular screw fixation versus screw and rod fixation. Operative Neurosurgery 2005;56:ONS 151–159.
178. Melcher RP, Puttlitz CM, Kleinstueck FS, Lotz JC, Harms J, Bradford DS. Biomechanical testing of posterior atlantoaxial fixation techniques. Spine 2002;27:2435–2440.
179. Richter M, Schmidt R, Claes L, Puhl W, Wilke H. Posterior atlantoaxial fixation. Biomechanical in vitro comparison of six different techniques. Spine 2002;27:1724–1732.
180. Lynch JJ, Crawford NR, Chamberlain RH, Bartolomei JC, Sonntag VKH. Biomechanics of lateral mass/pedicle screw fixation at C1-2. Presented at the 70th Annual Meeting of the Neurosurgical Society of America, Abstract No. 02, April 21–24, 2002.
181. Harms J, Melcher RP. Posterior C1-C2 fusion with polyaxial screw and rod fixation. Spine 2001;26:2467–2471.
182. Naderi S, Crawford NR, Song GS, Sonntag VKH, Dickman CA. Biomechanical comparison of C1-C2 posterior fixations: Cable, graft, and screw combinations. Spine 1998;23:1946–1956.
183. Paramore CG, Dickman CA, Sonntag VKH. The anatomical suitability of the C1-2 complex for transarticular screw fixation. J Neurosurg 1996;85:221–224.
184. Henriques T, Cunningham BW, Olerud C, Shimamoto N, Lee GA, Larsson S, McAfee PA. Biomechanical comparison of five different atlantoaxial posterior fixation techniques. Spine 2000;25:2877–2883.
185. Bell G, Bailey S. Anterior cervical fusion for trauma. Clin Orthop 1977;128:155–158.
186. Stauffer E, Kelly E. Fracture dislocations of the cervical spine: instability and recurrent deformity following treatment by anterior interbody fusion. J Bone Joint Surg 1977; 59A:45–48.
187. Douglas R, Hebert M, Zdeblick T. Radiographic comparison of plated versus unplated fusions for single ACDF. 26th

Annual Meeting of the Cervical Spine Research Society; Atlanta (GA): 1998.
188. McDonough P, Wang J, Endow K, Kanim L, Delamarter R. Single-level anterior cervical discectomy: plate vs. no plate. 26th Annual Meeting of the Cervical Spine Research Society; Atlanta (GA): 1998.
189. Bolesta M, Rechtine G. Three and four level anterior cervical discectomy and fusion with plate fixation: a prospective study. 26th Annual Meeting of the Cervical Spine Research Society; Atlanta (GA): 1998.
190. Doh E, Heller J. Multi-level anterior cervical reconstruction: comparison of surgical techniques and results. 26th Annual Meeting of the Cervical Spine Research Society; Atlanta (GA): 1998.
191. Panjabi M, Isomi T, Wang J. Loosening at screw-bone junction in multi–level anterior cervical plate construct. 26th Annual Meeting of the Cervical Spine Research Society; Atlanta (GA): 1998.
192. Swank M, Lowery G, Bhat A, McDonough R. Anterior cervical allograft arthrodesis and instrumentation: multilevel interbody grafting or strut graft reconstruction. Eur Spine J 1997;6(2):138–143.
193. Clausen J, et al. Biomechanical evaluation of Casper and Cervical Spine Locking Plate systems in a cadaveric model. J Neurosurg 1996;84:1039–1045.
194. Schulte K, Clark C, Goel V. Kinematics of the cervical spine following discectomy and stabilization. Spine 1989;14:1116–1121.
195. Traynelis V, Donaher P, Roach R, Kojimoto H, Goel V. Biomechanical comparison of anterior caspar plate and three-level posterior fixation techniques in a human cadaveric model. J Neurosur 1993;79:96–103.
196. Chen I. Biomechanical evaluation of subcortical versus bicortical screw purchase in anterior cervical plating. Acta Neurochirurgica 1996;138:167–173.
197. Ryken T, Clausen J, Traynelis V, Goel V. Biomechanical analysis of bone mineral density, insertion technique, screw torque, and holding strength of anterior cervical plate screws. J Neurosur 1995;83:324–329.
198. Wang J, Panjabi M, Isomi T. Higher bone graft force helps in stabilizing anterior cervical multilevel plate system. in 26th Annual Meeting of the Cervical Spine Research Society; Atlanta (GA): 1998.
199. Foley K, DiAngelo D, Rampersaud Y, Vossel K, Jansen T. The in vitro effects of instrumentation on multilevel cervical strut-graft mechanics. Spine 1999;24(22):2366–2376.
200. Yoganandan Y. Personal communication.
201. Coe J, Warden K, Sutterlin C, McAfee P. Biomechanical evaluation of cervical spinal stabilization methods in a human cadaveric model. Spine 1989;14:1122–1131.
202. Scifert J, Goel V, Smith D, Traynelis V. In vitro biomechanical comparison of a posterior plate versus facet wiring in quasi-static and cyclic modes. in 44th Annual Meeting of the Orthopaedic Research Society; New Orleans (LA): 1998.
203. Raftopoulos D, et al. Comparative stability of posterior cervical plates and interfacet fusion. Proc Orthop Res Soc 1991;16:630.
204. Gill K, Paschal S, Corin J, Ashman R, Bucholz R. Posterior plating of the cervical spine: a biomechanical comparison of different posterior fusion techniques. Spine 1988;13:813–816.
205. Lim T, An H, Koh Y, McGrady L. A biomechanical comparison between modern anterior versus posterior plate fixation of unstable cervical spine injuries. J Biomed Engr 1997;36:217–218.
206. Jonsson H, Cesarini K, Petren-Mallmin M, Rauschning W. Locking screw-plate fixation of cervical spine fractures with and without ancillary posterior plating. Arch Orthop Trauma Surg 1991;111:1–12.
207. Hacker R, Eugene O, Cauthen J, Gilbert T. Prospective randomized multi-center clinical trial of cervical fusion cages. in 14th Annual Meeting of the North American Spine Society; Chicago Hilton and Towers: 1999.
208. Yang K, Fruth I, Trantrizos A, Steffen T. Biomechanics of anterior cervical interbody fusion cages. in 14th Annual Meeting of the North American Spine Society Chicago Hilton and Towers: 1999.
209. Goel VK, Dick D, Kuroki H, Ebraheim N. Expulsion Resistance Of the RHAKOSS^{TM} C Spinal Implant In a Cadaver Model. 2002, Internal Report, University of Toledo (OH).
210. Totoribe K, Matsumoto M, Goel VK, Yang S-J, Tajima N, Shikinami Y. Comparative biomechanical analysis of a cervical cage made of unsintered hydroxyapatite particles/poly-L-lactide composite in a cadaver model. Spine 2003;28:1010–1015.
211. Gwon JK, Chen J, Lim TH, et al. In vitro comparative biomechanical analysis of transpedicular screw instrumentations in the lumbar region of the human spine. J Spine Disord 1991;4:437.
212. Rohlmann A, Calisse J, Bergmann G, Weber U, Aebi M. Internal spinal fixator stiffness has only a minor influence on stresses in the adjacent disks. Spine 24(12):1192.
213. Weinhoffer SL, et al. Intradiscal pressure measurements above an instrumented fusion–a cadaveric study. Spine 1995;20:526.
214. Lim T-H, et al. Biomechanical evaluation of diagonal fixation in pedicle screw instrumentation. Spine 2001;26(22):2498–2503.
215. Glazer PA, et al. Biomechanical analysis of multilevel fixation methods in the lumbar spine. Spine 1997;22:171.
216. Heller JG, Zdeblick TA, Kunz DA, et al. Spinal instrumentation for metastatic disease: In vitro biomechanical analysis. J Spinal Disord 1993;6:17.
217. Vaccaro AR, Chiba K, Heller JG, Patel TCh, Thalgott JS, Truumees E, Fischgrund JS, Craig MR, Berta SC, Wang JC. Bone grafting alternatives in spinal surgery. Spine J 2002;2:206–215.
218. Hitchon PW, et al. In vitro biomechanical analysis of three anterior thoracolumbar implants. J Neurosurgery – Spine 2 2000;93:252–258.
219. Vahldiek MJ, Panjabi MM. Stability potential of spinal instrumentations in tumor vertebral body replacement surgery. Spine 1998;23(5):543–550.
220. Oda I, Cunningham BW, Lee GA, Abumi K, Kaneda K, McAfee P, Mow VC, Hayes WC. Biomechanical properties of anterior throacolumbar multisegmental fixation–An analysis of construct stiffness and screw-rod strain. Spine 2000;25(8):2303–2311.
221. Shirado O, et al. Quantitative histologic study of the influence of anterior spinal instrumentation and biodegradable polymer on lumbar interbody fusion after corpectomy–a canine model. Spine 1992;17:795.
222. Heth JA, Hitchon PW, Goel VK, Rogge TN, Drake JS, Torner JC. A biomechanical comparison between anterior and transverse interbody fusion cages. Spine 2001;26:E261–E267.
223. Wang S-T, Goel VK, Fu T-U, Kubo S, Choi Woosung, Liu C-L, Tain-Hsiung Chen T-S. Posterior instrumentation reduces differences in spine stability due to different cage orientations – an in vitro study. Spine 2005;30:62–67.
224. Kanayama M, Cunningham BW, Haggerty CJ, Abumi K, Kaneda K. McAfee PC. In vitro biomechanical investigation

of the stability and stress-shielding effect of lumbar interbody fusion devices. J Neurosurg 2000;93(2 Suppl):259–265.
225. Pitzen T, Geisler FH, Matthis D, Muller-Storz H, Steudel WI. Motion of threaded cages in posterior lumbar interbody fusion. Eur Spine J 2000;9(6):571–576.
226. Kim Y. Prediction of Mechanical Behaviors at interfaces between bone and two interbody cages of lumbar spine segments. Spine 26(13):1437–1442.
227. Volkman T, Horton WC, Hutton WC. Transfacet screws with lumbar interbody reconstruction: Biomechanical study of motion segment stiffness. J Spinal Disord 1996;9(5):425–432.
228. Patwardhan A, Carandang G, Ghanayem A, et al. Compressive preload improves the stability of the anterior lumbar interbody fusion (ALIF) cage construct. J Bone Joint Surg Am 2003;85-A:1749–1756.
229. Smith KR, Hunt TR, Asher MA, et al. The effect of a stiff spinal implant on the bone mineral content of the lumbar spine in dogs. J Bone Joint Surg Am 1991;73:115.
230. Spivak JM, Neuwirth MG, Labiak JJ, et al. Hydroxyapatite enhancement of posterior spinal instrumentation fixation. Spine 1994;19:955.
231. McAfee PC, Farey ID, Sutterlin CE, et al. Device–related osteoporosis with spinal instrumentation. Spine 1989;14: 919.
232. McAfee PC, Farey ID, Sutterlin CE, et al. The effect of spinal implant rigidity on vertebral bone density: A canine model. Spine 1991;16:S190.
233. McAfee PC, Lubicky JP, Werner FW. The use of segmental spinal instrumentation to preserve longitudinal spinal growth–an experimental study. J Bone Joint Surg Am 1983;65:935.
234. Penta M. Sandhu A, Fraser RD. Magnetic resonance imaging assessment of disk degeneration 10 years after anterior lumbar interbody fusion. Spine 1995;20:743.
235. Kanayama M, Cunningham BW, Sefter JC, Goldstein JA, Stewart G, Kaneda K, McAfee PC. Does spinal instrumentation influence the healing process of posterolateral spinal fusion? An in vivo animal model. Spine 1999;24(11):1058–1065.
236. Kanayama M, Cunningham BW, Weis JC, et al. Maturation of the posterolateral fusion and its effect on load-sharing of spinal instrumentation. J Bone Joint Surg Am 1997;79: 1710.
237. Rohlmann A, Bergmann G, Graichen F. A spinal fixation device for in vivo load measurement. J Biomech 1994;27: 961.
238. Rohlmann A, Bergmann G, Graichen F, et al. Comparison of loads on internal spinal fixation devices measured *in vitro* and *in vivo*. Med Eng Phys 1997;19:539.
239. Rohlmann A, Graichen F, Weber U, Bergmann G. Biomechanical studies monitoring in vivo implant loads with a telemeterized internal spinal fixation device. Spine 2000; 25(23):2981–2986.
240. Goel VK, Konz RJ, Chang HT, et al. Load sharing comparison of a hinged vs. a rigid screw device in the stabilized lumbar motion segment: A finite element study. J Prosth Orthotoics 2002.
241. Hitchon PW, Goel VK, Rogge T, Grosland NM, Sairyo K, Torner J. Biomechanical studies of a dynamized anterior thoracolumbar implant. Spine 2000;25(3):306–309.
242. Scifert J, Sairyo K, Goel VK, Grobler LJ, Grosland NM, Spratt KF, Chesmel KD. Stability analysis of an enhanced load sharing posterior fixation device and its equivalent conventional device in a calf spine model. Spine 1999;24: 2206–2213.
243. Vishnubotla S, Goel VK, Walkenhorst J, Boyd LM, Vadapalli S, Shaw MN. Dyanmic fixation systems compared to the rigid spinal instrumentation–a finite element investigation. Presented at the 24th Annual meeting of the American Society of Biomechanics; Portand (OR): Sep 11–13, 2004.
244. Dooris AP. Experimental and Theoretical Investigations into the Effects of Artificial Disk Implantation on the Lumbar Spine, Ph.D. dissertation, University of Iowa, Iowa City (IA): 2001.
245. Dooris AP, Goel VK, Grosland NM, Gilbertson LG, Wilder DG. Load-sharing between anterior and posterior elements in a lumbar motion segment implanted with an artificial disk. Spine 2001;26(6):E122–E129.
246. Goel V, Dooris A, Grosland N, Drake J, Coppes J, Ahern A, Wolfe S, Roche K. Biomechanics of a lumbar spine segment stabilized using morselized bone as an interbody graft. 27th Annual Meeting International Society for the Study of the Lumbar Spine; Adelaide, Australia: April 9–13, 2000.
247. Boden SD, Schimandle JH. Biological enhancement of spinal fusion. Spine 1995;20:113S.
248. Bodern SD, Sumner DR. Biologic factors affecting spinal fusion and bone regeneration. Spine 1995;20:1029.
249. Grauer JN, Patel TC, Erulkar JS, Troiano NW, Panjabi MM, Friedlaender GE. Evaluation of OP-1 as a graft substitute for intertransverse process lumbar fusion. Spine 2001; 26(2):127–133.
250. Patel TC, Jonathan S. Erulkar JS, Jonathan N, Grauer JN, Nancy W, Troiano NW, Panjabi MM, Friedlaender GE. Osteogenic protein-1 overcomes the inhibitory effect of nicotine on posterolateral lumbar fusion. Spine 2001;26(15): 1656–1661.
251. Goel VK, Grosland NM, Todd DT, et al. Application of finite element models to predict clinically relevant biomechanics of lumbar spine. Semin Surg 1998;10:112.
252. Goel VK, Kim YE. Effects of injury on the spinal motion segment mechanics in the axial compression mode. Clin Biomech 1989;4:161–167.
253. Goel VK, Kim TE, Lim TH, et al. An analytical investigation of the mechanics of spinal instrumentation. Spine 1988; 13:1003.
254. Goel VK, Kong WZ, Han JS, Weinstein JN, Gilbertson LG. A combined finite element and optimization investigation of lumbar spine mechanics with and without muscles. Spine 1993;18:1531–1541.
255. Goel VK, Lim TH, Gilbertson LG, et al. Clinically relevant finite element models of a ligamentous lumbar motion segment. Sem Spine Surg 1993;5:29.
256. Goel VK, Lim TH, Gwon J, et al. Effects of rigidity of an internal fixation device–a comprehensive biomechanical investigation. Spine 1991;16:S155.
257. Grosland NM, Goel VK, Grobler LJ, et al. Adaptive internal bone remodeling of the vertebral body following an anterior interbody fusion: A computer simulation. The 24th International Society for the Study of the Lumbar Spine, Singapore, Singapore, June 3–6, 1997.
258. Klara PM, Ray CD. Artificial nucleus replacement clinical experience. Spine 27(12):1374–1377.
259. Ray CD, Corbin TP. Prosthetic disk and method of implanting. US pat 4,772,287;1990.
260. Dooris A, Hudgin G, Goel V, Bao C. Restoration of normal multisegment biomechanics with prosthetic intervertebral disk. 48th Annual Meeting, Orthopedic Research Society; Dallas, TX: Feb. 10–13, 2002.
261. Goel VK, Grauer J, Patel TG, Biyani A, Sairyo K, Vishnubhotla S, Matyas A, Cowgill I, Shaw M, Long R, Dick D, Panjabi MM, Serhan H. Effects of charite artificial disc on the implanted and adjacent spinal segments mechanics using a hybrid testing protocol. Spine (Accepted)

262. Lee CK, et al. Development of a prosthetic intervertebral disk. Spine 1991;16:S253.
263. Garfin SR, Yuan Hansen A, Reiley Mark A. New technologies in spine kyphoplasty and vertebroplasty for the treatment of painful osteoporotic compression fractures. Spine 2001;26(14):1511–1515.
264. Hedman TP, et al. Design of intervertebral disk prosthesis. Spine 1991;16:S256.
265. Lim T-H, T. Brebach T, Renner SM, Kim W-J, Kim JG, Lee RE, Andersson GBJ, An HS. Biomechanical evaluation of an injectable calcium phosphate cement for vertebroplasty. Spine 2002;27(12):1297–1302.
266. Belkof SM, John M. Mathis JM, Erik M. Erbe EM, Fenton C. Biomechanical evaluation of a new bone cement for use in vertebroplasty. Spine 2000;25(9):1061–1064.
267. Liebschner MAK, Rosenberg WS, Keaveny TM. Effects of bone cement volume and distribution on vertebral stiffnes after vertebroplasty. Spine 2001;26:1547–1554.
268. Hitchon PW, Goel V, Drake J, Taggard D, Brenton M, Rogge T, Torner JC. A Biomechanical comparison of Hydroxypatite and Polymethylmethacrylate Vertebroplasty in a Cadaveric Spinal Compression Fracture Model. J Neurosug (Spine 2) 2001;95:215–220.
269. Huntzinger J, Phares T, Goel V, Fournier R, Kuroki H, McGowan D. The effect of concentration on polymer scaffolds for bioartificial intervertebral disks. 49th Annual Meeting, Orthopedic Research Society; New Orleans (LA): Feb 2–5 2003.
270. Goel V, Miller S, Navarro R, Price J, Ananthan R, Matyas A, Dick D, Yuan H. Restoration of physiologic disk biomechanics with a telemeterized natural motion elastomer disk. SAS4, Vienna, Austria, May 4–8, 2004.

Reading List

Bao QB, et al. The artificial disk: Theory, design and materials. Biomaterials 1996;17:1157.
Langrana NA, Lee CK, Yang SW. Finite element modeling of the synthetic intervertebral disk. Spine 1991;16:S245.

See also BONE AND TEETH, PROPERTIES OF; LIGAMENT AND TENDON, PROPERTIES OF; SCOLIOSIS, BIOMECHANICS OF; SPINAL IMPLANTS.

ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition

VOLUME 4

Hydrocephalus, Tools for Diagnosis and Treatment of – Monoclonal Antibodies

ENCYCLOPEDIA OF MEDICAL DEVICES AND INSTRUMENTATION, SECOND EDITION

ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition
Volume 4

Hydrocephalus, Tools for Diagnosis and Treatment of – Monoclonal Antibodies

Edited by

John G. Webster
University of Wisconsin–Madison

The *Encyclopedia of Medical Devices and Instrumentation* is available online at http://www.mrw.interscience.wiley.com/emdi

WILEY-INTERSCIENCE

A John Wiley & Sons, Inc., Publication

Published by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Encylopedia of medical devices & instrumentation/by John G. Webster,

editor in chief. – 2nd ed.
p. ; cm.
Rev. ed. of: Encyclopedia of medical devices and instrumentation. 1988.
Includes bibliographical references and index.
ISBN-13 978-0-471-26358-6 (set : cloth)
ISBN-10 0-471-26358-3 (set : cloth)
ISBN-13 978-0-470-04069-0 (v. 4 : cloth)
ISBN-10 0-470-04069-6 (v. 4 : cloth)

1. Medical instruments and apparatus–Encyclopedias. 2. Biomedical engineering–Encyclopedias. 3. Medical physics–Encyclopedias. 4. Medicine–Data processing–Encyclopedias. I. Webster, John G., 1932- . II. Title: Encyclopedia of medical devices and instrumentation.

[DNLM: 1. Equipment and Supplies–Encyclopedias–English. W 13

E555 2006]
R856.A3E53 2006
610.2803–dc22
2005028946

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTRIBUTOR LIST

ABDEL HADY, MAZEN, *McMaster University, Hamilton, Ontario Canada*, Bladder Dysfunction, Neurostimulation of

ABEL, L.A., *University of Melbourne, Melbourne, Australia*, Ocular Motility Recording and Nystagmus

ABREU, BEATRIZ C., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

ALEXANDER, A.L., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

ALI, ABBAS, *University of Illinois, at Urbana-Champaign*, Bioinformatics

ALI, MÜFTÜ, *School of Dental Medicine, Boston, Massachusetts*, Tooth and Jaw, Biomechanics of

ALPERIN, NOAM, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

ANSON, DENIS, *College Misericordia, Dallas, Pennsylvania*, Environmental Control

ARENA, JOHN C., *VA Medical Center and Medical College of Georgia*, Biofeedback

ARIEL, GIDEON, *Ariel Dynamics, Canyon, California*, Biomechanics of Exercise Fitness

ARMSTRONG, STEVE, *University of Iowa, Iowa City, Iowa*, Biomaterials for Dentistry

ASPDEN, R.M., *University of Aberdeen, Aberdeen, United Kingdom*, Ligament and Tendon, Properties of

AUBIN, C.E., *Polytechniquie Montreal, Montreal Quebec, Canada*, Scoliosis, Biomechanics of

AYRES, VIRGINIA M., *Michigan State University, East Lansing, Michigan*, Microscopy, Scanning Tunneling

AZANGWE, G., Ligament and Tendon, Properties of

BACK, LLOYD H., *California Institute of Technology, Pasadena, California*, Coronary Angioplasty and Guidewire Diagnostics

BADYLAK, STEPHEN F., *McGowan Institute for Regenerative Medicine, Pittsburgh, Pennsylvania*, Sterilization of Biologic Scaffold Materials

BANDYOPADHYAY, AMIT, *Washington State University, Pullman, Washington*, Orthopedic Devices, Materials and Design for

BANERJEE, RUPAK K., *University of Cincinnati, Cincinnati, Ohio*, Coronary Angioplasty and Guidewire Diagnostics

BARBOUR, RANDALL L., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

BARKER, STEVEN J., *University of Arizona, Tucson, Arizona*, Oxygen Monitoring

BARTH, ROLF F., *The Ohio State University, Columbus, Ohio*, Boron Neutron Capture Therapy

BECCHETTI, F.D., *University of Michigan, Ann Arbor, Michigan*, Radiotherapy, Heavy Ion

BELFORTE, GUIDO, *Politecnico di Torino – Department of Mechanics*, Laryngeal Prosthetic Devices

BENKESER, PAUL, *Georgia Institute of Technology, Atlanta, Georgia*, Biomedical Engineering Education

BENNETT, JAMES R., *University of Iowa, Iowa City, Iowa*, Digital Angiography

BERSANO-BEGEY, TOMMASO, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

BIGGS, PETER J., *Harvard Medical School, Boston, Massachusetts*, Radiotherapy, Intraoperative

BIYANI, ASHOK, *University of Toledo, and Medical College of Ohio, Toledo, Ohio*, Human Spine, Biomechanics of

BLOCK, W.F., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

BLUE, THOMAS E., *The Ohio State University, Columbus, Ohio*, Boron Neutron Capture Therapy

BLUMSACK, JUDITH T., *Disorders Auburn University, Auburn, Alabama*, Audiometry

BOGAN, RICHARD K., *University of South Carolina, Columbia, South Carolina*, Sleep Laboratory

BOKROS, JACK C., *Medical Carbon Research Institute, Austin, Texas*, Biomaterials, Carbon

BONGIOANNINI, GUIDO, *ENT Division Mauriziano Hospital, Torino, Italy*, Laryngeal Prosthetic Devices

BORAH, JOSHUA, *Applied Science Laboratories, Bedford, Massachusetts*, Eye Movement, Measurement Techniques for

BORDEN, MARK, *Director of Biomaterials Research, Irvine, California*, Biomaterials, Absorbable

BORTON, BETTIE B., *Auburn University Montgomery, Montgomery, Alabama*, Audiometry

BORTON, THOMAS E., *Auburn University Montgomery, Montgomery, Alabama*, Audiometry

BOSE SUSMITA,, *Washington State University, Pullman, Washington*, Orthopedic Devices, Materials and Design for

BOVA, FRANK J., *M. D. Anderson Cancer Center Orlando, Orlando, FL*, Radiosurgery, Stereotactic

BRENNER, DAVID J., *Columbia University Medical Center, New York, New York*, Computed Tomography Screening

BREWER, JOHN M., *University of Georgia*, Electrophoresis

BRIAN, L. DAVIS, *Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, Ohio*, Skin, Biomechanics of

BRITT, L.D., *Eastern Virginia Medical School, Norfolk, Virginia*, Gastrointestinal Hemorrhage

BRITT, R.C., *Eastern Virginia Medical School, Norfolk, Virginia*, Gastrointestinal Hemorrhage

BROZIK, SUSAN M., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

BRUNER, JOSEPH P., *Vanderbilt University Medical Center, Nashville, Tennessee*, Intrauterine Surgical Techniques

BRUNSWIG NEWRING, KIRK A., *University of Nevada, Reno, Nevada*, Sexual Instrumentatio n

BRUYANT, PHILIPPE P., *University of Massachusetts, North Worcester, Massachusetts*, Nuclear Medicine, Computers in

BUNNELL, BERT J., *Bunnell Inc., Salt Lake City, Utah*, High Frequency Ventilation

CALKINS, JERRY M., *Defense Research Technologies, Inc., Rockville, Maryland*, Medical Gas Analyzers

CANNON, MARK, *Northwestern University, Chicago, Illinois*, Resin-Based Composites

CAPPELLERI, JOSEPH C., *Pfizer Inc., Groton, Connecticut*, Quality-of-Life Measures, Clinical Significance of

CARDOSO, JORGE, *University of Madeira, Funchal, Portugal*, Office Automation Systems

CARELLO, MASSIMILIANA, *Politecnicodi Torino – Department of Mechanics*, Laryngeal Prosthetic Devices

CASKEY, THOMAS C., *Cogene Biotech Ventures, Houston, Texas*, Polymerase Chain Reaction

CECCIO, STEVEN, *University of Michigan, Ann Arbor, Michigan*, Heart Valve Prostheses, In Vitro Flow Dynamics of

CHAN, JACKIE K., *Columbia University, New York, New York*, Photography, Medical

CHANDRAN, K.B., *University of Iowa, Iowa City, Iowa*, Heart Valve Prostheses

CHATZANDROULIS, S., *NTUA, Athens, Attiki, Greece*, Capacitive Microsensors for Biomedical Applications

CHAVEZ, ELIANA, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CHEN, HENRY, *Stanford University, Palo Alto, California*, Exercise Stress Testing

CHEN, JIANDE, *University of Texas Medical Branch, Galveston, Texas*, Electrogastrogram

CHEN, YAN, *Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, Ohio*, Skin, Biomechanics of

CHEYNE, DOUGLAS, *Hospital for Sick Children Research Institute*, Biomagnetism

CHUI, CHEN-SHOU, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

CLAXTON, NATHAN S., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

CODERRE, JEFFREY A., *Massachus etts Institute of Technology, Cambridge, Massachusetts*, Boron Neutron Capture Therapy

COLLINS, BETH, *University of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

COLLINS, DIANE, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CONSTANTINOU, C., *Columbia University Radiation Oncology, New York, New York*, Phantom Materials in Radiology

COOK, ALBERT, *University of Alberta*, *Edmonton, Alberta, Canada*, Communication Devices

COOPER, RORY, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CORK, RANDALL C., *Louisiana State University, Shreveport, Louisiana*, Monitoring, Umbilical Artery and Vein, Blood Gas Measurements; Transcuta neous Electrical Nerve Stimulation (TENS); Ambulatory Monitoring

COX, JOSEPHINE H., *Walter Reed Army Institute of Research, Rockville, Maryland*, Blood Collection and Processing

CRAIG, LEONARD, *Feinberg School of Medicine of Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

CRESS, CYNTHIA J., *University of Nebraska, Lincoln, Nebraska*, Communicative Disorders, Computer Applications for

CUMMING, DAVID R.S., *University of Glasgow, Glasgow, United Kingdom*, Ion-Sensitive Field-Effect Transistors

CUNNINGHAM, JOHN R., *Camrose, Alberta, Canada*, Cobalt 60 Units for Radiotherapy

D'ALESSANDRO, DAVID, *Montefiore Medical Center, Bronx, New York*, Heart–Lung Machines

D'AMBRA, MICHAEL N., *Harvard Medical School, Cambridge, Massachusetts*, Cardiac Output, Thermodilution Measurement of

DADSETAN, MAHROKH, *Mayo Clinic, College of Medicine, Rochester, Minnesota*, Microscopy, Electron

DALEY, MICHAEL L., *The University of Memphis, Memphis, Tennessee*, Monitoring, Intracranial Pressure

DAN, LOYD, *Linköping University, Linköping, Sweden*, Thermocouples

DAS, RUPAK, *University of Wisconsin, Madison, Wisconsin*, Brachytherapy, High Dosage Rate

DATTAWADKAR, AMRUTA M., *University of Wisconsin, Madison, Madison, Wisconsin*, Ocular Fundus Reflectometry

DAVIDSON, MICHAEL W., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

DE LUCA, CARLO, *Boston University, Boston, Massachusetts*, Electromyography

DE SALLES, ANTONIO A.F., *UCLA Medical School, Los Angeles, California*, Stereotactic Surgery

DECAU, SABIN, *University of Maryland, School of Medicine*, Shock, Treatment of

DECHOW, PAUL C., *A & M University Health Science Center, Dallas, Texas*, Strain Gages

DELBEKE, JEAN, *Catholique University of Louvain, Brussels, Belgium*, Visual Prostheses

DELL'OSSO, LOUIS F., *Case Western Reserve University, Cleveland, Ohio*, Ocular Motility Recording and Nystagmus

DELORME, ARNAUD, *University of San Diego, La Jolla, California*, Statistical Methods

DEMENKOFF, JOHN, *Mayo Clinic, Scottsdale, Arizona*, Pulmonary Physiology

DEMIR, SEMAHAT S., *The University of Memphis and The University of Tennessee Health Science Center, Memphis, Tennessee*, Electrophysiology

DEMLING, ROBERT H., *Harvard Medical School*, Skin Substitute for Burns, Bioactive

DENNIS, MICHAEL J., *Medical University of Ohio, Toledo, Ohio*, Computed Tomography

DESANTI, LESLIE, *Harvard Medical School*, Skin Substitute for Burns, Bioactive

DEUTSCH, STEVEN, *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

DEVINENI, TRISHUL, *Conemaugh Health System*, Biofeedback

DI BELLA EDWARD, V.R., *University of Utah*, Tracer Kinetics

DIAKIDES, NICHOLAS A., *Advanced Concepts Analysis, Inc., Falls Church, Virginia*, Thermography

DOLAN, PATRICIA L., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

DONOVAN, F.M., *University of South Alabama*, Cardiac Output, Indicator Dilution Measurement of

DOUGLAS, WILSON R., *Children's Hospital of Philadelphia, Philadelphia, Pennsylvania*, Intrauterine Surgical Techniques

DRAPER, CRISSA, *University of Nevada, Reno, Nevada*, Sexual Instrumentation

DRZEWIECKI, TADEUSZ M., *Defense Research Technologies, Inc., Rockville, Maryland*, Medical Gas Analyzers

DURFEE, W.K., *University of Minnesota, Minneapolis, Minnesota*, Rehabilitation and Muscle Testing

DYRO, JOSEPH F., *Setauket, New York*, Safety Program, Hospital

DYSON, MARY, *Herts, United Kingdom*, Heat and Cold, Therapeutic

ECKERLE, JOSEPH S., *SRI International, Menlo Park, California*, Tonometry, Arterial

EDWARDS, BENJAMIN, *University of Wisconsin-Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

EDWARDS, THAYNE L., *University of Washington, Seattle, Washington*, Chromatography

EKLUND, ANDERS, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

EL SOLH, ALI A., *Erie County Medical Center, Buffalo, New York*, Sleep Studies, Computer Analysis of

ELMAYERGI, NADER, *McMaster University, Hamilton, Ontario, Canada*, Bladder Dysfunction, Neurostimulation of

ELSHARYDAH, AHMAD, *Louisiana State University, Baton Rouge, Louisiana*, Ambulatory Monitoring; Monitoring, Umbilical Artery and Vein, Blood Gas Measurements

FADDY, STEVEN C., *St. Vincents Hospital, Sydney, Darlinghurst, Australia*, Cardiac Output, Fick Technique for

FAHEY, FREDERIC H., *Childrens Hospital Boston*, Computed Tomography, Single Photon Emission

FAIN, S.B., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

FELDMAN, JEFFREY, *Childrens Hospital of Philadelphia, Philadelphia, Pennsylvania*, Anesthesia Machines

FELLERS, THOMAS J., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

FERRARA, LISA, *Cleveland Clinic Foundation, Cleveland, Ohio*, Human Spine, Biomechanics of

FERRARI, MAURO, *The Ohio State University, Columbus, Ohio*, Drug Delivery Systems

FONTAINE, ARNOLD A., *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

FOUST, MILTON J., JR, *Medical University of South Carolina Psychiatry and Behavioral Sciences, Charleston, South Carolina*, Electroconvulsive Therapy

FRASCO, PETER, *Mayo Clinic Scottsdale, Scottsdale, Arizona*, Temperature Monitoring

FRAZIER, JAMES, *Louisiana State University, Baton Rouge, Louisiana*, Ambulatory Monitoring

FREIESLEBEN DE BLASIO, BIRGITTE, *University of Oslo, Oslo, Norway*, Impedance Spectroscopy

FRESTA, MASSIMO, *University of Catanzaro Magna Græcia, Germaneto (CZ), Italy*, Drug Delivery Systems

FREYTES, DONALD O., *McGowan Institute for Regenerative Medicine, Pittsburgh Pennsylvania*, Sterilization of Biologic Scaffold Materials

FROEHLICHER, VICTOR, *VA Medical Center, Palo Alto, California*, Exercise Stress Testing

FUNG, EDWARD K., *Columbia University, New York, New York*, Photography, Medical

GAGE, ANDREW A., *State University of New York at Buffalo, Buffalo, New York*, Cryosurgery

GAGLIO, PAUL J., *Columbia University College of Physicians and Surgeons*, Liver Transplantation

GARDNER, REED M., *LDS Hospital and Utah University, Salt Lake City, Utah*, Monitoring, Hemodynamic

GEJERMAN, GLEN, *Hackensack University Medical, Hackensack, New Jersey*, Radiation Therapy, Quality Assurance in

GEORGE, MARK S., *Medical University of South Carolina Psychiatry and Behavioral Sciences, Charleston, South Carolina*, Electroconvulsive Therapy

GHARIEB, R.R., *Infinite Biomedical Technologies, Baltimore, Maryland*, Neurological Monitors

GLASGOW, GLENN P., *Loyola University of Chicago, Maywood, Illinois*, Radiation Protection Instrumentation

GLASGOW, GLENN, *University of Wisconsin-Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

GOEL, VIJAY K., *University of Toledo, and Medical College of Ohio, Toledo, Ohio*, Human Spine, Biomechanics of

GOETSCH, STEVEN J., *San Diego Gamma Knife Center, La Jolla, California*, Gamma Knife

GOLDBERG, JAY R., *Marquette University Milwaukee, Wisconsin*, Minimally Invasive Surgery

GOLDBERG, ZELENNA, *Department of Radiation Oncology, Davis, California*, Ionizing Radiation, Biological Effects of

GOPALAKRISHNAKONE, P., *National University of Singapore, Singapore*, Immunologically Sensitive Field-Effect Transistors

GOPAS, JACOB, *Ben Gurion University of the Negev, Beer Sheva, Israel*, Monoclonal Antibodies

GORGULHO, ALESSANDRA, *UCLA Medical School, Los Angeles, California*, Stereotactic Surgery

GOUGH, DAVID A., *University of California, La Jolla, California*, Glucose Sensors

GOUSTOURIDIS, D., *NTUA, Athens, Attiki, Greece*, Capacitive Microsensors for Biomedical Applications

GRABER, HARRY L., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

GRAÇA, M., *Louisiana State University, Baton Rouge, Louisiana*, Boron Neutron Capture Therapy

GRANT, WALTER III, *Baylor College of Medicine, Houston, Texas*, Radiation Therapy, Intensity Modulated

GRAYDEN, EDWARD, *Mayo Health Center, Albertlea, Minnesota*, Cardiopulmonary Resuscitation

GREEN, JORDAN R., *University of Nebraska, Lincoln, Nebraska*, Communicative Disorders, Computer Applications for

HAEMMERICH, DIETER, *Medical University of South Carolina, Charleston, South Carolina*, Tissue Ablation

HAMAM, HABIB, *Université de Moncton, Moncton New Brunswick, Canada*, Lenses, Intraocular

HAMMOND, PAUL A., *University of Glasgow, Glasgow, United Kingdom*, Ion-Sensitive Field-Effect Transistors

HANLEY, JOSEPH, *Hackensack University Medical, Hackensack, New Jersey*, Radiation Therapy, Quality Assurance in

HARLEY, BRENDAN A., *Massachusetts Institute of Technology*, Skin Tissue Engineering for Regeneration

HARPER, JASON C., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

HASMAN, ARIE, *Maastricht, The Netherlands*, Medical Education, Computers in

HASSOUNA, MAGDY, *Toronto Western Hospital, Toronto, Canada*, Bladder Dysfunction, Neurostimulation of

HAYASHI, KOZABURO, *Okayama University of Science, Okayama, Japan*, Arteries, Elastic Properties of

HENCH, LARRY L., *Imperial College London, London, United Kingdom*, Biomaterials: Bioceramics

HETTRICK, DOUGLAS A., *Sr. Principal Scientist Medtronic, Inc., Minneapolis, Minnesota*, Bioimpedance in Cardiovascular Medicine

HIRSCH-KUCHMA, MELISSA, *University of Central Florida NanoScience Technology Center, Orlando, Florida*, Biosurface Engineering

HOLDER, GRAHAM E., *Moorfields Eye Hospital, London, United Kingdom*, Electroretinography

HOLMES, TIMOTHY, *St. Agnes Cancer Center, Baltimore, Maryland*, Tomotherapy

HONEYMAN-BUCK, JANICE C., *University of Florida, Gainesville, Florida*, Radiology Information Systems

HOOPER, BRETT A., *Areté Associates, Arlington, Virginia*, Endoscopes

HORN, BRUCE, *Kaiser Permanente, Los Angeles, California*, X-Rays Production of

HORNER, PATRICIA I., *Biomedical Engineering Society Landover, Maryland*, Medical Engineering Societies and Organizations

HOROWITZ, PAUL M., *University of Texas, San Antonio, Texas*, Fluorescence Measurements

HOU, XIAOLIN, *Risø National Laboratory, Roskilde, Denmark*, Neutron Activation Analysis

HOVORKA, ROMAN, *University of Cambridge, Cambridge, United Kingdom*, Pancreas, Artificial

HUANG, H.K., *University of Southern California*, Teleradiology

HUNT, ALAN J., *University of Michigan, Ann Arbor, Michigan*, Optical Tweezers

HUTTEN, HELMUT, *University of Technology, Graz, Australia*, Impedance Plethysmography

IAIZZO, P.A., *University of Minnesota, Minneapolis, Minnesota*, Rehabilitation and Muscle Testing

IBBOTT, GEOFFREY S., *Anderson Cancer Center, Houston, Texas*, Radiation Dosimetry, Three-Dimensional

INGHAM, E., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

ISIK, CAN, *Syracuse University, Syracuse, New York*, Blood Pressure Measurement

JAMES, SUSAN P., *Colorado State University, Fort Collins, Colorado*, Biomaterials: Polymers

JENSEN, WINNIE, *Aalborg University, Aalborg, Denmark*, Electroneurography

JIN, CHUNMING, *North Carolina State University, Raleigh, North Carolina*, Biomaterials, Corrosion and Wear of

JIN, Z.M., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

JOHNSON, ARTHUR T., *University of Maryland College Park, Maryland*, Medical Engineering Societies and Organizations

JONES, JULIAN R., *Imperial College London, London, United Kingdom*, Biomaterials: Bioceramics

JOSHI, ABHIJEET, *Abbott Spine, Austin, Texas*, Spinal Implants

JUNG, RANU, *Arizona State University, Tempe, Arizona*, Functional Electrical Stimulation

JURISSON, SILVIA S., *University of Missouri Columbia, Missouri*, Radionuclide Production and Radioactive Decay

KAEDING, PATRICIA J., *Godfrey & Kahn S.C., Madison, Wisconsin*, Codes and Regulations: Medical Devices

KAMATH, CELIA C., *Mayo Clinic, Rochester, Minnesota*, Quality-of-Life Measures, Clinical Significance of

KANE, MOLLIE, *Madison, Wisconsin*, Contraceptive Devices

KATHERINE, ANDRIOLE P., *Harvard Medical School, Boston, Massachusetts*, Picture Archiving and Communication Systems

KATSAGGELOS, AGGELOS K., *Northwestern University, Evanston, Illinois*, DNA Sequencing

KATZ, J. LAWRENCE, *University of Missouri-Kansas City, Kansas City, Missouri*, Bone and Teeth, Properties of

KESAVAN, SUNIL, *Akebono Corporation, Farmington Hills, Michigan*, Linear Variable Differential Transformers

KHANG, GILSON, *Chonbuk National University*, Biomaterials: Tissue Engineering and Scaffolds

KHAODHIAR, LALITA, *Harvard Medical School, Boston, Massachusetts*, Cutaneous Blood Flow, Doppler Measurement of

KIM, MOON SUK, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

KIM, YOUNG KON, *Inje University, Kimhae City, Korea*, Alloys, Shape Memory

KINDWALL, ERIC P., *St. Luke's Medical Center, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

KING, MICHAEL A., *University of Massachusetts, North Worcester, Massachusetts*, Nuclear Medicine, Computers in

KLEBE, ROBERT J., *University of Texas, San Antonio, Texas*, Fluorescence Measurements

KLEIN, BURTON, *Burton Klein Associates, Newton, Massachusetts*, Gas and Vacuum Systems, Centrally Piped Medical

KNOPER, STEVEN R., *University of Arizona College of Medicine*, Ventilatory Monitoring

KONTAXAKIS, GEORGE, *Universidad Politécnica de Madrid, Madrid, Spain*, Positron Emission Tomography

KOTTKE-MARCHANT, KANDICE, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Vascular Graft Prosthesis

KRIPFGANS, OLIVER, *University of Michigan, Ann Arbor, Michigan*, Ultrasonic Imaging

KULKARNI, AMOL D., *University of Wisconsin–Madison, Madison, Wisconsin*, Ocular Fundus Reflectometry, Visual Field Testing

KUMARADAS, J. CARL, *Ryerson University, Toronto, Ontario, Canada*, Hyperthermia, Interstitial

KUNICKA, JOLANTA, *Bayer HealthCare LLC, Tarrytown, New York*, Differential Counts, Automated

KWAK, KWANJ JOO, *University of Miami Miller School of Medicine, Miami, Florida*, Microscopy, Scanning Force

LAKES, RODERIC, *University of Wisconsin-Madison*, Bone and Teeth, Properties of

LAKKIREDDY, DHANUNJAYA, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

LARSEN, COBY, *Case Western Reserve University, Cleveland, Ohio*, Vascular Graft Prosthesis

LASTER, BRENDA H., *Ben Gurion University of the Negev, Beer Sheva, Israel*, Monoclonal Antibodies

LATTA, LOREN, *University of Miami, Coral Gables, Florida*, Rehabilitation, Orthotics in

LEDER, RON S., *Universidad Nacional Autonoma de Mexico Mexico, Distrito Federal*, Continuous Positive Airway Pressure

LEE, CHIN, *Harvard Medical School, Boston, Massachusetts*, Radiotherapy Treatment Planning, Optimization of; Hyperthermia, Interstitial

LEE, HAI BANG, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

LEE, SANG JIN, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

LEI, LIU, *Department of General Engineering, Urbana, Illinois*, Bioinformatics

LEI, XING, *Stanford University, Stanford, California*, Radiation Dose Planning, Computer-Aided

LEWIS, MATTHEW C., *Medical College of Wisconsin, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

LI, CHAODI, *University of Notre Dame, Notre Dame, Indiana*, Bone Cement, Acrylic

LI, JONATHAN G., *University of Florida, Gainesville, Florida*, Radiation Dose Planning, Computer-Aided

LI, QIAO, *University of Michigan, Ann Arbor, Michigan*, Immunotherapy

LI, YANBIN, *University of Arkansas, Fayetteville, Arkansas*, Piezoelectric Sensors

LIBOFF, A.R., *Oakland University, Rochester, Michigan*, Bone Ununited Fracture and Spinal Fusion, Electrical Treatment of

LIGAS, JAMES, *University of Connecticut, Farmington, Connecticut*, Respiratory Mechanics and Gas Exchange

LIMOGE, AIME, *The René Descartes University of Paris, Paris, France*, Electroanalgesia, Systemic

LIN, PEI-JAN PAUL, *Beth Israel Deaconess Medical Center, Boston, Massachusets*, Mammography

LIN, ZHIYUE, *University of Kansas Medical Center, Kansas City, Kansas*, Electrogastrogram

LINEAWEAVER, WILLIAM C., *Unive rsity of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

LIPPING, TARMO, *Tampere University of Technology, Pori, Finland*, Monitoring in Anesthesia

LIU, XIAOHUA, *The University of Michigan, Ann Arbor, Michigan*, Polymeric Materials

LLOYD, J.J., *Regional Medical Physics Department, Newcastle-upon-Tyne, United Kingdom*, Ultraviolet Radiation in Medicine

LOEB, ROBERT, *University of Arizona, Tuscon, Arizona*, Anesthesia Machines

LOPES DE MELO, PEDRO, *State University of Rio de Janeiro, Térreo Salas, Maracanã, Thermistors*

LOUDON, ROBERT G., Lung Sounds

LOW, DANIEL A., *Washington University School of Medicine, St. Louis, Missouri*, Radiation Therapy Simulator

LU, LICHUN, *Mayo Clinic, College of Medicine, Rochester, Minnesota*, Microscopy, Electron

LU, ZHENG FENG, *Columbia University, New York, New York*, Screen-Film Systems

LYON, ANDREW W., *University of Calgary, Calgary, Canada*, Flame Atomic Emission Spectrometry and Atomic Absorption Spectrometry

LYON, MARTHA E., *University of Calgary, Calgary, Canada*, Flame Atomic Emission Spectrometry and Atomic Absorption Spectrometry

MA, C-M CHARLIE, *Fox Chase Cancer Center, Philadelphia, Pennsylvania*, X-Ray Therapy Equipment, Low and Medium Energy

MACIA, NARCISO F., *Arizona State University at the Polytechnic Campus, Mesa, Arizona*, Pneumotachometers

MACKENZIE, COLIN F., *University of Maryland, School of Medicine*, Shock, Treatment of

MACKIE, THOMAS R., *University of Wisconsin, Madison, Wisconsin*, Tomotherapy

MADNANI, ANJU, *LSU Medical Centre, Shreveport, Louisiana*, Transcutaneous Electrical Nerve Stimulation (TENS)

MADNANI, SANJAY, *LSU Medical Centre, Shreveport, Louisiana*, Transcutaneous Electrical Nerve Stimulation (TENS)

MADSEN, MARK T., *University of Iowa, Iowa City, Iowa*, Anger Camera

MAGNANO, MAURO, *ENT Division Mauriziano Hospital, Torino, Italy*, Drug Delivery Systems

MANDEL, RICHARD, *Boston University School of Medicine, Boston, Massachusetts*, Colorimetry

MANNING, KEEFE B., *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

MAO, JEREMY J., *University of Illinois at Chicago, Chicago, Illinois*, Cartilage and Meniscus, Properties of

MARCOLONGO, MICHELE, *Drexel University, Philadelphia, Pennsylvania*, Spinal Implants

MAREK, MIROSLAV, *Georgia Institute of Technology, Atlanta, Georgia*, Biomaterials, Corrosion and Wear of

MARION, NICHOLAS W., *University of Illinois at Chicago, Chicago, Illinois*, Cartilage and Meniscus, Properties of

MASTERS, KRISTYN S., *University of Wisconsin, Madison, Wisconsin*, Tissue Engineering

MAUGHAN, RICHARD L., *Hospital of the University of Pennsylvania*, Neutron Beam Therapy

MCADAMS, ERIC, *University of Ulster at Jordanstown, Newtownabbey, Ireland*, Bioelectrodes

MCARTHUR, SALLY L., *University of Sheffield, Sheffield, United Kingdom*, Biomaterials, Surface Properties of

MCEWEN, MALCOM, *National Research Council of Canada, Ontario, Canada*, Radiation Dosimetry for Oncology

MCGOWAN, EDWARD J., *E.J. McGowan & Associates*, Biofeedback

MCGRATH, SUSAN, *Dartmouth College, Hanover, New Hampshire*, Oxygen Analyzers

MEEKS, SANFORD L., *University of Florida, Gainesville, Florida*, Radiosurgery, Stereotactic

MELISSA, PETER, *University of Central Florida NanoScience Technology Center, Orlando, Florida*, Biosurface Engineering

MENDELSON, YITZHAK, *Worcester Polytechnic Institute*, Optical Sensors

METZKER, MICHAEL L., *Baylor College of Medicine, Houston, Texas*, Polymerase Chain Reaction

MEYEREND, M.E., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

MICHLER, ROBERT, *Montefiore Medical Center, Bronx, New York*, Heart–Lung Machines

MICIC, MIODRAG, *MP Biomedicals LLC, Irvine, California*, Microscopy and Spectroscopy, Near-Field

MILLER, WILLIAM, *University of Missouri Columbia, Missouri*, Radionuclide Production and Radioactive Decay

MITTRA, ERIK, *Stony Brook University, New York*, Bone Density Measurement

MODELL, MARK, *Harvard Medical School, Boston, Massachusetts*, Fiber Optics in Medicine

MORE, ROBERT B., *RBMore Associates, Austin, Texas* Biomaterials Carbon

MORE, ROBERT, *Austin, Texas*, Heart Valves, Prosthetic

MORROW, DARREN, *Royal Adelaide Hospital, Adelaide, Australia*, Intraaortic Balloon Pump

MOURTADA, FIRAS, *MD Anderson Cancer Center, Houston, Texas*, Brachytherapy, Intravascular

MOY, VINCENT T., *University of Miami, Miller School of Medicine, Miami, Florida*, Microscopy, Scanning Force

MÜFTÜ, SINAN, *Northeastern University, Boston, Massachusetts*, Tooth and Jaw, Biomechanics of

MURPHY, RAYMOND L.H., Lung Sounds

MURPHY, WILLIAM L., *University of Wisconsin, Madison, Wisconsin*, Tissue Engineering

MURRAY, ALAN, *Newcastle University Medical Physics, Newcastle upon Tyne, United Kingdom*, Pace makers

MUTIC, SASA, *Washington University School of Medicine, St. Louis, Missouri*, Radiation Therapy Simulator

NARAYAN, ROGER J., *University of North Carolina, Chapel Hill, North Carolina*, Biomaterials, Corrosion and Wear of

NATALE, ANDREA, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

NAZERAN, HOMER, *The University of Texas, El Paso, Texas*, Electrocardiography, Computers in

NEUMAN, MICHAEL R., *Michigan Technological University, Houghton, Houghton, Michigan*, Fetal Monitoring, Neonatal Monitoring

NEUZIL, PAVEL, *Institute of Bioengineering and Nanotechnology, Singapore*, Immunologically Sensitive Field-Effect Transistors

NICKOLOFF, EDWARD L., *Columbia University, New York, New York*, X-Ray Quality Control Program

NIEZGODA, JEFFREY A., *Medical College of Wisconsin, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

NISHIKAWA, ROBERT M., *The University of Chicago, Chicago, Illinois*, Computer-Assisted Detection and Diagnosis

NUTTER, BRIAN, *Texas Tech University, Lubbock, Texas*, Medical Records, Computers in

O'DONOHUE, WILLIAM, *University of Nevada, Reno, Nevada*, Sexual Instrumentation

ORTON, COLIN, *Harper Hospital and Wayne State University, Detroit, Michigan*, Medical Physics Literature

OZCELIK, SELAHATTIN, *Texas A&M University, Kingsville, Texas*, Drug Infusion Systems

PANITCH, ALYSSA, *Arizona State University, Tempe, Arizona*, Biomaterials: An Overview

PAOLINO, DONATELLA, *University of Catanzaro Magna Græcia, Germaneto (CZ), Italy*, Drug Delivery Systems

PAPAIOANNOU, GEORGE, *University of Wisconsin, Milwaukee, Wisconsin*, Joints, Biomechanics of

PARK, GRACE E., *Purdue University, West Lafayette, Indiana*, Porous Materials for Biological Applications

PARMENTER, BRETT A., *State University of New York at Buffalo, Buffalo, New York*, Sleep Studies, Computer Analysis of

PATEL, DIMPI, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

PEARCE, JOHN, *The University of Texas, Austin, Texas*, Electrosurgical Unit (ESU)

PELET, SERGE, *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

PERIASAMY, AMMASI, *University of Virginia, Charlottesville, Virginia*, Cellular Imaging

PERSONS, BARBARA L., *University of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

PIPER, IAN, *The University of Memphis, Memphis, Tennessee*, Monitoring, Intracranial Pressure

POLETTO, CHRISTOPHER J., *National Institutes of Health*, Tactile Stimulation

PREMINGER, GLENN M., *Duke University Medical Center, Durham, North Carolina*, Lithotripsy

PRENDERGAST, PATRICK J., *Trinity College, Dublin, Ireland*, Orthopedics, Prosthesis Fixation for

PREVITE, MICHAEL, *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

PURDY, JAMES A., *UC Davis Medical Center, Sacramento, California*, Radiotherapy Accessories

QI, HAIRONG, *Advanced Concepts Analysis, Inc., Falls Church, Virginia*, Thermography

QIN, YIXIAN, *Stony Brook University, New York*, Bone Density Measurement

QUAN, STUART F., *University of Arizona, Tucson, Arizona*, Ventilatory Monitoring

QUIROGA, RODRIGO QUIAN, *University of Leicester, Leicester, United Kingdom*, Evoked Potentials

RAHAGHI, FARBOD N., *University of California, La Jolla, California*, Glucose Sensors

RAHKO, PETER S., *University of Wisconsin Medical School*, Echocardiography and Doppler Echocardiography

RALPH, LIETO, *University of Wisconsin–Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

RAMANATHAN, LAKSHMI, *Mount Sinai Medical Center*, Analytical Methods, Automated

RAO, SATISH S.C., *University of Iowa College of Medicine, Iowa City, Iowa*, Anorectal Manometry

RAPOPORT, DAVID M., *NYU School of Medicine, New York, New York*, Continuous Positive Airway Pressure

REBELLO, KEITH J., *The Johns Hopkins University Applied Physics Lab, Laurel, Maryland*, Micro surgery

REDDY, NARENDER, *The University of Akron, Akron, Ohio*, Linear Variable Differential Transformers

REN-DIH, SHEU, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

RENGACHARY, SETTI S., *Detroit, Michigan*, Human Spine, Biomechanics of

REPPERGER, DANIEL W., *Wright-Patterson Air Force Base, Dayton, Ohio*, Human Factors in Medical Devices

RITCHEY, ERIC R., *The Ohio State University, Columbus, Ohio*, Contact Lenses

RIVARD, MARK J., *Tufts New England Medical Center, Boston, Massachusetts*, Imaging Devices

ROBERTSON, J. DAVID, *University of Missouri, Columbia, Missouri*, Radionuclide Production and Radioactive Decay

ROTH, BRADLEY J., *Oakland University, Rochester, Michigan*, Defibrillators

ROWE-HORWEGE, R. WANDA, *University of Texas Medical School, Houston, Texas*, Hyperthermia, Systemic

RUMSEY, JOHN W., *University of Central Florida, Orlando, Florida*, Biosurface Engineering

RUTKOWSKI, GREGORY E., *University of Minnesota, Duluth, Minnesota*, Engineered Tissue

SALATA, O.V., *University of Oxford, Oxford, United Kingdom*, Nanoparticles

SAMARAS, THEODOROS, *Aristotle University of Thessaloniki Department of Physics, Thessaloniki, Greece*, Thermometry

SANGOLE, ARCHANA P., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

SARKOZI, LASZLO, *Mount Sinai School of Medicine*, Analytical Methods, Automated

SCHEK, HENRY III, *University of Michigan, Ann Arbor, Michigan*, Optical Tweezers

SCHMITZ, CHRISTOPH H., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

SCHUCKERS, STEPHANIE A.C., *Clarkson University, Potsdam, New York*, Arrhythmia Analysis, Automated

SCOPE, KENNETH, *Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

SCOTT, ADZICK N., *University of Pennsylvania, Philadelphia, Pennsylvania*, Intrauterine Surgical Techniques

SEAL, BRANDON L., *Arizona State University, Tempe, Arizona*, Biomaterials: An Overview

SEALE, GARY, *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

SEGERS, PATRICK, *Ghent University, Belgium*, Hemodynamics

SELIM, MOSTAFA A., *Cleveland Metropolitan General Hospital, Palm Coast, Florida*, Colposcopy

SETHI, ANIL, *Loyola University Medical Center, Maywood, Illinois*, X-Rays: Interaction with Matter

SEVERINGHAUS, JOHN W., *University of California in San Francisco*, CO_2 Electrodes

SHALODI, ABDELWAHAB D., *Cleveland Metropolitan General Hospital, Palm Coast, Florida*, Colposcopy

SHANMUGASUNDARAM, SHOBANA, *New Jersey Institute of Technology, Newark, New Jersey*, Polymeric Materials

SHARD, ALEXANDER G., *University of Sheffield, Sheffield United Kingdom*, Biomaterials, Surface Properties of

SHEN, LI-JIUAN, *National Taiwan University School of Pharmacy, Taipei, Taiwan*, Colorimetry

SHEN, WEI-CHIANG, *University of Southern California School of Pharmacy, Los Angeles, California*, Colorimetry

SHERAR, MICHAEL D., *London Health Sciences Centre and University of Western Ontario, London, Ontario, Canada*, Hyperthermia, Interstitial

SHERMAN, DAVID, *The Johns Hopkins University, Baltimore, Maryland*, Electroencephalography

SHI, DONGLU, *University of Cincinnati, Cincinnati, Ohio*, Biomaterials, Testing and Structural Properties of

SHUCARD, DAVID W.M., *State University of New York at Buffalo, Buffalo, New York*, Sleep Studies, Computer Analysis of

SIEDBAND, MELVIN P., *University of Wisconsin, Madison, Wisconsin*, Image Intensifiers and Fluoroscopy

SILBERMAN, HOWARD, *University of Southern California, Los Angeles, California*, Nutrition, Parenteral

SILVERMAN, GORDON, *Manhattan College*, Computers in the Biomedical Laboratory

SILVERN, DAVID A., *Medical Physics Unit, Rabin Medical Center, Petah Tikva, Israel*, Prostate Seed Implants

SINHA, PIYUSH, *The Ohio State University, Columbus, Ohio*, Drug Delivery Systems

SINHA, ABHIJIT ROY, *University of Cincinnati, Cincinnati, Ohio*, Coronary Angioplasty and Guidewire Diagnostics

SINKJÆR, THOMAS, *Aalborg University, Aalborg, Denmark*, Electroneurography

SLOAN, JEFFREY A., *Mayo Clinic, Rochester, Minnesota*, Quality-of-Life Measures, Clinical Significance of

SO, PETER T.C., *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

SOBOL, WLAD T., *University of Alabama at Birmingham Health System, Birmingham, Alabama*, Nuclear Magnetic Resonance Spectroscopy

SOOD, SANDEEP, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

SPECTOR, MYRON, *Brigham and Women's Hospital, Boston, Massachusetts*, Biocompatibility of Materials

SPELMAN, FRANCIS A., *University of Washington*, Cochlear Prostheses

SRINIVASAN, YESHWANTH, *Texas Tech University, Lubbock, Texas*, Medical Records, Computers in

SRIRAM, NEELAMEGHAM, *University of Buffalo, Buffalo, New York*, Cell Counters, Blood

STARKO, KENTON R., *Point Roberts, Washington*, Physiological Systems Modeling

STARKSCHALL, GEORGE, *The University of Texas*, Radiotherapy, Three-Dimensional Conformal

STAVREV, PAVEL, *Cross Cancer Institute, Edmonton, Alberta, Canada*, Radiotherapy Treatment Planning, Optimization of

STENKEN, JULIE A., *Rensselaer Polytechnic Institute, Troy, New York*, Microdialysis Sampling

STIEFEL, ROBERT, *University of Maryland Medical Center, Baltimore, Maryland*, Equipment Acquisition

STOKES, I.A.F., *Polytechniquie Montreal, Montreal Quebec, Canada*, Scoliosis, Biomechanics of

STONE, M.H., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

SU, XIAO-LI, *BioDetection Instruments LLC, Fayetteville, Arkansas*, Piezoelectric Sensors

SUBHAN, ARIF, *Masterplan Technology Management, Chatsworth, California*, Equipment Maintenance, Biomedical

SWEENEY, JAMES D., *Arizona State University, Tempe, Arizona*, Functional Electrical Stimulation

SZETO, ANDREW Y.J., *San Diego State University, San Diego, California*, Blind and Visually Impaired, Assistive Technology for

TAKAYAMA, SHUICHI, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

TAMUL, PAUL C., *Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

TAMURA, TOSHIYO, *Chiba University School of Engineering, Chiba, Japan*, Home Health Care Devices

TANG, XIANGYANG, *GE Healthcare Technologies, Wankesha, Wisconsin*, Computed Tomography Simulators

TAYLOR, B.C., *The University of Akron, Akron, Ohio*, Cardiac Output, Indicator Dilution Measurement of

TEMPLE, RICHARD O., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

TEN, STANLEY, *Salt Lake City, Utah*, Electroanalgesia, Systemic

TERRY, TERESA M., *Walter Reed Army Institute of Research, Rockville, Maryland*, Blood Collection and Processing

THAKOR, N.V., *Johns Hopkins University, Baltimore, Maryland*, Neurological Monitors

THIERENS, HUBERT M.A., *University of Ghent, Ghent, Belgium*, Radiopharmaceutical Dosimetry

THOMADSEN, BRUCE, *University of Wisconsin–Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

TIPPER, J.L., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

TOGAWA, TATSUO, *Waseda University, Saitama, Japan*, Integrated Circuit Temperature Sensor

TORNAI, MARTIN, *Duke University, Durham, North Carolina*, X-Ray Equipment Design

TRAN-SON-TAY, ROGER, *University of Florida, Gainesville, Florida*, Blood Rheology

YENI, YENER N., *Henry Ford Hospital, Detroit, Michigan*, Joints, Biomechanics of

YLI-HANKALA, ARVI, *Tampere University of Technology, Pori, Finland*, Monitoring in Anesthesia

YOKO, KAMOTANI, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

YOON, KANG JI, *Korea Institute of Science and Technology, Seoul, Korea*, Micropower for Medical Applications

YORKE, ELLEN, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

YOSHIDA, KEN, *Aalborg University, Aalborg, Denmark*, Electroneurography

YOUNGSTEDT, SHAWN D., *University of South Carolina, Columbia, South Carolina*, Sleep Laboratory

YU, YIH-CHOUNG, *Lafayette College, Easton, Pennsylvania*, Blood Pressure, Automatic Control of

ZACHARIAH, EMMANUEL S., *University of Medicine and Dentistry of New Jersey, New Brunswick, New Jersey*, Immunologically Sensitive Field-Effect Transistors

ZAIDER, MARCO, *Memorial Sloan Kettering Cancer Center, New York, New York*, Prostate Seed Implants

ZAPANTA, CONRAD M., *Penn State College of Medicine, Hershey, Pennsylvania*, Heart, Artificial

ZARDENETA, GUSTAVO, *University of Texas, San Antonio, Texas*, Fluorescence Measurements

ZELMANOVIC, DAVID, *Bayer HealthCare LLC, Tarrytown, New York*, Differential Counts, Automated

ZHANG, MIN, *University of Washington, Seattle, Washington*, Biomaterials: Polymers

ZHANG, YI, *University of Buffalo, Buffalo, New York*, Cell Counters, Blood

ZHU, XIAOYUE, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

ZIAIE, BABAK, *Purdue University, W. Lafayette, Indiana*, Biotelemetry

ZIELINSKI, TODD M., *Medtronic, Inc., Minneapolis, Minnesota*, Bioimpedance in Cardiovascular Medicine

ZIESSMAN, HARVEY A., *Johns Hopkins University*, Computed Tomography, Single Photon Emission

PREFACE

This six-volume work is an alphabetically organized compilation of almost 300 articles that describe critical aspects of medical devices and instrumentation.

It is comprehensive. The articles emphasize the contributions of engineering, physics, and computers to each of the general areas of anesthesiology, biomaterials, burns, cardiology, clinical chemistry, clinical engineering, communicative disorders, computers in medicine, critical care medicine, dermatology, dentistry, ear, nose, and throat, emergency medicine, endocrinology, gastroenterology, genetics, geriatrics, gynecology, hematology, heptology, internal medicine, medical physics, microbiology, nephrology, neurology, nutrition, obstetrics, oncology, ophthalmology, orthopedics, pain, pediatrics, peripheral vascular disease, pharmacology, physical therapy, psychiatry, pulmonary medicine, radiology, rehabilitation, surgery, tissue engineering, transducers, and urology.

The discipline is defined through the synthesis of the core knowledge from all the fields encompassed by the application of engineering, physics, and computers to problems in medicine. The articles focus not only on what is now useful but also on what is likely to be useful in future medical applications.

These volumes answer the question, "What are the branches of medicine and how does technology assist each of them?" rather than "What are the branches of technology and how could each be used in medicine?" To keep this work to a manageable length, the practice of medicine that is unassisted by devices, such as the use of drugs to treat disease, has been excluded.

The articles are accessible to the user; each benefits from brevity of condensation instead of what could easily have been a book-length work. The articles are designed not for peers, but rather for workers from related fields who wish to take a first look at what is important in the subject.

The articles are readable. They do not presume a detailed background in the subject, but are designed for any person with a scientific background and an interest in technology. Rather than attempting to teach the basics of physiology or Ohm's law, the articles build on such basic concepts to show how the worlds of life science and physical science meld to produce improved systems. While the ideal reader might be a person with a Master's degree in biomedical engineering or medical physics or an M.D. with a physical science undergraduate degree, much of the material will be of value to others with an interest in this growing field. High school students and hospital patients can skip over more technical areas and still gain much from the descriptive presentations.

The *Encyclopedia of Medical Devices and Instrumentation* is excellent for browsing and searching for those new divergent associations that may advance work in a peripheral field. While it can be used as a reference for facts, the articles are long enough that they can serve as an educational instrument and provide genuine understanding of a subject.

One can use this work just as one would use a dictionary, since the articles are arranged alphabetically by topic. Cross references assist the reader looking for subjects listed under slightly different names. The index at the end leads the reader to all articles containing pertinent information on any subject. Listed on pages xxi to xxx are all the abbreviations and acronyms used in the *Encyclopedia.* Because of the increasing use of SI units in all branches of science, these units are provided throughout the *Encyclopedia* articles as well as on pages xxxi to xxxv in the section on conversion factors and unit symbols.

I owe a great debt to the many people who have contributed to the creation of this work. At John Wiley & Sons, Encyclopedia Editor George Telecki provided the idea and guiding influence to launch the project. Sean Pidgeon was Editorial Director of the project. Assistant Editors Roseann Zappia, Sarah Harrington, and Surlan Murrell handled the myriad details of communication between publisher, editor, authors, and reviewers and stimulated authors and reviewers to meet necessary deadlines.

My own background has been in the electrical aspects of biomedical engineering. I was delighted to have the assistance of the editorial board to develop a comprehensive encyclopedia. David J. Beebe suggested cellular topics such as microfluidics. Jerry M. Calkins assisted in defining the chemically related subjects, such as anesthesiology. Michael R. Neuman suggested subjects related to sensors, such as in his own work—neonatology. Joon B. Park has written extensively on biomaterials and suggested related subjects. Edward S. Sternick provided many suggestions from medical physics. The Editorial Board was instrumental both in defining the list of subjects and in suggesting authors.

This second edition brings the field up to date. It is available on the web at http://www.mrw.interscience.wiley.com/emdi, where articles can be searched simultaneously to provide rapid and comprehensive information on all aspects of medical devices and instrumentation.

JOHN G. WEBSTER
University of Wisconsin, Madison

LIST OF ARTICLES

ABBREVIATIONS AND ACRONYMS

AAMI	Association for the Advancement of Medical Instrumentation
AAPM	American Association of Physicists in Medicine
ABC	Automatic brightness control
ABET	Accreditation board for engineering training
ABG	Arterial blood gases
ABLB	Alternative binaural loudness balance
ABS	Acrylonitrile–butadiene–styrene
ac	Alternating current
AC	Abdominal circumference; Affinity chromatography
ACA	Automated clinical analyzer
ACES	Augmentative communication evaluation system
ACL	Anterior chamber lens
ACLS	Advanced cardiac life support
ACOG	American College of Obstetrics and Gynecology
ACR	American College of Radiology
ACS	American Cancer Society; American College of Surgeons
A/D	Analog-to-digital
ADC	Agar diffusion chambers; Analog-to-digital converter
ADCC	Antibody-dependent cellular cytotoxicity
ADCL	Accredited Dosimetry Calibration Laboratories
ADP	Adenosine diphosphate
A-D-T	Admission, discharge, and transfer
AE	Anion exchange; Auxiliary electrode
AEA	Articulation error analysis
AEB	Activation energy barrier
AEC	Automatic exposure control
AED	Automatic external defibrillator
AEMB	Alliance for Engineering in Medicine and Biology
AES	Auger electron spectroscopy
AESC	American Engineering Standards Committee
AET	Automatic exposure termination
AFO	Ankle-foot orthosis
AGC	Automatic gain control
AHA	American Heart Association
AI	Arterial insufficiency
AICD	Automatic implantable cardiac defibrillator
AID	Agency for International Development
AIDS	Acquired immune deficiency syndrome
AL	Anterior leaflet
ALG	Antilymphocyte globulin
ALS	Advanced life support; Amyotropic lateral sclerosis
ALT	Alanine aminotransferase
ALU	Arithmetic and logic unit
AM	Amplitude modulation
AMA	American Medical Association
amu	Atomic mass units
ANOVA	Analysis of variance
ANSI	American National Standards Institute
AP	Action potential; Alternative pathway; Anteroposterior
APD	Anterioposterior diameter
APL	Adjustable pressure limiting valve; Applied Physics Laboratory
APR	Anatomically programmed radiography
AR	Amplitude reduction; Aortic regurgitation; Autoregressive
Ara-C	Arabinosylcytosine
ARD	Absorption rate density
ARDS	Adult respiratory distress syndrome
ARGUS	Arrhythmia guard system
ARMA	Autoregressive-moving-average model
ARMAX	Autoregressive-moving-average model with external inputs
AS	Aortic stenosis
ASA	American Standards Association
ASCII	American standard code for information interchange
ASD	Antisiphon device
ASHE	American Society for Hospital Engineering
ASTM	American Society for Testing and Materials
AT	Adenosine-thiamide; Anaerobic threshold; Antithrombin
ATA	Atmosphere absolute
ATLS	Advanced trauma life support
ATN	Acute tubular necrosis
ATP	Adenosine triphosphate
ATPD	Ambient temperature pressure dry
ATPS	Ambient temperature pressure saturated
ATR	Attenuated total reflection
AUC	Area under curve
AUMC	Area under moment curve
AV	Atrioventricular
AZT	Azido thymidine
BA	Biliary atresia
BAEP	Brainstem auditory evoked potential
BAPN	Beta-amino-proprionitryl
BAS	Boston anesthesis system
BASO	Basophil
BB	Buffer base
BBT	Basal body temperature

BCC	Body-centered cubic
BCD	Binary-coded decimal
BCG	Ballistocardiogram
BCLS	Basic cardiac life support
BCRU	British Commitee on Radiation Units and Measurements
BDI	Beck depression inventory
BE	Base excess; Binding energy
BET	Brunauer, Emmett, and Teller methods
BH	His bundle
BI	Biological indicators
BIH	Beth Israel Hospital
BIPM	International Bureau of Weights and Measurements
BJT	Bipolar junction transistor
BMDP	Biomedical Programs
BME	Biomedical engineering
BMET	Biomedical equipment technician
BMO	Biomechanically optimized
BMR	Basal metabolic rate
BOL	Beginning of life
BP	Bereitschafts potential; Break point
BR	Polybutadiene
BRM	Biological response modifier
BRS	Bibliographic retrieval services
BSS	Balanced salt solution
BTG	Beta thromboglobulin
BTPS	Body temperature pressure saturated
BUN	Blood urea nitrogen
BW	Body weight
CA	Conductive adhesives
CABG	Coronary artery by-pass grafting
CAD/CAM	Computer-aided design/computer-aided manufacturing
CAD/D	Computer-aided drafting and design
CADD	Central axis depth dose
CAI	Computer assisted instruction; Computer-aided instruction
CAM	Computer-assisted management
cAMP	Cyclic AMP
CAPD	Continuous ambulatory peritoneal dialysis
CAPP	Child amputee prosthetic project
CAT	Computerized axial tomography
CATS	Computer-assisted teaching system; Computerized aphasia treatment system
CAVH	Continuous arteriovenous hemofiltration
CB	Conjugated bilirubin; Coulomb barrier
CBC	Complete blood count
CBF	Cerebral blood flow
CBM	Computer-based management
CBV	Cerebral blood volume
CC	Closing capacity
CCC	Computer Curriculum Company
CCD	Charge-coupled device
CCE	Capacitance contact electrode
CCF	Cross-correlation function
CCL	Cardiac catheterization laboratory
CCM	Critical care medical services
CCPD	Continuous cycling peritoneal dialysis
CCTV	Closed circuit television system
CCU	Coronary care unit; Critical care unit
CD	Current density
CDR	Complimentary determining region
CDRH	Center for Devices and Radiological Health
CEA	Carcinoembryonic antigen
CF	Conversion factor; Cystic fibrosis
CFC	Continuous flow cytometer
CFR	Code of Federal Regulations
CFU	Colony forming units
CGA	Compressed Gas Association
CGPM	General Conference on Weights and Measures
CHO	Carbohydrate
CHO	Chinese hamster ovary
CI	Combination index
CICU	Cardiac intensive care unit
CIF	Contrast improvement factor
CIN	Cervical intraepithelial neoplasia
CK	Creatine kinase
CLAV	Clavicle
CLSA	Computerized language sample analysis
CM	Cardiomyopathy; Code modulation
CMAD	Computer managed articulation diagnosis
CMI	Computer-managed instruction
CMRR	Common mode rejection ratio
CMV	Conventional mechanical ventilation; Cytomegalovirus
CNS	Central nervous system
CNV	Contingent negative variation
CO	Carbon monoxide; Cardiac output
COBAS	Comprehensive Bio-Analysis System
COPD	Chronic obstructive pulmonary disease
COR	Center of rotation
CP	Cerebral palsy; Closing pressure; Creatine phosphate
CPB	Cardiopulmonary bypass
CPET	Cardiac pacemaker electrode tips
CPM	Computerized probe measurements
CPP	Cerebral perfusion pressure; Cryoprecipitated plasma
CPR	Cardiopulmonary resuscitation
cps	Cycles per second
CPU	Central Processing unit
CR	Center of resistance; Conditioned response; Conductive rubber; Creatinine
CRBB	Complete right bundle branch block
CRD	Completely randomized design
CRL	Crown rump length
CRT	Cathode ray tube
CS	Conditioned stimulus; Contrast scale; Crown seat
CSA	Compressed spectral array
CSF	Cerebrospinal fluid
CSI	Chemical shift imaging
CSM	Chemically sensitive membrane
CT	Computed tomography; Computerized tomography
CTI	Cumulative toxicity response index
CV	Closing volume

C.V.	Coefficient of variation
CVA	Cerebral vascular accident
CVP	Central venous pressure
CVR	Cardiovascular resistance
CW	Continuous wave
CWE	Coated wire electrodes
CWRU	Case Western Reserve University
DAC	Digital-to-analog converter
DAS	Data acquisition system
dB	Decibel
DB	Direct body
DBMS	Data base management system
DBS	Deep brain stimulation
dc	Direct current
DCCT	Diabetes control and complications trial
DCP	Distal cavity pressure
DCS	Dorsal column stimulation
DDC	Deck decompression chamber
DDS	Deep diving system
DE	Dispersive electrode
DEN	Device experience network
DERS	Drug exception ordering system
DES	Diffuse esophageal spasm
d.f.	Distribution function
DHCP	Distributed Hospital Computer Program
DHE	Dihematoporphyrin ether
DHEW	Department of Health Education and Welfare
DHHS	Department of Health and Human Services
DHT	Duration of hypothermia
DI	Deionized water
DIC	Displacement current
DIS	Diagnostic interview schedule
DL	Double layer
DLI	Difference lumen for intensity
DM	Delta modulation
DME	Dropping mercury electrode
DN	Donation number
DNA	Deoxyribonucleic acid
DOF	Degree of freedom
DOS	Drug ordering system
DOT-NHTSA	Department of Transportation Highway Traffic Safety Administration
DPB	Differential pencil beam
DPG	Diphosphoglycerate
DQE	Detection quantum efficiency
DRESS	Depth-resolved surface coil spectroscopy
DRG	Diagnosis-related group
DSA	Digital subtraction angiography
DSAR	Differential scatter-air ratio
DSB	Double strand breaks
DSC	Differential scanning calorimetry
D-T	Deuterium-on-tritium
DTA	Differential thermal analysis
d.u.	Density unit
DUR	Duration
DVT	Deep venous thrombosis
EA	Esophageal accelerogram
EB	Electron beam
EBCDIC	Extended binary code decimal interchange code
EBS	Early burn scar
EBV	Epstein–Barr Virus
EC	Ethyl cellulose
ECC	Emergency cardiac care; Extracorporeal circulation
ECCE	Extracapsular cataract extinction
ECD	Electron capture detector
ECG	Electrocardiogram
ECM	Electrochemical machining
ECMO	Extracorporeal membrane oxygenation
ECOD	Extracranial cerebrovascular occlusive disease
ECRI	Emergency Care Research Institute
ECS	Exner's Comprehensive System
ECT	Electroconvulsive shock therapy; Electroconvulsive therapy; Emission computed tomography
EDD	Estimated date of delivery
EDP	Aortic end diastolic pressure
EDTA	Ethylenediaminetetraacetic acid
EDX	Energy dispersive X-ray analysis
EEG	Electroencephalogram
EEI	Electrode electrolyte interface
EELV	End-expiratory lung volume
EER	Electrically evoked response
EF	Ejection fraction
EF	Electric field; Evoked magnetic fields
EFA	Estimated fetal age
EGF	Epidermal growth factor
EGG	Electrogastrogram
EIA	Enzyme immunoassay
EIU	Electrode impedance unbalance
ELF	Extra low frequency
ELGON	Electrical goniometer
ELISA	Enzyme-linked immunosorbent assay
ELS	Energy loss spectroscopy
ELV	Equivalent lung volume
EM	Electromagnetic
EMBS	Engineering in Medicine and Biology Society
emf	Electromotive force
EMG	Electromyogram
EMGE	Integrated electromyogram
EMI	Electromagnetic interference
EMS	Emergency medical services
EMT	Emergency medical technician
ENT	Ear, nose, and throat
EO	Elbow orthosis
EOG	Electrooculography
EOL	End of life
EOS	Eosinophil
EP	Elastoplastic; Evoked potentiate
EPA	Environmental protection agency
ER	Evoked response
ERCP	Endoscopic retrograde cholangiopancreatography
ERG	Electron radiography; Electroretinogram
ERMF	Event-related magnetic field
ERP	Event-related potential
ERV	Expiratory reserve volume

ESCA	Electron spectroscopy for chemical analysis
ESI	Electrode skin impedance
ESRD	End-stage renal disease
esu	Electrostatic unit
ESU	Electrosurgical unit
ESWL	Extracorporeal shock wave lithotripsy
ETO, Eto	Ethylene oxide
ETT	Exercise tolerance testing
EVA	Ethylene vinyl acetate
EVR	Endocardial viability ratio
EW	Extended wear
FAD	Flavin adenine dinucleotide
FARA	Flexible automation random analysis
FBD	Fetal biparietal diameter
FBS	Fetal bovine serum
fcc	Face centered cubic
FCC	Federal Communications Commission
Fct	Fluorocrit
FDA	Food and Drug Administration
FDCA	Food, Drug, and Cosmetic Act
FE	Finite element
FECG	Fetal electrocardiogram
FEF	Forced expiratory flow
FEL	Free electron lasers
FEM	Finite element method
FEP	Fluorinated ethylene propylene
FES	Functional electrical stimulation
FET	Field-effect transistor
FEV	Forced expiratory volume
FFD	Focal spot to film distance
FFT	Fast Fourier transform
FGF	Fresh gas flow
FHR	Fetal heart rate
FIC	Forced inspiratory capacity
FID	Flame ionization detector; Free-induction decay
FIFO	First-in-first-out
FITC	Fluorescent indicator tagged polymer
FL	Femur length
FM	Frequency modulation
FNS	Functional neuromuscular stimulation
FO	Foramen ovale
FO-CRT	Fiber optics cathode ray tube
FP	Fluorescence polarization
FPA	Fibrinopeptide A
FR	Federal Register
FRC	Federal Radiation Council; Functional residual capacity
FSD	Focus-to-surface distance
FTD	Focal spot to tissue-plane distance
FTIR	Fourier transform infrared
FTMS	Fourier transform mass spectrometer
FU	Fluorouracil
FUDR	Floxuridine
FVC	Forced vital capacity
FWHM	Full width at half maximum
FWTM	Full width at tenth maximum
GABA	Gamma amino buteric acid
GAG	Glycosaminoglycan
GBE	Gas-bearing electrodynamometer
GC	Gas chromatography; Guanine-cytosine
GDT	Gas discharge tube
GFR	Glomerular filtration rate
GHb	Glycosylated hemoglobin
GI	Gastrointestinal
GLC	Gas–liquid chromatography
GMV	General minimum variance
GNP	Gross national product
GPC	Giant papillary conjunctivitis
GPH	Gas-permeable hard
GPH-EW	Gas-permeable hard lens extended wear
GPO	Government Printing Office
GSC	Gas-solid chromatography
GSR	Galvanic skin response
GSWD	Generalized spike-wave discharge
HA	Hydroxyapatite
HAM	Helical axis of motion
Hb	Hemoglobin
HBE	His bundle electrogram
HBO	Hyperbaric oxygenation
HC	Head circumference
HCA	Hypothermic circulatory arrest
HCFA	Health care financing administration
HCL	Harvard Cyclotron Laboratory
hcp	Hexagonal close-packed
HCP	Half cell potential
HDPE	High density polyethylene
HECS	Hospital Equipment Control System
HEMS	Hospital Engineering Management System
HEPA	High efficiency particulate air filter
HES	Hydroxyethylstarch
HETP	Height equivalent to a theoretical plate
HF	High-frequency; Heating factor
HFCWO	High-frequency chest wall oscillation
HFER	High-frequency electromagnetic radiation
HFJV	High-frequency jet ventilation
HFO	High-frequency oscillator
HFOV	High-frequency oscillatory ventilation
HFPPV	High-frequency positive pressure ventilation
HFV	High-frequency ventilation
HHS	Department of Health and Human Services
HIBC	Health industry bar code
HIMA	Health Industry Manufacturers Association
HIP	Hydrostatic indifference point
HIS	Hospital information system
HK	Hexokinase
HL	Hearing level
HMBA	Hexamethylene bisacetamide
HMO	Health maintenance organization
HMWPE	High-molecular-weight polyethylene
HOL	Higher-order languages
HP	Heating factor; His-Purkinje
HpD	Hematoporphyrin derivative
HPLC	High-performance liquid chromatography
HPNS	High-pressure neurological syndrome
HPS	His-Purkinje system
HPX	High peroxidase activity

HR	Heart rate; High-resolution
HRNB	Halstead-Reitan Neuropsychological Battery
H/S	Hard/soft
HSA	Human serum albumin
HSG	Hysterosalpingogram
HTCA	Human tumor cloning assay
HTLV	Human T cell lymphotrophic virus
HU	Heat unit; Houndsfield units; Hydroxyurea
HVL	Half value layer
HVR	Hypoxic ventilatory response
HVT	Half-value thickness
IA	Image intensifier assembly; Inominate artery
IABP	Intraaortic balloon pumping
IAEA	International Atomic Energy Agency
IAIMS	Integrated Academic Information Management System
IASP	International Association for the Study of Pain
IC	Inspiratory capacity; Integrated circuit
ICCE	Intracapsular cataract extraction
ICD	Intracervical device
ICDA	International classification of diagnoses
ICL	Ms-clip lens
ICP	Inductively coupled plasma; Intracranial pressure
ICPA	Intracranial pressure amplitude
ICRP	International Commission on Radiological Protection
ICRU	International Commission on Radiological Units and Measurements
ICU	Intensive care unit
ID	Inside diameter
IDDM	Insulin dependent diabetes mellitus
IDE	Investigational device exemption
IDI	Index of inspired gas distribution
I:E	Inspiratory: expiratory
IEC	International Electrotechnical Commission; Ion-exchange chromatography
IEEE	Institute of Electrical and Electronics Engineers
IEP	Individual educational program
BETS	Inelastic electron tunneling spectroscopy
IF	Immunofluorescent
IFIP	International Federation for Information Processing
IFMBE	International Federation for Medical and Biological Engineering
IGFET	Insulated-gate field-effect transistor
IgG	Immunoglobulin G
IgM	Immunoglobulin M
IHP	Inner Helmholtz plane
IHSS	Idiopathic hypertrophic subaortic stenosis
II	Image intensifier
IIIES	Image intensifier input-exposure sensitivity
IM	Intramuscular
IMFET	Immunologically sensitive field-effect transistor
IMIA	International Medical Informatics Association
IMS	Information management system
IMV	Intermittent mandatory ventilation
INF	Interferon
IOL	Intraocular lens
IPC	Ion-pair chromatography
IPD	Intermittent peritoneal dialysis
IPG	Impedance plethysmography
IPI	Interpulse interval
IPPB	Intermittent positive pressure breathing
IPTS	International practical temperature scale
IR	Polyisoprene rubber
IRB	Institutional Review Board
IRBBB	Incomplete right bundle branch block
IRPA	International Radiation Protection Association
IRRAS	Infrared reflection-absorption spectroscopy
IRRS	Infrared reflection spectroscopy
IRS	Internal reflection spectroscopy
IRV	Inspiratory reserve capacity
IS	Image size; Ion-selective
ISC	Infant skin servo control
ISDA	Instantaneous screw displacement axis
ISE	Ion-selective electrode
ISFET	Ion-sensitive field effect transistor
ISIT	Intensified silicon-intensified target tube
ISO	International Organization for Standardization
ISS	Ion scattering spectroscopy
IT	Intrathecal
ITEP	Institute of Theoretical and Experimental Physics
ITEPI	Instantaneous trailing edge pulse impedance
ITLC	Instant thin-layer chromatography
IUD	Intrauterine device
IV	Intravenous
IVC	Inferior vena cava
IVP	Intraventricular pressure
JCAH	Joint Commission on the Accreditation of Hospitals
JND	Just noticeable difference
JRP	Joint replacement prosthesis
KB	Kent bundle
Kerma	Kinetic energy released in unit mass
KO	Knee orthosis
KPM	Kilopond meter
KRPB	Krebs-Ringer physiological buffer
LA	Left arm; Left atrium
LAD	Left anterior descending; Left axis deviation
LAE	Left atrial enlargement
LAK	Lymphokine activated killer
LAL	Limulus amoebocyte lysate
LAN	Local area network
LAP	Left atrial pressure
LAT	Left anterior temporalis
LBBB	Left bundle branch block
LC	Left carotid; Liquid chromatography

LCC	Left coronary cusp
LCD	Liquid crystal display
LDA	Laser Doppler anemometry
LDF	Laser Doppler flowmetry
LDH	Lactate dehydrogenase
LDPE	Low density polyethylene
LEBS	Low-energy brief stimulus
LED	Light-emitting diode
LEED	Low energy electron diffraction
LES	Lower esophageal sphincter
LESP	Lower esophageal sphincter pressure
LET	Linear energy transfer
LF	Low frequency
LH	Luteinizing hormone
LHT	Local hyperthermia
LL	Left leg
LLDPE	Linear low density polyethylene
LLPC	Liquid-liquid partition chromatography
LLW	Low-level waste
LM	Left masseter
LNNB	Luria-Nebraska Neuropsychological Battery
LOS	Length of stay
LP	Late potential; Lumboperitoneal
LPA	Left pulmonary artery
LPC	Linear predictive coding
LPT	Left posterior temporalis
LPV	Left pulmonary veins
LRP	Late receptor potential
LS	Left subclavian
LSC	Liquid-solid adsorption chromatography
LSI	Large scale integrated
LSV	Low-amplitude shear-wave viscoelastometry
LTI	Low temperature isotropic
LUC	Large unstained cells
LV	Left ventricle
LVAD	Left ventricular assist device
LVDT	Linear variable differential transformer
LVEP	Left ventricular ejection period
LVET	Left ventricular ejection time
LVH	Left ventricular hypertrophy
LYMPH	Lymphocyte
MAA	Macroaggregated albumin
MAC	Minimal auditory capabilities
MAN	Manubrium
MAP	Mean airway pressure; Mean arterial pressure
MAST	Military assistance to safety and traffic
MBA	Monoclonal antibody
MBV	Maximum breathing ventilation
MBX	Monitoring branch exchange
MCA	Methyl cryanoacrylate
MCG	Magnetocardiogram
MCI	Motion Control Incorporated
MCMI	Millon Clinical Multiaxial Inventory
MCT	Microcatheter transducer
MCV	Mean corpuscular volume
MDC	Medical diagnostic categories
MDI	Diphenylmethane diisocyanate; Medical Database Informatics
MDP	Mean diastolic aortic pressure
MDR	Medical device reporting
MDS	Multidimensional scaling
ME	Myoelectric
MED	Minimum erythema dose
MEDPAR	Medicare provider analysis and review
MEFV	Maximal expiratory flow volume
MEG	Magnetoencephalography
MeSH	Medline subject heading
METS	Metabolic equivalents
MF	Melamine-formaldehyde
MFP	Magnetic field potential
MGH	Massachusetts General Hospital
MHV	Magnetic heart vector
MI	Myocardial infarction
MIC	Minimum inhibitory concentration
MIFR	Maximum inspiratory flow rate
MINET	Medical Information Network
MIR	Mercury-in-rubber
MIS	Medical information system; Metal-insulator-semiconductor
MIT	Massachusetts Institute of Technology
MIT/BIH	Massachusetts Institute of Technology/ Beth Israel Hospital
MMA	Manual metal arc welding
MMA	Methyl methacrylate
MMECT	Multiple-monitored ECT
MMFR	Maximum midexpiratory flow rate
mm Hg	Millimeters of mercury
MMPI	Minnesota Multiphasic Personality Inventory
MMSE	Minimum mean square error
MO	Membrane oxygenation
MONO	Monocyte
MOSFET	Metal oxide silicon field-effect transistor
MP	Mercaptopurine; Metacarpal-phalangeal
MPD	Maximal permissible dose
MR	Magnetic resonance
MRG	Magnetoretinogram
MRI	Magnetic resonance imaging
MRS	Magnetic resonance spectroscopy
MRT	Mean residence time
MS	Mild steel; Multiple sclerosis
MSR	Magnetically shielded room
MTBF	Mean time between failure
MTF	Modulation transfer function
MTTR	Mean time to repair
MTX	Methotroxate
MUA	Motor unit activity
MUAP	Motor unit action potential
MUAPT	Motor unit action potential train
MUMPI	Missouri University Multi-Plane Imager
MUMPS	Massachusetts General Hospital utility multiuser programming system
MV	Mitral valve
MVO_2	Maximal oxygen uptake
MVTR	Moisture vapor transmission rate
MVV	Maximum voluntary ventilation
MW	Molecular weight

NAA	Neutron activation analysis
NAD	Nicotinamide adenine dinucleotide
NADH	Nicotinamide adenine dinucleotide, reduced form
NADP	Nicotinamide adenine dinucleotide phosphate
NAF	Neutrophil activating factor
NARM	Naturally occurring and accelerator-produced radioactive materials
NBB	Normal buffer base
NBD	Neuromuscular blocking drugs
N-BPC	Normal bonded phase chromatography
NBS	National Bureau of Standards
NCC	Noncoronary cusp
NCCLS	National Committee for Clinical Laboratory Standards; National Committee on Clinical Laboratory Standards
NCRP	National Council on Radiation Protection
NCT	Neutron capture theory
NEEP	Negative end-expiratory pressure
NEMA	National Electrical Manufacturers Association
NEMR	Nonionizing electromagnetic radiation
NEQ	Noise equivalent quanta
NET	Norethisterone
NEUT	Neutrophil
NFPA	National Fire Protection Association
NH	Neonatal hepatitis
NHE	Normal hydrogen electrode
NHLBI	National Heart, Lung, and Blood Institute
NIR	Nonionizing radiation
NIRS	National Institute for Radiologic Science
NK	Natural killer
NMJ	Neuromuscular junction
NMOS	N-type metal oxide silicon
NMR	Nuclear magnetic resonance
NMS	Neuromuscular stimulation
NPH	Normal pressure hydrocephalus
NPL	National Physical Laboratory
NR	Natural rubber
NRC	Nuclear Regulatory Commission
NRZ	Non-return-to-zero
NTC	Negative temperature coefficient
NTIS	National Technical Information Service
NVT	Neutrons versus time
NYHA	New York Heart Association
ob/gyn	Obstetrics and gynecology
OCR	Off-center ratio; Optical character recognition
OCV	Open circuit voltage
OD	Optical density; Outside diameter
ODC	Oxyhemoglobin dissociation curve
ODT	Oxygen delivery truck
ODU	Optical density unit
OER	Oxygen enhancement ratio
OFD	Object to film distance; Occiputo-frontal diameter
OHL	Outer Helmholtz layer
OHP	Outer Helmholtz plane
OIH	Orthoiodohippurate
OPG	Ocular pneumoplethysmography
OR	Operating room
OS	Object of known size; Operating system
OTC	Over the counter
OV	Offset voltage
PA	Posterioanterior; Pulmonary artery; Pulse amplitude
PACS	Picture archiving and communications systems
PAD	Primary afferent depolarization
PAM	Pulse amplitude modulation
PAN	Polyacrylonitrile
PAP	Pulmonary artery pressure
PAR	Photoactivation ratio
PARFR	Program for Applied Research on Fertility Regulation
PARR	Poetanesthesia recovery room
PAS	Photoacoustic spectroscopy
PASG	Pneumatic antishock garment
PBI	Penile brachial index
PBL	Positive beam limitation
PBT	Polybutylene terephthalate
PC	Paper chromatography; Personal computer; Polycarbonate
PCA	Patient controlled analgesia; Principal components factor analysis
PCG	Phonocardiogram
PCI	Physiological cost index
PCL	Polycaprolactone; Posterior chamber lens
PCR	Percent regurgitation
PCRC	Perinatal Clinical Research Center
PCS	Patient care system
PCT	Porphyria cutanea tarda
PCWP	Pulmonary capillary wedge pressure
PD	Peritoneal dialysis; Poly-p-dioxanone; Potential difference; Proportional and derivative
PDD	Percent depth dose; Perinatal Data Directory
PDE	Pregelled disposable electrodes
p.d.f.	Probability density function
PDL	Periodontal ligament
PDM	Pulse duration modulation
PDMSX	Polydimethyl siloxane
PDS	Polydioxanone
PE	Polyethylene
PEEP	Positive end-expiratory pressure
PEFR	Peak expiratory now rate
PEN	Parenteral and enteral nutrition
PEP	Preejection period
PEPPER	Programs examine phonetic find phonological evaluation records
PET	Polyethylene terephthalate; Positron-emission tomography
PEU	Polyetherurethane
PF	Platelet factor
PFA	Phosphonoformic add
PFC	Petrofluorochemical
PFT	Pulmonary function testing
PG	Polyglycolide; Propylene glycol

PGA	Polyglycolic add
PHA	Phytohemagglutinin; Pulse-height analyzer
PHEMA	Poly-2-hydroxyethyl methacrylate
PI	Propidium iodide
PID	Pelvic inflammatory disease; Proportional/integral/derivative
PIP	Peak inspiratory pressure
PL	Posterior leaflet
PLA	Polylactic acid
PLATO	Program Logic for Automated Teaching Operations
PLD	Potentially lethal damage
PLED	Periodic latoralized epileptiform discharge
PLT	Platelet
PM	Papillary muscles; Preventive maintenance
PMA	Polymethyl acrylate
p.m.f.	Probability mass function
PMMA	Polymethyl methacrylate
PMOS	P-type metal oxide silicon
PMP	Patient management problem; Poly(4-methylpentane)
PMT	Photomultiplier tube
PO	Per os
Po_2	Partial pressure of oxygen
POBT	Polyoxybutylene terephthalate
POM	Polyoxymethylene
POMC	Patient order management and communication system
POPRAS	Problem Oriented Perinatal Risk Assessment System
PP	Perfusion pressure; Polyproplyene; Postprandial (after meals)
PPA	Phonemic process analysis
PPF	Plasma protein fraction
PPM	Pulse position modulation
PPSFH	Polymerized phyridoxalated stroma-free hemoglobin
PR	Pattern recognition; Pulse rate
PRBS	Pseudo-random binary signals
PRP	Pulse repetition frequency
PRO	Professional review organization
PROM	Programmable read only memory
PS	Polystyrene
PSA	Pressure-sensitive adhesive
PSF	Point spread function
PSI	Primary skin irritation
PSP	Postsynaptic potential
PSR	Proton spin resonance
PSS	Progressive systemic sclerosis
PT	Plasma thromboplastin
PTB	Patellar tendon bearing orthosis
PTC	Plasma thromboplastin component; Positive temperature coefficient; Pressurized personal transfer capsule
PTCA	Percutaneous transluminal coronary angioplasty
PTFE	Polytetrafluoroethylene
PTT	Partial thromboplastin time
PUL	Percutaneous ultrasonic lithotripsy
PURA	Prolonged ultraviolet-A radiation
PUVA	Psoralens and longwave ultraviolet light photochemotherapy
P/V	Pressure/volume
PVC	Polyvinyl chloride; Premature ventricular contraction
PVI	Pressure–volume index
PW	Pulse wave; Pulse width
PWM	Pulse width modulation
PXE	Pseudo-xanthoma elasticum
QA	Quality assurance
QC	Quality control
R-BPC	Reverse bonded phase chromatography
R/S	Radiopaque-spherical
RA	Respiratory amplitude; Right arm
RAD	Right axis deviation
RAE	Right atrial enlargement
RAM	Random access memory
RAP	Right atrial pressure
RAT	Right anterior temporalis
RB	Right bundle
RBBB	Right bundle branch block
RBC	Red blood cell
RBE	Relative biologic effectiveness
RBF	Rose bengal fecal excretion
RBI	Resting baseline impedance
RCBD	Randomized complete block diagram
rCBF	Regional cerebral blood flow
RCC	Right coronary cusp
RCE	Resistive contact electrode
R&D	Research and development
r.e.	Random experiment
RE	Reference electrode
REM	Rapid eye movement; Return electrode monitor
REMATE	Remote access and telecommunication system
RES	Reticuloendothelial system
RESNA	Rehabilitation Engineering Society of North America
RF	Radio frequency; Radiographic-nuoroscopic
RFI	Radio-frequency interference
RFP	Request for proposal
RFQ	Request for quotation
RH	Relative humidity
RHE	Reversible hydrogen electrode
RIA	Radioimmunoassay
RM	Repetition maximum; Right masseter
RMR	Resting metabolic rate
RMS	Root mean square
RN	Radionuclide
RNCA	Radionuclide cineagiogram
ROI	Regions of interest
ROM	Range of motion; Read only memory
RP	Retinitis pigmentosa
RPA	Right pulmonary artery
RPP	Rate pressure product
RPT	Rapid pull-through technique
RPV	Right pulmonary veins
RQ	Respiratory quotient

RR	Recovery room
RRT	Recovery room time; Right posterior temporalis
RT	Reaction time
RTD	Resistance temperature device
RTT	Revised token test
r.v.	Random variable
RV	Residual volume; Right ventricle
RVH	Right ventricular hypertrophy
RVOT	Right ventricular outflow tract
RZ	Return-to-zero
SA	Sinoatrial; Specific absorption
SACH	Solid-ankle-cushion-heel
SAD	Source-axis distance; Statistical Analysis System
SAINT	System analysis of integrated network of tasks
SAL	Sterility assurance level; Surface averaged lead
SALT	Systematic analysis of language transcripts
SAMI	Socially acceptable monitoring instrument
SAP	Systemic arterial pressure
SAR	Scatter-air ratio; Specific absorption rate
SARA	System for anesthetic and respiratory gas analysis
SBE	Subbacterial endocarditis
SBR	Styrene-butadiene rubbers
SC	Stratum corneum; Subcommittees
SCAP	Right scapula
SCE	Saturated calomel electrode; Sister chromatid exchange
SCI	Spinal cord injury
SCRAD	Sub-Committee on Radiation Dosimetry
SCS	Spinal cord stimulation
SCUBA	Self-contained underwater breathing apparatus
SD	Standard deviation
SDA	Stepwise discriminant analysis
SDS	Sodium dodecyl sulfate
S&E	Safety and effectiveness
SE	Standard error
SEC	Size exclusion chromatography
SEM	Scanning electron microscope; Standard error of the mean
SEP	Somatosensory evoked potential
SEXAFS	Surface extended X-ray absorption fine structure
SF	Surviving fraction
SFD	Source-film distance
SFH	Stroma-free hemoglobin
SFTR	Sagittal frontal transverse rotational
SG	Silica gel
SGF	Silica gel fraction
SGG	Spark gap generator
SGOT	Serum glutamic oxaloacetic transaminase
SGP	Strain gage plethysmography; Stress-generated potential
SHE	Standard hydrogen electrode
SI	Le Système International d'Unités
SEBS	Surgical isolation barrier system
SID	Source to image reception distance
SIMFU	Scanned intensity modulated focused ultrasound
SIMS	Secondary ion mass spectroscopy; System for isometric muscle strength
SISI	Short increment sensitivity index
SL	Surgical lithotomy
SLD	Sublethal damage
SLE	Systemic lupus erythemotodes
SMA	Sequential multiple analyzer
SMAC	Sequential multiple analyzer with computer
SMR	Sensorimotor
S/N	Signal-to-noise
S:N/D	Signal-to-noise ratio per unit dose
SNP	Sodium nitroprusside
SNR	Signal-to-noise ratio
SOA	Sources of artifact
SOAP	Subjective, objective, assessment, plan
SOBP	Spread-out Bragg peak
SP	Skin potential
SPECT	Single photon emission computed tomography
SPL	Sound pressure level
SPRINT	Single photon ring tomograph
SPRT	Standard platinum resistance thermometer
SPSS	Statistical Package for the Social Sciences
SQUID	Superconducting quantum interference device
SQV	Square wave voltammetry
SR	Polysulfide rubbers
SRT	Speech reception threshold
SS	Stainless steel
SSB	Single strand breaks
SSD	Source-to-skin distance; Source-to-surface distance
SSE	Stainless steel electrode
SSEP	Somatosensory evoked potential
SSG	Solid state generator
SSP	Skin stretch potential
SSS	Sick sinus syndrome
STD	Source-tray distance
STI	Systolic time intervals
STP	Standard temperature and pressure
STPD	Standard temperature pressure dry
SV	Stroke volume
SVC	Superior vena cava
SW	Standing wave
TAA	Tumor-associated antigens
TAC	Time-averaged concentration
TAD	Transverse abdominal diameter
TAG	Technical Advisory Group
TAH	Total artificial heart
TAR	Tissue-air ratio
TC	Technical Committees
TCA	Tricarboxylic acid cycle
TCD	Thermal conductivity detector
TCES	Transcutaneous cranial electrical stimulation

TCP	Tricalcium phosphate
TDD	Telecommunication devices for the deaf
TDM	Therapeutic drug monitoring
TE	Test electrode; Thermoplastic elastomers
TEAM	Technology evaluation and acquisition methods
TEM	Transmission electron microscope; Transverse electric and magnetic mode; Transverse electromagnetic mode
TENS	Transcutaneous electrical nerve stimulation
TEP	Tracheoesophageal puncture
TEPA	Triethylenepho-sphoramide
TF	Transmission factor
TFE	Tetrafluorethylene
TI	Totally implantable
TICCIT	Time-shared Interaction Computer-Controlled Information Television
TLC	Thin-layer chromatography; Total lung capacity
TLD	Thermoluminescent dosimetry
TMJ	Temporomandibular joint
TMR	Tissue maximum ratio; Topical magnetic resonance
TNF	Tumor necrosis factor
TOF	Train-of-four
TP	Thermal performance
TPC	Temperature pressure correction
TPD	Triphasic dissociation
TPG	Transvalvular pressure gradient
TPN	Total parenteral nutrition
TR	Temperature rise
tRNA	Transfer RNA
TSH	Thyroid stimulating hormone
TSS	Toxic shock syndrome
TTD	Telephone devices for the deaf
TTI	Tension time index
TTR	Transition temperature range
TTV	Trimming tip version
TTY	Teletypewriter
TUR	Transurethral resection
TURP	Transurethral resections of the prostrate
TV	Television; Tidal volume; Tricuspid valve
TVER	Transscleral visual evoked response
TW	Traveling wave
TxB_2	Thrombozame B^2
TZ	Transformation zone
UES	Upper esophageal sphincter
UP	Urea-formaldehyde
UfflS	University Hospital Information System
UHMW	Ultra high molecular weight
UHMWPE	Ultra high molecular weight polyethylene
UL	Underwriters Laboratory
ULF	Ultralow frequency
ULTI	Ultralow temperature isotropic
UMN	Upper motor neuron
UO	Urinary output
UPTD	Unit pulmonary oxygen toxicity doses
UR	Unconditioned response
US	Ultrasound; Unconditioned stimulus
USNC	United States National Committee
USP	United States Pharmacopeia
UTS	Ultimate tensile strength
UV	Ultraviolet; Umbilical vessel
UVR	Ultraviolet radiation
V/F	Voltage-to-frequency
VA	Veterans Administration
VAS	Visual analog scale
VBA	Vaginal blood volume in arousal
VC	Vital capacity
VCO	Voltage-controlled oscillator
VDT	Video display terminal
VECG	Vectorelectrocardiography
VEP	Visually evoked potential
VF	Ventricular fibrillation
VOP	Venous occlusion plethysmography
VP	Ventriculoperitoneal
VPA	Vaginal pressure pulse in arousal
VPB	Ventricular premature beat
VPR	Volume pressure response
VSD	Ventricular septal defect
VSWR	Voltage standing wave ratio
VT	Ventricular tachycardia
VTG	Vacuum tube generator
VTS	Viewscan text system
VV	Variable version
WAIS-R	Weschler Adult Intelligence Scale-Revised
WAK	Wearable artificial kidney
WAML	Wide-angle mobility light
WBAR	Whole-body autoradiography
WBC	White blood cell
WG	Working Groups
WHO	World Health Organization; Wrist hand orthosis
WLF	Williams-Landel-Ferry
WMR	Work metabolic rate
w/o	Weight percent
WORM	Write once, read many
WPW	Wolff-Parkinson-White
XPS	X-ray photon spectroscopy
XR	Xeroradiograph
YAG	Yttrium aluminum garnet
ZPL	Zero pressure level

CONVERSION FACTORS AND UNIT SYMBOLS

SI UNITS (ADOPTED 1960)

A new system of metric measurement, the International System of Units (abbreviated SI), is being implemented throughout the world. This system is a modernized version of the MKSA (meter, kilogram, second, ampere) system, and its details are published and controlled by an international treaty organization (The International Bureau of Weights and Measures).

SI units are divided into three classes:

Base Units

length	meter[†] (m)
mass[‡]	kilogram (kg)
time	second (s)
electric current	ampere (A)
thermodynamic temperature§	kelvin (K)
amount of substance	mole (mol)
luminous intensity	candela (cd)

Supplementary Units

plane angle	radian (rad)
solid angle	steradian (sr)

Derived Units and Other Acceptable Units

These units are formed by combining base units, supplementary units, and other derived units. Those derived units having special names and symbols are marked with an asterisk (*) in the list below:

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
*absorbed dose	gray	Gy	J/kg
acceleration	meter per second squared	m/s^2	
*activity (of ionizing radiation source)	becquerel	Bq	1/s
area	square kilometer	km^2	
	square hectometer	hm^2	ha (hectare)
	square meter	m^2	

[†]The spellings "metre" and "litre" are preferred by American Society for Testing and Materials (ASTM); however, "−er" will be used in the Encyclopedia.

[‡]"Weight" is the commonly used term for "mass."

§Wide use is made of "Celsius temperature" (t) defined $t = T - T_0$ where T is the thermodynamic temperature, expressed in kelvins, and $T_0 = 273.15\,K$ by definition. A temperature interval may be expressed in degrees Celsius as well as in kelvins.

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
*capacitance	farad	F	C/V
concentration (of amount of substance)	mole per cubic meter	mol/m^3	
*conductance	siemens	S	A/V
current density	ampere per square meter	A/m^2	
density, mass density	kilogram per cubic meter	kg/m^3	g/L; mg/cm^3
dipole moment (quantity)	coulomb meter	C·m	
*electric charge, quantity of electricity	coulomb	C	A·s
electric charge density	coulomb per cubic meter	C/m^3	
electric field strength	volt per meter	V/m	
electric flux density	coulomb per square meter	C/m^2	
*electric potential, potential difference, electromotive force	volt	V	W/A
*electric resistance	ohm	Ω	V/A
*energy, work, quantity of heat	megajoule	MJ	
	kilojoule	kJ	
	joule	J	N·m
	electron volt†	eV†	
	kilowatt hour†	kW·h†	
energy density	joule per cubic meter	J/m^3	
*force	kilonewton	kN	
	newton	N	$kg·m/s^2$
*frequency	megahertz	MHz	
	hertz	Hz	1/s
heat capacity, entropy	joule per kelvin	J/K	
heat capacity (specific), specific entropy	joule per kilogram kelvin	J/(kg·K)	
heat transfer coefficient	watt per square meter kelvin	$W/(m^2·K)$	
*illuminance	lux	lx	lm/m^2
*inductance	henry	H	Wb/A
linear density	kilogram per meter	kg/m	
luminance	candela per square meter	cd/m^2	
*luminous flux	lumen	lm	cd·sr
magnetic field strength	ampere per meter	A/m	
*magnetic flux	weber	Wb	V·s
*magnetic flux density	tesla	T	Wb/m^2
molar energy	joule per mole	J/mol	
molar entropy, molar heat capacity	joule per mole kelvin	J/(mol·K)	
moment of force, torque	newton meter	N·m	
momentum	kilogram meter per second	kg·m/s	
permeability	henry per meter	H/m	
permittivity	farad per meter	F/m	
*power, heat flow rate, radiant flux	kilowatt	kW	
	watt	W	J/s
power density, heat flux density, irradiance	watt per square meter	W/m^2	
*pressure, stress	megapascal	MPa	
	kilopascal	kPa	
	pascal	Pa	N/m^2
sound level	decibel	dB	
specific energy	joule per kilogram	J/kg	
specific volume	cubic meter per kilogram	m^3/kg	
surface tension	newton per meter	N/m	
thermal conductivity	watt per meter kelvin	W/(m·K)	
velocity	meter per second	m/s	
	kilometer per hour	km/h	
viscosity, dynamic	pascal second	Pa·s	
	millipascal second	mPa·s	

†This non-SI unit is recognized as having to be retained because of practical importance or use in specialized fields.

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
viscosity, kinematic	square meter per second	m^2/s	
	square millimeter per second	mm^2/s	
	cubic meter	m^3	
	cubic decimeter	dm^3	L(liter)
	cubic centimeter	cm^3	mL
wave number	1 per meter	m^{-1}	
	1 per centimeter	cm^{-1}	

In addition, there are 16 prefixes used to indicate order of magnitude, as follows:

Multiplication factor	*Prefix*	*Symbol*	*Note*
10^{18}	exa	E	
10^{15}	peta	P	
10^{12}	tera	T	
10^{9}	giga	G	
10^{8}	mega	M	
10^{3}	kilo	k	
10^{2}	hecto	h[a]	[a]Although hecto, deka, deci, and centi are
10	deka	da[a]	SI prefixes, their use should be avoided
10^{-1}	deci	d[a]	except for SI unit-multiples for area and
10^{-2}	centi	c[a]	volume and nontechnical use of
10^{-3}	milli	m	centimeter, as for body and clothing
10^{-6}	micro	μ	measurement.
10^{-9}	nano	n	
10^{-12}	pico	p	
10^{-15}	femto	f	
10^{-18}	atto	a	

For a complete description of SI and its use the reader is referred to ASTM E 380.

CONVERSION FACTORS TO SI UNITS

A representative list of conversion factors from non-SI to SI units is presented herewith. Factors are given to four significant figures. Exact relationships are followed by a dagger (†). A more complete list is given in ASTM E 380-76 and ANSI Z210. 1-1976.

To convert from	*To*	*Multiply by*
acre	square meter (m^2)	4.047×10^{3}
angstrom	meter (m)	$1.0 \times 10^{-10\dagger}$
are	square meter (m^2)	$1.0 \times 10^{2\dagger}$
astronomical unit	meter (m)	1.496×10^{11}
atmosphere	pascal (Pa)	1.013×10^{5}
bar	pascal (Pa)	$1.0 \times 10^{5\dagger}$
barrel (42 U.S. liquid gallons)	cubic meter (m^3)	0.1590
Btu (International Table)	joule (J)	1.055×10^{3}
Btu (mean)	joule (J)	1.056×10^{3}
Bt (thermochemical)	joule (J)	1.054×10^{3}
bushel	cubic meter (m^3)	3.524×10^{-2}
calorie (International Table)	joule (J)	4.187
calorie (mean)	joule (J)	4.190
calorie (thermochemical)	joule (J)	$4.184^{\dagger}$
centimeters of water (39.2 °F)	pascal (Pa)	98.07
centipoise	pascal second (Pa·s)	$1.0 \times 10^{-3\dagger}$
centistokes	square millimeter per second (mm^2/s)	$1.0^{\dagger}$

To convert from	*To*	*Multiply by*
cfm (cubic foot per minute)	cubic meter per second (m^3/s)	4.72×10^{-4}
cubic inch	cubic meter (m^3)	1.639×10^{-4}
cubic foot	cubic meter (m^3)	2.832×10^{-2}
cubic yard	cubic meter (m^3)	0.7646
curie	becquerel (Bq)	$3.70 \times 10^{10\dagger}$
debye	coulomb-meter (C·m)	3.336×10^{-30}
degree (angle)	radian (rad)	1.745×10^{-2}
denier (international)	kilogram per meter (kg/m)	1.111×10^{-7}
	tex	0.1111
dram (apothecaries')	kilogram (kg)	3.888×10^{-3}
dram (avoirdupois)	kilogram (kg)	1.772×10^{-3}
dram (U.S. fluid)	cubic meter (m^3)	3.697×10^{-6}
dyne	newton(N)	$1.0 \times 10^{-6\dagger}$
dyne/cm	newton per meter (N/m)	$1.00 \times 10^{-3\dagger}$
electron volt	joule (J)	1.602×10^{-19}
erg	joule (J)	$1.0 \times 10^{-7\dagger}$
fathom	meter (m)	1.829
fluid ounce (U.S.)	cubic meter (m^3)	2.957×10^{-5}
foot	meter (m)	$0.3048^{\dagger}$
foot-pound force	joule (J)	1.356
foot-pound force	newton meter (N·m)	1.356
foot-pound force per second	watt(W)	1.356
footcandle	lux (lx)	10.76
furlong	meter (m)	2.012×10^2
gal	meter per second squared (m/s^2)	$1.0 \times 10^{-2\dagger}$
gallon (U.S. dry)	cubic meter (m^3)	4.405×10^{-3}
gallon (U.S. liquid)	cubic meter (m^3)	3.785×10^{-3}
gilbert	ampere (A)	0.7958
gill (U.S.)	cubic meter (m^3)	1.183×10^{-4}
grad	radian	1.571×10^{-2}
grain	kilogram (kg)	6.480×10^{-5}
gram force per denier	newton per tex (N/tex)	8.826×10^{-2}
hectare	square meter (m^2)	$1.0 \times 10^{4\dagger}$
horsepower (550 ft·lbf/s)	watt(W)	7.457×10^2
horsepower (boiler)	watt(W)	9.810×10^3
horsepower (electric)	watt(W)	$7.46 \times 10^{2\dagger}$
hundredweight (long)	kilogram (kg)	50.80
hundredweight (short)	kilogram (kg)	45.36
inch	meter (m)	$2.54 \times 10^{-2\dagger}$
inch of mercury (32 °F)	pascal (Pa)	3.386×10^3
inch of water (39.2 °F)	pascal (Pa)	2.491×10^2
kilogram force	newton (N)	9.807
kilopond	newton (N)	9.807
kilopond-meter	newton-meter (N·m)	9.807
kilopond-meter per second	watt (W)	9.807
kilopond-meter per min	watt(W)	0.1635
kilowatt hour	megajoule (MJ)	$3.6^{\dagger}$
kip	newton (N)	4.448×10^2
knot international	meter per second (m/s)	0.5144
lambert	candela per square meter (cd/m^2)	3.183×10^3
league (British nautical)	meter (m)	5.559×10^2
league (statute)	meter (m)	4.828×10^3
light year	meter (m)	9.461×10^{15}
liter (for fluids only)	cubic meter (m^3)	$1.0 \times 10^{-3\dagger}$
maxwell	weber (Wb)	$1.0 \times 10^{-8\dagger}$
micron	meter (m)	$1.0 \times 10^{-6\dagger}$
mil	meter (m)	$2.54 \times 10^{-5\dagger}$
mile (U.S. nautical)	meter (m)	$1.852 \times 10^{3\dagger}$
mile (statute)	meter (m)	1.609×10^3
mile per hour	meter per second (m/s)	0.4470

To convert from	*To*	*Multiply by*
millibar	pascal (Pa)	1.0×10^{2}
millimeter of mercury (0 °C)	pascal (Pa)	1.333×10^{2}†
millimeter of water (39.2 °F)	pascal (Pa)	9.807
minute (angular)	radian	2.909×10^{-4}
myriagram	kilogram (kg)	10
myriameter	kilometer (km)	10
oersted	ampere per meter (A/m)	79.58
ounce (avoirdupois)	kilogram (kg)	2.835×10^{-2}
ounce (troy)	kilogram (kg)	3.110×10^{-2}
ounce (U.S. fluid)	cubic meter (m^3)	2.957×10^{-5}
ounce-force	newton (N)	0.2780
peck (U.S.)	cubic meter (m^3)	8.810×10^{-3}
pennyweight	kilogram (kg)	1.555×10^{-3}
pint (U.S. dry)	cubic meter (m^3)	5.506×10^{-4}
pint (U.S. liquid)	cubic meter (m^3)	4.732×10^{-4}
poise (absolute viscosity)	pascal second (Pa·s)	0.10†
pound (avoirdupois)	kilogram (kg)	0.4536
pound (troy)	kilogram (kg)	0.3732
poundal	newton (N)	0.1383
pound-force	newton (N)	4.448
pound per square inch (psi)	pascal (Pa)	6.895×10^{3}
quart (U.S. dry)	cubic meter (m^3)	1.101×10^{-3}
quart (U.S. liquid)	cubic meter (m^3)	9.464×10^{-4}
quintal	kilogram (kg)	1.0×10^{2}†
rad	gray (Gy)	1.0×10^{-2}†
rod	meter (m)	5.029
roentgen	coulomb per kilogram (C/kg)	2.58×10^{-4}
second (angle)	radian (rad)	4.848×10^{-6}
section	square meter (m^2)	2.590×10^{6}
slug	kilogram (kg)	14.59
spherical candle power	lumen (lm)	12.57
square inch	square meter (m^2)	6.452×10^{-4}
square foot	square meter (m^2)	9.290×10^{-2}
square mile	square meter (m^2)	2.590×10^{6}
square yard	square meter (m^2)	0.8361
store	cubic meter (m^3)	1.0†
stokes (kinematic viscosity)	square meter per second (m^2/s)	1.0×10^{-4}†
tex	kilogram per meter (kg/m)	1.0×10^{-6}†
ton (long, 2240 pounds)	kilogram (kg)	1.016×10^{3}
ton (metric)	kilogram (kg)	1.0×10^{3}†
ton (short, 2000 pounds)	kilogram (kg)	9.072×10^{2}
torr	pascal (Pa)	1.333×10^{2}
unit pole	weber (Wb)	1.257×10^{-7}
yard	meter (m)	0.9144†

continued

HYDROCEPHALUS, TOOLS FOR DIAGNOSIS AND TREATMENT OF

Sandeep Sood
Anders Eklund
Noam Alperin
University of Illinois at Chicago
Chicago, Illinois

INTRODUCTION

Epidemiology

A congenital form of hydrocephalus occurs in roughly 50 in 100,000 live births (6). Hydrocephalus may also be acquired later in life as a result of a brain tumor, following meningitis, trauma, or intracranial hemorrhage. It has been estimated that prevalence of shunted hydrocephalus is about 40/100,000 population in the United States (7). Untreated hydrocephalus has a poor natural history with a mortality rate of 20–25% and results in severe physical and mental disabilities in survivors (8,9). There has been a significant reduction in mortality and morbidity with use of shunting. However, shunting is associated with a high failure rate; a 40% failure rate occurs within the first year after shunting (10). Advances in the technology have lead to the development of a diverse type of shunt systems to circumvent problems related to long-term shunting, such as obstruction, infection, and overdrainage. Yet, studies done to evaluate these devices have not shown a significant long- or short-term benefit from their use compared with the conventional devices (10). It is estimated that, during the year 2000, the cost associated with shunting exceeded one billion dollars in the United States alone (11). Shunt replacement accounted for 43% of shunt procedures. Endoscopic surgery has provided an alternative strategy in patients with obstructive hydrocephalus. However, limited data in the literature suggest that long-term survival of third ventriculostomy is not significantly superior to that of a shunt (12).

Physiology

The CSF flow in the craniospinal system is influenced by two separate processes: (1) the circulation of the CSF from its formation sites to its absorption sites (i.e., bulk flow) and (2) an oscillatory (back and forth) flow during the cardiac cycle (pulsatile flow). The first process governs the overall volume of CSF and thereby influences intracranial pressure (ICP). The second process, the oscillatory movement of the CSF within the craniospinal compartments, is caused by the pulsatile blood flow entering and leaving the intracranial compartment during the cardiac cycle. These two processes occur over different time scales; circulation and replenishing of CSF occurs over minutes, whereas the time scale of the pulsatile CSF flow is milliseconds.

CSF Circulation. Unlike other organ systems, the brain and the spinal cord are unique in being bathed in a clear fluid called cerebrospinal fluid. The exact role that it plays in maintaining the necessary environment for the functioning of the nervous system is unclear. It has been ascribed a role in providing nutrition, removing excess waste, circulating neurotransmitters, maintaining the necessary electrolyte environment, and acting as a shock absorber against trauma.

The distribution of nutrients, or neurotransmitters, and removal of waste products of metabolism, is an unlikely function of CSF, because these chemicals are present in very low concentrations in the CSF. The main function of CSF is to provide buoyancy to support the brain and act as a cushion against trauma. The normal brain weighs about 1500 g; however, supported by the buoyancy of the CSF, its apparent weight is reduced to about 50 g in the cranium. Support for its role in cushioning the brain and spinal cord against trauma comes from clinical conditions like severe spinal canal stenosis. The CSF cushion around at the site of stenosis is markedly reduced. As a result, spinal cord injury often occurs even with minor trauma as the shock waves are directly transmitted from the bone to the spinal cord.

Cerebrospinal fluid is made through a complex process that occurs in the cells of the choroid plexus, which lines the margin of the four fluid-filled spaces in the brain called the ventricles. First, an ultrafilterate of plasma is formed in the connective tissue surrounding the choroidal capillaries. Next, this is converted into a secretion by carbonic anhydrase enzyme present in the choroids epithelium. The CSF is made at a fairly constant rate of about 10 mL/h. Most of the CSF is made in the choroids plexus of the lateral ventricles. Roughly, 20% of the CSF comes from the ventricular walls. As most CSF is made in the lateral ventricles, it is traditionally believed that the CSF bulk flow occurs from the lateral ventricles to the third ventricle, fourth ventricle, and then through the foramen of Magendie and Lushka into the cerebello-pontine cistern and on to the surface of the brain and spinal cord (Fig. 1). A fifth of the CSF runs down around the spinal cord and then back to the cranial subarachnoid space.

The CSF is absorbed by the cells of the arachnoid granulations (13). These are present in the superior sagittal sinus. The process involves pinocytosis of a small quanta of CSF, on the subarachnoid side of the granulations, and discharge into the blood on the venous side. The process is driven by a pressure difference of at least 5 mm Hg between the subarachnoid CSF and the superior sagittal sinus. A small proportion of CSF is also absorbed along the perivascular spaces and along the nerve sheaths exiting the spinal canal (14).

This traditional view has been recently challenged. Johnston et al. in experimental and cadaveric studies have demonstrated that a large amount of CSF is present around the olfactory nerve and the cribriform plate area

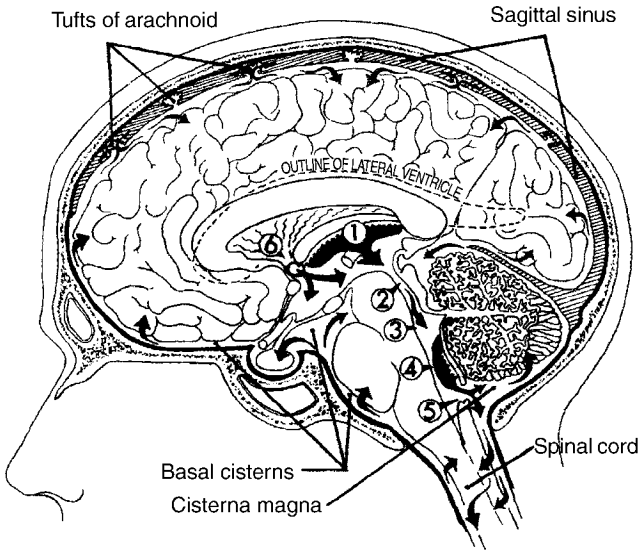


Figure 1. CSF is mainly formed by the choroids plexus in the lateral, third (1), and fourth (4) ventricles. The fluid flows (in the direction of the arrows) from lateral ventricles through the foramen of Monro (6) into the third ventricle. CSF then passes through the aqueduct of Sylvius (2) into the fourth ventricle (3) and exits the fourth ventricle through the foramen of Luschka and Magendie (5) into the cisterna magna and the basal cisterns. The flow is then into the subarachnoid space over the surface of the brain and about the spinal cord. Finally, the fluid is absorbed through the tufts of arachnoid (arachnoid villi) into the sagittal sinus.

and drains into the lymphatic system of the face (15,16). Others believe that CSF may be absorbed directly at the level of the capillaries and perivascular spaces (17).

CSF Pulsations. The pulsatile back and forth movement of CSF between the cranium and the spinal canal with each heartbeat plays a role in modulating the pulsatile cerebral blood flow. Blood flow in the arteries leading blood to the brain is pulsatile, whereas the pulsatility of blood flow in the cerebral veins is considerably attenuated. The displacement of CSF into the spinal canal during the systolic phase helps accommodate the temporary increase in blood volume in the intracranial space, which otherwise has only a limited capacity to accommodate additional volume due to the limited compliance of the intracranial compartment. Reducing the pulsatility of the blood flow through the brain may play a role in diffusion of nutrients to the brain cells from the blood and of waste products from the brain cell to the blood through a less pulsatile flow at the level of the capillaries. As discussed later, MRI measurements of the pulsatile arterial, venous, and CSF flows, to and from the cranium, can now be used to measure intracranial compliance and pressure (ICP), noninvasively. As one of the main roles of shunting is to protect the brain from increased ICP, diagnostic noninvasive measurement of ICP may aid in management decisions in hydrocephalous.

Pathophysiology

Hydrocephalus occurs if there is a mismatch between the CSF production and the absorption. Accumulation of CSF can occur from obstruction to the egress of CSF from the ventricles. This is referred to as obstructive or noncommunicating hydrocephalus. It may also result from impairment of absorption of the CSF at the level of the arachnoid villi or increased resistance to the flow of CSF in the subarachnoid spaces from fibrosis and scarring related to meningitis or previous subarachnoid hemorrhage. This is referred to as communicating hydrocephalus. Irrespective of the cause, the accumulation of CSF has two consequences. It results in an increase in the pressure in the cranium and may cause diltation of the ventricles (ventriculomegaly).

- **Increase in Intracranial Pressure:** Maintaining normal intracranial pressure is important for the functioning of the brain. The pressure in the intracranial cavity increases exponentially with an increase in the total volume of its content (brain tissue, blood, and the CSF) (18). Therefore, increase in intracranial volume, due to uncompensated accumulation of CSF, increases ICP and reduces intracranial compliance. Compliance quantifies the ability of a compartment to accommodate increase in volume for a given increase in pressure and is defined as the ratio of the changes in volume and pressure:

$$\text{Compliance} = \frac{\Delta v}{\Delta p} \tag{1}$$

  where, Δv is change in volume and, Δp is the change in pressure. Intracranial compliance decreases with increased ICP because of the exponential relationship between ICP and intracranial volume (ICV).

Normal ICP is about 1–5 mm Hg in an infant and up to 20 mm Hg in an adult. It is measured by inserting a needle into the spinal canal and recording the pressure using a manometer or by placing a catheter with miniature strain gauge transducer at its distal tip (Codman, Raynham, MA; Camino, Integra LifeSciences, Plainsboro, NJ) directly into the brain parenchyma or the ventricles through a small twist drill hole in the skull. Noninvasive means for measurement of ICP would be important for diagnosis and management of hydrocephalus. Over the last several decades, different approaches have been attempted (19). A method based on measurements of CSF and blood flows to and from the brain by MRI is described in more detail in this article. Increase in ICP can affect the brain in two ways. First, it reduces perfusion of blood into the brain due to the reduced cerebral perfusion pressure (i.e., arterial pressure minus ICP). Depending on the severity and duration, it may result in chronic ischemia causing impairment in higher mental functions, developmental delay in children, or an acute ischemic injury and stroke. Second, rise in pressure in any one of the compartments in the cranium, formed by the tough dural falx in the midline and the tentorium between the cerebral hemispheres superiorly and the cerebellum inferiorly, forces the brain to herniate. This often leads to infarction of the brain stem and death.

- **Symptoms:** Clinically, patients who have elevated ICP generally present with typical symptoms. Headache is the most common. It occurs especially in the

early hours of the morning in initial stages. Low respiratory rate during sleep results in buildup of blood CO_2 and vasodiltation. This aggravates the increased ICP in the early stages of the disease. Vomiting is the next common symptom and probably results either from the distortion of the brain stem vomiting center or its ischemia. Vomiting is often associated with retching and rapid respiration that lowers the blood CO_2 level. This in turn leads to vasoconstriction and lowers the ICP and often results in a transient relief in headaches. Diplopia or double vision is also commonly encountered in a setting of increased ICP. It is a result of stretch of the sixth cranial nerve, which controls the abduction of the eyes. Weakness of ocular abduction disturbs the normal axial alignment of the two eyes resulting in defective fusion of the two images by the brain. Blurring of vision and visual loss may occur in patients with long-standing intracranial hypertension. This results from edema of the optic nerve head as the axoplasmic flow in the neurons of the optic nerve is impaired by the high ICP that is transmitted to the nerve through the patent nerve sheath. Hearing deficits related to similar effect on the cochlea are, however, less frequent. Lethargy or sleepiness is frequently observed in patients with high ICP and is probably from a combination of decreased cerebral perfusion and distortion of the brain stem.

- **Ventricular Enlargement:** Depending on the pathophysiology of hydrocephalus, CSF may accumulate only in the ventricles as in obstructive hydrocephalus or in both the ventricles and the subarachnoid space in communicating hydrocephalus. The increased pressure within the ventricle is transmitted to the periventricular region and results, over time, in loss of neurons, increase in periventricular interstitial fluid, and subsequent gliosis with loss of white matter (20).

When onset of hydrocephalus occurs early in infancy, before the skull sutures have closed, the enlarging ventricles are associated with a progressive increase in head circumference and developmental delay. In later childhood and adults, the increasing ventricular size is associated with symptoms of increased ICP. However, ventricular enlargement may also occur with normal mean ICP, in the elderly patients (21). This is referred to as normal pressure hydrocephalus (NPH). The enlarging ventricle stretches the periventricular nerve fibers. The patient presents not with signs of increase in ICP but with progressive gait ataxia, bladder incontinence, and dementia. Similar presentation may also be observed in adolescents with aqueductal stenosis and obstructive hydrocephalus. These patients with compensated long-standing hydrocephalus have been referred to as long-standing hydrocephalus of adults (LOVA) (22).

It is not clear why the ventricles enlarge preferentially, compared with the subarachnoid space, even though the pressure distributes equally in a closed system. It has been argued that, rather than the actual mean ICP, it is the pulse pressure that determines ventricular diltation. Di Rocco et al. (23) have shown that ventricular enlargement could be induced by an intraventricular pulsatile balloon with a high pulse pressure, despite the mean ICP being normal. It may be argued that in a pulsatile system, it is the *root mean square (*RMS) of the pressure, rather than the mean pressure, that is the cause of enlarged ventricles. It has been suggested that, in communicating hydrocephalus, decrease in compliance may be responsible for preferential transmission of the pulsations to the ventricles (24,25). However, others have shown that in acute or chronic communicating hydrocephalus, the pulse pressure and the pressure waveforms in the SAS and the ventricles are similar (26,27). Alternative explanations offered are that the pia over the cortical surface is more resilient than ependyma that lines the ventricular wall; the venous pressure in the periventricular region is lower, making it more deformable than the subcortical area (28).

DIAGNOSTIC METHODS

Measurement of Resistance to CSF Reabsorption

The hydrodynamics of the craniospinal system is governed by patient-specific properties like CSF formation rate, CSF reabsorption resistance (historically termed as outflow resistance), venous pressure in the sinus, and craniospinal compliance. Together with the periodic variations in ICP, due to blood volume variation from the heartbeat and vasomotion, these properties describe the CSF dynamics, which provide the working environment of the brain. When this environment is disturbed, it affects the function of the brain resulting in the clinical symptoms of hydrocephalus. After shunting, symptoms are often eliminated or reduced. It shows that a clinical improvement can be accomplished by actively changing the brain's working environment. This link among CSF dynamics, brain function, symptoms, and shunting has made researchers look for CSF dynamical means to identify patients that would benefit from a shunt surgery. Outflow resistance has been suggested as a strong predictive parameter in communicating hydrocephalus. Invasive infusion tests in conjunction with a mathematical model of the craniospinal system can be used to estimate CSF absorption rate. The most accepted model for the system hydrodynamics has been proposed by Marmarou (29).

The basic assumptions for the model are as follows:

- CSF reabsorption rate is linearly dependent on the difference between the intracranial and venous pressures (the outflow resistance describes this linear relationship)
- A pressure-dependent compliance
- A constant formation rate of CSF, independent of ICP

The model can be displayed as an electrical analogy (Fig. 2). The model is described mathematically as a differential equation of the time-dependent ICP as a function of external infusion and the governing physical parameters:

$$\frac{dP_{\mathrm{IC}}(t)}{dt}+\frac{K}{R_{\mathrm{out}}}[P_{\mathrm{IC}}(t)]^2-\left(K\cdot I_{\mathrm{infusion}}(t)+\frac{K\cdot P_{\mathrm{r}}}{R_{\mathrm{out}}}\right)P_{\mathrm{IC}}(t)=0 \tag{2}$$

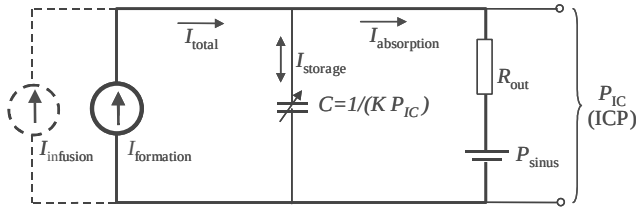


Figure 2. Model of the dynamics of the CSF system. $I_{formation}$ is the CSF formation rate, C is the pressure-dependent compliance described by the elastance parameter K, R_{out} is outflow resistance, P_{sinus} is the venous pressure in the sinus, and P_{IC} is the intracranial pressure. $I_{infusion}$ is the option of external infusion of artificial CSF.

where R_{out} is outflow resistance, P_{IC} is the intracranial pressure, P_r is the ICP at rest, and K is the elastance.

Estimation of CSF outflow resistance and the other system parameters requires perturbation of the system steady state by infusion of fluid into the craniospinal system, either through a lumbar or a ventricular route. Typically, one or two needles are placed in the lumbar canal. When two needles are used, one is connected to a pressure transducer for continuous recording of the dynamic changes in ICP following the infusion, and the other one for infusion and/or withdrawal of the fluid. Different protocols of infusion will lead to unique mathematical solutions. In addition to the resting pressure (also refereed to as opening pressure), which always is determined during these investigations, it is generally believed that the outflow resistance is the clinically most important parameter, but compliance has also been proposed as a predictor for outcome after shunting.

Bolus Infusion. An example of ICP recoding during the bolus infusion test is shown in Fig. 3. The approach is to first determine the compliance from the ratio of the injected volume and the magnitude of pressure increase (30). A pressure volume index (PVI), which describes compliance, is calculated through the expression:

$$\mathrm{PVI} = \frac{\Delta V}{\log(P_p/P_0)} \tag{3}$$

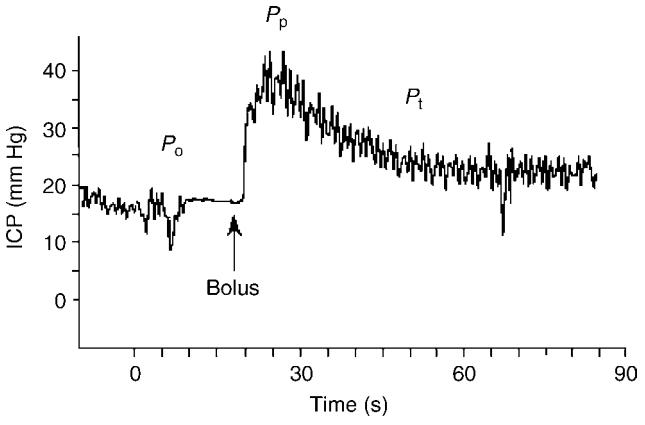


Figure 3. ICP curve from a bolus injection of 4 mL. Figure from Marmarou et al. (30).

where ΔV is the infused volume, P_p is the peak pressure and P_0 is the initial pressure just before the infusion. The next step is to determine R_{out} from the spontaneous relaxation curve when the ICP returns toward the resting pressure (Fig. 3). Solving the differential equation for the relaxation phase after the bolus infusion gives the following expression for R_{out} as a function of time (31):

$$R_{out} = \frac{tP_0}{\mathrm{PVI}\ \log\left[\dfrac{(P_t/P_p)(P_p - P_0)}{P_t - P_0}\right]} \tag{4}$$

where t is the time in seconds after the bolus and P_t is the measured pressure at time t on the relaxation curve. From each bolus, a number of values of R_{out} are calculated and averaged, for example at $t = 1$ min, 1.5 min, and 2 min. The bolus procedure is usually repeated a couple of times for increased measurement reliability.

Constant Pressure Infusion. In this infusion protocol, several constant ICP levels are created. This is done by using a measurement system that continuously records the ICP and regulates it by controlling the pump speed of an infusion pump (Fig. 4) (32). The net infusion rate needed to sustain ICP at each pressure level is determined, and a flow versus pressure curve is generated (Fig. 4). Using linear regression, the outflow resistance is then determined from the slope of that curve (33), because at steady state, the differential equation reduces to

$$I_{inf} = \frac{1}{R_{out}} P_{IC} - \frac{P_r}{R_{out}} \tag{5}$$

where P_{IC} is the mean intracranial pressure on each level, P_r is the resting pressure, and I_{inf} is the net infusion flow at each level.

The constant pressure method can also be used to estimate the CSF formation rate. This is done by lowering the ICP beneath the venous pressure, i.e., below 5 mm Hg. At that ICP, no CSF reabsorption should take place. Therefore, the net withdrawal of CSF needed to sustain that constant pressure level should equal the formation rate.

Constant Flow Infusion. In this method, both the static and the dynamic behavior of the CSF system can be used to estimate outflow resistance (34). In a steady-state analysis R_{out} can be calculated from the stable ICP value associated with a certain constant infusion rate. R_{out} is then estimated by the following expression:

$$R_{out,stat} = \frac{P_{level} - P_r}{I_{inf}} \tag{6}$$

where $R_{out,stat}$ is a static estimation of R_{out}, P_{level} is the new equilibrium pressure obtained at the constant infusion rate, P_r is the resting pressure, and I_{inf} is the infusion rate (Fig. 5).

R_{out} can also be estimated from the dynamic phase during the pressure increases toward the new equilibrium (Fig. 5). This procedure will also give an estimate of the craniospinal compliance (elastance). The differential equation is now solved for a condition of a constant infusion rate, and the solution is fitted against the recorded pressure

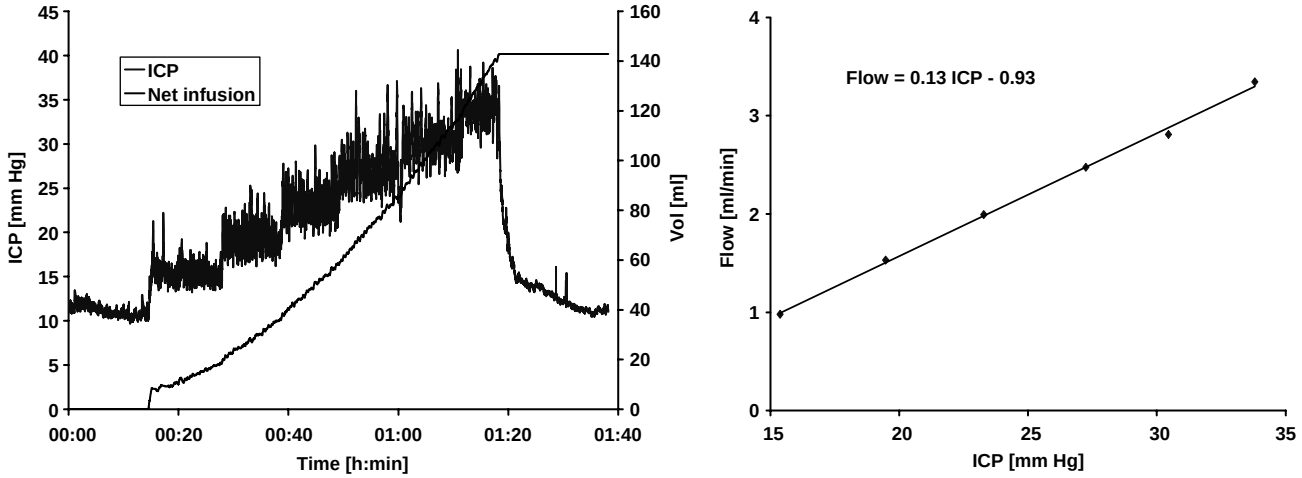


Figure 4. Pressure and flow curves for the constant pressure method. Left graph shows typical data for ICP and infused CSF volume versus time. Right graph shows the mean pressure and flow points determined from each steady-state level. Outflow resistance corresponds to the inverse of the slope.

curve. The time-dependent pressure increase is described by the following expression:

$$P(t) = \frac{\left(I_{\text{inf}} + \frac{P_{\text{r}} - P_0}{R_{\text{out,dyn}}}\right) \cdot (P_{\text{r}} - P_0)}{\frac{P_{\text{r}} - P_0}{R_{\text{out,dyn}}} + I_{\text{inf}} \left[e^{-K\left(\frac{P_{\text{r}} - P_0}{R_{\text{out,dyn}}} + I_{\text{inf}}\right) t} \right]} + P_0 \qquad (7)$$

where K is the elastance and P_0 is a reference pressure that is suggested to be equal to venous sinus pressure. Fitting against data will result in estimations of the unknown parameters $R_{\text{out,dyn}}$, K, and P_0.

In summary, CSF infusion tests are conducted to reveal parameters describing the hydrodynamics of the craniospinal system. An active infusion of artificial CSF is performed, and the resulting ICP response is recorded, and parameters such as outflow resistance, compliance, formation rate, and the venous sinus pressure are then estimated based on a proposed mathematical model. Outflow resistance values determined with the bolus method are usually lower than the values determined with the constant infusion and constant pressure methods. The reason for this difference is not well understood at this time. Determination of R_{out} is often used as a predictive test in hydrocephalus, and it has been stated that if the outflow resistance exceeded a certain threshold, it is an excellent predictor of clinical improvement after shunting (35). In a recent guideline for idiopathic normal pressure hydrocephalus, measurement of R_{out} is included as a supplementary test for selecting patients suitable for shunt surgery (36).

DIAGNOSIS WITH IMAGING

Cross-sectional imaging is routinely used in the diagnosis of hydrocephalous. CSF spaces are well visualized with CT and MRI. In CT images, CSF spaces appear darker due to the lower atomic density of the CSF compared with that of brain tissue. MRI provides an excellent soft-tissue contrast resolution and is considered the primary imaging modality for brain imaging. With MRI, CSF spaces can appear either darker or brighter compared with its surrounding tissues depending on the imaging technique. An example of a CT image and MRI images demonstrating abnormally large CSF spaces is shown in Fig. 6. Cross-sectional imaging enables quantitative assessment of the CSF spaces as well

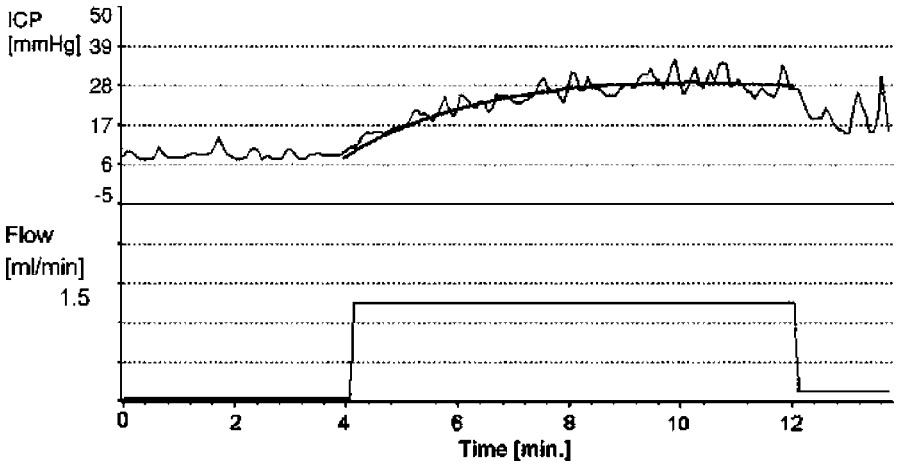


Figure 5. ICP data from a constant infusion investigation. Figure modified from Czosnyka et al. (34).

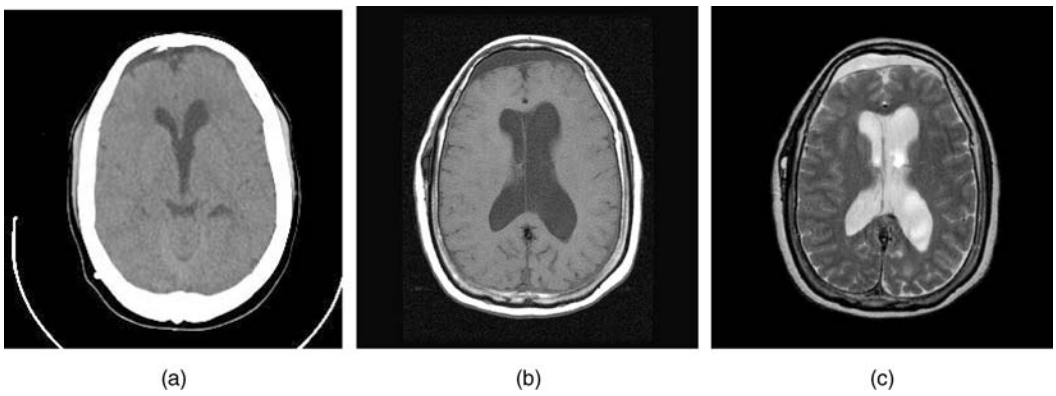


Figure 6. An example of a (a) CT image and (b and c) MRI images demonstrating abnormally large CSF spaces. The appearance of CSF in MRI depends on the technique used to acquire the images; its (b) dark with a T1 technique and (c) bright with a T2 technique.

as 3D reconstruction of the geometry of the ventricular system. The 3D model is obtained by segmentation of the CSF spaces in each of the 2D slices. An example of 3D models of the ventricular system from MRI data demonstrating normal size ventricles and enlarged ventricles are shown in Fig. 7a and b, respectively.

MRI-based motion-sensitive techniques capable of imaging flow are gaining an important role in the diagnosis of hydrocephalus. In particular, dynamic phase-contrast techniques provide images of velocities (velocity-encoded images). The degree of brightness in these images is proportional to the direction and the speed of the moving fluid or tissue. Dynamic (cine) phase contrast images are used to visualize the back and forth flow through the different CSF pathways. The cine phase contrast MRI (PCMRI) technique is also used to derive quantitative parameters such as CSF volumetric flow rate through the aqueduct of Sylvius, from which the CSF production rate in the lateral ventricles can be estimated (37), and intracranial compliance and pressure (19,30).

MRI-Based Measurement of Intracranial Compliance and Pressure

The noninvasive measurement of compliance and pressure uses the cardiac pulsations of the intracranial volume and pressure (30,38). This method is the noninvasive analogs to the measurement of intracranial compliance with the previously described bolus infusion method where the volume and pressure changes are calculated from the MRI measurements of CSF and blood flows to and from the brain. Intracranial elastance, i.e., a change in pressure due to a small change in volume, or the inverse of compliance, is derived from the ratio of the magnitudes of the changes in volume and pressure, and the pressure is then derived through the linear relationship between elastance and pressure. The MRI method measures the arterial, venous, and CSF flows into and out of the cranial vault. A small-volume change, on the order of 1 mL, is calculated from the momentary differences between inflow and outflow at each time points in the cardiac cycle. The pressure change is proportional to the pressure gradient change, which is calculated from time and spatial derivatives of the CSF velocities using fluid dynamics principles.

A motion-sensitive MRI technique, cine phase contrast, provides a series of images where the value at each picture element is proportional to the velocity at that location. The phase contrast MRI technique is based on the principle that the precession frequency of the protons is proportional to the magnetic field strength. Therefore, velocity can be phased-encoded by varying the magnetic field in space and time, i.e., generating magnetic field

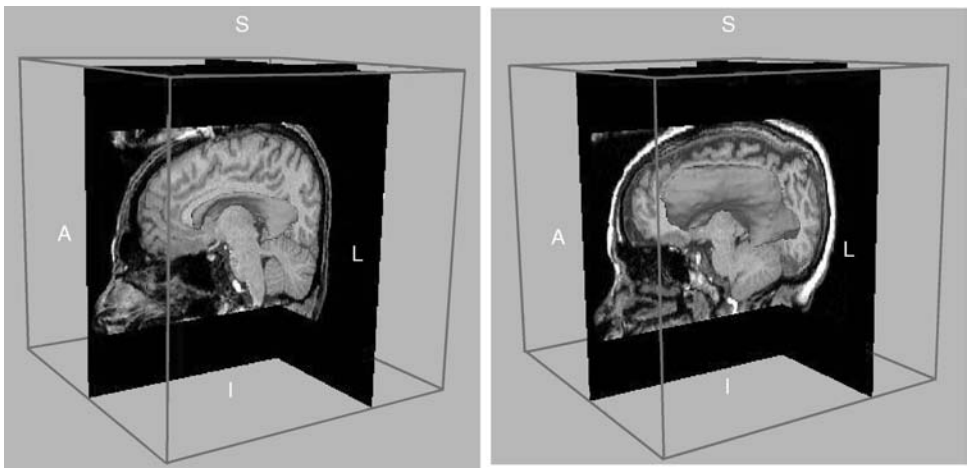


Figure 7. Volume rendering of the CSF spaces inside the brain (i.e., ventricles) generated using segmented MRI data from a (left) healthy volunteer and from a (right) hydrocephalic patient.

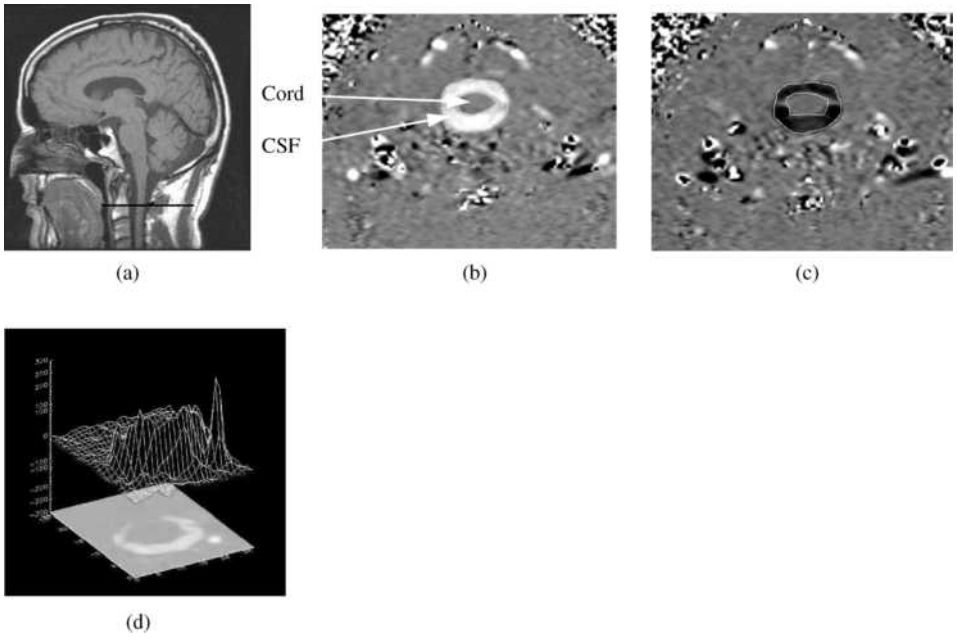


Figure 8. (a) Anatomical mid-sagittal T1-weighted MR image showing the location of the axial plane used for CSF flow measure- ment (dark line). (b and c) Phase-contrast MRI images of CSF flow in the spinal canal. (b) CSF flow during systole. (c) CSF flow during diastole. The pixel values in these images are proportional to velocities in a direction perpendicular to the image plane. Gray-static tissue, white-outward flow (caudal direction), and black-inward flow (cranial direction). (d) A 3D plot of the CSF velocities during systole.

gradients. When a gradient field is applied along an axis for a short time, the proton's phase will change based on its location along that axis. When a bipolar (positive and then negative) gradient field is applied, the phase of the stationary protons will increase during the positive portion (lobe) of the bipolar gradient and then will decrease during the negative lobe. If the lobes were of equal area, no net phase change would occur. However, moving protons, such as those in the blood or CSF, will experience different field strength during each lobe due to their change in position; this will result in a net phase change proportional to the proton velocity.

Examples of MRI phase contrast images of CSF and blood flow are shown in Figs. 8 and 9, respectively. The oscillatory CSF flow between the cranial and the spinal compartments is visualized in images taken in a transverse anatomical orientation through the upper cervical spinal canal. The location of this plane is indicated on a mid-sagittal scout MR image shown in Fig. 8a. Fig. 8b depicts outflow (white pixels) during systole, and Fig. 8c depicts

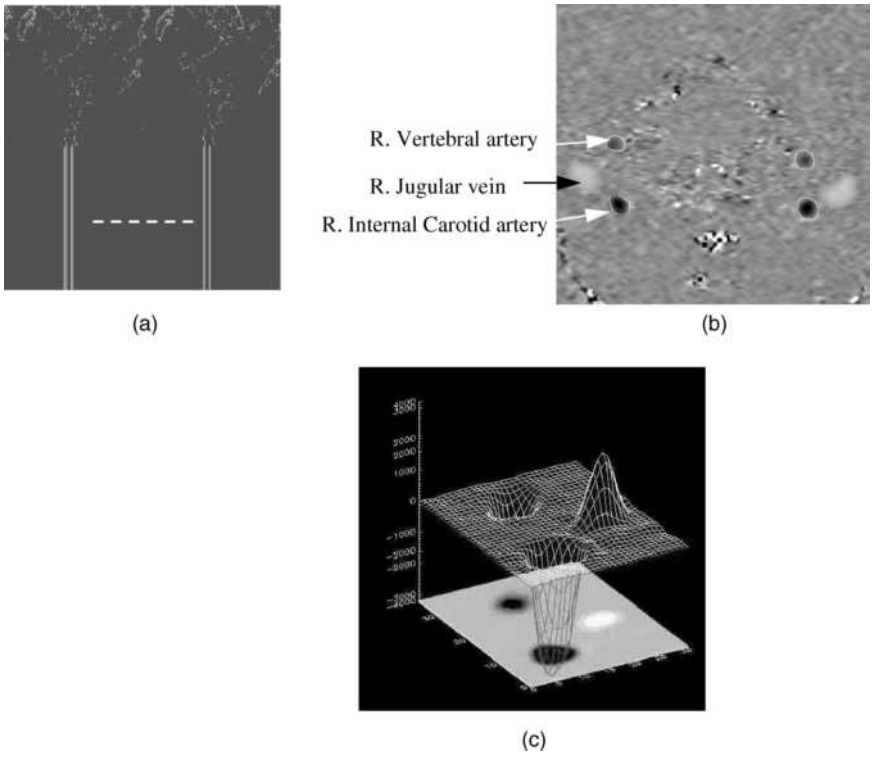


Figure 9. (a) A blood vessel MRI scout image showing the location of the axial plane for blood flow measurement (dash line). (b) A phase contrast MRI image of blood flow through that location. Black pixels indicate arterial inflow, and white are venous outflow. (c) A 3D plot of the blood flow velocities.

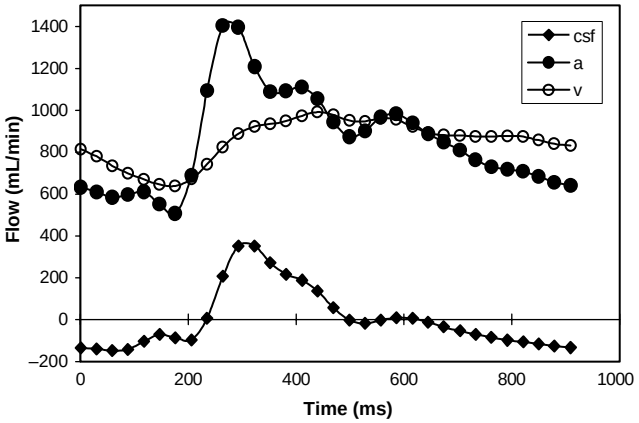


Figure 10. The volumetric flows into and out of the cranial vault during the cardiac cycle derived from the MRI phase contrast scans. Total arterial inflow (filled circles), venous outflow (open), and the cranial-to-spinal CSF volumetric flow rate (diamonds) during the cardiac cycle. Note that arterial inflow is greater than venous outflow during systole.

inflow (black pixels) during diastole. Fig 9d depicts a 3D plot of the velocities in a region of interest containing the CSF space and an epidural vein. The CSF flow is imaged with a low-velocity encoding, and the faster blood flow through the neck arteries and veins is imaged using high-velocity encoding. The location of the imaging plane used for blood flow measurement is shown in Fig. 9a, and a velocity encoded image of blood flow is shown in Fig. 9b. Fig 9c depicts a 3D plot of the velocities in a region of interest containing the internal carotid and vertebral arteries and the jugular vein.

Volumetric flow rates are obtained by integration of the velocities throughout a lumen cross-sectional area. The total volumetric arterial flow rate—that is, total cerebral blood flow—is calculated directly from the sum of the volumetric flow through the four vessels carrying blood to the brain (internal carotid and vertebral arteries). The venous blood outflow is obtained by summation of the flow through the jugular veins, and through secondary venous outflow channels such as the epidural, vertebral, and deep cervical veins when venous drainage occurs through these veins. An example of the volumetric flow waveforms for CSF, arterial inflow, and venous outflow measured in a healthy volunteer is shown in Fig. 10.

The rate of the time-varying intracranial volume change (net transcranial volumetric flow rate) is obtained by subtracting outflow rates from inflow rates at each time point. The intracranial volume change (delta of volume from a given reference point) is obtained by integrating that waveform with respect to time. Waveforms of the net transcranial volumetric flow rate and the change in the intracranial volume are shown in Fig. 11.

The magnitude of the change in intracranial pressure during the cardiac cycle (pulse pressure) is proportional to that of the CSF pressure gradient waveform. A method to measure pressure gradient of pulsatile flow in tubes with MRI was reported by Urchuk and Plewes (39). Pulsatile pressure gradients are derived from the MRI velocity-encoded phase contrast images using the Navier–Stokes relationship between pressure gradient and temporal and spatial derivatives of the fluid velocity for incompressible fluid in a rigid tube (40). Pressure traces from invasive recordings obtained invasively in patients with low and elevated ICP with an intraventricular pressure transducer and the corresponding CSF pressure gradient waveforms derived from the MRI measurements of the CSF velocities at low- and high-pressure states are shown in Fig 12. The ratio of the magnitude of the pressure and volume changes, i.e., intracranial elastance, is then expressed in terms of MR-ICP based on the linear relationship between elastance and ICP.

DEVICES FOR TREATMENT

Despite significant advances in understanding of the pathophysiology of hydrocephalus, the gold standard for the treatment of hydrocephalus still continues to be CSF diversion through a tube shunt to another body cavity. Unfortunately, treatment with CSF shunts is associated with multiple complications and morbidity. The rate of shunt malfunction in the first year of shunt placement is 40%, and, thereafter, about 10% per year. The cumulative risk of infection approaches 20% per person although the risk of infection per procedure is only 5–8% (41). The technological advances in shunt valve designs and materials have had only a marginal impact on the rate of complications. Third ventriculostomy has become popular in recent years for management of obstructive hydrocephalus, but many questions about its long-term permanence remain controversial. Choroid plexectomy (42,43) aimed at arresting hydrocephalus by reducing CSF production or pharmacotherapy with similar intentions have had very limited success in selected patients.

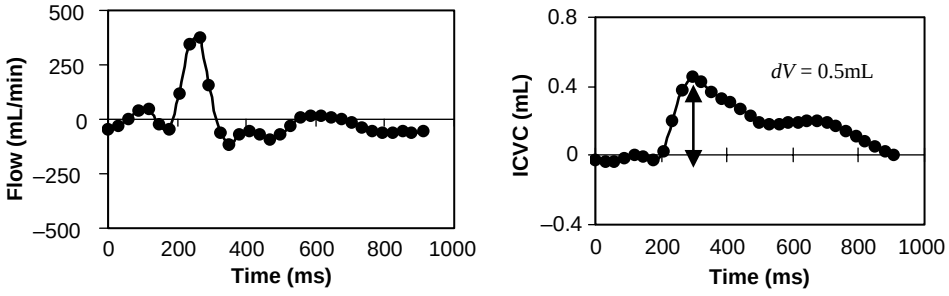


Figure 11. (Left) The MRI-derived net transcranial volumetric flow rate waveform. (Right) The intra cranial volume change during the cardiac cycle derived by integrating the net transcranial volumetric flow waveform on the left. Note that the maximal volume change in this subject is 0.5 mL.

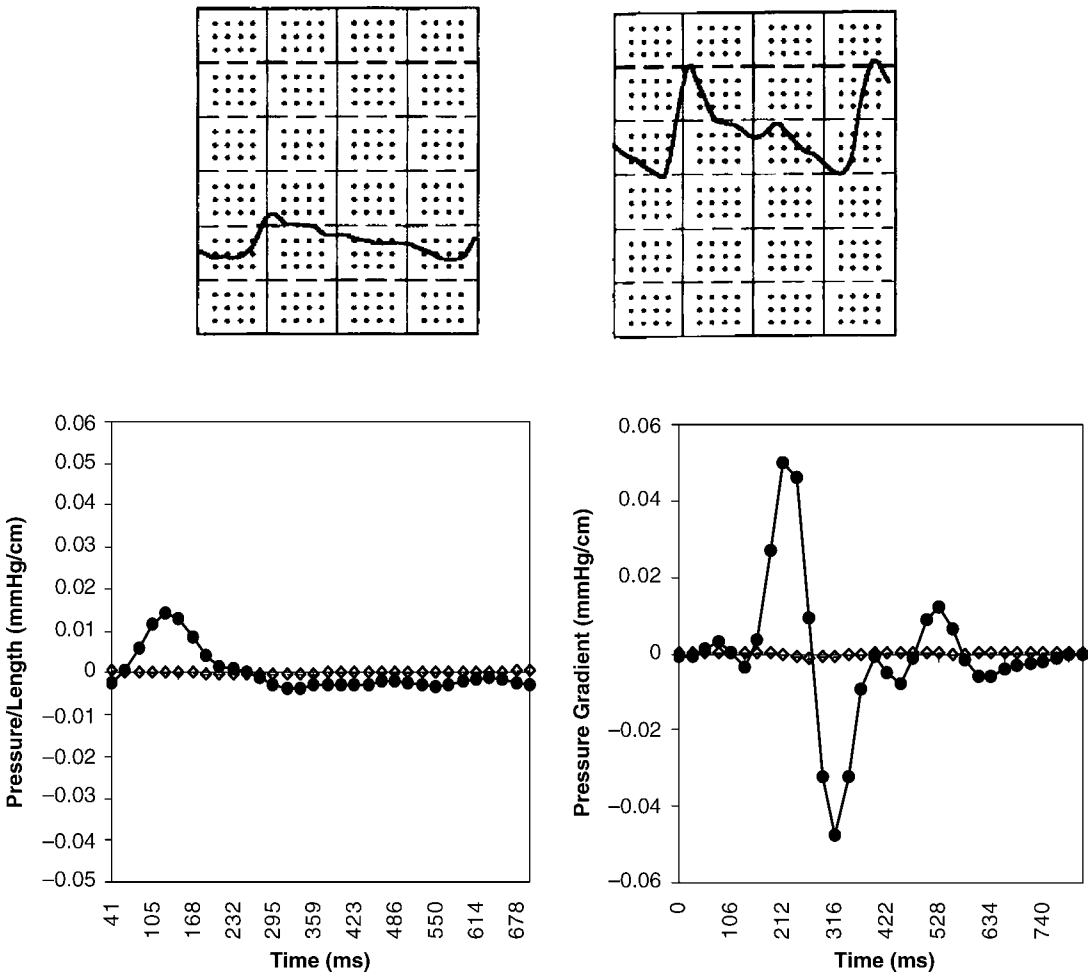


Figure 12. Invasive pressure traces (top) obtained with an intra-ventricular catheter from two patients with (left) low and (right) elevated ICP. The corresponding MRI-derived CSF pressure gradients are shown at the bottom.

Nonobstructive Hydrocephalus

No treatment other than CSF diversion has been effective in management of this form of hydrocephalus. The CSF may be diverted from the ventricles through a catheter that runs in the subcutaneous tissue into the abdominal cavity where it is absorbed by the peritoneum (*ventriculo- peritoeal shunt)* (Fig. 13). It may also be diverted from the spinal subarachnoid space by a lumbar shunt that diverts it to the peritoneum (*Lumbar-peritoneal shunt)*. Lumbar CSF diversion avoids the potential risk of brain injury by the ventricular catheter. Lumbar shunts have a lower risk of obstruction and infection (44) but are more prone to malfunction from mechanical failures (45), and, the development of hind brain herniation, over a period of time, has been well documented (46,47). Evaluation of a lumbar shunt for function is more cumbersome than that of a ventricular shunt. The lumbar shunt is usable in patients with communicating hydrocephalus, small ventricles, and patients who have had multiple ventricular shunt malfunctions.

In patients who cannot absorb CSF from the peritoneum due to scarring from previous operations or infections, the CSF may be diverted to the venous system through a catheter placed at the junction of superior vena cava and the right atrium (*ventriculo/lumbar-atrial shunt*).

A typical shunt system consists of three parts (Fig. 14). First, the proximal catheter, i.e, the catheter, is inserted into the ventricle or the lumbar subarachnoid space. Second, the valve controls the amount of CSF that flows through the shunt system, and third, the distal catheter drains the CSF from the valve to the peritoneum or the atrium.

Proximal Catheter

Three basic types of proximal catheter designs are available: simple with multiple perforations (Codman, Raynham, MA; PS Medical, Goleta, CA), simple Flanged (Heyer-Schulte), Integra, Plainsboro, NJ; Anti-Blok (Phoenix Vygon Neuro, Valley Forge, PA) with receded perforations. The last two have been designed to minimize the growth of choroid plexus into the perforations and causing obstruction. There is no controlled study to suggest that these two designs are in any way superior to simple perforations. The flanged catheters can get stuck, as choroid plexus grows around it, making removal of an obstructed catheter difficult (48).

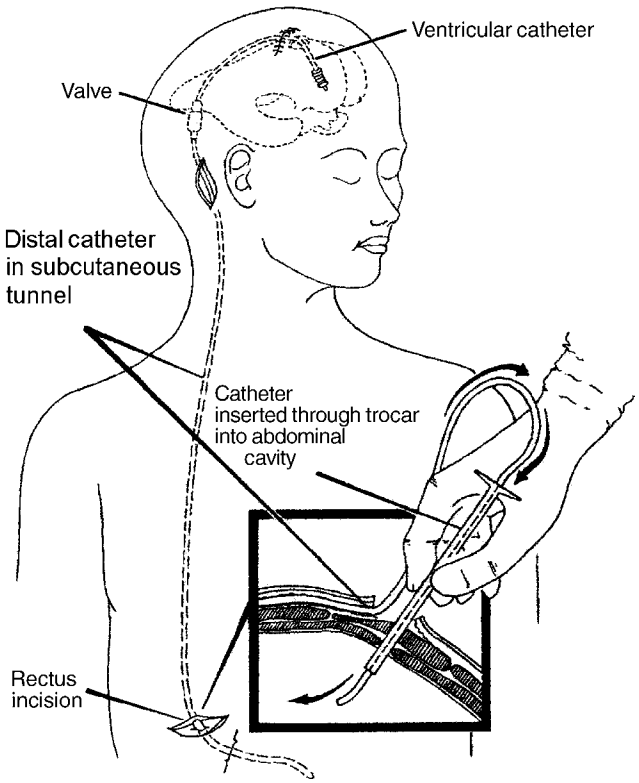


Figure 13. The shunt consists of three parts. The ventricular catheter enters the skull through a burr hole in the skull and passes through the brain into the lateral ventricle. It is connected to the valve that is placed in the subcutaneous tissue of the scalp. The valve in turn is connected to the distal catheter that runs in the subcutaneous tissue to enter the peritoneal cavity of the abdomen as shown in the inset (*ventriculo-peritoneal shunt*) or into the jugular vein and through it to the superior vena cava (*ventriculo-atrial shunt*).

Placement of the proximal catheter has generated considerable controversy in the literature (48–51). More recently, endoscopic placement of the proximal catheter into the frontal horn, away from the choroid plexus, has been advocated to minimize proximal malfunction (3,52,53). Again no controlled study has been done to confirm whether placement of the proximal catheter into frontal or occipital horn is superior to placement in the body of the lateral ventricle. Often catheters that are grossly malpositioned may continue to work, whereas those that are well positioned may fail. The choice of the site, frontal or parietal, may be made on the basis of the above although some studies have suggested a higher incidence of seizure with catheters placed via a frontal burr-hole (49). A study to evaluate use of endoscope to place the shunt catheter in the frontal horn failed to show any benefit (54). This suggests that no matter where the catheter is placed, the flow of CSF toward the catheter causes the choroids plexus to creep toward the catheter, ultimately causing ingrowth and obstruction of the catheter (55).

To remove an obstructed catheter, intraluminal coagulation of the choroid plexus is done using a stylet and low-voltage diathermy, at the time of shunt revision (56–58). Massive intraventricular hemorrhage may occur if the choroid plexus is torn while forcefully removing the catheter. Delayed subarachnoid hemorrhage from rupture of pseudoaneurysm resulting from diathermy of a catheter close to anterior cerebral artery has been reported (59). At times, if the ventricular catheter is severely stuck, it is advisable to leave it in position but occlude it by a ligature and clip. This may become necessary as sometimes an occluded catheter may become unstuck over time and begin to partially function, resulting in formation of subgaleal CSF collection. Replacing a new catheter into the ventricle in patients with small or collapsed ventricles can be sometimes challenging. In most instances, after removal of the old catheter, the new catheter can be gently passed into the ventricle through the same tract. Frameless stereotaxis (StealthStation, Medtronics, Goleta, PA) is now available and may offer an alternative to cumbersome and time-consuming frame-based stereotactic catheter placement (52).

Valve

The valve regulates the amount of CSF that is drained. The aim is to maintain normal ICP. The simplest valves are differential pressure valves. The CSF drainage in these valves is based on the pressure difference between the proximal and the distal ends. Three major configurations are available (Fig. 14): diaphragm, slit valve, and ball-spring mechanism in different pressure ranges (low, medium, and high). Recently, valves in which the pressure setting can be changed with a magnetic wand have become available. These programmable valves allow pressure changes over different pressure ranges based on the manufacturer. The pressure setting on the valve can be ascertained by X ray of the head in the Medos valve (Codman, Raynham, MA) or using a magnetic wand in the Strata valve (Medtronics, Goleta, CA). To prevent inadvertent changes in the valve setting by stray magnetic fields, the Polaris valve (Sophysa, Costa Mesa, CA) has an ingenious locking mechanism that allows changes only if the magnetic field has a certain configuration.

Slit valves tend to be the most inaccurate in their performance followed by ball and spring valves. The diaphragm valves proved to be most stable in long-term tests. Most valves, like the slit valves, ball–spring, and diaphragm valves, offer a lower resistance (<2.5 mm Hg/mL/min) than the normal physiological CSF outflow of 6–10 mm Hg/mL/min. The standard distal tubing of 110 cm increases the overall resistance to 50–80% of the physiological value (60).

Standard differential pressure valves are available in different pressure ranges. It is unclear whether it makes a difference in an ambulatory patient to use a low-, medium-, or high-pressure valve because in the upright position irrespective of the rating the hydrostatic column converts all differential pressure valves into "negative" pressure valves (61). The overdrainage results in persistent headaches from low ICP, ventricular collapse, and increased risk of shunt obstruction. Long-term changes in cerebrovenous physiology cause acute and severe increase in ICP without enlargement of ventricles at the time of shunt malfunction (62). To circumvent the overdrainage in the

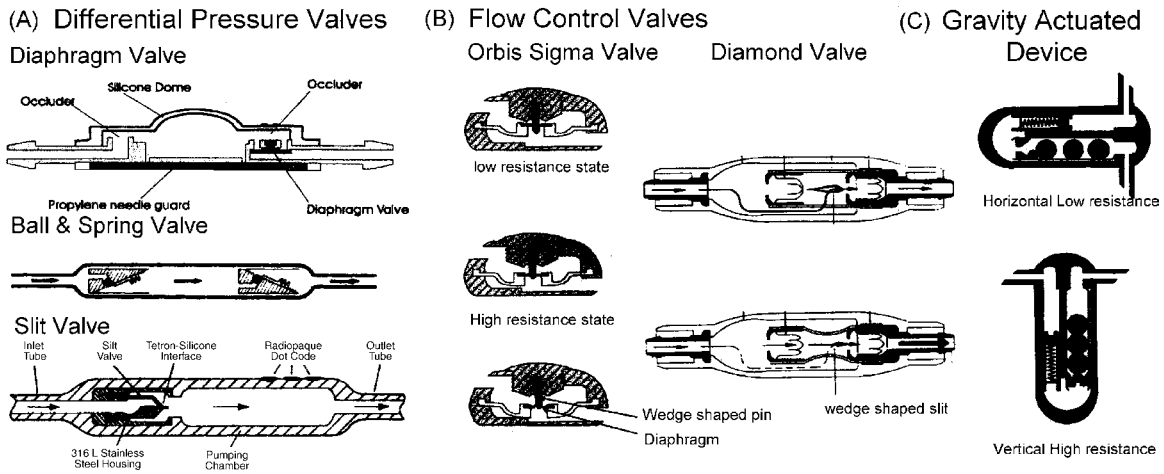


Figure 14. Three major types of valve designs are available. (a) Differential pressure valves allow flow in proportion to the pressure difference between the proximal catheter and the distal catheter. Configurations are a simple diaphragm, ball and spring, or slit valve. Programmable differential pressure valves can be programmed to a pressure setting using a magnetic wand. In the upright position, due to a negative pressure from the hydrostatic column of fluid in the distal catheter, these valves tend to overdrain causing negative pressure symptoms. (b) Flow control valves have the ability to limit overdrainage from a negative hydrostatic pressure gradient. The Orbis-Sigma Valve has a wedge-shaped pin over which the orifice of the diaphragm (arrows) rests. When the distal pressure becomes negative in the upright position, the diaphragm slides downward on the pin narrowing the drainage channel and hence reducing the flow rate. The Diamond valve has a wedge-shaped slit in the construct (arrows) that narrows as the distal pressure becomes increasingly negative again reducing the flow rate. (c) Gravity actuated devices reduce drainage in the upright position by increase in resistance to flow of CSF from the weight of metal balls in the drainage channel.

upright position, ingenious devices, also referred to as devices for reducing siphoning (DRS), have been developed (Fig. 14). The Anti-Siphon device (Integra LifeSciences) has a flexible diaphragm that mechanically senses atmospheric pressure and shuts off the drainage channel if the hydrostatic pressure in the fluid column becomes negative. Flow control valves (Orbis Sigma Valve, Integra LifeSciences and Diamond Valve, Vygon Neuro, Valley Forge, PA) have a drainage channel that narrows as the differential pressure increases in the upright position to reduce the flow. Gravity actuated devices (Gravity Compensating Accessory, Integra LifeSciences, CA, Chabbra Shunt) have metal balls that fall over one another in the upright position to increase resistance to flow. Double channel devices (Dual Switch Valve, Christoph Miethke GmBH & Co KG; SiphonGuard, Codman) have two channels; the low-resistance channel is shut off by a gravity actuated ball in the upright position.

There is no evidence to suggest that use of one type of valve is superior to the other, and several valve designs are available in the market today. A recent multicentric study, evaluating three basis types of valves, failed to confirm the utility of flow control or anti-siphon valves in children and infants over the differential pressure valves (10). Similarly, studies have failed to show that programmable devices are superior to fixed pressure valves (63). Over a period of time, the ventricle tended to become small irrespective of the type of valve used. The rate of proximal malfunction in a patient with flow control valves was 6.5% compared with 42–46% for the other two valves, although the overall rate of malfunction and shunt survival was not statistically different. The design of the flow control valves with a narrow orifice makes it sensitive to malfunction (64). Certainly, revising a valve has less morbidity and risk of neurological injury than revising the proximal catheter, especially in patients with slit ventricles. There is evidence that a significant number of patients do not tolerate flow control valves and, despite a radiologically functioning shunt, have high intracranial pressure from underdrainage through the valve. In patients with limited pressure–volume compensatory reserve, there can be an excessive increase in intracranial pressure during cardiovascular fluctuations, especially at night and be responsible for nighttime or early morning headaches, in patients with flow control devices (60). Self-adjusting diaphragm valves like the Orbis-Sigma (Integra LifeSciences), on bench test, have proved to be inaccurate and unstable at perfusion rates of 20–30 mL/h, which is the most important physiological range, leading to pre-valve pressures rapidly changing between 4 and 28 mm Hg. During long-term perfusion, these may resemble ICP pressure waves (60).

Diaphragm-based anti-siphon devices are prone to obstruction from encapsulation as has been shown in experimental animals and is often encountered in patients who have had recurrent malfunctions (65). Some patients are more prone to develop heavy scarring around the shunt system. Again, there is no evidence that using an open (ASD, Anti Siphon device, Integra LifeSciences) has any advantage over using a closed system that opens when the pressure exceeds the negative hydrostatic pressure (SCD, Siphon Control Device, Medtronics), although theoretically malfunctions in an open system would only result in loss of

anti-siphon function without obstruction to the flow of CSF. In the open system (ASD), the flow through the valve stops only after the intracranial pressure has become negative in the upright position, which is more physiological, than with SCD, in which the flow stops once the pressure reaches zero. In the multicentric shunt study, the incidence of overdrainage was 7.8% in the SCD group and 2.6% in the Standard valve group. The study suggests that diaphragm-based anti-siphon devices may not be any superior to differential pressure valves in reducing overdrainage (10). Considerable controversy also revolves around the most optimum site for placement for the anti-siphon devices (66,67). The classic position is at the level of the skull base: however, the bench test suggests a marked tendency to overdrain if the SCD is below the level of the proximal catheter. These factors may be minor when considered in light of the excessive sensitivity of the SCD to external pressure from scar or when the patient is lying on the device (64).

The gravity actuated device (GAD) is used in conjunction with a differential pressure valve to limit overdrainage (68). It is similar to the horizontal vertical valve used in lumbar shunts but constructed to fit in-line with a ventriculoperitoneal shunt. There is no literature to prove or disprove its utility; however, in individual cases, we have found it effective. Experimental evidence suggests that motion and vibration (35) make the mechanism of these devices ineffective although clinical studies are lacking. The position of the GAD device is critical for optimum functioning. Slight angulation of the device to vertical can cause underdrainage in the horizontal position and overdrainage in the vertical position. Examples of pressure flow characteristics of a standard differential pressure valve, a flow control valve, and a valve containing a GAD are shown in Fig. 15.

Distal Catheter

Distal shunt malfunction is reported to occur in 12% to 34% of shunts (51,69). Three types of distal catheters have been used: the closed ended with side slits, open ended with side slits, and open ended. A higher incidence of distal catheter obstruction has been noted in catheters with side slits whether closed ended or open ended (51,70). Omental ingrowth is responsible for the peritoneal catheter obstruction; possibly the distal slits act as collection points for the debris and provide a channel for trapping the omentum. It is unclear whether using open-ended distal catheters increases the likelihood of small ventricle malfunction. Use of extended length catheters (110–120 cm) is not associated with an increase in the complications and eliminates the need to lengthen the peritoneal catheter for growth of the patient (71). However, care must be taken to identify patients who may have enough length of tubing in the abdomen but may underdrain due to a narrow and taught segment of tubing from subcutaneous tethering as a result of scarring and calcification.

It is difficulty to justify use of atrial over the peritoneal site for distal absorption (72,73). Data on 887 patients suggested that atrial shunts have a higher rate of malfunction although some studies have not shown a significant difference. However, when the same information was stratified by age, shunt type, and time period, there was no significant difference in shunt durability. Cardio-pulmonary complication, such as irreversible pulmonary hypertension, endocarditis, and glomerulonephritis, are some of the more serious complications that may occur with atrial shunts (73). Alternative sites, like pleura, may result in significant negative pressures in the shunt system (74). Poor absorption from the pleura may result in large pleural effusions in small children (74). The gall bladder has also been effectively used in patients in whom peritoneal, atrial, or pleural sites have been exhausted (75,76). Potential complications of these shunts, notably biliary ventriculitis and biliary meningitis, have been reported in the literature (77,78). The ventriculo-femoral shunt may be tried in patients with a difficult access to the atrium from the subclavian or jugular route (79). Trans-diaphragmatic placement of the distal catheter in the sub-hepatic space worked successfully in one reported patient with poor peritoneal access due to scarring (80).

Shunt Material

Ideal shunt material should be completely biocompatible, be easy to handle, flexible, resistant to infection, and nonmetallic but radio-opaque (metals interfere with MRI imaging). From a manufacturing standpoint, it should be easy

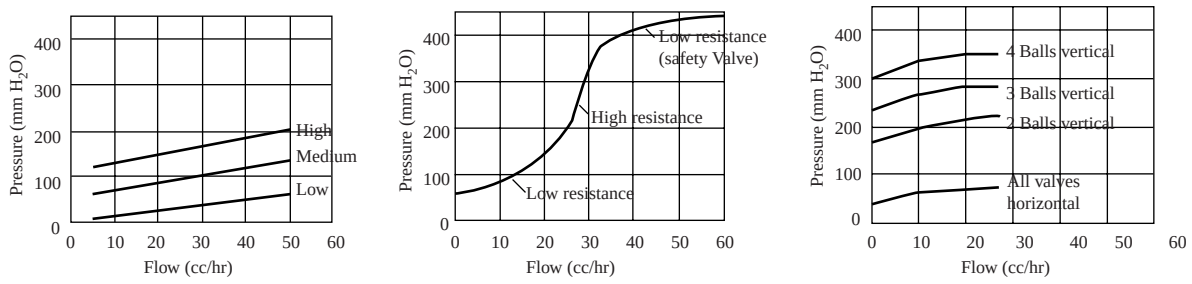


Figure 15. Pressure flow characteristics of (a). Standard differential pressure valve; note that with increasing differential pressure, such as from upright posture, there is an increase in the flow rate. (b) The flow control valve has a sigmoid flow-pressure relationship; in the upright position, the valve works at the high resistance stage and maintains a relatively steady flow rate despite increase in differential pressure. (c) Gravity actuated device, in vertical position acts as a very high-resistance differential pressure valve (depending on the number of balls in the device) and as a low-resistance differential pressure valve in the supine position.

to mold into tubing and making valve components. Silicone polymer is probably the best available material for this purpose.

Some studies have suggested development of silicone allergy in some patients with ventricular shunts (81–84). It is unclear whether it represents a true immunological reaction or a nonspecific foreign body type granulomatous reaction (85). In patients with suspected or documented silicone allergy, use of polyurethane (86) or, more recently CO_2 extracted silicone catheters has been postulated but not proven to offer some advantage in reducing risk of recurrent malfunctions.

Subcutaneous location of the distal catheters makes them susceptible to degradation from a foreign-body reaction mounted by the body (87). Scarring around the catheter, calcification, and stress fractures are long-term consequences of this reaction (88,89). Unless there is some amount of surface degradation, the adhesions to the subcutaneous tissues do not occur (87). Evidence suggests that barium used in the silicone catheters is probably not an important factor in promoting calcification and degradation (90). Use of barium-free catheters, however, makes it difficult to evaluate a shunt system on radiological imaging.

To minimize colonization of shunt catheters and infection, recently antibiotic-coated catheters have become available. The catheters are available coated with rifampin and minocycline (Medtronics, Goleta, CA) and another with rifampin and clindamycin (Bactiseal, Codman, Raynham, MA). The antibiotic is most active against Staph epidermidis, which is the cause of shunt infection in most patients. The antibiotic gradually leaches out of the catheter over a 30–60 day period providing added advantage. Control studies have shown a significant reduction in rate of infection with use of these catheters (91,92). The major drawback is the excessive cost of the antibiotic-coated catheters.

Shunt Malfunction

About 30–40% of the shunts malfunction within the first year of placement (10) and 80% of malfunctions are proximal malfunctions. Although most patients with a malfunctioning shunt will present with the classic features of raised pressure, headache, and vomiting, in 20%, there may be no signs of raised pressure (93). Instead, this group of patients present with a subtle change in behavior, decline in school performance, gait disturbances, and incontinence. Some patients may present with aggravation in the signs and symptoms of Chiari malformation or syringomyelia. Parents are often more sensitive to these subtle changes. In a study comparing the accuracy of referral source in diagnosing shunt malfunction, parents were more likely to be correct about the diagnosis as compared with a hospital or general practitioner (94).

At examination, a tense fontanelle, split sutures, and swelling at the shunt site are very strongly suggestive of a malfunctioning shunt. Shunt pumping has a positive predictive value of only 20% (95). A shunt valve that fails to fill up in 10 minutes is very strongly suggestive of shunt malfunction. Radiological assessment may demonstrate a fracture or dislocation. Presence of double-backing of the distal catheter, wherein the distal catheter tip loops out of the peritoneal through the same spot that it enters it, is diagnostic of distal malfunction (96). The shunt tap gives useful information about the proximal and distal shunt system. The absence of spontaneous flow and poor drip rate indicate proximal malfunction, whereas a high opening pressure is suggestive of distal malfunction (97). The presence of increase in size of the ventricles on CT scan confirms a malfunctioning shunt; however, a large number of patients with long-standing shunt have altered brain compliance and may not dilate the ventricles at the time of presentation. In children, similar symptoms occur in the common illnesses like otitis media; gastroenteritis of viral fevers often confound the diagnosis. Radiological assessment of shunt flow using radionuclide or iodide contrast media injected into the shunt may help (2,98–100). Unfortunately, although some studies have shown an accuracy of 99% with combined pressure and radionuclide evaluation (98,101), others have shown a 25–40% incidence of deceptive patency when evaluated by radionuclide cisternogram (97). This could stem from a partial but inadequately functioning shunt, intermittent malfunction, or presence of isolated ventricle. Similar problems are encountered with an iodide contrast-based shuntogram or shunt injection tests. In the absence of normative data with regard to adequate flow in the shunt, which may vary significantly with the individual, time of the day, and activity (102), use of Doppler-based flow devices, flow systems that work based on differential temperature-gradient or MRI-based flow systems becomes irrelevant for an individual patient. Lumbar infusion tests and shunt infusion tests to assess the outflow resistance through the shunt are cumbersome and require a laboratory-based setup and may not be possible in an ER setting (103–105). Infusion through a reservoir to assess outflow resistance through the shunt suggests a cutoff of less than 12 mm Hg/mL/min as reliable for distinguishing a clinically suspected high probability of malfunction from those with a low probability of shunt malfunction (104). However, this is the group of patients who may not really need the test, and patients who have a questionable malfunction on clinical grounds often have equivocal results on the infusion study.

In childhood hydrocephalus, ICP is the only accurate guide to shunt function other than the symptoms (105). Again the ability to measure ICP through the valve tap becomes unreliable with a partial proximal malfunction. A similar problem may be encountered with in-line telemetric ICP monitors (106,107). In-addition, the telemetric transducers may develop a significant drift over time. In difficult cases, the only way to resolve the issue may be to explore the shunt, to measure ICP through a lumbar puncture if the patient has communicating hydrocephalus, or to place an ICP monitor. Noninvasive monitoring of ICP is going to have a major role in assessment of these patients. For patients who have a very compliant brain, ventricular diltation on the CT scan easily confirms inadequate shunt function.

Despite advances in shunt technology, the incidence of shunt malfunction has not changed over the last 50 years. Nulsen and Becker (3) reported a rate of malfunction of 44%, in 1967, which is similar to that reported in recent studies. To improve on the existing shunt systems, it is

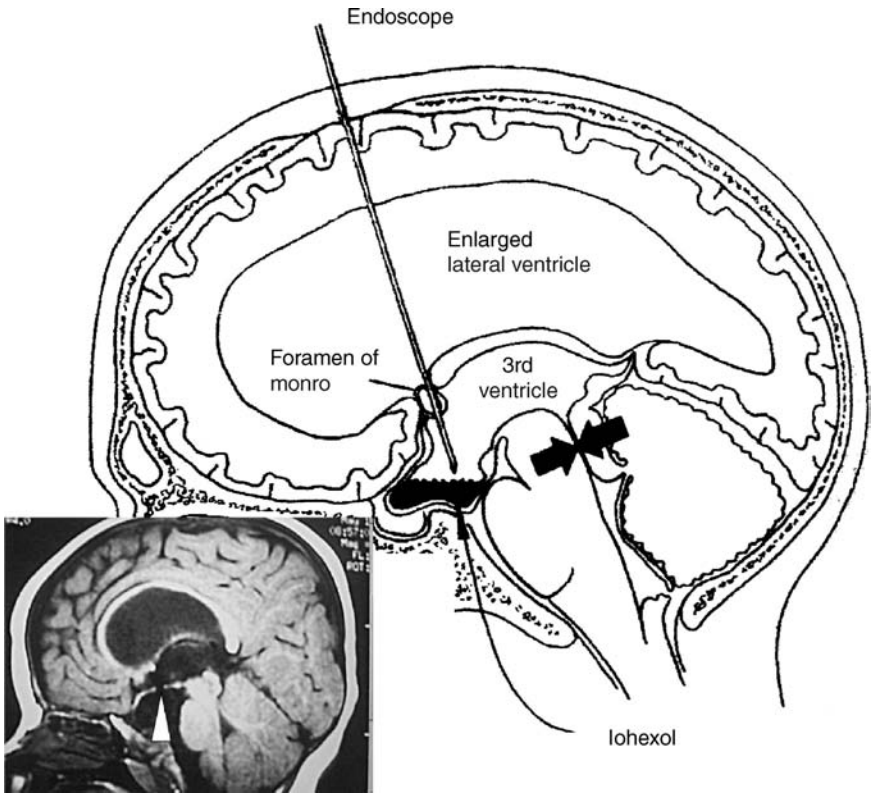


Figure 16. Third ventriculostomy is useful in patients with obstructive hydrocephalus such as observed with aqueductal stenosis (arrows). An endoscope is passed through a small hole in the skull into the frontal horn of the ventricle and navigated into the third ventricle. The floor of the third ventricle is then perforated under vision (arrow, in the inset) so as to bypass the obstruction at the aqueduct. Iohexol, a nonionic iodinated dye, is then instilled into the ventricle to demonstrate good communication between the ventricles and the subarachnoid space.

important to understand the factors that reduce shunt survival. Although the location of the proximal catheter has not been clearly shown to influence shunt survival (50,51), the presence of a small amount of fluid around the proximal catheter is associated with longer shunt survival. In a study that looked at shunt over a 11 year period, statistically significant differences were noted in shunt survival in patients with tumor versus post-hemorrhagic and aqueductal stenosis; shunts in infants and the pediatric age group survive shorter than in adults; shunt after multiple revisions survive shorter, and additional shunts placed for isolated ventricles have shorter survival (70,108). Chronic inflammatory changes of granular ependymitis often seen at the time of endoscopic shunt placement in patients with multiple revisions probably contribute to recurrent malfunction and progressive shortening of the interval between revisions as the number of surgeries increase (108).

The nature of the valve clearly influences the risk of proximal catheter malfunction. It is much lower with flow control valves (10,109). Overdrainage from the differential pressure valves pulls the choroid plexus toward the proximal catheter and may promote malfunction (55). However, the increased rate of valve malfunction in flow control devices balances out this advantage (10).

THIRD VENTRICULOSTOMY

In patients with an obstructive type of hydrocephalus, third ventriculostomy offers an alternative to shunt. The procedure involves making an opening in the relatively thin membrane of the floor of the third ventricle. This is accomplished by passing an endoscope through the lateral ventricle and guiding it through the foramen of Munro to the floor of the third ventricle (Fig. 16). The opening allows CSF to bypass the obstruction at the level of the aqueduct or the fourth ventricle and directly enter the subarachnoid space.

Although third ventriculostomy has been recommended as a procedure of choice for obstructive hydrocephalus; data from some prospective studies have failed to show an improved cure rate (110,111). A retrospective analysis of ventriculographic versus endoscopic third ventriculostomy in 213 cases does show the superiority of the endoscopic procedure over the ventriculographic operation both in terms of reduced risk and improved survival of the procedure (112). Despite the theoretical advantages, evidence suggests that third ventriculostomy may not be effective in controlling raised intracranial pressure in all patients (112,113). Early failures in a radiologically proven case of obstructive hydrocephalus may relate to multifactorial etiology of hydrocephalus; associated absorption defects, obliteration of subarachnoid space from long-standing ventricular diltation, and unidentified infectious cause of aqeductal stenosis may be responsible. Late failures may relate to gliotic scarring over the ventriculostomy, which has been visually confirmed by endoscope in some cases. Does the ventriculostomy close from scarring, or is it a secondary response to lack of flow through the ventriculostomy due to poor absorption, therefore, a lack of gradient between the ventricle and the subarachnoid space, is unclear. In a small prospective study comparing the shunt failure rate with the failure rate of third ventriculostomy,

no statistical difference was found between the two (12). Likewise, no controlled study has compared laser, blunt, or sharp fenestration of the floor or demonstrated usefulness of balloon diltation of the fenestration. The success rate of third ventriculostomy of 49–100%, as reported in the literature, may not be a true representative of the efficacy of third ventriculostomy (12). Evaluation after third ventriculostomy and defining success is difficult in the absence of documented reduction in ICP or improvement on neuropyschological tests. This is so because the ventricles may not reduce in size, and to and from motion through the patent fenestration may still be observed on MRI CSF flow studies even though the patient may be symptomatic. In the absence of clear evidence in literature, third ventriculostomy is often advocated for patients with obstructive hydrocephalus, and if failures occur, shunting is preferred over repeat fenestration.

It is hoped that close collaboration between the industry and medicine will help develop "smart shunts" that would be able to mimic physiological CSF dynamics. These devices will possibly incorporate nanotechnology and would be superior to the presently available devices. It is likely that better understanding of CSF drainage mechanism in the future may help develop alternatives such as drugs that improve drainage of CSF through lymphatic/arachnoidal venous channels or promote proliferation of new lymphatic/arachnoidal venous channels. Until that time, it seems that shunts are the best available alternative for management of communicating hydrocephalus.

BIBLIOGRAPHY

1. Dandy WE, Blackfan KD. Internal hydrocephalus: An experimental, clinical and pathological study. Am J Dis Child 1914;30:406–482.
2. Graham P, Howman-Giles R, Johnston I, Besser M. Evaluation of CSF shunt patency by means of technetium-99m DTPA. J Neurosurg 1982;57:262–266.
3. Nulsen FE, Becker DP. Control of hydrocephalus by valve-regulated shunt. J Neurosurg 1967;26:362–374.
4. Pudenz RH, Russell FE, Hurd AH, Shelden CH. Ventriculo-auriculostomy; a technique for shunting cerebrospinal fluid into the right auricle; preliminary report. J Neurosurg 1957;14:171–179.
5. Nulsen FE, Spitz EB. Treatment of hydrocephalus by direct shunt from ventricle to jugular vain. Surg Forum 1951;94: 399–403.
6. Fernell E, Hagberg B, Hagberg G, von Wendt L. Epidemiology of infantile hydrocephalus in Sweden. I. Birth prevalence and general data. Acta Paediatr Scand 1986;75:975–981.
7. Bondurant CP, Jimenez DF. Epidemiology of cerebrospinal fluid shunting. Pediatr Neurosurg 1995;23:254–258; discussion 259.
8. Foltz EL, Shurtleff DB. Five-year comparative study of hydrocephalus in children with and without operation (113 Cases). J Neurosurg 1963;20:1064–1079.
9. Laurence KM, Coates S. The natural history of hydrocephalus. Detailed analysis of 182 unoperated cases. Arch Dis Child 1962;37:345–362.
10. Drake JM, Kestle JR, Milner R, Cinalli G, Boop F, Piatt J, Jr. Haines S, Schiff SJ, Cochrane DD, Steinbok P, MacNeil N. Randomized trial of cerebrospinal fluid shunt valve design in pediatric hydrocephalus. Neurosurgery 1998;43:294–303; discussion 303–295.
11. Patwardhan RV, Nanda A. Implanted ventricular shunts in the United States: The billion-dollar-a-year cost of hydrocephalus treatment. Neurosurgery 2005;56:139–144; discussion 144–135.
12. Tuli S, Alshail E, Drake J. Third ventriculostomy versus cerebrospinal fluid shunt as a first procedure in pediatric hydrocephalus. Pediatr Neurosurg 1999;30:11–15.
13. Bell WO. Cerebrospinal fluid reabsorption. A critical appraisal. 1990. Pediatr Neurosurg 1995;23:42–53.
14. Edsbagge M, Tisell M, Jacobsson L, Wikkelso C. Spinal CSF absorption in healthy individuals. Am J Physiol Regul Integr Comp Physiol 2004;287:R1450–1455.
15. Koh L, Zakharov A, Johnston MG. Integration of the subarachnoid space and lymphatics: Is it time to embrace a new concept of cerebrospinal fluid absorption? Cerebrospinal Fluid Res 2005;2:6.
16. Johnston M, Zakharov A, Papaiconomou C, Salmasi G, Armstrong D. Evidence of connections between cerebrospinal fluid and nasal lymphatic vessels in humans, non-human primates and other mammalian species. Cerebrospinal Fluid Res 2004;1:2.
17. Greitz D, Greitz T, Hindmarsh T. We need a new understanding of the reabsorption of cerebrospinal fluid—II. Acta Paediatr 1997;86:1148.
18. Marmarou A, Shulman K, LaMorgese J. Compartmental analysis of compliance and outflow resistance of the cerebrospinal fluid system. J Neurosurg 1975;43:523–534.
19. Raksin PB, Alperin N, Sivaramakrishnan A, Surapaneni S, Lichtor T. Noninvasive intracranial compliance and pressure based on dynamic magnetic resonance imaging of blood flow and cerebrospinal fluid flow: Review of principles, implementation, and other noninvasive approaches. Neurosurg Focus 2003;14:e4.
20. McAllister JP2nd, Chovan P. Neonatal hydrocephalus. Mechanisms and consequences. Neurosurg Clin N Am 1998; 9:73–93.
21. Adams RD, Fisher CM, Hakim S, Ojemann RG, Sweet WH. Symptomatic occult hydrocephalus with "normal" cerebrospinal-fluid pressure. A treatable syndrome. N Engl J Med 1965; 273:117–126.
22. Oi S, Shimoda M, Shibata M, Honda Y, Togo K, Shinoda M, Tsugane R, Sato O. Pathophysiology of long-standing overt ventriculomegaly in adults. J Neurosurg 2000;92:933–940.
23. Di Rocco C, Pettorossi VE, Caldarelli M, Mancinelli R, Velardi F. Experimental hydrocephalus following mechanical increment of intraventricular pulse pressure. Experientia 1977;33:1470–1472.
24. Egnor M, Rosiello A, Zheng L. A model of intracranial pulsations. Pediatr Neurosurg 2001;35:284–298.
25. Egnor M, Zheng L, Rosiello A, Gutman F, Davis R. A model of pulsations in communicating hydrocephalus. Pediatr Neurosurg 2002;36:281–303.
26. Linninger AA, Tsakiris C, Zhu DC, Xenos M, Roycewicz P, Danziger Z, Penn R. Pulsatile cerebrospinal fluid dynamics in the human brain. IEEE Trans Biomed Eng 2005;52:557–565.
27. Stephensen H, Tisell M, Wikkelso C. There is no transmantle pressure gradient in communicating or noncommunicating hydrocephalus. Neurosurgery 2002;50:763–771; discussion 771–763.
28. Portnoy HD, Branch C, Castro ME. The relationship of intracranial venous pressure to hydrocephalus. Childs Nerv Syst 1994;10:29–35.
29. Marmarou A. A theoretical model and experimental evaluation of the cerebrospinal fluid system. 1973.
30. Alperin N, Lichtor T, Mazda M, Lee SH. From cerebrospinal fluid pulsation to noninvasive intracranial compliance and pressure measured by MRI flow studies. Curr Med Imaging Rev. In press.

31. Marmarou A, Shulman K, Rosende RM. A nonlinear analysis of the cerebrospinal fluid system and intracranial pressure dynamics. J Neurosurg 1978;48:332–344.
32. Ekstedt J. CSF hydrodynamic studies in man. 1. Method of constant pressure CSF infusion. J Neurol Neurosurg Psych 1977;40:105–119.
33. Lundkvist B, Eklund A, Kristensen B, Fagerlund M, Koskinen LO, Malm J. Cerebrospinal fluid hydrodynamics after placement of a shunt with an antisiphon device: A long-term study. J Neurosurg 2001;94:750–756.
34. Czosnyka M, Batorski L, Laniewski P, Maksymowicz W, Koszewski W, Zaworski W. A computer system for the identification of the cerebrospinal compensatory model. Acta Neurochirurgica 1990;105:112–116.
35. Borgesen SE, Gjerris F. The predictive value of conductance to outflow of CSF in normal pressure hydrocephalus. J Neurol 1982;105:65–86.
36. Marmarou A, Bergsneider M, Klinge P, Relkin N, Black PM. The value of supplemental prognostic tests for the preoperative assessment of idiopathic normal-pressure hydrocephalus. Neurosurgery 2005;57:17–28.
37. Huang TY, Chung HW, Chen MY, Giiang LH, Chin SC, Lee CS, Chen CY, Liu YJ. Supratentorial cerebrospinal fluid production rate in healthy adults: Quantification with two-dimensional cine phase-contrast MR imaging with high temporal and spatial resolution. Radiology 2004;233:603–608.
38. Alperin NJ, Lee SH, Loth F, Raksin PB, Lichtor T. MR-Intracranial pressure (ICP): A method to measure intracranial elastance and pressure noninvasively by means of MR imaging: Baboon and human study. Radiology 2000;217:877–885.
39. Urchuk SN, Plewes DB. MR measurements of pulsatile pressure gradients. J Magn Reson Imaging 1994;4:829–836.
40. Bird R, Stewart W, Lightfoot E. Transport Phenomena. New York: Wiley Sons; 1960.
41. Walters BC, Hoffman HJ, Hendrick EB, Humphreys RP. Cerebrospinal fluid shunt infection. Influences on initial management and subsequent outcome. J Neurosurg 1984;60: 1014–1021.
42. Pople IK, Ettles D. The role of endoscopic choroid plexus coagulation in the management of hydrocephalus. Neurosurgery 1995;36:698–701; discussion 701–692.
43. Weiss MH, Nulsen FE, Kaufman B. Selective radionecrosis of the choroid plexus for control of experimental hydrocephalus. J Neurosurg 1972;36:270–275.
44. Aoki N. Lumboperitoneal shunt: Clinical applications, complications, and comparison with ventriculoperitoneal shunt. Neurosurgery 1990;26:998–1003; discussion 1003–1004.
45. Selman WR, Spetzler RF, Wilson CB, Grollmus JW. Percutaneous lumboperitoneal shunt: Review of 130 cases. Neurosurgery 1980;6:255–257.
46. Chumas PD, Armstrong DC, Drake JM, Kulkarni AV, Hoffman HJ, Humphreys RP, Rutka JT, Hendrick EB. Tonsillar herniation: The rule rather than the exception after lumboperitoneal shunting in the pediatric population. J Neurosurg 1993;78:568–573.
47. Payner TD, Prenger E, Berger TS, Crone KR. Acquired Chiari malformations: Incidence, diagnosis, and management. Neurosurgery 1994;34:429–434; discussion 434.
48. Ausman JI. Shunts: Which one, and why? Surg Neurol 1998;49:8–13.
49. Albright AL, Haines SJ, Taylor FH. Function of parietal and frontal shunts in childhood hydrocephalus. J Neurosurg 1988;69:883–886.
50. Bierbrauer KS, Storrs BB, McLone DG, Tomita T, Dauser R. A prospective, randomized study of shunt function and infections as a function of shunt placement. Pediatr Neurosurg 1990;16:287–291.
51. Sainte-Rose C, Piatt JH, Renier D, Pierre-Kahn A, Hirsch JF, Hoffman HJ, Humphreys RP, Hendrick EB. Mechanical complications in shunts. Pediatr Neurosurg 1991;17:2–9.
52. Pang D, Grabb PA. Accurate placement of coronal ventricular catheter using stereotactic coordinate-guided free-hand passage. Technical note. J Neurosurg 1994;80:750–755.
53. Yamamoto M, Oka K, Nagasaka S, Tomonaga M. Ventriculoscope-guided ventriculoperitoneal shunt and shunt revision. Technical note. Acta Neurochir (Wien) 1994;129:85–88.
54. Kestle JR, Drake JM, Cochrane DD, Milner R, Walker ML, Abbott R, 3rd, Boop FA. Lack of benefit of endoscopic ventriculoperitoneal shunt insertion: A multicenter randomized trial. J Neurosurg 2003;98:284–290.
55. Hakim S. Observations on the physiopathology of the CSF pulse and prevention of ventricular catheter obstruction in valve shunts. Dev Med Child Neurol Suppl 1969;20:42–48.
56. Martinez-Lage JF, Lopez F, Poza M, Hernandez M. Prevention of intraventricular hemorrhage during CSF shunt revisions by means of a flexible coagulating electrode. A preliminary report. Childs Nerv Syst 1998;14:203–206.
57. Steinbok P, Cochrane DD. Removal of adherent ventricular catheter. Pediatr Neurosurg 1992;18:167–168.
58. Whitfield PC, Guazzo EP, Pickard JD. Safe removal of retained ventricular catheters using intraluminal choroid plexus coagulation. Technical note. J Neurosurg 1995;83: 1101–1102.
59. Handler MH. A complication in removing a retained ventricular catheter using electrocautery. Pediatr Neurosurg 1996; 25:276.
60. Czosnyka M, Czosnyka Z, Whitehouse H, Pickard JD. Hydrodynamic properties of hydrocephalus shunts: United Kingdom Shunt Evaluation Laboratory. J Neurol Neurosurg Psych 1997;62:43–50.
61. Trost HA. Is there a reasonable differential indication for different hydrocephalus shunt systems? Childs Nerv Syst 1995;11:189–192.
62. Sood S, Kumar CR, Jamous M, Schuhmann MU, Ham SD, Canady AI. Pathophysiological changes in cerebrovascular distensibility in patients undergoing chronic shunt therapy. J Neurosurg 2004;100:447–453.
63. Pollack IF, Albright AL, Adelson PD. A randomized, controlled study of a programmable shunt valve versus a conventional valve for patients with hydrocephalus. Hakim-Medos Investigator Group. Neurosurgery 1999;45:1399–1408; discussion 1408–1311.
64. Aschoff A, Kremer P, Benesch C, Fruh K, Klank A, Kunze S. Overdrainage and shunt technology. A critical comparison of programmable, hydrostatic and variable-resistance valves and flow-reducing devices. Childs Nerv Syst 1995;11:193–202.
65. Drake JM, da Silva MC, Rutka JT. Functional obstruction of an antisiphon device by raised tissue capsule pressure. Neurosurgery 1993;32:137–139.
66. Fox JL, Portnoy HD, Shulte RR. Cerebrospinal fluid shunts: An experimental evaluation of flow rates and pressure values in the anti-siphon valve. Surg Neurol 1973;1:299–302.
67. Tokoro K, Chiba Y. Optimum position for an anti-siphon device in a cerebrospinal fluid shunt system. Neurosurgery 1991;29:519–525.
68. Chabbra DK, Agarwal GD, Mittal P. "Z" flow hydrocephalus shunts, a new approach to the problem of hysrocephalus. The rationale behind its design and the initial results of pressure monitoring after "Z" flow shunt implantation. Acta Neurochir (Wien) 1993;121:43–47.
69. Sekhar LN, Moossy J, Guthkelch AN. Malfunctioning ventriculoperitoneal shunts. Clinical and pathological features. J Neurosurg 1982;56:411–416.
70. Cozzens JW, Chandler JP. Increased risk of distal ventriculoperitoneal shunt obstruction associated with slit valves or

distal slits in the peritoneal catheter. J Neurosurg 1997;87: 682–686.
71. Couldwell WT, LeMay DR, McComb JG. Experience with use of extended length peritoneal shunt catheters. J Neurosurg 1996;85:425–427.
72. Borgbjerg BM, Gjerris F, Albeck MJ, Hauerberg J, Borgesen SV. A comparison between ventriculo-peritoneal and ventriculo-atrial cerebrospinal fluid shunts in relation to rate of revision and durability. Acta Neurochir (Wien) 1998;140: 459–464; discussion 465.
73. Lam CH, Villemure JG. Comparison between ventriculoatrial and ventriculoperitoneal shunting in the adult population. Br J Neurosurg 1997;11:43–48.
74. Willison CD, Kopitnik TA, Gustafson R, Kaufman HH. Ventriculopleural shunting used as a temporary diversion. Acta Neurochir (Wien) 1992;115:67–68.
75. Ketoff JA, Klein RL, Maukkassa KF. Ventricular cholecystic shunts in children. J Pediatr Surg 1997;32:181–183.
76. Novelli PM, Reigel DH. A closer look at the ventriculogallbladder shunt for the treatment of hydrocephalus. Pediatr Neurosurg 1997;26:197–199.
77. Barami K, Sood S, Ham SD, Canady AI. Postural changes in intracranial pressure in chronically shunted patients. Pediatr Neurosurg 2000;33:64–69.
78. Bernstein RA, Hsueh W. Ventriculocholecystic shunt. A mortality report. Surg Neurol 1985;23:31–37.
79. Philips MF, Schwartz SB, Soutter AD, Sutton LN. Ventriculofemoroatrial shunt: A viable alternative for the treatment of hydrocephalus. Technical note. J Neurosurg 1997;86: 1063–1066.
80. Rengachary SS. Transdiaphragmatic ventriculoperitoneal shunting: Technical case report. Neurosurgery 1997;41: 695–697; discussion 697–698.
81. Goldblum RM, Pelley RP, O'Donell AA, Pyron D, Heggers JP. Antibodies to silicone elastomers and reactions to ventriculoperitoneal shunts. Lancet 1992;340:510–513.
82. Gower DJ, Lewis JC, Kelly DL, Jr. Sterile shunt malfunction. A scanning electron microscopic perspective. J Neurosurg 1984;61:1079–1084.
83. Snow RB, Kossovsky N. Hypersensitivity reaction associated with sterile ventriculoperitoneal shunt malfunction. Surg Neurol 1989;31:209–214.
84. Sugar O, Bailey OT. Subcutaneous reaction to silicone in ventriculoperitoneal shunts. Long-term results. J Neurosurg 1974;41:367–371.
85. Kalousdian S, Karlan MS, Williams MA. Silicone elastomer cerebrospinal fluid shunt systems. Council on Scientific Affairs, American Medical Association. Neurosurgery 1998; 42:887–892.
86. Jimenez DF, Keating R, Goodrich JT. Silicone allergy in ventriculoperitoneal shunts. Childs Nerv Syst 1994;10: 59–63.
87. Del Bigio MR. Biological reactions to cerebrospinal fluid shunt devices: A review of the cellular pathology. Neurosurgery 1998;42:319–325; discussion 325–316.
88. Echizenya K, Satoh M, Murai H, Ueno H, Abe H, Komai T. Mineralization and biodegradation of CSF shunting systems. J Neurosurg 1987;67 :584–591.
89. Elisevich K, Mattar AG, Cheeseman F. Biodegradation of distal shunt catheters. Pediatr Neurosurg 1994;21:71–76.
90. Irving IM, Castilla P, Hall EG, Rickham PP: Tissue reaction to pure and impregnated silastic. J Pediatr Surg 1971;6:724–729.
91. Aryan HE, Meltzer HS, Park MS, Bennett RL, Jandial R, Levy ML. Initial experience with antibiotic-impregnated silicone catheters for shunting of cerebrospinal fluid in children. Childs Nerv Syst 2005;21:56–61.
92. Govender ST, Nathoo N, van Dellen JR: Evaluation of an antibiotic-impregnated shunt system for the treatment of hydrocephalus. J Neurosurg 2003;99:831–839.
93. Fried A, Shapiro K. Subtle deterioration in shunted childhood hydrocephalus. A biomechanical and clinical profile. J Neurosurg 1986;65:211–216.
94. Watkins L, Hayward R, Andar U, Harkness W. The diagnosis of blocked cerebrospinal fluid shunts: A prospective study of referral to a paediatric neurosurgical unit. Childs Nerv Syst 1994;10:87–90.
95. Piatt JH, Jr. Physical examination of patients with cerebrospinal fluid shunts: Is there useful information in pumping the shunt? Pediatrics 1992;89:470–473.
96. Martinez-Lage JF, Poza M, Izura V. Retrograde migration of the abdominal catheter as a complication of ventriculoperitoneal shunts: The fishhook sign. Childs Nerv Syst 1993;9: 425–427.
97. Sood S, Canady AI, Ham SD. Evaluation of shunt malfunction using shunt site reservoir. Pediatr Neurosurg 2000; 32:180–186.
98. Hayden PW, Rudd TG, Shurtleff DB. Combined pressure-radionuclide evaluation of suspected cerebrospinal fluid shunt malfunction: A seven-year clinical experience. Pediatrics 1980;66:679–684.
99. Sweeney LE, Thomas PS. Contrast examination of cerebrospinal fluid shunt malfunction in infancy and childhood. Pediatr Radiol 1987;17:177–183.
100. Vernet O, Farmer JP, Lambert R, Montes JL. Radionuclide shuntogram: Adjunct to manage hydrocephalic patients. J Nucl Med 1996;37:406–410.
101. Savoiardo M, Solero CL, Passerini A, Migliavacca F. Determination of cerebrospinal fluid shunt function with water-soluble contrast medium. J Neurosurg 1978;49:398–407.
102. Kadowaki C, Hara M, Numoto M, Takeuchi K, Saito I. CSF shunt physics: factors influencing inshunt CSF flow. Childs Nerv Syst 1995;11:203–206.
103. Czosnyka M, Whitehouse H, Smielewski P, Simac S, Pickard JD. Testing of cerebrospinal compensatory reserve in shunted and non-shunted patients: A guide to interpretation based on an observational study. J Neurol Neurosurg Psych 1996;60:549–558.
104. Morgan MK, Johnston IH, Spittaler PJ. A ventricular infusion technique for the evaluation of treated and untreated hydrocephalus. Neurosurgery 1991;29:832–836; discussion 836–837.
105. Fouyas IP, Casey AT, Thompson D, Harkness WF, Hayward RD. Use of intracranial pressure monitoring in the management of childhood hydrocephalus and shunt-related problems. Neurosurgery 1996;38:726–731; discussion 731–722.
106. Woodford J, Saunders RL, Sachs E, Jr. Shunt system patency testing by lumbar infusion. J Neurosurg 1976;45:60–65.
107. Cosman ER, Zervas NT, Chapman PH, Cosman BJ, Arnold MA. A telemetric pressure sensor for ventricular shunt systems. Surg Neurol 1979;11:287–294.
108. Miyake H, Ohta T, Kajimoto Y, Matsukawa M. A new ventriculoperitoneal shunt with a telemetric intracranial pressure sensor: Clinical experience in 94 patients with hydrocephalus. Neurosurgery 1997;40:931–935.
109. Lazareff JA, Peacock W, Holly L, Ver Halen J, Wong A, Olmstead C. Multiple shunt failures: An analysis of relevant factors. Childs Nerv Syst 1998;14:271–275.
110. Decq P, Barat JL, Duplessis E, Leguerinel C, Gendrault P, Keravel Y. Shunt failure in adult hydrocephalus: Flow-controlled shunt versus differential pressure shunts–a cooperative study in 289 patients. Surg Neurol 1995;43:333–339.
111. Garton HJ, Kestle JR, Cochrane DD, Steinbok P. A cost-effectiveness analysis of endoscopic third ventriculostomy. Neurosurgery 2002;51:69–77; discussion 77–68.

112. Santamarta D, Diaz Alvarez A, Goncalves JM, Hernandez J. Outcome of endoscopic third ventriculostomy. Results from an unselected series with noncommunicating hydrocephalus. Acta Neurochir (Wien) 2005;147:377–382.
113. Cinalli G, Sainte-Rose C, Chumas P, Zerah M, Brunelle F, Lot G, Pierre-Kahn A, Renier D. Failure of third ventriculostomy in the treatment of aqueductal stenosis in children. J Neurosurg 1999;90:448–454.
114. Hirsch JF, Hirsch E, Sainte Rose C, Renier D, Pierre-Khan A. Stenosis of the aqueduct of Sylvius. Etiology and treatment. J Neurosurg Sci 1986;30:29–39.

See also INTRAUTERINE SURGICAL TECHNIQUES; MICRODIALYSIS SAMPLING; MONITORING, INTRACRANIAL PRESSURE.

HYPERALIMENTATION.

See NUTRITION, PARENTERAL.

HYPERBARIC MEDICINE

BARBARA L. PERSONS
BETH COLLINS
WILLIAM C. LINEAWEAVER
University of Mississippi
Medical Center
Jackson, Mississippi

INTRODUCTION

The goal of hyperbaric oxygen (HBO) therapy is to deliver high concentrations of oxygen under pressure to increase the amount of dissolved oxygen in the blood. The physiologic repercussions of this increased plasma oxygen have widespread effects that translate into a variety of clinical applications. Initially, the use of hyperbaric medicine surrounded acute decompression illness and gas embolism. Later, the increased oxygen under pressure was shown to have use in a variety of clinical situations. The delivered pressure can be two to six times ambient atmospheric pressure (ATM) or atmospheres absolute (ATA) depending on the indication. The current Undersea and Hyperbaric Medical Society approved uses of HBO are shown in Table 1, and many of these indications will be discussed individually.

Table 1. Approved Uses for Hyperbaric Oxygen Therapy

acute decompression illness
gas embolism
carbon monoxide poisioning
clostridial gas gangrene
necrotizing soft tissue infections
compromised skin grafts and skin flaps
crush injury
compartment syndrome
acute traumatic ischemias
radiation tissue damage
refractory osteomyelitis
selected problem wounds
acute exceptional blood loss anemia
acute thermal burns
intracranial abscess

HBO in the treatment of these conditions is supported by controlled medical trials published in peer-reviewed journals and, as such, is evidence-based. Numerous other experimental uses exist for HBO, such as for stroke and for cardiac ischemia, but these uses have not yet been sufficiently proven to be supported by the Undersea and Hyperbaric Medicine Society or by the American College of Hyperbaric Medicine. The goal of this chapter is to outline the physical principles underlying the use of HBO therapy, to discuss its medical indications for HBO, and to familiarize the reader with the mechanical, safety, and regulatory issues involved in operating a hyperbaric medicine program.

HISTORICAL BACKGROUND

British physician and clergyman Henshaw was the first to use alteration in atmospheric pressure to treat medical conditions when he used his domicilium chamber in 1862. Hyperbaric medicine also surrounded diving and diving medicine. Triger, in 1841, gave the first human description of decompression sickness (1). In 1934, U.S. Naval Submarine Officer Dr. Albert Behnke was the first to use oxygen recompression to treat decompression sickness in naval divers (2). Later, in 1943, Gagnon and Cousteau invented SCUBA (self-contained underwater breathing apparatus). Dr. Boerma, a Dutch thoracic surgeon, removed the blood cells from pigs in 1955 and found they could survive with the oxygen dissolved in plasma by use of HBO. An upsurge in hyperbaric surgery followed in 1956, when Boerma performed cardiovascular surgery in a hyperbaric chamber, which along with hypothermia, allowed for periods of circulatory arrest of 7–8 min. The large chamber developed at Wilhelmina Gasthuis in Amsterdam in 1959, headed by Boerma, allowed a wide variety of research to be carried out on the uses of HBO therapy on many diseases (3). By 1966, it was indicated for the treatment of protection during induced circulatory arrest, homotransplantation, clostridial infection, acute arterial insufficiency, chronic arterial insufficiency, and hypovolemic shock. Shortly thereafter, the advent of cardiopulmonary bypass obviated hyperbaric chambers for cardiac protecton. In 1967, the Undersea Medical Society was founded by the U.S. Navy diving and submarine medical officers. This organization originally focused on undersea and diving medicine but, later, came to include clinical hyperbaric medicine. In 1986, the name was changed to the Undersea and Hyperbaric Medical Society or UHMS with more than 2500 physician and scientist members in 50 countries. More recently, the American College of Hyperbaric Medicine has come to offer board certification to U.S. physicians in the specialty of hyperbaric medicine.

PHYSICS

To understand HBO therapy, one must understand a few basic laws of physics, namely, Boyle's law, Charles law, Dalton's law, and Henry's law. Boyles law explains how gas volume shrinks with increasing pressure. Charles law explains that the volume of a gas decreases with decreasing temperature. Dalton's law explains that each gas in a

mixture exerts its own partial pressure independently of the others. Henry's law explains that the number of gas molecules that will dissolve in a liquid depends on the partial pressure of the gas as well as on the mass of the liquid. These laws themselves will be explained in the first part of this section and the application of the laws to each area of hyperbaric medicine will be explained in the respective section.

Boyle's Law

Boyles law states if the temperature remains constant, the volume of a gas is inversely proportional to the pressure.

Boyle's law is stated as $PV = K$ or $P = K/V$, where P is pressure, V is volume, and K is a constant.

Thus, as in Table 2, the volume of a bubble at 1 Atmosphere or sea level shrinks to one-half of its original volume at 33 feet below sea level (10 m or 2 atmospheres). It shrinks to one-third of its original size at 66 feet below sea level (20 m or 3 atmospheres). Conversely, if a diver were to hold his breath at depth and then ascend, the air in his lungs will expand to three times the volume it occupied at 66 feet below sea level.

Charle's Law

Charles' law states that if the pressure remains constant, the volume of a fixed mass of gas is directly proportional to absolute temperature. The volume increases as the temperature increases. For example, a balloon has a volume of 1 L at 20 °C and its volume would expand to 1.1 L at 50 °C.

Charles' law is stated as $V/T = K$ or $V1/T1 = V2/T2$, where V is volume of gas 1 or 2 , T is temperature in Kelvin of gas 1 or 2, and K is a constant.

Dalton's Law

Dalton's law states that each gas in a mixture exerts its partial pressure independently, which is important in understanding human physiogy and HBO. For example, if a patient is given a high concentration of oxygen and a lower concentration of nitrogen, these gasses will diffuse across membranes and act in the body independently of each other.

Dalton's law is stated as $P(t) = P1+P2+P3$, where $P(t)$ is total pressure and $P1$, $P2$, and $P3$ are the individual gas pressures. Gases try to equalize their concentrations across a membrane, which explains how breathing a higher oxygen, lower nitrogen mixture can help nitrogen leave the blood through the lungs as nitrogen gas.

Henry's Law

Henry's law states that at a given temperature, the amount of gas dissolved in a solute is directly proportional to the pressure of the gas above the substance.

Table 2. Pressure vs. Volume at Depth

Depth (feet sea water)	Pressure (ATA) atmospheres	Gas Volume (%)
0(sea level)	1	100
33	2	50
66	3	33
165	6	17

Henry's law illustrates that when a liquid is exposed to a gas, some of the gas molecules dissolve into it. The number of moles that will dissolve in the liquid depends on the mass of the liquid, the partial pressure of the gas, its solubility in the liquid, the surface area of contact, and the temperature (as that changes with the partial pressure). Thus, more gas, oxygen, or nitrogen will dissolve in tissue fluid at a higher pressure because the partial pressure of each gas increases at a higher pressure.

Henry's law is stated as $p = Kc$, where p is the partial pressure of the gas 1, c is its molar concentration, and K is the Henry's law constant, which is temperature-dependent.

An example of Henry's law is dissolved carbon dioxide in soda, which bubbles out of solution as the pressure decreases. Another example is exemplified by water when it is heated. Long before it boils, bubbles of air form on the side of the pan, which is an example of the gas coming out of solution as the temperature is raised.

For treatment of decompression illness (DCI) and gas embolism (GE), increased pressure alone as well as the increased oxygen pressure facilitate treatment. Both of these conditions are a result of gas bubbles in the tissues or gas bubbles in the blood causing blockage of vessles or ischemia of tissues. Therefore, shrinking the bubbles with increased pressure allows them to be removed or minimized by the body, which is the principle of Boyle's law, that the volume of a gas varies inversely with pressure. If one has a bubble occluding an important vessel or lodged in a joint, it will shrink as the pressure increases as in Table 2. The blood can accommodate an increased amount of dissolved gas with increased atmospheric pressure as explained by Henry's law. Henry's law states that the amount of gas that will dissolve in a liquid is proportional to the partial pressure of the gas in contact with that liquid as well as to the atmospheric pressure. Atmospheric pressure at sea level is 760 mmHg, and the normal atmosphere consists of 21% oxygen and 79% nitrogen. (see Fig. 1) (4). The pressure of a gas dissolved in plasma relates to its solubility in a liquid as well as to its partial pressure. A gas bubble caught in the tissues or in the systemic circulation either because of an air embolus or because of decompression sickness is significantly decreased under hyperbaric conditions, as illustrated by Table 3. By Boyles law, the pressure alone shrinks the bubble to a fraction of its' original size. Dalton's law states that total pressure of a mixture of gases is equal to the sum of the pressures of each gas. Each gas is acting as if it alone were present, which explains how the oxygen under pressure creates a situation whereby the higher dissolved oxygen surrounds the bubble and causes the diffusion of nitrogen out of the bubble, called nitrogen washout. Simply put, each gas attempts to have equal concentration of particles, in this case, nitrogen and oxygen on each side of the gas bubble. Nitrogen, therefore, diffuses out of the bubble and shrinks in size. Hyperbaric oxygen enables treatment of the two conditions where air or nitrogen bubbles become lodged in the tissues.

PHYSIOLOGY

Hyperbaric oxygen therapy oxygen enters the systemic circulation through the alveoli in the lungs, and the

Table 3. Oxygen Levels During Hyperbaric Oxygen Treatment Breathing Air (4)

ATA Chamber	Pressure Chamber	PO_2 (mmHg) Chamber	PAO_2 (mmHg) Lung	O_2 ml/dl vol % Plasma
1	760	160	100	0.31
2	1520	319	269	0.83
2.36	1794	377	322	1.00
2.82	2143	450	400	1.24
3	2280	479	429	1.33
4	3040	638	588	1.82
5	3800	798	748	2.32
6	4560	958	908	2.81
Breathing 100% Oxygen				
1	760	760	673	2.08
2	1520	1520	1433	4.44
2.36	1794	1794	1707	5.29
2.82	2143	2143	2056	5.80
3	2280	2280	2193	6.80
4	3040	100% oxygen is not used above 3ATA to minimize the risk of oxygen toxicity.		
5	3800			
6	4560			

diffusion is mediated by the pressure differential between the alveolar oxygen content and the oxygen content of venous blood. Alveolar oxygen content is 100 mmHg and venous oxygen content is 40 mm Hg (5). Normally, 97% of oxygen iscarried in the arterial blood bound to hemoglobin molecules and only 3% of the oxygen is dissolved in plasma, illustrated by the formula for oxygen content in arterial blood. $CaO_2 = (1.34 \times Hb \times SaO_2) + (0.003 \times PaO_2)$, where CaO_2 is oxygen content, Hb is Hemoglobin in grams, and SaO_2 is arterial O_2 saturation expressed as a fraction not a percentage (0.95, not 95%). 1.34 is the realistic binding capacity of hemoglobin, although 1.39 is the actual binding capacity. 0.003 times the PaO_2 is the amount of oxygen soluble in plasma at normal atmospheric pressure. With hyperbaric oxygen, the amount of oxygen dissolved in arterial blood is dramatically increased. Conversely, the amount carried by hemoglobin remains about the same as that achieved by inspiring oxygen at 1 atmosphere absolute (ATA) (4). As measured in ml O_2 per deciliter of whole blood, the oxygen content increases significantly under hyperbaric conditions as shown in Table 3. The oxygen content of blood increases from 0.31 at 1 atmosphere to 6.80ml/dl vol% at 3 atmospheres. Note that 100% oxygen is not used at pressures greater than 3 ATA to minimize the risk of oxygen toxicity. The alveolar type 1 cells in the lungs and the neurons in the brain are sensitive to excessive concentrations of oxygen. Again, to prevent oxygen toxicity, which can lead to alveolar damage or seizure, 100% oxygen is not given at pressures greater than 3 ATA. Even in these pressure ranges, air breaks are given to patients and they breathe air as opposed to concentrated oxygen under pressure for 10 minutes during many of the protocols, which in theory, gives the xanthine oxidese system a chance to deal with the current load of free radicals in the lungs and tissues and the Gaba amino buteric acid (GABA) depletion in the brain a chance to normalize. As mentioned, the blood's ability to carry more dissolved oxygen molecules under higher atmospheric pressure is a function of Henry's law. It states that the amount of gas that will dissolve in a liquid is proportional to the partial pressure of the gas in contact with that liquid as well as to the atmospheric pressure. Thus, more gas, oxygen, or nitrogen will dissolve in tissue fluid at higher atmospheric pressure. %X is the percentage of gas X dissolved in the liquid, $P(t)$ is the atmospheric pressure, and $P(X)$ is the partial pressure of gas X.

TRANSCUTANEOUS OXYMETRY ($TcPO_2$ OR TCOM)

In order for the patient to benefit from HBO therapy, they must have adequate perfusion of blood to the affected area. This perfusion can be assessed prior to hyperbaric oxygen therapy by checking transcutaneous oxygen tension, $TcPO_2$. Transcutaneous oxygen tension values of less than 40, which increase to more than 100 mmHg while breathing 100% oxygen or to more than 200 during HBO therapy, will likely benefit from hyperbaric oxygen therapy (6). The detailed mechanisms by which the elevated oxygen tension is felt to improve wound healing and the body's ability to combat bacterial pathogens will be discussed under each indication for HBO therapy. Briefly, Hunt and Pa: (7) showed, in 1976, that increased oxygen tension stimulated collagen synthesis and fiberblast proliferation. Then, studies revealed improved ability of leukocytes to clear infected wounds of bacteria with hyperbaric oxygen (8,9). Then, in 1989, Zamboni et al. (10) showed that HBO therapy in the reduction of tissue flap necrosis was a systemic phenomenon that involved inhibition of neutrophil adherence and prevention of arteriolar vasoconstriction thought to be via a nitric oxide-mediated mechanism. In addition, it increases platelet-derived growth factor Beta (PDGF B), vascular endothelial growth factor (VEGF), Epidermal growth factor (EGF), and other factors.

APPROVED INDICATIONS

The following indications are approved uses of hyperbaric oxygen therapy as defined by the Hyperbaric Oxygen

Therapy Committee of the Undersea and Hyperbaric Medical Society (Table 1). Most of these indications are also covered by Medicare and many are covered by major insurance companies. The indications include air or gas embolism, decompression illness, carbon monoxide poisioning, Clostridial myositis and myonecrosis (gas gangrene), crush injury, compartment syndrome and other acute traumatic ischemias, decompression sickness, problem wounds, exceptional blood loss anemia, intracranial abscess, necrotizing soft tissue infections, refractory osteomyelitis, delayed radiation injury, compromised skin grafts and flaps, and thermal burns. Many of these indications will be specifically discussed in the following sections.

HYPERBARIC CHAMBER BASICS

Hyperbaric oxygen therapy is a feature offered in many hospitals, medical centers, and in specialty situations such as diver rescue stations and oil rigs. Over 500 hyperbaric chambers exist in the United States alone. When a patient is refered for one of the listed emergent or nonemergent indications to undergo hyperbaric oxygen therapy, a complete workup of the patient should be performed if possible prior to hyperbaric oxygen therapy. Emergency situations may necessitate an abbreviated exam. The patient should wear only cotton medical-center-provided clothing to prevent static electricity, which could cause a spark and a fire. All foreign appliances should be removed, including hearing aids, lighters, and jewelry. Internal appliances such as pacemakers are usually safe under hyperbaric conditions. The patient will then be premedicated with a benzodiazapine such as valium if needed for anxiety. Hyperbaric oxygen therapy can be administered through monoplace or multiplace chambers. Monoplace chambers accommodate a single patient within a pressurized environment of 100% oxygen (Fig. 1). These chambers are often constructed as acrylic cylinders with steel ends and can withstand pressures of up to 3 ATA. Multiplace chambers are usually constructed of steel and can withstand pressures up to 6 ATA (Fig. 2). They can accommodate two or more people and often have the capacity to treat ventilated or critically ill patients. Some are even large enough to accommodate operating teams, and the 100% oxygen is delivered to the patient via face mask, hood, or endotrachial tube. Depending on the indication, patients will require from 1 to 60 treatments. Most commonly, treatments are delivered to the patient for 60–90 min at 2.8–3.0 ATA five days a week for the protocol duration.

CONTRAINDICATIONS

Six absolute contraindications exist to hyperbaric oxygen therapy.

1. Untreated pneumothorax – An untreated pneumothorax can be converted to a tension pneumothorax with administration of HBO.
2. History of spontaneous pneumothorax.

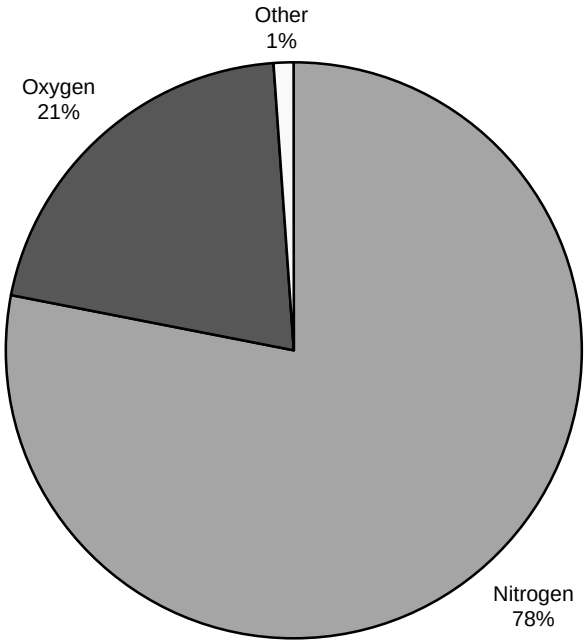


Figure 1. Relative composition of air.

3. Bleomycin – History of the chemotherapy agent Bleomycin, which can cause pneumonitis, especially if the patient is exposed to HBO.
4. Doxyrubicin.
5. Disulfiram – (antibuse) Blocks production of superoxide dismutase, which protects the patient from oxygen toxicity.
6. Cisplatin/Carboplatin – An anticancer agent that interferes with DNA synthesis.
7. Mefanide (Sulfamyelon) – It is a topical ointment for burns and wounds that is a carbonic anhydrase inhibitor and increases the risk of seizure during HBO therapy.

The relative considerations and contraindications are many. In these patients, the hyperbaric physician should consider each patient individually, including their history of thoracic surgery, seizure disorder, obstructive lung disease, congestive heart failure, pulmonary lesions on X-ray or CT scan, seizure disorder, upper respiratory infections

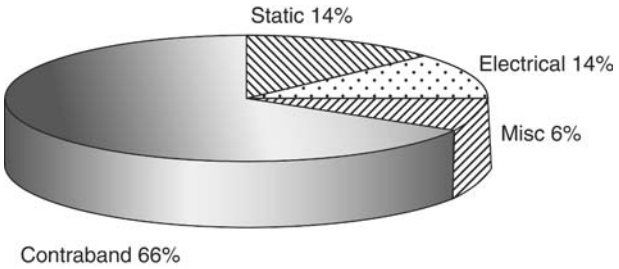


Figure 2. Fire risk.

Table 4. Arterial and Venous Air Embolism Etiology

Etiology of Air Embolism	
Arterial Air Embolism	*Venous Air Embolism*
Pulmonary Overpressure	Central Venous Catheters
Open Lung Biobsy	Infusion Pumps
Arterial Catheter	Neurosurgery
Angiography	Laparoscopy
Surgery	Liver Transplantation
Penetrating Chest Trauma	Neurosurgery
Pneumothorax	Pelvic Surgery
	Trendelenburg Position
Cardiopulmonary Bypass	Necrotizing Enterocolitis
Dialysis	Lumbar Spine Spine Surgery
Autotransfusion	Air Contrast Salpingogram
Neonatal Respiratory	Umbilical Venous Catheters
Distress Syndrome	
Paradoxical, Patent	
Foramen Ovale	

(due to risk of barotraumas to the ears), acute viral infections, uncontrolled high fever (as it increases CNS sensitivity to oxygen), reconstructive ear surgery, congenital spherocytosis, history of optic neuritis, recent retinal repair, claustrophobia, acidosis, nicotine, alcohol, and many others.

AIR EMBOLISM

Air embolism is a medical emergency. In the diving community, it is the main cause of death following diving accidents. Early diagnosis followed by definitive treatment are critical in determining the eventual outcome. Treatment is based on compression of the air bubbles by Boyle's law as well as oxygenation of ischemic tissues and treatment of ischemia reperfusion injury with hyperbaric oxygen. Air embolism can occur in the hospital setting by introduction of air into the systemic circulation by central venous and arterial catheters and other invasive procedures. Interestingly, the first report of a death from air embolism was from France in 1821. The patient was undergoing surgery on his clavicle when the surgeon noted a hissing sound in the area of operation and the patient yelled "my blood is falling into my heart- I'm dead" (11). The patient likely died of a venous air embolism obstructing the systemic circulation. Air entry in the systemic circulation occurs following violation of the systemic circulation by any number of mechanisms (Table 4), which can be either by introduction of air into the arterial circulation, as in a lung biopsy, chest trauma, and pulmonary overpressure (diving), or into venous circulation, as in air introduction via central venous catheters, liver transplantation, and neurosurgery. Venous air emboli are more common, whereas arterial emboli are tend to be more serious. Physiologically, the air bubble forms or is introduced into the circulation. The lung usually serves as an excellent filter for air emboli, and can protect the embolism from traveling to the brain. This protective filter may be bypassed by a patent foramen ovale. Approximately 30% of patients have a foramen ovale that is patent by probe. The lung as a filter may be overwhelmed by large quantities of air. The bubble can then lodge in the smaller arteries of the brain causing obstruction. An air embolism is immediately identified as a foreign body, and platelets are activated, which leads to an inflammatory cascade. Hypoxia then develops distal to the obstruction with associated swelling. The embolism is eventually absorbed by the body, but the fibrin deposition at the embolism site may prevent return of blood flow. In order to diagnose an air embolism, it needs to have a high index of suspicion. If air embolism is suspected, the patient should receive a number of immediate measures, including ACLS or ATLS protocols. The patient should be placed on the left side and one should consider draining air from the right atrium with a central venous catheter. Air in the heart can cause a machinery murmur. 100% oxygen should be administered via a face mask or endotrachial tube, and the patient should be hydrated to preserve intravascular volume. Per Dalton's law, administration of high oxygen will cause nitrogen to diffuse out of the air bubble and shrink in size. Dalton's law states that total pressure of a mixture of gases is equal to the sum of the pressures of each gas. Each gas acts as if it alone were present. Dalton's law is stated as $P(t) = P1 + P2 + P3$. The inspired 100% oxygen will also maximize oxygenation of the tissues as much as possible under normal atmospheric pressure. Next, the patient should be emergently treated with hyperbaric oxygen. If a chamber is available that can provide compression to 6 ATA with air or a mixture of 50% nitrogen and 50% air, treatment should immediately be performed following the U.S. Navy protocol (Fig. 3).

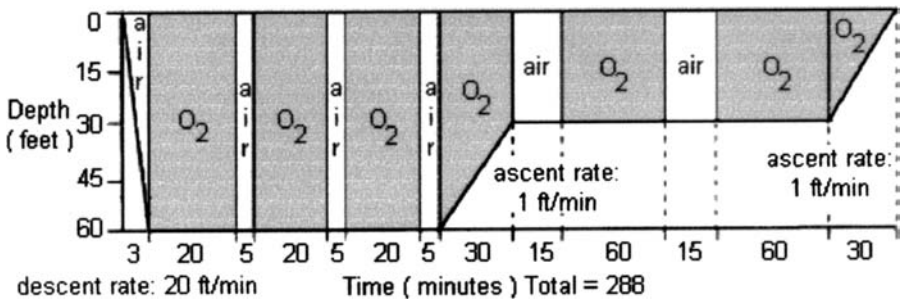


Figure 3. U.S. Navy decompression treatment.

DECOMPRESSION ILLNESS

By definition, decompression illness (DCI), also called bends or caisson disease, occurs when gas bubbles exit in the blood or body tissues. At depth, more gas can dissolve in the tissues than in the blood. When the diver ascends too quickly, the gas comes out of solution and forms bubbles in the tissues and in the blood, much like popping a soda can and releasing its pressure causes the carbon dioxide bubbles to come out of the solution. In 1994, Diver's Alert Network (DAN) recorded 1164 diving-related injuries and 97 diving-related deaths, many related to DCI (12). The severity of DCI depends on the volume and location of gas bubbles. The range of symptoms is from vague constitutional complaints or limb pain to cardiopulmonary arrest and coma, the pathophysiology of which is explained by Henry's law. As previously explained, Henry's law is $p = Kc$, where p is the partial pressure of the gas 1, c is its molar concentration, and K is the Henry's law constant, which is temperature-dependent. Thus, simply stated, more inert gas can be dissolved in a liquid at higher atmospheric pressure and, conversely, less under lower atmospheric pressure, which occurs on decompression when gas must be removed from tissues, and rapid decompression leads to bubble formation. Using the Henry's law equation, one can calculate the estimated amount of nitrogen a diver must clear from the bloodstream (about 5 L) in rising from 100 ft to the surface. The amout would be approximately 750 ml nitrogen assuming room temperature, which is a significant volume of nitrogen that must be eliminated from the divers bloodstream. The onset of symptoms is usually rapid and 75% of patients experience symptoms within 1 h of decompression and 90% within 12 h of decompression (13). A small number of patients may present even later, particularly if they have flown in commercial aircraft after diving and not followed the recommendation of the major diving organizations not to fly within 24 hours of one's last dive. Interestingly, up to 10% of the inert gas that is absorbed in the tissues is released as bubbles after the diver's decompression (14). Patients experience symptoms depending on the location and concentration of the bubbles (see Table 5). Bubbles forming in or near joints cause the joint pain of a classical "bend." These musculoskeletal effects are called type 1 DCS. When these effects occur in the spinal cord or brain, numbness, paralysis, and disorders of higher cerebral function may result. If large numbers of bubbles enter the venous bloodstream, congestive symptoms in the lung and circulatory shock can then occur. These pulmonary and neurologic effects are termed type 2 DCS. Treatment should involve immediate administration of 100% oxygen, which facilitates nitrogen washout by the previously explained principles of Dalton's law. Rehydration as well as advanced cardiac or trauma life-support protocols should be followed by transfer to a hyperbaric facility emergently. The patient should be treated with hyperbaric oxygen following U.S. Navy guidelines (Fig. 3), even if the inspired oxygen and rehydration alone have improved the patient's signs and symptoms because tiny bubbles may be left that can cause tissue necrosis. Hyperbaric oxygen shrinks the size of the mostly nitrogen-filled bubbles by the principles of Boyle's law, and the increased pressure also increases the partial pressure of the gas by Dalton's law, hastening complete elimination of the bubble (Table 2). If the U.S. Navy (Fig. 3) recompression regimen fails to lead to symptom resolution, the Diver's Alert Network or a medical expert on DCI should be contacted, and one of a number of recompression tables may be followed.

Table 5. Signs and Symptoms of Decompression Illness

Symptoms of Decompression Illness	Signs of Decompression Illness
Unusual fatigue	Blotchy skin rash
Skin itching	Paralysis, muscle weakness
Pain in joints/ muscles of arms, legs or torso	Difficulty urinating
Dizziness	Confusion, personality changes, bizarre behavior
Vertigo	Amnesia,
Ringing in the ears, (Tinnitus)	Staggering
Numbness, tingling and paralysis	Coughing up bloody, frothy sputum
Shortness of breath (Dyspnea)	Collapse or unconsciousness
	Tremors

PROBLEM WOUNDS

The management of problem wounds should always include infection control, debridement, aggressive wound care, and correction of perfusion and oxygenation deficiencies. When an oxygenation deficiency of the wound is found, in the face of nonreconstructable vascular disease, hyperbaric oxygen should be considered as an adjunctive therapy. An increase in tissue oxygen tension by HBO therapy enhances wound healing by increasing neutrophil bactericidal capacity, inhibiting toxin formation in and even killing some anaerobes, encouraging fibroblast activity, and promoting angiogenesis (15). In normal physiology, the oxygen gradient across a wound is essential to stimulate these components of healing. Oxygen consumption is relatively low in wounds, and microvasculature damage and peripheral vasoconstriction increase diffusion distances. Partial pressure via Dalton's law is the driving force of diffusion. Hyperbaric oxygen creates a steep tissue oxygenation gradient, providing a stronger stimulus than lactate or moderate hypoxia, to initiate and facilitate wound healing (16,17). These stimulated factors are thought to include platelet-derived growth factor B (PDGF-B), Vascular endothelial growth factor (VEGF), Epidermal growth factor, and others. Several clinical studies support the use of hyperbaric oxygen to promote wound healing. Perhaps the studies involving diabetic lower extremity wounds have been most informative. Several studies have shown an increased number of healed wounds, decreased wound size, and decreased rates of amputation among patients receiving hyperbaric

oxygen therapy as an adjunctive treatment (18,19). Baroni et al. reported in a controlled study that a significant number of subjects receiving HBO went on to heal their wounds and fewer required amputation when compared with subjects not receiving HBO (20). In another study involving 151 diabetic patients with wounds of the lower extremity, Oriani et al. showed that 130 of these patients completely healed their wounds with adjunctive HBO (21). When compared with conventionally treated wounds, HBO-treated patients had an accelerated rate of healing, reduced rate of amputation, and an increased rate of completely healed wounds on a long-term basis (21). Transcutaneous Oxymetry ($TcPO_2$) is currently the best tool available to evaluate tissue hypoxia and to select patients appropriate for HBO therapy. It can also be used to monitor progress during hyperbaric oxygen therapy. $TcPO_2$ measurements should be taken with the patient breathing room air. A value of greater than 50 mmHg around the wound site indicates that the wound has adequate oxygenation and hyperbaric oxygen is not likely to improve healing. Values below 40 at the wound site should be considered for HBO therapy. Patients with marginal $TcPO_2$ should be further tested while breathing 100% oxygen. $TcPO_2$ values of greater than 100 while breathing 100% oxygen is an indicator that they are likely to respond to HBO therapy. If this challenge $TcPO_2$ is less than 100, they still may benefit if the tested $TcPO_2$ at the wound site is greater than 200 mmHg while they are breathing 100% oxygen at 2.0 ATA in the hyperbaric chamber (22). A $TcPO_2$ value of less than 30 around a wound that does not exhibit this response, which indicates vascular compromise and the patient should be considered for revascularization if possible. Of note, 96% of limbs with $TcPO_2$ values below 30 mmHg had abnormal arteriograms. It is also important to follow $TcPO_2$ values weekly, and diabetic patients may have normal or falsely elevated noninvasive Doppler studies and a low $TcPO_2$, implying satisfactory perfusion and inadequate oxygenation of the wound and, as such, may pose a diagnostic delimma. The diabetic patient with normal noninvasive Doppler and low $TcPO_2$ will respond best to HBO. HBO therapy should be reserved for those diabetic wounds not responding to traditional management of debridement, antibiotics, and general wound care, including vascular reconstruction. The use of HBO therapy is necessary in only 15–20% of these patients. HBO therapy increases wound oxygen tension, enhancing host antibacterial mechanisms and promoting wound healing and is reserved for wounds in which the primary etiologies are tissue hypoxia or infection (13). Treatments are delivered at 2.0–2.4 atmospheres for 90–120 min once or twice daily. When serious infections are present, patients are typically hospitalized and given IV antibiotics and hyperbaric treatments twice daily five days a week. The $TcPO_2$ values should be checked weekly because hyperbaric oxygen facilitates angiogenesis by a nitric oxide and vascular endothelial growth factor Beta (VEGF-B). When the room air $TcPO_2$ is greater than 40 mmHg, the hyperbaric oxygen therapy can safely be discontinued. HBO is an adjuvant treatment; therefore, diabetic control, debridement, and aggressive wound treatment are given first priority. When the wound bed has adequate granulation tissue, application of grafts can shorten morbidity, hospital stay, and health-care costs. The underlying problem in failure of a wound to heal is usually hypoxia and infection. Hyperbaric oxygen treatments in selected patients can facilitate healing by increasing tissue oxygen tension, thus providing the wound with a more favorable environment for healing. Therefore, hyperbaric oxygen therapy can be an important component to any comprehensive wound care program.

COMPROMISED FLAPS AND GRAFTS

Skin grafts and flaps with adequate blood supply do not require HBO. Hyperbaric oxygen therapy is extremely useful in situations where the skin grafts or flaps suffer from compromised microcirculation or hypoxia.

Flaps

The benefits of HBO on flaps develop from a systemic elevation in oxygen tension (23–25). In addition, HBO therapy prevents neutrophil adherence and subsequent vasoconstriction following ischemia. Too often, a compromised flap is allowed to progress over the days following surgery until visible signs of necrosis obviate the use of HBO, because delayed treatment with HBO cannot revive dead tissue. The resulting disappointment, as well as the associated patient dissatisfaction, can be avoided by rapid diagnosis of the flap problem and early involvement of the hyperbaric physician. The keys to successful treatment of compromised flaps with HBO are accurate diagnosis of the specific flap problem and appropriate and expedient initiation of hyperbaric oxygen treatment. Awareness of the different etiologies of flap compromise is necessary to plan for effective HBO treatment. A random flap with distal necrosis is completely different from a free flap with total venous occlusion. Proper classification of flaps, different etiologies of flap compromise, and understanding of how HBO is thought to effect ischemia reperfusion injury defines which patients will benefit from HBO. Flap classification is based on an assessment of blood supply, tissue composition, and method of movement. Each of these elements must be evaluated, but it is blood supply that is most important. The blood supply to the flap is either axial, based on a named vessel, or random, based on the subdermal plexus. Commonly, flap compromise occurs when the surgeon tries to mobilize tissue outside the defined arterial supply, when there is a pedicle problem exists, or when free flaps are exposed to prolonged ischemia. The tissue composition of a flap may include skin, subcutaneous tissue, fascia, muscle, bone, other tissues, or a combination of these. Flap composition is very important because different tissue types have different tolerances to ischemia. For instance, a myocutaneous flap will be more susceptible to ischemia than a fasciocutaneous flap, because muscle is much more sensitive to ischemic injury than fascia and skin (26). In those circumstances where a prolonged primary ischemia or any secondary ischemia resulting from vessel thrombosis and revision anastomosis exists, the flaps will undergo ischemia reperfusion injury.

When treating compromised flaps, a multimodality approach should be initiated. This approach should include the use of vasodilators if arterial vasospasm is suspected, removal of sutures if tension or compression are suspected, dextran and pentoxifylline for rheological purposes, medicinal and chemical leeching for venous congestion, and the early use of hyperbaric oxygen if blood flow can be documented. The use of HBO therapy is appropriate only when the flap problem has been defined, documented perfusion of the flap exists, appropriate surgical salvage measures have been first considered, and HBO therapy can be performed in an expedient manner. Specifically with respect to free flaps, extended primary ischemia time greater than 2 h or any secondary ischemia time may result in partial or total flap necrosis. This injury is usually reversible if recognized early and treated expeditiously. Essentially, it is ischemia reperfusion injury. Numerous research studies support the use of HBO in the salvage of compromised free tissue transfers (27,28). A rat free-flap model showed similar improvement in flap survival (27). A clinical study evaluated free-flap salvage in the face of prolonged primary or any secondary ischemia (28). Salvage was significantly better in the HBO treatment group vs. controls, but only if initiated within 24 h. Free flaps compromised by prolonged primary or secondary ischemia have responded favorably to HBO treatment with complete salvage, in most cases, if HBO is started early. The treatment regimen is 2.0–2.4 ATA, 90 min q 8 h x 24 h, then q 8–12 h x 48 h (29). Treatment duration is based on clinical evaluation.

Grafts

Skin grafts are anatomically different from flaps in that skin grafts lack an inherent blood supply. Skin grafts are composed of avascular tissue that depends entirely on the recipient bed for oxygenation. HBO is useful in preparing the recipient bed and in promoting healthy granulation tissue to support split-thickness skin grafts. One controlled study showed a significant improvement in skin graft survival from 17% to 64% with the addition of HBO treatment. Although literature exists to support the use of HBO for composite grafts, a study by the University of Mississippi Medical Center found no significant effect of HBO on rat-ear composite grafts larger than 1 cm (30,31) Further research is needed to better understand the effects of HBO on composite graft survival. The rational for use of HBO in crush injury, compartment syndrome, frostbite, and other traumatic ischemias is similar to those for compromised flaps as they are all cases of ischemia and ischemia reperfusion injury.

CRUSH INJURY, COMPARTMENT SYNDROME, AND OTHER ACUTE TRAUMATIC ISCHEMIAS

These conditions are trauma-related situations in which the underlying pathophysiology is that of ischemia reperfusion (IR) injury. Ischemia times of greater than 4 h willresult in some degree of permanent necrosis. The physiologic basis of IR injury has become better understood in recent years. Most of the animal research centers around the production of oxygen-free radicals. Although the endothelial xanthine oxidase pathway has received much attention in the literature (32), more recent evidence supports the fact that neutrophils are a more important source of oxygen-free radicals via membrane NADPH oxidase and degranulation. Also, neutrophil adhesion is felt to cause ischemia reperfusion IR-associated vasoconstriction.

A perceived paradox exists related to HBO for IR injury. The less-informed observer often does not understand why HBO improves reperfusion injury and might think HBO instead increases free radical formation. (An oxygen-free radical is an oxygen molecules with an unpaired electron in its outer shell.) During ischemia, ATP is ultimately degraded to hypoxanthine and xanthine, which are anaerobic metabolites. With reperfusion, oxygenated blood is reintroduced into the ischemic tissue, and the hypoxanthine and xanthine plus oxygen creates oxygen-free radicals. Superoxide and hydroxyl oxygen-free radicals are formed, which can cause extensive tissue damage. The authors believe that the major mediator of damage is, in fact, neutrophil adherence to postcapillary venules significant and progressive vasoconstriction occurs in arterioles adjacent to leukocyte-damaged venules. Neutrophil adherence and vasoconstriction lead to a low flow state in the microcirculation and then vessel thrombosis, which is the endpoint of IR injury. The leukocyte-damaged venule is thought to be responsible for the arterial vasoactive response. HBO inhibits neutrophil adherence to the endothelial cells and thereby inhibits the ultimate thrombosis of microvessels, but the complete mechanism is still poorly understood, but is thought to involve the elevation in nitric oxide mediated by an increase in nitric oxide syntase (33). Free radical formation is not felt to be worsened with HBO as fewer adherent neutrophils actually exist to contribute to the neutrophil oxygen-free radical-generating system.

Treatment with hyperbaric oxygen in the face of IR injury carried the concern that that providing extra oxygen would increase free radical production and tissue damage. This query has been resolved by studies that have shown that HBO actually antagonizes the ill effects of IR injury in a variety of tissues (33–35). One of the first studies evaluating HBO and IR injury showed that HBO, immediately upon reperfusion, significantly improved skin flap survival following 8 h of global ischemia in a rat axial skin flap model with increased microvascular blood flow during reperfusion. Free-flap survival improves with HBO treatment during reperfusion even following ischemia times of up to 24 h (36). Hyperbaric oxygen administered during and up to 1 h following 4 h global ischemia significantly reduced neutrophil endothelial adherence in venules and also blocked the progressive arteriolar vasoconstriction associated with reperfusion injury (37). HBO inhibited *in vitro* beta-2-integrin (CD18)-induced neutrophil adherence function, but did not alter other important neutrophil functions such as oxidative burst or stimulus-induced chemotaxis and migration. This latter finding is very important, because HBO, through its action on the CD18 adhesion molecule,

blocks the neutrophil adherence associated with IR injury without interfering with other neutrophil functions that would increase the risk of infectious complications. Initially, the focus in acute ischemia caused by trauma should be restoration of blood supply. The authors, therefore, recommend HBO therapy for all patients with muscle ischemia time greater than 4 h and skin ischemia time greater than 8 h. The major effects of IR injury are felt to occur within the first 4–7 h of reperfusion. 2 ATA hyperbaric oxygen increases the tissue oxygen tension 1000%. Treatment protocol is 2.0–2.5 ATA for 60 min, q 8 h x 24 h, then q 8–12 h x 48 h with clinical re-evaluation. If progressive signs of ischemic injury are still present, the treatment is continued at 2.0 ATA, q 12 h for 2–3 more days. Usually, 72 h of treatment is adequate as long as the first treatment is initiated within 4 h of surgery.

RADIATION TISSUE DAMAGE AND OSTEORADIONECROSIS

1.2 million cases of invasive cancer are diagnosed yearly, half of which will receive radiation therapy and 5% of which will have serious radiation complications, which represents 30,000 cases per year of serious radiation sequellae (38). HBO is also well studied for its use in treating osteoradionecrosis in conjunction with adequate debridement of necrotic bone. Carl et al. also reported success is applying HBO to 32 women with radiation injury following lumpectomy and radiation compared with controls (39). Feldmeier and his colleagues reviewed the literature and found no evidence to support the potentiation of malignant cells or the engancement of cancer growth (40). The treatment protocol is 2.5 ATA for 90 min daily for 20–50 treatments. HBO can also be used as a radiosensitizer and are as much as three times more sensitive to radiation kisses than are hypoxic cells (41).

REFRACTORY OSTEOMYELITIS

Chronic refractory osteomyelitis (CROM) is infection of the medullary and cortical portions of the bone that persists or recurs following treatment with debridement and antibiotics. The principles of treatment are fairly simple. First, the dead bone is debrided and bone cultures should be taken along with administration of appropriate antibiotics. Next, the interface or cicatrix, which separates the compromised bone from adequate blood supply, is removed. Finally, hypoxia in the wound must be corrected, which may be accomplished by HBO. The treatment protocol is 2.0 ATA for 90 min daily for 20–60 treatments. Note that CROM and refractory osteomyelitis require the longest treatment protocols. HBO is believed to oxygenate hypoxic/ischemic tissues, augment host antimicrobial responses, augment osteoclastic activity, and induce osteogenesis in normal and infected bone and antibiotic synergism.

ACUTE THERMAL BURNS

HBO is approved by the USMS but it is not covered by Medicare. Gruber demonstrated in 1970 that the area around and under a third-degree burn was hypoxic and could only be raised by oxygen at increased pressure (42). HBO has been found to prevent extension, reduce edema, increase healing rates, and decrease total cost in several randomized studies (43,44). HBO is also thought to decrease the rate of burn sepsis based on several early studies. The controversy, in part, surrounds current guidelines for early debridement and grafting of burns. Once excised, a burn no longer exists and HBO will not be helpful. In case burns are not easily amenable to excision such as flash burns to the face or groin, HBO may be helpful to prevent extension of the burn and to aid healing. Treatment must be started within 24 h. The recommended regimen is 2.0 ATA for 90 min every 8 h on the first day, then every 12 hours for 5 or 6 days.

ACUTE EXCEPTIONAL BLOOD LOSS ANEMIA

Hyperbaric oxygen for treatment of acute blood loss anemia is reserved for those patients whose anemia is not immediately treatable for practical, disease process, or religious reasons, which may include warm antibody hemolytic disease, Jehova's Witnesses, those with rare blood types, and those who refuse transfusion for other personal reasons. As explained in the physiology section, HBO dramatically increases the amount of solublized oxygen the blood can carry. In fact, Boerema showed, in 1955, that pigs could be exsanguinated to four-tenths of one gram of hemoglobin per deciliter and be maintained in a hyperbaric environment of 3 ATA without hypoxia. The goal in HBO therapy for these conditions is to improve the oxygen depth with the daily or twice daily HBO treatments until the anemia can be improved. In between the treatments, the patients should be maintained a lower FIO_2 of inspired oxygen if possible to help reduce oxygen toxicity.

CARBON MONOXIDE POISONING

In 1966, Wada first used HBO to treat survivors of coal mine disasters with carbon monoxide poisoning and burns. The modern-day sources of carbon monoxide include automobile exhaust, home heaters, portable generators, propane engines, charcoal burners and camp stoves, and methylene chloride paint strippers. The initial treatment for carbon monoxide poisoning is 100% oxygen. The administration of 100% oxygen via a nonrebreather mask facilitates the dissociation of CO from hemoglobin to approximately 1.5 h. Hyperbaric oxygen delivered at 2.8–3.0 ATA reduced the halflife of CO-bound hemoglobin further to 23 min. In addition, patients who had one hyperbaric treatment for CO poisoning had 46% neuropsychiatric sequelae at discharge and 50% at one month versus two HBO treatments at 2.8–3.0 ATA having 13% at discharge and 18% at one month. The current recommendation is 3.0 ATA for 90 min with air breaks delivered every 8 h for a total of 3 treatments (called theWeaver protocol). Some authors still feel one treatment may be adequate (45).

CYANIDE POISONING

Hydrocyanide gas or HCN is formed when any number of substances burns, including furniture, asphalt, paper, carpeting (nylon), lighting baths (acrylic), plastic(polystyrene), and insulation (melamine resins). The antidote for cyanide poisoning begins with breathing 100% oxygen, ATLS protocols, and administration of IV sodium thiosulphate and is continued with a slow infusion of sodium nitrate and simultaneous HBO therapy if it is available. The sodium nitrate creates methemoglobin, which can impair the oxygen-carrying capacity of hemoglobin. HBO increases the amount of oxygen dissolved in plasma and may offer a direct benefit. The treatment regimin is 3.0 ATA with 30/10 airbreaks.

HYPERBARIC CHAMBER FACILITY DESIGN AND SAFETY

Over 500 hyperbaric facilities exist in the United States, and the number of hyperbaric chambers is steadily increasing worldwide. Hyperbaric chambers are classified as either monoplace or multiplace. They differ functionally in that the monoplace chamber instills oxygen into the entire chamber environment, whereas in a multiplace chamber, patients breathe 100% oxygen via a breathing mask or oxygen hood and exhaled gases are vented outside the chamber. Monoplace chambers are constructed either as an acrylic cylinder with metal ends or are primarily constructed of metal. Most commonly, the monoplace chambers are formed from an acrylic cylinder from 20 to 40 inches in diameter with tie rods connecting it to end caps. The opening is a rotating lock or a cam action lever closure. Separate oxygen and air sources provide the oxygen sources and air for air breaks during therapy. An oxygen vent must be exhausted outside the building. The through ports on the HBO chamber door allow passage of specially made intravenous monitoring devices and ventilators. The larger diameter monoplace chambers are more comfortable; however, they require more oxygen and can be heavier and more expensive to install. The acrylic chambers can provide a maximum of 3 ATA pressure. Alternatively, some monoplace chambers are constructed mostly of steel with acrylic view ports, which can accommodate pressures of up to 6 ATA and are often used in special situations such as offshore rigs where a compact chamber is needed to treat decompression illness required in U.S. Navy Table 5. Multiplace chambers are much larger and are designed to provide treatment to multiple people or to manage complex conditions. Some can even house operating rooms with special precautions. They are typically made of steel with acrylic view ports and are designed for operation up to 6 ATA or 165 feet of sea water. The gauges are reported in feet of sea water on these multiplace chambers to facilitate the use of dive tables for staff or patients. These multiplace chambers are, therefore, best-suited to treat deep water decompression illness. These chambers can accommodate from 2 to 20 people and have variable configurations including horizontal cylinders, spherical shapes, and rectangular chambers.

The primary professional hyperbaric medicine societies in the United States are the Undersea and Hyperbaric Medical Society (UHMS) and the American College of Hyperbaric Medicine. The UHMS has developed a clinical hyperbaric medicine facility accreditation program. This program can be accessed via the UHMS website at http://www.uhms.org, and it was designed to assure that clinical facilities are:

1. Staffed with well-trained specialists;
2. Using high quality equipment that is properly installed, maintained, and operated to the highest possible safety standards;
3. Providing high quality care;
4. Maintaining proper documentation of informed consent, treatment protocols, physician participation, training, and so on (46).

Safety Elements for Equipment and Facilities

The American Society of Mechanical Engineers (ASME) and the Pressure Vessel for Human Occupancy Committee (PVHO) define the design and fabrication guidelines for hyperbaric chambers. Although not required in all states or worldwide, it is accepted as the international standard (46). Next, the National Fire Protection Association (NFPA) has established a safety standard for hyperbaric facilities. The publication, NFPA 99, Safety Standard for Health Care Facilities, Hyperbaric Facilities, Chapter 20 explains the details of and criteria for equipment associated with a hyperbaric chamber facility. The requirements include fire abatement systems, air quality, and electrical requirements. These requirements apply to any hyperbaric chamber placed within a health-care facility. Each site must have a safety director. It is important to have only cotton clothing and to avoid any sources of sparks or static electricity given the 100% oxygen (Fig. 2). In addition to these guidelines, hyperbaric chambers are pressure vessels and, as such, are subject to boiler and pressure vessel laws. They are also medical devices and, in the United States, are also subject to FDA rules for class II medical devices. A chamber is required to have a clearance from the FDA before the device can be legally marketed or distributed, which is often called a 510 k clearance, denoting the form on which the clearance must be submitted. To check on whether a device has received clearance in the United States, one must contact the manufacturer or the Food and Drug Administration (FDA) most easily via their website, http://www.fda.gov/scripts/cdrh/cfdocds/cfpmn/dsearch.cfm.

Facilities must develop defined safety protocols and emergency plans that are available through both the Undersea and Hyperbaric Medicine Society (UHMS) and the American College of Hyperbaric Medicine (ACHM).

FRONTIERS AND INVESTIGATIONAL USES

The use of hyperbaric oxygen therapy has, at times, been surrounded with controversy and spurious claims from improving athletic performance to slowing the aging process. It is essential that the hyperbaric medicine

physician, staff, and potential patients understand and follow the principles of evidence-based practice, which means prescribing HBO therapy for the conditions proven to benefit from such treatment. The UHMS website, at www.UHMS. org, and AHCM are good resources for additional information as are numerous publications on hyperbaric medicine such as the hyperbaric medicine textbook available through the UHMS website. Investigational uses for hyperbaric oxygen therapy include carbon tetrachloride poisoning, hydrogen sulfide poisioning, sickle cell crisis, spinal cord injury, closed head injury, cerebral palsy, purpura fulminans, intraabdominal and intracranial abscess, mesenteric thrombosis, retinal artery occlusion, cystoid macular edema, bell's palsy, leprosy, lyme disease, stroke and traumatic brain injury, and brown recluse spider bite. Some of the many investigational uses for HBO therapy may have merit, but these must be rigorously studied using well-designed trials. As the field of hyperbaric medicine continues to advance, so will our understanding of the complex physiologic effects of delivering oxygen under pressure.

AKNOWLEDGMENT

Special thanks to Bob Bartlet, MD, for his help in preparing this chapter.

BIBLIOGRAPHY

1. Bakker DJ, Cramer FS. Hyperbaric surgery. Perioperative care. p 2.
2. Behnke AR, Shaw LA. Use of hyperbaric oxygen in treatment of compressed air illness. Nav Med Bull 1937;35:1–12.
3. Boerma I. High tension oxygen therapy. Proc Royal Soc Med 1964;57(9):817–818.
4. Bakker DJ, Cramer FS. Hyperbaric Surgery Perioperative Care. p. 67.
5. Jain KK. Physical Physiological, and Biochemical Aspects of Hyperbaric Oxygenation. Textbook of Hyperbaric Medicine. Toronto: Hogrefe & Huber Publishers; 1990. p 11.
6. Matos L, Nunez A. Enhancement of healing in selected problem wounds. In: Kindwall EP, Whelan HT, eds. Hyperbaric Medicine Practice, 2nd ed. Flagstaff AZ; Best; 1999.
7. Hunt TK, Pai MP. The effect of varying ambient oxygen tensions on wound metabolism and collagen synthesis. Surg Gynecol Obstet 1972;135:561–567.
8. Knighton DR, Halliday BJ, Hunt TK. Oxygen as an antibiotic: A comparison of the effects of inspired oxygen concentration and antibiotic administration on in vivo bacterial clearance. Arch Surg 1986;121:191–195.
9. Zamboni WA, Roth AC, Russell RC, Graham B, Suchy H, Kucan JO. Morphologic analysis of the microcirculation during reperfusion of ischemic skeletal muscle and the effect of hyperbaric oxygen. Plast Reconstr Surg 1993;91(6):1110–23.
10. Zamboni WA, et al. The effect of acute hyperbaric oxygen therapy on axial pattern skin flap survival when administered during and after total ischemia. J Reconstr Microsurg 1989;5:343–347.
11. Muth CM. Gas embolism (Review). New Eng J Med 342: 476–482.
12. Divers Alert Network: Report on 1994 Diving Accidents. Durham, NC: Duke University; 1995.
13. Barnard EEP, Hanson JM, et al. Post decompression shock due to extravasation of plasma. BMJ 1966;2:154.
14. Powel MR, Spencer MP, Von Ramm OT. Ultrasonic surveillance of decompression. In: Bennett PB, Elliott DH, eds. The Physiology of Diving and Compressed Air Work, 3rd ed. London Bailliere: Tindall; 1982. pp 404–434.
15. Zamboni WA. Applications of hyperbaric oxygen therapy in plastic surgery. In: Oriani G, Marroni A, eds. Handbook on Hyperbaric Medicine, 1st ed. New York: Springer; 1995. p. 443–484.
16. Hunt TK. The physiology of wound healing. Ann Emerg Med 1988;17:1265–1273.
17. Hunt TK, Hopf HW. Wound healing and wound infection. Surg Clinics N Am 1997;77(3):587–606.
18. Oriani G, Micheal M, Meazza D, et al. Diabetic foot and hyperbaric oxygen therapy: A ten-year experience. J Hyperbar Med 1992;7:213–221.
19. Wattel FE, Mathieu DM, Fossati P, et al. Hyperbaric oxygen in the treatment of diabetic foot lesions: Search for healing predictive factors. J Hyperbar Med 1991;6:263–267.
20. Baroni G, Porro T, Faglia E, Pizzi G, et al. Hyperbaric oxygen in diabetic gangrene treatment. Diabetes Care 1987;10:81–86.
21. Oriani G, Micheal M, Meazza D, et al. Diabetic foot and hyperbaric oxygen therapy: A ten-year experience. J Hyperbar Med 1992;7:213–221.
22. Strauss MB, Bryant BJ, Hart GB. Transcutaneous oxygen measurements under hyperbaric oxygen conditions as a predictor for healing of problem wounds. Foot Ankle Int. 2002 Oct; 23(10):933–7.
23. Zamboni WA, Roth AC, Russel RC, Nemiroff PM, Casas L, Smoot EC. Hyperbaric oxygen improves axial skin flap survival when administered during and after total ischemia. J Reconstr Micro 1989;5:343–347.
24. Hunt TK, Pai MP. The effect of varying ambient oxygen tensions on wound metabolism and collagen synthesis. Surg Gyn Obstet 1972;135:561–567.
25. Niinikoski J, Hunt TK. Oxygen Tension in Human Wounds. J Surg Res 1972;12:77–82.
26. Mathieu D, et al. Pedicle musculocutaneous flap transplantation: prediction of final outcome by transcutaneous oxygen measurements in hyperbaric oxygen. Plast Reconstr Surg 1993;91:329–334.
27. Waterhouse MA, et al. The use of HBO in compromised free tissue transfer and replantation: A clinical review. Undersea Hyperb Med 1993;20(Suppl):54 (Abstract).
28. Perrins DJD, Cantab MB. Influence of hyperbaric oxygen on the survival of split skin grafts. Lancet 1967;1:868–871.
29. Persons BL, Zamboni WA. Hyperbaric oxygen in plastic and reconstructive surgery. In: Bakker DJ, Cramer FS, eds. Hyperbaric Surgery Perioperative Care. Flagstaff, AZ: Best; 2002.
30. Mazelowski MC, Zamboni WA, Haws MF, Smoot EC, Stephenson LL. Effect of hyperbaric oxygen on composite graft survival in a rat ear model. Undersea and Hyperbaric Med Suppl 1995;22:50.
31. McFarlane RM, Wermuth RE. The use of hyperbaric oxygen to prevent necrosis in experimental pedicle flaps and composite skin grafts. Plast Reconstr Surg 1966; 37:422–430.
32. Angel MF, et al. Free radicals: Basic concepts concerning their chemistry, pathophysiology, and relevance to plastic surgery. Plast Reconstr Surg 79:990.
33. Lozano DD, Zamboni WA, Stephenson LL. Effect of hyperbaric oxygen and medicinal leeching on survival of axial skin flaps subjected to total venous occlusion. Undersea Hyperb Med suppl 1997;24:86.
34. Thom SR, Bhopale V, Fisher D, Manevich Y, Huang PL, Buerk DG. Stimulation of nitric oxide synthase in cerebral cortex due to elevated partial pressures of oxygen: an oxidative stress response. J Neurobiol 2002;51(2):85–100.

35. Jones S, Wang WZ, Natajaraj C, Khiabani, Stephenson LL, Zamboni WA. HBO inhibits IR induced Neutrophil CD 18 Polarization by a nitric oxide mechanism. Undersea Hyperb Med 2002;35 (Suppl):75.
36. Zamboni WA, Roth AC, Russel RC, Nemiroff PM, Casas L, Smoot EC. Hyperbaric oxygen improves axial skin flap survival when administered during and after total ischemia. J Reconstr Micro 1989;5:343–347.
37. Gimbel M, Hunt TK. Wound healing and hyperbaric oxygen. In: Kindwall EP, Whelan HT, eds. Hyperbaric Medicine Practice, 2nd ed. Flagstaff, AZ: Best; 1999. p 169–204.
38. Bartlett B. Hyperbaric therapy. Radiation Injury 1994:2–3. HBO has been shown to increase angiogenesis and blood flow in previously irradiated tissue or bone. (Marx RE, Ehler WJ, Taypongsak PT, Pierce LW. Relationship of oxygen dose to angiogenesis induction in irradiated tissue. Am J Surg 1990; 160:519–524.
39. Carl UM, Feldmeier JJ, Schmitt G, Hartmann KA. Hyperbaric oxygen therapy for late sequelae in women receiving radiation after breast conserving surgery. Int J Radiat Oncol Biol Phys 2001;49:1029–1031.
40. Feldmeier JJ. Hyperbaric oxygen: Does it have a cancer causing or growth enhancing effect. Proceedings of the consensus conference sponsored by the European society for therapeutic radiology and oncology and the European committee for hyperbaric medicine. Portugal, 2001:129–146.
41. Gray KH, Conger AD, Ebert M, Hornsey S, Scott OCA. The concentration of oxygen dissolved in tissues at the time of irradiation as a factor in radiotherapy. Br J Radiol 1953;26: 638–648.
42. Wada J, Ikeda T, Kamata K, Ebuoka M. Oxygen hyperbaric treatment for carbon monoxide poisoning and severe burn in coal mine gas explosion. Igakunoaymi (Japan) 1965;54–68.
43. Germonpre P, Reper P, Vanderkelen A. Hyperbaric oxygen therapy and piracetam decrease the early extension of deep partial thickness burns. Burns 1996;6:468–473.
44. Cianci P, Sato R, Green B. Adjunctive hyperbaric oxygen reduces length of hospital stay, surgery and the cost of care in severe burns. Undersea Biomed Research Suppl 1991;18:108.
45. Bartlett R. Carbon monoxide poisoning. In: Haddad M, Shannon M, Winchester J, eds. Poisoning and Drug Overdose, 3rd ed. New York: WB Saunders Company; 2002.
46. Bakker DJ, Cramer FS. Hyperbaric Surgery Perioperative Care. Flagstaff, AZ: Best Publishing; 2002.

See also BLOOD GAS MEASUREMENTS; HYPERBARIC OXYGENATION; OXYGEN MONITORING; PULMONARY PHYSIOLOGY; RESPIRATORY MECHANICS AND GAS EXCHANGE; VENTILATORY MONITORING.

HYPERBARIC OXYGENATION

HARRY T. WHELAN
JEFFREY A. NIEZGODA
MATTHEW C. LEWIS
Medical College of Wisconsin
Milwaukee, Wisconsin

ERIC P. KINDWALL
BERNADETTE CABIGAS
St. Luke's Medical Center
Milwaukee, Wisconsin

INTRODUCTION

Hyperbaric oxygen (HBO) is simply the delivery of molecular oxygen in very high dosage. Even though experience has shown HBO to be very useful in a number of conditions, the exact mechanism of action at the molecular level is not fully understood. Studies done by Thom et al. (1) demonstrated that elevated oxygen tensions stimulated neuronal nitric oxide synthase (NOS1) and increased steady-state nitric oxide concentration in their microelectrode-implanted rodents. Buras et al. (2) in their studies with human umbilical vein endothelial cells (HUVEC) and bovine aortic endothelial cells (BAEC) showed that hyperbaric oxygen (HBO) down-regulated intracellular adhesion molecule 1 (ICAM 1) expression via the induction of endothelial nitric oxide synthase (NOS3), which proved beneficial in treating ischemia reperfusion injuries. Other studies talk about interactions between nitric oxide and oxygen species and their role in various disease states. Clearly, interest in HBO is growing.

Boerema (3) introduced hospital use of the hyperbaric chamber in the late 1950s in Holland, simply to maintain a semblance of normoxia in patients undergoing cardiac surgery. Heart–lung machines had not yet been invented, and the use of the chamber made certain kinds of cardiac surgery possible for the first time. Boerema felt that if enough oxygen could be driven physically into solution in the tissues, which he termed "drenching", the circulation to the brain could be interrupted longer than 3–4 min. It also rendered surgery on many pediatric patients less risky. For example, if the normal arterial pO_2 in a patient with Tetralogy of Fallot was 38 mmHg, placing him in the chamber might raise it to 94 mmHg. Operating on the patient under hyperbaric conditions posed much less risk of ventricular fibrillation when the heart or great vessels were manipulated.

This idea caught on quickly, and soon large surgical hyperbaric chambers were built in Glasgow, New York, Los Angeles, Chicago, Minneapolis, and at Boston Children's Hospital. By the early 1960s, however, heart–lung machines became more common, and the need to do surgery in the hyperbaric chamber diminished substantially. Many large surgical chambers were left to gather dust or were dismantled, as hospital floor space is always at a premium. During this time the surgeons, who had been doing most of the research, left the field. Of the nondiving conditions, only carbon monoxide poisoning and gas gangrene seemed to be likely candidates for hyperbaric oxygen treatment based on credible research.

In 1969, however, a double-blind controlled study on the use of hyperbaric oxygen in senility was published in *The New England Journal of Medicine*. Results seemed promising, and this initiated the propagation of hyperbaric quackery. The original investigators made no sweeping claims for the research, but simply felt that the area merited further investigation. Eventually, further research showed that the results of the study reported in the *New England Journal* article were a statistical anomaly and could not be reproduced. However, hyperbaric enthusiasts seized upon the earlier report, and senility began to be treated in hyperbaric chambers, along with a host of other diseases. Most of these were not in medical centers. Fly-by-night "clinics" suddenly appeared claiming to cure anything and everything. Patients were treated for skin wrinkles, loss of sexual vigor, and a host of other

maladies. As there were few investigators doing good research in the area at that time, the field fell into disrepute.

Fortunately, a few legitimate investigators persisted in their work, looking at the effects of hyperbaric oxygen in greater detail. Soon it became clear that under hyperbaric conditions oxygen had some unusual effects. The Undersea and Hyperbaric Medical Society created a committee to investigate the field. After careful study, the committee laid down guidelines for what should be reimbursed by third-party payers and what conditions should be considered investigational. Their report appeared in 1977 and was adopted as a source document for Blue Cross/Blue Shield (4). About the same time, Jefferson C. Davis of the United States Air Force School of Aerospace Medicine edited the first textbook in hyperbaric medicine (5). It was only then that a firm scientific basis was reestablished for the field, leading to increased acceptance by the medical community. The number of chambers operating in hospitals has risen dramatically from only 37 in 1977 to > 500 today. The Undersea and Hyperbaric Medical Society (www.UHMS.org) and the American College of Hyperbaric Medicine (www.ACHM.org) have taken responsibility for setting standards in this field and for encouraging additional research. At this time, ~ 13 clinical disorders have been approved for hyperbaric treatment. They include air or gas embolism, carbon monoxide poisoning, clostridial myonecrosis, crush injury or compartment syndrome, decompression sickness, problem wounds, severe blood loss anemia, necrotizing soft tissue infections, osteomyelitis, radiation tissue damage, skin grafts or flaps, thermal burns and brain abscess.

Remember that hyperbaric oxygen was introduced initially into hospitals in order to simply maintain normoxia or near-normoxia in patients undergoing surgery. It was only later, and quite serendipitously that researchers discovered that oxygen under increased atmospheric pressure gained some of the attributes of a pharmacologic agent. Oxygen begins to act like a drug when given at pressures of 2 atm or greater. For example, oxygen under pressure can terminate lipid peroxidation *in vivo* (6), it can enhance the bacteriocidal capabilities of the normal leukocyte (7,8), and it can stimulate the growth of new capillaries in chronically ischemic tissue, such as in the diabetic foot, or in tissue that has undergone heavy radiation. It can reduce intracranial pressure on the order of 50% within seconds of its initiation, and this effect is additive to that of hypocapnia (9–11). HBOT can increase the flexibility of red cells, augmenting the effects of pentoxifylline (12). It can decrease edema formation by a factor of 50% in postischemic muscle and prevent second-degree burn from advancing to full-thickness injury (13–15). Hyperbaric oxygen has also been shown to hasten functional recovery of traumatized peripheral nerves by almost 30% following repair. Many of these discoveries have been made only in the last decade.

In a number of these areas, we are beginning to understand the basic mechanisms of action, but overall very little is understood at the molecular level. It is anticipated that studies involving nitric oxide synthase will provide insight regarding the elusive molecular mechanistic explanation. Also, many contributions to our understanding have come from advances made in the biochemistry of normal wound healing. We understand that normal oxygen pressures are 80–90-mmHg arterially, that oxygen enters our tissues from the capillaries, and that at this interface carbon dioxide (CO_2) is removed. Under hyperbaric conditions, all of this changes. At a chamber pressure of 2.4 atm (ATA), the arterial oxygen pressure (pO_2) reaches ~ 1500 mmHg, immediately saturating the red blood cells (RBCs). Upon reaching the tissues, these RBCs never unload their oxygen. At this high partial pressure of gas, oxygen diffuses into the tissues directly from the plasma. Returning to the heart, the RBCs are bathed in plasma with a pO_2 of 150–200 mmHg. Tissue oxygen requirements are completely derived from the plasma. In theory, one might think that this condition could prove fatal, as red cells no longer can carry CO_2 away from the tissues. However, we are fortunate that CO_2 is 50 times more soluble in plasma than are oxygen and nitrogen, and the body has a very capable buffering system which overcomes the loss of the Haldane effect, which is the increase in CO_2 carrying capacity of deoxygenated red cells (16).

Another factor to be considered is the actual part of the circulatory system that overcomes the loss of the Haldane effect. Traditionally, we think of this exchange occurring in the capillaries. Under very high pressures, however, computer modeling has shown that nitrogen exchange under pressure (as in deep sea divers) is probably complete by the time the blood reaches the arteriolar level. Whether this is true when hyperbaric oxygen is breathed has not yet been determined. The rate of metabolism under hyperbaric conditions appears to be unchanged, and the amount of CO_2 produced appears to be about the same as when breathing air. It would be interesting to know just at what level oxygen exchange is accomplished in the tissues, as this might have practical implications when treating people with severe capillary disease.

Oxygen can be toxic under pressure. Pulmonary toxicity and lung damage can be seen at oxygen pressures > 0.6 atm during chronic exposure. Central nervous system (CNS) toxicity can manifest as generalized seizure activity when oxygen is breathed at pressures of 3 atm or greater. The CNS toxicity was first observed by Paul Bert in 1878, and is termed the "Paul Bert Effect" (17). Despite years of research into this phenomenon, the exact underlying or molecular cause of the seizure has not yet been discovered. There is a generalized vasoconstriction that occurs when oxygen is breathed at high pressure, reducing blood flow to muscle, heart, and brain by a factor of $\sim 20\%$, as a defense against toxic quantities of oxygen. The exact mechanism responsible for this phenomenon is not fully understood.

Central nervous system oxygen toxicity was evaluated by the Royal Navy. The purpose of this research was to determine the time until convulsion so that combat swimmers would know their endurance limits under various conditions. Volunteer research subjects swam in a test tank using closed-circuit oxygen rigs until convulsion occurred and thus established safe oxygen tolerance boundaries.

Also related to the effect of oxygen, the "off" phenomenon (18) was first described by Donald in 1942. He observed that seizures sometimes occurred when the chamber pressure was reduced or when a diver surfaced and oxygen breathing under pressure was suddenly terminated. Lambertsen (19) provided a description of this type of seizure activity:

> The convulsion is usually but not always preceded by the occurrence of localized muscular twitching, especially about the eyes, mouth and forehead. Small muscles of the hands may also be involved, and incoordination of diaphragm activity in respiration may occur. After they begin, these phenomena increase in severity over a period which may vary from a few minutes to nearly an hour, with essentially clear consciousness being retained. Eventually an abrupt spread of excitation occurs and the rigid tonic phase of the convulsion begins. Respiration ceases at this point and does not begin again until the intermittent muscular contractions return. The tonic phase lasts for about 30 seconds and is accompanied by an abrupt loss of consciousness. It is followed by vigorous clonic contractions of the muscle groups of the head and neck, trunk and limbs. As the incoordinated motor activity stops, respiration can proceed normally.

Within the wound healing community, current doctrine holds that a tissue pO_2 of 30–40 mmHg is necessary for adequate wound healing (20,21). Below 30 mmHg, fibroblasts are unable to replicate or produce collagen. Additionally, when the pO_2 drops < 30 mmHg, leukocytes are unable to utilize oxidative mechanisms to kill bacteria. We have noted that the tissue pO_2 is critical, but that the actual quantity of oxygen consumed in wound healing is relatively small. The amount of oxygen used to heal a wound is only $\sim 10\%$ of that required for brain metabolism.

Production of new collagen is also a requirement for capillary ingrowth or proliferation (22). As capillaries advance, stimulated by angiogenic growth factor, they must be supported by an extracellular collagen matrix to facilitate ingrowth into tissue. In the absence of new collagen, capillary ingrowth cannot occur. This effect is crucial in treating radionecrosis (23–25), where the tissue is primarily hypovascular, and secondarily hypoxic and hypocellular. It has been discovered that when collagen production can be facilitated, new capillaries will invade the previously irradiated area, and healing will then occur. The tissue pO_2 rises to $\sim 80\%$ of normal and plateaus; however, this is sufficient for healing and will even support bone grafting. Historically, the only means of managing radionecrosis was to excise the radiated area and bring in fresh tissue with its own blood supply. New collagen formation and capillary ingrowth also account for the rise in tissue pO_2, which can be achieved in patients with diabetic foot lesions.

It is now well understood that the stimulus for growth factor production by the macrophage is hypoxia and/or the presence of lactic acid (26,27). Wounds managed in hyperbaric units are typically ischemic and hypoxic. Periods of relative hypoxia, required for the stimulation of growth factor production, exist between hyperbaric treatments.

Surprisingly, oxygen levels remain high in tissues for longer than one would expect following hyperbaric treatment. In a study by George Hart (28) at Long Beach Memorial Hospital, a mass spectrometer probe was inserted in the unanesthetized thigh tissues of normal volunteers. Muscle and subcutaneous tissue pO_2 values in study subjects remained significantly elevated for 2–3 h following hyperbaric oxygen treatment. Arterial pO_2 was also measured and found to rise immediately and significantly under hyperbaric conditions but returned to normal levels within a couple of minutes upon egress from the chamber (Fig. 1). Thus, multiple daily HBO treatments can maintain useful oxygen levels for up to 12 h/day.

Mention has been made of enhanced leukocyte killing of bacteria under hyperbaric conditions. Jon Mader of the University of Texas-Galveston (29) carried out a rather simple, but elegant, experiment to demonstrate this. The fascinating part of this study is that in the evolution of the human body, a leukocyte has never been exposed to a partial pressure of 150 mmHg while in tissues. This level is impossible to attain breathing air. Nevertheless, when one artificially raises the pO_2 far beyond the leukocyte's normal functional parameters, it becomes even more lethal to bacteria. This is an anomaly, as one rarely can improve on Mother Nature. Of some interest in this regard is that if one bites one's tongue, one is never concerned about possible infection, even though it is a human bite. Similarly, hemorrhoidectomies rarely, if ever, become infected. The reason is that the pO_2 of the tissues in and around the oral cavity are very high, and the pO_2 in hemorrhoidal veins is nearly arterial. Tom Hunt has shown it is impossible to infect tissue that is injected with raw staphylococci if the pO_2 in the same tissue is > 50 mmHg. Both he and David Knighton have described oxygen as an antibiotic (30,31).

The reduction of intracranial pressure is facilitated by vasoconstriction. Experimentally, Rockswold has shown that mortality can be halved in victims of closed head injury with Glasgow Coma Scales in the range of 4–6. One of the major mechanisms here is a reduction of intracranial pressure while continuing to oxygenate hypoxic brain (32–36). Sukoff et al. (37) administered 100% O_2, 1.5 ATA $\times$ 60 min every 24 h (maximum of 7 h) to severely brain injured patients. This resulted in a 50% reduction in mortality.

A paper published by Mathieu (38) has shown that the flexibility index of red cells can be changed from 23.2 to 11.3 within 15 hyperbaric treatments. This increase in flexibility can prove quite useful in people with narrowed capillaries. However, whether this phenomenon plateaus at 15 treatments, its duration and underlying mechanism are still unknown.

Nylander et al. (39) demonstrated that following complete occlusion of the blood flow to rat leg for 3 h, postischemic edema could be reduced by 50% if the animals are promptly treated with hyperbaric oxygen. He also demonstrated that the mechanism for this was preservation of adenosine triphosphate (ATP) in the cells, which provides the energy for the cells to maintain their osmolarity. Cianci

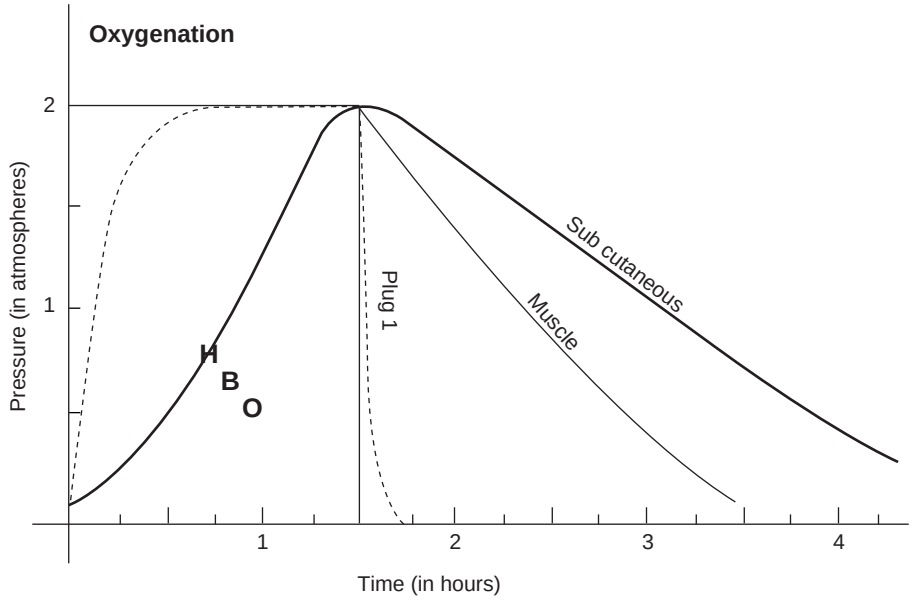


Figure 1. Arterial, muscle and subcutaneous pO_2 after HBO treatment.

(40,41). Yamaguchi, and others have underscored the importance of ATP in preventing edema in burn. Treatment twice daily has shown to be more efficacious than treatment once a day.

Zamboni (42) and Gingrass, working at the University of Southern Illinois, have shown the effects of hyperbaric oxygen on speeding functional return in peripheral nerve repair and grafting. At 6 weeks, there is a 28% improvement of function in the affected leg of these rats.

Niezgoda (43) performed a randomized and double-blinded study in human volunteers investigating the effect of hyperbaric oxygen in a controlled burn wound model. He demonstrated statistically significant decreases in edema formation and wound exudate in the hyperbaric oxygen treated group.

Finally, the mechanism for the effects of hyperbaric oxygen in carbon monoxide poisoning (44,45) is now better understood. Traditionally, it was felt that the mere presence of carboxyhemoglobin blocked transport of oxygen to the tissues. However, studies by Goldbaum et al. (46) at the Armed Forces Institute of Pathology in 1975 lead us to different conclusions. Impairment of cytochrome *A*3 oxidase and lipid peroxidation occurring following a reperfusion injury are now suggested as the primary pathways in the pathophysiology causing fatality. Stephen Thom (47–50) at the University of Pennsylvania pioneered research in this area. It appears that as carbon monoxide levels fall, the products of lipid peroxidation rise, indicating that brain damage is occurring only during the resuscitative phase, thus becoming reperfusion injury. Thom demonstrated that a period of hypotension (even though it may only be a matter of seconds) is enough to initiate lipid peroxidation. Oxygen at 1 atm has no effect on halting the process. However, oxygen at 3 atm terminates lipid peroxidation. Patients who have been treated acutely with hyperbaric oxygen rarely exhibit signs of delayed deterioration, reported in 30–40% of severe cases treated only with normobaric oxygen. The probable mechanism for this is the ability of hyperbaric oxygen at three ATA to terminate lipid peroxidation.

Finally, in many ways it seems paradoxical that oxygen at high pressure, which intuitively would seem to provide more substrate for free-radical formation, still benefits tissues from crush injury and postischemic states. But ironically, it is precisely this hyperbaric environment that promotes certain so-called reactive oxygen species with inherent protective qualities (51). More studies are certainly needed to investigate the underlying pharmacologic benefits afforded by hyperbaric oxygen. We have only just begun to explore and utilize a treatment modality whose time has come.

BIBLIOGRAPHY

1. Thom SR, Fisher D, Zhang J, Bhopale VM, Ohnishi ST, Kotake Y, Ohnishi T, Buerk DG. Stimulation of perivascular nitric oxide synthesis by oxygen. Am J Physiol Heart Circ Physiol 2003;284:H1230–1239.
2. Buras JA, Stahl GL, Svoboda KKH, Reenstra WR. Hyperbaric oxygen downregulates ICAM-1 expression induced by hypoxia and hypoglycemia: the role of NOS. Am J Physiol Cell Physiol 2000;278:C292–C302.
3. Boerema I, Kroll JA, Meijne NG, Lokin E, Kroon B, Huiskes JW. High atmospheric pressure as an aid to cardiac surgery. Arch Chir Neerl 1956;8:193–211.
4. Kindwall EP. Report of the committee on hyperbaric oxygenation. Bethesda: Undersea Medical Society; Bethesda 1977.
5. Davis JC, Hunt TK. Hyperbaric oxygen therapy. Undersea Medical Society; Bethesda 1977.
6. Thom SR. Molecular mechanism for the antagonism of lipid peroxidation by hyperbaric oxygen. Undersea Biom Res (Suppl) 1990;17:53–54.
7. Andersen V, Hellung-Larsen P, Sorensen SF. Optimal oxygen tension for human lymphocytes in culture. J Cell Physiol 1968;72:149–152.
8. Gadd MA, McClellan DS, Neuman TS, Hansbrough JF. Effect of hyperbaric oxygen on murine neutrophil and T-lymphocyte functions. Crit Care Med 1990;18:974–979.

9. Hayakawa T, Kanai N, Kuroda R, et al. Response of cerebrospinal fluid to hyperbaric oxygenation. J Neurol Neurosurg Psych 1971;34:580–356.
10. Miller JD, Fitch W, Ledingham IM, et al. Reduction of increased intracranial pressure. Neurosurgery 1970;33: 287–296.
11. Sukoff MH, Ragatz RE. Hyperbaric oxygenation for the treatment of acute cerebral edema. Neurosurgery 1982;10: 29–38.
12. Nemiroff PM. Synergistic effects of pentoxifylline and hyperbaric oxygen on skin flaps. Arch Otolaryngol Head Neck Surg 1988;114:977–981.
13. Cianci P, Lueders H, Shapiro R, Sexton J, Green B. Current status of adjunctive hyperbaric oxygen in the treatment of thermal wounds. In: Proceedings of the second Swiss symposium on hyperbaric medicine; Baker DJ, JS, editors. Foundation for Hyperbaric Medicine; Basel: 1988. p 163–172.
14. Grossman AR. Hyperbaric oxygen in the treatment of burns. Ann Plast Surg 1978;1:163–171.
15. Hart GB, O'Reilly RR, Broussard ND, Cave RH, Goodman DB, Yanda RL. Treatment of burns with hyperbaric oxygen. Surg Gynacol Obstet 1974;139:693–696.
16. Coburn RF, Forster RE, Kane PB. Considerations of the physiological variables that determine the blood carboxyhemoglobin concentrations in man. J Clin Invest 1965;44: 1899–1910.
17. Bert P. Barometric Pressure. 1879. p 579. (translated by Hitchcock MS, Hitchcock FA) Bethesda: Reprinted by the Undersea and Hyperbaric Medicine Society; 1978.
18. Donald KW. Oxygen poisoning in man. Br Med J 1947; 712–717.
19. Lambertsen CJ. In: Fenn WO, Rahn H, editors. Handbook of Physiology, Respiration. Washington, D.C.: American Physiological Society; Section 3, Volume II. p 1027–1046.
20. Knighton DR, Hunt TK, Scheuenstuhl H, Halliday B, Werb Z, Banda MJ. Oxygen tension regulates the expression of angiogenesis factor by macrophages. Science 1983;221:1283–1285.
21. Knighton DR, Oredsson S, Banda MJ, Hunt TK. Regulation of repair: hypoxic control of macrophage mediated angiogenesis. In: Hunt TK, Heppenstall RB, Pines E, Rovee D, editors. Soft and hard tissue repair. New York: Praeser; 1948. p 41–49.
22. Knighton DR, Hunt TK, Thakral KK, Goodson WH. Role of platelets and fibrin in the healing sequence, an in vivo study of angiogenesis and collagen synthesis. Ann Surg 1982;196: 379–388.
23. Davis JC. Soft tissue radiation necrosis: The role of hyperbaric oxygen. HBO Rev 1987;2(3):153–167.
24. Davis JC, et al. Hyperbaric oxygen: A new adjunct in the management of radiation necrosis. Arch Otol 1979;105:58–61.
25. Hart GB, Mainous EG. The treatment of radiation necrosis with hyperbaric oxygen (OHP). 1976;37:2580–2585.
26. Jensen JA, Hunt TK, Scheuenstuhl H, Banda MJ. Effect of lactate, pyruvate, and physician on secretion of angiogenesis and mitogenesis factors by macrophages. Lab Invest 1986;54: 574–578.
27. Knighton DR, Schumerth S, Fiegel VD. Microenvironmental regulation of macrophage growth factor production. In preparation.
28. Hart GB, Wells CH, Strauss MB. Human skeletal muscle and subcutaneous tissue carbon dioxide, nitrogen and oxygen gas tension measurement under ambient and hyperbaric conditions. J App Res Clin Exper Therap Spring 2003;3(2).
29. Wang J, Corson K, Mader J. Hyperbaric oxygen as adjunctive therapy in vibrio vulnificus septicemia and cellulites. Undersea Hyperbaric Med Spring 2004, 31(1):179–181.
30. Hunt TK, Linsey M, Grislis G, Sonne M, Jawetz E. The effect of differing ambient oxygen tensions on wound infections. Ann Surg 1975;181:35–39.
31. Knighton DR, Halliday B, Hunt TK. Oxygen as an antibiotic. A comparison of the effects of inspired oxygen concentration and antibiotic administration on in vivo bacterial clearance. Arch Surg 1986;121:191–195.
32. Miller JD, et al.The effect of hyperbaric oxygen on experimentally increased intracranial pressure. J Neurosurg 1970;32: 51–54.
33. Miller JD, Ledingham IM. Reduction of increased intracranial pressure: Comparison between hyperbaric oxygen and hyperventilation. Arch Neurol 1971;24:210–216.
34. Mogami H, et al.Clinical application of hyperbaric oxygenation in the treatment of acute cerebral damage. J Neurosurg 1969;31:636–643.
35. Sukoff MH, et al.The protective effect of hyperbaric oxygenation in experimental cerebral edema. J Neurosurg 1968;29: 236–241.
36. Sukoff MH, Ragatz RE. Hyperbaric oxygen for the treatment of acute cerebral edema. Neurosurgery 1982;10(1):29–38.
37. Sukoff MH. Effects of hyperbaric oxygenation [comment]. J Neurosurg 2001;94(3):403–411.
38. Mathieu D, Coget J, Vinkier L, Saulnier F, Durocher A, Wattel F. Red blood cell deformability and hyperbaric oxygen therapy. (Abstract) HBO Rev 1985;6:280.
39. Nylander G, Lewis D, Nordstrom H, Larsson J. Reduction of postischemic edema with hyperbaric oxygen. Plast Reconstr Surg 1985;76:596–601.
40. Cianci P, Lueders HW, Lee H, Shapiro RL, Sexton J, Williams C, Green B. Adjunctive hyperbaric oxygen reduces the need for surgery in 40-80% burns. J Hyper Med 1988;3: 97–101.
41. Cianci P, Lueders HW, Lee H, Shapiro RL, Sexton J, Williams C, Green B. Hyperbaric oxygen and burn fluid requirements: Observations in 16 patients with 40-80% TBSA burns. Undersea Biomed Res (Suppl) 1988;15:14.
42. Zamboni WA, Roth AC, Russell RC, Nemiroff PM, Casa L, Smoot C. The effect of acute hyperbaric oxygen therapy on axial pattern skin flap survival when administered during and after total ischemia. J Reconst Microsurg 1989;5: 343–537.
43. Niezgoda JA, Cianci P. The effect of hyperbaric oxygen on a burn wound model in human volunteers. J Plast Reconstruct Surg 1997;99:1620–1625.
44. Brown SD, Piantadosi CA. Reversal of carbon monoxide-cytochrome C oxidase binding by hyperbaric oxygen in vivo. Adv Exp Biol Med 1989;248:747–754.
45. End E, Long CW. Oxygen under pressure in carbon monoxide poisoning. J Ind Hyg Toxicol 1942;24:302–306.
46. Goldblum LR, Ramirez RG, Absalon KB. Joint Committee on Aviation Pathology XII. What is the mechanism of carbon monoxide toxicity? Aviat Space Environ Med 1975;46(10): 1289–1291.
47. Thom SR. Antagonism of carbon monoxide-mediated brain lipid peroxidation by hyperbaric oxygen. Toxicol Appl Pharmacol 1990;105:340–344.
48. Thom SR, Elbuken ME. Oxygen-dependent antagonism of lipid peroxidation. Free Rad Biol Med 1991;10:413–426.
49. Thom SR. Carbon-monoxide mediated brain lipid peroxidation in the rat. J Appl Physiol 1990;68:997–1003.
50. Thom SR. Dehydrogenase conversion to oxidase and lipid peroxidation in brain after carbon monoxide poisoning. J Appl Physiol 1992;73:1584–1589.
51. Thom SR, Bhopale V, Fisher D, Manevich Y, Huang PL, Buerk DG. Stimulation of nitric oxide synthase in cerebral cortex due to elevated partial pressures of oxygen: An oxidative stress response. J Neurobiol 2002;51:85–100.

See also HYBERBARIC MEDICINE; OXYGEN MONITORING.

HYPERTENSION. See BLOOD PRESSURE MEASUREMENT.

HYPERTHERMIA, INTERSTITIAL

MICHAEL D. SHERAR
London Health Sciences Centre and University of Western Ontario
London, Ontario, Canada

LEE CHIN
University of Toronto
Toronto, Ontario, Canada

J. CARL KUMARADAS
Ryerson University
Toronto, Ontario, Canada

INTRODUCTION

Interstitial hyperthermia or thermal therapy is a minimally invasive method for the treatment of cancer. Radio frequency (RF), microwave, laser light, or ultrasound energy is delivered through one or more thin needle devices inserted directly into the tumor.

Interstitial devices have the significant advantage over external devices of being able to deliver thermal energy directly into the target region, thereby avoiding depositing energy into intervening nontarget tissue. Their main disadvantage is that the needle devices employed often deposit energy over only a small volume. This can make it challenging to deliver an adequate thermal dose to large target regions. This problem was highlighted in an early radiation therapy oncology group (RTOG) phase III trial in which only 1 out of 86 patients was deemed to have received an adequate thermal treatment (1).

These early challenges in interstitial hyperthermia have been addressed, to some extent, through the development of improved heating devices and more detailed monitoring of applicator placement and dose delivery. Quality assurance guidelines have been developed by the RTOG to raise the quality of heating (2). The guidelines recommend pretreatment planning and equipment checks, the implantation of considerations and documentation, the use of thermometry, and the development of safety procedures. Treatment procedures have also been improved through the use of more detailed thermometry, especially using magnetic resonance imaging approaches (3,4).

THERMAL DOSE AND HEAT TRANSFER

The goal of interstitial thermal therapy is to deliver a prescribed dose to a target volume. Thermal dose is defined as equivalent minutes at 43 °C, or TD. The units of TD are minutes, which represents the time tissue would need to be maintained at a constant temperature of 43 °C to have the same effect as the particular time–temperature history that the tissue was exposed to. The thermal dose after □ minutes of heating can be calculated if the time–temperature history is known (5),

$$\mathrm{TD}(t) = \int_0^t R^{43-T(\tau)} d\tau \qquad \text{where}$$

$$R = \begin{cases} 0.25 \text{ for } T \le 43\,^{\circ}\mathrm{C} & (1) \\ 0.5 \text{ for } T > 43\,^{\circ}\mathrm{C} & (2) \end{cases}$$

The dose prescribed for treatment depends on whether the heating is being used as an adjuvant to radiation or systemic therapy, or whether it is being used as a stand-alone treatment to coagulate tissue. For the former use, the dose prescribed is typically 10–60 min (Eq. 1) and for the latter it is usually prescribed to be > 240 min (Eq. 2). This is because temperatures employed for adjuvant treatment (usually referred to as hyperthermia) are in the 40–45 °C range. For stand-alone coagulation (usually referred to as thermal therapy or thermal ablation), temperatures in the range of 55–90 °C are used.

The temperature (T) produced in tissue depends on the heat deposition by the applicator, heat conduction, and blood flow according to

$$\rho c \frac{\partial T}{\partial t} - \nabla \cdot (k \nabla T) + \mathbf{v} \cdot \nabla T = Q$$

where ρ is the tissue mass density, ∇ is the heat capacity of the tissue, k is the thermal conductivity of the tissue, $\mathbf{v}$ is the blood velocity profile, and Q is the heat absorbed per unit volume. Detailed knowledge of the blood velocity profile at the capillary level is generally unknown, and even if it were known the calculations would require impractically large computational resources. While several models have been proposed to calculate heat transfer due to perfusion, the Pennes bioheat transfer equation is most often employed (6)

$$\rho c \frac{\partial T}{\partial t} - \nabla \cdot (k \nabla T) + w c_{\mathrm{b}} (T - T_{\mathrm{b}}) = Q$$

where w is blood mass perfusion rate, c_{b} is the blood heat capacity, and T_{b} is the temperature of the blood entering the treatment field, and $\mathbf{v}$ is the velocity field of any convective flow (e.g., as the blood in large vessels). This equation can be used to predict the temperature in tissue, and therefore plan thermal therapy or hyperthermia treatments if the perfusion rate is known. Penne's equation does not accurately predict for the effect of large blood vessels that must be modeled individually.

ELECTROMAGNETIC HEATING

The heat absorbed (or deposited) in tissue is often described in terms of the power per unit mass. It is called the specific absorption rate or SAR. For electromagnetic devices heat is deposited by the motion of charges or ions. The movement of charge depends on the electric field produced by the applicator in tissue. In microwave and RF hyperthermia, the applicators are driven by sinusoidally time-varying signals. In this case, the electric field can be written in phasor form $\mathbf{E}$ such that the electric field is given by, $E(t) = \Re(\mathbf{E}e^{j\omega t})$, where $\Re(\mathbf{x})$ is the real part of the complex vector $\mathbf{x}$, and ω is the angular frequency of the driving

signal. The SAR is then

$$\mathrm{SAR} = \frac{Q}{\rho} = \frac{\sigma}{2\rho}(\mathbf{E}\cdot\mathbf{E}^*)$$

where σ is the electrical conductivity of the tissue.

The calculation of the electric field $\mathbf{E}$ is based on Maxwell's equations. For microwave devices, these equations are combined to produce the Helmholtz vector wave equation

$$\nabla\times\nabla\times\mathbf{E} - k^2\mathbf{E} = 0$$

where k is the complex-valued wavenumber given by $k^2 = \omega^2\mu\varepsilon - j\omega\mu\sigma$ and μ is the magnetic permeability of the medium, which for tissue is the same as the free-space value, and ε is the electrical permittivity of the medium. The divergence free condition, $\nabla\cdot\mathbf{E} = 0$, may have to also be explicitly imposed if the solution technique does not inherently do this.

For RF devices, the frequency is sufficiently low that the displacement currents can be ignored. In this case, it is usually simpler to determine the scalar electric potential V and from this derive the electric field, $\mathbf{E} = -\nabla V$. The electric potential obeys a Poisson-type equation

$$-\nabla\cdot(k\nabla V) = 0$$

For models of both microwave and RF devices, the governing Helmholtz or Poisson equation is imposed in a domain with a known electric field or electric potential specified as a boundary condition to represent the power source. Another condition that is often imposed on the surface of metals is that the tangential component of the electric field is zero, $\hat{n}\times\mathbf{E} = 0$.

The solution of the governing equations with appropriate boundary conditions is impossible for all but the simplest geometries. For most practical cases, numerical methods and computational tools are required. The finite difference time domain (FDTD) method (7), the finite element (FE) method (8,9), and the volume surface integral equation (VSIE) method (10) are the most commonly utilized methods for solving the governing equations in electromagnetic hyperthermia and thermal therapy. In the FDTD method, the domain is discretized into rectangular elements. The accuracy of a FDTD solution depends on the size of the mesh spacing. Smaller elements produce more accurate solutions, but also require more memory to store the system of equations. Since the grids are rectangular, their nonconformation to curved tissue boundaries produces a stair-casing effect. Therefore, a large number of elements are required to model such geometries accurately. Unlike the FDTD method, the FE method uses tetrahedral meshes in the domain and the VSIE method uses triangular meshes on domain surfaces. Tetrahedral and triangular meshes are more suitable than regular finite difference grids for three-dimensional (3D) modeling since they do not have the stair casing effect at tissue boundaries.

RADIO FREQUENCY DEVICES

In RF, thermal therapy tissue is heated by electrical resistive (or J) heating. The heating devices, or applicators, are inserted interstitially to produce currents in the tissue. The currents typically oscillate sinusoidally in the kilohertz or low megahertz frequency range. As a result, this modality is often referred to as radio frequency or RF heating. There devices have an advantage over other interstitial devices in their simplicity and low cost. They can operate at low frequency, and therefore do not require complex power generators. The RF probes, due their simplicity, tend to have the smallest diameter of all the types of interstitial heating probes. The RF heating technique has been extensively reviewed by others (11–15).

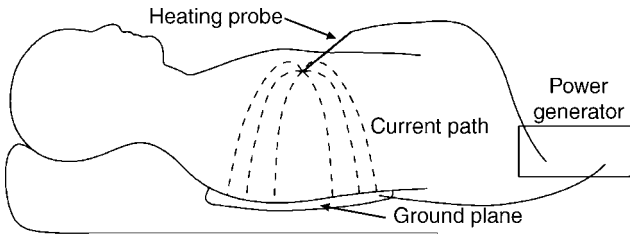


Figure 1. The heating in RF devices is caused by current flow. Since the current flows from the heating electrode to the ground pad, there is a high current density near the electrode due to its small size compared to the ground pad. This results in heating that is localized to the heating electrode.

There are several designs of RF interstitial devices, which may be categorized into three groups. The simplest design consists of a single electrode at a probe tip (often referred to as a needle electrode) (9,16–20). The current flows between a single electrode at the end of an applicator and a large ground plate placed at a distal site. Since the current flows between a small electrode and a large plate, the currents are concentrated near the electrodes resulting in SAR patterns that are localized to the electrodes as illustrated in Fig. 1.

With these single electrode probes the coagulation diameter is usually limited to $\sim$ 1.6 cm. Therefore several probes are needed to cover a larger area (21), or a single probe can be inserted into several locations, sequentially, during a treatment.

Since it is desirable to avoid the insertion of multiple interstitial probes, single probes that release multiple electrodes outward from the probe tip have been designed to produce large coagulation volumes. Two examples of these are the Boston Scientific (Watertown, MA; formerly Radio Therapeutics Corporation, Mountain View, CA) RF 3000 system in which 10–12 tines are deployed from a cannula to form an umbrella shape (Fig. 2) and the RITA Medical Systems (Mountain View, CA) Starburst probes with up to 9 tines. In some configurations, some of the tines in the Starburst probes are replaced with dedicated thermocouples while others are hollow electrodes through which saline can be infused into the target region to enhance heating. These multielectrode probes are able to produce coagulation regions with diameters up to 7 cm, although complete coverage of a large region can be difficult in high blood flow organs, such as the kidney (22).

The negative RTOG phase III trial, in which only 1 out of 86 patients was deemed to have received an adequate thermal dose (1) illustrated the need to not only increase the target volume coverage, but also to control the heating.

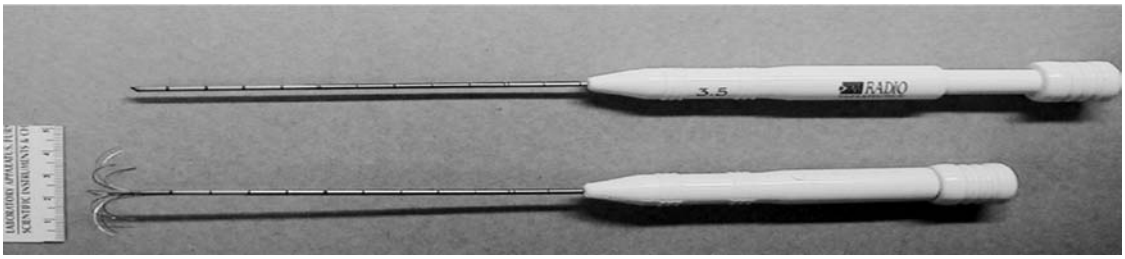


Figure 2. A Boston Scientific (Watertown, MA; formerly Radio Therapeutics Corporation, Mountain View, CA) insterstital RF probe with 10 tines that are deployed from the cannulus after insertion into a target region. The deployed tines produce a coagulation zone that is larger than the zone that can be produced by a single electrode probe. The top probe is shown with the tines undeployed (ready for insertion) and the bottom probe shows the probe with the tines deployed (as they would be after insertion).

Control is needed to enable the conformation of the heating to irregularly shaped target volumes while avoiding nearby organs at risk and to compensate for heterogeneous cooling by the vasculature (23). Partial control can be achieved by appropriate positioning of the probes and the adjustment their power. Further control along the direction of the probe is also needed (24,25) and multielectrode current source (MECS) applicators have been developed to provide this capability (26). The MECS applicators contain several electrodes placed along their length with the amplitude and phase of each electrode independently controlled. In the most common configuration, the electrodes are capacitively coupled (insulated) with the tissue. The electric fields induced by the electrodes produce currents in the tissue that cause heating. Since the electrodes are capacitively coupled, the probes can be inserted into brachytherapy catheters, for example, making it feasible to add interstitial heating as a simultaneous adjuvant to brachytherapy (interstitial radiation therapy). The electric field (and hence current) may be induced between electrodes on the same probe or on separate probes, or it may be induced between the probe electrodes and a grounding plane.

MICROWAVE DEVICES

Microwave applicators can produce larger coagulation regions than RF applicators due to their radiative nature. However, the construction of the power generator and matching circuitry makes these devices more complex, and therefore more expensive. Due to this, microwave interstitial hyperthermia has been used less often in the clinic than RF interstitial hyperthermia.

Ryan et al. reviewed and compared several types of microwave interstitial applicators (27) and several excellent reviews of microwave interstitial thermal therapy exist (28–32). The two most commonly used devices are the dipole antenna and the helical antenna. The dipole antenna is the simplest form of microwave interstitial antenna (7,8,33). It is usually constructed from a coaxial cable with the outer conductor removed from an end section (typically 1 or 2 cm in length) to expose the inner conductor (Fig. 3). A power generator feeds a sinusoidally oscillating signal into the cable at one of the ISM frequency bands between 400 MHz and 3 GHz. The inner- and outer-conductor electrodes at the tip of the coaxial cable act as an antenna that produces microwaves that radiate out into the tissue. Tissue is an attenuating medium that absorbs microwaves, and this absorbed energy is converted into heat in the tissue.

The radiative or active length of a typical dipole interstitial device is 1–3 cm. The devices produce a coagulation region that is ellipsoidal shaped with a large axis of up to 3 cm along the length of the antenna and a small axis of up to 2 cm diameter. The drawback of the dipole applicator is that the region of highest SAR, or hot spot, is located at the point at which the outer conductor is cut away. Therefore, the tips of these antennas have to be inserted past the center of the target region, and this can be a problem if the target region is located adjacent to a critical structure.

A further problem with dipole antennas is that the SAR patterns are sensitive to the depth to which the antenna is inserted into tissue (8). A second common microwave applicator design, referred to as a helical antenna (34–36), has been designed to make the applicator insensitive to its insertion depth. In this applicator, one electrode is wrapped in a helix pattern around an exposed coaxial cable (Fig. 4). The antennas are also designed to extend the heating pattern along the applicator and toward the tip of the antenna compared to the dipole antenna. The SAR pattern from a BSD Medical (Salt Lake City, UT) helical antenna is shown in (Fig. 5). The antenna was operating at 915 MHz. The measurement was performed using the thermographic imaging technique (37) and demonstrates

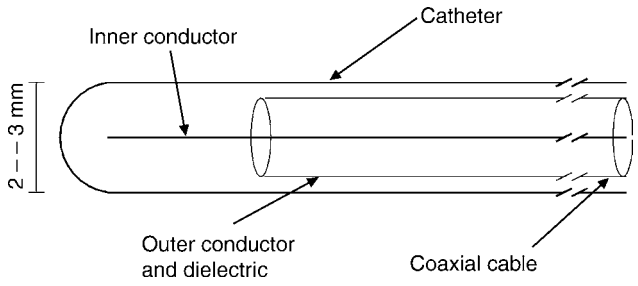


Figure 3. A schematic representation of a microwave interstitial dipole antenna applicator. The outer conductor of a coaxial cable is stripped away to produce a radiating section.

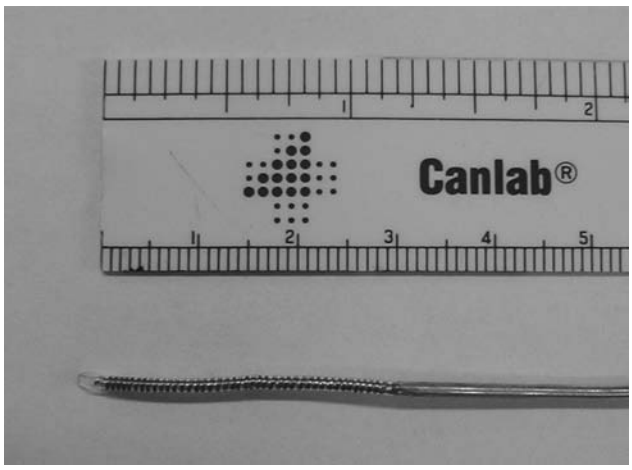


Figure 4. Shown here is a BSD Medical (Salt Lake City, UT) helical microwave applicator. The inner conductor of a coaxial cable is extended backward in a helical pattern around the dielectric insulator. There is no connection between the helical section and the outer conductor.

that the heating extends along the length of the helix and that the hot spot is close to the tip of the applicator.

Interstitial microwave applicators have the advantage over RF applicators in the ability to use arrays of applicators to dynamically steer the SAR pattern (33). For large target volumes, several applicators can be inserted. The heating pattern can then be adjusted by not only adjusting the power to each applicator, but also by adjusting the relative phase of the signal to each applicator. The phase can be adjusted such that the microwaves produced by the applicators interfere constructively in regions that require heating and interfere destructively in regions that should be spared. The predetermination of the phase required for each applicator can be calculated during treatment planning. This is a challenging calculation for applications in which tissue is electrically heterogeneous or the placement of the applicators cannot be accurately predicted. In these cases real-time monitoring of the treatment is required and a manual or computer run feedback control is used to set the phase of the applicators to produce the desired heating profile.

The size of the coagulation volume is limited by the maximum temperature in the treatment field. Since the maximum temperature is usually located at the applicator, it is possible to increase the coagulation volume by cooling adjacent to the applicator. Using this technique, the cross-section area of a coagulation volume has been noted to increase by a factor of 2.5 in one study (38) and the coagulation volume diameter was found to increase from 1.2 to 2.4 cm (39). In microwave heating, the cooling is usually done by passing water or air through the catheter containing the antenna (29,38,40). In RF heating, cooling water is passed inside the electrode to cool the tissue near the electrode (41,42).

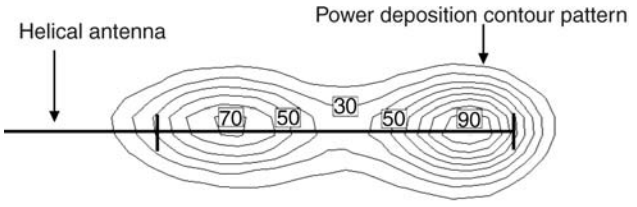


Figure 5. The normalized SAR pattern along the coronal plane of a BSD Medical (Salt Lake City, UT) helical applicator operating at 915 MHz. The image was provided courtesy of Claire McCann.

In RF heating, it is also possible to increase the coagulation volume by saline injection from the lumen of the electrode (43). Since saline is electrically conductive, injecting it into the tumor increases the electrical conductivity of the tumor, and hence the SAR in the tumor. This technique has not gained popularity due to the inability of control the flow of saline in the tumor, resulting in irregular and unpredictable coagulation regions being produced.

CLINICAL STUDIES WITH MICROWAVE AND RF DEVICES

Interstitial microwave and RF heating systems have been widely used to clinically treat a variety of tumors in phase I (34,44–47), phase II (18,32,44–49) and phase III trials (1,50). The RF systems have been used to treat a large range of sites, including brain (45), head and neck (1), breast (1), myocardium (51), lung (14), liver (11,52), pancreas (18), prostate (48), and kidney (44,53). Microwave systems have also been used to treat a large range of sites, including liver (4), prostate (both carcinoma and benign hyperplasia) (29,36), head and neck (1,32,49), brain (34,50), breast (1), and other pelvic areas (1). The heat treatments are used alone (29,36), or combined with external beam radiation (54), interstitial radiotherapy (brachytherapy) (1,32,46), and/or chemotherapy (17). The heat treatments are used alone (29,36), or combined with external beam radiation (29,36,48,54,55), combined with external beam radiation (54) or interstitial radiotherapy (brachytherapy) (1,32,46,48,55), and with chemotherapy (17).

The interstitial hyperthermia treatments are usually administered under ultrasound, CT or MR guidance. During the treatment the hyperechoginicity of microbubbles that can be produced at sufficiently high temperatures can provide some real-time ultrasound feedback of the treatment. Posttreatment evaluation can be performed using contrast enhanced ultrasound, CT or MR. The vasculature in the coagulated volume is destroyed and the destroyed volume can be identified as an unenhanced region in the image (41,56,57).

LASER DEVICES

First described in 1983 by Bown (58), Interstitial Laser Photocoagulation (ILP) [sometimes referred to as Laser Induced Thermal Therapy (LITT)] involves the use visible or near infra-red (IR) light delivered through fibre optic cables to heat tissue for therapeutic purposes. The ILP has been investigated as an experimental treatment for a variety of solid tumors including liver, breast, stomach, pancreas, kidney, lung, and bone (59). The tissue temperature is raised causing coagulation of the target volume. Similar to the microwave and RF cases, the production of

heat in a local volume of tissue results from the amount of absorbed laser energy, $S(r)$. In biomedical treatments, such as LITT, it is the total absorbed laser energy that typically determines the therapeutic outcome. It is equal to the product of the local fluence rate, $\phi(r)$ which is the total photon power over all directions that pass through a point area of space, and the absorbing characteristics, $\mu_a(r)$ of the tissue (60):

$$S(r) = \mu_a(r)\phi(r)$$

The absorbed optical energy deposition pattern is governed by the absorption and scattering characteristics of the tissue. An absorption event causes the interacting molecule to enter a vibrational–rotational state that results in a transfer of energy to surrounding molecules that manifests as a local increase in temperature (61). Absorption occurs due to interactions with native molecules called chromophores with examples including melanin, hemoglobin, and water. In a given tissue, the concentration weighted sum of the absorption of different chromophores leads to its bulk macroscopic absorption. Scattering refers to a directional change in light propagation and likely results from differences in the index of refraction in the various cellular components, such as the between cell membranes and the extracellular space. Here the scattering is assumed to be elastic with no change in energy occurring during the interaction. The statistical quantities that govern light interactions are the scattering coefficient, $\mu_a(\text{cm}^{-1})$ and absorption coefficient, $\mu_a(\text{cm}^{-1})$ and are defined, respectively, as the probability of scattering or absorption per average distance traveled (also known as the mean free path). In the case of scattering, consideration is given to the probability of scatter in a particular direction. An additional parameter known as the anisotropy factor, g, quantifies this directionality by integrating the average cosine of the scattering probability over all directions. When $g = 0$, scattering is isotropic. However, in the case of biological tissues g typically lies within the range of 0.7 and 0.99 meaning that scattering typically occurs in the forward direction. The reduced scattering coefficient, $\mu_s' = \mu_s(1-g)$, allows light scattering to be approximated as isotropic although scattering events are actually in the forward direction. The inverse of the reduced scattering coefficient is, therefore, the average distance that light travels before it changes direction from its original direction of propagation (62).

In theory, Maxwell's equations could be used to calculate the scattering and absorption of the EM vector fields due to the underlying tissue components (63). In this case, the tissue microstructure could be modeled as random perturbations, $\varepsilon_1(r)$ in the dielectric constant around a mean value, $\varepsilon_0(r)$, with the total dielectric constant, $\varepsilon(r)$, given by the sum of these quantities. However, in practice, due to the complex and random composition of tissue, a complete and accurate description of $\varepsilon(r)$ has yet to be realized. Instead a more commonly used solution is to consider light as a stream of neutral particles or photons with individual quanta of energy that propagate elastically throughout the medium. This formalism is governed by radiative transport theory (64), and assumes light to be monochromatic while ignoring its conventional wave characteristics, such as polarization, diffraction, interference, and fluorescence. Although incomplete, the photon model has been shown to be consistent with experimental measurements in turbid media (65).

A commonly employed model of photon propagation is the Monte Carlo (MC) method (66), which utilizes probability distributions to simulate the propagation of thousands to millions of individual photon packets based on the optical properties of tissue to arrive at a statistical representation of the overall light distribution. The MC is amenable to heterogeneous and arbitrary geometries and does not suffer from the limiting assumptions of analytical solutions. However, its primary disadvantage is the requirement of long computational times, on the order of hours to days, to achieve reasonable statistics. Regardless, with the increasing speed of modern computers, the Monte Carlo method remains a viable option for photon simulations. The reader is referred to an excellent review by Roggan and Muller (67) for the implementation of the MC model for treatment planning of LITT.

Alternatively, one may employ formal solutions to the governing equations for photon transport. The energy flow of photons in a scattering and absorbing medium is described by the radiative transfer equation (RTE) (64). The RTE is an integro differential equation that describes the energy conservation of photons within an infinitesimally small volume that result from losses due to absorption and scattering as well as gains arising from photons scattered from other directions and from the laser source. Analytical solutions to the RTE are difficult to obtain. Hence, various approximations have been proposed to convert the RTE to a more mathematically tractable and practical form. A standard technique, called the P_n approximation, expands the radiance and source as a finite series of spherical harmonics to nth order. The $P1$ approximation is the simplest of these expansions and in the steady state is also known as the diffusion approximation (63,64):

$$\nabla^2\phi(\vec{r}) - \frac{\mu_a}{D}(\vec{r})\phi(\vec{r}) = -\frac{1}{D}S(\vec{r})$$

Here $\phi(\vec{r})$ is the photon fluence rate, while D is the photon diffusion coefficient given by

$$D = \frac{1}{3[\mu_s' + \mu_a]}$$

The primary assumption of the diffusion equation, that is linear flux anisotropy, is only accurate when the scattering properties of the medium are much larger than the absorption properties and at locations $> 1/\mu_s'$ from the source. A number of analytical solutions to the diffusion equation exist for simple but practical geometries. The solution for a point source in an infinite homogeneous medium is given by (63)

$$\phi(\vec{r}) = \frac{P_0 e^{(-\mu_{\text{eff}} r)}}{4\pi r}$$

This solution is particularly useful as, assuming an infinite medium, it may be integrated numerically to provide the light distribution of cylindrical or extended source

of arbitrary geometries. However, it is well known that tissue optical properties often change from their native state after undergoing thermal coagulation. This results in heterogeneities in optical properties that effect the overall light distribution (68). In such cases, analytical solutions are available only for the simplest geometries and numerical methods such as the finite element (69), finite difference (70), and boundary element method (71) must be employed. A thorough discussion of these methods was given in the preceding section for microwaves and their implementation in the case of photon propagation is the same.

Initially, bare tipped optical fibers were used to deliver laser light to the tumor. High temperatures immediately adjacent to the fiber tip cause the tissue to char and form a zone of carbonization. The charred fiber then acts as a point heat source and the temperature of the fiber increases significantly leading to vacuolization of the surrounding tissue. The volume of coagulation around the fiber grows until thermal equilibrium is reached at the edges of the lesion. Here, the conduction of heat from the fiber is balanced by the tissue's ability to remove energy through blood flow and thermal conduction.

The size of the lesion depends on the thermal conduction properties of the tissue, but would normally be limited to $\sim$ 2 cm in diameter. Larger tumors require multiple optical fiber implants to enable complete coverage of the tumor volume. For example, a 4 cm diameter tumor would require at least eight fibers to fully coagulate the tumor.

The limitations of the bare tipped fibers have been addressed in two ways. The first was to employ a line source geometry instead of a point source. This can be achieved by using a diffusing tip fiber where light gradually leaks out of the fiber over an extended distance of a few centimeters. The second approach is to restrict the temperature of the fiber to lower than the charring threshold by controlling the power delivered to the fiber. If charring is avoided, light can propagate into the tissue resulting in heating at a distance from the fiber and a broader SAR pattern. These two approaches can be combined to achieve greater lesion volumes from single fibers. Heisterkamp et al. (72) demonstrated an almost doubling of the coagulated volume from 4.32 cm^3 (bare tipped) to 8.16 cm^3 (temperature restricted diffusing tip) using such an approach.

The other major factor that affects the lesion size is the wavelength of the light used. Somewhat counterintuitively, light that is less absorbed by tissue, results in greater lesion sizes. This is because the light can penetrate further into the tissue, and therefore directly heat at greater distances from the fiber. The availability of high power sources at two specific wavelengths (810 nm as produced by diode lasers and 1064 nm as produced by Nd:YAG lasers) has dominated the development of interstitial laser thermal therapy. Wyman et al. (73) have shown that 1064 nm light can enable the creation of greater lesion sizes due to its greater penetration. However, Nd:YAG lasers are large, generally immobile and inconvenient and so many have adopted 810 nm as the wavelength of choice due to the availability of compact and inexpensive sources. More recently 980 nm lasers have been employed to combine mobility with greater light penetration (74,75).

Differences between Nd:YAG and Diode lasers are only realized if charring is avoided. Once charring and carbonization has occurred the fiber acts as a point or line heat source. There is no further light propagation into the tissue and subsequent heating has no wavelength dependency. In order to exploit the penetration of light into the tissue, the fiber tip temperature must be controlled to avoid charring. Achieving such control is somewhat challenging as the temperature of the tip can rise very quickly in a positive feedback loop. As charring begins, the rate of temperature rise increases that causes an increasing rate of charring. Robust, automatic feedback control mechanisms are necessary to ensure controlled heating and lesion formation.

INTERSTITIAL ULTRASOUND

The possibility of developing interstitial ultrasound devices for hyperthermia applications was proposed by Hynynen in 1992 (76). The initial studies examined various design parameters including the choice of ultrasound frequency, electric and acoustic power, and catheter cooling. As Hynynen showed (76), thin interstitial ultrasound applicators were likely capable of heating perfused tissue to therapeutic temperatures.

Ultrasound is a high frequency longitudinal pressure wave that can pass relatively easily through soft tissue. Consequently, it has been useful as an energy source for diagnostic imaging where focussed ultrasound radiators are used to produce high resolution images of soft tissue abnormalities. During transmission through tissue energy is lost due to absorption and to a much lesser extent to scattering. The absorption is caused by friction as the pressure wave causes relative motion of the tissue components. These frictional forces cause heating that can be significant if the incident ultrasound power is high enough. The absorption, α is frequency dependent where

$$\alpha = a\, f^m$$

and a and m are coefficients that are variable between tissues although m is $\sim$ 1.5 for most soft tissues. Rapidly increasing absorption with frequency is the main reason that the penetration of diagnostic imaging is limited at very high ultrasound frequencies. Higher penetration is also the reason that relatively low ultrasound frequencies are used for ultrasound heating. Typically, frequencies in the range 0.5–2 MHz have been used in external focused ultrasound heating applications. However, this becomes problematic for interstitial devices that are small and resonate at high ultrasound frequencies.

Interstitial ultrasound applicators have since been developed and are usually designed as thin tubular radiators. The radiator consists of a piezoelectric material that will resonate acoustically at a frequency f determined by the wall diameter d:

$$f = \frac{v}{2d}$$

where v is the speed of sound in the piezoelectric material (e. g., 4000 $m{\cdot}s^{-1}$ in the piezoelectric material PZT 4A). For interstitial applicators, thin radiators are required. A wall

thickness of 0.2 mm, for example, would translate into an operating frequency of $\sim$ 10 MHz (76). The SAR for a cylindrical applicator is dependent on its dimensions and the frequency of operation as given by

$$\mathrm{SAR} = 2\alpha\, f I_0 \left(\frac{r}{r_0}\right) \mathrm{e}^{-2\mu\, f(r-r_0)}$$

where α is the ultrasound absorption coefficient in tissue, I_0 is the intensity of ultrasound at the applicator surface, r_0 is the radius of the applicator, r is the distance from the centre of the applicator to the point of interest and μ is the attenuation coefficient of ultrasound that includes absorption and scattering. Skinner et al. (77) have calculated and compared the SAR of ultrasound, laser, and microwave applicators assuming a simple cyclindrical radiation pattern for each. The SAR of all these applicators is dominated by the thin cylindrical geometry so that despite the larger penetration depth of ultrasound, only slightly larger diameter lesions can be produced. In order to overcome the limiting geometry, new interstitial ultrasound applicators have been developed that take advantage of the focusing ability of ultrasound (78) or that employs acoustic matching that can result in efficient transmission at multiple frequencies (79).

The development of interstitial ultrasound applicators is still at the preclinical stage (80,81) although larger, intracavitary applicators are being applied in the treatment of prostate cancer using a transrectal technique (82).

BIBLIOGRAPHY

1. Emami BC, et al. Phase III study of interstitial thermoradiotherapy compared with interstitial radiotherapy alone in the treatment of recurrent or persistent human tumors: A prospectively controlled randomized study by the Radiation Therapy Oncology Group. Inte J Rad Oncol Biol Phy 1996;34(5): 1097–1104.
2. Emami BP, et al. RTOG Quality Assurance Guidelines for Interstitial Hyperthermia. Inter J Rad Oncol Biol Phy 1991; 20(5):1117–1124.
3. Peters RD, et al. Magnetic resonance thermometry for predicting thermal damage: An application of interstitial laser coagulation in an in vivo canine prostate model. Magn Reson Med 2000;44(6):873–883.
4. Morikawa S, et al. MR-guided microwave thermocoagulation therapy of liver tumors: Initial clinical experiences using a 0.5 T open MR system. J Magn Reson Imaging 2002;16(5):576–583.
5. Sapareto SA, Dewey WC. Thermal dose determination in cancer therapy. Int J Radiat Oncol Biol Phys 1984;10(6): 787–800.
6. Pennes HH .Analysis of tissue and arterial blood temperatures in the resting human forearm. 1948. J Appl Physiol 1998; 85(1):5–34.
7. Camart JC, et al. New 434 MHz interstitial hyperthermia system monitored by microwave radiometry: theoretical and experimental results. Intern J Hypertherm 2000;16(2):95–111.
8. Mechling JA, Strohbehn JW. 3-Dimensional Theoretical SAR and Temperature Distributions Created in Brain-Tissue by 915 and 2450 MHz Dipole Antenna-Arrays with Varying Insertion Depths. Intern J Hypertherm 1992;8(4):529–542.
9. Uzuka T, et al. Planning of hyperthermic treatment for malignant glioma using computer simulation. Int J Hypertherm 2001;17(2):114–122.
10. Wust P, et al. Simulation studies promote technological development of radiofrequency phased array hyperthermia. Int J Hypertherm 1996;12(4):477–494.
11. Haemmerich D, Lee Jr FT. Multiple applicator approaches for radiofrequency and microwave ablation. Int J Hypertherm 2005;21(2):93–106.
12. McGahan JP, Dodd GD, 3rd. Radiofrequency ablation of the liver: current status. AJR Am J Roentgenol 2001;176(1):3–16.
13. Friedman MI, et al. Radiofrequency ablation of cancer. Cardiovasc Intervent Radiol 2004;27(5):427–434.
14. Lencioni RL, et al. Radiofrequency ablation of lung malignancies: where do we stand? Cardiovasc Intervent Radiol 2004;27(6): 581–590.
15. Gazelle GS, Goldberg SN, Solbiati L, Livraghi T. Tumor ablation with radio-frequency energy. Radiology 2000;217(3): 633–646.
16. Goletti O, et al. Laparoscopic radiofrequency thermal ablation of hepatocarcinoma: preliminary experience. Surg Laparosc Endosc Percutan Tech 2000;10(5):284–290.
17. Morita K, et al. Combination therapy of rat brain tumours using localized interstitial hyperthermia and intra-arterial chemotherapy. Inter J Hypertherm 2003;19(2):204–212.
18. Matsui Y, et al. Selective thermocoagulation of unresectable pancreatic cancers by using radiofrequency capacitive heating. Pancreas 2000;20(1):14–20.
19. Aoki H, et al. Therapeutic efficacy of targeting chemotherapy using local hyperthermia and thermosensitive liposome: evaluation of drug distribution in a rat glioma model. Int J Hypertherm 2004;20(6):595–605.
20. Lencioni R, et al. Radio-frequency thermal ablation of liver metastases with a cooled-tip electrode needle: results of a pilot clinical trial. Eur Radiol 1998;8(7):1205–1211.
21. Haemmerich D, et al. Large-volume radiofrequency ablation of ex vivo bovine liver with multiple cooled cluster electrodes. Radiology 2005;234(2):563–568.
22. Rendon RA, et al. The uncertainty of radio frequency treatment of renal cell carcinoma: Findings at immediate and delayed nephrectomy. J Urol 2002;167(4):1587–1592.
23. Crezee J, Lagendijk JJ. Temperature uniformity during hyperthermia: the impact of large vessels. Phys Med Biol 1992;37(6):1321–1337.
24. vanderKoijk JF, et al. Dose uniformity in MECS interstitial hyperthermia: The impact of longitudinal control in model anatomies. Phys Med Biol 1996;41(3):429–444.
25. VanderKoijk JF, et al. The influence of vasculature on temperature distributions in MECS interstitial hyperthermia: Importance of longitudinal control. Intern J Hypertherm 1997; 13(4):365–385.
26. Kaatee RSJP. Development and evaluation of a 27 MHz multielectrode current-source interstitial hyperthermia system. Med Phys 2000;27(12):2829–2829.
27. Ryan TP. Comparison of 6 Microwave Antennas for Hyperthermia Treatment of Cancer—SAR Results for Single Antennas and Arrays. Intern J Rad Oncol Biol Phys 1991; 21(2):403–413.
28. Roemer RB. Engineering aspects of hyperthermia therapy. Annu Rev Biomed Eng 1999;1:347–376.
29. Sherar MD, Trachtenberg J, Davidson SRH, Gertner MR. Interstitial microwave thermal therapy and its application to the treatment of recurrent prostate cancer. Intern J Hypertherm 2004;20(7):757–768.
30. Fabre JJ, et al. 915 MHz Microwave Interstitial Hyperthermia. 1. Theoretical and Experimental Aspects with Temperature Control by Multifrequency Radiometry. Intern J Hypertherm 1993;9(3):433–444.
31. Camart JC, et al. 915 MHz Microwave Interstitial Hyperthermia. 2. Array of Phase-Monitored Antennas. Intern J Hypertherm 1993;9(3):445–454.

32. Prevost B, et al. 915 MHz Microwave Interstitial Hyperthermia. 3. Phase-II Clinical-Results. Intern J Hypertherm 1993; 9(3):455–462.
33. Camart JC, et al. Coaxial Antenna-Array for 915 MHz Interstitial Hyperthermia—Design and Modelization Power Deposition and Heating Pattern Phased-Array. IEEE Trans Microwave Theory Tech 1992;40(12):2243–2250.
34. Fike JR, Gobbel GT, Satoh T, Stauffer PR. Normal Brain Response after Interstitial Microwave Hyperthermia. Intern J Hypertherm 1991;7(5): 795–808.
35. McCann C, et al. Feasibility of salvage interstitial microwave thermal therapy for prostate carcinoma following failed brachytherapy: studies in a tissue equivalent phantom. Phys Med Biol 2003;48(8):1041–1052.
36. Sherar MD, et al. Interstitial microwave thermal therapy for prostate cancer. J Endourol 2003;17(8):617–625.
37. Guy A. Analysis of Electromagnetic Fields Induced in Biological Tissues by Thermographic Studies on Equivalent Phanton Models. IEEE Trans Biomed Eng 1971;19:205–214.
38. Trembly BS, Douple EB, Hoopes PJ. The Effect of Air Cooling on the Radial Temperature Distribution of a Single Microwave Hyperthermia Antenna In vivo. Intern J Hypertherm 1991;7(2):343–354.
39. Goldberg SN, et al. Radiofrequency tissue ablation: increased lesion diameter with a perfusion electrode. Acad Radiol 1996;3(8):636–644.
40. Eppert V, Trembly BS, Richter HJ. Air Cooling for an Interstitial Microwave Hyperthermia Antenna—Theory and Experiment. IEEE Trans Biomed Eng 1991;38(5):450–460.
41. Goldberg SN, et al. Treatment of intrahepatic malignancy with radiofrequency ablation: Radiologic-pathologic correlation. Cancer 2000;88(11):2452–2463.
42. Solbiati L, et al. Hepatic metastases: percutaneous radiofrequency ablation with cooled-tip electrodes. Radiology 1997; 205(2):367–373.
43. Livraghi T, et al. Saline-enhanced radio-frequency tissue ablation in the treatment of liver metastases. Radiology 1997;202(1):205–210.
44. Michaels MJ, et al. Incomplete renal tumor destruction using radio frequency interstitial ablation. J Urol 2002;168(6):2406–2409.
45. Takahashi H, et al. Radiofrequency interstitial hyperthermia of malignant brain tumors: Development of heating system. Exper Oncol 2000;22(4):186–190.
46. Seegenschmiedt MH, et al. Clinical-Experience with Interstitial Thermoradiotherapy for Localized Implantable Pelvic Tumors. Am J Clin Oncol Cancer Clin Trials 1993;16(3): 210–222.
47. Seegenschmiedt MH, et al. Multivariate-Analysis of Prognostic Parameters Using Interstitial Thermoradiotherapy (Iht-Irt)—Tumor and Treatment Variables Predict Outcome. Intern J Rad Oncol Biol Phys 1994;29(5):1049–1063.
48. van Vulpen M, et al. Radiotherapy and hyperthermia in the treatment of patients with locally advanced prostate cancer: Preliminary results. Bju Inter 2004;93(1):36–41.
49. Engin K, et al. Thermoradiotherapy with Combined Interstitial and External Hyperthermia in Advanced Tumors in the Head and Neck with Depth Greater-Than-or-Equal-to 3 Cm. Intern J Hyperther 1993;9(5):645–654.
50. Sneed PK, et al. Thermoradiotherapy of Recurrent Malignant Brain-Tumors. Int J Radiat Oncol Biology Physics. 1992.
51. Wonnell TL, Stauffer PR, Langberg JJ. Evaluation of Microwave and Radio-Frequency Catheter Ablation in a Myocardium-Equivalent Phantom Model. IEEE Trans Biomed Eng 1992;39(10):1086–1095.
52. Buscarini L, Buscarini E. Therapy of HCC-radiofrequency ablation. Hepato-Gastroenterol 2001;48(37):15–19.
53. Rendon RA, et al. Development of a radiofrequency based thermal therapy technique in an in vivo porcine model for the treatment of small renal masses. J Urol 2001;166(1):292–298.
54. Blute ML, Larson T. Minimally invasive therapies for benign prostatic hyperplasia. Urology 2001;58(6A):33–40.
55. Van Vulpen M, et al. Three-dimensional controlled interstitial hyperthermia combined with radiotherapy for locally advanced prostate carcinoma—A feasibility study. Intern J Rad Oncol Biol Phy 2002;53(1):116–126.
56. Belfiore G, et al. CT-guided radiofrequency ablation: a potential complementary therapy for patients with unresectable primary lung cancer—a preliminary report of 33 patients. AJR Am J Roentgenol 2004;183(4):1003–1011.
57. Cioni D, Lencioni R, Bartolozzi C, Percutaneous ablation of liver malignancies: Imaging evaluation of treatment response. Eur J Ultrasound 2001;13(2):73–93.
58. Bown SG. Phototherapy in tumors. World J Surg 1983;7(6): 700–709.
59. Witt JD, et al. Interstitial laser photocoagulation for the treatment of osteoid osteoma. J Bone Joint Surg Br 2000; 82(8):1125–1128.
60. Welch A. The thermal response of laser irradiated tissue. IEEE J Quantum Electr 1984;20(12):1471–1481.
61. Boulnois J. Photophysical processes in recent medical laser developments: A review. Lasers Med Sci 1986;1(1):47–66.
62. Wyman D, Patterson M, Wilson B. Similarity relations for the interaction parameters in radiation transport and their applications. Appl Op 1989;28:5243–5249.
63. Ishimaru A. Diffusion Approximation, in Wave Propagation and Scattering in Random Media. New York: Academic Press; 1978. p. 178.
64. Duderstadt JH. Nuclear Reactor Analysis. New York: John Wiley & Sons; 1976.
65. Rinzema K, Murrer L. Direct experimental verification of light transport theory in an optical phantom. J Opt Soc Am A 1998;15(8):2078–2088.
66. Wilson BC, Adam G. A Monte Carlo model for the absorption and flux distributions of light in tissue. Med Phys 1983; 10(6):824–830.
67. Roggan A, Muller G. Dosimetry and computer based irradiation planning for laser-induced interstitial thermotherapy (LITT). In: Roggan A, Muller G, editors. Laser-Induced Interstitial Thermotherapy. Bellingham, (WA): SPIE Press; 114–156.
68. Jaywant S, et al. Temperature dependent changes in the optical absorptio nand scattering spectra of tissue. SPIE Proc 1882. 1993;
69. Arridge SR, Schweiger M, Hiraoka M, Delpy DT. A finite element approach for modeling photon transport in tissue. Med Phys 1993;20(2 Pt. 1):299–309.
70. Pogue BW, Patterson MS, Jiang H, Paulsen KD. Initial assessment of a simple system for frequency domain diffuse optical tomography. Phys Med Biol 1995;40(10):1709–1729.
71. Ripoll J, Nieto-Vesperinas M. Scattering Integral Equations for Diffusive Waves. Detection of Objects Buried in Diffusive Media in the Presence of Rough Interfaces. J Opt Soc of Am A 1999;16:1453–1465.
72. Heisterkamp J, van Hillegersberg R, Sinofsky E. Heat-resistant cylindrical diffuser for interstitial laser coagulation: Comparison with the bare-tip fiber in a porcine liver model. Lasers Surg Med 1997;20(3):304–309.
73. Wyman DR, Schatz SW, Maguire JA. Comparison of 810 nm and 1064 nm wavelengths for interstitial laser photocoagulation in rabbit brain. Lasers Surg Med 1997;21(1):50–58.
74. McNichols RJ, et al. MR thermometry-based feedback control of laser interstitial thermal therapy at 980 nm. Lasers Surg Med 2004;34(1):48–55.

75. Kangasniemi M, et al. Thermal therapy of canine cerebral tumors using a 980 nm diode laser with MR temperature-sensitive imaging feedback. Lasers Surg Med 2004;35(1):41–50.
76. Hynynen K. The Feasibility of Interstitial Ultrasound Hyperthermia. Med Phys 1992;19(4):979–987.
77. Skinner MG, Iizuka MN, Kolios MC, Sherar MD. A theoretical comparison of energy sources—microwave, ultrasound and laser—for interstitial thermal therapy. Phys Med Biol 1998; 43(12):3535–3547.
78. Chopra R, Bronskill MJ, Foster FS. Feasibility of linear arrays for interstitial ultrasound thermal therapy. Med Phys 2000; 27(6):1281–1286.
79. Chopra R, Luginbuhl C, Foster FS, Bronskill MJ. Multi-frequency ultrasound transducers for conformal interstitial thermal therapy. IEEE Trans Ultrason Ferroelectr Freq Control 2003;50(7):881–889.
80. Nau WH, et al. MRI-guided interstitial ultrasound thermal therapy of the prostate: A feasibility study in the canine model. Med Phys 2005;32(3):733–743.
81. Diederich CJ, et al. Catheter-based ultrasound applicators for selective thermal ablation: Progress towards MRI-guided applications in prostate. Int J Hyperther 2004;20(7):739–756.
82. Uchida T, et al. Transrectal high-intensity focused ultrasound for treatment of patients with stage T1b-2NOMO localized prostate cancer: A preliminary report. Urology 2002; 59(3): 394–398.

See also BRACHYTHERAPY, HIGH DOSAGE RATE; HEAT AND COLD, THERAPEUTIC; HYPERTHERMIA, SYSTEMIC; HYPERTHERMIA, ULTRASONIC; PROSTATE SEED IMPLANTS.

HYPERTHERMIA, SYSTEMIC

R. WANDA ROWE-HORWEGE
University of Texas Medical School
Houston, Texas

INTRODUCTION

Systemic hyperthermia is deliberate heating of the whole body to achieve an elevated core temperature for therapeutic purposes. Other terms used are whole-body hyperthermia, systemic or whole body thermal therapy, and hyperpyrexia. The goal of systemic hyperthermia is to reproduce the beneficial effects of fever. Typically, core body temperatures of 41–42 °C are induced for 1–2 h, or alternatively 39–40 °C for 4–8 h. Systemic hyperthermia, by virtue of application to the whole body, aims to alleviate systemic disease conditions, in contrast to local or regional hyperthermia that treats only a specific tissue, limb, or body region.

HISTORICAL BACKGROUND

The use of heat to treat disease goes back to ancient times. Application of fire to cure a breast tumor is recorded in an ancient Egyptian papyrus, and the therapeutic value of elevated body temperature in the form of fever was appreciated by ancient Greek physicians. Hippocrates wrote, "What medicines do not heal, the lance will; what the lance does not heal, fire will," while Parmenides stated, "Give me a chance to create a fever and I will cure any disease." In the first century AD, Rufus (also written as Refus or Ruphos) of Ephesus advocated fever therapy for a variety of diseases. Hot baths were considered therapeutic in ancient Egypt, Greece, Rome, China, and India as they still are in many aboriginal cultures today, along with burying diseased individuals in hot sand or mud. Hot baths and saunas are an integral part of health traditions throughout the Orient, in Indian Ayurvedic medicine, as well as in Eastern European and Scandinavian countries. Following several earlier anecdotal reports, several nineteenth century German physicians observed regression or cure of sarcoma in patients who suffered prolonged, high fevers due to infectious diseases. This led to efforts to induce infectious fevers in cancer patients, for example, by applying soiled bandages or the blood of malaria patients to wounds. The late nineteenth century New York physician, William Coley, achieved cancer cures by administration of erysipelas and other bacterial endotoxins, now known as Coley's toxins, and attempted to create standardized preparations of these pyrogens (1). At around the same time, treatment of syphilis by placing the patient in a stove-heated room, or a heat box, became commonplace. Successful hyperthermic treatment of other sexually transmitted diseases, such as gonorrhea, and neurological conditions, such as chorea minor, dementia paralytica, and multiple sclerosis along with arthritis, and asthma were widely reported. Interestingly, it was noted by Italian physicians that upon completion of the draining of the Pontine Swamps near Rome by Mussolini in the 1930s, not only was malaria eradicated, but the prevalence of cancer in the area was the same as in the rest of Italy, whereas earlier the whole malaria-infected region was noted for its absence of cancer. It was concluded that the frequent fever attacks common in malaria stimulated the immune system to prevent the development of cancers.

The science of hyperthermia became grounded in the first few decades of the twentieth century when some of the biological effects of elevated body temperature were elucidated and attempts were made to understand and control the therapeutic application of heat. Numerous devices were developed to produce elevated temperatures of the body, by a variety of physical means. After a shift in focus to local and regional hyperthermia, there is now a resurgence of interest in systemic hyperthermia for treatment of cancer, as well as other systemic diseases. Whole-body hyperthermia treatment is now carried out at several university centers in the United States, and Europe (Table 1), where controlled clinical trials are being carried out. Numerous private clinics, principally in North America, Germany, Austria, Eastern Europe, Japan, and China also perform systemic hyperthermia, mostly as part of holistic, alternative, treatment regimens.

PHYSICS OF SYSTEMIC HYPERTHERMIA

As shown schematically in Fig. 1, in order to achieve body temperature elevation, there must be greater deposition of heat energy in the body than heat energy lost from

Table 1. Clinical Academic/Regional Systemic Hyperthermia Centers

Country	City	Institution	Principal Investigator	Heat Type, Machine[a]	Protocol (time, temp)
Asia					
China	Baoding	Second Hospital of Baoding	Chunzhu Yin	RF	3.5–6 h, 40–40.5 °C
China	Changchun	Jilin Tumor Hospital	Changguo Hong	RF	3.5–6 h, 40–40.5 °C
China	Jiangmen	Guangdong Jiangmen Renmin Hospital	Wenping Wu	IR	
China	Shanghai	Changhai Hospital, Second Military Medical School	Yajie Wang		
China	Shanghai	Department of Tumor Hyperthermia Center	Kai-sheng Hou	IR, ET-Space extracorporeal	1–2 h, 41.8–42.5 °C
China	Shanghai	Shanghai Jingan Central Hospital	Weiping Tao	IR, ET-Space extracorporeal	2 h, 41.6 °C 4 h, 42.1 °C
China	Tai'an City	88th Hospital of PLA	Yong Peng		1–2 h, 41.8 °C 6 h, 39.5–40 °C
China	Zhengzhou	Modern Hospital, Zhengzhou	Dingjiu Li	RF	3.5–6 h, 40–40.5 °C
Japan	Tokyo	Luke Hospital	Akira Takeuchi	IR	
Europe					
Belarus	Minsk	Belarus Center for Pediatric Oncology and Hematology	Reimann Ismail-zade	HF EM, Yakhta-5	2 h, 41.8–42.5 °C 1 h, 42.5–43 °C
Germany	Berlin	Ludwig Maximilian University Charité Medical Center	Bert Hildebrandt, Hanno Riess, Peter Wust	IR, Iratherm	1 h, 41.8 °C
Germany	Frankfurt	Krankenhaus Nordwest	Elke Jäger, Akin Atmata	IR, Aquatherm	1 h, 41.8 °C
Germany	Munich	Ludwig Maximilian University Hospital Clinic	Harald Sommer	IR, Iratherm	1 h, 41.8 °C
Hungary	Kecskemét	Institute of Radiology of Kecskemét	Miklós Szücs	IR, OncoTherm	1 h, 41.8 °C
Norway	Bergen	University of Bergen, Haukeland University Hospital	Baard-Christian Schem	IR, Iratherm	1 h, 41.8 °C
Russia	Novosibirsk	Siberian Scientific Research Institute of Hyperthermia	Roman Tchervov	Water bath	43.5–44.0 °C
Russia	Obninsk	Medical Radiological Research Center of Russian Academy of Medical Sciences, Obninsk	Yuri Mardynsky	HF EM, Yakhta 5	1–2 h, 41.0–42.3 °C
North America					
United States	Galveston, TX	University of Texas Medical Branch	Joseph Zwischenberger	extracorporeal	2 h, 42.5 °C
United States	Houston, TX	University of Texas Medical School	Joan M. Bull	IR, Heckel	6 h, 40 °C
United States	Buffalo, NY	Roswell Park Cancer Institute	William G. Kraybill	IR, Heckel	6 h, 40 °C
United States	Durham, NC	Duke Comprehensive Cancer Center[b]	Zeljko Vujaskovic	IR, Heckel	6 h, 40 °C

[a]Radio frequency = RF; infrared = IR.
[b]Starting in 2006.

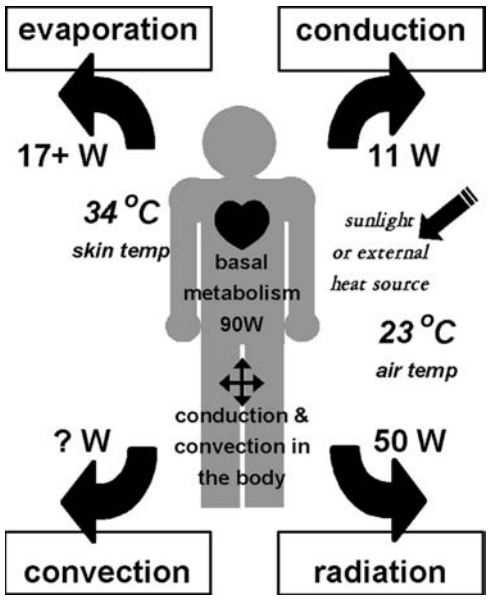


Figure 1. Schematic of heat balance mechanisms in the human body. Body temperature is determined by the balance of metabolic heat production plus heating from external sources, and heat losses by radiation, evaporation, convection, and conduction.

conduction, convection, radiation and evaporation, that is,

$$Q'_{\text{dep}}\Delta t > Q'_{\text{loss}}\Delta t \tag{1}$$

where $Q' = \Delta Q/\Delta t$ represents the change in heat energy, Q (measured in Joules or calories), over a time period Δt. Net heat energy deposition in a volume element ΔV of tissue of density ρ_{tis} results in an increase in temperature ΔT dependent on the specific heat of the tissue, c_{tis},

$$\left(\frac{Q'_{\text{dep}}}{\Delta V} - \frac{Q'_{\text{loss}}}{\Delta V}\right)\Delta t = (\rho_{\text{tis}}\Delta V)c_{\text{tis}}\Delta T$$

$$\Delta T = (Q'_{\text{dep}} - Q'_{\text{loss}}) \cdot \frac{\Delta t}{\rho_{\text{tis}}c_{\text{tis}}} \tag{2}$$

Heat deposition is the sum of the absorbed power density, P_{abs}, from external sources and heat generated by metabolism, Q_{met},

$$\frac{\Delta Q'_{\text{dep}}}{\Delta V} = P_{\text{abs}}(x,y,z,t) + \frac{\Delta Q'_{\text{met}}}{\Delta V}(x,y,z,t) \tag{3}$$

If the air temperature is higher than the body surface temperature, heat is absorbed from air surrounding the body by the skin, as well as during respiration. Power deposition in tissue from external electromagnetic fields depends on the coupling of the radiation field (microwave, RF, ultrasound, visible or IR light) with tissue. The body's metabolic rate depends on the amount of muscular activity, the temperature, pressure and humidity of the environment, and the size of the body. Metabolic rate increases nonlinearly with core body temperature, in part due to the exponential increase of the rate of chemical reactions with temperature (Arrhenius equation). An empirical relationship between basal metabolic rate and core temperature has been determined as

$$\text{Basal MR} = \frac{85 \times 1.07^{(T_{\text{core}})}}{0.5} \tag{4}$$

which can be exploited to maintain elevated body temperatures (2). At room temperature a human body produces $\sim$ 84 W, which increases to $\sim$162 W at a core temperature of 41.8 °C.

Heat losses from the body are often termed sensible (convective, conductive, radiative) and insensible (evaporative, latent). The primary mode of heat loss from the body is by radiation, as described by the Stefan–Boltzmann law,

$$\frac{Q'_{\text{rad}}}{\Delta V} = e_{\text{skin}}\sigma A_{\text{skin}}(T_{\text{skin}} - T_s)^4 \tag{5}$$

where $Q'_{\text{rad}}/\Delta V$ is the power radiated, e_{skin} is the emissivity of the skin (radiating material), σ is Stefan's constant = 5.6703×10^{-8} W·m^{-2}/K, A_{skin} is the skin surface area, T_{skin} is the temperature of the skin (radiator), and T_s is the temperature of the surroundings (e.g., air, water, wax). Human skin is a near perfect radiator in the IR, with an emissivity of 0.97. At room temperature, >50% of the heat generated by metabolism is lost by radiation; a clothed adult loses some 50 W at room temperature. This increases to $\sim$66% at a core temperature of 41.8 °C, as is targeted in a number of systemic hyperthermia protocols, when the skin temperature rises to 39–40 °C (3).

Direct transfer of body heat to the molecules around the body (typically air) occurs by conduction, or molecular agitation within a material without any motion of the material as a whole, which is described by Fourier's law,

$$\frac{\Delta Q'_{\text{cond}}}{\Delta V} = \kappa A_{\text{skin}}\frac{\Delta T}{\Delta x} \tag{6}$$

where ΔQ_{cond} is the heat energy transferred per unit volume in time Δt, κ is the thermal conductivity (W·mK^{-1}) of the material surrounding the body (air, water), and ΔT is the temperature difference across thickness Δx of the material. Air is a poor thermal conductor, therefore heat loss by conduction is relatively low. On the other hand, water has a thermal conductivity 20 times that of air at 0 °C, increasing further with temperature, therefore during hyperthermia it is important that any water in contact with the skin is not at a lower temperature. The relative thermal conductivity of body tissues is important in determining thermal conduction within the body from external sources of heat. For example, fat is a relative thermal insulator with a thermal conductivity one third of that of most other tissues, therefore fat bodies are slower to heat.

Convective heat transfer involves material movement and occurs principally via blood moving heat to, or from, the skin and other tissues, and air currents (respiratory and environmental) moving warm air to or from the body. Equation 7 is written for the blood,

$$\frac{\Delta Q'_{\text{conv}}}{\Delta V} = \rho_b c_b[w_b(x,y,z,T) \cdot (T - T_b) + U_b(x,y,z,T) \cdot \nabla T] \tag{7}$$

where w_b is the specific capillary blood flow rate, U_b is the specific blood flow through other vessels. In the context of

systemic hyperthermia, where a patient is in a closed chamber, environmental air currents can be minimized. Heat loss by respiration, however, can amount to almost 10% of metabolic heat generation.

Another route of heat loss from the body is evaporation of perspiration from the skin. Because of the very large heat of vaporization of water, cooling of the blood in skin capillaries occurs due to evaporation of sweat. Evaporation from exhaled moisture also results in cooling of the surrounding air.

$$\frac{\Delta Q'_{\text{evap}}}{\Delta V} = m_w \frac{L_v}{\Delta t} \tag{8}$$

where m_w is the mass of the water and L_v is the latent heat of vaporization (2.4×10^6 J·kg^{-1} at 34 °C). In hot conditions with maximal rates of evaporation, heat loss through evaporation of sweat can be as much as 1100 W. Heat loss in the lungs is ~10 W.

Combining the heat generation and heat loss terms leads to a general heat transfer equation, an extension of the classic Pennes bioheat transfer equation.

$$\left[\left(\frac{\Delta Q'_{\text{met}}}{\Delta V} + P_{\text{abs}}\right) - \left(\frac{\Delta Q'_{\text{rad}}}{\Delta V} + \frac{\Delta Q'_{\text{cond}}}{\Delta V} + \frac{\Delta Q'_{\text{conv}}}{\Delta V} + \frac{\Delta Q'_{\text{resp}}}{\Delta V} + \frac{\Delta Q'_{\text{evap}}}{\Delta V}\right)\right] = \rho_{\text{tis}} c_{\text{tis}} \Delta T \tag{9}$$

into which the expressions given in Eqs. 2–8 may be substituted. Precise solution of this equation for temperature distribution is complex and requires a number of simplifying assumptions which have generated significant controversy in bioheat transfer circles. Modeling of temperature distributions within a body subjected to hyperthermia is also complex because of the heterogeneity of thermal characteristics between and within tissue, the directionality of power application, and the dynamic nature of thermoregulation by human body. Nonetheless, the factors governing systemic heating of the body can be appreciated.

INDUCTION OF SYSTEMIC HYPERTHERMIA

Apart from the induction of biological fever by pathogens or toxins, all methods of hyperthermia involve transfer of heat into the body from an external energy source. The required net power to raise the temperature of a 70 kg human from 37 to 41.8 °C (2) is 400 W (5.7 mW). While the heat absorption from these sources is highly nonuniform, distribution of thermal energy by the vascular system quickly results in a uniform distribution of temperature. Indeed, systemic hyperthermia is the only way to achieve uniform heating of tissues. Because physiological thermoregulation mechanisms such as vasodilation and perspiration counteract attempts to increase core body temperature, careful attention must be paid to optimizing the physical conditions for heating such that there is efficient deposition of heat energy in the body and, even more importantly, minimization of heat losses. Wrapping the body in reflective blankets, foil, or plastic film to reduce radiative and evaporative losses, or keeping the surrounding air moist to minimize losses by perspiration are key techniques for achieving a sustained increase in body temperature.

Noninvasive methods of heating include immersing the body in hot water or wax, wrapping the body in a blanket or suit through which heated water is pumped, placing the patient on a heated water mattress, surrounding the body with hot air, irradiating with IR energy, and applying RF or microwave electromagnetic energy. These techniques may be applied singly or in combination. For example, the Pomp–Siemens cabinet used until recently throughout Europe, as well as in the United States, a modification of a device originally developed by Siemens in the 1930s, has the patient lying on a heated water mattress under which an inductive loop generates an RF field, all inside a chamber through which hot air is circulated. The Russian Yakhta-5 system applies a high frequency (13.56 MHz) electromagnetic field through a water-filled mattress to permit whole body heating up to 43.5 °C and simultaneous deep local hyperthermia through additional applicators providing 40.6 MHz electromagnetic radiation. The majority of whole-body hyperthermia systems currently in clinical use employ IR radiation to achieve systemic heating. Invasive approaches to systemic hyperthermia are extracorporeal heating of blood, removed from the body via an arteriovenous shunt, prior to returning it to the circulation, as well as peritoneal irrigation with heated fluid (4). A useful schematic summary of whole-body hyperthermia induction techniques along with references is provided by van der Zee (5).

All of these approaches involve a period of steady temperature increase, followed by a plateau or equilibrium phase where the target temperature is maintained for anywhere from 30 min to several hours, and finally a cool-down phase. Depending on the method of hyperthermia induction, the patient may be anesthetized, consciously sedated, administered analgesia, or not given any kind of medication at all. An epidural block is sometimes given to induce or increase vasodilation. During radiant heat induction, the temperature of the skin and superficial tissues (including tumors) is higher than the core (rectal) temperature whereas during the plateau (maintenance) phase, the skin–superficial tissue temperature drops below the core temperature. As already described, heat losses due to physiological mechanisms limit the rate of heating that can be achieved. When insulation of the patient with plastic foil was added to hot air heating, the heating time to 41.8 °C was decreased from 230 to 150 min (65%), and further to 110 min (48%) by addition of a warm water perfused mattress (5). The homogeneity of the temperature distribution was also significantly increased by the addition of insulation and the water mattress. Noninvasive systemic hyperthermia methodologies typically produce heating rates of 1–10 °C·h^{-1} with 2–3 °C·h^{-1} being most common. More rapid heating can be achieved by the invasive techniques, at the expense of greater risk of infection and morbidity.

COMMERCIALLY AVAILABLE WHOLE-BODY HYPERTHERMIA SYSTEMS

A number of commercially available devices have resulted from the development of these initially experimental

systems. The Siemens–Pomp system has already been mentioned, but is no longer commercially available. Similarly, neither the radiant heat chamber developed by Robins (3), and marketed as the Aquatherm system, nor the similar Enthermics Medical Systems RHS-7500 radiant heat device, both producing far IR radiation (IR C) in a moist air chamber, are currently being sold, though they are still in use in several centers. A close relative is the Iratherm2000 radiant heat chamber originally developed by von Ardenne and co-workers (6). In this device, water-filtered infrared radiators at 2400 °C emit their energy from above and below the patient bed, producing near-IR (IR A) radiation that penetrates deeper into tissue than far IR radiation, causing direct heating of the subcutaneous capillary bed. Thermal isolation is ensured by reflective foils placed around the patient. However, note that significant evaporative heat loss through perspiration can be a problem with this system. Also with a significant market share is the Heckel HT 2000 radiant heat device in which patients lie on a bed enclosed within a soft-sided rectangular tent whose inner walls are coated with reflective aluminum foil that ensures that the short wavelength infrared A and B radiation emitted by four radiators within the chamber uniformly bathes the body surface. Once the target temperature is reached, the chamber walls are collapsed to wrap around the body, thereby preventing radiative and evaporative heat loss, and permitting maintenance of the elevated temperature, as shown in Fig. 2.

Another radiant heat device, used mainly in Germany, is the HOT-OncoTherm WBH-2000 whole-body hyperthermia unit which is a chamber that encloses all but the patient's head. Special light-emitting diode (LED) radiators deliver computer-generated, alloy-filtered IR A wavelengths that penetrate the skin to deliver heat to the capillary bed. The manufacturer claims that these wavelengths also preferentially stimulate the immune system. Recently, Energy Technology, Inc. of China has released the ET-SPACE whole-body hyperthermia system, which produces IR A radiation in a small patient chamber into which warm liquid is infused to help increase the air humidity and thereby reduce perspiration losses. A number of low cost, far infrared, or dry, saunas are being sold to private clinics, health clubs, and even individuals for treatment of arthritis, fibromyalgia, detoxification, and weight loss. Examples are the Smarty Hyperthermic Chamber, the TheraSauna, the Physiotherm, and the Biotherm Sauna Dome. Table 2 summarizes features of these commercially available whole-body hyperthermia devices.

Figure 2. Heckel HT-2000 radiant heat whole body hyperthermia system. Unit at the University of Texas Medical School at Houston. Patient is in the heat maintenance phase of treatment, wrapped in the thermal blankets which form the sides of the chamber during active heating.

BIOLOGICAL EFFECTS OF SYSTEMIC HYPERTHERMIA

An understanding of the biological effects of systemic hyperthermia is critical to both its successful induction and to its therapeutic efficacy. Systemic responses to body heating, if not counteracted, undermine efforts to raise body temperature, while cellular effects underlie both the rationale for the use of hyperthermia to treat specific diseases, and the toxicities resulting from treatment. Although improved technology has allowed easier and more effective induction of systemic hyperthermia, most of the recent clinical advances are due to better understanding and exploitation of specific biological phenomena.

Physiological Effects of Elevated Body Temperature

The sympathetic nervous system attempts to keep all parts of the body at a constant temperature, tightly controlled by a central temperature 'set point' in the preoptic–anterior hypothalamus and a variety of feedback mechanisms. The thermostat has a circadian rhythm and is occasionally reset, for example, during fever induced by infectious agents and endotoxins, but not in endogenously induced hyperthermia. Occasionally, it breaks down completely as in malignant hyperthermia or some neurological disorders affecting the hypothalamus. Ordinarily, when core body temperature rises, the blood vessels initially dilate, heart rate rises, and blood flow increases in an effort to transport heat to the body surface where it is lost by radiation, conduction, and convection. Heart rate increases on average by 11.7 beats·min^{-1}· °C^{-1} and typically remains elevated for several hours after normal body temperature is regained. Systolic blood pressure increases to drive the blood flow, but diastolic pressure decreases due to the decreased resistance of dilated vessels, thus there is an increase in cardiac output. Heart rate and blood pressure must therefore be monitored during systemic hyperthermia, and whole-body hyperthermia is contraindicated in most patients with cardiac conditions. Interestingly, hyperthermia increases cardiac tolerance to ischemia/reperfusion injury probably due to activation of manganese superoxide dismutase (Mn-SOD) and involvement of cytokines.

Respiration rate also increases and breathing becomes shallower. Perspiration results in evaporation of sweat from the skin and consequent cooling, while the respiration rate increases in order to increase cooling by evaporation of moisture from expired air. Weight loss occurs despite fluid intake. There is a decrease in urinary output and the urine has a high specific gravity, concentrating urates and

Table 2. Commercially Available Clinical Whole-Body Hyperthermia Devices

Manufacturer	Website	Device Name	Heating Mechanism	Temperature Range, °C	Application
Energy Technology	http://www.eti.com.cn/EN/pro/product2.htm	ET-SPACE	Multiple IR radiators (IR A)	39–41.8	Oncology
Heckel Medizintechnik GmbH	http://www.heckel-medizintechnik.de/frameset_e.html	HT 2000 M	4 300W IR radiators (IR A, B)	38.5–40.5	Oncology, rheumatology
Hot-Oncotherm	http://www.hot-oncotherm.com/oncothermia.htm	WBH-2000	Multiple LED radiators (IR A)	37–42	Oncology
Von Ardenne Institut für Angewandte Medizinische Forschung, GmbH	http://www.ardenne.de/med_eng/	Iratherm 800	4 IR radiators (IR A)	37–38	Physical medicine, complementary medicine, oncology
		Iratherm 1000	6 IR radiators (IR A)	37–39	
		Iratherm 2000	10 IR radiators (IR A)	37–42	

phosphates. In endogenously induced hyperthermia, but not in fever, glomerular filtration, as evidenced by the creatinine clearance, decreases with increasing temperature. As already mentioned, metabolic rate increases nonlinearly with temperature, which leads to an increase in blood sugar, decreased serum potassium levels, and increased lactic acid production. All the above normal physiological effects may be enhanced or counteracted by anesthesia or sedation, as well as by disease states such as cancer because of drugs used in treatment or intrinsic pathophysiological consequences of the disease.

At ~42.5 °C, the normal thermocompensatory mechanisms break down and the body displays the symptoms of advanced heat stroke, namely, lack of sweating, rapid heart beat, Cheyne–Stokes breathing, central nervous system disfunction, and loss of consciousness. Ultimately, breathing ceases despite the continuation of a heart beat.

Cellular Thermal Damage

When temperature is increased by a few degrees Celcious, there is increased efficiency of enzyme reactions (Arrhenius equation), leading to increased metabolic rates, but at temperatures > 40 °C molecular conformation changes occur that lead to destabilization of macromolecules and multimolecular structures, for example, to the side chains of amino acids in proteins, which in turn inhibit enzyme action. Small heat shock proteins (HSP) interact with the unfolding proteins to stabilize them and prevent their aggregation and precipitation. Eventually, however, at ~42 °C, complete denaturation of proteins begins that totally disrupts many molecular processes, including deoxyribonucleic acid (DNA) repair. Thus systemic hyperthermia can have significant effects when paired with drugs that cause DNA damage (e.g., for chemotherapy of cancer).

Membranes are known to be extremely sensitive to heat stress because of their complex molecular composition of lipids and proteins. At a certain temperature, lipids change from the tightly packed gel phase to the less tightly packed liquid crystalline phase, and permeability of the cell membrane (membrane fluidity) increases. As temperature increases further, the conformation of proteins also becomes affected, eventually resulting in disorderly rearrangement of the lipid bilayer structure and receptor inactivation or loss. Temperature changes of ~5 °C are necessary to cause measurable changes in normal cell membrane permeability. Heat-induced cell membrane permeability can be exploited to increase drug delivery, for example, transdermally, or into tumor cells. Increased vascular permeability due to thermal increase of endothelial gap size also aids drug delivery into tumors. At higher temperatures, heat damage to membranes can cause cell death, but it will also interfere with therapeutic approaches that depend on membrane integrity (e.g., receptor targeted drug delivery, antibodies, etc.). Irreversible disruption of cytoplasmic microtubule organization and eventual disaggregation, as well as disruption of actin stress fibers and vimentin filaments, occur at high temperatures (43–45 °C) above those used in whole-body hyperthermia, but these cytoskeletal effects are of concern with loco-regional hyperthermia.

A variety of effects in the cell nucleus also occur at high temperatures (>41 °C) including damage to the nuclear membrane, increases in nuclear protein content, changes in the structure of nucleoli, inhibition of DNA synthesis and chromosomal damage in S-phase. These changes in nuclear structure compromise nuclear function and may cause cell death, though they are unlikely to be significant at the temperatures achieved in systemic hyperthermia. Disaggregation of the spindle apparatus of mitotic cells may be responsible for the high thermal sensitivity of cells in mitosis, as well as in S phase. Hyperthermic inactivation of polymerase β, an enzyme primarily involved in DNA repair, is sensitized by anesthetics and may have a role to play in the enhancement of the effects of ionizing radiation by systemic hyperthermia, as well as in augmenting the cytotoxic effect of drugs that cause DNA damage.

Metabolic Effects

Moderate increases in temperature lead to increased cellular reaction rates, which may be seen as increased oxygen consumption and glucose turnover. In consequence, cells may become deprived of nutrients, the intracellular ATP concentration falls, accumulation of acid metabolites increases pH, and thermal sensitivity increases. Such conditions are found in tumors and may contribute to their sensitivity to heat. Further acidifying tumor cells during hyperthermic treatment seems a promising approach as is discussed further below. At high temperatures, the citric acid cycle may be damaged leading to other acidic metabolites. Increased plasma acetate has been measured following clinical whole-body hyperthermia treatments, which reduces both release of fatty acids from adipose tissue into plasma and subsequent lipid oxidation.

Endocrine Function

Increases in plasma levels of an array of hormones have been noted after whole-body hyperthermia. Increased ACTH levels appear to be accompanied by increased levels of circulating endorphins. This may explain the sense of well-being felt by many patients after systemic hyperthermia treatment, and the palliative effect of hyperthermia treatments for cancer. Increased secretion of somatotropic hormone after systemic hyperthermia has also been measured (7).

Thermal Tolerance

Thermal tolerance is a temporary state of thermal resistance, common to virtually all mammalian cells, which develops after a prolonged exposure to moderate temperatures (40–42 °C), or a brief heat shock followed by incubation at 37 °C, and also certain chemicals. The decay of thermotolerance occurs exponentially and depends on the treatment time, the temperature, and the proliferative status of the cells. Several days are usually required for baseline levels of heat sensitivity to be regained, which has important implications for fractionated therapy. When pH is lowered, less thermal tolerance develops, and its decay is slower. Thus the long periods at moderate temperature achieved by clinical systemic hyperthermia systems should

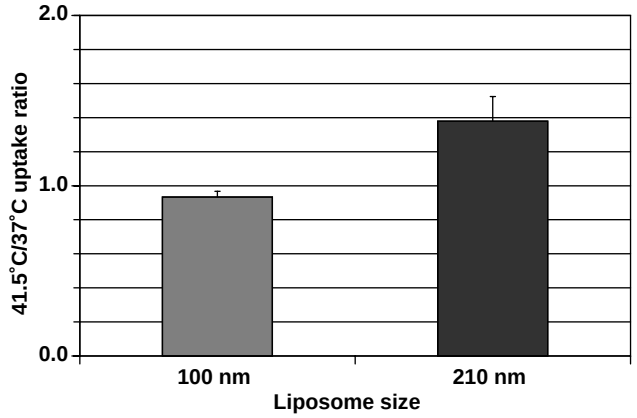


Figure 3. Increase in tumor uptake of large liposomes after 1 h of 41.5 °C whole-body hyperthermia. Systemic heat treatment increased the effective pore size from ~210 to 240 nm. Because of the large pore size in MTLn3 tumors, 100 nm (average diameter) liposomes were able to pass into the tumor equally well at normal and elevated temperatures. The increased effective pore size due to hyperthermia allowed larger 200 nm liposomes, which were partially blocked at normal temperatures, to pass more effectively into the tumor.

induce thermal resistance in normal cells, while the acidic parts of tumors should be relatively unaffected. This has not, however, been studied clinically. The mechanisms involved in the induction of thermotolerance are not well understood, but there is mounting evidence that heat shock proteins are involved.

Step-Down Sensitization

Another distinct phenomenon is step-down sensitization in which an exposure of cells to temperatures >43 °C results in increased sensitivity to subsequent temperatures of 42 °C or lower. This can be important clinically for local and regional hyperthermia if there are marked variations in temperature during the course of treatment, the magnitude of the effect depending on the magnitude of the temperature change. It has been suggested that this phenomenon could be exploited clinically by administering a short, high temperature treatment prior to a prolonged treatment at a lower temperature, thereby reducing pain and discomfort. Since temperatures >43 °C cannot be tolerated systemically, a local heat boost would be required to take advantage of this effect for whole body hyperthermia. So far, there is no evidence that tumor cells are differently sensitized by step-down heating than normal cells.

Effect of Hyperthermia on Tumors

It was initially thought that tumor cells have intrinsically higher heat sensitivity than normal cells, but this is not universally true. Although some neoplastic cells are more sensitive to heat than their normal counterparts, this appears to be the case at temperatures higher than those used in systemic hyperthermia. Tumors *in vivo*, on the other hand, often do have a higher thermal sensitivity than normal tissues because of abnormal vasculature (reduced blood flow), anaerobic metabolism (acidosis), and nutrient depletion. Due to their tortuous and poorly constructed vasculature, tumors have poor perfusion, thus heat dissipation by convection is reduced. At high temperatures (43 °C and up) this means that tumors become a heat reservoir with a consequent rise in temperature, which if maintained for too long damages the microcirculation and further impairs convective heat loss. Also increased fibrinogen deposition at damaged sites in the vascular wall leads to clusion of tumor microvessels. Significant heating of the tumor cells results, which may be directly cytotoxic. Additionally, the impaired blood flow brings about acidosis, increased hypoxia and energy depletion all of which increase the heat sensitivity of tumor cells (8). At lower temperatures, typical of those achieved in whole-body hyperthermia, blood flow increases (9) though the mechanism is not well understood. For these reasons, along with the historical evidence for antitumor effects of fever and the metastatic nature of malignant disease, cancer has become the main focus of systemic hyperthermia.

Systemic hyperthermia results in increased delivery of drugs to tumor sites because of increased systemic blood flow. It can also increase blood vessel permeability by increasing the effective pore size between the loosely bound endothelial cells forming tumor microvessels, permitting larger molecules, such as nanoparticles and gene therapy vectors, to pass into the interstitium (10). Figure 3 shows increased uptake of 210 nm liposomes in rat breast tumors after 1 h of 41.5 °C whole-body hyperthermia. Heat may also be toxic to endothelial cells, resulting in a transient normalization of vascular architecture and improvement in blood flow (11). Another barrier to drug delivery is the high interstitial pressure of many tumors. Since whole-body hyperthermia, even at fever-range temperatures, causes cell death (apoptosis and necrosis) within tumors it reduces the oncotic pressure allowing greater penetration of large molecules. Table 3 summarizes the interactions of systemic hyperthermia which facilitate nanoparticle delivery to tumors.

Table 3. Whole-Body Hyperthermia Facilitates Nanoparticle Therapy

Heat Interaction	Therapeutic Effect
↑ Blood flow	↑ Nanoparticle delivery to tumor
↑ In endothelial gap size	↑ Nanoparticles in interstitium
↑ Endothelial cell apoptosis/necrosis → transient normalization of vasculature	↑ Nanoparticles in interstitium
↑ Tumor cell apoptosis/necrosis ↓ oncotic pressure	↑ Nanoparticles in interstitium
Temperature-dependent ↑ in permeability of liposome bilayer	↑ And synchronization of drug release
Cellular and molecular effects in tumor	↑ Drug in tumor cell ↑ drug efficacy
Direct interactions with drug	↑ Drug efficacy

Whole-Body Hyperthermia and the Immune System

An increase in ambient temperature can serve as a natural trigger to the immune system and it appears that the thermal microenvironment plays a critical role in regulating events in the immune response. The early work of Coley on cancer therapy with infectious pyrogens implicated fever-induced immune stimulation as the mediator of tumor responses (1). While there have been numerous *in vitro* studies of the effect of temperature on components of the immune system, indicating that the thermal milieu regulates T lymphocytes, natural killer (NK) cells, and dendritic cells (DC), *in vivo* examinations of the immune effects of systemic hyperthermia are relatively few. Initial animal model studies concluded that whole-body hyperthermia resulted in immunosuppression, but high temperatures were used, tumors were mostly immunogenic, and immune response was merely inferred from the incidence of metastatic spread rather than from measurement of specific markers of immune system activation. The majority of *in vivo* studies in animals provide evidence of a nonspecific host reaction in response to hyperthermia in which both T and B lymphocytes, as well as macrophages, are involved (12). Although NK cells are intrinsically more sensitive *in vitro* to heat than B and T cells, their activation by systemic hyperthermia has been observed. Microwave induced whole-body hyperthermia of unrestrained, unanesthetized mice at 39.5–40 °C for 30 min, three or six times weekly, resulted in increased NK cell activity and reduced pulmonary metastasis in tumor-bearing mice, but none in normal mice (13). Evidence for hyperthermia-induced human tumor lysis by IL-2 stimulated NK cells activated by HSP72 expression also exists (14). Increased numbers of lymphocyte-like cells, macrophages, and granulocytes are observed in the tumor vasculature and in the tumor stroma of xenografts and syngeneic tumors in mice immediately following a mild hyperthermia exposure for 6–8 h. In the SCID mouse/human tumor system tumor cell apoptosis seen following treatment was due largely to the activity of NK cells. The investigators hypothesize heat dilatation of blood vessels and increased vessel permeability may also give immune effector cells greater access to the interior of tumors (15). In balb/C mice, fever-range whole-body hyperthermia increased lymphocyte trafficking, resulting in early responsiveness to antigen challenge (16). Thus systemic hyperthermia may be an effective, nontoxic adjuvant to immunotherapy.

A recent clinical study examined the effect of whole-body hyperthermia combined with chemotherapy on the expression up to 48 h later of a broad range of activation markers on peripheral blood lymphocytes, as well as serum cytokines and intracellular cytokine levels in T cells, and the capacity of these cells to proliferate. Immediately after treatment with 60 min of 41.8 °C WBH as an adjunct to chemotherapy, a drastic but transient, increase in peripheral NK cells and CD56+ cytotoxic T lymphocytes was observed in the patients' peripheral blood. The number of T cells then briefly dropped below baseline levels, a phenomeonon that has also been observed by others (17). A marked, but short-lived, increase in the patients' serum levels of interleukin-6 (IL-6) was also noted. Significantly increased serum levels of tumor necrosis factor-alpha (TNF-alpha) were found at 0, 3, 5 and 24 h posttreatment. Further immunological consequences of the treatment consisted of an increase in the percentage of peripheral cytotoxic T lymphocytes expressing CD56, reaching a maximum at 48 h post-WBH. Furthermore, the percentage of CD4+ T cells expressing the T cell activation marker CD69 increased nearly twofold over time, reaching its maximum at 48 h. Since similar changes were not observed in patients receiving chemotherapy alone, this study provided strong evidence for prolonged activation of human T cells induced by whole-body hyperthermia combined with chemotherapy (18).

Activation of monocytes has been observed following hot water bath immersion such that response to endotoxin stimulation is enhanced with concomitant release of TNF-α. Macrophage activation and subsequent lysosomal exocytosis were observed in the case of a patient treated for liver metastases by hyperthermia. Lysosomal exocytosis induced by heat may be an important basic reaction of the body against bacteria, viruses, and tumor growth and was proposed as a new mechanism of thermally induced tumor cell death mediated by an immune reaction (19).

Several investigators have suggested that the immune changes seen during *in vivo* whole-body hyperthermia are mediated by elevations in the plasma concentrations of either catecholamines, growth hormone, or beta-endorphins. In volunteers immersed in a heated water bath, neither recruitment of NK cells to the blood, nor the percentages or concentrations of any other subpopulations of blood mononuclear cells were altered by hormone blockade. However, somatostatin partly abolished the hyperthermia induced increase in neutrophil number. Based on these data and previous results showing that growth hormone infusion increases the concentration of neutrophils in the blood, it was suggested that growth hormone is at least partly responsible for hyperthermia induced neutrophil increase. A similar study suggested that hyperthermic induction of T lymphocytes and NK cells is due to increased secretion of somatotropic hormone (7).

The peripheral blood level of prostaglandin E_2 (PGE_2), which may act as an angiogenic switch, transforming a localized tumor into an invasive one by stimulating new blood vessel growth, and which also has an immunosuppressive effect, is elevated in patients with tumors compared to healthy control subjects. In a clinical study of cancer patients receiving 1–2 h of 41.8–42.5 °C whole-body hyperthermia, or extracorporeal hyperthermia, blood levels of PGE_2 decreased markedly after treatment and correlated with tumor response (20).

In addition to their role as protectors of unfolding proteins, extracellular heat shock proteins (HSP) can act simultaneously as a source of antigen due to their ability to chaperone peptides and as a maturation signal for dendritic cells, thereby inducing dendritic cells to cross-present antigens to CD8+ T cells (21). Heat shock proteins can also act independently from associated peptides, stimulating the innate immune system by eliciting potent proinflammatory responses in innate immune cells. The heat shock response also inhibits cyclooxygenase-2 gene expression at the transcriptional level by preventing the activation of

nuclear factor-kappaB (NFκB) (22). Thermal upregulation of HSPs (HSP70 and HSP110) is strongest in lymphoid tissues and may relate to the enhanced immune responses that are observed during febrile temperatures. It has been proposed that local necrosis induced by hyperthermic treatment induces the release of HSPs, followed by uptake, processing and presentation of associated peptides by dendritic cells. By acting as chaperones and as a signal for dendritic cell maturation, HSP70 might efficiently prime circulating T cells. Therefore, upregulating HSP70 and causing local necrosis in tumor tissue by hyperthermia offers great potential as a new approach to directly activate the immune system, as well as to enhance other immunotherapies (23,24).

CLINICAL TOXICITIES OF WHOLE-BODY HYPERTHERMIA TREATMENT

At fever-range temperatures, adverse effects of systemic hyperthermia treatment are minimal however, at higher temperatures they can be significant, even fatal. On the other hand, the teratogenic effects (birth defects, still births, spontaneous abortions) and °Cular damage (cataract induction) resulting from electromagnetic fields used in local hyperthermia are not seen in systemic hyperthermia. The transient cardiorespiratory effects of elevated temperature can, however, lead to severe toxicity. Elevated heart rate, especially at high temperatures may result in arrythmias or ischemic heart failure, consequently patients have to be very carefully screened with regard to their cardiac status. Beta blockade has generally been found to be deleterious although infusion of esmolol has been safely carried out (25). Pulmonary hypertension and edema due to capillary leak may also be seen, but like the cardiac effects, these return to baseline a few hours after treatment. Increased serum hepatic enzymes have been noted, but these may be cancer related. All these toxicities are less prevalent or less severe with radiant heat systems, particularly at lower temperatures, and when light conscious sedation is used rather than general anesthesia. For example, decreased platelet count, decreased plasma fibrinogen, and other factors leading to increased blood clotting have been noted, particularly in extra-corporeal hyperthermia, but also with other methods of heating carried out under inhalation-administered anesthesia drugs. On the other hand, with whole-body hyperthermia under conscious sedation there is no evidence of platelet drops (26) and animal studies even show platelet stimulation providing protection against radiation induced thrombocytopenia.

Since systemic hyperthermia is almost never used as a single treatment modality, it is important to recognize that whole-body hyperthermia combined with radiation and chemotherapy can enhance some of the toxicities associated with these modalities. For example, the cardiotoxicity of doxorubicin and both the renal toxicity and hematological toxicity of platinum agents may increase under hyperthermia (27), while the muscle and peripheral nervous system effects of radiation and some drugs can also be enhanced (28). Bone marrow suppression is the limiting toxicity of many chemotherapy drugs but there is little data to suggest that whole body hyperthermia exacerbates this effect. On the contrary, the synergy of hyperthermia with several chemotherapy agents may mean that lower doses can be used, resulting in less toxicity. For example, systemic hyperthermia combined with carboplatin achieves therapeutic results without elevation of myelosuppression and responses have occurred at lower than normal doses (29). Pressure sores can easily develop at elevated temperatures thus care must be taken not only in patient placement and support, but also with application of monitoring devices. If heat dissipation is locally impaired, for example, at pressure points, hot spots occur that can lead to burns. This is rarely a problem with fever-range whole-body hyperthermia, but in anesthetized patients undergoing high heat regimens burns are not uncommon.

Following systemic hyperthermia treatments, malaise and lethargy are almost universally experienced although these may be counteracted by pain relief and a sense of well-being due to released endorphins. However, the faster the target temperature is reached, the less the exhaustion (6), thus attention to minimizing heat dissipation during the heat-up phase and using efficient heating devices, such as those that generate heat by several mechanisms (e.g., radiant heat and EM fields), add a regional heat boost, or produce near-IR radiation that is preferentially absorbed, is advantageous to patient well being. Fever after treatment in the absence of infectious disease is not uncommon and may be associated with an inflammatory response to tumor regression. Nausea and vomiting during the first couple of days after treatment are also common. Outbreaks of herpes simplex (cold sores) in susceptible individuals have also been noted, but are easily resolved with acyclovir.

THERMAL DOSE

The definition of dose for systemic hyperthermia is problematic. An applied dose would be the amount of heat energy generated or delivered to the body but even if it can be measured, this quantity does not predict biological effects. By analogy with ionizing radiation, the absorbed dose would be amount of thermal energy absorbed per unit mass of tissue ($J \cdot kg^{-1}$), however, this is not a quantity that can be readily measured, or controlled, neither would it necessarily predict biological effects. As indicated in the previous sections, the effects of systemic hyperthermia depend on (*1*) the temperature, and (*2*) the duration of heating, but not on the energy required to produce the temperature rise. This leads to the concept of time at a given temperature as a practical measure of dose. In reality, however, temperature is seldom constant throughout a treatment, even in the plateau phase of systemic hyperthermia, so time at temperature is at best a crude measure. Nonetheless, it is the one that is used most often clinically for whole-body hyperthermia because of its simplicity. Ideally, the dose parameter should allow for comparison of treatments at different temperatures. Based on the Arrhenius relationship and measured cell growth inhibition curves, the heating time at a given temperature relative to the heating time at a standard temperature or

thermal dose equivalent (TDE), was defined empirically as,

$$T_1 = t_2 \cdot R^{(T_1 - T_2)} \tag{10}$$

A discontinuity occurs in the temperature-time curves between 42 and 43 °C for both cells in culture and heated tissues, thus the value of R changes for temperatures above the transition: $R \sim 2 <$ 42.5 °C and $R \sim 5 >$ 42.5 °C *in vitro* while for *in vivo* heating studies, $R = 2.1$ below the transition temperature and 6.4 above 42.5 °C. In practice, a finite time is required for the body or tissue of interest to reach the target temperature, temperature fluctuates even after the target temperature is reached, and there is a cooling period after heating ceases. If the temperature is measured frequently throughout treatment, the temperature–time curves can be integrated to provide the accumulated thermal dose that produces an equivalent effect to that resulting from holding the cells–tissue at a constant reference temperature for a given a period of time:

$$t_{43} = \int_{t_i}^{t_f} R^{43 - T(t)} dt \tag{11}$$

where t_i and t_f are the initial and final times of the heating procedure (30). This thermal isoeffect dose (TID) is usually expressed in minutes is sometimes known as the tdm43 or the cumulative equivalent minutes (CEM 43 °C). While a biological factor has now been built in to the dose measure, and the integrated TID allows for temperature variations during heat-up and cool-down phases, it does not take into account thermal tolerance and step-down sensitization. Nor is it particularly relevant to clinical whole-body hyperthermia where multiple physical and biological effects combine in a complex manner although for a given patient, time–temperature profiles are generally reproducible from one treatment to another. A further modification attempts to take into account temperature inhomogeneity through the measurement of temperature at multiple sites and defining T90, namely, that temperature exceeded by 90% of the measurements (or correspondingly 20%: T20; or 50%: T50). The TID is then expressed as cumulative equivalent minutes that T90 is equal to 43 °C (CEM 43 °C T90) (31).

The efficiency of adjuvant hyperthermia in enhancing the biological effectiveness of other treatments is often reported in terms of the thermal enhancement factor (TEF) or thermal enhancement ratio (TER). This quantity is defined in terms of the isoeffect dose as,

$$\text{TER} = \frac{\begin{array}{c}\text{dose of treatment to achieve} \\ \text{a given endpoint}\end{array}}{\begin{array}{c}\text{dose of treatment with heat} \\ \text{to achieve the same endpoint}\end{array}} \tag{12}$$

In clinical and laboratory studies, the TER is often computed on the basis of isodose rather than isoeffect, for example, in the case of hyperthermia plus drug induced arrest of tumor growth, $\text{TER} = \text{TGD}_{\text{HT}}/\text{TGT}_{\text{RT}}$, where TGD_{HT} is the tumor growth delay due to hyperthermia plus chemotherapy, and TGT_{RT} is the tumor growth delay resulting from chemotherapy at room temperature. Similarly, the enhancing effect of hyperthermia on radiation treatment may be expressed through $\text{TER} = \text{D0}_{\text{HT}}/\text{D0}_{\text{RT}}$ or $\text{TER} = \text{LD50}_{\text{HT}}/\text{LD50}_{\text{RT}}$, where D0 is the time required to reduce survival to $1/e$ of its initial value, and LD50 is the lethal dose to 50% of cells.

TEMPERATURE MEASUREMENT

Since systemic hyperthermia achieves a uniform temperature distribution, except for possible partial sanctuary sites, thermometry for systemic hyperthermia is much less challenging than for regional or intracavitary hyperthermia, but it is still important to prevent adverse effects, especially burns. Also, convection can induce steep thermal gradients, especially around major blood vessels, so that careful placement of temperature probes is required. Most practitioners of whole-body hyperthermia measure temperature in several locations, typically the rectum, the esophagus, and at several skin sites. During heat-up, the esophageal temperature is usually 1–2 °C higher than the rectal temperature, but during plateau phase it drops to 0.5–1.5 °C below the rectal temperature. Continuous and accurate temperature measurement is particularly important when temperatures >41°C are to be achieved, as critical, life-threatening changes can occur in minutes or even seconds and over changes in temperature of as little as 0.1–0.2 °C because of the nonlinear response to temperature. For moderate temperature systemic hyperthermia, temperature measurement to within 0.1 °C is usually adequate, but a precision of 0.01 °C is desirable when heating to >41 °C and also allows determination of the specific absorption rate from the slope of the temperature versus time curve. The temperature measuring device must be insensitive to all other influences, such as ambient temperature, moisture, nearby electromagnetic fields, and so on and satisfying this criterion can be difficult. Frequent calibration of thermometers in the working range of temperatures is important since some thermometers appear fine at 30 °C, but drift substantially at 40 °C and above. Stringent quality control of any thermometry system is required to monitor accuracy, precision, stability, and response time.

Table 4 summarizes the different types of thermometer probes available for internal and external body temperature measurements, and their relative merits and disadvantages for systemic hyperthermia. Thermistors are most often used for standard temperature monitoring sites while thermocouples are used for tumor or other intra-tissue measurements. Recently, noninvasive methods of temperature measurement have been developed that are beginning to see application in hyperthermia. Thermography provides a two-dimensional (2D) map of surface temperature by measurement of infrared emission from the body, though deep-seated hot structures may be visualized because of heat carried by blood flow from the interior heat source to the skin. It is useful to detect skin hotspots and therefore in burn prevention. Since temperature-induced changes in the mechanical properties of tissue lead to altered ultrasound propagation velocity, mapping of ultrasound velocity can also provide a visual map of temperature. Tomographic reconstruction of 2D or 3D temperature is theoretically possible, but it is difficult in practice because of the heterogeneity of tissue characteristics. A

Table 4. Temperatures Probes for Systemic Hyperthermia

Probe Type	Measurement Principle	Accuracy	Sensitivity	Stability	Advantages or Disadvantages
Clinical	Expansion of mercury or alcohol in glass	Moderate $\leq$ 0.1 °C	Low	High	Large size, inflexible. Slow response.
Platinum resistance thermometer	Linear resistance change with temperature	High $\sim$ 0.02 °C		Used as standard for calibration of other types of thermometers.	Expensive. Difficult to calibrate. Large size. Sensitive to shock.
Thermocouple	Seebeck effect: temperature dependent voltage difference between two conductors made of different metals	Moderate $\leq$ 0.1 °C	Moderate to high	Moderate	*Small sensor*. Nonlinear voltage change with temp. Sensitive to EM fields. Can't handle steep temp. gradients.
Thermistor (e.g., Bowman Loop Larsen probe)	Inverse relationship between temperature and semiconductor resistance	High $<$ 0.05 °C	High	Poor Require frequent recalibration	*Short time constant*. Not interchangeable. Sensitive to EM fields.
GaAs	Temperature specific absorption	Moderate	Low		*Small size*
Optical (fiber optic probe):	Change w/temp.:				*Not sensitive to EM fields*. *Small size*.
LCD birefringent crystal fluorescent phosphor	Color reflectance		Low	Low	Unstable
	Refraction of polarized light		Low		
	Decay of fluorescence	Very high			

Table 5. Summary of Clinical Trials of Whole-Body Hyperthermia[a]

First Author	Public Year	Study Type	Number of Patients	Disease	Protocol	Result of WBH	Reference, PMID
WBH Alone							
Kraybill, W.G.	2002	Phase I		Advanced solid tumors	3–6 h at 39.5–40.0 °C	Well tolerated No significant adverse events ↓ in circulating lymphocytes	16, 12028640
Steinhausen, D.	1994	Phase I	103	Advanced refractory or recurrent cancers	1 h at 41.8 °C + hyperglycemia + hyperoxemia	Minimal side effects 52 responses (50%)	8023241
WBH + Chemotherapy							
Bakshandeh, A.	2003	Phase II	25	Nonmetastatic malignant pleural mesothelioma	1 h at 41.8 °C + ifosfamide + carboplatin + etoposide	Grade III/IV neutropenia and thrombocytopenia 5 partial remissions (20%)	12609573
Bull, J.M.	1992	Phase II	17	Advanced metastatic sarcoma	2 h at 41.8–42.0 °C + BCNU	Limiting toxicity = thrombocytopenia 7 responses/SD (41%) ↑ survival	33, 1607734
Bull, J.M.	2002	Phase I	13	Various chemotherapy resistant cancers	6 h at 40.0 °C + doxil + 5-FU + metronomic interferon-α	Grade III toxicities 9 responses/SD (69%)	60
Bull, J.M.	2004	Phase I	33	Advanced metastatic cancers (GI, breast, head and neck, sarcoma, neuroendocrine)	6 h at 40.0 °C + cisplatin + gemcitabine + interferon-α	20 responses/SD (66%) ↑ survival ↑ quality of life	35
Douwes, F.	2004	Pilot	21	Ovarian cancer	1-2 h at 41.5–42.0 °C + cisplatin or carboplatin + hyperglycemia	18 responses/SD (86%) ↑ quality of life	15108039
Engelhardt, R.	1990	Pilot	23	Advance metastatic melanoma	1 h at 41.0 °C + cisplatin + doxorubicin	Slight ↑ in myelotoxicity 10 responses/SD Response rate = that in literature for chemo alone	52, 2198312
Guan, J.	2005	Phase II	32	Advanced cancers		94% responses/SD Pain reduction in all pts. Increased KPS Decreased tumor markers	65
Hegewisch-Becker, S.	2002	Phase II	41	Pretreated advanced metastatic colorectal cancer	1 h at 41.8 + oxaliplatin + leucovorin + 5FU	No excess toxicity 31 responses/SD (76%)	55, 12181242
Hildebrandt, B.	2004	Phase I/II	28	Metastatic colorectal cancer	1 h at 41.8–42.1 °C + hyperglycemia + hyperoxemia + folinic acid + 5-FU + mitomycin C	Grade III/IV toxicities 11 responses/SD (39%)	44, 15204528

Hou, K.	2004	Phase II	54	Advanced cancers	1–2 h at 41.8–42.5 °C, extracorporeal + chemotherapy vs. chemotherapy alone	75.3% responses/SD 72.6% ↓ tumor markers 70% pain relief improved sleep ↑ weight, appetite, KPS All signifly > control Reversible toxicities	68
Ismael-Zade, R.S.	2005	Pilot	5	Pediatric renal cell carcinoma	3 h at 41.8–42.5 °C + doxorubicin + interferon-α	No complications 5 responses (100%)	15700247
Kurpeshev, O.K.	2005	Phase II	42	Various disseminated cancers	1–2 h at 41.0–42.3 °C + poly-chemotherapy	Regression of metastases. Pain reduction	66
Richel, O.	2004	Phase II	21	Metastatic and recurrent cervical cancer	1 h at 41.8 °C + carboplatin	Grade III/IV leucopenia, thrombopenia, anemia, renal toxicity 16 responses/SD (76%)	57, 15581981
Robins, H.I.	1993	Phase I	30	Various refractory cancers	1 h at 41.8 °C + carboplatin	Myelotoxicity 9 responses (30%)	53, 8355046
Robins, H.I.	1997	Phase I	16	Various refractory cancers	1 h at 41.8 °C + L-PAM	Lower platelet nadir myelosuppression 8 responses/SD (50%)	48, 8996137
Strobl, B.	2004	Phase II	7	Metastatic cervical cancer	1 h at 41.5–41.8 °C + paclitaxel + carboplatin	Grade II alopecia Grade III/IV thrombopenia, neutropenia ↑ survival	58
Westermann A.M.	2001	Phase II	14	Platinum resistant ovarian cancer	1 h at 41.8 °C + carboplatin	Grade IV thrombocytopenia, grade III neutropenia 9 responses/SD (64%)	11378341
Westermann A.M.	2003	Phase II	95	Metastatic sarcoma	1 h at 41.8 °C + ifosfamide + carboplatin + etoposide	Neutropenia, thrombocytopenia, infection 58 responses/SD (61%)	50, 12759526
WBH + Radiation							
Overgaard, J.	1995	Randomized multicenter	70	Metastatic melanoma	1 h at 43 °C + fractionated RT vs. FRT alone	Improved local tumor control, ↑ survival	41, 7776772
Robins, H.I.	1990	Pilot	8	Nodular lymphoma, chronic lymphocytic leukemia	41.8 °C + TBI vs. LON + TBI	8 responses/SD ↑ survival	24, 2182581

[a]Published since 1990.

number of magnetic resonance (MR) techniques have been used for thermal mapping and BSD Medical and SIEMENS Medical Systems have collaborated to develop a hybrid hyperthermia/MRI system, although it is not a whole-body hyperthermia machine. Currently, the most widely accepted MR technique is the proton resonance frequency (PRF) method that exploits the temperature dependence of the chemical shift of water. Unlike the value of the water spin-lattice relaxation time or the molecular diffusion coefficient, both of which have been used for MRI temperature measurements, the thermal coefficient relating temperature to the water chemical shift has been shown to be essentially independent of tissue type and physiological changes induced by temperature (32). Recently an interleaved gradient echo–echo planar imaging (iGE-EPI) method for rapid, multiplanar temperature imaging was introduced that provided increased temperature contrast-to-noise and lipid suppression without compromising spatio-temporal resolution (33).

CLINICAL EXPERIENCE

Cancer

Systemic hyperthermia has been used mostly for treatment of cancer because of its potential to treat metastatic disease. Initial treatments aimed to produce direct killing of tumor cells based on the premise, now understood not to be universally true, that cancer cells are more susceptible to elevated temperatures than normal cells, and the higher the temperature the greater the tumor cell kill. Maximally tolerated temperatures of 41.5–42 °C were therefore maintained for 1–2 h as the sole treatment. Response rates were, however, disappointing. Tumor regressions were observed in less than half the cases, no tumor cures were achieved, and remissions were of short duration. It became apparent that the heterogeneity of cell populations within tumors, along with micro-environmental factors, such as blood/nutrient supply, pH, and oxygen tension prevent the thermotoxic results achieved in the laboratory. Consequently, the focus of research on systemic hyperthermia shifted to using hyperthermia as an adjunct to other cancer therapies, principally chemotherapy and radiotherapy. It is important to note that because of the experimental status of systemic hyperthermia treatment for cancer, almost all clinical trials, summarized in Table 5, have been performed on patients with advanced disease for whom whole-body hyperthermia, either as a sole therapy, or as an adjunct, is a treatment of last resort. In these cases, any response whatsoever is often remarkable. Nonetheless, a number of hyperthermia centers in Europe have discontinued systemic hyperthermia because the high temperature protocols required intensive patient care and led to unacceptable toxicities, especially in light of the efficacy and reduced toxicities of newer generation chemotherapies. Large, randomized, multicenter, Phase III trials are, however, needed to firmly establish the benefits of systemic hyperthermia in conjunction with chemotherapy and radiation. Also, validation and optimization of fever-range temperature protocols are much needed.

Systemic Hyperthermia and Chemotherapy. The beneficial interaction of hyperthermia with several classes of chemotherapy agents, acting via several mechanisms as summarized in Table 6, has spurred a variety of thermochemotherapy regimens and several clinical trials of systemic hyperthermia and chemotherapy are ongoing. While the results have been mixed, elevated response rates were recorded in the treatment of sarcoma when systemic hyperthermia was combined with doxorubicin and cyclophosphamide (54) or BCNU (34). Systemic hyperthermia is the only way to heat the lung uniformly, and impressive response rates and increased durations of response have been achieved in both small cell and nonsmall cell lung cancer treated with the combination of whole body hyperthermia at 41 °C for 1 h with adriamycin, cyclophosphamide, and vincristine (ACO protocol) (34). Neuroendocrine tumors also appear to have increased sensitivity to systemic hyperthermia and multidrug chemotherapy (51).

Optimal combination of whole-body hyperthermia with chemotherapy requires an understanding of the mechanisms of interaction of heat with individual drugs or drugs in combination. Preclinical data is consistent with the concept that the timing of chemotherapy during whole-body hyperthermia should affect therapeutic index. For example, Fig. 4 shows the effect on tumor cures in mammary carcinoma bearing rats of 6 h of 40 °C whole-body hyperthermia administered with, or 24 or 48 h after gemcitabine. A synergistic response was obtained when hyperthermia was begun with gemcitabine administration or 48 h later. The effect of gemcitabine was completely negated, however, when hyperthermia was administered 24 h after the start of heating, perhaps due to cell cycle effects. With cisplatin, the greatest therapeutic index is achieved if the drug is given 24 h before the start of whole-body hyperthermia, thereby preventing thermal augmentation of cisplatin induced nephrotoxicity (55). In a clinical investigation of multiple cycles of radiant heat whole-body hyperthermia combined with carboplatin, Ifosfamide, etoposide, and granulocyte colony stimulating factor, it was found that toxicity was minimized when carboplatin was

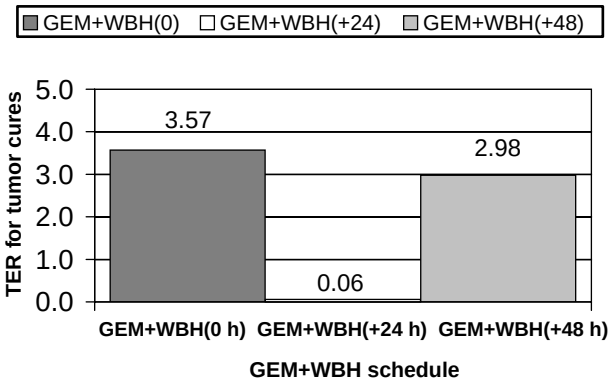


Figure 4. Schedule dependence of fever range whole-body hyperthermia enhanced gemcitabine tumor cures. A supraadditive cure rate occurred when whole-body hyperthermia (WBH) was given at the same time as gemcitabine administration or 48 h later. When hyperthermia followed gemcitabine by 24 h the number of cures dropped to almost zero, well below the number achieved with gemcitabine alone.

Table 6. Chemotherapy Agents Used with Whole-Body Hyperthermia

Class of Agent	Likely Mechanism of Heat Interaction	Drugs Used with WBH in Clinical Studies	Investigator References[a]
Alkylating agents	Impaired DNA repair Improved pharmacokinetics	Cyclophosphamide (CTX)	Parks, 1983 (4) Engelhardt, 1988 (34)
		Dacarbazine (DTIC)	Lange, 1983 (35)
		Melphalan (L-PAM)	Robins, 1997 (36)
		Ifosfamide (IFO)	Engelhardt, 1988 (34) Issels, 1990 (37) Westermann, 2003 (38)
Nitrosoureas	Impaired DNA repair Improved pharmacokinetics	BCNU	Parks, 1983 (4) Bull, 1992 (39)
		Me-CCNU	Bull, 1992 (39)
Platinum agents	Impaired DNA repair Altered plasma protein binding	Cisplatin (CDDP)	Parks, 1983 (4) Herman, 1982 (40) Engelhardt, 1990 (41) Robins, 1993 (42) Douwes, 2004 (43)
		Carboplatin (CBDCA)	Westermann, 2003 (38) Hegewisch-Becker, 2002 (44) Hegewisch-Becker, 2003 (45) Douwes, 2004 (43) Richel, 2004 (46) Strobl, 2004 (47)
		Oxaliplatin	Elias, 2004 (48) Hegewisch-Becker, 2002 (44)
Anthracyline antibiotics	Impaired DNA repair Enzyme activation	Adriamycin	Engelhardt, 1990 (41) Bull, 2002 (49)
		Bleomycin	Herman, 1982 (40)
Antimetabolites	Increased drug transport Cell cycle arrest Impaired DNA repair	5-FU	Lange, 1983 (35) Larkin, 1979 (50) Bull, 2002 (49) Hegewisch-Becker, 2002 (44)
		Gemcitabine	Bull, 2004 (51)
Antiproliferatives	Impaired DNA repair	Etoposide (VP-16)	Barlogie, 1979 (52) Issels, 1990 (37) Westermann, 2003 (42)
Topoisomerase inhibitors	Impaired DNA repair	Irinotecan (CPT-11)	Hegewisch-Becker, 2003 (45) Elias, 2004 (48)
Taxanes	Microtubule disruption Apoptosis	Paclitacel Docetaxel	Strobl, 2004 (49) Strobl, 2004 (49)
Biological response modifiers	Increased anti-viral and antiproliferative activity	Interferon	Robins, 1989 (53) Bull, 2002, 2004 (34,49)

[a]References prior to 1980, or not in English, are not provided in the Bibliography at the end of this article.

given during the plateau phase of WBH, 10 min after target temperature was reached (56).

A major rationale for whole-body hyperthermia in cancer treatment is the ability to treat metastases, but this is actually a controversial issue. There have been no clinical studies specifically designed to study the effect of systemic hyperthermia on either the efficacy against metastatic disease or prevention of development of metastases. Increased survival in advanced malignancies is often interpreted to mean a reduction in metastatic disease, but direct measurement of the incidence and response of metastases is rare. Based on some animal studies, it has been suggested that systemic hyperthermia could actually promote the metastatic spread of tumor cells, but this has not been confirmed. One clinical study found an increase of tumor cells in blood 24 h after 41.8 °C WBH, but there was no evidence that this caused metastatic spread of disease (57). Several animal experiments do support the efficacy of whole-body hyperthermia against metastases. In mouse models of lung cancer and melanoma, the number of lung metastases was scored after repeated systemic microwave heating. It was found that the number of lung metastases was significantly reduced, and NK-cell activity was higher, in treated animals. The authors hypothesized that WBH interferes with the spread of organ metastases, possibly through a mechanism involving NK cells (13). Another study of mouse Lewis lung carcinoma in which the animals were treated with 60 min of systemic hyperthermia at 42 °C, demonstrated a reduction in the number and percentage of large metastases (>3 mm) on day 20 post-tumor implantation. Addition of radiation led to a reduction to 50% of control of the number of lung metastases as

well as the percent of large metastases on day 20 (58). In a breast cancer ocult metastasis model in rats, 6 h of 40 °C whole-body hyperthermia combined with daily, low dose, metronomic irinotecan resulted in delayed onset, and reduced incidence, of axillary lymph node metastases compared to control in rats, as did treatment with 40 °C WBH alone. The combination therapy also reduced axillary metastasis volume. Interestingly, none of the therapies significantly affected inguinal lymph node metastases, but lung metastases were decreased in both the combination therapy and WBH alone groups. Rats treated with fever-range whole-body hyperthermia and metronomic irinotecan also survived significantly longer (36%) than control animals (59).

Systemic Hyperthermia and Radiotherapy. The augmentation of ionizing radiation induced tumor kill by hyperthermia is well documented for local hyperthermia and has led to numerous protocols combining whole-body hyperthermia with radiation therapy (60,61). Hyperthermia is complementary to radiation in several regards: ionizing radiation acts predominantly in the M and G_1 phases of the cell cycle while hyperthermia acts largely in S phase; radiation is most effective in alkaline tissues whereas hyperthermic cytotoxicity is enhanced under acidic conditions; radiation is not effective in hypoxic regions yet hyperthermia is most toxic to hypoxic cells. Thus when hyperthermia is combined with radiotherapy, both the hypoxic, low pH core of the tumor is treated as well as the relatively well perfused outer layers of the tumor. Furthermore, because of its vascular effects, hyperthermia enhances tumor oxygenation thus potentiating radiation cell kill. Hyperthermia also increases the production of oxygen radicals by radiation, and reduces the repair of DNA damage caused by ionizing radiation. Thus hyperthermia and radiotherapy together often have a synergistic effect, and this combination is now well accepted for treatment of a number of tumors.

Fever-Range WBH. Like systemic hyperthermia alone, combined modality treatments were initially aimed to achieve maximally tolerated temperatures. Such regimens, however, carry significant risk to the patient, require general anesthesia, and necessitate experienced, specialist personnel to provide careful monitoring of vital signs and patient care during the treatment. More recently, it has been appreciated that lower core body temperatures (39–40 °C) maintained for a longer time (4–8 h), much like fever, can indirectly result in tumor regression through effects on tumor vasculature, the immune response, and waste removal (detoxification). The optimum duration and frequency of mild hyperthermia treatment has, however, not yet been determined. Protocols range from single treatments of 4–6 h, or similar long duration treatments given once during each cycle of chemotherapy, to daily treatments of only 1 h. Several studies of mild, fever-range, whole-body hyperthermia with chemotherapy have demonstrated efficacy against a broad range of cancers (34,17) and clinical trials are currently being conducted at the University of Texas Health Science Center at Houston, Roswell Park Cancer Institute, New York, and by the German Interdisciplinary Working Group on Hyperthermia (62).

Systemic Hyperthermia and Metabolic Therapy. Increased rates of metabolic reactions lead to rapid turnover of metabolites, causing cellular energy depletion, acidosis, and consequent metabolic disregulation. Tumors, which have increased metabolic rates [glucose, adenomine triphosphate (ATP)] compared to normal cells, may be particularly sensitive to thermally induced energy depletion and this has been exploited in the Cancer Multistep Therapy developed by von Ardenne, which is a combined hyperthermia–chemotherapy–metabolic therapy approach to cancer (63). The core of this approach is systemic hyperthermia at 40–42 °C, sometimes with added local hyperthermia to achieve high temperatures within the tumor. A 10% solution of glucose is infused into the patient to achieve a high accumulation of lactic acid within the tumor that cannot be cleared because of sluggish blood flow and confers an increased sensitivity to heat to the tumor cells. Administration of oxygen increases the arterial oxygen pressure and stimulates lysozymal cytolysis. Finally low dose chemotherapy is added.

Palliation. Pain relief is reported by many patients receiving systemic hyperthermia treatment, whether with chemotherapy or radiation. Indeed, almost all patients undergoing thermoradiotherapy report pain relief. Immediate pain relief following treatment is likely to stem from an increased level of circulating β-endorphins, while longer term pain relief may be due to increased blood flow, direct neurological action, and disease resolution, for example, tumor regression in cancer patients, or detoxification. Meaningful improvements in quality of life typically result from such pain relief. Localized infrared therapy using lamps radiating at 2–25 μm is used for the treatment and relief of pain in numerous medical institutes in China and Japan.

Diseases Other than Cancer. Therapeutic use of heat lamps emitting IR radiation is commonplace throughout the Orient for rheumatic, neurological and musculoskeletal conditions, as well as skin diseases, wound healing, and burns. The improvements reported appear to be largely due to increased blood flow bringing nutrients to areas of ischemia or blood vessel damage, and removing waste products. Scientific reports of these treatments are, however, difficult to find. Application of heat via hot baths or ultrasound has long been standard in physical therapy for arthritis and musculoskeletal conditions, though ice packs are also used to counter inflammatory responses. Heat decreases stiffness in tendons and ligaments, relaxes the muscles, decreases muscle spasm, and lessens pain. Unfortunately, few clinical trials of efficacy have been performed, and methodological differences or lack of rigor in the studies hinder comparisons (64). A clinical trial in Japan reported a supposedly successful solution for seven out of seven cases of rheumatoid arthritis treated with whole-body IR therapy, and it is reported that the King of Belgium was cured of his rheumatoid arthritis in three months due IR treatments. Systemic hyperthermia with

whole-body radiant heat units is being carried out in clinical centers as well as many private clinics in Germany for the purpose of alleviating rheumatoid arthritis. It has been proposed that the induction of TNF receptors by WBH may induce a remission in patients with active rheumatoid arthritis. The use of heat packs has long been standard to relieve the pain of fibromyalgia. Again, the therapeutic effect is believed to be due to increased circulation flushing out toxins and speeding the healing process. Whole-body hyperthermia treatment for fibromyalgia and chronic fatigue syndrome (CFS) is to be found in a number of private clinics. Hyperthermia increases the number and activity of white blood cells, stimulating the depressed immune system of the CFS patient.

Because of its immune stimulating effects, whole-body hyperthermia is a strong candidate for treatment of chronic progressive viral infections, such as HIV and hepatitis C. A clinical trial at the University Medical Center Utrecht, The Netherlands has evaluated extracorporeal heating to induce systemic hyperthermia of 41.8 °C for 120 min under propofol anesthesia for treatment of hepatitis C (65). Human immunodeficiency virus (HIV)-infected T cells are more sensitive to heat than healthy lymphocytes, and susceptibility increases when the cells are presensitized by exposure to tumor necrosis factor. Thus, induction of whole-body hyperthermia or hyperthermia specifically limited to tissues having a high viral load is a potential antiviral therapy for acquired immunodeficiency syndrome (AIDS). An Italian study has found treatment of AIDS with beta-carotene and hyperthermia to be synergistic, preventing progression of early disease and also increasing the survival time in patients with severe AIDS. A single treatment of low flow extracorporeal hyperthermia was found effective against AIDS associated Kaposi's sarcoma, though there was significant toxicity. Core temperature was raised to 42 °C and held for 1 h with extracorporeal perfusion and *ex vivo* blood heating to 49 °C. Complete or partial regressions were seen in 20/29 of those treated at 30 days post-treatment, with regressions persisting in 14/29 of those treated at 120 days post-treatment. At 360 days, 4/29 maintained tumor regressions with 1 patient being in complete remission still at 26 months (66).

THE FUTURE OF SYSTEMIC HYPERTHERMIA

While there is a resurgence of interest in systemic hyperthermia, this modality has not yet been adopted as a mainstream therapy, and optimal clinical trials have not yet been carried out. Well-designed, well-controlled, multicenter clinical trials need to be conducted. In order to unequivocally demonstrate the utility of whole-body hyperthermia in the treatment of cancer as well as other diseases, it will be important to accrue a sufficiently large number of patients who do not have end-stage disease. Thanks to the commercial availability of systemic hyperthermia systems, the variability between induction techniques at different institutions can be removed. Newer instrumentation, particularly near-IR radiant heat devices, along with treatment at lower temperatures (fever-range thermal therapy) should lead to significantly reduced toxicity. Better exploitation of the narrow window of effective temperatures within which the cellular effects of heat can be exploited yet damage remains minimal, and improved understanding of the biological interactions *in vivo* of systemic heat with chemotherapy and radiation will be essential to optimize therapy.

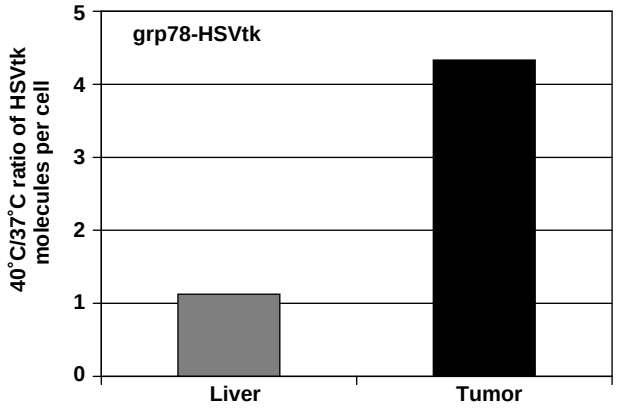


Figure 5. Fever range whole-body hyperthermia increases therapeutic gene (grp78-HSVtk) delivery in tumor.

The effects of systemic hyperthermia on tumor blood flow and vascular permeability have the potential to increase delivery of various small molecules, nanoparticles, and gene therapy vectors to tumors. Ferromagnetic nanoparticles can be heated by external magnetic fields and offer the potential for internal hyperthermia, both locally and systemically. Thermally sensitive liposomes that release their contents at designated temperatures are also of interest. The ability of systemic hyperthermia to aid in systemic delivery of gene therapy vectors (the holy grail of gene therapy) and enhance transfection of cells with therapeutic gene plasmids is under investigation in several laboratories (67,68), and shows potential along with targeted gene therapy via the heat shock response. For example, Fig. 5 shows a fourfold hyperthermic increase of therapeutic gene delivery to tumor when plasmid DNA was injected intravenously into mammary carcinoma bearing rats immediately after 6 h of whole-body hyperthermia at 40 °C. Thus systemic hyperthermia is likely to see increasing application as an enhancer of drug delivery.

There is a great deal of interest in the immunological consequences of whole-body hyperthermia, and as they become better understood, the combination of systemic hyperthermia with specific immunotherapies will undoubtedly be pioneered, not just for cancer but also, by analogy with fever, in a broad range of diseases.

SUMMARY

Systemic hyperthermia is founded on solid physical and biological principles and shows promise in the treatment of a number of diseases. Modern whole-body hyperthermia devices use IR-A radiation sources together with effective heat loss techniques to achieve a controlled, uniform temperature distribution throughout the body with minimal patient toxicity. A shift in paradigm has occurred away from achieving direct cell killing with short

bouts of maximally tolerated temperatures, to inducing indirect curative effects through longer duration treatments at lower temperatures, and synergy with other modalities, such as radiotherapy. Better understanding of the interactions of elevated temperature with metabolic and genetic pathways will allow thermally driven targeted therapies. Of particular promise is the use of systemic hyperthermia as an immune system stimulator and adjunct to immunotherapy. Application of systemic hyperthermia to nanoparticle delivery and gene therapy is emerging. Whole-body hyperthermia is moving from being dubbed an alternative therapy to becoming a standard treatment and clinical hyperthermia centers are to be found all over the world.

Useful Websites

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hstat6.section.40680	Techniques and Devices Used to Produce Hyperthermia
http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=cmed.section.7813	Physics and Physiology of Heating
http://www.duke.edu/~dr3/hyperthermia_general.html	Clinical Hyperthermia Background
http://www.eurekah.com/isbn.php?isbn=1-58706-248-8&bookid=143&catid=50	Online book: *Locoregional Radiofrequency-perfusional and Whole Body Hyperthermia in Cancer Treatment: New Clinical Aspects*, E.F. Baronzio and A. Gramaglia (eds.), Eurekah Bioscience Database
http://www.esho.info/professionals/	European Society for Hyperthermic Oncology
http://www.hyperthermie.org/index2.html	German Interdisciplinary Working group on hyperthermia
http://www.uth.tmc.edu/thermaltherapy/	Systemic Thermal Therapy at the University of Texas

BIBLIOGRAPHY

1. Nauts HC. Bacterial pyrogens: beneficial effects on cancer patients. In: Gautherie M, Albert E, editors. Biomedical Thermology, Progress in Clinical Biological Research. New York: Alan R. Liss; 1982. p 687–696.
2. Law HT, Pettigrew RT. Heat transfer in whole body hyperthermia. Ann NY Acad Sci 1980;335:298–310.
3. Robins HI, et al. A non-toxic system for 41.8 °C whole body hyperthermia: results of a phase I study using a radiant heat device. Cancer Res 1985;45:3937–3944.
4. Parks LC, Smith GV. Systemic hyperthermia by extracorporeal induction techniques and results. In: Storm FK, editor. Hyperthermia in Cancer Therapy. Boston: Hall; 1983. pp 407–446.
5. van der Zee J, van Rhoon GC, Faithfull NS, van den Berg AP. Clinical hyperthermic practice: whole body hyperthermia. In: Field SB, Hand JW, editors. An Introduction to the Practical Aspects of Clinical Hyperthermia. London: Taylor & Francis; 1990.
6. Wust P, et al. Feasibility and analysis of thermal parameters for the whole-body-hyperthermia system IRATHERM-2000. Int J Hyperthermia 2000 Jul–Aug; 16(4):325–339.
7. Blazickova S, Rovensky J, Koska J, Vigas M. Effect of hyperthermic water bath on parameters of cellular immunity. Int J Clin Pharmacol Res 2000;20(1–2):41–46.
8. Vaupel P, Kallinowski F. Physiological effects of hyperthermia. In: Streffer C, editor. Hyperthermia and the Therapy of Malignant Tumors. Berlin/Heidelberg, Germany: Springer-Verlag; 1987.
9. Dudar TE, Jain RK. Differential response of normal and tumor microcirculation to hyperthermia. Cancer Res 1984;44(2):605–612.
10. Kong G, Braun RD, Dewhirst MW. Characterization of the effect of hyperthermia on nanoparticle extravasation from tumor vasculature. Cancer Res 2001;61:3027–3032.
11. Jain RK. Normalization of tumor vasculature: an emerging concept in antiangiogenic therapy. Science 2005;307(5706): 58–62.
12. Urano M, et al. Effect of whole body hyperthermia on cell survival, metastasis frequency, and host immunity in moderately and weakly immunogenic murine tumors. Cancer Res 1983;43:1039–1043.
13. Shen RN, et al. Whole-body hyperthermia decreases lung metastases in lung tumor-bearing mice, possibly via a mechanism involving natural killer cells. J Clin Immunol 1987;7(3):246–253.
14. Roigas J, Wallen ES, Loening SA, Moseley PL. Heat shock protein (HSP72) surface expression enhances the lysis of a human renal cell carcinoma by IL-2 stimulated NK cells. Adv Exp Med Biol 1998;451:225–229.
15. Burd R, et al. Tumor cell apoptosis, lymphocyte recruitment and tumor vascular changes are induced by low temperature, long duration (fever-like) whole body hyperthermia. J Cell Physiol 1998;177(1):137–147.
16. Ostberg JR, Gellin C, Patel R, Repasky EA. Regulatory potential of fever-range whole body hyperthermia on Langerhans cells and lymphocytes in an antigen-dependent cellular immune response. J Immunol 2001;167(5):2666–2670.
17. Kraybill WG, et al. A phase I study of low temperature, long duration (fever-range) whole body hyperthermia (WBH) in patients with advanced solid tumors: correlation with mouse models. Int J Hyperthermia 2002;18:253–266.
18. Atanackovic D, et al. 41.8 degrees C whole body hyperthermia as an adjunct to chemotherapy induces prolonged T cell activation in patients with various malignant diseases. Cancer Immunol Immunother Epub 2002 Oct 18 2002;51(11–12):603–613.
19. Barni S, et al. Lysosomal exocytosis induced by hyperthermia: a new model of cancer death. III. effect on liver metastasis. Biomed Pharmacother 1996;50:79–84.
20. Hou K. The level changes of peripheral blood PGE_2 and clinical significance in patients with malignant tumor before and after whole body hyperthermia. Proceedings of the 26th Congress of the International Clinical Hyperthermia Society, September 9–12, 2004, Shenzhen, China, 2004; p 66.
21. Manjili MH, et al. Subjeck, Cancer immunotherapy: stress proteins and hyperthermia. Int J Hyperthermia 2002;18(6): 506–520.
22. Ialenti A, et al. Inhibition of cyclooxygenase-2 gene expression by the heat shock response in J774 murine macrophages. Eur J Pharmacol 2005;509(2–3):89–96.

23. Repasky EA, Issels R. Physiological consequences of hyperthermia: Heat, heat shock proteins and the immune response. Int J Hyperthermia 2002;18:486–489.
24. Ostberg JR, Repasky EA. Emerging evidence indicates that physiologically relevant thermal stress regulates dendritic cell function. Cancer Immunol Immunother [Epub ahead of print]; Apr 28, 2005.
25. Berry JM, Michalsen A, Nagle V, Bull JM. The use of esmolol in whole-body hyperthermia: cardiovascular effects. Int J Hyperthermia 1997;13(3):261–268.
26. Robins HI, et al. Adjunctive therapy (whole body hyperthermia versus Ionidamine) to total body irradiation for the treatment of favorable B-cell neoplasms: a report of two pilot clinical trials and laboratory investigations. Int J Radiat Onco Biophys 1990;18:909–920.
27. Ohno S, et al. Haematological toxicity of carboplatin and cisplatin combined with whole body hyperthermia in rats. Br J Cancer 1993;68:469–474.
28. Haveman J, et al. Effects of hyperthermia on the peripheral nervous system: a review. Int J Hyperthermia 2004;20(4):371–391.
29. Calvert AH, et al. Early clinical studies with cis-diammine-1,1-cyclobutane dicarboylate platinum II. Cancer Chemother Pharmacol 1982;9:140–147.
30. Sapareto SA, Dewey WC. Thermal dose determination in cancer therapy. Int J Radiat Oncol Biol Phys 1984;10(6):787–800.
31. Thrall DE, et al. Using units of CEM 43 degrees C T90, local hyperthermia thermal dose can be delivered as prescribed. Int J Hyperthermia 2000;16(5):415–428.
32. Webb AG. Temperature measurement using nuclear magnetic resonance. Ann Reports NMR Spectrosc 2001;45:1–67.
33. Stafford RJ, Hazle JD, Glover GH. Monitoring of high-intensity focused ultrasound-induced temperature changes *in vitro* using an interleaved spiral acquisition. Magn Reson Med 2000;43(6):909–912.
34. Engelhardt R. Summary of recent clinical experience in whole-body hyperthermia combined with chemotherapy. Recent Results Cancer Res 1988;107:200–224.
35. Lange J, Zanker KS, Siewert JR, Eisler K, Landauer B, Kolb E, Blumel G. and Remy, W. Extracorporeally induced whole-body hyperthermia in conventionally incurable malignant tumor patients Med Wochenschr. 1983;108(13):504–509.
36. Robins HI, et al. Phase I clinical trial of melphalan and 41.8 degrees C whole-body hyperthermia in cancer patients. J Clin Oncol 1997;15:158–164.
37. Issels RD, Wilmanns W, editors. Recent Results in Cancer Research, Vol. 107: Application of Hyperthermia in the Treatment of Cancer. Berlin/Heidelberg: Springer Verlag; 1988.
38. Westermann AM, et al. Systemic Hyperthermia Oncologic Working Group, A Systemic Hyperthermia Oncologic Working Group trial, Ifosfamide, carboplatin, and etoposide combined with 41.8 degrees C whole-body hyperthermia for metastatic soft tissue sarcoma. Oncology 2003;64(4):312–321.
39. Bull JM, et al. Chemotherapy resistant sarcoma treated with whole body hyperthermia (WBH) combined with 1-3-bis(2-chloroethyl)-1-nitrosourea (BCNU). Int J Hyperthermia 1992; 8(3):297–304.
40. Herman TS, Sweets CC, White DM, Gerner EW. Effect of heating on lethality due to hyperthermia and selected chemotherapeutic drugs. J Natl Cancer Inst 1982;68(3):487–491.
41. Engelhardt R, et al. Treatment of disseminated malignant melanoma with cisplatin in combination with whole-body hyperthermia and doxorubicin. Int J Hyperthermia 1990;6(3): 511–515.
42. Robins HI, et al. Phase I clinical trial of carboplatin and 41.8 degrees C whole-body hyperthermia in cancer patients. J Clin Oncol 1993;11:1787–1794.
43. Douwes F, et al. Whole-body hyperthermia in combination with platinum-containing drugs in patients with recurrent ovarian cancer. Int J Clin Oncol 2004;9(2):85–91.
44. Hegewisch-Becker S, et al. Whole-body hyperthermia (41.8 °C) combined with bimonthly oxaliplatin, high-dose leucovorin and 5-fluorouracil 48-hour continuous infusion in pretreated metastatic colorectal cancer: a phase II study. Ann Onc 2002;13(8):1197–1204.
45. Hegewisch-Becker S, et al. Whole body hyperthermia (WBH, 41.8 °C) combined with carboplatin and etoposide in advanced biliary tract cancer. Proc Am Soc Clin Oncol 2003; (abstr. 1247), 22:311.
46. Richel O, et al. Phase II study of carboplatin and whole body hyperthermia (WBH) in recurrent and metastatic cervical cancer. Gynecol Oncol 2004;95(3):680–685.
47. Strobl B, et al. Whole body hyperthermia combined with carboplatin/paclitaxel in patients with ovarian carcinoma–Phase-II-study. J Clin Oncol -Proc Am Soc Clin Oncol 2004;22(14S).
48. Elias D, et al. Heated intra-operative intraperitoneal oxaliplatin plus irinotecan after complete resection of peritoneal carcinomatosis: pharmacokinetics, tissue distribution and tolerance. Ann Oncol 2004;15:1558–1565.
49. Bull JM, et al. Phase I study of long-duration, low-temperature whole-body hyperthermia (LL-WBH) with liposomal doxorubicin (Doxil), 5-fluorouracil (5-FU), & interferon-α (IFN-α), Proc Amer. Soc Clin Oncol 2002;(Abst. 2126).
50. Larkin JM, A clinical investigation of total-body hyperthermia as cancer therapy. Cancer Res, 1979;39(6 Pt 2):2252–2254.
51. Bull JM, et al. Update of a phase I clinical trial using fever-range whole-body hyperthermia (FR-WBH) + cisplatin (CIS) + gemcitabine (GEM) + metronomic, low-dose interferon-alpha (IFN-alpha), Proceedings of the International Congress on Hyperthermic Oncology, 20.-24.04., St. Louis, (Session 21 and Poster 689); 2004.
52. Barlogie B, Corry PM, Lip E, Lippman L, Johnston DA, Tenczynski TF, Reilly E, Lawson R, Dosik G, Rigor B, Hankenson R, Freireich EJ Total-body hyperthermia with and without chemotherapy for advanced human neoplasms. Cancer Res 1979;39(5):1481–1489.
53. Robins HI, et al. Phase I trial of human lymphoblastoid interferon with whole body hyperthermia in advanced cancer. Cancer Res 1989;49(6):1609–1615.
54. Gerad H, van Echo DA, Whitacre M, Ashman M, Helrich M, Foy J, Ostrow S, Wiernik PH, Aisner J. Doxorubicin, cyclophosphamide, and whole body hyperthermia for treatment of advanced soft tissue sarcoma. Cancer. 1984 Jun 15;53(12):2585–91.
55. Baba H, et al. Increased therapeutic gain of combined cis-diamminedichloroplatinum (II) and whole body hyperthermia therapy by optimal heat/drug scheduling. Cancer Res 1989;49(24 Pt. 1):7041–7044.
56. Katschinski DM, et al. Optimization of chemotherapy administration for clinical 41.8 degrees C whole body hyperthermia. Cancer Lett 1997;115(2):195–199.
57. Hegewisch-Becker S, et al. Effects of whole body hyperthermia (41.8 degrees C) on the frequency of tumor cells in the peripheral blood of patients with advanced malignancies. Clin Cancer Res 2003;9(6):2079–2084.
58. Teicher BA, Holden SA, Ara G, Menon K. Whole-body hyperthermia and lonidamine as adjuvant therapy to treatment with cisplatin with or without local radiation in mouse bearing the Lewis lung carcinoma. Int J Hyperthermia 1995;11(5):637–645.
59. Sumiyoshi K, Strebel FR, Rowe RW, Bull JM. The effect of whole-body hyperthermia combined with 'metronomic' chemotherapy on rat mammary adenocarcinoma metastases. Int J Hyperthermia 2003;19(2):103–118.
60. Overgaard J, et al. Randomised trial of hyperthermia as adjuvant to radiotherapy for recurrent or metastatic

malignant melanoma, European Society for Hyperthermic Oncology. Lancet 1995;345(8949):540–543.
61. Hehr T, Wust P, Bamberg M, Budach W. Current and potential role of thermoradiotherapy for solid tumours. Onkologie 2003;26(3):295–302.
62. Hildebrandt B, et al. Current status of radiant whole-body hyperthermia at temperatures > 41.5 degrees C and practical guidelines for the treatment of adults. The German Interdisciplinary Working Group on Hyperthermia. Int J Hyperthermia 2005;21(2):169–183.
63. Hildebrandt B, et al. Whole-body hyperthermia in the scope of von Ardenne's systemic cancer multistep therapy (sCMT) combined with chemotherapy in patients with metastatic colorectal cancer: a phase I/II study. Int J Hyperthermia 2004;20(3):317–333.
64. Robinson V, et al. Thermotherapy for treating rheumatoid arthritis. Cochrane Database Syst Rev 1: CD002826; 2002.
65. van Soest H, van Hattum J. New treatment options for chronic hepatitis C. Adv Exp Med Biol 2003;531:219–226.
66. Pontiggia P, Rotella GB, Sabato A, Curto FD. Therapeutic hyperthermia in cancer and AIDS: an updated survey. J Environ Pathol Toxicol Oncol 1996;15(2–4):289–297.
67. Li CY, Dewhirst MW. Hyperthermia-regulated immunogene therapy. Int J Hyperthermia 2002;18(6):586–596.
68. Okita A, et al. Efficiency of lipofection combined with hyperthermia in Lewis lung carcinoma cells and a rodent pleural dissemination model of lung carcinoma. Oncol Rep 2004;11:1313–1318.

Further Reading

Bakhshandeh A, et al. Year 2000 guidelines for clinical practice of whole body hyperthermia combined with cytotoxic drugs from the University of Lübeck and the University of Wisconsin. *J Oncol Pharm Practice* 1999;5(3):131–134.
Field SB, Hand JW, editors. An Introduction to the Practical Aspects of Clinical Hyperthermia. London: Taylor & Francis; 1990.
Gautherie M, editor. Methods of External Hyperthermic Heating. Berlin/Heidelberg: Springer Verlag; 1990.
Gautherie M, editor. Whole Body Hyperthermia: Biological and Clinical Aspects. Berlin/Heidelberg: Springer Verlag; 1992.
Hahn GM. Hyperthermia and Cancer. New York: Plenum Press, 1982.
Hildebrandt B, et al. Current status of radiant whole-body hyperthermia at temperatures > 41.5 degrees C and practical guidelines for the treatment of adults. The German Interdisciplinary Working Group on Hyperthermia. *Int J Hyperthermia* 2005;21(2):169–183.
Issels RD, Wilmanns W, editors. Recent Results in Cancer Research, Vol. 107: Application of Hyperthermia in the Treatment of Cancer. Berlin/Heidelberg: Springer Verlag; 1988.
Nussbaum GH, editor. Physical Aspects of Hyperthermia. American Association of Physicists in Medicine Medical Physics Monograph No. 8. New York: American Institute of Physics; 1982.
Guan J, et al. The clinical study of whole-body hyperthermia (WBH) improving the survival state of patients with advanced cancer. Proc 26th Congress of the International Clinical Hyperthermia Society, Sept. 9–12, 2004, Shenzhen, China; 2004; p 66.
Kurpeshev OK, Tsyb AF, Mardynsky YS. Whole-body hyperthermia for treatment of patients with disseminated tumors- Phase II. In: P.H. Rehak, K.H. Tscheliessnigg, editors. Proceedings 22nd. Annual Meeting of the European Society for Hyperthermic Oncology, June 8–11, 2005, Graz, Austria, 2005; p 103.
Hou K. Assessment of the effects and clinical safety of the treatment of advanced malignant tumor with extracorporeal whole body hyperthermia. Proceedings of the 26th Congress of the International Clinical Hyperthermia Society, Sept. 9–12, 2004, Shenzhen, China; 2004 p 71.

See also BIOHEAT TRANSFER; HEAT AND COLD, THERAPEUTIC; HYPERTHERMIA, INTERSTITIAL; HYPERTHERMIA, ULTRASONIC; RADIATION DOSIMETRY FOR ONCOLOGY.

HYPERTHERMIA, ULTRASONIC

DIMPI PATEL
DHANUNJAYA LAKKIREDDY
ANDREA NATALE
The Cleveland Clinic Foundation
Cleveland, Ohio

INTRODUCTION

The use of elevated temperature as a form of medical treatment has been fairly ubiquitous across cultures throughout the course of time. The earliest record of heat for therapeutic use was found in an Egyptian surgical papyrus dated to 3000 BC (1). Hippocrates, considered by many to be the father of medicine, used heat to treat breast cancer. He based his practice of medicine on an ancient Greek ideology that advises using heat after trials of surgery and medications have failed (2). German physicians in the 1800s noted cases where cancer patients had developed high fevers secondary to infections that resulted in a miraculous disappearance of their tumors (3). These observations provided inspiration for the development of several techniques that attempted to induce hyperthermia. One such popular method entailed wrapping a patient's body in plastic and then dipping him in hot wax. Another popular technique involved removing a portion of the patient's blood, heating it, and then transfusing the warmed blood back to the patient's body, thereby creating systemic hyperthermia (4). These treatments had varied success rates, often culminating in fatality, and were subsequently discarded. Thus, the interest in hyperthermia lessened in the face of more conventional cancer treatments (e.g., chemotherapy and radiation). The current revival of interest in hyperthermia has resulted from a combination of clinicians searching for a therapeutic mode other than chemotherapy and radiation, in tandem with several preliminary randomized clinical trials in a small selected group of patients that have shown marked improvement in disease states with the use of either hyperthermia alone or particularly as an adjuvant to other more traditional modalities.

Traditionally, conventional hyperthermia has been defined as a therapeutic elevation of whole body temperature or target tissue while maintaining low enough temperatures to avoid tissue coagulation (3). This definition of hyperthermia can be broadened to include the therapeutic elevation of temperature to cause tissue destruction and coagulation, such as that implemented in HIFU (high intensity focus ultrasound) procedures. Classically, microwaves, radio frequency (RF), electromagnetic radiations, or ultrasounds have been used to heat tissue to 40–44 °C (5). This article compares and contrasts electromagnetic waves to ultrasonic waves as a heating modality, explain the physics behind ultrasound

generation, and explores the thermal and mechanical biophysics involved with ultrasound delivery to tissue. Then, the medical fields that are currently benefitting from conventional ultrasound hyperthermia and HIFU are considered, and finally some of the many applicators involved with thermal ultrasound delivery are evaluated.

ULTRASOUND VERSUS ELECTROMAGNETIC RADIATION

Electromagnetic waves were often used in various applications of conventional hyperthermia treatments. However, ultrasound has emerged as a better option because of its shorter wavelength and lower energy absorption rate, which make it easier to control and to localize the area that is being heated. For example, for a half-power penetration depth of 4 cm, the ultrasound wavelength in tissues (e.g., muscle) is 1 mm; however, electromagnetic wavelength required for the same transmission is 500 mm. Focusing energy into a volume smaller than a wavelength is generally not possible. Using 500 mm ($\sim$40 MHz) of electromagnetic waves to heat a tumor that is situated 4 cm below the skin with proportions of 6 cm in diameter in the upper abdomen results in a temperature elevation of the entire abdomen including the spleen, liver, and all major vessels. More than one-half of the body's cardiac output circulates through the abdominal area, and this widespread heating results in a systemic elevation of temperature, thereby limiting the use of electromagnetic radiation for tumors in the body cavity (3). Electromagnetic waves are currently limited to regional hyperthermia and treating very superficial tumors (6). Conversely, ultrasound that has a wavelength of 1 mm can be focused within the area of the tumor, thus allowing less energy to be radiated to other areas of the body, resulting in less damage to surrounding healthy tissue. The current fabrication technology allows for practical applicator dimensions and multiple transducer configurations that makes it possible to control and shape a wide variety of ultrasound beams. The use of focused transducers or electronically phased arrays allow for better localization and temperature control of the target tissue (7). In contrast to these positive attributes, high acoustic absorption at bone–soft tissue interface and reflection from gas surfaces may make certain therapeutic scenarios difficult.

GENERATION AND PROPAGATION OF ULTRASOUND

Ultrasonic Transducers

In order to generate ultrasonic waves for tissue warming, a transducer containing piezoelectric crystals is required. Piezoelectric crystals are found in Nature or can be artificially grown. Quartz and synthetic ferroelectric ceramics (e.g., lead metanobiate, lead zirconate, and titanates of barium) all have strong piezoelectric properties (8). The ceramic most commonly used in the fabrication of ultrasound transducers is synthetic plumbium zirconium titante (PZT). Transducers are manufactured by applying an external voltage to these ferroelectric materials to orient their internal dipole structure. They are then cooled to permanently maintain their dipole orientation. Finally, they are cut into any desired shape, such as spherical bowls for focused ultrasonic fields (3,8). Ultrasound transducers have electrodes attached to the front and back for application and detection of electrical fields. With the application of an alternating electrical field parallel to the surface of piezoelectric material, the crystals will contract and vibrate for a short time with their resonant frequency. The frequency at which the transducer is able to vibrate is indirectly proportional to its thickness; higher frequencies are a result of thinner transducers, lower frequencies a result of thicker transducers (8).

Piezoelectric crystals are able to contract or expand when an electrical field is applied to them because dipoles within the crystal lattice will realign themselves as a result of attractive and repulsive forces causing a change in physical dimension of the material in the order of nanometers (electrostriction or reverse piezoelectric effect). When echos are received, the ultrasound waves will compress and expand the crystals (8). This mechanical stress causes the dipoles to realign on the crystal surface creating a net charge (piezoelectric effect) (Fig. 1).

Transducers function optimally when there is broad bandwidth in the frequency domain and short impulse response in the time domain. Also, when there is little electroacoustical conversion inefficiency, and little mismatch between the electrical impedances of the generator and the transducer (3,9). A transducer's ability to transmit energy is dependent on the characteristics of acoustic impedances and its contact medium. Both the density of the material and propagation velocity of ultrasound waves will determine its impedance. When both impedances match, then less energy is lost through reflection back into the transducer. For example, at the interface between air and the transducer, most of the energy will be reflected back to the transducer and will travel to the opposite direction because air has $\sim$16 times less impedance than the transducer. If the transducer is a half wavelength in thickness, the reflected wave arrives at the opposite surface in phase with the direction of its motion and can then be transmitted into the medium. Since the efficiency at which a transducer transmits energy has a direct relationship to the degree of impedance match, efficiency can be increased significantly by adding an impedance matching layer of a quarter wavelength thickness, subsequently making the characteristic impedance equal to the geometric average of those of the transducer and the loading medium (3,8).

RADIATION FIELD OF ULTRASONIC TRANSDUCERS

The radiation field of an ultrasonic transducer depends on its physical properties and the transmission characteristics of the medium through which it will pass. Conventional planar transducers create a nonfocused field, whereas some modifications to the same transducer can create a focused field.

NONFOCUSED FIELDS

Planar Transducers

Ultrasonic waves that are radiated from the transducer surface can be described as a tightly packed array of

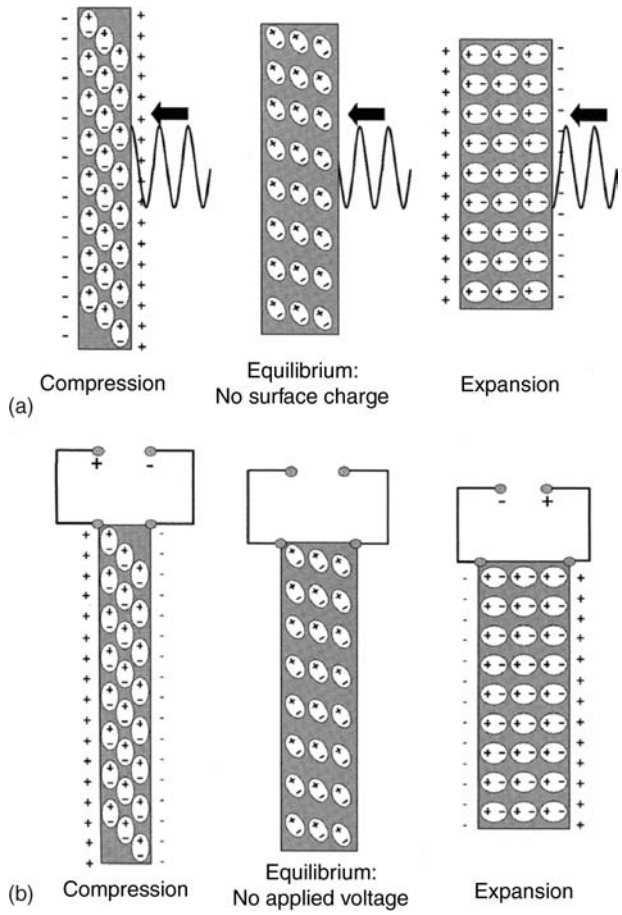


Figure 1. A piezoelectric compound consists of aligned molecular dipoles. (a) At equilibrium, the dipoles are arranged in a configuration that results in a neutral net charge. When the piezoelectric compound is mechanically stressed (e.g., an ultrasound wave) the element changes its physical dimensions. At peak pressure amplitudes, the element will contract. When no stress is placed upon the element it is in equilibrium. At peak rarefaction, the element will expand. This realignment of dipoles results in the production of a net positive or negative surface charge. (b) When an electrical field is applied to the piezoelectric element the dipoles can be realigned in response to attractive or repulsion forces. This rearrangement results in either expansion or contraction of the element. In the absence of an applied electrical field the element is in equilibrium and has a neutral net charge. (Published with the permission from Ref. 8).

separate point sources of sound energy (Fig. 2a). Each of these points emits a spherical wavelet (Fig. 3). These waves interact both constructively and destructively creating a diffraction pattern. Any point in the medium is a compilation of all the sources that reach that target at that period of time. This diffraction pattern can be calculated using Huygen's principle. Two separate transducers whose emission fields interact in the same media are subject to the same laws of construction and destruction. Planar transducers operating in continuous wave mode are shown in (Fig. 2). In the Fresnel zone or the near field, the beam energy distribution is collimated, which is a result of the many destructive and constructive interactions of the spherical wavelets (Figs. 2c and 3). The

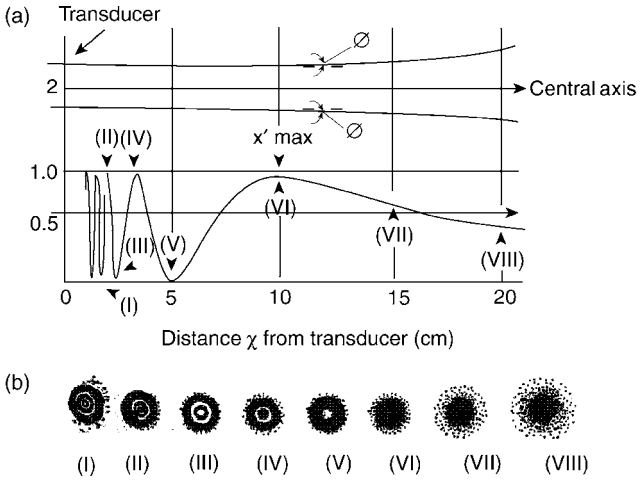


Figure 2. A planar transducer operating in a continuous wave mode. (a) The envelope containing almost all of the ultrasonic energy. (b) The relative intensity of the ultrasonic beam along a central axis. The intensity across the field fluctuates greatly at small distances from the surface of the transducer. At greater distances along the central axis the intensity distribution across the field stabilizes and deteriorates with beam divergence. (c) Ring diagrams illustrating the energy distribution at positions indicated in (b). In the near field, the ring beam pattern is collimated, but at greater distances from the transducer surface the beam diverges. (Published with permission from Ref. 3).

beam path is a function of the dimension of the active part of the transducer surface, thus the beam diameter that is converging at the end of the near field is approximately one-half of the size of transducer diameter. The intensity and pressure amplitudes fluctuate greatly at small distances from the surface transducer (Fig. 2b). As the distance from the transducer surface increases, the beam path diverges (Fig. 2c and 3). In large diameter, high frequency transducers, there is less beam divergence in the far field. After a certain distance from the transducer surface, the intensity stabilizes; however, intensity along the axis deteriorates along with beam divergence (Fig. 2b). Circular, square, and rectangular transducers have similar fields; albeit, circular transducers have more pronounced fluctuations of intensity in the near field (3,8).

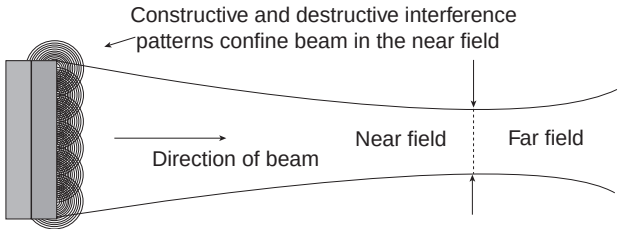


Figure 3. Ultrasonic waves that are radiated from the transducer surface are described as a tightly packed array of separate point sources. Each of these points radiates a spherical wavelet (Huygen's principle). These spherical wavelets will interact constructively and destructively. In the near field, these interactions result in a convergent beam pattern. In the far field, the beam pattern diverges. (Published with permission from Ref. 8).

FOCUSED FIELDS

Single-Element Transducers

When an ultrasonic wave travels through different media, the laws of geometric optics can be applied. Ultrasonic waves can be reflected, refracted, and scattered. When there is a high degree of impedance mismatch between the generator and transducer, ultrasonic waves will be reflected back into the transducer. The angle of reflection is equal to the angle of impedance, much like that of a mirror. Single element transducers can be focused by using a curved acoustic lens or a curved piezoelectric element (8). When an ultrasonic wave goes through two media with different propagation velocities there is a certain degree of refraction. Ultrasonic propagation through water is 1500 $m \cdot s^{-1}$. In order to focus the ultrasound field, a lens of plastic (e.g., polystyrene), which has a higher propagation velocity, is placed between the transducer and the water media; these converging lenses are then concave to the water media and at plane with the transducer interface. In an unfocused transducer, the focal length is directly proportional to the transducer frequency and diameter. In a focused single element transducer, the focal distance is brought closer to the transducer surface. The focal distance is defined as the distance between the transducer surface and the portion of the beam that is narrowest. The focal zone, which is the area of best lateral resolution, is defined as the area at which the width of the beam is less than two times the width at the focal distance (3,8) (Fig. 4). The focal zone is dependent on the aperture and the wavelength of the ultrasound. The focal area through which 84% of the ultrasound passes is two to four wavelengths in hyperthermia systems. With ultrasonic transducers, the intensity distribution dimensions are a function of frequency and aperture. Therefore, the larger the aperture, the shorter the focal region, the higher the frequency and the smaller the diameter of the beam (Fig. 5). Ceramic curved bowl-shaped transducers, while more efficient than a lens, do not have the versatility of a lens. Once a ceramic bowl is fabricated, the focal length is set. Lenses can be interchanged creating a variety of focal lengths with one transducer (8).

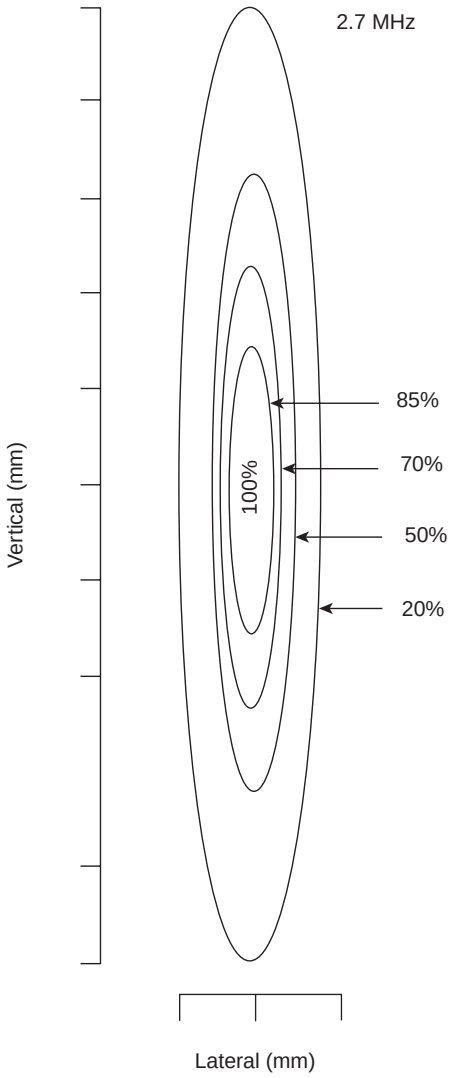


Figure 5. The intensity distribution in the focus of an ultrasound transducer. The diameter and the length are a function of frequency. The lower the frequency, the larger the diameter, and the smaller the aperture, the longer the focal region. (Published with permission from Ref. 3).

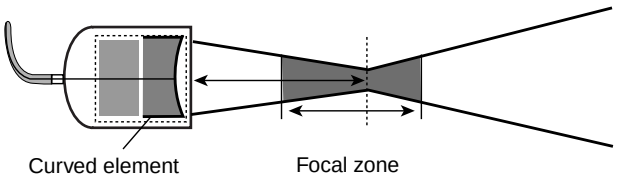


Figure 4. The focal zone is the area of optimal lateral resolution. The use of a curved element or an acoustic lens allows the focal distance to be brought closer to the transducer surface. The use of a curved element decreases the beam diameter at the focal distance and increases the angle of beam divergence far field. The focal zone, which is the are of best lateral resolution, is defined as the area at which the width of the beam is less than two times the width at the focal distance. (Published with permission from Ref. 8).

Multielement Transducers

Linear array transducers contain 256–512 narrow piezoelectric elements. They produce a beam by firing a portion of the total number of elements as a group. If a single element were fired the beam pattern would be divergent in the near field. By firing a group of elements, it is possible to focus and converge the beam. All the individual beams interact both constructively and destructively to produce a collimated beam. A phase array transducer is composed of 64–128 elements. The ultrasound beam can be steered and focused without moving the transducer by electrically activating the separate elements on the transducer surface at slightly different times (8).

Table 1. Acoustic Impedance[a]

Tissue	Z, rayls
Air	0.0004×10^6
Lung	0.18×10^6
Fat	1.34×10^6
water	1.48×10^6
Kidney	1.63×10^6
Blood	1.65×10^6
Liver	1.65×10^6
Muscle	1.71×10^6
Skull bone	7.8×10^6

[a]Measured in rayls. $z = pc$ (z = impedance, p = sound pressure, c = speed of sound), for air, water, and selected tissues. (Published with permission from Ref. (8).

Table 2. Density and Speed of Sound in Tissues and Materials for Medical Ultrasound.[a]

Material	Density, $kg \cdot m^{-3}$	c, $m \cdot s^{-1}$	c, $mm \cdot s^{-1}$
Air	1.2	330	0.33
Lung	300	600	0.60
Fat	924	1450	1.45
water	1000	1480	1.48
Soft tissue	1050	1540	1.54
Kidney	1041	1565	1.57
Blood	1058	1560	1.56
Liver	1061	1555	1.55
Muscle	1068	1600	1.60
Skull bone	1912	4080	4.08
PZT[b]	7500	4000	4.00

[a]Published with permission from Ref. 8.
[b]PZT, lead cisconate .inanate

PROPAGATION OF ULTRASOUND IN BIOLOGICAL TISSUES

The journey of an ultrasound wave through human tissue is sometimes arduous. As the waves propagate through different biological media, they are subject to reflection, refraction, scattering, and absorption (8). When there is a difference in acoustic impedance between the boundaries of two tissues, reflection occurs (Table 1). There is 100% reflection at the air–skin interface. However, if a coupling medium (e.g., gel) is used, reflection is reduced to 0.1%. When the beam is not perpendicular to tissue boundary, the transmitted ultrasound energy undergoes a change in direction at the boundary; this directional change is termed refraction. As waves propagate through tissue they must overcome internal friction resulting in a loss of energy. The mechanical energy that is lost is converted to heat, which is termed absorption. At higher frequencies, ultrasonic waves move quicker, thus forcing the molecules to move against each other creating friction. The more these molecules move, the more energy is consumed or absorbed, and subsequently will be converted to heat. The speed at which the ultrasonic wave travels is dependent on the arrangement of the molecules. If they are densely arranged, they will collide sooner when they are exposed to a stimulus, and will lose energy quickly and at shorter distances. Ultrasonic waves can travel through the skin without much absorption until they reach tissues with high collagen content, (e.g., bone, periosteum, ligaments, capsules, fascia, and tendons). Ultrasonic waves travel through most solid tissues at a speed of 1500–1600 $m \cdot s^{-1}$. Its velocity in fat and postmenopausal breast tissue may be as low as 1450 $m \cdot s^{-1}$ and the lens of the eye $\sim$1600 $m \cdot s^{-1}$. As a general rule, ultrasonic waves move through soft tissue with relatively little reflection or refraction (3,8) (Table 2).

Ultrasonic speed through bone is $\sim$4080 $m \cdot s^{-1}$. Bone easily absorbs ultrasonic energy and reflects it to tissues that are located at the bone–tissue interface. Since bone absorbs ultrasonic energy so readily, it heats up very quickly, consequently making it harder to control temperature. Thus, bone and its immediate surrounding tissue were once considered problematic for therapeutic use of ultrasonic hyperthermia (3,8). Nevertheless, in a recent study, Moros et al. noted that the presence of underlying bone in superficial unfocused ultrasound hyperthermia could actually be exploited to induce more uniform and enhanced temperature distributions in superficial target volumes. In particular, they have shown that the presence of bone in superficial target volumes enhances temperature elevation not only by additional direct power deposition from acoustic reflection, but also from thermal diffusion from the underlying bone (10).

The intensity at which an ultrasonic beam is transmitted has an effect on target tissue temperature. Intensity is defined as the amount of power per unit area. Doubling the amount of power used will result in quadrupling the intensity. Ultrasonic waves will lose intensity as they propagate further into the tissue. The attenuation coefficient is the relative intensity loss per centimeter of travel in a given medium (Table 3). Beam divergence, absorption, and scattering will also cause a loss in intensity of the propagating beam. The absorption coefficient of the tissue being exposed to us determines the target temperature that tissue will reach. The absorption coefficient is dependent on the density of the tissue and will linearly increase at higher frequencies. The absorption coefficient in soft tissue is 4–10 times lower than that of bone, and therefore bone heats more quickly (3). At short exposure times (e.g., < 0.1 s), temperature and intensity are directly propor-

Table 3. Attenuation Coefficients for Selected Tissues at 1 MHz.[a]

Tissue Composition	Attenuation Coefficient 1 MHz beam, $dB \cdot cm^{-1}$
Water	0.0002
Blood	0.18
Soft tissues	0.3–0.8
Brain	0.3–0.5
Liver	0.4–0.7
Fat	0.5–1.8
Smooth muscle	0.2–0.6
Tendon	0.9–1.1
Bone, cortical	13–26
Lung	40

[a]Published with permission from Ref. 8.

tional. However, as the time intervals increase, other factors in addition to intensity, (e.g., blood perfusion), must be considered. An approximate estimate of the ultrasonic energy requirements for heating a target volume to therapeutic temperature depends on assessing thermophysical properties of that tissue, intensity of the ultrasound beam, ultrasonic absorption coefficient, and additional factors (e.g., blood circulation to target tissue) (3,8,11). The thermal index is defined as the ratio of acoustic power created by the transducer to raise the target area by 1 °C. This is calculated by using an algorithm that takes into account the ultrasonic frequency, beam area, and the acoustic power output of the transducer (8).

Ultrasonic waves act on tissues thermally and mechanically. Mechanical effects on tissues via ultrasound include acoustic torque, acoustic streaming, radiation force, stable cavitation, and unstable cavitation (11). Any object situated within an acoustic field will be subject to acoustic pressure and acoustic force. Acoustic pressure in a standing wave field is inversely proportional to velocity. Acoustic torque results from variations in the acoustic field, which can be described as a time-independent twisting action. Acoustic torque causes a rotational movement of cells and intracellular organelles in the medium. Acoustic streaming describes the movement of fluid in an ultrasonic field. The compression phase of an ultrasonic wave deforms tissue molecules. Radiation force affects gas bubbles that are in the tissue fluids. Negative pressure induces the bubbles originally dissolved in the medium to fall out of solution. With positive and negative pressure wave fluctuations, these bubbles expand and contract without reaching critical size (stable cavitation). Unstable cavitation occurs when bubbles collapse violently under pressure after growing to critical size due to excessive energy accumulation. This implosion produces large, brief local pressure and temperature release, as well as causing the release of free radicals. Organs that are air-filled, (e.g., the lungs or intestines), are subject to greater chance of unstable cavitation. Unstable cavitation is somewhat random, and as such it may lead to uncontrollable tissue destruction (8,11). Bubble growth can be limited by low intensity, high frequency, and pulsed ultrasound. Higher frequency means shorter cycle duration so time for bubble growth is regulated. Pulsed ultrasound restricts the number of successive growth cycles and allows the bubble to regain its initial size during the off period. The mechanical index estimates the possibility of cavitation occurrence. The mechanical index is directly proportional to peak rarefaction pressure, and inversely proportional to the square root of the ultrasound frequency (8).

MEDICAL APPLICATIONS OF CONVENTIONAL HYPERTHERMIA

Ultrasound as a heating modality has been used in several different medical fields. It is used in treating sprains, bursitis, joint inflammation, cardiac ablations, and in gynecology. However, the main area conventional hyperthermia is currently used is in oncology. The use of conventional hyperthermia as an oncologic treatment is supported by a plethora of studies that demonstrate that heat on cell lines and on animal tumor transplant models can result in tumor regression; however, it is rarely used alone because its efficacy is greatly potentiated in combination with radiation or chemotherapy. Conventional hyperthermia treatments elevate target tissue temperatures to 42–46 °C (12). Treatment times are usually between 30 and 60 min. Treatment applications are administered once or twice a week and are applied in conjunction with or not long after radiation. In all of the recent phase III trials, a sequential delivery scheme was used. This means that radiation and hyperthermia were administered separately, with radiation preceding hyperthermia treatments (13). Tumoricidal effects *in vivo* are achieved at temperatures between 40 and 44 °C (5). Large tumors often have an inadequate blood supply and resultantly, have difficulty meeting their requirements for oxygen and nutrients. This situation creates a hypoxic environment that is low in pH (2–3) (3,5,14). When tumor cells are heated to therapeutic temperatures, their cellular metabolic processes are accelerated, thereby further increasing the demands for oxygen and nutrients in an already depleted environment. Most tumor cells are unable to reproduce in this hostile environment, resulting in termination of tumor growth and shrinkage of the tumor (5,15). In temperatures > 40 °C, protein denaturation has been observed as the main mechanism of cellular death. Widespread protein denaturation results in structural changes in the cytoskeleton and the cell membrane, and in enzymes that are necessary for deoxyribonucleic acid (DNA) synthesis, cellular division, and cellular repair (5). Hyperthermic efficacy is a function of temperature and exposure time. To quantify, at temperatures $>$ 42.5– 43 °C, the exposure time can be halved with each 1 ° temperature increase to give an equivalent cell kill (5,16). Healthy tissues remain undamaged at temperatures of 44 °C for a 1 h duration (5,17). The exceptions are central nervous tissues, which suffer irreversible damage after being exposed to heat at temperatures ranging from 42 to 42.5 °C for >40–60 min (5,18). Peripheral nervous tissue that has been treated for $>$ 30 min at 44 °C or an equivalent dose results in temporary functional loss that is reversed in 4 weeks (5,19). Therefore, since a small difference in temperature produces a large difference in the amount of cells killed, it is important to be able to have good control on the site and duration of heat delivery to reduce the damage to surrounding healthy tissue.

RADIATION COUPLED WITH CONVENTIONAL HYPERTHERMIA

While hyperthermia independently has been found to have antitumor effects, its efficacy is greatly potentiated when coupled with radiation. Cells that are in a low pH hypoxic environments, those that are in the S or M phases of cell division, and those that are malnourished are relatively resistant to radiation (5,7). Hyperthermia increases radiation damage and prevents cellular repair of damaged DNA (5,16). Hyperthermia increases blood perfusion via

vasodilation which results in increased oxygenation, thus allowing increased radiosensitivity (5,7,16). Response rates with hyperthermia alone are ~15%, with radiotherapy ~35%, with combined radiotherapy, and hyperthermia ~70% (20). There have been many U.S. and European clinical trials that support substantial improvement in patients who have been treated with a combination of radiation and hyperthermia. Examples of some recent trials include randomized multiinstitutional phase III trials for treating melanoma (20,21), glioblastoma multiforme (20,22), chest wall recurrence of breast cancer (20,23), head and neck cancer (20,24,25), head and neck in superficial measurable tumors (20,26,27), in various recurrent persistent tumors (20,28), cervical cancer (29), uterine cancer (30) and in locally advanced pelvic tumors (20,31) (Table 4). Trial success rates were very dependent on the uniformity of temperature delivery. In the past, trials had often provided mediocre results because temperatures were ~1–1.5 °C too low and consequently not able achieve adequate tumoricidal levels (7). It is often difficult to uniformly heat larger tumor (3,7). When radiation and hyperthermia are used simultaneously excellent radiation delivery is achieved, often resulting in tumor regression; however, its delivery is equally as toxic to healthy cells that necessitate the need for a very precise delivery system.

CHEMOTHERAPY IN CONJUNCTION WITH CONVENTIONAL HYPERTHERMIA

Chemotherapeutic efficacy is enhanced by hyperthermia (5,20,34,35) (Table 5). As areas are heated, perfusion is increased, thus allowing an increase in drug concentrations in areas of the tumor that are poorly vascularized, increased intracellular drug uptake, and enhanced DNA damage. Drugs (e.g., mitomycin C, nitrosureas, cisplatin, doxorubicin, and mitoxantrone) are subject to less drug resistance when used with heat. The synergistic effect of chemotherapy and hyperthermia was demonstrated in virtually all cell lines treated at temperatures >40 °C for alkylating drugs, nitroureas, and platin analogs dependent on exposure time and temperature. Chemotherapeutic agents can be revved up 1.2–10 times with the addition of heat (5). *In vivo*, experiments showed improvement when doxorubicin and mitoxantrone were combined with hyperthermia. However, antimetabolites vinblastine, vincristine, and etoposide did not show improvement with the addition of hyperthermia. In animal studies, increased toxicities were seen in skin (cyclophosphamide, bleomycin), heart (doxorubicin), kidney (cisplatin, with a core temperature >41 °C), urinary tract (carmustine, with core temperatures >41 °C), and bone marrow (alkylating agents and nitroureas) (5,34). Lethal toxicity was enhanced when systemic hyperthermia was applied in combination with cyclophosphamide, methyl-CCNU, and carmustine (5). The success of hyperthermia and chemotherapy combinations depends on the temperature increase in the organs for which the drug is used and its subsequent toxicity, all of which are can be influenced by the accuracy of the heating device and the operator.

MODES OF CONVENTIONAL HYPERTHERMIA APPLICATION

Externally Applied Techniques

In the past, single planar transducers were used to apply local heat. A disk shaped piezoelectric transducer (range from 0.3–6.0 MHz in frequency and up to 16 cm in diameter) is mounted above a chamber of cooled degassed water. This device has a coupling chamber which allows water to circulate (3) (Fig. 6). It is coupled to the body via a plastic or latex membrane. Unfortunately, these types of devices are unable to achieve homogenous therapeutic thermal levels. The reason is that this system uses an unfocused energy source. When an unfocused energy source is applied to the skin, the intensity and the temperature will be the highest at the contact point and will subsequently lose intensity as it travels deeper into the tissue. However, by cooling the skin, the "hot spot" is shifted to the subcutaneous fatty tissue that is poorly vascularized. Fat is an insulator and as a result much energy is conserved rather than lost. Furthermore, cooling the skin will produce vasoconstriction which conserves even more heat and facilitates the heating of deeper tissues (3) (Figs. 7, 8a and b). However, even with this strategy adequate temperatures could not be reached. The reason for this is that large tumors often consist of three zones, a central necrotic core, an intermediate zone that is normally perfused, and a marginal zone that has a greater number of vessels due to proliferation induced angiogenesis. Due to the abundance of vasculature on the marginal surface, much heat is dissipated to the surrounding tissue. The relatively avascular center will heat to a higher temperature than the marginal or intermediate zone because there is little dissipation of heat, creating hot spots (Fig. 9) (7,54). Thus, it is not possible to therapeutically heat a tumor with a single planar source. This theory is substantiated by a significant number of clinical trials (7,55–61). Most trials reported that patients had some difficulty with dose-limiting pain, extensive central tumor necrosis, blistering, and ulceration (7). Conversely, a focused ultrasound source localizes energy within the tumor volume while sparing the surrounding tissue (Fig. 10). The use of a focused beam allows for homogenous heating and higher intensity which allows the generation of greater temperatures within the target area. Attenuation and beam divergence cause rapid deterioration of intensity beyond the focal zone (3) (Fig. 11 a and b). Focused ultrasound sources overcome some of the limitations of planar heating. Focusing allows for controlling the amount of heat that is delivered to the poorly perfused areas thus limiting hot spots and some of the side effects. For a heating system to be successful clinically on a large scale, it must account for geometric and dimensional variations of target tissue, possess the ability to heat the sites that need it, and avoid side effects and complications as much as possible (3).

Technological advances in hyperthermia devices have paved the way for better therapeutic options. The use of mosaics or separately controlled transducers allowed better spatial and temperature control to target bulky irregularly shaped tumors. The multielement planar

Table 4. Hyperthermia and Radiation Clinical Trials

Reference/ name of trial	Tumor Entity (stage)	Type of Trial	No. of Patients	Type of Hyperthermia	Results of Control Arm (RT only)[a]	Results of Hyperthermia Arm (RT+HT)[a]	Significance of Results ($p<0.05$)
(26,32) RTOG	Head and neck (superficial measurable tumor)	Prospective randomized multicenter	106	Superficial (915 MHz microwave)	34% CR	34% CR	–
(25)	Head and neck untreated locoregional tumor	Prospective randomized	65	Superficial (27-12 MHz microwave)	32% DR	55% CR	+
					19% DFS at 1.5 years	33% DFS at 1.5 years	+
(24,33)	Head and neck (N3locoregional tumor)	Prospective randomized	41	Superficial (280 MHz microwave)	41% CR	83% CR	+
					24% LRFS	68% LRFS	+
					0% OS at 5 years	53% OS at 5 years	+
(21) ESHO-3	Melanoma (skin metastases or recurrent skin lesions)	Prospective randomized Multicenter	70	Superficial (various techniques)	35% CR	62% CR	+
					28% LRFS at 5 years	46% LRFS at 5 years	+
(23) MRC/ ESHO-5	Breast cancer (local recurrences or inoperable primary lesions)	Randomized multicenter	306	Superficial (various techniques)	41% CR	59% CR	+
					ca. 30% LRFS	ca. 50% LRFS	+
					ca. 40% AS at 2 years	ca. 40% AS at 2 years	–
(31)	Rectal cancer	Prospective randomized multicenter	143	Deep regional HT (various techniques)	15% CR	21% CR	–
					22% OS at 3 years	13% OS at 3 years	–
	Bladder cancer		101		51% CR	73% CR	+
					22% OS at 3 years	28% OS at 3 years	–
	Cervical cancer		114		57% CR	83% CR	+
					27% OS at 3 years	51% OS at 3 years	+
(28)	Various (recurrent or progressive lesions)	Prospective randomized multicenter	174	Interstitial HT (300-2450 MHz microwave or RF)	54% CR	57% CR	–
					34% OS at 2 years	35% OS at 2 years	
(25)	Gioblastoma (postoperative)	Prospective randomized	79	Interstitial HT	15% OS at 2 years	31% OS at 2 years	+
(29)	Stage IIIB uterine cervix	Prospective randomized	40	Deep regional HT	50% CR	80% CR	+
					45% CR at 3 years	79.7% CR at 3 years	
(27)	Superficial tumors	Prospective randomized	122	EM	42.3% CR	66.1% CR	+
(30)	Uterine cervical	Prospective randomized multicenter	110	RF	68.5% CR	73.2% CR	+

[a]AS = actuarial survival; CR = complete remission; DFS = disease free survival; HT = hyperthermia; LRFS = local relapse free survival; OS = overall survival; RF = radio frequency electric currents; RT = radiotherapy. (Published with permission from Ref 20).

ultrasound applicators met these demands and are capable of treating tumors at depths up to 8 cm. The multisector applicator allows for heating to the edge of the aperture and the acoustic beams are nondiverging in the near field, thus allowing large tumor heating with lateral measurements of 15 × 15 cm. Each of these 16 sectors can be varied from 0 to 100% power to uniformly heat across the tumor. If an area of the tumor is too difficult to treat, more energy

Table 5. Hyperthermia and Chemotherapy Clinical Trials

Reference	Tumor Entity	Type of Trial	No. of Patients	Type of Hyperthermia[a]	Type of Chemotherapy[a]	Results[a]
(36)	Oesophagus cancer (preoperative)	Phase II	32	localHT/ Endoluminal MW	CDDP + Bleo + Cyc	8 CR/13 PR (65% RR)
(37)	Oesophagus cancer (preoperative)	Phase III	20	localHT/ Endoradiotherm	CDDP + Bleo	1 CR/5 PR/4 MR (50% RR); FHR (41.2%)
			20	Control	CDDP + Bleo	0CR/5 PR/0 MR (25% RR); FHR (18.8%)
(38)	Stomach cancer	Phase II	33	RHT/thermotron	Mitomycin + 5FU	3 CR + 10 PR (39% RR)
	pancreatic cancer		22	8 MHz	Mitomycin + 5FU	3 CR + 5 PR (36% RR)
(39)	Pancreatic cancer	Phase II	77	RHT 13.5 MHz	Mitomycin + 5FU +/ - immunostimulation	27..3% survival at 1 year
(40,41)	Sarcomas (pretreated with chemotherapy)	Phase II (RHT 86)	38	RHT/BSD 1000 60-110 MHz	VP16 + IFO	6 pCR + 4PR + 4FHR (37% RR)
		Follow-up	65		VP16 + IFO	9pCR + 4PR + 8FHR (32% RR)
(42,43)	High risk soft tissue sarcomas	Phase II (RHT 91)	59	RHT/BSD 2000 80-110 MHz	VP16 + IFO + ADR	ICR/6pCR + 8PR + 13 MR (47%)
(44)	High risk soft tissue sarcoma	Phase III (EORTC 62961)	112	RHT/BSD 2000 80-110 MHz (randomized)	VP16 + IFO + ADR	OS: 46% at 5 years (08/00)
(45)	Soft tissue sarcoma	Phase II	55	ILP with HT	TNF + IFN + L-PAM	10CR/35PR (82% RR)
(46)	Sarcoma/ teratomas (metastatic)	Phase I/II	19	WBH	IFO + CBDCA	6PR (32% RR)
(47)	Sarcoma (metastatic)	Phase II	12	WBH	IFO + CBDCA + VP16	7PR (58% RR)
(48)	Refractory cancers (advanced or metastatic)	Phase I	16	WBH (Aquatherm)	L-PAM (dose-escalation)	ICR/2PR (19% PR)
(49)	Pediatric sarcomas	Phase II	34	RHT/BSD 2000 80-110 MHz	V16 + IFO + CBDCA	12 NED ('best response')/ 7 CR Duration: 7-64 months
(50)	Pediatric nontesticular germ cell tumours	Phase II	10	RHT/BSD 2000 80-110 MHz	CDDP + VP16 + IFO (=PEI)	5CR + 2PR (70% RR) Six patients alive without evidence of tumour (10-33 months)
(51)	Cervical cancer (recurrences)	Phase II	23	RHT/array-system 70 MHz	CDDP (weekly)	2pCR/ICR + 9PR (52% RR)
(52)	Rectal cancer (Dukes C preoperative)	Phase II	27	Intraoperative IHP	Mitomycin C	3 LR
			35	Control	Mitomycin C	13LR
(53)	Metastatic Sarcoma	Phase II	108	whole body Hyperthermia	IFO/CBDCA/VP16	68% success at 1 year

[a]P = intraoperitoneal hyperthermic perfusion; WBH = whole body hyperthermia; 5FU = 5-flurouracil; VP16 = etoposide; IFO = ifosfamide; ADR = Adriamycin = Doxorubin; CDDP = Cisplatin; CBDCA = Carboplatin; Bleo = Bleomycin; L-PAM = Melphan; TNF = tumor necrosis factor alpha; IFN = interferon gamma; p = pathohistological; RR = response rate; CR = complete remission; PR = partial remission; MR = minor response; FHR = favorable histological response >75%; LR = local recurrence; NED = no evidence of disease.(Published with permission from Ref. 20).

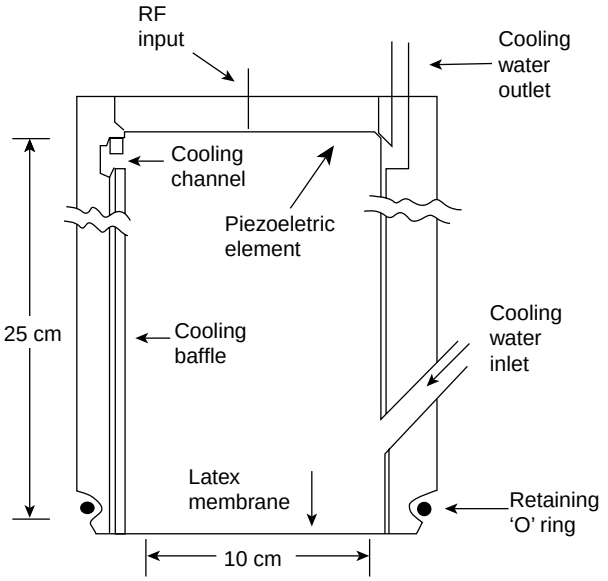


Figure 6. A cross-sectional diagram of a single planar element hyperthermia applicator. The chamber between the latex membrane and the piezoelectric element contains cooled degassed water. During local hyperthermia treatment an ultrasonic conducing gel is applied to the target site that is then coupled to the latex membrane. (Published with permission from Ref. 3).

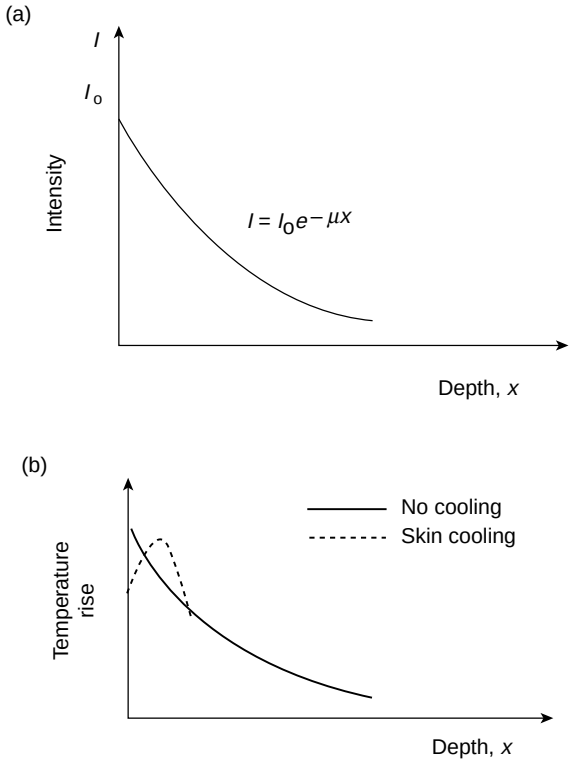


Figure 8. (a) Ultrasound intensity is greatest at the surface. Intensity will deteriorate exponentially due to attenuation as the depth from the surface increases. Published with the permission of (3). (b) Since temperature and intensity are directly proportional, temperature will decrease exponentially as depth increases. Cooling the skin will cause the "hot spot" to shift to the poorly perfused fatty tissue. (Published with permission from Ref. 3).

can be directed to just that target segment. The temperatures can be adjusted in relation to variations in temperature distribution due to blood flow, variations in target tissue morphology, and based on the patient's comfort level. These devices have the ability to contour the energy field to match the tumor outline. These systems generally have two frequencies: 1 MHz (used to heat 3–6 cm) and 3.4 MHz (used for more superficial 2–3 cm) (7,62). Examples of heating temperatures for different ultrasound hyperthermia devices are shown (Table 6). These planar array systems have been adapted to allow for thermoradiation in conjunction with an external beam radiation (7). An extended bolus configuration with an internal reflecting system was created to direct the ultrasound energy into desired tissue. This configuration allows the ultrasound transducer to be outside the radiation beam thus preventing potential interference of the two (7,70).

Another approach to achieving greater spatial control is to use a larger variety of small transducers in a nonplanar geometric configuration. This approach has been used in treating intact breast with cancer (7,71). The patient lies prone while the breast is immersed within the water filled cylindrical applicator (Fig. 12). The cylindrical applicator is composed of eight rings (each ring is 25 cm in diameter by 1.6 cm in height), with up to 48 transducers (1.5 $\times$1.5 cm plane transducers), which are interspersed around the

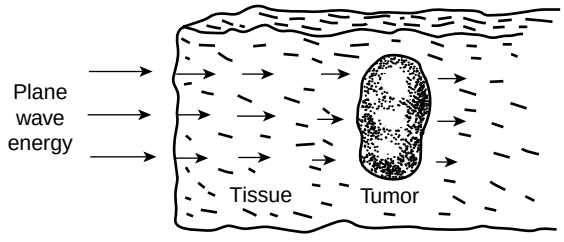


Figure 7. The pattern of ultrasound delivery via plane wave radiation targeting a tumor that is located deep within the tissue. (Published with permission from Ref. 3).

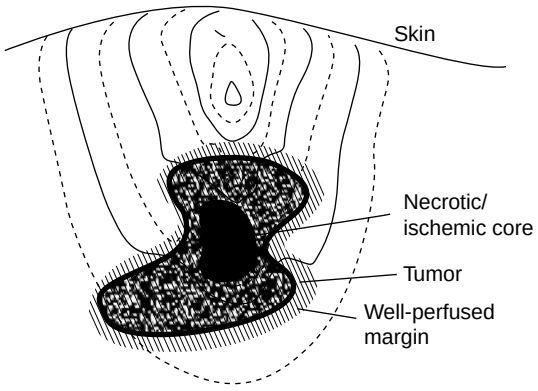


Figure 9. Temperature distribution in a subcutaneous tumor by plane wave transducer. The temperature at the necrotic zone is higher than in the surrounding tissues. (Published with permission from Ref. 3).

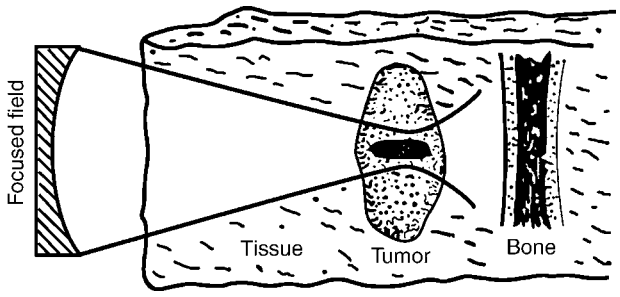


Figure 10. Pattern of radiation with a focused ultrasound beam. The contoured beam localizes the tumor within the focal area while sparing the surrounding tissue (Published with permission from Ref. 3).

ring. The frequency ranges from 1.8 to 2.3 and 4.3–4.8 MHz. The driving frequency and the power can be individually selected within each ring, which allows for better spatial and temperature control. This technique has not yet reached widespread clinical use (7).

The Scanning Ultrasound Reflector Linear Array System (SURLAS), which may soon be implemented in clinical practice allows for 3D power distribution while applying simultaneous external hyperthermia in conjunction with radiation to superficial areas (7,13,72–77). (Fig. 13). The SURLAS applicator consists of two parallel opposed ultrasound linear arrays that aim their sound waves to a V-shaped ultrasound reflector that further organizes and spreads the energy over the scanned target site (7,13). The two arrays operate at different frequencies (1.9 and 4.9). This allows for control of penetration depth through the exploitation of intensity modulation of the two beams (13). The applicator housing this transducer and the temperature regulated water bolus are placed on the patient. This system allows both the radiation and the ultrasonic waves to enter the patient's body concurrently. During the scanning interval, power levels and frequencies in each transducer can be individually regulated, thus allowing for good control over depth penetration and lateral heating (7). This system can treat superficial tumors that are 15×15 cm in area and with distal margins up to 3 cm deep (13). However, scan times must be limited to <20 s to avoid transient temperature variations >1 °C (7,73).

Large superficial tumors ranging from 3 to 4 cm deep 20×20 cm in surface area have been successfully treated with mechanically scanned planar transducers with 2D

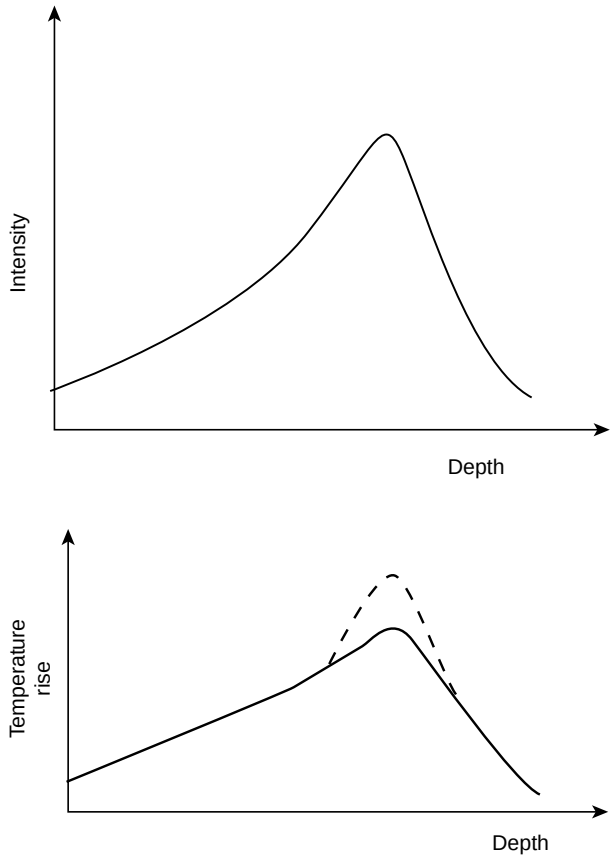


Figure 11. (a) With the use of a focused field, higher intensities can be achieved in target tissue at a greater depth. (Published with permission from Ref. (3)). (b) Since temperature and intensity are directly proportional greater temperatures can also be attained in target tissue at greater depths. (Published with permission from Ref. 3).

motion (7,63) (Fig. 14). This approach can be used in treating tumors in the chest region, which often have a heterogenous thickness and are situated close to bone. Once an ultrasound is launched into tissue, it cannot leave the body; consequently, it will just "bounce" around until it is completely absorbed. If the ultrasound is absorbed by bone or nerves, neuropathies and bone necrosis can occur. Mechanically scanned planar transducer frequencies can range from 1 to 6 MHz. Accurate spatial control has been achieved by controlling the operating frequency and

Table 6. Examples of Clinical Temperature and Response Rates of Certain Hyperthermia Systems[a]

Device	Reference	Number of Patients	Maximum Temperature, °C	Minimum Temperature, °C	Average Temperature, °C	Complete Response Rate, %	Partial Response Rate, %
Scanned ultrasound	(63)	5	45.9	41.1			
	(64)	149				34	36
	(65)	72	44.4	40.0		22	40
	(66)	17	43.1	39.9		24	70
	(67)	15	44	40.4	42.3		
Multielement ultrasound	(68)	147	42.7	38.5	40.4		
Transrectal ultrasound	(69)	14	43.2	40.5	42.2		

[a]Published with permission from Ref. 7.

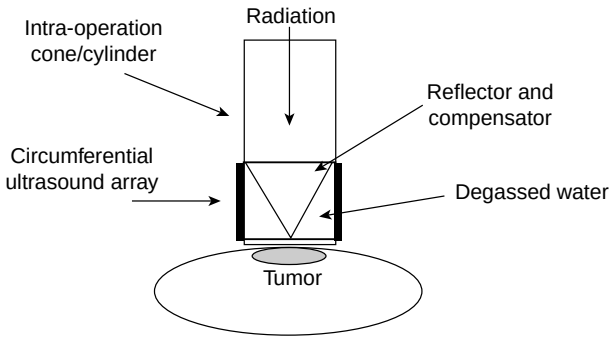


Figure 12. A schematic diagram of an intraoperative multielement transducer system with a circumferential transducer array and reflector configuration. (Published with permission from Ref. 7).

applied power levels, as a function of location, to account for variations of tumor thickness. Separate transducers, which are driven at different frequencies or by time multiplexing the driving frequency of a given transducer between its fundamental and odd harmonic frequencies, are able to create a situation that allows control over penetration depth (7). The penetration depth, as well as the temperature distribution resulting as a function of depth, can be controlled online during the scanning by regulating the frequency amplitude. In the clinical setting, all these biophysical properties must be coupled with the patient's ability to tolerate the treatment to create a functional algorithm (7,63).

Scanned focus ultrasound systems (SFUs) provide the most flexibility for clinical applications (7,64–67,78,79). These systems provide the greatest possibility of overcoming the challenges of tissue heating. The SFUs systems generally use four to six 1 MHz spherically focused transducers each overlapped so that a common focal zone of 3 mm o.d. to treat deep tissue. This focal zone is mechanically scanned in circular or octagonal patterns within the tumor at rates of 20–100 mm·s^{-1}. In order to guarantee that there is a consistency in temperature, scan cycles must be shorter than 10 s. During scanned focused ultrasound hyperthermia treatments, temperature distributions can be controlled by utilizing the measured temperatures to vary the power output as a function of the location. The resolution is determined by a variety of thermometry

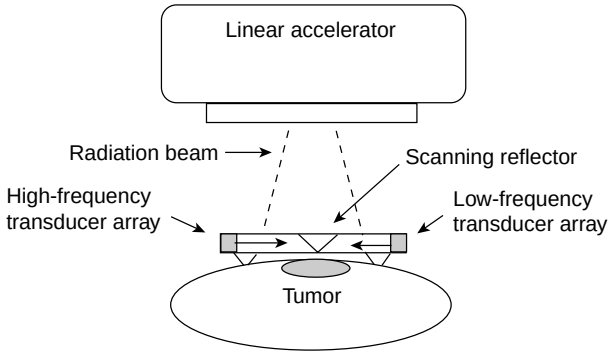


Figure 13. A schematic diagram of a multielement low profile scanning reflector system. (Published with permission from Ref. 7).

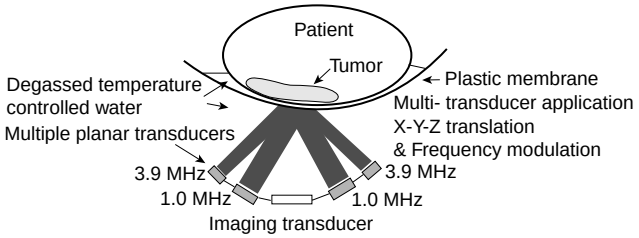


Figure 14. A schematic diagram of mechanically scanned ultrasound system for superficial hyperthermia. (Published with permission from Ref. 7).

points, scanning, and computer control speed (7). The regulation of temperature can be controlled by the clinician or the computer (7,80).

External applicator systems for hyperthermia have now been developed that use electrically phased focused transducer arrays. The advantages of using an electrically phased focused transducer array is that it allows for better synthesis of complex beam patterns and the ability to electronically focus and steer. The 3D complex beam-forming techniques result in higher scanning speeds, smaller applicators, and better reliability due to more static parts (7). Examples of electrically phased focused transducer arrays include concentric ring arrays (7,81), sector-vortex phased arrays (7,82), spherical and cylindrical arrays (7,83,84), and tapered phased arrays (7,85).

Intracavitary Techniques

Conventional ultrasonic hyperthermia can be used for intracavitary applications. This modality can be used to treat tumors that are situated deep within the body or with those that situated close to a body cavity. Clinically, prostrate cancer and benign prostate hyperplasia are the best suited for this treatment (7). The transrectal applicator consists of one-half cylindrical transducer segments 10–20 mm o.d. $\times$ 10 mm long. It is sectored for better angular control with frequency range of 1.0–1.6 MHz. The transducers are housed in a plastic head; also, a temperature regulated degassed water within an extendable bolus is attached (7,86–88) (Fig. 15). The heating energy is emitted radially from the length of each transducer segment, and the power is applied along the length of the applicator. This technique is able to heat tissues that are 3–4 cm deep from the cavity wall. The temperature controlled water bolus maintains a safe temperature for

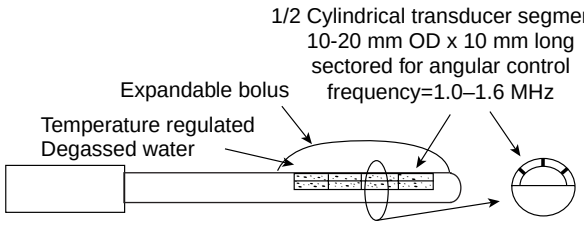


Figure 15. A nonfocused multielement applicator with longitudinal and angular power deposition abilities. This device is used in the treatment of the prostate cancer or BPH. (Published with permission from Ref. 7).

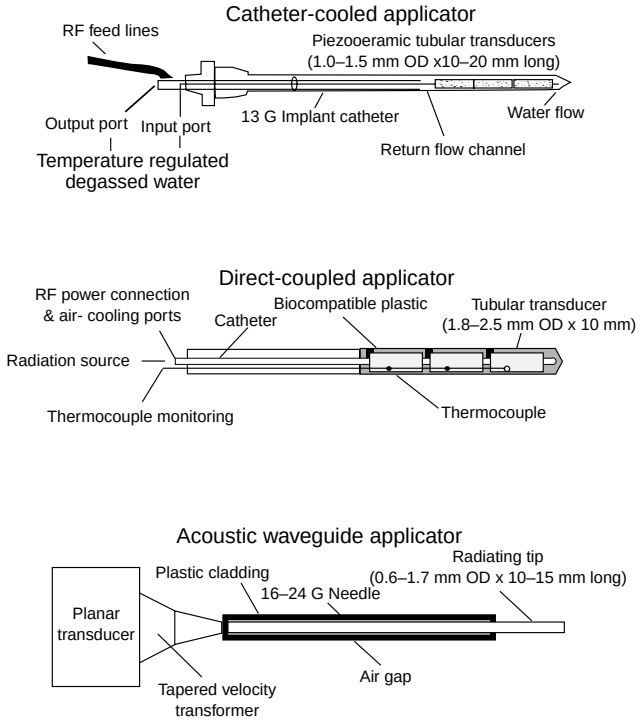


Figure 16. Interstitial hyperthermia catheters. (Published with permission from Ref. 7).

the rectal mucosa. Improved versions of this applicator have added four sectors on each tubular section for 16 channels total. These devices are being fabricated to be compatible with MRI guide protocols (7,89).

Interstitial Techniques

Interstitial techniques are used for treating deep-seated and/or large tumors that are not amenable for surgical resection. Heating sources are implanted into the tumor, thus focusing the energy directly to the site. Interstitial radiation is a standard practice in the treatment of tumors, therefore incorporating adjuvant heat is a logical progression to maximizing treatment. There are three basic designs of interstitial ultrasonic applicators: catheter cooled and direct coupled that consists of tubular piezoceramic transducers, and acoustic waveguide antennas (7) (Figs.16a–c).

Multielement ultrasound applicators with catheter cooling use piezoceramic tubular transducers (1.0–1.5 mm o.d. $\times$ 10–20 mm long, with frequency ranging from 7 to 10 MHz) have circulating coolant channels incorporated within the support structures to allow the applicator to be sealed in place within closed end implant catheters (13–14 gauge) (7) (Fig. 16a). These catheters are able to improve control of radial penetration of heat. In addition, it has the ability to control longitudinal power deposition along the length of the applicator (7,90–94). The power to each tubular transducer can be adjusted to control tissue temperature along the length of the catheter. The length and the number of transducers can be selected depending on the desired temperature and longitudinal resolution. This feature is very valuable in that it allows adjustability to tumor geometry variations, blood perfusion variations, and the variation within the tumor tissue. Another advantage of this device is that, unlike microwaves and RF hyperthermia, the power deposition pattern is not limited by the length of insertion or whether other catheters are within the implant. These catheters are more challenging than others for the operator to use skillfully because it is complicated to control both the electronics and the water cooling. Also, small transducers are less reliable. However, it is this complexity that allows for great plasticity in therapeutic temperature distributions (7).

Direct coupled applicators are used to deliver thermobrachy therapy via remote after- loading radiation sources (Fig. 16b). Larger applicator size limits these catheters to few clinical treatments. The implant catheter consists of the transducer and an acoustically compatible housing, which is biologically and electrically insulated. The implant catheter usually ranges from 2.2 to 2.5 mm in diameter. The inner lumen is formed from a catheter that is compatible with standard brachytherapy and commercial after loaders. The transducers have sensors that are able to monitor tissue temperature. In order to conserve size, a water cooling mechanism was not included as part of the catheter. This device is less efficient because transducer self-heating increases the wall temperature and thus reduces radial heating. Therefore, the thermal penetration is sensitive to acoustic frequency (7,95,96). Some studies have shown that integrating an air cooling system to this catheter will allow for better heating penetration (7,95).

The acoustic wave-guide antenna has a minimally invasive 16–24 gauge stainless steel needle that is coupled by a conical tapered velocity transformer to a piezoceramic disk transducer (1.3 cm o.d. operating at 1 MHz) (7,99) (Fig. 16c). The length of the radiating tip can be changed by adjusting the length of the plastic sleeve by 1–1.5 cm. The needle diameter size minutely fluctuates due to Raleigh surface waves propagating from the wave-guide generating flexural vibrations of the needle portion. Acoustic patterns that have been measured demonstrate peaks and nodes in adjacent tissue along the radiating aperture. The temperature of the tissue that is radiated matches the temperature of the radiating antennae. The disadvantages of this system are that the power output is potentially limited for larger or more perfused tumors, and it is difficult to control the longitudinal power deposition (7).

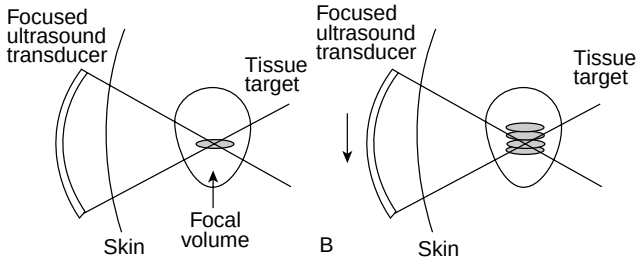


Figure 17. Schematic of HIFU. (a) Illustrates a formation of a single lesion. (b) Illustrates a confluent array of lesions required for a therapeutic effect. (Published with permission from Ref. 98).

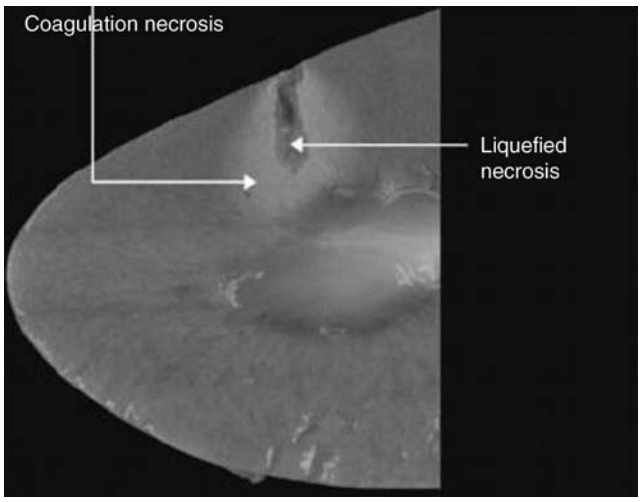


Figure 18. Image of coagulation and liquefied necrosis created with HIFU in an *ex vivo* porcine kidney. (Published with permission from Ref. 108).

A Brief History of HIFU

Using HIFU as an extracorporeal technique of creating coagulative necrosis was first conceptualized in 1942 by Drs. Lynn and Putnam (12,98,99) (Fig. 17a). In 1954, Dr. William Fry was the first to use HIFU to destroy central nervous tissue in the brains of cats and monkeys (12,98,100,101). Later, Frank Fry treated patients with Parkinson's disease and neuromata (12,98,102). Throughout the 1950s and 1960s, HIFU research continued, although it was often plagued with limited success due to lack of technology (103–106). In 1956, Dr. Burov suggested that HIFU can be implemented in the treatment of cancer (12,98,107). Since then, the popularity of HIFU has gradually increased with the advent of better devices and with the success of its use *in vitro* and *in vivo* experimental trials. In current literature, HIFU is categorized as a high temperature hyperthermia because higher temperatures than those used in conventional hyperthermia are required to achieve therapeutic goals.

BASIC PRINCIPLES OF HIFU

The concept of HIFU is similar to that of using a magnifying glass to focus the sun's beams to set fire to some dry leaves. Only the leaves that are in focus will be set on fire, the surrounding ones will be spared (12,98). Likewise, if an ultrasonic beam with sufficient energy is tightly focused, it can be used to elevate temperatures within a target tissue resulting in cell death and coagulative necrosis while sparing the skin and surrounding tissues (98,108) (Fig. 18). Histologically, there is a sharp demarcation between the necrotic tissue that was radiated with HIFU and the healthy surrounding tissue. In the liver, 2 h after exposure, the cells look normal, however, approximately a 10 cell wide rim of glycogen poor cells can be found. After 48 h, the entire area that was radiated will be dead (109).

During HIFU procedures, tissue temperature $>56\ ^{\circ}C$ are used because at that temperature irreversible cell death through coagulative necrosis occurs. The main mechanism used is coagulative necrosis via thermal adsorption (110). The other mechanism is cavitation induced damage that is caused by both thermal and mechanical properties of the ultrasound wave (110,111). However, recent studies have been investigating the use of cavitation to enhance the level of ablation and to reduce exposure times. It has been proposed that a focused ultrasound protocol that induces gas bubbles at the focus will enhance the ultrasound absorption and ultimately create larger lesions (110,112). Individual HIFU exposure times can be as little as 1–3 s, while larger volumes may require up to 30–60 s. Individual lesions can be linearly complied to create a clinically relevant lesion (Fig. 17 b). Since individual exposure time is quick, issues (e.g., the cooling effects of blood perfusion) can be considered negligible (7,98,113,114). Therefore, energy transfer and temperature elevation in tissue is considered proportional to acoustic field energy (100). The lesions are cigar-shaped or ellipsoid with the long axis parallel to the ultrasonic beam (12,98). In order to ablate tissue transducer frequency must be between 0.5 and 10 MHz. The higher the frequency, the narrower and shallower the lesion will be. The wavelength ranges from 3 to 0.25 mm. The size of the focal point is determined by the wavelength. Thus, the transverse diameter of the focus is limited to one wavelength and the axial diameter is eight times that wavelength. As a result of this, all generators create a focal size that is 10×1 mm. The shape of the lesion is determined by the acoustic properties of the tissue, ultrasound intensity in conjunction with exposure time, and transducer geometry (12). Lesion size is determined by power density at the focus, pulse duration, and the number of pulses. In order to create a well-demarcated lesion the intensity must be $>100\ W{\cdot}cm^{-2}$, thus being able to reach temperatures that are $>65\ ^{\circ}C$ in <5 s (11). Focal peak intensities generally range between 300 and 2000 $W{\cdot}cm^{-2}$ (7). The ultrasonic waves used in HIFU are generated by piezoelectric elements. In order to achieve high intensity focus ultrasound that is able to ablate tissues three techniques have been found to focus the ultrasound beam: (*1*). spherical arrangement of piezoelements (Fig. 19), (*2*) combination of a plane transducer with an acoustic lens (Fig. 20), (3). cylindrical piezoelements together with a parabolic reflector (11) (Fig. 21).

CURRENT EXTRACORPOREAL DEVICES, INTRACAVITARY DEVICES, AND IMAGING

While there are many devices that are used in experimental trials, few of those are currently used in widespread clinical practice. The two main categories of HIFU devices are extracorporeal and transrectal. Extracorporeal devices have been implemented in experimental trials in many medical fields. Extracorporeal devices use larger transducers, lower frequencies, and longer focal lengths than intracavitary devices (97).

An important factor in clinical application of these devices is the ability to monitor treatment accurately. In current practice, this is accomplished either by using real-time ultrasound (116–118) or MRI (119–122). When

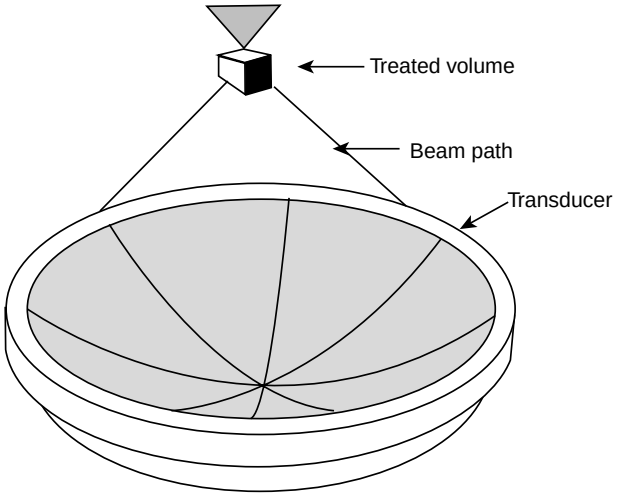


Figure 19. A single spherically curved focused transducer. (Published with permission from Ref. 110).

MR is used to guide HIFU treatments, sublesioning ultrasound exposures are used to identify the target region, local rise in temperatures are used to confirm the position of the ultrasound focus and then higher intensity therapeutic exposures are used for treatment. Currently, several groups are using ultrasound surgery systems that utilize MRI to map temperature elevations online during HIFU procedures (110,120–122). This technique has been used to treat breast tumors and uterine fibroids, and these treatments are in the process of being used clinically in several countries (110,123–125). The MR can effectively use temperature data to determine the parameter of thermal tissue damage (110) and is limited in that it is costly, has lower spin resolution, and because of its technology for producing MR compatible ultrasound equipment required for HIFU is lagging.

When ultrasound is used as a guide, the diagnostic transducer is arranged confocally with the therapeutic transducer and their relationship is fixed. The position

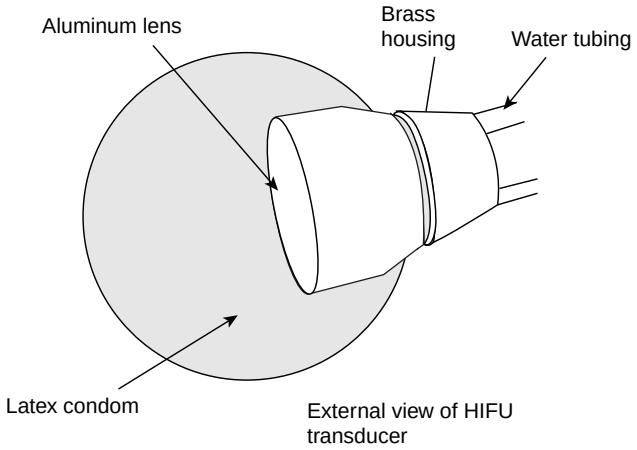


Figure 20. A plane transducer with an acoustic lens used for focusing the ultrasound waves. (Published with permission from Ref. 115).

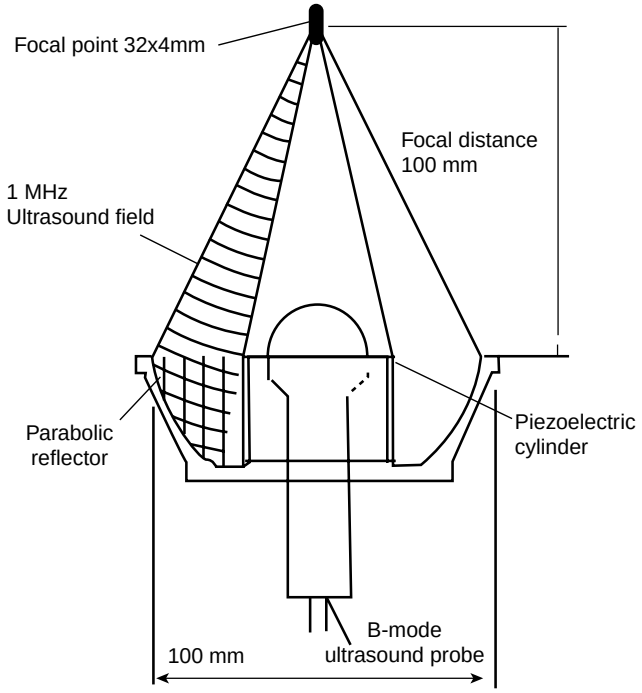


Figure 21. A cylindrical transducer with parabolic reflector. (Published with permission from Ref. 108).

of the therapeutic focus can be reliably identified on the diagnostic image. The extent of treatment can be monitored by recording post-treatment gray scale changes on the diagnostic image (98). Ultrasound as a guide is advantageous in that it is less expensive and is more readily accessible, it has faster treatment times, compact sized equipment, and provides a good correlation between observed ultrasound changes and the region of necrosis in the tissue. The disadvantage of using ultrasound as a guide is that image quality is not optimal (98,110,126). Furthermore, ultrasound waves are obstructed by bone and air-filled viscera.

MEDICAL APPLICATIONS OF HIFU

Liver Cancer

While hepatocellular carcinoma is frequently encountered in clinical practice, hepatic metastasis from other primary sources is much more common. Currently, the only definite treatment choice for hepatic metastases is surgery, however, 5 year survival rates are only 25–30%. Arterial embolization is another emerging technique. Therefore, the desire to find a noninvasive technique is preeminent (98). The Chongqing HAIFU device has been used for a couple of years in China to treat a variety of tumors, however, adequate data has not yet been collected (98,127,128) (Fig. 22). The JC-HIFU system (HAIFU Technology Company, Chongquing, PR China) uses an extracorporeal transducer that operates at 0.8–1.6 MHz, the aperture 12–15 cm, focal length 9–15 cm. It operates at Isp of 5–15 $kW \cdot cm^{-2}$. A diagnostic ultrasound probe (3.5 MHz) is aligned along the same axis as the therapeutic transducer. Both the treatment and diagnostic transducers are placed in a reservoir of degassed water in the center of the

Figure 22. A Hifu System that is used both clinically and experimentally in the treatment of liver metastates. (Published with permission from Ref. 128).

treatment table. The degassed water provides acoustic coupling between the patient and the transducer. Horizontal movement of the transducer is possible along three orthogonal axes of the bed because it is facilitated by the cylindrical gantry at one end of the table. All movement is controlled by the adjacent computer terminals (128). In a recent clinical trial carried out in Churchill Hospital in Oxford, England in conjunction with Chongqing University of Medical Sciences in Chongqing China, 11 patients with liver metastases were treated with the JC-HIFU device. While it is not possible to have a good statistical analysis with such a small subject pool, some general observations were made about the safety of this device. Of the 11 patients treated, 7 out of 11 patients complained of transient pain and 3 out of 11 complained of superficial burns. Out of the 7 patients that experienced pain, oral analgesia brought relief to 6. Burn sites were treated with ice-packs and aloe gel. Two of the three burn sites were only millimeters across. One of the burns was 2×3 cm and had healed by the 2 week follow-up period. It appears that from a safety standpoint the JC-HIFU is a feasible treatment option for hepatic metastases, however, larger trials will be needed to determine the true efficacy of the treatment (128).

Another study by Wu et al looked at 55 patients with hepatocellular carcinoma with cirrhosis. Tumor size ranged 4–14 cm in size with an average size of 8.14 cm. Patients were classified according to progression of disease: 15 patients had stage II, 16 had stage IIIA, and 24 had stage IIIC. All patients were treated with an extracorporeal HIFU device similar to the one previously mentioned for the treatment of liver metastases. The average number of treatment applications was 1.69. There were no serious side effects. Imaging following HIFU treatment evaluated for the absence of tumor vascular supply and shrinkage of treated lesions. Serum alpha-fetoprotein returned to normal in 34% of patients. At 6 months, 86.1% of the patients were still alive, at 12 months 61.5% of the patients were still alive, and at 18 months 35.3% of the patients were still alive. The survival rates were the highest in patients who were stage II. Therefore, this study demonstrated that HIFU is a safe option in the treatment of hepatocellular carcinoma (129).

Prostate Cancer

Prostate cancer is one of the common types of cancer in males, and it is frequently the cause of cancer-related death (130). Since physicians are able to detect prostate cancer early, there has been an increase in the number of patients needing treatment. Radical prostatectomy is the treatment of choice in patients who have organ-confined disease and a life expectancy of >10 years. Radical prostectectomy offers excellent results 5 and 10 years after the operation, although there is still risk of morbidity associated with the operation, thus precipitating the need for a noninvasive procedure. Currently, brachytherapy, cryosurgery, 3D conformal radiotherapy, and laparoscopic radical prostatectomy have been implemented with good results (130,131). However, if a cure is not achieved, these treatments cannot be repeated and there is high risk of morbidity associated with salvage radical prostatectomy, thus necessitating the need for another treatment option. In 1995, Madersbacher reported that they were able to destroy the entire tumor within the prostate (98,132). Early reports showed success rates of controlling local tumors at 50% at 8 months and then approaching 90% in later studies (98,133,134). In the later years, as clinicians gained more experience and as technology has improved, treatment of the entire gland was performed (98,135,136).

A recent report was published that looked at 5 year results with transrectal high intensity focused ultrasound in the treatment of localized prostate cancer. One hundred and forty six patients were treated with Ablatherm device (EDAP, Lyon, France). The tablespoon-shaped intracavitary applicator contains both a 7.5 MHz retractable ultrasound scanner for diagnosis and a piezoelectric therapeutic transducer that can be driven at frequencies of 2.25–3.0 MHz. The computer-controlled treatment head is able

to move three dimensionally. The applicator can be rotated 45 ° in the rectal ampulla. A cooling device that consists of a balloon containing degassed coupling fluid surrounds the treatment head. Energy can be released from the balloon rectal interface thereby maintaining rectal temperatures <15 °C. Out of 137 patients 6 reported symptomatic UTI, 2 reported chronic pelvic pain, 16 reported infravesicular obstruction, 8 reported grade1 stress incontinence, and 1 reported rectourethral fistula. The success rate of the Ablatherm system is between 56 and 73% (131).

Another study that was published by Uchida et al. performed 28 HIFU treatments on 20 patients to treat localized prostate carcinoma (T1b-2NOMO). A modified Sonoblate 200 HIFU device (Focus Surgery, Indianapolis Ind.) was used in this study. Sonoblate 200 uses a 4 MHz PZT transducer for both imaging and treatment. Each pulse delivery ablates a volume of $2 \times 2 \times 10$ mm^3 in a single beam with 2.5 and 4.5 cm focal length probes. Probes with focal lengths of 3.0, 3.5, 4.0 cm can be used in a split-beam conformation to create lesion sizes of $3 \times 3 \times 10$ mm^3. A cooling device maintains rectal temperatures at <22 °C. In this study, there was a 100% success rate. The UTI-like symptoms were common in the first 2 weeks post-HIFU, but were easily remedied with alpha-blockers and painkillers. One patient had a rectourethral fistula after a second HIFU treatment. Of 10 patients who were still able to attain tumescence prior to the procedure, 3 reported postoperative impotence. It is hypothesized that the reason the Sonoblate 200 is getting superior results to the Ablatherm system is that the treatable focal length is longer in the Sonoblate system. This allows the Sonoblate 200 to treat prostates <50 mL, whereas the Ablatherm can only treat prostates that are <30 mL. However, a controlled prospective study is needed to evaluate the potential reasons for this difference in efficacy (130).

Gynecology

The most common pelvic tumor in women of reproductive age is fibroids. The current surgical options available to manage fibroids are either hysterectomy or myomectomy. Hysterectomy is often not a viable option for women who wish to have children. Myomectomies often result in 50% tumor recurrence in 5 years. Hormone therapy results in temporary reduction in tumor size by 35–65% (115). Therefore, there is a need for a permanent, noninvasive technique to manage fibroids. A device was developed for treating uterine fibroids. The prototypic device aligns a commercial abdominal diagnostic ultrasound transducer with a therapeutic ultrasound intracavitary probe (Fig. 23). This device was constructed to accommodate the specific constraints of the female pelvic anatomy. The transducer contains a 3.5 MHz PZT-8 crystal, 25.4 mm in diameter bonded by an aluminum lens to focus the ultrasound beam. A water-filled latex condom is used for acoustic coupling of the transducer and also has the potential for transducer and tissue cooling. Ergonomics testing in humans demonstrated clear visualization of the HIFU transducer in relation to the uterus, thereby demonstrating a potential for HIFU to treat fibroids from the cervix to the fundus through the width of the uterus. However, this device needs to be tested in treating the uterus in large animal models prior to beginning human trials (115). Extracorporeal devices have been used in a small clinical phase I trials, but the results are still pending (98,137).

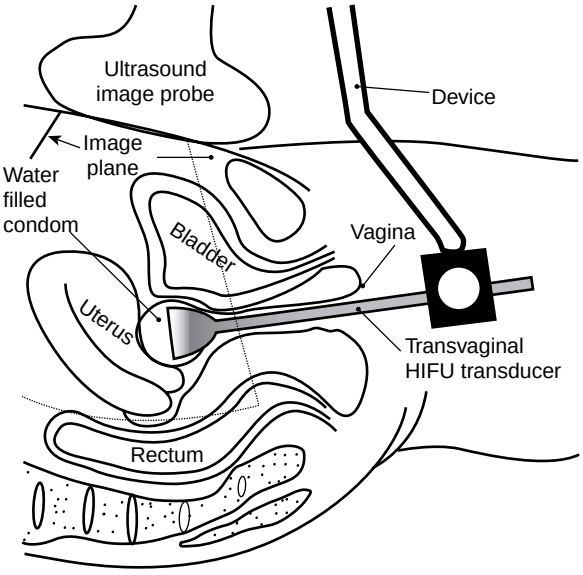


Figure 23. Conceptual diagram of the transvaginal device HIFU in use. The dotted line represents the image plane. Three factors determine what area can be treated by the HIFU transducer: focal length, the range of water stand-off, and the range of mobility once inside the vagina. (Published with permission from Ref. 115).

Breast Cancer

Every year >1 million new cases of breast cancer are diagnosed. Breast cancer is the most common malignancy in women (138,139). In the past, the only options available to women were radical and modified radical mastectomy that included dissection of the axillary lymph nodes. More recently, breast conservation surgery in conjunction with radiotherapy, chemotherapy, and hormone therapy has gained popularity in early stages of breast cancer. However, the change in approach toward a less radical surgery, while being better for a woman's body image, has not really increased long-term survival rates in breast cancer patients (138,140). Other options, such as cryoablation and laser frequency have been studied as minimally invasive approaches, however, these techniques are limited in that they require percutaneous access and that they are only able to treat small masses. In a recent study by Wu et al., (65) women with breast cancer (T1-2, N0-2, M0) verified with biopsy were studied. Patients were divided into a control group that had a modified radical mastectomy, and a group that had extracorporeal HIFU followed by a radical mastectomy. The HIFU system used is the same one described earlier in the treatment of liver malignancies, the JC-HIFU therapeutic system. The therapeutic U.S. beam was produced by a 12 cm in diameter PZT-4 ceramic transducer with a focal length of 90 mm that was driven at a frequency of 1.6 MHz. The ellipsoid focal region dimensions were 3.3 mm along the beam axis, and 1.1 mm

along the transverse axis. A real-time imaging U.S. device, the AU3 (Esaote, Genoa, Italy) was used at frequencies of 3.5–5.0 MHz. The diagnostic transducer is placed in the center of the therapeutic transducer. Real-time imaging can accomplish three separate tasks. It can locate the tumor that needs to be treated, it can guide the deposition of U.S. energy into the tumor, and it can provide real-time assessment of the coagulation necrosis during therapy. The results demonstrated that there were no severe side effects. Those that were reported included mild local pain, warmth, and a sensation of heaviness in the affected breast. However, only 4 of the 23 HIFU patients had significant pain to require a 3–5 day course of oral analgesics. Only one patient had a minor skin burn. Pathologic examination of the breast tissue revealed complete coagulative necrosis, and the tumor vasculature was damaged. The immunohistological staining revealed that no expression PCNA, MMP-9, and CD44v6 was found, indicating that the tumor cells had lost their ability to proliferate, invade, and metastasize. Therefore, this study demonstrated the safety and efficacy of HIFU in the treatment of breast cancer (138).

Neurology

Recently, there have been some published studies that propose using large array ultrasound transducers to overcome distortions caused by the skull (110). The goal has been to be able to create an array that can focus to destroy target tissue while preserving surrounding tissue. A 320 element array has been used to focus ultrasound through10 human skulls. This approach is completely noninvasive. This technique is modeled after a layered wave vector-frequency domain-model and uses a hemisphere-shaped transducer to propagate ultrasound through the skull using CT scans as a guide (110,141,142). The ability to focus energy has implications that are not limited to just tumor treatment. It has been shown that focused ultrasound can selectively and consistently open the blood brain barrier (BBB) (110).

Another neurological area that may benefit from HIFU is in the treatment of nerve spasticity and pain. Spasticity, which is signified by uncontrollable muscle contractions, is difficult to treat. In a recent study, HIFU was used to treat and suppress the sciatic nerve complex of rabbits *in vivo*. An image-guided HIFU device including a 3.2 MHz spherically curved therapeutic transducer and an ultrasound diagnostic device were used. A focal intensity of 1480–1850 W/CM2 was used to create a complete conduction block in the 22 nerve complexes. Treatment times averaged 36 s. Gross histological examination revealed blanched nerve complex with lesion dimensions of 2.8 cm^3. Further histological examination revealed the probable cause of nerve block as axonal demyelination and necrosis of Schwann cells. This study illustrates the potential that HIFU may have in the treatment of nerve spasticity (143).

Cardiovascular System

The role of ultrasound in cardiology has been instrumental to the increasing knowledge of the cardiovascular system. Diagnostic ultrasound technology has led to a greater understanding of the anatomy and physiology of the human heart and vascular systems. Over the years, in addition to diagnostic use, the role of ultrasound has been expanded to the therapeutic realm. Both conventional ultrasound and HIFU modalities have been used with varied success in many cardiovascular therapeutic applications. These applications range from harvesting the internal mammary artery for coronary artery bypass surgery to ablation of cardiac arrhythmias. An ultrasonically activated (vibrating up to 55,000 Hz) harmonic scalpel (Ethicon Endosurgery, Cincinnati, OH) produces low heat ($<100\ ^\circ C$) thereby effectively coagulating and dividing the tissue and has a wide range of applications in cardiothoracic surgery (144). By using a 1 MHz phased array transducer with an acoustic intensity of 1630 $W{\cdot}cm^{-2}$ or 22547 $W{\cdot}cm^{-2}$ one can successfully create precise defects ranging from 3 to 4 mm in diameter *ex vivo* in porcine valve leaflet, canine pericardium, human newborn atrial septum, and right atrial appendage (145). Cardiac arrhythmia is one area where significant work has been done using ultrasonic hyperthermia for therapeutic purposes. Strickberger et al. demonstrated an extracorporeal HIFU ablation of the atrioventricular junction of beating canine heart after thoracotomy (146). Their experimental system consisted of a polyvinyl membrane covering the heart and lungs. The thoracic cavity was filled with degassed water serving as a coupling medium. A 7.0 MHz diagnostic 2D ultrasound (Diasonics VST Master Series, Diasonics/Vingmed Ultrasound Inc) attached to a spherically focused single piezoelectric element therapeutic ultrasound transducer (1.4 MHz frequency; 1.1×8.3 mm focal length and 63.5 cm focal zone) with the maximum intensity of 2.8 $kW{\cdot}cm^{-2}$ was applied during the diastole for 30 s to achieve complete AV nodal junctional block (Fig. 24a–c). Experience with HIFU application is very preliminary and has not been tried for AV nodal ablation in humans yet.

Ultrasound had also been used clinically in the treatment of atrial fibrillation (AF), which is the most common arrhythmia affecting 0.5–2.5% of the population globally. Over the last decade, ablation procedures by isolating the pulmonary veins and eliminating electrical triggers from the atria has become a popular and effective mode of therapy for AF. Traditionally, RF energy has been used as an energy source for ablation. Radio frequency catheter ablation of AF requires good tissue contact, multiple lesions, significant experience and manual skills with long procedure time. Complications related to RF application in AF ablation include pulmonary vein stenosis, atrioesophageal fistula, left atrial rupture due catheter perforation or inappropriate amount of power. The limitations of the existing RF technology could be overcome with the use of HIFU balloon systems (147).

Our group performed the initial work on pulmonary vein isolation in humans using a through-the-balloon circumferential ultrasound (conventional-unfocused) ablation system for treatment of recurrent atrial fibrillation (148). Fifteen patients with drug refractory atrial fibrillation underwent a PVI using a novel transballoon ultrasound ablation catheter (Atronix, Inc) (Fig. 25a–c). The ablation system was composed of a 0.035 in. diameter

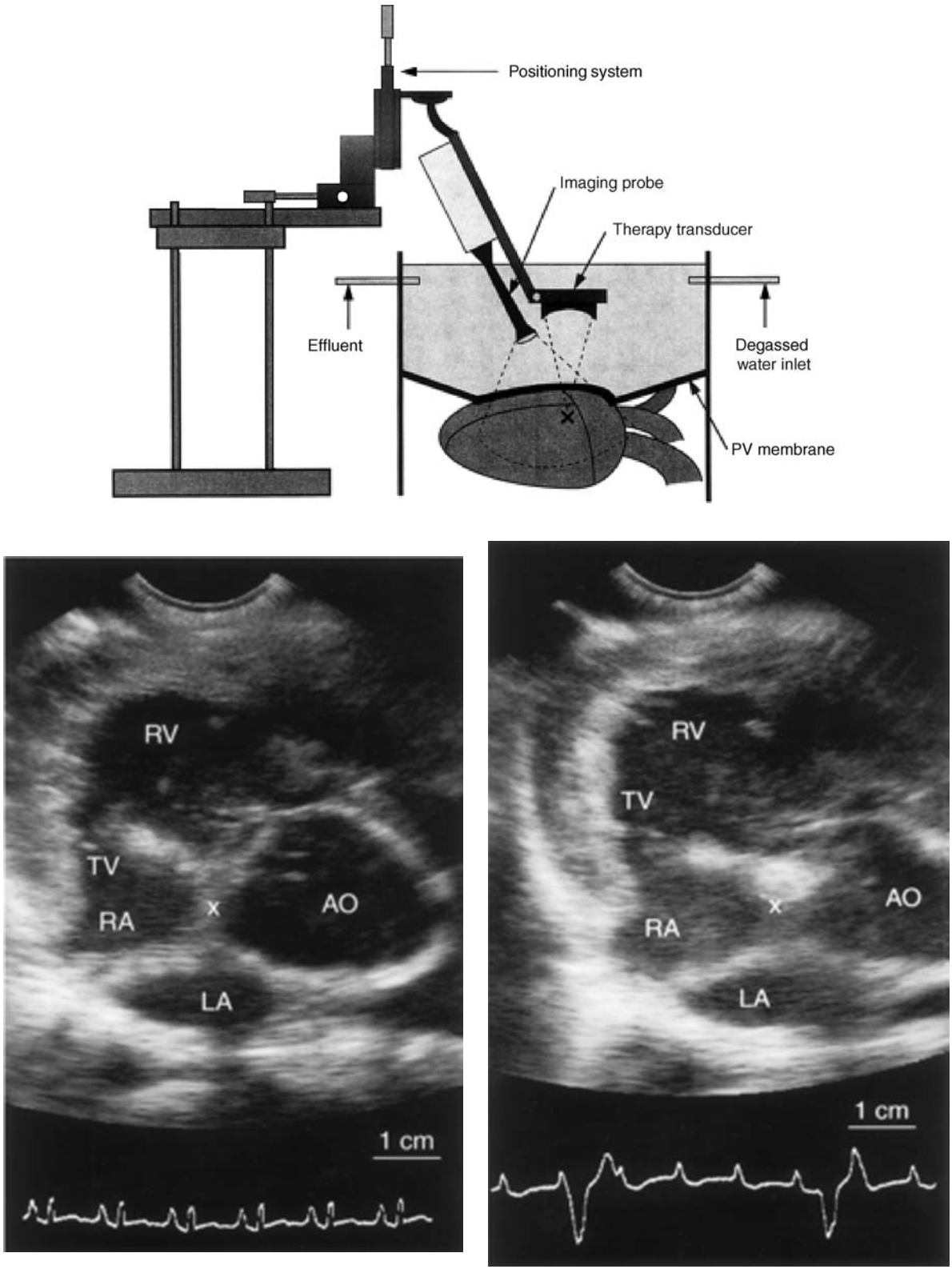


Figure 24. (a) Schematic of the experimental HIFU apparatus. The therapeutic ultrasound transducer is mounted 63.5 mm from the target (X). A polyvinyl chloride membrane covers the heart and the lungs. Degassed water flows in and out of the thoracic cavity at a rate of 600 mL·min^{-1}. Combined diagnostic/ablation transducers are placed into degassed water. (Published with permission from Ref. 146). (b) ECG and Echocardiogram of a canine heart. Prior to ablation of the AV node. (c) After ablation of the AV node using HIFU, the ECG shows complete AV block and the echo image depicts an increased density of the ablated tissue. (Published with permission from Ref. 146).

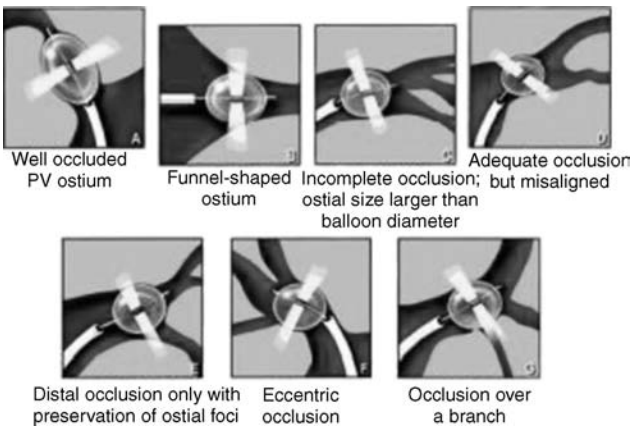


Figure 25. Anatomical pulmonary vein variations and technical limitations. (Published with permission from Refs. 148 and 149).

luminal catheter with a distal balloon (2.2 cm at maximum diameter) housing a centrally located ultrasound transducer (8 MHz). The ultrasound ablation system was advanced over a guide wire (0.035 in., 0.088 cm) to the intended pulmonary vein. The ablation performance and tissue temperature are monitored by thermocouples on the balloon and the therapeutic transducer. The ablation time was 2 min with an additional minute to deflate the balloon. The energy was delivered perpendicularly to the surface of the balloon and ablation at the funnel portion of the pulmonary vein (antrum) could not be achieved with the original design. Additionally, other anatomical characteristics of the target sites like ostial diameter larger than the balloon size, inability to reach the right inferior or other pulmonary vein ostia, ostial instability, early branching of the vein, and eccentric position of the ultrasound transducer in the vein made it difficult to deliver energy effectively (Fig. 26). These technical limitations have been addressed in some of the newer balloon systems where the energy delivery could be accomplished in a divergent angle enabling ablation around the antrum with the tip of the balloon sitting at the pulmonary vein ostium. Early animal studies on HIFU mediated AF ablation have shown promising results with an experimental device that focuses ultrasonic energy via a parabolic balloon, using gas or fluid as a reflector (ProRhythm INC.) (Fig. 25c–e). (149,150).

Since 2003 ~60 patients were treated using this system. With improved catheter design the success rate of AF ablation has increased from ~50 up to 80% without evidence of pulmonary vein stenosis. These preliminary human study results need to be confirmed in larger series. Since there are no large clinical trials on AF ablation with this technology, it is still somewhat premature to predict if HIFU is the complete answer. This device is expected to be released in Europe in 2005 (149,150).

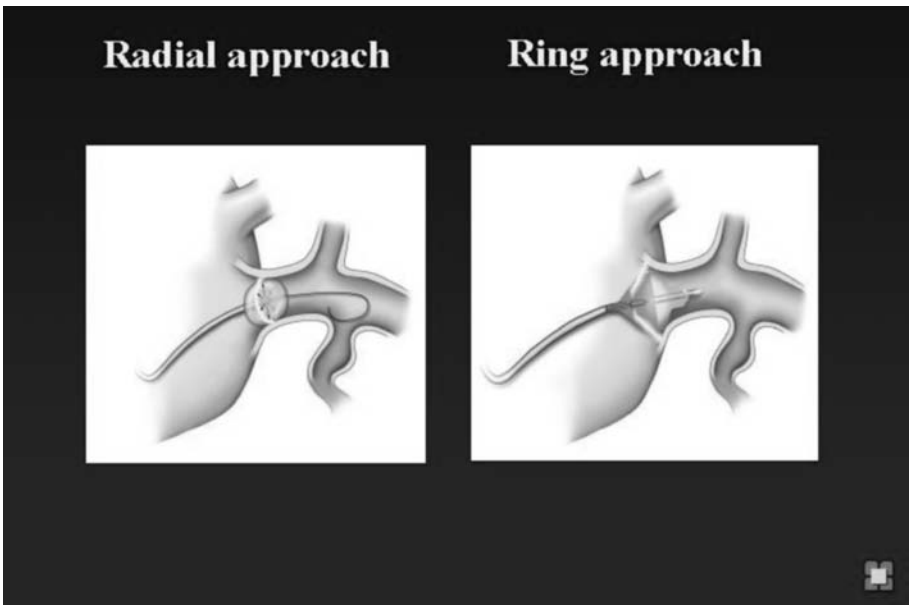
Attempts have been made to harness HIFU for transmyocardial revascularization (TMR) to improve blood supply to damaged myocardium caused by advanced heart disease. Using a 10 cm diameter transducer operating a frequency of 2.52 MHz, intensity of 2300 $W \cdot cm^{-2}$ and pulse repetition period of 40 ms at 50% duty cycle, small channels were successfully created in canine myocardium (151). This shows the potential for future application for HIFU in TMR in a noninvasive fashion.

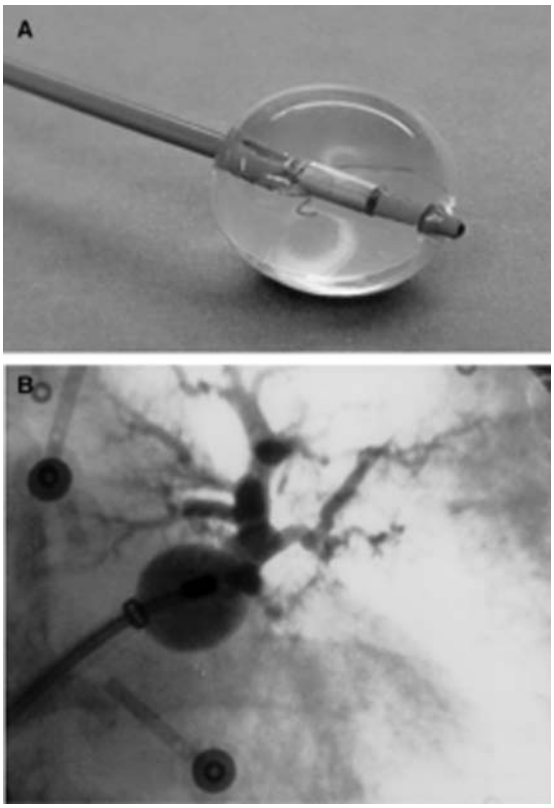
Other Applications of HIFU

Several branches of medicine have already begun to benefit from the use of HIFU with the prospect of many more applications in the future. Thus far the greatest influence of HIFU has been in oncology, with other fields now exploring and experimenting with HIFU to determine its potential utility. The HIFU has been proposed as a tool for synovectomy in the treatment of rheumatoid arthritis (RA) (98,152) and has been used to control opiate refractory pain in pancreatic cancer patients (98,127) and internal bleeding in organs and vessels (98,153). A hand-held HIFU device has been successfully used to perform vasectomies in dogs as a 1–2 min procedure (98,154).

Future Perspectives in Conventional Hyperthermia and HIFU Use

Heat as therapeutic entity has had a rich history punctuated with many successes and failures. The evolution and integration of therapeutic hyperthermia in the clinical setting have been the product of clinical trials, development of new devices, and education of the medical personnel. Conventional hyperthermia has been used for a long time with many energy sources such as electromagnetic, ultrasound, and microwaves. Similarly, it has been >50 years since HIFU was first conceptualized and actualized. Many subspecialties of medicine have benefitted from the use of hyperthermia. Some of the limitations that were miring conventional hyperthermia and HIFU have only recently begun to be overcome and now these therapies can reach a wider patient population. Both of these techniques share similar obstacles due to the limitations that are inherent to ultrasound. Indeed, ultrasound hyperthermia procedures are limited in that ultrasound cannot propagate through air-filled cavities (e.g., lung or bowel). Consequently, lung tumors other than those that are at the periphery are not likely to be amenable to treatment with HIFU or conventional hyperthermia. Also, tumors that are in close proximity to the bowel or within the bowel wall pose an increased chance of visceral perforation with HIFU use. In addition, other side effects such as pain, soft tissue and bone damage, and skin burns have been reported. Often these side effects can be minimized by varying scan paths, altering frequency, power deposition, or the applicator position (7). The success of both hyperthermia techniques is determined by whether ultrasound energy can be properly directed to the site of interest, or if therapeutic temperatures are achieved, and if other factors can be compensated for, such as hemodynamic changes. Other challenges facing ultrasound hyperthermia systems include the ability to gain control over spatial distribution of heat, tissue temperature monitoring, and improved diagnostic visualizations in order to better treat the tumor site. The HIFU treatment times also need to be shortened. Despite this, treatment times of 1 h for a 2 cm superficial tumor using HIFU is preferable to surgical resection; conversely, at present the same tumor can be treated much





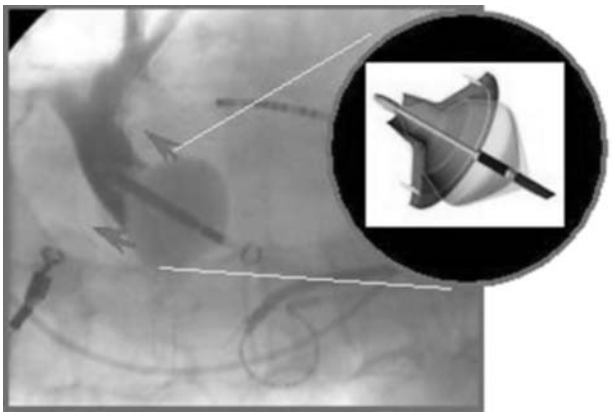


Figure 26. (a) A schematic of the first ultrasound balloon system illustrating the radial delivery of ultrasound energy that is transferred perpendicularly to the tissue. (b) 8F transballoon ultrasound ablation catheter with a central lumen that can accommodate a 0.035 in. guidewire. The distal end of the catheter lodges a cylindrical transducer axially with a saline-filled balloon inflated over it. (c) The balloon is advanced over a guidewire at the ostium of the left upper pulmonary vein. An occlusive pulmonary venogram is done to confirm that the transducer is located at the proximal portion of the vein. Published with permission from Refs. 148 and 149. (d) A schematic of the HIFU device illustrating a ring pattern of ultrasound transmission. The parabolic balloon focuses the ultrasound beam forward. The HIFU catheter is able to create lesions in the antrum away from the lumen of the pulmonary veins thus reducing the likelihood of pulmonary vein stenosis. (e) The HIFU catheter at the ostium of the pulmonary vein.

more quickly with radiofrequency ablation. However, in large tumors longer treatment times are justified because there are very low mortality and morbidity rates associated with the use of HIFU (98).

The benefits of both conventional hyperthermia and HIFU are limitless, but both need more testing in larger patient populations to fully delineate their clinical role. At present, the scarcity of available equipment, the required technical expertise and the lack of thoroughly trained providers, the increased complexity and difficulty involved in using conventional hyperthermia and HIFU when compared to other systems (e.g., electromagnetic) have limited the widespread use of this treatment modality (7). Issues, such as accurate delivery of focused ultrasound to the target tissue, the ability to monitor and control temperatures, and reasonable treatment times, have all been improved upon and will continue to be refined with further testing and research. Imaging modalities like MRI and high resolution CT scan will enhance the precision of HIFU in various system applications. In the future, HIFU could emerge as a good alternative to traditional surgery, and conventional hyperthermia may become a mainstay as an adjunct to chemotherapy and radiation to ultimately improve the arsenal of methodologies available to better treat patients.

BIBLIOGRAPHY

1. Seegenschmiedt MH, Vernon CC. A historical perspective on hyperthermia in oncology. In: Seegenschmiedt MH, Fessenden P, Vernon CC, editors. Thermoradiotherapy and Thermochemotherapy Vol 1. Berlin: Springer Verlag; 1995. pp 3–44.
2. Modern cancer treatment by oncothermia and hyperthermia. 2003. Available at http://www.hot-oncotherm.com/historical.htm. Accessed 2005, Feb 4.
3. Lele P. Hyperthermia Ultrasonics. Encyclopedia of Medical Devices. New York: John Wiley, Sons; 1998.
4. History of Hyperthermia. Available at http://www.starfarm.it/hyperthermia/history.htm. Accessed 2005, April 4.
5. van der Zee J. Heating the patient: A promising approach? Ann Oncol 2002;13:1173–1184.
6. Fesseden P, Hand JW. Hyperthermia therapy physics. In: Smith AR, editor. Medical radiology: Radiation therapy physics. Berlin: Springer Verlag; 1995. pp 315–363.
7. Diederich CJ, Hynynen K. Ultrasound technology for hyperthermia. Ultrasound Med Biol 1999;25:871–887.
8. Bushberg J, Seibert J, Leidholdt E, Boone J. The Essential Physics of Medical Imaging. Philadelphia Lippincott, Williams, and Wilkins; 2002. pp 469–553.
9. Chapelon JY, et al. New Piezoelectric transducers for therapeutic ultrasound. Ultrasound Med. Biol. 2000;26:153–159.
10. Moros E, et al. Thermal contribution of compact bone to intervening tissue-like media exposed to planar ultrasound. Phys Med Biol 2004; 869–886.
11. Madersbacher S, Marberger M. High-energy shockwaves and extracorporeal high-intensity focused ultrasound. J Endourol 2003;17:667–672.
12. ter Haar G. High intensity focused ultrasound for the treatment of tumors. Echocardiography 2001;18:317–321.
13. Novak P, et al. SURLAS: A new clinical grade ultrasound system for sequential or concomitant thermoradiotherapy of superficial tumors: Applicator description.
14. Reinhold H, Endrich B. Tumour microcirculation as a target for hyperthermia. Int J Hypertherma 1986;2:111–137.
15. Vaupel PW, Kelleher DK. Metabolic status and reaction to heat of normal and tumor tissue. In: Seegenschmiedt MH, Fessenden P, Vernon CC. editors. Thermoradiation and Thermochemotherapy, Vol. 1. Berlin: Springer Verlag; 1995. pp 157–176.
16. Raaphorst GP. Fundamental aspects of hyperthermic biology. In: Field SB, Hand JW, editors. An Introduction to the Practical Aspects of Clinical Hyperthermia. London: Taylor and Francis; 1990. pp 10–54.
17. Fajardo LF, Pathological effects of Hyperthermia in normal tissue. Cancer Res 1984;44:4826s–4835s.
18. Sminia P, et al. Effects of hyperthermia on the central nervous system: A review. Int J Hyperthermia 1994;10:1–130.
19. Wondergem J, et al. Effects of local hyperthermia on the motor function of the rat sciatic nerve. Int J Rad Bio 1988;53:429–439.
20. Falk MH, Issels RD. Hyperthermia in oncology. Int J Hyperthermia 2001;17:1–18.
21. Overgaard J, et al. Randomised trial of hyperthermia as adjuvant to radiotherapy for recurrent or metastatic malignant melanoma. The Lancet 1995;345:540–543.
22. Sneed PK, et al. Survival benefit of hyperthermia in a prospective randomized trial of brachytherapy boost (hyperthermia for glioblastoma multiforme). Int J Rad Oncol Biol Phys 1998;40:287–295.
23. Vernon CC, et al. Radiotherapy with or without hyperthermia in the treatment of superficial localized breast cancer: Results from five randomized controlled trials. Int J Rad Oncol Biol Phys 1996;35:713–744.
24. Valdagni R, Amichetti M, Pani G. Radical Radiation alone versus radical radiation plus microwave hyperthermia for N3 (TNM-UICC) neck nodes: A prospective randomized clinical trial. Int J Rad Oncol Biol Phys 1988;15:13–24.
25. Datta N, et al. Head and Neck cancers: Results of thermoradiotherapy versus radiotherapy. Int J Hyperthermia 1990;6:479–486.
26. Perez et al. Randomised phase III study comparing irradiation and hyperthermia with irradiation alone in superficial measurable tumors. Am J Clin Oncol 1991;14:133–141.
27. Jones E, et al. Randomized Trial of Hyperthermia and radiation for superficial tumors. J Clin Oncol 2005;23(13):3079–3085.
28. Emami B, et al. Phase III study of interstitial thermoradiotherapy compared with interstitial radiotherapy alone in the treatment of recurrent or persistent human tumors: A prospectively controlled randomized study by the radiation therapy oncology group. Int J Rad Oncol Biol Phys 1996;34:1097–1104.
29. Harima Y, et al. A randomized clinical trial of radiation therapy versus thermoradiotherapy in stage IIIB cervical carcinoma. Int J Hyperthermia 2001;17(2):97–105.
30. Vasanthan A, et al. Regional hyperthermia combined with radiotherapy for uterine cervical cancers: A multi-institutional prospective randomized trial of international atomic energy agency. Int J Rad Oncol Biol Phys 2005;61(1):145–153.
31. Van der Zee J, et al. Comparison of radiotherapy alone with radiotherapy plus hyperthermia in locally advanced pelvic tumors: a prospective, randomized, multicentre trial. Dutch Deep Heat Hyperthermic Group. Lancet 2000; 1119–1125.
32. Perez CA, Gillespie B, et al. Quality assurance problems in clinical hyperthermia and their impact on therapeutic outcome: a report by the radiation oncology group. International Journal of Radiation Oncology, Biology, and Physics, 16:551–558.
33. Voldagni R, Amichetti M. 1994 Report of long-term follow-up in a randomized trial comparing radiation therapy plus hyperthermia to metastatic lymphnodes in stage IV head and neck patients. International Journal of Radiation Oncology, Biology, and Physics, 28:163–169.

34. Dahl O. Interaction of heat and drugs in vitro and in vivo. In: Seegenschmiedt MH, Fessenden P, Vernon CC, editors. Thermoradiotherapy and Thermochemotherapy, Vol 1. Berlin: Springer Verlag; 1995. pp 103–121.
35. Westermann AM, Grosen EA, et al. A pilot study of whole body hyperthermia and carboplatin in platinum-resistant ovarian cancer. Eur J Cancer 2001;37:1111–1117.
36. Li DJ, Hou BS. A preliminary report on the treatment of esophageal cancer by intraluminal microwave hyperthermia and chemotherapy. Cancer Treatment Reports 1987;71: 1013–1019.
37. Sugemachi K, et al. Chemotherapy combined with or without hyperthermia for patients with oesophageal carcinoma: a prospective randomized trial. International Journal of Hyperthermia 10:485–493.
38. Kakehi M, et al. Multi-institutional clinical studies on hyperthermia combined with radiotherapy or chemotherapy in advanced cancer of deep-seated organs. International Journal of Hyperthermia 6:719–740.
39. Falk RE, et al. Combination therapy for resectable and unresectable adenocarcinoma of the pancres. Cancer 57: 685–688.
40. Issels RD, et al. Ifosfamide plus etoposide combined with regional hyperthermia in patients with locally advanced sarcomas: a phase II study. Journal of Clinical Oncology 1990;8:1818–1829.
41. Issels RD, et al. Improvement of local control by regional hyperthermia combined with systemic chemotherapy(ifosfamide plus etoposide) in advanced sarcomas: updated report on 65 patients. Journal of Cancer Research and Clinical Oncology 1991;117:141–147.
42. Issels RD, et al. Preoperative systemic etoposide/ifosfamide/doxorubicin chemotherapy combined with regional hyperthermia in high-risk sarcoma: a pilot study. Cancer Chemotherapy and Pharmacology 1993;31(suppl. 2) S233–S237.
43. Issels RD, et al. Neoadjuvant chemotherapy combined with regional hyperthermia (RHT) followed by surgery and radiation in primary and recurrent high-risk soft-tissue sarcomas (HR-STS) of adults (updated report). Journal of Cancer Research and Clinical Oncology 1998;124(suppl.) R105.
44. Issels RD. Soft Tissues Sarcomas-What is Currently Being Done. European Journal of Surgical Oncology 21(suppl) 471–474.
45. Eggermont A MM, et al. Isolated Limb Perfusion with high-Dose tumor necrosis factor alpha in combination with interferon gamma and melphalan for nonresectable extremity soft tissue sarcomas: a multicenter trial. Journal of Clinical Oncology 14:2653–2665.
46. Wiedemann GJ, et al. Ifosfamide and carboplatin combined with 41.8 °C whole body hyperthermia in patients with refractory sarcoma and malignant teratoma. Cancer Research 54:5346–5350.
47. Weidemann GJ, et al. Ifosfamide carboplantin, and etopside (ICE) combined with 41.8 °C whole-body hyperthermia in patients with refractory sarcoma. European Journal of Cancer 32A:888-891.
48. Robins HI, et al. Phase I Clinical Trail of Melphalan and 41.8 °C whole-body hyperthermia in cancer patients. Journal of Clinical Oncology 1993;15:154–164.
49. Romanowski R, Schott C. Regionale hyperthermie mit systemischer Chemotherapie bei Kindern und Jugendlichen: Durchfuhrbarkeit und Klinische Verlaufe bei 34 intensiv vorbehandelten Patienten mit prognostisch ungunstigen Tumorerkrankkungen. Klinische Padiatrie 205:249–256.
50. Wessalowski R, Kruck H. Hyperthermia for the treatment of patients with malignant germ cell tumors. A phase I/II study in ten children and adolescents with recurrent or refractory tumors. Cancer 82:793–800.
51. Rietbroek RC, Schilthuis MS. Phase II trial of weekly locoregional hyperthermia and cisplantin in patients with previously irridated recurrent carcinoma of the uterine cervix. Cancer 1997;79:935–942.
52. Takahashi M, Fujimot S. Clinical outcome of intraoperative pelvic hyperthermochemotherapy for patients with Dukes' C rectal cancer. International Journal of Hyperthermia 10:749–754.
53. Westermann AM, Wiedemann GJ. A Systemic Hyperthermia Oncologic Working Group trial. Ifosfamide, carboplatin, and etoposide combined with 41.8 degrees C whole-body hyperthermia for metastatic soft tissue sarcoma. Oncology. 2003;64(4):312–21.
54. Anhalt DP, Hynynen K, Roemer RB. Patterns of changes of tumour temperatures during clinical hyperthermia: Implications for treatment planning, evaluation and control. Int J Hyperthermia 1995;11:425–436.
55. Corry PM, Barlogie B, Tilchen EJ, Armour EP. Ultrasound induced hyperthermia or the treatment of human superficial tumors. Int J Rad Oncol Biol Phys 1982;8:1225–1229.
56. Corry PM, et al. Combined ultrasound and radiation therapy treatment of human superficial tumors. Radiology 1982b;145: 165–169.
57. Harrison GH. Ultrasound hyperthermia applicators: Intensity distributions and quality assurances. Int J Hyperthermia 1990;6:169–174.
58. Marmor JB, Hahn GM. Ultrasound heating in previously irradiated sites. Int J Rad Oncol Biol Phys 1978;4:1029–1032.
59. Marmor JB, Hahn GM. Combined radiation and hyperthermia in superficial human tumors. Cancer 1980;46:1986–1991.
60. Marmor JB, Pounds D, Hahn GM. Clinical studies with ultrasound induced hyperthermia. Natl Cancer Inst Monogram 1982;61:333–337.
61. Marmor JB, Pounds D, Hahn GM. Treatment of superficial human neoplasms by local hyperthermia induced by ultrasound. Cancer 1979;43:188–197.
62. Labthermics Technologies Online.1998–1999. Available at http://www.labthermics.com/hyper.html. Accessed 2005 Jan 26.
63. Anhalt DP, et al. Scanned ultrasound hyperthermia for treating superficial disease. Hyperthermic oncology. vol 2. Proceedings of the 6th international Congress on Hyperthermic Oncology, Tucson, (AZ); 1992. p 191-192.
64. Lele PP. Advanced ultrasonic techniques for local tumor hyperthermia. Radiol Clin N Am 1989;27:559–575.
65. Harari PM, et al. Development of scanned focused ultrasound hyperthermia: clinical response evaluation. Int J Rad Oncol Biol Phys 1991;21:831–840.
66. Hand JW, Vernon CC, Prior MV. Early experience of a commercial scanned focused ultrasound hyperthermia system. Int J Hyperthermia 1992;8:587–607.
67. Guthkelch AN, et al. Treatment of malignant brain tumors with focused ultrasound hyperthermia and radiation: Results of a phase I trial. J Neuro Oncol 1991;10:271–284.
68. Duthkelch et al.
69. Formine et al.
70. Straube WL, et al. An ultrasound system for simultaneous ultrasound hyperthermia and photon beam irradiation. Int J Rad Oncol Biol Phys 1996;36:1189–1200.
71. Lu XQ, et al. Design of an ultrasonic therapy system for breast cancer treatment. Int J Hyperthermia 1996;12:375–399.
72. Moros EG, Fan X, Straube WL. An investigation of penetration depth control using parallel opposed ultrasound hyperthermia. J Accoust Soc Am 1997;101:1734–1741.

73. Moros EG, Fan X, Straube WL, Myerson RJ. Numerical and in vitro evaluation of temperature fluctuations during reflected-scanned planar ultrasound hyperthermia. Int J Hyperthermia 1998;14:367–382.
74. Moros EG, Myerson RJ, Straube WL. Aperture size to therapeutic volume relation for a multi-element ultrasound system: Determination of applicator adequacy for superficial hyperthermia. Med Phys 1993;20:1399–1409.
75. Moros EG, Roemer RB, Hynynen K. Simulations of scanned focused ultrasound hyperthermia. The effects of scanning speed and pattern on the temperature fluctuations at focal depth. IEEE Trans Ultrason Ferroelec Frequency Control 1988;35:552–560.
76. Moros EG, et al. Simultaneous delivery of electronic beam therapy and ultrasound hyperthermia using scanning reflectors: a feasibility study. Int J Rad Oncol Biol Phys 1995;31: 893–904.
77. Moros EG, Straube WL, Myerson RJ. Potential for power deposition conformability using reflected-scanned planar ultrasound. Int J Hyperthermia 1996;12:723–736.
78. Lele PP, Parker KJ. Temperature distributions in tissues during local hyperthermia by stationary or steered beams of unfocused or focused ultrasound. Br J Cancer 1982;45 (Suppl):108–121.
79. Hynynen K, et al. A scanned, focused, multiple transducer ultrasonic system for localized hyperthermia treatments. Int J Hyperthermia 1987;3:21–35.
80. Lin W, Roemer RB, Hynynen K. Theoretical and experimental evaluation of a temperature controller for scanned focused ultrasound hyperthermia. Med Phys 1990;17:615–625.
81. Ibbini MS, Cain CA. The concentric-ring array for ultrasound hyperthermia: combined mechanical and electrical scanning. Int J Hyperthermia 1990;6:401–419.
82. Umemura S, Cain CA. The sector-vortex phased array: acoustic field synthesis for hyperthermia. IEEE Trans Ultraon Ferroelec Frequency Control 1989;36:249–257.
83. Ebbini ES, Cain CA. Experimental evaluation of a prototype cylindrical section ultrasound hyperthermia phased array applicator. IEEE Trans Ultrason Ferroelec Frequency Control 1991a;38:510–520.
84. Ebbini ES, Cain CA. A spherical-section ultrasound phased array applicator for deep localized hyperthermia. IEEE Trans Biomed Eng 1991b;38:634–643.
85. Benkeser PJ, Frizzell LA, Goss SA, Cain CA. Analysis of a multielement ultrasound hyperthermia applicator. IEEE Trans Ultrason Ferroelec Frequency Control 1989;36:319–325.
86. Diederich CJ, Hynynen K. Induction of hyperthermia using an intracavitary ultrasonic applicator. IEEE Ultrason Sympos Proc 1987;2:871–874.
87. Diederich CJ, Hynynen K. Induction of hyperthermia using an intracavitary multielement ultrasonic applicator. IEEE Trans Biomed Eng 1989;36:432–438.
88. Diederich CJ, Hynynen K. The development of intracavitary ultrasonic applicators for hyperthermia: A design and experimental study. Med Phys 1990;17:626–634.
89. Smith NB, Buchanan MT, Hynynen K. Transrectal ultrasound applicator for prostate heating monitored using MRI thermometry. Int J Rada Oncl Biol Phys 1999;33:217–225.
90. Diederich CJ. Ultrasound applicators with integrated catheter-cooling for interstitial hyperthermia: Theory and preliminary experiments. Int J Hyperthermia 1996;12:279–297.
91. Diederich CJ, Hynynen K. Ultrasound technology for interstitial hyperthermia. In: Seegenschmiedt MH, Sauer R, editors. Interstitial and intracavitary thermoradiotherapy. Berlin: Springer-Verlag; 1993. pp 55–61.
92. Hynynen K. The feasibility of interstitial ultrasound hyperthermia. Med Phys 1992;19:979–987.
93. Hynynen K, Davis KL. Small cylindrical ultrasound sources for induction of hyperthermia via body cavities or interstitial implants. Int J Hyperthermia 1993;9:263–274.
94. Lee RJ, Klein LJ, Hynynen K. A multi-element and multi-catheter ultrasound system for interstitial hyperthermia. IEEE Trans Biomed Eng 1999.
95. Deardorff DL, Diederich CJ, Nau WH. Air-cooling of direct-coupled ultrasound for interstitial hyperthermia and thermal coagulation. Med Phys 1998;25:2400–2409.
96. Diederich CJ, et al. Direct coupled interstitial ultrasound applicators for simultaneous thermobrachytherapy: A feasibility study. Int J Hyperthermia 1996;12:401–419.
97. Jarosz BJ. Feasibility of ultrasound hyperthermia with waveguide interstitial applicator. IEEE Trans Biomed Eng 1996;6:1106–1115.
98. Kennedy J, et al. High intensity ultrasound: Surgery of the future? Br J Rad 2003;76:590–599.
99. Lynn JG, Zwemer RL, Chick AJ, Miller AG. A new method for generation and use of focused ultrasound in experimental biology. J Gen Physiol 1942;26:179–193.
100. Fry WJ, et al. Ultrasonic lesions in the mammalian central nervous system. Science 1955;122:517–518.
101. Fry WJ, Mosberg WH, Barnard JW, Fry FJ. Production of focal destructive lesions in the central nervous system with ultrasound. J Neurosurg 1954;11:471–478.
102. Fry FJ. Precision high-intensity focusing ultrasonic machines for surgery. Am J Phys Med 1958;37:152–156.
103. Ballantine HT, Bell E, Manlapaz J. Progress and problems in the neurological application of focused ultrasound. J Neurosurg. 1960;17:858–876.
104. Warwick R, Pond JB. Trackless lesions in nervous tissues produced by HIFU (high-intensity mechanical waves). J Anat 1968;102:387–405.
105. Lele PP. Concurrent detection of the production of ultrasonic lesions. Med Biol Eng 1966;4;451–456.
106. Lele PP. Production of deep focal lesions by focused ultrasound-current status. Ultrasonics 1967;5:105–112.
107. Burov AK. High-intensity ultrasonic vibrations for action on animal and human malignant tumours. Dokl Akad Nauk SSSR 1956;106:239–241.
108. Kohrmann KU, et al. Technical characterization of an ultrasound source for noninvasive thermoablation by high-intensity focused ultrasound. BJU Int 2002;90:248–252.
109. ter Haar G. Robertson D: Tissue destruction with focused ultrasound *in vivo*. Eur Urol 1993;23(Suppl. 1):8–11.
110. Clement GT, Perspectives in clinical uses of high-intensity focused ultrasound. Ultrasonics 2004;42:1087–1093.
111. Holt RG, Roy RA, Edson PA, Yang X. Bubbles and HIFU: the good, the bad, and the ugly. In: Andrew MA, Crum LA, Vaezy S, editors. Proceedings of the 2nd International Symposium on Therapeutic Ultrasound; 2002. pp 120–131.
112. Sokka SD, King R, Hynynen K. MRI-guided gas bubble enhanced ultrasound heating in *in vivo* rabbit thigh. Phys Med Biol 2003;48:223–241.
113. Billard BE, Hynynen K, Roemer RB. Effects of physical parameters on high temperature ultrasound hyperthermia. Ultrasound Med Biol 1990;16:409–420.
114. Kolios MC, Sherar MD, Hunt JW. Blood Flow cooling and Ultrasonic lesion formation. Med Phys 1996;23:1287–1298.
115. Chan A, et al. An image-guided high intensity focused ultrasound device for uterine fibroids treatment. Med Phys 2002;29:2611–2620. Otsuka R, et al. In vitro ablation of cardiac valves using high intensity focused ultrasound. Ultrasound Med Biol 2005;31:109–114.

116. Wu F, et al. Pathological changes in human malignant carinoma treated with high-intensity focused ultrasound. Ultrasound Med Biol 2001;27:1099–1106.
117. Chaussy C, Thuroff S. High-intensity focused ultrasound in prostate cancer: Results after 3 years. Mol Urol 2000;4:179–182.
118. Madersbacher S, et al. Tissue ablation in benign hyperplasia with high-intensity focused ultrasound. Eur Urol 1993;23 (Suppl. 1):39–43.
119. Hynynen K, et al. MR imaging-guiding focused ultrasound surgery of fibroadenomas in the breast: A feasibility study. Radiology 2001;219:176–185.
120. Hynynen K, et al. A clinical noninvasive MRI monitored ultrasound surgery method. RadioGraphics 1996;16:185–195.
121. Hazel JD, Stafford RJ, Price RE. Magnetic resonance imaging-guided focused ultrasound thermal therapy in experimental animal models: Correlation of ablation volumes with pathology in rabbit muscle and VX2 tumors. J Magnet Reson Imag 2002;15:185–194.
122. Weidensteiner C, et al. Real time MR temperature mapping of rabbit liver in vivo during thermal ablation. Mag Reson Imag 2003;50:322–330.
123. Chan AH, et al. An Image-guided high intensity focused ultrasound device for uterine fibroids treatment. Med Phys 2002;29:2611–2620.
124. Tempany CMC, et al. MR imaging-guided focused ultrasound surgery of uterine leiomyomas: A feasibility study. Radiology 2003;226:897–905.
125. Stewart EA, et al. Focused Ultrasound treatment of uterine fibroid tumors: Safety and feasibility of a noninvasive thermoablative technique. Am J Obstet Gynecol 2003;189:48–54.
126. Wu F, et al. Changes in ultrasonic image of tissue damaged by high intensity ultrasound *in vivo*. J acoustic Soc Am 1998;103:2869.
127. Wu F, Wang ZB, Chen WZ, Zou JZ. Extracorporeal High-Intensity Focused Ultrasound for treatment of solid carcinomas: Four-year Chinese clinical experience. Proceedings of the 2nd International Symposium on Therapeutic Ultrasound; July 29-Aug1; Seattle; 2002.
128. Kennedy JE, et al. High- Intensity focused ultrasound for the treatment of liver tumours. Ultrasonics 2004;42:931–935.
129. Wu F, et al. Extracorporeal high intensity focused ultrasound ablation in the treatment of patients with large hepatocellular carcinoma. Ann Surg Oncol 2004;11(12): 1061–1069.
130. Uchida T, et al. Transrectal high-intensity focused ultrasound for treatment of patients with stage T1b-2NOMO localized prostate cancer: A preliminary report. Urology 2002;59:394–399.
131. Blana A, Walter B, Rogenhofer S, and Wieland W. High-Intensity focused ultrasound for the treatment of localized prostate cancer: 5-year experence. Urology 2004;63:297–300.
132. Madersbacher S, et al. Effect of high intensity focused ultrasound on human prostate cancer *in vivo*. Cancer Res 1995;55:3346–3351.
133. Gelet A, et al. Treatment of prostate cancer with transrectal focused ultrasound: Early clinical experience. Eur Urol 1996;29:174–183.
134. Chaussy C, Thuroff S, Lacoste F, Gelet A. HIFU and prostate cancer: The European experience. Proceedings of the 2nd International Symposium on Therapeutic Ultrasound; July 29–Aug 1; Seattle; 2002.
135. Beerlage HP, et al. High-intensity focused ultrasound followed after one to two weeks by radical retropubic prostatectomy: Results of a prospective study. Prostate 1999;39:41–46.
136. Beerlage HP, et al. Transrectal high-intensity focused ultrasound using the Ablatherm device in treatment of localised prostate carcinoma. Urology 1999;54:273–277.
137. Kohrmann KU, et al. High-intensity focused ultrasound for noninvasive tissue ablation in the kidney, prostate, and uterus. J Urol 2000;163 (4Suppl.):156.
138. Wu F, et al. A randomized clinical trial of high-intensity focused ultrasound ablation for the treatment of patients with localised breast cancer. BJC 2003;89:2227–2233.
139. Mc Pherson K, Steel CM, Dixon JM. Breast cancer epidemiology, risk factors, and genetics. Br J Med 2000;321: 624–628.
140. Curran D, et al. Quality of life in early stage breast cancer patients treated with radical mastectomy or breast-converging procedure: Results of EORTC trial 10801. The European Organization for Research and Treatment of Cancer (EORTC), Breast Cancer Cooperative Group (BCCG). Eur J Cancer 1998;34:307–314.
141. Aubry J, et al. Experimental demonstration of noninvasive transskull adaptive focused based on prior computed tomography scans. J Acoust Soc Am 2003;113:84–93.
142. Clement GT, Hynynen K. A noninvasive method for focusing ultrasound through the human skull. Phys Med Biol 2002;47:1219–1236.
143. Foley J, et al. Image-guided HIFU neurolysis of peripheral nerves to treat spasticity and pain. Ultrasound Med Bio 2004;30:1199–1207.
144. Ohtsuka T, et al. Thoracoscopic internal mammary artery harvest for MICABG using the Harmonic Scalpel. Ann Thorac Surg 1997 June; 63 (6Suppl):S10.
145. Lee LA, et al. High intensity focused ultrasound effect on cardiac tissues: Potential for clinical application. Echocardiography 2000 Aug; 17(6 Pt 1):563–566.
146. Strickberger SA, et al. Extracardiac ablation of the canine atrioventricular junction by use of high-intensity focused ultrasound. Circulation 1999;100:203–208.
147. Natale A. Cleveland Clinic Foundation. Personal Interview. March 5, 2005.
148. Natale A, et al. First Human Experience with Pulmonary vein isolation using a through-the-balloon circumferential ultrasound ablation system for recurrent atrial fibrillation. Circulation 2000;102:1879.
149. Saliba W, et al. Circumferential ultrasound ablation for pulmonary vein isolation: Analysis of acute and chronic failures. J of Cardiovasc Electrophysiol 2002;13:957–961.
150. Ayoma H, et al. Circumferential lesion characteristic of high intensity focused ultrasound balloon catheter for pulmonary vein isolation. Heart Rhythm 2004;1:S430.
151. Smith NB, Hynyen K. The feasibility of using focused ultrasound for transmyocardial revascularization Ultrasound Med Biol 1998;24:1045–1054.
152. Foldes K, et al. Magnetic resonance imaging-guided focused ultrasound synovectomy. Scand J Rheumal 1999;28:233–237.
153. Vaezy S, et al. Hemostasis of punctured blood vessels using high-intensity focused ultrasound. Ultrasound Med Biol 1998;24:903–910.
154. Roberts WW, et al. High-Intensity focused ultrasound ablation of the vas deferens in canine model. J Urol 2002; 167:2613–2617.

See also HYPERTHERMIA, INTERSTITIAL; HYPERTHERMIA, SYSTEMIC; THERMOMETRY.

HYPOTHERMIA.

See TEMPERATURE MONITORING.

I

IABP. See INTRAAORTIC BALLOON PUMP.

IMAGE INTENSIFIERS AND FLUOROSCOPY

MELVIN P. SIEDBAND
University of Wisconsin
Fitchburg, Wisconsin

INTRODUCTION

Early fluoroscopic systems used a phosphor-coated sheet or screen to convert incident X-ray photons to light. The radiologist observed the image through a lead glass protective screen. A film camera was often used to record the image. There are two serious disadvantages to this method: the radiologist had to be dark adapted so that image details were hard to see and the collection solid angle of the eye or camera lens was small. The small collection angle meant that the X-ray exposure had to be increased by almost 100 times to have the same diagnostic quality as a conventional radiograph. Photographing the fluoroscopic screen, photofluorography, is no longer used because of the high exposure to the patients. All X-ray images are noise limited by the finite number of X-ray quanta detected and seen. A 50 keV X-ray photon can, at best, produce ~1000 visible light photons, if absorbed by the old-type phosphor. About 5–10% of the incident X-ray photons are stopped and converted to light by the fluorescent screen. The quantum detection efficiency, (QDE) of the screen is the product of the ability to absorb the incident X-ray photon and the probability of emitting light. A thicker screen would absorb more photons, but would also cause more lateral spreading of the light and reduce the resolution of the image. The 5–10% QDE figure is a practical compromise between resolution and sensitivity. If it is assumed that the visible light photons are emitted isotropically, then the lens of the eye or camera subtends only a very small fraction of this light radiation hemisphere. A sheet of film in a radiographic cassette has a phosphor-coated sheet on either side and can collect light photons far more efficiently.

The invention of the image intensifier overcame these objections. The concept of the early image intensifiers was to use a thin, curved, glass meniscus, ~15 cm diameter, coated on the convex side with a scintillator, originally of the same composition as the zinc:cadmium sulfide fluoroscopic phosphor plates, and a photoemitter on the concave surface. Light produced by the scintillator did not have far to travel to excite the photoemitter. This assembly was placed in a vacuum tube and the photoelectrons were accelerated toward an output viewing screen where a small and very bright image was produced. Because the photoemitter was in close optical contact with the scintillator, the collection angle was very large so that a higher radiation exposure was not needed and the image at the viewing screen was bright enough so that dark adaptation was obviated and fine details could be seen. An optical viewer comprising an objective lens and relay lens, similar in design to a submarine periscope, was used to observe the image.

Modern image intensifiers use an epitaxially grown scintillator of CsI with the photoemitter deposited directly on the surface of that layer. Because the epitaxial layer can be made much thicker than the powdery deposit of the older type scintillator for the same resolution, the new image intensifiers require less exposure–image than conventional radiographs. The thicker layer has a higher QDE than the fluoroscopic screens (Fig. 1).

Many modern image intensifiers use a thin curved steel window on which the scintillator and photoemitter are deposited. This geometry eliminates the X-ray scatter produced by the glass window of the older tubes and improves image contrast. Larger tubes up to 35 cm sensor diameter have been made with variable magnification or zoom capability. The window of thin steel is the flash with a coating of aluminum or other metal and etched to create a domain structure, similar to the domains seen in the zinc coating of galvanized steel. The domains here are very small, ~ 100 μm diameter. Cesium iodide is then vapor deposited on this surface and forms epitaxially (i.e., crystal growth follows the orientation of the metallic substrate), and grows as a collection of optically isolated fibers. This scintillator can be made quite thick with little light spreading. The greater thickness means higher quantum efficiency (i.e., a measure of the fraction of incident X-ray photons converted to light photons), when compared to conventional phosphor plate scintillators. A thin layer of silver and antimony is vapor deposited on top of the scintillator and the final sensitization is accomplished by depositing cesium from heated tubes or reservoirs after assembly in the vacuum tube. The AgCs:Sb photoemitter on the surface of the scintillator is similar to the photocathode of a photomultiplier (PMT), tube.

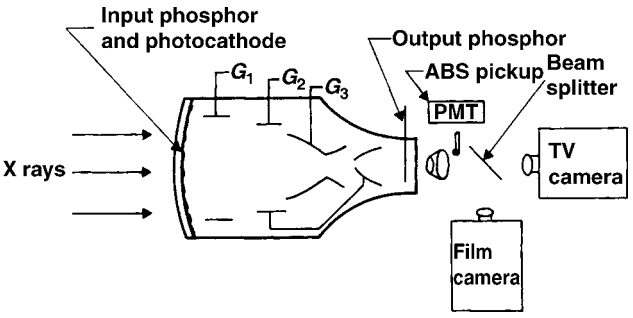


Figure 1. In the image intensifier, X rays strike the input phosphor screen, thus generating light. Light stimulates the photocathode to emit electrons, which are accelerated through 25 kV to strike the output phosphor screen. Brightness gain is due to both geometric gain and electronic gain.

The vacuum tube uses a series of metal cylinders between the photoemitter and the output phosphor. The photoemitter, cylinders, and output phosphor are connected to voltage sources to create shaped electric fields. The field potential at any point affects an electron beam in the same way that the index of refraction of a glass lens affects a beam of light. These cylinders and their potentials form electrostatic lenses to focus the photoelectrons to produce the small and bright image on the output phosphor. Like optical lens elements, the metal cylinders can make compound electrostatic lenses so that focal length can be varied to change the size of the output image: variable zoom.

The brightness gain of an image intensifier is a measure that compares the brightness of the image at the small output screen to that of the older fluoroscopic screens. This gain is a result of the added energy imparted to the photoelectrons, the higher probability of stopping an incident X-ray photon, and the effect of compressing the large input area signal to the small area of the output screen. The older term, brightness gain, has been replaced by conversion efficiency. This term is defined as the output luminance in candela·m^{-2} for an input exposure rate of 1 mR·s^{-1} or 10 μGy·s^{-1}. If assumed it is that a single 50 keV input X-ray photon has a 50% probability of interacting with the input scintillator and producing light (i.e., QDE is 50%) and produces 2000 visible light photons, of which ~ 1000 reach the photoemitter. About 100 of these will produce electrons that will be accelerated by the electric field across the tube and strike the output phosphor as 25 keV electrons. Each 25 keV photoelectron has about a 10% probability of converting its energy into 2.2 eV visible light photons. The QDE, quantum detection efficiency, of the image intensifier is assumed to be 50% because of the thicker CsI scintillator.

An input exposure rate in the diagnostic range produces ~200,000 X-ray photons·mm^{-2}·s^{-1}, converts to 100,000 light photons/X-ray photon. In the visible range, ~10^{10} photons/mm^2 have a light intensity of 1 candela/m^2. Substituting in the above, a 1 mR (10 μGy) input would produce ~1 candela·m^{-2} for an output screen the same size as the input scintillator. This is >200 times brighter than the older fluoroscopic screens. Because the output screen area of an image intensifier is 25 mm diameter for an input screen of diameter of 220 mm, the same number of output light photons are emitted from a smaller area resulting in an additional gain factor of ~75. This yields a total brightness gain of >15,000 over the old fluoroscopic screen!

A fluoroscopic television system is used for dynamic studies, to enable the viewer to see how a contrast agent is swallowed, how blood flows, to locate objects for a surgical procedure, and so on. During the study, an image record (formerly a film record) could be made for later examination using a video recorder or computer. A rapid succession of images could be recorded and then viewed to find the few images revealing the particular problem, for example, how a heart valve functioned or to diagnose an eroded region in the esophagus as the contrast medium sped by. Or a single image could be made to show the problem for later correction.

A fast optical lens optically collimates the small output image of the image intensifier. Collimation, in this case, means that the optical object is at the focal plane of the collimating lens and its image is focused on an infinite distance away. Any optical device with its own objective lens, such as a TV or film camera, could be aimed through the collimating lens and its image size would be the original intensified image size times the ratio of (objective focal length)/(collimator focal length). The lenses could be separated by several centimeters before image vignetting occurs. A beam splitting mirror can be interposed on the collimator and the objective lenses so that image light is simultaneously apportioned between a TV and film camera. A small mirror or prism and lens could sample the light and form an image over a small hole in the cover of a PMT tube. The hole would permit only light from the center of the image to reach the PMT. The output of the PMT is used to control the X-ray tube current to maintain constant image brightness for continuous viewing or for automatic exposure control of one or a sequence of recorded images. This automatic brightness stabilizer scheme is similar in concept to automatic exposure control of most digital cameras.

The quality of an X-ray image is a compromise between exposure to the patient and the noise or "graininess" of the image. Most images are made using the ALARA principal (As Low As Reasonably Achievable). For any given X-ray exposure, there are a finite number of X-ray photons incident on the patient and then, through the patient, incident on the image sensor. The statistics of photons–area follow the Poisson distribution, so that the variance (noise) is the square root of the average number of photons in a pixel (picture element). To produce a second image having twice the linear resolution (detail) as the first requires four times the exposure. Because the eye averages exposure time >0.2 s, to record an equivalent image in 0.02 s requires an exposure rate 10 times greater. X-ray exposure requirements are determined by the by the X-ray absorption of the patient, diagnostic needs and are different for continuous viewing (real-time fluoroscopy), a sequence of images (video or motion pictures), or single images for later diagnoses.

Before the rapid growth and improvement of digital cameras and computer technologies, still and motion picture film cameras were the only practical means to record images. Because of differences of integrating capability of the eye and detail required of the recorded film images, the X-ray beam current, pulse width (exposure time/image), and so on, the ratio of transmission/reflection of the beam splitting mirror, and other operating parameters, must, be adjusted to obtain the required image quality for each image application requirement while minimizing total exposure to the patient.

Most modern systems use charge-coupled image sensors with a high dynamic range and pulse the X-ray beam current to the required level while digitally recording video images. Computer display of selected images or a dynamic sequence of images has largely displaced motion picture film techniques. Video tape recorders are used in simple systems. The automatic brightness control–automatic exposure control of the digital system uses signals from selected image areas derived from the computer image to optimize image quality in the region of interest.

BIBLIOGRAPHY

Further Reading

Gebauer A, Lissner J, Schott O. Roentgen Television. New York: Grune & Stratton; 1966.

Siedband MP. Image storage subtraction techniques and contrast enhancement by electronic means. Symposium on the Physics of Diagnostic Radiology; University of California, June 1968.

Siedband MP. Image intensification and television. In: Taveras, Ferrucci, editors. Radiology, Diagnosis, Imaging, Intervention. Chapt. 10. New York: Lippincott; 1990.

Siedband MP, Duffy PA. Brightness Stabilizer with Improved Image Quality, US patent No. 3,585,391. Accessed 1971.

Siedband MP. X-ray image storage, reproduction and comparison system. US patent No. 3,582,651, 1971.

IMAGING, CELLULAR. See CELLULAR IMAGING.

IMAGING DEVICES

MARK J. RIVARD
Tufts New England Medical Center
FRANK VAN DEN HEUVAL
Wayne State University

INTRODUCTION

Historically, external radiation treatment of deep-seated malignancies was performed using ortho-voltage equipment. The radiological characteristics of these beams caused maximum dose deposition to occur on the skin of the patient. At that time, skin damage was the limiting factor for dose delivery to the tumor. When the skin turned red due to radiation damage (erythema), the physician had to find another area or portal through which to deliver radiation. The portal was then defined by its orientation and the surface of skin it irradiated.

Nowadays, the treatment is performed with higher photon energies that permit a skin-sparing effect (i.e., the dose at the skin is lower than that deposited a few centimeters deeper) due to the absence of electronic equilibrium. The historic name "portal" still denotes a radiotherapy treatment beam oriented for entry within a patient. Physicians verify whether the treatment is correct using megavoltage treatment beams (4–20 MV photons) as an imaging tool. A transmission image, obtained much like a diagnostic transmission image, provides information describing the patient anatomy and gives clues on the beam orientation and positioning, but also on the extent and shape of the treated area (or portal field). As such, portal imaging is the most direct manner to confirm accuracy of treatment delivery.

Traditional portal verification is done using radiographic films, much like the classical diagnostic films. Films are positioned at the beam exit side of the irradiated patient. Portal image analysis involves comparison with a simulation image that is typically obtained using diagnostic quality X rays (60–120 kV photons). The simulation image serves as the reference image, showing anatomical information clearly and delineating the intended treatment field. Comparison of the simulation image with the portal image is complicated due to the inherent poor quality obtained when imaging using high energy photons (1). The whole procedure of patient positioning, artifact removing, imaging processing, and evaluation using film represents a significant fraction of the total treatment time. This procedure increases the workload per patient, and as a result, the number of images taken is minimized due to economic concerns rather than concerns for efficiency or treatment quality. Indeed, studies have demonstrated that weekly portal image verification, which is the current clinical standard, does not guarantee accurate treatment setup for a population of patients (2).

Portal imaging differs considerably from diagnostic transmission imaging. The main difference is the photon energies used to generate the images. In diagnostic imaging, photons having energies ranging from 50 to 120 kV interact in patients primarily via the photoelectric effect. The cross-section for these interactions is highly dependent on the atomic number of the medium in which they traverse: A higher atomic number increases the probability of interaction. The average atomic number of bony anatomy is higher than that of soft-tissue, yielding good contrast for the bony anatomy. At treatment energies (1–10 MeV) the predominant photon interaction is Compton scattering. The cross-section for this interaction is largely dependent on the media density, and the resulting image will show the largest contrast when large differences in density are present. In practice, this means that differences in soft tissues will contribute most to the visible signal.

These considerations imply that the dynamic range of an electronic portal imaging detector (EPID) is used to obtain information on soft-tissue variations (3), divergence effects (4), scatter contributions (5), field-edge information, and in the case of fluoroscopic imagers: vignetting and glare. With the exception of field-edge information, all of these factors are nonlocalized and tend to change gradually within an image. Not only are these features slowly varying, but they also have a large contrast-to-noise ratio (CNR) compared to the clinically important bone–soft-tissue contrast.

The EPIDs permit the same tasks as film-based imaging, but increase the efficiency and provide added value by using digital imaging techniques. The EPIDs are devices that electronically capture the photon energy fluence transmitted through a patient irradiated during treatment, and allow direct digitization of the image. This image is then immediately available for visualization on a computer screen and electronic storage. When the treatment verification process uses EPIDs, departmental efficiency is increased and quality is improved at the same cost as when using film-based imaging.

Proposals to generate electronic images started in the beginning of the 1980s mainly through the work of Bailey et al. (6), who used systems based on video techniques. This seminal work was then further developed toward more clinically applicable systems by Shalev and co-workers (7), Leong (8), Munro et al. (9), and Visser et al. (10). All

of these systems were the basis for the first generation of commercially available EPIDs. They all combined an analog camera with a fluorescent screen generating the optical coupling using a mirror system. Wong et al. (11) replaced the mirror system with optical fibers (one for each pixel). The technology developed further, and is described below in greater detail.

PHYSICAL ASPECTS OF ELECTRONIC PORTAL IMAGING TECHNOLOGY

Camera-Based Detectors

The initial experience using EPIDs was obtained using camera-based systems. Again similar to film-based portal imaging, the camera-based systems measured the photon energy fluence exiting the patient. However, phosphorescent and fluorescent screens replaced the film, and a mirror was oriented at an angle of 45° to reflect the screen toward a video camera. Subsequently, the image was digitized. Because of the intrinsically low-detector efficiency, bulky detector size, and poor image quality, this technology has now become outdated in comparison with more sophisticated technologies.

The low-detector efficiency was due to many limitations in the signal path from screen to computer. Screen conversion efficiency was not ideal when using Gd_2O_2S. In addition, $<0.1\%$ of the light emitted reached the video camera, due to the poor light collection efficiency of the video camera lens. This low rate of signal collection was subsequently impacted by competing electronic noise from the camera in close proximity to the operating linear accelerator (linac). Also, image acquisition typically required a full treatment fraction as compared to the technique using partial fraction irradiation that is commonly used for radiographic portal imaging. Due to the camera-based system detector orientation, rigid positioning of the large mirror was crucial. Changes in linac gantry rotation could cause apparent changes in patient positioning due to physical sag of the camera and mirror mounting system. Furthermore, image quality was suboptimal due to the large lenses required to focus the light signal to the video camera. Degradation of spatial resolution, field uniformity, signal-to-noise (SNR), and field flatness all contributed to minimizing the utility of this detector type.

LIQUID IONIZATION CHAMBERS

The liquid ionization chamber (LIC) is based on a design proposed by Wickman (12), who proposed to use liquid as an ionization medium to increase the efficiency of ionization chambers. Indeed, the introduction of isooctane increased the signal level by over a factor of 10, but also deleteriously increased the recombination of the electrons due to their low mobility. A first prototype was built by Meertens et al. (13) using two printed circuit boards with perpendicular electrode strips. This resulted in a 30×30 matrix of ionization chambers, and was further refined by van Herk et al. (14) to include 128×128 and finally 256×256 matrices.

To obtain an image, the matrix is scanned row by row, by successively switching high voltage to different row electrodes and measuring all column electrodes. The ionization chamber polarizing voltage is typically 300 V, which is comparable to the voltage applied over a regular megavoltage ionization chamber. The typical current produced by the chamber is of the order of 100 pA. Due to the high voltage switching there is a limit on the speed with which the image may be obtained.

In most of the commercially available imagers, $\sim$1 s is required to readout the complete matrix. Figure 1 shows a schematic diagram of a EPID. The LIC is efficient in that it is able to obtain information constantly in between readouts. The low recombination rate ($\alpha \simeq 4.510^{-16}$ m^3/s) of the ions in the liquid makes that the signal accumulates during radiation and provides an averaging effect.

Multiple Detector Combinations

An alternative way to obtain two-dimensional (2D) transmission images is to use a line detector much like the ones found in computer–tomography devices. They consist of a

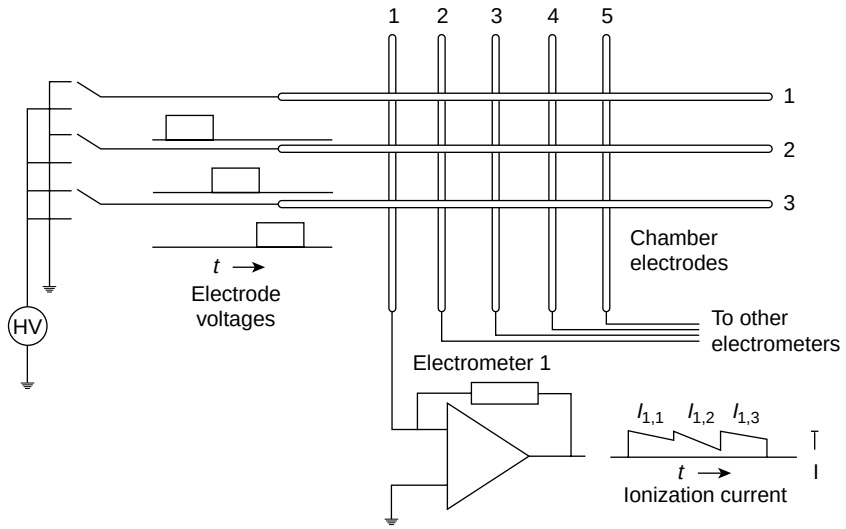


Figure 1. Implanted gold seeds, imaged using a flat-panel portal imager.

line of point detectors, which usually contain a phosphorescent material and an optical light detector (15). Alternatively, Lam et al. (16) constructed a device containing 256 silicon diodes. The line of detectors is scanned through the field in a mechanical fashion. However, this approach is time intensive and not appropriate for clinical techniques such as respiratory gated radiotherapy for treatment of lung cancer where the exiting photon energy fluence is of a dynamic nature.

Flat Panel Technology

The advance of flat-panel displays, where the use of amorphous silicon created surfaces that locally behaved as a crystalline material, allowed for lithography of integrated circuits. The same thin-film technology (TFT) was used to generate photodiode circuits detecting optical light. The TFT is deposited on a glass substrate of ~1 mm thick as is shown in Fig. 2. One of the major advantages of these circuits is that they are highly radiation resistant and can be placed directly in a radiation beam. As with computer integrated circuit chip technology, the TFT EPID can be etched with a resolution of a few micrometers, permitting construction of a large detector matrix. As the photodiodes only detect visible light, a phosphorescent screen is used to perform the conversion much as for camera-based EPIDs. The TFT EPID is nonconducting during the radiation. To read out the TFT, a voltage bias is applied to allow collected charge to flow between the photodiode and an external amplifier. An amplifier records this charge, which is proportional to the light intensity. The TFT EPID array has a maximum readout rate of 25 Hz. In comparison to the camera-based EPID system, the large TFT detectors are designed to be in direct contact with the conversion screen, thus eliminating the poor optical coupling and efficiency intrinsic to the camera-based systems.

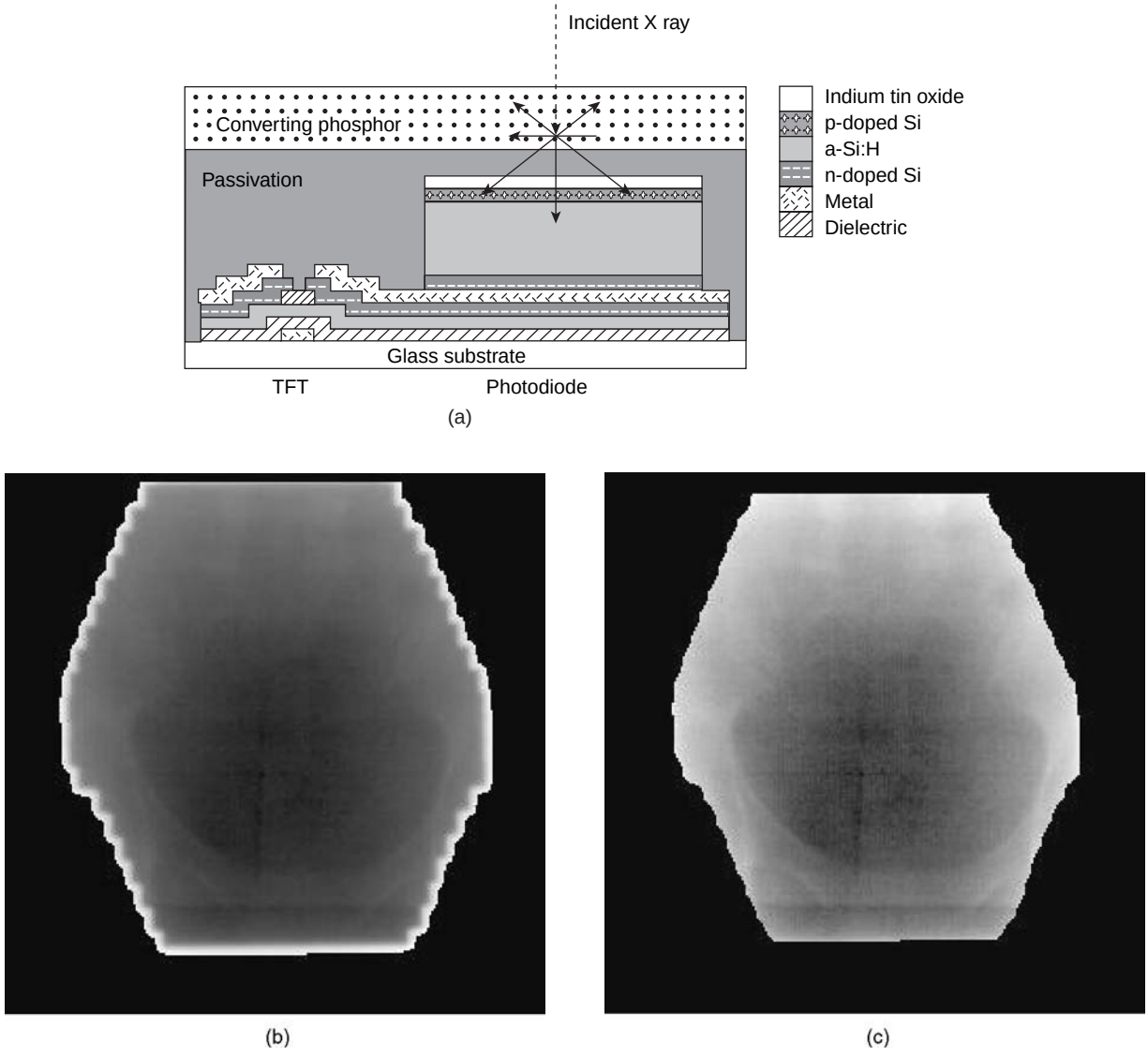


Figure 2. Schematic cross-section (not to scale) of a single a-Si:H imaging pixel.

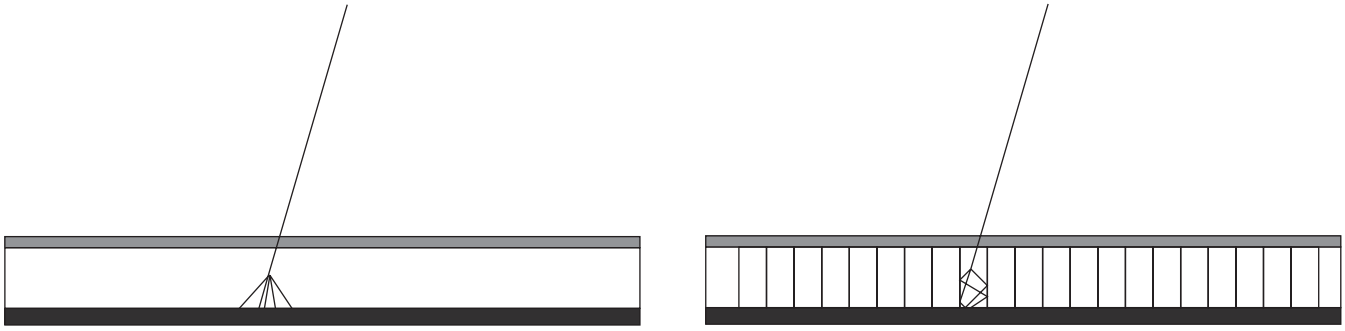


Figure 3. Illustration of EPID conversion screens used to trap optical light. Fig. 3a shows a regular Gd_2O_2S-screen with optical spread, while Fig. 3b shows a CsI screen which limits optical spread and increases detector resolution.

The resolution and efficiency of flat-panel imagers are theoretically superior to those from camera-based and LIC EPIDs. Research performed by Munro et al. (17) indicates that the amorphous silicon imager is X-ray quantum limited, and that the resolution is limited by the spread of the optical photons in the imager. The increase in quality and the compact size of TFT EPIDs means that they may be easily installed onto a linac using a robotic arm. Consequently, TFT EPIDs are now the only type of detectors commercially available. Efforts are underway to replace the current conversion screens, which usually are Gd_2O_2S phosphors, with CsI, which can be grown as single crystals the size of a pixel. The optical light is then trapped in a manner similar to an optical fiber, and therefore optical spread is eliminated, as shown in Fig. 3.

Figure 4 shows pelvic images taken with a camera-based detector, LIC, and flat-panel imager.

EPID APPLICATIONS

Replacement of Radiographic Film

Just as is common in radiology departments, electronic acquisition and management of imaging data is quickly becoming standard practice. Because of the fast pace of detector evolution in the past decade, EPID technology is facilitating hospital-wide imaging digitization. However, to understand why widespread implementation of EPID systems has not yet occurred, it is important to perform a brief cost analysis.

Compared to radiographic film-based portal imaging, up-front capital expenditures for EPID systems are about a factor of 5 larger (e.g., $25,000 vs. 125,000). However, on-going costs associated with an EPID system as compared to a radiographic film-based portal imaging program are much less. For example, regular purchasing of film and maintenance of a film processor (including silver harvesting) is not required with an EPID program. This cost-analysis makes EPID highly competitive given the digital direction facing modern health care.

With direct image digitization comes the ability to transmit data as required for telemedicine. Furthermore, imaging data storage and retrieval, as required for radiology picture archiving and communication systems (PACS), has additional advantages over conventional radiographic film storage. In radiation oncology departments, it is now commonplace for record-and-verify systems to be coupled to an electronic patient charting system. By storing the EPID images in this domain, many of the concerns for patient record keeping, retrieval, and preservation of treatment confidentiality are overcome.

Improvement of Patient Positioning

The original purpose of portal imaging is to reduce the incidence and extent of errors made during radiation treatment. The type of errors can be categorized as gross errors and stochastic errors. Examples of such errors are shown in Fig. 5. The gross errors occur only once and when detected can be removed from the treatment after one fraction. Stochastic errors contain a random component, which implies that the error changes from day to day. Errors and QA-problems can also introduce a systematic component hidden by the random component, which can only be corrected for if it's extent is known.

To reduce stochastic errors, there are two general methodologies, on- or off-line corrections. The most straightforward methodology uses on-line correction where an image is obtained in "localization mode" where minimal dose to the patient is applied. If a discrepancy is observed in the patient setup, and if this discrepancy is larger than a predetermined threshold or action level, efforts are taken to eliminate the error by changing the patient position or changing the treatment configuration. The aforementioned threshold is based on the precision to which position can be determined *and* with which the patient setup correction can be applied. Given the digital nature of EPID images, it is possible to increase the accuracy using computerized algorithms to objectively measure patient positioning with respect to the treatment field (18–21). This approach has not been widely adopted, mainly due to the perceived labor intensity and some medico-legal aspects. However, this may change with the advent of other on-line repositioning techniques (cf. ultrasound-based repositioning) and increased process automation.

Weekly port-filming is the standard procedure in the QA of external beam radiation therapy, which generally

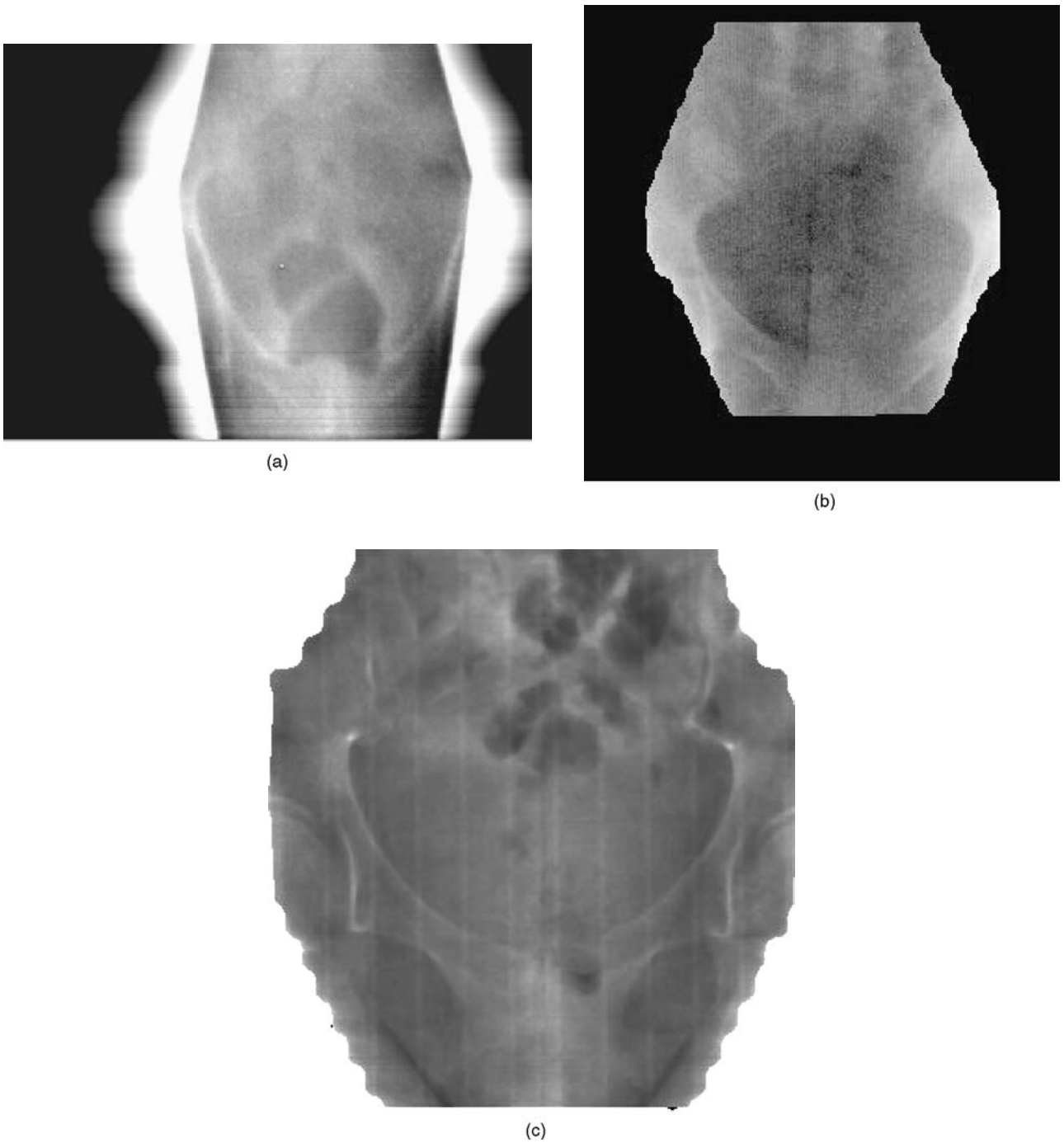


Figure 4. A comparison of pelvic images taken with three different types of EPIDs: 4a) a camera-based system, 4b) a liquid ionization chamber system, and 4c) a amorphous silicon (a-Si:H) flat panel imaging system.

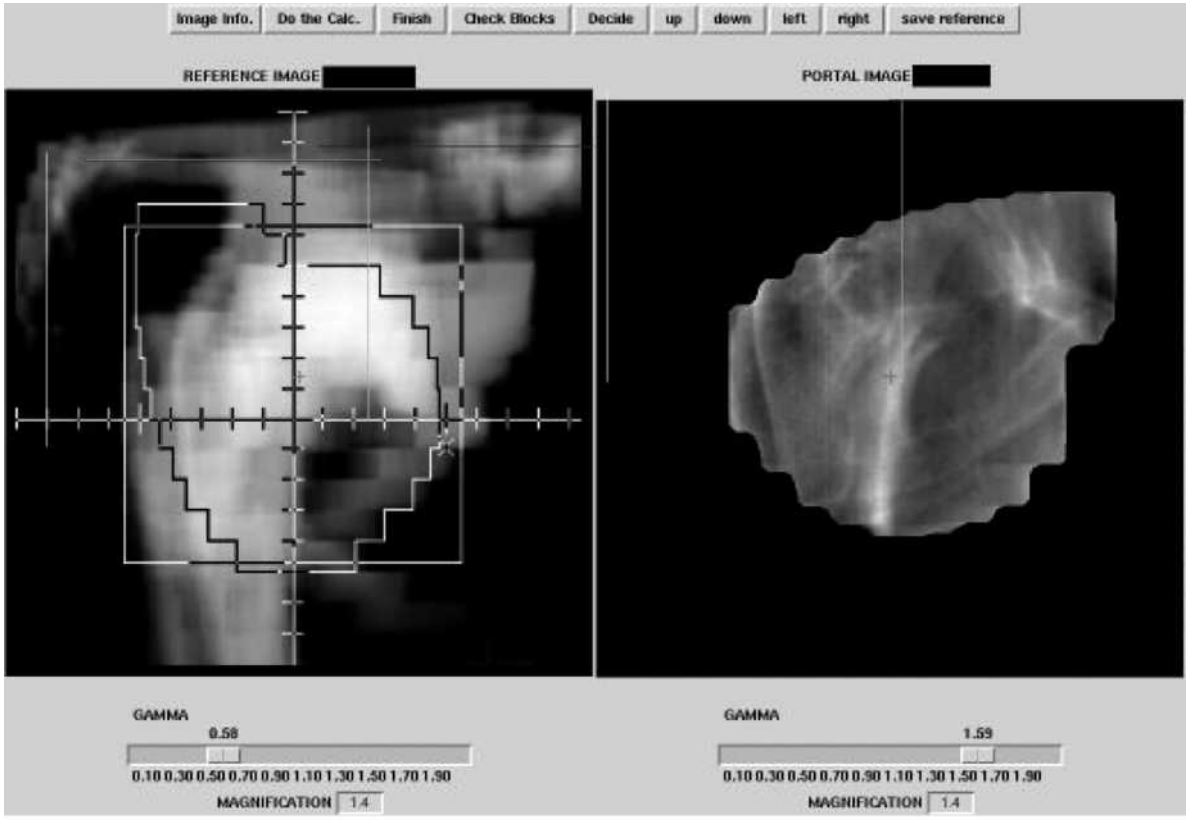
implies that a sample of 4 positions is taken out of a 20 fraction treatment. Errors are corrected after the first image. An interesting study by Valicente et al. (2), showed that this practice is suboptimal. The use of EPIDs allows us to obtain images in a more economical way. Most off-line correction strategies assume that the distribution formed by all consecutive errors is a normal distribution characterized by the mean error and the standard deviation calculated as in

Mean:

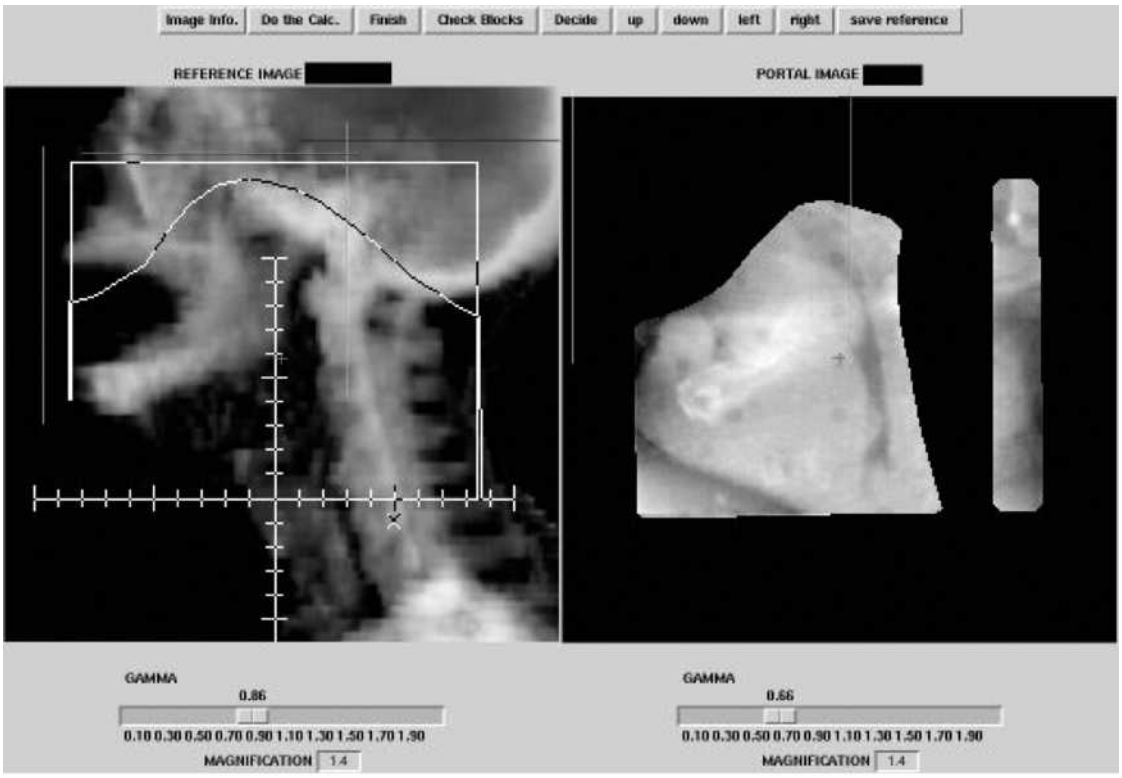
$$\langle x \rangle = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

Standard Deviation:

$$\sigma = \sqrt{\frac{1}{(N-1)}\sum_{i=1}^{N}(\langle x \rangle - x_i)^2}$$



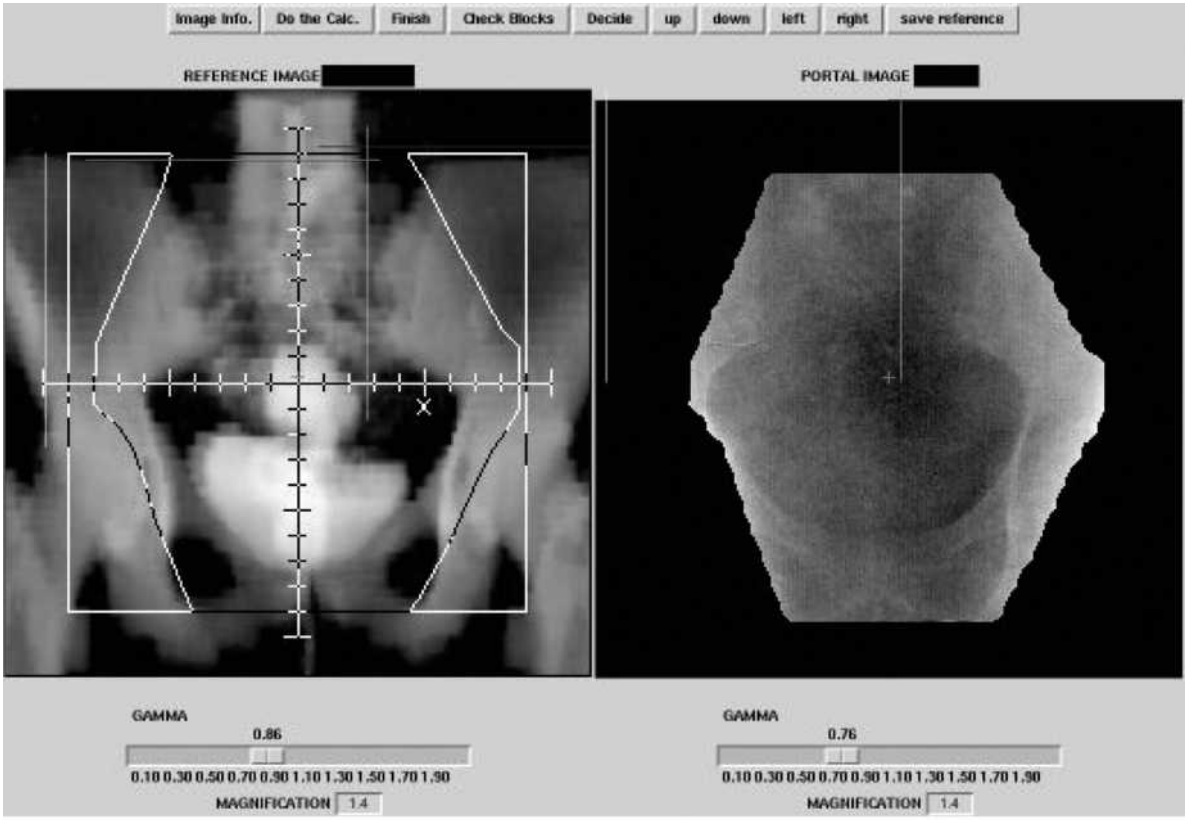
(a)



(b)

Figure 5. Examples of setup errors. Reference images on the right are digitally reconstructed radiographs with the correct setup. On the left are the measured portal images. (a) Faulty collimator angle. (b) Wrong blocking used. (c) Wrong MLC file. (d) Patient positioning error.

(c)



(d)

Figure 5. (*Continued*)

By repeated sampling of the distribution (e.g., taking port films), the strategy estimates the value of $\langle x \rangle$ as close as possible and corrects for the error, which will reduce the systematic error in the treatment. Several groups have studied the implications of this strategy and its variations. The most successful approach seems to be the following strategy: The set-up variations are recorded and averaged. This is compared to an action level that depends on how many samples have been taken already (the level shrinks as the the amount of information on the systematic error increases). These studies showed that systematic errors could be reduced to $\sim$2 mm (22).

Organ Motion

One of the major reasons for use of EPIDs is the fact that the patients anatomy and position vary from those used for treatment planning purposes. The factors involving this variation are

- Patient movement.
- Patient positioning inaccuracies.
- Organ motion.

Any of these factors will influence the actual dose distribution to be different from that obtained using treatment planning. It is straightforward to correct for the first two problems using portal imaging as the patients position is typically well-characterized using bony anatomy. An excellent compilation on the incidence, extent, and repercussions of organ movement was performed by Langen and Jones (23). Efforts to incorporate organ movement during radiotherapy treatment planning involves enlarging the target to be treated. Sophisticated algorithms that calculate the extent of these enlargements were developed independently by Stroom et al. (24) and by van Herk et al. (25). The general framework for this enlargement is given in ICRU 50 and ICRU 60 (26,27). In these reports the gross target volume (GTV) is defined as the volume containing demonstrated tumor. A margin is added to the GTV to account for suspected microscopic tumor involvement, and is defined as the clinical target volume (CTV). Finally, the planning target volume (PTV) is defined by the CTV and an additional margin to allow for geometrical variations such as patient movement, positioning errors, and organ motion. The margins added to GTV and CTV can substantially increase the PTV since the margins are applied in three dimensions. Because of volume effects of radiation therapy, there is a tendency to minimize the PTV by increasing the precision of the treatment. As explained above, EPIDs are able to minimize uncertainties caused by patient motion and positioning errors using off- or on-line correction strategies.

Except in a few cases like in lung or where air pockets are present (24,28), the target is virtually indiscernible using X rays. To solve this problem, other modalities, like CT (29) and ultrasound (30,31), have been used to determine the position of the organ. The EPIDs can also be used to image the position of organs if radioopaque markers are implanted. The markers need to be of high density and migration needs to be limited. The efficacy and feasibility of using markers with EPIDs was studied by Balter et al. (32), application of the use of markers have been extensively studied by Pouliot and co-workers (33). The use of markers is becoming more popular and EPID systems are being augmented with software to detect markers as well as perform the requisite positional calculations.

Quality Assurance and *In Vivo* Dosimetry

In 2001, Task Group No. 58 of the American Association of Physicists in Medicine Radiation Therapy Committee issued a protocol to define the standard-of-care for performing EPID QA on a daily, monthly, and annual basis (34). In this protocol, a quality assurance program is proposed where daily checks of EPID system performance, image quality, and safety interlocks are performed by a radiation therapy technologist. In addition to reviewing results and independently performing checks conducted by the technologist, a medical physicist should conduct the following checks on a monthly basis: perform constancy check of SNR, resolution, and localization; inspect images for artifacts; do a mechanical inspection of all EPID components; and maintain the computer system. The annual QA tasks are also performed by the medical physicist, and include all of the above tasks plus a full check of the EPID geometric localization accuracy. By performing these QA tasks, the radiotherapy department may be reasonably assured of a reliable EPID system for clinical use.

The QA tests typically utilize a vendor-supplied phantom designed to facilitate evaluation of the aforementioned tasks. Since clinical linacs are typically dual energy (e.g., 6 and 15 MV) in design, tests are applied to both photon energies. As expected, the lower photon energy will demonstrate improved image quality (e.g., SNR and spatial resolution). Since many EPID systems utilize sophisticated computer software utilities, testing of this software in a realistic setting is an integral component of the EPID quality management program. As can be expected with tests that are subjective in nature, it is recommended that multiple users be employed to evaluate the subjective criteria so as to minimize user bias.

In addition to the aforementioned advantages, an EPID system permits unique opportunities of radiotherapy treatment QA. Treatment fields are typically blocked with beam modifiers to account for irregularities in patient shape. These beam modifiers include compensating materials when minor changes are required, or beam wedges when gross changes are required. Because modern linacs have features like dynamic wedges and dynamic multileaf collimators, the nonintegral approach that EPIDs offer over radiographic film (i.e., the ability to obtain several images at different stages during a dynamic process) permits continued high-quality QA. Due to the electronic nature of EPID measurement of the photon energy fluence exiting a patient, one can perform exit dosimetry and quantitative comparisons with treatment planning intentions (35). However, these efforts are currently research driven, and widespread clinical implementation may not be expected for a few years.

BIBLIOGRAPHY

1. AAPM report No. 24: Radiotherapy portal imaging quality, 1987.
2. Valicenti RK, Michalski JM, Bosch WR, Gerber R, Graham MV, Cheng A, Purdy JA, Perez CA. Is weekly port filming adequate for verifying patient postion in modern radiation therapy? Int J Rad Oncol Biol Phys 1994;30(2):431–438.
3. Moseley J, Munro P. Display equalization: A new display method for portal images. Med Phys 1993;20(1):99–102.
4. Van den Heuvel F, Han I, Chungbin S, Strowbridge A, Tekyi-Mensa S, Ragan D. Development and clinical implementation of an enhanced display algorithm for use in networked electronic portal imaging. Int J Rad Oncol Biol Phys 1999;45:1041–1053.
5. Jaffray DA, Batista JJ, Fenster A, Munro P. X-ray scatter in megavoltage transmission radiography: Physical characteristics and influence on image quality. Med Phys 1994;21(1):45–60.
6. Bailey NA, Horn RA, Kamp TD. Fluoroscopic visualization of megavoltage therapeutic x-ray beams. Int J Radiat Oncol Biol Phys 1980;6:935–939.
7. Leszczynski KW, Shalev S, Cosby S. A digital video system for on-line portal verification. Medical Imaging IV: Image Formation. SPIE 1990;1231:401–405.
8. Leong J. Use of digital fluorscopy as an on-line verification device in radiation therapy. Phys Med Biol 1986;31:985–992.
9. Munro P, Rawlinson JA, Fenster A. A digital fluoroscopic imaging device for radiotherapy localization. Int J Rad Oncol Biol Phys 1990;18:641–649.
10. Visser AG, Huizenga H, Althof VGM, Swanenburg BN. Performance of a prototype fluoroscopic radiotherapy imaging system. Int J Rad Oncol Biol Phys 1990;18:43–50.
11. Wong JW, Slessinger ED, Hermes RE, Offutt CJ, Roy T, Vannier MW. Portal dose images I: Quantitative treatment plan verification. Int J Rad Oncol Biol Phys 1990;18:1455–1463.
12. Wickman GA. A liquid filled ionisation chamber with high spatial reslution. Phys Med Biol 1974;19:66–72.
13. Meertens H, van Herk M, Weeda J. A liquid ionisation detector for digitital radiography of therapeutic megavoltage photon beams. Phys Med Biol 1985;30:313–321.
14. Van Herk M, Meertens H. A matrix ionization chamber imaging device for on-line patient set-up verification during radiotherapy. Radiother Oncol 11:369–378.
15. Morton EJ, Swindell W, Evans PM. A linear array, scintillation crystal-photodiode detector for megavoltage imaging. Med Phys 1991;18:681–691.
16. Lam KS, Partowmah M, Lam WC. An on-line electronic portal imaging system for external beam radiotherapy. Br J Radiol 1986;59:1007–1013.
17. Munro P, Bouius DC. X-ray quantum limited portal imaging using amorphous silicon flat-panel arrays. Med Phys 1998;25(5):689–702.
18. De Neve W, Van den Heuvel F, De Beukeleer M, Coghe M, Verellen D, Thon L, De Roover P, Roelstraete A, Storme G. Interactive use of on-line portal imaging in pelvic radiation. Int J Rad Oncol Biol Phys 1993;25:517–524.
19. Van den Heuvel F, De Neve W, Verellen D, Coghe M, Coen V, Storme G. Clinical implementation of an objective computer-aided protocol for intervention in intra-treatment correction using electronic portal imaging. Radiother Oncol 1995;35:232–239.
20. Balter JM, Pelizarri CA, Chen GTY. Correlation of projection radiographs in radiation therapy using open curve segments and points. Med Phys Mar.–Apr. 1992;19(2): 329–334.
21. Bijhold J, Lebesque JV, Hart AAM, Vijlbrief RE. Maximizing setup accuracy using portal images as applied to a conformal boost technique for prostate cancer. Radiother Oncol 1992;24:261–271.
22. Bel A, Vos PH, Rodrigus PTR, Creutzberg CL, Visser AG, Stroom JC, Lebesque JV. High-precision prostate cancer irradiation by clinical application of an offine patient setup verification procedure, using portal imaging. Int J Rad Oncol Biol Phys 1996;35(2).
23. Langen KM, Jones DT. Organ motion and its management. Int J Radiat Oncol Biol Phys May 1 2001;50(1):265–278.
24. Stroom JC, Boer de HC, Huizenga H, Visser AG. Inclusion of geometrical uncertainties in radiotherapy treatment planning by means of coverage probability. Int J Radiat Oncol Biol Phys Mar 1 1999;43(4):905–919.
25. van Herk M, Remeijer P, Lebesque JV. Inclusion of geometric uncertainties in treatment plan evaluation. Int J Radiat Oncol Biol Phys Apr 1 2002;52(5):1407–1422.
26. ICRU and International Commission on Radiation Units and Measurements. Prescribing, recording and reporting photon beam therapy. ICRU Report 1993; 50.
27. ICRU and International Commission on Radiation Units and Measurements. Prescribing, recording and reporting photon beam therapy, Supplement to ICRU 50. ICRU Report 1999; 62.
28. Erridge SC, Seppenwoolde Y, Muller SH, van Herk M, De Jaeger K, Belderbos JSA, Boersma LJ, Lebesque JV. Portal imaging to assess set-up errors, tumor motion and tumor shrinkage during conformal radiotherapy of non-small cell lung cancer. Radiother Oncol 2003;66(1):75–85.
29. Jaffray DA, Drake DG, Moreau M, Martinez AA, Wong JW. Radiographic and tomographic imaging system integrated into a medical linear accelerator for localization of bone and soft-tissue targets. Int J Radiat Oncol Biol Phys Oct 1 1999;45(3):773–789.
30. Lattanzi J, McNeeley S, Hanlon A, Schultheiss TE, Hanks GE. Ultrasound-based stereotactic guidance of precision conformal external beam radiation therapy in clinically localized prostate cancer. Urology Jan 2000;55(1): 73–78.
31. Serago CF, Chungbin SJ, Buskirk SJ, Ezzell GA, Collie AC, Vora SA. Initial experience with ultrasound localization for positioning prostate cancer patients for external beam radiotherapy. Int J Radiat Oncol Biol Phys Aug 1 2002;53(5):1130–1138.
32. Balter JM, Sandler HM, Lam K, Bree RL, Lichter AS, ten Haken RK. Measurement of prostate movement over the course of routine radiotherapy using implanted markers. Int J Rad Oncol Biol Phys 1995;31(1):113–118.
33. Vigneault E, Pouliot J, Laverdiere J, Roy J, Dorion M. Electronic portal imaging device detection of radioopaque markers for the evaluation of prostate position during megavoltage irradiation: A clinical study. Int J Radiat Oncol Biol Phys Jan 1 1997;37(1): 205–212.
34. Herman MG, Balter JM, Jaffray DA, McGee KP, Munro P, Shalev S, van Herk M, Wong JW. Clinical use of electronic portal imaging: report of radiation therapy committee task group 58. Med Phys May 2001;28(5):712–737.
35. Boellaard R, Van Herk M, Mijnheer BJ. A convolution model to convert transmission dose images to exit dose distributions. Med Phys 1997;24(2):189–199.

See also COMPUTED TOMOGRAPHY; MAGNETIC RESONANCE IMAGING; PHOTOGRAPHY, MEDICAL; POSITRON EMISSION TOMOGRAPHY; ULTRASONIC IMAGING.

IMMUNOLOGICALLY SENSITIVE FIELD–EFFECT TRANSISTORS

EMMANUEL S. ZACHARIAH
University of New Jersey
New Brunswick, New Jersey

P. GOPALAKRISHNAKONE
National University of Singapore
Singapore

PAVEL NEUZIL
Institute of Bioengineering
and Nanotechnology
Singapore

INTRODUCTION

Diagnostics as a whole represent a large, well-established, and continually expanding market. Methods for the selective determination of analytes in biological fluids, such as blood and urine, are important. When a foreign substance (antigen) invades the human body, the immune system produces antibodies that interact with the antigen. Such a recognizing process involves the formation of an immunocomplex based on interactions between the immunospecies. The recognition is specific for the antibody-antigen system and, thus, for the measurement of antigen concentration (1,2). This determination is of importance for diagnosis because the antigens can be viruses, bacteria that are involved in many human illnesses, such as cancer and AIDS. The analytes detected and measured have also included many other medical diagnostic molecules such as hormones, clinical disease biomarkers, drugs, and environment pollutants such as pesticides. Antibody diversity is so great that virtually any biomolecule can be recognized (3). The range of analyte concentrations encountered is extremely large, from greater than $10^{-3}\,M$ for species such as glucose and cholesterol and to less than $10^{-12}\,M$ for certain drugs and hormones (4). It is for the detection of these low level analytes that the application of immunological techniques is essential.

An immunoassay is a multistep diagnostic test based on the recognition and binding of the analyte by the antibody. Most immunoassay techniques are based on the separation of free and bound immunospecies (5). In these techniques, one of the immunoagents (antibody or antigen) is immobilized on a solid phase. The solid phase facilitates the separation and washing steps required to differentiate bound and free fractions of the label. Quantification of a bound immunoagent is conducted by using labels covalently bound to the immunoagent with specific properties suitable for detection. The most common labels are radioactive markers, enzymes, and fluorescent labels. For many of the nonisotopic labels, the reagents have been designed such that binding of labeled antigen to antibody in some way modulates the activity of the label, resulting in a homogenous immunoassay without the need for a separation step. The most familiar type of enzyme immunoassay in clinical analysis is known as enzyme-linked immunosorbent assay (ELISA) (6). Different schemes of enzyme immunoassay exist, and, in clinical laboratory practice, the most popular are the "Sandwich" method for large analytes, and competitive binding immunoassay methods for the determination "haptens" (low molecular weight analytes).

The advent of biosensor technology, with the possibility of direct monitoring of immunoreactions, provides opportunity to gain new insight into antigen-antibody reaction kinetics and create rapid assay devices with wider applications. A biosensor is composed of (*1*) a biochemical receptor, which uses biosubstances such as enzymes, antibodies, or microbes to detect an analyte, (*2*) a transducer, which transforms changes in physical or chemical value accompanying the reaction into a measurable response, most often in the form of electrical signal (7–9). The term immunosensor is used when antibodies are immobilized to recognize their appropriate antigens (or vice versa) (10). Immunosensors possess several unique features, such as compact size, simplicity of use, one-step reagentless analysis, and absence of radioactivity, which make them attractive alternatives to conventional immunoassay techniques. Immunosensors can be divided, in principle, into two categories: nonlabeled and labeled (11). Nonlabeled or direct-acting immunosensors are designed in a way that the immunocomplex (i.e., the antibody-antigen complex) is directly determined by measuring physical changes induced by the formation of the complex. In contrast, labeled or indirect-sensing immunosensors have incorporated a sensitively detectable label. The immunocomplex is thus entirely determined through measurement of the label. In order to determine an antigen, the corresponding antibody is immobilized on the membrane matrix, which is held on an amperometric-or potentiometric-sensing transducer used to measure the rate of the enzymatic reaction (12–14).

Of the electrochemical technologies for biosensors, the Ion-Sensitive Field Effect Transistor (ISFET) has been the center of special attention as a transducer. ISFETs were introduced by Bergveld in 1970 (15), and were the first type of this class of sensor in which a chemically sensitive layer was integrated with solid-state electronics. A field effect transistor (FET) can be considered as a charge-sensitive device (i.e., any change in the excess interfacial charge at the outer insulator surface will be mirrored by an equal and opposite charge change in the inversion layer of the FET). By excluding the gate metal in a FET and using a pH-sensitive gate insulator, a pH-sensitive FET was constructed (16,17). After the invention of the ISFET, many different types of FET-based sensors have been presented. The application of enzymes as the selecting agent in ISFET-based sensing systems leads to the development of highly sensitive sensors. Such enzyme-modified ISFETs (EnFETs) can, in principle, be constructed with any enzyme that produces a change in pH on conversion of the concerning substrate (18). By combining the ISFET with a membrane that contains a biological substance, like an antibody, the sensor can detect a specific antigen (19). The ISFET immunosensors or Immunologically sensitive FETs (IMFETs) have several advantages over the conventional enzyme immunoassay. The ISFET could be mass-produced by an integrated circuits (IC) process, which makes it very small and economical. An electric circuit can be integrated on the same chip. The biosensor platform finds many applications in various

fields, such as medical diagnostics, fermentation process control, and environmental monitoring.

THEORY

In order to understand the operation of the IMFET, one must trace its origins back to the ISFET or ChemFET (Fig.1). The latter devices have been described in depth elsewhere (20–22). A packaged ISFET is shown in Fig. 2 (23). ISFETs and ChemFETs have, in turn, evolved from the Metal Oxide Semiconductor Field Effect Transistor (MOSFET), currently the most popular active device in the entire semiconductor industry. It is a unipolar device, where the current is given by the flow of majority carriers, either holes in PMOS type or electrons in NMOS type. The operation of the MOSFET can be considered as a resistor controlled by the status of a gate region, so-called MIS structure. It is a sandwich consisting of a stacked-gate

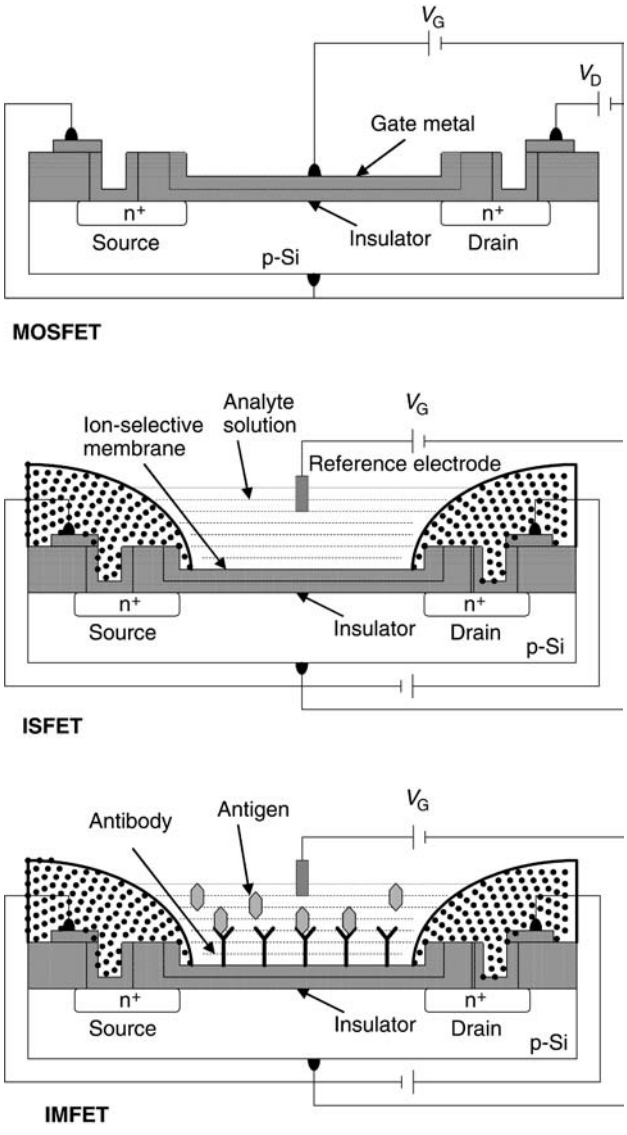


Figure 1. The hierarchy of field effect transistor. a. MOSFET. b. ISFET. c. IMFET.

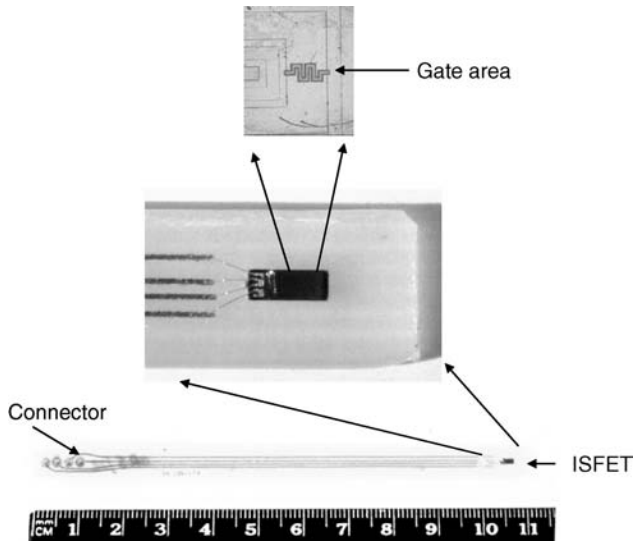


Figure 2. Photomicrograph of an ISFET device packaged on PCB (23).

metal layer, an insulator (typically silicon oxide), and a semiconductor. Assume a low-level doped p-type (NMOS device). Three different states of charge distribution can occur, depending on the voltage V_g, applied between the metal and a semiconductor. A negative value of V_g causes positive holes to accumulate at the semiconductor-insulator interface. A positive value of V_g of a low magnitude leads to the "depletion" condition in which mobile holes are driven away from the interface, resulting in a negative charge of low density due to the presence of immobile acceptor atoms. Finally, if the V_g exceeds a certain threshold voltage (V_{th}), electrons accumulate at the semiconductor-insulator interface at a density greater that the hole density, a situation opposite to that normally found with *p*-type semiconductors. This depletion of mobile charge carriers followed by surface inversion is known as the "field effect." It forms an electrically conductive channel between two other terminals, a source and a drain (see Figure 1a). The drain current I_d through the transistor is a functions of drain and gate voltage. Without surface inversion (i.e., $V_g < V_{th}$,) the drain current is negligible, because the drain-to-substrate PN junction is reverse biased.

The MOSFET and its descendants are charge-controlled devices. In analytical applications (e.g., ISFETs, ChemFETs, and IMFETs), the change in charge density is brought about by adsorption of one or more species present in the solution onto the FET structure. In the ISFET, the gate metal is replaced with a conventional reference electrode (Ag/AgCl or Hg/Hg_2Cl_2), a solution containing an ionic species of interest, and an electroactive material (membrane) capable of selective ion exchange with the analyte (Fig. 1b), which is an example of a nonpolarizable interface, that is, reversible charge transfer occurs between the solution and the membrane. The analyte generates a Nernst potential at the membrane-solution interface, which then modulates the drain current analogous to the manner in which changing the externally applied voltage does for the MOSFET.

Direct-Acting (Label-Free) IMFET

The structure of the direct-acting IMFET is similar to that of the ISFET, except that the solution-membrane interface is polarized rather than unpolarized. If the solution-membrane interface of the ISFET is ideally polarized (i.e., charge cannot cross the interface), then the ISFET can measure the adsorption of charged species at the interface as shown below. As antibodies, antigens, and proteins are generally electrically charged molecules, the polarized ISFET could be used to monitor their nonspecific adsorption at the solution-membrane interface. To render the polarized ISFET selective for a given antigen and thus create the so-called IMFET, the specific antibody for that antigen has to be immobilized on the surface of the ISFET (see Fig. 1c). The adsorption of this antigen would then be specifically enhanced over other molecules in the solution and the signal measured by the ISFET would be mostly due to the adsorption of that particular antigen. The ISFET interacts with the analyte through an ion-exchange mechanism, whereas the IMFET interaction is based on the antigen-antibody reaction.

This design for the measurement of the adsorption of charged molecules is practicable only if charge cannot cross the interface, which, thus, acts as an ideal capacitor. As will be seen, failure to achieve a perfectly polarizable interface has a detrimental effect on the specificity of the IMFET. Few reports exist on direct-acting IMFETs; a brief analysis on the work of Janata research group will be presented here (24–26). The capacitance of a polarized interface is described by electrical double-layer theory and is usually modeled as a series combination of two capacitors, C_G and C_H, where C_G is the capacitance of the diffuse Gouy–Chapman part of the double layer and C_H is the capacitance of the Helmholtz part of the double layer (27). The total capacitance, C_{dl}, is therefore

$$1/C_{dl} = 1/C_G + 1/C_H \tag{1}$$

The electrical circuit through the gate of an ISFET with an ideally polarized interface can be modeled, therefore, as a series combination of C_G, C_H, and C_0, as drawn in Fig. 3, where C_0 is the capacitance of the insulator. A gate voltage V_G is applied through a reference electrode between the solution and the semiconductor. The process of adsorption of charged molecules can be modeled as the transfer of a quantity of charge from the solution to the surface of the transistor as would occur if the switch were closed for a short time period allowing the current source to transfer the charge. As adsorption occurs, the charge on each plate of the capacitors will change to accommodate the new charge balance. The charge change on capacitor C_0 is the quantity of interest as it represents the charge in the inversion layer of the FET, Q_i, and will affect the drain current of the transistor, which can be directly measured. If a quantity of charge, Q_{ads}, is transferred by the adsorption of charged molecules, then the charge change on C_0, Q_i, can be represented by

$$Q_i = Q_{ads}\{C_0/(C_0 + C_{dl})\} \tag{2}$$

Hence, only a fraction of the adsorbed charge will be mirrored in the transistor. When adsorption occurs, because electroneutrality must be observed in the system, an equal quantity of the opposite charge must either enter the inversion layer of the FET or enter the double layer from the solution. Equation 2 predicts that part of the image charge will come from the solution as ions entering the double layer with the adsorbing molecules. This fraction of charge, which is mirrored in the inversion layer of the FET, will be defined as β, and it is defined as

$$\beta = Q_i/Q_{ads} = C_0/(C_0 + C_{dl}) \tag{3}$$

According to this model, only 0.3% of the charge on the adsorbing molecules will be mirrored in the inversion layer of the FET. The authors conservatively estimated β to be 10^{-4}. Considering the I_d current as a function of the potential at the solution-membrane interface, it is clear that a relationship between the adsorbed charge and interfacial potential, $\Phi_{Sol\text{-}mem}$, is necessary to describe the chemical response of the IMFET. This potential is merely the charge change induced in the inversion layer divided by the insulator capacitance:

$$\Phi_{Sol-mem} = Q_i/C_0 = \beta Q_{ads}/C_0 \tag{4}$$

Substitution of this expression in to Equations 5 and 6 yields the response equations for the polarized ISFET. The authors derived the following expressions for the polarized ChemFET relating to Q_i to the observed parameter, the drain current (I_d):

$$I_d = \frac{\mu n W C_0}{L}\left(V_g - V_t - E_r - \phi_{sol-mem} - \frac{V_d}{2}\right)V_d \qquad V_d < V_{dsat} \tag{5}$$

and

$$I_d = \frac{\mu n W C_0}{2L}(V_g - V_t - E_r - \phi_{sol-mem})^2 \quad V_d > V_{dsat} \tag{6}$$

where W is the width of the source-drain conducting channel, μ_n is the effective electron mobility in the channel, C_0 isthe capacitance per unit area of the gate insulator, L is the channel length, V_d is the drain-to-source voltage, V_g is the applied gate voltage, V_t is the threshold voltage (for surface inversion), and E_r is the potential of the reference electrode.

The antibody-antigen binding reaction in its simplest form can be expressed in terms of the following

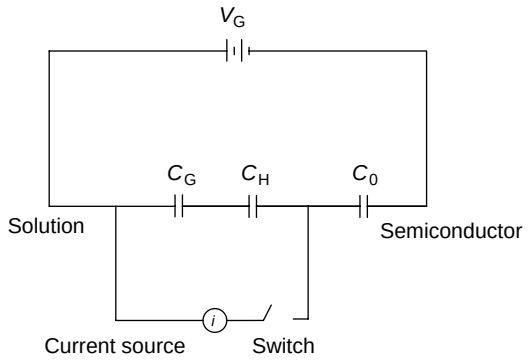


Figure 3. Electrical model for the measurement of charge adsorption with the CHEMFET.

biomolecular reaction:

$$\mathrm{Ab} + \mathrm{Ag} \rightleftarrows \mathrm{AbAg}$$

where Ab is the antibody, Ag is the antigen, and AbAg is the complex. The reaction is characterized by the equilibrium constant K,

$$K = [\mathrm{AbAg}]/[\mathrm{Ab}][\mathrm{Ag}] \tag{7}$$

The total charge change at the interface due to the binding, Q_i, can be shown to be

$$Q_\mathrm{i} = \beta Q_\mathrm{ads} = \beta z F \frac{K[\mathrm{Ag}][\mathrm{S}]}{1 + [\mathrm{Ag}]} \tag{8}$$

where z is the ionic charge of the antigen and [S] is the surface concentration of binding sites (the surface concentration of immobilized antibodies before binding). Substitution of this expression into Equation 4 yields

$$\Phi_\mathrm{Sol-mem} = \frac{\beta z F K[\mathrm{Ag}][\mathrm{S}]}{C_0(1 + [\mathrm{Ag}])} \tag{9}$$

From Equation 9, the limit and range of the detection for the IMFET can be predicted. Assume that the equilibrium constant is in typical range from 10^5 to 10^9 (28), which gives a value of $\beta = 10^{-4}$. If the antibodies are immobilized with a surface concentration of 1 molecule per 10 nm^2 and the charge on an antigen is five electronic charges of an antibody, the IMFET's detection limit would be in the range of $10^{-7} - 10^{-11}$M of concentration antibody concentration. The antigen concentration that gives 90% surface coverage can similarly be calculated to be in the range of $10^{-4} - 10^{-8}$ M. Similar equations can be derived for the case where the antigen is immobilized at the interface rather than the antibody. However, it has been argued by many researchers that a static measurement concerning the presence of a protein layer on an electrode is difficult, because the charged groups are, of course, neutralized by surrounding counter ions (29). In order to avoid interference from other charged species present in the solution, the substrate for immobilization should preferably be inert and nonionic (24–30), which in aqueous solutions implies a hydrophobic surface (31). Ideal conditions that are required in this coherence are a truly capacitive interface at which the immunological binding sites can be immobilized, a nearly complete antibody coverage, highly charged antigens, and a low ionic strength.

Schasfoort et al. (32) extensively studied the requirements for the construction of IMFET, which would operate on the direct potentiometric sensing of protein charges. The charge redistribution around immobilized proteins at the insulator-solution interface can be described by the double-layer theory (33). On adsorption, the diffuse layer of counter ions around the protein charges may overlap with the diffuse layer of the electrolyte-insulator interface. The thickness of diffuse-charge layers is described by the Debye theory (34) and defined by the distance where the electrostatic field has dropped to 1/e of its initial value:

$$\kappa^{-1} = \left(\frac{\varepsilon_0 \varepsilon k T}{2q^2 I}\right)^{1/2}$$

where κ^{-1} is the Debye length, q the elementary change, k Boltzmann's constant, T absolute temperature, ε_0 the permittivity of vacuum, ε the dielectric constant, and $I = 1/2 \sum c_i z_i^2$ represents the ionic strength in which c_i is the concentration of ion i with valency z (for 1-1 salt, I can be replaced by c).

It can be seen from the equation that the Debye length is strongly dependent on the ionic strength of the solution; more precisely, the Debye length is inversely proportional to the square root of the ionic strength. Therefore, one can expect that the chance of overlapping of the double layers of the substrate-solution interface and the adsorbed proteins can be substantial only if low electrolyte concentrations are used, owing to the dimensions of the proteins (Fig. 4). In a physiological salt solution, the Debye length is limited to ca. 0.8 nm. It is obvious that only charge density changes that occur within the order of a Debye length of the ISFET surface can be detected. With the macromolecules, such as protein, the dimensions are much larger (about 10 nm) than those of the double layer of the electrolyte-insulator interface, which means that, in such a case, most of the protein charge will be at a distance greater than the Debye length from the surface. If, moreover, on top of a monolayer of antibody molecules a second layer on antigens in coupled, it is obvious that the chance of overlap of the diffuse layers of antigens with electrolyte-substrate interface will decrease even more. At high ionic strength, the additional charges of the antigen are nearly always located far outside the diffuse layer at the ISFET surface and pure electrostatic detection of these antigenic charges, therefore, is impossible. In addition, a theoretical approach is

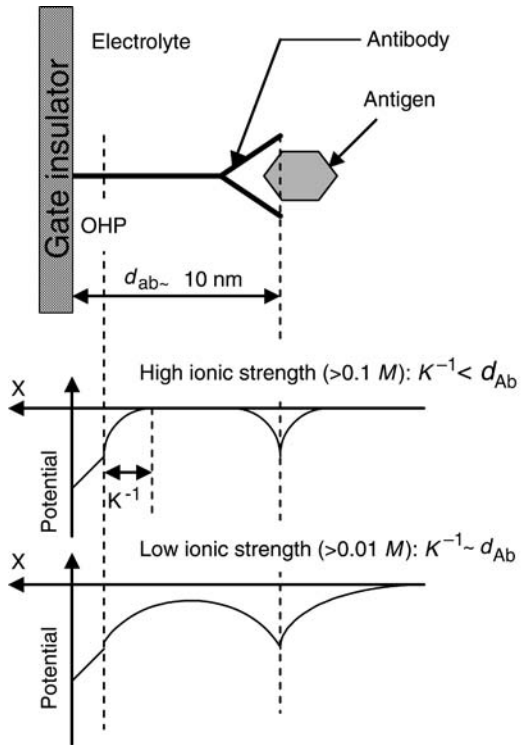


Figure 4. Schematic representation of the potential distribution in a direct-acting IMFET. K^{-1} is the Debye length; d_Ab, dimension of macromolecule (e.g., antibody).

given based on the Donnan equilibrium description, which provides an insight into the potential and ion distribution in the protein layer on the IMFET (32). It is shown that the Donnan potential and the internal pH shift, induced by the protein charges, compensate each other to a greater extent. If the ISFET shows Nernstian behavior, it can be concluded that a direct detection of protein charge is impossible. In order to construct an IMFET, a reference FET or ISFET with a low sensitivity would satisfy the detection of the partially compensated Donnan potential in the presence of an adsorbed protein layer. However, the application of such as IMFET is limited to samples with low ionic strength.

An alternative, indirect approach is proposed by Schasfoort et al. (35,36) for the detection of an immunological reaction taking place in a membrane, which covers the gate area of an ISFET (Figs. 5a and 5b). The protein layer on the gate is exposed to pulse-wise increases in electrolyte concentration. As a result, ions will diffuse into the protein layer and, because of a different mobility of anions and cations, transients in potential will occur at the protein-membrane solution interface. The ISFET, being a voltage-sensitive device, is suitable for the measurement of these transients. As the mobility of ions is a function of the charge density in the protein membrane, changes in the charge density will influence the size and direction of the transients. By exposing the ISFET to a pH gradient and a continuous series of ion concentration pulses, the isolectric point of the protein layer can be detected and, thus, changes as the result of an immunological reaction. When a membrane separates two

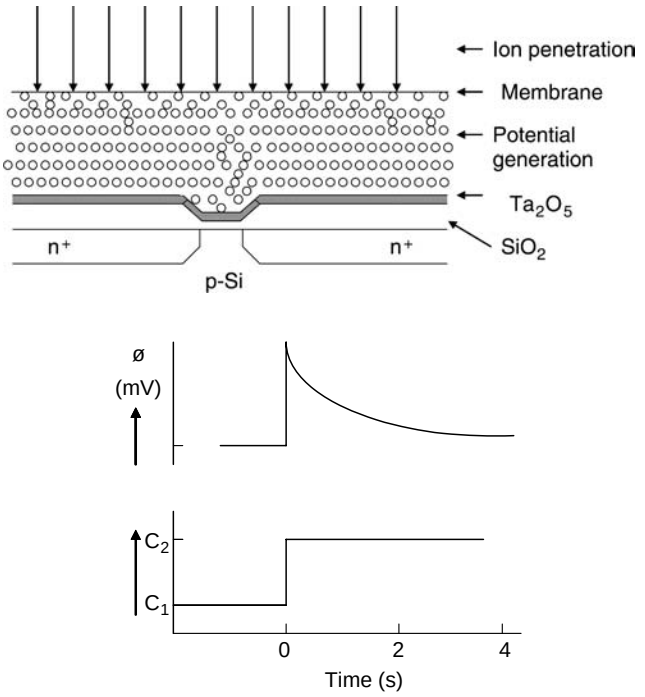


Figure 5. An ion-step arrangement: an ISFET exposed to an increased electrolyte concentration. Transient potential can bemeasured, developing from transient transport of ions across the membrane, which is caused by stepwise changes in electrolyte concentration. The ISFET response ϕ as a result of the stepwise changes in electrolyte concentration (C_1–C_2).

compartments with different electrolyte concentrations, a potential gradient can be measured. The different diffusion rates of anions and cations through the membrane set up a static membrane potential, which can be expressed by the Nernst–Planck equation (33):

$$\phi_m = \frac{RT}{F}.U.In\frac{a_2}{a_1}\ U = \frac{D_+ - D_-}{D_+ + D_-}$$

where ϕ_m = the membrane potential, RT and F have their common meaning, U = the ratio of the diffusion coefficients (D_+ and D_-) of cations and anions, and a_1 and a_2 are the electrolyte activities in the respective compartments. The ion-step method is further developed by Schasfoort and Eijsma (37), and a detailed theoretical understanding of the ion-step response has been presented by Eijiki et al. (38). Recently, an impedance spectroscopy method was used tocharacterize immobilized protein layers on the gate of an ISFET and to detect an antigen-antibody recognition event (39).

Indirect-Sensing IMFET

Although the ion-step method is an indirect way of measuring antigen-antibody reaction that occurs on the gate region of an ISFET, it does not involve any reagents that enhance or amplify the signal intensity. Many approaches to transduction of the antibody-antigen combining event are indirect. They are based on the ability of an enzyme label to produce electroactive substances within a short span of time. Antibody or antigen is immobilized on the gate area of pH-FET. In the competitive binding assay, the sample antigen competes with enzyme-labeled antigen for the antibody-binding sites on the membrane. The membrane is then washed, and the probe is placed in a solution containing the substrate for the enzyme. IMFETs based on the sandwich assay are applicable for measuring large antigens that are capable of binding two different antibodies. Such sensors use an antibody that binds analyte-antigen, which then binds an enzyme-labeled second antibody. After removal of the nonspecifically adsorbed label, the probe is placed into the substrate-containing solution, and the extent of the enzymatic reaction is monitored electrochemically. Gate voltage is supplied by reference electrode, such as Ag/AgCl or a Hg/Hg_2Cl_2 electrode, that is immersed in a sample solution. It is, however, difficult to make a small conventional electrode, which prevented the IMFET from being miniaturized as a whole. When a noble metal, such as platinum or gold, is used as a reference electrode, the potential between the metal electrode and sample solution fluctuates. The fluctuation makes stable measurement impossible. A method to cancel the fluctuation using a reference ISFET (REFET) is reported. A combination of two kinds of ISFET is used, one of which detects a specific substance whereas the other (REFET) does not detect it (Fig. 6). Thus, measuring the differential output between the two ISFETs can cancel the potential fluctuation in the sample solution and drift due ISFET (40–42).

Most of the indirect-sensing IMFET studies are carried out using urease-conjugated antibodies. Urea is used as a substrate. The immunosensor uses a reaction wherein urea is hydrolyzed by the urease-labeled second antibody. The

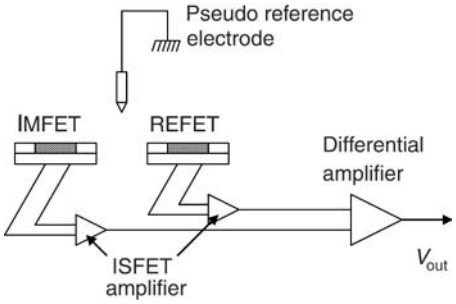


Figure 6. Differential measurement setup for an IMFET.

reaction is

$$H_2NCONH_2 + 2H_2O + H^+ \rightarrow 2NH_4^+ + HCO_3^-$$

According to the reaction, the pH value in the membrane becomes high. On the other hand, on the ISFET surface with inactive antibody membrane, the above reaction does not occur and pH remains constant. Hence, by measuring the differential output between two ISFETs, only pH changes due to urea hydrolysis are detected. In some cases, the authors used antibodies conjugated with the glucose oxidase. These sensors use oxidation of glucose by glucose oxidase. In the reaction, gluconic acid is produced and the pH value in the glucose oxidase immobilized membrane becomes low. To achieve a high sensitivity of horseradish peroxidase (HRP) detection, various substrates, either alone or in combination, are tested and the result is shown in Fig. 7.

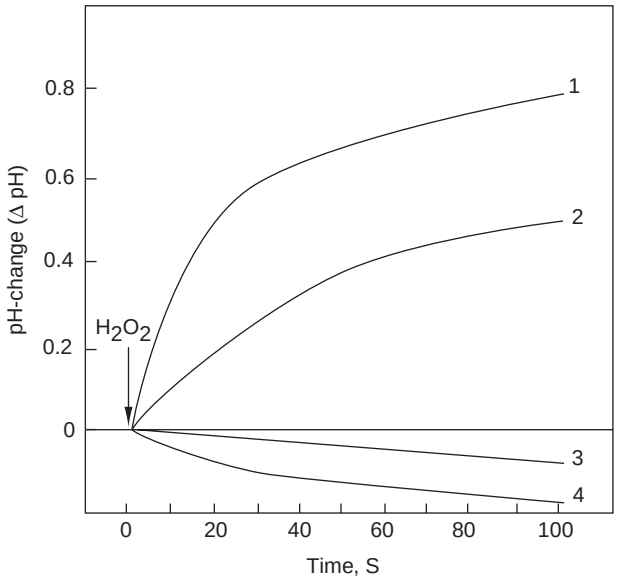


Figure 7. Typical ISFET responses for HRP (10^{-9} *M*) substrates (1) Ascorbic Acid + O-phenylenediamine (OPD); (2) OPD; (3) piodophenol+ luminol; (4) catechol.

PRACTICE

Direct-Acting IMFET

The rationale for attempting to combine the fields of immunology and electrochemistry in the design of analytical devices is that such a system should be sensitive due to the characteristics of the electrochemical detector while exhibiting the specificity inherent in the antigen-antibody reaction. The ideal situation would be to detect the binding of immunoreagents directly at an electrode, for example, by changes in surface potential, which could be truly described as an immunosensor (43,44). Much more effort has been committed to develop transducers, which rely on direct detection of antigen by the antibody immobilized on its surfaces (or vice versa). In 1975, Janata immobilized a sugar-binding protein Concanavalin A on a PVC-coated platinum electrode and studied its responses in the presence of sugar (30). The potential of the electrode with respect to an Ag/AgCl electrode changed owing to adsorption of the charged macromolecule. Although the system reported was not based on an immunochemical reaction, the finding of a potentiometric response stimulated further investigations in this field. Direct potentiometric sensing of antigen human choriogonadotropin (hCG) with an anti-hCG antibody sensitized titanium wire resulted in 5 mV shifts with respect to a saturated calomel electrode (45). The change in potential was explained by a simple charge transfer model.

In 1978, Schenck first proposed a concept of direct immunosensing by an ISFET (46,47). He suggested using FET with, on the gate region, a layer of antibody specific to a particular antigen. Replacement of electrolyte solution with another electrolyte solution-containing antigen should alter the charge of the protein surface layer due to the antigen-antibody reaction, thus affecting the charge concentration in the inversion layer of the transistor. The corresponding change in the drain current would then provide a measure of the antigenic protein concentration in the replacement solution. Many research groups have tried to realize the proposed concept of Schenck, but the results obtained are meager (48,49). Collins and Janata immobilized a PVC membrane containing cardiolipin antigen onto the gate of a previously encapsulated ChemFET (50). They demonstrated that the solution-membrane interface was somewhere between a polarized and a nonpolarized interface, based on the measured membrane exchange current density. The measured potential was therefore a mixed potential deriving out of the permeation of Na^+ and Cl^- ions into and out of the membrane. The change in potential following specific binding of antibody to the membrane was due primarily to a perturbation of the mixed potential, rather than to the adsorbed charge from the antibody itself. Therefore, the device could not be considered selective for the immunoreactive species of interest. Besides, Janata reported that it is impossible to construct an IMFET without having an ideal polarized solution-insulator interface. He proclaimed all of his earlier results as artifacts (51). In spite of these practical difficulties, Gotoh et al. (52) published results obtained with an IMFET sensitive to Human serum albumin (HSA).

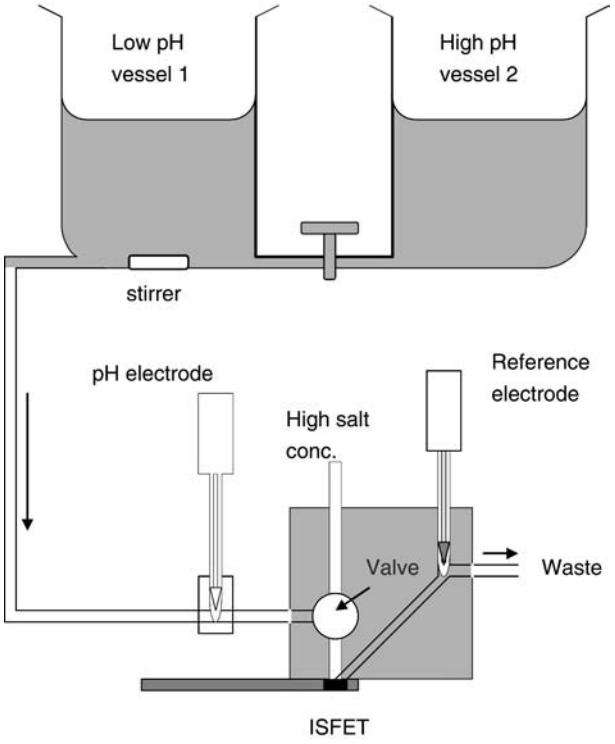


Figure 8. Outline of an Ion-step flow through system.

A 2 mV shift was detected with HSA containing polyvinylbutyral membrane deposited on an ISFET after reaction with its antibody. It appears that experimental results obtained with direct detection of protein on solid-state electrode or similar devices are, so far, limited to second-order effects. Nevertheless, a real theoretical explanation is absent. Therefore, until more experimental evidence is available, the true value of direct-acting IMFET concept remains to be established.

Schasfoort et al. (36) proposed an alternative approach to overcome the above-described difficulties of a direct detection of immunological reaction with ISFET. By stepwise changing the electrolyte concentration of the sample solution, a transient diffusion of ions through the membrane-protein layer occurs, resulting in a transient membrane potential, which can be measured by the ISFET. A flow-through system was used to carry out the experiments as schematically drawn in Fig. 8. The pH of the electrolyte can be changed by using a gradient vessel. When the solution flows under hydrodynamic pressure out of vessel 1, the pH will change through mixing with a solution of different pH from vessel 2. By opening the value for a few seconds, the ISFET can be exposed to a higher salt concentration. The step change in join concentration was completed within 50 ms. After 2 s the valve was closed and the membrane can gain equilibrate with the buffer flowing out of vessel 1. In order to exchange the electrolyte concentration rapidly, the volume between the valve and the ISFET was kept small. ISFETs with a polystyrene-agarose membrane were incubated with 10^{-5} M HSA for 3 h. The ISFET response was measured as a function of the pH, and the inversion point was determined to be pI $= 3.72 \pm 0.05$. Subsequently, the ISFETs were incubated in different concentrations of anti-HSA antibodies solution ranging from 0.06 to 64 μM. The anti-HSA antibody was able to change the inversion point of the HSA-coated membrane from 3.70 to 5.55. The above experiments clearly demonstrated that the net charge density in a protein layer deposited on an ISFET could be determined by exposing the membrane to a stepwise change in electrolyte concentration while measuring ISFET current change. The transient membrane potential observed is a result of the different mobilities of the positive and negative ions present in the protein layer. It is also observed that characteristic inversion points and slope are a function of the protein concentration and type of protein. Also isolectric points could be detected from the membrane potentials as a function of the pH. This detection of the isoelectric point of a protein complex is the basis for the development of an IMFET. An immunological reaction results in a change of the fixed-charge density in the membrane, which can be explained by a shift of the protein isoelectric point due to the immunological reaction.

The ion-step method was originally designed to measure immunoreaction via the change in charge density, which occurs in an antibody-loaded membrane, deposited on an ISFET, upon reaction with a charged antigen. The efficacy of ion-step method for the quantification of a non-charged antigen was demonstrated using progesterone as the model analyte (53). Progesterone is an uncharged molecule, hence, it cannot be detected directly by using the ion-step method. A competitive method was devised using a charged progesterone-lysozyme conjugate. To prepare the ISFETs for ion-step measurement, a membrane support was created by depositing a 1:1 mixture of polystyrene beads and agarose on the gate. The ISFETs were then cooled to 4 °C and the solvent was slowly evaporated, leaving a porous membrane with a thickness of approximately 4 μm. The ISFET was then heated to 55 °C for 1 h to immobilize the membrane onto the gate. The ISFET was placed in the flow-through system (see Fig. 8) and a monoclonal antibody specific to progesterone was incubated on the membrane (0.5 mg/ml, 4 °C for 20 h). A competitive assay method was used to detect progesterone levels, and the detection limit was approximately 10^{-8} M of progesterone in the sample solution. Recently, Besselink et al. (54) described an amino bead-covered ISFET technology for the immobilization of antibodies. HSA was immobilized onto the amino bead-coated ISFET, by covalent cross-linking method, and the anti-HSA antibodies were quantitated using the ion-step method. The antibody concentration was detected within 15 min, with yields up to 17 mV (Fig. 9).

Indirect-Sensing IMFET

The indirect-sensing IMFET concept emerged during the early 1990s in order to overcome the difficulties met with the direct-acting IMFET devices (55). Colapicchioni et al. (56) immobilized IgG using protein A onto the gate area of an ISFET. The efficacy of the IMFET was demonstrated using Human IgG and atrazine antibodies captured using protein A. As the atrazine is a small molecule (hapten), which does not induce an immunoresponse as such, it was linked to a carrier protein. Bovine Serum Albumin (BSA)

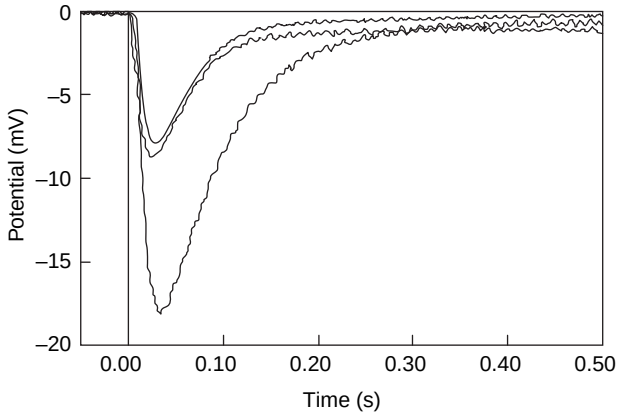


Figure 9. Ion-step responses of HSA-coated ISFET before (upper solid curve) and after incubation (for 15 min) with undiluted anti-HSA (lower solid curve) and anti-BSA (dashed curve). Ion stepping was performed at pH 4.02.

was conjugated to ametryn sulfoxide, which has structural similarity with atrazine, and the ametryn-BSA conjugate was injected into rabbit to raise antibodies. A sandwich assay format was used to detect Human IgG and a competitive assay format was used to quantitate atrazine concentration. The antigen-antibody reaction was monitored by the addition antihuman IgG-GOD conjugate and ametryn-GOD, respectively. Glucose was used as the substrate and the pH variation was detected by the ISFET. The sensitivity of the assay was 0.1 μg/ml and 1 ppb for human IgG and atrazine, respectively. An ISFET-based immunosensor was demonstrated for the detection of bacterial (*Clostridium thermocellum*) cells. The analysis included the reaction of antibacterial antibodies with cells in suspension or after covalent immobilization of cells on porous photoactivated membranes and, subsequently, the revelation of bound antibodies by the conjugate of protein A and HRP and the quantitation of enzyme activity with ISFET. The sensitivity of the sensor was within a range of 10^4–10^7 cells per ml (57). Selvanaygam et al. (23) reported ISFET-based immunosensors for the qunatitation of β-Bungarotoxin (β-BuTx), a potent presynaptic neurotoxin from the venom of *Bungarus multicinctus*. A murine monoclonal antibody (mAb 15) specific to β-BuTx was immobilized on the gate area, and the antigen-antibody reaction was monitored by the addition of urease-conjugated rabbit anti-β-BuTx antibodies. The sensor detected toxin level as low as 15.6 ng/ml. The efficacy of the sensor for the determination of β-BuTx from B. *multicinctus* venom was demonstrated in the mouse model.

An immunological *Helicobacter pylori* urease analyzer (HPUA), based on solid-phase tip coated with a monoclonal antibody specific to *H. pylori*'s urease and ISFET, was reported by Sekiguchi et al. (58). A solid-phase tip, with an inner diameter of 0.55 mm, coated with the monoclonal antibody, was incubated for 15 min at room temperature in an endoscopically collected gastric mucus sample. The activity of urease captured on the inner surface of the solid-phase tip was measured by coupling it with an ISFET in a measuring cell containing urea solution. The pH change of urea solution after 55 s of the enzymatic reaction inside the tip was measured by withdrawing 1.1 μl of solution toward the upstream of the tip, where the measuring ISFET was installed. One cycle of measurement was completed in 17.5 s, and the sensitivity of system was 0.2 m IU/ml. The calibration curve for the quantitation of urease is shown in Fig. 10. Clinical studies were carried out with 119 patients (75 males and 44 females with an average age of 51, ranging from 13 to 79) who underwent gatroduodenoscopy and judged necessary to evaluate the infection of *H. pylori* and urea breath test (UBT) was used as a gold standard. Thirty-three of the UBT positive 36 patients were positive, and 81 of UBT negative 83 patients were negative by HPUA resulting in the 92% sensitivity and 98% specificity.

An IMFET for the detection of HIV-specific antibodies based on a combination of ELISA principle and ISFET flow injection analysis setup was presented by Aberl et al. (59). The active sensing components consist of a reaction cartridge containing a carrier with the immobilized receptor layer and an ISFET sensor mounted in a flow-through cell. A flow cell was constructed using two ISFET sensors on one in a two-channel configuration (Fig. 11). The liquid

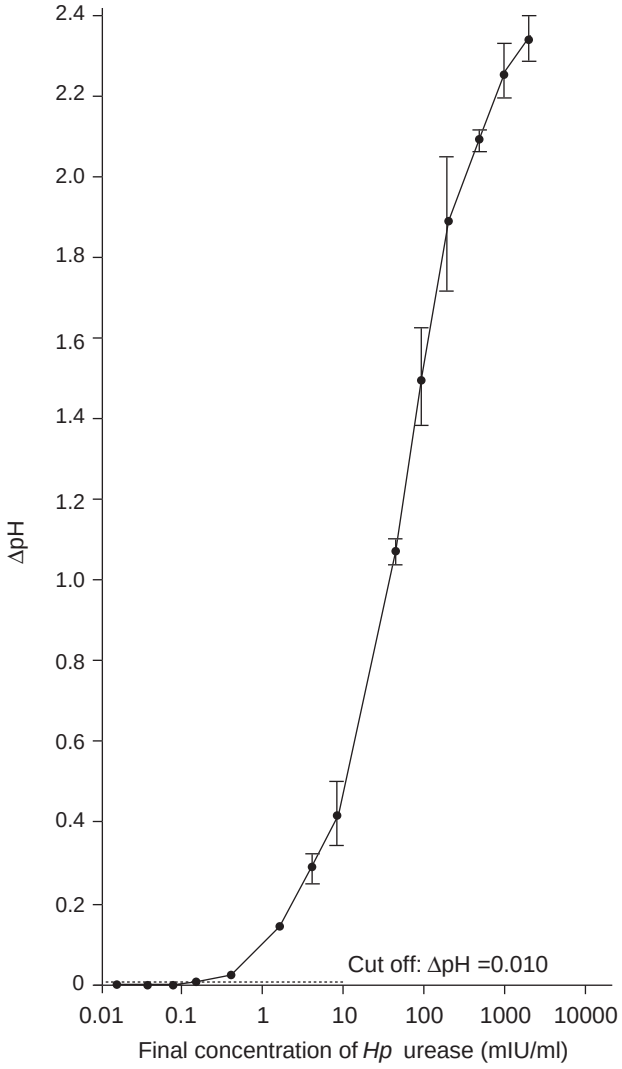


Figure 10. A standard curve for HPUA.

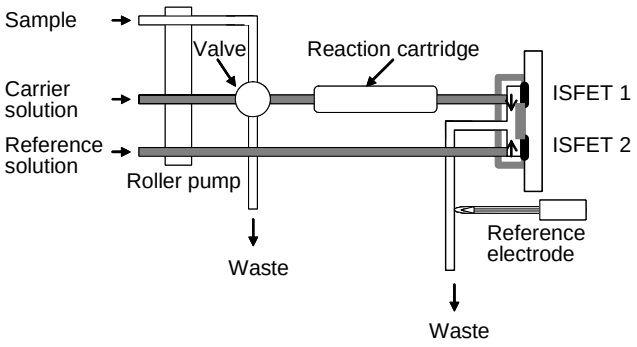


Figure 11. Diagrammatic representation of a flow injection system for indirect immunosensing.

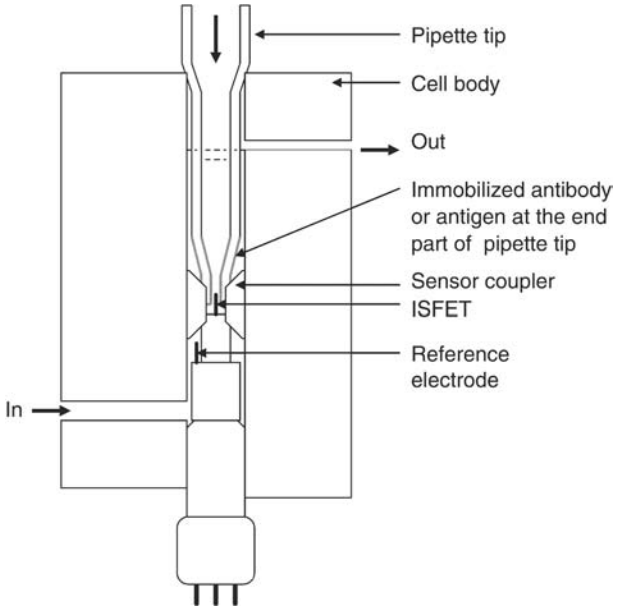


Figure 12. Cross-sectional view of a pH-measuring cell.

headspace on top of the ISFET sensors was reduced to about 1 μl, and the dead volume of the whole sensor cell was 7μl. The detection principle was realized according to the sandwich ELISA procedure using urease as a pH shifting marker enzyme. Antigen molecules (p24 or gp120) were immobilized on cellulose nitrate membranes mounted in a special flow-by cartridge or the inner surface of Borosilicate glass capillary tubing. After blocking the unspecific binding sites, the antigen was reacted with specific serum in different dilution or nonspecific serum as a negative control. In comparison with conventional ELISA, the ISFET-FIA ELISA showed a slight lower sensitivity. The antibodies were detected in a serum diluted more than 1:12,000 in ELISA, whereas the sensitivity of the ISFET– FIA ELISA was between a 1:1000 and a 1:10,000 dilution. Glass as a support material showed highly reproducible test results when compared with cellulose nitrate membrane.

Tsuruta et al. (60) reported a fully automated ISFET-based ELISA system using a pipette tip as a solid phase and urease as a detecting enzyme. The inner wall of the end part of a pipette tip was used as a solid phase, and the urease activity of the conjugate, captured after a two-step immunoreaction, was measured by coupling the pipette tip with the ISFET in a pH measuring cell (Fig. 12). A two-step sandwich assay procedure was used for the quantitation of AFP, CEA, HBsAg, and HBsAb, and a two-step competition assay was used for HBcAb, and second-antibody configuration was used for HTLV-1 Ab. After final incubation in conjugate solution, the pipette tip was washed and it was introduced into the pH measuring cell in order to couple it with ISFET. At the same time, feeding of the substrate solution was stopped, to read the pH change for 20 s. The output (source potential) of the ISFET was read and stored in the CPU during the above-mentioned 20s at 0.1 s intervals. The maximum changing rate of the source potential ($\Delta V/\Delta t$, mV/*s*) was calculated from these 200 data points. The total assay time was 21 min as the sum of 5, 10, 5 and 1 min for preheating of sample, First immunoreaction, Second immunoreaction, and pH measurements, respectively. The assay speed was 60 samples/h. Assay performance, such as within run CVs, between run CVs, detection limits, and correlation with the conventional ELISA kits, were satisfactory for all of six analytes. The detection limit for CEA, 0.09 μg/l was comparable to better than those reported for the most advanced chemiluminescent ELISA system (0.086 μg/l).

Polymerase chain reaction (PCR) has proven to be of great importance in clinical diagnosis (61). Usually, the PCR products have been detected by staining with ethidium bromide in qualitative methods, and fluorescent dyes in real-time quantitation. Although electrophoresis has the advantage of giving information on the molecular size of PCR products, it is not well-suited to mass screening or automation. On the other hand, real-time monitoring is well-suited for mass screening and automation but is expensive. One of the most promising methods for automatizing the detection system of PCR products is ELISA. An ISFET-based ELISA was used to quantitate PCR products (62). Double-stranded PCR products with digoxigenin and biotin at both terminals were obtained by using digoxigenin-and biotin-labeled primers. The PCR products were detected by a two-step sandwich ELISA. One μl of the solution after PCR was introduced into the end part of the solid-phase pipette tip coated with antidigoxigenin antibody. Biotin-labeled PCR products captured at the solid phase were detected with avidin-urease conjugate, and the enzyme activity was measured by the ISFET in a pH measuring cell containing urea solution. The detection limit of the system was determined using a known amount of purified PCR product labeled with digoxigenin and biotin, and it was found that 10 amol of the labeled DNA in 1 μl sample. The assay was used to detect HTLV-1 provirus gene integrated in the genome of human MT-Cell, and it was found that 100 pg of the genomic DNA was specifically detectable after 35 cycles of PCR. The apparent dynamic range for detection of MT-1 DNA was from 100 pg to 100 ng.

One of the most important targets in molecular biology is the quantitation of mRNA related to special disease by RT-PCR. The accuracy of quantitative RT-PCR has been remarkably improved by the introduction of competitive

RT-PCR, in which a synthetic RNA is used as an internal standard (63). Tsuruta et al. (64) developed a ISFET-ELISA method for the qantitatiion of mRNA in clinical samples. In this method, a fixed amount of a synthetic RNA, pRSET RNA, was added as internal standard to the solution of target RNA (IL-1β) after reverse transcription, PCR was carried out using digoxigenin-labeled sense primer and biotin-labeled antisense primer for IL-1β, and FITC-labeled sense primer and a biotin-labeled antisense primer for pRSET. The double-stranded PCR products of IL-1β and pRSET were captured by two solid-phase pipette tips, one coated with antidigoxigenin antibody and another with anti-FITC antibody, respectively, and sandwiched by an avidin-urease conjugate, whose activity was measured with ISFET. The ratio of the signal intensity for IL-1β to that for pRSET was used to quantitate the concentration of IL-1β. A calibration curve was obtained using a known amount of AW109 RNA as an external standard in place of IL-1β m RNA. It was found that 10^2–10^6 copies of IL-1β mRNA were measurable by the present method. Expression levels of IL-1β mRNA in clinical samples, such as monocytes of peripheral blood or synovial cells from patients with RA or OA, were determined.

Practical Limitations

In this section, we shall address some practical problems that have been limiting factors in the commercial application of IMFETs. The widespread use of IMFETs for applications ranging from medical diagnosis to industrial process control or environmental monitoring has not actually happened. The underlying reasons for this situation fall into two main categories, those that are inherent to the transistor, such as material, encapsulation, and reference electrode, and those problems common to its application as an immunosensor function, such as, antibody immobilization, stability, and durability. The pH sensing properties and drift behavior of the ISFET is the main limiting factor in the commercial breakthrough of ISFET. After the invention of the ISFET, initially the only gate material used was SiO_2. Although SiO_2 showed pH, sensitivity of 20 to 40 mV/pH, the thermally grown gate oxide loses its isolation property within a few hours of immersion in a solution. In order to isolate this gate oxide from the solution, another isolating layer, such as Si_3N_4, Al_2O_3, or Ta_2O_5, has to be placed on top of this gate oxide. A layer of Si_3N_4 on top of SiO_2 showed 45–50 mV/pH, and other layers, such as Al_20_3 and Ta_2O_5, showed even higher pH sensitivity, 53–57 mV/pH and 55–59 mV/pH, respectively (65). Drift rate for Si_3N_4 is reported as 1 mV/h and for Al_2O_3 and Ta_2O_5 0.1–0.2 mV/h after 1000 min of operation at pH 7.0. In most of the work on IMFETs published so far, these three gate materials, Si_3N_4, Al_2O_3, and Ta_2O_5, have been used. IMFETs are also sensitive to light and temperature (66).

The pH-sensitive ISFETs can be fabricated by means of standard MOS technology, except for the metallization step. However, after dicing the wafers into single chips, the substrate becomes exposed at the edges of the senor. Encapsulation and packaging are two final processing steps that determine reliability and durability (lifetime) of the IMFETs. In order to achieve high quality sensors, all electrical components have to be isolated from their surroundings. Several reports exist on the encapsulation and packaging of ISFET devices for pH application (21). The simplest method of isolating these sides is encapsulation with epoxy-type resins. The most important ISFET characteristics, such as stability, accuracy, and durability, also pertain to the reference electrode. One of the major hurdles in IMFETs is the lack of a solid-state reference electrode. The small IMFETs have to be combined with a conventional KCl-solution-filled reference electrode. In order to achieve miniaturized IMFET, it is important to miniaturize the reference electrode. In general, two approaches have been followed: reference FETs (REFETs), which are used in an ISFET/REFET/quasi-reference electrode setup, and miniaturized conventional reference electrodes. In the first approach, attempts have been made to cover the ISFET surface with a pH-insensitive layer or to render the surface pH insensitive by chemical modification. In the second approach, the structure of a conventional electrode (mostly Ag/AgCl type) is miniaturized partially or completely on a silicon wafer. Its potential is a function of concentration of chloride ions. They are supplied either from an internal electrolyte reservoir formed by an anisotropic etching in the silicon wafer or by adding chloride ions into the test solution.

Some of the technological factors such as pH sensitivity and drift can now be overcome with the existing technology. A hurdle peculiar to direct-acting IMFET is the need to provide a thin but fully insulating layer (membrane) between the antigen or antibody coating and the semiconductor surface. Such a membrane must be thin enough to allow a small charge redistribution occurring as a result of analyte (antigen-antibody) binding to exert a detectable change in electrical field. Conversely, it must also provide adequate insulation to prevent dissipation of the field by leakage of ions. Even assuming that the ideal insulating membrane can be developed, a further hurdle may need to be overcome. Surface charges and hydrogen binding sites of proteins cause a counter-ion shell (double-layer) and structured water shells to surround the molecules; these regions of structured charge will inevitably contribute to the electrical field affecting the FET gate. Pending these breakthroughs, the development of direct-acting IMFETs appears to be stagnant.

The immobilization methods used for immunosensors include a variety of adsorption, entrapment, cross-linking, and covalent methods. In general, a covalent immobilization method consisting of silanization step and subsequent coupling procedure via glutaraldehyde has been used to immobilize antibodies onto the gate region (67,68). However, no methodical investigation about antibody stability, storage, and lifetime exists. Reproducible regeneration of the sensing area is one of the major problems met with IMFETs that have been used for continual monitoring or repeat usage. The need for renewal of the sensing surface derives from the high affinity constants derived from the strong antigen-antibody reaction. Two different strategies have been used to achieve the renewal of the sensing surface, breakage of the antigen-antibody bond and reusing the immunologic reagent immobilized on the solid phase. A second alternative is the elimination of antigen-antibody

complex from the solid support and immobilization of fresh immunologic material. Dissociation of antigen from antibody is usually carried out in low pH and high ionic strength solutions. Protein A was chemically immobilized onto the gate surface by using a polysiloxane layer of [3-(2-aminoethyl)aminopropyl]trimethoxysilane (APTES) and cross-linking agent such as glutaraldehyde. Reversibility of the linkage between Protein A and antibodies in order to restore the device for the next measurement was studied by breaking the antibody-antigen complex formed on Protein A using a variety of reagents. Glycine buffer pH 2 and 3 and $MgCl_2$ 3.5 M were found to be more effective when compared with other tested reagents due to high ionic strength (55). Selvanayagam et al. (23) studied the reusability of an ISFET sensor by removing the antibody membrane from the gate area. The regenerated devices tested were reported to function normally five times, although a considerable amount of time was required for the regeneration process. Recently, IMFET using magnetic particle and integrated to flow injection system has been described to overcome the problem of regeneration (69,70). The immunological material was immobilized on the surface of magnetic particles and were transported by a flow system, and were retained on the gate area of the ISFET by a magnetic field produced by a magnet (Fig. 13). The regeneration of immunologic materials was achieved by releasing the magnetic field, thus freeing those particles that were washed by the flow system, and new magnetic particles were injected and retained on the surface of transducer by reacting the magnetic field. A fresh and reproducible surface was thus produced, ready for the next analytical cycle.

The main barrier to the successful introduction of IMFETs for clinical testing is undoubtedly the high performance and automation level of the machines that already exist in centralized laboratories. They have been developed specifically for use with either immunoassay or clinical chemistry. Immunoassay performance is continually being optimized and assay times have been reduced over the past few years. Depending on the parameter, the assay time can be as low as 6 min and the majority of the larger machines could carry out between 100 to 200 testes per hour. IMFETs must be compared with these methods with respect to assay time, sensitivity, and cost. The need for in-built calibration has been frequently encountered in sophisticated quantitative IMFETs. Although feasible and acceptable in laboratory-based instrumentation, it remains a major problem in small disposable IMFET devices. To facilitate the increased use of IMFETs, one should look for real tests and standards that prototypes can meet, based on current diagnostic needs and perceived future development. The progress of IMFET beyond the experimental laboratory level is mainly dependent on how skillfully its development and marketing are combined with parameter selection.

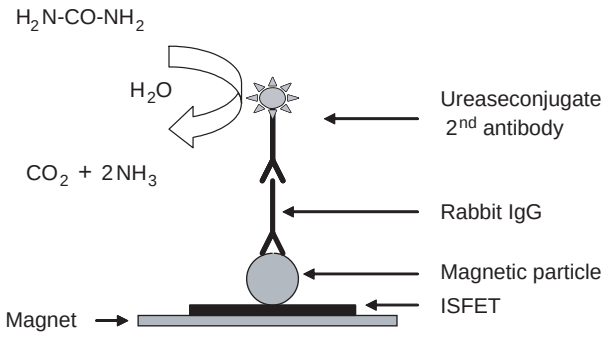


Figure 13. Schematic representation of the magnetoIMFET.

FUTURE DIRECTIONS

A key consideration in antibody immobilization to the gate area is to maintain, reproducibly, the highest possible binding activity after immobilization while conserving the amount of antibody used. However, many aspects, both fundamental and more applied, require in-depth study before IMFETs can become successful commercial device, including improved control of biomolecule immobilization and novel immobilization strategies; enhancement of biomolecule stability and retention of activity *in vitro*; and the ability to reproduce the high signal-to-noise ratios obtained in simple test solutions in "real" samples such as blood or water. The IMFETs tends to respond nonspecifically to any molecule bound to the surface; hence, it affects the measurement parameter to some extent. The specificity of analyte detection, therefore, relies entirely on achieving high ratios of specific to nonspecific binding, which, in the context of low concentrations of analyte in blood, can represent a formidable problem. Reduction of nonspecific binding is another area that will continue to be of major importance. The ability to immobilize ordered antibodies will maximize antigen binding to a given surface while reducing the availability of nonbinding site sections of the immobilized antibody, or uncovered surface areas, which can promote nonspecific interaction with other components in the sample. The potential advantages of using Fab fragments rather than more complex intact antibodies (such as IgG) could be explored.

IMFETs, similar to immunoassays, involve multistep procedures, including washing steps, incubation periods, and quite complex signal generation protocols. It is likely that research efforts into a novel signal amplification system, without the normally associated complications of multi-reagents or multistep protocol will be of increasing importance. The irreversibility of antigen-antibody interaction presents a major challenge in designing IMFETs for continual monitoring or repeated usage. Treatment of an antibody-antigen complex with a mildly denaturing medium for a short time interval has shown some promise in regenerating sensor surfaces. Development of enhanced denaturation conditions, which optimize dissociation of antigen while minimizing irreversible loss of antibody structural integrity, may be possible in the near future. The use of catalytic antibodies in immunosensors has been proposed. The ability of catalytic antibodies to catalyze the hydrolysis of phenyl acetate with the formation of acetic acid allows integration of pH-sensitive microelectrodes to give a potentiometric immunosensing system (71). The advantage of catalytic antibodies over normal antibodies is that reversibility of

response is readily achieved, because bound antigen reacts to form a product with a low affinity for the antibody, resulting in dissociation. As the binding event is followed immediately by a catalytic reaction and release of the reaction products, the molecular recognition site is regenerated with each molecular reaction; as a consequence, catalytic antibodies can be used to create reversible IMFETs for continuous monitoring of analyte concentrations. Improvements in sensitivity and cross-reactivity are likely to be achieved as a result of the increasing interest in this field of research.

CONCLUSION

Although the ISFET concept has existed for over 30 years, its practical applications such as the IMFETs are still emerging very slowly. The relatively slow rate of progress of IMFET technology from inception to fully functional commercial devices for these applications is a reflection of technology-related and market factors. In general, two approaches have been followed in the past to realize IMFETs. In the first approach, antigen-antibody reaction on an immonbilized membrane was monitored without any addition of labels. The second approach takes the advantage of an enzyme label to indirectly monitoring the antigen-antibody reaction using pH-sensitive FET. The development of IMFETs that directly detect antigen-antibody reaction is extremely challenging; only a few examples exist, the majority of which are without valid theoretical explanation. Although it shows enormous promise in the early stages of development, an effective, reliable, and analyte-selective direct-acting IMFET sensor is yet to be constructed. The ion-step method represents a novel measurement concept for potentiometric detection and quantification of an adsorbed antigen or antibody molecule in which modified ISFETs are used. Many approaches to transduction of the antibody-antigen combining event are indirect, necessarily involving the use of reagents admixed with analyte, and therefore cannot be seen as routes to development of "True" IMFETs. Nevertheless, such reagent-dependent, indirect-sensing IMFETs may offer real commercial advantages over the current generation direct-acting IMFET readout technologies. The clinical diagnostic field offers real opportunities for the exploitation of IMFET, but because it is a highly competitive and well-established market, those who wish to introduce new products must carefully target their market niche. IMFETs will have to compete with such technology on the basis of factors such as cost, ease of use, sensitivity, operational stability, robustness, and shelf-life.

BIBLIOGRAPHY

1. Catty D. editor, Antibodies Volume 1. A Practical Approach. Washington, DC: IRL; 1988.
2. Mayforth RD. Designing Antibodies. New York: Academic Press; 1993.
3. Johnstone AP, Turner MW. Immunochemistry 1. A Practical Approach. New York: IRL; 1997.
4. Bluestein BI, Walczak IM, Chen SY. Fiber optic evanescent wave immunosensors for medical diagnostics. Trends Biotechnol 1990;8:161–168.
5. Gosling JP, A decade of development in immunoassay methodology. Clin Chem 1990;36:1408–1427.
6. Kemeny DM, A Practical Guide to ELISA. New York: Pergamon Press; 1991.
7. Turner APF, Karube I, Wilson GS. editors. Biosensors: Fundamentals and Applications. New York: Oxford University Press; 1987.
8. Buerk DG, Biosensors: Theory and Applications. Lancester, PA: Technomic; 1993.
9. Eggins BR. Biosensors: An Introduction. New York: Wiley; 1996.
10. Aizawa M. Principles and applications of electrochemical and optical biosensors. Anal Chim Acta 1991;250:249–256.
11. North JR. Immunosensors: Antibody-based biosensors. Trends Biotechnol 1985;3:180–186.
12. Aizawa M. Immunosensors for clinical analysis. Adv Clin Chem 1994;31:247–275.
13. Morgan LC, Newman DJ, Price CP. Immunosensors: Technology and opportunities in laboratory medicine. Clin Chem 1996;42:193–209.
14. Ghindilis AL, Atanasov P, Wilkins M, Wilkins E. Immunosensors: Electrochemical sensing and other engineering approaches. Biosens Bioelecton 1998;13:113–131.
15. Bergveld P. Future applications of ISFETs. Sens Actuators B 1991;4:125–133.
16. Kuriyama T, Kimura J. An ISFET biosensor. In: Wise DL, editor, Applied Biosensors. Boston, MA: Butterworth; 1989.
17. Kimura J, Kuriyama T. FET biosensors. J Biotechnol 1990;15:239–254.
18. Van der Schoot BH, Bergveld P. ISFET based enzyme sensors. Biosensors 1987/88;3:161–186.
19. Yuquing M, Jianguo G, Jianrong C. Ion-sensitive field effect transducer-based biosensors. Biotechnol Advances 2003; 21:527–534.
20. Janata J, Moss SD. Chemically sensitive field-effect transistors. Biomed Eng 1976;11:241–245.
21. Bergveld P, Sibbald A. Comprehensive Analytical Chemistry, XXIII: Analytical and Biomedical Applications of Ion-Selective Field-Effect Transistors. New York: Elsevier; 1988.
22. Bergveld P. Thirty years of ISFETOLOGY. What happened in the past 30 years and what may happen in the next 30 years. Sens Actuators B 2003;88:1–20.
23. Selvanayagam ZE, Neuzil P, Gopalakrishnakone P, Sridhar U, Singh M, Ho LC. An ISFET-based immunosensor for the detection of β-Bungaraotoxin. Biosens Bioelectron 2003; 10:405–414.
24. Janata J, Huber RJ. Chemically sensitive field effect transistors. In: Freiser H, editor. Ion-Sensitive Electrodes in Analytical Chemistry, vol. II, New York: Plenum Press; 1980.
25. Janata J, Blackburn GF. Immunochemical potentiometric sensors. Ann N Y Acad Sci 1984;428:286–292.
26. Blackburn GF. Chemically sensitive Field effect transistors. In: Turner APF, Karube I, Wilson S. editors. Biosensors: Fundamentals and Applications. Oxford: Oxford Science Publications; 1987.
27. Bockris JOM, Reddy AKN. Modern Electrochemistry, vol. 2. New York: Plenum Press; 1970.
28. Eisen HN. Immunology. An Introduction to Molecular and Cellular Principles of the Immune Response. New York: Harper and Row; 1974.
29. Bergveld P. A critical evaluation of direct electrical protein detection methods. Biosens Bioelecton 1991;6:55–72.
30. Janata J. An immunoelectrode. J Am Chem Soc 1975;97:2914–2916.

31. Janata J, Janata J. Novel protein-immobilizing hydrophobic polymeric membrane, process for producing same and apparatus employing same. U.S. Patent 3,966,580, 1976.
32. Schasfoort RBM, Bergveld P, Kooyman RPH, Greve J. Possibilities and limitations of direct detection of protein changes by means of immunological field-effect transistor. Anal Chim Acta 1990;238:323–329.
33. Moore WJ. Physical Chemistry. London: Longman; 1976.
34. Davies JT, Rideal EK. Interfacial Phenomena. London: Academic Press; 1963.
35. Schasfoort RBM, Bergveld P, Bomer J, Kooyman RPH, Greve J. Modulation of the ISFET response by an immunological reaction. Sens Actuators 1989;17:531–535.
36. Schasfoort RBM, Kooyman RPH, Bergveld P, Greve J. A new approach to ImmunoFET operation. Biosens Bioelecton 1990; 5:103–124.
37. Schasfoort RBM, Eijsma B. The ions-step method applied to disposable membranes for the development of ion-responding immunosensors. Sens Actuators 1992;6:308–311.
38. Eijkel JCT, Olthuis W, Bergveld P. An ISFET-based dipstick device for protein detection using the ion-step method. Biosens Bioelectron 1997;12:991–1001.
39. Kharitonov AB, Wasserman J, Katz E, Willner I. The use of impedance spectroscopy for the characterization of protein-modified ISFET devices: Application of the method for the analysis of biorecognition processes. J Phys Chem 2001; B-105:4205–4213.
40. Sergeyeva TA, Soldatkin AP, Rachkov AE, Tereschenko MI, Piletsky SA, Elskaya AV. β-Lactamase label-based potentiometric biosensor for α-2 interferon detection. Anal Chim Acta 1999;390:73–81.
41. Yacoub-George E, Wolf H, Koch S, Woias P. A miniaturized ISFET-ELISA system with a pre-treated fused silica capillary as reaction cartridge. Sens Actuators B 1996;34:429–434.
42. Starodub NF, Dzantiev BB, Starodub VN, Zherdev AV. Immunosensor for the determination of the herbicide simazine based on an ion-selective field-effect transistor. Anal Chim Acta 2000;424:37–43.
43. Wang J. Electroanalytical Techniques in Clinical Chemistry and Laboratory Medicine, New York: VCH; 1988.
44. Luppa PB, Sokoll LJ, Chan DW. Immunosensors — principles and applications to clinical chemistry. Clin Chim Acta 2001;314:1–26.
45. Yamamoto N, Nagasawa Y, Sawai M, Sudo T, Tsubomura H. Potentiometric investigations of antigen-antibody and enzyme-enzyme inhibitor reactions using chemically modified metal electrodes. J Immunol Methods 1978;22:309–317.
46. Schenck JF. Technical difficulties remaining to the application of ISFET devices. In: Cheung PW, editor. Theory, Design and Biomedical Applications of Solid State Chemical Sensors. New York: CRC; 1978.
47. Schenck JF. Field effect transistor for detection of biological reactions. U.S. Patent 4,238, 757, 1980.
48. Schoning MJ, Poghossiana A. Recent advances in biologically sensitive field-effect transistors (BioFETs). Analyst 2002; 127:1137–1151.
49. Janata J, Thirty years of CHEMFETs — a personal view. Electroanalysis 2004;16:1831–1835.
50. Collins S, Janata J. A critical evaluation of the mechanism of potential response of antigen polymer membrane to the corresponding antiserum. Anal Chim Acta 1982;136:93–99.
51. Janata J. Proceedings of the 2nd International Meeting on Chemical Sensors; 1986:25–31.
52. Gotoh M, Tamiya E, Karube I. Micro-FET biosensors using polyvinylbutyral membrane. J Membrane Sci 1989;41:291–303.
53. Schasfoort RBM, Keldermans CEJM, Kooyman RPH, Bergveld P, Greve J, Competitive immunological detection of progesterone by means of the ion-step induced response of an immunoFET. Sens Actuators 1990;BI:368–372.
54. Besselink GAJ, Schasfoort RBM, Bergveld P. Modification of ISFETs with a monolayer of latex beads for specific detection of proteins. Biosens Bioelecton 2003;18:1109–1114.
55. Toshihide K. Electrochemical sensors in immunological measurement. E.U. Patent 0,328,380, 1998.
56. Colapicchioni C, Barbaro A, Porcelli F, Giannini I. Immunoenzymatic assay using CHEMFET devices. Sens Actuators B 1991;6:186–191.
57. Akimeno VK, Khomutov SM, Obraztsova AY, Vishnivetskii SA, Chuvilskaya NA, Laurinavichus KS, Reshetilov AN. A rapid method for detection of Clostridium thermocellum by field-effect transistor immunodetection. J Microbiol Methods 1996;24:203–209.
58. Sekiguchi T, Nakamura M, Kato M, Nishikawa K, Hokari K, Sugiyama T, Asaka M. Immunological Helicobacter pylori urease analyzer based on ion-sensitive field effect transistor. Sens Actuators 2000;B67:265–269.
59. Aberl F, Modrow S, Wolf H, Koch S, Woias P. An ISFET-based FIA system for immunosensing. Sens Actuators 1992;B6:186–191.
60. Tsuruta H, Yamada H, Motoyashiki Y, Oka K, Okada C, Nakamura M. An automated ELISA system using a pipette tip as a solid phase and a pH-sensitive field effect transistor as a detector. J Immunol Methods 1995;183:221–229.
61. Saiki R, Scharf S, Faloona F, Mullis K, Horn G, Erlich H, Arnheim N. Enzymatic amplification beta-globulin genomic sequences and restriction analysis for diagnosis of sickle cell anemia. Science 1985;230:1350–1354.
62. Tsuruta H, Matsui S, Hatanaka T, Namba T, Miyamoto K, Nakamura M. Detection of the products of polymerase chain reaction by an ELISA system based on an ion-sensitive field effect transistor. J Immunol Methods 1994;176:45–52.
63. Wang AM, Doyle MV, Mark DF. Quantitation of mRNA by the polymerase chain reaction. Proc Natl Acd Sci USA 1989; 86:9717–9721.
64. Tsuruta H, Matsui S, Oka K, Namba T, Shinngu M, Nakamura M. Quantitation of IL-1-beta mRNA by a method of RT-PCR and an ELISA based on ion-sensitive field effect transistor. J Immunol Methods 1995;180:259–264.
65. Gopel W, Hesse J, Zemel JN editorss. Sensors: A Comprehensive Survey, Chemical and Biochemical Sensors, Part I. Verlagsgesellschaft Mbh: VCH; 1991.
66. Neuzil P. ISFET integrated sensor technology. Sens Actuators 1995;B24-25:232–235.
67. Filippo P, Antonio NC. Sensor with antigen chemically bonded to a semiconductor device. E.U. Patent 0,395,137, 1990.
68. Starodub NF, Samodumova IM, Starodub VN. Usage of organosilanes for integration of enzymes and immunocomponents with electrochemical and optical transducers. Sens Actuators 1995;B24-25:173–176.
69. Santandreu M, Sole S, Fabregas E, Alegret S. Development of electrochemical immunosensing systems with renewable surfaces. Biosens Bioelectron 1998;13:7–17.
70. Sole S, Alegret S, Cespedes F, Fabregas E. Flow injection immunoanalysis based on a magnetoimmunosensor system. Anal Chem 1998;70:1462–1467.
71. Blackburn GF, Talley DB, Booth PM, Durfor CN, Martin MT, Napper AD, Rees AR. Potentiometric biosensor employing catalytic antibodies as the molecular recognition element. Anal Chem 1990;62:2211–2216.

See also ION-SENSITIVE FIELD EFFECT TRANSISTORS.

IMMUNOTHERAPY

QIAO LI
University of Michigan
Ann Arbor, Michigan

INTRODUCTION

Immunotherapy of cancer, infectious disease, and autoimmune disease has opened a new area for disease management. This approach has developed very fast lately due to the advances and involvements of modern technology in molecular biology, cell biology, immunology, biochemistry, and bioengineering. Adoptive T cell immunotherapy of cancer involves passive administration of lymphoid cells from one host to another, or back to itself in order to transfer tumor immunity for cancer treatment. It was first realized >20 years ago that adoptive immunotherapy may be feasible to treat human malignancies. However, the early form of this practice was quite simple. It could be as easy as a straightforward blood cell transfer. The apparent inefficiency of antitumor immune responses, and the failure to successfully combat the disease laid the foundation for current concepts of immunotherapy. It did not take too long before it was realized that boosting the antitumor immune response by deliberate vaccination could increase the potential benefits of immune cell-based therapies. In addition, activation of lymphoid cells with monoclonal antibodies (mAb) toward the molecules involved in T cell signaling pathways has resulted in therapeutic effector T cells. The use of immune adjuvants coadministrated with the cell infusion has enhanced the antitumor efficacy of the transferred cells and has made adoptive cellular immunotherapy a promising strategy for cancer treatment. Studies on the trafficking of adoptively transferred cells *in vivo* as well as the identification and characterization of T cell subsets responsible for antitumor reactivity have provided valuable insights toward the development of novel immunotherapeutic strategies. The adoptive immunotherapy of established tumors with the transfer of tumor-reactive lymphoid cells has now been shown to be highly effective against significant tumor burdens both in animal models and in clinical trials. This is, at least in part, due to recent developments in this area, such as treating cancer in special settings (e.g., in lymphopenic hosts induced by prior conditioning); redirecting the effector cells to tumor through genetic engineered chimerical T cell receptors (TCRs) or by transferred tumor antigen-specific TCRs; and the use of these strategies in combination. This article intends to review the above developments that have made adoptive T cell immunotherapy an attractive alternative for cancer treatment.

INDUCTION OF TUMOR-REACTIVE PRE-EFFECTOR T CELLS *IN VIVO*

Successful induction of tumor-reactive "pre-effector" cells in a tumor-bearing host represents the first step toward the conduct of an effective adoptive T cell immunotherapy of cancer. This procedure provides a source of "pre-effector" cells for subsequent T cell activation and expansion *in vitro* to generate large numbers of "effector" cells to be infused back to the tumor-bearing host or cancer patient for therapy. Due to the relative lack of immunogenicity and potential immunosuppressive mechanisms of human malignancies, application of tumor T cell therapy in the clinical setting has been hampered for a long time by difficulties to reliably isolate tumor-sensitized lymphoid cells from the cancer-bearing host. Nevertheless, recent observations in animal studies and clinic trials have led to the development of strategies to induce T cell sensitization *in vivo*.

Peripheral blood lymphocytes (PBL) represents a convenient source of pre-effector cells. However, in most cases, particularly in the case of solid tumors, the frequency of tumor-specific pre-effector cells in PBL is extremely low, generally far below what is observed in response to viral infection. Experimental studies and clinical experience with adoptive immunotherapy have demonstrated that tumor-draining lymph node (TDLN) cells are potentially effective antitumor reagents. Chang et al. was the first to evaluate vaccine-primed LN (VPLN) as a source of lymphoid cells that could be secondarily sensitized by *in vitro* methods to generate effector cells capable of mediating regression of established tumors upon adoptive transfer in clinical trials (1–3). These trials included subjects with metastatic melanoma, renal cell cancer, and head and neck squamous cell cancers, and have resulted in prolonged, durable, complete responses.

In murine models, it has been observed that TDLN harbor lymphoid cells that are functionally capable of mediating rejection of immunogenic tumors in adoptive transfer after *in vitro* activation (4,5). However, both tumor-infiltrating lymphocytes (TIL) and TDLN cells were found to be incapable of mediating the regression of poorly immunogenic tumors such as the B16–BL6 melanoma, a highly invasive tumor of spontaneous origin. It was then discovered that the subcutaneous inoculation of B16–BL6 tumor cells admixed with the bacterial adjuvant, *Corynebacterium parvum*, resulted in reactive TDLN cells that differentiated into therapeutic effector T cells upon activation *in vitro* (6,7). Upon adoptive transfer, these LN cells successfully mediated the regression of established tumors. In addition to the ability to mediate regression of experimentally induced pulmonary metastases, these activated cells were effective in the treatment of spontaneous visceral metastases originating from a primary tumor, a condition that more closely approximates human malignancy. These studies thus demonstrated that vaccination of animals with irradiated tumor cells admixed with a bacterial adjuvant was capable of inducing tumor-reactive T cells in the draining LN.

We have applied these methods to generate vaccine-primed LN in patients with advanced melanoma and renal cell cancer (RCC) for therapy (3,8). Patients with RCC or melanoma received intradermal inoculation of irradiated autologous tumor cells admixed with *Bacillus Calmette–Guerin* (BCG) as a vaccine. Seven to ten days later, draining LN were removed for *in vitro* activation and expansion. Activated LN cells were then administrated intravenously with the concomitant administration of IL-2 for immunotherapy with defined success (3).

Studies demonstrated that tumor cells genetically modified with immunostimulatory genes are capable of sensitizing T cells. Transfection of cytokine genes into murine tumor cells have resulted in reduced tumorigenicity following inoculation of the modified tumor cells into animals (9). In these studies, animals that rejected the inoculum of modified tumor cells also rejected a subsequent challenge of unmodified parental tumor cells, thus demonstrating the development of tumor immunity. We performed a clinical study of patients with melanoma to evaluate the immunobiological effects of GM–CSF transduced autologous tumor cells given as a vaccine to prime draining lymph nodes (10). There was an increased infiltration of dendritic cells (DCs) in the GM–CSF-secreting vaccine sites compared with the wild type (WT) vaccine sites. This resulted in a greater number of cells harvested from the GM–CSF–VPLNs compared with the WT–VPLNs. Patients received adoptively transferred GM–CSF–VPLN cells secondarily activated and expanded *in vitro*. A complete clinical response was observed in one of five patients. This work documented measurable immunobiologic differences of GM–CSF-transduced tumor cells given as a vaccine compared with WT tumor cells.

Collectively, these observations suggested that TDLN or VPLN cells may represent an ideal source of tumor-reactive T cells. They also established the rationale for developing tumor vaccines utilizing autologous tumors admixed with bacterial adjuvant or genetically modified with cytokine genes, which may prove useful in facilitating the generation of immune T cells for adoptive immunotherapy.

ACTIVATION AND POLARIZATION OF EFFECTOR T CELLS *IN VITRO*

A major challenge in T cell immunotherapy of cancer is how to activate and expand the relatively low numbers of tumor-specific T cells obtained from the tumor-bearing host. Previous studies demonstrated that freshly isolated TDLN cells had defects in TCR-mediated signal transduction and were not immediately competent in adoptive transfer models (11,12). It has therefore become a critical prerequisite in adoptive immunotherapy to expand the pre-effector cells into large numbers of effector cells while augmenting their antitumor reactivity.

In vitro T cell activation using monoclonal antibodies in the absence of antigen takes advantage of common signal transduction pathways that are ubiquitous to T cells. This principle has been used to expand tumor-primed T cells contained within TDLN or VPLN. The initial efforts involved the use of anti-CD3 mAb as a surrogate antigen to activate tumor-primed lymphoid cells, followed by expansion in IL-2 (12). This approach resulted primarily in the generation of $CD8^+$ effector cells that mediated tumor regression *in vivo*. Subsequent clinical studies utilizing this method to activate VPLN cells demonstrated that this cellular therapy can result in achieving durable tumor responses in subjects with advanced cancer (1,3). We have extended these investigations in animal models and with human samples by examining other mAbs that deliver costimulatory signals in concert with anti-CD3 to activate tumor-primed lymphoid cells. These other antibodies have involved anti-CD28 and anti-CD137 (13–16). The results of these investigations have indicated that costimulation can increase the proliferation of tumor-primed lymphoid cells and their ability to mediate tumor regression *in vivo*.

Several important principles in animal models that are relevant for the treatment of human malignancies have been identified. For example, the *in vitro* cytokine profiles released by effector T cells when cocultured with tumor cells are found to be predictive of their ability to mediate tumor regression *in vivo*. Effector cells that mediate a type 1 (i.e., IFNγ) and GM–CSF response to tumor antigen are capable of eradicating tumor upon adoptive transfer. In contrast, cells that demonstrate a type 2 profile (i.e., IL-10, IL-4) appear suppressive, and do not mediate tumor regression (16,17). We have determined the importance of IFNγ in mediating tumor regression both in animal studies (16) and in clinical trials (3). In a phase II adoptive cellular trial in patients with advanced renal cell cancer, we demonstrated that IFNγ secretion and the IFNγ: IL-10 ratio of cytokine released by effector T cells in response to tumor antigen was associated with clinical outcomes. Specifically, activated T cells that have an increased IFNγ:IL-10 ratio correlated with tumor response (3). Although effector T cells can be generated through antibody activation to mediate tumor regression in animal models, clinical responses in adoptive immunotherapy have been confined to a minority of patients. One potential reason for these limited responses is that antibody-activation procedures generally stimulate T cells broadly without discriminating between type1 and type 2 cells, presumably due to the polyclonal expansion characteristics of antibodies directed to the TCR common chain, for example, CD3ε of the TCR/CD3 complex or CD28. As a result, both type 1 cytokines, such as IL-2, IFNγ, and type 2 cytokines, for example, IL-4, IL-5, and IL-10, are modulated (13,18). Therefore, alternative protocols need to be defined that will preferentially stimulate the type 1 cytokine profile to generate more potent tumor-reactive T cells for cancer immunotherapy. Toward this end, various *in vitro* strategies have been investigated utilizing additional signaling stimuli to promote Th1/Tc1 cell proliferation and antitumor reactivity (19,20). We reported that costimulation of TDLN cells through newly induced 4-1BB and CD3/CD28 signaling can significantly increase antitumor reactivity by shifting T cell responses toward a type 1 cytokine pattern, while concomitantly decreasing type 2 response (16). Using the proinflammatory cytokines, we recently reported that IL-12 and IL-18 can be used to generate potent $CD4^+$ and $CD8^+$ antitumor effector cells by synergistically polarizing antibody-activated TDLN cells toward a Th1 and Tc1 phenotype, and that the polarization effect was NF-κB dependent (21).

The recognition and use of cell polarization strategies during and /or post antibody activation of T cells represents another significant change and addition to the traditional practice of adoptive therapy. While adoptive immunotherapy of cancer requires large numbers of therapeutic T cells for transfer into cancer patients, the phenotype and cytokine profile of these cells are crucial in determining the

outcomes of the therapy. The use of polarizing reagents to modulate T cell function toward the type 1 response provides a rational strategy to enhance the efficacy of cellular therapy.

USE OF IMMUNE ADJUVANT IN CONCERT WITH T CELL ADMINISTRATION

In the course of adoptive immunotherapy of cancer, administration of T cell growth factors accompanying T cell transfer may promote T cell activation, proliferation, and tumor killing, and therefore augment clinical outcomes for the therapy. These growth factors, as well as other immune modulatory reagents used in concert with T cell transfer, function as immune adjuvants in eliciting antitumor activities *in vivo*. The most useful adjuvant to T cell transfer to date has been the exogenous administration of IL-2 (1–3,22,23).

Nearly 20 years ago, Rosenberg and colleagues performed a pilot protocol to investigate the feasibility and practicality of immunotherapy of patients with advanced cancer using TIL and recombinant IL-2 (22). The study represents an initial attempt to use TIL plus IL-2 administration with enhanced tumoricidal capacity in the adoptive immunotherapy of human malignancies. Twelve patients with melanoma, renal cell carcinoma, breast carcinoma, or colon carcinoma were treated with varying doses and combinations of TIL, IL-2, and cyclophosphamide. Three partial responses (PR) to therapy were observed. No toxic effects were directly attributable to TIL infusions. However, the toxicities of therapy were similar to those ascribed to IL-2. Indeed, the use of IL-2 has resulted in significant morbidity associated with cellular therapies (3,24). Moreover, a few recent studies showed that IL-2 may negatively regulate effector cells through activation-induced cell death (23,25), expanding the frequency of $CD4^+CD25^+$ T cells, or cause cell redistribution secondary to Ag-induced cell death (26,27). These studies suggest that novel reagents need to be identified to serve as alternative immune adjuvants for adoptive T cell therapy.

In a recent study (25), failed adoptive T cell therapy could be reversed with low dose IL-15 administration, but not IL-2. A related T cell growth factor, IL-15, protected T cells against activation-induced cell death and promoted homeostatic maintenance of memory T cells and, therefore, may be advantageous to T cell-based cancer treatment. Similarly, the role of IL-15 in early activation of memory $CD8^+$ CTLs has been described (28). In this study, memory $CD8^+$ T cells expressing OVA-specific TCR were transferred into IL-15-transgenic (Tg) mice, IL-15 knockout (KO) mice, or control C57BL/6 mice followed by challenge with recombinant Listeria monocytogenes expressing OVA (rLM-OVA). *In vivo* CTL activities were significantly higher in the IL-15 Tg mice, but lower in the IL-15 KO mice than those in control mice at the early stage after challenge with rLM-OVA. *In vivo* administration of rIL-15 conferred robust protection against reinfection via activation of the memory $CD8^+$ T cells. In addition, IL-27 is a novel IL-12 family member that plays a role in the early regulation of Th1 initiation and synergizes with IL-12 in IFNγ production (29). Mechanistic studies revealed that although a comparable proliferative response to IL-27 was observed between STAT1-deficient and wild-type $CD4^+$ T cells, synergistic IFNγ production by IL-27 and IL-12 was impaired in STAT1-deficient $CD4^+$ T cells. IL-27 also augmented the expression of MHC class I on $CD4^+$ T cells in a STAT1-dependent manner (29).

Although the *in vivo* administration of proinflammatory cytokines has demonstrated antitumor efficacy, their potent antitumor activity is often achieved at the expense of unacceptable toxicity. For example, IL-12 and IL-18 administration was found to be associated with lethal organ damages, attributed in part to extremely high levels of host-induced IFNγ production (30). It is anticipated that administration of low doses of proinflammatory cytokines in the context of passively transferred TDLN cells will lead to increased therapeutic efficacy. To this end, the adjuvant effect of low dose cytokine administration over a long period of time can be compared with that of a high dose over a short period of time. These experiments should help to determine if prolonged administration of low dose cytokines can enhance the therapeutic efficacy by improving trafficking, survival, and proliferation of the adoptively transferred T cells.

While toxicity of traditionally used IL-2 limits its clinical utility at high doses, use of novel cytokines at tolerable low doses in conjunction with cellular therapy may provide alternative strategies that are less toxic. If the newly identified proinflammatory cytokines, such as IL-15 and IL-27 prove to be useful adjuvants to T cell therapy, they may result in more effective antitumor responses with reduced morbidity.

TRAFFICKING AND PROLIFERATION OF EFFECTOR T CELLS AFTER ADOPTIVE TRANSFER

Adoptive T cell therapy has been used for treatment of viral and malignant diseases with encouraging results. However, little is known about the fate and trafficking of the transferred effector cells. A study performed at NCI assessed the trafficking of gp100-specific pmel-1 cells to large, vascularized tumors that express or do not express the target Ag (31). It was found that approximately equal numbers of pmel-1 T cells infiltrated the Ag-positive and -negative tumors. Massive infiltration and proliferation of activated antitumor pmel-1 cells were observed in a variety of peripheral tissues, including lymph nodes, liver, spleen, and lungs, but not peripheral blood. However, T cell function, as measured by production of IFNγ, release of perforin, and activation of caspase-3 in target cells, was confined to Ag-expressing tumor. It was thus concluded that adoptively transferred $CD8^+$ T cells traffic indiscriminately and ubiquitously while mediating specific tumor destruction.

We recently characterized the infiltration of adoptively transferred TDLN cells in the host bearing pulmonary metastases (21). The TDLN cells were activated with anti-CD3/anti-CD28 followed by cell culture in IL-12 + IL-18 before transfer into tumor-bearing host. The TDLN

cells were labeled with CFSE immediately before infusion. Immunohistochemical evaluation of adoptively transferred TDLN cells accumulating in pulmonary tumor nodules was performed. Infused TDLN cells were observed to (1) attach to venules, (2) mix with host leukocytes in perivenular collections, and (3) infiltrate tumor nodules. Active migration of infused cells into pulmonary tumor nodules was found to be correlated with significant tumor regression. This corroborates a previous report by Plautz et al. showing that infused TDLN cells must infiltrate pulmonary nodules to suppress tumor growth (32).

Several other reports support the hypothesis that efficient tumor regression needs the *in situ* accumulation of transferred effector cells. Another study conducted at the University of Michigan demonstrated that the infused cells must accumulate in metastatic lesions to suppress tumor growth, and that the process is dynamic (33). In studies treating murine lung metastases with adoptively transferred TDLN cells, the TDLN donor cells were initially confined to alveolar capillaries with no movement into metastases after infusion. However, within 4 h, TDLN cells began migrating across pulmonary postcapillary venules and first appeared within metastases. After 24 h, most donor cells in the lung were associated with tumor nodules. Donor cell proliferation both within the lung and in the lymphoid organs was detected. Importantly, T cells that had proliferated in the lymphoid organs trafficked back to the tumor-bearing lungs, accounting for $\sim 50\%$ of the donor cells recovered from these sites. These studies demonstrate that adoptively transferred TDLN cells migrate directly into tumor-bearing organs and seed the recirculating pool of lymphocytes after infusion. Cells that have differentiated in lymphoid organs eventually migrate into the tumor site. Additionally, *in vitro*-generated Melan-A-specific CTLs were found to survive intact *in vivo* for several weeks and localize preferentially to tumor (34). Over all, these studies suggest that methods to improve trafficking and recruitment of donor T cells to the tumor may improve therapeutic efficacy of cellular therapy.

The availability of congenic strains of mice bearing T cell markers that differ by epitopes that can be identified by monoclonal antibodies allows us to track adoptively transferred cells in a semisyngeneic host. In order to perform quantitative tracking studies of the infused cells, the congenic strain of B6 mouse that expresses CD45.1 can be used to generate TDLN for transfer into CD45.2 hosts. Analysis of the infiltrate can be performed by mechanical dissociation of the tumors in order to recover viable lymphoid infiltrates. By FACS analysis, the number of transferred CD4/CD8 T cells can be quantified. Proliferation of infused cells can be assessed by labeling them with CFSE. Confirmed correlation between effective tumor regression and the infiltration of infused cells to tumor should encourage further attempts to modulate T cell trafficking by biochemical controls or by genetic modification of well-identified adhesion molecules, for example, LFA, ICAM, and selectins. Furthermore, a very recent study described the regulation of T cell trafficking by sphingosine 1-phosphate (S1P) receptor 1 (S1P1) (35). Mature T cells from S1P1 transgenic mice exhibited enhanced chemotactic response toward S1P, and preferentially distributed to the blood rather than secondary lymphoid organs, such as draining lymph nodes. This work suggests that S1P1 affects systemic trafficking of peripheral T cells, and therefore makes the S1P/S1P1 signaling pathway a novel target for T cell trafficking modulation.

IDENTIFICATION AND CHARACTERIZATION OF T CELL SUBSETS RESPONSIBLE FOR ANTITUMOR REACTIVITY

The $CD8^+$ CTLs have long been recognized as the effector cells that mediate tumor regression. In addition, $CD4^+$ effector T cells and NK cells have also been identified to directly or indirectly mediate tumor regression. We reported that CD28 costimulation of tumor-primed lymphoid cells promotes the generation of potent tumor-reactive effector cells, particularly $CD4^+$ T cells. These anti-CD3/anti-CD28 activated $CD4^+$ TDLN cells could independently mediate tumor regression in adoptive immunotherapy (13,14,21).

It has to be presumed that any source of antitumor reactive T cells derived from the tumor-bearing host, that is, TDLN, will represent a small percentage of the total population of retrieved cells. Therefore, a theoretically practical approach would be the identification, isolation, activation and expansion of subsets of T cells capable of mediating tumor regression. In this endeavor, Shu and coworkers found that the down-regulation of the homing molecule L-selectin could serve as a surrogate marker for the isolation of specific tumor-sensitized T cells (18). In adoptive immunotherapy of established intracranial MCA 205 tumors, L-selectin^{low} ($CD62L^{low}$) cells displayed at least 30-fold greater therapeutic efficacy than unfractionated cells. The L-selectin^{high} cells did not demonstrate any antitumor effects. These results demonstrate that the purification of L-selectin^{low} cells led to the generation of immune effector cells with unusually high therapeutic efficacy against chemically induced tumors. After that, Plautz et al. used advanced tumor models in a stringent comparison of efficacy for the L-selectin^{low} subset versus the total population of TDLN cells following culture in high dose IL-2. L-selectin^{low} subset comprised 5–7% of the TDLN cells. Adoptive transfer of activated L-selectin^{low} cells eliminated 14-day pulmonary metastases and cured 10-day subcutaneous tumors, whereas transfer of maximally tolerated numbers of unfractionated TDLN cells was not therapeutic (36). At the same time, it was identified that tumor-induced L-selectin^{high} cells were suppressor T cells that mediated potent effector T cell blockade and caused failure of otherwise curative adoptive immunotherapy (37). The treatment failure using unfractionated TDLN cells was due to cotransfer of the L-selectin^{high} suppressor T cells present in TDLN. However, the L-selectin^{high} suppressor T cells were only found in day-12 TDLN. In contrast, day-9 TDLN and normal spleens lacked L-selectin^{high} cells.

It was not long before a second surrogate marker was identified for the isolation of tumor-specific T cells. Stoolman et al. described that tumor-specific responses in TDLN were concentrated in cells expressing P-selectin ligand (Plig^{high} T cells) (38). This study found that the minor subset of TDLN T cells expressing binding sites for the

adhesion receptor P-selectin ($Plig^{high}$ T cells) produced T lymphoblasts with the most tumor-specific IFNγ synthesis *in vitro* and antitumor activity following adoptive transfer *in vivo*. The cultured $Plig^{high}$ TDLN cells were 10- to 20-fold more active against established pulmonary micrometastases than cultured, unfractionated TDLN, and >30-fold more active than cultured TDLN cells depleted of the $Plig^{high}$ fraction. The $Plig^{high}$ T cells expressed high levels of CD69 and low levels of CD62L (L-$selectin^{low}$), which agrees with the previous studies on L-selectin in TDLN. Further supporting these observations is a recent study indicating that recruitment of IFNγ-producing cells into the inflamed retina *in vivo* is preferentially regulated by P-selectin glycoprotein ligand (39).

In a different attempt to selectively activate tumor-sensitized T cells, superantigens were utilized *in vitro* to stimulate effector cell generation in a murine model (40). The TDLN cells stimulated with staphylococcal enterotoxins A (SEA), B (SEB) or C2 (SEC2) resulted in the selective expansion of Vβ3 and 11, Vβ3 and 8, or Vβ8.2 T cells, respectively. Adoptive transfer studies revealed that SEB- and SEC2-, but not SEA- stimulated cells mediated tumor-specific regression. These results suggested that T cells bearing Vβ8 may preferentially respond to the growing tumor than T cells bearing Vβ3 or 11 elements of the T cell receptor. Similarly, stimulating TDLN cells with different anti-Vβ mAbs instead of the pan-T cell reagent anti-CD3 mAb enabled the selective activation of Vβ T cell subsets (17). Enrichment of Vβ subsets of TDLN cells revealed that $Vβ8^+$ cells released high amounts of IFNγ and GM–CSF with minimal amount of IL-10 in response to tumor, and mediated tumor regression *in vivo*. In contrast, enriched population of $Vβ5^+$, $Vβ7^+$, and $Vβ11^+$ cells released low amounts of IFNγ and GM–CSF with high levels of IL-10, and had no *in vivo* antitumor reactivity. *In vitro* depletion of specific Vβ subsets from the whole TDLN pool confirmed that the profile of cytokine released correlated with *in vivo* antitumor function. These studies indicate that functional Vβ subpopulations of effector cells express differential antitumor reactivity, and that selective stimulation of tumor-sensitized T cells is feasible and may represent a more efficient method of generating therapeutic T cells for therapy.

Application of cell subsets for successful T cell therapy should include two approaches: identification of T cell subsets responsible for mediating antitumor reactivity as discussed above, and simultaneously, the elimination of those subsets that are non-reactive or even suppressive. Characterization of regulatory $CD4^+CD25^+$ T cell subpopulation in terms of their potential suppressive effects on anticancer effector cells would warrant further investigations in this area. A current study showed that $CD8^+$ T cell immunity against a tumor self-antigen is augmented by $CD4^+$ T helper cells, but hindered by naturally occurring $CD4^+CD25^+$ T regulatory cells (Treg cells)(41). Adoptive transfer of tumor-reactive $CD8^+$ T cells plus $CD4^+CD25^-$ Th cells into CD4-deficient hosts induced autoimmunity and regression of established melanoma. However, transfer of $CD4^+$ T cells that contained a mixture of $CD4^+CD25^-$ and $CD4^+CD25^+$ Treg cells or Treg cells alone prevented effective adoptive immunotherapy. These findings thus suggest that adoptive immunotherapy requires the absence of naturally occurring $CD4^+CD25^+$ Treg cells to be effective, and the optimal composition of a cellular agent should be composed of $CD8^+$ cells plus $CD4^+CD25^-$ cells.

ADOPTIVE T CELL IMMUNOTHERAPY OF CANCER IN LYMPHOPENIC HOST

Studies in the late 1970s demonstrated that the induction of lymphopenia by sublethal total body irradiation can be beneficial for the treatment of tumors in mice (42). Chang et al. reported that the adoptive transfer of immune cells in the irradiated host confers improved therapeutic effects compared to the normal host (43). The role of lymphodepletion on the efficacy of T cell therapy is incompletely understood and may depend on the destruction of $CD4^+CD25^+$ regulatory cells, interruption of homeostatic T cell regulation, or abrogation of other normal tolerogenic mechanisms. A report by Dummer et al. indicated that the reconstitution of the lymphopenic, sublethally irradiated murine host with syngeneic T cells triggered an antitumor autoimmune response that required expansion within lymph nodes (44).

There are several different animal models of lymphopenia that can be utilized. These include the use of whole body irradiation (WBI), chemotherapy-induced, or genetically altered hosts (i.e., RAG1 knockout mice) that are deficient of T and B cells. The use of various chemotherapeutic agents to induce lymphopenia would simulate the clinical setting. Cyclophosphamide (CTX) is an agent that has been extensively used in murine models and is actively used in the therapy of certain human cancers. It has been described to eliminate tumor-induced suppressor cells in both animal and human settings.

A few years ago, a report described a phase I study of the adoptive transfer of cloned melanoma antigen-specific T lymphocytes for therapy of patients with advanced melanoma (45). Clones were derived from peripheral blood lymphocytes or TILs of patients. Twelve patients received two cycles of cells. Peripheral blood samples were analyzed for persistence of transferred cells by TCR-specific PCR. Transferred cells reached a maximum level at 1 h after transfer, but rapidly declined to undetectable levels by 2 weeks. The lack of clinical effectiveness of this protocol suggested that transfer of different or additional cell types, or that modulation of the recipient host environment was required for successful therapy. Relevant to these studies is the clinical experience reported by Rosenberg and coworkers who infused tumor-reactive T cells in melanoma patients after a nonmyeloablative conditioning regimen (cyclophosphamide/fludarabine) (46). Conditioning with the nonmyeloablative chemotherapy before adoptive transfer of activated tumor-reactive T cells enhanced tumor regression and increased the overall rates of objective clinical responses. Six of thirteen patients demonstrated significant clinical responses as well as autoimmune melanocyte destruction. In a follow up of this experience in 25 patients, the conditioning regimen was given prior to adoptive T cell therapy as before (47). Examination of the T cell persistence through analysis of the specific TCR demonstrated

that there was a significant correlation between tumor regression and the degree of persistence in peripheral blood of adoptively transferred T cell clones. Transferred cells persisted for as long as 2 months in the lymphopenic setting induced by the conditioning regimen. In contrast, they presented in the blood for only 2 or 3 weeks without the prior chemotherapy. These series of studies strongly suggest that the lymphopenic host induced by the nonmyeloablative conditioning regimen may provide a better environment for the functioning of the transferred T cells, and hence improve their therapeutic efficacy. Examination of the mechanisms involved in the reconstitution of the lymphodepleted host after adoptive T cell transfer will be important in identifying methods to improve the efficacy of T cell therapies for cancer.

REDIRECT EFFECTOR T CELLS TO TUMOR

As mentioned earlier, the low precursor frequency of tumor-specific T cells in patients hampers routine isolation of these cells for adoptive transfer. To overcome this problem, "targeted adoptive immunotherapy" or "genetic adoptive immunotherapy" has become an attractive option for cancer treatment. This strategy can be approached in two ways: introduction of a chimeric TCR into effector cells; or introduction of a tumor-specific TCR into naïve cells.

The T-body approach uses patient-derived lymphocytes transfected with chimeric receptor genes constructed with the variable domains of monoclonal antibodies or cytokines linked to the constant regions of TCR. The rationale for this novel approach to redirect effector cells combines the effector functions of T lymphocytes with the ability of antibodies or cytokines to recognize predefined surface antigens or cytokine receptors with high specificity and in a non-MHC restricted manner.

Eshhar et al. (48) was one of the first to describe this approach by developing a chimeric receptor gene which recognized trinitrophenyl (TNP). Retroviral transduction of the anti-TNP/TCR chimeric gene into a T cell hybridoma line resulted in gene expression. These gene modified T cells were cytolytic and released IL-2 in response to TNP-labeled Daudi cells, but not unlabeled cells. Also among the pioneers in this area, Hwu et al. (49) developed a recombinant chimeric receptor against an epitope expressed on the majority of ovarian cancer cell lines. The TIL were transduced with this chimeric gene and evaluated for immunologic function. The gene modified TIL showed specific lysis of an ovarian carcinoma cell line, but not nonovarian cell lines. In a direct comparison, the gene modified TIL showed greater therapeutic efficacy *in vivo* than the nontransduced TIL (49). Pinthus et al. evaluated the therapeutic efficacy of anti-erbB2 chimeric receptor-bearing human lymphocytes on human prostate cancer xenografts in a SCID mouse model (50). Local delivery of erbB2-specific transgenic T cells to well-established subcutaneous and orthotopic tumors resulted in retardation of tumor growth and prolongation of animal survival. In a setting of metastatic cancer (51), anti-erbB2 chimeric receptor-modified T cells killed breast cancer cells and secreted IFNγ in an Ag-specific manner *in vitro*. Treatment of established metastatic disease in lung and liver with these genetically engineered T cells resulted in dramatic increases in survival of the xenografted mice. In another report, $CD4^+$ cells isolated from the peripheral blood and engrafted with a recombinant immunoreceptor specific for carcinoembryonic Ag (CEA) efficiently lysed target cells in a MHC-independent fashion, and the efficiency was similar to that of grafted $CD8^+$ T cells (52). In an attempt to further improve the therapeutic utility of redirected T cells, T lymphocytes were transferred with CEA-reactive chimeric receptors that incorporate both CD28 and TCR-zeta signaling domains. T cells expressing the single-chain variable fragment of Ig (scFv)-CD28-zeta chimera demonstrated a far greater capacity to control the growth of CEA^+ xenogeneic and syngeneic colon carcinomas *in vivo* compared with scFv-CD28 or scFv-zeta transfected T cells. This study has illustrated the ability of a chimeric scFv receptor capable of harnessing the signaling machinery of both TCR-zeta and CD28 to augment T cell immunity against tumors (53).

In addition to antibodies, cytokines could also be used to reconstruct chimeric TCRs. The IL-13 receptor alpha2 (IL-13Rα2) is a glioma-restricted cell-surface epitope not otherwise detected within the central nervous system. Kahlon et al. (54) described a novel approach for targeting glioblastoma multiforme (GBM) with IL-13Rα2-specific CTLs. The chimeric TCR incorporates IL-13 for selective binding to IL-13Rα2. This represents a new class of chimeric immunoreceptors that signal through an engineered immune synapse composed of membrane-tethered cytokine (IL-13) bound to cell-surface cytokine receptors (IL-13Rα2) on tumors. Human IL-13-redirected $CD8^+$ CTL transfectants display IL-13Rα2-specific antitumor effector function including tumor cell cytolysis and cytokine production. *In vivo*, the adoptive transfer of genetically modified CTL clones resulted in the regression of established human glioblastoma orthotopic xenografts.

The second genetic approach to redirect T cells involves the introduction of tumor-specific TCRs into naïve cells. Genes encoding tumor antigen-specific TCRs can be introduced into primary human T cells as a potential method of providing patients with a source of autologous tumor-reactive T cells. Several tumor-associated antigens have been identified and cloned from human tumors, such as melanoma, breast cancers, and RCC. The antigens have been identified by their ability to induce T cell reactivity by their binding to the TCR $\alpha\beta$ complex. The subsequent cloning of functional TCR genes capable of recognizing tumor-associated antigens offers a potential opportunity to genetically modify naive cells that have not been previously exposed to tumor antigen and to become competent in recognizing tumor. Cole et al. (55) transfected the cDNA for the TCR α and β chains of an HLA-A2 restricted, melanoma-reactive T cell clone into the human Jurkat T cell line. The transfected line was able to mediate recognition of the melanoma antigen, MART-1, when presented by antigen-presenting cells. This represented the first report of a naive cellular construct designed to mediate functional tumor antigen recognition. A recent study explored the simultaneous generation of $CD8^+$ and $CD4^+$ melanoma-reactive T cells by retroviral-mediated transfer of a TCR specific for HLA-A2-restricted epitope of the melanoma antigen tyrosinase

(56). The TCR-transduced normal human peripheral blood lymphocytes secreted various cytokines when stimulated with tyrosinase peptide-loaded antigen-presenting cells or melanoma cells in an HLA-A2-restricted manner. Rosenberg and co-worker (57) isolated the α and β chains of the TCR from a highly avid anti-gp100 CTL clone and constructed retroviral vectors to mediate gene transfer into primary human lymphocytes. The biological activity of transduced cells was confirmed by cytokine production following coculture with stimulator cells pulsed with gp100 peptides, but not with unrelated peptides. The ability of the TCR gene to transfer Ag recognition to engineered lymphocytes was confirmed by HLA class I-restricted recognition and lysis of melanoma tumor cell lines. In addition, nonmelanoma-reactive TIL cultures developed antimelanoma activity following anti-gp100 TCR gene transfer. Together, these studies suggest that lymphocytes genetically engineered to express melanoma antigen-specific TCRs may be of value in the adoptive immunotherapy of patients with melanoma.

The HPV16 (human papilloma virus type 16) infection of the genital tract is associated with the development of cervical cancer in women. The HPV16-derived oncoprotein E7 is expressed constitutively in these lesions and represents an attractive candidate for T cell mediated adoptive immunotherapy. In a recent study, Scholten et al. reported that HPV16E7 TCR gene transfer is feasible as an alternative strategy to generate human HPV16E7-specific T cells for the treatment of patients suffering from cervical cancer and other HPV16-induced malignancies (58). These TCR genes specific for HPV16E7 were isolated and transferred into peripheral blood-derived $CD8^+$ T cells. Biological activity of the transgenic CTL clones was confirmed by lytic activity and IFNγ secretion upon antigen-specific stimulation. Most importantly, the endogenously processed and HLA-A2 presented HPV16E7 CTL epitope was recognized by the TCR-transgenic T cells. In a separate study, ovalbumin (OVA)-specific $CD4^+$ cells were successfully generated. Chamoto et al. (59) prepared mouse antigen-specific Th1 cells from nonspecifically activated T cells after retroviral transfer of TCR genes. These Th1 cells transduced with the α and β genes of the I-A (d)-restricted OVA-specific TCR produced IFNγ in response to stimulation with OVA peptides or A20 B lymphoma cells expressing OVA as a model tumor antigen. The TCR-transduced Th1 cells also exhibited cytotoxicity against tumor cells in an antigen-specific manner. In addition, adoptive transfer of TCR-transduced Th1 cells exhibited potent antitumor activity *in vivo*.

Genetic alteration of T cells with chimeric receptor genes or antigen-specific TCR genes confers the redirection of effector cells to the tumor for its destruction. These approaches may offer novel opportunities to develop immunocompetent effector cellular reagents and improve the efficacy of adoptive immunotherapy of cancer.

COMBINED THERAPY

Cancer is a disease that involves multiple gene malfunctions and numerous biochemical and cellular event errors during its development and metastasis within an individual. Therefore, it is difficult to achieve success utilizing adoptive T cell transfer as a monotherapy. The above-reviewed use of vaccination to induce tumor-reactive pre-effector *in vivo*; the coadministration of immune adjuvant with T cell transfer; and the gene therapy to redirect T cells to tumor are all among the strategies taken to elicit and/or strengthen the efficacy of T cell therapy. Combination therapy is a very common practice during the treatment of diseases. Active vaccine therapy, for example, can be used in concert with chemotherapy, radiotherapy, or antibody therapy. Combining a glioma tumor vaccine engineered to express the membrane form of macrophage colony-stimulating factor with a systemic antiangiogenic drug-based therapy cured rats bearing 7 day old intracranial gliomas (60). We successfully demonstrated that local radiotherapy potentiates the therapeutic efficacy of intratumoral dendritic cell (DC) administration (61), and that anti-CD137 monoclonal antibody administration augments the antitumor efficacy of DC-based vaccines (62).

In order to enhance the efficiency of T cell therapy, various strategies have been employed accompanying cell transfer. These combined therapies include cell transfer in combination with intratumoral expression of lymphotactin (63), DC vaccination (64), or blockade of certain molecules expressed in tumor cells, such as B7-H1 (65).

One of the major obstacles to successful adoptive T cell therapy is the lack of efficient T cell infiltration of tumor. Combined intratumoral lymphotactin (Lptn) gene transfer into SP2/0 myeloma tumors and adoptive immunotherapy with tumor specific T cells eradicated well-established SP2/0 tumors in six of eight mice, and dramatically slowed down tumor growth in the other two mice (63). Cell tracking using labeled T cells revealed that T cells infiltrated better into the Lptn-expressing tumors than non-Lptn-expressing ones. These data provide solid evidence of a potent synergy between adoptive T cell therapy and Lptn gene therapy as a result of facilitated T cell targeting. Dendritic cells are well-known potent antigen-presenting cells. Hwu and co-workers (64) reported that DC vaccination could improve the efficacy of adoptively transferred T cells to induce an enhanced antitumor immune response. Mice bearing B16 melanoma tumors expressing the gp100 tumor antigen were treated with activated T cells transgenic for a TCR specifically recognizing gp100, with or without concurrent peptide-pulsed DC vaccination. Antigen-specific DC vaccination induced cytokine production, enhanced cell proliferation, and increased tumor infiltration of adoptively transferred T cells. The combination of DC vaccination and adoptive T cell transfer led to a more robust antitumor response than the use of each treatment individually. This work shows that in addition to their ability to initiate cell-mediated immune responses by stimulating naive T cells, dendritic cells can strongly boost the antitumor activity of activated T cells *in vivo* during adoptive immunotherapy. Certain cell surface molecules, expressed either on tumor cells or on T cells, have demonstrated have demonstrated potential suppressive impact on the adoptive T cell immunotherapy. For example, during the last few years, new members of the B7 family molecules have been identified, for example, B7-H1, which

is constitutively expressed on 66% of freshly isolated squamous cell carcinomas of the head and neck (SCCHN) (65). When B7-H1-negative mouse SCC line, SCCVII, was transfected to express B7-H1, all of the animals succumbed to B7-H1/SCCVII tumors even after adoptive T cell immunotherapy. However, the infusion of B7-H1 blocking monoclonal antibody with activated T cells cured 60% of animals. The data support B7-H1 blockade as a new approach to enhance the efficacy of T cell immunotherapy. These findings also illuminate a new potential application for the blockade of certain "negative costimulation molecules" on T cells, for example, CTLA-4 and programmed death-1 (PD-1) molecules. This kind of blocking may augment the therapeutic efficacy mediated by the transferred T cells. The blockade can be done using specific monoclonal antibodies, soluble ligands for CTLA-4 or PD-1, or by synthesized antagonists. In addition, effector cells can be derived from the animals deficient in the relevant molecules for preclinical investigations.

Immune tolerance of tumor-bearing host represents another major obstacle for the successful use of adoptive T cell immunotherapy. A recent study examined the requirement for assistance to the low affinity tumor-specific $CD8^+$ T cells transferred into tumor-bearing mice (66). The TCR transgenic mice expressing a class I-restricted hemagglutinin (HA)-specific TCR (clone 1 TCR) were generated. Upon transfer into recipient mice in which HA is expressed at high concentrations as a tumor-associated Ag, the clone 1 TCR $CD8^+$ T cells exhibited very weak effector function and were soon tolerized. However, when HA-specific $CD4^+$ helper cells were co-transferred with clone 1 cells and the recipients were vaccinated with influenza, clone 1 cells were found to exert a significant level of effector function and delayed tumor growth. This work shows that in order to optimize the function of low avidity tumor-specific T cells after adoptive transfer, additional measures need to be taken to help break the host tolerance.

Effective tumor therapy requires a proinflammatory microenvironment that permits T cells to extravasate and to destroy the tumor. Proinflammatory environment can be induced by various chemical, physical, and immunological protocols. Greater extent of success can be expected by combining adoptive T cell therapy with the traditional cancer treatment methods, for example, surgery, chemotherapy, and radiation therapy, as well as with different forms of immunotherapeutic strategies, such as vaccine, antibody, cytokines, gene therapy, and so on. The factors to be combined can involve two or more approaches.

In summary, adoptive immunotherapy utilizing tumor-reactive T cells offers a promising alternative approach for the management of cancer. Through the endeavors of clinical and basic research scientists during the last two decades, the process of adoptive T cell therapy of cancer has evolved from its original single-step approach into its current multiple-step procedure. Successful T cell immunotherapy of cancer is the outcome of this multi-step process that depends on successful Ag priming, numerical amplification of low frequency Ag-specific precursors, use of immune adjuvants, and efficient infiltration of tumors in all metastatic sites by effector T cells. New directions in this field include the identification and application of tumor-reactive subpopulation of T cells, creation of a lymphopenic environment in the recipient host, and the redirection of the effector cells toward the tumor. Development of these latter techniques and the combined use of different therapeutic strategies may further improve the efficacy of the immunotherapy of human cancer employing adoptive T cell transfer. Studies and developments of immunotherapy for cancer should accelerate the application of this strategy in infectious disease, autoimmune disease and other disease managements.

BIBLIOGRAPHY

1. Chang AE, et al. Adoptive Immunotherapy with vaccine-primed lymph node cells secondarily activated with anti-CD3 and interleukin-2. J Clin Oncol 1997;15:796.
2. Chang AE., et al. Generation of vaccine-primed lymphocytes for the treatment of head and neck cancer. Head and Neck 2003;25:198.
3. Chang AE, et al. Phase II trial of autologous tumor vaccination, Anti-CD3-activated vaccine-primed lymphocytes, and Interleukin-2 in stage IV renal cell cancer. J Clin Oncol 2003;21:884.
4. Shu S, Chou T, Rosenberg SA. Generation from tumor-bearing mice of lymphoid cells with in vivo therapeutic efficacy. J Immunol 1987;139:295–304.
5. Chou T, Chang AE, Shu S. Generation of therapeutic T lymphocytes from tumor-bearing mice by in vitro sensitization: Culture requirements and characterization of immunologic specificity. J Immunol 1988;140:2453–2461.
6. Geiger J, Wagner P, Shu S, Chang AE. A novel role for autologous tumor cell vaccination in the immunotherapy of the poorly immunogenic B16-BL6 melanoma. Surg Oncol 1992;1:199–208.
7. Geiger J, et al. Generation of T- cells reactive to the poorly immunogenic B16-BL6 melanoma with efficacy in the treatment of spontaneous metastases. J Immunother 1993;13:153–65.
8. Li Q, et al. Immunological effects of BCG as an adjuvant in autologous tumor vaccines. Clin Immunol 2000;94:64–72.
9. Restifo N, et al. A nonimmunogenic sarcoma induced with the cDNA for interferon-γ elicits $CD8^+$ T cells against the wild-type tumor: Correlation with antigen presentation capability. J Exp Med 1992;175:1423–1431.
10. Chang AE, et al. Immunogenetic therapy of human melanoma utilizing autologous tumor cells transduced to secrete GM–CSF. Hum Gene Ther 2000;11:839–850.
11. Liu J, et al. Ex vivo activation of tumor-draining lymph node T cells reverses defects in signal transduction molecules. Can Immunol Immunother 1998;46:268.
12. Yoshizawa H, Chang AE, Shu S. Specific adoptive immunotherapy mediated by tumor-draining lymph node cells sequentially activated with anti-CD3 and IL-2. J Immunol 1991;147:729–37.
13. Li Q, Furman SA, Bradford CR, Chang AE. Expanded tumor-reactive $CD4^+$ T-Cell responses to human cancers induced by secondary anti-CD3/anti- CD28 activation. Clin. Can. Res. 1999;5:461.
14. Li Q, et al. Therapeutic effects of tumor-reactive $CD4^+$ cells generated from tumor-primed lymph nodes using anti-CD3/anti-CD28 monoclonal antibodies. J Immunother 2002;25:304.
15. Ito F, et al. Antitumor therapy reactivity of Anti-CD3/Anti-CD28 bead-activated lymphoid cells: Implications for cell in a murine model. J Immunother 2003;26:222.

16. Li Q, et al. Polarization effects of 4-1BB during CD28 costimulation in generating tumor-reactive T Cells for cancer immunotherapy. Can. Res. 2003;63:2546.
17. Aruga A, et al. Type 1 versus type 2 cytokine release by Vβ T cell subpopulations determines in vivo antitumor reactivity: IL-10 mediates a suppressive role. J Immunol 1997;159:664–673.
18. Kagamu H, Shu S. Purification of L-selectin^{low} cells promotes the generation of highly potent CD4 antitumor effector T lymphocytes. J Immunol; 1998;160:3444–3452.
19. Hart-Meyers J, et al. Cutting Edge: CD94/NKG2 is expressed on Th1 but not Th2 cells and costimulates Th1 effector functions. J Immunol 2002;169:5382–5386.
20. Hou W, et al. Pertussis toxin enhances Th1 responses by stimulation of dendritic cells. J Immunol 2003;170:1728–1736.
21. Li Q, et al. Synergistic effects of IL-12 and IL-18 in skewing tumor-reactive T-cell responses towards a type 1 pattern. Can. Res. 2005;65:1063.
22. Topalian SL. et al. Immunotherapy of patients with advanced cancer using tumor-infiltrating lymphocytes and recombinant interleukin-2: a pilot study. J. Clin. Oncol. 1988;6:839.
23. Shrikant P, Mescher MF. Opposing effects of IL-2 in tumor immunotherapy: promoting CD8 T cell growth and inducing apoptosis. J. Immunol. 2002;169:1753.
24. Rosenberg SA, et al. Experience with the use of high-dose IL-2 in the treatment of 652 cancer patients. Ann Surg 1989;210:474.
25. Roychowdhury S, et al. Failed adoptive immunotherapy with tumor-specific T cells: Reversal with low-dose interleukin 15 but not low-dose interleukin 2. Can Res 2004;64:8062.
26. Nacsa J, et al. Contrasting effects of low-dose IL-2 on vaccine-boosted Simian Immunodeficiency Virus (SIV)-specific CD4$^+$ and CD8$^+$ T cells in macaques chronically infected in SIV-mac251. J Immunol 2005;174:1913.
27. Poggi A, et al. Tumor-induced apoptosis of human IL-2-activated NK cells: role of natural cytotoxicity receptors. J Immunol 2005;174:2653.
28. Yajima T, et al. A novel role of IL-15 in early activation of memory CD8$^+$ CTL after reinfection. J Immunol 2005;174: 3590–3597.
29. Kamiya S, et al. An indispensable role for STAT1 in IL-27-induced T-bet expression but not proliferation of naïve CD4$^+$ T cells. J Immunol 2004;173:3871–3877.
30. Osaki T, et al. IFN-γ-inducing factor/IL-18 administration mediates IFN-γ-and IL-12-independent antitumor effects. J Immunol 1998;160:1742.
31. Palmer DC, et al. Vaccine-stimulated adoptively transferred CD8$^+$ T cells traffic indiscriminately and ubiquitously while mediating specific tumor destruction. J Immunol 2004;173: 7209–7216.
32. Mukai S, Kjaergaard J, Shu S, Plautz GE. Infiltration of tumors by systemically transferred tumor-reactive T lymphocytes is required for antitumor efficacy. Can Res 59: 5245.
33. Skitzki J, et al. Donor cell cycling, trafficking, and accumulation during adoptive immunotherapy for murine lung metastases. Can Res 2004;64:2183.
34. Meidenbauer N, et al. Survival and tumor localization of adoptively transferred melan-A-specific T cells in melanoma patients. J Immunol 2003;170:2161–2169.
35. Chi H, Flavell RA. Cutting Edge: Regulation of T cell trafficking and primary immune responses by sphingosine 1-phosphate receptor 1. J Immunol 2005;174:2485–2488.
36. Wang LX, Chen BG, Plautz GE. Adoptive immunotherapy of advanced tumors with CD62 L-selectin^{low} tumor-sensitized T lymphocytes following ex vivo hyperexpansion. J Immunol 2002;169:3314–3320.
37. Peng L, et al. Tumor-induced L-selectinhigh suppressor T cells mediate potent effector T cell blockade and cause failure of otherwise curative adoptive immunotherapy. J Immunol 2002;169:4811–4821.
38. Tanigawa K, et al. Tumor-specific responses in lymph nodes draining murine sarcomas are concentrated in cells expressing P-selectin binding sites. J Immunol 2001;167:3089–3098.
39. Xu H, et al. Recruitment of IFN-γ-producing (Th1-like) cells into the inflamed retina in vivo is preferentially regulated by P-selectin glycoprotein ligand 1:P/E-selectin interactions. J Immunol 2004;172:3215–3224.
40. Shu S, et al. Stimulation of tumor-draining lymph node cells with superantigenic staphyloccocal toxins leads to the generation of tumor-specific effector cells. J Immunol 1994;152:1277–1288.
41. Antony PA, et al. CD8$^+$ T cell immunity against a tumor/self-antigen is augmented by CD4$^+$ T helper cells and hindered by naturally occurring T regulatory cells. J Immunol 2005;174: 2591–2601.
42. Mule JJ, Jones FR, Hellstrom I, Hellstrom KE. Selective localization of radio-labeled immune lymphocytes into syngeneic tumors. J Immunol 1979;123:600.
43. Chang AE, et al. Differences in the effects of host suppression on the adoptive immunotherapy of subcutaneous and visceral tumors. Can Res 1986;46:3426.
44. Dummer W, et al. T cell homeostatic proliferation elicits effective antitumor autoimmunity. J Clin Inves 2002;110:185.
45. Dudley ME, et al. Adoptive transfer of cloned melanoma-reactive T lymphocytes for the treatment of patients with metastatic melanoma. J Immunoth 2001;24:363–373.
46. Dudley ME, et al. Cancer regression and autoimmunity in patients after clonal repopulation with antitumor lymphocytes. Science 2002;298:850.
47. Robbins PF, et al. Cutting Edge: Persistence of transferred lymphocyte clonotypes correlates with cancer regression in patients receiving cell transfer therapy. J Immunol 2004;173: 7125–7130.
48. Eshhar Z, Waks T, Schindler DG, Gross G. Specific activation and targeting of cytotoxic lymphocytes through chimeric single chains consisting of antibody binding domains and the g or z subunits of the immunoglobulin and T cell receptors. Proc Natl Acad Sci USA 1993;90:720–724.
49. Hwu P, et al. In vivo antitumor activity of T cells redirected with chimeric antibody/T cell receptor genes. Can Res 1995;55:3369–3373.
50. Pinthus JH, et al. Immuno-gene therapy of established prostate tumors using chimeric receptor-redirected human lymphocytes. Can Res 2003;63:2470–2476.
51. Kershaw MH, et al. Gene-engineered T cells as a superior adjuvant therapy for metastatic cancer. J Immunol 2004;173: 2143–2150.
52. Hombach A, et al. CD4$^+$ T cells engrafted with a recombinant immunoreceptor efficiently lyse target cells in a MHC antigen-and-fas independent fashion. J Immunol 2001;167:1090–1096.
53. Haynes NM, et al. Rejection of syngeneic colon carcinoma by CTLs expressing single-chain antibody receptors codelivering CD28 costimulation. J Immunol 2002;169:5780–5786.
54. Kahlon KS, et al. Specific recognition and killing of glioblastoma multiforme by interleukin 13-zetakine redirected cytolytic T cells. Can Res 2004;64:9160–9166.
55. Cole DJ, et al. Characterization of the functional specificity of a cloned T cell receptor heterodimer recognizing the MART-1 melanoma antigen. Can Res 1995;55:748–752.
56. Roszkowski JJ, et al. Simultaneous generation of CD8$^+$ and CD4$^+$ melanoma-Reactive T cells by retroviral-mediated transfer of a single T-cell receptor. Can Res 2005;65:1570–1576.

57. Morgan RA, et al. High efficiency TCR gene transfer into primary human lymphocytes affords avid recognition of melanoma tumor antigen glycoprotein 100 and does not alter the recognition of autologous melanoma antigens. J Immunol 2003;171:3287–3295.
58. Scholten KB, et al. Preservation and redirection of HPV16E7-specific T cell receptors for immunotherapy of cervical cancer. Clin Immunol 2005;114:119–129.
59. Chamoto K, et al. Potentiation of tumor eradication by adoptive immunotherapy with T-cellreceptor gene-transduced T-helper type 1 cells. Can Res 2004;64:386–390.
60. Jeffes EWB, et al. Antiangiogenic drugs synergize with a membrane macrophage colony-stimulating factory-based tumor vaccine to therapeutically treat rats with an established malignant intracranial glioma. J Immunol 2005;174:2533–2543.
61. Teitz-Tennenbaum S. et al. Radiotherapy potentiates the therapeutic efficacy of intratumoral dendritic cell administration. Can Res 2003;63:8466–8475.
62. Ito F, et al. Anti-CD137 monoclonal antibody administration augments the antitumor efficacy of dendritic cell-based vaccines. Can Res 2004;64:8411.
63. Huang H, Li F, Gordon JR, Xiang J. Synergistic enhancement of antitumor immunity with adoptively transferred tumor-specific $CD4^+$ and $CD8^+$ T cells and intratumoral lymphotactin transgene expression. Can Res 2002;62:2043.
64. Lou Y, et al. Dendritic cells strongly boost the antitumor activity of adoptively transferred T cells in vivo. Can Res 2004;64:6783.
65. Strome SE. B7-H1 blockade augments adoptive T-cell immunotherapy for squamous cell carcinoma. Can Res 2003;63:6501.
66. Lyman MA, et al. The fate of low affinity tumor-specific $CD8^+$ T cells in tumor-bearing mice. J Immunol 2005;174:2563–2572.

See also BORON NEUTRON CAPTURE THERAPY; MONOCLONAL ANTIBODIES.

IMPEDANCE PLETHYSMOGRAPHY

HELMUT HUTTEN
University of Technology
Graz, Australia

INTRODUCTION

Plethysmography is a volumetric method, that is, a method for the assessment of a volume (the Greek words *plethys* and *plethora* mean full and fullness, respectively). Impedance plethysmography is based on the measurement of passive electrical properties of biological tissues. Those passive electrical properties are parameters of the so-called bioimpedance. The first publication about impedance plethysmography by Nyboer et al. (1) dates back to 1943. Pioneering contributions to the basic understanding of the relations between the assessment of volumes by impedance plethymosgraphy and the electrical properties of biological tissue have been provided by Schwan et al. (2) already in 1955. But already by the end of the nineteenth century Stewart had used the recording of electrical conductivity to study transit times between different sites of the body after injection of saline into the circulation (3). Blood flow recording is one of the most relevant fields for the clinical application of impedance plethysmography nowadays.

Impedance plethysmography is a volumetric method that aims to assess a volume or changes of a volume. Usually, a volume is the filling volume of a space that is enclosed by geometric boundaries. In this case, volumetry means the determination of the boundaries with subsequent assessment of the volume within the boundaries. Those boundaries can be determined by the impedance method if the electrical properties of the substances on both sides of the boundaries are different.

Impedance plethysmography, however, can also be applied to the assessment of volumes that are not lumped compartments within geometric boundaries, for example, it can be used for the volumetric measurement of a certain component within a mixture. Such components may be cells (e.g., the volume of cells in blood), tissues (e.g., the volume of fat tissue in the body), spaces with different composition (e.g., intra- and extracellular spaces), or the volume of the air that is enclosed in the alveoli of lung tissue. In that case, volumetry means the estimation of the space that would be occupied by the respective component if it would be concentrated in one single lumped compartment. Usually, this volume is estimated as a percentage of the whole distribution volume, for example, the volume of cells in blood or the content of water in the whole body. The electrical properties of the respective component must be different from those of all other components. The volumetric assessment does not require a homogeneous distribution of the considered component within the given space if the actual distribution can be taken into account, for example, by a model. Under certain conditions, a tissue can be identified by specific features like morphological structure and/or chemical composition if those features are related with its electrical properties.

The typical application of plethysmography in clinical routine is the diagnosis of those diseases for which the measurement of volumes or changes of volume renders possible the interpretation of functional disorders or functional parameters. The most widely and routinely applied diagnostic examinations are concerned with:

1. Heart: Cardiac mechanical disorders by impedance cardiography (i.e., pumping insufficiency by measuring cardiac stroke volume, including heart rate and other cardiac parameters like ejection period). This application is discussed in another article.
2. Peripheral circulation: Vascular disorders by impedance rheography (i.e., deep venous thrombosis by impedance phlebography, and estimation of blood flow in the brain or other peripheral vessels).
3. Lung: Ventilatory disorders by impedance pneumography (i.e., insufficient ventilation by monitoring the tidal volume and/or respiratory rate).

METHODOLOGY

Impedance plethysmography is a noninvasive method that employs contacting, usually disposable electrodes, in most cases metal-gel electrodes, for example, with Ag/AgCl for

the metal plate. Usually, the electrodes have circular geometry; however, other geometries might be preferable, for example, band-like geometry for segmental measurement at the extremities. It must be considered that the metal plates of electrodes are areas with the same potential, and therefore may affect the electromagnetic field distribution in the considered object. Electrodes with small areas help to reduce that effect, whereas electrodes with large areas render it possible to reach a more homogenous current field in the measured object.

Contacting electrodes can easily be attached to the skin or surface of the measurement object, usually by an adhesive material that is already fixed to the electrode. Only in special cases, for example, for research purposes, does the measurement require invasive application.

Different contactless measurement modes gain increasing attention, for example,

1. Microwave-based methods with antennas as applicators and measurement of the scattered electromagnetic field.
2. Methods based on exploiting the magnetic instead of the electrical properties: inductive plethysmography that uses coils and records the changes in the inductance and magnetic susceptibility plethysmography that employs strong magnetic fields and records the changes in the magnetic flux, for example, by superconducting quantum interference devices (SQUID).

All bioimpedance-based methods are aiming at recording either the effect on the applied electromagnetic field by the measurement object or the response of the measurement object to the application of the electromagnetic field. The directly measured quantities are electrical quantities, for example, voltages or currents. Those quantities are actually imaging electrical coefficients, for example, conductivity, permittivity, or resistivity, which are material-specific parameters of the tissue impedance and monitor its morphological structure and/or chemical composition. With these material-specific parameters the geometric boundaries are determined and used for the estimation of the volume or volume changes. Relations between the measured electrical parameters and the volume are usually based on models. The employed models can be very simple, for example, described by simple geometric boundaries like cylinders, spheres, or ellipsoids. More complex 3D models may be described by the finite element method (FEM) or similar approaches, which allows simulating the distribution of the electromagnetic field in the measured object, that is, the current pathways and the iso-potential planes. Sophisticated iterative optimization procedures are employed to match the simulated values with the measured ones (4).

Electrical impedance tomography (EIT) is a direct approach for determining the 3D geometry of those compartments with the same material-specific parameters in a biological object like the human torso or an extremity. This technique uses multiple electrodes, usually arranged in a plane. Mapping (or imaging) of the impedance distribution in the examined cross-sectional layer requires the solution of the inverse or back-projection problem. Actually, the obtained 2D image is a pseudo-3D image since the current pathways are not constrained to the examined layer. Electrical impedance tomography supplies a comparable near-anatomic cross-sectional image comparable to those of other CT-based procedures [e.g., X-ray CT, NMR, or ultrasound (US)], however, with very poor spatial resolution. Boundaries of compartments with the same material-specific coefficients are found by segmentation. Segmented areas with assumed thickness of the single layers are used for volume estimation. Changes in the volume can be assessed by comparing the volumes obtained in consecutive images.

It is a characteristic feature of all methods that record the electrical bioimpedance that the evoked response depends on the strength of the local electromagnetic field. For this reason, it has to be taken into account that the resulting current density may be inhomogeneous in the considered tissue volume. Causes for such nonhomogeneity may be geometric constraints (e.g., a nonregular shape of the considered volume); the composition of the tissue within the considered volume that may be a mixture of tissues with different electrical properties (e.g., blood with low resistivity, or bone with high resistivity, as compared with skeletal muscle). Those different tissues may electrically be switched in parallel or serial order; and the size and the location of the current-feeding electrodes. The current density is higher in regions near to the feeding electrodes than in distant regions. Consequently, the regions near to the feeding electrodes give the strongest contribution to the measured voltage.

This requires (1) careful selection of the current-feeding site. In the tetrapolar mode also the position of the voltage-sensing electrodes must be taken into account; and (2) appropriate consideration of the inhomogeneous distribution of the electrical parameters within the considered tissue volume, that is, the course of blood vessels.

Special electrode arrangements have been developed for certain applications in order to minimize the measurement errors. Concentric multielectrode arrangements with the outer electrodes on a potential different from that of the inner electrode have been proposed with the objective to optimize the current distribution in the measured volume.

The frequency that can be used for the measurement of the passive electrical properties of biological tissue ranges from very low frequencies to some gigahertz. The most popular frequency band for impedance plethysmography is between 1 kHz and 10 MHz. This frequency band encloses the so-called β-dispersion, which is actually a dielectric or structural relaxation process. The β-dispersion is also known as Maxwell–Wagner relaxation. It is characterized by a transition in the magnitude of the electrical parameters with frequency. This transition is caused by the fact that cellular membranes have high impedance below and low impedance above that β-dispersion. For frequencies distinctly below the β-dispersion, the current flow is restricted to the extracellular space. For frequencies distinctly above the β-dispersion, the current can pass through the cellular membrane. Consequently, with frequencies distinctly below the β-dispersion only the volume or volume changes of the extracellular space will be monitored, whereas with frequencies distinctly above the

Table 1. Compilation of Typical Values of Resistivity (Ω·m) of Various Body Tissues[a]

	Frequency							
	10 Hz	100 Hz	1 kHz	10 kHz	100 kHz	1 MHz	10 MHz	100 MHz
Muscle, skeletal	9.6	8.8	8.1	7.6	2.0	1.8	1.6	1.4
Muscle, heart	9.6	9.3	8.0	6.0	2.1	2.0	1.6	1.5
Liver	10.0	8.7	8.6	7.6	4.6	2.8	2.8	1.7
Kidney					1.9	1.8	1.4	1.3
Brain			6.1		6.0	5.3	3.7	1.5
Fatty tissue			23.2					
Blood			1.6	1.5	1.5	1.4	0.9	0.8

[a]Note that those values must not be assumed to represent exact figures since they do not consider important details like species, preparation of sample, time after excision, temperature, or the procedure and protocol for their measurement. The values are compiled from many different sources and, if necessary transformed to resistivity.

β-dispersion, the total volume (i.e., both extra- and intracellular space) or changes of this total volume can be recorded. Using at least one frequency below and another one above the β-dispersion allows determining the ratio of extra- and intracellular spaces, and hence also fluid shifts between these spaces.

Special applications of this approach are the monitoring of fluid exchange processes during hemodialysis (5) and orthostatic challenges (6), the control and management of fluid infusion therapy, the detection of lung edema, and the viability surveillance of organs after blood flow has been stopped during surgery or when the organs are preserved for transplantation (7,8). The viability surveillance is based on the fact that oxygen deficiency with the subsequent lack of energy-rich substrates causes a failure of the active transmembraneous ionic transport mechanisms and, as a consequence, leads to an intracellular edema (i.e., an increase of the intracellular volume). This approach has also been investigated for graft rejection monitoring.

The passive electrical properties are specific for each tissue. They are mainly depending on the content of water, the ratio of extra- and intracellular space, the concentration of electrolytes, and the shape of the cells and their orientation in the electrical field (e.g., of the fibers of skeletal and cardiac muscle). Table 1 shows a compilation of typical values of resistivity of various body tissues. It must be taken into account, however, that these values do not represent exact figures. Exact figures need detailed information about species, preparation of the sample, time after excision, measurement temperature, the employed method, and the protocol for the measurement. Comprehensive data compilations with the supplement of those details are found in Refs. 9 and 10.

These tissue-specific properties can be used for special applications, such as the analysis of the tissue composition or for tissue characterization by Impedance Spectroscopy. Those methods are the subject of another article and will not be discussed here in detail. A very popular application is the determination of total body water (11) or of whole body composition, for example, the determination of the percentage of body fat in order to support adequate nutrition management or control of physical exercises. Such approaches aim for the estimation of the compartmental volume of a certain tissue (e.g., fat) that is mixed with another tissue (e.g. fat-free tissue) in a common space (i.e., the body or an extremity).

FUNDAMENTALS OF BIOIMPEDANCE MEASUREMENT

The most important principle for bioimpedance measurements is the adequate modeling of the passive electrical behavior of the tissue by an equivalent electrical circuit. The validity of simple models is restricted to narrow frequency ranges (e.g., the β-dispersion) and/or to simple geometric shapes of the biological object (e.g., cylinders as a model for extremities). The most widely accepted models for the bioimpedance in the frequency range around the β- dispersion are the *RC*-networks shown in Fig. 1. These models represent the spatially distributed electrical properties by discrete components. Actually, they are only simplified 2D models. The network shown in Fig. 1a is mimicking the biological system and its histological structure. It represents both the extracellular and intracellular space by the resistors R_e and R_i, respectively, and the cell membrane by the capacitor C_m. Since the current passes twice the membrane when flowing through the cell, the two capacitors C_m* in series with R_i can equivalently be expressed by one single capacitor C_m in series with R_i. This network is usually replaced by the one shown in Fig. 1b in which R_s is arranged in series with the parallel circuit of R_p and C_p. These components have no relation with real histological structures. The parameter R_s corresponds to the parallel circuitry of R_e and R_i as can be demonstrated for high frequencies. The parameter R_s can be considered to be very small as compared with R_p. In many cases, R_S may

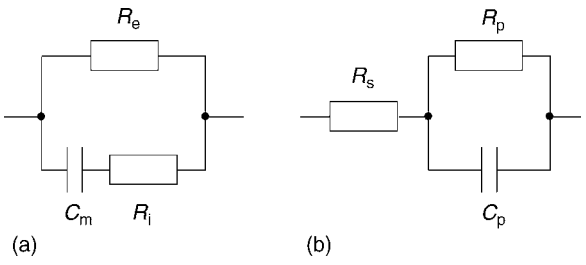


Figure 1. RC-networks modeling tissue impedance. The model in (a) mimicks morphological structures, whereas the model in (b) shows the electrically equivalent, but, more usual circuitry.

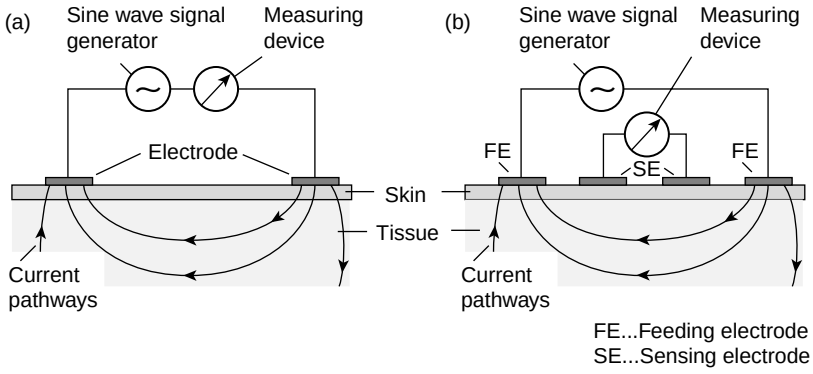


Figure 2. Bioimpedance measurement set-up. (a) Shows the two-electrode configuration, whereas the four-electrode configuration is depicted in (b).

even be neglected for cases of simplification, that is, the electrical model is simply a parallel circuit of a resistor and a capacitor (e.g., $R_p \| C_p$).

When using contacting electrodes, different approaches are possible for the measurement of the bioimpedance. The most important feature is the number of employed electrodes, usually two or four electrodes. More than 4 electrodes, up to 128 electrodes, are primarily used for CT-based impedance imaging like EIT. The two-electrode configuration is called the bipolar mode, and the four-electrode configuration is the tetrapolar mode (Fig. 2). The bipolar mode can be compared with the usual method for impedance measurement by feeding a current into the measurement object and recording the voltage or vice versa. In the tetrapolar mode, two electrodes are the feeding electrodes (usually the outer electrodes) and the other two electrodes are the sensing electrodes (usually the inner ones). In the tetrapolar mode, more than two sensing electrodes can be employed, for example, if monitoring of serial segments at the extremities are to be achieved.

The interface between the electrode with the metallic plate on the one side and the electrolyte on the other side is the boundary where a current carried by electrons is transformed into a current carried by ions. The electrolyte may either be contained in the gel of the electrodes or be the electrolytic fluid in the tissue. The basic process of the charge transfer from electrons to ions is a chemical reaction (12). The simplest model of such an interface is again an impedance consisting of a parallel circuit with a resistor R_F (the Faraday resistance) and a capacitor C_H (the Helmholtz capacitance), i.e., $R_F \| C_H$ (Fig. 3b). Real electrodes show a polarization effect that is caused by a double layer of opposite charges at the interface, actually the Helmholtz capacitance (Fig. 3a). Therefore, the electrode model with $R_F \| C_H$ has to be supplemented with an additional voltage source E_P. The steady-state condition is reached if the tendency of metallic ions to enter the electrolyte and leave behind free electrons is balanced by the electrostatic voltage originating from the double layer. After disturbances, for example, by charge transfer forced by an externally applied voltage, another equilibrium for the double-layer voltage is reached with a time constant depending on R_F and C_H. All these components may have poor stability with time, especially in the period immediately after attaching the electrode on the skin. For surface electrodes, it must also be taken into account that the impedance of the skin, especially the stratum corneum, which is the outmost epidermal layer, can be much larger than the impedance of the deeper tissue (e.g., skeletal muscle), which is in a complex parallel-serial arrangement with the skin. Sweating underneath the electrode lowers the electrode-tissue transimpedance. For the measurement of the impedance of deeper tissues the adequate preparation of the skin by abrasion, stripping, or puncturing at the site of the electrodes might be necessary in order to diminish the transimpedance. This transimpedance depends on the size of the electrode (i.e., the contacting area) and the measurement frequency, and ... additionally on the pressure with which the electrode is attached to the skin. For electrodes, a typical value for the transimpedance is $\sim$100–200 $\Omega \cdot cm^2$.

In the bipolar mode, the two electrode–electrolyte interfaces are in series with the actual bioimpedance of the measured object. Therefore, the recorded impedance is always the sum of at least three impedances. The impedance of the biological sample cannot be calculated as an individual quantity from the recorded impedance value. This is the most serious shortcoming of the bipolar mode.

In the tetrapolar mode, the electrode–electrolyte interface usually can be neglected if the measurement is performed with a device with high input impedance (i.e., with very low current passing across the electrode–tissue inter-

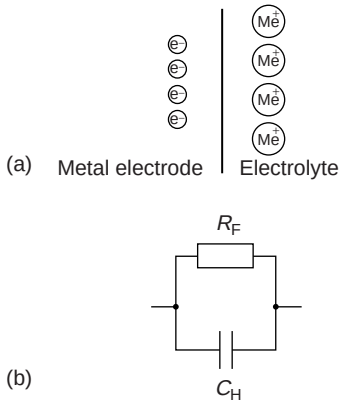


Figure 3. Metal electrode: electrolyte interface. (a) Illustrates the steady-state condition at the boundary. Following the tendency of metallic ions (Me^+) to enter the electrolyte and to leave behind free electrons, the steady state is reached when this tendency is balanced by the electrostatic voltage originating from the double layer. (b) Shows the simplified model with only the passive electrical components, that is, the Faraday resistance R_F and the Helmholtz capacitance C_H.

face). A drawback of the tetrapolar mode, however, is that the measured impedance value cannot be exactly assigned with a certain volume of the tissue between the sensing electrodes, even if the sensing electrodes are positioned on a straight line connecting the two feeding electrodes. Band electrodes that are attached at the whole circumference (e.g., at the extremities), yield more valid results. If circular electrodes are applied and both the feeding and the sensing electrodes are placed on the circumference of a cross-section (e.g., around the thorax or the head), it is nearly impossible to assign the actually measured impedance value with a certain volume within that cross-section due to the complex current field. In those cases, the monitored volume consists of a serial-parallel circuitry of different tissue layers with different electrical properties (e.g., skin, subcutaneous soft tissue, bone, deeper tissues). For this reason, the result of ventilation measurements with electrodes, either disk or band electrodes, placed on the thorax may be affected by the higher conductivity of extra-thoracic muscles as compared with the very low conductivity of the rib cage, which prevents the entrance of current into the pulmonary tissue that is actually the object of interest. Sweating may additionally cause another low impedance parallel circuit along the skin and, thus, yield considerable measurement errors. The situation is similar for the measurement of brain parameters, for example, brain–blood flow or brain edema, with electrodes placed on the scalp. Since the conductivity of the extracranial soft tissue (e.g., skin, muscle) is much higher than the conductivity of the bony skull, only few current pathways will pass through the brain.

The different instrumental approaches for measuring the bioimpedance in the frequency range of the β-dispersion are the impedance bridge (e.g., an extension of the classical Wheatstone bridge); the self-balanced active bridge; the resonance method that is mainly a compensation method; the pulse method that is a time domain procedure and not widely used; the voltage-current method, based either on feeding a constant voltage (i.e., from a generator with low source impedance) and monitoring the resulting current or on feeding a constant current (i.e., from a generator with a sufficiently high source impedance, $\sim$100 kΩ since the load impedance may be up to 1 kΩ) and monitoring the resulting voltage. If employing the bipolar mode, it would be more correct in this case to use the term transimpedance than impedance for the actually measured quantity between the electrodes.

For the tetrapolar configuration only the voltage-current method is applicable. Phase angle measurements employ an ohmic resistor in series with the measuring object as reference. Absolute phase angle measurements are questionable even for the bipolar configuration since the measured phase angle always includes the phase shifts that are caused by the two electrode–skin contacts and depend on the actual values of the Faraday resistance and the Helmholtz capacitance. If the Faraday resistance is small and the Helmholtz capacitance is fairly large, the phase shift by the electrode–skin interface may become negligible. This is one of the advantages of electrodes with artificially increased surfaces, for example, electrodes with porous or fractally coated surfaces that might be obtained by sputtering or chemical processes, as compared with polished surfaces.

Usually, the measurement is performed with a constant-voltage generator for technical reasons. The applied feeding signal, whether voltage or current, should be sinusoidal with a very low distortion factor (i.e., with a low content of harmonics) and with high stability both in amplitude and frequency. Any modulation of the feeding signal may provoke an additional response for this undesired modulation frequency that has to be avoided with regard to the frequency dependence of the impedance specific variables. The voltage amplitude is in the range of some volts, the current amplitude is in the range of some microamps to milliamps (μA to mA). The changes in the impedance caused by volume changes can be very small, $<0.001\%$. This means that very small changes in the measured current or voltage have to be processed. Hence, the sensitivity and stability of the input amplifier must be very high in order to detect such small changes in the measured signal.

Independent from the measurement method, careful consideration of measurement errors is necessary. A main source of measurement errors may be parasitic components, such as stray capacitances between neighboring wires leading to the sensing electrodes, or between wires and their shielding, or stray capacitances between metallic components of the measuring system and ground, which become the more effective the higher the measuring frequency.

The risk for undesired stimulation of the heart or peripheral nerves if such electrical voltages or currents are applied for monitoring purposes, is negligible, both with regard to the high frequency and the low current density. Furthermore, heating and heat-induced secondary effects can be neglected.

However, proper attention must be paid for the selection of the equipment and its performance data for the intended application. Furthermore, the employed devices must be safe even in case of technical failure. Patient-near devices are directly connected with the patient whereby the connecting impedance is rather low.

CHARACTERISTICS OF BIOIMPEDANCE

The microscopic electrical properties that describe the interaction of an electromagnetic wave with biological tissue are the complex conductivity σ^* with the unit $\Omega^{-1}\cdot m^{-1}$, mho$\cdot m^{-1}$, S$\cdot m^{-1}$, or $1\cdot(\Omega\cdot m)^{-1}$

$$\sigma^*(\omega) = \sigma'(\omega) + j\sigma''(\omega)$$

and the complex dielectric permittivity ε^* with the unit F/m

$$\varepsilon^*(\omega) = \varepsilon'(\omega) - j\varepsilon''(\omega)$$

ω is radian frequency with the unit hertz. The electrical properties are depending on the frequency with strong dependence in the range of a dispersion.

The relation between these two quantities can be described in accordance with Ref. 13 by

$$\sigma^*(\omega) = j\omega\,\varepsilon^*(\omega)$$

With the conduction current that is related with the basic conductivity σ_0, that is, the current carried by the mobility of ions in the extracellular space, and the polarization current (sometimes called displacement current) that is related with permittivity, the following equations are obtained

$$\sigma' = \sigma_0 + \omega\,\varepsilon''(\omega)$$
$$\sigma'' = \omega\,\varepsilon'(\omega) = \varepsilon_0\,\varepsilon_r(\omega)$$

where ε_0 is the dielectric permittivity of free space with $\varepsilon_0 = 8.85 \times 10^{-12}$ F·m^{-1}, and ε_r is the relative dielectric permittivity (with $\varepsilon_r = 1$ for the free space and $\varepsilon_r = 81$ for water in the low and medium frequency range).

Instead of the complex conductivity σ^*, the inverse complex resistivity ρ^* with the unit Ω·m can be used. The resistivity is usually preferred in the context of impedance plethysmography:

$$\rho^*(\omega) = \rho'(\omega) + j\,\rho''(\omega)$$

The complexity of these quantities considers the fact that in the alternating current (ac) range the biological tissue cannot adequately be described by a simple resistance (or its inverse conductance), but needs the extension to a complex quantity, that is, impedance or admittance. Some authors prefer the term Admittance Plethysmography instead of Impedance Plethysmography (14,15). The simplest adequate model for such an impedance is represented by a resistance and a reactance. The resistance causes the loss in power, whereas the reactance causes the delay (or 18 phase shift) between voltage and current. The dominating reactance of bioimpedance in the frequency range of interest is capacitive and becomes more relevant for higher frequencies. Only for very high frequencies that usually are not employed for impedance plethysmography, can the reactance be composed by both a capacitive and an inductive component.

Bioimpedance can be described like any technical impedance in different forms, for example, by its magnitude (or modulus) Z_0 and its phase angle (or argument) φ, (i.e., the delay between voltage and current):

$$Z = Z_0\,e^{j\varphi}$$

or by its real part (or resistance) Re{Z} and its imaginary part (or reactance) Im{Z}:

$$Z = \mathrm{Re}\{Z\} + j\,\mathrm{Im}\{Z\}$$

Alternating current voltages and ac currents, too, can be expressed as complex quantities, that is, by their magnitude and phase angle, or by their real and imaginary part although this is rather unusual. The magnitude of the impedance Z_0 corresponds to the quotient of the magnitudes of voltage V_0 and current I_0, that is,

$$Z_0 = V_0/I_0$$

Appropriate modeling of the electrical properties of biological tissue by discrete and lumped electrical components renders possible the proper consideration of multilayer or compartmentally composed tissues with different electrical properties of each layer or compartment. Such tissues can be modeled as serial, parallel, or serial–parallel equivalent circuits in 2D presentation. More recently, the modeling has been extended to 3D models using the FEM or comparable approaches.

The impedance parameters can be depicted in different modes as a function of frequency (Fig. 4). The presentation of the magnitude (usually in logarithmic scale with regard to its wide range) and the phase angle against the frequency over several decades, and therefore in logarithmic scale is known as the Bode plot. A similar presentation is used for both the real and imaginary part versus the frequency on the x axis. This mode of presentation is sometimes called the spectrum (e.g., modulus spectrum and phase angle spectrum). A different form of presentation is in a plane with the real part along the x axis and the imaginary part along the y axis, both in linear scaling, with the frequency as parameter. This mode of presentation is frequently called the Cole–Cole-plot (but also the Nyquist plot, locus plot, or Wessel graph). The same modes of

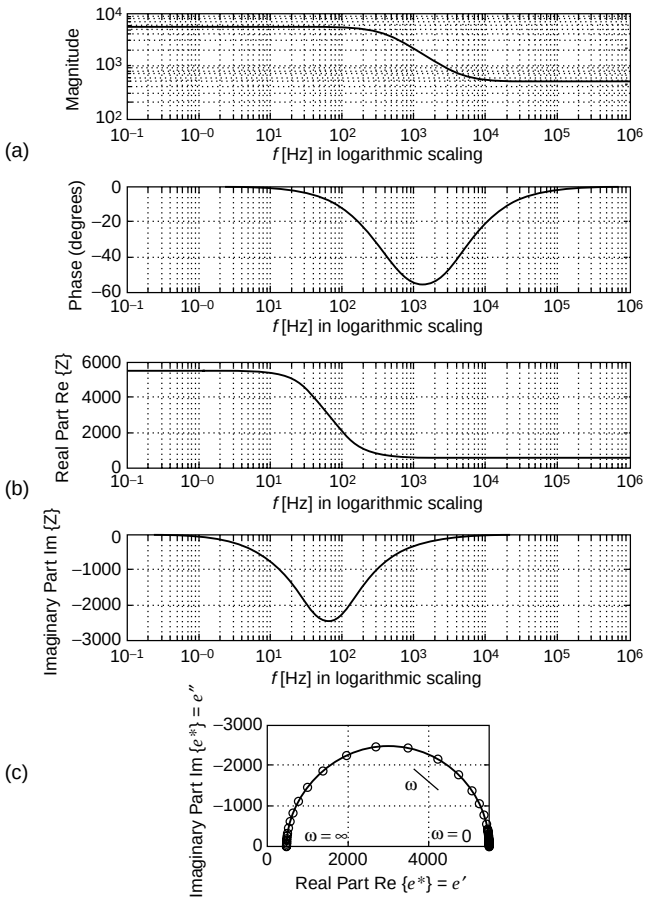


Figure 4. Different modes for the presentation of impedance quantities. (a) Shows the magnitude (in logarithmic scaling) and the phase angle of impedance as functions of frequency in logarithmic scaleing (Bode plot). (b) Shows the real and imaginary part of impedance depicted versus the frequency in logarithmic scaling. (c) Presents the Cole–Cole plot with the real part at the x axis, the imaginary part at the y axis, and the frequency as parameter along the curve. The results shown are not from biological tissue, but calculated for the circuit of Fig. 1b with $R_s = 500\ \Omega$, $R_p = 5\ \mathrm{k}\Omega$, and $C_p = 500$ nF.

presentation are possible for the complex quantities σ^*, ρ^*, and ε^*.

Usually, impedance plethysmography is accomplished employing only a single measuring frequency or a few discrete measuring frequencies. However, impedance spectroscopy with a multitude of measuring frequencies is gaining interest, especially for the determination of spatially distributed volumes. Typical examples are the determination of body composition, tissue, or organ vitality monitoring in combination with cellular edema as a result of hypoxemia, and monitoring of infusion therapy.

Certain forms of electrotherapy are also utilizing the passive electrical properties of biological tissue. Methodology and technology for these forms of electrotherapy, however, are not discussed here.

MODEL-BASED RELATIONS FOR VOLUME DETERMINATION

Valid relations between the monitored electrical quantities and the searched volumetric parameters have to be used to calculate a quantitative result that can be expressed in units of volume (e.g., mL or cm^3). Most of these relations are based on models. Possibly the first model for the interpretation of bioimpedance measurements has already been developed for a suspension of cells in a fluid by Fricke and Morse in 1925 (16).

The simplest model-based approach for establishing an impedance–volume relationship is a cylindrical volume conductor of radius r_0, length L, and resistivity ρ^*. It is assumed that this volume conductor is surrounded by soft material with significantly higher resistivity $\rho_e^* \gg \rho^*$, so that it must not be considered as a parallel circuit and its actual radial extension has no impact. This volume conductor may be a blood vessel (e.g., an artery or vein) surrounded by tissue that has higher resistivity than blood. Furthermore, it is assumed that the inflow of the volume ΔV into that cylinder expands the radius homogeneously by Δr over the total length. It is also assumed that neither the length L nor the resistivity ρ^* are affected by the volume injected into the volume conductor. For didactic simplicity, only the real part ρ' of the complex resistivity ρ^* is considered; that is, only the real part of the impedance is taken into account. However, despite this simplification, the variable is understood as impedance Z. This simplification is generally valid for frequencies much lower than the β-dispersion, since for these frequencies the phase angle is small ($< 10^\circ$). For higher frequencies, the calculation must be performed with proper consideration of the complex quantities.

With these assumptions the following relations for Z and V are valid:

$$Z_0 = \rho' L/[\pi r_0^2]$$
$$V_0 = L\pi r_0^2$$

From these two equations the following equation can easily be calculated

$$Z_0 = \rho' L^2/V_0$$

After the inflow of the volume ΔV and the increase of the radius by Δr, the following relations are valid:

$$Z_1 = Z_0 - \Delta Z = \rho' L/[\pi(r_0 + \Delta r)^2]$$
$$V_1 = V_0 + \Delta V = L\pi(r_0 + \Delta r)^2$$

with $Z_1 < Z_0$ and $V_1 > V_0$ for $\Delta r > 0$.

With some simple mathematical operations it can be shown that

$$\rho' L^2 Z = (\rho' L^2 + \Delta V Z)(Z - \Delta Z)$$

If the product $\Delta V\, \Delta Z\, Z$ is neglected as a product of small quantities, the result becomes:

$$\Delta V = \rho'(L/Z)^2 \Delta Z$$

This is the well-known Nyboer equation that relates the volume change ΔV with the change in impedance ΔZ as a consequence of the blood inflow into a peripheral artery, for example, into the aorta or carotid artery with each heart beat.

For proper application all included simplifications must carefully be taken into account. Only mathematical simplifications, but no methodological constraints have been mentioned here. Such a constraint may be that a part of the injected volume is already flowing out of the measured vessel segment before the inflow of the whole volume ΔV into this vessel segment has been completed. This constraint must especially be considered if the segment is short and the vessel wall rather stiff.

With regard to the surrounding tissue, a more realistic model would be an arrangement of two concentric cylinders of length L with different conductivities. The inner cylinder with radius r_1 and resistivity ρ_1' has the impedance $Z_1 = \rho_1' L/[\pi r_1^2]$, whereas the outer cylinder with radius r_2 and resistivity ρ_2' has the impedance $Z_2 = \rho_2' L/[\pi(r_2 - r_1)^2]$. Electrically both cylinders are arranged in parallel configuration. Hence, the total impedance is obtained by $Z_0 = Z_1 Z_2/(Z_1 + Z_2)$. The inner cylinder shall be a vessel into which blood of volume ΔV is pumped, causing a homogenous dilation of the vessel radius by Δr_1 and a lowering of its impedance to $Z_1^* = Z_1 - \Delta Z_1 = \rho_1' L/[\pi(r_1 + \Delta r_1)^2]$. Since Z_2 shall not be affected, the total impedance is $Z_0^* = Z_1^* Z_2/(Z_1^* + Z_2)$. The following steps are similar to those leading to the Nyboer equation. Since the resulting equation and its application to measurements are more complicated, they will not be discussed here in detail. Even this model is actually simplified, because in reality the tissue around the vessel will neither be a cylinder nor have a homogeneous resistivity. This last situation may become relevant with a vein or a bone in the vicinity of the artery.

With regard to the constraints of the Nyboer equation, another approach has been used that finally leads to the Kubicek equation (17). The model-based approach starts again with the single-vessel model of length L. However, in contrast to the Nyboer approach the assumption is not made that the inflow of the volume ΔV into the considered vessel segment is finished before the outflow starts. Here, the basic assumption is that the inflow is constant during the inflow time T_{inf} and that the outflow starts with some delay, however, temporal overlap of outflow with inflow

must not be excluded. With this assumption, the change in the intravascular volume and, hence, in the impedance, is maximal when there is only inflow into and no outflow from the segment. This maximal change of the impedance can be expressed by its first time derivative [i.e., by $(dZ/dt)_{\max}$]. The total inflowing volume ΔV can then be taken into account by multiplying $(dZ/dt)_{\max}$ with the inflow time $T_{\inf}$. With regard to the aorta this inflow time is equivalent with the ejection time of the left ventricle. In many cases even the inflow time can additionally be obtained from the impedance curve. This leads finally to the Kubicek equation:

$$\Delta V = \rho'(L/Z)^2 T_{\inf}(dZ/dt)_{\max}$$

Obviously, the only relevant difference in both approaches is the Nyboer assumption that the total volume change ΔV has already been injected into the measured vessel segment before the outflow starts against the Kubicek assumption that this volume ΔV is entering the measured vessel segment with constant rate during the whole inflow period. The Kubicek equation is more realistic for a short vessel segment with a rather stiff wall. For such vessels, the Nyboer equation leads to an underestimation of the real volume change. In contrast, if the inflow is decreasing at the end of the inflow period, for example, at the end of the ventricular ejection period, the Kubicek equation yields an overestimation of the volume change.

All other model-based assumptions are identical or comparable. Both approaches consider only a single vessel with homogeneous dilation over the total length within the measured tissue and neglect the surrounding tissue and its composition with regard to nonhomogeneous resistivity. Blood resistivity is taken as constant although there is some evidence that it depends on the flow velocity.

Although the Kubicek equation has primarily been proposed for the monitoring of cardiac output, both equations have also been applied to the monitoring of pulsatile peripheral blood flow. Both models, however, do not consider that in the peripheral circulation a basic or nonpulsatile blood flow may exist as well.

Different modifications have been suggested in order to overcome relevant drawbacks. Most of these modifications are optimized with regard to the monitored quantity, geometric constraints, modes of application, or positioning and shape of electrodes. They will not be discussed here.

No valid impedance–volume models have been proposed for the quantitative monitoring of ventilation by the application of impedance plethysmography. Statistical models are used for the impedance–volume relationship regarding body composition. Some first approaches have been suggested for the volume changes due to fluid shifts.

INSTRUMENTATION AND APPLICATIONS

The typical basic equipment for impedance plethysmography consists of the signal generator, either a constant voltage generator or a constant current generator; the frequency-selective measuring device, either for current or voltage, in combination with AD conversion. The equipment may be supplied with more than one signal channel for certain applications, for example, with two channels for simultaneous and symmetric monitoring at both extremities or one channel for each frequency in multifrequency measurements; the signal processor, for example, for processing the impedance quantities; the processing unit for calculating the volumetric quantities; the monitor and/or data recorder; multiple electrodes and shielded leads; specific auxiliary equipment, for example, venous occlusion machine with cuff and pump.

Devices for impedance plethysmography are small, light, usually portable, and battery powered. The devices for patient-near application are much cheaper than competitive equipment based on nuclear magnetic resonance (NMR), X ray, or US. Also, the running costs are much lower than for the competitive technologies, usually these costs are mainly required by the single-use electrodes.

Peripheral Hemodynamics

The objective is the detection of deficiencies either in the arterial or venous peripheral circulation. The application of impedance plethysmography to peripheral vascular studies has already been in the interest of Nyboer in 1950 (18).

In the peripheral circulation, the most interesting quantity is arterial blood flow or perfusion. Impedance measurement is performed either in the bipolar or, more frequently, in the tetrapolar configuration. The tetrapolar configuration requires a longer segment for measurement in order to place the sensing electrodes in proper distance from the feeding electrodes with the nonhomogenous current field in their vicinity. Electrodes are either of the circular or disk or the band type. Disk electrodes can be placed directly above the monitored vessel and therefore provide high sensitivity, but the magnitude and the reproducibility of the measured signal in repeated measurements are strongly dependent on exact electrode placement (19–21). Band electrodes are preferred for the measurements at extremities because they can be placed around the extremities. In this case, the measured object is the whole segment between the sensing electrodes including the extravascular tissue and may include more than only one vessel. Flow can be estimated by application of the Nyboer, the Kubicek or any modified impedance–volume equation. Competitive methods are utilizing ultrasound Doppler, contrast X-ray angiography, or NMR.

Some diagnostic information about the stiffness of the arterial vessel wall can be obtained by the impedance method from the measurement of the pulse wave propagation velocity, usually executed at two different sites of the same arterial pathway. The pulse that is actually recorded with the impedance method is the intravascular volume pulse, that is, the dilation of the vessel with each heart beat (22). Simple formalistic models are used to relate the pulse wave propagation velocity with the stiffness of the vessel wall. If the intravascular blood pressure is also monitored, then it is possible to calculate the stiffness or its inverse, the compliance as ratio of the volume change and pressure change $\Delta V/\Delta p$, directly.

Another problem is the diagnosis of proximal or deep venous thrombosis and of other obstacles for the venous return flow to the heart from the extremities (23). One

approach is actually a modification of the venous occlusion plethysmography that has already been introduced in 1905 by Brodie and Russel (24). A cuff is placed around the extremity and connected with a pump. The cuff pressure is enhanced abruptly so that it occludes the vein and stops venous outflow without affecting the arterial inflow. The volume increase following venous occlusion allows estimating the arterial inflow. When the occlusion is stopped after $\sim$ 20 s, the venous outflow starts again and thereby leads to a reduction in volume. The slope or the time constant of this postocclusion emptying process are used to assess the outflow resistance, for example, the hydrodynamically obstructive impact of deep venous thrombosis, or the venous wall tension. However, other pathological effects must be carefully considered, for example, increased central venous pressure. For this reason, the recording is frequently and simultaneously performed on both extremities, so that the results can be compared with each other. The measurement is usually executed with band electrodes. Competitive methods are ultrasound Doppler, contrast X-ray venography, or NMR.

A similar impedance-based approach is employed to test the performance of the drainage system in extremities. Changes in the hydrostatic pressure are used to shift volume between the trunk and an extremity, for example first by bringing down an arm before raising it above the head. The affected volume shifts can be recorded and render possible the assessment of the performance of the draining system. This approach is frequently used to study fluid shifts caused by tilting experiments, during microgravity experiments, or after long periods of bedrest.

Brain Perfusion and Edema

The most important objectives are monitoring of cerebral bloodflow and the detection of cerebral edema. First publications about rheoencephalography are dating back to 1965 (25,26).

The volume of the brain with its enclosed fluid spaces, for example, the intravascular volume and the cerebrospinal fluid volume, is kept constant by its encapsulation in the bony skull. The expansion of the volume of one compartment, for example, increase of the intravascular volume by augmented arterial blood pressure, the space-demanding growth of a brain tumor or intracerebral bleeding, can only be compensated by the diminution of the volume of other compartments. If the space-demanding process is of nonvascular origin, the most affected compartment will be the intravascular volume. Due to the compression of blood vessels, the cerebral bloodflow and thus metabolism will be reduced.

Impedance measurements aiming for the brain as organ are difficult because the encapsulating bony skull has a very high resistivity as compared with the soft extracranial tissue of the face and the scalp. If the tetrapolar mode is used, more than two sensing electrodes may be applied. Different electrode arrangements have been described to force the current pathways through the skull into the brain, but also the application of the feeding electrodes to the closed eyelids. However, the measurement of the transcephalic impedance has not become a routinely applied clinical method with the exception of neonates in which the thickness of the bony skull is very small. Competitive methods based on NMR, X ray, US, and even photoplethysmography have gained more attention in the recent past. Some expectations are related to the development of contactless applications, especially for the continuous monitoring of edema (27). This might allow control treatment by hyperosmolaric infusion therapy.

Ventilation and Lung Performance

Impedance pneumography were among the first applications of impedance plethysmography and had already been described by Geddes et al. in 1962 (28,29).

The objective of impedance pneumography is to record the tidal volume under resting conditions or during exercise. Additionally, the breathing rate can be obtained by the impedance method. The principle is based on the measurement of the transthoracic impedance that increases during inspiration as a consequence of increasing alveolar air filling, and decreases during expiration (30). The conductivity of lung tissue at the end of a normal expiration is $\sim$0.10 $\Omega^{-1}\cdot m^{-1}$ as compared with 0.05 $\Omega^{-1}\cdot m^{-1}$ at the end of normal inspiration. The application of impedance pneumography is very simple and also applicable for critically ill patients, since it allows continuous recording without requiring a breathing tube. However, the quantitative determination of the tidal volume is difficult and needs calibration by another spirometric method. No realistic model-based interpretation of quantitative assessment is available until now. Some expectations are related with multifrequency measurement (31).

The impedance measurement can be performed with the bi- or tetrapolar configuration. It is usually performed separately for each side of the lung in order to detect differences. Even with more electrodes, however, the spatial resolution is too poor to allow detection of regional inhomogeneities of alveolar air filling. For that reason, this field is gaining growing interest for the application of EIT (4). Also, EIT has limited spatial resolution, as compared with other CT-based imaging procedures. Despite this drawback, it has some potential for the detection of regional inhomogeneities in ventilation. Such an approach would be of high relevance for diagnostic purposes and for the efficiency control of artificial ventilation. Serious drawbacks for EIT, however, are the costs of the equipment and the necessity to attach up to 64 or 128 electrodes around the thorax.

Since pulmonary edema is usually a general and not a localized phenomenon, its monitoring might be possible by utilizing the transthoracic impedance measurement. Measurement of extravascular lung water is investigated as a methodological approach to guide the fluid management of patients with noncardiogenic pulmonary edema (32). The conductivity of lung edema fluid is $\sim$1 $\Omega^{-1}\cdot m^{-1}$, and therefore is distinctly different from alveolar tissue filled with more or less air.

The impedance coefficients of tumor tissue are different from normal lung tissue. Hence, cancer detection might be possible by bioimpedance measurement. But with regard to its poor spatial resolution, the bi- or tetrapolar transthor-

acic impedance measurement is not qualified, but EIT might become useful for some special applications. Other imaging procedures were superior until now, primarily due to the higher spatial resolution as compared with EIT.

Much work has been devoted to comparing impedance pneumography with inductive pneumography. There is some evidence that inductive pneumography is superior concerning ventilation monitoring in newborn infants, especially for risk surveillance with regard to SIDS. An interesting approach tries to utilize inductive pneumography in combination with the evaluation of the signal morphology for the monitoring of airway obstruction (33).

Intercompartmental Fluid Shifts

Intercompartmental fluid shifts occur during dialysis, for example, hemodialysis, but also during infusion therapy and emergence of edema. The assessment of fluid shifts, which are actually changes of volumes, by the measurement of electric variables has been an outstanding objective very early in the scientific research and medical utilization of bioimpedance and dates back to 1951 (34).

Hemodialysis is a therapeutic procedure that is employed in patients with renal insufficiency. The therapeutic objective is to remove water, electrolytes, urea, and other water-soluble substances in combination with the reestablishment of a normal acid–base status. In a simplified model, the water is first removed from the intravascular volume, (i.e., the blood plasma). This means that the hematocrit, and thereby the viscosity of blood, is increased cousing the work load for the heart is enhanced. Also, the osmotic pressure of the blood is raised, whereas the hydrostatic blood pressure is lowered. Both effects contribute to the refilling of the intravascular space by a fluid shift from the interstitial space (i.e., the extravascular extracellular space). This fluid shift finally causes another fluid shift from the intracellular space into the interstitial space. The dynamics of these fluid shifts is primarily controlled by the hydrostatic and osmotic pressure differences between the different spaces, but also by the substance-specific permeability of the different barriers between the spaces including the dialysis membrane. If removal of water is too fast or changes of ions like sodium, potassium, and calcium are too large, the hemodynamics of the patient or the excitability of tissues like the heart or central nervous system may become disturbed. Impedance plethysmographic methods have some potential for the control of those fluid shifts, that is, may help to avoid critical disequilibrium syndromes like hypotension, headache, and vomiting. The best results of the plethysmographic measurement are achieved if segmental measurement is performed at the extremities instead of whole body measurements (5). Figure 5 shows a schematic presentation of such segmented body. The forearm accounts only for $\sim$1% of body weight, but contributes 25% to whole body impedance.

Infusion therapy aims mainly to filling the intravascular volume by utilizing venous access. However, depending on the control variables (e.g., hydrostatic pressure, osmotic pressure, and permeability of the barriers), intercompartmental fluid shift cannot be avoided. Consequently, part of the infused volume will not remain in the intravascular system and help to stabilize blood pressure, but escape to the extravascular space.

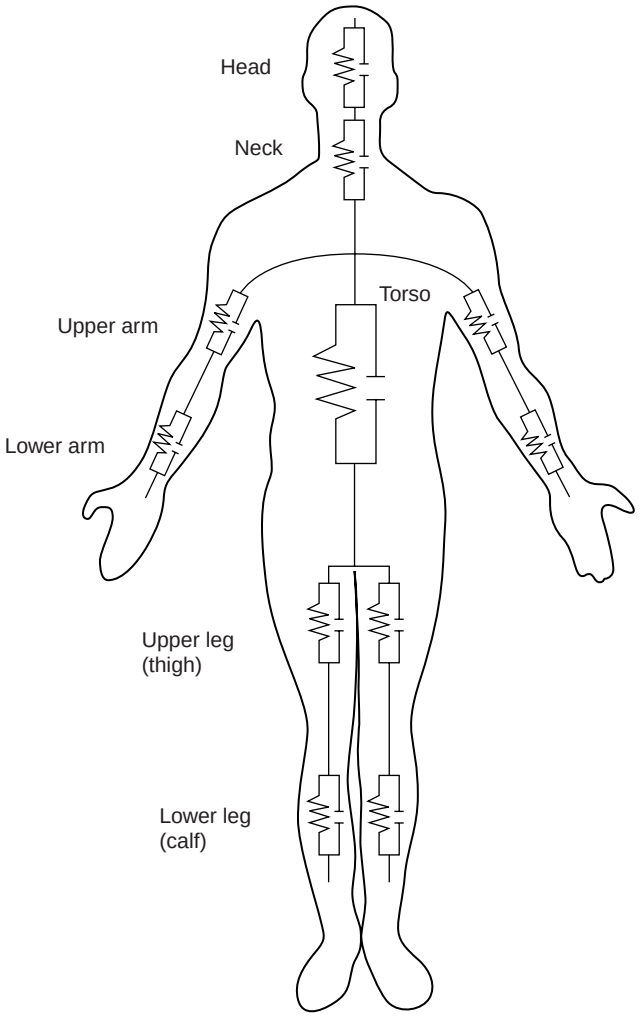


Figure 5. Schematic presentation of the whole body subdivided into 11 segments that are modeled by *RC*-networks.

A special objective of intravenous infusion therapy with hyperosmolaric fluid is the removal of fluid from the extravascular space in order to avoid the emergence of edemas. In the brain, such infusion may help to lower the intracranial pressure that is caused by edema.

Body Composition

Most frequently this method is called bioelectric impedance analysis (BIA). The interest in this methodological approach has increased with the increased interest in healthy life style in industrialized populations since the first reports at the end of the 1980s (35,36).

The body consists of different tissues, for example, muscle, fat, bones, skin, the nervous system, and connective tissue. All these tissues contain intra- and extravascular spaces. The extravascular space can be subdivided into the extracellular or interstitial space and the intracellular space. Chemically, the body consists of water, proteins, carbohydrates, lipids, ions, rare elements, and some other substances. The main objective of body composition analysis

is the assessment of total body water, of lean or fat free tissue mass, and of fat tissue. Those measurements have clinical relevance, but they are also gaining growing attention in the field of sports, physical fitness, wellness, nutrition, and for life science in aerospace research.

The methodological approaches are different, for example bipolar or tetrapolar mode, single-or multiple-frequency measurement, whole body or segmental measurement. In single-frequency measurements, a sinussoidal current with usually < 1 mA (range 0.02–10 mA) and a frequency with typically 50 kHz (range 20–100 kHz) is employed. The basic assumption is that resistivity of fat tissue is higher than that of the so-called lean tissue that has a higher content of water. In the nonclinical field, the determination of body composition is frequently combined with the measurement of body weight using a scale with two integrated electrodes in the foot platform. In that case, however, only the two legs and the lower part of the trunk will be included into the estimation. In more advanced devices, additional electrodes are available in the form of hand grips. Detailed segmental measurements are more valid since the trunk of the body may contribute $\sim$70% to the total body weight and up to 90% to the total body fat, but only 5% to the measured whole body impedance. More sophisticated equipment utilize multifrequency measurement (11).

The applied statistically based predictive equations are of the linear regression type and consider individual parameters like sex, age, height, and weight. They are primarily used to assess the lean muscle mass, the body fat, the fat free mass, the water content, and the body mass index. For physically based segmental calculations, the extremities, the body trunk, and even the head are modeled by cyclindric shape with uniform cross-section over the total length.

Competitive methods include antropometric measures like the girth, simple skin-fold measurements by mechanical calipers, but also highly advanced methods, such as NMR, X ray (dual energy X-ray absorptiometry, or nuclear imaging), and for certain purposes even the hydrostatic weighing in a water tank that is assumed to be the most accurate method.

Laboratory Applications

Blood Cell Counting. The employed methodological principle is frequently called the Coulter principle (37). Blood cells (e.g., erythrocytes, leucocytes) are passing through a capillary (diameter < 100 μm) filled with blood plasma. For frequencies below the β-dispersion the impedance of the cell is higher than that of the surrounding plasma. Consequently, the passage of each cell affects the recorded impedance. Those impedance changes are used for cell counting. In sophisticated devices the impedance changes are quantitatively measured and allow cell volume estimation, which also renders possible the determination of the cell volume distribution function (frequently called the Price–Jones distribution function with the cell diameter as variable). Since the number of leucocytes is very small compared with that of the erythrocytes (usually $< 0.1\%$), leucocytes do not really disturb the counting of erythrocytes. In contrast, however, the erythrocytes must be destroyed before the leucocytes can be counted. This is usually achieved by chemical hemolysis of the erythrocytes. Furthermore, chemical substances are utilized to render possible the differentiation between the different populations of leucocytes (e.g., granulocytes, monocytes, lymphocytes).

Hematocrit. The objective is the estimation of the intracellular volume of blood. The hematocrit is defined as the ratio of the volume of blood cells to total blood volume, although frequently the hematocrit is taken as a measure for the ratio of only the volume of erythrocytes to total blood volume. However, since the partial volume of all other blood cells (i.e., leucocytes, platelets) is very small compared with that of the erythrocytes, the results are not distinctly different. Determination of the ratio between the extracellular volume and the total blood volume is possible by application of at least one measuring frequency below and another one above the β-disperion. Problems that need further consideration as possible sources of error are the electrolytic conductivity of blood plasma, which is dependent on the actual concentration of ions, and the sedimentation of the cells as a consequence of their higher specific weight that may take place during the measuring period. The measurement requires the withdrawal of a sample of blood from a vein or another blood vessel. Measurement of the hematocrit can be achieved either with the impedance or the dielectric technique (38).

Several modifications of the impedance method for the noninvasive determination of the hematocrit have been proposed. In one approach, impedance measurement is performed at the finger by application of two different frequencies (i.e., 100 kHz and 10 MHz). The hematocrit is determined by an algorithm that uses both the pulsatile and the baseline component of both measuring frequencies (39). In another remarkable methodological approach, the patient puts a finger in a temperature-stabilized bath. A fluid with gradually increasing ionic concentration is pumped through the bath chamber, thus leading to a decrease in the impedance of the bath fluid. The pulsatile volume changes of the finger are recorded by impedance measurement in the bath chamber. These pulsatile impedance fluctuations disappear only if the impedance of the bath fluid is identical with that of the blood in the finger. The conclusion is made from the actual impedance of the bath fluid on the hematocrit (40).

Others

Cell Imaging. Nanotechnology-based on bioimpedance sensing, chip devices have been described that allows us to rapidly detect and image cells with a specific phenotype in a heterogeneous population of cells (41). This might be useful for screening purposes, recognition of cell irregularities, and detection of risk patients like human immunodeficiency virus (HIV)-infected individuals. Cell identification is made possible by administration of marker substances. A promising measurement procedure is electrochemical cyclic voltammetry. When a sinusoidal voltage with constant amplitude and a variable frequency in the range of some kilohertz is applied, the impedance is plotted

in the spectrographic mode. Among other effects, volume changes might be the dominating measured effect if the marker binds with a receptor in the cell membrane, and herewith affects membrane properties like its permeability for water or ions or the transmembraneous active ion transport mechanisms.

Inductive Plethysmography. Inductive plethysmography is employed for respiratory monitoring, so-called respiratory inductance plethysmography (RIP). It is based on the measurement of the thoracic crosssection that is enclosed by coils and includes both the rib cage and abdominal compartments (42–45). In the medium frequency range (30–500 kHz), changes in the volume are monitored by the influenced inductance. In the higher frequency range (~100 MHz), the inductively provoked signals (i.e., eddy currents depending on the alveolar air filling) are recorded by appropriately arranged coils.

Magnetic Susceptibility Plethysmography. Magnetic susceptibility plethysmography is a contactless method. It is based on the application of a strong magnetic field and monitors the variation of the magnetic flux. The measurement is accomplished with superconduting quantum interference device (SQUID) magnetometers. This approach may primarily be utilized for the assessment of blood volume changes in the thorax, but until now it is not employed for clinical routine (46).

SUMMARY

In comparison with biomedical engineering as a recognized discipline, the research activities in the field of bioimpedance are much older. It can be assumed that Nikola Tesla, a former student of physics in Graz (Austria) and the inventor of the ac technology, already knew about the passive electrical properties of biological tissues when he demonstrated his famous and public performances with the administration of high voltage pulses "···I demonstrated that powerful electrical discharges of several hundred thousand volts which at that time were considered absolutely deadly, could be passed through the body without inconvenience or hurtful consequences" in the 1880s. This knowledge was utilized by d'Arsonval since 1892 for therapeutic purposes, mainly aiming for heat induction in certain parts of the body. In 1913, Rudolf Hoerber, at that time a physiologist at the University of Kiel (Germany), measured the electrical conductance of frog muscle at 7 MHz and found that at this frequency the membrane resistance is short circuited.

Since its beginning, bioimpedance remained to be a challenge to physicists, medical doctors, and of course engineers. The most relevant basic research was performed in the second half of the twentieth century. The progress that has been reached has been and is still utilized both for diagnostic and therapeutic purposes in medicine. Impedance plethysmography is one of the different fields of bioimpedance application. If impedance plethysmography is correctly understood, it does not only mean the determination of a solid volume with well-defined boundaries, but also the volumetric determination of one component contained in a mixture of different components.

Progress in technology has rendered possible applications that are of great interest for medicine. The most relevant progress is in the field of signal acquisition, including advanced electrode technology, signal processing, and model-based signal interpretation. Not all attempts to utilize the passive electrical properties of biological tissue for diagnostic purposes have been successful. In many cases, other technologies have been shown to be superior. But there is no doubt that the whole potential of impedance plethysmography has not been exhausted. New challenges in the medical field are cellular imaging and possibly even molecular imaging. In all applications, however, impedance plethysmography will have to prove its validity and efficiency.

BIBLIOGRAPHY

1. Nyboer J, Bango S, Nims LF. The Impedance Plethysmograph and Electrical Volume Recorder. CAM Report OSPR 1943; 149.
2. Schwan HP. Electrical properties of body tissues and impedance plethysmography. IRE Trans Biomed Electron 1955; 3:32–46.
3. Stewart GN. Researches on the circulation time in organs and on the influence which affect it. J Physiol 1894;15:1–89.
4. Li J. Multifrequente Impedanztomographie zur Darstellung der elektrischen Impedanzverteilung im menschlichen Thorax. PhD thesis, University of Stuttgart, 2000. Available at http://elib.uni-stuttgart.de/opus/volltexte/2000/736/pdf/li_diss.pdf.
5. Kanai H, Haeno M, Sakamoto K. Electrical measurement of fluid distribution in legs and arms. Med Prog Technol 1987;12:159–170.
6. Osten H. Impedanz-Plethysmographie im Orthostasetest. Münchn Med Wochenschr 1977;119:897–900.
7. Gersing E. Measurement of electrical impedance in organs. Biomed Techn 1991;36:6–11.
8. Dzwonczyk R, et al. Myocardial electrical impedance responds to ischemia in humans. IEEE Trans Biomed Eng 2004;BME-51:2206–2209.
9. Durney CH, Massoudi H, Iskander MF. Radiofrequency Radiation Dosimetry. 4th ed. USAFSAM-TR-85–73; 1985.
10. Gabriel S. 1997. Appendix B: Part 1: Literature Survey. Available at http://niremf.ifac.cnr.it/docs/DIELECTRIC/AppendixB1.html.
11. Segal KR. Estimation of extracellular and total body water by multiple-frequency bioelectrical impedance measurement. Am J Clin Nutr 1991;V54-1:26–29.
12. Neuman MR. Biopotential electrodes. In: Webster JG, editor. Medical Instrumentation—Application and Design. 3rd ed. New York: Wiley; 1998.
13. Rigaud B, Morucci J-P, Chauveau N. Bioimpedance measurement. In: Morucci J-P, et al. edition. Bioelectrical Impedance Techniques in Medicine. Crit. Rev. Biomed. Eng. 1996; 24: 257–351.
14. Yamakoshi KI, Shimazu H, Togawa T, Ito H. Admittance plethysmography for accurate measurement of human limb blood flow. Am J Physiol 1978;235:H821–H829.
15. Shimazu H, et al. Evaluation of the parallel conductor theory for measuring human limb blood flow by electrical admittance plethysmography. IEEE Trans Biomed Eng 1982;BME-29: 1–7.
16. Fricke H, Morse S. The electrical resistance of blood between 800 and 4.5 million cyclces. J Gen Physiol 1925;9: 153–157.

17. Kubicek WG, et al. Development and evaluation of an impedance cardiac output system. Aerospace Med 1966;37:1208–1212.
18. Nyboer J. Electrical impedance plethysmography: a physical and physiological aproach to peripheral vascular studies. Circulation 1950;2:811–821.
19. Yamamoto Y, Yamamoto T, Öberg PA. Impedance plethysmography in human limbs. Part 1: On electrodes and electrode geometry. Med Biol Eng Comput 1991;29:419–424.
20. Yamamoto Y, Yamamoto T, Öberg PA. Impedance plethysmography in human limbs. Part 2: Influence of limb crosssectional areas. Med Biol Eng Comput 1992;30:518–524.
21. Lozano A, Rosell J, Pallas-Areny R. Errors in prolonged electrical impedance measurement due to electrode repositioning and postural changes. Physiol Meas 1995;16:121–130.
22. Risacher F, et al. Impedance plethysmography for the evaluation of pulse wave velocity in limbs. Med Biol Eng Comput 1992;31:318–322.
23. Hull R, et al. Impedance plethysmography: The relationship between venous filling and sensitivity and specificity for proximal vein thrombosis. Circulation 1978;58:898–902.
24. Brodie TG, Russel AE. On the determination of the rate of blood flow through an organ. J Physiol 1905;33:XLVII–XLVIII.
25. Seipel JH, Ziemnowicz SAR, O'Doherty DS. Cranial impedance plethysmography—rheoencephalography as a method for detection of cerebrovascular disease. In: Simonson E, McGavack TH, editors. Cerebral Ischemia. Springfield, IL: Charles C Thomas; 1965; p 162–179.
26. Hadjiev D. A new method for quantitative evaluation of cerebral blood flow by rheoencephalography. Brain Res 1968;8:213–215.
27. Netz J, Forner E, Haagemann S. Contactless impedance measurement by magnetic induction—a possible method for investigation of brain impedance. Physiol Meas 1993;14:463–471.
28. Geddes LA, Hoff HE, Hickman DM, Morre AG. The impedance pneumograph. Aerospace Med 1962;33:28–33.
29. Baker LE, Geddes LA, Hoff HE. Quantitative evaluation of impedance spirometry in man. Am J Med Elect 1965;4:73–77.
30. Nopp P, et al. Dielectric properties of lung tissue as a function of air content. Phys Med Biol 1993;38:699–716.
31. Brown BH, et al. Multifrequency imaging and modelling of respiratory related electrical impedance changes. Physiol Meas 1994;15(Suppl. A):1–12.
32. Nierman DM, et al. Transthoracic bioimpedance can measure extravascular water in acute lung injury. J Surg Res 1996;65:101–108.
33. Brack Th et al. Continuous and cooperation-independent monitoring of airway obstruction by a portable inductive plethysmograph. AJRCCM 2004;169:1.
34. Löfgren B. The electrical impedance of a complex tissue and its relation to changes in volume and fluid distribution. Acta Physiol Scand 1951;23(Suppl. 81):1–51.
35. Schloerb PR, et al. Bioimpedance as a measure of total body water and body cell in surgical nutrition. Eur Surg Res 1986;18:1.
36. van Loan MD, et al. Association of bioelectric resistive impedance with fat-free mass and total body water estimates of body composition. Amer J Human Biol 1990;2:219–226.
37. Coulter WH. High speed automatic blood cell counter and cell size analyzer. Proc Nat Electron Conf 1956;12:1034.
38. Treo EF, et al. Comparative analysis of hematocrit measurements by dielectric and impedance techniques. IEEE Trans Biomed Eng 2005;MBE-52:549–552.
39. http://www.patentalert.com/docs/001/z00120411.shtml.
40. Yamakoshi KI, et al. Noninvasive measurement of hematocrit by electrical admittance plethysmography technique. IEEE Trans Biomed Eng 1989;27:156–161.
41. Mishra NN, et al. Bio-impedance sensing device (BISD) for detection of human CD4+ cells. Nanotech 2004, vol. 1, Proc 2004 NSTI Nanotechnology Conf 2004, p 228–231.
42. Brouillette RT, Morrow AS, Weese-Mayer DE, Hunt CE. Comparison of respiratory inductive plethysmography and thoracic impedance for apnea monitoring. J Ped 1987;111:377–383.
43. Valta P, et al. Evaluation of respiratory inductive plethysmography in the measurement of breathing pattern and PEEP-induced changes in lung volume. Chest 1992;102:234–238.
44. Cohen KP, et al. Design of an inductive plethysmograph for ventilation measurement. Physiol Meas 1994;15:217–229.
45. Strömberg TLT. Respiratory Inductive Plethysmography. Linköping Studies in Science and Technologies Dissertations No. 417, 1996.
46. Malmivuo J, Plonsey R. Impedance plethysmography. In: Malmivuo J, Plonsey R, editors, Bioelectromagnetism. New York: Oxford University Press; 1995. Chapt. 25.

Further Reading

Foster KR, Schwan HP. Dielectric properties of tissue and biological materials. A critical review. Crit Rev Biomed Eng 1989;17:25–102.
Kaindl F, Polzer K, Schuhfried F. (1958, 1966, 1979): Rheographie. Darmstadt, Dr: Dietrich Steinkopff Verlag; 1958 (1st ed), 1966 (2nd ed), 1979 (3rd ed).
Morucci J-P, et al. Bioelectrical impedance techniques in medicine. Crit Rev Biomed Eng 1996;24:223–681.
Pethig R. Dielectric and Electronic Properties of Biological Materials. Chichester: Wiley; 1979.
Schwan HP. Determination of biological impedances. In: Nastuk WL, editor. Physical Techniques in Biological Research. Vol. VI (ptB) New York: Academic Press; 1963; p 323–407.
Schwan HP. Biomedical engineering: A 20th century interscience. Its early history and future promise. Med Biol Eng Comput 1999;37(Suppl. 2):3–13.
Stuchly MA, Stuchly SS. Dielectric properties of biological substances—tabulated. J Microw Power 1980;15:19–26.
Webster JG, editor. Medical Instrumentation—Application and Design. 3rd ed. New York: Wiley; 1998.
Gabriel C, Gabriel S. Compilation of the Dielectric Properties of Body Tissues at RF and Microwave Frequencies, 1996. Available at http://www.brooks.af.mil/AFRL/HED/hedr/reports/dielectric/Report/Report.html.

See also BIOIMPEDANCE IN CARDIOVASCULAR MEDICINE; PERIPHERAL VASCULAR NONINVASIVE MEASUREMENTS.

IMPEDANCE SPECTROSCOPY

BIRGITTE FREIESLEBEN DE BLASIO
JOACHIM WEGENER
University of Oslo
Oslo, Norway

INTRODUCTION

Impedance spectroscopy (IS), also referred to as electrochemical impedance spectroscopy (EIS), is a versatile approach to investigate and characterize dielectric and conducting properties of materials or composite samples (1). The technique is based on measuring the impedance

(i.e., the opposition to current flow) of a system that is being excised with weak alternating current or voltage. The impedance spectrum is obtained by scanning the sample impedance over a broad range of frequencies, typically covering several decades.

In the 1920, researchers began to investigate the impedance of tissues and biological fluids, and it was early known that different tissues exhibit distinct dielectric properties, and that the impedance undergoes changes during pathological conditions or after excision (2,3).

The advantage of IS is that it makes use of weak amplitude current or voltage that ensures damage-free examination and a minimum disturbance of the tissue. In addition, it allows both stationary and dynamic electrical properties of internal interfaces to be determined, without adversely affecting the biological system. The noninvasive nature of the method combined with its high information potential makes it a valuable tool for biomedical research and many medical applications are currently under investigation and development; this will be reviewed at the end of the article.

This article starts with providing a general introduction to the theoretical background of IS and the methodology connected to impedance measurement. Then, the focus will be on applications of IS, particularly devises for *in vitro* monitoring of cultured cell systems that have attracted widespread interest due to demand for noninvasive, marker-free, and cost-effective methods.

THEORY

Impedance, $\boldsymbol{Z}$, is a complex-valued vector that describes the ability of a conducting medium to resist flow of alternating current (ac). In a typical IS experiment (Fig. 1), a sinusoidal current $\boldsymbol{I}(\boldsymbol{t})$ signal with angular frequency ω ($\omega = 2\pi f$) is passed through the sample and the resulting steady-state voltage $\boldsymbol{U}(\boldsymbol{t})$ from the excitation is measured. According to the ac equivalent of Ohm's law, the impedance is given by the ratio of these two quantities

$$Z = \frac{U(t)}{I(t)} \tag{1}$$

The impedance measurement is conducted with use of weak excitation signals, and in this case the voltage response will be sinusoid at the same frequency ω as the applied current signal, but shifted in phase φ. Introducing complex notation, Eq. 1 translates into

$$Z = \frac{U_0}{I_0}\exp(i\varphi) = |Z|\exp(i\varphi) \tag{2}$$

with U_0 and I_0 being the amplitudes of the voltage and current, respectively, and $i = \sqrt{-1}$ being the imaginary unit. Thus, at each frequency of interest the impedance is described by two quantities: (*1*) its magnitude $|Z|$, which is the ratio of the amplitudes of $\boldsymbol{U}(\boldsymbol{t})$ and $\boldsymbol{I}(\boldsymbol{t})$; and (*2*) the phase angle φ between them. The impedance is measured over a range of frequencies between hertz and gigahertz, dependent on the type of sample and the problem to study.

The measured impedance can be divided into its real and imaginary components, that is, the impedance contribution arising from current in-phase with the voltage and 90° out-of-phase with the voltage

$$R = \mathrm{Re}(Z) = |Z|\cos(\varphi), \quad X = \mathrm{Im}(Z) = |Z|\sin(\varphi) \tag{3}$$

The real part is called resistance, R, and the imaginary part is termed reactance, X. The reactive impedance is caused by presence of storage elements for electrical charges (e.g., capacitors in electrical circuit).

In some cases it is convenient to use the inverse quantities, which are termed admittance $\boldsymbol{Y} = 1/\boldsymbol{Z}$, conductance $\boldsymbol{G} = \mathrm{Re}(\boldsymbol{Y})$, and susceptance $\boldsymbol{B} = \mathrm{Im}(\boldsymbol{Y})$, respectively. In the linear regime (i.e., when the measured signal is proportional to the amplitude of the excitation signal), these two representation are interchangeable and contain the same information. Thus, IS is also referred to as admittance spectroscopy.

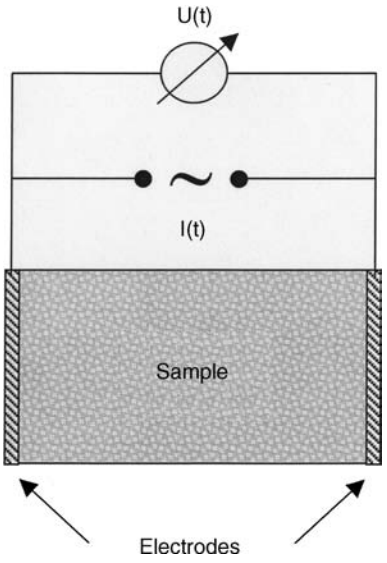


Figure 1. Schematic of a two-electrode setup to measure the frequency-dependent impedance of a sample that is sandwiched between two parallel plate electrodes.

INSTRUMENTATION

The basic devices for conducting impedance measurements consist of a sinusoid signal generator, electrodes, and a phase-sensitive amplifier to record the voltage or current. Commonly, a four-electrode configuration is used, with two current injecting electrodes and two voltage recording electrodes to eliminate the electrode–electrolyte interface impedance. As discussed below, some applications of IS make use of two-electrode arrangements in which the same electrodes are used to inject current and measure the voltage.

Since the impedance is measured by a steady-state voltage during current injection, some time is needed when changing the frequency before a new measurement can be performed. Therefore, it is very time consuming if each frequency has to be applied sequentially. Instead, it is common to use swept sine wave generators, or spectrum analyzers with transfer function capabilities and a white noise source. The white noise signal consists of the

superposition of sine waves for each generated frequency, and the system is exposed to all frequencies at the same time. Fourier analysis is then used to extract the real and imaginary parts of the impedance.

The electrodes used for impedance experiments are made from biocompatible materials, such as noble metals, which in general is found not to have deleterious effect on biological tissue function. Electrode design is an important and complicated issue, which depends on several factors, including the spatial resolution required, the tissue depth, and so on. It falls beyond the scope of this article to go further into details. The interested readers are referred to the book by Grimnes and Martinsen (4) for a general discussion.

Common error sources in the measurements include impedance drift (e.g., caused by adsorption of particles on the electrodes or temperature variations). Ideally, the system being measured should be in steady-state throughout the time required to perform the measurement, but in practice this can be difficult to achieve. Another typical error source is caused by pick-up of electrical noise from the electronic equipment, and special attention must be paid to reduce the size of electric parasitics arising, for example, from cables and switches.

DATA PRESENTATION AND ANALYSIS

The most common way to analyze the experimental data is by fitting an equivalent circuit model to the impedance spectrum. The model is made by a collection of electrical elements (resistors, capacitors) that represents the electrical composition in the system under study.

As a first step, it is useful to present the measured data by plotting log $|Z|$ and φ versus log f in a so-called Bode-diagram (Fig. 2a), and by plotting $\mathrm{Im}|Z|$ versus $\mathrm{Re}|Z|$ named a Nyquist diagram or an *impedance locus* (Fig. 2b). The examples provided in Fig. 2 are made for an electrical circuit (insert Fig. 2b). While the first way of presenting the data shows the frequency-dependence explicitly, the phase angle φ is displayed in the latter.

The impedance spectrum gives many insights to the electrical properties of the system, and with experience it is possible to make a qualified guess of a proper model based on the features in the diagrams (cf. Fig. 4). Similar to other spectroscopic approaches like infrared (IR) or ultraviolet (UV)/visible (vis), the individual components tend to show up in certain parts of the impedance spectrum. Thus, variations in the values of individual components alter the spectrum in confined frequency windows (Fig. 3).

For a given model, the total impedance (transfer function) is calculated from the individual components with use of Ohm's and Kirchhoff's laws. The best estimates for the parameters, that is, the unknown values of the resistors and capacitors in the circuit, are then computed with use of least-square algorithms. If the frequency response of the chosen model fits the data well, the parameter values are used to characterize the electrical properties of the system.

In order to fit accurately the equivalent circuit model impedance to the impedance of biomaterials, it is often necessary to include nonideal circuit elements, that is,

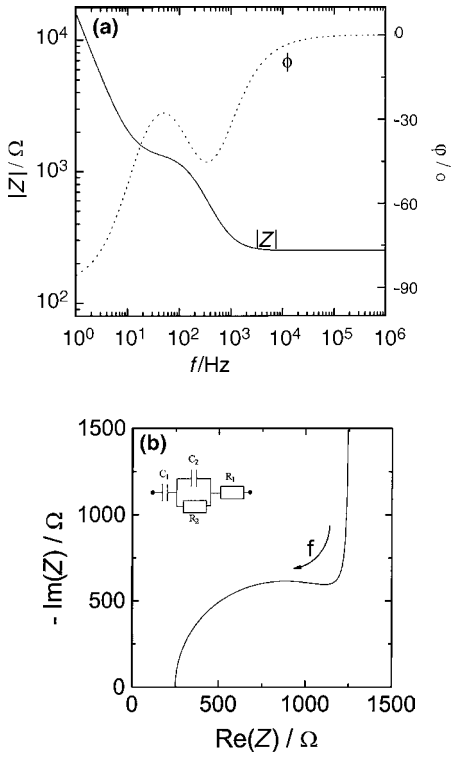


Figure 2. Different representations of impedance spectra. (a and b) visualize the frequency-dependent complex impedance of the electrical circuit shown in (b) with the network components: $R_1 = 250\ \Omega$; $R_2 = 1000\ \Omega$; $C_1 = 10\ \mu\mathrm{F}$; $C_2 = 1\ \mu\mathrm{F}$. (a) Bode-diagram presenting frequency-dependent impedance magnitude $|Z|$ together with the phase shift φ of the sample under investigation. (b) Wessel diagram locus of the same electrical network. The imaginary component of the impedance (reactance) is plotted against the real component (resistance). The arrow indicates the direction along which the frequency increases.

elements with frequency dependent properties. Such elements are not physically realizable with standard technical elements. Table 1 provides a list of common circuit elements that are used to describe biomaterials with respect to their impedance and phase shift. The constant phase element (CPE) portrays a nonideal capacitor, and was originally introduced to describe the interface impedance of noble metal electrodes immersed in electrolytic solutions (5). The physical basis for the CPE in living tissue (and at electrode interfaces) is not clearly understood, and it is best treated as an empirical element. Another example is the Warburg impedance σ that accounts for the diffusion limitation of electrochemical reactions (4).

It is important to place a word of caution concerning the equivalent circuit modeling approach. Different equivalent circuit models (deviating with respect to components or in the network structure) may produce equally good fits to the experimental data, although their interpretations are very different. It may be tempting to increase the number of elements in a model to get a better agreement between experiment and model. However, it may then occur that the model becomes redundant because the components cannot be quantified independently. Thus, an overly complex

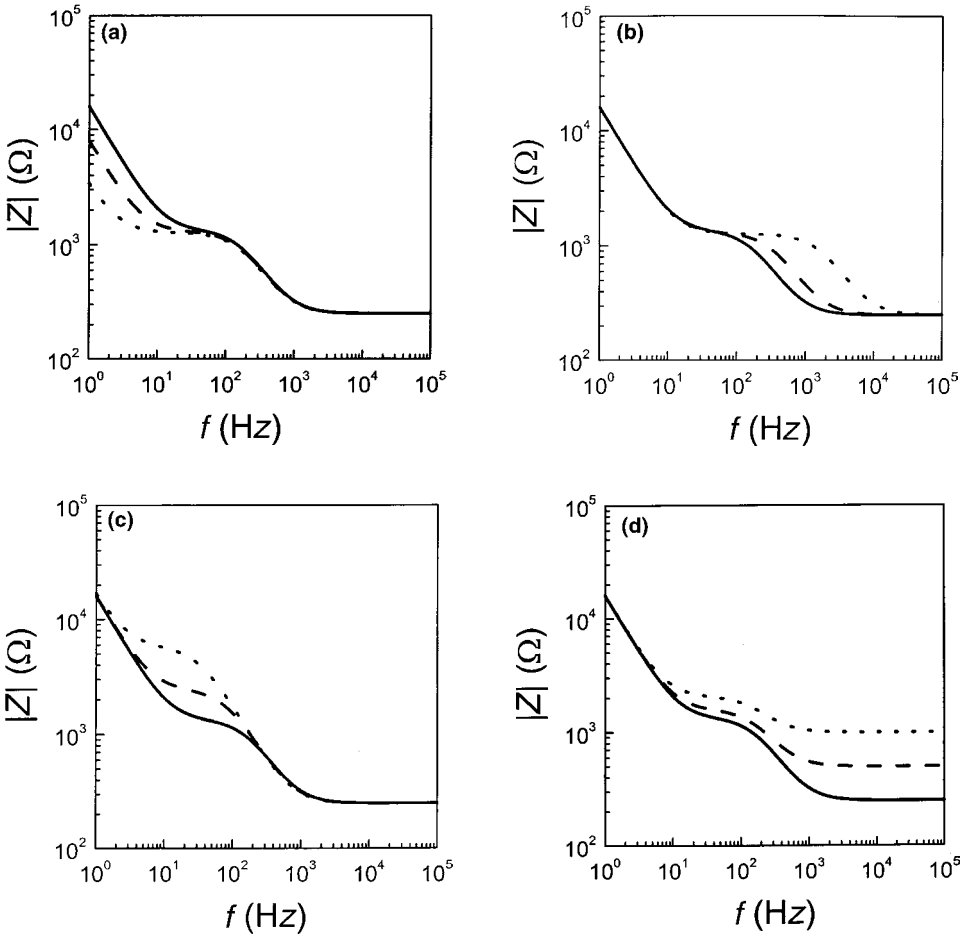


Figure 3. Calculated spectra of the impedance magnitude $|Z|$ for the electrical circuit shown in the insert of Fig. 2b when the network parameters have been varied individually. The solid line in each figure corresponds to the network parameters: $R_1 = 250\ \Omega$; $R_2 = 1000\ \Omega$; $C_1 = 10\ \mu F$; $C_2 = 1\ \mu F$. (a) Variation of C_1: 10 μF (solid), 20 μF (dashed), 50 μF (dotted). (b) Variation of C_2: 1 μF (solid), 0.5 μF (dashed), 0.1 μF (dotted). (c) Variation of R_1: 1000 Ω (solid), 2000 Ω (dashed), 5000 Ω (dotted). (d) Variation of R_2: 250 Ω (solid), 500 Ω (dashed), 1000 Ω (dotted).

model can provide artificially good fits to the impedance data, while at the same time highly inaccurate values for the parameters. Therefore, it is sound to use equivalent circuits with a minimum number of elements that can describe all aspects of the impedance spectrum (6).

Alternatively, the impedance data can be analyzed by deriving the current distribution in the system with use of differential equations and boundary values (e.g., the given excitation at the electrode surfaces). The parameters of the model impedances are then fitted to the data like described above. An example of this approach is be presented below, where it is used to analyze the IS of a cell-covered gold film electrode.

IMPEDANCE ANALYSIS OF TISSUE AND SUSPENDED CELLS

The early and pioneering work on bioimpedance is associated with the names of Phillipson, et al. (5). In these studies blood samples or pieces of tissue were examined in an experimental setup as shown in Fig. 1, and the dielectric properties of the biological system were investigated over a broad range of frequencies from hertz to gigahertz.

To understand the origin of bioimpedance, it is necessary to look at the composition of living material. Any tissue is composed of cells that are surrounded by an extracellular fluid. The extracellular medium contains proteins and polysaccharides that are suspended in an ionic solution and the electrical properties of this fluid are determined by the mobility and concentration of the ions, primarily Na^+ and Cl^-. The cell membrane marks the boundary between the interior and exterior environment, and consists of a 7–10 nm phospholipid bilayer. The membrane allows diffusion of water and small nonpolar molecules, while transport of ions and polar molecules requires the presence of integral transport proteins. On the inside, the cell contains a protein-rich fluid with specialized membrane-bound organelles, like the nucleus. For most purposes the fluid behaves as a pure ionic conductor. Thus, the cell membrane is basically a thin dielectric sandwiched between two conducting media and in a first approximation its impedance characteristics are mainly capacitive.

The simplest possible explanatory model for biological tissue (Fig. 4a-1) therefore consists of two membrane capa-

Table 1. Individual Impedance Contributions of Ideal and Empirical Equivalent Circuit Elements

Component of Equivalent Circuit	Parameter	Impedance, Z	Phase Shift, φ
Resistor	R	R	0
Capacitor	C	$(i\omega C)^{-1}$	$-\pi/2$
Coil	L	$i\omega L$	$+\pi/2$
Constant phase element (CPE)	$\alpha(0 \le \alpha \le 1)$	$1/(iC\omega)\alpha \cdot$	$-\alpha\pi/2$
Warburg impedance, σ	σ	$\sigma(1-i)\omega^{-0.5}$	$-\pi/4$

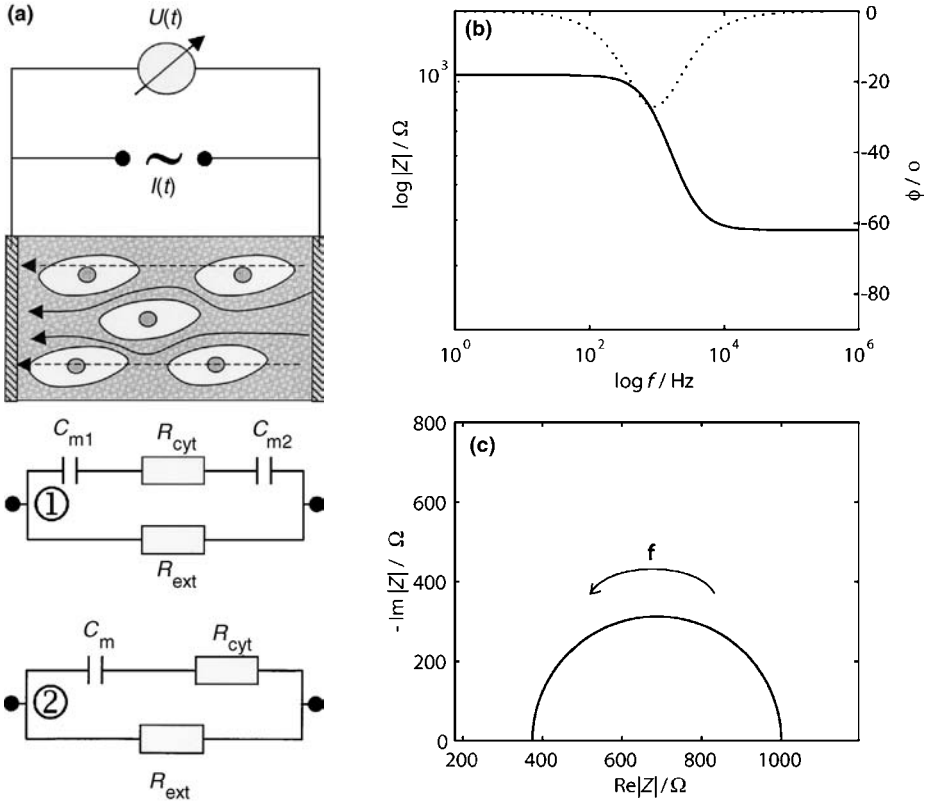


Figure 4. (a) Schematics of the impedance measurement on living tissue. The arrows indicate the pathway of current flow for low frequencies (solid line) and high frequencies (dashed). Only at high frequencies the current flows through the cells. The electrical structure of tissue can be directly translated into equivalent circuit **1**, which can be simplified to equivalent circuit **2**. (b) Bode-diagram for network 2 in Fig. 4a, using $R_{ext} = 1000\ \Omega$; $R_{cyt} = 600\ \Omega$; $C_m = 100$ nF. (c) Impedance locus generated with the same values.

citors C_{m1} and C_{m2} in series with a resistor of the cytosolic medium R_{cyt}, which in turn acts in parallel with the resistance of the conducting extracellular medium R_{ext}. Since the two series capacitances of the membrane cannot be determined independently, they are combined to an overall capacitance C_m. The equivalent circuit model of this simple scenario (Fig. 4a-2) gives rise to a characteristic impedance spectrum, as shown with the Bode-diagram (Fig. 4b) and the Nyquist diagram (Fig. 4c). Impedance data for biological tissue is also often modeled by the so-called Cole–Cole equation (7)

$$Z = R_\infty + \frac{\Delta R}{1 + \Delta R(i\omega C)^\alpha} \quad \Delta R = R_0 - R_\infty \tag{4}$$

This simple empirical model is identical to the circuit of Fig. 4, except that the capacitor is replaced by a CPE element. The impedance spectrum is characterized by four parameters $(\Delta R, R_\infty, \alpha, \tau)$, where R_0, R_∞ is the low- and high frequency intercepts on the x axis in the Nyquist plot (cf. Fig. 4d), τ is the time constant $\tau = \Delta R \cdot C$, and α is the CPE parameter. The impedance spectrum will be similar to Fig. 4b, c, but when $\alpha \neq 1$, the semicircle in the Nyquist diagram is centered below the real axis, and the arc will appear flattened. For macroscopically heterogeneous biological tissue, the transfer function is written as a sum of Cole–Cole equations.

The features of the impedance spectrum Fig. 4b can be intuitively understood: at low frequencies the capacitor prevents current from flowing through R_{cyt} and the measured impedance arises from R_{ext}. At high frequencies, with the capacitor having a very low impedance, the current is free to flow through both R_{cyt}, R_{ext}. Thus, there is a transition from constant-level impedance at low frequencies to another constant level. This phenomenon is termed dispersion, and will be discussed in the following.

A homogenous conducting material is characterized by a bulk property named the resistivity ρ' having the dimensions of ohm centimeters (Ω·cm). Based on this intrinsic parameter, the resistance may be defined by

$$R = \frac{\rho' L}{A} \tag{5}$$

where A is the cross-sectional area and L is the length of the material. Thus, by knowing the resistivity of the material and the dimensions of the system being studied, it is possible to estimate the resistance. Similarly, a homogeneous dielectric material is characterized by an intrinsic property called the relative permittivity ε', and the capacitance is defined by

$$C = \frac{\varepsilon' \varepsilon_0 A}{d} \tag{6}$$

where ε_0 is the permittivity of free space with dimension F/m, and A, d are the dimensions of the system as above. For most biological membranes, the area-specific capacitance is found to be quite similar, with a value of $\sim 1\ \mu F \cdot cm^{-2}$ (8).

For historical reasons the notation of conductivity σ' with dimensions $S \cdot m^{-1}$ and conductance ($G = \sigma' A/d$) has been preferred over resistance R and resistivity ρ, but the information content is the same, it is just expressed in a different way.

It is possible to recombine ε' and σ' by defining a complex permittivity $\varepsilon = \varepsilon' + \varepsilon''$, with $\mathrm{Re}(\varepsilon) = \varepsilon'$ and $\mathrm{Im}(\varepsilon) = \varepsilon''$. The

imaginary part accounts for nonideal capacitive behavior, for example, current within the dielectric due to bound charges giving rise to a viscous energy loss (dielectric loss). Therefore, ε'' is proportional to σ', when adjusted for the conductivity that is due to migration σ_0 (9)

$$\varepsilon'' = \frac{\sigma' - \sigma_0}{2\pi f \varepsilon_0} \tag{7}$$

When a piece of biological material is placed between two electrodes, it is possible to measure the capacitance of the system and thereby to estimated the tissue permittivity ε'. In general, ε' quantifies the ratio of the capacitance when a dielectric substance is placed between the electrodes, relative to the situation with vacuum in between. The increase of capacitance upon insertion of a dielectric material is due to polarization in the system in response to the electric field. For direct current (dc) or low frequency situations ε' is called the dielectric constant. When the frequency is increased, ε' often shows strong frequency dependence with a sigmoid character in a log–log plot of ε' versus frequency. This step-like decrease of the permittivity is referred to as a dielectric dispersion. The frequency f_c at which the transition is half-complete is called the characteristic frequency, and is often expressed as time constant τ with

$$\tau = \frac{1}{f_c} \tag{8}$$

Going back to Fig. 4c, the characteristic frequency is found directly as the point when the phase angle is at maximum.

The origin of dielectric dispersion in a homogeneous material is due to a phenomenon termed orientation polarization. Dipolar species within the material are free to move and orient themselves along the direction of the field, and therefore they contribute to the total polarization. However, when the frequency becomes too high, the dipoles can no longer follow the oscillation of the field, and their contribution vanishes. This relaxation causes the permittivity ε' to decrease.

For heterogeneous samples like tissue additional relaxation phenomena occur, leading to more complex frequency dependence. In 1957, Schwan (10) defined three prominent dispersion regions of relevance for bioimpedance studies called α, β, and γ, which is shown in Fig. 5. The dispersions are generally found in all tissue, although the time constant and the change in permittivity $\Delta\varepsilon'$ between the different regions may differ (9).

Briefly stated, the α-dispersion originates from the cloud of counterions that are attracted by surface charges of the cell membrane. The counterions can be moved by an external electric field, thereby generating a dipole moment and relaxation. The β-dispersion, which is also called Maxwell–Wagner dispersion, is found in a window between kilohertz and megahertz (kHz and MHz). It arises due to the accumulation of charges at the interface between hydrophobic cell membranes and electrolytic solutions. Since the aqueous phase is a good conductor, whereas the membrane is not, mobile charges accumulate and charge up the membrane capacitor, thus, contributing to polarization. When the frequency gets too high, the charging is not complete, causing a loss of polarization. Finally,

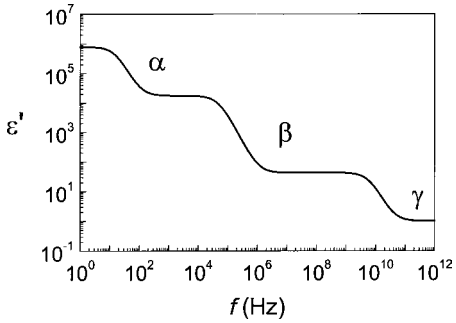


Figure 5. Frequency-dependent permittivity ε' of tissue. The permittivity spectrum $\varepsilon'(f)$ is characterized by three major dispersions: α-, β-, and γ-dispersion.

the γ-dispersion is due to the orientation polarization of small molecules, predominantly water molecules.

Most IS measurements are performed at intermediate frequencies in the regime of the β-dispersion. In this frequency window, the passive electrical properties of tissue are well described with the simple circuit shown in Fig. 4 or by the Cole–Cole equation. The measurements can be used to extract information about extra- and intercellular resistance, and membrane capacitance. For example, it has been shown that cells in liver tissue swell, when the blood supply ceases off (ischemia) and that the swelling of the cells can be monitored as an increase in the resistance of the extracellular space R_{ext} (11). Cell swelling compresses the extracellular matrix around the cells, and thereby narrows the ion pathway in this region. Based on experiments like these, there is a good perspective and prognosis that IS may serve as a routine monitoring tool for tissue vitality even during the surgery.

APPLICATION: MONITORING OF ADHERENT CELLS *IN VITRO*

The attachment and motility of anchorage dependent cell cultures is conveniently studied using a microelectrode setup. In this technique, cells are grown directly on a surface containing two planar metal electrodes, one microelectrode and one much larger counter electrode. The cells are cultured in normal tissue culture medium that serves as the electrolyte.

When current flows between the two electrodes, the current density, and the measured voltage drop, will be much higher at the small electrode. Therefore the impedance measurement will be dominated by the electrode polarization of the small electrode Z_{el}. Instead, no significant polarization takes place at the larger counter electrode and its contribution to the measured impedance may be ignored. The electrode polarization impedance Z_{el} acts physically in series with the resistance of the solution R_{sol}. Since the current density is high in a zone (the constrictional zone) proximal to the microelectrode, the electrolytic resistance will be dominated by the constriction resistance R_c in this region (Fig. 6). The total measured impedance may therefore be approximated by $\mathbf{Z} \sim \mathbf{Z}_{el} + R_c$ (4). If necessary, R_c may be determined from high frequency measurements where the electrode resistance is

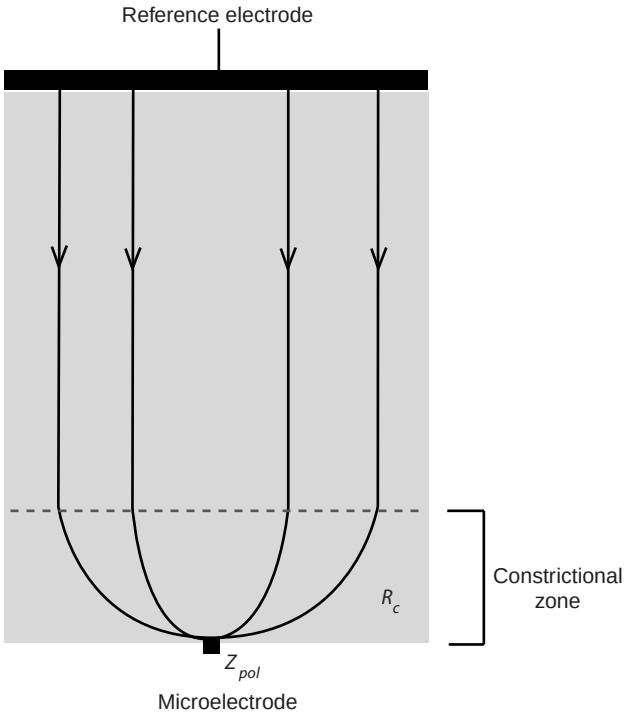


Figure 6. Schematic of two-electrode configuration.

infinitely small, that is, $\mathrm{Re}(Z_{\mathrm{el}}) \to 0$, and subtracted from the measured impedance to determine the impedance of the electrode–electrolyte interface.

When cells adhere on the small electrode, they constrain the current flow from the interface, increasing the measured impedance. The changes mainly reflects the capacitive nature of nonconducting lipid-bilayer membrane surrounding the cells. The cell membranes cause the current field to bend around the cells, much like if they were microscopic insulating particles. It is possible to follow both cell surface coverage and cell movements on the electrode, and morphological changes caused by physiological/pathological conditions and events may be detected. The technique may also be used to estimate cell membrane capacitances, and barrier resistance in epithelial cell sheets. In addition, the method is highly susceptible to vertical displacements of the cell body on the electrode with sensitivity in the nanometer range.

INSTRUMENTATION

The technique was introduced by Giaever and Keese in 1984 and referred to as Electrical Cell-Substrate Impedance Sensing (ECIS) (12,13). The ECIS electrode array consists of a microdisk electrode ($\sim 5 \times 10^{-4}\,\mathrm{cm}^2$) and a reference electrode ($\sim 0.15\,\mathrm{cm}^2$); depending on the cell type to be studied, the recording disk electrode may contain a population of 20–200 cells. The electrodes are made from depositing gold film on a polycarbonate substrate over which an insulating layer of photoresist is deposited and delineated. A 1 V amplitude signal at fixed frequency (0.1–100 kHz) is applied to the electrodes through a large resistor to create a controlled current of $1\,\mu A$, and the corresponding voltage across the cell-covered electrodes

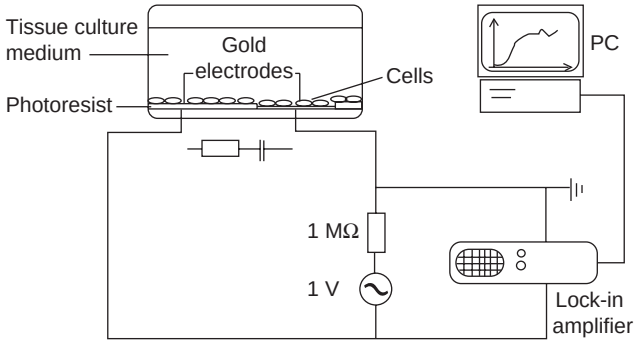


Figure 7. The ECIS measurement setup.

is measured by a lock-in amplifier, which allows amplification of the relatively small signals. The amplifier is interfaced to a PC for data storage. The impedance is calculated from the measured voltage displayed in real time on the computer screen (Fig. 7). During the measurements the sample is placed in an incubator at physiological conditions.

The ECIS system is now commercially available, and the electrode slides allow multiple experiments to be performed at the same time (14). Some modifications to the technique have been described, such as a two-chamber sample well, which permit simultaneous monitoring on a set of empty electrodes being exposed to the same solution (15), platinized single-cell electrodes (15), and inclusion of a voltage divider technique to monitor the impedance across a range of frequencies (16). More recently, impedance studies have been performed using other types of electrode design. One approach has been to insert a perforated silicon-membrane between two platinium electrodes, there by allowing for two separate electrolytic solutions to exist on either side of the membrane (17). The results obtained with these techniques are generally identical to those obtained by the ECIS system.

MODEL OF ELECTRODE–CELL INTERFACE

To interpret ECIS-based impedance data, a model of the ECIS electrode–cell interface has been developed that allows determination of (*1*) the distance between the ventral cell surface and the substratum, (*2*) the barrier resistance, and (*3*) the cell membrane capacitance of confluent cell layers (18). The model treats the cells as disk shaped objects with a radius r_c that are separated an average distance h from the substrate (Fig. 8). When cells cover the electrode, the main part of the current will flow through the thin layer of medium between the cell and the electrode, and leave the cell sheet in the narrow spacing between cells. However, the cell membrane, which is modeled as a capacitor (an insulating layer separating the conducting fluids of the solution and the cytosol) allows a parallel current flow to pass through the cells. The minor resistive component of the membrane impedance due to the presence of ionic channels is ignored in the calculations. By assuming that the electrode properties are not affected by the presence of cells, a boundary-value model of the current flow across the cell layer may be used to derive a relation

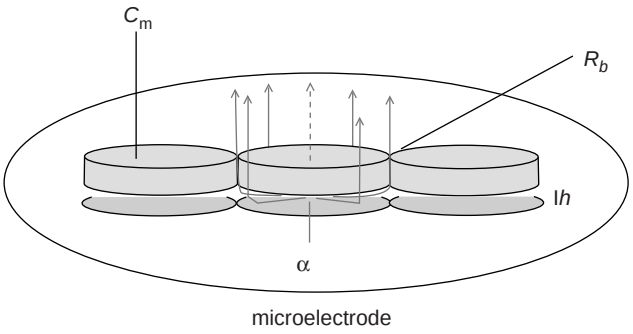


Figure 8. Model of current flow paths. The impedance changes associated with the presence of cells, arise in three different regions: from current flow under the cells quantified by α, from current flow in the narrow intercellular spacings causing the barrier resistance R_b. In parallel, some current will pass through the cell membranes giving rise to capacitive reactance C_m.

between the specific impedance of a cell-covered electrode Z_{cell} and the empty electrode Z_{el}

$$\frac{1}{Z_{cell}} = \frac{1}{Z_{el}}\left(\frac{Z_{el}}{Z_{el}+Z_m} + \frac{\frac{Z_m}{Z_{el}+Z_m}}{\frac{\gamma r_c}{2}\frac{I_0(\gamma r)}{I_1(\gamma r)} + R_b\left(\frac{1}{Z_{el}}+\frac{1}{Z_m}\right)}\right)$$

$$\gamma = \sqrt{\frac{\rho}{h}\left(\frac{1}{Z_{el}}+\frac{1}{Z_m}\right)} \qquad (9)$$

where I_0, I_1 are the modified Bessel functions of the first kind of order zero and one, R_b and ρ are the specific barrier resistance and resistivity of the solution, and $Z_m = -2i/(\omega C_m)$ is the specific membrane impedance of the cells. A parameter $\alpha = r_c(\rho/h)^{0.5}$ is introduced as an assessment of the constraint of current flow under the cells. The impedance spectrum of an empty electrode and a cell-covered electrode is used to fit the three adjustable parameters (R_b, α, C_m).

The model outlined above has been further refined to describe polar epithelial cell sheets, treating separately the capacitance of the apical, basal, and lateral membranes (19). Some applications of the model will be discussed in the following sections.

MONITORING ATTACHMENT AND SPREADING

As a cell comes into contact with a solid surface, it forms focal contacts, primarily mediated by transmembrane proteins that anchor structural filaments in the cell interior to proteins on the substrate. During this process, the rounded cell spreads out and flattens on the surface, greatly increasing its surface area in contact with the electrode. The cell will also establish contacts with neighboring cells through particular cell–cell junctions, such as tight junctions, where strands of transmembrane proteins sew neighboring cells together, and gap junctions formed by clusters of intercellular channels, connecting the cytosol of adjacent cells.

The attachment process is normally studied using single-frequency measurements. Figure 9a and b show Bode

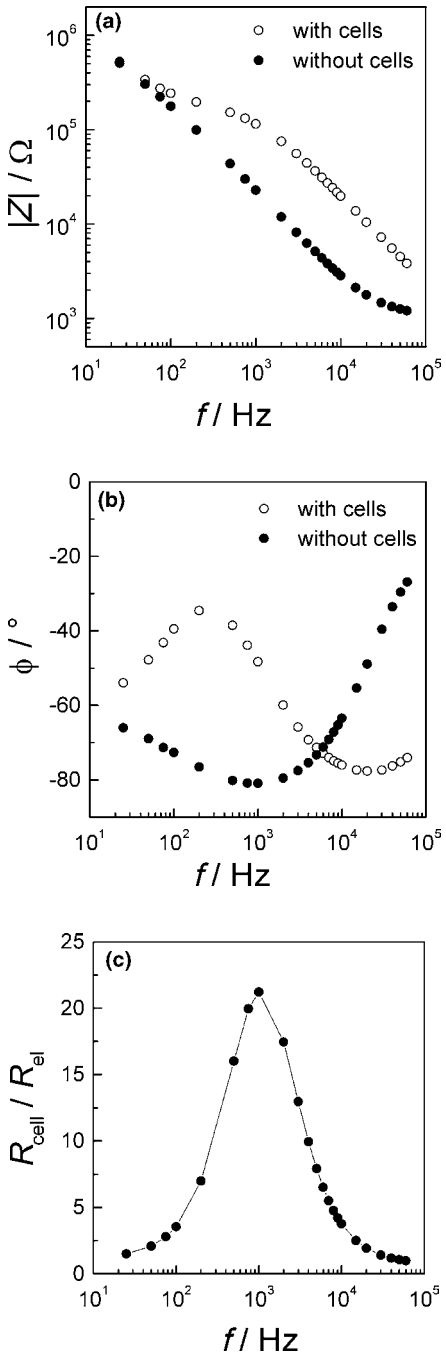


Figure 9. (a, b) Bode-diagrams of an ECIS electrode with confluent MDCK cells and an empty electrode. (c) Plot showing the division of the measured resistance of an ECIS electrode with confluent MDCK cells with the corresponding values of the empty electrode plotted versus log *f*.

plots for ECIS data of an empty electrode and an electrode with a confluent layer of epithelial MDCK cells. It is seen that the presence of cells primarily affects the impedance spectrum for intermediate frequencies between 1 and 100 kHz (Fig. 9a). At the highest frequencies, the two plots approach a horizontal plateau that represents the ohmic solution resistance between the working and the counter electrode. Within the relevant frequency window, the

phase-shift plot for the data of the cell-covered electrode displays two extrema. At frequencies $\sim$ 200 Hz, the phase shift φ is closest to zero, indicating that the contribution of the cells on the measured impedance is mainly resistive. At higher frequencies, the effect of the cell layer becomes more capacitive, and φ starts approaching -90°. The impedance spectrum of the empty electrode displays a single dispersion related to double-layer capacitance at the electrode interface.

The ideal measurement frequencies, where the presence of cells is most pronounced, are determined by dividing the impedance spectrum of a cell-covered electrode with the spectrum of a naked electrode. The same can be done for the resistance or capacitance spectrum, respectively. The most sensitive frequency for resistance measurements is typically found between 1 and 4 kHz (Fig. 9c), where the ratio $R_{cell}(f)$ / $R_{el}(f)$ is at maximum. The capacitive contribution peaks at much larger frequencies, typically on the order of 40 kHz, so that capacitance measurements are often performed at this higher frequency.

During the initial hours following the initial electrode–cell contact, the monitored impedance undergoes a characteristic steep increase. Once the spreading is complete, the impedance continues to fluctuate, reflecting the continuous morphological activities of living cells, for example, movements of cells on the electrode, either by local protrusions or directed movements of the entire cell body, or cell divisions (Fig. 10). The signal characteristics of the impedance during the spreading phase are generally found to be distinct for different cell cultures, both in terms of the duration of the initial gradient and its relative size in comparison to the impedance recorded from a the naked electrode (20). Also, characteristic impedance curves can be obtained by coating the electrode with different proteins (e.g., fibronectin, vitronectin) (21).

Simultaneous optical monitoring of a transparent ECIS electrode has allowed systematic comparison of cell confluence and measured impedance (22). Analysis of data from subconfluent MDCK epithelial cultures revealed a strong linear association between the two variables with cross-correlation coefficients > 0.9; the correlation was found to be equally strong in early and late cultures. This result indicates that $\sim$ 80% of the variance in the measured

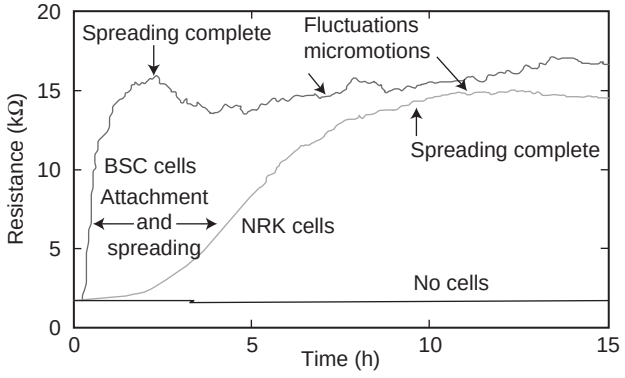


Figure 10. Attachment assay of BSC and NRK fibroblastic cells followed for an interval of 15 h. The graph shows the measured resistance (4 kHz) as function of time; the spreading phase is indicated with arrows.

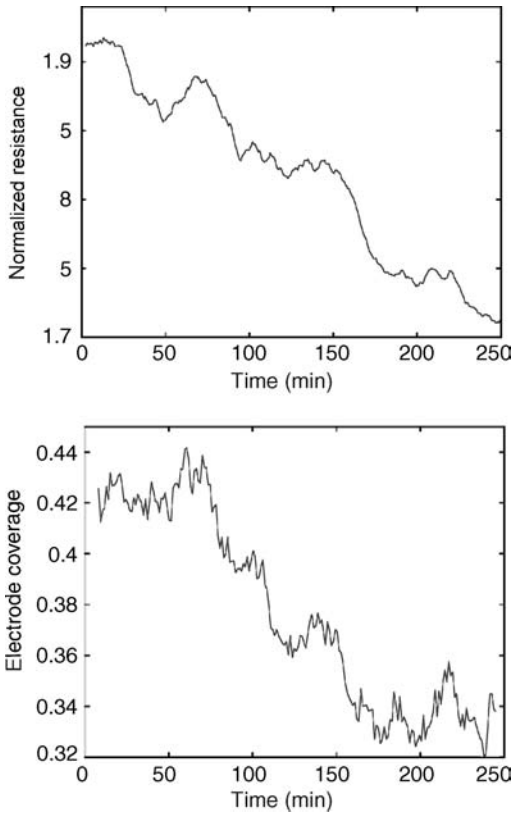


Figure 11. Correlation between resistance and cell coverage. The normalized resistance (4 kHz) versus time (upper panel), and the electrode coverage versus time (lower panel) during the same time interval. The measurement was started 32 h after the cells had been seeded out; the cross-correlation factor was $r = 0.94$.

resistance (4 kHz) can be attributed to changes in cell coverage area (Fig. 11). Moreover, it was possible to link resistance variations to single-cell behavior during cell attachment, including cell-division (temporary impedance plateau) and membrane ruffling (impedance increase). The measured cell confluence was compared to the theoretical model (Eq. 9), neglecting the barrier resistance (i.e., $R_b = 0$), and the calculated values were found to agree well with the data (Fig. 12). Studies like these might pave the way for standardized use of ECIS to quantify attachment and spreading of cell cultures.

IMPEDANCE SPECTROSCOPY AS A TRANDUCER IN CELL-BASED DRUG SCREENING

Another application of impedance spectroscopy with strong physiological and medical relevance is its use as transducer in ECIS-like experiments for cell-based drug screening assays. Here, the impedance readout can be used to monitor the response of cells upon exposure to a certain drug or a drug mixture. In these bioelectric hybrid assays the cells serve as the sensory elements and they determine the specificity of the screening assay while the electrodes are used as transducer to make the cell behavior observable. In the following example, endothelial cells isolated from bovine aorta (BAEC = bovine aortic endothelial cells) were grown to confluence on gold-film electrodes since they

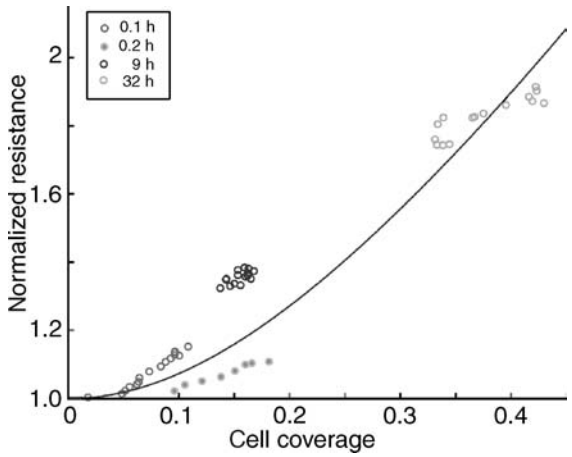


Figure 12. Theoretical prediction of cell coverage. Theoretical curve of normalized resistance plotted as function of cell coverage on the electrode. Normalized resistance and corresponding cell density are shown for four different registrations with circles. Time points indicate when the recordings were initiated with respect to the start of the culture; each circle corresponds to average values for 15 min time intervals.

express cell-surface receptors (β-adrenoceptors) that are specific for adrenalin and derivatives (23,24). These β-adrenoceptors belong to the huge family of G-protein coupled receptors (GPCR) that are of great pharmaceutical relevance and impact. By measuring the electrical impedance of the cell-covered electrode, the stimulation of the cells by the synthetic adrenaline analogue isoprenaline (ISO) can be followed noninvasively in real time without any need to apply costly reagents or to sacrifice the culture (25). Experimentally, the most sensitive frequency for time-resolved impedance measurements is first determined from a complete impedance spectrum along an extended frequency range as depicted in Fig. 13. The figure compares the impedance spectrum of a circular gold-film electrode ($d = 2$ mm) with and without a confluent monolayer of BAECs. The contribution of the cell layer to the total impedance of the system is most pronounced at frequencies close to 10 kHz, which is, thus, the most sensitive sampling frequency for this particular system. It is noteworthy that the most sensitive frequency may

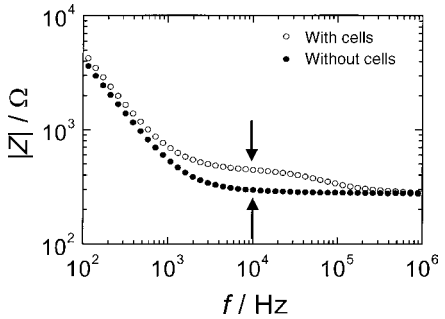


Figure 13. Frequency-dependent impedance magnitude for a planar gold-film electrode ($d = 2$ mm) with and without a confluent monolayer of BAEC directly growing on the electrode surface. The difference in impedance magnitude is maximum at a frequency of 10 kHz.

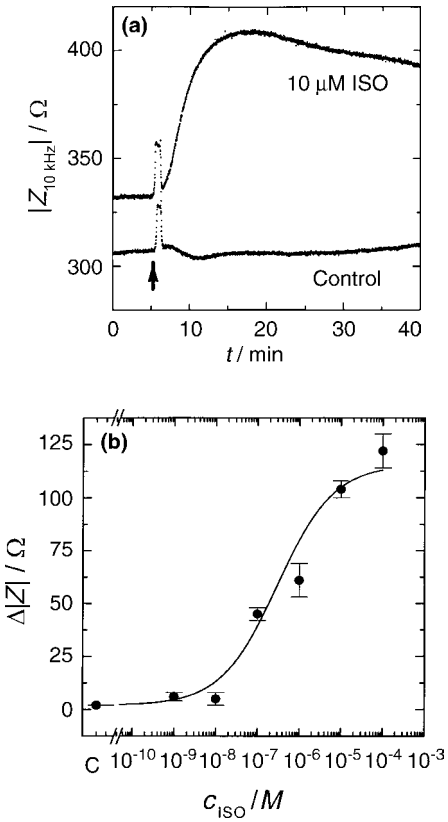


Figure 14. (a) Time course of the impedance magnitude at a sampling frequency of 10 kHz when a confluent monolayer of BAECs is exposed to 10 μM ISO or a corresponding vehicle control. (b) Dose-response relationship between the increase of impedance magnitude ΔZ and the concentration of isoproterenol applied. Quantitative analysis reveals an EC_{50} of 0.3 μM similar to the binding constant of ISO to β-adrenoceptors.

change with the electrode size and the individual electrical properties of the cells under study.

Figure 14a traces the time course of the impedance magnitude at a frequency of 10 kHz when confluent BAEC monolayers were either challenged with 10 μM ISO or a vehicle control solution at the time indicated by the arrow. The exchange of fluids produces a transient rise of the impedance by 10–20 Ω that is not caused by any cellular response, but mirrors the reduced fluid height within the measuring chamber. As expected, no response of the cells is seen in the control experiment. The cell population exposed to 10 μM of ISO shows a significant increase in electrical impedance that goes through a maximum 10 min after ISO application, and then slowly declines. The reason for the increase in impedance as observed after ISO stimulation is similar to what has been described for three-dimensional (3D) tissues above. The adrenaline derivative induces a relaxation of the cytoskeleton that in turn makes the cells flat out a bit more. As a consequence the extracellular space between adjacent cells narrows and increases the impedance of the cell layer. Note that the time resolution in these measurements is ~ 1 s so that even much faster cell responses than the one studied here can be monitored in real time. Moreover, no labeled probe had to be applied and

the sensing voltages used for the measurement ($U_0 = 10$ mV) are clearly noninvasive.

From varying the ISO concentration, a dose-response relationship (Fig. 14b) can be established which is similar to those derived from binding studies using radiolabeled ligands. Fitting a dose-response transfer function to the recorded data returns the concentration of half-maximum efficiency EC_{50} as (0.3 ± 0.1) μM, which is in close agreement to the binding constant of ISO to β-adrenoceptors on the BAEC surface as determined from binding assays with radiolabeled analogs (23).

These kind of electrochemical impedance measurements are also used to screen for potent inhibitors of cell-surface receptors. Staying with the example discussed in the preceding paragraph, the blocking effect of Alprenolol (ALP), a competitive inhibitor of β-adrenoceptors (β-blocker), is demonstrated. Preincubation of BAEC with ALP blocks the stimulating activity of ISO, as shown in Fig. 15. The figure compares the time course of the impedance magnitude at a frequency of 10 kHz when BAEC monolayers were stimulated with 1 μM ISO either in absence of the β-blocker (a) or after preincubation (b).

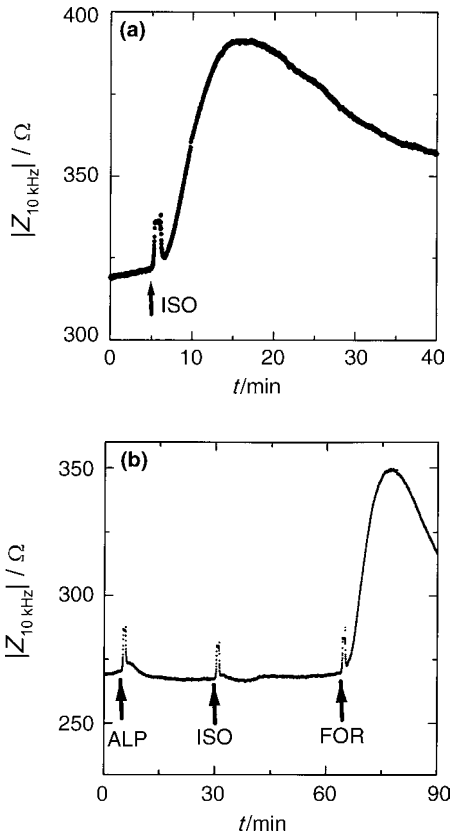


Figure 15. (a) Time course of the impedance magnitude at a sampling frequency of 10 kHz, when confluent BAEC are exposed to 1 μM ISO. (b) Time course of the impedance magnitude of a confluent monolayer of BAECs upon sequential exposure to 10 μM of the β-blocker ALP and 1 μM ISO 20 min later. The β-adrenergic impedance increase is omitted by the β-blocker. Intactness of the signal transduction cascade is verified by addition of forskolin (FOR), a receptor independent activator of this signal transduction pathway.

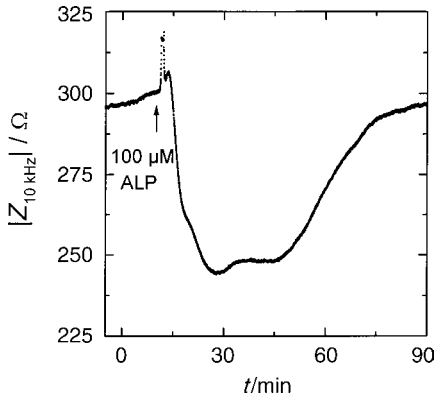


Figure 16. Time course of the impedance magnitude at a sampling frequency of 10 kHz when a confluent BAEC monolayer is exposed to an over dose of the β-blocker alprenolol (100 μM ALP). Addition of ALP is indicated by the arrow.

When the cell layers were incubated with 10 μM ALP prior to the addition of 1 μM ISO, the cells do not show any ISO response indicating that a 10-fold increase of ALP was sufficient to block activation of the receptors. To prove that the cells were not compromised in any way during these experiments the same signal transduction cascade was triggered via a receptor-independent way at the end of each experiment. This can be easily done by application of FOR, a membrane permeable drug that intracellularly activates the same enzyme that is triggered by ISO binding to the receptor. Forskolin stimulation of those cells that had been blocked with ALP earlier in the experiment induces a strong increase of electrical impedance indicating that the intracellular transduction pathways are functional (Fig. 15b).

Besides screening for the activity of drugs in cell-based assays, these kind of measurements are also used to check for unspecific side effects of the compounds of interest on cell physiology. Dosage of ALP and many of its fellow β-blockers has to be adjusted with great care since these lipophilic compounds are known to integrate nonspecifically into the plasma membrane. As shown in Fig. 15b, application of 10 μM ALP does not show any measurable side effects. Using ALP in concentrations of 100 μM induces a transient but very pronounced reduction of the electrical impedance (Fig. 16). This decrease in impedance may be the result of the interaction of ALP with the plasma membrane or an induced contraction of the cell bodies.

The preceding examples showed that impedance measurements of cell-covered gold electrodes in which the cells serve as sensory elements can be used in screening assays for pharmaceutical compounds, but also for cytotoxicity screening. The interested reader is referred to Ref. 26 and 27.

SUMMARY AND OUTLOOK

Impedance spectroscopy is a general technique with important applications in biomedical research and medical diagnostic practice. Many new applications are currently under investigation and development. The potential of the

technique is obviously great, since it is noninvasive, easily applied, and allows on-line monitoring, while requiring low cost instrumentation. However, there are also difficulties and obstacles related to the use of IS. Foremost, there is no separate access to the individual processes and components of the biological system, only the total impedance is measured, and this signal must be interpreted by some chosen model. There are many fundamental issues yet to be solved, both connected with understanding the origin of bioimpedance, methodological problems with finding standardized ways of comparing different samples, as well as technical issues connected with the equipment used to probe bioimpedance.

Prospective future *in vivo* applications include quantification of ischemia damage during cardiac surgery (28) and organ transplantation (29), as well as graft rejection monitoring (30). Impedance spectroscopy are also used for tissue characterization, and recently a device for breast cancer screening became commercially available. Multifrequency electrical impedance tomography (EIT) performing spatially resolved IS is a potential candidate for diagnostic imaging devices (31), but due to poor resolution power compared to conventional methods like MR, only few clinical applications are described.

The use of impedimetric biosensor techniques for *in vitro* monitoring of cell and tissue culture is promising. With these methods, high sensitivity measurements of cell reactions in response to various stimuli have been realized, and monitoring of physiological–pathological events is possible without use of marker substances. The potential applications cover pharmaceutical screening, monitoring of toxic agents, and functional monitoring of food additives. Microelectrode-based IS is interesting also for scientific reasons since it allows studying the interface between cells and technical transducers and supports the development of implants and new sensor devices (32).

Finally, affinity-based impedimetric biosensors represent an interesting and active research field (33) with many potential applications, for example, immunosensors monitoring impedance changes in response to antibody–antigen reactions taking place on electrode surfaces.

BIBLIOGRAPHY

1. Macdonald JR. Impedance Spectroscopy. New York: John Wiley & Sons; 1987.
2. Fricke H, Morse S. The electrical capacity of tumors of the breast. J Cancer Res 1926;10:340–376.
3. Schwan H. Mechanisms Responsible for Electrical Properties of Tissues and Cell Suspensions. Med Prog Technol 1993;19: 163–165.
4. Grimnes S, Martinsen ØG. Bioimpedance and Bioelectricity basics. Cornwall: Academic Press; 2000.
5. McAdams E, Lackermeier A, McLaughlin J, Macken D, Jossinet J. The linear and non-linear electrical properties of the electrode-electrolyte-interface. Biosens Bioelectron 1995;10: 67–74.
6. Kottra G, Fromter E. Rapid determination of intraepithelial resistance barriers by alternating current spectroscopy. II. Test of model circuits and quantification of results. Pflugers Arch 1984;402:421–432.
7. Cole K, Cole R. Dispersion and adsorption in dielectrics. I.. alternating current characteristics. J Chem Phys 1941;9: 341–351.
8. Cole Ks. Membrane, Ions and Impulses. Berkeley (CA): University of California Press; 1972.
9. Kell D. Biosensor. Fundamentals and Applications. Turner A, Karube I, Wilson G, editors. Oxford Science Publications; 1987. pp 427–468.
10. Schwan H. Electrical properties of tissue and cell suspensions. Advances in biological and medical physics. Lawrence J, Tobias C, editors. New York: Academic Press; 1957. pp 147–209.
11. Gersing E. Impedance spectroscopy on living tissues for determination of the state of organs. Bioelectrochem Bioenenerg 1998;45:149.
12. Giaever I, Keese CR. Monitoring fibroblast behavior in tissue culture with an applied electric field. Proc Natl Acad Sci 1984;81:3761–3764.
13. Giaever I, Keese CR. A morphological biosensor for mammalian cells. Nature (London) 1993;366:591–592.
14. Applied Biophysics, Inc. 2002.
15. Connolly P, et al. Extracellular electrodes for monitoring cell cultures. IOP Publishing; 1989.
16. Wegener J, Sieber M, Galla HJ. Impedance analysis of epithelial and endothelial cell monolayers cultured on gold surfaces. J Biochem Biophys Methods 1996;76:327–330.
17. Hagedorn R, et al. Characterization of cell movement by impedance measurements on fibroblasts grown on perforated Si-membranes. Biochem Biophys Acta—Molecular Cell Res 1955;1269:221–232.
18. Giaever I, Keese C. Micromotion of mammalian cells measured electrically. Proc Natl Acad Sci USA 1991;88:7896–7900.
19. Lo CM, Keese CR, Giaever I. Impedance analysis of MDCK cells measured by electric cell-substrate impedance sensing. Biophys J 1995;69:2800–2807.
20. Giaever I, Keese CR. Use of electric fields to monitor the dynamical aspect of cell behavior in tissue culture. IEEE Trans Biomed Eng 1986;33:242–247.
21. Mitra P, Keese CR, Giaever I. Electrical measurements can be used to monitor the attachment and spreading of cells in tissue culture. BioTechniques 1991;11:504–510.
22. De Blasio BF, Laane M, Walmann T, Giaever I. Combining optical and electrical impedance techniques for quantitative measurements of confluence in MDCK-I cell cultures. BioTechniques 2004;36:650–662.
23. Zink S, Roesen P, Sackmann B, Lemoine H. Regulation of endothelial permeability by beta-adrenoceptor agonists: Contribution of beta 1- and beta 2-adrenoceptors. Biochim Biophys Acta 1993;1178:286–298.
24. Zink S, Roesen P, Lemoine H. Micro- and macrovascular endothelial cells in beta-adrenergic regulation of transendothelial permeability. Am J Physiol 1995;269:C1209–C1218.
25. Wegener J, Zink S, Roesen P, Galla H. Use of electrochemical impedance measurements to monitor beta- adrenergic stimulation of bovine aortic endothelial cells. Pflugers Arch 1999;437:925–934.
26. Arndt S, et al. Bioelectrical impedance assay to monitor changes in cell shape during apoptosis. Biosens Bioelectron 2004;19:583–594.
27. Keese C, Karra N, Dillon B, Goldberg A, Giaever I. Cell-substratum interactions as a predictor of cytotoxity. *In Vitro* Mol Toxicol 1998;11:183–191.
28. Benvenuto, et al. Impedance microprobes for myocardial ischemia monitoring. 1st Annual International IEEE-EMBS. Lyon, France; 2000. p 234–238.

29. Haemmerich D, et al. Changes in electrical resistivity of swine liver after occlusion and postmortem. Med Biol Eng Comput 2002;40:29–33.
30. Ollmar S. Noninvasive monitoring of transplanted kidneys by impedance spectroscopy—a pilot study. Med Biol Eng Comput 1997;35:1–336.
31. Brown B. Electrical impedance tomography (EIT): A review. J Med Eng Technol 2003;27:97–108.
32. Borkholder D. Cell based biosensors using microelectrodes. Ph.D. Dissertation. 1998. Stanford University.
33. Katz E, Wilner I. Probing biomolecular interactions at conducting and semiconducting surfaces by impedance spectroscopy: routes to impedimetric immunosensors. Electroanalysis 2003;15:913–947.

See also CAPACITIVE MICROSENSORS FOR BIOMEDICAL APPLICATIONS.

IMPLANT, COCHLEAR. See COCHLEAR PROSTHESES.

INCUBATORS, INFANTS

ROBERT WHITE
Memorial Hospital South Bend
South Bend, Indiana

INTRODUCTION

Providing newborn infants with appropriate thermal protection is known to improve their growth rates (1–3), resistance to disease (4,5) and survival (6–11). Keeping premature, sick, or otherwise high risk babies warm is particularly critical and, when their care precludes covering them with protective swaddling clothing, especially difficult. Incubators are devices used during the care of such high-risk infants and are designed with the intent of producing environmental conditions that are consistently suitable to each unique infant's particular needs. There are many different kinds of incubators that differ in detail in the way they are constructed, heated, and controlled (12–16). All provide a mattress for the infant to lie upon, surrounded by a warmed microclimate that is controlled by a logical system governing the amount of heat needed to keep the environmental temperature within a limited range. In some incubators, this microclimate is produced within a rigid walled chamber; such devices are called closed incubators. When they are heated by using a fan to force air over a metallic heating coil prior to its entry into the infant's chamber, these closed incubators are also called forced convection incubators. There also are open incubators; those have no walls and, therefore, no chamber surrounding the mattress. There is nothing delimiting the convective environment in an open device, so they need to be heated by using a radiant warmer directed to the mattress area. These devices, therefore, are commonly called open radiant incubators, radiant warmer beds, or radiant heaters.

Each of these types of incubators provides certain unique advantages. The convectively heated incubator provides a caretaker with a far easier method for controlling the humidification of the infant's microclimate, when compared to the open radiant warmer bed. Therefore, a baby under an open radiant heater loses more body fluid than does an infant within a closed convectively heated chamber (17). But conversely, a baby in an open incubator, while more complicated to care for in terms of medical fluids administration, is physically more accessible in terms of other kinds of care that sometimes are equally important to the well being of sick babies. Current "top of the line" incubators incorporate the advantages of both types, utilizing a radiant warmer bed with a removable enclosure that allows full physical access to the infant when the incubator is operated in the radiant heater mode, and better control of humidification and noise when the enclosure is placed around the baby and operated in the convectively heated mode.

An incubator, in many respects, is just a very little house sized to fit the space and functional requirements of an infant occupant. As choices must be made when conditioning the environment in any house, different options must be considered when designing the climate control system in an incubator. In the following review, some of these considerations will be explained from the perspective of how environmental manipulators affect newborn infants who are not just little human adults, but also developing individuals with special physical, physiologic, metabolic, and neurological capabilities and limitations that make them unique. In great measure incubator manufacturers have been successful in translating present day knowledge of babies and their special needs into technical solutions that make today's incubators remarkably functional. But any infant caretaker or incubator manufacturer can attest to the limitations of today's devices which, as they are approximate to our present scientific knowledge and the existing level of technology, are flawed by our considerable remnant ignorance and the failure of existing technology to meet certain imperative needs already known.

HISTORY

It is ancient knowledge that infants who are allowed to get cold have a greater chance of dying than do infants kept warm. Prior to the nineteenth century, keeping small babies warm meant swaddling with multiple layers of cloth, providing body contact with the mother, or placement of the infant near a warm, roaring fireplace. Such classic thermal care served lusty, healthy babies well, but was inadequate to provide for the special needs of premature or otherwise enfeebled newborns. These special needs were not met because, until the last century, there was almost no recognizable major medical or social commitment toward enhancing the survival of babies born prematurely. The infant mortality rate was high and accepted. However, in response to various politicosocial events that occurred in the late 1700s and early 1800s, the value of improving premature infant survival increased, stimulating the development of special incubators in which to care for these fragile, newly valued babies.

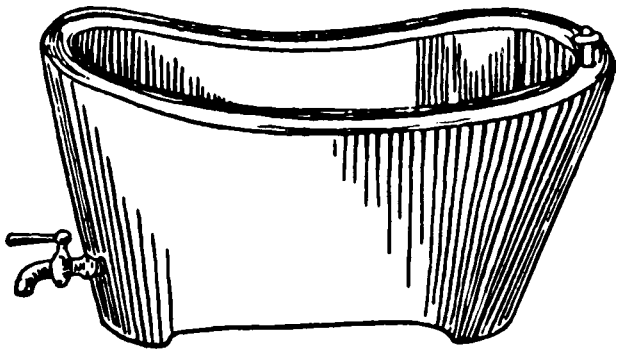


Figure 1. von Ruehl warming tub (1835). Reprinted with permission from T. E. Cone, Jr., *History of the Care and Feeding of the Premature Infant,* Boston, MA: Little, Brown, 1985.

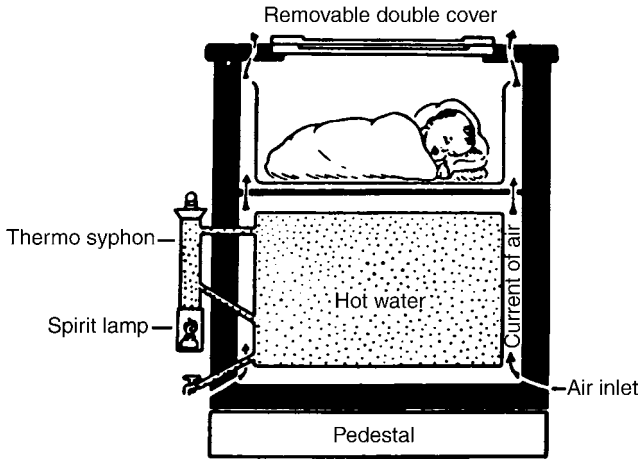


Figure 2. Tarnier incubator (1880). Reprinted with permission from T. E. Cone, Jr., *History of the Care and Feeding of the Premature Infant,* Boston, MA: Little, Brown, 1985.

The first serious attempt to improve on the thermal protection provided newborns was reflected in a warming tub developed in 1835 by Johann Georg von Ruehl (1769–1846) in Russia (Fig. 1). The von Ruehl tub was simply a double-walled sheet-iron open cradle that was kept warm by filling the space between the walls with warm water. Variations on von Ruehl's design were subsequently developed throughout Europe, and this type of primitive open incubator remained a standard device for care until 1878.

Although the von Ruehl device must be recognized as a developmental milestone, to be truly accurate, the developmental history of modern infant incubators must be traced back centuries to Egypt where the artificial incubation of eggs was refined and remained a closely guarded secret and uniquely Egyptian profession. Not until 1799 were these secrets introduced into Europe by members of Napoleon's expedition. Professor Stephane Tarnier (1828–1897) of the Paris Maternity Hospital in 1878 saw a chicken incubator at the Paris Zoo. The incubator, based on old Egyptian designs, had been constructed by Odile Martin. Dr. Tarnier perceived how such a device, with modifications, could be used to keep premature infants warm. Odile Martin subsequently built the first approximation of the modern enclosed infant incubator initially used at the Paris Maternity Hospital in 1880 (Fig. 2)

The Tarnier incubator was simple in its design. The infant lay in the upper chamber of a two-chambered double-walled box insulated to slow the loss of heat. The infant chamber was topped with a removable cover through which the infant could be observed while remaining protected from cooling room drafts. The heating of the upper chamber was achieved by warming a large supply of water contained in the lower chamber of the incubator. The water was heated by an alcohol or gas lamp thermosyphon that was external to the incubator chambers and connected by piping that allowed convection driven water flow between the heater and the water reservoir. Cool room air freely flowed into the lower warming chamber where the air picked up heat from the surface of the warm water reservoir and then, by natural convection, rose to enter and warm the upper chamber containing the infant.

The Tarnier incubator was neither elegant nor efficient, and even when within the device, infants needed the extra protection of swaddling blankets. It did, however, reflect the technology of its day and, in the climate of a new commitment toward the study and welfare of feeble infants, stimulated others to refine the basic Tarnier design to correct deficiencies discovered through acquired practical experience with the incubator in clinical settings. The historical progression in this development has been illuminated by Dr. Thomas Cone, and the reader is referred to Dr. Cone's excellent treatise for a more detailed description of the many design variations introduced and tested along this early path leading to today's equipment (18).

The modernization of incubators has not only led to marked improvement in the thermal protection provided infants, but it also has been pivotal in increasing our knowledge of diseases unique to newborn babies. In turn, each increment of knowledge has required manufacturers to modify incubators in order to provide solutions to problems the new scientific discovery imposed. For example, when electric fans became available and forced air convection systems were developed, incubator heating became so improved that infants for the first time could be undressed during care. This along with the use of new transparent plastics in the construction of incubator walls allowed clinicians to make observations that led to detailed descriptions of illnesses about which little was known when infants were hidden within the predominately opaque wooden and metal chambers of the past. But while the employment of clear plastic in incubator construction enhanced the ability to observe infants, its poor insulating qualities made the task of maintaining the incubator chamber in a warm and stable state more difficult. And as improved visibility led to new lifesaving therapies, such as the administration of intravenous fluids, the use of respirators in care, and the development of new diagnostic procedures and surgical interventions, the transparent plastic walls that provided caretakers with visual access to sick babies during these therapeutic processes also served as impediments. Incubators had to be modified so that catheters, tubes, and wires could be connected to an infant's body. Increasing numbers of access holes were

drilled through the walls of the incubator to provide portals of entry for these therapeutic tools, but the new fenestrations also produced new exits for life-sustaining heat. More modifications were needed and, as each problem was solved, a new dilemma emerged.

Even today, infant caretakers and incubator manufacturers continue to struggle with these and other problems contributing to the strengths and weaknesses in incubator devices. In this article, the physiologic, clinical, and technical factors leading to the design of existing incubators will be outlined further and some of the limitations of incubators explained in greater detail. Throughout, we hope that it remains clear that incubator development is an ongoing process requiring frequent and critical review of existing methods to assure that the thermal protection being provided is still appropriate during the delivery of other and especially newer forms of care also deemed necessary to a baby's well being. The ideal incubator of tomorrow is one that neither impedes care nor can itself, when providing thermal protection, be impeded by other forms of care. This has always been and remains the major challenge to health care providers and engineers committed to incubator development.

FACTORS CONSIDERED IN INCUBATOR DESIGN

Physiological Heat Balance

Even though some controversy exists concerning the exact definition of the body temperature limits within which a newborn's body functions optimally, in general, any temperature between 35.5 and 37.5 °C is probably within that normal range and can be referenced to published data available on the subject. Body temperature is determined by the balance between the heat produced within and lost from the body tissues. In order to design or even understand the design and limitations of modern incubators, a knowledge of these basic factors is required to provide the context for how infants differ from adults in their thermoregulatory capabilities.

Heat Production

All animals, including human babies, produce body heat as a by-product of the biochemical processes that sustain life. The basic amount of heat produced by a newborn infant is $\sim$1.5–2 W·kg^{-1}. During the first weeks of life, this minimal rate of heat production is both weight and age related, with smaller and younger babies capable of producing less heat than larger and older infants (19–23).

In addition to this basic capacity, most healthy babies have the capability to generate additional heat to a maximum production rate of $\sim$4.5–5 W·kg^{-1} (21–23). This additional heat-producing capacity is often called upon for protective purposes, as, for example, when the infant is challenged to fight off infection or when stressed by situations that cause an exorbitant amount of heat to be lost from the body. The capability to increase the amount of heat produced to replace body heat losses is called homeothermy. In contrast to homeotherms, some creatures, such as lizards, reptiles, and fish, are poikilotherms that do not produce more heat when cooled, but actually decrease their metabolic rates when exposed to cold.

When considering thermoregulatory problems associated with newborn care, both homeothermy and poikilothermy must be understood, because under some circumstances, it is possible for a homeothermic animal to behave like a poikilotherm. Sometimes during the medical care of humans this possibility is used to a patient's benefit; for example, during some heart operations patients are given drugs that prevent their nervous systems from responding to the cold. In this circumstance, it is desirable to slow the body's metabolic rate, which can be achieved by cooling the drug treated patient who, by intent, has been changed to a temporary poikilotherm.

At times, a homeotherm may revert temporarily to a poikilothermic state. This is particularly common in immature or very sick newborns, and especially those with neurologic damage (24) or with breathing problems that lead to inadequate blood and tissue oxygen levels (25,26). Poikilothermy in a newborn can also occur because of drugs administered to a mother in labor with subsequent placental transport of those drugs to the infant (27).

In spite of their occasional reversion to poikilothermy, it nonetheless is commonly advised that newborns should be thought of as homeotherms and protected from environments that would unduly stimulate homeothermic tendencies. This is because homeotherms increase heat production by increasing metabolic work which can cause excess utilization of finite fat, sugar, and protein stores needed to sustain other vital body functions and to meet growth and developmental milestones. Moreover, extra metabolic work produces more than just extra heat; acidic and other metabolic by-products produced at the same time can cause severe imbalances in the body's critical acid–base balance (4). As a consequence of the self-protective reaction to cold stress a newborn may, therefore, be faced with an equally life-threatening biochemical stress (Fig. 3).

It has been suggested that one reason cold-exposed infants have higher mortality rates is that they become metabolically exhausted and incapable, in part because of the consequent acidosis, to fight off the stresses placed on their bodies by other threatening events. But problems can arise when attempts are made to provide infants with protection from cold stress, since it is unclear how to be sure that a given environment minimizes the infant's metabolic efforts to maintain homeothermy. Theoretically, this could be achieved by measuring the infant's metabolic rate continuously and making adjustments in the environmental supports whenever the infant's rate of metabolism changed, but measurement of heat production by a newborn is very difficult in the clinical setting.

In any case, few infant caretakers would consider the infant's metabolic rate as the only measure of success when providing a baby with thermal protection. The infant's actual body temperature is also considered important (21) and it is well known that metabolic rate is only one factor contributing to the body temperature measured. Body temperature also is influenced by the rate at which heat is lost from the body and, if one wishes to produce a device that contributes to the maintenance of an infant's body temperature, it is necessary, by virtue of its balancing

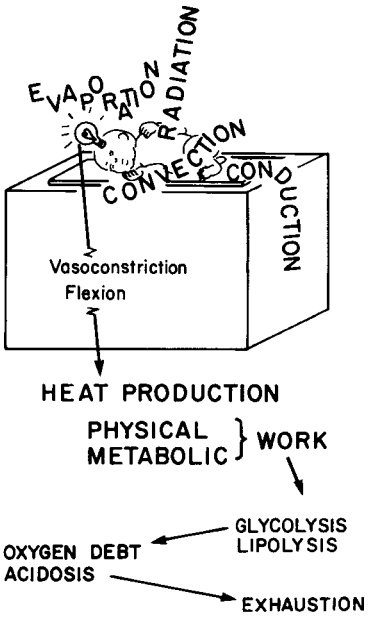


Figure 3. Schematic of homeothermy in newborns. On sensing loss of body heat, the infant minimizes heat loss from the skin by vasoconstricting blood vessels, changing body positions, and increasing metabolic rate. The increase in metabolism can produce acidosis and depletion of energy substrate stores. Reproduced by permission from P. H. Perlstein, "Routine and special care–Physical environment." In A. A. Fanaroff and R. J. Martin (Eds.), *Behrman's Neonatal-Perinatal Medicine*, 3rd ed., St. Louis, MO: C. V. Mosby Co., 1983.

effect on the final body temperature achieved, to understand how heat loss occurs in newborns.

Heat Loss

In the final analysis, incubators actually protect infants only by modifying the factors that contribute to heat loss via the well-known conductive, convective, radiant, and evaporative mechanisms.

The flow of heat occurs only when there is a difference in the temperatures of adjacent structures. Heat can only be lost by a warmer object to a cooler one. Conductive heat losses occur when an infant comes in physical contact with a cooler solid surface. A baby loses heat by conduction to a cooler mattress, blanket, diaper, or other clothing. Convective losses are similar to, but independent of, conductive losses and occur when a baby is exposed to air currents that are cooler than the infant. Convective losses are influenced not only by temperature differences, but also by the wind chill factor determined by speed at which the air is flowing. In a practical sense, this means that if an incubator has a fan as part of a forced convection heating system, air movement produced by the fan can cause a cooling effect in excess that which would occur if air at the same temperature was still.

Heat loss in infants also occurs by radiation in the infrared (IR) spectrum to cooler solid objects surrounding, but not in contact with their skin. The incubator and room walls, windows, and furniture all contribute to heat loss via radiant mechanisms. Finally, evaporative heat losses occur as infants transpire water from their skin into the surrounding environment. They can also lose heat by evaporation as residual amniotic fluid or bath water dries from their skin and hair, and they lose heat from their lungs as they exhale warm humid air.

Gross estimates of the magnitude of heat loss can be calculated by physical heat-transfer equations for each mechanism. The reader is referred to thermal transfer books for the details of these mechanisms, but examination of these equations here in a simplified form is a useful way to discover some of the special features that influence heat transfer as applied to newborn care.

All nonevaporative heat losses are quantitatively proportional to the magnitude of the temperature difference between the warmer object (T_o) losing heat and the cooler environmental feature (T_e) that will receive the heat in transfer:

$$\text{Heat loss} = \propto (T_0 - T_e)$$

This equation becomes equality by adding to it an object specific constant called a thermal transfer coefficient (k):

$$\text{Heat loss} = k(T_0 - T_e)$$

Different materials at the same temperature lose heat at different rates when exposed at the same thermal environment; for example, a block of wood has a lower thermal transfer coefficient than a block of steel. Newborn infants have higher thermal transfer coefficients than do adults and therefore lose body heat more rapidly than adults when exposed to any environment that is cooler than body temperature, so in a room temperature that feels comfortable to an adult, a newborn can get cold.

It must be emphasized that the heat loss equation as written above is grossly simplified and omits numerous other factors important to actual heat exchange. A more accurate equation would have to include a factor accounting for the infant's exposed surface area, which is a quantity that changes with changes in an infant's position and is modified if the infant is swaddled in blankets, wears a diaper, booties, hat, or, if in the course of surgical care, has a bandage applied. The equation as simplified particularly fails to reflect the true degree of complexity describing the thermal relationship between an infant's skin and the radiant surfaces of a room or incubator. The relationship is modified, for example, by complex factors describing the exact path and distance traveled by infrared (IR) waves in their transfer from object to object.

Radiant heat loss is also modified by the emissivities of the objects exchanging energy. Like black carbon, an infant's skin, no matter what its actual color, is presumed to have an emissivity of 1, which means it absorbs and emits IR rays perfectly and completely. The emissivities of the materials used in incubator manufacture or in nursery wall coverings also modify the amount of radiant exchange occurring with the infant's radiant surface. The emissivities of these objects become particularly important in an incubator chamber in which the surface area of the interior walls surrounds the exposed surface area of the infant's radiating skin.

The following equation, although still simplified, provides a better approximation of expected radiant losses (H_r)

from an infant in an incubator (28). In this equation, A_b is the exposed surface area of the infant, A_r is the area of the walls surrounding the infant, E_s is the emissivity of the infant's skin, and E_r is the emissivity of the walls. The symbol o is the Stefan–Boltzmann constant, 5.67×10^{-8} $W^{-1}{\cdot}m^{-2}{\cdot}K^{-4}$. When using this equation, temperatures are expressed in kelvin and the heat loss in watts.

$$H_r = A_b\left[\frac{1}{Es} + \frac{A_b}{A_r}\left(\frac{1}{E_r} - 1\right)\right]^{-1} \sigma(T_s^4 - T_r^4)$$

Radiant exchange relationships are so variable because of differences between infants and different incubator environments that, even using this more complex equation, only poor and static quantitative approximation of actual radiant flux can be made in clinical settings. This proves to be a practical problem when considering incubator designs, since it has been documented that in many situations radiant losses can account for >60% of the total heat loss from an infant (29).

Evaporative heat losses are not specifically related to temperature difference and occur instead because of differences that exist between the partial pressures of water in the boundary layer of air next to the infant's skin and that in the environment beyond the boundary layer limits.

$$\text{Evap loss} = K(\text{partial pressure skin} - \text{partial pressure air})(\text{ares})$$

Partial pressures are not the same as relative humidities, so even in an environment at 100% relative humidity, an infant can lose water and heat if the skin surface is warmer than the environment. For each milliliter of fluid lost from the body, ~580 g·cal (2.4 kJ) of heat are lost in the vaporization process. This route of heat loss accounts for ~25% of the total heat loss when an infant is dry. When lying unclothed on an open bed heated only by a radiant heater, up to 300 $mL^{-1}{\cdot}kg^{-1}{\cdot}day^{-1}$ of fluid can be lost by evaporation from the skin of very immature infants in the first days of life. In an enclosed incubator that is poorly humidified, up to 150 $mL^{-1}{\cdot}kg^{-1}{\cdot}day^{-1}$ of water can be lost by this mechanism in very immature infants. Following birth when the infant is wet with amniotic fluid, or following a bath, this can become the predominant route of heat loss (30–34,34).

Environmental Temperature

It should be obvious from the previous discussion that since conduction, convection, radiation, and evaporation are each relatively independent mechanisms, no single measurable quantity can be used to calculate their combined contribution to heat loss. Air temperature, for example, can be used to estimate only the convective component of heat loss from a baby, while measurements of incubator inside wall temperatures can only be helpful in determining approximate losses due to radiation. This means that if the incubator walls are cold, a baby in an incubator can lose heat even if the air temperature is warmer than the infant. The only feature necessary for this to be true is for radiant losses to be higher than convective heat gains.

Environmental temperature, although frequently used loosely to describe any ambient thermal value, must be understood to be a specific reference to the combination of temperatures actually experienced by an infant in thermal exchange relationships via multiple mechanisms. Unfortunately, few guidelines exist at present to help caretakers know the true environmental temperature for an infant within an incubator in a clinical setting. When certain conditions are met, however, some of the guidelines seem to be useful. Dr. Hey, for example, determined that in a well-humidified enclosed convectively heated incubator with single-layer Plexiglas walls, the environmental temperature perceived by a contained infant is ~1 °C lower than the measured midincubator chamber air temperature for every 7 °C difference that exists between the incubator air temperature and the air temperature of the room within which the incubator stands (35).

Heat Transfer within the Body

Since the skin of the newborn is the major heat-losing surface in exchange with the environment, mechanisms by which heat transfers from interior tissues to the skin play an important part in the heat loss process. The rate at which internally produced heat is transferred from the body core temperature T_B through body tissues to the outer body skin surface at temperatures T_s is computed using the following equation:

$$\text{Heat transfer} = C(T_B - T_s)$$

Where C is an individual's specific thermal conductance coefficient, which is affected by the absolute thickness and character of the skin, subcutaneous fat, and other subcutaneous tissue, and by the blood flow rate from the body core to its surface. Obviously, babies are smaller and have thinner body coverings than do adults, and, therefore, as they lose heat more rapidly from their surfaces than do adults, they also transfer heat more rapidly to their surfaces from their internal heat-producing organs. In addition, an infant's blood vessels are relatively close to the body surface. Since the vascularity of a particular body surface determines the rate at which blood will shunt core heat around intervening insulating tissues to the skin surface, such shunting contributes heavily to the high thermal conductance of a baby.

Heat can also be lost rapidly from an infant's body core via the respiratory system. This route of loss is of relatively minor significance in a healthy and spontaneous breathing infant, but in a baby who is ill and especially one being mechanically assisted by a respirator, this can become the major route by which body heat is lost or gained. Body heat transferred by this route is dependent on the infant's temperature, breathing rate, the flow rate and temperature of gases reaching the lungs, and the water content of the gas delivered to the airway. If temperatures and humidification are not properly monitored and controlled, the heat losses from the respiratory passages can be so great that they exceed the capacity of the incubator heater.

The Concept of a Neutral Thermal Environment

Theoretically and as demonstrated by several authors (21,36,37), it is possible for a competent homeothermic baby to have a body temperature that is below the normal

range at a time when the measured metabolic rate of the infant is at a minimal unstimulated level. For example, Brück (36) documented that during a period of cooling associated with a falling environmental temperature, an infant's metabolic rate and heat production increased, but, as soon as the cooling environment was caused to reheat, the infant's metabolic rate decreased to minimal heat-producing levels, and this decrease occurred even before the infant's cooled body temperature returned to normal. This was confirmed by Adamsons et al. (21) and again in the study by Grausz (37).

The study by Adamsons et al. in particular provided some insight into why homeothermic reactions are not predicted only by examination of static body temperatures. In this study, the metabolic rates of infants in various thermal environments were determined and correlations computed to determine the relative value of measuring only rectal temperature, skin temperature, or environmental temperature, or only the difference between the skin and environmental temperatures, in reliably predicting what the infant's metabolic rates actually were at the time the temperature measurements were made. The study determined that no correlation existed between rectal temperature and metabolic rate, a slightly better correlation existed between environmental temperature and metabolic rates, a still better correlation with skin temperature existed, and an almost perfect correlation was demonstrated between metabolic rate and the difference between skin and incubator environmental temperatures (Fig. 4). These results can be understood by recalling that when body temperature is stable, heat production or metabolic rate must be equal to the rate of heat loss from the infant. If this balance does not exist, the infant will get either warmer or cooler, depending on the direction of the imbalance. So if heat production equals heat loss and heat loss is proportional only to the difference between the magnitude of the skin and environmental temperatures and not to the magnitudes themselves, it follows that heat production must also be related to the same temperature difference and, similarly, should be relatively independent of any single absolute temperature value.

These discoveries led to an approximate definition of what might constitute an optimal thermal environment in which to raise small infants. This environment is called a neutral thermal environment, referring to that set of thermal conditions existing when an infant is in a minimal metabolic state and has a body temperature that is within a normal range.

From the previous discussion, it might seem reasonable that to provide an infant with a neutral thermal environment it is necessary then only to establish skin-to-environment temperature gradients documented in various published studies to be associated with minimal metabolic rates in infants with normal temperature. Unfortunately, such references can provide only very rough guidelines, since any one infant can achieve different minimal rates of metabolism at different times and, if body temperature is to be kept stable, each change in this minimal achievable heat production rate must be balanced with a change in the amount of heat loss allowed. When the minimal heat production increases, the skin environmental temperature

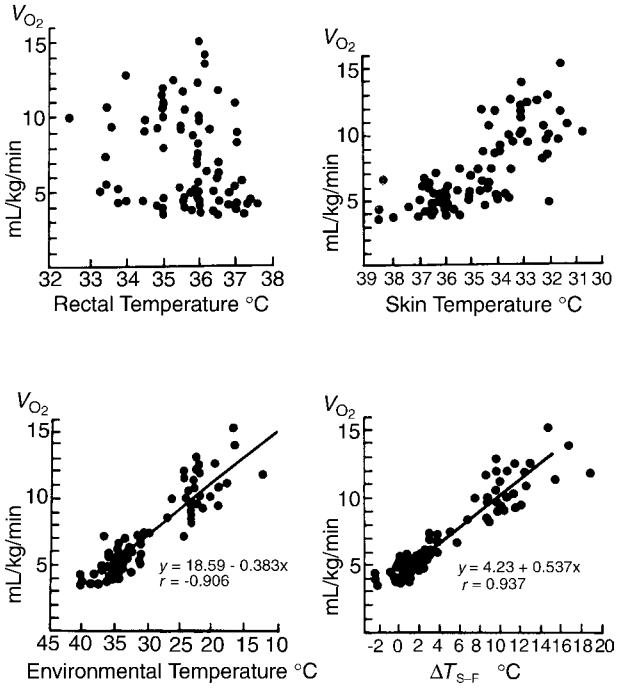


Figure 4. Metabolic rate expressed as oxygen consumption (Vo_2) as correlated with rectal temperatures, skin temperature, incubator environmental temperature, or the difference between the skin and environmental temperature (ΔT_{s-e}). Reproduced by permission from P. H. Perlstein, "Routine and special care—Physical environment." In A. A. Fanaroff and R. J. Martin (Eds.), *Behrman's Neonatal-Perinatal Medicine,* 3rd ed., St. Louis, MO: C. V. Mosby Co., 1983. Adapted from Adamsons et al. (21).

difference needs to be increased, and when the minimal heat production falls, the gradient needs to be decreased. Although concepts such as the neutral thermal environment can be discussed using static equations, it must be remembered that they actually are used only in dynamic settings.

In any case, it is very difficult to provide an infant with truly neutral thermal conditions and becomes practically impossible in some common situations, such as when head hoods or other auxiliary gas delivery devices are used during care. Head hoods are small plastic boxes or tents made to enclose the infant's head when resting on a mattress. The hoods are used to deliver and control the concentration of humidified oxygen to the infant. Since the head of an infant can represent 20% of the infant's body surface, a significant amount of body heat can be lost if the head if the head hood temperature is not carefully controlled. Even if the temperature is controlled by prewarming the oxygen prior to its delivery into the head hood, the infant can lose heat if the gas is delivered at a flow rate that produces an excessive wind chill component to the convective heat flux. It also has been documented that even when total body heat losses are less than usually needed to stimulate a homeothermic response, any local cooling of facial skin tissue can stimulate a baby to become hypermetabolic (36,38).

Since no studies have been published to provide guidelines for establishing neutral thermal conditions in an incubator containing auxiliary sources of heating and cooling, and because such sources are very commonly used during infant care, no incubator manufacturer can guarantee, when an auxiliary devices are used, that such conditions can be produced by any incubator on the market today. As a corollary, unless both body temperatures and infant metabolic rates are continuously monitored, infant caretakers and medical researchers are similarly constrained from claiming precision in their delivery of continuous and certifiable thermal protection that is consistent with the concept of thermoneutrality.

Before we leave this subject, it should also be noted that a significant number of knowledgeable baby care specialists disagree with the idea that a neutral thermal environment represents an optimal goal for incubator control. Their arguments are numerous, but most often include the irrefutable fact that no one has ever documented scientifically that such protection is really beneficial. They also cite the studies by Glass et al. (2,3) in which it was documented that babies tend to lose their very important self-protective ability to react as homeotherms if not exposed to periodic cold stresses. This means that one price paid by an infant raised in a neutral thermal environment is adaptation to that environment and, much as a prolonged stay in the tropics diminishes an adult's capacity to tolerate the northern winters, a baby so adapted may be more susceptible to the damaging effects of unavoidable occasional exposures to the cold.

It must be emphasized that the arguments against the neutral thermal environment are not arguments in favor of letting all babies stay cold; the debate primarily concerns whether a baby is better off in an environment that theoretically maximizes growth by reducing metabolic work to an absolute minimum but might increase the infant's susceptibility to subsequent stresses, or better off in an environment that very mildly stimulates the infant to metabolically contribute to his own ongoing thermal welfare so that important self-protective capabilities are not forgotten. As yet there are insufficient scientific data to resolve this issue. A more recent observation is that the body temperatures of the fetus and older infant are higher than the typical neutral thermal environment proposed for preterm infants, and in both cases, follow a circadian rhythm that is not observed or supported in the typical infant incubator. These considerations imply that while the "neutral thermal environment" is a useful concept for current incubator design, it is not yet known how the "optimal thermal environment" should be defined, especially for the preterm infant.

INCUBATOR STUDIES

In spite of the difficulties encountered when attempting to define, let alone achieve, the optimal environmental conditions that are protective for small babies, it is clear that babies raised in different environments do have different survival rates. The scientific studies documenting these differences in survival in different environments are worth reviewing, since they have provided insight into features distinguishing some incubators from others and ways in which these features may produce environments that can prove to be both protective to some infants and dangerous for others. These studies have also been the major impetus to changes in designs that have resulted in the kinds of incubator devices in use today.

With few exceptions, until the early 1970s, incubator designers relied only on convective heaters to warm the chamber within which a baby was contained. Such devices were relatively simple to construct and provided a method whereby the chamber mattress could be kept warm thereby limiting conductive heat losses, and a method to keep the surrounding air warm, limiting convective losses. The use of wet sponges early in the history, and later evaporation pans, over which the convective currents of warmed air passed before entering the infant's chamber, provided the humidity needed to limit evaporative losses. Additionally, the warmed air contained in the chamber produced some warming of the chamber walls thereby reducing to some degree radiant heat losses from the infant. The heating units in early models of these incubators were controlled only by simple air temperature-sensitive thermostat mechanisms.

Such an incubator with clear plastic walls for enhancing visualization of the contained infant was used in a famous series of infant survival studies published between 1957 and 1965. In this incubator, a fan was used to force the convective air currents into the infant's chamber after passage over a heating element through a turbulence producing baffle resting in a humidifying pan of water. A highlight of these studies was published in 1958 by Silverman et al. (7) who compared the survival rates of two groups of premature infants cared for in this convectively heated and humidified device. For one group of infants the incubator air was heated to 28 °C and for the other group the air was heated to a warmer 32 °C. The study resulted in a conclusion that infants cared for in the warmer incubator had a higher survival rate than did infants cared for in the cooler incubator. During the study it was observed that although babies in the warmer incubator had a greater chance of surviving, not all of the infants who survived in the warmer incubator had warm body temperatures; in fact, 10% of the babies in the 32 °C incubator had body temperatures ~35.5 °C. Dr. Silverman deduced that the reason some of the babies got cold was due to excessive radiant heat losses to the thin plastic chamber walls that were cooled by virtue of their exterior surfaces being exposed to the cooler nursery environment.

Dr. Silverman wished to test whether even better survival rates could be achieved by assuring that an infant was not only in a warm incubator, but that the infant's body temperature also was always kept within a normal range. Along with Dr. Agate and an incubator manufacturer he helped develop a new incubator that was radiantly heated to reduce the radiant losses observed when the incubator was only convectively heated (39). To assure that the contained infant's temperature was maintained within normal range, the new incubator employed an electronic feedback servo-control mechanism that responded to changes in the temperature of a thermistor attached to

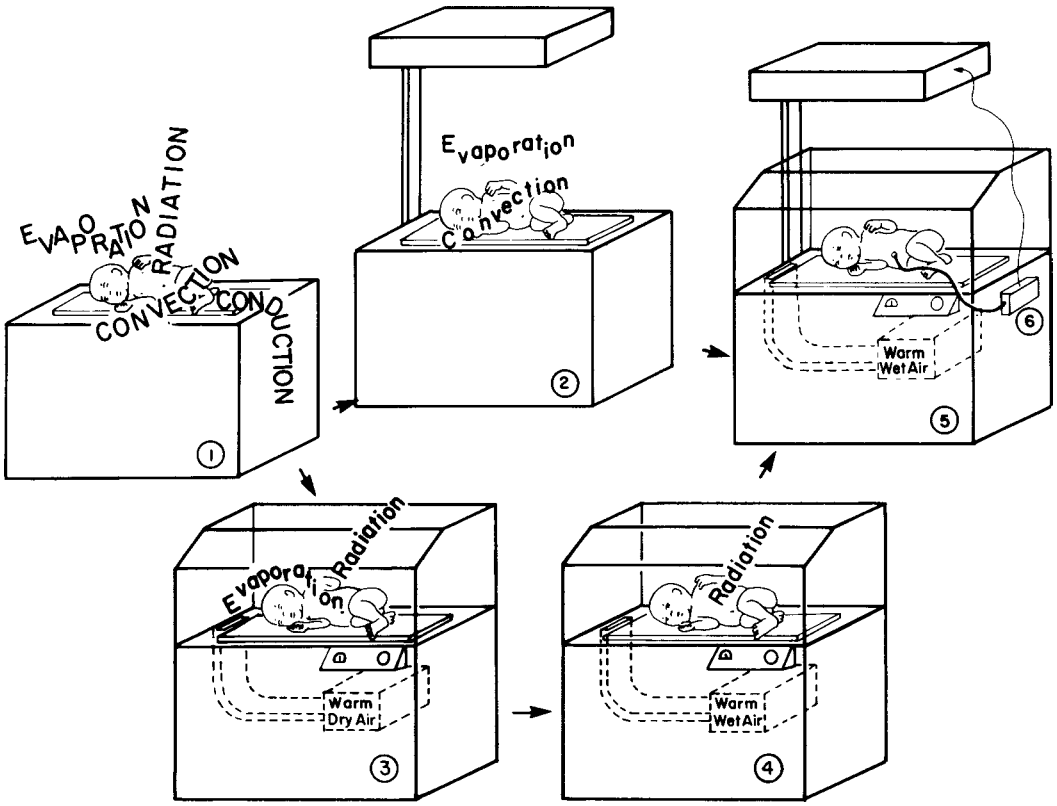


Figure 5. Logic leading to development of skin servo-controlled radiantly heated convectively ventilated incubator. (1) An unprotected baby loses heat from skin surfaces by conduction, convection, evaporation, and radiation. (2) A radiant heater eliminates radiant and conductive losses, but not those caused by convection and evaporation. (3) An unhumidified convectively heated incubator eliminates convective and conductive losses, but not those caused by radiation and evaporation. (4) Humidifying a convectively heated incubator eliminates all major losses except for the losses by radiation. (5) Using a radiant heater to warm a convectively ventilated and humidified incubator should eliminate all sources of heat loss from the infant's skin. (6) Normal infant temperature can be ensured by adding a controller to the incubator so that power is delivered to the radiant heater whenever the infant's skin temperature falls below a preset value. Reproduced by permission from P. H. Perlstein, "Routine and special care—Physical environment." In A. A. Fanaroff and R. J. Martin (Eds.), *Behrman's Neonatal-Perinatal Medicine,* 3rd ed., St. Louis, MO: C. V. Mosby Co., 1983.

the infant's skin surface, causing the incubator's radiant heater to turn on or off, depending on whether the transduced skin temperature value was below or above an absolute temperature value considered normal (Fig. 5).

Note that before settling on a servo-controlled, radiantly heated and convectively ventilated system Agate and Silverman did explore alternative methods whereby an incubator could be equipped to guarantee that an infant's temperature was maintained within a normal range. In particular, they considered simply using the well-established convective heating system in the servo-control loop but rejected this approach when they discovered that when servo controlled in response to changes in skin temperature, the convective heater produced massive and unacceptable changes in air temperature within the incubator chamber. The servo-controlled radiant heater, however, produced an environment in which the air temperature was quite stable, especially when compared to the thermal cycling recorded within the convectively heated servo-controlled system.

The radiantly heated, convectively ventilated, and skin servo-controlled enclosed incubator was evaluated in two independent studies, with results published in 1964 (9,10). In these controlled trials, premature infants were divided into two groups: one group of infants was provided care in the new radiantly heated incubator that was servo controlled to maintain the contained infant's skin temperature at 36 °C, while the other group of like babies was cared for using the simpler 32 °C air temperature thermostat-controlled convectively heated incubator that Silverman's group concluded was the best incubator setup in their previous study published in 1958. The two studies in 1964 reached a common conclusion; the skin temperature-controlled radiantly heated system produced higher survival rates when used during the care of the infants studied. Because of fabricating difficulties, though, this radiantly heated incubator model was commercially marketed for only a short period of time; soon after its introduction, it was replaced on the commercial market by a skin servo-controlled convectively heated enclosed

incubator that was easier to fabricate and, like the radiantly heated device, was also capable of keeping an infant's skin temperature at a value considered normal.

The introduction of this convectively heated servo-controlled device on the market was justified by an extrapolated interpretation of the studies reported in 1964. This common interpretation led to a conclusion that the studies simply demonstrated that, in terms of survival, it was only important to keep a baby's skin temperature warm and stable. The interpretation ignored the fact that more than just a difference in skin temperatures distinguished the two study groups. Infants in the two different study environments, the one produced by an air temperature referenced thermostat controlling a convective heater and the other by a skin temperature referenced servo system controlling a radiant heater, were also, as previously well discussed by Agate and Silverman, exposed to environments that differed in the frequency and amplitude of thermal cycling produced by the different systems (39). The radiantly protected infants, who survived better, not only had warmer and more stable skin temperatures as a group, but were also exposed to a much more stable environment than were the convectively protected infants with the less favorable group outcomes.

The commercially released convectively heated and skin temperature referenced servo-controlled incubator was the most common incubator in clinical use during the late 1960s; not until 1970 was the characteristic rapidly changing air temperatures within the incubator chamber redescribed and shown to cause some sick small babies to stop breathing (40). These episodes of respiratory arrest, called apneic spells, were specifically observed during the incubator's heating cycles. The mechanism whereby a sudden rise in temperature causes some babies to become apneic remains unknown, but was an observation well reported even prior to the 1970 publication. Even without knowing the mechanism of this relationship, it remains undisputed, so incubator manufacturers have continued to search for ways to produce stabilization of the incubator environmental temperatures (Fig. 6).

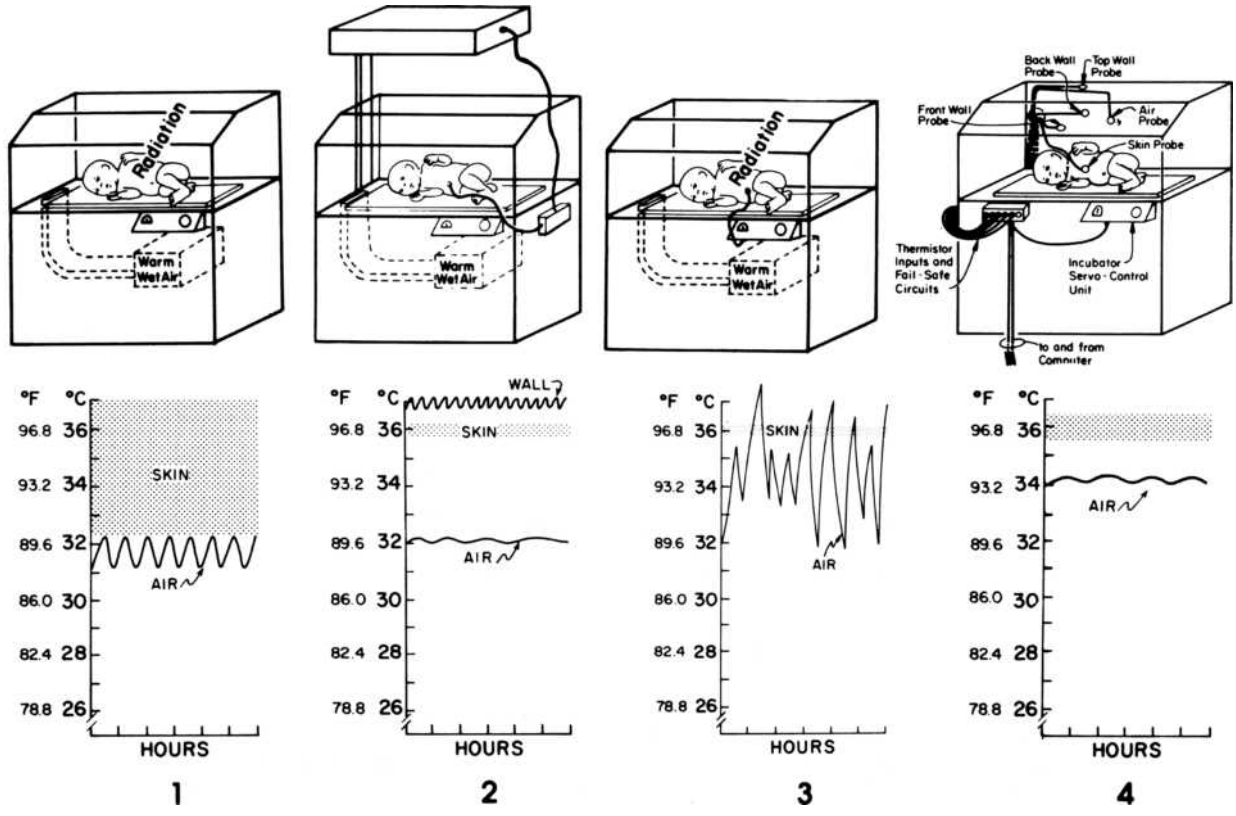


Figure 6. Skin and air temperature characteristics recorded using four different incubator systems.(1) A convectively heated and humidified incubator in which air temperature is thermostatically controlled. This was the device studied by Silverman in 1958 (7). Note cyclie variations in air temperature and wide variation in recorded skin temperatures.(2) A radiantly heated convectively ventilated and humidified incubator that is servo controlled to maintain skin temperature at specified value. This was the device studied by Day (9) and Beutow (10) in 1964. Note that the walls are warm, limiting radiant heat losses and that air temperature is stable and skin temperature variations are minimal. (3) A convectively heated and humidified incubator in which the air temperature heating is servo controlled to maintain skin temperature at specified value. This was the device reported to cause some babies to stop breathing (40) as a response to erratic heating cycles that produces rapid increases in air temperature. (4) A convectively heated and humidified incubator that is computer controlled using Alcyon algorithm (11). Note the stable air temperature and minimal variability in skin temperature.

INCUBATOR DYNAMICS

There are many reasons why attempts to stabilize incubator heating have been only partially successful. It is fairly simple to build a box within which the temperatures can be predictably controlled and stabilized for prolonged periods of time if the box is never opened and the thermal characteristics of both the box and its contents never change, but these simplifying conditions never exist when caring for a sick infant. When infants are cleaned, fed, examined, or otherwise cared for, they must be touched, and the incubator box must be entered. Infant's body positions frequently change, exposing different surface areas with different shapes to the incubator environment causing changes in convective flow patterns and altering the view factors influencing radiant heat flow to the incubator walls. Incubator openings necessitated by the need to touch a contained infant cause both environmental cooling and increased infant heat loss that, in an incubator heated in response to either air temperature or infant temperature changes, causes a logical increase in the incubator's heat output. If any such heating requirement is sustained, the incubator heating element warms to the point where it will retain and release heat even after the incubator is closed and the need for additional heating has passed. Such heat retention and release contributes to what is commonly referred to as the thermal lag characteristic of a heating system and can cause temperatures to overshoot, that is to rise above the temperature level targeted when the heater was activated. This is the same phenomenon observed when an electric stove is turned off, and yet the heating element continues to glow brightly prior to cooling. As with an electric stove heating element, the heater in an incubator is not only slow to cool, but also relatively slow to warm up when first energized after the heater is turned on. Again, just as unintended overheating can occur, the thermal lag due to the mass of the heater can result in an incubator getting colder than intended because of this characteristic delay between action and reaction.

Besides the heater element, there are numerous other places where heat is stored in the incubator system. For example, heat storage occurs in the water used for humidification and in the air masses between the heater and the incubator chamber. Since these must be heated to a temperature higher than the incubator chamber in order to raise the chamber temperature, the heat stored in these parts also will continue to raise the chamber temperatures even after the heater power supply has been turned off. Conversely, when the heater power is turned back on, not only the heater but also the air in the spaces leading to the chamber must heat up before the temperature of the air in the chamber can rise.

It is these delays between the time the heater power is changed and the time the air temperature responds that determines the magnitude and frequency of the air temperature cycles to which an incubated infant is exposed. Thermal lag obviously contributes to the tendency for incubator environments to become unstable and, thereby, potentially threatening to the contained infant. Many hardware and logical software solutions to this problem have been tried in commercially available devices, but all have been frustrated by the complex nature of newborn care, which results in a degree of unpredictability beyond the compensating capability of any solution thus far tried. The implementation of feedback control on incubator heating is an example of one way to attempt to respond to many of these problems. However, examining how servo mechanisms actually work in a little more detail provides a deeper appreciation of why this logical technology often fails in the dynamic setting of an incubator and may even contribute to instability in the incubator environment.

Feedback Control

Feedback control systems are commonly referred to as closed loop control systems, as contrasted to open loop systems. A cooking stove again can be used to give an example of each type of control system. Stove top heaters are typically controlled by an open loop system. That is, a dial is adjusted controlling the quantity of gas or electricity going to the heater unit. In this manner, a fixed rate of heat production by the heater unit is specified. This is called an open-loop control system because after once setting the rate of heat production, the heater will continue to produce the same heat output regardless of how hot the object on the stove gets.

In contrast, the oven of a modern stove is equipped with a closed-loop temperature control system, in which a dial is adjusted to specify the desired oven temperature. In control system parlance, this specified temperature is referred to as the set point. A temperature sensor inside the oven works in conjunction with the temperature setting dial to control the rate of heat production in the oven heating unit and when the temperature measured by the oven temperature sensor rises to the set point value on the control dial, the oven heat production is reduced or stopped entirely. After the heat production is stopped, the oven temperature slowly falls as the heat escaped from the oven to the surrounding area. At some point, the oven temperature will fall below the temperature set on the control and the heater will again be turned on; this on–off cycling will continue as long as the oven is in operation.

Feedback Control and Incubators

An incubator heated in automatic feedback response to changes in electronically transduced infant temperature is called an infant skin servo-controlled (ISC) incubator. In one type of ISC incubator the heater is instructed to turn completely on when the infant's skin temperature falls below a preset lower limit or turn completely off when skin temperature exceeds a defined upper limit. Because power applied to the heater is either maximal or zero, this form of servo mechanism is called nonlinear or "on–off control". Another form of control is designated as linear proportional servo control. In a proportional control system the amount of power applied to the incubator heater is graduated in a manner to be proportional in some linear fashion to the transduced skin temperature deviation from a predetermined value. The amount of power actually applied to the heater for each incremental change in skin temperature can be different in different realizations of this control method, and this increment determines the amount of

deviation in skin temperature that can occur before full or zero power is applied.

The theoretical advantage of a proportional over an on–off servo-control system is the possibility of limiting the tendency that a large heating element has to overshoot or undershoot a desired heater temperature. In a proportional system, the heater is turned off slowly as the infant's skin temperature warms toward the set point temperature. This contrasts with an on–off system that remains fully energized until the set point is reached. In an incubator system, a properly designed skin temperature referenced proportional control unit will produce very stable environmental temperatures as long as the infant is stable, the temperature sensing the thermistor attached to the skin remains undisturbed, and the incubator chamber is kept closed and protected from outside environmental perturbations. In a clinical setting such stable and undisturbed conditions exists infrequently, so the theoretical advantages of proportional control over on–off control are difficult to demonstrate. In fact, the cycling of temperatures recorded within a proportionally controlled incubator in a dynamic clinical setting can be indistinguishable from that recorded in an on–off controlled incubator, and this functional behavior is predicted in basic feedback control theory (Fig, 7).

The thermal cycles recorded in servo-controlled incubators are produced as a combined consequence of (*1*) an inherent characteristic of closed-loop feedback systems; (*2*) periodic extreme perturbations of the skin thermistor that trigger either full or zero energization of the incubator heater; and (3) the incubator's thermal lag characteristic, which causes repetitive over-and undershoot of temperatures targeted by the control logic. Following the initiation of a cycling pattern, the time it takes the system to settle down and reestablish stable control is variable and related to the heat-dissipating speed of the heater material and the thermal buffering capability of the homeothermic infant. In some situations, the cycling, when induced by a single perturbation, has been observed to continue for many hours and, when an incubator is repeatedly perturbated, for many days. The published evidence that some babies react to these thermal cycles by becoming apneic justifies a reminder that these characteristic thermal cycles represent a profound problem negating some of the advantages intended when feedback control is applied in incubator designs. At least in incubator servo systems that use skin temperature as a reference value to determine heater status, even the theoretical negative effects often outweigh any theoretical or demonstrable effects that are positive. This can be appreciated by recalling that a *sine qua non* of optimal control in a skin temperature referenced servo system is the reliable and accurate measurement of an infant's skin temperature and, using today's technology in a clinical setting, the ability to transduce skin temperatures accurately for prolonged periods of time is marginal at best and, perhaps, even impossible.

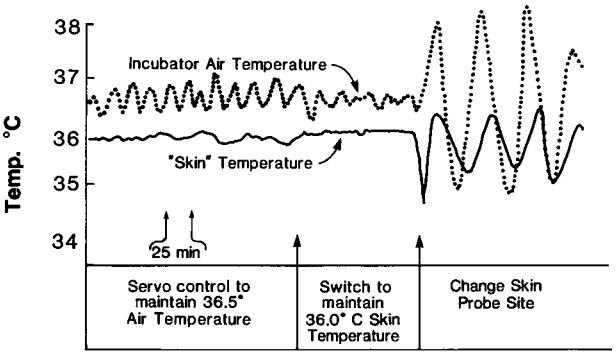


Figure 7. Example of variations in dynamic air temperature changes when an enclosed incubator that was initially servo controlled to maintain a stable air temperature was switched to a skin temperature referenced servo control mode and then was perturbated by changing the site of attachment of the skin temperature-sensing thermistor. The wide thermal cycling illustrated as a consequence of this sequence continued for 3 h before returning to a more stable pattern.

Skin Temperature Measurement

Skin temperature measurement accuracy is limited by variability in the characteristics of infant, transducers, and infant care techniques. The surface temperatures of infants are not homogeneous because of difference in (1) skin and subcutaneous tissue thickness over different body parts, (2) differences in structures underlying different skin surfaces, and (3) difference in the vascularity and vasoreactivity-characterizing different body regions. Different skin surfaces have different temperatures, and when temperature transducers are connected to the skin surface, they measure the temperature only at the specific site of their connection. Moreover, thermistors are attached using devices that can compression of superficial skin vessels underlying the thermistor element. Thus, by their very attachment, thermistors modify both the absolute temperature they are expected to measure and the spontaneous dynamic variability in the measured skin temperature that is normally affected by the changing amounts of warm blood flowing through the skin over different time periods. Additional factors also affect thermistor accuracy. Thermistors are faulted as precise skin temperature measuring devices because they are manufactured in various shapes and sizes and are protected using different materials so that each affects transduction in a specific and different way. Thermistors also generally measure temperature three dimensionally (3D). They are affected not only by the temperature of the surface to which they are attached, but also by the environment to which their unattached surfaces are exposed. Depending on the amount and type of insulation used in their manufacture, they are also affected by the temperature of the wires used to connect the thermistor to electronic signal conditioners.

These inherent characteristics of thermistors, added to the fact that they are freely moved during clinical care from one site of attachment to another, provide sufficient cause to explain why the skin temperature-dependent heater-controlling servo mechanism in an incubator can easily be directed into an unstable state that produces thermal cycling in the environment. This environmental instability is even further exacerbated by infant care practices that,

for examples, cause thermistors to be removed from the skin when X rays are taken, or to be covered with sterile towels during surgical procedures. During such care-related events, the thermistor is fooled into measuring environmental and not skin temperature. Obviously, if the thermistor provides the servo electronics with misinformation, the servo control decisions based on this information can be nonsensical.

THE NONTHERMAL ENVIRONMENT

Although infant incubators were initially designed solely to maintain body temperature in high risk infants, they are now seen in a much more complex role, as devices that provide a complete "microenvironment" for high risk infants. To one extent or another, they are used to modify the sensory input an infant receives through visual, auditory, olfactory, and kinesthetic pathways. In addition, incubators are now appreciated as the source of exposure to potentially unwanted environmental toxins, such as electromagnetic radiation (EMR) and chemical compounds used in the manufacture and operation of the incubator. Closed incubators are often used as delivery systems for oxygen and humidification, sometimes in ways unanticipated at the time of their design and manufacture. Incubators have been developed for specialized purposes, such as transport, use in an magnetic resonance imaging (MRI) suite, or cobedding twin infants. Incubators have increasingly been designed as a platform for a modular system of support equipment including ventilators, IV pumps, and monitors. As these devices become increasingly controlled by digital components, they also gain the capability of integrating their data output, so that incubator-derived information, such as air temperature and the infant's skin temperature, can be continuously recorded in an electronic medical record. Incubators have had a role in infection control since their earliest days, but this function is now being increasingly emphasized as infection has emerged as the leading cause of late morbidity in premature infants. These multiple functions of modern incubators increase their complexity exponentially, since many of these factors interact with one another, often in ways unanticipated by either the designers or the users. Because of the recent nature of these nonthermal applications of the infant incubator, there is only a limited scientific foundation to guide designers and caregivers, so not all of these topics will be discussed in greater depth below.

The Infant Incubator as a Sensory Microenvironment

Although the comparison of an incubator to a uterus, the environment it replaces, has been noted since the earliest days of incubator design, it will be evident from the preceding sections of this article that temperature regulation has been the first and primary design consideration: and necessarily so, since it had immediate implications for infant survival. When incubators became used as devices for delivery of supplemental oxygen in the mid-twenteth century, morbidity and mortality were again the primary endpoints: first for improved survival as the benefits of oxygen supplementation were identified, and then for increased morbidity as an epidemic of oxygen-induced blindness from retinopathy of prematurity followed, again chronicled most eloquently by Dr. Silverman (41). Only recently have the other environmental features of the uterus been compared to the micro and macro environments of the NICU and the implications for design been considered.

Taste, smell, and touch are the earliest fetal senses to develop, beginning in the second trimester of pregnancy, followed closely by auditory development, and finally by visual development as the baby approaches term gestation. While thus far there appears to be no reason to suspect that the sense of taste is stimulated or influenced by the incubator, there is accumulating evidence that the other senses are indeed affected by this microenvironment.

Infants can develop a conditioned response to certain odors that they come to associate with painful procedures, such as alcohol, whereas a pleasant odor has been shown to reduce apnea, and babies will orient preferentially to odors from their mother (42).

In utero, infants are exposed to a fluid environment, frequent movement with circadian rhythmicity, and as they approach term, increasing contact with a firm boundary, features absent in the typical incubator. After-market adaptations that caused the bed or the entire incubator to move in one fashion or another have been introduced sporadically, often with the intent of reducing infant apnea, but none have been documented to be efficacious. Additional modifications of the infant mattress to make it more suitable for the skin and developmental needs of preterm infants have been introduced, again without clear benefit to this point.

Sound is a constant although variable stimulus in the uterus, and different in many ways from that of the modern incubator. *In utero* sounds are delivered via a fluid medium where lower pitched sounds predominate, and the mother's voice, heartbeat and bowel sounds are far more prevalent than any extraneous noise. As such, there is a definite circadian rhythm both to the sounds and to movement associated with them. In the closed incubator, the predominant sound is that of the incubator fan that produces a constant white noise, usually in excess of 50 dB, a level that has been shown to interfere with infant sleep (43) and speech recognition (44). Depending on the NICU in which it is used, this may be louder or softer than the NICU ambient noise level, so the closed incubator may be used as a haven from a noisy NICU, or itself could be noisier than the external environment, in which case the open radiant warmer might provide a more suitable auditory environment. A number of after-market modifications have been suggested to reduce both fan noise and intrusion of noise from outside the incubator, including blankets or quilts to cover the incubator and sound-absorbing panels (45,46), but their use is problematic to the degree that they affect air flow and other operating characteristics of the incubator.

The visual environment of the incubator may be presumed to be neutral, but incubators are often used to control the visual environment of the NICU, particularly in conjunction with the use of incubator covers, to produce a dimly lit environment which may enhance infant sleep (47). Light penetration into the incubator may also affect

circadian rhythmicity in the preterm infant (48), and may be a source of heat gain through the greenhouse effect.

Electromagnetic Radiation in the Infant Incubator

Any electrically powered device emits electromagnetic radiation (EMR), usually at intensities far below those considered to constitute a risk. Since the organs of preterm infants are in a crucial and rapid phase of growth, however, concern about EMR emitted by incubator and radiant warmer heaters and other components has merited special attention. Several studies have documented EMR levels in incubators, with proposed strategies to reduce exposure including shielding panels (49) and increasing the distance between the baby and the EMR source (50).

Incubators for Specialized Purposes

The conventional closed incubator or radiant warmer is used as a static device in the NICU, but the same needs for temperature control and a safe microenvironment exist for infants who require transport from one hospital to another, or to an MRI suite. Transport incubators place a premium on space (so that multiple modular components can be mounted) and weight (especially those used for air transport). Incubators developed for use in an MRI suite must be similarly portable, but use almost exclusively plastic materials and have an integrated coil for scanning (51,52).

SUMMARY

Infant incubators are specially heated devices that provide a bed surface or chamber within which an infant can be cared for and kept warm. Throughout this article both the positive and negative features of today's incubators have been noted and placed into both a context of what is theoretically desirable and of what is practically feasible. It is apparent that, at present, our knowledge of infant physiology and the availability of technical solutions are severely limited, and that all existing incubator devices reflect and are faulted by these limitations. In historical perspective, however, it is clear that incubators over the past 100+ years have steadily been improved by manufacturers to incorporate new items of knowledge and technology as they become available. This historical path has been fruitful and provides, in its continuing course, direction for the future in incubator development. Future iterations of the infant incubator will need to incorporate new information on the optimal microenvironment for the high-risk newborn as well as new capabilities made possible by the ongoing quantum changes in digital technology.

BIBLIOGRAPHY

1. Meystán J, Járai I, Fekete M. The total energy expenditure and its components in premature infants maintained under different nursery and environmental conditions. Pediatr Res 1968;2:161.
2. Glass L, Silverman WA, Sinclair JC. Effect of the thermal environment on cold resistance and growth of small infants after the first week of life. Pediatrics 1968;41:1033.
3. Glass L, Silverman WA, Sinclair JC. Relationship of thermal environment and caloric intake to growth and resting metabolism in the late neonatal period. Biol Neonate 1969;14:324.
4. Gandy GM, et al. Thermal environment and acid base homeostasis in human infants during the first few hours of life. J Clin Invest 1964;43: 751.
5. Cornblath J, Schwartz R. Disorders of carbohydrate metabolism in infancy. In: Shaffer JA, editors. Major Problems in Clinical Pediatrics. Vol. 3. Philadelphia: Saunders; 1966. p 34.
6. Silverman WA, Blanc WA. The effect of humidity on survival of newly born premature infants. Pediatrics 1957;20:477.
7. Siverman WA, Ferrig JW, Berger AP. The influences of the thermal environment upon the survival of newly born premature infants. Pediatrics 1958;22:876.
8. Silverman WA, Agate FJ, Ferrig JW. A sequential trial of the nonthermal effect of atmospheric humidity on survival of newborn infants of low birth weight. Pediatrics 1964;34:171.
9. Day RL, Caliguiri L, Kamenski C, Ehrlich F. Body temperature and survival of premature infants. Pediatrics 1964;34: 171.
10. Beutow KC, Klein SW. Effect of maintenance of normal skin temperature on survival of infants of low birth weight. Pediatrics 1964;34:163.
11. Perlstein PH, Edwards NK, Atherton HD, Sutherland JM. Computer assisted newborn intensive care. Pediatrics 1976; 57:494.
12. Evaluation of infant radiant warmers. Health Devices, 1973;3:4.
13. Perstein PH. Thermal control. Rep Ross Conf Pediatr Res 1976;69:75.
14. Edwards NK. Radiant warmers. Rep Ross Conf Pediatr Res, 1976;69:79.
15. Evaluation of infant incubators. Health Devices. 1981;11:47.
16. Evaluation of infant incubators. Health Devices 1982;11:191.
17. Wu PYR, Hodgman JE. Insensible water loss in preterm infants. Pediatrics 1974;54:704.
18. Cone Jr. TE, History of the Care and Feeding of the Premature Infant. Boston: Little, Brown; 1985.
19. Scopes JW. Metabolic rate and temperature control in the human body. Br Med Bull 1966;22:88.
20. Scopes JW, Ahmed I. Minimal rates of oxygen consumption in sick and premature newborn infants. Arch Dis Child 1966;41:407.
21. Adamsons K Jr., Gandy GM, James LS. The influence of thermal factors upon oxygen consumption of newborn infants. J Pediatr 1965;66:495.
22. Hill JR, Rahimtulla KA. Heat balance and the metabolic rate of newborn babies in relation to environmental temperature: And the effect of age and weight on basal metabolic rate. J Physiol(London) 1965;180:239.
23. Dawes GS. Oxygen consumption and temperature regulation in the newborn. I Foetal and Neonatal Physiology. Chicago: Year Book Medical Publishersh; 1968. p 191.
24. Cross K, et al. Lack of temperature control in infants with abnormalities of the central nervous system. Arch Dis Child 1971;46:437.
25. Dawkins MJR, Hull D. Brown fat and the response of the newborn rabbit to cold. J Physiol (London) 1963;169:101.
26. Hill JR. Oxygen consumption of newborn and adult mammals: Its dependence on oxygen tension in inspired air and on environmental temperatures. J Physiol (London) 1959;149: 346.
27. Cree JE, Meyer J, Hailey DM. Diazepam in Labour: Its metabolism and effect on the clinical condition and thermogenesis of the newborn. Br Med J 1973;3:251.

28. Brück K. Heat production and temperature regulation. In: Stave U, editor. Pernatal Physiology. New York: Plenum Press; 1978. p 474.
29. Day RL. Respiratory metabolism in infancy and childhood. Am J Child 1943;65:376.
30. Hey EN, Katz G. Evaporative water loss in the newborn baby. J Physiol (London) 1969;200:605.
31. Sulyok E, Jéquier E, Ryser G. Effect of relative humidity on thermal balance of the newborn infant. Biol Neonate 1972; 21:210.
32. Belgaumkar TR, Scott KE. Effects of low humidity on small premature infants in servo control incubators. Biol Neonate 1975;26:337.
33. Hammarlund K, Nilsson GE, Öberg PA, Sedin G. Transepidermal water loss in newborn infants. Acta Paediatr Scand 1977;66:553.
34. Hammarlund K, Nilsson GE, Öberg PA, Sedin G. Transepidermal water loss in newborn infants: Evaporation from the skin and heat exchange during the first hours of life. Acta Paediatr Scan 1980;69:385.
35. Hey EN, Mount LE. Heat losses from babies in incubators. Arch Dis Child 1967;42:75.
36. Brück K. Temperature regulation in the newborn infant. Biol Neonate 1961;3:65.
37. Grausz JP. The effects of environmental temperature changes on the metabolic rate of newborn babies. Acta Paediatr. Scand 1968;57:98.
38. Mestán J, Jrai I, Bata G, Fekete M. The significance of facial skin temperature in the chemical heat regulation of premature infants. Biol Neonate 1964;7:243.
39. Agate FJ, Silverman WA. The control of body temperature in the small newborn infant by low-energy infra-red radiation. Pediatrics 1963;37:725.
40. Perlstein PH, Edwards NK, Sutherland JM. Apnea in premature infants and incubator air temperature changes. N Engl J Med 1970;282:461.
41. Silverman WA. The lesson of retrolental fibroplasias. Sci Am 1977;236:100.
42. Schaal B, Hummel T, Soussignan R. Olfaction in the fetal and premature infant: Functional status and clinical implications. Clin Perinatol. 2004;31:261.
43. Morris BH, Philbin MK, Bose C. Physiologic effects of sound on the newborn. J Perinatol 2000;20:S55.
44. Robertson A, Stuart A, Walker L. Transmission loss of sound into incubators: implications for voice perception by infants. J Perinatol 2001;21:236.
45. Johnson AN. Neonatal response to control of noise inside the incubator. Pediatr Nurse 2001;27:600.
46. Bellini CV, et al., Use of sound-absorbing panel to reduce noisy incubator reverberating effects. Biol Neonate 2003;84:293.
47. Hellstrom-Westas L, et al., Short-term effects of incubator covers on quiet sleep in stable premature infants. Acta Paediatr 2001;90:1004.
48. Rivkees SA. Emergence and influences of circadian rhythmicity in infants. Clin Perinatol 2004;31:217.
49. Bellieni CV, et al. Reduction of exposure of newborns and caregivers to very high electromagnetic fields produced by incubators. Med Phys 2005;32:149.
50. Bellieni CV, et al. Increasing the engine-mattress distance in neonatal incubators: A way to decrease exposure of infants to electromagnetic fields. Ital J Pediatr 2005;29:74.
51. Blumi S, et al. MR imaging of newborns by using an MR-compatible incubator with integrated radiofrequency coils: Initial experience. Radiology, 2004;231:594.
52. Whitby EH, et al. Ultrafast magnetic resonance imaging of the neonate in a magnetic resonance-compatible incubator with a built-in coil. Pediatrics 2004;113:e150.

Further Reading

Adamsons K. The role of thermal factors in fetal and neonatal life. Pediatr Clin North Am 1966;13:599.
Ahlgren EW. Environmental control of the neonate receiving intensive care. Int Anesthesiol Clin 1974;12:173.
Brück K. Heat production and temperature regulation. In: Stave U, editor. Perinatal Physiology New York: Plenum Press; 1978 p 455.
Dawes GS, Oxygen consumption and temperature regulation in the newborn. I Foetal and Neonatal Physiology. Chicago: Year Book Medical Publisher; 1968. p 191.
Delue NA. Climate and environment concepts. Clin Perinatal 1976;3:425.
Hey EN, Katz G. The optimum thermal environment for naked babies. Arch Dis Child 1970;45:328.
Holman JP. Heat Transfer, New York: McGraw-Hill; 1981.
Klaus M, Fanaroff A, Martin RJ. The physical environment. In: Klaus MH, Fanaroff AA, editors. Care of the High Risk Neonate. Philadelphia: Saunders; 1979. p 94.
Lutz L, Perlstein PH. Temperature control in newborn babies. Nurs Clin North Am 1971;6:15.
Mayr O. The Origins of Feedback Control. Cambridge, (MA): MIT Press; 1970.
Ogata K. Modern Control Engineering, Englewood Cliffs (NJ): Prentice-Hall; 1970.
Oliver TK. Temperature regulation and heat production in the newborn. Pediatr Clin North Am 1965;12:765.
Oppenheim AV, Willsky A, Young IT. Signals and Systems, Englewood Cliffs (NJ): Prentice-Hall; 1983.
Perstein PH. Thermal regulation. In: Fanaroff AA, Martin RJ, editors. Behrman's Neonatal-Perinatal Medicine, 3rd ed. St. Louis (MO): Mosby; 1983. p 259–277.
Scopes JW. Thermoregulation in the newborn. In: Avery GB, editors. Neonatology, Philadelphia: Lippincott; 1975. p 99.
Sinclair JC. The effect of the thermal environment on neonatal mortality and morbidity. In: Adamson K, Fox HA, editors. Preventability of Perinatal Injury. New York: Alan R. Liss; 1975. p 147.
Sinclair JC. Metabolic rate and temperature control. In: Smith CA, Nelson NM, editors. The Physiology of the Newborn Infant. Springfield (IL) : Thomas; 1976. p 354.
Todd JP, Ellis HB. Applied Heat Transfer. New York: Harper & Row; 1982.

See also BIOHEAT TRANSFER; NEONATAL MONITORING; TEMPERATURE MONITORING.

INFANT INCUBATORS.

See INCUBATORS, INFANT.

INFORMATION SYSTEMS FOR RADIOLOGY.

See RADIOLOGY INFORMATION SYSTEMS.

INFUSION PUMPS.

See DRUG INFUSION SYSTEMS.

INTEGRATED CIRCUIT TEMPERATURE SENSOR

TATSUO TOGAWA
Waseda University
Saitama, Japan

INTRODUCTION

Temperature can affect the electronic characteristics of semiconductor devices. Although this is a disadvantage

in many applications, especially for analogue devices, it may be turned into an advantage if such a device is used as a temperature sensor. In principle, any parameter in such a device having a temperature coefficient can be used for temperature measurement. For example, a temperature telemetry capsule, in which a blocking oscillator frequency varies with temperature, has been developed for measuring gastrointestinal temperature (1). In this system, the temperature affects the reverse-bias base-collector current, which determines the period of relaxation oscillation. However, it has been shown that the voltage across a p–n junction of a diode or transistor under a constant forward-bias current shows excellent linear temperature dependency over a wide temperature range. Many conventional or specially designed diodes or transistors composed of Ge, Si, or GaAs have been studied for use as thermometers (2–4).

The advantages of diodes and transistors as temperature sensors are their high sensitivity and low nonlinearity. The temperature sensitivity under normal operation is ca -2 mV/K, which is $\sim$50 times higher than that of a copper-constantan thermocouple. The nonlinearity is low enough for many applications, although its value depends on the structure and material of the device. It is known that a Schottky diode, which has a structure composed of a rectifying metal-semiconductor contact, possesses good voltage–temperature linearity (5). Some transistors used as two-terminal devices by connecting the base to the collector also possess good linearity (6,7), and a transistor that has been especially developed for temperature sensing is commercially available (8). This has a linearity that is comparable to that of a platinum-resistance temperature sensor.

It is advantageous to have a diode and a transistor temperature sensor fabricated on a chip with associated interfacing electronics using integrated circuit (IC) technology. Several integrated temperature sensors that provide either analogue or digital outputs have been developed and are commercially available. A diode or transistor temperature sensor fabricated on a central processing unit (CPU) chip is especially useful when used to monitor the temperature of the chip. Such a sensor has been used to detect overheating, and to protect the CPU system by controlling a fan used to cool the chip or to slow down the clock frequency.

THEORY

The characteristics of p–n junctions are well known (9,10). In p–n junction diodes, the current flowing through the forward-biased junction is given by

$$I = I_s(e^{qV/mkT} - 1) \tag{1}$$

where I_s is the saturation current, q is the electron charge, V is the voltage across the junction, k is the Boltzmann constant, m is the ideality factor having a value between 1 and 2, which is related to the dominant current component under the operating conditions used, and T is the absolute temperature. At a temperature close to room temperature, and when the current is relatively high, so that the current due to the diffusion of the carrier dominates, $m = 1$, and so the second term in Eq. 1 given in parentheses can be neglected. Equation 1 can then be simplified to

$$I = I_s e^{qV/kT} \tag{2}$$

The temperature dependence of the saturation current, I_s, is given by

$$I_s = Ae^{-E_g/kT} \tag{3}$$

where E_g is the bandgap energy at $T = 0$ K, and A is a constant dependent on the geometry and material of the device. Strictly speaking, A also depends on the temperature. However, the temperature dependency is very weak compared to the exponential term in Eq. 3. Thus,

$$I = A\mathrm{e}^{(qV-E_g)/kT} \tag{4}$$

For a constant current, I, $(qV–E_g)/kT$ is constant. Thus, the voltage across a p–n junction, V, is a linear function of the absolute temperature, T. On extrapolating to $T = 0$, then $qV = E_g$.

The temperature coefficient of V can be derived from Eq. 4 as

$$\left.\frac{dV}{dT}\right|_{I=\text{const}} = \frac{V - E_g/q}{T} \tag{5}$$

Since the value of $qV–E_g$ is always negative, V decreases with increasing T. For silicon, $E_g \sim$1.17 eV, and for $T \sim$300 K, $V \sim$600 mV, and $dV/dT \sim -1.9$ mV/K. In actual diodes, the current–voltage characteristics have been studied in detail over a wide temperature range. The forward voltage exhibits a linear dependence for $T > 40$ K for a constant current (11). The observed value of dV/dT in a typical small signal silicon p–n junction diode ranges between -1.3 and -2.4 mV/K for $I = 100$ μA (11). In germanium and gallium arsenide p–n junction diodes, and for silicon Schottky diodes, the forward voltage exhibits a similar sensitivity (3–5).

In most p–n junctions, the current through the junction contains components other than those due to carrier diffusion, and therefore, Eq. 4 does not hold. The base-emitter p–n junction in transistors is advantageous in this respect. Here, the diffusion component forms a larger fraction of the total current than that in diodes, even for a diode connection in which the base is connected to the collector. The nonlinear temperature dependence in the forward voltage in diode-connected transistors is lower than that of most diodes (7). Further improvement in linearity is attained under constant collector current operation, since only the diffusion component flows to the collector, while other components flow to the base (12).

From Eq. 2, one can obtain the following expression

$$\ln I = \ln I_s + qV/kT \tag{6}$$

The value of T can be obtained from the gradient of a plot of $\ln I$ versus V, as q and k are known universal constants. This implies that the current–voltage characteristics can be used as an absolute thermometer (6). If $\ln I$ is a linear function of V, only two measurements are required to determine the gradient. If V_1 and V_2 are voltages corre-

sponding to different current levels, I_1 and I_2, the difference between these two voltages is calculated using

$$V_1 - V_2 = (kT/q)\ln(I_1/I_2) \quad (7)$$

Thus, the difference in voltage corresponding to the different current levels for a constant ratio is proportional to the absolute temperature, without any offset. Using this relationship, a thermometer providing an output proportional to the absolute temperature can be realized, either by applying a square wave current to a p–n junction (12), or by using two matched devices operating at different current levels (13).

FUNDAMENTAL CIRCUITS AND DEVICES

A schematic drawing of the fundamental circuit of the thermometer with a short-circuited transistor or a diode is shown in Fig. 1. A constant current is applied to the transistor or diode in the forward bias direction, and the voltage across the junction is amplified using a differential amplifier. By adjusting the reference voltage applied to another input of the differential amplifier, an output voltage proportional to either the absolute temperature in kelvin or in degrees Celsius or any other desired scale can be obtained. The operating current of small signal diodes and transistors is typically 40–100 A. If the current becomes too high, a self-heating error may be produced due to the power dissipated in the junction. If the current becomes too small, problems due to leakage and the input current of the first stage amplifier may become significant (7).

The nonlinearity in the temperature dependency of the forward voltage is not a serious problem for most applications, and it can be reduced by appropriate circuit design. In a Schottky diode, this nonlinearity is < 0.1 K over the

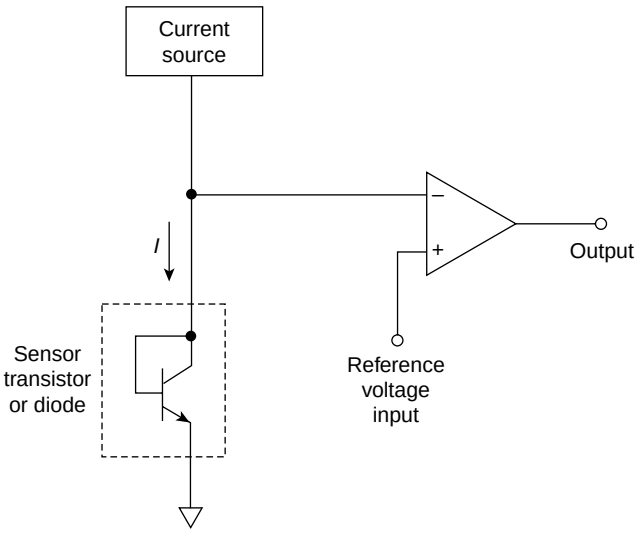


Figure 1. A fundamental interfacing circuit of a thermometer making use of a transistor or a diode as a temperature sensor to provide a voltage output proportional to temperature, with a zero voltage output at a specific temperature dependent on the reference voltage selected.

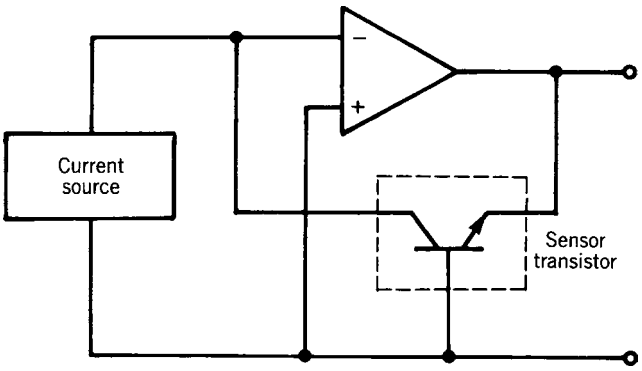


Figure 2. A circuit for constant collector current operation in a sensor transistor.

temperature range −65 to 50 °C (5), and a comparable performance is expected for diode-connected silicon transistors (7). Further improvement in the linearity can be attained by linearization of the circuit. Linearization using a logarithmic ratio module reduces the error to <0.05 °C in the temperature range −65 to 100 °C (7). Linearity is also improved using a constant collector current, as pointed out previously. An example of an actual circuit is shown in Fig. 2. In this circuit, the operational amplifier drives the base-emitter voltage to maintain a constant collector current. By applying a square-wave current and measuring the amplitude of the resulting square-wave base-emitter voltage, a linear output proportional to the absolute temperature is obtained, as expected from Eq. 7(12). Further improvement in accuracy can be attained by employing a curve fitting with three-point calibration, the error due to the nonlinearity can be reduced to 0.01 °C in the temperature range of −50 to 125 °C (14).

Three-terminal monolithic IC temperature sensors that provide a voltage output proportional to temperature using the Celsius scale are commercially available, examples being LM45 (National Semiconductor) and AD22100/22103 (Analog Devices). The LM45 device operates using a single power supply voltage in the range 4–10 V, and provides a voltage output that corresponds to the temperature in degrees Celsius multiplied by a factor of 10 mV, for example, 250 mV = 25 °C. The AD22100 and AD22103 devices provide a ratiometric output, that is, the output voltage is proportional to the temperature multiplied by the power supply voltage. For example, AD22100 has a sensitivity of 22.5 mV/ °C giving output voltages of 0.25 V at −50 °C and 4.75 V at 150 °C when the power supply voltage is 5.0 V.

Two matched transistors operated using different collector currents can be used to obtain an output proportional to the absolute temperature (15). The difference in the base-emitter voltages of the two transistors is a linear function of temperature, as shown in Eq. 7. Convenient two-terminal current-output devices using this technique are commercially available. Figure 3 shows an idealized scheme representing such devices. If the transistors Q_1 and Q_2 are assumed to be identical and have a high common-emitter current gain, their collector currents will be equal, and will constrain the collector currents Q_3 and Q_4. If Q_3 has r-fold base-emitter junctions, and each one is identical

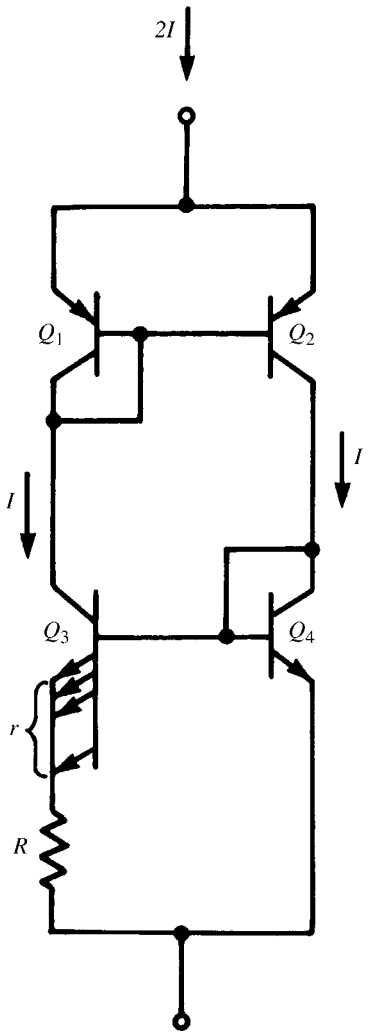


Figure 3. An idealized scheme of a two-terminal IC temperature sensor that provides a current output proportional to the absolute temperature.

to that of Q_4, the emitter current of a junction in Q_3 is $1/r$ that of Q_4. From Eq. 7, the voltage across resistance R is obtained from

$$RI = (kT/q)\ln r \tag{8}$$

Thus, the total current, $2I$, is proportional to the absolute temperature. Although the actual components are not ideal, practical devices are available as monolithic ICs, such as AD590 and AD592 (Analog Devices) (16). In these devices, $r = 8$ and R is trimmed to have a sensitivity of about 1 A/K. The output current is unchanged in the supply-voltage range 4.0 to 30 V. A voltage output proportional to the absolute temperature can be obtained by connecting a resistor in series with the ICs. For example, a sensitivity of 1 mV/K is obtained by connecting 1 kΩ resistor in series. By trimming the series resistor, the error in temperature reading can be adjusted to zero at any desired temperature. After trimming, the maximum error depends on the range in temperature under consideration. For example, a maximum error of <0.1, 0.2, and 0.3 °C is obtained for temperature ranges of 10, 25, and 50 °C, respectively (17).

Monolithic temperature sensors that provide a digital output are also commercially available. For example, TMP06 (Analog Devices) sensors provide a pulse-width modulated output. The output voltage assumes either a high or low level, so that the high period ($T1$) remains constant at 40 ms for all temperatures, while the low period ($T2$) varies with temperature. In the normal operation mode, the temperature on the Celsius *scale*, T, is given by

$$T = 406 - [731 \times (T1/T2)] \tag{9}$$

According to Analog Devices' TMP06 data sheet, for an operating supply voltage between 2.7 and 5.5 V, the absolute temperature accuracy is ±1 °C in the temperature range 0–70 °C, with a temperature resolution of 0.02 °C.

The National Semiconductor LM75 device is also a monolithic temperature sensor that provides a digital output. It includes a nine-bit analog-to-digital converter, and provides a serial output in binary format so that the least significant bit corresponds to a temperature difference of 0.5 °C.

Newer devices will come along in the future that may be more appropriate than the ones mentioned here. Information about such devices, together with their data sheets, will be available from the internet sites of manufactures.

APPLICATIONS

Although thermistors are still widely used for thermometry in the medical field, IC temperature sensors have potential advantages over thermistors. Integrated circuit sensors can be fabricated using IC technology encompassing interfacing electronics on a single IC chip, and many general purpose IC temperature sensors are now commercially available.

Current-output-type IC temperature sensors, such as AD590, are convenient for use as thermometer probes for body core and skin temperature measurements. Figure 4 shows a scheme for such a simple thermometer. According to the manufacturer's data sheet, although the sensitivity and zero offset are adjustable independently in this circuit, an accuracy of 0.1 °C is attainable with L- or M-grade AD590 devices using a single-trim calibration if the temperature span is 10 °C or less. If a regulating resistor is included in the probe, interchangeability can be realized. Because of the current output capacity, the resistance of

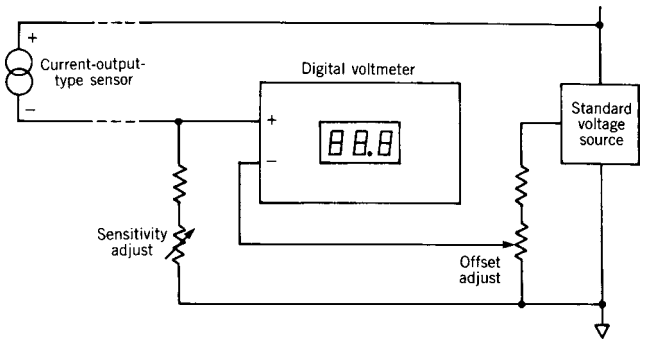


Figure 4. A simple thermometer that makes use of a two-terminal current output-type IC temperature sensor.

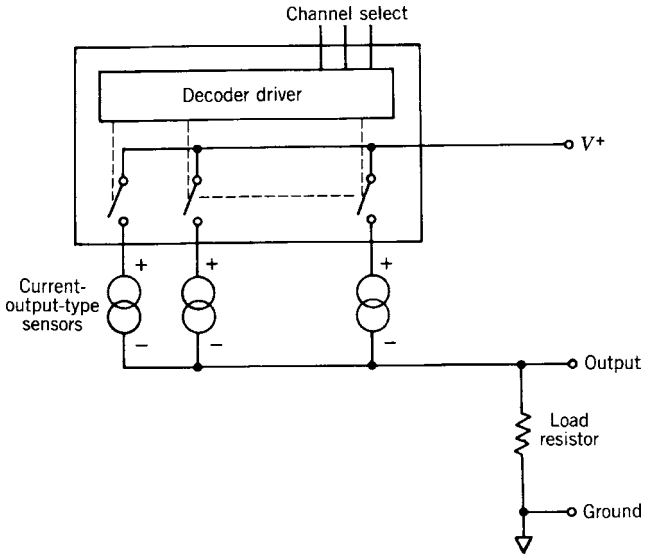


Figure 5. A multiplexing scheme for a current output-type IC temperature sensor.

the cable or connector does not affect the temperature measurement.

This type of device is also convenient for temperature measurements at many other points, especially when the output data are processed using a PC. All the sensors can be connected to a single resistor, as shown in Fig. 5, and by switching the excitation the outputs from each sensor can be multiplexed. To calibrate each sensor individually, all the sensors are maintained at an appropriate temperature, together with a standard thermometer. The outputs from each sensor as well as that from a standard thermometer are input into a PC. Then, the temperature offsets for each sensor can be stored, and all the measurement data can be corrected using these correction factors. Two-point calibration is also realized by using data at two known temperatures. A matrix arrangement of the sensors can be formed using two decoder drivers.

Temperature measurements at many different points can be performed easier using IC temperature sensors that generate serial digital outputs, such as TMP05/TMP06. Connecting these devices as shown in Fig. 6 allows for the realization of a daisy chain operation. When a start pulse is applied to the input of the first sensor, the temperature data from all the sensors is generated serially, so that the temperatures of each sensor are represented in a ratiometric form, which is the ratio of the duration of the high and low output levels for each period. It is a remarkable advantage of this sensor that a thermometer can be realized without using any analogue parts.

An important application of IC temperature sensors is the monitoring of CPU temperatures to protect a system from overheating. The temperature of a CPU chip can be detected by a p–n junction fabricated on the same silicon chip as the CPU, as shown in Fig. 7. The advantage of fabricating the temperature sensor on the CPU chip is to make the temperature measurement accurate enough and to minimize the time delay due to heat conduction so as to prevent overheating. The CPU can be protected from overheating by controlling a cooling fan or by slowing down the clock speed. Interfacing devices for this purpose are commercially available. For example, the MAX6656 (Dallas Semiconductor) device can detect temperatures at three locations, such as the CPU, the battery, and the circuit board, and the output can be used to control a cooling fan. To control the clock frequency, a specially designed frequency generator can be used. For example, the AV9155 (Integrated Circuit Systems) device allows for a gradual transition between frequencies, so that it obeys the CPU's cycle-to-cycle timing specifications.

FUTURE

It is ~25 years since convenient IC temperature sensors were introduced for scientific and industrial temperature measurements. In medicine, the application of this type of sensor is in its infancy. There are many applications where

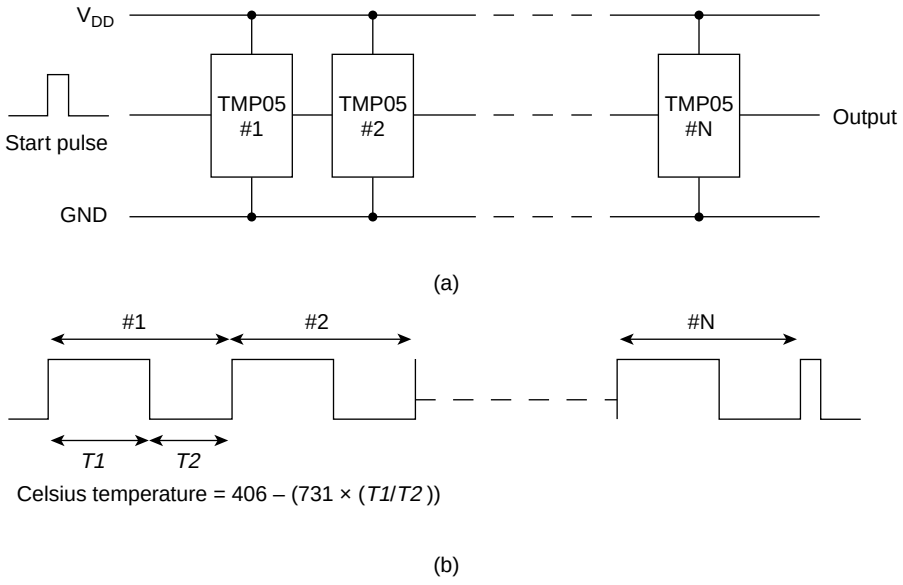


Figure 6. (a) The connecting scheme for a daisy chain operation of a serial-digital-output-type temperature sensor, and (b) the output waveform. The temperature using the Celsius scale at each sensor can be determined from the ratio of the duration of the highest and lowest points in each cycle.

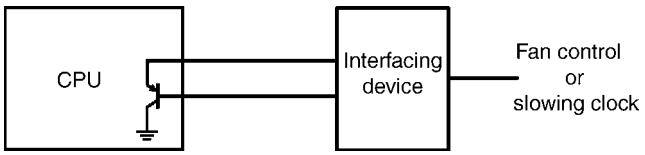


Figure 7. A scheme for monitoring the temperature of a CPU to protect it from overheating by fan control or by slowing down the clock.

these sensors can be used effectively, and undoubtedly their use will be wide spread in the near future.

Digital output IC temperature sensors show the most promise. Using such sensors, thermometers can be made without using analog components, and digital signals are convenient when a photocoupler is used for isolation.

Medical thermometry requires a relatively high degree of accuracy within a narrow temperature range. An absolute accuracy of 0.1 °C is required for body temperature measurements. However, this is hard to attain without individual calibration using most temperature sensors. While adjustment of the trimmer resistor has been used in many thermometer units, correcting data using a PC employing initially obtained correction factors will be much simpler, especially when many sensors are used, and digital output IC temperature sensors are advantageous for such a purpose.

Fabricating different types of sensors, such as force and temperature sensors, in one chip, and then applying them in robot hands to mimic all the sensing modalities of human skin, is another promising field. In such applications, the digital output capability will be a great advantage.

BIBLIOGRAPHY

1. Mackay RS. Endoradiosonde. Nature (London) 1957;179: 1239–1240.
2. Harris H. Concerning a thermometer with solid-state diodes. Sci Amer 1961;204(6):192.
3. MacNamara AG. Semiconductor diodes and transistors as electrical thermometers. Rev Sci Instrum 1963;33: 1091–1093.
4. Cohen BG, Snow WB, Tretola AR. GaAs p-n junction diodes for wide range thermometry. Rev Sci Instrum 1963; 34:1091–1093.
5. Griffiths B, Stow CD, Syms PH. An accurate diode thermometer for use in thermal gradient chambers. J Phys E 1974; 7:710–714.
6. Felimban AA, Sandiford DJ. Transistors as absolute thermometers. J Phys E 1974;7:341–342.
7. Davis CE, Coates PB. Linearization of silicon junction characteristics for temperature measurement. J Phys E 1977;10:613–619.
8. O'Neil P, Derrington C. Transistors—a hot tip for accurate temperature sensing. Electronics 1979;52(21):137–141.
9. Sah C, Noyce RN, Shockley W. Carrier generation and recombination in *p-n* junctions and p-n junction characteristics. Proc IRE 1957;45:1228–1243.
10. Sah C. Effect of surface recombination and channel on p-n junction transistor characteristics. IRE Trans Electron Devices 1962;ED9:94–108.
11. Sclar N, Pollock DB. On diode thermometers. Solid State Electron 1972;15:473–480.
12. Verster TC. p-n junction as an ultralinear calculable thermometer. Electron Lett 1968;4:175–176.
13. Ruhle RA. Solid-state temperature sensor outperforms previous transducers. Electronics 1975;48(6):127–180.
14. Ohte A, Yamagata M. A precision silicon transistor thermometer. IEEE Trans Instrum Meas 1977;IM-26:335–341.
15. Vester TC. Dual transistor as thermometer probe. Rev Sci Instrum 1969;40:174–175.
16. Timko MP. A two-terminal IC temperature transducer. IEEE J Solid-State Circuits 1976;SC-11:784–788.
17. Sheingold DH, editor. Transistor Interfacing Handbook, A Guide to Analog Signal Conditioning. Norwood, MA: Analog Devices; 1980. p 153–177.

Further Reading

Sze SM. Semiconductor Devices—Physics and Technology. New York: John Wiley & Sons; 1985.
Togawa T, Tamura T, Öberg PA. Biomedical Transducers and Instruments. Boca Raton, FL: CRC Press; 1997.
Moore BD. IC temperature sensors find the hot spot. EDN July 2/98, 1998; 99–110.
Frank R. Semiconductor junction thermometers. In: Webster JG, editor. The Measurement, Instrumentation, and Sensors Handbook. Boca Raton, FL: CRC Press; 1999. p 32/74–32/87.

See also CAPACITIVE MICROSENSORS FOR BIOMEDICAL APPLICATIONS; ION-SENSITIVE FIELD EFFECT TRANSISTORS; THERMOMETRY.

INTERFERONS. See IMMUNOTHERAPY.

INTERSTITIAL HYPERTHERMIA. See HYPERTHERMIA, INTERSTITIAL.

INTRAAORTIC BALLOON PUMP

PETER WELLER
City University
London, United Kingdom

DARREN MORROW
Royal Adelaide Hospital
Adelaide, Australia

INTRODUCTION

The heart is a pump made of cardiac muscle or myocardium. It has four pumping chambers, namely, a right and left atrium and a right and left ventricle. The atria act as primer pumps for the ventricles. The right ventricle pumps deoxygenated blood returning from the body through the pulmonary artery and into the lungs. This is called the pulmonary circulation. The left ventricle pumps oxygenated blood returning from the lungs through the aorta and into the rest of the body. This is called the systemic circulation.

The heart also has four one-way valves that prevent the backward flow of blood. The tricuspid valve lies between the right atrium and right ventricle while the pulmonary valve lies between the right ventricle and the pulmonary artery. Similarly, the mitral valve lies between the left atrium and the left ventricle while the aortic valve lies between the left ventricle and the aorta.

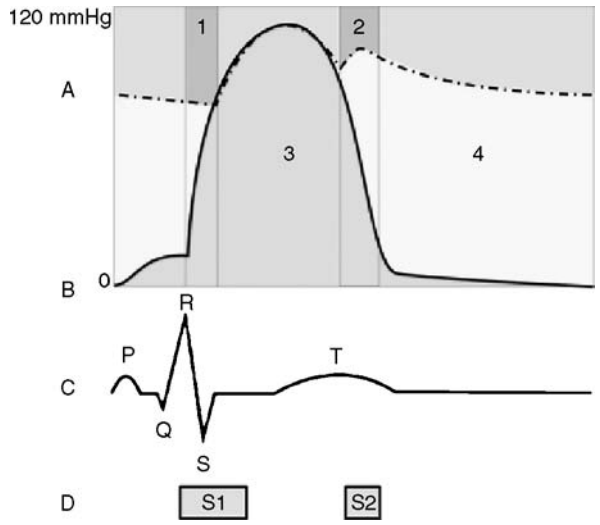


Figure 1. The relationship between the aortic pressure (A – dashed line), the ventricular pressure (B – solid line), the electrocardiogram (C) and the heart sounds (D). Region 1 = isovolumetric contraction and Region 2 = isovolumetric relaxation, Region 3 (green) = tension time index (TTI) and Region 4 (yellow) = diastolic pressure time index (DPTI). S1 represents the closing of mitral and tricuspid valves, S2 represents the closure of aortic and pulmonary valves.

CARDIAC CYCLE

The heart pumps rhythmically. The cardiac cycle is the sequence of events that take place in the heart during one heartbeat (Fig. 1). Thus, the duration of the cardiac cycle varies inversely with the heart rate. At a typical resting heart rate of 60 beats per minute (bpm), the cardiac cycle lasts 1 s or 1000 ms.

Mechanical Events

One cardiac cycle consists of a period of contraction called systole followed by a period of relaxation called diastole. The duration of systole, called the systolic time interval (STI), is relatively constant, but the duration of diastole, called the diastolic time interval (DTI), varies with the heart rate. Thus, when the heart rate increases, the DTI shortens.

When the left ventricle contracts, the pressure in the left ventricle rises above the pressure in the left atrium and the mitral valve closes. Soon afterward, the pressure in the left ventricle rises above the pressure in the aorta and the aortic valve opens. Blood flows from the left ventricle into the aorta. The period between closing of the mitral valve and opening of the aortic valve is called isovolumetric contraction.

When the left ventricle relaxes, the pressure in the left ventricle falls below the pressure in the aorta and the aortic valve closes. This causes a momentary drop in pressure in the aorta called the dichrotic notch. The period between the opening and closing of the aortic valve is called ventricular ejection. Soon afterward, the pressure in the left ventricle falls below the pressure in the left atrium and the mitral valve opens. Blood flows from the left atrium into the left ventricle. The period between the closure of the aortic valve and opening of the mitral valve is called isovolumetric relaxation.

The left atrium contracts and relaxes just before the left ventricle. This boosts the blood flow from the left atrium into the left ventricle.

These events are mirrored in the right ventricle and right atrium. However, the pressures in the pulmonary circulation are much lower than the pressures in the systemic circulation.

The movements of the chambers, valves and blood can be imaged noninvasively using ultrasound and this is called an echocardiogram.

Electrical Events

The rhythmical pumping of the heart is caused by waves of electrical impulses that spread through the myocardium from the atria to the ventricles. A recording of these waves is called an electrocardiogram (ECG). The P wave represents atrial contraction. The R wave represents ventricular contraction and signals the beginning of systole. The T wave represents ventricular relaxation and signals the beginning of diastole.

Acoustic Events

The opening and closing of the valves in the heart creates sounds that can be heard at the surface of the chest using a stethoscope. A recording of these sounds is called a phonocardiogram. The first heart sound (S1) represents closure of the mitral and tricuspid valves and signals the beginning of systole. The second heart sound (S2) represents closure of the aortic and pulmonary valves and signals the beginning of diastole.

MYOCARDIAL OXYGEN BALANCE

The systemic circulation delivers oxygenated blood to the body. Body tissues use oxygen to generate energy from the oxidation of fuels. All tissues, including the myocardium, need energy to function. The net delivery of oxygen to the myocardium is called the myocardial oxygen balance.

$$M_{OB} = M_{OS} - M_{OD}$$

where M_{OB} = myocardial oxygen balance, M_{OS} = myocardial oxygen supply, M_{OD} = myocardial oxygen demand. In the healthy heart the myocardial oxygen balance is positive, that is supply exceeds demand. In the failing heart the balance can be negative, that is demand exceeds supply.

Myocardial Oxygen Supply

The main blood supply of the myocardium comes from the two coronary arteries and their branches. The small amount of blood that reaches the myocardium transmurally from within the chambers of the heart is insignificant. The coronary arteries arise from the aorta just beyond the aortic valve and ramify within the myocardium. Myocardial oxygen supply depends on the coronary blood flow and the amount of oxygen that can be extracted from the blood.

When the heart contracts the coronary arteries are compressed and the coronary blood flow is decreased. The net driving force for coronary blood flow is called the coronary perfusion pressure.

$$C_{PP} = A_P - V_P$$

where C_{PP} = coronary perfusion pressure, A_P = aortic pressure, V_P = ventricular pressure. The coronary circulation is unique because more blood flows during diastole when the ventricular pressure is low than during systole when the ventricular pressure is high. Thus, the coronary blood flow depends on the coronary perfusion pressure, the diastolic time interval and the patency of the coronary arteries. Myocardial oxygen supply is represented by the area between the aortic pressure wave and the left ventricular pressure wave, called the diastolic pressure time index (DPTI).

Myocardial Oxygen Demand

The myocardium uses energy to perform the work of pumping. The work performed by the heart can be estimated by the mean aortic blood pressure multiplied by the cardiac output. Myocardial oxygen demand depends on the heart rate, the systolic wall tension and the cardiac contractility. Systolic wall tension is developed during isovolumetric contraction and depends upon the preload, the afterload and the wall thickness. The preload is the degree to which the left ventricle is filled before it contracts, that is the left ventricular end diastolic volume. The afterload is the pressure in the aorta or the systemic vascular resistance against which the left ventricle contracts. Myocardial oxygen demand is represented by the area under the left ventricular pressure curve, called the tension time index (TTI).

The myocardial oxygen balance is represented by the ratio DPTI:TTI.

THE PATHOPHYSIOLOGY OF LEFT VENTRICULAR PUMP FAILURE

When the left ventricle begins to fail as a pump, the cardiac output falls. Compensatory physiological mechanisms bring about an increase in left ventricular end diastolic volume, heart rate, and systemic vascular resistance. The result is an increase in preload and afterload with a decrease in coronary blood flow. Thus, the myocardial oxygen demand increases while the myocardial oxygen supply decreases. There may come a point when demand exceeds supply resulting in a negative myocardial oxygen balance. The left ventricle is then deprived of oxygen and cannot generate sufficient energy to do the work required of it. The pump failure is therefore exacerbated and this can precipitate a downward spiral of decline eventually ending in death. The therapeutic goal is to reverse this decline and help the failing left ventricle to recover by restoring a positive myocardial oxygen balance. Diuretics to decrease the preload, inotropic drugs to increase the myocardial contractility and vasodilators to decrease the preload and afterload are the mainstay of treatment. However, in the most severely ill patients, pharmacological measures alone may be insufficient and it is in these extreme circumstances that counterpulsation therapy may be effective.

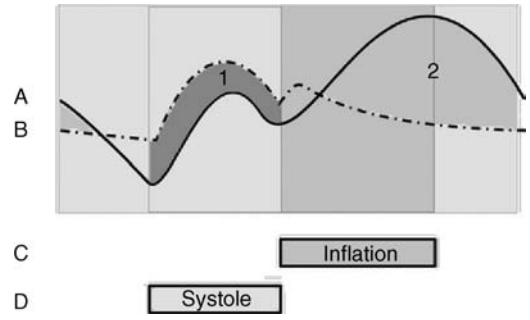


Figure 2. The relationship between the ventricular pressure with counterpulsation (A – solid line), without counterpulsation (B-dashed line), IABP balloon inflation (C) and ventricular systole (D). Region 1 (green) = systolic unloading and Region 2 (yellow) = diastolic augmentation.

THE PRINCIPLE OF COUNTERPULSATION

The principle of counterpulsation is the incorporation of an additional pump into the systemic circulation in series with the left ventricle. The pump is operated in synchrony, but out of phase, with the cardiac cycle. Pump systole occurs during ventricular diastole and pump diastole occurs during ventricular systole.

The primary physiological effects of counterpulsation are twofold (Fig. 2): A decrease in the aortic pressure during systole (called systolic unloading). This is evidenced by a decrease in the end diastolic pressure (EDP), the peak systolic pressure (PSP) and the mean systolic pressure (MSP). An increase in the aortic pressure during diastole (called diastolic augmentation). This is evidenced by an increase in the mean diastolic pressure (MDP).

Systolic unloading reduces the work of the left ventricle because it pumps against a lower pressure. This decreases myocardial oxygen demand. Diastolic augmentation increases coronary blood flow because it increases the coronary perfusion pressure. This increases myocardial oxygen supply. Thus, the myocardial oxygen balance is improved.

Among the secondary physiological effects of counterpulsation are increases in the stroke volume (SV, the volume of blood pumped with each heartbeat), the CO (equal to the SV multiplied by the heart rate) and the blood flow to the other vital organs.

HISTORICAL PERSPECTIVE

Counterpulsation was first described in theory in 1958 by Harken (1). It was to be achieved by cannulating the femoral arteries, rapidly withdrawing a set volume of blood during systole and rapidly reinfusing the same volume of blood during diastole. Clauss et al. (2) reported this in clinical practice in 1961 but it was unsuccessful because the rapid movements of blood were difficult to implement and caused severe hemolytic damage to the red blood cells.

In 1958, Kantrowitz and McKennin (3) described counterpulsation achieved by wrapping a part of the diaphragm around the thoracic aorta and stimulating the phrenic nerve, causing contraction of the diaphragm, during diastole. Moulopoulos et al. (4) and Clauss et al.(5) described counterpulsation achieved by intra-aortic balloon pumping in 1962. Operative insertion of the balloon through a surgically exposed femoral artery was necessary. It was inflated during diastole and deflated during systole. In 1968, Kantrowitz et al. (6) reported a successful clinical study.

In 1963, Dennis et al.(7) described external counterpulsation achieved using a pneumatic compression garment that enclosed the legs and lower torso. It was inflated during diastole and deflated during systole. It was reported to be as successful as the IABP in clinical studies but it is not commonly used. Kantrowitz et al. (8) described counterpulsation achieved by a permanently implantable intra-aortic balloon in 1972. It was unsuccessful in clinical practice because there remained the need for a connection to an external pump and this provided a portal of entry for infection.

In 1979, following the development of thinner catheters, percutaneous insertion of the intra-aortic balloon through a femoral artery puncture was introduced. This could be performed at the bedside and avoided the need for a surgical operation in most patients. Consequently, intra-aortic balloon pumping became the most widely adopted method of counterpulsation.

CLINICAL APPLICATIONS

Indications

The IABP was first used clinically in 1968 by Kantrowitz to support patients with cardiogenic shock after acute myocardial infarction(9). During the 1970s the indications broadened (Table 1) and by 1990 ~70,000 pump procedures were performed worldwide each year (10) although there is wide variation between different countries and centres. The IABP support has been used successfully in patients with left ventricular failure or cardiogenic shock from many causes including myodarditis, cardiomyopathy, severe cardiac contusions and drug toxicity but the commonest are myocardial infarction and following cardiac surgery. The trend has been a move away from hemodynamic support in pump failure towards the treatment, and even prophylaxis of, acute myocardial ischaemia. Patients can be maintained on the IABP for hours, days or even weeks, particularly when used as a bridge to cardiac transplantation or other definitive treatment (11). Of those who survive to hospital discharge, long-term survival is satisfactory (12).

An early series of 747 IABP procedures in 728 patients between 1968 and 1976 was reported by McEnany et al. (13). Over the course of the study, they observed that cardiogenic shock or chronic ischaemic left ventricular failure as the indication for IABP fell from 79 to 26% of patients whilst overall in-hospital survival rose from 24 to 65% of patients. They also noted an increase from 38 to 58% of patients undergoing cardiac surgery following IABP insertion. They postulated that broadened indications for, and earlier insertion of, the IABP together with more aggressive surgical treatment of any underlying cardiac lesion led to the improvement in survival. In the later Benchmark Registry of nearly 17,000 IABP procedures performed in 203 hospitals worldwide between 1996 and 2000, the main indications were support for coronary angioplasty (21%), cardiogenic shock (19%), weaning from cardiopulmonary bypass (16%), preoperative support in high risk patients (13%), and refractory unstable angina (12%) (14). The overall in-hospital mortality was 21%.

High risk patients undergoing cardiac surgery may have a better outcome if treated preoperatively with IABP therapy. In a series of 163 patients with a left ventricular ejection fraction of <0.25 and undergoing coronary artery bypass grafting (CABG), the 30 day mortality was reduced from 12 to 3% (15). Similar results were obtained in a small randomized study (16). In a series of 133 patients who underwent CABG off cardiopulmonary bypass between 2000 and 2003, the use of adjuvant preoperative IABP therapy in the 32 highest risk patients led to outcomes comparable with the lower risk patients (17). The use of IABP therapy to improve outcome after coronary angioplasty for acute myocardial infarction remains controversial. Early studies suggested an improved outcome (18–20), but two recent large randomized trials have shown no benefit in haemodynamically stable patients (21,22). A report from the SHOCK Trial Registry showed that the in-hospital mortality in patients with cardiogenic shock after acute myocardial infarction could be reduced from 77% to 47% by combined treatment with thrombolysis and IABP, particularly when followed by coronary revascularization (23).

IABP therapy is used infrequently in children, who commonly suffer from predominantly right ventricular failure associated with congenital heart disease. The greater elasticity of the aorta may limit diastolic augmentation and the more rapid heart rate may make ECG triggering difficult. Echocardiographic triggering has been used as an effective alternative (24,25). Survival rates of 57% (26) and 62% (27) have been reported in small series of carefully selected patients.

Contraindications

The only absolute contraindications to IABP therapy are severe aortic regurgitation and aortic aneurysm or dissection. In patients with severe aorto-iliac vascular disease

Table 1. Indications for IABP Therapy

Left ventricular failure or cardiogenic shock	Preoperative support before cardiac or non-cardiac surgery
Refractory unstable angina or ischaemic ventricular arrhythmias	Adjunct to coronary angioplasty or thrombolyis
Weaning from cardiopulmonary bypass	Adjunct to off-bypass cardiac surgery
Bridge to cardiac transplantation	

Table 2. Complications of IABP Therapy

Vascular	Balloon-Related	Other
Hemorrhage	Gas embolism	Infection
Aortoiliac dissection or perforation	Entrapment	Thrombocytopenia
		Paraplegia
Limb ischaemia		
Visceral ischaemia		

the balloon should not be inserted through the femoral artery.

Complications

The IABP therapy continues to cause a significant number of complications (Table 2), but serious complications are uncommon and directly attributable deaths are rare. Nevertheless, some have argued against its indiscriminate use, feeling that for many patients the risks outweigh the benefits. Kantrowitz reported rates of 41 and 4% for minor and major complications, respectively, in his series of 733 patients. Of these, 29% were vascular (including 7% hemorrhagic) and 22% were infections (28). Vascular complications include haemorrhage from the insertion site and lower limb ischaemia caused by the balloon catheter or sheath occluding the iliac or femoral artery. The vascular status of the lower limbs should be observed closely in patients on IABP therapy. Ischaemia may resolve when the catheter or sheath is removed but surgical intervention including femoral thromboembolectomy, femorofemoral bypass or even amputation is required in up to half of cases (29).

Several risk factors for vascular complications have been identified. They are female gender, diabetes, hypertension, peripheral vascular disease, obesity, old age, sheathed insertion, percutaneous insertion, and insertion via the femoral artery compared to directly into the ascending aorta (13,14,19,28–30). In one study of patients with peripheral vascular disease, the rate of vascular complications was 39% for percutaneous insertion compared to 18% for open insertion (29).

Complications caused by perforation of the balloon are rare, but potentially serious. Embolization of the helium shuttle gas can result in stroke or death. Coagulation of blood within the balloon can result in balloon entrapment. In this situation the instillation of thrombolytic agents may allow the balloon to be retrieved percutaneously but otherwise open surgery is required. It is therefore mandatory to remove the balloon immediately if any blood is detected within the pneumatic system.

Thrombocytopenia (a reduction in the number of platelets) developed in one-half of patients but they rapidly recovered when the balloon was removed (31).

EQUIPMENT FOR CLINICAL APPLICATION

The IABP consists of a balloon catheter and movable drive console. A monitor on the drive console displays the arterial pressure wave and the ECG. Commercial consoles have

Figure 3. A commercial IABP device. (Courtesy of Datascope Corp.)

controls that allow the operator to select the assist ratio and trigger mode and adjust the timing of inflation and deflation and the inflation volume of the balloon. The drive console also contains a helium tank for balloon inflation and a battery as a backup power source in the event that the mains electricity supply is interrupted. In common with medical equipment the IABP console conforms to international safety standards. Figure 3 shows a current commercial model.

The balloon is made of inelastic polyurethane and is cylindrical in shape. Balloons are available in volumes from 25 to 40 cm^3 and the correct size is selected according to the height of the patient. The balloon is mounted at the end of a double-lumen catheter. Modern catheters have an outer diameter of 7–8 French gauge. The inner lumen is open at the tip to allow insertion over a guidewire and direct measurement of the aortic blood pressure after the guidewire is removed. The outer lumen forms a closed system connecting the balloon to a pneumatic pump chamber within the drive console. Two views of a current balloon catheter are shown in Fig. 4.

The balloon catheter is most commonly inserted percutaneously through the femoral artery in the groin over a guidewire using a traditional or modified Seldinger technique. An intra-arterial sheath is used to secure and

Figure 4. An IABP catheter. (Courtesy of Datascope Corp.) In both inflated (top image) and uninflated (bottom image) modes.

protect the access site before insertion of the balloon catheter. However, because the sheath has a larger outer diameter than the balloon catheter it can increase the risk of vascular complications and sheathless insertion has now been introduced. Under fluoroscopic guidance the balloon is positioned in the descending thoracic aorta just beyond the origin of the left subclavian artery. Alternatively, the balloon can be inserted through the iliac, axillary or subclavian arteries or directly into the ascending aorta during open surgery.

A shuttle gas is pumped back and forth between the pump chamber and the balloon to cause inflation and deflation. The ideal shuttle gas would have a low density combined with a high solubility in blood. A dense gas is slow to move along the catheter. This introduces a significant delay between the opening of the valves in the pump chamber and the inflation or deflation of the balloon making correct timing more difficult to achieve. An insoluble gas is unsafe if the balloon was to leak or burst allowing it to escape into the blood. There it may form bubbles leading to potentially fatal gas embolism. Originally, carbon dioxide was used as the shuttle gas. It is a dense gas but dissolves easily in blood. More recently, helium has been used as the shuttle gas. It is a less dense gas but dissolves less easily in blood.

Pumping is initiated with an assist ratio of 1:2, which means that one in every two heartbeats is assisted. This allows the arterial pressure wave of each assisted beat to be compared with an unassisted beat to facilitate correct timing.

Pumping is continued with an assist ratio of 1:1, which means that every beat is assisted. As the patient recovers, they can be weaned from the IABP by periodically decreasing the assist ratio until pumping is eventually discontinued.

Weaning can also be achieved by periodically decreasing the inflation volume of the balloon. However, this can lead to problems with blood clotting in the folds of the underinflated balloon and it is not commonly used.

CONTROL OF IABP

The inflation/deflation cycle of the intra-aortic balloon pump is controlled by a closed loop circuit as illustrated in Fig. 5.

The physiological variables required for determining triggering are measured, filtered and converted into digital signals. These are used in enable the control strategy to determine the appropriate inflation and deflation times of the balloon. This information is then passed to the pneumatic circuit for balloon operation. The results of this strategy are then feed back to the console via a new set of patient variables. The actions of the controller are displayed on the console monitor. Additionally the control strategy can be set by the clinician.

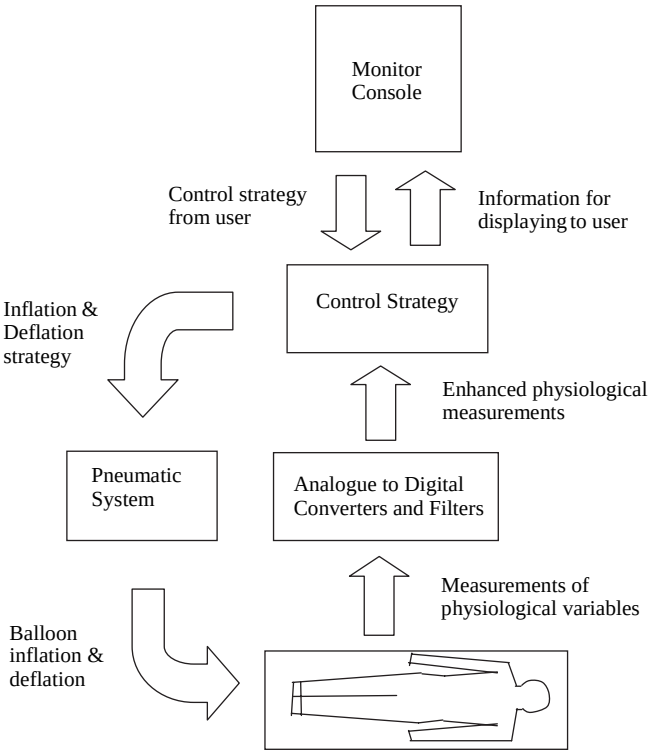


Figure 5. Closed-loop circuit for IABP control. The arrows indicate the flow of information for controlling the IABP.

TRIGGERING

The drive console requires a trigger signal to synchronise with the cardiac cycle. The trigger signal indicates the start of ventricular systole.

Commercial consoles have several modes of triggering: The ECG trigger, where the trigger signal is the R wave of the ECG. This is the most commonly used mode of triggering. Blood pressure trigger, where the trigger signal is the upstroke of the arterial pressure wave. This is used when the ECG signal is too noisy to allow reliable R wave recognition. Pacer trigger, where the trigger signal is the pacing spike of the ECG. This is used when the patient has an implanted cardiac pacemaker. Internal trigger, where the trigger signal is generated internally by the console at a set rate. This is used during cardiac surgery when the heartbeat is temporarily arrested.

TIMING

The safety and efficacy of balloon pumping is dependent upon the correct timing of balloon inflation and deflation during the cardiac cycle. Inflation should occur during the isovolumetric relaxation period of ventricular diastole. Too early inflation is unsafe because it overlaps the ejection period of ventricular systole, impedes ejection and increases the work of the heart. Too late inflation is less effective because it reduces diastolic augmentation.

Safe and effective inflation timing is fairly easy to achieve because the systolic time interval is relatively constant, regardless of heart rate or rhythm. The operator is trained to adjust the time of inflation until it visibly corresponds with the dichrotic notch on the arterial pressure wave display.

Deflation should occur at the end of ventricular diastole or during the isovolumetric contraction period of ventricular systole. Too late deflation is unsafe because it overlaps the ejection period of ventricular systole, impedes ejection and increases the work of the heart. Too early deflation is less effective because it reduces diastolic augmentation.

Safe and effective deflation timing is more difficult to achieve because the length of diastole is variable, depending on the heart rate and rhythm. The operator is trained to adjust the time of deflation until it visibly reduces the EDP and PSP on the arterial pressure wave display. Commercial consoles have two modes of timing:

Conventional Timing

Under conventional timing, the balloon is inflated and deflated at set time intervals after the R wave is detected (called manual timing). This copes badly with alterations in heart rate of >10 bpm. The problem is that when the heart rate increases, the diastolic time interval shortens and the balloon now deflates too late. Conversely, when the heart rate decreases, the diastolic time interval lengthens and the balloon deflates too early.

Conventional timing can be improved by predicting the duration of the current cardiac cycle by averaging the R–R interval of the previous 5–20 heartbeats. The balloon is then deflated at a set time interval before the next R wave is predicted to arrive (called predictive timing). This copes fairly well with alterations in heart rate and is the most commonly used mode of timing.

Both manual and predictive timing cope badly with alterations in heart rhythm, when the length of diastole can vary from beat to beat in an entirely unpredictable way. Unfortunately, cardiac arrhythmias such as atrial fibrillation and frequent ectopic beats are common in these patients making optimal timing difficult to achieve.

Real Timing

Under real timing (also called R wave deflation), the balloon is deflated when the R wave is detected and inflated a set time interval later. This copes very well with alterations in heart rate and rhythm but it tends to cause too late deflation. The problem is that when the R wave is detected there may be insufficient time to fully deflate the balloon before overlapping the ejection period.

The R wave deflation is usually used as a safety mechanism in conjunction with conventional timing. It ensures that the balloon is deflated if an R wave arrives unexpectedly due to a sudden change in heart rate or rhythm.

OPTIMIZATION

As was seen above the standard timing strategies both suffer from problems in certain scenarios and thus control of IABPs are not efficient in providing the best patient treatment. A natural solution to this is to develop timing strategies that optimize the balloon inflation regime according to the current patient condition. Several teams have reported work in this area.

Jaron et al. in 1979 (32) developed a multielement mathematical model of the canine circulation and the IABP. They expressed inflation and deflation times in terms of the total duration of inflation (DUR) and the time from the R wave to the middle of pump systole (TMPS). Duration of inflation and TMPS were expressed in terms of percentages of the duration of the cardiac cycle.

The model was validated by comparison with anesthetized dogs. They varied DUR and TMPS and measured the effects on EDP, MDP, and CO. Each of the three dependent variables (z axis) were plotted against the two independent variables (x and y axes) to create a three-dimensional (3D) urface. In general, there was considerable similarity between the surfaces obtained from the model and from the dogs. In particular, the similarity was closer for measurements of pressure (EDP, MDP) than for measurements of flow (CO).

Their results indicated that the locations of the desired optima for EDP (75% DUR, 45% TMPS) & CO (55% DUR, 75% TMPS) did not coincide. Furthermore, some combinations of DUR and TMPS within the range used clinically produced detrimental effects. Because not all variables could be optimised at same time, they suggested that the choice of timing settings involved balancing the clinical needs of the patient.

Jaron et al. in 1983 (33) subsequently developed a lumped model of the canine circulation and the IABP. It was validated by comparison with their previous model. They varied the timing of inflation and deflation, the speed of inflation and deflation and the volume of the balloon. The speed was either fast or slow, taking 7% or 33% of the duration of the cardiac cycle, respectively. The fast speed represented the ideal console with near instantaneous balloon inflation and deflation. The slow speed represented commercial consoles with significant inflation and deflation delay due to the movement of the shuttle gas. They measured the effects on EDP, SV, and coronary blood flow.

Inflation at end systole maximized SV and coronary blood flow for fast and slow speeds. Deflation at end diastole minimized EDP for fast speeds. At slow speed, deflation timing involved a trade-off between decreased EDP with early deflation and increased SV and coronary blood flow with late deflation. The overall benefit of the IABP was greater with fast speeds than slow speeds. It was also proportional to balloon volume.

In later experiments (34), they classified dependent variables as either internal, reflecting myocardial oxygen demand, or external, reflecting myocardial oxygen supply. Internal variables measured were TTI and EDP. External variable measured were SV and MDP. They showed that early deflation minimizes internal variables while late deflation maximizes external variables.

Niederer and Schilt in 1988(35) used a mechanical mock circulation and a mathematical model to investigate then influence of timing of inflation and deflation, speed of inflation and deflation and balloon volume on the efficacy

of the IABP. Timing was again expressed in terms of percentage of cardiac cycle duration. The default settings of the model were fast inflation at 30% time, fast deflation at 90% time and a volume of 40 cm^3. These were found to produce an increase in SV of 25%, a decrease in left ventricular systolic pressure of 10% and an increase in aortic diastolic pressure of 50%.

The time of inflation was varied between 20% and 50%. This had little effect on SV when the speed was fast, but inflation after 30% caused a slight decrease in SV compared to default when the speed was slow. A time of deflation before 80% caused a slight decrease in SV compared to default. Fast inflation and deflation speeds caused opening and closing shock waves that could be harmful were they to occur in humans. Balloon volume was varied between 10 and 50 mL. There was a nonlinear relationship with SV, but 40 mL was adequate for optimal performance in the mock circulation.

Barnea et al. in 1990 (36) developed a sophisticated computer simulation of the normal, failing and IABP-assisted failing canine circulation. It included simple physiological reflexes involved in the regulation of the cardiovascular system. The failing heart was simulated by reducing the contractility of the normal heart. Myocardial oxygen supply and demand were calculated from the model. They were balanced in the normal circulation and imbalanced in the failing circulation. IABP assistance of the failing circulation was shown to restore the balance.

Sakamoto et al. in 1995 (37) investigated the effect of deflation timing on the efficiency of the IABP in anaesthetized dogs. They compared a deflation time before the R wave (during late diastole) with deflation times after the R wave (during isovolumetric contraction). Deflation during the middle of isovolumetric contraction was the most effective in obtaining optimal systolic unloading.

Morrow and Weller (38) successfully used genetic algorithms and the fitness function proposed by Kane et al.(39) to evolve a fuzzy controller that optimized cardiac assistance in a computer simulation of the IABP-assisted failing heart. The inputs were MDP and PSP and the output was deflation time.

Automatic Control

Kane et al. in 1971 (39) proposed a performance index or fitness function that reflected the overall benefit of IABP assistance at different timing combinations. Inflation and deflation times were expressed in terms of the delay before inflation after the R wave and the duration of inflation. The function included weighted MDP, MSP, and EDP.

$$\text{Fitness} = k_1\text{MDP} + k_2\text{MSP} + k_4\delta(k_3 - \text{EDP})^2$$

$$\text{where,}\quad k_1 = \frac{100}{\text{MDP}_0},\quad k_2 = \frac{100}{\text{MDP}_0},\quad k_3 = \text{EDP}_0,$$

$$k_4 = -500\left(\frac{1}{k_3}\right)^2,\quad \delta = \begin{cases} 1 & \text{if}(k_3 - \text{EDP}) < 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{MDP}_0 = \text{unassisted MDP},\ \text{EDP}_0 = \text{unassisted EDP}$$

It was tested in a mechanical mock circulation and fitness was found to be a unimodal function of delay and duration. An automatic controller was developed that used a gradient descent search algorithm to improve the fitness by adjusting the delay and duration at each heartbeat. The performance of the controller was compared with the performance of R wave deflation in simulated cases of heart failure. The controller was considerably better in moderate heart failure and marginally better in severe failure. In severe failure the trade-off between systolic unloading or diastolic augmentation was particularly apparent. In moderate failure, the fitness function emphasized systolic unloading (early deflation) while in severe failure it adapted to emphasize diastolic augmentation (late deflation). Thus, in severe failure the controller tended to simulate R wave deflation.

Martin and Jaron in 1978 (40) developed a manual controller for the IABP that allowed DUR and TMPS to be adjusted. It was tested successfully on anesthetized dogs and was capable of linking to a computer for automatic control.

Jaron et al. in 1985 (34) suggested that SV and TTI index were suitable choices for a fitness function for the fine adjustment of inflation time. Both MDP and EDP were suitable choices for the adjustment of deflation time. However, later work showed that PSP correlated better with myocardial oxygen demand than EDP.

Barnea (41,42), Smith (43) and their co-workers in 1989 proposed a fitness function for optimal control of deflation time. It included weighted MDP and PSP. The permitted interval for deflation time was -200 ms to $+100$ ms relative to the predicted arrival of the next R wave. An automatic controller was developed that used a search and approximation algorithm to converge upon the optimum fitness after a number of heartbeats. It was tested successfully on computer simulations and anaesthetised dogs. It was able to follow a moving optimum, both within the same patient over time and between different patients.

Zelano et al. in 1990 (44) developed an automatic controller for the IABP that used different trigger signals. Balloon inflation occurred either upon detection of S2 or at a set time prior to the predicted time of the next S2. Balloon deflation occurred either during the P–R interval at a set time after the P wave or after the R wave. The advantage of using the P wave, R wave and S2 for triggering is that all can be detected in real time. This allows the controller to follow changes in heart rate and rhythm. It was tested successfully in a semiautomatic open loop operation in anaesthetised dogs with a coronary artery tied to simulate a myocardial infarction. They proposed a fitness function for automatic closed loop control. It used weighted MSP and MDP.

Kantrowitz et al. in 1992 (45) reported a clinical trial of automatic closed loop control of the IABP. They used a rule-based algorithm to adjust the time of inflation and deflation. Its safety was verified in anaesthetized dogs and it was then tested on 10 human patients. Their aims were for inflation to occur at dichrotic notch and for deflation to overlap the first half of ventricular ejection. They were successful in 99 and 100% of recordings respectively. Eight of the patients survived. Neither of the two deaths was attributed to the controller.

Sakamoto et al. in 1995 (46) developed a new algorithm to cope with atrial fibrillation, the most unpredictable cardiac arrhythmia sometimes described as irregularly irregular. The aim was for inflation to occur at the dichrotic notch. They were able to predict the time of arrival of the dichrotic notch from mathematical analysis of the R–R interval from the previous 60 heartbeats. Deflation occurred at the R wave. It was tested on ECG recordings from real patients and performed better than conventional timing.

FUTURE DEVELOPMENTS

The use of IABP therapy for preoperative support and as an adjunct to coronary angioplasty or bypass and off-bypass cardiac surgery is likely to increase although larger studies are required to identify which patients will benefit most. The role of the IABP in left ventricular failure is likely to decrease with the increasing use of left ventricular assist devices.

Manufacturers recent research and development has focused on: Improved automatic control algorithms that better cope with alterations in heart rate and rhythm and adjust inflation and deflation times to optimize cardiac assistance. Better catheter designs that cause fewer vascular complications and permit more rapid movement of the shuttle gas.

BIBLIOGRAPHY

1. Harken DE. Presentation at the International College of Cardiology, Brussels; 1958.
2. Clauss RH, et al. Assisted circulation. I. The arterial counterpulsator. J Thorac Cardiovasc Surg 1961;41:447.
3. Kantrowitz A, McKinnen WMP. Experimental use of diaphragm as experimental myocardium. Surg Forum 1958;9:266.
4. Moulopoulos SD, Topaz S, Kolff WJ. Diastolic balloon pumping (with carbon dioxide) in the aorta. A mechanical assistance to the failing circulation. Am Heart J 1962;63:669.
5. Clauss RH, Missier P, Reed GE, Tice D. Assisted circulation by counterpulsation with intra-aortic balloon: Methods and effects. Proceeding of the 4th ACEMB; 1962.
6. Kantrowitz A, et al. Initial clinical experience with intraaortic balloon pumping in cardiogenic shock. J Am Med Ass 1968;203:113.
7. Dennis C, Moreno JR, Hall DP. Studies on external counterpulsation as a potential measure for acute left heart failure. Trans Am Soc Artif Intern Organs 1963;9:186.
8. Kantrowitz A, et al. Initial clinical experience with a new permanent mechanical auxiliary ventricle: The dynamic aortic patch. Trans Am Soc Artif Intern Organs 1972;18(0):159–167, 179.
9. Kantrowitz A, et al. Jr. Initial clinical experience with intraaortic balloon pumping in cardiogenic shock. JAMA 1968;203(2):113–118.
10. Kantrowitz A. Origins of intraaortic balloon pumping. Ann Thoracic Surg 1990;50(4):672–674.
11. Freed PS, Wasfie T, Zado B, Kantrowitz A. Intraaortic balloon pumping for prolonged circulatory support. Am J Cardiol 1988;61(8):554–557.
12. Lund O, et al. Intraaortic balloon pumping in the treatment of low cardiac output following open heart surgery–immediate results and long-term prognosis. Thoracic Cardiovascular Surg 1988;36(6):332–337.
13. Meharwal ZS, Trehan N.Vascular complications of intra-aortic balloon insertion in patients undergoing coronary reavscularization: Analysis of 911 cases.[See comment]. Eur J Cardio-Thoracic Surg 2002;21(4):741–747.
14. Ferguson JJ. 3rd, et al. The current practice of intra-aortic balloon counterpulsation: Results from the benchmark registry. J Am College Cardiol 2001;38(5):1456–1462.
15. Dietl CA, et al. Efficacy and cost-effectiveness of preoperative iabp in patients with ejection fraction of 0.25 Or less. Ann Thorac Sur 1996;62(2):401–408.
16. Christenson JT, Simonet F, Badel P, Schmuziger M. Evaluation of preoperative intra-aortic balloon pump support in high risk coronary patients. Eur J Cardio-Thoracic Sur 1997;11(6): 1097–1103.
17. Suzuki T, et al. Usefulness of preoperative intraaortic balloon pump therapy during off-pump coronary artery bypass grafting in high-risk patients. Ann Thoracic Surg 2004;77 (6):2056–2059.
18. Brodie BR, Stuckey TD, Hansen C, Muncy D. Intra-aortic balloon counterpulsation before primary percutaneous transluminal coronary angioplasty reduces catheterization laboratory events in high-risk patients with acute myocardial infarction. Am J Cardiol 1999;84(1):18–23.
19. Ishihara M, Sato H, Tateishi H, Uchida T, Dote K. Intraaortic balloon pumping as the postangioplasty strategy in acute myocardial infarction. Am Heart J 1991;122(2):385–389.
20. Ohman EM, et al. Use of aortic counterpulsation to improve sustained coronary artery patency during acute myocardial infarction. Results of a randomized trial. The randomized iabp study group Circulation 1994;90(2):792–799.
21. Stone GW, et al. A prospective, randomized evaluation of prophylactic intraaortic balloon counterpulsation in high risk patients with acute myocardial infarction treated with primary angioplasty. Second primary angioplasty in myocardial infarction (pami-ii) trial investigators. J Am College Cardiol 1997;29(7):1459–1467.
22. van't Hof AW, et al. A randomized comparison of intra-aortic balloon pumping after primary coronary angioplasty in high risk patients with acute myocardial infarction. [See comment]. Eur Heart J 1999;20(9):659–665.
23. Sanborn TA, et al. Impact of thrombolysis, intra-aortic balloon pump counterpulsation, and their combination in cardiogenic shock complicating acute myocardial infarction: A report from the shock trial registry. Should we emergently revascularize occluded coronaries for cardiogenic shock? J Am College Cardiol 2000;36(3 Suppl. A):1123–1129.
24. Minich LL, et al. Intra-aortic balloon pumping in children with dilated cardiomyopathy as a bridge to transplantation. J Heart Lung Transplant 2001;20(7):750–754.
25. Pantalos GM, et al. Estimation of timing errors for the intraaortic balloon pump use in pediatric patients. ASAIO Journal 1999;45(3):166–671.
26. Akomea-Agyin C, et al. Intraaortic balloon pumping in children. Ann Thoracic Surg 1999;67(5):1415–1420.
27. Pinkney KA, et al. Current results with intraaortic balloon pumping in infants and children. Ann Thoracic Surg 2002;73(3):887–891.
28. Kantrowitz A, et al. Intraaortic balloon pumping 1967 through 1982: analysis of complications in 733 patients. Am J Cardiol 1986;57(11):976–983.
29. Miller JS, Dodson TF, Salam AA, Smith RB3rd. Vascular complications following intra-aortic balloon pump insertion. Am Surg 1992;58(4):232–238.
30. Macoviak J, et al. The intraaortic balloon pump: an analysis of five years' experience. Ann Thoracic Surg 1980;29(5):451–458.

31. Vonderheide RH, Thadhani R, Kuter DJ. Association of thrombocytopenia with the use of intra-aortic balloon pumps. Am J Med 1998;105(1):27–32.
32. Jaron D, Ohley W, Kuklinski W. Efficacy of counterpulsation: model and experiment. Trans Am Soc Artificial Inter Organs 1979;25:372–377.
33. Jaron D, Moore TW, He P. Theoretical considerations regarding the optimization of cardiac assistance by intraaortic balloon pumping. IEEE Trans Biome Eng 1983;30(3):177–185.
34. Jaron D, Moore TW, He P. Control of intraaortic balloon pumping: theory and guidelines for clinical applications. Ann Biomed Eng 1985;13(2):155–175.
35. Niederer P, Schilt W. Experimental and theoretical modelling of intra-aortic balloon pump operation. Med Biol Eng Comput 1988;26(2):167–174.
36. Barnea O, Smith B, Moore TW, Jaron D. Simulation and optimization of intra-aortic balloon pumping. Proceedings. Computers in Cardiology (Cat. No.89CH2932-2). Los Alamitos (CA): IEEE Computer Society Press; 1990. p 237–240.
37. Sakamoto T, et al. Effects of timing on ventriculoarterial coupling and mechanical efficiency during intraaortic balloon pumping. ASAIO Journal 1995;41(3):M580–MM583.
38. Weller PR, Morrow DR, LeFèvre JE. Evolution of a fuzzy controller for the intra-aortic balloon pump. Proceedings of 2nd European Medical and Biological Engineering Conference, EMBEC'02, Vienna, Austria, Hutten H, Krosl P. editors. ISBN 3-901351-62-0;4–8. December 2002; p 1588–1589.
39. Kane GR, Clark JW, Bourland HM, Hartley CJ. Automatic control of intra-aortic balloon pumping. Trans Am Soc Artif Int Organs 1971;17:148.
40. Martin PJ, Jaron D. New controller for in-series cardiac-assist devices. Med Biol Eng Computing 1978;16(3):243–249.
41. Barnea O, Smith B, Moore TW, Jaron D. An optimal control algorithm for intra-aortic balloon pumping. Images of the Twenty-First Century. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No.89CH2770-6). New York: Vol. 5. IEEE; 1989. p 1419–1420.
42. Barnea O, et al. Optimal controller for intraaortic balloon pumping. IEEE Trans. Biomed Eng 1992;39(6):629–634.
43. Smith B, Barnea O, Moore TW, Jaron D. An algorithm for optimal control of the intra-aortic balloon pump. Proceedings of the Fifteenth Annual Northeast Bioengineering Conference (Cat. No.89-CH2689-8). New York; IEEE. 1989; p 75–76.
44. Zelano JA, Li JK, Welkowitz W. A closed-loop control scheme for intraaortic balloon pumping. IEEE Trans Biomed Eng 1990;37(2):182–192.
45. Kantrowitz A, et al. Initial clinical trial of a closed loop, fully automatic intra-aortic balloon pump. ASAIO Journal 1992; 38(3):M617–M621.
46. Sakamoto T, Arai H, Maruyama T, Suzuki A. New algorithm of intra aortic balloon pumping in patients with atrial fibrillation. ASAIO Journal 1995;41(1):79–83.

Further Reading

Goldberger M, Tabak SW, Shah PK. Clinical experience with intra-aortic balloon counterpulsation in 112 consecutive patients. *Am Heart J* 1986;111(3):497–502.

McEnany MT, et al. Clinical experience with intraaortic balloon pump support in 728 patients. *Circulation* 1978;58(3 Pt 2):I124–I132.

See also ARTERIES, ELASTIC PROPERTIES OF; HEMODYNAMICS; VASCULAR GRAFT PROSTHESIS.

INTRACRANIAL PRESSURE MONITORING. See MONITORING, INTRACRANIAL PRESSURE.

INTRAOCULAR LENSES. See LENSES, INTRAOCULAR.

INTRAOPERATIVE RADIOTHERAPY. See RADIOTHERAPY, INTERAOPERATIVE.

INTRAUTERINE DEVICES (IUDS). See CONTRACEPTIVE DEVICES.

INTRAUTERINE SURGICAL TECHNIQUES

JOSEPH P. BRUNER
Department of Obstetrics and Gynecology
Nashville, Tennessee

R. DOUGLAS WILSON
N. SCOTT ADZICK
University of Pennsylvania
Philadelphia, Pennsylvania

LOUISE WILKINS-HAUG
AUDREY C. MARSHALL
Women's and Children's Hospital
Boston, Massachusetts

RUBEN A. QUINTERO
Florida Institute for Fetal Diagnosis and Therapy
Tampa, Florida

M. YAMAMOTO
Y. VILLE
University of Paris
Paris, France

ANTHONY JOHNSON
University of North Carolina
Chapel Hill, North Carolina

JULIE S. MOLDENHAUER
Wayne State University
Detroit, Michigan

INTRODUCTION

Intrauterine surgery of the fetus or fetal adnexae is spreading rapidly throughout the world. In a broad sense, intrauterine surgery includes any procedure in which a medical device is purposely placed within the uterine cavity. For most physicians, however, the concept of intrauterine surgery excludes such commonly performed procedures as amniocentesis and chorionic villous sampling. Rather, the term usually refers to techniques requiring specialized knowledge, experience, and most especially, instrumentation. Procedures most commonly mentioned in discourses on intrauterine surgery include those specifically developed for the treatment of such serious anatomic defects as congenital cystic adenomatoid malformation (CCAM), sacrococcygeal teratoma (SCT), lower urinary tract obstruction (LUTO), myelomeningocele, aortic or pulmonic stenosis, gastroschisis, iatrogenic amniorhexis, twin-to-twin transfusion syndrome (TTTS), and twins discordant for severe anomalies. As no one person in the world is an expert in every one of these procedures, the remainder of

this chapter is divided into individual sections, each authored by someone widely recognized as a leader in the treatment of that particular anomaly.

CONGENITAL CYSTIC ADENOMATOID MALFORMATION AND SACROCOCCYGEAL TERATOMA

Most prenatally diagnosed malformations are managed by appropriate medical and surgical evaluation and treatment following planned delivery near term, with some cases requiring transfer of the mother to a tertiary referral center with obstetrics, maternal fetal medicine, medical genetics, neonatology, and pediatric surgery subspecialties. Certain anatomic abnormalities can have significant fetal developmental consequences, with emergency *in utero* therapy being required due to gestational age and mortality risks for the fetus. Open maternal fetal surgery poses additional risks to the mother. Maternal fetal surgery (1–4) should not be attempted until (1) the natural history of the fetal disease is established by following up on untreated cases, (2) selection criteria for cases requiring intervention are developed, (3) pathophysiology of the fetal disorder and its correction are defined in fetal animal models, and (4) hysterotomy and fetal surgery can be performed without undue risk to the mother and her reproductive potential.

Congenital cystic adenomatoid malformation of the lung (CCAM) (5) is a rare lesion characterized by a multicystic mass of pulmonary tissue with proliferation of bronchial structures. CCAM is slightly more common in males and is unilobar in 80–95% of cases. CCAMs derive their arterial blood supply from the normal pulmonary circulation. CCAMs are divided clinically into cystic and solid lesions, but have been divided traditionally into three types based on their pathological characteristics. Type 1 CCAM lesions account for 50% of postnatal CCAM cases and consist of single or multiple cysts lined by ciliated pseudostratified epithelium (5). These cysts are usually 3–10 cm in size and 1–4 in number (5). Type II CCAM lesions account for 40% of postnatal cases of CCAM and consist of more numerous cysts of smaller diameter, usually less than 1 cm. They are lined by ciliated, cuboidal, or columnar epithelium (5). Type III CCAM lesions account for only 10% of CCAM cases and are usually large, homogenous, microcystic masses that cause mediastinal shift. These lesions have bronchiolar-like structures lined by ciliated cuboidal epithelium separated by masses of alveolar-sized structures lined by nonciliated cuboidal epithelium (5). Prognosis in Type III CCAM is related to size.

Sacrococcygeal teratoma (SCT) (6) is defined as a neoplasm composed of tissues from either all three germ layers or multiple foreign tissues lacking an organ specificity. SCT is thought to develop from a totipotent somatic cell originating in Hensen's node. SCT has been classified by the relative amounts of presacral and external tumor present, with Type I completely external with no presacral component, Type II external component with internal pelvic component, Type III external component with internal component extending into the abdomen, and Type IV completely internal with no external component (7). A Type I SCT is evident at birth and is usually easily resected and has a low malignant potential (6). Type II and III SCTs are recognized at birth, but resection may be difficult requiring an anterior and posterior approach (6). A Type IV SCT can have a delayed diagnosis until symptomatic at a later age (6). SCT is one of the most common tumors in newborns and has an incidence of 1 per 35,000 to 40,000 live births (6).

Evaluation of the fetal status for CCAM and SCT requires multiple imaging and functional techniques (8–10), including fetal ultrasound, fetal MRI, and fetal echocardiogram with arterial and venous Doppler assessments (umbilical artery, umbilical vein, ductus venosus). Measurements include combined cardiac output, cardiothoracic ratio, descending aortic blood flow, inferior vena cava diameter, placental thickness, umbilical artery systolic to diastolic Doppler ratio, and amniotic fluid index. Presence of ascites, pleural or pericardial effusion, and skin or scalp edema are important markers for the extent of fetal hydrops and its overall effect on fetal stability. Specific ultrasound imaging of the CCAM and SCT looks for the percentage of cystic and solid components in the tumors as well as an overall mass volume (cc) estimate (AP cm $\times$ transverse cm $\times$ height cm $\times$ 0.52). The SCT consistency and size can be reflected directly in the combined cardiac output and amount of vascular shunting. The CCAM overall size can cause mediastinal shift with cardiac dysfunction and pulmonary deformation. Validated ratio of CCAM/head circumference (CVR) can be used for prognosis and follow-up planning (10). The specific lobar location for the CCAM may have a differential impact on cardiac function. The development of fetal hydrops is due mainly to cardiac dysfunction secondary to compression.

The physiologic changes required in the fetal status to move from expectant management to open maternal fetal surgery is generally dictated by fetal (gestational age and extent of fetal hydrops) and maternal factors (8–10). Criteria for consideration of maternal fetal surgery for CCAM resection (fetal lobectomy) require the absence of maternal risk factors for anesthesia and surgery, a singleton pregnancy with a normal karyotype (amniocentesis, chorionic villus sampling, or percutaneous umbilical blood sampling), no other anatomical abnormalities beyond the associated hydrops, gestational age of 21–31 weeks, and massive multicystic or predominantly solid CCAM (CVR $>$ 1.6) (8–10). In selected cases, the failure of *in utero* therapy techniques, such as thoracoamniotic shunting or cyst aspiration for the large Type I lesions, to reverse the fetal hydrops would be required. Criteria for consideration of maternal fetal surgery for debulking of a SCT require the absence of maternal risk factors for anesthesia and surgery, a singleton pregnancy with a normal karyotype, the absence of significant associated anomalies, evidence of impending high output cardiac failure, gestational age of 21–30 weeks, and favorable SCT anatomy classification (Type I or II) (9).

The technique for maternal hysterotomy to allow access to the fetus has been well described and has evolved over 25 years of experimental and clinical work (1,8,9). The uterus is exposed through a maternal low transverse abdominal incision. If a posterior placenta is present, superior and inferior subcutaneous flaps are raised and a vertical midline fascial incision is made to expose the uterus for a convenient anterior hysterotomy with the uterus remaining

in the abdomen. Conversely, the presence of an anterior placenta necessitates the division of the rectus muscles so the uterus can be tilted out of the abdomen for a posterior hysterotomy. A large abdominal ring retractor (Turner–Warwick) is used to maintain exposure and prevent lateral compression of the uterine vessels. Sterile interoperative ultrasound is used to delineate the fetal position and placental location. The edge of the placenta is marked under sonographic guidance using electrocautery or a marking pen. The position and orientation of the hysterotomy is planned to stay parallel to and at least 6 cm from the placental edge but still allow exposure of the appropriate fetal anatomy. The hysterotomy is facilitated by the placement of two large monofilament sutures (PDS II 1 Ethicon; Somerville, NJ) parallel to the intended incision site and through the full thickness of the uterine wall and membranes under sonographic guidance. The electrocautery is used to incise the myometrium between the two stay sutures down to the level of the amniotic membranes. A uterine stapler device (US Surgical Corporation; Norwalk, CT) with absorbable Lactomer staples is then directly introduced through the point of fixation and into the amniotic cavity by using a piercing attachment on the lower limb of the stapler. The stapler is fired, thereby anchoring the amniotic membranes (chorion, amnion) to the uterine wall creating a hemostatic hysterotomy. Careful evaluation for the membrane adhesion status and for any myometrial bleeding sites is undertaken. If required, interrupted PDS sutures are used to control bleeding and membrane separation. The fetus and the internal uterine cavity are continually bathed in warmed lactated Ringers at 38–40°C using a level I warming pump connected to a red rubber catheter that is placed in the uterine cavity through the hysterotomy.

For CCAM resection (1,8,11), once the appropriate fetal area is visualized in the hysterotomy site, the fetal arm is brought out for pulse oximeter monitoring, IV access, and fetal position control. Intraoperative fetal echocardiography is used throughout to monitor cardiac function. The fetal chest is entered by a fifth intercostal space thoracotomy. The lesion usually decompresses out through the thoracotomy wound consistent with the increase in the thoracic pressure from the mass (8). Using techniques initially developed on experimental animals, the appropriate pulmonary lobes containing the lesion are resected (1,11). Fetal resuscitation is performed if needed through intravenous administration of crystalloid, blood, and code-blue medications with fetal echocardiography providing functional information. The fetal thoracotomy is closed and the fetal arm is returned to the uterus.

The technique for debulking of an external fetal SCT has been described in detail previously (1,9,12,13). The fetal foot is used for pulse oximeter monitoring and IV access with intraoperative echocardiography. The fetal SCT is exposed and a Hagar dilator is placed in the rectum. Fetal skin is incised circumferentially around the base of the tumor and a tourniquet is applied to constrict blood flow. The tumor is debulked externally, usually with a 90 mm thick tissue stapler (US Surgical Corporation; Norwalk, CT). The objective of the fetal SCT resection is to occlude the tumor vascular supply and remove the low resistance tumor vascular bed from the fetal circulation. No attempt is made to dissect the intrapelvic component of the tumor or to remove the coccyx (done with a second procedure after birth). Fetal resuscitation is performed if needed through intravenous administration of crystalloid, blood, and code-blue medications with fetal echocardiography providing functional information. The fetal sacral wound is closed.

Repair of the hysterotomy after fetal surgery (1–4) uses a water-tight two-layered uterine closure, with interrupted full thickness stay sutures placed first and untied using PDS II 1 (Ethicon; Somerville, NJ), and the uterus is then closed with a running continuous stitch PDSII 0 (Ethicon; Somerville, NJ) including the chorion-amnion membrane layer. The interrupted stay sutures are then tied after the amniotic fluid volume has been corrected with warm lactated Ringers through a red rubber catheter and volume confirmed by ultrasound visualization. The omentum is sutured in place over the hysterotomy closure to help seal the hysterotomy site with vascularized tissue and to prevent bowel adherence to the site, especially when a posterior hysterotomy is performed. The maternal laparotomy incision is closed in layers. It is important to use a subcuticular skin closure covered with a transparent dressing so that monitoring devices can be placed on the maternal abdomen postoperatively.

In some specific cases, when the CCAM lesion is not resected *in utero*, it continues to be a large space-occupying lesion with mediastinal shift. Thus, it might be anticipated that respiratory compromise will be present at birth, the delivery may be facilitated with an EXIT procedure (*ex utero* intrapartum therapy) (14). Uterine relaxation is maintained by high concentration inhalational anesthetics, with additional tocolysis if necessary. The EXIT requires only the head and chest to be initially delivered through, preferably, a low transverse hysterotomy wound thereby preserving uterine volume with the lower fetal body and continuous warmed lactated Ringers infusion to prevent cord compression. These maneuvers preserve the uterine-placental circulation and continue placental gas exchange. The EXIT procedure can be done through an anterior or posterior hysterotomy, but its location in the uterus may require that all future pregnancies be delivered by cesarean section with no trial of labor if a low anterior transverse location is not available.

All future pregnancies following maternal hysterotomy for maternal-fetal surgery require cesarean section at term with no trial of labor. Maternal obstetrical risks in a subsequent pregnancy are similar to risks following for a classic cesarean section (15).

LOWER URINARY TRACT OBSTRUCTION

The diagnosis and treatment of fetal lower urinary tract obstruction (LUTO) requires knowledge of the differential diagnosis and the natural history of the condition, a thorough understanding of the criteria for therapy, and management expertise. Fetal LUTO is one of the most commonly diagnosed birth defects. Untreated, and depending on the level of the obstruction, it may lead to hydronephrosis, renal dysplasia, pulmonary hypoplasia, and

perinatal death (16,17). The prognosis depends on the extent of preexisting renal damage and the effectiveness of therapy. Treatment with fetal urinary diversion procedures is aimed at preventing renal damage and pulmonary hypoplasia (18–20).

Obstruction to urine flow has been shown in animal models to result in hydronephrosis and renal dysplasia (21). Release of the obstruction is associated with no or variable renal damage depending on the timing of the release or the creation of the defect (21,22). Pulmonary hypoplasia is another major potential complication of fetuses with obstructive uropathy (23). The association probably results from the attendant oligohydramnios.

Urethral obstruction may result from posterior urethral valves (PUV), anterior urethral valves, megalourethra, urethral duplications, urethral atresia, obstructive ureterocele, or cloacal dysgenesis. Posterior urethral valves (PUV), first described by Young et al. (24), constitute the most common cause of lower urinary tract obstruction in male neonates, with an incidence of 1:8000 to 1:25,000 livebirths (25). The lesions occur only in males because the female counterpart of the verumontanum, from which the valves originate, is the hymen.

In utero therapy is usually limited to fetuses with bladder outlet obstruction. Fetuses with unilateral obstruction are not typically considered candidates for *in utero* therapy, regardless of the magnitude of the obstruction or renal findings. In these patients, the risk/benefit ratio of *in utero* intervention favors expectant management, even if it means loss of the affected renal unit.

Fetal renal function may be assessed by analysis of fetal urinary parameters via vesicocentesis. Patients are considered candidates for *in utero* therapy if fetal urinary parameters are below the threshold for renal cystic dysplasia. If the values are above the threshold, therapy should not be offered.

The application of selection criteria in patients with fetal LUTO for possible *in utero* therapy results in a significant attrition rate. Disqualification from therapy may result both from "too healthy" or "too sick" conditions. Examples of too healthy conditions include normal amniotic fluid volume or suggestion of nonobstructive dilatation of the urinary tract. Examples of too sick conditions include sonographic evidence of renal cystic dysplasia, abnormal fetal urinary parameters, abnormal karyotype, or the presence of associated major congenital anomalies. Of 90 patients referred to the Florida Institute for Fetal Diagnosis and Therapy from October 1996 to October 2003, more than one-half were disqualified from therapy from single or overlapping conditions.

Percutaneous ultrasound-guided vesicoamniotic shunting of fetuses with LUTO began in the early 1980s (16,19,23). The goal of therapy is to avoid development of pulmonary hypoplasia from the attendant oligohydramnios as well as to preserve renal function. Fetal bladder shunting should be offered only to patients without sonographic or biochemical evidence consistent with renal cystic dysplasia, normal karyotype, and lack of associated major congenital anomalies.

The procedure can be performed under local, regional, or general anesthesia. A minimal skin incision is made. Ultrasound is used to identify the ideal site of entry into the fetal bladder, below the level of the umbilicus. Color Doppler ultrasonography is used to identify the umbilical vessels around the distended bladder and avoid them. Under ultrasound guidance, the trocar is directed through the maternal tissues and up to the fetal skin. Fetal analgesia is achieved with pancuronium 0.2 mg/kg and fentanyl 10 mcg/kg. The trocar stylet is used to enter the fetal bladder with a sharp, swift, and controlled maneuver. If a prior vesicocentesis had been performed, it is advisable to obtain a sample of fetal urine for microbiological purposes to rule out preexisting infection. A sample of fetal urine is sent for further biochemical testing. Placement of the double-pigtail catheter is monitored with ultrasound. After the distal loop is deployed in the bladder, the trocar is retrieved to the level of the bladder wall. A small amount of the straight portion of the catheter may be advanced into the bladder to avoid retracting the distal loop into the bladder wall. The trocar shaft is retrieved slowly while simultaneously maintaining pressure on the catheter to deploy the straight portion within the bladder wall and fetal skin. Once the shaft of the trocar reaches the fetal skin, entrance of the catheter, including the proximal loop, can be safely deployed. If complete anhydramnios is present prior to insertion of the catheter, it is advantageous to attempt an amnioinfusion with an 18 gauge needle prior to shunting to create the space for deployment of the proximal loop. Amnioinfusion is aimed at preventing misplacement of the proximal loop within the myometrium and fetal membranes.

Despite adequate placement, malfunction of vesicoamniotic shunting may occur up to 60% of the time (26). The shunt may pull from the skin into the fetal abdomen, resulting in iatrogenic ascites, or out of the fetal bladder, with no further drainage of urine. The shunt may pull out of the fetus altogether as well. Replacement of the shunt is associated with an additive risk of fetal demise, chorioamnionitis, premature rupture of membranes, and miscarriage or preterm delivery, for a total perinatal loss rate of approximately 5% per instance.

In 1995, we proposed the use of endoscopy to assess the fetal bladder for diagnostic and surgical purposes (27,28). Endoscopic visualization of the fetal bladder with a larger endoscope can be justified during vesicoamniotic shunting. Currently, we use a 3 mm or a 3.9 mm trocar with a 2.7 mm or 3.3 mm diagnostic or operating endoscope. This diameter is slightly larger than the 14 gauge (approximately 2.1 mm) needle used for the insertion of the double-pigtail catheter. Access to the fetal bladder allows remarkable evaluation of the bladder, ureteral orifices, and urethra as well as the opportunity to perform surgical procedures.

In normal fetuses, the urethra is not dilated, appearing as a small hole within the bladder. In patients with a true urethral obstruction, endoscopy will show a variable dilatation of the urethra at the level of the bladder neck. The urethra is located using a 25° or a 70° diagnostic rigid endoscope. Alternatively, a flexible/steerable endoscope may be used. The anatomical landmarks to identify at this level include the verumontanum and the urethral valves. The diagnostic endoscope is then exchanged for a rigid

operating endoscope. A 600 μm YAG-laser fiber is passed through the operating channel of the endoscope, and then ablated using 5–10 w and 0.2 s pulses in successive steps. The fiber is placed as anterior and medial as possible. It is not necessary to evaporate the entire valvular tissue. Instead, only a few defects to either side of the midline are necessary to establish urethral patency (27,29). The dilated urethra may collapse intraoperatively once patency is re-established, which may obscure the field of view and require frequent instillation of saline to the side port of the trocar to distend it. Color Doppler may also be used to document fetal urination through the penis.

A urethrorectal fistula may occur from thermal damage beyond the posterior wall of the urethra into the perirectal space. To avoid this complication, only 5–10 w of energy in short bursts should be used while ablating the valves.

The management of fetuses with lower obstructive uropathy continues to be one of the most challenging subjects in fetal therapy. The difficulties include establishing the correct differential diagnosis, accurately predicting subsequent renal function, and providing the best treatment.

MYELOMENINGOCELE

Myelomeningocele results from the failure of caudal neural tube closure during the fourth week of gestation. The lesion is characterized by protrusion of the meninges through a midline bony defect of the spine, forming a sac containing cerebrospinal fluid and dysplastic neural tissue. Affected infants exhibit varying degrees of somatosensory loss, neurogenic sphincter dysfunction, paresis, and skeletal deformities (30). Virtually all such infants also have the Chiari II malformation, and up to 95% develop hydrocephalus (31). Although myelomeningocele is not a lethal disorder, the neurologic sequelae are progressive, and worsen until the lesion is closed. Observational and cohort studies have demonstrated improvement of the Chiari II malformation (32,33), decreased hydrocephalus (34), and improved lower extremity function after intrauterine repair of myelomeningocele (35).

On the day of surgery, the pregnant patient is taken to a standard obstetrical operating room. An epidural catheter is placed and, after induction of general endotracheal anesthesia, she is prepared as if for a cesarean section. Many of the general anesthetic agents cross the placenta and provide analgesia for the fetus, and the epidural catheter enables the administration of continuous postoperative analgesics if needed. The gravid uterus is exposed with a Pfannenstiel incision and exteriorized. The uterine contents are then mapped with a sterile ultrasound transducer, and the location of the fetus and the placenta are determined. Initial uterine entry is obtained with a specialized trocar developed at Vanderbilt University Medical Center (Cook Incorporated; Bloomington, IN). The Tulipan–Bruner trocar consists of a tapered central introducer covered by a peel-away Teflon sheath. Use of this trocar has demonstrated to reduce operative time and blood loss while providing atraumatic entry into the uterine cavity (36). Two through-and-through chromic sutures are passed through the uterine wall and membranes on either side of the selected entry point. The introducer is then passed into the uterine cavity under direct ultrasonographic guidance using a modified Seldinger technique. The central introducer is then removed, leaving only the trocar sheath. Excess amniotic fluid may be aspirated and stored in sterile, warm syringes. The footplate of a U.S. surgical CS-57 autostapling device (United States Surgical Corporation; Norwalk, CT) is then inserted through the peel-away sheath, and the sheath is removed, leaving the stapler in proper position. When activated, the stapler creates a 6–8 cm uterine incision. At the same time, all the layers of the uterine wall are held together, much like the binding of a book.

The fetus is directly visualized and manually positioned within the uterus so that the myelomeningocele sac is located in the center of the hysterotomy. Proper position is maintained by grasping the fetal head and trunk through the flaccid uterine wall. During the procedure, the fetal heart rate is monitored by continuous ultrasonographic visualization.

The myelomeningocele is closed in routine neurosurgical fashion. Approximately 20% of patients will not have a well-formed myelomeningocele sac, but a crater-like lesion termed myeloschisis. As fetuses with myeloschisis have less viable skin for closure, it may be necessary to use bilateral vertical relaxing incisions in the flanks to create bipedicular flaps that can be advanced and closed over the dural sac. The resulting full-thickness cutaneous defects are covered with cadaveric skin (37).

After repair of the spina bifida lesion, the uterus is closed in layers using #1 PDS sutures. The first layer incorporates the absorbable polyglycolic acid staples left by the autostapling device. As the last stitches of this layer are placed, the reserved amniotic fluid or physiologic crystalloid solution, mixed with 500 mg of nafcillin or an equivalent dosage of an antibiotic effective against Staphylococcus species, is replaced in the uterus. The sterile, warm fluid is added until the uterine turgor, as determined by manual palpation, is restored to the preoperative level, which is followed by an imbricating layer. A sheet of Interceed absorbable adhesion barrier (Johnson & Johnson Medical, Inc.; Arlington, TX) or omentum is attached over the incision to prevent adhesion formation. The uterus is returned to the abdomen. The fascial layer is closed in routine fashion, and the dermis closed with a running subcuticular suture or staples. The fetus is monitored postoperatively using continuous electronic fetal monitoring (EFM) and intermittent transabdominal ultrasonography.

Postoperative uterine contractions are monitored using continuous EFM. Uterine contractions are initially controlled with intravenous magnesium sulfate and oral or rectal indomethacin, and subsequently with subcutaneous terbutaline or oral nifedipine, supplemented by indomethacin as needed. Patients are monitored with weekly transabdominal ultrasonographic examinations. Delivery of each child is accomplished via standard cesarean section. Although the same abdominal incision is used for the cesarean section as for the fetal surgery, the fetus is preferably delivered via a lower uterine segment incision.

The uterus and abdominal incisions are closed in routine fashion.

VALVULOPLASTY

Severe aortic stenosis in midgestation may lead to left ventricular myocardial damage and can ultimately result in hypoplastic left heart syndrome. Paradoxically, in these fetuses, which are likely to progress to HLHS, the left ventricle initially appears normal in size, or even enlarged, in the setting of left ventricular systolic dysfunction. As gestation progresses, diminished flow through the diseased left ventricle leads to decreased flow, and the ventricle experiences growth arrest, resulting in left heart hypoplasia at birth. Early relief of fetal aortic stenosis may preserve left heart function and growth potential by maintaining flow through the developing chamber. To this end, a number of operators have developed techniques to perform fetal aortic valvuloplasty in second trimester fetuses.

The mother is placed under general anesthesia in a supine position with left lateral uterine displacement. Transabdominal ultrasound imaging and external manipulation are employed to achieve ideal fetal position. In this position, a line of approach from the anterior abdominal surface traverses the apex of the fetal left ventricle (LV), paralleling the LV outflow tract, and crossing the valve into the ascending aorta. The fetus is given intramuscular anesthetic and muscle relaxant prior to catheterization. If unable to position the fetus using external maneuvers, the operators perform a limited laparotomy to enable direct uterine manipulation and transuterine imaging.

A low profile, over-the-wire coronary angioplasty catheter is chosen with a balloon diameter based on the measurement of the aortic annulus, using a balloon:annulus ratio of 1:2. The balloon catheter is mounted on a floppy-tipped guidewire, with 3 cm of distal wire exposed. The wire/catheter assembly is then advanced through the 19G 12 cm stainless-steel introducer cannula until the balloon emerges. Affixing a visible and palpable marker on the proximal catheter shaft allows the operator to reproduce this balloon/cannula relationship during the procedure without relying wholly on the ultrasound imaging.

The introducer is advanced through the fetal chest wall and to the LV epicardium under ultrasound guidance. The LV is entered with the introducer, and the obturator removed with the tip of the cannula just below the aortic valve (Fig. 1). Blood return through the cannula confirms an intracavitary position.

The wire/catheter assembly is passed through the cannula, and the tip of the wire is identified as it emerges. While maintaining imaging of the aortic valve and ascending aorta, the precurved wire tip is manipulated to probe for the valve. Valve passage, confirmed echocardiographically by imaging the wire in the ascending aorta, is followed by catheter insertion to the premarked depth. The balloon is then inflated, by hand or by pressure gauge, to a pressure at which it achieves the intended balloon:annulus ratio. Upon completion of the dilation, the entire apparatus is removed from the fetus.

Figure 1. Transabdominal ultrasound imaging during introducer cannula insertion into dilated fetal left ventricle. The tip of the introducer is positioned in the left ventricular outflow tract directed at the stenotic aortic valve.

When first reported in the 1990s, fetal aortic valve dilation was performed with minimal technical success (38). Using the technique described above, technical success rates are now over 80% in fetuses between 21 and 26 weeks (39). As the technical aspects of the procedure continue to be refined, and safety is established, issues of patient selection will become the major focus of ongoing research. Anatomic and physiologic variables predicting left ventricular normalization following successful fetal aortic valvuloplasty remain poorly understood.

AMNIOEXCHANGE

Gastroschisis is a paraumbilical defect of the anterior abdominal wall associated with intrauterine evisceration of the fetal abdominal organs. The incidence of gastroschisis is approximately 1:4000 births, with a 1:1 male:female ratio. Most cases are sporadic and aneuploidy is uncommon.

Gastroschisis is characterized by a full-thickness defect of the abdominal wall, usually located to the right of the umbilical cord, which has a normal insertion. The defect in the abdominal wall is generally quite small (3–5 cm). The herniated organs include mainly bowel loops, although, in rare cases, the spleen and liver may be involved. Intestinal atresias and other gastrointestinal disruptions are found in as many as 15% of cases, and malrotation is also universal.

Although the prognosis is excellent with an ultimate survival of greater than 90%, many factors may jeopardize the outcome of these infants. A chronic aseptic amniotic fluid peritonitis (perivisceritis) often occurs. The herniated organs become covered by an inflammatory peel in the third trimester, resulting from chemical irritation by exposure to digestive enzymes in the amniotic fluid. Thickening, edema, and matting together of the intestines occurs in these cases, and may result in a secondary ischemic injury

to the bowel as the abdominal defect becomes too small. Meconium is frequently found in the amniotic fluid of affected fetuses. Its presence probably reflects intestinal irritation. Intrauterine growth restriction (IUGR) is frequent, occurring in up to 60% of fetuses. Oligohydramnios can occur, and may lead to fetal stress by cord compression. Premature birth is a frequent and still poorly understood complication. At birth, infants have low serum albumin and total protein levels, which probably results from chronic peritonitis.

The obstetrical management of the fetus with gastroschisis is controversial. Some studies have shown no clear benefit of cesarean delivery over vaginal delivery, where as others demonstrate an improved perinatal outcome in infants delivered by elective cesarean section prior to labor. Postoperative infection and delayed total enteral nutrition are the major acute complications of the newborn. Although all neonates with gastroschisis require surgery shortly after birth, repair may be by primary fascial closure, or by delayed fascial closure using temporary coverage with a silastic/Dacron intra-abdominal pouch. The repair as a primary or secondary procedure depends on the degree of chemical peritonitis with matting of the bowel that is present. Delayed intestinal function with poor enteral nutrition is expected in most patients. Central venous access and early total parenteral nutrition are therefore usually required.

In a study by Luton et al. (40), gastroschisis was created at mid gestation in 21 lamb fetuses. Saline was amnioinfused in some fetuses every 10 days until term. Thickness of the bowel muscularis, thickness of the serous fibrosis, and plasma cell infiltration were all significantly improved in amnioexchanged animals when compared with fetal lambs that were not amnioexchanged (40). Histologic analysis of appendices removed from human newborns demonstrated increased fibrosis in those with gastroschisis; after amnioinfusion, the serosa was still edematous, but no inflammation was seen (41). In a pilot study in human pregnancies (42), the same authors investigated the effect of amnioinfusion on the outcome of prenatally diagnosed gastroschisis. Following up on their work showing that an inflammatory response exists in the amniotic fluid of fetuses with gastroschisis, they hypothesized that amniotic fluid exchange would improve the outcomes of prenatally diagnosed cases. The outcome of 10 amnioinfused fetuses with gastroschisis was compared with 10 nonamnioinfused matched controls. Results showed that fetuses undergoing amnioinfusion had a shorter duration of curarization after surgical repair (2.2 ± 1.9 versus 6.8 ± 6.9 days, $+ = 0.019$), a shorter delay before full oral feeding (49.7 ± 21.5 versus 72.3 ± 56.6 days, NS) and a shorter overall length of hospitalization (59.5 ± 19.7 versus 88.5 ± 73.6 days, NS). The authors confirmed their previous data showing that amniotic fluid displays a chronic inflammatory profile, and they speculated that a reduction of the inflammatory response could improve the outcome of human fetuses with gastroschisis (42).

Amnioexchange for treatment of gastroschisis begins after 30 weeks' gestation, and is repeated approximately every two weeks until delivery. A complete obstetrical ultrasound examination is performed prior to the amnioexchange. The patient is admitted to the labor and delivery suite, and a nonstress test is performed before and after the amnioexchange. An intravenous line or a heplock is placed, and a single vial of blood is obtained and held for routine admission laboratory studies, in the case that urgent delivery is required. Prophylactic tocolysis may be given in the form of intravenous or subcutaneous terbutaline. Light IV sedation may also be given if desired.

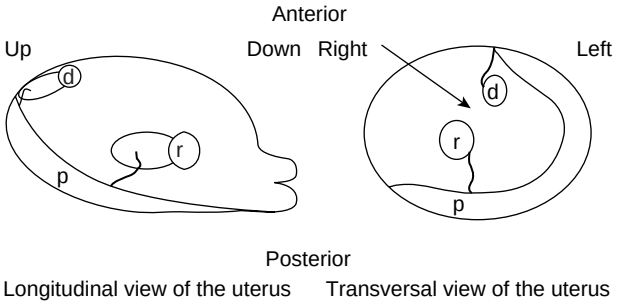


Figure 2. Views of the uterus.

After performing a sterile abdominal prep and drape, amnioexchange is performed using a "closed system." The closed tubing system materials are illustrated in Fig. 2. All of the materials are sterile except the graduated cylinder. Two sterile three-way stopcocks (b and c) are connected end-to-end. Sterile IV solution (a) is connected to stopcock (c) by way of sterile IV tubing (a). Three lengths of sterile connecting IV tubing (b) are connected to the remaining exposed ports of the stopcock assembly. Tubing from the side port of stopcock (b) is allowed to drain to the graduated cylinder (d). A 60 mL syringe (e) is connected to the tubing attached to the inline port of stopcock (b). The tubing attached to the inline port of stopcock (c) will connect to the therapeutic amniocentesis needle. Stopcock (c) is closed off to the IV solution (a). Stopcock (b) is closed off to the graduate (d). Amniocentesis is performed under continuous ultrasound guidance. Once access is obtained, the stylet is removed from within the needle lumen and the connection tubing is attached. With the stopcocks positioned as noted above, amniotic fluid is withdrawn into the syringe (e) until the syringe is filled. Stopcock (b) is then closed off to the patient and the fluid is expelled into the graduate (d). Stopcock (b) is then turned off to the graduate (d), and this step is repeated until the desired amount of amniotic fluid has been withdrawn (300–900 ml). With stopcock (b) closed to the graduate (d), stopcock (c) is closed off to the patient. The syringe is then filled with sterile warmed saline. Stopcock (c) is then closed to the IV solution (a) and the fluid is infused into the patient. This step is repeated until the desired amount of fluid is infused. These steps are repeated serially until the infusion procedure is complete. If the amniotic fluid volume falls within normal range, the amniotic fluid volume at the end of the amnioexchange should be the same as at the beginning of the procedure. In the presence of oligohydramnios, additional sterile warmed fluid can be added to the uterine cavity in order to achieve a normal fluid volume by the end of the procedure.

If the fluid volume is normal, the placenta is located posteriorly or fundally, and the fetus is quiescent, all of the toxic amniotic fluid planned for removal might be aspirated in one step. With an anteriorly implanted placenta, however, or in the presence of oligohydramnios or an active fetus, it may be necessary to remove a small amount of amniotic fluid, and then replace it with warmed normal saline, repeating the procedure serially until the amnioexchange is completed.

After completion of the amnioexchange, electronic monitoring of the fetal heart rate and uterine activity continues until the patient fulfills the usual criteria for discharge.

AMNIOPATCH

Iatrogenic preterm premature rupture of membranes (PPROM) occurs in approximately 1.2% of patients after genetic amniocentesis (43), 3–5% of patients after diagnostic fetoscopy (44), and approximately 5–8% of patients after operative fetoscopy. Although the membranes might seal spontaneously in this setting (45,46), most patients continue to leak fluid and are at a risk for pregnancy loss.

The overall perinatal mortality of previable PPROM managed expectantly is 60% (47,48). Nearly one-third of these deaths occurs *in utero*. Pulmonary hypoplasia occurs in 50% of cases diagnosed before 19 weeks (49). Serious sequelae in surviving infants include blindness, chronic lung disease, and cerebral palsy.

Patients with iatrogenic PPROM between 16 and 24 weeks gestation who do not have clinical evidence of intraamniotic infection are candidates for amniopatch therapy. PPROM is confirmed with a sterile speculum examination showing vaginal pooling of fluid, ferning, and a positive Nitrazine test. The maximum vertical pocket of amniotic fluid is measured sonographically. Patients are placed on intravenous antibiotics and bed rest for one week to allow for spontaneous sealing of the membranes. If spontaneous sealing does not occur, 1 unit of autologous platelets and cryoprecipitate are prepared if the patient is eligible for autologous donation. Otherwise, donor platelets and cryoprecipitate are prepared.

After informed consent, an amniocentesis is performed using a 22 gauge needle. The needle is directed into an available pocket of fluid regardless of the site of the previous invasive procedure. A K-51 tubing extension attached to a three-way stopcock is connected to the hub of the needle. Platelets are administered first, followed by cryoprecipitate. In our original protocol, 1 whole unit of platelets was injected. We have subsequently reduced the dose of platelets to one-half a unit because of an unexplained fetal death demise, an adverse effect probably caused by sudden activation of a large number of platelets.

In our series of 28 cases, the average gestational age at the time of the procedure was 19 weeks and 3 days. The average gestational age of delivery in patients who did not have an intrauterine fetal demise was 33 weeks and 4 days. Overall, membrane sealing occurred in 19 of 28 patients (67.9%). Of the 28 patients treated, 11 had a large membrane detachment but no overt leakage of fluid. The detachment of the membrane occurred from fluid escaping the amniotic cavities through the membrane defect, causing dissection of the chorionic cavity. In these patients, only the chorion separates this fluid from leaking grossly to the vagina. In this group, the amniopatch was successful in resealing the amniotic membrane in 7 of the 11 patients (63.6%).

The precise mechanism by which the amniopatch works is unknown. Presumably, platelet activation at the site of rupture and fibrin formation initiates a healing process that enables the membranes to seal.

TTTS

Feto-fetal transfusion syndrome can be described in all monochorionic multiple pregnancies but has been extensively reported in twins. Twin-to-twin transfusion syndrome (TTTS) develops in approximately 15% of all monochorionic pregnancies (50), and carries a high perinatal mortality rate (50). The fetuses are morphologically normal, and inter-twin vascular communications on the chorionic plate are thought to be responsible for the development of the disease through unidirectional blood transfusion from the donor to the recipient twin. Besides the primary hemodynamic imbalance between the twins, the disease may lead to disruptive lesions in both twins. Before the development of antenatal ultrasound, TTTS was diagnosed at birth as a discordance of at least 20% in weight and 5 g/dL in the hemoglobin concentrations of two twins of the same sex (52). These criteria were abandoned because these features could not be consistently recognized in utero. With the development of ultrasound, the polyhydramnios-oligohydramnios sequence has been found to be the condition carrying one of the highest perinatal mortality rates in obstetrics, up to 90% without treatment.

Laser coagulation of placental anastomoses by fetoscopy is the most effective first-line treatment of FFTS, which leads to at least one survivor at birth and intact survival at 6 months of age in 76% and 76% respectively, as compared with 56% and 51% in cases treated by serial amnioreduction in the Eurofetus randomized trial (53).

The selection criteria to qualify for percutaneous endoscopy-directed laser coagulation of placental anastomoses include:

1. Gestational age of less than 26 weeks.
2. Ultrasound diagnosis of a single monochorionic placenta by ultrasound in the first trimester of pregnancy.
3. Polyhydramnios in the recipient's amniotic cavity with a deepest vertical pool $\geq$8 cm or $\geq$10 cm before or after 20 weeks of gestation, respectively.
4. Oligohydramnios in the donor's amniotic sac with a deepest vertical pool $\leq$2 cm.

Preoperative evaluation consists of ultrasound examination, including morphological examination, fetal Doppler, cardiothoracic index, identification of placental location, and cord insertions. Amniocentesis or amniore-

duction prior to laser may cause intra-amniotic bleeding and therefore make the procedure more difficult, or even impossible, due to impaired visualization. The site of entry is chosen as demonstrated in Fig. 1, for the scope to be entered at a right angle to the long axis of the small twin in order to maximize the chance to ensure adequate visualization of the placental surface and intertwin membranes. Ideally, the scope should also be entered alongside a virtual line joining the two cord insertions. When these criteria are met, the vascular equator of the placenta as well as the vascular anastomoses on the chorionic plate are more likely to be visualized in the operative field.

Prophylactic cefazolin 2 g, indomethacin suppository 100 mg, and oral flunitrazepam are given before surgery and local anesthesia with nonadrenalinized xylocaine is injected down to the myometrium. A 10 Fr cannula for a central venous catheter loaded with a trocar is introduced percutaneously under continuous ultrasound guidance. A 2 mm 0° fetoscope (Storz 26008 AA) is passed down a straight or curved sheath to operate on posterior or anterior placentas, respectively. The sheath also has a working channel carrying a 1 mm diode laser fiber.

A systematic examination of the chorionic plate alongside the insertion of the inter-twin membrane is performed. Identification of crossing vessels and of their arterial or venous nature is possible because arteries cross over veins and show a darker red color than veins owing to a lower oxygen saturation in the circulating blood (54). Selective coagulation of anastomotic vessels is performed with the aim of separating the monochorionic placenta into two distinct fetal-placental circulations, sparing the normal cotyledons of each placental territory. Nonselective coagulation of crossing vessels is only performed when the distal end or the origin of the vessel cannot be identified. The power of the diode laser is set at 30–50 w. At the end of theprocedure, excessive amniotic fluid is drained through the sheath of the fetoscope until normal amniotic fluid volume is obtained with a deepest vertical pool of 5–6 cm.

BIPOLAR UMBILICAL CORD OCCLUSION

Selective reduction in complicated monochorionic (MC) multifetal pregnancies is performed to prevent the delivery of an anomalous or severely compromised fetus and improve the perinatal outcome for the surviving co-twin by delaying delivery or risk associated with spontaneous loss of the affected. The use of cardiotoxic agents, such as potassium chloride, is contraindicated in MC pregnancies because of the potential vascular transmission of the agent and compromise of the co-twin due to the presence of placental vascular anastomoses. Thermal vascular occlusive techniques, such as bipolar umbilical cord occlusion (BPC), have been shown to achieve the stated goals with minimal maternal morbidity. Indications for BPC that are unique to MC gestations include twin reverse arterial perfusion, discordant fetal anomalies, and isolated severe growth lag. BPC has also been used as a primary intervention in advanced twin-twin transfusion syndrome or as a secondary procedure when alternative therapies such as amnioreduction or laser have failed to correct the disease process.

The procedure was originally described by Deprest et al. (55). In brief, using the standard sterile technique, the patient's abdomen is properly prepared and draped. An abdominal ultrasound is performed to confirm fetal position, viability, and umbilical cord locations. General and conduction anesthesia may be used; however, intravenous sedation with local infiltration of 1% lidocaine or 0.25% bupivicane for subcutaneous, deep muscle, and fascia anesthesia is usually sufficient and is associated with less maternal morbidity. A small skin incision is made to allow insertion of an endoscopic trocar. Under continuous ultrasound guidance, the instrument is inserted through a placental free window toward the targeted umbilical cord, ideally avoiding the gestational sac of the normal co-twin. Once the trocar is secured in the amniotic sac, the obturator is removed. The bipolar forceps are inserted and advanced to the umbilical cord.

The cord is grasped and positioned away from the amnion before thermal energy is applied. The duration and wattage (W) necessary for occlusion will vary, from 20–60 s and 20–50 W, respectively, based on the gestational age and umbilical cord thickness. When a full-thicknessgrasp exists, application of the thermal energy will result in turbulence and "streaming" of amniotic fluid adjacent to the forceps. It is not uncommon to have an audible "pop" secondary to the heating of Wharton's jelly and subsequent rupture of amnion at the site of occlusion, which should not be perceived as a sign of completed coagulation. As a result of the natural spiral of the umbilical cord, complete occlusion of all vessels requires 2–3 applications of the forceps at adjacent sites. Pulse and color flow Doppler blood flow studies are performed to confirm cord occlusion at each site.

The size of the BPC forceps that have been used for these procedures has varied from 2.2–5.0 mm. The majority of procedures have been performed with commercially available single-use 3.0 mm bipolar diathermy forceps (55–58).

Intravenous prophylactic antibiotics and indomethacin for tocolysis are generally given prior to the procedure. Postoperative monitoring for uterine contractions and, depending on the gestational age, continuous or intermittent fetal heart rate should be done for at least 2 hours. The majority of programs will observe patients for an extended period of 12–24 h with limited activity. Subsequent doses of antibiotics and tocolytic treatment are given during this time. Prior to discharge, a limited ultrasound is performed to determine the amniotic fluid volume and assess for signs of hydrops and anemia, including Doppler velocemitry of the middle cerebral artery and, where appropriate, similar studies of the umbilical artery and ductus venosus. If no evidence of preterm labor, leaking of amniotic fluid, or bleeding exists, the patient is discharged with instructions to continue with modified bed rest at home for 7–10 days, take her temperature bid, and report an elevation, leaking of vaginal fluid, bleeding, or contractions. An ultrasound is performed in 10–14 days and then at a minimum every 4 weeks thereafter. Additional ultrasounds and fetal monitoring should be performed as clinically indicated by the primary disease and gestational age.

BIBLIOGRAPHY

1. Harrison MR, Adzick NS. Open Fetal Surgery Techniques. The Unborn Patient: The Art and Science of Fetal Therapy, 3rd ed. Harrison MR, Evan MI, Adzick NS, Holzgreve W, editors. New York: WB Saunders Company; 2001. pp 247–255.
2. Harrison MR, Anderson J, Rosen MA, et al. Fetal surgery in the primate I. Anesthetic, surgical and tocolytic management to maximize fetal-neonateal survival. J Pediatr Surg 1982;17:115–122.
3. Nakayama DK, Harrison MR, Seron-Ferre M, et al. Fetal surgery in the primate II. Uterine electromyographic response to operative procedure and pharmacologic agents. J Pediatr Surg 1984;19:333–339.
4. Adzick NS, Harrison MR, Glick PL, et al. Fetal surgery in the primate III. Maternal outcome after fetal surgery. J Pediatr Surg 1986;21:477–480.
5. Bianchi DW, Crombleholme TM, D'Alton ME. Cystic Adenomatoid Malformation. In: Bianchi DW, Crombleholme TM, D'Alton ME, editors. Fetology-Diagnosis & Management of the Fetal Patient. New York: McGraw-Hill; 2000. pp 289–297.
6. Bianchi DW, Crombleholme TM, D'Alton ME. Sacrococcygeal teratoma. In: Bianchi DW, Crombleholme TM, D'Alton ME, editors. Fetology-Diagnosis & Management of the Fetal Patient. New York: McGraw-Hill; 2000. pp 867–877.
7. Altman RP, Randolph JG, Lilly JR. Sacrococcygeal teratoma: American Academy of Pediatrics Surgical Section Survey 1973. J Pediatr Surg 1974;9:389–398.
8. Adzick NS. Management of fetal lung lesions. Clin Perinatol 2003;30:481–492.
9. Hedrick HL, Flake AW, Crombleholme TM, et al. Sacrococcygeal teratoma: Prenatal assessment, fetal intervention, and outcome. J Pediatr Surg 2004;39(3):430–438.
10. Crombleholme TM, Coleman B, Hedrick H, et al. Cystic adenomatoid malformation volume ratio predicts outcome in prenatally diagnosed cystic adenomatoid malformation of the lung. J Pediatr Surg 2002;27(3):331–338.
11. Rice HE, Estes JM, Hedrick MH, et al. Congenital cystic adenomatoid malformations: A sheep model. J Pediatr Surg 1994;29:692–696.
12. Flake AW. Fetal sacrococcygeal teratoma. Sem Pediatr Surg 1993;2:113–120.
13. Adzick NS, Crombleholme TM, Morgan MA, et al. A case report. A rapidly growing fetal teratoma. Lancet 1997;349:538.
14. Hedrick HL. Ex utero intrapartum therapy. Semi Ped Surg 2003;10(3):190–195.
15. Wilson RD, Johnson MP, Flake AW, et al. Reproductive outcomes after pregnancy complicated by maternal-fetal surgery. Am J Obstet Gynecol 2004;191:1430–1436.
16. Harrison MR, Filly RA, Parer JT, et al. Management of the fetus with a urinary tract malformation. JAMA 1981; 246(6):635–639.
17. Nakayama D, Harrison M, deLorimier A. Prognosis of posterior urethral valves present at birth. J Ped Surg 1986;21:43–45.
18. Golbus MS, Harrison MR, Filly RA, et al. In utero treatment of urinary tract obstruction. Am J Obstet Gynecol 1982; 383–388.
19. Berkowitz RL, Glickman MG, Smith GJ, et al. Fetal urinary tract obstruction: What is the role of surgical intervention in utero? Am J Obstet Gynecol 1982;144(4):367–375.
20. Rodeck C, Nicolaides K. Ultrasound guided invasive procedures in obstetrics. Clin Obstet Gynecol 1983;10:515.
21. Beck AD. The effect of intra-uterine urinary obstruction upon the development of the fetal kidney. Urol 1971;105:784–789.
22. Pringle KC, Bonsib SM. Development of fetal lamb lung and kidney in obstructive uropathy: A preliminary report. Fetal Ther 1988;3(1-2):118–128.
23. Manning FA, Harman CR, Lange IR, et al. Antepartum chronic fetal vesicoamniotic shunts for obstructive uropathy: A report of two cases. Am J Obstet Gynecol 1983;145(7):819–822.
24. Young H, Frontz W, Baldwin J. Congenital obstruction of the posterior urethra. J Urol 1919;3:289–365.
25. Reuss A, Wladimiroff J, Niermeyer M. Antenatal diagnosis of renal tract anomalites by ultrasound. Pediat Nephrol 1987;1:546–552.
26. Johnson MP, Bukowski TP, Reitleman C, et al. In utero surgical treatment of fetal obstructive uropathy: A new comprehensive approach to identify appropriate candidates for vesicoamniotic shunt therapy. Am J Obstet Gynecol 1994;170(6):1770–1776; discussion 1776–1779.
27. Quintero RA, Hume R, Smith C, et al. Percutaneous fetal cystoscopy and endoscopic fulguration of posterior urethral valves [see comments]. Am J Obstet Gynecol 1995;172(1 Pt 1):206–209.
28. Quintero RA, Johnson MP, Romero R, et al. In-utero percutaneous cystoscopy in the management of fetal lower obstructive uropathy. Lancet 1995;346(8974):537–540.
29. Quintero RA, Shukla AR, Homsy YL, et al. Successful in utero endoscopic ablation of posterior urethral valves: A new dimension in fetal urology. Urology (Online) 2000;55(5): 774.
30. Steinbok P, Irvine B, Cochrane DD, Irwin BJ. Long-term outcome and complications of children born with myelomeningocele. Child's Nerv Syst 1992;8:92–96.
31. McLone DG. Continuing concepts in the management of spina bifida. Pediare Neurosurg 1992;18:254–257.
32. Tulipan N, Hernanz-Schulman M, Bruner JP. Reduced hindbrain herniation after intrauterine myelomeningocele repair: A report of four cases. Pediatr Neurosurg 1998;29: 274–278.
33. Tulipan N, Hernanz-Schulman M, Bruner JP. Intrauterine myelomeningocele repair reverses preexisting hindbrain herniation. Pediatr Neurosurg 1999;31:137–142.
34. Bruner JP, Tulipan N, Paschall RL, Boehm FH, Walsh WF, Silva SR, Hernanz-Schulman M, Lowe LH, Reed GW. Fetal surgery for myelomeningocele and the incidence of shunt-dependent hydrocephalus. JAMA 1999;282:1819–1825.
35. Johnson MP, Sutton LN, Rintoul N, Crombleholme TM, Flake AW, Howell LJ, Hedrick HL, Wilson RD, Adzick NS. Fetal myelomeningocele repair: Short-term clinical outcomes. Am J Obstet Gynecol 2003;189:482–487.
36. Bruner JP, Boehm FH, Tulipan N. The Tulipan-Bruner trocar for uterine entry during fetal surgery. Am J Obstet Gynecol 1999;181:1188–1191.
37. Mangels KJ, Tulipan N, Bruner JP, Nickolaus D. Use of bipedicular advancement flaps for intrauterine closure of myeloschisis: Technical report. Pediatr Neurosurg 2000; 32:52–56.
38. Kohl T, Sharland G, Allan LD, Gembruch U, Chaoui R, Lopes LM, Zielinsky P, Huhta J, Silverman NH. World experience of percutaneous ultrasound-guided balloon valvuloplasty in human fetuses with severe aortic valve obstruction. Am J Cardiol 2000;85:1230–1233.
39. Tworetzky W, Wilkins-Haug L, Jennings RW, van der Velde ME, Marshall AC, Marx GR, Colan SD, Benson CB, Lock JE, Perry SB. Balloon dilation of severe aortic stenosis in the fetus: Potential for prevention of hypoplastic left heart syndrome: candidate selection, technique, and results of successful intervention. Circulation 2004;110:2125–2131.

40. Luton D, de Lagausie P, Guibourdenche J, Peuchmaur M, Sibony O, Aigrain Y, Oury IF, Blot P. Influence of amnioinfusion in a model of in utero created gastroschisis in the pregnant ewe. Fetal Diagn Ther 2000;15:224–228.
41. Luton D. Etude de L'inflammation, dans le Laparoschisis, Humain et dans un Modele de Laparoschisis de Bredis. These; 2001.
42. Luton D, de Lagausie P, Guibourdenche J, Oury IF, Sibony O, Vuillard E, Boissinot C, Aigrain Y, Beaufils F, Navarro J, Blot P. Effect of amnioinfusion on the outcome of prenatally diagnosed gastroschisis. Fetal Diagn Ther 1999;14:152.
43. The NICHD National Registry for Amniocentesis Study Group. Midtrimester amniocentesis for prenatal diagnosis. Safety and accuracy. JAMA 1976;236:1471–1476.
44. Rodeck C. Fetoscopy guided by real-time ultrasound for pure fetal blood samples, fetal skin samples, and examination of the fetus in utero. Br J Ob Gyn 1980;87:449–456.
45. Quintero R, Reich H, Puder K, et al. Brief report: Umbilical-cord ligation of an Acardiac twin by fetoscopy at 19 weeks of gestation. N Engl J Med 1994;33:469–471.
46. Gold R, Goyert G, Schwartz D. Conservative management of second-trimester post-amniocentesis fluid leakage. Obstet Gynecol 1989;74:745–747.
47. Bengtson J, VanMarter L, Barss V. Pregnancy outcome after premature rupture of the membranes at or before 26 weeks' gestation. Obstet Gynecol 1989;73:921–927.
48. Beydoun S, Yasin S. Premature rupture of the membranes before 28 weeks: Conservative management. Am J Obstet Gynecol 1986;155:471–479.
49. Rotschild A, Ling E, Puterman M. Neonatal outcome after prolonged preterm rupture of the membranes. Am J Obstet Gynecol 1990;162:46–52.
50. Sebire NJ, Snijders RJ, Hughes K, et al. The hidden mortality of monochorionic twin pregnancies. Br J Obstet Gynecol 1997;104(10):1203–1207.
51. Patten RM, et al. Disparity of amniotic fluid volume and fetal size: Problem of the stuck twin-US studies. Radiology 1989;172:153–157.
52. Danskin FH, Neilson JP. Twin-to-Twin transfusion syndrome: What are appropriate diagnostic criteria? Am J Obstet Gynecol 1989;161:365–369.
53. Senat MV, Deprest J, Boulvain M, Paupe A, Winer N, Ville Y. Endoscopic laser surgery versus serial amnioreduction for severe twin-to-twin transfusion syndrome. N Engl J Med 2004;351:136–144.
54. Benirschke K, Driscoll S. The Pathology of the Human Placenta. New York: Springer-Verlag; 1967.
55. Deprest JA, Audibert F, Van Schoubroeck D, Hecher K, Mahieu-Caputo D. Bipolar coagulation of the umbilical cord in complicated monochorionic twin pregnancy. Am J Obstet Gynecol 2000;182:340–345.
56. Nicolini U, Poblete A, Boschetto C, Bonati F, Roberts A. Complicated monochorionic twin pregnancies: Experience with bipolar cord coagulation. Am J Obstet Gynecol 2001;185:703–707.
57. Taylor MJO, Shalev E, Tanawattanacharoen S, Jolly M, Kumar S, Weiner E, Cox PM, Fisk NM. Ultrasound-guided umbilical cord occlusion using bipolar diathermy for Stage III/IV twin-twin transfusion syndrome. Prenat Diagn 2002; 22:70–76.
58. Johnson MP, Crombleholme TM, Hedrick H, et al. Bipolar umbilical cord cauterization for selective termination of complicated monochorionic pregnancies. Am J Obstet Gynecol 2001;185(6):S245.

See also FETAL MONITORING; HYDROCEPHALUS, TOOLS FOR DIAGNOSIS AND TREATMENT OF; MONITORING, UMBILICAL ARTERY AND VEIN.

OIL. See LENSES, INTRAOCULAR.

ION-EXCHANGE CHROMATOGRAPHY. See CHROMATOGRAPHY.

IONIZING RADIATION, BIOLOGICAL EFFECTS OF

ZELENNA GOLDBERG
Department of Radiation Oncology
Davis, California

INTRODUCTION

Radiation biology refers to all biologic responses induced in cells and tissues by ionizing radiation, a term that encompasses the study of all action of ionizing radiation on living things. Ionizing radiation (IR) is radiation that has sufficient energy to cause the ejection of an orbital electron from its path around the nucleus, which indicates that the photon or charged particle can release large amounts of energy in a very small space. IR is separated into electromagnetic radiation and particulate radiation. Electromagnetic radiation consists of X rays or γ rays, which differ only in their mode of production, not in their physical effects in interacting with biologic tissue. X rays are produced by machines that accelerate electrons and then focus them to hit a target usually made of tungsten or gold. The kinetic energy is transferred from the electrons to the target and then is released as photons. These are the most common medical exposures through diagnostic or therapeutic radiation. In contrast, γ-radiation is produced within the nucleus from radioactive isotopes. The γ rays result from the unstable nuclear configuration decaying to a more stable state and releasing the "extra" energy in this transition in the form of the γ ray. Natural background radiation is of this type. Particulate radiation is those radiation sources that are not photons and consist of electrons, protons, α-particles, neutrons, and heavy charged particles such as the nuclei of carbon, neon, argon, or iron. These particles are positively charged because the orbiting electrons have been stripped from them.

As noted, IR is defined by its causing the ejection of an orbital electron when the radiation interacts with tissue. These ionizing events can be further subdivided by being directly ionizing or indirectly ionizing. Directly ionizing events are those in which a charged particle with sufficient kinetic energy interacts with the tissue, directly resulting in the ejection of the orbital electron. These events happen frequently within an exposed tissue, so the charged particle rapidly transfers its energy to the tissue, a concept codified as linear energy transfer (LET). Charged particles have a high LET. In contrast, indirectly ionizing radiation such as γ- or X-radiation is first absorbed in the material and secondary, energetic charged particles (electrons) are released. Because this latter process only occurs when the photon is close enough to an atom to interact, it is referred to as sparsely ionizing radiation and has a low linear energy transfer (LET).

A longstanding paradigm in radiation biology has been that many effects induced by IR, including its carcinogenic effects and ability to kill cancer cells, are the result of DNA damage arising from the actions of IR in cell nuclei, especially interactions of IR and its products with nuclear DNA (1–3). When a charged particle or secondary electron produced by the interaction of a photon with an orbital electron damages DNA itself, it is known as the direct action of IR. Yet, DNA represents a small fraction of the actual size of the cell. Therefore, it was recognized that most of the interaction of IR within a cell would be with more abundant biomolecules, specifically water. If the action is mediated through the intracellular production of radiolytic reactive products, e.g., $OH^{\cdot}$, $H^{\cdot}$, O_2, and H_2O_2, that are generated in aqueous fluid surrounding DNA, then the DNA damage is called indirect action. These processes unquestionably can result in a variety of types of DNA damage, including DNA single- and double-strand breaks, modifications of deoxyribose rings and bases, intra- and inter-strand DNA–DNA cross-links, and DNA–protein cross-links (1,4,5). About a third of all DNA damage is caused by the direct effects of sparsely ionizing γ and X rays, with the remaining balance being attributable to the indirect actions of IR. With high-LET radiation such as the more densely ionizing α-particles that are emitted by radon and radon progeny, the direct actions of IR on DNA become more predominant and the nature of the DNA modifications become much more complex. Regardless of the type of IR, all of the above forms of DNA damage can lead to untoward effects in cells if unrepaired or misrepaired. With specific regard to carcinogenesis, genomic mutations caused by IR are widely thought to arise from DNA damage that is subsequently converted into a mutation as a result of misprocessing by DNA repair mechanisms or that is converted into a heritable mutation when DNA undergoes replication.

This classic view of radiation biology is giving way to a more complex and complete understanding that the cell membrane and other intracellular compartments, as well as the tissue vasculature, are important targets of IR. Furthermore, as detailed below, although intracellular effects are better defined, we now recognize that the extracellular environment and cell–cell contact play important roles in modulating the effects of IR at the cellular and tissue levels.

MOLECULAR RADIATION BIOLOGY/BYSTANDER EFFECTS

IR is a cellular toxin that is sensed at the cellular level through the ATM-p53-p21 pathway (6,7). Although the upstream sensor of ATM remains to be definitively elucidated, the initial radiation sensing protein likely has a redox sensor and undergoes a conformational change, or possibly is phosphorylated, resulting in the phosphorylation of ATM (8). This, in turn, activates the p53 pathway leading to cell cycle arrest, nuclear translocation of transcription factors such as NF-κB, and either cellular repair or apoptosis. How an individual cell commits to a given fate (repair or apoptosis) remains unclear, but undoubtedly it represents the final integrated response to many simultaneous intracellular events. Several excellent reviews on this topic have been written (9,10). It should be noted that IR also affects the 26s proteosome, which is another level of non-nuclear intracellular response affecting cell survival, presumably through alterations in the removal of activated (phosphorlyated) proteins, or proteins active in apoptotic processes such as Bcl2 (11,12).

Radiation effects in cells not directly hit by a radiation ionizing event are called bystander effects. Although initial interest in this effect can be found in the medical literature as far back as into the 1950s, the current interest in the area was stimulated by a study published in 1992 in which Nagasawa and Little (13) observed increases in the frequency of sister chromatid exchanges in $\sim$30% of immortalized Chinese hamster ovary cells that received low-dose exposure to α-particles, even though less than 1% of the cells' nuclei were estimated to actually receive direct nuclear hits by an α-particle. That a relatively low percentage of the cells experienced one or more direct "hits" by the α-particles, be they in the cytoplasm or in the nuclei, suggested the possibility that some mechanism was conveying a radiation-associated response to unirradiated cells. Other groups went on to confirm and extend on this finding. It is now well recognized that cells do not require a direct nuclear traversal to result in radiation changes. There are extracellular responses, predominantly TGF-β mediated through media (cell culture experiments) or extracellular fluids (tissues), but other factors cannot be excluded (14). Furthermore, it is recognized that cell–cell contact and gap junctional communications are critical in transmitting the signal from the directly irradiated cell to the neighboring, bystander cells (15).

TIME, DOSE, AND FRACTIONATION

The biologic effects of IR in tissue relate to the size of the dose delivered, time between radiation exposures and the total dose of IR given. Although environmental exposures are chronic and (usually) low level, medical exposures are acute and can be repetitive when given for the treatment of cancer. Radiation therapy for the treatment of malignancy remains the most effective anti-cancer agent discovered, and treatment schedules are predicated on the "4 R's of radiobiology": repair, repopulation, redistribution, and reoxygenation.

Repair of radiation-induced damage underlies the intrinsic radiation sensitivity of the cell to radiation cell killing. Cells can be broadly grouped into those that can repair significant amounts of damage and those with more limited repair capacity. The former descriptive category corresponds to tissue types where there is limited normal cell turnover, such as lung, kidney, or brain, and these are tissues that display damage after a more prolonged time and are therefore known as late responding tissues. The cells with more limited intrinsic repair capacity are known as acute responding tissues and are typified by skin or gut. These cell types have a limited life span and are constantly being replaced within normal physiologic functioning.

Repopulation refers to the generation of new cells to replace those killed by the IR exposure. Although therapeutically beneficial for containing normal tissue toxicity,

repopulation is also active in malignant tissue and thereby allows greater numbers of tumor clonogens to develop after treatment. For some types of tumors, an acceleration of repopulation begins after 4 weeks of radiation therapy, which can compromise clinical outcome if prolonged therapeutic IR fractionation schemes are used.

Redistribution refers to the changes in cell assortment across the cell cycle. It has long been established that cells vary in their sensitivity to the cytotoxic effects of IR depending on where in the cell cycle they are at the time of irradiation (16,17). Cells are most sensitive to IR effects in the G2/M phase and are most resistant during the S-phase. G1 is intermediate between these two. Thus, clinical radiation therapy schedules use multiple fractions to overcome the relative resistance of the cells in S-phase of the cell cycle on a given treatment day. This difference in radioresistance across stages in the cell cycle also underlies one mechanism of the synergy between radiation therapy and many chemotherapeutic agents in the treatment of malignant disease. Although S-phase cells resist killing from the IR, they are more sensitive to some chemotherapeutics that are active during S-phase when the DNA is most exposed and is replicating. Furthermore, this alteration of cell cycle sensitivity to IR cytotoxicity also has been an area of research for IR-biologic response modifiers. If one can increase the percentage of cells in G2/M at the time of IR, then the relative cell kill per fraction of IR will increase.

Reoxygenation refers to the presence or absence of molecular oxygen within the cell at the time of IR delivery. As detailed at the beginning, most of the cellular damage from IR is mediated through the production of free radicals within the cell. If molecular oxygen is present, these free radicals can be transformed into more complex peroxide radicals, which are more difficult for the cell to repair. This is classically known as "fixing the radiation damage" in the British use of the term "fix" (make more solid), not the American (repair). The increase in cell killing from IR in the presence of oxygen versus under hypoxic conditions is known as the oxygen enhancement ratio, and it is approximately a factor of 2–3, dependent on where in the cell cycle the irradiated cell sits. Although normal tissues have a well-maintained vascular supply so that oxygenation is essentially constant, tumors have a tortuous and unstable vasculature, so that the vessels can open and close erratically. This type of hypoxia is referred to as "acute hypoxia." Chronic hypoxia occurs when the tumor outgrows its blood supply and the tumor cells are simply beyond the diffusion capability of molecular oxygen. Attempts to exploit this differential, either by sensitizing the hypoxic cells or by specifically targeting them, have been explored and remain active areas of research. Furthermore, identification of the presence of hypoxia remains a significant clinical strategy (18).

The four R's of radiobiology reflect intracellular controls of overall radiation response. The tissue level effects from IR in the treatment of cancer are dependent on several parameters: volume, dose per fraction, total dose, and time between fractions. The volume of tissue irradiated is critical for determining the long-term repair by repopulation by normal cells to fill in for those irreparably damaged by the IR. Each cell type has its own intrinsic radiation sensitivity, and the tissue organization (serial or parallel cellular arrangement) as well as the tissue vasculature play critical roles in tissue repair and thus radiosensitivity. Tissues are subdivided into two categories of radiosensitivity based on when they display their damage. Tissues that display their radiation induced damage during a standard course of medical radiation therapy are known as "acute responding tissues." These tissues naturally have a large amount of cellular turnover, such as skin or gut, and at a cellular level, this corresponds to lesser ability to repair sublethal DNA damage (i.e., a larger proportion of DNA damage is lethal and nonrepairable). In contrast, tissues where there is little or no normal cellular turnover, such as brain, kidney, lung, or spinal cord, there is a substantial ability to repair sublethal damage, and tissue toxicity from radiation therapy is displayed late, long after therapy has finished. These tissues are therefore labeled "late responding tissues." Therapeutic strategies are designed to separate these two types of responses as malignant tumors are models of acute responding tissues, whereas the dose limiting side effects from radiation therapy are secondary to late responding tissue effects (19,20).

THE LINEAR NO-THRESHOLD MODEL OF RADIATION EFFECTS

Based on the data collected from the victims of the bomb detonations at Hiroshima and Nagasaki, a linear no-threshold model of radiation effects was developed and adopted for public policy applications. In essence, the model states that (1) all radiation exposures are biologically active, (2) the response in the cells/tissue is linear with dose, and (3) there is no threshold below which there is no or negligible effects. The model was developed from the moderately low-dose exposure ranges of 0.5–2 Gy and the medically significant sequelae of increased mortality (carcinogenic and noncarcinogenic, predominantly cardiovascular) (21–23). Hiroshima and Nagasaki data represent the effects of a single acute dose exposure delivered to the whole body with the nutritional deprivation from WW2; as such, they are subject to criticism and questions of their applicability to modern healthy populations where low-dose radiation exposure is often of a more chronic nature. Nevertheless, they remain the best available population-based data, and they have been painstakingly collected and analyzed. As detailed below, however, the shape of the response curve at the lowest doses remains controversial.

LOW-DOSE EXPOSURES

Low-dose ionizing radiation (LDIR) in the 1–10 cGy range has largely unknown biological activity in the human. Current modeling for health and safety regulations, as well as prediction of carcinogenesis, presupposes a linear, no-threshold model of radiation effects based on the nuclear bomb explosion data, which estimates the effect and risk at low dose by extrapolation from measured effects at high doses. Yet the scientific literature presents a more complex picture, and few data clearly support a linear

dose-response model and none in humans. Numerous studies suggest some effects of LDIR may be benign or even beneficial under some circumstances (24,25). Reported benefits include stimulated growth rates in animals, increased rates of wound healing, reductions in cell apoptosis, enhancements in the repair of damaged DNA, and increases in radioresistance via the induction of an adaptive response, among others (26–28). Other lines of evidence, however, suggest LDIR can be hazardous, and if a threshold for detrimental responses does exist, it is operational only at very low dose levels, e.g., $\leq$1 cGy. Schiestl et al. (29) found that 1–100 cGy doses of X rays could cause genetic deletions in mice in a linear dose-response manner. This remains an active area of research (30).

Little is known regarding individual variability in sensitivity to radiation exposure. Studies are actively ongoing to develop methods to best assess interindividual variability. This understanding will have both a health risk assessment and medical applications (31).

GENOMIC INSTABILITY

Radiation-induced genomic instability encompasses a range of measurable endpoints such as chromosome destabilization, sister chromatid exchanges, gene mutation and amplification, late cell death, and aneuploidy, all of which may be causative factors in the development of clinical disease, including carcinoma.

Kadhim et. al. identified the persistence of radiation-induced chromosomal instability following α-particle irradiation in clonal populations of murine bone marrow cells (32). The same group followed up this seminal work with an examination of four human bone marrow samples, subjected to an *ex vivo* α-particle IR (33). Further research then demonstrated that the genomic instability phenotype could be transmitted *in vivo* when murine hemopoietic cells that had been irradiated *in vitro* were transplanted into mice that had previously had their native bone marrow purged (34). However, when Whitehouse and Tawn examined radiation workers in Sellafield, England, who had bone marrow plutonium deposition evidence for genomic instability was not found (35). Whether γ-IR could induce genomic instability was then examined. The original reports from Kadhim et al. were negative (32,33), but further research examining hprt locus mutations convincingly demonstrated that genomic instability was inducible by γ-irradiation (36). Thus, although genomic instability can clearly be demonstrated in the laboratory, whether it occurs in humans after IR exposure remains uncertain (37).

Our understanding of the biologic effects of IR is evolving and ever growing. Research has been active in this field for over 100 years, and it remains a vibrant research area with the new molecular tools now available. The target of interest and concern is as small as the individual DNA base or as large as the whole organism. Mechanistic studies of the subcellular targets of IR and the cellular response signaling cascades must be matched with more complex system evaluations so that the summative effect of these pathways becomes known. Cell signaling exists within and between cells, and this cross-talk affects the ultimate response to IR exposure. Improving our understanding of radiation response at the cellular and tissue levels will undoubtedly yield advances for medical/therapeutic radiation as well as general cellular biology.

BIBLIOGRAPHY

1. Goodhead DT. Initial events in the cellular effects of ionizing radiations: Clustered damage in DNA. Int J Radiation Biol 1994;65(1):7–17.
2. Iliakis G. The role of DNA double strand breaks in ionizing radiation-induced killing of eukaryotic cells. Bioessays 1991;13(12):641–648.
3. Sutherland BM, et al. Clustered DNA damages induced by x rays in human cells. Radiat Res 2002;157(6):611–616.
4. Ward JF. DNA damage produced by ionizing radiation in mammalian cells: identities, mechanisms of formation, and reparability. Prog Nucleic Acid Res Mol Biol 1988;35:95–125.
5. Sutherland BM, et al. Clustered DNA damages induced in isolated DNA and in human cells by low doses of ionizing radiation. Proc Natl Acad Sci USA 2000;97(1):103–108.
6. Fernandes N, et al. DNA damage-induced association of ATM with its target proteins requires a protein interaction domain in the N terminus of ATM. J Biol Chem 2005;280(15):15158–15164.
7. Kang J, et al. Functional interaction of H2AX, NBS1, and p53 in ATM-dependent DNA damage responses and tumor suppression. Mol Cell Biol 2005;25(2):661–670.
8. Lavin MF, et al. ATM signaling and genomic stability in response to DNA damage. Mutat Res 2005;569(1–2):123–132.
9. McBride WH, et al. A sense of danger from radiation. Radiat Res 2004;162(1):1–19.
10. Li L, Zou L. Sensing, signaling, and responding to DNA damage: Organization of the checkpoint pathways in mammalian cells. J Cell Biochem 2005;94(2):298–306.
11. Ghobrial IM, Witzig TE, Adjei AA. Targeting apoptosis pathways in cancer therapy. CA Cancer J Clin 2005;55(3):178–194.
12. Pervan M, et al. Molecular pathways that modify tumor radiation response. Am J Clin Oncol 2001;24(5):481–485.
13. Nagasawa H, Little JB. Induction of sister chromatid exchanges by extremely low doses of alpha- particles. Cancer Res 1992;52(22):6394–6396.
14. Barcellos-Hoff MH. Integrative radiation carcinogenesis: Interactions between cell and tissue responses to DNA damage. Semin Cancer Biol 2005;15(2):138–148.
15. Goldberg Z, Lehnert BE. Radiation-induced effects in unirradiated cells: A review and implications in cancer. Int J Oncol 2002;21:337–349.
16. Brown JM. The effect of acute x-irradiation on the cell proliferation kinetics of induced carcinomas and their normal counterpart. Radiat Res 1970;43(3):627–653.
17. Brown JM, Berry RJ. Effects of X-irradiation on the cell population kinetics in a model tumour and normal tissue system: Implications for the treatment of human malignancies. Br J Radiol 1969;42(497):372–377.
18. Brown JM, Wilson WR. Exploiting tumour hypoxia in cancer treatment. Nat Rev Cancer 2004;4(6):437–447.
19. Fowler JF. The eighteenth Douglas Lea lecture. 40 years of radiobiology: Its impact on radiotherapy. Phys Med Biol 1984;29(2):97–113.
20. Fowler JF. Potential for increasing the differential response between tumors and normal tissues: Can proliferation rate be used? Int J Radiat Oncol Biol Phys 1986;12(4):641– 645.
21. Hayashi T, et al. Long-term effects of radiation dose on inflammatory markers in atomic bomb survivors. Am J Med 2005;118(1):83–86.

22. Preston DL, et al. Effect of recent changes in atomic bomb survivor dosimetry on cancer mortality risk estimates. Radiat Res 2004;162(4):377–389.
23. Yamada M, et al. Noncancer disease incidence in atomic bomb survivors, 1958–1998. Radiat Res 2004;161(6):622–632.
24. Loken MK, Feinendegen LE. Radiation hormesis. Its emerging significance in medical practice. Invest Radiol 1993;28(5):446–450.
25. Jaworowski Z. Hormesis: The beneficial effects of radiation. 21st Century Sci Tech 1994;7:22–27.
26. Shadley JD, Afzal V, Wolff S. Characterization of the adaptive response to ionizing radiation induced by low doses of X rays to human lymphocytes. Radiat Res 1987;111(3):511–517.
27. Liu SZ, Liu WH, Sun JB. Radiation hormesis: Its expression in the immune system. Health Phys 1987;52(5):579–583.
28. Sagan LA, Cohen JJ. Biological effects of low-dose radiation: Overview and perspective. Health Phys 1990;59(1): 11–13.
29. Schiestl RH, Khogali F, Carls N. Reversion of the mouse pink-eyed unstable mutation induced by low doses of x-rays. Science 1994;266(5190):1573–1576.
30. Goldberg Z, et al. Effects of low-dose ionizing radiation on gene expression in human skin biopsies. Int J Radiat Oncol Biol Phys 2004;58(2):567–574.
31. Rocke DM, et al. A Method for Detection of Differential Gene Expression in the Presence of Inter-Individual Variability in Response. Bioinformatics. In press.
32. Kadhim MA, et al. Transmission of chromosomal instability after plutonium alpha-particle irradiation. Nature 1992; 355(6362):738–740.
33. Kadhim MA, et al. Alpha-particle-induced chromosomal instability in human bone marrow cells. Lancet 1994; 344(8928):987–988.
34. Watson GE, et al. Chromosomal instability in unirradiated cells induced in vivo by a bystander effect of ionizing radiation. Cancer Res 2000;60(20):5608–5611.
35. Whitehouse CA, Tawn EJ. No evidence for chromosomal instability in radiation workers with in vivo exposure to plutonium. Radiat Res 2001;156(5 Pt 1):467–475.
36. Kadhim MA, Marsden SJ, Wright EG. Radiation-induced chromosomal instability in human fibroblasts: Temporal effects and the influence of radiation quality. Int J Radiat Biol 1998;73(2):143–148.
37. Goldberg Z. Clinical implications of radiation-induced genomic instability. Oncogene 2003;22(45):7011–7017.

See also CODES AND REGULATIONS: RADIATION; NONIONIZING RADIATION, BIOLOGICAL EFFECTS OF; RADIATION DOSIMETRY FOR ONCOLOGY; RADIATION PROTECTION INSTRUMENTATION; RADIATION THERAPY, QUALITY ASSURANCE IN.

ION-PAIR CHROMATOGRAPHY.

See CHROMATOGRAPHY.

ION–SENSITIVE FIELD-EFFECT TRANSISTORS

PAUL A. HAMMOND
DAVID R.S. CUMMING
University of Glasgow
Glasgow, United Kingdom

INTRODUCTION

The pH of a solution is defined as

$$\mathrm{pH} = -\log[\mathrm{H}^+] \tag{1}$$

where $[\mathrm{H}^+]$ is the concentration of hydrogen ions in the solution. The standard laboratory method of measuring the pH of a solution uses a glass electrode with a thin-walled, bulb-shaped membrane at the bottom. The electrode is filled with a standard solution (usually 0.1 *M* HCl), into which a silver wire coated with silver chloride is dipped (Fig. 1). Hydrolysis of both the inside and outside of the glass membrane forms thin gel layers, separated by dry glass inside the membrane. Hydrogen ions from the test solution are able to diffuse into the outside gel layer. The glass is doped with mobile ions (e.g., Li^+), which are able to cross the membrane and "relay" the concentration of H^+ ions in the test solution to the inside of the bulb. To make pH measurements, the potential of the internal silver wire is measured with respect to a reference electrode. Since the glass membrane is selective toward $[\mathrm{H}^+]$ ions, the overall cell potential (at room temperature) is given by the Nernst equation:

$$\psi = \mathrm{const} + \frac{kT}{q}\ln[\mathrm{H}^+] \tag{2a}$$

$$= \mathrm{const} + 25.7 \times 10^{-3} \ln 10 \log[\mathrm{H}^+] \tag{2b}$$

$$= \mathrm{const} - 0.0592\,\mathrm{pH} \tag{2c}$$

where k is the Boltzmann constant, T the absolute temperature, and q the electronic charge.

A conventional reference electrode consists of a glass tube containing a filling solution of known composition (e.g., saturated KCl). In an Ag/AgCl electrode, electrical contact is made to the filling solution by a silver wire that has been coated with silver chloride. If saturated KCl is used, the wire develops a potential of 199 mV versus the standard hydrogen electrode (SHE). This potential remains constant as long as the chloride concentration remains constant. The internal filling solution is separated from the test solution by a permeable membrane or "liquid junction", usually made from porous glass or ceramic. Diffusion of ions through the junction provides the conductive path between the reference electrode and the test solution. The KCl is usually used as the filling solution, as the mobility of the K^+ and Cl^- ions are nearly equal, minimising any build-up of junction potential.

The glass-electrode, with its complicated materials and internal reference solution, does not lend itself to miniaturization. However, with the arrival of the metal oxide semiconductor field-effect transistor (MOSFET) in 1960, another method of measuring the interface potential became available. The MOSFET is a three-terminal device in which the voltage on a gate electrode controls the current flowing between source and drain electrodes. The MOSFET has a metal or polysilicon gate electrode, separated from the bulk silicon by a thin, insulating gate oxide layer to provide an extremely high input impedance (Fig. 2a). In an n-channel MOSFET, a positive voltage applied to the gate electrode attracts electrons from the bulk of the p-type silicon to the surface beneath the oxide and creates an inversion region that is rich in mobile electrons. This inversion region forms a channel that allows current to flow between source and drain. The minimum gate voltage

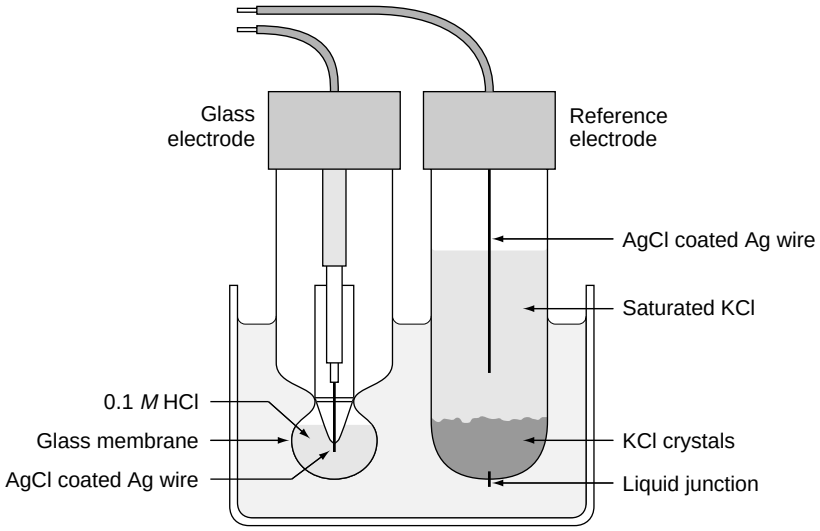


Figure 1. Diagram of the glass electrode together with a silver–silver chloride reference electrode.

that is required to produce "strong" inversion is termed the threshold voltage and is given by

$$V_T = \frac{\Phi_M - \Phi_S}{q} - \frac{Q_I}{C_I} - \frac{Q_S}{C_I} + 2\psi_F \qquad (3)$$

The threshold voltage incorporates the work function difference between the metal gate and the silicon ($\Phi_M - \Phi_S$), any fixed charge in the oxide or in the oxide–silicon interface Q_I and the charge in the depletion region of the silicon Q_S. The term $2\psi_F$ is twice the Fermi level of the p-type silicon and arises from the definition of strong inversion, that is the inverted material must have a free-electron density that is equivalent to the acceptor density in the p-type material (1). The parameter C_I is the capacitance of the oxide (insulator) layer, and all charges and capacitances are expressed per unit area.

In 1970, Bergveld (2) recognized that the MOSFET structure could be adapted to create an ion-sensitive FET (ISFET) by omitting the metal gate. He found that by applying a voltage between the source and drain of this device and placing it in solution, the current flowing would vary with the pH of the solution. This first ISFET used the native gate oxide (SiO_2) as the ion-sensing layer and was sensitive to the concentration of both Na^+ and H^+ ions in the solution (3). The silicon dioxide, used to insulate the metal gate from the silicon substrate, has a similar structure to the permselective membrane used in the glass electrode, and will therefore respond to the concentration of hydrogen ions in a solution. A reference electrode, the electrical contact of which can be regarded as the gate terminal, is used to bias the ISFET (Fig. 2b).

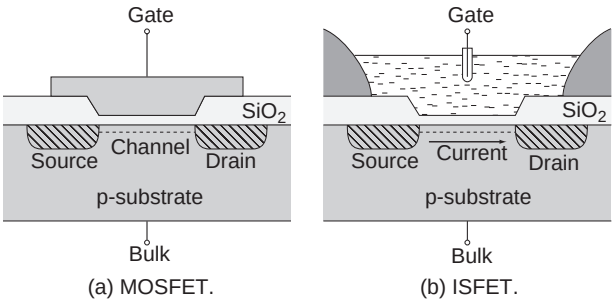


Figure 2. Diagrams showing the similarity in cross-section between a MOSFET and an ISFET.

A model for an ISFET, made by excluding the metal gate from a MOSFET, will include all the contributions to its threshold voltage considered in equation 3. There are an additional two terms, one due to the potential of the reference electrode ψ_{REF}, and the other due to the potential at the solution–oxide interface $\Delta\psi + \chi_{SOL}$. Here, χ_{SOL} is the constant surface dipole potential of the solvent (usually water), and $\Delta\psi$ is the concentration-dependent interface potential. The threshold voltage of the ISFET is then

$$V_T = \psi_{REF} - \Delta\psi + \chi_{SOL} - \frac{\Phi_S}{e} - \frac{Q_I}{C_I} - \frac{Q_S}{C_I} + 2\psi_F \qquad (4)$$

since the gate metal of the MOSFET is replaced by the metal in the reference electrode and its work function has been included in ψ_{REF}.

The only term in equation 4 that varies with ionic concentration is $\Delta\psi$, so measuring the concentration is simply a matter of measuring V_T. The ISFET is then an ideal transducer with which to measure ionic concentrations since it has an extremely high input impedance, and hence does not require any bias current to flow in the solution. It also uses the same fabrication process as a MOSFET, suggesting that not only can it be made very small, but that it can be integrated on the same substrate as the sensor electronics.

THE SITE-BINDING MODEL

Initially, it was assumed that the silicon dioxide insulator used in the ISFET would obey the Nernst equation, in the same way as the membrane used in the glass electrode did. This meant that, for a pH-ISFET, $\Delta\psi$, and hence V_T, would be a linear function of pH with a response of 59 $mV \cdot pH^{-1}$. The model for the ISFET simply substituted the Nernst equation in place of $\Delta\psi$ in equation 4 (4). While this

suggested how to operate the ISFET in a circuit, it did not provide any insight into the chemical processes that occur at the solution–oxide interface. Nor did it provide an explanation for the sub-Nernstian and pH dependent sensitivities that were measured for SiO_2 ISFETs (5).

The so-called site-binding model assumes that the insulator surface has ionizable sites that react directly with the electrolyte to bind or release hydrogen ions. The surface becomes more or less charged depending on the concentration of ions in the solution. The charged surface induces a layer of complementary charge in the solution that forms a double-layer capacitance across which the interface potential is developed. The solution side of the insulator–electrolyte interface is thought to be made up of several "layers" as shown in Fig. 3. The solvent molecules and any ions that are specifically adsorbed onto the surface make up an inner layer. The locus of the electrical centers of the adsorbed ions is called the inner Helmholtz plane (IHP) or Stern layer. The total surface charge density in this plane due to the ions is σ_0. Solvated ions in the solution cannot approach as close to the surface, and their locus of closest approach forms the outer Helmholtz plane (OHP). The interaction of the solvated ions with the surface is purely electrostatic and they are referred to as nonspecifically adsorbed ions. Because of thermal mixing in the solution, these ions are distributed throughout the diffuse layer that extends from the OHP into the bulk of the electrolyte. The excess charge density in the diffuse layer is σ_D so that the total charge in the solution is $\sigma_D + \sigma_0$.

The charge density in the semiconductor region is σ_S, so applying the requirement of charge neutrality:

$$\sigma_D + \sigma_0 + \sigma_S = 0 \tag{5}$$

If the potential in the semiconductor bulk is defined to be zero and the potential of the electrolyte bulk is fixed at ψ_{REF}, then

$$\psi_{REF} + (\psi_D - \psi_{REF}) + (\psi_0 - \psi_D) + (\psi_S - \psi_0) - \psi_S = 0 \tag{6}$$

In addition,

$$\psi_0 - \psi_S = -\frac{\sigma_S}{C_I} \tag{7}$$

$$\psi_0 - \psi_D = -\frac{\sigma_D}{C_H} \tag{8}$$

$$\psi_D - \psi_{REF} = -\frac{2kT}{e}\sinh^{-1}\frac{\sigma_D}{\sqrt{8\varepsilon kTc}} \tag{9}$$

where k is the Boltzmann constant, ϵ is the permittivity, and c is the total ionic concentration of the solution. The last equality (eq. 9) is the Gouy–Chapman model for the diffuse layer (6), and C_H is the capacitance formed between the inner and outer Helmholtz planes. Combining equations 6–9 produces:

$$\psi_{REF} + \underbrace{\frac{-2kT}{e}\sinh^{-1}\left(\frac{\sigma_D}{\sqrt{8\varepsilon kTc}}\right) - \frac{\sigma_D}{C_H}}_{\psi_{EI} \equiv -\Delta\psi} + \underbrace{\frac{\sigma_S}{C_I} - \psi_S}_{\psi_{IS}} = 0 \tag{10}$$

This equation provides a link between the interface potential $\Delta\psi$ and the charge density of the EIS system.

The "site-binding" model was developed by Yates et al. (7) to describe the interactions at a general oxide–electrolyte interface. It was later applied to the electrolyte–SiO_2–Si system by Siu and Cobbold (8) and Bousse et al. (9), whose approach is outlined here. When an SiO_2 surface is in contact with an aqueous solution, it hydrolyzes to form surface silanol (SiOH) groups. These groups are amphoteric, meaning that they can react with either an acid or a base. The acidic and basic character of a neutral SiOH site is described by the following reactions (Fig. 4) and dissociation constants. (The dissociation constant is the equilibrium constant for a reversible dissociation reaction. It expresses the amount by which the equilibrium favors the

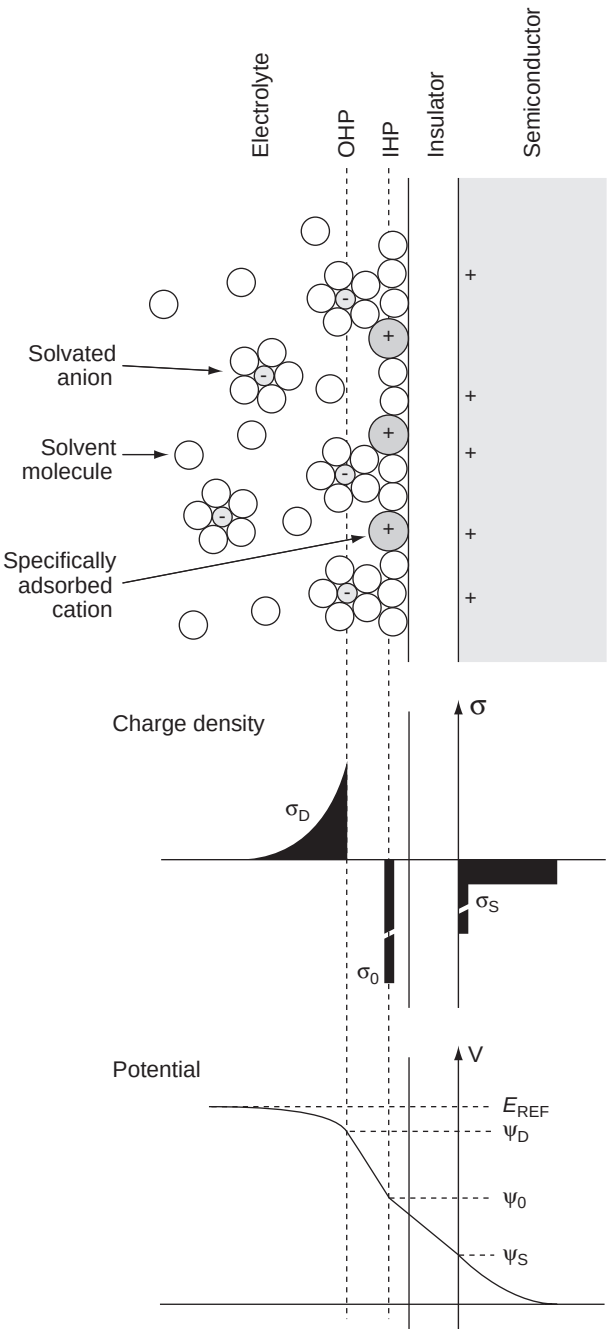


Figure 3. Model of the electrolyte–insulator–semiconductor (EIS) system.

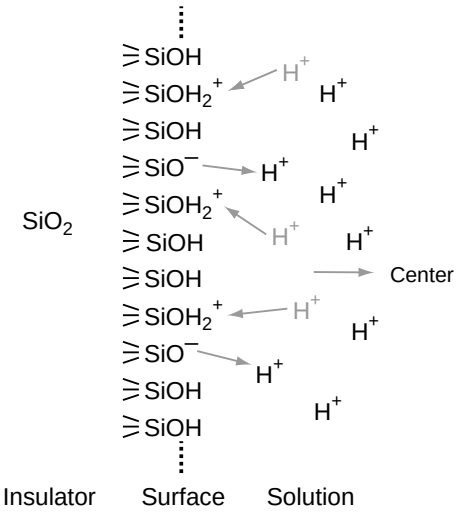


Figure 4. Diagram illustrating the surface sites and hydrogen-ion reactions for SiO_2.

products over the reactants.)

$$SiOH_2^+ \rightleftharpoons SiOH + H^+, \quad K_1 = \frac{[SiOH][H^+]_0}{[SiOH_2^+]} \tag{11}$$

$$SiOH \rightleftharpoons SiO^- + H^+, \quad K_2 = \frac{[SiO^-][H^+]_0}{[SiOH]} \tag{12}$$

where the square brackets indicate concentrations and the subscript 0 is used to indicate a surface quantity. Reactions between surface sites and other ions in the supporting electrolyte (e.g., Na^+, Cl^-) are ignored since they have been shown to have a negligible effect on the interface potential (10).

The density of sites on the surface is

$$N = [SiOH_2^+] + [SiO^-] + [SiOH] \tag{13}$$

and the surface charge per unit area is

$$\sigma_0 = e([SiOH_2^+] - [SiO^-]) \tag{14}$$

Due to thermal mixing, the surface concentration of H^+ ions can be related to the bulk H^+ concentration by Boltzmann statistics:

$$[H^+]_0 = [H^+]\exp\left(\frac{-e\Delta\psi}{kT}\right) \tag{15}$$

as $\Delta\psi$ is the potential difference from the electrolyte bulk to the insulator surface. Multiplying equation 11 by equation 12 and substituting for $[H^+]_0$ using equation 15 gives

$$[H^+] = \sqrt{K_1 K_2}\,\exp\left(\frac{e\Delta\psi}{kT}\right)\sqrt{\frac{[SiOH_2^+]}{[SiO^-]}} \tag{16}$$

For the case that $\Delta\psi = 0$ and $\sigma_0 = 0$ $\left(\text{i.e., } [SiOH_2^+] = [SiO^-]\right)$, we can see from equation 16 that $[H^+] = \sqrt{K_1 K_2}$. This is the hydrogen ion concentration in the solution required to produce an electrically neutral surface, and is called the point of zero charge (pzc). The pH at this point is denoted pH(pzc) and this can be substituted in 16 to yield

$$2.303(\text{pH(pzc)} - \text{pH}) = \frac{e\Delta\psi}{kT} + \ln\mathcal{F} \tag{17}$$

This equation provides the link between charge, potential, and pH. The function

$$\mathcal{F} = \sqrt{[SiOH_2^+]/[SiO^-]} \tag{18}$$

plays a key role in the response of the surface. It can be written in terms of the "normalized" net charge on the surface $\hat{\sigma}_0 = \sigma_0/eN$, and the parameter $\delta = 2\sqrt{K_2/K_1}$:

$$\mathcal{F} = \frac{\hat{\sigma}_0/\delta + \sqrt{(\hat{\sigma}_0/\delta)^2(1-\delta^2)+1}}{1-\hat{\sigma}_0} \tag{19}$$

Equations 17 and 19 give the solution pH as a function of both $\Delta\psi$ and σ_0, so now it remains to find the relationship between $\Delta\psi$ and σ_0. This is done by using the definition of $\Delta\psi$ in equation 10, the charge neutrality condition in equation 5:

$$\sigma_D + \sigma_0 = \Delta\sigma = -\sigma_S \tag{20}$$

It is usually assumed that $\Delta\sigma = 0$ (9), so that $\sigma_0 = -\sigma_D$. Finally, the interface potential $\Delta\psi$ can be found as a function of the solution pH, by using a parametric method in $\hat{\sigma}_0$. Values of the dissociation constants and surface site density of SiO_2 obtained from the literature are shown in Table 1, and used to generate the SiO_2 pH response curve in Fig. 5. Not only does the SiO_2 surface have a low sensitivity of -46.3 mV/pH (at pH 7), it also has a nonlinear response, especially in the acid pH range.

Table 1. Values for the Parameters of pH Sensitive Insulators Found in the Literature

	K_1	K_2	N_A (sites·m^{-2})	Reference
SiO_2	$10^{1.8}$	$10^{-6.2}$	5×10^{18}	11
Al_2O_3	10^{-6}	10^{-10}	8×10^{18}	12
Ta_2O_3	10^{-2}	10^{-4}	10×10^{18}	12

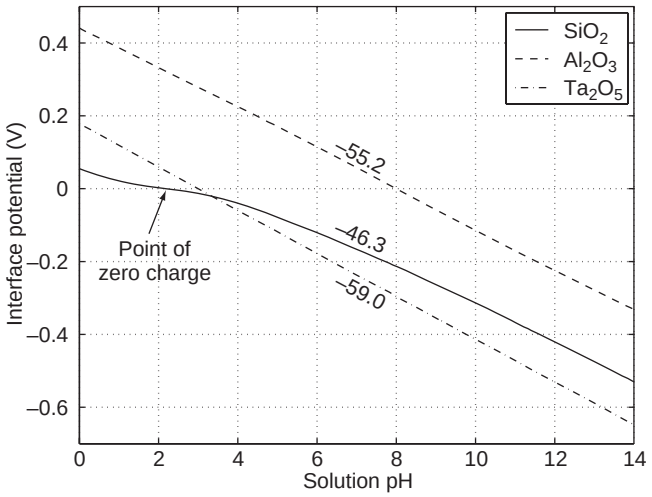


Figure 5. Graph of the theoretical pH response for SiO_2, Al_2O_3, and Ta_2O_5 surfaces.

The insulators Al_2O_3 and Ta_2O_5 also have amphoteric surface groups, so can be modelled using the same parametric method. As shown in Fig. 5, these surfaces produce a linear response with sensitivities much closer to the Nernstian ideal of -59.2 mV·pH^{-1}. As a result, Al_2O_3 and Ta_2O_5 have been widely used as the pH sensitive layer for fabricating ISFETs. The measured sensitivities for these insulators (53–57 mV·pH^{-1} for Al_2O_3, 56–57 mV·pH^{-1} for Ta_2O_5; see Table 2) are close to the theoretical values shown in Fig. 5.

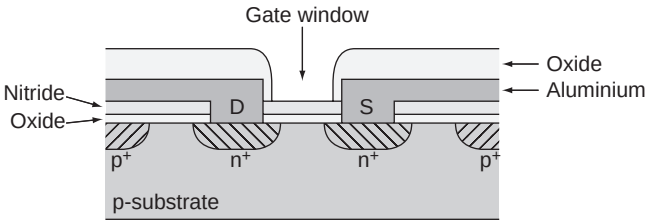


Figure 6. Diagram of the cross-section through an ISFET formed by depositing silicon nitride directly on top of the gate oxide (13).

DEVELOPMENT OF THE ISFET

Much of the subsequent work on ISFETs has concentrated on measuring pH, as this plays a vital role in many biochemical systems. Because silicon dioxide has a low pH sensitivity, the magnitude of which varies with pH (5), Matsuo and Wise (13) experimented with the use of silicon nitride (Si_3N_4) instead. Their ISFET was made by depositing a layer of nitride on top of the thermally grown gate oxide. The ISFET was located at the tip of a needle-shaped probe, which was covered in a thick layer of SiO_2 to insulate it from the solution (Fig. 6). The ISFET was found to have an almost ideal pH response and very low sensitivity to sodium and potassium ion concentrations. Selectivity between ion species is important to differentiate changes in pH from changes in the total ionic concentration of the solution.

In a sense, silicon nitride was an obvious material to investigate as it was, and still is, widely used as a passivation layer to protect devices, such as integrated circuits from the ingress of moisture. Other well-known materials included the oxides of aluminium and tantalum (Al_2O_3 and Ta_2O_5). The pH sensitivity, ion selectivity, response times, and drift rates for these materials have been extensively studied (5). Silicon oxynitride (SiO_xN_y) and other more exotic insulators, such as zirconium and tin oxides (ZrO_2, SnO_2), and even diamond-like carbon (DLC) have all been used to make pH ISFETs. Oxides and nitrides of metals have also been investigated to try and improve parameters, such as response time and maximum operating temperature. These include platinum oxide (PtO_2), titanium nitride (TiN), iridium oxide (Ir_2O_3), and indium tin oxide (ITO). The published pH sensitivity and drift rates (where available) for these materials are summarized in Table 2.

Apart from the choice of pH sensitive material, development of the ISFET has mostly focused on achieving compatibility with the CMOS process. CMOS devices use complementary pairs of n-type and p-type MOSFETs to implement circuits. The CMOS process is the dominant technology for integrated circuits (ICs), so achieving compatibility would allow complex devices containing ISFETs to be fabricated by a standard, industrial process. Figure 7 is a simplified cross-section of a CMOS invertor, showing the polysilicon gate electrodes, the source and drain regions and the metal interconnections.

ISFETs have been made using both NMOS and PMOS transistors. However, in a p-substrate process, the bulk terminal of a PMOS ISFET can be biased above the substrate ground potential, permitting more flexibility in circuit design. The reverse is true for an n-substrate CMOS process. It has also been shown that the noise performance of an ISFET is dominated by $1/f$ or "flicker" noise (23), which is lower in PMOS transistors (24).

Initial attempts to integrate ISFETs involved significant modifications to the standard CMOS process. Wong and White (15) followed the standard sequence of CMOS process steps, until the metallization stage. They then removed the oxide, the polysilicon gate electrode, and the gate oxide above the ISFETs by wet etching. A thinner

Table 2. Values of pH Sensitivity and Drift Rates for ISFETs Made Using a Arange of Materials, Obtained from the Literature

Material	pH Range	Sensitivity, (mV·pH^{-1})	Drift, (mV·h^{-1})	Reference
SiO_2	4–10	25–35 (pH $<$ 7) 37–48 (pH $>$ 7)	Unstable	5
SiO_xN_y	2–8.3	57.4 $\pm$ 0.4	0.8 (pH 7)	14
	4–9	18–20	Not mentioned	15
PtO_2	1–10	40.5 $\pm$ 4.0	0.5 (pH 6.86)	16
Si_3N_4	1–13	45–56	1.0 (pH 7)	5
ZrO_2	2–10	50	Slow response	17
Al_2O_3	1–13	53–57	0.1–0.2 (pH 7)	5
Ta_2O_5	1–13	56–57	0.1–0.2 (pH 7)	5
DLC	1–12	54–59	3 μV/h (pH 3)	18
SnO_2	2–10	55–58	Not mentioned	19
ITO	2–12	58	Not mentioned	20
TiN	1.68–10.01	59	$<$ 1	21
Ir_2O_3	3–10	59.5	Unclear	22

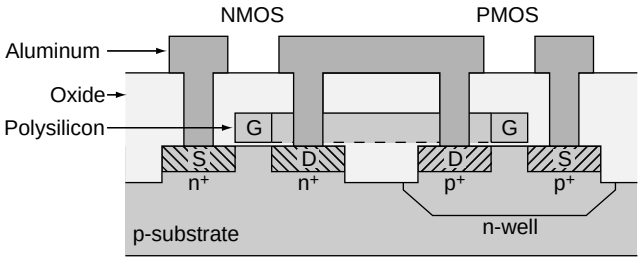


Figure 7. Diagram of the cross-section through a CMOS invertor.

oxide layer was regrown and the sensing layer of Si_3N_4 or Ta_2O_5 was deposited on top of this. The contact windows were then opened and the aluminum deposited to form the interconnects as for a normal CMOS process.

Bousse et al. (25) took a similar approach, but completed the CMOS process before etching away the deposited SiO_2 above the polysilicon gate of the ISFET. They then deposited Si_3N_4 over the whole wafer so that the polysilicon was retained as a floating electrode in the ISFET. This floating electrode did not reduce the ISFET sensitivity to pH, and had the additional benefit of making the ISFET less sensitive to changes in light levels. It does this by shielding the channel from photons that can generate electron–hole pairs, which contribute to the ISFET current. However, since the nitride they used was deposited by low pressure chemical vapor deposition (LPCVD) at 785 °C, the aluminium interconnect had to be replaced with tungsten silicide, which was able to withstand this high temperature step. This meant that a specially modified CMOS process had to be used to fabricate the ISFETs.

Bausells et al. (26) extended the floating electrode idea to a two-metal CMOS process by connecting the polysilicon gate and both metal layers together. In addition, they used the silicon oxynitride passivation layer as the pH sensitive material for the ISFET (Fig. 8). This meant that the ISFET could be fabricated by a commercial foundry using a standard CMOS process, without the need for any process modifications. The fabricated ISFETs had a sensitivity of 47 $mV \cdot pH^{-1}$ and a lifetime of > 2 months. As well as the advantages of using of a well-characterized industrial process, there are additional "system-on-chip" design benefits. These include access to libraries of components such as amplifiers and digital gates that greatly ease the design of the whole system.

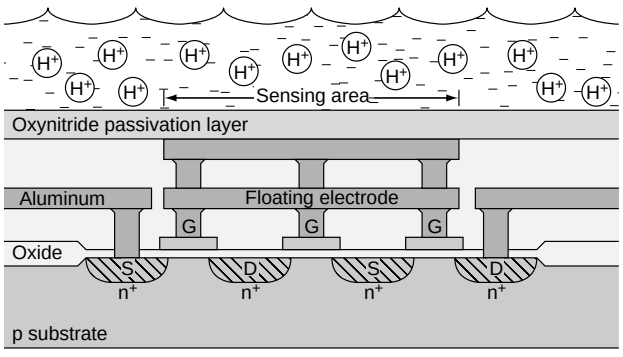


Figure 8. Diagram of the cross-section through an ISFET fabricated by an unmodified commercial CMOS process (26). The floating electrode allows the passivation layer to be used as a pH sensitive insulator.

Unfortunately, the unmodified CMOS ISFETs were found to have large threshold voltages; they were also based on silicon oxynitride, a material that has been found to have a widely varying sensitivity, depending on the deposition conditions (Table 2). The large threshold voltage has been shown to be caused by trapped charge left on the floating electrode during fabrication (27). To avoid these problems, Jakobson et al. (28) removed the passivation layer by using the aluminum of the floating electrode as an etch-stop. They then experimented with low temperature evaporation of pH sensitive layers onto the exposed aluminium. The best performance was obtained by using a layer of platinum (to prevent the aluminum from corroding), followed by a layer of tantalum oxide.

The floating-electrode ISFETs can be considered as special cases of the extended-gate ISFET that was first proposed in 1983. The idea was to separate the electronics from the chemically active region, and by doing so make the device easier to passivate and package than a standard ISFET with an exposed gate insulator. A coaxial polysilicon structure was used to provide a screened connection between the chemically sensitive area and the gate of a MOSFET (29). A more recent CMOS extended-gate ISFET design used a long (unscreened) aluminium track with one end connected to the polysilicon gate and the other exposed by the final pad-etch step (21). This idea has been taken to its limit by using a discrete, off-the-shelf MOSFET and connecting the gate terminal to the pH-sensitive material with a length of wire (20). This method is clearly not applicable to a sensor system-on-chip design, but it does provide a simple method of characterizing the behavior of the material.

The floating-electrode ISFET has also been used to protect against electrostatic discharge (ESD). In the first ESD-protected devices, the polysilicon gate was left intact and connected to a terminal via a MOSFET switch. This provided a reverse-biased diode between the floating electrode and the substrate that would breakdown (reversibly) before the gate insulator was damaged (30). However, current leakage through the "off" MOSFET was such that any response to changing pH decayed to zero in a matter of seconds. To achieve a steady-state response, a large platinum electrode was connected to the ISFET gate to supply current from the solution to replace that being lost though the MOSFET. To avoid the problem of leakage current altogether, ESD-protected ISFETs have been fabricated with a separate platinum ring electrode around the sensitive gate area (31). The platinum electrode is a preferential discharge path to the substrate, protecting the ISFET in the same manner that a lightning conductor protects a building.

ISFETs have also been adapted to create chemically modified FETs (CHEMFETs), which are sensitive to the concentration of ions other than hydrogen. This is achieved by attaching a polymer membrane containing a suitable ionophore to the pH sensing surface of the ISFET. The stability of the ISFET–membrane interface is improved by the addition of an intermediate hydrogel layer. In this way, CHEMFETs sensitive to K^+ (32), Na^+ (33), and other cations have been developed.

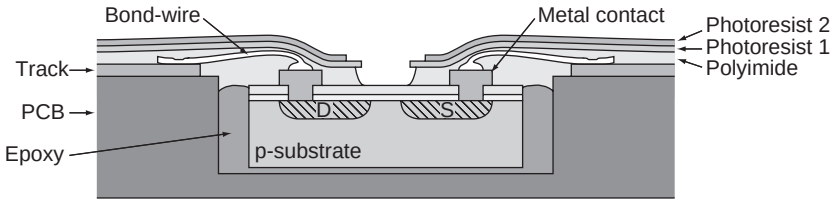


Figure 9. Diagram of the cross-section through an ISFET-based device in a recessed PCB, encapsulated using a layer of polyimide and two layers of photoresist (34).

PACKAGING

One of the main obstacles that has prevented the commercialization of ISFET based devices is the repeatability and reliability of the encapsulation procedure. It is normal for the encapsulant to be applied by hand, covering the chip and bond wires, but leaving a small opening above the sensing area. Epoxy is the most extensively used material although it is important to select a composition that is stable, a good electrical insulator, and does not flow during encapsulation. Many commercially available epoxies have been assessed for their suitability by making measurements of their electrical impedence over time (34–37).

By using ultraviolet (UV) curable polymers, it is possible to increase the automation of the packaging process using a standard mask aligner. A lift-off technique was developed using a sacrificial layer of photosensitive polyimide to protect the ISFET gates. Alumina-filled epoxy was applied by screen printing and partially cured, before the polyimide was etched away leaving a well in the epoxy (38). After 10 days in solution, leakage currents of 200 nA were observed. Better results were achieved by direct photopolymerization of an epoxy-based encapsulant. The ISFETs packaged using this method showed leakage currents of 35 nA after 3 months in solution (38). To avoid polarizing the reference electrode, a leakage current of < 1 nA is desirable (5). This photolithographic patterning of the encapsulant was done at the wafer-level, to all the devices simultaneously. Subsequently, the wafer was diced up and the individual chips were wire-bonded and coated with more encapsulant by hand. At the chip-level, wire-bonded ISFET chips have been covered (again by hand) with a 0.5–1 mm thick photosensitive, epoxy-based film, then exposed and developed (39).

Some degree of automation was introduced by Sibbald et al. (34) who used a dip-coating method to apply the polymers. They first recessed the chip into a PCB and wire-bonded the connections, before coating it with a layer of polyimide. Two layers of photoresist followed, before the underlying polyimide was etched away (Fig. 9). The packaged devices showed < 10 pA leakage current after 10 days in solution. However, the encapsulation did exhibit electrical breakdown for applied bias voltages in excess of 1.5–2 V, which was attributed to the high electric field in the thin layer of resist covering the bond wires. In a separate study, photosensitive polyimide has also been used to create the wells that separate the ion-selective membranes on a multiple ISFET chip (40).

The structure of the ISFET has also been modified to improve the lifetime and ease of manufacture of the packaged device. One solution was to make the ISFET chip long and thin (1.2×12 mm) with the sensing area at one end and the bond pads at the other so that the bond-wires did not enter the solution (14). The chip itself was encapsulated with a thick layer of silica. More radical solutions involved bulk micromachining to form back-side contacts to the ISFET so that the bond wires were on the opposite side of the chip to the solution. The front side of the chip is protected by anodic bonding of glass (Fig. 10). A review of back-side contact ISFETs is provided by Cané et al. (41), but the technique is not particularly suited to a CMOS chips, which have many bond-pads arranged around the perimeter.

ISFET CIRCUITS

When an ISFET is placed in solution, a concentration-dependent potential ($\Delta\phi$) is formed at the interface

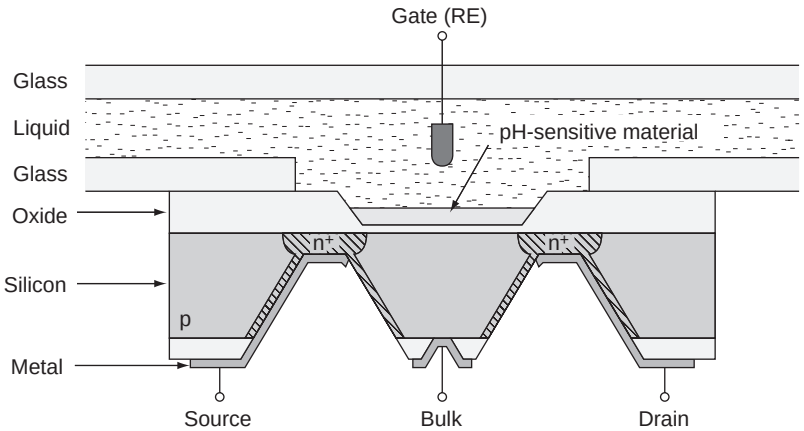
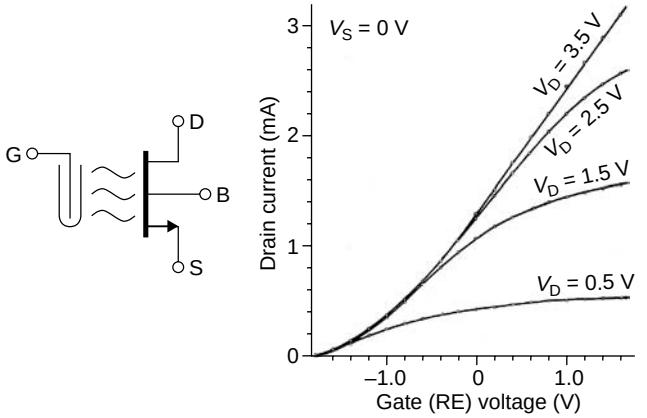


Figure 10. Diagram of the cross-section through a back-side contacted ISFET chip with liquid and electrical contacts on opposite sides (41).



(a) Symbol. (b) *I–V* characteristics.

Figure 11. Diagram of the circuit symbol and a graph of the *I–V* transfer characteristics (4) for an n-type ISFET.

between the gate insulator and the solution. This potential modulates the threshold voltage (V_T) of the ISFET (eq. 4), an internal quantity that cannot be directly measured. Some circuitry is therefore required to convert changes in the threshold voltage into changes of a measurable signal. The solution to this problem lies in the use of feedback control to maintain a constant drain current in the ISFET. For a FET, operating in the linear region[3], the drain current is given by:

$$I_D = k' \frac{W}{L} \left[(V_{GS} - V_T) V_{DS} - \frac{V_{DS}^2}{2} \right] \qquad (21)$$

where k' is a process-dependent constant, and W, L are the width, length of the device. (In the linear region, $0 < V_{DS} \leq (V_{GS} - V_T)$ and I_D varies with V_{DS}.) The parameters V_{GS} and V_{DS} are, respectively, the gate-source and drain-source voltages applied to the FET. For an ISFET, a reference electrode in the solution acts as the gate terminal as shown symbolically in Fig. 11a. The I_D vs. V_{GS} curves for an ISFET as measured by Moss et al. (4) are shown in Fig. 11b. It is clear from this graph that biasing an ISFET at a constant drain current is only possible if both V_{DS} and V_{GS} ($= V_G - V_S$) are maintained constant. If a reference electrode is used to control V_G, then from equation 21, as V_T changes with I_D held constant, V_S must change by an equal and opposite amount to compensate. Measuring $\Delta\psi$ (and hence pH) then becomes a straightforward matter of measuring the terminal voltage V_S.

In his original paper on the operation of the ISFET, Bergveld (3) stated that one of the important advantages of ISFETs, compared with conventional pH electrodes, is that there is no need for a reference electrode. Instead, he used a feedback circuit to control the bulk terminal of the ISFET and maintain a constant drain current. However, this makes the assumption that the solution is perfectly isolated from the ISFET source and drain terminals, as well as the circuit. Any current (even leakage current through the packaging), that flows into the solution will affect its potential. To a bulk-feedback circuit, this will be indistinguishable from a change in solution concentration. It is therefore safer to assume that the solution is grounded, or at least at some well-defined potential with respect to the circuit, and use a reference electrode to ensure that this is the case. For this reason, all of the subsequently published circuits relating to ISFETs include a reference electrode.

The probe fabricated by Matsuo and Wise (13) contained an ISFET and a MOSFET of identical dimensions. The ISFET was configured as a source follower with the MOSFET acting as a constant current source. A saturated calomel electrode (SCE, shown in Fig. 1) was used as a reference, and the output voltage measured at the source terminal of the ISFET. Bergveld (42) used a grounded reference electrode to avoid the problem of a short circuit if the solution is already at ground potential (e.g., in an earthed metal container). He used an instrumentation amplifier arrangement to maintain a constant current at a constant drain-source voltage (Fig. 12). Amplifiers A_1 and A_2 set V_{DS} as determined by the fixed current flowing in R_1. Current feedback from amplifier A_4 adjusts the drain voltage via R_2 to keep the current constant as the threshold voltage changes. One disadvantage of this circuit is that the output (measured across R_9) is not referenced to a fixed voltage such as ground.

There have been many other circuit topologies proposed to keep the drain current and/or the drain-source voltage constant. A straightforward circuit that achieves both of

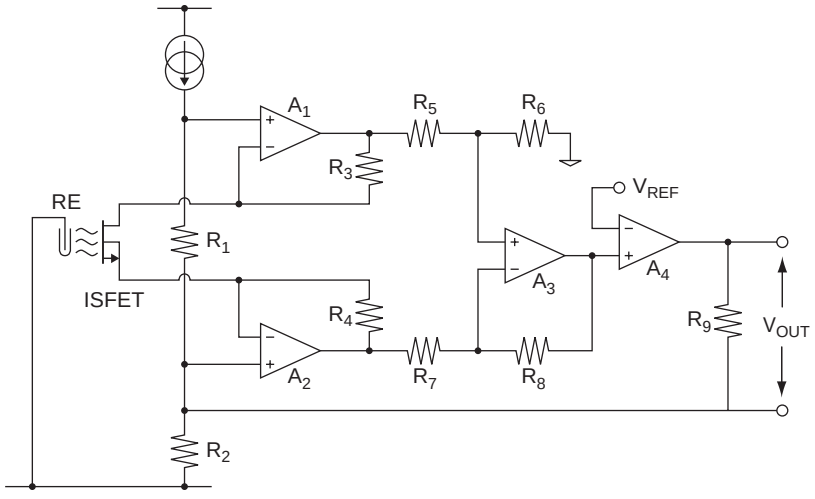


Figure 12. Circuit diagram of an ISFET source and drain follower circuit used to maintain a constant drain current at a constant drain-source voltage (42).

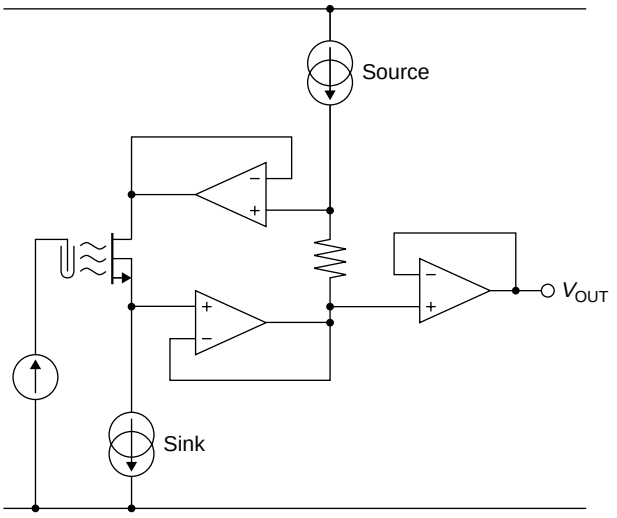


Figure 13. Circuit diagram of constant drain current, constant drain–source voltage ISFET bias circuit (43).

these objectives was presented by Ravezzi and Conci (43). It used a pair of unity-gain, noninverting operational amplifiers (op-amps) to ensure that the voltage dropped across a resistor is also dropped across the ISFET (Fig. 13). Constant current was maintained in the resistor by a current source, and in the ISFET by a current sink. This arrangement allows the source and drain potentials to adjust as the threshold voltage changes, allowing the reference electrode to remain at a fixed potential.

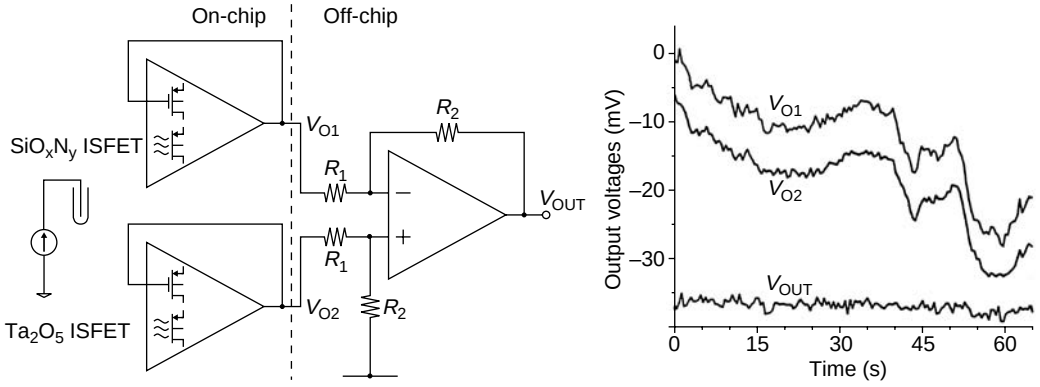
The first integrated ISFET circuit was the so-called operational transducer published by Sibbald (44) in 1985. It used an ISFET and a MOSFET with identical geometries as the active devices in the "long-tailed pair" input stage of an amplifier. Feedback was used to control the gate voltage of the MOSFET to ensure that the same current flowed in both devices. The MOSFET gate voltage tracked that of the ISFET and so measured the changes in threshold voltage. The key advantage of this circuit was that changes in temperature affected both devices equally and were canceled out.

Wong and White (15) recognized that there was little to be gained from an integrated, miniaturized sensor if it relied on a large, external reference electrode. Instead, they used an on-chip gold contact as a quasi-reference electrode (qRE) and a differential circuit to make measurements between two ISFETs with different pH sensitivities. The potential difference between the solution and the qRE will depend on the solution composition. However, like temperature, this is a common-mode signal, which will affect both ISFETs equally. Hence, it can be rejected by means of a differential circuit. Tantalum oxide and silicon oxynitride were used as sensing layers for the two ISFETs, which formed the input stages of op-amps integrated onto a CMOS chip (Fig. 14a). The outputs from the two op-amps were fed into an off-chip differential amplifier. The overall circuit gave a response of 40–43 $mV \cdot pH^{-1}$. The benefit of the differential approach can be seen in Fig. 14b, which shows the single-ended (V_{O1} and V_{O2}) and differential (V_{OUT}) responses as electrical noise was deliberately applied to the solution. Two copies of the bias circuit in Fig. 13 have also been used to create a differential system with Si_3N_4 and SiO_2 ISFETs (45).

The concept of integration has been extended by Hammond et al. (46) who created a complete digital pH meter on a single CMOS chip. This design makes use of the libraries of components provided by the CMOS foundry to integrate not only the ISFET, but also analog bias circuits, digital signal processing, and storage onto the same chip (Fig. 15a). The chip was mounted in a recessed PCB and covered with a thick layer of photoresist so that only the ISFET area was exposed. The digital response of the device to the changing pH of the solution in which it is immersed is shown in Fig. 15b.

MINIATURE REFERENCE ELECTRODES

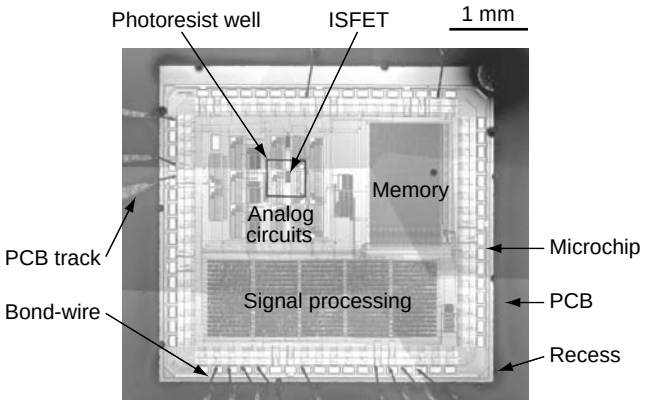
The first attempt to incorporate a reference electrode on an ISFET chip used a thin-film Ag/AgCl electrode (47). Electrodes like this, with no internal reference solution, are sensitive to changes in concentration of their primary ion (in this case Cl^-), and are referred to as quasi-reference electrodes. To solve this problem, Smith and Scott (48) also



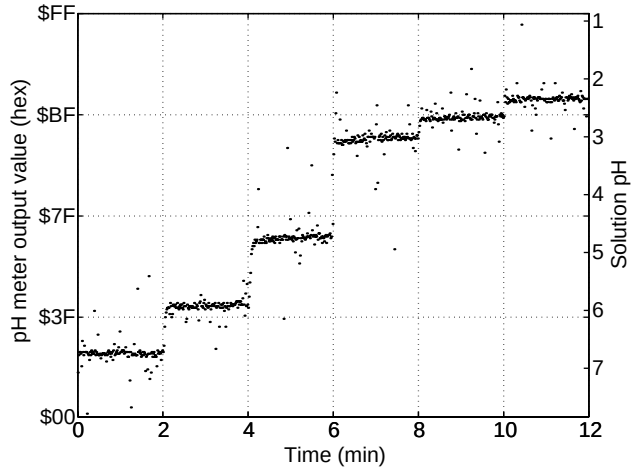
(a) Circuit diagram.

(b) Output voltage.

Figure 14. Circuit diagram and graph of output voltage for a differential integrated amplifier based on Ta_2O_5 and SiO_xN_y ISFETs (15).



(a) Chip layout



(b) pH response

Figure 15. Chip layout and pH response of a single-chip digital pH meter.

integrated the reference solution and porous membrane into the chip. Wells were etched into the back-side of a silicon wafer, to leave a membrane 10–70 μm thick. The membranes were anodized to porous silicon in a bath of concentrated hydrofluoric acid. The wells were filled with saturated KCl and sealed with glass coverslips that had been coated with thin films of Ag/AgCl. The reference electrode exhibited a low drift rate of 80 $\mu V \cdot h^{-1}$ (worst-case) and a lifetime of > 2 weeks. However, a method of mass producing an integrated, liquid-filled electrode has yet to be developed.

Recent developments of miniature reference electrodes have focused on the use of polymer-supported electrolyte gels, to replace the liquid filling solution. Suzuki et al. (49) developed an electrode that uses finely ground KCl powder supported in a matrix of poly(vinylpyrrolidone) (PVP). An exploded diagram of the electrode is shown in Fig. 16. First, a layer of silver was evaporated onto a glass substrate, inside a U-shaped gold backbone. A layer of polyimide was applied to protect the silver and to define the electrode structure. The AgCl was grown through a slit in the polyimide, and a liquid junction was formed by casting a hydrophilic polymer into the square recess. The electrolyte layer, containing the KCl powder, was then screen-printed over the AgCl and the liquid junction. Finally, a passivating layer of silicone rubber was applied. The electrode can be stored dry, and activated when required by the injection of a saturated solution of KCl and AgCl through the silicone. This miniature reference electrode showed a stability of ±1.0 mV over a period of 100 h. No difference was observed between experimental data obtained with the miniature reference electrode and with a large, commercial reference electrode.

REFERENCE FETs

The sensitivity of the differential circuits already discussed can be increased if one of the devices has no response to

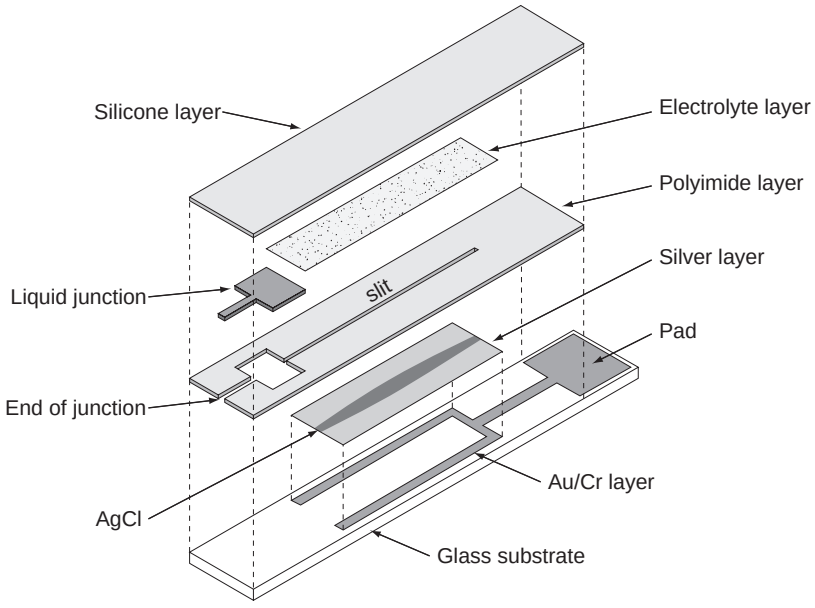


Figure 16. Exploded diagram showing the construction of a miniature Ag/AgCl reference electrode based on a polymer-supported electrolyte gel (49).

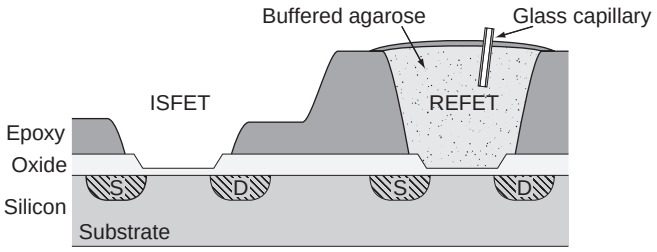


Figure 17. Diagram of the cross-section through a REFET created using encapsulated pH buffer solution (50).

changes in pH. Such a device is called a reference FET (REFET). The first REFET was presented by Comte and Janata (50) in 1978. It consisted of an ISFET surrounded by a well of epoxy that was filled with a buffered agarose gel. A glass capillary was inserted into the gel and sealed in place with more epoxy (Fig. 17). This acted as a liquid junction between the internal reference gel and the external solution. The response of the REFET was only 3 $mV \cdot pH^{-1}$ and it provided temperature compensation of ± 0.01 $pH \cdot C^{-1}$ when used in a differential arrangement. However, the techniques required to prepare this REFET are not well suited to mass production.

The pH sensitivity of an ISFET is due to the presence of chemical sites on the surface of the insulator (Fig. 4) that can exchange hydrogen ions with the solution. Attempts to reduce the density of these sites, and hence the sensitivity, by chemical modification of the surface proved unsuccessful (51). Instead a thin membrane of parylene was used to cover an ISFET and convert it into a REFET (52). Parylene has an extremely low density of surface sites and the REFET was found to have a very low pH sensitivity of only 0.5 $mV \cdot pH^{-1}$. However, parylene forms an insulating (or ion-blocked) layer that affects the electrical properties of the underlying ISFET. Even a very thin membrane reduces the gate capacitance, and hence transconductance, dramatically. This is a problem for differential circuits, which rely on ISFET and REFET having well-matched electrical properties. If a membrane could be found whose ion-conducting properties were sufficient to pass the electrical voltage of the solution to the sensing layer of the underlying ISFET, the transconductance would not be changed. Clearly, such "ion-unblocking" membranes must be insensitive to variations in pH.

Bergveld et al. (53) investigated several candidate polymers for REFET membranes and found that polyacrylate gave the best performance. An intermediate layer of buffered poly(hydroxyethyl methacrylate) (p-HEMA) hydrogel was necessary to fix the interface potential and to improve the adhesion of the membrane (54). The REFET showed < 2 $mV \cdot pH^{-1}$ sensitivity, good mechanical properties, and its transconductance matched that of the ISFET. Despite these useful properties, the acrylate membrane was selectively permeable for cations (e.g., potassium ions). This problem was solved by optimizing the amount of didodecyldimethylammonium bromide (DDMAB) added to the membrane (Fig. 18a). This large immobile cation repels mobile cations (e.g., potassium), from the membrane. Figure 18b shows the performance of the ISFET–REFET

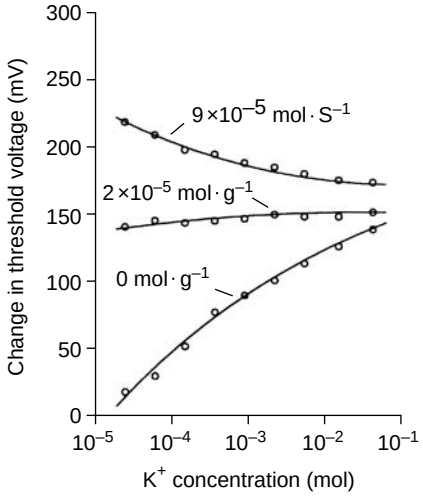


Figure 18. Graphs of the response of an acrylate REFET, and an ISFET, to changes in potassium and hydrogen ion concentrations (53).

differential system, as compared to the individual devices, measured using a platinum qRE.

An alternative approach was developed by van den Vlekkert et al. (55) who used a thick layer of p-HEMA hydrogel to produce an ISFET with a retarded pH response. By controlling the diffusion coefficient of hydrogen ions, a pseudo-REFET with a nonresponse time of ~ 10 min was created. Such a REFET is only really useful in a flow-through system where the analyte is pumped through in short bursts followed by rinsing and calibration solutions (34). The p-HEMA hydrogel was also used by Chudy et al. (56) as the base layer for chemically sensitive FETs (CHEMFETs) and REFETs. The CHEMFET uses an additional ion-selective membrane containing an ionophore, selected to target the ion of interest. Exclusion of the ionophore from the membrane enabled the creation of a REFET that showed almost no response to pH, potassium, or calcium ions.

APPLICATIONS

According to Bergveld (57), ~ 20 companies have commercialised ISFETs based on 150 patents. Product offerings and specifications vary, but in general terms ISFETs have found applications where small size, fast response, robustness, wide operating temperature range, and operation in nonaqueous environments are desirable. The ISFET can also be stored dry, making it more convenient to use than traditional glass electrodes, since the lifetime of a stored ISFET is almost indefinite. ISFETs are used for high resolution applications and products are available with specified resolutions < 0.01 pH units. However, many of the devices on the market have a resolution of ~ 0.1 pH units.

The ISFET has made a significant impact on a number of industries where it is a requirement to monitor and control the addition of reagents. It made an early appearance in the food industry where its robust construction allowed measurement to be made in foodstuffs. In addition,

its ability to operate in nonaqueous environments is an advantage when working with meat and dairy products. The ISFET has also found applications in the pharmaceutical industry, where a fast re-sponse and high operating temperatures are required. It has also been used in electroplating where it is able to withstand the corrosive plating solutions. In addition to manufacturing industries, the ISFET is also used in waste water management and in the treatment of effluent.

Early researchers considered medical applications ranging from monitoring the pH in the mouth (particularly in the context of tooth decay), to direct measurement of ionic balance in the blood. ISFETs been built into dental prosthetics to enable the direct measurement of pH in the presence of plaque (58). Recent data on tooth decay was obtained using a pH imaging microscope (59) although, unlike the ISFET device, this does not allow for *in situ* observations to be made. ISFETs are also used indirectly for dental applications, for example, in the evaluation of potential prosthetic materials (60). The ISFET has been modified in a manner similar to that for the REFET for use in blood analysis beyond measuring ionic concentrations. In this embodiment, the ISFET can be made into a sensitive enzyme sensor. In their paper, Sant et al. (61) demonstrated a sensor for creatinine based on an ISFET with a sensitivity of 30 mV·pCreatinine^{-1}. The application is in renal failure and haemodialysis in particular.

Recently, there has been a growth of interest in the use of ISFETs in a device known as a diagnostic pill. The concept of a diagnostic pill was developed in the 1950s and 1960s (62). Such devices are small enough to be swallowed by a patient, and once inside the gastrointestinal tract, can measure and wirelessly transmit data to an external receiver over a period of many hours or even days. The earliest devices were used to measure pressure changes in the gut, but glass-electrode-based pills were also made to measure gut pH. The diagnostic pill has made something of a comeback in recent years, particularly with the invention of the video pill (63), a device that is capable of wirelessly transmitting images of the gut with considerably less patient discomfort than would be caused by an endoscopic procedure. Various new examples of pH measuring pills have also been developed. The Bravo capsule, which does not actually use an ISFET, is designed for use in the esophagus (64). During operation it is pinned to the lining of the esophagus so that it can monitor conditions such as gastro-esophageal reflux disease (GERD) over a period of 2–3 days. The IDEAS capsule (65) that does use an ISFET, has been built for use in the lower gastrointestinal tract. The device is designed pass naturally through the gut with as little intervention as possible and is required to be able to operate in difficult conditions where biofouling of the sensor is a potential problem.

CONCLUSIONS

The ISFET first appeared in 1970 as an offshoot of the rapidly growing capability of the microelectronics industry. In its simplest form, the ISFET is a modification of the traditional MOSFET in which the gate electrode is removed and the gate oxide exposed to solution. The ISFET uses changes in the surface chemistry of the gate oxide to modify the threshold voltage of the transistor and produce an electronic signal. The sensitivity and performance of the ISFET is highly dependent on the material used to form the gate oxide layer. The behavior of such materials is best modeled using a site-binding approach to calculate the variation in surface charge with solution pH. A wide variety of materials have been experimented with, those found to give the best performance are the oxides of aluminium and tantalum. However, in a standard CMOS manufacturing process, the passivation layer is made of silicon nitride, which is also a good pH sensitive material. It is therefore possible to make good ISFETs using a standard foundry process with little or no additional effort. As a result it has become possible to implement integrated ISFET circuits. A number of circuit topologies for detecting the change in ISFET threshold voltage have been designed, those using voltage followers to maintain the ISFET bias conditions produce the best results. Further development of ISFET integrated circuits has enabled complete instruments to be fabricated on a single IC. To avoid the use of a bulky reference electrode, differential circuits using matched pairs of ISFET and REFET have been designed. However, difficulties in creating a good REFET have increased interest in developing miniature reference electrodes that are compatible with IC processing. The ISFETs have already found widespread application in manufacturing industries, environmental monitoring and medicine. It is expected that with improved miniaturization, integration, and packaging technologies, new applications will emerge.

BIBLIOGRAPHY

1. Streetman BG, Banerjee S. Solid State Electronic Devices. 5th ed. New Jersey: Prentice Hall; 2000.
2. Bergveld P. Development of an ion-sensitive solid-state device for neurophysiological measurements. IEEE Trans Biomed Eng 1970;BM17:70.
3. Bergveld P. Development, operation, and application of ion-sensitive field effect transistor as a tool for electrophysiology. IEEE Trans Biomed Eng 1972;BM19:342.
4. Moss SD, Johnson CC, Janata J. Hydrogen, calcium, and potassium ion-sensitive FET transducers—preliminary-report. IEEE Trans Biomed Eng 1978;25:49–54.
5. Matsuo T, Esashi M. Methods of ISFET fabrication. Sens Actuator 1981;1:77–96.
6. Bard AJ, Faulkner LR. Electrochemical Methods—Fundamentals and Applications. Hoboken (NJ), John Wiley; 2001.
7. Yates DE, Levine S, Healy TW. Site-binding model of electrical double-layer at oxide-water interface. J Chem Soc Faraday Trans 1974;70:1807–1818.
8. Siu WM, Cobbold RSC. Basic properties of the electrolyte-SiO_2-Si system-physical and theoretical aspects. IEEE Trans Electron Devices 1979;26:1805–1815.
9. Bousse L, de Rooij NF, Bergveld P. Operation of chemically sensitive field-effect sensors as a function of the insulator-electrolyte interface. IEEE Trans Electron Devices 1983;30: 1263–1270.
10. Bousse L, de Rooij NF, Bergveld P. The influence of counterion adsorption on the ψ_0/pH characteristics of insulator surfaces. Surf Sci 1983;135:479–496.

11. Harame DL, Bousse LJ, Shott JD, Meindl JD. Ion-sensing devices with silicon-nitride and borosilicate glass insulators. IEEE Trans Electron Devices 1987;34:1700–1707.
12. van Hal REG, Eijkel JCT, Bergveld P. A general model to describe the electrostatic potential at electrolyte oxide interfaces. Adv Colloid Interface Sci 1996;69:31–62.
13. Matsuo T, Wise KD. Integrated field-effect electrode for biopotential recording. IEEE Trans Biomed Eng 1974;BM21: 485–487.
14. Rocher V, et al. An oxynitride ISFET modified for working in a differential-mode for pH detection. J Electrochem Soc 1994; 141:535–539.
15. Wong HS, White MH. A CMOS-integrated ISFET-operational amplifier chemical sensor employing differential sensing. IEEE Trans Electron Devices 1989;36:479–487.
16. Tsukada K, Miyahara Y, Miyagi H. Platinum-platinum oxide gate pH ISFET. Jpn J Appl Phys Part 1-Regul Pap Short Notes Rev Pap 1989;28:2450–2453.
17. Akiyama T, et al. Ion-sensitive field-effect transistors with inorganic gate oxide for pH sensing. IEEE Trans Electron Devices 1982;29:1936–1941.
18. Voigt H, et al. Diamond-like carbon-gate pH-ISFET. Sens Actuator B-Chem 1997;44:441–445.
19. Liao HK, et al. Study of amorphous tin oxide thin films for ISFET applications. Sens Actuator B-Chem 1998;50:104–109.
20. Yin LT, et al. Study of indium tin oxide thin film for separative extended gate ISFET. Mater Chem Phys 2001;70:12–16.
21. Chin YL, et al. Titanium nitride membrane application to extended gate field effect transistor pH sensor using VLSI technology. Jpn J Appl Phys Part 1-Regul Pap Short Notes Rev Pap 2001;40:6311–6315.
22. Katsube T, Lauks I, Zemel JN. pH-sensitive sputtered iridium oxide-films. Sens Actuator 1982;2:399–410.
23. Jakobson CG, Nemirovsky Y. 1/f noise in ion sensitive field effect transistors from subthreshold to saturation. IEEE Trans Electron Devices 1999;46:259–261.
24. Jolly RD, McCharles RH. A low-noise amplifier for switched capacitor filters. IEEE J Solid-State Circuit 1982;17:1192–1194.
25. Bousse L, Shott J, Meindl JD. A process for the combined fabrication of ion sensors and CMOS circuits. IEEE Electron Device Lett 1988;9:44–46.
26. Bausells J, Carrabina J, Errachid A, Merlos A. Ion-sensitive field-effect transistors fabricated in a commercial CMOS technology. Sens Actuator B-Chem 1999;57:56–62.
27. Hammond PA, Ali D, Cumming DRS. Design of a single-chip pH sensor using a conventional 0.6 μm CMOS process. IEEE Sens J 2004;4:706–712.
28. Jakobson CG, Dinnar U, Feinsod M, Nemirovsky Y. Ion-sensitive field-effect transistors in standard CMOS by post processing. IEEE Sens J 2002;2(4):279–287.
29. van der Spiegel J, Lauks I, Chan P, Babic D. The extended gate chemically sensitive field-effect transistor as multi-species microprobe. Sens Actuator 1983;4:291–298.
30. Smith R, Huber RJ, Janata J. Electrostatically protected ion sensitive field-effect transistors. Sens Actuator 1984;5: 127–136.
31. Baldi A, Bratov A, Mas R, Domínguez C. Electrostatic discharge sensitivity tests for ISFET sensors. Sens Actuator B-Chem 2001;80:255–260.
32. Reinhoudt DN, et al. Development of durable K^+-selective chemically-modified field-effect transistors with functionalized polysiloxane membranes. Anal Chem 1994;66:3618–3623.
33. Brunink JAJ, et al. Chemically modified field-effect transistors - a sodium-ion selective sensor based on calix[4]arene receptor molecules. Anal Chim Acta 1991;254:75–80.
34. Sibbald A, Whalley PD, Covington AK. A miniature flow-through cell with a 4-function CHEMFET integrated-circuit for simultaneous measurements of potassium, hy-drogen, calcium and sodium-ions. Anal Chim Acta 1984;159:47–62.
35. Chovelon JM, Jaffrezic-Renault N, Fombon JJ, Pedone D. Monitoring of ISFET encapsulation aging by impedance measurements. Sens Actuator B-Chem 1991;3:43–50.
36. Grisel A, Francis C, Verney E, Mondin G. Packaging technologies for integrated electrochemical sensors. Sens Actuator 1989;17:285–295.
37. Gràcia I, Cané C, Lora-Tamayo E. Electrical characterization of the aging of sealing materials for ISFET chemical sensors. Sens Actuator B-Chem 1995;24:206–210.
38. Muñoz J, et al. Planar compatible polymer technology for packaging of chemical microsensors. J Electrochem Soc 1996;143:2020–2025.
39. Bratov A, Muñoz J, Dominguez C, Bartrolí J. Photocurable polymers applied as encapsulating materials for ISFET production. Sens Actuator B-Chem 1995;25:823–825.
40. Tsukada K, Sebata M, Miyahara Y, Miyagi H. Long-life multiple-ISFETs with poly-meric gates. Sens Actuator 1989;18: 329–336.
41. Cané C, Gràcia I, Merlos A. Microtechnologies for pH ISFET chemical sensors. Micro-electron J 1997;28:389–405.
42. Bergveld P. The operation of an ISFET as an electronic device. Sens Actuator 1981;1:17–29.
43. Ravezzi L, Conci P. ISFET sensor coupled with CMOS readout circuit microsystem. Electron Lett 1998;34:2234–2235.
44. Sibbald A. A chemical-sensitive integrated-circuit - the operational transducer. Sens Actuator 1985;7:23–38.
45. Palán B, et al. New ISFET sensor interface circuit for biomedical applications. Sens Actuator B-Chem 1999;57:63–68.
46. Hammond PA, Ali D, Cumming DRS. A system-on-chip digital pH meter for use in a wireless diagnostic capsule. IEEE Trans Biomed Eng 2005;52:687–694.
47. Harame DL, Shott JD, Bousse L, Meindl JD. Implantable ion-sensitive transistors. IEEE Trans Biomed Eng 1984;31:572–572.
48. Smith RL, Scott DC. An integrated sensor for electrochemical measurements. IEEE Trans Biomed Eng 1986;33:83–90.
49. Suzuki H, Shiroishi H, Sasaki S, Karube I. Microfabricated liquid junction Ag/AgCl reference electrode and its application to a one-chip potentiometric sensor. Anal Chem 1999;71:5069–5075.
50. Comte PA, Janata J. Field-effect transistor as a solid-state reference electrode. Anal Chim Acta 1978;101:247–252.
51. van den Berg A, Bergveld P, Reinhoudt DN, Sudhölter EJR. Sensitivity control of ISFETs by chemical surface modification. Sens Actuator 1985;8:129–148.
52. Matsuo T, Nakajima H. Characteristics of reference electrodes using a polymer gate ISFET. Sens Actuator 1984;5:293–305.
53. Bergveld P, et al. How electrical and chemical requirements for REFETs may coincide. Sens Actuator 1989;18:309–327.
54. Sudhölter EJR, et al. Modification of ISFETs by covalent anchoring of poly(hydroxyethyl methacrylate) hydrogel-introduction of a thermodynamically defined semiconductor-sensing membrane interface. Anal Chim Acta 1990;230:59–65.
55. van den Vlekkert HH, de Rooij NF, van den Berg A, Grisel A. Multi-ion sensing system based on glass-encapsulated ph-ISFETs and a pseudo-REFET. Sens Actuator B-Chem 1990;1:395–400.
56. Chudy M, Wróblewski W, Brzózka Z. Towards REFET. Sens Actuator B-Chem 1999;57:47–50.
57. Bergveld P. Thirty years of ISFETOLOGY-what happened in the past 30 years and what may happen in the next 30 years. Sens Actuator B-Chem 2003;88:1–20.
58. Visch LL, Bergveld P, Lamprecht W, Sgravenmade EJ. pH measurements with an ion sensitive field-effect transistor in

the mouth of patients with xerostomia. IEEE Trans Biomed Eng 1991;38:353–356.
59. Hiraishi N, et al. Evaluation of active and arrested carious dentin using a pH-imaging microscope and an X-ray analytical microscope. Oper Dent 2003;28:598–604.
60. De Aza PN, Luklinska ZB, Anseau M. Bioactivity of diopside ceramic in human parotid saliva. J Biomed Mater Res Part B 2005;73B:54–60.
61. Sant W, et al. Development of a creatinine-sensitive sensor for medical analysis. Sens Actuator B-Chem 2004;103:260–264.
62. Mackay RS. Radio telemetering from within the body. Science 1961;134:1196–1202.
63. Iddan G, Meron G, Glukhovsky A, Swain P. Wireless capsule endoscopy. Nature (London) 2000;405:417–417.
64. Pandolfino JE, et al. Ambulatory esophageal pH monitoring using a wireless system. Am J Gastroenterol 2003;98:740–749.
65. Johannessen EA, et al. Implementation of multichannel sensors for remote biomedical measurements in a microsystems format. IEEE Trans Biomed Eng 2004;51:525–535.

See also IMMUNOLOGICALLY SENSITIVE FIELD-EFFECT TRANSISTORS; INTEGRATED CIRCUIT TEMPERATURE SENSOR.

ISFET. See ION-SENSITIVE FIELD-EFFECT TRANSISTORS.

JOINTS, BIOMECHANICS OF

George Papaioannou
University of Wisconsin
Milwaukee, Wisconsin

Yener N. Yeni
Henry Ford Hospital
Detroit, Michigan

INTRODUCTION

The human skeleton is a system of bones joined together to form segments or links. These links are movable and provide for the attachment of muscles, ligaments, tendons, and so on. to produce movement. The junction of two or more bones is called an articulation. There are a great variety of joints even within the human body and a multitude of types among living organisms that use exo- and endoskeletons to propel. Articulation can be classified according to function, position, structure and degrees of freedom for movement they allow, and so on. Joint biomechanics is a division of biomechanics that studies the effect of forces on the joints of living organisms.

Articular Anatomy, Joint Types, and Their Function

Anatomic and structural classification of joints typically results in three major categories, according to the predominant tissue or design supporting the articular elements together, that is, joints are called fibrous, cartilaginous, or synovial.

Synovial joints are cavitated. In general, two rigid skeletal segments are brought together by a capsule of connective tissue and several other specialized tissues, that form a cavity. The joints of the lower and upper limbs are mainly synovial since these are the most mobile joints. Mobility varies considerably and a number of subcategories are defined based on the specific shape or architecture and topology of the surfaces involved (e.g., planar, saddle, ball and socket) and on the types of movement permitted (e.g., flexion and extension, medial and lateral rotation) (Table 1). The basic structural characteristics that define a synovial joint can be summarized in four features: a fibrous capsule that forms the joint cavity, a specialized articular cartilage covering the articular surfaces, a synovial membrane lining the inner surface of the capsule that also secretes a special lubricating fluid, the synovial fluid. Additional supportive structures in synovial joints include disks, menisci, labra, fat pads, tendons, and ligaments.

Cartilaginous joints are also solid and are more commonly known as synchondroses and symphyses, a classification based on the structural type of cartilage that intervenes between the articulating parts (Table 2). This cartilage is hyaline and fibrocartilage for synchondroses and *symphyses*, respectively. *Synchondroses* allow very little movement as in the case of the rib cage that contributes to the ability of this area to expand with respiration. Most *symphyses* are permanent; those of sacrum and coccyx can, however, degenerate with subsequent fusion between adjacent vertebral bodies as part of the normal development of these bones.

Fibrous joints are solid. The binding mechanism that dominates the connectivity of the articulating elements is principally fibrous connecting tissue, although other tissue types also may be present. Length, specific arrangement, and fiber density vary considerably according to the location of the joint and its functional requirements. Fibrous joints are classified in three groups: sutures, gomphoses, and syndesmoses (Table 3);

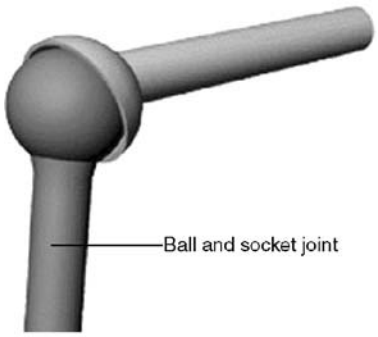
In addition to the obligatory components that all the synovial joints possess, several joints contain intraarticular structures. Discs and menisci are examples of such structures. They differ from one another mainly in that a disc is a circular structure that may completely subdivide a joint cavity so that it is, in reality, two joints in series, whereas a meniscus is usually a crescent-shaped structure that only partially subdivides the joint. Complete discs are found in the sternoclavicular and in the radiocarpal joint. A variety of functions have been proposed for intraarticular discs and menisci. They are normally met at locations where bone congruity is poor, and one of their main functions is to improve congruity and, therefore stability between articular surfaces. Shock absorption facilitation and combination of movements are among their likely roles. They may limit a movement or distribute the weight over a larger surface or facilitate synovial fluid circulation throughout the joint.

The labrum is another intraarticular structure. In humans, this structure is only found in the glenohumeral and hip joints. They are circumferential structures attached to the rim of the glenoid and acetabular sockets. Labra are distinct from articular cartilage because they consist of fibrocartilage and are triangular in their middle section. Their bases are attached to the articular margins and their free apical surfaces lined by synovial membrane. Like discs, their main function is to improve fit and protect the articular margins during extremes of movement.

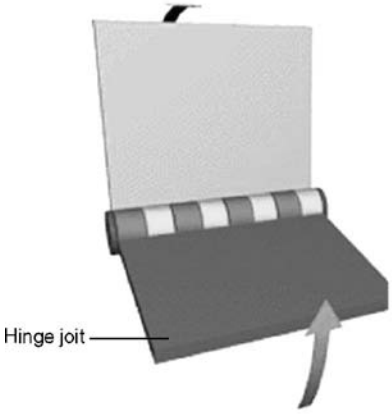
Fat pads are localized accumulations of fat that are found around several synovial joints, although only those in the hip (acetabular pad) and the knee joint (infrapatellar pad) are named. Suggested functions for fat pads include protection of other intraarticular structures (e.g., the round ligament of the head of the femur) and serving as cushions or space-fillers thus facilitating more efficient movement throughout the entire available range.

Bursae are enclosed, self-contained, flattened sacs typically with a synovial lining. They facilitate movement of musculoskeletal tissues over one another and thus are located between pairs of structures (e.g., between ligament and tendon, two ligaments, two tendons or skin, and bone). Deep bursae, such as the illiopsoas bursa or the deep

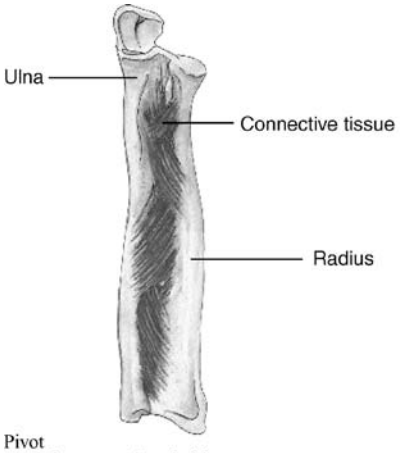
Table 1. Diarthroses: Synovial Joints



Ball and socket
Other names: Spheroidal; endarthroses
Description: Ball-shaped head fits into concave socket
Movement: Widest range of all joints; triaxial
Example: Shoulder and hip joints

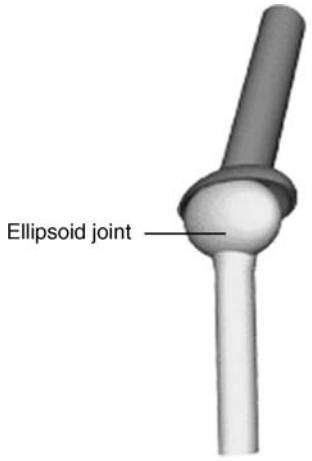


Hinge
Other name: Ginglymus
Description: Spool-shaped head fits into concave surface
Movement: In one plane about single axis (uniaxial); like hinged-door movement (namely, flexion and extension)
Examples: Elbow, knee, ankle, and interphalangeal joints

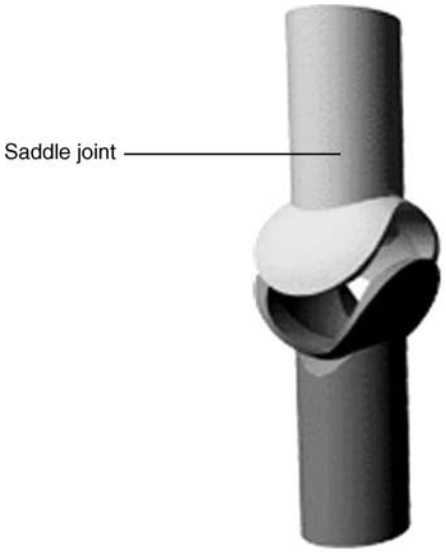


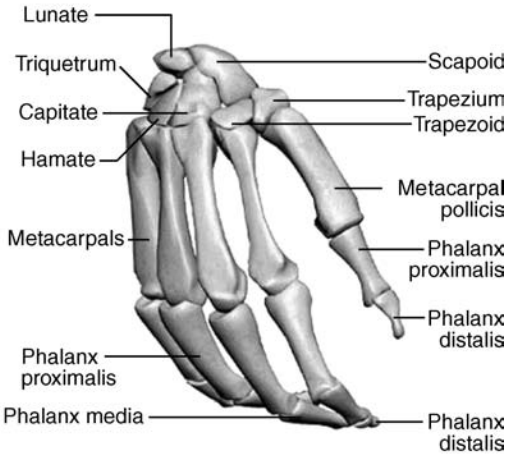
Pivot
Other name: Trochoid
Description: Arch-shaped surface rotates about rounded or peglike pivot
Movements: Rotation: uniaxial
Example: Between axis and atlas; between radius and ulna

Table 1. (*Continued*)



Ellipsoidal
Other names: Condyloid, ovoid
Description: Arch-shaped condyle fits into elliptical cavity
Movements: In two planes at right angles to each other – specifically, flexion, extension, abduction, and adduction; biaxial
Example: Between radius and carpals





Saddle
Other name: Reciprocal
Description: Saddle-shaped bone fits into socket that is concave-convex in opposite direction; modification of condyloid joint
Movements: Same kinds of movement as condyloid joint but freer; like rider in saddle; biaxial
Example: Thumb, between first metacarpal and trapezium

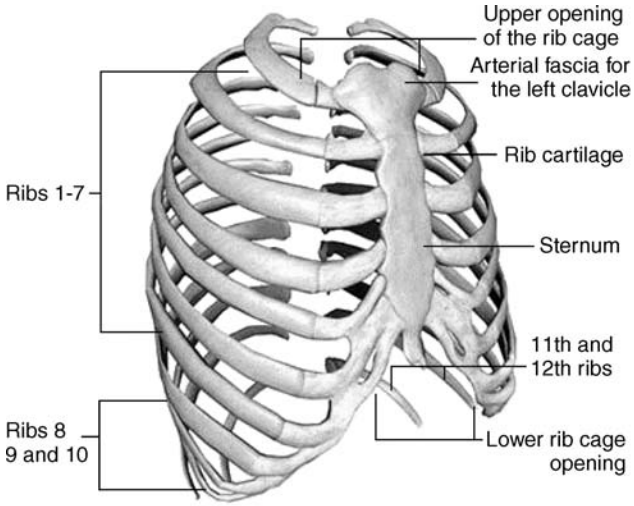
Gliding
Other name: Arthroidal
Description: Articulating surfaces; usually flat
Movement: Gliding, a nonaxial movement
Example: Between carpal bones; between sacrum and ilium (sacroiliac joints)

retrocalcaneal bursa, develop along with joints and by a similar series of events during the embryonic period.

Tendons are located at the ends of many muscles and are the means by which these muscles are attached to bone or other skeletal elements. The primary structural component of tendons is type I collagen. Tendons almost exclusively operate under tensile forces.

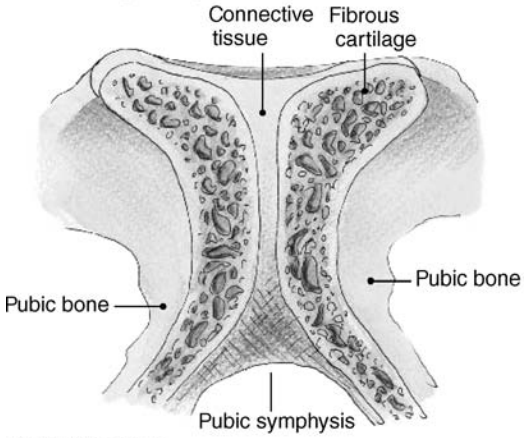
Ligaments are dense bands of connective tissue that connect skeletal elements to each other, either creating (as in the case of syndesmoses) or supporting joints. According to their location they are classified as intracapsular, capsular, or extracapsular. Structurally, they resemble the tendons in that they consist predominantly of type I collagen.

Table 2. Amphiarthroses: Cartilaginous Joints



Cartilaginous
Other name: Synchondroses
Description: The joint formed by each of the costal cartilage
Movement: Bending and twisting, or slight compression
Example: Between the ribs and the sternum; between carpal and tarsal bones

Fibrocartilaginous joint



Fibrocartilaginous
Other name: Symphyses
Description: Within the joint, separating the bones, is a fibrocartilaginous pad – these pads, or discs, serve as shock absorbers
Movement: Compression, flexion, extension, and rotation
Example: Intervertebral and pubic joints

Articular Cartilage

Articular cartilage, the resilient load-bearing tissue that forms the articulating surfaces of synovial joints functions through load distribution mechanism by increasing the area of contact (thereby reducing the stress) and provides these surfaces with the low friction, lubrication, and wear characteristics required for repetitive gliding motion.

Biomechanically, cartilage is another intraarticular absorption mechanism that dampens mechanical shocks and spreads the applied load onto subchondral bone (Fig. 2). Articular cartilage should be viewed as a multiphasic material. It consists primarily of a large extracellular matrix (ECM) with a sparse population of highly specialized cells (chondrocytes) distributed throughout the tissue. The primary components of the ECM are water, proteoglycans, and collagens, with other proteins and glycoproteins present in lower amounts (1). The solid phase is comprised by this porous-permeable collagen-PG matrix filed with freely movable interstitial fluid (fluid phase) (2). A third phase is the ion phase, necessary to describe the electromechanical behaviors of the system. The structure and composition of the articular cartilage vary throughout its depth (Fig. 2), from the articular surface to the subchondral bone. These differences include cell shape and volume, collagen fibril diameter and orientation, proteoglycan concentration, and water content. These all combine to provide the tissue with its unique and complex structure and mechanical properties. A fine mechanism of interstitial fluid pressurization

Table 3. Synarthroses: Fibrous Joints

Sutures
- Other name: —
- Description: The edges of the bones have interdigitations or grooves that fit very closely and firmly together; the connecting fibers are very short
- Movement: None
- Examples: Between the flat bones of the skull

Syndesmoses
- Other name: Ligamentous
- Description: Two bones, that may be widely separated tied together by ligaments; the ligaments may be in the form of cords, bands, or flat sheets
- Movement: None (some give)
- Example: Between the distal ends of the tibia and fibula

Gomphosis
- Other name: —
- Description: A joint in which the surfaces of bony components are adapted to each other like a peg in a hole
- Movement: None
- Example: The conical process of a tooth is inserted in the bony socket of the mandible or maxilla

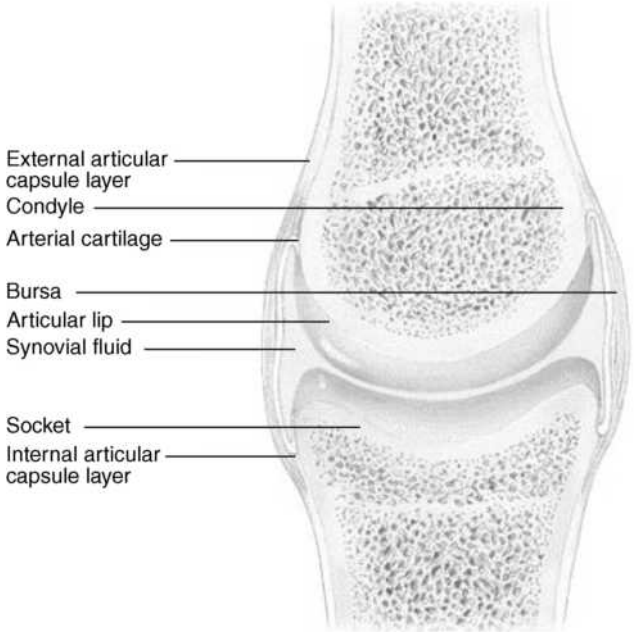


Figure 1. Basic structure and components of a synovial joint (also called diarthroses).

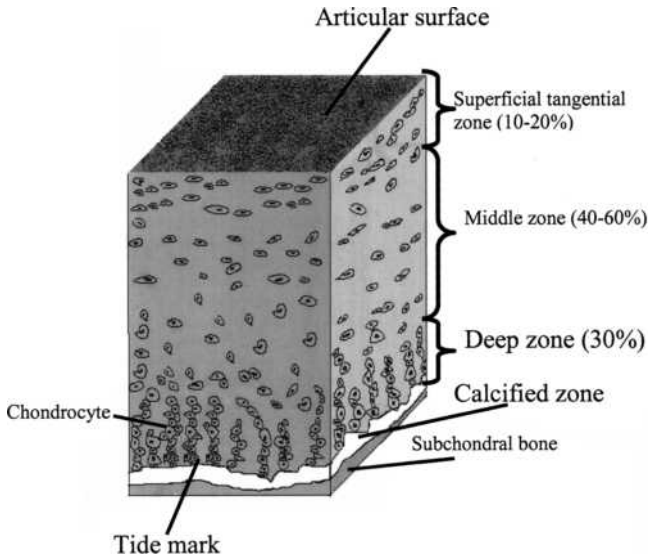


Figure 2. Zones of articular cartilage.

results from the flow of interstitial fluid through the porous-permeable solid matrix that in turn defines the rate dependent load-bearing response of the material. It is noteworthy that articular cartilage provides its essential biomechanical functions for eight decades or more in most of the human synovial joints and no synthetic material performs this well as a joint surface.

The frictional characteristics between two surfaces sliding over each other are significantly influenced by the topography of the given surfaces. Anatomical shape changes affect the way in which loads are transmitted across joints, altering the lubrication mode in that joint and, thus, the physiologic state of cartilage. Articular surfaces are relatively rough, compared to machined bearing surfaces, at the microscopic level. The natural surfaces are surprisingly much rougher than joint replacement prostheses. The mean of the surface roughness for articular cartilage ranges from 1 to 6 μm, while the metal femoral head of a typical artificial hip has a value of ~0.025 μm, indicating that the femoral head is apparently much smoother. Topographic features on the joint surfaces are characterized normally by primary anatomic contours, secondary roughness (<0.5 mm in diameter and <50 μm deep), tertiary hollows on the order of 20–45 μm deep; and, finally, quaternary ridges 1–4 μm in diameter and 0.1–0.3 μm deep. Scanning electron micrographs (SEMs)of arthritic cartilage usually depict a large degree of surface irregularity and anomalous microtopography. These surface irregularities have profound effects on the lubrication mechanism. They accelerate the effects of friction and the rate of degradation of the articular cartilage. The types of joint surface interactions vary greatly between different joints in the body, different animals, between different size animals of the same species, different genders, and different ages. For example, the human hip joint is a deep congruent ball and socket joint (where the cartilage thickens peripherally at the acetabulum); this differs greatly from the bcondylar nature of the distal femur in the knee joint, and the saddle shape of the carpometacarpal joint in the thumb. The degree of shape matching between the various bones and articulating cartilage surfaces composing a joint is a major factor affecting the distribution of stresses in the cartilage and subchondral bone.

Effects of Motion and External Loading on Joints

The articular joint is viewed as an organ with complicated mechanisms of memory and adaptation that accommodates changes in its function. Joint loading results in motion and the couple load–motion is required to maintain normal adult articular cartilage composition, structure, and mechanical properties. The type, intensity, and frequency of loading necessary to maintain normal articular cartilage vary over a broad range. The intensity or frequency of loading should not exceed or fall below these necessary levels, since this will disturb the balance between the processes of synthesis and degradation. Changes in the composition and microstructure of cartilage will result. Reduced joint loading, as has been observed in cases of immobilization by casting or external fixation, leads to atrophy or degeneration of the cartilage. The changes affect both the contact and noncontact areas. Changes in the noncontact areas resulting from rigid immobilization include fibrillation, decreased proteoglycan content and synthesis, and altered proteoglycan conformation, such as a decrease in the size of aggregates and amount of aggregate. Normal nutritive transport to cartilage from the synovial fluid by means of diffusion and convection has been diminished, resulting in these changes. Increased joint loading, either through excessive use, increased magnitudes of loading, or impact, also may affect articular cartilage. Catabolic effects can be induced by a single-impact or repetitive trauma, and may serve as the initiating factor for progressive degenerative changes. Osteoarthritis, a joint disease of epidemic proportions in the western world, is characterized by erosive cartilage lesions, cartilage loss and destruction, subchondral bone sclerosis and cysts, and large osteophyte formation at the margins of the joint (3).

Moderate running exercise may increase the articular cartilage proteoglycan content and compressive stiffness, decrease the rate of fluid flux during loading, and increase the articular cartilage thickness in skeletally immature animals. However, no significant changes in articular cartilage mechanical properties were observed in dogs in response to lifelong increased activity that did not include high impact or torsional loading of their joints. Disruption of the intraarticular structures (e.g., menisci or ligaments) will alter the forces acting on the articular surface in both magnitude and areas of loading. The resulting joint instability is associated with profound and progressive changes in the biochemical composition and mechanical properties of articular cartilage. In experimental animal models, responses to transection of the anterior cruciate ligament or meniscectomy have included fibrillation of the cartilage surface, increased hydration, changes in the proteoglycan content, reduced number and size of proteoglycan aggregates, joint capsule thickening, and osteophyte formation. It seems likely that some of these changes result from the activities of the chondrocytes, because their rates of synthesis of matrix components, breakdown of matrix components, and secretion of proteolytic enzymes are all increased. *In vitro* studies have shown that loading of the cartilage matrix can cause all of these mechanical, electric, and physicochemical events, but thus far it has not been clearly demonstrated which signals are most important in stimulating the anabolic and catabolic activity of the chondrocytes. A holistic physicochemical and biomechanical model of cartilage function in health and disease remains a challenge in the scientific community.

KINEMATICS OF JOINTS

General Comments

Mechanical analysis can refer to kinetics (forces) and/or kinematics (movement), with kinetics being the cause and kinematics the result. Mechanical analysis can develop models proceeding from forces to movements or vice versa. The analysis that starts from the cause (force) is called direct or forward dynamics, and produces a defined set of forces that caused the unique movement. This approach has one solution, and hence is deterministic. Starting from

the movement the analysis is called inverse dynamics. In this case, an infinite number of combinations of individual forces acting on the system can be the causes of the same unique movement, which makes the inverse dynamics approach not deterministic. The simplest and most essential system of mechanical formulations for explaining and describing motion is the Newton's second law. More advanced techniques include the Lagrange, d'Alembert, and Hamilton's methods. In general all of these methods start by describing equations of motion for a rigid body for translation, rotation, or combinations of them for both two (2D) and three-dimensional (3D) space. If the model assumes that the articulated segments that create the articulation are modeled as rigid bodies the remaining task is to calculate the relative motion between the two segments by applying graphics or joint kinematic analysis.

Kinematics is the study of the movements of rigid structures, independent of the forces that might be involved. Two types of movement, translation (linear displacement) and rotation (angular displacement), occur within three orthogonal planes; that is, movement has six degrees of freedom. Humans belong to the vertebrate portion of the phylum chordata, and as such possess a bony endoskeleton that includes a segmented spine and paired extremities. Each extremity is composed of articulated skeletal segments linked together by connective tissue elements and surrounded by skeletal muscle. Motion between skeletal segments occurs at joints. Most joint motion is minimally translational and primarily rotational. The deviation from absolute rotatory motion may be noted by the changes in the path of a joint's "instantaneous center of rotation". These paths have been measured for most of the joints in the body and vary only slightly from true arcs of rotation. For human motion to be effective, not only must a comparatively rigid segment rotate its position relative to an adjacent segment, but many adjacent limb movements must interact as well. Whether the hand is trying to write or the foot must be lifted high enough to clear an obstacle on the ground, the activity is achieved via coordinated movements of multiple limb segments. To provide for the greatest possible function of an extremity, the proximal joint must have the widest range of motion to position the limb in space. This joint must allow for rotatory motions of large degrees in all three planes about all three axes. A means is also provided to translate the limb, so that an extremity can function at all locations within its global range. Rotational motion of the elbow and knee joints allows such overall changes as adjacent limb segments move. Finally, to fine-tune the use of this mechanism with respect to the extremities, for their functional purposes, the hand and foot are required to have a vast amount of movement about all three axes, although the rigid segments are relatively small. Such movement requires the presence of relatively universal joints at the terminal aspect of each extremity.

Characterization of the General Mechanical Joint System: Terminology and Definitions

The displacement of a point is simply the difference between its position after a motion and its position before that motion. It can be represented by a 3D vector drawn from the initial position of the point to its final position. The components of the displacement vector will be the changes in the coordinates of the point's position from measurement in the reference coordinate system. It is apparent that not only the positions, but also displacements measured are relative to some reference. Rigid body (RB) displacements are more complicated than point displacements since for a rigid body a displacement is a change in its position relative to some reference, but more than three parameters are needed to describe it. Two simple types of RB displacement can be described: translation and rotation. An important property of pure translation of a RB is that the displacement vectors of all points in the body are identical and are nonzero. In pure rotation of a RB, although points in the body experience nonzero displacements, one point in that body experiences zero displacement. In addition to that rule, Euler's theorem shows that in pure rotation all points along a particular line through that undisplaced point also experience zero displacement. This line is also known as the axis of rotational displacement. Chasles theorem further states that any displacement of a RB can be accomplished by a translation along a line parallel to the axis of rotation that is defined by Euler's theorem plus a rotation about that same parallel axis. Simply that suggests that any displacement in 3D is equivalent to the motion of a nut, representing the body, on an appropriate stationary screw that was centered on the line described above. Indeed, it can be shown that any displacement in 3D is equivalent to a translation plus a rotation.

Degrees of Freedom

The biological organisms capable of propelling themselves through different media consist of more than one rigid body. A system consisting of a 3D reference frame and an isolated rigid body in space has six degrees of freedom (DOF). To describe the position of each body relative to the ground reference frame it would be necessary to use six parameters, so for two unconnected rigid bodies 12 parameters would be necessary. The system consisting of these two unconnected bodies and the fixed -ground reference would have 12 DOF. The human–animal body consists of a combination of suitably connected bodies. The connections, joints between the bodies, serve to constrain the motions of the bodies so that they are not free to move with what would otherwise be six DOF for each body. Therefore, we can define the number of DOF that the joint removes as the number of degrees of constraint that it provides. It can be shown that every time a joint is added to a system, the number of degrees of freedom in that system is reduced by the number of degrees of constraint provided by that joint. This suggests the following generic formula for the calculation of the degrees of freedom of a system:

$$F = 6(L-1) - 5J_1 - 4J_2 - 3J_3 - 2J_4 - J_5$$

where $F =$ is the number of degrees of freedom in the system of connected joints; $L =$ is the number of joints in the system, including the ground joint (which has no degrees of freedom), and $J_n =$ the number of joints having n degrees of freedom each.

Table 4 contains a description of the major joints in the human body along with the segments-bones that they

Table 4. Characteristics of Major Human Joints

Joint	Bones	Type	DOF	Type of motion	Range of Motion (deg)
Shoulder	Humerus-Scapula	Diarthrosis (spheroidal)	3	Flexion	150
				Extension	50–60
				Abduction	90–120
				Abduction	Complete
				Rotation	
Elbow	Humerus-ulna	Diarthrosis (ginglymus)	2	Flexion	145–160
				Extension	0–5
				Rotation (radius)	
Radioulnar	Superior radius-ulna	Diarthrosis (trochoid)	1	Pronation	70–75
				Supination	85–90
Wrist	Radius-carpal	Diarthrosis (condyloid)	2	Flexion	90–95
				Extension	60–70
				Radial deviation	20–25
				Ulnar deviation	55–65
				Circumduction	Complete
Metacarpal-phalangeal	Metacarpal-phalanges	Diarthrosis (condyloid)	2	Flexion	80–90
				Extension	20–30
				Radial deviation	20–25
				Ulnar deviation	15–20
Finger	Interphalanges	Diarthrosis (ginglymus)	1	Flexion	80–90
				Extension	0–10
Thumb	First metacarpal-carpal	Diarthrosis (reciprocal)	2	Flexion	80–90
				Extension	20–25
				Abduction	40–45
				Abduction	0–10
				Circumduction	Complete
Hip	Femur-acetabulum	Diarthrosis (spheroidal)	3	Flexion	90–120
				Extension	10–20
				Abduction	30–45
				Abduction	30
				Medial rotation	30–40
				Lateral rotation	60
				Circumduction	Complete
Knee	Tibia-femur	Diarthrosis (ginglymus)	2	Flexion	120–140
				Extension	0
				Medial rotation	30
				Lateral rotation	40
Ankle	Tibia-fibula-talus	Diarthrosis (ginglymus)	1	Flexion	20–30
				Extension	40–45
Intertarsal	Tarsals	Diarthrosis (arthroidal)	2	Gliding	Limited motion
Metatarsal-phalangeal	Metatarsals-phalanges	Diarthrosis (condyloid)	2	Flexion	25–30
				Extension	80–90
				Abduction	15–20
				Abduction	Limited
Interphalangeal	Phalanges	Diarthrosis (arthroidal)	1	Flexion	90
				Extension	0
Tibio-fibular	Distal tibia-fibula	Synarthrosis (syndesmosis)	0	Slight movement	Give
Skull	Cranial	Synarthrosis (suture)	0	No movement	
Sterno-costal	Ribs-sternum	Amphiarthrosis (synchondrosis)	0	Slight movement	
Sacroiliac	Sacrum-ilium	Amphiarthrosis (synchondrosis)	0	No movement	Elastic
Intervertebral	Cervical vertebrae	Diarthrosis (arthroidal)	3	Flexion	40
				Extension	75
				Lateral flexion	35–45
				Axial rotation	45–50
	Thoracic vertebrae	Diarthrosis (arthroidal)	3	Flexion	105
				Extension	60
				Lateral flexion	20
				Axial rotation	35
	Lumbar vertebrae	Diarthrosis (arthroidal)	3	Flexion	60
				Extension	35
				Lateral flexion	20
				Axial rotation	5
	Atlas axis	Diarthrosis (trochoid)	1	Pivoting motion	

articulate, their respective type DOF and type/range of motion they provide.

Planar Motion

Some human joints move predominantly in one plane (e.g., the knee joint) in which case the motion can be approximated and analyzed by graphical methods. Here the rotation is characterized by the motion of all points on concentric cycles with an identical angle of rotation around the undisplaced center of rotation (CR). The CR may be located inside and outside of the boundaries of the rotating body. The most common graphical method for the calculation is the so-called bisection technique. If the initial and final states of the body are known, the position of the center or rotation and the angle of rotation may be reconstructed (Fig. 3).

Instantaneous Center of Rotation

When a 2D body is rotating without translation, for example, a rotating stationary bicycle gear, any marked point P on the body may be observed to move in a circle about a fixed point called the axis of rotation or center of rotation. When a rigid body is both rotating and translating, for example, the motion of the femur during gait, its motion at any instant of time, can be described as rotation around a moving center of rotation. The location of this point at any instant, termed the instantaneous center of rotation (ICR), is determined by finding the point which, at that instant, is not translating. Then by definition, at that instant, all points on the rigid body are rotating about the ICR. For practical purposes, the ICR is determined by noting the paths traveled by two points, S and Q, on the object in a very short period of time to S' and Q'. The paths SS' and QQ' will be perpendicular to lines connecting them to the ICR because they approximate, over short periods, tangents to the circles describing the rotation of the body around the ICR at that instant. Perpendicular bisectors to these two paths will intersect at the instantaneous axis of rotation.

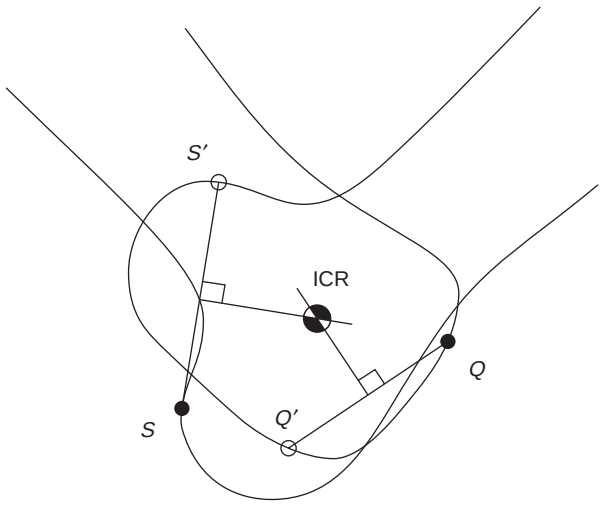


Figure 3. Points S and S′ as well as Q and Q′, lie on the arcs of circles around the center of rotation ICR (used synonymously with CR after the section Instantaneous Center of Rotation). If lines SS′ and QQ′ are bisected perpendicularly, the center of rotation CR is located at the intersection of these perpendicular bisectors. This construction assumes that the perpendicular bisectors are differently orientated but a special case arises if the bisectors are identically oriented. Then the points S, Q and the center of rotation ICR lie on a straight line.

If the ICR is considered to be a point on a moving body, its path on the fixed body is called a fixed centrode. If the ICR is considered to be a point on a fixed body, its path on the moving body is called the moving centrode.

Although in principle two objects may move relative to one another in any combination of rotation and translation, diarthroidal joint surfaces are constrained in their relative motion. The articular surface geometry, the ligamentous restraints, and the action of muscles spanning the joint are the main constraining systems. In general, joint surface separation (or gapping-proximal-distal) and impaction are small compared to overall joint motion. Mechanically, when surfaces are adjacent to each other they may move relative to each other in either sliding or rolling contact. In rolling contact, the contacting points on the two surfaces have zero relative velocity, that is, no slip. Rolling and sliding contacts occur together when the relative velocity at the contact point is not zero. The instant center will then lie between the geometric center and the contact point. All diarthroidal joint motion consists of both rolling and sliding motion. In the hip and shoulder, sliding motion predominates over rolling motion. In the knee, both rolling and sliding articulation occur simultaneously. These simple concepts affect the design of total joint prostheses. For example, some total knee replacements have been designed for implantation while preserving the posterior cruciate ligament, which appears to help maintain the normal kinematics of rolling and sliding in the knee. Other knee prostheses substitute for ligament control of kinematics by alterations in articular surface contour through constraining congruity.

Analytical Methods

Simple kinematic analysis of pure planar translations and rotation or combinations of the two as well as complicated 3D analysis of a rigid body requires the positional information of a minimum of three noncolinear points to describe this motion uniquely. If the position of three points at two instants is known, the displacement from one position to another may be interpreted as translation, rotation or both. Therefore, the first task is to continually monitor the positions of three points on each rigid body. This analysis is conveniently divided into data collection and data analysis.

Data Collection. A constant challenge for the experimental motion analyst is the collection of accurate spatial displacement kinematics of a joint. Several methods have been employed. A review is presented here.

Video and digital optical motion capture (tracking) systems offer state-of-the-art, high resolution, accurate motion capture options to acquire, analyze, and display 3D motion data. The systems are integrated with analog

data acquisition systems to enable simultaneous acquisition (1–300 Hz) of force plate and electromyographic data. Clinically validated software analysis packages are used to analyze and display kinematic, kinetic, and electromyographic data in forms that are easy to interpret.

The major components of a video and digital motion capture system are the cameras, the controlling hardware modules, the software to analyze and present the data, and the host computer to run the software. These systems are designed to be flexible, expandable (from 3 to up to 200 cameras in motion analysis tracking for Hollywood animation movies) and easy to integrate into any working environment. This system collects and processes coordinate data in the least amount of time and requires minimal operator intervention. This system uses motion capture cameras to rapidly acquire 3D coordinate positions from reflective markers placed on subjects. Illuminating strobes with differing wavelengths are used to track the spatial displacement (between 1 and 10 mm resolution) of spherical reflective markers attached to the subject's skin at appropriately chosen locations, preferably on bony landmarks on the human body to minimize skin movement. They can be infrared (IR), visible red, or near-IR strobes to fit the lighting conditions of the capture environment. Also, the lenses can be of fixed or variable focal length for total adaptability. Images are processed within the optical capture cameras where markers are identified and coordinates are calculated before being transferred to the computers. After the completion of the movement, the system provides 3D coordinate and kinematic data. The disadvantages of the system include the skin movement error whose effect is more prominent (3 cm error) at high movement speeds. These high speed motion tasks (impact biomechanics, e.g.) are handled by high speed cine cameras with data acquisition rates several orders of magnitude greater than clinical motion analysis systems. The processing method that is almost real time uses combinations of skin markers (minimum three at each segment) to produce coordinate systems for each segment and eventually describe intersegmental relative motion or relate all the different segment motions to the laboratory fixed-coordinate system. Recently, methods employing clusters of markers have shown to somewhat reduce the skin marker artifact but are yet to be adopted in the clinical practice.

A more accurate method (<1 mm translation and up to 1000 Hz) is the cineradiographical method, which employs an X-ray machine and uses special cameras for capture of sequences of the digital radiographs. In addition to accuracy, these systems directly access the *in vivo* skeletal kinematics so that the resulting analysis can be directly related to bony landmarks. Radiation issues, magnification, and distortion factors are some drawbacks that can be overcome by appropriate image analysis techniques. This method is, however, prone to occlusion errors when two segments overlap and simultaneously cross the field of view of the X-ray source. Stereosystems with more than one X-ray sources can limit this artifact. A biplane radiographic system consists of two X-ray generators and two image intensifiers optically coupled to synchronized high speed video cameras that can be configured in a custom gantry to enable a variety of motion studies. The system can be set up with various set-up modes (e.g., a 60° interbeam angle), an X-ray source to object distance of 1.3 m, and an object to intensifier distance of 0.5 m. Images are acquired with the generators in continuous radiographic mode (typically 100 mA, 90 kVp). The video cameras are electronically shuttered to reduce motion blur. Short (0.5 s) sequences are recorded to minimize radiation exposure. X-ray exposure and image acquisition are controlled by an electronic timer–sequencer to capture only the desired phase of movement.

CODA is an acronym of Cartesian Opto-electronic Dynamic Anthropometer, a name first coined in 1974 to give a working title to an early research instrument developed at Loughborough University, United Kingdom. The 3D capability is an intrinsic characteristic of the design of the sensor units, equivalent to but much more accurate than the stereoscopic depth perception in normal human vision. The system is precalibrated for 3D measurement, which means that the lightweight sensor can be set up at a new location in a matter of minutes, without the need to recalibrate using a space-frame. Each sensor unit must be independently capable of measuring the 3D coordinates of skin markers in real-time. As a consequence, there is great flexibility in the way the system can be operated. For example, a single sensor unit can be used to acquire 3D data in a wide variety of projects, such as unilateral gait. Up to six sensor units can be used together and placed around a capture volume to give "extra sets of eyes" and maximum redundancy of viewpoint. This enables the system to track 360° movements that often occur in animation and sports applications. The calculation of the 3D coordinates of markers is done in real-time with an extremely low delay of 5 ms. Special versions of the system are available with latency shorter than 1 ms. This opens up many applications that require real-time feedback, such as research in neurophysiology and high quality virtual reality systems, as well as tightly coupled real-time animation. It is also possible to trigger external equipment using the real-time data. The automatic intrinsic identification of markers combined with processing of all 3D coordinates in real-time means that graphs and stick figures of the motion and many types of calculated data can be displayed on a computer screen during and immediately after the movement occurs. The data are also immediately stored to file on the hard drive.

A new concept in measuring movement disorders utilizes a unique miniature solid-state gyroscope, not to be confused with gravity sensitive accelerometers. The instrument is fixed with straps directly on the skin surface of the structure whose motion is of interest. It has been successfully used to quantify: tremor (resting, posture, kinetic), rapid pronation–supination of the hand, arm swing, lateral truncal sway, leg stride, spasticity (pendulum drop test), dyskinesia, and alternating dystonia. The system (Motus) senses rotational motion only and is ideal for quantifying human movement since most skeletal joints produce rotational motion. This disadvantage is outweighed by its miniature size that allows it to be of great value for certain types of studies. A different system (Gypsy Gyro) uses 18 small solid-state inertial sensors (gyros) to accurately measure the exact rotations of the actor's bones in real-time for motion capture. The system can easily be worn beneath

normal clothing. With wireless range these systems–suits can be used to record up to 64 actors simultaneously.

Another concept for 3D motion analysis is the measurement system CMS10 (Zebris) designed as a compact device for everyday use. The measurement procedure is based on the travel time measurement of ultrasonic pulses that are emitted by miniature transmitters (markers placed on the skin) to the three microphones built into the compact device. A socket for the power pack (supplied with the device) as well as the interface to a computer are located on the back of the device. The evaluation and display of the measurement data are carried out in real-time. It is possible to use either a table clamp or a mobile floor stand with two joints to support the measurement system.

Data Analysis

Coordinate Systems and Transformation. In the analysis of experimental joint mechanics data, the transformation of point coordinates from one coordinate system to another is a frequent task (4). A typical application of such a transformation would be gait analysis data recorded in a laboratory fixed coordinate system (by means of film or video sequences) that must be converted to a reference system fixed to the skeleton of the test subject. The laboratory fixed coordinate system may be designated by *xyz* and the body reference system by *abc* (Fig. 4). The location of a point $S(a/b/c)$ in the body reference system is defined by the radius vector $s = a \cdot e_a + b \cdot e_b + c \cdot e_c$. Consider the reference system to be embedded into the laboratory system. Then the radius vector $r_m = x_m \cdot e_x + y_m \cdot e_y + z_m \cdot e_z$ describes the origin of the reference system in the laboratory system. The location of $S(x/y/z)$ is now expressed by the coordinates *a, b, c*. The vector equation $r = r_m + s$ gives the radius vector for point *S* in the laboratory system (Fig. 4). Employing the full notation we have: $r = (x \cdot e_x + y \cdot e_y + z \cdot e_z) = (x_m \cdot e_x + y_m \cdot e_y + z_m \cdot e_z) + (a \cdot e_a + b \cdot e_b + c \cdot e_c)$. A set of transformation equations results after some intermediate matrix algebra to describe the coordinates. The scalar products of the unit vectors in the *xyz* and *abc* systems produce a set of nine coefficients C_{ij}. The cosine of the angle between the coordinate axes of the two systems corresponds to the value of the scalar products. Three "direction cosines" define the orientation of each unit vector in one system with respect to the three unit vectors of the other system. Due to the inherent properties of orthogonality and unit length of the unit vectors, there are six constraints on the nine direction cosines, which leaves only three independent parameters describing the transformation. Employing the matrix notation of the transformation equation we have

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x_m \\ y_m \\ z_m \end{bmatrix} + \begin{bmatrix} c_{11}\, c_{12}\, c_{13} \\ c_{21}\, c_{22}\, c_{23} \\ c_{31}\, c_{32}\, c_{33} \end{bmatrix} * \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

In coordinate transformations, the objects remain unchanged and only their location and orientation are described in a rotated and possibly translated coordinate system. If a measurement provides the relative spatial location and orientation of two-coordinate systems the relative translation of the two systems and the nine coefficients C_{ij} can be calculated. The coefficients are adequate to describe the relative rotation between the two coordinate systems.

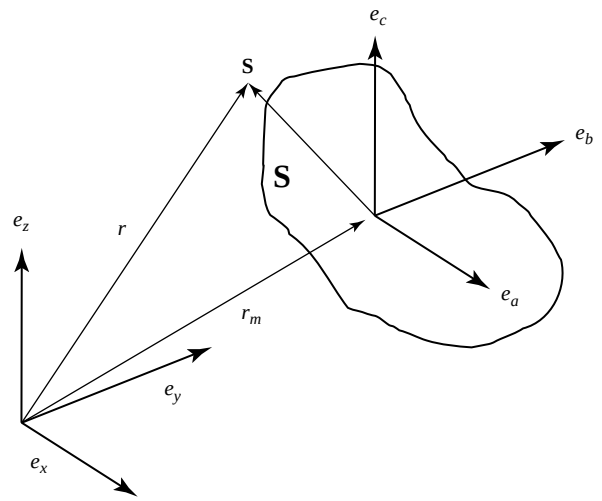


Figure 4. Changing the coordinate systems, transformation of point coordinates from one coordinate system to another.

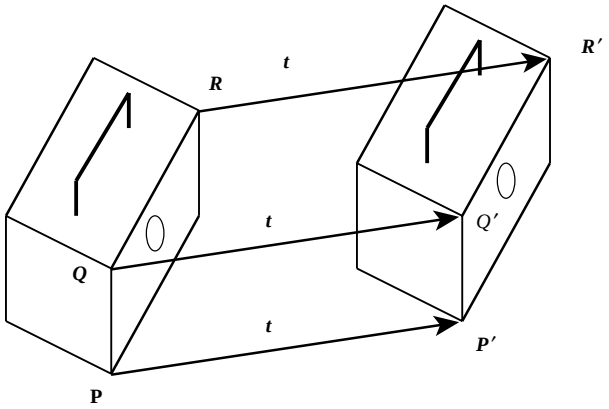


Figure 5. A rigid body (shoebox) moves parallel to itself. The radius vectors from *O* to *P* and from *O* to *P′* are designated by *r* and *r′* so that $r' = r + t$, where *t* is the difference vector.

Translation in Three-Dimensional Space. In translation in 3D space, the rigid object moves parallel to itself (Fig. 5). Pure translation in 3D space leaves the orientation of the body unchanged as in the case of pure 2D translation.

Rotations about the Coordinate Axes. A rotation in 3D space is defined by specifying an axis and an angle of rotation (Fig. 6). The axis can be described by its 3D orientation and location (5). A rotation, as does the translation explained earlier, leaves all the points on the axis unchanged; all other points move along circular arcs in planes oriented perpendicular to the axis (6,7).

This rotation moves an arbitrary point *P* to location *P′* with constant distance *z* from the *xy* plane ($z = z'$). This produces the following matrix notation for the respective equations for the rotation that changes *x* and *y* coordinates

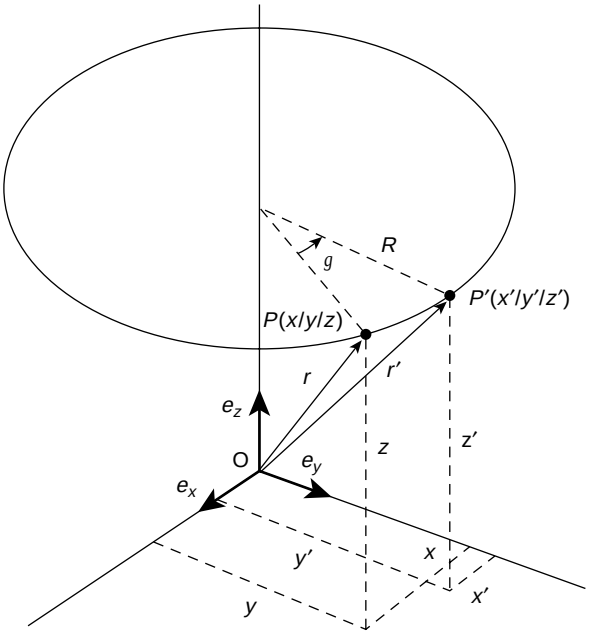


Figure 6. Rotation about the z axis of the coordinate system.

but leaves the z coordinate unchanged.

$$r' = \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = D_z(\gamma)r$$

The matrix describing a rotation about the z axis is designated $\boldsymbol{D}_z(\lambda)$. The matrices describing a rotation about the y axis through angle β and about x axis through angle α are similar.

$$r' = \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = D_y(\beta)r$$

$$r' = \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = D_y(\alpha)r$$

Combined Rotations as a Result of a Sequence of Rotations. Assume that the first rotation of a rigid body occurs about the z axis of a coordinate system. The rotation matrix related to the unit vectors e_x, e_y, e_z, is

$$D_z(\gamma = 90^\circ) = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The second rotation occurs supposedly about the x' axis, that is, about a body-fixed axis on the body (previously rotated about its z axis). The rotation matrix related to the unit vectors e'_x, e'_y, e'_z, is

$$D_{x'}(\alpha = 90^\circ) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}$$

Matrix intermediate calculation here gives

$$r'' = D_{z'} * D_{x'} * r$$

In this calculation the sequence of the matrices is very important especially as this sequence differs from what one might expect. First, the matrix of the second partial rotation acts on the vector $\boldsymbol{r}$ and then, in a second step on the matrix of the first partial rotation. If the sequence of the two partial rotations is interchanged, the combined rotation is described by

$$r'' = D_x * D_{z'} * r$$

For rotations about body-fixed axes it is true that in general, the matrix of the last rotation in the sequence of rotations is the first one to be multiplied by the vector to be rotated. The matrix $\boldsymbol{B}$ describing the image resulting from n partial rotations about body-fixed axes is composed according to the formula:

$$B_{\text{body - fixed}} = D_1 * D_2 * D_3 * \ldots * D_{n-1} * D_n$$

where the indexes indicate the sequence of the rotations. Alternatively, if the n rotation were to be produced about axes fixed in space (i.e., fixed in the ground, laboratory frame) and not about body-fixed axes, the sequence of the matrexes in the matrix product would be different:

$$B_{\text{space - fixed}} = D_n * D_{n-1} * \ldots * D_2 * D_1$$

Euler and Bryant–Cardan Angles. Any desired orientation of a body can be obtained by performing rotations about three axes in sequence. There are, however, many ways of performing three such rotations. One can do this task at random, but for reasons of clarity two conventions are frequently used: the Euler's and Bryant–Cardan's rotations. In the Euler notation, the general rotation is decomposed of three rotations about body-fixed axes in the following manner:

Rotation 1: about the z axis through the angle φ rotation matrix $D_z(\varphi)$ (Fig. 7).
Rotation 2: about the x' axis through the angle θ rotation matrix $D_{x'}(\theta)$.
Rotation 3: about the z'' axis through the angle ψ rotation matrix $D_{z''}(\psi)$.

The matrix describing Euler's combined rotation is given by the matrix product

$$B = D_z(\varphi) * D_{x'}(\theta) * D_{z''}(\psi)\,(\text{Euler})$$

According to the Bryant and Cardan the general rotation is decomposed of three rotations about body-fixed axes in the following manner:

Rotation 1: about the x axis through the angle φ_1 rotation matrix $D_x(\varphi_1)$ (Fig. 7).
Rotation 2: about the y' axis through the angle φ_2 rotation matrix $D_{y'}(\varphi_2)$.
Rotation 3: about the z'' axis through the angle φ_3 rotation matrix $D_{z''}(\varphi_3)$,

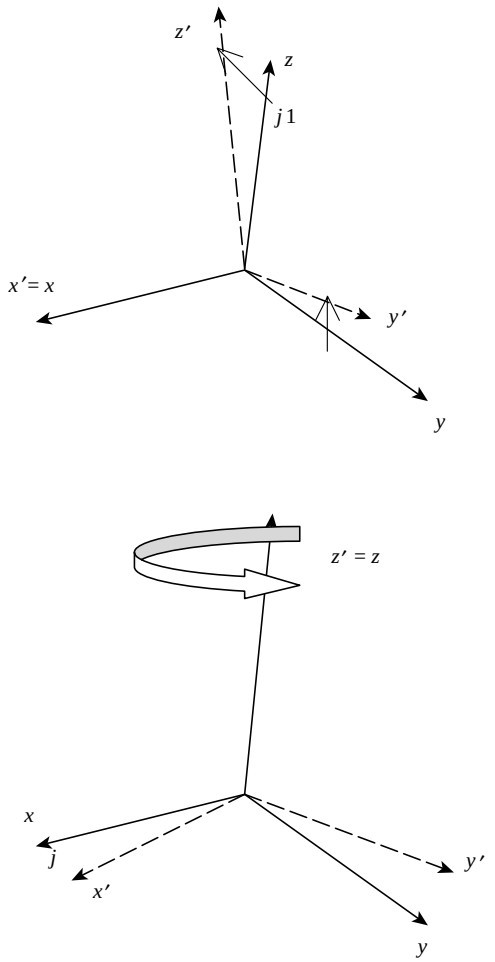


Figure 7. General rotation composed of three partial rotations. The first rotation according to the Bryant–Cardan convention (above). The first of the general rotations using Euler as the selection of the axes and angles of rotation (below).

in which case the matrix of combined rotation is given by

$$B = D_x(\varphi_1) * D_{y'}(\varphi_2) * D_{z''}(\varphi_3)\,(\text{Bryant} - \text{Cardan})$$

For reasons of simplicity, we have presented single or combined rotations about coordinate axes, but more complicated rotational laws can be applied as we deal with rotations about arbitrary axes. Rotation and translation can also be integrated into one single motion with Chasles theorem. Chasles theorem states that "the general motion in 3D space is helical motion", or "the basic type of motion adapted to describe any change of location and orientation in 3D space is helical motion". The relevant axis of rotation is designated the "helical axis". Chasles theorem is also known as the "helical axis" theorem.

KINETICS OF JOINTS

The study of the forces that bring about the movements discussed above is called kinetics. Because kinetics provides insights into the cause of the observed motion, it is essential to the proper interpretation of human movement processes. Forces and loads are not visually observable; they must be either measured with instrumentation or calculated from kinematics data. Kinetic quantities studied include such parameters as the forces produced by muscles; reaction loads between body parts as well as their interactions with external surfaces; the load transmitted through the joints; the power transferred between body segments; and the mechanical energy of body segments. Inherent to such studies are the functional demands imposed on the body. The structure and stability of each extremity and its joints reflect different systems and functional demands. The functional demands on the upper extremity are quite different from those on either the upper and lower axial skeleton or those on the lower extremity. Depending on which joint and/or structure is addressed, different types and degrees of rotational motion are allowed and are functional. How much structural strength is needed versus how much movement is allowed in each area dictates the nature of the material, size, shape, and infrastructure of the joint system established to perform a given movement.

Equations of Motion

The kinetics deal with the effects of forces on the motion of a body. When the motion is known, the problem is then to find the force system acting on the body. There are joint forces and joint moments. With all the kinematic quantities known, it is possible to find the joint forces and moments from the resulting force system that acts on each element. This is done by solving a system of simultaneous equations at successive time intervals. Since muscles are an unknown force system, the resolved muscle force and the real joint force are treated as totally unknown joint forces in the analysis. The three equations of motion for linear motions are

$$\sum F = ma_x \quad \sum F = ma_y \quad \sum F = ma_z$$

The three equations of motion for rotation are

$$\begin{aligned}\sum M_x &= I_{xx}\alpha_x - (I_{yy} - I_{zz})\omega_y\omega_z - I_{xy}(\alpha_y - \omega_x\omega_z)\\ &\quad - I_{yz}(\omega_y^2 - \omega_z^2) - I_{xx}(\alpha_z + \omega_x\omega_y)\\ \sum M_y &= I_{yy}\alpha_y - (I_{zz} - I_{xx})\omega_z\omega_x - I_{yz}(\alpha_z - \omega_y\omega_x)\\ &\quad - I_{xx}(\omega_z^2 - \omega_x^2) - I_{xy}(\alpha_x + \omega_y\omega_z)\\ \sum M_z &= I_{zz}\alpha_z - (I_{xx} - I_{yy})\omega_x\omega_y - I_{zx}(\alpha_x - \omega_z\omega_y)\\ &\quad - I_{xy}(\omega_x^2 - \omega_y^2) - I_{yz}(\alpha_y + \omega_z\omega_x)\end{aligned}$$

where M is the moment, I is the mass moment of inertia, α the angular acceleration, and ω is the angular velocity. The moment equations can be simplified if the axes of the reference frames coincide with the principal axes, with the origin at the center of gravity. These equations, called Euler equations, are

$$\begin{aligned}\sum M_x &= I_x\alpha_x - (I_y - I_z)\omega_y\omega_z\\ \sum M_y &= I_y\alpha_y - (I_z - I_x)\omega_z\omega_x\\ \sum M_z &= I_z\alpha_z - (I_x - I_y)\omega_x\omega_y\end{aligned}$$

Continuity conditions are derived based on the fact that equal and opposite forces and moments occur at the joint between the two segments.

The anthropometric data for the mass, the center of gravity, the moment of inertia, and so on for the different parts of the human body are available in the literature (8,9).

Motion and Forces on Diarthroidal Joints

In vivo experimental measurements on the relative motions between articulating surfaces of a joint, which correspond to daily activities, are limited. Most quantitative information is obtained from gait studies that do not provide the accuracy and precision for the detailed information required for lubrication studies. However, even simple calculations show that translational speeds between two articulating surfaces can range from $\sim 0.06\,\mathrm{m.s^{-1}}$ between the femoral head surface and the acetabulum surface during normal walking, to $\sim 0.6\,\mathrm{m.s^{-1}}$ between the humeral head surface and the glenoid surface of the shoulder when a baseball pitcher throws a fastball. Cartilage to cartilage contact or fluid-film layers, or a mixture of both are normally the contact mechanisms at the joint. During a normal walking cycle, the human hip, knee, and ankle joints can be subjected to loads on the order of six times body weight, with these peak loads occurring just after heel-strike and just before toe-off. The average load on the joint is approximately three to five times body weight, which lasts as long as 60% of the walking cycle. During the swing phase of walking, only light loads are carried. During this phase, the articular surfaces move rapidly over each other. In addition, extremely high forces occur across the joints in the leg during jumping. Descending stairs can load the knee with up to 10 times body weight, suggesting that the load on the joint surface is dependent on the task performed, that is, the loading sites change drastically as the articulating surfaces move relative to each other.

MATHEMATICAL AND MECHANICAL MODELS OF JOINTS

Locomotion results from complex, high dimensional, nonlinear, dynamically coupled interactions between an organism and its environment. Simple models called templates have been and can be made to resolve the redundancy of multiple legs, joints, and muscles by seeking synergies and symmetries. The simplest model (least number of variables and parameters) that exhibits a targeted behavior is called a template (10). Templates can be used to test control strategies against empirical data. Templates must be based in more detailed morphological and physiological models to ask specific questions about multiple legs, the joint torques that actuate them, the recruitment of muscles that produce those torques and the neural networks that activate the ensemble. These more elaborate models are called anchors. They introduce representations of specific biological details of the organism. The control of slow, variable-frequency locomotion appears to be dominated by the nervous system, whereas during rapid, rhythmic locomotion, the control may reside more within the mechanical system. Anchored templates of many-legged, sprawled-postured animals may suggest that passive, dynamic self-stabilization from a feedforward, tuned mechanical system can reject rapid perturbations and simplify control. Future progress would benefit from the creation of a field embracing comparative neuromechanics. Both templates and anchors are part of a system of mathematical and structural definitions and standard methods of description and dissemination of knowledge. In the next few sessions an attempt is made to describe some of those methods.

The human musculoskeletal system is often modeled by joining rigid links with continuous mass distribution. The joint may be of the revolute or spherical type, with restrictions consistent with body construction. Usually, the human segments form an open linkage.

Knowing all the kinematics at the center of mass of the segments, the joint force and the moment analysis proceeds by drawing free-body diagrams of the segments involved. The free body diagrams for the hip joint, the knee joint, and the ankle joint in a sagittal plane are illustrated in Fig. 8 as an example:

Assessment of Mechanical Factors Associated With Joint Degeneration: Limitations and Future Work

Joint degeneration results from complex, multidimensional, nonlinear, dynamically coupled interactions between the organism and its environment. The assessment of mechanical factors associated with joint degeneration has traditionally combined longitudinal clinical studies with carefully designed experimental techniques and theoretical computational analyses. The quality of

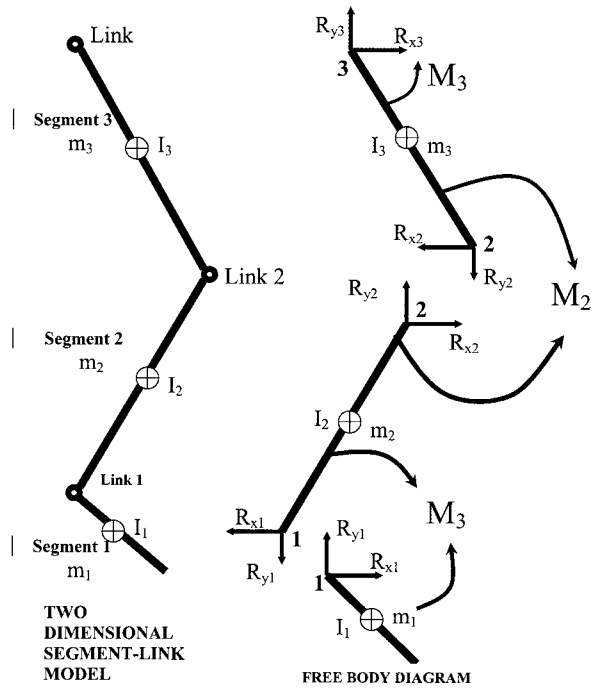


Figure 8. Relationship between the free body diagram and the link-segment model. Each segment is "broken" at the joints, and the reaction forces and moments of force acting at each joint are shown.

such assessments depends both on the accuracy–precision of the measurement methodology and the theoretical framework for its interpretation (i.e. joint mathematical models). To capitalize on the increasing level of measurement accuracy, theoretical analysis requires more detailed morphological and physiological models (11). For example, because of variations between individuals, the detailed, morphological analysis required for accurate modeling of cartilage stresses must be patient specific. In addition, the dynamics of the human task to be modeled (for e.g., human jumping) as expressed in the strain rate of tissue deformation must be accounted for in the analysis. Once the error estimate (simulation versus experiment) is established, model predictions can address specific clinical-biological questions. Traditionally, *in situ* methodology (cadaveric experiments and *in vitro* tests) is applied when an *in vivo* measurement is impossible. This limitation presents a number of implications and assumptions that weaken the theoretical analysis. Recent developments have further improved accuracy in the experimental measurement of *in vivo* knee kinematics. These developments allow significant improvements upon previous limitations by applying patient-specific task-dependent models in the study of joint pathogenesis.

The vast majority of dynamic knee studies have been performed with conventional motion analysis techniques, using markers attached to the skin. Conventional motion analysis is not sufficiently accurate to enable analysis of cartilage stress. Previous studies have shown that skin markers move substantially relative to underlying bone, with RMS errors of 2–7 mm and peak errors as large as 14 mm in estimates of tibial position during gait (12). A study of four subjects during running and hopping (using 250 frame$\cdot$s^{-1} stereo radiography) has demonstrated skin marker motion relative to the femur averaging 1–5 mm throughout the motion, with peak-to-peak errors of the oscillation at impact averaging 7–14 mm (13). Errors were both subject and activity specific (14). Techniques have been developed for improving estimates of bone dynamics from skin markers using large numbers of markers and optimization–modeling of soft tissue deformation (15,16) but the performance of these methods for *in vivo* studies has only been validated for the tibia (during a slow, impact-free 10 cm step-up movement of a single patient with an external fixator). Average errors were low, but peak errors routinely exceeded 1 mm. Errors would likely increase significantly during faster movements and movements involving impact, and also where the soft tissue layer between skin and bone is thicker (e.g., for the femur). However, if the kinematic measurements are to be used in conjunction with musculoskeletal models to estimate dynamic loads and stresses on joint tissues, then even errors as small as 1 mm may be unacceptable. For example, when estimating strains in the ACL, a $\pm$1 mm error in tibio–femoral displacement could introduce uncertainty in the ligament length of approximately $\pm$3% (assuming a nominal ligament length of 30 mm). This error is similar in magnitude to estimated peak ligament elongation occurring during common activities, such as stair climbing (17). For investigating cartilage deformation, this error magnitude would be even less acceptable. A 1 mm displacement would be equivalent to a cartilage strain of $\sim$ 25%, relative to the average thickness of healthy tibio–femoral cartilage (18). A displacement error of this magnitude would translate into huge differences in estimates of contact forces. Thus, efforts to model, predict and correct for soft tissue deformation are unlikely to achieve sufficient accuracy for assessing soft tissue behavior. Alternatively, kinematics from a high speed stereoradiographic system capable of tracking implanted tantalum markers *in vivo* with 3D accuracy and precision better than 0.072 mm in translation and 0.35° in rotation (19) are more appropriate in use with advanced computational models. This accuracy is an order of magnitude or greater of improvement over conventional motion analysis techniques, and is uniquely capable of providing the accuracy necessary to model joint stresses.

From Experimental to Advanced Theoretical Analysis in Joint Mechanics

In addition to measuring joint kinematics and contact areas, investigators have attempted to measure articular contact stresses and pressures. However, stresses throughout the cartilage layer cannot be measured experimentally. Direct measurements of stress can be made at the articular surface using pressure sensing devices (20–22) (e.g., pressure sensitive Fuji film, piezoresistive contact pressure transducers, dye staining, silicone rubber casting). For cadaver studies, Fuji film sheets (Fuji Prescale Film; Itoh, New York, NY) are inserted in a joint and if pressed produce a stain whose intensity depends on the static applied pressure. Alternatively, digital electronic pressure sensors (e.g., K-scan, Tekscan, Boston, MA) can be placed onto the articular surface. These sensors are thin and flexible, and can be made to conform to the anatomy of the medial and lateral knee compartments. They consist of printed circuits divided into grids of load-sensing regions. Each load-sensing region within the grid has a piezoresistive pigment that can be used to determine the total compressive load within that region. After appropriate calibration procedures, dynamic pressure distributions can be calculated. In addition to providing a continuous, dynamic readout, K-scan has been reported to more accurately estimate contact areas than Fuji film (23,24).

There are significant concerns with the use of these sensors for estimating actual contact pressures. These techniques measure only surface-layer stresses, they alter the nature of cartilage surface interactions and are too invasive for *in vivo* human use. Thus, the clinical validity of articular pressure measurement with such sensors is questionable. They can, however, be important tools for the evaluation of the predictive power of joint models (2). By including the sensor in a finite element model, the effects of the sensor film on the actual contact mechanics can be accounted for (25). Thus, contact pressure predictions from such models can be directly compared to the pressure sensor measurements for finite element (FE) model validation.

Many *in situ* experimental studies have been conducted to obtain 3D knee joint kinematics and force-displacement data (21,26–30). Cadaver studies, however, cannot reproduce the complex loading seen by the joint during strenuous movements, since the muscle forces driving the

movement cannot be simulated. Because of these fundamental limitations of experimental measures, mathematical models are favored for obtaining comprehensive descriptions of the spatial and temporal variations of cartilage stresses. A numerical model could be used to perform parametric studies of geometry, loading or material properties in controlled ways that would not be possible with tissue samples.

Theoretical Analysis of Joint Mechanics

During the last two decades, a number of theoretical joint mechanics studies with different degrees of accuracy and predictive power have been presented in the literature (31–39). Computational modeling work has included anatomical or geometrical observation (40,41) and analytical mathematical modeling (42–46). More recently, advanced FE modeling approaches allowed for improvements in the predictive power of localized tissue deformation (47–54). Joint biomechanics problems are characterized by moving contacts between two topologically complex soft tissue layers separated by a thin layer of non-Newtonian synovial fluid. A prime example is the multibody sliding contact problem between the tibia, femur, and menisci. The complexity of such problems requires implementation of sophisticated numerical methods for solutions (55–58). The finite element method is ideally suited for obtaining solutions to joint contact problems. Thus far, much of the finite element analysis has been applied to the study of hard tissue structures, often as it relates to prosthetic devices (59,60). When addressed, soft tissue layers are treated as single-phase elastic materials. As a consequence of the relative dearth of precise patient specific geometric data, material properties and insufficiency in accuracy of *in vivo* kinematics for input, no patient specific computational models have been reported for longitudinal clinical joint studies.

Surface Modeling

Surface modeling methods calculate the shape variations of joints and visualize the proximity of subchondral bone surfaces during static loading or dynamic movement. These methods can combine *in situ* data, motion analysis optical system data or high speed biplane radiographic image data and 3D bone surface information derived from computed tomography to determine subchondral bone motion. This method can be used to identify the regions of contact during static loading or dynamic motion, to calculate the surface area of subchondral bone within close contact, and to determine the changing position of the close contact area during dynamic activities (Fig. 9).

In vivo dynamic joint surface interaction information would be useful in the study of osteoarthritis changes in joint space and contact patterns over time, in biomechanical modeling to assist in finite element modeling, and in identifying normal and pathological joint mechanics pre- and postsurgery. Previous attempts to quantify the interaction between bones have utilized various methods including castings (62,63), pressure sensitive film (64), mathematical surface modeling (65,66), implant registration (67) and cine phase contrast magnetic resonance imaging (MRI) (68). The casting method can only be applied to cadaver models under static loading conditions. Pressure sensitive film also requires a cadaver model and necessitates inserting material into the joint space. Mathematical surface modeling allows analysis of dynamic motions *in vivo*, however, the joint must be disarticulated after testing. Implant registration requires either surgical implants or nonsubject specific image matching algorithms. Cine phase contrast MRI requires repeatedly performing the same motion pattern during testing and is limited to a small range of motion. The process described below is an improvement on these previous techniques because it utilizes live subjects performing dynamic tasks with unrestricted motion. Direct measurement of articular cartilage behavior *in vivo* during dynamic loading is problematic. In order to estimate the behavior of articular cartilage, the surface proximity interaction method that precisely tracks the motion of subchondral bone surfaces in vivo. Articular cartilage behavior is then estimated from these subchondral bone measurements.

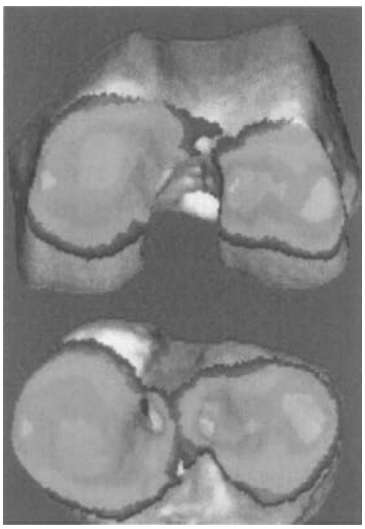


Figure 9. Example applications showing dynamic *in vivo* tibiofemoral bone surface motion using joint proximity (Euclidian distance) mapping during human one-legged hopping.

Anderst et al. (61) described a method to estimate *in vivo* dynamic articular surface interaction by combining joint kinematics from high-speed biplane radiography with 3D bone shape information derived from computed tomography (CT). Markers implanted in the bones were visible in both the CT scans and the radiographic images, and were used to register the subchondral bone surfaces with the 3D bone motion. Joint surface interactions were then estimated by analyzing the relative proximity of the subchondral bone surfaces during the rendered movement. Computed Tomography data can be also used for joint geometry–shape characterization. The method is referred as reconstruction of volumetric models into rendered joint surface geometry models.

Computed Tomography data are typically collected for this method with slice spacing between the different images of $\sim 0.625 - 1.25$ mm and the in-plane resolution

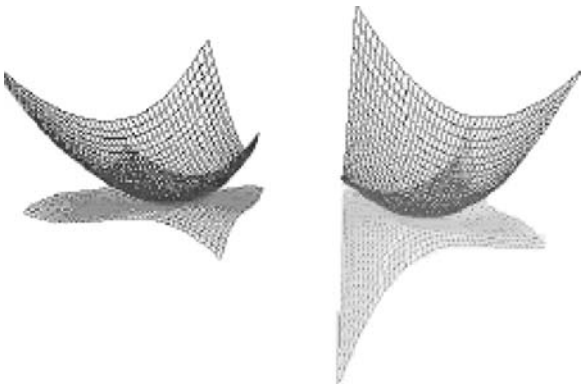


Figure 10. Articular surface matching (femoral condyle on top and tibial plateau below) using geometrical objects (69).

is $\sim$ 0.293–0.6 mm depending on the size of the bone. The CT scans are reconstructed into 3D solid figures using software that employs reconstruction techniques, that is, the regularized marching tetrahedra algorithm by Treece et al. (70). If necessary, threshold values are adjusted to ensure the entire bone surface appeared in the reconstruction and the opposing bone surfaces never overlapped in computer animations of the motion.

Anterior–posterior and lateral radiographs are commonly used to preoperatively determine prosthetic size and proper donor selection for osteochondral allografts. By using 3D computer aided design tools and the reconstructed 3D joint geometry from CT described above, size determination is less prone to out-of-plane imaging errors associated with sagittal and coronal roentgenograms. Assessment of surface size, curvature analysis and knee incongruity is possible with *in vivo* CT [Fig. 10 (69,71,72)]. After the 3D joint surface reconstruction models the distal articular femur ($n = 16$) can be represented by six circles, the diameters of these circles, their angular arcs, and the distances between their centers varied with the size of the femur (Fig. 11; Table 5). There is a statistically significant association between several geometry parameters when the lateral or the medial distal femur is studied independently. These associations do not exist when we correlate medial versus lateral compartments across the population.

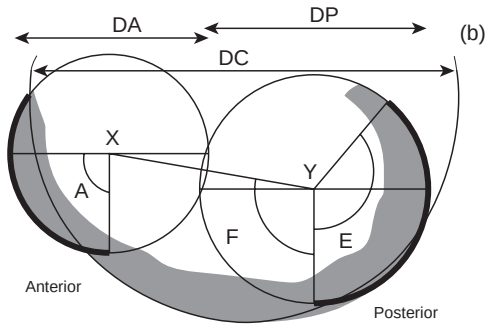


Figure 11. All the sagittal view measured parameters in the study of femoral head congruity (69).

The Joint Distribution Problem

Much attention has been devoted to the solution to what has become known as the general joint distribution problem that is, the problem of estimating the *in vivo* forces transmitted by the individual anatomical structures in the joint neighborhood during some activity of interest (73). The prediction of forces in joint structures has many applications. In the field of medicine, these predictions are useful for obtaining a better understanding of muscle and ligament function, mechanical environment within which prosthetic components must operate, and mechanical effects of musculoskeletal diseases. In the realm of sport biomechanics, these predictions are useful for better understanding of the kinetic demands, performance constraints, and mechanisms for improving athletic performance. Industrial applications include the optimization of occupational performance and safety considerations. Although the general techniques for predicting forces in joint structures may be used throughout this broad range of applications, the particular method of choice and the details of the analysis depend on the application.

The general force distribution problem normally arises in the following way. The musculoskeletal system or a relevant portion thereof is modeled as a mechanical system consisting of a number of essentially rigid elements (body segments) subjected to forces due to the presence of a gravitational field, and segmental contact with external objects, neighboring segments, and soft tissue structures that produce and constrain system motion. The associated inverse dynamics problem is then formulated and solved to determine the variable intersegmental (joint) force and moment resultants during the activity of interest. The joint resultants are abstract kinetic quantities that represent the net effect of all the forces transmitted by the anatomic structures crossing the joint. At a typical joint, these forces normally include the forces transmitted by the muscles, ligaments, and articular (bony) contact surfaces.

The unknown forces transmitted by the joint structure are next related to the known intersegmental resultants by writing the joint equilibrium equations. These equations express the fact that the vector sum of all the forces in the individual anatomic structures, and the vector sum of all the moments about the joint center produced by those forces, are equal to the intersegmental resultant force and moment, respectively. Assuming that all joint geometry (point of application and orientation of forces) is known and that these two independent vector equations (or six independent scalar equations) involve as unknowns the M muscle and L ligament forces, together with the $3C$ scalar components of the C bony contact forces, these joint equilibrium equations are indeterminate whenever the sum $(M + L + 3C)$ of the unknown forces exceeds six. Thus, if the system model includes only one bony contact force $(C = 1)$ and more than three muscle and/or ligament forces $(M + L > 3)$, the corresponding joint distribution problem will be indeterminate and therefore have an infinite number of solutions.

Finally, the joint resultants are decomposed or distributed to the individual joint structures at each instant of

Table 5. Femoral ($n = 16$) Medial and Lateral Compartment Measurements[a]

	Medial	SD	Lateral	SD	Ratio M/L
DC mm	68.28	5.003	67.839	5.865	1.006
DA mm	42.45	10.086	44.41	4.608	0.955
DP mm	40.364	1.231	41.212	3.069	0.979
XY mm	24.519	2.686	23.529	3.069	1.042
F°	100.374	10.572	102.631	5.834	0.978
E°	151.168	10.9	139.629	12.509	1.0824

interest during the activity, using some appropriate solution methodology.

The general joint distribution problem may thus be stated in the following way. At any instant of time when the joint resultants are known, the forces transmitted by the individual joint structures are determined such that the equilibrium equations, and all relevant constraints on the forces in the individual joint structures, are simultaneously satisfied. The classical studies of joint distribution problems use essentially two different methods to solve the indeterminate joint distribution problem: the reduction method, and the optimization method.

The mathematical modeling of human anatomy and its functions has been influenced by two main simulation approaches or philosophies. In the first the joint structures are of no importance in the mathematical modeling while in the second simulation of the geometry and structural relationships of the joint components in addition to their behavioral properties are the main tasks. Hefzy et al. categorized these different approaches as phenomenological and anatomical, respectively (74).

Phenomenological Joint Models. The phenomenological models include two groups: the rheological models and the advanced figure animation models. The rheological models analyze the dynamic behavior of a system by treating it as viscoelastic, being composed of springs and dampers. However, the noncorrespondence of these components to the structure of the components in the knee leads to no structural information in the model output.

The advanced figure animation models provide information on body dynamics by taking into account body segment dimensions, masses, moments of inertia, and so on, but do not model the detailed geometry of joints.

Anatomical Models. Two different approaches to modeling categorization exist in the experimental literature. According to the first, the categorization of the models depends on the type of motion reproduced by the mathematics. The second approach categorizes models according to their structural basis. There is a vast number of studies attempting to model specific component structure and behavior that will be evaluated in order to identify the optimum method for the modeling of each specific component.

The Reduction Method. The reduction method reduces the number of unknown forces to correspond with the number of equations governing the distribution problem (or increases the number of equations to agree with the number of unknowns, e.g., the deformation-force relations for the unknown forces). For the general 3D distribution problem, this implies that the number of unknown scalar forces ($M + L + 3C$) must be reduced to six to allow for a unique solution. Previous investigators have reduced the number of unknowns by (*1*) grouping muscles and ligaments with apparently similar functional roles, (*2*) grouping multiply connected bony contact force regions, (*3*) assuming a direction for the unknown bony contact force, (*4*) using EMG data to determine when a muscle is active, (*5*) ignoring ligament forces except near the limits of the range of motion, and (*6*) ignoring antagonistic muscular activity. Several models [(73,75); Fig. 12a, b] predicted muscle forces and the bony contact force at the hip during locomotion. In these studies, the indeterminate distribution problem at the hip or the knee was made determinate through several simplifying assumptions. The hip muscles were combined into six functional groups (long flexors, short flexors, long extensors, short extensors, abductors, and adductors), and ligament function at the hip was assumed to be negligible. The forces transmitted by these six muscle groups, combined with the three components of bony contact force, comprised nine unknown scalar quantities. Only two of the three components of the resultant hip moment were considered; analysis of the component tending to internally or externally rotate the femur relative to the pelvis was rejected as inaccurate. Previous reports of EMG activity were to demonstrate that there is little antagonistic muscle action, and only muscle agonists were considered. Despite these simplifications, the possibility of activity in both the long and short flexors and extensors still made the problem indeterminate. A solution was obtained, however, by assuming activity in only one flexor (either long or short, but not both) and one extensor.

The Optimization Method. The previous discussion indicates that the distribution problem at a joint is typically an indeterminate problem, since the number of muscles, ligaments and bony contact regions available to transmit forces across a joint in many cases exceeds the minimum number required to generate a determinate solution to the joint equilibrium equations. Determinate solutions are obtainable only with significant simplification of joint function or anatomy. In contrast, the optimization method of solving the general joint distribution problem does not require such simplification. Rather, it retains many of the anatomical complexities incorporated in defining the problem, and seeks an optimum solution (i.e., a solution that maximizes some process or action). Optimization techniques may be divided into linear and

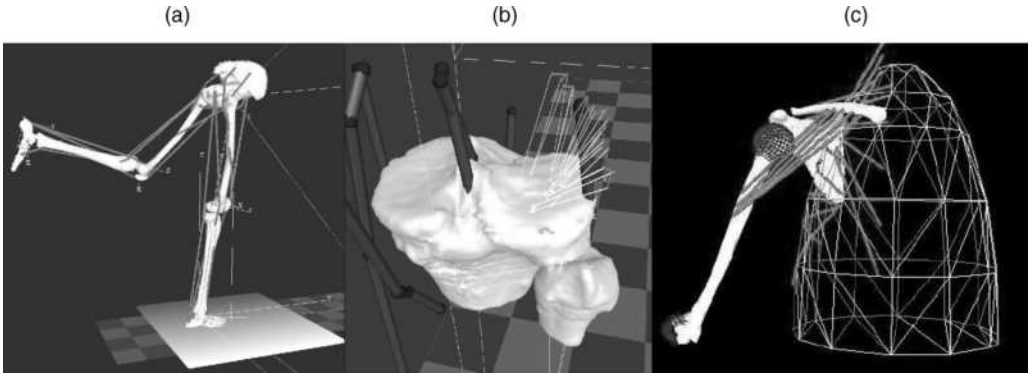


Figure 12. Forward and inverse analysis driven mathematical models, (a) the Strathclyde model animated using SIMM Musculographics-Motion Analysis Software. (From Ref. (77)) (b) Force distribution method-inverse dynamics, the vector of bony contact force is shown on the tibial plateau. (From Refs. (71) and (77)). (c) the Delft forward and Inverse shoulder model (78).

nonlinear methods Simultaneous use of linear cost and constraint functions in the model formulation constitutes the so-called linear optimization method. In contrast, if the cost function and/or one or more of the constraint functions is nonlinear, solution of the problem requires the use of nonlinear optimization methods. Both of these optimization methods have been made practical with high speed computers, since the techniques require many iterations of equation-solving to find an optimum solution. Neuromuscular function is complex and poorly understood at the conceptual, if not detailed level. However, it is generally assumed that physiological functions are optimized in some way. In 1836, the Weber brothers commented that we walk and run in "the way that affords us the least energy expenditure for the longest time and with the best results". Experimental evidence suggests that oxygen consumption (and presumably energy expenditure) is minimal at freely selected walking speeds, supporting the assumptions of optimal physiological neuromuscular function.

The optimization criteria that the neuromuscular control system "chooses", either consciously or unconsciously, to select muscle action may vary considerably with the nature of the physical activity to be performed and the physical capabilities of the individual. For example, muscle control in sprint running may serve to maximize velocity, while in walking the control process may serve to maximize endurance. In a painful pathological situation, such as degenerative joint disease, muscular control may serve to minimize pain. If this pain occurs due to the joint surface pressure, the appropriate optimization criterion may be to minimize to bony contact force. Muscular control may also serve to minimize the forces transmitted by passive joint structures such as ligaments. These examples indicate that there are possible optimization criteria to choose from, and the choice of a criterion to solve a particular distribution problem may not be obvious.

Given that gait presents a relatively unambiguous performance criterion, one can claim that it fits into the framework of optimum control theory. Additionally, gait presents a characteristic bilateral symmetry that leads to relatively simple representation of the dynamic system. Activity in the stance phase of gait is described by equality constraints on the "states", knowing that each gait stage involves dynamic constraints that reflect the particular nature of the phasic activity. However, our motivation in the use of optimum control for the study of movement relies upon the belief that it is currently the most sophisticated methodology available for solving extremely complex problems. Optimal control theory requires not only that the system dynamics are precisely determined and formulated but also that an appropriate performance criterion is chosen. Therefore, deficiencies in the modeling that are present in either the system dynamics or the performance criterion are indicated by differences between model and experiment.

Forward Analysis. The technique of forward simulation allows the study of the causal relationship between forces acting on a mechanical structure and the resulting movement (79). A first approach requires the description of joint moments, initial positions and velocities for each body segment as input variables. The forward dynamics problem then is expressed as a set of differential equations of motion with associated restraints and solved to yield angular accelerations. Angular velocities and displacements are then determined by integration (80). Modifications to input variables, either joint moment profiles or initial segmental configuration, can then be introduced and the resulting changes in the movement pattern evaluated.

Due to recent improvements in musculoskeletal modeling, forward simulation driven by muscle activity rather than joint moments is now possible. The definition of muscle activation sequences and the initial configuration of the mechanical system (angles and angular velocities) are entered into the forward simulation. A physiological model describing muscle excitation characteristics as well as muscle mechanics is used for calculating the individual muscle force production and the resulting motion is calculated.

Because of the inherent complexity of the musculoskeletal model and the multiple interaction of parameters, forward simulation alone is unlikely to contribute to the

identification of aberrant muscle action which affects timing or force production, causing an aberrant gait pattern. Matching the system response to the joint movement profiles observed in an individual patient would require extensive trial and error with no guarantee of ever finding the parameter setting responsible for the aberrant joint movement.

Optimization has been used previously as a curve-fitting tool in addition to forward simulation techniques to reproduce the movement pattern observed in an individual subject. In the studies of Chao and Rim (81), and also Townsend (82), optimization techniques were used to determine the applied joint moments by iteratively varying them until the theoretical limb displacement fits those measured in the laboratory. Chou et al. (83) predicted the minimum energy consumption trajectory of the swing limb using a similar approach. Using approximate muscle-force and/or joint moment trajectories, Pandy and Berme (84,85) evaluated body segment motion and ground reaction forces during single stance, based on a 3D model incorporating 7 DOF. In Jonkers et al. (76), initial inputs to the forward simulation process were the normalized quantified muscle activation patterns of 22 muscles, and the initial segmental configuration (both angles and angular velocity) derived from Winter (86). Two distinct musculoskeletal models (one including 6 DOF, the other 7 DOF) were defined and a muscle driven forward simulation was implemented. A series of optimization sequences then were executed to modify the muscle activation patterns and initial segmental configuration, until the system output of the forward simulation approximated the angle data reported by Winter (86). The accuracy and effectiveness of the analysis sequence proposed and the model response obtained using two distinct musculoskeletal models were verified and analyzed with respect to the kinesiology of normal walking.

Based on the integrated use of optimization and forward simulation techniques, a causal relation between muscle action, initial segmental configuration and resulting joint kinematics during single limb stance phase of gait was successfully established for a musculoskeletal model incorporating 22 muscles and 7 DOF (Fig. 12). Despite the inherent simplifications of the planar models used in this study, several kinesiological principles of normal walking were confirmed by the analysis and a reference base for exploring the causal relation between muscle function and resulting movement pattern was established.

Finite Element Analysis of Human Joints. Adaptation of finite element analysis (FEA) to the analysis of stress in human joints, requires fundamental studies in the following areas: (*1*) three dimensional FEA of moving contact problems utilizing finite deformation laws (linear elastic versus biphasic) for cartilage and non-Newtonian laws for synovial fluid; (*2*) automated adaptive methods for the generation and control of 3D computational models using error estimates and controls that account for nonlinearities, singularities, and boundary layer effects; (*3*) improvements in geometry of FE models by using patient specific MRI and CT imaging data. Recently available semiautomatic 3D mesh generation tools (reconstructed mesh from the volumetric imaging data), and numerical methods for processing the material and geometric data required for the contact analysis considerably reduce the time requirements of such efforts; (*4*) implementation of recently available high accuracy 3D joint *in vivo* kinematics to calibrate the FE models, and to assist in simulating strenuous activities; (*5*) Parallel solution algorithms for the nonlinear time-dependent problems utilizing high performance computer architectures.

Several models of articular cartilage have been proposed to describe its mechanical behavior; however, none of these models have been able to address the full spectrum of this tissue's complex mechanical responses. Typically, these models implement a subset of known tissue behavior, for which material properties must be determined from experiments. Experimental results indicate that cartilage exhibits flow-dependent viscoelasticity, anisotropy, and tension compression nonlinearity due to its ultrastructure and composition (87). These characteristics are further compounded by the inhomogeneity of the tissue through its thickness. It is widely accepted that the time-dependent response of cartilage can be accurately represented by the biphasic theory derived by Mow et al. (21). Under isotonic conditions, this biphasic theory of incompressible solid and fluid phases is appropriate for most applications involving cartilage modeling for infinitesimal or finite deformations (26,88). Numerical methods are required to solve these nonlinear problems, even for relatively simple geometries. Linear 3D elements (89) and nonlinear 3D formulations have been presented for the biphasic theory (90–92). However, full nonlinear 3D analysis for joints of realistic geometry and strains remains a computationally challenging problem.

Hirsch (1944) proposed to use the Hertz contact theory for contacting elastic spheres to model cartilage indentation (28). Askew and Mow (1978) analyzed the problem of a stationary parabolically-distributed normal surface traction acting on a layered transversely isotropic elastic medium to assess the function of the stiff surface layer of cartilage (29). More recently, complex, biphasic models have been developed that target slow, quasistatic loading response (89,90,93–95). Under high loading rates, however, the cartilage demonstrates a mechanical behavior that does not deviate substantially from a linear elastic model (38,96), so complex and numerically unwieldy biphasic models may not be necessary for rapid loading events. Eberhardt et al. (1990) (36) and (1991) (97) developed a solution for the contact problem of normal and tangential loading of elastic spheres, with either one or two isotropic elastic layers to model cartilage. A modified Hertzian (MH) theory that takes into account the thickness effect of the elastic layer has been used for articular joint contact analysis by several investigators (23,32,37,97). The results from the studies listed above have demonstrated that considerable differences may be found in cartilage stress predictions depending on the particular cartilage constitutive model being employed, that is linear elasticity theory versus linear biphasic theory. In view of the above, it has been proposed that it is adequate to use a transversely

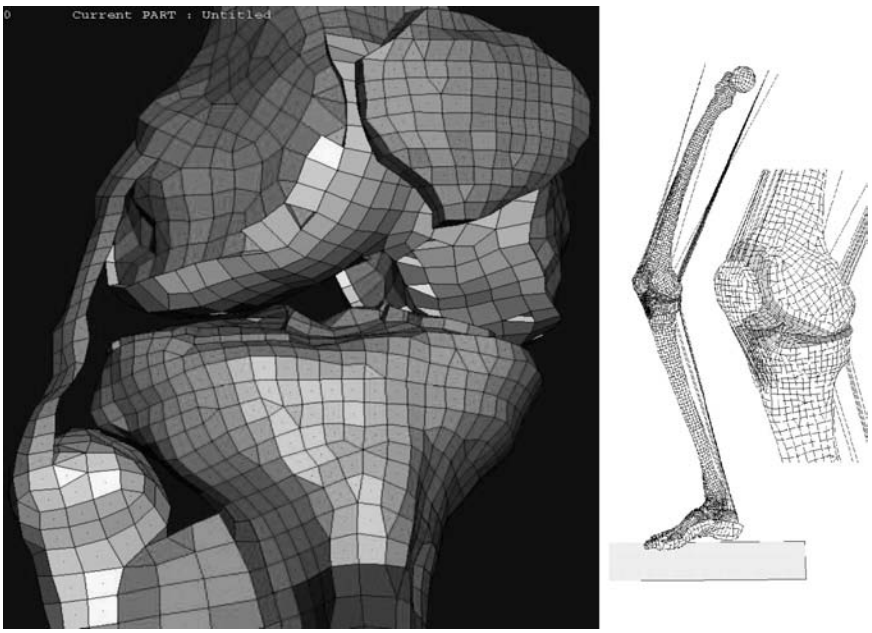


Figure 13. (a) Knee FE model components shown in different color-implicit formulation (102) (b) the knee FE model during simulation of single leg-hopping-explicit formulation (103).

isotropic, linearly elastic, homogeneous constitutive relationship to model the normal contact pressure distribution of the tibial plateau (33,98).

In the case of FE modeling applied to impacts and very rapid movements, model formulation problems must be accounted for by the solution. Implicit techniques are not widely used as they can be limited by problems of convergence. This method does not use kinematics to verify whether contact occurs at the calculated points at the associated load levels. For those applications, explicit techniques have proven useful in their ability to simulate deformations and contact conditions simultaneously as well as large segmental displacements (70,99–101) (Figs. 13b and 14). Explicit finite element analysis was also proven to be a valuable tool when simulating total knee replacement motions due to loads applied by a knee simulator (44,45). Stability of solution and low computational cost are the two main advantages of the explicit methods over classic implicit techniques when forces are used to drive the models (44,98,103–109). The main limitation of the explicit technique is that the method is only conditionally stable. Combination of both techniques is suggested so that the explicit formulation will be used to "educate" the implicit algorithms. Implicit formulations are used when the localized effect of articular stress needs to be addressed at the cartilage level with increased subject specific refined geometry since these techniques are more appropriate for the task (Fig. 13a).

Toward Patient-Specific and Task-Dependent Morphological FE Models. In order to apply FE computational methods to problems of joint mechanics, precisely measured anatomical data must be used to construct a 3D solid model from which the appropriate finite mesh can be constructed and analysis performed (110,111). Computed tomography (112) and MRI (113,114), or reshaped digital calipers and machine–controlled contact digitization (23) are commonly used methods to obtain soft–hard tissue geometry. Clinical modalities of these imaging techniques have resolutions that are typically limited to 500 μm. Considering that the articular cartilage of the knee joint is only $\sim$ 4 mm thick (14), a resolution of 500 μm is generally inadequate and special protocols are required for the imaging procedure (19). Note that all properly calibrated 3D FE knee joint models to date, have been developed from images of cadaveric knees (43,115). Models with patient specific geometry are possible using

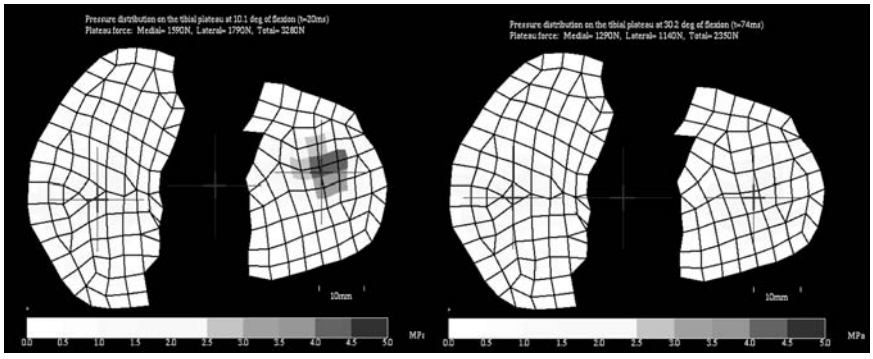


Figure 14. Tibial plateau pressure at 10 and 30 degrees flexion during single-legged hopping (103).

combined appropriate MR and CT imaging modalities that will result in increased model detail (100).

With recent technological advances in 3D imaging, computer hardware, and FE formulations, it has become possible to estimate human tissue properties (100,103). Estimation of the stiffness of hard tissue (i.e., cancellous bone) using large-scale finite element (LSFE) models or microFEs at both the apparent and trabecular levels is therefore possible (47,52,70,116–122). This direct analysis approach has advantages over traditional statistical methods in that (*1*) all morphometric measurements available from the 3D image data are contained in the FE model, and (*2*) the FE model of the structure is a mechanistic rather than statistical approach to the prediction of biological tissue properties. Application to these techniques in the clinical environment with *in vivo* data is gaining considerable ground. These recent improvements in imaging, allow the generation of 3D models in a semi-automatic fashion (23,69,123,124). Models now provide the opportunity to study the mechanical effects of individual geometrical variation (125). Such ***subject specific finite element models*** are mainly useful under two conditions: (*1*) In the statistical analysis, if the differences between subjects are large relative to the differences in model predictions related to the choice of input parameters that can only be estimated. (*2*) In a contact FE problem with high accuracy 3D kinematics refined meshes are needed in the vicinity of the contact zone (114).

The inherent ability of FEM to overcome geometrical and topological modeling problems is no longer burdened by some of the limitations apparent in early studies. Such limitations have been overcome by recent computational and experimental advances. These limitations included insufficient accuracy of *in vivo* kinematics–kinetics input data, inadequate hard–soft tissue imaging, excessive computational cost and the inherently complex process of material characterization in relation to some of the components of the stress tensor in biological material.

JOINT STABILITY

Newton's third law states that forces always exist in pairs that are equal and opposite in direction, such that if one body pushes against another, the second body will push back against the first with a force of equal magnitude if the state of motion is constant. This would mean that in order for muscles to be able to pass movement on to a limb segment, there has to be an interaction between the segment and another bone, and a joint structure that will allow the desired rotational direction and force. Such a mechanism prevents translational movements, which can "dislocate" one rigid bony limb segment from another.

Bones, ligaments, and muscles all contribute to joint stability. The extent each structure participates in maintaining joint stability differs between joints. Because bones are rigid relative to the other tissues in the joint, they can provide great stability to the joint. In general, the greater the circumference of the bony segment enclosed by its counterpart, the greater the amount of translational stability that exists in the joint. For example, the femoral head is almost completely enclosed by the acetabulum in the hip joint, whereas the humeral head is only slightly enclosed by the glenoid in the shoulder. It is obviously easier to dislocate the shoulder than the hip.

Ligaments, much like cables, can restrict rotational or translational motions depending on their location and the direction of the force. For example, the cruciate ligaments limit the anterioposterior translation of the tibia on the femur at the knee joint, while talocalcaneofibular ligaments prevent rotational motions as well as translational motions between talus, fibula and calcaneus. However, ligaments, unlike bones, cannot provide rigid constraints as they are relatively soft and flexible.

Muscles–tendons are also soft and flexible and, like ligaments, provide nonrigid constraints to the joint. A significant difference between the two, however, is that muscles are active in controlling the joint motion whereas ligaments are only passive stabilizers. Muscle action is usually complementary to that of ligaments and, in fact, can protect ligaments from damage or can protect the joint from further damage in case that the ligament is ruptured. Tensile forces exerted on muscles are counterbalanced by compressive forces across the joint surfaces providing stability to the joint against forces acting to open up the joint space. Thus, muscles can support or limit motions and serve the dual function of providing desired movements while contributing to joint stability.

The Hip Joint

The hip joint is the link of the upper body and the pelvis–trunk with the lower limbs, the main locomotion facility of the body. It is a ball-and-socket joint (Table 4) in which the head of the femur resides in the acetabulum of the pelvis, making one of the largest and most stable joints in the body. The surface area and the radius of curvature of the articular surface of the acetabulum closely match that of the articular surface of the femoral head. The hip joint is structurally a highly constrained joint. Because of the inherent stability conferred by its bony architecture, this joint is well suited for performing the weightbearing supportive tasks that are imposed on it.

The femoral ball is embraced by the acetabular socket, allowing rotation to occur with virtually no translation. The cartilage that covers the acetabulum thickens peripherally (126). A plane through the circumference of the acetabulum immediately at its opening would project with the sagittal plane intersections at an angle of 40° (opening posterior) and 60° (opening laterally). This architectural constraint imparted by the bony shapes almost eliminates the need for ligamentous and soft-tissue constraints to maintain the stability of the hip articulation. Although this increased constraint provides stability to the hip, there is a structural drawback. Such constraint limits the global range of motion of this joint at the fulcrum of the lower extremity. Fortunately, human biomechanics and everyday tasks performed by the hip do not violate these limitations and the hip's range of motion is very rarely subjected to extremes. During most ambulatory activities, such as normal bipedal locomotion, the lower extremity is positioned anteriorly in the sagittal plane with only small rotations necessary in the other two

planes. Hip flexion of at least 120°, abduction of at least 20°, and external rotation of at least 20° are necessary for carrying out normal daily activities. Activities such as descending stairs sitting, rising from a chair, and dressing require greater degrees of flexion and rotation at the hip joint. For example descending stairs requires 36° of motion whereas squatting requires 122° of motion in the sagittal plane (127).

The hip joint reaction force (at a neutral position) can reach three times body weight (BW) in single legged stance, but could get up to six times BW during the stance phase of gait and increases significantly with gait velocity. The main mechanism influencing this magnitude is the ratio of the abductor muscle force and the effect of the gravitational force moment arms. The rule normally suggests greater reaction forces expected at low ratios (128). Bracing and use of a cane can decrease the hip joint force.

The Knee Joint

The knee consists of a two joint structure: the femorotibial joint and the patellofemoral joint. The femorotibial joint is the largest joint in the body and is considered to be a modified hinged joint containing the articulating ends of the femur and tibia. The patellofemoral joint consists of the patella, the largest sesamoid bone, and the trochlea of the femur. Taken together, the knee joints function to control the distance between the pelvis and the foot as a control link. Because of the role of the knee in weight-bearing, its surrounding of very strong musculature and its location between the two longest bones in the body, tremendous forces are generated across it. Surface motion occurs simultaneously in both sagittal and the transverse plane with the first being the dominant plane of motion (129). During activities such as running, landing, and pivoting, the knee functions to maintain a given leg length and acts as a shock absorber. During stair climbing, crouching, and jumping, large propulsive forces at the knee (several times BW) are generated to control the degree and speed of shortening and lengthening of the leg. In these situations, knee stability is a dynamic process maintained through fixed bony and ligamentous constraints and modified by the action of the muscles crossing the joint. The higher the rate of the loading the more demanding the role of the knee in sustaining stability. The bony morphology of the femur, tibia, menisci, cruciate–collateral ligaments, and patella contribute to joint stability, but to a lesser extent than that of other more constrained joints such as the hip. Static constraints play a very significant role in knee stability as compared with the shoulder for example where stability is maintained through the dynamic action of the surrounding muscles. The cruciate and collateral ligaments are the major structures limiting motion at the knee. The posteromedial and posterolateral capsular complexes augment the four primary ligaments, with the menisci playing a lesser role. The muscles crossing the knee contribute to dynamic stability and are particularly important in the presence of pathologic laxity. A method that describes the constraining mechanism of the anterior and posterior ligament has used the notion of the 2D "four bar linkage". The instant center of rotation, designated primarily by the femoral condyle surface shape, follows a semicircular pathway, and the direction of displacement of the femorotibial contact points is tangential to the surface of the tibia, indicating gliding throughout the range of motion. However, the axis of rotation at the knee does not remain fixed during flexion. Indeed, as the knee flexes the screw axis will sweep out a ruled surface in space, known as the axode. This fluctuation in the screw axis signifies that the knee is not truly a hinge joint, for which the axode would degenerate to a fixed line in space. Usually, the knee is approximated as a hinge joint, a simplification that may be acceptable for flexion angles between 45 and 90° where the moving screw axis remains very close to the line passing through the centers of curvature of the two posterior femoral condyles. The motion at the articular surfaces is not one of pure rolling, but a combination of rolling and sliding as indicated by the screw axis that never lies near the articular surfaces of the tibiofemoral joint.

The Foot Structure; the Ankle Joint

The ankle joint can be described as a saddle-shaped lower end structure of a long bone (tibia and fibula) Fig. 15. Its inferior transverse ligament encloses the superior aspect of the body of the talus (the trochlea). It is the joint that first receives the transient impact that travels through the tibia in gait or other movement. It alternates in both form and function to receive load as a shock absorbing

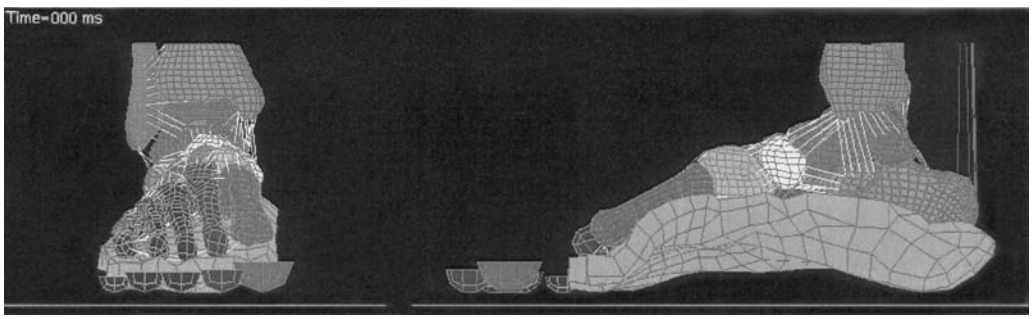


Figure 15. The Wayne State University Ankle joint and foot Finite element model (130).

mechanism and to propel significant leverage based force during fast locomotion. The subtalar and ankle joint act like a mitered hinge. The tibial surface forming the superior dome of the ankle is concave sagittally, is slightly convex from side to side, and is oriented $\sim 93^{\circ}$ from the long axis of the tibia (it is higher on the lateral than the medial side). The upper articulating surfaces of the talus appear to match closely that of the cavity formed by the tibiofibular mortise. The superior part of the body of the talus is wedge-shaped. It is about one-fourth wider in front than behind, with an average difference of 2.4 mm; anteriorly a minimal difference of 1.3 mm and a maximal difference of 6 mm. From front to back, the articular surface spans an arc of $\sim 105^{\circ}$. This surface contour, having a smaller diameter medially than laterally, has been compared to a section or rostrum of a cone. The primary motion of the ankle joint is dorsiflexion–plantarflexion. Its axis of rotation is obliquely oriented with respect to all three anatomic planes with ankle dorsiflexion and tibial internal rotation being associated with subtalar eversion (pronation) whereas the ankle plantarflexion and tibial external rotation are associated with subtalar inversion (supination). The axis extends from anterior, superior, and medial to inferior, posterior, and lateral as it is passing through the inferior tips of the malleoli. It is at angles of 93° with respect to the long axes of the tibia and $\sim 12^{\circ}$ to the joint surface. However, rather than a true single ICR, the ankle has been noted to have multiple instant centers, all of which perturbate and fall very close to a single point within the body of the talus. Through a complete arc of ankle rotation, the center may be displaced anywhere from 4 to 7 mm. The oblique orientation of the axis of rotation to the sagittal, coronal, and transverse planes, translation of the talus in the mortise can occur in all three directions. The talus has been observed *in vitro* to rotate easily in the ankle mortise implying relative movement between the malleoli. Because the trochlea is wider anteriorly than posteriorly, it has been suggested that lateral play of the talus within its mortise occurs only when the ankle is in plantarflexion. Subtalar motion has been described as screw-like influencing the flexibility of the transverse tarsal joint. Others suggest that instability exists in dorsiflexion, while others yet believe that with intact ligaments translation, occurs only in the sagittal direction. These differences can be explained by behavior of >100 ligaments and by the roles played by the subtalar joint, the kinematic chain of the hindfoot, and the muscles that traverse this area in transmitting forces across this area during plantarflexion and dorsiflexion. The talus is unique because this bone has seven articulations that connect it to four other bones, it lies between the foot and the leg, and contains no muscular attachments. The stability of the talus and its articulations, therefore, relies heavily on the ligamentous attachments and musculotendinous complexes that traverse the talus and attach distally. Therefore the main characteristic of ankle joint is its strong passive stability attributed to a variety of factors. First is the bony stability provided by contact of the trochlea with the tibial plafond. Second are the medial and lateral cartilaginous slightly concave surfaces that articulate with the two malleoli. Third are the ligamentous connections between the tibia, fibula, talus, and calcaneus. Ankle stability increases during weight bearing and depends more on articular surface congruency.

The tarsometatarsal joints are relatively mobile and intrinsically stable joints that produce the arch-like configuration allowing wide range of motion at the first metatarsophalangeal joint with gliding during most of its range and jamming artful extension. The medial longitudinal arch functions as a beam and a truss.

The mode of foot–ankle mobility and muscular control is the most significant determinant of both limb stability and body progression (131).

Standing barefooted we load our heels with two-thirds of the load while the other third is loading the forefoot. During walking and at the early phase of stance, the center of pressure moves from the posterolateral heel rapidly across the midfoot, a phenomenon coupled with the firing of the anterior tibial musculature to slow foot plantarflexion and prevent foot slap. Then, at mid- and late stance the posterior calf musculature fires, propelling the body over the foot towards toe-off phase where the hallux bears the most pressure.

At higher speeds/rates of mobility it is the intrinsic mechanical properties that provide control rather than the neuromuscular control system. The force at the ankle joint can reach magnitudes up to six times body weight during walking and thirteen times BW during running. The heel fat pad is a very effective shock absorbing mechanism and it has been shown that high heels, or narrow shoes, narrow toebox, can lead to altered foot mechanics that ultimately result in foot deformities and pain.

The Spine

The spine is composed of a series of vertebrae (7 cervical, 12 thoracic, 5 lumbar, the sacrum and coccyx) and intervening soft tissues, such as intervertebral disks, ligaments, and muscle attachments. It provides important functions including support of the body structure and protection of vital tissues such as the spinal cord, nerves and arteries. Yet it is flexible and allows mobility to the torso. Several studies have described the passive and active range of motion of spinal segments differently (Table 4). The motion of the spine is usually analyzed through consideration of motion segments that are comprised of two adjacent vertebrae with the intervertebral disk and other intervening soft tissues. These structures, also called functional spinal units (FSU), move with 6 DOF, however, the motion is quite complex due to six articulate faces between the two bony segments and attachment of multiple ligaments and muscles. Normally simultaneous translations and rotations of FSU are coupled in the analyses (134). Creep, relaxation and hysteresis are the three prevailing viscoelastic characteristics of the intervertebral disks (1). At high rates of loading the disks serve as a shock absorber with compression strength being higher from upper cervical to lower lumbar levels. Although this is true for most joints, vibrational properties of spinal segments have attracted particular attention as they are thought to be related to injury and pain. The spinal cord exhibits some longitudinal

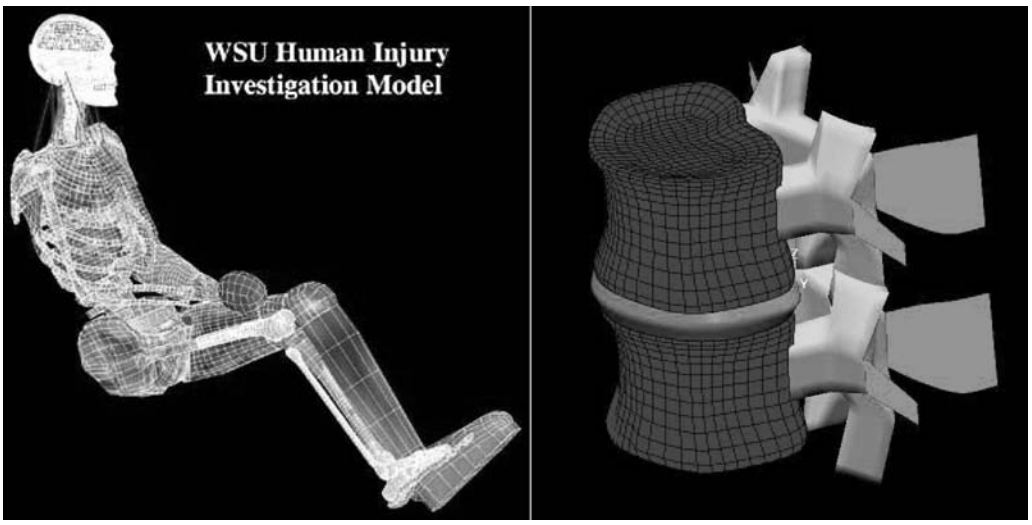


Figure 16. The Wayne State University full body impact model and the spine model (130,135).

elasticity but very poor axial translation. These translational forces are the ones primarily associated with neurological injury. Apparently muscular support is of vital importance is stability. If for example bilateral facet resection exceeds 50% the significant increase in annulus stresses may occur. Seatbelts can adequately protect front seat occupants against frontal impact in an airbag equipped vehicle (135).

Several models of the spine have been developed (130,135). A nonlinear finite element model of lumbar spine segment L3–L5 is shown in Fig. 16. The effects of upper-body mass, nucleus injury, damping, and different vibration frequency loads were analyzed for the whole body vibration (WBV) using this model. Anterior regions of L3–L5 segment show small vibration amplitudes, but posterior regions show large amplitudes. The posterior regions of intervertebral discs of lumbar spine are easy to injury during long-term WBV compared with anterior regions. The vibration of the human spine is more dangerous to facets, especially during WBV approximating a sympathetic vibration, which may lead to abnormal remodeling and disorders of lumbar spine.

The Shoulder

The shoulder complex allows the arm to move with respect to the thorax. The biomechanics of the shoulder involves the study of four different articulations, namely, the acromioclavicular, sternoclavicular, glenohumeral joints, and the scapulothoracic joints. The shoulder complex has the greatest range of motion among joints in the body (Table 4). Traditional descriptions of humeral motion are based on the angle between the humerus and the thorax in the sagittal plane (flexion and extension) and the coronal plane (abduction), with one very common characteristic: the glenohumeral articulation is inherently unstable due to the shallow nature of the glenoid fossa, containing only one-third of the diameter of the humeral head. The ligaments, capsular and muscular structures are the main stability mechanisms. Axial rotation of the humerus is conventionally described by the degrees of internal or external rotation when the humeral axis is parallel to the thorax, or perpendicular to the thorax (abducted 90°). The primary anterior stabilizers of the shoulder when the arm is abducted at 90° are the inferior glenohumeral ligaments. Horizontal abduction and adduction, also known as horizontal extension and flexion, respectively, are commonly used to describe arm position when the axis of the humerus is perpendicular to the thorax. The muscles surrounding the structure produce a barrier effect by applying compressive forces and by eccentric contraction resulting is active stability. The level of elevation of the normal arm is 167–168° and 171–175° for men and women, respectively. Average extension or posterior elevation is approximately 60°. When the arm is adducted by the side of the body, it is possible to achieve approximately 180° of rotation. With abduction of the arm to 90°, however, the total arc of rotation is reduced to 120°. The range of motion of the shoulder decreases with normal aging. The maximum loads that are characteristic for the glenohumeral joint can reach magnitudes of one and a half body weight (132).

The Elbow

The elbow joint consists of the proximal end of ulna and the capitulum of the humerus. Three articulations are present: Humeroulnar, humeroradial, and proximal radioulnar (133). It is a hinge-type joint whose functions include positioning and stabilizing the hand, providing lever-type support during lifting and weight-bearing during activities, such as a pommel horse exercise. The humerus-ulna joint itself has only one degree of freedom, however, the rotational motion of the radius over ulna provides another degree of freedom to the elbow. Thus the elbow can be considered to have two degrees of freedom. Basically its changing center of rotation during flexion-extension makes it more than a hinge joint: flexion-extension (0–160°)

and axial rotation, or pronation–supination (70–80° and 80–85°, respectively). The functional range of motion for most activities of daily living is 100° for flexion–extension (30–130°) and pronation–supination (±50°) at the elbow joint. The instantaneous axes of rotation for flexion–extension are localized at the center of the lateral projected curvature of the trochlea and capitulum. The size of this region measures only 2–3 mm.

The axis of forearm rotation passes through the capitulum and head of the radius, extending to the distal ulna. The carrying angle, the angle between the long axis of the humerus and the long axis of the ulna, averages 7° in men and 13° in women.

Motion in the plane of varus–valgus rotation at the elbow indicates joint instability and such motions are restricted. The rotational forces are mainly resisted by the anterior and posterior oblique fibers of the medial collateral ligament; the former is stretched during flexion and extension, whereas the latter is stretched only during flexion. In addition to the medial collateral ligament, the shape of the articular surface and the anterior capsule contribute to the resistance to valgus stress while the articular congruity and the radial head provide stability in flexion. The same structures resist valgus stress in extension as well. It is the lateral collateral ligament, anconeus, and joint capsule that provide stability under varus stress. Force generated at the elbow can reach up to three times body weight in everyday activities (136).

The Wrist

The wrist or carpus is a very complicated joint consisting of multiple articulations. It provides a stable support for the hand, allowing for the transmission of grip forces as well as positioning of the hand and digits for fine movements. Wrist position affects the ability of the fingers to operate (flex-extend) and to effectively grasp. The main function of the wrist is to fine-tune grasp by controlling the length–tension relationship in the extrinsic muscles to the hand (137). The hand is the principal instrument of touch and its combination of sensibility and motor function has classified it as the ultimate supportive organ of information, accomplishment, and evolution. The self-selected wrist position for maximal power grip has been shown to be 35° of extension with 7° of ulnar deviation. In addition, full wrist flexion deteriorates the efficiency of the finger flexors and the grip strength to 25% of that available when the wrist is in extension. Martial arts experts have exploited this knowledge for a long time to disarm assailants. When the wrist is positioned in flexion, the combination of a passive extensor tenodesis effect, increasing resistance to flexion, and decreasing excursion of the flexors results in a weakened grip. The assailant, therefore, is unable to maintain a grip on the weapon. The strength of the finger flexors is more than twice that of extensors. The unique feature of the metacarpophalangeal (MCP) joints is their structural asymmetry, which assisted by the ligament positions and the bony configuration of the metacarpal heads results in effective stabilization and finely tuned motion.

OVERVIEW

Human diarthroidal joints can support high levels of mechanical load for several decades, yet degenerative joint diseases occur in epidemic proportions every year. Understanding the role of mechanics in diseases, such as osteoarthritis, requires analysis of the behavior of both healthy and pathological joints under a full range of physiological loading conditions. Relationships between joint function, meniscus injury, and osteoarthritis after trauma (e.g., ACL injury–reconstruction) requires detailed knowledge of internal tissue stresses. Such knowledge is necessary to advance our understanding of cartilage–ligament–meniscus failure mechanisms at the macroscopic level, as well as mechanotransduction at the cellular level.

Experimental measurements alone are not sufficient to delineate the detailed biomechanics of the human joint. Recent publications have confirmed that mathematical modeling can be an effective tool for the simulation and analysis of complex biological structures (e.g., the human knee). However, existing modeling strategies, based on generic geometry and/or driven with simplistic kinematics and loading assumptions, have been of limited value in addressing clinical hypotheses. Research efforts in the biomechanics community have focused on cadaveric studies for the characterization of tolerances on material, geometry, and attachment parameters that will restore contact pressure distribution to within a specified deviation from accepted normal values. Models driven directly by *in vivo* kinematic data have not been possible due to the relatively poor accuracy available from traditional motion analysis methods, limiting the predictive power of FE modeling techniques in the study of knee pathogenesis. New advances in modeling problem formulations that take advantage of patient-specific geometry and unique motion analysis systems (cineradiography) for high accuracy 3D knee kinematics can overcome limitations of previous approaches. This should generate high quality estimates of internal tissue stresses, providing new insight into the role of tissue loading in the development and progression of musculoskeletal diseases such as osteoarthrosis.

BIBLIOGRAPHY

1. Maroudas A. Physical chemistry of articular cartilage and the intervertebral disc. In: Sokoloff L, ed. The Joints and Synovial Fluid. New York: Academic; 1980: 239–291.
2. Ateshian GA, Lai WM, Zhu WB, Mow VC. An asymptotic solution for the contact of two biphasic cartilage layers. J Biomech 1994;27:1347–1360.
3. Mow VC, Ratcliffe A. Structure and function of articular cartilage and meniscus. In: Mow VC, Hayes WC, eds. Basic Orthopedic Biomechanics. 2nd ed. Philadelphia: Lippincott-Raven; 1997. p 113–177.
4. Wittenburg J. Dynamics of Systems of Rigid Bodies. Stuttgart, Germany: B.G. Teubner; 1977.
5. Woltring HJ. Planar control in multi-camera calibration for 3-D gait studies. J Biomech 1980;13:39–48.
6. Woltring HJ. Representation and calculation of 3-D joint movement. Human Movement Sci 1991;10:603–616.
7. Goldstein H. Classical Mechanics. Reading (MA): Addison-Wesley; 1980.

8. Dempster WT. Space Requirements of the Seated Operator. Wright Patterson Air Force Base, OH, 1955.
9. Drillis R, Contini R, Bluestein M. Body segment parameters; A survey of measurement techniques. Artif Limbs 1964;25: 44–66.
10. Full RJ, Koditschek DE. Templates and anchors: neuromechanical hypotheses of legged locomotion on land. J Exp Biol 1999;202:3325–3332.
11. Nigg BM, Herzog W. Biomechanics of the musculoskeletal system. West Sussex (UK): Wiley; 1999.
12. Manal K, McClay Davis I, Galinat B, Stanhope S. The accuracy of estimating proximal tibial translation during natural cadence walking: bone vs. skin mounted targets. Clin Biomech (Bristol, Avon) 2003;18:126–131.
13. Tashman S, Anderst W. Skin motion artifacts at the knee during impact movements. Seventh Annu Meet, Gait and Clinical Movement Analysis Soc. Chattanooga (TN); 2002.
14. Tashman S, Anderst W. Skin motion artifacts at the knee during impact movements. Gait Posture 2002;16:11–12.
15. Alexander EJ, Andriacchi TP. Correcting for deformation in skin-based marker systems. J Biomech 2001;34:355–361.
16. Andriacchi TP, Toney MK. A point cluster method for in vivo measurement of limb segment movement. Adv Bioeng 1994; 28:185–186.
17. Fleming BC et al. The strain behavior of the anterior cruciate ligament during stair climbing: An in vivo study. Arthroscopy 1999;15:185–191.
18. Shepherd DE, Seedhom BB. Thickness of human articular cartilage in joints of the lower limb. Ann Rheum Dis 1999;58: 27–34.
19. Tashman S, Anderst W. In-vivo measurement of dynamic joint motion using high speed biplane radiography and CT: Application to canine ACL deficiency. J Biomech Eng 2003;125:238–245.
20. Jurvelin JS, Arokoski JP, Hunziker EB, Helminen HJ. Topographical variation of the elastic properties of articular cartilage in the canine knee. J Biomech 2000;33: 669–675.
21. Mow VC, Kuei SC, Lai WM, Armstrong CG. Biphasic creep and stress relaxation of articular cartilage in compression? Theory and experiments. J Biomech Eng 1980;102: 73–84.
22. Mow VC, Lai WM, Holmes MH. Advanced theoretical and experimental techniques in cartilage research. Biomechanics: Principles and Applications. Boston: Martinus Nijhoff Publishers; 1982. p 47–74.
23. Fregly BJ, Sawyer WG. Estimation of discretization errors in contact pressure measurements. J Biomech 2003;36: 609–613.
24. Harris ML, Morberg P, Bruce WJ, Walsh WR. An improved method for measuring tibiofemoral contact areas in total knee arthroplasty: A comparison of K-scan sensor and Fuji film. J Biomech 1999;32:951–958.
25. Wu JZ, Herzog W, Epstein M. Articular joint mechanics with biphasic cartilage layers under dynamic loading. J Biomech Eng 1998;120:77–84.
26. Vermilyea ME, Spilker RL. Hybrid and mixed-penalty finite-elements for 3-D analysis of soft hydrated tissue. Int J Numer Methods Eng 1993;36:4223–4243.
27. Prendergast PJ, van Driel WD, Kuiper JH. A comparison of finite element codes for the solution of biphasic poroelastic problems. Proc Inst Mech Eng [H] 1996;210:131–136.
28. Hirsch C. A contribution to the pathogenesis of chondromalacia of the patella. Acta Chir Scand 1944;83:1–106.
29. Askew MJ, Mow VC. The Biomechanical Function of the Collagen Ultrastructure of articular cartilage. J Biomech Eng 1978;100:105–115.
30. Spilker RL, Maxian TA. A mixed-penalty finite-element formulation of the linear biphasic theory for soft-tissues. Int J Numer Methods Eng 1990;30:1063–1082.
31. van der Voet A. A comparison of finite element codes for the solution of biphasic poroelastic problems. Proc Inst Mech Eng [H] 1997;211:209–211.
32. Wu JZ, Herzog W, Ronsky J. Modeling axi-symmetrical joint contact with biphasic cartilage layers–an asymptotic solution. J Biomech 1996;29:1263–1281.
33. Donzelli PS, Spilker RL, Ateshian GA, Mow VC. Contact analysis of biphasic transversely isotropic cartilage layers and correlations with tissue failure. J Biomech 1999;32: 1037–1047.
34. Wu JZ, Herzog W, Epstein M. Evaluation of the finite element software ABAQUS for biomechanical modelling of biphasic tissues. J Biomech 1998;31:165–169.
35. Blankevoort L, Kuiper JH, Huiskes R, Grootenboer HJ. Articular contact in a three-dimensional model of the knee. J Biomech 1991;24:1019–1031.
36. Eberhardt AW, Keer LM, Lewis JL, Vithoontien V. An analytical model of joint contact. J Biomech Eng 1990;112: 407–413.
37. Hirokawa S. Three-dimensional mathematical model analysis of the patellofemoral joint. J Biomech 1991;24: 659–671.
38. Oloyede A, Flachsmann R, Broom ND. The dramatic influence of loading velocity on the compressive response of articular cartilage. Connect Tissue Res 1992;27:211–224.
39. Buschmann MD et al. Confined compression of articular cartilage: linearity in ramp and sinusoidal tests and the importance of interdigitation and incomplete confinement. J Biomech 1998;31:171–178.
40. Bursac PM, Obitz TW, Eisenberg SR, Stamenovic D. Confined and unconfined stress relaxation of cartilage: appropriateness of a transversely isotropic analysis. J Biomech 1999;32:1125–1130.
41. Haut Donahue TL, Hull ML, Rashid MM, Jacobs CR. How the stiffness of meniscal attachments and meniscal material properties affect tibio-femoral contact pressure computed using a validated finite element model of the human knee joint. J Biomech 2003;36:19–34.
42. Beaugonin M, Haug E, Cesari D. Improvement of numerical ankle/foot model: Modeling of deformable bone. Proc 1997 41st Stapp Car Crash Conf. Lake Buena Vista (FL): SAE, Warrendale (PA); 1997. p 225–237.
43. Beillas P, et al. Limb: Advanced FE model and new experimental data. Stapp Car Crash J 2001;45:469–494.
44. Godest AC et al. Simulation of a knee joint replacement during a gait cycle using explicit finite element analysis. J Biomech 2002;35:267–275.
45. Halloran J, Petrella A, Rullkoetter P. Explicit finite element model predicts TKR mechanics. 49th Annu Meet, Orthopaedic Res Soc. New Orleans (LA); 2003. p 1312.
46. Benvenuti J-F. Modélisation tridimensionnelle du genou humain. Laboratoire de Génie Médical. Lausanne, Switzerland: EPF Lausanne; 1998.
47. Beillas P et al. Foot and ankle finite element modeling using CT-scan data. 43rd Stapp Car Crash Conf. San Diego; 1999. p 217.
48. Keyak JH, Skinner HB. Three-dimensional finite element modelling of bone: Effects of element size. J Biomed Eng 1992;14:483–489.
49. Ulrich D et al. The quality of trabecular bone evaluated with micro-computed tomography, FEA and mechanical testing. Stud Health Technol Inform 1997;40:97–112.
50. Ulrich D, van Rietbergen B, Weinans H, Ruegsegger P. Finite element analysis of trabecular bone structure: a comparison

of image-based meshing techniques. J Biomech 1998;31: 1187–1192.
51. Beillas P. Modélisation des membres inférieurs en situation de choc automobile. Laboratoire de Biomécanique. Paris, France: École Nationale Supérieure d'Arts et Métiers; 1999.
52. Noailles J. Modélisation mécanique par éléments finis de l'articulation du genou. Laboratoire de Biomécanique. Paris, France: École Nationale Supérieure d'Arts et Métiers; 1999.
53. Limbert GMT, Freeman MAR. Three dimensional finite element model of the human ACL. Simulation of a passive knee flexion cycle. Analysis of deformations and stresses. 47th Annu Meet Orthopaedic Res Soc. San Francisco (CA); 2001. p 794.
54. Hirokawa S, Tsuruno R. Hyper-elastic model analysis of anterior cruciate ligament. Med Eng Phys 1997;19:637–651.
55. Butler DL et al. Location-dependent variations in the material properties of the anterior cruciate ligament. J Biomech 1992;25:511–518.
56. Yamamoto K, Hirokawa S, Kawada T. Strain distribution in the ligament using photoelasticity. A direct application to the human ACL. Med Eng Phys 1998;20:161–168.
57. Pioletti DP. Viscoelastic properties of soft tissues: Application to knee ligaments and tendons. Departement de physique Ecole polytechnique federale de Lausanne. Lausanne: Ecole polytechnique federale de Lausanne; 1997.
58. Hirokawa S, Tsuruno R. Three-dimensional deformation and stress distribution in an analytical/computational model of the anterior cruciate ligament. J Biomech 2000;33: 1069–1077.
59. Taylor M, Tanner KE, Freeman MA, Yettram AL. Cancellous bone stresses surrounding the femoral component of a hip prosthesis: An elastic-plastic finite element analysis. Med Eng Phys 1995;17:544–550.
60. Aspden RM. A model for the function and failure of the meniscus. Eng Med 1985;14:119–122.
61. Anderst WJ, Tashman S. A method to estimate in vivo dynamic articular surface interaction. J Biomech 2003;36: 1291–1299.
62. Walker PS, Hajek JV. The load-bearing area in the knee joint. J Biomech 1972;5:581–589.
63. Fukubayashi T, Kurosawa H. The contact area and pressure distribution pattern of the knee. A study of normal and osteoarthrotic knee joints. Acta Orthop Scand 1980;51: 871–879.
64. Warner JJ. Articular contact patterns of the normal glenohumeral joint. J Shoulder Elbow Surg 1998;7:381–388.
65. Scherrer PK, Hillberry BM, Van Sickle DC. Determining the in vivo areas of contact in the canine shoulder. J Biomech Eng 1979;101:271–278.
66. Soslowsky LJ et al. Quantitation of in situ contact areas at the glenohumeral joint: A biomechanical study. J Orthop Res 1992;10:524–534.
67. Dennis DA, Komistek RD, Hoff WA, Gabriel SM. In vivo knee kinematics derived using an inverse perspective technique. Clin Orthop Relat Res 1996: 107–117.
68. Sheehan FT, Zajac FE, Drace JE. Using cine phase contrast magnetic resonance imaging to non-invasively study in vivo knee dynamics. J Biomech 1998;31:21–26.
69. Papaioannou G, Tashman S, Nelson F. Morphology proportional differences in the medial and lateral compartment of the distal femur. 50th Ann Meet, Orthopaedic Res Soc. San Francisco, (CA); 2004. p 1256.
70. Treece GM, Prager RW, Gee AH. Regularised marching tetrahedra: Improved iso-surface extraction. Comput Graph 1999;23:583–598.
71. Papaioannou G, Tashman S. Validation of a lower limb model based on 3D knee kinematics from a high speed biplane dynamic radiography. In: Fotiadis DI, Dassios G, Kiriaki K, Massalas CV, eds. Scattering Theory and Biomedical Engineering Modelling and Applications. World Scientific; 2001.
72. Ateshian GA. A B-spline least-squares surface-fitting method for articular surfaces of diarthrodial joints. J Biomech Eng 1993;115:366–373.
73. Paul JP, Poulson J. The analysis of forces transmitted by joints in the human body. 5th Int Conf Experimental Stress Analysis. Udine, Italy; 1974.
74. Hefzy MS, Zoghi M, Jackson WT, DiDio LJA. Method to measure the three-dimensional patello-femoral tracking. Adv Bioeng 1988;8:47–49.
75. Papaioannou G, Daly D, Spaepen A. Use of new optimalization tools in knee joint modelling. In: Dassios G, Fotiadis DI, Kiriaki K, Massalas CV, eds. Scattering Theory and Biomedical Engineering Modelling and Applications. Singapore: World Scientific; 2000. pp 282–295.
76. Jonkers I, Spaepen A, Papaioannou G, Stewart C. An EMG-based, muscle driven forward simulation of single support phase of gait. J Biomech 2002;35:609–619.
77. Papaioannou G. A Three Dimensional Mathematical Model of the Knee Joint. Bioengineering Ph.D. Thesis. Glasgow, (UK): University of Strathclyde; 2000.
78. Van der Helm FC, Veeger HE, Pronk GM, Van der Woude LH, Rozendal RH. Geometry parameters for musculoskeletal modelling of the shoulder system. J Biomech 1992; 25:129–144.
79. Zajac FE. Muscle coordination of movement: a perspective. J Biomech 1993;26(Suppl 1):109–124.
80. Pandy MG, Berme N. A numerical method for simulating the dynamics of human walking. J Biomech 1988;21:1043–1051.
81. Chao EY, Rim K. Application of optimization principles in determining the applied moments in human leg joints during gait. J Biomech 1973;6:497–510.
82. Townsend MA, Tsai TC. Biomechanics and modelling of bipedal climbing and descending. J Biomech 1976;9:227–239.
83. Chou LS, Song SM, Draganich LF. Predicting the kinematics and kinetics of gait based on the optimum trajectory of the swing limb. J Biomech 1995;28:377–385.
84. Pandy MG, Berme N. Quantitative assessment of gait determinants during single stance via a three-dimensional model—Part 2. Pathological gait. J Biomech 1989;22:725–733.
85. Pandy MG, Berme N. Quantitative assessment of gait determinants during single stance via a three-dimensional model—Part 1. Normal gait. J Biomech 1989;22:717–724.
86. Winter DA. The Biomechanics and Motor Control of Human Gait. Waterloo: University of Waterloo Press; 1987.
87. Huang CY, Stankiewicz A, Ateshian GA, Mow VC. Anisotropy, inhomogeneity, and tension-compression nonlinearity of human glenohumeral cartilage in finite deformation. J Biomech 2005;38:799–809.
88. Chan B, Donzelli PS, Spilker RL. A mixed-penalty biphasic finite element formulation incorporating viscous fluids and material interfaces. Ann Biomed Eng 2000;28:589–597.
89. Zhang H, Totterman S, Perucchio R, Lerner AL. Magnetic resonance image based 3D poroelastic finite element model of tibio-menisco-femoral contact. Proc 23rd Annu Meet Am Soc Biomechanics. Pittsburgh; 1999.
90. Perie D, Hobatho MC. In vivo determination of contact areas and pressure of the femorotibial joint using non-linear finite element analysis. Clin Biomech (Bristol, Avon) 1998;13: 394–402.
91. Eckstein F et al. Quantitative relationships of normal cartilage volumes of the human knee joint—assessment by

magnetic resonance imaging. Anat Embryol (Berlin) 1998;197: 383–390.
92. Eckstein F, Reiser M, Englmeier KH, Putz R. In vivo morphometry and functional analysis of human articular cartilage with quantitative magnetic resonance imaging—from image to data, from data to theory. Anat Embryol (Berlin) 2001;203:147–173.
93. Rudert MJ, et al. Articular cartilage thickness measurement with MRI and multi-detector computed tomography (MDCT). 49th Annu Meet, Orthopaedic Res Soc. New Orleans (Lo); 2003. p 0571.
94. Bendjaballah MZ, Shirazi-Adl A, Zukor DJ. Biomechanics of the human knee joint in compression: Reconstruction, mesh generation and finite element analysis. Knee 1995;2:69–79.
95. Xia Y. Magic-angle effect in magnetic resonance imaging of articular cartilage: A review. Invest Radiol 2000;35:602–621.
96. DiSilvestro MR, Zhu Q, Suh JK. Biphasic poroviscoelastic simulation of the unconfined compression of articular cartilage: II—Effect of variable strain rates. J Biomech Eng 2001;123:198-200.
97. Eberhardt AW, Lewis JL, Keer LM. Normal contact of elastic spheres with two elastic layers as a model of joint articulation. J Biomech Eng 1991;113:410–417.
98. Donahue TL, Hull ML, Rashid MM, Jacobs CR. A finite element model of the human knee joint for the study of tibio-femoral contact. J Biomech Eng 2002;124:273–280.
99. Weinans H, Sumner DR, Igloria R, Natarajan RN. Sensitivity of periprosthetic stress-shielding to load and the bone density-modulus relationship in subject-specific finite element models. J Biomech 2000;33:809–817.
100. Papaioannou G, Yang K, Fyhrie D, Tashman S. Validation of a subject specific finite element model of the human knee developed for in-vivo tibio-femoral contact analysis. 50th Annu Meet, Orthopaedic Res Soc. San Francisco; 2004. p 0358.
101. Couteau B et al. Finite element modelling of the vibrational behaviour of the human femur using CT-based individualized geometrical and material properties. J Biomech 1998;31:383–386.
102. Papaioannou G, Anderst W, Tashman S. Elevated joint contact forces in ACL-reconstructed knees: A finite element analysis driven by in vivo kinematic data. IMECE'03:2003 ASME Int Mech Eng Cong Exposition. Washington, DC; 2003.
103. Beillas P, Papaioannou G, Tashman S, Yang KH. A new method to investigate in vivo knee behavior using a finite element model of the lower limb. J Biomech 2004;37:1019–1030.
104. Li G, Lopez O. Reliability of a 3D finite element model constructed using magnetic resonance images of a knee for joint contact stress analysis. 23rd Proc Am Soc Biomech. Pittsburgh; 1999.
105. Charras GT, Guldberg RE. Improving the local solution accuracy of large-scale digital image-based finite element analyses. J Biomech 2000;33:255–259.
106. Dunbar WL, Jr., Un K, Donzelli PS, Spilker RL. An evaluation of three-dimensional diarthrodial joint contact using penetration data and the finite element method. J Biomech Eng 2001;123:333–340.
107. Spilker RL, Donzelli PS, Mow VC. A transversely isotropic biphasic finite element model of the meniscus. J Biomech 1992;25:1027–1045.
108. Morrison JB. The mechanics of the knee joint in relation to normal walking. J Biomech 1970;3:51–61.
109. Kettelkamp DB, Jacobs AW. Tibiofemoral contact area—determination and implications. J Bone Joint Surg Am 1972;54:349–356.
110. Radin EL et al. Relationship between lower limb dynamics and knee joint pain. J Orthop Res 1991;9:398–405.
111. Collins JJ, Whittle MW. Impulsive forces during walking and their clinical implications. Clin Biomech 1989;4:179–187.
112. Tashman S, Leisen JC, Sherlitz C, Radin EL. Methods for the reduction of heelstrike impulsive loading. Second World Cong Biomech. Amsterdam, The Netherlands; 1994.
113. Rolf C et al. An experimental in vivo method for analysis of local deformation on tibia, with simultaneous measures of ground reaction forces, lower extremity muscle activity and joint motion. Scand J Med Sci Sports 1997;7:144–151.
114. Tashman S, Anderst W. Internal/external and varus/valgus knee rotations are different in ACL-reconstructed and contralateral (intact) limbs during running. 49th Annu Meet, Orthopaedic Res Soc. New Orleans (LA); 2003. p 0124.
115. Papaioannou G, Fyhrie D, Tashman S. Effects of patient-specific cartilage geometry on contact pressure: An in-vivo finite element model Of ACL reconstruction. 50th Annu Meet, Orthopaedic Res Soc. San Francisco; 2004. p 1289.
116. Hollister SJ, Fyhrie DP, Jepsen KJ, Goldstein SA. Application of homogenization theory to the study of trabecular bone mechanics. J Biomech 1991;24:825–839.
117. Jacobs CR et al. NACOB presentation to ASB Young Scientist Award: Postdoctoral. The impact of boundary conditions and mesh size on the accuracy of cancellous bone tissue modulus determination using large-scale finite- element modeling. North American Congress on Biomechanics. J Biomech 1999;32:1159–1164.
118. van Rietbergen B et al. Tissue stresses and strain in trabeculae of a canine proximal femur can be quantified from computer reconstructions. J Biomech 1999;32:443–451.
119. Yeni YN, Fyhrie DP. Finite element calculated uniaxial apparent stiffness is a consistent predictor of uniaxial apparent strength in human vertebral cancellous bone tested with different boundary conditions. J Biomech 2001;34: 1649–1654.
120. Blankevoort L, Huiskes R, de Lange A. The envelope of passive knee joint motion. J Biomech 1988;21:705–720.
121. Ahmed AM, Burke DL. In-vitro measurement of static pressure distribution in synovial joints—Part I: Tibial surface of the knee. J Biomech Eng 1983;105:216–225.
122. Besnault B, et al. A parametric finite element of the human pelvis. 42nd Stapp Car Crash Conf Proc. Tempe (AZ); 1998. p 337.
123. Anderst W, Tashman S. A unique method to determine dynamic in vivo articular surface interaction. 4th World Cong Biomecha. Calgary, Alberta, Canada; 2002. p 520.
124. Anderst WJ, Les C, Tashman S. In vivo serial joint space measurements during dynamic loading in a canine model of osteoarthritis. Osteoarth Cartilage 2005;13(9):808–816.
125. Demetropoulos CK. Dynamic evaluation of contact pressure and effect of graft harvest at osteochondral donor site in the knee, personal communication. 2003.
126. Kempson GE, Spivey CJ, Swanson SA, Freeman MA. Patterns of cartilage stiffness on normal and degenerate human femoral heads. J Biomech 1971;4:597–609.
127. Johnston RC, Smidt GL. Hip motion measurements for selected activities of daily living. Clin Orthop Relat Res 1970;72:205–215.
128. Bergmann G, Graichen F, Rohlmann A. Is staircase walking a risk for the fixation of hip implants? J Biomech 1995;28:535–553.
129. Andriacchi TP, Mikosz RP, Hampton SJ, Galante JO. Model studies of the stiffness characteristics of the human knee joint. J Biomech 1983;16:23–29.
130. King AI. A review of biomechanical models. J Biomech Eng 1984;106:97–104.
131. Cavanagh PR et al. The relationship of static foot structure to dynamic foot function. J Biomech 1997;30:243–250.

132. Matsen FA, Fu FH, Hawkins RJ. The Shoulder: A Balance of Mobility and Stability. Rosemont: AAOS; 1992.
133. Chao EY, Morrey BF. Three-dimensional rotation of the elbow. J Biomech 1978;11:57–73.
134. Panjabi MM, Krag MH, Goel VK. A technique for measurement and description of three-dimensional six degree-of-freedom motion of a body joint with an application to the human spine. J Biomech 1981;14:447–460.
135. Yang KH, Latouf BK, King AI. Computer simulation of occupant neck response to airbag deployment in frontal impacts. J Biomech Eng 1992;114:327–331.
136. Walker PS. Human Joints and Their Artificial Replacements. Springfield (IL): Thomas; 1977.
137. Youm Y, Yoon YS. Analytical development in investigation of wrist kinematics. J Biomech 1979;12:613–621.

See also CARTILAGE AND MENISCUS, PROPERTIES OF; HIP JOINTS, ARTIFICIAL; HUMAN SPINE, BIOMECHANICS OF; LIGAMENT AND TENDON, PROPERTIES OF.

JOINT REPLACEMENT. See MATERIALS AND DESIGN FOR ORTHOPEDIC DEVICES.

LAPARASCOPIC SURGERY. See MINIMALLY INVASIVE SURGERY.

LARYNGEAL PROSTHETIC DEVICES

GUIDO BELFORTE
MASSIMILIANA CARELLO
Politecnico di Torino
Torino, Italy

GUIDO BONGIOANNINI
MAURO MAGNANO
ENT Division Mauriziano Hospital
Torino, Italy

INTRODUCTION

The larynx is a uniquely complicated organ strategically located at the division of the upper aerodigestive tract into the gastrointestinal tract and the airways. Alteration of its function can have a significant impact on vocal, digestive, and respiratory physiology (1–6). The hyoid bone, the thyroid cartilage, and the cricoid cartilage form the outside framework of the larynx. The mobile interior framework consists of the leaf-shaped epiglottis and the arytenoid cartilages. Each vocal cord stretches from an anterior projection of the arytenoid to the anterior midline of the inside thyroid cartilage. The arytenoid cartilages move in both a rocking and sliding motion on the cricoid cartilage to abduct and adduct the true vocal cords. The intrinsic muscles of the larynx control the movement of the vocal cords.

A section of the larynx is shown in Fig. 1, which is a posterior view of the cartilages and membranes (skeleton of larynx).

Figure 1. Section of the larynx: posterior view of cartilages and membranes.

The larynx has three important functions: protection of the lower airways during swallowing; respiration; and phonation.

Phonation is a complicated process in which sound is produced for speech. During phonation, the vocal folds are brought together near the center of the larynx by muscles attached to the arytenoids. As air is forced through the vocal folds, they vibrate and produce sound. The tone and level of the sound can be changed by contracting or relaxing the muscles of the arytenoids. As the sound produced by the larynx travels through the throat and mouth, it is further modified to produce speech.

Cancer of the larynx represented $\sim$0.7% of the total cancer risk in 2001, and is the most common of all head and neck cancers. However, head and neck cancers account for only $\sim$9% of all cancers diagnosed annually.

Laryngeal cancer occurs about five times more frequently in males than females. It is rare prior to age 40, after which the incidence in males increases rapidly with age. Cigarette smoking is the most important cause of laryngeal cancer, with smokers having a roughly 10-fold higher risk than nonsmokers. Heavy alcohol consumption also is a well-established risk factor.

Though laryngeal cancer is infrequent compared to cancer of the breast, lung, and prostate, the literature regarding this disease is substantial. This apparently disproportionate body of writing reflects the perceived importance of this neoplasm, which is in turn related to its potential impact on people's communicative ability: the threat to a patient's vocal organ is associated with profound psychological and socioeconomic overtones.

Originally, larynx cancer treatment focused primarily on cure by relentless and aggressive surgery. That era has been followed by the emergence of conservative partial laryngectomy, the development of more sophisticated radiation methods, and organ-sparing strategies in which chemotherapeutic, radiotherapeutic, and surgical techniques are used in a variety of combinations.

Total laryngectomy is one of the standard operations for laryngeal carcinomas. The prognosis associated with laryngeal carcinoma has improved: As the curability of laryngeal carcinoma is now $>60\%$, many patients thus survive for a long time after surgery.

Laryngectomy changes the anatomy: The lower respiratory tract is separated from the vocal tract and from the upper digestive tract, and the laryngectomee breathes through a tracheostoma. The direct connection between the vocal tract and the upper digestive tract remains unchanged.

Figure 2 shows the anatomic structure before laryngectomy (a) and after laryngectomy (b). Before laryngectomy (Fig. 2a), air can travel from the lungs to the mouth (as represented by the arrows), and the voice can be produced and modulated. After the laryngectomy (Fig. 2b), air issuing from the tracheostoma cannot reach the mouth, and sounds cannot be produced.

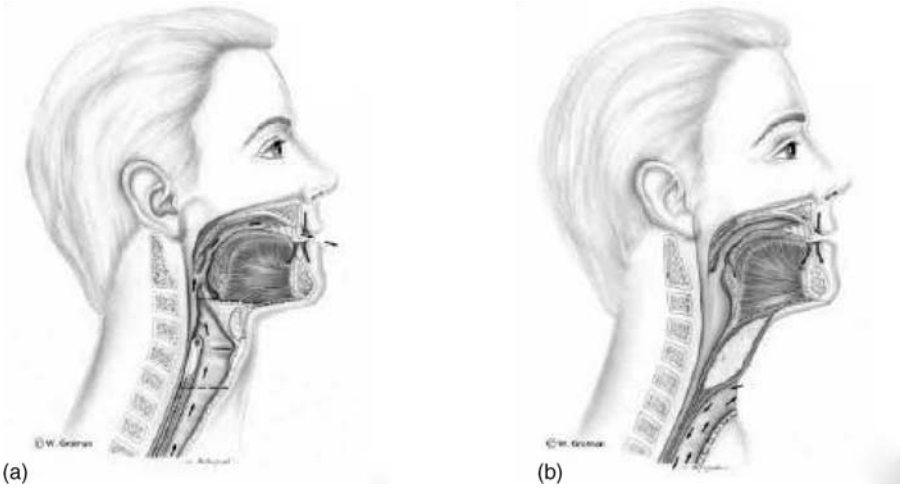


Figure 2. Anatomy before (a) and after (b) laryngectomy. (From Ref. 6.)

THE VOICE AFTER A LARYNGECTOMY

After laryngectomy, the patient is deprived of both the vibrating sound source (the vocal folds) and the energy source for voice production, as the air stream from the lungs is no longer connected to the vocal tract (1–9).

For some laryngectomy patients, the loss of speech is more important than survival itself. Consequently, a number of different methods for recovering phonation have been developed, including: esophageal speech; artificial larynx; and surgical laryngoplasty.

Numerous internal or external mechanical vibration sources have been developed that cause the air in the vocal tract to vibrate. These vibration sources can be powered by air pressure (expired air from the tracheostoma in some cases) or by an electric source with battery, and are thus classified as pneumo-larynges or electrical larynges.

ESOPHAGEAL SPEECH

Rehabilitation of the laryngectomized patient is usually a delayed process following recovery from surgery. After surgery, some patients try aphonic lip speech enhanced by buccal air trapping, while others choose written language as the method of communication.

Though most laryngectomized patients begin to learn esophageal speech, this method of speech rehabilitation requires a sequence of training sessions to develop the ability to insufflate the esophagus by inhaling or injecting air through coordinated muscle activity of the tongue, cheeks, palate, and pharynx.

Patients are encouraged to attempt esophageal sound soon after they are able to swallow food comfortably and learn to produce esophageal sound by trapping air in the mouth and forcing it into the esophagus. This produces a "burp-like" tone that can be developed into the esophageal voice.

There are various techniques for transporting air into the esophagus.

With the injection technique, the tongue forces air back into the pharynx and esophagus. This takes place in two stages, with the tongue forcing the air from the mouth back into the pharynx in the first stage, and the back of the tongue propelling the air into the esophagus in the second stage. For air to be transported into the esophagus, it is extremely important that these two stages be correctly synchronized.

With the inhalation method of esophageal speech, the patient creates a pressure in the esophagus that is lower than atmospheric pressure. As a result of this pressure difference, air will flow through the mouth past the upper segment of the esophagus into the lower esophagus. The patient will need to inhale air to be able to create a low endothoracic and esophageal pressure.

The last technique of capturing air is by swallowing air into the stomach.

Voluntary air release or "regurgitation" of small volumes vibrates the cervical esophageal inlet, hypopharyngeal mucosa, and other portions of the upper aerodigestive tract to produce a "burp-like" sound. Articulation by the lips, teeth, palate, and tongue produces intelligible speech.

Esophageal speech training is time consuming, frustrating, and sometimes ineffective.

Its main disadvantage is the low success rate in acquiring useful voice production, which varies from 26 to 55%. In addition, esophageal speech results in low-pitched (60–80 Hz) and low intensity speech, whose intelligibility is often poor. Age is the most important factor in determining success or failure: older patients are less successful in learning esophageal speech.

The airway used to create the esophageal voice is shown in Fig. 3. Direction of flow is indicated by the arrows, making it possible to distinguish between the pulmonary air used for breathing and the mouth air used for speech.

THE ELECTRONIC ARTIFICIAL LARYNX

Wright introduced the first electrolarynx in 1942. The most widely used electronic artificial larynx is the handheld transcervical device, or electrolarynx. This electrical device contains a vibrating diaphragm, which is held against the throat and activated by a button to inject vibratory energy through the skin and into the hypopharynx. By mouthing

Figure 3. Oesophageal speech. (From Ref. 6.)

words, the laryngectomee converts the vibrations into a new, low frequency voice for speech.

An example of electrolarynx application is shown in Fig. 4, where it is possible to distinguish the airway from the sound way and the positioning of the device in contact with the neck. Examples of commercial electrolarynges are shown in Fig. 5.

The same operating principle has been used in recent years to develop a transoral device (Fig. 6), which is placed in the mouth, where it is housed in an upper denture or an orthodontic retainer. The system consists of a control circuit, a loud speaker, and rechargeable batteries positioned inside the denture, as well as a charging port so that the batteries can be recharged outside the mouth.

FROM THE PNEUMOLARYNX TO THE VOICE PROSTHESIS

The artificial larynx has undergone many transformations over the years, and continues to do so today. The first types

Figure 4. Electrolarynx speech. (From Ref. 6.)

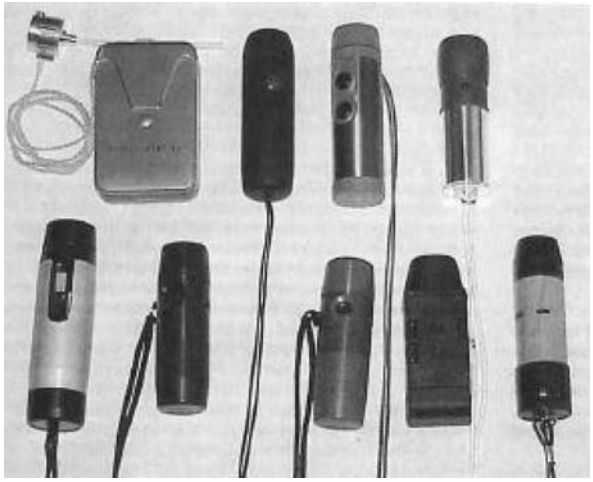


Figure 5. Examples of commercial electrolarynx. (From Ref. 5.)

of pneumolarynges, which included neck or mouth types and internal or external types were developed in 1860, when the first attempts at voice rehabilitation through surgery or prosthetization were made. A device was used to direct air from the lungs via a small tube to the mouth (1,3,5,7,9).

Experimental research started in 1860 with Ozermack and Burns, though it was not until 1874 that Billroth used an internal prosthesis designed by Gussenbauer (Fig. 7). This device was fitted in the trachea via the stoma with a cannula featuring a branch that entered the oral cavity.

Other similar prostheses were developed and used (Gosten, Gluck, etc.). Results, however, were not particularly satisfactory, as these devices were plagued by problems, such as tissue necrosis and leakage of food and liquids into the windpipe and lungs, thus causing infections (i.e., pneumonia).

Figure 8 shows the external prosthesis developed by Caselli in 1876. Figure 9 shows the application of the external prosthesis developed by Briani in 1950.

Figure 6. Transoral device. (From Ref. 5.)

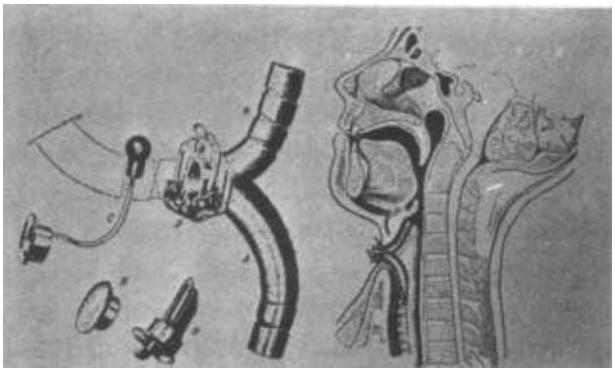


Figure 7. The Gussembauer prosthesis. (From Ref. 3.)

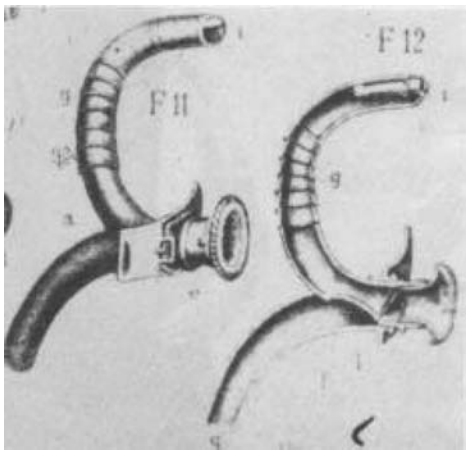


Figure 8. The Caselli prosthesis. (From Ref. 3.)

Figure 9. The Briani prosthesis. (From Ref. 3.)

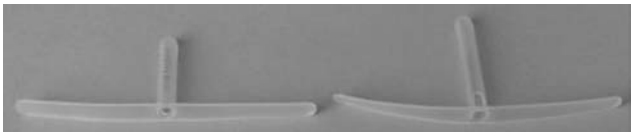


Figure 10. Example of removable prosthesis (Bivona).

Since 1960, surgical technique has been improved and a number of different types of prosthesis have been invented that take the importance of reproducing a voice that is similar to the original voice into account.

Tracheoesophageal puncture (TEP) or fistula with a voice prosthesis is the most popular of the new techniques, and has become the standard method of voice restoration after laryngectomy. However, it is usable only for selected patients.

Singer and Blom introduced the first TEP in 1980 (7,8). The aim was to create a permanent fistula (puncture or shunt) into the posterior tracheoesophageal wall between the trachea and the esophagus tract, so the pulmonary air can shunt from the airway into and up the esophagus. In this way, the vibration of the anatomic structures produces a noise.

A removable or fixed one-way type prosthesis can be positioned in the fistula.

The removable, or nonindwelling, prosthesis (Bivona, Blom-Singer duckbill) can be removed daily for cleaning by the patient, and is taped to the peritracheostomal skin. Figure 10 shows a photo of two Bivona valves; lengths differ in order to adapt the prosthesis to the patient.

The fixed, or indwelling, prosthesis cannot be removed daily, but is surgically placed in the fistula under local anesthesia. The operating principle of this type of prosthesis (known as a phonatory valve or voice button) can be explained with reference to Fig. 11.

Pulmonary air can be pushed through the valve into the esophagus for speech during expiration in two ways: by the laryngectomee, who covers the tracheal stoma with a finger (bottom left, Fig. 11) or automatically by a tracheostoma breathing valve (bottom right, Fig. 11).

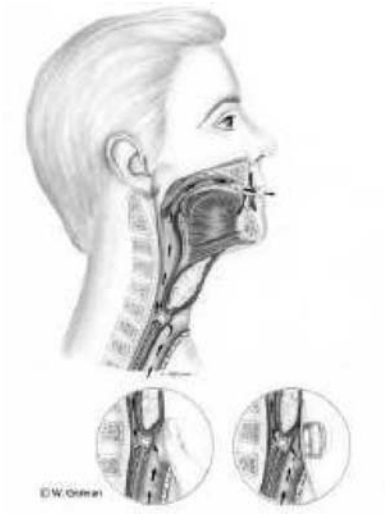


Figure 11. Speech with phonatory prosthesis. (From Ref. 6.)

The tracheostoma breathing valve contains an elastic diaphragm installed in a peristomal housing, which permits normal respiration during silent periods. Expiratory air for speech shuts off the pressure-sensitive diaphragm, and is thus diverted through the valve into the esophagus. This device eliminates the need for manual stoma occlusion during speech.

In both cases, air from the lung reaches the esophagus and causes the mucosal tissue tovibrate. The resulting sound can be modulated by the mouth, teeth, oral cavity, and so on, to produce the new voice.

Most speakers that have prosthesis do not have difficulties with articulation, rate of speech, or phonatory duration.

If esophageal speech depends on gulping or trapping air using the phonatory valve, the resulting speech depends on expiratory capacity. Voice quality is very good, and may resemble the "original" voice.

Poor digital occlusion of the tracheostoma as well as poor tracheostoma valve adherence allows pulmonary air to escape from the stoma prior to its diversion through the prosthesis into the oesophagus for voice production. In fact, the bleeding off of pulmonary air limits phonatory duration and, therefore, the number of words that can be spoken.

The one-way valve design of the prosthesis prevents aspiration (of food and liquid) from the esophagus to the trachea. An example of a phonatory valve is shown in Fig. 12 (11,12). The prosthesis illustrated in this sketch consists of an air-flow tracheal entry, whereby an endotracheal retention collar is connected to the mucosa (or tracheal flange), a hollow cylindrical tube (whose length depends on the patient's physical characteristics) connecting the trachea to the esophagus, an endoesophageal flange, and a dome (or hat) that closes the proximal endoesophageal end of the tube. Via the razor-thin slit (or esophagus exit), the hat enables airflow to pass from the trachea to the esophagus when there is a positive differential pressure, and prevents the reverse flow of liquid (or food) when the differential pressure becomes negative. The arrows represent the airflow paths.

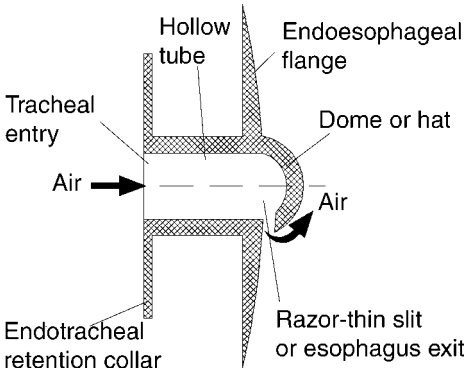


Figure 12. Sketch of phonatory valve or prostheses.

Figure 13. Examples of fixed commercial prosthesis. (Staffieri, Groningen standard, Groningen low pressure, Panje, Provox).

Hat shape and the extension of the razor-thin slit can differ according to valve type. The razor-thin slit may be located at the base of the hat (Staffieri, Groningen low pressure), at the center of the hat (Panje, Groningen standard), or inside the hollow tube (Provox, Blom-Singer).

Though valve geometry and shape may vary, the operating principle remains the same.

Several commercial prostheses are shown in Fig. 13: from left to right, they include the Staffieri, Groningen standard, Groningen low pressure, Panje, and Provox types.

Fixed and removable prostheses are available in different lengths, which usually range from 6 to 12 mm to enable the surgeon to select the dimensions that are best suited to the patient's physical characteristics, for example, posterior tracheoesophageal wall thickness.

To compare valve performance in the laboratory, most authors use valve airflow resistance (7,11), which is defined as the ratio of pressure to flow-rate. Figure 14 show an example of resistance versus flow-rate characteristics obtained with experimental tests on commercial valves (11).

Low resistance allows air to pass freely though the prosthesis with little effort on the part of the patient, and

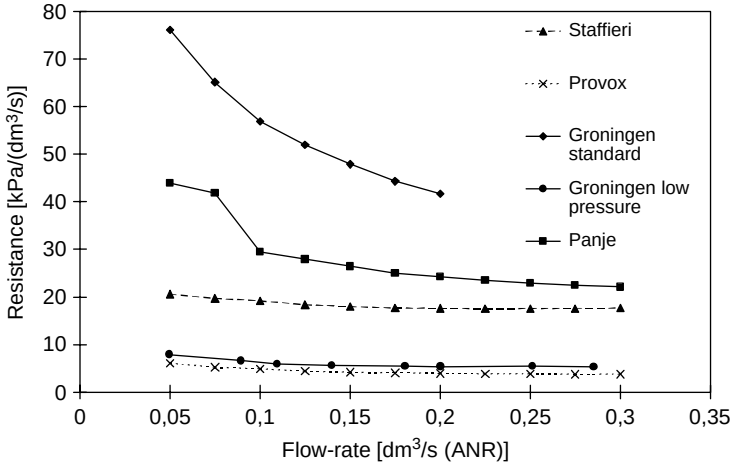


Figure 14. Resistance of commercial valves.

is thus the most favorable condition. Different materials have been used for phonatory prostheses. For indwelling prostheses, silicone rubber is the material of choice, though other materials such as polyurethane have also been used. The most significant problem affecting voice prostheses is the formation of a thick biofilm on the esophageal surface, as the esophageal environmental around the prosthesis provides ideal growing conditions for bacteria, fungi, and yeast.

In this area, secretions from the trachea, the mucous membranes in the oropharynx and esophagus, and saliva from the oral cavity create an optimal "pabulum" for microorganisms, which can thus adhere to the prosthesis surface. Though biofilm does not lead immediately to valve, malfunction, colonies growing around the prosthesis can increase airflow resistance, blocking the valve or preventing it from closing correctly. This causes saliva to leak from the oral cavity through the esophagus lumen. The biofilm also grows into the valve material (silicone rubber).

Massive microorganism growth is more frequent in indwelling prostheses than in non-indwelling types, as the latter are regularly removed and cleaned.

The first step of this colonization is the formation of a very thin layer of organic origin, called a conditioning film. Microorganisms adhere to the thin layer of saliva conditioning film on the inner esophageal surface of the device and form the first stratum of biofilm, which is nearly impossible to eradicate. There are few methods for reducing the growth of bacteria and fungi: One possibility is to apply (through oral instillations) antimycotic drugs that reduce biofilm formation and increase prosthesis life (1,2,7,8,13,14). The valve has a limited life, varying from ~4 to 8 months. After this period, the valve must be removed and changed in an outpatient procedure because of a dysfunctional mechanism caused by the biofilm, which causes food and liquids to leak from the esophageal area to the trachea.

Figure 15 shows a new Provox valve (left) and a Provox valve after eight months of use by a patient (right). Silicon rubber deterioration caused by microbial action and adherent deposits of microorganisms is clearly apparent.

Current work with voice prostheses focuses on improving their aerodynamic characteristics (low airflow resistance) and developing more resistant materials that can extend the life of the device. For the patient, in any case, the most important thing after a total laryngectomy is to be hopeful: a prosthesis can provide a new voice, improving the quality of life.

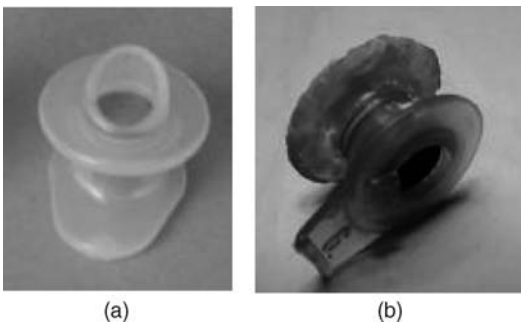


Figure 15. Provox valve: (a) new and (b) old.

BIBLIOGRAPHY

1. Mahieu HF. Voice and speech rehabilitation following laryngectomy. Ph.D. dissertation, Rijksuniversiteit Groningen; 1988.
2. Pighi GP, Cristoferi V. Protesi tacheo-esofagee. Bologna: Arianna Editrice; 1998.
3. Societé Francaise d'Oto-Rhino-Laryngologie et de Pathologie cervico-faciale. Réhabilitation de la voix et de la déglutition après chirurgie du larynx. Arnette, Paris; 1992.
4. Testut L, Jacob O. Trattato di anatomia topografica. Vol. II Collo - Torace –Addome, Utet, Torino, 1998.
5. Harrison LB, Sessions RB, Hong WK. Head and neck cancer. Lippinocott Williams & Wilkins; 2004; 178–189.
6. http://www.orl.nl
7. Blom ED. Tracheoesophageal voice restoration: origin, evolution, state of the art. Folia Phonatrica Logopaedica 2000; 52:14–23.
8. Blom ED, Singer MI, Hamaker RC. Tracheoesophageal voice restoration following total laryngectomy. San Diego: Singular Publishing Group Inc.; 1998.
9. Brown DH, et al. Postlaryngectomy voice rehabilitation: state of the art at the millennium. World J Surg 2003;27(7):824–831.
10. Nakamura T, Shimizu Y. Thacheal, laryngeal and esophageal replacement device. The Biomedical Engineering Handbook CRC and IEEE Press; 2000, p 136/1–136/13.
11. Belforte G, Carello M, Miani C, Staffieri A. Staffieri tracheo-oesophageal prosthesis for voice rehabilitation after laryngectomy: an evaluation of characteristics. Med Biol Eng Comput 1998;36:754–760.
12. Staffieri M, Staffieri A. A new voice button for post-total laryngectomy speech rehabilitation. Laryngoscope 1988; 98(9):1027–1029.
13. Schwandt LQ, et al. Prevention of biofilm formation by dairy products and N-acetylcysteine on voice prostheses in an artificial throat. Acta Otolaryngol 2004;124:726–731.
14. Leunisse C, et al. Biofilm formation and design features of indwelling silicone rubber tracheoesophageal voice prostheses—An electron microscopical study. J Biomed Mater Res 2001;58(5):556–563.

See also COMMUNICATION DEVICES; PULMONARY PHYSIOLOGY.

LASER SURGERY. See ELECTROSURGICAL UNIT (ESU).

LASERS, IN MEDICINE. See FIBER OPTICS IN MEDICINE.

LENSES, CONTACT. See CONTACT LENSES.

LENSES, INTRAOCULAR

HABIB HAMAM
Université de Moncton
Moncton, New Brunswick,
Canada

INTRODUCTION

Ridley's implantation (1949) of the first intraocular lens (IOL) marked the beginning of a major change in the practice of ophthalmology. The IOLs are microlenses placed inside the human eye to correct cataracts, nearsightedness, farsightedness, astigmatism, or presbyopia. There are two types of IOLs: anterior chamber lenses,

which are placed in the anterior chamber of the eye between the iris and the cornea, and posterior chamber IOLs, which are placed in the posterior chamber behind the iris and rest against the capsular bag. Procedures for implanting the IOLs and technologies for manufacturing them in various sizes, thicknesses, and forms as well as with various materials progressed tremendously in the last decade. Multifocal IOLs are one of the important signs of this progress. While monofocal IOLs, the most commonly used, are designed to provide clear vision at one focal distance, the design of multiple optic (multifocal) IOLs aims to allow good vision at a range of distances.

INTRAOCULAR LENSES: WHAT AND WHY?

An intraocular lens, commonly called IOL, is a tiny artificial lens implanted in the eye. It usually replaces the faulty (cataractous) cristalline lens. The most common defect of the natural lens is the cataract, when this optical element becomes clouded over. Prior to the development of IOLs, cataract patients were forced to wear thick coke bottle glasses or contact lenses after the surgery. They were essentially blind without their glasses. In addition to IOLs replacing the crystalline lenses, a new family of IOLs, generally referred to as phakic lenses, is nowadays subject of active research and development. (Phakos is the Greek word for lens. Phakic is the medical term for individuals who have a natural crystalline lens. In Phakic IOL surgery, an intraocular lens is inserted into the eye without removal of the natural crystalline lens.) These IOLs are placed in the eye without removing the natural lens, as is completed in cataract surgery. They are used to correct high levels of nearsightedness (myopia) or farsightedness (hyperopia).

An IOL usually consists of a plastic lens with plastic side struts called haptics to hold the lens in place within the capsular bag. The insertion of the IOL can be done under local anesthesia with the patient awake throughout the operation, which usually takes <30 min in the hands of an experienced ophthalmologic surgeon (Fig. 1).

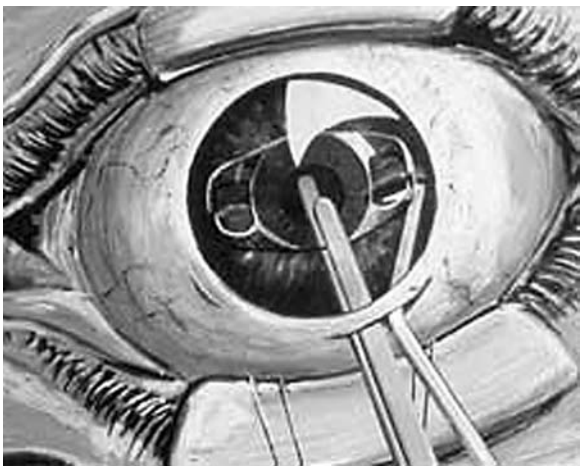


Figure 1. Implantation of an IOL: Since the natural lens is left undistrubed, the operation is much simpler than a cataract operation. The entire procedure consists of making a small incision at the edge of the cornea and placing the appropriate tiny plastic lens in the space between the iris and the cornea, (the anterior chamber). Stitches are used to close the incision.

HISTORICAL OVERVIEW

The idea of the IOL dates back to the beginning of modern cataract surgery when Barraquer developed keratomileusis (1). However, the first implantation of an artificial lens in the eye was probably attempted in 1795 (2). References to the idea of the IOL before World War II in ophthalmic literature are rare. There has been mention of limited animal experiments using both quartz and plastic material performed in the 1940s, but nothing had come of these efforts (3).

The most important step toward the implantation of IOLs came as a result of World War II pilots, and the injuries sustained when bullets would strike the plastic canopy of their aircraft (Fig. 2), causing small shards of plastic to go into their eye. In the late 1940s, Howard Ridley was an RAF ophthalmologist looking after these unfortunate pilots and observed, to his amazement, little or no reaction in cases in which the material had come from Spitfire planes. He then concluded the poly(methyl methacrylate) (PMMA) material of the canopies (windshield) was compatible with eye tissue (4). This observation sparked the idea for inserting an artificial lens in the eye. Ridley, who was convinced this lens should be placed in the posterior chamber, designed a disk-shaped lens, much like the natural lens, with a small peripheral flange allowing him to hold the lens with forceps (4). The artificial lens, made entirely of PMMA, weighed slightly >100 mg in air and was $\sim$8.35 mm in diameter. In several cases, he attempted to implant the lens following intracapsular surgery using the vitreous base for support (5). On November 29, 1949, the first successful IOL implantation was performed at St. Thomas Hospital in London (6,7). While far from perfect, the procedure worked well enough to encourage

Figure 2. Invention of the IOL: During World War II, Fighter pilots were sometimes involved in accidents where the plastic windshield (canopy) of their aircraft was shattered. Doctors found that fragments of the canopy that entered the eye were tolerated by the eye tissues. They might remain inside the eye for years, and the eye would not react.

further refinement. Then, over a decade, Ridley implanted several hundred IOLs (8).

Though Ridley was ahead of his time, his method was subject to serious criticism. Complications were common and failure rates $> 50\%$ were often contemporaneously quoted. Fixation was dependent on the formation of adhesions between the iris and the capsule. Several ophthalmologists strongly opposed to his procedure. Implantation in the anterior chamber was technically easier and was compatible with intracapsular surgery. Also, fixation could be achieved at the time of implantation by adding haptic struts to the lens that could be wedged into the angle. The first anterior chamber lens was implanted by Baron in 1952 (8).

To make intraocular lens implantation safe, developments in lens design and surgical techniques were required. Lens implantation did not become widely adopted as an integral part of cataract surgery until the 1980s. Key advances were the introduction of viscoelastic fluids to protect the cornea during implantation and flexible haptics to enhance long term stability of the IOL (9).

With traditional single vision monofocal IOLs, people generally experience good vision after surgery at a single focal point, either near or at a distance. The multifocal IOL (10) was designed in the mid-1990s to provide a full range of vision with independence from glasses in most situations.

Besides, the invention of phakic lenses is no less important than Ridley's invention. These IOLs were introduced by Strampelli (11) and later popularized by Barraquer in the late 1950s (12). Phakic IOLs are becoming more popular because of their good refractive and visual results and because they are easy to implant in most cases (13). In the beginning, the design was a biconcave angle-supported lens. These lenses were abandoned following serious angle- and endothelium-related complications. By the late 1980s, Baikoff (14,15) introduced a myopia lens with Kelman-type haptics (16). This design had many problems, leading to its design modification a number of times. Fyodorov, the inventor of radial keratotomy (17), introduced the concept of a soft phakic lens in the space between the iris and the anterior surface of the crystalline lens (18).

Based on earlier works of Worst, winner of the Binkhorst Award for innovation in ophthalmology in 1976, Fechner introduced phakic myopia lens of iris claw design in 1986 (19). This IOL is then referred to as Worst–Fechner lens (20). Many companies around the world manufactured it in various models. Today, people usually identify it by the name of Artisan IOL.

MATERIAL OF THE IOL

Many factors, such as the optical quality of the system of the eye (aberrations, . . .), presence of inflammation, cost, and wound size, depend on the material and the form of the IOL.

From the point of view of flexibility, there are two families of IOLs: foldable and inflexible lenses. Foldable IOLs are generally made of acrylic or silicone. They can be rolled up and inserted through a tube with a very small incision not requiring any stitches. Inflexible IOLs, typically made of PMMA, require a larger incision because they are unfoldable. Most lenses are biconvex, thus optically equivalent upside down. However, most lenses have haptics which are generally angled to push the posterior optics.

Four basic materials are used for IOLs: PMMA, silicone, acrylic and collamer. Other materials are also used. For example, some manufacturers replace silicon by a hydrophilic biocompatible polymer, called collamer. Many IOLs have been made from PMMA plastic, the same plastic the original hard contact lenses were made of. Silicon IOLs are foldable. Folding an IOL allows it to be inserted through a smaller incision. A smaller incision heals faster and induces less postop astigmatism. Some foldable lenses are now being made of acrylic plastic. While acrylic and silicone lenses are very popular, PMMA is the time-tested material but, as stated above, requires a large incision. Silicone oil is also a problem for silicone IOLs in that the view of the fundus can be severely degraded. This is less so for PMMA and hydrophobic acrylic IOLs and least for hydrophilic acrylic. Although this is a relative contraindication for silicone IOLs in the face of significant vitreoretinopathy, a solvent exists that eliminates the problem (21). Collamer is a new hydrophilic material just recently released that has shown some interesting properties. It has been shown to exhibit less internal reflectance than other lens materials including silicone, acrylic, and PMMA (22). It reduces the risk of long-term inflammation (23). Table 1 summarizes the characteristics of the four materials.

VARIOUS PARAMETERS FOR IOLs

Are age, race, and sex important parameters for IOL implantation? Age at which surgery is performed turned out to be of great importance (24–28). The ideal age should be at around 18 years when the refraction stabilizes. However, in specific circumstances, in the interest of the minor patient, the parents and the surgeon can opt to perform phakic lens implantation at an earlier age. Studies

Table 1. Four Commonly Used IOLs Materials and their Advantages and Drawbacks

Material	Flexibility	Advantages	Drawbacks
PMMA	Rigid	Low cost, less inflammation, long-term experience, good bicompatibility	larger incision, not foldable
Silicone	Foldable	Smaller incision, injectable	high cost, more inflammation, cannot use with silicon oil
Collamer	Foldable	Smaller incision, less inflammation, very good bicompatibility	high cost, short term experience
Acrylic	Foldable	Smaller incision, less inflammation, high refraction index (thin IOL), good bicompatibility	high cost

of the suitable age of IOL implantation in children have been carried out (24). A 3-year-old child has been qualified with IOL implantation, the child younger than 9 years old should be implanted with a normal adult IOL and then corrected with glasses, and a child after 10 years old should be directly implanted with a proper dioptric IOL (24). Some researchers evaluated the influence of cataract surgery on the eyes of children between 1 and 5 years old. They concluded that cataract surgery, either extraction with or without IOL implantation, did not retard axial elongation in children above 1 year old (25). Comparisons between children with congenital or developmental lens opacities who underwent extracapsular cataract extraction and children with normal eyes have been carried out (26). The pattern of axial elongation and corneal flattening was similar in the congenital and developmental groups to that observed in normal eyes. No significant retardation or acceleration of axial growth was found in the eyes implanted with IOLs compared with normal eyes. A myopic shift was seen particularly in eyes operated on at 4–8 weeks of age and it is recommended that these eyes are made 6 D hyperopic initially with the residual refractive error being corrected with spectacles (26).

To our knowledge, IOL implantation does not depend on race and sex.

OPTICAL QUALITY

Two optical qualities are distinguinshed: the intrinsic optical quality of the IOL and the optical quality of the system of the eye including the IOL. Many factors, such as the material and the geometrical profile of the IOL, influence the intrinsic quality of this optical element. Axial shift (decentration), transversal rotation (not around the optical axis), and deformation (mechanical stresses, . . .) of the IOL are examples of factors affecting the optical quality of the whole system of the eye even in case the IOL alone is a perfect optical element. Thus, the optical quality of the IOL is certainly important. However, for vision, the determinant factor is the optical quality of the whole system of the eye in which the IOL is implanted. Several studies have been undertaken to assess the optical quality of the IOL and the optical quality of the whole system of the eye. Before progressing in this section, let us briefly introduce the notion of optical quality.

Aberrations

Stated in wave optics, the system of the eye should transform the input wavefront into a perfect convergent spherical wavefront that has the image point as center (29–31). Note that an optical wavefront represents a continuous surface composed of points of equal phase. Thus all image-forming rays, which travel normal to the exit spherical wavefront, meet in the focal point in phase, resulting in maximum radiant energy being delivered to that point. In reality, this situation never occurs. The rays modified by the optical system do not converge entirely to a common point image. For one object point correspond several image points that form a blurred image. This deviation from the ideal case is called optical aberration, or merely aberration, and is a measure of the optical quality of the system. Aberration can be quantified either with respect to the expected image point or to the wavefront corresponding to this ideal point. If the real output wavefront is compared to the ideal one, it is called the difference between them wavefront aberration (29). All human eyes suffer from optical aberrations that limit the quality of the retinal image (32–36). Several metrics have been proposed to measure the optical quality of the system of the eye (37–41). Let us return back to IOLs now. Optical quality of multifocal IOLs will be treated in the section devoted to this kind of lens.

Optical Quality of the IOL

The optical quality of IOL was the subject of intensive studies. Several common but some contrasted results have been obtained. An exhaustive study goes beyond the scope of this document. We limit our attention to some recent results. Tognetto et al. (42) evaluated the optical quality of different IOLs by using an optical test bench. The purpose of the study was to evaluate the optical quality of IOLs and not to evaluate the optical performance of these lenses once implanted. Three randomly acquired samples of 24 different models of foldable IOLs were compared. The conclusion is that different IOLs can transmit different spectra of spatial frequencies. The best frequency response was provided by acrylic IOLs, particularly those with an asymmetrically biconvex profile. This could be due to a reduction of optical degradation provided by this type of profile. A lens with a higher frequency response should create a better quality of vision once implanted, and the frequency response should therefore be considered when choosing the intraocular lens model (42).

Negishi et al. (43) evaluated the effect of chromatic aberrations in pseudophakic eyes with various types of IOLs. Their results show that longitudinal chromatic aberrations of some IOLs may degrade the quality of the retinal image. They concluded that attention must be paid to the detailed optical performance of IOL materials to achieve good visual function.

In a comparative study (44), Martin found that the collamer IOL reduced the number of induced higher order aberrations when compared with acrylic and silicone lenses. Indeed, he found that the collamer IOL has 55–117% fewer induced higher order aberrations than acrylic or silicone materials. As a consequence, it produces less postop glare. He concluded the collamer lens provides clearer vision than the other lenses.

Optical Quality of ARTISAN Lenses

Brunette et al. (45) evaluated the optical quality of the eye before and after the insertion of an ARTISAN phakic intraocular lens for the treatment of high myopia (range −20.50 to −9.75D). Consecutive patients implanted with the ARTISAN lens by a single surgeon were prospectively enrolled. One eye per subject was tested. The wavefront aberration was calculated from images recorded with a Hartmann-Shack sensor (46,47). The PSF and the MTF were also computed from the wavefront aberration. It was concluded that preliminary data using the Hartmann–Shack wavefront sensor have not revealed a tendency

toward deterioration of the optical performance following the insertion of an ARTISAN lens for the treatment of high myopia. The Hartmann–Shack sensor is a useful tool for the objective assessment of the image optical quality of eyes with a phakic intraocular lens.

MULTIFOCAL IOLs

Unlike the natural lens, the curvature of current intraocular lenses cannot be changed by the eye. Standard intraocular lenses provide good distance vision, and the patient needs reading glasses for near vision. Newer bifocal intraocular lenses give distance vision in one area and near vision in another area of the vision field. How does it work?

The basic idea consists in providing a lens with two posterior focal points instead of one. The IOL is no longer monofocal. It becomes bifocal. Two solutions are possible: refractive or diffractive bifocal lens. A third solution consists in combining both approachs together.

Diffractive Lenses

The idea comes from the principle of the Fresnel zone plate. It consists in designing a binary diffractive phase element so when the incident wave comes across this diffractive element, all the resulting waves, coming from all points of the zone plate, arrive in phase at a certain (focal) point. They then superimpose constructively, yielding a focusing behavior. As shown in Fig. 3, waves traveling along various segments arrive in phase at the focal point since the optical paths differ by a multiple of the wavelength. To fulfill this condition, the thickness d is chosen so that it introduces a phase shift of π: $d = (2k+1)\,\lambda/[2(n-1)]$ (optical path: $(2k+1)\,\lambda/2$). In Fig. 3, $k = 0$, yielding $d = \lambda/[2(n-1)]$, where n is the refraction index of the diffractive lens. The radii of the rings (Fig. 3) verify the following rule:

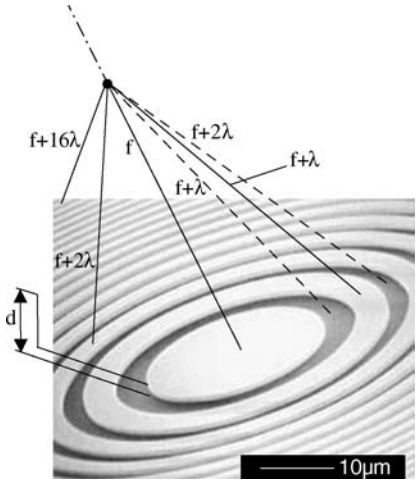


Figure 3. Binary diffractive phase element: Fresnel zone plate. Waves traveling along various segments arrive in phase at the focal point since the optical paths differ by a multiple of the wavelength. To fulfill this condition, the thickness d is chosen so that it introduces a phase shift of π: $d = (2k + 1)\,\lambda/[2(n-1)]$ (optical path: $(2k + 1)\,\lambda/2$): In the figure, $k = 0$.

$r_m = \sqrt{m} \cdot r_1 = \sqrt{m\lambda f}$, where r_1 is the radius of the smallest circle and f is the focal length. Without grooves on the diffractive, the waves traveling along the segments, represented by dashed lines in Fig. 3, would arrive in phase opposition with respect to other waves. In reality, the binary form (two phase levels) of the diffractive lens does not fully ensure the condition of coming in phase at the focal point. For rigor, several phase levels are required (Fig. 4). In general, the diffraction efficiency η, which is defined as the ratio of the focused energy to the incident energy, increases with the number of phase levels L according to the following formula (48): $\eta = \sin^2(\pi/L)/(\pi/L)^2$. Using two phase levels (binary), only 41% of the energy is focused. The rest is scattered in space. A four level diffractive lens focuses 81% of the input energy (it scatters only 19% of the incident energy).

To obtain a bifocal diffractive lens, we need to focus rays on two focal points at distances f_1 and f_2. It can be done by providing two series of zones (rings). The first series, the inner one, involves radii verifying $r_m^{(1)} = \sqrt{m\lambda f_1}$ (with $m = 1,2,\ldots, M$), whereas the radii in the second series, the outer one, satisfy the condition $r_p^{(2)} = \sqrt{p\lambda f_2}$ (with $p = M+1,\ldots, P$). To obtain a multifocal diffractive lens, we need more series of radii.

Refractive Lenses

An alternative to obtain a multifocal length consists in modifying the surface profil of a conventional biconvex lens so that it provides two or more different focal points for light convergence. The technique consists in designing a spherical refractive surface that has additional refracting surfaces to give a near add (Fig. 5), or a near and intermediate add. The principle of multifocal refractive lenses is illustrated in Fig. 6a. Refractive IOLs with several focal points are commercialized in various models. For example, one of the models includes five refractive zones targeting distance, intermediate and near vision (Fig. 6b). The IOL uses continuous aspheric optics to ensure that 100% of the light entering the eye reaches the retina. The lens uses five concentric zones with the first, third, and fifth zones being far dominant and second and fourth zones being near dominant (49). The light distribution is arranged so that 50% of light is distant focussed, 13% is focussed for intermediate vision and 37% for near vision. The near add

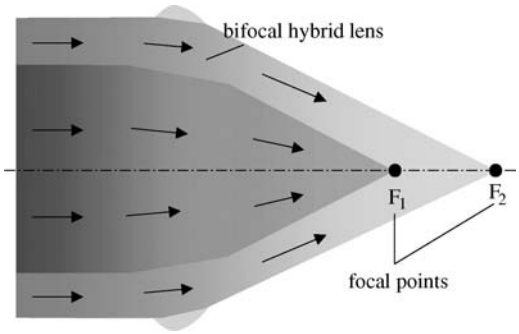


Figure 4. Bifocal hybrid refractive/diffractive IOL: Anterior surface broken up into a refractive zone and a second zone composed of concentric diffractive rings.

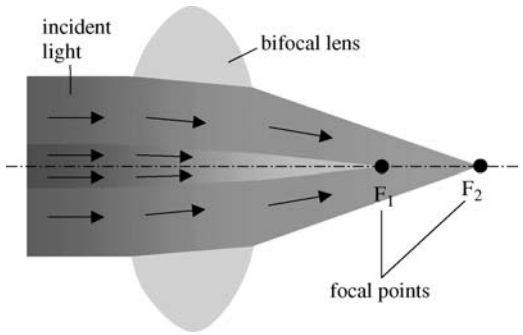


Figure 5. Bifocal refractive IOL: Spherical refractive surface that has additional refracting surfaces to give a near add.

comprises of 3.5 dioptre intraocular power equivalent to a 2.75–2.85 add in the spectacle plane (49).

Optical Quality of Refractive and Diffractive Lenses

This discussion will be limited to some recent results. Pieh et al. (50) compared the optical properties of bifocal diffractive and multifocal refractive intraocular lenses. The IOLs were manufactured by different companies. A model eye with a pupil 4.5 mm in diameter was used to determine the point spread function (PSF) (30,51) of the distance focus and near focus of the IOLs to compare them with PSFs of foci of corresponding monofocal lenses. For interpreting the PSFs the through focus response, the modulation transfer function (MTF) (51), and the Strehl ratio (51) were evaluated. They concluded the modulation transfer functions

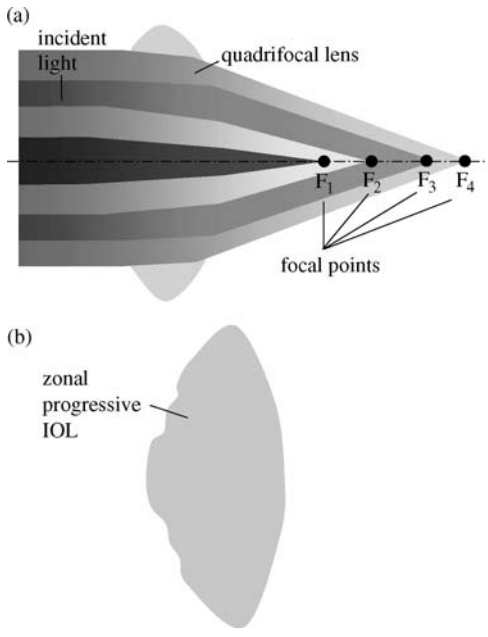


Figure 6. Multifocal refractive IOLs: (a) several refractive surfaces with different curvatures. Each one provides a focal point (b) a commercial zonal progressive IOL with five concentric zones on the anterior surface. Each zone repeats the entire refractive sequence corresponding to distance, intermediate and near vision, resulting in wide-range vision.

reveal comparable properties for distance vision and a superiority of the bifocal diffractive lens over the refractive multifocal lens for near vision. This mean due to the fact that the incoming light is distributed over different zones in the refractive lenses.

Hybrid Lenses

Diffractive and refractive optics (Fig. 4) can be combined. In this type of lens, the basic spherical refractive surface is broken up into a refractive zone and a second zone composed of concentric diffractive rings. This combination of zones creates two different focal points for light convergence, one for near objects and one for distant objects. Hybrid IOLs are basically bifocal lenses (Fig. 4). The usual strategy has been to have a distance component with a near component targeting the usual near distance. However, multifocal hybrid lenses are possible.

IOLs AND ACCOMODATION

New research into accommodating intraocular lenses indicates that many patients who get these implants will enjoy good distance and near vision (52). The first hint that intraocular lens implants could accommodate came back in 1986, when Thornton used A-scan biometry to report on anterior movement of a three-piece loop IOL (53). He found that this forward movement allowed some patients to have good uncorrected distance and near vision simultaneously.

The mechanisms of presbyopia remain incompletely understood. A review of the variety of such mechanisms has been presented by Atchison (54). Accommodation in the youthful, phakic human eye is accomplished by contraction of the ciliary body and subsequent release in the resting tension of the zonular fibers by which the crystalline lens is suspended, resulting in increased lens curvature (55–57). The weight of current evidence seems to suggest that although some loss of ciliary body action might contribute to reduced accommodation (58), significant ciliary body function persists into advanced maturity, and that loss of lens and capsule elasticity in concert with changes in the geometry of zonular attachments are probably most culpable in producing the distress of presbyopia (59). If so, replacement of the crystalline lens with a lens that responds to ciliary body contraction should restore accommodative function (60). Attempts to replace the crystalline lens by refilling the capsular bag with appropriately deformable gels have been made (59,61,62). McLeod et al. aimed at designing an accommodating intraocular lens with extended accommodative range that can be adapted to current standard phacoemulsification and endocapsular implantation technique (60). They concluded that a dual optic foldable IOL design can increase the optical effect of a given displacement and suggests improvements for accommodating intraocular lenses.

BIBLIOGRAPHY

1. Barraquer JI. Keratomileusis for the correction of myopia. Arch Soc Amer Oftalmol Optom 1964;5:27–48.

2. Ascher K. Prothetophakia two hundred years ago. Am J Ophthalmol 1965;59:445–446.
3. Nordlohne ME. The Intraocular Lens Development and Results with Special Reference to the Binkhorst Lens, 2nd ed. The Hague: Dr. W. Junk b.v.; 1975: p 14–17
4. Ridley H. Intraocular acrylic lenses—past, present and future. Trans Ophthal Soc UK 1964;84:5–14.
5. Kwitko ML, Kelman CD, editors. Intraocular lens implantation: The Beginnings: The History of Modern Cataract Surgery. The Hague: Kugler Publications; 1998: p 35–52.
6. Apple DJ, Sims J. Harold Ridley and the invention of the intraocular lens. Surv Ophthalmol 1996;40:279–292.
7. Rosen E. History in the making. J Cataract Refract Surg 1997;23:4–5.
8. Choyce P. Intraocular Lenses and Implants. London; H K Lewis & Co. Ltd; 1964.
9. Apple DJ et al. Complications of intraocular lenses: A historical and histopathological review. Surv Ophthalmol 1984;29: 1–54.
10. Javitt JC, et al. Outcomes of cataract extraction with multifocal intraocular lens implantation. Functional status and quality of life. Ophthalmology 1997;104:589–599.
11. Strampeli B. Sopportabilità di lenti acrilichi in camera anteriore nella afachia e nei vizi di refrazione. Ann Oftalmol Clin Ocul 1954;80:75–82.
12. Barraquer J. Anterior chamber plastic lenses. Results and conclusions from five years experience. Trans Ophthalmol Soc UK 1959;79:393–424.
13. Neuhann T, et al. Phakic intraocular lenses, J Refract Surg 1998;14:272–279.
14. Baikoff G, Joly P. Correction chirurgicale de la myopie forte par un implant de chambre anterior dans l'oeil phake. Bull Soc Belge Ophthalmol 1989;233:109–125.
15. Baikoff G. Phakic anterior chamber intraocular lenses. Int Ophthalmol Clin 1991;31:75–86.
16. Rosen ES, Haining WM, Arnott EJ editors. Intraocular Lens Implantation. London, New York: Mosby; 1984.
17. Fyodorov SN, Durnev VV. Operation of dosaged dissection of corneal circular ligament in cases of myopia of mild degree. Ann Ophthalmol 1979;11:1885–1890.
18. Fyodorov SN, Zuev VK, Aznavayez VM. Intraokuliarnaia korrektsia miopii vysokoi stepeni zadnekamernmi otritsatelimi linzami. Oftalmochirugia 1991;3:57–58.
19. Fechner PU, Alpar J. Iris Claw Lens or Lobster Claw Lens of Worst; 1986.
20. Fechner P, Worst J. A New concave intraocular lens for the correction of myopia. Eur J Implant Refract Surg 1989;1:41–43.
21. Hoerauf H, Menz DH, Dresp J, Laqua H. Use of 044 as a solvent for silicone oil adhesions on intraocular lenses. J Cataract Refract Surg 1999;25:1392–1396.
22. Ossipov A. Comparison of internal reflectance patterns of Collamer, acrylic and silicone. 1997. Data on file, STAAR Surgical.
23. Davis EA. Study of post-cataract surgery inflammation with 3 different IOLs (Collamer, SI40NB, AR40). Summary of data found in all patients. Presented at OSN Meeting: New York: October 2003.
24. Jia S, Wang X, Wang E. A study of suitable age for intraocular lens implantation in children according to ocular anatomy and development, Zhonghua Yan Ke Za Zhi. 1996 Sept; 32(5): 336–338.
25. Zou Y, Chen M, Lin Z, Yang W, Li S. Effect of cataract surgery on ocular axial length elongation in young children. Yan Ke Xue Bao 1998;14(1) : 17–20.
26. Flitcroft DI, Knight-Nanan D, Bowell R, Lanigan B, O'Keefe M. Intraocular lenses in children: changes in axial length, corneal curvature, and refraction. Br J Ophthalmol 1999;83(3):265–269.
27. Kora Y, et al. Eye growth after cataract extraction and intraocular lens implantation in children. Ophthalmic Surg 1993;24(7): 467–475.
28. Pan Y, Tang P. Refraction shift after intraocular lens implantation in children. Zhonghua Yan Ke Za Zhi 2001;37(5):328–331.
29. Welford W. Aberrations of Optical Systems. Bristol: Adam Hilger; 1962.
30. Born M, Wolf E. The Diffraction Principles of Optics: Electromagnetic Theory of Propagation, Interference, and Diffraction of Light, 6th ed. New York: Pergamon Press; 1989.
31. Hamam H. Aberrations and their impact on image quality. Wavefront Analysis, Aberrometers & Corneal Topography, Agarwal's edition, 2003.
32. Castejon-Mochon JF, Lopez-Gil N, Benito A, Artal P. Ocular wave-front aberration statistics in a normal young population. Vision Res 2002;42(13):1611–1617.
33. Howland HC, Howland B. A subjective method for the measurement of monochromatic aberrations of the eye. J Opt Soc Am 1977;67(11):1508–1518.
34. Porter J, Guirao A, Cox IG, Williams DR Monochromatic aberrations of the human eye in a large population. J Opt Soc Am A Opt Image Sci Vis 2001;18(8):1793–1803.
35. Paquin MP, Hamam H, Simonet P, Objective measurement of the optical aberrations for myopic eyes, Opt Vis Sci 2002;79: 285–291.
36. Thibos LN, Hong X, Bradley A, Cheng X. Statistical variation of aberration structure and image quality in a normal population of healthy eyes. J Opt Soc Am A 2002;19:2329–2348.
37. Françon M Vision dans un instrument entaché d'aberration sphérique. Thèse, éditions de la Revue d'Optique, 1945.
38. Smith WJ Modern optical engineering, the design of optical system. New York: Me Graw-Hill, 1990.
39. Maréchal A. Étude des effets combinés de la diffraction et des aberrations géométriques sur l'image d'un point lumineux. Revue d'Optique; 1947.
40. Hamam H, New metric for optical performance. Opt Vis Sci 2003;80:175–184.
41. Marsack JD, Thibos LN, Applegate RA Metrics of optical quality derived from wave aberrations predict visual performance. J Vis 2004;4(4):322–328.
42. Tognetto D, et al. Analysis of the optical quality of intraocular lenses. Inv. Ophthalmol & Vis Sci (IOVS) 2004;45/8:2682–2690.
43. Negishi K, Ohnuma K, Hirayama N, Noda T. Effect of chromatic aberration on contrast sensitivity in pseudophakic eyes. Arch Ophthalmol 2001;119:1154–1158.
44. Matin RG. Higher-Order Aberrations and Symptoms with Pseudophakia, Symposium on Cataract, IOL and Refractive Surgery, April 12–16, 2003 San Francisco, CA.
45. Brunette I, et al. Optical quality of the eye with the artisan phakic lens for the correction of high myopia. Optomol Vis Sci 2003 Feb; 80(2):167–174.
46. Liang J, Grimm B, Goelz S, Bille JF. Objective measurement of wave aberrations of the human eye with the use of a Hartmann-Shack wave-front sensor. JOSA A 1994;11:1949–1957.
47. Hamam H. An apparatus for the objective measurement of ocular image quality in clinical conditions. Opt Commun 2000;173:23–36.
48. Hamam H, de Bougrenet JL. Efficiency of programmable quantized diffractive phase elements. Pure and Appl Opt 1996;5:389–403.
49. Wilson K. New Technology Removes Cataract and Improves Vision. Geriatr and Aging 1998;1: p 15.
50. Pieh S, et al. Quantitative performance of bifocal and multifocal intraocular lenses in a model eye: Point spread function in multifocal intraocular lenses. Arch-Ophthalmol. 2002;120(1): 23–28.
51. Malacara D, Malacara Z. Handbook of Lens Design. New York; Marcel Dekker 1994.
52. Karpecki PM. The future of IOLs that accommodate. Rev Opt Dec 2002.

53. Thornton S. Lens implantation with restored accommodation. Curr Cana Ophthal Prac 1986;4:60–62.
54. Atchison DA. Accommodation and presbyopia. Ophthal Physi Opt 1995;15:255–272.
55. Fisher RF. Presbyopia and the changes with age in the human crystalline lens. J Physiol (London) 1973;228:765–779.
56. Koretz JF, Handelman GH. Modeling age-related accommodative loss in the human eye. Math Mod 1986;7:1003–1014.
57. Schachar RA. Zonular function: A new model with clinical implications. Ann Ophthalmol 1994;26:36–38.
58. Hara T, et al. Accommodative intraocular lens with spring action part 1. Design and placement in an excised animal eye. Ophthal Surg 1990;21:128–133.
59. Gilmartin B. The aetiology of presbyopia: A summary of the role of lenticular and extralenticular structures. Ophthal Physiol Opt 1995;15:431–437.
60. McLeod SD, Portney V, Ting A. A dual optic accommodating foldable intraocular lens. British J Ophthal 2003;87:1083–1085.
61. Cumming JS, Slade SG, Chayet A. AT-45 Study Group. Clinical evaluation of the model AT-45 silicone accommodating intraocular lens. Results of feasibility and the initial phase of a Food and Drug Administration clinical trial. Ophthalmology 2001;108:2005–2009.
62. Kuchle M, et al. Implantation of a new accommodating posterior chamber intraocular lens. J Refract Surg 2002;18:208–216.

See also BIOMATERIALS: POLYMERS; CONTACT LENSES; VISUAL PROSTHESES.

LIFE SUPPORT.

See CARDIOPULMONARY RESUSCITATION.

LIGAMENT AND TENDON, PROPERTIES OF

G AZANGWE
RM ASPDEN

INTRODUCTION

Tendons and ligaments are fibrous connective tissues that play a mechanical role in the stability and locomotion of the body by transmitting tension. Unlike muscle, which actively contracts, ligaments and tendons are passive. Tendons transmit mechanical forces from muscle to bone, whereas ligaments join bone to bone. Both tendons and ligaments contain relatively few cells (1), and their extracellular matrices are made up of several components. These components are combined in various proportions, and with different organizations to give mechanical properties appropriate to the function of the particular tendon or ligament. There have been a number of reviews in recent years covering specific ligaments, for example, in the rabbit (2), or tendons, such as the human achilles (3), or aspects of their behavior such as healing and repair (4). In this article, the emphasis is on properties the ligaments have in common, which will provide an insight into how and why they behave as they do. This will be based around their functioning as fiber-reinforced materials whose properties are regulated by the cells they contain that produce and maintain the extracellular matrix.

First, this article considers the components of the tissue, not from a biochemical point of view, but as components that may be combined to produce mechanically stable materials. The constituents are considered in terms of the matrix in which are embedded fibers of collagen and varying amounts of elastin. Following this is a discussion of the ways these components interact in ligaments and tendons to yield composite materials with the required mechanical properties. A consideration of some of the ways in which ligaments and tendons may be damaged, and the mechanisms by which they might recover or be repaired, leads to a final, brief review of their surgical replacement. A small section on work being conducted in order to produce a tissue engineered ligament and tendon is also included.

COMPONENTS

Tendons and ligaments are composed primarily of collagen fibers surrounded by a matrix. Here the matrix refers to all the materials that surround the collagen fibers providing both structural support and a medium for diffusion of nutrients and gases. Note this is in contrast to its use in biological terms in which it generally includes the fibrous components. The matrix contains proteoglycans and adhesive glycoproteins and is highly hydrated (typically 65–80% water) (5,6).

Collagen

Collagen fibrils are able to reinforce the weak matrix because of their much greater stiffness and strength in tension (7). The collagen molecule is a long, stiff rod made up of three polypeptide chains wound as a triple helix structure (8). Each fibril is like a rope in which linear molecules are packed together with their axes parallel, within a few degrees, to the fibril axis. Molecules are held together by covalent cross-links so that the fibril is very strong in tension. The regular arrangement of molecules along the axial direction in a fibril gives rise to the characteristic periodicity of 67 nm, which may be seen in electron micrographs (Fig. 1). Although to date, 21 genetically different types of collagen have been identified, types

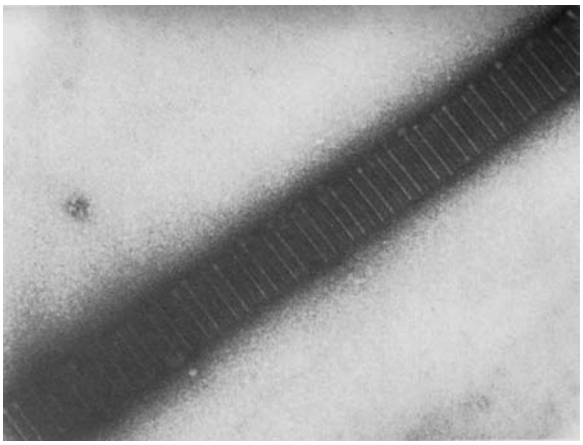


Figure 1. Electron micrograph of collagen fibril from rat tail tendon stained with uranyl formate showing characteristic banding pattern repeated every 67 nm. Micrograph courtesy of Dr D. P. Knight.

I and III fibrous collagens dominate in tendons and ligaments.

There is a heirarchical structure within tendons (9) that has been reviewed often [e.g., (10)], and that appears to have been based on an earlier description of keratin. In this model, collagen molecules are arranged successively into microfibrils, subfibrils, fibrils, and fascicles that are finally packed into the tendon sheath. Evidence for micro- and subfibrils is still equivocal, but a principal mechanical structure is the fibril. These generally have a unimodal distribution of diameters in the newborn, typically a few tens of nanometers, but assume a bimodal distribution in mature tendons and ligaments with mode diameters typically ~100–300 nm (11,12). The ends of fibrils are rarely seen in electron micrographs, and even when they are seen, it is not clear whether they are artifacts of sectioning. Fibrils appear to grow by end-to-end addition of short fibrils and there is evidence from 12-week old mouse skin that fibril tips may eventually fuse with the central region of other fibrils to create a meshwork (13). It is not known whether this happens in tendon or ligament or whether the fibrils are as long as the tendon or ligament. Fibrils are arranged into fascicles, which are ~80–300 μm in diameter and these have a "crimped" morphology that is seen most clearly using polarized light (6,9). This crimp is generally described as a planar zigzag with a sharp change in direction of the collagen fibrils with a periodicity of ~200–300 μm (Fig. 2). On application of a tensile load, initial elongation of the fiber requires a relatively low stress because it simply leads to removal of the crimp (14,15). Once the crimp is removed, the fibers become very stiff. This crimp structure, which is also found in ligaments, explains the characteristic shape of the stress–strain curve for ligaments and tendons (14).

There are many studies of the mechanical properties of tendon and ligament. Most tendons have similar properties, because of their role transmitting forces from muscle to bone and the need to do this most efficiently (16–19). In contrast, every ligament has unique properties that fit it to its function at that particular site in the body. These may range, for example, from the highly extensible ligamentum flavum, helping control spinal posture, to the relatively stiff cruciate ligaments in the knee. Most information on the mechanical properties of collagen has been inferred from experiments on tendon; and though they contain a large proportion of collagen, ~70–80% of the dry weight, the fibrils are surrounded by matrix, and therefore the tissue is really a composite material (6). This makes it difficult to separate the behavior of the individual components.

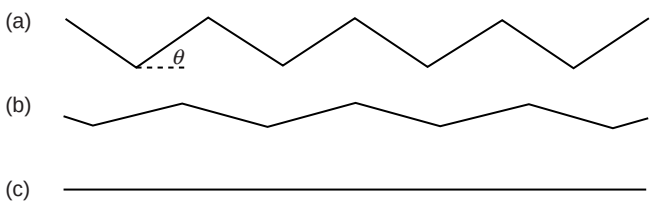


Figure 2. Schematic diagram of the crimp structure in collagen fibers seen from the side. (a) relaxed state [this is grossly exaggerated; the true crimp angle (θ) is ~15°]. As the fiber is stretched, the crimp straightens out (b) until at strains of a ~0.03, the fiber becomes straight (c), at which point it becomes much stiffer.

Proteoglycans

The proteoglycans found predominantly in tendon and ligament belong to the small leucine-rich proteoglycan (SLRP) family; decorin, biglycan, lumican, and fibromodulin though there are small amounts of the large proteoglycans aggrecan (20,21) and versican (21,22). The SLRPs all comprise a repeating structure that is rich in leucine residues, between 13 and 16% of all residues (23). They are present in different tissues in differing amounts and appear at different stages of development (24). Their function is poorly understood though gene knockout studies in mice have shown marked osteopenic effects on bone, skin laxity, and irregular collagen fibers in tendon (25). Most of these SLRPs have one or two glycosaminoglycan (GAG) chains of varying lengths attached, generally either dermatan sulfate or chondroitin sulfate, which can interact electrostatically with collagen (26,27). The proteoglycan decorin has been found to be localized in the tissue to a specific region of a collagen fibril (28). It is also reported to play a regulatory role in collagen fibrillogenesis, by affecting fibril radius (13), and increases the strength of uncrosslinked fibers (29,30). In regions where tendons pass over bone and are subjected to compressive loading in addition to tension, a fibrocartilaginous tissue containing large amounts of aggrecan and biglycan develops (31,32). The adhesive glycoproteins include fibronectin and thrombospondin (33–35), both of which contain possible attachment sites for cells. In humans, these were reported to be more common in the tendon sheath than in the fibrous bulk (36) and fibronectin was found to be up-regulated at sites of injury (34).

The matrix is weak in shear; that is, if it is loaded in compression, it tries to slide out sideways unless it is contained. This behavior may be readily seen in jelly (Jello in the United States) and is not surprising given its high water content, since fluids simply flow when sheared. It is not easy to measure the shear strength of matrix. Proteoglycans at similar concentrations have a very low shear strength (37); however, matrix may be stiffer than this because of the interactions between its macromolecular components. An analysis of the behavior of tendon suggests that its matrix would have a shear modulus of ~100 kPa (38). Because of this low stiffness in shear, the matrix alone is not well suited to bearing loads. Also, its proportion in ligament and tendon is quite low, ~10–20% of the dry weight. The ways in which matrix may transmit stress to the fibers and prevent crack propagation will be discussed later.

Elastic Fibers

Electron microscopy reveals the presence of elastic fibers in ligaments and tendons (39,40). Elastic fibers have two components; elastic fiber microfibrils and elastin. The microfibrils have a diameter of 10–12 nm and consist of glycoproteins. Elastin is composed of mainly hydrophobic nonpolar amino acids with a high content of valine (41).

Elastic fibers are highly extensible, they do not creep and their extension is reversible at high strains. Their mechanical properties are thus very different from collagen. Most of our knowledge of elastic fibers comes from experiments on ligamentum nuchae, a ligament from the cervical spine, which contains ~70% elastin by dry weight (42). Elastin closely resembles a rubber in many respects and its mechanical properties are certainly very similar (43). Purified samples of ligamentum nuchae will extend to roughly twice their resting length before breaking.

The extensibility of a tendon or ligament depend in part on the elastin content of the tissue. Ligamentum flavum, from the spine, which may typically be strained to ~50% contains roughly 70% elastin, by dry weight (44), whereas tendon, which works at strains <4% contains only 2% elastic fibers by dry weight (1). It is fairly easy to see why highly extensible tissues have a high proportion of elastin, but not quite as easy to explain the presence of elastic fibers in a relatively inextensible tissue such as tendon. A clue is provided by some synthetic fibrous composite materials that contain two different kinds of fiber (45). Here a small proportion of strong, low stiffness fibers added to the composite produces a material that is less susceptible to failure under sudden application of load than one that contains only stiff fibers; that is, it makes the material less brittle. It may be that the small proportion of elastic fibers in tendon provide some protection against the sudden application of load that may occur, for example, if an animal is startled.

Fiber–Matrix Interactions

The combination of strong fibers in a weak matrix leads to materials that are less susceptible to mechanical damage while maintaining a high proportion of the strength of the fibers. In particular, they are less susceptible to sudden failure than a homogeneous material would be; a property called "toughness" (46). This composite nature has been recognized for many years (6) and provides a theoretical framework for understanding the properties of the tissues. It also enables some useful comparisons to be made between relatively simple synthetic composites and biological tissues in which the complexities of composition and structure make modeling very difficult. The aim is to obtain an understanding of how the similarities in the tissues, fibers in a matrix, enable them to function in general terms before considering the differences (in composition, and organization), which give them their specific properties. The function of collagen fibrils and fibers in such a composite is to withstand axial tension, since, like any rope, they have little resistance compression and flexion (7). As the tissue is stretched the matrix will try to flow and this will exert a shear force along the surface of the collagen fibers tending to stretch and orient them (7,47). This length increase, which is normally expressed as a fraction of the original length and is then termed "strain", leads to a restoring force in the fiber that balances the applied force. The behavior is rather like a loaded spring that stretches to enable it to bear load, but returns to its relaxed length on removing the load. Similarly, collagen fibers are able to reinforce a tissue if they are oriented so that an applied load tends to stretch them. The nature of the shear force exerted by the gel is unknown, but two simple models, those of elastic and plastic (or frictional) stress transfer, have been used to investigate stresses in the fibers and the force that has to be generated at the fiber surface to enable them to function in this way (47–49). Fibers that are shorter than the tissue are still able to provide reinforcement (50). Some fibrils observed in tissues (51) and those grown *in vitro* (52,53) appear to be tapered, rather than having a uniform radius. Analytical and finite element models of idealized single-fiber composites have shown that tapered fibers have two distinct advantages over uniform fibers: the axial stresses within the fiber are more uniformly distributed and they contain a much smaller amount of material, though their effectiveness at reinforcing is just as great. A more uniform stress within the fiber means more of the fiber is carrying a significant stress, thus making better use of the fiber, and avoids the generation of stress concentrations that are potentially damaging and could lead to fiber fracture. In addition, the volume of material in a cone, for example, is only one-third of that in a straight cylinder and, therefore, a tapered fiber incurs a far smaller metabolic cost by the cells to produce it. In straight-cylindrical fibers it has been calculated that interactions at the fiber surface do not have to be great in order to load fully the fiber. In tendon, assuming conservatively only one interaction per 67 nm D-period, it was estimated that fiber–matrix interaction forces of the order of only 10 pN was sufficient to load fully the fiber (47). These forces are similar in magnitude to van der Waals forces or hydrogen bonds. This suggests that permanent bonds or covalent interactions between fiber and matrix are not essential for the mechanical functioning of the tissue though, of course, it does not preclude them. Regulating the interaction between fibers and matrix is clearly important in this model of how the tissues function and decorin, as described above, is a prime candidate for a role in this. Changes in the concentration and orientation of collagen and its interactions with the matrix have been used to explain the dramatic changes in a similar fibrous tissue, the uterine cervix, that occur during parturition (54). The presence of the matrix around the collagen fibrils is also important when it comes to preventing crack propagation. This will be considered in more detail in the context of the tissues themselves.

LIGAMENT

Ligaments are short bands of tough, but flexible, fibrous connective tissue that bind bones together and guide joint motion, or support organs in place. The word ligament is derived from the Latin word "ligare," which means to bind. Generally, ligaments can be classified into two major subgroups. There are those connecting the elements of the skeletal system (usually crossing joints) and those connecting other soft tissues, such as the suspensory ligaments in the abdomen. This section only considers skeletal ligaments. The main function of the skeletal ligaments, such as the anterior cruciate ligament (ACL) of the knee joint, is to stabilize and control normal kinematics, to prevent

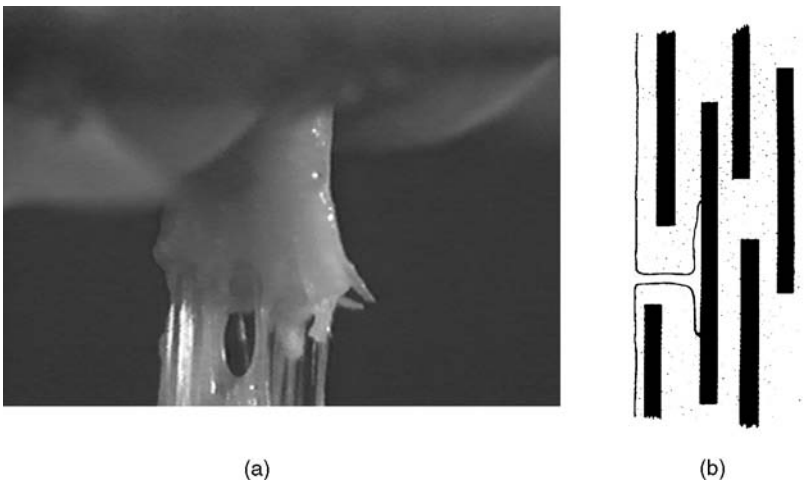


Figure 3. (a) Video image showing a ligament failing under tensile strain. (b) A schematic diagram showing a longitudinal section through a fiber composite illustrating how a weak matrix prevents crack propagation from one fiber to another and dissipates its energy by creating cracks in directions other than across the composite.

abnormal displacements and rotation that may damage the articular surfaces (2,55). At the insertion to bone, the ligaments change from flexible ligamentous tissue to rigid bone, mediated by a transitional zone of fibrocartilage and mineralized cartilage. This helps to prevent stress concentration at the attachment site by allowing gradual change in stiffness (56,57).

Similar to tendon, ligaments are primarily composed of collagen embedded in a weak matrix (7). The collagen molecules in ligaments pack together to form fibrils and the fibrils aggregate into larger fibrous bundles (2). As described previously, the function of collagen fibrils is to provide tensile reinforcing for the matrix. The proportion of fibers within the ligamentous structure and the orientation of the collagen fibers are the main factors that govern the mechanical behavior of the tissue (7). In contrast to tendon, there is commonly less collagen, which is less highly oriented than in tendon, conferring a generally greater extensibility to these tissues.

The collagen fibrils also prevent damaged tissues from failing suddenly. For example, most ligaments do not tear straight across when they are damaged (Fig. 3). Instead, small tears in the matrix are diverted when they encounter the strong collagen fibrils (see below on failure). There is then a possibility that a damaged ligament can heal while, in the meantime, retaining the ability to withstand some load. Some ligaments, however, (e.g., the ACL), have a limited ability to heal when ruptured and this means that it often needs to be replaced or reconstructed when ruptured. A brief summary of different options available for treating ruptured ligaments will be presented in a later section.

Unlike tendons that all have very similar composition, structure, and function, the same cannot be said of ligaments, and it is far harder to make general comments about their properties. Much less is known, too, about the relationship between their structures and functions as the arrangements of their collagen fibrils are more complex than those in tendons (58,59). This greater complexity is understandable when it is realized that the function of a ligament is very dependent on its position in the body; for example, the medial collateral ligament in the knee of a sheep operates at strains of ~0.02 (60), whereas the ligamentum flavum of the human spine operates at strains of up to 0.6. Figure 4 shows that ligamentum flavum is much less stiff than tendon.

It is not surprising that some ligaments contain a high proportion of elastin (~60–70% of the dry weight), which enables them to withstand the high strains to which they are subjected without fracture (61). Ligaments are viscoelastic, that is their properties are time dependent and they appear stiffer if stretched more rapidly. These ligaments exhibit hysteresis, that is, they lose energy on being taken through a cycle of stretching and relaxing. Tkaczuk (61) published a detailed account of the mechanical properties of longitudinal ligaments from the human spine. These deform elastically up to strains of ~0.25, when the stress is ~5 MPa, and rupture at a stress of ~20 MPa. Shah et al.(62) also showed that, like tendons, the collagen fibers are crimped and this crimp disappears at strains of ~1.2–2.8% depending on the ligament. When ligaments are cut from the joint, they can often be seen to contract rapidly, suggesting that they are held in a state of tension even when the joint is in a relaxed state.

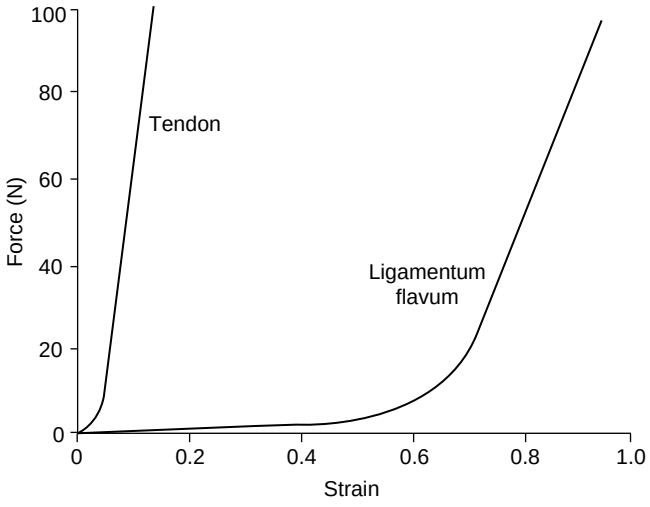


Figure 4. Comparison of force–strain curves obtained for extension of tendon and ligamentum flavum.

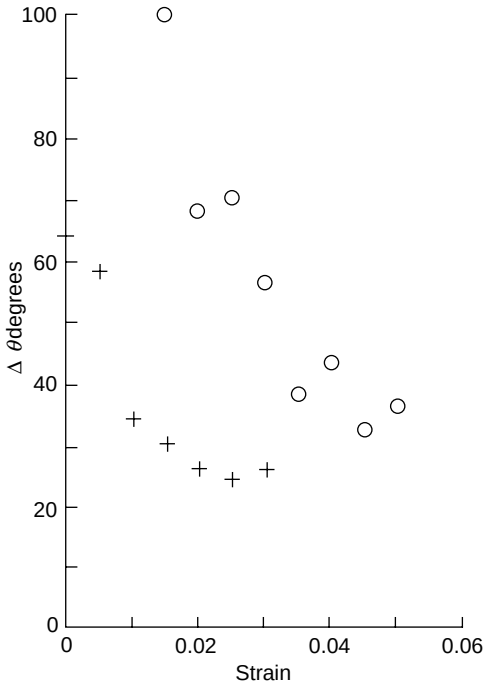


Figure 5. Full width at half-maximum (fwhm), Δ, of the distribution of orientations of collagen fibers in ligamentum flavum. (o) and posterior longitudinal ligament, (+) as a function of strain.

Spinal ligaments provide a good example of the mechanical function of ligaments in a joint (58). The longitudinal ligaments and the ligamentum flavum act together with the intervertebral disc to achieve a mechanically stable joint and serve to limit its mobility. The ligamentum flavum is almost twice as far from the axis of rotation in forward bending as the posterior longitudinal ligament, and hence it can be seen that it needs to be roughly twice as extensible (58,63). This is partly explained by a higher elastin and lower collagen content in ligamentum flavum (61), but also by a less highly aligned organization of collagen fibers (63). As the ligament is stretched, the fibers become more highly aligned and this will increase the stiffness of the tissue, that is its resistance to extension. X-ray diffraction experiments have measured the spread of orientations of fibrils in these ligaments and modelling has shown how this decreases with increasing extension (47). Figure 5 shows that the width of the distribution, $\Delta\theta$, as measured at half the peak height, is greater for ligamentum flavum than posterior longitudinal ligaments. This mechanism provides an explanation for the form of the force–strain curve, shown in Fig. 4, One final point about ligaments is that they have a nerve supply that makes them potential sources of pain. It has also been suggested that they may function as proprioceptors as part of a reflex arc, that is, the ligaments would act as sensors to detect the position of a joint and the information would then be used to control the muscles around the joint thereby controlling its movement and stability (64).

In summary, ligaments are composite materials containing crimped collagen fibers that are prestressed in the relaxed joint. They have a nonlinear stress–strain curve and are viscoelastic. The collagen fibers are relatively disoriented in the unstretched tissue and become more highly aligned as the tissue is stretched. They often contain a proportion of elastin. Their composition and structure depend on their position in the body and their dynamic behavior, that is, the change in structure with strain, becomes more important.

TENDON

The function of tendon is to transmit the force generated by a contracting muscle to the correct point of application on a bone so as to manipulate a joint. Tendons are often preferable to direct attachment of muscle to bone because of various functional requirements. Muscles have a low tensile strength, defined as load at fracture per unit cross-sectional area. This means that they must have a large cross-sectional area in order to transmit sufficient force without tearing. Around many joints (e.g., the fingers), there is insufficient space to attach many muscles. The muscle, therefore, is located further away and attachment made by a tendon, which may be tens of centimeters long in the hand and forearm. Tendons, therefore, need to be strong, so that they can be relatively slender, and stiff, so that the force developed by the muscle is transmitted to the bone without energy being wasted on stretching the tendon. A graph of force as a function of strain for a tendon is shown in Fig. 6. This type of curve is obtained by subjecting the tendon to various forces and recording the amount by which the tendon is stretched, and many of the early studies have never been bettered (6,9,18,40,65). The steepness of the stress–strain curve is a measure of the stiffness of the material. Figure 6 shows that for small strains the tendon requires very little force to stretch it but thereafter it stiffens considerably. The stress at this point is ~10 MPa, which is several orders of magnitude greater than the stresses needed to shear the matrix (6). If the stiffness did not increase, a muscle would continue to stretch the tendon and the force would not be transmitted to the bone.

Tendons are believed to function in the body at strains up to ~0.04 (66,67). Beyond this strain, a tendon does not return to its original length when the applied stress is removed. A tendon will break at strains of ~0.1 (67). The

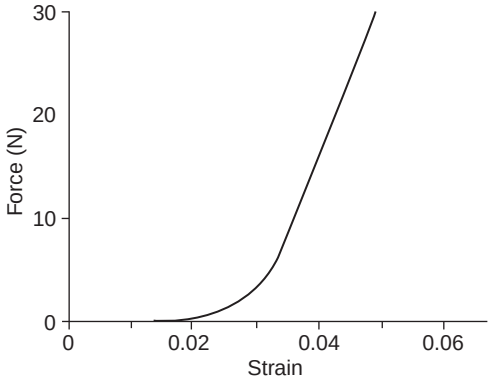


Figure 6. Force–strain curve for tendon.

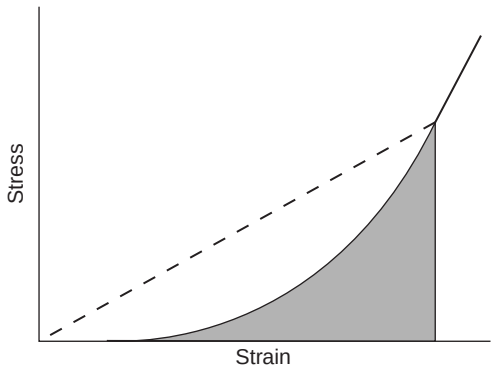


Figure 7. Schematic of stress–strain showing energy stored in a stretched tendon.

initial stages of tendon extension involve straightening the crimp of the collagen fibers described previously.

The energy stored in the stretched tendon is given by the area under the stress–strain curve, as shown in Fig. 7, and it is this energy that must be used to create a tear in the tissue. This energy is lower in a material with a J-shaped stress–strain relationship than if, for example, it were linear. Minimizing the energy available to cause a fracture in this way gives the tissue a property known as resilience, that is, a tendon does not suddenly fail if it is overloaded—unlike a steel wire. While the crimped collagen fibers are being straightened, the weak matrix must be sheared. Because of the fluid-like nature of the matrix, it tends to flow. The rate of flow depends on the force applied to it and the amount of flow is greater the longer the force is applied. The result is that tendons are "viscoelastic" (18,38,68). The effect of the rate at which force is applied to the tendon is shown in Fig. 8.

This time dependence of mechanical behavior leads to a phenomenon known as "creep". For example, when a load of 10 N was applied to a human flexor digitorum tendon, the initial strain was 0.015, but 100 s later under the same load, the strain had increased to 0.016 (69). Thus, if a tendon is stretched rapidly, the matrix has less chance to flow and creep in the material is less resulting in the tendon being stiffer as shown in Fig. 8. Viscous flow of the matrix also provides a mechanism for dissipating energy; more work is done in stretching a tendon than is recovered when the tendon is allowed to relax. This phenomenon is known as hysteresis and is illustrated in Fig. 9. The behavior of a tendon is therefore intermediate between that of a steel wire that stores all the energy used to stretch it, and a viscous liquid that simply flows to a new position and does not store any of the energy put in to cause it to flow.

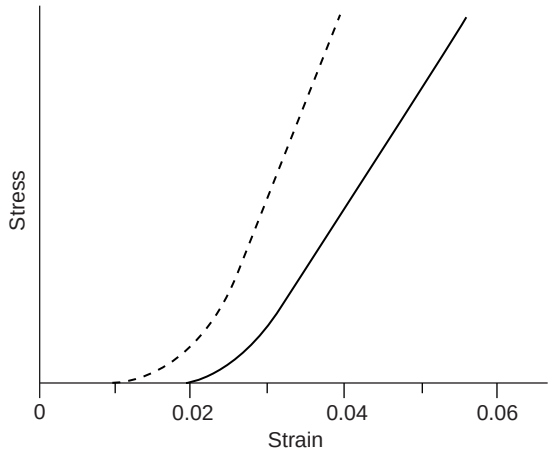


Figure 8. Schematic stress–strain curves for tendon to show the effects of different loading rates. These curves correspond to slow loading (continuous curve) and rapid loading (dashed curve).

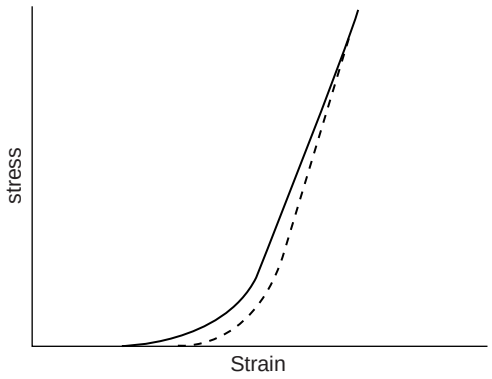


Figure 9. Stress–strain curves for tendon that is stretched (continuous curve) and then relaxed (dashed curve), showing hysteresis; that is, not as much energy is recovered on relaxing as was initially used to stretch the tissue.

MEASURING THE PROPERTIES OF LIGAMENTS AND TENDONS

To quantify the physical properties of ligaments and tendons, mechanical testing of bone–ligament/tendon–bone complexes is often performed (70–73). That this method is often used is partly due to the difficulty in testing isolated ligaments and tendons. Ideally, testing isolated ligaments and tendons would provide measures of the material properties of the tissue alone, but such tests are complicated by difficulties in effectively securing the cut ends (2). Putting the free ends in clamps often results in stress concentrations at the grips, which may contribute to premature failure. Although the use of bone–ligament/tendon–bone complexes still provides more secure clamping, it also increases the difficulty of separating the properties of the ligament from those of the insertion sites. When such complexes are subjected to tensile loading, the resulting load-displacement curve represents the mechanical properties of the bone–ligament/tendon–bone complex as whole rather than specifically about the material that makes up the ligament or tendon.

In order to obtain the material properties of the ligaments, one needs to measure their length (to calculate strain from deformation divided by original length) and cross-sectional area (to calculate stress from applied force divided by original area). From stress–strain curves material properties such as Young's modulus (slope of stress–strain curve), maximum stress, maximum strain, and energy density (area under stress–strain curve) can be

determined. Special devices such as buckle transducers (74) and Hall-effect displacement transducers (75) have been used to measure ligament strains during testing. The drawback of such devices is that they rely on direct contact with the tissue sample and that may influence the results. Optical analysers have also been employed as a noncontact method to measure ligament strains (2). However, inaccuracies may occur because the irregular dye blobs used as markers change shape on stretching making it difficult to define unique points. Another technique that has been employed to measure strain is the use of video dimension analyser (VDA) (76,77). This method requires no direct contact with the specimen, but relies on a recorded video image.

The irregular and complex shape and geometry of ligaments and tendons also make it difficult to measure their cross-sectional area. Although flexible callipers, which are able to follow contours better, have been used to measure ligament cross-sectional areas (2), they still require contact, and this results in errors in measurements. Other investigators have calculated the cross-sectional area of a known length of ligament from measurements of its density by floatation in a mixture of xylene and carbon tetrachloride (78). A number of noncontact methods, such as the use of a rotating microscope (79), use of the VDA (80) and the laser micrometer (81) have also been employed to measure the cross-sectional area.

When mechanical properties of ligaments and tendons are being determined, it is important to consider the rate at which they are loaded since they are viscoelastic, that is, their mechanical properties depend on the rate at which they are deformed (82–86). This sensitivity to strain rate means that ligaments and tendons exhibit properties of stress relaxation (decreased stress with time under constant deformation) and creep (increased deformation with time under constant load) (87,88).

Because of the complex geometry of tendons and ligaments, the orientation of the specimens during mechanical testing affects their physical properties and the manner in which they fail, and should therefore be taken into account when performing mechanical tests. Torsion has been implicated as a factor in the rupture of the ACL during sporting injuries (89–91). Cyclic loading has also been found to lower the yield point or soften the ligament by increasing its compliance (decrease the slope of the linear region of the stress–strain curve) (55). Azangwe et al. (92) showed that, when combined, tension–torsion loading affects both structural and mechanical properties of anterior cruciate ligaments.

LIGAMENT AND TENDON FAILURE MECHANISMS

Since ligaments and tendons consist of collagen fibers reinforcing a weak matrix, it is reasonable to compare their behavior under tensile loading with that of synthetic fiber-reinforced composites, since their failure mechanisms are well established. This section describes some of the modes of failure of fiber-reinforced composites and how they are related to those of the ligaments and tendons. Detailed accounts of failure modes of synthetic fiber-reinforced composites can be found in textbooks, for example, Agarwal and Broutman (45), and Kelly and Macmillan (50). When a material is subjected to any kind of loading, it can absorb energy by two basic mechanisms:

1. Material deformation.
2. Creation of new surfaces.

Material deformation occurs whenever a material is subjected to load. However, if the energy supplied is sufficiently large, cracks may be initiated. Whether they propagate depends on the relative amounts of energy required to create new surfaces compared with that stored in the deformed matrix. For brittle materials such as glass, the energy required to create a new fracture surface is small and though only a small amount of deformation takes place, this is elastic and the associated energy is sufficient to propagate the crack. This means that brittle materials have a low energy-absorption capability. On the other hand, for ductile materials, large plastic deformations occur, which dissipate energy, resulting in large energies being absorbed rather than being available to drive a fracture. This finding shows that the total energy-absorbing capability or "toughness" of a material can be enhanced by increasing either the length of the path of the crack during separation or the material-deformation capability. In metals, the latter mechanism frequently occurs and metallurgical processes are developed to maintain ductility. In composites, replacing low energy-absorbing constituents with greater energy-absorbing constituents can enhance the toughness.

As is the case with many materials, failure in a fiber-reinforced composite emanates from small inherent defects in the material. Several failure events may occur during the fracture of a fiber reinforced composite material, such as,

1. Microcracking of the matrix.
2. Separation of fibers from the matrix (debonding and pull-out).
3. Breaking of fibers,

Forcing a crack to take a longer path is the main mechanism encountered in fiber composites to increase toughness. In fibrous materials, when a matrix crack encounters a strong fiber placed perpendicular to the direction of crack propagation, if the crack cannot cross the fiber because it is too strong, the crack is forced to branch to run parallel with the fiber (Fig. 10). If the crack goes right around the fiber, this may result in fiber debonding, and pull-out, that is, becoming detached from the matrix and pulled out leaving fiber ends showing as the material is stretched. In many cases the surface area produced by secondary cracks is much larger than the area of the primary cracks. This may increase the fracture energy many times and is an effective way of increasing the toughness of composites or the total energy absorbed during fracture. Fibers may eventually fracture when their strength is exceeded. For most synthetic fiber-reinforced composites, fibers are separated by matrix and therefore are unable to pass energy directly

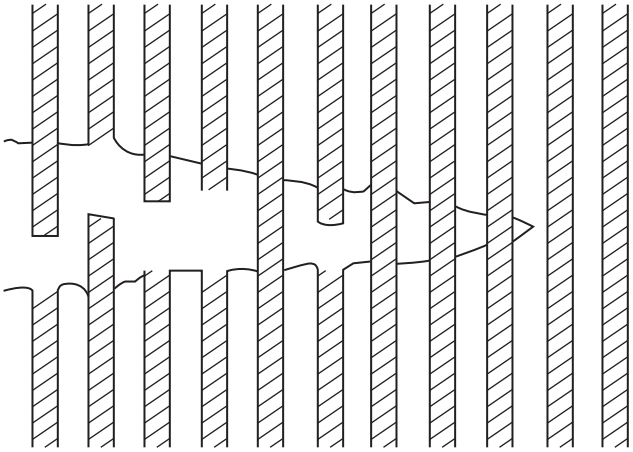


Figure 10. Model of crack tip in a fiber-reinforced composite showing how a crack may propagate Fibers become debonded from the matrix and are pulled out as the crack widens.

from one to another, hence it is unlikely that they all fail together. Collagen fibers in ligaments are similarly separated by the matrix, and hence the transmission of stress from fiber to fiber in the ligament is indirect (7). The situation is more complicated in biological tissues because of the complexity of the components and their arrangement within the tissue and because of the possibility of repair. It seems reasonable to assume that a crack will start first in the weak matrix rather than in the strong collagen fibers. It is unlikely that the crack will spread into a fibril, as previously discussed, but it will be deflected by the fibrils into new directions.

It appears that tendons and ligaments can continue to withstand stress long after they are damaged, but before complete fracture occurs (70,71,93). However, since damage implies an irreversible change, at least until the biological repair process begins, the tissue will not return to its original dimension when the stress is removed. When testing bone–tendon/ligament–bone complexes, an additional failure mode may occur at the ligament insertion to bone. For example, the ligamentum flavum tears at the enthesis, the junction with the bone, leaving virtually no possibility of natural healing (72). Fortunately, this injury does not appear to occur in vivo and can occur in isolation or in combination with other failure modes. As mentioned previously, most ligaments appear to be prestressed within the body so that if they are severed or avulsed they retract and the damaged ends are no longer in contact. This makes it very difficult for cells within the tissue then to bridge the gap by synthesizing new matrix.

The tearing of tendons is a fairly common injury, though rupture of the tendon tends to occur at the junction with bone. Healing of torn tendons in humans is slow, and is not always improved by surgical intervention (3). Healing starts with the invasion of cells into the damaged area that first lay down fine, poorly oriented collagen fibrils (94). These fibrils later increase in diameter and become increasingly oriented as the cell population drops and the structure becomes more akin to that of the original tendon (95).

Damage to and repair of ligaments follows a very similar pattern to that described for tendons. A point to note in all this is that there are differences in healing characteristics between different ligaments. The ACL of the knee joint, for example, appears to have a poor healing capacity when injured prompting the need for reconstruction. At the junctions of ligaments to bone, the fibers of the ligament become more compact, then cartilaginous and finally calcified before finally merging into the bone and there are changes in cell phenotype and expressed proteins (96,97). This complex structure reflects the difficulty of attaching a tough, flexible material to a hard, brittle one.

LIGAMENT AND TENDON REPAIR

Because of the slow rate of healing of certain ligaments and tendons, their reconstruction with synthetic materials has been attempted, with varying degrees of success, for many years (98). The area of prosthetic materials and methods is so vast that only a very brief survey will be attempted here. The main approaches that have been tried are replacement using tissues taken from another part of the body, or from an animal, and complete substitution by a synthetic material. None of these approaches has so far proved entirely satisfactory. Success rates of 80–90% are reported for both techniques (99–101), but some long-term studies have shown that this success rate may fall to 40–50% after $\sim$5 years (102) and there are reports of high, $\sim$50%, incidence of degenerative change (101,103). Friedman et al. (104), reviewed the autogenous reconstruction of the ACL using patellar tendon, iliotibial band, gracilis tendon, semitendinosus, or meniscus that gives some idea of the range of tissues that have been tried. Unfortunately, most repairs stretch with time resulting in loss of stability (105); this is probably due to the differences in structure and mechanical properties between the original tissue and the replacement, described earlier. If the replacement tissue is stretched too much while being fixed in position, it may be irreversibly strained.

Early attempts at the synthetic replacement of ligaments and tendons, using silk, for example, were not successful. These prostheses were intended to be permanent, but had not the strength and fatigue resistance to withstand the millions of cycle of loading imposed on them during the lifetime of the recipient. More recently polyester (106), carbon fiber (107,108), or various combinations of synthetic materials and autogenous tissue have all been tried but still seem not to overcome this particular problem (103,109,110).

Another approach that is currently being explored for augmenting or reproducing ligaments and tendons is that of tissue engineering (111,113), though this has not yet reached the stage of clinical utility. These techniques may include the development of biodegradable scaffolds, on which it is hoped to encourage cells from the patient to grow a replacement tissue (113), and growth factors (114) and their introduction into the tissue using gene transfer (112,115).

SUMMARY

Tendons and ligament are connective tissues subject primarily to tensile forces. They comprise crimped collagen

fibrils and some elastin embedded in a weak matrix. Collagen fibers in tendon, composed of bundles of fibrils, are highly aligned along the direction of applied force. The initial structural response to the application of force is straightening of the crimp, which occurs at a strain of ~2%, after which the tendon stiffens considerably. Higher strains are not immediately reversibly and may lead to structural damage to the tissue. The function of the crimp appears to be to minimize the energy stored in the stretched tissue thus reducing the energy available to cause fracture. Many ligaments can be stretched more than tendons, because of their more complex collagen fibril organization and, sometimes, the high proportion of elastin present. Tendons and ligaments are viscoelastic; they do not store all the energy used to stretch them, their response to load depends on the rate at which it is applied, and they continue to deform even if the applied load remains constant. Because a lot of energy is required to produce a large fracture surface, they do not break easily, that is, they are tough materials. Because some ligaments do not heal well when injured, there is a need to replace them. However, success in this area is still limited. Future replacements may include tissue engineered ligaments.

ACKNOWLEDGMENTS

We thank Professor D.W.L. Hukins for valuable discussions over many years and the many other colleagues who have contributed to some of the work we have described here. We also thank the Medical Research Council, the Engineering and Physical Sciences Research Council, and the Arthritis Research Campaign for financial support.

BIBLIOGRAPHY

1. Fank CB, Shrive NG. Ligament. In: Nigg BM, Herzog W, editors. Biomechanics of the Musculo-Skeletal System. Chichester: J Wiley; 1999.
2. Frank CB, Hart DA, Shrive NG. Molecular biology and biomechanics of normal and healing ligaments—a review. Osteoarthritis Cartilage 1999;7:130–140.
3. Maffulli N. Rupture of the Achilles tendon. J Bone Joint Surg 1999;81-A:1019–1036.
4. Woo SL, Debski RE, Zeminski J, Abramowitch SD, Saw SS, Fenwick JA. Injury and repair of ligaments and tendons. Annu Rev Biomed Eng 2000;2:83–118.
5. Nachemson AL, Evans JH. Some mechanical properties of the third human lumbar interlaminar ligament (ligamentum flavum). J Biomech 1968;1:211–220.
6. Elliott DH. Structure and function of mammalian tendon. Biol Rev 1965;40:392–421.
7. Hukins DWL, Aspden RM. Composition and properties of connective tissues. Trends Biochem Sci 1985;10:260–264.
8. Brodsky B, Ramshaw JA. The collagen triple-helix structure (Review). Matrix Biol 1997;15:545–554.
9. Kastelic J, Galeski A, Baer E. The multicomposite structure of tendon. Connect Tissue Res 1978;6:11–23.
10. Vincent JFV. Structural Biomaterials Princeton. 2nd ed. New Jersey: Princeton University Press; 1990.
11. Parry DA, Barnes GR, Craig AS. A comparison of the size distribution of collagen fibrils in connective tissues as a function of age and a possible relation between fibril size distribution and mechanical properties. Proc R Soc London Ser B Biol Sci 1978;203:305–321.
12. Patterson-Kane JC, Wilson AM, Firth EC, Parry DA, Goodship AE. Comparison of collagen fibril populations in the superficial digital flexor tendons of exercised and nonexercised thoroughbreds. Equine Vet J 1997;29:121–125.
13. Kadler KE, Holmes DF, Graham H, Starborg T. Tip-mediated fusion involving unipolar collagen fibrils accounts for rapid fibril elongation, the occurrence of fibrillar branched networks in skin and the paucity of collagen fibril ends in vertebrates. Matrix Biol 2000;19:359–365.
14. Buckley CP, Lloyd DW, Konopasek M. On the Deformation of Slender Filaments with Planar Crimp: Theory, Numerical Solution and Applications to Tendon Collagen and Textile Materials. Proc R Soc London Ser A, Math Phys Sci 1980; 372:33–64.
15. Diamant J, Keller A, Baer E, Litt M, Arridge RGC. Ultrastructure of collagen as a function of ageing. Proc R Soc London Ser B Biol Sci 1972;180:293–312.
16. Betsch DF, Baer E. Structure and mechanical properties of rat tail tendon. Biorheology 1980;17:83–94.
17. Haut RC, Lancaster RL, Decamp C. Mechanical properties of the canine patellar tendon—some correlations with age and the content of collagen. J Biomech 1992;25:163.
18. Woo SL. Mechanical properties of tendons and ligaments. I. Quasi-static and nonlinear viscoelastic properties. Biorheology 1982;19:385–396.
19. Woo SLY, Sites TJ. Current advances on the study of the biomechanical properties of tendons and ligaments. In: Nimni ME, editor. Collagen, Volume II, Biochemistry Biomechanics. Boca Raton: CRC Press; 1988.
20. Campbell MA, Tester AM, Handley CJ, Checkley GJ, Chow GL, Cant AE, Winter AD, Cain WE. Characterization of a large chondroitin sulfate proteoglycan present in bovine collateral ligament. Arch Biochem Biophys 1996;329:181–190.
21. Robbins JR, Vogel K. Regional expression of mRNA for proteoglycans and collagen in tendon. Eur J Cell Biol 1994;64: 264–270.
22. Thomopoulos S, Hattersley G, Rosen V, Mertens M, Galatz L, Williams GR, Soslowsky LJ. The localized expression of extracellular matrix components in healing tendon insertion sites: an in situ hybridization study. J Orthop Res 2002;20:454–463.
23. Hocking AM, Shinomura T, McQuillan DJ. Leucine-rich repeat glycoproteins of the extracellular matrix. Matrix Biol 1998;17:1–19.
24. Wilda M, Bachner D, Just W, Geerkens C, Kraus P, Vogel W, Hameister H. A comparison of the expression pattern of five genes of the family of small leucine-rich proteoglycans during mouse development. J Bone Miner Res 2000;15:2187–2196.
25. Corsi A, Xu T, Chen XD, Boyde A, Liang J, Mankani M, Sommer B, Iozzo RV, Eichstetter I, Robey PG, Bianco P, Young MF. Phenotypic effects of biglycan deficiency are linked to collagen fibril abnormalities, are synergized by decorin deficiency, and mimic Ehlers-Danlos-like changes in bone and other connective tissues. J Bone Miner Res 2002;17:1180–1189.
26. Gelman RA, Blackwell J. Collagen-mucopolysaccharide interactions at acid pH. Biochim Biophys Acta 1974;342: 254–261.
27. Lindahl U, Hook M. Glycosaminoglycans and their binding to biological macromolecules. Annu Rev Biochem 1978;47: 385–417.
28. Scott JE, Orford CR. Dermatan sulphate-rich proteoglycan associates with rat tail tendon collagen at the d band in the gap region. Biochem J 1981;197:213–216.
29. Weber IT, Harrison RW, Iozzo RV. Model structure of decorin and implications for collagen fibrillogenesis. J Biol Chem 1996;271:31767–31770.

30. Pins GD, Christiansen DL, Patel R, Silver FH. Self-assembly of collagen fibers. Influence of fibrillar alignment and decorin on mechanical properties. Biophys J 1997;73:2164–2172.
31. Evanko SP, Vogel KG. Proteoglycan synthesis in fetal tendon is differentially regulated by cyclic compression *in vitro*. Arch Biochem Biophys 1993;307:153–164.
32. Koob TJ, Clark PE, Hernez DJ, Thurmond FA, Vogel KG. Compression loading *in vitro* regulates proteoglycan synthesis by tendon fibrocartilage. Arch Biochem Biophys 1992; 298:303–312.
33. Cockburn CG, Barnes MJ. Characterization of thrombospondin binding to collagen (type I) fibers: role of collagen telopeptides. Matrix 1991;11:168–176.
34. Amiel D, Gelberman R, Harwood F, Siegel D. Fibronectin in healing flexor tendons subjected to immobilization or early controlled passive motion. Matrix 1991;11:184–189.
35. Kannus P, Jozsa L, Jarvinen TA, Jarvinen TL, Kvist M, Natri A, Jarvinen M. Location and distribution of non-collagenous matrix proteins in musculoskeletal tissues of rat. Histochem J 1998;30:799–810.
36. Jozsa L, Lehto M, Kannus P, Kvist M, Reffy A, Vieno T, Jarvinen M, Demel S, Elek E. Fibronectin and laminin in Achilles tendon. Acta Orthop Scand 1989;60:469–471.
37. Mow VC, Mak AF, Lai WM, Rosenberg LC, Tang LH. Viscoelastic properties of proteoglycan subunits and aggregates in varying solution concentrations. J Biomech 1984;17:325–338.
38. Hooley CJ, Cohen RE. A model for the creep behaviour of tendon. Int J Biol Macromol 1979;1:123–132.
39. Kannus P. Structure of the tendon connective tissue. Scand J Med Sci Sports 2000;10:312–320.
40. Parry DA, Craig AS, Barnes GR. Tendon and ligament from the horse: an ultrastructural study of collagen fibrils and elastic fibers as a function of age. Proc R Soc London Ser B Biol Sci 1978;203:293–303.
41. Ross R. The elastic fiber. J Histochem Cytochem 1973;21: 199–208.
42. Minns RJ, Soden PD, Jackson DS. The role of the fibrous components and ground substance in the mechanical properties of biological tissues: a preliminary investigation. J Biomech 1973;6:153–165.
43. Gosline JM. The elastic properties of rubber-like proteins and highly extensible tissues. Symp Soc Exp Biol 1980;34:331–357.
44. Evans JH, Nachemson AL. Biomechanical study of human lumbar ligamentum flavum. J Anat 1969;105:188–189.
45. Agarwal BD, Broutman LJ. Analysis and performance of fiber composites New York: Wiley; 1980.
46. Gordon JE. The new science of stronmg materials. Harmondsworth: Penguin Books Ltd.; 1976.
47. Aspden RM. Fiber reinforcing by collagen in cartilage and soft connective tissues. Proc R Soc London Ser B Biol Sci 1994;B-258:195–200.
48. Goh KL, Aspden RM, Mathias KJ, Hukins DWL. Effect of fiber shape on the stresses within fibers in fiber-reinforced composite materials. Proc R Soc London Ser 1999;A-455: 3351–3361.
49. Goh KL, Mathias KJ, Aspden RM, Hukins DWL. Finite element analysis of the effect of fiber shape on stresses in an elastic fiber surrounded by a plastic matrix. J Mater Sci 2000;35:2493–2497.
50. Kelly A, Macmillan NH. Strong Solids. 3rd ed.; Oxford: Oxford University Press; 1986.
51. DeVente JE, Lester GE, Trotter JA, Dahners LE. Isolation of intact collagen fibrils from healing ligament [letter]. J Electron Microsc 1997;46:353–356.
52. Holmes DF, Chapman JA, Prockop DJ, Kadler KE. Growing tips of type-I collagen fibrils formed in vitro are near paraboloidal in shape, implying a reciprocal relationship between accretion and diameter. Proc Natl Acad Sci USA 1992;89: 9855–9859.
53. Kadler KE, Holmes DF, Trotter JA, Chapman JA. Collagen fibril formation. Biochem J 1996;316 (Pt 1): 1–11.
54. Aspden RM. The theory of fiber reinforced composite materials applied to changes in the mechanical properties of the cervix during pregnancy. J Theor Biol 1988;130:213–221.
55. Cabaud HE. Biomechanics of the anterior cruciate ligament. Clin Orthop 1983; 26–31.
56. Dodds JA, Arnoczky SP. Anatomy of the anterior cruciate ligament: a blueprint for repair and reconstruction. Arthroscopy 1994;10:132–139.
57. Noyes FR, DeLucas JL, Torvik PJ. Biomechanics of anterior cruciate ligament failure: an analysis of strain-rate sensitivity and mechanisms of failure in primates. J Bone Joint Surg Am 1974;56:236–253.
58. Hukins DWL, Kirby MC, Sikoryn TA, Aspden RM, Cox AJ. Comparison of structure, mechanical properties and functions of lumbar spinal ligaments. Spine 1990;15:787–795.
59. Viidik A. Functional properties of collagenous tissues. Int Rev Connect Tissue Res 1973;6:127–215.
60. Claes L, Neugebauer R. In vivo and in vitro investigation of the long-term behaviour and fatigue strength of carbon fiber ligament replacement. Clin Orthop 1985;186:99–111.
61. Tkaczuk H. Tensile properties of human lumbar longitudinal ligaments. Acta Orthop Scand Suppl 1968; 115.
62. Shah JS, Jayson MIV, Hampson WGJ. Mechanical implications of crimping in collagen fibers of human spinal ligaments. Proc Instn Mech Eng [H], J Eng Med 1979;8: 95–102.
63. Kirby MC, Sikoryn TA, Hukins DWL, Aspden RM. Structures and mechanical properties of the longitudinal ligaments and ligamenta flava of the spine. J Biomed Eng 1989;11:192–196.
64. Brand RA. Knee ligaments: a new view. J Biomech Eng 1986;108:106–110.
65. Viidik A. Tensile strength properties of Achilles tendon systems in trained and untrained rabbits. Acta Orthop Scand 1969;40:261–272.
66. Rigby J, Hirai N, Spikes JD, Eyring M. The mechanical properties of rat tail tendon. J Gen Physiol 2003;43:265–283.
67. Haut RC, Little RW. A constitutive equation for collagen fibers. J Biomech 1972;5:423–430.
68. Dorrington KL. The theory of viscoelasticity in biomaterials. Symp Soc Exp Biol 1980;34:289–314.
69. Cohen RE, Hooley CJ, McCrum NG. Viscoelastic creep of collagenous tissue. J Biomech 1976;9:175–184.
70. Kennedy JC, Hawkins RJ, Willis RB, Danylchuck KD. Tension studies of human knee ligaments. Yield point, ultimate failure, and disruption of the cruciate and tibial collateral ligaments. J Bone Joint Surg Am 1976;58:350–355.
71. Neumann P, Keller TS, Ekstrom L, Perry L, Hansson TH, Spengler DM. Mechanical properties of the human lumbar anterior longitudinal ligament. J Biomech 1992;25:1185–1194.
72. Sikoryn TA, Hukins DWL. Mechanism of failure of the ligamentum flavum of the spine during in vitro tensile tests. J Orthop Res 1990;8:586–591.
73. Azangwe G, Mathias KJ, Marshall D. Macro and microscopic examination of the ruptured surfaces of anterior cruciate ligaments of rabbits. J Bone Joint Surg Br 2000;82:450–456.
74. Barry D, Ahmed AM. Design and performance of a modified buckle transducer for the measurement of ligament tension. J Biomech Eng 1986;108:149–152.
75. Beynnon B, Howe JG, Pope MH, Johnson RJ, Fleming BC. The measurement of anterior cruciate ligament strain in vivo. Int Orthop 1992;16:1–12.
76. Woo SL, Newton PO, MacKenna DA, Lyon RM. A comparative evaluation of the mechanical properties of the rabbit

medial collateral and anterior cruciate ligaments. J Biomech 1992;25:377–386.
77. Lam TC, Shrive NG, Frank CB. Variations in rupture site and surface strains at failure in the maturing rabbit medial collateral ligament. J Biomech Eng 1995;117:455–461.
78. Sikoryn TA, Hukins DWL. Failure of the longitudinal ligaments of the spine. J Mater Sci Lett 1988;7:1345–1349.
79. Gupta P, Subramanian KN, Brinker WO, Gupta AN. Tensile strength of canine cruciate ligaments inthe dog. Am J Vet Res 1971;32:183–190.
80. Njus GO, Njus NM. A non-contact method for determining cross-sectional area of soft tissue. Trans Orthopaed Res Soc 1984;32:126–131.
81. Woo SL, Danto MI, Ohland KJ, Lee TQ, Newton PO. The use of a laser micrometer system to determine the cross-sectional shape and area of ligaments: a comparative study with two existing methods. J Biomech Eng 1990;112:426–431.
82. King GJ, Pillon CL, Johnson JA. Effect of in vitro testing over extended periods on the low-load mechanical behaviour of dense connective tissues. J Orthop Res 2000;18:678–681.
83. Kwan MK, Lin TH, Woo SL. On the viscoelastic properties of the anteromedial bundle of the anterior cruciate ligament. J Biomech 1993;26:447–452.
84. Provenzano P, Lakes R, Keenan T, Vanderby, Jr. R. Nonlinear ligament viscoelasticity. Ann Biomed Eng 2001;29: 908–914.
85. Silver FH, Christiansen DL, Snowhill PB, Chen Y. Role of storage on changes in the mechanical properties of tendon and self-assembled collagen fibers. Connect Tissue Res 2000; 41:155–164.
86. Woo SL, Gomez MA, Akeson WH. The time and history-dependent viscoelastic properties of the canine medical collateral ligament. J Biomech Eng 1981;103:293–298.
87. Thornton GM, Oliynyk A, Frank CB, Shrive NG. Ligament creep cannot be predicted from stress relaxation at low stress: a biomechanical study of the rabbit medial collateral ligament. J Orthop Res 1997;15:652–656.
88. Thornton GM, Shrive NG, Frank CB. Altering ligament water content affects ligament pre-stress and creep behaviour. J Orthop Res 2001;19:845–851.
89. Speer KP, Warren RF, Wickiewicz TL, Horowitz L, Henderson L. Observations on the injury mechanism of anterior cruciate ligament tears in skiers. Am J Sports Med 1995;23:77–81.
90. Fischer JF, Leyvraz PF, Bally A. A dynamic analysis of knee ligament injuries in alpine skiing. Acta Orthop Belg 1994;60:194–203.
91. Emerson RJ. Basketball knee injuries and the anterior cruciate ligament. Clin Sports Med 1993;12:317–328.
92. Azangwe G, Mathias KJ, Marshall D. The effect of torsion on the appearance of the rupture surface of the ACL of rabbits. Knee 2002;9:31–39.
93. Woo SL, Orlando CA, Gomez MA, Frank CB, Akeson WH. Tensile properties of the medial collateral ligament as a function of age. J Orthop Res 1986;4:133–141.
94. Davison PF. The organization of collagen in growing tensile tissues. Connect Tissue Res 1992;28:171.
95. Greenlee, Jr. TK, Pike D. Studies on tendon healing in the rat. J Plast Reconstr Surg 1071;48:260–270.
96. Matyas JR, Anton MG, Shrive NG, Frank CB. Stress governs tissue phenotype at the femoral insertion of the rabbit MCL. J Biomech 1995;28:147–157.
97. Moriggl B, Kumai T, Milz S, Benjamin M. The structure and histopathology of the "enthesis organ" at the navicular insertion of the tendon of tibialis posterior. J Rheumatol 2003;30:508–517.
98. Cotton FJ, Morrison GM. Artifical ligaments at the knee: a technique. N Engl J Med 1934;210:1331–1332.
99. Fujikawa K, Kobayashi T, Sasazaki Y, Matsumoto H, Seedhom BB. Anterior cruciate ligament reconstruction with the Leeds-Keio artificial ligament. J Long Term Eff Med Implants 2000;10:225–238.
100. Chen CH, Chen WJ, Shih CH. Arthroscopic reconstruction of the posterior cruciate ligament: a comparison of quadriceps tendon autograft and quadruple hamstring tendon graft. Arthroscopy 2002;18:603–612.
101. Ruiz AL, Kelly M, Nutton RW. Arthroscopic ACL reconstruction: a 5-9 year follow-up. Knee 2002;9:197–200.
102. Schroven IT, Geens S, Beckers L, Lagrange W, Fabry G. Experience with the Leeds-Keio artificial ligament for anterior cruciate ligament reconstruction. Knee Surg Sports Traumatol Arthrosc 1994;2:214–218.
103. Drogset JO, Grontvedt T. Anterior cruciate ligament reconstruction with and without a ligament augmentation device: results at 8-Year follow-up. Am J Sports Med 2002;30:851–856.
104. Friedman MJ, Sherman OH, Fox JM, Del PW, Snyder SJ. Ferkel RJ. Autogeneic anterior cruciate ligament (ACL) anterior reconstruction of the knee. A review. Clin Orthop Relat Res 1985;9–14.
105. Alexander H, Weiss AB. Editorial Comment. Clin Orthop Relat Res 1985;2–3.
106. Fujikawa K, Ohtani T, Matsumoto H, Seedhom BB. Reconstruction of the extensor apparatus of the knee with the Leeds-Keio ligament. J Bone Joint Surg Br 1994;76:200–203.
107. Jenkins DH, Forster IW, McKibbin B, Ralis ZA. Induction of tendon and ligament formation by carbon implants. J Bone Joint Surg Br 1977;59:53–57.
108. Turner IG, Thomas NP. Comparative analysis of four types of synthetic anterior cruciate ligament replacement in the goat: in vivo histological and mechanical findings. Biomaterials 1990;11:321–329.
109. Marumo K, Kumagae Y, Tanaka T, Fujii K. Long-term results of anterior cruciate ligament reconstruction using semitendinosus and gracilis tendons with Kennedy ligament augmentation device compared with patellar tendon autografts. J Long Term Eff Med Implants 2000;10:251–265.
110. Fukubayashi T, Ikeda K. Follow-up study of Gore-Tex artificial ligament—special emphasis on tunnel osteolysis. J Long Term Eff Med Implants 2000;10:267–277.
111. Woo SL, Hildebrand K, Watanabe N, Fenwick JA, Papageorgiou CD, Wang JH. Tissue engineering of ligament and tendon healing. Clin Orthop 1999; S312–S323.
112. Huard J, Li Y, Peng H, Fu FH. Gene therapy and tissue engineering for sports medicine. J Gene Med 2003;5:93–108.
113. Gentleman E, Lay AN, Dickerson DA, Nauman EA, Livesay GA, Dee KC. Mechanical characterization of collagen fibers and scaffolds for tissue engineering. Biomaterials 2003;24: 3805–3813.
114. DesRosiers EA, Yahia L, Rivard CH. Proliferative and matrix synthesis response of canine anterior cruciate ligament fibroblasts submitted to combined growth factors. J Orthop Res 1996;14:200–208.
115. Martinek V, Latterman C, Usas A, Abramowitch S, Woo SL, Fu FH, Huard J. Enhancement of tendon-bone integration of anterior cruciate ligament grafts with bone morphogenetic protein-2 gene transfer: a histological and biomechanical study. J Bone Joint Surg Am 2002;84-A:1123–1131.

Further Reading

Akeson WH, Pedowitz R, O'Connor JJ, editors. Knee Ligaments: Structure, Function, Injury and Repair. 2nd ed. New York: Lippincott Williams & Wilkins; 2003.
Mow VC, Hayes WC, editors. Basic Orthopaedic Biomechanics. New York: Raven Press; 1991.

Evans CH, Scully SP, guest editors. Orthopaedic Gene Therapy. Clinical Orthopaedics and Related Research. Volume 379 Suppl., 2000.

See also BONE AND TEETH, PROPERTIES OF; CARTILAGE AND MENISCUS, PROPERTIES OF.

LINEAR VARIABLE DIFFERENTIAL TRANSFORMERS

SUNIL KESAVAN
Akebono Corporation
Farmington Hills, Michigan

NARENDER REDDY
The University of Akron
Akron, Ohio

INTRODUCTION

Several methods of transduction are available to convert physiological events into electrical signals. Basic physiological variables are first converted by sensing elements into variables that can easily be measured by available transducers. One such transducer, the linear variable differential transformer, commonly abbreviated as LVDT (some manufacturers designate it as LDVT — linear differential voltage transformer), is used to convert mechanical displacement into proportional electronic signals. LVDTs are capable of measuring physiological variables, such as displacement, pressure, force, and acceleration, which are either available in the form of a linear displacement or can be converted into such movement.

THEORY

An LVDT is an inductive electromechanical transducer that uses a primary (energizing) coil and two series-opposed secondary coils. This mode of connecting the secondaries serves to mutually cancel out the secondary voltages. In this popular configuration, due to Shaevitz (1), the primary winding is symmetrically placed with respect to the secondary windings on a cylindrical former. The former surrounds a free-moving rod-shaped magnetic core, which provides a path for the magnetic flux linking the coils (Fig. 1). The magnetic core is connected to a sensing device like a movable diaphragm. Movement of the sensor induces core movement, which in turn produces voltage variations that are measured directly.

When the sliding magnetic core is in the central (null) position, the electromotive forces (emfs) generated in the secondaries are equal, and the net output voltage, e_0 is, therefore, zero. Movement of the core from this central position causes the mutual inductance (coupling) for one coil to increase and the other coil to decrease. The amplitude of the output voltage, e_0, being the difference between the emfs in the two secondaries, varies approximately linearly with the position of the core on either side of the null position (Fig. 2). The differential secondary connection in the LVDT causes the phase of the output voltage to change by 180° as the core passes through the null position. The output voltage, e_0, is generally out of phase with the excitation voltage, e_i. The phase shift is dependent on the frequency of e_i, and each LVDT has a particular frequency at which phase shift is zero.

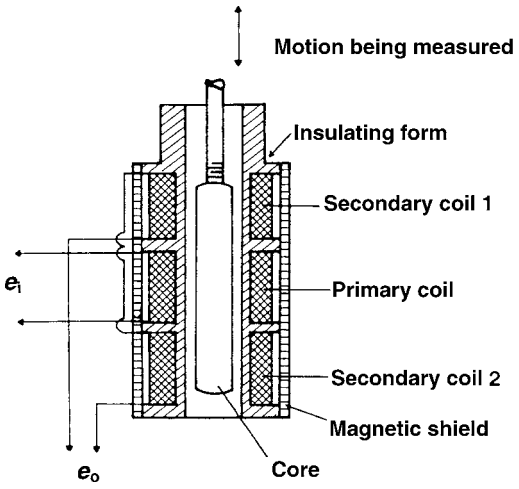


Figure 1. Schematic of the cutaway view of an LVDT showing the core, primary coil, and two secondary coils: (e_i) excitation voltage; (e_0) output voltage.

FABRICATION

The LVDT features essentially frictionless measurement and long mechanical life, because there is no mechanical

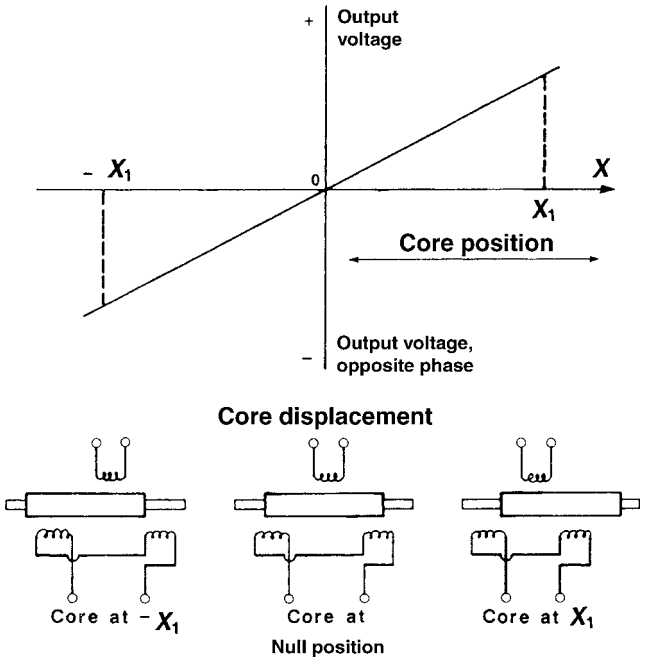


Figure 2. Output voltage of an LVDT as a linear function of core position.

contact between the enclosed coil assembly and the separate freely moving core within the coil assembly (Fig. 1).

A typical alternating current (ac) LVDT core consists of a uniformly dense cylindrical slug of a high permeability nickel–iron alloy. The core is internally threaded to accept nonmagnetic core rods from an external sensing or actuating element. The core moves within a cylindrical coil assembly. Hollow cores are employed when a low mass core is desired. Recently, researchers have developed lightweight glass-covered amorphous wire cores that can be used to fabricate high sensitivity LVDTs with good mechanical and corrosion resistance (2). The primary and secondary windings are spaced symmetrically by winding them on a slotted cylindrical former. To avoid material corrosion or insulation leakage, the windings are impregnated with an insulating varnish under a vacuum. The coils are encapsulated in epoxy for further mechanical and moisture protection. Magnetic or ferromagnetic materials in the proximity of an LVDT can disrupt its magnetic field. The magnetic field of an LVDT may also induce eddy currents into nonmagnetic materials in its vicinity. These currents in turn would create a magnetic flux that would interfere with the LVDT output. These problems are avoided in practice by enclosing the LVDT in a case fabricated from an alloy, a high permeability iron, or a stainless steel. The LVDT assembly is then mounted in a C- or split block. LVDTs that can measure rotational movement are also available.

In addition to the ac LVDTs described in the previous sections, direct current (dc) LVDTs are also available (3,4). These LVDTs, in addition to having all the advantages of ac LVDTs, possess the simplicity of dc operation. They consist of two integral parts: an ac-operated LVDT and a carrier generator–signal conditioning module. The small carrier system eliminates the need for the ac excitation, demodulation, and amplification equipment required for conventional ac LVDTs. This cuts down the cost and reduces the volume of LVDT instrumentation; dc units can be battery operated or be supplied by a simple dc power supply (3,4). Also, any dc meter can be employed to read the LVDT output. These advantages, coupled with the small size of the dc LVDTs, make them attractive for use in hospitals and other medical environments.

The LVDTs have several advantages and a few disadvantages (3–6) as briefly reviewed next.

1. Essentially frictionless operation and long mechanical life: As described in the previous section, the LVDT has no moving mechanical contact between the moving core and the windings. This ensures that LVDTs have a fast dynamic response as no additional load apart from the core mass is imposed on the measured event. In addition, this helps LVDTs to have a long, essentially infinite, mechanical life.
2. Good in hostile environments: LVDTs can be manufactured to withstand the vagaries of chemical corrosion and extremes of temperature and pressure. This is facilitated by the separation between the core and windings of the LVDT. Only a static seal is required to isolate the coil assembly from hostile environments.
3. Extremely high resolution: LVDTs can respond to extremely small displacements. Microdisplacement LVDT transducers capable of measuring displacements down to 100 pm have been fabricated (7).
4. Null repeatability: The null position of an LVDT is very repeatable, even with large temperature variations.
5. Input–Output isolation: Since the primary and secondary windings are isolated from each other, the signal ground can be isolated from the excitation ground.
6. Cross-axis rejection: The LVDT is only responsive to axial core motion. Cross-axis motion induced by conditions such as jarring or continuous vibration will not affect the LVDT output.
7. Overtravel damage resistance: As the LVDT core can pass completely through the coil assembly, the transducer is inherently immune to damage from unanticipated overtravel that can be encountered in applications where materials or structures can yield or fail.
8. Absolute output: Unlike a lot of other transducers that are incremental output devices, LVDTs are absolute output devices, that is, the displacement information from an LVDT is not lost if the system loses power. When the measuring system is restarted, the LVDTs output value will be the same as it was before the power failure occurred.

All these advantages, in addition to their reasonable cost, have made the LVDT an attractive displacement measurement technique. However, LVDTs for use in medical applications have the following disadvantages: (1) They require a constant amplitude excitation of high frequency. (2) They cannot be used in the vicinity of equipment that creates strong magnetic fields.

LVDT INSTRUMENTATION

Instrumentation normally used with an ac LVDT should perform the following functions (3,4).

Excitation

An LVDT needs an ac input of constant amplitude at a frequency that is not readily available. Hence, an oscillator of the appropriate frequency has to be connected to an amplifier with amplitude regulation on its output.

Amplification

As in the case of most transducers, the low level outputs of LVDTs require amplification. One procedure for amplification employs two steps: (1) use of an ac carrier–amplifier before demodulation; and (2) a dc amplifier after the demodulator (3,4).

Demodulation

As discussed earlier, the output of an LVDT remains proportional to the displacement while it undergoes a

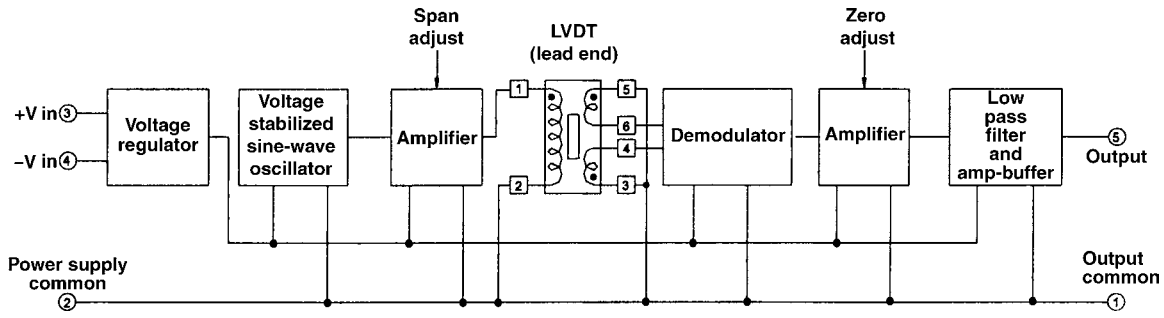


Figure 3. Block diagram for an LVDT employing a series 1000 oscillator–demodulator supplied by Trans-Tek, Inc. (Courtesy of Trans-Tek Inc.)

phase shift of 180° when the core goes through the null. When this LVDT output is connected to a voltmeter, the meter will register the same reading for equal amounts of core displacement on either side of the null position. This lack of directional sensitivity has to be overcome if one has to tell to which side of the null the core is displaced. Two techniques can be used to confer directional sensitivity on an LVDT output. In one technique, the core is offset, and the operation is centered on a position other than the null point. In this case, the output signal either increases or decreases. The other procedure uses a phase-sensitive demodulator (also called a detector). Several such devices are available and are discussed in detail elsewhere (8). The simplest forms employ diode rectification while the complex forms involve synchronous demodulation. Figure 3 shows the block diagram for an LVDT employing a series 1000 oscillator–demodulator supplied by Trans-Tek, Inc. (9).

The demodulator confers directional sensitivity on its input (output of the LVDT), which is either in phase with, or 180° out of phase with, the carrier signal (10). The demodulator output e_0 is usually sent to a low pass filter that will pass only the frequencies present in x and reject all higher frequencies created by the modulation procedure. Obviously, demodulation is not required if the LVDT transducer is to be used only on one side of the null position.

Recent developments allow all LVDT support circuitry to be accomplished using an inexpensive flexible field programmable analog array (FPAA). The FPAA consists of "configurable analog blocks" consisting of switched-capacitor op-amp cells surrounded by a programmable interconnect and I/O structure (11).

dc Power

Stable dc voltage sources are required for operation of the electronics associated with LVDTs. The dc LVDTs available at the present time employ a microcircuit module including all the electronics needed to provide ac excitation to the primary of the LVDT and to demodulate and amplify the analog LVDT signal. The module is mounted in tandem with the LVDT and only increases the effective LVDT length slightly.

Figure 4 shows the block diagram for a dc LVDT (9). The oscillator produces a constant amplitude sine wave excitation for the primary of the LVDT. A phase sensitive demodulator and an RC filter network process the secondary coil output. Some dc LVDT modules are furnished with a reverse polarity protector for the dc power input. dc LVDTs are becoming increasingly popular due to their advantages in the areas of calibration and signal conditioning.

SELECTION CRITERIA

Several criteria have to be considered in selecting a particular LVDT for a certain application (12). The manufacturer supplies several of these parameters as specification criteria.

Total Stroke

Stroke-length specification in the selection of an LVDT for a particular application is governed by the displacement to be measured. LVDTs can be custom-made for either short- (up to 0.01 m) or long-stroke (up to 1.5 m) operation; however, cost of fabrication increases greatly with increase in length, and lengths over 0.03 m may not be cost effective.

Linearity and the Nominal Linear Range

The output of an LVDT is a nearly linear function of core position for a rather wide range on either side of the balance (null) position (Fig. 2). A nominal linear range is defined for an LVDT as the core displacement on either side of the balance position for which the LVDT output as a function of displacement remains a straight line. Outside this range, the output starts to deviate gradually from the ideal straight line in the form of a smooth curve. Linearity of an LVDT is defined as "the maximum deviation from a best-fit straight line (applied to a plot of LVDT output voltage vs. core displacement) within the nominal linear

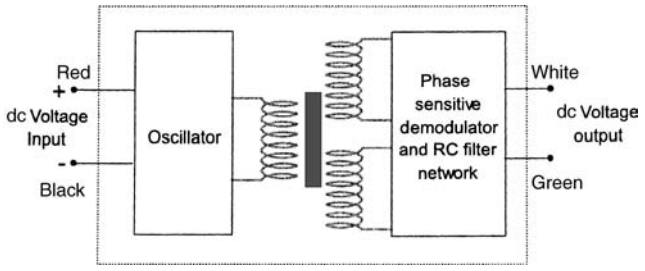


Figure 4. Block diagram of a dc LVDT. (Courtesy of Trans-Tek Inc.)

range" (12). Linearity is usually expressed as a percentage of the full scale. A typical LVDT has a linearity of about $\pm 0.25\%$.

Sensitivity

The sensitivity of an LVDT is usually expressed as the output in millivolts (or V) per 0.001 m core displacement per volt input. Normally, both the input voltage and the frequency are specified as well, because voltage sensitivity may vary with frequency over a limited frequency range. A typical miniature LVDT transducer has a sensitivity of ~8–200 mV out/0.001 m/V input.

Resolution

Resolution of an LVDT is the smallest core movement that can produce an observable change in the voltage output (12). With careful circuit design, displacements smaller than 100 nm can be detected.

Armature Mass

The mass of the armature (core) of the LVDT should be small so as not to unduly load the measured event. A reduction in the length of the LVDT results in a reduction in either the linearity or the maximum linear range, whereas sensitivity increases.

Excitation Frequency and Voltage

The sensitivity of the LVDT depends on both the excitation voltage and frequency. Normally, a sinusoidal voltage of 3–15 V rms amplitude and a frequency of 60 Hz–20 kHz is used for the excitation of LVDTs. The sensitivity of an LVDT increases with the excitation frequency, particularly at the lower part of the operating frequency range (12). Normally, an excitation frequency range of 1–5 kHz produces optimal LVDT operation.

Operating Environment

LVDTs have the advantage of being available in hostile-environment-proof format. Transducers designed to withstand both high and cryogenic temperatures and high pressures are available. Immersion-type LVDTs resistant to corrosive liquids are also available. Normally, specification criteria for an LVDT include information on the temperature range of operation and the temperature coefficient.

Residual Voltage Output

The residual voltage output is the LVDT output when the core is in the null position. This should ideally be zero; however, the null voltages and the harmonics of the excitation source do not cancel, resulting in a nonzero residual output (12). In practice, the residual voltage is about 1% of that obtained with maximum displacement.

Repeatability

Repeatability, the ability of the LVDTs to give the same output if the core is displaced and returned to the original position is an important consideration. LVDTs with repeatability better than 100 nm are available for some critical applications.

MEDICAL APPLICATIONS

LVDTs are used in medical applications and research to measure physiological variables that are either available in the form of a linear displacement or can be converted into such movement. LVDTs for medical applications can be readily fabricated in very small sizes with low mass cores. This will ensure that only a negligible force is imposed on the measured physiological event. Also, due to the low alternating currents in the windings, negligible magnetic load is imposed. When not in use, the core remains in the null position, and no force is imposed on the measured event. Even when the core is displaced from null, the load imposed on the event is small. These advantages, coupled with the general advantages of LVDTs discussed in the previous paragraphs, make these transducers very attractive for physiological measurements.

One early application of LVDTs was in the fabrication of invasive blood pressure measurement transducers (13). These transducers consisted of three essential parts: (1) a dome with pressure fittings, (2) a stainless steel diaphragm and core assembly, and (3) the LVDT coils. Pressure transmitted via the catheter exerts a force on the diaphragm. This causes a movement of the diaphragm, which in turn manifests itself in a movement of the core attached to it. Movement of the core of the LVDT creates a proportional output that can then be recorded after suitable electronic circuit processing. Catheter tip and implantable transducers employed the same principle (13). However, these rugged LVDT blood pressure transducers have been supplanted by cost-effective microelectromechanical system (MEMS) type transducers (14).

Another application for LVDT transducers is in indentation tests on tissue to determine mechanical properties. The authors have developed an LVDT indenter for the characterization of the mechanical properties of skin and the underlying soft tissue (Fig. 5). The indenter uses a loaded hemispherical tip coupled with a load cell-LVDT system for simultaneously measuring both the force and displacement during indentation tests. This information in turn was used to evaluate soft tissue properties. Walsh and Schettini (15) used a similar indenter to measure the *in vivo* viscoelastic response properties of brain tissue. Oculotonometers operating on the same principle and designed to indent the corneoscleral shell use LVDTs to measure deflections in the micrometer range (16). In another similar application, Gunner et al. (17) used an LVDT transducer-mounted extensometer to measure the *in vivo* recoil characteristics of human skin. The device consisted of two flat rectangular tabs, one fixed and the other capable of rectilinear sliding motion, attached to the test skin surface with double-sided adhesive tape. This combination was attached to an LVDT displacement transducer. Behavior of the skin resulting from the movement of the tabs was converted by the LVDT into electronic signals that were then analyzed to characterize the skin.

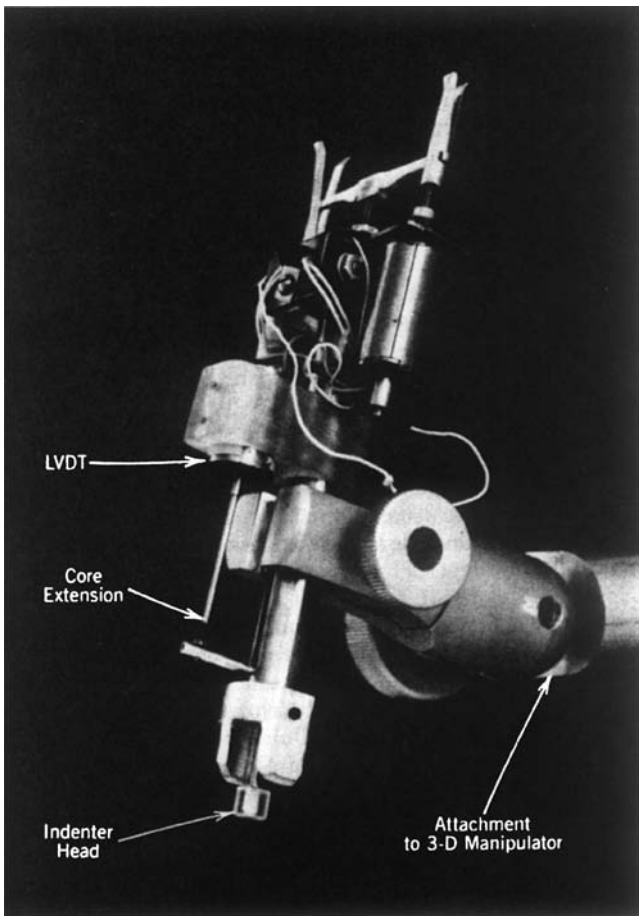


Figure 5. An LVDT skin and tissue indenter. (Courtesy of N. P. Reddy, J. Kagan and G. V. B. Cochran.)

Christiansen et al. (18) used a similar device for the viscoelastic characterization of skin. These examples are just a few of the myriad potential applications for LVDT transducers in soft tissue characterization.

Several radiological and neurological devices have incorporated LVDTs. For example, Laser Diagnostic Technologies of San Diego, CA, incorporated a position-sensing DC-DC LVDT into a scanning laser tomography instrument designed for retinal topography (9). The stable and repeatable DCDT output is part of a continuous feedback loop in the scanner's on-board logic control system. Radionics, Inc., employed an LVDT in a sophisticated modular probe drive used to support the precise implantation of deep brain stimulating electrodes (9). The device uses a push–pull cable drive mechanism to move the carrier that guides the probe to the desired location in the brain. The LVDT is mounted at the top of the mechanism and is used to accurately monitor probe position (Fig. 6). The LVDT used in this application was sealed to resist moisture and was modified to withstand the rigors of steam sterilization.

LVDTs have been used in endocrinology and pharmacology to evaluate *in vitro* and *in vivo* contractile properties of vascular smooth muscle. Erdos et al. (19) designed an

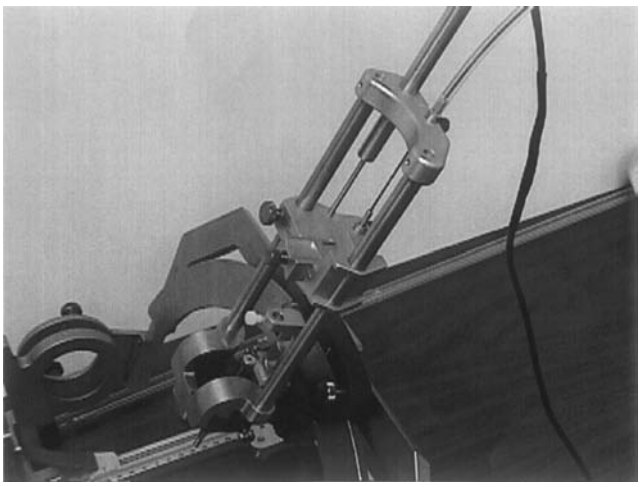


Figure 6. Mechanism of modular brain probe drive showing the LVDT used to help in precise electrode placement. [Courtesy of Radionics (a division of Tyco Healthcare).]

in vitro isotonic myograph employing an LVDT. The device resembled a beam-type balance. One arm of the device was connected to the contracting muscle specimen, and the other arm was counterbalanced by a suspended weight. Motion of the weight was translated via the movement of an LVDT core into electronic signals.

Gow (20) employed a novel LVDT electronic caliper for the continuous monitoring of arterial diameter changes during pulsation. This low mass device was found to possess a rapid response time with a natural resonant frequency greater than 180 Hz. Shykoff et al. (21) used LVDT measurements of changes in diameter of dorsal hand veins to establish diameter, pressure, and compliance relationships. LVDT-based devices have been used to evaluate the *in vivo* vascular effects of drugs with the dorsal hand vein technique (22). For example, Landau et al. (23) used an LVDT-based device to evaluate the magnesium-induced vasodilation in the dorsal hand vein. A similar technique was used by Streeten and Anderson (24) to measure venous contractile responses to locally infused norepinephrine.

LVDTs have been used in numerous orthopedic and dental devices. Chen et al. (25) used an LVDT bonded to the mandibular first molars to quantify mandibular deformation during mouth opening. Other possible medical applications are the mapping of facial contours before and after maxillofacial surgery and the profiling of spinal deformation in abnormalities like scoliosis. Buhler et al. (26) and Flamme et al. (27) used LVDTs to quantify micromotion in orthopedic implants. Recently, Dong et al.(28) incorporated an LVDT into a device for quantitative assessment of tension in wires of fine-wire external orthopedic fixators.

An LVDT was used to calibrate finger movements and to correlate these movements with surface electromyograms from the flexor digitorum superficialis muscles in the forearm (29,30). This work was designed to develop techniques for control of anthropomorphic teleoperator fingers

using surface electromyographic signals obtained from the forearm.

Wang et al. (31) used six spring-loaded LVDTs in an experimental technique to measure three-dimensional (3D), six-degrees-of-freedom motion of human joints. Rotary LVDTs are useful for measuring joint angles. For measuring 3D rotations, rotary LVDTs are incorporated into six-degree-of-freedom motion linkages.

As illustrated in the applications discussed above, LVDTs are highly suited for biomedical device and research applications requiring accurate displacement measurements with high resolution, input–output isolation, and cross-axis rejection. Although LVDTs are being replaced by miniaturized, cost-effective transducers utilizing advanced fabrication technologies in many applications, their advantages still render them excellent candidates for biomedical applications.

ACKNOWLEDGMENT

The authors would like to acknowledge the help provided by Rema Menon in the preparation of this manuscript.

BIBLIOGRAPHY

1. Schaevitz H. The linear variable differential transformer. Proc Soc Stress Anal 1947;4:79–88.
2. Chiriac H, Hristoforou E, Neagu M, Pieptanariu M. Linear variable differential transformer sensor using glass-covered amorphous wires as active core. J Magn Magn Mater 2000; 215:759–761.
3. Schaevitz Engineering. Technical bulletins 1002D and 7007. Pennsauken (NJ) 1986.
4. Schaevitz Sensors. Shaevitz Sensor Solutions. Catalog No. SCH-2001 Hampton (VA) 2000.
5. [Anonymous]. No date. An LVDT Primer. [Online]. Macro Sensors. Available at www.macrosensors.com. 2005 March 8.
6. Weinstein E. LVDTs on the factory floor. Instrum Control Syst 1982;55:59–61.
7. Sydenham PH. Microdisplacement transducers. J Phys E 1972;6:721–733.
8. Szczyrbak J, Schmidt EDD. LVDT signal conditioning techniques. Meas Control 1997;183:103–111.
9. [Anonymous]. No date. LVDT application. [Online]. Trans-Tek Inc. Available at www.transtekinc.com. 2005 March 8.
10. Doebelin EO. Measurement Systems: Application and Design. 4th ed. New York: McGraw-Hill; 1990.
11. Severn J, October. 2001. New analog interface for LVDTs. [Online]. Industrial Technology. Available at www.industrialtechnology.co.uk. 2005 March 8.
12. Anonymous, Finding the right LVDT. Instrum Control Syst 1977;50:61–62.
13. [Anonymous]. No date. LVDT Applications. [Online]. Macro Sensors. Available at www.macrosensors.com. 2005 March 8.
14. Seeley RS. 1996. The future of medical microelectromechanical systems. [Online]. Available at www.devicelink.com/mem/archive/96/01/003.html. 2005 March 8.
15. Walsh EK, Schettini A. A pressure-displacement transducer for measuring brain tissue properties *in vivo*. J Appl Physiol 1975;38:187–189.
16. Stepanik J. The Mackay-Marg Tonometer. Acta Ophthal 1970;48:1140.
17. Gunner CW, Hutton WC, Burlin TE. An apparatus for measuring the recoil characteristics of human skin *in vivo*. Med Biol Eng Comput 1979;17:142–144.
18. Christiansen MS, Hargens III CW, Nacht S, Gans EH. Viscoelastic properties of intact human skin: Instrumentation, hydration effects, and the contribution of the stratum corneum. J Invest Dermatol 1977;69:282–286.
19. Erdos EG, Jackman V, Barnes WC. Instrument for recording isotonic contractions of smooth muscles. J Appl Physiol 1962;17: 307–308.
20. Gow BS. An electrical caliper for measurement of pulsatile arterial diameter changes *in vivo*. J Appl Physiol 1966;21: 1122–1126.
21. Shykoff BE, Hawari FI, Izzo JL. Diameter, pressure and compliance relationships in dorsal hand veins. Vasc Med 2001;6(2):97–102.
22. Pang YC. Autonomic control of venous system in health and disease: effect of drugs. Pharmacol Therapeut 2001;90:179–230.
23. Landau R, Scott JA, Smiley RM. Magnesium-induced vasodilation in the dorsal vein. BJOG (Bri J Obst Gyn) 2004;111: 446–451.
24. Streeten DHP, Anderson GH. Mechanisms of orthostatic hypotension and tachycardia in patients with pheochromocytoma. AJH 1996;9:760–769.
25. Chen DC, Lai YL, Chi LY, Lee SY. Contributing factors of mandibular deformation during mouth opening. J Dent 2000;28(8):583–588.
26. Buhler DW, Oxland TR, Nolte LP. Design and evaluation of a device for measuring three-dimensional motions of press-fit femoral stem prosthesis. Med Eng Phys 1999;19:187–199.
27. Flamme CH, Kohn D, Kirsch L, Hurschler C. Primary stability of different implants used in conjunction with high tibial osteomy. Arch Orth Trauma Surg 1999;119:450–455.
28. Dong Y, Saleh M, Yang L. Quantitative assessment of tension in wires of fine-wire external fixators. Med Eng Phys 2005;27:63–66.
29. Gupta V, Reddy NP. Surface electromyogram for the control of anthropometric teleoperator fingers. Weghorst SJ, Soeburg HB, Morgan KS, editors. Medicine Meets Virtual Reality: Healthcare in the Information Age. Amsterdam: IOP Press; 1996.
30. Devavaram A, Reddy NP. Intelligent systems for control of telemanipulators using surface EMG signals, submitted for publication.
31. Wang M, Bryant JT, Dumas GA. A new *in vitro* measurement technique for small three-dimensional joint motion and its application to the sacroiliac joint. Med Eng Phys 1996;18(6): 495–501.

Further Reading

Herceg ED. Handbook of Measurement and Control, Pennsauken (NJ): Schaevitz Engineering; 1972.

Anonymous. LVDTs remain "State-of-the-art. Meas Inspect Technol 1982;4(2):13–16.

Anonymous. Displacement transducers (linear variable differential transformer products review). Control Instrum 1984;16(8): 23–25.

Geddes LA, Baker LE. Principles of Applied Biomedical Instrumentation. 3rd ed. New York: Wiley-Interscience; 1989.

Webster JG. Medical Instrumentation: Application and Design. 3rd ed. New York: John Wiley & Sons; 1998.

See also BLOOD PRESSURE MEASUREMENT; INTEGRATED CIRCUIT TEMPERATURE SENSOR.

LITERATURE, MEDICAL PHYSICS. See MEDICAL PHYSICS LITERATURE.

LITHOTRIPSY

ALON Z. WEIZER
GLENN M. PREMINGER
Duke University Medical Center
Durham, North Carolina

INTRODUCTION

The clinical introduction of shock wave lithotripsy (SWL) by Chaussy in 1980 has revolutionized the way in which patients with renal and ureteral calculi are treated. Shock wave lithotripsy is a noninvasive method of fragmenting stones located inside the urinary tract. Since its initial introduction, SWL technology has advanced rapidly in terms of the means for shock wave generation, shock wave focusing, patient coupling, and stone localization. Despite rapid technological advances, most current commercial lithotripters are fundamentally the same; they produce a similar pressure waveform at the focus, which can be characterized by a leading shock front with a compressive wave followed by a trailing tensile wave (Fig. 1). The acoustic fields produced by different lithotripters differ from each other in terms of the peak amplitudes of the pressure waveform, pulse duration, beam size, total acoustic energy, and therefore, their overall performance.

Clinical experience has guided the technical development of second and third generation lithotripters with the aim of providing user convenience and multifunctionality of the device, rather than on further understanding of how SWL fragments calculi or injures surrounding renal tissue. Furthermore, the evolution of lithotripter design thus far has overwhelmingly relied upon the importance of the compressive wave component of the shock wave (positive portion of the sound wave), with almost total neglect of the contribution of the tensile component of the waveform. Consequently, current lithotripters have suffered from inferior fragmentation rates compared to the original HM3 lithotripter.

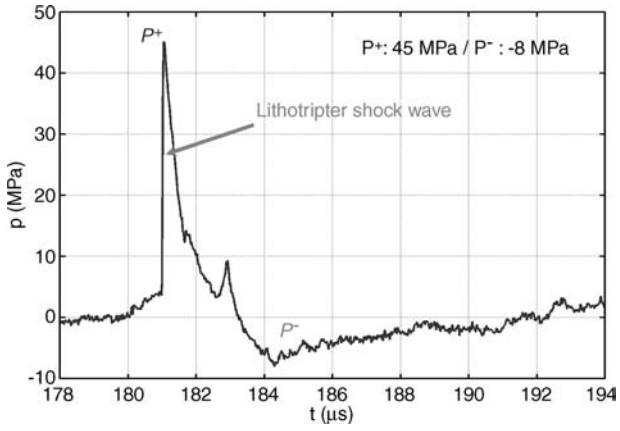


Figure 1. Pressure–time relationship of typical shock waves. Typical pattern of a lithotripter-generated shock wave. The shock wave is characterized by a leading positive or compressive component (P+) followed by a negative or tensile component (P−).

In contrast, significant progress in SWL basic science research has been made in the past 5 years to improve our understanding of the primary mechanisms for stone comminution (fragmentation) and tissue injury. It is now recognized that the disintegration of renal calculi in a lithotripter field is the consequence of dynamic and synergistic interaction of two fundamental mechanisms: stress wave-induced dynamic fracture in the form of nucleation, growth, and coalescence of preexisting microcracks inside the stone (1) and cavitation erosion caused by the violent collapse of bubbles near the stone surface (2,3). Similarly, two different mechanisms have been proposed for SWL-induced tissue injury: shear stress due to shock front distortion (4) and cavitation induced inside blood vessels, especially the expansion of intraluminal bubbles (5).

To understand how SWL fragments stones and causes tissue injury, the basic components of current lithotripters, the mechanisms behind stone fragmentation and kidney injury, and clinical results of the original electrohydraulic lithotripter in fragmenting kidney stones in patients will be described. In addition, the future directions of SWL will be reviewed based on current research that is investigating ways to make lithotripters more efficient and safer.

HISTORY AND EVOLUTION OF SWL

Physicists at Dornier Systems, Ltd. and Friedrich Shafen, Germany began experimenting with shock waves and their travel through water and tissue in 1963. Throughout the 1970s, numerous experimental lithotripters were developed that used new methods of shock wave generation and focusing as well as different techniques of stone localization. In addition, experimental studies were being performed *in vitro* and *in vivo* (in animal models) examining the effects of shock waves on various organs and tissues.

In 1980, Chaussy and associates successfully treated the first human and reported their first series of 72 patients in 1982 (6). Subsequently, $>$ 1800 articles have been published in the peer-reviewed literature, detailing the use of SWL for the management of renal and ureteral calculi. Moreover, numerous second and third generation devices have been introduced and are currently being used throughout the world. To understand how SWL results in stone fragmentation, the fundamentals of this technology are reviewed.

SWL PRINCIPLES

Despite the tremendous number of lithotripters currently available for fragmentation of renal and ureteral stones, all of these devices rely on the same laws of acoustic physics. Shock waves (i.e., high pressure sound waves) consisting of a sharp peak in positive pressure followed by a trailing negative wave are generated extracorporeally and passed through the body to fragment stones. Sounds waves readily propagate from a water bath or water medium into the human body, due to similar acoustic impedances.

As a consequence, all lithotripters share four main features: an energy source to generate the shock wave, a

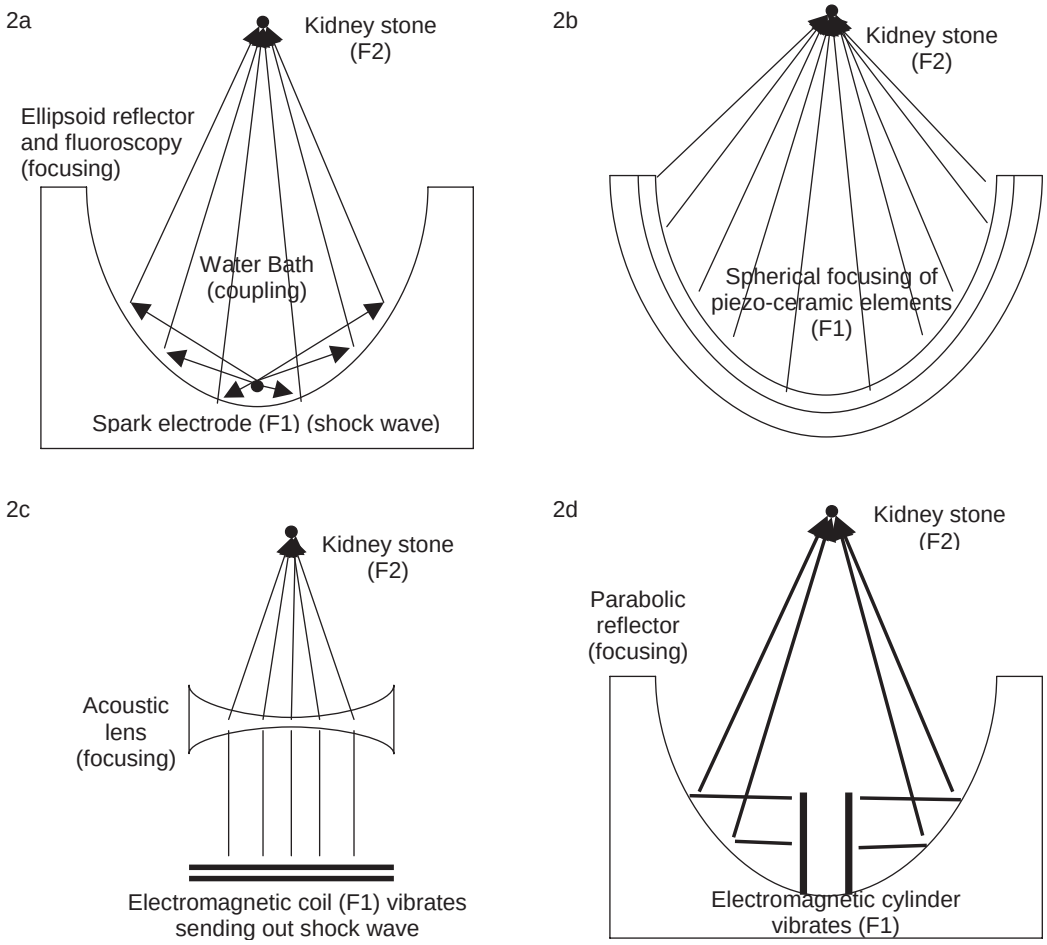


Figure 2. Schematic of different shock wave lithotripters. (2a) Electrohydraulic lithotripsy. Spark electrode generates shock wave which is focused at F2 by an ellipsoid reflector. In the original electrohydraulic lithotripter, a water bath served as a coupling medium to allow passage of the shock wave from the source (F1) to the patient and stone (F2). Fluoroscopy or ultrasound can be used to focus to identify and focus the shock waves on the stone. 2b) Piezoelectric lithotripsy. Many ceramic elements are arranged in a spherical pattern. When energy passes through them, they vibrate and send shock waves through a coupling medium to the stone. 2c and d) Electromagnetic lithotripsy. An electromagnetic coil (2c) or cylinder (2d) are stimulated to vibrate by passage of electric current. Their shock waves are focused on the stone (F2) by an acoustic lens (2c) or a parabolic reflector.

device to focus the shock wave at a focal point, a coupling medium, and a stone localization system. (Fig. 2a) The original electrohydraulic lithotripter utilized a spark plug energy generator with an elliptical reflector for focusing the shock waves. A water bath transmitted the shock waves to the patient with stone localization provided by biplanar fluoroscopy. Modification of the four basic components of this first generation lithotripter has provided the development of second and third generation devices that are currently available.

Shock Wave Generation and Focusing

While all lithotripters share the four aforementioned features, it is the mode of shock wave generation that determines the actual physical characteristics of that particular device. The types of energy sources differ on how efficient they are in fragmenting stones and how many treatments are required to adequately treat the stone. Figure 2 diagrammatically summarizes the three different shock wave generation sources used in commercially available lithotripters.

Electrohydraulic Generator. In the original Dornier HM3 lithotripter, the electrohydraulic generator (Fig. 2a) was located at the base of a water bath and produced shock waves by electric spark discharges of 15,000–25,000 V of 1 μs duration. This high voltage spark discharge produced the rapid evaporation of water that created a shock wave by expanding the surrounding fluid. The generator was located in an ellipsoidal reflector that concentrated the reflected shock waves at a second focal point, F2, with F1 being the point of origin of the primary shock waves. While the HM3 continues to be the gold standard lithotripter for stone fragmentation, the short half-life of its electrode results in variable pressure between shocks the longer it is used. In addition, minimal displacement of the electrode, as it deteriorates at F1, results in displacement of the F2 resulting in inaccurate focusing of the shock wave on the

stone. The need for frequent replacement of the electrode increases the cost of electrohydraulic lithotripsy.

Piezoelectric Generator. Piezoelectric shock waves are generated (Fig. 2b) by the sudden expansion of ceramic elements excited by a high frequency, high voltage pulse. Thousands of these elements are placed along the inner surface of a hemisphere at the base of a pool of water. While each of these elements moves only slightly in response to a pulse of electrical energy, the summation of the simultaneous expansion of multiple elements results in a high energy shock wave directed to the focal area, at the center of the sphere. The shock wave is propagated through either a small water basin or a water-filled bag to the focal point, F2. The spherical focusing mechanism of the piezoelectric lithotripters provides a wide area of shock wave entry at the skin surface, which causes minimal patient discomfort, but a very small focal area with the smallest amount of energy at F2 compared to other energy sources (7). The small focal area necessitates greater precision to focus and fragment the stone inside the kidney.

Electromagnetic Generator. In electromagnetic devices (Fig. 2c and d), shock waves are generated when an electrical impulse moves a thin, spherical metallic membrane, which is housed within a cylindrical shock tube. The resulting shock wave, produced in the water-filled shock tube, passes through an acoustic lens and is thereby directed to the focal point, F1 (Fig. 2c). The shock wave is coupled to the body surface with a moveable water cushion and coupling gel (8). Alternatively, when energy is passed through a cylindrical coil, the resulting magnetic field pushes away the surrounding cylindrical membrane producing a shock wave that can be focused by a parabolic reflector (Fig. 2d). While these devices produce reliable shock waves with consistent pressures and focus on F2, they also produce small focal regions that may result in reduced stone fragmentation and higher tissue parenchymal injury.

Shock Wave Focusing. Shock waves must be focused in order to concentrate their energy on a calculus. The type of shock wave generation dictates the method of focusing used. Machines that utilize point sources, such as spark-gap electrodes (electrohydraulic lithotripters), generate shock waves that travels in an expanding circular pattern. All of these machines use ellipsoid reflectors for focusing shock waves at the second focal point, F2.

Since a single piezoelement produces a very small amount of energy, larger transducers with multiple ceramic elements are required for piezoelectric lithotripters. The array of elements is positioned in a spherical dish that allows focusing in a very small focal region, F1. Finally, the vibrating metal membranes of the electromechanical lithotripters produce an acoustical plane wave that uses an acoustic lens for focusing the shock wave at F1.

Coupling Medium

The original Dornier HM3 machine utilized a 1000 L water bath to transmit shock waves into the patient. This method of patient coupling required unique positioning of the patient, since the anesthetized subject had to be lowered into the tub and the calculus accurately positioned at the second focal point. Second generation lithotripters were designed to alleviate the physiologic, functional, and economic problems of a large water bath. Current models utilize an enclosed water cushion, or a totally contained shock tube, to allow simplified positioning and "dry" lithotripsy (9).

Stone Localization

Stone localization during lithotripsy is accomplished with either fluoroscopy or ultrasonography. Fluoroscopy provides the urologist with a familiar modality and has the added benefits of effective ureteral stone localization. However, fluoroscopy requires more space, carries the inherent risk of ionizing radiation to both the patient and medical staff and is not useful in localizing radiolucent calculi (certain noncalcium-containing stones are not seen on fluoroscopy or conventional radiographs).

Ultrasonography utilizes sound waves to locate stones. These sound waves are generated at a source. When they encounter different tissue densities, part of the sound wave is reflected back and these reflected waves are used to generate a two dimensional image that can be used to focus the shock wave on a calculus. Sonography-based lithotripters offer the advantages of stone localization with continuous monitoring and effective identification of radiolucent stones, without radiation exposure (7). Additionally, ultrasound has been documented to be effective in localizing stone fragments as small as 2–3 mm, and is as good or better than routine radiographs to assess patients for residual stone fragments following lithotripsy (8). The major disadvantages of ultrasound stone localization include the basic mastery of ultrasonic techniques by the urologist and the inherent difficulty in localizing ureteral stones. While there are several systems that utilize both ultrasound and fluoroscopy to aid in stone localization, many commercially available lithotripters now use a modular design in which the fluoroscopy unit is not attached to the lithotripter reducing storage space as well as allowing use of the fluoroscopy unit for other procedures.

MECHANISMS OF STONE FRAGMENTATION

Due to recent advances made in the basic science of shock wave lithotripsy, there is a better understanding of how shock waves result in stone fragmentation. Both the positive and negative portions of the shock wave (Fig. 1) are critical in stone fragmentation and also play a role in renal tissue injury. The four mechanisms described for stone comminution include compressive fracture, spallation, cavitation, and dynamic fatigue. As the calculus develops *in vivo*, it is formed by both crystallization of minerals as well as organic matrix material. This combination forms an inhomogeneous and imperfect material that has natural defects. When the shock wave encounters this inhomogeneous structure, the force generated in the plane of the shock wave places stress on these imperfections resulting in compression-induced tensile cracking. This mechanism is known as compressive fracture. Spallation, another mechanism of shock wave induced stone fragmentation, occurs when the shock wave encounters fluid behind the

stone and part of the wave is reflected back onto the stone placing tensile stress on the same imperfections.

Shock waves cause bubbles to form on the surface of the stone. These bubbles grow during the negative or tensile component of the shock wave. When the positive component of the next wave passes, these bubbles violently collapse releasing their energy against the stone surface as secondary shock waves and/or microjets. This phenomenon, known as cavitation, represents the third mechanism of stone fragmentation. Dynamic fatigue describes the sum of accumulated damage to the stone that coalesce to result in stone fragmentation and eventually destruction of the stone (10,11).

CLINICAL RESULTS OF SHOCK WAVE LITHOTRIPSY

Because many experts continue to consider the Dornier HM3 (the original electrohydraulic lithotripter, Dornier MedTech America, Inc., Kennesaw, GA) the gold standard in lithotripsy, the available clinical literature comparing the Dornier HM3 (first generation) lithotripter to other commercially available lithotripters is summarized in Table 1. This type of summary does not truly allow a comparison between different lithotripters currently in use. Yet, this comparison demonstrates that modifications to second and third generation lithotripters have traded better patient comfort for a lessening of stone free rates. This current clinical data has been one of the driving forces behind the modifications that are currently being investigated to improve SWL.

Summarized results of five large early series on the clinical efficacy of SWL using the unmodified Dornier HM3 lithotripter demonstrate stone free rates for renal pelvis (RP), upper calyx (UC), middle calyx (MC), and lower calyx (LC) stones were 76% (48–85), 69% (46–82), 68% (52–76), and 59% (42–73), respectively. Stone free rates in these series were better for smaller stones (0–10 mm) and RP stones with relatively poorer stone free rates for LC stones. In comparison, results of five large published series on the Siemens Lithostar, a lower power machine demonstrated stone free rates for RP, UC, MC, and LC stones of 69% (55–80), 67% (46–90), 63% (43–82), and 60% (46–73). A comparison of integrated stone free rates stratified by size in a regression model of the HM3 and Lithostar found significantly greater stone free rates across all stone sizes for the HM3 lithotripter (11). While most studies have evaluated the efficacy of SWL for adults, stone free rates with the HM3 are similar for children (21). The limited number of comparative studies of newer machines and the explosion in the number of commercially available lithotripters makes it difficult to assess their clinical efficacy.

While the HM3 has been shown to produce excellent stone free rates for renal calculi, there continues to be debate on the clinical efficacy of SWL for ureteral stones. The main problem is that stones in the ureter are more difficult to locate and therefore more difficult to target with the shock wave. However, several studies have demonstrated stone free rates close to 100% for the treatment of proximal ureteral stones with SWL (22). However, stone free rates appear to decline to 70% for mid-ureteral stones for many lithotripters (23). Treatment of distal ureteral stones with SWL typically involves a prone or a modified sitting position to allow shock wave targeting of the stone below the pelvic brim. Stone free rates of distal ureteral stones with the HM-3 lithotripter have been as high as 85–96% (24). Endoscopic methods (ureteroscopy) can also be employed for ureteral stones, especially those located in the distal ureter with equivalent or better stone free rates.

While many of the lithotripters sighted in Table 1 are no longer in clinical use or have been updated, these studies clearly demonstrated several key points. The HM3 continues to provide equivalent and likely superior stone free rates when compared to other lithotripters in studies. While most commercially available lithotripters provide acceptable stone free rates for stones located within the kidney, these stones often require more treatment sessions and adjunctive procedures to achieve the stone free rates of the HM3 device. Additionally, success rates with all lithotripters declines the further the stone progresses down the ureter and poses positioning challenges with alternative methods indicated to remove ureteral stones.

TISSUE INJURY WITH CLINICAL LITHOTRIPSY

Clinical experience treating patients with SWL has demonstrated that while SWL is generally safe, shock waves can cause acute and chronic renal injury. This concept has been confirmed by multiple animal studies and a few human clinical studies (11). Originally, shear stress to the tissue was believed to be the main cause of this renal injury. However, recent studies have sugggested that SWL induced renal injury is a vascular event induced by vasoconstriction or cavitation-induced injury to the microvasculature (25). Additionally, this initial tissue damage may promote further renal injury via the effects of free radical production (26). Acute damage to the kidney appears to be mainly a vascular insult.

While clinicians have long recognized the acute effects of SWL, most have believed that there was no long-term sequela to shock wave treatment. The $>$ 25 years of clinical experience with SWL serves as a testament to its safety and effectiveness. However, several chronic problems may develop as a consequence of SWL. Table 2 summarizes the acute and chronic effects of SWL. Perhaps the most serious long-term problem of SWL is the increased risk of hypertension. A review of 14 studies on the impact of SWL on blood pressure demonstrated that when stratified according to the number of shock waves administered, higher doses of shock waves seem to correlate with a greater likelihood of increased rates of new-onset hypertension or changes in diastolic blood pressure (11). The impact of hypertension on cardiovascular disease including the risk of myocardial infarction, stroke, and renovascular disease make this a serious long-term effect that needs further investigation.

Three mechanisms of SWL induced tissue injury have been reported: cavitation, vasoconstriction, and free radical induced tissue injury. Investigators have demonstrated *in vitro* that cavitation bubble expansion can rupture artificial blood vessels (5). Other investigators have shown in animal models that cavitation takes place in kidneys exposed to shock waves (27). While cavitation bubbles that form on the stone surface contribute to stone fragmentation,

Table 1. Literature Comparison of Lithotripters to Gold Standard Dornier HM3

Study Type	Reference	HM3 Compared to:	Stone Location	Number of Patients	Stone Free Rates (SFR), %	Auxiliary Procedures	Retreatment Rate, %	Comment
Prospective	12	Wolf Piezolith	Kidney	HM3: 334 Wolf: 378	HM3: 75Wolf: 72		HM3: 15.5 Wolf: 45	Wolf required more retreatment, more shocks, treatment rates decreased dramatically for ureteral stones with Wolf
	13	EDAP LT01	Kidney	HM3: 500 EDAP: 500	HM3:77.2-90.4 EDAP: 42.5-87.5		> with EDAP	More sessions, increased shocks required with EDAP
	14	MFL 5000	Kidney	198 total	HM3: 80MFL: 56			Increased subcapsular hematoma and longer treatment times with MFL
	15	Wolf Piezolith 2300	Ureter	70 total	HM3: 74Wolf: 76.6			Comparable 3 month SFR but used plain radiographs for comparison
	16	Siemens Lithostar	Kidney	HM3: 91Siemens:85	HM3: 91Siemens: 65		HM3: 4 Siemens: 13	SFR comparable at 3 months, Increased tissue injury with HM3 by urinary enzymes
Retrospective	17	EDAP LT01 Sonolith 2000	Kidney and ureter	HM3: 70EDAP: 113Sono: 104	HM3: 79EDAP: 82Sono: 79	HM3: 12EDAP: 13Sono: 9	HM3: 4EDAP: 42Sono: 26	SFR at 3 months comparable with HM3 and EDAP, lower for Sonolith, lower retreatment with HM3
	18	Siemens Lithostar, Dornier HM4, Wolf Piezolith 2300, Direx Tripter X-L, Breakstone	Kidney and ureter	Multicenter	comparable between 2nd generation			All were deemed inferior to HM3 in terms of stone free rates
	19	Medstone STS	Kidney	HM3: 5698Med: 8166	HM3: 70Med: 81.5	HM3: 3.1Med: 5.5	HM3: 4.4 Med: 5.2	Slightly better retreatment and need for auxiliary procedures with HM3
	20	Lithotron	Kidney	38 matched pairs	HM3: 79Lithotron: 58	> for Lithotron	> for Lithotron	HM3 superior to Lithotron using matched pair analysis
	11	Siemens Lithostar		Meta-analysis	HM3: 59-76 Siemens: 60-69			Using regression model, SFR better with HM3 across all stone sizes

Table 2. Acute and Chronic Injury with SWL

Acute	Chronic
Renal edema (swelling)	Hypertension(elevated blood pressure)
Hematuria (blood in urine)	Decreased renal function
Subcapsular hematoma	Accelerated stone formation (in animal models)
Decreased renal blood flow	Renal scar formation
Altered renal function:	
Impaired urine concentration	
Impaired control of electrolytes	

their formation in other locations (tissue, blood vessel lumen) is an unwanted end product that results in tissue injury.

Recent investigations have elucidated yet another potential mechanism of renal injury secondary to high energy shock waves. Evidence suggests that SWL exerts an acute change in renal hemodynamics (i.e., vasoconstriction) that occurs away from the volume targeted at F2, as measured by a transient reduction in both glomerular filtration rate (GFR) and renal plasma flow (RPF) (28). Prolonged vasoconstriction can result in tissue ischemia and permanent renal damage.

Vasoconstriction and cavitation both appear to injure the renal microvasculature. However, as the vasoconstriction induced by SWL abates, reperfusion of the injured tissue might also result in further tissue injury by the release of free radicals. These oxidants produced by the normal processes of cellular metabolism and cellular injury cannot be cleared and injure the cell membrane destroying cells. Free radical formation has been demonstrated in animal models (26).

It appears that the entire treated kidney is at risk of renal damage from SWL-induced direct vascular and distant vasoconstrictive injury, both resulting in free radical formation. Although previous studies have suggested that the hemodynamic effects are transient in nature in normally functioning kidneys, patients with baseline renal dysfunction may be at significant risk for permanent renal damage (28). Patients of concern may be pediatric patients, patients undergoing multiple SWL treatments to the same kidney, patients with solitary kidneys, vascular insufficiency, glomerulosclerosis, glomerulonephritis, or renal tubular insult from other causes.

SHOCK WAVE LITHOTRIPSY ADVANCES

Based on a better understanding of cavitation in stone fragmentation as well as the role of cavitation, vasoconstriction, and free radical formation in SWL-induced tissue injury, several groups are investigating ways in which SWL can be clinically more effective and safe. In general, these advancements involve changes to the shock wave itself, by modifying the lithotripter, alterations in treatment technique, improvements in stone fragmentation / passage or the reduction in tissue injury through medical adjuncts, and improved patient selection.

Changes to the Lithotripter

There are two major mechanical modifications that can improve stone comminution, based on our current understanding of acoustic physics. One is to enhance the compressive component of the shock wave. The original HM3 relies on a high energy output and thus the compressive component to achieve stone fragmentation. The downside of this effect, from clinical experience, is patient discomfort and potential renal tissue injury. Alternatively, one can improve stone comminution by altering the tensile component of the shock wave and thus better control cavitation. Below, several ways are describe in which investigators are modifying the shock wave to improve comminution, with decreased renal tissue injury.

Several investigators have modified lithotripters to alter the negative portion of the shock wave that is responsible for cavitational-induced stone fragmentation and tissue injury. In one study, a reflector insert is placed over the original reflector of an electrohydraulic lithotripter to create a second shock wave that arrives behind the original shockwave, thus partially canceling out the negative component of the shock wave. These investigators found that this modification reduced phantom blood vessel rupture, while preserving stone fragmentation *in vitro* (29). Similarly, an acoustic diode (AD) placed over the original reflector, has the same impact as this modified reflector (30).

However, because reducing the tensile component of the shock wave weakens that collapse of bubbles at the stone surface, two groups have designed piezoelectric inserts into an electrohydraulic lithotripter that send small, focused shock waves at the time of bubble collapse near the stone surface thus, intensifying the collapse of the cavitation bubble without injuring surrounding tissue (29).

Another way in which investigators have modified the shock wave is by delivering shock waves from two lithotripters to the same focal point, F2. Dual pulse lithotripsy has been evaluated by several investigators both *in vitro*, animal models and in clinical trials. Several investigators have demonstrated in an *in vitro* model that the cavitation effect became more localized and intense with the use of two reflectors. Also, the volume and rate of stone disintegration increased with the use of the two reflectors, with production of fine (< 2 mm) fragments (31). In both animal models and clinical studies, dual pulse lithotripsy has been shown to improve stone fragmentation with reduced tissue injury.

Modifications to Treatment Strategy

The original Dornier HM3 lithotripter rate was synchronized with the patient's electrocardiogram so that the shock rate did not exceed the physiologically normal heart rate. Experience with newer lithotripters has revealed that ungating the shock wave delivery rate results in few cardiac abnormalities. As a result, there has been a trend to deliver more shock waves in a shorter period of time. However, increasing doses of shock wave energy at a higher rate may have the potential to increase acute and chronic renal damage.

As a result, several investigators have evaluated ways in which the treatment strategy can be modified to optimize SWL treatment. One proposed strategy is altering the rate of shock wave delivery. Several investigators have reported that 60 shocks $\cdot$ min^{-1} at higher intensities resulted in the

most efficient stone fragmentation than 120 shocks · min^{-1}. This has been confirmed *in vitro*, in animal models and also in randomized clinical trials. These studies speculate that at increased rates, more cavitation bubbles were formed in both the fluid and tissue surrounding the stone that did not dissipate between shocks. As a result, these bubbles scattered the energy of subsequent shocks resulting in decreased efficiency of stone fragmentation (32).

In order to acclimate patients to shock waves in clinical treatment, lower energy settings are typically used and gradually increased. A study investigating whether increasing voltage (and thus increasing treatment dose) impacted on stone fragmentation has been performed *in vitro* and in animal models. Stones fragmented into smaller pieces when they were exposed to increasing energy compared to decreasing energy. The authors speculate that the low voltage shock waves "primed" the stone for fragmentation at higher voltages (33). In addition, animals exposed to an increasing voltage had less tissue injury than those kidneys exposed to a decreasing or stable dose of energy. While this treatment strategy has not been tested clinically, it might be able to improve *in vivo* stone comminution while decreasing renal parenchymal injury (34).

In the same vein, several studies have reported that pretreating the opposite pole of a kidney with a low voltage dose of shock waves (12 kV), prior to treating a stone in the other pole of a kidney with a normal dosage of shock waves, reduced renal injury when as little as 100 low voltage shocks were delivered to the lower pole. It is believed that the low voltage shock waves causes vasoconstriction which protects the treated pole of the kidney from hemorrhagic injury (35).

Other SWL treatment modifications being tested include aligning the shock wave in front of the stone in order to augment cavitation activity at the stone surface or apply overpressure in order to force cavitation bubble collapse. While these techniques have only been investigated *in vitro*, these alterations in shock wave delivery, as well as the previous treatment strategies demonstrate how an improved understanding of the mechanisms of SWL can enhance stone comminution and potential reduce renal tissue injury (36,37).

Adjuncts to Improve SWL Safety and Efficacy

Antioxidants. A number of studies have investigated the role of antioxidants in protecting against free radical injury to renal parenchyma (38). Table 3 summarizes the results of various *in vitro* and *in vivo* studies on the use of antioxidants to protect against SWL-induced renal injury due to free radicals. While these studies are intriguing, further clinical trials will be needed to evaluate potential antioxidants for use in patients undergoing SWL.

Improving Stone Fragmentation. Another potential way to improve stone fragmentation is to alter the stone's susceptibility to shock wave energy. One group has demonstrated that after medically pretreating stone in an *in vitro* environment, one could achieve improved stone fragmentation. These data suggest that by altering the chemical environment of the fluid surrounding the stones it is possible to increase the fragility of renal calculi *in vitro* (49). Further studies are warranted to see if calculi can be clinically modified, prior to SWL therapy, in the hopes of enhanced stone fragmentation.

Table 3. Investigated Antioxidants Providing Protection against SWL-Induced Free Radical Injury

Reference	Study Type	Antioxidant
39	*In vitro*	nifedipine, verapamil, diltiazem
40	*In vitro*	Vitamin E, citrate
41	*In vitro*, animal	Selenium
42	Animal	Verapamil
43	Animal	Verapamil
26	Animal	Allopurinol
44	Animal	Astragulus membranaceus, verapamil
45	Human	Antioxidant vitamin
46,47	Human	Verapamil, nifedipine
48	Human	Mannitol

Improving Stone Expulsion. Several reports have demonstrated that calcium channel blockers, steroids, and alpha blockers all may improve spontaneous passage of ureteral stones. (ref) Compared to a control group, investigators found improved stone clearance and shorter time to stone free in patients treated with nifedipine or tamsulosin following SWL compared to the control group. Additionally, retreatment rates were lower (31%) for the medical treatment group compared to the control group (51%). While expulsive therapy appears to offer improved outcomes following SWL for ureteral stones, confirmation with a randomized controlled study is needed (50). Another intriguing report from the same group involves the use of Phyllanthus niruri (Uriston) to improve stone clearance of renal stones following SWL. This medication is derived from a plant used in Brazilian folk medicine that has been used to treat nephrolithiasis. Again, stone free rates were improved with the administration of Uriston following SWL compared to a control group with apparently the greatest effect seen in those patients treated with lower pole stones (51).

Improving Stone/Patient Selection for SWL. Another way to enhance the efficacy of SWL is improve patient selection. Advances in computed tomography (CT) have allowed better determination of internal stone architecture (52). As a consequence, a few studies have demonstrated that determining the Hounsefield units (i.e., density unit of material on CT) of renal stones on pretreatment, noncontrasted CT could predict stone free rates of patients treated with SWL (53). Current micro-CT and newer multidetector CT scanners have the potential to identify stone composition based on CT attenuation. Therefore, stone compositions that are SWL resistant, such as calcium oxalate monohydrate or cystine stones, can be identified and those patients can be treated with endoscopic modalities, therebye avoiding additional procedures in these patients (54). Clinical trials utilizing this concept will need to be performed.

Other factors, such as the distance of the stone from the skin, weight of the patient, and other imaging modalities, are being investigated to help determine who is likely to benefit the most from SWL and which patients should be treated initially with other modalities.

CONCLUSIONS

Shock wave lithotripsy has revolutionized the way in which urologists manage urinary calculi. Patients can now be treated with minimal discomfort using an outpatient procedure. While all lithotripters rely on the same fundamental principles of acoustic physics, second and third generation lithotripters appear to have traded patient comfort and operator convenience for reduced stone free rates, as compared to the original HM3 lithotripter. In addition, mounting evidence demonstrates that SWL has both acute and chronic impact on renal function and blood pressure as a result of renal scarring.

Basic science research has provided insight into how SWL results in stone comminution as well as renal tissue injury. While the compressive component of the shock wave causes stone comminution, it is apparent that the tensile component plays a critical role in creating passable stone fragments. These same forces cause tissue injury by damaging the microvasculature of the kidney. This knowledge has resulted in several novel modifications to improve both stone free rates as well as SWL safety. Mechanical modifications to lithotripsy have focused on controlling cavitation. Preventing bubble expansion in blood vessels while intensifying bubble collapse near the stone surface has been demonstrated to achieve improved stone comminution with decreased tissue injury *in vitro* and in animal models. Many of these designs could be adapted to conventional lithotripters. Modification of treatment techniques have also stemmed from our better understanding of SWL. Slowing treatment rates may limit the number of cavitation bubbles that can interfere with the following shock wave. Voltage stepping and alternative-site pretreatment with low dose shock waves, may cause global renal vasoconstriction that prevents cavitational injury to small vessels during treatment. In addition, our understanding that free radicals may be the end culprit in parenchymal damage has suggested that pretreatment with antioxidants may prevent SWL-induced renal injury. Finally, improved CT imaging may allow us to predict which stones and patients are best suited for SWL versus endoscopic stone removal. Further advances will continue to make SWL a major weapon in the war against stone disease for years to come.

BIBLIOGRAPHY

1. Lokhandwalla M, Sturtevant B. Fracture mechanics model of stone comminution in ESWL and implications for tissue damage. Phys Med Biol 2000;45:1923–1940.
2. Coleman AJ, Saunders JE, Crum LA, Dyson M. Acoustic cavitation generated by an extracorporeal shockwave lithotripter. Ultrasound Med Biol 1987;13:69–76.
3. Zhu S, Cocks FH, Preminger GM, Zhong P. The role of stress waves and cavitation in stone comminution in shock wave lithotripsy. Ultrasound Med Biol 2002;28:661–671.
4. Howard D, Sturtevant B. *In vitro* study of the mechanical effects of shock-wave lithotripsy. Ultrasound Med Biol 1997;23:1107–1122.
5. Zhong P, Zhou Y, Zhu S. Dynamics of bubble oscillation in constrained media and mechanisms of vessel rupture in SWL. Ultrasound Med Biol 2001;27:119–134.
6. Chaussy C, et al. First clinical experience with extracorporeally induced destruction of kidney stones by shock waves. J Urol 1982;127:417–420.
7. Preminger GM. Sonographic piezoelectric lithotripsy: More bang for your buck. J Endourol 1989;3:321–327.
8. Abernathy BB, et al. Evaluation of residual stone fragments following lithotripsy: Sonography versus KUB. In: Lingeman JE, Newman DM., editors. Shock Wave Lithotripsy II. New York: Plenum Press; 1989. pp 247–254.
9. Cartledge JJ, Cross WR, Lloyd SN, Joyce AD. The efficacy of a range of contact media as coupling agents in extracorporeal shockwave lithotripsy. BJU Int 2001;88:321–324.
10. Eisenmenger W. The mechanisms of stone fragmentation in ESWL. Ultrasound Med Biol 2001;27:683–693.
11. Lingeman JE, Lifshitz DA, Evan AP. Surgical Management of Urinary Lithiasis. In: Walsh PC, Retik AB, Vaughan ED, Jr., Wein AJ., editors. Campbell's Urology. Philadelphia: Saunders; 2002. p 3361–3451.
12. Rassweiler J, et al. Wolf Piezolith 2200 versus the modified Dornier HM3. Efficacy and range of indications. Eur Urol 1989;16:1–6.
13. Sofras F, et al. Extracorporeal shockwave lithotripsy or extracorporeal piezoelectric lithotripsy? Comparison of costs and results. Br J Urol 1991;68:15–17.
14. Chan SL, et al. A prospective trial comparing the efficacy and complications of the modified Dornier HM3 and MFL 5000 lithotriptors for solitary renal calculi. J Urol 1995;153:1794–1797.
15. Francesca F, et al. Ureteral lithiasis: In situ piezoelectric versus in situ spark gap lithotripsy. A randomized study. Arch Esp Urol 1995;48:760–763.
16. Graber SF, Danuser H, Hochreiter WW, Studer UE. A prospective randomized trial comparing 2 lithotriptors for stone disintegration and induced renal trauma. J Urol 2003;169: 54–57.
17. Tan EC, Tung KH, Foo KT. Comparative studies of extracorporeal shock wave lithotripsy by Dornier HM3, EDAP LT 01 and Sonolith 2000 devices. J Urol 1991;146:294–297.
18. Bierkens AF, et al. Efficacy of second generation lithotriptors: A multicenter comparative study of 2,206 extracorporeal shock wave lithotripsy treatments with the Siemens Lithostar, Dornier HM4, Wolf Piezolith 2300, Direx Tripter X-1 and Breakstone lithotriptors. J Urol 1992;148:1052–1056. Discussion 1056–1057.
19. Cass AS. Comparison of first generation (Dornier HM3) and second generation (Medstone STS) lithotriptors: Treatment results with 13,864 renal and ureteral calculi. J Urol 1995;153:588–592.
20. Portis AJ, et al. Matched pair analysis of shock wave lithotripsy effectiveness for comparison of lithotriptors. J Urol 2003;169:58–62.
21. Cass AS. Comparison of first-generation (Dornier HM3) and second-generation (Medstone STS) lithotripters: Treatment results with 145 renal and ureteral calculi in children. J Endourol 1996;10:493–499.
22. Robert M, A'Ch S, Lanfrey P, Guiter J, Navratil H. Piezoelectric shockwave lithotripsy of urinary calculi: Comparative study of stone depth in kidney and ureter treatments. J Endourol 1999;13:699–703.
23. Marguet CG, Springhart WP, Auge BK, Preminger GM. Advances in the surgical management of nephrolithiasis. Minerva Urol Nefrol 2004;56:33–48.
24. Rodrigues Netto Junior N, Lemos GC, Claro JF. *In situ* extracorporeal shock wave lithotripsy for ureteral calculi. J Urol 1990;144:253–254.
25. Evan AP, et al. Shock wave lithotripsy-induced renal injury. Am J Kidney Dis 1991;17:445–450.
26. Munver R, et al. *In vivo* assessment of free radical activity during shock wave lithotripsy using a microdialysis system: The renoprotective action of allopurinol. J Urol 2002;167:327–334.

27. Evan AP, et al. *In vivo* detection of cavitation in parenchyma of the pig kidney during shock wave lithotripsy. American Urological Association Annual Meeting. Orlando (FL): 2002. p 1500.
28. Willis LR, et al. Effects of SWL on glomerular filtration rate and renal plasma flow in uninephrectomized minipigs. J Endourol 1997;11:27–32.
29. Zhou Y, Cocks FH, Preminger GM, Zhong P. Innovations in shock wave lithotripsy technology: updates in experimental studies. J Urol 2004;172:1892–1898.
30. Zhu S, et al. Reduction of tissue injury in shock-wave lithotripsy by using an acoustic diode. Ultrasound Med Biol 2004; 30:675–682.
31. Sheir KZ, et al. Evaluation of synchronous twin pulse technique for shock wave lithotripsy: determination of optimal parameters for *in vitro* stone fragmentation. J Urol 2003; 170:2190–2194.
32. Paterson RF, et al. Stone fragmentation during shock wave lithotripsy is improved by slowing the shock wave rate: studies with a new animal model. J Urol 2002;168:2211–2215.
33. McAteer JA, et al. Voltage-Stepping During SWL Influences Stone Breakage Independent of Total Energy Delivered: *In vitro* studies with model stones. American Urological Association Annual Meeting. Chicago: 2003. p 1825.
34. Maloney M, et al. Treatment strategy improves the *in vivo* stone comminution efficiency and reduces renal tissue injury during shock wave lithotripsy. American Urological Association Annual Meeting. San Antonio (TX): 2005. p 1108.
35. Willis LR, et al. Same-pole application of low- and high-energy shock waves protects kidney from swl-induced tissue injury. American Urological Association Annual Meeting. San Francisco: 2004. p 1114.
36. Sapozhnikov OA, et al. Effect of overpressure and pulse repetition frequency on cavitation in shock wave lithotripsy. J Acoust Soc Am 2002;112:1183–1195.
37. Sokolov DL, et al. Prefocal alignment improves stone comminution in shockwave lithotripsy. J Endourol 2002;16:709–715.
38. Preminger GM. Review: *in vivo* effects of extracorporeal shock wave lithotripsy: animal studies. J Endourol 1993;7: 375–378.
39. Jan CR, Chen WC, Wu SN, Tseng CJ. Nifedipine, verapamil and diltiazem block shock-wave-induced rises in cytosolic calcium in MDCK cells. Chin J Physiol 1998;41:181–188.
40. Delvecchio F, et al. Citrate and Vitamin E Blunt the SWL Induced Free radical surge in an in-vitro MDCK cell culture model. American Urological Association Annual Meeting. San Francisco: 2004. p 1120.
41. Strohmaier WL, Lahme S, Bichler KH. Amelioration of high energy shock wave induced renal tubular injury by selenium-an in vivo study in rats. American Urological Association Annual Meeting. Anaheim (CA): 2004. p 1529.
42. Yaman O, et al. Protective effect of verapamil on renal tissue during shockwave application in rabbit model. J Endourol 1996;10:329–333.
43. Willis LR, et al. Effects of extracorporeal shock wave lithotripsy to one kidney on bilateral glomerular filtration rate and PAH clearance in minipigs. J Urol 1996;156:1502–1506.
44. Sheng BW, et al. Astragalus membranaceus reduces free radical-mediated injury to renal tubules in rabbits receiving high-energy shock waves. Chin Med J (Engl) 2005;118:43–49.
45. Kehinde EO, et al. The effects of antioxidants on renal damage occuring during treatment of renal calculi by lithotripsy. American Urological Association Annual Meeting. San Antonio (TX): 2005. p 1698.
46. Strohmaier WL, et al. Protective effect of verapamil on shock wave induced renal tubular dysfunction. J Urol 1993;150:27–29.
47. Strohmaier WL, et al. Limitation of shock-wave-induced renal tubular dysfunction by nifedipine. Eur Urol 1994;25:99–104.
48. Ogiste JS, et al. The role of mannitol in alleviating renal injury during extracorporeal shock wave lithotripsy. J Urol 2003;169: 875–877.
49. Heimbach D, et al. The use of chemical treatments for improved comminution of artificial stones. J Urol 2004;171: 1797–1801.
50. Micali S, et al. Efficacy of expulsive medical therapy using nifedipine or tamsulosin after shock wave lithotripsy of ureteral stones. American Urological Association Annual Meeting. San Antonio (TX): 2005. p 1680.
51. Antonio C, et al. May Phyllanthus niruri (Uriston) affect the efficacy of ESWL on renal stnoes? A prospective, randomised short term study. American Urological Association Annual Meeting. San Antonio (TX): 2005. p 1696.
52. Zarse CA, et al. Nondestructive analysis of urinary calculi using micro computed tomography. BMC Urol 2004;4:15.
53. Saw KC, et al. Calcium stone fragility is predicted by helical CT attenuation values. J Endourol 2000;14:471–474.
54. Williams JC, et al. Progress in the use of helical CT for imaging urinary calculi. J Endourol 2004;18:937–941.

See also MINIMALLY INVASIVE SURGERY; ULTRASONIC IMAGING.

LIVER TRANSPLANTATION

PAUL J. GAGLIO
Columbia University College
of Physicians and Surgeons
New York, New York

INTRODUCTION

From a conceptual perspective, liver transplantation involves the replacement of a diseased or injured liver with a new organ. Historically, liver transplantation has emerged from an experimental procedure deemed "heroic" therapy for patients not expected to survive, to the treatment of choice with anticipated excellent long-term outcomes for patients with end stage liver disease. This article will outline the history of and indications for liver transplantation, delineate short- and long-term complications associated with the procedure, and discuss the role of immunosuppressive therapy, intrinsic to the long-term success of the procedure.

HISTORY

Historically, the most significant and persistent impediment to liver transplantation has been the availability of suitable organs. Up until the early 1960s, "death" was defined as cessation of circulation, and thus, donation from deceased donors was thought to be both impractical and impossible, as organs harvested from pulseless, nonperfusing donors would not function when transplanted, due to massive cellular injury. The concept of "brain death" and ability to harvest organs from individuals defined as such first occurred at Massachusetts General Hospital in the early 1960s, when a liver was harvested from a patient whose heart was beating despite central nervous system failure. This seminal event led to the development of a new concept; death was defined when cessation of brain function occurred, rather than the cessation of circulation. Thus, brain dead donors with stable blood pressure and the absence of comorbid disease could serve as potential organ donors. Improvements in the ability to preserve and transport organs dramatically increased organ availability, necessitating a

centralized system to facilitate procurement and allocation of organs to individuals waiting for transplantation. This was initially provided by SEOPF (the Southeast Organ Procurement Foundation), founded in 1968, from which UNOS (the United Network for Organ Sharing) arose. At present, UNOS operates the OPTN (Organ Procurement and Transplantation Network), providing a centralized agency that facilitates recovery and transportation of organs for transplantation, and appropriately matches donors and recipient.

LIVER TRANSPLANTATION: INITIAL RESULTS

The first reported liver transplantation occurred in 1955, in the laboratory of Dr. Stuart Welch (1). In a dog model, an "auxiliary" liver was transplanted into the abdominal cavity, leaving the native liver *in situ*. Between 1956 and 1960, various investigators initiated experiments in different animal models whereby "orthotopic" liver transplantation was performed, achieved by removal of the native liver and implantation of a "new" liver in its place, requiring anastamoses of the donor and recipient hepatic vein and artery, bile duct, and portal vein (see Fig. 1). These initial attempts at liver transplantation refined the surgical procedure, however, graft dysfunction and death of the animals occurred quickly, due to ineffective immunosuppression and eventual rejection of the liver mediated by the animal's immune system (2).

The first human liver transplants were performed by Dr. Thomas Starzl in 1963, at the University of Colorado (3). These early attempts at transplantation highlighted the difficulties associated with extensive abdominal surgery in desperately ill patients, and were associated with poor outcomes, largely due to technical difficulties and the inability to effectively prevent rejection. Similar negative experiences at other centers led to a worldwide moratorium on liver transplantation. However, a major breakthrough in the ability to prevent rejection and prolong the survival of the transplanted liver occurred following the availability of Cyclosporine in 1972 (described below). With continued refinement of the surgical techniques required to perform liver transplantation, combined with the ability to minimize organ rejection, posttransplant outcomes improved significantly. From 1963 to 1979, 170 patients underwent liver transplantation at the University of Colorado; 56 survived for 1 year, 25 for 13–22 years, and several remain alive today 30 years after their surgery. Continued improvement in posttransplantation outcomes were achieved, and thus, in 1983, the National Institutes of Health (NIH) established that liver transplantation was no longer considered an "experimental" procedure, rather, as definitive therapy for appropriately selected patients with end-stage liver disease. Additional advances in immunosuppression (reviewed below) including the discovery of polyclonal and monoclonal antibodies to T-cells or their receptors, and other agents such as Tacrolimus, Mycophenolate Mofetil, and Sirolimus have further improved outcomes.

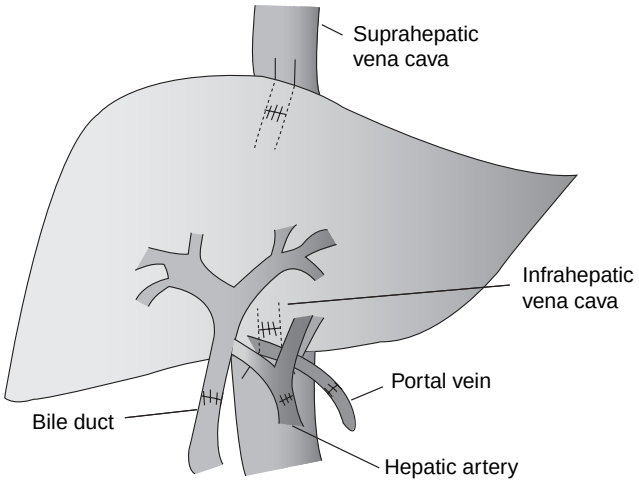


Figure 1. Schematic representation of an orthotopic liver transplant.

INDICATIONS FOR LIVER TRANSPLANTATION

Liver transplantation is an accepted therapeutic modality for complications of chronic liver disease, or acute liver failure. In general, liver transplantation is recommended when a patient with end stage liver disease manifests signs and symptoms of hepatic decompensation, not controlled by alternative therapeutic measures. This is evidenced by

1. Esophageal and/or gastric variceal bleeding, or bleeding from portal hypertensive gastropathy.
2. Hepatic encephalopathy.
3. Spontaneous bacterial peritonitis.
4. Significant ascites.
5. Coagulopathy.

Patients with liver disease complicated by early stage hepatocellular carcinoma (HCC), often defined as either a single lesion <5 cm or not more than three lesions, each <3 cm are also considered candidates for liver transplantation irrespective of evidence of concomitant hepatic decompensation (4).

If a patient meets these initial criteria, further requirements must be realized. It is generally accepted that liver transplantation is indicated if the patient is not moribund and the transplant is likely to prolong life with a $>50\%$ chance of 5-year survival. Furthermore, it is anticipated that the transplant will restore the patient to a range of physical and social function suitable for the activities of daily living. Patients who are suitable candidates should not have comorbid disease with involvement of another major organ system, which would preclude surgery or indicate a poor potential for rehabilitation.

Transplant candidates undergo a thorough psychological assessment prior to liver transplantation. Adequate family and social support must be demonstrated to ensure adherence to the difficult long-term medical regimen that will be required posttransplant. In addition, if a history of substance abuse is present, most transplantation programs require that the patient complete at least 6 months of documented rehabilitation with displayed freedom from alcohol and/or drug recidivism. "Psycho-social" assessment is usually performed by several individuals, including a Psychiatrist or Psychologist and an experienced social worker. In addition, living

Table 1. Diseases Associated with Fulminant Hepatic Failure

Viral Infection	
Frequent	Hepatitis A, B, D, E, Hepatitis Non A-G
Rare	Hepatitis C
	Cytomegalovirus
	Epstein Barr virus
	Herpes simplex virus
Metabolic	Acute fatty liver of pregnancy
	Reye's SX
Toxin, Drugs	Acetaminophen
	Nsaid's
	CCL4
	Isoniazid
	Sodium valproate
	Methyl DOPA
	Tetracycline
	Halothane
	Amanita phalloides (mushroom poisoning)
	Yellow Phosphorus
	"Herbal Medication"
Drug Combos	Acetaminophen and ETOH
	Acetaminophen and barbiturates, isoniazid trimethoprim, and sulamethoxazole amoxicillin and clavulinic acid
Ischemic	Hepatic artery thrombosis
	Budd-Chiari Syndrome
	Right ventricular failure, cardiac tamponade shock
Miscellaneous	Hyperthermia
	Hellp SX

Table 2. Indications for Liver Transplantation

Diseases Effecting Hepatic Parenchyma
Viral hepatitis with cirrhosis (Hepatitis B with or without Delta Virus, Hepatitis C, Non A-E hepatitis)
Autoimmune hepatitis
Alcoholic cirrhosis
Metabolic disorders (Wilson's disease, hemochromatosis, alpha 1 Antitrypsin, Tyrosinemia, protoporphyria, Cystic fibrosis, familial amyloidosis, Neiman-Pick disease)
Fulminant hepatic failure due to any cause
Drug induced liver disease
Diseases Effecting Biliary System
Primary and secondary biliary cirrhosis
Sclerosing cholangitis
Caroli's disease
Relapsing cholangiohepatitis
Choledochal cysts with obstruction and bilary cirrhosis
Hepatic Neoplasia/Malignancies
Patients with nonmetastatic primary hepatocellular carcinoma, with;
A single tumor not $>$ 5 cm
No more than three lesions with the largest lesion $<$ 3 cm
No thrombosis of the portal or hepatic vein
Hemangioendothelioma (confined to the liver)
Neuro endocrine tumors with hepatic involvement
Large hepatic Hemangioma
Miscellaneous Causes
Hepatic vein thrombosis (Budd-Chiari syndrome)
Portal vein thrombosis
Hepatic artery thrombosis
Trauma

donor liver transplantation (LDLT), discussed in greater detail below, requires a detailed psychosocial assessment of both recipient and potential donor. In most transplantation centers, an independent donor advocate team consisting of a social worker, internist, and surgeon who are independent of the team evaluating the recipient performs the difficult task of educating a potential donor regarding the risks and benefits of LDLT, assessing motivation to be a donor, and determining if coercion is present.

Another indication for liver transplantation is fulminant hepatic failure, defined as hepatic encephalopathy (confusion) arising in the setting of massive liver injury in a patient without preexisting liver disease. This condition is rapidly fatal unless recovery of hepatic function occurs spontaneously, and thus, emergent liver transplantation may be required. Conditions associated with fulminant hepatic are listed in Table 1.

ETIOLOGY OF LIVER DISEASES REQUIRING LIVER TRANSPLANTATION

Diseases associated with hepatic dysfunction in adults and children are outlined in Tables 2 and 3, respectively. In general, any disease process in adults or children that induces either acute or chronic hepatocellular, biliary, or vascular injury may necessitate liver transplantation. The indications for liver transplantation in children are identical to those in adults, that is, liver transplantation is indicated in the presence of progressive liver disease in patients who fail medical management. Many of the disease processes in adults that induce liver failure are recapitulated in children. However, specific disease states seen in children including metabolic diseases and congenital biliary anomalies represent additional indications for liver transplant. Moreover, liver transplantation is indicated in infants and children if the transplant will prevent or attenuate derangements in cognition, growth, and nutrition. Therefore, children should be considered for liver transplantation when there is evidence that hepatic decompensation is either unavoidable (based on knowledge of the history of the disease itself), imminent, or has already occurred. The clinical scenarios that determine when liver transplantation is required in children can include one or more of the following:

1. Intractable cholestasis.
2. Portal hypertension with or without variceal bleeding.
3. Multiple episodes of ascending cholangitis.
4. Failure of synthetic function (coagulopathy, low serum albumin, low cholesterol).
5. Failure to thrive or achieve normal growth, and/or the presence of cognitive impairment due to metabolic derangements, and malnutrition.
6. Intractable ascites.
7. Encephalopathy.
8. Unacceptable quality of life including failure to be able to attend school, intractable pruritis.
9. Metabolic defects for which liver transplantation will reverse life-threatening illness and/or prevent irreversible central nervous system damage.

Table 3. Additional Indications for Liver Transplantation in Infants and Children

Disease	Defect	Inheritance	Comments
Cholestatic Liver Disease			
Obstructive: Biliary Atresia (most common indication for liver transplantation in children)			
Intrahepatic: Alagille's Syndrome, Bylers disease, familial cholestatic symptoms			
Other			
Congenital hepatic fibrosis			
Metabolic Diseases			
Alpha 1 antitrypsin	Decreased serum A1AT	Codominant	May reverse both liver and lung disease
Wilsons's Disease	Decreased Ceruloplasmin	Autosomal Recessive (AR)	
Tyrosinemia	Fumarylacetoacetate hyrolase	AR	Transplant in fulminant liver failure, or to prevent hepatic neoplasia
Urea cycle defects	Example: ornithine transcarbamylase	x-linked dominant	Prevent CNS injury
	Arginosuccinate synthetase	AR	
Galactosemia	Galactose phosphate uridyl transferase	AR	Prevent development of cirrhosis and Hepatoma
Glycogen storage Diseases	Glucose 6 phosphatase	AR	Consider transplant if dietary management not successful
Type 1A	Brancher enzyme		
Type IV			
Familial hypercholesteroloemia	Type 2 A-LDL receptor deficiency	AR	Avoids ASHD
Gaucher's Disease	Glucocerebrosidase	AR	May need combined liver/bone marrow t-plant
Nieman-Pick disease	Sphingomyelinase	AR	
Crigler-Najjar type 1	Uridine diphosphate glucoronly transferase	AR	prevents fatal Kernicterus
Cystic fibrosis	Chloride ion transfer gene	AR	May need combined liver/lung transplant
Hyperoxaluria type 1	Alanine glyoxalate aminotransferase	AR	Usually requires combined liver/kidney
Neonatal Fe storage	Unknown	Varies	Transplant as infant
Hemophilia A and B	Factor VIII/IX	x-linked	Transplant indication varies (?iron overload, factor inhibitor present)
Disorders of bile acid synthesis (Bylers disease)	Unknown	Varies	Transplant indicated if associated with end stage liver disease

10. Life threatening complications of stable liver disease (e.g., hepatopulmonary syndrome).

CONTRAINDICATIONS TO LIVER TRANSPLANTATION (ADULTS AND CHILDREN)

At present, "absolute exclusion" criteria for liver transplantation are evolving. In general, patients with advanced cardiac or pulmonary disease, severe pulmonary hypertension, active substance abuse, coma with evidence of irreversible central nervous system injury, sepsis, or uncorrectable congenital abnormalities that are severe and life threatening are not transplant candidates. In addition, individuals with evidence of extrahepatic malignancy do not meet criteria for transplantation, unless the patient meets standard oncologic criteria for "cure". "Relative" exclusion criteria include renal insufficiency when renal transplantation is not feasible, prolonged respiratory failure requiring $> 50\%$ oxygen, advanced malnutrition, primary biliary malignancy, inability to understand the risk/benefits of the procedure, and inability to comply with medications and conform to follow-up regimens. Recent data indicates successful outcomes in HIV infected patients who undergo liver transplantation, a population formerly considered noncandidates for the procedure. However, initial enthusiasm regarding successful transplantation outcomes must be restrained by evidence that HCV recurrence in HIV–HCV coinfected patients may be problematic (5).

RECIPIENT CHARACTERISTICS AND PRIORITIZATION FOR TRANSPLANTATION

Given the relatively stable number of available donor organs in the setting of a rapidly expanding pool of potential recipients, the timing of transplantation is critical. Liver transplantation in a stable patient who is anticipated to do well for many years while waiting for an available organ may not be appropriate, while liver transplantation in a moribund patient with a low probability of posttransplantation survival is similarly inappropriate. Prior to 1997, prioritization for liver transplantation was based on the location where patients received their care (i.e., home, hospital, intensive care unit) and waiting time on the transplant list. In 2002, several policies were instituted by UNOS in an attempt to produce a more equitable organ allocation scheme. Waiting time and whether the patient was hospitalized were eliminated as determinants of prioritization of organ allocation. The "MELD" score (Model for End Stage Liver Disease) a logarithmic numerical score based on the candidate's renal function (creatinine), total bilirubin, and INR (international normalized ratio for prothrombin time) has been shown to be the best predictor of

mortality among cirrhotic patients, including those on the transplant waiting list. It was therefore adopted by UNOS as a mechanism to prioritize waiting list candidates. MELD had been validated as a predictor of 3-month survival in diverse groups of patients with various etiologies and manifestations of liver disease (6). Presently, a patient's position on the liver transplantation waiting list is now determined by their MELD score; patients with highest MELD scores are ranked highest on the list. Prospective analysis of the impact of MELD indicates improvement in both the rate of transplantation, pretransplantation mortality, and short-term posttransplantation mortality rates (7). However, retrospective analysis has suggested that posttransplantation survival may be reduced in patients with very high pretransplantation MELD score, particularly in Hepatitis C infected patients (8). Conversely, MELD score effectively delineates when a patient is "too well" for transplantation. A recent review indicates that posttransplantation survival in patients transplanted with a MELD score of <15 is lower than a nontransplanted cohort with similar MELD score (9). Thus, it is clear that careful recipient selection, with attention to pressor and ventilatory requirements, need for dialysis, age, and MELD score are important factors in selecting appropriate candidates for liver transplantation.

LIVER TRANSPLANTATION: SOURCE OF ORGANS

At present, there are three potential types of organ donors specific to liver transplantation, identified as deceased, living, or non-heart beating. Deceased donors (DD) comprise the majority of liver donors. Either by self-identification while living, or after discussion with "next of kin" when donor brain death has been declared, individuals are acknowledged as potential organ donors. Recent data from UNOS indicate that 1- and 3-year patient survival in recipients of DD liver transplant is 81 and 71%, respectively (10). However, despite efforts to maximize utilization of organs acquired from DD including the use of older donors, steatotic (fatty) livers, and livers infected with Hepatitis C or B, a growing disparity exists between the number of available livers and the number of individuals waiting for transplantation. This critical shortage of organs has resulted in both an increase in the waiting time for liver transplantation and death rate among patients on the waiting list. In response, the modalities of adult-to-child and adult-to-adult LDLT have emerged as alternatives to deceased donor liver transplantation (11,12). Adult-to-child LDLT usually involves the removal of the left lateral segment of the liver (~20% of hepatic mass) from an adult donor for implantation into a child, while adult to adult living donor liver transplantation requires that the larger, right lobe of the liver (which accounts for ~50–60% of the hepatic mass) be removed from the donor to ensure adequate hepatic mass in the recipient. Rapid regeneration of the liver remnant in the donor and the partial allograft transplanted into the recipient occurs, to the extent that appropriate liver volume is restored within 1–2 months in both donor and recipient following surgery. Since most pediatric LDLT recipients are <2-years old, they receive a liver graft of adequate or even excessive size, and thus liver insufficiency due to the receipt of inadequate liver mass is rare. In contradistinction, as the recipient of an adult-to-adult living donor liver transplantation receives a graft that must over time grow to an appropriate volume, selection of recipients best able to tolerate transplantation of a "partial" graft is necessary. In appropriately selected pediatric and adult recipients, 1- and 3-year graft and patient survival in individuals who undergo LDLT is similar or superior to DD (10). However, when comparing postoperative complications in recipients of DD versus LDLT, recipients of LDLT have a greater rate of biliary complications including bile leaks and biliary strictures, which occur in 15–32% of patients (13). In addition, the "small-for-size syndrome" manifested as prolonged posttransplantation cholestasis with or without portal hypertension may occur following LDLT, if the graft is of inadequate size (14). Fortunately, the majority of patients who experience this syndrome recover without the requirement of retransplantation.

Recently, significant interest in the utilization of "non-heart beating" donors (donation after cardiac death, DACD) as a potential modality to further increase the pool of available organs has emerged. In contrast to DD who are declared brain dead, DACD are critically ill patients who are not brain dead, but have no expectation of recovery and who based on their own prior wishes or families request are removed from life support. Following cardiac arrest and declaration of death, organs are harvested. There are two types of DACD, "controlled" and "uncontrolled". In the controlled DACD (Maastricht category 3 "death anticipated") the patient is removed from life support and death occurs in the operating room. Once death has been declared, organs deemed suitable for transplantation are rapidly perfused with preservation solution and removed surgically. The uncontrolled DACD (Maastricht category 1 and 2 "death not anticipated") is declared dead after cardiac arrest, rushed to the operating room, and organs are harvested. Uncontrolled DACD are usually not utilized for liver transplantation due to the high rate of primary nonfunction (defined below), usually due to prolonged ischemia of the graft. When utilizing controlled DACD for transplantation, emerging data indicates that recipient and graft survival are diminished when compared to deceased and living donor liver transplantation with a higher incidence of primary nonfunction, biliary injury, and requirement for retransplantation. However, several centers have reported acceptable outcomes when utilizing controlled DACD organs, particularly those without significant ischemia in well-selected recipients (15).

Finally, "domino" transplantation is an option for patients afflicted with familial amyloidotic polyneuropathy (FAP). Familial amyloidotic polyneuropathy is a fatal disease caused by an abnormal amyloidogenic transthyretin (TTR) variant generated by the liver. Liver transplantation in these patients removes the source of the variant TTR molecule, and represents the only known curative treatment. As no intrinsic liver disease exists in patients affected by FAP, the liver explanted from a patient with FAP may be transplanted into another patient, thus, allowing "domino" transplantation. Survival in both recipients of FAP livers and transplanted FAP patients has

been reported to be excellent and comparable to survival with OLT performed for other chronic liver disorders (16).

POSTTRANSPLANTATION MANAGEMENT

The complex nature of the surgical procedure utilized to both explant (remove) the diseased, cirrhotic liver and implant (transplant) the new allograft into the recipient make it intuitive that the majority of the early complications following liver transplantation are technical and related to the surgical procedure itself. However, following the first postoperative days, and as patients progress to the first month posttransplantation and beyond, the nature and variety of complications change. Early complications (within the first 2 months) and late complications (beyond 2 months) may negatively affect patient and graft survival (Table 4). Complications specific to the surgical procedure and those that directly affect the transplanted organ are discussed below.

EARLY COMPLICATIONS

Primary Nonfunction and Early Graft Dysfunction

A major threat to the newly transplanted liver is primary graft nonfunction (PNF). This syndrome defined as acidosis, rising INR, progressive elevation in liver transaminases and creatinine, and decreases in mentation occurs when the newly transplanted liver allograft fails to function normally. The mechanisms responsible for this phenomenon are complex, and relate to donor factors, inadequate preservation of the liver, prolonged ischemia, extensive steatosis of the graft, hepatic artery thrombosis (see below) or immune response to the implanted organ (17). In the setting of PNF, a rapid assessment of hepatic artery flow needs to occur, as immediate surgical repair of a thrombosed hepatic artery may reverse PNF. In the absence of hepatic artery thrombosis, emergent retransplantation is required for PNF.

In contrast to PNF, early graft dysfunction (EGD) is manifested by an early rise in serum transaminases to values $>$ 2000–3000 IU/L, cholestasis with rising Bilirubin levels, without marked coagulopathy or impairment in mental status and renal function. EGD may occur in the setting of ischemic injury or steatosis in the graft, and typically occurs within the first 24–48 h after the transplant. Unlike PNF, the manifestations of EGD usually improve with supportive care, and emergency retransplantation is not necessary.

Table 4. Early and Late Complications Following Liver Transplantation

Early
Graft Specific
Primary nonfunction
Early graft dysfunction
Hepatic artery thrombosis
Hepatic and portal vein thrombosis
Preservation injuryBiliary complications: bile leak, biliary stenosis
Acute cellular rejection
Other
Bacterial and fungal infection
CMV infection
Recurrent Hepatitis B and C
Late
Graft Specific
Chronic rejection
Recurrence of primary disease
Other
Hypertension
Hyperlipidemia
Diabetes
Obesity
Cardiac disease
Renal dysfunction
Fungal infection (Cryptococcus, Aspergillus)
CMV
Posttransplant lymphoproliferative disorder
Nonhepatic malignancy: i.e., skin cancer

Hepatic Artery Thrombosis

A potentially devastating posttransplantation complication is hepatic artery thrombosis (HAT). Hepatic artery thrombosis occurs more commonly in pediatric transplant recipients compared to adults due to the technical difficulties associated with the anastomosis of smaller size vessels. In HAT, the immediate postoperative period may be associated with graft failure, elevation in serum liver transaminases, bile leak, hepatic necrosis, and sepsis. Since the blood supply to the biliary tree in the early posttransplant period is principally from the hepatic artery, HAT is frequently associated with irreversible injury to the biliary tract (18). Thus, HAT in the first 7 days after liver transplantation is an indication for emergent artery repair or retransplantation.

Due to the potentially devastating consequences of HAT, most transplant centers screen for this complication with duplex-ultrasound (US) in the immediate posttransplant period. If duplex-US suggests HAT, angiography is usually performed to confirm the diagnosis, and if present, surgical revision of the hepatic artery is required. If surgical repair cannot be achieved, liver retransplantation may be necessary.

PORTAL AND HEPATIC VEIN THROMBOSIS

Though less common than HAT, thrombosis of the portal and/or hepatic veins in the immediate posttransplant period can also adversely affect patient and graft survival. Acute "Budd-Chiari" syndrome due to hepatic vein or vena cava thrombosis is associated with abdominal pain, peripheral edema, and the threat of graft failure, as hepatic congestion in the newly transplanted liver is poorly tolerated. In this circumstance, emergency thrombectomy and surgical revision is required. Acute portal vein occlusion may be associated with exacerbation of preexisting portal hypertension, associated with gastrointestinal bleeding from porto-systemic collateral vessels such as esophageal and gastric varices. Acute portal vein thrombosis is managed by surgical repair, while chronic portal vein thrombosis may be well tolerated. A potential alternative to surgical repair for both hepatic and portal vein stenosis or occlusion is thrombolysis and/or the placement of

endovascular stents by an experienced interventional radiologist (19).

ACUTE CELLULAR REJECTION

Rejection of any transplanted organ is a constant threat, as immunologic recognition of the graft as "foreign" may be associated with injury. However, compared to other organs, liver allografts are relatively privileged immunologically, and thus, the incidence and consequences of acute cellular rejection (ACR) are diminished when compared to other solid organs utilized for transplantation. The reported incidence of ACR within the first posttransplant year is 30–50%, in most cases, usually occurring within the first 2–3 weeks postoperatively. The clinical presentation is variable; ACR may be asymptomatic, or associated with fever or abdominal pain. Laboratory findings include elevation or failure of normalization of serum transaminases, usually in association with a rising alkaline phosphatase and/or bilirubin. The diagnosis of acute liver graft rejection is confirmed by liver biopsy and examination of liver histology (20). Conventional histologic criteria associated with ACR include the presence of periportal lymphocytic infiltrate, as well as bile duct and hepatic vascular endothelial cell injury. Most cases of ACR respond to treatment with intravenous bolus glucocorticoids. Approximately 10% of patients with ACR will not improve with intravenous glucocorticoids, requiring the administration of monoclonal or polyclonal anti-T cell antibodies (reviewed below). Mild and moderate ACR may also respond to either increasing the dose of the primary immunosuppressive agent, or switching to an alternate calcineurin inhibitor. This approach has been used with increasing frequency, particularly in patients transplanted for HCV and HBV due to concerns regarding the negative impact of over-immunosuppression on viral recurrence.

BILIARY COMPLICATIONS

Bile leaks and strictures generally occur at the anastomosis of the donor and recipient bile ducts, recognized by a rise in serum bilirubin and/or alkaline phosphatase or by the presence of bile in surgical drains in the immediate posttransplantation period. The incidence of biliary complications is between 5 and 15% following deceased donor liver transplantation. However, between 15 and 30% of patients who undergo living donor liver transplantation develop biliary complications, due to the complexity of the biliary reconstruction required (21). In both deceased and living donor recipients, the majority of bile leaks resolve spontaneously without the need for reoperation. As previously stated, the biliary tree receives the vast majority of its blood supply from the hepatic artery, and thus, the adequacy of hepatic artery blood flow needs to be evaluated in the setting of any biliary injury. If spontaneous resolution of the bile leak does not occur, endoscopic or radiologic placement of a biliary stent across the biliary anastamoses is often successful (22). In some cases, surgical exploration and revision of the biliary anastamoses with a Roux-en-Y choledochojejunostomy may be required.

Anastamotic biliary strictures require careful attention, as if left untreated, cholangitis, graft dysfunction, and eventually secondary biliary cirrhosis may occur. Techniques for management include dilatation and stenting via biliary endoscopy or percutaneous transhepatic cholangiogram by an interventional radiologist. If these modalities are unsuccessful, surgical revision of the biliary anastamosis with a Roux-en-Y choledochojejunostomy may be required. In rare cases with diffuse stricturing, retransplantation may be necessary.

ISCHEMIC AND PRESERVATION INJURY

The newly transplanted liver is always subjected to some degree of ischemic injury (23). Cold (or hypothermic) ischemia is unavoidable, as it occurs prior to transplantation while the liver is cooled in preservation solution, awaiting implantation. Warm (normothermic) ischemia occurs during the transplantation procedure itself, when hepatic blood flow is interrupted to minimize blood loss during transplantation, or when the formerly "cooled" liver is subjected to body temperature during transplantation. Cold ischemia is usually well tolerated, while in contrast, warm ischemia often leads to death of hepatocytes, with resultant elevation in serum transaminases, apoptosis and centrilobular necrosis. In the setting of significant warm ischemia, graft failure may result. Several investigators have noted improvement in ischemic injury and enhanced graft and patient outcomes by employing a technique described as "ischemic preconditioning" defined as a brief period of controlled ischemia followed by a short interval of reperfusion before the actual surgical procedure (24). This is accomplished during liver transplantation by transiently interrupting hepatic inflow by placing a vascular clamp or a loop around the portal triad (i.e., portal vein, hepatic artery, and bile duct), rendering the whole organ ischemic for 10–15 min, after which the clamp is removed and the liver is reperfused. This technique may be of particular benefit in organs with significant steatosis.

Complications Beyond Two Months

Progress in the surgical techniques required to perform transplantation, the treatment of postoperative complications and prevention of rejection have been associated with significant improvements in short-term morbidity and mortality following transplantation. Coincident with improvements in short-term outcomes has been a rise in long-term complications. These complications, including side effects of chronic immunosuppression, neoplasia, and infections are discussed in detail elsewhere. Long-term complications that affect the transplanted liver are discussed below.

CHRONIC REJECTION

Chronic allograft rejection or "vanishing bile duct syndrome" is rare, but in contradistinction to acute cellular rejection, a much more difficult to treat complication. Diagnostic criteria for chronic rejection include bile duct atrophy affecting the majority of bile ducts, with or without bile duct loss. Arterial and venous injury affecting the

large branches of the hepatic artery or portal vein (foamy arteriopathy) may also be present (24). Risk factors for chronic liver rejection include transplantation for primary sclerosing cholangitis, primary biliary cirrhosis, HLA mismatch between donor and recipient, and cytomegalovirus infection. Chronic rejection is usually a harbinger of poor outcomes, often resulting in the requirement for retransplantation; altering immunosuppression is rarely associated with improvement.

Recurrence of Primary Disease Following Liver Transplantation

A major challenge to the liver transplant community is recurrence of the primary disease that caused the patients native liver to fail. Diseases that do not recur following liver transplantation include congenital anatomic anomalies (e.g., biliary atresia, polycystic liver disease, Caroli's disease, Alagilles syndrome, congenital hepatic fibrosis) and metabolic diseases of the liver (e.g., Wilson's disease, alpha 1 antitrypsin deficiency). However, all other causes of liver disease including primary biliary cirrhosis, primary sclerosing cholangitis, autoimmune hepatitis, nonalcoholic fatty liver disease, hemochromatosis and alcohol related liver disease have been reported to recur after liver transplantation. In some cases, recurrent disease may lead to significant liver injury with resultant graft failure (26–30). Disease processes most commonly associated with recurrence include viral hepatitis B (HBV) and C (HCV). The recurrence of HBV is associated with uniformly poor outcomes with graft failure and death. Fortunately, recurrence of HBV after liver transplantation can be prevented by administering hepatitis B immune globulin (HBIG) at the time of transplantation and at regular intervals thereafter, with or without the use of antiviral agents such as Lamivudine and Adefovir. In contradisctinction to HBV, HCV recurrence following liver transplantation remains a significant source of morbidity and mortality, with negative impact on post-transplantation outcomes. In patients with active HCV replication prior to transplantation, reacquisition of viremia following transplantation is universal, and histologic injury due to HCV occurs in up to 90% of patients followed for 5 years (31). Although histologic injury in the allograft due to HCV is exceedingly common, disease progression after the development of hepatitis is variable, with some patients experiencing indolent disease and others rapidly progressing to cirrhosis and liver failure. In patients that develop HCV associated cirrhosis posttransplantation, up to 42% will experience decompensation manifested as ascites, encephalopathy, or hepatic hydrothorax, and $<50\%$ of patients survive >1 year after the development of decompensation (32). It is important to contrast the natural history of HCV before and after transplant; prospective and retrospective data are emerging which indicate that the progression of HCV following liver transplantation is accelerated when compared to the nonimmunosuppressed pretransplant patient population.

Whether HCV recurrence is more severe in recipients of LDLT than in DD recipients is controversial. Although several recent reports indicate that HCV recurrence may be more problematic in recipients of LDLT when compared to DD (33), particularly the cholestatic variant of HCV (34), other authors have noted no differences in outcomes in HCV infected patients who undergo LDLT when compared to DD (35,36). At present, both the optimal timing for transplant in HCV patients and the therapy for recurrent HCV following liver transplantation are incompletely described. Theoretically, eradication of HCV prior to liver transplantation in patients with decompensated liver disease would be beneficial, although in practice, this strategy has been marred by exacerbation of encephalopathy, infections, and other serious adverse events, particularly in patients treated with high dose Interferon and ribavirin (37). A novel approach including initiating therapy with low dose interferon (including Pegylated interferon preparations) and ribavirin with slow escalation in dose may be associated with improved tolerability and efficacy (38). Following liver transplantation, both preemptive therapy prior to the development of histologic injury and directed therapy after the onset of liver injury have been attempted with varying degrees of success. It is important to note, however, that posttransplantation, tolerability of interferon preparations, and ribavirin is suboptimal. Significant leucopenia and anemia are common, likely due to drug induced bone marrow suppression and renal insufficiency potentiating ribavirin induced hemolysis (39).

Immunosuppressive Medications

A cornerstone to posttransplantation management is the ability to prevent or attenuate immunologic rejection of the transplanted graft, which when left untreated, can be associated with graft failure. From a conceptual standpoint, understanding how recognition of the newly engrafted liver as "foreign" occurs, how to modulate immune mediated injury, and at the same time prevent "overimmunosuppression" are critical to achieve optimal post transplantation outcomes. The various immunosuppressive medications and their mechanism of action currently utilized in liver transplant recipients are listed in Table 5. Unfortunately, all immunosuppressive therapy is associated with undesired effects, with the potential for additive effects when agents are combined. In general, most transplant centers utilize three agents to prevent allograft rejection in the immediate posttransplant period, utilizing a combination of a calcineurin inhibitor such as Cyclosporine (CYA) or Tacrolimus (TAC), a second agent such as Mycophenolate mofetil (MMF) or Azathioprine (AZA), and a glucocorticoid such as Prednisone. As patients achieve adequate liver function and freedom from rejection beyond 6-months posttransplantation, satisfactory immunosuppression can be achieved in many patients with monotherapy, usually with a calcineurin inhibitor, although in patients who are at increased risk of rejection such as those with autoimmune hepatitis, primary biliary cirrhosis, or sclerosing cholangitis, long-term immunosuppression is achieved with a combination of a calcineurin inhibitor with either low dose MMF or Prednisone (40).

Corticosteroids

Corticosteroids achieve their desired immunosuppressive affects by the suppression of leukocyte, macrophage, and cytotoxic T-cell activity, and diminution of the effect of

Table 5. Immunosuppressive Agents

Agent	Mechanism of Action	Side Effects
Antilymphocyte globulin	Depletes circulating lymphocytes	Flu-like symptoms Anaphylaxsis
Antithymocyte globulin		Lymphoproliferative disorders
OKT3	Depletes circulating T cells	Flu-like symptoms Anaphylaxsis Lymphoproliferative disorders
Basiliximab Daclizumab	IL-2 receptor blockade	Infections Gastrointestinal distress Pulmonary edema and bronchospasm (rare)
Cyclosporine	Inactivates calcineurin, decreases IL2 production, Inhibits T-cell activation	Hypertension Renal insufficiency Neuropathy Hyperlipidemia Gingival hyperplasia HirsutismInsulin resistance
Prednisone	Suppression of leukocyte, macrophage, and cytotoxic T-cell activity Decrease cytokines, prostoglandins, and leukotrienes	Hypertension Dyslipidemia Glucose intolerance Bone abnormalities Peptic ulcers Psychiatric disorders
Azathioprine	Inhibits adenosine and guanine production Inhibits DNA and RNA synthesis in rapidly proliferating T cells	Leukopenia Anemia Thrombocytopenia Pancreatitis
Tacrolimus	Inactivates calcineurin, decreases IL2 production, Inhibits T-cell activation	Hypertension Renal insufficiency Insulin resistance Neuropathy Hyperlipidemia
Mycophenolate mofetil	Inhibits of inosine monophosphate dehydrogenase (IMPDH) Prevents T- and B-cell proliferation	Leukopenia Anemia Thrombocytopenia GI side effects
Sirolimus	inhibiting mTOR (target of Rapamycin) Prevents T-cell replication.	Hepatic artery thrombosis Bone marrow suppression Hyperlipidemia Pneumonitis Inhibits wound healing

cytokines, prostaglandins, and leukotrienes. However, hypertension, dyslipidemia, glucose intolerance, bone loss, peptic ulcers and psychiatric disorders are often associated with therapy. Therefore, a strategy to taper and discontinue glucocorticoids within the first 6 months–1 year following transplantation while maintaining adequate levels of calcineurin inhibitor is employed by many transplant centers. This tactic is often altered in patients who undergo liver transplantation secondary to an immunologic disorder such as autoimmune hepatitis, primary biliary cirrhosis and sclerosing cholangitis due to an enhanced risk of acute cellular rejection. In these patients, either long-term use of corticosteroids with an attempt to minimize doses is advocated, or chronic use of MMF or AZA in combination with a calcineurin inhibitor is required.

T-Cell Depleting Agents

In the past, "induction therapy" with antilymphocyte agents such as antilymphocyte globulin or antithymocyte globulin or monoclonal antibody preparations such as OKT3 was utilized immediately after liver transplantation to rapidly induce an immune suppressed state via the rapid destruction of the host's T cells. However, due to significant systemic side effects including fevers, allergic reactions, serum sickness, and thrombocytopenia, the use of these agents is now usually reserved for the treatment of glucocorticoid resistant rejection, or less commonly, in patients with severe renal insufficiency in an attempt to delay the use of either CYA or TAC, which may be associated with worsening of renal function (41).

IL-2 Receptor Blockers

T-cell activation and proliferation following presentation of a foreign antigen requires the induction of several cytokines, including IL-2 (interleukin 2). Antibodies directed against the interleukin (IL)-2 receptor are effective for initial immunosuppression, as IL-2 receptor blockade

down regulates IL-2 mediated T-cell proliferation. The IL-2 receptor antibodies such as Basiliximab and Daclizumab, given intravenously at the time of transplant and during the first posttransplantation week can reduce the incidence of acute liver graft rejection when utilized in combination with a calcineurin inhibitor, although these agents may not be sufficient to prevent rejection when utilized alone. The IL-2 receptor antibodies are generally well tolerated, although side effects may include infections, gastrointestinal distress, and rarely, pulmonary edema and bronchospasm. As these agents rarely induce renal dysfunction, many transplant programs utilize IL-2 receptor antibodies as "induction" therapy in individuals with renal insufficiency at the time of transplantation (42), in an attempt to delay initiation or diminish dose of calcineurin inhibitors, which may exacerbate renal insufficiency.

Calcineurin Inhibitors

IL-2 inhibition effectively suppresses T-Cell activation. Cyclosporine and TAC achieve this by binding to cytoplasmic receptors, forming complexes which inactivate calcineurin, a key enzyme in T-cell signaling. The major side effects of both CYA and TAC include hypertension, renal insufficiency, and neurologic complications. However, there is evidence to suggest that obesity, hyperlipidemia, hirsutism, and gingival hyperplasia occur more commonly in patients who receive CYA, while a higher rate of diarrhea, insulin resistance, and diabetes is seen in patients who receive TAC. In response to inconsistent absorption of standard Cyclosporine, the development of a microemulsified formulation of cyclosporine (e.g., Neoral) has allowed consistent blood levels (43). Given their efficacy and oral administration, calcineurin inhibitors have a central role in posttransplant immunosuppression.

Safety and efficacy of calcineurin inhibitors is generally assessed by monitoring blood levels drawn prior to the dose (trough), although several investigators describe that blood levels drawn 2 h after a dose of Cyclosporine (i.e., C2 levels) rather than trough levels more accurately indicate exposure to drug. At many transplantation centers, the definition of appropriate target level of calcineurin inhibitor is linked to the patients time posttransplantation; in general, higher levels are required in the first several months postoperatively while the threat of rejection is acute. The target levels for calcineurin inhibitors can be appropriate adjusted downward as patients achieve both normal liver function and freedom from rejection months to years following surgery. In addition, a philosophy of minimizing exposure to high levels of calcineurin inhibitors in HBV or HCV infected patients is adopted by many transplant centers, due to the negative impact of "overimmunosuppression" on viral replication and disease recurrence.

Antiproliferative Agents

Antiproliferative agents such as AZA and MMF prevent the expansion of activated T cells and B cells and regulate immune mediated injury. Azathioprine, a purine analogue, is metabolized in the liver to its active compound, 6-mercaptopurine, which inhibits adenosine and guanine production, thus inhibiting DNA and RNA synthesis in rapidly proliferating T cells. Mycophenolate Mofetil is a potent noncompetitive inhibitor of inosine monophosphate dehydrogenase (IMPDH), an enzyme necessary for the synthesis of guanine, a purine nucleotide. Mycophenolate Mofetil, when used in combination with a calcineurin inhibitor and steroids has been shown to be associated with lower rejection rates in the first 6 months posttransplantation when compared to AZA (44). The major toxicities associated with the use of either MMF or AZA are bone marrow suppression with resultant leukopenia, anemia, and thrombocytopenia, though this is more marked with AZA. Mycophenolate Mofetil has been associated with a greater incidence of dyspepsia, peptic ulcers, and diarrhea when compared to AZA, while pancreatitis may occur in individuals prescribed AZA. These side effects usually abate by dose reduction or discontinuation. The majority of transplant centers utilize a combination of a Calcineurin inhibitor with either MMF or, less commonly, AZA for at least the first 3–6 months posttransplantation. Since AZA and MMF do not cause renal insufficiency, they can be utilized in a strategy to minimize or avoid calcineurin inhibitor use, particularly in patients with renal dysfunction.

Other Immunosuppressive Agents

The limitations and untoward effects of available immunosuppressive agents have induced research and development of alternative agents. Sirolimus (Rapamycin) (RAPA) and its derivative Everolimus represent a new class of compounds, which achieve their immuosuppressive effect by inhibiting mTOR (target of Rapamycin). Inhibition of mTOR diminishes intracellular signaling distal to the IL-2 receptor and prevents T-cell replication. As the lymphoproliferative pathways inhibited by RAPA and Everolimus are distinct from those affected by calcineurin inhibitors, investigators have utilized these agents in combination with calcineurin inhibitors to achieve synergistic effect. However, enthusiasm for RAPA has been tempered by recent data showing higher rates of hepatic arterial thrombosis in patients who receive RAPA in the weeks immediately following transplantation (45). In addition, impaired wound healing has been noted in patients who receive RAPA, potentially due to impairment of granulation mediated by inhibition of TGF-β. Leukopenia, thrombocytopenia, and hyperlipidemia are the principal toxicities associated with RAPA and Everolimus. Recent reports of pneumonitis in RAPA treated patients have also emerged. A positive attribute of both RAPA and Everolimus is the absence of renal toxicity; some data suggest that post transplantation renal insufficiency can be reversed when calcineurin inhibitors are withdrawn and RAPA is initiated (46). Newer immunosuppressive agents will continue to be developed; it is hoped that these agents will be associated with diminished short- and long-term toxicity and facilitate a state of "immune tolerance" of the graft that will ultimately allow minimization of the requirement for immunosuppressive medications.

SUMMARY

Liver transplantation is the treatment of choice for appropriately selected patients with end stage liver disease.

Over the last several decades, significant advances in surgical technique and immunosuppression, selection of appropriate donors, grafts, and recipients, and improved therapies to prevent and treat postoperative complications have greatly improved posttransplantation outcomes. Despite these impressive achievements, many challenges remain. It is becoming increasingly apparent that the growing disparity between the number of liver transplant candidates and available organs will be associated with escalating death rates on the transplant waiting list. Enhanced posttransplantation survival has led to the emergence of complications associated with patient longevity, including nonhepatic disease, complications of immunosuppression, infections, neoplasia, and recurrence of the primary disease for which the liver transplantation was indicated. Further progress in liver transplantation will be achieved by maximizing the use of available organs, refinement and exploration of alternatives to deceased donor liver transplantation, improvements in immunosuppression, and enhanced recognition and treatment of long-term complications, particularly recurrent liver disease.

BIBLIOGRAPHY

1. Welch CS. A note on transplantation of the whole liver in dogs. Transplant Bull 1955;2:54.
2. Kukral JC, Littlejohn MH, Williams RK, Pancer RJ, Butz GW Jr, Starzl TE. Hepatic function after canine liver transplantation. Arch Surg 1962;85:157–165.
3. Starzl TE, Marchioro TL, Porter KA, Brettschneider L. Related Articles, Homotransplantation of the liver. Transplantation 1967;5(Suppl):790–803.
4. Mazzaferro V, Regalia E, Doci R, Andreola S, Pulvirenti A, Bozzetti F, Montalto F, Ammatuna M, Morabito A, Gennari L. Liver transplantation for the treatment of small hepatocellular carcinomas in patients with cirrhosis. N Engl J Med 1996;334: 693–699.
5. Moreno S, Fortun J, Quereda C, Moreno A, Perez-Elias MJ, Martin-Davila P, de Vicente E, Barcena R, Quijano Y, Garcia M, Nuno J. Martinez Liver transplantation in HIV-infected recipients. Liver Transpl 2004;11:76–81.
6. Kamath PS, Wiesner RH, Malinchoc M, Kremers W, Therneau TM, Kosberg CL, D'Amico G, Dickson ER, Kim WR. A model to predict survival in patients with endstage liver disease. Hepatology 2001;33:464–470.
7. Freeman RB, Wiesner RH, Edwards E, Harper A, Merion R, Wolfe R. United Network for Organ Sharing Organ Procurement and Transplantation Network Liver and Transplantation Committee. Results of the first year of the new liver allocation plan. Liver Transpl 2004;10:7–15.
8. Onaca NN, Levy MF, Sanchez EQ, Chinnakotla S, Fasola CG, Thomas MJ, Weinstein JS, Murray NG, Goldstein RM, Klintmalm GB. A correlation between the pretransplantation MELD score and mortality in the first two years after liver transplantation. Liver Transpl 2003;9:117–123.
9. Merion RM, Schaubel DE, Dykstra DM, Freeman RB, Port FK, Wolfe RA. The survival benefit of liver transplantation. Am J Transpl 2005;5:307–313.
10. http://www.UNOS.org.
11. Broelsch CE, Whitington PF, Emond JC, Heffron TG, Thistlethwaite JR, Stevens L, Piper J, Whitington SH, Lichtor JL. Liver transplantation in children from living related donors. Surgical techniques and results. Ann Surg 1991;214: 428–437.
12. Marcos A. Right-lobe living donor liver transplantation. Liver Transpl 2000;6:S59–S63.
13. Shiffman ML, Brown RS Jr., Olthoff KM, Everson G, Miller C, Siegler M, Hoofnagle JH. Living donor liver transplantation: summary of a conference at The National Institutes of Health. Liver Transpl 2002;8:174–188.
14. Emond JC, Renz JF, Ferrell LD, Rosenthal P, Lim RC, Roberts JP, Lake JR, Ascher NL. Functional analysis of grafts from living donors. Implications for the treatment of older recipients. Ann Surg 1996;224:544–552.
15. Otero A, Gomez-Gutierrez M, Suarez F, Arnal F, Fernandez-Garcia A, Aguirrezabalaga J, Garcia-Buitron J, Alvarez J, Manez R. Liver transplantation from maastricht category 2 non-heart-beating donors: A source to increase the donor pool? Transpl Proc 2004;36:747–750.
16. Ericzon BG, Larsson M, Herlenius G, Wilczek HE. Familial Amyloidotic Polyneuropathy World Transplant Registry. Report from the Familial Amyloidotic Polyneuropathy World Transplant Registry (FAPWTR) and the Domino Liver Transplant Registry (DLTR). Amyloid 2003;10:67–76.
17. Schemmer P, Mehrabi A, Kraus T, Sauer P, Gutt C, Uhl W, Buchler MW. New aspects on reperfusion injury to liver–impact of organ harvest. Nephrol Dial Transpl 2004;19:26–35.
18. Bhattacharjya S, Gunson BK, Mirza DF, Mayer DA, Buckels JA, McMaster P, Neuberger JM. Delayed hepatic artery thrombosis in adult orthotopic liver transplantation-a 12-year experience. Transplantation 2001;71:1592–1596.
19. Vignali C, Cioni R, Petruzzi P, Cicorelli A, Bargellini I, Perri M, Urbani L, Filipponi F, Bartolozzi C. Role of interventional radiology in the management of vascular complications after liver transplantation. Transpl Proc 2004;36:552–554.
20. Lefkowitch JH. Diagnostic issues in liver transplantation pathology. Clin Liver Dis 2002;6:555–570.
21. Fondevila C, Ghobrial RM, Fuster J, Bombuy E, Garcia-Valdecasas JC, Busuttil RW. Biliary complications after adult living donor liver transplantation. Transpl Proc 2003;35: 1902–1903.
22. Denys A, Chevallier P, Doenz F, Qanadli SD, Sommacale D, Gillet M, Schnyder P, Bessoud B. Interventional radiology in the management of complications after liver transplantation. Eur Radiol 2004;14:431–439.
23. Selzner N, Rudiger H, Graf R, Clavien PA. Protective strategies against ischemic injury of the liver. Gastroenterology 2003;125:917–936.
24. Clavien PA, Yadav S, Sindram D, Bentley RC. Protective effects of ischemic preconditioning for liver resection performed under inflow occlusion in humans. Ann Surg 2000;232: 155–162.
25. Demetris A, Adams D, Bellamy C, Blakolmer K, Clouston A, Dhillon AP, Fung J, Gouw A, Gustafsson B, Haga H, Harison D, Hart J, Hubscher S, Jaffe R, Khettry U, Lassman C, Lewin K, Martinez O, Nakazawa Y, Neil D, Pappo O, Parizhskaya M, Randhawa P, Rasoul-Rockenschaub S, Reinholt F, Reynes M, Robert M, Tsamandas A, Wanless I, Wiesner R, Wernerson A, Wrba F, Wyatt J, Yamabe H. Update of the international banff schema for liver allograft rejection: Working recommendations for the histopathologic staging and reporting of chronic rejection. An International Panel. Hepatology 2000 Mar; 31(3):792–799.
26. Neuberger J. Recurrent primary biliary cirrhosis. Baillieres Best Pract. Res Clin Gastroenterol 2000;14:669–680.
27. Wiesner RH. Liver transplantation for primary sclerosing cholangitis: timing, outcome, impact of inflammatory bowel disease and recurrence of disease. Best Pract Res Clin Gastroenterol 2001;15:667–680.
28. Molmenti EP, Netto GJ, Murray NG, Smith DM, Molmenti H, Crippin JS, Hoover TC, Jung G, Marubashi S, Sanchez EQ,

Gogel B, Levy MF, Goldstein RM, Fasola CG, Gonwa TA, Klintmalm GB. Incidence and recurrence of autoimmune/alloimmune hepatitis in liver transplant recipients. Liver Transpl 2002;8:519–526.
29. Burke A, Lucey MR. Non-alcoholic fatty liver disease, non-alcoholic steatohepatitis and orthotopic liver transplantation. Am J Transplant 2004;4:686–693.
30. Mackie J, Groves K, Hoyle A, Garcia C, Garcia R, Gunson B, Neuberger J. Orthotopic liver transplantation for alcoholic liver disease: A retrospective analysis of survival, recidivism, and risk factors predisposing to recidivism. Liver Transpl 2001; 7:418–427.
31. Berenguer M, Prieto M, Rayon J, Mora J, Pastor M, Vicente O, et al. Natural history of clinically compensated hepatitis C virus related graft cirrhosis after liver transplantation. Hepatology 2000;32:852–858.
32. Gane E. The natural history and outcome of liver transplantation in hepatitis C virusinfected recipients. Liver Transpl 2003;9:S28–S34.
33. Garcia-Retortillo M, Forns X, Llovet JM, Navasa M, Feliu A, Massaguer A, Bruguera M, Fuster J, Garcia-Valdecasas JC, Rimola A. Hepatitis C recurrence is more severe after living donor compared to cadaveric liver transplantation. Hepatology 2004;40:699–707.
34. Gaglio PJ, Malireddy S, Levitt BS, Lapointe-Rudow D, Lefkowitch J, Kinkhabwala M, Russo MW, Emond JC, Brown RS Jr. Increased risk of cholestatic hepatitis C in recipients of grafts from living versus cadaveric liver donors. Liver Transpl 2003;9:1028–1035.
35. Shiffman ML, Stravitz RT, Contos MJ, Mills AS, Sterling RK, Luketic VA, Sanyal AJ, Cotterell A, Maluf D, Posner MP, Fisher RA. Histologic recurrence of chronic hepatitis C virus in patients after living donor and deceased donor liver transplantation. Liver Transpl 2004;10:1248–1255.
36. Russo MW, Galanko J, Beavers K, Fried MW, Shrestha R. Patient and graft survival in hepatitis C recipients after adult living donor liver transplantation in the United States. Liver Transpl 2004;10:340–346.
37. Crippin JS, McCashland T, Terrault N, Sheiner P, Charlton MR. A pilot study of the tolerability and efficacy of antiviral therapy in hepatitis C virus-infected patients awaiting liver transplantation. Liver Transpl 2002;8:350–355.
38. Everson GT. Treatment of chronic hepatitis C in patients with decompensated cirrhosis. Rev Gastroenterol Disord 2004;4: S31–S38.
39. Gane E. Treatment of recurrent hepatitis C. Liver Transpl 2002;8:S28–S37.
40. Conti F, Morelon E, Calmus Y. Immuosuppressive therapy in liver transplantation. J Hepatol 2003;39:664–678.
41. Tector AJ, Fridell JA, Mangus RS, Shah A, Milgrom M, Kwo P, Chalasani N, Yoo H, Rouch D, Liangpunsakul S, Herring S, Lumeng L. Promising early results with immunosuppression using rabbit anti-thymocyte globulin and steroids with delayed introduction of tacrolimus in adult liver transplant recipients. Liver Transpl 2004;10:404–407.
42. Liu CL, Fan ST, Lo CM, Chan SC, Ng IO, Lai CL, Wong J. Interleukin 2 receptor antibody (basiliximab) for immunosuppressive induction therapy after liver transplantation: A protocol with early elimination of steroids and reduction of Tacrolimus dosage. Liver Transpl 2004;10:728–733.
43. Lilly LB, Grant D. Optimization of cyclosporine for liver transplantation. Transpl Proc 2004;36:267S–270S.
44. Wiesner R, Rabkin J, Klintmalm G, McDiarmid S, Langnas A, Punch J, McMaster P, Kalayoglu M, Levy G, Freeman R, Bismuth H, Neuhaus P, Mamelok R, Wang W. A randomized double-blind comparative study of mycophenolate mofetil and azathioprine in combination with cyclosporine and corticosteroids in primary liver transplant recipients. Liver Transpl 2001;7:442–450.
45. Trotter JF. Sirolimus in liver transplantation. Transplant Proc 2003;35:193–200.
46. Nair S, Eason J, Loss G. Sirolimus monotherapy in nephrotoxicity due to calcineurin inhibitors in liver transplant recipients. Liver Transpl 2003;9:126–129.

See also DIFFERENTIAL COUNTS, AUTOMATED; PHARMACOKINETICS AND PHARMACODYNAMICS; IMMUNOTHERAPY.

LONG BONE FRACTURE. See BONE UNUNITED FRACTURE AND SPINAL FUSION, ELECTRICAL TREATMENT OF.

LUNG MECHANICS. See RESPIRATORY MECHANICS AND GAS EXCHANGE.

LUNG PHYSIOLOGY. See PULMONARY PHYSIOLOGY.

LUNG SOUNDS

ROBERT G. LOUDON
RAYMOND L. H. MURPHY

INTRODUCTION

Medical devices and instrumentation have developed rapidly in the last few decades, yet the first diagnostic medical instrument, the stethoscope, is still the most widely used, and it has changed only superficially in design and function.

The lungs, as we breathe, produce sounds that are transmitted to the body surface and to the mouth. The characteristics of these sounds convey information about the sound-producing and -transmitting structures. This information often has diagnostic value. Auscultation of the lungs is therefore widely taught and practiced. Textbooks of physical diagnosis present a body of information that has been derived by careful workers since the introduction of the stethoscope by R.T.H. Laennec in 1819. Much of that information was indeed presented by Laennec himself in his remarkable treatise, *De l'Auscultation Mediate*(1,2).

In this article, the medical devices and instruments that have been applied to the study of lung sounds, including the traditional acoustic stethoscope are reviewed. This survey will include sound transducers and their placement, methods, and equipment used for the recording and analysis of lung sounds, results obtained by the use of these techniques, and their clinical meaning. Recent work on this subject helps in the understanding of what we hear with the stethoscope; some is aimed at answering specific questions in physiology or pathology, and some is designed to provide new diagnostic and monitoring tools. Much of this work has been done in the past three decades, reflecting the enormous increase in the availability and quality of sound recording and processing techniques during that period. Reviews of lung sounds (3–5) and the success of the International Lung Sounds Association and its annual meetings bear witness to the upsurge of interest in the subject. Recommended standards for terms and techniques used in computerized respiratory sound

analysis (CORSA) have been prepared by a Task Force of the European Respiratory Society and published in the European Respiratory Review series (6). Better understanding of the meaning of current and future observations promises a larger place in the future for clinical and research applications.

THE STETHOSCOPE

The introduction of the stethoscope is an interesting story, well described in a bicentenary appreciation of Laennec's birth (7). Laennec, a young physician practicing in Paris, had occasionally found it useful to listen directly to a patient's chest, as had been done by physicians at least since the time of Hippocrates. In 1816, he wished to listen to an obese young lady's heart, but was reluctant to do so. He recollected (and this part of the story may be apocryphal) having seen boys playing on a park bench, one listening to the wooden bench at one end with his ear, and the other scratching the other end. Laennec's own words were that "he happened to recollect a simple and well know fact in acoustics, that sound could be transmitted through solid material or along a tube. He rolled a quire of paper into a sort of cylinder", placed one end over her heart, and listened at the other end. He was "not a little surprised and pleased" to hear the sounds more clearly in this "mediate" fashion than he had ever been able to do by the immediate application of his ear (2). Over the next 3 years he amassed an enormous amount of information about the sounds heard over the chests of his patients. As he did all of the autopsies at the Hopital Necker in Paris where he worked, he could often relate these sounds to the underlying pathology.

The first edition of Laennec's book (1) cost 13 francs for the two volumes; for an extra 2.50 francs, one received a wooden stethoscope. This "cylinder" served its purpose well. Modifications were introduced over the years, such as earpieces, flexible tubing, binaural stethoscopes, and a diaphragm on the chest piece. The relative merits of diaphragm and bell, the effect of the length and bore of the tubing, and the convenience of different patterns have been debated over the years, and the design of modern stethoscopes has been largely empirical, better models surviving because of their popularity with auscultators. Some characteristics that acousticians might think of as defects may indeed be advantageous from the physician's point of view. Those using them tend to feel comfortable listening to sounds with which they are familiar and may reject a stethoscope that lets them hear too much.

The assessment of acoustical performance of stethoscopes is not as simple as it might seem, and approaches to this problem have been described by several authors (8–10). The value of the traditional stethoscope is in no way reduced by the recent introduction of devices and instruments that can record and analyze the sounds that we hear. Rather, its value is increased. Appropriate use on new medical devices and instrumentation adds science to art, measurement to impression, and recordings to memory. Better understanding of what lung sounds mean, and of how much the simple stethoscope can tell us and how much it cannot, will make the use of the simple stethoscope in examining rooms or on clinical rounds more important than ever.

SOUND TRANSDUCERS

Microphones transform mechanical energy to electrical energy, in the sound frequency range. Mechanical movement at the chest wall, resulting from the transmission of vibrations representing lung sounds to the chest wall surface, may be detected by any one of several devices. The main categories are ceramic, condenser (capacitor), and electret microphones. Ceramic microphones use a piezoelectric ceramic element that produces voltage when it is stressed. They tend to be stable and rugged and do not need a bias voltage for operation. Condenser microphones of the conventional type act as a variable capacitor that requires a bias voltage. They have good sensitivity and frequency-response characteristics. Electret microphones are a more recent type; a permanent charge on the diaphragm and no free electrostatic charge on its surface relieve the need for a polarizing (bias) voltage and reduce sensitivity to humidity.

Most microphones are designed to receive sound transmitted through air. Air coupling has been used by several investigators recording sounds from the surface of the chest wall, or from the trachea, and it is not surprising that stethoscope chest-pieces have been used for this purpose. The sound transmitted through the air column in stethoscope tubing can be applied to a microphone just as it can to an auscultating eardrum. Direct mechanical coupling of the transducer to the signal site (chest wall or tracheal surface) is an alternative to air coupling.

Several authors have reviewed the relative advantages and disadvantages of the various types of microphones as lung sound transducers (11,12). Desirable characteristics include sensitivity, rejection of ambient noise and surface noise, appropriate frequency response, insensitivity to variation in pressure of application, ease of attachment, ruggedness, and low price. Sensitivity is necessary because of the low level of the sound signal. Vesicular breath sounds will on occasion be virtually inaudible, for example, when airflow rates at the mouth are <0.27 L/s (13).

It is not always possible to study lung sounds in ideal circumstances, and rejection of ambient noise is important for many applications. Microphone housing can be helpful in this regard. Heart sounds are often of greater amplitude than lung sounds and may obscure them. They can be made less troublesome by the frequency response of the microphone because heart sounds are in a lower frequency range. Air coupling or inherent microphone characteristics may help by increasing the high frequency response. Microphone placement can also reduce the interference from heart sounds, which are, of course, loudest over the front of the chest, particularly in the left lower zone, and are less obtrusive on the right side, especially at the base of the right lung posteriorly. One method that has been adopted to reduce contamination of lung sounds by the heart sounds is to record the electrocardiogram simultaneously and to use some form of gating to delete segments where the heart sounds are present (14). The periodicity of

the heart sound makes this an attractive alternative for some purposes. Muscle noise can also contaminate lung sounds; again the frequency content of muscle noise is considerably lower than that of lung sounds, and a microphone that is insensitive to low frequency noise, or subsequent filtration of the signal, can be helpful. Muscle noise has the disadvantage of being timed with respiration because it arises from respiratory muscle activity, and this prevents it from being gated out on a time base. Most investigators have found that the frequency range of most interest in the recording and analysis of lung sounds lies between 100 and 1000 Hz, well within the frequency range of most microphones.

Surface noise is another important source of difficulty that can arise in recording and interpreting lung sounds. The movements associated with respiration make it easy for the microphone to slide over the skin surface in phase with respiration, producing sounds that are in phase with respiration, may be in the same frequency range as lung sounds, and may be very difficult to distinguish from friction sounds such as a pleural friction rub. Surface noise is more likely to arise when the microphone is mechanically in contact with, but not firmly fixed to, the chest wall. Air coupling may have advantages over mechanical coupling in this respect, but not always if the chest piece is of the diaphragm type commonly used in stethoscopes. Respiratory movement may also cause changes in the pressure with which a microphone is applied to the chest wall; if the microphone is strapped to the chest by a circumferential band, pressure on the microphone will increase as inspiration occurs and the chest diameter increases. Variation in pressure of the microphone against the chest wall is liable to alter the acoustic coupling, particularly if mechanical coupling is used to transmit surface movement to the sensitive microphone element. If the pressure exerted is sufficient, the deformation of the sensitive element may approach the limit of its range, damping the signal. Air-coupled microphones are less sensitive to changes in pressure of application, provided that the air chamber between chest wall surface and microphone element is vented to theoutside, usually by a small-bore needle; but too large a vent may increase the amount of ambient sound recorded (15).

For some purposes, the sound transducer is applied only briefly at a specific site on the chest wall while a few breaths are recorded. For monitoring purposes, attachment of a sound recording device for a period of hours or overnight may be necessary. Lightness and small bulk are important in this type of application, and in some cases two-sided adhesive tape or an adhesive patch similar to that used for electrocardiograph electrodes is adequate for attachment.

If chest wall surface movement is unimpeded, the vibrations that correspond to the lung sound do not involve actual mechanical displacement of the chest wall surface by more than a few micrometers. A sensor applied to the surface may measure displacement or, if it applies a load to the chest wall surface, it may measure pressure rather than displacement, or a combination of the two. Some sound transducers measure acceleration rather than actual physical displacement. In each case, the reaction of the sensor to the signal being sensed will influence its characteristics. Inertia, rigidity, or counterpressure by the sensing element may cause distortion of the sound. Particularly in the case of accelerometers, the mass of the sensing element will determine its frequency response characteristics. It is not always clear what criteria are used in making a decision about microphone type. The human ear is remarkably good at separating out the different sounds that may be combined to form a mixed signal, and often the final judgment may be made by listening to replay of a recorded signal. The efficiency of a particular sound system depends on the purpose for which it is intended, but unless the signal is listened to with an educated ear it is easy to be misled by, for example, frequency components whose origin is not obvious from inspection of a graphic or calculated spectrum.

RECORDING AND DISPLAY SYSTEMS

Those using devices and instruments to study lung sounds will choose recording and display systems appropriate to their purpose. Audio tape and strip-chart recorders have now virtually all been replaced by computers or systems designed or modified for the purpose. The signals of interest may be presented to the observer audibly, visually, or in a variety of forms during and after analysis. Standard physical examination of the chest does, in a sense, present audible and visual displays to the clinician. The stethoscope presents an audible signal at the earpieces, and the clinician observes his patient breathe to get a visual display of respiratory movement.

For teaching purposes at the bedside, an electronic stethoscope or microphone may be connected to several headsets worn by students, by telemetry if preferred, giving the instructor an opportunity to share the sounds with them. In this way, a realistic learning experience is provided with less imposition on the patient's patience. Recording of sounds for teaching purposes usually involves a computer system, or an electronic stethoscope and audio tape recorder. Standard audiovisual equipment has been used for editing, for adding comments, and for preparation of cassettes or disks for distribution (16,17) for teaching purposes.

For research purposes, arrays of microphones are now available with computer recording, analysis, and display systems to show the distribution of sound signals over the surface of the chest (18–20). Brief differences in time of sound signals have clinical relevance by allowing comparison in timing of the same sound signal of a crackle or the start of a wheeze arriving at different surface sites in the same patient. And on a longer time base, in asthmatics, for example, the site, the frequency pattern, and the sound amplitude of wheezing may change during exercise, sleep, exposure to cold air or to inhalants such as pollen, or industrial exposure, or in response to drug treatment. Sleep disorders such as nocturnal asthma, the sleep apnea syndrome, and snoring, may be studied by sound monitoring. Nocturnal asthma and snoring are present in the same patient more often than would be expected as a result of chance alone, especially in asthmatics under the age of 40 (21). Snoring is a respiratory, but not a lung sound, as it

rises in the upper airways, at or above the larynx. Possible explanations for the association with asthma, and sound monitoring methods and devices, have been reviewed (22).

Comparisons over long periods of time were once made by recording the results of analyses of wheezes, rather than by comparing the actual recorded sounds. The development and proliferation of computers with rapidly increasing audiovisual capability and storage capacity are now, however, changing the situation to allow storage of original data on tape or disk together with derived values. Kraman et al. (23) evaluated minidisk recorders, with their considerable increase of storage capacity for music, for lung sound recording. They found no distortion of frequency or waveforms that would interfere with this use. For some studies, analyzing sound signals in real time as they are being acquired makes it simpler to monitor results as they accrue, and helps direct the course of an experiment.

An early example of audiovisual recording is in a paper by Krumpe et al. (24), in which the authors discuss the evaluation of bronchial air leaks by auscultation and phonopneumography. They describe three patients who develop air leaks from the bronchi after resectional lung surgery and in whom "videophonopneumography" provided more precise correlation of abnormal sounds with the underlying visibly leaking bronchial abnormalities. Audiovisual tapes or disks are useful for teaching or demonstration purposes, by providing examples of classical or of unusual sounds.

Simultaneous sound recordings at several sites have been used to study the spatial distribution of lung sounds. This has provided information on regional ventilation, and on the localization of abnormalities in disease such as pneumonia, airways obstruction, bullae, or small areas of infarction, atelectasis, fibrosis, or interstitial lung disease. Indeed, lung imaging by sound production provides a potential alternative to chest X rays and computed tomography (CT) scans, without the need to inject possibly damaging energy or drugs.

For research purposes, analysis of sound signals and any associated physiological measurements were formerly conducted off-line. The signals were recorded on tape or disk and replayed for analysis. This allowed editing for selection of relevant segments of data and for quality control and signal conditioning, such as amplification, filtering, or attenuation. The purpose of each study will determine the equipment needs, but most current lung sound research uses computers with high speed audiovisual capabilities. These can be adapted to record lung sounds along with physiological respiratory variables, such as airflow, lung volume, and esophageal pressure, measured simultaneously, which can then be related to the lung sounds. If relationships in time are to be studied with any precision, it is necessary to record signals together on one medium and it is necessary to know the frequency characteristics of the items of equipment used, and the time delays introduced by filters, envelope detectors, integrators, frequency analyzers, and other acquisition or processing devices.

SOUND ANALYSIS

Sound amplitude and frequency content are the two measurements that most commonly form the basis of lung sound analysis systems. Early studies presented the sound signal as a time-amplitude plot. If such plots represent a respiratory cycle on a few centimeters of paper, the result is a compressed representation that superimposes many successive sound signal cycles to form an envelope. Simple integrating and rectifying circuits can provide the outline of the envelope as a single line, thus acting as an envelope detector, ac–dc converter, or sound-level meter. Filters incorporated in such circuitry can yield a method for comparing sound amplitude in different frequency bands (14,17) or to provide a signal believed to represent the important band range of vesicular sound from the ventilation point of view (25).

The sound spectrogram is really an extension of this principle, the signal of interest being passed repetitively through a narrow bandpass filter with slowly changing center frequency and the signals passed being assembled to present a graphic display of time on the horizontal axis, sound frequency on the vertical axis, and sound amplitude by the degree of blackening of the paper. Sound spectrograms of this type, used routinely in the speech sciences, were applied to heart and lung sounds extensively by McKusick et al. (26) and are still widely used to good effect.

The time-amplitude plot of a sound signal has been used to advantage in a different way by Murphy et al. (27). Features of the sound waveform cannot be studied in detail without using a rapid time sweep on an oscilloscope, and only a brief (a few milliseconds) segment can be viewed in this way. By digitizing a sound signal at a rapid rate and playing the signal back through a digital-to-analogue converter (DAC), a "time-expanded" waveform was prepared. This has proved of particular value in studying crackles (rales), the brief sounds heard over fibrotic, edematous, consolidated, or atelectatic lung. Measurable characteristics of these crackles, such as the initial or the largest deflection width, show diagnostic value and automatic methods for their measurement are now being applied.

The sound characteristics of rhonchi, as opposed to crackles (continuous versus discontinuous adventitious sounds) require an additional approach. Essentially, they are longer in duration, possessed of perceptible pitch, and have a repetitive waveform pattern. Waveform analysis is a rapidly moving field. Sound frequency spectrum analysis of lung sounds has most frequently been reported in terms of discrete Fourier analysis. Several workers have used a fast Fourier transform algorithm to measure frequency content of signal segments. One way of representing time-variant sound signals is to assemble a sequence of spectra with frequency on the horizontal axis, sound amplitude or power on the vertical axis, and time on an oblique axis. Usually, some overlapping of the sequential segments and appropriate windowing (e.g., Hanning) are used. The resulting "bird's-eye view" has proved to be readily related to sounds, providing a mental image that can evoke a mental image of the sounds represented. Individual peaks on a frequency spectrum may be related to individual wheezes coming from the chest, and peak detection programs have been used (28,29) to compare them statistically. The fast Fourier transform is the most frequently reported type of waveform analysis, but other techniques, such as those of linear predictive coding (LPC), the

maximal entropy method of waveform analysis, fractal-dimension analysis, wavelet networks, and artificial neural networks, are being explored. They are most likely to prove useful in brief sounds, in timing the onset or rapid changes in complex sounds, or in noting time relationships among sounds recorded at separate or at adjacent sensors. Any graphic form of waveform analysis is more readily interpreted when it can be combined with visual examination of a simultaneous time-amplitude plot.

RESULTS AND CLINICAL APPLICATIONS

Increasing attention and techniques for more exact representation have led to a rapid growth in information available about lung sounds. The meaning of these various items of information will emerge more slowly, as will clinical applications. The objective, quantitative study of lung sounds, is still at an interesting rapid growth phase of development. It is clear that a good deal of information is contained in the signals that we hear emerging from the chest (2) and that auscultation is one of the safest of diagnostic procedures, since no external energy or chemical is inserted into the body. It is also clear that some of the information conveyed would be difficult to obtain in any other way. Much of it is regional or local and may be able to tell us about mechanical events and structural characteristics at specific sites in the chest (30,31). The vesicular lung sounds have been studied in sufficient detail that we now know more about the probable general range of bronchial dimensions involved in the production of these sounds, but not the exact site; the effects of flow rate and lung volume, but not the exact nature of the relationships; and we know that there are relationships between vesicular lung sound intensity and regional ventilation, but not their exact nature. The roles of production and of transmission of these sounds are not always easy to distinguish from one another in the end-product, sensed at the site of their detection, but recent work by Kiyokawa and Pasterkamp (32) shows progress in this distinction.

We know that wheezing indicates airflow obstruction and roughly its levels in the bronchial tree. We know that several factors, such as airway dimensions, geometry, and compressibility, are important. Endobronchial surface characteristics and the presence and nature of secretions may also have some effect. We know that flow rates and intrathoracic pressure and volume history affect wheezes; but we do not know the relative importance of these factors and the extent of variation from one disease state to another. Crackles are known to be associated with certain diseases and not with other radiographically similar diseases, but we are not sure why. We know that crackles from different types of abnormal lungs have different characteristics, but a great deal of clinical observation will be needed to test their diagnostic value: and physiological or pathological studies to understand the basic mechanisms involved.

Laennec's stethoscope—and for that matter the stethoscope pulled currently from the pocket of a white coat—allows the user to consider the sound of one breath at one place at one time. Medical devices and equipment are now being developed that can expand the observations in time, in space, in content, and in information; for example, from one or two breaths to hundreds of breaths, and from one specific point on the chest to the entire chest. From one breath described or remembered as vesicular, reduced in volume, with a few end-expiratory crackles the information may expand to an assembly of pages of tables and graphs showing a variety of measured features. These can include diagrams of the chest showing where and how the lungs and ventilation vary, where and how much airflow obstruction or lung collapse is present, and can offer a regional description of airways' diameters and other characteristics.

Que et al. (33) developed a system to measure tracheal flow from tracheal sounds, and to use this to estimate tidal volume, minute ventilation, respiratory frequency, mean inspiratory flow rate, and duty cycle. Careful observations and comparison of the results with simultaneously recorded pneumotachygraph-derived volumes in various postures allowed them to address the problems inherent in the adverse signal/noise ratio and the low level of the flow-derived sound at flow rates seen in quiet breathing. The system that they developed suggests that their method of phonospirometry measures overall ventilation reasonably accurately without mouthpiece, noseclip, or rigid postural constraints.

The study by Kiyokawa and Pasterkamp (32) in a sense complements this by measuring lung sounds at two closely spaced sensors on the chest surface. In five healthy subjects, volume-dependent variations in phase and amplitude of signals recorded over the lower lobe might reflect spatial variations of airways and diaphragm during breathing. These authors noted similar variations in phase and amplitude on passive sound transmission, suggesting that a difference in sound transmission was a more likely cause of the variations than a difference in sound generation. Their observations compare local sounds that reflect local circumstances; the observations discussed in the previous paragraph concern central sounds that reflect total ventilation.

Several systems are now available or under development that can record sound signals simultaneously from a number of sites, with or without associated physiological signals, and present the observations for read-out by the physician. Lung sound documentation and analysis can now be done on personal digital assistants (PDAs) as well as on laptop computers. Stethoscopes can be connected to these devices wirelessly or by a short cable. This allows objective quantification of these sounds at the bedside (34,35). A personal computer based "telemedicine" system has been described in which two remote hemodialysis sites were connected by high speed telephone lines to allow video and audio supervision of dialysis from a central site (36). Such equipment may eventually be used to supplement—or perhaps in some circumstances replace—other diagnostic devices such as fluoroscopy or other types of radiographic imaging. They have the great advantage of avoiding the subjection of a patient to any potentially harmful radiation or other energy, and can therefore be used over prolonged periods of time.

Transthoracic speed of sound introduced at the mouth or the supraclavicular space (35) can be mapped at several sites on the chest using sound input with specific

characteristics. This may allow noninvasive monitoring of conditions such as pneumonia, congestive heart failure, or pleural effusion that increase intrathoracic density. Chronic obstructive lung disease may be detected by reading lung sound maps showing time intensity plots at several sites over the chest; this appears to be more accurate than current clinical diagnostic methods. The ability to detect diaphragmatic movement by multichannel lung sound analysis suggests that it may prove to be an inexpensive bedside test. It may also have useful applications in ventilator management.

It seems clear that wider application of these new developments in lung sound analysis will lead to safe, useful, and rewarding forms of clinical and physiological information that can answer many imaging, diagnostic, and monitoring problems.

BIBLIOGRAPHY

1. Laennec RTH. De l'auscultation mediate ou traite du diagnostic des maladies des poumon et du coeur, fonde principalement sur ce nouveau moyen d'exploration. 1st French ed., Volumes 2, Paris: Brosson et Chaude; 1819.
2. Laennec RTH, trans. by Forbes J 1st American ed. Philadelphia: James Webster; 1823. p 211.
3. Mikami R, Murao M, Cugell DW, Chretien J, Cole P, Meier-Sydow J, Murphy RL, Loudon RG. International symposium on lung sounds. Synopsis of proceedings. Chest 1987;92:342–345.
4. Bettencourt PE, Del Bono EA, Spiegelman D, Herzmark E, Murphy RL. Clinical utility of chest auscultation in common pulmonary diseases. Am J Respir Crit Care Med 1994;150: 1291–1297.
5. Pasterkamp H, Kraman SS, Wodicka GR. Respiratory sounds. Advances beyond the stethoscope. Am J Respir Crit Care Med 1997;156:974–987.
6. Sovijarvi AHA, Vanderschoot J, Earis JE. Computerized Respiratory Sound Analysis (CORSA): Recommended standards for terms and techniques. Eur Respir Rev 2000;10: 77:585–649.
7. Sakula A. Laennec RTH 1781–1926. His life and work. A bicentenary appreciation. Thorax 1981;36:81.
8. Ertel PY. Stethoscope acoustics and the engineer: Concepts and problems. J Audio Eng Soc 1971;19:182–188.
9. Charbonneau G, Sudraud M. Measurement of the frequency response of some commonly used stethoscopes. Consequences to cardiac and pulmonary auscultation. Bull Eur Physiol 1985;21:49–55.
10. Abella M, Formolo J, Penney DG. Comparison of the acoustic properties of six popular stethoscopes. J Acoust Soc Am 1992;91:2224–2228.
11. Charbonneau G, Racineux JL, Sudraud M, Tuchais E. An accurate recording system and its use in breath sounds spectral analysis. J Appl Physiol 1983;55:1120–1127.
12. Pasterkamp H, Kraman SS, DeFrain PD, Wodika GR. Measurement of respiratory acoustical signals. Comparison of sensors. Chest 1993;104:1518–1993.
13. Kraman SS. Lung sounds: Relative sites of origin and comparative amplitude in normal subjects. Lung 1983;161:57–64.
14. Pasterkamp H, Fenton R, Tal A, Chernick V. Interference of cardiovascular sounds with phonopneumography in children. Am Rev Respir Dis 1985;131:61–64.
15. Kraman SS, Wodicka GR, Oh Y, Pasterkamp H. Measurement of respiratory acoustic signals. Effect of microphone air cavity width, shape, and venting. Chest 1995;108:1004–1008.
16. Cugell DW. Use of tape recordings of respiratory sound and breathing pattern for instruction in pulmonary auscultation. Am Rev Respir Dis 1971;104:948–950.
17. Banaszak EF, Kory RC, Snider GL. Phonopneumography. Am Rev Respir Dis 1973;107:449–455.
18. Kompis M, Pasterkamp H, Wodicka GR. Acoustic imaging of the human chest. Chest 2001;120:1309–1321.
19. Sun X, Cheetam BM, Earis JE. Real time analysis of lung sounds. Technol Health Care 1998;6:11–22.
20. Bergstresser T, Ofengeim D, Vyshedskiy A, Shane J, Murphy R. Sound transmission in the lung as a function of lung volume. J Appl Physiol 2002;93:667–674.
21. Fitzpatrick MF, Martin K, Fossey E, Shapiro CM, Elton RA, Douglas NJ. Snoring, asthma and sleep disturbance in Britain: a community-based survey. Eur Respir J 1993;6:531–535.
22. Dalmasso F, Prota R. Snoring: analysis, measurement, clinical implications and applications. Eur Respir J 1996;9:146–159.
23. Kraman SS, Wodicka GR, Kiyokawa H, Pasterkamp H. Are minidisc recorders adequate for the study of respiratory sounds?. Biomed Instrum Technol 2002;36:177–182.
24. Krumpe PE, Hadley J, Marcum RA. Evaluation of bronchial air leaks by auscultation and phonopneumography. Chest 1984;85:777–781.
25. Ploysongsang Y, Martin RR, Ross RD, Loudon RG, Macklem PT. Breath sounds and regional ventilation. Am Rev Respir Dis 1977;116:187–199.
26. McKusick VA, Jenkins JT, Webb GN. The acoustic basis of the chest examination: Studies by means of sound spectrography. Am Rev Tuberc 1955;72:122–134.
27. Murphy RLH, Holford SK, Knowler WC. Visual lung sound characterization by time-expanded wave-form analysis. N Engl J Med 1977;296:968–971.
28. Baughman RP, Loudon RG. Lung sound analysis for continuous evaluation of airflow obstruction in asthma. Chest 1985;88:364–368.
29. Pasterkamp H, Tal H, Leahy F, Fenton R, Chernik V. The effect of anticholinergic treatment on post exertional wheezing in asthma studied by phonopneumography and spirometry. Am Rev Respir Dis 1985;132:16–21.
30. Nath AR, Capel LH. Inspiratory crackles and the mechanical events of breathing. Thorax 1974;29:695–698.
31. Murphy RLH, Jr., Gaensler EA, Holford SK, Delbono EA, Eppler G. Crackles in the early detection of asbestosis. Am Rev Respir Dis 1984;129:375–379.
32. Kiyokawa H, Pasterkamp H. Volume-dependent variations of regional lung sound, amplitude, and phase. J Appl Physiol 2002;93:1030–1038.
33. Que C-L, Kolmaga C, Durand L-G, Kelly SM, Macklem PT. Phonospirometry for noninvasive measurement of ventilation: Methodology and preliminary results. J Appl Physiol 2002;93:1515–1526.
34. Bergstresser T, Ofengeim D, Vyshedskiy A, Shane J, Murphy R. Sound transmission in the lung as a function of lung volume. J Appl Physiol 2002;93:667–674.
35. Paciej R, Vyshedskiy A, Shane J, Murphy R. Transpulmonary speed of sound input into the supraclavicular space. J Appl Physiol 2002;94:604–611.
36. Winchester JF, Tohme WG, Schulman KA, Collman J, Johnson A, Meissner MC, Rathore S, Khanafer N, Eisenberg JM, Mun SK. Hemodialysis patient management by telemedicine: Design and implementation. ASAIO J 1997;43:M763–766.

See also PULMONARY PHYSIOLOGY; RESPIRATORY MECHANICS AND GAS EXCHANGE.

LVDT. See LINEAR VARIABLE DIFFERENTIAL TRANSFORMERS.

MAB. See MONOCLONAL ANTIBODIES.

MAGNETIC RESONANCE IMAGING

W. F. BLOCK
A. L. ALEXANDER
S. B. FAIN
M. E. MEYEREND
C. J. MORAN
S. B REEDER
K. K. VIGEN
O. WIEBEN
University of
Wisconsin–Madison
Milwaukee
Madison, Wisconsin

J. H. BRITTAIN
General Electric Healthcare
Milwaukee, Wisconsin

INTRODUCTION

The principle of nuclear magnetic resonance (NMR) was discovered by Felix Bloch and Edward Purcell independently in 1946. The two were awarded the Nobel Prize in Physics for the discovery, which had numerous applications in studying molecular structure and diffusion. Atomic nuclei with an odd number of protons or an odd number of neutrons behave like spinning particles, which, in turn, create a small nuclear spin angular momentum. This angular momentum of an electrically charged particle such as the nucleus of a proton leads to a magnetic dipole moment. In the absence of an external magnetic field, the orientation of these magnetic moments is random due to thermal random motion. These magnetic moments are referred to as spins, because the fundamentals of the phenomena can be explained using classical physics where the moments act similarly to toy tops or gyroscopes. The NMR phenomenon exists in several atoms and is used today to study metabolism via imaging. However, hydrogen is the simplest and most imaged nucleus in MR examinations of biological tissues because of its prevalence and high signal compared with other nuclei.

NMR imaging was renamed Magnetic resonance imaging (MRI) to remove the word nuclear, which the general public associated with ionizing radiation. MRI can be explained as the interaction of spins with three magnetic fields: a large static field referred to as B_0, which organizes the orientation of the spins; a radio frequency (RF) magnetic field referred to as B_1, which perturbs the spins so that a signal can be created; and spatially varying magnetic fields referred to as gradients, which encode the spatial location of the spins. These subsystems are shown in Fig. 1.

When an external magnetic field is present, the distribution of the magnetic moments is no longer random. Current technology allows large, homogenous static magnetic fields to be created using superconducting magnets, whereas smaller fields are possible with permanent magnets. In most conventional systems, the static field is aligned along the longitudinal axis or the long axis of the body, as shown in the z axis in Fig. 1. Clinical MRI scanners have been built with static fields ranging from 0.1 to 7 T, but the vast majority of scanners are between 0.5 and 3.0 T.

THEORY

Creating Net Magnetization

Consider a static field oriented along the z axis with magnitude B_0, or represented as a vector $\mathbf{B} = B_0\mathbf{k}$. Hydrogen protons have a quantum operator whose z component is quantized to $\pm$½. According to quantum mechanics, only two discrete sets of orientations exist for the magnetic dipole of each hydrogen nucleus. In the parallel energy state, the magnetic moment vector μ orients itself so that its projection on the z axis aligns with the direction of the main magnetic field B_0. In the antiparallel energy state, this projection aligns in the opposite direction of the main field. It can be shown that the two allowed angles between magnetic dipoles and the static field are $\theta = \pm 54^\circ$ (1), and thus a population of spins will be oriented as in Fig. 2b.

The ratio of spins in the parallel state n^- to the spins in antiparallel state n^+ is given by the Boltzmann equation

$$\frac{n^-}{n^+} = e^{\frac{-\Delta E}{kT}} = e^{-\frac{\gamma 2\pi}{h} B_0 kT} \tag{1}$$

where γ is a nuclei-specific constant referred to as the gyromagnetic ratio, k denotes the Boltzmann constant, and T is the absolute temperature. There are only slightly more spins in the parallel state than in the antiparallel state because this state is of lower energy; however, the prevalence of water in biological tissue can create an adequate signal with this differential. This distribution of spin orientations in a small volume element results in an average or net magnetization **M**, which aligns along the longitudinal or z axis, as shown in Fig. 2b. The entire process is referred to as polarization. The contributions in the transverse $(x-y)$ plane sum to zero. As the argument of the exponential in Equation 1 is small and the difference in energy levels varies proportionally with field strength, the length of the net magnetization vector varies linearly with field strength. A quantum mechanics description of the spin distribution can be found elsewhere (2).

Signal Generation

The behavior of the net magnetization vector in an external field can be described by the classical model according to

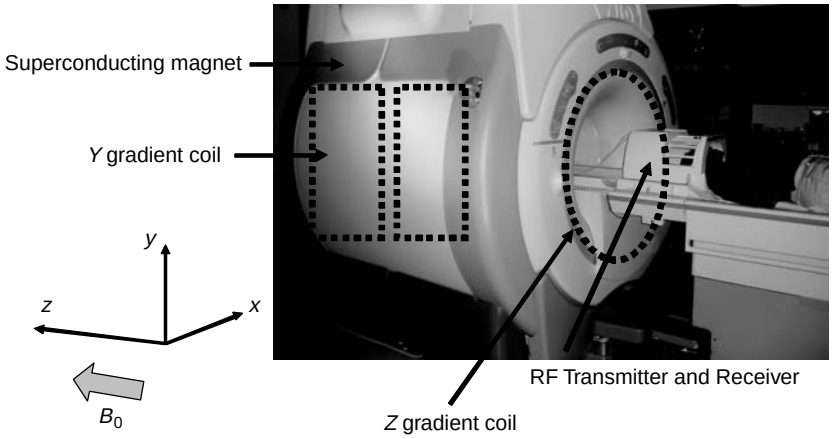


Figure 1. Clinical 1.5 T MRI scanner with static field oriented along long axis of the body (z). Patient's head lies in RF coil, which is used to both perturb and receive MR signal. Scanner bed will move patient into middle of cylinder before imaging begins. MR gradient coils for y dimension are shown, which have mirrored coils on the opposite side of the magnet. A portion of the z gradient, based on solenoid design, is also shown.

the Bloch equation.

$$\frac{d\mathbf{M}}{dt} = \gamma(\mathbf{M} \times \mathbf{B}) \tag{2}$$

A useful parallel description is a spinning toy top where the axis of the top is analogous to **M** and gravity is analogous to **B**. In the equilibrium state, the net magnetization **M** and the static magnetic field $\mathbf{B}_0$ are parallel so that **M** does not experience a torque and consequently the direction of **M** does not change. Similarly, the axis of a spinning top oriented vertically remains vertical.

The second magnetic field in MRI is an RF field that is created using an RF amplifier that supplies oscillating current into a coil that surrounds the patient. The coil is designed to create a magnetic field, referred to as $\mathbf{B}_1$ field, oriented in the transverse plane and approximately on the order of 10 T. By having the RFenergy oscillate at the resonant frequency of the nuclei, this relatively low field can perturb and rotate the net magnetization away from its orientation along the longitudinal axis. The resonant or Larmor frequency ω_0 is related to the static field strength such that $\omega_0 = \gamma B_0$. For protons, the gyromagnetic ratio $\gamma/2\pi = 42.57$ MHz/T. The field created by the tuned RF coil, referred to as an excitation, can be viewed as an applied torque that tips or flips spins away from the longitudinal axis by an angle referred to as the flip angle. The strength of the $\mathbf{B}_1$ field and the length of time it is applied determine the flip angle. The flip angle usually varies between 5 and 180° depending on the application.

Once the magnetization is no longer parallel to the static field, the right-hand side of Equation 5 is no longer zero and the direction of the net magnetization will change. In fact, it will begin to precess about the axis of the static magnetic field and at the Larmor frequency. In general, the precessional frequency is directly proportional to the magnetic field experienced by the spin, such that $\omega = \gamma B$. Similar to a toy top that is tipped an angle θ off its vertical axis, the top will maintain an angle of θ as it rotates about the vertical force supplied by gravity.

The net magnetization can be described by its longitudinal component M_z and its transverse component M_{xy}, a complex value whose magnitude describes the component's strength and whose angle describes the location of the component in the x–y plane. The rapid rotation of the transverse component will create a time-varying magnetic flux. A properly oriented receiver coil will detect this time-varying flux as a time-varying voltage, in agreement with Faraday's law of induction. Often, the same coil used for excitation can also be used for reception. This voltage signal, known as a Free Induction Decay, or FID, is shown after a 90° excitation in Fig. 3, which is the most basic form of a MR signal. Although the entire magnetization vector is tipped into the transverse plane in this example, smaller flip angles will also create a transverse magnetization and thus an FID.

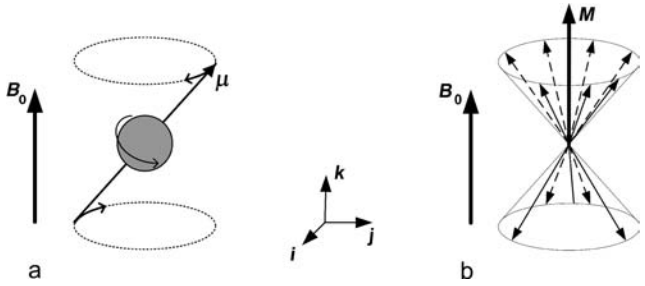


Figure 2. (a) shows the precession of a spin with a magnetic moment μ in a static field with magnetic flux density of B_0. An assembly of spins in parallel and antiparallel states is shown in (b). The Boltzmann equation determines the ratio of the spins in the two states. As the components in x and y compensate each other, the net magnetization **M** has a component in the z direction only (parallel to B_0). The coordinate system is shown with its unit vectors **i**, **j**, and **k** along x, y, and z.

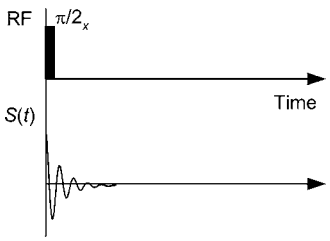


Figure 3. Generation of a free induction decay after a 90° RF excitation.

The complex motion of the net magnetization, and thus the recorded FID signal, can be described in a simplified manner by using a rotating reference frame that rotates at the Larmor frequency about the static B_0 field. In this rotating frame, the FID will decay as a simple exponential. The causes of this decay, and its use as potential image contrast mechanism, will be described after spatial encoding is described. Most MR signals are demodulating using the Larmor frequency and, thus are effectively acquired in the rotating frame.

Spatial Encoding

The first two magnetic fields described above allow biological tissue to be polarized, perturbed, and measured. In terms of clinical imaging, however, these fields merely allow us to integrate the signal derived throughout the body, a measure of little value. The field of MRI developed only when a third spatially varying magnetic field, referred to as a gradient field, was invented to spatially encode the MRI signal. This method allows us to achieve submillimeter resolution while using RF energy whose wavelengths are on the order of tens of centimeters to meters.

Dr. Paul Lauterbur realized, in 1973, that, instead of working like others to build a more homogenous field for NMR spectroscopy, spatially varying the strength of the magnetic field could provide a means to build an imaging system. For this work, he won the Nobel Prize in Medicine along with Sir Peter Mansfield in 2003.

The three gradient coils in a cylindrical MRI system, two of which are shown in Fig. 1, are laid out concentrically on a cylinder. The cylinder surrounds the patient as he or she lies inside of the static B_0 field. The three coils are designed to create longitudinal magnetic fields in z that vary in strength linearly with the x, y, and z dimensions, respectively. The digital scanner hardware controls the current waveforms, which are amplified by three respective gradient amplifiers before being sent to the gradient coils. The strength of each component gradient field, G_x,G_y, or G_z, is measured in G/cm or mT/m and is directly proportional to the current supplied to the coil. Changing gradient strengths quickly on clinical scanners is possible with amplifiers capable of slew rates of approximately 200 mT/m/s.

As the resonant frequency of an MR spin is directly proportional to the magnetic field it experiences, a gradient coil allows us to linearly vary the frequency of spins according to their position within the magnet. For example, a gradient of strength G_x, which does not vary in time, causes the frequency of spins to vary linearly with the x coordinate.

$$w(x) = \gamma[B_0 + G_x x] \tag{3}$$

Spins to the left of the magnet center rotate slower, spins at the exact magnet center remain unchanged, and spins to the right rotate faster than they did without the gradient.

Gradients can be used to selectively excite only spins from a slice or slab of tissue. Slice thicknesses in 2D MRI range from 1 to 20 mm. To select a transverse slice, the z gradient can be applied during RF excitation, which will cause the resonant frequency to vary as a function of z in the magnet, such that $w(z) = \gamma[B_0 + G_z z]$. Instead of exciting all the spins within the magnet, only spins whose frequency matches the narrow bandwidth of a pulsed-RF excitation will be excited within a slice at the center of the magnet. Modulating the frequency of the RF pulse up will move the slice superior in the body, whereas modulating it down will excite an inferior slice. Likewise, slices perpendicular to the x or y axis can be excited by applying a G_x or G_y gradient, respectively, simultaneously with RF excitation. In fact, an oblique slice orientation can be achieved with a combination of two or more gradients. The ability to control from which tissue signal is obtained, without any patient movement, is a major advantage of MRI.

Simplified MR Spatial Encoding. Once a slice of tissue is selected, the two remaining spatial dimensions must be encoded. A somewhat simplified method of visualizing encoding follows. For a transverse slice, receiver data could be obtained after RF excitation while a constant gradient was applied in the x direction. Tuning a receiver to select a very narrowband frequency range would determine which spins were present within a spatial range $x_1 < x < x_1 + \Delta x$. By repeating the experiment while changing the narrowband frequency range, a projection of the spin densities along the x axis could be determined. Likewise, the same experiment could be repeated while applying a constant y gradient to obtain a projection of spin densities along the y axis. Likewise, projections along arbitrary axes could be achieved by acquiring data while applying a combination of x and y gradients after RF excitation. In a matter very similar to computed tomography (CT) imaging, an image could be reconstructed from this set of acquired projections.

MR Spatial Encoding in the Fourier Domain. Although possible, the proposed method would be very slow because each sample point within each projection would require its own MR experiment or excitation. Time between excitations in MR vary in duration from 2 ms to 4 s depending on the desired image contrast. Even with the shortest excitation, each slice would require over 3 min of scan time. All the data for an entire projection can be acquired within milliseconds by considering how the phase of the transverse magnetization varies, instead of the frequency, with spatial position. This description also uses the concept that position in MR is encoding using an alternative Fourier domain where signal location is mapped onto spatial frequencies.

Integrating the frequency expression in Equation 3 indicates how the spin phase, or location of the transverse magnetization within the transverse plane, will vary with the x coordinate during a general time-varying gradient $G_x(t)$ applied after excitation. Ignoring the phase term due to the B_0 field, which will be removed during demodulation of the received signal, gives a phase term for each spin that varies with the spatial position x and the integral of the applied gradient at each point in time.

$$M_{xy}(x,t) = M_{xy}(x) e^{-j2\pi \frac{\gamma}{2\pi} \int_{t'=0}^{t} [Gx(t')x] dt'} \tag{4}$$

The signal received by the MR coil can be expressed as an integration of all the excited spins in the x–y plane using the following equation:

$$S(t) = \int_y \left[\int_x M_{xy}(x,y) e^{-j2\pi k_x(t)x} dx \right] dy \qquad (5)$$

where a Fourier spatial frequency, termed $k_x(t)$ in MR, has been substituted for $\frac{\gamma}{2\pi}\int_{t'=o}^{t}(G_x(t')x)dt'$. In this example, the coil simply integrates all the spins in the y dimension, and thus only spatial information in x is available. The received signal can be seen as a 1D Fourier transform of the projection of transverse magnetization $M_{xy}(x, y)$ onto the x axis. The corresponding coordinate in the Fourier domain at each point in time t is determined by the ongoing integral of the gradient strength. Thus, we can acquire an entire projection in one experiment rather than numerous MR experiments as in the simplified example with the narrowband receiver.

The last spatial dimension for this 2D imaging example y has a corresponding Fourier dimension termed k_y, which can be similarly traversed by designing the integral of the G_y gradient current.

$$S(t) = \int_y \int_x M_{xy}(x,t) e^{-j2\pi k_x(t)x} e^{-j2\pi k_y(t)y} dxdy \qquad (6)$$

where $k_y(t) = \frac{\gamma}{2\pi}\int_{t'=o}^{t}(G_x(t')x)dt'$. The integral of the gradients determines location in the Fourier space known as k space in MRI. Numerous strategies can be used to traverse and sample k space before transforming the data, often with a Fast Fourier Transform (FFT), into the image domain. The method can be extended to three dimensions by exciting a slab of tissue and using the G_z gradient to encode the third dimension.

As in the simplified example, a combination of G_x and G_y can be used to sample the 1D Fourier expression of projections of M_{xy} at arbitrary angles. This data can be transformed into the actual projections using 1D inverse Fourier transforms. Methods very similar to computed tomography can translate the projection data into an image. Although acquiring data in this manner, known as radial imaging, has interesting properties, by far the most popular method in clinical imaging traverses the Fourier space in a Cartesian raster pattern known as spin-warp imaging.

This sampling is typically completed in a series of experiments, where the time between consecutive excitations is referred to as the repetition time TR. In many MR acquisition schemes, a complete k space line is acquired along k_x, known as the frequency-encoding or readout direction, for each TR. During each subsequent TR, a line parallel to the previous one is sampled after applying a short, pulsed G_y gradient. By changing the strength of the pulsed G_y gradient by equal increments during each MR experiment, a different phase shift is placed on spins depending on their position in y. In terms of the k space formalism, the area under the G_y gradient pulse causes a vertical displacement in k space such that a different horizontal line in k space is acquired in each MR experiment, as shown in Fig. 4. Here, the vertical direction in k space is known as the phase-encoding direction. By applying 1D Fourier transforms in the k_x direction, spin position is resolved based on their frequency during readout. An image is formed by following these horizontal transforms with 1D Fourier transforms in the k_y dimension. Here, the y position of spins is resolved due to the different phase shifts experienced in each experiment prior to the readout gradient.

The image coverage, or field of view, in MRI decreases as the sampling rate decreases. As MR samples in the frequency domain, failure to sample fast enough in k space leads to aliasing in the image domain. Higher resolution in MRI requires obtaining higher spatial frequencies or larger extents of k space. Achieving adequate resolution and coverage then increases the amount of k space sampling that is required and increases imaging time. Unlike other modalities where hundreds to thousands of detectors can be used at a time, encoding spatial position in this method only allows one point of data to be taken at a time, which explains MR's relatively slow acquisition speed. Industrial scanners have only recently determined how to partially bypass this limitation by using up to 32 different receivers who have different spatial sensitivities to different tissues. The differences of each receiver in proximity, and thus sensitivity to each spin, can be used to synthesize unacquired regions of k space.

Image Contrast Through Varying Decay Rates

Imaging the spatial density distribution of hydrogen often produces a very low contrast image, as the density of hydrogen is relatively consistent in soft tissue. However, the imaging experiments described above can be easily modified to exploit the differences in time for which the

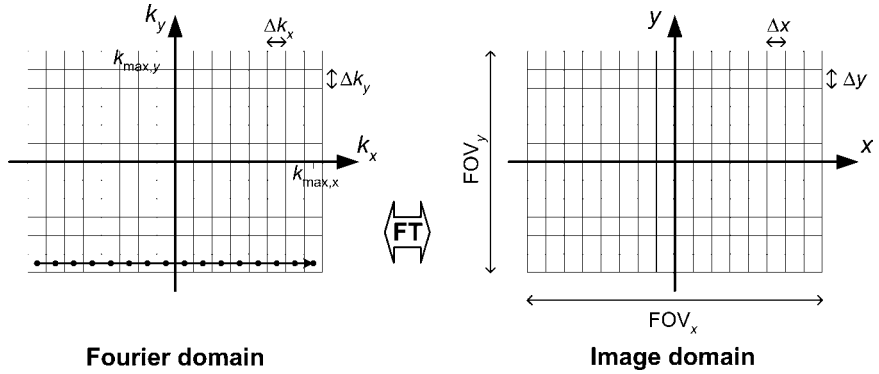


Figure 4. 2D spin-warp imaging with one readout per TR. One complete line is sampled along the readout direction on a rectilinear grid in the Fourier domain (k space) with a resolution of Δk_x (circles on the arrow). The next line is acquired parallel at a distance of Δk_y on the grid by increasing the phase-encoding gradient. This scheme is repeated until the desired grid is sampled. Images are reconstructed by an inverse 2D Fourier transform (FT).

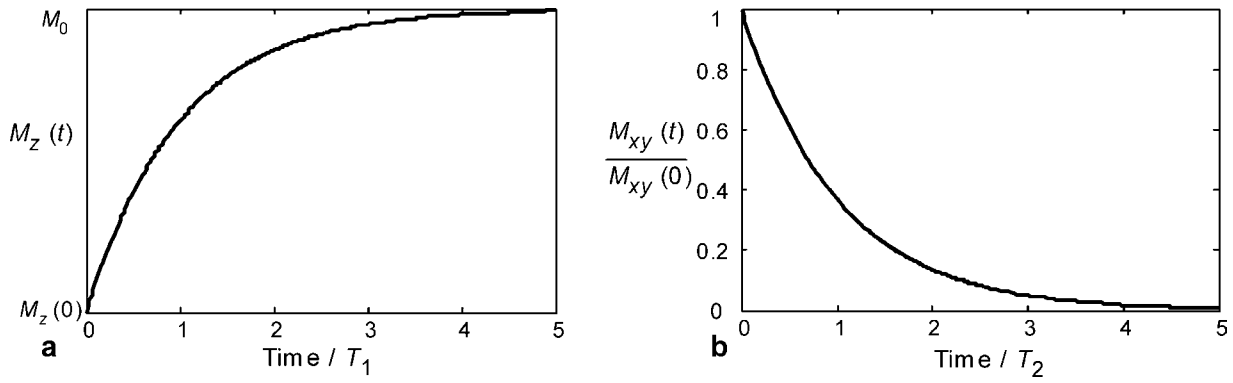


Figure 5. The regrowth of the longitudinal magnetization M_z (a) and the decay of the transverse magnetization M_{xy} (b) after an RF excitation.

MR spins remain perturbed. The differences account for the vast majority of image contrast in standard clinical MRI.

After the spins are perturbed, the transverse magnetization decays toward zero, whereas, the longitudinal magnetization returns toward its equilibrium magnetization. As more mechanisms exist for the loss of transverse magnetization than for the regrowth of longitudinal magnetization, the length of the magnetization vector **M** does not remain constant after excitation. Although related, the rate of longitudinal relaxation time, termed T_1, is always larger than the rate of transverse relaxation time, termed T_2.

If the magnetization has been completely tipped in the transverse plane with a flip angle of 90°, then the longitudinal magnetization recovers as

$$M_z = M_0[1 - e^{-t/T_1}] \tag{7}$$

T_1 is also called the spin–lattice relaxation time, because it depends on the properties of the nucleus and its interactions with its local environment. The transverse relaxation time T_2 is also referred to as the spin–spin relaxation time, reflecting dephasing due to interactions between neighboring nuclei.

$$M_{xy} = M_{xy}(0)e^{-t/T_2} \tag{8}$$

where $M_{xy}(0)$ is the initial transverse magnetization ($M_{xy}(0) = M_0$ for a 90° pulse). The temporal evolution of the longitudinal and transverse magnetization is shown in Fig. 5. In general, hydrogen protons in close proximity to macromolecules have lower relaxation times than bulk water that is freer to rotate and translate its position.

Delaying the encoding and acquisition of the transverse magnetization until some time after RF excitation generates T_2 image contrast. As injured and pathological tissues generally have higher T_2 relaxation rates, T_2-weighted images have positive image contrast. T_1-weighting can be achieved by using an interval between MR experiments, the TR parameter, which does not allow enough time for tissue to fully recover its longitudinal magnetization. Thus, tissues with shorter T_1 relaxation rates will recover more quickly and thus have more signal present in the subsequent experiments used to build the image than tissues with longer T_1. In general, T_1-weighting provides negative contrast for pathological tissue. An example is shown in Fig. 6 for a human brain tumor. The differing rates of

Figure 6. The flexibility of MR to image in different planes with different types of image contrast is shown in these sagittal and axial brain tumor (arrows) images.

Table 1. Longitudinal (T_1) and Transverse (T_2) Proton Nuclear Magnetic Resonance Relaxation Times for Several Tissues and Blood at 1.5 T

Tissue	T_1/ms	T_2/ms
Gray brain matter (3)	950	100
White brain matter (3)	600	80
Cerebrospinal fluid (CSF) (3)	4500	2200
Muscle (3)	900	50
Fatty tissue (3)	250	60
Oxygenated blood	1200	220
De-oxygenated blood	1200	120

recovery can also be used to null out an unwanted tissue, such as fat, by inverting all the spins 180° prior to imaging. As the point where unwanted tissue passes through the null of the recovery phase, an imaging experiment is begun. This technique is referred to as inversion recovery magnetization preparation or simply inversion recovery. Table 1 lists representative relaxation times for some tissues (3). Extensive reviews of the relaxations times (4) and methods for their measurement (3,4) are available.

Spin Echoes

Ideally, the transverse magnetization decays according to T_2. However, the signal dephasing in the transverse plane is significantly accelerated by field inhomogeneities due to difference in magnetic susceptibility between tissue types or the presence of paramagnetic iron. The largest inhomogeneities occur at air/tissue interfaces, such as near the sinuses or near the diaphragm. These effects lead to different precession frequencies and loss of coherence that are described by a T_2^* relaxation time

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T'_2} \tag{9}$$

where T'_2 represents the decay due to the effects described above.

A method to reverse these often undesirable dephasing effects uses a 90° pulse followed by a 180° spin refocusing pulse after a time delay τ_d, as shown in Fig. 7b. The first pulse rotates the longitudinal magnetization into the transverse plane as in the case of the FID. Prior to the second pulse, the magnetization dephases in the transverse plane due to T_2^* effects, with some spins rotating faster than the Larmor frequency and others spinning slower. The second pulse flips all magnetic moments about an axis in the transverse plane, effectively inverting the phase accruals due to different rotational frequencies. Over the second interval τ_d, the faster spins will catch up with the slower spins. As a result, a spin echo is said to form at the time $2\tau_d$, also known as the echo time or TE time. The amplitude of the signal at time TE is only decreased due to T_2 decay whereas the T_2^* effects have been reversed. A simplified pulse sequence for the generation of a spin echo is shown in Fig. 7b without the gradient waveforms necessary for spatial encoding.

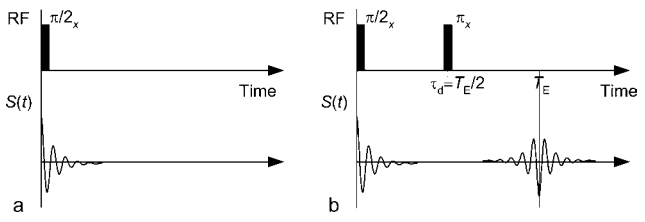


Figure 7. Generation of a (a) free induction decay and (b) a spin echo. In (a), other local factors dephase signal faster according to a T_2^* decay rate. If a second RF pulse is applied at time $\tau_d = TE//2$ to flip the magnetization by 180°, the spins will refocus and form an echo at $TE = 2\tau_d$, which is only subject to T_2 decay.

Signal-to-Noise Ratios

Signal in MR is generally proportional to the number of nuclei and, thus, to the volume of the image voxel. Noise in MR is caused by the random fluctuations of electrons in the patient, and thus the source of noise is independent from the signal generating sources. The data acquisition system is designed such that the noise level from properly designed MR electronics will be dominated by patient noise. Overall, $SNR = \text{voxel volume} * \sqrt{\text{total data sampling time}}$.

Imaging Sequences

Ideally, all MRI would be performed with high spatial resolution, a high signal-to-noise ratio (SNR), ultrashort imaging time, and no artifacts. The difficulty in achieving all of these properties simultaneously has led to the development of many acquisition methods that differ in image contrast, acquisition speed, SNR, susceptibility to and type of artifacts, energy deposited in the imaged patient, and suppression of unwanted signal such as fat. Their corresponding images represent a combination of tissue-specific parameters T_1, T_2, proton density ρ, and scan-specific parameters such as repetition time (TR), echo time (TE), flip angle, field of view (FOV), spatial resolution, and magnetization preparation.

Gradient Recalled Echo (GRE) Imaging

Spin-echo imaging is desirable because signal voids due to magnetic field inhomogeneity are avoided that could mask pathological tissue or injury. Long repetition times, and thus long scan times, are necessary in spin-echo imaging to allow longitudinal magnetization to return after the relatively high flip angles used. Long scan times hinder the capture of dynamic processes such as the beating heart, cause discomfort to the patient, and limit the number of patients who can be imaged with this expensive resource. Thus, other methods of imaging have been developed. In gradient recalled echo (GRE) imaging, the echo is formed by dephasing and rephasing of the signal with gradient fields as shown in Fig. 8. In these diagrams, known as pulse sequence diagrams, plots of the time-varying gradient and RF waveforms are shown as function of time. Compared with the spin-echo sequences, gradient recalled echo imaging does not have a refocusing RF pulse. The absence of this pulse allows for a reduced minimal repetition time and echo time compared with spin-echo imaging, but the signal becomes susceptible to T_2^* decay rather than T_2 decay.

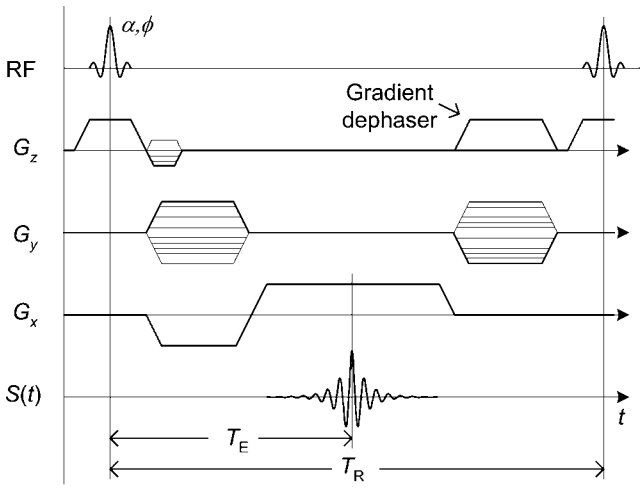


Figure 8. Basic 2D gradient echo pulse sequence. First, the magnetization is tipped into the transverse plane by an angle α during the application of a slice select gradient G_z. Then, the gradients G_y and G_x are used for phase encoding and the readout gradient. An echo forms at $t = T_E$ when the area under the readout gradient is zero. This experiment is then repeated every T_R with a different phase encode.

After RF excitation, the signal is dephased along the readout direction x with a prewinding gradient lobe. The amplitude of this gradient is then inverted to rephase the spins. When the area under the readout gradient is zero, the trajectory passes through the origin of k space and the echo forms with maximum amplitude. During the prewinder along the x axis, a gradient in y is played out to produce y depending on phase shifts for phase encoding.

By using a flip angle less than 90°, significant amounts of transverse magnetization are available without the need for long repetition times needed for recovering longitudinal magnetization. For example, after a single 30° excitation, the transverse magnetization contains sin(30°) or one-half of the available magnetization. Meanwhile, the longitudinal magnetization still contains cos(30°) or nearly 87% percent of the equilibrium magnetization. In fast GRE imaging, the repetition time is significantly reduced and generally less than the T_2 values of biological tissues. Under this condition, the transverse magnetization from preceding RF pulses is not completely dephased and generally results in a complex superposition of echoes from multiple RF pulses. Under certain conditions, a steady-state can be reached from repetition to repetition for one or more components of the magnetization (6).

GRE sequences can be used to generate T_1, T_1/T_2, T_2, T_2^*, and proton density-weighted contrast, depending on the choice of TR, TE, the flip angle α, and the phase φ of the RF pulse. By altering the phase of the RF transmit pulse in a pseudo-random method, the steady state of the transverse magnetization can be scrambled while the beneficial aspects of the longitudinal steady state are maintained. Although the signal from the transverse steady state is lost and only the signal from the current RF pulse is available, strongly T_1-weighted images are available with this technique, known as RF spoiling or spoiled gradient recalled (SPGR) imaging. This technique is popular with contrast-enhanced MR angiography, where an intravenously injected paramagnetic contrast agent significantly decreases the T_1 of blood while the T_1 of static tissues remains unchanged.

In an opposite approach, known as steady-state free precession (SSFP), the maximum amount of the transverse magnetization is maintained by rewinding all gradients prior to each RF pulse. The method provides T_2-like contrast very quickly and has proven very popular when fast imaging is essential such as in cardiac imaging.

Other Rapid MR Imaging Methods

In many applications, a short scan time is required to reduce artifacts from physiological motion or to observe dynamic processes. Many techniques exist to reduce the scan time while preserving high spatial resolution. One way to decrease spin-echo imaging time is to acquire multiple or all k space lines after a single preparation of the magnetization as explored with RARE (Rapid Acquisition with Relaxation Enhancement) (7). Also referred to as fast or turbo spin echo, the method works by creating a train of spin reversal echoes for which one line of k space is acquired for each. In echo-planar imaging (EPI) (8), an oscillating G_x gradient is used to quickly create many gradient echoes. By adding small blip gradients in between the negative and positive pulses of G_x, different horizontal lines in k space can be acquired.

Although the first MRI method proposed the acquisition of projections (9) as in CT, acquiring k space data on a Cartesian grid is fairly robust to magnetic field inhomogeneities and other system imperfections. Although spin-warp imaging (10) is predominant today, k space can be sampled along numerous 2D or 3D trajectories. The PROPELLER technique (11) acquires concentric rectangular strips that rotate around the origin, as shown in Fig. 8c. This method offers some valuable opportunities for motion correction due to the oversampling of the center of k space. K space can be sampled more efficiently with fewer echoes using spiral trajectories (12), as shown in Fig. 8d. Here, the amount of k space that can be acquired in one excitation is limited only by T_2 decay and possible blurring due to off-resonance spins. In nonCartesian acquisitions, phase errors due to off-resonance spins cause blurring. The sampling trajectories for these acquisitions schemes are shown in Fig. 9.

Applications

MRI is quickly moving beyond morphological and anatomical imaging. The advent of new functional image contrast mechanisms is making MR a tool for a much wider group of people than radiologists. Psychology, psychiatry, neurology, and cardiology are just some of the new areas where MR is being applied. A description of application areas in functional brain, diffusion-weighted brain, lung, MR angiography, cardiac, breast, and musculoskeletal imaging follows.

Functional Magnetic Resonance Imaging (fMRI). Functional Magnetic Resonance Imaging (fMRI) is a method of

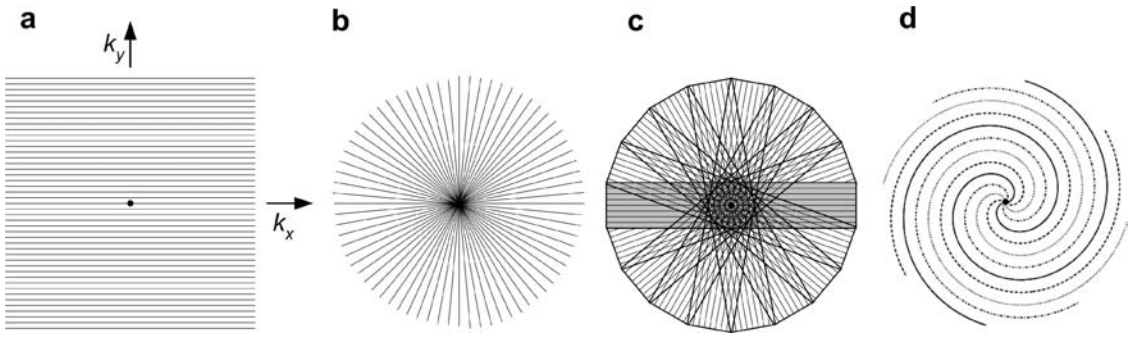


Figure 9. 2D k space sampling trajectories. Shown are the spin-warp (a), radial sampling (b), PROPELLER (c), and interleaved spiral imaging (d) examples.

measuring the flow of oxygenated blood in the brain (13–15). FMRI is based on the blood oxygen-level-dependent, or BOLD, effect. BOLD MRI is accomplished by first exposing a patient or volunteer to a stimulus or having them engage in a cognitive activity while acquiring single-shot images of their brain. The region of the brain that is responding to the stimulus or is engaged in the activity will experience an increase in metabolism. This metabolic increase will require additional oxygen. Therefore, an increase in oxygenated blood flow will occur (oxyhemoglobin) to the local brain area that is active. Oxyhemoglobin differs in its magnetic properties from deoxyhemoglobin. Oxyhemoglobin is diamagnetic like water and cellular tissue. Deoxyhemoglobin is more paramagnetic than tissue, so it produces a stronger MR interaction. These differences between oxyhemoglobin and deoxyhemoglobin in BOLD imaging are exploited by acquiring images during an "active" state (more oxyhemoglobin) and in a "resting" state (more deoxyhemoglobin), which creates a signal increase in the "active" state and a signal decrease in the resting state. Figure 10 shows a typical BOLD time course (shown in black) where four "active" states and four "resting" states exist. With prior knowledge of the activation timing (shown in red), we can perform a statistical test on the data to determine which areas of the brain are active. This statistical map (shown in color) is superimposed on a high resolution MR image so that one can visualize the functional information in relation to relevant anatomical landmarks.

Diffusion Imaging. The random motion of water molecules may cause the MRI signal intensity to decrease. The NMR signal attenuation from molecular diffusion was first observed more than a half century ago by Hahn (1950) (16).

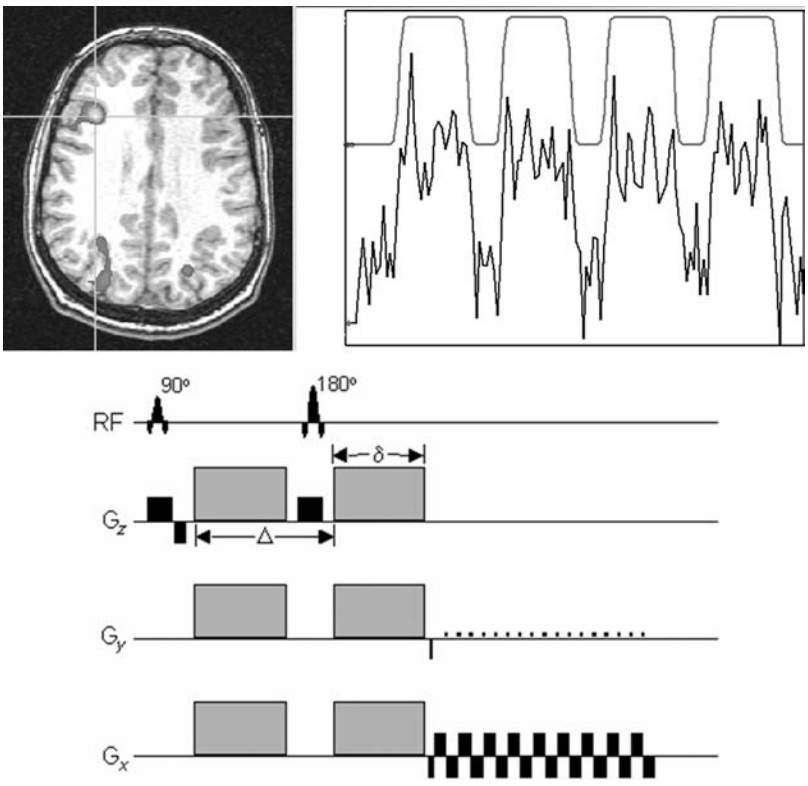


Figure 10. Color brain activation map is superimposed on high resolution MR image. Signal levels of the activated pixels are shown to increase during cognitive activity periods, whereas they fall off during periods of rest.

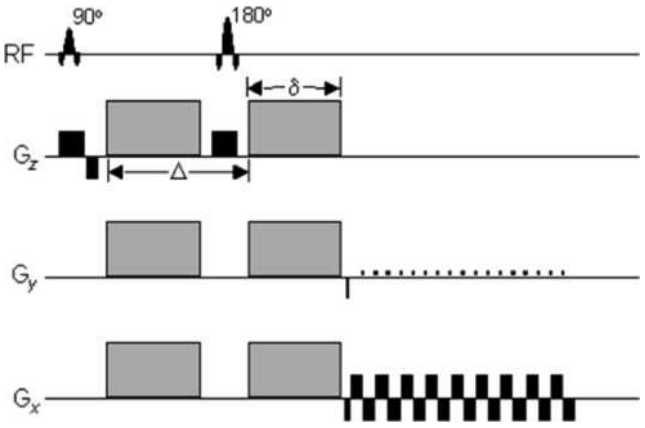


Figure 11. Temporal schematic of a diffusion-weighted, spin-echo pulse sequence with an EPI readout. The diffusion gradient pulses are shown as gray boxes on the gradient axes. The direction of diffusion-weighting can be changed by changing the relative weights of the diffusion gradients along G_x, G_y, and G_z.

Subsequently Stejskal and Tanner (1965) described the NMR signal attenuation in the presence of field gradients (17). More recently, field gradient pulses have been used to create diffusion-weighted MR images (18).

Diffusion-Weighted Pulse Sequences. Typically, diffusion-weighting is performed using two gradient pulses with equal magnitude and duration on each side of a 180° refocusing pulse, as shown in Fig. 11. The first gradient pulse dephases the magnetization as a function of position, and the second pulse rephases the magnetization. For stationary (e.g., no flow or diffusion) molecules, the phases induced by both gradient pulses will completely cancel, no signal attenuation will occur. In the case of motion in the direction of the applied gradient, a net phase difference will occur, $\Delta\phi = \gamma v G \delta \Delta$, which is proportional to the velocity v, the area of the gradient pulses defined by the amplitude G, and the duration δ, and the spacing between the pulses Δ. For the case of diffusion, the water molecules are also moving, but in arbitrary directions and with variable effective velocities. Thus, in the presence of diffusion gradients, the signal from each diffusing molecule will accumulate a different amount of phase, which, after summing over a voxel, will cause signal attenuation. For simple isotropic Gaussian diffusion, the signal attenuation for the diffusion gradient pulses in Fig. 11 is described by $S = S_o e^{-b\mathrm{D}}$ where S is the diffusion-weighted signal, S_o is the signal without any diffusion-weighting gradients (but otherwise identical imaging parameters), D is the apparent diffusion coefficient, and b is the diffusion-weighting described by the properties of the pulse pair $b = (\gamma G \delta)^2(\Delta\text{-}\delta/3)$.

Diffusion Tensor Imaging. The diffusion of water in fibrous tissues (e.g., white matter, nerves, and muscle) is anisotropic, which means the diffusion properties change as a function of direction. A convenient mathematical model of anisotropic diffusion is using the diffusion tensor (19), which uses a 3×3 matrix to describe diffusion using a general 3D multivariate normal distribution. The diffusion tensor matrix describes the magnitude, anisotropy, and orientation of the diffusion distribution. In a diffusion tensor imaging (DTI) experiment, six or more diffusion-weighted images are acquired along noncollinear diffusion gradient directions. Maps of the apparent diffusivity for each encoding direction are calculated by comparing the signal in an image without diffusion-weighting and the signal with diffusion-weighting. The diffusion tensor may then be estimated for each voxel, and maps of the mean diffusion, anisotropy, and orientation may be constructed, as shown in Fig. 12.

The primary clinical applications of diffusion-weighted imaging and DTI are ischemic stroke (20,21) and mapping the white matter anatomy relative to brain tumors and other lesions (22). DTI is also highly sensitive to subtle changes in tissue microstructure and, therefore, has become a popular tool for investigating changes or differences in the microstructure as a function of brain development and aging, as well as disease.

Vascular Imaging. Magnetic Resonance Angiography (MRA) describes a series of techniques that can be used to image vascular morphology and provide quantitative blood flow information in high detail. Two widely used techniques, phase contrast angiography and time-of-flight angiography, use the inherent properties of blood flow in the MR environment to create angiograms. A third technique, contrast-enhanced angiography, uses the injection of a

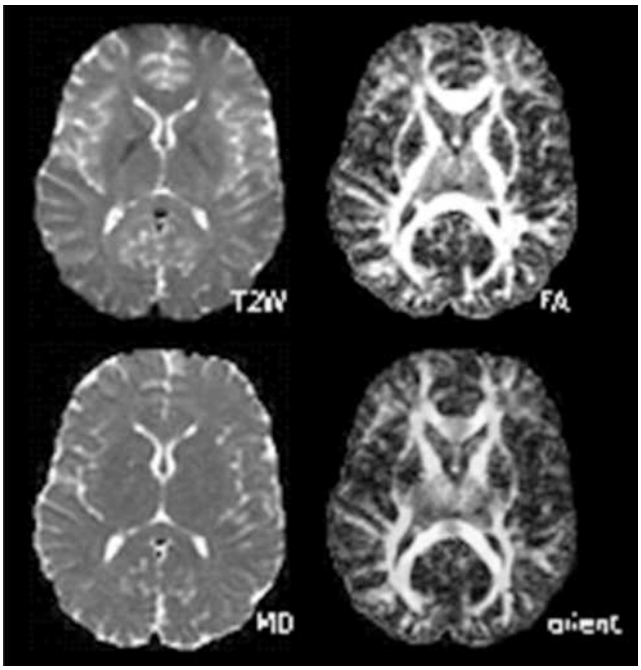


Figure 12. Representative diffusion tensor images. The images are (top-left): a T_2-weighted (or nondiffusion-weighted) image; (bottom-left): a mean diffusivity map (note similar contrast to T_2-weighted image with cerebral spinal fluid appearing hyperintense); (top-right): a fractional anisotropy map (hyperintense in white matter); and (bottom-right) the major eigenvector direction indicated by color (red = R/L, green = A/P, blue = S/I) weighted by the anisotropy (note that specific tract groups can be readily identified).

paramagnetic contrast agent into the vascular system to specifically alter the magnetic properties of the blood in relation to the surrounding tissue.

Phase-contrast (PC) angiography (23) usually uses a pair of gradient pulses of equal strength and opposite polarity, placed in the MRI sequence between the RF excitation pulse and the data acquisition window. During the imaging sequence, stationary nuclei accumulate phase during the first gradient pulse, and accumulate the opposite phase during the second gradient pulse, resulting in zero net phase. Moving nuclei accumulate phase during the first gradient pulse, but during the second pulse are in different positions, and accumulate phase different from that obtained during the first pulse. The net accumulated phase is proportional to the strength of the gradient pulses and the velocity of the nuclei. From the resulting data, images can be formed of both blood vessel morphology and blood flow.

TOF angiography techniques (24) (more accurately called "inflow" techniques) typically use a conventional gradient-echo sequence to acquire a thin 3D volume or a series of 2D slices. The nuclei in stationary tissue are excited by many consecutive slice-selective RF pulses. As a short TR is used, the longitudinal magnetization is not able to return to equilibrium, resulting in saturation of magnetization and low signal. Moving nuclei in the blood flow into the slice during each TR period, having been excited by zero or very few RF pulses. As these nuclei arrive in the imaging slice at or near full equilibrium magnetization, high signal is obtained from blood. Figure 13 shows a projection image of a 3D TOF dataset acquired in the head. The TOF technique can produce high quality MRA images in many situations, but slow or in-plane blood flow can result in blood signal saturation and reduced the quality of the images.

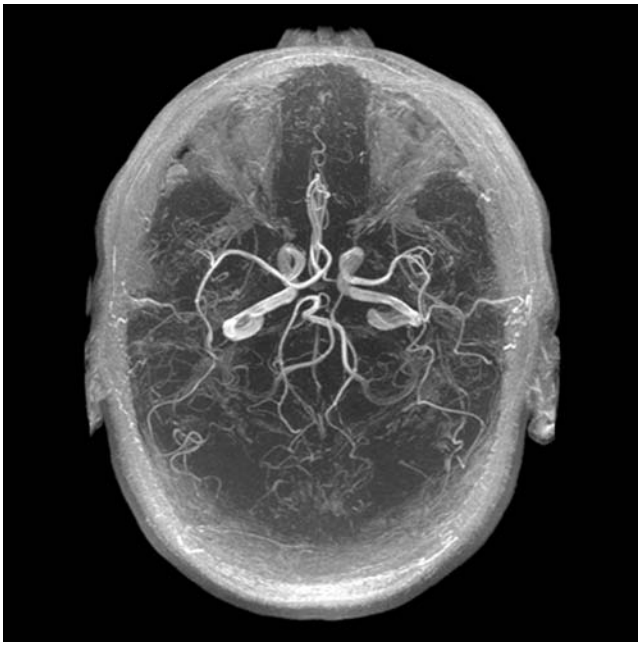


Figure 13. Time-of-flight (TOF) angiography in the head uses inflow of fresh blood to produce contrast between blood and the surrounding tissue. A Maximum Intensity Projection (MIP) reformatted image is used to compress the acquired volume data into a single slice for display.

Contrast-enhanced MRA (CE-MRA) is performed using an injection of a paramagnetic contrast agent into the intravenous bloodstream (25). Although several transition and rare-earth metal ions can be used, the most common is Gadolinium (Gd^{+3}) chelated to a biologically compatible molecule. The compound is paramagnetic, having a strong dipole moment and generating strong local magnetic field perturbations, which increase the transfer of energy between the excited hydrogen nuclei and the lattice, promoting T_1 relaxation and return to equilibrium of the longitudinal magnetization.

The contrast agent is injected intravenously in a limb away from the area of interest and circulates into the arterial system. The longitudinal relaxation rate is typically enhanced by a factor of 15 to 25 during this initial arterial phase, resulting in a much shorter T_1 for blood compared with the surrounding tissue. As the longitudinal magnetization in blood is much higher after each TR period, background tissue is suppressed in a manner similar to TOF imaging, and blood vessels have a comparably bright signal on the resulting images. Imaging is typically performed so that the central k space lines, which contain most of the image contrast information, are acquired while the contrast agent is distributed in the arteries of interest, but before it can circulate into the neighboring veins.

MRA data consist of large volumetric sets of image data, which are stored in the format of contiguous image slices. Specialized image display techniques are used to display the data in a manner that can be interpreted by the radiologist. Maximum Intensity Pixel (MIP) projections are widely used and are formed by projecting the volume set of data onto a single image plane. Here, each image pixel is obtained as the maximum value along the corresponding projection, as shown in Fig. 13. Volume rendering is beginning to be used more often to display MR angiograms. The individual slices of data are always available for detailed review by the radiologist and can be reformatted into any plane on the computer workstation to optimally display the vasculature of interest.

Cardiac MRI. Cardiac magnetic resonance (CMR) imaging is an evolving technique with the unprecedented ability to depict both detailed anatomy and detailed function of the myocardium with high spatial and temporal resolution. The past decade has seen tremendous development of phased array coil technology, ultra-fast imaging sequences, and parallel imaging techniques, all of which have facilitated ultra-fast imaging methods capable of capturing cardiac motion during breath-holding. The ability to perform imaging in arbitrary oblique planes, the lack of ionizing radiation, and the excellent soft tissue contrast of MR make it an ideal method for cardiac imaging. Comprehensive cardiac imaging is performed routinely in both in-patient and out-patient settings across the country and is widely considered the gold standard for clinical evaluation of many cardiac diseases (26).

Ischemic heart disease caused by atherosclerotic coronary artery disease (CAD) is the leading cause of mortality,

morbidity, and disability in the United States, with over 7 million myocardial infarctions and 1 million deaths every year (27). Consequently, ischemic heart disease is the primary indication for CMR. Accurate visualization of wall thickness and global function (ejection fraction), as well as focal wall motion abnormalities, is performed with retrospectively ECG-gated ultra-fast short TR pulse sequences, especially steady-state gradient recalled echo imaging (28), as shown in Fig. 14. Breath-held cinemagraphic or CINE images have high SNR, excellent blood to myocardial contrast, and excellent temporal resolution (< 40–50 ms) capable of detecting subtle wall motion abnormalities. Areas of myocardial infarction (nonviable tissue) are exquisitely depicted with inversion recovery (IR) RF-spoiled gradient echo imaging, acquired 10–20 minutes after intravenous injection of gadolinium contrast (29), as shown in Fig. 14. Areas of normal myocardium appear dark, whereas regions of nonviable myocardium appear bright (delayed hyper-enhancement). Accurate depiction of subtle myocardial infarction is possible because of good spatial resolution across the heart wall. The combination of motion and viability imaging is a powerful combination. Areas with wall motion abnormalities but without delayed hyper-enhancement may be injured or under-perfused from a critical coronary artery stenosis but are viable and may benefit from revascularization.

Cardiac "stress testing" using CMR has seen increasing use for the evaluation of hemodynamically significant coronary artery stenoses (30). Imaging of the heart during the first pass of a contrast bolus injection using rapid T_1-weighted RF-spoiled gradient echo sequences is a highly sensitive method for the detection of alterations

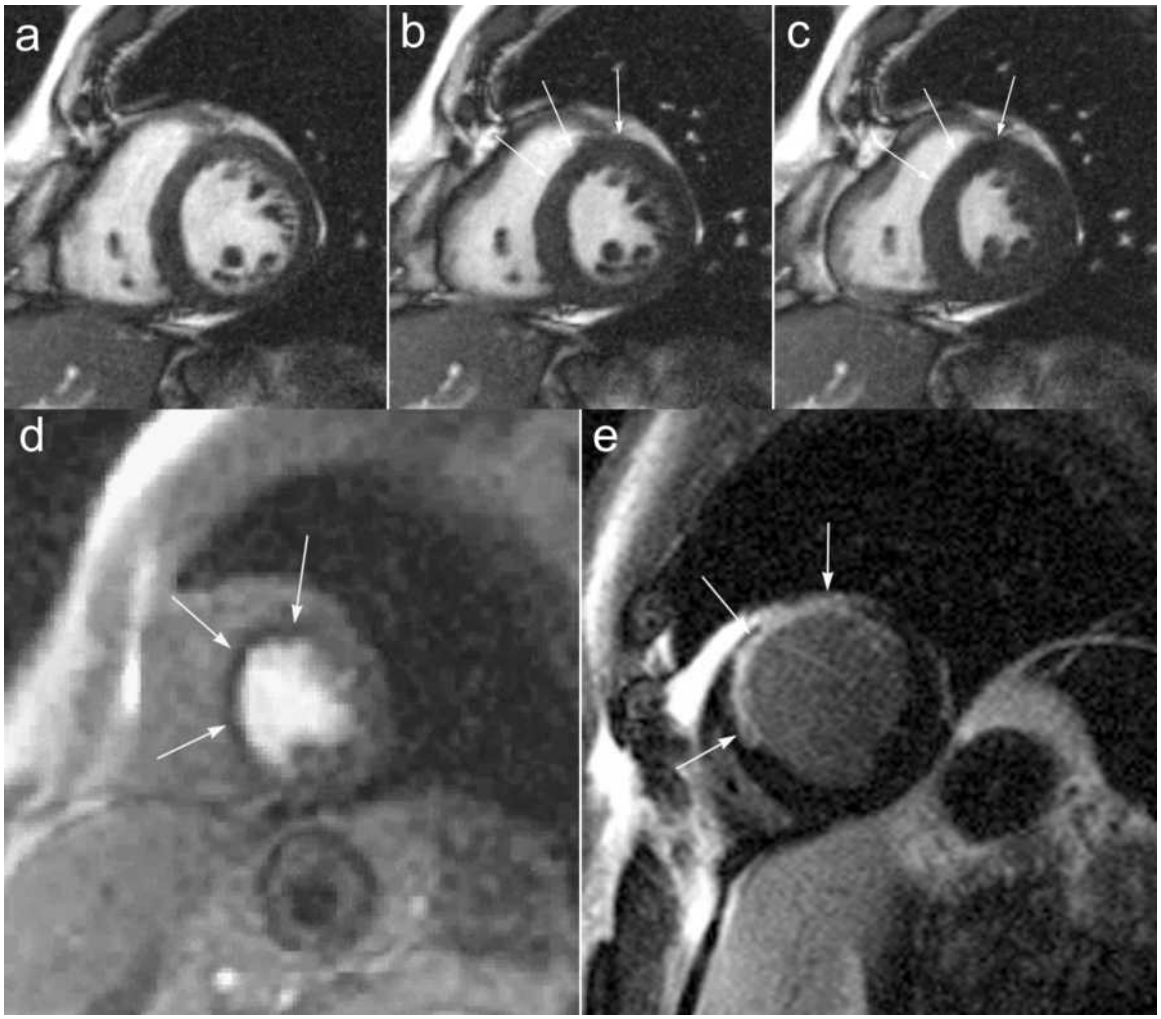


Figure 14. End-diastolic (a), mid-systolic (b), and end-systolic (c) short axis CINE images of the heart in a patient with a myocardial infarction in the anterior wall and septum, demonstrated by decreased wall thickening (arrows in b, c). The corresponding T_1-weighted RF-spoiled gradient recalled echo first-pass perfusion image (d) shows a fixed perfusion deficit (darker myocardium) in the corresponding territory (arrows). Finally, an inversion recovery RF-spoiled gradient echo image acquired at the same location (e) demonstrates a large region of delayed hyper-enhancement (arrows) indicating a full wall thickness region of nonviable myocardium that corresponds to the region of decreased perfusion and decreased contraction.

in myocardial blood flow (perfusion). Perfusion imaging during both stress (pharmacologically induced) and rest can reveal "reversible" perfusion defects that reflect a relative lack of perfusion during stress. In this way, coronary "reserve" can be evaluated and CAD can be uncovered, leading to further evaluation with coronary catheterization and possible angioplasty and stenting. Direct imaging of coronary arteries with CMR has shown tremendous technical advances, but is not commonly used, except for imaging of proximal coronary arteries in the evaluation of anomalous coronary arteries.

Other important indications of CMR include congenital heart disease, primarily, but not exclusively, in the pediatric population (31). Accurate diagnosis of a wide variety of congenital abnormalities requires high resolution, high contrast imaging that permits depiction of complex anatomical variants seen with congenital heart disease. Although anatomic imaging can be performed accurately with cardiac-gated CINE sequences, conventional sequences such as cardiac-gated black-blood fast spin-echo (FSE) and T_1-weighted spin-echo imaging are invaluable tools. Equally important to accurate anatomical imaging is functional imaging. With altered anatomy comes radically altered hemodynamics, requiring visualization of myocardial function with CINE imaging. Phase-contrast velocity imaging permits flow quantification through the heart, including the great vessels (pulmonary artery, aorta, etc.). An important example includes quantification of left-to-right "shunts" with resulting over-circulation of the pulmonary circulation. With a wide variety of pulse sequences, flexible scan plane prescription and the lack of ionizing radiation, CMR is ideally suited for evaluation of congenital heart disease.

Other important applications of CMR include visualization of valvular disease, pericardial disease, valvular disease, and cardiac masses. The latter two are particularly well evaluated with CMR; however, they are relatively uncommon and will not be discussed here.

Hyperpolarized Contrast Agents in MRI. Conventional MR imaging measures the resonant signal from the hydrogen nuclei of water, the most ubiquitous and highly concentrated component of the body. However, many other nuclei exist with magnetic dipole moments that produce MR signals. Many of these nuclei, such as phosphorous-31 and sodium-23, are biologically important in disease processes. However, these species typically exist at a very low concentration in the body, making them difficult to image with sufficient signal. One approach is to align, or polarize, the nuclei preferentially using physical processes other than the intrinsic magnetic field of the MR scanner. In some cases, these polarization processes can align many more nuclei than otherwise possible. These hyperpolarized nuclei can then act as contrast agents to better visualize blood vessels or lung airways on MRI. For example, helium-3 and xenon-129 are inert gases whose magnetic dipole moments can be hyperpolarized using spin-exchange optical pumping—a method of generating a preferred alignment of the nuclear dipoles using polarized laser light (32). As they are inert gases, polarized helium-3 and xenon-129 are used as inhaled contrast agents for visualizing the lung airspaces (upper-right panel) using MRI (33),(34). Unlike other parts of the body, conventional MRI of the lungs suffers from poor signal due to low water proton density and the multiple air-tissue interfaces that further degrade the MR signal in the upper left of Fig. 15. Hyperpolarized gas MRI has been particularly useful for depicting airway obstruction in several lung diseases including asthma (lower panel of Fig. 15) (35), emphysema (36), and cystic fibrosis (37). Additional techniques based on this technology show promise for MR imaging of blood vessels using injected xenon-129 dissolved in lipid emulsion (38), gas-filled microvesicles (39) and liquid-polarized carbon-13 (40). Hyperpolarized carbon-13 agents are of particular interest because of the wide range of biologically active carbon compounds in the body. Another important advantage of this technology is its ability to maintain high signal using low magnetic field (0.1–0.5 T) scanners (41). These systems are much cheaper to purchase and maintain than the high field (1.5–3.0 T) MRI scanners in common clinical use today.

Breast MRI. Breast MRI is presently used as an adjunct to mammography and ultrasound for the detection and diagnosis of breast cancer. Dynamic contrast-enhanced (DCE) MRI has been shown to have high sensitivity (83%–96%) to breast cancer but has also demonstrated variable levels of specificity (37–89%) (42). DCE-MRI requires an injection of a contrast agent and acquisition of a subsequent series of images to enable the analysis of the time course of contrast uptake in suspect lesions. Lesion morphology is also important in discerning benign from malignant lesions in breast MRI. Standard in-plane spatial resolution is sub-millimeter. A typical clinical breast MRI includes a spoiled gradient echo (SPGR) T_1-weighted sequence both precontrast (Fig. 16a) and, at minimum, at 30 second intervals postcontrast (Fig. 16b). Along with their morphologic characteristics, lesions can be further described by the shape of their contrast uptake curve. The three general categories of contrast uptake are (*1*) slow, constant contrast uptake (*2*) rapid uptake and subsequent plateau of contrast, and (*3*) rapid uptake and rapid washout of contrast (43). Although the slowly enhancing lesions are usually benign and fast uptake and washout is a strong indication of malignancy, time course lesion characterization is not absolute. The ambiguity of time course data for certain classes of lesions drives the investigation into higher temporal resolution imaging methods. A standard clinical breast MRI also includes acquisition of a T_2-weighted sequence for the identification of cysts (Fig. 16c). Present research in breast DCE-MRI is focused on development and application of pulse sequences that provide high temporal and spatial resolution. Also, investigation is ongoing into more specific characterization of uptake curves. Diffusion-weighted MRI, blood-oxygen-level-dependent imaging, and spectroscopy are also being investigated as possible methods to improve the specificity of DCE-MRI in the breast. In some circumstances, the high sensitivity of breast DCE-MRI outweighs the variable specificity leading to the present use of DCE-MRI to determine the extent of disease, with equivocal mammographic findings, and for the screening of high risk women.

Figure 15. Upper left shows normal lack of signal in parenchyma of lungs in MRI. Ventilated areas are clearly seen after imaging inhaled hyperpolarized helium. Rapid imaging during inhalation and exhalation shows promise for capturing dynamic breathing processes.

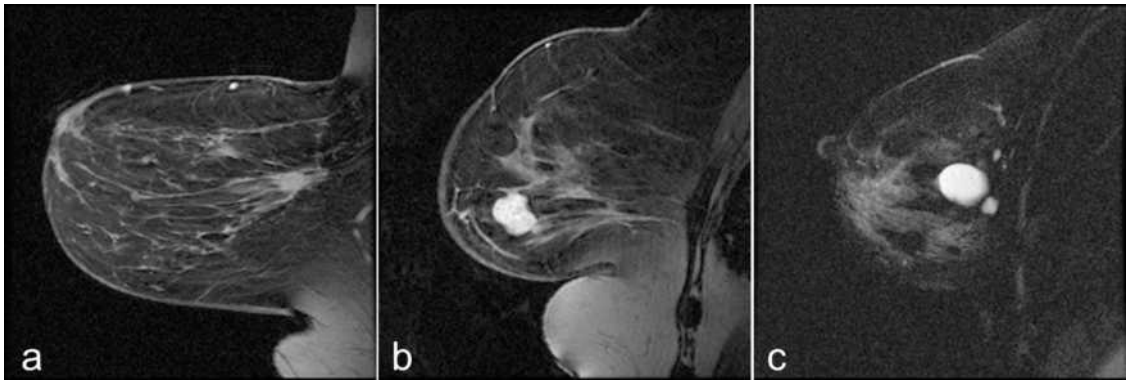


Figure 16. Fat-suppressed (a) pre-contrast T_1-weighted image (b) postcontrast T_1-weighted image (c), and noncontrast T_2- weighted image. Images are from different patients.

MRI of Musculoskeletal Disease. Musculoskeletal imaging studies traumatic injury, degenerative changes, tumors, and inflammatory conditions of the bones, tendons, ligaments, muscles, cartilage, and other structures of joints. Although X-ray imaging is the work-horse imaging modality for many musculoskeletal diseases, MRI plays a critical role in several aspects of diagnosis, staging, and treatment monitoring.

Fast spin-echo (FSE) pulse sequences are typically used to acquire T_1, T_2, and proton density-weighted images of the joints. For the assessment of joint structures, images are acquired in multiple planes to ensure adequate spatial resolution in all dimensions. These MR images can be used to evaluate tissues including ligaments, bone, cartilage, meniscus, and labrum. As a result of their high spatial resolution and excellent soft tissue contrast, MR images can provide accurate diagnosis that can prevent unnecessary surgeries and can facilitate pre-operative planning when surgical intervention is required.

Osteoarthritis is a degenerative disease that affects approximately 20 million Americans and countless others around the world. Currently, this debilitating disease is often not detected until the patient experiences pain that can be reflective of morphologic changes to joint cartilage. MR can be used to accurately measure cartilage thickness and volume. New MR techniques are also under development that may provide insight into biochemical changes in cartilage at the earlier stages of osteoarthritis that precede gross morphologic changes and patient pain.

Fortunately, primary bone tumors are relatively rare. However, bone is a common site for metastatic disease, which is especially true for breast, lung, prostate, kidney, and thyroid cancers. MR is a sensitive test for metastatic bone disease and is being adopted as a standard of care in some parts of the world, replacing nuclear scintigraphy. A typical approach employs inversion recovery pulse sequences to generate fat-suppressed, T_2-weighted images. Diffusion-weighted imaging also shows promise to detect hemotalogic cancers such as multiple myeloma, leukemia, and lymphoma.

Inflammatory diseases include infection and inflammatory forms of arthritis. Infection of the foot is a common complication of microvascular disease often seen with diabetes, a disease afflicting 18 million Americans. MR can be used to assess the vasculature of the foot as well as diagnose infection and evaluate treatment efficacy. Two million Americans have rheumatoid arthritis, a common form of inflammatory arthritis. Inflammation from a condition known as synovitis, which often occurs in rheumatoid arthritis patients, is shown in Fig. 17. New MR methods

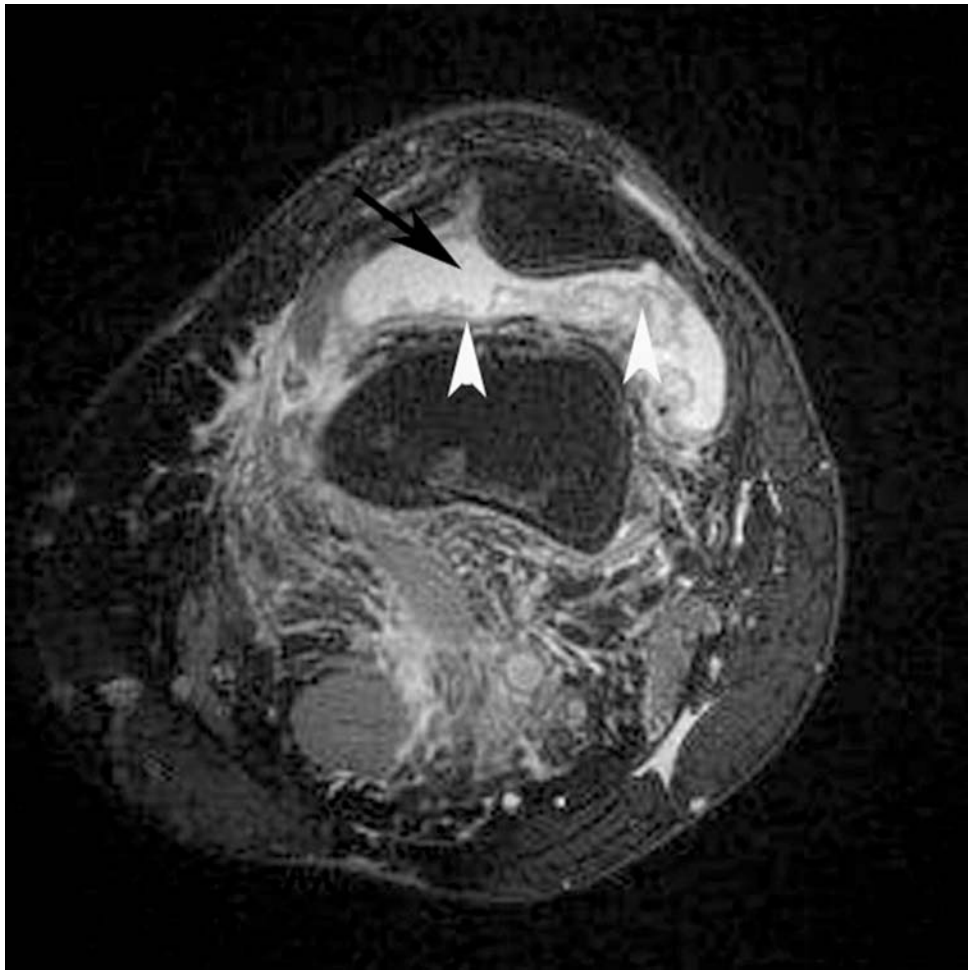


Figure 17. Axial knee image in a patient with inflamed synovium shows excellent soft tissue contrast available in MRI. The thickened intermediate signal intensity for synovium (arrowheads) is well distinguished from the adjacent high signal intensity of joint fluid (arrow).

are being developed to detect rheumatoid arthritis earlier and to gauge treatment success.

BIBLIOGRAPHY

1. Liang Z-P. Lauterbur PC. In: Akay M, ed. Principles of Magnetic Resonance Imaging—A Signal Processing Perspective. IEEE Press Series in Biomedical Engineering. New York: IEEE Press; 2000. p 416.
2. Abragam A. Principles of Nuclear Magnetism. Oxford, UK: Oxford University Press; 1994.
3. Haacke EM, et al. Magnetic Resonance Imaging-Physical Principles and Sequence Design. New York: Wiley; 1999. p 914.
4. Bottomley PA, et al. A review of H-1 nuclear-magnetic-resonance relaxation in pathology—are T1 and T2 diagnostic. Med Phys 1987;14(1):1–37.
5. Kingsley PB. Methods of measuring spin-lattice (T-1) relaxation times: An annotated bibliography. Concepts Magn Reson 1999;11(4):243–276.
6. Scheffler K. A pictorial description of steady-states in rapid magnetic resonance imaging. Concepts Magn Reson 1999; 11(5):291–304.
7. Hennig J, Nauerth A, Friedburg H. RARE imaging: A fast imaging method for clinical MR. Magn Reson Med 1986; 3(6):823–833.
8. Mansfield P. Multi-planar image formation using NMR spin echoes. J Phys C: Solid State Phys 1977;10:L55–L58.
9. Lauterbur PC. Image formations by induced local interactions: Examples employing nuclear magnetic resonance. Nature 1973;242:190–191.
10. Edelstein WA, et al. Spin warp NMR imaging and applications to human whole-body imaging. Phys Med Biol 1980;25(4):751–756.
11. Pipe JG. Motion correction with PROPELLER MRI: Application to head motion and free-breathing cardiac imaging. Magn Reson Med 1999;42(5):963–969.
12. Meyer CH. et al. Fast spiral coronary artery imaging. Magn Reson Med 1992;28(2):202–213.
13. Ogawa S, Lee T-M, Nayak A, Glynn P. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. Magn Reson Med 1990;14:68–78.
14. Ogawa S, et al. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proc Natl Acad Sci USA 1990;87:9868–9872.
15. Bandettini PA, et al. Time course EPI of human brain function during task activation. Magn Reson Med 1992;25(2): 390–397.
16. Hahn E. Spin echoes. Phys Rev 1950;80(4):580–594.
17. Stejskal E, Tanner J. Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. J Chem Phys 1965;42(1):288–292.
18. Le Bihan D, Breton E, Lallemand D, Grenier P, Cabanis E, Laval-Jeantet M. MR imaging of intravoxel incoherent motions: Application to diffusion and perfusion in neurologic disorders. Radiology 1986: 161(2):401–407.
19. Basser PJ, Mattiello J, LeBihan D. MR diffusion tensor spectroscopy and imaging. Biophys J 1994;66:259–267.
20. Moseley ME, Kucharczyk J, Mintorovitch J, Cohen Y, Kurhanewicz J, Derugin N, Asgari H, Norman D. Diffusion-weighted MR imaging of acute stroke: Correlation with T2-weighted and magnetic susceptibility-enhanced MR imaging in cats. AJNR Am J Neuroradiol 1990;11(3):423–429.
21. Warach S, Dashe JF, Edelman RR. Clinical outcome in ischemic stroke predicted by early diffusion-weighted and perfusion magnetic resonance imaging: A preliminary analysis. J Cereb Blood Flow Metab 1996;16(1):53–59.
22. Witwer BP, Moftakhar R, Hasan KM, Deshmukh P, Haughton V, Field A, Arfanakis K, Noyes J, Moritz CH, Meyerand ME, Rowley HA, Alexander AL, Badie B. Diffusion-tensor imaging of white matter tracts in patients with cerebral neoplasm. J Neurosurg 2002;97(3):568–575.
23. Dumoulin CL, Hart HR. Magnetic resonance angiography. Radiology 1986;161:717–720.
24. Keller PJ, Drayer BP, Fram EK, Williams KD, Dumoulin CL, Souza SP. MR angiography with two-dimensional acquisition and three-dimensional display. Radiology 1989;173:527–532.
25. Prince MR, Yucel EK, Kaufman JA, Harrison DC, Geller SC. Dynamic gadolinium-enhanced three-dimensional abdominal MR arteriography. J Magn Reson Imaging 1993;3: 877–881.
26. Gibbons RJ, Araoz PA. The year in cardiac imaging. J Am Coll Cardiol 2005;46(3):542–551.
27. Association AH, ed. 2005: Heart Disease and Stroke Statistics —2005 Update. Dallas, TX: A.H. Association; 2005.
28. Carr JC, Simonetti O, Bundy J, Li D, Pereles S, Finn JP. Cine MR angiography of the heart with segmented true fast imaging with steady-state precession. Radiology 2001;219(3):828–834.
29. Kim RJ, Wu E, Rafael A, Chen EL, Parker MA, Simonetti O, Klocke FJ, Bonow RO, Judd RM. The use of contrast-enhanced magnetic resonance imaging to identify reversible myocardial dysfunction. N Engl J Med 2000;343(20):1445–1453.
30. Ray T. Magnetic resonance imaging in the assessment of coronary artery disease. Curr Atheroscler Rep 2005;7(2): 108–114.
31. Rickers C, Kraitchman D, Fischer G, Kramer HH, Wilke N, Jerosch-Herold M, et al. Cardiovascular interventional MR imaging: A new road for therapy and repair in the heart. Magn Reson Imaging Clin N Am 2005;13(3):465–479.
32. Bouchiat M, Carver T, Varnum C. Nuclear polarization in 3He gas induced by optical pumping and dipolar exchange. Phys Rev Lett 1960;5:373–375.
33. Albert MS, Cates GD, Driehuys B, et al. Biological magnetic resonance imaging using laser-polarized 129Xe. Nature 1994;370:199–201.
34. van Beek E, Wild J, Kauczor H, Schreiber W, Mugler J, Lange E. Functional MRI of the lung using hyperpolarized 3-helium gas. J Mag Reson Imag 2004;20:540–554.
35. Samee S, Altes T, Powers P, et al. Imaging the lungs in asthmatic patients by using hyperpolarized helium-3 magnetic resonance: Assessment of response to methacholine and exercise challenge. J Allergy Clin Immunol 2003;111: 1205–1211.
36. Salerno M, Lange E, Altes T, Truwit J, Brookeman J, Mugler J. Emphysema: Hyperpolarized helium 3 diffusion MR imaging of the lungs compared with spirometric indexes—initial experience. Radiology 2002;222:252–260.
37. Altes TA, de Lange EE. Applications of hyperpolarized helium-3 gas magnetic resonance imaging in pediatric lung disease. Top Magn Reson Imaging 2003;14:231–236.
38. Moller HE, Chawla MS, Chen XJ, et al. Magnetic resonance angiography with hyperpolarized 129Xe dissolved in a lipid emulsion. Magn Reson Med 1999;41:1058–1064.
39. Callot V, Canet E, Brochot J, et al. MR perfusion imaging using encapsulated laser-polarized 3He. Magn Reson Med 2001;46:535–540.
40. Goldman M, Johannesson H, Axelsson O, Karlsson M. Hyperpolarization of 13C through order transfer from parahydrogen: A new contrast agent for MRI. Magn Reson Imag 2005;23:153–157.
41. Parra-Robles J, Cross AR, Santyr GE. Passive shimming of the fringe field of a superconducting magnet for ultra-low

field hyperpolarized noble gas MRI. J Magn Reson 2005;174: 116–124.
42. Kvistad KA, et al. Breast lesions: Evaluation with dynamic contrast-enhanced T1-weighted MR imaging and with T2*-weighted first-pass perfusion MR imaging. Radiology 2000; 216:545–553.
43. Kuhl CK, Schild HH. Dynamic interpretation of MRI of the breast. JMRI 2000;12:965–974.

MAGNETOCARDIOGRAPHY. See BIOMAGNETISM.

MANOMETRY, ANORECTAL. See ANORECTAL MANOMETRY.

MANOMETRY, ESOPHAGEAL. See ESOPHAGEAL MANOMETRY.

MAMMOGRAPHY

PEI-JAN PAUL LIN
Beth Israel Deaconess
Medical Center
Boston, Massachusetts

INTRODUCTION

It has been shown that reduced breast cancer mortality in the past decade can be attributed to the high sensitivity of screening mammography in detecting nonpalpable lesions (1,2). There has been a wide variety of equipment and imaging modalities employed in the breast cancer detection and imaging; ranging from ultrasound imager, X-ray mammography, computed tomography scanner (CT), to magnetic resonance imager (MRI). Additionally, thermography, light diaphanography, electron radiography, and microwave radiometry have also been utilized, experimentally, to detect breast cancer without much success. Brief explanations of these imaging modalities have been described in the first edition of this Encyclopedia, NCRP Report No.85, and in a review article by Jones (3–5). Among those modalities; ultrasound, X-ray mammography, CT, and MRI, X-ray mammography is the most practical and relatively inexpensive, and is the only main stream of equipment available for breast cancer detection. At present, the ultrasound imager is often employed as an adjunct to the X-ray mammography and is not the primary screening imaging device. However, when mammography is mentioned, it is normally meant to say "X-ray mammography". For these reasons, this article will devote all of its effort to X-ray mammography.

THE MAMMOGRAPHY QUALITY STANDARDS ACT OF 1992 (MQSA), (PUBLIC LAW 102–539)

Breast cancer was a major public health issue in the early 1990s; U.S. Congress enacted MQSA "to ensure that all women have access to quality mammography for the detection of breast cancer in its earliest, most treatable stages". Thus, it is required by law that facilities providing mammography services be properly accredited and be certified by the U.S. Food and Drug Administration (FDA). "Accreditation and Certification" of mammography facilities are beyond the scope of this article, and interested readers are requested to refer to the FDAs WEB Site (6), and the American College of Radiology (ACR) WEB Site (7).

X-RAY MAMMOGRAPHY

Conventional X-ray equipment was initially employed for breast cancer imaging with industrial (thick) emulsion film, or portal imaging film for use in radiation therapy as the image receptor in order to visualize the small microcalcifications, prior to, as late as 1970s. The breast entrance dose of this imaging process exceeded well over 85 mGy per film ($\sim$10 R per film) and the radiographic techniques were typically in the range of 45–55 kVp, and $\sim$1000 mAS with a radiation beam quality of half-value layer (HVL) = 1.0–1.5 mm of aluminum (mmAl) (8). The X-ray beam spectrum produced by a conventional X-ray tube, equipped with tungsten anode, is not necessarily optimized for breast cancer detection. The mammography images obtained in this manner had the desired spatial resolution ($\sim$20 lp·mm^{-1}), but had a less than desirable radiographic contrast. This combination of "high entrance dose" with "low radiographic contrast" was not an acceptable approach. The radiology community was searching for a new breast imaging solution. In the 1970s, xeromammography imaging plates provided the much needed improvement in image quality and lowered the breast entrance dose by a factor of two thirds to one half compared to using the thick emulsion industrial type film (9).

Due to its unique patient positioning of breast imaging, the geometrical arrangement of dedicated mammography units should be pointed out. As shown in Fig. 1, with the

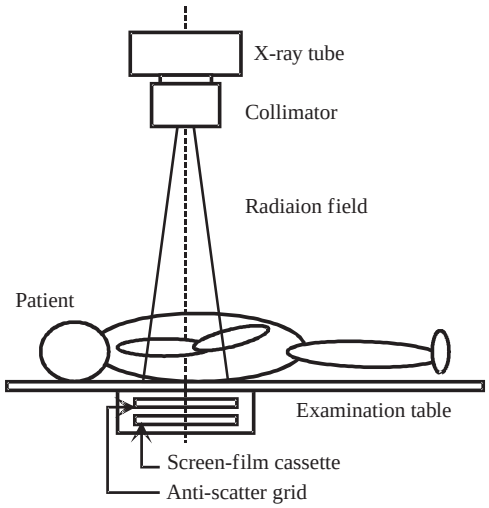
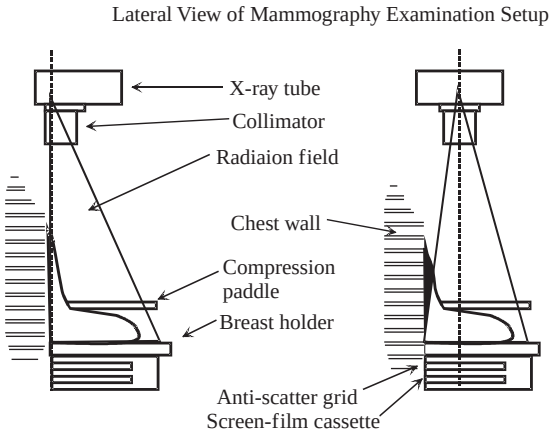


Figure 1. Geometric arrangement of conventional radiography. The X-ray system, the anatomy of interest, and the screen-film cassette are centered and aligned for exposure.



(a) Correct alignment of X-ray system with the patient's chest wall included in the field-of-view.

(b) Improper setup. The "black" triangular shaped area is not in the field-of-view.

Figure 2. Geometry of dedicated mammography systems. On the left side of (a), the geometrical arrangement of a dedicated mammography unit is correctly setup where the center of the radiation beam is aligned to the chest wall of the patient with the compression cone pressing down on the breast being imaged. On the right side (b), improper setup for mammography examination is evident. The "black" triangular area is not included in the image captured by the image receptor, potentially missing the suspected cancer site.

conventional radiography system, the X-ray tube is setup in the center of the anatomy of interest. The radiation field is a projection of diverging X rays restricted by the collimator, centered to the area of interest. The geometry of a dedicated mammography system on the other hand is offset so as to maximize inclusion of breast tissue close to the chest wall (see Fig. 2a). If the geometry were setup in the same manner as the conventional radiography as depicted in Fig. 2b, the black triangular area of the breast would not be included in the imaging field possibly missing a suspected cancer site.

THE PHYSICS OF MAMMOGRAPHIC IMAGING

Considering that various tissues in the breast are radiologically similar, but not identical, the task of differentiating the fibrous, ductal, and glandular tissues is extremely difficult. This is due to the fact that (1) water; the main ingredient of human tissue, has a density of $\rho = 1.0\,\mathrm{g{\cdot}cm^{-3}}$ and effective atomic number of $Z_{\mathrm{eff}} = 7.4$–7.6, and (2) fat has a density of $= 0.9\,\mathrm{g{\cdot}cm^{-3}}$ and effective atomic number of $Z_{\mathrm{eff}} = 5.9$–6.5 (10). Thus, various breast tissues with varying degrees of fat and water contents are very similar from radiological point of view.

The development of screen-film mammography was, in reality, paired with the redesigning of dedicated X-ray equipment for breast imaging. Breasts are relatively thin (physical thickness and X-ray attenuation property) compared to other body parts. Radiographs of extremities, for example, are obtained with X-ray tube potential in the range of 50–60 kVp. Breast tissues contain no high attenuation anatomy, such as bones, a lower tube potential (< 35 kVp) would be more suitable for breast imaging (11). Use of lower tube potential has a potential benefit of taking advantage of the photoelectric effect in differentiating the subtle differences of breast tissues.

It should be pointed out that there are five basic ways that X-ray photons interact with matter; they are (1) coherent scattering, (2) photoelectric effect, (3) Compton effect, (4) pair production, and (5) photodisintegration (12). Of these five interactions, photoelectric effect and Compton effect predominantly contribute to the image formation in diagnostic radiology. The probabilities of photoelectric effect and the Compton effect can be expressed as following;

$$\text{Photoelectric effect} \sim Z^3/E^3 \tag{1}$$

$$\text{Compton effect} \sim E * Z \tag{2}$$

In equations (1) and (2), E is the photon energy, and Z is the atomic number. Clearly, from Eq. 1, the lower the X-ray photon energy, and the higher the atomic number the probability of the photoelectric effect will be higher. Since the photoelectric effect is proportional to the "cube" of (Z/E), the differential of breast tissues in Z_{eff} is "enhanced" in the image formation. Thus, a better approach to differentiate and image the breast tissues is to employ a lower X-ray tube potential for imaging. While this approach is "good" for imaging, the radiation dose absorbed by the breast would still be a major concern.

Conventional X-ray tubes employ tungsten as the target material and produce a broad X-ray spectrum through the bremsstrahlung process. At X-ray tube potential of 35 kVp and lower, the tube potential used on most dedicated mammography systems, the tungsten target X-ray tubes would produce broad energy spectrum that contribute less to the image formation and more to the radiation dose. Tungsten, rhodium, and molybdenum are ideal for use in X-ray production due to their relatively high melting points than other metallic elements. Typically, dedicated mammography equipment is equipped with molybdenum target X-ray tube with beryllium window (port), and 30-μm thick molybdenum filter. The X rays, generated at 25–30 kVp tube potential, produced by the molybdenum target X-ray tube contain a large fraction of characteristic X rays at energies $\sim$17–20 keV (13,14). It is, therefore, quite natural to optimize and utilize these characteristic X rays for image formation and, consequently, patient exposure reduction at the same time. Most dedicated mammography equipment employ a combination of molybdenum target with aluminum, molybdenum, or rhodium filters. The characteristic X rays generated at the molybdenum target have X-ray photon energies of 17.4 and 19.6 keV, just below the K-absorption edge of 20 keV, thus are quite transparent to the molybdenum filter. This is illustrated in Fig. 3. On the left of Fig. 3a, is a schematic drawing of the X-ray spectrum generated at 30 kVp with a molybdenum target and aluminum filter. In the middle of Fig. 3b, is the same system with 30 μm thick molybdenum filter. Notice that the K-absorption edge curve (dashed curve) of molybdenum shows that the X-ray photons with energies just above

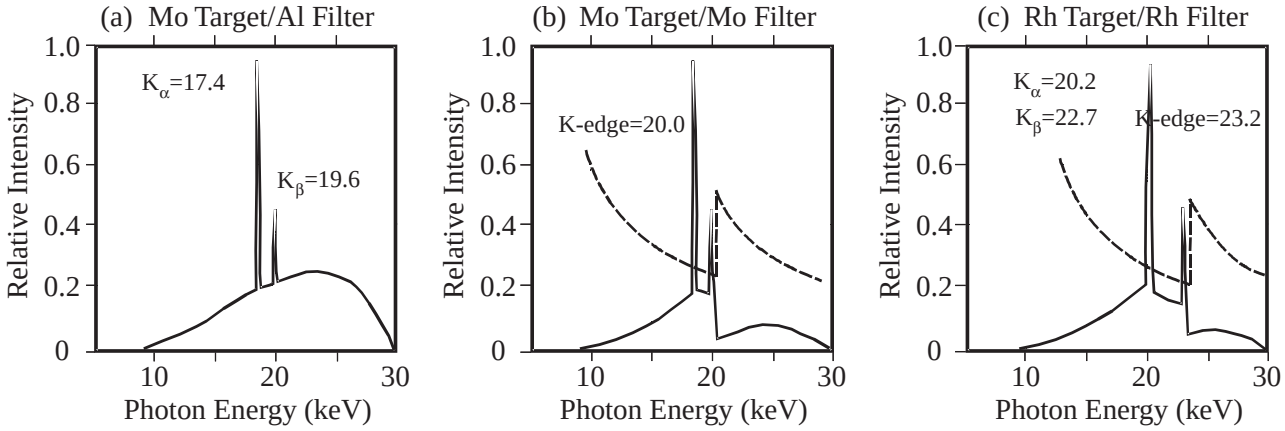


Figure 3. The spectrum of X rays generated with molybdenum and rhodium. On the far left (a) is the X-ray spectrum generated by a molybdenum target at 30 kVp and filtered with aluminum filter. The characteristic X rays K_α (17.4 keV), and K_β (19.6 keV) appear as the two peaks in the graph. In the middle (b) is the X-ray spectrum generated by the same X-ray system in part a, but is filtered with 30 μm thick molybdenum. The K-absorption edge curve is shown in dashed line. On the far right (c) is an X-ray spectrum generated by a rhodium target at 30 kVp and filtered with 20 μm thick rhodium. The general shapes of middle figure and far right figure are similar with differences in the peak energies of K-characteristic X rays ($K_\alpha = 20.2$ keV, and $K_\beta = 22.7$ keV), and the K-absorption edge energy (23.2 keV).

20 keV would be preferentially absorbed than those just below 20 keV. Similarly, the same can be said for the rhodium target with rhodium filter as shown in Fig. 3c. Figure 3b and c are graphically speaking quite similar. It is noteworthy to point out that a careful study of Fig. 3b, and c will reveal that X-ray beams generated from rhodium target X-ray tube "must not" be filtered with molybdenum! The intensity of rhodium K-characteristic X rays (K_α, and K_β) would be in the energy range where the molybdenum K absorption is high. Taking advantage of the spectral information, in 1991,GE introduced the Senographe DMR unit equipped with a dual track and dual filter (molybdenum and rhodium) X-ray tube for mammography applications. Some of the physical characteristics and the spectral energy data of molybdenum, rhodium, and tungsten are summarized in Table 1.

For illustration purposes and descriptions of imaging components, following this paragraph, the screen-film mammography (SFM) system manufactured by GE, the Senogaphe DMR mammography unit, is shown in Fig. 4. The photograph in Fig. 4 represents the overall external and mechanical design of a typical "dedicated" X-ray mammography unit. It represents "the overall" design with respect to the tilting gantry with the X-ray tube housing at the top and the image receptor at the bottom. And, the gantry is attached to a column (or stand), which houses the elevation mechanism of entire gantry.

Table 1. K-Characteristic X Rays and K-Absorption Edge of Molybdenum, Rhodium, and Tungsten

	Molybdenum	Rhodium	Tungsten
Atomic Number	42	45	74
K_α[a]	17.4	20.2	59.3
K_β[a]	19.6	22.7	69.1
K-edge[a]	20.0	23.7	69.5

[a]Energy in keV.

THE SCREEN-FILM MAMMOGRAPHY

The screen-film mammography employs the same basic image receptor system as conventional radiographic imaging. While conventional radiography employs a double-emulsion film sandwiched between two intensifying \ screens yielding a spatial resolution of (up to) 8 $lp \cdot mm^{-1}$ for a detailed screen (15), a typical SFM image receptor consists of a single-emulsion film and a single thin rare-earth phosphor intensifying screen yielding a spatial resolution of $\sim 20\ lp \cdot mm^{-1}$ (16).

In order to optimize the efficiency of the SFM, manufacturers including Agfa, Kodak, Konica, and Fuji, have produced matching pairs of the intensifying screen and film specifically for use with mammography. And, to further improve the sensitometric characteristics of the screen-film, the processing chemistry, particularly the developer, have also been carefully prescribed along with its development conditions. Note that the sensitometric characteristics of screen-film system refers to the "photographic effect or blackening effect" of the screen-film system in response to the X-ray absorption (17). An example of this matching pair of intensifying screen and film is depicted in Fig. 5, the emission and absorption

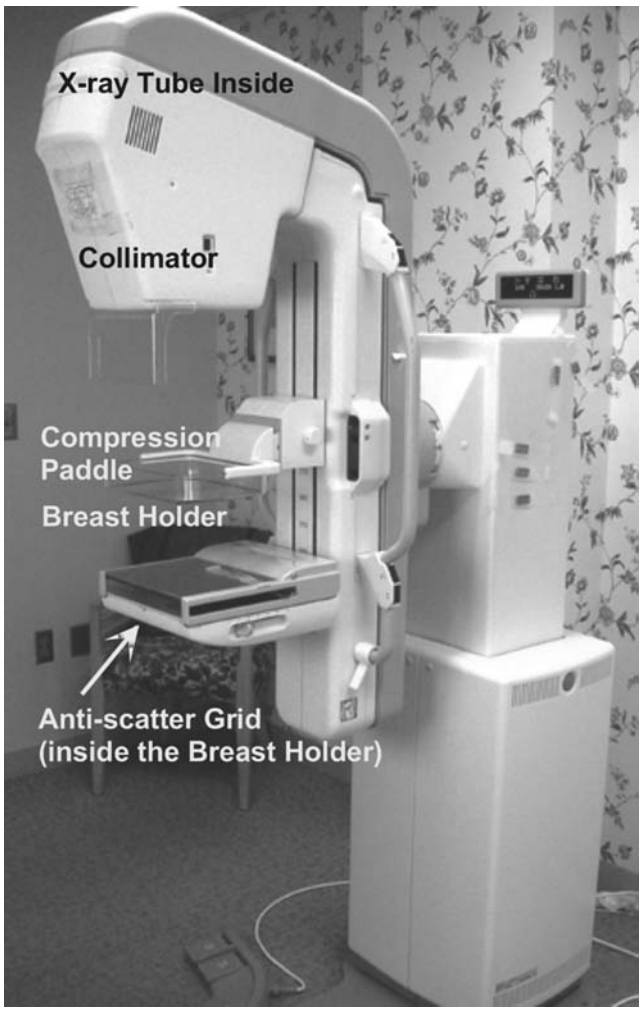


Figure 4. A typical dedicated mammography unit; GE Senographe DMR. (Courtesy of GE Healthcare.) The overall mechanical design of a typical dedicated X-ray mammography showing the tilting gantry with the X-ray tube housing at the top and the image receptor at the bottom. The gantry (or the elongated C arm) is normally attached to a column or a stand in which the elevation mechanism of entire gantry is housed.

characteristics of intensifying screen and film employed in SFM. Note that the emission of 550 nm wavelength light from AD Mammo Screens is matched by the absorption spectra of the AD-M Film.

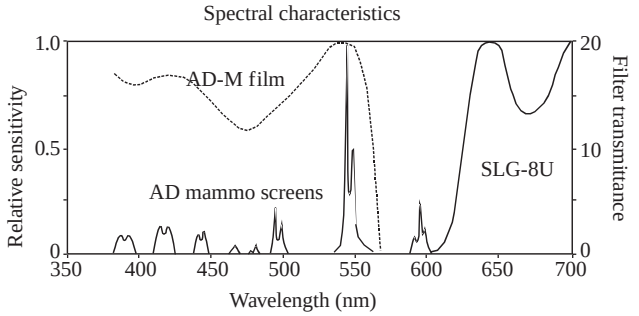


Figure 5. Spectral characteristics of screen-film mammography system. The AD Mammo screens are green-emitting and the AD-M film is orthochromatic. The figure shows the light-emitting spectrum of the AD Mammo Screens, the spectral sensitivity curve of the AD-M film, and the transmittance spectrum of SLG-8U safety light filter. (Courtesy of Fujifilm Medical Systems USA, Inc.)

In order to accommodate the varying size of breasts, two imaging cassette sizes are available with dedicated mammography equipment, they are 18×24 cm (8×10 in.), and 24×30 cm (10×12 in.). The film is loaded on the topside of the cassette with the emulsion side facing the intensifying screen. Figure 6 is a schematic drawing showing the cross-section of a typical screen-film mammography cassette (the single emulsion film, single intensifying screen) with enlarged views of the film (top, right), and the screen (bottom, right). The "sponge", in the cassette, is employed to assure good screen-film contact.

THE ANTISCATTER GRID

The antiscatter grid is placed between the breast holder, and the cassette slot, refer to Fig. 4. Scattered radiation is one of the main causes of degrading the image quality in X-ray imaging. The antiscatter grid is employed to minimize the scattered radiation from reaching the screen-film system while allowing the primary radiation to pass through. Although, mammography examinations are typically conducted with X-ray potentials <30 kVp, would still require a moving (reciprocating) antiscatter grid to cleanup the scattered radiation. Typically, the antiscatter grid used in mammography has a grid ratio of 5 : 1, or 4 : 1. While the exposure time in X-ray mammography imaging is relatively long ($\sim$1 s), the speed of the moving grid must be

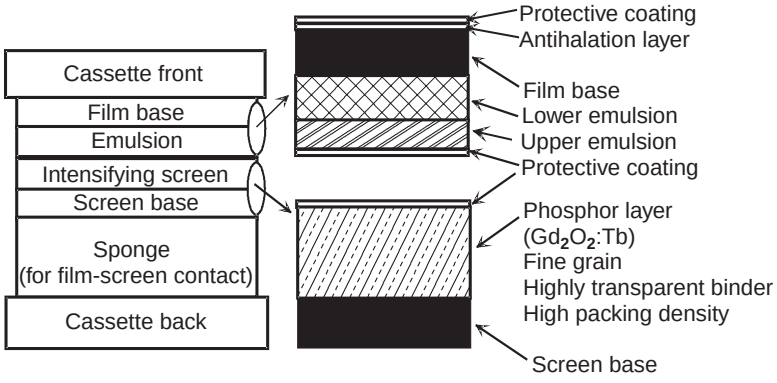


Figure 6. The cross-sectional view of screen-film cassette. (Courtesy of Fujifilm Medical Systems USA, Inc.)

carefully adjusted to avoid artifact associated with grids. For the antiscatter grids, the grid line rate must be sufficiently high, or the attenuation material structure must be so designed to avoid being imaged as grid artifacts on the mammograms. The honeycomb design antiscatter grid; Cellular (HTC) Grid utilized in LoRad mammography systems and marketed by Hologic is well known for its high transmission of useful primary radiation with increased absorption of scattered radiation (18).

THE COMPUTED RADIOGRAPHY FOR MAMMOGRAPHY

Computed radiography was introduced to radiology in 1981 and the "digitization" of routine radiographic examinations had started in earnest. The CR image plate housed in cassette replaced, directly, the screen-film cassette in routine radiography applications. This direct replacement design required no mechanical modifications on the existing radiographic equipment. With the introduction of CR, a whole new set of technology including the optical CR readers, image processing software programs, image display subsystems, and so on, made the filmless radiology department within an achievable reality (19).

The difference between the routine CR and the CR for mammography (CR-M) is largely due to the optical response of the CR phosphor; thus, the radiation dose required to reduce the image noise, and the spatial resolution capability for mammography applications. In 2001, Fuji introduced the FCR 5000MA, which was used in the America College of Radiology Imaging Network (ACRIN); Digital versus Screen-Film Mammography (DMIST) study (20,21). While the study had already been concluded, the official results are in the final preparation stage, and have not been released yet. The FCR 5000MA differentiated itself from the previous generation of CR products by utilizing a 50 μm pixel spot size for the laser and introduced dual side IP reading to the market.

In January 2004, Fuji introduced the Profect CS, or "ClearView-CS" in the United States. Clear View CS is pending FDA approval and not yet commercially available in the United States. Fuji recently finished the Premarketing Approval (PMA) application to FDA, and the approval is anticipated sometime during the second half of 2005 (22). Note, however, that the 100 μm pixel spot size digital mammography system has been in use for the past few years in Europe, Australia, and Japan. The mechanism in which "how" the IP and the CR reader work together to produce an image is beyond the scope of this article and readers are suggested to turn to publications available (23,24), or corresponding articles in this Encyclopedia.

The obvious advantage of CR-M, and the full field digital mammography (FFDM) systems, is its wide dynamic range and its linear response to radiation. In screen-film mammography, over-or underexposure was one of the main causes of "repeat" examinations. Such exposure related repeats can be minimized due to the wide latitude in exposure acceptable for imaging. Furthermore, the images are no longer "fixed" or "limited"; the subject contrast and the overall image brightness can be adjusted for image interpretation by adjusting the display window "level", and

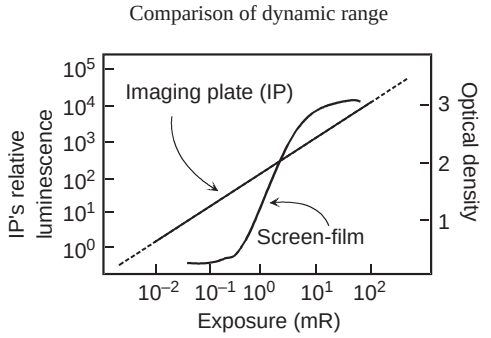


Figure 7. Comparison of dynamic range of CR and screen-film image receptors. (Courtesy of Fuzifilm Medical Systems, USA, Inc.) The image plate (IP) has wider linear response to radiation extends over four decades of dose and eliminates the inherent limitations of the "toe" and "shoulder" portions of the film's H&D Curve. (Courtesy of Fujifilm Medical Systems USA, Inc.)

"width". The display "window level" corresponds to the film optical density in SFM, and the "widow width" corresponds to the contrast. In digital imaging systems, both window level and width may be adjusted to optimize the image rendered on the monitor. Figure 7 shows the response differences of these two image receptor systems to the radiation exposure. The exposure range for the CRs "Image Plate" is wider than that of a typical screen-film combination by an order of 2 to 3 in magnitude.

THE STEREOTACTIC BREAST BIOPSY MAMMOGRAPHY

Breast biopsy is often required when an area of suspicious malignancy site is revealed after a mammography is obtained. In order to accurately extract the specimen, a stereo mammogram may be obtained so that the biopsy needle can be accurately inserted to the site where it is suspected of malignancy; core biopsy (25). Most dedicated mammography equipment are designed to perform routine (screen-film, or FFDM) mammography, and with an attachment to perform stereotactic mammography and core biopsy in an "up right" posture; either the patient is standing or sitting on a chair.

From a pair of stereo images separated by 30° (±15° from the centerline), using the triangulation technique, the three-dimensional (3D) coordinates of the suspicious site can be calculated within an accuracy of 1 mm (25), see the schematic diagram of stereotactic imaging geometry in Fig. 8. It is essential that the patient remain still before and after the acquisition of the stereo images. The core biopsy can be localized accurately only if the patient positioning remain the same.

Stereotactic mammography requires that the breast being examined is compressed (just as in routine mammography), but the patient should remain standing still while the mammogram is being processed. The processing time alone takes at least 90 s with typical automatic film processor. Thereafter, the best location of the biopsy needle entrant site, and the coordinates of the sampling must be determined. The prolonged time represents substantial

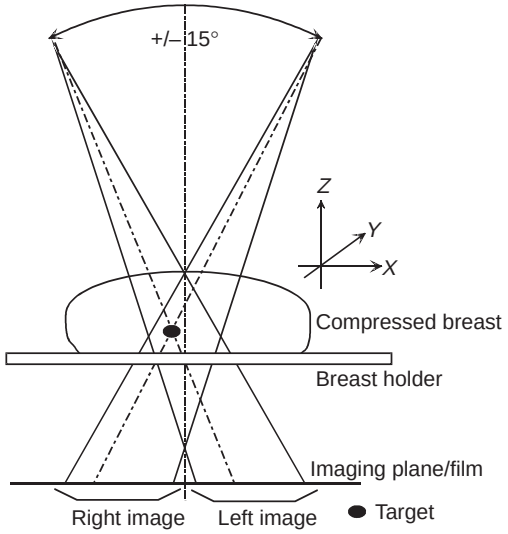


Figure 8. Schematic drawing of geometrical setup for stereotactic mammography. From the stereo images, with known geometry and the *X*–*Y* coordinates, the depth in *Z* direction is calculated via triangulation.

discomfort on the part of the patient not to mention the anxiety of fear for the possibility of being diagnosed and confirmed as having breast cancer.

DIGITAL SPOT MAMMOGRAPHY

Before the introduction of FFDM, in order to ease the discomfort of undergoing the biopsy process, a small detector with imaging area of 5×5 cm had been developed by LoRad, for example, so that the stereo images of the breast can be displayed immediately after exposures are made. This is one of the successful biomedical applications that have resulted from collaborative technology transfer programs between the National Aeronautics and Space Administration (NASA), the National Cancer Institute (NCI), and the U. S. Department of Health and Human Services Office on Women's Health (OWH) (26). Since the imaging area is only 5×5 cm, it is referred to as Digital Spot Mammography. The biopsy coordinate is then calculated via the built-in software program with a shorter overall examination time.

In addition, dedicated breast biopsy systems have been developed where the patient would be lying on her stomach (recumbent). The breast under examination is positioned through a hole in the examination table and aligned with the X-ray equipment and the biopsy needle gun. The recumbent core biopsy mammography unit manufactured by LoRad is depicted in Figs. 9 and 10. LoRad is now one of the brand names sold under Hologic. A similar recumbent core biopsy unit, called Mammo Test Biopsy System, is available from Fisher Imaging.

FULL FIELD DIGITAL MAMMOGRAPHY

The CR-M is a direct replacement of the screen-film cassette. In other words, the IP cassette replaced the screen-film cassette. Therefore, no major mammography equipment modification would be necessary. In DR on the other hand, just as in conventional radiography and fluoroscopy applications, the image receptor system may be built into the image receptor compartment as an integral part of the imaging assembly (27,28). Manufacturers had attempted to physically fit the DR detector assembly, or more accurately; the Flat Panel (FP) detector assembly, into the space occupied by the SFM cassette. To date, no commercial product of FP detector assembly has been fitted as a *direct physical replacement* of the SFM cassette is available (29). In other words, due to technical difficulties, there is no FP detector assembly that is sufficiently compact to fit into the space designed to accommodate the SFM cassette. The image acquired with FP detector is transmitted to and processed by the image processing chain thereafter.

While there are a hand full of FFDM systems either under development or manufactured for clinical applications, three FFDM units are currently available in the United States, and are described here (27). The fact that

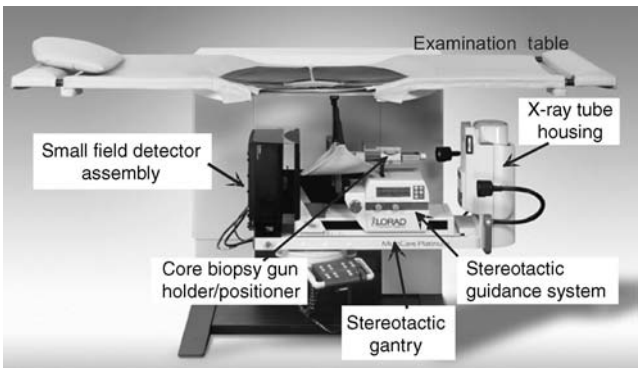


Figure 9. The recumbent stereotactic core biopsy system. A recumbent design provides a stable patient positioning to acquire a pair of stereo images for accurate localization of biopsy site while the patient lays flat on her stomach with comfort during the entire examination. (Courtesy of Hologic.)

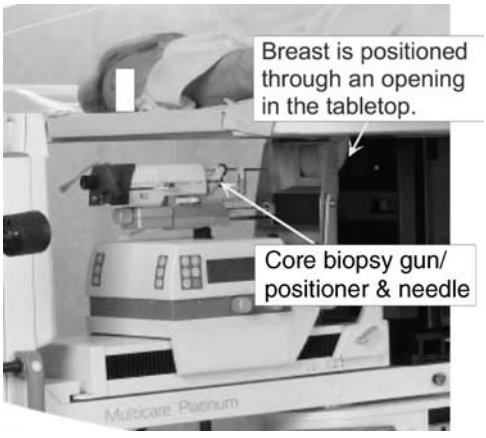


Figure 10. The zoomed view of the recumbent stereotactic core biopsy system. In the zoomed view, the breast under the examination is positioned through the opening in the tabletop. (Courtesy of Hologic.)

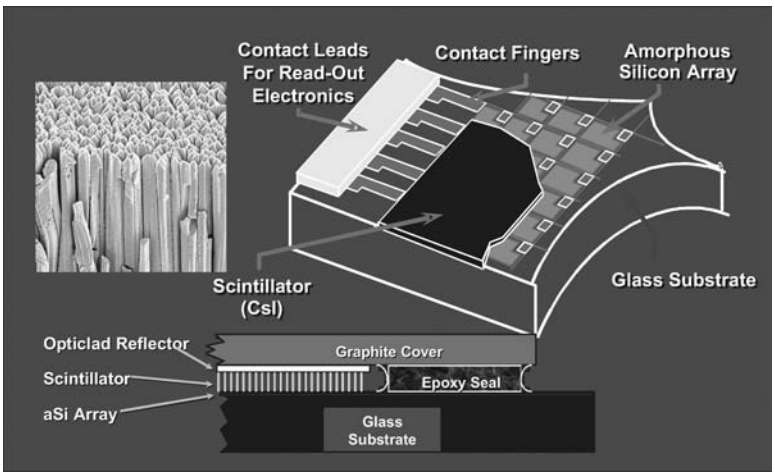


Figure 11. Schematic drawing of an amorphous silicon digital detector. (Courtesy of GE Healthcare.)

these three systems are designed to cover an imaging field of at least ($\sim$) 18×24 cm, the smaller cassette size of SFM, the name; FFDM was assigned. This is to signify the full size imaging filed coverage as opposed to the small field detector of LoRad's Digital Spot Mammography, or Fischer Imaging's Mammo Test Biopsy System.

There is a trend that mammography equipment is being transformed to "digital" format. One reason of the slow pace of this transformation is the large initial capital expenditure of FFDM. Both, analog and digital formats are expected to coexist for many more years to come. The impact of Fuji's CR-M cannot be ignored due to the fact that it is possible to convert and replace the screen-film cassette directly. Any of the existing SFM units will be able to join the "digital era".

Currently, there are three FFDM systems that have received FDA approval and are sold in the United States. The GE Senogaphe 2000D was approved by FDA on January 28, 2000 and "accredited" by ACR in February, 2003. Fischer Senoscan received its FDA approval on September 25, 2001 with subsequent ACR accreditation in August, 2003. Hologic/LoRad's Selenia received FDA approval and ACR accreditation, respectively, on October 2, 2002, and in September of 2003 (27).

GEs FFDM: SENOGRAPHE 2000D, AND SENOGRAPHE DS

Figure 11 depicts the FP detector developed by General Electric Medical Systems, and the photo inset (top left) shows the Cesium–Iodide (CsI) scintillator crystals. The incident X rays are absorbed by the CsI crystals, which in turn, convert the X-ray photon energy to light. The CsI crystals are deposited in columnar shape so as to minimize the scatter, and optical diffusion of scintillation lights from one column to the other. The underlying photodiode–transistor amorphous Silicon (a-Si) arrays is connected to control data lines, then, converts the light to electrical signals for further processing (30).

GEs FP detector is an indirect-conversion digital detector system. The detective quantum efficiency (DQE) of this FP detector system has been shown to be $\sim$60% at Zero Frequency (31,32). Essentially, being the first FFDM system introduced to the commercial market, GEs Senographe 2000D ushered in the digital mammography era to the radiology community. Depicted in Fig. 12 is the second-generation FFDM unit, Senographe DS, with the X-ray gantry set to $+15^\circ$ with the stereotactic biopsy needle assembly attached. Both Senographe systems (2000D and DS) are equipped with CsI scintillator on an a-Si photodiode–transistor arrays with pixel size of 100 μm and an image matrix size of 1920×2304, covering a field of view 19×23 cm.

FISCHER IMAGING'S SENOSCAN FFDM

On the other hand, Fischer Imaging Corporation's answer to the FFDM is the Senoscan FFDM unit. The unique feature of Senoscan unit is that it employs a slot mechanism for the X-ray filed collimation that is synchronized to the detector as the slot and detector assembly sweeps across the imaging field, see Fig. 13. The detector of

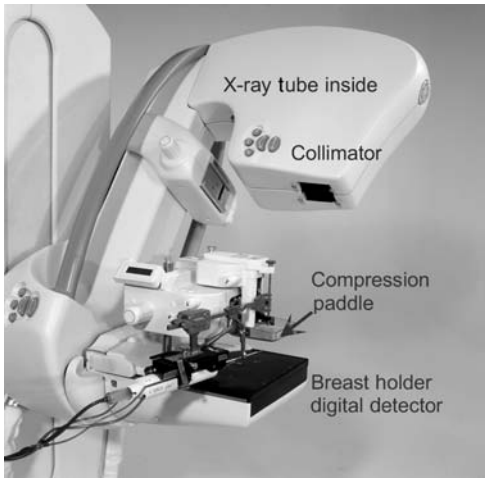


Figure 12. Photograph of GE FFDM unit; Senographe DS with stereotactic attachment installed. The gantry is tilted to $+15^\circ$ and prepped for Stereotactic imaging. Note, this is the FFDM version of similar unit shown in Fig. 1, but with zoomed up view for the stereotactic attachment. (Courtesy of GE Healthcare.)

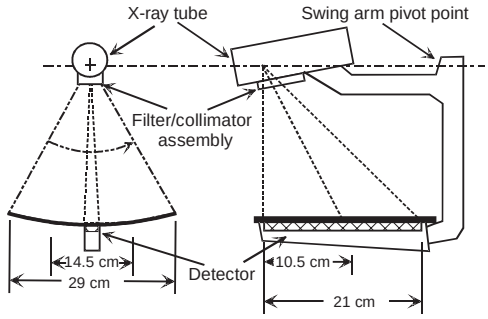


Figure 13. A schematic drawing of the scanning mechanism of Fischer Senoscan FFDM. (Adapted from Fisher Imaging Sanoscan user's manual with permission. Courtesy of Fischer Imaging.)

Senoscan unit consists of a layer of thallium activated CsI scintillator connected to charge-coupled-device (CCD) chips via fiber optics (33). Hence, this is also an indirect-conversion digital detector system.

The use of slot mechanism also means "restricting" the X-ray beams for exposure. It offers an advantage of less scattered radiation for improved image contrast (34). Therefore, no antiscatter grid is necessary with Senoscan unit. The absence of the antiscatter grid compensate for a less efficient use of the available X rays. In addition, a tungsten anode X-ray tube (35) is employed with this unit, since the tungsten anode X-ray tube would produce X rays more efficiently than a molybdenum or rhodium target. Senoscan is equipped with CCD chips having pixel size of (1) 27 μm, covering a field of view 11×15 cm under the high resolution mode, and (2) pixel size of 54 μm, covering a field of view 19×29 cm for the normal mode. The image matrix size of this system is 4096×5625.

HOLOGIC/LORAD'S FFDM; SELENIA

Selenia's image receptor is a 250 μm thick amorphous Selenium (a-Se) photoconductor. Underneath a layer of a-Se photoconductor is a thin-film transistor (TFT) arrays that serves as an active readout mechanism. The TFT arrays are typically deposited onto a glass substrate, which provides a physical support for the entire detector components. Selenia uses a 250 μm thick a-Se photoconductor to capture the X-ray photons impinging on the detector surface without the aid of a scintillator. It is said that a thickness of 250 μm a-Se photoconductor is adequate to stop 95% of the X-ray photons in the mammographic energy range (36). The photon energy is converted to a pair of electron, which is negatively charged, and a positively charged "hole". With bias voltage applied, the signal is read off by the TFT arrays. Thus, this is a direct-conversion digital detector system. The detector is an a-Se photoconductor and TFT arrays with pixel size of 70 μm, covering a field of view 24×29 cm, resulting in an image matrix size of 3328×4096. A summary of the detector characteristics for these three FFDM systems is given in Table 2, (37).

READING MAMMOGRAPHY IMAGES

Initially, the FFDM images were printed on dry laser printers and read on high luminance view boxes (ACR accreditation required of a luminance of $> 3000\,\mathrm{cd \cdot m^{-2}}$) (38). All three FFDM systems are equipped with workstations where images are soft-copy read. The workstations are commonly equipped with two 5-megapixel high resolution monitors. With the soft-copy reading, an assortment of image manipulation are possible including, but not limited to, basic window width and window level adjustments, zooming, image reversal, and so on for viewing the details of the pathology. More importantly, the impact of image processing in transforming the acquired raw data set to the image displayed for soft-copy reading should be recognized (39). For example, as can be seen in Fig. 14, substantial differences not only in the brightness and contrast of the image but also in the impression of pathology in the images can be noticed. While all three images are acceptable and diagnostic, the two enhanced images are easier to recognize various details.

SUMMARY

A brief history of mammography was described as the introduction to the X-ray mammography. The SFM continues to play its major role in breast cancer detection while FFDM is gaining its installed base. Upon the anticipated FDA approval, the CR-M is poised for introduction for clinical applications in the second-half of 2005. However, all three modalities employed in breast cancer detection, namely; the SFM, FFDM, and the CR-M are expected to

Table 2. Summary of Imaging Characteristics of FFDM Units

Manufacturer	Model Name	Scintillator	Detector	Pixel Size, μm	Image Matrix Size	Imaging Size, $L \times W$ cm
GE	Senographe 2000D, DS	CsI (Tl)[a]	TFT	100	1920×2304	19×23
Fisher Imaging	Senoscan	CsI (Tl)	CCD	24/48	4096×5625	22×30
Hologic/LoRad	Selenia	a-Se[b]	TFT	70	3328×4096	25×29
	DSM	CsI (Tl)	CCD	48	1025×1024	5×5
Fuji	CR-M	BaFBr(Eu)[c]	Computed radiography	50	1770×2370 2364×2964	18×24 24×30

[a]CsI(Tl) = Talium Activated Cesium Iodide.
[b]a-Se = Amorphous Selenium.
[c]BaFBr(Eu) = Barium Fluorobromide with Europium.

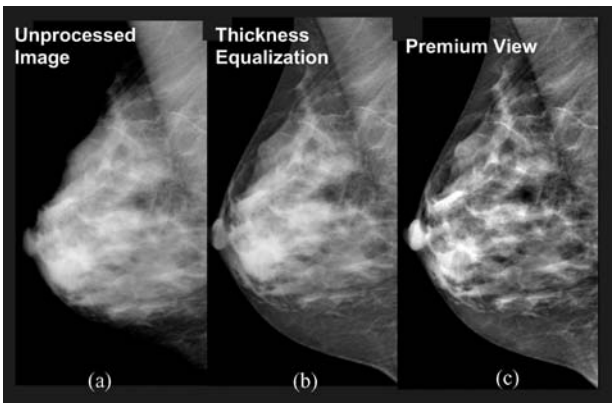


Figure 14. Comparison of "processed" images. The unprocessed image (a) can be dramatically improve its appearance with "Thickness Equalzation" processing (b), or the "Premium View" processing (c). The impact of image processing is quite evident. (Courtesy of GE Healthcare.)

play their roles in contributing to reduction of breast cancer deaths through the early detection of mammography screening.

Display of FFDM images and its workstations are important components of the total picture of FFDM (40,41). However, monitors and workstations associated with diagnostic radiology imaging including mammography are in a domain of their own and no attempt is made to include these subjects in this article. Such subject matters belong to digital image processing and display. Finally, archiving of the acquired digital images is also a very important aspect of digitized diagnostic images in the overall operation of radiology. However, this subject is better handled in Picture Archiving and Communication Systems (PACS), and readers are refered to articles under PACS for additional information.

ACKNOWLEDGMENT

The information and materials presented here had been provided by numerous numbers of representatives and scientists from respective manufacturers mentioned in the main text. The author would like to express his thanks and sincere appreciation by listing the manufacturers here, (in alphabetical order) in lieu of listing the individuals who had provided their supports: General Electric Healthcare, Fischer Imaging Corporation, Fujifilm Medical Systems USA, Inc., and Hologic Inc. (LoRad).

BIBLIOGRAPHY

1. Nystrom L, et al. Breast cancer screening with mammography: Overview of Swedish randomised trial, Lancet 1993; 341:973–978.
2. Hendrick RE, et al. Benefit of screening mammography in women aged 40–49: A new meta-analysis of randomized controlled trial. J Nat Cancer Inst Monograph 1997; (22):87–92.
3. Mammography—A User's Guide, Washington, DC: National Council on Radiation Protection and Measurements, 1986. NCRP Report No. 85.
4. Zermenhof RA. Mammography, Encyclopedia of Medical Devices and Instrumentation, Vol 3. New York: John Wiley & Sons; 1988.
5. Jones CH. Methods of breast imaging. Phys Med Biol 1982;27(4)463–499.
6. FDA Home page. [Online]. Available at http://www.fda.gov/cdrh/mammography/frmamcom2.html.
7. ACR Home page. [Online]. Available at http://www.acr.org/s_acr/sec.asp?CID=589&DID=14253.
8. Speiser RC, Zanrosso EM, Jeromin LS. Dose comparisons for mammographic systems. Med Phys 1986;13(5):667–673.
9. Boag JW. Xeroradiography. Phys Med Biol 1973;18:3.
10. Wolbarst AB. Dependence of attenuation on atomic number and photon energy. Physics of Radiology. Appleton & Lange; 1992. Chapt. 14.
11. Hoeffken W. Soft tissues—mammography. The invisible light applied, the advancing diagnostic evaluation of the skeleton and thorax following Roentgen's discovery. In: Rosenbusch G, Oudekerk M, Ammann E, editors. Radiology in Medical Diagnostics, Evolution of X-ray Applications 1895–1995. Oxford: Blackwell Science Ltd.; 1995. Chapt. 2.
12. Curry III TS, Dowdey JE, Murry Jr. RC. Christensen's Physics of Diagnostic Radiology, 4th ed., Philadelphia: Lea & Febiger; 1990.
13. Fewell TR, Shuping RE. Handbook of Mammographic X-ray Spectra. Rockville, MD: HEW Publication (FDA) 79–8071; 1978.
14. Boone JM, Fewell TR, Jennings RJ. Molybdenum, rhodium, and tungsten anode spectral modeling using interpolating polynomials with application to mammography. Med Phys Dec. 1997:24(12).
15. Sprawls Jr. P. Chapter 18: Blur, Resolution, and Visibility of Detail. Physical Principles of Medical Imaging. Aspen Publishers, Inc.; 1987.
16. Haus AG. Technologic improvements in screen-film mammography. Radiology 1990;174:628–637.
17. Chapter V. Film Response. The Fundamentals of Radiography, 4th ed., Rochester, N.Y.: Eastman Kodak Co.; 1980.
18. Rezentes PS, de Almeida A, Barnes GT. Mammography grid performance. Radiology 1999;210:227.
19. Sonoda M, Takano M, Miyahara J, Kato H. Computed radiography utilizing laser stimulated luminescence. Radiology 1983;148:833–838.
20. Pisano ED, et al. American College of Radiology Imaging Network Digital Mammographic Imaging Screening Trial: Objectives and Methodology. Radiology published online. June 16, 2005.
21. ACRIN Home Page. [Online]. Available at http://www.acrin.org.
22. Communication with National Program Manager, Women's Healthcare Imaging Systems, Fuji Photo Film Ltd.
23. FCR (Fuji Computed Radiography) General Description of Image Processing, Fuji Photo Film Co., Ltd.
24. Tateno Y, Iinuma T, Takano M, editors. Computed Radiography. Tokyo: Springer-Verlag; 1987.
25. Carr JJ, et al. Stereotactic localization of breast lesions: How it works and methods to improve accuracy. RadioGraphics 2001;21:463.
26. Winfield DL. Aerospace technology transfer to breast cancer imaging. Acta Astronau 1997;41(4–10):515–523.
27. Pisano ED, Yaffe MJ. Digital Mammography. Radiology 2005;234:353–362.
28. Pisano ED. Current status of full-field digital mammography. Radiology 2000;214:26.
29. Personal communications with various X-ray industry marketing departments.

30. Vedantham S, et al. Full Breast digital mammography with anamorphous silicon-based flat panel detector: Physical characteristics of a clinical prototype. Med Phys 2000; 27(3):558–567.
31. Albagli D, et al. Performance of advanced a-Si/CsI-based flat panel X-ray detectors for mammography. In: Yaffe MJ, Antonuk LE, editors. Proceedings of SPIE. Medical Imaging 2003: Physics of Medical Imaging, Vol. 5030, June 2003. p 553–563.
32. Shaw J, Albagil D, Wei C-Y. Enhanced a-Si/CsI-based flat panel X-ray detector for mammography. In: Yaffe MJ, Flynn MJ, editors. Proceeding of SPIE. Medical Imaging 2004: Physics of Medical Imaging. Vol. 5368, May 2004. p 370–378.
33. Tesic MM, Picaro MF, Munier B. Full field digital mammography scanner. Eur J Rad 1997;31:2–17.
34. Boone JM, et al. Grid and slot scan scatter reduction in mammography: Comparison by using Monte Carlo techniques. Radiology 2002;222:519–527.
35. Operator Manual, SenoScan: Full Field Digital Mammography System, Fischer Imaging Corporation, Denver, CO, Dec. 2002.
36. Smith AP. Fundamentals of digital mammography: Physics, technology and practical considerations. Radiol Manager 2003, Sep.–Oct.; 25(5):18–24, 26–31.
37. Mahesh M. AAPM/RSNA physics tutorial for residents, Digital mammography: an overview. Radiographics 2004;24: 1747–1760.
38. The ACR Mammography Quality Control Manual. Preston, VA: ACR, 1999.
39. Pisano ED, Yaffe MJ. Digital mammography. Radiology 2005;234:353–362.
40. Hemminger BM, et al. Evaluation of the effect of display luminance on the feature detection of simulated masses in mammograms. Proc SPIE 1997;3036:12.
41. Pisano ED, et al. Radiologists' preferences for digital mammographic display. Radiology 2000;216:820–830.

Further Reading

Mahesh M, AAPM/RSNA physics tutorial for residents, Digital mammography: An overview. *Radiographics* 2004;24:1747–1760.

Smith AP, Hall PA, Marcello DM. Emerging technologies in breast cancer detection. *Radiology Manager* 2004, July.–Aug.; 26(4): 16–24.

Shramchencko N, Blin P, Mathey C, Klausz R. Optimized exposure control in digital mammography. In: Yaffe MJ, Flynn MJ, editors. Proceedings of SPIE. Medical Imaging 2004: Physics of Medical Imaging. May 2004; Vol. 5368, p 445–456.

Curry III TS, Dowdey JE, Murry Jr. RC. Christensen's Physics of Diagnostic Radiology. 4th ed., Philadelphia: Lea & Febiger; 1990. (This is a textbook used in various radiology residency programs across the United States.)

Online References

FDA Home page. [Online]. Available at http://www.fda.gov/cdrh/mammography/frmamcom2.html.

ACR Home page. [Online]. Available at http://www.acr.org/.

ACRIN Home Page. [Online]. Available at http://www.acrin.org.

GE Healthcare Home Page. [Online]. Available at http://www.gehealthcare.com/usen/whc/whcindex.html.

Fischer Imaging Home Page. [Online]. Available at http://www.fischerimaging.com/default/default.asp.

Fujifilm Medical Systems USA Home Page. [Online]. Available at http://www.fujimed.com.

Hologic Home Page. [Online]. Available at http://www.hologic.com/.

MATERIALS, BIOCOMPATIBILITY OF. See BIOCOMPATIBILITY OF MATERIALS.

MATERIALS, PHANTOM, IN RADIOLOGY. See PHANTOM MATERIALS IN RADIOLOGY.

MATERIALS, POLYMERIC. See POLYMERIC MATERIALS.

MATERIALS, POROUS. See POROUS MATERIALS FOR BIOLOGICAL APPLICATIONS.

MEDICAL EDUCATION, COMPUTERS IN

ARIE HASMAN
Maastricht
The Netherlands

INTRODUCTION

The amount of knowledge is increasing rapidly in many disciplines. Medicine is not an exception. Because of scientific research new knowledge comes available at such a pace, that physicians should read 19 articles a day, every day of the week, to keep up to date. Since that is not possible the results of scientific research often are applied clinically only years later. Computers can support physicians in finding relevant recent information. In the next section, the reasons for using computers in medical education are presented. Then the roles computers can play in medical education are reviewed. Since health professionals increasingly use computer systems for their work, they need to know the benefits and limitations of these systems. The discipline of medical informatics is responsible for developing these systems and therefore is discussed. The next sections discuss the use of Internet and electronic patient records. Also, it is discussed why knowledge of information systems is important for health professionals.

THE REASONS FOR USING COMPUTERS IN MEDICAL EDUCATION

The goal of academic medical education is to educate students to become physicians. During the study knowledge, skills and attitudes have to be mastered. Students are taught academic skills like critical reading, they are acquainted with the principles of research methods, and they should know the scientific background of the basic disciplines (like anatomy, molecular cell biology and genetics, endocrinology and metabolism, immunology and inflammation, growth, differentiation and aging) as far as they are related to the study of abnormalities and to diagnosis and therapy. Because of the explosive growth of biomedical knowledge, it is not possible anymore to teach all the currently available knowledge. This does not have to be a problem since part of the knowledge presented to medical students during their formal education may be more or less obsolete by the time they are in their main professional practice. Moreover, it is difficult to teach medicine for the coming era, since most of the future's

technology is probably nonexistent today. Students therefore must be taught how to become life-long learners. Computers can support this process.

Computers play an increasing role in the practice of medicine too. No doctor—whether a general practitioner or a specialist in advanced or social care—will be able to escape the confrontation with some form of information processing. The work of health professionals is dominated by information collection, storage, retrieval, and reasoning. Health professionals both use individual patient data and general medical or nursing knowledge. The amount of medical and nursing knowledge increases so quickly that health professionals cannot stay fully up to date. Tools are therefore needed to acquire relevant knowledge at the time it is needed.

Computer systems are installed in many hospital departments and physician offices. Hospital information systems support, for example, financial, administrative, and management functions. Clinical departmental systems are used to collect, store, process, retrieve, and communicate patient information. Clinical support systems are used in function laboratories [for electroencephalogram (EEG), electrocardiogram (ECG), electromyogram (EMG), and spirometry analysis], for imaging [magnetic resonance imaging (MRI), computerized tomography (CT), nuclear medicine, ultrasound], and in clinical laboratories (analysis of electrolytes, etc.). The results of clinical support systems are increasingly stored in so-called electronic patient records, together with the medical history, results of physical examination, and progress notes. Electronic patient records gradually replace the paper patient record. Apart from the fact that electronic paper records are better readable they also support functions (like decision support) paper records cannot provide. Students must learn the benefits and limitations of these kinds of systems.

Decision support systems are used to support clinicians during the diagnostic or therapeutic process and for preventive purposes to prevent either errors of omission (when, eg, the physician does not order a mammography when indicated) or commission (when, eg, the physician prescribes the wrong drug).

Clinical patient data are increasingly stored in the above mentioned electronic patient records (EPRs), from which they can be later retrieved by physicians or nurses who are in need of the data. Also, information systems can retrieve relevant data from the electronic patient record when a suitable interface between system and EPR is available. When a standard vocabulary is used for representing data values in the EPR, decision support systems can interpret these data and remind, alert, or critique the physician or provide access to relevant knowledge, based on patient data available in the EPR. Health professionals should have insight in and knowledge of the principles, concepts, and methods underlying electronic patient records.

Also, patients become active players in the field and increasingly demand access to the EPR.

Patient management increasingly has become the combined task of a group of healthcare workers. Therefore the memory-aid role of the patient record more and more changes into a communication role. Paper records have several limitations in this respect. In addition, since the appearance of the report "To err is human" of the IOM (1) it is apparent that due to miscommunication (among which are problems with reading handwritten information, with incomplete information, etc.) medical errors are made that even may lead to the death of patients. Electronic patient records and order entry systems can reduce the number of errors because they are not only more readable, but because they can also be interfaced with decision support systems when standardized terminology is used.

Decision support can be passive in the sense that the decision support system contains information that has to be searched by the physician. In this case, the healthcare professional takes the initiative to search for information, for example, via PubMed or the Cochrane Library. Decision support can also be active. In this case, the decision support system volunteers advice based on information it can retrieve from the EPR. Decision support systems can either proactively suggest the next step in a diagnostic or treatment process or reactively remind the healthcare professional that a (preventive) procedure was not performed or a step in the protocol was not carried out.

ROLES FOR COMPUTERS IN MEDICAL EDUCATION

What roles can computers play in medical education? In the first place, information systems can be used to manage the learning process. Students can get access to the curriculum via so-called learning environments (e.g., Blackboard or WebCT), can get overviews of their marks, can access computer aided instruction programs, can access PowerPoint presentations of their teachers, and so on. Computers provide access to the internet so that they can search for content knowledge.

Computer-aided instruction can be used to teach students certain subjects. For example, computers are used to simulate regulatory mechanisms occurring in the human body. With a simulation program, students can treat "patients" without risking their patient lives. The simulation is often based on models that present (patho)physiologic processes in the form of mathematical equations. When the models become increasingly accurate they can even be used in patient care. Also, patient management problems can be simulated. In this case, usually no mathematics is involved, but the patient's signs and symptoms as they develop as a function of time are expressed as text.

There are simulation tools that allow users to evaluate, plan, or redesign hospital departments, or parts of other healthcare systems. Physicians will be confronted with the results of simulations. The model that is used in the simulation has to be checked for validity. Some tools present the modeled processes visually so that physicians or nurses can easily determine the correctness of the model. An example is the modeling of a phlebography service. On the screen, the cubicles and other rooms are displayed and the movements of patients, physicians, and nurses can be followed. In this way, the users can judge whether the model represents the situation of a phlebography service in an adequate way.

Note that computers should not be considered as surrogate teachers controlling students' learning. Computers should enrich the learning environment by expanding the student's control over their self-learning and by providing a better learning environment as a supplement to traditional methods of learning. The effectiveness of CAI has always been a subject of controversy. Studies have claimed both that CAI is superior and that CAI is inferior to traditional methods. The majority of the publications, however, support the notion that CAI is as effective as traditional educational methods (2).

Although decision making is the pre-eminent function of a physician, hardly anywhere is the student confronted with a systematic exposition of procedures of good decision making. The use of computers can facilitate the teaching of these procedures. Computer support offers new possibilities to teach problem solving techniques and analytical methods that are presently learned by the student through practice and the observation of mentors.

MEDICAL INFORMATICS

Education concerning the advantages and limitations of the use of computers for supporting the work of health professionals is the responsibility of medical informatics departments. Medical informatics can also be instrumental in developing computer-aided instruction programmes and simulation packages. Medical informatics is the discipline that deals with the systematic processing of data, information, and knowledge in the domains of medicine and healthcare. The objects of study are the computational and informational aspects of processes and structures in medicine and healthcare (3). Medical informatics is a very broad discipline covering subjects like applied technology (bioinformatics, pattern recognition, algorithms, human interfaces, etc.) and services and products (quality management, knowledge-based systems, electronic patient records, operations–resource management, etc.). Also human and organizational factors (managing change, legal issues, needs assessment, etc.) should be taken into account.

Informatics can be either a systems-oriented discipline in which computer systems, operating systems and programming languages are the object of study or a methods-oriented discipline in which the methods are studied that can be used to create algorithms that solve problems in some application domain. In the case of the methods-oriented approach, a problem is studied and formalized solutions are determined. Medical informatics is an example of this approach: it studies the processing of data, information, and knowledge in medicine and healthcare. Medical informatics focuses on the computational and informational aspects of (patho)physiological processes occurring in the patient, cognitive processes going on in the brain of physicians, and organizational processes that control healthcare systems.

The resulting knowledge can be used to design information systems that can support healthcare professionals. It is clear that healthcare professionals need to have some medical informatics knowledge in order to optimally use information systems. Medical informatics education should therefore be part of the medical curriculum.

Various groups of professionals with quite different backgrounds can be identified who carry out medical informatics tasks ranging from the use of IT to developing information systems. Users of information systems naturally need less medical informatics knowledge than health informatics experts who develop health information systems or support other healthcare workers in designing terminology servers, coding systems, and so on.

There exists a wide range of job opportunities in the field of medical informatics. These jobs require various medical informatics capabilities. In addition to medical informatics, students (and graduate students) with other backgrounds may prefer a job in the field of medical informatics. In order to obtain the relevant capabilities these students have to learn additional subjects depending on their previous education and the type of specialization they want to achieve. These students can be graduates from healthcare related programs or from informatics–computer science programs. Graduates from healthcare related programs possess the relevant medical knowledge, but need to increase their medical informatics knowledge. Graduates with an informatics or computer science background must learn how the healthcare system is organized and how healthcare professionals are working in order to develop systems that are appreciated by healthcare professionals. Medical informatics is therefore taught in different ways (4) depending on the type of students and the type and extent of specialization that they want to achieve.

USE OF THE INTERNET

Much knowledge can be found on the internet. Browsers allow health professionals and patients to access sites containing (references to) medical knowledge. PubMed is an example. It contains references to the medical literature. The web contains a lot of information of which the quality is not always guaranteed. Especially in the medical arena, this is a big disadvantage. The internet has become one of the most widely used communication media. With the availability of Web server software, anyone can set up a Web site and publish any kind of data that is then accessible to all. The problem is therefore no longer finding information, but assessing the credibility of the publisher as well as the relevance and accuracy of a document retrieved from the net. The Health On the Net Code of Conduct (HONcode) has been issued in response to concerns regarding the quality of medical and health information (5). The HONcode sets a universally recognized standard for responsible self-regulation. It defines a set of voluntary rules to make sure that a reader always knows the source and the purpose of the information they are reading. These rules stipulate, for example, that any medical or health advice provided and hosted on a site will only be given by medically trained and qualified professionals unless a clear statement is made that a piece of advice is offered from a nonmedically qualified individual or organization. Another guideline states that support for the Web site should be clearly identified, including the identities of

commercial and noncommercial organizations that have contributed funding, services or material for the site. Students searching for information should be introduced to these guidelines.

Searching can be carried out by entering keywords. These keywords can be connected by Boolean operators like AND, OR, and NOT. A user can for example enter: Diuretics and Hypertension to search for documents that discuss the use of diuretics in hypertension. The NLM (National Library of Medicine) uses the Medical Subject Headings (MeSH) vocabulary for indexing most of their databases. Students should be taught how to efficiently search in bibliographic databases using, for example, the MeSH vocabularies.

We speak of e-learning when content is accessible via Web browsers. Some characteristics of e-learning follow: Internet is the distribution channel. Access to the content is possible 24 h/7 days a week. It is learner-centered. The student determines the learning environment, the speed of learning, the subjects to consider, the learning method. A mix of learning methods can be used (blended learning): for example, virtual classroom, simulations, cooperation, communities, and "live" learning.

Virtual learning environments aim to support learning and teaching activities across the internet. Blackboard and WebCT are examples of such environments. These environments offer many possibilities. New or modified educational modules can be announced or teachers can give feedback regarding the way a module is progressing. Also general information about a module can be provided. Staff information can be presented with photo, email address, and so on. Assignments can be posted, and so on. The virtual classroom allows students to communicate online, whereas discussion boards allow asynchronous communication. Also, links to other websites can be provided. The internet is a source of information for patients. They can retrieve diagnostic and therapeutic information from the internet. Increasingly, patients present this information to their care providers. Health professionals must know how to cope with this new situation and must be able to assess the quality of the information.

KNOWLEDGE OF INFORMATION SYSTEMS

Information systems are increasingly used in healthcare. They not only support administrative and financial, but also clinical and logistic processes. Since healthcare workers have to use information systems they should know the possibilities, but also the limitations, of information systems. Since in information systems, for example, data can be easily retrieved, the quality of entered data determines the quality of the results: garbage in, garbage out. In addition, they should have the skills to work with information systems. Information systems relevant for healthcare professionals include hospital information systems, departmental systems, electronic patient record systems, order entry and result reporting systems, and so on. But healthcare workers should also be proficient in the use of productivity tools like word processing systems, bibliographic search systems, and so on.

Logistics is becoming more important these days. Hospitals have to work not only effectively, but also more efficiently, thereby taking the preferences of patients into account. Planning systems can reduce the time that ambulatory patients have to spend in the hospital for undergoing tests, but also the length of stay of hospitalized patients can be reduced by planning both the patients and the needed capacity (6).

It is important for healthcare workers to know what support they can expect from information systems and to know which conditions have to be satisfied in order that information systems can really be of help. Optimal use of information systems therefore does not only depend on acquired skills, but also on the insight in and knowledge of the principles, concepts, and methods behind information systems. This is true for all types of healthcare professionals. When hospitals or physicians consider the purchase of information systems they must be able to specify their requirements so that they will not be confronted with systems that do not perform as expected.

ELECTRONIC PATIENT RECORDS

Physicians store information about their patients in patient records. The patient record frequently is a paper record in which the physician writes his notes. Paper records have several advantages because they are easy to use, easy to carry, and so on. But there are also limitations: they may be difficult to read and are totally passive: the physician records the information with little support (e.g., headings in a form) and therefore the recordings are often incomplete. Not only are the data incomplete, they also contain errors, for example, due to transcription or because the patient gave erroneous information. The readers of patient records can interpret the data in the patient record incorrectly. The fact that data are recorded in chronological order makes the retrieval of facts sometimes difficult: such data are not recorded on standard positions in the record. A study showed that because of the constrained organization of paper records, physicians could not find 10% of the data, although these data were present in the paper record (7). The patient data are usually stored in more than one type of record, because each department uses its own records, each with a different lay-out. Paper records are not always available at the time they are needed. Paper records are passive: they will not warn the physician if he overlooks some results. If the results of a lab request are unexpectedly not yet available, the paper record will not indicate that. Despite these drawbacks physicians are usually very positive about the use of the paper record, because they do not recognize their shortcomings.

Electronic patient records have some advantages over paper records. The legibility of electronic patient records is per definition good. Data can be presented according to different views (time-, source-, and problem-oriented), making the data easier to access. Electronic patient records, when interfaced with decision support systems, can provide reminders when a physician forgets something.

The use of computers in medical education is diverse. They can be used for managing the learning process, for distributing content, for assessing the student's knowledge or skills, and so on. In this case, they can be regarded as educational tools. As is clear from the above information systems are extensively used in the practice of health professionals. Students should be taught the benefits, but also the limitations of the use of information systems. Finally the information systems have to be developed. To be able to do so students need additional education in medical informatics.

BIBLIOGRAPHY

1. Institute of Medicine. To err is human: Building a safer health system. The National Academies Press; 2000.
2. Qayumi AK, et al. Comparison of computer-assisted instruction (CAI) versus traditional textbook methods for training in abdominal examination (Japanese experience). Med Ed 2004;38:1080–1088.
3. Hasman A, Haux R, Albert A. A systematic view on medical informatics. Comp Meth Prog Biomed 1996;51: 131–139.
4. Haux R, Grant A, Hasman A, Hovenga E, Knaup P. Recommendations of the International Medical Informatics Association (IMIA) on education in health and medical informatics. Methods Inf Med 2001;40:78–82.
5. http://www.hon.ch (last visited 21 December 2004).
6. van Merode GG, Groothuis S, Hasman A. Entreprise resource planning for hospitals. Int J Med Inform 2004;73: 493–501.
7. Fries JF. Alternatives in medical record formats. Med Care 1974;12:871–881.

MEDICAL ENGINEERING SOCIETIES AND ORGANIZATIONS

ARTHUR T JOHNSON
University of Maryland
College Park, Maryland

PATRICIA I HORNER
Landover, Maryland

INTRODUCTION

Modern technology has transformed the practice of medicine. We can now see where we could not before, conduct surgery with minimal trauma, intervene at the genetic level, replace whole natural organs with functional artificial ones, make rapid diagnoses, and peer into the workings of the brain. More patients are surviving, and those who do are living better. Much of the credit for these advances goes to the engineers, physicians, and physiologists who together decided what needed to be done, the science required to support it, and how it could be made practical. Medical engineers are now very much involved in the process of developing medical advances. They bring to medicine the abilities of conceptualization, computation, and commercialization. They use varied tools such as biophysics, applied mathematics, physiological modeling, bioinstrumentation and control, imaging, and biomechanics to accomplish their advances.

The result is that there are nearly as many subspecialties of medical engineering as there are medical specialties. Tissue engineers, for instance, grow bioartificial tissues and organs as replacements; metabolic engineers find means to adjust cellular metabolic pathways to produce greater quantities of biochemicals and hormones; and rehabilitation engineers design new prostheses or modify existing units to reestablish adequate function in patients who have lost ability usually as the result of trauma. There are medical engineers working with biosensors, bioprocess optimization, multiple imaging modes, pancreatic function, vascular replacement, and drug delivery. Biomaterials engineers have produced materials that can function in different regional corporal environments. Indeed, there is no part of the human body that has not been studied by medical engineers to improve or replace lost function.

As the body of medical knowledge has increased overall and has been repeatedly split more and more finely into specialties, there has been a concomitant proliferation of organizations to communicate, share, and advocate action related to their particular specialties. Some of these would be recognized as chiefly engineering organizations with application interests in medicine; some are medical societies with significant engineering contributions. There is almost no significant human disease, physiological system, organ, or function without a group or organization representing associated interests. There is even a group interested in developing synthetic biological forms that, although it is too premature to link with medicine, may someday have a profound effect on medicine. All of these groups can be found by searching the Internet, and any attempt to enumerate them here would be outdated very quickly.

DEFINITIONS

Progress in biological science and engineering has not been made with a clear distinction between medical and nonmedical applications. Advances in human medicine often find applications as well in veterinary medicine. Genetic coding techniques have been applied equally to humans and fruit flies. Prospective biomaterials are modeled on computer without regard for the ultimate specific application, and they are tested in animals, plants, or fungi before approval for human use. Progress toward better nutrition through science, and toward purer environments through improved pollutant detection monitoring, have resulted in better human health for most humans living in the developed world. Biology is biology, whether applied to human health care or not, so a convergence of basic knowledge and methods between medical and biological engineers is expected to continue.

Several relevant definitions attempt to distinguish among various fields where engineers have and will continue to contribute.

The U.S. National Institutes for Health (NIH) has the following definition of bioengineering:

> Bioengineering integrates physical, chemical, mathematical, and computational sciences and engineering principles to study biology, medicine, behavior, and health. It advances fundamental concepts; creates knowledge from the molecular to the organ systems levels; and develops innovative biologics, materials, processes, implants, devices, and informatics approaches for the prevention, diagnosis, and treatment of disease, for patient rehabilitation, and for improving health.

The U.S. National Science Foundation program in Biochemical Engineering and Biotechnology (BEB) describes its program in the following way:

> Advances the knowledge base of basic engineering and scientific principles of bioprocessing at both the molecular level (biomolecular engineering) and the manufacturing scale (bioprocess engineering). Many proposals supported by BEB programs are involved with the development of enabling technologies for production of a wide range of biotechnology products and services by making use of enzymes, mammalian, microbial, plant, and/or insect cells to produce useful biochemicals, pharmaceuticals, cells, cellular components, or cell composites (tissues).

The Whitaker Foundation definition of biomedical engineering is as follows:

> Biomedical engineering is a discipline that advances knowledge in engineering, biology, and medicine, and improves human health through cross-disciplinary activities that integrate the engineering sciences with the biomedical sciences and clinical practice. It includes: 1) The acquisition of new knowledge and understanding of living systems through the innovative and substantive application of experimental and analytical techniques based on the engineering sciences, and 2) The development of new devices, algorithms, processes, and systems that advances biology and medicine and improve medical practice and health care delivery.

And, finally, the Institute of Biological Engineering (IBE) defines biological engineering as follows:

> Biological engineering is the biology-based engineering discipline that integrates life sciences with engineering in the advancement and application of fundamental concepts of biological systems from molecular to ecosystem levels. The emerging discipline of biological engineering lies at the interfaces of biological sciences, engineering sciences, mathematics and computational sciences. It applies biological systems to enhance the quality and diversity of life.

HISTORICAL DEVELOPMENTS

In 1948, in New York City, a group of engineers from the Instrument Society of America (ISA) and the American Institute of Electrical Engineers (AIEE), with professional interests in the areas of X-ray and radiation apparatus used in medicine, held the First Annual Conference on Medical Electronics. Soon thereafter the Institute of Radio Engineers (IRE), joined with the ISA and AIEE, and the series of annual meetings continued. Subsequent years witnessed a remarkable growth of interest in biomedical engineering and participation by other technical associations. By 1968 the original core group evolved into the Joint Committee on Engineering in Medicine and Biology (JCEMB), with five adherent national society members: the Instrument Society of America (ISA), the Institute of Electrical and Electronics Engineers, Inc. (IEEE), the American Society of Mechanical Engineers (ASME), the American Institute of Chemical Engineers (AIChE), and the Association for the Advancement of Medical Instrumentation (AAMI), who jointly conducted the Annual Conference on Engineering in Medicine and Biology (ACEMB).

Professional groups responded vigorously to the demands of the times. Attendance at the annual conference by natural scientists and medical practitioners grew to approximately 40% of the total; medical associations requested formal participation with their technical counterparts on the JCEMB. New interdisciplinary organizations were formed. New intrasociety and intersociety groups, committees, and councils became active; meetings filled the calendar; and publications overflowed the shelves.

In 1968, a document was prepared that read as follows:
WHEREAS:

1. Common interdisciplinary purposes cannot be well served by individual groups working independently from each other;
2. Certain associations have developed in attempts to meet the need;
3. Conferences and publications have proliferated in attempts to meet the needs;
4. At present, no mutually satisfactory mechanism exists for the coordination of the relevant groups and functions;
5. There does exist an annual meeting and proceedings publication sponsored by a limited number of societies through the Joint Committee on Engineering in Medicine and Biology (JCEMB);
6. The JCEMB is formally structured with a constitution, plural societal representati9on, and an established pattern of operation. This structure and pattern of operation, however, are not deemed adequate to fulfill present and future needs. To the best of our knowledge, there exists no other single organization that seems capable of fulfilling these needs.

THEREFORE, it is appropriate that a new organization be established.

On July 21, 1969, at the 22nd ACEMB in Chicago, Illinois, representatives of 14 national engineering, scientific, and medical associations founded the Alliance for Engineering in Medicine and Biology (AEMB). It was incorporated on December 24, 1969, in Washington, D.C. Lester Goodman, Ph.D., served as Founder President in

1970–197l; Arthur C. Beall, MD(1972); Alan R. Kahn, MD(1973); Harry S. Lipscomb, MD(1974); Anthony Sances, Jr., Ph.D. (1975); Charles Weller, MD(1976–1977); Edward J. Hinman, MD MPH(1978–1979); Paul W. Mayer, MD(1980–1982); Francis M. Long, Ph.D., (1983–1984); Arthur T. Johnson, PE, Ph.D. (1985–1988); and Alfred R. Potvin, PE Ph.D. served as the final President in 1989–1990.

The Alliance operations were determined by an Administrative Council composed of delegates from each of its affiliates. Later the Alliance was to consist of more than 20 such organizations:

Aerospace Medical Association (ASMA)
American Academy of Orthopaedic Surgeons (AAOS)
American Association of Physicists in Medicine (AAPM)
American College of Chest Physicians (ACCP)
American College of Physicians (ACP)
American College of Radiology (ACR)
American College of Surgeons (ACP)
American Institute of Aeronautics and Astronautics (AIAA)
American Institute of Biological Sciences (AIBS)
American Institute of Chemical Engineers (AIChE)
American Institute of Ultrasound in Medicine (AIUM)
American Society for Artificial Internal Organs (ASAIO)
American Society for Engineering Education (ASEE)
American Society for Hospital Engineers of the American Hospital Association (ASHE)
American Society for Testing and Materials (ASTM)
American Society of Agricultural Engineers (ASAE)
American Society of Civil Engineers (ASCE)
American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE)
American Society of Internal Medicine (ASIM)
American Society of Mechanical Engineers (ASME)
Association for the Advancement of Medical Instrumentation (AAMI)
Biomedical Engineering Society (BMES)
Institute of Electrical and Electronics Engineers (IEEE)
Instrument Society of America (ISA)
National Association of Bioengineers (NAB)
Neuroelectric Society (NES)
RESNA—Rehabilitation Engineering & Assistive Technology Society of North America
Society for Advanced Medical Systems, now American Medical Informatics Association (AMIA)
Society for Experimental Stress Analysis (SESA)
SPIE—International Society for Optical Engineering
Alpha Eta Mu Beta—National Biomedical Engineering Student Honor Society, established under the auspices of A EMB.

The Alliance headquarters office opened on November 1, 1973. John H. Busser served as the first Executive Director. Patricia I. Horner served as Assistant Director, as Administrative Director, and succeeded Busser as the Executive Director. Among its goals, is the following excerpted in part from its constitution, bylaws, and recorded minutes:

> to promote cooperation among associations that have an active interest in the interaction of Engineering and the physical sciences with medicine and the biological sciences in enhancement of biomedical knowledge and health care.
>
> to establish an environment and mechanisms whereby people from relevant various disciplines can be motivated and stimulated to work together
>
> to respond to the needs of its member societies, as expressed by their delegates, rather than to seek authoritative preeminence in its domain of interest. . .
>
> to support and enhance the professional activities of its membership. . .

The 23rd ACEMB in Washington, D.C., in 1970, was the first held under the aegis of the Alliance. From 1979 to 1984, the IEEE Engineering in Medicine and Biology Society (EMBS) held their conferences immediately preceding the ACEMB. The Society for Advanced Medical Systems, later to become AMIA, and the Biomedical Engineering Society also held their meetings for several years in conjunction with the ACEMB.

The accomplishments of the Alliance far outstripped the expectations of its founders. The Alliance more than fulfilled responsibilities for the annual conference inherited from the predecessor JCEMB, but the Alliance made important contributions through a variety of studies and publications ranging from a 5-year ultrasound research and development agendum to a guideline for technology procurement in health-care institutions:

- First International Biomedical Engineering Workshop Series held in Dubrovnik, Yugoslavia, under the sponsorship of the National Science Foundation. This project was in cooperation with AIBS and the International Institute of Biomedical Engineering in Paris. Five workshops were held, and planning handbooks were completed.
- Assessment of selected medical instrumentation; Tasks 1–4, ultrasonic diagnostic imaging; Task 5, radiologic and radionuclide imaging technology.
- Summary guidelines and courses on technology procurement; practices and procedures for improving productivity in research and health-care institutions.
- Information exchange and problem assessments in medical ultrasound, including preparation and distribution of a directory of federal activities, conducted instrumentation conferences, delineated training needs, assessed technology transfer potential, and prepared guidelines for the establishment of clinical ultrasound facilities.

- Joint U.S.–Egypt international technology transfer project in medical diagnostic ultrasound, including international workshops and the design and support of a focus laboratory for ultrasonic diagnosis at Cairo University Medical School.
- Short courses for continuing education at the annual conference on engineering in medicine and biology.
- International directory of biomedical engineers.

Before long, the proliferation of medical engineers, and competing interests among societies, led to a fragmentation of the field. It became clear that the Alliance no longer represented positions of the entire field. No organized group could speak for the entire profession, and the spirit of unity that had led to the development of AEMB no longer existed. It was time for a new beginning.

AMERICAN INSTITUTE FOR MEDICAL AND BIOLOGICAL ENGINEERING

In 1988, the National Science Foundation funded a grant to develop an infrastructure for bioengineering in the United States. The AEMB, jointly with the U.S. National Committee on Biomechanics (USNCB), was to develop a unifying organization for bioengineering in the United States. The co-principal investigators were Robert M. Nerem and Arthur T. Johnson, and Patricia Horner served as Project Director. The AEMB/USNCB Steering Committee consisted of Robert M. Nerem, Arthur T. Johnson, Michael J. Ackerman, Gilbert B. Devey, Clifford E. Brubaker, Morton H. Friedman, Dov Jaron, Winfred M. Phillips, Alfred R. Potvin, Jerome S. Schultz, and Savio L-Y Woo. The Steering Committee met in January and March 1989, and the first workshop was held in August 1989. Two more Steering Committee meetings were held in December 1989 and March 1990, and the second workshop was held in July 1990. The outcome of these two workshops was to establish the American Institute for Medical and Biological Engineering (AIMBE). All AEMB members voted to cease operation of the Alliance for Engineering in Medicine and Biology in 1990 and to transfer the AEMB assets and 501(c)3 status to AIMBE in 1991.

AIMBE opened an office in Washington, D.C., in 1995 with Kevin O'Connor as Executive Director. He was succeeded by Arthur T. Johnson in 2004 and Patricia Ford-Roegner in 2005. AIMBE Presidents have been as follows: Robert Nerem (1992–1993), Pierre Galletti (1994), Jerome Schultz (1995), Winfred Phillips (1996), Larry McIntire (1997), William Hendee (1998), John Linehan (1999), Shu Chien (2000), Peer Portner (2001), Buddy Ratner (2002), Arthur Coury (2003), Don Giddens (2004), Thomas Harris (2005), and Herbert Voigt (2006).

Representing over 75,000 bioengineers, the AIMBE seeks to serve and coordinate a broad constituency of medical and biological scientists and practitioners, scientific and engineering societies, academic departments, and industries. Practical engagement of medical and biological engineers within the AIMBE ranges from the fields of clinical medicine to food, agriculture, and environmental bioremediation.

AIMBE's mission is to

- Promote awareness of the field and its contributions to society in terms of new technologies that improve medical care and produce more and higher quality food for people throughout the world.
- Work with lawmakers, government agencies, and other professional groups to promote public policies that further advancements in the field.
- Strive to improve intersociety relations and cooperation within the field.
- Promote the national interest in science, engineering, and education.
- Recognize individual and group achievements and contributions to medical and biological engineering.

AIMBE is composed of four sections:

- The College of Fellows—1000 Persons who are the outstanding bioengineers in academic, industry, and government. These leaders in the field have distinguished themselves through their contributions in research, industrial practice, and/or education. Most Fellows come from the United States, but there are international Fellows.
- The Academic Council—Universities with educational programs in bioengineering at the graduate or undergraduate level. Currently there are approximately 85 member institutions. Representative to the Council generally are chairs of their departments. Many also are members of the College of Fellows. The Council considers issues ranging from curricular standards and accreditation to employment of graduates and funding for graduate study.
- The Council of Societies—The AIMBE's mechanism coordinating interaction among 19 scientific organizations in medical and biological engineering. The purposes of the Council are to provide a collaborative forum for the establishment of society member positions on issues affecting the field of medical and biological engineering, to foster intersociety dialoge and cooperation that provides a cohesive public representation for medical and biological engineering, and to provide a way to coordinate activities of member societies with the activities of academia, government, the health-care sector, industry, and the public and private biomedical communities.
- The Industry Council—A forum for dialog among industry, academia, and government to identify and act on common interests that will advance the field of medical and biological engineering and contribute to public health and welfare. Industrial organizations may be members of the Industry Council if they have substantial and continuing professional

interest in the field of medical and biological engineering.

Current members of the Council of Societies are as follows:

American Association of Physicists in Medicine
American College of Clinical Engineering
American Institute of Chemical Engineers; Food, Pharmaceutical and Bioengineering Division
American Medical Informatics Association
American Society of Agricultural and Biological Engineers
American Society for Artificial Internal Organs
American Society for Biomechanics
American Society of Mechanical Engineers, Bioengineering Division
Biomedical Engineering Society
Controlled Release Society
IEEE Engineering in Medicine and Biology Society
Institute of Biological Engineering
International Society for Magnetic Resonance in Medicine
Orthopaedic Research Society
Rehabilitation Engineering and Assistive Technology Society of North America
Society for Biomaterials
SPIE: The International Society for Optical Engineering
Surfaces in Biomaterials Foundation

Current members of the Industry Council are as follows:

Biomet, Inc.
Boston Scientific Corporation
Genzyme Corporation
Medtronic, Inc.
Pequot Ventures
Smith + Nephew
Vyteris, Inc.
Wright Medical Technology, Inc.
Zimmer, Inc.

The AIMBE Board of Directors oversees the work of the College of Fellows and the three councils. The Board consists of a President who is assisted by two Past Presidents, the President-Elect, four Vice-Presidents at Large, a Secretary-Treasurer, and the Chair of the College of Fellows—all of whom are elected by the Fellows. The Board also includes chairs of the other councils and chairs of all standing committees. AIMBE's day-to-day operations are supervised by the Executive Director in the Washington headquarters.

AIMBE's Annual Event each winter in Washington, D.C., provides a forum on the organization's activities and is a showcase for key developments in medical and biological engineering. The annual event includes a 1-day scientific symposium sponsored by the College of Fellows, a ceremony to induct the newly elected Fellows, and a 1-day series of business meetings focused on public policy and other issues of interest to AIMBE's constituents. For additional information about AIMBE's mission, memberships, and accomplishments, visit http://www.aimbe.org.

The AIMBE has focused on public policy issues associated with medical and biological engineering. The AIMBE enjoys high credibility and respect based on the stature of its Fellows, support from constituent societies, and its intention to be a forum for the best interests of the entire field. The AIMBE has taken positions on several important issues and advocated that they be adopted by various agencies and by Congress. A few of the AIMBE'S public policy initiatives that have met with success are as follows:

- National Institute of Biomedical Imaging and Bioengineering (NIBIB)—Created in 2000 with the help of AIMBE advocacy, the NIBIB has received strong support from the AIMBE and other institutions that value the role of technology in medicine, particularly the Academy of Radiological Research. The NIBIB has experienced rapid growth and development in all areas, including scientific programs, science administration, and operational infrastructure. The prognosis for the near future is continued growth and development especially in bioengineering, imaging, and interdisciplinary biomedical research and training programs.
- FDA Modernization Act (FDAMA)—Enacted in 1997, this legislation amended the Federal Food, Drug, and Cosmetic Act relation to the regulation of food, drugs, devices, and biological products. FDAMA enhanced the FDA's mission in ways that recognized the Agency would be operating in a twenty-first century characterized by increasing technological, trade, and public health complexities.
- Biomaterials Access Assurance Act—The 1998 legislation provides relief for materials suppliers to manufacturers of implanted medical devices by allowing those suppliers to be dismissed from lawsuits in which they are named if they meet the statutory definition of a "biomaterials supplier."
- National Institutes of Health Bioengineering Consortium (BECON)—This is the focus of bioengineering activities at the NIH. The Consortium consists of senior-level representatives from all NIH institutes, centers, and divisions plus representatives of other Federal agencies concerned with biomedical research and development. The BECON is administered by NIBIB.

The AIMBE Hall of Fame was established in 2005 to recognize and celebrate the most important medical and biological engineering achievements contributing to the quality of life. The Hall of Fame provides tangible evidence of the contributions of medical and biological engineering during the following decades:

1. *1950s and earlier*
   - Artificial kidney
   - X ray
   - Cardiac pacemaker
   - Cardiopulmonary bypass
   - Antibiotic production technology
   - Defibrillator
2. *1960s*
   - Heart valve replacement
   - Intraocular lens
   - Ultrasound
   - Vascular grafts
   - Blood analysis and processing
3. *1970s*
   - Computer-assisted tomography (CT)
   - Artificial hip and knee replacement
   - Balloon catheter
   - Endoscopy
   - Biological plant/food engineering
4. *1980s*
   - Magnetic resonance imaging (MRI)
   - Laser surgery
   - Vascular stents
   - Recombinant therapeutics
5. *1990s*
   - Genomic sequencing and micro-arrays
   - Positron emission tomography
   - Image-guided surgery

The AIMBE has now turned its attention to Barriers to Further Innovation. It is providing forums and platforms for identification and discussion of obstacles standing in the way of advances in medical and biological engineering. Barriers could be procedures, policies, attitudes, or information and education, anything that can yield when AIMBE constituents apply pressure at appropriate levels.

OTHER SOCIETIES

These are other general biomedical engineering societies that operate within the United States Among these, the Biomedical Engineering Society (BMES), Engineering in Medicine and Biology Society (EMBS), and the Institute for Biological Engineering (IBE) are probably the most inclusive. Others direct their attentions to specific parts of the discipline. There are trade organizations that have an industry perspective (such as AdvaMed for the medical device industry and the Biotechnology Industry Organization (BID) for the biotech industry), and there are peripheral organizations that deal with public health, the environment, and biotechnology. Many of these organizations publish excellent journals, newsletters, and information sheets. Those from trade organizations are often distributed free of charge, but they do not include peer-reviewed articles. Information about these can be found on the Internet.

Internationally, an organizational hierarchy exists. National and transnational organizations can belong to the International Federation for Medical and Biological Engineering (IFMBE), and that confers membership privileges to all AIMBE members and constituent society members. The IFMBE and the International Organization for Medical Physics (IOMP) together jointly sponsor a World Congress on Medical Physics and Biomedical Engineering every 3 years. The IOMP and IFMBE are members of the International Union for Physical and Engineering Sciences in Medicine (IUPESM), and the IUPESM, in turn, is a member of the International Council for Science (ICSU). ICSU members are national and international scientific unions and have a very broad and global outreach.

THE FUTURE

At least for the foreseeable future, new groups will be formed representing medical engineering specialties. Whether these groups organize formally and persist will depend on the continuing importance of their areas of focus. The organizations with a more general foci will continue to function and may spawn splinter groups. Given the political importance of concerted effort, organizations such as the AIMBE will continue to be active in promoting policy. Competitive pressures among different organizations, especially when expectations of continuing growth cannot be sustained, will always be a threat to the current order. Given that the cycle of competition and disorder leading to a realization that some ordered structure is preferable has been repeated at least once, there will continue to be some undercurrent of turmoil within the community of medical engineering organizations.

U.S. PROFESSIONAL SOCIETIES AND ORGANIZATIONS

Biomedical Engineering Associations and Societies

Advamed. 1200 G Street NW, Suite 400, Washington, D.C. 20005. 202-783-8700, http://www.advamed.org. Stephen J. Ubl, President. 1300 Members.

Represents manufacturers of medical devices, diagnostic products, and medical information systems. AdvaMed's members manufacture nearly 90% of the $80 billion of health-care technology purchased annually in the United States. Provides advocacy, information, education, and solutions necessary for success in a world of increasingly complex medical regulations.

Alpha Eta Mu Beta. 8401 Corporate Drive, Suite 140, Landover, MD 20785. 301-459-1999, http://www.ahmb.org.

Herbert F. Voigt, National President; Patricia I. Horner, Executive Director. 20 chapters.

Alpha Eta Mu Beta, the National Biomedical Engineering Honor Society, was founded by Daniel Reneau at Louisiana Tech University in 1979. This organization was sponsored by the AEMB. The AEMB was established to mark in an outstanding manner those biomedical engineering students who manifested a deep interest and marked ability in their chosen life work to promote an understanding of their profession and to develop its members professionally.

American Institute for Medical and Biological Engineering. 1901 Pennsylvania Ave NW, Suite 401, Washington, D.C. 20006. 202-496-9660, http://www.aimbe.org. Patricia Ford Roegner, Executive Director. 1000 Fellows; 18 Scientific Organizations; 85 Universities.

Founded in 1991 to establish an identity for the field of medical and biological engineering, which is the bridge between the principles of engineering science and practice and the problems and issues of biological and medical science and practice. The AIMBE comprises four sections. The College of Fellows with over 1000 persons who are the outstanding bioengineers in academia, industry, and government. The Academic Council is 85 universities with educational programs in bioengineering at the graduate or undergraduate level. The Council of Societies is 18 scientific organizations in medical and biological engineering. The Industry Council is a forum for dialog among industry, academia, and government. Principal activities include participation in formulation of public policy, dissemination of information, and education. Affiliated with the International Federation for Medical and Biological Engineering. Annual event each winter in Washington, D.C.

American Association of Engineering Societies. 1828 L Street NW, Suite 906, Washington, D.C. 20036. 202-296-2237, http://www.aaes.org. Thomas J. Price, Executive Director. 26 Engineering Societies.

Founded in 1979 in New York City. Member societies represent the mainstream of U.S. engineering with more than one million engineers in industry, government, and academia. The AAES has four primary programs: communications, engineering workforce commission, international, and public policy. Governance consists of two representatives from each of 26 member societies. Convenes diversity summits, publishes engineering and technology degrees, and holds annual awards ceremony.

American Academy of Environmental Engineers. 130 Holiday Court, Suite 100, Annapolis, MD 21401. 410-266-3311, http://www.aaee.net. David A. Asselin, Executive Director.

The American Sanitary Engineering Intersociety Board incorporated in 1955 became the American Academy of Environmental Engineers in 1966; and in 1973, it merged with the Engineering Intersociety Board. Principal purposes are improving the practice, elevating the standards, and advancing public recognition of environmental engineering through a program of specialty certification of qualified engineers.

American Academy of Orthopaedic Surgeons. 600 North River Road, Rosemont, IL 60018. 847-823-7186, http://www.aaos.org. Karen L. Hackett, Chief Executive Officer. 24,000 Members.

Founded in Chicago in 1933. Provides education and practice management services for orthopedic surgeons and allied health professionals. Maintains a Washington, D.C., office. Annual spring meeting.

American Academy of Orthotists and Prosthetists. 526 King Street, Suite 201, Alexandria, VA 22314. 703-836-0788, http://www.oandp.org. Peter D. Rosenstein, Executive Director. 3000 Members.

Founded in 1970 to further the scientific and educational attainments of professional practitioners in the disciplines of orthotics and prosthetics. Members have been certified by the American Board for Certification in Orthotics and Prosthetics. Annual spring meeting.

American Association of Physicists in Medicine. One Physics Ellipse, College Park, MD 20740. 301-209-3350, http://www.aapm.org. Angela R. Keyser, Executive Director. 4700 Members.

Founded in Chicago in 1958 and incorporated in Washington in 1965. Promotes the application of physics to medicine and biology. Member society of the American Institute of Physics. Annual summer meeting.

American Chemical Society. 1155 Sixteenth Street NW, Washington, D.C. 20036. 800-227-5558, http://www.chemistry.org. Madeleine Jacobs, Executive Director, and CEO. 159,000 Members.

Founded in 1877 in New York City. Granted a national charter by Congress in 1937. Encourages the advancement of chemistry. Semiannual spring and fall meetings.

American College of Nuclear Physicians. 1850 Samuel Morse Drive, Reston, VA 20190. 703-326-1190, http://www.acnponline.org. Virginia M. Pappas, Executive Director. 500 Members.

Established in 1974, the organization provides access to activities that encompass the business and economics of nuclear medicine as they impact nuclear medicine physicians. Semiannual meetings fall and winter.

American College of Physicians. 190 N. Independence Mall West, Philadelphia, PA 19106. 215-351-2600, http://www.acponline.org. John Tooker, Executive Vice President. 119,000 Members.

Founded in New York City in 1915. Merged with the Congress of Internal Medicine in 1923 and merged in 1998 with the American Society of Internal Medicine. Patterned after England's Royal College of Physicians. Members are physicians in general internal medicine and related subspecialties. Maintains a Washington, D.C., office. Annual spring meeting.

American College of Radiology. 1891 Preston White Drive, Reston, VA 20191. 703-648-8900, http://www.acr.org. Harvey L. Neiman, Executive Director. 30,000 Members.

Founded in 1923 in San Francisco and incorporated in California in 1924. Purpose is to improve the health of patients and society by maximizing the value of radiology and radiologists by advancing the science, improving patient service, and continuing education. Annual fall meeting.

American College of Surgeons. 633 N. St. Clair Street, Chicago, IL 60611. 312-202-5000, http://www.facs.org. Thomas R. Russell, Executive Director. 64,000 Fellows, 5000 Associate Fellows.

Founded in 1913 and incorporated in Illinois. U.S. member of the International Federation of Surgical Colleges. Members are Fellows who must meet high standards established by the College. Purpose is to improve the quality of care for the surgical patient by setting high standards for surgical education and practice. Annual fall clinical congress.

American Congress of Rehabilitation Medicine. 6801 Lake Plaza Drive, Suite B-205, Indianapolis, IN 46220. 317-915-2250, http://www.acrm.org. Richard D. Morgan, Executive Director. 1700 Members, 15 Companies.

Founded in 1923 as the American College of Radiology and Physiotherapy. Name changed in 1926 to American College of Physical Therapy and in 1930 to American Congress of Physical Therapy. Changed again in 1945 to American Congress of Physical Therapy and in 1953 became American Congress of Physical Medicine and Rehabilitation. Adopted its current name in 1967. Provides education for professionals in medical rehabilitation. Fall annual meeting.

American Institute of Biological Sciences. 1444 I Street NW, Suite 200, Washington, D.C. 20005. 202-628-1500, http://www.aibs.org. Richard O'Grady, Executive Director. 80 Societies, 6000 Members.

Founded in 1947 as part of the National Academy of Sciences. Incorporated as an independent nonprofit since 1954. Absorbed America Society of Professional Biologists in 1969. Represents more than 80 professional societies with combined membership exceeding 240,000 scientists and educators. Also more than 6000 individual members. Purpose is to better serve science and society. Annual meeting in August.

American Institute of Chemical Engineers. 3 Park Avenue, New York, NY 10016. 212-591-8100, http://www.aiche.org. John Sofranko, Executive Director. 40,000 Members.

Organized in 1908 and incorporated in 1910. Member of Accreditation Board for Engineering and Technology, American National Standards Institute, American Association of Engineering Societies, and other related organizations. Purpose is to advance the chemical engineering profession. Annual meeting in November.

American Institute of Physics. One Physics Ellipse, College Park, MD 20740. 301-209-3131, http://www.aip.org. Marc H. Brodsky, Executive Director, and Chief Executive Officer. 10 Societies and 24 Affiliates.

Chartered in 1931 to promote the advancement of physics and its application to human welfare. Federation of 10 Member Societies representing spectrum of physical sciences.

American Institute of Ultrasound in Medicine. 14750 Sweitzer Lane, Suite 100, Laurel, MD 20707. 301-498-4100, http://www.aium.org. Carmine Valente, Chief Executive Officer.

Began in 1951 at a meeting of 24 physicians attending the American Congress of Physical Medicine and Rehabilitation. Membership includes biologists, physicians, and engineers concerned with the use of ultrasound for diagnostic purposes. Provides continuing education, CME tests, and accreditation of ultrasound laboratories. Annual fall meeting.

American Medical Informatics Association. 4915 St. Elmo Avenue, Suite 401, Bethesda, MD 20814. 301-657-1291, http://www.amia.org. Don Detmer, President, and CEO. 3000 Members.

Founded in 1990 through a merger of three existing health informatics associations. Members represent all basic, applied, and clinical interests in health-care information technology. Promotes the use of computers and information systems in health care with emphasis on direct patient care. Semiannual meetings: spring congress in the West and fall annual symposium in the East.

American Society for Artificial Internal Organs. P.O. Box C, Boca Raton, FL 33429. 561-391-8589, http://www.asaio.net. 1400 Members.

Established in 1955 in Atlantic City, NJ. Annual June conference.

American Society for Engineering Education. 1818 N Street NW, Suite 600, Washington, D.C. 20036. 202-331-3500, http://www.asee.org. Frank L. Huband, Executive Director. 12,000 Members, 400 Colleges, 50 Corporations.

Founded in 1893 as the Society for Promotion of Engineering Education. Incorporated in 1943 and merged in 1946 with Engineering College Research Association. Members include deans, department heads, faculty members, students, and government and industry representatives from all disciplines of engineering and engineering technology. Member of the American Association of Engineering Societies, Accreditation Board for Engineering and Technology, American Institute for Medical and Biological Engineering, and American Council on Education. Participating society of World Federation of Engineering Associations. Purpose is to further education in engineering and engineering technology. Annual June meeting.

American Society for Healthcare Engineering of the American Hospital Association. One North Franklin, 28th Floor, Chicago, IL 60606. 312-422-3800, http://www.ashe.org. Albert J. Sunseri, Executive Director.

Affiliate of the American Hospital Association. Annual June meeting.

American Society for Laser Medicine and Surgery. 2404 Stewart Avenue, Wausau, WI 54401. 715-845-9283, http://www.aslms.org. Dianne Dalsky, Executive Director. 3000 Members.

Founded in 1980 to promote excellence in patient care by advancing laser applications and related technologies. Annual spring meeting.

American Society of Agricultural and Biological Engineers. 2950 Niles Road, St. Joseph, MI 49085. 269-429-0300, http://www.asabe.org. Melissa Moore, Executive Vice President. 9000 Members.

Founded in 1907 as the American Society of Agricultural Engineers and changed its name in 2005. Dedicated to the advancement of engineering applicable to agricultural, food, and biological systems. Annual meeting in July.

American Society of Civil Engineers. 1801 Alexander Bell Drive, Reston, VA 20191. 703-295-6000, http://www.asce.org. Patrick J. Natale, Executive Director. 137,500 Members.

Founded in 1852 as the American Society of Civil Engineers and Architects. Dormant from 1855 to 1867, but it revived in 1868 and incorporated in 1877 as the American Society of Civil Engineers. Over 400 local affiliates, 4 Younger Member Councils, 230 Student Chapters, 36 Student Clubs, and 6 International Student Groups. Semiannual spring and fall meetings.

American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc. 1791 Tullie Circle NE, Atlanta, GA 30329. 404-636-8400, http://www.ashrae.org.

Incorporated in 1895 as the American Society of Heating and Ventilating Engineers, known after 1954 as American Society of Heating and Air-Conditioning Engineers. Merged in 1959 with American Society of Refrigerating Engineers to form American Society of Heating, Refrigerating and Air-Conditioning Engineers. Annual summer meeting.

American Society of Mechanical Engineers. Three Park Avenue, New York, NY 10016. 212-591-7722, http://www.asme.org. Virgil R. Carter, Executive Director. 120,000 Members.

Founded in 1880 and incorporated in 1881. Focuses on technical, educational, and research issues of engineering and technology. Sets industrial and manufacturing codes and standards that enhance public safety. Conducts one of the world's largest technical publishing operations. Semiannual summer and winter meetings.

American Society of Neuroradiology. 2210 Midwest Road, Suite 207, Oak Brook, IL 60523. 630-574-0220, http://www.asnr.org. James B. Gantenberg, Executive Director/CEO. 2700 Members.

Founded in 1962. Supports standards for training in the practice of neuroradiology. Annual spring meeting.

American Society of Safety Engineers. 1800 E. Oakton Street, Des Plaines, IL 60018. 847-699-2929, http://www.asse.org. Fred Fortman, Executive Director. 30,000 Members.

Founded in 1911 as the United Association of Casualty Inspectors and merged with the National Safety Council in 1924, becoming its engineering section. Became independent again in 1947 as the American Society of Safety Engineers and incorporated in 1962. There are 13 practice specialties, 150 chapters, 56 sections, and 64 student sections. Annual spring meeting.

Association for Computing Machinery. 1515 Broadway, New York, NY 10036. 212-626-0500, http://www.acm.org. John R. White, Executive Director. 80,000 Members.

Founded in 1947 at Columbia University as Eastern Association for Computing Machinery and incorporated in Delaware in 1954. Affiliated with American Association for Advancement of Science, American Federation of Information Processing Societies, Conference Board of Mathematical Sciences, National Academy of Sciences-National Research Council, and American National Standards Institute. Advancing the skills of information technology professionals and students. Annual fall meeting.

Association for the Advancement of Medical Instrumentation. 1100 North Glebe Road, Suite 220, Arlington, VA 22201. 703-525-4890, http://www.aami.org. Michael J. Miller, President. 6000 Members.

Founded in 1967, the AAMI is an alliance of over 6000 members united by the common goal of increasing the understanding and beneficial use of medical instrumentation and technology. Annual spring meeting.

Association of Biomedical Communications Directors. State University of New York at Stony Brook, Media Services L3044 Health Sciences Center, Stony Brook, NY 11794. 631-444-3228. Kathleen Gebhart, Association Secretary. 100 Members.

Formed in 1974 as a forum for sharing information; adopted a Constitution and Bylaws in 1979, and incorporated in April 1979 in North Carolina. Members are directors of biomedical communication in academic health science settings. Annual spring meeting.

Association of Environmental Engineering and Science Professors. 2303 Naples Court, Champaign, IL 61822. 217-398-6969, http://www.aeesp.org. Joanne Fetzner, Secretary. 700 Members.

Formerly, in 1972, the American Association of Professors in Sanitary Engineering. Professors in academic programs throughout the world who provide education in the sciences and technologies of environmental protection. Biennial conference in July.

Biomedical Engineering Society. 8401 Corporate Drive, Suite 140, Landover, MD 20785. 301-459-1999, http://www.bmes.org. Patricia I. Horner, Executive Director. 3700 Members.

Founded in 1968 in response to a need to give equal representation to both biomedical and engineering interests. The purpose of the Society is to promote the increase of biomedical engineering knowledge and its use. Member

of American Institute for Medical and Biological Engineering. Annual fall meeting.

Biophysical Society. 9650 Rockville Pike, Bethesda, MD 20814. 301-634-7114, http://www.biophysics.org. Ro Kampman, Executive Officer. 7000 Members.

Founded in 1957 in Columbus, OH, to encourage development and dissemination of knowledge in biophysics. Annual winter meeting.

Health Physics Society. 1313 Dolley Madison Boulevard, Suite 402, McLean, VA 22101. 703-790-1745, http://www.hps.org. Richard J. Burk, Jr, Executive Secretary. 7000 Members.

Founded in 1956 in the District of Columbia and reincorporated in Tennessee in 1969. Society specializes in occupational and environmental radiation safety. Affiliated with International Radiation Protection Association. Annual summer meeting.

Human Factors and Ergonomics Society. P.O. Box 1369, Santa Monica, CA 90406. 310-394-1811, http://www.hfes.org. Lynn Strother, Executive Director. 50 Active Chapters and 22 Technical Groups.

Founded in 1957, formerly known as the Human Factors Society. An interdisciplinary organization of professional people involved in the human factors field. Member of the International Ergonomics Association. Annual meeting in September-October.

Institute for Medical Technology Innovation. 1319 F Street NW, Suite 900, Washington, D.C. 20004. 202-783-0940, http://www.innovate.org. Martyn W.C. Howgill, Executive Director.

The concept was developed in 2003 by leaders in the medical device industry, and the Institute was incorporated and opened its offices in 2004. Purpose is to demonstrate the role, impact, and value of medical technology on health care, economy, and society, for the benefit of patients.

Institute of Electrical and Electronics Engineers. 3 Park Avenue, 17th Floor, New York, NY 10016. 212-419-7900, http://www.ieee.org. Daniel J. Senese, Executive Director. 365,000 Members.

The American Institute of Electrical Engineers was founded in 1884 and merged in 1963 with the Institute of Radio Engineers. Three Technical Councils, 300 local organizations, 1300 student branches at universities, and 39 IEEE Societies including the Engineering in Medicine and Biology Society with 8000 members and meets annually in the fall. Maintains Washington, D.C. office.

Institute of Environmental Sciences and Technology. 5005 Newport Drive Suite 506, Rolling Meadows, IL 60008. 847-255-1561, http://www.iest.org. Julie Kendrick, Executive Director.

Formed by a merger of the Institute of Environmental Engineers and the Society of Environmental Engineers in 1953. Annual spring meeting.

Instrument Society of America. 67 Alexander Drive, Research Triangle Park, NC 27709. 919-549-8411, http://www.isa.org. Rob Renner, Executive Director. 30,000 Members.

Founded in Pittsburgh in 1945. Charter member of American Automatic Control Council, affiliate of American Institute of Physics, member of American Federation of Information Processing Societies, member of American National Standards Institute, and U.S. representative to the International Measurement Confederation. Develops standards, certifies industry professionals, provides education and training, publishes books and technical articles, and hosts largest conference for automation professionals in the Western Hemisphere. Annual meeting in October.

International Biometric Society ENAR. 12100 Sunset Hills Road, Suite 130, Reston, VA 22090. 703-437-4377, http://www.enar.org. Kathy Hoskins, Executive Director. 6500 Members.

Founded in September 1947. Became the International Biometric Society in 1994. Annual March and June meetings.

International College of Surgeons. 1516 North Lake Shore Drive, Chicago, IL 60610. 312-642-3555, http://www.icsglobal.org. Max C. Downham, Executive Director. 10,000 Members.

Founded in Geneva, Switzerland, in 1935 and incorporated in the District of Columbia in 1940. Federation of general surgeons and surgical specialists. Annual spring meeting of U.S. section and biennial international meetings.

International Society for Magnetic Resonance in Medicine. 2118 Milvia Street, Suite 201, Berkeley, CA 94704. 510-841-1899, http://www.ismrm.org. Roberta A. Kravitz, Executive Director. 6,000 Members.

Formed as a merger of the Society for Magnetic Resonance Imaging and Society of Magnetic Resonance in Medicine in 1995. Promotes the application of magnetic resonance techniques to medicine and biology. Annual meetings in April/May.

Medical Device Manufacturers Association. 1919 Pennsylvania Avenue NW, Suite 660, Washington, D.C. 20006. 202-349-7171, http://www.medicaldevices.org. Mark B. Leahey, Executive Director. 140 Companies.

Created in 1992. Supersedes Smaller Manufacturers Medical Device Association. Represents manufacturers of medical devices. Annual May meeting.

Radiation Research Society. 810 East 10th Street, Lawrence, KS 666044. 800-627-0629, http://www.radres.org. Becky Noordsy, Executive Director. 2025 Members.

Founded in 1952 as a professional society of persons studying radiation and its effects. Annual spring-summer meetings.

Radiological Society of North America. 820 Jorie Boulevard, Oak Brook, IL 60523. 630-571-2670, http://

www.rsna.org. Dave Fellers, Executive Director. 37,577 Members.

Founded as Western Roentgen Society and assumed its current name in 1918. Members are interested in the application of radiology to medicine. Holds the largest medical meeting in the world annually in November with more than 60,000 in attendance.

RESNA—Rehabilitation Engineering & Assistive Technology Society of North America. 1700 N. Moore Street, Suite 1540, Arlington, VA 22209. 703-524-6686, http://www.resna.org. Larry Pencak, Executive Director. 1000 Members.

Founded in 1979 as the Rehabilitation Engineering Society of North America. In June 1995, the name was changed to the Rehabilitation Engineering and Assistive Technology Society of North American—RESNA. Twenty-one special interest groups and seven professional specialty groups. Annual meeting in June.

SPIE—International Society for Optical Engineering. P.O. Box 10, Bellingham, WA 98227. 360-676-3290, http://www.spie.org. Eugene G. Arthurs, Executive Director. 14,000 Members, 320 Companies.

Founded in 1956 in California as the Society of Photographic Instrumentation Engineers, it later became the Society of Photo-Optical Instrumentation Engineers and assumed its current name in 1981. Members are scientists, engineers, and companies interested in application of optical, electro-optical, fiber-optic, laser, and photographic instrumentation systems and technology. Semiannual meetings.

Society for Biological Engineering of the American Institute of Chemical Engineers. 3 Park Avenue, New York, NY 10016. 212-591-7616, http://www.bio.aiche.org.

Established by the AIChE for engineers and applied scientists integrating biology with engineering.

Society for Biomaterials. 15000 Commerce Parkway, Suite C, Mt. Laurel, NJ 08054. 856-439-0826, http://www.biomaterials.org. Victoria Elliott, Executive Director. 2100 Members.

Founded in 1974. Promotes biomaterials and their uses in medical and surgical devices. Annual spring meeting.

Society for Biomolecular Screening. 36 Tamarack Avenue, #348, Danbury, CT 06811. 203-743-1336, http://www.sbsonline.org. Christine Giordano, Executive Director. 1080 Members, 230 Companies.

Supports research in pharmaceutical biotechnology and the agricultural industry that use chemical screening procedures. Annual fall meeting.

Society for Modeling and Simulation International. P.O. Box 17900, San Diego, CA 92177. 858-277-3888, http://www.scs.org. Steve Branch, Executive Director.

Established in 1952 as the Simulation Council and incorporated in California in 1957 as the Simulation Councils. Became Society for Computer Simulation in 1973 and later changed its name to the current one. A founding member of the Information Processing Societies and National Computer Confederation Board, and affiliated with American Association for the Advancement of Science. Holds regional simulation multiconferences.

Society of Interventional Radiology. 3975 Fair Ridge Drive, Suite 400 North, Fairfax, VA 22033. 703-691-1805, http://www.sirweb.org. Peter B. Lauer, Executive Director.

Society of Nuclear Medicine. 1850 Samuel Morse Drive, Reston, VA 20190. 703-709-9000, http://www.interactive.snm.org. Virginia Pappas, Executive Director.

Society of Rheology. Suite 1N01, 2 Huntington Quadrangle, Melville, NY 11747. 516–2403, http://www.rheology.org. Janis Bennett, Executive Director.

Permanent address is at the American Institute of Physics and is one of five founding members of the AIP. Members are chemists, physicists, biologists, and others concerned with theory and precise measurement of flow of matter and response of materials to mechanical force. Annual meeting held in October or November.

Professional Societies and Organizations of Other Countries

Pick up from IFMBE Affiliates on www.ifmbe.org.

Further Reading

Biomedical Engineers, Brief 519, G.O.E. 02.02.01; D.O.T. (4th ed.) 019. Chronicle Guidance Publications, Moravia, NY 13118, 1994.

Biomedical Engineers. *Occupational Outlook Handbook, 2004-05 Edition*. Bureau of Labor Statistics, US. Department of Labor. http://www.bls.gov/oco/ocos262.htm.

Boykin D. Biomedical engineering takes center stage. Engineering Times 2004; 26 (9). National Society of Professional Engineers, 1420 King Street, Alexandria, VA 22314.

Collins CC. The retrospectroscope: Notes on the history of biomedical engineering in America. IEEE Eng Med Biol Mag Dec. 1988. IEEE Engineering in Medicine & Biology Society, 445 Hoes Lane, Piscataway, NJ 08854.

Enderle J, editor. Charting the milestones of biomedical engineering. *IEEE Eng Med Biol Mag*, May 2002. IEEE Engineering in Medicine and Biology Society, 445 Hoes Lane, Piscataway, NJ 08854.

Fagette PH je, Homer PI, editor. The Biomedical Engineering Society: An Historical Perspective. Landover, MD: The Biomedical Engineering Society; 2004.

Goodman L. The International Federation for Medical and Biological Engineering—20 years on. Med Biol Eng Comput Jan. 1978. International Federation for Medical and Biological Engineering.

Katona P. The Whitaker Foundation: The end will be just the beginning. IEEE Trans Med Imaging 2002;21(8). IEEE Engineering in Medicine & Biology Society, 445 Hoes Lane, Piscataway, NJ 08854.

Johnson AT. Executive Director's Report: What is ASMBE all about? ASMBE Newslett 2004:1.

Planning a Career in Biomedical Engineering (2004). Biomedical Engineering Society, 8401 Corporate Drive, Suite 140, Landover, MD 20785. http://www.bmes.org

Tompkins W. From the President. IEEE Eng Med Biol Mag, December 1988. IEEE Engineering in Medicine & Biology Society, 445 Hoes Lane, Piscataway, NJ 08854.

MEDICAL GAS ANALYZERS

TADEUSZ M. DRZEWIECKI
JERRY M. CALKINS
Defense Research Technologies, Inc.
Rockville, Maryland

INTRODUCTION

Medical gas monitoring has been so successful in improving patient safety and reducing patient risk that it has become standard medical practice today in every part of medical practice from hospital to home. The argument for providing additional patient safety will continue to be a powerful incentive to improve and enhance the methods and techniques to provide increased knowledge of the monitoring of respiratory and anesthetic gases. Research on gas markers to aid in diagnosis is an equally important application, and with the capability to measure gases at parts per billion or even trillion, heretofore unobserved gases can point to early diagnosis of such nearly always fatal neonatal diseases as necrotizing enterocolitis and other difficult-to-diagnose states.

Medical gas analyzers are used to sample and measure gases of medical importance, such as anesthesia and respiratory monitoring, and detection of trace gases for diagnostic purposes. This article predominantly discusses these two cases. The estimation of arterial blood gases is considered only in terms of measurement of respired gases. Two basic categories of sensor/analyzers exist: continuous and batch. Gas analyzers are further broken down by their sensing mechanisms into two fundamental modes of operation, specific and analytic.

Continuous devices are used where real-time information is needed. Batch systems operate on a bolus of gas, usually when real-time information is not needed. Many applications may have to live with the offline, longer duration of a batch test because nothing else is available.

Specific-type sensors rely on particular physical phenomena that are activated in the presence of a particular gas. Electrochemical devices, for example, are representative of a specific sensor wherein a voltage is developed by a chemical reaction between the sensor material and the gas being analyzed in some identified or known proportion to the amount of gas present. The same is true of fuel cells and other galvanic devices where an electric potential is developed in the presence of a difference in partial pressures across a conducting medium.

Specificity is a major issue and is of particular importance to the medical community. At this point in time, no truly specific sensors exist. All sensors exhibit some form of cross-sensitivity to a variety of gases, some in the same family, some with similar physical properties, and some for extraneous reasons not always obvious to the user. Nitrous oxide (N_2O) and carbon dioxide (CO_2) have practically the same molecular weight, 44.0128 versus 44.0098. Consequently, they have near-identical physical characteristics such as specific heat and viscosity. Interestingly, they also have almost exactly the same absorption wavelengths, although not necessarily because the atomic weights are the same but rather because the orbital electron transition energetics are similar. Thus, it is difficult to distinguish the two with most conventional techniques, and a carbon dioxide sensor that is based on absorption at 4.3 mm will be affected by the presence of nitrous oxide with its peak absorption at 4.5 mm. Unless a very narrow wavelength light source is used, such as a laser, the nitrous oxide will absorb some of the energy and make it appear that more carbon dioxide is present than there really is.

Analytic devices imply an ability to assay a gas or gas mixture and tell the user not only how much of a particular gas is present, but also which gases or elements are present and in what relative quantities, of which optical spectroscopy is a good example where a large number of absorption lines exist for different gases, so one can scan the entire spectrum from ultraviolet (UV) to far infrared (IR) and compare absorption lines to see what is present. This example is interesting because IR spectroscopy can also be the basis for a specific sensor when only a single or a particular wavelength of light is monitored looking only for a gas at that absorption line. Gas chromatography (GC) is also a batch process where a bolus of gas to be assayed is separated into its constituent parts in time by a molecular sieve and the binary pairs (the carrier and the separated gas) are detected and quantized upon exiting the GC column.

Perhaps the most common use of and need for continuous gas analysis or sensing is in real-time respiratory applications where inspired and expired (end-tidal) concentrations of respiratory gases are measured to validate that the appropriate standards of care are being applied and that proper ventilation and oxygenation of a patient is being achieved. An example would be monitoring of a ventilated patient in the ICU or recovery room. In addition, the monitoring of anesthetic gases during and immediately following anesthetic administration in the operating room is a critical application that can mean the difference between life and death. Too much can lead to brain damage or death, and too little can result in unnecessary pain and memory recall. And, of course, the detection and warning of the presence of toxic gases such as carbon monoxide released from desiccated soda lime CO_2 scrubbers due to interactions between the anesthetic agents and various scrubbers, or the production of the highly toxic Compound A, is critical to patient safety.

Continuous medical gas monitoring provides the clinician with information about the patient's physiologic status, estimates of arterial blood gases, verifies that the appropriate concentrations of delivered gases are administered, and warns of equipment failure or abnormalities in the gas delivery system. Monitors display inspired and expired gas concentrations and sound alarms to alert clinical personnel when the concentration of oxygen (O_2), carbon dioxide (CO_2), nitrous oxide (N_2O), or volatile anesthetic agent falls outside the desired set limits.

Medical gas analysis has been driven by a need for safety and patient risk reduction through respiratory gas analysis and identification and quantification of volatile anesthetic vapors. Perhaps one of the earliest anesthetic agent sensors was the Drager Narkotest, which comprised a polymer rubber membrane that contracted and moved a needle as it absorbed agent. It did not require

an electrical power supply and its slow response was not any slower than the rate of change of gas composition. Much has transpired since those early days.

Currently, numerous methods and techniques of gas monitoring are in place, and new techniques and paradigms are constantly being developed to meet a new need or to serve the community with better performance or lower cost. In this review, many of the intrinsic advantages and disadvantages of these methods and techniques are discussed. A brief comparison, which includes stand-alone and multioperating room gas monitors that can determine concentrations of anesthetic and respiratory gases in the patient breathing circuit during anesthesia, is also presented.

Much of the research and development of these monitors have followed the long use of similar detector principles from analytical chemistry. As a result of the fast pace of sensor development, to a great extent driven by the need for hazardous gas sensors in the face of terrorist threats and being spearheaded by agencies such as the Defense Department's Defense Advanced Research Projects Agency (DARPA), an attempt is made to cover the most common systems and provide insights into the future based on solid technological developments.

The current development of gas analyzers is described in the extensive anesthesia and biomedical engineering literature. Complete and specific historical information about the principles and applications of these devices is well reviewed in several texts [e.g., Ref. (1)], manufacturers' and trade publications [(2) (ECRI)], and an extensive open literature describing equipment and operating principles, methods, and techniques that is available on the Internet. Societies and professional associations also exist that deal with just one method of gas analysis that can provide in-depth information about their particular interests. The Chromatographic Society is one such organization. It is the purpose of this article to concisely summarize such a large selection of information sources to a manageable few, but with enough references and pointers to allow even the casual reader to obtain whatever relevant information at whatever level is required.

CURRENT GAS MONITOR METHODS AND TECHNIQUES

As a result of the chemically diverse substances to be measured, medical gas analyzers commonly combine more than one analytical method. Methods of interest to the medical practitioner, clinician, researcher, or operator include, in alphabetical order:

- Colorimetry
- Electrochemistry
  - Fuel cells
  - Polarography
- Gas chromatography
  - Flame ionization
  - Photoionization Detectors
  - Thermal conductivity
- Infrared/Optical Spectroscopy
- Luminescence/fluorescence
- Mass spectrometry
- Nuclear Magnetic Resonance
- Paramagnetism
- Radioactive ionization
- Raman Laser Spectroscopy
- Solid-state sensors
  - Semiconductor metal oxides
  - ChemFETs
  - Solid-state galvanic cells
  - Piezoelectric/Surface Acoustic Wave

Each of these methods will be described in the following text. Illustrative examples of typical devices may be mentioned.

COLORIMETRY

Colorimetry is one of the oldest methods of gas analysis that is typically used to detect the presence of carbon dioxide in a breath as a means of determining if proper tracheal intubation has been performed. Basically, it works on the principle of changing the color of a material such as paper or cloth impregnated with a reagent in the presence of a known analyte. An example is the changes in litmus paper from purple to red in the presence of an acid and blue in the presence of a base. Carbon dioxide (CO_2) in the presence of water vapor in the exhaled breath produces carbonic acid, which turns the litmus paper toward red, usually some shade of pink. The degree of color change can be quite subjective, but a color scale usually accompanies most devices so that a coarse estimate, roughly $\pm 0.5\%$ CO_2 by volume, can be made. More expensive, sophisticated devices offer an electronic colorimetric analyzer that does the comparison automatically and can even provide a digital output.

Reagents may be tuned to a variety of specific gases and are quite commonly used in the semiconductor business to monitor levels of hydride gases to include arsine and phosphene. Hydrogen sulfide (H_2S) is also a common analyte for colorimetric sensors (1).

Cross-sensitivity can be additive or subtractive with colorimetric devices. For example, a colorimetric capnometer (CO_2 sensor) may register false-positives and false-negatives. Color may change in the presence of acidic reflux or the ingestion of acidic liquids (lemonade, wine, etc.). Conversely, the presence of bases could negate the acid response, which is true in other applications as well. For example, in hydrogen sulfide detection, the presence of methyl mercaptan (CH_4S) compounds can cancel out any reading of H_2S. Clearly, any chemical reactions between the selected analyte and a reactive contaminant can affect readings one way or another.

Figure 1 shows a photograph of one of the newer colorimetric capnometers on the market manufactured by Mercury Medical (www.mercurymed.com). A color-changing tape used in this device turns yellow in the

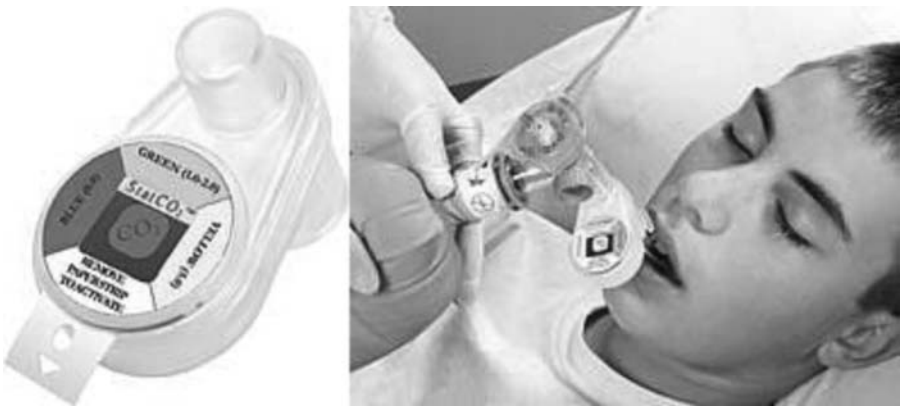


Figure 1. Mercury medical colorimetric end tidal CO_2 detector.

presence of CO_2, but returns to green when no CO_2 is present. The same piece of paper will register changes as the patient breathes. As the tape is consumed, it can be pulled through the device, and a fresh piece exposed to the breath.

ELECTROCHEMISTRY

Electrochemical gas sensors operate on the principle that a current is generated when the selected gas reacts at an electrode in the presence of an electrolyte, not unlike a battery. For this reason, these devices are often called amperometric gas sensors or microfuel cells. Electrochemical gas sensors are found ubiquitously in industry because of their excellent sensitivity to toxic gases (often low parts per million, ppm) and relatively low cost. They are, however, consumable devices and have a limited operating and shelf life, typically a few years. They are not particularly susceptible to poisoning, that is, being contaminated or degraded by absorption of particular contaminant gas species. Gases of medical importance that can be sensed with electrochemical devices are ammonia, carbon monoxide, nitric oxide, oxygen, ozone, and sulfur dioxide. The most prominent medical use is in the measurement of oxygen.

Three basic elements exist in any electrochemical gas sensor. The first is a gas-permeable, hydrophobic membrane, which allows gas to diffuse into the cell but keeps the liquid or gel electrolyte inside. It is the slow diffusion process that limits the time response of these sensors, although, if they are made very small, they can be quite responsive. Typically, however, an oxygen sensor may take as long as 20 s to equilibrate, making these devices impractical for real-time monitoring of respiration other than monitoring some average oxygen level.

The second element is the electrode. Selection of the electrode is critical to the selectivity of an appropriate reaction. Typically, electrodes are catalyzed noble metals such as gold or platinum. Gases such as oxygen, nitrogen oxides, and chlorine, which are electrochemically reducible, are sensed at the cathode while those that are electrochemically oxidizable, such as carbon monoxide, nitrogen dioxide, and hydrogen sulfide, are sensed at the anode.

The third element is the electrolyte that carries the ions between the electrodes. The electrolyte must be kept encapsulated in the cell as leakage would cause dysfunction.

In many cases, a fourth element exists which is a filter/scrubber mounted across the face of the sensor and the permeable membrane, which helps with the specificity by eliminating some interfering gases. In an oxygen sensor, this element is often an activated charcoal molecular sieve that filters out all but carbon monoxide and hydrogen. Other filters can be tailored to allow only the selected analyte through.

An oxygen fuel cell gas detector uses a lead anode that is oxidized during operation. It is powered by the oxygen it is sensing with a voltage output proportional to the oxygen concentration in the electrolyte. In this case, the electrolyte is potassium hydroxide (KOH) solution. With the recent explosion in research on fuel cells, they have become almost ubiquitous in medical practice, supplanting the Clark electrodes to a great extent.

Among the most recognizable oxygen-sensing devices are the Clark electrodes (Ag/AgCl anode, Pt cathode). One of the first applications of this device was monitoring oxygen concentrations in the inspiratory limb of the breathing circuit of an anesthesia machine. Clark electrodes differ slightly from the above-described fuel-cell-type devices. Clark electrodes require an external voltage source to generate a bias voltage against which the oxygen-induced potential operates. These devices are, therefore, called polarographic because of the bias voltage that is required for its operation, contrasting with a galvanic or amperometric cell, which produces current on its own proportional to the amount of analyte present. Oxygen from the sample fluid equilibrates across a Teflon membrane with a buffered potassium chloride (KCl) solution surrounding a glass electrode. The electrode has a platinum cathode and a silver/silver chloride anode. With between 0.5 V and 0.9 V applied across the electrodes, the consumption of O_2 at the cathode, and hence the current in the circuit, is dependent on the O_2 concentration in the solution, which rapidly equilibrates with the sample. In practice, 0.68 V is used. Performance is adversely affected by the presence of N_2O and halogenated anesthetic agents such as halothane. Protection of the platinum

Figure 2. Datex-Ohmeda Smart Vent 7900 ventilator monitor uses a galvanic cell O_2 sensor.

cathode and the need for semipermeable membranes reduces their effectiveness.

Fuel cell and polarographic devices both require temperature and pH compensation and, as indicated before, have limited life spans because of the consumable nature of the reaction and the propensity of the permeable membranes to eventually lose the effectiveness of the Clark electrodes.

Datex-Ohmeda, one of the largest manufacturers of anesthesia equipment, sells many of their systems, included with which is a galvanometric fuel cell oxygen sensor. Figure 2 shows a Smart Vent 7900^{TM} (3) display showing the level of oxygen being delivered. The oxygen sensor life is specified at 18 months.

GAS CHROMATOGRAPHY

Gas chromatography (GC) is an analytic tool that provides the user with an assay of what is in the gas sample of interest. It consists of two parts: (1) separation of different species by differential travel times through a separation column and (2) analytic detection of the quantity of each analyte at the end of the column using any one of a variety of methods. That is, GC provides a quantification of the constituent gases as well as an identification of what the constituent gases are. It is actually a very simple process, and, were it not for the relatively long analysis time, usually on the order of several minutes to as long as an hour, and its batch nature, it would be used more frequently.

GC requires the separation of a gas sample into its constituent gases, which is accomplished by mixing the sample with a carrier gas, usually some inert gas like helium or nitrogen, which is not adsorbed by the GC column, and passing it through a column or long impermeable, nonreacting (e.g., stainless steel) capillary filled with a zeolite, molecular sieve, or other material that separates the sample gases according to their physical or chemical properties. The different gas constituents travel through the column at different speeds and exit as binary pairs (a constituent and the carrier) in order of their adsorption properties. The time it takes for a constituent to exit the column identifies the constituent. The capillary columns can be as short as a few centimeters to tens and even hundreds of meters long. The capillary columns are heated to maintain a constant temperature, as the transport characteristics tend to be temperature-dependent.

Calibration is required to determine the transit times for each analyte, which is done by injecting known gas samples at the start of the column and physically timing them at the exit. At the exit of the column, a gas detector exists usually a flame ionization or thermal conductivity detector that measures the amount of constituent relative to the carrier. After all the constituents have been accounted for, the assay of the original sample can be made by summation of the relative amounts of each constituent and then taking the ratio of each constituent relative to the summation, which then gives the concentrations and the final assay.

Usually, this summation is accomplished by summing the areas under the detector output peaks and then ratioing the areas under the individual peaks relative to the total area. The choice of gas chromatographic detectors depends on the resolution and accuracy desired and includes (roughly, in order from most common to the least): the flame ionization detector (FID), thermal conductivity detector (TCD or hot wire detector), electron capture detector (ECD), photoionization detector (PID), flame photometric detector (FPD), thermionic detector, and a few more unusual or expensive choices like the atomic emission detector (AED) and the ozone- or fluorine-induced chemiluminescence detectors.

The Flame Ionization Detector (FID) is widely used to detect molecules with carbon-hydrogen bonds and has good sensitivity to low ppm. Basically, in operation, the analyte is injected into a hydrogen carrier and ignited inside a grounded metal chamber. Hydrogen is used because it burns clean and is carbon-free. The latter is important because output is proportional to the ionized carbon atoms. An electrode is situated just above the flame and a voltage potential is applied. The current produced is proportional to the number of carbon atoms in the analyte. When applied at the exit of a GC, very accurate measures of hydrocarbon gases can be made.

Photoionization Detectors (PIDs) are used to detect the volatile organic compound (VOC) outputs of GCs but are also widely used in industry and science to detect environmental and hazardous gases. They operate on a similar principle to that of the FID, but use ultraviolet light (UV) as opposed to a flame to ionize the flowing gas between insulated electrodes. As UV energy is a much higher frequency (lower wavelength) than IR or visible light, it can be larger and, consequently, can readily ionize gases. The ionization potentials of the analyte gases are matched by adjusting the frequency of the emitted light. The output power of the lamp is roughly the product of the number of photons and the energy per photon divided by the area and time, although changing the output frequency will change the photon energy ($E = h\upsilon$), thereby changing the power. The output power can be changed independently by increasing the fluence of photons.

Figure 3. Seito ToxiRae personal PID gas monitor.

An inert gas lamp provides the UV light, (e.g., xenon lamps emit UV light at 147.6 nm, krypton at 123.9 nm, and argon at 105.9 nm). An advantage is that the sensitivity to particular species or groups of compounds can be adjusted by adjusting the output power to match the distinct ionization potentials of analyte gases. Consequently, different sensor sets can be achieved. For example, amines, aromatic compounds, and benzene are highly detectable at 9.5 eV. Disease and other anomaly marker gases often found in the breath, such as acetone, ammonia, and ethanol, are detectable at 9.5 eV as well as 10.6 eV. Other, more complex, marker gases such as acetylene, formaldehyde, and methanol can be detected at 10.6 eV and 11.7 eV. Typically, the PID devices are fairly responsive, on the order of a few seconds, and do well with moderately low concentrations (e.g., 0.1 ppm isobutylene).

One of the nice things about PIDs is that they can be made very small in size, as shown in Fig. 3, which shows the Rae Systems, Inc. ToxiRae personal gas monitor (www.raesystems.com). Depending on the UV source, CO, NO, SO_2, or NO_2 can be read. It works with rechargeable batteries.

Thermal Conductivity Detectors (TCD) are used to detect and quantify gases that have large variations in thermal conductivity. Gases that are discriminated well are sulfur dioxide and chlorine, which have roughly one-third the conductivity of air to helium and hydrogen, which have six and seven times the conductivity of air. As heat transfer depends on three mechanisms, radiation, convection, and conduction, the actual TCD sensor itself must be designed in such a way that conduction dominates, which implies a very slow, constant, moving flow to minimize or stabilize convection effects and a radiation-shielded enclosure. Most arrangements use two identically heated coils of wire comprising two legs of a Wheatstone bridge, one coil in a reference gas tube and the other in the sample tube. When the thermal conductivity of the sample increases, the sample coil is cooled more than the reference, and its resistance changes (usually decreases), thereby generating a voltage difference across the bridge. TCDs are usually used in high concentration applications, as they do not have the sensitivity of other techniques. TCDs do very well when mounted at the exit of a GC where the separated gas analytes are expected to have large variations in thermal conductivity.

Gas chromatographs have come a long way over the last decade as far as size and cost are concerned. Although laboratory-grade devices such as the HP stand-alone system shown in Fig. 4 still are fairly common, portability is being stressed in order to get the almost incomparable

Figure 4. An HP/Agilent laboratory-grade gas chromatograph.

detectibilty of the GC to where the real gas problems exist, such as in the emergency or operating rooms, in the field, and at sites where toxins and suspected hazardous gases may be present. Bringing the GC to the patient or taking data without having subjects come into the lab has spawned systems such as Mensanna's VOC (volatile organic compound) GC system that uses a PID (Fig. 5) to check trace gases in the breath, and HP's has introduced a briefcase-sized micro-GC (Fig. 6). Lawrence Livermore National Laboratory has taken the recent developments in micromachining, MEMS (micro-electromechanical systems), and microfluidics and developed a real micro-GC. Researchers at LLNL (4) have micro-machined a very long capillary on a silicon chip, which serves as the separating column.

Figure 5. Mensanna portable GC for measuring breath VOCs.

Figure 6. Agilent (formerly HP) Micro-GC.

Figure 7 shows the implementation of this device that is reported to have a response time of less than 2 min.

INFRARED/OPTICAL SPECTROSCOPY

Gases absorb light or photon energy at different wavelengths depending on the complexity of their molecular structure. When a molecule is subjected to light energy of a particular frequency, the atoms involved in the atomic bonds will vibrate at the light frequency. If the frequency matches their resonant frequency, or a harmonic, they will resonate, thereby becoming highly absorbent as more of the light energy is used to feed the motion of the resonating molecules. The more complex a molecule, the greater number of different atomic bonds it will have and, consequently, the more absorption frequencies it will have. Table 1 provides some guidelines for absorption for different molecules.

Most infrared analyzers measure concentrations of volatile fluorocarbon halogenated anesthetic agents, carbon dioxide, and nitrous oxide using nondispersive infrared (NDIR) absorption technology. The transduction means may differ. Most use an electronic IR energy detector of one sort or another, such as a bolometer, solid-state photon detectors, and thermopiles; however, one monitor uses

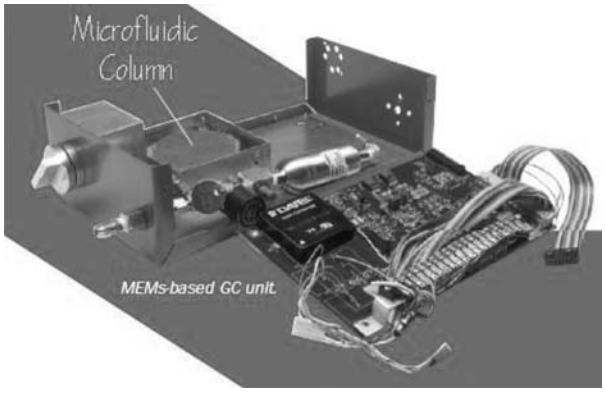


Figure 7. LLNL MEMS-based micro-GC.

Table 1. Infrared Absorption Bands for Gases of Medical Interest

Wavelength (microns)	Elements of Atomic Bonds	Typical Gases
2.7–4 μm	X-H (X = C, N, O, S)	H_2O, CH_4,
4.3–5 μm	C-X (X = C, N, O)	CO_2, N_2O
5.26–6.6 μm	C-X (X = C, N, O)	fluorocarbons
7.7–12.5 μm	C-X (X = C, N, O)	fluorocarbons

another IR detection principle, photoacoustic spectroscopy, based on the level of sound produced when an enclosed gas is exposed to pulsed/modulated IR energy.

Infrared analyzers have been used for many years to identify and assay compounds for research applications. More recently, they have been adapted for respiratory monitoring of CO_2, N_2O, and halogenated anesthetic agents.

Dual-chamber NDIR spectrometers pass IR energy from an incandescent filament through the sample chamber and an identical geometry but air-filled reference chamber. Each gas absorbs light at several wavelengths, but only a single absorption wavelength is selected for each gas to determine the gas concentration. The light is filtered after it passes through the chambers, and only that wavelength selected for each gas is transmitted to the detector. The light absorption in the analysis chamber is proportional to the partial pressure (concentration) of the gas. Most manufacturers use a wavelength range around 3.3 μm, the peak wavelength at which the hydrogen-carbon bond absorbs light, to detect halogenated anesthetic hydrocarbons (halothane, enflurane, isoflurane, etc.).

In one monitor that identifies and quantifies halogenated anesthetic agents, the analyzer is a single-channel, four-wavelength IR filter photometer. Each of four filters (one for each anesthetic agent and one to provide a baseline for comparison) transmits a specific wavelength of IR energy. Each gas absorbs differently in the selected wavelength bands so that the four measurements produce a unique signature for each gas. In another monitor, potent anesthetic agents are assessed by determining their absorption at only three wavelengths of light. Normally, only one agent is present so this process reduces totagent ID. However, the use of "cocktails," mixtures of agents, usually to reduce undesired side effects of one or another agent, require very special monitoring because of the possibility of accidental overdosing.

The Datex-Ohmeda Capnomac (www.us.datex-ohmeda.com), a multigas anesthetic agent analyzer, is based on the absorption of infrared radiation. This unit accurately analyzes breath-to-breath changes in concentrations of CO_2, NO_2, and N_2O and anesthetic vapors. It is accurate with CO_2 for up to 60 breaths/min, and 30 breaths/min for O_2 (using a slower paramagnetic sensor), but N_2O and anesthetic vapors show a decrease in accuracy at frequencies higher than 20 breaths/min. The use of narrow wave-band filters to increase specificity for CO_2 and N_2O makes the identification of the anesthetic vapors, which are measured in the same wave band more difficult. It is interesting to note that IR spectroscopy can also be used on

liquids, as exemplified by the Inov 3100 near-infrared spectroscopy monitor that has been offered as a monitor for intracerebral oxygenation during anesthesia and surgery. Studies with this monitor indicate that it needs a wide optode separation and the measurements are more likely those of the external carotid flow rather than the divided internal carotid circulation (5).

A subset of NDIR is photoacoustic spectroscopy, which measures the energy produced when a gas expands by absorption of IR radiation, which is modulated at acoustic frequencies. A rotating disk with multiple concentric slotted sections between the IR source and the measurement chamber may be used tmodulate the light energy. The acoustic pressure fluctuations created occur with a frequency between 20 and 20,000 Hz, producing sound that is detected with a microphone and converted totan electrical signal. Each gas (anesthetic agent, CO_2, N_2O) exhibits this photoacoustic effect most strongly at a different wavelength. This method cannot distinguish which halogenated agent is present, however. The microphone detects the pulsating pressures from all four gases simultaneously and produces a four-component magnetic signal. A monitor using IR photoacoustic technology has been developed that can quantify all commonly respired/anesthetic gases except N_2 and water vapor. Similarly, a microphone detects the pulsating pressure changes in a paramagnetic oxygen sensor (magnetoacoustics).

The Bruel & Kjaer Multigas Monitor 1304 (6) measurements use photoacoustic spectroscopy and also incorporate a pulse oximeter. It has some advantages over the Datex Ohmeda Capnomac because it uses the same single microphone for detection of all gases, displaying gas concentration in real-time.

With the development of both fixed frequency and tunable solid-state lasers, a revolution in IR spectroscopy has occured with new technical approaches appearing every year. Tuned diode laser spectroscopy, and Laser-induced Photo Acoustic Spectroscopy (a DARPA initiative) are developments that bear close watching as they mature. The ability to produce IR energy at extremely narrow bandwidths allows discrimination of very closely related gas species such as CO_2 and N_2O. In addition, most of the volatile anesthetic agents such as halothane, desflurane, isoflurane, and sevoflurane can also be thus distinguished.

An advantage of NDIR is that the sensing mechanism does not interfere or contact the sample, thus minimal chance exists that the sample would be affected. The costs of these devices have continued to decrease, with numerous companies competing to keep the prices low and attractive. Disadvantages are the susceptibility to dirt and dust in the optical path and cross-sensitivities to interfering gases.

Companies marketing anesthesia and respiratory mechanics monitors are involved in either development or promotion of NDIR. Figure 8 shows a Datex-Ohmeda Capnomac Ultima that uses NDIR for CO_2, N_2O and anesthetic analysis, and agent identification. The top waveform is the plethysmograph, the next down is the O_2 (measured with a fast paramagnetic sensor), and the bottom waveform is the capnographic CO_2 waveform. As an added capability beyond gas analysis, to the right of the capnogram is the pressure-volume loop that is used to assess the lung compliance.

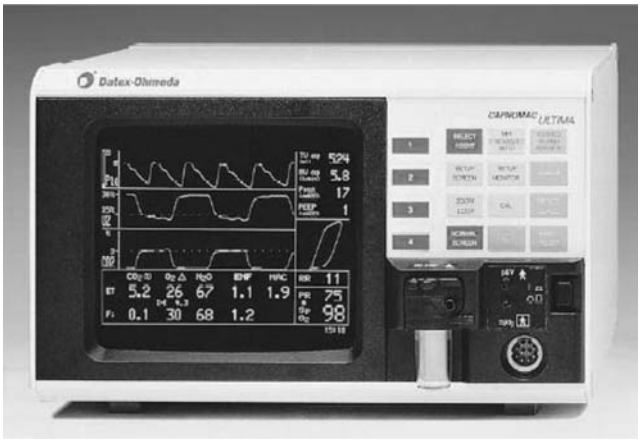


Figure 8. A Datex-Ohmeda Capomac Ultima multiple gas analyzer.

Another limitation of NDIR is its relatively low sensitivity due, for the most part, to the short path length over which the IR energy is absorbed. The short path length and small chamber size is dictated by the need for fast response in order to be able to monitor human physiological and respiratory response.

Breath rates are normally about 10 breaths per min (bpm) but, under acute hyperventilation conditions, can reach 100 bpm and higher. Also, neonates and small animals naturally exhibit high breathing rates, which requires a sensor with a millisecond response in order to be able to detect breath-by-breath variations. However, in cases where a fast response is not the driving factor, the sampling chamber may be lengthened or the path length increased.

Notwithstanding this limitation, the recent invention of Cavity Ring-Down Spectroscopy (CRDS) by Princeton chemist Kevin Lehmann (7,8) is based on absorption of laser energy over huge path lengths but is extremely fast. By bouncing laser energy between near-perfect mirrors in a sample or test chamber, the light can pass through the gas of interest multiple times, often for total distances of up to 100 km, which creates the opportunity to detect miniscule trace amounts of the gas species of interest. The time it takes for the light energy to get attenuated to zero provides a measure of the amount of the gas species present. The shorter the Ring-Down time, the more of the gas is present. These times are, however, on the order of only milliseconds. In fact, Tiger Optics of Warrington, PA, the company that has licensed Dr. Lehmann's technological development, claims that trace gases can be detected in the hundreds of parts per trillion. The LaserTrace multi-point, multi-species, multi-gas analyzer (shown in Fig. 9) is capable of detecting many species such as H_2O, O_2, CH_4, H_2, CO, NH_3, H_2S, and HF. The O_2 module measures down to 200 parts-per-trillion (ppt), in milliseconds, noninvasively and can readily be adapted to respiratory breath measurements. Methane can be detected at the parts per billion level.

Figure 9. Tiger Optics Cavity Ring-Down Spectrometer.

Although of interest in the detection and quantification of oxygen, spectroscopy has not been widely considered for this application. However, an oxygen absorption line exist in the center of the visible spectrum at 760 nm. Often neglected because of the problems associated with the narrowness of the absorption band (0.01 nm versus 100 nm for CO_2 in the IR) as well as with spurious interference from visible light, it is nonetheless an opportunity because no interference exists from any other gases of medical interest. The development of low cost, very narrow-band lasers has resulted in the successful introduction of Laser-based Absorption Spectroscopy from Oxigraf (www.oxigraf.com). With 100 ms response and $\pm 0.02\%$ resolution traces, such as that shown in Fig. 10, are possible. Cost is relatively low in comparison with other technical approaches with similar capabilities.

LUMINESCENCE/FLUORESCENCE

Gas sensors that use luminescence or fluorescence basically take advantage of the phenomenon of excitation of a molecule and the subsequent emission of radiation. Photoluminescence implies optical excitation and re-emission of light at a different, lower frequency. Chemiluminescence implies the emission of light energy as a result of a chemical reaction. In both cases, the emitted light is a function of the presence of the gas species of interest and is detected by optical means. Most industrial sensors use photomultiplier tubes to detect the light, but the needs of the medical community are being met with more compact fiber-optic systems and solid-state photodetectors.

Fluorescence quenching is a subset of luminescence-based sensors, the difference being that the presence of the analyte, rather than stimulating emission of light, actually diminishes the light output. Fundamentally, fluorescence occurs when incoming light excites an electron in a fluorescent molecule to a higher energy state, and, in turn, when the electron returns to its stable state, it releases energy in the form of light.

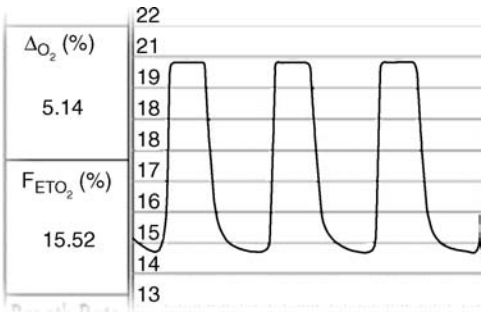


Figure 10. Oxigraf Fast Oxygen Sensor trace of respired O_2.

Two important characteristics of fluorescence are that the light necessary to excite a fluorescent molecule has a shorter wavelength than that of the fluorescent emission, and that the fluorescence of a particular molecule may be suppressed (quenched) or enhanced (dequenched) by the presence of one or more specific molecules. Consequently, the presence of such other molecules (called analytes) may be detected.

A few companies exist that use one form or another of luminescence sensing, in particular, of oxygen in the medical arena, although the sensors are ubiquitous for other gases. For example, Ocean Optics (www.oceanoptics.com) FOXY Fiber Optic Oxygen Sensors use the fluorescence of a ruthenium complex in a sol-gel to measure the partial pressure of oxygen. First, a pulsed blue LED sends light, at ~475 nm, to and through an optical fiber, which carries the light to the probe. The distal end of the probe tip consists of a thin layer of a hydrophobic sol-gel material. A ruthenium complex is trapped in the sol-gel matrix, effectively immobilized and protected from water. The light from the LED excites the ruthenium complex at the probe tip and the excited ruthenium complex fluoresces, emitting energy at ~600 nm. When the excited ruthenium complex encounters an oxygen molecule, the excess energy is transferred to the oxygen molecule in a nonradiative transfer, decreasing or quenching the fluorescence signal. The degree of quenching is a function of the level of oxygen concentration pressure in the film, which is in dynamic equilibrium with oxygen in the sample. Oxygen as a triplet molecule is able to efficiently quench the fluorescence and phosphorescence of certain luminophores. This effect is called "dynamic fluorescence quenching." When an oxygen molecule collides with a fluorophore in its excited state, a nonradiative transfer of energy occurs. The degree of fluorescence quenching relates to the frequency of collisions and, therefore, to the concentration, pressure, and temperature of the oxygen-containing media. The energy is collected by the probe and carried through an optical fiber to a spectrometer where an analog-to-digital (A/D) converter converts the data to digital data for use with a PC.

MASS SPECTROSCOPY

Mass spectroscopy provides, what many consider, the best accuracy and reliability of all of the gas analyzing/assaying schemes. The basic concept is to assay the analyte by reducing it into ionized component molecules and separating them according to their mass-to-charge ratio. By this technique, the constituents of the sample gas are ionized. The resulting ions are accelerated through an electrostatic field and then passed through a deflecting magnetic field. The lighter ions will deflect more than the heavier ions. Detecting the displacement and counting the ions can achieve the assay of the gas sample.

Ion detectors usually comprise an electric circuit where the impinging ions generate a current, which can be measured with a conventional circuit. The more current, the more ions are impinging. In practice, the ionization must be conducted in a high vacuum of the order of 10^{-6} torr. The gas is ionized with a heated filament, or by other means as have been discussed before.

A number of different mass spectroscopic configurations have been developed over the years. Time-of-flight (TOF) systems differ from the magnetically deflected devices in that the ions are free to drift across a neutrally charged evacuated flight chamber after having been accelerated electrostatically by a series of gratings that separate the ions. The time it takes for the ions to travel across the chamber is a function of their mass. An output not unlike that of a GC is developed. Quadrupole mass spectrometers focus the ions through an aperture onto a quadrupole filter. The ion-trap mass spectrometer traps ions in a small volume using three electrodes. An advantage of the ion-trap mass spectrometer over other mass spectrometers is that it has a significantly increased signal-to-noise ratio because it is able to accumulate ions in the trap. It also does not require the same kind of large dimensions that the TOF and magnetically deflected devices need, so, as a consequence, it can be made in a fairly compact size. Finally, the Fourier-transform mass spectrometer takes advantage of an ion-cyclotron resonance to select and detect ions. Single-focusing analyzers use a circular beam path of 180°, 90°, or 60°. The various forces influencing the particle separate ions with different mass-to-charge ratios. Double-focusing analyzers have an electrostatic analyzer added to separate particles with difference in kinetic energies.

A particular advantage of the mass spectrometer is that it can operate with an extremely small gas sample and can detect minute quantities. Response was long compared with most continuous sensors, but with the development of high speed microprocessors, analysis times have steadily decreased to where, today, it is not unusual to have assays in less than one minute. With the development of MEMS TOF devices, the time-of-flight is measured in microseconds.

Mass spectrometers have always tended to be bulky and expensive and, thus, rarely used on a single patient basis. Multiplexing up to 30 patients using a complex valving switching system has been shown to be feasible and has made the system much more cost-effective. Figure 11 shows a conventional ThermoElectron laboratory-grade mass spectrometer setup.

The move to miniature mass spectrometers has been rapid over the last decade, from a suitcase-sized miniature TOF mass spectrometer, developed at Johns Hopkins Applied Physics laboratory (Fig. 12) (9), to a micro-electro-mechanical system (MEMS) device smaller than a penny, developed in Korea at the MicroSystems Lab of Ajou University (Fig. 13) (10).

Mass spectroscopy is often used as the detector in combination with gas chromatography to enhance the sensitivity down to ppb, because in a GC, detection limits the capability.

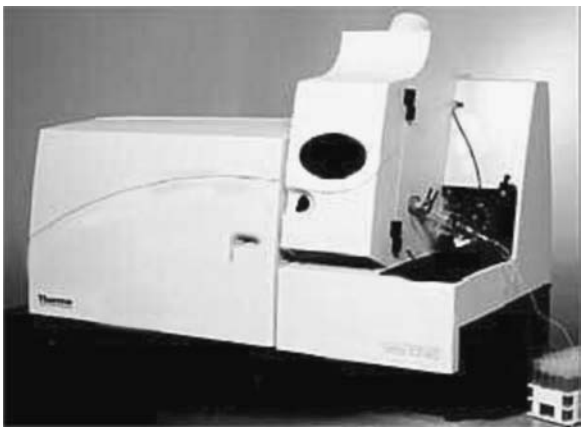


Figure 11. A ThermoElectron laboratory-grade quadrupole mass spectrometer.

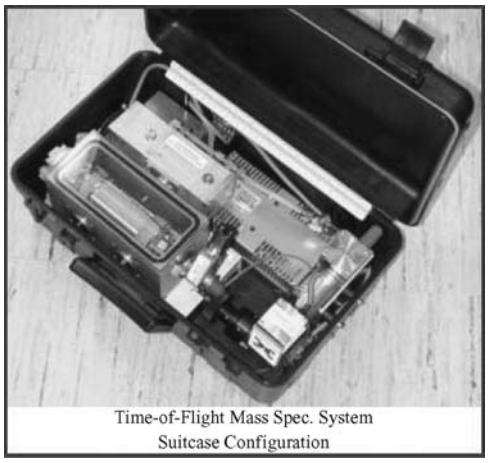


Figure 12. JHU Teeny mass spectrometer.

NUCLEAR MAGNETIC RESONANCE

Nuclear Magnetic Resonance (NMR) is the process by which a relatively low intensity radio-frequency (RF) signal at the resonant frequency of the species of gas of interest interacts with the atoms in a gas and aligns them momentarily, which requires some energy. When the RF

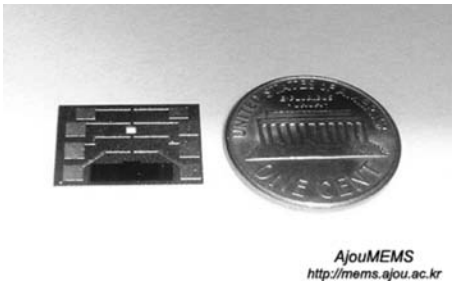


Figure 13. MEMS TOF mass spectrometers developed at Ajou University in Korea.

signal is removed, the atoms release the energy that had been stored in the alignment resonance, return to their chaotic orientations, and re-emit an RF signal at the same resonant frequency at which they were excited. Each atomic bond has its own characteristic frequency so that a spectroscopic analysis can be made by scanning through a large number of frequencies. A variant of NMR, nuclear quadrupole resonance (NQR) is used for detecting explosive vapors, which could be very useful in military medicine where explosives in clothing during triage pose a major hazard. The hydrogen bonds in TNT have a resonance at $\sim$760 kHz and the vapors of plastic explosives have resonances at low MHz frequencies. NMR is very attractive because gas analysis can be performed without having to physically take a sample of the analyte in question. As the initiating and re-emitted RF signals can pass through most nonferrous materials with little attenuation, gas, liquid, and solid chemical species can be interrogated noninvasively. By summing the returns from a number of signals, sensitivity to the low ppm can be achieved.

In dirty environments where optical or infrared devices suffer significant degradation of performance, NMR is particularly useful. Compared with chromatographic approaches, NMR eliminates the need for solvents, columns, carrier gases, or separations. Also, NMR analysis can be performed in real-time because typical atomic relaxation times, the time it takes for the atomic spin axes to return to their original orientations, is on the order of milliseconds for many gases of interest.

PARAMAGNETISM

Many gases exhibit magnetic sensitivity, paramagnetism, due to their polar nature, which means that they are attracted to a magnetic field. For oxygen, its paramagnetic sensitivity is due to its two outer electrons in unpaired orbits. Most of the gases used in anesthesia are repelled by a magnetic field (diamagnetism).

Paramagnetic sensors are typically used specifically for measuring oxygen concentration. The high degree of sensitivity of oxygen (compared with other gases) to magnetic forces reduces the cross-sensitivity to other gases of paramagnetic sensors. Sensors come in two variants, the older balance type and the newer pressure type. The balance types of sensors are relatively frail and have been replaced by the pressure types. Nevertheless, some of these older devices are still being used. The balance type of sensor uses a mechanical approach that has a dried gas sample flowing through a chamber in which a nitrogen-filled dumbbell is balanced in a magnetic field. The paramagnetic force on the oxygen in the sample puts a torque on the dumbbell. The output can be read either as a spring-loaded displacement or, in newer devices, electronically by measuring the current required to keep the dumbbell centered.

Most modern paramagnetic oxygen sensors consist of a symmetrical, two-chambered cell with identical chambers for the sample and reference gas (often air or nitrogen). These cells are joined at an interface by a responsive differential pressure transducer or microphone. Sample and reference gases are pumped through these chambers in which a strong, usually varying, magnetic field surrounding the region acts on the oxygen molecules and generates a static pressure or a time-varying (acoustic) difference between the two sides of the cell, causing the transducer to produce a DC or AC voltage proportional to the oxygen concentration. When the magnetic field is modulated at acoustic frequencies, these devices may sometimes be referred to as magnetoacoustic.

Paramagnetic oxygen analyzers are very accurate, highly sensitive, and responsive, often with a step response of 200 ms to 90% of maximum. However, they require calibration, usually with pure nitrogen and oxygen. A major drawback is that they are adversely affected by water vapor and, consequently, require a water trap incorporated into their design. The frequency response makes then useful for measurement of oxygen on a breath-by-breath basis. The Datex Ohmeda Capnomac Ultima that was previously shown in Fig. 8 uses a paramagnetic oxygen sensor, as do many of the other mainline medical gas monitor manufacturers.

RADIOACTIVE IONIZATION

The ubiquitous smoke detector found in every house, hospital, and facility has spawned detectors for other gases such as carbon monoxide, a very important marker gas in medical diagnosis. Although usually used as devices that are set to alarm when a preset level is detected, they are also used as calibrated sensors. A very low level radioactive alpha particle source (such as Americium-241) can ionize certain gases so that, in the presence of an analyte, a circuit can be completed and current caused to flow in the detector circuit.

Ionization detectors detect the presence of invisible particles (less than 0.01 micron in size) in the air. Inside the detector, a small ionization chamber exists that contains an extremely small quantity of radioactive isotope. Americium-241 emits alpha particles at a fairly constant rate. The alpha particles, which travel at an extremely high rate of speed, knock off an electron from the oxygen and nitrogen molecules in the air passing through the ionization chamber. The free electron (negative charge) is then attracted to a positively charged plate, and the positively charged oxygen or nitrogen is attracted to a negatively charged plate, which creates a very small but constant current between the plates of a detector circuit, which in itself is a gas detection mechanism much in the same way that the other ionization detectors operated. However, when particles, such as soot particles, dust, fumes, or steam, enter the ionization chamber, the current is disrupted. If the current decreases too mid, an alarm is triggered.

The disadvantage of these devices is clearly the health hazard associated with the presence of the radioactive material. However, because the detector contains only a tiny amount of radioactive material, exposure is unlikely with proper care in handling. Another disadvantage of these sensitive detectors is the false-positive alarms that can be triggered by spurious dust and other nontoxic fumes. However, the big advantage is that ionization detectors are very sensitive and, given that false alarms

are tolerable, should be considered in most alarm situations.

Another form of radioactive detector is the electron capture detector, which uses a radioactive Beta emitter (electrons) to ionize some of the carrier gas and produce a current between a biased pair of electrodes. When organic molecules that contain electronegative functional groups, such as halogens, phosphorous, and nitro groups, pass by the detector, they capture some of the electrons and reduce the current measured between the electrodes.

RAMAN LASER SPECTROSCOPY

In the 1980s, Raman scattering was first heralded as an improvement to mass spectrometry (11), although some individuals had reservations (12). Although no longer manufactured but still serviced, Ohmeda Rascal II multi-gas analyzer uses a Raman scattering of laser light to identify and quantify O_2, N_2, CO_2, N_2O, and volatile anesthetic agents. It is stable and can monitor N_2 directly and CO_2 accurately for a wide range of concentrations. One of the acknowledged disadvantages is that a possibility of some destruction of volatile anesthetic agent exists during the analysis because the concentration of halothane does appear to fall when recirculated and as much as 15% must be added. Some concern exists over the reliability of the hardware, software, and laser light source (13) that is currently being addressed by others.

Raman scattering occurs when a gas sample is drawn into an analyzing chamber and is exposed to a high intensity beam from an argon laser. The laser energy is absorbed by the various molecules in the sample and are then excited into unstable vibrational or rotational energy states, which is the Raman scattering. The low intensity Raman scattered, or re-emitted, light signals are measured at right angles to the laser beam, and the spectrum of Raman scattering lines can be used to identify various types of gas molecules. Spectral analysis allows identification of known compounds by comparison with their Raman spectra. This technique is of similar accuracy to mass spectrometry.

SOLID-STATE SENSORS

At least four types of solid-state gas sensors exist: semiconductor metal oxide (SMO) sensors; chemically sensitive field effect transistors (ChemFETs); galvanic oxide sensors; and piezoelectric or surface acoustic wave (SAW) crystal sensors.

Semiconductor metal sensors are an outgrowth of the development of semiconductor devices. Early in the development of transistors and integrated circuits, it was observed that the characteristics would change in the presence of different gases. Recalling that a transistor is basically a voltage-controlled resistor, it was discovered that the p-n junction resistance was being changed by chemical reaction with the semiconductor materials (14). Commercially available Taguchi Gas Sensors (TGS) tin oxide sensors have found a niche as electronic noses. Walmsley et al. (15) used arrays of TGS sensors to develop patterns for ether and chloroform and other vapors of medical interest.

Hydrocarbons were among the first gases to be detected, and later, hydrogen sulfide was found to be detectable. Since the first tin oxide sensors appeared in the late 1960s, it has been found that by doping transition metal oxides, such as tin and aluminum, with other oxides, that as many as 150 different gases could be specifically detected (1) at ppm levels. The heated oxide adsorbs the analyte and the resistance change is a function of the concentration of the analyte. The semiconducting material is bonded or painted in a paste to a nonconducting substrate and mounted between a pair of electrodes. The substrate is heated to a temperature such that the gas being monitored reversibly changes the conductivity of the semiconducting metal oxide material. When no analyte is present, the current thinking is that oxygen molecules capture the free electrons in the semiconductor material when they are absorbing on the surface, thereby preventing the mobility of the electron flow. Analyte molecules replace the oxygen, thereby releasing the free electrons and, consequently, reducing the SMO resistance between the electrodes.

The ChemFET is derived from a field effect transistor where the normal gate metal has been replaced with a catalytic metal or chemically sensitive alloy. The gaseous analyte interacts with the gate metal and changes the FET characteristics to include gain and resistance.

Solid-state galvanic cells are based on the semiconductor galvanic properties of certain oxides or hydrides. The zirconium oxide galvanic cell oxygen sensor is probably one of the most ubiquitous sensors in daily life. It is the sensor mounted in every automobile catalytic converter to measure its effectiveness. Zirconium oxide, when heated to a temperature of about 700 °C, becomes an oxygen ion conductor, so that, in the presence of a difference in partial pressure on either side of a tube with metallized leads coming off each side, a voltage potential (Nernst voltage) is developed. These devices are commonly called fugacity sensors. As the process is reversible, a voltage applied will cause oxygen ions to flow. This process may also be applicable to the hydrogen ions in hydrides. An advantage of these oxygen sensors over other types is that no consumable exists. Hence, life is long. However, the need for heating tends to make these devices difficult to handle and they, as well as SMOs, require significant power to power the heating elements. However, because these sensors can be very small, they can have fast response times, often less than 100 ms, which makes them suitable for use for respiration monitoring. The electronics associated with the detection circuits is simple and should be very reliable.

Piezoelectric sensors use the change in a crystal vibrational frequency or propagation of surface acoustic waves to measure the concentration of a selected analyte. Most often, an analyte sample is passed through a chamber containing two piezoelectric crystals: a clean reference crystal and a second crystal that has been coated with a compound that specifically adsorbs specific analyte gases. Organophillic coatings are used for hydrocarbons such as anesthetic vapors. The resulting increase in mass changes the coated crystal's resonant frequency or the speed of propagation in direct proportion to the concentration of

anesthetic gas in the sample. Using either some form of beat frequency measurement or detection of the phase shift between the two crystals, a detection circuit can generate a signal that can be processed and displayed as a concentration.

EMERGING TECHNOLOGIES—MEMS—MICROFLUIDICS–NANOTECHNOLOGY

The development of a variety of microfluidic labs-on-a-chip is leading the charge in the state-of-the-art in gas sensing. It was noted in the gas chromatography section that microfluidic channels are being used as GC columns and in the mass spectrometry section that microfluidics plays a major role in the TOF passages and chambers. The ability to miniaturize classic technologies has opened the door to mass production as well as the ability to mix-and-match sensors and technologies. The development of electronic noses that can discriminate between thousands of different chemicals and gases is driving the need to detect odors and minute quantities of dangerous or toxic gases.

Patient safety in medicine continues to be a major driver. MEMS (micro-electromechanical systems) and nanotechnology have become the enabling technologies for placing thousands of sensors in microdimensional arrays that can be placed inside a capsule and swallowed or implanted to monitor physiological and metabolic processes. IR bolometers that can sense incredibly small temperature differences as low as 0.02 °C are already a part of today's inventory (Fig. 14), and in the future, nanotechnology elements that are merely one molecule thick and dust particle-sized may provide IR spectroscopic capability in implantable or inhaled micro-packages.

A new paradigm for gas analysis that has been enabled by the development of microfluidics was originally suggested in the 1960s by Mapleson (16), who suggested measuring a physical gas property as a way of inferring binary gas mixture composition. This concept has been extended and implemented for ternary and quaternary gas mixtures with a microfluidic gas multiple gas property analyzer (17–20). Ternary mixtures are assayed with a chip that measures viscosity with a microcapillary viscometer and density with a micro-orifice densitometer. An early prototype microfluidic lab-on-a-chip showing the microcapillaries is shown in Fig. 15. By measuring properties possessed in common by all gases, such as density, viscosity, and specific heat, a single chip can be used to analyze mixtures of any four gases. The concentrations of the constituents can be determined by simultaneously solving the equations that relate the mixture properties to the concentrations, thereby determining the relative concentrations of the mixture gases required to produce the measured properties. The only limitation to such an approach is that one must know what at least all but one of the constituents are. Microfluidic property sensors such a capillary viscometers, orifice densitometers, and speed-of-sound calorimeters can provide real-time simultaneous assays of respiratory gases (O_2, CO_2, and N_2), the simultaneity of which then enables the reduction to practice of a variety of physiologic analyzers such as metabolic rate, cardiac output, and cardio-pulmonary function that had been postulated back in the 1960s (21) but never practically implemented.

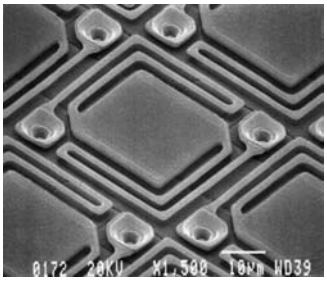


Figure 14. 25 micron wide longwave IR microbolometer detector.

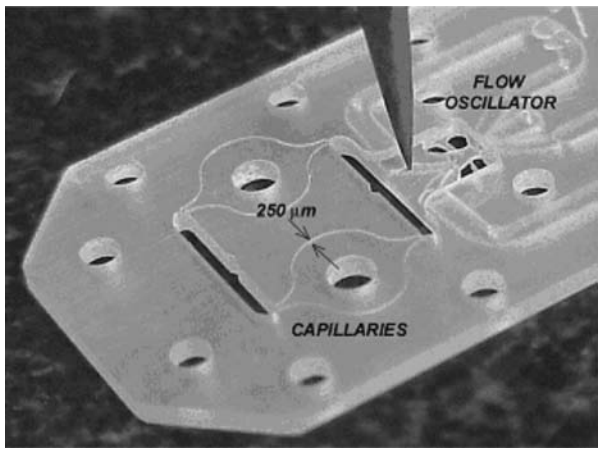


Figure 15. Microfluidic chip that measures the density and viscosity of three-component respired gases from which the constituent gas concentrations of oxygen, nitrogen, and carbon dioxide are deduced.

Advances in optics and optical fiber technology, greater expansion of solid-state sensing, and the revolutionary aspects of nanotechnology will provide gas analysis and sensing with new capabilities, provided that the lessons learned from the past are appropriately heeded.

OTHER CONSIDERATIONS IN GAS ANALYSIS

Most continuous gas analysis devices are side-stream monitors that acquire gas samples from a breathing circuit through long, narrow-diameter tubing lines. These lines may incorporate moisture removal to allow moisture to pass through the sampling line and into the atmosphere, as is the case with Nafion tubing. A water trap or filter may also be used to remove condensation from the sample in order to reduce water vapor before the gas sample reaches the analysis chamber. Gas samples are aspirated into the monitor at either an adjustable or a fixed-flow rate, typically from 50 to 250 ml/min. Lower rates minimize the amount of gas removed from the breathing circuit and, therefore, from the patient's tidal volume; however, lower sampling flow rates increase the response time and typically reduce the accuracy of conventional measurements.

Gas monitors have to eliminate the exhaust gas through a scavenging system or back to the patient's breathing circuit.

DISPLAYS, ALARMS, CALIBRATION, AND CONTROLS

Many gas monitors provide a graphic display of breath-by-breath concentrations and a hardcopy of trends of gas concentrations from the beginning to the end of an event (e.g., anesthesia delivery during an operation). The user typically calibrates or verifies calibration of the sensors with a standard gas mixture from an integral or external gas cylinder. Gas monitors are usually microprocessor-controlled and have user-adjustable alarms that typically include factory-preset default alarms or alarm-set programs for both high and low concentrations of the gases measured. Some monitors also have alarms for system malfunctions, such as disconnection from a gas source, and leaks can often be identified from trending of O_2 and CO_2. Occlusion, apnea, or inadvertent rebreathing can also be identified. Most monitors typically have a method for temporarily, but not indefinitely, silencing audible alarms for low O_2, high N_2O, and high agent, whereas other, less critical audible alarms can be permanently silenced. Most monitors typically display real-time waveforms and long-term trends. They have integral display capability and are also commonly equipped with output jacks to interface with computerized record-keeping systems or with additional analog or digital display units such as chart recorders and printers.

PATIENT SAFETY

A review of the background and significance of medical gas sensors and monitors would be incomplete without an expression of the context that patient safety has had on the impetus for recent gains in technology and the need for additional improvements. Clearly the intrinsic dangers in the conduct of anesthesia have been long understood. It became evident in the early 1980s that patient safety and reduction to risk was possible if respiratory and anesthetic gas monitoring was routinely available and used. As a result of improved and increased availability of medical gas monitoring technology and professional activity lead by the Anesthesia Patient Safety Foundation (APSF) with the support of the American Society of Anesthesiologists (ASA), a standard for monitoring has been adopted and is routinely used in today's clinical practice. This standard requires assessment of the patient's ventilation and oxygenation in addition to circulation and temperature. The use of such monitors has resulted in a significant decrease in the risk of anesthesia-related deaths and morbidity in the ICU and other critical care situations.

CONCLUSION

Medical gas monitoring has been so successful in improving patient safety and reducing patient risk that medical malpractice liability insurance companies have lowered their risk liabilities and premiums to anesthesiologists who guarantee the routine implementation of these standards whenever possible (22). The argument for providing additional patient safety will continue to be a powerful incentive to improve and enhance the methods and techniques to provide increased knowledge of the monitoring of respiratory and anesthetic gases.

The availability of gas sensors and monitors is a boon to the medical profession from both a clinical as well as a research point of view. In addition to patient safety, new diagnostic capabilities are emerging every year. In research, new gas-sensing capabilities are enhancing the discovery of markers for all kinds of disease states and metabolic functions.

Looking to the future, MEMS, microfluidics, and nanotechnology will provide growth in our understanding of physiologic processes at levels of detail never before conceived of, from inside the body as well as supplanting today's conventional techniques.

BIBLIOGRAPHY

1. Chou J, Hazardous Gas Monitors, A Practical Guide to Selection, Operation and Applicationsy. New York: McGraw-Hill; 2000.
2. ECRI. Multiple medical gas monitors, respired/anesthetic. Product Comparison System, ECRI Product Code 17–445, August 1993.
3. Datex Ohmeda Smart Vent 7900. Product Description AN2842-B/7 01 1999 Datex-Ohmeda, Inc.
4. Yu CM, Sheem SK. (1995). Miniature gas chromatography sensor. Lawrence Livermore National Lab. www.llnl.gov/sensor_technology/STR72.html.
5. Harris D, Bailey S. Near infrared spectroscopy in adults. Anaesthesia 1993;48:694–696.
6. McPeak HB, Palayiwa E, Robinson GC, Sykes MK. An evaluation of the Bruel and Kjaer monitor 1304. Anaesthesia 1992; 47(1):41–47.
7. Lehmann KK, Rabinowitz P. High-finesse optical resonator for cavity ring-down spectroscopy based upon Brewster's angle prism retroreflectors. U.S. Patent 5,973,864, October 26, 1999.
8. Lehmann KK, Romanini D. The superposition principle and cavity ring-down spectroscopy. J Chem Phys 1996;105(23): 10263–10277.
9. Ecelberger SA, Cornish TJ, Bryden W. The improved teeny-TOF mass spectrometer for chemical and biological sensing, 3rd Harsh-Environment Mass Spectrometry Workshop, March 25–28, 2002 Pasadena, CA.
10. Yoon HY, Kim JH, Choi ES, Yang SS, Jung KW. Fabrication of a novel micro time-of-flight mass spectrometer. Sens Actuators A 2002;97–98:441–447.
11. Westenskow DR, Smith KW, Coleman DL, et al. Clinical evaluation of a Raman scattering multiple gas analyzer. Anesthesiology 1989;70:350–355.
12. Severinghaus JW, Ozanne GM. Multi-operating room monitoring with one mass spectrometer. Acta Anaesthesiol Scan 1987; (Suppl) 70:186–187.
13. Lockwood G, Landon MJ, Chakrabarti MK, Whitwam JG. The Ohmeda Rascal II. Anaesthesia 1994;49:44–53.
14. Kress-Rogers E, ed. Handbook of Biosensors and Electronic Noses—Medicine, Food and the Environment. Boca Raton, FL: CRC Press; 1996.

15. Walmsley AD, Haswell SJ, Metcalfe E. Methodology for the selection of suitable sensors for incorporation into a gas sensor array. Analytica Chimica Acta 1991;242:31.
16. Mapleson WW. Physical methods of gas analysis. Brit J Anaesth 1962;34:631.
17. Drzewiecki TM. Fluidic multiple medical gas monitor. NIH BioEngineering Symposium, Bethesda, MD, February 1998.
18. Drzewiecki TM, Polcha M, Koser M, Calkins J. A novel inexpensive respiratory and anesthetic gas monitor. Society for Technology in Anesthesia 1999 Annual Meeting. San Diego, CA: January 1999.
19. Drzewiecki TM, Calkins J. Real time, simultaneous analysis of multiple gas mixtures using microfluidic technology. Proc Instrument Society of America AD 2000 Symposium, Charleston, WV: April 2000.
20. Drzewiecki TM. Method and apparatus for real time gas analysis. U.S. Patent 6,076,392, June 2000.
21. Kim TS, Rahn H, Farhi LE. Estimation of true venous and arterial PCO_2 by gas analysis of a single breath. J Appl Physiol 1966;21(4):1338–1344.
22. Swedlow DB. Respiratory gas monitoring. In: Saidman L, Smith N, eds. Monitoring in Anesthesia. 3rd ed. Boston, MA: Butterworth-Heinemann; 1993. pp 27–50.

See also BLOOD GAS MEASUREMENTS; RESPIRATORY MECHANICS AND GAS EXCHANGE.

MEDICAL PHOTOGRAPHY.

See PHOTOGRAPHY, MEDICAL.

MEDICAL PHYSICS LITERATURE

COLIN ORTON
Harper Hospital and Wayne State University
Detroit, Michigan

INTRODUCTION

Medical physicists are responsible for the application of physics to the diagnosis and treatment of disease and other disabilities although, in some countries, the treatment of patients with disabilities is a separate field, often referred to as "biomedical engineering" or words to that effect. Here, we restrict ourselves to applications in the diagnosis and treatment of disease.

The major applications in diagnosis are the use of X-rays for imaging (diagnostic radiology, including computerized tomography (CT), etc.); radioactive isotopes for imaging and uptake measurements [nuclear medicine, single photon emission computed tomography (SPECT), positron emission tomography (PET), etc.]; magnetic resonance magnetic resonance imaging (MRI) and spectroscopy (MRS); ultrasound (ultrasonography).

Applications of physics to the treatment of disease include the following: external beams of X-rays, gamma-rays, or electrons for the treatment of cancer radiation oncology, including stereotactic radiosurgery, intensity modulated radiation therapy (IMRT), total body irradiation (TBI), etc.; external beams of heavy particles for the treatment of cancer (neutrons, protons, heavy ions); internal radioisotope treatments for cancer (brachytherapy, systemic radiotherapy, and radioimmunotherapy) and other problems (intravascular brachytherapy, hyperthyroidism, etc.); hyperthermia for the treatment of cancer.

Other topics of major interest for medical physicists include various medical applications of light and lasers, radiation protection, radiation measurements, radiation biology, and the applications of computers to all of the above.

Throughout the world there are medical physics organizations that represent medical physicists and provide information to help them in their profession. Most of these are national associations, with about 70 of these represented by the International Organization for Medical Physics (IOMP). In terms of publications, by far the two most prolific organizations are the Institute of Physics and Engineering in Medicine (IPEM) in the United Kingdom, and the American Association of Physicists in Medicine (AAPM) in North America. These two organizations between them have published hundreds of reports, monographs, and meeting proceedings that are used as reference materials throughout the world. From the IPEM many of these are published by the Institute of Physics Publishing (IOPP) in Bristol, UK, and from the AAPM many are published by either the American Institute of Physics (AIP) or Medical Physics Publishing (Madison, WI).

JOURNALS

Several national and international organizations and independent publishers publish journals used by medical physicists. Some of these journals are used extensively for medical physics papers in which at least 25% of the manuscripts are medical physics articles. These are categorized as "Primary" journals below. Others contain some medical physics articles (<25%) and are categorized as "Secondary".

PRIMARY MEDICAL PHYSICS JOURNALS

Australasian Physical & Engineering Sciences in Medicine, Australasian College of Physical Scientists and Engineers in Medicine and the College of Biomedical Engineers.

Journal of Applied Clinical Medical Physics, American College of Medical Physics (http://www.jacmp.org).

Journal of Medical Physics, Association of Medical Physicists of India.

Medical Dosimetry, American Association of Medical Dosimetrists (Elsevier).

Medical Physics, American Association of Physicists in Medicine (AIP).

Physica Medica, Istituti Editoriali e Poligrafici Internazionali Casella Postale n.1, Succursale n.8, 56123 Pisa, Italy (http://www.iepi.it).

Physics in Medicine and Biology, Institute of Physics and Engineering in Medicine (IOPP).

Polish Journal of Medical Physics and Engineering, Polish Society of Medical Physics.

Zeitschrift für Medizinische Physik, Deutschen, Österreichischen und Schweizerischen Gesellschaft für Medizinische Physik (Elsevier).

SECONDARY MEDICAL PHYSICS JOURNALS

Acta Radiologica, Scandinavian Society of Radiology (Taylor & Francis).

American Journal of Roentgenology, American Roentgen Ray Society.

Applied Radiation and Isotopes (Elsevier).

Australasian Radiology, Royal Australian and New Zealand College of Radiologists (Blackwell).

Biomedical Imaging and Interventional Journal (University of Malaya: http://www.biij.org).

Biomedical Instrumentation & Technology, Association for the Advancement of Medical Instrumentation.

Brachytherapy, American Brachytherapy Society (Elsevier).

British Journal of Radiology, British Institute of Radiology.

Canadian Association of Radiologists Journal, Canadian Association of Radiologists.

Cancer Radiothérapie, Société Française de Radiothérapie Oncologique (Elsevier).

Cardiovascular and Interventional Radiology, Cardiovascular and Interventional Radiological Society of Europe, Japanese Society of Angiography and Interventional Radiology, and British Society of Interventional Radiology (Springer).

Cardiovascular Radiation Medicine (Elsevier).

Clinical Imaging (Elsevier).

Clinical Radiology, Royal College of Radiologists (Elsevier).

Critical Reviews in Computed Tomography (Taylor & Francis).

Computerized Medical Imaging and Graphics, Computerized Medical Imaging Society (Elsevier).

Computers in Biology and Medicine (Elsevier).

Current Problems in Diagnostic Radiology (Mosby).

European Journal of Nuclear Medicine and Molecular Imaging, European Association of Nuclear Medicine (Springer).

European Journal of Radiology (Elsevier).

European Journal of Ultrasound, European Federation of Societies for Ultrasound in Medicine and Biology (Elsevier).

European Radiology, European Congress of Radiology (Springer).

Health Physics, Health Physics Society (Lippincott Williams & Wilkins).

IEEE Transactions on Medical Imaging, IEEE.

International Journal of Radiation Biology (Taylor & Francis).

International Journal of Radiation Oncology, Biology, Physics, American Society of Therapeutic Radiology and Oncology (Elsevier).

Investigative Radiology (Lippincott Williams & Wilkins).

Journal of Biomedical Optics, International Society for Optical Engineering (SPIE).

Journal of Cardiovascular Magnetic Resonance, Society for Cardiovascular Magnetic Resonance (Taylor & Francis).

Journal of Clinical Ultrasound (Wiley).

Journal of Computer Assisted Tomography (Lippincott Williams & Wilkins).

Journal of Diagnostic Radiography and Imaging, Royal Society of Medicine.

Journal of Digital Imaging, Society for Computer Applications in Radiology (Springer).

Journal of Electronic Imaging, International Society for Optical Engineering (SPIE).

Journal of Labelled Compounds and Radiopharmaceuticals (Wiley).

Journal of Magnetic Resonance Imaging, International Society for Magnetic Resonance Medicine (Wiley)

Journal of Neuroimaging, American Society of Neuroimaging (Sage Publications).

Journal of Nuclear Cardiology, American Society of Nuclear Cardiology (Elsevier).

Journal of Nuclear Medicine, Society of Nuclear Medicine.

Journal of Nuclear Medicine Technology, Society of Nuclear Medicine.

Journal of Radiological Protection, Society for Radiological Protection (IOPP, Bristol, U.K.).

Journal of the Acoustical Society of America, Acoustical Society of America (American Institute of Physics, New York).

Journal of the American Society of Echocardiography, American Society of Echocardiography (Mosby).

Journal of Thoracic Imaging, Society of Thoracic Radiology (Lippincott Williams & Wilkins).

Journal of Ultrasound in Medicine, American Institute of Ultrasound in Medicine.

Journal of Vascular and Interventional Radiology, Society of Interventional Radiology (Lippincott Williams & Wilkins).

Journal of X-Ray Science and Technology (IOS Press, Amsterdam).

Lasers in Medical Science (Springer).

Lasers in Surgery and Medicine, American Society for Laser Medicine and Surgery (Wiley).

Medical Engineering & Physics (Elsevier).

Magnetic Resonance Imaging (Elsevier).

Magnetic Resonance in Medicine, International Society for Magnetic Resonance in Medicine (Wiley).

Medical Engineering & Physics, Institute of Physics and Engineering in Medicine (Elsevier).

Medical Image Analysis (Elsevier).

Molecular Imaging and Biology, Academy of Molecular Imaging (Springer).

Neuroradiology, European Society of Neuroradiology (Springer).

NMR in Biomedicine (Wiley).

Nuclear Medicine and Biology, Society of Radiopharmaceutical Sciences (Elsevier).

Pediatric Radiology, European Society of Pediatric Radiology, Society for Pediatric Radiology, Asian and Oceanic Society for Pediatric Radiology (Springer).

Photomedicine and Laser Surgery, World Association for Laser Therapy (Mary Ann Liebert, Inc.).

Physiological Measurement, Institute of Physics and Engineering in Medicine (IOPP).

Progress in Nuclear Magnetic Resonance Spectroscopy (Elsevier).

Radiation Measurements (Elsevier).

Radiation Physics and Chemistry (Elsevier).

Radiation Research, Radiation Research Society.

Radiographics, Radiological Society of North America.

Radiography, College of Radiographers (Elsevier).

Radiology, Radiological Society of North America.

Radiotherapy and Oncology, European Society for Therapeutic Radiology and Oncology (Elsevier).

Seminars in Interventional Radiology (Thieme).

Seminars in Nuclear Medicine (Elsevier).

Seminars in Radiation Oncology (Saunders).

Seminars in Roentgenology (Elsevier).

Seminars in Ultrasound, CT and MRI (Elsevier).

Techniques in Vascular and Interventional Radiology (Elsevier).

Topics in Magnetic Resonance Imaging (Lippincott Williams & Wilkins).

Ultrasound in Medicine & Biology, World Federation for Ultrasound in Medicine and Biology (Elsevier).

Ultrasound in Obstetrics and Gynecology (Wiley).

Year Book of Diagnostic Radiology (Elsevier).

Year Book of Nuclear Medicine (Elsevier).

BOOKS AND REPORTS

As with journals, books and reports are published by both medical physics organizations, especially the AAPM and the IPEM, and independent publishers. Most of the books and reports in the following lists can be purchased directly from the publishers or, alternatively, through bookstores using the ISBN number provided.

MEDICAL AND RADIOLOGICAL PHYSICS

General

A Century of X-Rays and Radioactivity in Medicine: With Emphasis on Photographic Records of the Early Years, Richard F. Mould, ISBN 0-7503-0224-0, 1993, 236 pp, IOPP, Bristol (UK).

Essentials of Radiology Physics, Charles A. Kelsey, ISBN: 0875273548, 1985, 467 pp, W.H. Green.

Introduction to Radiological Physics and Radiation Dosimetry, Frank Herbert Attix, ISBN: 0-471-01146-0, 1986, 640 pp, Advanced Medical Publishing, Madison (WI).

Meandering in Medical Physics: A Personal Account of Hospital Physics, J.E. Roberts, N.G. Trott, ISBN: 0750304944, 1999, 181 pp, IOPP, Bristol (UK).

Medical Physics and Biomedical Engineering, Brown BH, Smallwood RH, Barber DC, Lawford PV; Hose DR, ISBN: 0750303670, 1998, 768 pp, IOPP, Bristol (UK).

Medical Physics Handbook of Units and Measures, Freim J, Jr, ISBN: 0944838308, 1992, 47 pp, Medical Physics Publishing, Madison (WI).

Medical Radiation Physics: Roentgenology, Nuclear Medicine & Ultrasound, Hendee WR, ISBN: 0815142404, 1979, 517 pp, Year Book Medical Publishers.

Physics in Medicine and Biology Encyclopedia (2 Volume Set), T.F. McAinsh, (editor), ISBN: 0080264972, 1986, Pergamon, Elmsfood (NY).

Physics and Engineering in Medicine in the New Millennium, Sharp PF, Perkins AC editors., ISBN: 0904181952, 2000, 156 pp, IPEM, York, (UK).

Physics of Radiology, 4th ed., Johns H E, John Robert Cunningham, ISBN: 0398046697, 1983, 796 pp, Charles C. Thomas.

Physics of Radiology, second edition, Wolbarst A B, ISBN: 1-930524-22-6. Published: 2005, 660 pp, Medical Physics Publishing, Madison (WI).

Physics of the Body, Cameron JR, Skofronick JG, Grant RM, ISBN: 094483891X, 1999, 394 pp, Medical Physics Publishing, Madison (WI).

Principles and Practice of Clinical Physics & Dosimetry, Michael L.F. Lim, ISBN: 1-883526-11-6, 2005, 500 pp, Advanced Medical Publishing, Madison (WI).

Principles of Radiological Physics, 4th ed., Graham D, Cloke P, ISBN: 0443070733, 2003, 576 pp, Churchill Livingstone.

Radiation Biophysics, Alpen EL, ISBN: 0120530856, 1998, 484 pp, Academic Press.

Radiation Physics Handbook for Medical Physicists, Ervin B. Podgorsak, ISBN: 3540250417, 2005, 360 pp, Springer, New York.

Review of Radiological Physics, Walter Huda, Richard M. Slone, ISBN: 0781736757, 2002, 350 pp, Lippincott Williams & Wilkins, Philadelphia.

Topical

How the Body Works, Lenihan J, ISBN: 0944838-48-0, 1995, 200 pp, Medical Physics Publishing, Madison (WI).

Physics of the Body 2nd ed., Cameron J, et al., ISBN: 0-944838-91-X, 1999, 394 pp, Medical Physics Publishing, Madison (WI).

Medical Applications of Nuclear Physics, Bethge K, Kraft G, Kreisler P, Walter G, ISBN: 3540208054, 2004, 208 pp, Springer.

Progress in Medical Radiation Physics, Orton CG, ISBN: 0306417898, 1985, 248 pp, Plenum, New York.

RADIATION ONCOLOGY PHYSICS

General

AAPM Monograph No. 15, Radiation Oncology Physics, Kereiakes J, Elson H, Born C, editors, ISBN: 0-883185-33-4, 1986, 812 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 26, General Practice of Radiation Oncology Physics in the 21st Century, Almon Shiu & David Mellenberg, ISBN: 0-944838-98-7, 2000, 368 pp, Medical Physics Publishing, Madison (WI).

Applied Physics for Radiation Oncology, Robert Stanton & Donna Stinson ISBN: 0-944838-60-X, 1996, 375 pp, Medical Physics Publishing, Madison (WI).

Biomedical Particle Accelerators, Scharf WH, Siebers JV, ISBN: 1563960893, 1994, 480 pp, Springer.

Blackburn's Introduction to Clinical Radiation Therapy Physics, Benjamin Blackburn ISBN: 0-944838-06-5, 1989, 218 pp, Medical Physics Publishing, Madison (WI).

Clinical Radiotherapy Physics, 2nd ed., Jayaraman S, Lanzl LH, Lanzl EF, ISBN: 3540402845, 2004, 523 pp, Springer.

Handbook of Radiotherapy Physics: Theory and Practice, Mayles P, Nahum A, Rosenwald J-C, ISBN: 0750308605, 2005, 700 pp, IOPP, Bristol (UK).

Modern Technology of Radiation Oncology, Van Dyk J, ISBN: 0-944838-38-3, 1999, 1072 pp, Medical Physics Publishing, Madison (WI).

Practical Radiotherapy: Physics and Equipment, Cherry P, Duxbury A, ISBN: 1900151065, 1998, 224 pp, Cambridge University Press, New York.

Radiation Therapy Physics, Hendee WR, Ibbott GS, Hendee EG, ISBN: 0471394939, 2004, 450 pp, Wiley, New York.

Radiotherapy Physics and Equipment, Morris S, Williams A, ISBN: 0443062110, 2001, 176 pp, Churchill Livingstone.

Radiotherapy Physics: In Practice, Williams JR, Thwaites DI, editors, ISBN: 0-19-262878-X, 2000, 362 pp, Oxford University Press, New York.

Review of Radiation Oncology Physics, Prasad SC, ISBN: 1-930524-08-0, 2002, 95 pp, Medical Physics Publishing, Madison (WI).

Study Guide for Radiation Oncology Physics Board Exams, Berman B, ISBN: 0-944838-94-4, 2000, 112 pp, Medical Physics Publishing, Madison (WI).

The Physics of Radiation Therapy, Hardbound 3rd ed., Khan F, ISBN: 0-7817-3065-1, 2003, 511 pp, Wiley, New York.

Walter & Miller's Textbook of Radiotherapy Radiation Physics, Therapy and Oncology, 6th ed., Bomford CK et al., ISBN: 0443062013, 2003, 660 pp, Elsevier.

Topical

3-D Conformal and Intensity Modulated Radiation Therapy: Physics and Clinical Applications, Purdy JA, Grant W III, Palta JR, Butler EB, Perez CA, editors, ISBN: 1-883526-10-8, 2001, 650 pp, Advanced Medical Publishing, Madison (WI).

A Practical Guide to 3-D Planning and Conformal Radiation Therapy, Purdy JA, Starkschall G, ISBN: 1-883526-07, 1999, 400 pp, Advanced Medical Publishing, Madison (WI).

A Practical Guide to Intensity-Modulated Radiation Therapy, Memorial Sloan-Kettering Cancer Center, ISBN: 1-930524-13-7, 2003, 450 pp. Medical Physics Publishing, Madison (WI).

AAPM Manual No. 2: Workbook on Dosimetry and Treatment Planning for Radiation Oncology Residents, Wu RK, Gerbi BJ, Doppke KP, editors, ISBN: 0-88318-916-X, 1991, 32 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 2, Practical Aspects of Electron Beam Treatment Planning, Orton CG, Bagne F, editors, ISBN: 0-88318-247-5, 1978, 109 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 7, Recent Advances in Brachytherapy Physics, Shearer DR, editors, ISBN: 0-88318-285-8, 1981, 202 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 8, Physical Aspects of Hyperthermia, Nussbaum GH, editors, ISBN: 0-88318-414-1, 1982, 656 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 9, Advances in Radiation Therapy Treatment Planning, Wright AE, Boyer A, editors, ISBN: 0-883184-23-0, 1982, 626 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 16, Biological, Physical and Clinical Aspects of Hyperthermia, Paliwal BR, Hetzel FW, Dewhirst M, editors, ISBN: 0-88318-558-X, 1988, 483 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 28, Intravascular Brachytherapy/Fluoroscopically Guided Interventions, Balter S,

Chan RC, Shope TB Jr, editors, ISBN: 1-930524-10-2, 2002, 930 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 29, Intensity-Modulated Radiation Therapy: The State of the Art, Palta JR, Rockwell Mackie T, editors, ISBN: 1-930524-16-1, 2003, 904 pp. Medical Physics Publishing, Madison (WI).

AAPM Proceedings No. 2, Proceedings of the Symposium on Electron Dosimetry and Arc Therapy, Paliwal BR, editor, ISBN: 0-88318-404-4, 1981, 384 pp, Medical Physics Publishing, Madison (WI).

AAPM Proceedings No. 3, Proceedings of a Symposium on Quality Assurance of Radiotherapy Equipment, Starkschall G, editor, ISBN: 0-88318-422-2, 1982, 242 pp, Medical Physics Publishing, Madison (WI).

AAPM Proceedings No. 5, Optimization of Cancer Radiotherapy, Paliwal BR, Herbert DE, Orton CG, editors, ISBN: 0-88318-483-4, 1984, 556 pp, Medical Physics Publishing, Madison (WI).

AAPM Proceedings No. 12, Biological & Physical Basis of IMRT & Tomotherapy, Paliwal BR, Herbert DE, Fowler JF, Mehta M, editors, ISBN: 1-930524-11-0, 2002, 390 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 7, Protocol for Neutron Beam Dosimetry, Radiation Therapy Committee Task Group 18; ISBN: 0-88318-276-9, 1980, 51 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 13, Physical Aspects of Quality Assurance in Radiation Therapy, Radiation Therapy Committee Task Group 24, with contribution from Task Group 22, ISBN: 0-88318-457-5, 1984, 63 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 17, The Physical Aspects of Total and a Half Body Photon Irradiation, Radiation Therapy Committee Task Group 29; ISBN: 0-88318-513-X, 1986, 55 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 21, Specification of Brachytherapy Source Strength, Radiation Therapy Committee Task Group 32; ISBN: 0-88318-545-8, 1987, 21 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 23, Total Skin Electron Therapy: Technique and Dosimetry Radiation Therapy Committee Task Group 30; ISBN: 0-88318-556-3, 1987, 55 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 24, Radiotherapy Portal Imaging Quality, Radiation Therapy Committee Task Group 28; ISBN: 0-88318-557-1, 1987, 29 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 26, Performance Evaluation of Hyperthermia Equipment, Hyperthermia Committee Task Group 1; ISBN: 0-88318-636-5, 1989, 46 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 27, Hyperthermia Treatment Planning, Hyperthermia Committee Task Group 2; ISBN: 0-88318-643-8, 1989, 57 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 32, Clinical Electron-Beam Dosimetry, Radiation Therapy Committee Task Group 25, ISBN: 0-88318-905-4, 1990, 40 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 40, Radiolabeled Antibody Tumor Dosimetry (Reprinted from Medical Physics, Vol. 20, Issue 2), Nuclear Medicine Committee Task Group 2; ISBN: 1-56396-233-0, 1993, 112 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 41, Remote Afterloading Technology, Remote Afterloading Technology Task Group 41; ISBN: 1-56396-240-3, 1993, 107 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 45, Management of Radiation Oncology Patients with Implanted Cardiac Pacemakers (Reprinted from Medical Physics, Vol. 21, Issue 1), Task Group 34. ISBN: 1-56396-380-9, 1994, 6 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 46, Comprehensive QA for Radiation Oncology (Reprinted from Medical Physics, Vol. 21, Issue 4), Radiation Therapy Committee Task Group 40; ISBN: 1-56396-401-5, 1994, 37 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 47, AAPM Code of Practice for Radiotherapy Accelerators (Reprinted from Medical Physics, Vol. 21, Issue 4), Radiation Therapy Task Group 45; ISBN: 1-56396-402-3, 1994, 37 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 48, The Calibration and Use of Plane-Parallel Ionization Chambers for Dosimetry of Electron Beams (Reprinted from Medical Physics, Vol. 21, Issue 8) Radiation Therapy Committee Task Group 39; ISBN: 1-56396-461-9, 1994, 10 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 50, Fetal Dose from Radiotherapy with Photon Beams (Reprinted from Medical Physics, Vol. 22, Issue 1), Radiation Therapy Committee Task Group 36; ISBN: 1-56396-453-8, 1995, 20 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 54, Stereotactic Radiosurgery, Radiation Therapy Committee Task Group 42; ISBN: 1-56396-497-X, 1995, 100 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 55, Radiation Treatment Planning Dosimetry Verification, AAPM ISBN: 1-56396-534-8, 1995, 200 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 56, Medical Accelerator Safety Considerations (Reprinted from Medical Physics, Vol. 20, Issue 4), Radiation Therapy Committee Task Group 35; ISBN: 1-888340-01-0, 1993, 15 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 59, Code of Practice for Brachytherapy Physics (Reprinted from Medical Physics, Vol. 24, Issue 10), Radiation Therapy Committee Task Group 56; ISBN: 1-888340-14-2, 1997, 42 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 61, High Dose-Rate Brachytherapy Treatment Delivery (Reprinted from Medical Physics, Vol. 25, Issue 4), Radiation Therapy Committee Task Group 59; ISBN: 1-888340-17-7, 1998, 29 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 62, Quality Assurance for Clinical Radiotherapy Treatment Planning (Reprinted from Medical Physics, Vol. 25, Issue 10), Radiation Therapy Committee Task Group 53; ISBN: 1-888-340-18-5, 1998, 57 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 66, Intravascular Brachytherapy Physics (Reprinted from Medical Physics, Vol. 26, Issue 2), Radiation Therapy Committee Task Group 60; ISBN: 1-888340-23-1, 1999, 34 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 67, Protocol for Clinical Reference Dosimetry of High-Energy Photon and Electron Beams (Reprinted from Medical Physics, Vol. 26, Issue 9) Task Group 51; ISBN: 1-888340-25-8, 1999, 24 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 68, Permanent Prostate Seed Implant Brachytherapy (Reprinted from Medical Physics, Vol. 26, Issue 10), Task Group 64; ISBN: 1-888340-26-6, 1999, 23 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 69, Recommendations of the AAPM on ^{103}Pd Interstitial Source Calibration and Dosimetry: Implications for Dose Specification and Prescription, AAPM, ISBN: 1-888340-27-4, 2000, 9 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 71, A Primer for Radioimmunotherapy and Radionuclide Therapy Task Group 7, ISBN: 1-888340-29-0, 2001, 73 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 72, Basic Applications of Multileaf Collimators, Task Group 50 ISBN: 1-888340-30-4, 2001, 54 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 75, Clinical Use of Electronic Portal Imaging, Radiation Therapy Committee Task Group 58, ISBN: 1-888340-34-7, 2001, 26 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 76, AAPM Protocol for 40-300 kV X-ray Beam Dosimetry in Radiotherapy and Radiobiology, AAPM, ISBN: 1-888340-35-5, 2001, 26 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 81, Dosimetric Considerations for Patients with Hip Prostheses Undergoing Pelvic Irradiation, Radiation Therapy Committee Task Group 63 (Reprinted from Medical Physics, Vol. 30, Issue 6), ISBN: 1-888340-42-8, 2003, 21 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 82, Guidance Document on Delivery, Treatment Planning, and Clinical Implementation of IMRT, AAPM Radiation Therapy Committee, ISBN: 1-888340-43-6, 2003, 25 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 83, Quality Assurance for Computed-Tomography Simulators and the Computed-Tomography-Simulation Process, Radiation Therapy Committee Task Group 66, ISBN: 1-888340-44-4, 2003, 31 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 84, A Revised AAPM Protocol for Brachytherapy Dose Calculations (Reprinted from Medical Physics, Vol. 31, Issue 3, pp. 633-674), Radiation Therapy Committee, ISBN: 1-888340-46-0, 2004, 42 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 85, Tissue Inhomogeneity Corrections for Megavoltage Photon Beams, Task Group No. 65 of the Radiation Therapy Committee, ISBN: 1-888340-47-9, 2004, 124 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 86, Quality Assurance for Clinical Trials: A Primer for Physicists, Subcommittee on Quality Assurance Physics for Cooperative Trials of the Radiation Therapy Committee, ISBN: 1-88340-48-7, 2004, 68 pp, Medical Physics Publishing, Madison (WI).

Achieving Quality in Brachytherapy, BR Thomadsen, ISBN: 0750305541, 1999, pp, 268, Advanced Medical Publishing, Madison (WI).

A Practical Guide to CT-Simulation, Edited by: Coia L, Schultheiss T, Hanks G, ISBN: 1-883526-04-3, 1995, 216 pp, Advanced Medical Publishing, Madison (WI).

Brachytherapy Physics (1994 AAPM Summer School), Williamson J et al., ISBN: 0-944838-50-2, 1995, 715 pp, Medical Physics Publishing, Madison (WI).

Clinical Target Volumes in Conformal and Intensity Modulated Radiation Therapy, Gregorie V, Scalliet and P, Ang KK, ISBN: 3540413804, 2003, 300 pp, Springer Verlag.

Contemporary IMRT: Developing Physics and Clinical Implementation, Webb S, ISBN: 0750310049, 2004, 478 pp, IOPP, Bristol (UK).

CT Simulation for Radiotherapy, Jani S, ISBN: 0-944838-32-4, 1993, 172 pp, Medical Physics Publishing, Madison (WI).

Geometric Uncertainties in Radiotherapy, BIR, 2003, ISBN: 0905749537, The British Institute of Radiology.

Intraoperative Irradiation, Techniques and Results, Gunderson LL, Wilett CG, Harrison LB, Calvo FA, editors ISBN: 0-89603-523-9, 1999, 560 pp, Advanced Medical Publishing, Madison (WI).

Intraoperative Radiation Therapy, Ralph Dobelbower, Jr. and Mitsuyuki Abe, ISBN: 0849368464, 1989, 432 pp, CRC Press, Boca Raton (FL).

Introduction to Clinical Radiation Oncology, 3rd ed., Coia L, Moylan D, ISBN: 0-944838-70-7, Published: March 1998, 568 pp, Medical Physics Publishing, Madison (WI).

IPEM Report No. 68, A Guide to Commissioning & Quality Control of Treatment Planning Systems, Shaw J, editor, ISBN: 0904181839, 1996, IPEM, York (UK).

IPEM Report No. 81, Physics Aspects of Quality Control in Radiotherapy, Mayles WPM, Lake RA, McKenzie AL, Macaulay EM, Morgan HM, Powley SK, ISBN: 0904181928, 1998, IPEM, York (UK).

IPEM Report No. 83, Targeted Radiotherapy, Fleming JS, Perkins AC, ISBN: 0904181979, Published: 2000, 112 pp, IPEM, York (UK).

Linac and Gamma Knife Radiosurgery, Isabelle M. Germano, ISBN: 1879284707, 2000, 295 pp, Advanced Medical Publishing, Madison (WI).

Linear Accelerators for Radiation Therapy, D. Greene, ISBN: 0750304766, 1997, 288 pp, IOPP, Bristol (UK).

Monitor Unit Calculations for External Photon and Electron Beams, Gibbon JP, editor, ISBN: 1-883526-08-6, 2000, 152 pp, Advanced Medical Publishing, Madison (WI).

Physical Aspects of Brachytherapy, Godden TJ, ISBN: 0852745117, 1988, 304 pp, IOPP, Bristol (UK).

Physical Aspects of Stereotactic Radiosurgery, Phillips M H, ISBN: 0306445352, 1993, 286 pp, Plenum, New York.

Physics and Technology of Hyperthermia, Field SB, Franconi C, ISBN: 9024735092, 1999, 668 pp, Springer.

Physics of Electron Beam Therapy, Klevenhagen SC, ISBN: 0852747810, 1985, 214 pp, IOPP, Bristol (UK).

Physics of Radiotherapy X-Rays from Linear Accelerators, Metcalfe P et al., ISBN: 0-944838-76-6, 1997, 493 pp, Medical Physics Publishing, Madison (WI).

Practical Essentials of Intensity Modulated Radiation Therapy, Chao KSC, Smith Apisarnthanarax, and Gokhan Ozyigit, ISBN: 0-7817-5279-5, 2004, 324 pp, Advanced Medical Publishing, Madison (WI).

Practical Manual of Brachytherapy, Pierquin B, Marinello G, ISBN: 0-944838-73-1, 1997, 296 pp, Medical Physics Publishing, Madison (WI).

Primer on Theory and Operation of Linear Accelerators, 2nd ed., Karzmark CJ, Morton R, ISBN: 0-944838-66-9, 1998, 50 pp, Medical Physics Publishing, Madison (WI).

Principles and Practice of Brachytherapy, Nag S, editor, ISBN: 0879936541, 1997, 752 pp, Futura.

Principles and Practice of Brachytherapy Using Afterloading Systems, Joslin CA, Flynn A, Hall EJ, editors, ISBN: 0-340-74209-7, 2001, 464 pp, Edward Arnold, London.

Protocol and Procedures for Quality Assurance of Linear Accelerators, Constantinou, ISBN: 0-9638266-0-3, 1993, 92 pp, Medical Physics Publishing, Madison (WI).

Quality Assurance in Radiotherapy Physics, Starkschall G, ISBN: 0944838219, 1991, 387 pp, Medical Physics Publishing, Madison (WI).

Radiation Therapy Planning, Bentel GC, ISBN: 0070051151, 1995, 643 pp, McGraw-Hill, New York.

Radiotherapy In Practice – Brachytherapy, Hoskin PJ, Coyle C, ISBN: 0198529406, 2005, 224 pp, Oxford University Press, New York.

Study Guide for Radiation Oncology Physics Board Exams, Berman B, Thomadsen B, ISBN: 0-944838-94-4, 2000, 112 pp, Medical Physics Publishing, Madison (WI).

The Physics and Radiobiology of Fast Neutron Beams, Bewley DK, ISBN: 085274093x, 1989, 192 pp, IOPP, Bristol (UK).

The Physics of Conformal Radiotherapy: Advances in Technology, Webb S, ISBN: 0750303972, 1997, 382 pp, IOPP, Bristol (UK).

The Physics of Modern Brachytherapy for Oncology, Baltas D, Kreiger H, Zamboglou N, ISBN: 0750307080, 2005, 450 pp, IOPP, Bristol (UK).

The Physics of Three Dimensional Radiation Therapy: Conformal Radiotherapy, Radiosurgery and Treatment Planning, Webb S, ISBN: 075030247x, 1993, 373 pp, IOPP, Bristol (UK).

The Q Book – The Physics of Radiotherapy X-Rays: Problems and Solutions Metcalfe P et al., ISBN: 0-944838-86-3, 1998, 100 pp, Medical Physics Publishing, Madison (WI).

The Theory & Practice of Intensity Modulated Radiation Therapy, Sternick S, editor, ISBN: 1-883526-05-1, 1997, 256 pp, Advanced Medical Publishing, Madison (WI).

The Use of Computers in Radiation Therapy: Schlegel W, Bortfeld T, editors, ISBN: 3540671765, 2000, 604 pp, Springer, New York.

The Use of Plane Parallel Ionization Chambers in High Energy Electron and Photon Beams: An International Code of Practice for Dosimetry, IAEA, ISBN: 9201048963, 1997, 125 pp, IAEA.

Therapy Physics Review, Paliwal B, ISBN: 0-944838-67-7, 1996, 65 pp, Medical Physics Publishing, Madison (WI).

Three-Dimensional Radiation Treatment: Technological Innovations and Clinical Results, Kneschaurek P, Molls M, Feldmann HJ, ISBN: 3805569475, 2000, S. Karger.

Topics in Dosimetry & Treatment Planning for Neutron Capture Therapy, Zamenhof RG, Solares GR, Harling OK, editors, ISBN: 1-883526-02-7, 1994, 245 pp, Advanced Medical Publishing, Madison (WI).

Treatment Planning in Radiation Oncology, Khan F M, Potish R, editors, ISBN: 0-683-04607-1, 1997, 608 pp, Lippincott, New York.

DIAGNOSTIC RADIOLOGICAL PHYSICS

General

AAPM Monograph No. 3, The Physics of Medical Imaging: Recording System Measurements and Techniques (1979 Summer School), Haus AG, editor, ISBN: 0-88318-260-2, 624 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 23, The Expanding Role of Medical Physics in Diagnostic Imaging, Frey GD,

Sprawls P, editors, ISBN: 1-888340-09-6, 1997, 583 pp, Medical Physics Publishing, Madison (WI).

Christensen's Physics of Diagnostic Radiology, Curry T S III, Dowdey JE, Murry RC Jr, ISBN: 0812113101, 1990, 522 pp, Lippincott, New York.

IPEM Report 61, Physics in Diagnostic Radiology, Faulkner K, Cranley K, Starritt HC, Wankling PF, editors, ISBN: 090418160X, 1990, 150 pp, IPEM, York (UK).

Physics for Diagnostic Radiology, Dendy P P, Heaton B, ISBN: 0750305916, 1999, 446 pp, IOPP, Bristol (UK).

Practical Radiography, Robert Ward, ISBN: 0-944838-49-9, (1996 reprint), 112 pp, Medical Physics Publishing, Madison (WI).

The Physics of Diagnostic Imaging, Dowsett DJ, Johnston RE, Kenny PA, ISBN: 0412460602, 1998, 609 pp, Edward Arnold, London.

Topical

AAPM Monograph No. 4, Quality Assurance in Diagnostic Radiology, Waggener R, Wilson C, editors, ISBN: 0-883182-68-8, 1977, 190 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 30, Specifications, Performance Evaluation and Quality Assurance of Radiographic and Fluoroscopic Systems in the Digital Era, Goldman L, Yester M, editors, ISBN: 1-930524-21-8, 2004, 300 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 4, Basic Quality Control in Diagnostic Radiology, Task Force On Quality Assurance Protocol; ISBN: 0-88318-251-3, 1977, 57 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 14, Performance Specifications and Acceptance Testing for X-Ray Generators and Automatic Exposure Control Devices, AAPM, ISBN: 0-88318-461-3, 1985, 96 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 15, Performance Evaluation and Quality Assurance in Digital Subtraction Angiography, Diagnostic X-Ray Imaging Committee/ DigitalRadiography/Fluorography Task Group; ISBN: 0-88318-482-6, 1985, 36 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 29, Equipment Requirements and Quality Control for Mammography, Diagnostic X-Ray Imaging Committee Task Group 7, ISBN: 0-88318-807-4, 1990, 72 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 31, Standardized Methods for Measuring Diagnostic X-Ray Exposures, Diagnostic X-Ray Imaging Committee Task Group 8, ISBN: 0-88318-874-0, 1990, 22 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 58, Managing the Use of Fluoroscopy in Medical Institutions, Radiation Protection Committee Task Group 6; ISBN: 1-888340-13-4, 1998, 42 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 60, Instrumentation Requirements of Diagnostic Radiological Physicists, Diagnostic X-Ray Imaging Committee Task Group 4; ISBN: 1-888340-15-0, 1998, 40 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 70, Cardiac Catheterization Equipment Performance, Diagnostic X-ray Imaging Committee, Task Group 17, ISBN: 1-888340-28-2, 2001, 71 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 74, Quality Control in Diagnostic Radiology, Task Group 12 ISBN: 1-888340-33-9, 2002, 77 pp, Medical Physics Publishing, Madison (WI).

Advances in Film Processing Systems Technology and Quality Control in Medical Imaging, Haus AG, ISBN: 1-930524-01-3, 2001, 245 pp., Medical Physics Publishing, Madison (WI).

Basics of Film Processing in Medical Imaging, Haus A, Jaskulski S, ISBN: 0-944838-78-2, 1997, 338 pp, Medical Physics Publishing, Madison (WI).

Digital Mammography Proceedings, Yaffe M, ISBN: 1-930524-00-5, 2001, 856 pp Medical Physics Publishing, Madison (WI).

Interventional Fluoroscopy: Physics, Technology, Safety, Balter S, ISBN: 0471390100, 2001, 284 pp, Wiley, New York.

IPEM Report No. 32, Measurement of the Performance Characteristics of Diagnostic X-Ray Systems Used in Medicine. Part 1: X-Ray Tubes and Generators, Cranley K ISBN: 090418174X, 1995, 28 pp, IPEM, York (UK).

IPEM Report No. 32, Measurement of the Performance Characteristics of Diagnostic X-Ray Systems Used in Medicine. Part II: X-Ray Image Intensifier Television Systems, Starritt H C, ISBN: 0904181758, 1996, 61 pp, IPEM, York (UK).

IPEM Report No. 32, Measurement of the Performance Characteristics of Diagnostic X-Ray Systems Used in Medicine. Part IV: X-Ray Intensifying Screens, Films, Processors and Automatic Exposure Control Systems, Holubinka M R, ISBN: 0904181774, 1996, 43 pp, IPEM, York (UK).

IPEM Report No. 32, Measurement of the Performance Characteristics of Diagnostic X-Ray Systems Used in Medicine. Part V: Conventional Tomographic Equipment, ISBN: 0904181782, 1996, 18 pp, IPEM, York (UK).

IPEM Report No. 32, Measurement of the Performance Characteristics of Diagnostic X-Ray Systems Used in Medicine. Part VI: X-Ray Image Intensifier Fluorography Systems, Robertson J, ISBN: 0904181790, 1995, 21 pp, IPEM, York (UK).

IPEM Report No. 59, The Commissioning & Routine Testing of Mammographic X-Ray Systems 2nd ed., Law J, Dance DR, Faulkner K, Fitzgerald MC,

Ramsdale ML, Robinson A, ISBN: 0904181723, 1994, 119 pp, IPEM, York (UK).

IPEM Report No. 67 Quality Assurance in Dental Radiology, Starritt HC, Faulkner K, Wankling PF, Cranley K, Robertson J, Young K, ISBN: 0904181677, 1994, 25 pp, IPEM, York (UK).

IPEM Report No. 77, Recommended Standards for Routine Testing of Diagnostic X-Ray Imaging Systems, ISBN: 0904181871, 1997, 64 pp, IPEM, York (UK).

IPEM Report No. 78, Catalogue of Diagnostic X-Ray Spectra & Other Data Cranley K, Gilmore BJ, Fogarty GWA, Desponds L, ISBN: 090418188X, 1997, IPEM, York (UK).

IPEM Report No. 79, The Critical Examination of X-Ray Generating Equipment in Diagnostic Radiology, ISBN: 0904181898, 1998, 17 pp, IPEM, York (UK).

Mammography Quality Control: The Why & How Book, Calvin Myers, ISBN: 0-944838-83-9, 1997, 43 pp, Medical Physics Publishing, Madison (WI).

Medical Imaging and Radiation Protection for Medical Students and Clinical Staff, Martin CJ, Dendy PP, Corbett RH, 2003, The British Institute of Radiology.

Practical Digital Imaging and PACS (1999 AAPM Summer School), Seibert A et al., ISBN: 0-944838-92-8, 1999, 577 pp, Medical Physics Publishing, Madison (WI).

Principles of Radiographic Imaging: An Art and a Science, Carlton RR, Adler A, ISBN: 0766813002, 2000, 752 pp, Thomson Delmar Learning.

Radiation Exposure and Image Quality in X-Ray Diagnostic Radiology: Physical Principles and Clinical Applications, Aichinger H, Dierker J, Joite-Barfuß S, Säbel M, ISBN: 3540442871, 2003, 212 pp, Springer, New York.

Radiology Review: Radiologic Physics, Nickoloff EL, Naveed Ahmad, ISBN: 1416022600, 2005, 272 pp, Saunders, Philadelphia.

Screen Film Mammography: Imaging Considerations and Medical Physics Responsibilities, Gary Barnes and G. Donald Frey, ISBN: 0-944838-12-X, 1991, 127 pp, Medical Physics Publishing, Madison (WI).

IMAGING

General

3D Imaging in Medicine, 2nd ed., Udupa JK, Herman GT, ISBN: 084933179X, 1999, 384 pp, CRC Press, Boca Raton (FL).

Essentials of Diagnostic Imaging, Guebert G M, Pirtle OL, Yochum TR, ISBN: 0801674557, 1995, 252 pp, Mosby, Philadelphia.

Foundations of Image Science, Barrett HH, Myers K, ISBN: 0471153001, 2003, 1100 pp, Wiley, New York.

Fundamentals of Medical Imaging, Paul Suetens, ISBN: 0521803624, 2002, 294 pp, Cambridge University Press, New York.

Handbook of Medical Imaging, Volume 1: Physics and Psychophysics, Beutel J, Kundel HL, Van Metter RL, ISBN: 0819436216, 2000, 968 pp, SPIE.

Introduction to Biomedical Imaging, Webb AG, ISBN: 0471237663, 2002, 264 pp, Wiley, New York.

Introduction To The Principles of Medical Imaging, Guy C, ISBN: 1860945023, 2005, 400 pp, Imperial College Press, London.

Medical Imaging 2004: Physics of Medical Imaging (SPIE Proceedings), Yaffe MJ, ISBN: 0819452815, 2004, SPIE.

Medical Imaging Physics, 4th ed., Hendee WR, Ritenour ER, ISBN: 0-471-38226-4, 2002, 536 pp, Wiley, New York.

Physics for Medical Imaging, Farr RF, Allisy-Roberts PJ, ISBN: 0702017701, 1997, 276 pp, Bailliere Tindall.

Physics of Medical Imaging, Dobbins JT, Boone JM, ISBN: 0819427810, 1998, 842 pp, SPIE.

Principles of Imaging Science and Protection, Thompson MA, Hattaway MP, Hall JD, ISBN: 0721634281, 1994, 522 pp, Saunders, Philadelphia.

Principles of Medical Imaging, Kirk Shung K, Smith MB, Tsui BMW, ISBN: 0126409706, 1992, 289 pp, Academic Press.

The Essential Physics of Medical Imaging, Hardbound, 2nd ed., Bushberg JT, Seibert JA, Leidholdt EM Jr, Boone JM, ISBN: 0-683-30118-7, 2001, 965 pp, Lippincott.

The Physics of Diagnostic Imaging, 2nd ed., Dowsett DJ, Kenny PA, Johnston RE, ISBN: 0412460602, 2005, Hodder Headline (Arnold).

The Physics of Medical Imaging, Webb S, ISBN: 0852743491, 1988, 633 pp, IOPP, Bristol (UK).

Topical

AAPM Monograph No. 11, Electronic Imaging in Medicine, Fullerton GD, Hendee W, Lasher J, Properzio W, Riederer S, editors, ISBN: 0-88318-454-0, 1983, 484 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 12, Recent Developments in Digital Imaging (1984 Summer School), Doi K, Lanzl L, Lin P-J P, editors, ISBN: 0-88318-463-X, 1984, 576 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 25, Practical Digital Imaging and PACS (1999 AAPM Summer School), Seibert A et al., ISBN: 0-944838-92-8, 1999, 577 pp, Medical Physics Publishing, Madison (WI).

Electrical Impedance Tomography: Methods, History and Applications, Holder DS, ISBN: 0750309520, 2005, 456 pp, IOPP, Bristol (UK).

Mathematics of Medical Imaging, Epstein CL, ISBN: 0130675482, 2003, 768 pp, Prentice Hall, New York.

Medical Imaging Signals and Systems, Prince JL, Links J, ISBN: 0130653535, 2005, 550 pp, Prentice Hall, New York.

Wavelet Analysis with Applications to Image Processing, Lakshman Prasad and S. Sitharama Iyengar, ISBN: 0849331692, 1997, 304 pp, CRC Press, Boca Raton (FL).

COMPUTERIZED TOMOGRAPHY

General

AAPM Monograph No. 6, Medical Physics of CT and Ultrasound: Tissue Imaging and Characterization (1980 Summer School), Fullerton GD, Zagzebski J, editors, ISBN: 1-888340-08-8, 1980, 717 pp, Medical Physics Publishing, Madison (WI).

Computed Tomography: Fundamentals, System Technology, Image Quality, Applications, Kalender WA, ISBN: 3-8957-8081-2, 2000, 220 pp, Advanced Medical Publishing, Madison (WI).

CT Physics: The Basics, Villafana T, ISBN: 0683307118, 2002, 250 pp, Lippincott.

Topical

AAPM Report No. 1, Phantoms for Performance Evaluation and Quality Assurance of CT Scanners, AAPM, ISBN: 1-888340-04-5, 1977, 23 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 39, Specification and Acceptance Testing of Computed Tomography Scanners, Diagnostic X-Ray Imaging Committee Task Group 2; ISBN: 1-56396-230-6, 1993, 95 pp, Medical Physics Publishing, Madison (WI).

IPEM Report No. 32, Measurement of the Performance Characteristics of Diagnostic X-Ray Systems Used in Medicine. Part III: Computed Tomography X-Ray Scanners, ISBN: 0904181766, 2003, 94 pp, IPEM, York (UK).

NUCLEAR MEDICINE

General

AAPM Monograph No. 10, Physics of Nuclear Medicine: Recent Advances (1983 Summer School), Rao D, Chandra R, Graham M, editors, ISBN: 0-88318-440-0, 1983, 560 pp, Medical Physics Publishing, Madison (WI).

Diagnostic Nuclear Medicine: A Physics Perspective, Hamilton DI, ISBN: 3540006907, 2004, 465 pp, Springer, New York.

Essentials of Nuclear Medicine Physics, Powsner RA, Powsner ER, ISBN: 0632043148, 1998, 199 pp, Blackwell, Cambridge (MA).

Handbook of Nuclear Medicine, Madsen M, Ponto J, ISBN: 0-944838-14-6, 1992, 114 pp, Medical Physics Publishing, Madison (WI).

Introductory Physics of Nuclear Medicine, Ramesh Chandra, ISBN: 0812114426, 1992, 221 pp, Lea & Febiger, Philadelphia (PA).

Nuclear Medicine and PET: Technology and Techniques, Christian PE, Bernier D, Langan JK, ISBN: 0323019641, 2003, 640 pp, Mosby.

Nuclear Medicine Physics: The Basics, Ramesh Chandra, ISBN: 068330092X, 1998, 182 pp, Lippincott, New York.

Physics in Nuclear Medicine, Cherry SR, Sorenson J, Phelps M, ISBN: 072168341X, 2003, 523 pp, Saunders, Philadelphia.

Practical Nuclear Medicine, 3rd ed., Sharp PF, Gemmell HG, Murray AD, ISBN: 185233875X, 2005, 352 pp, Springer, New York.

Principles and Practice of Nuclear Medicine, Early PJ, Bruce D,. Sodee MD, ISBN: 0801625777, 1995, 877 pp, Mosby.

Topical

AAPM Manual No. 1: Nuclear Medicine Instrumentation Laboratory Exercises for Radiology Residency Training, Van Tuinen R J, Grossman LW, Kereiakes JG, editors, ISBN: 0-88318-0001, 1994, 81 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 1, Biophysical Aspects of Medical Use of Technetium-99m, Kereiakes JG, Corey KR, editors, ISBN: 1-888340-05-3, 1976, 126 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 18, Expanding the Role of Medical Physics in Nuclear Medicine (1989 Summer School), Frey GD, Yester MV, editors, ISBN: 0-883189-15-1, 1989, 368 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 6, Scintillation Camera Acceptance Testing & Performance Evaluation, Nuclear Medicine Committee; ISBN: 0-88318-275-0, 1980, 23 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 9, Computer-Aided Scintillation Camera Acceptance Testing, Nuclear Medicine Committee Task Group; ISBN: 0-88318-407-9, 1981, 40 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 22, Rotating Scintillation Camera SPECT Acceptance Testing and Quality Control, Nuclear Medicine Committee Task Group; ISBN: 0-88318-549-0, 1987, 26 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 52, Quantitation of SPECT Performance (Reprinted from Medical Physics, Vol. 2, Issue 4), Nuclear Medicine Committee Task Group 4; ISBN: 1-56396-485-6, 1995, 10 pp, Medical Physics Publishing, Madison (WI).

IPEM Report No. 65, Quality Standards in Nuclear Medicine, Edited by Hart GC and Smith AH, ISBN: 0904181642, 1992, 123 pp, IPEM, York (UK).

IPEM Report No. 66, Quality Control of Gamma Cameras & Associated Computer Systems, Hannan J, editor, ISBN: 0904181650, 1992, 62 pp, IPEM, York (UK).

IPEM Report No. 85, Radioactive Sample Counting – Principles and Practice, Driver I, editor, ISBN: 0904181995, 2002, 63 pp, IPEM, York (UK).

IPEM Report No. 86, Quality Control of Gamma Camera Systems, Bolster A, ISBN: 1903613132, 2003, 130 pp, IPEM, York (UK).

IPEM Report No. 87, Basics of Gamma Camera Positron Emission Tomography Hillel P, editor, ISBN: 1903613183, 2004, 73 pp, IPEM, York (UK).

Positron Emission Tomography: Basic Sciences, Bailey DL, Townsend DW, Valk PE, Maisey MN, ISBN: 1852337982, 2005, 382 pp, Springer, New York.

Principles and Practice of Positron Emission Tomography, Wahl RL, ISBN: 0781729041, 2002, 442 pp, Lippincott, New York.

Therapeutic Applications of Monte Carlo Calculations in Nuclear Medicine, Habib Zaidi, ISBN: 0750308168, 2003, 363 pp, IOPP, Bristol (UK).

MAGNETIC RESONANCE IMAGING AND SPECTROSCOPY

General

AAPM Monograph No. 14, NMR in Medicine: The Instrumentation and Clinical Applications (1985 Summer School), Thomas SR, Dixon RL, editors, ISBN: 0-88318-497-4, 1985 595 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 21, The Physics of MRI, Michael Bronskill, Sprawls P, editor, ISBN: 1-563962-05-5, 1992, 784 pp, Medical Physics Publishing, Madison (WI).

Magnetic Resonance Imaging: Physical Principles and Sequence Design, Haacke EM et al., ISBN: 0471351288, 1999, 914 pp, John Wiley & Sons, Inc., New York.

Magnetic Resonance Imaging: Principles, Methods, and Techniques, Sprawls P, ISBN: 0-944838-97-9, 2000, 200 pp, Medical Physics Publishing, Madison (WI).

Magnetic Resonance Imaging: Theory and Practice, Vlaardingerbroek MT, Den Boer JA, ISBN: 3540600809, 1996, 347 pp, Springer, New York.

Magnetic Resonance in Medicine, 4th ed., Peter Rinck, ISBN: 0632059869, 2001, 245 pp, Blackwell.

Non-Mathematical Approach to Basic MRI, Smith H, Ranallo F, ISBN: 0-944838-02-2, Published: 1989, 203 pp, Medical Physics Publishing, Madison (WI).

Topical

AAPM Report No. 20, Site Planning for Magnetic Resonance Imaging Systems, Nuclear Magnetic Resonance Committee Task Group 2; ISBN: 0-88318-530-X, 1986, 60 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 28, Quality Assurance Methods and Phantoms for Magnetic Resonance Imaging (Reprinted from Medical Physics, Vol. 17, Issue 2), Nuclear Magnetic Resonance Committee Task Group 1, ISBN: 0-88318-800-7, 1990, 9 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 34, Acceptance Testing of Magnetic Resonance Imaging Systems (Reprinted from Medical Physics, Vol. 19, Issue 1), Nuclear Magnetic Resonance Committee Task Group 6; ISBN: 1-56396-028-1, 1992, 13 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 77, Practical Aspects of Functional MRI, Nuclear Medicine Committee Task Group 8, ISBN: 1-888340-37-1, 2002, 22 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 78, Proton Magnetic Resonance Spectroscopy in the Brain, Magnetic Resonance Task Group 9, ISBN: 1-888340-38-X, 2002, 21 pp, Medical Physics Publishing, Madison (WI).

Handbook of MRI Pulse Sequences, Matt Bernstein, Kevin King, Xiaohong Joe Zhou, ISBN: 0120928612, 2004, 1040 pp, Academic Press.

Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques, Buxton RB, ISBN: 0521581133, 2001, 536 pp, Cambridge University Press, New York.

IPEM Report No. 80, Quality Control in Magnetic Resonance Imaging, RA Lerski, J De Wilde, D Boyce and J Ridgeway, ISBN: 0904181901, 1999, 50 pp, IPEM, York (UK).

Principles of Magnetic Resonance Imaging: A Signal Processing Perspective, Liang Z-P, Lauterbur PC, ISBN: 0780347234, 1999, 416 pp, Wiley, New York.

ULTRASOUND PHYSICS

General

AAPM Monograph No. 6, Medical Physics of CT and Ultrasound: Tissue Imaging and Characterization (1980 Summer School), Fullerton GD, Zagzebski J, editors, ISBN: 1-888340-08-8, 1980, 717 pp, Medical Physics Publishing, Madison (WI).

Advances in Ultrasound Techniques and Instrumentation, Wells PNT, ISBN: 0443088535, 1993, 192 pp, Saunders, Philadelphia.

Clinical Ultrasound Physics: A Workbook for Physicists, Residents, and Students, Kofler JMJr, et al., ISBN: 1-930524-06-4a, 2001, 85 pp, Medical Physics Publishing, Madison (WI).

Diagnostic Ultrasound: Physics and Equipment, Hoskins P, Thrush A, Martin K, Whittingam T, Hoskins PR, Thrush A, Whittingham T, ISBN: 1841100420, 2002, 208 pp, Cambridge University Press, New York.

Essentials of Ultrasound Physics, Zagzebski J A, ISBN: 0815198523, 1996, 220 pp, Mosby.

Physical Principles of Medical Ultrasonics, Hill CR, Bamber JC, ter Haar GR, ISBN: 0471970026, 2002, 528 pp, Wiley, New York.

Physics and Instrumentation of Diagnostic Medical Ultrasound, Fish P, ISBN: 0471958956, 2005, 250 pp, Wiley, New York.

Principles and Applications of Ultrasound, Langton CM, ISBN: 0750308052, 2005, 252 pp, IOPP, Bristol (UK).

The Physics of Clinical MR Taught Through Images, Runge VM, Nitz WR, Schmeets SH, Faulkner WH, Desai NK, ISBN: 1588903222, 2004, 221 pp, Thieme Med. Pub.

Ultrasound in Medicine, Duck FA, Baker AC, Starritt HC, editors, ISBN: 0750305932, 1998, 314 pp, IOPP, Bristol (UK).

Ultrasound Physics Mock Exam, Owen CA, Zagzebski JA, ISBN: 0941022633, 2004, Davies, Inc.

Topical

AAPM Report No. 8, Pulse Echo Ultrasound Imaging Systems: Performance Tests and Criteria, General Medical Physics Committee Ultrasound Task Group; ISBN: 0-88318-283-1, 1980, 73 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 65, Real-Time B-Mode Ultrasound Quality Control Test Procedures (Reprinted from Medical Physics, Vol. 25, Issue 8), Ultrasound Task Group 1; ISBN: 1-888340-22-3, 1998, 22 pp, Medical Physics Publishing, Madison (WI).

Basic Doppler Physics, Smith H, Zagzebski J, ISBN: 0-944838-15-4, 1991, 136 pp, Medical Physics Publishing, Madison (WI).

Doppler Ultrasound: Physics, Instrumental, and Clinical Applications, 2nd ed., Evans DH, McDicken WN, ISBN: 0471970018, 2000, 456 pp, Wiley, New York.

IPEM Report No. 70, Testing of Doppler Ultrasound Equipment, Hoskins PR, Sherriff SB, Evans JA, ISBN: 0904181715, 1994, 136 pp, IPEM, York (UK).

IPEM Report No. 71, Routine Quality Assurance of Ultrasound Imaging Systems ISBN: 0904181820, 1995, 66 pp, IPEM, York (U.K.).

IPEM Report No. 84, Guidelines for the Testing and Calibration of Physiotherapy Ultrasound Machines, Pye S, Zequiri B, ISBN: 0904181987, 2001, 67 pp, IPEM, York (UK).

Safety of Diagnostic Ultrasound, Barnett SB, Kossoff G, editors, ISBN: 1850706468, 1997, 147 pp, Taylor & Francis.

The Safe Use of Ultrasound in Medical Diagnosis, ter Haar G, Duck FA, ISBN 0-905749-42-1, 2000, 120 pp, British Medical Ultrasound Society and British Institute of Radiology, London (UK).

Three-Dimensional Ultrasound, Downey B, Pretorius DH, Fenster A, ISBN: 0-7817-1997-6, 1999, 272 pp, Lippincott Williams & Wilkins, Philadelphia.

LIGHT AND LASERS

General

AAPM Report No. 3, Optical Radiations in Medicine: A Survey of Uses, Measurement and Sources, AAPM, ISBN: 1-888340-06-1, 1977, 28 pp, Medical Physics Publishing, Madison (WI).

An Introduction to Biomedical Optics, Splinter R, ISBN: 0750309385, 2005, 350 pp, IOPP, Bristol (UK).

Applied Laser Medicine, Breuer H, Krasner N, Okunata T, Sliney D, Berlien H-P, Müller GJ, ISBN: 354067005X, 2004, 740 pp, Springer, New York.

Laser-Tissue Interactions: Fundamentals and Applications, Niemz MH, ISBN: 3540405534, 2003, 305 pp, Springer, New York.

Light, Visible and Invisible, and Its Medical Applications, Newing A, ISBN: 1860941648, 1999, 212 pp, World Scientific, River Edge (NJ).

Topical

AAPM Report No. 57, Recommended Nomenclature for Physical Quantities in Medical Applications of Light, General Medical Physics Committee Task Group 2; ISBN: 1-888340-02-9, 1996, 6 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 73, Medical Lasers: Quality Control, Safety, Standards, and Regulations, Task Group 6, ISBN: 1-888340-31-2, 2001, 68 pp, Medical Physics Publishing, Madison (WI).

IPEM Report No. 76, Ultraviolet and Blue-Light Phototherapy-Principles, Sources, Dosimetry and Safety, Diffey B, Hart G, ISBN: 0904181863, Published: 1997, 56 pp, IPEM, York (U.K.).

Laser Safety, Henderson R, ISBN: 0750308591, 2003, 480 pp, IOPP, Bristol (U.K.).

Laser Systems for Photobiology and Photomedicine, Chester AN, Martellucci S, Scheggi AM, ISBN: 0306438860, 1991, 311 pp, Plenum, New York.

Ultraviolet Radiation in Medicine, Diffey BL, ISBN: 0852745354, 1982, 172 pp, IOPP, Bristol (U.K.).

RADIATION PROTECTION

General

Atoms, Radiation, and Radiation Protection, 2nd ed., Turner JE, ISBN: 0-471-59581-0, 1995, 576 pp, Wiley, New York.

Basic Health Physics: Problems and Solutions, Bevelacqua J, ISBN: 0471297119, 1999, 559 pp, Wiley, New York.

Basic Radiation Protection Technology (2nd edition), Gollnick DA, ISBN: 0916339033, 1988, Pacific Radiation Corp.

Contemporary Health Physics: Problems and Solutions, Bevelacqua J, ISBN: 0471018015, 1995, 456 pp, Wiley, New York.

CRC Handbook of Management of Radiation Protection Programs, 2nd ed., Miller KL, ISBN: 0849337704, 1992, 496 pp, CRC Press, Boca Raton (FL).

Exposure Criteria for Medical Diagnostic Ultrasound: 2. Criteria Based on All Known Mechanisms (NCRP Report, No. 140), ISBN: 0929600738, 2003, NCRP.

Handbook of Health Physics and Radiological Health, Shleien B, Slaback LA Jr, Birky B, ISBN: 0683183346, 1997, 700 pp, Lippincott.

Introduction to Health Physics, 3rd ed., Cember H, ISBN: 00-71054618, 1996, 731 pp, McGraw-Hill, New York.

Medical Effects of Ionizing Radiation, 2nd ed., Mettler FA Jr, Upton AC, ISBN: 0721666469, 1995, 440 pp, Elsevier.

Physics for Radiation Protection, Martin JE, ISBN: 0-471-35373-6, 2000, 713 pp, Wiley, New York.

Practical Radiation Protection and Applied Radiobiology, 2nd ed., Dowd SB, Tilson ER, editors, ISBN: 0721675239, 1999, 368 pp, Saunders, New York.

Practical Radiation Protection in Healthcare, Martin CJ, Sutton DG, editors, ISBN: 0192630822, 2002, 440 pp, Oxford University Press, New York.

Radiation Protection, Seeram E, Travis E, ISBN: 0-397-55032-4, 1996, 320 pp, Lippincott.

Radiation Protection, Kathren RL, ISBN: 0852745540, 1985, 212 pp, IOPP, Bristol (UK).

Radiation Protection, Hallenbeck WH, ISBN: 0873719964, 1994, 288 pp, CRC Press, Boca Raton (FL).

Radiation Protection: A Guide for Scientists and Physicians, Shapiro J, ISBN: 0674745868, 1990, 494 pp, Harvard University Press, Cambridge (MA).

Radiofrequency Radiation Standards: Biological Effects, Dosimetry, Epidemiology, and Public Health Policy, Klauenberg BJ, Grandolfo M, Erwin DN, ISBN: 0306449196, 1995, 476 pp, Kluwer, Norwell (MA).

Topical

AAPM Proceedings No. 4, Radiotherapy Safety, Thomadsen B, editor, ISBN: 0-88318-443-5, Published: 1984, 169 pp, Medical Physics Publishing. Madison (WI).

AAPM Report No. 19, Neutron Measurements Around High Energy X-Ray Radiotherapy Machines, Radiation Therapy Committee Task Group 27, ISBN: 0-88318-518-0, 1986, 34 pp, Medical Physics Publishing. Madison (WI).

AAPM Report No. 25, Protocols for the Radiation Safety Surveys of Diagnostic Radiological Equipment, Diagnostic X-Ray Imaging Committee Task Group 1; ISBN: 0-88318-574-1, 1988, 55 pp, Medical Physics Publishing. Madison (WI).

AAPM Report No. 35, Recommendations on Performance Characteristics of Diagnostic Exposure Meters (Reprinted from Medical Physics, Vol. 19, Issue 1), Diagnostic X-Ray Imaging Committee Task Group 6; ISBN: 1-56396-029-X, 1992, 11 pp, Medical Physics Publishing. Madison (WI).

Exposure of the Pregnant Patient to Diagnostic Radiations, Wagner L, et al., ISBN: 0-944838-72-3, 1997, 259 pp, Medical Physics Publishing, Madison (WI).

How Clean is Clean, How Safe is Safe? Eisenbud M, ISBN: 0-944838-33-2, 1993, 63 pp, Medical Physics Publishing. Madison (WI).

IPEM Report No. 50, Chernobyl: Response of Medical Physics Departments in the UK, Haywood JK, editor, ISBN: 0904181456, 1986, 99 pp, IPEM, York (UK).

IPEM Report No. 63, Radiation Protection in Nuclear Medicine & Pathology Goldstone KE, Jackson PC, Myers MJ, Simpson A E, editors, ISBN: 0904181626, 1991, 190 pp, IPEM, York (UK).

IPEM Report No. 69, Recommendations for the Presentation of Type Test Data for Radiation Protection Instruments in Hospitals, The Radiation Protection Instrument Calibration Working Party of the IPEM, ISBN: 0904181707, 1994, 12 pp, IPEM, York (UK).

IPEM Report No. 72, Safety in Diagnostic Radiology, ISBN: 904181812, 1995, 119 pp, IPEM, York (UK).

IPEM Report No. 75, The Design of Radiotherapy Treatment Room Facilities, Stedeford B, Morgan HM, Mayles WPM, ISBN: 1903613855, 1997, 161 pp, IPEM, York (UK).

IPEM Report No. 82, Cost-Effective Methods of Patient Dose Reduction in Diagnostic Radiology, ISBN: 0904181944, 2001, 50 pp, IPEM, York (UK).

IPEM Report No. 88, Guidance on the Establishment and Use of Diagnostic Reference Levels for Medical X-Ray Examinations, ISBN: 1903613205, 2004, 44 pp, IPEM, York (UK).

Management and Administration of Radiation Safety Programs (HPS 1998 Summer School), Charles Roessler, ISBN: 0-944838-01-4, 1998, 603 pp, Medical Physics Publishing, Madison (WI).

Public Protection from Nuclear, Chemical, and Biological Terrorism, Brodsky A, Johnson RH, Goans RE, editors, ISBN: 1-930524-23-4, Published: July 2004, 872 pp, Medical Physics Publishing, Madison (WI).

Radiation Injuries, Gooden D, ISBN: 33333, 1991, 239 pp, Medical Physics Publishing, Madison (WI).

Radiation Protection Dosimetry: A Radical Reappraisal, Simmons J, Watt D, ISBN: 0-944838-87-1, 1999, 160 pp, Medical Physics Publishing, Madison (WI).

Radiation Protection in Medical Radiography, Statkiewicz Sherer MA, Visconti PJ, Ritenour RE, ISBN: 0323014526, 2002, 336 pp, Mosby.

Radiation Safety and ALARA Considerations for the 21st Century, HPS Midyear Symposium, 2001, ISBN: 1-930524-02-1, 2001, 280 pp, Medical Physics Publishing. Madison (WI).

Shielding Techniques for Radiation Oncology Facilities, 2nd ed., McGinley PH, ISBN: 1-930524-07-2, 2002, 184 pp, Medical Physics Publishing. Madison (WI).

Subject Dose in Radiological Imaging, Ng K-H, Bradley DA, Warren-Forward HM, ISBN: 0-444-82989-x, 1998, Elsevier.

The Invisible Passenger, Radiation Risks for People Who Fly, Barish RJ, ISBN: 188352606X, 1999, 119 pp, Advanced Medical Publishing, Madison (WI).

University Health Physics, Belanger R, Papin PJ, editors, ISBN: 1-930524-15-3, 2003, 408 pp, Medical Physics Publishing. Madison (WI).

RADIATION MEASUREMENTS

General

Applications of New Technology: External Dosimetry, Higginbotham JF, ISBN: 0-944838-68-5, 1996, 464 pp, Medical Physics Publishing, Madison (WI).

Fundamentals of Radiation Dosimetry, Green S, ISBN: 075030913x, 2005, 275 pp, IOPP, Bristol (UK).

Medical Radiation Detectors: Fundamental and Applied Aspects, Kember NF, ISBN: 0750303190, 1994, 236 pp, IOPP, Bristol (UK).

Radiation Detection and Measurement, 3rd ed., Knoll GF, ISBN: 0-471-07338-5, 1999, 816 pp, Wiley, New York.

Radiation Dosimetry: Physical and Biological Aspects, Orton CG, ISBN: 0306420562, 1986, 344 pp, Plenum, New York.

Radiation Instruments, Cember H, ISBN: 1-930524-03-X, 2001, 472 pp, Medical Physics Publishing, Madison (WI).

Topical

A Procedural Guide to Film Dosimetry, Yeo IJ, Kim JO, ISBN: 1-930524-19-6, 2004, 65pp, Medical Physics Publishing, Madison (WI).

AAPM Proceedings No. 11, Kilovoltage X-Ray Dosimetry for Radiotherapy and Radiobiology, Ma C-M, Seuntjens J, editors, ISBN: 1-888340-16-9, 1998, 220 pp, Medical Physics Publishing, Madison (WI).

AAPM Proceedings No. 13, Recent Developments in Accurate Radiation Dosimetry, Seuntjens JP, Mobit PN, editors, ISBN: 1-930524-12-9, 2002, 353 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 12, Evaluation of Radiation Exposure Levels in Cine Cardiac Catheterization Laboratories, Diagnostic Radiology Committee Cine Task Force; ISBN: 0-88318-439-7, 1984, 28 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 63, Radiochromic Film Dosimetry (Reprinted from Medical Physics, Vol. 25, Issue 11), Radiation Therapy Committee Task Group 55; ISBN: 1-888340-20-7, 1998, 23 pp, Medical Physics Publishing, Madison (WI).

Instrumentation, Measurements and Electronic Dosimetry, (HPS Midyear 2000) ISBN: 0-944838-93-6, 2000, 271 pp, Medical Physics Publishing, Madison (WI).

Internal Radiation Dosimetry (1994 HPS Summer School), Raabe O, ISBN: 0-944838-47-2, 1994, 667 pp, Medical Physics Publishing, Madison (WI).

Microdosimetry and Its Applications, Rossi HH, Zaider M, ISBN: 3540585419, 1996, 321 pp, Springer, New York.

Practical Applications of Internal Dosimetry, Bolch WE, editor, ISBN: 1-930524-09-9, 2002, 480 pp, Medical Physics Publishing, Madison (WI).

RADIATION BIOLOGY

General

An Introduction to Radiobiology, Nias AHW, ISBN: 0471975907, 1998, 400 pp, Wiley, New York.

Basic Clinical Radiobiology, Steel GG, ISBN: 0340807830, 2002, 266 pp, Edward Arnold, London.

Biological Risks of Medical Irradiations, Fullerton G, ISBN: 0883182793, 1987, 335 pp, American Institute of Physics, New York.

Effects of Atomic Radiation, Schull WJ, ISBN: 0471125245, 1995, 97 pp, Wiley, New York.

Handbook of Radiobiology, Prasad KN, ISBN: 0849325013, 1995, 352 pp, CRC Press, Boca Raton (FL).

Primer of Medical Radiobiology, Travis E, ISBN: 0815188374, 1989, 302 pp, Mosby, Philadelphia.

Radiobiology for the Radiologist, 5th ed., Hall E J, ISBN: 0-7817-2649-2, 2000, 608 pp, Lippincott.

Topical

A Compilation of Radiobiology Practice Examinations for Residents in Diagnostic Radiology and Radiation Oncology, Chapman JD, Shahabi S, Chapman BA, ISBN: 1 883526 09 4, 2000, 163 pp, Advanced Medical Publishing, Madison (WI).

AAPM Proceedings No. 7, Prediction of Response in Radiation Therapy: Analytical Models and Modelling, Paliwal B et al., ISBN: 0-883186-24-1, 1989, 757 pp, Medical Physics Publishing, Madison (WI).

AAPM Proceedings No. 9, Prediction of Response in Radiation Therapy: Radiosensitivity, Repopulation, Paliwal BR, Herbert D, Fowler JF, Kinsella TJ, editors, ISBN: 1-56396-271-3, 1993, 383 pp, Medical Physics Publishing, Madison (WI).

AAPM Proceedings No. 10, Volume & Kinetics in Tumor Control & Normal Tissue Complications: 5th International Conference on Dose, Time, and Fractionation in Radiation Oncology, Paliwal BR, Herbert D, editors, ISBN: 1-888340-11-8, 1998, 483 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 18, A Primer on Low-Level Ionizing Radiation and Its Biological Effects, Biological Effects Committee; ISBN: 0-88318-514-1, 1986, 103 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 43, Quality Assessment and Improvement of Dose Response Models: Some Effects of Study Weaknesses on Study Findings, Biological Effects Committee Task Group 1; ISBN: 0-944838-45-6, 1993, 351 pp, Medical Physics Publishing, Madison (WI).

Applied Radiobiology and Bioeffect Planning, Wigg D, ISBN: 1-930524-05-6, 2001, 486pp, Medical Physics Publishing, Madison (WI).

Biological Models and Their Applications in Radiation Oncology, Olsen KJ, ISBN: 10883526-03-5, 1994, 61 pp, Advanced Medical Publishing, Madison (WI).

BJR Supplement 26: Chronic Irradiation: Tolerance and Failure in Complex Biological Systems, 2002, The British Institute of Radiology.

Health Effects of Exposure to Low-level Ionizing Radiation, Hendee WR, Edwards FM, editors, ISBN: 0750303492, 1996, 640 pp, IOPP, Bristol (UK).

Health Effects of Exposure to Low Levels of Ionizing Radiation: BEIR V, National Research Council, ISBN: 0309039959, 1989, 436 pp, National Academies Press.

Physics and Radiobiology of Nuclear Medicine, Gopal Saha, 2nd ed., ISBN: 0387950214, 2003, 253 pp, Springer, New York.

Prediction of Tumor Treatment Response, Chapman DJ, Peters LJ, Withers RH, ISBN: 0080346898, 1989, 336 pp, Elsevier.

Radiation Oncology – Radiobiological and Physiological Perspectives, Awwad H K, ISBN: 0792307836, 1990, 688 pp, Kluwer, Norwell (MA).

The Radiation Biology of the Vascular Endothelium, Rubin DB, ISBN: 0849348404, 1997, 272 pp, CRC Press, Boca Raton (FL).

RADIOLOGICAL PHYSICS FOR RADIOLOGICAL TECHNOLOGISTS

General

Introduction to Radiologic Sciences and Patient Care, 3rd ed., Adler AM, Carlton RR, ISBN 0721697828, 2003, 520 pp, Saunders, Philadelphia.

Radiologic Physics and Radiographic Imaging, Bushong SC, ISBN: 0323032648, 2004, Mosby, Philadelphia.

Radiologic Science for Technologists: Physics, Biology and Protection, Bushong SC, ISBN: 0323025552, 2004, 704 pp, Mosby.

The Fundamentals of X-Ray and Radium Physics, Joseph Selman, ISBN: 0398058709, 1994, 637 pp, Charles C. Thomas.

Topical

Magnetic Resonance Imaging: Physical and Biological Principles, Bushong SC, ISBN: 0323014852, 2003, 528 pp, Mosby, Philadelphia.

MRI for Technologists, Woodward P, ISBN: 0071353186, 2000, 408 pp, McGraw-Hill, New York.

Principles and Practice of Radiation Therapy, Washington CM, Leaver D, ISBN: 0323017487, 2004, 992 pp, Elsevier, New York.

Rad Tech's Guide to Mammography: Physics, Instrumentation, and Quality Control, Jacobson DR, Seeram E, ISBN: 0632044993, 2001, 120 pp, Blackwell.

Radiation Protection: Essentials of Medical Imaging Series, Stewart C. Bushong, ISBN: 0070120137, 1998, 288 pp, McGraw-Hill, New York.

The Basic Physics of Radiation Therapy, Selman J, ISBN: 0398056854, 1990, 749 pp, Charles C. Thomas.

The Fundamentals of Imaging Physics and Radiobiology for the Radiologic Technologist, Selman J, ISBN: 0398069875, 2000, 484 pp, Charles C. Thomas.

MATHEMATICS AND STATISTICS

AAPM Monograph No. 13, Multiple Regression Analysis: Applications in the Health Sciences, Herbert DE, Meyers R H, editors, ISBN: 0-88318-490-7, 1984, 598 pp, Medical Physics Publishing, Madison (WI).

Chaos and the Changing Nature of Science and Medicine: An Introduction, Herbert D, editor, ISBN: 1-56396-442-2, 1995, American Institute of Physics, New York.

Mathematical and Computer Modeling of Physiological Systems, Rideout V, ISBN: 0-13-563354-0, 1991, 155 pp, Prentice Hall, New York.

COMPUTERS

General

AAPM Monograph No. 17, Computers in Medical Physics (1988 Summer School), Benedetto A R, Huang HK, Ragan DP, editors, ISBN: 0-88318-802-3, 1988, 417 pp, Medical Physics Publishing, Madison (WI).

Topical

AAPM Report No. 10, A Standard Format for Digital Image Exchange, Task Force on Digital Image Data Exchange of the Science Council; ISBN: 0-88318-408-7, 1982, 11 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 30, E-Mail and Academic Computer Networks, Computer Committee Task Group 1, ISBN: 0-88318-806-6, 1990, 54 pp, Medical Physics Publishing, Madison (WI).

PACS: Basic Principles and Applications, Huang HK, ISBN: 0471253936, 1998, 519 pp, Wiley, New York.

PUBLIC EDUCATION

AAPM Report No. 53, Radiation Information for Hospital Personnel, Radiation Safety Committee; ISBN: 1-56396-480-5, 1995, 24 pp, Medical Physics Publishing, Madison (WI).

Cancer Patient's Guide to Radiation Therapy, Steeves R, ISBN: 0-944838-26-X, 1991, 87 pp, Medical Physics Publishing, Madison (WI).

Radiation and Health, Thormod Henrikson, H. David Maillie, David H. Maillie, ISBN: 0415271622, 2002, 240 pp, Taylor & Francis.

MEDICAL PHYSICS EDUCATIONAL AND PROFESSIONAL ISSUES

AAPM Monograph No. 27, Accreditation Programs and the Medical Physicist: 2001 AAPM Summer School Proceedings, Dixon R, Butler P, Sobol W, editors, ISBN: 1-930524-04-8, 2001, 364 pp. Medical Physics Publishing, Madison (WI).

AAPM Report No. 33, Staffing Levels and Responsibilities of Physicists in Diagnostic Radiology, Diagnostic X-Ray Imaging Committee Task Group 5; ISBN: 0-88318-913-5, 1991, 30 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 36, Essentials and Guidelines for Hospital Based Medical Physics Residency Training Programs, Presidential Ad Hoc Committee on Clinical Training of Radiological Physicists; ISBN: 1-56396-032-X, 1990, 147 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 38, The Role of a Physicist in Radiation Oncology, Professional Information and Clinical Relations Committee Task Group 1; ISBN: 1-56396-229-2, 1993, 12 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 42, The Role of the Clinical Medical Physicist in Diagnostic Radiology, Professional Information and Clinical Relations Committee; ISBN: 1-56396-311-6, 1994, 20 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 64, A Guide to the Teaching of Clinical Radiological Physics to Residents in Diagnostic and Therapeutic Radiology, Committee on the Training of Radiologists; ISBN: 0-944838-09-X, 1999, 32 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 79, Academic Program Recommendations for Graduate Degrees in Medical Physics, Education and Training of Medical Physicists Committee, ISBN: 1-888340-39-8, 2002, 72 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 80, The Solo Practice of Medical Physics in Radiation Oncology, Professional Information and Clinical Relations Committee Task Group 11, ISBN: 1-888340-41-X, 2003, 18 pp, Medical Physics Publishing, Madison (WI).

Current Regulatory Issues in Medical Physics, Martin M, Smathers J, ISBN: 0-944838-29-4, 1992, 458 pp, Medical Physics Publishing, Madison (WI).

Medical Physicists and Malpractice, Shalek R, Gooden D, ISBN: 0-944838-64-2, 1996, 140 pp, Medical Physics Publishing, Madison (WI).

RADIATION PHYSICS

AAPM Proceedings No. 8, Biophysical Aspects of Auger Processes, Howell RW, Narra V, Sastri K, Rap D, editors, ISBN: 1-56396-095-8, 1992, 418 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 37, Auger Electron Dosimetry (Reprinted from Medical Physics, Vol. 19, Issue 6), Nuclear Medicine Committee Task Group 6; ISBN: 1-56396-186-5, 1992, 25 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 49, Dosimetry of Auger-Electron-Emitting Radionuclides (Reprinted from Medical Physics, Vol. 21, Issue 12), Nuclear Medicine Committee Task Group 6; ISBN: 1-56396-451-1, 1994, 15 pp, Medical Physics Publishing, Madison (WI).

Bluebells and Nuclear Energy, Reynolds A, ISBN: 0-944838-63-4, 1996, 302 pp, Medical Physics Publishing, Madison (WI).

Chernobyl Record: The Definitive History of the Chernobyl Catastrophe, Mould RF, 075030670x, 2000, 350 pp, IOPP, Bristol (UK).

My Life with Radiation, Ralph Lapp, ISBN: 0-944838-52-9, 1995, 168 pp, Medical Physics Publishing, Madison (WI).

Radiological Physicists, del Regato JA, ISBN: 0-88318-469-9, 1985, 188 pp, Medical Physics Publishing, Madison (WI).

Understanding Radiation, Wahlstrom B, ISBN: 0-944838-62-6, 1996, 120 pp, Medical Physics Publishing, Madison (WI).

Medical Physics Online Resources

Electronic Medical Physics World, International Organization for Medical Physics (http://www.medphysics.wisc.edu/~empw).

Medical Physics World, International Organization for Medical Physics (http://www.iomp.org); also available in hardcopy.

Global Online Medical Physics Book (http://www.iomp.org).

Organizations That Publish in Medical Physics

American Association of Physicists in Medicine (AAPM), One Physics Ellipse, College Park, MD 20740 (http://www.aapm.org).

American Institute of Physics, 2 Huntington Quadrangle, Melville, NY 11747-4502 (http://aip.org).

American Institute of Ultrasound in Medicine, 14750 Sweitzer Lane, Suite 100, Laurel, MD 20707 (http://www.aium.org).

American Roentgen Ray Society, 44211 Slatestone Court, Leesburg, VA 20176-5109 (http://www.arrs.org).

American Society for Therapeutic Radiology and Oncology, 12500 Fair Lakes Circle, Suite 375, Fairfax, VA 22033-3882 (http://www.astro.org).

British Institute of Radiology, 36 Portland Place, London, W1B 1AT, (http://www.bir.org.uk).

Health Physics Society, 1313 Dolley Madison Boulevard, Suite 402, McLean, Virginia 22101 (http://hps.org).

Institute of Physics and Engineering in Medicine (IPEM), Fairmount House, 230 Tadcaster Road, York, YO24 1ES, UK (http://www.ipem.ac.uk).

International Atomic Energy Agency, P.O. Box 100, Wagramer Strasse 5, A-1400 Vienna, Austria (http://www.iaea.org).

International Commission on Radiation Units and Measurements, Inc., ICRU, 7910 Woodmont Avenue, Suite 400, Bethesda, MD 20814-3095, USA (http://www.icru.org).

International Commission on Radiological Protection, SE-171 16 Stockholm Sweden (Elsevier).

International Society for Magnetic Resonance in Medicine, 2118 Milvia Street, Suite 201, Berkeley, CA 94704, USA (http://www.ismrm.org).

International Society for Optical Engineering (SPIE), PO Box 10, Bellingham WA 98227-0010 USA (http://spie.org).

National Council on Radiation Protection and Measurements, 7910 Woodmont Avenue, Suite 400, Bethesda, MD 20814-3095 (http://www.ncrponline. org).

Radiation Research Society, 10105 Cottesmore Court, Great Falls, VA 22066 (http://www.radres.org).

Radiological Society of North America, Inc., 820 Jorie Boulevard, Oak Brook, IL 60523-2251 (http://rsna.org).

Society of Nuclear Medicine, 1850 Samuel Morse Dr. Reston, VA 20190 (http://snm.org).

See also BIOMEDICAL ENGINEERING EDUCATION; MEDICAL EDUCATION, COMPUTERS IN; MEDICAL ENGINEERING SOCIETIES AND ORGANIZATIONS.

MEDICAL RECORDS, COMPUTERS IN

YESHWANTH SRINIVASAN
BRIAN NUTTER
Texas Tech University
Lubbock, Texas

INTRODUCTION

Computers play a vital role in almost every aspect of modern medical science. If we begin with a primitive definition of a computer, as a system with a processor unit capable of performing arithmetic and logical operations and a memory unit to store programs and data, then almost every sophisticated piece of modern medical equipment comes with a computer. Most standard medical procedures [e.g., magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), ultrasonography] use computers to record and to process data. Computer-based medical records would seem a natural choice for documentation of the recorded data. Moreover, the association of this data with computers presents numerous advantages including easy and random access to the data by multiple distributed users, globally searchable databases, and automated batch processing. However, the use of computer-based medical records is not as prevalent in today's medical community as one would expect. In this article, we explain the reasons for this anomaly, and we present methods to circumvent the problems associated with computer-based medical records and to allow medical practitioners to achieve the real potential of computer-based systems.

PAPER-BASED MEDICAL RECORD SYSTEMS

Medical records chronicle the transactions between healthcare professionals and their patients. They record every detail of a patient's visit to a hospital, clinic, or office from the time of the patient's arrival to the time the patient is discharged. Details recorded include the condition of the patient at the time of examination, procedures and medications administered to the patient, symptoms and observations, diagnoses and test results, together with the exact time of each relevant activity.

Traditionally, medical records have consisted of handwritten documents called *charts*. This remains the most prevalent form of recording patient information, especially in small hospitals and clinics. These charts typically record every piece of information that is judged potentially relevant to a patient's case. When a patient visits a hospital or a primary care physician for the first time, a new chart is made, and the patient name and contact information are recorded on it. When the patient is discharged from the hospital, a discharge note is made on the chart. During subsequent visits, new charts are added. All patient charts are maintained in a labeled file bearing the patient's name and other information to enable easy identification and retrieval.

However, paper-based records exhibit shortcomings in six important areas that are necessary to ensure adequate long-term care to patients and easy preservation of treatments administered.

1. *Accessibility*: Accessibility refers to the immediate availability of patient-related information at all times to all parties with a need to know. It is important to review a patient's medical history before commencing any new treatment. If a patient decides to visit a specialist, emergency room, clinic, or hospital other than the primary care physician, a request may be made for the patient's medical history to be faxed or couriered to the second facility. This information transfer may take between hours to days to complete, depending on the distance and arrangements between the two locations, available modes of communication, availability of authorized personnel to promptly respond to the request, and so on. During

emergencies, a delay of just a few minutes in administering the correct treatment to a patient could have very serious ramifications. A situation in which adverse reactions to particular medications known by a primary care physician simply are not made available in a timely fashion to emergency room personnel is all easy to imagine. The delay in disseminating accurate detailed background information is perhaps the biggest drawback of paper-based medical record systems.

2. *Legibility*: Paper-based records are generally handwritten documents, and one person's handwriting is often not readily understood by another. The fact that entries may be made by many different people makes understanding a chart even more difficult. Furthermore, paper-based records are subject to wear and tear, accidental immersion in water or other liquids due to leaks, misfortune or negligence, smudging of ink, and so on. These factors lead to reduced legibility of the entries on the charts and consequently make interpreting the charts more subjective.
3. *Bulkiness*: Paper-based records for a single patient initially consist of just a few charts, but they can quickly become a thick file with subsequent visits. Assuming that a small clinic sees $\sim$ 10 new patients daily, that the physicians in the clinic take rotating holidays and that it is not a leap year, $\sim$ 3650 new files will be added over a period of 1 year. The files of many of the existing patients also become bulkier and subsequently occupy more space. All of this paper forces the addition of storage space in the clinic for these new records, even though the space could often be better utilized for clinical purposes through the addition of much more profitable and useful treatment or diagnostic capabilities.
4. *Data Dimensionality*: Data dimensionality refers to the ability to interpret the available information in multiple, complementary fashions. The data contained in paper-based records simply provide information about the medical history of a particular patient. In order to obtain trends, distributions or statistical patterns about a particular patient or even groups of people in a geographical area, the required information has to be manually compiled from each record. This can be extremely laborious and time consuming. Furthermore, even a simple text search for a particular word or phrase can take impractically long if the search must be conducted visually through paper-based handwritten records.
5. *Integrity*: A patient will have multiple medical files, maintained simultaneously at every hospital or clinic ever visited, yet none of these records contain the complete medical history of the patient. Usually, the patient's primary care physician is the one who has the records that will provide the most accurate long-term information, and during times of emergency the primary care physician will frequently be the first one contacted for patient records. However, medical facilities rarely have in place any mechanism to update the primary care physician. If, for example, during a visit to a specialist it was found that a patient was allergic to a particular class of medications, the primary care physician's records will often not be updated to show this information. Even if the patient records from multiple sources were summoned, there will still remain a very realistic chance of missing one or more sources, leaving the completeness of paper-based records questionable. Paper-based systems also risk inadvertent loss or misfile of complete pages in copying, transporting, photocopying, faxing, and handling of records.
6. *Replication*: Paper-based medical records are prone to destruction by natural disasters, fires, leaks, and pests in storage areas. Because paper-based records are bulky, it is both expensive and tedious to replicate them and store back-up copies. Hence, if they are destroyed, patient medical histories can be completely lost.

Several research papers have been devoted to explaining the shortcomings of paper-based medical records (1,2). Nevertheless, paper-based records are still the most widely used method for maintaining patient information. The process of recording information in charts is a natural process of making notes, and it requires little specialized training. Once recorded, the information is relatively permanent and cannot be distorted easily. Furthermore, the chart often bears the signatures of both the patient and the appropriate medical personnel who treated the patient, which may be an important legal requirement. Such legal issues are a significant reason that many organizations resist wholesale changes to their paper-based record systems.

COMPUTER-BASED MEDICAL RECORD SYSTEMS

Paper-based records were the only method of storing patient medical information in use until $\sim$ 25 years ago. Since that time, computer technology has experienced exponential growth, creating endless possibilities for developments within allied fields. The fields of medicine in general and medical records in particular have been no exception. Computer-based records were introduced to overcome the inherent limitations and problems of paper-based records and to alter patient medical record keeping in order to incorporate advancements in computer technology. The computer-based medical record is commonly given several names within differing environments, including Computer-based Patient Record (CPR), Electronic Health Record (EHR), and Electronic Medical Record (EMR). Throughout this article, we will use the name computer-based patient record and the abbreviation CPR to refer to computer-based medical records. The CPRs are fundamental to the role of computers in medical records, and a significant portion of this article deals with them.

A CPR is an electronic patient record that resides in a computer-based information system. Such systems are often specifically designed to augment medical practitioners by providing accessibility to complete and accurate long-term data, alerts, reminders, clinical decision support

systems, links to medical knowledge databases, and other aids (3,4). The CPRs are essentially digital equivalents of paper-based records. Because this patient information is stored digitally, the enormous information processing and networking capabilities of computers can be utilized to drastically increase the efficacy of patient treatment. They easily overcome the shortcomings of paper-based records on the six key issues outlined in the previous section, and they provide several other fascinating features exclusive to them.

1. *Accessibility*: With the advent of the Internet and the World Wide Web, information on virtually any topic can be quickly and easily obtained. CPR databases can be stored in networked servers, which can be searched from any properly networked location. With properly designed databases and network systems, the complete medical history of a patient can be retrieved in a matter of seconds. This helps in ensuring the most prompt and accurate treatment possible, which is one of the fundamental reasons for maintaining a medical record.
2. *Legibility*: The information recorded on CPRs is presented using digitally reproduced fonts, which are then independent of the handwriting of the person entering the data into the system. Anytime patient information is required, it can be readily visualized on a viewing device, such as a CRT monitor or PDA, which does not introduce any wear and tear in the record itself. If a paper copy of a patient record is required for a particular activity, a new printout can be generated every time. Thus there is no subjective element in understanding the information.
3. *Bulkiness*: CPRs can be stored on a variety of media, including Compact Disc (CD), Digital Versatile Disc (DVD), and Hard Disk Drive (HDD). A single CD, 11.4 cm in diameter and weighing $\sim$15 g, can store 700 megabytes (MB) of data. This translates into $>$100,000 pages of raw text or two hundred 500-page books, which would occupy $>$10 m of shelf space. The DVDs have $>$10 times the density of CDs, and HDDs offer $>$100 times the information content of a CD. This means that with proper database design, all of the medical records of all of 1 year's new patients at the aforementioned small clinic can now be stored on a single CD, and several years of patient data for all ongoing patient activities can be stored on just one DVD.
4. *Data Dimensionality*: Data on CPRs can be exploited and interpreted in many more ways than can a simple paper-based record. Entire databases can be subjected to automated search, and specific information from a single record or a collection of records can be retrieved in a matter of seconds. Trends and patterns in disease occurrence and treatment administration and effectiveness can readily be extracted and processed for limited or wide-ranging populations. Database Management System (DBMS) software can be used to plot the available information on graphs and pie charts, which greatly simplify the contents of medical records into a form that can be interpreted by laymen. All these operations can be done automatically using state-of-the-art software and require little, if any, human intervention. Techniques in data mining, for example, allow automated determination and analysis of data correlations among very large numbers of variables.
5. *Integrity*: CPRs are generally stored on local systems within hospitals, clinics, and physician offices, which means that medical records of a particular patient could be distributed across many systems throughout the world. Using modern networking techniques, these systems can be easily yet securely networked, and widely dispersed CPRs can be located and made available for inclusion in patient diagnosis and treatment. It would also be practical to develop a system in which all patient records are stored in a central repository, and any time new records of the patient are created or updated, they can be added to this central repository. Using either of these methods in a well-designed system, the complete medical history of a patient can be very rapidly retrieved from any of the authorized locations connected to the network.
6. *Replication*: Due to the high densities for computerized storage of data, CPRs occupy very little space when compared to paper-based systems. It is very easy and inexpensive to create back-up copies and to store these copies in geographically diverse locations. In a well-designed computer-based system with automated, networked backup, the complete medical history of every patient can still be retrieved even if a record storage facility is completely destroyed.

Exclusive Features of CPR

With considerable manual effort, the performance of paper-based records can be made comparable to that of CPRs within the six major requirements of a medical record outlined above. The efficiency advantages of computer-based medical record systems will nevertheless offer significant cost advantages over paper-based systems in achieving comparable performance levels. Significant improvements in patient treatment can result when CPRs are incorporated into expert systems and DBMS frameworks that engender an entirely new set of features for medical records.

1. *Content-Based Data Retrieval*: If a large hospital system is informed that a popular drug has been determined to cause serious side effects and that the manufacturer has consequently issued an immediate recall, the hospital system may hesitate to issue a broad public warning, because that action would lead to unnecessary panic and to patients flooding the system seeking confirmation on their prescriptions, even when their medications are completely unrelated to the recall. If the hospital maintained only paper-based records, it would take weeks to pore through patient charts to obtain the names

and locations of the patients to whom the drug was prescribed. Such a situation presents a perfect case for content-based data retrieval. With properly constructed CPRs, the same information can be much more rapidly retrieved through querying the database to fetch all patient records with the name of the recalled drug in the 'Drugs Prescribed' field. Then, only the patients who actually have been prescribed the recalled medication can be individually contacted with explicit instructions by a qualified staff member. This data search technique is known as Content-Based Data Retrieval (CBDR).

2. *Automatic Scheduling*: With CPR-based systems, routine periodic check-up or follow-up visits can be automatically scheduled, and e-mail reminders or automated telephone recordings can be sent directly to the patients. This automation will relieve healthcare professionals of a significant administrative burden.
3. *Knowledge-Based Systems*: Every physician attempts to solve a particular condition to the best of his limited knowledge (5). It is reported that there are $>5{,}000{,}000$ medical facts, which are constantly being appended, updated and questioned by over 30,000 medical journal publications each year (6). It is humanly impossible to remember all of these facts and to reproduce them accurately as needed. However, CPR databases all over the world can be rapidly and effectively searched in order to find information about patients who were previously diagnosed with the same condition, how they were treated and the end result. While the presiding doctor still makes the decisions and retains responsibility, he can have experimental evidence and experiential knowledge from other experts to assist his analysis (7). Diagnostic decision support systems like Quick Medical Reference (QMR) and Massachusetts General Hospital's DXplain provide instant access to knowledge bases of diseases, diagnoses, findings, disease associations and lab information.
4. *Expert Systems*: Human errors in recording information can be greatly reduced by programming the data-entry interface to reject anomalous values and to restrict the input to a finite set of possible choices. For example, a pulse rate of 300 beats min^{-1} would be highly unlikely. Potentially dangerous spelling mistakes in generating a prescription can be identified. Dates for follow-up activities can be checked against reasonable parameters. A well-designed interface would prompt a message to recheck such entries.
5. *In situ Data Access*: A medical records system in which a physician accesses a wireless PDA while conversing with a patient could allow direct bidirectional communication between a physician and a pharmacy, a laboratory, or a specialist. Collaborators on a complex case can simultaneously access each other's progress, even when the collaborators are geographically separated.

In 1991, the United States Institute Of Medicine (IOM) conducted a study on the advantages and disadvantages of CPRs and published its findings (3). The IOM concluded that CPRs greatly enhance the health of citizens and greatly reduce costs of care, and it called for widespread implementation of CPRs by 2001. However, 4 years past that deadline, widespread implementation still remains only a concept. Figures for CPR adoption rates in hospitals range from 3% to 21%, while CPR adoption rates for physician offices range from 20% to 25% (8). It is safe to say that these figures are not representative of the popularity that one would reasonably expect from the benefits of CPRs. The following section is devoted to explaining this anomaly.

ROADBLOCKS IN CPR IMPLEMENTATION

Recording and processing of information with CPRs requires installation of computer systems and software, and, in general, more training of healthcare professionals is required to get them acquainted with CPR systems than with paper-based record systems. Furthermore, realizing the full potential of CPR systems requires not only that patient records issued in the future should be CPRs but also that previously taken paper-based records should be converted into CPRs. Some of the important factors affecting this onerous large-scale conversion will now be discussed.

1. *Cost*: Expense is usually the major obstacle to the full-fledged incorporation of CPRs. Depending on whether the application is for the office of a single physician or for a large hospital, the basic infrastructure needed to establish a CPR system could vary from a single PC to a complete network of workstations and servers loaded with expensive network, database and records management software. Both hardware and software computer technologies become outmoded very quickly, and these technologies require constant update, increasing maintenance and operation costs. There is also a significant cost factor involved in hiring data-entry operators to convert existing paper-based records to CPRs and in training healthcare professionals to then operate the new systems and software. Although several studies have shown that CPRs are effective in the long run (9,10), both hospitals and physician offices remain apprehensive about investing many thousands of dollars on a new technology while the tried and tested paper-based records work reasonably effectively.
2. *Abundance of Choices and Lack of Standards*: There are >100 CPR management software systems available in the market today, and additional entries arrive in the marketplace annually. Each software approach will have its advantages and disadvantages, and it is never easy to decide which one is best suited for a particular application. Furthermore, different software interfaces use different templates for the patient record, recording slightly different patient data within very different structures. A

standardized CPR format independent of the specific hardware and software running on a particular system is yet to be developed. This lack of standardization in CPR formats severely cripples cross-system interaction and comprehensive integration of CPR systems. The marketplace has, to date, failed to produce a clear market leader or an industrial body with the ability to enforce such standards. A customer contemplating a purchase and the subsequent investments in training and maintenance may have serious reservations concerning the ability to change from the products of one manufacturer to another.

3. *Confidentiality*: The United States government passed the Health Insurance Portability and Accountability Act (HIPAA) in 1996, protecting a patient's right to confidentiality and specifying the information on a medical record that can be shared and the information that must be considered confidential. A record can be made available only on a "need to know" basis, wherein only personnel who absolutely need to know the data are granted access. For example, physicians participating in the diagnosis and treatment process of a specific patient are granted complete access to the entire medical record, while health agencies involved in conducting demographic studies are granted access only to information that does not and can not reveal the identity of the patient.
4. *Security*: In the case of paper-based records, it is relatively easy to control access and to ensure security by keeping those records under lock and key. The security of CPRs will be constantly threatened by hackers, trying to break into a system in order to retrieve confidential patient information. Unless properly protected, CPRs are also vulnerable to being compromised during transmission across a network. Unlike paper-based records, CPRs can be easily manipulated, intentionally or unintentionally, in fashions that do not appear obvious. These vulnerabilities represent an enormous security problem, and effective solutions are required before CPRs can be put to widespread use. Furthermore, efforts to provide network security have significant costs, requiring manpower, equipment, and frequent software upgrades.
5. *Apprehension Among Patients*: Even in developed countries like the United States, considerable uneasiness exists among patients about the use of computers to maintain patient medical records. While some of this phenomenon can be ascribed to the reluctance of the older generation in accepting new technology, the most important reason for this apprehension is the lack of clear understanding by patients about the benefits of CPRs. One of the easiest ways to overcome this fear is for healthcare professionals to directly reach out to the patients and explain the long-term advantages of CPRs.
6. *Legal Issues*: The validity of CPRs in a court of law is still questionable. The Millennium Digital Commerce Act, otherwise known as the Electronic Signature Act, was passed in 2000 (although the portion relating to records retention was passed in 2001). The act established that the electronic signature and electronic records cannot be denied legal effect simply because they are electronic and are not signed in ink on paper. However, legal implications in CPR transactions extend far beyond this act. For example, if it is found that the software system used to manage CPRs does not comply with stipulated security requirements, the validity of a CPR can be questioned in court, even if no indication of tampering can be identified. The CPR systems are also prone to computer viruses, worms and Trojans, which exploit loopholes in the computer security defenses to gain unauthorized access to sensitive data and/or to maliciously corrupt the information.

Despite significant progress toward overcoming these hurdles, complete transition from paper-based records to CPRs is a time-consuming and capital-intensive process, and during the midst of this transition, the benefits are unlikely to be immediately apparent. In the following section, some of the options available to healthcare professionals in implementing a CPR system, and several technologies available to ensure secure storage and transmission of confidential medical information are reviewed.

FEATURES OF CPR SYSTEMS

A complete CPR system can consist of individual workstations, data input devices (keyboards, mice, light pens, scanners, cameras), output devices (printers, plotters, and display monitors), Database Management System (DBMS) software, text, image and video processing hardware and software, and the networking infrastructure for intra- and intersystem communication. The security and reliability requirements of these systems should never be understated. With so many factors involved, there is always room for improvement, and a universally perfect combination has yet to be implemented. Nevertheless, a wide variety of CPR systems have been successfully implemented, tested and put to everyday use, with most such systems working successfully in the environment for which they were designed. We will now explore some of the general features of CPR systems and their components, and we will present advantages and disadvantages of the choices available.

System Architecture

A fundamental decision that must be made before implementing a CPR system is the choice of the system architecture to be used. There are two broad choices, which we now discuss.

1. *Centralized Architecture*: A block schematic of a centralized architecture as implemented in CPR systems is shown in Fig. 1. A CPR system based on the centralized architecture has a central server that contains complete medical records of all patients in the system. Any modification to a record must be

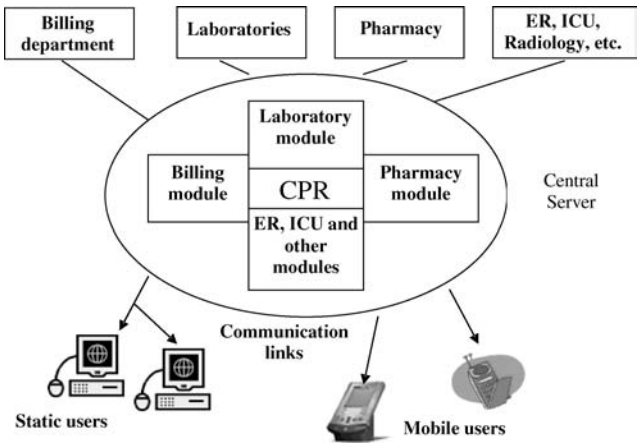


Figure 1. Centralized Architecture (Reprinted with permission from (The Computer-Based Patient Record: An Essential Technology for Health Care, Revised Edition) © (1997) by the National Academy of Sciences, courtesy of the National Academies Press, Washington, D.C.)

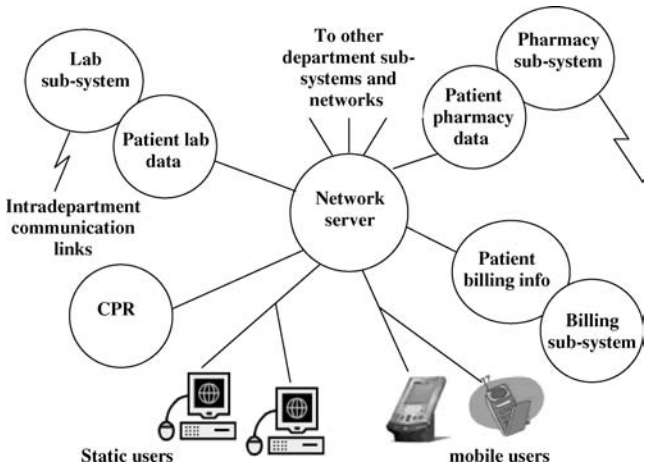


Figure 2. Distributed Architecture (Reprinted with permission from (The Computer-Based Patient Record: An Essential Technology for Health Care, Revised Edition) © (1997) by the National Academy of Sciences, courtesy of the National Academies Press, Washington, D.C.)

done to the CPR stored in this central repository to have any lasting effect. Communication links may be provided to other departments that will immediately communicate any new information to the server, thereby keeping the CPR on the central server fully updated. The principal advantage of this architecture lies in the immediate availability of the complete record for use elsewhere within the system. It is also easier to manage and to maintain the resources of the system, because they are located in one single location. However, a poorly designed system can be prone to server overloading, leading to delayed responses. Server or network failures can bring the system to a complete standstill, and if the data on the server is corrupted or destroyed, the medical record activity subsequent to system backup may be completely lost. These issues can be resolved, but the solutions can be expensive.

2. *Distributed Architecture*: The alternative to the centralized approach is the distributed architecture. A block schematic is shown in Fig. 2. In this approach, the load of the server is distributed among smaller subsystems, distributed through many departments. No department contains the complete medical record of all patients. The records are completed "on demand", that is, when a user at a system workstation must obtain complete information, the workstation sends requests to each individual subsystem, receives data back from them, and arranges them into a record. This system continues to function even if one or more of the subsystems become inoperative. However, the response time, defined as the time elapsed between the user request to fetch a record and the arrival of the complete record back to the requesting workstation, might become significant if even one of the subsystems is overloaded. The architecture also provides more opportunities for security breaches than the centralized approach. There may also be difficulties related to each subsystem using different data structures to manage and to store records, making the design of an effective DBMS potentially quite complex. These issues have been successfully addressed in specific implementations of distributed CPR systems.

DBMS and Media Processing

The general components of a CPR are shown in Fig. 3. An ideal DBMS would be able to seamlessly integrate different media formats (text, images, three-dimensional (3D) models, audio and video). Four important types of DBMS commercially available are hierarchical, relational, text- and object oriented. Each of these is better suited for one or more different media and system architectures than the others. Many modern CPR systems, especially those based on a distributed architecture, use a combination of commercial DBMSs or utilize a custom DBMS.

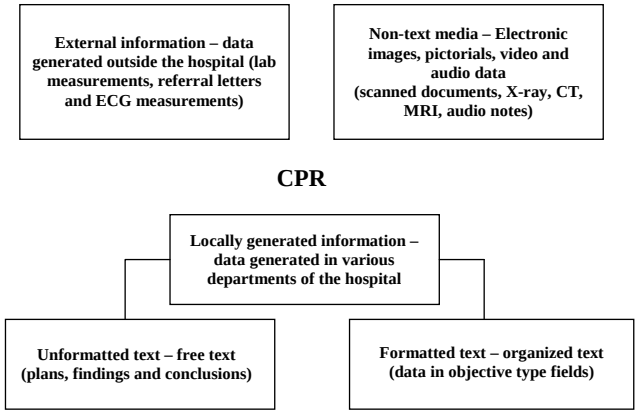


Figure 3. Components of a CPR.

An important aspect of a DBMS is the mechanism that it uses for data retrieval. The details of record construction can play a significant role in creation of databases that can be efficiently and accurately searched. Entries in formatted fields can be searched much more easily than scanned documents, images or nonformatted text. A DBMS should also provide organization of the data in a manner that is sensible and easy to comprehend. Well-organized databases will prove easier to manage and faster to search.

Security and Legal Issues

A core issue in the design of CPR systems is security. A single security breach could be used to access thousands of complete CPRs, which are then vulnerable to many malicious activities. Thus, it is extremely critical to protect individual CPRs from unauthorized access, as well as to protect the CPR system itself from hackers attempting to access the system or attempting to cause it to fail. In addition, the CPR data must be protected before it can be sent across a network, due to potential communication leaks that can land sensitive information in the wrong hands. Tools such as network packet sniffers are extremely powerful if an unauthorized party can achieve physical access to network interconnection hardware or cabling. In general, five conditions must be satisfied before patients or health care personnel are granted access to the records on a network (11):

1. The patient must not have recorded an explicit prohibition of CPR transmission across the network.
2. The patient must have consented to the transfer of the partial or complete CPR to the recipient institution.
3. The identity of the patient must be accurately verified and authenticated.
4. The identities of the healthcare provider and the recipient institution must be accurately verified.
5. It must be verified that the patient is mentally competent. If the patient is found to be mentally incompetent, then the need for the patient record must be sufficiently urgent that if it is not made available immediately, serious harm could result to the patient's health.

For a CPR produced by a CPR system to be a legally valid document, the CPR should be able to answer the following questions (12).

1. *Who Wrote It*? A single CPR is frequently handled by many different people, with several healthcare providers entering, updating, or removing information. A CPR must include the identities of all the people who have modified the record, together with the information that they have modified. This involves secure access control using smart cards, PINs, fingerprints, passwords, etc. to verify the identity of requestors before they will be allowed to access or to modify the information on the records.
2. *When Was It Written*? Time and date information is maintained by automatically attaching timestamps to the CPRs. The system must be designed to ensure that unauthorized personnel cannot change the system time or any timestamp in any CPR.
3. *What Does It Say*? Computer technology is reinvented roughly every 5 years. Data storage and retrieval media become obsolete very quickly. For example, the magnetic tapes that were very popular 25 years ago can be found only in museums today. It is thus important for CPR systems to incorporate these technological developments, and yet not become outmoded too quickly, so that the contents of current CPRs and archived CPRs can both be read and processed to obtain the same information.
4. *Has the Information Been Altered*? Undocumented alteration of information will prove to be the most important legal issue, and hence, the most difficult one to solve. A CPR system must be designed to allow determination of whether the information is indeed as originally recorded or whether it has been subsequently modified in a manner that would affect possible legal proceedings. Several effective solutions, based on cryptographic techniques, exist.

A very reliable solution to ensure that a particular CPR provides answers to all of these four questions is to attach digital signatures to them. Just as an ink-based signature provides legal acceptability for paper-based records, digital signatures can ensure the legal acceptability of a CPR. Digital signatures offer both signatory and document authentication, and they have been proven to be more effective than the ink-based signature (13). Signer authentication provides the capability to identify the person who signed the document, while document authentication offers the capability to determine whether a document was altered after it was signed.

Digital signatures use public-key encryption schemes to provide this functionality. In public-key encryption schemes (14), every user gets a pair of *keys*: a public key, which everybody is allowed to know, and a private key, which is kept as a secret known only by the individual and the authority that issues and verifies the keys. A digital signature is computed using the data bytes of the document and both the private and public keys and attached to the document. Consider that person A creates a CPR and attaches a digital signature to it using his public and private keys. Anybody with access to A's public key can verify that the CPR was actually created by A using the CPR, the digital signature and A's public key. If the result is correct, according to a straightforward mathematical relationship, A is authenticated as the signatory of the CPR, and the CPR is authenticated as accurate. Any alteration of the CPR after the calculation and attachment of the digital signature would corrupt the digital signature and would then cause the CPR to be identifiable as modified. In this case, the CPR would be considered fraudulent, although the source of the fraud would not necessarily be identifiable.

Public-key encryption can also be used to secure CPRs before transmitting them across a network. Consider that person A desires to send a CPR to person B. Person A looks up the public key of B and encrypts the CPR using B's public key. The CPR can only be decrypted using B's private key. Since only B has access to his private key, only he can decrypt the CPR. The principle of public-key encryption can be extended to CPRs to protect the authenticity of the information, to identify malicious alterations of the CPR, and to secure transmission across a network.

POINTERS TO THE FUTURE

Much of this article has been devoted to explaining the benefits of the CPRs and the factors hindering their widespread implementation. Despite their tremendous potential, the development of commercial CPR systems has not reflected the progress made in many related fields in computer technology. In this section, some of these related technologies that, in the future, the authors feel will have great potential to broaden the range of advantages of computer-based medical record services and to make them sufficiently secure are presented.

Radio Frequency Identification Tags

Radio Frequency (RF) identification (RFID) refers to an automatic ID system that uses small RF devices to identify and to track individual people, pets, livestock, and commercial products. These systems are used to automatically collect highway tolls and to control access to buildings, offices and other nonpublic sites. An RFID tagging system includes the tags themselves, a device to write information to the tags, one or more devices to read the data from the tags, and a computer system with appropriate software to collect and to process information from the tag. Many applications can use the same devices to both read and write the RFID tags. While the physical layout of the tag may vary according to the application, its basic components will include an intelligent controller chip with some memory and an antenna to transmit and receive information. According to its power requirements, an RFID tag will be classified into one of two types. Active tags have their own power source and hence provide greater range. Passive tags will be powered by RF pulses from the tag reader (or writer) and thus exhibit no shelf life issues due to battery exhaustion.

While plans are underway to replace bar codes with RFID tags on virtually every single commercial product, the field of medical informatics has found some unique applications for RFID tags. The U.S. Food and Drug Administration (FDA) recently approved the implantation of RFID tags on humans (15). These tags are similar to the tags being used on animals, and they are implanted under the skin in the triceps area. The chip contains a 16-digit number that can be readily traced back to a database containing the patient's information. One such chip made by VeriChip costs $\sim$ \$125, exclusive of the cost of implantation. This technology is expected to be a boon to individuals with life-threatening medical conditions and to lower medical costs by reducing errors in medical treatment.

Testing of replacement of bar codes with RFID tags in patient bracelets is ongoing. Unlike bar codes, RFID tags do not require clear line of sight between the bar code and the bar code reader, nor do they require active operator intervention. This ensures that healthcare workers will not fail to scan a patient ID bracelet. One such successfully tested system is Exavera's eShepherd, which combines RFID tags with Wireless Fidelity (Wi-Fi) networks and Voice over IP (VoIP) to implement a single system that will track patients, staff and hospital assets (16). The Wi-Fi routers can deliver patient information directly to a physician's handheld PDA every time any RFID transceiver detects a patient. Physicians and hospital staff can have the patient information whenever and wherever they want, and they do not have to refer repeatedly to a secure physical filing area to retrieve patient records.

A major factor impeding widespread implementation of RFID tags is the legal and ethical issue of keeping such detailed records of people, their activities, and all the things they buy and use, without their consent. However, once this issue has been resolved, RFID tags will change the way patient records are processed.

Biometrics

Biometrics refers to automatic recognition of people based on their distinctive anatomical and behavioral characteristics including facial structure, fingerprint, iris, retina, DNA, hand geometry, signature, gait and even the chemical composition of sweat (17). Biometric systems have been used in wide ranging applications like user authentication before entry into buildings and offices, criminal investigation, and identification of human remains.

Biometric systems will also find an excellent application in management of medical records. Individual traits of people who are authorized to access the record can be stored along with the record. When the system receives a request from a user to retrieve the record, the system will attempt to match the biometrics of the user to the database of biometrics of all the people who are authorized to access the record. The record is fetched only if the match is positive. In this manner, a CPR can be viewed and modified only by authorized personnel. This is potentially an effective application for biometric systems, because the system can be trained with as many samples as needed from authorized personnel. Biometric system applications where a strong need exists to verify the authenticity of claimed membership in a preidentified population have been successfully tested with almost 100% accuracy. Such systems also offer a deterrent to would-be intruders because the biometric information associated with an unsuccessful access attempt can be retained and used to identify the intruder.

Generally, the biometric information of a user is encoded on a smart card that the user must use to gain access into the system. This provides a far better authentication solution than using just a user name and a password, because both of these can be forged fairly easily. Most current biometric systems are based on the uniqueness of human fingerprints, and systems based on more complicated biometric features are currently undergoing

investigation. The development of superior biometric systems holds the key to more secure CPR authentication.

Content-Based Image Retrieval

Content-Based Image Retrieval (CBIR) is the process of retrieving images from a database based on the degree of similarity in content between an example image (the query image) and the various images contained in the database. Traditional non-CBIR data retrieval mechanisms apply text-based methods, in which a string query is matched to user-generated descriptions of documents or to contents of specific fields in a record. When patient images are integrated into patient records, simple text-based searches will not serve the same purpose. Text descriptions of images are very subjective and require data entry by a knowledgeable user. If that same image is to be added to a different database, the accompanying description also must be added if the image is to be retrievable.

These techniques utilize both global properties, such as image statistics and color distributions, and local properties, such as position and texture, to retrieve images that appear "similar" to the query image. This is a powerful tool for physicians, because CBIR enables them to compare the image of a current patient (query image) to similar images of other patients in the database, to examine descriptions attached to those images in their corresponding CPRs, and to study the treatment administered in those cases, all in a matter of seconds.

More and more frequently, scanned documents such as referral letters and scanned paper-based records are being added to CPRs. These documents can be treated as images, and CBIR systems based on character recognition technologies can search for text descriptions on these documents, making these documents compatible with text-based searches. This technology can save a great deal of time, money and errors in converting historical or ongoing paper-based records to CPRs, because the records can simply be scanned and need not be entered again into the CPR system. As more and more visual aids are added to medical diagnostics, CBIR will become indispensable to CPR management.

Steganography and Watermarking

Steganography and cryptography are related but distinct methods that are used to ensure secure transmission and secure archival of information. In cryptography, messages are encrypted in such a manner that unauthorized recipients may not decrypt the messages easily, using long, secure passwords and complex mathematical algorithms to drastically alter the data. Often, however, the encrypted messages themselves may be obtained fairly easily, because they are transmitted over insecure networks and archived on insecure servers. In steganography, the very presence of secure information is masked by hiding that information inside a much larger block of data, typically an image.

If an oncologist wishes to get a second opinion from a gynecological specialist concerning a diagnosis of cervical cancer, he might elect to send a digital image of the cervix of the patient. He would also include his diagnosis, relevant patient details including other complications the patient may have, and a referral letter, all of which are sensitive and confidential information. The traditional method to send the supplemental information electronically would be to encrypt the information using public-key encryption techniques and to transmit it as separate files together with the image. If the transmission is intercepted by a malicious party seeking private data, the transmission would garner interest, because it would be quite obvious that any encrypted information following the image information would likely be related to that image. The data thief must still work to decrypt the transmission and to reveal the confidential information. Decryption itself is difficult without extensive computational capacity, but data thieves will often find getting the necessary password to be a much easier method to access the data. The disadvantage of encryption is that the data thief knows what to try.

An alternate data transmission method would use steganography. The sensitive information can be embedded in inconspicuous areas of the image (18). Thus, no additional information or files are transmitted, and the image will raise less suspicion that associated sensitive data can be located. Although the embedded data is computationally easier to decode than the encrypted messages, it will only be decoded by people with foreknowledge that there is information embedded in the image. Furthermore, the embedded data can be encrypted before it is embedded, so that the data thief has another level of barrier. The information is much more difficult to locate, and the thief still has to decrypt or to steal the password.

Usually, the clinically important sections of the image will be identified, and only the remaining regions will be used for embedding the secret information. Several sophisticated steganography techniques, like BPCS, ABCDE, and modified ABCDE (19–21) have been successfully implemented. These methods can hide significant amounts of information without clinically altering the image, although hiding too much data leaves evidence that can be detected mathematically using tools from the field of steganalysis. Figure 4 shows an example of data hiding using the

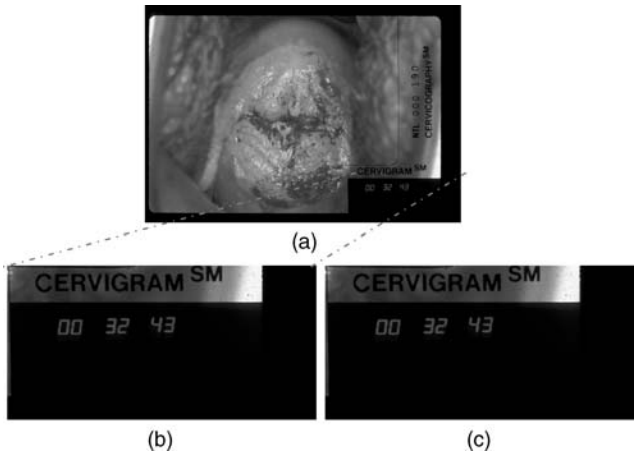


Figure 4. Steganography. (a) Original digital image of the cervix. (b) Section of clinically unimportant segment of image in a. (c) Section in b encoded with secret information.

modified ABCDE scheme. Figure 4a is a digital image of a cervix, and Fig. 4b shows the bottom right section of this image, which contains no clinically important information. This section forms an 864×432 RGB image, which requires 1,119,744 bytes to represent directly or 295,727 bytes to represent when losslessly compressed to PNG format. Figure 4c is the section in Fig. 4b after 216,326 bytes of sensitive information have been embedded. This hidden data corresponds to >40 pages of pure text, at 5000 characters per page. Figure 4c visually appears very similar to the section in Fig. 4b, even with this extensive additional data embedded. An even more secure solution would encrypt the sensitive information before embedding it inside the carrier image. This way, even if steganalysis methods detected that the carrier image contains embedded information, the decryption problem still remains.

Watermarking techniques are similar to digital signatures in the sense that they can provide owner and document authentication for digital media. They are usually nonvisibly detectable signals embedded into the media that can be checked, by authorized personnel, to verify the validity of the media and to trace its copyright holders. Watermarking and steganography techniques can also be used to identify times, dates, locations, and information that could be used to cross-reference to a patient, so that misidentification or misfiling of image data under the wrong patient file can be detected and corrected. It is the opinion of the authors that both these technologies will provide valuable solutions for secure CPR systems.

CONCLUDING REMARKS

This article attempted to provide a brief overview of the applications of computers in medical records, the benefits of computerization, and issues concerning widespread deployment of computer-based medical record systems. The computerization of medical records is a vast area of research, employing many people from diverse fields of medicine, science and engineering. While we have attempted to summarize and present the material from our own perspective, the topic itself is explored in great detail by several worthy publications. There are literally thousands of publications that attempt to explain or to solve one or more of the issues presented in this article and to provide much greater detail than is possible in this article. Thus, the authors felt it appropriate to leave the reader with a few resources, both in print and on the World Wide Web (WWW), to obtain more information about computers in medical records.

Much of the material in this article has been inspired from the extensive discussions in Refs. 3 and 22. These books provide interesting observations and refer to a wealth of references that pioneered the computer revolution in the medical field. Two more recent books, (23 and 24), also provide excellent insight into the various aspects of using computers to manage medical records. The WWW contains many websites that provide solid insight into various aspects of medical records management. Ref. 6 is one such site that is constantly updated with information about upgrading to CPR systems, cost benefit calculators, software vendor surveys, and from basic tutorials on every aspect of CPR management. Refs. 25 and 26 provide excellent evaluations of commercially available software for CPR management.

ACKNOWLEDGMENT

The authors would like to thank Mr. Rodney Long at the U.S. National Library of Medicine, Bethesda, Maryland, for providing the images of the cervix used in the example on steganography.

BIBLIOGRAPHY

1. Shortliffe EH. The evolution of electronic medical records. Acad Med 1999;74(4):414–419.
2. Burnum JF. The misinformation era: The fall of the medical record. Ann Intern Med 1989;110:482–484.
3. Dick RS, Steen EB, Detmer DE, editors. Institute of Medicine. The Computer-Based Patient Record—An Essential Technology for Health Care. Washington (DC): National Academy Press; 1997.
4. Shortliffe EH, Perreault LE, Fagan LM, Wiederhold G. Medical Informatics Computer Applications in Health Care and Biomedicine. 2nd ed. New York: Springer-Verlag; 2000.
5. Covell DG, Uman GC, Manning PR. Information needs in office practice: Are they being met? Ann Intern Med 1985; 103:596–599.
6. Voelker KG. (None). Electronic Medical Records [online]. http://www.emrupdate.com. Accessed 2005, April 10.
7. Weed LL. Medical records that guide and teach. New Engl J Med 278(11):593–600 and 278(12):652–657.
8. Brailer DJ, Terasawa EL. Use and adoption of computer-based patient records, California Healthcare Foundation report, Oct. 2003, ISBN 1-932064-54-0.
9. Wang SJ, et al. A cost-benefit analysis of electronic medical records in primary care. Am J Med Apr 1 2003;114(5):397–403.
10. Classen DC, Pestonik SL, Evans RS, Burke JP. Computerized surveillance of adverse drug events in hospital patients. J Am Med Assoc 1991;266:2847–2851.
11. David MR, et al. Maintaining the confidentiality of medical records shared over the Internet and the World Wide Web. Ann Intern Med 1992;127(2):138–141.
12. Cheong I. The legal acceptability of an electronic medical record. Aust Fam Phys Jan. 1997;26(1).
13. Askew RA. Understanding Electronic Signatures [online]. Real Legal. Available at http://www.reallegal.com/downloads/pdf/ESigAskewWhitePaper.pdf. Accessed 2005. April 10.
14. Dent AW, Mitchell CJ. User's Guide to Cryptography and Standards. Boston: Artech House; 2004.
15. Sullivan L. FDA approves RFID tags for humans. Inform Week Oct. 2004.
16. Collins J. RFID remedy for medical errors. RFID J May 2004.
17. Jain AK, et al. Biometric: A grand challenge. IEEE Conference on Pattern Recognition, Vol. 2; Aug. 2004. pp 935–26.
18. Johnson NF, Duric Z, Jajodia S. Information hiding: Steganography and Watermarking—Attacks and Countermeasures. Norwell (MA): Kluwer Academic; 2001.
19. Kawaguchi E, Eason RO. Principle and Applications of BPCS-Steganography. Proc SPIE Int Symp Voice, Video Data Communications; 1998.
20. Hirohisa H. A Data Embedding Method Using BPCS Principle With New Complexity Measures. Proc Pacific Rim Workshop on Digital Steganography. July 2002; p 30–47.

21. Srinivasan Y, et al. Secure Transmission of Medical Records using High Capacity Steganography. Proc IEEE Conf Comput.-Based Medical Systems; 2004. p 122–127.
22. Dick RS, Steen EB, editors. Institute of Medicine, The Computer-Based Patient Record—An Essential Technology for Health Care. Washington (DC): National Academy Press; 1991.
23. Hartley CP, Jones III ED. EHR Implementation: A Step-by Step Guide for the Medical Practice. Am Med Assoc Feb. 2005.
24. Carter JH. Electronic Medical Records: A Guide for Clinicians and Administrators. American College of Physicians March 2001.
25. Rehm S, Kraft S. Electronic medical records—The FPM vendor survey, Family Practice Management. Am Acad Family Phys Jan. 2001.
26. Anderson MR. EMR frontrunners, Healthcare informatics online, May 2003.

See also EQUIPMENT ACQUISITION; OFFICE AUTOMATION SYSTEMS; PICTURE ARCHIVING AND COMMUNICATION SYSTEMS; RADIOLOGY INFORMATION SYSTEMS.

MICROARRAYS

NEIL WINEGARDEN
University Health Network
Microarray Centre, Toronto
Ontario, Canada

INTRODUCTION

Microarrays allow for the simultaneous, parallel, interrogation of multiple biological analytes. Originally, microarrays were devised as a method by which gene expression could be measured in a massively parallel manner (all the genes in the genome at once), however, recent advances have demonstrated that microarrays can be used to interrogate epigenetic phenomena, promoter binding, protein expression, and protein binding among other processes. The overall process is reliant upon the manufacture of a highly ordered array of biological molecules, which are typically known entities. The features of this array behave as probes, which react with and bind to the unknown, but complimentary material present in a biological sample. Here we will focus specifically on gene expression (deoxyribonucleic acid, DNA) microarrays, which can be used to assay the activity of thousands of genes at a time.

In 1993, Affymetrix published a novel method of using light directed synthesis to build oligonucletide arrays that could be used for a variety of biological applications (1). Shortly thereafter, a group lead by Patrick Brown and Ron Davis at Stanford University demonstrated that robotically printed cDNA arrays could be used to assay gene expression (2). Now, more than a decade after this initial work was made public, both types of DNA array are commonly found in genomics laboratories.

BASIC PRINCIPLES

A DNA microarray contains a highly ordered arrangement (array) of several discrete probe molecules. Generally, the identity of these probes, be they cDNA or oligonucleotides, is either known or can be determined readily. The probes are deposited by some means (see the section Fabrication of Microarrays) onto a solid-support substrate such as glass or silicon. DNA microarrays take advantage of a basic characteristic of DNA, namely, the ability of one strand of DNA to find its complementary strand in solution and bind (hybridize) to it. This hybridization event is highly specific following standard Watson–Crick base pairing rules (Fig. 1).

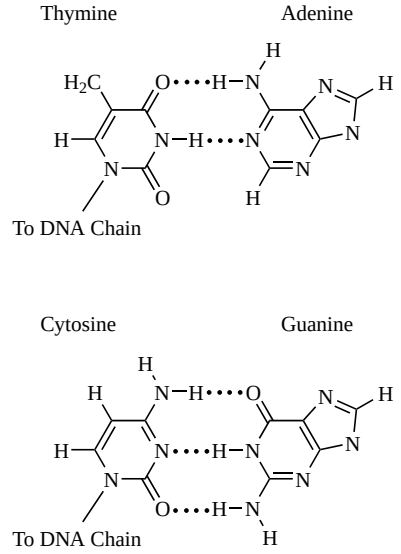


Figure 1. Watson–Crick base pairing interactions. During hybridization, specific base-paring interactions occur by which Thymine (T) binds specificly to Adenine (A) and Cytosine (C) binds specifically to Guanine (G). The binding of these bases to one another is mediated by hydrogen bonding as shown. The GC base pairs are stronger by virtue of the three hydrogen bonds formed compard to only two for AT.

Gene Expression

With some exceptions, the genetic makeup of every cell in an organism is the same. Each cell has the same complement of genes, which comprise the organism's genome. The subset of genes that are active in a particular cell dictate that cell's function. When we say a gene is active or *expressed*, we mean that particular gene is being transcribed. Transcription is the process by which ribonucleic acid (RNA) polymerase II (an enzymatic complex) reads a gene and creates a complementary copy of messenger RNA (mRNA). The more a gene is transcribed, the more copies of mRNA will be present in a cell. Thus genes that are highly active in the cell will be represented by multiple copies of mRNA, whereas genes that are inactive in the cell will have very few or no copies of mRNA in the cell. Microarrays function to measure the amount of mRNA present in the cells of a biological sample such as a tumor biopsy. The activity of the genes is inferred from this measure.

Gene Structure

In higher eukaryotes, somatic cells (diploid) have two copies of every gene: one maternally and the other

paternally derived. In the context of the diploid cell, each copy is termed an allele. In the case where both inherited alleles are the same for a given gene, that gene is said to be homozygous. If the two alleles are different, then the gene is heterozygous. Alleles may be either dominant (phenotypically manifested regardless of what the other allele is), or recessive (phenotypically manifested only in the absence of a dominant allele). In the case of a heterozygous gene, the dominant allele will be phenotypically manifested and the recessive allele will not. If both alleles are different, but dominant, they are termed codominant and both alleles will elicit a phenotype. The gene is comprised of DNA, which is double stranded. One strand is the sense strand or the strand that encodes the information, which will be ultimately represented in mRNA. The other strand is said to be anti-sense and is the strand of DNA that is actually read by the RNA polymerase to generate the mRNA. DNA has directionality: A gene is transcribed starting at the 3′ end of the antisense strand of the DNA and is read toward the 5′ end. The resultant mRNA is made from the 5′ to the 3′ end.

Genes are regulated by specific sequences of DNA that lie outside the coding region of the gene. The first such sequence is the promoter. Promoters bind the transcriptional machinery (RNA polymerase II) that performs transcription. Promoters are found 5′ (upstream) of the gene and are proximal to the transcription start site. An additional class of regulatory sequence called an enhancer may be associated with the gene. Enhancers may lie upstream, downstream, or internal (usually in noncoding regions termed introns) to the gene (3). Specific transcription factors bind enhancers and promote recruitment or activation of the basal transcriptional machinery. It is the coordinated function of the promoter and enhancer, with the transcription factors that bind them, that control if a gene is active or not within the cell. Thus, genes are regulated, and can be turned on, off, or modulated up or down by the regulatory mechanisms of the cell.

RNA Isolation

Ribonucleic acid must be isolated from cells in order to prepare the material for hybridization to the array. A cell contains three major species of RNA: mRNA, transfer RNA (tRNA), and ribosomal RNA (rRNA). Together they are refered to as total RNA. For the purpose of gene expression experiments with microarrays, the mRNA is the species we are interested in and represents $\sim 1\%$ of total RNA. In order to isolate total RNA from cells, one of two main modalities is used: solution- or solid-phase extraction. In solution-phase methods, cells are lysed in the presence of isothiocyanate in order to inactivate any RNases (naturally occurring enzymes that nonspecifically degrade RNA). The lysate is then extracted with an acidified phenol: chlorophorm:isoamyl alcohol solution. The RNA selectively partitions to the aqueous phase of this mixture away from proteins and DNA. The aqueous phase is removed and RNA is precipitated out of solution using isopropyl alcohol at high salt concentrations. Solid-phase methods make use of the variable binding activity of RNA to a silica matrix at high and low salt conditions. Cells are again lysed in the presence of isothiocyanate. The high concentration of isothiocyante used in this methodology not only inactivates the RNases, it also selectively precipitates proteins out of solution. The lysate is applied to a column containing a silica filter at the bottom. The lysate is pulled through the column via vacuum or centrifugation, thereby removing the proteins and cellular debris. In this method, DNA may also bind to the column, and as such contaminating DNA is removed by the application of DNase. The column is washed to remove any further contaminants, and then the RNA is eluted from the filter using water.

mRNA Structure

In eukaryotic cells, mRNA has a unique feature that allows researchers to either purify it away from the rest of the RNA or to direct enzymes to it specifically while avoiding the other RNA species. This feature is the polyA tail. The polyA tail is a long stretch of adenine nucleotides found at the 3′ end of mRNA, which is added post-transcriptionally. Such stretches of adenine nucleotides do not typically occur naturally in genes or other RNA species. The polyA tail will hybridize to an artificially generated oligonucleotide made up of a series of deoxythymine nucleotides (oligo-dT). If the oligo-dT is coupled to a support matrix (e.g., beads) the mRNA can be pulled out of solution thereby purifying it away from the rest of the total RNA. While some researchers prefer to include this step in their process, it is generally not a requirement for microarray analysis. Rather than purify the mRNA, the oligo-dT can be used as a primer for creating an enzymatically labeled complement of the mRNA.

Labeling

In order to render the RNA visible to a detection system, it is necessary to label it in some manner. While some laboratories choose a direct methodology of chemically labeling the mRNA itself, it is most common to work via a cDNA or cRNA intermediate that is labeled enzymatically.

The simplest methodology involves creating labeled cDNA. In this technique, the RNA is reverse-transcribed (DNA is made from an RNA template) by an enzyme named reverse transcriptase (RT) (for sample protocols, see Ref. 4). Reverse transcriptase requires a small oligonuclotide primer that binds to the RNA creating a short double-stranded region (an RNA:DNA hybrid). In order to ensure that the RT enzyme reads only the mRNA, the polyA tail of mRNA is exploited by using a primer made of a stretch of several (usually 20–25) thymine residues. The resultant DNA is the complement of the RNA and it is thus referred to as complementary DNA (cDNA). The RT reaction requires that free nucleotides (each of A, C, G, and T) are present to create the DNA. If one of these nucleotides is chemically modified with some detectable molecule (such as a fluorophore), then it will be incorporated into the cDNA strand, and that cDNA will be detectable with a fluorescent reader. Alternatively, it is possible to use a reactive molecule (such as amino-allyl) in place of a fluorescent molecule. After incorporation into the DNA, the DNA is then coupled to a reactive form of a fluorophore

(usually a reactive ester). This latter implementation of the method has an advantage in that the amino-allyl modifier is a much smaller chemical group that is incorporated much more efficiently into DNA than a bulky fluorescent moiety.

Often the amount of RNA available is limiting and cannot be detected by standard means. In this case, it is generally necessary to amplify the amount of material present. A typical microarray experiment usually requires 5–10 μg of total RNA in order to be able to obtain useful data. When researchers are working with diminishingly small samples, such as from a needle biopsy or a fine needle aspirate, it is often not possible to obtain this amount of total RNA. To overcome this limitation, various amplification strategies have been adopted. The most popular method of amplification is based on the protocols of Dr. James Eberwine from the University of Pennsylvania (5). In this technique, RNA is converted into cDNA using the same method described above with two key differences: (1) there is no labeled nucleotide incorporated and (2) the oligo-dT primer has another short sequence of DNA appended to it that represents a T7 promoter region. The T7 promoter is a bacteriophage-derived sequence that initiates transcription by T7 polymerase. After the cDNA is created, a second strand is generated creating a double-stranded artificial gene with a T7 promoter on one end. This artificial gene is then transcribed by the addition of T7 polymerase, which is allowed to make numerous transcripts of the gene. The transcripts that are obtained can either be labeled directly, or they in turn can be turned into labeled cDNA using standard methodologies described above. The resultant RNA is now actually the opposite sequence of the original mRNA, so it is said to be cRNA (complementary RNA).

The Affymetrix GeneChips utilize an amplification system based on T7 transcription as described above. During the production of cRNA, biotin modified nucleotides are incorporated. Posthybridization (see the section on Hybridization) the arrays are stained with a streptavidin bound fluorophore. Streptavidin is a protein that specifically and tightly binds to biotin molecules, allowing the fluorophore to be attached to the cRNA.

A clean-up step is required to remove any free, unbound detection molecules. This step helps to ensure that background signal is kept to a minimum. There are two main methods by which such purification is performed, one is based on standard nucleic acid purification systems, similar to the RNA isolation method described earlier, and the other is based on size exclusion. For the first method, a nucleic acid purification column is utilized. The cRNA or cDNA binds to the silica filter, but the less charged free nucleotides flow through. After a series of washes, the cRNA or cDNA is eluted from the column. The second methodology utilizes a membrane filter (usually incorporated into a column) that has a defined pore size. The large cRNA and cDNA molecules are retained on the membrane; where as the small free nucleotides flow through. The column is then inverted and the cDNA or cRNA is then eluted off the column by flowing wash buffer in the opposite direction. This purified labeled material is then ready for hybridization to the array.

Hybridization

Microarray technology relies on the natural ability of single-stranded nucleic acids to find and specifically bind complementary sequences. Purified labeled material is exposed to the spotted microarray and the pool of labeled material "self-assembles" onto the array, with each individual nucleic acid (cDNA or cRNA) species hybridizing to a specific spot on the array containing its complement. The specificity of this interaction needs to be controlled, as there may be several similar and related sequences present on the array. The control of hybridization specificity is accomplished through the adjustment of the hybridization stringency. Highly stringent conditions promote exact matches where as low stringency will allow some related, but nonexact matches to occur. In a microarray experiment, stringency is typically controlled by two factors: the concentration of salt in the hybridization solution and the temperature at which hybridization is allowed to occur.

High salt concentrations tend to lead to lower stringency of hybridization. Both strands of nucleic acid involved in the hybridization event contain a net negative charge. As such, there is a small repulsion between these two strands, which needs to be overcome to bring the labeled nucleic acid into proximity of the arrayed probe. The salt ions cluster around the nucleic acid strands creating a mask and shielding the electrostatic forces. Higher salt concentrations have a greater masking effect, thus allowing hybridization to occur more easily. If salt concentrations are high enough, the repulsion effects are completely masked and even strands of DNA that have low degrees of homology may bind to one another.

Temperature is another important factor. Every double-stranded nucleotide has a specific temperature at which the two strands will "melt" or separate. The temperature at which exactly 50% of a population of pure double-stranded material separates is termed the melting temperature (T_m). The T_m of a nucleic acid is controlled partially by the length of the strand and partially by the percentage of G and C residues (termed the GC content). The G and C residues bind to one another as a Watson–Crick base pair. This pairing interaction is the result of three hydrogen bonds forming. The other potential base pair in a DNA hybrid, A:T, only has two such hydrogen bonds and thus the greater the GC content of the nucleotide, the more stable the hybrid. At very low temperatures, nonstandard Watson–Crick base pair interactions can also occur causing noncomplementary sequences or sequences that are $<100\%$ matched to form hybrids. It is necessary therefore to find a temperature that will prevent or melt nonspecific hybrids, but allow the specific interactions to occur. For a microarray, this presents a challenge as there are thousands of specific interactions that must be accommodated. In the case of oligonucleotide arrays, the design of the oligonucleotides to be spotted takes this issue into account and probes are designed that tend to fall within a narrow window of potential melting temperatures. cDNA arrays are more difficult because the sequences spotted vary greatly in both GC content and length. In such cases, it is often true that conditions that represent somewhat of a "compromise" are necessary.

Hybridization kinetics can generally be modeled as shown in Eq. 1(6). The change in the amount of hybridization product LS over time is a function of the decrease in the concentration of labeled target L and free spotted DNA S over time. To simplify the equation, the rate of hybridization is equal to some rate constant k multiplied by the product of the concentrations of L and S. Thus hybridization rate is a direct function of the concentrations of the labeled target molecule and the DNA probe in the spot.

$$\frac{d[\mathrm{LS}]}{d\mathrm{T}} = -\frac{d[\mathrm{L}]}{d\mathrm{T}} - \frac{d[\mathrm{S}]}{dT} = \frac{d[\mathrm{L}-\mathrm{S}]}{dT} = k\,[\mathrm{L}][\mathrm{S}] \tag{1}$$

In the case of an oligonucleotide microarray, it is often the case that the number of spotted DNA molecules is in great excess to the number of target molecules. As such, the concentration of the spotted DNA probe remains fairly constant and can be considered part of the constant k. Thus the equation for hybridization can be simplified as shown in Eq. 2 (6), where the rate of hybridization is typically driven by the concentration of the labeled target molecules alone.

$$\frac{d[\mathrm{LS}]}{d\mathrm{T}} = k'[\mathrm{L}] \tag{2}$$

In the case of two color oligonucleotide arrays, the two labeled samples compete for hybridization to the probe that remains in excess and thus hybridization is simply a reflection of the concentrations of each of the two labeled targets L_1 and L_2 [Eq. 3(6)].

$$\frac{d[\mathrm{L_1S}]}{d[\mathrm{L_2S}]} = \frac{k_1'[\mathrm{L_1}]}{k_2'[\mathrm{L_2}]} \tag{3}$$

The situation becomes somewhat more complex when the probe molecules are not in excess of the target molecules. This is often the case with cDNA arrays. In these cases, the concentration of the spotted probe does change significantly as hybridization occurs and thus each of the labeled targets L_1 and L_2 hybridize in a manner described by Eqs. 4 and 5 (7).

$$\frac{d[\mathrm{L_1S}]}{d\mathrm{T}} = k_1[\mathrm{S}][\mathrm{L_1}] = k_1([\mathrm{S^0}] - [\mathrm{L_1S}] - [\mathrm{L_2S}])([\mathrm{L_1^0}] - [\mathrm{L_1S}]) \tag{4}$$

$$\frac{d[\mathrm{L_2S}]}{d\mathrm{T}} = k_2[\mathrm{S}][\mathrm{L_2}] = k_2([\mathrm{S^0}] - [\mathrm{L_2S}] - [\mathrm{L_1S}])([\mathrm{L_2^0}] - [\mathrm{L_2S}]) \tag{5}$$

In such a case, the rate of hybridization is affected by the change in the concentrations of the spotted probe from the initial concentration S^0, where S^0 changes as the probe molecules are bound by either L_1 and L_2.

When looking at differential hybridization between the two targets, we can represent the kinetics as shown in Eq. 6 (7).

$$\frac{d[\mathrm{L_1S}]}{d[\mathrm{L_2S}]} = \frac{k_1([\mathrm{L_1^0}] - [\mathrm{L_1}S])}{k_2([\mathrm{L_2^0}] - [\mathrm{L_2}S])} \tag{6}$$

If one is to assume that the two fluorescent molecules used in a two-color experiment behave similarly, and that the rate of hybridization of the two labeled targets is the same, we can say $k_1 = k_2$. It has been demonstrated that under ideal conditions and when the hybridization reaction is allowed to continue to equilibrium that the ratio of the concentrations of each possible hybrid L_1S and L_2S is equivalent to the ratio of the original concentrations of the two targets L_1 and L_2 [Eq. 7 (7)]. This point is important because it is the basis for microarrays to work, assuming that the ratios read from the scans during data analysis are reflective of an actual biological condition.

$$\frac{[\mathrm{L_1S}]}{[\mathrm{L_2S}]} = \frac{[\mathrm{L_1^0}]}{[\mathrm{L_2^0}]} \tag{7}$$

The goal of microarray hybridization is to produce a result for which the signal obtained from specific hybridization is very strong when compared to any background signal that may be obtained by a nonspecific adsorption of labeled material to the substrate, or nonspecific binding to spotted elements. To reach this goal, it is common to use certain nonspecific blocking reagents in the hybridization solution. Frequently, nucleic acids from sources known not to contain any sequences that will interfere with specific hybridization are used. For example, in a hybridization of a human sample to an array, one might use yeast tRNA and salmon sperm RNA as competitors to bind any regions of the substrate or probes that have a generic nucleic acid binding capacity. These nucleic acids are nonlabeled and will therefore not contribute any signal when the array is scanned.

Washing

Unlike traditional northern blots, the majority of the stringency of a microarray assay is accomplished at the hybridization step. The washing step of a microarray experiment is a critical operation, but is important more as a means to remove unbound material in order to reduce background signal than it is to control the specificity of the signal obtained.

Wash buffers generally contain two components: a salt solution and a detergent. The salt solution, frequently sodium chloride sodium citrate (SSC), is set to a concentration that supports the maintenance of the hybridized molecules. This concentration most frequently falls in the $1\times$ to $2\times$ concentration range with some labs using as low of a concentration as $0.1\times$ ($1\times$ SSC contains 0.15 M NaCl and 0.015 M Na-citrate).

The detergents used in wash buffers help to remove the unbound fluorescent molecules that would normally stick to the surface of the slide. The detergent acts as a surfactant and helps to isolate and remove the unbound fluorescent material. Typically, an anionic detergent such as sodium dodecyl sulfate (SDS) is used for this purpose.

The temperature for the washes varies depending on the stringency of the wash solution being used. As with hybridization, the combination of temperature and salt concentration determines the overall stringency of the washes.

After washing the microarrays, it is generally necessary to perform a rinse. The rinse is typically a solution similar to the wash solutions without the detergent. If detergent remains on the slide after drying, the solution may fluoresce particularly if the labeled material has been trapped in detergent micelles.

Scanning

It is necessary to use an imaging device to detect the fluorescent labels present on the hybridized microarray. In general, the imaging device must contain an excitation light source, an emission filter, and a light gathering device.

During scanning, the labeled material, be it fluorescent or some other form of detectable molecule, is imaged and the resultant data is converted to a digital image. The optimal resolution at which the image is scanned is dependent on the size of the features and on their interspot spacing. A general rule of thumb is that the resolution of the image should be such that the pixels represent one-tenth of the diameter of the spot. For spotted arrays, for example, the features tend to be on the order of 100 μm in diameter and thus 10 μm resolution is frequently used. Affymetrix's technology, however, can generate features that are 11 μm square; in this case, a much higher resolution of down to 1 μm is required.

Most commonly, the image that is generated is a 16-bit grayscale TIFF (Tagged Image File Format) image (Fig. 2). The 16-bit depth of the image provides a total of 65,536 gray levels providing a possibility of more than five orders of magnitude range. The TIFF format is important because it is a universally accepted format that is LOSSLESS; that is, even with compression, this format retains all image information. The images can then be imported into the appropriate image quantification software.

Image Quantification

After scanning, it is necessary to extract data from the images. Image quantification generally starts with segmentation. Segmentation is the process by which pixels that represent the signal are isolated from those that represent background. During segmentation, the discrete areas of the image that represent the spotted DNA material are identified and digitally isolated from the remainder of the image. The intensities of all of the pixels in the individual spot are averaged to determine the overall spot intensity. This spot intensity is proportional to the amount of material hybridized to that region, with higher intensities resulting from increased numbers of hybridized molecules. Each spot, for each channel (in the case of two color microarrays) is quantified, and the resultant data are tabulated. Other data may also be extracted at this stage. It is common to also obtain intensity data for the area outside of the individual spots. This value represents the background of the image and indicates the amount of signal that would have been obtained regardless of a specific hybridization event. It is common, however, not universal, to subtract the background values from the signal intensities of the spots.

There are several means by which segmentation can be carried out. In the most basic setup, a fixed shape (usually a circle) is placed over each spot. The entire complement of pixels lying within the circle is used to determine the average intensity. Pixels lying outside of one of these circles are deemed to be background signal. More advanced segmentation algorithms attempt to account for the fact that most of the spotted features on a microarray are in fact not perfectly uniform. Spots may deviate from a true circular shape, or may have regions within the circle in which DNA was not attached (creating a spot that is reminiscent of a doughnut). In addition, it is not uncommon for each of the spots to have some degree of variance in their diameter. The more advanced methods utilize various algorithms and statistics to determine which pixels actually represent signal and which are more representative of background.

Image quantitation software then processes the entire image and produces a table of results that represents the signal, and the background for each feature on the array. These packages may also export various other data, which can be used in quality control analysis such as standard deviations, coefficients of variance, circularity, or uniformity of the spot, and so on. This data table can then be processed as part of the data analysis.

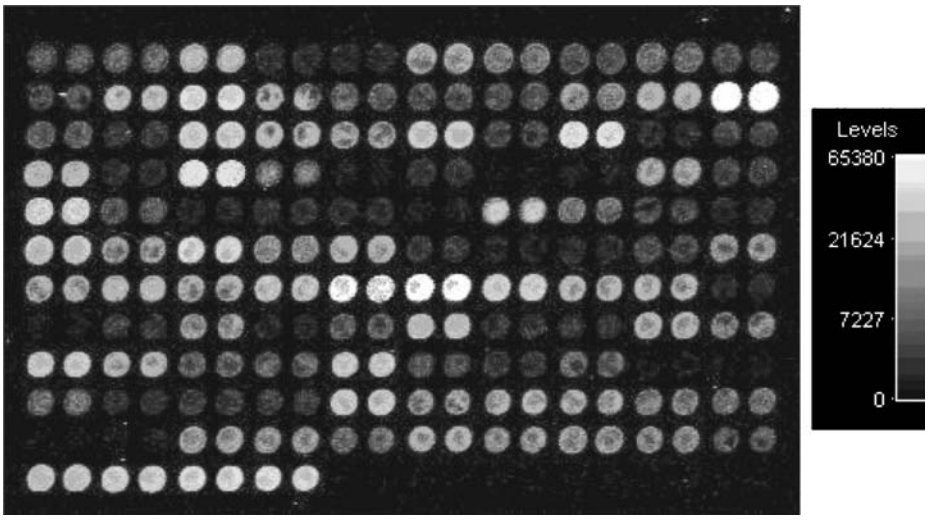


Figure 2. Arrays imaged on a microarray scanner are presented as 16-bit grayscale TIFF images. The picture shown represents a small subsection of a larger array. Each spot is 100 μm in diameter and the spot-to-spot spacing is 200 μm in this image. The image was scanned at 10-μm resolution.

Data Analysis

An exhaustive description of the process of DNA microarray data analysis is far beyond the scope of this article (for an excellent review see Ref. 8). The exact process followed depends greatly on the experimental design and the question being addressed. There are, however, some basic principles that tend to be fairly common in dealing with microarray data: statistical analysis of data, supervised and/or nonsupervised data mining, data visualization, and validation are all key components.

Statistical analysis of microarray data comes into play in two main areas. The first is to determine which spots are reliable and provide sufficient data. Spots that have a high degree of variance across replicates, for example, are likely not able to provide reliable data. These hypervariable genes or signals need to be filtered from the data so as to not skew the results of data mining. Statistics may also play a role in supervised data analysis.

There are two major categories of data mining: supervised and nonsupervised. Supervised data mining utilizes algorithms in which the user imparts restrictions on how the data is grouped. For example, in an experiment where a cohort of patients was tested in which one group was healthy and the other group was afflicted with a particular disease, one would indicate to the algorithm which arrays were from the healthy patients and which were from the patients with disease. The algorithm then tests the data to find genes that are markers for the diseases. Specifically, each gene is tested to see if the expression levels for that gene are statistically significantly different in each of the two patient groups. The goal is to find a series of genes that can act as markers that are diagnostic of the disease.

In nonsupervised clustering, the algorithm is not given any indication as to how the individual samples are related. In true nonsupervised clustering, the algorithm is not even told how many groups exist. The data are analyzed and the samples are grouped based on similarity metrics. The classical methods of nonsupervised clustering include hierarchical clustering and principal components analysis (PCA). The algorithms generally display the data via some visualization pattern such as the canonical "plaid" expression patterns seen from hierarchical clustering. The researcher then overlays the grouping information onto the patterns provided to see if the individual groups naturally separate from one another. In other cases, this methodology may be being used to determine how many groups there truly are, as the researcher may not have this information a priori. In such cases, the groups can then be further examined to see if there are differences in treatment response, survival, or any other characteristic desired. Generally, after this technique is performed one will attempt to look for clusters of genes in the patterns that distinguish between the different groups and again use these genes as markers.

Regardless of the methodology utilized, it is extremely important to validate the data. Cross-validation strategies are various, but in their most basic form, one obtains a cohort of patients to profile. A subset of this cohort is used to look for potential markers. Once the markers have been identified, the remaining patients are tested and only the identified markers are used to try and group the patients. If the markers are able to stratify the patients into their appropriate groups, then the markers are considered to be viable and may provide beneficial diagnostic ability. On occasion, however, the validation set is not properly grouped. In such cases, the markers are only useful for the narrow set of patients used in the initial tests and more testing is required to find a viable set of markers.

FABRICATION OF MICROARRAYS

There are two main methodologies for manufacturing microarrays, which differ in the means by which the probe material spotted onto the arrays is prepared. In one methodology, the DNA to be spotted is generated *in situ* using either standard or modified phosphoramidite chemistry. (Phospohoramidites are reactive forms of each of the nucleotides that make up DNA. Phosphoramidite chemistry is a well-defined process by which moderate length stretches of DNA can be created with any specific sequence.) This method is used by Affymetrix and Agilent, the two largest commercial suppliers of microarrays, although both groups use a different approach to the *in situ* synthesis.

Other groups use *ex situ* synthesis, whereby the DNA material is either prepared as PCR products (cDNA) or oligonucleotides manufactured using standard phosphoramidite synthesis. Once this material is prepared it is spotted onto the array substrate using either contact or noncontact printing methodologies. Amersham (now GE Healthcare) and Applied Biosystems use this methodology to make microarrays as do almost all of the "homebrew" laboratories that make microarrays in house.

Fabrication of DNA Arrays *In Situ*

There are two main approaches to the generation of microarrays by *in situ* synthesis of DNA: photolithography and inkjetting. Affymetrix, the industry leader uses a proprietary photolithography process to mask off areas of the array, protecting some areas, and leaving others available for the DNA synthesis reaction to occur (1). This is a multistep process requiring several masks per array to be made. Each synthesis reaction is performed sequentially. For each nucleotide position, there are four possible masks (one for each of A, G, C, and T). Thus, an array comprised of 25-mer oligonucleotides would require $\sim$100 masks to complete the process (typically $\sim$70 are required for an array due to the sequences used). Affymetrix uses a modified phosphoramidite chemistry for synthesis of the oligonucleotide chains; whereas standard phospohoramidite chemistry uses acid labile protection groups, the Affymetrix technology utilizes groups that can be removed by ultraviolet (UV) light. The Affymetrix technology allows for extremely high density arrays of hundreds of thousands of features to be prepared on very small substrates of <1 cm^2.

Other groups have developed technologies that allow them to get around the need for multiple masks to be made for each array design. The pioneer in this area was Nimblegen, who uses digital light processor (DLP)

micromirrors to create the masks (9). Each of these DLP units (used typically in AV projectors and large screen televisions) comprises thousands of tiny (10 μm^2) micromirrors. The micromirrors can be individually addressed and the angle of the mirrors changed to allow light to pass through. In the "open state", the micromirror directs light onto the surface of the microarray, allowing DNA synthesis to occur. In the "closed state", the micromirror reflects light away from the surface, disallowing DNA synthesis. A computer controls the mirrors and thus each DLP unit has a near infinite number of combinations that can each be controlled, and as such, a single unit can create any pattern desired on the array. Nimblegen uses the same chemistry as Affymetrix, using light activated deprotection of the phosphoramidites. A somewhat newer entry into this area is Xeotron (now part of Invitrogen). Xeotron also uses micromirror DLPs to address the masks, however, they have also incorporated small microfluidic channels on their chips. Each feature is placed in a microscopic well on the chip. Rather than using the modified phosphoramidite chemistry of Affymetrix and Nimblegen, Xeotron uses standard chemistry, but has instead employed a caged acid that can be freed by light (10,11). As such, the acid that controls deprotection of the nascent oligonucleotide can be directed to specific locations by light. The Nimblegen and Xeotron technologies have the advantage of being highly amenable to custom array generation, however the Affymetrix technology is particularly well suited to mass production of a standard array. Each of these approaches has found customers in the marketplace.

A third approach to *in situ* synthesis of the oligonucleotides involves ink-jet spotting. Agilent uses this technology (developed by Rosetta Inpharmatics) in which each of the reactive phosphoramidites (A, G, C, and T) are loaded in to a separate "ink-cartridge" to allow for control of which nucleotide is added to each spot during the synthesis stage (12,13). This methodology eliminates the need for masks, but does require very high precision robotics as the print head must return to the same spot many times, within micron accuracy, during the course of synthesis. This technology draws from the strength of each of the others mentioned in that it is relatively easy to customize the design of arrays, and yet, mass production of arrays is possible using a large robotic system.

Fabrication of DNA Arrays *Ex Situ*

Some of the commercial vendors and nearly all of the "homebrew" microarray centers utilize and approach of spotting DNA that was prepared *ex situ*. In the case of cDNA arrays, the spotted material is prepared by polymerase chain reaction (PCR), whereas oligonucleotide arrays are generated using oligos created via high throughput oligo synthesis. The DNA material is purified and placed into a specific spotting buffer that is compatible with the substrates being used.

The DNA is typically aliquoted out into multiwell plates (96, 384, or 1536 wells /plate) to facilitate transfer by the arraying robot. The buffer that the DNA is placed in has several functions. First, the buffer stabilizes the DNA to prevent it from degradation. Second, the buffer must provide an appropriate surface tension to ensure that the spots that are placed on the substrate are of a controllable size and uniform in shape. Of similar importance, however, is that the buffer must provide conditions that are compatible with the attachment chemistry that is going to be utilized.

The DNA may either be coupled to the slide through rather simple electrostatic interactions or via a specific coupling reaction. Electrostatic interactions are mediated by using a uniform positively charged substrate that attracts the negatively charged DNA. Often the substrates used are silylated to provide reactive amine groups on the surface. Alternatively, one may coat the slides with a chemical such as poly-L-lysine, which simply adsorbs onto the substrate and provides a net positive charge. This type of interaction is mass based. As such, there is a maximum mass of DNA that can bind to any one spot on the substrate. Longer DNAs will be represented by fewer copies than shorter DNAs. To overcome this, it is possible to use more specific interactions by using modifiers on the DNA that will react with certain groups on the slide. The two most common such modalities involve aldehyde or epoxide chemistry. In this method, the DNA is modified with a primary amine group. The substrate has reactive aldehydes or epoxides that will react specifically with the primary amine to form a covalent bond (Fig. 3). This type of interaction is molarity based, and as such, with the exception of steric effects, the number of DNAs that bind per spot is relatively equivalent regardless of length.

EQUIPMENT

The manufacture of microarrays, and their subsequent use requires some very specialized equipment. Generally, a facility that produces microarrays will require some advanced robotics for fabrication. A laboratory that uses arrays will require scanning devices to read the arrays. Due to the relatively high costs of these pieces of equipment it is common for many people to rely on core facilities for some or all of the process.

Arraying Robots

Ex situ prepared DNAs are spotted onto the microarray substrates via robotics (Fig. 4). Robotics are required to accurately position the printing devices over the slides to create the arrays. The majority of systems utilize pins and direct contact to deposit the DNA material. In this system, a printhead with several spotting pins in a defined arrangement is used to dip into the multiwell plates and pick up the material to be spotted. The typical operation sequence of an arrayer robot may include:

1. Dipping the printing applicators (pins) into a source plate to pick up DNA samples. Each applicator picks up a separate DNA sample from an individual well in the plate. Typically 32–48 pins are used at one time.
2. Movement to a blot-station to preprint from the pins. This step removes excess solution from the pins to ensure that the spots that are printed onto the arrays

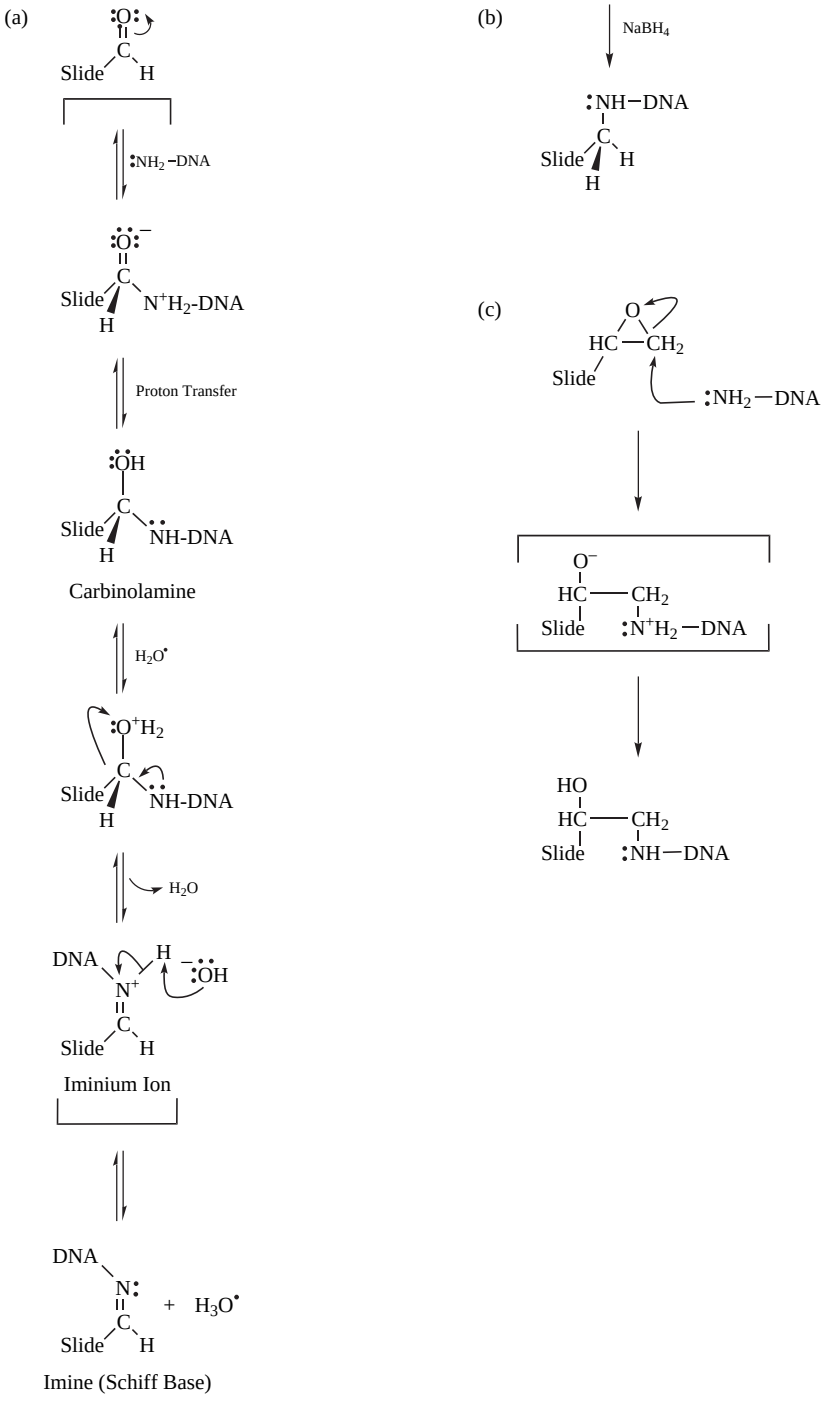


Figure 3. Covalent attachment of amino-modified DNAs to aledhyde (a) or epoxide (b) slides is possible. An amino-modified DNA reacts with an aldehyde surface by a Schiff's base reaction. The resultant Schiff base must be reduced with an agent such as sodium borohydride ($NaBH_4$) to prevent reversal of the reaction.

are uniform in size and do not run into one another causing contamination.

3. Movement to the slide platform. The print head then moves over the slide platform taking position over the first slide.
4. Printing onto the arrays. The print head moves down bringing the pins in contact with the slide. The DNA solution held in the pins by capillary action is spotted onto the slide. The printhead then moves to the next slide position and again spots onto the slide. This process is repeated until all of the slides on the platform have been printed.
5. Washing the pins. The print head then moves the pins to a wash station. Although there are many configurations possible, the basic principle is to use water or some other solution to remove the excess liquid from the pins and then to dry the pins (under vacuum or stream of air). This process may be repeated several times to make sure there is no carryover.

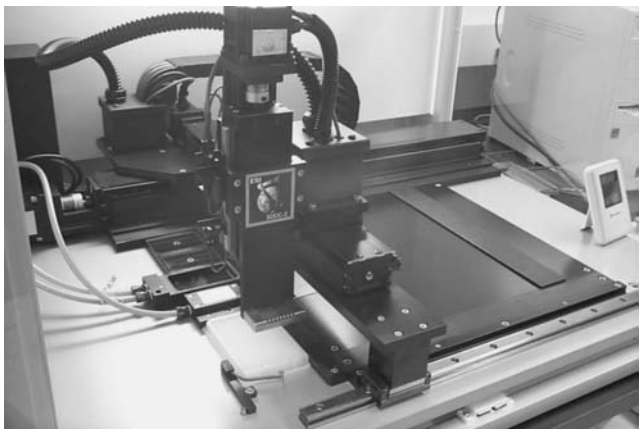


Figure 4. A microarraying robot. The robotic arrayer prints DNA onto glass slides with very high precision. Robots such as this have extremely high accuracy, on the order of 10 μm or less.

6. Loading the next sample. The print head returns to the source plate to pick up the next set of samples.

In a typical high throughput system, such as those offered by Bio-Rad, BioRobotics, GeneMachines, Genetix, and Telechem International, 48 pins are used at one time. The entire operation sequence described above may take 3–4 min to complete for 100 arrays. Often arrays may contain 20,000–40,000 spots. As such, a typical print run may require 600 or more cycles through the operation sequence, which can take as long as 30 h or more to complete.

Hybridization and Fluidics Stations

Certain array platforms require that a specific hybridization and/or fluidics station be utilized. In the case of spotted arrays (home-brew in particular), this is usually an option and often a case of personal preference. In these cases, a hybridization station may be utilized to improve mixing of the hybridization solution over the array. The rate of diffusion of a labeled nucleic acid in solution is actually very low, and as such, some researchers prefer to use an automated station that performs mixing of the solution.

In the case of Affymetrix GeneChip technology, a specific hybridization and fluidics station are required. The hybridization station is simply a rotating incubator in which the chips are placed. A bubble that is introduced into the sealed array cartridge moves around during rotation creating a mixing effect. The fluidics station is a more advanced system that is required to introduce the various labeling components and wash solutions required. This station allows the user to keep the cartridge sealed without having to attempt to pipette solutions in and out.

Scanners

While some microarray imagers such as the Perkin Elmer ScanArray and GeneFocus DNAScope are confocal scanners, this is not a strict requirement. Confocal imaging serves to eliminate extraneous signals, but reduces the light gathering ability of the device. There are >10,000 commercial microarray scanners in the field capable of reading standard glass microarrays. The leading scanner makers include Agilent, Axon, Bio-Rad, GeneFocus, PerkinElmer, and other vendors. The laser scanner uses one or more lasers with wavelengths appropriate to the fluorophores being used. The most commonly used fluorophores for microarrays are cyanine 3 and cyanine 5 (or fluors with equivalent spectra). Cyanine 3 has an absorbance maximum of 550 nm and emission maximum of 570 nm. There are 2 main lasers used in scanners to excite this fluorphore: "Gre-Ne" (green neon) gas lasers and Nd:YAG (neodymium doped yttrium aluminum garnet) frequency doubled solid-state diode lasers. Cyanine 5 has an absorbance maximum of 650 nm and an emission maximum of 670 nm. There are two main lasers used in scanners to excite this fluorophore: standard He–Ne gas lasers and red diode lasers. Table 1 shows some of the characteristics of these two dyes, along with two other popular dyes, Alexa 555 and Alexa 647, which have spectra that are very similar to those of Cy3 and Cy5 respectively (Fig. 5).

Cyanine 3 and 5 have some important features that make these dyes particularly suitable for use in microarray analysis. The spectra of these dyes have little over lap and can generally be separated from one another with little to no cross-talk. In addition, these fluors have a somewhat unique property in that they are brighter when dry than when wet. Most fluorophores have the opposite behavior, which is impractical for microarrays because the scanners generally cannot handle wet preparations.

The other major class of microarray imager is a CCD (charge coupled device) based system. In general, these imagers use a white light source to excite the fluorophores. The fluorescent light that is emitted is captured by the CCD and converted into a digital image. Rather than scanning the slide, a CCD based imager tiles together several sections of the slide to create an image of the entire surface. This tiling can create a stitching effect whereby the "seams" of the images may not be completely smooth.

Table 1. Key Characteristics of the Most Commonly Used Fluorophores for Microarray Analysis

Fluorophore	Excitation Max, nm	Emission Max, nm	Molar Extinction Coefficient	Molecular Weight
Cy3	550	570	150,000	766
Cy5	649	670	250,000	792
Alexa555	555	565	150,000	1,250
Alexa647	650	668	239,000	1,250
Phycoerytherin	566	575	19,600,000	240,000

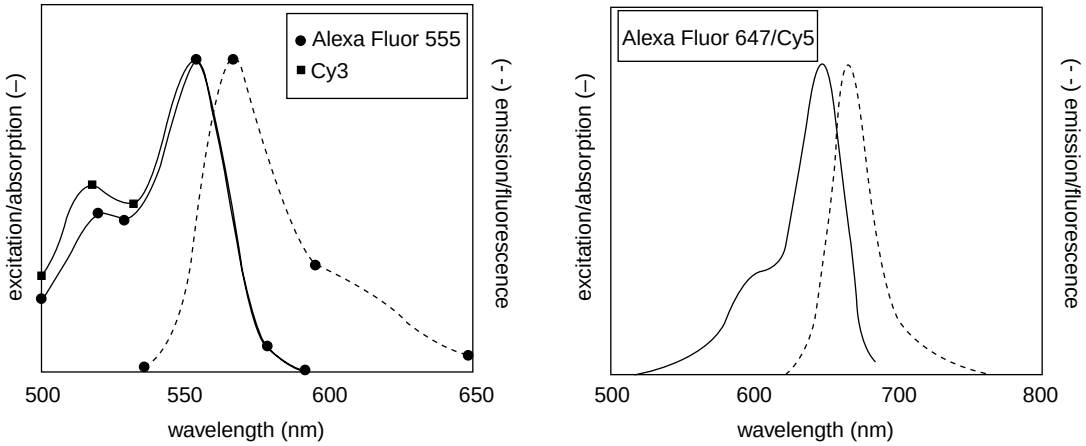


Figure 5. Representative spectra of the fluors commonly used in spotted microarray experiments. Alexa Fluor 555 and Cy3 are excited by green wavelengths of light whereas Alexa Fluor 647 and Cy5 are excited by red wavelengths of light. One green excited and one red excited fluor may be used at the same time as there is little overlap in their excitation spectra.

This problem can be overcome with advanced lighting systems and software.

Affymetrix arrays use a different labeling chemistry for detection relying on the naturally occurring fluorescent protein phycoerytherin. Phycoerythrin is a naturally occurring pigment protein from light harvesting algae that absorbs strongly at 566 nm and has an emission peak at 575 nm. It is a very bright fluorophore having a molar extinction coefficient that is 80 times as high as the standard Cy3 and Cy5 molecules. The limitation of this molecule is that it is also 200 times larger, making the number of molecules that can be incorporated per sequence much less. As such, this molecule can only be applied to the DNA posthybridization for fear that it would create steric interference.

MICROARRAYS AS MEDICAL DEVICES

To date, microarrays have mostly found use in basic research applications, and have yet to make a strong impact on the diagnostic market. [During the preparation of this text, Roche received FDA clearance for the first ever array based diagnostic chip. The AmpliChip CYP450 based on the Affymetrix platform was approved in January of 2005 (see http://www.roche.com/med-cor-2005-01-12).] Microarrays have indeed been used to study many diseases including various cancers, cardiovascular disease, inflammatory disease, psychiatric disorders and infectious disease. This basic research will ultimately lead to the identification of potential therapeutic markers for drugs of for diagnostics. The potential of microarrays extends beyond target discovery, however, and will eventually impact on the way that medical care is performed.

Target Discovery

The use of microarrays in basic research laboratories has often focused on target discovery. In these applications, microarrays are used to profile a particular disease where disease tissues are compared to healthy tissues either from the same patient or from a separate test population. In such experiments, the goal is to find genes that are differentially regulated (either up or down) in the disease state compared to a healthy tissue. Such genes are thought to be involved in the disease state or in the cellular response to the disease. As such, these genes are potential diagnostic markers and may also represent drug targets.

Drug/Lead Discovery

Microarrays can also be used once the target has been identified. It is possible to use microarrays to screen potential therapeutic compounds, for example, to determine which candidates reverse the pattern of gene expression that is indicative of disease. Microarrays have been even more effective in looking at toxicity of lead compounds. One of the leading contributors to failure of a pharmaceutical compound is toxic or off target events. Microarrays have proven useful in screening for the up-regulation in toxicity related genes. In addition, it is possible to determine if the compound creates other effects that while not toxic *per se* could cause undesirable side effects from nonspecific interactions. Often toxicity models are tested in model organisms such as rats or dogs. Several toxicity specific arrays have been developed that allow for profiling of genes in these model systems rather than human cells.

Diagnostics and Prognostics

One of the more promising areas for microarrays to have direct impact as a medical device is in the area of diagnostics and prognostics. As mentioned under target discovery, basic research has often strived to look for a panel of genes that can be used as a molecular fingerprint of a disease. There are numerous publications in which researchers have attempted to use molecular profiles to correlate to patient outcome, disease state, tumor type, or any of several other factors. DNA

microarrays are particularly well suited to this type of analysis. Many complex diseases are multifactoral; rather than a single prognostic or diagnostic marker being present, it may be necessary to look at several genes at one time. Microarrays allow for identification of a panel of genes, which when looked at together may provide diagnostic or prognostic power. Although it has not become common practice yet, there are examples of microarrays being used to prescreen patients on the basis of a molecular profile (14).

Other attempts are being made at using microarrays to study infectious disease. Often times a patient may present with a set of symptoms that could be indicative of several different infectious agents. It is possible to prepare a microarray that would identify the agent as well as to subtype the bacterium or virus on the basis of pathogenicity. This particular application may prove very useful in identifying not only the infectious agent, but also the best course of treatment.

Pharmacogenomics and Theranostics

A concept that is gaining in popularity is pharmacogenomics or theranostics (15). Both of these terms refer to the idea of tailoring a patient's treatment or therapy on the basis of their genetic makeup. Many pharmaceuticals on the market have not known any potentially serious side effects in a subset of patients. In addition, there are typically at least some patients that are nonresponders to a particular treatment. These effects are often times the result of the patient's genetic make-up. Most of the work in this area has focused on genotyping: looking at certain variable regions of DNA and determining which variants are present in people who have negative reactions or in people who respond well to a treatment. It is hoped that in the near future it will be possible to screen a patient and determine which of a panel of drugs will be most beneficial. Perhaps even more important, it will be possible to prevent serious negative outcomes by avoiding treatment of a patient that will have a poor reaction to a drug. Theranostics also involves monitoring a patient through a course of treatment. It is possible that a patient can be screened during treatment to ensure that the therapy is working as expected. If a change occurs, the physician would be able to alter the therapy to ensure that the disease is treated in the most effective way possible.

SUMMARY

Microarrays provide a means to screen hundreds to thousands of biological analytes in parallel. These analytes can be DNA, RNA, or protein. DNA microarrays allow for rapid profiling of gene expression. While there are a few competing platforms that can be utilised, the basic principles are the same: RNA from a biological sample is extracted, labeled and applied to an array of DNA probes. Signals generated from the array indicate which genes are active and which are not. The ability to screen multiple tissues or patients make microarrays particularly well suited to uncovering the complex gene networks involved in disease. While typically used in basic research applications for target or marker discovery, the future will most likely see microarrays used in diagnostic applications and for tailoring medical treatment.

BIBLIOGRAPHY

1. Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL. Multiplexed biochemical assays with biological chips. Nature (London) 1993;364:555–556.
2. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270:467–470.
3. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. Cell 2004;116: 499–509.
4. Hegde P, Qi R, Abernathy K, Gay C, Dharap S, Gaspard R, Hughes JE, Snesrud E, Lee N, Quackenbush J. A concise guide to cdna microarray analysis. Biotechniques 2000;29: 548–550, 552–544, 556 passim.
5. Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH. Amplified rna synthesized from limited quantities of heterogeneous CDNA. Proc Natl Acad Sci USA 1990;87:1663–1667.
6. Schena M. Microarray analysis. Hoboken: John Wiley & Sons; 2003.
7. Wang Y, Wang X, Guo SW, Ghosh S. Conditions to ensure competitive hybridization in two-color microarray: A theoretical and experimental analysis. Biotechniques 2002;32: 1342–1346.
8. Quackenbush J. Computational analysis of microarray data. Nature Rev Genet 2001;2:418.
9. Singh-Gasson S, Green RD, Yue Y, Nelson C, Blattner F, Sussman MR, Cerrina F. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. Nat Biotechnol 1999;17:974–978.
10. Gao X, LeProust E, Zhang H, Srivannavit O, Gulari E, Yu P, Nishiguchi C, Xiang Q, Zhou X. A flexible light-directed DNA chip synthesis gated by deprotection using solution photogenerated acids. Nucleic Acids Res 2001;29:4744–4750.
11. LeProust E, Pellois JP, Yu P, Zhang H, Gao X, Srivannavit O, Gulari E, Zhou X. Digital light-directed synthesis. A microarray platform that permits rapid reaction optimization on a combinatorial basis. J Comb Chem 2000;2:349–354.
12. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley PS. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol 2001;19:342–347.
13. Hughes TR, Shoemaker DD. DNA microarrays for expression profiling. Curr Opin Chem Biol 2001;5:21–25.
14. Schubert CM. Microarray to be used as routine clinical screen. Nat Med 2003;9:9.
15. Picard FJ, Bergeron MG. Rapid molecular theranostics in infectious diseases. Drug Discov Today 2002;7:1092–1101.

See also DNA SEQUENCE; MICROBIOREACTORS; POLYMERASE CHAIN REACTION.

MICROBIAL DETECTION SYSTEMS

PATRICIA L. DOLAN
JASON C. HARPER
SUSAN M. BROZIK
Sandia National Laboratories
Albuquerque, New Mexico

INTRODUCTION

Infectious diseases accounted for 25–33% of the estimated 54 million deaths worldwide in 1988 (1), of which more than half are attributed to tuberculosis, malaria, chronic hepatitis B, diarrheal diseases, and human immunodeficiency virus/Acquired Immune Deficiency Syndrome HIV/AIDS. The incidence of more than 30 diseases identified since the mid-1970s continues to grow, which include HIV/AIDS, liver disease due to hepatitis C virus, cholera, tick-transmitted Lyme disease, foodborne illness caused by *E. coli* O157:H7 and *Cyclospora*, waterborne disease due to *Cryptosporidium*, and the hantavirus pulmonary syndrome. Additionally, the first known cases of human influenza caused by the avian influenza virus, H5N1, were identified in Hong Kong in 1997 (2).

Although death due to infectious diseases in the United States remains low relative to that of noninfectious diseases, their occurrence is increasing. In 2000, the Federation of American Scientists reported that infectious-disease-related death rates nearly doubled from 1980 to 170,000 annually (1). Many of these diseases, most recently the West Nile virus, were introduced from outside the U.S. borders by international travelers, immigrants, returning U.S. military personnel, or imported animals and foodstuffs. Still, the most dangerous infectious microbes reside within U.S. borders. Four million Americans are chronic carriers of the hepatitis C virus, a significant cause of liver cancer and cirrhosis. It is predicted that the death rate due to hepatitis C virus infection may surpass that of HIV/AIDS in the next five years. Influenza viruses are responsible for approximately 30,000 deaths annually. In addition, hospital-acquired infections are surging due to highly virulent and resistant pathogens such as *Staphylococcus aureus*.

The burden of identifying and treating infected individuals and controlling disease outbreaks generally lies with physicians, hospitals, and first responders. Table 1 contains important characteristics for several of the more common pathogenic microorganisms. As evidenced by this noninclusive table, a wide variety of microorganisms exists from which the specific diseasecausing microbe must be identified. In addition, the number of cells or particles that can provide an infectious dose is often extremely low. For example, the infectious dose of *E. coli* O157:H7 is as low as 10 cells (3), which poses a significant challenge to healthcare professionals and first responders who must quickly identify the infectious agent. Antimicrobial treatments that attempt to neutralize all possible infectious pathogens are often not possible or safe. Depending on the nature and severity of the infection, a delay of only a few hours in providing the proper therapy may lead to death.

Medical Microbiology

Medical microbiology is the discipline of science devoted to identifying microbial agents that are responsible for infectious disease and elucidating the mechanism of interaction between the microorganism and human host. Historically, microbiologists have used plating, microscopy, cell culture, and susceptibility tests to identify and study microorganisms. In the hospital and clinical diagnostic laboratory, these, methods are still widely used and will be briefly discussed in this article. The general procedure for isolation and identification of infectious and parasitic microbes is (1) specimen collection and streaking onto culture plates for production of isolated bacteria colonies, (2) staining and microscopic analysis, (3) cell culture in various media, and (4) antibiotic susceptibility testing.

Plating. Plating entails the streaking of a specimen onto a solid nutrient media-filled Petri dish and incubation at 35–37 °C. Under these conditions, a single bacterium divides and eventually produces a colony that is visible to the eye (Fig. 1). A visible colony generally contains more than 10^7 organisms. The colony morphology, color, time required for growth, appropriate media, and other growth conditions are used to characterize the microbe. The incubation time required for growth of a colony from a single cell is dependent on the growth rate of the microorganism. Fast-growing organisms such as *E. coli*, with a doubling time of 30 minutes, would require approximately 13 hours to produce a colony of 10^7 organisms. More typically, microorganisms require several days to a week to generate visible colonies. The plated specimen is often a complex solution such as blood, urine, feces, or sputum containing diverse native flora in addition to the infectious microbe. Plating serves as a method to isolate the infectious microbe as each

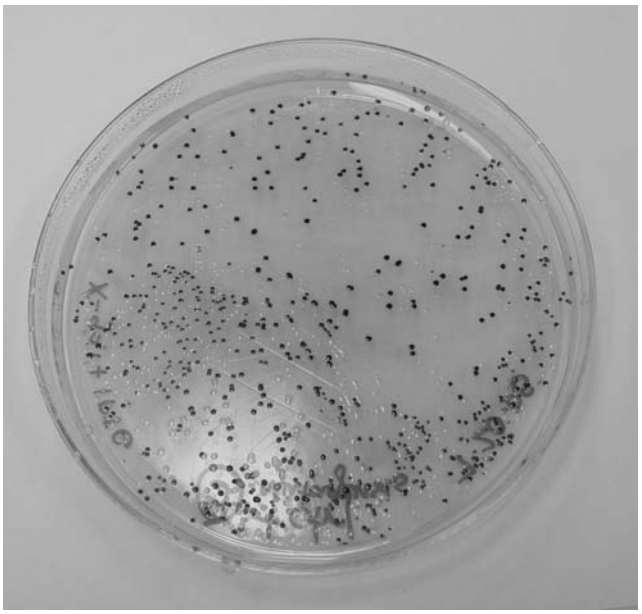


Figure 1. Colonies of *E. coli* grown on LB +ampicillin and X-gal +IPTG agar media in a Petri dish.

Table 1. Characteristics of common pathogenic organisms

Organism	Site(s)	Disease(s)	Incubation period	Mode of transmission	Infectious Dose
Bacteria, Gram positive					
Corynebacterium diphtheriae	upper respiratory track, skin	diphtheria	2–7 days	direct contact, droplet spread	unknown
Streptococcus pneumoniae	lower respiratory tract	pneumonia, meningitis	1–3 days	direct contact, droplet spread	unknown
Listeria monocytogenes	monocytes, leukocytes, blood, intestine, skin	listeriosis, meningoencephalitis	variable; 3–70 days	ingestion, direct contact, neonatal contact (birth), transplacental	unknown, <1000 cells
Bacillus anthracis	lower respiratory tract, skin lesions, intestine	anthrax	few hours −7 days	direct contact, airborne, ingestion	8,000–50,000 spores
Staphylococcus aureus	skin, osteomyelitis, blood, heart	boils, furuncles, abscesses, impetigo, osteomyelitis, sepsis, toxic shock syndrome	variable; 4–10 days	direct contact, ingestion, autoinfection, neonatal contact (birth)	$10^2 - 10^6$ cells
Streptococcus pyogenes	throat, skin, blood, middle ear	pharyngitis, septicemia, erysipelas, rheumatic fever, scarlet fever, otitis media, foodborne illness	1–3 days	direct contact, droplet spread, injection	<1000 cells
Bacteria, Gram negative					
Bordetella pertussis	oropharynx	whooping cough	6–20 days	direct contact, droplet spread, airborne	unknown
Escherichia coli(O157:H7)	large intestine	hemorrhagic colitis	2–8 days	ingestion	~10 cells
Legionella spp.	lower respiratory tract	Legionellosis, Pontiac fever	2–10 days	airborne (non–communicable)	unknown
Neisseria gonorrhoeae	genitourinary tract, eye	gonorrhoea, pelvic inflammatory disease, septicemia, pharyngitis	2–7 days	direct contact (usually sexual), neonatal contact (birth)	unknown
Neisseria meningitidis	meninges	meningitis	2–10 days	direct contact, droplet spread	unknown
Salmonella spp.	small intestine	gastroenteritis	6–72 hours	ingestion	10–100,000 cells
Salmonella typhi	small intestine	typhoid fever	1–3 weeks	ingestion	1,000–100,000 cells
Shigella spp.	intestine	shigellosis (enteritis); Bacillary dysentery	1–7 days	ingestion	10–200 cells
Yersinia spp.	intestine	enterocolitis, acute mesenteric lymphadenitis	3–7 days	ingestion	10^6 cells
Vibrio spp.	intestine	cholera, enteritis	4–96 hours	ingestion	~10^6 cells
Anaerobes, Gram positive					
Actinomyces spp.	jaw, thorax, abdomen	chronic abcesses, draining sinuses	variable; days or months	direct contact (mouth), airborne, fomites	unknown
Clostridium botulinum	nerves, muscle, skin, intestine (infants)	botulism, acute bilateral cranial nerve impairment	12–36 hours	direct contact, ingestion (non–communicable)	unknown
Clostridium difficile	intestine	pseudomembranous colitis	6–24 hours	environmental	unknown
Clostridium perfringens	skin lesion, large intestine	gas gangrene, food poisoning	1–4 days	ingestion	10^5–10^8 cells
Clostridium tetani	nerves, muscle	tetanus	3–21 days	direct contact (non–communicable)	potent toxin

Table 1. (*Continued*)

Organism	Site(s)	Disease(s)	Incubation period	Mode of transmission	Infectious Dose
Anaerobes, Gram negative					
Bacteriodes spp.	mouth, respiratory track, large intestine	peritonitis, endometritis, abscesses, septicemia	unknown	endogenous	unknown
Bacteria, acid–fast					
Mycobacterium tuberculosis	lower respiratory tract, laryngeal, meningeal	Tuberculosis	4–12 weeks	direct contact, droplet spread, airborne	10 cells
Mycobacterium avium complex	lower respiratory tract, lymph nodes	pulmonary, lymphadenitis	unknown	ingestion, skin lesions (non–communicable)	10^4–10^7 cells
Yeasts					
Candida albicans	mucous membranes, skin	oral thrush, intertrigo, vulvovaginitis, paronychia	variable; 2–5 days in infants	direct contact (sexual), neonatal contact (birth)	unknown
Cryptococcus neoformans	meninges, lower respiratory tract	meningitis, pneumonia	unknown	airborne	unknown
Molds					
Aspergillus spp.	lower respiratory tract	aspergillosis	variable; days to weeks	airborne (non–communicable)	unknown
Blastomyces dermatitidis	lower respiratory tract, skin	blastomycosis	few weeks to months	airborne (non–communicable)	unknown
Coccidioides immitis	lower respiratory tract, skin	coccidioidomycosis	1–4 weeks	airborne (non–communicable)	unknown
Histoplasma capsulatum	lower respiratory tract, skin	histoplasmosis	3–17 days	airborne (non–communicable)	10 spores
Viruses					
Acquired immunodeficiency syndrome	progressive damage to immune and other organ systems	HIV/AIDS	6 months −7+ years	direct contact (sexual, contact with blood or bodily fluids)	unknown
Hepatitis A	liver	hepatitis, type A	10–50 days	ingestion	10–10^3 virus particles
Hepatitis B	liver	hepatitis, type B	24–180 days	direct and indirect contact (bodily fluids), fomites	unknown
Hepatitis C	liver	hepatitis, type C	6–10 weeks	direct contact (contaminated blood)	10^2–10^3 particles/mL blood
Herpes simplex I	skin	herpes (vesicular lesions)	7–10 days	direct contact (saliva)	unknown
Herpes simplex II	skin	herpes (genital)	2–12 days	direct contact (usually sexual)	unknown
Influenza	upper respiratory tract	flu	1–4 days	direct contact, droplet spread, airborne	2–800 virus particles
Measles	skin	rubeola	8–13 days	direct contact, droplet spread	~10 virus particles
Rubella	skin	German measles	12–23 days	direct contact, droplet spread	10–60 virus particles
Varicella–Zoster	skin	chicken pox, shingles	13–17 days	direct contact, droplet spread, airborne	unknown

[†]Infectious Disease Information. (2005, April 29). Infectious disease information, NCID, CDC. [Online]. Centers for Disease Control and Prevention. http://www.cdc.gov/ncidod/diseases/index.htm [2005, August 21]; The "Bad Bug Book." (2003, January 30). FDA/CFSAN Bad Bug Book: Introduction to Foodborne Pathogenic Microorganisms and Natural Toxins. [Online]. U.S. Food and Drug Administration Center for Food Safety and Administration. www.cfsan.fda.gov/~mow/intro.html [2005, August 21].

Infectious Agents MSDS Index. (2003, July 31). Index to Material Safety Data Sheets (MSDS) for Infectious Substances. [Online]. Public Health Agency of Canada. www.phac-aspc.gc.ca/msds-ftss/index.html [2005, August 31].

colony originates from a single cell and is therefore pure of any other cell types. A colony can subsequently be used as a pure sample for microscopy, cell culture, and other analytical tests. Additionally, as only viable cells divide, plating can differentiate between dead microbes and those that are viable and may be the source of infection.

Staining. Microscopic observation of microorganisms is generally preceded by staining a specimen on a microscope slide. The microbe response to various stains (gram-positive/negative, acid-fast, etc.), size, grouping (single, double, chains), and morphology (bacillus, coccus, spirillum, pleomorphic) provide characteristic information helpful in identifying the microorganism. However, several pathogenic species appear similar, or are indistinguishable, under the microscope. For effective observation under a microscope, at least 10^5 cells per milliliter of sample should be present. A colony specimen usually meets this requirement, and preconcentration is generally necessary for viewing a nonplated specimen. Still, very small cells can be difficult to observe and may be overlooked. Finally, microscopic observation usually cannot distinguish between dead and live cells.

Cell Culture. Cell culture is used to ascertain the biochemical properties of a microorganism. A single colony is inoculated into a liquid media broth and incubated. Incubation is usually performed near 37 °C with agitation via shaking or gas sparging to facilitate gas transport into the liquid media for uptake by the microbes. Signs of microbial growth in liquid media include turbidity and gas formation. Turbidity can be used as a simple and nondestructive method to measure cell growth. An optical density measurement provides the degree of light scattering at a particular wavelength through a given path length of liquid media. Increasing cell density due to growth usually increases the degree of light scattered. The measured optical density can, therefore, be directly related to the total cell mass. A calibration curve for each bacterial species is required as various sizes and shapes of different microbes scatter light to varying extents.

Microbial growth in several different media is used to determine a specific microbe's biochemical and physiological characteristics. Definitive identification can require 20 or more media tests. Such tests often use selective media. A media can be made selective by addition of chemicals that inhibit microbe and native flora growth while allowing growth of a specific organism. For example, Thayer–Martin medium selectively isolates pathogenic *Neisseria gonorrhea* and *Neisseria meningitides* (4). The medium contains vancomycin to inhibit growth of gram-positive bacteria, anisomycin to inhibit fungi growth, colistin to inhibit most gramnegative bacilli growth, and trimethoprim-sulfamethoxazole to inhibit *Proteus* growth. The *Neisseria* species are resistant to these inhibitors at the concentrations present in the medium and grow freely.

Antibiotic Susceptibility Testing. Upon isolation and identification of the infectious microbe, antibiotic susceptibility testing can be performed to identify antimicrobial agents that inhibit growth. Additionally, the minimal inhibitory concentration (MIC) is determined by exposing bacteria in media broth to various concentrations of an antimicrobial agent. The lowest antibiotic concentration that inhibits growth is the MIC. A concentration of the antibiotic in the blood at or above the MIC should successfully treat an infection.

Development

Plating, microscopy, cell culture, and susceptibility testing techniques for identifying and treating infectious microorganisms have proven effective against a plethora of pathogens, hence its continued use today. However, these clinical microbiology methods have changed very little over the past century, often require days to obtain confirmed results, and cannot be used successfully to characterize several significant infectious agents including the hepatitis virus. However, with the recent and significant advances in molecular biotechnology, two additional microbe identification methods have found wide use, immunoassay and polymerase chain reaction.

Developed in 1959, the utility of the immunoassay was not fully realized by the medical diagnostic community until the late 1970s and early 1980s. The immunoassay takes advantage of an immune system reaction, the highly specific and strong binding of antibody to antigen. Antibodies are developed that specifically bind a given microorganism, chemical byproducts or proteins produced by a given microorganism, or antibodies produced by the host in response to infection caused by a given microorganism. The developed antibodies are tagged with a reporter molecule. Reporters can be radioisotopes, chemiluminescent or fluorescent molecules, or enzymes (i.e., alkaline phosphatase, horseradish peroxidase) that can produce a radiographic, colorimetric, or fluorescent signal. In the presence of the antigen, the antibody will bind and will remain bound through washes that remove unbound antibody. Detection of the reporter after washes indicates that the antigen was present, as bound antibody was not removed during washing. Although rapid, highly specific, and sensitive, immunoassays cannot differentiate between viable and dead cells and are limited to tests for which antibodies can be developed. They also can be affected by contaminants in the test specimen and do not provide quantitative information regarding the number of pathogenic agents present.

Serological assays are the most commonly used immunoassay in the medical laboratory and by the Centers for Disease Control and Prevention (CDC). The mechanism of Serodiagnosis entails binding of lab-developed antibodies to antibodies produced in the host in response to a specific infection. This indirect method of detecting infectious agents allows identification of microbes that are currently difficult or impossible to isolate and culture. For example, because HIV-1 virus requires advanced containment facilities and is difficult to isolate and culture, it is serologically diagnosed via detection of antibodies produced by the host against the virus. Additionally, a method to effectively isolate and culture hepatitis virus has not yet been devised. Therefore, diagnosis of hepatitis virus infection is done serologically. A lag phase of several weeks often exists

between onset of infection and production of antibodies by the host against the microbial agent and, thus, possibly leads to false negatives. False positives are also a concern as antibodies produced by the host during a previous infection may be present and detected.

Immunoassays were the primary diagnostic method used for microbial detection by the CDC until the development of polymerase chain reaction (PCR) (5). PCR is a technique that specifically amplifies DNA sequences (for more information, see page 8). This technology has transformed molecular biology and genetics and has changed diagnostic approaches to the identification, detection, and characterization of infectious agents. With PCR, extremely small quantities of DNA from a microorganism can be amplified and detected. Detection of amplified DNA can occur through gel electrophoresis or via genetic probes. Based upon the highly specific binding between complimentary nucleobases of DNA and RNA, genetic probes are nucleic acid sequences that bind to DNA or RNA unique to a given microorganism. Genetic probes are marked with radioisotopes, chemiluminescent or fluorescent molecules, or enzymes and will produce a quantitative signal only when the complimentary microorganism DNA or RNA is present in the sample (for more information, see page 9). Ou et al. (6) used PCR to amplify and detect HIV sequences from seropositive individuals. Subsequently, PCR amplification and sequence analysis of HIV amplicons (amplified DNA sequences) became the first use of comparative nucleic acid sequence information in a disease outbreak setting (7). Although PCR is very sensitive and sequence analysis provides specific identification capability, these technologies are expensive, time-consuming, labor-intensive, and require expertise in molecular biology. Consequently, use of PCR and genetic probes for identification of microbes is common in research laboratories and academic institutions, but, to date, is not widely used in hospital or medical diagnostic laboratories.

To address rising national and worldwide public health needs, it is desirable that a sensitive, specific, fast, and simple-to-operate device be employed to detect infectious agents. Microbial detection systems that attempt to meet this need have been commercially available since the late 1970s and have progressed significantly with the molecular biotechnology revolution. Still, microbial detection systems face three major challenges: time, sensitivity, and specificity of analysis. Microbial testing and detection must be rapid to allow adequate time for treating the infection and be highly sensitive as a single pathogenic organism may be infectious. Additionally, as a low number of pathogenic microbes may be present in complex biological samples, such as blood or urine, high specificity remains an essential requirement. To tackle these problems, alternative nucleic acid-based approaches have been integrated into user-friendly microbial detection systems that are commercially available for diagnostic purposes.

Contemporary microbial detection systems or biosensors typically consist of a selective biorecognition molecule connected to a transducer that converts a biochemical interaction into a measurable signal. Recognition molecules include nucleic acids, antibodies, and peptides. Commonly used transducers include electrochemical, optical, and piezoelectric. The following sections will discuss numerous commercially available microbial detection systems used in clinical and field settings, including (1) nucleic acid-based, optical technologies and systems; (2) fiber-optic, waveguide-based fluoroimmunoassay systems; (3) a chip- and nanoparticle-based bio-barcode optical technology; (4) an electronic microchip-based technology; and (5) an electronic nose microbial detector.

NUCLEIC ACID-BASED OPTICAL TECHNOLOGIES

Line Immunoprobe Assay (LIPA)

The line immunoprobe assay (LIPA) is a nucleic acid recombinant immunoblotting assay (RIBA) (i.e., oligonucleotides that differentiate different genetic variants are transferred onto a nitrocellulose membrane in a straight line) (8,9). PCR is performed from the clinical sample using primers that selectively amplify a DNA region containing nucleotide differences. The amplicons are hybridized with the immobilized oligonucleotides on the membrane, and an enzyme-based colorimetric method is used to detect binding and positive reactivity. The nucleotide differences contained within the amplified sample DNA provide a unique signature that differentiates target genotypes or mutant microorganisms. These assays were among the first commercially available assays using nucleic acid hybridization for diagnostic purposes.

COBAS AMPLICOR Analyzer

A second commercially available system using PCR technology is the COBAS AMPLICOR Analyzer (Roche Diagnostics; Rotkreuz, Switzerland). This system automates amplification and detection of target DNA from infectious agents by combining five instruments into one: a thermal cycler, automatic pipettor, incubator, washer, and reader. Amplified biotinylated products are captured on oligonucleotide-coated magnetic microparticles and detected colorimetrically with use of an avidin-horseradish peroxidase (HRP) conjugate. The system can detect a broad range of agents including *Bacillus anthracis, Chlamydia trachomatis, Neiserria gonorrhea, Mycobacterium tuberculosis*, cytomegalovirus, hepatitis B and hepatitis C viruses, and HIV in clinical specimens including serum, urine, and sputum. The manufacturer reports that more than 4000 COBAS AMPLICOR Analyzers are currently used in clinical settings worldwide (10).

Real-Time PCR (RT-PCR)

A number of commercially available systems for the diagnosis of infectious diseases make use of a third nucleic acid-based approach (i.e., RT-PCR). As the name implies, RT-PCR, pioneered by Applied Biosystems (Foster City, CA) in the mid-1990s (11), amplifies and measures agent-specific DNA as the reaction proceeds in real-time. It is used to quantify the amount of agent-specific input DNA or cDNA by correlating the amount of DNA with the time it takes to detect a fluorescent signal. This technology uses fluorescent reporter probes (i.e., molecular beacons) that are

detected and quantitated at each cycle of the PCR. Molecular beacons are single-stranded, dual-labeled fluorogenic DNA or RNA probes that form a stem loop structure. The loop hybridizes to the target nucleic acid, whereas the stem is end-labeled with a fluorophore at the 5′-end adjacent to a quencher at the 3′-end. Fluorescence resonance energy transfer (FRET) is the process by which energy from an excited fluorophore (donor) is transferred to the adjacent fluorophore (acceptor) at close proximity, resulting in the quenching of fluorescence. Hybridization of the target sequence to the loop separates fluorophore and quencher, and the fluorescence is measured.

The GeneXpert System (Cepheid; Sunnyvale, CA) fully automates and integrates sample preparation with the RT-PCR detection processes. It uses microfluidics technology integrated into disposable assay cartridges. The cartridges contain all the specific reagents required to detect disease organisms such as *Bacillus anthracis, Chlamydia trachomatis*, or foodborne pathogens. The system provides quantitative results from unprocessed clinical samples in 30 minutes or less and is capable is self-cleaning and decontamination before its next use. The GeneXpert module forms the core of the Biohazard Detection System deployed nationwide by the United States Postal Service for anthrax testing in mail-sorting facilities (12). It is also used in hospital laboratories, physician offices, and public health clinics.

Idaho Technology (Salt Lake City, UT) manufactures an automated, field-ready RT-PCR instrument, the R.A.P.I.D. (ruggedized advanced pathogen identification device) system, which is based on the Light Cycler Instrument from Roche Diagnostics (Alameda, CA). The R.A.P.I.D. is developed for military field hospitals and first responders in harsh field environments. Amplification of DNA in real-time can be performed on environmental and blood samples. Idaho Technology claims a 15 minute set-up time and a 20 minute PCR run for a total of 35 minutes using the R.A.P.I.D. Pathogen Test Kit. This instrument is reported to be very sensitive (i.e., *Pseudomonas aeruginosa* was detected in blood culture samples at 10 cfu (colony-forming units)/ml (13). (A colony-forming unit is a single viable cell that forms a colony of identical cells when plated.) This technology is well-established and has been in use worldwide since 1998.

The R.A.P.I.D. technology was recently put to the test at the Prince Sultan Air Base in Saudi Arabia (14). Medical personnel observed a clustering of diarrhea cases and thought them to be due to influenza. However, testing of patient samples with the R.A.P.I.D. identified the cause to be foodborne *Salmonella* within hours of sample submission. Due to the prompt response by medical and services personnel, the outbreak was limited to less than 3% of the base population.

Nucleic Acid Sequence-Based Amplification (NASBA)

A fourth approach for nucleic acid-based detection of infectious organisms is nucleic acid sequence-based amplification (NASBA), a bioMérieux, Inc. (Marcy-l'Etoile, France) proprietary isothermal amplification technology. This method is based on specific amplification of RNA by the simultaneous activity of three RNA-specific enzymes, AMV-reverse transcriptase, T7 RNA polymerase, and RNase H, generating single-stranded RNA as the endproduct (15). NASBA is an isothermal amplification procedure, carried out at 41°C.

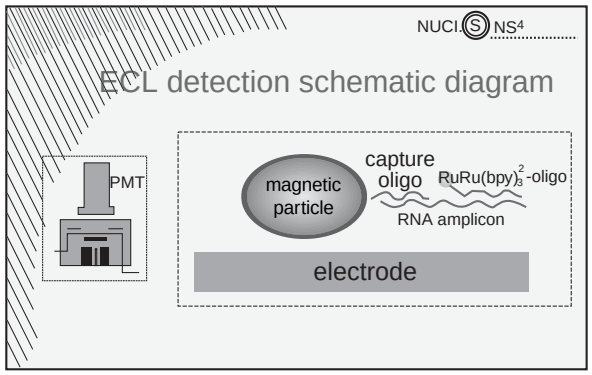


Figure 2. NucliSens Reader ECL detection scheme (16).

Two commercially available systems, NucliSens Reader and NucliSens EasyQ Analyzer (bioMérieux, Inc.), make use of the NASBA technology. Although both systems use NASBA for selective amplification of RNA, as well as a bioMérieux proprietary Boom silica-based nucleic acid extraction method, the NucliSens Reader relies on an electrochemiluminescence (ECL) detection technology, and the NucliSens EasyQ uses fluorescent detection by incorporating specific molecular beacons to which amplicons hybridize. With this method, amplification and detection occur simultaneously in a single tube.

The ECL-based NucliSens Reader employs a sandwich hybridization method for the detection of amplified target RNA (Fig. 2) (16). Two target-specific DNA probes are used: a capture probe bound to magnetic beads and a detection probe labeled with tris (2,2′-bipyridine) ruthenium (Ru). Each of these probes bind to a different region of the target RNA. After the hybridized sample is drawn into the ECL flow cell and the beads are magnetically immobilized on the electrode, a voltage is applied, and the resulting emitted light is detected by a photomultiplier tube (PMT). According to the manufacturer, measurement of 50 reactions takes approximately 50 minutes.

Real-time NASBA and fluorescent detection of target-bound molecular beacons are accomplished by the NucliSens Basic Kit and EasyQ Analyzer (Fig. 3) (17). This technique is most often used to detect RNA viruses. Using a multiplexed NASBA technique to detect four human immunodeficiency virus type 1 (HIV-1) subtypes, DeBaar et al. (18) reported an 89% correct subtype identification relative to sequence analysis and a sensitivity of 92%. The limit of detection was approximately 10^3 copies of HIV-1 RNA per reaction. Lanciotti and Kerst (19) conducted a study comparing TaqMan RT-PCR (Applied Biosystems) and standard reverse-transcription PCR (Roche Molecular Biochemicals) assays with NucliSens NASBA assays

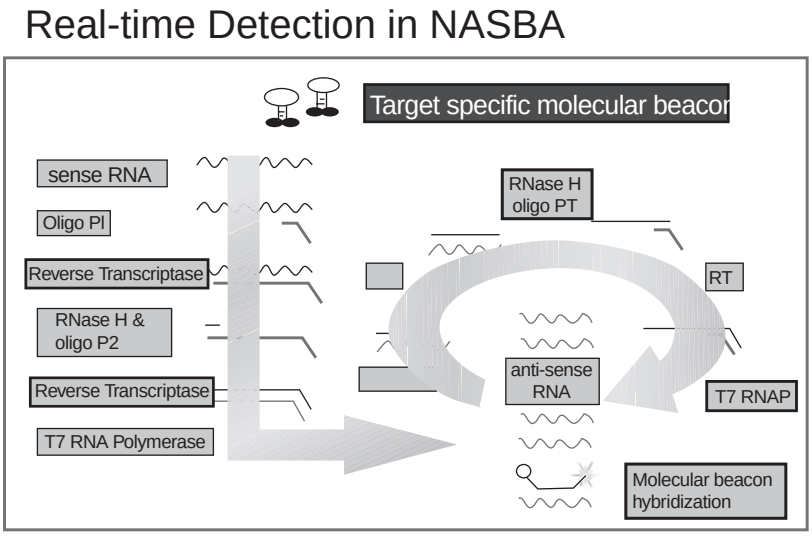


Figure 3. NucliSens EasyQ detection scheme (17).

for detecting West Nile (WN) virus and St. Louis encephalitis virus. The ECL-based and molecular beacon-based NASBA assays demonstrated equal or greater sensitivities and specificities than reverse-transcription PCR in human cerebral spinal fluid. The NASBA-ECL assay for WN virus was 10-fold more sensitive than either the TaqMan or NASBA-molecular beacon assay, detecting 0.01 pfu of WN virus. Moreover, the NASBA molecular beacon-based assay performed significantly faster than either PCR procedures (i.e., a positive signal was detected within 14–45 minutes).

Strand Displacement Amplification (SDA)

A fifth approach for nucleic acid-based detection of infectious organisms is strand displacement amplification (SDA). SDA, first reported by Walker et al. in 1992 (20), is an isothermal process that amplifies DNA or RNA using a restriction enzyme and a DNA polymerase plus several primers, without requiring temperature cycling. Available since 1999, the BDProbeTecET System (Becton, Dickinson & Co.; Franklin Lakes, NJ) couples the proprietary technology, SDA, and real-time fluorescent detection in a rapid one-hour format. This high throughput, chip-based, closed system was developed for detection of *Chlamydia trachomatis* and *Neisseria gonorrhea* in urine samples, endocervical swabs, and male urethral swabs in a one-hour assay time. The complete system includes a sample preparation module, a priming and warming heater unit, and an amplification and fluorescence detection unit. The optical system consists of a fiber-optic bundle with eight branches, and the fluorescent detection reader monitors real-time fluorescence by FRET. Emitted light passes through a custom optical band-pass filter, is detected by a PMT, and is analyzed by software.

Little et al. (1999) (21) reported a sensitivity of 10–15 *N. gonorrhea* cells or *C. trachomatis* elementary bodies. Akduman et al. (22) reported that out of 3544 urine samples tested, 152 were positive using the BDProbeTecET System, and 130 were positive by standard culture techniques resulting in a sensitivity of 99.2% and a specificity of 99.3%.

FIBER-OPTIC FLUOROIMMUNOASSAY SYSTEMS

Analyte 2000 and RAPTOR

The Analyte 2000 and its sister field model, RAPTOR (Research International; Monroe, WA), detection systems use a fiber-optic, waveguide-based sandwich fluoroimmunoassay for the near real-time detection of pathogens in a variety of raw fluid samples (23). Optical fibers are long, thin strands of either glass or plastic that can transmit light over long distances. In the RAPTOR, a monolayer of capture antibodies are immobilized on the surface of a cylindrical waveguide (Fig. 4)(23). The waveguide is incubated with a clinical sample for three to five minutes, washed, and re-incubated with a fluorophore-labeled antibody to form an antibody/antigen/labeled-antibody "sandwich." Excitation light, injected into the waveguide, creates an evanescent wave electric field in the fluid and generates an optical emission from the antibody-antigen complexes. The fluorescent signals are then monitored by a photodetector.

Using the Analyte 2000, the detection limits for *Bacillus anthracis* (vegetative cells) was reported as 30 cfu/ml in water, and for the avirulent strain of *B. anthracis* (i.e., Sterne strain), 100 cfu/ml in whole blood. For spores, the detection limit was 5×10^4/ml (23). The infectious dose of *B. anthracis* in a healthy individual requires inhalation of about ~8,000–50,000 spores (24). This number is reduced in more vulnerable individuals, such as the elderly or those with respiratory problems. Vaccinia virus (a surrogate of the Smallpox virus) from throat swabs was detected at 2.1×10^4 pfu (plaque-forming units, the viral equivalent of bacterial colonies)/ml (25). The infectious dose of smallpox is thought to be low (i.e., 10–100 organisms) (26).

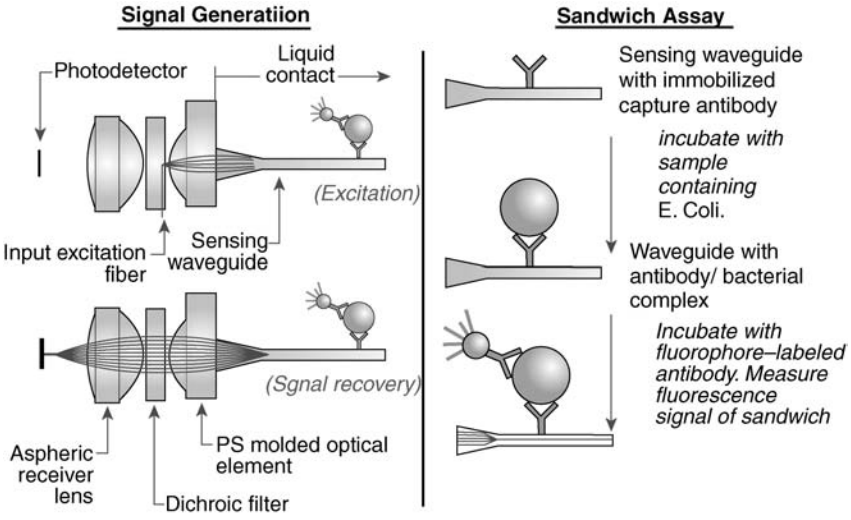


Figure 4. Optical and biomolecular processes of RAPTOR technology (23).

NANOPARTICLE-BASED BIO-BARCODE TECHNOLOGY

Verigene System

The Verigene System (Nanosphere; Northbrook, IL) is an automated device for the chip-based detection of proteins and nucleic acids using an innovative gold nanoparticle-based bio-barcode technology. For proteins, the assay uses two types of probes (Fig. 5 (27): (1) magnetic microparticles (MMPs) functionalized with monoclonal antibodies (mAbs) specific for a target antigen and (2) gold nanoparticles (NP) functionalized with polyclonal antibodies specific for the same target and DNA oligonucleotides (the "bio-barcodes") with a sequence that is a unique identification tag for the target. The Au nanoparticles and the MMPs sandwich the target, generating a complex with a large ratio of barcode DNA to protein target. A magnetic field is applied, allowing the separation of all the MMP/target/NP complexes from the reaction mixture. After a wash to dehybridize the barcode DNA from the nanoparticles, another magnetic field removes the NPs, leaving only the barcode DNA. Detection and identification of the barcodes occurs next through a PCR-less process of amplification. Chip-immobilized capture DNA, complementary with half of the target barcode DNA sequence, is used to bind the barcode DNA. Then, gold nanoparticles, functionalized with oligonucleotides that are complementary to the other half of the barcode DNA, are hybridized to the captured barcode strands. The signal is amplified by the catalytic electrodeposition of Ag onto the Au nanoparticles, and the results are recorded with the Verigene ID system, which measures scattered light intensity from each barcode/Au/Ag complex.

Like protein detection, DNA detection via the nanoparticle bio-barcode approach uses two types of probes (Fig. 6)(28): (1) magnetic microparticles functionalized with oligonucleotides that are complementary to one-half

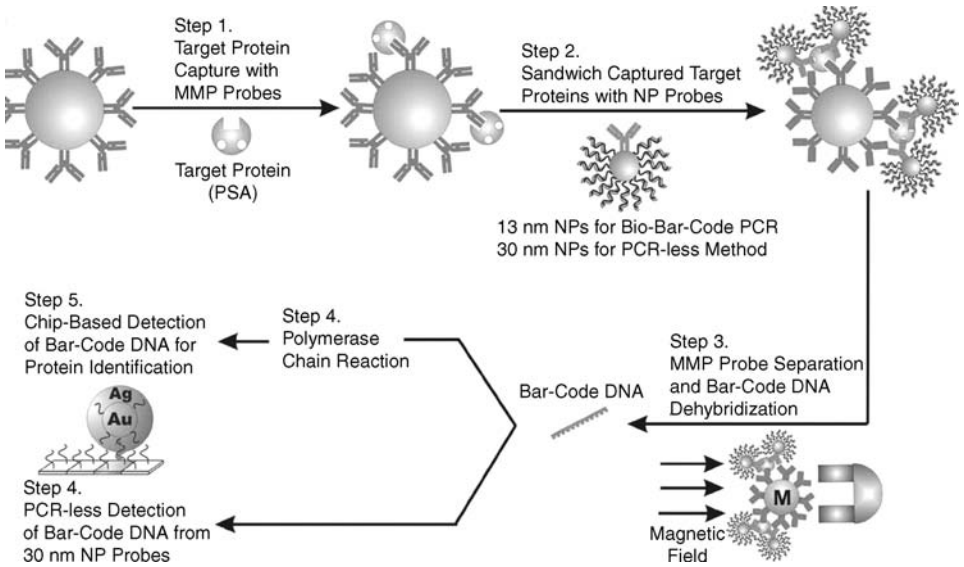


Figure 5. Prostate-specific antigen (PSA) detection and barcode DNA amplification and identification (27).

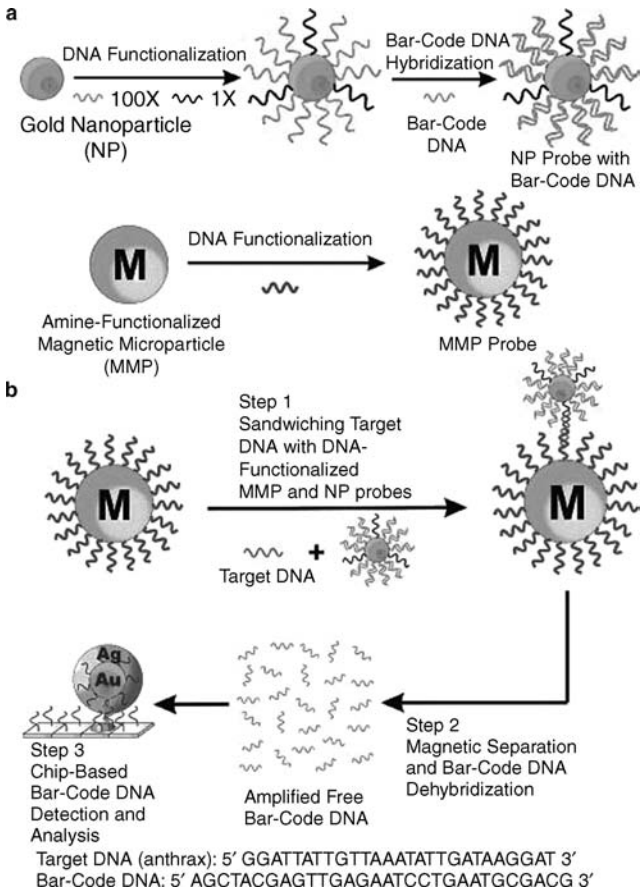


Figure 6. The DNA-bio-barcode assay. (a) Nanoparticle and magnetic microparticle probe preparation. (b) Nanoparticle-based PCR-less DNA amplification scheme (28).

of a target sequence and (2) gold nanoparticles functionalized with two types of oligonucleotides, one that is complementary to the other half of the target sequence and one that is complementary to a barcode sequence that is a unique identification tag for the target sequence. The assay proceeds in the same manner as with protein targets, with the analysis also accomplished by the scanometric method with a Verigene ID system.

The nanoparticle-based bio-barcode approach is reported to provide a sensitivity of 500 zeptomolar, approximately 10 target DNA strands in a 30 μl sample (27). Prostate-specific antigen was detected at 30 attomolar levels with this method, and PCR on the DNA barcodes boosted sensitivity to 3 attomolar (28). The entire assay can be carried out in 3–4 h.

MICROCHIP TECHNOLOGY

NanoChip System

The NanoChip System (Nanogen; San Diego, CA) is an electronic microarray device based on the electrophoretic transport of proteins and nucleic acids on a microchip to specific sites where traditional immunoassays or nucleic acid hybridization reactions occur (Fig. 7) (29). The electronic microchip is a planar array of microelectrodes that electrophoretically transport-charged biomolecules to any individually-electrically-addressed microsite on the surface of the device. Each microsite has an agarose-streptavidin permeation layer coated on top of a platinum microelectrode to bind biotinylated capture molecules. The microchips are referred to as "active electronic microchips" because electric fields are generated for the purpose of transporting biomolecules to and from specific

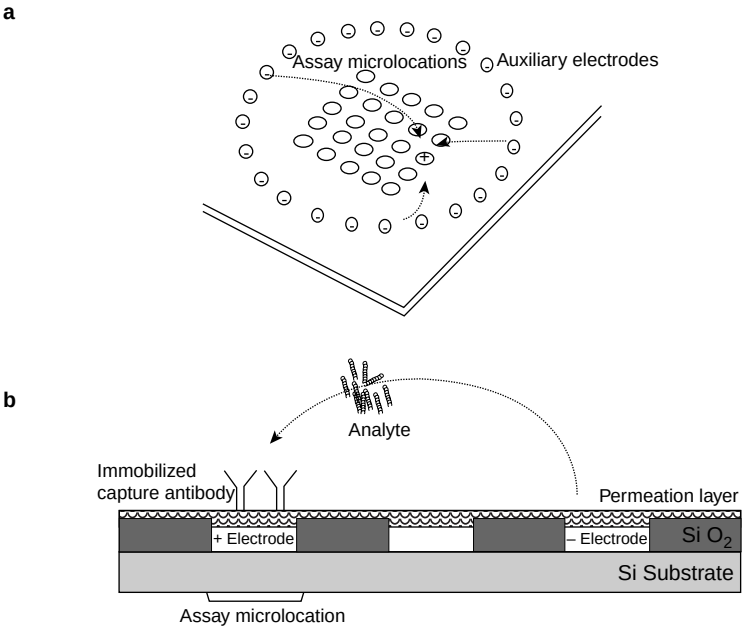


Figure 7. Active electronic microchip technology. (a) Basic chip layout; (b) Cross-section of microchip for electrophoresis of proteins (29).

microsites in a process the manufacturer terms "electronic addressing" (29). As the electric field is generated only in the immediate vicinity of the electrodes, not affecting the solution in other parts of the device, each microsite is an independent assay site allowing for the detection of multiple analytes. Also, the generated electric fields can be used to selectively dehybridize nonspecifically bound analytes from assay sites, which greatly improves the selectivity of the assay. When the biotinylated capture probes are attached to the array, fluorescently labeled analytes are introduced, and further electrical adjustments are made to direct the analytes to concentrate at the microsites for rapid hybridization or antibody-antigen interactions. The fluorescent signal is monitored in a laserinduced fluorescence scanner, and the analytes are identified based on the microlocation of the fluorescence.

Nanogen researchers performed a diagnostic immunoassay for two fluorescently labeled toxins simultaneously, staphylococcal enterotoxin B (SEB) and cholera toxin B. They reported a sensitivity of better than 20 nM concentrations of toxins (29). High specificity was also demonstrated by low nonspecific binding and cross-binding. This assay took 6 minutes to perform, 1 minute for electronic addressing to bind analytes and 5 minutes for washing to reduce nonspecific binding.

More recently, Nanogen researchers reported on an integrated "stacked" microlaboratory for performing automated electric field-driven immunoassays and DNA hybridization assays (30). This device is composed of a CMOS-based electronic microarray chip, a dielectrophoresis microchip, and several modules for DNA sample preparation, strand displacement amplification, and hybridization. *E. coli* bacteria and Alexa-labeled staphylococcal enterotoxin B were detected in the device with specific-to-nonspecific signal ratios of 4.2:1 and 3.0:1, respectively. Identification of the Shiga-like toxin gene from *E. coli* was accomplished in a 2.5 h comprehensive protocol including the dielectrophoretic concentration of intact bacteria, DNA amplification, electronic DNA hybridization to fluorescently-labeled probes, and detection with a fluorescent microscope. This experiment used bacteria cell suspensions of 10^9 cells/ml with a specific-to-nonspecific signal ratio of 22.5:1, showing outstanding specificity.

ELECTRONIC NOSE

Osmetech Microbial Analyzer

An electronic nose is a device that consists of an array of gas sensors with different selectivity patterns, a signal collecting unit, and data analysis by pattern recognition software. When microorganisms grow and metabolize, they emit volatile organic compounds and gases that can be monitored by a biosensor array. The Osmetech Microbial Analyzer (OMA; Osmetech; London, UK) is an automated headspace analyzer using arrays of organic conducting polymers as sensors. The device samples the headspace above the surface of the specimen and detects volatile compounds with an array of up to 48 conducting polymer sensors. Each polymer has unique adsorptive properties, and, once adsorbed, the volatile components modulate the conduction mechanism of the polymer resulting in reversible changes in resistance (Fig. 8) (31). The signal is measured as a percentage change of the original resistance of the polymer. Multivariate data algorithms are used to compare the responses and establish a diagnosis.

When 534 clinical urine samples were analyzed by the OMA, 22.5% had significant bacteriuria (i.e., $>10^5$ cfu/ml), resulting in a sensitivity of 84% and a specificity of 88% relative to standard culture methods (32). Alhough less than optimal, this device shows promise for automated, rapid screening. The company's second FDA approval for detection of bacterial vaginosis was secured in January 2003. Clinical trials with more refined versions of the instrument are in progress. Although electronic nose technology is still in its infancy, it clearly has the potential for providing rapid, sensitive, and simultaneous detection of different strains of bacteria.

FUTURE TRENDS IN MICROBIAL DETECTION SYSTEMS

In the development of the microbial detection systems mentioned, researchers have begun to focus on building integrated devices that combine a pre or post-processing step such as PCR-based amplification, with post-derivatization (fluorescent labeling) and detection. Microarray technologies are being developed to overcome limitations of sample volume and high throughput analysis. In the

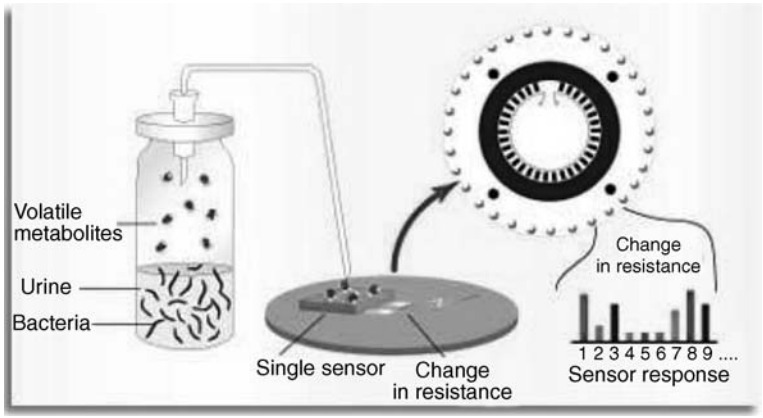


Figure 8. Osmetch Microbial Analyzer detection technology (31).

near future, biomedical science can realize the integration of all laboratory equipment used in molecular biology on a chip-based platform arrayed to detect large numbers of pathogens in a high throughput, portable device. Biological Micro-Electro-Mechanical systems (BioMEMS), also referred to as lab-on-a-chip and micro total analytical systems (μ-TAS), is an area rapidly advancing due to the integration of micro and nanotechnology with biotechnology. Current reviews for this technology are abundant in the literature (33,34). Microfluidic-based devices have been on the market since 1999, but much work is still underway to build modular-type systems with complete integration of sample collection, concentration, pre and post-processing steps, separation, selective capture, viability detection, lysing, and protein and DNA analysis. Development of such systems in a high throughput fashion capable of detecting and discriminating between hundreds of pathogenic agents would impact not only medical diagnostics but homeland security and public health, including home monitoring, medicine, and veterinary diagnostics. As the field moves toward lab-on-a-chip systems, cost, limited sample throughput, ease-of-use, and limited waste production (reagentless systems) will be considered in design strategies. The second progression toward advanced microbial detection systems will be the incorporation of nanotechnology. Current nanotechnologies such as quantum dots, nanoparticles, and synthetic nanopores are already being incorporated into current chip-based diagnostic systems. CellTracks technology (Immunicon Corporation, Huntingdon Valley, PA) has developed magnetic nanoparticles called ferrofluids, which consist of a magnetic core encompassed by a polymer coating tagged with antibodies for whole cell and pathogen detection. Up-Converting Phosphor Technology (UPT), by OraSure Technologies, Inc., makes use of proprietary ceramic nanoparticles for DNA detection. These particles have been shown to be a 1000 times more sensitive than fluorescent technologies. Finally, a trend exists to build detection systems from the bottom up rather than the top down. Small building blocks such as protein motors are being designed to move cargo including peptides and antibody fragments as a method of patterning arrays. "Switchable" materials such as poly-n-isopropylacrylamide (PNIPAM) are used to pattern antibodies, capture proteins, and move fluids, replacing mechanical components of BioMEMS systems. PNIPAM has a thermally activated lower critical solubility temperature (LCST) of 32 °C. At temperatures below the LCST, the polymer swells in water to create a hydrophilic surface that resists protein adsorption. Above the LCST, the polymer collapses to form a hydrophobic surface that promotes protein adsorption. Whether the bottom up approach based solely on nanomaterials will hold in the long run remains to be seen. However, it is clear that nanotechnology that complements and extends current MEMS detection methods will revolutionize the field of medical diagnostics. Early examples of lab-on-a-chip technologies integrating nanotechnologies already exists, which address the current limitations of detection systems. The approach is toward development of portable microsystems that are reagentless; handle small sample size; eliminate the need of labels and probes; are specific, sensitive, and high throughput; perform multiple functions from sample concentration to final detection; and are easy to use.

Briefly, some of these technologies include cantilever arrays, which operate by a slight bending of the cantilever beam at the nanoscale level upon analyte binding. Protiveris, Inc. (Rockville, MD) is developing microcantilever arrays for combined detection of DNA and protein. Capture molecules are attached to the beams and, as samples moves across the device, binding of a target molecule results in nm bending of the beam. These devices can be integrated into microfluidic systems, require no labels or reagents, and are very sensitive and specific. Nanowires, nanoneedles, and nanoelectrode arrays are additional technologies that can detect multiple analytes simultaneously. Electronic signals can be averaged over thousands of electrodes eliminating the need for PCR amplification, and no reagents are required. These devices are coated with selective molecular recognition molecules and change in conductance occurs during a binding/recognition event. Aside from integration into lab-on-a-chip systems, these technologies have applications in *in vivo* medical diagnostics.

Over the next few years, nanotechnologies will continue to evolve and become integrated into chip-based microsytems for detection, diagnostics, and drug delivery. Later, in perhaps 20– 30 years, the introduction of nanomachines for *in vivo* diagnostics and treatment may well emerge, changing the current way of conducting medical and healthcare practice.

BIBLIOGRAPHY

1. Gannon JC. The Global Infectious Disease Threat and Its Implications for the United States. [Online]. Federation of American Scientists. http://www.fas.org/irp/threat/nie99-17d.htm.
2. Fauci AS. Global Health: The United States Response to Infectious Diseases, Testimony before the U.S. Senate Labor and Human Resources Subcommittee on Public Health and Safety. [Online]. National Institute of Allergy and Infectious Diseases, National Institutes of Health. http://www3.niaid.nih.gov/about/directors/congress/1998/0303/default.htm.
3. Ivnitski D, Abdel-Hamid I, Atanasov P, Wilkins E. Biosensors for detection of pathogenic bacteria. Biosens Bioelectron 1999;14:599–624.
4. Washington JA. Principles of Diagnosis. [Online]. University of Texas Medical Brach. http://gsbs.utmb.edu/microbook/ch010.htm.
5. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 1988;239:487–491.
6. Ou CY, Kwok S, Mitchell SW, Mack DH, Sninsky JJ, Krebs JW, Feorino P, Warfield D, Schochetman G. DNA amplification for direct detection of HIV-1 in DNA of peripheral-blood mononuclear-cells. Science 1988;239:295–297.
7. Ou CY, Ciesielski CA, Myers G, Bandea CE, Luo CC, Korber BTM, Mullins JI, Schochetman G, Berkelman RL, Economou AN, Witte JJ, Furman LJ, Satten GA, Macinnes KA, Curran JW, Jaffee HW. Molecular epidemiology of HIV transmission in a dental practice. Science 1992;256:1165–1171.
8. Stuyver L, Van Geyt C, De Gendt S, Van Reybroeck G, Zoulim F, Leroux-Roels G, Rossau R. Line probe assay for monitoring

drug resistance in hepatitis B virus-infected patients during antiviral therapy. J Clin Microbiol 2000;38:702–707.
9. Stuyver L, Wyseur A, Vanarnhem W, Lunel F, Laurentpuig P, Pawlotsky JMM, Kleter B, Bassit L, Nkengasong J, Vandoorn LJ, Maertens G. Hepatitis C virus genotyping by means of 5′-UR/core line probe assays and molecular analysis of untypeable samples. Virus Res 1995;38:137–157.
10. Roche Diagnostics - Roche Molecular Diagnostics - Products. Automated PCR Systems. [Online]. Roche Molecular Diagnostics. http://www.rochediagnostics. com/ba_rmd/rmd_ products_automated_pcr_systems23.html.
11. Heid CA, Stevens J, Livak KJ, Williams PM. Real time quantitative PCR. Genome Res 1996;6:986–994.
12. Cepheid – Products. (No date). GeneXpert Technology, System Overview. [Online]. Cepheid. http://www.cepheid.com/Sites/cepheid/content.cfm?id=164.
13. Jaffe RI, Lane JD, Bates CW. Real-time identification of *Psuedomonas aeruginosa* direct from clinical samples using a rapid extraction method and polymerase chain reaction (PCR). J Clin Lab Analysis 2001;15:131–137.
14. Idaho Technology - R.A.P.I.D. Successes using the R.A.P.I.D. [Online]. Idaho Technology. http://www.idahotechnology.-com/rapid/success.htm.
15. NucliSens Key Technologies. (2005). bioMérieux – Clinical Microbiology Products. [Online]. bioMérieux, Inc. http://www.biomerieuxusa. com/clinical/nucleicacid/technology. htm.
16. NucliSens Reader. (2005). bioMérieux – Clinical Microbiology Products. [Online]. bioMérieux, Inc. http://www.biomerieux-usa.com/clinical/nucleicacid/reader.htm.
17. NucliSens EasyQ. (2005). bioMérieux – Clinical Microbiology Products. [Online]. bioMérieux, Inc. http://www.biomerieux-usa.com/clinical/nucleicacid/easyq/easyq_technology.htm.
18. DeBaar MP, Timmermans EC, Bakker M, Rooij E, van Gemen B, Goudsmit J. One-tube real-time isothermal amplification assay to identify and distinguish human immunodeficiency virus type 1 subtypes A, B, and C and circulating recombinant forms AE and AG. J Clin Microbiol 2001;39:1895–1902.
19. Lanciotti RS, Kerst AJ. Nucleic acid sequence-based amplification assays for rapid detection of West Nile and St. Louis encephalitis viruses. J Clin Microbiol 2001;39:4506–4513.
20. Walker GT, Little MC, Nadeau JG, Shank DD. Isothermal *in vitro* amplification of DNA by a restriction enzyme/DNA polymerase system. Proc Natl Acad Sci USA 1992;89:392–396.
21. Little MC, Andrews J, Moore R, Bustos S, Jones L, Embres C, Durmowicz G, Harris J, Berger D, Yanson K, Rostkowski C, Yursis D, Price J, Fort T, Walters A, Collis M, Llorin O, Wood J, Failing F, O'Keefe C, Scrivens B, Pope B, Hansen T, Marino K, Williams K, Boenisch M. Strand displacement amplification and homogeneous real-time detection incorporated in a second-generation DNA probe system, BDProbeTecET. Clin Chem 1999;45:777–784.
22. Akduman D, Ehret M, Messina K, Ragsdale S, Judson FN. Evaluation of a strand displacement amplification assay (BD ProbeTec-SDA) for detection of *Neisseria gonorrhoeae* in urine specimens. J Clin Microbiol 2002;40:281–283.
23. RAPTOR. RAPTOR, Portable, Multianalyte Bioassay System. [Online]. Research International. http://www.resrch-intl.com/raptor.html.
24. Bacillus anthracis. (June 2, 2003). *Bacillus anthracis* (anthrax). [Online]. http://microbes.historique.net/anthracis.html.
25. Donaldson KA, Kramer MF, Lim DV. A rapid detection method for Vaccinia virus, the surrogate for smallpox virus. Biosens Bioelectron 2004;20:322–327.
26. Franz DR, Jahrling PB, Friedlander AM, McClain DJ, Hoover DL, Bryne WR, Pavlin JA, Christopher GW, Eitzer EM, Jr.. Clinical recognition and management of patients exposed to biological warfare agents. JAMA 1997;278:399–411.
27. Nam J-M, Thaxton CS, Mirkin CA. Nanoparticle-based bio-barcodes for the ultrasensitive detection of proteins. Science 2003;301:1884–1886.
28. Nam J-M, Stoeva SI, Mirkin CA. Bio-bar-code-based DNA detection with PCR-like sensitivity. J Am Chem Soc 2004;126:5932–5933.
29. Ewalt KL, Haigis RW, Rooney R, Ackley D, Krihak M. Detection of biological toxins on an active electronic microchip. Anal Biochem 2001;289:162–172.
30. Yang JM, Bell J, Huang Y, Tirado M, Thomas D, Forster AH, Haigis RW, Swanson PD, Wallace RB, Martinsons B, Krihak M. An integrated, stacked microlaboratory for biological agent detection with DNA and immunoassays. Biosens Bioelectron 2002;17:605–618.
31. Osmetech. (2005). eNose Technology. [Online]. Osmetech. http://www.osmetech.plc.uk/enose.htm.
32. Aathithan S, Plant JC, Chaudry AN, French GL. Diagnosis of bacteriuria by detection of volatile organic compounds in urine using an automated headspace analyzer with multiple conducting polymer sensors. J Clin Microbiol 2001;39:2590–2593.
33. Bashir R. BioMEMS: State-of-the-art in detection, opportunities and prospects. Adv Drug Delivery Rev 2004;56:1565–1586.
34. Lee SJ, Lee SY. Micro total analysis system (μ-TAS) in biotechnology. Appl Microbiol Biotechnol 2004;64:289–299.

See also COLORIMETRY; COMPUTER-ASSISTED DETECTION AND DIAGNOSIS; COMPUTERS IN THE BIOMEDICAL LABORATORY; FLUORESCENCE MEASUREMENTS; SAFETY PROGRAM, HOSPITAL.

MICROBIOREACTORS

XIAOYUE ZHU
TOMMASO BERSANO-BEGEY
YOKO KAMOTANI
SHUICHI TAKAYAMA
University of Michigan
Ann Arbor, Michigan

INTRODUCTION

This article defines microbioreactors as micrometer scale reaction vessels in which biological reactions are performed. Whereas large-scale bioreactors mainly focus on efficiently producing desired end products, microbioreactors have additional targeted applications such as studying cellular processes under simulated physiological microenvironments, and functioning as portable cellular biosensors or implantable devices inside the body to restore tissue functions. Increasing needs in developing personalized cell-based therapies and medicines together with rapid advances in micro- and nanotechnologies ensure that the field of microbioreactors will continue to flourish. Although most microbioreactors are still in their infancy, relatively simple compared to their macroscopic counterparts, and often not fully integrated or packaged into compact self-contained platforms, they already have started to have an

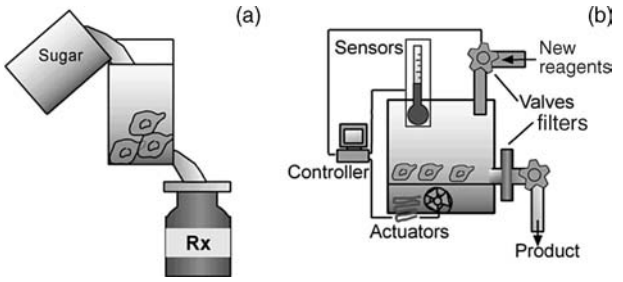


Figure 1. What are bioreactors and microbioreactors? Part (a) shows a generalized bioreactor application, feeding cells to produce complex biomolecules for pharmaceutical applications. Part (b) shows the main components of both bioreactors and microbioreactors: a reaction chamber containing main reagents and cells, sensors (for temperature, flow rate, pH, oxygen, glucose, etc.), actuators (pumps, valves, heaters, mixers, filters), and controllers to regulate actuators based on sensor readings.

impact in pharmaceutical, medical, clinical, and biological fields (Fig. 1).

Microbioreactors enable low cost high throughput investigation of bioreactions in ways not possible with macroscopic bioreactors. For example, the development of a biopharmaceutical requires extensive optimization processes prior to scale-up for industrial manufacture: The bacterial strain or mammalian cell that produces the highest yield of a particular product is constructed or identified; cultural parameters (e.g., temperature, pH, oxygen concentration, and media composition) are systematically evaluated to maximize cell growth and product formation; the design of the reactors themselves are adjusted iteratively as the development process gradually scale-up from benchtop to production scale. With microfabrication techniques, hundreds and thousands of small reaction chambers can be produced simultaneously and operated in parallel allowing vast combinations of parameters to be screened concurrently in short periods of times (Fig. 2a). The consumption of reagents and resources are reduced dramatically because for a given reaction chamber geometry, every 10-fold reduction in linear dimension would lead to a 1000 time reduction in its volume.

To enhance the quality of information obtained from cellular studies, rather than simply increasing throughput, some microbioreactors are designed to simulate physiological cellular microenvironments. Using microbioreactors for studying cellular physiology makes intuitive sense when one considers that much of the bioreactions within living organisms occur at the microscale. For example, capillary blood vessels, lung small airways, livers sinusoids, kidney nephrons, and reproductive tracts are all networks of small sacs, ducts, and tubes. Cells in these and other similar systems are constantly perfused with nutrients and oxygen and exposed to shear stresses and gradients of chemicals (1–3). Conventional *in vitro* cell culture studies, such as culture dishes or 96 well plates, often fail to present many of these dynamic physiological parameters. Microbioreactors, however, can be designed to simulate many such physical and chemical conditions that cells experience inside the body. In the body, different tissues and cell types also interact and communicate with each other. Microbioreactors can be designed to network multiple reaction chambers together to capture the complexity of living organisms and used as animal and human surrogates in pharmacokinetic and toxicology studies. Microbioreactors can also work as cell-based biosensors, where effects on cell behaviors serve as readouts for selective and sensitive detection (Fig. 2c).

Some microbioreactors aim to continuously produce and deliver small quantities of biopharmaceuticals (e.g., insulin for regulation of blood sugar) inside the body (Fig. 2e). Implantable devices that continuously produce drugs based on physiological demands would eliminate the need for repeated injections and blood tests, and allow for delivery of stable, safe, and effective doses that resemble physiological concentration profiles. Such devices contain cells that use nutrients from the body to produce and secrete drugs (4).

Besides cell-based microbioreactors, enzymatic microbioreactors are also useful for detection and analysis. Enzyme-linked immunosorbent assays (ELISA), for example, are useful in high sensitivity detection and analysis of a wide variety of proteins and chemicals related to pollution, disease, and basic biology. Microscale systems often require shorter incubation time due to the short distances that analytes and reagents have to diffuse. Microbioreactors are also more portable and thus suitable for field and point-of-care use.

The development, integration, and packaging of microbioreactors present many challenges, as well as unique opportunities. Because of scaling issues and fabrication limitations, functional microscopic devices often require totally new designs instead of simply "shrinking" their macroscopic counterparts. Thus microscale pumps, valves, and mixers not only look different from their macroscopic counterparts, but may also operate with different mechanisms. Microscale sensors, another crucial component of microbioreactors, is an active area of research that has yielded a variety of useful systems based on optical, electrochemical, ultrasonic, and mechanical detection schemes. Although many microscale components have been developed to date, functional microbioreactors that integrate multiple pumping, valving, mixing, sensing, and control features are still relatively uncommon. Microbioreactors generally have fewer components integrated into their systems compared to their larger counterparts.

The organization of the remainder of this article is as follows. First, the principles that govern microscale reactions and microbioreactor operation are discussed. Second, elements of microbioreactor components are described, along with a brief description of the microfabrication processes that can be used to create them. Finally, highlights of state-of-the-art microbioreactors are given with focuses on production optimization, clinical treatment, toxicology testing, development of implantable systems, and basic cell biology studies. Because of space limitations, this article is representative rather than comprehensive. We also focus mainly on cell-based bioreactors rather than enzyme-based ones. Interested readers are referred to reviews on immobilized microfluidic enzymatic reactors (5). We also exclude important bioreactor categories that are not directly

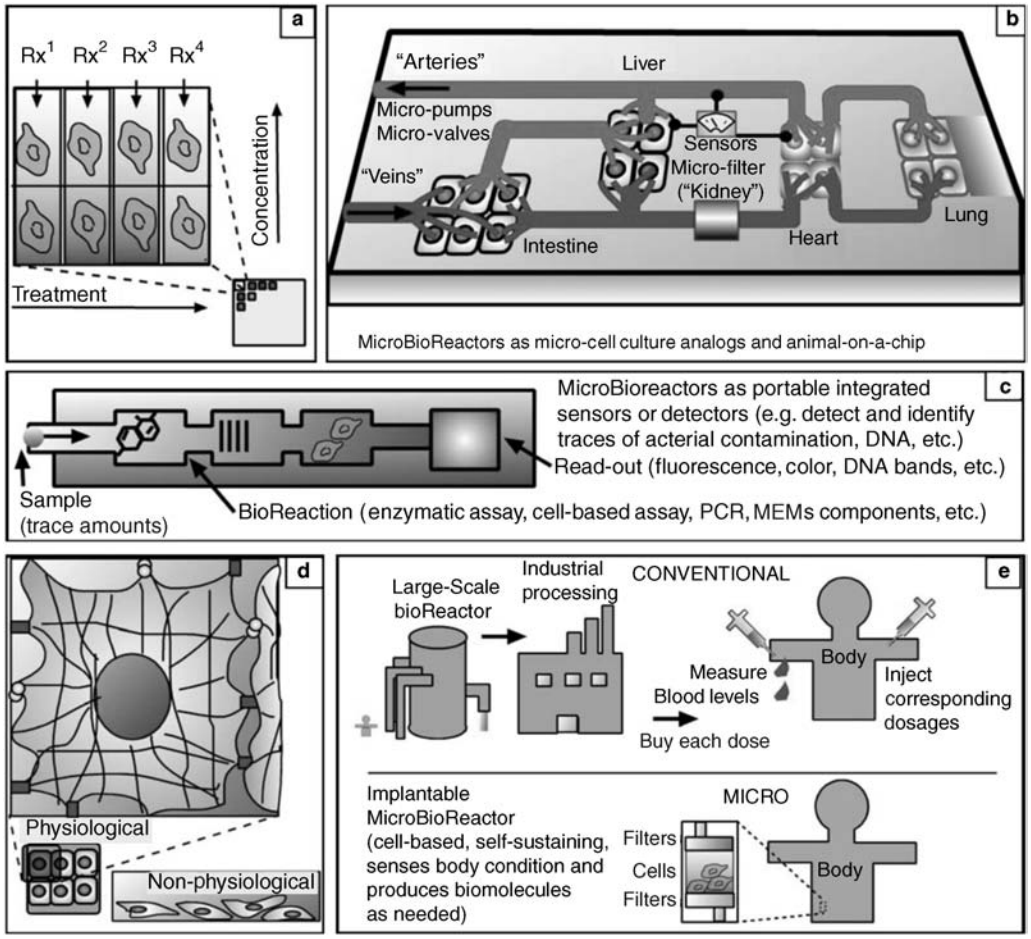


Figure 2. Advantages and new applications of microbioreactors over large scale bioreactors: (a) microscale parallel cell culture and testing in a microchip configuration can be used to test and compare hundreds of conditions and parameters simultaneously, to optimize a reaction or screen potential treatments. (b) Microdevices containing interconnected chambers and microcultured cells from different tissues can be used to test drug effects on entire organisms, taking into account systemic impact of organs such as liver on drug metabolism, possibly replacing some stages of animal and human testing. (c) Microbioreactors can be used as detectors for bioactive compounds (e.g., toxins, endocrine disruptors, drugs) or biodisruptive conditions (e.g., radiation) with the advantage that all the steps of a test can be integrated in a single device that requires only trace amounts of samples and reagents. Cell-based sensors may be able to detect unknown reagents that affect living organisms. (d) Microfabrication can also provide microbioreactors with more physiologically accurate microenvironments, thus making *in vitro* testing more reliable and closer to *in vivo* conditions: for example, in actual tissues, each cell is held in its three-dimensional (3D) shape by tension through connected cytoplasmic fibers (the cytoskeletons), and cells have many surface interactions with other surface microfeatures and other cells. In contrast, in conventional cell culture that cannot microfabricate these features, cells configuration, and environment conditions, such as the amount of surface contacts, are drastically different. (e) Implantable devices containing cells such as insulin-producing pancreatic islets can be used as an improved treatment that continuously produce insulin based on the body's minute-by-minute needs, eliminating the need for periodical blood tests and consequent injections of insulin mass-produced in conventional bioreactors.

related to medicine, such as those used for food testing, plant cell culture, and wastewater treatment.

MICROBIOREACTOR DESIGN PRINCIPLES

Basic principles of bioreactor operations, such as thermodynamics, kinetics, mass transfer, sterilization, and structural considerations, are similar for both macro- and microbioreactors. The implementation of these principles into actual practice, however, can differ greatly.

Scaling Effects

At first glance, development of microbioreactors may seem to be a matter of simply shrinking macroscopic components

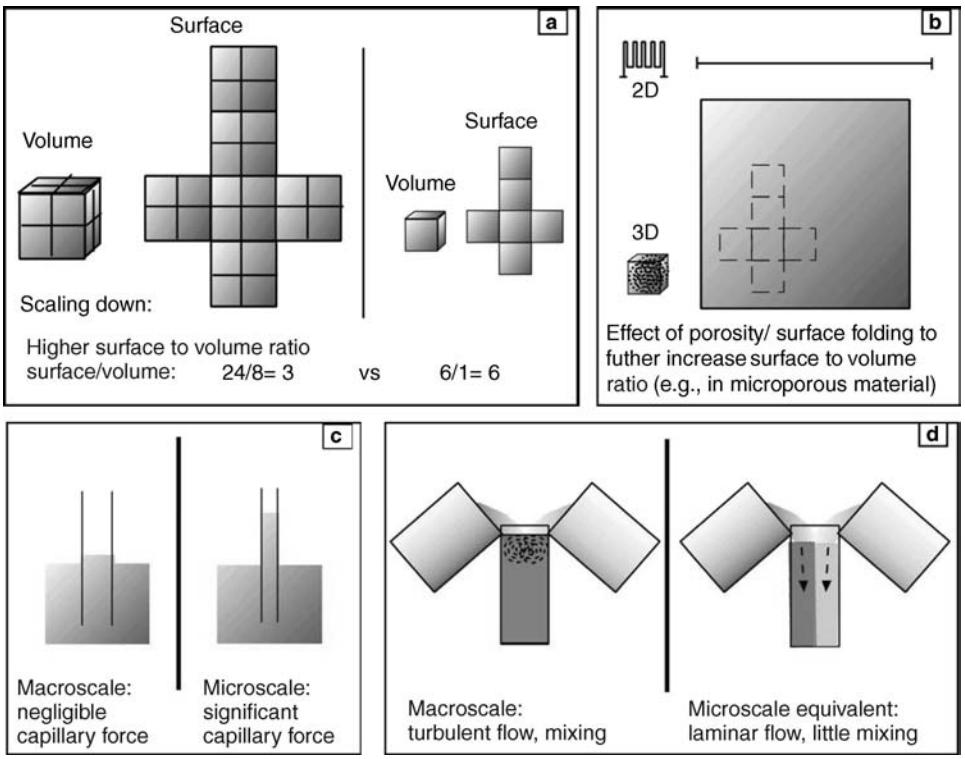


Figure 3. Microscale physics useful for microbioreactors. (a) As dimensions of devices scale down to microscopic scales, surfaces and volumes scale at different rate, so that smaller devices have a much greater surface/volume ratio. This change affects many physical phenomena such as capillary force (c) and produces effects such as laminar flow (d), which allows two liquids to flow side by side with minimal turbulent mixing. These microscale phenomena can be useful for microbioreactor opera tion. For example, capillary force can be used to pump liquid through microchannels, and laminar flows can be used to selectively deliver different reagents to different parts of single cells. (b) Microfabrication can further increase the surface/volume ratio, in a biomimetic manner, by creating microporous structures. This is useful since many biological reactions occur on surfaces.

into smaller ones. Directly downsizing systems and components, however, compromise their functions. Operations at microscales are often altered in unexpected ways compared to those at macroscales because the relative importance of physical effects that determine the functions of the microbioreactor components is size dependent. For example, if the linear dimensions of an object are reduced equally by a factor of 100, the surface area decreases by a factor of 10,000, and the volume decreases by a factor of 1,000,000 (Fig. 3a). The resulting increase in surface/volume ratio leads to prominence of surface effects such as surface tension and viscous drag over gravity and momentum. Thus, at the microscale, laminar flow, diffusive mixing, and capillary forces dominate over turbulence, convective mixing, and gravitational forces (Fig. 3c, d) (6). Scaling laws affect almost every aspect of microbioreactor operation. For example, it is more difficult to perform turbulent mixing at the microscale compared to at the macroscale (7–9), but more convenient to use electroosmosis flow to transport fluid samples through microchannels (10). More details of how scaling laws affect the design and operation of the microbioreactor components are noted in each section below.

Microbioreactor Operation Principles

The main challenge in bioreactor operation lies in its dynamic nature. Optimal reaction conditions must be maintained even as multiple parameters, such as concentration of molecules, activity of enzymes, quantity and quality of cells, and heat production are changing. Understanding the reaction mechanisms is helpful for optimizing reactor designs. It is also desirable to constantly monitor and control reaction environments in real time.

Mass And Heat Balance. In bioreactor processes, it is common to assume steady state, that is, a balance of input, output flow of materials, and thermal homeostasis. As substrate input is converted into product, mass balances need to be closely monitored and controlled. In many enzymatic processes, for example, the reactions are reversible or are inhibited by products. In such cases, it is beneficial to favor the forward enzymatic reaction by keeping the upstream reagent concentrations high and constantly removing downstream products. This type of balance can be achieved to different extents by fed-batch systems or continuous flow systems that constantly replenish and remove reactor contents. Even in so-called batch reactors, where there is no flow of liquid in and out of the reaction chamber, it is often crucial to aerate and maintain sufficient levels of oxygen at all times.

As dimensions are diminished, the conductive resistance at the channel wall/fluid interface decreases, and temperature gradient increases, leading to a greater efficiency of thermal conductivity (11). Because heat flux scales with surface area while heat capacity scales with volume, thermal time constants and heat flux scale linearly with decreasing dimensions. The small thermal mass and associated fast thermal response allows for quick temperature control. On the other hand, since the microsystems can be heated and cooled quickly, it is difficult to maintain a constant temperature; external temperatures must be set at the desired temperature, or constant heating and cooling

of the microbioreactor will be required to avoid fluctuations.

A phenomenon that may not impact macroscopic bioreactors considerably, but has a significant effect on microbioreactors, is evaporation. Unless carefully monitored, evaporation will greatly affect reactant concentrations and osmolarity because of the small volume that a microbioreactor can hold.

A variety of mathematical models have been developed to analyze and control bioreactor thermodynamics and stoichiometry. Interested readers are referred to existing texts on the topic (12). For example, Roels (13,14) developed correlations to estimate some of the important thermodynamic parameters to predict yields, substrate and oxygen requirements, and heat dissipation for cellular reactions.

Reaction Kinetics. Bioreaction kinetics depends critically on temperature, pH, dissolved oxygen concentrations, and presence of inhibitors or enhancers. Oxygen fluctuation affects the metabolic and signaling pathways of cells. For bacterial bioreactors, the main concern is often to achieve a high enough oxygen transfer rate to match the oxygen uptake rate and maintain sufficiently high dissolved oxygen concentrations. For mammalian cell cultures, it is necessary to maintain an appropriate dissolved oxygen concentration. In expansion of stem cells, for example, lower oxygen concentrations that mimic physiological values gives higher yields and purity of cells (15,16). Microbioreactors present unique challenges and opportunities in terms of oxygen transfer. Microbial and eukaryotic cell cultures require constant replenishment of dissolved oxygen into the medium because of the low solubility of oxygen in aqueous solutions (8 mg/L^{-1} at 35 °C in distilled water) (17). For larger bioreactors, bubbling oxygen or passing ambient gas through the cell culture suspension are commonly used (18). Microscale gas formation is more challenging, but has been demonstrated elegantly using electrolysis by Maharbiz et al. (19) (see Components section for details). For microbioreactors with submilliliter volumes, bubbles are generally avoided because they can clog channels and are difficult to dislodge from microchannel walls. For such systems, it is common to use a gas-permeable membrane to oxygenate the culture medium (20,21). The large surface/volume ratio of microdevices provides an advantage for such membrane-based oxygen transfer processes. Poly(dimethylsiloxane) (PDMS), a biocompatible polymer material commonly used for microdevice fabrication, is particularly advantageous in this regards due to its high permeability to oxygen.

Enzyme performance and cell growth and function are particularly sensitive to pH changes. When the reaction is not within the optimal pH range, the reaction rate declines dramatically (12). Substrates, products, or contaminants can reversibly or irreversibly affect enzyme activity. Surface interactions can also alter the kinetics of enzyme reactions, stability of enzymes (22), as well as growth and function of cells (23). Precisely dispensing nano- or picoliters of acids and bases into the tiny reaction chamber to adjust pH is technically challenging. Furthermore, because it is difficult to perform convective mixing in microbioreactors, there may be local variations in the pH and in the reaction kinetics.

Mechanical Considerations. When microbioreactors are coupled with microfluidic systems, cells are subjected to shear stresses. For a given maximum flow velocity, smaller channels will give rise to larger shear stresses (1). Although excess forces can damage cells, moderate shear stresses on genetically engineered cells have actually enhanced the production yield of recombinant proteins (24). Cells in the body are also often exposed to stretching and/or compression. These forces are critical factors that regulate function of cells in muscles, hearts, blood vessels, lungs, and other tissues (25,26). For example, shear stress regulates morphology, gene expression and function of endothelial cells, the monolayers of cells lining the inside of blood vessels (27–30). Shear stress is physiologically important for maintaining vascular homeostasis (31), regenerate bone and healing fractures (32), differentiate embryonic stem cells into cardiovascular fate, leukocytes rolling and tethering to endothelial cells (33). Nuclear factor-κβ (NF-κβ), for example, has been identified as a shear stress- responsive transcription factor that enhances the transcription of many genes including cytokines, growth factors, adhesion molecules, and immunoreceptors in response to shear stress (34).

Some of the biochemical and structural responses of cells to stretch include enhanced expression of endothelin (35), nitric oxide (36), and integrin (37) in vascular endothelial cells, and increased extracellular matrix production in cardiac fibroblasts (38) and in smooth muscle cells (39). Mechanical loading influence cellular functions *in vitro*, including proliferation (40,41), differentiation (42), hypertrophy (40,43), alignment (41,44), G protein activation, second messenger activity (45), and gene expression (43,46).

Cells in the body are constantly subjected to physical forces, such as cyclic mechanical deformation involving tension, compression, shear stress, or all three. Cell proliferation, differentiation, migration, signal transduction, and gene expression are all affected by such mechanical forces (47). Although it is largely unknown how mechanical stimuli are converted into intracellular signals of gene expression (48), there are many efforts to recreate physiological mechanical signals *in vitro*, either in the form of shear stresses from fluid flow, hydrodynamic pressures, or mechanical stresses applied to cells through the substrates. An ideal microbioreactor would provide optimal biomechanical and biodynamic stimuli for cell and tissue growth.

Material Considerations. The material used to fabricate the reaction chambers is crucial because they directly contact cells, enzymes, and products of bioreactors. Sometimes, it is also necessary to mimic physiological cellular environments (Fig. 2d) by altering the properties of the chamber surface to resemble that of the extracellular matrix (ECM). Proteins or enzymes can be immobilized onto surfaces to manipulate cell growth. Topographical features can also be incorporated to further mimic the ECM environment.

Surface Chemistry. As the size of devices decreases, their surface/volume ratio increases, making surface properties increasingly important in defining the performances of smaller bioreactors because cell function is intimately linked to the properties of the surfaces to which cells attach. Depending on the application, it is necessary to promote specific adsorption of proteins that mediate cell attachment and growth onto surfaces, or to prevent adsorption of proteins and cells. Because cellular receptors that bind to surfaces are nanoscale in size, it is also important to be able to pattern adhesive surfaces with resolutions from microns to nanometers (2,49–51). Surface properties are also important for enzyme-based microbioreactors because surface properties alter enzyme activity and stability. Enzymes are commonly immobilized through physical adsorption or covalent binding onto high surface area materials, such as carbon, silica, and polymers. Use of immobilized enzymes is often favored over free enzymes because of reduction in enzyme costs, ease of recovery, stability, and ability to be incorporated into microsystems. Ratner's recent review covers important topics including surface modification of materials to prevent nonspecific protein adsorption, immobilizing functional groups on surface, and development of synthetic materials (52).

A useful model surface for studying biomaterials interactions is a self-assembled monolayer (SAM). Often formed on gold using alkane thiols, SAMS are highly ordered arrays of linear molecules that have one end attached to the surface of gold or other bulk materials and the other end exposed to the environment. A useful feature of SAMs is that their surface properties are determined predominantly by the very terminal functional group. By altering the nature of these terminal groups, SAMs can prevent protein and cell attachment or promote binding of specific ones (53). Other types of systems that are useful for controlling protein adsorption include covalent bonding, via silanes, to silica or metal. Micropatterns of biopolymers can also be generated using photolithography, which uses patterns of light to induce region-selective chemical reactions on a surface. A more recent technique is microcontact printing, which involves transfer of small molecules or proteins onto solid substrates from an elastomeric stamp (53). Microfluidic networks have also been used for protein patterning: elastomers with embedded channel features are used to direct small volumes of protein solutions into networks of channels to create protein patterns corresponding to the path of fluid flow (54–56). Surface patterned microfluidic channels can then be used directly as microbioreactors in which micropatterned cell culture can be performed.

Topography Control. Living organisms are not flat. The endothelial lining of blood vessels, for example, exhibit an irregular wave-like topography to prevent build-up of fatty deposits (57). Endothelial monolayers have been modeled as a wavy surface by computational methods to estimate the influence of the waviness on local flow forces (58). Muscle fibers form microscale ridges and grooves onto which myoblasts can attach and proliferate (59). Microfabricated topographical features, regardless of whether they mimic physiological microtopographies or not, can be used in microbioreactors to modulate adhesion, align or orient cells, and even affect cell growth and differentiation. Examples of topographical features that have been studied include single cliffs, grooves/ridges, spikes, hills, tunnels and tubes, fibers, cylinders, mesh, waves, and random roughness. Materials used for topographical controls include gold, silicon, carbon, inorganic compounds, such as silica, lithium niobate, silicon nitride, and polymers, such as polymethylmethacrylate, silicones, cellulose acetate, collagen, fibrin, and PDMS (51). Microtopography can also affect surface wetting and fluidic flow patterns. Even for a simple groove structure, variations of the aspect ratio and contact angle of the underlying substrate materials can dramatically alter the morphology of liquid droplets contacting the surface (60).

Sterilization. Similar to larger scale bioreactors, microbioreactors and their solutions and gases can be sterilized using heat, chemicals, radiation, or through the filtration of agents. The small reaction volumes of microbioreactors provide an advantage in terms of sterilization, because for a given concentration, the total number of cells, spores, or other contaminants depends on the volume. Then, assuming that the "death rate" of the contaminant is independent of reactor size or contaminant numbers, the sterilization time needed to extinct the contaminants will decrease with decreasing reactor size. In this respect, microbioreactors are more time efficient in batch sterilization than macroscopic ones. Ultraviolet (UV) sterilization is particularly convenient for transparent microbioreactors, such as those fabricated in PDMS. When pH sensors or dissolved oxygen sensors are packaged into the microbioreactor, autoclave and UV radiation may not be feasible, and alternative methods may need to be used. Flowing 70% ethanol through the chambers/channels and subsequently drying is also sufficient for many microbioreactor applications.

Other Considerations. Phototrophic microorganisms consume light energy to survive; they can grow in simple and inexpensive nutrient media. The light requirements of phototrophic microorganisms, however, impose other significant constraints on photobioreactor designs (61). These constraints are due to an exponential attenuation of the light flow passing through an optically absorbing medium. Microbioreactors, given their large surface/volume ratios, are promising for photobioreactor applications. Cultivation of phototrophic microorganisms in optimized photobioreactors would increase the product yield several folds by maintaining the culture under appropriate conditions.

Ultrasound irradiation can change both the structure and function of biologic molecules such as proteins and deoxyribonucleic acid (DNA) (62). At mild intensity, ultrasound irradiation can increase the activity of free enzymes (63). For example, significant enhancement of reaction rate can be achieved when exposing subtilisin power to ultrasound irradiation (64). Low energy ultrasound wave irradiation can also optimize the efficiency in ethanol production by yeast from mixed office waste paper in bioreactors (65). A variety of microscale ultrasound systems have been reported. Advanced microbioreactor

systems with ultrasound components may be developed to make bioreactors more efficient.

MICROBIOREACTOR COMPONENTS

Components For Heat Transfer

External heating elements can be incorporated into microbioreactors to control temperatures. External controls of temperature include the use of standard cell culture incubators that can accommodate the whole microbioreactor (66), heating tapes, and water-to-water heat exchangers (67). Internal, embedded heaters can be comprised of materials such as thin platinum films (68) or optically transparent indium–tin oxide (ITO) (69). Temperature measurements can also be made on chip using metallic, semiconducting, or optical materials.

An important application where heaters are crucial and microbioreactors have an advantage is the polymerase chain reaction (PCR), a temperature-controlled and enzyme-mediated DNA amplification technology (70). Polymerase chain reaction requires multiple cycles of high, low, and medium temperatures to separate DNA strands, anneal the primer to the template DNA, and make complementary copies of the template DNA. Since microsystems usually have high heat conductivity and low heat capacity, the time for raising and lowering the temperature is shortened and time will be saved when cyclic temperature fluctuations during the PCR reaction is necessary. Northrup et al. (71) integrated microfabricated polysilicon heaters into a micromachined silicon reaction chamber. Schneegass et al. (68) used a thermocycler chip with integrated thin platinum film heaters and sensors for temperature. Kopp et al. (72) used external copper blocks and heating cartridges with the surface temperature monitored by a platinum thin-film resistor. Burns and co-workers (73) developed a very simple and elegant PCR device that utilizes Rayleigh–Benard convection—a steady, buoyancy-driven circulatory flow that occurs between two surfaces, one on top and one on the bottom, maintained at two fixed temperatures—to perform temperature cycling.

Components For Aeration

Electrolytic gas generation provides a compact, scalable approach for gas delivery to microbioreactors (19). In brief, a pair of interdigitated Ti/Pt electrodes hydrolyzes electrolyte to generate oxygen gases at the narrow end of a gradually widened hydrophilic microchannel. The oxygen bubbles move along the conical microchannels and transfer into culture medium because of the positive pressure built up during gas generation and the different surface tension forces at the front and back of the bubbles formed in the gradually widening channel. The rate of oxygen generation can be precisely controlled by pulse width modulation of the electrode potential.

The smaller the bioreactor, the more crucial it is to avoid bubble formation to prevent blockage of microchannels. For such systems, it is advantageous to use gas permeable membranes that allow diffusion of gases through it without introduction of bubbles. Because the surface/volume ratio is large in microsystems, oxygen transfer through gas-permeable membranes is often sufficient to ensure adequate oxygenation for biomass production. A straightforward method to fabricate gas-permeable membranes is to spincoat PDMS prepolymer onto silanized silicon wafers, cure and harden to generate a thin membrane, then peel it off (20,21).

Components For Fluid Control

Valves. Valves can be categorized into active and passive valves. A passive valve is a flow-dependent obstruction that functions without any external actuation. Passive valves are mostly unidirectional. In an active valve, fluid flow is directed by active actuation (74). The advantage that active valves have over passive valves is the degree of control one has over the timing, rate, and direction of fluid flow. This type of control is necessary to make real-time adjustments and for feedback control of microbioreactors. Although a large variety of valves have been reported, it is still a challenge to integrate multiple valves into a functional bioreactor system. Difficulties arise because many valves are incompatible with other components to be integrated, or because the valves require large external systems for actuation.

Passive valves have been used for restricting flow to one direction, removing air from liquid, or making flows stop at select channel regions. Although the level of control is lower compared to active valves, passive valves have the advantages of having few or no moving parts, less complexity, easy fabrication, and less chance to break due to fatigue (75). Recent approaches for passive control valves involve the use of hydrophobic materials, surface patterning, and changing channel fluid resistance (by changing channel geometry) (11). Passively moving micropiston valves have also been fabricated *in situ* inside microchannels using laser polymerization of a nonstick polymer (76).

Most conventional active microvalves couple a flexible diaphragm (77,78) to, thermopneumatic (79), piezoelectric (80), electrostatic, electromagnetic (81), bimetallic, or other types of actuators. The scaling of these actuation forces to the microscale, however, is often unfavorable and requires macroscale external actuators for operation. An interesting alternative to active valves is the use of autonomously regulated valves made of hydrogels that swell in response to pH or other specific chemical or thermal environment (11,82). The volume changes of the hydrogel can valve or obstruct fluid flow directly, or indirectly, through deformation of a thin PDMS membrane. The PDMS, when deformed, partially occludes an orifice to regulate the feedback stream of compensating buffer solution (17). By altering their chemistry, hydrogels can also valve in response to the changes of temperature, light, electric fields, carbohydrates, or antigens. Ehrick et al. (83) incorporated genetically engineered proteins within hydrogels that swell in response to various ligands as potential valves for microfluidic channels. Other types of valves include the use of commercially available Braille displays (84) or a pneumatic valve system to deform flexible microchannels (85).

Pumps. A micropump should ideally be able to pump a wide range of fluids and gases, be self-priming, and be programmable. Ideal micropumps are still lacking; thus, many microfluidic applications use macroscopic pumps, such as syringe pumps. Micropumps can be classified into two main types: mechanical pumps and nonmechanical pumps (71). Mechanical pumps use electromagnetic, piezoelectric, pneumatic, shape memory, electrostatic, thermopneumatic, or thermomechanic components to deliver fluid. Mechanical pumps provide higher control over average flow rates, but the flow is often pulsed and the fabrication relatively complex (53,85). Many types of nonmechanical micropumps have also been successfully developed. The flow from nonmechanical pumps is usually pulse-free with a wide range of flow rates at low pressures and the fabrication is often less complex compared to mechanical pumps. An electrokinetic pump that utilizes an electric field for pumping conductive fluids by electrophoresis or electroosmotic flow (EOF) is the most common method to control flow in microfluidic systems (71). Electrokinetic pumps have the advantages of direct control, fast response, and simplicity. However, the substrate material, joule heating effect, and microchannel charge have to be considered. Other types of nonmechanical pumps include electrohydrodynamic pumps that use electrostatic forces acting on dielectric fluids, phase-transfer pumps that use pressure gradient between gas and liquid phases, electrowetting fluid actuation systems that use interfacial forces, electrochemical pumps that use the pressure of gas bubbles generated by electrolysis of water (71), magnetohydrodynamic pumps (86), capillary force, gravity-driven pumps (87), pneumatic pumps (85), and pumps that use action of piezoelectric pin arrays in refreshable Braille displays (84).

Multiple Laminar Streams. For most flows in small channels, viscous forces dominate; thus flow is laminar and lacks turbulence. When two or more streams pass through microchannels, they flow in parallel as if they are separated by physical boundaries. Laminar flow is a challenge for mixing, but a useful phenomenon for microscale fluid patterning (Fig. 3d) (55,88). By taking advantage of diffusive mixing (but not turbulent mixing) and laminar flows, spatiotemporally defined gradients can be generated and have been used to study chemotaxis (89,90). Multiple laminar flows have also been used for developing microfluidic assay systems [i.e., T-sensor (91)], and studying subcellular processes when the interfaces of the laminar streams are positioned over a single cell (92).

Mixers. Mixing is challenging in microfluidic systems because laminar flows preclude turbulent mixing (Fig. 3d). Microscale mixers, therefore, generally use elongational flows or laminar shears to increase interfacial areas between different fluids and mixing by diffusion. Distributive mixing physically splits the fluid streams into smaller segments and redistributes them to reduce the striation thickness. Passive and active mixers have been developed for microfluidic systems, including laminating mixers (93), rotary mixers (94), mixing based on out of phase forward and backward pumping of different liquids (84), plume mixers (nozzle arrays), chaotic advection mixers (9,95), movement of liquid plugs (96), and an electroosmotically driven micromixer that uses multiple intersecting channels to enhance lateral transport (97). Thorough reviews on micromixers have been given by Hessel et al. (98) and Nguyen and Wu (99).

Components For Mechanical Stimulation

Shear. Methods commonly used to impart shear stress on cells include cone viscometers, parallel plates, and capillary tube flow chambers (100,101). Gradients of shear stress can also be generated in a curved D-shape microchannel (102). Flow-induced cytoskeleton rearrangements were shown to depend on the geometry of the channel (D-shape channel versus flat surfaces, representing experience in microcirculation and large veins, respectively) and the presence of inflammatory drugs (103).

Stretch. When subjecting cultured cells to mechanical stretch, proper design and application of a strain device are required to provide a well-defined and reproducible strain field to study mechanotransduction. Information from such *in vitro* models, which facilitate systematic variations in mechanical conditions and allow rapid analyses, would yield tremendous insights into mechanical parameters that may be important *in vivo* (104). Biaxial cell strain devices have been used to strain lung cells (105,106) by repeated mechanical deformations of a membrane on which cells are attached. Strain could also be applied using a magnetic force (107), or via uniaxial cyclical stretch (108).

Components For Separation And Purification

Lysing. Cell contents are separated from their surrounding environment by a cell membrane. The membrane and an underlying cytoskeletal network provide mechanical strength to the cell and preserve its integrity. The first step in many analyses or isolation of cell contents is to disrupt the cell membrane. There are a variety of mechanical (homogenization, milling, ultrasonic disruption, and blenders) and nonmechanical (detergent, organic solvent, osmotic shock, enzymatic permeabilization, electrical discharge, heating, and pressure cycling) methods for disrupting cell membranes on the macroscopic scale. Demonstration of cell lysis on the microscopic scale inside microfluidic devices, however, has mostly been with nonmechanical methods that use detergents (96,109), electrical discharges, or lasers (110). Miniaturized cell electrolysis devices can work with small amounts of cells and reduce the amount of purification compared to other protocols (111,112). For example, Waters et al. (113) developed a microchip that is capable of performing *Escherichia coli* lysis, PCR amplification, and eletrophoretic analysis on a single device.

Separation. Cell separation techniques are fundamental to clinical diagnosis, therapy, and biotechnological production (114). For example, it is crucial to have purified cells before proliferation and production of cellular products. Current approaches include optical tweezers (115), centrifugation (116,117), filtration

(109,116,118), fluorescence-based cell sorting (FACS) (119,120) or magnetically activated cell sorting (MACS) (121), electric field-based manipulations and separations (114,122–124), and cell-motility based sorting (125,126). Microtechnology opens new opportunities in cell and biomolecule sorting that take advantage of laminar flow behaviors (125), electrical field properties (127), or other microscale phenomena. It also provides the opportunity to combine multiple modes of separation into an integrated system. Researchers have used physiological fluid mechanical phenomenon observed in blood microcirculation (plasma skimming and leukocyte margination) to filter leukocytes from whole blood (128). Petersson et al. (129) combined acoustic wave forces and laminar flow to continuously sort erythrocytes from lipid particles in the whole blood. Because these two components were different in density and responsiveness to pressure, erythrocytes were enriched at the pressure node (along the center of the channel) and lipid particles were gathered at the pressure antinodes (along the side walls). Finally, the erythrocytes and lipid microemboli separated into different branches at the end of the main channel. Because of the variety of different properties by which cells of interest need to be sorted, it is important to develop multiple modes of cell sorting. Reviews about microfluidic cell analysis and sorting devices can be found elsewhere (130,131).

Filtration. Microfabrication techniques have been developed to integrate filters and membranes inside microbioreactors. Zhao et al. (132) and Hisamoto et al. (133) successfully produced semipermeable nylon membranes inside microfluidic channels, by taking advantage of laminar flow so that a polymerization reaction would occur at the liquid interface of the two flows and produce a thin membrane in predetermined areas of a microfluidic device.

Components For Monitoring And Control

A key requirement for microbioreactors is the ability to measure parameters, such as temperature, dissolved oxygen, pH, and flow rate. For systems involving cells, mass balances are even more complex than with enzyme bioreactors because of the larger number of products and byproducts produced and the complex responses of cells to the changes in material concentrations. In small volumes, it is difficult or impossible to use standard, macroscopic, industrial probes. Miniaturized sensors must be developed (17). Sensors that do not consume the analytes are also preferred to avoid depletion and also because sample extraction is hard to achieve in closed and compact microsystems without disrupting the devices.

Optical, electrochemical, and thin-film solid-state conductivity are the three main categories of microsensing schemes. Many of the most useful microdetection schemes are based on optical measurements such as fluorescence intensity (134), fluorescence lifetime, chemiluminescence (135), and bioluminescence (136). Optical sensing is convenient because molecular or nanoscale "sensors" are simple to introduce inside microchannels and readout can be detected from a distance without direct external contacts. In addition, many optical probes can sense without consuming oxygen or perturbing pH. Nonperturbing sensors are important for maintaining a constant microenvironment because the quantities of chemicals are small in microbioreactors. Bioluminescence and chemiluminescence are sensitive with the detection limits down to 10^{-18}–10^{-21} mol, which offers great advantage over other spectroscopic-based detection mechanisms. Laser-induced fluorescence detections in microsystems have been reviewed by Johnson and Landers (137). Other optical detection methods for microfluidic systems have been reviewed by Mogensen et al. (138).

Sol–gel-based probes encapsulated by biologically localized embedding (PEBBLEs) allow real-time measurement of molecular oxygen, pH, and ions inside and around living cells (139). Kostov et al. (17) used an optical sensing system integrated with semiconductor light sources and detectors to perform continuous measurements of pH, optical density, and dissolved oxygen in miniature bioreactors. A drawback of optical sensing is the need for light sources, lenses for focusing, and detectors, which are more challenging to miniaturize compared to the sensor probes themselves. A useful solution is to use optical fiber-based systems, which allows decoupling of the probes from the light sources and detectors, and enables detection at sites inaccessible by conventional spectroscopic sensors (140). Compared with silicon micromachining techniques, the fabrication and integration of polymeric optical elements (waveguides, lenses and fiber-to-waveguide couplers) with microfluidic channels are fast and simple (141). An oxygen-sensitive fluorescent dye has been developed to monitor dissolved oxygen levels in a system. This provides advantages over electrochemically based sensors [for a review, see Ref. 142] due to their size, ease of fabrication, and sensitivity (141).

Electronic microsensor is another category that contains a large number of useful biochemical detectors (143). Electronic microsensors usually require direct hard wiring of sensors to a readout system but can be easier to multiplex and are often smaller overall compared to optical systems. For example, Walther et al. (144) integrated a pH-ISFET (ion-sensitive field-effect transistor), a temperature-sensitive diode, and a thin-film platinum redox electrode on a single chip. The chip is mounted on a carrier and inserted into the chamber to monitor various biological parameters such as pH and redox potential. Brown and coworkers developed polymer membrane-based solid-state sensors to measure pH, and ions (145).

Microfabricated ultrasensitive nanocalorimeters can measure heat generation to monitor cell metabolic activity in response to agonist and antagonist using as few as 10 cells and without prior knowledge of the mode of action of these drugs. This measurement is noninvasive and quantitative and is envisioned to be useful for pharmaceutical companies to find drug candidate (146).

Li et al. (147) developed a microfabricated acoustic wave sensor to measure the stiffness of a single cell. This sensor is promising for drug screening and toxicology studies. For example, the acoustic wave sensor is envisioned to measure the rigidity of a heart cell to distinguish the effect of positive and negative ionotropic drugs. Positive ionotropic drugs are useful for treating congestive heart failure and

negative ionotrpic drugs for hypertension. The acoustic wave sensor is also interesting for single cell muscle physiology.

Cells themselves can be used to sense and amplify biological signals. Several groups have developed histamine sensors by integrating cells on microfluidic devices (148,149). This chip-based detector caters to the need for a simple, rapid, and safe method for allergy identification. Cells can be engineered to have a variety of biological recognition events coupled to reporter genes to specifically sense analytes of interest (150).

Some new exciting prospects for future biosensors are in the area of nanotechnology. For example, quantum dots (151) and nanoscale PEBBLES (139) are extending the limits of sensitivity, stability, biocompatibility, and flexibility in optical sensing. Quantum dots are small semiconductor nanocrystals (on the order of nanometers to a few micrometers). Fluorescent quantum dots are able to detect biological species by fluorescing only when coming in contact with viable cells, making them useful probes for many types of labeling studies. These quantum dots are photobleached very slowly and can be manufactured to emit a wide range of wavelengths. They can be used in cell biology for the labeling of cellular structures, tracking the fate of individual cells, or as contrast agents (152). Nanowire-based sensors with their small size and higher sensitivity would also be ideal for integration into microbioreactors (153). Many microsensors have been developed and are ready for integration into microbioreactors.

Fabrication

A variety of methods and types of substrates are available for microfabrication. The most traditional and widely used method of microfabrication is photolithography. Originally developed for the microelectronics industry, photolithography is precise, highly reproducible, and capable of mass production. Photolithography uses patterns of UV light coming through a photomask to area-selectively induce chemical reactions in a polymeric, light-sensitive photoresist coated onto a semiconductor substrate. During development, the light exposed regions of the photoresist are selectively removed or selectively left behind generating micropatterned photoresist structure. The exposed areas of the substrate are then chemically etched to provide features of various depths and shapes.

Due to the high equipment costs involved in photolithography and because silicon and glass substrates used in electronic and mechanical devices are not necessarily the best materials for biological applications, alternative types of microfabrication have been developed. A cost-effective and experimentally straightforward method called soft lithography has been particularly useful for biological applications (2,53). Soft lithography uses elastomeric materials, such as PDMS to create microstructures, and to pattern and manipulate surfaces. The process involves casting PDMS against a photolithographically defined master mold to yield a polymeric replica. The PDMS replica is then sealed against another material to form channels and reservoirs. Alternatively, the replica can be used in a technique called microcontact printing, where the PDMS mold is used as a stamp to transfer protein or molecular ink to a substrate. The PDMS has several properties, which make it useful for biological applications: (1) biocompatibility allows cell culture on and inside PDMS structures, (2) optical transparency allows optical inspection and sensing, (3) gas permeability allows long-term growth of cells without depletion of oxygen, (4) flexibility allows cell cultured on PDMS to be mechanically stretched (2,50,53).

Other polymer based microfabrication techniques include hot embossing, injection molding, and laser ablation. A hot embossing technique called nano-imprint lithography developed by Chou et al. (154) has the ability to fabricate sub-10 nm nanometer features. This process creates nanostructures in a resist by deforming the resist shape with embossing (155). Moriguchi et al. (156) developed a unique photothermal microfabrication technique, where agar microchamber arrays with living cells inside them can be remolded *in situ* during cell cultivation.

Integration

The development of a microbioreactor requires assembling multiple functional units (for electronic, mechanical, biological, and chemical processing) into a compact device. Integration and packaging poses a whole new challenge on top of the challenges to develop individual components. With macroscopic bioreactors, it is relatively straightforward to connect different components together with little or no worries of space organization. As one scales down, the placement of components must be performed strategically as the room around the reactor chamber decreases dramatically (because volume scales as length cubed, a 10 time reduction in linear dimensions, e.g., will lead to a 1000 time reduction in the available space). There are also technical challenges to fabricate microscale fittings and connectors, or to join two components via connectors even if they could be fabricated. A variety of processes used to build integrated circuits and microelectromechanical systems have played a major role in fabricating integrated microfluidic systems and microbioreactors. These technologies, known collectively as micromachinging, selectively etch away or deposit structural layers on silicone wafers.

Recent efforts to reduce costs, enhance material biocompatibility, and needs for diverse chemical and mechanical properties have also led to the use of a wide variety of polymeric materials in microfabricated devices. Integration, unless planned carefully, can lead to material incompatibilities in fabrication or operation. Notable accomplishments of integrated microfluidic systems include DNA analysis chips (78,157), pneumatically driven microfluidic cell sorters and protein recrystallization chips (77,119,158,159), portable cell-based biosensor systems (160), microfermentors with integrated sensors (17,20,21), and a computerized microfluidic cell culture system actuated using the pins of a refreshable Braille display (84). Integrating fluid control components for cell-culture microbioreactors is rapidly progressing, but still underdeveloped. Interested readers are referred to recent review articles on integrated microfluidic devices (161).

SPECIFIC EXAMPLES OF MICROBIOREACTORS

This section presents select examples of microbioreactors. The examples are not exhaustive, but are meant to show representative examples in each of the following four categories: (1) microbioreactors that are used to optimize bioproduction, (2) microbioreactors that provide cells with physiological microenvironments to more accurately predict physiological drug kinetics and toxicity, (3) microbioreactors that are used to develop cell-based therapies, and (4) microbioreactors for mechanistic studies. Because the field is still young, the devices are relatively simple and many are still prototypes rather than refined products ready for real world applications. The rapid advances, however, promise an increasing role of microbioreactors in the clinic, laboratory, and at home or other points of need.

Microbioreactors For Optimizing Production Conditions

With the development of microtechnologies, more and more bioprocess optimization is performed in small volume bioreactors with integrated detection systems. For example, Rao and co-workers (17) have demonstrated parallel fermentations of *E. coli*. in a milliliter-size microbioreactor. A smaller, microliter-size fermentor has been developed recently by the Jensen's group (20). The performances of these microbioreactors are comparable to traditional liter-size fermentors: Measurements of pH, dissolved oxygen (DO) and optical density (OD) of biomass in these microbioreactors have similar profiles as those in benchtop fermentors. Similarities in cellular metabolism and growth show the potential of using microbioreactors for bioprocess optimization. Arrays of microfermentors with integrated sensors and actuators are envisioned to drastically reduce the cost and time for developing new bioprocesses.

Microbioreactors For Toxicological And Drug Testing

Drugs need to be tested for toxicity and efficacy before administration to humans. Accurate prediction, however, is a challenge because most current drug tests are performed only on human cells or animals. Animal tests are expensive and time consuming, and efficacy of a drug obtained from an animal surrogate study can still be difficult to extrapolate to humans (162). Even when one uses human cells for analysis, the result may be totally different from what occurs physiologically because cells cultured in flasks or dishes experience a totally different microenvironment. Therefore, a microscale human surrogate with microcirculatory systems, three-dimensional (3D) tissue organizations, and appropriate cell–cell and tissue–tissue interactions would be beneficial in predicting human responses to drug treatment more precisely. Below are two notable examples of such efforts.

Bioartificial Livers. There are two major applications for which artificial livers are developed: one is to replace organ functions in patients with liver failure, and the other is to perform toxicology testing of drug candidates. Organ replacement functions require large bioartificial livers (BALs), whereas the toxicology studies would benefit from microscale BALs capable of conducting high throughput analyses. Microscale BALs are promising as convenient and low cost *in vitro* models for screening drug toxicities particularly in light of the fact that approximately one-half of all drug toxicities involve the liver.

Liver failure is the seventh leading cause of death by disease in the United States. About 26, 000 people died each year because of liver failure. Transplantation is limited by the supply of donor organs and the cost of the surgery (163). Extracorporeal BAL devices have been proposed as substitutes for transplantation. Macroscopic BALs have been tested in clinical trials (164). In an attempt to maximize the efficacy of the BALs, and to develop *in vitro* liver systems for biological and toxicological studies, several groups have microfabricated liver cell culture systems (165).

Microbioreactors for liver cell cultures have several configurations, ranging from flat-plate (163) or matrix-sandwiched monolayer designs (166) to 3D perfusion cultures (165). A flat-plate microbioreactor with an oxygen permeable membrane was shown to support viability and synthetic functions of hepatocytes cocultured with 3T3-J2 fibroblasts (163). This microchannel bioreactor was also functional when connected extracorporeally to a rat (162). The results indicate that this device can potentially be used as a liver support device and for the eventual scale-up to clinical devices. Compared with other configurations, monolayer designs excel in mass transfer, easy fabrication, and easy optical analysis of cells (165), although cells might be damaged by exposure to shear stress (167). Griffith and co-workers (165) developed an array of microbioreactors (with each channel $300 \times 300 \times 230$ μm, $L \times W \times H$ in dimension) that support 3D culture of liver cells by perfusion. Liver cells cultured in this device showed viable tissue structures and tight junctions, glycogen storage, and bile canaliculi. Membrane-based 3D perfusion hepatocyte culture systems have also been developed (66).

Micro CCA. It is important to integrate cells from different tissues together to simulate physiological drug metabolism. Shuler and co-workers developed Cell Culture Analogs (CCAs) that combine mathematical pharmacokinetics models with cell culture-based experimental studies to mimic human responses. The CCAs have compartmentalized cell cultures representing different tissues. Interconnections between these compartments allow circulation of media and metabolites. Since the CCAs mimic the time-dependent exposure and the metabolic interaction between multiple types of tissues and cells, predictions from the CCAs may be more accurate compared to existing *in vitro* models. Shuler and coworkers proved the concept of CCA with a macroscopic three-compartment (liver, lung, and other tissues) system and showed the feasibility and potential usefulness of such devices in testing naphthalene toxicity (168,169). MicroCCAs with three (liver, lung, and other tissues) and four (liver, lung, fat, and other tissues) compartments have also been reported (170,171).

Microbioreactors For Therapeutical Applications

Microfluidic Systems As Assisted Reproductive Technologies. Microdevices provide unique platforms for artificial reproduction and may ultimately increase the efficiency, safety, and cost-effectiveness of *in vitro* fertilization procedures. Currently, many embryo experiments are performed in macroscopic culture dishes, where human or animal oocytes (eggs) and embryos are manipulated manually. Use of pipettes for cellular manipulations is labor intensive and low in accuracy, reproducibility, and efficiency. In addition, the practice of transferring embryos from one type of media into another is abrupt and may shock the embryo due to the sudden change of environment (172–174); inside the female tract, the supply of nutrients, growth factors, and hormones changes gradually as the embryo development progresses. An alternative to manual pipetting is the use of microfluidic channels. Microfluidics is ideal for use in artificial reproduction, because it is a procedure that occurs physiologically inside small tubes and ducts, the size and numbers of cells (oocytes, sperms) required match well with dimensions of microsystems.

Beebe et al. (175) demonstrated manipulations of embryos and oocytes within microfluidic channels. The microfluidic systems can transport single mouse embryos through a channel network (176), remove the zona pellucida by chemical treatment (177), remove cumulus cells from oocytes via mechanical suction (178), and culture embryos in static or dynamic fluid environments (179). In some cases, embryos cultured in microfluidic channels develop faster compared to embryo grown in culture dishes and with growth kinetics that are closer to what is observed *in vivo*.

Cho et al. (125) developed a Microscale Integrated Sperm Sorter (MISS) that isolates motile sperms based on the ability of the motile sperms, but not the nonmotile ones, to cross-laminar flow streamlines. The device allows small volume samples that are difficult to handle with conventional sperm-sorting techniques to be sorted efficiently using a mild biomimetic sorting mechanism. The MISS integrates power source, sample injection ports, and sorting channel into one disposable polymer device making the device potentially useful not only clinically, but also as an at-home male infertility test (180).

Implantable Microcapsules. Microbioreactors that produce and release natural or recombinant bioagents are useful for delivering therapeutic agents *in vivo* to enhance metabolic function (181) or treat neurological disorders (182) and cancers (183,184). Mammalian cells, plant cells, microorganisms, enzymes, and biochemical compounds have been encapsulated in a variety of semipermeable containers mainly made of synthetic and natural hydrogels (e.g., poly(vinyl alchohol), poly(hydroxyl ethyl methacrylate), calcium alginate κ-carrageenan, chitosan, collagen, and gelatin). Here we focus on the application of microencapsulated mammalian cells. The outer membranes of the microspheres encapsulating the cells serve as selective barriers that allow exchange of nutrients, wastes, and therapeutic agents but block the passage of encapsulated cells as well as macromolecular components of the immune system. This approach has been effective in delivering genetically engineered cells that secrete growth hormone to partially correct growth retardation (185), recombinant human bone morphogenetic protein-2 (rhBMP-2) to induce bone formation and regeneration (186), interleukin-2 to delay tumor progression and prolong survival (187), coagulation factor IX for treatment of hemophilia B (188), dopamine to treat Parkinson's disease (182), insulin to maintain blood glucose level (189), and analgesic substances to relief pain (190). Most of these works are aided by using murine and canine models. Some of the most advanced systems are in early clinical trials. Biocompatibility (191,192) mechanical stability of the capsule material (193–195), efficacy (196), safety (197), and cost are still under evaluation and optimization. Reviews of therapeutic uses of microencapsulated genetically engineered cells can be found elsewhere (198).

Microbioreactors For Understanding Biological Responses

Use Of Multiple Laminar Streams To Study Subcellular Biology And Chemotaxis. Physiological cell environments are heterogeneous with local production and consumption of key growth factors, nutrients, and signaling molecules. Microfluidic systems are useful for mimicking such micropatterns of chemicals around cells. A particularly simple and useful method is to take advantage of small channel dimensions to generate multiple laminar streams that flow in parallel and adjacent to each other inside the same microchannel with minimum mixing (Fig. 3d). Such techniques can be even used to treat different parts of single living cells with different small molecular drugs, proteins, and small particles such as low density lipoprotein (LDL) (88).

Cancer and normal cells receive similar local stimuli inside the body, but behave totally differently. A critical question is why these differences arise. Taking advantage of multiple laminar flows to perform subcellular epidermal growth factor (EGF) stimulation, Sawano et al. (92) revealed differences in signal propagation between carcinoma and normal cells in response to local EGF stimulation.

Many cells direct their motion in response to chemical gradients. This phenomenon, called chemotaxis, protects microorganisms by allowing them to move toward more favorable conditions. In mammals, chemotaxis is important for guiding cell migration during development, embryogenesis, cancer metastasis, and inflammation. A challenge for analyzing chemotaxis is the lack of methods to generate well-defined chemical gradients that are stable and do not change with time. This difficulty arises due to the diffusivity of molecules and resulting changes of concentration profiles over time. Jeon et al. (90) demonstrated the use of microfluidic systems with branched networks of channels to generate stable gradients of interleukin-8 (IL-8) with linear, parabolic, and periodic concentration profiles. The position and shape of the concentration gradients were controlled by adjusting the flow rates and the positions to which reagent of interest were added into the channel network (89,199). The well-defined gradients allowed straightforward quantification of chemotaxis

coefficients as well as to observe complex migration behaviors of leukocytes in response to different concentration gradient profiles.

Studies Of Vascular Diseases In Microfluidic Channels. The mechanical forces that accompany blood flow and pressure fluctuation influence vascular cellular biology and pathology in many ways. As the blood flow along a vessel, the viscous drag forces constantly expose endothelial cells (ECs) to shear stress. The pulsatility of the blood flow also induces a periodic change of circumferential strain on ECs and their underlying smooth muscle cells (SMCs). These mechanical forces have been found to cause important biological changes in endothelial cell morphology and function, such as alignment and elongation (30,200), low density lipoprotein uptake (201), tissue plasminogen activator synthesis and secretion (202), and proliferation (67). There has also been studies about the combined effect of shear stress and cyclic strain on ECs (203–205) and SMCs (206–208). While much has been revealed about the alterations in EC function induced by mechanical stresses, relatively little is known about the mechanism mechanical signaling.

Capillary-size microfluidic channels were used to model malaria, a potentially vital disease caused by loss of deformability of red blood cells due to *P. falciparum* parasites infection. Erythrocytes exhibited increased rigidity and decreased deformability with the progression of the disease, as demonstrated by their increased difficulties to transverse through 2 to 8 μm wide PDMS channels. This type of microfluidic system may potentially be useful to screen antimalaria drugs (209).

CONCLUSION AND FUTURE PROSPECTS

The convergence of bioreactors with advances in microtechnology is starting an exciting revolution in medicine. With microfabrication technologies, many copies of a device can be generated and operated with quick procedures and reduced cost. The ability to perform parallel assays allows high throughput optimization of bioprocess conditions, opening the way for cost-efficient biopharmaceuticals production.

Humans and other living organisms are inherently microscopic in their essence, being comprised of networks of microscopic reactors (i.e., the cells), and interconnected by microfluidic vasculatures. Efforts to miniaturize bioreactors would lead to not only the development of smaller pharmaceutical production and screening systems, but also the construction of more physiological *in vitro* cell culture systems where the ultimate goal is to develop microbioreactors as animal or human surrogates. Even for cullture of single cell types, the ability of microfluidic systems to simulate physiological microenvironments is useful for revealing disease mechanisms, drug testing, use as biosensors, and single-cell-based clinical procedures such as *in vitro* fertilization. More complex microbioreactor systems with multiple cell types are being developed for toxicology studies. So-called animals-on-a-chip or minihumans provides exciting prospects for efficient drug discovery and personalized medicine.

Current state-of-the-art microbioreactors are still relatively simple with few components and limited sensing and control. Many are highly specialized and nonroutine in their use. The overall footprints of the systems are also often still macroscopic. Current advances in micro- and nanotechnologies as well as in medicine, however, promise rapid improvements in performance, accessibility, and sophistication of microbioreactors. Current trends point to a future where the gap between manmade devices and living organisms will narrow and applications of microbioreactors to medicine will grow.

ACKNOWLEDGMENTS

We thank the Whitaker Foundation, National Science Foundation (BES-0238625, DMI-0403603), National Institutes of Health (EB00379-01, HD049607-01), NASA (NNC04AA21A), and the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number DAAD19-03-1-0168 for financial support. We thank G. D. Smith, S. C. Chang, H. Chen, P.Y. Choi, B. H. Chueh, T. Kable, E. O. Kass, K. Ke, J.H. Kim, S. M. K. Lau, E.C. Lee, W.H. Lee, S.R. Mandal, J.A. Miller, Y. Murgha, S.O. Charoen, A.B. Ozel, S.J. Segvich, P.D. Settimi, D. Sud, B. Teply, M. Zhang, C. Zhong for assistance.

BIBLIOGRAPHY

1. Walker GM, Zeringue HC, Beebe DJ. Microenvironment design considerations for cellular scale studies. Lab Chip 2004;4:91–97.
2. Shim J, et al. Micro- and nanotechnologies for studying cellular function. Curr Top Med Chem 2003;3:687–703.
3. Bhadriraju K, Chen CS. Engineering cellular microenvironments to cell-based drug testing improve. Drug Discov Today 2002;7:612–620.
4. Desai TA, et al. Microfabricated immunoisolating biocapsules. Biotechnol Bioeng 1998;57:118–120.
5. Krenkova J, Foret F. Immobilized microfluidic enzymatic reactors. Electrophoresis 2004;25:3550–3563.
6. Madou MJ. Fundamentals of Microfabrication. Boca Raton (FL): CRC Press; 2002.
7. Tabeling P, et al. Chaotic mixing in cross-channel micromixers. Philos Trans R Soc London Ser A-Math Phys Eng Sci 2004;362:987–1000.
8. Stremler MA, Haselton FR, Aref H. Designing for chaos: Applications of chaotic advection at the microscale. Philos. Trans R Soc Lond Ser A-Math Phys Eng Sci 2004;362:1019–1036.
9. Stroock AD, et al. Chaotic mixer for microchannels. Science 2002;295:647–651.
10. Squires TM, Bazant MZ. Induced-charge electro-osmosis. J Fluid Mech 2004;509:217–252.
11. Ehrfeld W, Hessel V, Lowe H. Microreactors: New Technology for Modern Chemistry. Chichester : Wiley-VCH; 2000.
12. McDuffie NG. Bioreactor Design Fundamentals. Boston: Butterworth-Heinemann; 1991.
13. Roels JA. Macroscopic Thermodynamics and the Description of Growth and Product Formation in Microorganisms. Washington, (D.C.): American Chemical Society; 1983.
14. Roels JA. Energetics and Kinetics in Biotechnology. Amsterdam: Elsevier Biomedical Press; 1983.

15. Ezashi T, Das P, Roberts RM. Low O-2 tensions and the prevention of differentiation of hES cells. Proc Natl Acad Sci U S A 2005;102:4783–4788.
16. Csete M, et al. Oxygen-mediated regulation of skeletal muscle satellite cell proliferation and adipogenesis in culture. J Cell Physiol 2001;189:189–196.
17. Kostov Y, Harms P, Randers-Eichhorn L, Rao G. Low-cost microbioreactor for high-throughput bioprocessing. Biotechnol Bioeng 2001;72:346–352.
18. Blanch HW, Clark DS. Biochemical Engineering. New York: Marcel Dekker; 1997.
19. Maharbiz MM, et al. A microfabricated electrochemical oxygen generator for high-density cell culture arrays. J Microelectromech Syst 2003;12:590–599.
20. Zanzotto A, et al. Membrane-aerated microbioreactor for high-throughput bioprocessing. Biotechnol Bioeng 2004; 87:243–254.
21. Maharbiz MM, Holtz WJ, Howe RT, Keasling JD. Microbioreactor arrays with parametric control for high-throughput experimentation. Biotechnol Bioeng 2004;85:376–381.
22. Irazogui G, Villarino A, Batista-Viera F, Brena BM. Generating favorable nano-environments for thermal and solvent stabilization of immobilized beta-galactosidase. Biotechnol Bioeng 2002;77:430–434.
23. Hara M, Adachi S, Higuchi A. Enhanced production of carcinoembryonic antigen by CW-2 cells cultured on polymeric membranes immobilized with extracellular matrix proteins. J Biomater Sci-Polym Ed 2003;14:139–155.
24. Keane JT, Ryan D, Gray PP. Effect of shear stress on expression of a recombinant protein by Chinese hamster ovary cells. Biotechnol Bioeng 2003;81:211–220.
25. Dardik A, et al. Shear stress-stimulated endothelial cells induce smooth muscle cell chemotaxis via platelet-derived growth factor-BB and interleukin-1 alpha. J Vasc Surg 2005;41:321–331.
26. Kher N, Marsh JD. Pathobiology of atherosclerosis—a brief review. Semin Thromb Hemostasis 2004;30:665–672.
27. Butcher JT, Penrod AM, Garcia AJ, Nerem RM. Unique morphology and focal adhesion development of valvular endothelial cells in static and fluid flow environments. Arterioscler Thromb Vasc Biol 2004;24:1429–1434.
28. Kladakis SM, Nerem RM. Endothelial cell monolayer formation: Effect of substrate and fluid shear stress. Endothelium 2004;11:29–44.
29. Imberti B, Seliktar D, Nerem RM, Remuzzi A. The response of endothelial cells to fluid shear stress using a co-culture model of the arterial wall. Endothelium-New York 2002;9: 11–23.
30. Song J, et al. A Computer-Controlled Microcirculatory Support System for Endothelial Cell Culture and Shearing. Anal Chem 2005, In press.
31. Krizanac-Bengez L, Mayberg MR, Janigro D. The cerebral vasculature as a therapeutic target for neurological disorders and the role of shear stress in vascular homeostatis and pathophysiology. Neurol Res 2004;26:846–853.
32. Richards M, et al. Bone regeneration and fracture healing—Experience with distraction osteogenesis model. Clin Orthop Related Res 1998; S191–S204.
33. Kim MB, Sarelius IH. Role of shear forces and adhesion molecule distribution on P-selectin-mediated leukocyte rolling in postcapillary venules. Amer J Physiol-Heart Circ Phy 2004;287:H2705–H2711.
34. Blackwell TS, Christman JW. The role of nuclear factor-kappa B in cytokine gene regulation. Amer J Respir Cell Molec Biol 1997;17:3–9.
35. Awolesi MA, Sessa WC, Sumpio BE. Cyclic Strain up-Regulates Nitric-Oxide Synthase in Cultured Bovine Aortic Endothelial-Cells. J Clin Invest 1995;96:1449–1454.
36. Suzuki N, et al. Up-regulation of integrin beta (3) expression by cyclic stretch in human umbilical endothelial cells. Biochem Biophys Res Commun 1997;239:372–376.
37. Booz GW, Baker KM. Molecular signaling mechanisms controlling growth and function of cardiac fibroblasts. Cardiovasc Res 1995;30:537–543.
38. Kolpakov V, et al. Effect of mechanical forces on growth and matrix protein-synthesis in the in-vitro pulmonary-artery—analysis of the role of individual cell-types. Circ Res 1995;77:823–831.
39. Desrosiers EA, Methot S, Yahia LH, Rivard CH. Response of ligamentous fibroblasts to mechanical stimulation. Ann Chir 1995;49:768–774.
40. Komuro I, et al. Mechanical loading stimulates cell hypertrophy and specific gene-expression in cultured rat cardiac myocytes—possible role of protein-kinase-C activation. J Biol Chem 1991;266:1265–1268.
41. Neidlingerwilke C, Wilke HJ, Claes L. Cyclic stretching of human osteoblasts affects proliferation and metabolism—a new experimental—method and its application. J Orthopaed Res 1994;12:70–78.
42. Vandenburgh HH, Karlisch P. Longitudinal growth of skeletal myotubes in vitro in a new horizontal mechanical cell stimulator. In Vitro Cell Develop Biol 1989;25:607–616.
43. Reusch P, et al. Mechanical strain increases smooth muscle and decreases nonmuscle myosin expression in rat vascular smooth muscle cells. Circ Res 1996;79:1046–1053.
44. Stauber WT, Miller GR, Grimmett JG, Knack KK. Adaptation of rat soleus muscles to 4-wk of intermittent strain. J Appl Physiol 1994;77:58–62.
45. Gudi SRP, Lee AA, Clark CB, Frangos JA. Equibiaxial strain and strain rate stimulate early activation of G proteins in cardiac fibroblasts. Amer J Physiol-Cell Physiol 1998; 43:C1424–C1428.
46. Vandenburgh HH, Hatfaludy S, Sohar I, Shansky J. Stretch-induced prostaglandins and protein-turnover in cultured skeletal-muscle. Amer J Physiol 1990;259:C232–C240.
47. Cowan DB, Lye SJ, Langille BL. Regulation of vascular connexin43 gene expression by mechanical loads. Circ Res 1998;82:786–793.
48. Davies PF, Flow-mediated endothelial mechanotransduction. Physiol Rev 1995;75:519–560.
49. Blawas AS, Reichert WM. Protein patterning. Biomaterials 1998;19:595–609.
50. Zhu XY, et al. Fabrication of reconfigurable protein matrices by cracking. Nat Mater 2005.
51. Zhu XY, Bersano-Begey TF, Takayama S. Nanomaterials for Cell Engineering. In: Nalwa HS. editor, Encyclopedia of Nanoscience and Nanotechnology. American Scientific Publishers; 2004. p 857—878.
52. Ratner BD, Bryant SJ. Biomaterials: Where we have been and where we are going. Annu Rev Biomed Eng 2004;6:41–75.
53. Whitesides GM, et al. Soft lithography in biology and biochemistry. Annu Rev Biomed Eng 2001;3:335–373.
54. Delamarche E, et al. Microfluidic networks for chemical patterning of substrate: Design and application to bioassays. J Am Chem Soc 1998;120:500–508.
55. Takayama S, et al. Patterning cells and their environments using multiple laminar fluid flows in capillary networks. Proc Natl Acad Sci U S A 1999;96:5545–5548.
56. Folch A, Toner M. Cellular micropatterns on biocompatible materials. Biotechnol Prog 1998;14:388–392.

57. Grigioni M, et al. Pulsatile flow and atherogenesis: Results from in vivo studies. Int J Artif Organs 2001;24:784–792.
58. Satcher RL, Bussolari SR, Gimbrone MA, Dewey CF. The distribution of fluid forces on model arterial endothelium using computational fluid-dynamics. J Biomech Eng 1992;114:309–316.
59. Evans DJR, Britland S, Wigmore PM. Differential response of fetal and neonatal myoblasts to topographical guidance cues in vitro. Dev Genes Evol 1999;209:438–442.
60. Seemann R, et al. Wetting morphologies at microstructured surfaces. Proc Natl Acad Sci USA 2005;102:1848–1852.
61. Ogbonna JC, Tanaka H. Night biomass loss and changes in biochemical composition of cells during light/dark cyclic culture of Chlorella pyrenoidosa. J Ferment Bioeng 1996;82: 558–564.
62. Macleod RM, Dunn F. Ultrasonic irradiation of enzyme solutions. J Acoust Soc Amer 1966;40:1202.
63. Ishimori Y, Karube I, Suzuki S. Acceleration of immobilized alpha-chymotrypsin activity with ultrasonic irradiation. J Mol Catal 1981;12:253–259.
64. Vulfson EN, Sarney DB, Law BA. Enhancement of subtilisin-catalyzed interesterification in organic-solvents by ultrasound irradiation. Enzyme Microb Technol 1991;13:123–126.
65. Wood BE, Aldrich HC, Ingram LO. Ultrasound stimulates ethanol production during the simultaneous saccharification and fermentation of mixed waste office paper. Biotechnol Prog 1997;13:232–237.
66. Ostrovidov S, Jiang JL, Sakai Y, Fujii T. Membrane-based PDMS microbioreactor for perfused 3D primary rat hepatocyte cultures. Biomed Microdevices 2004;6:279–287.
67. Geiger RV, Berk BC, Alexander RW, Nerem RM. Flow-induced calcium transients in single endothelial-cells—spatial and temporal analysis. Amer J Physiol 1992;262:C1411–C1417.
68. Schneegass I, Brautigam R, Kohler JM. Miniaturized flow-through PCR with different template types in a silicon chip thermocycler. Lab Chip 2001;1:42–49.
69. Shivashankar GV, Liu S, Libchaber A. Control of the expression of anchored genes using micron scale heater. Appl Phys Lett 2000;76:3638–3640.
70. Shen KY, Chen XF, Guo M, Cheng J. A microchip-based PCR device using flexible printed circuit technology. Sensors and Actuators B-Chem 2005;105:251–258.
71. Northrup MA, Ching MT, White RM, Watson RT. DNA amplification with a microfabricated reaction chamber. Proc IEEE Int Conf Solid-state Sensors Actuators 1993; 924–926.
72. Kopp MU, de Mello AJ, Manz A. Chemical amplification: Continuous-flow PCR on a chip. Science 1998;280:1046–1048.
73. Krishnan M, Ugaz VM, Burns MA. PCR in a Rayleigh-Benard convection cell. Science 2002;298:793–793.
74. Elwenspoek M, Lammerink TSJ, Miyake R, Fluitman JHJ. Towards integrated microliquid handling systems. J Micromechanic Microengineer 1994;4:227–245.
75. Lao AIK, Lee TMH, Hsing IM, Ip NY. Precise temperature control of microfluidic chamber for gas and liquid phase reactions. Sensors and Actuators A-Phys 2000;84:11–17.
76. Altman GH, et al. Advanced bioreactor with controlled application of multi-dimensional strain for tissue engineering. J Biomech Eng 2002;124:742–749.
77. Thorsen T, Maerkl SJ, Quake SR. Microfluidic large-scale integration. Science 2002;298:580–584.
78. Grover WH, et al. Monolithic membrane valves and diaphragm pumps for practical large-scale integration into glass microfluidic devices. Sensors and Actuators B-Chem 2003;89: 315–323.
79. Vandepol FCM, Wonnink DGJ, Elwenspoek M, Fluitman JHJ. A thermo-pneumatic actuation principle for a micro-miniature pump and other micromechanical devices. Sensors and Actuators 1989;17:139–143.
80. Smits GJ. A piezoelectric micropump with three valves working peristalitically. Sensors and Actuators 1990;15:203–206.
81. Bosch D, et al. A silicon microvalve with combined electromagnetic/electrostatic actuation. Sensors and Actuators A-Phys 1993;37-8:684–692.
82. Liu RH, Yu Q, Beebe DJ. Fabrication and characterization of hydrogel-based microvalves. J Microelectromech Syst 2002; 11:45–53.
83. Ehrick JD, et al. Genetically engineered protein in hydrogels tailors stimuli-responsive characteristics. Nat Mater 2005;4: 298–302.
84. Gu W, et al. Computerized microfluidic cell culture using elastomeric channels and Braille displays. Proc Natl Acad Sci U S A 2004;101:15861–15866.
85. Unger MA, et al. Monolithic microfabricated valves and pumps by multilayer soft lithography. Science 2000;288: 113–116.
86. Lemoff AV, Lee AP. An AC magnetohydrodynamic micropump. Sens Actuators B-Chem 2000;63:178–185.
87. Zhu XY, et al. Arrays of horizontally-oriented mini-reservoirs generate steady microfluidic flows for continuous perfusion cell culture and gradient generation. Analyst 2004;129:1026–1031.
88. Takayama S, et al. Selective chemical treatment of cellular microdomains using multiple laminar streams. Chem Biol 2003;10:123–130.
89. Dertinger SKW, Chiu DT, Jeon NL, Whitesides GM. Generation of gradients having complex shapes using microfluidic networks. Anal Chem 2001;73:1240–1246.
90. Jeon NL, et al. Neutrophil chemotaxis in linear and complex gradients of interleukin-8 formed in a microfabricated device. Nat Biotechnol 2002;20:826–830.
91. Hatch A, et al. A rapid diffusion immunoassay in a T-sensor. Nat Biotechnol 2001;19:461–465.
92. Sawano A, Takayama S, Matsuda M, Miyawaki A. Lateral propagation of EGF signaling after local stimulation is dependent on receptor density. Dev Cell 2002;3:245–257.
93. Bessoth FG, deMello AJ, Manz A. Microstructure for efficient continuous flow mixing. Anal Commun 1999;36:213–215.
94. Hong JW, et al. A nanoliter-scale nucleic acid processor with parallel architecture. Nat Biotechnol 2004;22:435–439.
95. Liu RH, et al. Passive mixing in a three-dimensional serpentine microchannel. J Microelectromech Syst 2000;9:190–197.
96. Song H, Tice JD, Ismagilov RF. A microfluidic system for controlling reaction networks in time. Angew Chem Int Ed Engl 2003;42:768–772.
97. Burke BJ, Regnier FE. Stopped-flow enzyme assays on a chip using a microfabricated mixer. Anal Chem 2003;75:1786–1791.
98. Hessel V, Lowe H, Schonfeld F. Micromixers—a review on passive and active mixing principles. Chem Eng Sci 2005;60:2479–2501.
99. Nguyen NT, Wu ZG. Micromixers—a review. J Micromechanic Microengineer 2005;15:R1–R16.
100. Frangos JA, McIntire LV, Eskin SG. Shear-Stress Induced stimulation of mammalian-cell metabolism. Biotechnol Bioeng 1988;32:1053–1060.
101. Blackman BR, Barbee KA, Thibault LE. In vitro cell shearing device to investigate the dynamic response of cells in a controlled hydrodynamic environment. Ann Biomed Eng 2000;28:363–372.
102. Frame MDS, Chapman GB, Makino Y, Sarelius IH. Shear stress gradient over endothelial cells in a curved microchannel system. Biorheology 1998;35:245–261.

103. Frame MD, Sarelius IH. Flow-induced cytoskeletal changes in endothelial cells growing on curved surfaces. Microcirculation 2000;7:419–427.
104. Chien S, Li S, Shyy JYJ. Effects of mechanical forces on signal transduction and gene expression in endothelial cells. Hypertension 1998;31:162–169.
105. Clark CB, Burkholder TJ, Frangos JA. Uniaxial strain system to investigate strain rate regulation in vitro. Rev Sci Instr 2001;72:2415–2422.
106. Buck RC. Reorientation response of cells to repeated stretch and recoil of the substratum. Exp Cell Res 1980;127:470–474.
107. Mourgeon E, et al. Mechanical strain-induced posttranscriptional regulation of fibronectin production in fetal lung cells. Amer J Physiol-Lung Cell M Ph 1999;277:L142–L149.
108. Kato T, et al. Up-regulation of COX2 expression by uni-axial cyclic stretch in human lung fibroblast cells. Biochem Biophys Res Commun 1998;244:615–619.
109. Schilling EA, Kamholz AE, Yager P. Cell lysis and protein extraction in a microfluidic device with detection by a fluorogenic enzyme assay. Anal Chem 2002;74:1798–1804.
110. Soughayer JS, et al. Characterization of cellular optoporation with distance. Anal Chem 2000;72:1342–1347.
111. Lee SW, Tai YC. A micro cell lysis device. Sens Actuators A-Phys 1999;73:74–79.
112. Heo J, Thomas KJ, Seong GH, Crooks RM. A microfluidic bioreactor based on hydrogel-entrapped E. coli: Cell viability, lysis, and intracellular enzyme reactions. Anal Chem 2003;75:22–26.
113. Waters LC, et al. Microchip device for cell lysis, multiplex PCR amplification, and electrophoretic sizing. Anal Chem 1998;70:158–162.
114. Huang Y, et al. Electric manipulation of bioparticles and macromolecules on microfabricated electrodes. Anal Chem 2001;73:1549–1559.
115. Leitz G, Weber G, Seeger S, Greulich KO. The laser microbeam trap as an optical tool for living cells. Physiol Chem Phys Med Nmr 1994;26:69–88.
116. Hogman CF. Preparation and preservation of red cells. Vox Sang 1998;74:177–187.
117. Bauer J. Advances in cell separation: Recent developments in counterflow centrifugal elutriation and continuous flow cell separation. J Chromatogr B 1999;722:55–69.
118. Cheng J, Kricka LJ, Sheldon EL, Wilding P. Sample preparation in microstructured devices, microsystem technology in chemistry and life science. Top Curr Chem 1998; 215–231.
119. Fu AY, et al. A microfabricated fluorescence-activated cell sorter. Nat Biotechnol 1999;17:1109–1111.
120. Rathman M, et al. The development of a FACS-based strategy for the isolation of Shigella flexneri mutants that are deficient in intercellular spread. Mol Microbiol 2000;35:974–990.
121. Handgretinger R, et al. Isolation and transplantation of autologous peripheral CD34(+) progenitor cells highly purified by magnetic-activated cell sorting. Bone Marrow Transplant 1998;21:987–993.
122. Volkmuth WD, Austin RH. DNA electrophoresis in microlithographic arrays. Nature(London) 1992;358:600–602.
123. Li PCH, Harrison DJ. Transport, manipulation, and reaction of biological cells on-chip using electrokinetic effects. Anal Chem 1997;69:1564–1568.
124. Wang XB, et al. Cell separation by dielectrophoretic field-flow-fractionation. Anal Chem 2000;72:832–839.
125. Cho BS, et al. Passively driven integrated microfluidic system for separation of motile sperm. Anal Chem 2003;75: 1671–1675.
126. Horsman KM, et al. Separation of sperm and epithelial cells in a microfabricated device: Potential application to forensic analysis of sexual assault evidence. Anal Chem 2005;77: 742–749.
127. Chou CF, et al. Electrodeless dielectrophoresis of single- and double-stranded DNA. Biophys J 2002;83:2170–2179.
128. Shevkoplyas SS, Yoshida T, Munn LL, Bitensky MW. Biomimetic autoseparation of leukocytes from whole blood in a microfluidic device. Anal Chem 2005;77:933–937.
129. Petersson F, et al. Continuous separation of lipid particles from erythrocytes by means of laminar flow and acoustic standing wave forces. Lab Chip 2005;5:20–22.
130. Minc N, Viovy JL. Microfluidics and biological applications: the stakes and trends. C R Phys 2004;5:565–575.
131. Huh D, et al. Microfluidics for flow cytometric analysis of cells and particles. Physiol Meas 2005;26:R1–R26.
132. Zhao B, Viernes NOL, Moore JS, Beebe DJ. Control and applications of immiscible liquids in microchannels. J Am Chem Soc 2002;124:5284–5285.
133. Hisamoto H, et al. Chemicofunctional membrane for integrated chemical processes on a microchip. Anal Chem 2003; 75:350–354.
134. Kawabata T, Washizu M. Dielectrophoretic detection of molecular bindings. IEEE Trans Ind Appl 2001;37:1625–1633.
135. Wu XZ, Suzuki M, Sawada T, Kitamori T. Chemiluminescence on a microchip. Anal Sci 2000;16:321–323.
136. Roda A, et al. Biotechnological applications of bioluminescence and chemiluminescence. Trends Biotech 2004;22:295–303.
137. Johnson ME, Landers JP. Fundamentals and practice for ultrasensitive laser-induced fluorescence detection in microanalytical systems. Electrophoresis 2004;25:3513–3527.
138. Mogensen KB, Klank H, Kutter JP. Recent developments in detection for microfluidic systems. Electrophoresis 2004;25: 3498–3512.
139. Xu H, et al. A real-time ratiometric method for the determination of molecular oxygen inside living cells using sol-gel-based spherical optical nanosensors with applications to rat C6 glioma. Anal Chem 2001;73:4124–4133.
140. Wolfbeis OS. Fiber-optic chemical sensors and biosensors. Anal Chem 2002;74:2663–2677.
141. Chang-Yen DA, Gale BK. An integrated optical oxygen sensor fabricated using rapid-prototyping techniques. Lab Chip 2003;3:297–301.
142. Vandaveer WR, et al. Recent developments in electrochemical detection for microchip capillary electrophoresis. Electrophoresis 2004;25:3528–3549.
143. Bakker E, Telting-Diaz M. Electrochemical sensors. Anal Chem 2002;74:2781–2800.
144. Walther I, et al. Development of a miniature bioreactor for continuous-culture in a space laboratory. J Biotechnol 1994;38:21–32.
145. Yoon HJ, et al. Solid-state ion sensors with a liquid junction-free polymer membrane-based reference electrode for blood analysis. Sens Actuators B-Chem 2000; 64:8–14.
146. Johannessen EA, et al. Micromachined nanocalorimetric sensor for ultra-low-volume cell-based assays. Anal Chem 2002;74:2190–2197.
147. Li PCH, Wang WJ, Parameswaran M. An acoustic wave sensor incorporated with a microfluidic chip for analyzing muscle cell contraction. Analyst 2003;128:225–231.
148. Kurita R, et al. Differential measurement with a microfluidic device for the highly selective continuous measurement of histamine released from rat basophilic leukemia cells (RBL-2H3). Lab Chip 2002;2:34–38.

149. Matsubara Y, et al. Application of on-chip cell cultures for the detection of allergic response. Biosens Bioelectron 2004;19: 741–747.
150. Daunert S, et al. Genetically engineered whale-cell sensing systems: Coupling biological recognition with reporter genes. Chem Rev 2000;100:2705–2738.
151. Wu XZ, et al. Immunofluorescent labeling of cancer marker her2 and other cellular targets with semiconductor quantum dots. Nat Biotechnol 2003;21:41–46.
152. Parak WJ, Pellegrino T, Plank C. Labelling of cells with quantum dots. Nanotechnology 2005;16:R9–R25.
153. Cui Y, Wei QQ, Park HK, Lieber CM. Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species. Science 2001;293:1289–1292.
154. Chou SY, Krauss PR, Renstrom PJ. Imprint lithography with 25-nanometer resolution. Science 1996;272:85–87.
155. Raiteri R, Grattarola M, Butt HJ, Skladal P. Micromechanical cantilever-based biosensors. Sens Actuators B-Chem 2001;79:115–126.
156. Moriguchi H, et al. An agar-microchamber cell-cultivation system:flexible change of microchamber shapes during cultivation by photo-thermal etching. Lab Chip 2002;2:125–130.
157. Burns MA, et al. An integrated nanoliter DNA analysis device. Science 1998;282:484–487.
158. Hansen CL, Skordalakes E, Berger JM, Quake SR. A robust and scalable microfluidic metering method that allows protein crystal growth by free interface diffusion. Proc Natl Acad Sci USA 2002;99:16531–16536.
159. Chou HP, Spence C, Scherer A, Quake S. A microfabricated device for sizing and sorting DNA molecules. Proc Natl Acad Sci USA 1999;96:11–13.
160. DeBusschere BD, Kovacs GTA. Portable cell-based biosensor system using integrated CMOS cell-cartridges. Biosens Bioelectron 2001;16:543–556.
161. Erickson D, Li DQ. Integrated microfluidic devices. Anal Chim Acta 2004;507:11–26.
162. Shito M, et al. In vitro and in vivo evaluation of albumin synthesis rate of porcine hepatocytes in a flat-plate bioreactor. Artif Organs 2001;25:571–578.
163. Tilles AW, et al. Effects of oxygenation and flow on the viability and function of rat hepatocytes cocultured in a microchannel flat-plate bioreactor. Biotechnol Bioeng 2001;73:379–389.
164. Nyberg SL, et al. Primary hepatocytes outperform Hep G2 cells as the source of biotransformation functions in a bioartificial liver. Ann Surg 1994;220:59–67.
165. Powers MJ, et al. Functional behavior of primary rat liver cells in a three-dimensional perfused microarray bioreactor. Tissue Eng 2002;8:499–513.
166. Bader A, et al. Development of a small-scale bioreactor for drug metabolism studies maintaining hepatospecific functions. Xenobiotica 1998;28:815–825.
167. Allen JW, Bhatia SN. Improving the next generation of bioartificial liver devices. Semin Cell Dev Biol 2002; 13:447–454.
168. Ghanem A, Shuler ML. Combining cell culture analogue reactor designs and PBPK models to probe mechanisms of naphthalene toxicity. Biotechnol Prog 2000;16:334–345.
169. Ghanem A, Shuler ML. Characterization of a perfusion reactor utilizing mammalian cells on microcarrier beads. Biotechnol Prog 2000;16:471–479.
170. Park TH, Shuler ML. Integration of cell culture and microfabrication technology. Biotechnol Prog 2003;19:243–253.
171. Viravaidya K, Shuler ML. Incorporation of 3T3-L1 cells to mimic bioaccumulation in a microscale cell culture analog device for toxicity studies. Biotechnol Prog 2004;20:590–597.
172. Kruip TAM, Bevers MM, Kemp B. Environment of oocyte and embryo determines health of IVP offspring. Theriogenology 2000;53:611–618.
173. Bavister BD. Interactions between embryos and the culture milieu. Theriogenology 2000;53:619–626.
174. Fukui Y, Lee ES, Araki N. Effect of medium renewal during culture in two different culture systems on development to blastocysts from in vitro produced early bovine embryos. J Anim Sci 1996;74:2752–2758.
175. Beebe D, et al. Microfluidic technology for assisted reproduction. Theriogenology 2002;57:125–135.
176. Glasgow IK, et al. Handling individual mammalian embryos using microfluidics, IEEE Trans Biomed Eng 2001;48:570–578.
177. Zeringue HC, Wheeler MB, Beebe DJ. Zona pellucida removal of mammalian embryos in a microfluidic systems. Micro Total Analysis Syst 2000; 214–217.
178. Zeringue HC, Beebe DJ, Wheeler MB. Removal of cumulus from mammalian zygotes using microfluidic techniques. Biomed Microdevices 2001;3:219–224.
179. Hickman DL, Beebe DJ, Rodriguez-Zas SL, Wheeler MB. Comparison of static and dynamic medium environments for culturing of pre-implantation mouse embryos. Comparative Med 2002;52:122–126.
180. Suh RS, et al. Rethinking gamete/embryo isolation and culture with microfluidics. Hum Reprod Update 2003;9:451–461.
181. Korbutt GS, et al. Improved survival of microencapsulated islets during in vitro culture and enhanced metabolic function following transplantation. Diabetologia 2004;47:1810–1818.
182. Vallbacka JJ, Nobrega JN, Sefton MV. Tissue engineering as a platform for controlled release of therapeutic agents: implantation of microencapsulated dopamine producing cells in the brains of rats. J Control Release 2001;72:93–100.
183. Takenaga M, et al. A single treatment with microcapsules containing a CXCR4 antagonist suppresses pulmonary metastasis of murine melanoma. Biochem Biophys Res Commun 2004;320:226–232.
184. Yu BL, Chang TMS. Effects of long-term oral administration of polymeric microcapsules containing tyrosinase on maintaining decreased systemic tyrosine levels in rats. J Pharm Sci 2004;93:831–837.
185. AlHendy A, Hortelano G, Tannenbaum GS, Chang PL. Growth retardation—an unexpected outcome from growth hormone gene therapy in normal mice with microencapsulated myoblasts. Hum Gene Ther 1996;7:61–70.
186. Zilberman Y, et al. Polymer-encapsulated engineered adult mesenchymal stem cells secrete exogenously regulated rhBMP-2, and induce osteogenic and angiogenic tissue formation. Polym Adv Technol 2002;13:863–870.
187. Cirone P, Bourgeois JM, Austin RC, Chang PL. A novel approach to tumor suppression with microencapsulated recombinant cells. Hum Gene Ther 2002;13:1157–1166.
188. Hortelano G, Wang L, Xu N, Ofosu FA. Sustained and therapeutic delivery of factor IX in nude haemophilia B mice by encapsulated C2C12 myoblasts: Concurrent tumourigenesis. Haemophilia 2001;7:207–214.
189. Chen JP, et al. Microencapsulation of islets in PEG-amine modified alginate-poly(L-lysine)-alginate microcapsules for constructing bioartificial pancreas. J Ferment Bioeng 1998;86:185–190.
190. Xue YL, et al. Pain relief by xenograft of subarachnoid microencapsulated bovine chromaffin cells in cancer patients. Prog Nat Sci 2000;10:919–924.

191. Cole DR, et al. Transplantation of microcapsules (a potential bioartificial organ)—biocompatibility and host-reaction. J Mater Sci-Mater Med 1993;4:437–442.
192. Lou WH, Qin XY, Wu ZG. Preliminary research on biocompatibility of alginate-chitosan-polyethyleneglycol microcapsules. Minerva Biotecnol 2000;12:235–240.
193. Van Raamsdonk JM, Cornelius RM, Brash JL, Chang PL. Deterioration of polyamino acid-coated alginate microcapsules in vivo. J Biomater Sci-Polym Ed 2002;13:863–884.
194. Sakai S, Ono T, Ijima H, Kawakami K. Behavior of enclosed sol- and gel-alginates in vivo. Biochem Eng J 2004;22:19–24.
195. Koch S, et al. Alginate encapsulation of genetically engineered mammalian cells:comparison of production devices, methods and microcapsule characteristics. J Microencapsul 2003;20:303–316.
196. Arica B, et al. Carbidopa/levodopa-loaded biodegradable microspheres:in vivo evaluation on experimental Parkinsonism in rats. J Control Release 2005;102:689–697.
197. Leblond FA, et al. Studies on smaller (similar to 315 (μM) microcapsules: IV. Feasibility and safety of intrahepatic implantations of small alginate poly-L-lysine microcapsules. Cell Transplant 1999;8:327–337.
198. Chang TMS, Prakash S. Therapeutic uses of microencapsulated genetically engineered cells. Mol Med Today 1998;4: 221–227.
199. Jeon NL, et al. Whitesides GM. Generation of solution and surface gradients using microfluidic systems. Langmuir 2000;16:8311–8316.
200. Sprague EA, Steinbach BL, Nerem RM, Schwartz CJ. Influence of a laminar steady-state fluid-imposed wall shear-stress on the binding, internalization, and degradation of low-density lipoproteins by cultured arterial endothelium. Circulation 1987;76:648–656.
201. Diamond SL, Eskin SG, McIntire LV. Fluid-flow stimulates tissue plasminogen-activator secretion by cultured human-endothelial cells. Science 1989;243:1483–1485.
202. Levesque MJ, Sprague EA, Schwartz CJ, Nerem RM. The influence of shear-stress on cultured vascular endothelial-cells—the stress response of an anchorage-dependent mammalian-cell. Biotechnol Prog 1989;5:1–8.
203. Gomes N, et al. Shear stress modulates tumour cell adhesion to the endothelium. Biorheology 2003;40:41–45.
204. Davies PF, Tripathi SC. Mechanical-stress mechanisms and the cell—an endothelial paradigm. Circ Res 1993;72:239–245.
205. Ikeda M, et al. Extracellular signal-regulated kinases 1 and 2 activation in endothelial cells exposed to cyclic strain. Am J Physiol-Heart Circul Physiol 1999;276: H614–H622.
206. Smith PG, Roy C, Zhang YN, Chauduri S. Mechanical stress increases RhoA activation in airway smooth muscle cells. Am J Respir Cell Mol Bio 2003;28:436–442.
207. Lee T, Kim SJ, Sumpio BE. Role of PP2A in the regulation of p38-MAPK activation in bovine aortic endothelial cells exposed to cyclic strain. J Cell Physiol 2003;194:349–355.
208. Han O, Takei T, Basson M, Sumpio BE. Translocation of PKC isoforms in bovine aortic smooth muscle cells exposed to strain. J Cell Biochem 2001;80:367–372.
209. Shelby JP, et al. A microfluidic model for single-cell capillary obstruction by Plasmodium falciparum infected erythrocytes. Proc Natl Acad Sci USA 2003;100:14618–14622.

See also MICROARRAYS; MICROFLUIDICS; NANOPARTICLES; TISSUE ENGINEERING.

MICRODIALYSIS SAMPLING

JULIE A. STENKEN
Rensselaer Polytechnic Institute
Troy, New York

MICRODIALYSIS SAMPLING: NON-SPECIALIST VIEW

Microdialysis sampling devices are minimally invasive miniature dialyzers that can be implanted into a distinct tissue region to obtain a chemical snapshot over an integrated time period. In combination with appropriate chemical detection methods for the targeted substances, a microdialysis sampling device may be considered to be a universal biosensor. Obtaining chemical information from different tissues can often lead to either a greater understanding of the underlying chemistry involved with the physiological function of the organ or the origin of a particular disease process. A simplified view of the microdialysis sampling device is shown in Fig. 1. The central part of the microdialysis sampling device is a single semipermeable hollow fiber membrane with dimensions that range between 200 and 500 μm for its external diameter and 1 and 30 mm in length. A perfusion solution is passed through the device at microliter per minute flow rates. Compounds diffuse from the tissue space into the dialysis probe and are carried to an outlet to undergo chemical analysis. Originally microdialysis sampling was developed to obtain real-time chemical information from rodent brain and was termed intracranial dialysis. Microdialysis sampling has now been applied for chemical collection from nearly every single organ. In addition to neurotransmitter collection, the device has been used for endocrinology, immunology, metabolism, and pharmacokinetic applications as shown in Table 1. The biomedical literature cites

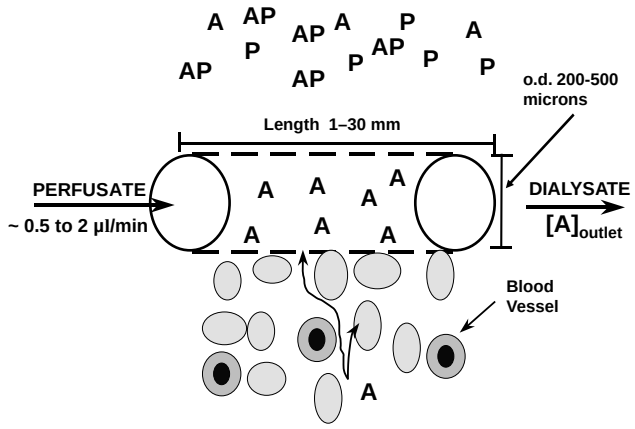


Figure 1. Microdialysis sampling process. A perfusion fluid that closely matches the ionic strength and composition of the fluid external to the microdialysis membrane is passed through at flow rates between 0.5 and 2.0 $\mu L \cdot min^{-1}$. Analytes, A, that are not protein bound, AP, diffuse through the extracellular fluid space (wavy lines) and can pass through the pores of the semipermeable membrane are collected into the device. The analyte outlet concentration $[A]_{outlet}$ can then be quantified using a suitable detection method.

Table 1. Typical Microdialysis Sampling Uses

Collection of endogenous analytes from brain including neurotransmitters (dopamine, norepinephrine, serotonin, glutamate, GABA[a], glucose, peptides and proteins (cytokines).
Collection of endogenous small hydrophilic analytes from other tissues (e.g., glucose for diabetics).
Collection of peptides and proteins from perphiral tissues, for example, glutathione, neuropeptides, and different cytokines and growth factors (e.g., VEGF).
Collection of xenobiotic analytes for pharmacokinetic or pharmacodynamic studies from numerous tissue sites (brain, dermis, muscle, and tumors) in both animals and humans.
Localized delivery of analytes followed by concomitant recovery of endogenous analytes or metabolized products (e.g., spin traps, metabolites).

[a]GABA = r-aminobutyric acid

thousands of research articles that have used microdialysis sampling devices to study many different basic and clinical research problems with perhaps 90% or greater of these applications focused in neuroscience. As more life scientists realize that microdialysis sampling devices exist, this device will be used more often to solve many additional clinical problems outside of the neurosciences.

INTRODUCTION

Mammalian organs are highly complex systems that achieve their functions through a multifaceted chemical communication network. For laboratory studies focused on understanding these chemical networks, the organ is often broken down into its component parts (cells, subcellular components, extracellular matrix components, etc.) to create more controlled conditions. These types of studies have allowed for a great understanding of the individual component parts, but do not address how the chemical communication may occur within the intact organ. Common diagnostic collection methods, such as blood or urine sampling, are too far removed from organs to be able to provide desired information about localized organ biochemistry. Gaining chemical information from an organ system prior to the creation of microdialysis sampling devices required either organ dissection or noninvasive analysis methods. Removing organs to access chemical content is fraught with many concerns including the preparation of the sample that involves in most cases sacrifice of the animal (or a biopsy for humans) as well as concerns about chemical stability and loss during the sample preparation process. For nearly all organs, the changes in localized tissue chemistry caused during sacrifice may severely alter some chemical communication systems. Finally, organ dissection does not allow for temporal studies of targeted analytes.

An increasing number of medical devices and instruments are becoming available to noninvasively measure *in vivo* chemical composition at spatially defined sites. Noninvasive measurements typically use spectroscopic instruments that are highly analyte specific. In particular, the most well-known medical devices for achieving these tasks are positron emission tomography (PET) and magnetic resonance imaging (MRI). Positron emission tomography scanning requires production of special isotopes (^{11}C, ^{13}N, ^{15}O, ^{18}F) with limited half-lives. Similarly, MRI uses ^{1}H to create an image and is a useful imaging technique because the concentration of hydrogen nuclei in water and fat is roughly 100 M ($mol \cdot /L^{-1}$). Magnetic resonance spectroscopy (MRS) can be used to noninvasively detect other isotopes (^{1}H, ^{13}C, ^{15}N, ^{19}F, ^{31}P), but requires very high concentrations of these nuclei for detection. Fluorescence imaging has also been used for noninvasive measurements in conjunction with near-infrared (IR) tags or with enzyme substrates that become fluorescent when cleaved (1–3). While these spectroscopic techniques hold significant promise for some areas of clinical medicine (cardiovascular and oncology), they are quite limited with respect to the range of analytes that can be detected and thus potential applications. While not noninvasive, minimally invasive microdialysis sampling devices coupled with appropriate dialysate analytical detection methods allow for measurements of localized tissue chemistry with significantly greater analyte flexibility and spatial resolution.

Microdialysis sampling originated as an alternative to push–pull perfusion devices for use in mammalian brain. Push–pull perfusion devices were used to collect fluid relevant to synaptic transmission (4). During a push–pull perfusion, a solution that closely matches the ionic chemical composition of the extracellular fluid (ECF) is loaded into a syringe and this is gently infused into the implantation site. This perfusion fluid mixes with the existing ECF and then is pulled back into the device. The analysis of push–pull samples became a useful tool for tracking neurotransmitter activities coupled with allowing for a remarkable understanding of the underlying chemical events associated with a wide variety of behavioral and physiological stimuli. Insertion of push–pull cannula could potentially cause tissue damage or lesions, which raised many concerns among researchers since this type of damage could limit the usefulness of the neurochemical data collected. In other words, researchers were often concerned that sampling was really taking place in a "lake" of fluid that may not contain the representative chemical components of the extracellular fluid space.

The push–pull sampling method was principally applied to neurochemical sampling. For other tissues, other extracellular fluid sampling methods have been described. These other methods include open-flow microperfusion and wick methods. Both of these methods have been used in muscle, as well as dermal sampling. Additionally, blister methods are common for obtaining interstitial fluid from the skin. Open-flow microperfusion is similar to push–pull perfusion. A cannula device contains an inner tube and an outer tube with open pores (typically $\sim$500 μm) is placed over this inner tube. A peristaltic pump then simultaneously delivers and withdraws fluid through the device (5). Open-flow microperfusion has primarily been used for sampling the extracellular fluid in muscle and skin. Alternatives to perfusion methods are wick methods, which use nylon wicks to sample the extracellular fluid space in animals and humans (6–8). To obtain multiple samples over a specified time period would require insertion and

removal of individual wick devices. This repeated insertion and removal might cause additional trauma to the sampling site.

To overcome the tissue damage concerns associated with push–pull perfusion for neuroscience applications, the use of semipermeable dialysis membranes at the tips of the push–pull cannula were used (9,10). These original dialysis bags eventually led to flow-through microdialysis probes introduced by Pycock and Ungerstedt in 1974 (11). The advantage of microdialysis sampling over the push–pull cannula is that fluid is not pushed into sensitive brain tissue. Unlike a push–pull perfusion, microdialysis sampling is a continuous process that provides a sample that excludes many of the components from the ECF. This exclusion process serves to provide a relatively clean sample for chemical analysis. Today, this technique has been widely used by life scientists to attain site-specific access to numerous tissue sites to study and solve many problems, which include, but are not limited to the following applications: (*1*) To elucidate the role of different neurotransmitters and neuropeptides in specifically defined brain regions; (*2*) To collect glucose continuously over many days to give a better chemical picture to diabetes specialists to determine the efficacy of an insulin regimen for individual patients; (*3*) To collect energy metabolites (glucose, lactate, pyruvate) as well as administered drugs (morphine) to understand diseased energy metabolism and blood-brain barrier transport from head trauma patients; (*4*) To determine if an efficacious drug concentration is reaching a specific infection site or diseased tissue sites for antineoplastic therapy or antimicrobial therapy; (*5*) To determine pharmacokinetic parameters in single animals; (*6*) To determine metabolite formation and accumulation in various tissues after a drug dose in a single animal over time; (*7*) To determine the extent of drug blood–brain barrier permeation of new drugs in animal models; (*8*) To collect various endogenous peptides and proteins (e.g., cytokines and growth factors) from different peripheral sites in animals and humans (*12*).

To newcomers to this device, the principles of microdialysis sampling operation at first seem deceptively simple. At the most basic level, the microdialysis sampling device may be considered to behave as an artificially implanted blood vessel that allows free analyte diffusion into the inner fiber lumen. However, as will be discussed in this article, numerous considerations that involve both an understanding of the localized biology and physiology, as well as the underlying mass transport processes are essential to microdialysis sampling data interpretation.

MICRODIALYSIS SAMPLING PRINCIPLES OF OPERATION

In principle, as long as the microdialysis probe can be implanted, it can be used for sampling localized tissue biochemistry. Microdialysis sampling requires only a few pieces of equipment. This equipment is not cost-prohibitive, which is the reason that many different researchers can perform microdialysis sampling experiments in their own laboratories. For most experiments, the necessary equipment includes a perfusion pump to deliver the perfusion fluid and a microdialysis probe. In addition to the pump and probe, for animal studies, a device to hold either an anesthetized animal (e.g., stereotaxic unit for neuroscience procedures) or a bowl with appropriate swivels to prevent tangled tubing for freely moving animals may be necessary.

Microdialysis Sampling Instrumentation Components

The basic components needed to perform microdialysis sampling experiments in an awake-freely moving animals has been described (13). The components required for this type of experiment includes the microperfusion pump, an inlet and outlet fluid swivel that prevents the fluid lines from becoming tangled, and a bowl system to allow the animal to freely move. Additional components can include refrigerated fraction collectors to allow collection and storage of sensitive samples. Microperfusion pumps used to deliver the perfusion fluid through the microdialysis probe are capable of delivering volumetric flow rates between 0.1 and 20 $\mu L \cdot min^{-1}$. Flow rates between 0.5 and 2.0 $\mu L \cdot min^{-1}$ are commonly used during most microdialysis sampling experiments.

Microdialysis sampling perfusion fluids are chosen to closely match the ionic strength and composition of the external tissue extracellular fluid surrounding the microdialysis probe. Perfusion fluids passed through microdialysis sampling probes are a form Ringer's solution for which there are numerous published chemical compositions (14,15). Typical Ringer's solutions contain ~150 m*M* NaCl, 4 m*M* KCl, and 2.4 m*M* $CaCl_2$ and can also be supplemented with glucose and other ionic salts ($MgCl_2$). These solutions are used both to maintain fluid balance, as well as ion balance across the dialysis membrane. Maintaining fluid balance across the dialysis membrane is important so that large osmotic pressures are not created. Significant osmotic pressure differences will cause fluid to be gained or lost during microdialysis sampling (16). Fluid loss is often undesirable for analytical as well as biological reasons. From an analytical perspective, oftentimes the analysis requires a set volume. For example, a liquid chromatographic analysis may require 10 μL of sample and a standard enzyme-linked immunosorbent assay (ELISA) may require 100 μL of sample. From a biological perspective, fluid loss can be undesirable in some tissues that are particularly sensitive, such as the brain. Furthermore, brain tissue is also highly sensitive to ionic concentration alterations since such changes can alter neurotransmitter release (17). By maintaining an osmotic balance, the fundamental mass transport mechanism for moving an analyte from the extracellular fluid space (ECF) to the dialysate lumen is principally diffusion.

Probe Geometry

Microdialysis sampling is typically considered to be synonymous with the term intracranial dialysis sampling because of its neuroscience origins. The first intracranial dialysis device was a linear design that traversed longitudinally through different brain regions. A variety of different probe designs have been described in many different review articles (18–20). Microdialysis probe design

has evolved to allow use of the device for biomedical applications beyond neuroscience. Linear geometry microdialysis sampling devices for neuroscience were not as useful as the push–pull cannula that could be inserted into known brain regions (e.g., striatum, hippocampus, *substantia nigra*) based on known stereotaxic coordinates that in some cases are < 1 mm wide in rat brain. To overcome this challenge for neurochemistry studies, more rigid cannula designs were created that can be inserted into specific brain regions and are now commercially available from a variety of sources (21).

As microdialysis sampling devices became a standard tool used by neuroscientists, researchers in other fields began to realize its great *in vivo* analysis potential. Principally, the use of the probes for collection of endogenous or xenobiotic components in blood and peripheral tissues became of interest (22). In these tissues, a rigid stainless steel cannula causes tissue damage and may make awake and freely moving experiments with animals quite difficult. Cannula designs using Teflon or fused silica are commonly used for to make flexible probes for either sampling in soft peripheral tissues (e.g., skin or liver) or for blood sampling. Linear probe designs have also been reintroduced after originally being applied to brain studies and are now used for insertion into soft peripheral tissues. An additional advantage of these flexible designs is they also allow for studies in awake and freely moving animals in peripheral tissues. Recent research interests in transgenic mice have forced the creation of smaller microdialysis sampling devices (23).

Probe Materials

Semipermeable hollow fiber membranes used for microdialysis sampling are the same as those used for kidney dialysis. Different polymeric semipermeable membranes have been used in microdialysis sampling probes and are listed in Table 2. Typical materials include cellulose-based membranes (cuprophan or cellulose acetate), polycarbonate/polyether blends, polyacrylonitrile, and polyethersulfone. These membranes span a wide range of molecular weight cutoffs (MWCO) from 5000 to 100,000 Da. Choice of the membrane to be used during microdialysis sampling requires both analyte molecular weight information, as well as where the probe will be implanted as some tissue regions (particularly in the brain) are too narrow for > 500 μm external diameter membranes.

Table 2. Commercially Available Microdialysis Membrane Dimensions[a]

	PC	PES	PAN	CUP
Outer radius, μm	250	250	170	120
Inner radius, μm	200	205	120	95
Wall thickness, μm	50	45	50	25
Molecular weight cutoff	20,000	100,000	29,000	6,000
Outer surface area, mm^2	6.28	6.28	4.27	3.01

[a]The data provided here is that given by the manufacturers of the microdialysis probes. It is not known if the radii are for dry or wet membranes. The abbreviations are as follows: PC = polyether/polycarbonate, PES = polyethersulfone, PAN = polyacrylonitrile (or AN-69), CUP = cuprophan. CMA Microdialysis, Inc sells PC, PES, and CUP membranes. Bioanalytical Systems, Inc sells PAN membrane probes.

Semipermeable hollow fiber dialysis membranes can be obtained with a known molecular weight cut off (MWCO). The MWCO can be experimentally determined for a hollow fiber using several different experimental methods. The primary method used to determine MWCO for hollow fiber membranes is to continuously pass through the fiber lumen over a long period of time (24 h or greater) a solution containing known molecular weight markers. Known solutes that are rejected by the membrane are then used to calculate membrane MWCO. In practice, the MWCO is really not an absolute number, but rather the median of a range. This molecular weight rejection range is highly dependent on the semipermeable membrane materials pore distribution and can exhibit either a narrow or broad MWCO range (24).

Originally, the purpose of microdialysis sampling was to use the dialysis membrane as a means to provide a sample for chemical analysis that did not require further sample preparation steps such as protein removal. Intracranical dialysis applications typically target hydrophilic analytes with molecular weights < 500 Da. For these applications, dialysis membranes with low MWCO of ~ 5000–6000 Da were commonly used to reject larger analytes and proteins so to allow liquid chromatographic analysis without further sample purification.

Recently, there has been a greater interest of applying microdialysis sampling to collect peptides and proteins. There have only been a few reports describing the use of different types of dialysis membranes towards the collection of large molecules, such as peptides and proteins (25). This is unfortunate as the types of commercially available membranes that are capable of providing the performance characteristics necessary for protein collection are relatively few. Kendrick extensively compared the recovery performance of different amino acids and peptides among different types of dialysis membranes (26). Torto *et al*. compared the dialysis collection efficiency for a series of saccharides [glucose (DP1), maltose (DP2), though maltoheptaose (D7)] among many different types of dialysis membranes (polyamide, polyethersulfone, and polysulfone) as well as different MWCO between 6 and 100 kDa (27). In some cases, membranes with similar MWCO and different chemistry exhibited similar recovery values. Whereas, some of the polysulfone membranes with 100 kDa MWCO exhibited quite low recovery for these low molecular weight analytes when compared to membranes with similar chemistry, but lower MWCO.

A set of model proteins including insulin (5.7 kDa), cytochrome *c* (12.4 kDa), ribonuclease A (13.7 kDa), lysozyme (14.4 kDa), and human serum albumin (67 kDa) were tested with different polymeric membranes and molecular weight cutoffs ranging between 20 and 150 kDa (28). All the membranes had similar external diameters (500 μm). Among the different membranes, only the polyethersulfone (100 kDa MWCO) commercially available from CMA Microdialysis, Inc and a Fresenius polysulfone membrane (150 kDa MWCO) exhibited similar recovery characteristics for the above-mentioned set of model proteins.

Membrane MWCO cannot be used as a means to specifically predict how well an analyte will be recovered during a microdialysis sampling procedures. New microdialysis sampling practitioners sometimes mistakenly believe that analytes near the membrane MWCO will be recovered. Although a membrane with 100 kDa MWCO allows some transport of molecules of this molecular weight, the recovery of an analyte of this size will be significantly $< 1\%$ (if at all) of the external sample concentration during microdialysis sampling. Dialysate analyte concentrations rarely reach equilibrium with the external sample concentrations except under unique conditions (very low flow rates and long membranes). For dialysate concentrations to reach those of the tissue medium surrounding the probe and thus approach equilibrium with the surrounding tissue concentrations, low perfusion fluid flow rates or long membranes are required in order to achieve residence times sufficient to obtain equilibration. During microdialysis sampling, the perfusion fluid only passes once through the inner membrane lumen with residence times on the order of seconds. Since this is in contrast to the methods used to obtain membrane MWCO, it is not surprising that sampled analyte molecular weight range is reduced due to the perfusion fluid making only one pass through the device. In general, the analyte molecular weight that easily passes through the membrane with 10% or greater recovery is roughly one tenth of the MWCO as shown in Fig. 2 (29). However, this is not an absolute value and different analytes have been reported to be difficult to dialyze across particular membranes. With rare exception (30), hydrophobic analytes typically are poorly dialyzed during microdialysis sampling (31,32). Furthermore, AN-69 membranes (polyacrylonitrile), which are sometimes used for kidney dialysis and have been used for in-house microdialysis probes, carry a negative charge that may cause rejection of certain negatively charged analytes (33,34).

Microdialysis sampling requires an inlet and outlet tube to be attached to the membrane. The length and inner diameter of the outlet tube attached to the membrane affects membrane backpressure. For some hollow fiber membranes, convective fluid loss (ultrafiltration) across

$$J_v = P(\Delta p - \Delta\pi)/l \tag{1}$$

these hollow fiber semipermeable membranes is possible and the extent of this ultrafiltration is related to the volumetric flux (J_v) shown in Eq. 1, where P is the permeability coefficient for the membrane, l is the length across the membrane (e.g., the membrane thickness), and Δp and $\Delta\pi$ the hydrostatic and osmotic pressure differences (35). Different membranes have difference permeability coefficients. The physical manifestation of this fluid loss is that the probe appears as if it is sweating during the dialysis procedure and is often observed with larger MWCO membranes.

Peptides and proteins diffuse very slowly across the small pores of membranes with MWCO between 5000 and 30,000 Da. To improve the relative recovery of these analytes, larger 100 kDa dialysis membranes have become commercially available. The disadvantage of these larger MWCO membranes is that they often exhibit ultrafiltration due to their larger pore sizes. Ultrafiltration fluid losses across 100 kDa or larger MWCO dialysis membranes should be determined prior to *in vivo* experiments. In particular, the ultrafiltration is exacerbated by the use of long outlet tubing with narrow diameter (a common need with awake and freely moving animal experiments). Osmotic balancing agents, such as dextrans or albumin, are commonly added to microdialysis perfusion fluids passed through 100 kDa or larger MWCO membranes to prevent excessive ultrafiltration as well as to prevent nonspecific adsorption on the device materials (36,37). While these agents are passed through the membrane, their potential loss to the surrounding tissue space has not been reported.

Recovery, Delivery, and Localized Infusion

Sensor devices are highly specific analytical detectors and can only be used to detect analytes that physically contact

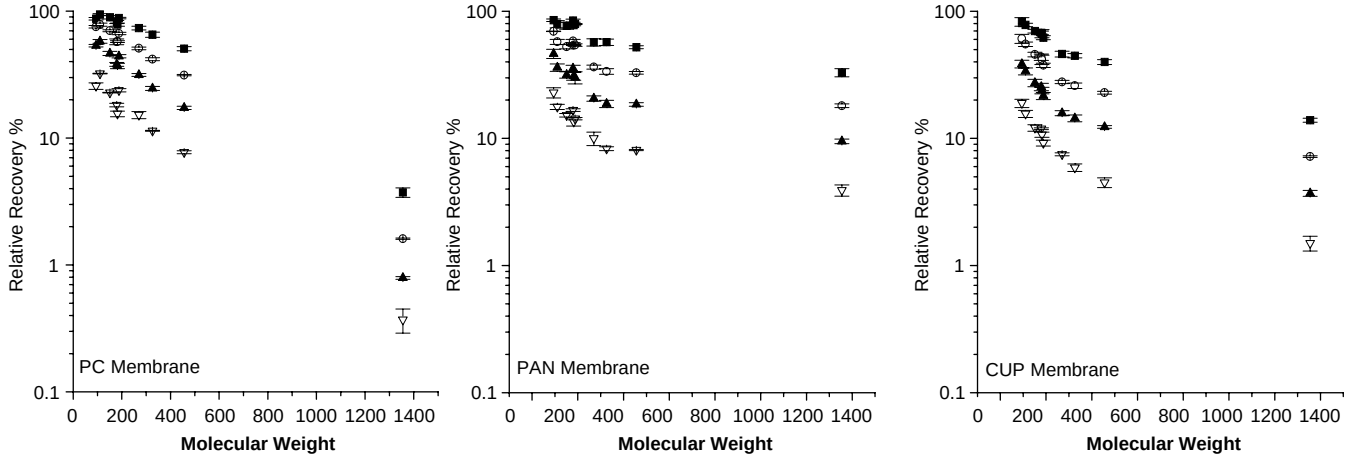


Figure 2. Semilog graph of relative recovery versus molecular weight for PC, PAN, and CUP membranes using different perfusion fluid flow rates. Flow rates are 0.5 $\mu L \cdot min^{-1}$ (■), 1.0 $\mu L \cdot min^{-1}$ (○), 2.0 $\mu L \cdot min^{-1}$ (▲) and 5.0 $\mu L \cdot min^{-1}$ (▲). Adapted from Ref. 29).

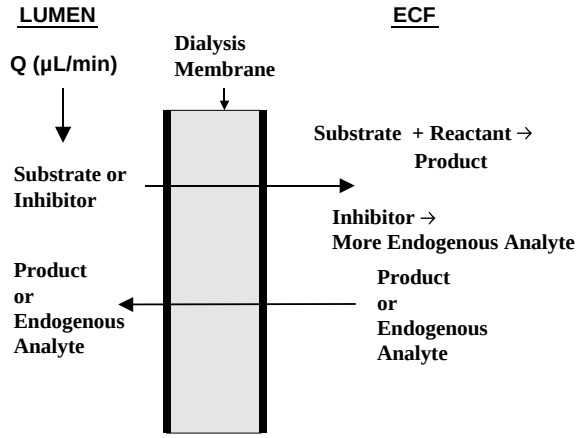


Figure 3. Localized infusion. A substrate or drug is locally infused through the microdialysis sampling probe. It diffuses into the ECF and then either reacts to form a product or causes a biochemical event to increase the concentrations of other molecules.

the sensor. Microdialysis sampling provides extensive flexibility with respect to ways in which it can be applied. Typical microdialysis sampling applications use what is termed the recovery mode where the device is placed into a sample matrix and analytes diffuse into the inner-fiber lumen of the probe (Fig. 1). Alternatively, the device can be used as a delivery device where a compound is infused through the probe causing either some alteration in a biochemical event (enzymatic reaction, enzymatic inhibition, or receptor binding) followed by collection of an endogenous analyte or enzymatic product. The probe can also be used to pass a substrate for a chemical or biochemical reaction and the products of that reaction can then be locally sampled as shown in Fig. 3. Unlike a specific sensor, microdialysis sampling devices allow for many physiological and biochemical processes to be studied within living tissue. Different examples of this approach for a variety of different tissues are shown in Table 3.

DEVICE CALIBRATION

Theoretical Foundations

Typical microdialysis sampling operating conditions (flow rates between 0.5 and 2.0 $\mu L \cdot min^{-1}$) will yield a representative fraction of the analyte concentration in the surrounding ECF. Since most microdialysis conditions are such that equilibrium between the dialysate and the sample is not obtained, a calibration has to be used to relate dialysis concentrations to external sample concentrations. Extraction efficiency (E_d) is used to relate the dialysis concentration to the sample concentration. The steady-state E_d equation is

$$E_d = \frac{C_{\text{outlet}} - C_{\text{inlet}}}{C_{\text{tissue},\infty} - C_{\text{inlet}}} = 1 - \exp\frac{-1}{Q_{\text{d}}(R_{\text{d}} + R_{\text{m}} + R_{\text{e}} + R_{\text{t}})} \quad (2)$$

shown below in Eq. 2, where C_{outlet} is the analyte concentration exiting the microdialysis probe, C_{inlet} is the analyte concentration entering the microdialysis probe, $C_{\text{tissue},\infty}$ is the analyte tissue concentration far away from the probe, Q_{d} is the perfusion fluid flow rate and R_{d}, R_{m}, R_{e}, and R_{t} are a series of mass transport resistances for the dialysate, membrane, external sample, and a trauma layer that exists at the interface of the probe membrane and the tissue as defined in Scheme 1 (48,49).

The resistance terms are additive and understanding how these resistance terms affect E_{d} is vitally important with respect to experimental design. Throughout the microdialysis sampling process, collected analytes must diffuse through at least three regions (tissue, membrane and dialysate) in order to exit the microdialysis probe. Each mass transport resistance term defined in Scheme 1 has a diffusive component, which indicates that analytes with smaller diffusion coefficients will exhibit much lower E_{d}. The combined resistance contributions from R_{d} and R_{m} can be experimentally determined *in vitro* by collecting dialysates at different flow rates. A plot of the natural log of $(1\text{-}E_{\text{d}})$ versus $1/Q_{\text{d}}$ should yield a straight line, which can be regressed to determine the additive values for R_{d} and R_{m}. In addition to this information, an *in vitro* E_{d} experiment performed at 37 °C with stirring to cause the sample resistance, R_{e}, to approach a zero value, will yield the highest possible *in vivo* E_{d}.

The variables shown in Eq. 2 illustrate that a combination of perfusion fluid flow rate (Q_{d}), as well as mass transport resistances for the dialysate, membrane, and tissue medium external to the microdialysis probe affect E_{d}. Decreasing Q_{d} allows for a greater fluid residence time within the dialysis membrane thus allowing analyte concentration to increase along the membrane axis.

Table 3. Some Examples of Localized Infusion Using Microdialysis Sampling

Type (substrate or inhibitor)	Infused Compound	Measured Analyte or Application	Tissue	References
Inhibitor	Cocaine	Dopamine	Brain	38
Substrate	Substance P and other neuropeptides	Proteolytic Products	Brain	39,40
Substrate	Salicylic acid or 4-hydroxybenzoic acid	2,3-DHBA, 2,5-DHBA, 3,4-DHBA	*In Vitro*, Brain	41–43
Substrate	Phenol or acetaminophen	Metabolites	Liver	44,45
Substrate and inhibitor	Angiotensin, phosphoramidon, captopril	Metabolites and enzymatic inhibition	Renal Cortex	46
Substrate	Suc-$(Ala)_3$-pNA	Elastase (protease) activity	*In Vitro*	47

$$R_d = \frac{13(r_i - r_\alpha)}{70\pi L r_i D_d} \quad ; R_m = \frac{\ln(r_o / r_i)}{2\pi L D_m \phi_m}; R_e = \frac{\Gamma[K_o(r_o/\Gamma)/K_1(r_o/\Gamma)]}{2\pi r_o L D_s \phi_s}$$

$$\Gamma = \sqrt{\frac{D_s}{(k_{ep}(r) + k_m(r) + k_c(r))}}$$

Scheme 1. The multiple mass transport equations used to describe microdialysis sampling. D is the diffusion coefficient through the dialysate, D_d, membrane, D_m, and sample, D_s. The parameter L is the membrane length; Γ (cm) is a composite function; $k_{ep}(r)$, $k_m(r)$, and $k_c(r)$ are kinetic rate constants as a function of radial position (r) from the microdialysis probe. Additional term definitions can be found in Ref. 48.

Fig. 4 shows a typical E_d curve simulated using the above equations (only R_d and R_m assuming a well-stirred system) for analytes with different aqueous diffusion coefficients. Fig. 4 clearly shows how microdialysis sampling membranes even for a hypothetical case perform in a manner that is consistent with diffusion being the major contributor affecting recovery. This scenario for the diffusivity is especially true for protein collection as many proteins of interest such as the cytokines have molecular weight values that can begin to approach the molecular weight cutoff limit for the dialysis membrane. In these cases, the protein diameter can begin to approach the values for the pore diameters resulting in restricted diffusion through the membrane, higher membrane mass transport resistances and thus reduced analyte recovery.

The parameter E_d is highly dependent on several physiochemical parameters (analyte diffusion coefficient, perfusion fluid flow rate, membrane pore size, and membrane surface area), kinetic uptake into cells (50) and the microvasculature (51), as well as the overall ECF volume fraction. The mass transport, resistances shown in Scheme 1 include analyte diffusivity terms for all three regions of mass transport, as well as kinetic terms for the tissue space, as shown in the above equations. Tissue diffusive and kinetic properties of the sample surrounding an implanted microdialysis probe will dictate the how reduced the *in vivo* E_d will be from the maximum possible *in vitro* E_d value at any particular flow rate. For hydrophilic analytes, it is generally assumed they diffuse only in the ECF that surrounds the tissue cellular components. This ECF space comprises approximately 20% of the overall tissue volume (52). Typically hydrophilic analytes have to diffuse around the cells en route to the microdialysis probe, the overall effective diffusive path length is increased due to the tortuous path traversed by the analyte. This tortuosity alters the tissue diffusion coefficient which can be approximated using $D_{ecf} = D_{aq}/\lambda^2$, where λ has a value of $\sim$ 1.5. In addition to the alteration in diffusive characteristics, tissues are vascularized and have active cellular components. Depending on the analyte, the active components will affect the overall microdialysis E_d.

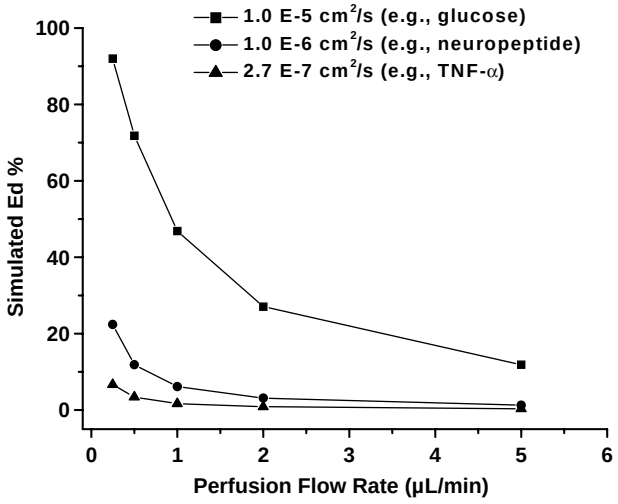


Figure 4. Simulated E_d using Bungay et al. (48) model with the following aqueous diffusion (D_{aq}) coefficients and with a membrane diffusion value of $0.2D_{aq}$ with length of 10 mm and R_i (200 μm) and R_o (250 μm).

In addition to these parameters influencing microdialysis E_d, analyte properties affect the shape and the time to reach steady state for the concentration profile to the microdialysis probe. Analytes that diffuse rapidly and have a rapid supply to the tissue have narrow concentration profiles to the dialysis probe. Conversely, analytes that slowly diffuse and are not readily supplied to the tissue space will have concentration profiles to the dialysis probe that are not as steep. During microdialysis material is removed from the sampling site and the extent to which matter is removed is a function diffusive and kinetic parameters applied to that particular analyte, for example, how rapidly it the analyte replenished to the ECF from either capillaries (drugs) or cellular release processes. This has been a concern by others particularly as it relates to the understanding of dopamine transmission in the brain (53). However, dopamine is a special case and has rapid release and uptake kinetics in the ECF. For analytes that are poorly transported across the capillary space in the brain along with analytes that do not undergo significant uptake (e.g., drugs), their relative recoveries are generally lower than those with higher uptake/kinetic rates. It may appear to be counterintuitive to think that analytes with very rapid kinetic removal from the space surrounding the microdialysis probe have increased relative recovery. However, the higher removal rates cause the concentration profile to the dialysis probe to have a much greater gradient to the device as compared to a poorly removed analyte which would have a much shallower concentration profile to the device. For analytes with similar ability to diffuse through the membrane, that is, their membrane diffusion coefficients are nearly equal, the flux should be greater for the sharper concentration gradient thus causing greater relative recovery.

The theoretical foundations for microdialysis sampling during steady state operations derived by Bungay et al. have been widely used to corroborate many different *in vivo* experimental observations. In a series of papers focused on neurotransmitters, Justice's group has studied how uptake inhibition decreases microdialysis E_d 38,54,55. Stenken et al. 56 showed that since kinetic removal of targeted analyte may in some cases be additive, the inhibition of a particular cytochrome P450 isoform for phenacetin and

antipyrine metabolism did not significantly alter E_d. This suggested that multiple kinetic components (capillary permeability plus metabolism) are important for analyte removal from liver tissue. Elmquist has shown that transporter inhibition in brain causes alterations in E_d for different drugs (57).

Calibration Methods

If the different variables shown in Eq. 2 and Scheme 1 are known prior to experimentation, then it is possible to predict the *in vivo* E_d. However, the difficulty with *in vivo* experiments is that obtaining values for variables is highly challenging. Furthermore, it is difficult to obtain an absolute value for C_{tissue} and an *in vivo* calibration requires that C_{tissue} (Eq. 2) be known. Even today, many experiments employing microdialysis sampling are semiquantitative and dialysate concentrations obtained are approximations of C_{tissue} at best. Since quite often ratios of dialysate analyte concentrations are obtained before and after some input (e.g., pharmacological or physical) with no attempt to measure the C_{tissue} during the experiment. To date, the *in vivo* calibration of implanted devices, including microdialysis probes, is an active area of research in the analytical chemistry and bioengineering communities and many authors have reviewed this subject (58,59). The principal difficulty with respect to obtaining a reliable device calibration is the inability to fully reproduce *in vitro* all the salient physiological features of tissue including permeation across capillaries and uptake processes (60,61).

Initial microdialysis sampling calibration focused on using an *in vitro* E_d calculation to estimate tissue analyte concentration. This approach gives a rough estimate of E_d and may provide an incorrect calibration factor. Furthermore, *in vitro* methods used for E_d measurement are affected by temperature (microdialysis sampling again is inherently a diffusion separation method), as well as sample stirring. A well-stirred buffer medium provides a mass transport external medium mass transport resistance (R_e) that approaches a value of zero. It is important to note that a quiescent medium does provide diffusional mass transport resistance and thus relative recoveries performed *in vitro* under stirred conditions will be different than those performed using quiescent conditions (48). How close a quiescently determined *in vitro* E_d is to the *in vivo* E_d would be wholly dependent upon the tissue kinetic properties for the targeted analyte. In other words, an analyte, such as dopamine, may exhibit higher *in vivo* E_d than *in vitro* quiescent E_d due to its extensive uptake kinetics causing a steeper concentration gradient to the dialysis probe as compared to the *in vitro* quiescent E_d measurement. Differences in the ability of the analyte to diffuse through the tissue space due to increased tortuosity and decreased volume fraction led to empirical methods that could be used to amend *in vitro* relative recovery calibration determinations (62). These methods focused on differences in tissue diffusion properties, but did not include the role of kinetic affects on microdialysis E_d causing significant errors for estimating *in vivo* values for C_{tissue}.

Jacobson et al. (63) were the first to try to create a more analyte-specific calibration procedure for microdialysis sampling. In their work, varying the perfusion fluid flow rates through the dialysis probe derived an analyte-specific membrane mass transport coefficient, K, shown below in Eq. 3, where A is the membrane surface area and Q_d is the dialysate volumetric flow rate. Eq. 3

$$\frac{C_{outlet}}{C_{tissue}} = 1 - \exp(-KA/Q_d) \tag{3}$$

is similar to Eq. 2, showing how the model of Bungay et al. incorporated previously known experimental results. In this case, the product (-KA) is related to the sum of the fraction of the mass transport resistance terms. Experimental results from this work immediately showed that understanding the underlying *in vivo* mechanisms affecting microdialysis E_d was more complicated than initially expected. These researchers found that different amino acids exhibited different *in vivo* mass transport coefficients. This was unexpected since the amino acids would be expected to have very similar diffusion coefficients due to their similar molecular weight. This data began to lead to the understanding that analyte properties (diffusion and kinetics) in the tissue play a major role with respect to microdialysis sampling calibration. An extension of calibration approach of Jacobson et al. is to pass the perfusion fluid through the dialysis probe so slowly that zero flow is approached and nearly 100% relative recovery as shown in Fig. 5. In this case, the goal is to calculate C_{sample} by attempting to reach an equilibrium state across the microdialysis membrane (64).

The most widely used calibration method for microdialysis sampling is based on knowing that diffusive flux should not occur across the dialysis membrane when the analyte concentration inside the perfusion fluid matches the concentration external to the microdialysis probe. This method was originally demonstrated by Lönnroth and has been called by a variety of names including Lönnroth plot,

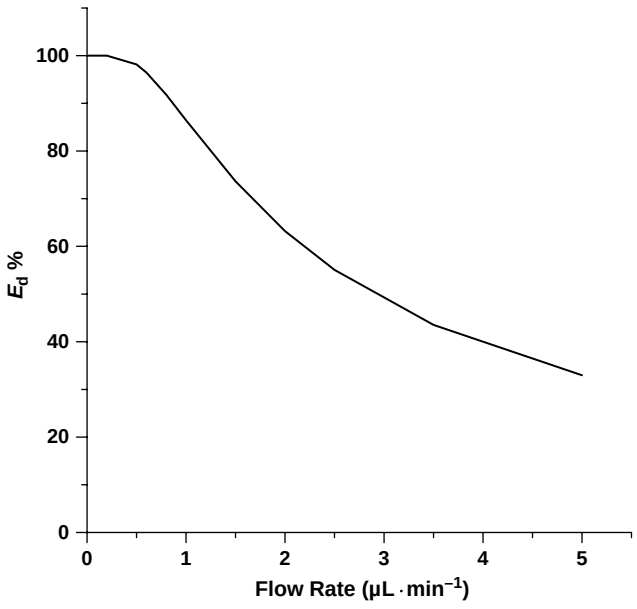


Figure 5. Mathematically modeled E_d for an approach to zero flow.

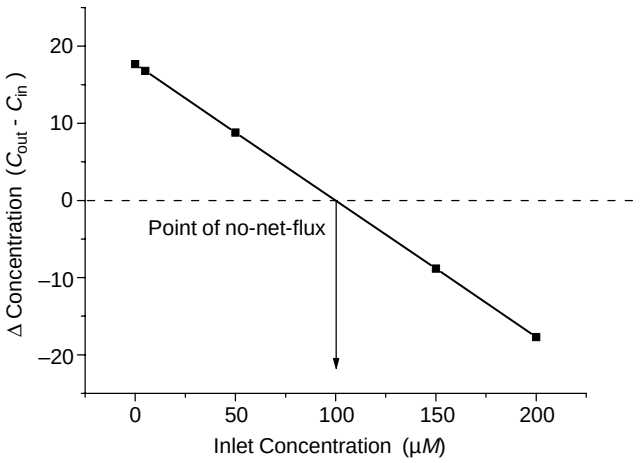


Figure 6. Lonnroth plot using hypothetical data.

method of no net flux (NNF) and method of zero net flux (ZNF). The ZNF method requires the tissue analyte concentration be at a steady state and has been used to determine basal concentrations for many analytes. The probe is perfused with different analyte concentrations that are either above or below the targeted analyte concentration. With these different perfusion studies, the loss or gain of analyte across the microdialysis probe can be determined and then plotted versus the inlet concentration as shown in Fig. 6. The analyte concentration is determined by the *x*-axis intercept and the relative recovery is the absolute value of the calibration line slope. A major drawback with this approach is the extensive amount of research time that has to be invested during the different perfusion studies. In particular, this is quite difficult to achieve with exogenous analytes (drugs) that would require a continuous infusion to achieve steady-state concentrations.

Quite often what is desired from a microdialysis sampling procedure is the analyte concentration during some sort of pharmacological challenge to an animal. This much needed analyte temporal information is not possible to collect with the requirement of steady state for the ZNF method. To overcome this problem and to gain information regarding the probe calibration from sample to sample, internal standards have been used during microdialysis sampling. Typically, internal standards have been chemicals with similar physicochemical properties as the targeted analyte. Some early work in internal standards proposed using one standard, such as antipyrine or ^{3}H–water to allow assessment of sample-to-sample differences should they arise throughout the duration of the microdialysis sampling events (65). Antipyrine would be a suitable reference for probe-to-probe variability because of its highly hydrophilic nature and ease of chemical detection. Urea has also been used as a microdialysis calibration reference for different metabolism studies (66–68). It is again important to note that the extraction efficiency of any particular analyte is a combined function of the different mass transport regions : dialysate, membrane, and most importantly tissue. Most likely during an *in vivo* microdialysis sampling experiment, the variability from sample to sample will occur due to alterations in tissue physiology, such as blood flow, metabolism, and uptake, which would serve to alter the tissue resistance and thus E_d. For this reason, it is generally preferred to have internal standards with similar tissue diffusion and kinetic properties as the analyte. Finding an appropriate internal standard is not a trivial task since analytes with similar structure and diffusive properties may also compete for enzymatic sites and may inhibit metabolic pathways that are important to removal and thus E_d values. However, this possibility must be considered in context of the tissue being sampled, as well as other additive kinetic properties (e.g., uptake into cells or capillaries) that may have a much greater impact on the E_d. For example, in the brain, the kinetic process that has been shown for several neurotransmitters to be most weighted towards affecting E_d are the neuronal uptake processes rather than metabolism processes. Additionally, in the liver, it appears that capillary blood flow and permeability are the primary contributors toward the E_d value obtained. Internal standards for peptides and proteins may be much harder to devise as receptor binding or for the cytokines, binding to the proteoglycan components of the ECF space may affect E_d and thus using molecular weight markers, such as inulin (69,70) or higher molecular weight fluorescein-labeled dextrans (e.g., FITC-Dextran 3000, FITC-Dextran 10,000) may only serve to report back diffusional mass transport differences during the duration of microdialysis sampling.

Effect of Probe Insertion Trauma

Insertion of microdialysis probes causes tissue damage (71,72). Although this has been known for quite some time, it has generally been overlooked by many microdialysis sampling users. The extent to which this insertion trauma affects the integrity of the microdialysis sampling concentrations and its true overall importance has been debated in the literature. The biomaterials literature is full of descriptions of the cellular events that occur after a foreign body implantation (73). It is known that edema occurs at the site of probe implantation (74) along with the recruitment of polymorphonuclear leukocytes (75,76) and matrix metalloproteinases (extracellular matrix remodeling enzymes) (77). Moderately reduced analyte flux to microdialysis probe chronically implanted has been reported for glucose (78).

The validity of the ZNF calibration methods for *in vivo* calibration, as well as determination of C_{tissue} for some analytes, has recently been a concern for neuroscientists interested in dopamine. Many careful studies performed by Michael's group illustrated that dopamine concentration measurements obtained with microelectrodes and microdialysis sampling devices were quite different (79,80). In particular, microdialysis sampling devices often exhibited much lower basal concentrations of dopamine than microelectrodes. Additional concerns have been raised for drug blood–brain barrier studies (81,82). Between these two examples, dopamine collection via microdialysis sampling appears to be the most severely affected because of its release and uptake sites being compromised due to the

insertion trauma (49,83,84). In essence, the creation of a trauma layer creates four separate mass transport regions during microdialysis sampling : the dialysate, membrane, trauma layer, and normal tissue that need to be accounted for during data interpretation.

ANALYSIS OF MICRODIALYSIS SAMPLES

Microdialysis sampling is essentially married to appropriate detection methods for the collected dialysates. In addition to providing a means to sample from an *in vivo* site, microdialysis sampling also provides a relatively protein free or clean sample for chemical analysis. The only selectivity imparted into a microdialysis membrane is its molecular weight cutoff. For this reason, as long as a targeted analyte can diffuse through the membrane, the microdialysis sampling probe can be used as an *in vivo* chemical collection device. Thus, assuming the targeted analyte can pass through the dialysis membrane pores coupled with the appropriate analytical methods, a microdialysis sampling device could be considered to be essentially an all-purpose *in vivo* sensor (85). There is an extensive literature that has reviewed the associated analytical chemistry for making measurements in microdialysis samples (86). Additional reviews include: Adell et al. (87), Chaurasia (88), Church and Justice (89), Davies and Lunte (90), Horn (91), Kennedy (92), Kennedy et al. (93). Lunte et al. (94), Lunte and Lunte (95), Obrenovitch (96), Parkin et al. (97), and Parrot et al. (98).

Sample Volume Limitations

The major bottleneck 25 years ago for microdialysis sampling gaining more wide-spread and universal acceptance had to do with the analytical detection method sample volume limitations. During the early stages of microdialysis sampling, the primary analytical detection methods used for analyte quantification were liquid chromatography (LC) coupled with various types of detectors [ultraviolet–visible (UV–Vis), fluorescence, and electrochemical]. In addition to LC methods, radioimmunoassay (RIA) was occasionally used for peptides and proteins. Twenty-five years ago, it was not uncommon to require 25–50 μL of sample for LC analyses. Today, 50–100 μL of sample is still needed for standard immunoassays. The trade off that had to occur became one of either obtaining higher concentration recovery across the membrane by using low perfusion flow rates (1 $\mu L \cdot min^{-1}$ or less) or gaining sufficient temporal resolution by going to faster flow rates to achieve sufficient sample volumes for chemical analysis.

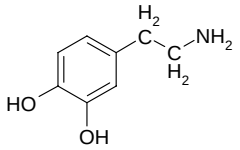
With the exception of glucose and lactate, many of the endogenous as well as xenobiotic analytes sampled using microdialysis had either micromolar (μM; 10^{-6} M) to nanomolar (nM 10^{-9} M) concentrations. These low concentrations often pushed the limitations of common analytical equipment since for most analytes an approximate detection limit with most UV-Vis detectors is roughly in the low μM range and for fluorescence and electrochemical detectors their detection limits are approximately in the nM range. The need to be able perform analytical measurements from such low volume dialysates drove analytical method development in multiple directions towards systems that could accommodate the low volumes without sacrificing method sensitivity, as well as development of high throughput methods that allowed for increased temporal resolution. Presently, there are many commercially available technologies that allow for samples that are <1 μL (e.g., capillary electrophoresis) or have duty cycles that are <1 min.

Separations-Based Methods for Microdialysis Sample Quantitation

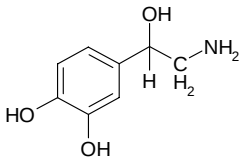
Using separation methods, such as LC or capillary electrophoresis, for the quantitation of microdialysis samples is highly advantageous since these methods can be quickly adapted to many different analytes. Before the extensive use of microdialysis sampling for studies of neurochemical transmission, the use of *in vivo* voltammetry for analysis of neurotransmitters was just beginning to be described as a method for catecholamine (dopamine and norepinephrine) (99,100). The difficulty with using these methods was that electrode potentials needed to oxidize the catecholamines, as well as their metabolites (3, 4-dihydroxyphenylacetic acid, DOPAC) were similar. Furthermore, it was soon discovered that during vesicular release of dopamine, very high concentrations of ascorbic acid were released (101). For this reason, *in vivo* voltammetry of these important neurochemicals became more challenging since all of these chemicals can be oxidized at or below the same potential. The advantage of separations methods with appropriate detectors is that components including targeted analytes, as well as endogenous and exogenous interferences (see Fig. 7) can be appropriately separated and quantified. Thus, an additional advantage of using chromatographic methods is that chromatographic methods provide intrinsic multiplexing capabilities for the chemical analysis of microdialysis samples if different analytes are expected in the same samples.

Liquid Chromatography. Liquid chromatographic methods have been used for analyzing a broad class of analytes from microdialysis samples including catecholamines, amino acids, pharmaceuticals, and their metabolites. Several articles are available that describe the necessary requirements for microdialysis sample analysis using LC (102,103). Liquid chromatographic separations methods are well suited to microdialysis samples because of the high salt content contained in the perfusion fluids. Salts are generally not retained by the LC stationary phase and are therefore eluted in the chromatographic void volume. The resolving power of LC stationary phases allows for multiple analytes to be quantified during a single chromatographic run. Different detectors have been applied to LC separations for quantitation of dialysis samples.

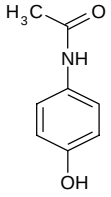
Capillary Electrophoresis. Capillary electrophoresis (CE) is a separation method that involves passing an electric field across a micron-sized (~25 to 75 μm internal diameter) capillary so as to allow separation of analytes based on their additive electrophoretic and electrosmotic mobilities. Neutral components in capillary electrophoresis



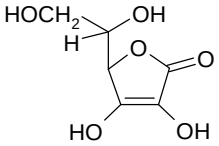
DOPAMINE



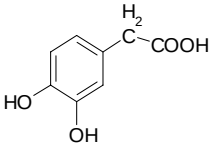
NOREPINEPHRINE



ACETAMINOPHEN (PARACETAMOL)



ASCORBIC ACID



3,4-DIHYDROXYPHENYLACETIC ACID (DOPAC)

Figure 7. Example of different analytes that can be oxidized at approximately the same potential using a carbon electrode under physiological conditions. The formal potential for dopamine, norepinephrine, and DOPAC are $\sim$ 0.7 V versus Ag/AgCl. For acetominophen, the formal potential is $\sim +0.8$ V versus Ag/AgCl. For ascorbic acid, the oxidation formal potential is $\sim +$ 0.15 V versus Ag/AgCl.

will elute based on their electroosmotic mobility. Compared to LC analysis, CE provides greater separation efficiency and the possibility of faster separations. An additional advantage of CE is that an enormous research effort has been placed into microfabrication of CE devices onto microchips. This suggests the possibility for point-of-care technologies to be integrated with microdialysis sampling.

Like LC, CE has been used for a wide range of microdialysis sampling analyses including catecholamine neurotransmitters, amino acid neurotransmitters, and pharmaceutical compounds. Capillary electrophoresis has one main disadvantage and that has to do with the poor UV detection limits due to the path length being significantly reduced. However, sensitive detection can be achieved using electrochemical or laser-induced fluorescence detection approaches.

Although there are several disadvantages with CE detection, that is, no standard equipment, requirement of expertise, there is one advantage to this detection method for microdialysis samples. By using pH-stacking methods, significant on-column preconcentration (100 or greater) can be achieved (104). The mismatch of pH serves to greatly concentrate the sample zone on the head of the capillary column. This allows for online preconcentration to occur during the CE experiment. For collection and detection of peptides or proteins this preconcentration can serve to be highly useful. An additional advantage of CE is that extremely fast separations on the order of seconds can be achieved making this technology similar to that of a separations-based biosensor (105).

Fast Separations. Microdialysis sampling can be a slow temporal process due to the need to collect sufficient sample with sufficient relative recovery. To make microdialysis sampling more sensor like in terms of its response time, a number of groups have worked on achieving high speed separations, as well as direct coupling dialysate outflow the detection method. This is particularly important for studies of neurotransmitter dynamics. While electrochemical approaches for studies of catecholamine neurotransmission provide millisecond time resolution, microdialysis sampling is hindered by the turn-around time for the analytical method. High speed detection capability of cycle times of $<$1 min have been reported by many different groups using both liquid chromatographic, as well as capillary electrophoresis separations including 1 s time resolution with neurotransmitters (106) and $<$1 min resolution with pharmacokinetic analyses (107).

Examples of Different Types of Detection

Electrochemical Detection. Liquid chromatography coupled with electrochemical detection (LC–EC) is both a highly sensitive and selective method for analysis of compounds that can undergo an electrochemical reaction (oxidation or reduction). In this sense, LC–EC is highly suited to the chemical analysis of important biogenic amines (dopamine, norepinephrine, etc) obtained from microdialysis probes implanted into the brain (108). For these measurements, the electrochemical detector has a glassy carbon electrode that allows for oxidation of the amines at potentials of roughly 700 mV versus a Ag/AgCl reference electrode. The LC–EC excels in this analysis task because these analytes can be oxidized and furthermore their basal concentrations are in the low to mid-nanomolar range. The advantage of the separation is evident when considering that catecholamine metabolites (DOPAC, HVA) have basal concentrations that are $\sim$ 100–1000 times greater than the clinically relevant catecholamines (dopamine, norepinephrine, serotonin).

In addition to catecholamine detection of microdialysis samples, LC–EC analysis has been applied to low concentration analytes obtained under varying conditions of oxidative stress. In these cases, typically either salicylic acid or 4-hydroxybenzoic acid is directly infused through the microdialysis probe to locally deliver a trapping agent as shown in Fig. 3 (109). These benzoic acids react with hydroxyl radical to form catechols which can then be separated and detected by LC–EC (110). The LC–EC has also been used to quantify the DNA oxidative damage biomarker, 8-dOH–dGuanosine (111,112).

Electrochemical detection can be made to be more selective by altering the electrode surface. Gold electrodes coated with Hg to create an amalgam are highly selective toward thiols such as cysteine and glutathione with low potentials needed for oxidation $\sim$ 150 mV versus Ag/AgCl (113). Lunte and O'Shea used this approach for glutathione detection using CE (114).

In addition to electrode modification, packed enzyme beds containing specific oxidase can be used prior to electrochemical detection. This is commonly applied to detection of the neurotransmitters choline and acetylcholine. Acetylcholine and choline can be separated chromatographically and then an enzymatic bed containing acetylcholine oxidase and choline oxidase is placed at the end of the column. These specific enzymatic reactions produce hydrogen peroxide which is then detected downstream at a platinum electrode (115). Note that more recent developments have attempted to immobilize the enzymes specifically to the electrode (116).

Examples of Fluorescence for Dialysates. Fluorescence detection is often employed when a known derivatization method can be applied to dialysate samples to improve method detection limits or to create a molecule that has a better handle for detection. For microdialysis samples, fluorescence derivatization is commonly applied to important amino acid neurotransmitters such as glutamate and GABA (117–119) and occasionally to biogenic amines (e.g., dopamine and norepinephrine) (120).

Examples of Mass Spectrometry for Dialysates. Mass spectrometry (MS) is a unique LC detector in that as long as the analyte has the ability to form an ion, MS can be used for analysis. However, mass spectrometric detection can be difficult with microdialysis samples because of the high salt content. This method has been particularly useful with neuropeptides because of their low concentrations. A problem with mass spectrometric detection is that salts from the dialysis perfusion fluid can cause analyte ionization suppression that leads to dramatically decreased detection capability for the method (121). Salts from dialysates are often removed via a column-switching technique that preconcentrates the analyte onto a C18 phase followed by desorption and detection (122). However, the use of nanoelectrospray devices can also reduce some of the problems associated with salt adducts (123–125). A particularly powerful method of LC-MS has been the ability to perform ionization in stages, which allows for structural elucidation of unknowns. The Kennedy group has been particularly successful with this approach for sequencing neuropeptides obtained from microdialysis samples (39,126).

Sensor Attachment to Microdialysis Probes

The microdialysis sampling process results in a relatively analytically clean (little to no protein) sample. Separations methods provide extensive analysis flexibility because they can be quickly optimized to the targeted analytes. However, there are *in vivo* monitoring situations where only one or a few analytes are targeted and highly specific analysis methods are available. A particular case in point is the continuous detection of glucose or lactate from diabetic humans (127). The primary advantage of coupling a sensing device to the end of the microdialysis sampling device is the sample matrix is simply saline passing across the analytical sensor. This prevents many of the difficulties associated with biofouling of implanted sensors (128). However, a critical problem for glucose sensing using this approach is that it can only provide information regarding the glucose concentration fluctuations throughout the day, but cannot really serve as an alarm system because of the lag times that are $\sim$20–30 min as compared to normal glucose sensors of a few minutes (129). Despite this concern, there is great value in using specialized sensors to a microdialysis device because of the clean sample delivered to the sensing device.

The use of biosensors attached to the end of microdialysis probes has become highly useful for clinical neuroscience where measuring glucose and lactate and in some cases other neurotransmitters are needed to understand homeostatic mechanisms (130,131). Most biosensors attached to dialysis probes have been for glucose, lactate, or glutamate detection (132–134). Cook has published an interesting approach combining immunoassay with electrochemical detection for specific measurements of coritsol (135).

Immunoassay for Peptide and Protein Detection

Peptide and protein analysis of microdialysis samples is challenging since the concentrations of these targeted analytes are often in the $ng \cdot mL^{-1}$ to $pg \cdot mL^{-1}$ levels. This requires either highly sensitive fluorescence derivatization techniques for use with capillary electrophoresis (136) or sensitive immunoassays. Conventional immunoassays require 50–100 μL of sample. To obtain these volumes requires the use of high flow rates (2 $\mu L \cdot min^{-1}$ or greater) or very long collection times. In most cases, because highly sensitive radioimmunoassays (RIA) are used, higher flow rates are used to achieve moderate temporal resolution.

It is becoming increasingly evident that cellular communication in biological systems is highly complex and networked. Despite the tremendous growth in microdialysis sampling to monitor cellular biochemistry and an increased interest in peptide and protein detection in dialysates, there has been relatively little research towards new analytical methods that can detect peptides and proteins in low volume dialysate samples. A few approaches have been published that require 80–100 μL of sample for detection of several different proteins (137,128). Multiplexed assays (up to 25 analytes or more) that can be performed on a single sample have been recently created for immunology studies of the important inflammatory mediator class of cytokine proteins.

Highly sensitive multiplexed immunoassay platforms based on particle-based flow cytometry has become commercially available that allows cytokine measurements in 50-μL sample volumes (139,140). The limit of detection for these assays fall into the low pg/mL range comparable and have been compared and validated against standard ELISA methods (141). The use of these particle-based

immunoassays is highly advantageous to the sample-limited microdialysis process and the sample volume needed has been decreased to < 25 μL by our group for cytokine detection. The advantage of these bead-based immunoassays for microdialysis samples is that several analytes can be analyzed in a single low volume sample. To illustrate the significant advantage that the bead-based immunoassay provides, if six separate cytokines were to be quantified in microdialysis samples using standard ELISA techniques more than 600 μL of sample would be needed. Using a flow rate of 1 $\mu L \cdot min^{-1}$, this would require 10 h of microdialysis sampling.

Mass versus Concentration Recovery

Microdialysis sampling E_d is a concentration recovery term and E_d increases as fluid flows decrease through the dialysis fiber creating longer residence times. Conversely, overall mass recovery typically increases as the flow rate increases as shown in Fig. 8. For some analytical applications, this increase in mass recovery may prove to be highly beneficial since it opens up possibilities for analytical preconcentration methods for the increased dialysate volumes.

MICRODIALYSIS SAMPLING APPLICATIONS

Microdialysis sampling applications have now been widely used in many different mammalian species including humans. The applications in humans have included studies in cancer, dermatology, immunology, pharmacokinetics and neuroscience. Many of these applications have been extensively reviewed by others and therefore will not be extensively discussed here. It is again important to note that microdialysis sampling has to date been principally applied to applications in neuroscience for the past three decades. However, over the past decade, more microdialysis sampling applications in other areas are now being described.

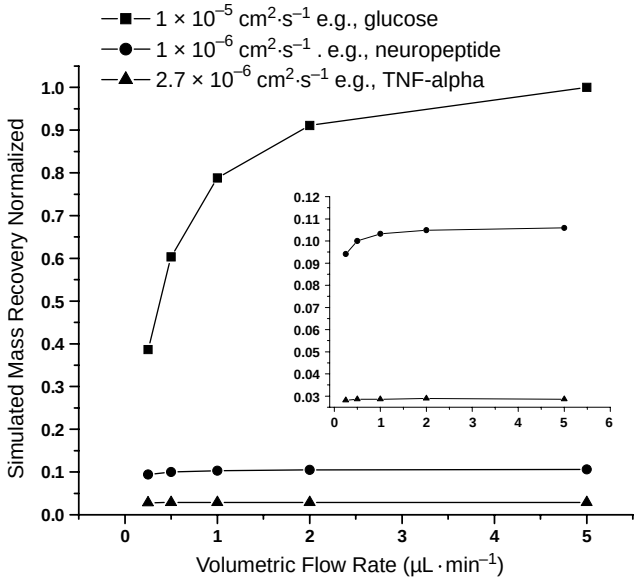


Figure 8. Modeled microdialysis sampling mass recovery for different analytes with different diffusion coefficients.

Neuroscience Applications

Microdialysis sampling has been in the neuroscientist toolbox for > 25 years. This device has been principally applied to neurotransmitter collection. Early on, the primary focus was in rat and more recently with probe redesigns and the biomedical value of knockouts, additional studies have been performed in mice (23,142). Current research interests focus on bridging the gap between animal models and human studies.

Reviewing all the microdialysis literature for neuroscience applications is a daunting task since the microdialysis sampling technique is now in wide use. Some of the applications have already been mentioned in the Analysis section of this article. However, several reviews and a book (see Bibliography section) are available as background. These reviews have covered general aspects of neurochemical collection with microdialysis sampling (143–146), microchemical analysis (147,148), and controversial aspects of neurotransmitter collection (84,149–151)

Neuropeptides. With successful sampling of hydrophilic neurotransmitters with microdialysis sampling, the next logical analyte class to target was neuropeptides. Like neurotransmitters, the quantitation of neuropeptides is challenging with microdialysis sampling due to their low concentrations. Furthermore, their lower diffusion coefficients cause their E_d values to be low. Temporal resolution can also be an issue since quite often immunoassays that require 50–100 μL are often used for detection. Despite these limitations, many neuropeptides have been sampled using microdialysis sampling and have been reviewed > 15 years ago (152,153). With the increased use of mass spectrometry for neuropeptide detection of dialysates (154,155) coupled with additional bead-based immunoassays, the application space for microdialysis sampling of neuropeptides should increase tremendously.

Pharmacokinetics

Microdialysis sampling has been applied for pharmacokinetic studies in animals and humans. The great advantage here is that microdialysis sampling tremendously decreases the overall number of animals used for a pharmacokinetics study. Typical pharmacokinetic studies in rodents require the animal to be sacrificed at each time point used for the analysis. Microdialysis sampling allows for collection throughout the time course of the experiment because it can be easily inserted into the jugular vein. Again, because of the highly flexible nature of liquid chromatographic analysis for drug studies, the microdialysis sampling technique can be rapidly applied to new drugs and their metabolites.

Microdialysis sampling applications in pharmacokinetics have been extensively reviewed. In addition to general reviews of the subject (156–158), there have been reviews focused on data analysis (159,160) and calibration (161,162). One of the more important points to consider

when working with pharmacokinetic data obtained using microdialysis sampling is that a microdialysis sample represents a concentration average over the collection time period. This is in contrast to blood sampling that represents the analyte concentration directly at the sample time.

Clinical Applications

Clinical applications of microdialysis sampling have grown tremendously over the past decade and will continue to grow (163). At the present time, microdialysis sampling methods in human subjects have focused on studies in peripheral tissues for glucose collection (164–167) or drug distribution (168–170). Additional applications have been to determine gut barrier dysfunction (171). Microdialysis sampling applications in tumors has spanned both pharmacokinetic investigations (172–174), as well as collection of growth factors and cytokines (175). Microdialysis sampling has also been applied to human brain studies that have focused on understanding the underlying altered biochemistry that occurs when head trauma to drug distribution (176–180). Microdialysis sampling has been used as a means to monitor pharmacokinetics in the human dermis and has been compared to blister suction techniques (181). While both sampling methods produced similar data, it was found that microdialysis sampling was much easier to handle for both the patient and clinician.

Monitoring different growth factors and cytokines is becoming more important in clinical medicine since these proteins are known to affect cell-to-cell signaling and communication and are therefore becoming important biomarkers to measure. In particular, a group of proteins that are of great *in vivo* interest are the cytokines. Cytokines are potent, transient, and highly localized soluble messenger proteins (~6–80 kDa) produced by T-cells and macrophages that control nearly every aspect of the immune system (182). Cytokines exhibit complex interactions and therefore it is often more important to determine the concentration and cytokine profile after an immune challenge rather than the concentration of one single cytokine.

Microdialysis sampling is an ideal technique to achieve real time *in situ* monitoring of these important protein mediators and also has been recently described for proteomics applications (183). The application of microdialysis to this area is now emerging as potential approach for clinical *in vivo* studies in both healthy and diseased subjects to recover targeted cytokine molecules from the exact action sites and has recently been reviewed by Clough (25). Commercially available microdialysis probes with a 100 kDa MWCO membrane have been used for *in vivo* microdialysis of some cytokines (184,185). It is important to note that microdialysis sampling provides localized sampling and thus insight into localized concentrations of cytokines that cannot be achieved via sampling from blood plasma. This has been recently demonstrated with the cytokine IL-6 where its interstitial fluid concentration was 100-fold higher than that found in the plasma (186).

Cytokines have low E_d through 100 kDa membranes. To improve cytokine E_d larger MWCO membranes (3000 kDa) typically used for plasmaphoresis have been used (187). Others are beginning to use the 3000 kDa MWCO membrane for collection of IL-6 and TGF- β_1 (186,188–190). The range of *in vitro* recoveries for different cytokine proteins is shown in Table 4.

EVALUATION AND FUTURE USE

Microdialysis sampling has become a mature technology for neurotransmitter collection and pharmacokinetic determinations in animals. Clinical microdialysis sampling applications provide the greatest opportunity for growth. Despite the extensive biomedical use of conventional microdialysis sampling, there are still aspects of the device that could be tremendously improved.

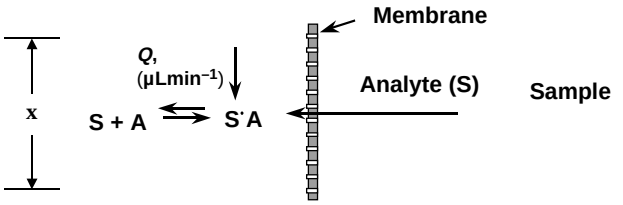
As currently practiced, microdialysis sampling in animals can be cumbersome due to the tubing lines required. Work in Lunte's group has focused on making micropumps using osmotic pumps as means to create line-free dialysis device (191,192). Reducing the microdialysis size by creating it on a microchip also has some advantages given that a decreased volume flow chamber may allow rapid equilibration across the device allowing E_d to approach nearly 100% (193–195).

A common problem with microdialysis sampling is the difficulty incurred when sampling hydrophobic analytes. This is an area with great promise with respect to either new device development or improvements to existing microdialysis sampling technology. Albumin is commonly included in the perfusion fluid to block

Table 4. Cytokine In Vitro Relative Recovery and Relevant Physicochemical Properties[a]

Cytokine	E_d% 0.5 $\mu L \cdot min^{-1}$	E_d% 1.0 $\mu L \cdot min^{-1}$	MW, kDa	Conformation	Active Protein, kDa
IL-2	4.5 ± 2.9 (12)	2.9 ± 1.1 (15)	17.2	Monomer	17.2
IL-4	12.0 ± 6.5 (12)	7.5 ± 2.6 (15)	13.6	Monomer	13.6
IL-5	1.3 ± 1.0 (12)	1.0 ± 0.5 (15)	13.1	Homodimer	26.2
IL-6	4.8 ± 1.4 (6)	2.7 ± 0.8 (6)	21.7	Monomer	21.7
IL-10	Not performed	1.1 ± 0.3 (3)	18.8	Homodimer	37.6
IL-12p70	N.D.	N.D.	35 & 40	Heterodimer	75
IFN-γ	2.0 ± 1.4 (18)	1.5 ± 0.8 (21)	15.9	Homodimer	31.8
MCP-1	24.5 ± 4.8 (6)	13.1 ± 3.9 (6)	13.1	Homodimer	26.2
TNF-α	8.0 ± 2.9 (15)	4.3 ± 1.1 (18)	17.3	Homotrimer	51.9

[a]Cytokine standards were either 1250 or 2500 $pg \cdot mL^{-1}$. All solutions were quiescent at room temperature. A CMA/20 10-mm 100-kDa PES membrane was used for these studies performed in our laboratory. The numbers in parentheses after the RR values are the number of trials (n). All data are reported as mean $\pm$ SD. N.D. is not detected.



$$\left(\frac{dC_L}{dx}\right) = \frac{2pD_{\text{mem}}(C_o - C_L)}{Q\ln(R_o/R_i)} - \frac{pR_i^2kC_L}{Q}$$

At $x = 0$, $C = C_o$
At $x = L$, $C = C_L$

Figure 9. Schematic of microdialysis enhanced transport. The analyte (S) or substrate for an affinity agent (A) binds with the affinity agent in the perfusion fluid. Affinity agents include antibodies, cyclodextrins or other supramolecular agents. The symbols in the equation are as follows: C_o = Concentration at $x = 0$, C_L = Concentration at $x = L$ (membrane length), D_{mem} = membrane diffusion coefficient, k = first-order rate constant for the reaction between the analyte, S, and the affinity agent, A, Q = perfusion fluid flow rate and R_i, R_o = inner and outer radii.

nonspecific adsorption sites within the microdialysis polymeric materials (196,197), and is still widely practiced (198). Unfortunately, this adds protein back into the dialysis perfusion that for small analytes causes difficulties with CE and LC analyses. A second approach involves using lipids, such as Intralipid, a material used to suspend hydrophobic drugs, to coat the nonspecific adsorption sites on the dialysis membrane or tubing (196,199). It is not known how compatible this approach is for CE or LC analysis.

Cyclodextrins have been used as perfusion-fluid additives for improving the microdialysis recovery for different analytes (200,201). A general scheme for the enhancement process is shown in Fig. 9. Cyclodextrins are well-known cyclic oligosacharides that have the capability to form inclusion complexes with various organic molecules by the capture of the guest molecule into a hydrophobic central cavity (202). Dialysate samples that contain cyclodextrin can be injected into an LC column without alteration to the chromatographic separation parameters, for example, plate number and peak width (203). Cyclodextrins are not selective, which for this enhancement approach is beneficial given that they can be applied to a wide variety of low molecular weight hydrophobic organic molecules (204,205). However, a difficulty with this approach is that cyclodextrins can diffuse out of the dialysis membrane that may complicate some applications. An additional difficulty is that cyclodextrins can interfere with affinity-based detection assays (immunoassays) due to competition between the analyte and cyclodextrin versus analyte and an antibody.

To overcome the free diffusion of cyclodextrin out of the dialysis tubing, different type of solid supports have been used for enhancement. Markides group has described the use of a solid-vehicle support for improving neuropeptide microdialysis relative recovery coupled with LC–MS detection (206,207). More specific enhancements can be achieved using antibody-immobilized beads for flow cytometry (208). With the antibody-enhancement approach, increases in microdialysis sampling relative recovery of 4–12-fold were achieved for different cytokines.

Microdialysis sampling is already beginning to be used in metabolomic studies using LC–MS detection methods (209). With the creation of microcoil nuclear magnetic resorance (NMR) that can measure nanoliter samples (210,211) the detection possibilities for dialysate samples are greatly increased. This approach has been recently applied to metabolomic studies with microdialysis sampling in brain (212).

The combination of these new discoveries for microdialysis sampling shows the enormous potential for solving many clinical biomedical problems with this device. The future for microdialysis sampling and device spin-offs is quite bright. New developments in instrumentation and collection procedures will provide much clinical benefit.

Microdialysis sampling has moved from a sampling device exclusively used to collect low molecular weight hydrophilic neurotransmitters to applications requiring the collection of larger proteins. Many cellular signaling processes occur via small molecule (nitric oxide, norepinephrine, acetylcholine, eicosanoids, etc.), peptide (angiotensin, etc.) as well as proteins (cytokines). Cytokine profiling for different disease states has become quite important. For the applications of both neuropeptide and larger protein collection such as the cytokines, the principal limitation to collecting these molecules is the diffusion properties of these analytes through the dialysis membrane. Proteins are difficult to collect through dialysis membranes since they can exhibit both nonspecific adsorption to the polymeric materials as well as diffusion restrictions through the membrane pores. Despite the variety of applications of microdialysis sampling to medical and scientific research, there have been few new developments with respect to improving the overall sampling efficiency for difficult to dialyze samples. As clinical proteomics and biomarker collection becomes better understood for making clinical predictions, it will be necessary to have sampling methods that can meet these clinical needs.

ACKNOWLEDGMENTS

Microdialysis sampling research in the Stenken laboratory is currently funded by NIH EB001441. The support from the John Wiley & Sons, Inc. editorial staff is greatly appreciated.

BIBLIOGRAPHY

1. Frangioni JV. *In vivo* near-infrared fluorescence imaging. Curr Opin Chem Biol 2003;7:626–634.
2. Ntziachristos V, Tung CH, Bremer C, Weissleder R. Fluorescence molecular tomography resolves protease activity *in vivo*. Nature Med 2002;8:757–761.
3. Sevick-Muraca EM, et al. Near-infrared imaging with fluorescent contrast agents. Handbook Biomed Fluoresc 2003; 445–527.
4. Myers RD, Adell A, Lankford MF. Simultaneous comparison of cerebral dialysis and push–pull perfusion in the brain of

rats: a critical review. Neurosci Biobehav Rev 1998;22:371–387.
5. Ellmerer M, et al. Measurement of interstitial albumin in human skeletal muscle and adipose tissue by open-flow microperfusion. Am J Physiol 2000;278:E352–E356.
6. Haaverstad R, Romslo I, Larsen S, Myhre HO. Protein concentration of subcutaneous interstitial fluid in the human leg. A comparison between the wick technique and the blister suction technique. Inter J Microcirculation, Clin Exper 1996;16:111–117.
7. Wiig H, Sibley L, DeCarlo M, Renkin EM. Sampling interstitial fluid from rat skeletal muscles by intermuscular wicks. Am J Physiol 1991;261:H155–H15.
8. Nedrebo T, et al. (2004). Differential cytokine response in interstitial fluid in skin and serum during experimental inflammation in rats. J Physiol 2004;556:193–202.
9. Bito LZ, et al. Concentration of free amino acids and other electrolytes in cerebrospinal fluid, in vivo dialyzate of brain, and blood plasma of the dog. J Neurochem 1966;13:1057–1067.
10. Delgado JM, et al. Dialytrode for long term intracerebral perfusion in awake monkeys. Arch Inter Pharmacodyn Ther 1972;198:9–21.
11. Ungerstedt U, Pycock C. Functional correlates of dopamine neurotransmission. Bull Schweizerischen Akad Medizinischen Wisse 1974;30:44–55.
12. Plock N, Kloft C. Microdialysis - theoretical background and recent implementation in applied life-sciences. Eur J Pharmaceut Sci 2005;25:1–24.
13. Davies MI, et al. Analytical considerations for microdialysis sampling. Adv Drug Deliv Rev 2000;45:169–188.
14. Benveniste H, Hüttemeier C. Microdialysis-theory and application. Prog Neurobiol 1990;35:195–215.
15. de Lange ECM, de Boer AG, Breimer DD. Methodological issues in microdialysis sampling for pharmacokinetic studies. Ad Drug Del Rev 2000;45:125–148.
16. Borg N, Stahle L. Recovery as a function of the osmolality of the perfusion medium in microdialysis experiments. Anal Chim Acta 1998;375:135–141.
17. Moghaddum B, Bunney BS. Ionic compositions of microdialysis perfusing solution alters the pharmacological responsiveness and basal outlflow of striatal dopamine. J Neurochem 1989;53:652–654.
18. Davies MI. A review of microdialysis sampling for pharmacokinetic applicaitons. Analyt Chim Acta 1999;379:227–249.
19. De Lange CMA, De Boer AG, Breimer DD. Intracerebral microdialysis. Intro Blood–Brain Barrier 1998; 94–112.
20. Bourne JA (2003). Intracerebral microdialysis: 30 years as a tool for the neuroscientist. Clin Exp Pharmacol Physiol 2003;30:16–24.
21. North American and International Commercial Sources of Microdialysis Probes. Available at CMA Microdialysis Inc. www.microdialysis.se; Bioanalytical Systems Inc., www. bioanalytical.com; SciPro www.scipro.com .
22. De la Peña A, Liu P, Derendorf H. Microdialysis in peripheral tissues. Adv Drug Deli Rev 2000;45:189–216.
23. Boschi G, Scherrmann JM. Microdialysis in mice for drug delivery research. Adv Drug Deliv Rev 2000;45:271–281.
24. Mulder M. Basic Principles of Membrane Technology. Dordrecht, The Netherlands: Kluwer Academic Publishers 1991.
25. Clough G. Microdialysis sampling of large molecules. The AAPS Journal E-Pub May 16, 2005.
26. Kendrick KM. Use of microdialysis in neuroendocrinology. Methods Enzymol 1989;168:182–205.
27. Torto N, et al. Optimal membrane choice for microdialysis sampling of oligosaccharides. J Chromatog A 1998;806:265–278.
28. Kjellstrom S, et al. Microdialysis—a membrane based sampling technique for quantitative determination of proteins. Chromatogr 1999;50:539–546.
29. Snyder KL, et al. Diffusion and calibration properties of microdialysis sampling membranes in biological media. Analyst 2001;126:1261–1268.
30. Schuck VJA, Rinas I, Derendorf H. *In vitro* microdialysis sampling of docetaxel. J Pharmaceutic Biomed Anal 2004;36: 807–813.
31. Groth L, Jírgensen A. *In vitro* microdialysis of hydrophilic and lipophilic compounds. Analyt Chim Acta 1997;355:75–83.
32. Mary S, et al. Assessment of the recovery of three lipophilic psoralens by microdialysis: An in vitro study. Inter J Pharmaceut 1998;161:7–13.
33. Patterson SL, Sluka KA, Arnold MA. A novel transverse push–pull microprobe: in vitro characterization and *in vivo* demonstration of the enzymatic production of adenosine in the spinal cord dorsal horn. J Neurochem 2001;76:234–246.
34. Patterson SL, Sluka KA, Arnold MA. A novel transverse push–pull microprobe: *In vitro* characterization and in vivo demonstration of the enzymatic production of adenosine in the spinal cord dorsal horn. [Erratum to document cited in A]. J Neurochem 2001;76:1955.
35. Ho WSW, Sikar KK, editors. Membrane Handbook New York; Van Nostrand Reinhold: 1992.
36. Rosdahl H, Hamrin K, Ungerstedt U, Henriksson J, A microdialysis method for the in situ investigation of the action of large peptide molecules in human skeletal muscle: Detection of local metabolic effects of insulin. Intern J Biol Macromole 2000;28:69–73.
37. Hamrin K, Rosdahl H, Ungerstedt U, Henriksson J. Microdialysis in human skeletal muscle: Effects of adding a colloid to the perfusate. J Appl Physiol 2002;92:385–393.
38. Smith AD, Justice JB. The effect of inhibition of synthesis, release, metabolism and uptake on the microdialysis extraction fraction of dopamine. J Neurosci Methods 1994;54:75–82.
39. Haskins WE, et al. Capillary LC-MS^2 at the attomole level for monitoring and discovering endogenous peptides in microdialysis samples collected *in vivo*. Analy Chem 2001;73:5005–5014.
40. Freed AL, Cooper JD, Davies MI, Lunte SM. Investigation of the metabolism of substance P in rat striatum by microdialysis sampling and capillary electrophoresis with laser-induced fluorescence detection. J Neurosc Methods 2001;109:23–29.
41. Ste-Mari L, Boismenu D, Vachon L, Montgomery J. Evaluation of sodium 4-hydroxybenzoate as an hydroxyl radical trap using gas chromatography-mass spectrometry and high-performance liquid chromatography with electrochemical detection. Ana Bioch 1996;241:67–74.
42. Chen R, Stenken JA. An in vitro hydroxyl radical generation assay for microdialysis sampling calibration. Anal Biochem 2002;306:40–49.
43. Marklund N, Clausen F, Lewander T, Hillered L. Monitoring of reactive oxygen species production after traumatic brain injury in rats with microdialysis and the 4-hydroxybenzoic acid trapping method. J Neurotrauma 2001;18: 1217–1227.
44. Scott DO, Lunte CE *In vivo* microdialysis sampling in the bile, blood, and liver of rats to study the disposition of phenol. Pharmaceut Res 1993;10:335–342.

45. Stenken JA, Ståthle L, Lunte CE, Southard MZ. Monitoring in situ liver metabolism in rats using microdialysis. Comparison of a microdialysis mass-transport model predictions to experimental metabolite generation data. J Pharmaceut Sci 1998;87:311–320.
46. Kajiro T, Nakajima Y, Fukushima T, Imai K. A method to evaluate the renin-angiotensin system in rat renal cortex using a microdialysis technique combined with HPLC-fluorescence detection. Anal Chem 2002;74:4519–4525.
47. Steuerwald AJ, Villeneuve JD, Sun L, Stenken JA. *In vitro* characterization of an in situ, microdialysis sampling assay for elastase activity detection. J Pharmaceut Biome Analy. In press.
48. Bungay PM, Morrison PF, Dedrick RL. Steady-state theory for quantitative microdialysis of solutes and water *in vivo* and in vitro. Life Sci 1990;46:105–119.
49. Bungay PM, et al. Microdialysis of dopamine interpreted with quantitative model incorporating probe implantation trauma. J Neurochem 2003;86:932–946.
50. Justice JB. Quantitative microdialysis of neurotransmitters. J Neurosc Methods 1993;48:263–276.
51. Clough GF, et al. Effects of blood flow on the *in vivo* recovery of a small diffusible molecule by microdialysis in human skin. J Pharmacol Exp Therapeut 2002;302:681–686.
52. Nicholson C. Diffusion and related transport mechanisms in brain tissue. Rep Prog Phys 2001;64:815–884.
53. Newton AP, Justice JB. Temporal response of microdialysis probes to local perfusion of dopamine and cocaine followed with one-minute sampling. Analy Chem 1994;66:1468–1472.
54. Cosford RJO, Vinson AP, Kukoyi S, Justice JBJ. Quantitative microdialysis of serotonin and norepinephrine: Pharmacological influences on *in vivo* extraction fraction. J Neurosci Methods 1996;68:39–47.
55. Vinson PN, Justice JB. Effect of neostigmine on concentration and extraction fraction of acetylcholine using quantitative microdialysis. J Neurosci Methods 1997;73:61–67.
56. Stenken JA, Lunte CE, Southard MZ, Ståhle L. Factors that influence microdialysis recovery. Comparison of experimental and theoretical microdialysis recoveries in rat liver. J Pharmaceut Sci 1997;86:958–966.
57. Dai H, Elmquist WF. Drug transport studies using quantitative microdialysis. Methods Mol Med 2003;89:249–264.
58. Stenken JA. Methods and issues in microdialysis calibration. Anal Chim Acta 1999;379:337–357.
59. Chen KC, et al. Theory relating in vitro and in vivo microdialysis with one or two probes. J Neurochem 2002;81:108–121.
60. Rice ME, Nicholson C. Diffusion and ion shifts in the brain extracellular microenvironment and their relevance for voltammetric measurements. The brain is not a beaker: *In vivo* vs. *In vitro* voltammetry. In: Boulton A, Baker G, Adams RN, editors. Neuromethods, Vol. 27, Voltammetric Methods in Brain Systems, ed. New York: Humana Press Inc.; 1995. pp 27–79.
61. Reach G, Wilson GS. Can continuous glucose monitoring be used for the treatment of diabetes? Analy Chem 1992;64: 381A–386A.
62. Benveniste H, Hüttemeir C. Microdialysis-theory and application. Progr Neurobiol 1990;35:195–215.
63. Jacobson I, Sandberg M, Hamberger A. Mass transfer in brain dialysis devices—a new method for the estimation of extracellular amino acids concentration. J Neurosci Methods 1985;15: 263–268.
64. Menacherry S, Hubert W, Justice JB. *In vivo* calibration of microdialysis probes for exogenous compounds. Analyt Chem 1992;64:577–583.
65. Yokel RA, Allen DD, Burgio DE. McNamara JP. Antipyrine as a dialyzable reference to correct differences in efficiency among and within sampling devices during in vivo microdialysis. J Pharmacol Toxicol Methods 1992;27:135–142.
66. Strindberg L, Lonnroth P. Validation of an endogenous reference technique for the calibration of microdialysis catheters. Scand J Clin Lab Investigation 2000;60:205–212.
67. Ronne-Engstrom E, et al. Intracerebral microdialysis in neurointensive care: The use of urea as an endogenous reference compound. J Neurosur 2001;94:397–402.
68. Ettinger SN, et al. Urea as a recovery marker for quantitative assessment of tumor interstitial solutes with microdialysis. Cancer Res 2001;61:7964–7970.
69. Leypoldt JK, Burkart JM. Small-solute and middle-molecule clearances during continuous flow peritoneal dialysis. Adv Peritoneal Dialysis 2002;18:26–31.
70. Krejcie TC, et al. Modifications of blood volume alter the disposition of markers of blood volume, extracellular fluid, and total body water. J Pharmacol Exp Therap 1999;291: 1308–1316.
71. Benveniste H, Diemer NH. Cellular reactions to implantation of a microdialysis tube in the rat hippocampus. Acta Neuropathol 1987;74:234–238.
72. Grabb MC, et al. Neurochemical and morphological responses to acutely and chronically implanted brain microdialysis probes. J Neurosci Methods 1998;82:25–34.
73. Anderson JM. Biological responses to materials. Ann Rev Mater Res 2001;31:81–110.
74. Dykstra KH, et al. Quantitative examination of tissue concentration profiles associated with microdialysis. J Neurochem 1992;58:931–940.
75. Davies MI, Lunte CE. Microdialysis sampling for hepatic metabolism studies: impact of microdialysis probe design and implantation technique on liver tissue. Drug Metabolism Disposition 1995;23:1072–1079.
76. Clapp-Lilly KL, et al. An ultrastructural analysis of tissue surrounding a microdialysis probe. J Neurosci Methods, 1999;90:129–142.
77. Planas AM, et al. Certain forms of matrix metalloproteinase-9 accumulate in the extracellular space after microdialysis probe implantation and middle cerebral artery occlusion/reperfusion. J Cerebral Blood Flow Metabolism 2002;22: 918–925.
78. Wisniewski N, et al. Analyte flux through chronically implanted subcutaneous polyamide membranes differs in humans and rats. Am J Physiol, Endocrinol Metabolism 2002;282:E1316–E1323.
79. Lu Y, Peters JL, Michael AC. Direct comparison of the response of voltammetry and microdialysis to electrically evoked release of striatal dopamine. J Neurochem 1998;70:584–593.
80. Borland LM, Shi G, Yang H, Michael AC. Voltammetric study of extracellular dopamine near microdialysis probes acutely implanted in the striatum of the anesthetized rat. J Neurosci Methods 2005;146:149–158.
81. Morgan ME, Singhal D, Anderson BD. Quantitative assessment of blood-brain barrier damage during microdialysis. J Pharmacol Exp Therap 1996;277:1167–1176.
82. Groothuis DR, et al. Changes in blood-brain barrier permeability associated with insertion of brain cannulas and microdialysis probes. Brain Res 1998;803:218–230.
83. Chen KC. Preferentially impaired neurotransmitter release sites not their discreteness compromise the validity of microdialysis zero-net-flux method. J Neurochem 2005;92: 29–45.

84. Khan SA, Michael AC. Invasive consequences of using microelectrodes and microdialysis probes in the brain. TrAC, Trends Anal Chem 2003;22:503–508.
85. Ballerstadt R, Schultz JS. Sensor methods for use with microdialysis and ultrafiltration. Adv Drug Del Rev 1996; 21:225–238.
86. Davies MI, et al. Analytical considerations for microdialysis sampling. Adv Drug Del Rev 2000;45:169–188.
87. Adell A, Artigas F. *In vivo* brain microdialysis: Principles and applications. Neuromethods 1998;32:1–33.
88. Chaurasia CS. *In vivo* microdialysis sampling: Theory and applications. Biomed Chromatog 1999;13:317–332.
89. Church WH, Justice Jr JB. On-line small-bore chromatography for neurochemical analysis in the brain. Adv Chromatog 1989;28:165–194.
90. Davies IM, Lunte CE. Microdialysis sampling coupled on-line to microseparation techniques. Chem Soc Rev 1997;26:215–222.
91. Horn TFW, Engelmann M. *In vivo* microdialysis for nonapeptides in rat brain—a practical guide. Methods 2001;23:41–53.
92. Kennedy RT. Bioanalytical applications of fast capillary electrophoresis. Anal Chim Acta 1999;400:163–180.
93. Kennedy RT, et al. *In vivo* neurochemical monitoring by microdialysis and capillary separations. Curr Opin Chem Biol 2002;6:659–665.
94. Lunte CE, Scott DO, Kissinger PT. Sampling living systems using microdialysis probes. Anal Chem, 1991;63:773A–780A.
95. Lunte SM, Lunte CE. Microdialysis sampling for pharmacological studies: HPLC and CE analysis. In: Brown PR, Grushka E, editors. Advances in Chromatography. New York: Marcel Dekker; 1996. pp 383–432.
96. Obrenovitch TP, Zilkha E. Microdialysis coupled to online enzymatic assays. Methods 2001;23:63–71.
97. Parkin MC, Hopwood SE, Boutelle MG, Strong AJ. Resolving dynamic changes in brain metabolism using biosensors and on-line microdialysis. TrAC, Trends Analy Chem 2003;22:487–497.
98. Parrot S, et al. Microdialysis monitoring of catecholamines and excitatory amino acids in the rat and mouse brain: Recent developments based on capillary electrophoresis with laser-induced fluorescence detection-A mini-review. Cellular Mol Neurobiol 2003;23:793–804.
99. Wightman RM, Strope E, Plotsky PM, Adams RN. Monitoring of transmitter metabolites by voltammetry in cerebrospinal fluid following neural path stimulation. Nature London 1976;262:145–146.
100. Adams RN. Probing brain chemistry with electroanalytical techniques. Anal Chem 1976;48:1126A–1138A.
101. Nagy G, Rice ME, Adams RN. A new type of enzyme electrode: The ascorbic acid eliminator electrode. Life Sci 1982; 31:2611–2616.
102. Wages SA, Church WH, Justice Jr JB. Sampling considerations for on-line microbore liquid chromatography of brain dialysate. Analy Chem 1986;58:1649–1656.
103. Kennedy RT, German I, Thompson JE, Witowski SR. Fast analytical-scale separations by capillary electrophoresis and liquid chromatography. Chem Rev 1999;99:3081–3131.
104. Arnett SD, Lunte CE. Investigation of the mechanism of pH-mediated stacking of anions for the analysis of physiological samples by capillary electrophoresis. Electrophoresis 2003;24:1745–1752.
105. Davies M, et al. Studies on animal to instrument hyphenation: Development of separation-based sensors for near real-time monitoring of drugs and neurotransmitters. Sample Preparation for Hyphenated Analyt Tech 2004; 191–220.
106. Rossell S, Gonzalez LE, Hernandez L. One-second time resolution brain microdialysis in fully awake rats. Protocol for the collection, separation and sorting of nanoliter dialyzate volumes. J Chromatogr B: Analy Technol Biomed Life Sci 2003;784:385–393.
107. Chen A, Lunte CE. Microdialysis sampling coupled on-line to fast microbore chromatography. J Chromatog Analy Appl 1995;691:29–35.
108. Kissinger PT, Refshauge C, Dreiling R, Adams RN. Electrochemical detector for liquid chromatography with picogram sensitivity. Analyt Lett 1973;6:465–477.
109. Acworth IN, Bogdanov MB, McCabe DR, Beal MF. Estimation of hydroxyl free radical levels *in vivo* based on liquid chromatography with electrochemical detection. Methods Enzymol 1999;300:297–313.
110. Acworth IN, Bailey BA, Maher TJ. The use of HPLC with electrochemical detection to monitor reactive oxygen and nitrogen species, markers of oxidative damage and antioxidants: application to the neurosciences. Prog HPLC-HPCE 1998;7:3–56.
111. Bogdanov MB, et al. A carbon column-based liquid chromatography electrochemical approach to routine 8-hydroxy-2'-deoxyguanosine measurements in urine and other biologic matrices: A one-year evaluation of methods. Free Rad Biol Med 1999;27:647–666.
112. Bogdanov MB, et al. Increased oxidative damage to DNA in a transgenic mouse model of Huntington's disease. J Neurochem 2001;79:1246–1249.
113. Stenken JA, Puckett DL, Lunte SM, Lunte CE. Detection of *N*-acetylcysteine, cysteine and their disulfides in urine by liquid chromatography with a dual-electrode amperometric detector. J Pharmaceut Biomed Analysis 1990;8:85–89.
114. Lunte SM, O'Shea TJ. Pharmaceutical and biomedical applications of capillary electrophoresis/electrochemistry. Electrophoresis 1994;15:79–86.
115. Zackheim JA, Abercrombie ED. HPLC/EC detection and quantification of acetylcholine in dialysates. Methods Mol Med 2003;79:433–441.
116. Kehr J, Dechent P, Kato T, Ogren SO. Simultaneous determination of acetylcholine, choline and physostigmine in microdialysis samples from rat hippocampus by microbore liquid chromatography/electrochemistry on peroxidase redox polymer coated electrodes. J Neurosci Methods 1998;83:143–150.
117. Kehr J. Determination of glutamate and aspartate in microdialysis samples by reversed-phase column liquid chromatography with fluorescence and electrochemical detection. J Chromatog B: Biomed Sci App 1998;708:27–38.
118. Kehr J. Determination of g-aminobutyric acid in microdialysis samples by microbore column liquid chromatography and fluorescence detection. J Chromatog B: Biomed Sci App 1998;708:49–54.
119. Tao L, Thompson JE, Kennedy RT. Optically gated capillary electrophoresis of o-phthalaldehyde/b-mercaptoethanol derivatives of amino acids for chemical monitoring. Analy Chem 1998;70:4015–4022.
120. Yamaguchi M, et al. Determination of norepinephrine in microdialysis samples by microbore column liquid chromatography with fluorescence detection following derivatization with benzylamine. Analy Biochem 1999;270: 296–302.
121. Mann M, Fenn JB. Electrospray mass spectrometry: principles and methods. Mass Spectrom 1992;1:1–35.
122. Emmett MR, Andren PE, Caprioli RM. Specific molecular mass detection of endogenously released neuropeptides using *in vivo* microdialysis/mass spectrometry. J Neurosci Methods 1995;62:141–147.

123. Wilm M, Mann M. Analytical properties of the nanoelectrospray ion source. Analy Chem 1996;68:1–8.
124. Juraschek R, Dulcks T, Karas M. Nanoelectrospray-more than just a minimized-flow electrospray ionization source. J Am Soc Mass Spectrom 1999;10:300–308.
125. Gangl ET, Annan M, Spooner N, Vouros P. Reduction of signal suppression effects in ESI-MS using a nanosplitting device. Analy Chem 2001;73:5635–5644.
126. Haskins WE, et al. Discovery and neurochemical screening of peptides in brain extracellular fluid by chemical analysis of *in vivo* microdialysis samples. Analy Chem 2004;76:5523–5533.
127. Heinemann L. Continuous glucose monitoring by means of the microdialysis technique: Underlying fundamental aspects. Diabetes Technol Therap 2003;5:545–561.
128. Wisniewski N, Moussy F, Reichert WM. Characterization of implantable biosensor membrane fouling. Fresenius J Analy Chem 2000;366:611–621.
129. Ward WK. Subcutaneous glucose monitoring: Microdialysis vs. intracorporeal. Diabetes Care 2002;25:410–411.
130. O'Neill RD, Lowry JP, Mas M. Monitoring brain chemistry *in vivo*: Voltammetric techniques, sensors, and behavioral applications. Crit Rev Neurobiol 1998;12:69–127.
131. Jones DA, et al. On-line monitoring in neurointensive care Enzyme-based electrochemical assay for simultaneous, continuous monitoring of glucose and lactate from critical care patients. J Electroanalyt Chem, 2002; 538–539, 243–252.
132. Boutelle MG, Fellows LK, Cook C. Enzyme packed bed system for the on-line measurement of glucose, glutamate, and lactate in brain microdialysis. Analy Chem 1992;64: 1790–1794.
133. Volpe G, Moscone D, Compagnone D, Palleschi G. *In vivo* continuous monitoring of ***L-lactate coupling subcutaneous microdialysis and an electrochemical biocell. Sensors Actuators B 1995; 24–25 138–141.
134. Ryan MR, Lowry JP, O'Neill RD. Biosensor for neurotransmitter L-glutamic acid designed for efficient use of L-glutamate oxidase and effective rejection of interference. Analyst 1997;122:1419–1424.
135. Cook CJ. Real-time measurement of corticosteroids in conscious animals using an antibody-based electrode. Nature Biotechnol 1997;15:467–471.
136. Sandberg M, Weber SG. Techniques for neuropeptide determination. TrAC, Trends Analy Chem 2003;22:522–527.
137. O'Connor KA, et al. A method for measuring multiple cytokines from small samples. Brain, Behavior, Immunity 2004; 18:274–280.
138. Li Y, Schutte RJ, Abu-Shakra A, Reichert WM. Protein array method for assessing *in vitro* biomaterial-induced cytokine expression. Biomaterials 2005;26:1081–1085.
139. Vignali DAA. Multiplexed particle-based flow cytometric assays. J Immunol Methods 2000;243:243–255.
140. Kellar KL, et al. Multiplexed fluorescent bead-based immunoassays for quantitation of human cytokines in serum and culture supernatants. Cytometry 2001;45:27–36.
141. Kellar KL, Iannone MA. Multiplexed microsphere-based flow cytometric assays. Exp Hemato 2002;30:1227–1237.
142. Xie R, Hammarlund-Udenaes M, De Boer AG, De Lange ECM. The role of P-glycoprotein in blood-brain barrier transport of morphine: Transcortical microdialysis studies in mdr1a (−/−) and mdr1a (+/+) mice. Br J Pharmacol 1999; 128:563–568.
143. Ungerstedt U. Introduction to intracerebral microdialysis. Tech Behav Neural Sci 1991;7:3–22.
144. Sharp T, Zetterstrom T. *In vivo* measurement of monoamine neurotransmitter release using brain microdialysis. Monit. Neuronal Act 1992; 147–179.
145. Westerink BHC, Timmerman W. Do neurotransmitters sampled by brain microdialysis reflect functional release? Anal Chim Acta 1999;379:263–274.
146. Salamone JD. The behavioral neurochem of motivation: methodological and conceptual issues in studies of the dynamic activity of nucleus accumbens dopamine. J Neurosc Methods 1996;64:137–149.
147. Justice Jr JB. Microchemical analysis in the brain. Microchem J 1986;34:11–14.
148. Kennedy RT, et al. *In vivo* neurochemical monitoring by microdialysis and capillary separations. Curr Opin Chem Biol 2002;6:659–665.
149. Fuxe K, Ferre S, Zoli M, Agnati LF. Integrated events in central dopamine transmission as analyzed at multiple levels. Evidence for intramembrane adenosine A2A/dopamine D2 and adenosine A1/dopamine D1 receptor interactions in the basal ganglia. Brain Res Rev 1998;26:258–273.
150. Del Arco A, Segovia G, Fuxe K, Mora F. Changes in dialysate concentrations of glutamate and GABA in the brain: An index of volume transmission mediated actions? J Neurochem 2003;85:23–33.
151. Di Chiara G, Tanda G, Carboni E. Estimation of in-vivo neurotransmitter release by brain microdialysis: The issue of validity. Behav Pharmacol 1996;7:640–657.
152. Kendrick KM. Use of microdialysis in neuroendocrinology. Methods Enzymol 1989;168:182–205.
153. Kendrick KM. Microdialysis measurement of *in vivo* neuropeptide release. J Neurosc Methods 1990;34:35–46.
154. Andren PE, Lin-S. N, Caprioli RM. Microdialysis/mass spectrometry. Mass Spectrom 1994;2:237–254.
155. Andren PE, Farmer TB, Klintenberg R. Endogenous release and metabolism of neuropeptides utilizing *in vivo* microdialysis microelectrospray mass spectrometry. Mass Spectrom Hyphenated Tech Neuropeptide Res 2002; 193–213.
156. Ståhle L, Microdialysis in pharmacokinetics. Eur J Drug Metabol Pharmacokine 1993;18:89–96.
157. de Lange ECM, Danhof M, de Boer AG. Breimer DD. Methodological considerations of intracerebral microdialysis in pharmacokinetic studies on drug transport across the blood-brain barrier. Brain Res Rev 1997;25:27–49.
158. Hansen DK, et al. Pharmacokinetic and metabolism studies using microdialysis sampling. J Pharmaceut Sci 1999;88: 14–27.
159. Ståhle L. Pharmacokinetic estimations from microdialysis data. Eur J Clin Pharmacol 1992;43:289–294.
160. Ståhle L. Zero and first moment area estimation from microdialysis data. Eur J Clin Pharmacol 1993;45:477–481.
161. Hammarlund-Udenaes M, Paalzow LK. de Lange ECM. Drug equilibration across the blood-brain barrier—pharmacokinetic considerations based on the microdialysis method. Pharmaceut Res 1997;14:128–134.
162. Bungay PM, Dedrick RL, Fox E, Balis FM. Probe calibration in transient microdialysis *in vivo*. Pharmaceut Res 2001;18: 361–366.
163. Muller M. Science, medicine, and the future: Microdialysis. Br Med J 2000;324:588–591.
164. De Boer J, Korf J, Plijter-Groendijk H. *In vivo* monitoring of lactate and glucose with microdialysis and enzyme reactors in intensive care medicine. Inter J Artif Organs 1994;17:163–170.
165. Weintjes KJ, et al. Microdialysis of glucose in subcutaneous adipose tissue up to 3 weeks in healthy volunteers. Diabetes Care 1998;21:1481–1488.
166. Gudbjoernsdottir S, et al. Direct measurements of the permeability surface area for insulin and glucose in human

skeletal muscle. J Clin Endocrino Metab 2003;88:4559–4564.
167. Lonnroth P. Microdialysis in adipose tissue and skeletal muscle. Hormone Metabol Res 1997;29:344–346.
168. Blochl-Daum B, et al. Measurement of extracellular fluid carboplatin kinetics in melanoma metastases with microdialysis. Br J Cancer 1996;73:920–924.
169. Muller M, et al. *In vivo* drug-response measurements in target tissues by microdialysis. Clin Pharmacol Therapeut 1997;62:165–170.
170. Lindberger M, Tomson T. Ståhle L. Validation of microdialysis sampling for subcutaneous extracellular valproic acid in humans. Therapeut Drug Monitoring 1998;20:358–362.
171. Solligård E, et al. Gut barrier dysfunction as detected by intestinal luminal microdialysis. Int Care Med 2004;30: 1188–1194.
172. Chu J, Gallo JM. Application of microdialysis to characterize drug disposition in tumors. Adv Drug Del Rev 2000;15:243–253.
173. Mader RM, et al. Penetration of capecitabine and its metabolites into malignant and healthy tissues of patients with advanced breast cancer. Br J Cancer 2003;88:782–787.
174. Johansen MJ, Thapar N, Newman RA, Madden T. Use of microdialysis to study platinum anticancer agent pharmacokinetics in preclinical models. J Expe Therap Oncol 2002; 2:163–173.
175. Dabrosin C. Microdialysis — an *in vivo* technique for studies of growth factors in breast cancer. Frontiers Biosci 2005;10: 1329–1335.
176. Hamani C, Luer MS, Dujovny M. Microdialysis in the human brain: review of its applications. Neurolog Res 1997;19:281–288.
177. Vespa P, et al. Increase in extracellular glutamate caused by reduced cerebral perfusion pressure and seizures after human traumatic brain injury: A microdialysis study. J Neurosur 1998;89:971–982.
178. Sherwin AL. Neuroactive amino acids in focally epileptic human brain: A review. Neurochem Res 1999;24:1385–1395.
179. Bradberry CW. Applications of microdialysis methodology in nonhuman primates: Practice and rationale. Crit Rev Neurobiol 2000;14:143–163.
180. Laruelle M. Imaging synaptic neurotransmission with *in vivo* binding competition techniques: A critical review. J Cerebral Blood Flow Metab 2000;20:423–451.
181. Benfeldt E, Serup J, Menne T. Microdialysis vs. suction blister technique for *in vivo* sampling of pharmacokinetics in the human dermis. Acta Dermato-Venereolog 1999;79: 338–342.
182. Lefkowitz DL, Lefkowitz SS. (2001) Macrophage-neutrophil interaction: A paradigm for chronic inflammation revisited. Immunol Cell Biol 2001;79:502–506.
183. Maurer MH, et al. The proteome of human brain microdialysate. Proteome Sci 2003;1:7–15.
184. Sjögren F, Svensson C, Anderson C. Technical prerequisites for *in vivo* microdialysis determination of interleukin-6 in human dermis. Br J Dermatol 2002;146:375–382.
185. Ao X, Rotundo RF, Loegering DJ, Stenken JA. *In vivo* microdialysis sampling of cytokines produced in mice given bacterial lipopolysaccharide. J Microbiol Methods 2005;62:327–336.
186. Sopasakis VR, et al. High local concentrations and effects on differentiation implicate interleukin-6 as a paracrine regulator. Obesity Res 2004;12:454–460.
187. Winter CD, et al. A microdialysis method for the recovery of IL-1beta, IL-6 and nerve growth factor from human brain *in vivo*. J Neurosc Methods 2002;119:45–50.
188. Rosendal L, et al. Increase in interstitial interleukin-6 of human skeletal muscle with repetitive low-force exercise. J Appl Physiol 2005;98:477–481.
189. Heinemeier K, Langberg H, Olesen JL, Kjaer M. Role of TGF-beta1 in relation to exercise-induced type I collagen synthesis in human tendinous tissue. J Appl Physiol 2003;95:2390–2397.
190. Langberg H, Olesen JL, Gemmer C, Kjaer M. Substantial elevation of interleukin-6 concentration in peritendinous tissue, in contrast to muscle, following prolonged exercise in humans. J Physiol 2002;542:985–990.
191. Lin Y-C, Hesketh PJ, Lunte SM, Wilson GS. A micromachined diaphragm micropump. Proc—Electrochem Soci 1995; 95–27. 67–72.
192. Hesketh PJ, et al. Biosensors and microfluidic systems. Tribology Issues and Opportunities in MEMS, Proceedings of the NSF/AFOSR/ASME Workshop on Tribiology Issues and Opportunities in MEMS, Columbus, Ohio, Nov. 9–11, 1997, 1998; pp 85–94.
193. Zahn JD, Trebotich D, Liepmann D. Microfabricated microdialysis microneedles for continuous medical monitoring. Proceedings of the 1st Annual International IEEE/EMBS Special Topics Conference on Microtechnologies in Medicine & Biology. October 12–14, 2000, Lyon, France; 2000 pp 375–380.
194. Talbot D, Liepmann D, Pisano AP. Microfabricated polysilicon microneedles for minimally invasive biomedical devices. Biomed Microdevices 2000;2:295–303.
195. Bergveld P, et al. Microdialysis based lab-on-a chip, applying a generic MEMS technology. Comprehen Analy Chem 2003;39:625–663.
196. Carneheim C, Ståhle L. Microdialysis of lipophilic compounds: a methodological study. Pharmacol Toxicol 1991; 69:378–380.
197. Mueller M, et al. *In vivo* characterization of transdermal drug transport by microdialysis. J Controlled Release 1995;37: 49–57.
198. Trickler WJ, Miller DW. Use of osmotic agents in microdialysis studies to improve the recovery of macromolecules. J Pharmaceut Sci 2003;92:1419–1427.
199. Kurosaki Y, Nakamura S, Shiojiri Y, Kawasaki H. Lipomicrodialysis: a new microdialysis method for studying the pharmacokinetics of lipophilic substances. Biolog Pharmaceut Bull. 1998;21:194–196.
200. Khramov AN, Stenken JA. Enhanced microdialysis extraction efficiency of ibuprofen in vitro by facilitated transport with β-cyclodextrin. Analy Chem 1999;71:1257–1264.
201. Kjellstrom S, et al. Online coupling of microdialysis sampling with liquid chromatography for the determination of peptide and non-peptide leukotrienes. J Chromatog A 1998;823:489–496.
202. Rekharsky MV, Inoue Y. Complexation thermodynamics of cyclodextrins. Chem Rev 1998;98:1875–1917.
203. Stenken JA, Chen R, Yuan X. Influence of geometry and equilibrium chemistry on relative recovery during enhanced microdialysis. Anal Chim Acta 2001;436:21–29.
204. Khramov AN, Stenken JA. Enhanced microdialysis recovery of some tricyclic antidepressants and structurally related drugs by cyclodextrin-mediated transport. Analyst 1999; 124:1027–1033.
205. Ward KW, et al. Enhancement of in vitro and *in vivo* microdialysis recovery of SB-265123 using intralipid and encapsin as perfusates. Biopharmaceut Drug Disposition 2003;24:17–25.
206. Pettersson A, Markides K, Bergquist J. Enhanced microdialysis of neuropeptides. Acta Biochim Polon 2001;48: 1117–1120.

207. Pettersson A, et al. A feasibility study of solid supported enhanced microdialysis. Analy Chem 2004;76:1678–1682.
208. Ao X, Sellati TJ, Stenken JA. Enhanced microdialysis relative recovery of inflammatory cytokines using antibody-coated microspheres analyzed by flow cytometry. Anal Chem 2004;76:3777–3784.
209. Kissinger CB, Kissinger PT. Can preclinical ADMET-PK now be done more efficiently and effectively? Preclinica 2004;2: 319–323.
210. Olson DL, Lacey ME, Sweedler JV. High-resolution microcoil NMR for analysis of mass-limited, nanoliter samples. Anal Chem 1998;70:645–650.
211. Wolters AM, Jayawickrama DA, Sweedler JV. Microscale NMR. Curr Opin Chem Biol 2002;6:711–716.
212. Khandelwal P, et al. Studying rat brain Neurochem using nanoprobe NMR spectroscopy: A metabonomics approach. Analy Chem 2004;76:4123–4127.

Further Reading

The most comprehensive sources for microdialysis sampling are the book and the two separate journal issues shown below.

Robinson T, Justice JB, editors. Microdialysis in the Neurosciences. Amsterdam (The Nethernand): Elsevier; 1991.
Lunte CE *Anal Chim Acta* 1999;379:227–369.
Elmquist WF, Sawchuk RJ. Microdialysis sampling in drug delivery. *Adv Drug Del Res* 2000;45:123–307.

See also ELECTROPHORESIS; GLUCOSE SENSORS; HYDROCEPHALUS, TOOLS FOR DIAGNOSIS AND TREATMENT OF; PHARMACOKINETICS AND PHARMACODYNAMICS.

MICROFLUIDICS

GLENN M. WALKER
North Carolina State University
Raleigh, North Carolina

INTRODUCTION

Microfluidics is the study and application of fluids at the microscale. The most common definition of the microscale is that one or more device dimension be in the range of 1–1000 μm. For reference, the diameter of an average human head hair is ~150 μm, the average thickness of a human fingernail is 360 μm, and the diameter of a human red blood cell is ~7 μm. Miniaturization technology, originally developed by the microelectronics industry, has been used to create microscale fluid components and complete microfluidic systems with pumps, valves, and filters, incorporated onto single microchips have been demonstrated.

By applying the analogy of the microelectronics industry (i.e., continuously incorporating more features into smaller areas) a logical application of microfluidics is to create lab-on-a-chip (LOC) systems. Lab-on-a-chip systems, also known as micro-total-analysis systems (μTAS), incorporate the functionality of biology or chemistry laboratories onto a single microfabricated chip. Ideally, a LOC system would be able to execute all of the tasks routinely performed in a biology or chemistry laboratory, such as sample preconditioning, mixing, reaction, separation, and analysis. Labor- and time-intensive procedures would be reduced to instant results derived from a series of automated steps performed on a LOC.

Microscale fluid handling confers many advantages over traditional lab operations (1). First, fluid quantities ranging from picoliters to microliters are used, thus reducing the amount of sample required for tests. Second, the amount of time required to perform some analyses (e.g., capillary electrophoresis) is reduced to seconds, which means analyses can be conducted many times faster than with traditional methods. Third, devices can be manufactured using microfabrication technology, which translates into reduced cost per device; disposable LOC systems can easily be envisioned.

In general, microfluidic devices are in early stages of development and are most often found in academic research laboratories. However, the benefits of these systems have been exploited to develop new medical devices for clinical diagnostics and point-of-care testing. Commercial examples of devices that make use of LOC concepts are discussed at the end of this article.

THEORY

Fluid Mechanics

The term microfluidics encompasses both liquid and gas behavior at the microscale, even though in most applications the working fluid is a liquid. All of the concepts discussed here are directed toward liquids. Other works are available which provide information on gas behavior at the microscale (2).

Fluid behavior at the microscale is different from that commonly observed in everyday experiences at the macroscale, owing primarily to the very low Reynolds (Re) numbers of the flow regime plus the large surface area/volume (SAV) ratios of the flow domain. As a consequence, viscous forces and surface tension effects become dominant over fluid inertia, and transport phenomena are purely diffusive.

Fluid flow at the microscale is typically laminar. Fluid flows are classified based on their flow regime, which can be predicted with the Re number. The Re number is the ratio of inertial forces to viscous forces and can be calculated with the equation

$$\mathrm{Re} = \frac{\rho V D_h}{\mu} \tag{1}$$

where ρ is the fluid density, V is the characteristic fluid velocity, D_h is the hydraulic diameter of the microchannel, and μ is the fluid viscosity. Fully developed fluid flow in a channel of circular cross-section is considered laminar if the Re number is <2100. For Re numbers between 2100 and 2300, the flow is considered transitional: it shows signs of both laminar and turbulent flow. A Re number >2300 indicates turbulent flow.

Laminar flow is predictable in the sense that the trajectories of microscopic particles suspended in it can be accurately predicted (Fig. 1a). Particles suspended in a turbulent fluid flow behave chaotically and their position as a function of time cannot be accurately predicted

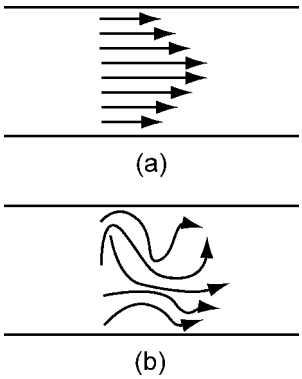


Figure 1. (a) Particles suspended in a laminar flow within a straight microchannel follow straight trajectories. (b) Particles suspended in a turbulent flow within a straight microchannel do not follow straight trajectories unless they are very close to the wall.

(Fig. 1b). Fluid flow that has a Re number <1 is also known as viscous flow, creeping flow, boundary-layer flow, or Stokes flow.

Low Re number flow is best visualized by imagining how honey (or any viscous substance) behaves when poured or stirred. For example, water flowing in microchannels will generally have a Re number <1. In this case, water will behave like a very viscous liquid (i.e., like honey). An important point to make here is that the properties of water do not change at the microscale; rather the microscale dimensions involved make the water *appear* more viscous than what we are accustomed to at the macroscale. An excellent description of low Re number environments has been given by Purcell (3). Very viscous fluid flows have certain characteristics: the flow is reversible, mixing is difficult, and flow separation does not occur (4).

Reversibility is the ability of a suspended particle in a fluid to retrace its path if the flow is reversed. This is a result of the minimal inertia (i.e., low Re number) present in fluid flows at the microscale. Figure 2 shows the path a suspended microscopic particle might take in forward and reverse flow.

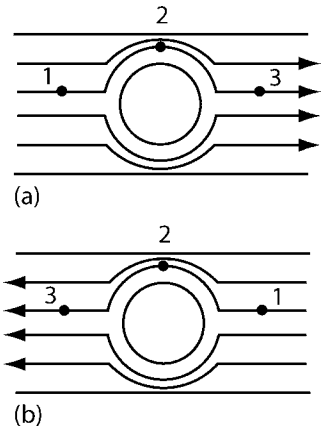


Figure 2. (a) A suspended particle in laminar flow around an obstacle in a microchannel. (b) If the flow is reversed, the particle will retrace the same path.

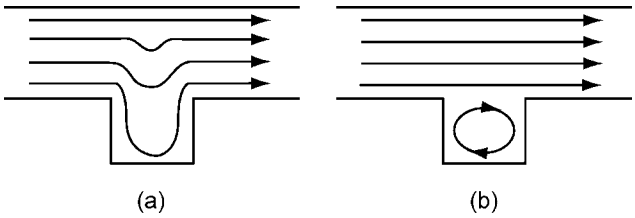


Figure 3. (a) At low velocities (low Re numbers), flow separation will not occur in microchannels. (b) At higher velocities (larger Re numbers), flow separation may become apparent.

A second characteristic of microscale fluid flow is a lack of flow separation. Flow separation is commonly observed in the form of vortices, which are recirculating flows separate from the main flow. Because of the low Re number environment, vortices usually will not form within microfluidic channels, as shown in Fig. 3a. Separation will only occur in flows wherein inertial forces are significant relative to viscous forces (Re >1). Figure 3b is a qualitative sketch of flow separation in a cavity.

The third characteristic of microscale fluid flows is inefficient mixing as a result of very low *Re* number flow, and thus negligible inertia. Low inertia means that stirring is not effective and that mixing must be accomplished by diffusion. At the macroscale, stirring minimizes the diffusion distances between two or more liquids by distributing "folds" of the liquids throughout the volume. Microscale methods of mixing have been developed that take advantage of the unique properties of the scale and improve the efficiency of mixing over simple diffusion; examples include using three-dimensional (3D) channel geometries, patterned channel surfaces, and pulsatile flow (5).

Figure 4 shows two streams flowing down a microchannel side-by-side. Because of the low Re number environment the streams will only mix by diffusion. If the flowrate is slow enough, the streams will eventually become uniformly mixed across the whole microchannel width.

The hydraulic diameter, D_h, of a microchannel is determined by its cross-sectional geometry and can be calculated with the equation

$$D_h = \frac{4A}{P} \tag{2}$$

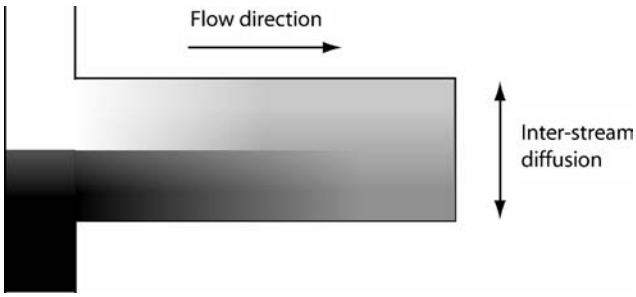


Figure 4. Two streams flowing in a microchannel will only mix by diffusion. Note that the concentration across the width of each half of the main microchannel is not constant as diffusional mixing progresses.

Table 1. Diffusion Coefficients for Biologically Important Molecules in Water[a]

Molecule	T, °C	D, $cm^2 \cdot s^{-1}$	Diffusion time, s
Cl^-	25	2.03×10^{-5}	0.02
O_2	18	2×10^{-5}	0.03
K^+	25	1.96×10^{-5}	0.03
Na^+	25	1.33×10^{-5}	0.04
Glucose	20	6×10^{-6}	0.08
Lactose	20	4.3×10^{-6}	0.12
Insulin	20	1.5×10^{-6}	0.33
Hemoglobin	20	6.3×10^{-7}	0.79
Urease	20	3.4×10^{-7}	1.47

[a]All values are from Ref. 6. The time for each particle to diffuse 10 μm is shown for comparison.

where A and P are the microchannel cross-sectional area and wetted perimeter, respectively. The hydraulic diameter is often used to calculate important flow characteristics for noncircular microchannel cross-sections.

At microscale dimensions diffusion is an effective mechanism for transporting molecules because of the relatively short distances involved. Particles diffuse from areas of high concentration to areas of low concentration and will eventually diffuse to uniform concentration throughout a given volume. The mean distance, d, a particle travels in a time, t, can be predicted with the equation

$$d^2 = 2Dt \tag{3}$$

where D is the diffusion coefficient of the particle. Diffusion times are proportional to the square of distance, which means that particles can diffuse across microscale distances within a particular medium in a matter of seconds. Table 1 lists representative molecules of biological significance and their diffusion coefficients.

The SAV ratios become very large at the microscale. Typical SAV ratios for macroscale containers such as Petri dishes or culture flasks are $\sim$10 cm^{-1}, while they are $\sim$800 cm^{-1} for microfluidic channels. Increased SAV ratios allow diffusion-limited processes, such as immunoassays to become much more efficient at the microscale because of the increased surface area available for binding. Large SAV ratios also allow rapid heat radiation from microscale fluid volumes and efficient gas exchange with the ambient atmosphere and fluid in microchannels (assuming the microchannel is made of a gas-permeable material). Enhanced gas transport is a critical ingredient for cell culture in microscale environments. One drawback of large SAV ratios is that evaporation becomes a significant problem.

The surface tension of a liquid becomes increasingly important at very small dimensions. To visualize this, think of a liquid surface as an elastic skin. If a slit were made in that skin, a certain amount of force per unit length would be required to hold the two sides of the slit together. The amount of force required to hold the two sides together is called the surface tension. Because the liquid surface is under tension, liquid confined by the surface (e.g., a raindrop) will experience an internal pressure. This pressure is called the Young–LaPlace pressure. Smaller fluid volumes result in larger SAV ratios, thus increasing the internal pressure. The pressure within a drop of liquid can be calculated with the formula

$$\Delta P = \frac{2\gamma}{R} \tag{4}$$

where γ is the liquid surface energy and R is the radius of the drop. At microscale dimensions, significant pressures can be created by surface tension. A common result of the pressure difference of an air/liquid interface is the capillary effect: a pressure difference across the interface propels liquid through a small diameter capillary or microchannel.

The capillary effect also depends on the contact angle of the microchannel surface. Hydrophobic surfaces (e.g., polymers) have contact angles $>90^\circ$ and hydrophilic surfaces (e.g., glass) have contact angles $<90^\circ$. Microfluidic devices with hydrophilic surfaces can be filled via capillary action. The pressure difference at an air–liquid interface within a microchannel with square cross-sectional area can be calculated with the formula

$$\Delta P = 2\gamma\left(\frac{\cos(\theta_c)}{W} + \frac{\cos(\theta_c)}{H}\right) \tag{5}$$

where W and H are the microchannel width and height, respectively, and θ_c is the contact angle of the liquid on the internal microchannel walls. Conversely, equation 5 gives the pressure required to force water into a hydrophobic microchannel of rectangular cross-section.

Microfluidic Modeling

Microscale fluid flow can be modeled from either a macroscopic or microscopic vantage point. Macroscopic modeling treats the fluid as a well-mixed volume while the microscopic view looks at how particles suspended in the fluid would behave under different flow conditions.

Macroscopic modeling, also called lumped modeling, uses conservation of mass to predict microfluidic system behavior. A pressure drop, ΔP, applied across a microchannel (or other conduit) with fluidic resistance Z, will induce a volumetric flow rate Q:

$$\Delta P = QZ \tag{6}$$

All microchannels have a fluidic resistance associated with them that depends on the geometry of the microchannel and the viscosity of the fluid. The fluidic resistance of a microchannel with a circular cross-section is given by

$$Z = \frac{8\mu L}{\pi R^4} \tag{7}$$

where μ is the fluid viscosity, L is the microchannel length, and R is the microchannel radius. The fluidic resistance of a microchannel with a rectangular cross-section is given by

$$Z = \frac{4\mu L}{ab^3} f\left(\frac{a}{b}\right)^{-1} \tag{8}$$

where $f(a/b)$ is calculated with the formula

$$f\left(\frac{a}{b}\right) = \frac{16}{3} - \frac{1024b}{a\pi^5}\sum_{n=0}^{\infty}\frac{\tanh ma}{(2n+1)^5} \tag{9}$$

When calculating the resistance of microchannels with rectangular cross-section, μ is the fluid viscosity, L is the microchannel length, $2a$ and $2b$ are the microchannel width and height, respectively, and m is calculated with

$$m = \frac{\pi(2n+1)}{2b} \tag{10}$$

If the aspect ratio of the microchannel is very small (i.e., $2b \ll 2a$) then the simplified formula

$$Z = \frac{3\mu L}{4ab^3} \tag{11}$$

can be used. The general rule of thumb is that equation 11 should be used for microchannels with $b/a < 0.1$. The resistance of other geometries can be found elsewhere (7).

In predicting microfluidic system behavior, the analogies to Kirchhoff's laws are used. The sum of pressure drops in a fluidic loop must be equal to zero; the total volumetric flowrate entering a node must be equal to the total volumetric flowrate leaving a node.

In contrast to the macroscopic view, microscopic modeling allows fluid behavior to be predicted. Specifically, the microscopic view allows the velocity profiles of a fluid flow to be calculated. Velocity profiles are plots that show the relative velocities of different portions of a fluid within a microchannel. Figure 1a is an example of a velocity profile.

The velocity of flow in a microchannel with circular cross-section varies radially and can be predicted with the formula

$$v(r) = \frac{R^2\Delta P}{4\mu L}\left(1 - \frac{r^2}{R^2}\right) \tag{12}$$

where μ is the fluid viscosity, L is the microchannel length, ΔP is the pressure drop, and R is the microchannel radius. The velocity profile of flow in a microchannel with rectangular cross-section varies along the height and width axes and can be predicted with the formula

$$v(x,y) = \frac{\Delta P}{2\mu L}\left(b^2 - y^2 - \frac{4}{b}\sum_{n=0}^{\infty}(-1)^n\frac{1}{m^3}\frac{\cos my\cosh mx}{\cosh ma}\right) \tag{13}$$

where μ is the fluid viscosity, L is the microchannel length, ΔP is the pressure drop, m is calculated from equation 10, and $2a$ and $2b$ are the microchannel width and height, respectively.

Microscopic modeling is performed when precise modeling of fluid behavior is needed. For example, cells attached to the wall of a microchannel might affect flow; modeling at the microscopic level would reveal any perturbations of the flow caused by the cell. In contrast, macroscopic modeling is performed when the behavior of the entire microfluidic system is needed. For example, fluid flow in many parallel microchannels might be required. Macroscopic modeling would reveal the relative flowrates through each microchannel and provide the microchannel dimensions needed to guarantee equal flow through each.

PUMPING FLUIDS

Fluids are pumped through microfluidic channels by creating gradients; the two most common types being pressure and electrical. Other types of gradients and their applications are discussed elsewhere (8).

Pressure gradients are the most common method used to pump fluid. Pressure is applied to one end of a microchannel which causes the fluid to flow down the pressure gradient. Common methods for creating a pressure gradient include pumps or gravity. Most methods for creating pressure-driven flow use macroscale pumps attached to the microfluidic device via tubing. Ideally, pumps should be incorporated on-chip to realize the ultimate vision for LOC devices. Many types of microfluidic pumps have been demonstrated and they presently constitute an active area of research (9).

Pressure-driven flow is attractive for use in microfluidics because it is easy to set up and model. Some drawbacks for using pressure-driven flow are sensitivity to bubbles, sensitivity to motion (via the tubing connecting pumps to the microfluidic device), and parabolic flow profiles. Shear stress is proportional to the pressure drop across a microchannel, which should be taken into account when manipulating cells.

The other common way to pump fluids is to use electrical gradients. This method of pumping fluid is only practical at the microscale level because of the large electric fields and SAV ratios required. Pumping via electric fields is called electrokinetic flow and is based on two phenomena: electrophoresis and electroosmosis. Electrophoresis operates on the principle that charged particles in an electric field will feel a force proportional to the field strength and their charge. The particles will move through the electric field toward the pole of opposite charge. Larger particles move slower than smaller particles because of the drag produced by moving through a fluid. Figure 5a shows an example of electrophoretic flow.

A charged particle in an electric field of strength E will travel with a velocity equal to

$$v = m_{\text{ep}}E \tag{14}$$

where m_{ep} is the electrophoretic mobility. Electrophoretic velocities are typically much smaller than the velocities caused by electroosmosis.

Electroosmosis will only function in the presence of an electric double layer at the surface of the microchannel.

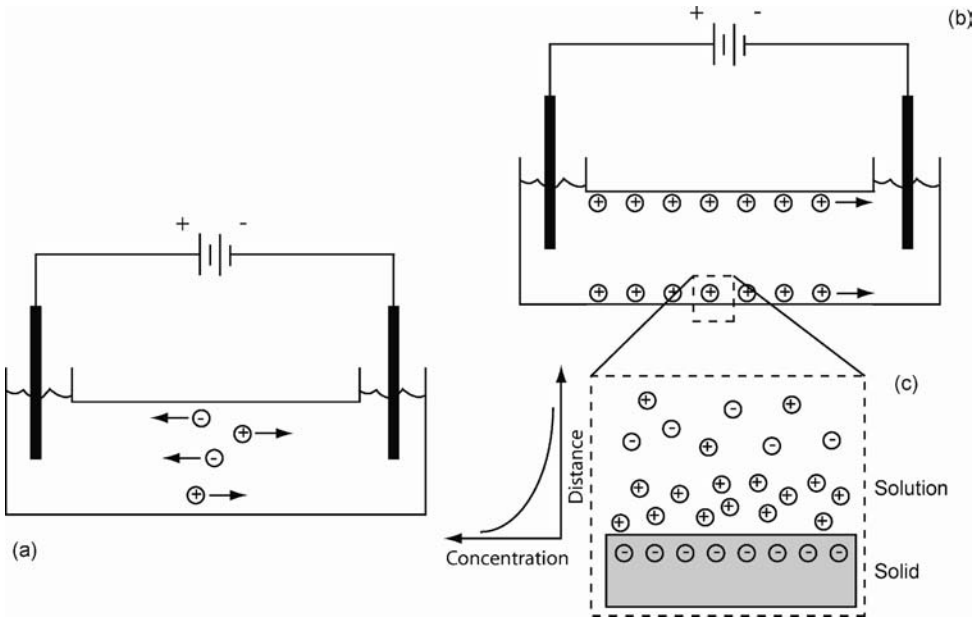


Figure 5. (a) Electrophoresis. Charged particles will move toward oppositely charged poles in an elec- tric field. (b) Electroosmosis. Charges lining a microchannel sur- face will move with an applied electric field, thus inducing bulk flow via momentum transfer within the liquid. (c) An electric double layer forms at charged microchannel surfaces; the layer thickness is called the Debye length.

An electric double layer forms at a charged surface when oppositely charged particles from an electrically neutral liquid gather at the surface. The thickness of the electric double layer is known as the Debye length; the concentration of charged particles at a surface falls off rapidly as a function of distance and is shown in Fig. 5c. When an electric field is applied across the length of the microchannel, the ions gathered at the microchannel surface begin to slide toward the oppositely charged pole, as shown in Fig. 5b. As the ions slide, they drag their neighbors within the bulk liquid, toward the middle of the microchannel. The friction between subsequent sliding layers of ions causes the bulk fluid to begin moving.

If the Debye length is much less than the characteristic dimensions of the microchannel, then the bulk fluid velocity can be predicted with the equation

$$v = m_{\text{eo}}E \tag{15}$$

where E is the electric field strength and m_{eo} is calculated with

$$m_{\text{eo}} = \frac{\varepsilon\zeta}{4\pi\mu} \tag{16}$$

where ε is the dielectric constant of the fluid, ζ is the zeta potential of the surface, and μ is the fluid viscosity.

Electrokinetic flow is attractive because it only requires the integration of electrodes in a microfluidic device, which is straightforward by microfabrication standards. Electrokinetic flow is therefore amenable to interfacing with electronic control circuitry. Electrokinetic flow also results in a blunt flow profile, which reduces the distortion of transported samples. Lastly, electrokinetic flow has very rapid response times, since the electrodes are integrated on chip, and are generally insensitive to movement off chip. Drawbacks to electrokinetic flow include fouling of the electrodes, which reduces electric field strength and therefore flowrate. Also, protein adsorption to microchannel surfaces affects the Debye layer, and in turn flow. Unintended side effects from electric fields on biological cells within the microfluidic device may also exist. Figure 6 shows the difference between the parabolic flow profile of pressure-driven flow and the blunt profile of electrokinetic flow.

Other methods of fluid flow based on surface tension, heat, and evaporation, have also been demonstrated. However, these methods have yet to be widely adopted and it is

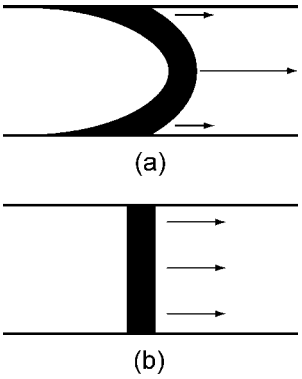


Figure 6. (a) Pressure-driven flow is parabolic; the middle of the stream moves faster than regions near the wall. (b) Electrokinetic flow is blunt; all parts of the stream move at equal velocity. Note that residual pressures can cause the profile to become slightly parabolic in the direction of the negative pressure gradient.

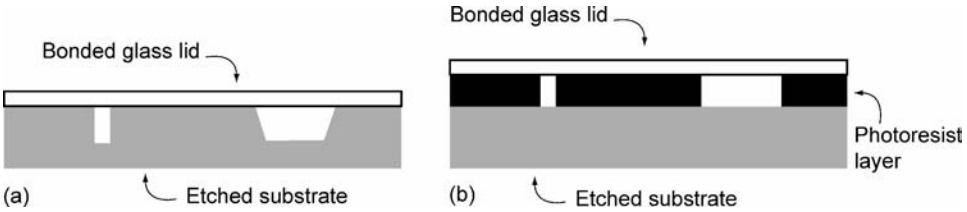


Figure 7. (a) Material is etched from a substrate and an enclosed structure is formed by bonding the etched substrate to a glass lid. (b) Walls of microchannels are built on top of a substrate and a glass lid is placed on top of the photoresist to form an enclosed structure.

unclear if they will prove more attractive than pressure-driven or electrokinetic flow.

FABRICATION

Fabricating microfluidic channels in traditional microfabrication materials, such as silicon and glass, can be achieved in two different ways. In the first, material is selectively removed, or etched, from a bulk substrate. The etched substrate is then bonded to another material (e.g., glass or silicon), which may have access holes or other features embedded in it. The result is an enclosed channel structure, as shown in Fig. 7a. The second method is to selectively add material to a substrate, and then bond another substrate to it. This method will also form enclosed channel structures as shown in Fig. 7b.

Photolithography is a fundamental part of all microfabrication (Fig. 8). Light is used to project patterns onto a photosensitive chemical, called a photoresist. The photoresist can be either positive or negative. Light chemically alters positive photoresist and makes it soluble in a developer. Negative photoresist is cross-linked by light, which makes it insoluble in developer. The patterned photoresist can be used as an etch mask for substrates, producing microchannels like those shown in Fig. 7a. Patterned photoresist can also be used in subsequent steps to direct the patterning of other materials that cannot be directly patterned with light. In-depth treatments of microfabrication techniques can be found elsewhere (10,11).

Polymers have recently become popular alternatives to traditional (e.g., silicon or glass) microfabrication materials. Polymers can be used to make microchannels by using the same methods mentioned previously for silicon and glass. Polymers also have the advantage that they can be molded that makes them a cheaper alternative (relative to silicon or glass) for mass production.

Polymer microfluidic devices can be created by molding, hot embossing, injection molding, photopolymerization, and laser ablation or laser cutting. An attractive method for fabricating polymer microfluidic devices is to use a process known as micromolding (12). In this process, photolithography is used to make a pattern of the microchannels (called a master). The photoresist provides a positive relief from which a polymer mold can be cast. The polymer is poured over the master and allowed to cure. The polymer mold is then peeled from the master and either placed on a substrate or incorporated into a multilayer device. The two advantages of this microfabrication method are (1) no special microfabrication equipment is required, and (2) many inexpensive copies of a microfluidic device can be rapidly manufactured.

Drawbacks to using polymers include leaching of material into microfluidic channels, solvent incompatibility,

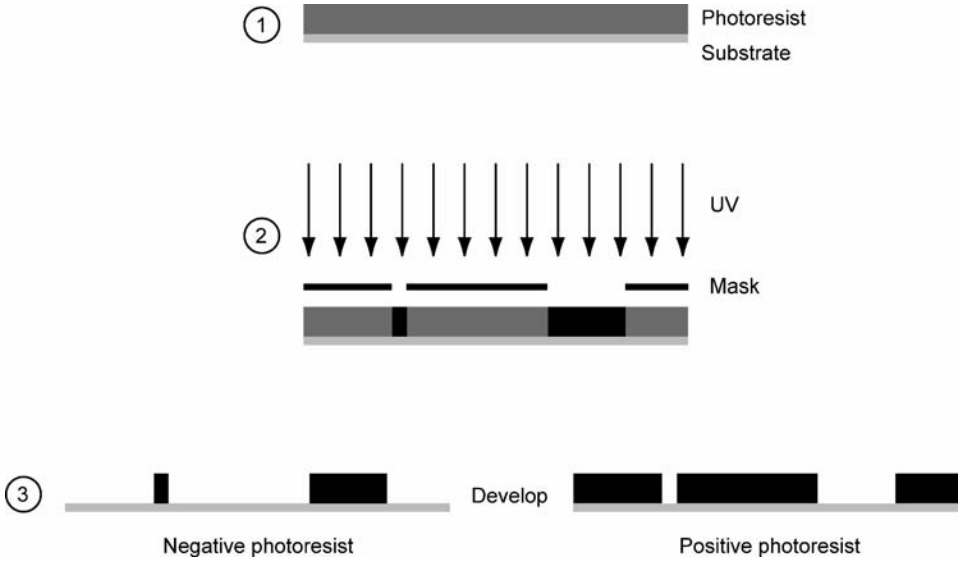


Figure 8. Photolithography requires an ultraviolet (UV) light source, a mask, and a photosensitive layer of material (i.e., photoresist). The photoresist is patterned with the UV light via a mask.

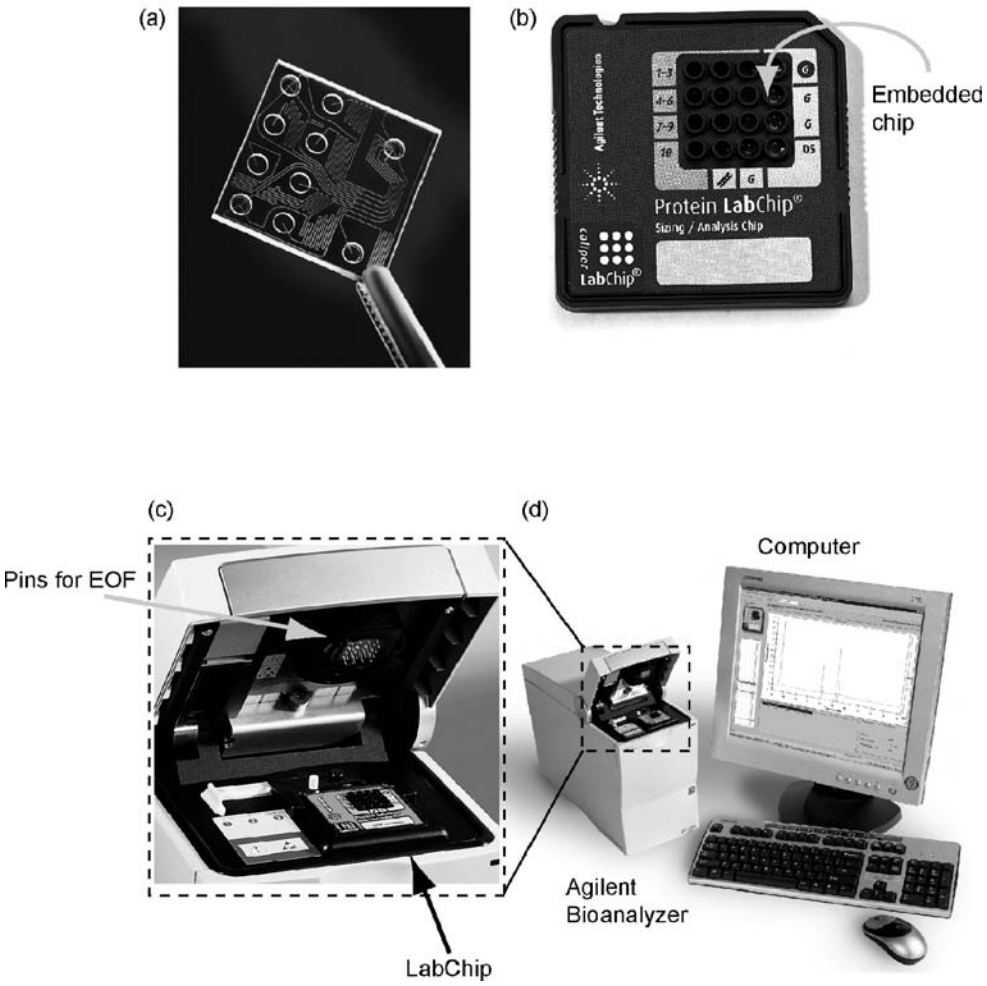


Figure 9. (a) The Bioanalyzer uses microfluidic chips etched from glass. (Images courtesy of Agilent Technologies). (b) The etched glass chips are encased in a plastic assembly that facilitates handling and limits contamination. Images courtesy of Agilent Technologies. (c) Samples are loaded onto the chip and the chip is loaded into the Bioanalyzer. (Images courtesy of Agilent Technologies). (d) The Bioanalyzer then performs an analysis on the samples. (Images courtesy of Agilent Technologies.)

and the ability of some substances to diffuse into the polymer. Also, surface treatments are occasionally required to make polymers compatible with electrokinetic flow; silicon or glass have an inherent surface charge that allows them to be used in electrokinetic flow applications.

BIOMEDICAL APPLICATIONS OF MICROFLUIDICS

Microfluidic concepts have already been incorporated in a variety of biomedical devices (13). One example of a microfluidic device now found in many biomedical labs is Agilent's Bioanalyzer. The Bioanalyzer system uses disposable chips etched in glass. The samples to be separated and reagents for the separation are loaded onto the chip via reservoirs. The chip is then placed in a reader, where electrokinetic flow is used to manipulate the samples in the reservoirs (Fig. 9).

Microfluidic capillary electrophoresis systems are becoming commonplace in laboratories. By using very small volumes for separation, joule heating from electrophoresis is rapidly radiated away from the gel. Efficient heat radiation allows larger voltages to be used which translates into faster separations. The shorter separation distances used also contribute to reduced analysis times. Microfluidic capillary electrophoresis systems allow DNA to be rapidly analyzed, and have highly reproducible results since the entire process is automated. Lastly, contamination is minimized because the devices are disposable.

Because of their small size, another attractive aspect of capillary electrophoresis (CE) systems is that they can be incorporated into LOC devices and made part of a complete system. An example that has been demonstrated in several research labs is a system that takes cells as inputs, lyses them, performs all necessary preprocessing, DNA amplification, and so on, and then performs the DNA separations, all with no human intervention (14).

Microfluidics are also being used in clinical devices; devices for hematology and disposable assays for point-of-care diagnostics are among those now being researched and brought to market. A handheld point-of-care device made by i-STAT is an example of a clinical microfluidic device (Fig. 10). The handheld device quantifies analytes in

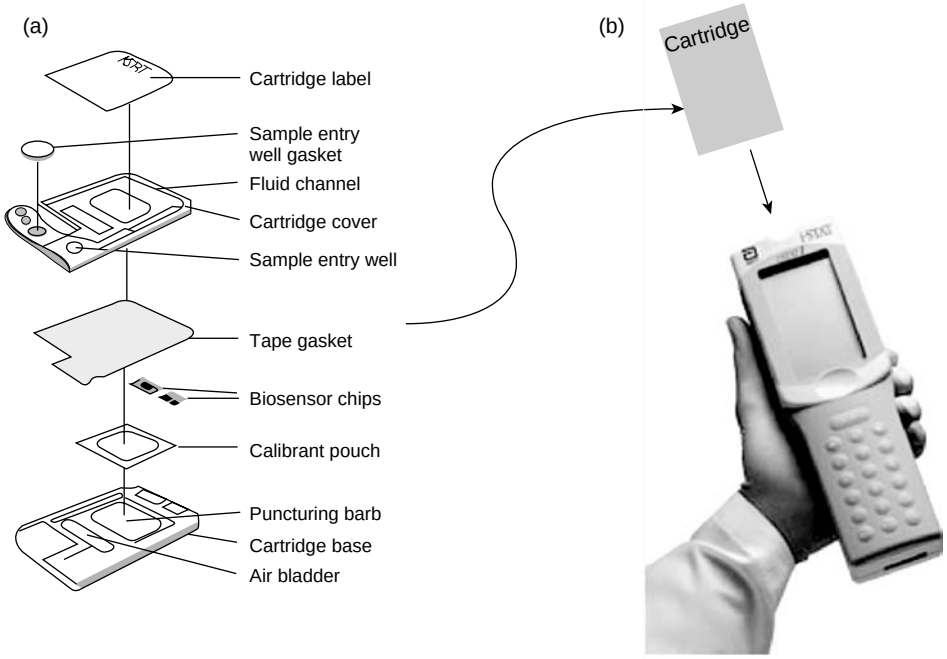


Figure 10. (a) The handheld point-of-care device manufactured by i-STAT performs analyses on blood samples contained in a disposable cartridge. Reprinted with permission from ACS (from Ref 15). Copyright 1998 American Chemical Sa. (b) Microfluidic cartridges are loaded with a sample and then plugged into the i-STAT handheld device. Image courtesy of Abbott Point-of-Care. The cartridge contains all necessary microfluidic control and sensing components that are then actuated by the handheld device.

blood samples that have been deposited on a disposable chip. The general procedure for operation is given below (15).

A patient's blood sample is deposited in a well on the disposable chip and then the well gasket is snapped shut (Fig. 10a). The microfluidic chip is then inserted into a handheld reader that performs an automated analysis of the blood sample, as shown in Fig. 10b. On-chip biosensors are automatically calibrated and checked for accuracy with an on-chip packet of calibrated solution. Once their accuracy has been determined, the calibration solution is flowed to the waste compartment. The blood sample is then flowed over the biosensors and the concentrations of different analytes are displayed on the handheld device screen within a few minutes. Diaphragm pumps are used to move the fluid. A variety of chips are available for different assays, including electrolytes and blood gases.

CONCLUSION

Microfluidics is the study and application of fluids at the microscale. Techniques used by the microelectronics industry have been adapted to facilitate the creation of micron-size channels capable of carrying fluid. The physical behavior of fluid at the microscale differs from behavior observed at the macroscale in everyday experience. The miniaturization of fluid handling has allowed LOC devices to be created in which all of the procedures of a traditional chemistry or biology lab are performed automatically in a single microfabricated chip. Lab-on-a-chip devices will allow new clinical and research tools to be developed.

BIBLIOGRAPHY

1. Brody JP, et al. Biotechnology at low Reynolds numbers. Biophys J 1996;71(6):3430–3441.
2. Karniadakis GE, Beskok A. Micro Flows. 2nd ed. New York: Springer-Verlag; 2001. p 360.
3. Purcell EM. Life at low Reynolds number. Am J Phys 1977;45(1):3–11.
4. Meldrum DR, Holl MR. Tech.Sight. Microfluidics. Microscale bioanalytical systems. Science 2002;297(5584):1197–1198.
5. Nguyen NT, Wu ZG. Micromixers—a review. J Micromech Microeng 2005;15(2):R1–R16.
6. Stein WD. Channels, Carriers, and Pumps: An Introduction to Membrane Transport. San Diego: Academic; 1990.
7. Shah RK, London AL. Laminar Flow Forced Convection in Ducts. New York: Academic; 1978.
8. Stone HA, Stroock AD, Ajdari A. Engineering flows in small devices: Microfluidics toward a lab-on-a-chip. Ann Rev Fluid Mech 2004;36:381–411.
9. Laser DJ, Santiago JG. A review of micropumps. J Micromech Microeng 2004;14(6):R35–R64.
10. Kovacs G. Micromachined Transducers Sourcebook. Boston: WCB McGraw-Hill; 1998.
11. Madou M. Fundamentals of Microfabrication. 2nd ed Boca Raton(FL): CRC Press; 2002.
12. McDonald J, et al. Fabrication of microfluidic systems in poly(dimethylsiloxane). Electrophoresis 2000;21(1):27–40.
13. Beebe DJ, Mensing GA, Walker GM. Physics and applications of microfluidics in biology. Ann Rev Biomed Eng 2002;4:261–286.
14. Waters L, et al. Microchip device for cell lysis, multiplex PCR, amplification, and electrophoretic sizing. Anal Chem 1998; 70(1):158–162.
15. Lauks IR. Microfabricated biosensors and microanalytical systems for blood analysis. Acc Chem Res 1998;31(5):317–324.

See also DRUG DELIVERY SYSTEMS; NANOPARTICLES

MICROPOWER FOR MEDICAL APPLICATIONS

JI YOON KANG
Korea Institute of Science and Technology
Seoul, Korea

INTRODUCTION

Generally, micropower is the local generation of electricity by small-scale generators, which locates the end point. As the recent development of the microelectromechanical system (MEMS), as well as CMOS electronics technology, has been reducing the size and cost of biomedical devices, the research of micropower became important for implantable biomedical devices since they require internal self-sustained power sources.

As for biomedical devices, micropower is an internal or external power source to supply energy for active devices, which replaces an organ's function or treats diseases. The examples of active implantable devices that consume energy are cardiac pacemakers, cardiac defibrillators, muscle stimulators, neurological stimulators, cochlear implants, and drug pumps (1). Hence, in this article the term micropower describes rather tiny power supplying devices for miniaturized sensors, actuators, and electric devices, whose size is $>$ or $< 1\ \text{cm}^3$.

The low power electrical actuator, such as a pacemaker or neuronal stimulator, requires tens of microwatts intermittently and their power source is usually a lithium iodine battery that lasts from 5–8 years. Usually, the stand-by current of a pacemaker is $\sim 1\ \mu\text{A}$ in waiting mode and its pulse current is ~ 6 mA. One example of a specification for a pacemaker pulse is in the range of 25 μJ ($\sim$11 mA at 2.2 V with a 1 ms discharge) and the capacity of the battery is 2 Ah at typical rating (2). The volume of a pacemaker is $\sim$20 mL and the volume occupied by the battery is about one-half of the total volume, 10 mL. Hence, the energy density (energy/volume) and reliability are important factors in the lifetime of the device.

An internal battery that is hermetically sealed in these devices can operate them with low power consumption; however, other implantable devices have radically different power requirements. Implantable cardioverter defibrillators demand the energy of 15–40 J providing six orders of magnitude larger than that of a pacemaker even though the pulses are less frequent. The current from a lithium silver vanadium battery is charged in an internal capacitor and the pulse of 1–2 A of current is fired. Electromechanical actuator like drug pumps demand more current than a lithium ion battery can deliver since it needs to overcome the high pressure in the chamber. The examples of drug pumps are insulin pumps, pain reliever, and an cerebrospinal fluid pump. A high current implantable battery should have low source impedance, such as lithium thionyl chloride, lithium carbon monofluoride, or lithium silver vanadium oxide.

Other future application are in wireless sensors, including physiological, chemical, or physical sensors embedded in an encapsulated environment. Miniaturized sensor consumes $< 100\ \mu\text{W}$ and a radio frequency (rf) transmitter consumed $\sim$10 mW intermittently. Since most of the power is used for communication, some research groups are developing several low power wireless transmission protocols (3,4). Hence, less power will be necessary for a sensor transmitter as technology evolves.

Some groups investigated more efficient and reliable batteries. To enhance their efficiency and lifetime, potential alternatives of the conventional batteries studied, such as a microfabricated battery, microfabricated fuel cell, and biofuel cell. Microelectrical system technology reduces the size of the primary battery and microfluidic galvanic cell (5), water activated microbatteries (6), and Li-ion microbatteries were demonstrated. The fuel cell attracts much attention since it has a high efficiency, high power, and low pollution rate. Research on fuel cells focus on high power applications, such as the automobile and portable electronics, like laptop computers and cellular phones (7). Recently, micromachining technologies employed as a method to fabricate miniature fuel cells (8–11) and their size became smaller than a button cell battery (12) with high power. Enzyme-based glucose/O_2 biofuel cells were reported by several groups (13) and a miniaturized all (14) was reported that is $< 1\ \text{mm}^3$, with although a power of $4.3\ \mu\text{W}$. Since glucose is available in all tissues and organs, it is advantageous in implanted medical sensortransmitters.

Although a primary battery as well as a rechargeable battery is an important tool that supplies reliable power to implant devices, the continuous power of the battery decreases with time, and after 5 years they will not supply enough power (15). Power delivery with an rf transmission can extend the lifetime and continuously deliver high power. In the case of a cochlear implant, an external device provides power and data through electromagnetic field coupling; however, it needs accurate positioning of the external device and may cause rf interference and heating of the tissue.

Therefore, many research groups are paying attention to microfabricated power scavenging devices as an auxiliary power source to recharge the battery with no external power. Ambient energy sources are body heat or movement of the human body. Piezoelectric material (16–18), capacitance change (19–24), and inductive coil (25–27) convert vibration or human motion into electrical energy. The generation by high frequency vibration is not suitable for implant devices since vibration of the human body is in the range of tens of hertz. Hence, energy conversion using vibration of low frequency can be integrated with implant devices.

Thermoelectric generators that convert temperature differences of the human body or combustion engine to electricity were reported (28,29). Another conversion method is the thermophotovoltaic power generator (30,31), which combines the combustion engine and solar cell. However, integrating a power generation device into an implant device has some limitations due to biocompatibility. Power generation with a high temperature like the combustion engine or thermovoltaic metoid, cannot be implemented in the inside of the human body due to the heating of tissues. Thermoelectric generation using body heat is promising for the implantable device. However, this article includes the review on the other portable power sources like the micro-

combustion engine, because those are also useful for a portable diagnosis system. The recent development of microfluidics and miniaturized biosensors enables point-of-care testing devices to be on the market in the near future. A microheat engine is highly efficient in energy conversion and will be useful as a portable medical equipment.

A good review article on micropower for wireless sensor networks was reported (20,32). It lists the candidates of portable power sources and compares the energy density for the battery and power density for a power generator. This article reviews the existing and potential micropower sources in view of medical applications from tiny sensors embedded in the human body to portable medical electronic devices.

MICROBATTERY

For a long time, the battery was a major energy source in portable electronic devices and it has evolved from Zn/MnO_2 to the Zn/air cell since 1900. Electrochemical power sources were developed in response to the needs of the flashlight, automotive starter, mobile electronics, and laptop computers. These days, there is a tremendous need for portable electronics demanding smaller, lighter, and longer lived batteries. Hence, many researches are on the way to making microfuel cell or microfabricated batteries using various kinds of electrochemical power. This section will briefly review microbatteries including fuel cells, biofuel cells, and micromachined batteries.

Microfuel Cell

Although the energy density of conventional batteries has been increasing from 500 $Wh \cdot L^{-1}$ for the Ni/Cd battery to 1500 $Wh \cdot L^{-1}$ for the Li/C–CoO_2 battery, there is a large jump for the air-cathode fuel cells of 4500 $Wh \cdot L^{-1}$ using hydrogen, hydrocarbon, and metals (7). The proton exchange membrane fuel cell (PEMFC) depicted in Fig. 1 is one of the promising techniques for fuel cell, which was implemented in miniaturization. Power sources for the automobile or the portable electronics of a huge market have been a main concern of fuel cell research because of the advantage of high efficiency and easy rechargeability. However, these days, in response to the demand of low power application, many miniaturized fuel cells are under study. Miniature fuel cells with a series path in a flipflop configuration was fabricated in a planar array and a four-cell prototype was produced, 40 $mW \cdot cm^{-1}$ (33). Yu (11) added microfluidic channels using anisotropic wet etching of silicon to the flipflop configuration and measured a peak power density of 190 $mW \cdot cm^{-2}$. They also reported that the flipflop fuel cell was constructed on printed circuit board (PCB) and that they achieved the area power density of >700 $mW \cdot cm^{-2}$. Wainright (12) fabricated on-board hydrogen storage with multiple coplanar fuel cells in series on ceramic substrates. They stored hydrogen in the form of the stabilized aqueous solutions of sodium borohydride ($NaBH_4$) or a metal hydride material, such as $LaAl_{0.3}Ni_{4.7}$. The energy density of microfabricated fuel cells did not exceed that of a Li/MnO_2 coin cell, but it had the advantage of higher power and compatibility with other microelectronic, microelectromechanical, or microfluidic devices. A polymer microfluidic channel was applied to a fuel cell using PDMS (10) and poly (methyl methacrylate) (PMMA) (8). They achieved a comparable area power density with a silicon based one. Whitesides and co-workers (9) also reported a membraneless vanadium redox fuel cell using a laminar flow property in a microfluidic channel with 200 $mW \cdot cm^{-2}$. As for the commercialization, MTI microfuel cells and Manhattan Scientifics are making portable fuel cells using direct methanol fuel cell (DMFC) and Medis technology announced direct liquid fuel cell (DLFC) for handheld devices. Miniaturized fuel cells are promising for implantable medical devices with a high power requirement and it has the advantage of a longer lifetime and less charging time than a conventional battery.

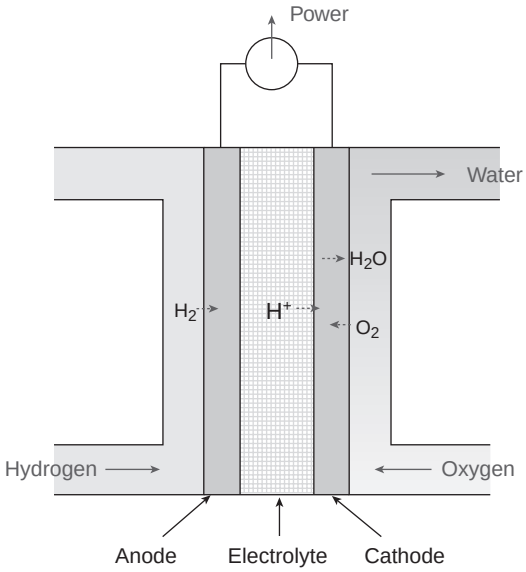


Figure 1. Schematic of a PEM fuel cell.

However, PEM fuel cells do not fit well with the implantable microdevices with high power and a long life application because the refill of a hydrogen fuel cell is not easy when it is sealed in the human body. As an alternative fuel cell, the biofuel cell is a strong candidate for a low power embedded device like a microbiosensor. Although a biofuel cell is not capable of high power, the easy availability of fuel (glucose) gives it a long operation time. Enzyme-based glucose/O_2 biofuel cells were studied since 1980s (13) and Heller and co-worker (34) demonstrated the feasibility of a membraneless miniature biofuel cell as an implanted micropower source. Its chemical reaction is described in equations 1 and 2.

$$\text{glucose} \rightarrow \text{glucolactone} + 2H^+ + 2e^- \quad (1)$$

$$O_2 + 4H^+ + 3e^- \rightarrow 2H_2O \quad (2)$$

Some fuel cell components, such as case, membrane, ion conductive electrolyte, and plumbing was removed in the biofuel cell and its size became <1 mm^2. The power of the biofuel is 4.3 μW with 0.52 V and it is suitable for an implanted devices because of tiny volume and abundant glucose inside of human body. Moor et al. (35) developed a microfluidic chip based ethanol–oxygen biofuel cell, which

produced 18 μW with 0.34 V and is applicable to integrate the biofuel cell and the microfluidic chip.

Micromachined Battery

A microfabricated battery usually refers to a thin-film battery, however, recently some research groups are studying MEMS-based microbatteries. Integration of microbatteries with CMOS electronics or an MEMS device is easy and can be fabricated on a chip with a device. Since the microbatteries main concern is power output due to limitation of surface rather than the capacity, most research activities are focused on increasing power to out perform maintaining capacity.

As for thin-film batteries, Bates et al. (36) at Oak Ridge National Laboratory reported a thin-film secondary battery, which was made up of lithium and lithium ion. The thickness is tens of μm and the area is in the cm^2 range. Its continuous current output is 1 $mA \cdot cm^{-2}$. Pique et al. (37) constructed primary Zn–Ag_2O and secondary Li ion microbatteries in plane using laser direct-write with a capacity of 100 $\mu Ah \cdot cm^{-2}$.

Other than classical thin-film batteries, several micromachined MEMS batteries were developed. They try to integrate power sources with microelectronic circuits and microsensors. Prof. Lin at UC Berkeley proposed a water-activated battery with 1.86 mWh in the area of 12×12 mm for lab-on-a-chip application (6) that overcomes the corrosiveness of the micromachined batteries with sulfuric acid and hydrogen peroxide (38). Andres (5) devised a pump integrated with a micropower source, in which microfluidic galvanic cells supplied power to heat up the two-phase fluid for pumping. The capacity of the micromachined battery is lower than that of a thin-film battery yet, its application is restricted to the integrated power for a micromachined implantable device.

MICROPOWER GENERATOR

A micropower generator scavenges energy from devices. The energy sources are mechanical (vibration and human body movement), thermal (temperature difference), and solar energy. Thermoelectric devices convert the temperature difference to electricity and the photovoltaic cell collects solar energy. A MEMS-based power generator scavenges energy from the vibration in the mechanical structure or human body movements.

Thermoelectric Generator

Thermoelectric generation using the Seebeck effect was widely studied. This effect was discovered in 1821 by the physicist, Jonn Seebeck. It is the same phenomenon with a thermocouple where the temperature difference produces electricity or work. Although a thermoelectric power generator is an old technique and is commercially available in various sizes in the market, recent research focuses on making miniaturized low power thermoelectric microgenerators using microfabrication. Cost-effective fabrication technology was developed using electroplated structures with an epoxy film (28) and nanowire arrays by electrochemical deposition was implemented to improve thermoelectrical properties (29). Reportedly, several companies announced a thermoelectric generator using body heat. Applied Digital Solutions (39) announced a thermoelectric generator called ThermoLife, which produced a power of 49 μW from a temperature difference of 5 °C in 0.5 cm^2. Biophan technologies (40) also announced a biothermal power source as shown in Fig. 2 for implantable devices like a pacemaker and defibrillator (41). They aim to produce 100 μW at 3 V with 1 °C temperature difference. A different scheme converting thermal energy to electricity is piezoelectric generator actuated by thermal expansion of two-phase working fluid (42). Although they produced up to 56 μW at its resonance frequency of 370 Hz, a practical application needs the careful design of a heat-transfer mechanism because the temperature is required to oscillate at a resonance frequency of structure. A thermoelectric generator has a well-established technology and its operation is relatively stable; however, note that the temperature difference inside the human body is < 1 °C. The temperature gradient is maximum at the skin surface and will limit the location of the thermoelectric power generator.

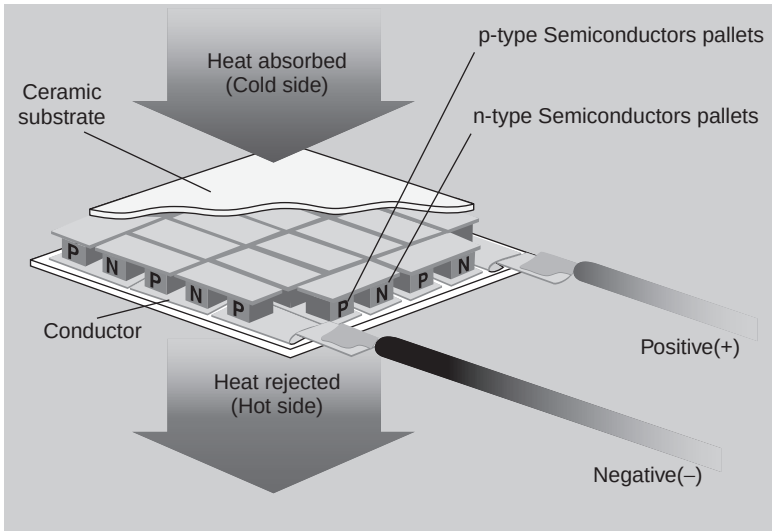


Figure 2. Principles of thermoelectric power source (Biophan).

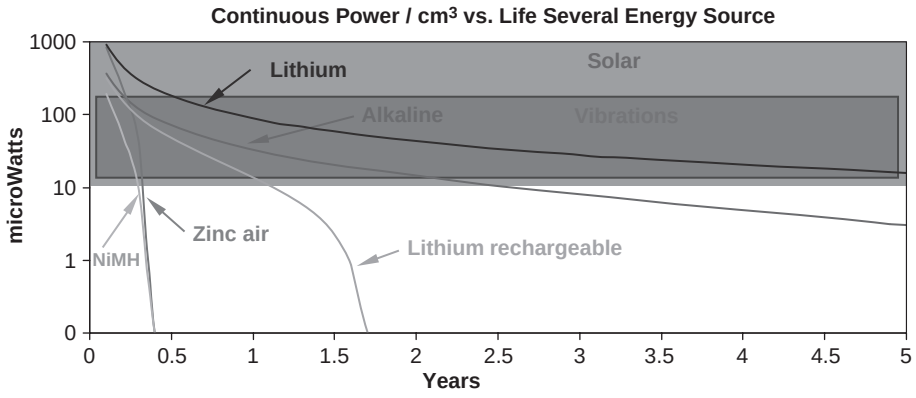


Figure 3. Comparison of power density from vibrations, solar and batteries (21).

Power Generation with Ambient Vibration

Conversion of a mechanical vibration to electric power has been studied from the mid 1990s (43) using MEMS, while thermal and solar energy were exploited to generate electricity long ago. They changed vibration to electrical energy using piezoelectric, electromagnetic, and electrostatic generations.

The frequency of vibration in an ambient environment ranges from 60 to 400 Hz and for the microwave oven the acceleration was 2.25 $\mathrm{m}\cdot\mathrm{s}^{-2}$ at a resonance frequency of 120 Hz (15). Shad (21) suggested a graph comparing the power density of power scavenging and batteries as a function of time (Fig. 3). Power density in the vibration of machining center or microwave oven becomes larger than conventional batteries after 3 or 4 years, which means the waste vibration energy is not negligible. Williams and Yates (43) presented a general model in equation 3 for the power of external vibration as depicted in Fig. 4, when a mass is moved at a resonant frequency of ω_{n}.

$$P = \frac{\zeta_t m Y_{\mathrm{o}}^2 \omega_{\mathrm{n}}^3}{4(\zeta_{\mathrm{t}} + \zeta_{\mathrm{o}})^2}\zeta \tag{3}$$

where m is mass, and Y_{o} is the maximum extent that the mass can move, ζ_{t} and ζ_{o} denote a damping factor both in the transducer structure and in the environment (e.g., air).

Electromagnet Conversion. Motion between the inductor and the permanent magnet induces an electromagnetic current in the inductor coil, as shown in Fig. 5 (25). The induced voltage, V, in the coil is given by equation 4.

$$V = NBl\frac{\omega_{\mathrm{n}}}{2\zeta} \tag{4}$$

where N is the number of turns in the coil, B is the strength of the magnetic field, l is the length of coil, and z is the displacement of the magnet in the coil. This type of generator was fabricated with laser micromachining by Ching et al. (27) in 1 cm^3 volume and generated a 4.4 V peak-to-peak with a maximum rms power of 830 μW. Glynne-Jones et al. (44) at the University of Southampton, derived 157 μW on average new car engine. Perpetuum Ltd., a spin-off company from the University of Southampton, produced an electromechanical microgenerator. That generated up to 4 mW and its operation frequency was 30–350 Hz. The vibration amplitude was 200 μm with 60–110 Hz and demonstrated a wireless temperature sensor transmitter system. This result showed electromagnetic conversion is feasible in low frequency vibration and is promising if it is compatible with silicon micromachining.

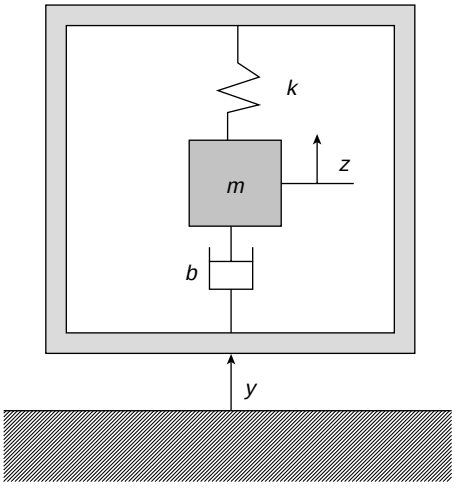


Figure 4. Schematic analysis for mechanical movement of a power generator with external vibration.

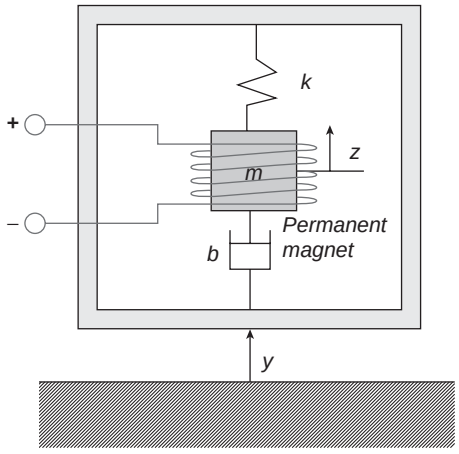


Figure 5. Schematic of an electromagnetic conversion device.

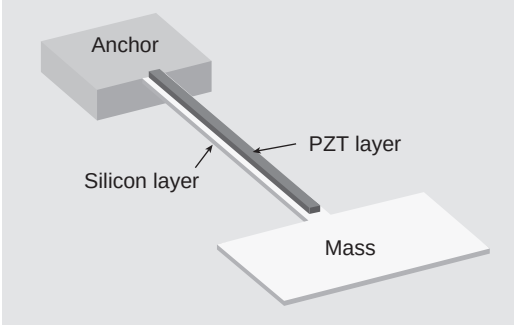


Figure 6. Structure of a piezoelectric cantilever for power generation.

Piezoelectric Conversion. The piezoelectric effect states that the deformation in the material produces an electrical charge due to the separation of charge within crystal structures. The most widely used piezoelectric material is PZT (lead zirconate titanate) in ceramic materials and PVDF [poly (vinylidene fluoride)] in polymers. Several groups are studying the piezoelectric cantilever with seismic mass (Fig. 6). The constitutive equations of piezoelectric materials are expressed in equation 5.

$$\sigma = Y(\delta - d_{31}E) \qquad D = d_{31}\sigma + \varepsilon E \tag{5}$$

where σ is the mechanical stress, δ is the mechanical strain, Y is Young's modulus, d_{31} is the piezoelectric strain coefficient, D is the charge density, E is the electric field, and ε is the dielectric constant of the piezoelectric material. The piezoelectric coefficient links the mechanical stress–strain to the electrical charge equation. If the circuit is open ($D = 0$), the voltage across the piezoelectric layer is described in equation 6.

$$V = \frac{-d_{31}\sigma}{\varepsilon} t_{\text{piezo}} \tag{6}$$

where t_{piezo} is the thickness of the piezoelectric layer. The charge collected on the electrode is integrated on the area of the surface with no load condition as in equation 7

$$Q = \int D\, dA = \int d_{31}\sigma\, dA \tag{7}$$

When the impedance in the load circuit is pure resistance, the time-averaged power can be derived in Ref. 17 with the geometry of a cantilever. White and co-workers (16) presented a thick-film PZT generator, and the maximum power is $\sim 2\ \mu$W. According to the analysis of Lu et al. (17), a 5 mm long PZT cantilever can generate $>100\ \mu$W with the amplitude of $>20\ \mu$m at ~ 3 kHz resonance. Roundy (15) demonstrated a piezoelectric converter of $1\,\text{cm}^3$ in volume, that is 1.75 cm in length. It generated 200 μW and is driven with vibrations of 2.25 $\text{m}\cdot\text{s}^{-2}$ at 120 Hz. A microfabricated PZT cantilever generator driven by a bubble was studied by Kang et al. (44), and a few picowatt was generated with one tiny cantilever at 30 Hz and tens of μW is expected on a $1\,\text{cm}^2$ surface (46). If the design, material, and fabrication are optimized, piezoelectric powergeneration will produce hundreds of microwatts with a volume of $1\,\text{cm}^3$.

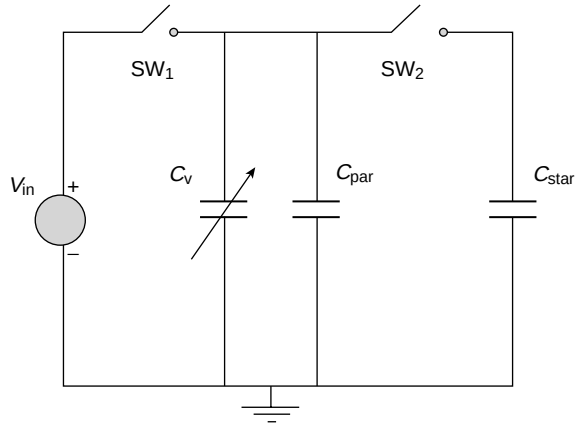


Figure 7. Circuit representation for an electrostatic converter (15).

Electrostatic Conversion. The electrical energy in a capacitor is given in equation 8.

$$E = \frac{1}{2}QV = \frac{1}{2}C_{\text{v}}V^2 = \frac{1}{2}\frac{Q^2}{C_{\text{v}}} \tag{8}$$

When the charge, Q, is constant, if the variable capacitance, C_{v}, is decreased the total energy E in the capacitor will increase. The MEMS structure can change the capacitance C_{v} with an external vibration, and stored energy in the capacitor transfers to energy storage. In the beginning, the external power source V_{in} initiates the charging process as in Fig. 7 (15). When C_{v} is maximum, SW1 is closed and the variable capacitance C_{v} is charged. While vibration changes capacitance C_{v}, all switched are open. When C_{v} reaches a minimum, SW2 is turned on and the energy in C_{v} is transferred to a storage capacitance C_{stor}. The disadvantage of electrostatic conversion is that it needs an external voltage source and switching circuit. The voltage across the storing capacitance is given in equation 9.

$$E_{\text{stor}} = \frac{1}{2}(C_{\max} - C_{\min})V_{\max}V_{\text{in}} \qquad (\text{Ref. } 47) \tag{9}$$

where $V_{\max}$ is the maximum voltage across the capacitor C_{v}. Switching the circuit is realized using a diode and field effect transistor (FET) switch. Meninger et al. (47) made a comb-type variable capacitance with an in-plane overlap type with a 7 μm gap and a 500 μm depth using the 0.6 μm CMOS process. They produced a power of 8 μW with an ultralow power delay locked loop (DLL)-based system. Miao et al. (23) reported an out-of-plane variable capacitor with a gap closing type that varies from 100 pF to 1 pF. A periodic voltage output of 2.3 kV (10 Hz) was generated when the charging voltage was 26 V, which implies that a power of 24 μW (2.4 $\mu\text{J}\cdot\text{cycle}^{-1}$) can be produced. Mitchenson et al. analyzed architectures for vibration-driven micropower generators (26) and they fabricated a prototype of an electrostatic power generator producing 250 $\text{V}\cdot\text{cycle}^{-1}$ that corresponds to 0.3 $\mu\text{J}\cdot\text{cycle}^{-1}$ (22). Other studies demonstrated polymer capacitor (24) and a liquid rotor power generator with a variable permittivity producing 10 μW (19). Recent developments in electrostatic generators

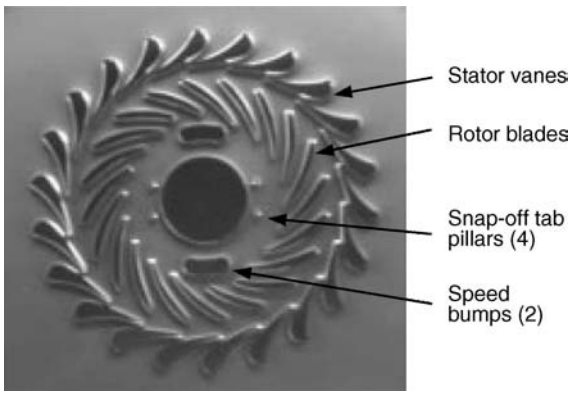


Figure 8. SEM of microturbine of MIT (52).

demonstrated that it is feasible and manufacturable with MEMS and that it has the advantage of a low frequency application like human body movement. However, the generated voltage is very high and should be managed for the implant application. Since it is compatible with the CMOS process and the variable capacitor is a well-established MEMS device, it is very promising as an integrated power generator with a sensor and transmitter.

Microheat Engine

A microheat engine is hard to be implanted in the human body, but it is promising as an external power source for portable medical equipment. A tiny internal combustion engine, made by precise machining, such as electrical discharge machining (EDM) or MEMS, may have a much higher density than a primary battery since the energy density of fossil fuel is $\sim 45\ \mathrm{MJ \cdot kg^{-1}}$, while that of Li ion batteries are at most $0.5\ \mathrm{MJ \cdot kg^{-1}}$ (48). Microturbine by EDM (48), Microrotors by deep reactive ion etching (DRIE) (49), heat engine (42) and reciprocating devices (50) was reported for electric power generation.

Several groups are working on a microheat engine, since fossil fuel offers a much higher energy density. Since the generated power is expected to generate 10–20 W, it is suitable for high power application. The Stirling engine (51), the reciprocal combustion engine (50), the Wankel motors, and gas turbines are reported. One of the first microengine projects was started at MIT (52) and the microfabricated turbine with a 4.2 mm diameter was illustrated in Fig. 8. Massachusetts Institute of Technology has been working on making microheat engines with a turbo charger and Georgia tech is collaborating with MIT on a magnetic generator (53). Allen and co-workers (54) at Georgia Institute of Technology generated a direct current (dc) electric power of 1.1 W with microfabricated windings at 120,000 rpm, although it was not integrated with a heat engine. Peirs et al. (48) made a microturbine by EDM and that tested to speeds of 160,000 rpm and produced a mechanical power of 28 W and an electrical power of 16 W. A miniaturized heat engine is still in the initial stage and no demonstration of power generation using a heat engine was reported. The heat engine would be very useful for high power applications such as portable analytical equipment. Another scheme of a heat engine is thermophotovoltaic power generation (30,31). They converted the heat radiation in a SiC microcombustor to electric energy using photovoltaic cells, which is $<1\ \mathrm{cm^2}$ and produced a power of 1.02 W with 2.28 V.

CONCLUSION

This article reviews micropower devices for medical implantable devices and portable medical device. Since the micropower devices have their own characteristics, there is no winner among them. When a selecting a micropower system for a specific application, one should consider the energy capacity, power, volume, voltage, and compatibility of fabrication with microelectronic device.

Currently, primary or secondary batteries with external power transmission are a main power storage for a low power implanted device. As the application of implant devices is diversified and requires a high power output and longer lifetime, new battery like fuel cell, or biofuel will replace conventional battery system in the future. Furthermore, when power transmission through the skin is impossible due to the attenuation of transmission, integrated power generation device will be an alternative.

Most micropower generators are still in their infancy and they need much more study to be implemented in implant device. Research results on micropower generators showed only the feasibility of concept and their power is much less than the requirement. Hybrid micropower supplies (55) or integrated power system would be strong candidates for long battery life applications. The promising application of active implant device will be glucose sensor and the artificial pancreas for the treatment of diabetes and the ubiquitous bio- or environmental sensor network. Since there is strong demand in the market, it is believed that micropower system will be available in near future.

BIBLIOGRAPHY

1. Soykan O. Power sources for implantable medical devices. Business briefing: Medical device manufacturing and technology. 2002, p 76–79.
2. Mallela VS, Ilankumaran V, Rao NS. Trends in Cardiac Pacemaker Batteries. Indian Pacing Electrophysiol J 2004;4: 201–212.
3. Zhong LC, Shah R, Guo C, Rabaey J. An Ultra-Low Power and Distributed Access Protocol for Broadband Wireless Sensor Networks. Presented at IEEE Broadband Wireless Summit, Las Vegas (NV); 2001.
4. David Culler DE, Srivastava M. Overview of Sensor Networks. IEEE Comput Special Issue Sensor Networks 2004; 41–49.
5. Cardenas-Valencia A, et al. A microfluidic galvanic cell as an on-chip power source. Sensors Actuator B 2003;95:406–413.
6. Sammoura F, Lee Kb, Lin L. Water-activated disposable and long shelf life microbatteries. Sensors Actuators A 2004;111: 79–86.
7. Dyer CK. Fuel cells for portable applications. J Power Sources 2002;106:31–34.
8. Chan SH, Nguyen N-T, Xia Z, Wu Z. Development of a polymeric micro fuel cell containing laser-micromachined flow channels. J Micromech Microeng 2005;15:231–236.

9. Ferrigno R, et al. Membraneless Vanadium Redox Fuel Cell uisng Laminar Flow. J Am Chem Soc 2002;124: 12930–12931.
10. Shah K, Shin WC, Besser RS. Novel microfabrication approaches for directly patterning PEM fuel cell membranes. J Power Sources 2003;123:172–181.
11. Yu J, Cheng P, Maa Z, Yi B. Fabrication of a miniature twin-fuel-cell on silicon wafer. Electrochim Acta 2003;48: 1537–1541.
12. Wainright JS, Saniell RF, Liu CC, Lit M. Microfabricated fuel cells. Electrochem Acta 2003;48:2869–2877.
13. Barton SC, Gallaway J, Atanassov P. Enzymatic Biofuel Cells for Implantable and Microscale Devices. Chem Rev 2004;104: 4867–4886.
14. Heller A. Miniature biofuel cells. Phys Chem Chem Phys 2004;6:209–216.
15. Roundy SJ. Energy Scavenging for Wireless Sensor Nodes with a Focus on Vibration to Electricity Conversion. Mechanical Engineering. Berkeley: The University of California; 2003 p 287.
16. Glynne-Jones P, Beeby SP, White NM. Towards a piezoelectric vibration-powred microgenerator. IRR Proc-Sci Meas Technol 2001;148:68–72.
17. Lu F, Lee HP, Lim SP. Modeling and analysis of micro piezoelectric power generators for micro-electromechanical-systems applications. Smart Mat and Structure 2004;13: 57–63.
18. Shenck N, Paradiso JA. Energy Scavenging with Shoe-mounted piezoelectronics. IEEE Micro 2001;21:30–42.
19. Boland JS, Messenger JDM, Lo HW, Tai YC. Arrayed liquid rotor electret power generation. Presented at IEEE MEMS 2005, Miami(FL) 2005.
20. Roundy S, Steingart D, et al. Power Sources for Wireless Networks. Presented at Proc. 1st European Workshop on Wireless Sensor Networks (EWSN'04), Berlin(Germany). 2004.
21. Roundy S, Wright PK, Rabaey J. A study of low level vibrations as a power source for wireless sensor nodes. Comp Commun 2003;26:1131–1144.
22. Mitcheson PD, et al. MEMS electrostatic micropower generator for low frequency operation. Sensors and Actuator A 2004;115:523–529.
23. Miao P, Holmes AS, Yeatman EM, Green TC. Micro-Machined Variable Capacitors for Power Generaton. Presented at Electrostatics'03, Edinburgh(UK). 2003.
24. Arakawa Y, Suzuki Y, Kasagi N. Micro seismic power generator using electret polymer film. Presented at The Fourth International workshop on Micro and Nanotechnology for power generation and energy conversion applications Power MEMS 2004, Kyoto(Japan). 2004.
25. Li WJ, et al. A micromachined vibration-induced power generator for low power sensors of robotic systems. Presented at World Automation Congress: 8th International Symposium on Robotics with Applications, Hawaii 2000.
26. Mitcheson PD, Green TC, Yeatman EM, Holmes AS. Architectures for Vibration-Driven Micropower Generators. J Microelectomech Systems 2004;13:429–440.
27. Ching NNH, et al. A laser-micromachined multi-modal resonating power transducer for wireless sensing systems. Sensors Actuator A 2002;97:685–690.
28. Qu W, Plotner M, Fischer W-J. Microfabrication of thermoelectric generators on flexible foil substrates as a power source for autonomous microsystems. J Micromechan Microeng 2001; 11:146–152.
29. Wang W, Jia F, Huang Q, Zhang J. A new type of low power thermoelectric micro-generator fabricated by nanowire array thermoelectric material. Proc 22nd Int Conf Thermoelectrics 2003;682–684.
30. Wenming Y, et al. Effect of wall thickness of micro-combustor on the performance of micro-thermophotovoltaic power generators. Sensors Actuator A; in press, 2005.
31. Yang WM, et al. A prototype microthermophotovoltaic power generator. Appl Phys Lett 2004;84:3864–3866.
32. Pescovitz D. The power of small tech. Smalltimes 2002; 2.
33. Lee SJ, et al. Design and fabrication of a micro fuel cell array with flip-flop interconnection. J Power Sources 2002;112: 410–418.
34. Chen T, et al. A Miniature Biofuel Cell. J Am Chem Soc 2001;123:8630–8631.
35. Moore CM, Minteer SD, Martin RS. Microchip-based ethanol/oxygen biofuel cell. Lab Chip 2005;5:218–225.
36. Bates JB, et al. Thin-film lithium and lithium-ion batteries. Solid State Ionics 2000;135:33–45.
37. Pique A, et al. Rapid prototyping of micropower sources by laser direct write. Appl Phys A Mat Sci Proc 2004;79:783–786.
38. Lee KB, Lin L. Electrolyte based on-demand disposable microbattery. Presented at IEEE MEMS 2002, Las Vega. 2002.
39. Applied Digital solutions, www.adsl.com.
40. Biophan technologies. Available at www.biophan.com/biothermal.php.
41. MacDonald SG, Biothermal power source for implantable devices. US Patent 6,640,137, 2003.
42. Whalen S, et al. Design, Fabrication and testing of the P3 micro heat engine. Sensors Actuator A 2003;104:290–298.
43. Williams CB, Yates RB. Analysis of a micro-electric generator for microsystems. Sensors Actuator A 1996;52:8–11.
44. Glynne-Jones P, et al. An electromagnetic, vibration-powered generator for intelligent sensor systems. Sensors Actuators A 2004;110:344–349.
45. Kang J-Y, Kim H-J, Kim J-S, Kim T-S. Optimal design of piezoelectric cantilever for a micro power generator with microbubble. Presented at Microtechnologies in Medicine & Biology 2nd Annual International IEEE-EMB Special Topic Conference, Madison (WI). 2002.
46. Kang JY, Kim JS, Kim HY, Kim TS. Micro Power Generator with Piezoelectric Cantilever Driven By Micro Bubble. Sensors Actuator A submitted, 2005.
47. Meninger S, et al. Vibration-to-Electric Energy Conversion. IEEE Trans VLSI Systems 2001;9:64–76.
48. Peirs J, Reynaerts D, Verplaetsen F. A microturbine for electric power generation. Sensors Actuator A 2004;113: 86–93.
49. Miki N, Teo CJ, Ho LC, Zhang X. Enhancement of rotordynamic performance of high-speed micro-rotors for power MEMS applications by precision deep reactive ion etching. Sensors Actuator A 2003;104:263–267.
50. Lee DH, et al. Fabrication and test of a MEMS combustor and reciprocating device. J Micromech 2002;12:26–34.
51. Backhaus S, Swift GW. A thermoacustic Stirling heat engine. Nature (London) 1999;399:335–338.
52. Frechette LG, et al. Demonstration of a microfabricated high-speed turbine supported on gas bearings. Presented at Solid-state sensors and actuator workshop, Hilton head island, (SC). 2000.
53. Jacobson SA, et al. Progress toward a icrofabricated gas turbine generator for soldier portable power applications. Presented at 24th Army Science Conference, Orlando (FL). 2004.
54. Das S, et al. Multi-Watt electric power from a microfabricated permanent magnet generator. Presented at IEEE MEMS 2005, Miami(FL). 2005.
55. Harb J, LaFollete R, Selfridge R, Howell L. Microbatteries for self-sustained hybrid micropower supplies. J Power Sources 2002;104:46–51.

See also BIOTELEMETRY; COMMUNICATION DEVICES; MICROFLUIDICS.

MICROSCOPY AND SPECTROSCOPY, NEAR-FIELD

MIODRAG MICIC
MP Biomedicals LLC
Irvine, California

INTRODUCTION

The explosion of knowledge in life sciences is enabled by the ability to visualize beyond the capability of the human eye and by the capability to identify chemical compositions and the structure of matter. This was enabled by Levenhook's discovery of the new device for looking at the world of the small: the microscope. He found that by combining two lenses, it was possible to see much smaller objects than by the naked eye alone. This led to his subsequent discovery of the cell, which has spurred an explosion of knowledge in life science and medicine that continues at a dramatic rate of growth even today. Even 400 years after the Levenhook discovery, one of the first tools of choice for the visualization of small objects is the optical microscopy. What is known to a lesser extent is that the microscopic histochemical studies (i.e., staining of the tissues) with tissues and organelle-specific dyes in the late 1800s, initiated the modern pharmaceutical industry, when chemists and histologists alike envisioned an opportunity to selectively inhibit or kill pathogens by organic molecules in the same way organic dyes selectively label certain types of tissues, cells, and organelles. While the optical microscopy methods were able to uncover morphology and the structure and nature of the cells, they were faced with the ultimate physical limit of magnification, which is dictated by the spatial resolution limited by the diffraction limit. This limit was approximately the size of half of the wavelength of light used to perform the imaging.

The breakthrough in imaging small structures and further understanding the machinery of life and cell biology comes with the application of the deBroglie's postulate of particle-wave equality in order to use the electron beam with a shorter associated wavelength as an investigative imaging probe. Ruska's development of the first transmission electron microscope in late 1930s and the development of scanning electronic microscopies in the 1950s, opened the door to detailed investigations of the organization of subcellular assemblies, viruses, and even imaging of the individual biomolecules, at a resolution far beyond the diffraction limit of visible light. The rapid advances in the tools and techniques of ultramicroscopy, especially of scanning probe microscopies, which for the first time enabled routine molecular imaging, greatly contribute to enabling completely new multidisciplines, like nanoscience and nanotechnology, as well as an opening a door for an entirely new way of looking into the machinery of life.

While scanning probe microscopy in the 1990s allowed imaging at the unprecedented resolution of the unaltered samples; in general, it lacked the ability to uniquely identify the chemical composition of samples or unveil their physicochemical properties. For the investigation of chemical structures and fingerprinting the material composition, the tool of choice is optical spectroscopy. However, the problem with classical optical spectroscopic techniques is that it provides average, bulk results with no specific information linking certain morphological features with spectra. This can ultimately identify chemical composition and/or physicochemical properties. The ability of doing molecular fingerprinting and, in a raster pattern, subsequent molecular specific imaging at the nanoscale with the spatial resolution of modern ultramicroscopy techniques is the holy grail for many aspects of today's life sciences disciplines.

This goal is partially fulfilled with electron microscopy combined with energy-dispersive X-ray analysis (EDX), wherein semiquantitatively, it is possible to associate topographical structures with elemental composition. However, for most of the problems in life and materials sciences, simple knowledge of elemental composition is not sufficient, as it is necessary to identify molecular structure. Plus, the electron microscopy is, in most cases, a destructive method of analysis, since the sample needs to be prepared to be vacuum compatible, and be either electrically conductive, in the case of scanning electron microscopy, or have contrasts with heavier metals, in the case of transmission or scanning transmission electron microscopy. Furthermore, the physics of generating characteristic X rays (i.e., the minimum size of the excitation volume from which the signal is emerging) is in the tens of micrometers, thereby limiting elemental compositional analysis with spatial resolution only for the large structure in the tens-of-microns-sized range. The ideal technique will be one that will allow imaging of the unaltered sample, in a way similar to the way atomic force microscopy (AFM) allows, while at the same time providing a way for spectroscopic identification of the chemical structure.

There are several techniques currently in their infancy that promise spectroscopic probing with electromagnetic spectroscopic information carrier signals imposed over topography. They are near-field scanning optical microscopy, microthermal analysis, scanning nuclear magnetic resonance (NMR) microscopy, and scanning electron paramagnetic resonance (EPR) microscopy.

However, for solving any of the practical problems, it will be of great benefit that the ultrastructure's information probes are photons of visible, and near-infrared (IR) and ultraviolet (UV) light, as a great deal of both morphological (based on the photon's position and intensity/count) and compositional (based on adsorption, fluorescence, Raman shift, etc.) information can be simultaneously obtained as optical microscopy relies on light as an information carrier. This is due to the fact that the same photons, which are in standard imaging configurations used to generate images, carry much more information on composition and the various physical and chemical properties of the observed spot on the sample that can be extracted through different spectroscopic methods.

The technique that has evolved over the last decade and promises to fulfill the above-goals at the nanoscale level, is near-field scanning optical microscopy (NSOM or SNOM), which effectively breaks the physical limits imposed by the optical-diffraction-limited resolution by using the near-field evanescent waves and scanning mechanisms similar to those in the scanning probe microscopies.

THEORETICAL PRINCIPLES OF NSOM MICROSCOPY

Abbe's equation (Eq. 1) describes the resolution of the classical, far-field optics, (i.e., the minimum separations between two points that can be distinguished) (1). From this equation it is easy to conclude that the maximum attainable resolution in the far field is $\sim \frac{1}{2}$ of the applied wavelength, which means that the best optical microscope cannot be used to visualize details smaller than 200–400 nm.

$$p_{\min} = \frac{\lambda}{2n \sin \alpha'} \tag{1}$$

In Abbe's equation, n is the diffraction of the imaging index and α' is the aperture angle in the medium. While Abbe's equation describes the limiting resolution in the world of conventional optics, the Fourier optics approach can provide us with the same conclusion. Following Abbe's principle, Rayleigh (2) derived that the objects in a lens system in the far optical field are resolved only when the maximum of one pattern coincides with the first minimum of the neighboring features. What resulted was the discovery that the resolution criteria that describe the maximum resolution of the optical system based on the size of the numeric aperture was

$$d = 0.61\,\lambda/\mathrm{NA} \tag{2}$$

wherein λ is the applied wavelength and NA is the numerical aperture of the lens, again bringing the maximum theoretical resolution to $\sim$ 200 nm.

A similar observation can be derived from the Fourier formalism in optics. In Fourier optics, the information content embedded in the spatial frequency f, in the case when f is higher then $1/\lambda$, decays rapidly toward zero from the object and thereby, no data on the subwavelength features can be efficiently collected with standard far-field optics. However, it is well known that it is possible to receive an electromagnetic signal with antenna that is smaller in size than the wavelength. The solution of the Maxwellian equations that govern the behavior of electromagnetic radiation differs in the distance smaller than the wavelength than in the distance, much larger than the considered wavelengths. When the waves propagate within the distance much smaller than its wavelength, such situation is called the near field. The pragmatic definition of near-field optics will be a division of optics that deal with the elements of the subwavelength features scales, which are intended for passing the light through, from or near, to another element with subwavelength features positioned within the subwavelength distances. The spatial resolution in near-field optics depends on the feature's size and is limited to about one-half of the aperture size. Furthermore, the near-field system must be considered as a complete system consisting of two features (probe and sample in the case of microscopy) and the resolution of the system will be dependent on both sample and probe. Thus, it is impossible to speak of the unique or standard near-field resolution, as is done with a far-field instrument.

NEAR-FIELD IMAGING EQUIPMENT

The near-field scanning optical microscopy or scanning near-field optical microscopy (NSOM or SNOM) is a technique that enables users to work with standard optical tools that are integrated with scanning probe microscope (SPM) technology to obtain the optical image at a resolution in the range of tens of nanometers. This is quite comparable with the resolution of scanning electron or SPM. The integration of scanning probe and optical methods allows for the collection of optical information at resolutions well below the optical diffraction limit, which overlap real topography information obtained by scanning probe feedback. For spectroscopy applications, NSOM offers the potential for characterizing the spectroscopic signature of material on a submicron-to-nanometer scale, thereby affording new insights into nanoscopic structure and composition.

The principles of NSOM microscopy were theoretically founded by Synge in 1929 (2), and in his subsequent paper he described an imaginary device that closely resembles today's NSOM setup (3), including the use of piezoactuators. Due to technical difficulties at the time to implement such a device, the idea was forgotten until theoretician O'Keefe rediscovered the idea in 1956 (4). The first practical demonstration of the imaging of a structure smaller than the one-sixtieth of the applied wavelength was done in 1971 using the microwave in near-field scanning over the grid (5). The basic idea behind this method was to create evanescence, a standing wave, by light diffraction through an aperture that was much smaller than the wavelength, and then to use this evanescent light source to scan a sample in a raster-scan pattern in close proximity, and collect transmitted or reflected signal in the far field. The first demonstration of near-field optical imaging was implemented independently by Pohl (6) in 1982 and Lewis groups (7) in 1983, and described as an optical stethoscopy, in what is now considered the beginning of NSOM microscopy. The method grew rapidly during the 1990s and the trend is continuing to this day, as described in recent reviews (8–12). Furthermore, several companies are offering commercial instruments (13–16) that enable ordinary users, who are not inclined toward the instrumentation development, to apply NSOM in solving their research problems.

There are two fundamentally different ways to achieve near-field optical imaging. They are apertured-base and apertureless NSOMs with their principles of operation depicted in Fig. 1a and b (17). In the case of the apertured NSOM (Fig. 1a), the light passes through an aperture that is much smaller than the wavelength of applied light, and ultimately the resolution is defined by the size of the aperture. In the case of the apertureless NSOM (Fig. 1b), a sharp metallic tip is irradiated by a laser perpendicular (or as close as possible) to the tip along the axis. The irradiation excites the plasmons on the metallic surface of the tip and the field is concentrated by combining antenna and plasmonic effects at the top of the tip. The resolution of apertureless NSOM is thus defined by the size of the near-field excitation formed at the apex of the metallic tip, and is determined by tip sharpness, tip materials, and real and imaginary parts of the refraction index of the used metal.

The practical advantage of apertured NSOM is in its easy implementation. While the advantage of the apertureless

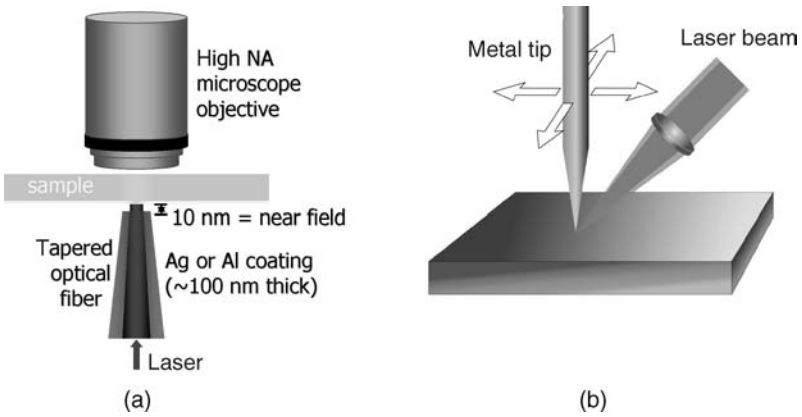


Figure 1. Two major principles of achieving the near-field scanning optical microscopy: (a) the aperture-based NSOM; (b) scatter-field, apertureless NSOM.

NSOM is theoretically a higher achievable resolution and possibly higher field strength, the technical difficulties in implementing the apertureless setup have not permitted those advantages to be realized. Thus, as of this writing, all of the existing commercial instruments are based on the apertured NSOM approach.

Regardless of the type of NSOM, the experimental setup always consists of a piezo *X–Y* scanner, whose role is to execute the raster-scan pattern scanning of the sample by the near-field probe; a *Z* piezo, whose role is to modulate the image by keeping the sample-tip distance; an optional, but for most cases necessary, noncontact modulation element, such as the tuning-fork for shear force feedback, or a piezo- or electromagnetic oscillator for AFM-like noncontact feedback; a near-field probe, which can be aperture or sharp tip; a laser light source; a far-field optical signal collection system, sample and sample holder; a system for coarse probe approach and sample–probe alignment; a scanner controller and computer for the image acquisition and reconstruction; and a vibration isolation system. In this manner, the NSOM most closely resembles the mechanism of the AFM, and almost all of the existing commercial setups, as well as many of the in-house, lab-made NSOMs, share these common components with the AFM and more general SPM platforms.

Piezo Scanner

The piezo scanner principle of work is based on the piezo effect, which is reversible internal stress induction within the crystal when exposed to the electric field (18). This stress induces crystal expansion. The piezo scanner in the NSOM is a more critical part than in the standard AFM. Ideally, it should be the perfect closed-loop scanner, due to stringent requirements of keeping the probe in a particular place during the raster scan, in order to achieve sufficient optical signal/noise ratio. Figure 2 depicts typical scanner configurations. The scanners are usually implemented in the form of the stacked piezo crystals (Fig. 2a), tube scanners (Fig. 2b), and bimorph (Fig. 2c) (19). Furthermore, scanners are often grouped into the so-called tripod configuration. The same material used for the SPM scanner, lead-zirconate-titanate ceramic (commonly referred as a PZT ceramic), is commonly used for the NSOM piezo scanner. The typical piezo-electric constant for PZT materials is about $-1.7\,\mathrm{V}\cdot\mathrm{nm}^{-1}$. However, in order to practically achieve the linearity over the whole range of scan, it is necessary to calibrate each individual scanner periodically to compensate for crystal nonlinearity, creeping, and drift. Those effects are further minimized by using active, real-time feedback, which can be implemented either through some form of the strain gauge, capacitance, or by optical means. The active feedback adjusts the voltage applied to the scanner to keep linearity and to secure the probe above the scanning position within the raster scan.

Optical Signal Acquisition System

The optical signal collection system is made up of optical and optoelectronic parts. The optical portion usually consists of the far-field microscope objective with a high NA lens. The tip-sample working distance, as well as the sample thickness and sample holder accessibility, define the maximum NA of the objective that can be used. Oil immersion objectives are used to enhance the NA by many times. Besides the objective, the collection system may contain filters, notch-filters polarizers, and beam splitters, depending on the particular configuration and imaging mode. The optoelectronic part of the collection system converts optical information to an electrical signal for further processing. It is usually either a highly sensitive photomultiplier tube (PMT), or, for ultimate sensitivity and single-photon counting, an avalanche photodiode detector (APD) array. For the PMT tube, the output signal can be either voltage or counts, and for the APD, it is only TTL (transistor–transistor logic) counts. The high sensitivity PMT tubes can satisfy most of the imaging requirements, however, for extremely weak signals, such as in single-molecular imaging or single-molecular spectroscopy, an APD detector is more desirable. Due care does need to be paid when using the APD detector, as overexposing the detector can damage it in an extremely short period of time. For the purpose of correlated experiments, many detectors are attached to the system. In the case of spectroscopy applications, the most commonly used dispersive detector is a highly sensitive CCD imaging camera, either solid-state or liquid-gas cooled. However, many setups use the other, more economical, wavelength or energy-dispersive detection systems, which are based on filters, prisms, or gratings in conjunction with either a PMT or APD detector.

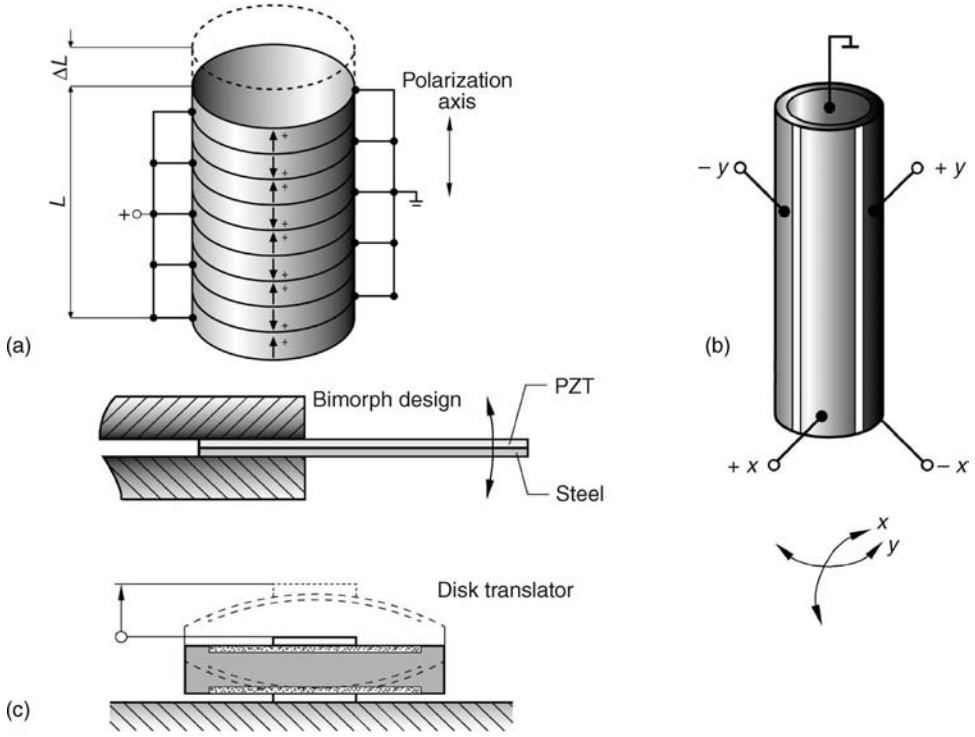


Figure 2. Schematics representation of the piezo actuators: (a) stacked piezo; (b) tube scanner; (c) bimorph scanner. [Courtesy PI (Physik Instru-mente) LP, www.pi.ws.

Light Sources for Near-Field Imaging and Spectroscopy

Illumination sources for near-field microscopy are always lasers. Selection criteria for the laser depends on the desired wavelength(s). The most economical are the solid-state lasers, followed by ion lasers, such as the Ar laser, which can selectively produce multiple wavelengths of light from 450 to 514 nm (17). Besides the two most common types, many setups use liquid lasers as well as optical parametric oscillators to produce specific wavelengths that are not available with a standard laser. The transduction path can consist of mirrors or single-mode optical fibers. The mirror-based path has better throughput, however, it is more complex to adjust and requires periodic readjustment. The optical-fiber-based transduction path has some higher attenuation then the mirror-based path, but it is very convenient for use, especially if it is implemented with the standard FC or similar connectors.

Microscope Head

The NSOM head usually consists of a probe holder, one or more piezo scanners, a system for coarse probe approach, and a case. The head is positioned at the top of the sample holder. The probe holder is directly attached to the Z direction piezo. A video system, which shows the image of the approaching tip and substrate, and a software- controlled stepper motor with micrometric screws provide coarse approach of the probe to the sample surface. When the probe tip is brought to close proximity to the sample, the fine approach mechanism is engaged. The fine approach mechanism is essentially a stepped mechanism wherein the probe is brought in a small piezo steps in the vertical direction to close the gap between the tip of the probe and sample substrate. The contact is achieved when the signal indicates the deflection of the probe in AFM-like setups, or, in shear-force mode, when its interaction with the sample reaches the user-prescribed "set point" voltage or current level. The determination of the appropriate set point varies for different samples and systems, and is more a result of art or tacit knowledge than an exact science. If the Z piezo is completely extended and contact has not been achieved, the piezo constricts to its neutral position and the stepper motor is activated to bring the probe to the approximate max extension distance of the piezo, and the process is repeated. Besides the Z piezo, sometimes the X–Y scanners can be positioned in the head.

Sample Holder/Stage

The sample holding stage can contain the X–Y piezo scanner, if it is not in the head. Its moving frame consists of micrometric screw positioners that push the sample holder in the X–Y direction under the probe, thus allowing sample "pan" operation. These screws can be manually or stepper-motor operated. The sample stage can be stand-alone, or it can be positioned at the top of an inverted, epi-fluorescence microscope. There are many advantages to having the NSOM sitting at the top of the standard inverted microscope. This configuration is able to combine far- and near-field microscopy, exploit the operation familiarity of the inverted fluorescence microscope, and deliver superior images to those acquired via a dedicated, stand-alone NSOM stage. However, a drawback of such a configuration is a larger mechanical circuit with a higher level of vibrational noise than in a dedicated system.

Controller

Controllers for NSOM are usually derived from the AFM/SPM controllers. In all of today's setups, they are digital.

The controller's role is to generate the high voltage signals necessary to feed the piezo scanner and move the probe in the raster scan pattern. It also controls the vertical, *Z* position of the probe via the PID control-loop model mechanism (proportional-integral-differential), and thus topographically modulates the signal; it maintains the non–contact feedback; it controls the coarse probe approach, and in some instances coarse sample positioning; and it acquires the signals coming from the probe (both optical and topographical) and forwards them to the computer for further processing. The controller usually consists of a series of analog-to-digital and digital-to-analog converters, precision operational amplifiers, and high voltage amplifiers. Due to the complexity of the tasks, often the controller is designed using high end digital signal processors and other high end embedded systems.

The role of the control software is to control the controller, acquire the image, and store it in some of editable and exportable format. Furthermore, the control software almost always possesses image processing capabilities, such as different image filters, Fourier transform, 3D representations and rendering, and so on. All of the commercially available NSOMs share the same software with their AFM/SPM "cousins". Many of the homemade systems, on the other hand, have software modified from existing commercial SPM/AFM controller software, or software that is independently written, as in cases where the users have designed the whole control electronics by themselves. Many times, the results are processed in third party software. For example, for image processing a very popular solution is to use shareware NIH Image software or its PC cousin, Scion Image, and for advanced applications, to use scripts written for IgroPro, LabView, MathLab, or for spectroscopy experiments, WinSpec. Use of higher level software like Igor Pro dramatically reduces development time of applications, as compared to the time required to write the script in C or C++ code.

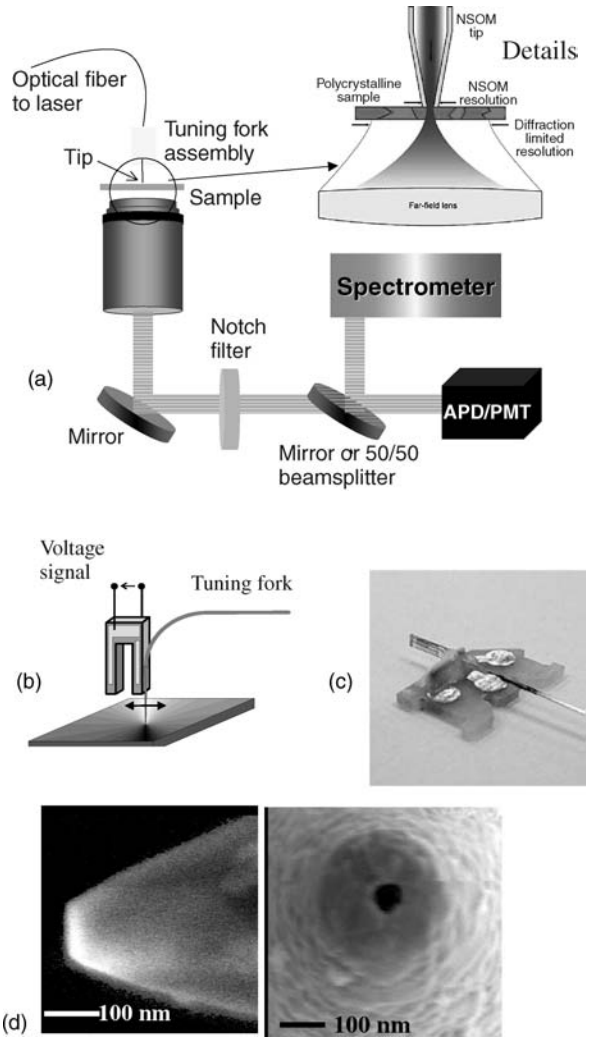


Figure 3. Example of the aperture-based, straight-fiber NSOM: (a) schematics representation of the scanning probe and collection; (b) schematics representation of the tapered fiber NSOM probe, attached to the tuning fork; (c) photography of the probe glued to the tuning fork; (d) SEM image of the end of the tapered, aluminum-coated fiber optic based NSOM probe. (Courtesy of Veeco Instruments Inc., Santa Barbara, CA.)

Apertured NSOM

The principle of the apertured NSOM is to use the aperture as a scanning probe. This is the first (5,6) and up-to-today most commonly implemented NSOM setup. In basic principle, the instrument consists of the *XYZ* piezo scanner(s) that moves the apertured probe over the raster scan pattern at the controlled aperture-sample height, and a non-contact feedback mechanism, the best-suited being the shear-force based one (20). An aperture in the tens of nanometer size can be formed by the tapering, heating, and pooling process borrowed from biophysics labs, where it is utilized for creating micropipettes (21) or for etching optical fibers (22). Additionally, as an aperture, it is possible to use the hollow cantilever (23).

Schematic representation of the instrument implementation, for both imaging and spectroscopy–hyperspectral imaging, is presented in Fig. 3a, configured for the most common, transmission mode (looking through the sample) operation. The laser light is coming from the optical fiber. Optical fiber is mounted on the tuning fork assembly, which is held onto the *Z*-piezo scanner and is constricted at the end to a tens of nanometer size range (see insert) and the evanescent field is formed at its aperture, as represented in the figure insert. The light is passing through the sample, interacting with sample matter, and is collected under the sample in the far-field with the high numerical aperture microscopy objective. Such a signal is further subjected to collection and processing. For the hyperspectral imaging or for the spectroscopy or spectral imaging purposes, the signal is first passed through the holographic notch filter, which eliminates excitation light and through the beam splitter is directed to the wavelength-dispersive detector, such as a CCD spectrometer, and to the summary, imaging detector, such as a PMT, or avalanche photodiode detector (APD). In the simplified setup, if the apparatus is used just for the optical imaging, the light is passed directly from the objective into the imaging detector, (i.e. APD or PMT).

Figure 3b schematically represents a typical apertured probe, mounted on the tuning fork. Micrographies at

Fig. 3c and d represent the frontal and lateral view of the metallic-coated, laser-pulled, tapered, fiber-based tip. An evanescent, standing wave is formed at the end of the aperture, and the size of the aperture approximately defines the optical resolution. Light is either brought through the aperture, as in transmission and reflection imaging mode, or collected through the aperture, as in the collection mode. The tip is scanned across the sample in a raster-scan pattern, and for imaging purposes; the signal is collected in the far field, either by sensitive photomultiplier tube, or by sensitive avalanche photo-diode counter.

The three distinctive different modes of operations of the aperture-based NSOM are illustrated in Fig. 4, which also graphically depicts the different kinds of information that can be extracted from the optical signal emanating from the sample. The origin of the optical contrast, as depicted in Fig. 4, can be due to topographic differences (different path length change the adsorption), material birefringence, reflectivity, sample extinction coefficients for the particular excitation wavelength, index of refraction, fluorescence emission properties, nonlinear spectroscopical properties of materials, and mechanical and magnetic stress in the sample. However, at the moment of this writing, for life sciences and biomedical applications, only transmitivitty, reflectivity and fluorescence properties are of significance. The mode that is the most useful for biological applications is the transmission mode or the "looking through" mode, and is most similar to classical biological microscopy In this case, as described above, light is brought through the fiber-based tip, the near field interacts with the sample, and the signal is collected in the far field as it passes through the sample. In this mode, it is possible to do transmission imaging, as well as fluorescence or other wavelength-resolved imaging, by application of adequate filters or wavelength-selective elements. In a reflection mode, which can be described as "looking on the surface mode", the near field interrogates the surface of the sample, and the scattered signal is collected in the far field. This mode allows imaging and spectroscopy of the nontransparent samples; however, the imaging efficiency is much lower than with transmission mode. In a collection mode, the light is passed either through the sample, or illuminated on the sample, and the signal is collected through the fiber in the near field. This mode is very burdensome to use, has low signal collection efficiency, and is used mainly in photonics research.

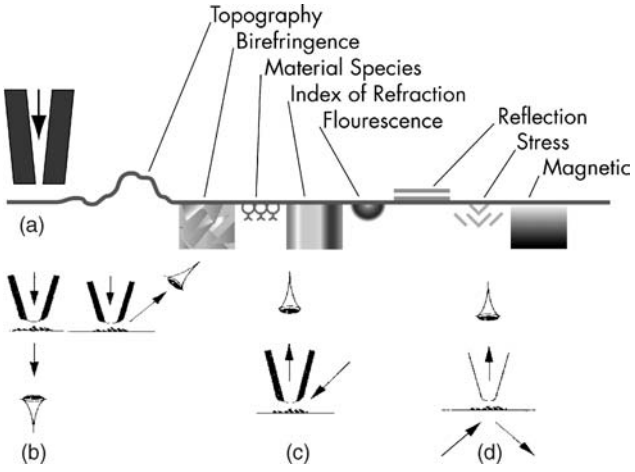


Figure 4. Typical apertured NSOM configuration: (a) illustration of information that is carried and can be extracted from the optical signal; (b) transmission and reflection mode NSOM; (c) collection mode NSOM; (d) total internal reflection or dark field mode. (Figure 3b–d adapted from Ref. 17.)

Besides these three modes, there are more exotic modes of operation, such as combined collection and illumination, where both sample illumination and resulting signal are passed and collected through the same probe; dark field imaging, where the probe tip is in close proximity to a sample that is illuminated from underneath with total internal reflection from the substrate, and wherein the probe acts as a second, tunneling prism. Besides pure optical operation modes, there are also a combined optomagnetic NSOM, which explores Kerr's effect (24); nanomass spectroscopy (25), where near field is used for ablation; and optoelectrochemical NSOM (26). The later two modes have a lot of potential applications in physiology with their ability to simultaneously image and record potential at subwavelength resolution.

To secure the probe in a near field, the aperture must be kept in close proximity to the sample with a distance much smaller than the applied wavelength. The fiber is kept at the nanometer-range distance from the sample by means of noncontact feedback. There are several ways of achieving feedback, mainly shear force, AFM-like normal force contact, and noncontact force feedback. The shear-force feedback provides gentler, lateral touching of the sample, thereby reducing the possibility of aperture contamination or tip–aperture mechanical failure.

In shear-force feedback (20), the fiber tip is oscillated laterally to the sample surface. The NSOM tip is rigidly premounted on a quartz tuning fork (Fig. 3b and c), which is a few millimeters in size. The tuning fork is mechanically vibrated at resonance frequency, usually in tens to hundreds of kilohertz, resulting in a few nanometers of lateral motion at the distal end of the NSOM tip. When the tip is in a close lateral proximity to the sample, the resonance frequency of the tip-tuning fork system is disturbed due to electrostatic, van der Waals, hydrogen bonding, and other kinds of attractive and/or repulsive interactions between the tip and the sample. This disturbance is read as an electrical signal that is processed, and the tip is moved accordingly in the vertical direction to achieve its preset resonance frequency, thus keeping the same distance from the sample.

Optical resolution, which is typically achieved by fiber-based apertured NSOM, is in the range of 50 nm, with maximum resolution being in the range of 20 nm. The improvement in tip fabrication procedures and in the control of the tip-sample separation distance will ultimately lead to better resolution. For apertured NSOM, fiberoptic or pipette-based tips are fabricated by constricting the core of the optical fiber to a 50–20 nm diameter. This is achieved by a heating–pulling method (21), wherein the fiber is transversally irradiated by CO_2 laser and simultaneously

stretched on the pipette puller until the fiber is broken, or by chemical etching (22) at the phase boundaries using the HF solution with oil on the top. To enhance the efficiency of the light transmission and to avoid the light leakage through the fiber shell, the probes are usually coated with either aluminum or silver. The coating is done by vacuum evaporation, and its role is to prevent light leaking out of the probe. The metallic coating is especially beneficial in the near-field surface-enhanced Raman spectroscopy.

Another way of performing apertured NSOM is by bending the fiber or pipette. In this case, the force feedback can be achieved either by shear force, or by AFM-like normal force in both contact or noncontact mode. The disadvantage of this approach is that such bended fibers are more vulnerable to mechanical failure, and if used for spectroscopy purposes, there may be problems with Raman scattering lines coming from the fiber shell materials.

The other way of achieving apertured NSOM imaging is by using the hollow AFM cantilevers (23). In this kind of setup, the light from the excitation laser is focused on the top aperture on the center of the hollow AFM tip, and the near field is formed at its bottom. The feedback mechanism used therein is the same as in noncontact AFM. Presently, the resolution (in the range of 100 nm) of such hollow-cantilevered-based apertured NSOMs is inferior to that of pulled-fiber-based NSOMs. Another disadvantage of the AFM-like force-feedback setup in NSOM applications is in that the AFM uses a laser beam to follow the bending of the cantilever. In NSOM, when many applications are counting the individual photons, the optical noise introduced by the AFM-like laser-based feedback may be several folds stronger than the signal. Considerable improvement is to be expected with piezo-actuated hollow cantilevers, which will avoid using laser feedback.

Apertureless NSOM

In the apertureless NSOM (Figs. 1b and 5), a sharp metallic tip is irradiated by the laser light orthogonally to the long tip axis, and the near-field excitation is scattered from the tip (27,28). The light scattering from the feature is much smaller than the applied wavelength, which also generates the strong evanescent field. The best strength of the scattering field is achieved if the excitation laser frequency corresponds to the surface-plasmon resonance of the metal from which the tip is made. Incoming beam scattering produces the evanescent field at the tip; however, the physics of the process is a combination of the near-field antenna effects and surface plasmon resonance. The laser induces the plasmons in the tip, which oscillates in parallel to the tip axis and amplifies the evanescent standing wave at the tip apex. The standing wave interacts with the sample and the signal is collected either in transmission or reflection mode in the far field. The tip is scanned across the sample in the same raster-pattern manner as with apertured NSOM. Feedback is provided in either the noncontact AFM manner, preferably with a tuning fork or some other nonlaser based Z-deflection feedback, and the signal collection is modulated by oscillating the probe in a vertical direction in order to avoid static and scattering artifacts. Furthermore, in modulated apertureless NSOM, the signal from the photodetector is also modulated with the same modulation signal source as a tip, in exactly the same frequency and phase (with possibilities of higher harmonics modulation. This is done in order to avoid inbound laser light nonnear-field induced scattering; static-scattering artifacts from sample features and to achieve optical signal acquisition always in a same sample-tip separation distance position.

In order to distinguish between the near-field scattering and inbound laser light, most of the apertureless NSOMs are used mainly for fluorescence, Raman, or for different nonlinear optical phenomena applications. Furthermore, because of the rapid decoy of the scattered field, modulation of the scanning probe, and control of the sample-tip separation distance in the apertureless configuration is much more critical than in the aperture-based NSOM.

Figure 5 is a schematic representation of a typical, homemade, apertureless NSOM setup; in this particular case used for fluorescence and fluorescence-lifetime (FLIM)

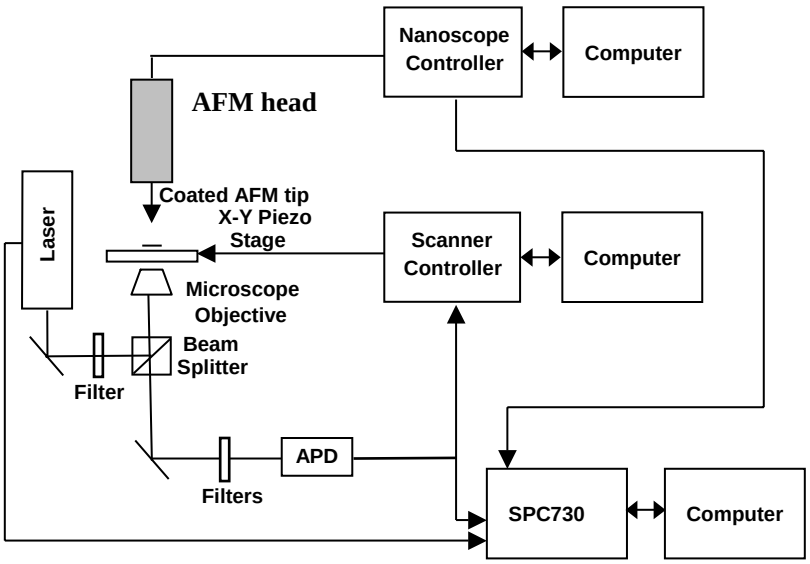


Figure 5. Schematic representation of the example of the apertureless NSOM. (Adapted from Ref. 29.)

near field imaging. It consists of a commercially available AFM head, mounted on the top of the inverted epi-fluorescence microscope, positioned at the optical table. The lateral irradiation of the tip, which is necessary in order to achieve the high intensity evanescent field generation, is produced by offsetting the position of the AFM tip in relationship to the high numerical apex microscope objective. In this particular case, the fluorescence imaging measurements were conducted in an inverted fluorescence-imaging microscope (Nikon Diaphot 300), the excitation light from a mode-locked YAG laser (Coherent Antares) at 532 nm wavelength, 10 ps pulse width, and 76 MHz repetition rate was focused on a diffraction-limited spot through an objective (Nikon 60X NA 1.4), and the emission from the sample was collected by the same objective, in the epi-fluorescence manner. The emission band-pass filters were HQ565/25 and D570/20 (Chroma Technology) to ensure that the excitation light and the feedback laser of AFM (650 nm) were both blocked. The emission was detected by an avalanche photodiode (APD) (Perkin Elmer, SPCM-AQR-15). The background photon counts with AFM feedback on were ~150 Hz. The sample cover slip was mounted on a closed-loop two-dimensional (2D) piezoelectric scanner (Polytec PI, P-731). The AFM (Veeco Instruments Inc, D3100) head and inverted microscope were coupled at an over–under position. The AFM tapping-mode tips used in this work are commercially available Si tips (Digital Instrument, OTESP7) coated with Au and Ag, by sputter coating. Image density of 128×128 pixels and scan rate of 1 Hz. As the quenching effect is highly distance dependent, the tip oscillation amplitude was reduced by reducing the driving voltage to the tip as much as possible without sacrificing image quality. Based on the force calibration curve, the tip oscillation amplitude during the imaging was estimated at $\sim$ 30 nm.

The sample-scanning confocal fluorescence image was recorded by a home-built computer control interface that counted the APD signal and raster-scanned the piezoelectric scanner. The fluorescence decay traces were recorded by a time-correlated single photon counting (TCSPC) module (Becker & Hickl SPC730, Germany). The start signal was from the APD and the stop signal was from synchronization of the YAG laser at one-half of the laser repetition rate. For the lifetime imaging mode, the TCSPC module reads the line-synchronization signal of the Digital Instrument Nanoscope IIIa controller to achieve a synchronized recording of the AFM signals and fluorescence signals.

Figure 6 depicts the FEM simulation (30) of near-field enhancement around a metallic tip positioned in close proximity to the sample and irradiated with a laser beam. Figure 6 shows the rapid dependence of the near-field excitation on the sample-tip separation, and emphasizes necessity of accurate, sub nanometer sample-tip separation control mechanism for any widespread, commercial applications. This is even more important for the Raman spectroscopy or hyperspectral imaging (or in this sense for any other, nonlinear optical applications), as the strength of the Raman emission is proportional to the fourth power of the strength of the electric field of applied light the small changes in the strength of the local near-field enhancement

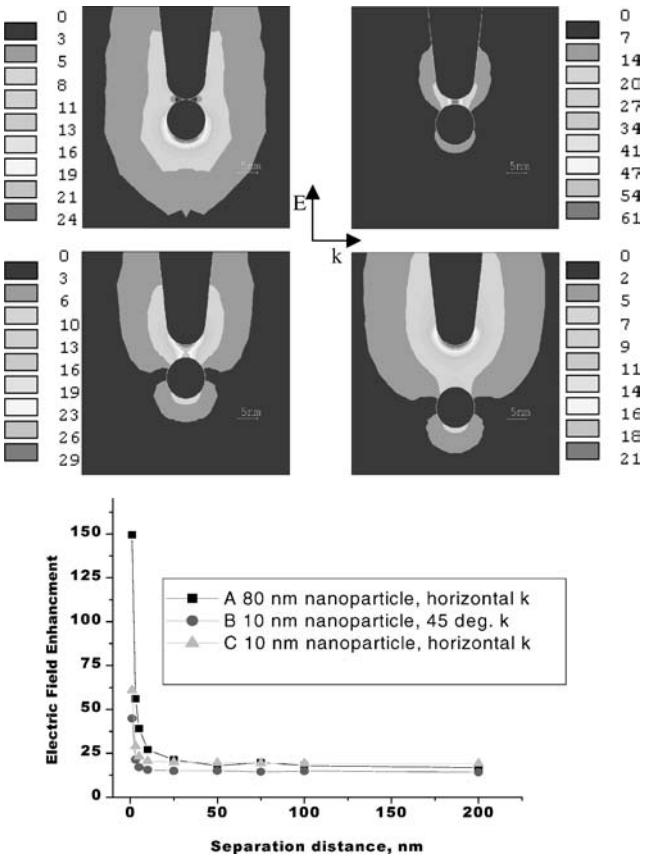


Figure 6. Finite-element methods (FEM) simulation of the electromagnetic field scattering and near field enhancement at the apex of the tip of the apertureless NSOM, and its behavior with change in tip-sample separation distance. (Adapted from Ref. 30.)

can produce the dramatic fluctuation in the strength of the optical signal.

Besides all of the difficulties in implementation, the advantage of the apertureless mode is in theoretically better resolution than the apertured mode, and in the higher surface-enhanced Raman signal, due to plasmon coupling, which makes this method, theoretically, an ideal molecular Raman nanoprobe, a "holy grail" for life scientists (31,32). The hyperspectral subwavelength Raman imaging is extremely important for further studies in system biology, proteomics, and metabolomics, as it is expected it will for the first time allow identification and spatial positioning of biomolecules within a cell, without the introduction of fluorescence labels. Aside from Raman imaging, this approach is expected to be better for the purpose of fluorescence lifetime imaging (FLIM) (29). The apertureless approach also has significant advantages because the surface plasmon enhancement is driven by the metallic tip and a higher local intensity of the scattered field.

However, to date, there is no commercial instrument available based on the apertureless NSOM principle, because of many technical problems, a significantly smaller photon flux, a low signal/noise ratio, and extreme signal dependance on the tip-sample separation distance. It is expected that with improvements in the control mechanism

for keeping the sample-tip separation in the subnanometer range, as well as improvements in laser positioning and further enhancement of the photodetector efficiencies, the apertureless NSOM will become more widespread, with the first commercial instruments expected to be introduced to the market in the near future.

The tip for the apertureless NSOM can be manufactured in three different ways. The simplest way is by electrochemically etching the metallic wire in the same way as for production of the STM tips. However, controlled and optimized shapes, which can provide more efficient field enhancement, can be better achieved by using the free ion bombardment (FIB) techniques for both tip growth and etching.

NSOM Operations

The typical operation of the NSOM consists of positioning the tip above the sample feature of interest, visually using the reflection and transmission video system, respectively. After executing a manual approach procedure, the tip is subsequently placed under PID control and is automatically maintained in the near-field region. The nonoptical, shear-force feedback relies on measuring the voltage generated by a quartz tuning fork onto which the NSOM tip is rigidly mounted, thus avoiding feedback laser. Having a feedback without the feedback laser is of great advantage, as avoidance of that voltage is a direct measure for the oscillation amplitude of the tip-tuning fork assembly, which varies with tip-to-sample distance over a range of $\sim$25 nm.

Unlike conventional AFM cantilever designs, the spring constant of the straight-fiber NSOM tip in the vertical direction is extremely high, thus avoiding damaging snap-to-contact. A feedback algorithm monitors the amplitude of the tuning fork by appropriately adjusting the tip-sample distance. Using this method, the NSOM tip is engaged and maintained within $\sim$5 nm of the surface in the near-field region throughout the NSOM scanning or spectroscopic measurements. Another advantage of using shear-force feedback is in the absence of a feedback laser, which is especially important in the low photon-count applications (e.g., in spectroscopy–hyperspectral imaging and single-molecular studies).

Other methods of maintaining the tip in the near field have not proven nearly as sensitive or reliable as tuning-fork-based shear-force feedback. Some methods originally developed for AFM applications may require actual surface contact and, consequently, possible surface or tip transformation, ultimately resulting in either damage or having to move the tip in and out of the near-field during data acquisition. The other disadvantage of AFM-like feedback is in the great technical difficulties to form a self-actuated, piezo-based AFM hollow tip, thus forcing the use of laser for feedback control. The AFM-like force feedback with pulled fiber tip requires a bent fiber, which is much less mechanically stable than a straight fiber. In addition, the bent fiber has problems associated with circular light paths and higher Raman scattering from glass–fiber substrates, interferences that carry especially negative consequences for near-field spectroscopy, as they can significantly increase the optical noise level.

The NSOM can be used both in air and in liquid. Most of the work to date has been done in air. While it has been demonstrated many times that the method can be successfully used in liquid operation, there are problems associated with the menisc force formed between the probe and liquid surface. For work in air, shear force is the superior method of feedback. However, for work in liquid, the AFM-like noncontact feedback has advantages. With further improvements in the fiber-based probes coating in the near future, it is expected that shear-force-based topographic imaging and feedback will become equal with the AFM-like noncontact-based topographic imaging in liquid.

Near-Field Spectroscopy

For spectroscopic studies, NSOM can be considered as a controlled light collector, with tens-of-nanometer spatial positioning resolution. Thus, many of the standard spectroscopic techniques could be applied, depending on the amount of signal available. For example, it was successfully demonstrated that NSOM can be used for fluorescence and photoluminescence spectroscopy; electroluminescence spectroscopy, time-resolved spectroscopy; polarization studies, and in early stage infrared (IR) and Raman spectroscopies. The latter of the two has the greatest promise in becoming the ultimate molecular nanoprobe. Some of the applications of *in situ* hyperspectral, that is, composition-specific nanoscale studies include chemical identification of observed samples; studies of optical properties of materials at nano-scale levels; detection of phase differences and impurities in materials; protein studies, and many more. Ultimately, it opens the door for a plethora of both fundamental and applied studies in the fields of physics, material science, chemistry, life sciences, and nanotechnology.

Examples of a near-field spectroscopy application are shown in Fig. 7. While Fig. 7a represents the topography image of the PIC dye crystal (33), Fig. 7b is the NSOM fluorescence image of the same area. In order to explore the origin of inhomogeneities, near-field fluorescence spectroscopy, with spectra presented at Fig. 7c, has been done at different points, labeled 1–4, on Fig. 7b. Finally, fluorescence spectra resolved the inhomogenieties of emission sources, pointing to the two different allotropes of crystals having emissions peaking at 645 and 690 nm, respectively.

The near-field signal, which is by default very weak, gets even weaker if we want to do the wavelength-resolved spectroscopic analysis, or full hyperspectral cube imaging, so longer exposition time is necessary to acquire usable spectra. For a near-field spectroscopy system to be successful, it needs to have an extremely stable probe position control, in all three axes, to keep the optimal sample-probe distance, and to keep the probe above the point of interest, for a prolonged time of signal collection.

Figure 8 represents the typical modern commercial NSOM microscopy–spectroscopy setup, in this particular case, an Aurora-3 for spectroscopy made by Veeco Instruments Inc., Santa Barbara, CA (34). In general, such a near-field microscopy–spectroscopy package consists of the NSOM microscope, an objective lens for signal collection, optical filters for elimination of the excitation laser light,

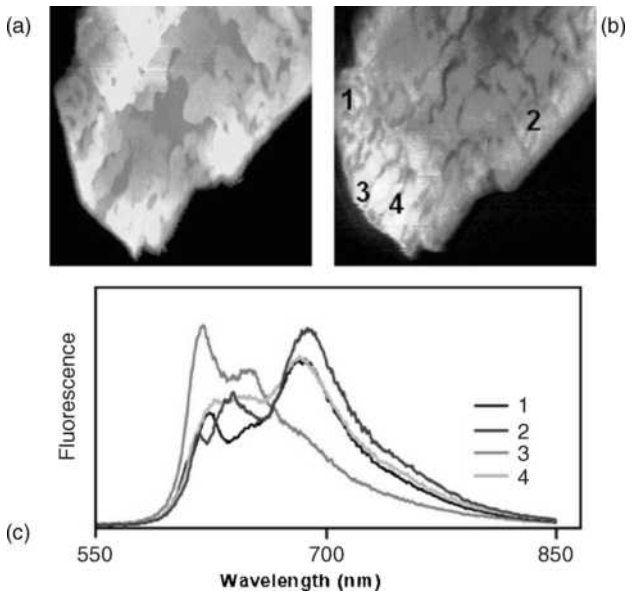


Figure 7. Example of fluorescence-based hyperspectral near-field imaging. (a) Shear-force topography image; (b) NSOM fluorescence image; (c) spatially resolved fluorescence spectra, numbers 1–4 corresponds to the different position on the crystal, demonstrating chemically specific imaging and material inhomogenities at the nanoscale. (Courtesy of David Vanden Bout and Paul Barbara, University of Minnesota, Minneapolis, MI.)

an optical pathway for signal transduction to the detector, and a wavelength dispersive detector. As the signal is generally very weak, the spectrometer needs to have a very sensitive detection system, in the best case, single-photon sensitivity. With today's solid-state detectors technology, the best detector to use is the CCD camera with a larger stack of sensitive pixels coupled to the imaging spectrometer. Depending on applications, either a Peltier cooled camera will be satisfactory (for most bright fluorescence samples) or a liquid-nitrogen-cooled camera for ultimate sensitivity, such as in single-molecular experiments and near-field Raman spectroscopy. Furthermore, interfacing and communication between spectrometer and NSOM scanning probe control software needs to be established.

Today, many spectrometers have semiopen control software, thus allowing triggering of the spectral acquisition with the TTL handshake signal, which can be produced by the NSOM microscope controller. In this way, the microscope controller initiates spectral acquisition and the system can be used with many different commercial spectrometers, allowing customization of the microprobe. Filtering out excitation photons is of extreme importance to increase the signal/noise ratio, and the best filters available today are notch, interferometric filters. Another consideration when designing the NSOM spectroscopy package is that not one formula will fit all of the requirements. Furthermore, virtually any application will require some special design consideration, so building the system

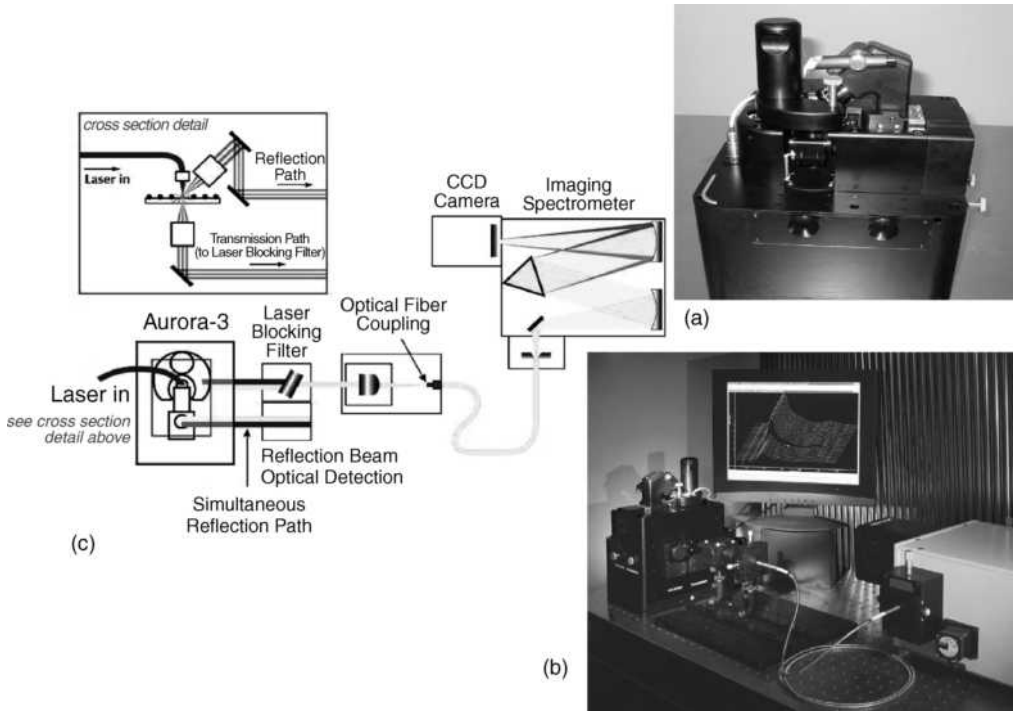


Figure 8. Typical, third generation commercially available NSOM setup, for imaging and spectroscopy: (a) photo of the NSOM head and sample holder; (b) complete system with attached spectrometer; (c) optical path schematics. (Courtesy of Veeco Instruments Inc., Santa Barbara, CA.)

as versatile and modular as possible is of the utmost importance. The ability to incorporate additional standard optical components and easy system reconfiguration should be first in the designer's mind. At the same time, once a system is set up for operation, it should allow for easy, straightforward, simple operation, and robustness, which are sometimes contradictory requirements. The design of the NSOM head, when having the spectroscopy package in mind, must incorporate extremely accurate closed-loop scan linearization. While shear-force feedback keeps superb control of sample-tip distance, the closed-loop *X–Y* scanning is necessary in order to keep the probe at the selected spatial position during sample collection.

The near-field spectroscopy package is an even more promising breakthrough technology for life sciences than for NSOM imaging alone. For the first time, it will allow qualitative identification of the composition of an observed sample, via optical spectroscopies, at the spatial resolution of scanning probe microscopy. The current commercial systems can distinguish differences in chemical composition of the samples, at the spatial resolution of 50 nm or better. For combined near-field spectroscopy and imaging, the signal may be split into the dual-beam path, wherein one path is used for the imaging detector and another one for spectroscopy. Such dual-beam solution minimizes removal, replacement, and realignment of components when a different mode is desired. The benefit is the ease of use and robust, reliable operation for separate or simultaneous transmission and reflection measurements.

In the example of the Aurora-3 optical path system, the light from both objectives is redirected out the side ports by two front-surface mirrors, which are reflection-coated for optimal visible/near-IR (NIR) operation. The near-field transmission and reflection light is collected in the far field by precisely aligned microscope objectives to provide high quality collimated (parallel) beams, with a nominal 7 mm diameter. This system design allows for NSOM spectroscopic operation with standard one-half in. diameter optics, though for ease of alignment and handling, 1 in. diameter optics is generally recommended. Furthermore, the reflection path is carefully aligned to match the transmission path. Such integrated solutions minimize removal, replacement, and realignment of components when a different mode is desired. The integrated reflection path is always available for use with any sample and never requires removal or realignment in the microscope with normal use. As there are many variations in the spectroscopy setup, it is important for an NSOM system that is intended for spectroscopic use to be capable of utilizing the standard optical element, so users can customize the system using standard optical poles and optical bench mounting systems.

Evaluation

While the first NSOM was invented just 2 years after the AFM, its widespread use has just started to pick up in the last several years. The reasons for this lag are severalfold: the first NSOM images were hard to interpret, and as they were achieved without topography modulation, the contrast in the images was not intuitive; there were no commercial instruments available until the mid-1990s, thus all users needed to build their own systems; the use of the home-built and first commercial instrument and its alignment procedures were cumbersome and complicated for any user who was not skilled in optoelectronics development; and the resolution of the systems depended on each individual sample. However, the field is changing rapidly, and with introduction of the latest, third generation of commercial systems (13–16), such as the Aurora-3 from Veeco Instruments Inc., MultiView 400 from Nanonics Imaging Ltd, Smena from NT-MDT, and AlphaSNOM from WiTec GmbH, the ease of use is comparable with the standard scanning probe microscope and is within the skill set of average life sciences user. Furthermore, with improvement in the serial production of the probes and quality control in recent years, the resolution of the NSOM system is becoming more uniform, and resolution expectations can be met with most samples.

The way to evaluate and measure resolution of the NSOM instrument is by utilizing Fisher's projection masks (35). It is virtually accepted as a standard for evaluating NSOM resolution and quality of image topographic modulation, the latter one by comparing the topographic with the optical image. The Fisher project mask is a regular hexagonal array of metallic spikes (Fig. 9). It is produced by having the monolayer of the monodispersed polymer spheres coated with metallic coating, and the spheres subsequently dissolved with organic solvents. What is left is the regular, closed-packed hexagonal matrix of metallic spikes. In transmission mode, the spike is seen as a dark spot on the optical micrography, while in reflection mode, the spike is a bright spot, and the void space is dark.

The near-field optical imaging obviously provides two great advantages over other types of imaging: its ability to simultaneously acquire topography, in a scanning-force manner, and an optical image. The optical image carries a plethora of different information that can be furthermore extracted. The most important advantage is that there are many different ways to extract direct or indirect information on spatial distribution of chemical composition of the observed sample, even on the single molecular level. Furthermore, at current state-of-the-art commercial NSOM instrumentation, the near-field imaging and spectroscopy can be performed at a resolution at least four times as high as the resolution of the best optical confocal microscopes. The obvious applications in the biomedical field are all of the applications for which the standard confocal and inverted fluorescence microscope is used today, but at the same time, done at much higher resolution (36–40). Examples of life sciences and biomedical applications involve optical ultramicroscopy of cells; optical imaging of cell organelles; imaging and spectroscopy of individual molecules and macromolecules; *in vivo* tracking of molecular events and endocytosis; and high resolution chromosome labeling [i.e., high resolution fluorescence *in situ* hybridization (FISH) (39) applications]. Some of these applications, like molecular tracking, and high resolution FISH are unachievable with other methods.

The molecular tracking applications are based on fluorescence, and in the future, Raman SERS applications will revolutionize our way of understanding how cellular

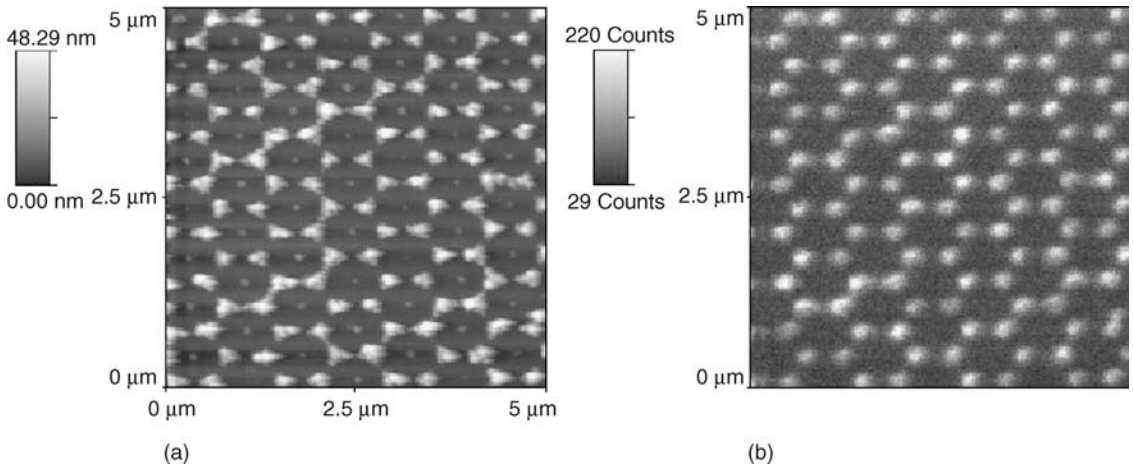


Figure 9. Fisher mask, typical structure for NSOM evaluation and calibration: (a) topographic image produced with shear-force feedback; (b) NSOM transmission mode image of the same area. The images were produced on the Aurora-3 NSOM, (Courtesy of Veeco Instruments Inc., Santa Barbara, CA.)

mechanisms work, and will bring significant contributions to practical pharmacological applications, such as drug candidates, where their target will be able to be imaged, measured, and tracked during action. Additionally, this application will add considerably to the body of knowledge of macromolecular interactions and in the mapping of the interactome of proteins, enzymes, and nucleic acids within the cell. The simplest way will be by expressing differently colored fluorescence proteins, fused with ligand and target, and tracking their spatial positions within the cell, and from the fluorescence resonance energy transfer or fluorescence intermittency, following their interactions. No other method is capable of achieving the resolution necessary to understand molecular machines *in vivo*. Another futuristic application of NSOM is in the ultrahigh density genomics and proteomics array, which theoretically can be packed at the density range higher than the wavelengths.

Some of the more futuristic life science applications will come with the integration of NSOM with mass spectrometry. Zenobi's group (25) demonstrated that the apertured NSOM is capable of doing the nanoablation of structures and thus can feed the mass spectrometer with material ablated with the resolution of a few tens of nanometer. This can dramatically improve the knowledge of spatial locations of proteins, and protein–protein complexes and interactions.

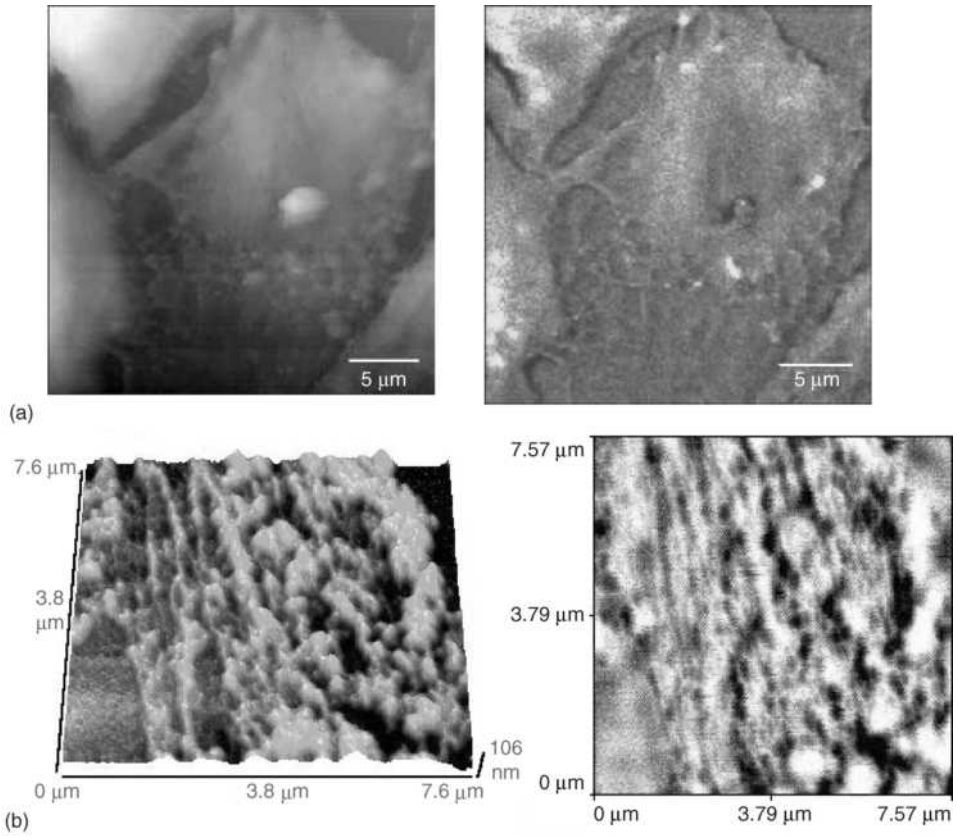
The NSOM-based FISH (38) can have a large impact on the future of molecular *in vitro* diagnostics because it will allow FISH to be applied on the shorter segments of DNA. This will permit many additional applications by painting shorter genes of this simple, chromosome painting technique, which are unachievable with far-field fluorescence or confocal-fluorescence equipment. If designed in a simple and high throughput manner, this may become a standard diagnostic instrument for molecular cytogenetics diagnostics in pathology labs.

It is expected that the next, fourth-generation instrumentation, currently in the design stage, will allow more routine imaging with a level of technical expertise, which is necessary for practical applications comparable with running today's confocal microscope.

Examples of the NSOMs biological applications are presented in Fig. 10a–c. Figure 10a is an example of the protein localization imaging beyond diffraction limit. In this particular image, the fibroblast cells were labeled with green fluorescence protein (GFP). The image on the left represents the shear-force micrography, which corresponds to the topographic image, while the image on the right is the GFP fluorescence image. Spatial distribution of the GFP can easily be observed within the cell at a resolution far exceeding the diffraction limits. This technique can be used to track the protein synthesis and trafficking within the cell if the targeted protein is fused with the fluorescence label, such as GFP or YFP. Another unique NSOM application is in optical characterization of the supramolecular and macromolecular assemblies. Figure 10b represents topographic, shear force (left) and NSOM transmission image (right) of the interband region of a polytene chromosome. In the optical image, the chromatin matter can be distinguished from the DNA based on the optical contrast, which is not possible based on the pure topography. Finally, Fig. 10c represents far-field optical transmission image of slice of the muscle tissue (left) and shear-force topography and near-field optical image, on the right top and bottom, respectively. The near-field imaging reveals the fine structure of the muscular fiber, its cell membrane, myofibrils, and endoplasmatic reticulum structure, in a similar manner as using the transmission electron microscopy.

Besides imaging, the near-field optical microscopy-like setup has promise for use in the nanoscale lithography (41), and for high density data storage (42). Nanoscale lithography will have applications in the preparation of tissue growth matrix and scaffolds, especially for the growth of neurons, while the high voluminous data storage will have a plethora of medical applications for storing ever-growing informational content of both imaging and high throughput diagnostics data.

In conclusion, near-field optical imaging is still a developing technique that shows much promise for biomedical



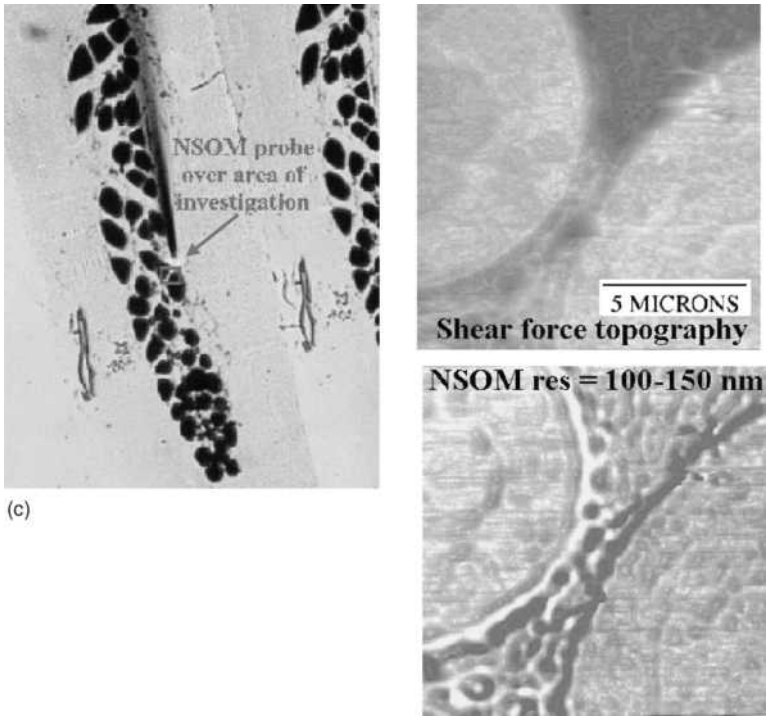


Figure 10. Examples of biomedical and life-sciences applications of NSOM imaging: (a) Shear-force topography and near-field fluorescence from GFP-labeled fibroblast cells. Images of the fibrorblast cells are prepared by growing them directly on glass cover slides and subsequent labeling. They are imaged in air at a scan speed of 1 Hz, no photodegradation was observed throughout the measurement. (Courtesy of Renato Zenobi, ETH, Zurich, © Renato Zenobi and ETH Zurich, Switzerland). (b) Shear force and NSOM image of the interband region of a polytene chromosome. (Courtesy of Sid Ragona and Phil Haydon, Laboratory of Cellular Signaling, Dept. of Zoology and Genetics, Iowa State University, Ames, IA, and Veeco Instruments Inc.). (c) Far-field, differential contrast optical microscopy of the muscle tissue, and details of the muscle cell by shear force and NSOM. (Courtesy of Sid Ragona and Phil Haydon, Laboratory of Cellular Signaling, Department of Zoology and Genetics, Iowa State University, Ames, IA and Veeco Instruments Inc.)

applications. This review presents the state-of-the-art NSOM technology up to the second quarter of 2005. It is certain that by the time of the next edition of this *Encyclopedia*, there will be many other technological and advancements in the field of biomedical applications of near-field optics.

BIBLIOGRAPHY

1. Abbe E. Beitrsge zur theorie des mikroskops und mikroskopichen wahrnehmung. Arch Mikroskop Anat 1873;9:413.
2. Synge EH. A suggested method for extending the microscopic resolution into the ultramicroscopic region. Phil Mag 1928;6: 356–362.
3. Synge EH. An application of piezoelectricity to microscopy. Phil Mag 1932;13:297–300.
4. O'Keefe JA. Resolving power of visible light. J Opt Soc Am 1956;46:359–362.
5. Ash A, Nichols G. Super-resolution aperture scanning microscope. Nature London 1972;237:510–511.
6. Pohl DW, Denk W, Lanz M. Optical stethoscopy—Image recording with resolution lambda/20. Appl Phys Lett 1984;44: 651–653.
7. Lewis A, Isaacson M, Harootunian A, Murray A. Development of 500A spatial-resolution light-microscope. 1. Light is efficiently transmitted through gamma-16 diameter apertures. Ultramicroscopy 1984;13:227–231.
8. Rasmussen A, Deckert V. New dimension in nano-imaging: breaking through the diffraction limit with scanning near-field optical microscopy. Anal Bioanal Chem 2005;381:165–172.
9. Edidin M. Near-field scanning optical microscopy, a siren call to biology. Traffic 2001;2:797–803.
10. Dunn RC. Near-field scanning optical microscopy. Chem Rev 1999;99:2891–2899.
11. Micic M. Near-field scanning optical microscopy and spectroscopy advance. Photonics Spectra 2005;38:124–125.
12. De Serio M, Zenobi R, Deckert V. Looking at the nanoscale: scanning near-field optical microscopy. TRAC-Trends Anal Chem 2003;22:70–77.
13. Product info: Aurora-3 NSOM , Veeco Instruments Inc., Santa Barbara, CA. Available at http://www.veeco.com.
14. Product info: Multiview Series, Nanonics Imaging Ltd., Jerusalem, Israel. Available at http://www.nanonics.co.il.
15. Product info: Alpha SNOM, Witec Wissenschaftliche Instrumente und Technologie. Available at GmbH GmbH,Ulm, Germany, http://www.witec.de.
16. Product info: Solver SNOM, NT-MDT Co, Zelenogorod, Russia. Available at http://www.ntmdt.ru.
17. Paesler MA, Moyer P. Near-Field Optics Theory, Instrumentation and Applications. New York: John Wiley & Sons; 1996.
18. Curie P, Curie J. Developement, par pression de l'electricite polaire dans les cristaux hemiedres a faces inclinees. Comp Ren 1880;91:291–295.
19. Spanner K. Micro Positioning, Nano Positioning, Nano Automation: Solution for Cutting Edge Technologies (product catalog), Karslruhe, Physik Instrumente (PI) GmbH & Co KG. Available at http://www.pi.ws. Accessed 2005.
20. Betzig E, Finn PL, Weiner JS. Combined shear force and near-field scanning optical microscope. Appl Phys Lett 1992;60: 2484–2486.
21. Garcia-Parajo M, Tate T, Chen Y. Gold-coated parabolic tapers for scanning near-field optical microscopy: Fabrication and optimisation. Ultramicroscopy 1995;61:155–163.
22. Pangaribuan T, et al. Reproducible fabrication technique of nanometric tip diameter fiber probe for photon scanning tunneling microscope. Jpn J Appl Phys 1992;31:L1302–L1304.
23. Radojewski R, Grabijec P. Combined SNOM/AFM microscopy with micromachined nanoapertures. Mater Sci–Poland 2003;21:321–332.
24. Takahashi S, Dickson W, Pollard R, Zaytas A. Near-field magneto-optical analysis in reflection mode SNOM. Ultramicroscopy 2004;100(3–4):443–447.
25. Stockle R, et al. Nanoscale atmospheric pressure laser ablation mass spectrometry. Anal Chem 2001;73:139–1402.
26. Chovin A, Garrigue P, Servant L, Sojic N. Electrochemical modulation of remote fluorescence imaging at an ordered opto-electrochemical nanoaperture array. Chemphyschem 2004;5:1125–1132.
27. Inoye Y, Kawata S. Near-field scanning optical microscope with a metallic probe tip. Opt Lett 1994;19:159–161.
28. Keilmann F, Hillenbrand R. Near-field microscopy by elastic light scattering from a tip. Philas Trans R Soc Sci A 2004;362:787–805.
29. Hu DH, et al. Correlated topographic and spectroscopic imaging beyond diffraction limit by atomic force microscopy metallic tip enhanced near-field fluorescence lifetime microscopy. Rev Sci Instr 2003;74:3347–3355.
30. Micic M, Klymyshin N, Suh YD, Lu HP. Finite element method simulation of the field distribution for AFM tip enhanced surface enhanced Raman scanning microscopy. J Phys Chem B 2003;107:1574–1584.
31. Sun WX, Shen ZX. Near-field scanning Raman microscopy using apertureless probes. J Raman Spectrosc 2003;34:668–676.
32. Richards D. Near-field microscopy: Throwing light on the nanoworld Philas Trans R Soc Sci A 2003;361:2843–2857.
33. Vanden Bout DA, Kerimo J, Higgins DA, Barbara PF. Spatially Resolved Spectral Inhomogeneities in Small Molecular Crystals Studied by Near Field Scanning. Opt Microsc J Phys Chem 1996;100:11843–11850.
34. Puestow R. Configuring Aurora-3 for Spectroscopy, application note, Veeco Instruments Inc, Santa Barbara, CA, 2003.
35. Fischer UC, et al. Latex bead projection nanopatterns. Surf Interface Anal 2002;33:75–80.
36. de Lange F, et al. Cell biology beyond the diffraction limit: Near-field scanning optical microscopy. J Cell Sci 2001;114: 4153–4160.
37. Subramaniam V, Kirsch AK, Jovin TM. Cell biological applications of scanning near-field optical microscopy (SNOM). Cell Molec Biol 1998;44:689–700.
38. Lewis A, et al. Near-field scanning optical microscopy in cell biology. Trends Cell Biol 1999;9:70–73.
39. Fukushi D, et al. Scanning near-field optical/atomic force microscopy detection of fluorescence in situ hybridization signals beyond the optical limit. Exp Cell Res 2003;289:237–244.
40. Krishnan RV, Varma R, Mayor S. Fluorescence methods to probe nanometer-scale organization of molecules in living cell membranes. J Fluoresc 2001;11:211–226.
41. Dryakhulshin VF, Klimov AY, Rogov VV, Vostkov NV. Near-field optical lithography method for fabrication of nanodimensional objects. Appl Surf Sci 2005;248:200–203.
42. Ferri V, et al. Near-field optical addressing of luminescent photoswitchable supramolecular system embedded in inert polymer matrices. Nano Lett 2004;4:835–859.

Further Reading

Prasad PN. Nanophotonics. New York: John Wiley & Sons; 2004.

Courion D. Near Field Microscopy and Near Field Optics. London: Imperial College Press; 2003.

Paul DW, Courion D. Near Field Optics. Arc-et Senans: Kulwer Academic Publisher; 1993.

Taatjes DJ, Brooke MT. Cell Imaging Techniques: Methods and Protocolos. Totowa: Humana Press; 2005.

See also MICROSCOPY, CONFOCAL; MICROARRAYS; NANOPARTICLES.

MICROSCOPY, CONFOCAL

NATHAN S. CLAXTON
THOMAS J. FELLERS
MICHAEL W. DAVIDSON
The Florida State University
Tallahassee, Florida

INTRODUCTION

The technique of laser scanning and spinning disk confocal fluorescence microscopy has become an essential tool in biology and the biomedical sciences, as well as in materials science due to attributes that are not readily available using other contrast modes with traditional optical microscopy (1–12). The application of a wide array of new synthetic and naturally occurring fluorochromes has made it possible to identify cells and submicroscopic cellular components with a high degree of specificity amid nonfluorescing material (13). In fact, the confocal microscope is often capable of revealing the presence of a single molecule (14). Through the use of multiply labeled specimens, different probes can simultaneously identify several target molecules simultaneously, both in fixed specimens and living cells and tissues (15). Although both conventional and confocal microscopes cannot provide spatial resolution below the diffraction limit of specific specimen features, the detection of fluorescing molecules below such limits is readily achieved.

The basic concept of confocal microscopy was originally developed by Minsky in the mid-1950s (patented in 1961) when he was a postdoctoral student at Harvard University (16,17). Minsky wanted to image neural networks in unstained preparations of brain tissue and was driven by the desire to image biological events as they occur in living systems. Minsky's invention remained largely unnoticed, due most probably to the lack of intense light sources necessary for imaging and the computer horsepower required to handle large amounts of data. Following Minsky's work, Egger and Petran (18) fabricated a multiple-beam confocal microscope in the late-1960s that utilized a spinning (Nipkow) disk for examining unstained brain sections and ganglion cells. Continuing in this arena, Egger went on to develop the first mechanically scanned confocal laser microscope, and published the first recognizable images of cells in 1973 (19). During the late-1970s and the 1980s, advances in computer and laser technology, coupled to new algorithms for digital manipulation of images, led to a growing interest in confocal microscopy (20).

Fortuitously, shortly after Minsky's patent had expired, practical laser-scanning confocal microscope designs were translated into working instruments by several investigators. Dutch physicist Brakenhoff developed a scanning confocal microscope in 1979 (21), while almost simultaneously, Sheppard contributed to the technique with a theory of image formation (22). Wilson, Amos, and White nurtured the concept and later (during the late-1980s) demonstrated the utility of confocal imaging in the examination of fluorescent biological specimens (20,23). The first commercial instruments appeared in 1987. During the 1990s, advances in optics and electronics afforded more stable and powerful lasers, high efficiency scanning mirror units, high throughput fiber optics, better thin-film dielectric coatings, and detectors having reduced noise characteristics (1). In addition, fluorochromes that were more carefully matched to laser excitation lines were beginning to be synthesized (13). Coupled to the rapidly advancing computer processing speeds, enhanced displays, and large-volume storage technology emerging in the late-1990s, the stage was set for a virtual explosion in the number of applications that could be targeted with laser scanning confocal microscopy.

Modern confocal microscopes can be considered as completely integrated electronic systems where the optical microscope plays a central role in a configuration that consists of one or more electronic detectors, a computer (for image display, processing, output, and storage), and several laser systems combined with wavelength selection devices and a beam scanning assembly. In most cases, integration between the various components is so thorough that the entire confocal microscope is often collectively referred to as a digital or video imaging system capable of producing electronic images (24). These microscopes are now being employed for routine investigations on molecules, cells, and living tissues that were not possible just a few years ago (15).

Confocal microscopy offers several advantages over conventional widefield optical microscopy, including the ability to control depth of field, elimination, or reduction of background information away from the focal plane (that leads to image degradation), and the capability to collect serial optical sections from thick specimens. The basic key to the confocal approach is the use of spatial filtering techniques to eliminate out-of-focus light or glare in specimens whose thickness exceeds the immediate plane of focus. There has been a tremendous explosion in the popularity of confocal microscopy in recent years (1–4,6,7), due in part to the relative ease with which extremely high quality images can be obtained from specimens prepared for conventional fluorescence microscopy, and the growing number of applications in cell biology that rely on imaging, both fixed and living cells and tissues. In fact, confocal technology is proving to be one of the most important advances ever achieved in optical microscopy.

In a conventional widefield optical epi-fluorescence microscope, secondary fluorescence emitted by the specimen often occurs through the excited volume and obscures resolution of features that lie in the objective focal plane (25). The problem is compounded by thicker specimens (>2 μm), which usually exhibit such a high degree of fluorescence emission that most of the fine detail is lost. Confocal microscopy provides only a marginal improvement in both axial (z; parallel to the microscope optical axis) and lateral (x and y; dimensions in the specimen plane) optical resolution, but is able to exclude secondary fluorescence in areas removed from the focal plane from resulting images (26–28). Even though resolution is somewhat enhanced with confocal microscopy over conventional widefield techniques (1), it is still considerably less than that of the transmission electron microscope (TEM). In this regard, confocal microscopy can be considered a bridge between these two classical methodologies.

Illustrated in Fig. 1 are a series of images that compare selected viewfields in traditional widefield and laser

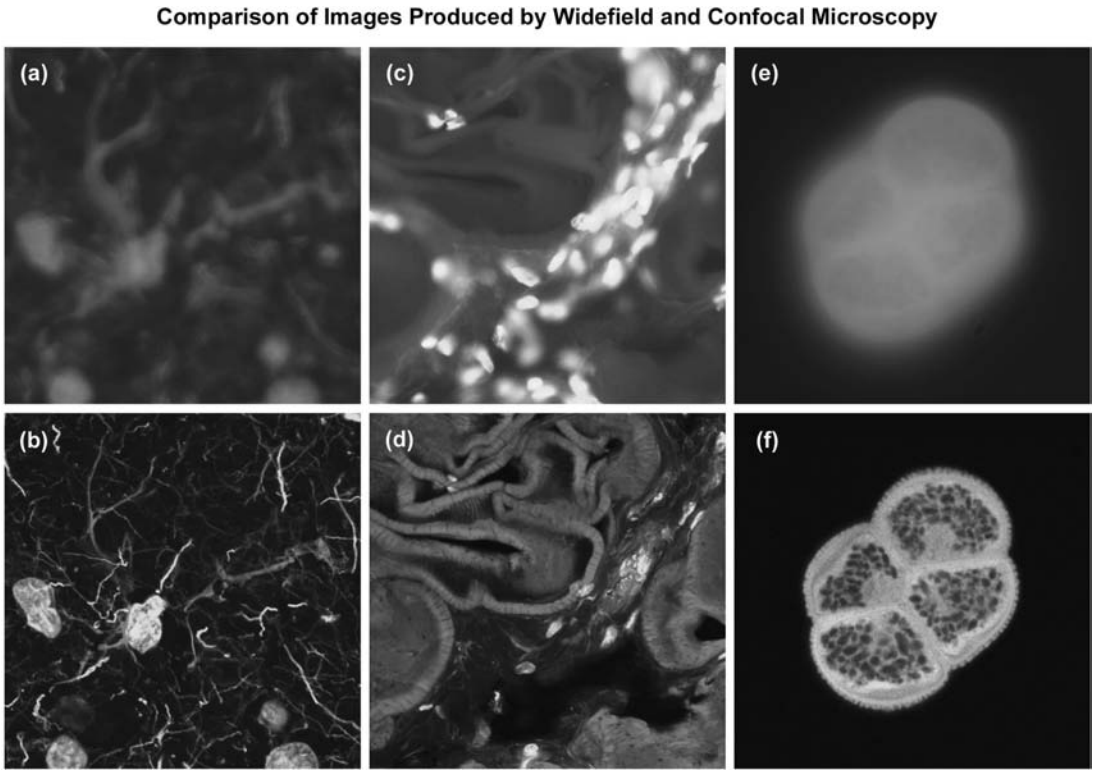


Figure 1. Comparison of widefield (upper row) and laser scanning confocal fluorescence microscopy images (lower row). Note the significant amount of signal in the widefield images from fluorescent structures located outside of the focal plane. (a) and (b) Mouse brain hippocampus thick section treated with primary antibodies to glial fibrillary acidic protein (GFAP; red), neurofilaments H (green), and counterstained with Hoechst 33342 (blue) to highlight nuclei. (c) and (d) Thick section of rat smooth muscle stained with phalloidin conjugated to Alexa Fluor 568 (targeting actin; red), wheat germ agglutinin conjugated to Oregon Green 488 (glycoproteins; green), and counterstained with DRAQ5 (nuclei; blue). (e) and (f) Sunflower pollen grain tetrad autofluorescence.

scanning confocal fluorescence microscopy. A thick (16 μm) section of fluorescently stained mouse hippocampus in widefield fluorescence exhibits a large amount of glare from fluorescent structures located above and below the focal plane (Fig. 1a). When imaged with a laser scanning confocal microscope (Fig. 1b), the brain thick section reveals a significant degree of structural detail. Likewise, widefield fluorescence imaging of rat smooth muscle fibers stained with a combination of Alexa Fluor dyes produce blurred images (Fig. 1c) lacking in detail, while the same specimen field (Fig. 1d) reveals a highly striated topography when viewed as an optical section with confocal microscopy. Autofluorescence in a sunflower (*Helianthus annuus*) pollen grain tetrad produces a similar indistinct outline of the basic external morphology (Fig. 1e), but yields no indication of the internal structure in widefield mode. In contrast, a thin optical section of the same grain (Fig. 1f) acquired with confocal techniques displays a dramatic difference between the particle core and the surrounding envelope. Collectively, the image comparisons in Fig. 1 dramatically depict the advantages of achieving very thin optical sections in confocal microscopy. The ability of this technique to eliminate fluorescence emission from regions removed from the focal plane offsets it from traditional forms of fluorescence microscopy.

PRINCIPLES OF CONFOCAL MICROSCOPY

The confocal principle in epi-fluorescence laser scanning microscope is diagrammatically presented in Fig. 2. Coherent light emitted by the laser system (excitation source) passes through a pinhole aperture that is situated in a conjugate plane (confocal) with a scanning point on the specimen and a second pinhole aperture positioned in front of the detector (a photomultiplier tube). As the laser is reflected by a dichromatic mirror, and scanned across the specimen in a defined focal plane, secondary fluorescence emitted from points on the specimen (in the same focal plane) pass back through the dichromatic mirror, and are focused as a confocal point at the detector pinhole aperture.

The significant amount of fluorescence emission that occurs at points above and below the objective focal plane is not confocal with the pinhole (termed out-of-focus light rays in Fig. 2) and forms extended Airy disks in the aperture plane (29). Because only a small fraction of the out-of-focus fluorescence emission is delivered through the pinhole aperture, most of this extraneous light is not detected by the photomultiplier and does not contribute to the resulting image. The dichromatic mirror, barrier filter, and excitation filter perform similar functions to identical components in a widefield epi-fluorescence microscope (30).

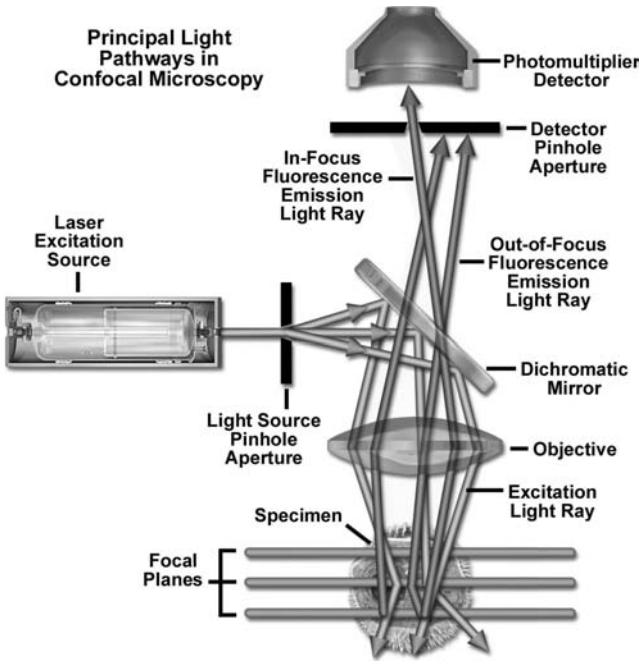


Figure 2. Schematic diagram of the optical pathway and principal components in a laser scanning confocal microscope.

Refocusing the objective in a confocal microscope shifts the excitation and emission points on a specimen to a new plane that becomes confocal with the pinhole apertures of the light source and detector.

In traditional widefield epi-fluorescence microscopy, the entire specimen is subjected to intense illumination from an incoherent mercury or xenon arc-discharge lamp, and the resulting image of secondary fluorescence emission can be viewed directly in the eyepieces or projected onto the surface of an electronic array detector or traditional film plane. In contrast to this simple concept, the mechanism of image formation in a confocal microscope is fundamentally different (31). As discussed above, the confocal fluorescence microscope consists of multiple laser excitation sources, a scan head with optical and electronic components, electronic detectors (usually photomultipliers), and a computer for acquisition, processing, analysis, and display of images.

The scan head is at the heart of the confocal system and is responsible for rasterizing the excitation scans, as well as collecting the photon signals from the specimen that are required to assemble the final image (1,5–7). A typical scan head contains inputs from the external laser sources, fluorescence filter sets and dichromatic mirrors, a galvanometer-based raster scanning mirror system, variable pinhole apertures for generating the confocal image, and photomultiplier tube detectors tuned for different fluorescence wavelengths. Many modern instruments include diffraction gratings or prisms coupled with slits positioned near the photomultipliers to enable spectral imaging (also referred to as emission fingerprinting) followed by linear unmixing of emission profiles in specimens labeled with combinations of fluorescent proteins or fluorophores having overlapping spectra (32–38). The general arrangement of scan head components is presented in Fig. 3 for a typical commercial unit.

Figure 3. Three-channel spectral imaging laser scanning microscope confocal scan head with SIM scanner laser port. The SIM laser enables simultaneous excitation and imaging of the specimen for photobleaching or photoactivation experiments. Also illustrated are ports for a visible, ultraviolet (UV), and infrared (IR) laser, as well as an arc discharge lamp port for widefield observation.

In epi-illumination scanning confocal microscopy, the laser light source and photomultiplier detectors are both separated from the specimen by the objective, which functions as a well-corrected condenser and objective combination. Internal fluorescence filter components (e.g., the excitation and barrier filters and the dichromatic mirrors) and neutral density filters are contained within the scanning unit (see Fig. 3). Interference and neutral density filters are housed in rotating turrets or sliders that can be inserted into the light path by the operator. The excitation laser beam is connected to the scan unit with a fiber optic coupler followed by a beam expander that enables the thin laser beam wrist to completely fill the objective rear aperture (a critical requirement in confocal microscopy). Expanded laser light that passes through the microscope objective forms an intense diffraction-limited spot that is scanned by the coupled galvanometer mirrors in a raster pattern across the specimen plane (point scanning).

One of the most important components of the scanning unit is the pinhole aperture, which acts as a spatial filter at the conjugate image plane positioned directly in front of the photomultiplier (39). Several apertures of varying diameter are usually contained on a rotating turret that enables the operator to adjust pinhole size (and optical section thickness). Secondary fluorescence collected by the objective is descanned by the same galvanometer mirrors that form the raster pattern, and then passes through a barrier filter before reaching the pinhole aperture (40). The aperture serves to exclude fluorescence signals from out-of-focus features positioned above and below the focal plane, which are instead projected onto the aperture as Airy disks having a diameter much larger than those forming the image. These oversized disks are spread over

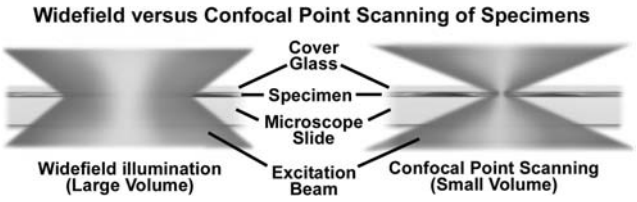


Figure 4. Widefield versus confocal microscopy illumination volumes, demonstrating the difference in size between point scanning and widefield excitation light beams.

a comparatively large area so that only a small fraction of light originating in planes away from the focal point passes through the aperture. The pinhole aperture also serves to eliminate much of the stray light passing through the optical system. Coupling of aperture-limited point scanning to a pinhole spatial filter at the conjugate image plane is an essential feature of the confocal microscope.

When contrasting the similarities and differences between widefield and confocal microscopes, it is often useful to compare the character and geometry of specimen illumination utilized for each of the techniques. Traditional widefield epi-fluorescence microscope objectives focus a wide cone of illumination over a large volume of the specimen (41), which is uniformly and simultaneously illuminated (as illustrated in Fig. 4a). A majority of the fluorescence emission directed back toward the microscope is gathered by the objective (depending on the numerical aperture) and projected into the eyepieces or detector. The result is a significant amount of signal due to emitted background light and autofluorescence originating from areas above and below the focal plane, which seriously reduces resolution and image contrast.

The laser illumination source in confocal microscopy is first expanded to fill the objective rear aperture, and then focused by the lens system to a very small spot at the focal plane (Fig. 4b). The size of the illumination point ranges from $\sim$0.25 to 0.8 μm in diameter (depending on the objective numerical aperture) and 0.5 to 1.5 μm deep at the brightest intensity. Confocal spot size is determined by the microscope design, wavelength of incident laser light, objective characteristics, scanning unit settings, and the specimen (41). Figure 4 presents a comparison between the typical illumination cones of a widefield (Fig. 4a) and point scanning confocal (Fig. 4b) microscope at the same numerical aperture. The entire depth of the specimen over a wide area is illuminated by the widefield microscope, while the sample is scanned with a finely focused spot of illumination that is centered in the focal plane in the confocal microscope.

In laser scanning confocal microscopy, the image of an extended specimen is generated by scanning the focused beam across a defined area in a raster pattern controlled by two high speed oscillating mirrors driven with galvanometer motors. One of the mirrors moves the beam from left to right along the x lateral axis, while the other translates the beam in the y direction. After each single scan along the x axis, the beam is rapidly transported back to the starting point and shifted along the y axis to begin a new scan in a process termed flyback (42). During the flyback operation, image information is not collected. In this manner, the area of interest on the specimen in a single focal plane is excited by laser illumination from the scanning unit.

As each scan line passes along the specimen in the lateral focal plane, fluorescence emission is collected by the objective and passed back through the confocal optical system. The speed of the scanning mirrors is very slow relative to the speed of light, so the secondary emission follows a light path along the optical axis that is identical to the original excitation beam. Return of fluorescence emission through the galvanometer mirror system is referred to as descanning (40,42). After leaving the scanning mirrors, the fluorescence emission passes directly through the dichromatic mirror and is focused at the detector pinhole aperture. Unlike the raster scanning pattern of excitation light passing over the specimen, fluorescence emission remains in a steady position at the pinhole aperture, but fluctuates with respect to intensity over time as the illumination spot traverses the specimen producing variations in excitation.

Fluorescence emission that is passed through the pinhole aperture is converted into an analog electrical signal having a continuously varying voltage (corresponding to intensity) by the photomultiplier. The analog signal is periodically sampled and converted into pixels by an analog-to-digital (A/D) converter housed in the scanning unit or the accompanying electronics cabinet. The image information is temporarily stored in an image frame buffer card in the computer and displayed on the monitor. Note that the confocal image of a specimen is reconstructed, point by point, from emission photon signals by the photomultiplier and accompanying electronics, yet never exists as a real image that can be observed through the microscope eyepieces.

LASER SCANNING CONFOCAL MICROSCOPE CONFIGURATION

Basic microscope optical system characteristics have remained fundamentally unchanged for many decades due to engineering restrictions on objective design, the static properties of most specimens, and the fact that resolution is governed by the wavelength of light (1–10). However, fluorescent probes that are employed to add contrast to biological specimens and, and other technologies associated with optical microscopy techniques, have improved significantly. The explosive growth and development of the confocal approach is a direct result of a renaissance in optical microscopy that has been largely fueled by advances in modern optical and electronics technology. Among these are stable multiwavelength laser systems that provide better coverage of the uv, visible, and near-IR spectral regions, improved interference filters (including dichromatic mirrors, barrier, and excitation filters), sensitive low noise wide-band detectors, and far more powerful computers. The latter are now available with relatively low cost memory arrays, image analysis software packages, high resolution video displays, and high quality digital image printers. The flow of information through a modern confocal microscope is presented diagrammatically in Fig. 5 (2).

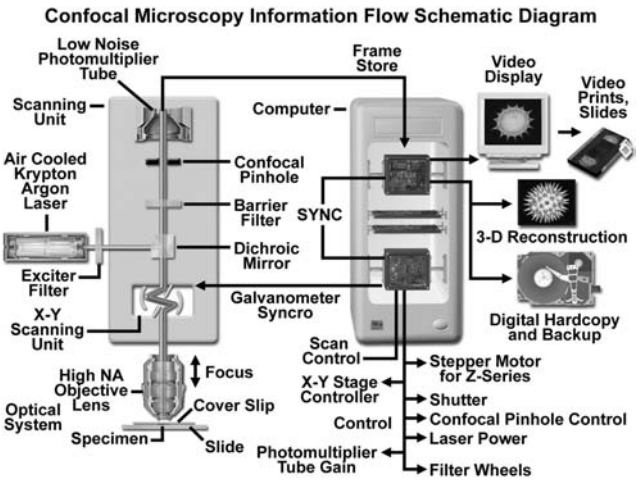


Figure 5. Confocal microscope configuration and information flow schematic diagram.

Although many of these technologies have been developed independently for a variety of specifically targeted applications, they have been incorporated gradually into mainstream commercial confocal microscopy systems. In current microscope systems, classification of designs is based on the technology utilized to scan specimens (7). Scanning can be accomplished either by translating the stage in the x, y, and z directions while the laser illumination spot is held in a fixed position, or the beam itself can be raster-scanned across the specimen. Because three-dimensional (3D) translation of the stage is cumbersome and prone to vibration, most modern instruments employ some type of beam-scanning mechanism.

In modern confocal microscopes, two fundamentally different techniques for beam scanning have been developed. Single-beam scanning, one of the more popular methods employed in a majority of the commercial laser scanning microscopes (43), uses a pair of computer-controlled galvanometer mirrors to scan the specimen in a raster pattern at a rate of approximately one frame per second. Faster scanning rates (to near video speed) can be achieved using acoustooptic devices or oscillating mirrors. In contrast, multiple-beam scanning confocal microscopes are equipped with a spinning Nipkow disk containing an array of pinholes and microlenses (44–46). These instruments often use arc-discharge lamps for illumination instead of lasers to reduce specimen damage and enhance the detection of low fluorescence levels during real-time image collection. Another important feature of the multiple-beam microscopes is their ability to readily capture images with an array detector, such as a charge-coupled device (CCD) camera system (47).

All modern laser scanning confocal microscope designs are centered on a conventional upright or inverted research level optical microscope. However, instead of the standard tungsten–halogen or mercury (xenon) arc-discharge lamp, one or more laser systems are used as a light source to excite fluorophores in the specimen. Image information is gathered point by point with a specialized detector, such as a photomultiplier tube or avalanche photodiode, and then digitized for processing by the host computer, which also controls the scanning mirrors and/or other devices to facilitate the collection and display of images. After a series of images (usually serial optical sections) has been acquired and stored on digital media, analysis can be conducted utilizing numerous image processing software packages available on the host or a secondary computer.

ADVANTAGES AND DISADVANTAGES OF CONFOCAL MICROSCOPY

The primary advantage of laser scanning confocal microscopy is the ability to serially produce thin (0.5–1.5 μm) optical sections through fluorescent specimens that have a thickness ranging up to 50 μm or more (48). The image series is collected by coordinating incremental changes in the microscope fine focus mechanism (using a stepper motor) with sequential image acquisition at each step. Image information is restricted to a well-defined plane, rather than being complicated by signals arising from remote locations in the specimen. Contrast and definition are dramatically improved over widefield techniques due to the reduction in background fluorescence and improved signal to noise (48). Furthermore, optical sectioning eliminates artifacts that occur during physical sectioning and fluorescent staining of tissue specimens for traditional forms of microscopy. The noninvasive confocal optical sectioning technique enables the examination of both living and fixed specimens under a variety of conditions with enhanced clarity.

With most confocal microscopy software packages, optical sections are not restricted to the perpendicular lateral (x–y) plane, but can also be collected and displayed in transverse planes (1,5–8,49). Vertical sections in the x–z and y–z planes (parallel to the microscope optical axis) can be readily generated by most confocal software programs. Thus, the specimen appears as if it had been sectioned in a plane that is perpendicular to the lateral axis. In practice, vertical sections are obtained by combining a series of x–y scans taken along the z axis with the software, and then projecting a view of fluorescence intensity as it would appear should the microscope hardware have been capable of physically performing a vertical section.

A typical stack of optical sections (often termed a *z series*) through a Lodgepole Pine tree pollen grain revealing internal variations in autofluorescence emission wavelengths is illustrated in Fig. 6. Optical sections were gathered in 1.0 μm steps perpendicular to the z axis (microscope optical axis) using a laser combiner featuring an argon ion (488 nm; green fluorescence), a green helium–neon (543 nm; red fluorescence), and a red helium–neon (633 nm; fluorescence pseudocolored blue) laser system. Pollen grains from this and many other species range between 10 and 40 μm in diameter and often yield blurred images in wide-field fluorescence microscopy (see Fig. 1c), which lack information about internal structural details. Although only 12 of the >36 images collected through this series are presented in the figure, they represent individual focal planes separated by a distance of ~3 μm and provide a good indication of the internal grain structure.

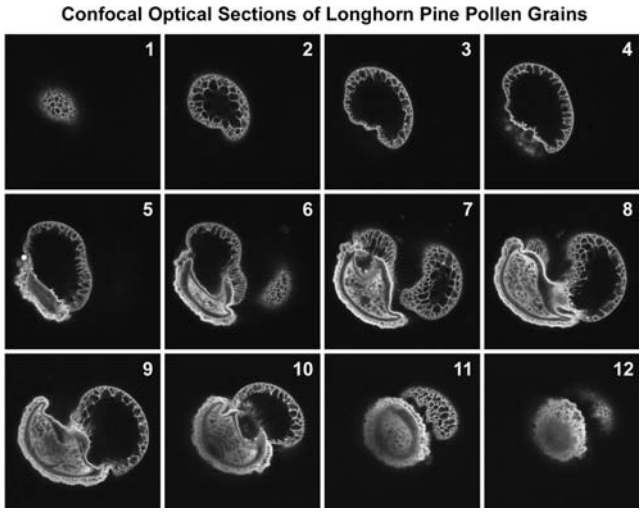


Figure 6. Lodgepole pine (*Pinus contorta*) pollen grain optical sections. Bulk pollen was mounted in CytoSeal 60 and imaged with a 100× oil immersion objective (no zoom) in 1 μm axial steps. Each image in the sequence (1–12) represents the view obtained from steps of 3 μm.

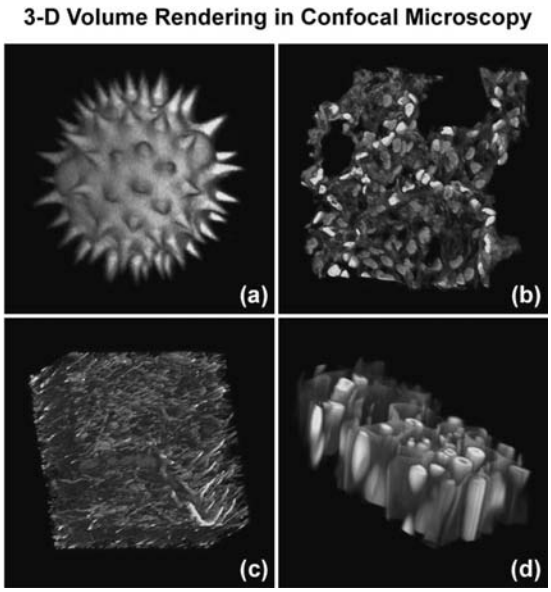


Figure 7. Three-dimensional volume renders from confocal microscopy optical sections. (a) Autofluorescence in a series of sunflower pollen grain optical sections was combined to produce a realistic view of the exterior surface. (b) Mouse lung tissue thick (16 μm) section. (c) Rat brain thick section. These specimens were each labeled with several fluorophores (blue, green, and red fluorescence) and the volume renders were created from a stack of 30–45 optical sections. (d) Autofluorescence in a thin section of fern root.

In specimens more complex than a pollen grain, complex interconnected structural elements can be difficult to discern from a large series of optical sections sequentially acquired through the volume of a specimen with a laser scanning confocal microscope. However, once an adequate series of optical sections has been gathered, it can be further processed into a 3D representation of the specimen using volume-rendering computational techniques (50–53). This approach is now in common use to help elucidate the numerous interrelationships between structure and function of cells and tissues in biological investigations (54). In order to ensure that adequate data is collected to produce a representative volume image, the optical sections should be recorded at the appropriate axial (z step) intervals so that the actual depth of the specimen is reflected in the image.

Most of the software packages accompanying commercial confocal instruments are capable of generating composite and multidimensional views of optical section data acquired from z-series image stacks. The 3D software packages can be employed to create either a single 3D representation of the specimen (Fig. 7) or a video (movie) sequence compiled from different views of the specimen volume. These sequences often mimic the effect of rotation or similar spatial transformation that enhances the appreciation of the specimen's 3D character. In addition, many software packages enable investigators to conduct measurements of length, volume, and depth, and specific parameters of the images, such as opacity, can be interactively altered to reveal internal structures of interest at differing levels within the specimen (54).

Typical 3D representations of several specimens examined by serial optical sectioning are presented in Fig. 7. A series of sunflower pollen grain optical sections was combined to produce a realistic view of the exterior surface (Fig. 7a) as it might appear if being examined by a scanning electron microscope (SEM). The algorithm utilized to construct the 3D model enables the user to rotate the pollen grain through 360° for examination. Similarly, thick sections (16 μm) of lung tissue and rat brain are presented in Fig. 7b and 7c, respectively. These specimens were each labeled with several fluorophores (blue, green, and red fluorescence) and created from a stack of 30–45 optical sections. Autofluorescence in plant tissue was utilized to produce the model illustrated in Fig. 7d of a fern root section.

In many cases, a composite or projection view produced from a series of optical sections provides important information about a 3D specimen than a multidimensional view (54). For example, a fluorescently labeled neuron having numerous thin, extended processes in a tissue section is difficult (if not impossible) to image using wide-field techniques due to out-of-focus blur. Confocal thin sections of the same neuron each reveal portions of several extensions, but these usually appear as fragmented streaks and dots and lack continuity (53). Composite views created by flattening a series of optical sections from the neuron will reveal all of the extended processes in sharp focus with well-defined continuity. Structural and functional analysis of other cell and tissue sections also benefits from composite views as opposed to, or coupled with, 3D volume rendering techniques.

Advances in confocal microscopy have made possible multidimensional views (54) of living cells and tissues that include image information in the x, y, and z dimensions as a function of time and presented in multiple colors (using two or more fluorophores). After volume processing of individual image stacks, the resulting data can be displayed as 3D multicolor video sequences in real time. Note that unlike conventional widefield microscopy, all fluorochromes in multiply labeled specimens appear in register

using the confocal microscope. Temporal data can be collected either from time-lapse experiments conducted over extended periods or through real-time image acquisition in smaller frames for short periods of time. The potential for using multidimensional confocal microscopy as a powerful tool in cellular biology is continuing to grow as new laser systems are developed to limit cell damage and computer processing speeds and storage capacity improves.

Additional advantages of scanning confocal microscopy include the ability to adjust magnification electronically by varying the area scanned by the laser without having to change objectives. This feature is termed the zoom factor, and is usually employed to adjust the image spatial resolution by altering the scanning laser sampling period (1,2,8,40,55). Increasing the zoom factor reduces the specimen area scanned and simultaneously reduces the scanning rate. The result is an increased number of samples along a comparable length (55), which increases both the image spatial resolution and display magnification on the host computer monitor. Confocal zoom is typically employed to match digital image resolution (8,40,55) with the optical resolution of the microscope when low numerical aperture and magnification objectives are being used to collect data.

Digitization of the sequential analog image data collected by the confocal microscope photomultiplier (or similar detector) facilitates computer image processing algorithms by transforming the continuous voltage stream into discrete digital increments that correspond to variations in light intensity. In addition to the benefits and speed that accrue from processing digital data, images can be readily prepared for print output or publication. In carefully controlled experiments, quantitative measurements of spatial fluorescence intensity (either statically or as a function of time) can also be obtained from the digital data.

Disadvantages of confocal microscopy are limited primarily to the limited number of excitation wavelengths available with common lasers (referred to as laser lines), which occur over very narrow bands and are expensive to produce in the UV region (56). In contrast, conventional widefield microscopes use mercury- or xenon-based arc-discharge lamps to provide a full range of excitation wavelengths in the UV, visible, and near-IR spectral regions. Another downside is the harmful nature (57) of high intensity laser irradiation to living cells and tissues, an issue that has recently been addressed by multiphoton and Nipkow disk confocal imaging. Finally, the high cost of purchasing and operating multiuser confocal microscope systems (58), which can range up to an order of magnitude higher than comparable widefield microscopes, often limits their implementation in smaller laboratories. This problem can be easily overcome by cost-shared microscope systems that service one or more departments in a core facility. The recent introduction of personal confocal systems has competitively driven down the price of low end confocal microscopes and increased the number of individual users.

CONFOCAL MICROSCOPE LIGHT DETECTORS

In modern widefield fluorescence and laser scanning confocal optical microscopy, the collection and measurement of secondary emission gathered by the objective can be accomplished by several classes of photosensitive detectors (59), including photomultipliers, photodiodes, and solid-state CCDs. In confocal microscopy, fluorescence emission is directed through a pinhole aperture positioned near the image plane to exclude light from fluorescent structures located away from the objective focal plane, thus reducing the amount of light available for image formation, as discussed above. As a result, the exceedingly low light levels most often encountered in confocal microscopy necessitate the use of highly sensitive photon detectors that do not require spatial discrimination, but instead respond very quickly with a high level of sensitivity to a continuous flux of varying light intensity.

Photomultipliers, which contain a photosensitive surface that captures incident photons and produces a stream of photoelectrons to generate an amplified electric charge, are the popular detector choice in many commercial confocal microscopes (59–61). These detectors contain a critical element, termed a photocathode, capable of emitting electrons through the photoelectric effect (the energy of an absorbed photon is transferred to an electron) when exposed to a photon flux. The general anatomy of a photomultiplier consists of a classical vacuum tube in which a glass or quartz window encases the photocathode and a chain of electron multipliers, known as dynodes, followed by an anode to complete the electrical circuit (62). When the photomultiplier is operating, current flowing between the anode and ground (zero potential) is directly proportional to the photoelectron flux generated by the photocathode when it is exposed to incident photon radiation.

In a majority of commercial confocal microscopes, the photomultiplier is located within the scan head or an external housing, and the gain, offset, and dynode voltage are controlled by the computer software interface to the detector power supply and supporting electronics (7). The voltage setting is used to regulate the overall sensitivity of the photomultiplier, and can be adjusted independently of the gain and offset values. The latter two controls are utilized to adjust the image intensity values to ensure that the maximum number of gray levels is included in the output signal of the photomultiplier. Offset adds a positive or negative voltage to the output signal, and should be adjusted so that the lowest signals are near the photomultiplier detection threshold (40). The gain circuit multiplies the output voltage by a constant factor so that the maximum signal values can be stretched to a point just below saturation. In practice, offset should be applied first before adjusting the photomultiplier gain (8,40). After the signal has been processed by the analog-to-digital converter, it is stored in a frame buffer and ultimately displayed on the monitor in a series of gray levels ranging from black (no signal) to white (saturation). Photomultipliers with a dynamic range of 10 or 12 bits are capable of displaying 1024 or 4096 gray levels, respectively. Accompanying image files also have the same number of gray levels. However, the photomultipliers used in a majority of the commercial confocal microscopes have a dynamic range limited to 8 bits or 256 gray levels, which in most cases, is adequate for handling the typical number of photons scanned per pixel (63).

Changes to the photomultiplier gain and offset levels should not be confused with postacquisition image processing to adjust the levels, brightness, or contrast in the final image. Digital image processing techniques can stretch existing pixel values to fill the black-to-white display range, but cannot create new gray levels (40). As a result, when a digital image captured with only 200 out of a possible 4096 gray levels is stretched to fill the histogram (from black to white), the resulting processed image appears grainy. In routine operation of the confocal microscope, the primary goal is to fill as many of the gray levels during image acquisition and not during the processing stages.

The offset control is used to adjust the background level to a position near 0 V (black) by adding a positive or negative voltage to the signal. This ensures that dark features in the image are very close to the black level of the host computer monitor. Offset changes the amplitude of the entire voltage signal, but since it is added to or subtracted from the total signal, it does not alter the voltage differential between the high and low voltage amplitudes in the original signal. For example, with a signal ranging from 4 to 18 V that is modified with an offset setting of 4 V, the resulting signal spans 0–14 V, but the difference remains 14 V.

Figure 8 presents a series of diagrammatic schematics of the unprocessed and adjusted output signal from a photomultiplier and the accompanying images captured with a confocal microscope of a living adherent culture of Indian Muntjac deer skin fibroblast cells treated with MitoTracker Red CMXRos, which localizes specifically in the mitochondria. Figure 8a illustrates the raw confocal image along with the signal from the photomultiplier. After applying a negative offset voltage to the photomultiplier, the signal and image appear in Fig. 8b. Note that as the signal is shifted to lower intensity values, the image becomes darker (upper frame in Fig. 8b). When the gain is adjusted to the full intensity range (Fig. 8c), the image exhibits a significant amount of detail with good contrast and high resolution.

The photomultiplier gain adjustment is utilized to electronically stretch the input signal by multiplying with a constant factor prior to digitization by the analog-to-digital converter (40). The result is a more complete representation of gray level values between black and white, and an increase in apparent dynamic range. If the gain setting is increased beyond the optimal point, the image becomes grainy, but this maneuver is sometimes necessary to capture the maximum number of gray levels present in the image. Advanced confocal microscopy software packages ease the burden of gain and offset adjustment by using a pseudocolor display function to associate pixel values with gray levels on the monitor. For example, the saturated pixels (255) can be displayed in yellow or red, while black-evel pixels (0) are shown in blue or green, with intermediate gray levels displayed in shades of gray representing their true values. When the photomultiplier output is properly adjusted, just a few red (or yellow) and blue (or green) pixels are present in the image, indicating that the full dynamic range of the photomultiplier is being utilized.

Established techniques in the field of enhanced night vision have been applied with dramatic success to photomultipliers designed for confocal microscopy (63,64). Several manufacturers have collaborated to fabricate a head-on photomultiplier containing a specialized prism system that assists in the collection of photons. The prism operates by diverting the incoming photons to a pathway that promotes total internal reflection in the photomultiplier envelope adjacent to the photocathode. This configuration increases the number of potential interactions between the photons and the photocathode, resulting in an increase in quantum efficiency by more than a factor of 2 in the green spectral region, 4 in the red region, and even higher in the IR (59). Increasing the ratio of photoelectrons generated to the number of incoming photons serves to increase the electrical current from the photomultiplier, and to produce a higher sensitivity for the instrument.

Photomultipliers are the ideal photometric detectors for confocal microscopy due to their speed, sensitivity, high signal/noise ratio, and adequate dynamic range (59–61).

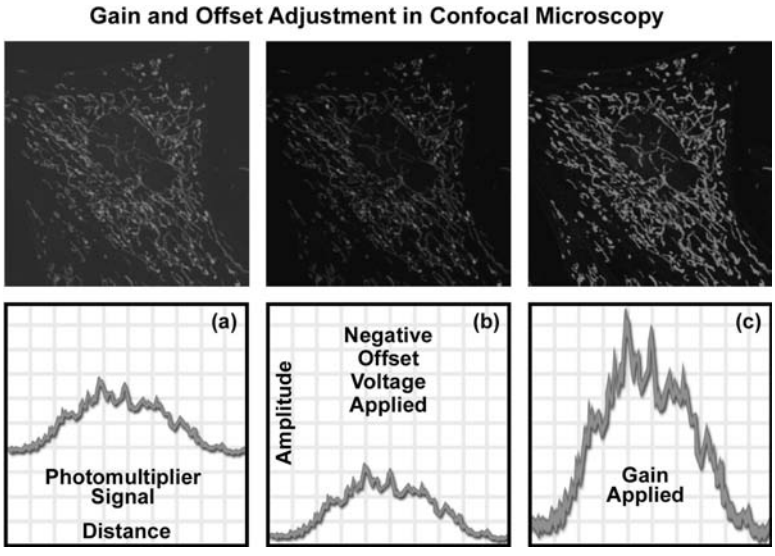


Figure 8. Gain and offset control in confocal microscopy photomultiplier detection units. The specimen is a living adherent culture of Indian Muntjac deer skin fibroblast cells treated with MitoTracker Red CMXRos. (a) The raw confocal image (upper frame) along with the signal from the photomultiplier. (b) Signal and confocal image after applying a negative offset voltage to the photomultiplier. (c) Final signal and image after the gain has been adjusted to fill the entire intensity range.

High end confocal microscope systems have several photomultipliers that enable simultaneous imaging of different fluorophores in multiply labeled specimens. Often, an additional photomultiplier is included for imaging the specimen with transmitted light using differential interference or phase-contrast techniques. In general, confocal microscopes contain three photomultipliers for the fluorescence color channels (red, green, and blue; each with a separate pinhole aperture) utilized to discriminate between fluorophores, along with a fourth for transmitted or reflected light imaging. Signals from each channel can be collected simultaneously and the images merged into a single profile that represents the real colors of the stained specimen. If the specimen is also imaged with a brightfield contrast-enhancing technique, such as differential interference contrast (65), the fluorophore distribution in the fluorescence image can be overlaid onto the brightfield image to determine the spatial location of fluorescence emission within the structural domains.

ACOUSTOOPTIC TUNABLE FILTERS IN CONFOCAL MICROSCOPY

The integration of optoelectronic technology into confocal microscopy has provided a significant enhancement in the versatility of spectral control for a wide variety of fluorescence investigations. The acoustooptic tunable filter (AOTF) is an electrooptical device that functions as an electronically tunable excitation filter to simultaneously modulate the intensity and wavelength of multiple laser lines from one or more sources (66). Devices of this type rely on a specialized birefringent crystal whose optical properties vary upon interaction with an acoustic wave. Changes in the acoustic frequency alter the diffraction properties of the crystal, enabling very rapid wavelength tuning, limited only by the acoustic transit time across the crystal.

An acoustooptic tunable filter designed for microscopy typically consists of a tellurium dioxide or quartz anisotropic crystal to which a piezoelectric transducer is bonded (67–70). In response to the application of an oscillating radio frequency (RF) electrical signal, the transducer generates a high frequency vibrational (acoustic) wave that propagates into the crystal. The alternating ultrasonic acoustic wave induces a periodic redistribution of the refractive index through the crystal that acts as a transmission diffraction grating or Bragg diffracter to deviate a portion of incident laser light into a first-order beam, which is utilized in the microscope (or two first-order beams when the incident light is nonpolarized). Changing the frequency of the transducer signal applied to the crystal alters the period of the refractive index variation, and therefore, the wavelength of light that is diffracted. The relative intensity of the diffracted beam is determined by the amplitude (power) of the signal applied to the crystal.

In the traditional fluorescence microscope configuration, including many confocal systems, spectral filtering of both excitation and emission light is accomplished utilizing thin-film interference filters (7). These filters are limiting in several respects. Because each filter has a fixed central wavelength and passband, several filters must be utilized to provide monochromatic illumination for multispectral imaging, as well as to attenuate the beam for intensity control, and the filters are often mechanically interchanged by a rotating turret mechanism. Interference filter turrets and wheels have the disadvantages of limited wavelength selection, vibration, relatively slow switching speed, and potential image shift (70). They are also susceptible to damage and deterioration caused by exposure to heat, humidity, and intense illumination, which changes their spectral characteristics over time. In addition, the utilization of filter wheels for illumination wavelength selection has become progressively more complex and expensive as the number of lasers being employed has increased with current applications.

Rotation of filter wheels and optical block turrets introduces mechanical vibrations into the imaging and illumination system, which consequently requires a time delay for damping of perhaps 50 ms, even if the filter transition itself can be accomplished more quickly. Typical filter change times are considerably slower in practice, however, and range on the order of 0.1–0.5 s. Mechanical imprecision in the rotating mechanism can introduce registration errors when sequentially acquired multicolor images are processed. Furthermore, the fixed spectral characteristics of interference filters do not allow optimization for different fluorophore combinations, nor for adaptation to new fluorescent dyes, limiting the versatility of both the excitation and detection functions of the microscope. Introduction of the AOTF to confocal systems overcomes most of the filter wheel disadvantages by enabling rapid simultaneous electronic tuning and intensity control of multiple laser lines from several lasers.

As applied in laser scanning confocal microscopy, one of the most significant benefits of the AOTF is its capability to replace much more complex and unwieldy filter mechanisms for controlling light transmission, and to apply intensity modulation for wavelength discrimination purposes (67,70). The ability to perform extremely rapid adjustments in the intensity and wavelength of the diffracted beam gives the AOTF unique control capabilities. By varying the illumination intensity at different wavelengths, the response of multiple fluorophores, for example, can be balanced for optimum detection and recording (71). In addition, digital signal processors along with phase and frequency lock-in techniques can be employed to discriminate emission from multiple fluorophores or to extract low level signals from background.

A practical light source configuration scheme utilizing an acoustooptic tunable filter for confocal microscopy is illustrated in Fig. 9. The output of three laser systems (violet diode, argon, and argon–krypton) are combined by dichromatic mirrors and directed through the AOTF, where the first-order diffracted beam (green) is collinear and is launched into a single-mode fiber. The undiffracted laser beams (violet, green, yellow, and red) exit the AOTF at varying angles and are absorbed by a beam stop (not illustrated). The major lines (wavelengths) produced by each laser are indicated (in nm) beneath the hot and cold mirrors. The dichromatic mirror reflects wavelengths < 525 nm and transmits longer wavelengths. Two longer wavelength lines produced by the argon–krypton laser

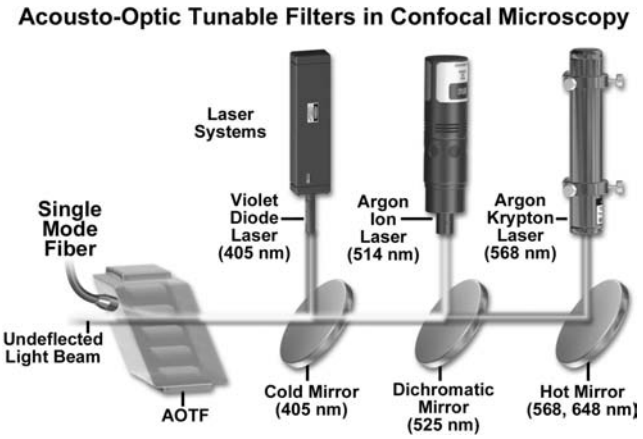


Figure 9. Configuration scheme utilizing an AOTF for laser intensity control and wavelength selection in confocal microscopy.

(568 and 648 nm) are reflected by the hot mirror, while the output of the argon laser (458, 476, 488, and 514 nm) is reflected by the dichromatic mirror and combined with the transmitted light from the argon–krypton laser. Output from the violet diode laser (405 nm) is reflected by the cold mirror and combined with the longer wavelengths from the other two lasers, which are transmitted through the mirror.

Because of the rapid optical response from the AOTF crystal to the acoustic transducer, the acoustooptic interaction is subject to abrupt transitions resembling a rectangular rather than sinusoidal waveform (66). This results in the occurrence of sidelobes in the AOTF passband on either side of the central transmission peak. Under ideal acoustooptic conditions, these sidelobes should be symmetrical about the central peak, with the first lobe having 4.7% of the central peak's intensity. In practice, the sidelobes are commonly asymmetrical and exhibit other deviations from predicted structure caused by variations in the acoustooptic interaction, among other factors. In order to reduce the sidelobes in the passband to insignificant levels, several types of amplitude apodization of the acoustic wave are employed (66,67), including various window functions, which have been found to suppress the highest sidelobe by 30–40 dB. One method that can be used in reduction of sidelobe level with noncollinear AOTFs is to apply spatial apodization by means of weighted excitation of the transducer. In the collinear AOTF, a different approach has been employed, which introduces an acoustic pulse, apodized in time, into the filter crystal.

The effective linear aperture of an AOTF is limited by the acoustic beam height in one dimension (ID) and by the acoustic attenuation across the optical aperture (the acoustic transit distance) in the other dimension (67). The height of the acoustic beam generated within the AOTF crystal is determined by the performance and physical properties of the acoustic transducer. Acoustic attenuation in crystalline materials, such as tellurium dioxide, is proportional to the square of acoustic frequency, and is therefore a more problematic limitation to linear aperture size in the shorter wavelength visible light range, which requires higher RF frequencies for tuning. Near-IR and IR radiation produces less restrictive limitations because of the lower acoustic frequencies associated with diffraction of these longer wavelengths.

The maximum size of an individual acoustic transducer is constrained by performance and power requirements in addition to the geometric limitations of the instrument configuration, and AOTF designers may use an array of transducers bonded to the crystal in order to increase the effective lateral dimensions of the propagating acoustic beam, and to enlarge the area of acoustooptic interaction (66,67,70). The required drive power is one of the most important variables in acoustooptic design, and generally increases with optical aperture and for longer wavelengths. In contrast to acoustic attenuation, which is reduced in the IR spectral range, the higher power required to drive transducers for infrared AOTFs is one of the greatest limitations in these devices. High drive power levels result in heating of the crystal, which can cause thermal drift and instability in the filter performance (66). This is particularly a problem when acoustic power and frequency are being varied rapidly over a large range, and the crystal temperature does not have time to stabilize, producing transient variations in refractive index. If an application requires wavelength and intensity stability and repeatability, the AOTF should be maintained at a constant temperature. One approach taken by equipment manufacturers to minimize this problem is to heat the crystal above ambient temperature, to a level at which it is relatively unaffected by the additional thermal input of the transducer drive power. An alternative solution is to house the AOTF in a thermoelectrically cooled housing that provides precise temperature regulation. Continuing developmental efforts promise to lead to new materials that can provide relatively large apertures combined with effective separation of the filtered and unfiltered beams without use of polarizers, while requiring a fraction of the typical device drive power.

In a noncollinear AOTF, which spatially separates the incident and diffracted light paths, the deflection angle (the angle separating diffracted and undiffracted light beams exiting the crystal) is an additional factor limiting the effective aperture of the device (67). As discussed previously, the deflection angle is greater for crystals having greater birefringence, and determines in part the propagation distance required for adequate separation of the diffracted and undiffracted beams to occur after exiting the crystal. The required distance is increased for larger entrance apertures, and this imposes a practical limit on maximum aperture size because of constraints on the physical dimensions of components that can be incorporated into a microscope system. The angular aperture is related to the total light collecting power of the AOTF, an important factor in imaging systems, although in order to realize the full angular aperture without the use of polarizers in the noncollinear AOTF, its value must be smaller than the deflection angle. Because the acoustooptic tunable filter is not an image-forming component of the microscope system (it is typically employed for source filtering), there is no specific means of evaluating the spatial resolution for this type of device (70). However, the AOTF may restrict the attainable spatial resolution of the imaging system

because of its limited linear aperture size and acceptance angle, in the same manner as other optical components. Based on the Rayleigh criterion and the angular and linear apertures of the AOTF, the maximum number of resolvable image elements may be calculated for a given wavelength, utilizing different expressions for the polar and azimuthal planes. Although diffraction limited resolution can be attained in the azimuthal plane, dispersion in the AOTF limits the resolution in the polar plane, and measures must be taken to suppress this factor for optimum performance. The dependence of deflection angle on wavelength can produce one form of dispersion, which is typically negligible when tuning is performed within a relatively narrow bandwidth, but significant in applications involving operation over a broad spectral range. Changes in deflection angle with wavelength can result in image shifts during tuning, producing errors in techniques, such as ratio imaging of fluorophores excited at different wavelengths, and in other multispectral applications. When the image shift obeys a known relationship to wavelength, corrections can be applied through digital processing techniques (1,7). Other effects of dispersion, including reduced angular resolution, may result in image degradation, such as blurring, that requires more elaborate measures to suppress.

SUMMARY OF AOTF BENEFITS IN CONFOCAL MICROSCOPY

Considering the underlying principles of operation and performance factors that relate to the application of AOTFs in imaging systems, a number of virtues from such devices for light control in fluorescence confocal microscopy are apparent. Several benefits of the AOTF combine to greatly enhance the versatility of the latest generation of confocal instruments, and these devices are becoming increasing popular for control of excitation wavelength ranges and intensity. The primary characteristic that facilitates nearly every advantage of the AOTF is its capability to allow the microscopist control of the intensity and/or illumination wavelength on a pixel-by-pixel basis while maintaining a high scan rate (7). This single feature translates into a wide variety of useful analytical microscopy tools, which are even further enhanced in flexibility when laser illumination is employed.

One of the most useful AOTF functions allows the selection of small user-defined specimen areas (commonly termed regions of interest; ROI) that can be illuminated with either greater or lesser intensity, and at different wavelengths, for precise control in photobleaching techniques, excitation ratio studies, resonance energy-transfer investigations, or spectroscopic measurements (see Fig. 10). The illumination intensity can not only be increased in selected regions for controlled photobleaching experiments (71–73), but can be attenuated in desired areas in order to minimize unnecessary photobleaching. When the illumination area is under AOTF control, the laser exposure is restricted to the scanned area by default, and the extremely rapid response of the device can be utilized to provide beam blanking during the flyback interval of the galvanometer scanning mirror cycle, further limiting unnecessary specimen exposure. In practice, the regions of excitation are typically defined by freehand drawing or using tools to produce defined geometrical shapes in an overlay plane on the computer monitor image. Some systems allow any number of specimen areas to be defined for laser exposure, and the laser intensity to be set to different levels for each area, in intensity increments as small as 0.1%. When the

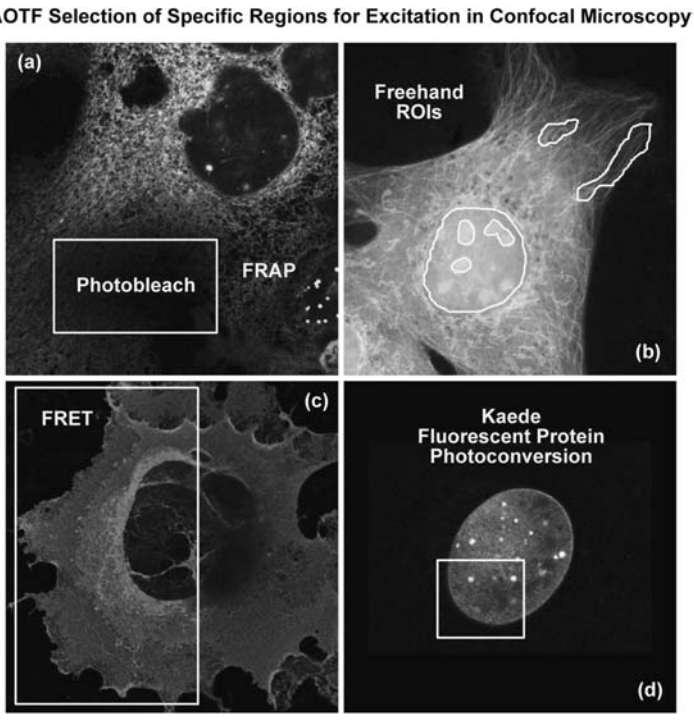


Figure 10. AOTF selection of specific regions for excitation in confocal microscopy. (a) Region of Interest (ROI) selected for fluorescence recovery after photobleaching (FRAP) experiments. (b) Freehand ROIs for selective excitation. (c) ROI for fluorescence resonance energy-transfer (FRET) analysis. (d) ROI for photoactivation and photoconversion of fluorescent proteins.

AOTF is combined with multiple lasers and software that allows time course control of sequential observations, time-lapse experiments can be designed to acquire data from several different areas in a single experiment, which might, for example, be defined to correspond to different cellular organelles.

Figure 10 illustrates several examples of several user-defined ROIs that were created for advanced fluorescence applications in laser scanning confocal microscopy. In each image, the ROI is outlined with a yellow border. The rat kangaroo kidney epithelial cell (PtK2 line) presented in Fig. 10a has a rectangular area in the central portion of the cytoplasm that has been designated for photobleaching experiments. Fluorophores residing in this region can be selectively destroyed by high power laser intensity, and the subsequent recovery of fluorescence back into the photobleached region monitored for determination of diffusion coefficients. Several freehand ROIs are illustrated in Fig. 10b, which can be targets for selective variation of illumination intensities or photobleaching and photoactivation experiments. Fluorescence resonance energy-transfer emission ratios can be readily determined using selected regions in confocal microscopy by observing the effect of bleaching the acceptor fluorescence in these areas (Fig. 10c; African green monkey kidney epithelial cells labeled with Cy3 and Cy5 conjugated to cholera toxin, which localizes in the plasma membrane). The AOTF control of laser excitation in selected regions with confocal microscopy is also useful for investigations of protein diffusion in photoactivation studies (74–76) using fluorescent proteins, as illustrated in Fig. 10d. This image frame presents the fluorescence emission peak of the Kaede protein as it shifts from green to red in HeLa (human cervical carcinoma) cell nuclei using selected illumination (yellow box) with a 405 nanometer violet–blue diode laser.

The rapid intensity and wavelength switching capabilities of the AOTF enable sequential line scanning of multiple laser lines to be performed in which each excitation wavelength can be assigned a different intensity in order to balance the various signal levels for optimum imaging (77). Sequential scanning of individual lines minimizes the time differential between signal acquisitions from the various fluorophores while reducing crossover, which can be a significant problem with simultaneous multiple-wavelength excitation (Fig. 11). The synchronized incorporation of multiple fluorescent probes into living cells has grown into an extremely valuable technique for study of protein–protein interactions, and the dynamics of macromolecular complex assembly. The refinement of techniques for incorporating green fluorescent protein (GFP) and its numerous derivatives into the protein-synthesizing mechanisms of the cell has revolutionized living cell experimentation (78–80). A major challenge in multiple-probe studies using living tissue is the necessity to acquire the complete multispectral data set quickly enough to minimize specimen movement and molecular changes that might distort the true specimen geometry or dynamic sequence of events (32–34). The AOTF provides the speed and versatility to control the wavelength and intensity illuminating multiple specimen regions, and to simultaneously or sequentially scan each at sufficient speed to accurately monitor dynamic cellular processes.

A comparison between the application of AOTFs and neutral density filters (78) to control spectral separation of fluorophore emission spectra in confocal microscopy is presented in Fig. 11. The specimen is a monolayer culture

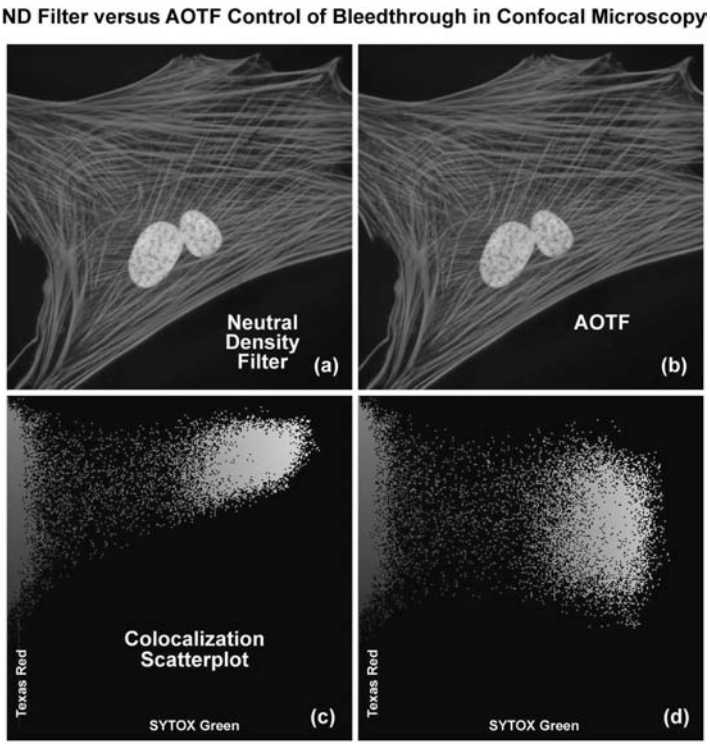


Figure 11. Fluorophore bleedthrough control with neutral density filters and sequential scanning using AOTF laser modulation. Adherent human lung fibroblast (MRC-5 line) cells were stained with Texas Red conjugated to phalloidin (actin; red) and counterstained with SYTOX green (nuclei; green). (a) Typical cell imaged with neutral density filters. (b) The same cell imaged using sequential line scanning controlled by an AOTF laser combiner. (c) and (d) Colocalization scatterplots derived from the images in (a) and (b), respectively.

of adherent human lung fibroblast (MRC-5 line) cells stained with Texas Red conjugated to phalloidin (targeting the filamentous actin network) and SYTOX Green (staining DNA in the nucleus). A neutral density filter that produces the high excitation signals necessary for both fluorophores leads to a significant amount of bleedthrough of the SYTOX Green emission into the Texas Red channel (Fig. 11a; note the yellow nuclei). The high degree of apparent colocalization between SYTOX Green and Texas Red is clearly illustrated by the scatterplot in Fig. 11b. The two axes in the scatterplot represent the SYTOX Green (abscissa) and the Texas Red (ordinate) channels. In order to balance the excitation power levels necessary to selectively illuminate each fluorophore with greater control of emission intensity, an AOTF was utilized to selectively reduce the SYTOX Green excitation power (Argon-ion laser line at 488 nm). Note the subsequent reduction in bleedthrough as manifested by green color in the cellular nuclei in Fig. 11c. The corresponding scatterplot (Fig. 11d) indicates a dramatically reduced level of bleed-through (and apparent colocalization) of SYTOX Green into the Texas Red channel.

The development of the AOTF has provided substantial additional versatility to techniques, such as fluorescence recovery after photobleaching (FRAP; 81,82), fluorescence loss in photobleaching (FLIP; 83), as well as in localized photoactivated fluorescence (uncaging; 84) studies (Fig. 10). The FRAP technique (81,82) was originally conceived to measure diffusion rates of fluorescently tagged proteins in organelles and cell membranes. In the conventional FRAP procedure, a small spot on the specimen is continuously illuminated at a low light flux level and the emitted fluorescence is measured. The illumination level is then increased to a very high level for a brief time to destroy the fluorescent molecules in the illuminated region by rapid bleaching. After the light intensity is returned to the original low level, the fluorescence is monitored to determine the rate at which new unbleached fluorescent molecules diffuse into the depleted region. The technique, as typically employed, has been limited by the fixed geometry of the bleached region, which is often a diffraction-limited spot, and by having to mechanically adjust the illumination intensity (using shutters or galvanometer-driven components). The AOTF not only allows near-instantaneous switching of light intensity, but also can be utilized to selectively bleach randomly specified regions of irregular shape, lines, or specific cellular organelles, and to determine the dynamics of molecular transfer into the region.

By enabling precise control of illuminating beam geometry and rapid switching of wavelength and intensity, the AOTF is a significant enhancement to application of the FLIP technique in measuring the diffusional mobility of certain cellular proteins (83). This technique monitors the loss of fluorescence from continuously illuminated localized regions and the redistribution of fluorophore from distant locations into the sites of depletion. The data obtained can aid in the determination of the dynamic interrelationships between intracellular and intercellular components in living tissue, and such fluorescence loss studies are greatly facilitated by the capabilities of the AOTF in controlling the microscope illumination.

The method of utilizing photoactivated fluorescence has been very useful in studies, such as those examining the role of calcium ion concentration in cellular processes, but has been limited in its sensitivity to localized regional effects in small organelles or in close proximity to cell membranes. Typically, fluorescent species that are inactivated by being bound to a photosensitive species (referred to as being caged) are activated by intense illumination that frees them from the caging compound and allows them to be tracked by the sudden appearance of fluorescence (84). The use of the AOTF has facilitated the refinement of such studies to assess highly localized processes such as calcium ion mobilization near membranes, made possible because of the precise and rapid control of the illumination triggering the activation (uncaging) of the fluorescent molecule of interest.

Because the AOTF functions, without use of moving mechanical components, to electronically control the wavelength and intensity of multiple lasers, great versatility is provided for external control and synchronization of laser illumination with other aspects of microscopy experiments. When the confocal instrument is equipped with a controller module having input and output trigger terminals, laser intensity levels can be continuously monitored and recorded, and the operation of all laser functions can be controlled to coordinate with other experimental specimen measurements, automated microscope stage movements, sequential time-lapse recording, and any number of other operations.

RESOLUTION AND CONTRAST IN CONFOCAL MICROSCOPY

All optical microscopes, including conventional widefield, confocal, and two-photon instruments are limited in the resolution that they can achieve by a series of fundamental physical factors (1,3,5–7,24,85–89). In a perfect optical system, resolution is restricted by the numerical aperture of optical components and by the wavelength of light, both incident (excitation) and detected (emission). The concept of resolution is inseparable from contrast, and is defined as the minimum separation between two points that results in a certain level of contrast between them (24). In a typical fluorescence microscope, contrast is determined by the number of photons collected from the specimen, the dynamic range of the signal, optical aberrations of the imaging system, and the number of picture elements (pixels) per unit area in the final image (66,86–88).

The influence of noise on the image of two closely spaced small objects is further interconnected with the related factors mentioned above, and can readily affect the quality of resulting images (29). While the effects of many instrumental and experimental variables on image contrast, and consequently on resolution, are familiar and rather obvious, the limitation on effective resolution resulting from the division of the image into a finite number of picture elements (pixels) may be unfamiliar to those new to digital microscopy. Because all digital confocal images employing laser scanners and/or camera systems are recorded and processed in terms of measurements made within discrete pixels (66),

some discussion of the concepts of sampling theory is required. This is appropriate to the subject of contrast and resolution because it has a direct bearing on the ability to record two closely spaced objects as being distinct.

In addition to the straightforward theoretical aspects of resolution, regardless of how it is defined, the reciprocal relationship between contrast and resolution has practical significance because the matter of interest to most microscopists is not resolution, but visibility. The ability to recognize two closely spaced features as being separate relies on advanced functions of the human visual system to interpret intensity patterns, and is a much more subjective concept than the calculation of resolution values based on diffraction theory (24). Experimental limitations and the properties of the specimen itself, which vary widely, dictate that imaging cannot be performed at the theoretical maximum resolution of the microscope.

The relationship between contrast and resolution with regard to the ability to distinguish two closely spaced specimen features implies that resolution cannot be defined without reference to contrast, and it is this interdependence that has led to considerable ambiguity involving the term resolution and the factors that influence it in microscopy (29). As discussed above, recent advances in fluorescent protein technology have led to an enormous increase in studies of dynamic processes in living cells and tissues (71–76,78–83). Such specimens are optically thick and inhomogeneous, resulting in a far-from-ideal imaging situation in the microscope. Other factors, such as cell viability and sensitivity to thermal damage and photobleaching, place limits on the light intensity and duration of exposure, consequently limiting the attainable resolution. Given that the available timescale may be dictated by these factors and by the necessity to record rapid dynamic events in living cells, it must be accepted that the quality of images will not be as high as those obtained from fixed and stained specimens. The most reasonable resolution goal for imaging in a given experimental situation is that the microscope provides the best resolution possible within the constraints imposed by the experiment.

THE AIRY DISK AND LATERAL RESOLUTION

Imaging a point-like light source in the microscope produces an electromagnetic field in the image plane whose amplitude fluctuations can be regarded as a manifestation of the response of the optical system to the specimen. This field is commonly represented through the amplitude point spread function, and allows evaluation of the optical transfer properties of the combined system components (29,86–88). Although variations in field amplitude are not directly observable, the visible image of the point source formed in the microscope and recorded by its imaging system is the intensity point spread function, which describes the system response in real space. Actual specimens are not point sources, but can be regarded as a superposition of an infinite number of objects having dimensions below the resolution of the system. The properties of the intensity point spread function (PSF; see Fig. 12) in the image plane as well as in the axial direction are major factors in determining the resolution of a microscope (1,24,29,40,85–89).

It is possible to experimentally measure the intensity point spread function in the microscope by recording the image of a subresolution spherical bead as it is scanned through focus (a number of examples may be found in the literature). Because of the technical difficulty posed in direct measurement of the intensity point spread function, calculated point spread functions are commonly utilized to evaluate the resolution performance of different optical systems, as well as the optical-sectioning capabilities of confocal, two-photon, and conventional widefield microscopes. Although the intensity point spread function extends in all three dimensions, with regard to the relationship between resolution and contrast, it is useful to consider only the lateral components of the intensity distribution, with reference to the familiar Airy disk (24).

The intensity distribution of the point-spread function in the plane of focus is described by the rotationally symmetric Airy pattern. Because of the cylindrical symmetry of the microscope lenses, the two lateral components

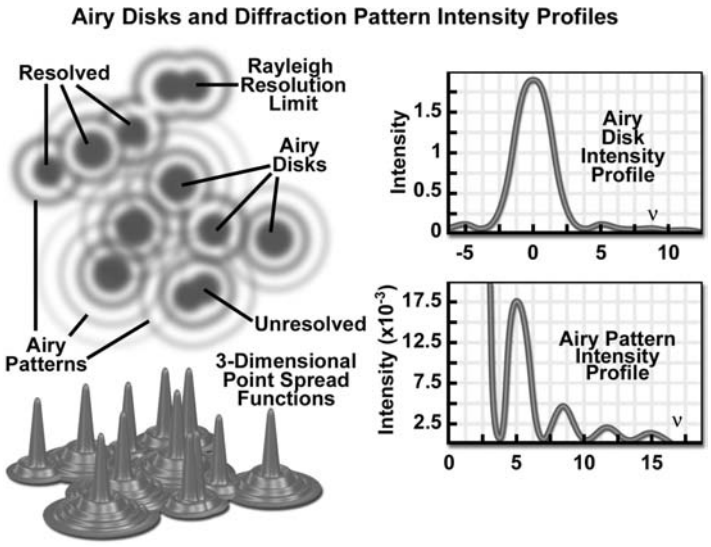


Figure 12. Schematic diagram of an Airy disk diffraction pattern and the corresponding three-dimensional point spread functions for image formation in confocal microscopy. Intensity profiles of a single Airy disk, as well as the first and higher order maxima are illustrated in the graphs.

(x and y) of the Airy pattern are equivalent, and the pattern represents the lateral intensity distribution as a function of distance from the optical axis (24). The lateral distance is normalized by the numerical aperture of the system and the wavelength of light, and therefore is dimensionless. Figure 12 (Airy disk and intensity function) illustrates diagrammatically the formation and characteristics of the Airy disk, the related 3D point spread function, and Airy patterns in the fluorescence microscope. Following the excitation of fluorophores in a point-like specimen region, fluorescence emission occurs in all directions, a small fraction of which is selected and focused by the optical components into an image plane where it forms an Airy disk surrounded by concentric rings of successively decreasing maximum and minimum intensity (the Airy pattern). The Airy pattern intensity distribution is the result of Fraunhofer diffraction of light passing through a circular aperture, and in a perfect optical system exhibits a central intensity maximum and higher order maxima separated by regions of zero intensity (85). The distance of the zero crossings from the optical axis, when the distance is normalized by the numerical aperture and wavelength, occur periodically (Fig. 12). When the intensity on the optical axis is normalized to one (100%), the proportional heights of the first four higher order maxima are 1.7%, 0.4%, 0.2%, and 0.08%, respectively.

A useful approach to the concept of resolution is based on consideration of an image formed by two point-like objects (specimen features), under the assumption that the image-forming process is incoherent, and that the interaction of the separate object images can be described using intensity point spread functions. The resulting image is then composed of the sum of two Airy disks, the characteristics of which depend on the separation distance between the two points (24,86). When sufficiently separated, the intensity change in the area between the objects is the maximum possible, cycling from the peak intensity (at the first point) to zero and returning to the maximum value at the center of the second point. At decreased distance in object space, the intensity distribution functions of the two points, in the image plane, begin to overlap and the resulting image may appear to be that of a single larger or brighter object or feature rather than being recognizable as two objects. If resolution is defined, in general terms, as the minimum separation distance at which the two objects can be sufficiently distinguished, it is obvious that this property is related to the width of the intensity peaks (the point spread function). Microscope resolution is directly related, therefore, to the full-width at half maximum (fwhm) of the instrument's intensity point spread function in the component directions (29,86,87).

Some ambiguity in use of the term resolution results from the variability in defining the degree of separation between features and their point spread functions that is sufficient to allow them to be distinguished as two objects rather than one. In general, minute features of interest in microscopy specimens produce point images that overlap to some extent, displaying two peaks separated by a gap (1,24,29,40,86). The greater the depth of the gap between the peaks, the easier it is to distinguish, or resolve, the two objects. By specifying the depth of the dip in intensity between two overlapping point spread functions, the ambiguity in evaluating resolution can be removed, and a quantitative aspect introduced.

In order to quantify resolution, the concept of contrast is employed, which is defined for two objects of equal intensity as the difference between their maximum intensity and the minimum intensity occurring in the space between them (55,86,89). Because the maximum intensity of the Airy disk is normalized to one, the highest achievable contrast is also one, and occurs only when the spacing between the two objects is relatively large, with sufficient separation to allow the first zero crossing to occur in their combined intensity distribution. At decreased distance, as the two point spread functions begin to overlap, the dip in intensity between the two maxima (and the contrast) is increasingly reduced. The distance at which two peak maxima are no longer discernible, and the contrast becomes zero, is referred to as the contrast cut-off distance (24,40). The variation of contrast with distance allows resolution, in terms of the separation of two points, to be defined as a function of contrast.

The relationship between contrast and separation distance for two point-like objects is referred to as the contrast/distance function or contrast transfer function (31,90). Resolution can be defined as the separation distance at which two objects are imaged with a certain contrast value. It is obvious that when zero contrast exists, the points are not resolved; the so-called Sparrow criterion defines the resolution of an optical system as being equivalent to the contrast cut-off distance (24). It is common, however, to specify that greater contrast is necessary to adequately distinguish two closely spaced points visually, and the well-known Rayleigh criterion (24) for resolution states that two points are resolved when the first minimum (zero crossing) of one Airy disk is aligned with the central maximum of the second Airy disk. Under optimum imaging conditions, the Rayleigh criterion separation distance corresponds to a contrast value of 26.4%. Although any contrast value >0 can be specified in defining resolution, the 26% contrast of the Rayleigh criterion is considered reasonable in typical fluorescence microscopy applications, and is the basis for the common expression defining lateral resolution according to the following equation (24), in which the point separation (r) in the image plane is the distance between the central maximum and the first minimum in the Airy disk:

$$r_{\text{lateral}} = 1.22\,\lambda/(2{\cdot}\text{NA}) = 0.6\,\lambda/\text{NA}$$

where λ is the emitted light wavelength and NA is the numerical aperture of the objective.

Resolution in the microscope is directly related to the fwhm dimensions of the microscope's point spread function, and it is common to measure this value experimentally in order to avoid the difficulty in attempting to identify intensity maxima in the Airy disk. Measurements of resolution utilizing the fwhm values of the point spread function are somewhat smaller than those calculated employing the Rayleigh criterion. Furthermore, in confocal fluorescence configurations, single-point illumination scanning and

single-point detection are employed, so that only the fluorophores in the shared volume of the illumination and detection point spread functions are able to be detected. The intensity point spread function in the confocal case is, therefore, the product of the independent illumination intensity and detection intensity point spread functions. For confocal fluorescence, the lateral (and axial) extent of the point spread function is reduced by ~30% compared to that in the wide-field microscope. Because of the narrower intensity point spread function, the separation of points required to produce acceptable contrast in the confocal microscope (29,31) is reduced to a distance approximated by

$$r_{\text{lateral}} = 0.4\lambda/\text{NA}$$

If the illumination and fluorescence emission wavelengths are approximately the same, the confocal fluorescence microscope Airy disk size is the square of the wide-field microscope Airy disk. Consequently, the contrast cut-off distance is reduced in the confocal arrangement, and equivalent contrast can be achieved at a shorter distance compared to the widefield illumination configuration. Regardless of the instrument configuration, the lateral resolution displays a proportional relationship to wavelength, and is inversely proportional to the objective lens numerical aperture.

As noted previously, lateral resolution is of primary interest in discussing resolution and contrast, although the axial extent of the microscope intensity point spread function is similarly reduced in the confocal arrangement as compared to the widefield fluorescence configuration (86,89). Reasonable contrast between point-like objects lying on the optical axis occurs when they are separated by the distance between the central maximum and the first minimum of the axial point spread function component.

Figure 13 presents the axial intensity distributions (89) for a typical widefield (Fig. 13a) and confocal (Fig. 13b) fluorescence microscope. Note the dramatic reduction in intensity of the wings in the confocal distribution as a function of distance from the central maximum.

A variety of equations are presented in the literature that pertains to different models for calculating axial resolution for various microscope configurations. The ones most applicable to fluorescence emission are similar in form to the expressions evaluating depth of field, and demonstrate that axial resolution is proportional to the wavelength, and refractive index of the specimen medium, and inversely proportional to the square of the numerical aperture. Consequently, the NA of the microscope objective has a much greater effect on axial resolution than does the emission wavelength. One equation (89) commonly used to describe axial resolution for the confocal configuration is given below, with η representing the index of refraction, and the other variables as specified previously:

$$r_{\text{axial}} = 1.4\ \lambda \cdot \eta/\text{NA}^2$$

Although the confocal microscope configuration exhibits only a modest improvement in measured axial resolution over that of the widefield microscope, the true advantage of the confocal approach is in the optical sectioning capability in thick specimens, which results in a dramatic improvement in effective axial resolution over conventional techniques. The optical sectioning properties of the confocal microscope result from the characteristics of the integrated intensity point spread function, which has a maximum in the focal plane when evaluated as a function of depth. The equivalent integral of intensity point spread function for the conventional widefield microscope is constant as a function of depth, producing no optical sectioning capabilities.

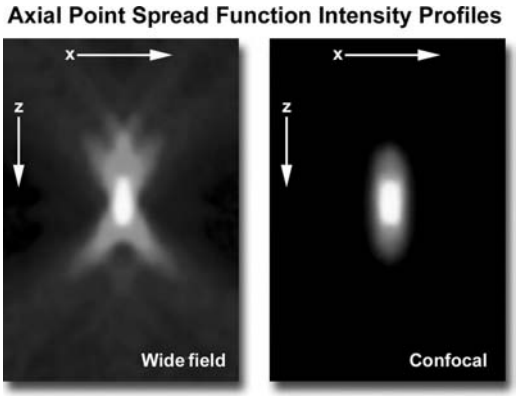


Figure 13. Comparison of axial (x–z) point spread functions for widefield (left) and confocal (right) microscopy.

FLUOROPHORES FOR CONFOCAL MICROSCOPY

Biological laser scanning confocal microscopy relies heavily on fluorescence as an imaging mode, primarily due to the high degree of sensitivity afforded by the technique coupled with the ability to specifically target structural components and dynamic processes in chemically fixed as well as living cells and tissues. Many fluorescent probes are constructed around synthetic aromatic organic chemicals designed to bind with a biological macromolecule (e.g., a protein or nucleic acid) or to localize within a specific structural region, such as the cytoskeleton, mitochondria, Golgi apparatus, endoplasmic reticulum, and nucleus (90). Other probes are employed to monitor dynamic processes and localized environmental variables, including concentrations of inorganic metallic ions, pH, reactive oxygen species, and membrane potential (91). Fluorescent dyes are also useful in monitoring cellular integrity (live versus dead and apoptosis), endocytosis, exocytosis, membrane fluidity, protein trafficking, signal transduction, and enzymatic activity (92). In addition, fluorescent probes have been widely applied to genetic mapping and chromosome analysis in the field of molecular genetics.

The history of synthetic fluorescent probes dates back over a century to the late-1800s when many of the cornerstone dyes for modern histology were developed. Among these were pararosaniline, methyl violet, malachite green, safranin O, methylene blue, and numerous azo (nitrogen) dyes, such as Bismarck brown (93). Although these dyes were highly colored and capable of absorbing selected bands of visible light, most were only weakly fluorescent and would not be useful for the fluorescence microscopes that would be developed several decades later. However,

several synthetic dye classes synthesized during this period, based on the xanthene and acridine heterocyclic ring systems, proved to be highly fluorescent and provided a foundation for the development of modern synthetic fluorescent probes. Most notable among these early fluorescent dyes were the substituted xanthenes, fluorescein and rhodamine B, and the biaminated acridine derivative, acridine orange.

Fluorochromes were introduced to fluorescence microscopy in the early twentieth century as vital stains for bacteria, protozoa, and trypanosomes, but did not see widespread use until the 1920s when fluorescence microscopy was first used to study dye binding in fixed tissues and living cells (7,93). However, it was not until the early 1940s that Coons developed a technique for labeling antibodies with fluorescent dyes, thus giving birth to the field of immunofluorescence (94). Over the past 60 years, advances in immunology and molecular biology have produced a wide spectrum of secondary antibodies and provided insight into the molecular design of fluorescent probes targeted at specific regions within macromolecular complexes.

Fluorescent probe technology and cell biology were dramatically altered by the discovery of the GFP from jellyfish and the development of mutant spectral variants, which have opened the door to noninvasive fluorescence multicolor investigations of subcellular protein localization, intermolecular interactions, and trafficking using living cell cultures (79,80,95). More recently, the development of nanometer-sized fluorescent semiconductor quantum dots has provided a new avenue for research in confocal and widefield fluorescence microscopy (96). Despite the numerous advances made in fluorescent dye synthesis during the past few decades, there is very little solid evidence about molecular design rules for developing new fluorochromes, particularly with regard to matching absorption spectra to available confocal laser excitation wavelengths. As a result, the number of fluorophores that have found widespread use in confocal microscopy is a limited subset of the many thousands that have been discovered.

BASIC CHARACTERISTICS OF FLUOROPHORES

Fluorophores are catalogued and described according to their absorption and fluorescence properties, including the spectral profiles, wavelengths of maximum absorbance and emission, and the fluorescence intensity of the emitted light (92). One of the most useful quantitative parameters for characterizing absorption spectra is the molar extinction coefficient (denoted with the Greek symbole, see Fig. 14a), which is a direct measure of the ability of a molecule to absorb light. The extinction coefficient is useful for converting units of absorbance into units of molar concentration, and is determined by measuring the absorbance at a reference wavelength (usually the maximum, characteristic of the absorbing species) for a molar concentration in a defined optical path length. The quantum yield of a fluorochrome or fluorophore represents a quantitative measure of fluorescence emission efficiency, and is expressed as the ratio of the number of photons emitted to the number of photons absorbed. In other words, the quantum

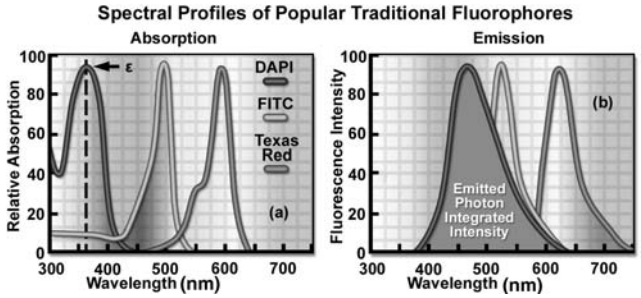


Figure 14. Fluorescent spectral profiles, plotted as normalized absorption or emission as a function of wavelength, for popular synthetic fluorophores emitting in the blue, green, and red regions of the visible spectrum. Each profile is identified with a colored bullet in (a), which illustrates excitation spectra. (b) The emission spectra for the fluorophores according to the legend in (a).

yield represents the probability that a given excited fluorochrome will produce an emitted (fluorescence) photon. Quantum yields typically range between a value of 0 and 1 and fluorescent molecules commonly employed as probes in microscopy have quantum yields ranging from very low (0.05 or less) to almost unity. In general, a high quantum yield is desirable in most imaging applications. The quantum yield of a given fluorophore varies, sometimes to large extremes, with environmental factors, such as metallic ion concentration, pH, and solvent polarity (92).

In most cases, the molar extinction coefficient for photon absorption is quantitatively measured and expressed at a specific wavelength, whereas the quantum efficiency is an assessment of the total integrated photon emission over the entire spectral band of the fluorophore (Fig. 14b). As opposed to traditional arc-discharge lamps used with the shortest range (10–20 nm) bandpass interference filters in wide-field fluorescence microscopy, the laser systems used for fluorophore excitation in scanning confocal microscopy restrict excitation to specific laser spectral lines that encompass only a few nanometers (1,7). The fluorescence emission spectrum for both techniques, however, is controlled by similar bandpass or longpass filters that can cover tens to hundreds of nanometers (7). Below saturation levels, fluorescence intensity is proportional to the product of the molar extinction coefficient and the quantum yield of the fluorophore, a relationship that can be utilized to judge the effectiveness of emission as a function of excitation wavelength(s). These parameters display an approximate 20-fold range in variation for the popular fluorophores commonly employed for investigations in confocal microscopy with quantum yields ranging from 0.05 to 1.0, and extinction coefficients ranging from 10,000 to 0.25 million ($L \cdot mol^{-1}$). In general, the absorption spectrum of a fluorophore is far less dependent on environmental conditions than the fluorescence emission characteristics (spectral wavelength profile and quantum yield; 92).

Fluorophores chosen for confocal applications must exhibit a brightness level and signal persistence sufficient for the instrument to obtain image data that does not suffer from excessive photobleaching artifacts and low signal/noise ratios. In widefield fluorescence microscopy, excitation illumination levels are easily controlled with neutral

Table 1. Laser and Arc-Discharge Spectral Lines in Widefield and Confocal Microscopy

Laser Type	Ultraviolet	Violet	Blue	Green	Yellow	Orange	Red
Argon-ion	351, 364		457, 477, 488	514			
Blue diode		405, 440					
Diode-pumped solid state	355	430, 442	457, 473	532	561		
Helium–cadmium	322, 354	442					
Krypton–argon			488		568		647
Green helium–neon				543			
Yellow helium–neon					594		
Orange helium–neon						612	
Red helium-neon							633
Red diode							635, 650
Mercury arc	365	405, 436	546		579		
Xenon arc		467					

density filters (40), and the intensity can be reduced (coupled with longer emission signal collection periods) to avoid saturation and curtail irreversible loss of fluorescence. Excitation conditions in confocal microscopy are several orders of magnitude more severe, however, and restrictions imposed by characteristics of the fluorophores and efficiency of the microscope optical system become the dominating factor in determining excitation rate and emission collection strategies (1,7,92).

Because of the narrow and wavelength-restricted laser spectral lines employed to excite fluorophores in confocal microscopy (Table 1), fluorescence emission intensity can be seriously restricted due to poor overlap of the excitation wavelengths with the fluorophore absorption band. In addition, the confocal pinhole aperture, which is critical in obtaining thin optical sections at high signal/noise ratios, is responsible for a 25–50% loss of emission intensity, regardless of how much effort has been expended on fine-tuning and alignment of the microscope optical system (7). Photomultiplier tubes are the most common detectors in confocal microscopy, but suffer from a quantum efficiency that varies as a function of wavelength (especially in the red and IR regions), further contributing to a wavelength-dependent loss of signal across the emission spectrum (59–62). Collectively, the light losses in confocal microscopy can result in a reduction of intensity exceeding 50 times of the level typically observed in widefield fluorescence instruments. It should be clear from the preceding argument that fluorophore selection is one of the most critical aspects of confocal microscopy, and instrumental efficiency must be carefully considered, as well, in order to produce high quality images.

In confocal microscopy, irradiation of the fluorophores with a focused laser beam at high power densities increases the emission intensity up to the point of dye saturation, a condition whose parameters are dictated by the excited state lifetime (97). In the excited state, fluorophores are unable to absorb another incident photon until they emit a lower energy photon through the fluorescence process. When the rate of fluorophore excitation exceeds the rate of emission decay, the molecules become saturated and the ground state population decreases. As a result, a majority of the laser energy passes through the specimen undiminished and does not contribute to fluorophore excitation. Balancing fluorophore saturation with laser light intensity levels is, therefore, a critical condition for achieving the optimal signal/noise ratio in confocal experiments (1,7,92,97). The number of fluorescent probes currently available for confocal microscopy runs in the hundreds (90,93), with many dyes having absorption maxima closely associated with common laser spectral lines (90). An exact match between a particular laser line and the absorption maximum of a specific probe is not always possible, but the excitation efficiency of lines near the maximum is usually sufficient to produce a level of fluorescence emission that can be readily detected.

Instrumentally, fluorescence emission collection can be optimized by careful selection of objectives, detector aperture dimensions, dichromatic and barrier filters, as well as maintaining the optical train in precise alignment (63). In most cases, low magnification objectives with a high numerical aperture should be chosen for the most demanding imaging conditions because light collection intensity increases as the fourth power of the numerical aperture, but only decreases as the square of the magnification. However, the most important limitations in light collection efficiency in confocal microscopy arise from restrictions imposed by the physical properties of the fluorophores themselves. As previously discussed, fluorescent probe development is limited by a lack of knowledge of the specific molecular properties responsible for producing optimum fluorescence characteristics, and the design rules are insufficiently understood to be helpful as a guide to the development of more efficient fluorophores. The current success in development of new fluorescent probes capable of satisfactory performance in confocal microscopy is a testament to the progress made through use of empirical data and assumptions about molecular structure extrapolated from the properties of existing dyes, many of which were first synthesized over a hundred years ago.

TRADITIONAL FLUORESCENT DYES

The choice of fluorescent probes for confocal microscopy must address the specific capabilities of the instrument to excite and detect fluorescence emission in the wavelength regions made available by the laser systems and detectors. Although the current lasers used in confocal microscopy (Table 1) produce discrete lines in the UV, visible, and

near-IR portions of the spectrum, the location of these spectral lines does not always coincide with absorption maxima of popular fluorophores. In fact, it is not necessary for the laser spectral line to correspond exactly with the fluorophore wavelength of maximum absorption, but the intensity of fluorescence emission is regulated by the fluorophore extinction coefficient at the excitation wavelength (as discussed above). The most popular lasers for confocal microscopy are air-cooled argon and krypton–argon ion lasers, the new blue diode lasers, and a variety of helium–neon systems (7,40). Collectively, these lasers are capable of providing excitation at 10–12 specific wavelengths between 400 and 650 nm.

Many of the classical fluorescent probes that have been successfully utilized for many years in widefield fluorescence (92,93), including fluorescein isothiocyanate, Lissamine rhodamine, and Texas red, are also useful in confocal microscopy. Fluorescein is one of the most popular fluorochromes ever designed, and has enjoyed extensive application in immunofluorescence labeling. This xanthene dye has an absorption maximum at 495 nm, which coincides quite well with the 488 nm (blue) spectral line produced by argon-ion and krypton–argon lasers, as well as the 436 and 467 principal lines of the mercury and xenon arc-discharge lamps, respectively. In addition, the quantum yield of fluorescein is very high and a significant amount of information has been gathered on the characteristics of this dye with respect to the physical and chemical properties (98). On the negative side, the fluorescence emission intensity of fluorescein is heavily influenced by environmental factors (e.g., pH), and the relatively broad emission spectrum often overlaps with those of other fluorophores in dual and triple labeling experiments (92,98,99).

Tetramethyl rhodamine (TMR) and the isothiocyanate derivative (TRITC) are frequently employed in multiple labeling investigations in widefield microscopy due to their efficient excitation by the 546 nm spectral line from mercury arc-discharge lamps. The fluorochromes, which have significant emission spectral overlap with fluorescein, can be excited very effectively by the 543 nm line from helium–neon lasers, but not by the 514 or 568 nanometer lines from argon-ion and krypton–argon lasers (99). When using krypton-based laser systems, Lissamine rhodamine is a far better choice in this fluorochrome class due to the absorption maximum at 575 nm and its spectral separation from fluorescein. Also, the fluorescence emission intensity of rhodamine derivatives is not as dependent upon strict environmental conditions as that of fluorescein.

Several of the acridine dyes, first isolated in the nineteenth century, are useful as fluorescent probes in confocal microscopy (93). The most widely utilized, acridine orange, consists of the basic acridine nucleus with dimethylamino substituents located at the 3 and 6 positions of the trinuclear ring system. In physiological pH ranges, the molecule is protonated at the heterocyclic nitrogen and exists predominantly as a cationic species in solution. Acridine orange binds strongly to DNA by intercalation of the acridine nucleus between successive base pairs, and exhibits green fluorescence with a maximum wavelength of 530 nm (92,93,100). The probe also binds strongly to ribonucleic acid (RNA) or single-stranded deoxyribonucleic acid (DNA), but has a longer wavelength fluorescence maximum ($\sim$ 640 nm; red) when bound to these macromolecules. In living cells, acridine orange diffuses across the cell membrane (by virtue of the association constant for protonation) and accumulates in the lysosomes and other acidic vesicles. Similar to most acridines and related polynuclear nitrogen heterocycles, acridine orange has a relatively broad absorption spectrum, which enables the probe to be used with several wavelengths from the argon-ion laser.

Another popular traditional probe that is useful in confocal microscopy is the phenanthridine derivative, propidium iodide, first synthesized as an antitrypanosomal agent along with the closely related ethidium bromide). Propidium iodide binds to DNA in a manner similar to the acridines (via intercalation) to produce orange-red fluorescence centered at 617 nm (101,102). The positively charged fluorophore also has a high affinity for double-stranded RNA. Propidium has an absorption maximum at 536 nm, and can be excited by the 488 or 514-nm spectral lines of an argon-ion (or krypton–argon) laser, or the 543 nm line from a green helium–neon laser. The dye is often employed as a counterstain to highlight cell nuclei during double or triple labeling of multiple intracellular structures. Environmental factors can affect the fluorescence spectrum of propidium, especially when the dye is used with mounting media containing glycerol. The structurally similar ethidium bromide, which also binds to DNA by intercalation (102), produces more background staining, and is therefore not as effective as propidium.

The DNA and chromatin can also be stained with dyes that bind externally to the double helix. The most popular fluorochromes in this category are 4′,6-diamidino-2-phenylindole (DAPI) and the bis (benzimide) Hoechst dyes that are designated by the numbers 33258, 33342, and 34580 (103–106). These probes are quite water soluble and bind externally to AT-rich base pair clusters in the minor groove of double-stranded DNA with a dramatic increase in fluorescence intensity. Both dye classes can be stimulated by the 351 nm spectral line of high power argon-ion lasers or the 354 nm line from a helium–cadmium laser. Similar to the acridines and phenanthridines, these fluorescent probes are popular choices as a nuclear counterstain for use in multicolor fluorescent labeling protocols. The vivid blue fluorescence emission produces dramatic contrast when coupled to green, yellow, and red probes in adjacent cellular structures.

ALEXA FLUOR DYES

The dramatic advances in modern fluorophore technology are exemplified by the Alexa Fluor dyes (90,107,108) introduced by Molecular Probes (Alexa Fluor is a registered trademark of Molecular Probes). These sulfonated rhodamine derivatives exhibit higher quantum yields for more intense fluorescence emission than spectrally similar probes, and have several additional improved features, including enhanced photostability, absorption spectra matched to common laser lines, pH insensitivity, and a high degree of water solubility. In fact, the resistance to photobleaching of Alexa Fluor dyes is so dramatic (108)

that even when subjected to irradiation by high intensity laser sources, fluorescence intensity remains stable for relatively long periods of time in the absence of antifade reagents. This feature enables the water soluble Alexa Fluor probes to be readily utilized for both live-cell and tissue section investigations, as well as in traditional fixed preparations.

Alexa Fluor dyes are available in a broad range of fluorescence excitation and emission wavelength maxima, ranging from the UV and deep blue to the near-IR regions (90). Alphanumeric names of the individual dyes are associated with the specific excitation laser or arc-discharge lamp spectral lines for which the probes are intended. For example, Alexa Fluor 488 is designed for excitation by the blue 488 nm line of the argon or krypton–argon ion lasers, while Alexa Fluor 568 is matched to the 568 nm spectral line of the krypton–argon laser. Several of the Alexa Fluor dyes are specifically designed for excitation by either the blue diode laser (405 nm), the orange/yellow helium–neon laser (594 nm), or the red helium–neon laser (633 nm). Other Alexa Fluor dyes are intended for excitation with traditional mercury arc-discharge lamps in the visible (Alexa Fluor 546) or UV (Alexa Fluor 350, also useful with high power argon-ion lasers), and solid-state red diode lasers (Alexa Fluor 680). Because of the large number of available excitation and emission wavelengths in the Alexa Fluor series, multiple labeling experiments can often be conducted exclusively with these dyes.

Alexa Fluor dyes are commercially available as reactive intermediates in the form of maleimides, succinimidyl esters, and hydrazides, as well as prepared cytoskeletal probes (conjugated to phalloidin, G-actin, and rabbit skeletal muscle actin) and conjugates to lectin, dextrin, streptavidin, avidin, biocytin, and a wide variety of secondary antibodies (90). In the latter forms, the Alexa Fluor fluorophores provide a broad palette of tools for investigations in immunocytochemistry, neuroscience, and cellular biology. The family of probes has also been extended into a series of dyes having overlapping fluorescence emission maxima targeted at sophisticated confocal microscopy detection systems with spectral imaging and linear unmixing capabilities. For example, Alexa Fluor 488, Alexa Fluor 500, and Alexa Fluor 514 are visually similar in color with bright green fluorescence, but have spectrally distinct emission profiles. In addition, the three fluorochromes can be excited with the 488 or 514 nm spectral line from an argon-ion laser and are easily detected with traditional fluorescein filter combinations. In multispectral (*x–y–l*; referred to as a lambda stack) confocal imaging experiments, optical separation software can be employed to differentiate between the similar signals (32–35). The overlapping emission spectra of Alexa Fluor 488, 500, and 514 are segregated into separate channels and differentiated using pseudocolor techniques when the three fluorophores are simultaneously combined in a triple label investigation.

CYANINE DYES

The popular family of cyanine dyes, Cy2, Cy3, Cy5, Cy7, and their derivatives, are based on the partially saturated indole nitrogen heterocyclic nucleus with two aromatic units being connected via a polyalkene bridge of varying carbon number (92,109). These probes exhibit fluorescence excitation and emission profiles that are similar to many of the traditional dyes, such as fluorescein and tetramethylrhodamine, but with enhanced water solubility, photostability, and higher quantum yields. Most of the cyanine dyes are more environmentally stable than their traditional counterparts, rendering their fluorescence emission intensity less sensitive to pH and organic mounting media. In a manner similar to the Alexa Fluors, the excitation wavelengths of the Cy series of synthetic dyes are tuned specifically for use with common laser and arc-discharge sources, and the fluorescence emission can be detected with traditional filter combinations.

Marketed by a number of distributors, the cyanine dyes are readily available as reactive dyes or fluorophores coupled to a wide variety of secondary antibodies, dextrin, streptavidin, and eggwhite avidin (110). The cyanine dyes generally have broader absorption spectral regions than members of the Alexa Fluor family, making them somewhat more versatile in the choice of laser excitation sources for confocal microscopy (7). For example, using the 547 nm spectral line from an argon-ion laser, Cy2 is about twice as efficient in fluorescence emission as Alexa Fluor 488. In an analogous manner, the 514 nm argon-ion laser line excites Cy3 with a much higher efficiency than Alexa Fluor 546, a spectrally similar probe. Emission profiles of the cyanine dyes are comparable in spectral width to the Alexa Fluor series.

Included in the cyanine dye series are the long-wavelength Cy5 derivatives, which are excited in the red region (650 nm) and emit in the far-red (680 nm) wavelengths. The Cy5 fluorophore is very efficiently excited by the 647 nm spectral line of the krypton–argon laser, the 633 nm line of the red helium–neon laser, or the 650 nm line of the red diode laser, providing versatility in laser choice. Because the emission spectral profile is significantly removed from traditional fluorophores excited by UV and blue illumination, Cy5 is often utilized as a third fluorophore in triple labeling experiments. However, similar to other probes with fluorescence emission in the far-red spectral region, Cy5 is not visible to the human eye and can only be detected electronically (using a specialized CCD camera system or photomultiplier). Therefore, the probe is seldom used in conventional widefield fluorescence experiments.

FLUORESCENT ENVIRONMENTAL PROBES

Fluorophores designed to probe the internal environment of living cells have been widely examined by a number of investigators, and many hundreds have been developed to monitor such effects as localized concentrations of alkali and alkaline earth metals, heavy metals (employed biochemically as enzyme cofactors), inorganic ions, thiols, and sulfides, nitrite, as well as pH, solvent polarity, and membrane potential (7,90–93,111,112). Originally, the experiments in this arena were focused on changes in the wavelength and/or intensity of absorption and emission spectra exhibited by fluorophores upon binding calcium

ions in order to measure intracellular flux densities. These probes bind to the target ion with a high degree of specificity to produce the measured response and are often referred to as spectrally sensitive indicators. Ionic concentration changes are determined by the application of optical ratio signal analysis to monitor the association equilibrium between the ion and its host. The concentration values derived from this technique are largely independent of instrumental variations and probe concentration fluctuations due to photobleaching, loading parameters, and cell retention. In the past few years, a number of new agents have been developed that bind specific ions or respond with measurable features to other environmental conditions (7,90).

Calcium is a metabolically important ion that plays a vital role in cellular response to many forms of external stimuli (113). Because transient fluctuations in calcium ion concentration are typically involved when cells undergo a response, fluorophores must be designed to measure not only localized concentrations within segregated compartments, but should also produce quantitative changes when flux density waves progress throughout the entire cytoplasm. Many of the synthetic molecules designed to measure calcium levels are based on the nonfluorescent chelation agents EGTA and BAPTA, which have been used for years to sequester calcium ions in buffer solutions (7,114,115). Two of the most common calcium probes are the ratiometric indicators fura-2 and indo-1, but these fluorophores are not particularly useful in confocal microscopy (7,116). The dyes are excited by UV light and exhibit a shift in the excitation or emission spectrum with the formation of isosbestic points when binding calcium. However, the optical aberrations associated with UV imaging, limited specimen penetration depths, and the expense of ultraviolet lasers have limited the utility of these probes in confocal microscopy.

Fluorophores that respond in the visible range to calcium ion fluxes are, unfortunately, not ratiometric indicators and do not exhibit a wavelength shift (typical of fura-2 and indo-1) upon binding, although they do undergo an increase or decrease in fluorescence intensity. The best example is fluo-3, a complex xanthene derivative, which undergoes a dramatic increase in fluorescence emission at 525 nm (green) when excited by the 488 nm spectral line of an argon-ion or krypton–argon laser (7,117). Because isosbestic points are not present to assure the absence of concentration fluctuations, it is impossible to determine whether spectral changes are due to complex formation or a variation in concentration with fluo-3 and similar fluorophores.

To overcome the problems associated with using visible light probes lacking wavelength shifts (and isosbestic points), several of these dyes are often utilized in combination for calcium measurements in confocal microscopy (118). Fura red, a multinuclear imidazole and benzofuran heterocycle, exhibits a decrease in fluorescence at 650 nm when binding calcium. A ratiometric response to calcium ion fluxes can be obtained when a mixture of fluo-3 and fura red is excited at 488 nm and fluorescence is measured at the emission maxima (525 and 650 nm, respectively) of the two probes. Because the emission intensity of fluo-3 increases monotonically while that of fura red simultaneously decreases, an isosbestic point is obtained when the dye concentrations are constant within the localized area being investigated. Another benefit of using these probes together is the ability to measure fluorescence intensity fluctuations with a standard FITC/Texas red interference filter combination.

Quantitative measurements of ions other than calcium, such as magnesium, sodium, potassium, and zinc, are conducted in an analogous manner using similar fluorophores (7,90,92). One of the most popular probes for magnesium, mag-fura-2 (structurally similar to fura red), is also excited in the ultraviolet range and presents the same problems in confocal microscopy as fura-2 and indo-1. Fluorophores excited in the visible light region are becoming available for the analysis of many monovalent and divalent cations that exist at varying concentrations in the cellular matrix. Several synthetic organic probes have also been developed for monitoring the concentration of simple and complex anions.

Important fluorescence monitors for intracellular pH include a pyrene derivative known as HPTS or pyranine, the fluorescein derivative, BCECF, and another substituted xanthene termed carboxy SNARF-1 (90,119–122). Because many common fluorophores are sensitive to pH in the surrounding medium, changes in fluorescence intensity that are often attributed to biological interactions may actually occur as a result of protonation. In the physiological pH range (pH 6.8–7.4), the probes mentioned above are useful for dual-wavelength ratiometric measurements and differ only in dye loading parameters. Simultaneous measurements of calcium ion concentration and pH can often be accomplished by combining a pH indicator, such as SNARF-1, with a calcium ion indicator (e.g., fura-2). Other probes have been developed for pH measurements in subcellular compartments, such as the lysosomes, as described below.

ORGANELLE PROBES

Fluorophores targeted at specific intracellular organelles, such as the mitochondria, lysosomes, Golgi apparatus, and endoplasmic reticulum, are useful for monitoring a variety of biological processes in living cells using confocal microscopy (7,90,92). In general, organelle probes consist of a fluorochrome nucleus attached to a target-specific moiety that assists in localizing the fluorophore through covalent, electrostatic, hydrophobic, or similar types of bonds. Many of the fluorescent probes designed for selecting organelles are able to permeate or sequester within the cell membrane (and therefore, are useful in living cells), while others must be installed using monoclonal antibodies with traditional immunocytochemistry techniques. In living cells, organelle probes are useful for investigating transport, respiration, mitosis, apoptosis, protein degradation, acidic compartments, and membrane phenomena. Cell impermeant fluorophore applications include nuclear functions, cytoskeletal structure, organelle detection, and probes for membrane integrity. In many cases, living cells that have been labeled with permeant probes can subsequently be fixed and counterstained with additional fluorophores in multicolor labeling experiments.

Mitochondrial probes are among the most useful fluorophores for investigating cellular respiration and are often employed along with other dyes in multiple labeling investigations. The traditional probes, rhodamine 123 and tetramethylrosamine, are rapidly lost when cells are fixed and have largely been supplanted by newer, more specific, fluorophores developed by Molecular Probes (90,123,124). These include the popular MitoTracker and MitoFluor series of structurally diverse xanthene, benzoxazole, indole, and benzimidazole heterocycles that are available in a variety of excitation and emission spectral profiles. The mechanism of action varies for each of the probes in this series, ranging from covalent attachment to oxidation within respiring mitochondrial membranes.

MitoTracker dyes are retained quite well after cell fixation in formaldehyde and can often withstand lipophilic permeabilizing agents (123). In contrast, the MitoFluor probes are designed specifically for actively respiring cells and are not suitable for fixation and counterstaining procedures (90). Another popular mitochondrial probe, entitled JC-1, is useful as an indicator of membrane potential and in multiple staining experiments with fixed cells (125). This carbocyanine dye exhibits green fluorescence at low concentrations, but can undergo intramolecular association within active mitochondria to produce a shift in emission to longer (red) wavelengths. The change in emission wavelength is useful in determining the ratio of active to nonactive mitochondria in living cells.

In general, weakly basic amines that are able to pass through membranes are the ideal candidates for investigating biosynthesis and pathogenesis in lysosomes (90–92,112). Traditional lysosomal probes include the nonspecific phenazine and acridine derivatives neutral red and acridine orange, which are accumulated in the acidic vesicles upon being protonated (92,93). Fluorescently labeled latex beads and macromolecules, such as dextran, can also be accumulated in lysosomes by endocytosis for a variety of experiments. However, the most useful tools for investigating lysosomal properties with confocal microscopy are the LysoTracker and LysoSensor dyes developed by Molecular Probes (90,92,126). These structurally diverse agents contain heterocyclic and aliphatic nitrogen moieties that modulate transport of the dyes into the lysosomes of living cells for both short- and long-term studies. The LysoTracker probes, which are available in a variety of excitation and emission wavelengths (91), have high selectivity for acidic organelles and are capable of labeling cells at nanomolar concentrations. Several of the dyes are retained quite well after fixing and permeabilization of cells. In contrast, the LysoSensor fluorophores are designed for studying dynamic aspects of lysosome function in living cells. Fluorescence intensity dramatically increases in the LysoSensor series upon protonation, making these dyes useful as pH indicators (91). A variety of Golgi apparatus specific monoclonal antibodies have also been developed for use in immunocytochemistry assays (90,127–129).

Proteins and lipids are sorted and processed in the Golgi apparatus, which is typically stained with fluorescent derivatives of ceramides and sphingolipids (130). These agents are highly lipophilic, and are therefore useful as markers for the study of lipid transport and metabolism in live cells. Several of the most useful fluorophores for Golgi apparatus contain the complex heterocyclic BODIPY nucleus developed by Molecular Probes (90,92,131). When coupled to sphingolipids, the BODIPY fluorophore is highly selective and exhibits a tolerance for photobleaching that is far superior to many other dyes. In addition, the emission spectrum is dependent upon concentration (shifting from green to red at higher concentrations), making the probes useful for locating and identifying intracellular structures that accumulate large quantities of lipids. During live-cell experiments, fluorescent lipid probes can undergo metabolism to derivatives that may bind to other subcellular features, a factor that can often complicate the analysis of experimental data.

The most popular traditional probes for endoplasmic reticulum fluorescence analysis are the carbocyanine and xanthene dyes, DiOC (6) and several rhodamine derivatives, respectively (90,92). These dyes must be used with caution, however, because they can also accumulate in the mitochondria, Golgi apparatus, and other intracellular lipophilic regions. Newer, more photostable, probes have been developed for selective staining of the endoplasmic reticulum by several manufacturers. In particular, oxazole members of the Dapoxyl family produced by Molecular Probes are excellent agents for selective labeling of the endoplasmic reticulum in living cells, either alone or in combination with other dyes (90). These probes are retained after fixation with formaldehyde, but can be lost with permeabilizing detergents. Another useful probe is Brefeldin A (131), a stereochemically complex fungal metabolite that serves as an inhibitor of protein trafficking out of the endoplasmic reticulum. Finally, similar to other organelles, monoclonal antibodies (127–129) have been developed that target the endoplasmic reticulum in fixed cells for immunocytochemistry investigations.

QUANTUM DOTS

Nanometer-sized crystals of purified semiconductors known as quantum dots are emerging as a potentially useful fluorescent labeling agent for living and fixed cells in both traditional widefield and laser scanning confocal fluorescence microscopy (132–136). Recently introduced techniques enable the purified tiny semiconductor crystals to be coated with a hydrophilic polymer shell and conjugated to antibodies or other biologically active peptides and carbohydrates for application in many of the classical immunocytochemistry protocols (Fig. 15). These probes have significant benefits over organic dyes and fluorescent proteins, including long-term photostability, high fluorescence intensity levels, and multiple colors with single-wavelength excitation for all emission profiles (136).

Quantum dots produce illumination in a manner similar to the well-known semiconductor light emitting diodes, but are activated by absorption of a photon rather than an electrical stimulus. The absorbed photon creates an electron-hole pair that quickly recombines with the concurrent emission of a photon having lower energy. The most useful semiconductor discovered thus far for producing biological quantum dots is cadmium selenide (CdSe), a material in

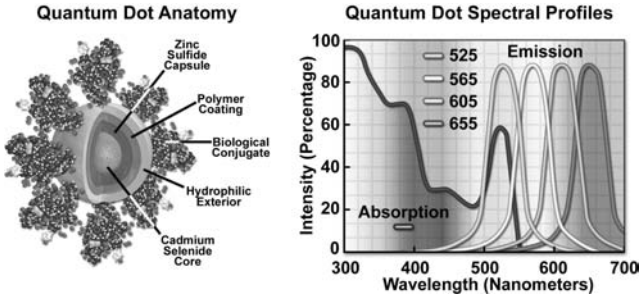


Figure 15. Anatomy and spectral profiles of quantum dot conjugates. The cadmium selenide core is encapsulated with zinc sulfide, and then a polymer coating is applied followed by a hydrophilic exterior to which the biological conjugate is attached (left). The absorption profile displays a shoulder at 400 nm, while the emission spectra all feature similar symmetrical profiles.

which the energy of the emitted photons is a function of the physical size of the nanocrystal particles. Thus, quantum dots having sizes that differ only by tenths of a nanometer emit different wavelengths of light, with the smaller sizes emitting shorter wavelengths, and vice versa.

Unlike typical organic fluorophores or fluorescent proteins, which display highly defined spectral profiles, quantum dots have an absorption spectrum that increases steadily with decreasing wavelength (Fig. 15). Also, in contrast, the fluorescence emission intensity is confined to a symmetrical peak with a maximum wavelength that is dependent on the dot size, but independent of the excitation wavelength (135). As a result, the same emission profile is observed regardless of whether the quantum dot is excited at 300, 400, 500, or 600 nm, but the fluorescence intensity increases dramatically at shorter excitation wavelengths. For example, the extinction coefficient for a typical quantum dot conjugate that emits in the orange region (605 nm) is approximately five-fold higher when the semiconductor is excited at 400 versus 600 nm. The fwhm value for a typical quantum dot conjugate is ~ 30 nm (135), and the spectral profile is not skewed towards the longer wavelengths (having higher intensity tails), such is the case with most organic fluorochromes. The narrow emission profile enables several quantum dot conjugates to be simultaneously observed with a minimal level of bleed through.

For biological applications, a relatively uniform population of cadmium selenide crystals is covered with a surrounding semiconductor shell composed of zinc sulfide to improve the optical properties. Next, the core material is coated with a polymeric film and other ligands to decrease hydrophobicity and to improve the attachment efficiency of conjugated macromolecules. The final product is a biologically active particle that ranges in size from 10 to 15 nm, somewhere in the vicinity of a large protein (133). Quantum dot conjugates are solubilized as a colloidal suspension in common biological buffers and may be incorporated into existing labeling protocols in place of classical staining reagents (such as organic fluorochrome-labeled secondary antibodies).

In confocal microscopy, quantum dots are excited with varying degrees of efficiency by most of the spectral lines produced by the common laser systems, including the argon-ion, helium–cadmium, krypton–argon, and the green helium–neon. Particularly effective at exciting quantum dots in the UV and violet regions are the new blue diode and diode-pumped solid-state lasers that have prominent spectral lines at 442 nm and below (135,136). The 405 nm blue diode laser is an economical excitation source that is very effective for use with quantum dots due to their high extinction coefficient at this wavelength. Another advantage of using these fluorophores in confocal microscopy is the ability to stimulate multiple quantum dot sizes (and spectral colors) in the same specimen with a single excitation wavelength, making these probes excellent candidates for multiple labeling experiments (137).

The exceptional photostability of quantum dot conjugates is of great advantage in confocal microscopy when optical sections are being collected. Unlike the case of organic fluorophores, labeled structures situated away from the focal plane do not suffer from excessive photobleaching during repeated raster scanning of the specimen and yield more accurate 3D volume models. In widefield fluorescence microscopy, quantum dot conjugates are available for use with conventional dye-optimized filter combinations that are standard equipment on many microscopes. Excitation can be further enhanced by substituting a shortpass filter for the bandpass filter that accompanies most filter sets, thus optimizing the amount of lamp energy that can be utilized to excite the quantum dots. Several of the custom fluorescence filter manufacturers offer combinations specifically designed to be used with quantum dot conjugates.

FLUORESCENT PROTEINS

Over the past few years, the discovery and development of naturally occurring fluorescent proteins and mutated derivatives have rapidly advanced to center stage in the investigation of a wide spectrum of intracellular processes in living organisms (75,78,80). These biological probes have provided scientists with the ability to visualize, monitor, and track individual molecules with high spatial and temporal resolution in both steady-state and kinetic experiments. A variety of marine organisms have been the source of >100 fluorescent proteins and their analogs, which arm the investigator with a balanced palette of noninvasive biological probes for single, dual, and multispectral fluorescence analysis (75). Among the advantages of fluorescent proteins over the traditional organic and new semiconductor probes described above is their response to a wider variety of biological events and signals. Coupled with the ability to specifically target fluorescent probes in subcellular compartments, the extremely low or absent photodynamic toxicity, and the widespread compatibility with tissues and intact organisms, these biological macromolecules offer an exciting new frontier in live-cell imaging.

The first member of this series to be discovered, GFP, was isolated from the North Atlantic jellyfish, *Aequorea Victoria*, and found to exhibit a high degree of fluorescence without the aid of additional substrates or coenzymes (138–142). In native green fluorescent protein, the fluorescent moiety is a tripeptide derivative of serine, tyrosine, and glycine that requires molecular oxygen for activation, but no additional cofactors or enzymes (143). Subsequent

investigations revealed that the GFP gene could be expressed in other organisms, including mammals, to yield fully functional analogs that display no adverse biological effects (144). In fact, fluorescent proteins can be fused to virtually any protein in living cells using recombinant complementary DNA cloning technology, and the resulting fusion protein gene product expressed in cell lines adapted to standard tissue culture methodology. Lack of a need for cell-specific activation cofactors renders the fluorescent proteins much more useful as generalized probes than other biological macromolecules, such as the phycobiliproteins, which require insertion of accessory pigments in order to produce fluorescence.

Mutagenesis experiments with green fluorescent protein have produced a large number of variants with improved folding and expression characteristics, which have eliminated wild-type dimerization artifacts and fine tuned the absorption and fluorescence properties. One of the earliest variants, known as enhanced green fluorescence protein (EGFP), contains codon substitutions (commonly referred to as the S65T mutation) that alleviates the temperature sensitivity and increases the efficiency of GFP expression in mammalian cells (145). Proteins fused with EGFP can be observed at low light intensities for long time periods with minimal photobleaching. Enhanced green fluorescent protein fusion products are optimally excited by the 488 nm spectral line from argon and krypton–argon ion lasers in confocal microscopy. This provides an excellent biological probe and instrument combination for examining intracellular protein pathways along with the structural dynamics of organelles and the cytoskeleton.

Additional mutation studies have uncovered GFP variants that exhibit a variety of absorption and emission characteristics across the entire visible spectral region, which have enabled researchers to develop probe combinations for simultaneous observation of two or more distinct fluorescent proteins in a single organism (see the spectral profiles in Fig. 16). Early investigations yielded the blue fluorescent protein (BFP) and cyan fluorescent protein (CFP) mutants from simple amino acid substitutions that shifted the absorption and emission spectral profiles of wild-type GFP to lower wavelength regions (146–148). Used in combination with GFP, these derivatives are useful in resonance energy transfer (FRET) experiments and other investigations that rely on multicolor fluorescence imaging (73). Blue fluorescent protein can be efficiently excited with the 354 nm line from a high power argon laser, while the more useful cyan derivative is excited by a number of violet and blue laser lines, including the 405 nm blue diode, the 442 nm helium–cadmium spectral line, and the 457 nm line from the standard argon-ion laser.

Another popular fluorescent protein derivative, the yellow fluorescent protein (YFP), was designed on the basis of the GFP crystalline structural analysis to red-shift the absorption and emission spectra (148). Yellow fluorescent protein is optimally excited by the 514 nm spectral line of the argon-ion laser, and provides more intense emission than enhanced green fluorescent protein, but is more sensitive to low pH and high halogen ion concentrations. The enhanced yellow fluorescent protein derivative (EYFP) is useful with the 514 argon-ion laser line, but can also be excited with relatively high efficiency by the 488 nm line from argon and krypton–argon lasers. Both of these fluorescent protein derivatives have been widely applied to protein–protein FRET investigations in combination with CFP, and in addition, have proven useful in studies involving multiprotein trafficking.

Attempts to shift the absorption and emission spectra of *Aequorea Victoria* fluorescent proteins to wavelengths in the orange and red regions of the spectrum have met with little success. However, fluorescent proteins from other marine species have enabled investigators to extend the available spectral regions to well within the red wavelength range. The DsRed fluorescent protein and its derivatives, originally isolated from the sea anemone *Discosoma striata*, are currently the most popular analogs for fluorescence analysis in the 575–650 nm region (149). Another protein, HcRed from the *Heteractis crispa* purple anemone, is also a promising candidate for investigations in the longer wavelengths of the visible spectrum (150). Newly developed photoactivation fluorescent proteins, including photoactivatable green fluorescent protein (PA-GFP;74), Kaede (76), and kindling fluorescent protein 1 (KFP1; 151), exhibit dramatic improvements over GFP (up to several 1000-fold) in fluorescence intensity when stimulated by violet laser illumination. These probes should prove useful in fluorescence confocal studies involving selective irradiation of specific target regions and the subsequent kinetic analysis of diffusional mobility and compartmental residency time of fusion proteins.

QUENCHING AND PHOTOBLEACHING

The consequences of quenching and photobleaching are suffered in practically all forms of fluorescence microscopy, and result in an effective reduction in the levels of emission (152,153). These artifacts should be of primary

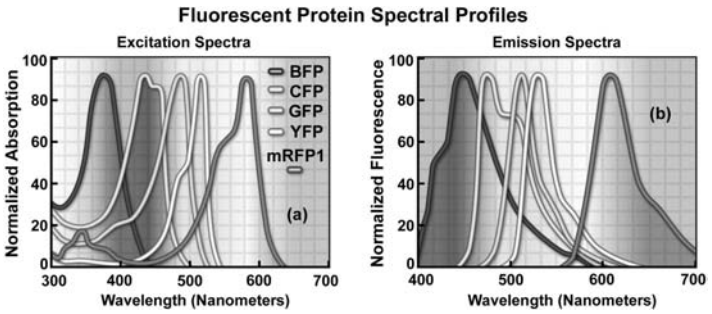


Figure 16. Fluorescent spectral profiles, plotted as normalized absorption or emission as a function of wavelength, for fluorescent proteins emitting in the blue to orange-red regions of the visible spectrum. Each profile is identified with a colored bullet in (a), which illustrates excitation spectra. (b) The emission spectra for the proteins according to the legend in (a).

consideration when designing and executing fluorescence investigations. The two phenomena are distinct in that quenching is often reversible whereas photobleaching is not (154). Quenching arises from a variety of competing processes that induce nonradiative relaxation (without photon emission) of excited-state electrons to the ground state, which may be either intramolecular or intermolecular in nature. Because nonradiative transition pathways compete with the fluorescence relaxation, they usually dramatically lower or, in some cases, completely eliminate emission. Most quenching processes act to reduce the excited state lifetime and the quantum yield of the affected fluorophore.

A common example of quenching is observed with the collision of an excited state fluorophore and another (nonfluorescent) molecule in solution, resulting in deactivation of the fluorophore and return to the ground state. In most cases, neither of the molecules is chemically altered in the collisional quenching process. A wide variety of simple elements and compounds behave as collisional quenching agents, including oxygen, halogens, amines, and many electron-deficient organic molecules (154). Collisional quenching can reveal the presence of localized quencher molecules or moieties, which via diffusion or conformational change, may collide with the fluorophore during the excited state lifetime. The mechanisms for collisional quenching include electron transfer, spin–orbit coupling, and intersystem crossing to the excited triplet state (154,155). Other terms that are often utilized interchangeably with collisional quenching are internal conversion and dynamic quenching.

A second type of quenching mechanism, termed static or complex quenching, arises from nonfluorescent complexes formed between the quencher and fluorophore that serve to limit absorption by reducing the population of active, excitable molecules (154,156). This effect occurs when the fluorescent species forms a reversible complex with the quencher molecule in the ground state, and does not rely on diffusion or molecular collisions. In static quenching, fluorescence emission is reduced without altering the excited state lifetime. A fluorophore in the excited state can also be quenched by a dipolar resonance energy transfer mechanism when in close proximity with an acceptor molecule to which the excited-state energy can be transferred nonradiatively. In some cases, quenching can occur through non molecular mechanisms, such as attenuation of incident light by an absorbing species (including the chromophore itself).

In contrast to quenching, photobleaching (also termed fading) occurs when a fluorophore permanently loses the ability to fluoresce due to photon-induced chemical damage and covalent modification (153–156). Upon transition from an excited singlet state to the excited triplet state, fluorophores may interact with another molecule to produce irreversible covalent modifications. The triplet state is relatively long lived with respect to the singlet state, thus allowing excited molecules a much longer timeframe to undergo chemical reactions with components in the environment (155). The average number of excitation and emission cycles that occur for a particular fluorophore before photobleaching is dependent on the molecular structure and the local environment (154,156). Some fluorophores bleach quickly after emitting only a few photons, while others that are more robust can undergo thousands or even millions of cycles before bleaching.

Figure 17 presents a typical example of photobleaching (fading) observed in a series of digital images captured at different time points for a multiply stained culture of normal Tahr ovary (HJ1.Ov line) fibroblast cells. The nuclei were stained with DAPI (blue fluorescence), while the mitochondria and actin cytoskeleton were stained with MitoTracker Red CMXRos (red fluorescence) and an Alexa Fluor phalloidin derivative (Alexa Fluor 488; green fluorescence), respectively. Time points were taken in 2 min intervals using a fluorescence filter combination with bandwidths tuned to excite the three fluorophores simultaneously while also recording the combined emission signals. Note that all three fluorophores have a relatively high intensity in Fig. 17a, but the DAPI (blue) intensity starts to drop rapidly at two min and is almost completely gone at six min (Fig. 17f). The mitochondrial and actin stains are more resistant to photobleaching, but the intensity of both drops dramatically over the course of the timed sequence (10 min).

An important class of photobleaching events is represented by events that are photodynamic, meaning they involve the interaction of the fluorophore with a combination of light and oxygen (157–161). Reactions between fluorophores and molecular oxygen permanently destroy fluorescence and yield a free-radical singlet oxygen species that can chemically modify other molecules in living cells. The amount of photobleaching due to photodynamic events is a function of the molecular oxygen concentration and the proximal distance between the fluorophore, oxygen molecules, and other cellular components. Photobleaching can be reduced by limiting the exposure time of

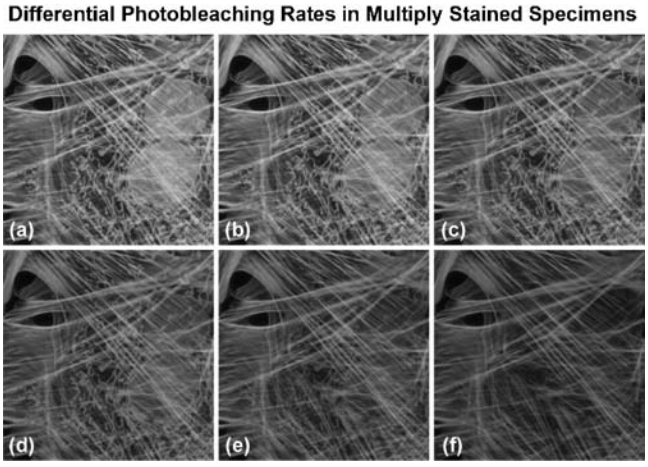


Figure 17. Photobleaching in multiply stained specimens. Normal Tahr ovary fibroblast cells were stained with MitoTracker Red CMXRos (mitochondria; red fluorescence), Alexa Fluor 488 conjugated to phalloidin (actin; green fluorescence), and subsequently counterstained with DAPI (nuclei; blue fluorescence). Time points were taken in two-minute intervals over a 10 min period using a fluorescence filter combination with bandwidths tuned to excite the three fluorophores simultaneously while also recording the combined emission signals. (a–f) Time = 0, 2, 4, 6, 8, 10 min, respectively.

fluorophores to illumination or by lowering the excitation energy. However, these techniques also reduce the measurable fluorescence signal. In many cases, solutions of fluorophores or cell suspensions can be deoxygenated, but this is not feasible for living cells and tissues. Perhaps the best protection against photobleaching is to limit exposure of the fluorochrome to intense illumination (using neutral density filters) coupled with the judicious use of commercially available antifade reagents that can be added to the mounting solution or cell culture medium (153).

Under certain circumstances, the photobleaching effect can also be utilized to obtain specific information that would not otherwise be available. For example, in FRAP experiments, fluorophores within a target region are intentionally bleached with excessive levels of irradiation (82). As new fluorophore molecules diffuse into the bleached region of the specimen (recovery), the fluorescence emission intensity is monitored to determine the lateral diffusion rates of the target fluorophore. In this manner, the translational mobility of fluorescently labeled molecules can be ascertained within a very small (2–5 μm) region of a single cell or section of living tissue.

Although the subset of fluorophores that are advantageous in confocal microscopy is rapidly growing, many of the traditional probes that have been useful for years in widefield applications are still of little utility when constrained by fixed-wavelength laser spectral lines. Many of the limitations surrounding the use of fluorophores excited in the ultraviolet region will be eliminated with the introduction of advanced objectives designed to reduce aberration coupled to the gradual introduction of low cost, high power diode laser systems with spectral lines in these shorter wavelengths. The 405 nm blue diode laser is a rather cheap alternative to more expensive ion and Noble gas based ultraviolet lasers, and is rapidly becoming available for most confocal microscope systems. Helium–neon lasers with spectral lines in the yellow and orange region have rendered some fluorophores useful that were previously limited to widefield applications. In addition, new diode-pumped solid-state lasers are being introduced with emission wavelengths in the UV, violet, and blue regions.

Continued advances in fluorophore design, dual-laser scanning, multispectral imaging, endoscopic instruments, and spinning disk applications will also be important in the coming years. The persistent problem of emission crossover due to spectral overlap, which occurs with many synthetic probes and fluorescent proteins in multicolor investigations, benefits significantly from spectral analysis and deconvolution of lambda stacks. Combined, these advances and will dramatically improve the collection and analysis of data obtained from complex fluorescence experiments in live-cell imaging.

BIBLIOGRAPHY

1. Pawley JB, editor. Handbook of Biological Confocal Microscopy. New York: Plenum Press; 1995.
2. Paddock SW, editor. Confocal Microscopy: Methods and Protocols. Totowa (NJ): Humana Press; 1999.
3. Diaspro A, editor. Confocal and Two-Photon Microscopy: Foundations, Applications, and Advances. New York: Wiley-Liss; 2002.
4. Matsumoto B, editor. Cell Biological Applications of Confocal Microscopy. Methods in Cell Biology, Vol. 70. New York: Academic Press; 2002.
5. Sheppard CJR, Shotton DM. Confocal Laser Scanning Microscopy. Oxford (UK): BIOS Scientific Publishers; 1997.
6. Müller M. Introduction to Confocal Fluorescence Microscopy. Maastricht, The Netherlands: Shaker; 2002.
7. Hibbs AR. Confocal Microscopy for Biologists. New York: Kluwer Academic; 2004.
8. Conn PM. Confocal Microscopy. Methods in Enzymology, Vol. 307. New York: Academic Press; 1999.
9. Corle TR, Kino GS. Confocal Scanning Optical Microscopy and Related Imaging Systems. New York: Academic Press; 1996.
10. Wilson T, editor. Confocal Microscopy. New York: Academic Press; 1990.
11. Gu M. Principles of Three-Dimensional Imaging in Confocal Microscopes. New Jersey: World Scientific; 1996.
12. Masters BR, editor. Selected Papers on Confocal Microscopy. SPIE Milestone Series, Vol. MS 131. Bellingham (WA): SPIE Optical Engineering Press; 1996.
13. Mason WT. Fluorescent and Luminescent Probes for Biological Activity. New York: Academic Press; 1999.
14. Peterman EJG, Sosa H, Moerner WE. Single-Molecule Fluorescence Spectroscopy and Microscopy of Biomolecular Motors. Ann Rev Phys Chem 2004;55:79–96.
15. Goldman RD, Spector DL. Live Cell Imaging: A Laboratory Manual. New York: Cold Spring Harbor Press; 2005.
16. Minsky M. Microscopy Apparatus. US Patent 3,013,467. 1961.
17. Minsky M. Memoir on Inventing the Confocal Scanning Microscopy. Scanning 1988;10:128–138.
18. Egger MD, Petran M. New Reflected-Light Microscope for Viewing Unstained Brain and Ganglion Cells. Science 1967;157:305–307.
19. Davidovits P, Egger MD. Photomicrography of Corneal Endothelial Cells *in vivo*. Nature (London) 1973;244:366–367.
20. Amos WB, White JG. How the Confocal Laser Scanning Microscope entered Biological Research. Biol Cell 2003;95:335–342.
21. Brakenhoff GJ, Blom P, Barends P. Confocal Scanning Light Microscopy with High Aperture Immersion Lenses. J Micros 1979;117:219–232.
22. Sheppard CJR, Wilson T. Effect of Spherical Aberration on the Imaging Properties of Scanning Optical Microscopes. Appl Opt 1979;18:1058.
23. Hamilton DK, Wilson T. Scanning Optical Microscopy by Objective Lens Scanning. J Phys E: Sci Instr 1986;19:52–54.
24. Spring KR, Inoué S. Video Microscopy: The Fundamentals. New York: Plenum Press; 1997.
25. Wilson T. Optical Sectioning in Confocal Fluorescence Microscopes. J Micros 1989;154:143–156.
26. Lichtmann JW. Confocal Microscopy. Sci Am Aug. 1994; 40–45.
27. White JG, Amos WB, Fordham M. An Evaluation of Confocal versus Conventional Imaging of Biological Structures by Fluorescence Light Microscopy. J Cell Biol 1987;105:41–48.
28. Swedlow JR, et al. Measuring Tubulin Content in Toxoplasma gondii: A Comparison of Laser-Scanning Confocal and Wide-Field Fluorescence Microscopy. Proc Natl Acad Sci USA 2002;99:2014–2019.
29. Stelzer EHK. Practical Limits to Resolution in Fluorescence Light Microscopy. In: Yuste R, Lanni F, Konnerth A, editors. Imaging Neurons: A Laboratory Manual. New York: Cold Spring Harbor Press; 2000. pp 12.1–12.9.
30. Rost FWD. Fluorescence Microscopy. Vol. 1. New York: Cambridge University Press; 1992.

31. Murray J. Confocal Microscopy, Deconvolution, and Structured Illumination Methods. In: Goldman RD, Spector DL, editors. Live Cell Imaging: A Laboratory Manual. New York: Cold Spring Harbor Press; 2005. pp 239–280.
32. Dickinson ME, et al. Multi-Spectral Imaging and Linear Unmixing Add a Whole New Dimension to Laser Scanning Fluorescence Microscopy. Biotechniques 2001;31:1272–1278.
33. Zimmermann T, Rietdorf J, Pepperkok R. Spectral Imaging and its Applications in Live Cell Microscopy. FEBS Lett 2003;546:87–92.
34. Lansford R, Bearman G, Fraser SE. Resolution of Multiple Green Fluorescent Protein Variants and Dyes using Two-Photon Microscopy and Imaging Spectroscopy. J Biomed Opt 2001;6:311–318.
35. Hiraoka Y, Shimi T, Haraguchi T. Multispectral Imaging Fluorescence Microscopy for Living Cells. Cell Struct Funct 2002;27:367–374.
36. Gu Y, Di WL, Kelsell DP, Zicha D. Quantitative Fluorescence Energy Transfer (FRET) Measurement with Acceptor Photobleaching and Spectral Unmixing. J Microsc 2004;215:162–173.
37. Ford BK, et al. Computed Tomography-Based Spectral Imaging for Fluorescence Microscopy. Biophys J 2001;80:986–993.
38. Bach H, Renn A, Wild UP. Spectral Imaging of Single Molecules. Single Mol 2000;1:73–77.
39. Wilson T, Carlini AR. Three-Dimensional Imaging in Confocal Imaging Systems with Finite Sized Detectors. J Microsc 1988;149:51–66.
40. Murphy DB. Fundamentals of Light Microscopy and Electronic Imaging. New York: Wiley-Liss; 2001.
41. Wright SJ, Wright DJ. Introduction to Confocal Microscopy. In: Matsumoto B, editor. Cell Biological Applications of Confocal Microscopy. Methods in Cell Biology, Vol. 70. New York: Academic Press; 2002. pp 1–85.
42. Webb RH. Confocal Optical Microscopy. Rep Prog Phys 1996;59:427–471.
43. Wilhelm S, Gröbler B, Gluch M, Hartmut H. Confocal Laser Scanning Microscopy: Optical Image Formation and Electronic Signal Processing. Jena, Germany: Carl Zeiss Advanced Imaging Microscopy; 2003.
44. Ichihara A, et al. High-Speed Confocal Fluorescence Microscopy using a Nipkow Scanner with Microlenses for 3-D Imaging of Fluorescent Molecules in Real-Time. Bioimages 1996;4:57–62.
45. Inoué S, Inoué T. Direct-View High-Speed Confocal Scanner—The CSU-10. In: Matsumoto B, editor. Cell Biological Applications of Confocal Microscopy. Methods in Cell Biology, Vol. 70. New York: Academic Press; 2002. p 88–128.
46. Nakano A. Spinning-Disk Confocal Microscopy —A Cutting-Edge Tool for Imaging of Membrane Traffic. Cell Struct Funct 2002;27:349–355.
47. Chong FK, et al. Optimization of Spinning Disk Confocal Microscopy: Synchronization with the Ultra-Sensitive EMCCD. In: Conchello JA, Cogswell CJ, Wilson T, editors. Three-Dimensional and Multidimensional Microscopy: Image Acquisition and Processing XI. Proc. SPIE. 2004; 5324: 65–76.
48. Sandison D, Webb W. Background Rejection and Signal-to-Noise Optimization in the Confocal and Alternative Fluorescence Microscopes. Appl Op 1994;33:603–610.
49. Centonze VE, White JG. Multiphoton Excitation Provides Optical Sections from Deeper within Scattering Specimens than Confocal Imaging. Biophys J 1998;75:2015.
50. Boccacci P, Bertero M. Image-Restoration Methods: Basics and Algorithms. In: Diaspro A, editor. Confocal and Two-Photon Microscopy: Foundations, Applications, and Advances. New York: Wiley-Liss; 2002. pp 253–269.
51. Verveer PJ, Gemkow MJ, Jovin TM. A Comparison of Image Restoration Approaches Applied to Three-Dimensional Confocal and Wide-Field Fluorescence Microscopy. J Micros 1998;193:50–61.
52. Conchello JA, Hansen EW. Enhanced 3-D Reconstruction from Confocal Scanning Microscope Images. 1: Deterministic and Maximum Likelihood Reconstructions. Appl Op 1990;29: 3795–3804.
53. Al-Kofahi O, et al. Algorithms for Accurate 3D Registration of Neuronal Images Acquired by Confocal Scanning Laser Microscopy. J Micros 2003;211:8–18.
54. Conchello JA, et al., editors. Three-Dimensional and Multidimensional Microscopy: Image Acquisition and Processing. Vol. I–XII, Bellingham (WA): SPIE International Society for Optical Engineering; 1994–2005.
55. Centonze V, Pawley J. Tutorial on Practical Confocal Microscopy and use of the Confocal Test Specimen. In: Pawley JB, editor. Handbook of Biological Confocal Microscopy. New York: Plenum Press; 1995. p 549–570.
56. Gratton E, vandeVen MJ. Laser Sources for Confocal Microscopy. In: Pawley JB, editor. Handbook of Biological Confocal Microscopy. New York: Plenum Press; 1995. p 69–98.
57. Ashkin A, Dziedzic JM, Yamane T. Optical Trapping and Manipulation of Single Cells using Infrared Laser Beams. London 1987;330:769–771.
58. DeMaggio S. Running and Setting Up a Confocal Microscope Core Facility. In: Matsumoto B, editor. Cell Biological Applications of Confocal Microscopy. Methods in Cell Biology, Vol. 70. New York: Academic Press; 2002. p 475–486.
59. Spring KR. Detectors for Fluorescence Microscopy. In: Periasamy A, editor. Methods in Cellular Imaging. New York: Oxford University Press; 2001. p 40–52.
60. Art J. Photon Detectors for Confocal Microscopy. In: Pawley JB, editor. Handbook of Biological Confocal Microscopy. New York: Plenum Press; 1995. p 183–196.
61. Amos WB. Instruments for Fluorescence Imaging. In: Allan VJ, editor. Protein Localization by Fluorescence Microscopy: A Practical Approach. New York: Oxford University Press; 2000. pp 67–108.
62. Hergert E. Detectors: Guideposts on the Road to Selection. Photonics Design and Applications Handbook. 2001. pp H110–H113.
63. Piston DW, Patterson GH, Knobel SM. Quantitative Imaging of the Green Fluorescent Protein (GFP). In: Sullivan KF, Kay SA, editors. Green Fluorescent Proteins, Methods in Cell Biology, Vol. 58. New York: Academic Press; 1999. pp 31–47.
64. Carter DR. Photomultiplier Handbook: Theory, Design, Application. Lancaster (PA): Burle Industries, Inc.; 1980.
65. Cody SH, et al. A Simple Method Allowing DIC Imaging in Conjunction with Confocal Microscopy. J Microsc 2005;217: 265–274.
66. Chang IC. Acousto-Optic Devices and Applications. In: Bass M, Van Stryland EW, Williams DR, Wolfe WL, editors. Optics II: Fundamentals, Techniques, and Design. New York: McGraw-Hill; 1995. pp 12.1–12.54.
67. Wachman ES. Acousto-Optic Tunable Filters for Microscopy. In: Yuste R, Lanni F, Konnerth A, editors. Imaging Neurons: A Laboratory Manual. New York: Cold Spring Harbor Press; 2000. p 4.1–4.8.
68. Shonat RD, et al. Near-Simultaneous Hemoglobin Saturation and Oxygen Tension Maps in Mouse Brain using an AOTF Microscope. Biophy J 1997;73:1223–1231.
69. Wachman ES, Niu W, Farkas DL. Imaging Acousto-Optic Tunable Filter with 0.35-Micrometer Spatial Resolution. App Opt 1996;35:5220–5226.
70. Wachman ES. AOTF Microscope for Imaging with Increased Speed and Spectral Versatility. Biophys J 1997;73:1215–1222.

71. Chen Y, Mills JD, Periasamy A. Protein Localization in Living Cells and tissues using FRET and FLIM. Differentiation 2003;71:528–541.
72. Wallrabe H, Periasamy A. Imaging Protein Molecules using FRET and FLIM Microscopy. Curr Opin Biotech 2005;16: 19–27.
73. Day RN, Periasamy A, Schaufele F. Fluorescence Resonance Energy Transfer Microscopy of Localized Protein Interactions in the Living Cell Nucleus. Methods 2001;25:4–18.
74. Patterson GH, Lippincott-Schwartz J. A Photoactivatable GFP for Selective Photolabeling of Proteins and Cells. Science 2002;297:1873–1877.
75. Verkhusha VV, Lukyanov KA. The Molecular Properties and Applications of Anthozoa Fluorescent Proteins and Chromoproteins. Nature Biotechnol 2004;22:289–296.
76. Ando R, et al. An Optical Marker Based on the UV-Induced Green-to-Red Photoconversion of a Fluorescent Protein. Proc Natl Acad Sci USA 2002;99:12651–12656.
77. Sharma D. The Use of an AOTF to Achieve High Quality Simultaneous Multiple Label Imaging. Bio-Rad Technical Notes. San Francisco: Bio-Rad, Note 4; 2001.
78. Miyawaki A, Sawano A, Kogure T. Lighting up Cells: Labeling Proteins with Fluorophores. Nature Cell Biol 2003;5:S1–S7.
79. Zhang J, Campbell RE, Ting AY, Tsien RY. Creating New Fluorescent Probes for Cell Biology. Nature Rev Mol Cell Bio 2002;3:906–918.
80. Lippincott-Schwartz J, Patterson G. Development and Use of Fluorescent Protein Markers in Living Cells. Science 2003;300:87–91.
81. Klonis N, et al. Fluorescence Photobleaching Analysis for the Study of Cellular Dynamics. Eur Biophys J 2002;31:36–51.
82. Lippincott-Schwartz J, Altan-Bonnet N, Patterson GH. Photobleaching and Photoactivation: Following Protein Dynamics in Living Cells. Nature Cell Biol 2003;5:S7–S14.
83. Phair RD, Misteli T. Kinetic Modelling Approaches to *in vivo* Imaging. Nature Rev Mol Cell Bio 2002;2:898–907.
84. Politz JC. Use of Caged Fluorophores to Track Macromolecular Movement in Living Cells. Trends Cell Biol 1999;9:284–287.
85. Born M, Wolf E. Principles of Optics. New York: Cambridge University Press; 1999.
86. Stelzer EHK. Contrast, Resolution, Pixelation, Dynamic range, and Signal-to-Noise Ratio: Fundamental Limits to Resolution in Fluorescence Light Microscopy. J Microsc 1997;189:15–24.
87. Pawley J. Fundamental Limits in Confocal Microscopy. In: Pawley JB, editor. Handbook of Biological Confocal Microscopy. New York: Plenum Press; 1995. p 19–37.
88. Webb RH, Dorey CK. The Pixelated Image. In: Pawley JB, editor. Handbook of Biological Confocal Microscopy. New York: Plenum Press; 1995. p 55–67.
89. Jonkman JEN, Stelzer EHK. Resolution and Contrast in Confocal and Two-Photon Microscopy. In: Diaspro A, editor. Confocal and Two-Photon Microscopy: Foundations, Applications, and Advances. New York: Wiley-Liss; 2002. p 101–125.
90. Haugland RP. The Handbook: A Guide to Fluorescent Probes and Labeling Technologies. Chicago: Invitrogen Molecular Probes; 2005.
91. Lemasters JJ, et al. Confocal Imaging of Ca^{2+}, pH, Electrical Potential and Membrane Permeability in Single Living Cells. Methods Enzymol 1999;302:341–358.
92. Johnson I. Fluorescent Probes for Living Cells. Histochem J 1998;30:123–140.
93. Kasten FH. Introduction to Fluorescent Probes: Properties, History, and Applications. In: Mason WT, editor. Fluorescent and Luminescent Probes for Biological Activity. New York: Academic Press; 1999. p 17–39.
94. Coons AH, Creech HJ, Jones RN, Berliner E. Demonstration of Pneumococcal Antigen in Tissues by use of Fluorescent Antibody. J Immunol 1942;45:159–170.
95. Tsien RY. Building and Breeding Molecules to Spy on Cells and Tumors. FEBS Lett 2005;579:927–932.
96. Bruchez Jr M, et al. Semiconductor Nanocrystals as fluorescent Biological Labels. Science 1998;218:2013–2016.
97. Tsien RY, Waggoner A. Fluorophores for Confocal Microscopy. In: Pawley JB, editor. Handbook of Biological Confocal Microscopy. New York: Plenum Press; 1995. p 267–280.
98. Wessendorf MW, Brelje TC. Which Fluorophore is Brightest? A Comparison of the Staining Obtained Using Fluorescein, Tetramethylrhodamine, Lissamine Rhodamine, Texas Red and Cyanine 3.18. Histochemistry 1992;98:81–85.
99. Entwistle A, Noble M. The use of Lucifer Yellow, BODIPY, FITC, TRITC, RITC and Texas Red for Dual Immunofluorescence Visualized with a Confocal Scanning Laser Microscope. J Microsc 1992;168:219–238.
100. Darzynkiewicz Z. Differential Staining of DNA and RNA in Intact Cells and Isolated Cell Nuclei with Acridine Orange. Methods Cell Biol 1990;33:285–298.
101. Waring MJ. Complex Formation Between Ethidium Bromide and Nucleic Acids. J Mol Biol 1965;13:269–282.
102. Arndt-Jovin DJ, Jovin TM. Fluorescence Labeling and Microscopy of DNA. Fluores Microsc Living Cells Culture Part B Methods Cell Biol 1989;30:417–448.
103. Kubista M, Akerman B, Norden B. Characterization of Interaction between DNA and 4′,6-Diamidino-2-phenylindole by Optical Spectroscopy. Biochemistry 1987;26:4545–4553.
104. Loewe H, Urbanietz J. Basic Substituted 2,6-Bisbenzimidazole Derivatives: A Novel Series of Substances with Chemotherapeutic Activity. Arzneim-Forsch 1974;24:1927–1933.
105. Arndt-Jovin DJ, Jovin TM. Analysis and Sorting of Living Cells According to Deoxyribonucleic Acid Content. J Histochem Cytochem 1977;25:585–589.
106. Durand RE, Olive PL. Cytotoxicity, Mutagenicity and DNA Damage by Hoechst 33342. J Histochem Cytochem 1982; 30:111–116.
107. Panchuk-Voloshina N, et al. Alexa Dyes, A Series of New Fluorescent Dyes that Yield Exceptionally bright, Photostable Conjugates. J Histochem Cytochem 1999;47:1179–1188.
108. Berlier JE, et al. Quantitative Comparison of Long-Wavelength Alexa Fluor Dyes to Cy Dyes: Fluorescence of the Dyes and their Conjugates. J Histochem Cytochem 2003;51:1699–1712.
109. Mujumdar RB, et al. Cyanine Dye Labeling Reagents: Sulfoindocyanine Succinimidyl Esters. Bioconjugate Chem 1993;4:105–111.
110. Ballou B, et al. Tumor Labeling *in vivo* using Cyanine-Conjugated Monoclonal Antibodies. Cancer Immunol Immunother 1995;41:257–263.
111. Zorov DB, Kobrinsky E, Juhaszova M, Sollott SJ. Examining Intracellular Organelle Function Using Fluorescent Probes. Circul Res 2004;95:239–252.
112. Stephens DG, Pepperkok R. The Many Ways to Cross the Plasma Membrane. Proc Natl Acad Sci USA 2001;98:4295–4298.
113. Rudolf R, Mongillo M, Rizzuto R, Pozzan T. Looking Forward to Seeing Calcium. Nature Rev Mol Cell Biol 2003;4:579–586.
114. Martin H, Bell MG, Ellis-Davies GC, Barsotti RJ. Activation Kinetics of Skinned Cardiac Muscle by Laser Photolysis of Nitrophenyl-EGTA. Biophys J 2004;86:978–990.
115. White C, McGeown G. Imaging of Changes in Sarcoplasmic Reticulum $[Ca^{2+}]$ using Oregon Green BAPTA 5N and Confocal Laser Scanning Microscopy. Cell Calcium 2002;31:151–159.
116. Helm PJ, Patwardhan A, Manders EM. A Study of the Precision of Confocal, Ratiometric, Fura-2-Based $[Ca^{2+}]$ Measurements. Cell Calcium 1997;22:287–298.
117. Rijkers GT, Justement LB, Griffioen AW, Cambier JC. Improved Method for Measuring Intracellular Ca^{++} with Fluo-3. Cytometry 1990;11:923–927.

118. Schild D, Jung A, Schultens HA. Localization of Calcium Entry through Calcium Channels in Olfactory Receptor Neurons using a Laser Scanning Microscope and the Calcium Indicator Dyes Fluo-3 and Fura-Red. Cell Calcium 1994;15: 341–348.
119. Willoughby D, Thomas RC, Schwiening CJ. Comparison of Simultaneous pH Measurements made with 8-Hydroxypyrene-1,3,6-trisulphonic acid (HPTS) and pH-Sensitive Microelectrodes in Snail Neurons. Pflugers Arch 1998;436:615–622.
120. Ozkan P, Mutharasan R. A Rapid Method for Measuring Intracellular pH using BCECF-AM. Biochim Biophys Acta 2002;1572:143.
121. Cody SH, et al. Intracellular pH Mapping with SNARF-1 and Confocal Microscopy. I: A Quantitative Technique for Living Tissue and Isolated Cells. Micron 1993;24:573–580.
122. Dubbin PN, Cody SH, Williams DA. Intracellular pH Mapping with SNARF-1 and Confocal Microscopy. II: pH Gradients within Single Cultured Cells. Micron 1993;24:581–586.
123. Poot M, et al. Analysis of Mitochondrial Morphology and Function with Novel Fixable Fluorescent Stains. J Histochem Cytochem 1996;44:1363–1372.
124. Keij JF, Bell-Prince C, Steinkamp JA. Staining of Mitochondrial Membranes with 10-Nonyl Acridine Orange, MitoFluor Green, and MitoTracker Green is Affected by Mitochondrial Membrane Potential Altering Drugs. Cytometry 2000;39: 203–210.
125. Reers M. et al. Mitochondrial Membrane Potential Monitored by JC-1 Dye. Methods Enzymol 1995;260:406–417.
126. Price OT, Lau C, Zucker RM. Quantitative Fluorescence of 5-FU-Treated Fetal Rat Limbs using Confocal Laser Scanning Microscopy and LysoTracker Red. Cytometry 2003;53A:9–21.
127. Kumar RK, Chapple CC, Hunter N. Improved Double Immunofluorescence for Confocal Laser Scanning Microscopy. J Histochem Cytochem 1999;47:1213–1217.
128. Suzuki T, Fujikura K, Higashiyama T, Takata K. DNA Staining for Fluorescence and Laser Confocal Microscopy. J Histochem Cytochem 1997;45:49–53.
129. Haugland RP. Coupling of Monoclonal Antibodies with Fluorophores. In: Davis WC, editor. Monoclonal Anibody Protocols, Methods in Molecular Biology. Vol. 45. Totowa, (NJ): Humana Press; 1995. p 205–221.
130. Pagano RE, Martin OC. Use of Fluorescent Analogs of Ceramide to Study the Golgi Apparatus of Animal Cells. In: Celis JE, editor. Cell Biology: A Laboratory Handbook. Vol. 2, 1998. p 507–512.
131. Cole L, Davies D, Hyde GJ, Ashford AE. ER-Tracker Dye and BODIPY-Brefeldin A Differentiate the Endoplasmic Reticulum and Golgi Bodies from the Tubular-Vacuole System in Living Hyphae of Pisolithus tinctorius. J Microsc 2000;197:239–249.
132. Jaiswal JK, Mattoussi H, Mauro JM, Simon SM. Long-Term Multiple Color Imaging of Live Cells using Quantum Dot Bioconjugates. Nature Biotechnol 2003;21:47–52.
133. Larson DR, et al. Water Soluble Quantum Dots for Multiphoton Fluorescence Imaging *in vivo*. Science 2003;300: 1434–1436.
134. Watson A, Wu X, Bruchez M. Lighting up Cells with Quantum Dots. Biotechniques 2003;34:296–303.
135. Michalet X, et al. Quantum Dots for Live Cells, in vivo Imaging, and Diagnostics. Science 2005;307:538–544.
136. Gao X, et al. *In vivo* Molecular and Cellular Imaging with Quantum Dots. Curr Opin Biotech 2005;16:63–72.
137. Lacoste TD, et al. Ultrahigh-Resolution Multicolor Colocalization of Single Fluorescent Probes. Proc Natl Acad Sci USA 2000;97:9461–9466.
138. Tsien RY. The Green Fluorescent Protein. Ann Rev Biochem 1998;67:509–544.
139. Sullivan KF, Kay SA, editors. Green Fluorescent Proteins, Methods in Cell Biology. Vol. 58. New York: Academic Press; 1999.
140. Conn PM, editor. Green Fluorescent Protein, Methods in Enzymology. Vol. 302. New York: Academic Press; 1999.
141. Hicks BW, editor. Green Fluorescent Protein, Methods in Molecular Biology. Vol. 183. Totowa (NJ): Humana Press; 2002.
142. Chalfie M, Kain S, editors. Green Fluorescent Protein: Properties, Applications, and Protocols. New York: Wiley-Liss; 1998.
143. Zimmer M. Green Fluorescent Protein: Applications, Structure, and Related Photophysical Behavior. Chem Rev 2002;102:759–781.
144. Chalfie M, et al. Green Fluorescent Protein as a Marker for Gene Expression. Science 1994;263:802–805.
145. Heim R, Cubitt AB, Tsien RY. Improved Green Fluorescence. Nature (London) 1995;373:664–665.
146. Heim R, Prasher DC, Tsien RY. Wavelength Mutations and Posttranslational Autoxidation of Green Fluorescent Protein. Proc Natl Acad Sci USA 1994;91:12501–12504.
147. Heim R, Tsien RY. Engineering Green Fluorescent Protein for Improved Brightness, Longer Wavelengths, and Fluorescence Resonance Energy Transfer. Curr Biol 1996;6: 178–182.
148. Wachter RM, et al. Structural Basis of Spectral Shifts in the Yellow-Emission Variants of Green Fluorescent Protein. Structure 1998;6:1267–1277.
149. Matz MV, et al. Fluorescent Proteins from Nonbioluminescent Anthozoa Species. Nature Biotechnol 1999;17:969–973.
150. Matz MV, Lukyanov KA, Lukyanov SA. Family of the Green Fluorescent Protein: Journey to the End of the Rainbow. BioEssays 2002;24:953–959.
151. Natural Animal Coloration can be Determined by a Nonfluorescent Green Fluorescent Protein Homolog. J Biol Chem 2000;275:25879–25882.
152. Song L, Hennink EJ, Young IT, Tanke HJ. Photobleaching Kinetics of Fluorescein in Quantitative Fluorescence Microscopy. Biophys J 1995;68:2588–2600.
153. Berrios M, Conlon KA, Colflesh DE. Antifading Agents for Confocal Fluorescence Microscopy. Methods Enzymol 1999; 307:55–79.
154. Lakowicz JR. Principles of Fluorescence Spectroscopy. New York: Kluwer Academic/Plenum Publishers; 1999.
155. Song L, Varma CA, Verhoeven JW, Tanke HJ. Influence of the Triplet Excited State on the Photobleaching Kinetics of Fluorescein in Microscopy. Biophys J 1996;70:2959–2968.
156. Herman B. Fluorescence Microscopy. New York: BIOS Scientific Publishers; 1998.
157. Bunting JR. A Test of the Singlet Oxygen Mechanism of Cationic Dye Photosensitization of Mitochondrial Damage. Photochem Photobiol 1992;55:81–87.
158. Byers GW, Gross S, Henrichs PM. Direct and Sensitized Photooxidation of Cyanine Dyes. Photochem Photobiol 1976; 23:37–43.
159. Dittrich PS, Schwille P. Photobleaching and Stabilization of Fluorophores used for Single Molecule Analysis with One- and Two-Photon Excitation. Appl Phys B 2001;73:829–837.
160. Gandin E, Lion Y, Van de Vorst A. Quantum Yield of Singlet Oxygen Production by Xanthene Derivatives. Photochem Photobiol 1983;37:271–278.
161. Kanofsky JR, Sima PD. Structural and Environmental Requirements for Quenching of Singlet Oxygen by Cyanine Dyes. Photochem Photobiol 2000;71:361–368.

Further Reading

Willison JR. Signal Detection and Analysis, In: Bass, M. Van Stryland, E. W. Williams DR, Wolfe WL. Optics I: Fundamentals, Techniques, and Design. New York: McGraw-Hill; pp 1995. 18.1–18.16.

See also CELLULAR IMAGING; FLUORESCENCE MEASUREMENTS; ION-SENSITIVE FIELD EFFECT TRANSISTORS.

MICROSCOPY, ELECTRON

MAHROKH DADSETAN
LICHUN LU
MICHAEL J. YASZEMSKI
Mayo Clinic,
College of Medicine
Rochester, Minnesota

INTRODUCTION

Invention of the light microscope by Janssens in 1590 was the first milestone in the microscopic world. Janssens' microscope magnified objects up to 20–30 times their original size. By the beginning of the twentieth century, objects could be magnified only up to 1000 times with a resolution of 0.2 μm. In the early 1930s, the limitations of light microscopes and the scientific desire to see intracellular structural details, such as mitochondria and nuclei led to the development of electron microscopes. The electron microscope took advantage of the much shorter wavelength of the electron compared to that of visible light. With the electron microscope, another 1000-fold increase in magnification was accomplished with a concomitant increase in resolution, allowing visualization of viruses, deoxyribonuclic acid (DNA), and smaller objects, such as molecules and atoms. The transmission electron microscope (TEM) was the first type of electron microscope, and was developed by Ruska and Knoll in Germany in 1931. Electron microscopy is based on a fundamental physics concept stated in the de Broglie theory (1924). This concept is that moving electrons have the properties of waves. The second major advancement in electron microscopy was made by Busch, who demonstrated in 1926 that electrostatic or magnetic fields could be used as a lens to focus an electron beam. In 1939, Siemens Corp. began commercial production of a microscope developed by Von Borries and Ruska in Germany. Hiller, Vance and others constructed the first TEM in North America in 1941. This instrument had a resolution of 2.5 nm.

About the same time that the first TEM was nearing completion in the 1930s, a prototype of the scanning electron microscope (SEM) was constructed by Knoll and Von Ardenne in Germany. However, the resolution of this microscope was no better than that of the light microscope. Following several improvements made by RCA in the United States, as well as Cambridge University in England, a commercial SEM became available in 1963. A later version of the SEM made by the Cambridge Instrument Co. had a resolving power of ~20–50 nm and a useful magnification of 75,000×. Recent models of the SEM have a resolving power of 3.0 nm and magnifications up to 300,000×.

Although the design of TEM and SEM is similar in many ways, their applications are very different. The TEM is patterned after as light microscope, except that electrons instead of light pass through the object. The electrons are then focused by two or more electron lenses to form a greatly magnified image onto photographic film or a charge coupled device (CCD) camera. The image produced by TEM is two-dimensional (2D) and the brightness of a particular region of the image is proportional to the number of electrons that are transmitted through the specimen at that position on the image. The SEM produces a three-dimensional (3D) image by scanning the surface of a specimen with a 2–3 nm spot of electrons to generate secondary electrons from the specimen that are then detected by a sensor. The resolution of an SEM is limited by two quite different sets of circumstances. One of these is concerned with the physics of electron optics, while the other depends on the penetration of electrons into the object being imaged.

A third type of electron microscope, the scanning transmission electron microscope (STEM) has features of both the transmission and scanning electron microscopes. This microscope is an analytical tool that determines the presence and distribution of the atomic elements in the specimen. Recently, two groups of researchers have accomplished a subangstrom resolution (0.06 nm) for STEM using an aberration corrector. They have reported that columns of atoms in a silicone crystal that are 0.078 nm apart can be distinguished at this resolution (1). The image of Si shown in Fig. 1 has been recorded in a high angle annular dark field (HAADF) mode, and the pairs of atomic columns are seen directly resolved. The HAADF detector collects electrons scattered by the sample to angles greater than the detector inner radius. Such high angle scattering is largely incoherent thermal diffuse scattering, which means that the resolution observed in the image is determined by the intensity distribution of the illuminating probe. With this advantage over conventional coherent high resolution transmission electron microscopy (HRTEM), HAADF–STEM has enabled imaging not only of individual atomic columns in crystals, but single dopant atoms on their surface and within their interior.

In 1982, another type of electron microscope, the scanning tunneling microscope (STM) was developed by two scientists, Rohrer and Binnig, for studying surface structure. This invention was quickly followed by the development of a family of related techniques classified as scanning probe microscopy (SPM). These techniques are based upon moving a probe (typically called a tip in STM, which is literally a sharp metallic object) just above a specimen's surface while monitoring some interaction

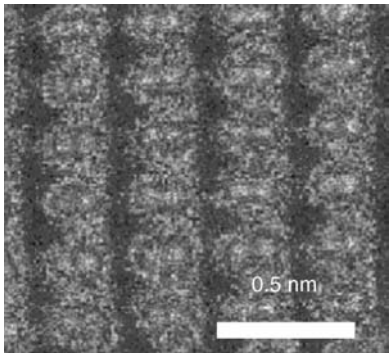


Figure 1. Image of a silicone crystal observed in the [112] orientation recorded with an aberration corrected STEM. (From Ref. 1). Reproduced by courtesy of American Association for the Advancement of Science.)

between the probe and the surface. Atomic force microscopy (AFM) is another important technique of this kind but is not categorized as electron microscopy. Bennig and Rohrer were awarded one-half of the 1986 Nobel Prize in physics for invention of the STM, while Ernst Ruska was awarded the other one-half of that same Nobel Prize for his 1931 invention of the electron microscope. In STM, electrons are speeded up in a vacuum until their wavelength is extremely short, only one hundred-thousandth that of white light. Beams of these fast-moving electrons are focused on a cell sample and are absorbed or scattered by the cell's parts so as to form an image on an electron-sensitive photographic plate. The STM is widely used in both industrial and academic research to obtain atomic scale images of metal surfaces. It provides a 3D profile of the surface, which is useful in characterizing surface roughness and determining the size and conformation of surface molecules. (2) Invention of electron microscopy had an enormous impact in the field of biology, specifically in cell and tissue analysis. Almost 15 years after the invention of the first electron microscope by Ruska, many efforts were made to apply this technique to biological problems. Using the electron microscope, cell organelles and cell inclusions were discovered or resolved in finer details. Electron microscopy, specifically TEM, is now among the most important tools in cell biology and diagnostic pathology.

The latest advancement in electron microscopy is 3D reconstruction of cellular components at a resolution that is on the order of magnitude of atomic structures defined by X-ray crystallography. The method for reconstruction of 3D images of single, transparent objects recorded by TEM is called electron tomography (ET). In order to generate 3D images of individual molecules, one needs to obtain as many tilt images as possible, covering the widest possible angular range. The representative images of particles obtained from different orientations is then analyzed and combined by a software program to reconstruct the molecule in 3D. With improvements in instrumentation, data collection methods and techniques for computation, ET may become a preferred method for imaging isolated organelles and small cells. So far, the electron tomography method covers the resolution range of 2.5–5.0 nm. Data obtained via electron tomography furnish a rich source of quantitative information about the structural composition and organization of cellular components. It offers the opportunity to obtain 3D information on structural cellular arrangements with a significantly higher resolution than that provided by any other method currently available (e.g., confocal laser microscopy) (3).

THEORY OF ELECTRON MICROSCOPY

According to electromagnetic theory, a light source initiates a vibrational motion that transmits energy in the direction of propagation. The wave motion of light is analogous to that produced by a stone thrown into a pool of water. When the waves generated from throwing a stone strike an object that has an opening or aperture, another series of waves is generated from the edge of the object. The result is a new source of waves that emerges with the

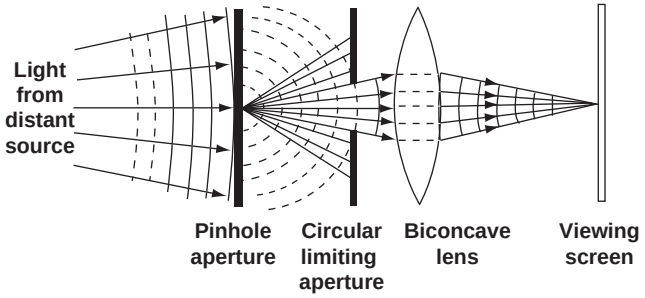


Figure 2. Diffraction of light waves.

original waves. This bending or spreading phenomenon is known as diffraction. Diffracted waves interfere with the initial waves, and the result is an image of the edge of the object. The edge appears to have a series of bands or fringes called Fresnel fringes running parallel to the edge (Fig. 2). Thus, if a strong beam of light illuminates a pinhole in a screen and thus a pinhole serves as a point source and the light passing through is focused by an apertured "perfect" lens on a second screen, the image obtained is not a pinpoint of light, but rather a bright central disk surrounded by a diffuse ring of light. Even if monochromatic light was used to illuminate the point source and was to pass through a perfect lens, the image will not be a sharp one, but rather a diffuse disk composed of concentric rings. This type of image is known as an Airy disk after Sir George Airy, who first described this pattern during the nineteenth century (Fig. 3) (4). To determine resolving power (RP), it is important to know the radius of the Airy disk. The radius of the Airy disk as measured to the first dark ring (r) is expressed by following equation:

$$r = \frac{0.612\,\lambda}{n(\sin\alpha)} \tag{1}$$

In Eq. 1, λ = wavelength of illumination; n = refractive index of the medium between the point source and the lens, relative to free space; α = half the angle of the cone of light from the specimen plane accepted by the front lens of objective

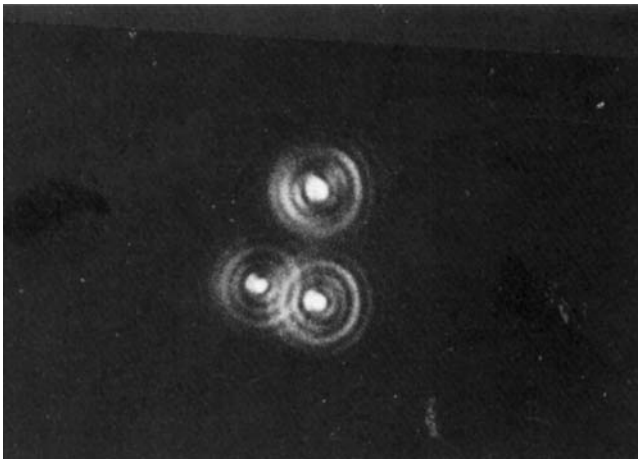


Figure 3. Airy disks generated by viewing three pinholes in a light microscope. Magnification of micrograph is 1000×. (From Ref. 4. Reproduced by courtesy of Jones and Bartlett Publishers.)

The above equation can be shown in another form as:

$$d = \frac{0.612\,\lambda}{\mathrm{NA}} \qquad (2)$$

where NA (numerical aperture) $= n$ sin α and represents the light gathering power of the lens aperture.

From the above equation, RP is defined as the minimum distance that two objects can be placed apart and still be seen as separate entities. Consequently, the shorter the distance, the better (or higher) is the RP of the system. For example, consider a light microscope using ultraviolet (UV) light, which lies beyond the lower end of the visible spectrum (400 nm). Further specifications of this system include a glass slide with standard immersion oil (refractive index, $n = 1.5$), and sinα= 0.87 (sine of a 64° angle, representing one-half of the 128° acceptance angle of a glass lens. The theoretical resolution that can be attained by this system is ~ 0.2 μm. In other words, two points in the specimen that are not separated by at least this distance will not be seen as two distinct points, but will be observed as a single blurred image. Since the values of sin α and n cannot be significantly increased beyond the stated values, the RP can most effectively be improved by reducing wavelength.

Electron Beams And Resolution

The concept that moving electrons might be used as an illumination source was suggested by a tenet of the de Broglie theory, that moving electrons have wave properties. The wavelength of this particle-associated radiation is given by following equation:

$$\lambda\,\frac{h}{mv} \qquad (3)$$

where m is the mass of the particle, v the velocity of particle, and h is Planck's constant ($6.626 \times 10^{-34}\mathrm{J}^{-1}$). For an electron accelerated by a potential of 60,000 V (60 kV), the wavelength of the electron beam would be ~0.005 nm, which is 100,000 times shorter than that for green light. By using Eq. 1, a TEM with perfect lenses would therefore in theory be able to provide a resolution of 0.0025 nm. In practice, the actual resolution of a modern high resolution transmission electron microscope is closer to 0.2 nm. The reason we are not able to achieve the nearly 100-fold better resolution of 0.002 nm is due to extremely narrow aperture angles (~1000 times smaller than that of the light microscope) needed by the electron microscope lenses to overcome a major resolution limiting phenomenon called spherical aberration. In addition, diffraction, chromatic aberration and astigmatism all contribute to decreased resolution in TEM, and need to be corrected to achieve higher resolution. (5)

Magnification

The maximum magnification of any microscope is simply the ratio of the microscope's resolution to the resolution of the unaided human eye. The resolution of the eye viewing an object at 25 cm is generally taken to be 0.25 mm. Since the resolution of a light microscope is ~0.25 μm, maximum useful light magnification is ~1000×, obtainable from an objective lens of 100× followed by an eyepiece of 10×. The magnification of TEM would be ~0.25mm/0.25 nm. This is a $10^6\times$ magnification, and corresponds to a 1000-fold increase in resolution compared to a light microscope. An objective lens of 100× is followed by an "intermediate" lens of 25×, and the final image is projected by a projector lens of 100×. Further magnification for critical focusing is obtained by viewing the image on the fluorescent screen with a long working distance binocular microscope of 10×. The final image is photographed at 250,000×. The processed negative is then enlarged a further 4× in a photographic enlarger. This result in a final prints (the electron micrograph) at the desired magnification of $10^6\times 6$.

Electromagnetic Lenses

An electromagnetic lens is generated by a coil of wire with a direct current (dc) that passes through the coil. This electromagnetic coil is called a solenoid. It forms an axially and radially symmetric magnetic field that converges to a point. A divergent cone of electrons enters from a point source, and thus forms a real image on the lens axis. An advantage of electromagnetic lenses is that the focal length can be made infinitely variable by varying the coil current. Therefore, both magnification and image focus can be adjusted by controlling the lens current (Fig. 4) (7).

Lens Aberrations

Electron lenses are affected by all the aberrations of optical lenses, such as spherical aberration, chromatic aberration, astigmatism, and distortion. Spherical aberration results from the geometry of both glass and electromagnetic lenses such that rays passing through the periphery of the lens are refracted more than rays passing along the axis. Spherical aberration may be reduced by using an aperture to eliminate some of the peripheral rays. Although this aperture is attractive for reducing spherical aberration, it decreases the aperture angle and thereby prevents the electron microscope from achieving the theoretical resolution predicted by Eq. 1.

Chromatic aberration results when electromagnetic radiations of different energies converge at different focal

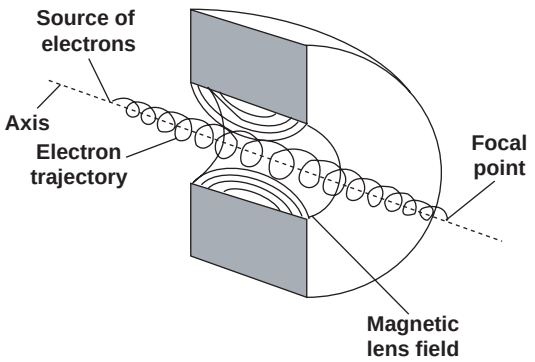


Figure 4. Single electron passing through electromagnetic lens. The electron is focused by the magnetic field to follow a trajectory that will converge at a defined focal point after it emerges from the lens.

planes. Chromatic aberration results in the enlargement of a focal point with a consequential loss of resolution. It can be corrected by using a monochromatic source of electromagnetic radiation. This entails stabilizing the accelerating voltage to generate the electrons with same levels of energy, and having a good vacuum to minimize the energy loss of the electrons during their passage through the transmission specimen. This effect can also be reduced by decreasing the aperture of the objective lens. (6)

Astigmatism is caused by radial asymmetry in a lens, giving rise to a focal length in one plane that is different from that in another plane. The fabrication and maintenance of a lens that has a perfectly symmetric lens field is not feasible in practice. Thus, it is necessary to correct astigmatism by applying a radial symmetry compensator device called a stigmator. This consists of an adjustable electric or magnetic field that can be applied across the lens in any chosen direction, thus compensating for astigmatism. Image distortion, due to magnification changing across the field from the value at the center, may be positive (called barrel distortion) or negative (called pincushion distortion). These effects can be compensated by operating two lenses in series, arranging for barrel distortion in one to be compensated by pincushion distortion in the other. The lens system in modern electron microscopes is designed to automatically counterbalance the various types of distortions throughout a wide magnification range. (4)

DESIGN OF THE TRANSMISSION ELECTRECTRON MICROSCOPE

Both the light and electron microscopes are similar so far as the arrangement and function of their components are concerned. Thus, both microscopes, when used for photographic purposes, can be conveniently divided into the following component systems.

Illuminating System

This system serves to produce the required radiation and to direct it onto the specimen. It consists of source and condenser lenses.

Source Lens. The source of electrons, or cathode, is a hairpin of fine tungsten wire about 2 mm long, maintained at ~2500 K by ~2 W of alternting current (ac) or dc power. Electrons boil off the white-hot tungsten surface, and are shaped into a conical beam by an electrode system called the gun. Two further electrodes, the shield and the anode, combine to form an electrostatic collimating lens and accelerator. A suitable accelerating voltage (20–100 kV) is chosen for the specimen under examination, and is applied to the cathode as a negative potential so that the anode may remain at earth potential. The cathode and shield are therefore carried on an insulator. The filament-shield voltage (cathode bias) is made variable to adjust the total current drawn from the filament, which in turn varies the brightness of the final image (4).

The energy of the electrons in the TEM determines the relative degree of penetration of electrons into a specific sample, or alternatively, influences the thickness of material from which useful information may be obtained. Thus, a high energy TEM (400 kV) not only provides the highest resolution but also allows for the observation of relatively thick samples (e.g., ~0.2 μm) when compared with the more conventional 100 kV or 200 kV instruments. Because of the high spatial resolution obtained, TEMs are often employed to determine the detailed crystallography of fine-grained, or rare, materials (6,8).

Condenser Lens. The condenser lens regulates the convergence (and thus the intensity) of the illuminating beam on the specimen. The divergent electron beam emerging from the anode aperture can, in simple instruments, be used to illuminate the specimen directly. However, sufficient image brightness for high magnification is difficult to obtain. As in the light microscope, a condenser system is almost invariably interposed between the gun and specimen to concentrate the beam on the region of the specimen under examination. A single condenser lens suffices for electronoptical work up to 50,000×. However, for high resolution work a double condenser system is always used, which will concentrate the beam into an area as small as 1μ diameter.

Specimen Manipulation System

The pierced metal grid carrying the specimen proper is clamped at its periphery to a suitable holder designed to conduct heat rapidly away. The specimen temperature in a TEM may rise to 200 °C. The holder, including its attached specimen grid is introduced into the evacuated specimen chamber through an airlock by means of an insertion tool. This tool is then generally withdrawn after the holder has been placed on the translation stage. The holder is then free to move with the stage, which is driven from outside the column through airlocks, by means of levers and micrometer screws. Two mutually perpendicular stage movements, each of about ±1 mm, allow any part of the grid area to be brought to the microscope axis and viewed at the highest magnification. Most instruments provide a scan magnification so that the whole grid area may be viewed at ~100×. Suitable specimen areas are then chosen, and centered for study at higher magnifications.

Imaging System

This part of the microscope includes the objective, intermediate, and projector lenses. It is involved in the generation of the image and the magnification and projection of the final image onto a viewing screen or camera system. Electrons transmitted by the specimen enter the objective lens. Those passing through the physical aperture are imaged at ~100× in the intermediate lens object plane, 10–20 cm below the specimen. The position of this primary image plane is controlled by the objective lens current (focus control). A second image, which may be magnified or diminished, is formed by the intermediate lens, the current through which controls overall magnification (magnification control). This secondary image, formed in the objective plane of the projector lens, is then further magnified, and the overall magnification is determined by the position of the fluorescent screen or film.

Image Recording System

The final image is projected onto a viewing screen coated with a phosphorescent zinc-activated cadmium sulfide powder. This powder is attached to the screen with a binder, such as cellulose nitrate. Most electron microscopes provide for an inclination of the viewing screen so that the image may be conveniently examined either with the unaided eye or through a stereomicroscope (the binoculars). Although the stereomicroscope image may appear to be rough due to the 100 μm sized phosphorescent particles that make up the screen, it is necessary to view a magnified image in order to focus accurately. Some microscopes may provide a second, smaller screen that is brought into position for focusing. In this case, the main screen remains horizontal, except during exposure of the film. All viewing screens will have areas marked to indicate where to position the image so that it will be properly situated on the film. Preevacuated films are placed into an air lock (*camera chamber*) under the viewing screen and the chamber evacuated to high vacuum. The chamber is then opened to the column to permit exposure of the film. In modern electron microscopes (Fig. 5), exposure is controlled by an electrically operated shutter placed below the projector lens. As one begins to raise the viewing screen, the shutter blocks the beam until the screen is in the appropriate position for exposure. The shutter is then opened for the proper interval, after which the beam is again blocked until the screen is repositioned.

Figure 5. Image of a 300 kV TEM (FEI-Tecnai G^2 Polara) for cryoapplications at liquid nitrogen and liquid helium temperatures. (Reproduced by courtesy of FEI Company.)

DESIGN OF THE SCANNING ELECTRON MICROSCOPE

The SEM is made up of two basic systems, and the specimen is at their boundary. The first system is the electron optical column that provides the beam of illumination that is directed to the specimen. The second system consists of the electron collection, signal amplification, and image display units, which converts the electrons emitted from the specimen into a visible image of the specimen.

Electron Optical Column

The electron gun and electron lenses are present in the electron optical column of the SEM in an analogous fashion to their presence in the TEM.

1. Electron Gun: The electron source is most commonly the hairpin tungsten filament located in a triode electron gun. The electrons are emitted by the filament (also called the cathode), and accelerated by a field produced by the anode. The anode is usually at a positive potential on the order of 15 kV with respect to the cathode. A third electrode, the shield, lies between the anode and cathode and is negative with respect to the cathode. After leaving the bias shield and forming an initial focused spot of electrons of ~ 50 μm in diameter, a series of two to three condenser lenses are used to successively demagnify this spot sometimes down to ~ 2 nm. These small spot sizes are essential for the resolutions required at high magnifications. A heated tungsten filament is the conventional electron source for most SEMs; other special sources are lanthanum hexaboride (LaB_6) and the field emission guns (FEG). Both of these latter sources produce bright beams of small diameter and have much longer lifetimes than heated tungsten filaments. Schottky emission has largely replaced earlier source technologies based on either tungsten and LaB6 emission or cold-field emission in today's focused electron beam equipment including SEM, TEM, Auger systems, and semiconductor inspection tools. Schottky and cold-field emission are superior to thermionic sources in terms of source size, brightness and lifetime. Both are up to 1000 times smaller and up to 100 times brighter than thermionic emitters.
2. Electron Lenses: Most SEMs have three magnetic lenses in their column: the first, second, and final condenser lenses. The first condenser lens begins the demagnification of the 50 μm focused spot of electrons formed in the region of the electron gun. As the amount of current running through the first condenser lens is increased, the focal length of the lens becomes progressively shorter and the focused spot of electrons becomes smaller. In our earlier discussion of electron lenses, it was noted that focusing takes place by varying the focal length. This is accomplished by changing the intensity of the lens coil current, which in turn alters the intensity of the magnetic field that is generated by the lens. As the lens current increases, the lens strength increases

and the focal length decreases. A short focal length lens consequently causes such a wide divergence of the electrons leaving the lens that many electrons are not able to enter the next condenser lens. The overall effect of increasing the strength of first condenser lens is to decrease the spot size, but with a loss of electrons. An aperture is positioned in the lenses to decrease the spot size and reduce spherical aberration by excluding the more peripheral electrons. Each of the condenser lenses behaves in a similar manner and possesses apertures.

In designing the final condenser lens, several performance characteristics must be considered:

(a) Aberrations. Since the intermediate images of the crossover produced by the condenser lenses have significantly larger diameters than the final spot size, the effect of aberrations on these lenses are relatively small. It is thus the effects of spherical and chromatic aberration as well as the astigmatism of the final condenser lens that are critical in the design and performance of the objective lens of SEM.

(b) Magnetic Field. As a result of electron bombardment, secondary electrons in the SEM are emitted over a wide solid angle. These have energies of only few electron volts, yet they must be able to reach the detector to produce the necessary signal. As a result, the magnetic field at the specimen must be designed so that it will not restrict effective secondary electron collection.

(c) Focal Length. The extent of lens aberrations is dependent upon the focal length. Thus, it is desirable to keep the latter as short as possible in order to help minimize the effects of aberrations.

The final lens usually has externally adjustable apertures. Normally, final apertures on the order of 50–70 μm are used to generate smaller, less electron dense spots for secondary electron generation and imaging. Larger apertures, for example, 200 μm, are used to generate larger spots with greater numbers of electrons. These large spots contain a great deal of energy and may damage fragile specimens. They are used primarily to generate X rays for elemental analysis rather than for imaging purposes.

Specimen Manipulation System

The specimen is normally secured to a metal stub and is grounded to prevent the build up of static high voltage charges when the beam electrons strike the specimen. In order to orient the specimen precisely, relative to the electron beam and electron detectors, all SEMs have controls for rotating and traversing the specimen in x, y, and z directions. It is also possible to tilt the specimen in order to enhance the collection of electrons by a particular detector. These movements have a large effect on magnification, contrast, resolution and depth of field. Some improvement can be made in imaging by reorientation of the specimen.

Interaction Of Electron Beam With Specimen

Three basic possibilities exist as to the nature of the beam–specimen interaction used to generate the image:

1. Some primary electrons, depending on the accelerating voltage, penetrate the solid to depths as much as 10 μm. The electrons scatter randomly throughout the specimen until their energy is dissipated by interaction with atoms of the specimen.
2. Some primary electrons collide with or pass close to the nucleus of an atom of the specimen such that there is a change in the electron's momentum. This results in electron scatter through a large angle and electron reflection from the specimen. Such elastically reflected primary electrons are known as backscattered electrons.
3. Some primary electrons interact with the host atoms so that as a result of collisions, a cascade of secondary electrons is formed along the penetration path. Secondary electrons have energy ranges of 0–50 eV and are the electrons most commonly used to generate the 3D image. The mean path length of secondary electrons in many materials is $\sim$1 nm. Thus, although electrons are generated throughout the region excited by the incident beam, only those electrons that originate $<$1 nm deep in the sample escape to be detected as secondary. The shallow depth of production of detected secondary electrons makes them very sensitive to topography.

In addition to producing backscattered and secondary electrons, specimen–beam interactions also produce photons, specimen currents, Auger electrons and X rays that are characteristic of the probed specimen. These emanations can be detected by X ray or electron spectroscopy for elemental analysis of the specimen surface. However, it is rarely used for biological specimens.

Signal Versus Noise

The signals generated as a result of the electron beam striking a specimen are used to convey different types of information about the specimen. In the usual SEM imaging mode, signals consist of the secondary electrons generated from the spot struck by the electron beam and noise consists of secondary electrons originating at locations away from where the beam struck the specimen. The image quality is eventually expressed by the signal-to-noise (S/N) ratio. In a poor quality image the signal to noise ratio is low. One may achieve a better image by either reducing the noise or raising the signal. Since it is more difficult to reduce the noise level, the signal is usually raised by increasing the electron emissions from the gun. Several methods to accomplish increased electron emissions include: altering the bias settings, decreasing the distance between the anode and the filament, decreasing the distance between the filament and the shield aperture, and using either a lanthanum hexaboride filament or a cold-field emissions gun. A second method to increase the signal is to use slower scan rates on the

specimen. Longer dwell times of the beam on the specimen will generate more secondary electrons from the spot where the beam strikes the specimen. This increase in current, however, carries with it an increased risk of damage to sensitive specimens (9).

Secondary Electron Detection

To collect the secondary electrons, a suitable electrode is held at a positive potential and serves to attract them and produce an emission current. The strength of this signal is proportional to the number of electrons striking the collector. This signal is used, after amplification; to modulate the intensity of the cathode-ray tube (CRT) beam as it moves across the tube face, synchronously with the path of the electron probe across the specimen surface.

Typically, the secondary electron collector is based on the original 1960 scintillator-photomultiplier design of Everhart and Thoronley. In this system, the secondary electrons are accelerated towards the scintillator by a potential difference of a few hundred to a few thousands volts. Upon hitting the scintillator, each electron produces many photons that are guided by the light pipe to the photomultiplier. Each photoelectron triggers a release of two or more secondary electrons at the first electrode (dynode) and process cascades, yielding from 100,000 to 50 million additional electrons. Thus, the photomultiplier reconverts the light to an electron current and provides a high degree of amplification that can be controlled by variation of the voltage applied to the dynodes (8).

Image Recording System

The final magnified image in the SEM is formed on a CRT or monitor. Unlike TEM, in which the electrons interact directly with the photographic medium, SEM images are most often photographed directly from the monitor through the lens of either a 35 mm roll film camera or a larger 4 in. $\times$ 5 in. (10.6 cm $\times$ 12.7 cm) sheet film camera. The camera shutter remains open as the electron beam slowly scans across the specimen. A valuable addition to most SEMs is the automatic data display that permits the generation of informational data on the viewing and recording monitors. With this accessory, experiment numbers, dates, accelerating voltages and magnifications may be displayed.

DIAGOSTIC ELECTRON MICROSCOPY

Electron microscopy excels as a diagnostic tool with respect to the detection and identification of both abnormal tissue anatomy and the pathogens responsible for the disease. The value of electron microscopy in difficult diagnostic situations has been demonstrated repeatedly, particularly when there is close coordination between the pathologist and the attending clinician. Ultrastructural study may be applied to a variety of substances including biological materials. By examination of specially prepared tissue sections, changes not perceived by light microscopy can be identified, leading to improved diagnostic interpretations. For example, in certain kidney diseases, such as nephrotic syndrome and Nil disease, the correct diagnosis can be made only by these means, and this in turn affects the selection of therapy. Similarly, certain neoplasms can be identified definitively only through ultrstructural studies, with obvious implications for treatment and prognosis. Another area of growing importance is the identification of viral particles in biological material. In some instances, ultrastructural study is the only way to establish the presence of a viral infection, and in other instances a diagnosis may be made earlier than by serological methods. The costs of diagnostic electron microscopy are relatively small in light of the benefits to patient care. The following are examples highlighting the use of electron microscopy in the diagnosis of certain diseases.

Neoplasms

Many neoplasms appear undifferentiated by light microscopy, but most show differentiation along one cell line or another at the ultrastructural level. However, it is noteworthy that not all of the ultrastructureal criteria for identifying the cell type may be present in every neoplasm. As expected, the more differentiated the neoplasm, the more likely will its cells contain a broad complement of diagnostic morphologic features. Usually, the ultrastructural findings do allow the pathologist to make a definitive diagnosis when interpreted in conjunction with the light microscopic picture, and in some cases with the histochemical and immunohistochemical results. The example that follows will highlight the use of diagnostic electron microscopy in the identification of carcinomas. Various types of carcinomas have a number of distinguishing features, but one common characteristic of all carcinomas is the presence of intracellular junctions, usually desmosomes and/or intermediate junctions. The presence of lumens, microvilli, tight junctions, junctional complexes, basal lamina, secretory granules, prominent Golgi apparatus, and moderately prominent rough endoplasmic reticulum are all suggestive of adenocarcinoma in the differential diagnosis of a neoplasm (10). Figure 6 shows a pancreatic carcinoma, which may arise from acinar cells, centrocinar cells, intercalated duct cells, interlobular duct cells, interlobular duct cells and main pancreatic duct cells. Adenocarcinomas arising from main and interlobular ducts (mucinous cystadenocarcinomas) have cells similar to those of bile ducts and intestinal epithelium; that is, the cytoplasm contains mucin granules, and the free surface has microvilli filled with thin filaments that anchor into the subjacent cytoplasm (11–13).

Infectious Diseases

Bacteria. Diagnostic criteria for bacterial rods and cocci are (*1*) the presence of an outer-cell wall, (*2*) the presence flagella or pili (fimbria) on the outer surface of the cell, (*3*) the presence of an inner-cell membrane, (*4*) a central nuclear region (nucleoid), without a limiting membrane, (*5*) dense cytoplasm composed mostly of ribosomes, and (*6*) a varying number of vesicles formed from the inner-cell membrane (mesosomes), storage vacuoles and endospores (14). Figure 7 shows the bacteria in Whipple disease. The rods are present both free and within macrophages.

Viruses. The distinct morphology of members of different viral families usually allows an agent to be assigned to

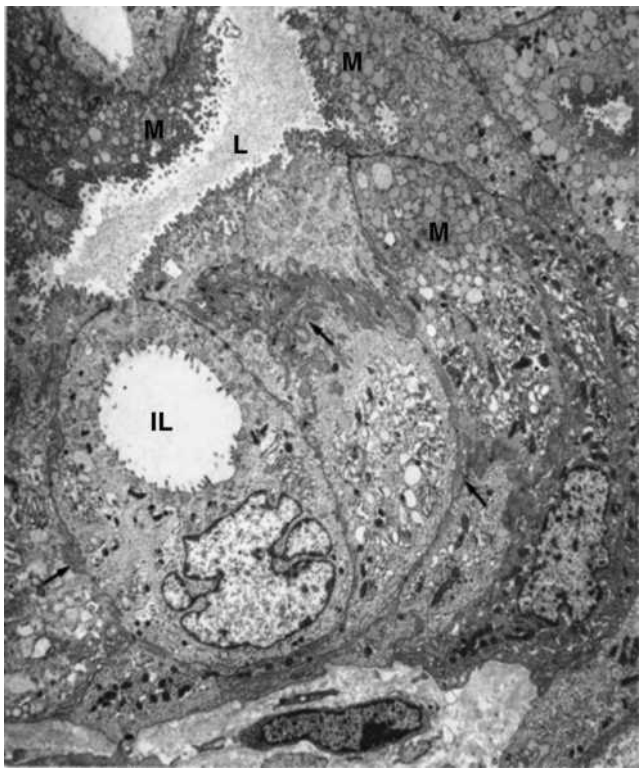


Figure 6. Ductal, mucinous cystadenocarcinoma (pancreas). In this field, the neoplastic cells form a cystic lumen (L) lined by innumerable microvilli. An intracytoplasmic lumen (IL), without junctional com- plexes, is present in one cell. Some of the cells lining the lumen have a rich collection of mucinous granules (M) in their apical cytoplasm. Lateral cell borders show a switch-backing pattern of interdigitation (arrows). (6800×) (From Ref. 10. Reproduced by courtesy of Springer-Verlag GmbH.)

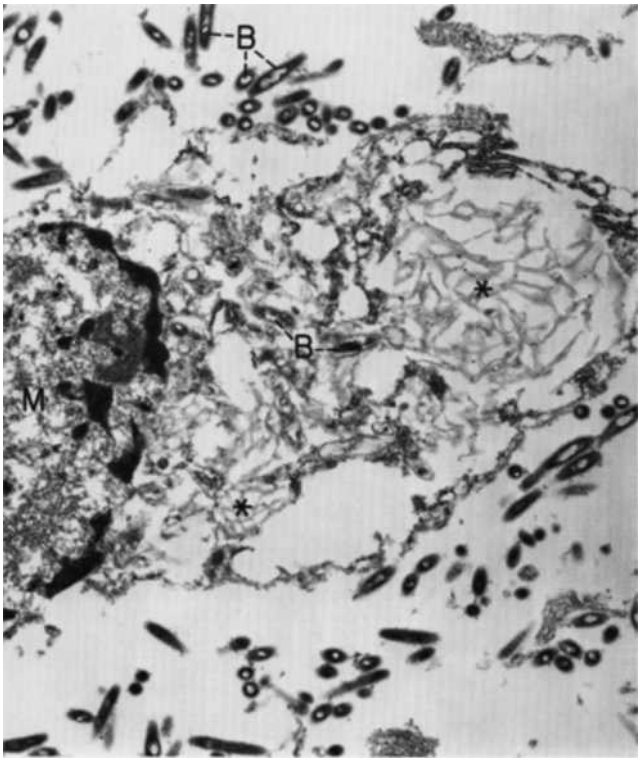


Figure 7. Whippl's disease. The lamina propria of the jejunal mucosa contained numerous Whipple's type macrophages [M= macrophage nucleus) and bacteria rods (B)]. Most of the intact bacilli are extracellular, whereas those in the macrophage are in various stages of degeneration, including the end-stage of serpiginous membrane (*). (16,500 ×) (From Ref. 10. Reproduced by courtesy of Springer-Verlag GmbH.)

a particular family. This morpho-diagnosis, combined with clinical information is often sufficient to permit a provisional diagnosis and to initiate treatment and containment protocols while waiting for other test results. Diagnostic criteria that apply to viruses in general are (*1*) the presence of intracellular and/or extracellular elliptical, stand-like, round or polygonal structures measuring 20–300 nm in diameter; and (*2*) the identification of viral morphology, consisting of a central, electron dense core (DNA-containing nucleoid) and an outer shell (capsid), which may have more than one layer (Fig. 8) (15).

Fungi. The electron microscopic diagnostic criteria for fungi include the identification of mononucleated oval yeast forms measuring 2–4 μm in diameter, with a thin cell wall and no true capsule. These can be located either extracellularly or intracellularly (16). A representative fungus, *Histoplasma capsulatum*, is shown in Fig. 9. The organisms have a clear halo between their visible cytoplasm and their thin cell wall.

Skeletal Muscle Diseases

The skeletal muscle responses to injury that are visible with the electron microscope can be categorized as follows: (*1*) alterations in the sarcolemma (e.g., discontinuities of the plasma membrane or the basement membrane); (*2*) alterations in myofilaments; (*3*) Z-band alterations (e.g., streaming and nemaline bodies); (*4*) nuclear changes (e.g., abnormal location of the nucleus within the muscle fiber and nuclear inclusions); (*5*) abnormalities of the sarcoplasmic reticulum and the T-system (e.g., tubular aggregates), (*6*) abnormal accumulations of metabolites (e.g., glycogen and lipids); (*7*) abnormal cytoplasmic structures (e.g., vacuoles, cytoplasmic bodies, concentric laminated bodies, fingerprint bodies, curvilinear bodies). In general, many of these ultrastructural abnormalities are not specific for a single disease. Electron microscopy can be a valuable adjunct to help the pathologist arrive at the proper interpretation of a muscle biopsy when taken together with all other available clinical, electrophysiologic, and histopathological data. In addition to the pathologic changes that might involve the muscle fibers themselves, many diseases of muscle also simultaneously affect adjoining connective tissue components, blood vessels and intramuscular nerves. It is therefore important to pay particular attention to these structures when examining muscle with the light and electron microscope (10). The light micrograph shown in Fig. 10 demonstrates centrally placed nuclei in the majority of the muscle fibers. The central nuclei often are surrounded by a clear area that is devoid of adenosine triphosphatase (ATPase) activity. Ultrastructural features of the paranuclear clear zone in Fig. 11 include (*1*) the

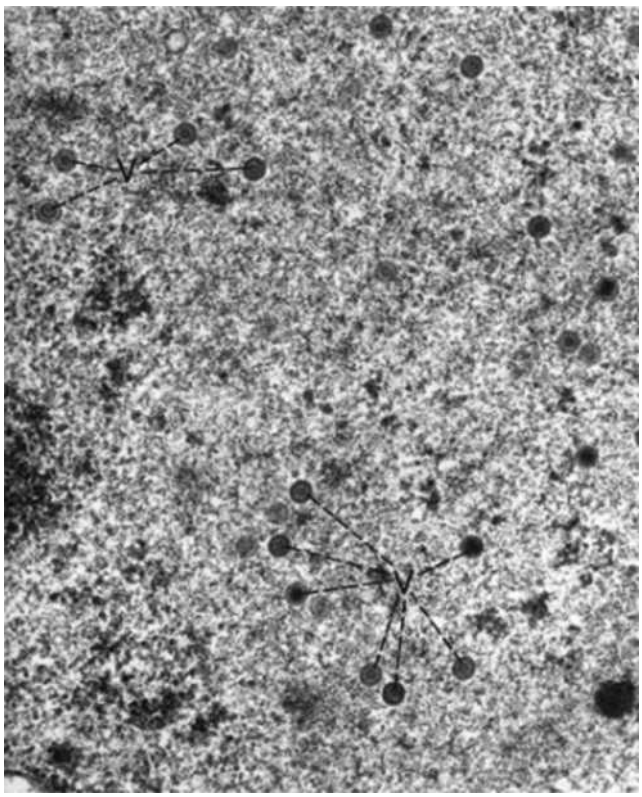


Figure 8. Herpes simplex encephalitis (cerebrum). Virions (V) have a central dense nucleoid and an outer three-layered capsid. (63,800×) (From Ref. 10. Reproduced by courtesy of Springer-Verlag GmbH.)

absence of myofilaments, (*2*) numerous mitochondria, (*3*) glycogen accumulation. In some patients, especially infants, the central clear zone may be more evident than the nuclei within that zone when examining the tissue in cross-section (10).

Peripheral Nerve Diseases

Wallerian Degeneration. Ultrastructural changes detected in peripheral nerve specimens include either the general pathologic responses of the peripheral nerve to either the injury or the specific disease entity that afflicts the patient. The general pathologic processes involving peripheral nerve can be divided into two broad categories: those that indicate a process primarily affecting the axon and those that indicate a process primarily affecting the myelin sheath. Examination of peripheral nerve biopsies by electron microscopy therefore must include evaluation of the axons, the interstitium, and the myelin and Schwann cells.

The sequence of structural changes following nerve injury that are collectively called Wallerian degeneration is shown in Fig. 12. Wallerian degeneration specifically refers to degeneration of the distal segments of a peripheral nerve after severance of the axons from their cell bodies (17). When the nerve injury is a contusion, the basement membrane of the Schwann cell is preserved, allowing regeneration within the endoneurial tube. In contrast, when the nerve injury is a transection, the endoneurial tube (composed of denervated Schawnn cells and extracellular matrix) may not be appropriately aligned with the regenerating axons. Axonal regeneration is therefore less efficient after nerve degeneration that follows a transection injury compared to that following a crush injury (10,17).

PROSPECTS OF ELECTRON MICROSCOPY

In the 1980s, electron microscopy lost much of its former role in the life sciences due to the introduction of modern, highly effective molecular analytical techniques, such as immunohistochemistry, chip technology, and confocal laser scan microscopy. In recent years, however, substantial technical improvements were made in specimen preparation, instrumentation and software, allowing the

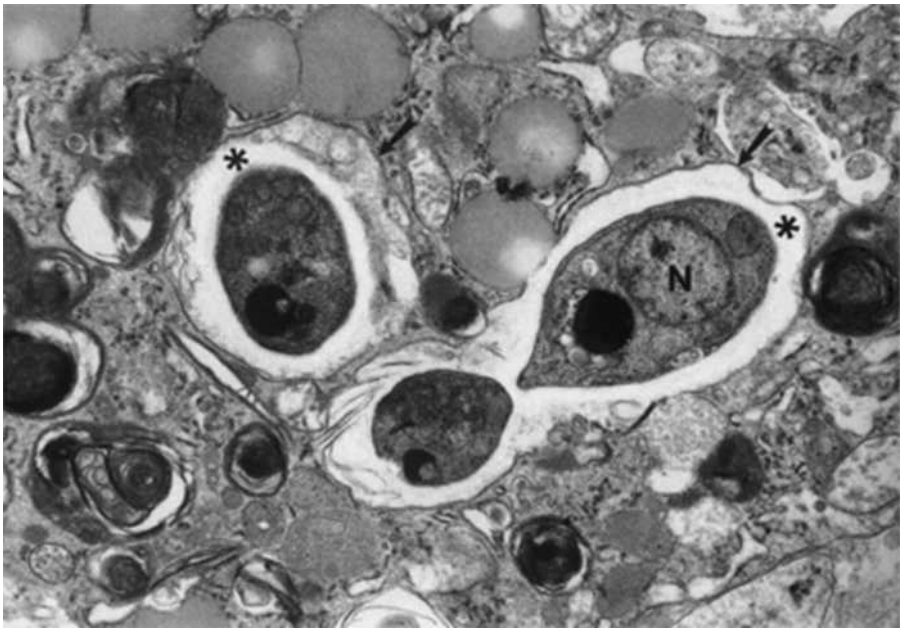


Figure 9. *Histoplasma capsulatum* (supraclavicular lymph node). High magnification of parasitic yeast forms illustrates details of their internal structure. N = nucleus; * = clear, peripheral, cytoplasmic halo; arrows = parasitic cell membrane. (20,000×) (From Ref. 10. Reproduced by courtesy of Springer-Verlag GmbH.)

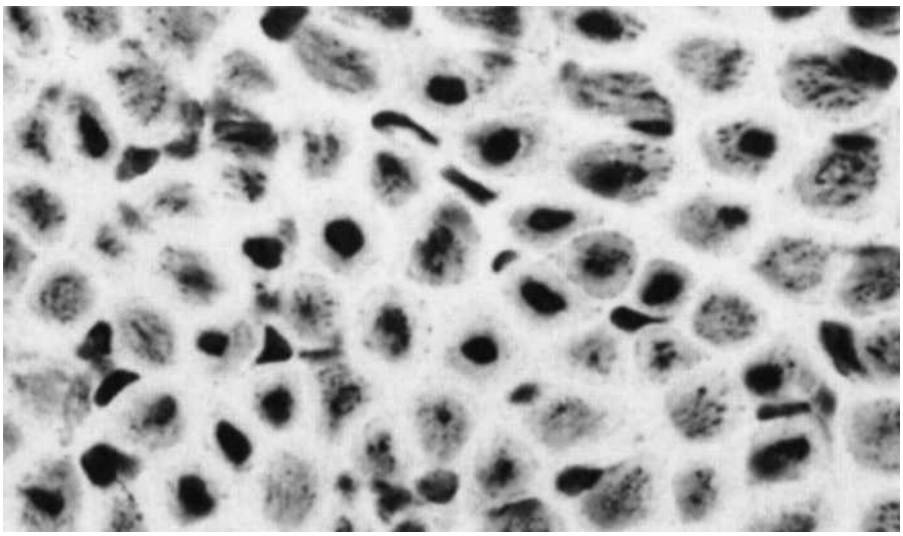


Figure 10. Centronuclear (myotubular) myopathy. The majority of small fibers contain centrally placed nuclei. (H&E, 150×) (From Ref. 10. Reproduced by courtesy of Springer-Verlag GmbH.)

electron microscope to reemerge as a valuable tool for analyzing molecular complexes.

Electron microscopy as an imaging technique allows a direct view of biological objects, while some of the other available techniques are indirect and in some instances nonspecific. Using electron microscopy, all components of the object and their mutual relationships at the molecular level can be analyzed. This information provides insight toward an understanding of structure–function relations.

The possibility of 3D reconstruction of cellular components via electron microscopy, along with the ease and speed with which newer instruments can provide data, have given the way to what many in the field are referring as a revolution. Recent technological advancements have made automated data acquisition possible, and have thus allowed a reduction of the total electron dose needed to image a specimen. Specimen preparation advances, such as embedding biological specimens in vitreous ice, have enabled studies of the macromolecular organization of cells. Whole prokaryotic and small eukaryotic cells can be directly grown and hydrated frozen on electron microscopy grids. Examination of the naturally preserved cells delivers images of the cellular structures in their functional environment. Such so-called tomograms contain all available information about the spatial relationships of macromolecular structures within the cell. However, due to their poor S/N and the generally highly crowded nature of the cytoplasm, the interpretation of

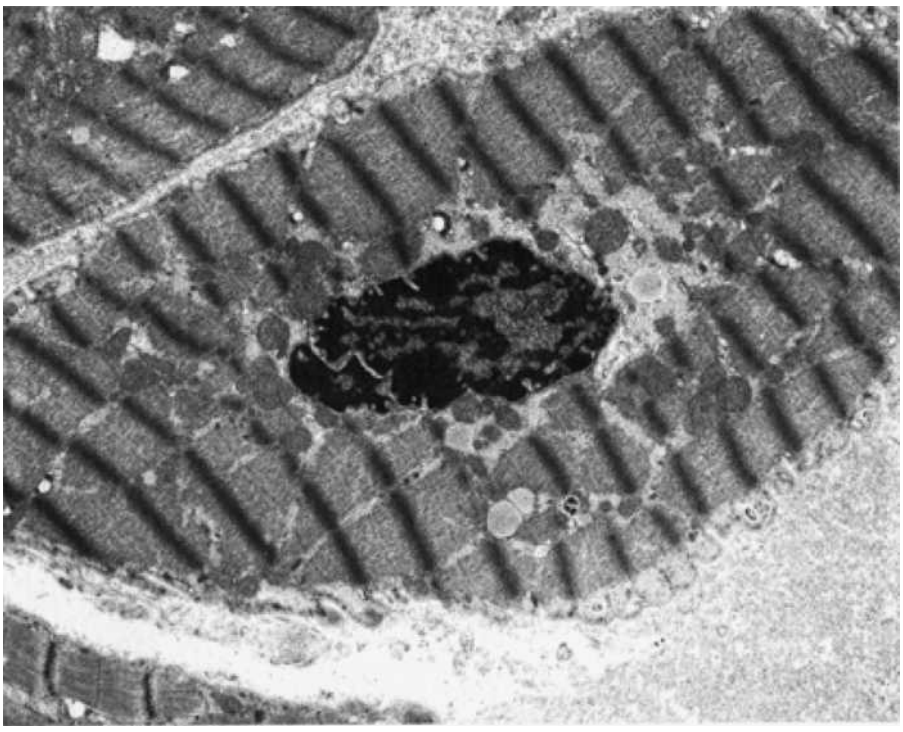


Figure 11. Muscle fiber with degenerated nucleus. Note absence of myofilaments and accumulation of glycogen in the paranuclear region to the right (15,000×). (From Ref. 10. Reproduced by courtesy of Springer-Verlag GmbH.)

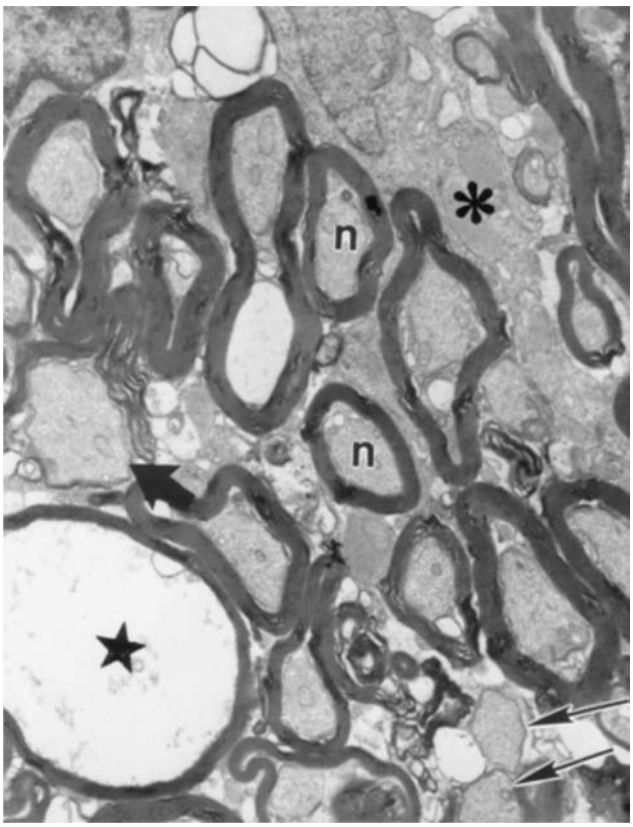


Figure 12. Ultrathin sections of opossum's optic nerve fibers 24 h after crash. Normal fibers (*n*) are seen among some altered fibers, which exhibit watery degeneration (star) and myelin sheath breakdown (thick arrow). Note demyelinated fibers (thin arrows) with an apparently intact axoplasmic cytoskeleton. Asterisk, astrocytic processes. (From Ref. 17. Reproduced by courtesy of Anais da Academia Brasileira de Ciencias.)

these tomograms remains difficult. To get significant information about specific structures in the cell, the images have to be evaluated using advanced pattern recognition methods. Existing structural models of cellular constituents at lower resolutions can guide the systematic evaluation of the tomograms. The aim is to visualize the complete 3D organization of the cell at molecular resolution. Structural evaluation by single particle analysis, electron crystallography and electron tomography is slow compared to other structure determination technologies, in particular X-ray crystallography. Processing time for the electronic technique is typically in the range of several months per solved structure, depending on the resolution achieved. The same task can be accomplished in the range of hours or days for X-ray crystallography, once suitable crystals are available. Continued joint efforts between the research community and manufacturers to develop user-friendly, universal interfaces between electron crystallography, single-particle analysis and electron tomography would improve this situation, and further expand the usefulness of of these electronic technologies.

BIBLIOGRAPHY

1. Nellist PD, et al Direct sub-angstrom imaging of a crystal lattice. Science 2004;305(5691):1741.
2. Freeman MR. Time -resolved scanning tunneling microscopy through tunnel distance modulation. Appl Phys Lett 1993;68(19):2633–2635.
3. Bonetta L. Zooming in on electron tomography. Nature Methods 2005;2(2):139–44.
4. Bozzola JJ, Russell LD. Electron Microscopy: Principals and techniques for biologists. Sudbury (MA): Jones and Bartlett Publishers; 1998.
5. Hayat MA. Principels and techniques of electron microscopy: Biological application. New York: Van Nostrand Reinhold Company; 1973.
6. Wischnitzer S. Introduction to electron microscopy. New York: Pergamon Press; 1981.
7. Meek GA. Practical electron microscopy for biologists. New York: John Wiley & Sons inc; 1976.
8. Joy DC. Beam interactions, contrast and resolution in the SEM. J Microsc 1984;136:241–58.
9. Haine M. The electron microscope: The present state of the Art. London: Spon; 1961.
10. Dickersin GR. Diagnostic Electron Microscopy: A text/ atlas. New York: Springer-Verlag; 1999.
11. Franchina M, Del Borrello E, Caruso A, Altavilla G. Serous tumors of the ovary: Ultrastructural observations. Eur J Gynaecol Oncol 1992;13(3):268–76.
12. Wolf HK, Garcia JA, Bossen EH. Oncocytic differentiation in intrahepatic biliary cystadenocarcinoma. Modern Pathol 1992;5(6):665–866.
13. Kobayashi TK, et al. Effects of Taxol on ascites cytology from a patient with fallopian tube carcinoma: Report of a case with ultrastructural studies. Diagn Cytopathol 2002;27(2):132–134.
14. Yogi T, et al Whipple's disease: The first Japanese case diagnosed by electron microscopy and polymerase chain reaction. Intern Med 2004;43(7):566–570.
15. Jensen HL, Norrild B. Herpes simplex virus-cell interactions studied by immunogold cryosection electron microscopy. Methods Mol Biol 2005;292:143–160.
16. Garrison RG, Boyd KS. Electron microscopy of yeastlike cell development from the microconidium of Histoplasma capsulatum. J Bacteriol 1978;133(1):345–353.
17. Narciso MS, Hokoc JN, Martinez AM. Watery and dark axons in Wallerian degeneration of the opossum's optic nerve: Different patterns of cytoskeletal breakdown? An Acad Bras Cienc 2001;73(2):231–243.

See also ANALYTICAL METHODS, AUTOMATED; CELLULAR IMAGING; CYTOLOGY, AUTOMATED.

MICROSCOPY, FLUORESCENCE

SERGE PELET
MICHAEL PREVITE
PETER T. C. SO
Massachusetts Institute of Technology
Cambridge, Massachusetts

INTRODUCTION

Fluorescence microscopy quantifies the distribution of fluorophores and their biochemical environment on the

micron length scale and allows *In vivo* measurement of biological structures and functions (1–3). Heimstädt developed one of the earliest fluorescence microscopes in 1911. Some of the first biochemical applications of this technique include the study of living cells by the protozoologist Provazek in 1914.

Fluorescence microscopy is one of the most ubiquitous tools in biomedical laboratories. Fluorescence microscopy has three unique strengths. First, the fluorescence microscope has high biological specificity. Based on endogenous fluorophores or exogenous probes, fluorescence microscopy allows the association of a fluorescence signal with a specimen structural and biochemical state. While fluorescence microscopy has comparable resolution to white light microscopes, their range of applications in biomedicine is much broader.

Second fluorescence microscopy is highly sensitive in the imaging of cells and tissues. The high sensitivity of fluorescence microscopy originates from two factors. One factor is the significant separation between the fluorophores' excitation and emission spectra. This separation allows the fluorescence signal to be detected by efficiently rejecting the excitation radiation background using bandpass filters. The fluorescence microscope has the sensitivity to image even a single fluorophore. The other factor is the weak endogenous fluorescence background in typical biological systems. Since there is minimal background fluorescence, weak fluorescence signal from even a few fluorescent exogenous labels can be readily observed.

Third, fluorescence microscopy is a minimally invasive imaging technique. *In vivo* labeling and imaging procedures are well developed. While photodamage may still result from prolonged exposure of shorter excitation radiation, long-term observation of biological processes is possible. Today, a single neuron in the brain of a small animal can be imaged repeatedly over a period of months with no notable damage.

SPECTROSCOPIC PRINCIPLES OF FLUORESCENCE MICROSCOPY

Fluorescence Spectroscopy

An understanding of spectroscopic principles is essential to master fluorescence microscopy (4–6). Fluorescence is a photon emission process that occurs during molecular relaxation from electronic excited states. Historically, Brewster first witnessed the phenomenon of fluorescence in 1838 and Stokes coined the term fluorescence in 1852. These photonic processes involve transitions between electronic and vibrational states of polyatomic fluorescent molecules (fluorophores) by the absorption of either one or more photons. Electronic states are typically separated by energies on the order of $10{,}000\,\mathrm{cm}^{-1}$ and vibrational sublevels are separated by $\sim 10^2$–$10^3\,\mathrm{cm}^{-1}$. In a one-photon excitation process, photons with energies in the ultraviolet (UV) to the blue–green region of the spectrum are needed to trigger an electronic transition, whereas photons in the infrared (IR) spectral range are required for two-photon excitation. The molecules from the lowest vibrational level of the electronic ground state are excited to an accessible vibrational level in an electronic excited state. The molecule is quickly relaxed to the lowest vibrational level of the excited electronic state after excitation on the time scale of femtoseconds to picoseconds via vibrational processes. The energy loss in the vibrational relaxation process is the origin of the Stoke shift where fluorescence photons have longer wavelengths than the excitation radiation. The coupling of the ground and excited – state both for the absorption and emission process is governed by the Franck–Condon principle, which states that the probability of transition is proportional to the overlap of the initial and final vibrational wave function. Since the vibrational level structures of the excited and ground states are similar, the fluorescence emission spectrum is a mirror image of the absorption spectrum, but shifted to lower wavelengths. The shift between the maxima of the absorption and emission spectra is referred to as the Stokes' shift. The residence time of a fluorophore in the excited electronic state before returning to the ground state is called the fluorescence lifetime. The fluorescence lifetime is typically on the order of nanoseconds. The Jablonski diagram represents fluorescence excitation and deexcitation processes (Fig. 1).

Fluorescence deexcitation processes can occur via radiative and nonradiative pathways. Radiative decay describes molecular deexcitation processes accompanied by photon emission. Molecules in the excited electronic states can also relax by nonradiative processes where excitation energy is not converted into photons, but are dissipated by thermal processes, such as vibrational relaxation and collisional quenching. Let Γ and k be the radiative and nonradiative decay rates, respectively, and N be the number of fluorophore in the excited state. The temporal evolution of the excited state can be described by

$$\frac{dN}{dt} = -(\Gamma + k)N \tag{1}$$

$$N = N_0 e^{-(\Gamma+k)t} = N_0 e^{-t/\tau} \tag{2}$$

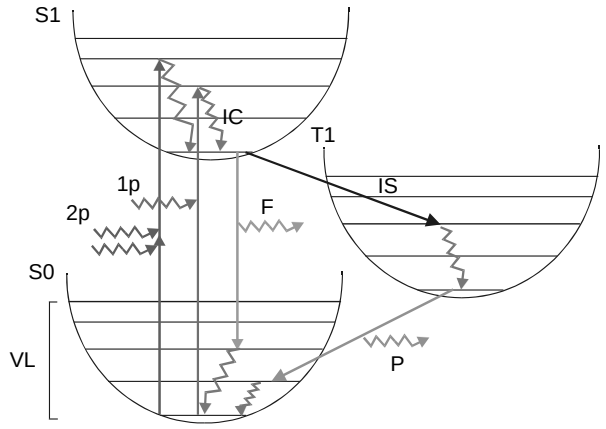


Figure 1. A Jablonski diagram describing fluorescence (F) and phosphorescence (P) emission and excitation processes based on one-photon (1p) and two-photon (2p) absorption. The parameters S0, S1, and T1 are the electronic singlet ground state, singlet excited state, and triplet excited state, respectively. Here VL denotes vibrational levels, IC denotes internal conversion, and IS denotes intersystem crossing.

The fluorescence lifetime, τ, of the fluorophore is the combined rate of the radiative and nonradiative pathways:

$$\tau = \frac{1}{\Gamma + k} \tag{3}$$

One can define the intrinsic lifetime of the fluorophore in the absence of nonradiative decay processes as, τ_0:

$$\tau_0 = \frac{1}{\Gamma} \tag{4}$$

The efficiency of the fluorophore can then be quantified by the fluorescence quantum yield, Q, which measures the fraction of excited fluorophore relaxing via the radiative pathway:

$$Q = \frac{\Gamma}{\Gamma + k} = \frac{\tau}{\tau_0} \tag{5}$$

Environmental Effect on Fluorescence

A number of factors contributes to the nonradiative decay pathways of the fluorophores and reduces fluorescence intensity. In general, the nonradiative decay processes can be classified as

$$k = k_{\mathrm{ic}} + k_{\mathrm{ec}} + k_{\mathrm{et}} + k_{\mathrm{is}} \tag{6}$$

where k_{ic} is the rate of internal conversion, k_{ec} is the rate of external conversion, k_{et} is the rate of energy transfer, and k_{is} is the rate of intersystem crossing.

Internal conversion describes the process where the electronic energy is converted to thermal energy via a vibrational process. The more interesting process is external conversion, where fluorophores lose electronic energy in collision process with other solutes. Several important solute molecules, such as oxygen, are efficient fluorescence quenchers. The external conversion process provides a convenient mean to measure the concentration of these molecules in the microenvironment of the fluorophore. The fluorophore is deexcited nonradiatively upon collision. The collisional quenching rate can be expressed as

$$k_{\mathrm{ec}} = k_0[Q] \tag{7}$$

where $[Q]$ is the concentration of the quencher and k_0 is related to the diffusivity and the hydrodynamics radii of the reactants.

When collisional quenching is the dominant nonradiative process, equation 1 predicts that fluorescence lifetime decreases with quencher concentration.

$$\frac{\tau_0}{\tau} = (1 + k_0\tau_0[Q]) \tag{8}$$

Collision quenching also reduces the steady-state fluorescence intensity, F, relative to the fluorescence intensity in the absence of quencher, F_0. The Stern–Volmer equation describes this effect:

$$\frac{F_0}{F} = 1 + k_0\tau_0[Q] \tag{9}$$

A related process is steady-state quenching, where fluorescence signal reduction is due to ground-state processes. A fluorophore can be chemically bound to a quencher to form a dark complex, a product that does not fluoresce. In this case, steady-state fluorescence intensity also decreases with quencher concentration as

$$\frac{F_0}{F} = 1 + K_{\mathrm{s}}[Q] \tag{10}$$

where K_{s} is the association constant of the quencher and the fluorophore. However, since steady-state quenching is a ground-state process that only reduces the fraction of fluorophores available for excitation, fluorescence lifetime is not affected.

Resonance energy-transfer rate, k_{et}, becomes significant when two fluorophores are in close proximity within ~5–10 nm as during molecular binding. The energy of an excited donor can be transferred to the accepted molecule via an induced dipole–induced dipole interaction. Let D represents the donor and A, the acceptor. Under illumination at the donor excitation wavelength, the number of excited donors and acceptors are N^{D}, N^{A}, respectively. Further, define the donor and acceptor deexcitation rates as k_{D} and k_{A}. The excited-state population dynamics of the donor and acceptor can be described as

$$\frac{dN^{\mathrm{D}}}{dt} = -(k_{\mathrm{D}} + k_{\mathrm{et}})N^{\mathrm{D}} \tag{11}$$

$$\frac{dN^{\mathrm{A}}}{dt} = -k_{\mathrm{A}}N^{\mathrm{A}} + k_{\mathrm{et}}N^{\mathrm{D}} \tag{12}$$

Solving these equations provides the dynamics of donor and acceptor fluorescence:

$$N^{\mathrm{D}} = N_0^{\mathrm{D}}\exp[(-k_{\mathrm{D}} - k_{\mathrm{et}})t] \tag{13}$$

$$N^{\mathrm{A}} = N_0^{\mathrm{D}}\frac{k_{\mathrm{et}}}{k_{\mathrm{A}} - k_{\mathrm{D}} - k_{\mathrm{et}}}[\exp(-k_{\mathrm{D}}t - k_{\mathrm{et}}t) - \exp(-k_{\mathrm{A}}t)] \tag{14}$$

The donor decay is a shortened single exponential, but the acceptor dynamics is more complex with two competing exponential processes.

The intersystem crossing rate, k_{is}, describes transitions between electronic excited states with wave functions of different symmetries. The normal ground state is a singlet state with an antisymmetric wave function. Excitation of the ground-state molecule via photon absorption results in the promotion of the molecule to an excited state with an antisymmetric wavefunction, another singlet state. Due to spin–orbit coupling, the excited molecule can transit into a triplet sate via intersystem crossing. The subsequent photon emission from the triplet state is called phosphorescence. Since the decay of the triplet state to the singlet ground state is forbidden radiatively, the triplet excited state has a very long lifetime on the order of microseconds to milliseconds.

FLUORESCENCE MICROSCOPE DESIGNS

The components common to most fluorescence microscopes are the light sources, the optical components, and the detection electronics. These components can be configured to create microscope designs with unique capabilities.

Fluorescence Excitation Light Sources

Fluorescence excitation light sources need to produce photons with sufficient energy and flux level. The ability to collimate the emitted rays from a light source further determines its applicability in high resolution imaging. Other less critical factors, such as wavelength selectivity, ease of use, and cost of operation, should also be considered.

Mercury arc lamps are one of the most commonly used light sources in fluorescence microscopy. The operation of a mercury arc lamp is based on the photoemission from mercury gas under electric discharge. The photoemission from a mercury arc consists of a broad background punctuated by strong emission lines. A mercury lamp can be considered as a quasimonochromatic light source by utilizing one of these strong emission lines. Since mercury lamps have emission lines throughout the near-UV and visible spectrum, the use of a mercury lamp allows easy matching of the excitation light spectrum with a given fluorophore by using an appropriate bandpass filter. Mercury arc lamps are also low cost and easy to use. However, since the emission of mercury lamps are difficult to collimate, they are rarely used in high resolution techniques, such as confocal microscopy. The advent of high power, energy efficient, light-emitting diodes (LEDs) with a long operation life allows the design of new light sources that are replacing arc lamps in some microscopy applications.

Laser light sources are commonly used in high resolution fluorescence microscopes. Laser light sources have a number of advantages including monochromatcity, high radiance, and low divergence. Due to basic laser physics, the laser emission is almost completely monochromatic. For fluorescence excitation, a monochromatic light source allows very easy separation of the excitation light from the emission signal. While the total energy emission from an arc lamp may be higher than some lasers, the energy within the excitation band is typically a small fraction of the total energy. In contrast, lasers have high radiance: the energy of a laser is focused within a single narrow spectral band. Therefore, the laser emission can be more efficiently used to trigger fluorescence excitation. Furthermore, laser emission has very low divergence and can be readily collimated to form a tight focus at the specimen permitting high resolution imaging. Gas lasers, such as the argon–ion laser and helium–neon lasers, are commonly used in fluorescence microscopy. Nowadays, they tend to be replaced by solid-state diode lasers that are more robust and fluctuate less. Lasers can further be characterized as continuous wave and pulsed. While continuous wave lasers are sufficient for most applications, pulsed lasers are used in two-photon microscopes where high intensity radiation is required for efficient induction of nonlinear optical effects.

Microscope Optical Components

The optical principle underlying fluorescence microscopes can be understood using basic ray tracing (7,8). The ray tracing of light through an ideal lens can be formulated into four rules: (1) A light ray originated from the focal point of a lens will emerge parallel to the optical axis after the lens. (2) A light ray propagating parallel to the optical axis will pass through the focal point after the lens. (3) Light rays originated from the focal plane of a lens will emerge collimated. (4) Collimated light rays incident upon a lens will focus at its focal plane (Fig. 2). From these rules, one can see that a simple microscope can be formed using two lenses with different focal lengths (Fig. 3). The lens, L1, with focal length, $f1$, images the sample plane and is called the objective. The lens, L2, with focal length, $f2$, projects the image onto the detector plane and is called the tube lens. From simple geometry, two points P1 and P2 separated by x in the sample plane will be separated by $x(f2/f1)$ at the detector plane, where the ratio $M = f2/f1$ is called the magnification. One can see that the image in the sample plane is enlarged by the magnification factor at the detector.

By using the common wide-field fluorescence microscope as an example, we can further examine the components of a

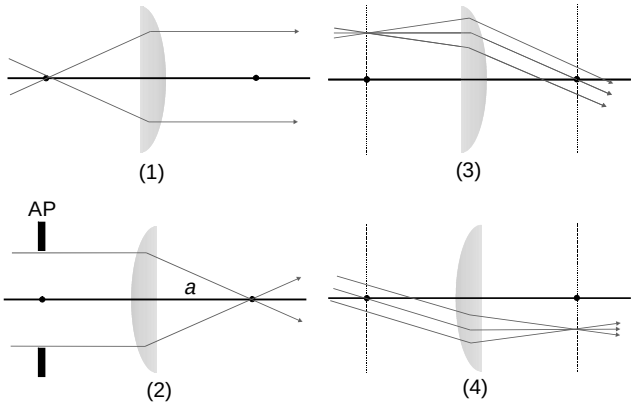


Figure 2. Four basic rules of optical ray tracing. (1) Light emerging from the focal point will become collimated parallel to the optical axis after the lens. And inversely (2) a collimated beam parallel to the optical axis will be focused at the focal plane of the lens. (3) A light source in the focal plane of the lens will become collimated after passing through the lens with an oblique angle determined by the distance from the optical axis and inversely (4) An oblique collimated beam will be focused in the focal plane by the lens. The numerical aperture of an imaging system is a function of the maximum convergence angle, α, as defined in rule 2. The maximum convergence angle is a function of the lens property and its aperture (AP) size.

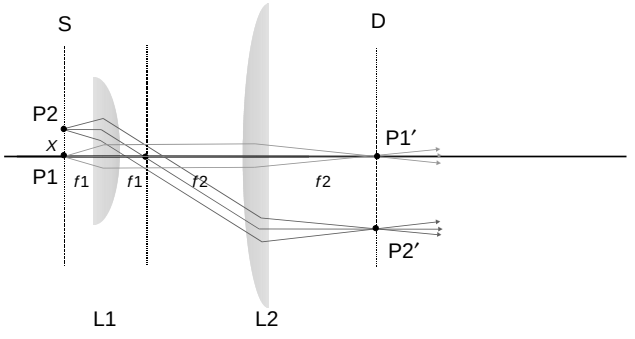


Figure 3. The detection path of a microscope. Lenses L1 and L2 are the objective and the tube lens, respectively. L1 has focal length $f1$ and L2 has focal length $f2$. For two points, P1 and P2 with separation, x, on the sample plane (S), these points are projected to points P1' and P2' on the detector plane (D).

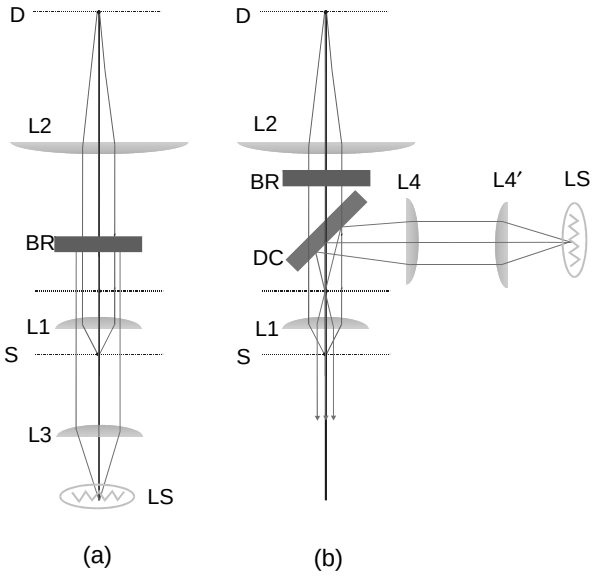


Figure 4. Two configurations of fluorescence microscopy (a) trans-illumination and (b) epi-illumination. The objective and detection tube lenses are L1 and L2. The condenser is L3. The excitation relay lenses are L4 and L4′. The sample and detector planes are S and D respectively. The light source is LS. The dichroic filter and the barrier filter are DC and BR.

complete fluorescence microscope system (Fig. 4a) (9). In addition to the detection optical path, fluorescence microscope requires an excitation light source. The excitation light source is typically placed in the focal point of a third lens, L3. The lens collimates the excitation light and projects it uniformly on the specimen (Koehler illumination). The lens, L3, is called the condenser. Since the excitation light is typically much stronger than the fluorescence emission, a bandpass filter is needed to block the excitation light. In this trans-illumination configuration, it is often difficult to select a bandpass filter with sufficient blocking power without also losing a significant portion of the fluorescence signal. To overcome this problem, an alternative geometry, epi-illumination, is commonly used (Fig. 4b). In this geometry, lens L1 functions both as the imaging objective and the condenser for the excitation light. A couple of relay lenses (L4, L4') are used to focus the excitation light at the back aperture plane of the objective via a dichroic filter that reflects the excitation light but transmits the fluorescence signal. The excitation light is collimated by L1 and uniformly illuminates the sample plane. The fluorescence signal from the sample is collected by the objective and projected onto the detector via the tube lens L2. Since the excitation light is not directed at the detector, the task of rejecting excess excitation radiation at the detector is significantly easier. A barrier filter is still needed to eliminate stray excitation radiation from the optical surfaces.

From Fig. 2, one may assume that arbitrarily small objects can be imaged by increasing the magnification ratio. However, this is erroneous as the interference of light imposes a resolution limit on an optical system (10). The smallest scale features that can be resolved using fluorescence microscopy are prescribed by the Abbe limit. For an infinitely small emitter at the sample plane, the image at the detector, the point spread function (PSF), is not a single point. Instead, the intensity is distributed according to an Airy function with a diameter, d:

$$d = M\frac{1.22\lambda}{\mathrm{NA}} \tag{15}$$

where M is the magnification of the system, λ is the emission wavelength, and NA is the numerical aperture of the objective, which is defined as (Fig. 2):

$$NA = n\sin\alpha \tag{16}$$

where α is the half-convergence angle of the light and n is the index of refraction of the material between the lens and the sample. Therefore, the images of two objects on the sample plane will overlap if their separation is $<1.22\lambda/\mathrm{NA}$. Since NA is always on the order of 1, an optical system can only resolve two separate objects if their separation is on the order of the wavelength of light.

Fluorescence Detectors and Signal Processing

Since the fluorescence signal is relatively weak, sensitive detectors are crucial in the design of a high performance fluorescence microscope. For a wide field microscope, the most commonly used detectors are charged couple device (CCD) cameras, which are area detectors that contain a rectilinear array of pixels. Each pixel is a silicon semiconductor photosensor called a photodiode. When light is incident upon an individual photodiode, electrons are generated in the semiconductor matrix. Electrodes are organized in the CCD camera such that the charges generated by optical photons can be stored capacitatively during the data acquisition. After data acquisition, manipulating the voltages of the electrodes on the CCD chip allows the charges stored in each pixel to be extracted from the detector sequentially and read out by the signal conditioning circuit. These cameras are very efficient devices with a quantum efficiency up to $\sim$80% (i.e., they can detect up to 8 out of 10 incident photons). Furthermore, CCD cameras can be made very low noise such that even four to five photons stored in a given pixel can be detected above the inherent electronic noise background of the readout electronics.

While CCDs are the detector of choice for wide-field microscopy imaging, there are other microscope configurations (discussed below) where an array detector is not necessary and significantly lower cost single element detectors can be used. Two commonly used single element detectors are avalanche photodiodes (APDs) and photomultiplier tubes (PMTs).

Avalanche photodiodes and photomultiplier tubes have been used in confocal and multiphoton microscopes. Avalanche photodiodes are similar to the photodiode element in a CCD chip. By placing a high voltage across the device, the photoelectron generated by the photon is accelerated across the active area of the semiconductor and collide with other electrons. Some of these electrons gain sufficient mobility from the collision and are accelerated toward the anode of the device themselves. This results in an avalanche effect with a total electron gain on the order

of hundreds to thousands. A sizable photocurrent is generated for each input photon. A normal photodiode or a CCD camera does not have single photon sensitivity because the readout electronic noise is higher than the single electron level. The gain in the avalanche photodiode allows single photon detection. Photomultiplier tubes operate on a similar concept. A photomultiplier is not a solid-state device, but a vacuum tube where the photons impact the cathode and generates a photoelectron using the photoelectric effect. The electron generated is accelerated by a high voltage toward a second electrode, called a dynode. The impact of the first electron results in the generation of a cascade of new electrons that are then accelerated toward the next dynode. A photomultiplier typically has ~5–10 dynode stages. The electron current generated is collected by the last electrode, the anode, and is extracted. The electron gain of a photomultiplier is typically >1–10 million. While APDs and PMTs are similar devices, they do have some fundamental differences. The APD are silicon devices and have a very high quantum efficiency (~80%) from the visible to the near-IR spectral range. The PMT photocathode material has a typical efficient of 20%, but can reach ~40% in the blue–green spectral range. However, PMTs are not sensitive in the red–IR range with quantum efficiency dropping to a few percent. On the other hand, PMT have significantly higher gain and better temporal resolution.

Advanced Fluorescence Microscopy Configurations

In addition to wide-field imaging, fluorescence microscopy can be implemented in other more advanced configurations to enable novel imaging modes. We will cover four other particularly important configurations: wide-field deconvolution microscopy, confocal microscopy, two-photon microscopy, and total internal reflection microscopy.

Wide-Field Deconvolution Microscopy. Wide-field microscopy is a versatile, low cost, and widely used technique. However, cells and tissues are inherently three dimensional (3D). In a thick sample, the signals from multiple sample planes are integrated to form the final image. Since there is little correlation between the structures at different depths, the final image becomes fuzzy. The need for 3D resolved imaging has long been recognized. The iterative deconvolution approach has worked well for relatively thin specimen, such as in the imaging of organelle structures in cultured cells (11) (Fig. 5). In terms of instrument modifications, the main difference between deconvolution microscopy and wide-field microscope is the incorporation of an automated axial scanning stage allowing a 3D image stack to be acquired from the specimen. An initial estimate of the 3D distribution of fluorophores is convoluted with the known PSF of the optical system. The resultant image is then compared with the measured 3D experimental data. The differences allow a better guess of the actual fluorophore distribution. This modified fluorophore distribution is then convoluted with the system PSF again and allows another comparison with experimental data. This process repeats until an acceptable difference between the convoluted image and the experimental data

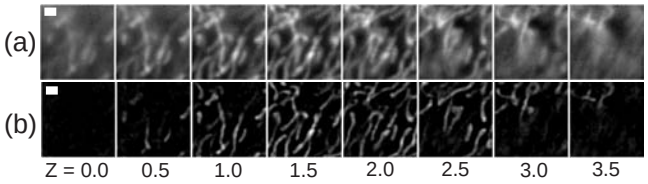


Figure 5. A comparison between (a) normal wide-field images and (b) deconvoluted images (11). Green fluorescent protein labeled mitochondria of a cultured cell was imaged by a wide-field fluorescence microscope as a 3D image stack. The image stack is deconvoluted and the significantly improved result is shown. The axial position of the image stack is shown below in units of micrometers.

is achieved. The deconvolution process in a wide-field fluorescence microscope belongs to the class of mathematical problems called ill-posed problems (12–14). An ill-posed problem does not have a unique solution, but depends on the selection of approach constraints to reach a final solution. One should consider the deconvoluted images only as the best estimate of the real physical structure given the available data. Furthermore, deconvolution algorithm is computationally intensive and often fails in thick specimens.

Confocal Fluorescence Microscopy. Confocal fluorescence microscopy is a powerful method that can obtain 3D resolved sections in thick specimens by completely optical means (15–18). The operation principle of confocal microscopy is relatively straightforward. Consider the following confocal optical system in the transillumination geometry (Fig. 6). Excitation light is first focused at an excitation pinhole aperture. An excitation tube lens collimates the rays and projects them toward the condenser. The excitation light is focused at the specimen. The emitted light from the focal point is collected by the objective and collimated by the emission tube lens. The collimated light is subsequently refocused at the emission pinhole aperture. The detector is placed behind the aperture. As it is clear in the ray tracing illustration, the fluorescence signal produced at the specimen position defined by the excitation

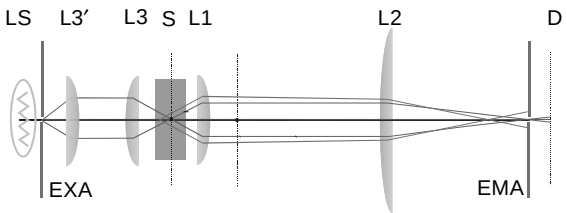


Figure 6. The configuration of a simple confocal microscope. The objective and the detection tube lenses are L1, L2. The light source is LS. The excitation aperture placed in front of the light source is EXA. The relay lens that images the excitation aperture and projects the image of the pinhole onto the specimen (S) are L3 and L3′. The fluorescence emission from the focal point (red rays) are projected onto the emission aperture (EMA) by L1 and L2. The signal is transmitted through EMA and is detected by the detector (D). Fluorescence generated outside the focal plane in the specimen (blue rays) are defocused at EMA and are mostly blocked.

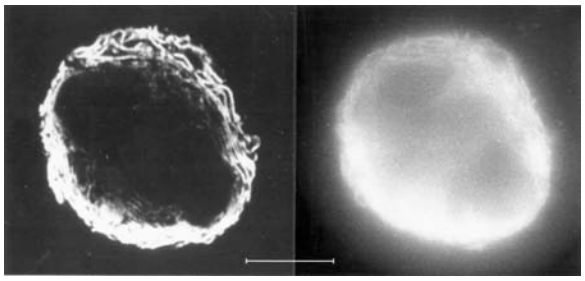


Figure 7. A comparison between confocal (a) and wide field (b) imaging of a plasmacytoma cell labeled with fluorescent antiendoplasmin that binds mainly to the endoplasmic reticulum. In the wide-field image, it is not possible to determine whether the central nucleic region contains endoplasmin end the structure of the cisternae are unclear (19).

pinhole aperture is exactly transmitted through the conjugate pinhole in the emission light path. However, for a fluorescence signal generated above or below the focal plane, the light is defocused at the emission pinhole aperture and is largely rejected. Hence, a pair of conjugate pinholes allows the selection of a 3D defined volume. One can show that a confocal microscope can image structures in 3D with a volume resolution of 0.1 fl. This system achieves 3D resolution, at the cost of obtaining fluorescence signal from only a single point in the specimen. It is necessary to raster scan the excitation focus to cover a 3D volume. Confocal microscopy has been used extensively to investigate microstructures in cells and in the imaging of tissues (19) (Fig. 7).

Two-Photon Fluorescence Microscopy. A two-photon microscope is an alternative to confocal microscopy for the 3D imaging of thick specimens. Denk, Webb, and coworkers in 1990 introduced two-photon excitation microscopy (18,20). Fluorophores can be excited by the simultaneous absorption of two photons each having one-half of the energy needed for the excitation transition. Since the two-photon excitation probability is significantly less than the one-photon probability, two-photon excitation occurs only at appreciable rates in regions of high temporal and spatial photon concentration. The high spatial concentration of photons can be achieved by focusing the laser beam with a high numerical aperture objective to a diffraction-limited spot. The high temporal concentration of photons is made possible by the availability of high peak power pulsed lasers (Fig. 8). Depth discrimination is the most important feature of multiphoton microscopy. In the two-photon case, $>80\%$ of the total fluorescence intensity comes from a 1 μm thick region about the focal point for objectives with numerical aperture of 1.25. For a 1.25 NA objective using excitation wavelength of 960 nm, the typical point spread function has a fwhm of 0.3 μm in the radial direction and 0.9 μm in the axial direction (Fig. 8). Two-photon microscopy has a number of advantages compared with confocal imaging: (*1*) Since a two-photon microscope obtains 3D resolution by limitation the region of excitation instead of the region of detection as in a confocal system, photodamage of biological specimens is restricted to the focal point. Since out-of-plane chromophores are not excited, they are not subject to photobleaching. (*2*) Two-photon excitation wavelengths are typically redshifted to about twice the one-photon excitation wavelengths in the IR spectral range. The absorption and scattering of the excitation light in thick biological specimens are reduced. (*3*) The wide separation between the excitation and emission spectra ensures that the excitation light and Raman scattering can be rejected without filtering out any of the fluorescence photons. An excellent demonstration of the ability of two-photon imaging for deep tissue imaging is in the neurobiology area (21) (Fig. 9).

Total internal reflection microscopy. Confocal and two-photon microscopy can obtain 3D resolved images from specimens up to a few hundred micrometers in thickness. However, both types of microscopy are technically challenging, require expensive instrumentation, and only can acquire data sequentially from single points. Total internal reflection microscopy (TIRM) is an interesting alternative if 3D-resolved information is only required at the bottom surface of the specimen, such as the basal membrane of a cell (22–24). Total internal reflection occurs at an interface between materials with distinct indices of refraction (Fig. 10). If light ray is incident from a high index prism, n_2, toward the lower index region, n_1, at an angle θ, the light will be completely reflected at the interface if θ is $>\theta_c$, the critical angle.

$$\sin\theta_c = \frac{n_1}{n_2} \tag{17}$$

While the light is completely reflected at the interface, the electric field intensity right above the interface is nonzero, but decays exponentially into the low index medium. The decay length of the electric field is on the order of tens to hundreds of nanometers. Compared with other forms of 3D resolved microscopy, TIRM allows the selection of the thinnest optical section, but only at the lower surface of the sample. While prism launch TIRM as described is simpler to construct, the bulky prism complicates the routine use of TIRM for cell biology studies. Instead, ultrahigh numerical aperture objectives have been produced (1.45–1.6 N). Light rays focus at the back aperture plane of the objective that are sufficiently off axis will emerge collimated, but at an oblique angle. If a specimen grown on a high index coverglass is placed upon the objective, total internal reflection can occur at the specimen-coverglass interface if the oblique angle is sufficiently large. This approach has been described as the objective launch TIRM and has been very successful in the study of exocytosis processes (23) (Fig. 11).

FLUORESCENT PROBES

Fluorescence microscopy has found many applications in biomedicine. This wide acceptance is a direct result of the availability of an ever growing set of fluorescence probes designed to measure cell and tissue structure, metabolism, signaling processes, gene expression, and protein distribution (25,26). The synthesis of fluorescent probes dates back to 1856, when William Perkin made the first synthetic

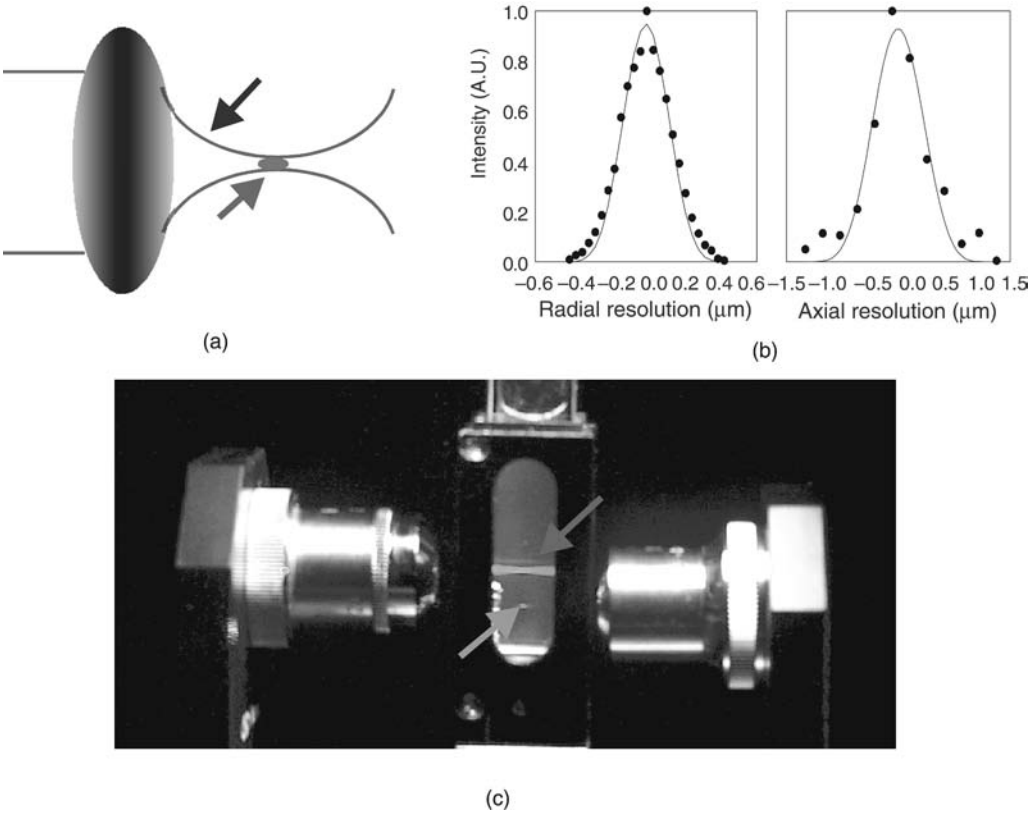


Figure 8. Two-photon microscopy optically sections and produces a fluorescent signal originating only from the focal point (a) the geometry of two-photon fluorescence. In traditional one-photon excitation, fluorescence is generated throughout the double inverted cones (blue arrow). Two-photon excitation generates fluorescence only at the focal point (red arrow). (b) The submicron PSF of two-photon excitation at 960 nm: The full-widths at half maximum (fwhm) are 0.3 μm radially and 0.9 μm axially. (c) An experimental visualization of the small excitation volume of two-photon fluorescence. One- and two-photon excitation beams are focused by two objectives (equal numerical aperture) onto a fluorescein solution. Fluorescence is generated all along the path in the one-photon excitation case (blue arrow), whereas fluorescence is generated only in a 3D confined focal spot for two-photon excitation (red arrow) The reduced excitation volume is thought to lead to less photodamage. (Please see online version for color figure)

probe from coal tar dye. Thereafter, many more synthetic dyes became available: pararosaniline, methyl violet, malachite green, safranin O, methylene blue, and numerous azo dyes. While most of these early dyes are weakly fluorescent, more fluorescent ones based on the xanthene and acridine heterocyclic ring systems soon became available.

Optical Factors in the Selection of Fluorescent Probes

Before providing a survey of the wide variety of fluorescent probes, it is important to first discuss the optical properties of fluorescent probes that are important for microscopic imaging: extinction coefficient, quantum yield, fluorescent lifetime, photobleaching rate, and spectral characteristics.

One of the most important characteristic of a fluorescent probe is its extinction coefficient. Extinction coefficient, ϵ, measures the absorption probability of the excitation light by the fluorophore. Consider excitation light is transmitted through a solution containing fluorophore at concentration c with a path length l. The light intensities before and after the solution are I_0 and I. The extinction coefficient can then be defined by Beer's law:

$$\log_{10}\frac{I_0}{I} = \varepsilon c l \tag{18}$$

Fluorescent probes with high extinction coefficients can be excited by lower incident light intensity allowing the use of lowest cost light sources and reducing the background noise of the images originated from scattered excitation light.

Quantum yield, Q, measures the relative contributions of the radiative versus nonradiative decay pathways. High quantum efficiency maximizes the fluorescent signal for each photon absorbed. The combination of probe extinction coefficient and quantum efficiency quantifies the total conversion efficiency of excitation light into fluorescent signal.

While ϵ and Q determines excitation light conversion efficiency, the maximum rate of fluorescent photon generation also depends on the lifetime, τ, of the probe. Since a molecule that has been excited cannot be reexcited until it returns to the ground state, fluorescent lifetime

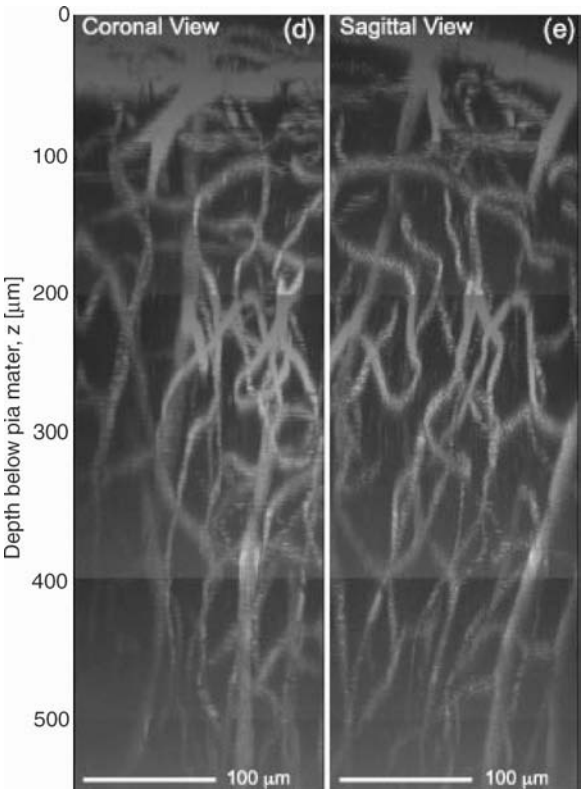


Figure 9. Fluorescein dextran labeled blood vessels in the primary vibrissa cortex of a living rat brain imaged using two-photon microscope down to a depth of $>500\,\mu m$, which demonstrates the ability of this technique to image deep into tissue (21).

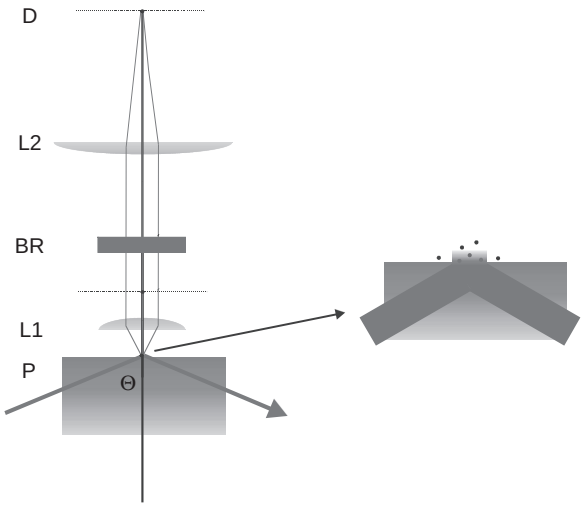


Figure 10. The configuration of a total internal reflection fluorescence microscope. L1 and L2 are objective and tube lens, respectively. The barrier filter is BR and the detector is D. The prism is P. The excitation light (green) is incident up the prism at angle, θ, greater than the critical angle. The excitation light is totally internally reflected from the surface. A magnified view is shown on the left. The evanescence electric field induced by the excitation light above the prism surface decays exponentially and only induces strong fluorescence signal for probes close to the surface of the prism. Please see online version for color figure.

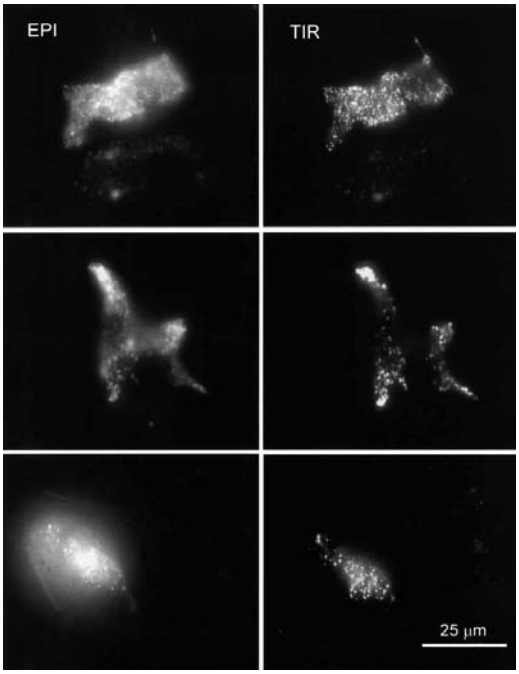


Figure 11. Epi-illuminated wide field (EPI) and total internal reflection (TIR) microscopy of bovine chromaffin cells containing secretory granules marked with GFP atrial naturetic protein (23). Only the lower plane of the cells contributes to the fluorescence signal in TIR set-up.

determines the rate at which a single probe molecule can be recycled. In general, for fluorescent probes with equal ε and Q, fluorescent photon production rate is an inverse function of probe lifetime. Further intersystem cross-rate also plays a role in determining photon generation rate. Since the triplet state has a very long lifetime, probes with high intersystem cross-rates are trapped in the triplet state with a relatively lower photon generation rate.

Photobleaching rate measures the probability that a probe will undergo an excited-state chemical reaction and become nonfluorescent irreversibly. Therefore, the photobleaching rate of a probe limits the maximum number of fluorescence photons that can be produced by a single fluorophore. Photobleaching rates of fluorophores vary greatly. For example, rhodamine can survive up to 100,000 excitation, fluorescein a few thousand, and tryptophan can only sustain a few excitation events. Photobleaching can be caused by a variety of processes. Generally, it is the result of a photochemical reaction in the excited state of the probe. For example, a common bleaching pathway is the generation of a triplet state that reacts with oxygen dissolved in solution to generate singlet oxygen and an oxidized molecule incapable of undergoing the same electronic transition as before.

Spectral properties are also important in probe selection for a number of reasons. First, selecting fluorescent probes with well-separated excitation and emission spectra allow more efficient separation of the fluorescence signal from the excitation light background. Second, fluorescent probes should be selected to match the detector used in the microscope that may have very different sensitivity across the

spectral range. For example, most PMTs have maximum efficiency in the green spectral range, but very low efficiency in the red. Therefore, green emitting probes are often better matches for microscopes using PMTs as detectors. Third, probes with narrow emission spectra allow a specimen to be simultaneously labeled with different colors providing a method to analyze multiple biochemical components simultaneously in the specimen.

Classification of Fluorescent Probes

There is no completely concise and definitive way to classify the great variety of fluorescent probes. A classification can be made based on how the fluorophores are deployed in biomedical microscopy: intrinsic probes, extrinsic probes, and genetic expressible probes.

Intrinsic Probes. Intrinsic probes refer to the class of endogenous fluorophores found in cells and tissues. Many biological components, deoxyribonuclic acid such as (DNA), proteins, and lipid membrane are weakly fluorescent. For example, protein fluorescence is due to the presence of amino acids: tryptophan, tyrosine, and phenylalanine. Among them, tryptophan is the only member with marginal quantum yield for microscopic imaging. However, fluorescent imaging based on tryptophan provides very limited information due to the prevalence of this amino acid in many proteins distributed throughout cellular systems and provides no specificity or contrast. The most useful intrinsic probes for microscopy imaging are a number of enzymes and proteins, such as reduced pyridine nucleotides [NAD(P)H], flavoproteins, and protoporphyrin IX. Both NAD(P)H and favoproteins are present in the cellular redox pathway. The NAD(P)H becomes highly fluorescent when reduced, whereas flavoprotein becomes fluorescent when oxidized, while their redox counterparts are nonfluorescent. These enzymes thus provide a powerful method to monitor cell and tissue metabolism. Protoporphyrin IX (PPIX) is a natural byproduct in the heme production pathway that is highly fluorescent. Certain cancer cells have been shown to have upregulate PPIX production relative to normal tissue and may be useful in the optical detection of cancer. Another class of important intrinsic fluorophores includes elastin and collagen, which resides in the extracellular matrix allowing structural imaging of tissues. Finally, natural pigment molecules, such as lipofuscin and melanin, are also fluorescent and have been used in assaying aging in the ocular system and malignancy in the dermal system respectively.

Extrinsic Probes. Many extrinsic fluorescent probes have been created over the last century. A majority of these extrinsic fluorophores are small aromatic organic molecules (25–28). Many probe families, such as xanthenes, canines, Alexas, coumarines, and acrinides have been created. These probes are designed to span the emission spectrum from near UV to the near-IR range with optimized optical properties. Since these molecules have no intrinsic biological activity, they must be conjugated to biological molecules of interest, which may be proteins or structure components, such as lipid molecules. Most common linkages are through reactions to amine and thiol residues. Reactions to amine are based on acylating reactions to form carboxamides, sulfonamides, or thioureas. Targeting thiol residue in the cysteines of proteins can be accomplished via iodoacetamides or maleimides. Other approaches to conjugate fluorophores to biological components may be based on general purpose linker molecules, such as biotin-avidin pairs or based on photoactivable linkers. A particularly important class of fluorophores conjugated proteins is fluorescent antibodies that allow biologically specific labeling.

In addition to maximizing the fluorescent signal, the greater challenge in the design of small molecular probes is to provide environmental sensitivity. An important class of environmentally sensitive probes distinguishes the hydrophilic versus hydrophobic environment and results in a significant quantum yield change or spectral shift based on solvent interaction. This class of probes includes DAPI, laurdan, and ANS, which have been used to specifically label DNA, measure membrane fluidity, and sense protein folding states, respectively. Another important class of environmentally sensitive probes senses intracellular ion concentrations, such as pH, Ca^{2+}, Mg^{2+}, Zn^{2+}. The most important members of this class of probes are calcium concentration sensitive because of the importance of calcium as a secondary messenger. Changes in intracellular calcium levels have been measured by using probes that either show an intensity or a spectral response upon calcium binding. These probes are predominantly analogues of calcium chelators. Members of the Fluo-3 series and Rhod-2 series allow fast measurement of the calcium level based upon intensity changes. More quantitative measurement can be based on the Fura-1 and Indo-1 series that are ratiometric. These probes exhibit a shift in the excitation or emission spectrum with the formation of isosbestic points upon calcium binding. The intensity ratio between the emission maxima and the isosbestic point allows a quantitative measurement of calcium concentration without influence from the differential partitioning of the dyes into cells.

Quantum dots belong to a new group of extrinsic probes that are rapidly gaining acceptance for biomedical imaging due to a number of their very unique characteristics (29–31). Quantum dots are semiconductor nanoparticles in the size range of 2–6 nm. Photon absorption in the semiconductor results in the formation of an exciton (an electron-hole pair). Semiconductor nanoparticles with diameters below the Bohr radius exhibit strong quantum confinement effect, which results in the quantization of their electronic energy level. The quantization level is related to particle size where smaller particles have a larger energy gap. The radiative recombination of the exciton results in the emission of a fluorescence photon with energy corresponding to the exciton's quantized energy levels. The lifetime for the recombination of the exciton is long, typically on the order of a few tens of nanoseconds. Quantum dots have been fabricated from II–VI (e.g., as CdSe, CdTe, CdS, and ZnSe) and III–V (e.g., InP and InAs) semiconductors. Due to the compounds involved in the formation of these fluorescent labels, toxicity studies have to be realized prior to any experiments. Recent research works have been devoted

to the better manufacture of these semiconductor crystals including methods to form a uniform crystalline core and to produce a surface capping layer that enhances the biocompability of these compounds, prevents their aggregation, and can maximize their quantum efficiency. Furthermore, coating the surface of quantum dots with convenient functional groups, including common linkages, such as silane or biotin, has been accomplished to facilitate linkage to the biological molecules. Quantum dots are unique in their broad absorption spectra, very narrow (~15 nm) emission spectrum, and extraordinary photostability. In fact, quantum dots have been shown to have photobleaching rates orders of magnitude below that of organic dyes. Quantum dots also have excellent extinction coefficients and quantum yield. While there are significant advantages in using quantum dots, they also have a number of limitations including their relative larger size compared with organic dyes and their lower fluorescence flux due to their long lifetime. Quantum dots have been applied for single receptor tracking on cell surface and for the visualization of tissue structures, such as blood vessels.

Genetic Expressible Probes. The development of genetically expressible probes has been rapid over the last decade (32). The most notable of these genetic probes is green fluorescent protein, GFP (33). The GFP was isolated and purified from the bioluminescent jellyfish *Aequorea Victoria*. Fusion proteins can be created by inserting GFP genes into an expression vector that carries a gene coding for a protein of interest. This provides a completely noninvasive procedure and perfectly molecular specific approach to track the expression, distribution, and trafficking of specific proteins in cells and tissues. In order to better understand protein signaling processes and protein–protein interactions, fluorescent proteins of different colors have been created based on random mutation processes. Today, fluorescent proteins with emission spanning the spectral range from blue to red are readily available. Expressible fluorescent proteins that are sensitive to cellular biochemical environment, such as pH and calcium, have also been developed. Novel fluorescent proteins with optically controllable fluorescent properties, such as photoactivatable fluorescent proteins, PA-GFP, photoswitchable CFP, and pKindling red have been created and may be used in tracing cell movement or protein transport. Finally, protein–protein interactions have been detected based on a novel fluorescent protein approach in which each of the interacting protein pairs carries one-half of a fluorescent protein structure that is not fluorescent. Upon binding of the protein pairs, the two halves of the fluorescent protein also recombine, which results in a fluorescent signal.

ADVANCED FUNCTIONAL IMAGING MODALITIES AND THEIR APPLICATIONS

A number of functional imaging modalities based on fluorescent microscopy have been developed. These techniques are extremely versatile and have found applications ranging from single molecular studies to tissue level experiments. The implementation of the most common imaging modalities will be discussed with representative examples from the literature.

Intensity Measurements

The most basic application of fluorescence microscopy consists in mapping fluorophore distribution based on their emission intensity as a function of position. However, this map is not static. Measuring intensity distribution as a function of time allows one to follow the evolution of biological processes. The fastest wide-field detectors can have a frame rate in the tens of kilohertz range, unfortunately at the expense of sensitivity and spatial resolution. They are used to study extremely fast dynamics, such as membrane potential imaging in neurons. For 3D imaging, point scanning techniques are typically slower than wide-field imaging, but can reach video rate speed using multifoci illumination.

Dynamic intensity imaging has been used at the tissue level to follow cancer cells as they flow through blood vessels and extravasate to form metastases, or in embryos to track the expression of a regulatory protein at different developmental stages. One commonly used technique to follow the movements of protein in cellular systems is fluorescent recovery after photobleaching (FRAP). In FRAP studies, a small area of a cell expressing a fluorescently labeled protein is subjected to an intense illumination that photobleaches the dye and leads to a drastic drop in fluorescence intensity. The rate at which the intensity recovers provides a measure of the mobility of the protein of interest.

An important aspect of the fluorescent microscopy technique lies also in the image analysis. Particle tracking experiments are an excellent example. Zuhang and coworkers (34) studied the infection pathway of the influenza virus labeled with a lipophilic dye in CHO cells. Each frame of the movie recorded was analyzed to extract the position of the virus particles with 40 nm accuracy. Three different stages of transport after endocytosis of the virus particle were separated, each involving different transport mechanisms transduced by a different protein as shown on Fig. 12. The first stage is dependant on actin and results in an average transport distance of 2 μm from the initial site of binding at the cell periphery. The second stage is characterized by a sudden directed displacement that brings the virus close to the nucleus with a speed of 1–4 $\mu m\, s^{-1}$ that is consistent with the velocity of dynein motors on microtublues. The last stage consists of back and forth motion in the perinuclear region. This is followed by the fusion of the endosome with the virus and the liberation of the genetic material. This event is identified by a sudden increase in the fluorescence intensity due to the dequenching of the fluorescent tags on the virus.

Spectral Measurements

An extremely important feature of fluorescent microscopy is the ability to image many different fluorescent species based on their distinct emission spectra. Dichroic bandpass filters optimized for the dyes used in the experiment can discriminate efficiently between up to four or five different fluorophores.

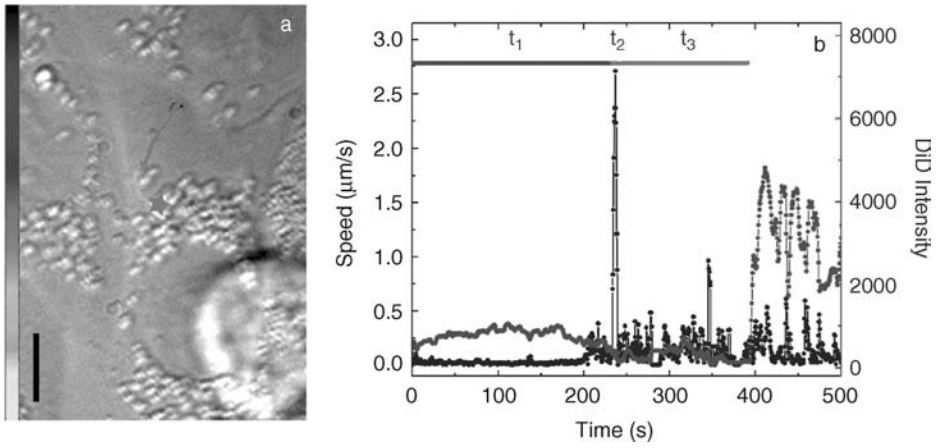


Figure 12. Particle tracking of virus infecting a cell. (a) Trajectory of the virus. The color of the trajectory codes time from 0 s (black) to 500 s (yellow). The star indicates the fusion site of the virus membrane with the vesicle. (b) Time trajectories of the velocity (black) and fluorescence (blue) of the virus particle (34). Please see online version for color figure.

In a study of connexin trafficking, Ellisman and coworkers (35) used a dual labeling scheme to highlight the dynamics of these proteins. Using a recombinant protein fused to a tetracystein receptor domain, the connexin was stably labeled with a biarsenical derivate of fluorescein or resorufin (a red fluorophore). The cells expressing these modified proteins were first stained with the green fluorophore and incubated 4–8 h. The proteins produced during this incubation period are fluorescently tagged in a second staining step with the red fluorophore. The two-color images highlight the dynamics of the connexin refurbishing at the gap junction. As shown on Fig. 13, the older proteins are found in the center and are surrounded by the newer proteins.

For wide-field fluorescence imaging using a CCD camera, spectral information is collected sequentially while position information is collected at once. Bandpass filters can be inserted to select the emission wavelength in between image frames. This procedure is relatively slow and can result in image misregistration due to the slight misalignment of the filters. This problem can be overcome by the use of electronically tunable filter. Two types of electronically tunable filters are available based either on liquid-crystal technology or on electrooptical crystals. Liquid-crystal tunable filters are made of stacks of birefringent liquid-crystal layers sandwiched between polarizing filters. Polarized light is incident upon the device. The application of a voltage on the liquid-crystal layer produces a wavelength dependent rotation of the polarization of light as the light is transmitted through the liquid-crystal layers. After cumulative rotations through the multiple layers, only the light at a specific spectral range is at the correct polarization to pass through the final polarizer without attenuation. The second type is acoustooptics tunable filters (AOTFs). An AOTF works by setting acoustic vibration at radio frequency (rf) through an electrooptical crystal to create a diffraction grating that singles out the appropriate wavelength with a few nanometer bandwidth. The main advantage of AOTF is that the wavelength selection is realized by tuning the acoustic wave frequency, which can be done in a fraction of a millisecond while the liquid-crystal tunable filters operate with a time constant of hundreds of milliseconds. The latter, however, have a larger clear aperture and selectable bandwidth ranging from a fraction of a nanometer up to tens of a nanometer. Liquid-crystal filters are more often used for emission filtering while the acoustooptic filters are more commonly used for excitation wavelength selection.

Typical emission spectra from molecular fluorophores have a sharp edge at the blue end of the spectrum, but have a long tail extending far into the red due to electronic relaxation from the excited state into a vibrationally excited ground state. When imaging with a few color channels, where each channel represents a single chromophore, one has to be careful to take into account the spectral bleedthrough of each dye into the neighboring channels. Collecting signal in a larger number of channels allows the use of a linear unmixing technique to account for the real shape of the emission spectra of each dye and accounts more precisely for their contributions in each pixel of the image. This technique can be implemented using tunable filters with a narrow bandwidth and CCD camera detectors. It has also been shown that an interferometer can be used to encode the spectral information in the image on the CCD camera. An image is then recorded for each step of the interferometer and a Fourier transform analysis allows the recovery of the spectral information. Although it requires more advanced postprocessing of the image data, this approach offers a large spectral range and a variable spectral resolution unmatched by the tunable filters.

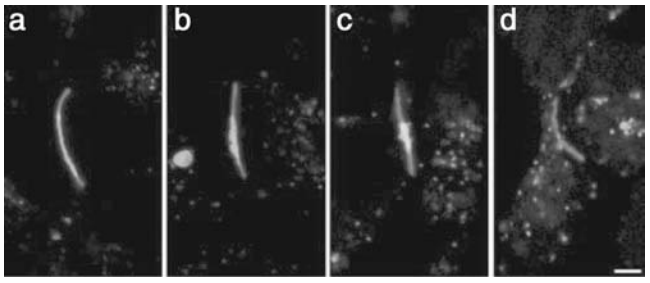


Figure 13. Connexin trafficking at gap junction. The newly produced proteins are labeled in red after 4 h (a and b) or 8 h (c and d) hours after the first staining step with green. The older proteins occupy the periphery of the gap junction, while the new ones are localized in its center (35). Please see online version for color figure.

In scanning systems, such as confocal microscopes, the use of dichroic beamsplitters can be readily constructed to simultaneously resolve two or three spectral channels in parallel at each scan position. If more spectral channels are desired for spectral decomposition measurement, the emission can be resolved in a multichannel detector using a grating or a prism to separate the different wavelength components. This has been used to separate the contribution of dyes with very similar emission spectra like GFP and fluorescein, or resolve the different intrinsic fluorophores contained in the skin where many fluorophores with overlapping spectra are present.

A particularly promising class of probes for spectral imaging are quantum dots. As discussed previously, the emission spectra of quantum dots are very narrow and can be tuned by changing their size. Further, all quantum dots have a very broad excitation spectrum and a single excitation wavelength larger than the band gap energy can lead to the efficient excitation of many different colored quantum dots simultaneously. In their report, Simon and coworkers (36) used these particles to track metastatic cells injected in the tail of a mouse as they extravasate into lung tissue. Using spectrally resolved measurements, they demonstrate their ability to recognize at least five different cell populations each labeled with different quantum dots. Figure 14 shows an image of cells labeled with different quantum dots and the emission spectra from each of these particles. The difference in emission spectra allows an easy identification of each cell population.

Lifetime Resolved Microscopy

Measurement of the fluorescence lifetime in a microscope provides another type of contrast mechanism and can be used to discriminate dyes emitting in the same wavelength range. It is also commonly used to monitor changes in the local environment of a probe measuring the pH or the concentration of cations *In situ*. The fluorescence lifetime can be shortened by interaction of the probe with a quencher, such as oxygen. Another type of quenching is induced by the presence of the transition dipole of other dyes, which are in close vicinity and lifetime measurements can be used to quantify energy-transfer processes (discussed further in a later section).

There are two methods to measure the fluorescence lifetime in a microscope. One is in the time domain and the other is in the frequency domain. In the time domain, a light pulse of short duration excites the sample and the decay of the emission is timed. The resulting intensity distribution is a convolution between the instrument response and the exponential decay of the fluorophore.

$$I(t) = I_0 \int_0^t G(t - T) \cdot \exp\left(\frac{T}{\tau}\right) dT \tag{19}$$

In the frequency domain, the excitation light is modulated at frequency ω. The intrinsic response time of the fluorescence acts as a low pass filter and the emitted signal is phase shifted and demodulated. Both the demodulation and the phase shift can be linked to the fluorescence lifetime.

$$\Delta\phi = a \tan(\omega\tau) \tag{20}$$

$$M = \frac{1}{\sqrt{1 + \omega^2\tau^2}} \tag{21}$$

In order to obtain a measurable phase shift and modulation, the frequency has to be on the same order of magnitude as the lifetime (i.e., 10^8 Hz). However, it is difficult to measure these two parameters at such high frequencies. Therefore, one typically uses a heterodyne detection to lower the frequency to the kilohertz range by modulating the detector at a frequency close to the excitation frequency.

For wide-field microscopy, an image intensifier is placed in front of the CCD camera to modulate the gain of detection. In the time domain, a short time gate is generated to collect the emission at various times after the excitation. In the frequency domain, the image intensifier is modulated at high frequencies and a series of images at different phases are acquired. In laser scanning confocal and multiphoton microscopes, time correlated single-photon counting is the method of choice for lifetime measurements in the time domain because it offers an excellent signal/noise ratio at low light levels. Every time a photon is detected, the time elapsed since the excitation of the sample is measured. A histogram of all the times of arrival yields a decay curve of the fluorescence in each pixel of the image. For brighter

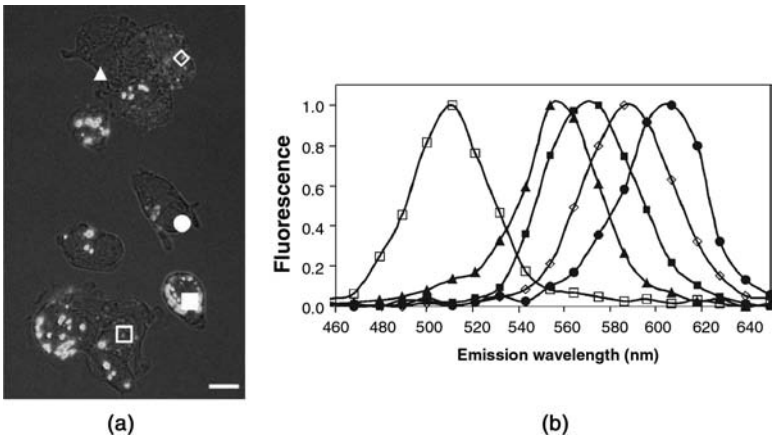


Figure 14. Spectral imaging of cells labeled by quantum dots. Cells were labeled with five different quantum dots and imaged in a multiphoton microscope. Each symbol represents a different quantum dot. The symbols on the image match the emission spectra seen on the graph. The spectral imaging set-up allows to discriminate between the different cell populations (36).

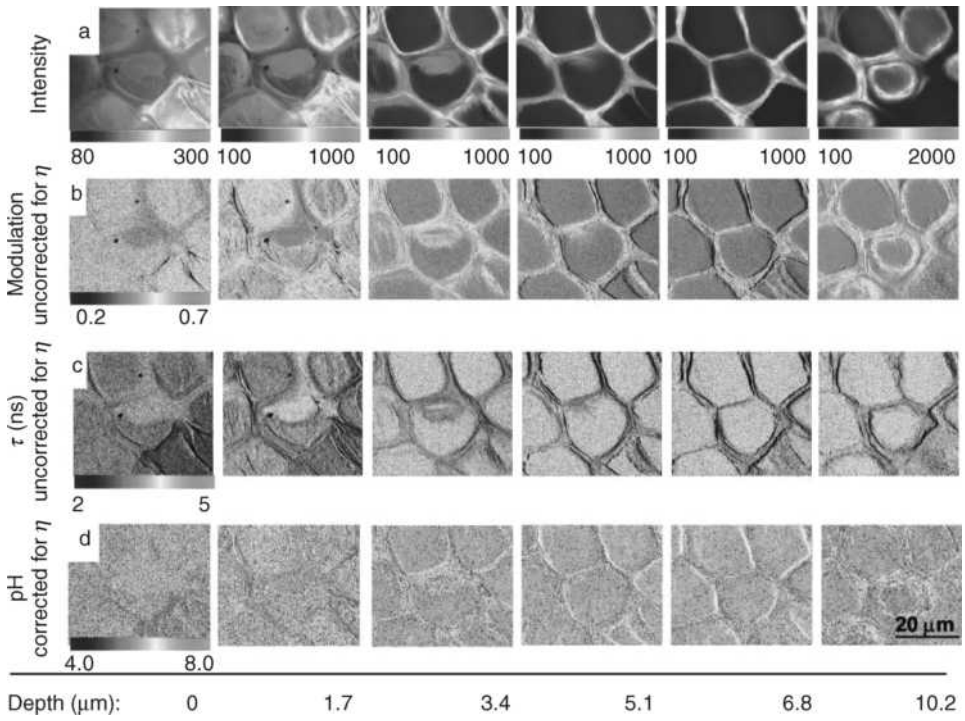


Figure 15. Quantification of the pH of the skin by lifetime imaging. (a) Intensity, (b) modulation, (c) lifetime, and (d) pH maps of a mouse skin at different depth. The lifetime measurements allow a determination of the pH independently of the intensity changes recorded between different imaging depth (37).

samples, a frequency domain approach using modulated detectors can also be used to measure the lifetime.

To measure the pH in the skin of a mouse, Clegg and co-workers (37) used a modified fluorescein probe whose lifetime varies from 2.75 ns at pH 4.5 to 3.9 ns at pH 8.5. As they image deeper in the skin, they observe that the average pH increases from 6.4 at the surface up to >7 at 15 μm depth. The extracellular space is mostly acidic (pH 6), while the intracellular space is at neutral pH. Typically, pH is measured in solution by correlating fluorescence intensities with specific pH levels. This approach is not suitable for tissues, as in the skin, since the dye is unevenly distributed throughout the tissue (Fig. 15) due to differential partitioning. A measurement of pH based on fluorescence lifetime is not dependent on probe concentration and thus the pH can be measured in the intra and extracellular space at various depths in the skin.

Polarization Microscopy

Polarization microscopy is a technique that provides information about the orientation or the rotation of fluorophores. Linearly polarized excitation light results in preferential excitation of molecules with their transition dipole aligned along the polarization. If the molecule is in a rigid environment, the emitted fluorescence will mostly retain a polarization parallel to the excitation light. However, if the molecule has time to rotate before it emits a photon, this will randomize the emission polarization. The anisotropy r is a ratio calculated from the intensity parallel $I_{\parallel}$ and perpendicular $I_{\perp}$ to the incident polarization and is a measure of the ability of the molecule to rotate.

$$r = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + 2I_{\perp}} \tag{22}$$

This ratio is mostly governed by two factors, which are the fluorescence lifetime τ and the rotational correlation time θ.

$$r = \frac{r_0}{1 + (\tau/\theta)} \tag{23}$$

where r_0 is the fundamental anisotropy. Molecules with a short fluorescence lifetime and a long rotational correlation time ($\tau < \theta$) will have a high anisotropy. In the opposite case, where molecules can freely rotate during the time they reside in the excited state, the anisotropy will be low. An approximate measurement of the mobility of a molecule can be obtained by exciting the sample at different polarization angles. A proper measurement of the anisotropy requires both a linearly polarized excitation light source and the detection of the parallel and perpendicular component of the fluorescence light using a polarizer. This technique has been used to measure viscosity and membrane fluidity *In vivo*. It has been applied to quantify enzyme kinetics, relying on the fact that the cleavage of a fluorescently labeled substrate leads to a faster tumbling and thus a decrease in anisotropy.

Goldstein and co-workers (38) used polarization microscopy at the single-molecule level to study the orientation of the kinesin motor on a microtubule. A thiol reactive rhodamine dye was attached to cysteines on the motor protein. Microtubules decorated with the modified kinesin were imaged under a different polarization angle. In the presence of adenosine monophosphate (AMP)–(PNP) [a nonhydrolyable analogue of adenonine triphosphats (ATP)], the fluorescence intensity depends strongly on the angle of polarization of the excitation light (Fig. 16) proving that the kinesin maintains a fixed orientation. In the presence of adenonine5–diphosphate (ADP), however, the anisotropy is lower (no dependence on excitation polarization angle),

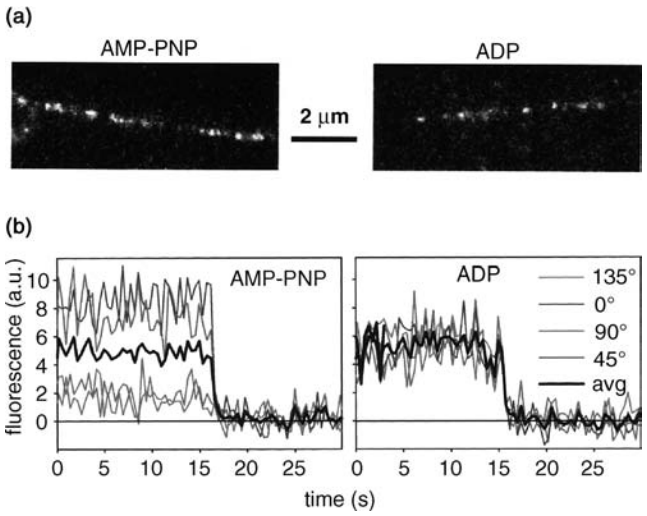


Figure 16. Mobility of single kinesin motors on microtubules probed by polarization microscopy. (a) Image of microtubules sparsely decorated with kinesin motors in presence of AMP–PNP and ADP. (b) Time course of the fluorescent intensity recorded from single molecule excited with linearly polarized light at four different angles. The large fluctuations of the fluorescence intensity as function of the excitation polarization in the AMP–PNP case demonstrate the rigidity of the kinesin motor on the microtubule (38).

leading to the conclusion that the kinesin is very mobile, while still attached to the microtubule.

Fluorescence Resonance Energy Transfer

Förster resonance energy transfer (FRET) is a technique used to monitor interaction between two fluorophores on the nanometer scale. When a dye is promoted to its excited state, it can transfer this electronic excitation by dipole–dipole interaction to a nearby molecule. Due to the nature of this interaction, Förster predicted a dependence of the FRET efficiency on the sixth power of distance that was demonstrated experimentally by Stryer with a linear polypeptide of varying length (39). The efficiency E of the process varies as function of the distance, R, between the two molecules as

$$E = \frac{R_0^6}{R_0^6 + R^6} \tag{24}$$

Where R_0 is called the Förster distance, which depends on Avogadro's number N_A, the index of refraction of the medium n, the quantum yield of the donor molecule Q_D, the orientation factor κ, and the overlap integral J.

$$R_0^6 = \frac{9000 \ln(10) \kappa^2 Q_D}{128 \pi^5 N_A \, n^4} J \tag{25}$$

κ represents the relative orientation of the transition dipole of the donor and acceptor molecules. In most cases, a random interaction is presumed and this factor is set to two-thirds. The overlap integral J represents the energy overlap between the emission of the donor and the absorption of the acceptor. For well-matched fluorophore pairs, R_0 is on the order of 4–7 nm.

Most FRET experiments are based on the measurement of the intensity of the donor and of the acceptor because the presence of FRET in a system is characterized by a decrease in the emission of the donor and an increase in the acceptor signal. Thus, in principle, a two color channel microscope is sufficient to follow these changes. However, experimental artifacts, such as concentration fluctuation and spectral bleed, complicate the analysis of these images and many different correction algorithms have been developed.

FRET measurements have been used in molecular studies to measure distances and observe dynamic conformational changes in proteins and ribonuclic acid (RNA). In cellular studies, FRET is often used to map protein interactions. By labeling one protein with a donor dye and its ligand with an acceptor dye, energy transfer will occur only when the two proteins are bound such that the dyes come in close proximity of each other.

The addition of fluorescence lifetime imaging provides the additional capability of retrieving the proportion of fluorophore undergoing FRET in each pixel of an image independently of concentration variations. This is possible because the fluorescence lifetime of a FRET construct is shorter than the natural decay of the dye. Thus if one has a mixture of interacting and free protein, fitting a double exponential to the fluorescence decay allows us to retrieve the proportion of interacting protein. This has been applied by Bastiaens and co-workers (40) to the study the phosphorylation of the EGF receptor ErB1. The transmembrane receptor is fused with a GFP protein. The phosphrylation of the protein is sensed by an antibody labeled with a red dye (Cy3). When the Erb1 is phosphorylated, the antibody binds to the protein and FRET occurs due to the short distance between the antibody and the GFP. The ErB1 receptors can be stimulated by EGF coated beads leading to phosphorylation and FRET. The time course of the stimulation is followed and for each cell and the fraction of phosphorylated receptors at various time interval is shown in Fig. 17. After 30 s, the FRET events are localized at discrete locations. But after 1 min, the whole periphery of the cell displays high FRET, demonstrating the lateral signaling of the receptor after activation at discrete location by EGF coated beads.

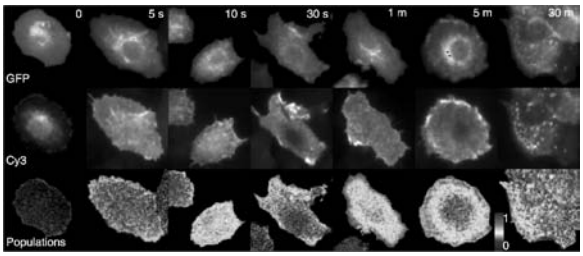


Figure 17. Time course of the phosphorylation of EGF receptors ERB1 after stimulation by EGF coated beads observed by FRET between a GFP modified EGF receptor and a phosphorylation antibody labeled with Cy3. While the GFP intensity is remains relatively constant, the concentration of the Cy3 tagged antibody clearly increases after the stimulation. This leads to an increase FRET signal as function of time (40).

CONCLUSION

The utility of fluorescence microscopy lies in its ability to study biological structure and function *In vivo*. The exquisite sensitivity and image contrast of fluorescence microscopy allow biological structures to be imaged on the submicron length scale. The greatest power of fluorescence microscopy lies in its ability to determine biochemical functions using assays based on fluorescence spectroscopy. With the availability of more versatile instruments, more fluorophores unit greater molecular and environmental specificity, the impact of fluorescence microscopy technology on biomedical science will only increase.

BIBLIOGRAPHY

1. Wang XF, Herman B. Fluorescence Imaging Spectroscopy and Microscopy. New York: Wiley; 1996.
2. Inoué S, Spring KR. Video Microscopy: the Fundamentals. New York: Plenum Press; 1997.
3. Herman B. Fluorescence Microscopy. Oxford: Bios Scientific Publishers / Springer in Association with the Royal Microscopy Society; 1998.
4. Cantor CR, Schimmel PR. Biophysical Chemistry. San Francisco: Freeman; 1980.
5. Valeur B. Molecular Fluorescence: Principles and Applications. Weinheim; New York: Wiley-VCH; 2002.
6. Lakowicz JR. NetLibrary Inc. Topics in Fluorescence Spectroscopy. Kluwer Academic; 2002.
7. Klein MV, Furtak TE. Optics. New York: Wiley; 1986.
8. Hecht E. Optics. Reading, (MA): Addison-Wesley; 2002.
9. Gu M. Advanced Optical Imaging Theory. Berlin: New York: Springer; 2000.
10. Born M,Wolf E. Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light. Cambridge: New York: Cambridge University Press; 1999.
11. Rizzuto R, Carrington W, Tuft RA. Digital imaging microscopy of living cells. Trends Cell Biol 1998;8:288–292.
12. Agard DA, Hiraoka Y, Shaw P, Sedat JW. Fluorescence microscopy in three dimensions. Methods Cell Biol 1989;30:353–377.
13. Carrington WA, et al. Superresolution three-dimensional images of fluorescence in cells with minimal light exposure. Science 1995;268:1483–1487.
14. Krishnamurthi V, Liu YH, Bhattacharyya S, Turner JN, Holmes TJ. Blind Deconvolution Of Fluorescence Micrographs By Maximum-Likelihood-Estimation. Appl Opt 1995;34:6633–6647.
15. Wilson T, Sheppard CJR. Theory and Practice of Scanning Optical Microscopy. New York: Academic; 1984.
16. Pawley JB. Handbook of Confocal Microscopy. New York: Plenum; 1995.
17. Gu M. Principles of Three-Dimensional Imaging in Confocal Microscopy. Singapore: World Scientific; 1996.
18. Masters BR. Selected Papers on Multiphoton Excitation Microscopy. Bellingham: SPIE Optical Engineering Press; 2003.
19. Amos WB, White JG. How the confocal laser scanning microscope entered biological research. Biol Cell 2003;95:335–342.
20. Denk W, Strickler JH, Webb WW. Two-photon laser scanning fluorescence microscopy. Science 1990;248:73–76.
21. Kleinfeld D, Mitra PP, Helmchen F, Denk W. Fluctuations and stimulusinduced changes in blood flow observed in individual capillaries in layers 2 through 4 of rat neocortex. Proc Natl Acad Sci U. S. A. 1998;95:15741–15746.
22. Axelrod D. Total internal reflection fluorescence. Ann Rev Biophys Bioeng 1984;13:247–268.
23. Axelrod D. Total internal reflection fluorescence microscopy in cell biology. Traffic 2001;2:764–774.
24. Axelrod D. Selective imaging of surface fluorescence with very high aperture microscope objectives. J Biomed Opt 2001;6:6–13.
25. Mason WT. Fluorescent and Luminescent Probes for Biological Activity: a Practical Guide to Technology for Quantitative Real-time Analysis. San Diego: Academic; 1999.
26. Slavic J. Fluorescence Microscopy and Fluorescent Probes. New York: Plenum Press; 1996.
27. Tsien RY. Fluorescent probes of cell signaling. Annu Rev Neurosci 1989;12:227–253.
28. Tsien RY. Fluorescent indicators of ion concentrations. Methods Cell Biol 1989;30:127–156.
29. Bruchez Jr M, Moronne M, Gin P, Weiss S, Alivisatos AP. Semiconductor nanocrystals as fluorescent biological labels. Science 1998;281:2013–2016.
30. Chan WC, et al. Luminescent quantum dots for multiplexed biological detection and imaging. Curr Opin Biotechnol 2002;13:40–46.
31. Michalet X, et al. Quantum dots for live cells, in vivo imaging, and diagnostics. Science 2005;307:538–544.
32. Zhang J, Campbell RE, Ting AY, Tsien RY. Creating new fluorescent probes for cell biology. Nat Rev Mol Cell Biol 2002;3:906–918.
33. Chalfie M, et al. Green Fluorescent Protein as a Marker for Gene Expression. Science 1994;263:802–805.
34. Lakadamyali M, Rust MJ, Babcock HP, Zhuang X. Visualizing infection of individual influenza viruses. Proc Natl Acad Sci U. S. A. 2003;100:9280–9285.
35. Gaietta G, et al. Multicolor and electron microscopic imaging of connexin trafficking. Science 2002;296:503–507.
36. Voura EB, Jaiswal JK, Mattoussi H, Simon SM. Tracking metastatic tumor cell extravasation with quantum dot nanocrystals and fluorescence emissionscanning microscopy. Nat Med 2004;10:993–998.
37. Hanson KM, et al. Two-photon fluorescence lifetime imaging of the skin stratum corneum pH gradient. Biophys J 2002;83:1682–1690.
38. Sosa H, Peterman EJ, Moerner WE, Goldstein LS. ADP-induced rocking of the kinesin motor domain revealed by single-molecule fluorescence polarization microscopy. Nat Struct Biol 2001;8:540–544.
39. Stryer L, Haugland RP. Energy Transfer: a spectroscopic ruler. Proc Natl Acad Sci U. S. A. 1967;58:712–726.
40. Verveer PJ, Wouters FS, Reynolds AR, Bastiaens PIH. Quantitative imaging of lateral ErbB1 receptor signal propagation in the plasma membrane. Science 2000;290:1567–1570.

See also FLUORESCENCE MEASUREMENTS; MICROSCOPY, CONFOCAL

MICROSCOPY, SCANNING FORCE

EWA P. WOJCIKIEWICZ
KWANJ JOO KWAK
VINCENT T. MOY
University of Miami
Miami, Florida

INTRODUCTION

Recent advances in technology have allowed us to study our world at molecular, even subatomic, resolution. One of the devices in the forefront of such studies is the atomic force microscope (AFM), which is a relatively complex device

with two major applications. It can be used as an imaging device, which allows for the acquisition of atomic-level images of biological structures as well as to measure forces of interactions between two opposing surfaces down to the single-molecule level.

Imaging AFM

The AFM was originally designed as an imaging tool (1). It was modified from the design of the scanning tunneling microscope (STM). The AFM acquires topographic images by methodically scanning a sample with a flexible probe, called a cantilever, which bends according to the contours of the sample's surface. The bending of the cantilever is translated into an image map, which reveals the height differences in the surface being scanned. It is possible to image biological samples under physiological conditions as imaging can be done in both air and liquid. The resulting resolution of such maps is at the atomic level (2,3).

The imaging AFM has been used to image many biological samples ranging from genetic material to cells to bone. A few of these studies will be highlighted. One of the earliest biological materials to be imaged was DNA, which has been imaged in many forms to date, including double- and single-stranded forms as well as more complex structures. The AFM has also been used for many applications including DNA sizing, previously only achieved using gel electrophoresis, DNA mapping, hybridization studies, and examinations of protein-DNA interactions (4). AFM studies of RNA were also conducted. Unlike DNA, which mainly forms a double-helical structure, RNA has the ability to form more advanced structures that do not rely solely on Watson–Crick base pairing. One example are the so-called kissing-loop structures imaged by Hansma et al. (4) (Fig. 1). Not only was the AFM used in imaging of such structures, many of them 3D, but also played an important role in designing them (5). Unlike other imaging techniques, AFM sudies can be done under physiological conditions allowing for the imaging of biological processes. Images of transcription complexes have been obtained, for example, *E.coli* RNA polymerase in complex with DNA. These studies are the only of their kind that can answer certain specific questions as to how the RNA transcription process takes place. One is able to visualize how the DNA does not get entangled in the nascent RNA strands. Such studies detailing the structure-function relationship of the transcription process are key in furthering our understanding of gene expression (6,7).

Also, imaging of cells was conducted to examine the structure of the cellular cortex in detail. The cell cytoskeleton is known to be involved in affecting cell shape as well as movement and other cellular responses to biochemical and biophysical signals. At present, relatively little is known about the mechanical organization of cells at a subcellular level. Pesen et al. (8) studied the cell cortex of bovine pulmonary artery endothelial cells (BPAECs) using AFM and confocal fluorescence microscopy (CFM). They were able to identify a coarse and fine mesh that make up the cortical cytoskeleton. These two types of mesh appear to be intertwined (Fig. 2) (8). Such details are not distinguished in imaging studies using fixed cells.

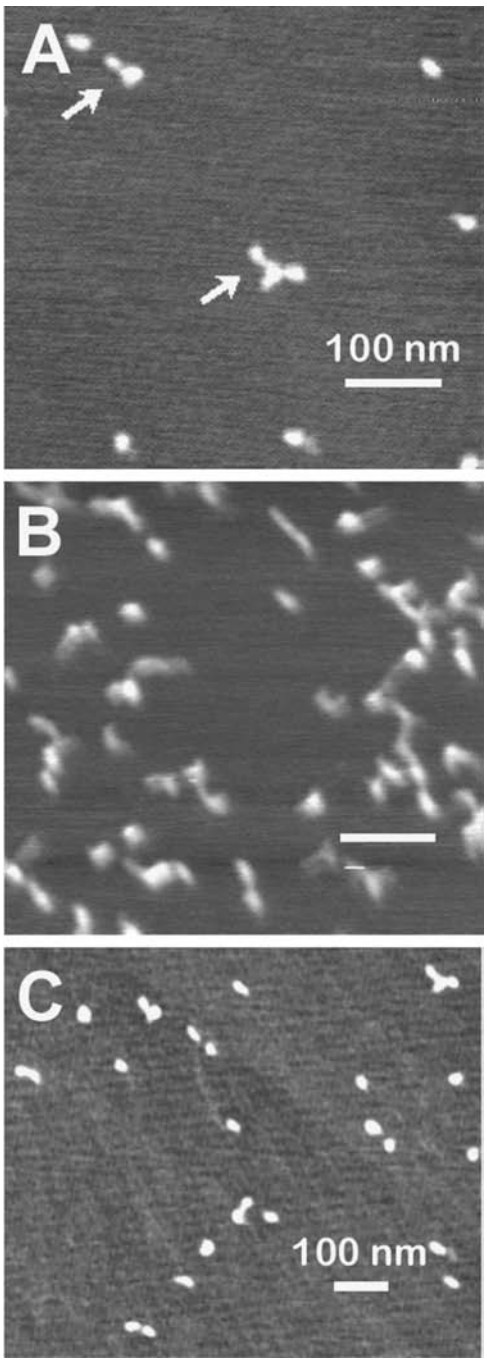


Figure 1. Kissing-loop RNA structures. (a,c) Kissing-loop RNAs at low concentrations. The three arms of the individual RNAs are visible. (B) Kissing-loop RNAs at a concentration 10-fold higher than in a and c. The individual structures are less well-defined. Scale bars = 100 nm for all images.

Other imaging studies have looked at tendon and bone, both of which are composed of type I tropo-collagen, which was done by acquisition of high resolution AFM images of type I collagen in conjunction with force spectroscopy studies, namely protein unfolding, which is described in the following section. Figure 3 reveals these high resolution collagen type I images. They were acquired using two different concentrations of collagen, which resulted in

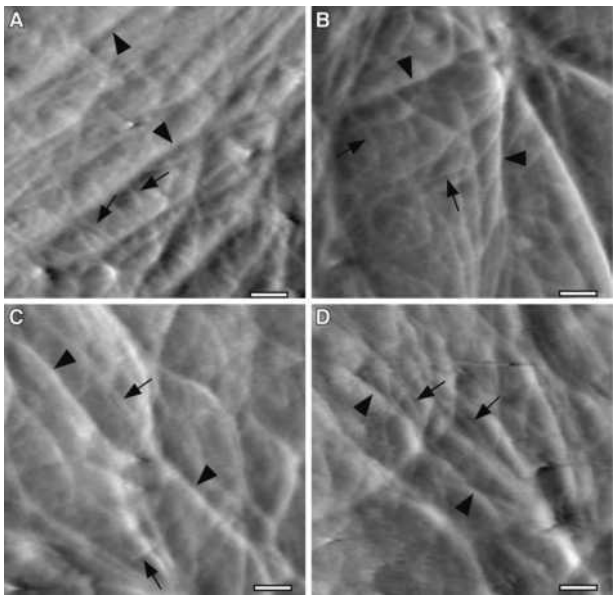


Figure 2. AFM Images of the cortical mesh of bovine pulmonary artery endothelial cells (BPAEC). (a–d) High magnification deflection AFM images of the cortical mesh of living BPAECs in a physiological saline. The filamentous mesh appears to be organized on two length scales, with coarse mesh (arrowheads) and fine mesh filaments (arrows). The two meshes are likely to be intertwined, although it is possible that the fine mesh is layered over the coarse mesh. Lateral resolution in these images is ~125 nm.

slightly different orientations of collagen: random at the lower concentration (Fig. 3a) and oriented unidirectionally in the higher concentration (Fig. 3b). In these studies, the AFM was used to investigate the mechanical properties of this collagenous tissue, which are altered in diseases such as osteoporosis. Being familiar with such properties is important for gaining further understanding as well as preventing and curing bone diseases (9).

Force Spectroscopy

The AFM can also be operated in the force scan mode, which allows for the measurement of adhesion forces between receptors and their corresponding binding partners, or ligands. In studies of ligand-receptor forces, the receptor is immobilized on the surface of a flexible AFM cantilever whereas the ligand is attached to a suitable substrate. The deflection of the cantilever during the approach and withdrawal of the cantilever from the substrate allow for the force of the interaction to be measured. These type of experiments provide information simulating the influence of internal and external forces that these receptors would experience in the body, for example, the shear stress experienced by a blood cell attached to the endothelium while blood rushes past it. Such information was previously unavailable when receptor-ligand interactions were examined using more traditional biochemical techniques. AFM has made it possible to acquire measurements that reveal the mechanical properties of biomolecules under applied force. Measurements of the unbinding force of a single ligand-receptor interaction can now be acquired with the AFM (10–12).

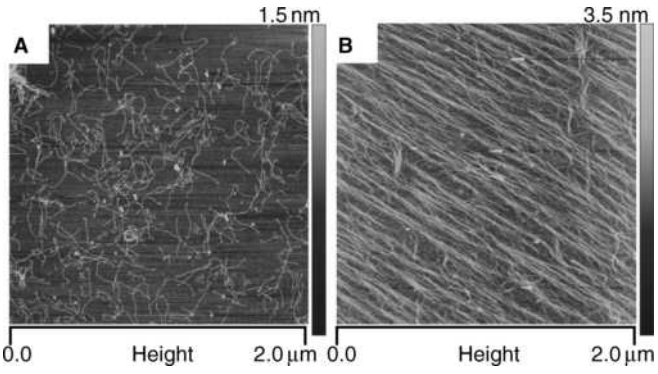


Figure 3. Topographical images (height; tapping mode in air) of type I collagen monomers on a mica substrate. (a) low (1 μg/ml) concentration of collagen, (b) high (10 μg/ml) concentration of collagen.

The AFM can also be used in adhesion studies involving whole cells (13,14). In these studies, the interaction between a cell expressing a particular receptor of interest and its ligand protein or another cell expressing the ligand is measured. The cell adhesion experiments allow for the acquisition of both single-molecule measurements, like in the above-mentioned studies, as well as multiple-bond interactions. The advantage of using the AFM in cell adhesion studies is the high specificity and wealth of information that is obtained. The AFM force scans provide information about the individual bond strengths as well as the force and work that is required to separate the entire complex. Combining single-molecule and multiple-bond data allows us to describe the thermodynamic model of the separation of a particular complex in addition to the mechanism of its action on the cellular scale (15,16).

The AFM can also serve as a microindenter that probes soft samples, including cells revealing information about their mechanical properties. The mechanical properties of cells play an important role in such essential physiological processes such as cell migration and cell division. Understanding these properties can later help scientists to identify when certain processes may be taking place. The mechanical properties of cells are chiefly determined by their actin cytoskleleton, which is the cell's "backbone." This type of information, which cannot be obtained using standard cell biology methods, allows for the estimation of the Young's modulus of living cells (16). The Young's modulus is a calculated value, which provides information regarding the compliance or elasticity of a cell. Such experiments may be done with either the imaging AFM or using force spectroscopy. Manfred Radmacher's group has conducted such measurements with an imaging AFM in force mapping mode. The advantage of such measurements is the wealth of information that they provide. These experiments reveal not only the elastic properties of the cells being examined but also topographical information. They can also be performed in conjunction with video microscopy to further confirm what one is visualizing and that cells are not undergoing damage (17,18). In force spectroscopy experiments, the Young's modulus is obtained by poking the cell cantilever tip. This type of

information can be correlated with adhesion data to determine whether elasticity changes in response to drugs or other stimulation have an effect on the strength of cell adhesion (19).

Another application allows scientists to study protein folding. Proteins are composed of amino acid chains, which constitute the primary protein structure. These amino acid chains have to undergo folding into tertiary and secondary, 3D structures, which is essential for the proper functioning of these proteins. Recently, advances in AFM have made it possible to study this fascinating biological process and bring new insight into the energy landscapes of protein folding. Proteins involved in mechanical functions are composed of multiple domains, which fold individually. One of these domains is fibronectin. The most common form of fibronectin is fibronectin type III (FN-III), which is found in an estimated 2% of all animal proteins and has thus been studied extensively. FN-III domains are found in the muscle protein titin, which is responsible for the passive elasticity of muscle. These domains have been found to unravel during forced extension. Understanding the forces required in unfolding events as well as the time it takes for unfolding to happen can be critical for the physiological functions of mechanical proteins such as titin (9,20–22).

Now that a brief overview of the potential applications of the AFM has been provided, it is important to understand the principles behind its operation. The following section focuses on the theory of data acquisition using AFM as well as descriptions of the equipment itself. The last section of this article provides a more in-depth evaluation of the technique in addition to discussing the most recent advances in the field.

THEORY AND EXPERIMENTAL APPROACH

This section focuses on the theory and experimental approaches of AFM. The section begins with a description of the imaging AFM as well as its different modes of operation, which allow for its applications in various experimental protocols. Later, force spectroscopy is described. Focus is placed on the force apparatus used in our laboratory, which relies on the same basic principals as the commercially available AFMs. In addition to a description of its operation, this section also discusses the different applications of the force apparatus.

Imaging AFM

Optical Beam Deflection. An AFM has a force sensor called a cantilever to measure the forces between a sharp tip and the surface (23). Unlike the optical microscope that relies on 2D images, the images acquired with the AFM are obtained in three dimensions: the horizontal *xy*-plane and the vertical *z*-direction. As shown in Fig. 4, the tip at the end of the cantilever is brought in close proximity to the sample mounted on a piezoelectric element. The AFM can be compared with a record player such as an old stylus-based instrument (1). It combines the principles of a scanning tunneling microscope (STM) and the stylus profiler. However, the probe forces in the AFM are much smaller than those ($\sim 10^4$ N) achieved with a stylus profiler.

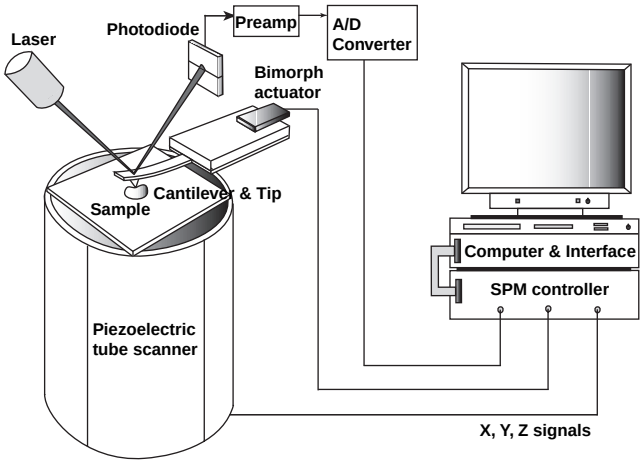


Figure 4. Schematic illustration of an AFM. A force sensor consists of a flexible cantilever with an extremely sharp tip at its end. A ceramic (Si_3N_4) or semiconductor (Si) tip on the cantilever can be brought into close proximity to a sample surface. As the tip is close to the sample surface, it either continuously touches or periodically vibrates on the surface, and bends or changes in its vibration amplitude and frequency. A laser spot is reflected off the top of the cantilever. When the cantilever bends, the laser light is deflected onto a two-panel photodiode. The detected signals are amplified and transformed by electronic circuits and sent to an SPM controller. The SPM controller, the computer, and their interfaces generate an image.

The AFM belongs to the family of scanning probe microscopes (SPMs). Like all other SPMs, the AFM uses an extremely sharp tip that moves over the sample surface with a raster scan (like the movement of the electron beam on the TV screen). In the first AFM, the bending of the cantilever was detected using an STM, but now a sensitive and simple optical method is used in most AFMs (24). As shown in Fig. 4, a laser beam is reflected off the cantilever onto a two-panel photodiode. As the cantilever bends, the position of the reflected laser light changes. Measurements are obtained as a result of the differences in the signal between the two segments of this photo-detector.

Feedback Operation. The AFM uses a piezoelectric element to position and scan the sample with high resolution. A tube-shaped piezoelectric ceramic that has a high stability is used in most SPMs. Application of voltage results in the stretching or bending of the piezoelectric tube, allowing it to move in all three dimensions and to position the cantilever probe with very high precision. For example, by applying a voltage to one of the two electrodes (*xy*-axis) the tube scanner expands and tilts away from a center position (*xy*-origin). A corresponding negative voltage applied to the same electrode causes the tube scanner contract, resulting in movements on the *xy*-plane relative to the origin. The magnitude of the movement depends on the type of piezoelectric ceramic, the shape of the element, and the applied voltage.

Feedback control is used for many common applications, such as thermostats, which are used to maintain a particular temperature in buildings, and autopilot, commonly used in airplanes. In the AFM, a feedback loop is used to

keep the force acting on the tip in a fixed relationship with the surface while a scan is performed. The feedback loop consists of the piezoelectric tube scanner, the cantilever and tip, the sample, and the feedback circuit. The feedback circuit consists of proportional and integral gain controls and provides an immediate response to scanning parameter changes. A computer program acts as a compensation network that monitors the cantilever deflection and attempts to keep it at a constant level.

Contact Mode (Static Mode). The AFM operates by measuring the intermolecular forces between the tip and sample. The most common method used in imaging AFM is contact mode, where the piezoelectric element slightly touches the tip to the sample. The experimental setup is shown in Fig. 4. As a result of the close contact, the tip and sample remain in the repulsive regime of the tip-sample interaction shown in Fig. 5. Thus, the AFM measures repulsive force between the tip and sample. As the raster scan moves the tip along the sample, the two-panel photodiode measures the vertical deflection of the cantilever, which reveals the local sample height. Each contour of the surface results in a movement of the tip in the *xyz*-direction, resulting in a change in the deflection angle of the laser beam. This change is measured through the photodiode and translated finally to an image.

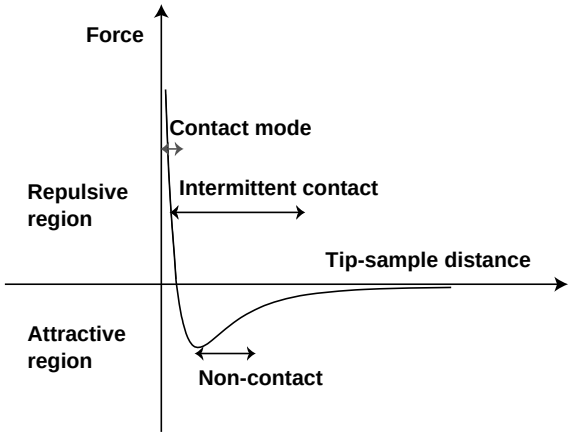


Figure 5. Relationship between the operating modes of AFM and the force regions. The *x*-axis of the graph is the tip-sample distance and the *y*-axis is the force or potential. The graph shows the force or potential, as a function of the distance, simply calculated by the Lennard–Jones potential and the DMT approximation. In contact mode AFM, the tip-sample interaction lies in the repulsive region of the curve above the *x*-axis. As the cantilever is pushed upward, a resulting restoring force occurs, which can be given by Hooke's Law. The difference between this graph and the measured force-distance curve occurs when the force measurement is not static but dynamic and is quickly affected by the capillary force in the thin surface water layer. In intermittent contact mode, the cantilever is operated by a vibration of an amplitude of 10 to 100 nm. The tip is still in contact with the surface, but it just contacts or "taps" on the surface for a very small fraction of its oscillation period. In this operation mode, the tip-sample interaction is broad, ranging from the repulsive region of the curve to the attractive region due to the long-range van der Waals force. The tip-sample interaction in noncontact mode is much weaker than the one in contact and intermittent contact mode. The force between the tip and sample is several orders of magnitude weaker than in the contact regime. As the cantilever in noncontact mode is vibrated near its resonance frequency with an amplitude less than 10 nm, the spacing between the tip and sample is on the order of several to tens of nanometers.

Tip-sample Interaction: The cantilever in the AFM is a critical component (1). The force produced by a spring always tends to restore the spring to its equilibrium position. When the spring is pushed upward by a distance z, it has to be pulled downward. This restoring force is given by Hooke's Law as:

$$F(z) = -k * (z - z_0) \qquad ()$$

where k is a spring constant and depends on the material and dimensions of the cantilever, z is the vertical position of the cantilever, and z_0 is the equilibrium position. As a cantilever with a spring constant of 0.1 newton/meter (N/m) is moved by 1 nm, the resulting force is 0.1 nanonewton (nN). The first tip used by the inventors of the AFM was made from diamond glued to a lever of gold foil (1). Microfabricated cantilever tips are now commercially used.

Electromagnetic forces determine the properties of solids, liquids, and gases; the behavior of particles in solution; and the organization of biological structures (25). These forces are also the source of all intermolecular interactions including covalent bonds, Coulomb forces, ionic forces, ion-dipole interaction, and dipole-dipole interaction. In the AFM, the intermolecular interactions between the tip and the sample surface include van der Waals forces, electrostatic forces, water capillary force, and material properties including elasticity. The most common force in the tip-sample interaction is the van der Waals force. The force is calculated using the Lennard–Jones potential, which combines the attractive van der Waals and the repulsive atomic potentials (25). The force depends on the distance between the tip and the sample, as shown in Fig. 5. This calculated force is an estimate of the van der Waals forces and is usually a few nanonewtons in magnitude.

Force-distance Curve: Since the invention of the AFM, many researchers have used it to measure the tip-sample interaction force on the atomic scale. The AFM records the force as the tip is brought in close proximity to the sample surface, even indented into the surface, and then pulled off. The measured force curve is a plot of cantilever deflection versus the extension of the *z*-piezoelectric scanner (*z*-piezo). A force-distance curve is a kind of interpretation of the force measurements. It needs a simple relationship between the cantilever deflection and the tip-sample distance. Thus, the force-distance curve describes the tip-sample interaction force as a function of the tip-sample distance rather than as a function of the *z*-piezo position. It is difficult to measure the quantitative forces with this technique because the spring constant of the cantilever and the shape of the tip are not accurately known. However, this technique has been used to study adhesion, elasticity, bond rupture length, and even thickness of adsorbed layers. These studies of the fundamental interactions between the sample surfaces have extended across basic

science, chemistry, biology, and even material science. The interaction force between tip and sample is typically on the order of tens of pN for biomolecular interactions. Force measurements in solution have the advantages of the AFM due to the lower tip-sample interaction.

Constant Force and Constant Height: In contact mode, the tip is scanned across the surface in contact either at a constant force or at a constant height above the sample. Constant force mode is achieved by use of a *z*-feedback loop from the deflection signal. The feedback circuits serve to maintain a constant force between the tip and the sample while the tip follows the contours of the surface. The piezoelectric tube can respond to any changes in the cantilever deflection. A computer program acts to keep the cantilever deflection at a constant level. Then, the tip-sample interaction can be kept at a predetermined restoring force. This technique is used to observe the precise topography of the sample surface. If the *z*-feedback loop is switched off, then the *z*-direction of the piezoelectric tube is kept constant, and an image is generated based on the cantilever deflection. Using constant height can be useful for imaging very flat samples.

Lateral Force Microscopy: Lateral force microscopy (LFM) is an extension of contact mode, where an additional detected parameter is the torsion of the cantilever, which changes according to the friction force (26). This lateral force induces a torsion of the cantilever, which, in turn, causes the reflected laser beam to undergo a change in a perpendicular direction to that resulting from the surface corrugation. The LFM uses a photodiode with four segments to measure the torsion of the cantilever. When the cantilever is scanned across the surface in contact, differences in friction between tip and sample cause the tip to stick-slip on the surface. This stick-slip behavior creates a characteristic saw-tooth waveform of atomic level in the friction image (27). The LFM can provide material-sensitive contrast because different components of a composite material exert different friction forces. Researchers often call this operation mode friction force microscopy (27,28). Increasing wear with decreasing sliding velocity on the nanometer scale has been observed with this technique. It has been demonstrated with LFM that, on the atomic scale, frictional properties are sensitive to changes in surface properties on chemical modification. The LFM can also be applied to chemical force microscopy (CFM) by a modified tip with chemical functionality (29). It has been demonstrated with CFM that mapping the spatial arrangement of chemical functional groups and their interactions is of significant importance to problems ranging from lubrication and adhesion to recognition in biological systems.

Capillary Force: The thin surface water layer that exists on the sample surface will form a small capillary bridge between the tip and the sample. The capillary force is important when the AFM is operated in air. Examine the effect of surface tension on AFM measurements. At the moment of tip contact with a liquid film on a flat surface, the film surface reshapes producing a ring around the tip. The water layer wets the tip surface because the water-tip contact (if it is hydrophilic) is energetically advantageous as compared with the water-air contact. If the tip radius is 10 nm and the contact angle is small (i.e., hydrophilic), a capillary force of about 10 nN can result. Thus, the capillary force is the same order of magnitude as the van der Waals interaction. An AFM tip has been used to write alkanethiols with a 30 nm line-width resolution on a gold thin film in a manner analogous to that of a dip pen (30). Recently, this dip-pen nanolithography has also been applied to direct nanoscale patterning of biological materials such as DNA, peptides, and proteins on glass substrates.

Vibration Mode (Dynamic Mode). In dynamic mode, the cantilever is oscillated close to its resonance frequency. This vibration mode operates at a frequency-modulation (FM) mode or the more common amplitude-modulation (AM) mode, which are basically the same as the frequencies used in radio communication. In the FM mode, a *z*-feedback loop keeps a constant force between the tip and the sample while the tip follows the contours of the surface by maintaining the resonance frequency. In the AM mode, the *z*-feedback loop keeps the constant tip-sample interaction by maintaining the amplitude of oscillation.

Intermittent Contact Mode: The cantilever in dynamic mode can easily be vibrated by a piezoelectric ceramic called a bimorph actuator. In air, the cantilever is oscillated close to its resonance frequency and positioned above a sample surface. When the vibrating cantilever comes close to the surface, its oscillation amplitude may change and can be used as the control signal. In this AM mode, the tip is still in contact with the surface, but it just contacts or "taps" on the surface for a very small fraction of its oscillation period. This operation mode is best known as tapping mode in commercial AFMs and, more generally, as intermittent contact mode.

As a raster scan moves the tip on the sample, the four-segment photodiode measures the vibration signal of the cantilever. The detected signal can be changed to root mean-square values by an analog-to-digital converter. In constant force mode, the *z*-feedback loop adjusts so that the averaged amplitude of the cantilever remains nearly constant. Each contour of the surface causes a movement of the tip in the *xyz*-direction, resulting in a change of the oscillation amplitude of the cantilever. This change is measured through a photodiode and finally translated to an image. In air, friction forces due to the surface water layer are dramatically reduced as the tip scans over the surface. Tapping mode may be a far better choice than contact mode for imaging of biological structures due to their inherent softness. In tapping mode, the cantilever can be vibrated at an amplitude of less than 100 nm. Additionally, changes in the phase of oscillation under tapping mode can be used to discriminate between different types of materials on the surface.

Tip-Sample Interaction: The mechanical resonance of the cantilever plays a major role in the response of the system for an interaction between a tip mounted on a vibrating cantilever and a non-homogeneous external force (23). Although basic equations governing the operation of a

bimorph actuator used to vibrate the cantilever are not introduced here, the position of the bimorph is given by:

$$u = u_0 + A_{ex}\cos(\omega t+)$$

where u_0 is the equilibrium position and the excitation is done with amplitude A_{ex}, a frequency ω, and a phase shift ϕ. The fundamental resonance frequency of the cantilever can be approximately calculated from equating its strain energy at the maximum deflection to the kinetic energy at the point of zero deformation. A more accurate method, which takes into consideration all the resonance frequencies of the cantilever together with the modes of vibration, can be obtained by solving the equation of motion subject to the boundary conditions (23). A basic equation to describe the motion of the cantilever is briefly introduced. If the tip-sample interaction is uniform and includes dissipative force in Newton's second law, the vibration system including the cantilever can be described as follows:

$$F(z) = k(z-u) + y(dz/dt) + m(d^2z/dt^2) \qquad ()$$

where $F(z)$ is the tip-sample interaction force, k is a spring constant of the cantilever, z is the vertical position of the cantilever, u is the motion of the bimorph, γ is the dissipation term (i.e., the friction coefficient of the material or the environment), and m is the effective mass of the cantilever. For the constant amplitude mode, we assume that the frictional force $\gamma\ (dz/dt)$ is compensated for by the driving force $F_{ex} = k\,A_{ex}\ \cos(\omega t + \phi)$. Then, the equation of motion is reduced to $F(z) = k\,z + m(d^2z/dt^2)$. If a strong tip-sample interaction occurs only at the point of contact, the motion of the cantilever tip can be almost perfect harmonic oscillation, $z = zo + A\sin\omega t$.

Resolution and Tip Effects: The resolution obtained by an AFM depends greatly on the sharpness of the cantilever tip. Broadening effects usually develop when imaging biological structures having extremely small surface features like a DNA strand (4). If a tip with a radius of curvature of about 20 nm is used to image DNA on a substrate surface, the observed width is about 20 nm, which is considerably greater than the expected value of 2.4 nm deduced from the van der Waals radii of DNA. When the tip radius is comparable with the size of the feature being imaged, it is important to evaluate the radius of the tip end. As such, the development of sharper tips is currently a major concern for commercial vendors, which is also of interest for biologists whose work would greatly benefit from much faster scanning. Recently, improvement of the scanning speed in AFM is one of the most important topics. The tip-sample interaction also tends to distort biological structures because they are relatively soft (31).

Phase Imaging: Phase imaging is an extension of tapping mode based on the measurement of the cantilever phase lag (32). The dependence of phase angles in tapping mode AFM on the magnitude of tip-sample interactions has been demonstrated. The phase images of several hard and soft samples have been recorded as a function of the free amplitude and the reference of the tapping amplitude. It is thought that the elastic deformation associated with the tip-sample repulsive force can be estimated by the repulsive contact interaction. In many cases, phase imaging complements the LFM and force modulation techniques, often providing additional information along with a topographic image. Phase imaging like LFM can also be applied to CFM by using a modified tip with chemical functionality.

Pulsed Force Mode: Pulsed force mode (PFM) is a nonresonant and intermittent contact mode used in AFM imaging (33). It is similar to tapping mode in that the lateral shear forces between the tip and the sample are also reduced. In contrast to tapping mode, the maximum force applied to the sample surface can be controlled, and it is possible to measure more defined surface properties together with topography. This mode is similar to the force modulation techniques of CFM in that a chemically modified tip is used. A series of pseudo force-distance curves can be achieved at a normal scanning speed and with much lower expenditure in data storage. A differential signal can be amplified for imaging of charged surfaces in terms of an electrical double-layer force.

Noncontact Mode: A reconstructed silicon surface has been imaged in a noncontact mode by AFM with true atomic resolution (34). The operation of the AFM is based on bringing the tip in close proximity to the surface and scanning while controlling the tip-sample distance for the constant interaction force. The tip-sample interaction forces in noncontact mode are much weaker than those in contact mode, as shown in Fig. 5. The cantilever must be oscillated above the sample surface at such a distance as is included in the attractive regime of the intermolecular force. Most surfaces are covered with a layer of water, hydrocarbons, or other contaminants when exposed to air, which makes it very difficult to operate in ambient conditions with noncontact mode. Under ultrahigh vacuum, clean surfaces tend to stick together, especially when the materials are identical. The FM mode used in noncontact mode can keep the constant tip-sample interaction by maintaining the resonance frequency of oscillation through the z-feedback loop. Nearly ten years following the invention of the AFM, a few groups achieved true atomic resolution with a noncontact mode (35). After that, several groups succeeded in obtaining true atomiclevel resolution with noncontact mode on various surfaces. Many important yet unresolved problems, such as determining the tip-sample distance where atomic-level resolution can be achieved, still remain. Experimentally, atomic-level resolution can be achieved only between 0 and 0.4 nm. A stiff cantilever vibrates near resonance frequency (300–400 kHz) with amplitude of less than 10 nm.

In covalently bound materials, the charge distribution of surface atoms reflects their bonding to neighboring atoms (36). These charge distributions have been imaged by noncontact mode with a light-atom probe such as a graphite atom. This process revealed features with a lateral distance of only 77 picometers (pm). However, all of the atomic-scale images have been generated in ultrahigh vacuum, which has limited applications in biology. Recently, several groups have reported obtaining atomic-scale images with FM mode in ambient conditions

and liquid environments. In the near future, true atomic-level imaging by AFM will be commercially available in various environments.

Force Spectroscopy

Equipment

AFM Instrumentation. The AFM that is used in the author's laboratory is a homemade modification of the standard AFM design that is used for imaging and is shown in Fig. 5 (19). It operates on the same basic principles as a commercial AFM. In the author's design, improvement of the signal quality by reducing mechanical and electrical noise and improvement of the instrument's sensitivity by uncoupling the mechanisms for lateral and vertical scans was achieved. The cantilever is moved vertically up and down using a piezoelectric translator (Physik Instrumente, model P-821.10) that expands or contracts in response to applied voltage. The vertical range of the piezo is 0–15 μm. A dish coated with substrate is placed below the cantilever, and the cantilever with a cell or protein attached can be lowered onto that dish using the piezo allowing for the receptor-ligand interaction to take place. During the acquisition of a force scan, the cantilever is bent (Fig. 6) causing the beam of a 3 mW diode laser (Oz Optics; em. 680 nm) that is focused on top of the cantilever to be deflected. A two-segment photodiode (UDT Sensors; model SPOT-2D) monitors these deflections of the laser beam. An 18 bit optically isolated analog-to-digital converter (Instrutech Corp., Port Washington, NY) then digitizes the signal from the photodiode. Custom software is used to control the piezoelectric translator and to time the measurements. The AFM is shielded inside of an acoustic/vibration isolation chamber in order to reduce vibration and aid in keeping a stable temperature. The detection limit of the AFM system is in the range of 20 pN.

Cantilever Calibration. It is necessary to determine the spring constant of the cantilever k_C (i.e., $F = k_C x$) in order to translate the deflection of the cantilever x to units of force F. Calibrating the cantilever can be achieved through theoretical techniques that provide an approximation of k_C (37) or through empirical methods. Using empirical methods to determine k_C involves taking measurements of cantilever deflection with application of a constant known force (38) or measuring the cantilever's resonant frequency (39). The method the author's use for calibrating cantilevers is based on Hutter and Bechhoefer (39). The author's use triangleshaped unsharpened gold-coated silicon-nitride cantilever tips that have spring constants ranging from 10 mN/m to 50 mN/m for ligand-receptor force measurements (TM Microscopes, Sunnyvale, CA). The cantilever tip can be treated as a simple harmonic oscillator whose power spectrum of thermal fluctuation can be used to derive the spring constant, which can be achieved by raising the cantilever a few microns from the surface of the experimental dish and monitoring its natural vibrational frequency for 2–3 s. Each vibration mode of the cantilever receives the thermal energy commensurated to one degree of freedom, $k_B T/2$. The measured variance of the deflection $\langle x \rangle^2$, can then be used to calculate the spring constant (i.e., $k_B T = k_C \langle x \rangle^2$, where k_B and T are Boltzmann's constant and temperature, respectively). To separate deflections belonging to the basic (and predominant) mode of vibration from other deflections or noise in the recording system, the power spectral density of the temperature-induced deflection is determined. The spring constant is estimated using only the spectral component corresponding to the basal mode of vibration. The spring constant can be calibrated in either air or solution using this approach. The calculated spring constant k_C can then be used to calculate rupture force F by $F = k_C C D V$. DV is the change in voltage detected by the photodiode just prior to and immediately after the rupture event. C is a calibration constant that relates deflection and photodiode voltage and is determined from the deflection of the cantilever when it is pressed against a rigid surface, such as the bottom of a plastic petri dish (19).

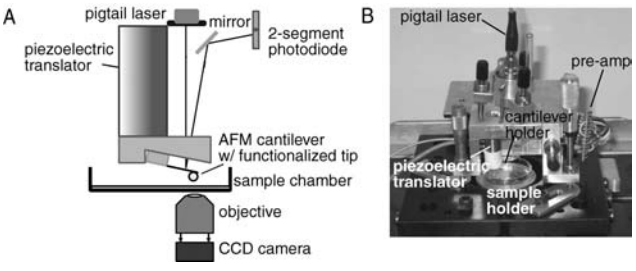


Figure 6. AFM experimental set-up. (a) Schematic diagram of the AFM. (b) Photograph of the author's AFM setup. The CCD camera is not in view.

Applications

Receptor-Ligand Adhesion Measurements. *Bell Model:* AFM force measurements (Fig. 7) of ligand-receptor interactions can be used to determine the dynamic strength of a complex and characterize the changes in free energy that the particular complex undergoes (i.e., energy landscape) during its breakage. The Bell model can be used to interpret these measurements (40). The Bell model is based on the assumption that the application of an external mechanical force to a receptor-ligand interaction bond will reduce the activation energy that needs to be overcome in order to break this bond. This unbinding force should increase with the logarithm of the rate at which an external mechanical force is applied toward the unbinding of adhesion complexes (i.e., loading rate), which was confirmed by a number of studies. For example, studies using the biomembrane force probe (BFP) (40) and the AFM have shown that increases in loading rate cause an increase in rupture force between individual complexes of streptavidin/biotin (12,15,41).

AFM Measurements of Adhesive Forces: In order to carry out force measurements, a cell is first attached to a cantilever tip and another cell or substrate proteins are plated on a dish. The method employed to attach cells to the cantilever tip works best on nonadherent items. A cell is attached to the AFM cantilever via concanavalin A (con A)-mediated linkages (15). Most cells have receptors for con A on their surface and will attach to the tip. To prepare the con A-functionalized cantilever, the cantilevers are soaked

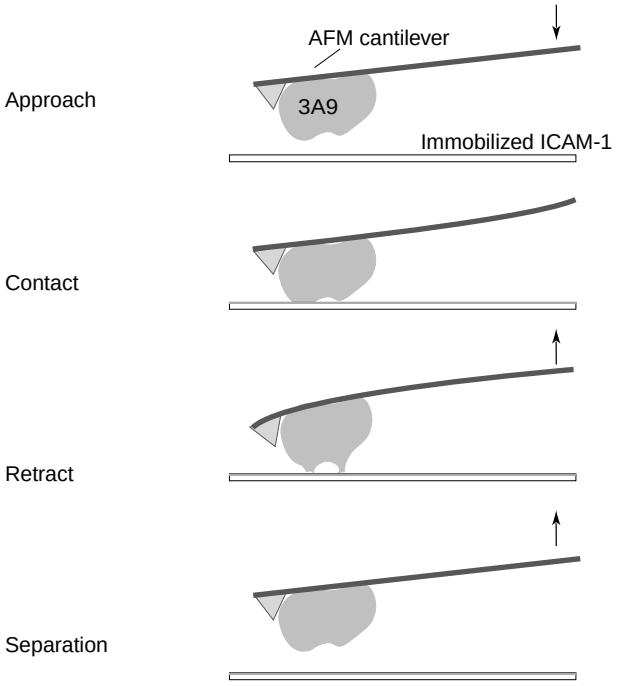


Figure 7. Steps in the acquisition of an AFM force measurement. The first step is the approach of the cantilever with a cell bound to the substrate, followed by contact between the cell and substrate and retraction of the cantilever, which results in the separation of the cell from the substrate. The cantilever is bent during this process. The arrows indicate the direction of cantilever movement.

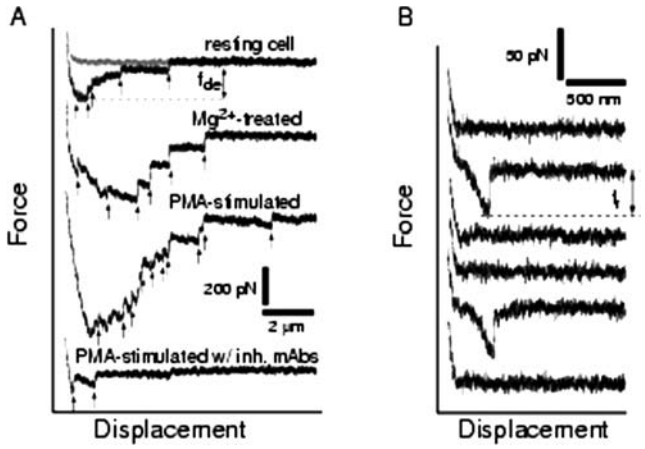


Figure 8. AFM force versus displacement traces of the interaction between cells expressing LFA-1 and immobilized ICAM-1. (a) Multiple-bond measurements acquired with a compression force of 200 pN, 5 s contact, and a cantilever retraction speed of 2 μm/s. The measurements were carried out with a resting, untreated cell (1st trace), a Mg^{2+}-treated cell (2nd trace), and a PMA-stimulated cell (3rd trace). The 4th trace corresponds to a measurement acquired from a PMA-stimulated cell in the presence of LFA-1 (20 μg/ml FD441.8) and ICAM-1 (20 μg/ml BE29G1) function-blocking monoclonal antibodies (mAbs). Arrows point to breakage of LFA-1/ICAM-1 bond(s). f_{de} is the detachment force, and the shaded area estimates the work of de-adhesion. (b) Single-molecule measurements of LFA-1/ICAM-1 unbinding forces. Traces 2 and 5 show adhesion. Measurements were obtained under conditions that minimized contact between the LFA-1-expressing cell and the ICAM-1-coated surface. The compression force was reduced to $\sim$60 pN and the contact time to 50 ms. An adhesion frequency of less than 30% in the force measurements ensured that a $>$85% probability exists that the adhesion event is mediated by a single LFA-1/ICAM-1 complex (42,43). The frequency of adhesion in test and control experiments was examined to confirm the specificity of the interaction. The addition of monoclonal antibodies against either LFA-1 or ICAM-1 significantly lowered the frequency of adhesion of both resting cells and activated cells under identical experimental conditions.

in acetone for 5 min, UV irradiated for 30 min, and incubated in biotinamidocaproyl-labeled bovine serum albumin (biotin-BSA, 0.5 mg/ml in 100 mM $NaHCO_3$, pH 8.6; Sigma) overnight. The cantilevers are then rinsed three times with phosphate-buffered saline (PBS, 10 mM PO_4^{3-}, 150 mM NaCl, pH 7.3) and incubated in streptavidin (0.5 mg/ml in PBS; Pierce, Rockford, IL) for 10 min at room temperature. Following the removal of unbound streptavidin, the cantilevers are incubated in biotinylated Con A (0.2 mg/ml in PBS; Sigma) and then rinsed with PBS.

The actual process of attaching a cell to a cantilever tip is reminiscent of fishing. A cantilever tip is positioned above the center of the cell. The largest triangular cantilever (320 μm long and 22 μm wide) with a spring constant of 0.017 N/m on the cantilever chip is usually used in our measurements. The cell is brought into focus, with the cantilever slightly out of focus. Then, the tip is lowered onto the center of the cell and held there motionless for approximately 1 s. When attached, the cell is positioned right behind the AFM cantilever tip. The force required to dislodge the cell from the tip is greater than 2 nN, which is much greater than the forces measured in the receptor-ligand studies that were, at most, 300 pN (15).

A piezoelectric translator is used during measurement acquisition to lower the cantilever with an attached cell onto the sample. The interaction between the attached cell and the sample is given by the deflection of the cantilever. This deflection is measured by reflecting a laser beam off the cantilever into a position sensitive two-segment photodiode detector, as described in the instrumentation section above.

As a result of this process, a force scan is obtained. The studies shown in Fig. 8 (42,43) were conducted on cells expressing the adhesion receptor LFA-1 (leukocyte function-associated antigen-1), an integrin expressed on the surface of T-cells, and its ligand ICAM-1 (intercellular adhesion molecule-1), expressed on the surface of APCs. In these experiments, LFA-1-expressing cells and ICAM-1 protein were used. An example of a few force scans from multiple bond cell adhesion studies can be seen in Fig. 8a. The red trace is the approach trace and the black is the retract trace. As the cantilever is lowered and contact is made between the cell and substrate, an initial increase in force occurs. As the cantilever is retracted back up, the force returns to zero and becomes negative as bonds are stretched and begin to break. The jumps in the force scan, which are pointed out by the arrows, represent bonds breaking. Two parameters can be used in such measurements to assess the level of cell adhesion. One is the detachment force, which is the maximum force required to dislodge the cell. Another is the work of deadhesion, which is the amount of work required to pull and stretch

the cell and for the bonds to break. It is derived by integrating the adhesive force over the distance traveled by the cantilever. In this example, Mg^{2+} and PMA are used, which enhance the adhesion of the cells studied through various mechanisms. It is easily observed that a very pronounced increase occurs in the area under the curve as well as the number of bonds that were broken following the application of these adhesion stimulators (16).

FM Force Measurements of Individual Receptor/Ligand Complexes: A different set of information can be derived from single-molecule adhesion measurements. These type of studies offer insight into the dissociation pathway of a receptor-ligand interaction and the changes in free energy that are associated with this process, which is achieved by measuring single-bond interactions between a receptor and ligand at increasing loading rates (20 pN/s–50,000 pN/s). In the author's setup, it translates to using rates of retraction of the cantilever from 0.1 to 15 μm/s.

In order to obtain unbinding forces between a single receptor-ligand pair, the experiments have to be carried out in conditions that minimize contact between the cantilever tip and substrate. A > 85% probability exists that the adhesion event is mediated by a single bond if the frequency of adhesion is maintained below 30% (15,42). An example of such measurements can be seen in Fig. 8b.

Depending on the speed at which the cantilever is retracted during the measurements, the collected data usually needs to be corrected for hydrodynamic drag, which is due to the hydrodynamic force that acts in the opposite direction of cantilever movement, and its magnitude is proportional to the cantilever movement speeds. The hydrodynamic force may be determined based on the method used by Tees et al. and Evans et al. (42,43). In single-bond AFM studies, it is found that the data obtained with cantilever retraction speeds higher than 1 μm/s needed to be corrected by adding the hydrodynamic force.

The damping coefficient can be determined by plotting the hydrodynamic force versus the speed of cantilever movement. The damping coefficient is the slope of the linear fit and was found to be about 2 pN □s/μm in the author's work (15).

AFM Measurements of Cell Elasticity. The AFM can also be used as a microindenter that probes the mechanical properties of the cell. In these measurements, which enable assesment of cell elasticity, a bare AFM tip is lowered onto the center of the cell surface at a set rate, typically 2 μm/s. Following contact, the AFM tip exerts a force against the cell that is proportional to the deflection of the cantilever. The indentation force used is below 1 nN (∼600 pN) in order to prevent damage to the cell. The deflection of the cantilever is recorded as a function of the piezoelectric translator position during the approach and withdrawal of the AFM tip. The force-indentation curves of the cells are derived from these records using the surface of the tissue culture dish to calibrate the deflection of the cantilever. Then, one can estimate the Young's modulus, which is a measure of elasticity. Estimates of Young's modulus are made on the assumptions that the cell is an isotropic elastic solid and the AFM tip is a rigid cone (44–46). According to this model, initially proposed by Hertz, the force (F) indentation (α) relation is a function of Young's modulus of the cell, K, and the angle formed by the indenter and the plane of the surface, θ, as follows:

$$F = \frac{K}{2(1-\upsilon^2)} \frac{4}{\pi \tan\theta} \alpha^2 \qquad (1)$$

Young's modulus are obtained in the author's laboratory by least square analysis of the forceindentation curve using routines in the Igor Pro (WaveMetrics, Inc., Lake Oswego, OR) software package. The indenter angle θ and Poisson ratio ν are assumed to be 55° and 0.5, respectively.

In order to determine the cell's elasticity, the force versus indentation measurements are fitted to the curves of the Hertz model. Figure 9 (44–46) illustrates an example of such measurements acquired on cells of varying elasticity. Cells with the greatest degree of indentation at a particular applied force will have the lowest Young's modulus values and will therefore be the "softest."

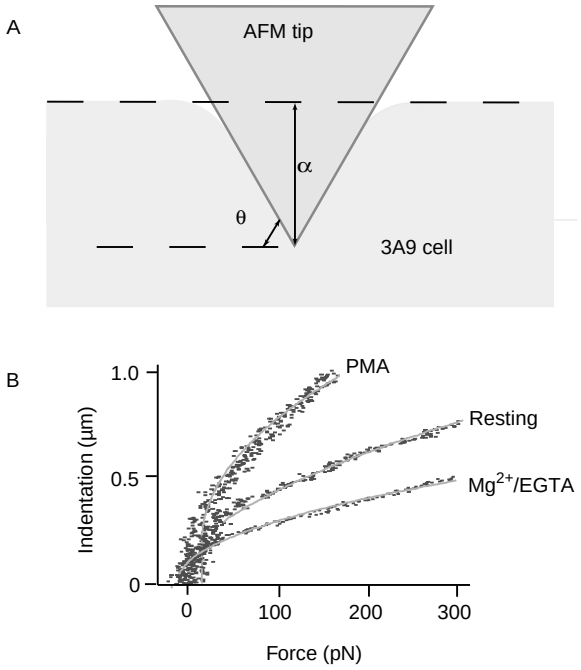


Figure 9. Acquisition of cell compliance measurements. (a) Tip of the AFM cantilever indenting a 3A9 cell. The cell compliance measurements were based on the assumption that the cell is an isotropic elastic solid and the AFM tip is a rigid cone (44–46). According to this model, initially proposed by Love and Hertz, the force (F)-indentation (α) relation (shown) is a function of Young's modulus of the cell, K, and the angle formed by the indenter and the plane of the surface, θ, as in equation (1). The indenter angle, θ, is assumed formed by the AFM tip and the 3A9 cell to be 55° and the Poisson ratio ν to be 0.5. (b) Force versus indentation traces of resting, PMA-stimulated and Mg^{2+}-treated 3A9 cells.

Protein Folding/Unfolding. The AFM can also be used to study protein unfolding. A cantilever tip is used to pick up proteins attached to a surface, which is followed by retrac-

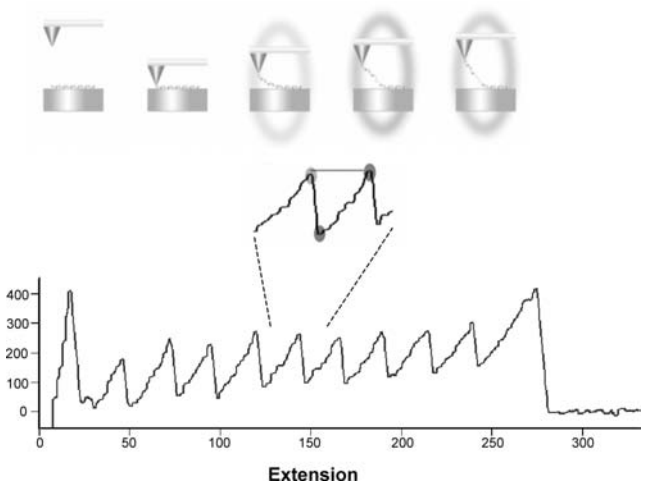


Figure 10. Consecutive unfolding peaks of a titin polyprotein, composed of FNIII domains. The inset demonstrates the corresponding steps of the unfolding process in correlation to the AFM data.

tion of the cantilever, which results in protein unfolding. The length of the unfolded protein can be over ten times its folded length, depending on the protein being studied (9).

This forced protein unfolding generates an opposing force due to the sudden drop in entropy as the protein is unfolded. Although a lower force is required to begin unfolding the protein, the force required to continue the unfolding is increased rapidly as the protein approaches its full, unfolded length. This phenomenon has been described by the worm-like chain model (WLC) of elasticity. The WLC model is based on two parameters, the total or contour length of the polymer being stretched and the persistence length. The persistence length reflects the polymer flexibility and is the length attained when a polymer is bent. A smaller persistence length is an indication of higher entropy and a polymer that is more difficult to unfold (47). When a multidomain protein is extended using the AFM, the first domain is unfolded at a certain pulling force, followed by a return of the cantilever to zero. Further unfolding meets with resistance once again, resulting in a characteristic saw-tooth profile of the unfolding, with each domain that was unfolded being represented by a peak. Figure 10 from Andreas F. Oberhauser illustrates this process. It is the unfolding of a titin polyprotein, which is composed of FNIII domains (22). The protein can also be refolded following unfolding, which is done by bringing the cantilever back down to the substrate and once again retracting it. If force curves representative of unfolding are observed once again, then refolding most likely took place. It is a much slower process (on the order of seconds) than the forced unfolding (48).

EVALUATION

Imaging AFM

The AFM is an exciting novel technology that enables the study of biological structures under physiological conditions. The AFM is probably the only technique of its kind that enables image dynamic processes taking place in real time. A number of other techniques are currently available in the biological sciences for imaging studies, however, most result in modifications to the biological sample. One such technique is electron microscopy (EM), which, until recently, provided images of the highest resolutions. In recent years, a number of modifications to the AFM have brought the resolution up to par and even surpassed those of EM.

In recent years, many advances have been made in the field of AFM. Significant improvements in resolution have been gained through cantilever tip modification. The currently available cantilevers are relatively "soft" and flexible with spring constants of 0.01–0.5 N/m. Tip deformation is one aspect that limits resolution. Recently, stiffer cantilevers have been designed improving resolution. One example are quartz cantilevers with spring constants on the order of 1 kN/m allowing for possibly subatomic-level resolution (34). Smaller cantilevers have been designed that increase the possible scanning speed of the AFM. Images of 100×100 pixels (240 nm scan size) have been achieved in 80 ms. A sharper, finer tip can also improve resolution, which has been achieved through the use of carbonanotubes, probably the greatest probe improvement to date, which are seamless cylinders composed of sp^2-bonded Carbon (49).

Several characteristics exist that make Carbon nanotubes improved AFM tip materials, including small diameter, a high aspect ratio, large Young's modulus, and mechanical robustness. They are able to elastically buckle under large loads. All these properties translate into higher sample resolution. Chemical vapor deposition has made it easier to grow Carbon nanaotubes on cantilever surfaces, a process that replaces previously more laborious and time-consuming attachment techniques (4,50).

A few techniques also exist worth mentioning that have improved AFM imaging. One such method is cryoAFM. This method addresses the previously mentioned problem of tip and sample flexibility. In this case, samples are imaged at extremely cold temperatures in their cryogenic states, which provides a rigid surface that exhibits a high Young's modulus, thus reducing cantilever deformation and increasing resolution. CryoAFM images match and surpass EM images. Other improvements address the problem resulting from the vibration induced by the commonly used piezoelectric translator. This vibration is translated to the cantilever holder and the liquid containing the sample being imaged. Magnetic mode (MAC) eliminates the cantilever holder entirely and replaces it with a magnetic cantilever. The cantilever is manipulated via a magnetic field. Photothermal mode (PMOD) uses a bimetallic cantilever that is oscillated via a pulsed diode laser (50,51).

Advances have also been made in single-molecule manipulation with a nanomanipulator. This method relies on a force feedback pen that actually allows the user to touch and manipulate the sample being studied. For example, one can dissect DNA from a chromosome. The interaction forces involved during manipulation of samples can also be studied through oscillating mode imaging. The nanomanipulator can now measure forces in the pN–μN range. For excellent reviews on this technique, see Yang et al. (50) Fotiadis et al. (52).

Progress has also been made in imaging of membrane proteins, which are not ideal candidates for X-ray crystallography, as they do not readily form 3D crystals. Atomic-level resolution images have been obtained of membrane proteins complexed with lipids using EM. However, AFM images of these proteins offer an improvement in that they can be carried out in near physiological conditions and allow for the acquisition of functional and structural information (50,52).

The continued improvements leading to the enhanced imaging capabilities of AFM are reflected in the most recent work being done in the field. We would like to highlight one area where a great deal of progress has been made, which is the filed of DNA and RNA assembly of nanostructures in which the imaging AFM plays a pivotal role. Some of the earlier successes in this area included the construction of 2D DNA arrays, which were assembled in a predictable manner (53). Much progress has also been made with RNA in an attempt to design self-assembling building blocks. The goal of such studies is to generate molecular materials, the geometry of which can be determined for applications in nanobiotechnology (54–56). Chworos et al. were able to form stable RNA structures termed "tectosquares" from RNA helices without the presence of proteins (5). "TectoRNAs" can be thought of as RNA Lego blocks that can be used for the formation of supramolecular structures. In order for these structures to be assembled, the right conditions have to be met in terms of divalent ion concentrations, temperature, as well as the strength, length, and orientation of the RNA. The AFM is an essential tool used in this process as it allows the researcher to obtain detailed images of the assembled tectosquares providing the opportunity to compare predicted structures with those that actually formed. The structures formed by Chworos et al. were in good agreement with the predicted structures. Figure 11 demonstrates the amazing predictability of assembly of these structures, where nine different types of RNA patterns were created (5). Such DNA and RNA structures may have applications in nanotechnology and the material sciences as they could be used to generate nanochips, nanocircuits, and nanocrystals (57). For an excellent recent review of DNA nanomechanical devices, please see Seeman (58).

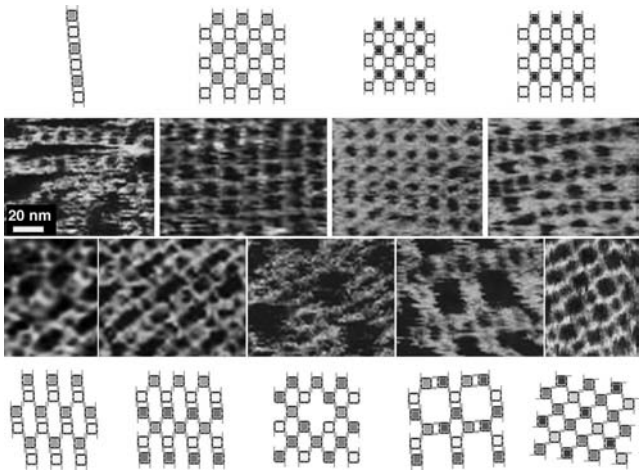


Figure 11. Diagram and AFM images of tectosquare nanopatterns generated from 22 tectosquares. One micrometer square scale AFM images obtained in solution (clockwise from the upper leftmost image) for the ladder pattern, fish net pattern, striped velvet pattern, basket weave pattern, cross pattern, tartan pattern, polka dot pattern, lace pattern, and diamond pattern. Scale bar, 20 nm.

Force Spectroscopy

Force spectroscopy allows us to measure interaction forces between receptors and their respective ligands. These studies can be carried out with purified proteins or cells, or a combination of both. Traditionally, adhesion measurements have been conducted using adhesion assays, which involve the attachment of cells to dishes coated with substrate. The cells are later dislodged manually or using a centrifuge, which are older yet still viable techniques that provide basic kinetic information about the interaction of a receptor-ligand pair. More advanced techniques for conducting force measurements include the use of microneedles, optical tweezers, magnetic beads, and the biomembrane force probe. These techniques, much like the AFM, provide more advanced information from which energy landscapes of an interacting receptor-ligand pair may be determined (11).

The AFM is also a powerful tool for determining the mechanical properties of cells, which was traditionally done using micropipettes or the cell poker, which offered much less precision than the AFM. More recently, methods such as the scanning acoustic microscope, optical tweezers, and magnetic tweezers have also been used in addition to the AFM (59).

An important advantage of the AFM over other methods is that it can be used in conjunction with other techniques through relatively simple modifications. Recently, it has been combined with the patch clamp technique to study the mechanically activated ion channels of sensory cells of the inner ear. This strategy allowed the researchers to use the AFM tip to stimulate the mechanosensory hair bundles by exerting force on them and measure the electrical output of the patch clamp simultaneously (9,60). Another example is combining an AFM with a confocal microscope, which could allow one to monitor cellular responses to AFM measurements using fluorescent reporter systems. One could monitor calcium levels, expression of caspases, and so on (61,62). The AFM could also be combined with RICM microsopy as well as FRET.

Other recent advances involve modifications that would allow for more efficient and effective receptor-ligand studies, including the use of more than one cantilever on the same chip simultaneously. In this case, multiple proteins could be attached and their interaction with their ligand could be measured. So far, this approach has been done with two cantilevers, which involves the use of two laser beams. Further modifications could allow for measurements with even more proteins. Improvements can also be mad in plating of the ligands proteins. In the ideal scenario, different proteins could be plated so that interaction between different receptor-ligand pairs cools be

carried out simultaneously. Current improvements also involve finding better ways to attach cells and proteins to the cantilever that would result in covalent attachment to the tip. Another area that requires improvement is data analysis. The currently available analysis program involves days of rather tedious computer time. Automating analysis would greatly reduce the time required to study a particular interaction. Also, some improvements can be made in data acquisition, where still frequent adjustments of cantilever retraction speed and contact time are required throughout the course of the experiment. Automating data acquisition would allow experiments to be carried out during the night, when noise levels are also minimal.

The applications of AFM technology are vast and too numerous to describe in one review article. The author's have attempted to summarize the technology that was deemed to be of great importance in the developing field of AFM. AFM technology is still limited to a relatively small number of laboratories, which is most likely due to the lack of familiarity with the field, limited expertise in operation, as well as the expense involved in acquiring an AFM. However, it is changing as more and more people are discovering the possibilities that become open to them if they acquire and familiarize themselves with this technology.

BIBLIOGRAPHY

1. Binnig G, Quate CF, Gerber C. Atomic force microscope. Phys Rev Lett 1986;56:930–933.
2. Heinz WF, Hoh JH. Spatially resolved force spectroscopy of biological surfaces using the atomic force microscope. Trends Biotechnol 1999;17(4):143–150.
3. Engel A, Lyubchenko Y, Muller D. Atomic force microscopy: A powerful tool to observe biomolecules at work. Trends Cell Biol 1999;9(2):77–80.
4. Hansma HG. Surface biology of DNA by atomic force microscopy. Annu Rev Phys Chem 2001;52:71–92.
5. Chworos A, et al. Building programmable jigsaw puzzles with RNA. Science 2004;306(5704):2068–2072.
6. Kasas S, et al. Escherichia coli RNA polymerase activity observed using atomic force microscopy. Biochemistry 1997; 36(3):461–468.
7. Rivetti C, et al. Visualizing RNA extrusion and DNA wrapping in transcription elongation complexes of bacterial and eukaryotic RNA polymerases. J Mol Biol 2003;326(5):1413–1426.
8. Pesen D, Hoh JH. Modes of remodeling in the cortical cytoskeleton of vascular endothelial cells. FEBS Lett 2005;579(2): 473–476.
9. Horber JKH. Local probe techniques. Methods Cell Biol 2002;68:1–32.
10. Lee GU, Kidwell DA, Colton RJ. Sensing discrete streptavidin-biotin interactions with AFM. Langmuir 1994;10(2):354–361.
11. Zhang X, Chen A, Wojcikiewicz E, Moy VT. Probing ligand-receptor interactions with atomic force microscopy. In: Protein-Protein Interactions: A Molecular Cloning Manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2002. p 241–254.
12. Yuan C, et al. Energy landscape of streptavidin-biotin complexes measured by atomic force microscopy. Biochemistry 2000;39(33):10219–10223.
13. Benoit M, et al. Discrete interactions in cell adhesion measured by single molecule force spectroscopy. Nat Cell Biol 2000;2(6):313–317.
14. Benoit M. Cell adhesion measured by force spectroscopy on living cells. Methods Cell Biol 2002;68:91–114.
15. Zhang X, Wojcikiewicz E, Moy VT. Force spectroscopy of the leukocyte function-associated antigen-1/intercellular adhesion molecule-1 interaction. Biophys J 2002;83(4):2270–2279.
16. Wojcikiewicz EP, et al. Contributions of molecular binding events and cellular compliance to the modulation of leukocyte adhesion. J Cell Sci 2003;116(12):2531–2539.
17. Matzke R, Jacobson K, Radmacher M. Direct, high-resolution measurement of furrow stiffening during division of adherent cells. Nat Cell Biol 2001;3(6):607–610.
18. Hassan AE, et al. Relative microelastic mapping of living cells by atomic force microscopy. Biophys J 1998;74(3):1564–1578.
19. Wojcikiewicz EP, Zhang X, Moy VT. Force and compliance measurements on living cells using atomic force microscopy (AFM). Biol Proced Online 2004;6:1–9.
20. Rief M, et al. Reversible unfolding of individual titin immunoglobulin domains by AFM [see comments]. Science 1997;276(5315):1109–1112.
21. Oesterhelt F, et al. Unfolding pathways of individual bacteriorhodopsins [see comments]. Science 2000;288(5463):143–146.
22. Li H, et al. Reverse engineering of the giant muscle protein titin. Nature 2002;418(6901):998–1002.
23. Sarid D. Scanning Force Microscopy. New York: Oxford University Press; 1991.
24. Meyer G, Amer NM. Novel optical approach to AFM. Appl Phys Lett 1988;53:1045–1047.
25. Israelachvili JN. Intermolecular and Surface Forces. 2nd ed. London: Academic Press; 1992.
26. Meyer G, Amer NM. Simultaneous measurement of lateral and normal forces with an optical-beam-deflection AFM. Appl Phys Lett 1990;57(20):2089–2091.
27. Mate CM, et al. Atomic-scale friction of a tungsten tip on a graphite surface. Phys Rev Lett 1987;59(17):1942–1945.
28. Overney RM, et al. Friction measurements on phasesegregated thin films with a modified atomic force microscope. Nature 1992;359:133–135.
29. Frisbie CD, et al. Functional group imaging by chemical force microscopy. Science 1994;265:2071–2074.
30. Piner RD, et al. "Dip-Pen" nanolithography. Science 1999; 283(5402):661–663.
31. Kwak KJ, et al. Topographic effects on adhesive force mapping of stretched DNA molecules by pulsed-force-mode atomic force microscopy. Ultramicroscopy 2004;100(3–4):179–186.
32. Magonov SN, EV, Whango MH. Phase imaging and stiffness in tapping mode AFM. Surf Sci 1997;375:L385–L391.
33. Miyatani T, Horii M, Rosa A, Fujihira M, Marti O. Mapping of electrical double-layer force between tip and sample surfaces in water with pulsed- forcemode atomic force microscopy. Appl Phys Lett 1997;71:2632–2634.
34. Giessibl FJ, et al. Subatomic features on the silicon (111)-(7×7) surface observed by atomic force microscopy. science 2000;289(5478):422–426.
35. Morita SR, Wiesendanger R, Meyer E. Noncontact AFM. New York: Springer; 2002.
36. Hembacher S, Giessibl FJ, MJ. Force microscopy with light-atom probes. Science 2004;305(5682):380–383.
37. Sader JE. Parallel beam approximation for V-shaped atomic force micrscope cantilevers. Rev Sci Instrum 1995;66:4583–4587.
38. Senden TJ, Ducker WA. Experimental determination of spring constants in atomic force microscopy. Langmuir 1994;10: 1003–1004.

39. Hutter JL, Bechhoefer J. Calibration of atomic-force microscope tips. Rev Sci Instrum 1993;64(7):1868–1873.
40. Bell GI. Models for the specific adhesion of cells to cells. Science 1978;200:618–627.
41. Merkel R, et al. Energy landscapes of receptor-ligand bonds explored with dynamic force spectroscopy [see comments]. Nature 1999;397(6714):50–53.
42. Tees DFJ, Woodward JT, Hammer DA. Reliability theory for receptorligand bond dissociation. J Chem Phys 2001;114: 7483–7496.
43. Evans E. Probing the relation between force—lifetime—and chemistry in single molecular bonds. Ann Rev Biophys Biomolec Struc 2001;30:105–128.
44. Hoh JH, Schoenenberger CA. Surface morphology and mechanical properties of MDCK monolayers by atomic force microscopy. J Cell Sci 1994;107:1105–1114.
45. Radmacher M, et al. Measuring the viscoelastic properties of human platelets with the atomic force microscope. Biophys J 1996;70(1):556–567.
46. Wu HW, Kuhn T, Moy VT. Mechanical properties of L929 cells measured by atomic force microscopy: Effects of anticytoskeletal drugs and membrane crosslinking. Scanning 1998;20(5): 389–397.
47. Fisher TE, et al. The study of protein mechanics with the atomic force microscope. Trends Biochem Sci 1999;24(10):379–384.
48. Altmann SM, Lenne P-F. Forced unfolding of single proteins. Methods Cell Biol 2002;68:312–336.
49. Hafner JH, et al. Structural and functional imaging with carbon nanotube AFM probes. Prog Biophys Molec Biol 2001;77(1):73–110.
50. Yang Y, Wang H, Erie DA. Quantitative characterization of biomolecular assemblies and interactions using atomic force microscopy. Methods 2003;29(2):175–187.
51. Sheng S, Zhifeng S. Cryo-atomic force microscopy. Methods Cell Biol 2002;68:243–256.
52. Fotiadis D, et al. Imaging and manipulation of biological structures with the AFM. Micron 2002;33(4):385–397.
53. Winfree E, et al. Design and self-assembly of two-dimensional DNA crystals. Nature 1998;394(6693):539–544.
54. Hansma HG, Kasuya K, Oroudjev E. Atomic force microscopy imaging and pulling of nucleic acids. Curr Opin Struct Biol 2004;14(3):380–385.
55. Jaeger L, Westhof E, Leontis NB. TectoRNA: Modular assembly units for the construction of RNA nano-objects. Nucl Acids Res 2001;29:455–463.
56. Seeman NC. DNA in a material world. Nature 2003;421:427–431.
57. Yan H, et al. DNA-templated self-assembly of protein arrays and highly conductive nanowires. Science 2003;301(5641): 1882–1884.
58. Seeman NC. From genes to machines: DNA nanomechanical devices. Trends Biochem Sci 2005;30(3):119–125.
59. Radmacher M. Measuring the elastic properties of living cells by AFM. Methods Cell Biol 2002;68:67–90.
60. Langer MG, Koitschev A. The biophysics of sensory cells of the inner ear examined by AFM and patch clamp. Methods Cell Biol 2002;68:142–171.
61. Charras GT, Lehenkari P, Horton M. Biotechnological applications of AFM. Methods Cell Biol 2002; 68.
62. Charras GT, Horton MA. Single cell mechanotransduction and its modulation analyzed by atomic force microscope indentation. Biophys J 2002;82(6):2970–2981.

See also BIOMAGNETISM; NANOPARTICLES.

MICROSCOPY, SCANNING TUNNELING

VIRGINIA M. AYRES
Michigan State University
East Lansing, Michigan

INTRODUCTION

Four years after its invention in 1982 (1), the scanning tunneling microscope (STM) was awarded the 1986 Nobel Prize for physics, one of only four such prestigious awards given for a truly significant contribution to scientific instrumentation. Since then, the family of scanning probe microscopy (SPM) techniques, which includes scanning tunneling microscopy, atomic force microscopy (2–4), magnetic force microscopy (5), near-field optical microscopy (6), scanning thermal microscopy (7), and others, has revolutionized studies of semiconductors, polymers, and biological systems. The key capability of SPM is that, through a controlled combination of feedback loops and detectors with the raster motion of piezoelectric actuator, it enables direct investigations of atomic-to-nanometer scale phenomena.

Scanning probe microscopy is based on a piezoelectric-actuated relative motion of a tip versus sample surface, while both are held in a near-field relationship with each other. In standard SPM imaging, some type of tip-sample interaction (e.g., tunneling current, Coulombic forces, magnetic field strength) is held constant in z through the use of feedback loops, while the tip relative to the sample undergoes an x–y raster motion, thereby creating a surface map of the interaction. The scan rate of the x–y raster motion per line is on the order of seconds while the tip-sample interaction is on the order of nanoseconds or less. The SPM is inherently cable of producing surface maps with atomic scale resolution, although convolution of tip and sample artifacts must be considered.

Scanning tunneling microscopy is based on a tunneling current from filled to empty electronic states. The selectivity induced by conservation of energy and momentum requirements results in a self-selective interaction that gives STM the highest resolution of all scanning probe techniques. Even with artifacts, STM routinely produces atomic scale (angstrom) resolution.

With such resolution possible, it would be highly desirable to apply STM to investigations of molecular biology and medicine. Key issues in biology and medicine revolve around regulatory signaling cascades that are triggered through the interaction of specific macromolecules with specific surface sites. These are well within the inherent resolution range of STM.

The difficulty when considering the application of STM to molecular biology is that biological samples are nonconductive. It may be more accurate to describe biological samples as having both local and varying conductivities. These two issues will addressed in this article, and examples of conditions for the successful use of STM for biomedical imaging will be discussed. We begin with an overview of successful applications of STM in biology and medicine.

SCANNING TUNNELING MICROSCOPY IN BIOLOGY AND MEDICINE: DNA AND RNA

The STM imaging for direct analysis of base pair arrangements in DNA was historically the first biological application of the new technique. An amusing piece of scientific history is that the first (and widely publicized) images (8–12) of (deoxyribonucleic acid) DNA were subsequently shown to correspond to electronic sites on the underlying graphite substrate! However, more careful investigations have resulted in an authentic body of work in which the base pairings and conformations of DNA and RNA are directly investigated by STM. One goal of these investigations is to replace bulk sequencing techniques and crystal diffraction techniques, which both require large amounts of material, with the direct sequencing of single molecules of DNA and RNA. Two examples of DNA and RNA investigation by STM are presented here. One is an investigation of DNA and RNA structures, and the other is an investigation of DNA biomedical function.

Recently reported research from the group at The Institute for Scientific and Industrial Research at Osaka University in Japan (13) has shown detailed STM images of well-defined guanine-cytosine (G-C) and adenine-thymine (A-T) base pairings in double- and single-stranded DNA. Four simple samples involving only G-C and only A-T base pairs in mixed (hetero) and single sided (homo) combinations were chosen for analysis (Fig. 1). These were deposited on a single-crystal copper (111)-orientation [Cu(111)] substrate using a technique developed specially by this group to produce flat, extended strands for imaging. An STM image showing the individual A-T base pairs in the hetero A-T sample is shown in Fig. 2. Images of the overall structures indicated repeat distances consistent with interpretation as the double helix. Images from mixed samples of hetero G-C and hetero A-T are shown in Fig. 3. The larger structure is interpreted as hetero G-C and the smaller as hetero A-T, which is consistent with X-ray diffraction data that indicates the A-T combination is more compact.

Only the double helix structure was observed for the hetero G-C samples. However, the homo G-C structures,

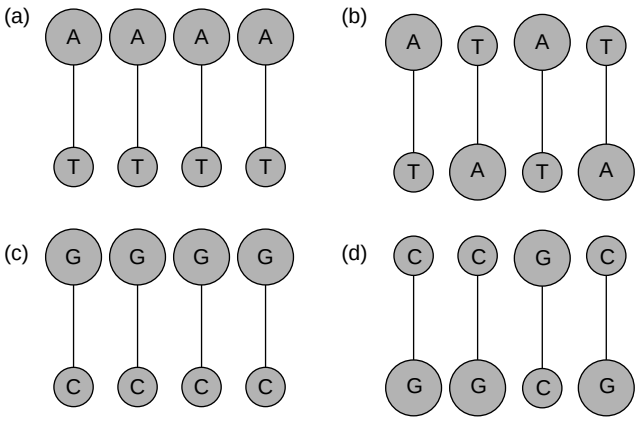


Figure 1. (a) Homo A-T, (b) Hetero A-T, (c) Homo G-C, and (d) Hetero G-C. (Figure adapted from Ref. 9, used with permission.)

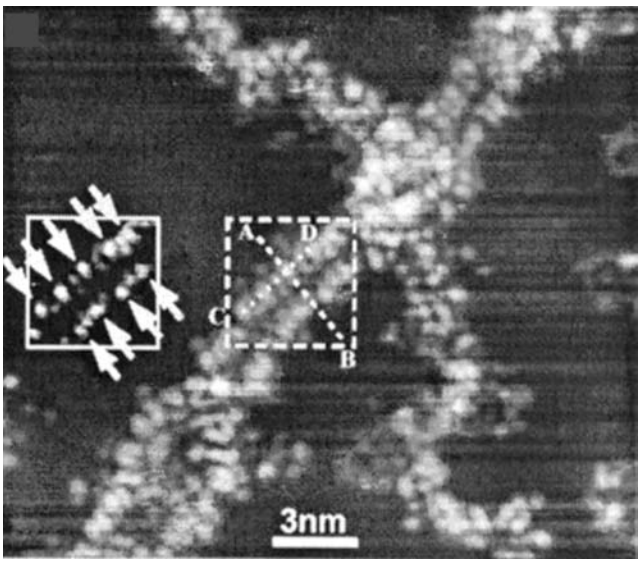


Figure 2. STM image of portion of Hetero A-T double helix of showing base pairs. (Reproduced from Ref. 9, used with permission.)

hetero A-T structures, and homo A-T structures were observed in two types, and the spot spacings and sizes of the second type would be consistent with interpretation as single-stranded DNA. The observed presence or lack of single-stranded configurations among the samples is consistent with the fact that hetero G-C has a higher melting (unraveling) temperature than the homo G-C and thus is more difficult to unwind. Both hetero and homo A-T pairs have lower melting temperatures than either of the G-C pairs. Images of both hetero A-T and Homo A-T samples often showed sizing and spacings consistent with interpretation as single-stranded DNA, in addition to observed double helix specimens. Thus, the presence/lack of single-stranded versus double helix images is consistent with known melting temperature data for the C-G and A-T base pairings.

The same group has also reported successful STM investigations of transfer-ribonuclic acid (t-RNA) (14). In

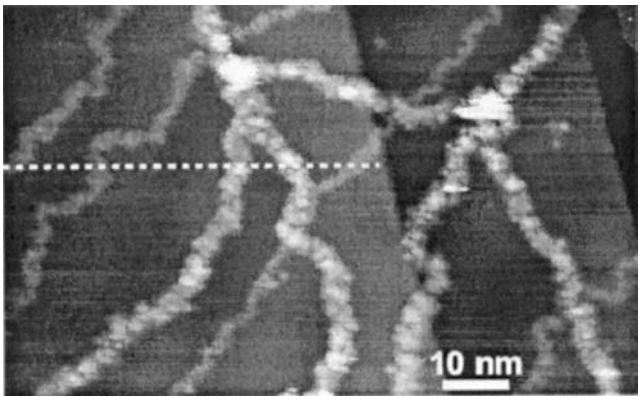


Figure 3. Hetero G-C and Hetero A-T mixed sample. The larger specimens are identified as Hetero G-C, and the smaller specimens are identified as Hetero A-T. Both are in a double helix configuration. (Reproduced from Ref. 9, used with permission.)

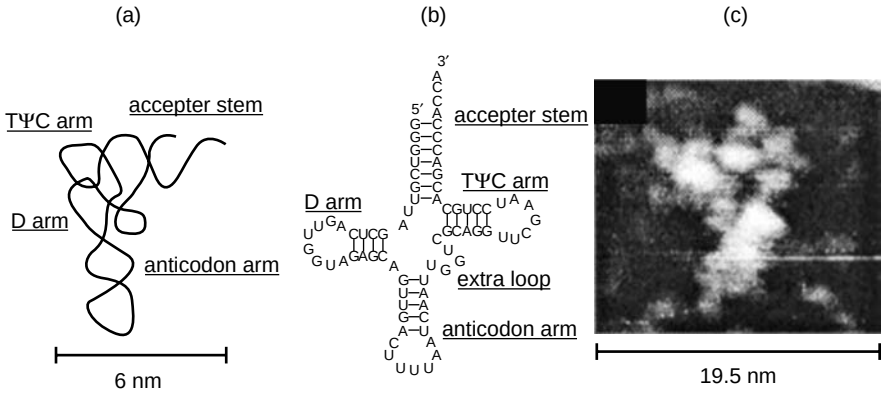


Figure 4. (a) Model of t-RNA L-shaped conformation. (b) Model base pair arrangement in L-shaped conformation. (c) STM image of L-shaped conformation at physiological pH. (Reproduced from Ref. 10, used with permission.)

RNA, the base pairing is adenine-uracil (A-U) instead of adenine-thymine (A-T). Also the backbone sugars are ribose rather than deoxyribose, but are still linked by phosphate groups. The RNA is very difficult to synthesize as a single crystal and consequently there is a very limited amount of X-ray diffraction data available for RNA. Little is known about its variations, and therefore direct investigations of single molecule RNA would add much to our knowledge.

Transfer RNA is a small RNA chain of $\sim$ 74–93 nucleotides that transfers a specific amino acid to a growing polypeptide chain at the ribosomal site of protein synthesis during translation (15). It has sites for amino acid attachment, and an anticodon region for codon recognition that binds to a specific sequence on the messenger RNA (m-RNA) chain. It has a partial double-helix structure even though it has only one chain, because the single RNA chain folds back, and loops back, on itself, as shown in Fig. 4a.

X-ray diffraction studies (16) have indicated that the t-RNA structure may often assume an L-shaped conformation with a long and a short arm. A model of the Escherichia Coli lysine t-RNA macromolecule used by the group for its STM studies is shown in Fig. 4a and b. It shows both the L conformation and the underlying loop and base pair chemistry.

Using STM, the group was able to directly image the L conformation as shown in Fig. 4c. In addition to the first direct statistical data on the lengths of the long and short arms, obtained from analysis of several STM images, an analysis of the influence of pH on conformation was also carried out. Current investigations are focusing on biofunction research issues in addition to structural research issues, using STM to directly image the coupling of the important amino acid molecules at specific t-RNA sites.

The STM investigations of nanobiomedical rather than structural issues are an important emerging research area. One example is the recently reported research from the University of Sydney group in which the local binding of retinoic acid, a potent gene regulatory molecule, to plasmid p-GEM-T easy (596 base pair Promega) DNA fragments on a single-crystal graphite substrate, was directly imaged and reported (17). Retinoic acid has been documented as responsible for a number of profound effects in cell differentiation and proliferation, and is known to accomplish its functions through selective site binding during the transcription process. The STM images of retinoic acid by itself on a single-crystal graphite substrate were investigated first. These showed sizes consistent with the retinoic acid molecular structure, and a bright head area with a darker tail area. A molecular model of retinoic acid, also shown in Fig. 5a, shows its aliphatic carbon ring head and polymeric tail. For reasons further discussed below, the aliphatic ring head may be expected to have a higher tunneling current associated with it than the polymeric tail, and therefore the observed bright and dark areas are consistent with the expected structure.

At low concentrations, retinoic acid was observed to bind selectively at minor groove sites along the DNA, with some clustering of retinoic acid molecules observed, as shown in Fig. 5b. High resolution STM imaging provided direct evidence for alignment of the retinoic acid molecules head-to-tail structure edge-on with the minor groove and also in steric alignment with each other. From STM height studies, it could also be inferred that the aliphatic ring head was attached to a ring partner along the minor groove surface, but that the tail was not attached. This may suggest a loosely bound on–off functional mechanism. At high concentrations, retinoic acid was observed to bind along the whole length of the DNA double helix, but again selecting the minor grooves. These first direct studies of selective site binding of retinoic acid with the minor groove of DNA should serve as a template for further direct investigations of other known minor groove binders, thereby opening up the direct investigation of an entire

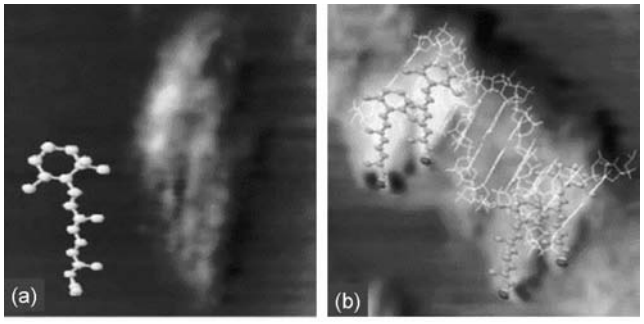


Figure 5. (a) STM image of retinoic acid on a graphite substrate compared with its molecular model showing the aliphatic ring head and polymeric tail. (b) STM image of retinoic acid binding to t-RNA with molecular model overlay. (Reproduced from Ref. 13, used with permission.)

class of regulatory molecule–DNA interactions. The interactions of related structures that are candidate therapeutic drug molecules could be receive similar direct investigation.

Note that both of the above groups have also made important contributions to sample preparation techniques for successful STM analysis of DNA and RNA. These sample preparation techniques will be discussed below in the context of the basic physics of the STM interaction, and the basic chemistry and conductivity of DNA and RNA samples.

BASIC PHYSICS OF THE STM INTERACTION

The STM is based on tip–sample interaction via a tunneling current between filled electronic states of the sample (or tip) into the empty electronic states of the tip (or sample), in response to an applied bias, as shown in Fig. 6. The bias may be positive or negative, and different and valuable information may often be obtained by investigation of the how the sample behaves in accepting, as well as in giving up, electrons. In STM imaging, it is important to recognize that the feature map or apparent topography of the acquired image is really a map of the local density of electronic states. Bright does not correspond to a raised topography; it corresponds to a region with a high density of electronic states. Therefore, in STM imaging of biological samples, an important consideration is that a differential conductivity will be observed from regions, such as rings (usually high) versus regions, such as alkane backbones (usually low). As in all SPM techniques, a z-direction feedback loop maintains some aspect of the tip samples interaction constant (Fig. 6). The readily available choices on commercial machines are to hold either the tunneling distance d constant (constant height mode) or the magnitude of the tunneling current content (constant current mode).

The current in question is a tunneling current, which is a quantum mechanical phenomenon. It is well documented that all electrons within the atomic planes of any material are in fact in such tight quarters that they display the characteristics of a wave in a waveguide, in addition to

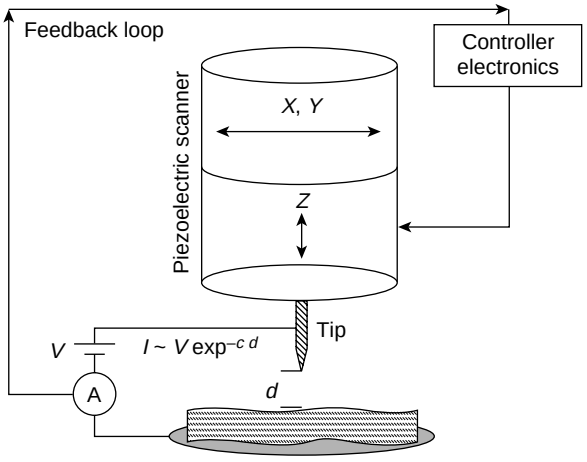


Figure 6. Important features of an STM system.

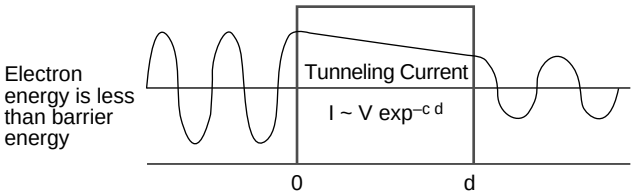


Figure 7. A particle penetrating into and through a wall.

their particle-likeness. An electron at the surface of a material faces a wall (barrier) created by the dissimilar material (e.g., air, vacuum, or a liquid). While a particle would run into a wall and bounce back, a wave can penetrate into and indeed through a wall (as light goes through glass). This is illustrated in Fig. 7. Additionally, all materials have precise energy levels within them, and therefore, electrons will move by going from one energy level at location 0 to another at location d, meeting conservation of energy requirements.

In STM, a tip with empty electronic states is brought physically close to a sample surface. The electrons are given a direction through the application of the bias (positive in this example). Because they are wavelike, when they reach the sample surface, they can tunnel through the barrier created by the 0-to-d gap and reach the empty states of the tip, where they are recorded as a current proceeding from sample to tip. A tunneling current has the known mathematical form: $I \sim V \exp^{-cd}$, where I is the tunneling current, V is the bias voltage between the sample and the tip, c is a constant and d is the tip-sample separation distance. The tunneling current depends sensitively on the size of the 0-to-d gap distance. To observe a tunneling current, the gap must be on the order of tens of nanometers. This is the case in any commercial STM system. It is remarkable, that with the addition of a simple feedback loop, a tip can be easily maintained within nanometers of a sample surface without touching it. Typical STM tunneling currents are on the order of 10^{-9}–10^{-12} A. With special preamplifiers, currents on the order of 10^{-14} A can be detected.

Because STM is a current-based technique, some situations that can interfere with its current will be briefly discussed. Very common in STM imaging of biological samples is for the tip to acquire a layer of biological material, possibly by going too close to the sample surface while passing over an insulating region where the feedback loop has little to work on. This usually just introduces image artifacts, discussed below, but it can sometimes insulate the tip from the sample, thus terminating the tip–sample interaction. The problem can be minimized through careful consideration of the expected chemistry and local conductivity of the biological specimen to be investigated.

CHEMISTRY, CONFORMATION, AND CONDUCTIVITY OF BIOLOGICAL SAMPLES

Consideration of the basic chemistry involved in a biological sample can help to determine its appropriateness for STM imaging. The building blocks for DNA and RNA are

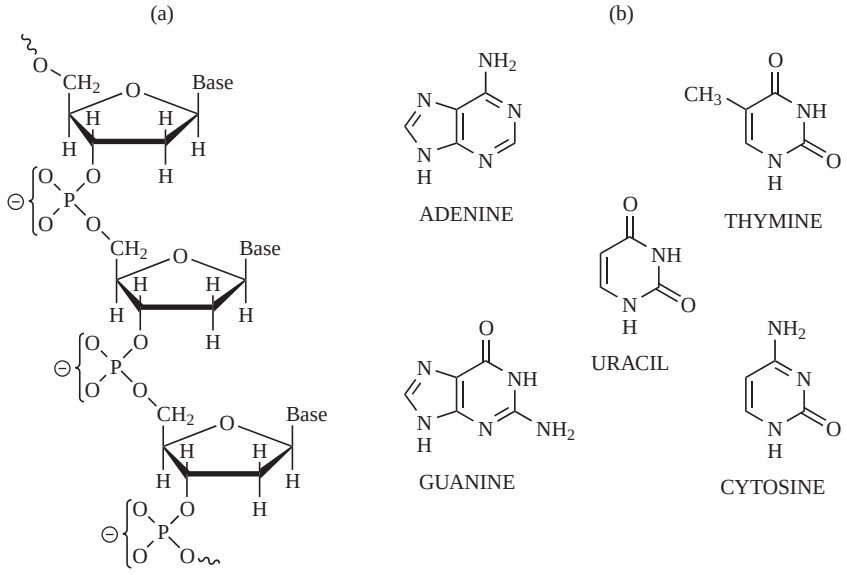


Figure 8. (a) The deoxyribose (ribose) sugar/phosphate backbone for DNA (RNA) is negatively charged due to phosphate groups. (b) DNA and RNA bases are nitrogenous ring systems.

shown Fig. 8. The sugar-phosphate backbone contains negatively charged phosphate groups for both DNA and RNA. The bases adenine, thymine, uracil, guanine, and cytosine are all nitrogenous ring systems. Thymine, cytosine, and uracil are six-member ring pyrimidine systems, and adenine and guanine are purines, the fusion of a six-member pyrimidine ring to a five-member imidazole ring. Successful STM imaging of monolayers of the individual bases has been reported (18,19). Examples of the high resolution STM imaging that is possible for monolayers of the individual bases are shown in Figs. 9 and 10.

The nitrogenous ring systems, like the classic benzene ring system, which has also been imaged (20), contain *p*-orbital electrons above and below the ring structure plane, which create a conductive electron cloud. Hence, the successful STM imaging of the DNA and RNA systems by the Osaka University and University of Sydney groups might be expected from the charged phosphate groups in the backbones and the ring systems in the base pairs.

However, there are also very difficult issues to resolve in making the local conductivity of, especially, the signature DNA and RNA base pairs available to the STM tip. These are enclosed within the sugar-phosphate backbones, and only partially exposed by the twisting of the helix, as shown in Fig. 11a and b (21,22). Also, the choice of substrate will powerfully influence the molecular structure deposited on it, especially if it is small. An example of this is shown in Fig. 12, taken from Ref. 16. The behaviors of pyridine (a single-nitrogen close relation to pyrimidine) and benzene on a single crystal (001) orientation copper, Cu(001), substrate were investigated. The pyrimidine monolayers (thymine, cytosine, and uracil) in Figs. 9 and 10 had rings oriented parallel to the substrate surface, but individual pyridine molecules on Cu(001) had rings perpendicular to the surface, due to the strong nitrogen-copper atom interaction, as shown in Fig. 12a. Also, if a single hydrogen atom was dissociated from the pyridine molecule, as can happen during routine scanning, the molecule would shift its position on the copper substrate (Fig. 12b). The STM imaging of an individual benzene molecule indicated a ring system parallel to the copper substrate (Fig. 12c), but hydrogen dissociation would cause the benzene molecule to become perpendicular to the substrate surface (Fig. 12d). Therefore both the substrate choice and interactions with the imaging tip can influence the conformation of the biomolecule and whether its locally conductive portions are positioned to produce a tunneling current.

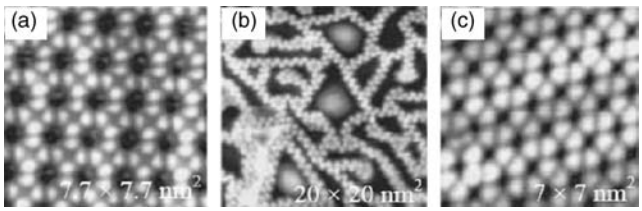


Figure 9. STM images of (a) guanine, (b) cytosine, and (c) adenine monolayers on a single crystal (111)-orientation gold substrate. (Reproduced from Ref. 14, used with permission.)

Now consider the situation of a molecule with a difference in local conductivity, like retinoic acid. The aliphatic ring head would similarly be expected to have a high local conductivity, and separate investigations of just retinoic acid by the University of Sydney group confirmed that this is the case (Fig. 5a). The polymeric tail is basically an alkane system without any *p*-type orbitals. Its conductivity is therefore expected to be less than the ring system and this is experimentally observed. However, results such as those shown in Fig. 13 from a group at California Institute of Technology, demonstrate that high resolution STM imaging even of low conductivity alkane systems is possible (23–26). Therefore, one aspect of STM biomolecular imaging is that there may be large differences in the conductivities of two closely adjacent regions. It then becomes an issue of whether the STM feedback loop will be able to sufficiently respond to the differences to maintain the tip-sample tunneling current interaction throughout the investigation. Prior consideration of the imaging parameters necessary for successful STM imaging of the *least* conductive part of the bio molecule can help.

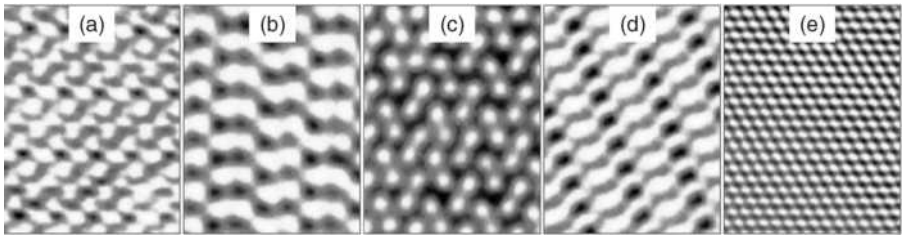


Figure 10. STM images of (a) guanine, (b) adenine, (c) uracil, and (d) thymine monolayers on (e) a single crystal (0001)-orientation molybdenum dissulfide substrate. (Adapted from Ref. 15, used with permission.)

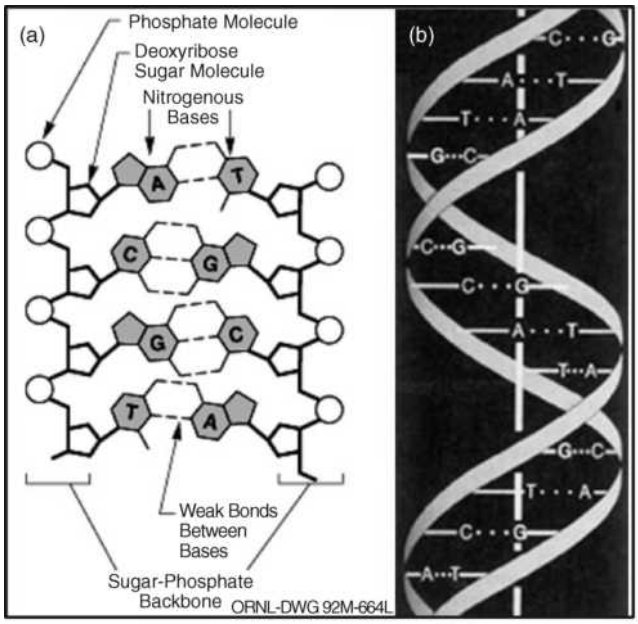


Figure 11. The three-dimensional conformation of DNA. (a) The base pairs are positioned between the sugar-phosphate backbones. (b) The overall structure is a double helix. (Reproduced from Refs. 17,18, used with permission.)

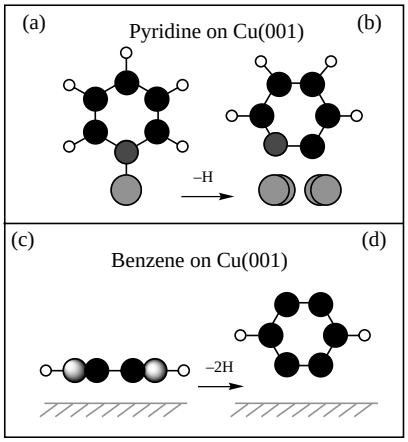


Figure 12. Influence of the sample-substrate interaction on sample orientation. (a) An individual pyridine molecule on a copper (001)-orientation, (Cu(001)) substrate is perpendicular to the surface due to the strong nitrogen–copper atom interaction. (b) An individual pyridine molecule from which a hydrogen atom has dissociated is also perpendicular to a Cu(001) surface but has a shifted location. (c) An individual benzene molecule on a Cu(001) substrate is parallel to the surface but (d) may become perpendicular if hydrogen dissociation occurs. (Adapted from Ref. 16, used with permission.)

Biomolecules, with only nanometer dimensions, always should be deposited on atomically flat single-crystal substrates. Substrates can also be selected to supply electrons to the biomolecule, for positive bias scanning, or to manipulate the biomolecule into a desired position. Another important sample preparation issue is that biomolecules often have multiple available conformations, including globular conformations that self-protect the molecule under nonphysiological conditions. While STM imaging may be performed in vacuum, air, and even in a liquid-filled compartment (liquid cell), the best resolution may be achieved in vacuum, which is a nonphysiological condition. The less physiological the imaging conditions, the more it will be necessary to use special molecular stretching techniques to investigate an open conformation. A special pressure jet injection technique was developed by the Osaka University group to deposit stretched DNA and RNA on single-crystal copper for vacuum STM imaging, without giving them the chance to close into globular conformations (13,14).

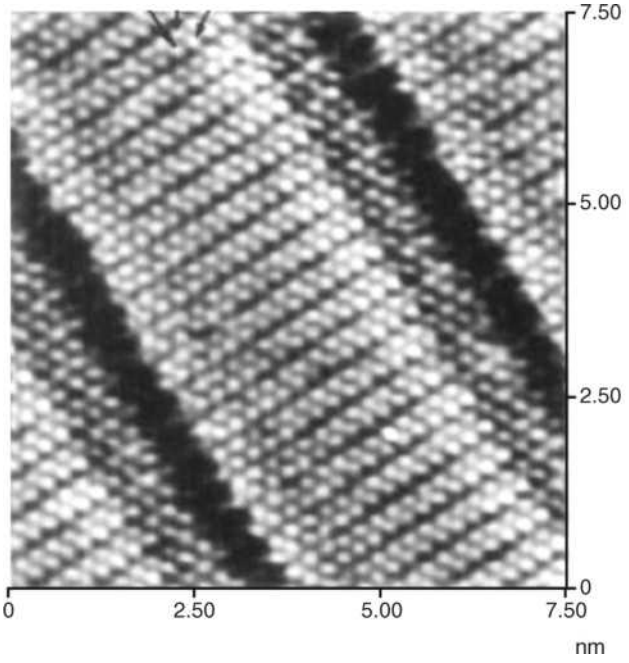


Figure 13. High resolution STM images of an alkane (pentatracontane) monolayer on graphite. (Reproduced from Ref. 19, used with permission.)

IMAGING ARTIFACTS AND DATA RESTORATION USING DECONVOLUTION

Examination of Fig. 5a shows the ring head of retinoic acid as a large blurred bright spot. Greater resolution of detail would clearly be desirable. As in all SPM imaging systems, tip artifacts versus the surface features will limit the resolution of the experiments performed. This is often cited as an ultimate barrier in STM studies of macromolecular structures and in scanning probe microscopy in general (27). It is therefore necessary to develop techniques for deconvolution of STM tip artifacts for enhancing the resolution of measured STM image data.

A commonly used approach for data restoration or eliminating the smearing effect of tip sample interaction is to assume that the observed signal is a convolution of the true image and the probe response function (PRF). The following equation gives a general degradation model due to the convolution of tip artifacts with true data resulting in the measurement *g(x,y)*. Neglecting the presence of the additive noise, the data can be modeled as

$$g(x,y) = f(x,y) * h(x,y) = \sum_{n,m} f(n,m)h(x-n,y-m)$$

where *g(x, y)*, *f(x, y)*, and *h(x,y)* are the observed or raw signal, true image, and PRF, respectively. One can then use deconvolution methods to extract the true image from the knowledge of measured data and probe PRF.

Theoretically, the probe response function is derived from the underlying physics of the tip sample interaction process. Hence, there is a need for a theoretical model for the tip sample interaction. Recent advances in formulation and modeling of tip sample interactions allow development of accurate compensation algorithms for deconvolving the effect of tip-induced artifacts.

Figure 14 shows an example of applying a deconvolution algorithm on synthetic degraded images. The degraded image in Fig. 14c is generated from a synthetic image in Fig. 14a blurred by a Gaussian PRF in Fig. 14b. Figure 14d shows the enhanced result obtained using deconvolution. Although the theoretical treatment of STM and related SPM techniques provide major challenges because the atomic structures of the tip and sample have to be modeled appropriately, its potential is clear and this is a strongly developing research area at the present time.

CONCLUSIONS

The STM imaging has the highest resolution of all SPM imaging techniques. As such, it would be highly desirable to apply STM to investigations of molecular biology and medicine. An often described difficulty when considering the application of STM to molecular biology is that biological samples are nonconductive. It would be more accurate to describe biological samples as having both local and varying conductivities. Design of STM experiments in which ring systems are exploited, and/or imaging parameters are set for the least conductive portion of the biomolecules may help produce successful imaging results. New research in applications of powerful deconvolution techniques to STM imaging will also open up the field of direct STM investigations of the structure and function of important biomolecules.

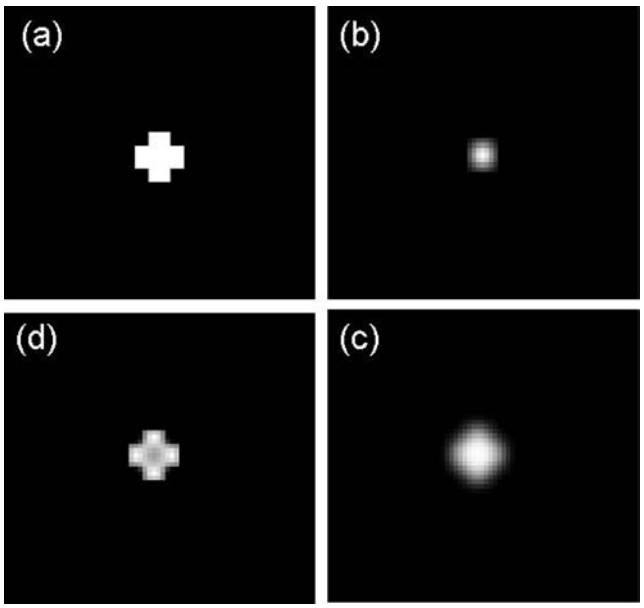


Figure 14. Clockwise from upper left: (a) synthetic true image (b) Gaussian PRF, (c) degraded measurement, and (d) restored image.

BIBLIOGRAPHY

1. Binning G, Rohrer H. Helv Phys Acta 1982;55:726–735.
2. Hansma HG, Oroudjev E, Baudrey S, Jaeger L. TectoRNA and kissing loops: Atomic force microscopy of RNA structures. J Microsc 2003;212:273–279; Sitko JC, Mateescu EM, Hansma HG. Sequence-dependent DNA condensation and the electrostatic zipper. Biophys J 2003;84:419–431; Hansma HG, Vesenka J, Siegerist C, Kelderman G, Morrett H, Sinsherimer RL, Bustanmante C, Elings V, Hansma PK. Reproducible imaging and dissection of plasmid DNA under liquid with the atomic force microscope. Science 1992;256:1180.
3. Hartmann U. Magnetic force microscopy. Annu Rev Mater Sci 1999;29:53–87.
4. Paesler MA, Moyer PJ. Near-Field Optics: Theory, Instrumentation, and Applications. New York: Wiley-Interscience; 1996.
5. Majumdar A. Scanning thermal microscopy. Annu Rev Mater Sci 1999;29:505–585.
6. Beebe TP, Jr., et al. Science 1989;243:370–371.
7. Lee G, Arscott PG, Bloomfiled VA, Evans DF. Science 1989;244:475–478.
8. Driscoll RJ, Youngquist MG, Baldescwieler JD. Atomic scale imaging of DNA using scanning tunnelling microscopy. Nature (London) 1990;346:294–296.
9. Tanaka H, Kawai T. Visualization of detailed structures within DNA. Surface Sci Lett 2003;539:L531–L536.
10. Nishimura M, Tanaka H, Kawai T. High resolution scanning tunneling microscopy imaging of Escherichia coli lysine transfer ribonucleic acid. J Vac Sci Technol B 2003;21:1265–1267.
11. Wikipedia, The free encyclopedia. Available at http://en.wikipedia.org/wiki/RNA.

12. Giege R, Puglisi JD, Floentz C. In: Cohn WE, Moldave K, eds., Progressin Nucleic Acid Research and Molecular Biology, Vol. 45. Amsterdam: Elsevier; 1993.
13. Hadi Zareie M, Lukins PB. Atomic-resolution STM structure of DNA and localization of the retinoic acid binding site. Biochem Biophy Res Commun 2003;303:153–159.
14. Otero R, et al. Proceedings of the 5th Trends In Nanotechnology (TNT04) CMP Cientifica; 2004. Available at http://www.phantomsnet.net/files/abstracts/TNT2004/AbstractKeynoteBesenbacherF.pdf.
15. Available at http://biochem.otago.ac.nz/staff/sowerby/periodicmonolayers.htm. Multiple references listed.
16. Lauhon LJ, Ho W. Single molecule chemistry and vibrational spectroscopy: Pyridine and benzene on Cu(001). J Phys Chem A 2000;104:2463–2467.
17. Image credit in Fig. 11(a): U.S. Department of Energy Human Genome Program, Available at http://www.ornl.gov/hgmis. This image originally appeared in the 1992 U.S. DOE Primer on Molecular Genetics.
18. Image credit in Fig. 11(b): Mathematical Association of America, Available at http://www.maa.org/devlin/devlin0403.html.
19. Claypool CL, et al. Source of image contrast in STM images of functionalized alkanes on graphite: A systematic functional group approach. J Phys Chem B 1997;101:5978–5995.
20. Faglioni F, Claypool CL, Lewis NS, Goddard WA III. Theoretical description of the STM images of alkanes and substituted alkanes adsorbed on graphite. J Phys Chem B 1997;101:5996–6020.
21. Claypool CL, et al. Effects of molecular geometry on the STM image contrast of methyl- and bromo-substituted alkanes and alkanols on graphite. J Phys Chem B 1999;103:9690–9699.
22. Claypool CL, Faglioni F, Goddard WA III, Lewis NS. Tunneling mechanism implications from an STM study of $H_3C(CH_2)15HC=C=CH(CH_2)15CH_3$ on graphite and $C_{14}H_{29}OH$ on MoS_2. J Phys Chem B 1999;103:7077–7080.
23. Villarubia JS. Algorithms for scanned probe microscope image simulation, surface reconstruction and tip estimation. J Res Nat Inst Standards Technol 1997;102:425–454.

See also BIOSURFACE ENGINEERING; MICROSCOPY, ELECTRON.

MICROSENSORS FOR BIOMEDICAL APPLICATIONS.

See CAPACITIVE MICROSENSORS FOR BIOMEDICAL APPLICATIONS.

MICROSURGERY

KEITH J. REBELLO
The Johns Hopkins University
Applied Physics Lab
Laurel, Maryland

INTRODUCTION

Microsurgery is a specialized surgical technique whereby a microscope is used to operate on tiny structures within the body. Fine precision microinstruments are used to manipulate tissue with or without robotic or computer control. This technique has allowed for significant advances in surgery especially for operations involving the inner ear, eye, brain, nerves, and small blood vessels.

HISTORY OF THE SURGICAL MICROSCOPE

The compound microscope is generally agreed upon to have been invented in the 1590s by the two Dutch opticians Hans and Zacharias Janssen. Their device consisted of a sliding tube with two aligned lenses. In 1624 Galileo Galilei, the famous astronomer and mathematician, demonstrated an inverted telescope to his colleagues of the Lincean Academy. One of them, Giovanni Faber, named it a microscope from the Greek words micro, meaning small, and scope, meaning to aim or shoot. Microscope lenses were improved in the seventeenth century by Antonie van Leeuwenhoek, a Dutch linen draper who originally was interested in counting the number of threads per square inch in his reams of cloth. How he constructed his spherical lenses still remains a mystery to this day. In the eighteen century, fine and course adjustments as well as tube inclination were added by Robert Hooke, who first discovered the cell. The microscope was further improved by Joseph Jackson Lister a British wine merchant, school visitor, and histologist, the father of Lord Joseph Lister whom is credited as stating the era of modern surgery. His innovations included the developed achromatic objective lens corrected for chromatic and spherical aberrations and stands designed to reduce vibrations. His jointly published work with Dr. Joseph Hodgkins in 1827, redefined the understanding at the time of arteries, muscles, nerves, and the brain.

In 1848 Carl Zeiss, a German machinist opened a microscope workshop. Ernst Abbé, a physicist working with Zeiss, derived new mathematical formulas and theories that allowed the optical properties to be mathematically predicted for the first time. Prior lenses had always been made by craftsmen who learned their trade by trial and error. Abbé's advancements enabled Zeiss to become the first mass producer of high quality microscopes.

USE IN SURGERY

Edwin Theodor Saemisch, a German ophthamologist, used loupes in surgery in 1876, but although the microscope was being used in the laboratory medical research environment it was not used the operating room. Zeiss manufactured a binocular microscope specifically designed for dissecting which was used for ophthalmological examinations of the cornea and anterior chamber of the eye. It was not until 1921 that Carl Olof Nylen, a Swedish, otologist and tennis olympian, used his homebuilt monocular microscope for the first time in ear surgery on a case of chronic otis media, a type of ear infection. His monocular microscope was quickly replaced in 1922 by a binocular microscope developed by adding a light source to a Zeiss dissecting microscope by his chief surgeon, Gunnar Holmgren. He used it to treat diseases otosclerosis, the abnormal growth of temporal bone in the middle ear.

Despite these early successes the surgical microscope was seldom used due to its limited field of vision, very short focal distance, poor light quality, and instability. It was not until the 1950s that the surgical microscope started to become more widely adopted. In 1953, Zeis released the Zeiss OpMi 1(Zeiss Operating Microscope Number One), which was specially designed for otological procedures. Its superior coaxial lighting, stability, and ease of operation enabled the advent of tympanoplasty operations to repair ruptured ear drums as well as widespread use in temporal bone surgery.

The success the microscope was having in otology soon spread to other disciplines as well. In the early 1950s, José Ignacio Barraquer adapted a slip lamp to the Zeiss otological surgical microscope for ocular microsurgery. By the 1960s, J. I. Barraquer, Joaquín Barraquer, and Hans Littman of Zeiss, had further modified the surgical microscope and refined microsurgical techniques to make ocular maneuvers in glaucoma microsurgery easier to perform. During this same time frame, Richard Troutman also had Zeiss make a special microscope for his ophthalmic procedures. He made many advances and is credited as adding electric and hydraulic control to surgical microscopes, but possibly his greatest innovation was the first variable magnification, or zoom, surgical microscope.

Around this time neurosurgeons also began using the surgical microscope in the operating room. In 1957, Theodore Kurze removed a neuriloma tumor from the seventh cranial nerve, and then later anastomized it to the hypoglossal nerve. He also developed techniques to use the surgical microscope for aneurysm surgeries. Recognizing that sterilization was a major problem, he developed the use of ethylene oxide gas to sterilize his surgical microscopes for use in the operating room. Dr. R. M. Peardon Donaghy established the first microsurgical training lab at the University of Vermont, where many surgeons were trained. He collaborated with the vascular surgeon Julius Jacobson to remove occlusions from cerebral arteries. Jacobson and his colleague Ernesto Suarez, were responsible for developing small vessel anastomoses techniques. Their procedures required another surgeon's assistance. To meet this need Jacobson invented the diploscope that allowed two surgeons to view the same operative field. Later he and his colleagues worked with Hans Littman of Zeiss to develop a commercial surgical microscope with a beamsplitter enabling two surgeons to operate at the same time. A modern day version of the Zeiss dual head microscope is shown in Fig. 1.

Inspired by the work of the neuroscientists plastic surgeon also started using the surgical microscope. Harold Buncke was one of the first plastic surgeons to use the microscope for digit/limb replantation and free-flap autoplantation. Buncke also developed many of the tools microsurgeons use by borrowing technology from the jewelry, watchmaking, and microassembly industries (Fig. 2.)

The 1960s saw the techniques applied to neurosurgery. Breakthroughs in microneurosurgery included the repair of peripheral nerve injuries, intracranial aneurysm surgeries, embolectomies of middle cerebral arteries, middle cerebral artery bypasses. One on the visionaries of this time was M. G. Yasargil a Turkish neurosurgeon from Switzerland. Trained in Donaghy's training lab he further refined and improved the surgical techniques.

Figure 1. Zeiss OpMi Vario S8 surgical microscope. (Courtesy of Carl Zeiss.)

The next advancements came with the development of minimally invasive surgical techniques. Traditional surgical techniques used relatively large incisions to allow the surgeon full access to the surgical area. This type of operation, called open surgery, enables the surgeon's hands and instruments to come into direct contact with organs and tissue, allowing them to be manipulated freely. These operations are classified as first generation surgical techniques and most surgeons are trained in this manner. While the large incision gives the surgeon a wide range of motion to do very fine controlled procedures, it also causes substantial trauma to the patient. In fact, the majority of trauma is caused by the incisions the surgeon uses to get access to the surgical site instead of the actual

Figure 2. Modern day microsurgical tools. (Courtesy of WPI Inc.)

surgical procedure itself. For example, in a conventional open-heart cardiac operation, the rib cage must be cracked and split exposing the heart muscle. This trauma not only increases pain to the patient, but adds to recovery times increasing hospital stays, in turn increasing costs.

In 1985, Muhe performed the first laparoscopic cholecystectomy, or gallbladder removal surgery with a fiberoptic scope. The technique he performed is commonly called minimally invasive surgery but also goes by other names, such as micro, keyhole, microscopic, telescopic, less invasive, and minimal access surgery. This microsurgical technique is based on learnings from gynecological pelviscopies and arthroscopic orthopedic operations along with the previous advances made in otology, opthamology, neurosurgery, and reconstructive microsurgeries. It has subsequently been applied to many other surgical areas, such as general surgery, urology, thoracic surgery, plastic surgery, and cardiac surgery. These procedures are classified as second generation surgeries as trauma to the patient is drastically reduced by the reducing or eliminating incisions. The shorter hospital stays and faster recovery times for the patient reduce the cost of a minimally invasive procedure 35% compared to its open surgery counterpart.

In a minimally invasive cardiac operation a few small holes, access points, or ports are punctured into the patient and trocars are inserted. A trocar consists of a guiding cannula or tube with a valve–seal system to allow the body to be inflated with carbon dioxide. This is done so that the surgeon has enough room to manipulate his instruments at the surgical site. An endoscope is inserted into one of the trocar ports to allow the surgeon a view the surgical site. Various other surgical instruments, such as clippers, scissors, graspers, shears, cauterizers, dissectors, and irrigators were miniaturized and mounted on long poles so that they can be inserted and removed from the other trocar ports to allow the surgeon to perform the necessary tasks at hand.

While minimally invasive surgery has many advantages to the patient, such as reduced postoperative pain, shorter hospital stays, quicker recoveries, less scarring, and better cosmetic results, there are a number of new problems for the surgeon. The surgeon's view is now restricted and does not allow him to see the entire surgical area with his eyes. While the operation is being performed he must look at a video image on a monitor rather than at his hands. This is not very intuitive and disrupts the natural hand–eye coordination we all have been accustomed to since childhood. The video image on the monitor is also only two dimensional (2D) and results in a loss of our binocular vision eliminating the surgeon's depth perception. While performing the procedure the surgeon does not have direct control of his own field of view. A surgical assistant holds and maneuvers the endoscopic camera. The surgeon has to develop his own language to command the assistant to position the scope appropriately, which often leads to orientation errors and unstable camera handling, especially during prolonged procedures. Since the images from the camera are magnified, small motions, such as the tremor in a surgical assistant's hand or even their heartbeat can cause the surgical team to experience motion induced nausea. To combat the endoscopic problems, some surgeons choose to manipulate the endoscope themselves. This restricts them to using only one hand for delicate surgical procedures and makes procedures even more complicated.

The surgeon also loses the freedom of movement he has in open surgery. The trocar ports are fixed to the patient's body walls by pressure and friction forces. This constrains the instrument's motion in two directions and limits the motion of the tip of the instrument to four degrees of freedom (in–out, left–right, up–down, and rotation). The trocars also act as pivot points and cause the surgical instruments to move in the opposite direction to the surgeon's hands. When the surgeon is moving left, the image on the monitor is moving to the right. The amount of this opposite movement also depends on the depth of the introduction of the instrument. Again because of the pivot point the deeper an instrument is inserted into the body the more the surgeon's movement is amplified. Even a small movement made by the surgeon on the outside of the patient can translate to a very large movement on the inside of the patient. The seals and valves in the trocars also impede movements which hinders the smoothness of motions into and out of the patient and greatly reduces the already limited tactile feedback the surgeon experiences. These movement behaviors and lack of tactile feedback are counter to what the surgeon is used to in open surgery and require long training to develop the technical skills to perform these operations.

Performing a minimally invasive procedure has been likened to writing your name holding the end of an 18 in. (45.72 cm) pencil (1). The surgeon loses three-dimensional (3D) vision, dexterity, and the sense of touch. The instruments are awkward, counterintuitive, and restricted in movement. The lack of tactile feedback prevents the surgeon from knowing how hard he or she is pulling, cutting, twisting, suturing, and so on. These factors cause a number of adjustments to be made by the surgeon, which requires significant retraining on how to do the procedures in a minimally invasive manner. These difficulties encountered by the surgeon cause degradation in surgical performance compared to open surgery which limits surgeons to performing only simpler surgical procedures.

In an attempt to address some of these shortcomings and allow the surgeon more control during operations a third generation of surgical procedures, robotic surgery was developed. Although these types of procedures are commonly referred to as robotic surgery, the operations themselves are not completely automated and are still carried out by a surgeon. For this reason, robotic surgery is also referred to as computer aided or computer assisted surgery.

The robotic technology was originally developed for telerobotic applications in the late 1980s for the Defense Advanced Research Project Administration (DARPA) by researchers at SRI International. The surgeon of the future would allow surgeons from remote command centers to operate on injured soldiers in the battlefield. In 1995, this technology was spun off into a new company named Intuitive Surgical to commercialize the technology for use in the hospital environment. Near the same time Dr. Yulan Wang was developing robotic technology for NASA to allow

Figure 3. Intuitive Surgical da Vinci robotic surgery system. (Copyright ©2005 Intuitive Surgical, Inc.)

surgeons on earth to deal with medical emergencies on the international space station. He formed Computer Motion in 1989. Both of these companies merged in 2003, and Intuitive Surgical is now the leader in Robotic Surgery. In 2002, Dr. Fredric Mol, one of the original founders of Intuitive Surgical, founded Hansen Medical which brings computerized robotic control of catheters to electrophysiology and interventional cardiac procedures.

Current robotic surgery systems have a number of benefits over conventional minimally invasive surgery. Figure 3 shows an Intuitive Surgical da Vinci robotic system. In this arrangement the surgeon sits comfortably at a computer console instead of having to stand throughout the entire procedure, which can last up to 5 h long. A three-armed robot takes his place over the patient. One arm holds an endoscope while the other two hold a variety of surgical instruments. The surgical team can also look at a video monitor to see what the surgeon is seeing. The surgeon looks into a stereo display in much the same way as looking though a surgical microscope and manipulates joystick actuators located below the display. This simulates the natural hand–eye alignment he is used to in open surgery, (Fig. 4). Since computers are used to control the robot and are already in the operating room, they can be used to give the surgeon superhuman like abilities. Accuracy is improved by employing tremor cancellation algorithms to filter the surgeon's hand movements. This type of system can eliminate or reduce the inherent jitter in a surgeon's hands for operations where very fine precise control is needed. Motion scaling also improves accuracy by translating large, natural movements into extremely precise, micromovements. A wide variety of surgical instruments or end effectors are available including graspers, cutters, cauterizers, staplers, and so on. Both companies provide end effectors that have special wrist like joints at their tips enabling full seven degree of freedom movements inside the patient, (Fig. 5), but still lack tactile feedback.

These robotic advances allow surgeons to perform more complex procedures, such as reconstructive cardiac operations like coronary bypass and mitral valve repair that cannot be performed using other minimally invasive techniques.

Figure 4. Intuitive Surgical stereo display and joysticks. (Copyright ©2005 Intuitive Surgical, Inc.)

Figure 5. Multidegrees-of-freedom end effector. (Copyright ©2005 Intuitive Surgical, Inc.)

MEMS

Around the same time that minimally invasive surgery was being developed, there was a turning point in microelectromechanical systems (MEMS). This a technology was developed from the integrated circuit industry to create miniature sensors and actuators. Originally, these semiconductor processes and materials were used to build electrical and mechanical systems, but have now expanded to include biological, optical, fluidic, magnetic, and other systems as well. The term MEMS originated in the United States and typically contain a moving or deformable object. In Europe, this technology goes by the name microsystems technology or microstructures technology (MST) and also encompasses the method of making these devices, which is referred to as micromachining. In Japan and Asia MEMS are called micromachines when mechanisms and motion are involved. The MEMS devices first were used in medical applications in the early 1970s with the advent of the silicon micromachined disposable blood pressure sensor (2), but it was not until the mid-1980s when more complicated mechanical structures, such as gears and motors, were able to be fabricated.

FABRICATION TECHNOLOGIES

The fabrication of MEMS devices is based on the merger of semiconductor microfabrication processes and micromachining techniques to create the desired microstructural components. There are four major processes that are used to fabricate MEMS devices: bulk micromachining, surface micromachining, LIGA, and precision machining. Combinations of these technologies are what allow MEMS to be highly miniaturized and integratable with microelectronics. These processes are very sensitive to impurities and environmental conditions such as temperature, humidity, and air quality. Typically, these fabrication steps are performed inside a cleanroom (Fig. 6). Bulk micromachining, surface micromachining, and LIGA have the added advantage of being able to be batch fabricated. This allows many devices to be made in parallel at the same time greatly reducing device cost.

Bulk micromachining utilizes wet- or dry-etch processes to form 3D structures out of the substrate. These subtractive processes produce isotropic or anisotropic etch profiles in material substrates, which are typically but not limited to silicon wafers. Bulk micromachining can create large MEMS structures on the micrometers (μm) to millimeters (mm) scale (tens of μm-to-mm thick). Commercial applications of bulk micromachining have been available since the 1970s. These applications include pressure sensors, inertial sensors such as accelerometers and gyros, and microfluidic channels and needles for drug delivery.

In surface micromachining, MEMS are formed on the surface of the substrate using alternating layers of structural and sacrificial materials. These film materials are repeatedly deposited, patterned, and etched to form structures that can then be released by removing sacrificial layers. The release process allows for the fabrication of complex movable structures that are already assembled,

Figure 6. Cleanroom fabrication facility. (Courtesy of Intel Corp.)

such as motors, switches, resonators, cantilevers, and so on. Surface micromachined structures are typically limited to thicknesses of 2–6 μm and because they use much of the same technology as is used in the integrated circuit industry are readily integrated with electronics. Because so much technology is shared with the IC industry silicon wafers are typical substrates with thousands of devices being able to be fabricated at once (Fig. 7).

Lithographie, Galvanik, Abformung (LIGA) is a German acronym that means lithography, electroforming, and molding. The technology was originally developed in the late-1970s to fabricate separation nozzles for uranium enrichment. This technology uses X rays to fabricate devices with very high aspect ratios. A synchrotron radiation source is used to define small critical dimensions in a poly(methyl methyacrylate) (PMMA), mold that can then be electroplated to form high aspect ratio metallic structures. Many parts can be batch fabricated in this manner, but assembly is usually still a serial process.

Precision machining technology, such as micro-EDM (microelectro discharge machining), laser micromachining, and micro stereo lithography, is also used to form complex structures out of metal, plastic, and ceramics that the previous fabrication technologies may be incapable of. Precision machining is typically a serial process, but is

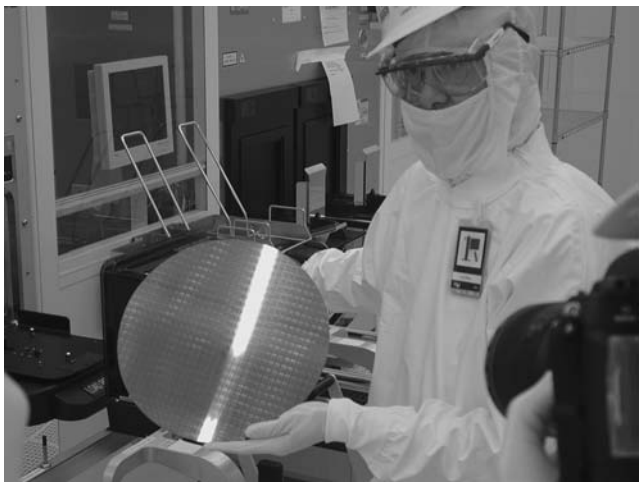


Figure 7. Silicon wafer in the fabrication process. (Courtesy of Intel Corp.)

often better able to deal with the varied shapes and substrates of microsurgical instruments.

Micro EDM is a form of spark machining used to shape conductive materials, such as silicon and metals. An EDM erodes material by creating a controlled electric discharge between an electrode and the substrate. It is a noncontact process and there is no direct mechanical cutting force applied to the substrate. Dielectric fluid is used to remove the erosion particles, as well as to keep the substrate material from oxidizing. Micro-EDMs can be used to make holes, channels gears, shafts, molds, dies, stents, as well as more complex 3D parts such as accelerometers, motors, and propellers (3).

Lasers can be used to both deposit and remove material. Laser ablation vaporizes material through the thermal noncontact interaction of a laser beam with the substrate. It allows for the micromachining of silicon and metals, as well as materials that are difficult to machine using other techniques such as diamond, glass, soft polymers, and ceramics. Laser direct writing and sintering is a maskless process where a laser beam is used to directly transfer metal materials onto a substrate. This can be used to form metal traces on nonplaner surfaces, which reduces the need for wires on surgical tools (4).

Micro stereo lithography processes generate 3D structures made out of ultraviolet (UV) cured polymers. It is an additive process where complex 3D structures are made from stacks of thin 2D polymer slices that have been hardened from a liquid bath. Conventional systems were limited in that they were a serial process where only one part could be made at a time. MicroTEC has developed a batch fabricated wafer level process called rapid material product development (RMPD), which is capable of constructing structures out of 100 different materials including plastics, sol–gels, and ceramics (5).

APPLICATIONS

The inclusion of MEMS technology in microsurgery, will allow for smaller more miniaturized surgical tools that not only overcome many of the limitations of microsurgical procedures, but allow for new more advanced operations to be performed. MEMS are just now being incorporated into microsurgical tools and coming on the market. Most are still at the research level, but the industry is moving in this direction as the need for smaller smarter tools increases.

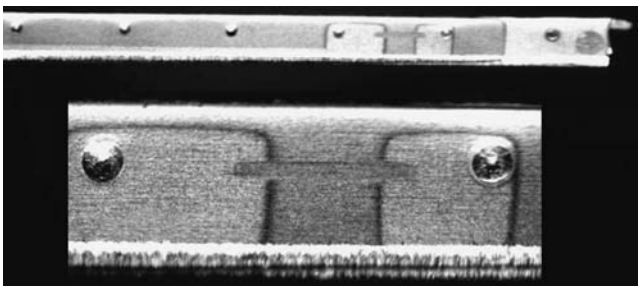


Figure 8. Strain gauges fabricated on surgical sharps. (Courtesy of Verimetra, Inc.)

HAPTIC FEEDBACK

One of the key areas for improvement in microsurgery is tactile feedback. The lack of tactile sensing limits the effectiveness these procedures. Recent work in robotic feedback for minimally invasive surgery has concentrated on force feedback techniques using motors and position encoders to provide tactile clues to the surgeon. In these approaches, the sense element is far removed from the sense area. Verimetra, Inc. has developed strain gauge force sensor fabrication technology which uses the surgical tools themselves as a substrate (6). Prior efforts have focused on fabrication of sensors on silicon, polyimide, or some other substrate followed by subsequent attachment onto a surgical tool with epoxy, tape, or some other glue layer. Attaching a sensor in this manner limits performance, introduces sources of error, limits the sensor's size, and further constrains where the sensor can be placed. By eliminating glue and adhesion layers improved sensitivity and reduces errors due to creep. Figure 8 shows strain gauges fabricated on surgical sharps. Figure 9 is a cut away SEM image of a strain gauge and temperature sensor embedded inside of a robotic microforcep. While this microfabrication technology is an improvement in sensor technology, wires are still used to connect the sensor to the outside world. Reliability and the added complexity of adding wires to surgical end effectors with high degrees

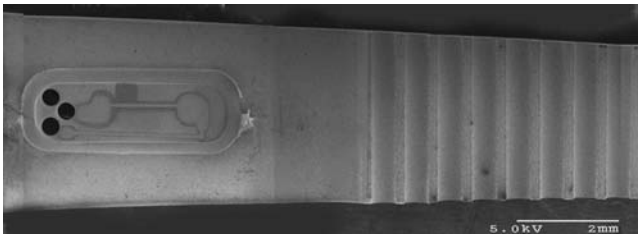


Figure 9. Strain gauges and temperature sensor embedded in robotic microgripper. (Courtesy of Verimetra, Inc.)

of freedom limit the effectiveness of the technology. Short-range wireless technology compatible with the operating room environment need to be developed to overcome these limitations.

Recently tactile feedback has been shown to be able to be added to noncontact lasers. Based on optical distance measurements, the systems synthesize haptic feedback through a robotic arm held by the surgeon when the focal point of the laser is coincident with a real surface. This gives the operator the impression of touching something solid. By increasing the power of the laser such a system could also be used for cutting or ablation.

TISSUE SENSING

Taking haptic feedback one step further is the ability to distinguish between different types of tissue in the body. Tissue sensing is of vital importance to a surgeon. Before making an incision into tissue, the surgeon must identify what type of tissue is being incised, such as fatty, muscular, vascular, or nerve tissue. This becomes more complicated because the composition and thickness of different human tissues varies from patient to patient. Failure to properly classify tissue can have severe consequences. For example, if a surgeon fails to properly classify a nerve and cuts it, then the patient can suffer effects ranging from a loss of feeling to loss of motor control. If a neurosurgeon cuts into a blood vessel while extracting a tumor severe brain damage may occur. The identification and classification of different types of tissue during surgery, and more importantly during the actual cutting operation, will lead to the creation of smart surgical tools. If a surgical tool senses that it is too close to or about to cut the wrong type of tissue it can simply turn itself off.

Verimetra, Inc. has developed a device called the data knife, Fig. 10. It is a scalpel, which is outfitted with different strain sensors along the edges of the blade to sense the amount of force being applied. The resistance of the tissue is one of the signals used for classifying tissue. Pressure sensors are used to measure the characteristics of material surrounding the blade. The pressure of the surrounding fluid can be used to help classify the type or location of tissue. Electrodes are used to measure the impedance of different types of tissue, as well as being used to identify nerves by picking up their electrical signals. The tool provides the real-time feedback surgeons have been asking for during surgery, and can also be used to record data for later use for tracking purposes.

Sensing the density of tissue can also be used to assist the surgeon in identifying tissue. In open cardiac bypass operations, the surgeons insert their hands inside the body to palpate arteries. For cardiac bypass surgery, surgeons select the bypass location by feeling where the fat and fatty plaque is located in your arteries with their fingers. The lack of tactile feedback in minimally invasive surgery, prevents them from using this technique. The MEMS devices have been developed for the palpation of tissue using strain gauges (7), diaphragms (8), micropositioners (9,10), and load cells (11) and have shown the ability to measure blood pressure, pulse, different kinds of arterial plaque, and distinguish between colon, bowel, stomach, lung, spleen, and liver tissue.

Piezoelectric transducers can also be used to measure density. Macroscale transducers are frequently used in imaging applications to differentiate between tumors, blood vessels, and different types of tissue. These transducers both emit and receive sound waves. By vibrating at a high frequency sound waves are emitted in the direction of the object of interest. The density of the impinged object can then be measured based on the signal that is reflected back by that object. Sound waves are reflected off the interfaces between different types of tissue and returned to the transducer. Present ultrasound systems tend to be large and are not well suited for incorporation into microsurgical devices. The MEMS technology is well suited for this application and many ultrasonic MEMS sensors have been developed for imaging (12–16).

Microelectromechanical systems ultrasound devices for density measurements are shown in Fig. 11. They have been shown to be able to detect the location of bone in tissue and are being applied to atrial fibrillation surgeries. Atrial fibrillation is what causes irregular heartbeats and leads to one out of every six strokes. Drugs can be used to treat this condition, but have dangerous side effects including causing a switch from atrial fibrillation to the more dangerous ventricle fibrillation. Pacemakers and other electrical control devices can be used, but they do not always work for all patients. The most effective treatment is the surgical

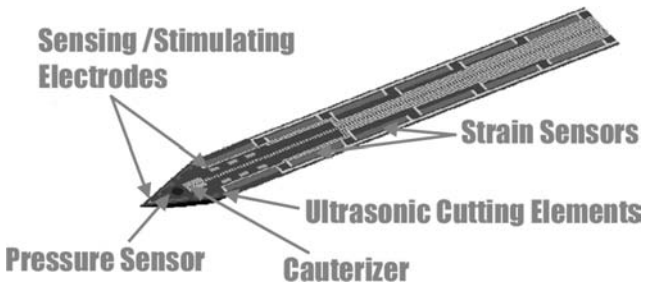


Figure 10. Data Knife smart scalpel. (Courtesy of Verimetra, Inc.)

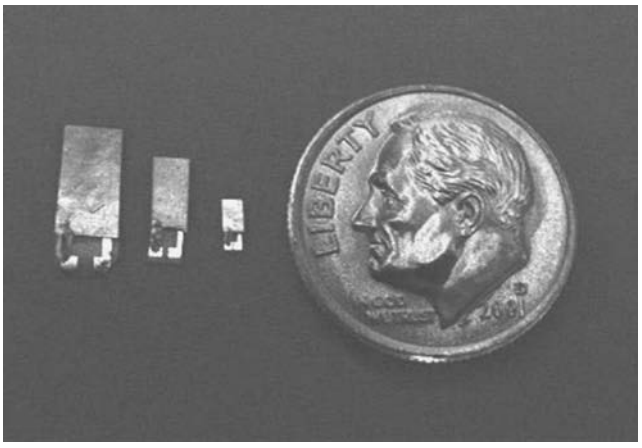


Figure 11. Ultrasound transducers next to a dime. (Courtesy of Verimetra, Inc.)

MAZE procedure, but it is an incredibly invasive treatment. The patient is put on a heart lung machine and then their heart is stopped. Next, the surgeon takes a scalpel and actually cuts all the way through the heart making lesions, which physically separate the heart muscle. These lesions break the electrical connections in the heart. The heart is then sewn back together. Recently, there have been a variety of different methods used to address the problem. Instead of physically cutting all the way through the heart with a scalpel, surgeons are using radio frequency, microwave, cryo, and laser energy to create transmural lesions. Transmurality means that the lesions go all the way through the tissue, breaking the heart's electrical connections. One of the problems surgeons encounter is to know how deep the ablation is or if it is transmural. If the lesions are not completely transmural or just partially transmural then the undesirable electrical signals may still be transmitted to the heart muscle. The MEMS ultrasound technology or micromachined electrodes can be used to measure the transmurality.

Temperature can be used to detect if a surgical device is close to a blood vessel, or if the surgical tool is in a diseased or infected area. Temperature can also be used to monitor the usage of a surgical device, by monitoring the time at which the device is at body temperature. Usage is just one of many areas where Auto-ID technologies will benefit microsurgery (17). They can be used to make sure that only the correct surgical tool is used for a procedure and if that tool has been properly sterilized. Keeping track of how many times, how long, and what was done with a surgical tool will improve the reliability and effectiveness of surgical procedures and will greatly impact the entire medical industry.

TRACKING SYSTEMS

Traditionally, a surgeon uses an endoscope in minimally invasive surgery to determine where the surgical instruments are located in the patient. The view the surgeon has of the surgical area is not ideal and the position of surgical instruments outside of the camera view is not known. Ideally, the surgeon would like to know the position and orientation of each of his instruments. Computer-aided surgery has enabled the surgeon to overlay magnetic resonance imaging (MRI) or computed artial tomography (CAT) scan images of the patient with position and orientation data taken during surgery to create 3D models that the surgeon can use to better visualize the surgical procedure. Computers can be used to simulate the procedure beforehand allowing the surgeon to practice difficult operations ahead of time.

Current technology in this area is predominately optical in nature. Markers are placed on the ends of the surgical instruments that are located outside of the body, as well as on specific locations on the patient's body. A computer registers the location of the surgical tools with the reference markers on the patient so that images of the patient's body can be aligned with the surgical tools. This is done through the use of visible and infrared (IR) cameras. The tips of the surgical tools that are located inside of the body are then extrapolated. The markers must not interfere with the surgery in any way, and therefore should be as small and lightweight as possible. While these systems are wireless and do not have cords that can get tangled on the surgeon or on the surgical tools, there must be an unobstructed path from the markers to the camera systems. The surgeon must be careful not to block the markers themself or with other surgical instruments. Precision is compromised because the location of the surgical tips is extrapolated and does not take into account bending of the surgical tools. Markers on the outside of the body do not take into account compression of the tissue.

MEMS based ultrasound tracking systems have been developed to address these issues (18). Constellations of ultrasound sensors can also be placed on the surgical tools themselves to determine position thereby eliminating errors from extrapolation. Reference markers can now be placed inside of the body, closer to the surgical area so that they are less affected by compression and movement of the patient.

Position and orientation can also be estimated using accelerometers and gyroscopes. The signal outputs can be integrated to determine or predict the distance traveled by a surgical tool. Conventional MEMS accelerometers have accuracies in the milligram range, which are not sufficient for measuring accurately the relatively small displacements made during surgery (19). More accurate inertial sensors need to be developed before they can be integrated into surgical tools.

Magnetic field sensors can also be used to determine position and orientation of surgical tools (20,21). A three axis magnetoeffect sensor has been developed for determining the location of catheters. Currently this is done by continually taking X rays of the patient. However X rays only provide a 2D snapshot, a 3D image would be more helpful for navigation.

EYE SURGERY

The leading cause of vision loss in adults > 60 are cataracts. The word cataract comes from the Greek meaning waterfall and was originally thought to be caused by opaque material flowing, like a waterfall, into the eye. The condition is actually caused by the clouding of the eye's intraocular lense. In the eye, light passes through the lens that focuses it onto the retina. The retina converts the light into electrical signals that are then interpreted by the brain to give us our vision. The lens is a hard crystalline material made mostly of water and protein. The protein is aligned in such a way to allow light to pass through and focus on the retina. When proteins in the lens clump together, the lens begins to cloud and block light from being focused on the retina. This causes vision to become dull and blurry, which is commonly referred to as a cataract.

Much like other surgery in other parts of the body, cataract surgery has followed down the minimally invasive path for the same benefits. Cataract surgery is one of the earliest known surgical procedures. The earliest evidence is the written Sanskrit writings of the Hindu surgeon

Susrata dating from the fifth century BC. He practiced a type of cataract surgery known as couching or reclination, in which a sharp instrument was inserted into eye and pushed the clouded lens out of the way. This displacement of the lens enabled the patient to see better. Although vision was still blurred without corrective lenses, many famous people underwent this procedure including the artists Michelangelo, Rembrandt, and Renoir. Couching was still performed until the mid-twentieth century in Africa and Asia.

In 1748, Jacques Daviel of Paris introduced extracapsular surgery where the lens was removed from the eye. Later, very thick pairs of glasses were used to focus the light onto the retina and restore sight, but the glasses were cumbersome and caused excessive magnification and distortion.

By 1949, Dr. Harold Ridley of England, used PMMA as the first intraocular lens. He discovered that PMMA was biocompatible with the eye while treating WWII fighter pilots whose eyes were damaged by shattering plastic from their windshields. In the 1960s and 1970s, extracapsular surgery became the preferred treatment. A large incision (10–12 mm) was made in the eye to remove and replace the lense. This procedure minimized problems with image size, side vision, and depth perception, but the large incisions required longer hospitalization, recovery time, and stitches.

Today, cataracts are removed with a procedure called phacoemulsication with ~1.5 million operations performed yearly. A hollow ultrasonically driven titanium needle is inserted into the anterior chamber of the eye. Ultrasonic energy is then used to liquefy the hard lens and it is then aspirated out of the eye. A foldable lens made of acrylic or silicone is inserted through a (1–3 mm hole) as a replacement. Since the incision size has been reduced compared to conventional extracapsular surgery, hospitalization, general anesthesia, sutures and bandages have all been eliminated. The reduction in incision size has also reduced the risk of infection and postoperative refractions.

During the procedure the surgeon cannot see directly under the needle as the lens is broken up and aspirated. A thin clear membrane or capsule surrounds the lense. The posterior capsule tissue underneath the lens is very delicate and easily cut compared the crystalline lens. To prevent the soft underlying tissue from damage requires a skilled surgeon whom has performed many procedures. If the posterior capsule is ruptured it can lead to glaucoma, infection, or blindness. As the size of the incision has decreased, heat damage to surrounding tissue from the ultrasonic tip has increased that can alter the characteristics of the tissue and change its appearance. In addition positive intraocular pressure must be maintained by balancing the flow of infusion fluid at positive pressure and the aspirated cataract lens fragments. If pressure is not maintained the anterior chamber can collapse. Pressure is currently maintained by sensors located many feet from the surgical area. This distance creates delays in the feedback loop, which can cause dangerous pressure fluctuations leading to damage to the iris and cornea.

Recently, micromachined silicon ultrasonic surgical tools for phacoemulsifiaction have been developed by Lal's research group (22,23) (Fig. 12). Piezoelectric material is attached to a micromachined silicon needle. The needle has a fluid channel for aspiration as well as a horn for amplifying the ultrasonic displacement. These silicon devices are able to generate higher stroke velocities and lower heat generation than their conventional titanium counterparts. High levels of integration has been achieved by integrating pressure and flow sensors directly on the needle for maintaining intraocular pressure, reducing delays and making the phacoemulsification procedure safer.

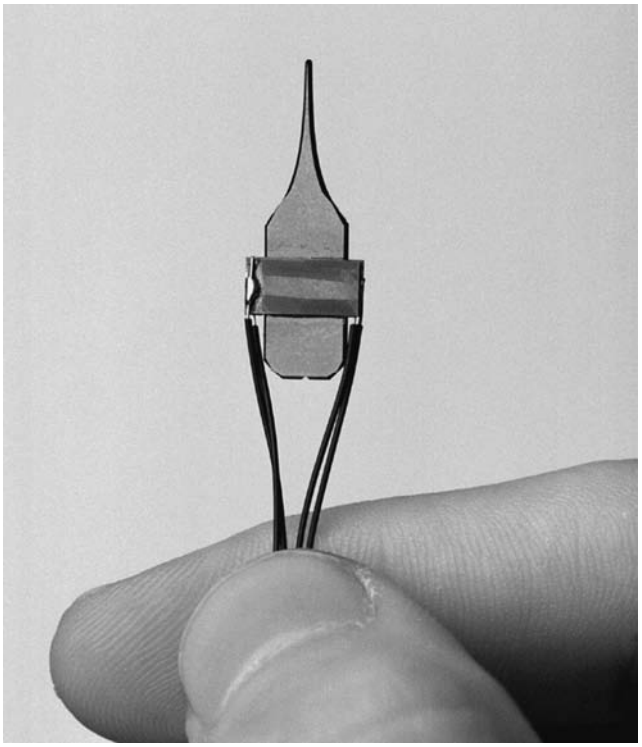


Figure 12. Ultrasonic phacoemulsification tool. (Courtesy of Jeff Miller/University of Wisconsin-Madison.)

To prevent damage to the posterior capsule a piezoelectric sensor has been integrated into a phacoemulsification hand piece and undergone clinical trials (24). The device can determine tissue hardness by measuring the impressed loading on the needle tip or by monitoring the resonant frequency at which the ultrasonic system oscillates. Both of these approaches have proven successful in determining when a hard to soft tissue transition has occurred during a phacoemulsification procedure. This technology enables a surgeon to get real-time feedback on the type of tissue he is cutting, and can be applied to other types of surgical procedures such as tumor extraction as well.

Insertion of a replacement lense requires precise movements by the surgeon. Serious postoperative vision problems may occur if the lens is inserted incorrectly and needs to be removed. Precision piezoelectric micromotors have been developed for intraocular delivery of replacement lenses after cataract removal (25). These inchworm actuators use a glider and clamp arrangement to generate large forces over small displacements. An electrostatic clamp made of an oxide dielectric layer sandwiched

between two silicon wafers layer locks the micromotor in place while a PZT actuator generates the force. The inertia of a mass is used to move the clamp. Forces of 3.0 *N* and step sizes of 100 nm to 10 μm have been reported.

In eye surgery there are many times when precision cutting is required. Highly sharpened steel, ceramic, or diamond scalpel blades are used, but are expensive to produce costing up to $1000 a piece. Disposable silicon micromachined scalpels are an attractive alternative. They can be batch fabricated to reduce cost and sharpened to an atomic edge along their crystal planes. They are already being used in Russia at the Fyodorov Eye Care Center in nonpenetrating deep sclerectomy operations for the treatment of glaucoma (26). BD (Bekton, Dikenson, and Company) is producing Atomic Edge silicon blades for cataract surgery (27). Soon they will be used for eye operations and eventually migrate to procedures on other parts of the body. Smaller incisions made by sharper blades result in less bleeding and tearing. An added advantage of silicon blades is that sensors and electronics can be directly fabricated on them during fabrication. Integrating cauterizing electrodes on the blade itself will prevents the patient from bleeding as well as let the surgeon more clearly see the surgical area.

CATHETERS/GUIDEWIRES/STENTS

Cardiac catheterizations can be referred to as a noninvasive surgical procedure. Specialized tubes or catheters are threaded up through blood vessels in the groin, arm, or neck to an area of the body which needs treatment. The problem is then treated from the inside of the blood vessel. The advantage of these approaches is that the procedures require very small incisions, hospital stays are usually one night or less and the discomfort and recovery times afterwards are minimal. For patients with more complicated ailments, catheter treatments can be used in combination with minimally invasive or open surgery to give the best possible results at the lowest possible risk. Catheters, guidewires, and stents currently represent the most widespread use of MEMS technology in surgery.

Diagnostic catheterizations, which are used to measure pressure in different parts of the body, take blood samples, and to perform detailed angiograms of the heart, can be performed by injecting X-ray dye through the catheters. The MEMS pressure sensors are now commonly found on catheter devices and are the most mature MEMS technology in this area. Even smaller designs are being sold for placement on guidewires, such as those made by Silex Microsystems, shown next to a 30 gauge needle (Fig. 13). Each sensor is but 100 μm thick, 150 μm wide, and 1300 μm long (28). MEMS ultrasound sensors are also starting to be used for both forward looking (16) and side looking intravascular imaging (12,13).

To provide the doctor with more information to make better decisions, additional MEMS sensors are needed to gather additional data for the diagnosis, monitoring of procedures, as well as for checking results of completed operations. Many other types of MEMS sensors are being researched to measure blood flows, pressures, temperatures, oxygen content, and chemical concentrations for placement on diagnostic catheters (29,30).

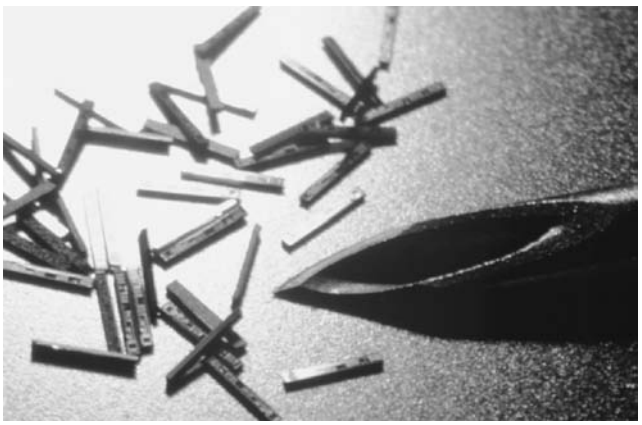


Figure 13. MEMS pressure sensors next to a 30-gauge needle (Courtesy of Silex Microsystems, Jarfalla, Sweden.)

Heart disease continues to be the leading cause of death in the United States. Interventional catheterization is an increasingly more common way to treat blood vessels which have become occluded (blocked) or stenotic (narrowed) by calcified artherosclerotic plaque deposits. Blood vessels that have become occluded or stenotic may interrupt blood flow, which supplies oxygen and cause heart attacks or strokes. Occluded or stenotic blood vessels may be treated with a number of medical procedures including angioplasty and atherectomy. Angioplasty techniques, such as percutaneous transluminal coronary angioplasty (PTCA), also known as balloon angioplasty are relatively noninvasive methods of treating restrictions in blood vessels. A balloon catheter is advanced over a guidewire until the balloon is positioned in the restriction of a diseased blood vessel. The balloon is then inflated compressing the atherosclerotic plaque. Frequently, the wall of the vessel is weakened after inflation and a metal stent in is expanded and deployed against the plaque. The stent helps keep the vessel open. During an atherectomy procedure, the stenotic lesion is mechanically cut or abraded away from the blood vessel wall using an atherectomy catheter.

Microelectrochemical systems pressure sensors can be used to measure the pressure in the balloons during inflation, to make sure damage due to over inflation is minimized. The MEMS temperature sensors can be integrated on catheters and guidewires to determine the location of inflamed plaque. The inflammation causes artery walls in the damaged area to have an increased temperature up to 3°C higher than healthy tissue. Approximately 20–50% of all patients undergoing these therapeutic procedures to clear blocked coronary arteries will suffer restenosis (reblockage) within 6 months of the initial procedure. Drug coated stents have significantly lowered these rates and have been approved for use in Europe for a few years. They are expected to be approved in the United States later this year. The MEMS laser micromachining technology is used in the fabrication of conventional stents and drug coated stents (31). Stainless steel micromachining technology has also been developed at the University of Virginia for piercing structure drug delivery/gene therapy stents for the treatment of restenosis (32). There is potentially a large opportunity for MEMS in embedding sensors into stents to

create a smart stents, which would be able to alert doctors when restenosis occurs or other problems occur (33). The MEMS rotary cutting devices have been fabricated for atherectomy procedures (34), but are not widely used because cut up plaque particles can flow downstream and cause strokes.

FETAL SURGERY

Fetal surgical techniques were first pioneered at the University of California San Francisco (UCSF) in the 1980s to operate on babies while they were still in the womb. The success rate of treating certain birth defects is higher the earlier they are addressed in fetal development. Initially open surgical techniques were used with an incision through the mother's abdomen to allow direct access for the surgeon's hands. After the surgery was complete, the womb was sutured and the mother delivered the baby weeks or months later. In 1981, the first fetal urinary tract obstruction procedure was performed at UCSF. Lately, minimally invasive surgical and robotic surgical techniques have been used that have reduced the risk of complications and premature labor. Other fetal procedures have also been developed to treat cojoined twins, hernias, spina bifida, tumors, and heart defects. One area of interest is in fetal heart.

The development of the cardiovascular system in a developing fetus is typically completed by the twelfth week of gestation. At this stage primary heart defects are small and if repaired will prevent defects from becoming more severe as the heart changes to adapt to the normalization blood flows and pressures in the later periods of gestation.

Fetal heart surgery can be used to treat hypoplastic heart syndrome (HHS), which was often considered fatal. This syndrome causes the heart's left side to fail to grow properly and is caused by an obstruction which restricts blood flow. The heart requires the mechanical stress of blood flow to grow properly, and thus fails to develop normally. Today, three open surgeries are required to allow the remaining single (right) ventricle to support both the pulmonary and systemic circulations. The long-term survival for these patients into adulthood is significantly $< 50\%$. Preserving both ventricles would result in the best chances of long-term survival.

An interventional catheter can be used to treat HHS in a minimally invasive manner. The balloon catheter must first penetrate in through the mother's abdominal wall, the placenta and then the fetus's chest into its heart. It must then locate the fetus's tiny heart and expand to open the blockage. Afterward, the surgeon needs to know whether the operation has succeeded enough for the baby to develop normally. Verimetra, Inc., Carnegie Mellon University, and Children's Hospital of Pittsburgh have embedded a MEMS flow sensor and a series of micromachined bar codes on the tip of a balloon catheter. The bar codes enable the catheter to be visualized with ultrasound allowing surgeons to know its exact position. The flow sensor is a thermistor. As blood flow increases the temperature measured by the thermistor decreases and in this manner blood flow changes as the catheter progresses through the heart's vessels can be monitored. This allows for the measurement of blood flow at or near a constriction and then again after the procedure to open the constriction has been performed.

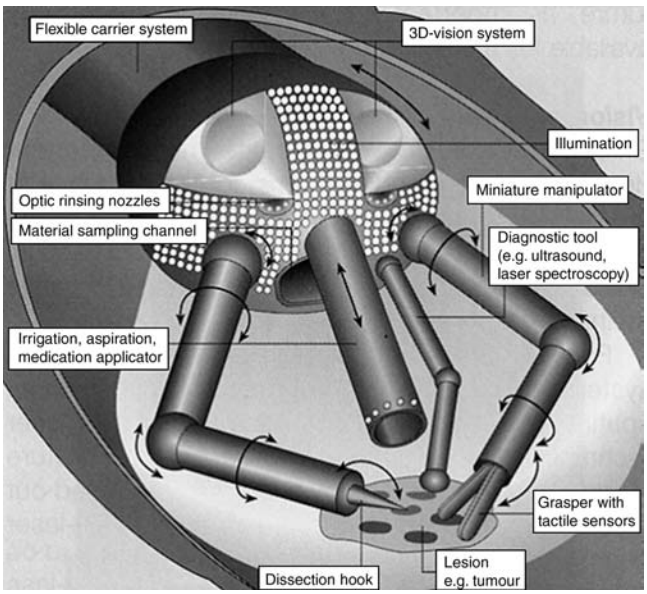


Figure 14. Highly integrated microsurgical probe. (Courtesy of Prof. Dr. M. Schurr, novineon Healthcare Technology Partners GmbH and Prof. G. Buess, Universitätsklinikums Tbingen.)

FUTURE SYSTEMS

In 1959, Richard Feynman gave his famous talk There's Plenty of Room at the Bottom (35). In it, he talked about being able to put a surgeon in a blood vessel which would be able to look around ones heart. Initially, new microsurgical tools will focus on measuring or detecting a specific parameter be it pressure, blood flow, velocity, temperature, and so on. The MEMS sensor systems will continue to be refined and improved leading to the integration of multiple sensor systems on surgical tool followed by tools with multiple functions. Multifunction surgical tools will reduce the number of tool insertions and removals reducing patient risks. Eventually, this will lead to highly integrated probes, which will do everything a surgeon needs, such as the concept shown in Fig. 14 (36). These tools will fit through a standard 5-mm port and have built in 3D cameras for visualization, biopsy samplers with microfluidic processing capability to do tissue analysis, ultrasound transducers, and tactile sensors for feedback to the surgeon.

BIBLIOGRAPHY

1. Diamiano R. Next up: Surgery by remote control. NY Times, Apr. 4, 2000; pp. D1.
2. Blazer E, Koh W, Yon E. A miniature digital pressure transducer. Proceedings of the 24th Annual Conference on Engineering Medicine and Biology. 1971. pp 211.
3. Peirs J, et al. A microturbine made by micro-electro-discharge machining. Proc 16th Eur Conf Solid-State Transducers 2002; 790–793.
4. Pique A, et al. Laser direct-write of embedded electronic components and circuits. Presented at Photon Processing

in Microelectronics and Photonics IV, Jan 24–27 2005, San Jose, (CA); 2005.
5. Götzen R. Growing up, additive processes in MEMS fabrication and packaging. Presented at Machines and Processes for Micro-scale and Meso-scale Fabrication, Metrology and Assembly, Gainesville, FL; 2003.
6. Rebello KJ, Lebouitz K, Migliuolo M. MEMS tactile sensors for surgical instruments. MRS Symp. Proc.: Biomicroelectromech. Systems (BioMEMS). 2003;773:55–60.
7. Menciassi A, et al. Force feedback-based microinstrument for measuring tissue properties and pulse in microsurgery. Presented at 2001 IEEE International Conference on Robotics and Automation (ICRA), May 21–26 2001, Seoul; 2001.
8. Rebello KJ, et al. MEMS based technology for endoscopic assessment of coronary arterial hardness. Presented at the 5th Annual NewEra Cardiac Care Conference, Dana Point (CA); 2002.
9. Bicchi A, et al. Sensorized minimally invasive surgery tool for detecting tissutal elastic properties. Presented at Proceedings of the 1996 13th IEEE International Conference on Robotics and Automation. Pt. 1 (of 4), Apr. 22–28 1996, Minneapolis, (MN); 1996.
10. Rosen J, Hannaford B, MacFarlane MP, Sinanan MN. Force controlled and teleoperated endoscopic grasper for minimally invasive surgery—experimental performance evaluation. IEEE Trans Biomed Eng 1999;46:1212–1221.
11. Scilingo EP, Bicchi A, De Rossi D, Iacconi P. Haptic display able to replicate the rheological behaviour of surgical tissues. Presented at Proceedings of the 1998 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Pt. 4 (of 6), Oct. 29-Nov. 1 1998, Hong Kong, China; 1998.
12. Fleischman A. et al. Miniature high frequency focused ultrasonic transducers for minimally invasive imaging procedures. Sensors Actuators, A: Phys 2003;103:76–82.
13. Zara JM, Bobbio SM, Goodwin-Johansson S, Smith SW. Intracardiac ultrasound scanner using a micromachine (MEMS) actuator. IEEE Trans Ultrasonics, Ferroelectrics, and Frequency Control 2000;47:984–993.
14. Daft C, et al. Microfabricated ultrasonic transducers monolithically integrated with high voltage electronics. Presented at 2004 IEEE Ultrasonics Symposium, Aug. 23–27 2004, Montreal, Queeue, Canada; 2004.
15. Chang JK, et al. Development of endovascular microtools. J Micromech Microeng 2002;12:824–831.
16. Degertekin FL, Guldiken RO, Karaman M. Micromachined capacitive transducer arrays for intravascular ultrasound. Presented at MOEMS Display and Imaging Systems III, Jan. 24–25 2005, San Jose, (CA); 2005.
17. Brock D. Smart medicine: The application of Auto-ID technology to healthcare. Auto-ID Center, MIT, Cambridge, (MA); 2002.
18. Tatar F, Mollinger JR, Bastemeijer J, Bossche A. Time of flight technique used for measuring position and orientation of laparoscopic surgery tools. Presented at Sensors, 2004. Proceedings of IEEE; 2004.
19. Fang C-M, Lee S-C. A research of robotic surgery technique by the use of MEMS accelerometer. Presented at Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society. EMBS/BMES Conference; 2002. Proceedings of the Second Joint, 2002.
20. Tanase D, et al. 3D position and orientation measurements with a magnetic sensor for use in vascular interventions. Presented at Biomedical Engineering, 2003. IEEE EMBS Asian-Pacific Conference on; 2003.
21. Totsu K, Haga Y, Esashi M. Three-axis magneto-impedance effect sensor system for detecting position and orientation of catheter tip. Sensors Actuators, A: Phys 2004;111:304–309.
22. Chen X, Lal A. Integrated pressure and flow sensor in silicon-based ultrasonic surgical actuator. Presented at Ultrasonics Symposium, 2001 IEEE; 2001.
23. Son I-S, Lal A. Feedback control of high-intensity silicon ultrasonic surgical actuator. Presented at Solid-State Sensors, Actuators and Microsystems, 2005. Digest of Technical Papers. TRANSDUCERS '05. The 13th International Conference on; 2005.
24. Polla DL, et al. Microdevices in medicine. Annu Rev Biomed Eng 2000;2:551–76.
25. Polla D, et al. Precision micromotor for surgery. Presented at Microtechnologies in Medicine and Biology, 1st Annual International, Conference On. 2000; 2000.
26. Kozlova TV, Shaposhnikova NF, Scobeleva VB, Sokolovskaya TV. Non-penetrating deep sclerectomy: Evolution of the method and prospects for development (review). Ophthalmosurgery 2000;3:39–54.
27. Angunawela R, Von Mohrenfels CW, Marshall J. A new age of cataract surgery. Cataract & Refractive Surgery Today I 2005; 36–38.
28. Kalvesten JE, Smith L, Tenerz L, Stemme G. First surface micromachined pressure sensor for cardiovascular pressure measurements. Presented at Proceedings of the 1998 IEEE 11th Annual International Workshop on Micro Electro Mechanical Systems, Jan. 25–29 1998, Heidelberg, Ger; 1998.
29. Tanase D, Goosen JFL, Trimp PJ, French PJ. Multi-parameter sensor system with intravascular navigation for catheter/guide wire application. Presented at Transducers'01 Eurosensors XV, Jun. 10–14 2001; Munich, 2002.
30. Haga Y, Esashi M. Biomedical microsystems for minimally invasive diagnosis and treatment. Proc IEEE Biomed App Mems Microfluidics 2004;92:98–114.
31. Kathuria YP. An overview on laser microfabrication of biocompatible metallic stent for medical therapy. Presented at Laser-Assisted Micro- and Nanotechnologies 2003, Jun. 29–Jul. 3 2003, St. Petersburg, Russian Federation; 2004.
32. Reed ML. Micromechanical systems for intravascular drug and gene delivery. Presented at BioMEMS 2002 Conference, Boston; 2002.
33. Goldschmidt-Clermont P, Kandzari D, Khouri S, Ferrari M. Nanotechnology needs for cardiovascular sciences. Biomed Microdevices 2001;3:83–88.
34. Ruzzu A, et al. A cutter with rotational-speed dependent diameter for interventional catheter systems. Presented at Micro Electro Mechanical Systems, 1998. MEMS 98. Proceedings., The Eleventh Annual International Workshop on, 1998.
35. Feynman RP. There's plenty of room at the bottom. J Microelectromech Systs 1992;1:60–66.
36. Schurr MO, Heyn S-P, Ment W, Buess G. Endosystems—Future perspectives for endoluminal therapy. Minimally Invasive Ther Allied Technol 1998;7:37–42.

Further Reading

Taylor RH, Lavallee S, Burdea GC, Mosges R. Computer Integrated Surgery: Technology and Clinical Application. Cambridge, (MA): MIT Press; 1996.

Zenati M. Robotic heart surgery. Cardiol Rev 2001;9(5):1–8.

Davies B. A review of robotics in surgery. Proc Inst Mech Eng 2000;214:129–140.

Madou M. Fundamentals of Microfabrication: The Science of Miniturization. 2nd ed. Boca Raton, (FL): CRC Press; 2002.

Kovacs GTA. Micromachined Transducers Sourcebook. Boston, MA: McGraw-Hill; 2002.

See also ENDOSCOPES; FIBER OPTICS IN MEDICINE; INTRAUTERINE SURGICAL TECHNIQUES; NANOPARTICLES; RADIOSURGERY, STEREOTACTIC.

MINIMALLY INVASIVE SURGICAL TECHNOLOGY

JAY R. GOLDBERG
Marquette University
Milwaukee, Wisconsin

INTRODUCTION

Most surgical procedures involve the invasion and disruption of body tissues and structures by surgical instrumentation and/or implantable medical devices, resulting in trauma to the patient. Diagnostic imaging procedures, such as magnetic resonance imaging (MRI), computed tomography (CT), X ray, positron emission tomography (PET), and ultrasound do not require disruption of body tissues, and are thus considered to be noninvasive. Extra corporeal shockwave lithotripsy (ESWL) used to disintegrate kidney stones is an example of a noninvasive therapeutic procedure. As shown in Fig. 1, it focuses acoustic energy, generated outside of the patient's body, on kidney stones. No trauma to the patient occurs during this procedure.

Open heart coronary artery bypass and organ transplantation surgery are examples of highly invasive surgery. These procedures require a high level of invasion and disruption of body tissues and structures. Bones, muscle tissue, and blood vessels are cut, and tissue from other parts of the body may be grafted, resulting in a high level of surgical trauma to the patient.

Minimally invasive surgical procedures are less traumatic than corresponding conventional surgical procedures. The use of small instruments placed through intact natural channels, such as the esophagus, urethra, and rectum, is less invasive than conventional open surgical approaches requiring large incisions, significant loss of blood, and trauma to tissues. The use of small instruments, such as biopsy guns and angioplasty balloons placed through small incisions, results in minor trauma to the patient, but is much less invasive than open surgical procedures used to accomplish the same goals. Procedures using small instruments through intact natural channels or small incisions are classified as minimally invasive.

A minimally invasive surgical procedure can be defined as surgery that produces less patient trauma and disruption of body tissues than its conventional open surgical counterpart. For example, conventional appendectomies require a 4 cm long incision made through layers of skin, muscle, and other tissues to gain access to the appendix. Once the appendix is removed, the layers of the wound are sutured together to allow them to heal. This invasive procedure requires a significant level of invasion and disruption of tissues and other structures. The minimally invasive appendectomy is performed with small surgical instruments placed into small incisions in the patient's abdomen. Once the appendix is removed, the small incision is closed with sutures. This procedure results in much less trauma to the patient than the open surgical approach.

Minimally invasive surgery (MIS) is performed with small devices inserted through intact natural orifices or channels, or small incisions used to create an orifice. Some procedures are performed with devices located outside the body, and thus are noninvasive. The MIS procedures are an alternative to open surgery. The benefits of MIS include reduced patient trauma, postoperative recovery time, and healthcare costs.

INSTRUMENTATION FOR MINIMALLY INVASIVE SURGERY

Many MIS procedures involve flexible or rigid fiber optic endoscopes for imaging surgical sites and delivering instrumentation for diagnostic or therapeutic applications. The term "endoscope" is a generic term used to describe tubular fiber optic devices placed into the body to allow visualization of anatomical structures.

Endoscopes consist of a fiber optic light guide, high intensity light source, coherent fiber optic bundle, steering mechanism, and various working channels for insertion of endoscopic instrumentation and infusion of irrigation fluids or insufflating gases. Glass fibers comprise the fiber optic light guide and coherent bundle and are surrounded by a flexible polyurethane sheath or rigid stainless steel tube. Figure 2 shows the components of a typical endoscope. The light source provides light that is transmitted through the light guide and projected onto the anatomical site to create an image. The coherent fiber bundle transmits the image back to the focusing eyepiece and into the surgeon's eye. The surgeon can move the endoscope proximally (toward the patient) and distally (away from the patient) by pushing and pulling the endoscope into and out of the body, respectively. Steering is accomplished with a handle attached to a cable that pulls and bends the tip of the endoscope in the desired direction. The proximal (closest to the patient) end of the endoscope (shown in Fig. 3) contains lumens for the fiber optic light guide and coherent fiber bundle, and one or more working channels for passage of instruments and irrigation fluid into and out of the operative site. Some endoscopes contain video cameras that display endoscopic images on a video monitor in the operating room, as shown in Fig. 4.

Specialized endoscopes have different names depending on what part of the body they are used to image. Table 1 lists the names of various endoscopes, the procedures for which they are used, and the anatomical location where they are used.

Endoscopic Procedures and Instrumentation

To perform an endoscopic procedure, the surgeon inserts a sterilized endoscope into a natural channel or orifice, such as the urethra, rectum, esophagus, bronchial tube, or nose. If there is no natural channel to provide access to the operative site (as in abdominal surgery), then the surgeon will create one with a small incision, and may insert a hollow trocar into the incision, as shown in Fig. 5. The trocar is left in the wound to provide access to the abdominal cavity. The endoscope is inserted into the access site (natural channel, incision, or trocar) and as it is advanced toward the operative site the surgeon monitors the image produced by the endoscope, either through the objective eyepiece or video monitor (as shown in Fig. 4). Rigid endoscopes are typically used when access is obtained

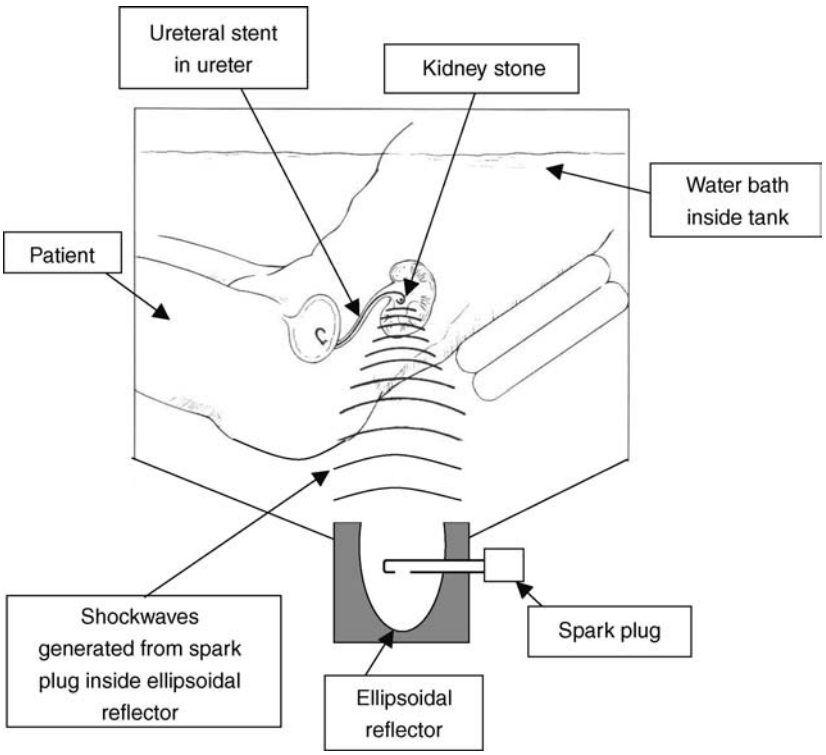


Figure 1. Schematic diagram of noninvasive extra corporeal shockwave lithotripsy (ESWL) procedure used to disintegrate kidney or biliary stones. Acoustic waves generated from spark plugs are focused with ellipsoidal reflectors on the kidney stone.

through an incision. Flexible endoscopes are used when access is obtained through natural channels (2). Once in position, the surgeon can then manipulate the endoscope to inspect tissues and anatomical structures (Table 1).

If a procedure other than visual inspection is required, the surgeon has a variety of instruments available to grasp, cut, remove, suture, and cauterize tissues, and remove debris through the endoscope. Forceps (Fig. 6) and graspers are used to grasp tissue and other objects such as kidney stones. Scalpels and scissors (Fig. 7) are used for cutting tissue. Suturing devices are used to repair internal wounds resulting from endoscopic procedures such as appendectomies, cholecystectomies (gallbladder removal), and arthroscopies. Morcellators are used to reduce the size and change the shape of a mass of tissue, such as a gallbladder, allowing it to be removed through a small incision. Electrohydraulic lithotriptor (EHL) and laser probes are used to disintegrate objects, such as kidney

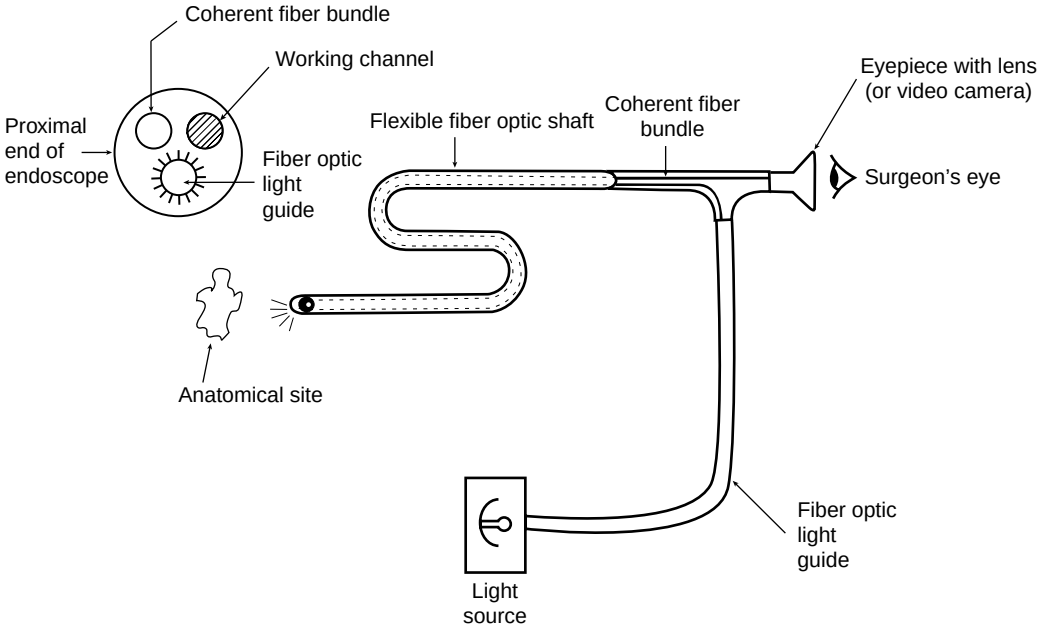


Figure 2. Components of a typical endoscope.

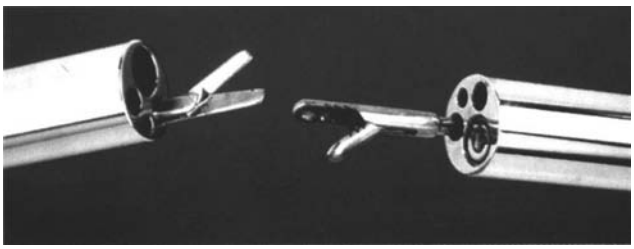


Figure 3. Proximal ends of typical endoscopes. Endoscopic instruments (left:scissors, right:forceps) are shown placed through working channels. Two other channels contain the fiber optic light guide and coherent fiber bundle, respectively. A fourth channel can be used for instrumentation or delivery of irrigation fluids to operative site (1). (Reprinted from Endoscopic Surgery, Ball, K., page 62, Copyright 1997, with permission from Elsevier.)

stones, with acoustic and laser energy, respectively, allowing the debris to be passed through the ureter. Stone baskets are used to trap kidney or ureteral stones for endoscopic removal, as shown in Fig. 8. These instruments are placed through the working channel of an endoscope or trocar and are operated by manipulating handles and controls located outside the patient.

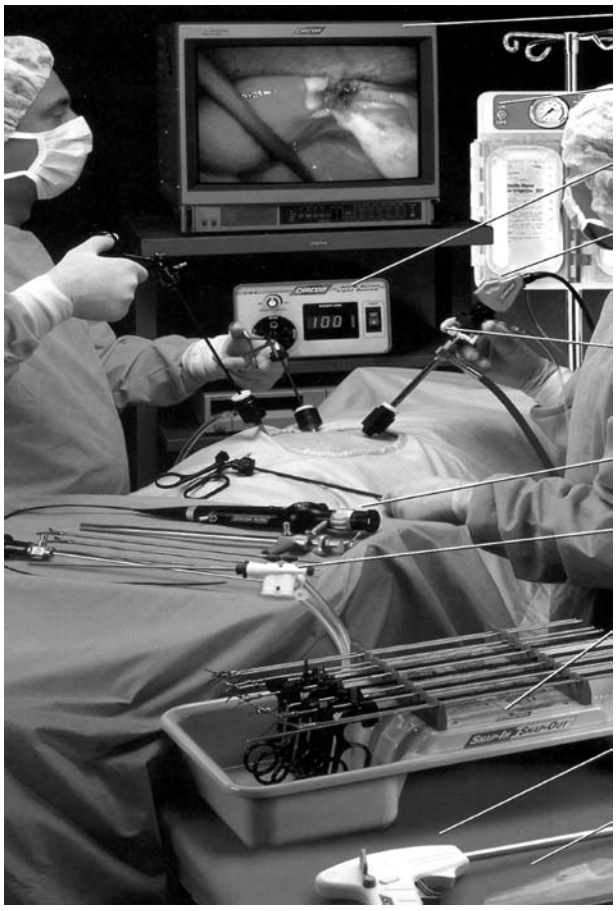


Figure 4. Surgeon viewing laparoscopic images of gallbladder on video monitor in operating room. Note use of three trocars in patient's abdomen; one each for video camera and two instruments. (Courtesy ACMI, Southborough, MA.)

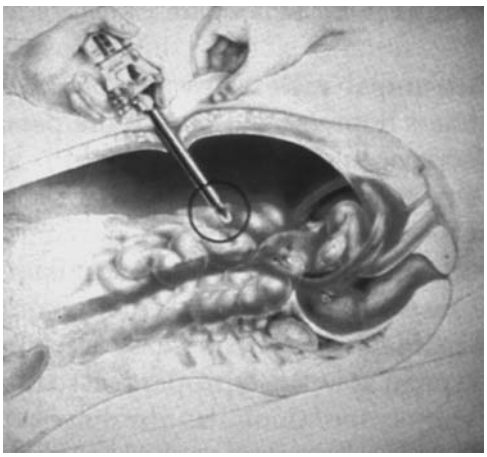


Figure 5. Trocar inserted into abdominal incision to provide access to abdominal cavity. Safety shield covers sharp end of trocar (circled) upon insertion to prevent damage to abdominal organs. (1) (Courtesy Ethicon Endo-Surgery Inc., Cincinnati, OH.)

Endoscopes may contain multiple working channels that allow for irrigation of fluids to irrigate and flush out clots and other surgical debris. The presence of the fluid improves visibility through the endoscope by keeping the visual field clear and the lens clean. Laparoscopic surgery requires insufflation of CO_2 or N_2O through the working channel or trocar and into the abdominal cavity to cause the cavity to expand, separating organs, and enlarging the operative site and visual field.

There are typically not enough working channels in a laparoscope to accommodate all of the instruments needed for a particular procedure. In these cases, multiple access sites are created with additional trocars. Typically, a camera is placed through one trocar and used to visualize the work being performed with instruments placed through other trocars.

When an endoscopic procedure is completed, the endoscope and instrumentation are removed from the patient via the natural channel or incision. When a laparoscopic procedure is completed, the laparoscope, camera, instruments, and trocars are removed from the patient. The wounds created by the trocars are sutured and the patient begins the recovery process. Although some of the CO_2 from insufflation may escape the abdominal cavity when all instrumentation is removed, some will be absorbed by body tissues and eliminated via respiration. Patients typically recover and are discharged from the hospital within 24–48 h.

Non-Endoscopic Procedures and Instrumentation

Not all minimally invasive procedures require endoscopic devices for imaging and placement of instruments. Some MIS procedures (listed in Table 2), such as stereotactic radiosurgery, use lasers or gamma radiation in place of scalpels. Others use small catheters to deliver medication, devices, or energy to specific anatomical locations. Balloons and stents, delivered via catheters, are used to access and dilate constricted vessels and maintain patency. Catheter mounted electrodes are used to deliver thermal, micro-

Table 1. Names of Endoscopes, MIS Procedures with Which they are Used, and Anatomical Location Where they are Used

Medical Specialty	Type of Endoscope Used	MIS Procedures	Anatomical Location
Urology	Cystoscope Ureteroscope Nephroscope	Cystoscopy, Transurethral resection of the prostate (TURP) Ureteroscopy, stone removal Nephroscopy, stone removal	Urethra, bladder ureter, kidney
Gastroenterology	Gastroscope Colonoscope Sigmoidoscope	Gastroscopy, gastric bypass Colonoscopy, Sigmoidoscopy	Stomach, colon Sigmoid colon
General surgery	Laparoscope	Laparoscopy, hernia repair, appendectomy, cholecystectomy (gallbladder removal)	Abdomen
Orthopedics	Arthroscope	Arthroscopy	Knee and other joints
Ob/Gyn	Hysterescope	Tubal ligation, hysterectomy	Female reproductive tract
Ear, nose, and throat	Laryngoscope, Bronchoscope Rhinoscope	Laryngoscopy, bronchoscopy Rhinoscopy, sinuscopy	Larynx, bronchus, nose, sinus cavities

wave, or radio frequency energy to selectively destroy tissue in ablation procedures used to treat cardiac arrhythmias (3), benign prostatic hypertrophy (BPH) (4), and other conditions. Many of these devices are guided through the body with the help of imaging or surgical navigation equipment.

Image Guided Surgery–Surgical Navigation

Image guided surgery (IGS) allows surgeons to perform minimally invasive procedures by guiding the advancement of instrumentation through the patient's body with increased accuracy and better clinical outcomes. Preoperatively, an IGS system is used to produce computerized anatomical maps of the surgical site from MRI or CT images of the patient. These maps are then used to plan the safest, least invasive path to the site. During an image guided procedure, the IGS system provides surgeons with a three dimensional image showing the location of instruments relative to the patient's anatomical structures. It tracks the movement of surgical instruments in the body, correlates these movements with the patient's preoperative images, and displays the location of the instruments on a monitor in the operating room. This feedback helps the surgeon safely and accurately guide instruments to the surgical site, reducing the risk of damaging healthy tissue (4,5).

An IGS system includes a computer workstation, localization system, display monitor, and specialized surgical instruments capable of being tracked by the system. Image processing and surgical planning software are also used. Tracking of instruments is accomplished through optical, electromagnetic, or mechanical means. Optical tracking systems use a camera mounted to view the surgical field and optical sensors attached to surgical instruments. These systems required line of sight between the camera and sensor to function properly. Electromagnetic tracking systems include a transmitter located close to the surgical site, and receivers attached to surgical instruments. Line of sight is not an issue with these systems, however, nearby metallic objects may produce interference to signals used to track instruments (4).

To ensure that the patient's anatomical features (and location of instruments) are accurately displayed by the IGS system, actual anatomical locations must be registered to the preoperative images. This can be accomplished by touching a probe to a marker on the patient's body and then assigning the location of this marker to its corresponding point in preoperative images (4).

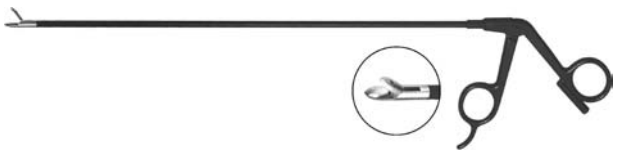


Figure 6. Endoscopic forceps. (Courtesy ACMI, Southborough, MA.)

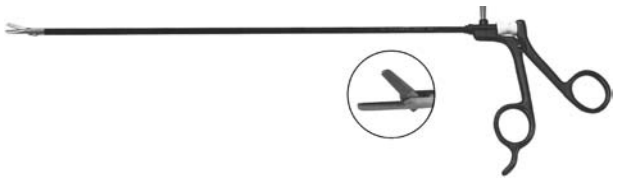


Figure 7. Endoscopic scissors. (Courtesy ACMI, Southborough, MA.)

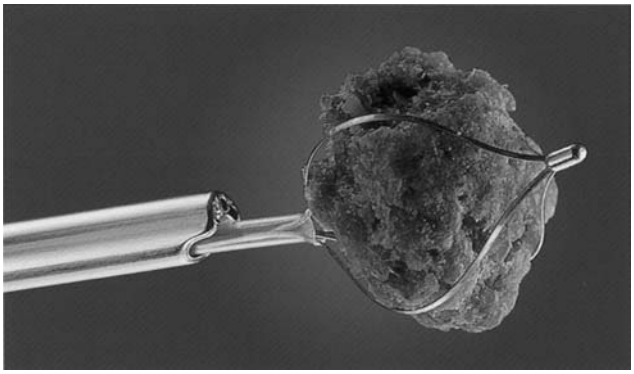


Figure 8. A stone basket used to trap and remove a large ureteral stone. (Courtesy ACMI, Southborough, MA.)

Table 2. Examples of Non-Endoscopic Minimally Invasive Surgical Procedures

Medical Specialty	Non-Endoscopic Minimally Invasive Surgical Procedure
Cardiovascular surgery	Minimally invasive direct coronary artery bypass (MIDCAB)
	Percutaneous transluminal coronary angioplasty (PTCA)
	Coronary stenting
	Radio frequency cardiac ablation
	Laser angioplasty
	Microwave catheter ablation for arrhythmias
	Chemical ablation for ventricular tachycardia
Ophthalmology	Laser photorefractive keratotomy (PRK)
	Laser ablation of cataracts
	Laser glaucoma surgery
Orthopedics	Total joint arthroplasty (hips, knees, and others)
Neurosurgery	Stereotactic radiosurgery
	Stereotactic radiotherapy
	Laser ablation of brain tissue, tumor tissue
Radiology	Clot removal
	Aneurism repair
	Cerebral arterial venous malformation repair
	Transjugular hepatic portal systemic shunt creation

Most surgical instruments must be adapted for use in image guided surgery by mounting sensors and other devices to allow detection of the instrument's position by the IGS system. Some medical device manufacturers are developing navigation-ready surgical instruments that contain small reflective spheres to act as reference arrays for cameras used in image guided surgery (5).

MINIMALLY INVASIVE SURGICAL PROCEDURES

This section contains a few examples of minimally invasive surgical procedures used in urology, orthopedics, neurosurgery, and general and cardiovascular surgery.

Ureteroscopy

Flexible ureteroscopy is used to perform a variety of diagnostic and therapeutic urological procedures. It involves entry into the body via intact natural channels (urethra and ureter) and does not require an incision. Local anesthesia and sedation of the patient are required.

Ureteroscopy is commonly used to remove ureteral or kidney stones. Initially, a cystoscope is inserted into the urethra. Next, a guidewire is placed through a cystoscope into the ureter, and advanced up into the kidney. While maintaining the position of the guidewire, the cystoscope is removed, and the distal end of the guidewire is placed into a working channel at the proximal end of the ureteroscope. The ureteroscope is then advanced along the guidewire into the ureter or kidney, as shown in Fig. 9. Active (controlled by the cable and handle) and passive deflection of the shaft, along with rotation of the flexible ureteroscope allows visual inspection of the renal calices as shown in Fig. 10. Ureteroscopic instruments are then placed through the working channel into the ureter or kidney. Figure 11 shows two devices used for ureteroscopic removal of ureteral stones. The stone grasper is used to physically grab the stone (Fig. 11). If the stone is small enough to fit into the working channel, it is pulled out of the patient through the ureteroscope. Large stones that will not fit into the working channel are pulled out with the ureteroscope. The laser lithotripter (Fig. 11) disintegrates the stone into small particles that can easily be passed naturally through the ureter, bladder, and urethra. Collapsed stone baskets can be placed alongside a stone and moved proximally and distally as they are expanded, until the stone is trapped in the basket (as shown in Fig. 8) and pulled out of the urethra.

Laparoscopy

Laparoscopy is commonly used for removal of the gallbladder and appendix, hernia repair, and other abdominal procedures. The basic steps involved in laparoscopy have been previously described. The lack of natural channels located in the abdomen requires the use of trocars to gain access to the operative site. Laparoscopic procedures require insufflation of gas to separate organs and expand the visual field. This is controlled by a separate insufflator that controls the pressure inside the abdomen produced by the flow of insufflating gases (1).

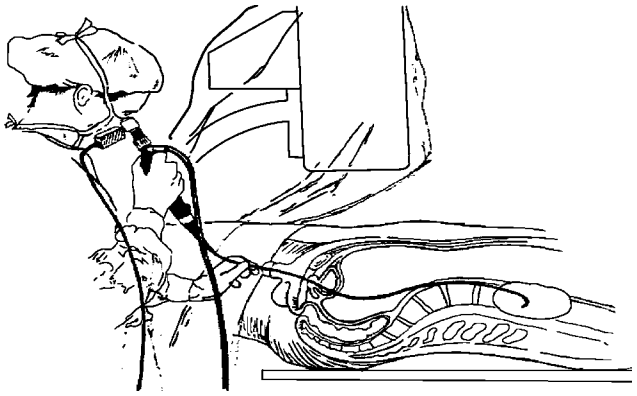


Figure 9. A surgeon views images through a ureteroscope placed through the urethra, bladder, ureter, and into the kidney. (Courtesy ACMI, Southborough, MA.)

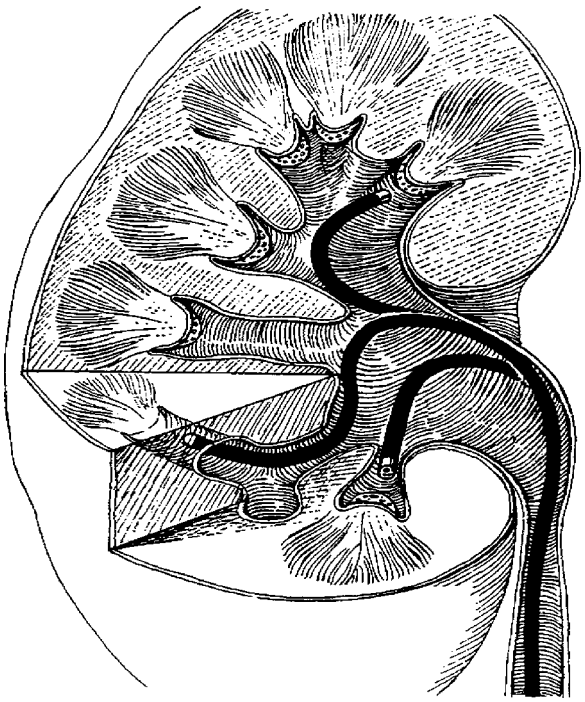


Figure 10. Flexibility and steerability of proximal end of ureteroscope allow inspection of upper, middle, and lower renal calices. (Courtesy ACMI, Southborough, MA.)

Total Joint Replacement

Total hip and knee replacements are typically performed with large incisions to expose the joint, allowing for complete visualization of and access to the joint and soft tissues. New surgical approaches using smaller incisions that result in less damage to muscles and other soft tissue limit the view of the joint. To compensate for the limited view, fluoroscopy and IGS systems are often used. Some existing surgical instruments have been modified to enable surgery through smaller incisions (6).

Traditional hip arthroplasty requires a 30–46 cm long incision, which is highly disruptive to muscles and surrounding tissues. Two different techniques can be used for minimally invasive total hip arthroplasty. One technique uses two 5 cm long incisions, one each for preparation and insertion of the acetabular and femoral components, respectively. The other technique involves one 8–10 cm long incision. Modified retractors and elevators are used to gain access and expose the joint. Fluoroscopy and IGS systems are used to provide the surgeon with a view of instruments and components (as shown in Fig. 12) and assist in positioning of instruments designed to enable accurate component alignment and placement. Minimally invasive hip arthroplasty results in less disruption of muscles and tissues, smaller and less noticeable scars, less blood loss and pain, and fewer blood clots and dislocations (6).

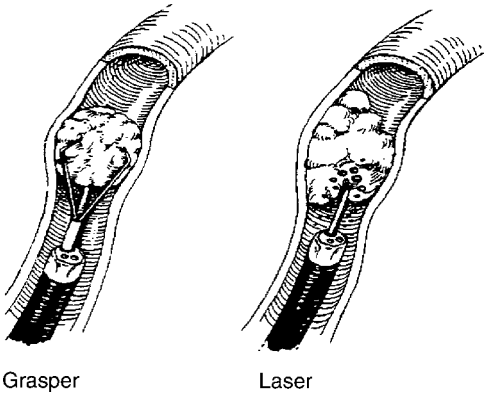


Figure 11. Devices used for endoscopic removal of ureteral stones. (Courtesy ACMI, Southborough, MA.)

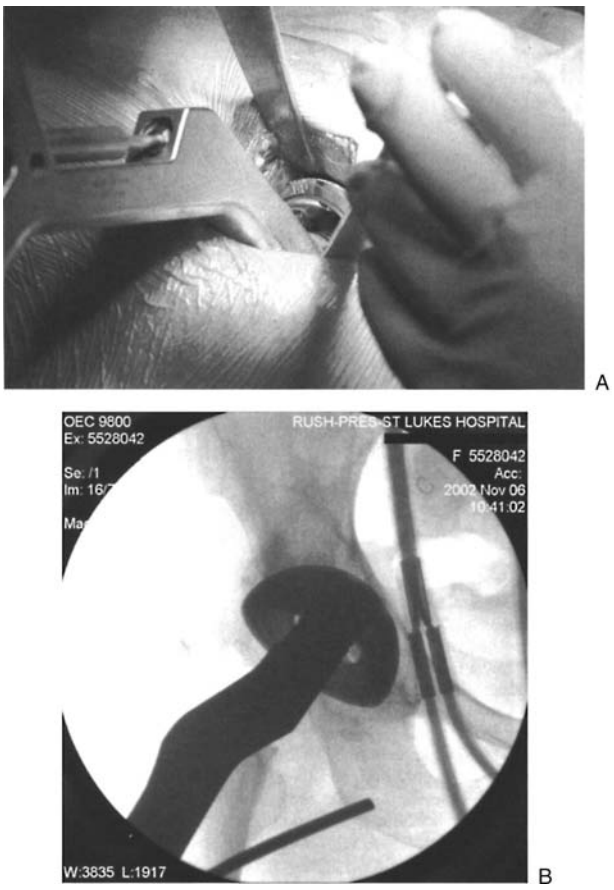


Figure 12. Acetabular component inserted with inserter through small incision (top image). Fluoroscopic image of inserter and acetabular component seated in acetabulum (bottom image). Images such as these allow surgeons to ensure proper placement of instruments and implantable components within hip joint when small incisions prevent viewing of entire device (6). (From MIS for the Hip and Knee: A Clinical Perspective, 2004, pg. 20, Minimally Invasive Total Hip Arthroplasty: The Two Incision Approach, Berger, R.A. and Hartzband, M.A., Fig. 2.11. Copyright 2004, with kind permission of Springer Science and Business Media.)

Most minimally invasive knee arthroplasties performed through smaller incisions involve a single compartment of the knee. These procedures typically use existing unicompartmental knee implants for resurfacing the medial or lateral femoral condyle and corresponding tibial plateau. Existing instrumentation has been modified to obtain

access to the joint and enable accurate placement and alignment of the unicompartmental knee components through smaller incisions.

Minimally Invasive Direct Coronary Artery Bypass

Coronary arteries may become blocked due to fatty deposits (plaque), blood clots, or other causes. This reduces blood flow to cardiac muscle, depriving it of needed oxygen and other nutrients. This condition can result in a myocardial infarction (heart attack) that can damage cardiac muscle.

Traditional open heart coronary artery bypass graft surgery has been used to restore normal blood flow to cardiac muscle. This procedure involves grafting a new vessel to points on both sides of the blocked coronary artery, thereby bypassing the blockage and restoring normal blood flow. It requires open access to a still heart to allow suturing of the graft to the blocked coronary artery. A sternotomy and separation of the sternum is required to provide access to the heart. The heart is stopped and the patient is attached to a heart–lung machine to maintain circulation of oxygenated blood through the body during the surgical procedure. This procedure is highly invasive and can result in a variety of complications, many of which are associated with use of a heart–lung machine. Inflammatory responses negatively affecting multiple organ systems have been observed in patients who were perfused with a heart–lung machine during traditional open heart coronary bypass surgery (7). These responses are due to reactions between circulating blood and material surfaces present in the heart–lung machine.

Minimally invasive direct coronary artery bypass is an alternative to open heart surgery. A small 5–10 cm incision made between the ribs replaces the sternotomy to gain access to the heart. A retractor is used to separate the ribs above the heart to maximize access to the operative site (8). Heart positioners and stabilizers (Fig. 13) are used to minimize the motion of the heart, allowing surgeons to perform the procedure on a beating heart, eliminating the need for the heart–lung machine. Some stabilizers grasp the heart with suction cups. Others use a fork like device to apply pressure to the beating heart to keep it steady for anastomosis of the graft (8). A thorascope and small endoscopic instruments are used to visualize the surgical site and perform the surgical procedure. The left internal mammary artery is commonly used as the grafted

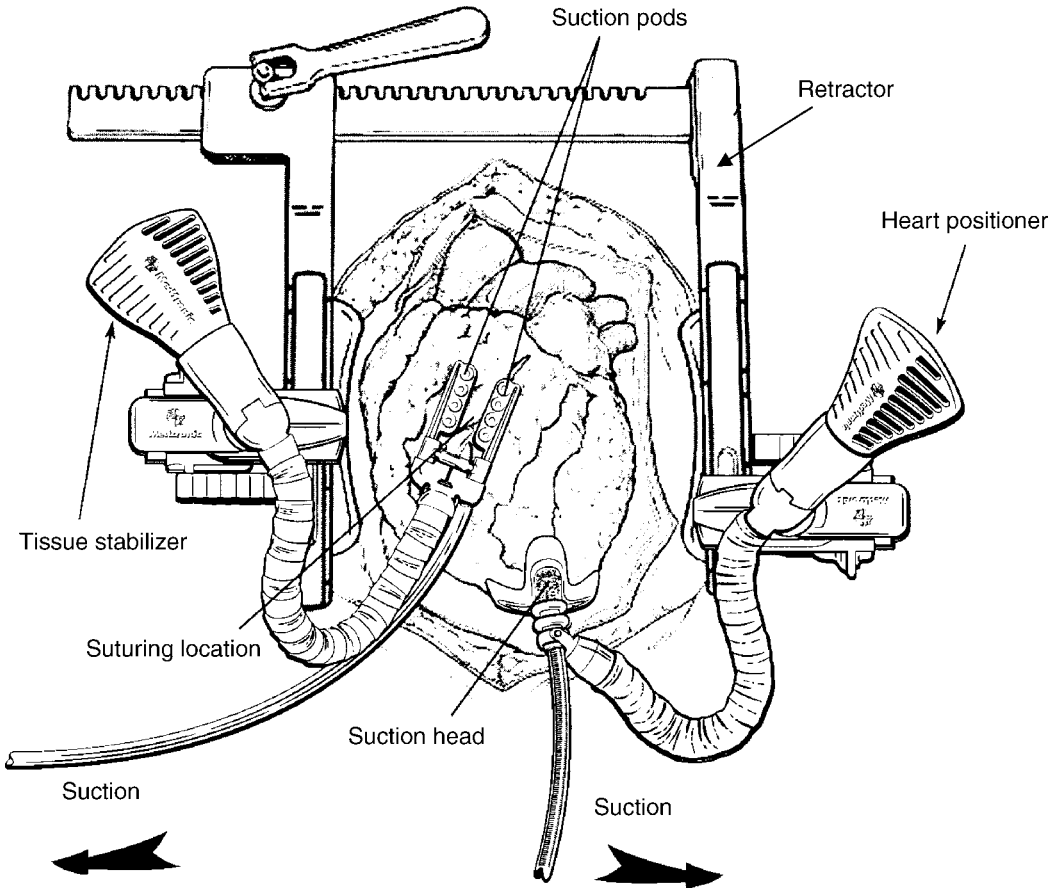


Figure 13. Heart positioner and stabilizer used in MIDCAB procedures. The positioner attaches to the sternal retractor and holds the heart in position using suction. It provides greater access to the blocked coronary artery. The tissue stabilizer, attached to the sternal retractor, straddles the blocked artery and holds the suturing site steady. This allows surgery on a beating heart, eliminating the need for a heart–lung machine along with its associated complications. (Courtesy of Medtronic, Inc., Minneapolis, MN.)

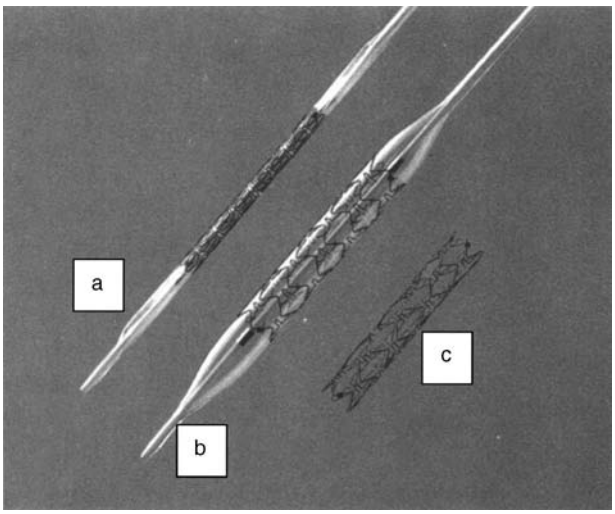


Figure 14. A PTCA catheter and stent. Stent and balloon in collapsed configuration for insertion and placement into coronary artery (a). Inflated balloon causing stent to expand (b). Expanded stent after collapse of balloon and removal of catheter (c). (Courtesy of Sorin Biomedica Cardio SpA, Italy.)

vessel to bypass the blockage of the coronary artery. The MIDCAB results in fewer complications, less blood loss, and shorter hospital stays.

Percutaneous Transluminal Coronary Angioplasty with Stent Placement

The PTCA method is used to open a blocked, constricted coronary artery instead of bypassing the blockage with a graft. Under local anesthesia, a steerable balloon catheter containing a stent mounted to the balloon (Fig. 14a) is inserted into the patient's femoral artery and guided to the constricted coronary artery under fluoroscopy. Once in position, the balloon is inflated to compress and flatten the plaque against the arterial wall, creating a larger opening for blood to flow through the coronary artery. The balloon is constructed with materials of different stiffness such that pressure from the inflating balloon is radially applied to the constricted area, and not to tissue outside of the constricted region (3). During balloon inflation, the stent is expanded to a larger diameter Fig. 14b, which is maintained after deflation of the balloon. The catheter is removed, and the expanded stent (Fig. 14c) is left in place to maintain a larger opening and prevent restenosis of the coronary artery.

Most coronary stents are made of stainless steel, nitinol, cobalt chrome molybdenum alloy, or gold (3). Some stents contain coatings to improve biological performance (biocompatibility and resistance to clot formation) and/or elute drugs into surrounding tissues to prevent restenosis of the coronary artery.

The PTCA method does not remove plaque from the coronary artery; it flattens it so it no longer restricts blood flow. Plaque removal methods involve lasers and rotational cutting devices (3).

Stereotactic Radiosurgery

In this noninvasive procedure, focused external beams of radiation are delivered to specific locations in the brain to treat tumors (9). The accuracy of the delivery prevents damage to healthy brain tissue. The patient's head is constrained in a mask or frame during the 30–45 min procedure.

Newer radiotherapy systems include robotic linear accelerators for delivery of radiation at any angle to the patient's head, a beam shaper to match the shape of the beam to the three-dimensional (3D) shape of the tumor, and imaging equipment to provide real-time tracking of tumor location and patient positioning (9).

Treatment of Benign Prostatic Hypertrophy

Benign prostatic hypertrophy causes the prostate to enlarge. An enlarged prostate applies pressure to the prostatic urethra that can occlude the urethra, reduce urinary flow rate, and make urination difficult. The enlarged prostate can be removed with an open surgical approach or less invasive transurethral resection of the prostate (TURP). The TURP procedure involves cutting and removing segments of the prostate through a cystoscope or with a resectoscope inserted into the urethra, and can result in complications, such as incontinence and impotence. Prostatic balloon dilators and prostatic stents inserted into the urethra have been used to expand the narrowed urethra. Transurethral laser incision of the prostate (TULIP) has been used to cut and remove prostate tissue. In this procedure, a catheter mounted laser facing radially outward toward the prostate delivers laser energy that cuts through the enlarged prostatic tissue, relieving pressure on the urethra.

Newer approaches to treating BPH include transurethral microwave therapy (TUMT) and transurethral needle ablation (TUNA) (4). These procedures use microwave and radio frequency energy, respectively, to heat and destroy unwanted tissue without actually removing the tissue. The TUMT method involves insertion of a urethral catheter containing a microwave antenna into the bladder neck. Channels contained in the catheter carry cooling water to prevent thermal damage to the urethra while microwave energy is used to heat the prostate for ~ 1 h. The TUNA method uses a cystoscope to insert two shielded electrode needles through the urethra and into the prostate. Radio frequency energy is delivered to the prostate, heating the tissue and destroying it. Thermal energy causes the prostate tissue to stiffen and shrink, relieving the pressure on the urethra caused by the enlarged prostate. Both TUMT and TUNA deliver controlled thermal energy to a targeted area to selectively destroy prostate tissue (4).

NEW DEVELOPMENTS IN TECHNOLOGY FOR MINIMALLY INVASIVE SURGERY

New devices and technologies are being developed and clinically evaluated for use in MIS. Robots can be used to enable surgeons to perform more intricate surgical

procedures than can be done with endoscopic devices (10). In robot-assisted surgery, the surgeon operates a console to control mechanical arms positioned over the patient. The arms become an extension of the surgeon's hands. While observing the procedure on viewing screens (one for each eye) that produce a 3D image, the surgeon uses hand, finger, and foot controls to move the mechanical arms containing interchangeable surgical instruments through a small opening in the patient's body (10).

Other new technologies for MIS include 3D video imaging, integrated robotic and surgical navigation systems, devices for mechanical retraction of the abdominal wall (eliminating the need for insufflation), and telerobotics (8).

OUTCOMES OF MINIMALLY INVASIVE SURGERY

Comparison of MIS to Conventional Approaches

Minimally invasive surgical procedures result in reduced patient trauma, less postoperative pain and discomfort, and decreased complication rates. Hospital stays and postoperative recovery times are reduced, resulting in lower hospital costs and allowing patients to return to work earlier.

Studies comparing various MIS procedures to their traditional open surgical counterparts have been conducted. One such retrospective study compared the clinical and economic aspects of laparoscopic and conventional cholecystectomies (11). Results of 153 consecutive traditional procedures and 222 consecutive laparoscopic procedures performed in a German teaching hospital were evaluated. Researchers found that although laparoscopic cholecystectomies required longer operative times (92 vs. 62 min), they resulted in fewer complications (6 vs. 9), shorter hospital stays (3 vs. 8 days), and an 18% decrease in hospital costs, when compared to traditional procedures.

In a Canadian study, the direct costs of conventional cholecystectomy, laparoscopic cholecystectomy, and biliary lithotripsy were compared (12). Researchers concluded that laparoscopic cholecystectomy provided a small economic advantage over the other two procedures and attributed this to a shorter hospital stay.

In the United States, hospital stays following laparoscopic cholecystectomies typically ranged from 3 to 5 days. Now, many of these procedures are performed on an outpatient basis. This additional reduction in hospital stay further reduces hospital costs, resulting in a greater economic advantage over the conventional procedure.

A study comparing results of 125 consecutive off-pump coronary bypass (OPCAB) procedures to a matched, contemporaneous control group of 625 traditional coronary artery bypass graft (CABG) procedures was conducted (7). The OPCAB procedure is a beating heart procedure that does not involve a heart–lung machine. Partial sternotomies were used with some patients in the OPCAB group. Researchers found that the OPCAB procedure resulted in a lower mortality rate (0 vs. 1.4%), reduced hospital stays (3.3 vs. 5.5 days), 24% lower hospital costs, and a reduced transfusion rate (29.6 vs. 56.5%), when compared to the traditional CABG procedure. Excellent graft patency rates and clinical outcomes were also reported with the OPCAB procedure.

In another study of 67 MIDCAB patients, it was found that average hospital charges for MIDCAB were $14,676 compared to $22,817 for CABG and $15,000 for coronary stenting (13). The significantly lower charges for MIDCAB were attributed to shorter hospital stays, elimination of perfusion expenses, and reduction in ICU, ventilation, and rehabilitation times.

Limitations of Minimally Invasive Surgery

There are several problems and limitations associated with MIS procedures. First, some minimally invasive surgical procedures can take longer to perform than their more invasive counterparts (10). Second, open surgical procedures allow the surgeon to palpate tissue and digitally inspect for tumors, cysts, and other abnormalities. Tactile feedback also assists in locating anatomical structures. The inability of the surgeon to actually touch and feel tissue and structures at the operative site limits the diagnostic ability of some MIS procedures (10). Third, a two-dimensional (2D) image is produced by the endoscope and video monitor. The resulting loss of depth perception combined with restricted mobility of instruments through a small incision make manipulation of endoscopic instruments challenging. Fine hand movements are difficult due to the long distances between the surgeon's hand and working ends of these instruments. Fourth, there is a limit to the number of instruments that can be used at one time through trocars and working channels of a laparoscope. Finally, instrumentation for MIS procedures is more expensive.

Insufflation presents a small risk not associated with open surgical procedures (1,10). If the gas pressure inside the abdominal cavity becomes too high or there is a small defect in a blood vessel, then a gas bubble can enter the bloodstream and travel to the heart or brain, causing unconsciousness or death.

Laparoscopic removal of cancer tissue carries a very small risk of transporting cancer cells to other parts of the body. As tumor tissue is removed through the laparoscope, some cancer cells may remain in the body. They may be displaced from their original location resulting in the spread of cancer cells to other parts of the body (10).

SUMMARY

Minimally invasive surgery is made possible through the use of specialized devices and technologies. These include endoscopes, surgical instruments, imaging and navigation systems, catheters, energy delivery systems, and other devices. The MIS procedures tend to require more time, are more difficult to perform than conventional procedures, and present some unique risks not associated with conventional procedures. Compared to conventional open surgical procedures, MIS procedures demonstrate similar clinical efficacy and reduce trauma and disruption of body tissues and structures. Patients experience less pain and discomfort, fewer complications, and shorter recovery periods. Hospital stays are reduced, resulting in lower hospital costs.

BIBLIOGRAPHY

1. Ball K, Endoscopic Surgery. St. Louis: Mosby Year Book; 1997.
2. Holland P, The Fundamentals of Flexible Endoscopes, Biomedical Instrumentation and Technology. Association for the Advancement of Medical Instrumentation, p 343–348, Sep./Oct. 2001.
3. Webster J G, ed. Minimally Invasive Medical Technology. Institute of Phys Publishing Ltd., 2001.
4. Spera G, The Next Wave in Minimally Invasive Surgery, Medical Device and Diagnostic Industry, Canon Communications, August 1998.
5. Gill B, Navigation Surgery Changing Medical Device Development, Medical Device and Diagnostic Industry, Canon Communications, December 2004.
6. Scuderi R G, Tria A J, eds., MIS of the Hip and the Knee: A Clinical Perspective; New York: Springer-Verlag, 2004.
7. Puskas J, et al. Clinical outcomes and angiographic patency in 125 consecutive off-pump coronary bypass patients. Heart Surg Forum May 1999;2(3):216–221.
8. Spera G, The kindest cut of all, Medical Device and Diagnostic Industry, Canon Communications, July 1998.
9. Technology Trends: Radiation System Could Be an Alternative to Surgery, Biomedical Instrumentation and Technology, Association for the Advancement of Medical Instrumentation, p. 18, January/February 2005.
10. Comarow A, Tiny holes, big surgery. U.S. News & World Report, July 22, 2002.
11. Bosch F, Wehrman U, Saeger H, Kirch W. Laparoscopic or open cholecystectomy: Clinical and economic considerations. Eur J Surg 2002;168(5):270–277.
12. Conseil d'evaluation des technologies de la sante du Quebec (CETS). The Costs of Conventional Cholecystectomy, Laparoscopic Cholecystectomy, and Biliary Lithotripsy. Montreal: CETS, 1993.
13. Oz M, Goldstein D, Minimally Invasive Cardiac Surgery. Humana Press; 1999.

See also ENDOSCOPES; GAMMA KNIFE; MICROSURGERY; STEREOTACTIC SURGERY; TISSUE ABLATION.

MOBILITY AIDS

RORY COOPER
ERIK WOLF
DIANE COLLINS
ELIANA CHAVEZ
JON PEARLMAN
AMOL KARMARKAR
ROSEMARIE COOPER
University of Pittsburgh
Pittsburgh, Pennsylvania

INTRODUCTION

The National Center for Health Statistics estimates that 12.4 million adults are unable to walk a quarter of a mile in the United States, and that 28.3 million adults have moderate mobility difficulty (1). Approximately 4 million Americans use wheelchairs, and about one-half of them use their wheelchairs as their primary means of mobility. About 1.25 million people wear a prosthetic limb due to injuries, birth anomalies, and disease. Once fatal injuries are now survivable due to advancing medical achievements, prolonging life spans, and the staggering growth in the aging population. Because of this growth, the population of individuals who use mobility aids is sure to grow in the coming decades.

The goal of issuing wheelchairs, prosthetics, walkers or rollators to individuals with mobility impairments independence remains the top priority when prescribing one of these devices, other main concerns include safety, not causing secondary injury (i.e., pressure sores, carpal tunnel syndrome, rotator cuff tear), and the physical design of the device (e.g., weight, size, ease of use). Research has shown that manual wheelchair users commonly report shoulder, elbow, wrist, and hand pain (2). Low back pain and fatigue are also common secondary ailments experience due to exposure of whole-body vibrations (3). Safety research shows that proper fitting and wheelchair skill can reduce injurious tips and fall in wheelchairs (4).

In order to fully maintain an active lifestyle, including recreational activities, participating in the community, and going to work, transportation for people with mobility impairments is essential. With the added technology that is necessary to allow people to use public or private transportation, added safety features must also be included to maintain the security of both the drivers and the passengers.

Besides performing normal activities of daily living, sports, and recreational activity are an important physical and psychosocial aspect of any human being. The case is no different with people who use assitive technology. With dramatic advances in technology, people with mobility impairments are able to go places and participate in activities that were once nearly impossible.

In recent years, wheelchairs, prosthetics, walkers, and rollators have been designed to be stronger, safer, lighter, more adjustable, and smarter than in the past, through the use of materials like titanium and carbon fiber, using advanced electronics, and so on. The technology of wheelchairs, prosthetics, walkers, and rollators has improved dramatically in the past 25 years due largely in part to the increased demand of consumers, their loved ones and others who assist consumers, and the people who recommend and issue these technology devices. The term"recommend" is used because it is crucial that these devices, especially wheelchairs and prosthetics, be issued by a team of professionals to provide the highest levels of independence and safety, and that this team is centered around the client. All of these components are important to the further development of the technology and, in turn, they may result in the increased independence of people with mobility impairments.

CLINICAL ASSESSMENT OF MOBILITY

The ultimate goal and outcome for a clinical assessment of mobility should drive toward a successful wheelchair, prosthetics, walker, or rollator recommendation that enhances the quality of life expectations and their effectiveness as reported by the consumer. Quality of life is specific to and defined by each person and/or family receiving clinical services. The consumer, their family, and care

givers, must be actively included in this process, as they will be most affected by the choice of the mobility aid. Also, people chose their mobility devices based on the features available that facilitate activities or address needs (5), and the clinician should be aware of the consumer preferences and the features of various devices.

The complexity of some of the mobility device components combined at times with involved disease processes can make it virtually impossible for a single clinician to act independently when recommending assistive technology. Therefore, involving *an interdisciplinary team* is recommended in the decision making process (6,7). This team, with the consumer as an involved member and a variety of rehabilitation professionals, includes a physiatrist or similarly trained physician who understands the importance of Assistive Technology, addresses medical issues and assists with mobility decisions; the Occupational or Physical Therapist with RESNA (www.resna.org) certified Assistive Technology Practitioner (ATP) credentials, who is the point person for evaluation and prescription; and the Rehabilitation Engineering (with RE Training and RET Credential) who is a technology expert with the ability to design–modify–tune devices, and who also understands the capabilities and applications of various technologies. Another important team partner is the Assistive Technology Supplier (ATS) or Certified Rehabilitation Supplier (CRTS, National Association of Rehabilitation Technology Suppliers, www.narts.org), who provides or manages demonstration equipment, does routine home and workplace assessments, and orders, assembles, and delivers the equipment. All team members involved in the mobility aid selection process should have knowledge about the technology available on the market. Peer reviewed journal articles (Assistive Technology, Archives of Physical Medicine and Rehabilitation, Journal of Rehabilitation Research and Development etc.), magazine articles and commercial database sources such as ABLEDATA (http://www.abledata.com/), or WheelchairNet (www.wheelchairnet.org) are good places to research devices or to direct consumers who want to inform and educate themselves.

Resources available to the team and its members includes a defined and dedicated space for demonstration equipment, assessments, and evaluations, access to common activities and tasks (e.g., ramps, curb cuts, bathroom, countertop), an electronic tracking system to follow clients and their technology, assessment resources (e.g., pressure mapping, SMART^{Wheel}, gait force plate, actigraph), IT Resources (e.g., email, web, databases, medline, paging, cell, wireless), and the facilities–hospital commitment to continuing education.

A mechanism for Quality Measures for AT Clinics will provide valuable feedback on performance quality and areas in need of improvement. An important tool to measure patient satisfaction is the information gained through a satisfaction survey provided to every patient, in order to find out whether the goals and desired outcomes have been met. Feedback on performance quality is provided through tracking mechanism of primary clinician credentials (ATS, ATP, RET), dedicated staffing levels, continuing education (CEUs and CMEs), compliance with the commission on Accreditation of Rehabilitation Facilities (CARF) AT Clinic Accreditation, as well as tracking of continuous quality improvement.

Assessment

The occupational or physical therapist conducts the initial assessment process and obtains critical information about the consumer and their environment. This part usually involves a structured interview with the consumer and then a physical motor assessment. Establishing a medical diagnosis that requires the mobility aid is vital to assure no ongoing medical problems exist that are not being adequately addressed. To properly specify a mobility device, the intentions and abilities of the consumer must be ascertained (8,9). The intentions and abilities may include how people perform tasks, where the deficits are, and how mobility systems can compensate for the deficits to augment task performance. Some outcome tools that are clinically used to measure the functional impact of mobility aids are the QUEST, the FEW and the Wheelchair Skills Test (WST).

Additional necessary information includes type of insurance, method of transportation, and physical capabilities. Also, if the consumer has been using a chair, historical information about their current chair should be addressed, including problems they are having. The mobility device chosen should also be compatible with the public and/or private transportation options available to the consumer, such as a bus, car, or van. The regularity of the surface, its firmness and stability are important, when, for example, recommending a wheelchair in determining the tire size, drive wheel location, and wheel diameter. The performance of a wheelchair is often dictated by the need to negotiate grades, as well as height transitions, such as thresholds and curbs. The clearance widths in the environment will determine the overall dimensions of the wheelchair. The climates the chair will be operated in, and the need to be able to operate in snow, rain, changing humidity and temperature levels, and other weather conditions, are important considerations as well.

A physical–motor assessment of strength, range of motion, coordination, balance, posture, tone, contractures, endurance, sitting posture, cognition, perception, and external orthoses is an important first step to obtain a basic understanding of an individual's capacity to perform different activities. The history likely provided significant insight related to their physical abilities. To verify this, a physical examination should focus on aspects of the consumer that (1) help justify the mobility aid, (2) help determine the most appropriate mobility aid, and (3) assure that medical issues are appropriately addressed.

Once the examination documents the need, or potential lack of need for the mobility device the remainder of the examination can focus on the appropriate technology. This is best assessed by giving the consumer the opportunity to try equipment to determine how functional and safe they maneuver/operate the device within the clinical space. During this part of the assessment, the consumer and family must be informed of the positive and negative characteristics of devices and services. The team needs to educate the consumer or family to participate in

choosing the device that will meet their needs (Locus of Control) and assure the provision of follow-up services.

The in-home evaluation conducted by the ATS verifies that the device is compatible and will fit within the home environment of the consumer; that may also included recreational and work environment. Once the appropriateness of a device is established, final measurements are taken. For many people, a few simple measurements can be used to determine the proper dimensions for a wheelchair (10). Body measurements are typically made with the consumer in the seated position. A Rehabilitation Technology Supplier, therapist, or other member of the rehabilitation team often completes this.

MANUAL WHEELCHAIRS

When most individuals think of manual wheelchairs they envision the boxy, steel framed, standard wheelchairs commonly seen at airports and shopping malls. These wheelchairs may be acceptable for transport of short distances, but are good for little else. Their heavy weight and complete lack of adjustability makes them poor choices for anyone using a manual wheelchair for an extended period of time.

The development of the lightweight and ultralight wheelchairs evolved in the late 1970s with a select few, extremely motivated, manual wheelchair users choosing to perform modifications on their own wheelchairs to make them faster and easier to propel (11). After these initial steps the demand became far greater for lightweight, adjustable wheelchairs and several companies were created to meet that need.

Research conducted on the new lightweight and then ultralight manual wheelchairs have quantitatively shown the benefits over the heavier, nonadjustable standard style wheelchairs. Cooper et al. (12) reported that when subjected to the fatigue tests described in the ANSI/RESNA standards, ultralight manual wheelchairs were 2.3 times more cost effective than lightweight wheelchairs and 3.4 times more cost effective than depot style wheelchairs.

The benefits of lightweight and ultralight manual wheelchairs do not end with higher cost efficiency and longevity. They are also crucial in preserving the upper extremities of their users. Because of the constant use of the upper extremities by manual wheelchair users, they tend to experience secondary injuries such as joint pain, repetitive strain injury, and nerve damage (2). Compared to standard and lightweight wheelchairs, the ultralight wheelchairs have two very distinct advantages when attempting to prevent secondary injury in manual wheelchair users: the primary advantage is the lower weight (Fig. 1).

Standard style wheelchairs tend to be greater than 36 lb(16 kg), while lightweight wheelchairs tend to be ~30–34 lb(14 + 18 kg) and ultralight wheelchairs ~20–30 lb(9–14 kg). The second advantage presented by the ultralight wheelchairs is adjustability. Lightweight wheelchairs do have some adjustability, but not to the extent of an ultralight wheelchair. Masse et al. (13) showed that moving the horizontal axle position toward the front of the wheelchair and moving the seat downward created a more efficient position for wheelchair propulsion and resulted in less exertion without loss of speed. Although some studies have been conducted to asses the importance of wheelchair setup in reducing upper extremity injury in manual wheelchair users, the solution has not been clearly defined. However, what is certain is that the adjustability and weight of manual wheelchairs are crucial parameters when selecting a wheelchair for a user that will be propelling for extended periods of time.

Another recent addition to manual wheelchairs has been suspension elements, such as springs or dampeners to reduce the amounts of vibration transmitted to manual wheelchair users. During a normal day of activity for a wheelchair user, they encounter bumps, oscillations, and other obstacles that may subject them to whole-body vibration levels that are considered harmful. VanSickle et al. (3) demonstrated that when propelling over certain obstacles, vibrations experienced at the seat of the wheelchair and the head of the user exceed the safety levels prescribed by the ISO 2631-1 Standard for evaluation of human exposure to whole-body vibration (14). Wheelchair companies have attempted to reduce the amounts of transmitted vibration by adding suspension to manual wheelchairs. Research has been done to evaluate effectiveness of suspension at reducing vibration. Results show that on average, suspension does reduce the vibration levels, however, the designs are not yet optimally effective and may not be as effective based on the orientation of the suspension elements (15,16).

POWERED ASSIST WHEELCHAIRS

Often, people using manual wheelchairs are required to transition to using powered wheelchairs or are at a level of capacity where they must choose between the two. This may be because of increased amounts of upper extremity

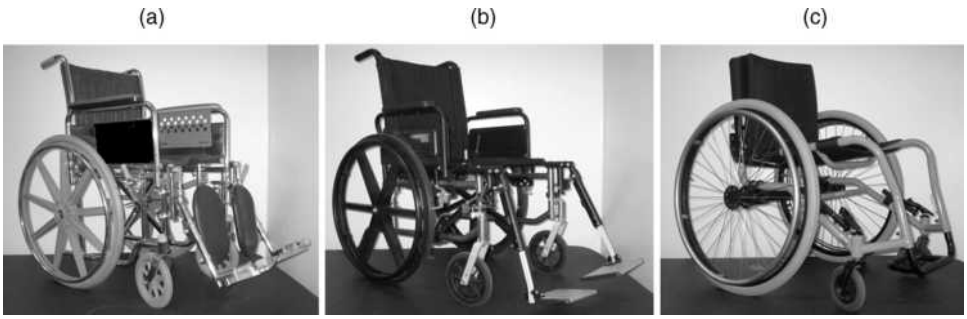


Figure 1. (a) Standard, (b) lightweight, and (c) ultralight wheelchairs.

pain, or the progression of a disease such as Multiple Sclerosis or Muscular Dystrophy. Recently, the development of Pushrim Activated Power Assist Wheelchairs (PAPAWs) has provided an alternative for these users. The PAPAW works through the use of a battery and a motor mounted directly into the wheel. The motor provides a supplement to the user so that very little force input from the user will still afford normal momentum. Transitioning from a manual wheelchair to a powered wheelchair may be difficult both physically and psychologically for some people. Users may not want to modify their homes or their cars to accommodate a powered wheelchair and may be used to providing their own mobility. Although the PAPAW provides a good intermediate wheelchair for users who may still benefit from a manual wheelchair, but do not have the strength or stamina, it also has its disadvantages. The added weight of the motor driven wheels dramatically increases the overall weight of the wheelchair and can also be difficult for users during transfers. Additionally, algorithms for the control of the PAPAWs are not yet refined and can lead to difficulty propelling the wheelchair.

POWERED WHEELCHAIRS

Powered wheelchairs represent devices that can provide an incredible amount of independence for individuals who may have extremely limited physical function. Perhaps even more than manual wheelchairs, having the proper elements and setup of a power wheelchair is vital. The lifestyle of the user, the activities in which they would like to participate, the environments to which they will be subjected, and ability level have all contribute to the correct prescription of a powered wheelchair. Many adjustments can be made to any particular powered wheelchair to specifically fit the needs of the user. Seating systems, cushions, added assistive technology to name a few. This section will focus on the characteristics that differentiate certain powered wheelchairs from one another (Fig. 2).

Powered wheelchairs come in three drive wheel setups: front, mid, and rear wheel. Each of these setups has different advantages and shortcomings. Powered wheelchair users are each unique and have specific requirements for their activities of daily living, such as maneuverability, obstacle climbing, or driving long distances. Mid-wheel drive wheelchairs tend to provide greater maneuverability because the drive wheels are located directly beneath the user's center of mass. Front-wheel drive wheelchairs are often associated with greater obstacle climbing ability. Rear-wheel drive powered wheelchairs are good for speed and outdoor travel, but may not provide the maneuverability or stability for some users. The different wheelchair setups provide the means for users to achieve their goals.

For the user, the joystick is probably the most important part of the powered wheelchair. However the standard displacement joystick is not acceptable for all users. Current technologies have allowed almost any user to operate a powered wheelchair as well as other assistive technologies, such as computers. Even the smallest abilities of the user can be translated to powered wheelchair mobility such as foot control, head control, finger control, tongue control and so on.

Like manual wheelchairs, powered wheelchairs also have different classifications. Medicare defines powered wheelchairs in three major categories: Standard weight frame powered wheelchair (K0010), Standard weight framed powered wheelchair with programmable control parameters for speed adjustment, tremor dampening, acceleration control and braking (K0011), other motorized powered wheelchair base (K0014) (17). Pearlman et al. (18) recently conducted a study examining the reliability and safety of standard weight frame powered wheelchairs. Of the 12 wheelchairs tested, only 3 passed the Impact and Fatigue section of the ANSI/RESNA Standards (19,20). Medicare has recently proposed to stop funding these powered wheelchairs, recommending the addition of a programmable controller that would place it in the second category (K0011). However, this may not be acceptable since the problems exist mainly with the drive train or the frame of the wheelchair and these parameters would not change. In order to adequately serve the interests of powered wheelchair users, the frames, drive trains, and control systems all need to perform within the standards put forward by ANSI/RESNA.

WALKERS AND ROLLATORS

Some individuals may have the ability to ambulate, however, they may get tired very easily or have visual

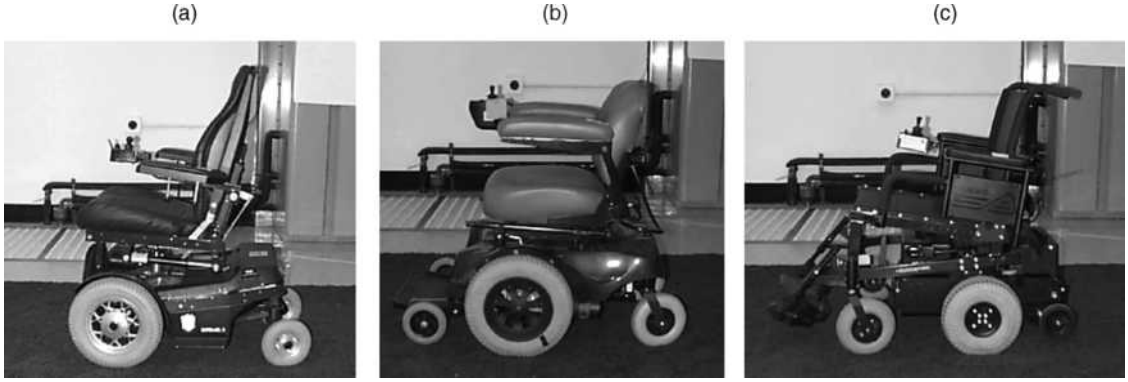


Figure 2. (a) Front-, (b) mid-, and (c) rear-wheel drive powered wheelchairs.

Figure 3. Robotic walker for users with visual impairments.

impairment problems that may result in a fall and injury. Fuller (21) reported that 33% of community-dwelling elderly people and 60% of nursing home residents fall each year. Walkers and rollators represent useful assistive technology devices for these individuals by lending support and weight relief during mobility. They may have zero, two, or four wheels and possibly have hand brakes. They may also contain a small area for sitting if the user becomes fatigued or a basket for carrying items for the user.

Some elderly persons, in addition to having mobility impairment, also have a visual impairment. Recent research has investigated a new robotic walker (Fig. 3) that through the use of sonar and infrared (IR) sensors can detect obstacles as well as provide guidance along a preprogrammed path (i.e., maneuvering around an assisted living home) (22).

SPORTS AND RECREATION DEVICES

As quality of life receives more attention, sports and recreational activities have become more important to individuals with disabilities. Sport and recreational activity participation provides many benefits to individuals with disabilities. Physical activity reduces or slows down the development of cardiovascular disease as well as modifies risk factors including high blood pressure, blood lipid levels, insulin resistance, and obesity (23). In addition, the development of muscular strength and joint flexibility gained through regular exercise improves the ability to perform activities of daily living (24). Regular exercise may help reduce clinical depression and days spent as an in-patient in a hospital, and may improve social interactions and prolong life expectancy (25). With the positive benefits of sports, exercise, and recreational activities in mind, the purpose of this section is to describe some of the more popular sports played by individual with disabilities.

Wheelchair Basketball

Wheelchair users who play basketball may have various diagnoses, such as paraplegia, cerebral palsy, amputations, post-polio syndrome, or a disabling injury. Participants are not required to use a wheelchair for their primary means of mobility or in their activities of daily living. Prior to the actual game, persons who want to play basketball must have their player classification level determined by a qualified referee. To equalize the capability of each team, the classification levels of the competitors are matched (26).

Whether players play zone or person-to-person basketball, the basic rules apply to both. Because different players have varying degrees of disability, rules have been developed that all players need to abide by. Keep firmly seated in the wheelchair at all times. A player may not use a functional leg or leg stump for physical advantage. An infraction of this rule constitutes a physical advantage foul (27).

Wheelchair basketball is similar to an everyday wheelchair, but incorporates features that enhance maneuverability (Fig. 4). Basketball wheelchairs are lightweight to allow for speed, acceleration and quick braking. The wheelchair must have four wheels. Two large, rear wheels and two front casters. The front casters are 2 in. (50 mm) in diameter and typically made from hard plastics, similar to the material used to make inline skate wheels. The rear wheels must be larger than or equal to 26 in. (338 mm) in diameter. The rear wheels must have handrims. Basketball wheelchairs use spoke guards made of high impact plastic. These guards cover the rear wheel spokes to prevent wheel damage and illegal ramming and picking. The spoke guards provide several benefits: First, spoke guards can be used to pick up the ball from the floor. Using a hand, the player pushes the ball against the spoke guard and rolls it onto their lap, Second, spoke guards protect hands and fingers from injury when reaching for the ball near another player's rear wheel. Third, they provide space to identify team affiliations and sponsor names. Camber is an important feature of basketball wheelchair as well. Camber is defined as"the angle of the main wheel to the vertical", or as a situation in which"the spacing between the top

Figure 4. Players competing in a friendly game of wheelchair basketball.

points of the wheels may be less than the spacing between the bottom points". Increasing camber slightly reduces the height of the seat, while it proportionally increases the wheelbase, which corresponds to the width of the wheelchair. In the same way, with negative camber, the center of gravity of the occupied wheelchair moves backward. From a practical point of view, increased wheel camber improves hand protection as chairs pass through doors and, in terms of basketball, camber makes a wheelchair more responsive during turns and protects players' hands when two wheelchairs collide from the sides, by limiting the collision to the bottom of the wheels and leaving a space at the top to protect the hands. Basketball wheelchair seats typically have a backward seat angle slope of 5–15°. The angle of the seat compared to the ground is known as "seat angles". Guards are an exception. Guards are allowed to have lower seat heights and greater seat angles. These modifications make chairs faster and more maneuverable for ball handling.

Wheelchair Racing

Individuals with all levels of SCI as well as lower limb amputees can participate in competitive races. The preferred racing chair among racers is the three-wheel chair (Fig. 5). The three-wheel design is constructed from high pressure tubular tires, light weight rims, precision hubs, carbon disk/spokes wheels, compensator steering, small push rings, ridged aluminum frame, and 2–15° of wheel camber. The camber in a racing chair makes the chair more stable and allows the athlete to reach the bottom of the pushrim without hitting the top of the wheels or pushrim.

Hand-Cycling

Cycling has been a popular outdoor sport for several years. The adaptability of cycling to different terrains makes it a favorite for many. Adaptive equipment for bicycles consists of a hand cycle allows individuals with limited use of their legs to utilize the strength of their arms (28). A handcycle typically consists of a three-wheel setup to compromise for the balance required when riding a two-wheeled bicycle. Two-wheeled handcycles do exist but require a great deal of skill and balance. Handcycle designs allow the user to propel, steer, break, and change gears, all with the upper extremities and trunk. Two types of handcycle designs are readily available (1) the upright and (2) the recumbent. In an upright handcycle, the rider remains in an upright position similar to the position the body takes when seated in a touring bike. Upright handcycles use a pivot steer to turn. Only the front wheel turns while the cycle remains in an upright position. Transferring and balancing tend to be easier on the upright cycle. In the recumbent handcycle, the rider's torso reclines and the legs are positioned out in front of the cyclist. These cycles use a lean-to-steer mechanism. The rider leans to turn, causing the cycle to pivot at hinge points. Leaning to turn can be challenging if the rider lacks trunk stability, in which case a pivot steering recumbent handcycle may be more appropriate. Recumbent handcycles are lighter and faster, making them the choice for hand cycle racing. Relatively minimal modifications are needed to accommodate individuals with tetraplegia. Some of the modifications include hand cuffs that can be mounted to the arm crank handles and elastic abdominal binders which can be fitted around the user and the handcycle seat to increase trunk stability.

Wheelchair Rugby

Rugby is played indoors on a large gym on a basketball court surface. Players use manual wheelchairs specifically designed for the sport. Due to the level of contact, the chairs have protective side bars on them and players are strapped in to prevent injury. Most chairs are made of titanium or steel to handle the hits that they sustain. In addition, the low pointers have a high camber (angle of the wheels) so that they can turn fast, as well as "red push rim covers so they can actually stick to the other person's chair and hold them." The high pointers have armor on the front of their chairs resembling a cow catcher so that they can push through the other players without getting stuck (Fig. 6).

To be eligible to play rugby, players must have a combination of upper and lower extremity impairment. Most of the players have sustained cervical level spinal injuries and have some degree of tetraplegia. Like in basketball, players receive a classification number based on there level of impairment (29). Rugby consists of two teams comprised of four players. The object of the game is for a player to have possession of the ball and cross the opponent's goal line.

Rugby wheelchairs are strictly regulated to ensure fairness. However, chairs may vary considerably depending on

Figure 5. Shows a racing three-wheeled wheelchair.

Figure 6. Wheelchair rugby.

a player's preferences, functional level and team role. Team roles may be assigned according to ability. Players with upper body limitations tend to perform the defensive blocking and picking roles. They use chairs that have additional length and hardware. All rugby chairs have extreme amounts of camber, 16–20°, significant bucketing, and antitip bars. The camber provides lateral stability, hand protection, and ease in turning. The bucketing (knees held high relative to rear end) helps with trunk balance and protection of the ball.

Tennis

Tennis players compete in both singles and doubles games. Players are required to have a permanent mobility-related physical disability that requires a wheelchair as the primary means of mobility. Tennis is played on the traditional tennis court using the tradition size and height tennis net. However, unlike traditional tennis, the ball is permitted two bounces on the court before it must be returned. Brakes are not permissible as stabilizers and the athlete must keep one buttock in contact with the seat at all times.

Tennis players use a three-wheeled chair with a large amount of camber to maximize mobility around the court. The seat is situated at a steep backwards seat angle slope. The angle helps with balance, keeps players against the seat backs, and gives them greater control over the wheelchair. The knees tend to be flexed with the feet on the footrest behind the player's knees. With the body in a relatively compact position, the combined inertia of rider and wheelchair is reduced, making the chair more maneuverable (30). Handles and straps can also be added to the chair. Many players incorporate plastic rigid handles into the front of the seat. Players use these handles when leaning for a shot or making quick directional changes. Straps can be used around the waist, knees and ankles, to help with balance (31).

Adaptive Skiing

Skis for skiers with disabilities have advanced state-of-the-art skis that offer shock absorption systems, frames molded to body shape, and quick release safety options. Skiers with disabilities can maintain a similar pace to that of unimpaired athletes with the development of adaptive seating, backrests, cushions, tethering ropes, roll bars and outriggers. Outriggers, an adapted version of a forearm crutch with a shortened ski, provide extra balance and steering maneuverability (32). Two types of sit-down adaptive skies are available: Bi and Mono. Bi skis are appropriate for skiers with limited trunk stability. With Bi skis, the skier balances on two skies and angulates and shifts to put the skis on edge. Bi skis have wider base of support, can usually be mastered quickly with few falls and are easier to control than a mono ski. The Mono Ski is the ski of choice for individuals who want high end performance, maneuverability and speed. With a mono ski, the skier sits relatively high on the seat of the ski over the snow. The skier uses upper body, arm and head to guide their movement down the hill. Sit-, Mono- and Bi-skis have loading mechanisms, usually hydraulic, that enable the individual to raise themselves to a higher position for transferring onto a ski lift.

TRANSPORTATION SAFETY AND ADAPTIVE DRIVING FOR WHEELCHAIR USERS

Wheelchair users, like the entire population, use several forms of transportation to travel from place to place: They are passengers in public transportation systems (bus, subways, and vans) and private vehicles, and are potential drivers of each of these types of vehicles. To ensure the safety of *all* passengers and drivers, certain safety mechanisms must be in place: drivers must be able to safely control the vehicle, and all seated passengers require seats securely fastened to the vehicle and passenger restraints (e.g., seatbelts) which can secure the passenger to the seat. The requirement of passenger restraints is relaxed for passengers in large vehicles, like busses and subways, because of the low likelihood of high velocity crashes (33). In many cases, adaptation of a vehicle is necessary when the original equipment manufacturer (OEM) control and/or securement mechanism cannot provide adequate safety for wheelchair users because either (1) sensory and/or motor impairments of the user requires adaptive driving equipment so the vehicle can be safely controlled, or (2) the user cannot safely or effectively use the OEM seat or passenger restraint system in the vehicle.

Vehicle Control Systems

The complexity of the adaptive equipment required for safe control of the vehicle is correlated to the type and level of impairment of the wheelchair user. Adaptive driving equipment can be as low tech as attaching a knob on a steering wheel, and as high tech as fly-by-wire technology, where computer-controlled actuators are added to all of the controls, and the drive interfaces with the computer (via voice and/or low or no-effort sensors). Adding actuators to all of the driving and operating controls of the vehicle through computer controls.

An example of this range is the types of steering adaptations available for people with disabilities. Figure 7 demonstrates both a typical steering knob for a person with little to no loss of hand sensory-motor function (a), and (b) a knob for someone with loss of some sensory motor function. Both types of steering knobs serve the same purpose: They allow the driver to safely steer the vehicle with one hand while (typically) their other hand is operating a hand control which actuates the fuelaccelerator and brake pedals. When upper-extremity sensory-motor function does not allow for safe turning of the OEM steering system (even with a knob), actuators can be used in lieu of upper-extremity function. These systems are named "low-" or "no-effort" steering systems, depending on the type of assistance that the actuators provide. Retrofitted controls for these systems are used and typically require removal of the OEM equipment (e.g., the steering wheel and/or column). Consequently, when this level of technology is used, it usually becomes unsafe for an unimpaired individual to drive the vehicle (without significant training).

Figure 7. Steering knobs.

Common fuelaccelerator and braking system hand controls are bolted to the OEM steering column, and actuate each pedal with mechanical rods (Fig. 8). These types of controls require nearly complete upper extremity function to operate. When upper extremity function is substantially impaired actuators are added to the braking and fuelaccelerator systems and are operated through some switching methods. The types of switches depend on the most viable control mechanism for the user: in some cases, a simple hand-operated rheostat variable resistance switch (i.e., dimmer switch or rheostat) may be used, and in other cases, a breath-activated pressure switch (sip-and-puff) system may be used.

The above steering, fuelaccelerator, and brake adaptive equipment are focused on the primary control systems of the vehicle (those which are required to drive the automobile). Various adaptive equipment can control the secondary control system of the vehicle also (e.g., ignition switch, climate control, windows). Like the primary control adaptive equipment, equipment to modify the secondary controls of the vehicle range from low to high tech. For example, users with impaired hand function may require additional hardware to be bolted to the ignition key so they can insert the key and turn it to start the vehicle. Alternatively, hardware can be added to allow a drive to start the vehicle via a switch, either remotely or from within the cabin. Power windows and door-lock switches can be rewired to larger switches, or in more accessible locations for the driver.

Figure 8. Common hand controls.

Both primary and secondary control systems must be placed in locations easily accessible to the driver. In some cases, wheelchair riders will transfer out of their wheelchair directly into the OEM seating system, allowing for most controls to remain in their OEM locations. If a wheelchair user remains in their wheelchair and drives the vehicle the controls must be made accessible to their seated position and posture. In these cases, along with the case a wheelchair users remaining in their wheelchair as passenger, provisions must be made to safely secure the wheelchair to the vehicle, and to provide adequate passenger restraints.

Wheelchair Tie-Down and Occupant Restraint Systems (WTORS)

For both practical and safety reasons, when a wheelchair user remains in wheelchair while riding in a vehicle they must be safely secured to the vehicle. An unsecured wheelchair will move around, causing the user to be unstable while the vehicle is moving. Being that power wheelchairs can be in excess of 200 lb (91 kg) in weight, This instability could cause harm to the wheelchair rider and/or the surrounding other passengers vehicle occupants. If the wheelchair user is driving, they may lose control of the vehicle if they accidentally roll away from the vehicle controls. Another important concern for an unsecured wheelchair rider is in the case of an accident. An unsecured wheelchair and user can easily be ejected out of the vehicle if they are not secured. The WTORS systems are currently governed by the ISO 7176-19 (34).

Several types of tie-down systems exist, including a four-point belt system and various latching-type mechanisms which typically require hardware to be attached to the wheelchair. The four-point belt systems are most common, and are found on public busses, and also private vehicles. Theses systems are the most widely used tie-down system because they can be attached to a wide variety of wheelchairs. In some cases, manufacturers incorporate attachment rings for these tie-down systems into their wheelchair. When no points of attachment are available (most common situation), points at the front and rear of the seat or frame be used. These attachment points must be sufficiently strong to secure the wheelchair in the event

of a crash, and be in locations which will allow the straps to be oriented within a specified range of angles with the horizontal (front straps: 30–60°, rear: 30–45°).

Unfortunately, tie-down systems are they are not convenient to use: a second person (other than the wheelchair user) is typically needed to help secure the wheelchair, making the operation laborious, and in some cases awkward for the wheelchair users who may not be comfortable with another person touching their wheelchair or encroaching on their personal space. Consequently, and especially on public transportation, these systems are commonly unused and the wheelchair user relies on their brakes wheel-locks for stability, risking their own safety in a crash.

Other mechanisms have been used to secure a wheelchair to the vehicle. These included wheel-clamps and t-bar systems. With these mechanisms, wheelchairs are secured to the vehicle through a mechanical clamp that adjusts to the wheelchair size. These systems are quicker to attach to the wheelchair, but are still difficult or impossible for a user to use independently.

A variety of wheelchair tie-down systems have been developed that allow the wheelchair users to independently lock and unlock their wheelchair to the vehicle. A common one used for personal vehicles is the EZ-Lock System, which is a hitch system for the wheelchair. This system allows the wheelchair user to maneuver the wheelchair so a specialized hitch attached to the wheelchair is captured into a latch bolted to the vehicle; both electric and manual release mechanisms can be used to unhitch the wheelchair from the device, allowing for custom placement of a hitch for easy accessibility to the wheelchair user. A drawback to this system is the specialized hardware that must be attached to the wheelchair that restricts folding a manual wheelchair and reduces ground clearance.

Because this system is designed to allow the user to drive forward into the device, it works well and is common in private vehicles where the wheelchair user drives the vehicle. In larger vehicles, such as public busses, it is typically more convenient for a user to back into a spot and lock their wheelchair.

An ideal system for public transportation would be one that a user can operate independently and that does not require specific hardware to be attached to the wheelchair that may not work on *all* wheelchair models. A system is currently being developed at the University of Pittsburgh that tries to achieve these goals.

Occupant restraint systems are the last requirement to allow a wheelchair user to safely travel in a vehicle. These restraint systems mimic the function of a seat belt, and can be either attached to the wheelchair (integrated restraint) or to the vehicle (Fig. 4). In both cases, the placement of the restraints with respect to the body is critical to prevent injury in a crash—either through direct insult of the seatbelt with the body, or because of submarining (where the torso slides down under the pelvic belt).

To ensure these WTORS systems and the wheelchair themselves can safely survive a crash, standards testing is in place. Rehabilitation Engineering and Assistive Technology Society of North America (RESNA), International Standards Organization (ISO), and Society of Automotive Engineers (SAE) have worked in parallel to establish minimum standards and testing methods to evaluate wheelchairs and WTORS systems (35). These tests mimic those performed on OEM seat and occupant restraint systems, which suggest the system should be able to withstand a 20 g crash (36). To encourage and guide wheelchair manufacturers to build their wheelchairs to these standards researchers have developed a website to inform all relevant stakeholders of the latest information (http://www.rercwts.pitt.edu/WC19.html).

LOWER EXTREMITY PROSTHETICS

Prosthetics are devices that replace the function of a body organ or extremity, unlike orthotic devices, which support existing extremities. Prosthetics range from simple cosmetic replacements to complicated structures that contain microprocessors for controlling hydraulic and pneumatic components. Commonly used prosthetic devices primarily include artificial limbs, joint implants, and intraocular lenses. Approximately, 29.6–35.4% of the U.S. population use prosthetic limbs (37) with >2% of them aged between 45 and 64 years using lower extremity (LE) prosthetics for mobility (U.S. Bureau of Census, 2000). Amputation, resulting from peripheral vascular diseases in the older population (60 years and older) and trauma in young population can be considered factors for the use of LE prosthetics.

Research and development in clinical practice has resulted in recent advances in the area of prosthetics designs and controls technology. Examples of these advances include the use of injection molding technology for socket manufacturing, shock absorbing pylons, the incorporation of neuro-fuzzy logic microprocessor-based controllers for myoelectric prostheses, and microprocessor-controlled prosthetic knees and ankles (38,39).

Prosthetic feet classified as"uniaxial" allow for movement at a single axis in one plane, such as plantarflexion and dorsiflexion of the ankle. In this type of prosthetic foot, the heel is typically composed of the same density materials as the rest of the foot, with an option of different heel height. Also, uniaxial feet have different options at the rubber toe section in terms of flexibility, which depends on the weight of the individual. Multiaxial prosthetic feet (MPFs) have five degrees of motion in three planes: plantarflexion–dorsiflexion, inversion/eversion, and rotation. This feature provides stability to the user, while walking on uneven surfaces and also aid in shock absorption lowering intensity of shear forces on residual limb. Elastomer or rubberized material is used to alter, resist, or assist with the different degrees of motion in the prosthetic foot. MPFs also provide options for different heel heights [0.5–1 in. (13–26 mm)] and different degrees of toe material resistance, while split internal structures in the heel assist with inversion/eversion on uneven ground. The MPFs are prescribed by the weight and shoe size of the consumer.

The solid ankle, cushion heel (SACH) prosthetic foot is the most commonly prescribed prosthetic foot for lower extremity amputations. The SACH foot is constructed out of eurothene (plastic) materials with a less dense material

Table 1. Functional Level and Devices

	Functional Level	Type of Device
K0	No ability to ambulate or transfer safely; prosthesis does not enhance mobility	Cosmesis
K1	Transfers and ambulates on level surfaces; household use	SACH
K2	Able to negotiate over low level environmental barriers; limited community ambulation	Low level energy storage feet
K3	Prosthetic usages are beyond simple ambulation; able to traverse MOST environmental barriers and is a community ambulator	Energy storage prosthesis
K4	Able to perform prosthetic ambulation exceeding basic skills (i.e., high impact); child, active adult and athlete	Energy storage prosthesis

incorporated at the heel. Use of materials with different densities permits proper positioning while standing, as softer heels aides in enhancement of the walking efficiency after heel strike, by shifting center of gravity forward. A device known as durometer is used to measure the density of plastics used in prosthetic devices. The weight and activity level of the individual using the prosthesis determines which heel density is selected, as heavier user require firmer heel cushion. The stiffness of heel, also determine amount of knee flexion and shock absorption. Greater the heel stiffness more the knee flexion and lower shock absorption during heel strike and vice versa. The SACH foot also contains a keel made out of a hard wood or composite material. Belting material is applied to the keel, which prevent the keel it from breaking through the eurothene cover. During ambulation, the foot simulates plantar flexion movement and prevents the loss of anterior support during the push off at the toe.

Individuals who use foot prosthetic devices are assessed for weight, potential activity levels, and type of use for which they anticipate using their prosthetic devices. Based on this assessment, clients are then categorized into four functional levels:

Energy Storage and Return (ESAR) prosthetic feet are fabricated to assist with the dynamic response of feet, acting as a diving board from which a person can push off during walking. These feet have capability to store energy during stance phase and return it to the user to assist in forward propulsion in late stance phase. The ESAR has flexible keels and are prescribed by the anticipated activity level and weight of the person. Also, limited evidence suggests use of ESAR as their use results in increasing ambulation speed and stride length ~7–13% greater than with a conventional (SACH) foot in both traumatic and vascular transtibial amputees (40).

Macfarlane et al. (40) compared energy expenditure of individuals with transfemoral amputations who walked with a SACH foot versus a Flex-Foot prosthetic. The SACH has a solid ankle and cushioned heel construction, while the Flex-Foot prosthetic has a hydraulic knee joint. The authors determined that Flex-Foot walking resulted in significantly lower exercise intensity, reduced energy expenditure and improved gait efficiency. These findings are significant considering the SACH foot is the most commonly used foot prosthetic in the U.S. today (41). Lower energy expenditure was also reported for individuals with trans-tibial amputation with the use of Flex-Foot as compared with a SCAH foot.

Higher level of limb loss results in addition of more prosthetic components. Prostheses for transfemoral amputations comprised of four basic components: the socket, the knee joint, the pylon, and the foot. Pylons are classified as: exoskeleton in which the weight of the individual is supported by the external structure of the prostheses (i.e., a crustacean shank), or endoskeleton that is comprised of an internal, weight-bearing pylon encased in moldable or soft plastics (i.e., modular pylon). The knee mechanism use, a conventional damping system, where a flow of (fluid or air) is controlled by a valve and its operation is set for a particular walking speed according to user's preference. The system described as intelligent prosthesis (IP), where a diameter of damping controlling valve is changeable according to varying speed of walking (42). Romo provided guidance on the selection of prosthetic knee joints and indicated that proper alignment impacts the effectiveness of matched and adjusted knee joints for smooth and reliable gait (43). Taylor et al. (44) compared effectiveness of an intelligent prosthesis (IP), and pneumatic swing-phase, control-dampening systems while walking on a treadmill at three speeds of 1.25, 1.6, and 2 mph (2, 2.6, and $3.2\,\mathrm{km}\cdot\mathrm{h}^{-1}$). The results indicated lower $V\mathrm{O}_2$ consumption for individuals using IP compared to controls-damping system at 2 mph ($3.2\,\mathrm{km}\cdot\mathrm{h}^{-1}$). The question often raised by critiques concerns the cognitive demands by the high end technology on the users. The results of the study by Heller et al. (42) that investigated cognitive demand when using the IP compared to a conventional prosthesis indicated no significant differences while using these prostheses for ambulation. Though not uncommonly prescribed high rejection rates has been described for prostheses after hip disarticulation and hemipelvectomy. These prostheses consist of an addition of hip joint mechanism to other parts similar to prostheses prescribe after transfemoral amputation.

Modular systems were first developed in the 1960s by Otto Bock, which consisted of shock absorbing pylons that contained with shock absorbers. Also, a reverse-pyramid fixture at the ends of the pylon permits angular adjustments to the alignment of these devices with the residual limb. Modular systems are lighter than the earlier wooden

systems, allow for 15° of movement gain in either the frontal or sagittal plane, and also permit internal and external rotational adjustments. Modular systems can extend the life of a prosthetic device, as worn parts can be replaced. In addition, individuals using these systems experienced less need for maintenance.

A significant improvement in the design procedure of the prosthetics considers the interaction of forces between prosthesis and residual limb can be found in the literature. Jia et al. (45) studied the exchange of loads and forces between the residual limb and prosthetic socket in transtibial amputation using the Finite Element Analysis (FEA) method. Lee et al. (46) used FEA to determine contact interface between the transtibial residual limb and prosthetic socket. The study determined the need for sameness of shapes for both the residual limb and socket in order to decrease peak normal and shear stresses over the patellar tendon, anterolateral and anteromedial tibia, and popliteal fossa. Winson et al. investigated the interaction between socket and residual limb during walking using a FEA model for transtibial prosthesis. Pylon deformities and stress distribution over the shank were problems identified during walking and results indicated need for pylon flexibility for better optimization and need of future studies identifying fatigue life of these prostheses (47).

With advancement in the area of prosthetics designs and development, simultaneous factors that need to be considered, use of these devices in clinical practice for targeted population and cost containment. Premature abandonment of mobility assistive devices, which might be due to poor performance and/or changes in the need of the user, is not uncommon and adds to the expense of these devices (48). Improved quality of service delivery for LE prostheses, which include identifications of reasons for successful use or nonuse of LE prostheses, is needed (49). Also, incorporation of standardized performance testing procedure to ensure durability of LE prosthetics is vital to the appropriate prescription of, and satisfaction with, prosthetic devices.

Prosthetic devices of today incorporate advancements from the aerospace and engineering fields and include the use of new materials, such silicone elastomer gel sleeves in to assist in the fit of prosthetic sockets, prosthetic feet made from carbon-fiber composite components that are lighter in weight, and surgical implantation of titanium prosthetic attachment devices directly to bones of residual limbs (50,51). Neuro- and microprocessors and sensors are now incorporated on-board the prosthetic device to control knee joint movement to improve the symmetry of different gait patterns across a variety of cadence speeds. Hydraulic or pneumatic devices are also used to dampen the swing-through phase of walking with the prostheses to assist with walking at difference cadences (52,53). Manufacturers are now using computer-aided design and manufacturing techniques to improve the fit of the prosthetic sockets as well as component designs (54,55).

Because of the growing population of people in need of mobility aids, and their demand to maintain their lifestyle, whether that includes going to and from work, participating in extracurricular activities, or maneuvering around their environment, continuing information must be gathered and disseminated to make these goals achievable. Through technological advancements people who require mobility aids can accomplish more of their goals than ever before, however there are still people for whom the technology is not yet developed enough or cannot obtain the proper devices to meet their needs. It is for this reason that problems must continually be studied and innovations must advance so that mobility aids will serve anyone who requires them to meet their goals.

BIBLIOGRAPHY

1. Schiller JS, Bernadel L. Summary Health Statistics for the U.S. Population: National Health Interview Survey, National Center for Health Statistics. Vital Health Stat 10(220):2004.
2. Sie IH, Waters RL, Adkins RH, Gellman H. Upper extremity pain in the postrehabilitation spinal cord injured client. Arch Phys Med Rehab 1992;73:44–48.
3. VanSickle DP, Cooper RA, Boninger ML, DiGiovine CP. Analysis of vibrations induced during wheelchair propulsion. J Rehab R&D 2001,38:409–421.
4. Calder CJ, Kirby RL. Fatal wheelchair-related accidents in the United States. Am J Phys Med Rehab 1990;69:184–190.
5. Mills T, et al. Development and consumer validation of the Functional Evaluation in a Wheelchair (FEW) instrument. Disabil Rehab 2002;24(1–3):38–46.
6. Chase J, Bailey DM. Evaluating potential for powered mobility. Am J Occup Therapy 1990;44(12):1125–1129.
7. Cooper RA, Cooper R. Electric Powered Wheelchairs on the Move. Physical Therapy Products; July/August, 1998, p 22–24.
8. Galvin JC, Scherer MJ. Evaluating, Selecting, and Using Appropriate Assistive Technology. Gaitherburg (MD):Aspen Publishers Inc.; 1996.
9. Cooper RA. Rehabilitation Engineering Applied to Mobility and Manipulation. Bristol (UK): Institute of Physics; 1995.
10. Grieco A. Sitting posture: An old problem and a new one. Ergonomics 1986;29:345–362.
11. Cooper RA. A perspective on the ultralight wheelchair revolution. Tech Disab 1996;5:383–392.
12. Cooper RA, et al. Performance of selected lightweight wheelchairs on ANSI/RESNA tests. Arch Phys Med Rehabil 1997;78:1138–1144.
13. Masse LC, Lamontagne M, O'Riain. Biomechanical analysis of wheelchair propulsion for various seating postitions. J Rehab R&D 1992;29:12–28.
14. International Standards Organization, Evaluation of Human Exposure to Whole-Body Vibration—Part 1: General Requirements. ISO 2631-1, Washington (DC): ANSI Press; 1997.
15. Wolf EJ, et al. Analysis of whole-body vibrations on manual wheelchairs using a Hybrid III test dummy. Proceedings of the annual RESNA conference; 2001. p 346–348.
16. Kwarciak A, Cooper RA, Wolf EJ. Effectiveness of rear suspension in reducing shock exposure to manual wheelchair users during curb descents. Proceedings of the annual RESNA conference; 2002. p. 365–367.
17. Centers for Medicare and Medicade Services (CMS), http://www.cms.hhs.gov/providers/pufdownload/anhcpcdl.asp, Accessed 2005. Jan 12.
18. Pearlman J, et al. Economical (K0010) Power Wheelchairs Have Poor Reliability and Important Safety Problems: An ANSI/RESNA Wheelchair Standards Comparison Study. Proceedings of the Annual RESNA Conference; 2005.
19. American National Standard for Wheelchairs, Volume 2, Additional Requirements for Wheelchairs (Including Scooters) With Electrical Systems. Virginia: Rehabilitation Engineering and Assistive Technology Society of North America; 1998.

20. American National Standard for Wheelchairs, Volume 1, Requirements and Test Methods for Wheelchairs (Including Scooters). Virginia: Rehabilitation Engineering and Assistive Technology Society of North America; 1998.
21. Fuller GF. Falls in the Elderly Am Fam Physician 2000;61(7):2159–2168.
22. Rentschler AJ, et al. Intelligent walkers for the elderly: Performance and safety testing of the VA-PAMAID robotic walker. J Rehab R&D 2003;40(5):423–432.
23. Abel T, et al. Energy Expenditure in wheelchair racing and hand biking- a basis for prevention of cardiovascular diseases in those with disabilities. Eur J Cardio Prev Rehab 2003;10(5):371–376.
24. Franklin BA, Bonsheim K, Gordon S. Resistance training in cardiac rehabilitation. J Cardiopulm Rehab 1991;11:99–107.
25. Kenedy DW, Smith RW. A comparison of past and future leisure activity participation between spinal cord injured and non-disabled persons. Paraplegia 1990;28:130–136.
26. National wheelchair basketball association (2003–2004) official rules and case book retrieved from National Wheelchair Basketball Association. http://www.mwba.org. Accessed.
27. Vanlandewijck Y, Daily C, Theisen DM. Field test evaluation of aerobic, anaerobic, and wheelchair basketball skill performances. Int J Sports Med 1999;20(8):548–554.
28. Janssen TWJ, Dallmeijer AJ, Van der Woude LHC. Physical capacity and race performance of handcycle users. Jour Rehab R&D 2001;38(1):33–40.
29. Lapolla T. International rules for the sport of wheelchair rugby. http://quadrugby.com/rules.htm. Accessed 2000.
30. Wheelchair tennis handbook. International Tennis Federation. http://www.itfwheelchairtennis.com. Accessed 2004.
31. Kurtz M. Difference makers. Sports' N Spokes 2002;28(2):10–14.
32. Russell JN, et al. Trends and Differential use of Assistive Technology Devices: United States, 1994. Adv Data 1997; 292:1–9.
33. Shaw G, Gillispie B. Appropriate portection for wheelchair riders on public transit buses. J Rehab R&D 2003;40(4):309–320.
34. International Standards Organization, Wheeled mobility devices for use in motor vehicles. ISO 7176-19, Vol. 31. 2004.
35. Hobson D. Wheelchair transport safety - the evolving solutions. J Rehab R&D 2000;37(5).
36. Bertocci G, Manary M, Ha D. Wheelchair used as motor vehicle seats: Seat loading in frontal impact sled testing. Med Eng & Phys 2001;23:679–685.
37. Mak AF, Zhang M, Boone DA. State-of-the-art research in lower-limb prosthetic Biomechanics socket interface: a review. J Rehab R&D 2001;38(2):161–174.
38. Weir RF, Childress DS, Grahn EC. Development of Externally-Powered Prostheses for Persons with Partial Hand Amputations. Proceedings of the Chicago 2000 World Congress on Medical Physics and Biomedical Engineering; 2000 July 23rd–28, Chicago (IL):.
39. van der Linde H. A systematic literature review of the effect of different prosthetic components on human functioning with a lower-limb prosthesis. J Rehab R&D 2004;41(4):555–570.
40. Macfarlane PA, et al. Transfemoral amputee physiological requirements: comparisons between SACH foot walking and flex-foot walking. J Prosthe & Ortho 1997;9(4):138–143.
41. Nielsen DH, Shurr DG, Golden JC, Meier K. Comparison of Energy Cost and Gait Efficiency During Ambulation in Below-Knee Amputees Using Different Prosthetic Feet - a Preliminary Report. J Prosthet Orthotics 1989;1(1):24–31.
42. Heller BW, Datta D, Howitt J. A pilot study comparing the cognitive demand of walking for transfemoral amputees using the Intelligent Prosthesis with that using conventionally damped knees. Clin Rehab 2000;14(5):518–522.
43. Romo HD. Prosthetic knee. Phys Med Rehab Clin N Am 2000;11(3):595–607.
44. Taylor MB, Clark E, Offord EA, Baxter C. A comparison of energy expenditure by a high level trans-femoral amputee using the Intelligent Prosthesis and conventionally damped prosthetic limbs. Prosthet Ortho Int 1996;8:116–121.
45. Jia X, Zhang M, Lee WC. Load Transfer Mechanics Between Trans-Tibial Prosthetic Socket and Residual Limb - Dynamic Effects. J Biomech 2004;37(9):1371–1377.
46. Lee WC, Zhang M, Jia X, Cheung JT. Finite Element Modeling of the Contact Interface Between Trans-Tibial Residual Limb and Prosthetic Socket. Med Eng Phys 2004;26(8):655–662.
47. Winson CCL, Zhang M, Boone D, Contoyannis B. Finite-Element Analysis to Determine Effect of Monolimb Flexibility on Structural Strength and Interaction Between Residual Limb and Prosthetic. J Rehab R&D 2004;41(6a):775–786.
48. Phillips B, Zhao H. Predictors of Assistive Technology Abandonment. Assis Technol 5(1):1993; 178–184.
49. Scherer MJ. The change in emphasis from people to person: introduction to the special issue on assistive technology. Disabil Rehab 2002;24(1–3):1–4.
50. Marks LJ, Michael JW. Science, Medicine, and the Future: Artificial Limbs. BMJ 2001;323(7315):732–735.
51. Beil TL. Interface Pressures During Ambulation Using Suction and Vacuum-Assisted Prosthetic Sockets. J Rehab R&D 2002;39(6):693–700.
52. Michael JW. Modern Prosthetic Knee Mechanisms. Clin Orthop Relat Res 1999;361:39–47.
53. Buckley JG, Spence WD, Solomonidis SE. Energy Cost of Walking: Comparison of"Intelligent Prosthesis" With Conventional Mechanism. Arch Phys Med Rehabil 1997;78(3):330–333.
54. Twiste M, Rithalia SV, Kenney L. A Cam-Displacement Transducer Device for Measuring Small Two-Degree of Freedom Inter-Component Motion in a Prosthesis. Med Eng Phys 2004;26(4):335–340.
55. Lee WC, Zhang M, Jia X, Cheung JT. Finite Element Modeling of the Contact Interface Between Trans-Tibial Residual Limb and Prosthetic Socket. Med Eng Phys 2004;26(8):655–662.

See also BLIND AND VISUALLY IMPAIRED, ASSISTIVE TECHNOLOGY FOR; ENVIRONMENTAL CONTROL; LOCOMOTION MEASUREMENT, HUMAN; REHABILITATION AND MUSCLE TESTING.

MODELING OF PHYSIOLOGICAL SYSTEMS.

See PHYSIOLOGICAL SYSTEMS MODELING.

MODELS, KINETIC.

See TRACER KINETICS.

MONITORING IN ANESTHESIA

TARMO LIPPING
VILLE JäNTTI
ARVI YLI-HANKALA
Tampere University of Technology
Pori, Finland

INTRODUCTION

Anesthesia is one of the most complex and mysterious phenomena in clinical work. The main feature of anesthesia

is the loss of consciousness, which suggests its relatedness to sleep, epilepsy, and various kinds of brain trauma. In the case of anesthesia and sedation, consciousness is manipulated deliberately to prevent the patient from being aware of their state and the medical procedures carried through.

Recent decades have seen significant advancements in mapping various psychological functions to corresponding brain areas, however, the knowledge of the formation of human consciousness is still based on uncertain hypothesis. This complexity makes anesthesia monitoring extremely challenging.

This article first addresses the problem of anesthesia monitoring from the clinical, as well as from the physiological, point of view. The main emphasis is on monitoring anesthetic depth as this is the most discussed topic in anesthesia monitoring today. It starts with clinical indicators of anesthetic depth, gives an overview on the methods used in commercially available depth-of-anesthesia monitors, and describes some new algorithms proposed and evaluated for anesthesia electrocardiograms (EEG) monitoring in recently published works. Finally, the feasibility of monitoring brain function is argued using neurophysiological parameters like EEG, Auditory Evoked Potentials (AEPs), and so on, in the Intensive Care Unit (ICU) and the Emergency Room (ER).

ANESTHESIA AS A PROCESS AND A PROCEDURE

Anesthesia can be seen from the clinical point of view as a procedure, carried out according to certain protocol. From the physiological point of view anesthesia is a process evolving in the nervous system as the dose of an anesthetic agent increases.

Anesthesia as a Procedure

The goal of general anesthesia in the operating room (OR) is to render the patient unaware so that they do not feel pain during the surgery or recall the events afterward. It is also important that the patient does not react to surgical stimuli by movement. In the ICU, the goal of sedation is to keep the patient calm and painless. Too deep anesthesia causes prolonged awakening times after surgery in OR and longer treatment times in the ICU. The goals of anesthesia and sedation can be achieved by hypnotics (unconsciousness producing drugs), analgesics (antinociceptive drugs), and neuromuscular blocking agents. The choice of drugs is mainly made based on experience and clinical signs during the treatment.

Although general anesthesia is considered a safe procedure, various complications like postoperative nausea, vomiting, and pain are relatively frequent. The incidence of recall of events and awareness during anesthesia is rare ($\sim 0.1\%$), however, the consequences may be traumatic for the patient (1). Anesthesia-related mortality has decreased significantly during past decades having recently been estimated at 1 death per 200,000–300,000 cases of anesthesia (2–4).

Anesthetic agents can be divided into inhalation anesthetics (e.g., halothane and isoflurane) and intravenous anesthetics (e.g., thiopental and propofol). Intravenous drugs are becoming more popular as they are short acting, do not cause gas pollution, are easy to administer, and do not cause airway irritation. Desirable properties of anesthetic agents include rapid, smooth and safe induction of and emergence from anesthesia, no accumulation in the body, minimal effects on cardiovascular functions, no irritation to tissues and veins, low potential to hypersensitivity reactions.

Anesthesia as a Process

During the last decades, it has become obvious that different anesthetic and sedative agents produce their effects with different mechanisms, and therefore from the physiological point of view depth of anesthesia is a vague notion (5). It is more meaningful to talk about components forming the state we usually call anesthesia. These include amnesia, unconsciousness (hypnosis), antinociception, and neuromuscular blockade (paralysis). Different neurophysiological modalities should be used in order to assess these components. In the operating room, the patient is said to be anesthetized while the term sedation is used in the ICU. Some drugs, like propofol, are useful in producing both anesthesia and sedation at different concentrations, while some others are most useful as anesthetics or sedatives. There is evidence that anesthesia and sedation can be produced via different structures in the brainstem. Hypnosis and sedation cause similar changes in the EEG signal (described in the section changes in Neurophysiological Variables During Anesthesia). As all the available depth-of-anesthesia monitors are either fully or partly based on the EEG, the terms depth of hypnosis and depth of sedation are used depending on the corresponding clinical situation and are, in the context of available monitoring devices, roughly equivalent.

The action of anesthetics can be studied at various levels of neuronal function (6). The models underlying these studies can be divided into those operating at the molecular–cellular level and those explaining anesthetic processes at higher levels (generator models). State-of-the-art knowledge on the molecular and neuronal substrates for general anesthesia has recently been reviewed in Ref. 7. The model proposed by Flohr describes the action of anesthetics as disruption of computational processes dependent on the NMDA receptors (8). The activation state of these receptors in cortical neurons determines the complexity of representational structures that can be built-up in the brain and thus the level of consciousness. Steyn-Ross et al. performed numerical simulations of a single-macrocolumn model of the cerebral cortex and found that the effects of anesthetics can be modeled by cortical phase transition (9). Their simulations explain several trends in the EEG caused by anesthetic actions and predict the decrease in spectral entropy of the EEG signal with deepening anesthesia. This model has supported the development of the Entropy module, described in the section Recently Developed Methods, Applied in Commercial Anesthesia Monitors. Great cautiousness must be taken, however, in interpretation of the models operating at molecular level, because they include only a small part of the neurophysiological functions known to be involved in consciousness and generation of anesthesia-induced EEG patterns.

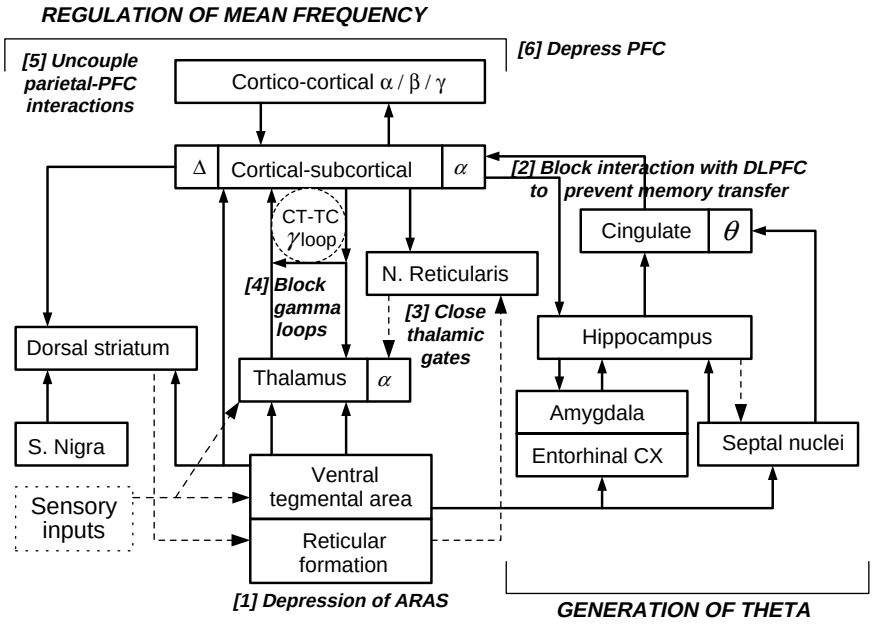


Figure 1. The Anesthetic Cascade model (redrawn with permission from: E. R. John and L. S. Prichep, "The anesthetic cascade: A theory on how anesthesia suppresses consciousness," Anesthesiology, Vol. 102, Fig. 11 (p. 468), 2005).

John et al. developed a higher level model based on a complex neuroanatomical system described in (10). This model, described and thoroughly discussed in volume 10 of Ref. 11, incorporates and explains an extensive bulk of results obtained from EEG, evoked potential and magnetic resonance imaging (MRI) image analysis, as well as laboratory studies. Loss of consciousness is described as the following cascade of events, called the Anesthetic Cascade by the authors, involving various brain regions (Fig. 1) (6): (*1*) depression of the brainstem; (*2*) depression of mesolimbic-dorsolateral prefrontal cortex interactions leading to blockade of memory storage; (*3*) inhibition of the nucleus reticularis of the thalamus, resulting in closure of thalamic gates (seen as increasing θ rhythm in the EEG); (*4*) blocking of thalamocortical reverberations (γ loop) and perception; (*5*) uncoupling of parietal-frontal transactions (coherence in γ frequency band decreases); (*6*) depression of prefrontal cortex to reduce awareness (increase in frontal θ and δ rhythm).

Definitions of the EEG rhythms used in the description of the Anesthetic Cascade are given in Table 1.

The model by John et al. underlies the Patient State Index for depth-of-anesthesia monitoring, described in the section Recently Developed Methods, Applied in Commercial Anesthesia Monitory. Although the Anesthetic Cascade model covers a large set of neurophysiological functions, it does not explain patterns like burst suppression, for example.

Table 1. Definition of the EEG Rhythms[a]

EEG Rhythm	Frequency Range, Hz
Delta (δ)	< 4
Theta (θ)	4–8
Alpha (α)	8–12
Beta (β)	12–25
Gamma (γ)	25–50

[a]Exact frequencies may vary slightly from source to source.

MONITORING ADEQUACY OF ANESTHESIA

Clinical Indicators and Measures of Anesthetic Depth

The verb monitor originally means to check systematically or to keep watch. Thus, monitoring actually does not necessarily involve medical equipment, but refers also to clinical inspection. As clinical indicators of anesthetic depth are often used as a reference for automated depth-of-anesthesia monitoring methods, they are shortly described here.

In the case of inhalation anesthetics, drug concentration can be monitored by measuring the partial pressure of the anesthetic in exhaled air (end-tidal concentration). Due to the variation of the potency of anesthetic agents, a universal unit of minimum alveolar concentration (MAC) has been applied. The value 1 MAC is the partial pressure of an inhaled anesthetic at which 50% of the unparalyzed subjects cease to express protective movement reaction to skin incision. The primary rationale behind the development of the term MAC was the need to compare the potency of different volatile anesthetics, not the effort to monitor the anesthetic state of an individual patient.

For intravenous anesthetics, no such direct measure can be derived and the effect of anesthetics can be estimated using pharmacokinetic models (effect-site concentration). In this case, the accuracy of the estimate depends on the adequacy of the model. If all subjects would react to anesthetics in exactly identical ways these concentration measures would provide a perfect indicator of adequacy of anesthesia. However, there is an intersubject variability in the effect of anesthetics, and therefore other indicators are needed.

Clinical indicators of the adequacy of surgical anesthesia can be divided into those measuring hypnosis and those measuring nociceptive–antinociceptive balance. The indicators measuring hypnosis include pupillary light reflex, tested by allocating a flashlight to one eye and observing both pupils for constriction; corneal reflex,

Table 2. Ramsay Score for Assessment of Level of Sedation[a]

Score	Clinical Status
1	Patient anxious and/or agitated
2	Patient cooperative
3	Patient responds to commands only
4	Brisk response
5	Sluggish response
6	No response to loud auditory stimulus

[a]See Ref. 13.

tested by applying a wisp of cotton wool to the cornea or by electrical stimulation using special electrodes (12); Eyelash reflex, tested by brushing the eyelashes with a moving object or by electrical stimulation; loss of counting, tested by letting the subject count slowly as long as they can from the onset of infusion–injection; syringe dropping, tested by letting the subject hold a syringe between their thumb and forefinger as long as they can; loss of obeying verbal commands.

The indicators measuring nociceptive–antinociceptive balance include avoidance reaction to nociception. This is mainly a spinal reflex, however, it correlates well with the concentration of most anesthetics; electrical tetanic stimulation, applied using needle electrodes or adhesive skin electrodes to the upper or lower limb; autonomic nervous system—mediated reactions or motor reactions to laryngoscopy and endotracheal intubation. This is a natural stimulus in many clinical situations in the operating room.

These indicators test the functioning of different neural pathways and their applicability depends on the anesthetic used. For example, ketamine leaves corneal and pupillary light reflexes intact.

For more graded and standardized clinical assessment of sedation and hypnosis, several scoring systems have been developed. Probably the most widely used such systems are the Ramsay score (Table 2) and the Observer's Assessment of Alertness and Sedation Scale (OAAS; Table 3). These scoring systems are developed for use in the ICU as they include scores for agitated states and cover mainly lighter levels of anesthesia. Therefore, they do not necessarily indicate the adequacy of anesthesia for surgical procedures. Also, the assessment obtained using these scoring systems is subjective.

Table 3. OAAS Score for Assessment of Level of Sedation[a]

Score	Clinical Status
5	Responds readily to command spoken in normal tone
4	Lethargic response to command spoken in normal tone
3	Lethargic response to command spoken loudly and repeatedly
2	Appropriate response to loud tone and mildly painful stimulus
1	Appropriate response to loud tone and moderately painful stimulus
0	No response

[a]See Ref. 14.

Changes in Neurophysiological Variables with Deepening Anesthesia

All the commercial monitors of hypnosis employ the EEG signal. Although different anesthetic agents induce specific features and patterns in the EEG, certain common trends in signal properties with deepening anesthesia can be seen. At subanesthetic levels, several agents produce oscillations at beta frequency range, sometimes called beta buzz. This activity is seen dominantly in the frontal brain areas. With increasing anesthetic concentrations, the activity becomes more widespread, decreases in frequency and increases in amplitude. Around concentrations, causing the subjects to stop responding to stimuli (1 MAC for inhalation anesthetics), the EEG activity slows further and high amplitude theta and delta waves occur. With further increasing concentration, the burst-suppression (BS) pattern occurs, finally turning into continuous suppression. The dynamics of this pattern, as well as the waveforms of bursts, varies for different anesthetic agents (Fig. 2). Several anesthetic agents tend to induce epileptiform seizure activity in patients with a prior history of seizures and even in subjects with no previous history of seizures (15,16).

In addition to the EEG signal, AEPs have been used for anesthesia monitoring. The latency of early cortical responses Pa and Nb increases and the amplitude decreases with deepening anesthesia (17). Also, late cortical responses to auditory stimuli, specifically the amplitude and latency of the N100 peak have been found to improve the assessment of the level of consciousness in ICU patients (18). A commercially available brain monitor, the AEP Monitor/2 by Danmeter A/S, combines AEPs with EEG parameters to calculate the cAAI index (see the next section).

In most commercially available monitoring devices, the EEG signal is obtained from the electrodes placed at the forehead. This makes the recording procedure easier in clinical situations. The electrodes tend to pick up frontal EMG, which is an artifact from the point of view of the EEG signal but may be used as a valuable indicator of nociception in light anesthesia (19). The EMG component of the measurement is either explicitly or implicitly incorporated into most of the available monitoring devices (see the section Discussion).

Another neurophysiological variable proposed for anesthesia monitoring is the respiratory sinus arrhythmia (RSA) component of the heart rate (HR) signal (20). Although potentially valuable addition to the assessment of the level of consciousness, this variable has not made its way to anesthesia monitoring devices to date.

Short History of Brain Monitoring in Anesthesia

Since the first measurements of human electroencephalogram, performed by Hans Berger in 1920s, this modality has been applied to studying the effects of various drugs, including anesthetics. The emergence of microprocessors and digital techniques for signal analysis opened new perspectives for anesthesia monitoring.

The first commercial brain monitoring device based on digital signal analysis was the *Cerebral Function Analyzing Monitor* (CFAM1), developed in 1975 by Prior and

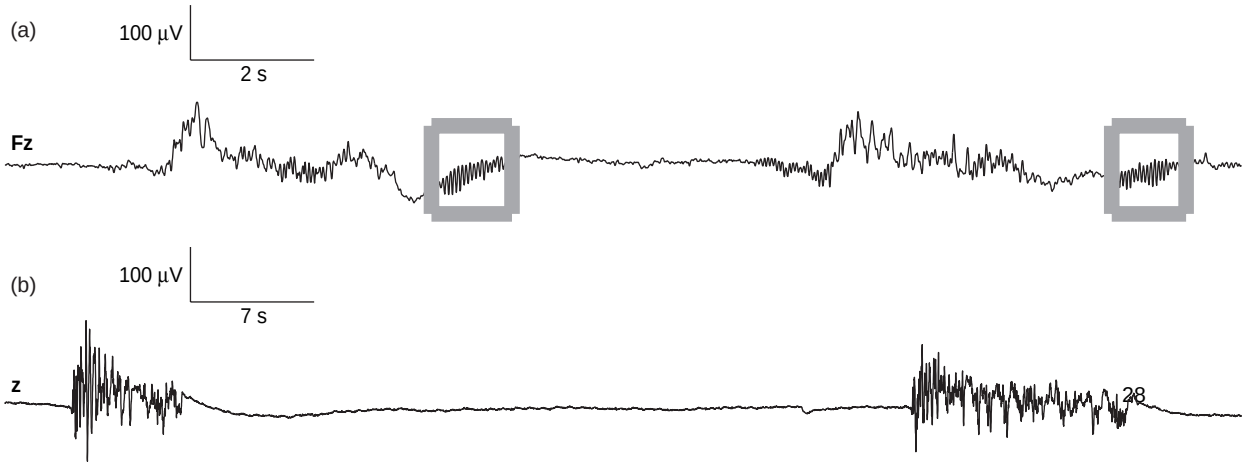


Figure 2. Samples of BS pattern in EEG during deep propofol (a) and sevoflurane (b) anesthesia. Detection of BS suppression and calculation of BS ratio is an important part of all modern depth-of-hypnosis monitors. The pattern varies significantly among anesthetic drugs. In the case of propofol anesthesia, spindles can be observed (marked by boxes in the figure). Note that the scale of the time axes is different for upper and lower curves.

Maynard (21). This device used the Motorola 6808 8-bit microprocessor. The display of the CFAM1 was divided into two sections, one showing the 10th and 90th percentile as well as the mean of the EEG amplitude distribution while the other showing the percentage of weighted (prewhitened) EEG activity per herz in the beta, alpha, theta, and delta frequency bands (Fig. 3). In addition, muscle activity, EEG suppression ratio, and electrode impedance were displayed. An important feature of the CFAM1 was the possibility of monitoring averaged evoked potentials. Since the introduction of CFAM1, the CFAM family of brain monitors has been continuously developed with the recently introduced CFAM4 being the latest member of this product family. Comprehensive list of publications referring to the CFAM family can be found at www.cfams.-com/references/a4a.htm.

In 1982 Datex-Ohmeda (Helsinki, Finland) introduced its first EEG monitor for anesthesia, *the Anesthesia Brain Monitor* (ABM). Like in most of the later monitoring devices, the location of the EEG electrodes in the ABM monitor was on the forehead. The monitor displayed the root-mean squared (rms) value of the EMG and the RMS, as well as the zero-crossing frequency of the EEG signal. The EMG and EEG signals were obtained from the same electrodes–bandpass filter of 65–300 Hz was applied to obtain the EMG while frequencies 1,5–25 Hz were used to obtain the EEG. The ABM monitor is described in (22).

At the beginning of 1990s Thomsen et al. took a different approach to anesthesia monitoring in their *Advanced Depth of Anesthesia Monitor* (ADAM) (23). They divided the signal into consecutive 2 s segments, applied a prewhitening filter, and derived 11 parameters: the rms value and 10 correlation coefficients from each segment. Either the values of the first 10 autocorrelation lags or the coefficients of the 10th-order autoregressive model were suggested as features. To create a set of reference classes, an unsupervised repetitive hierarchical cluster analysis was applied to the data bank of preannotated recordings of halothane and isoflurane anesthesia. Six clusters were defined, corresponding to anesthetic levels from drowsiness to very deep anesthesia. The classification was adjusted according to the anesthetic agent used. Burst-suppression was detected separately and the suppression ratio in 2 s segments was incorporated into classification. Anesthetic depth was displayed as the class probability histogram: A plot where each line represented the clusters

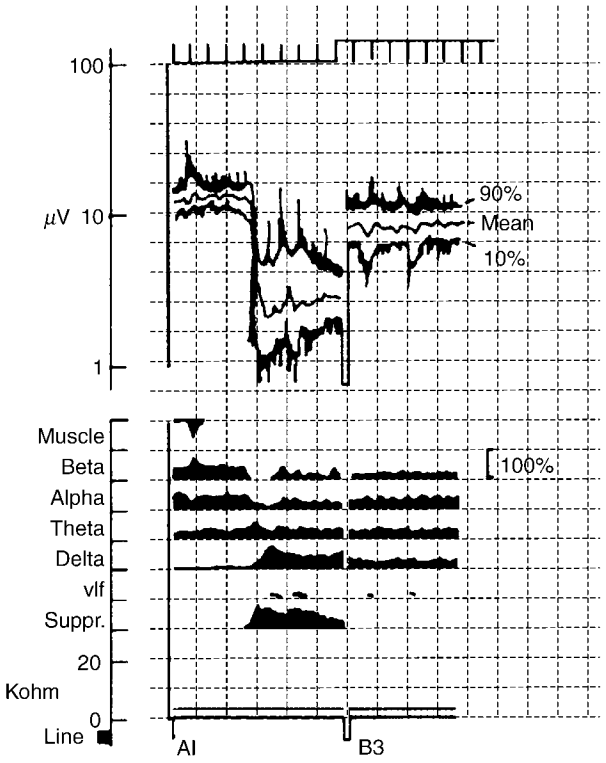


Figure 3. Layout of the screen of the CFAM1 monitor (with permission from D. Maynard).

obtained for 10 s period of the recording. The clusters were color coded. In spite of its advanced approach, ADAM was never implemented in a commercial anesthesia monitoring device.

Recently Developed Methods, Applied in Commercial Anesthesia Monitors

The *Bispectral Index Score* (BIS), developed by Aspect Medical Systems Inc. in 1997, marked a breakthrough in anesthesia monitoring. The output of the BIS monitor is a single number between 0 and 100 achieved by combining in a nonlinear fashion from the following parameters (24): relative beta ratio calculated in spectral domain as $\log\left(\frac{P_{30-47}}{P_{11-20}}\right)$, where P_{30-47} and P_{11-20} denote signal power in frequency ranges 30–47 and 11–20 Hz, respectively; SynchFastSlow measure calculated in bispectral domain as $\log\left(\frac{B_{0.5-47.0}}{P_{40.0-47.0}}\right)$, where $B_{0.5-47.0}$ and $B_{40.0-47.0}$ denote the sum of magnitudes of the bispectrum values in the corresponding frequency ranges; BS ratio. Bispectrum (the third-order spectrum), is defined as the two-dimensional (2D) Fourier transform (FT) of the third-order cumulant sequence $c_3(k_1, k_2)$ of the signal:

$$B(\omega_1, \omega_2) \overset{\text{FT}}{\leftrightarrow} c_3(k_1, k_2) \tag{1}$$

If the direct current (dc) component of the signal has been removed (as is usually the case), $c_3(k_1, k_2)$ equals to the third order moment sequence $m_3(k_1, k_2)$, defined as:

$$m_3(k_1, k_2) = \varepsilon\{s(n)s(n+k_1)s(n+k_2)\} \tag{2}$$

where $\varepsilon\{\cdot\}$ denotes expected value. Overview on the estimation of higher order spectra can be found in (25).

The weighting of the three parameters forming the BIS depends on signal properties and is not disclosed. In light anesthesia, relative beta ratio is dominating while SynchFastSlow measure becomes more important with deepening anesthesia. The function combining the parameters was developed empirically, based on thousands of EEG records. An important part of BIS is its careful artifact rejection scheme, dealing with heartbeat artifacts, eyeblinks, wandering baseline, muscle artifacts, and so on BIS has become very popular among anesthesiologists; the bulk of literature dealing with the behavior of BIS in various clinical situations, discussing its advantages as well as disadvantages, incorporates more than 1000 papers. Comprehensive bibliography can be found on the web-pages of Aspect Medical Systems Inc.

At the beginning of this decade Physiometrix Inc. brought to market the PSA 4000 depth-of-hypnosis monitor, based on the *Patient State Index* (PSI) (26). The development of the PSI was based on a library of 20,000 cases of EEG records. In addition, a library of surgical cases, a library of artifacts and results from volunteer studies (for calibration), were used. In PSI, the EEG signal is measured from four electrodes: Fp1, Fpz, Cz, and Pz, with the reference at linked ear electrodes. Signal analysis is based on power in standard EEG frequency bands (see Table 1) and incorporates the calculation of the following parameters: absolute power gradient between Fp1 and Cz leads in the gamma frequency band (25–50 Hz); absolute power changes between Fpz and Cz leads in beta (12–25 Hz) and between Fpz and Pz; leads in alpha (8–12 Hz) frequency bands; total spectral power (0.5–50 Hz) at the Fp1 lead; mean frequency of the total spectrum at Fpz lead; absolute power in delta frequency band (0.5–4 Hz) at Cz; relative power at Pz lead in slow delta frequency band;

The calculated parameters go through a mathematical transformation that guarantees their Gaussian distribution in order to be rescaled into the Z-score (Fig. 4). The Z-score sets the calculated parameters into relation with the parameter values obtained for reference population giving the percentage of the reference population that lies more standard deviations away from the mean than the calculated parameter (6). The Z-scored parameters are fed into discriminant analysis with adaptive discriminant functions. EEG suppression is detected separately: The suppression ratio is included in the discriminant analysis. The discriminant analysis yields the PSI index: a scalar between 0 and 100 with higher level of hypnosis corresponding to lower PSI value.

The *Narcotrend* anesthesia monitoring system was developed by a German group and first introduced in 2000 (27,28). This system has its roots in sleep analysis: A five-stage sleep scoring system was further developed into a system of 6 stages and 14 substages for level-of-hypnosis monitoring. These stages are mapped to a scale of 0–100 in the Narcotrend algorithm. The EEG signal is derived from one or two electrodes; the most common electrode location is on the forehead, however, according to the authors other electrode locations are possible. The signal is sampled at 128 Hz and prefiltered using lower and upper cutoff frequencies of 0.5 and 45 Hz, respectively. The principal idea underlying the method is similar to that of the PSI: Several variables calculated from the EEG signal are fed to discriminant analysis with separate detection of BS (Fig. 5). The variables are classified as time- and frequency-domain ones and contain signal power, autoregressive coefficients, relative power in standard EEG frequency bands, median frequency (the frequency dividing the signal spectrum into two parts of equal energy), spectral edge frequency (SEF95, the frequency below which 95% of signal energy is contained), and spectral entropy. The algorithm also contains plausibility check to ensure that the signal segment is actually similar to a typical EEG sample of corresponding stage and to detect patterns in the EEG signal untypical for general anesthesia (e.g., epileptic activity). The detailed algorithm of the Narcotrend index is proprietary.

Another EEG-based depth-of-anesthesia monitoring device is the recently introduced *M-Entropy* module for the Datex-Ohmeda S/5 anesthesia monitor. As the name indicates, the method is based on the idea that the entropy of the EEG signal decreases with deepening anesthesia. Signal entropy can be defined and calculated in many different ways (see also the next section) of which spectral entropy is employed in the M-Entropy module. Spectral entropy in the frequency range f_1–f_2 is expressed as

$$S(f_1, f_2) = \sum_{f_i=f_1}^{f_2} P_n(f_i) \log \frac{1}{P_n(f_i)} \tag{3}$$

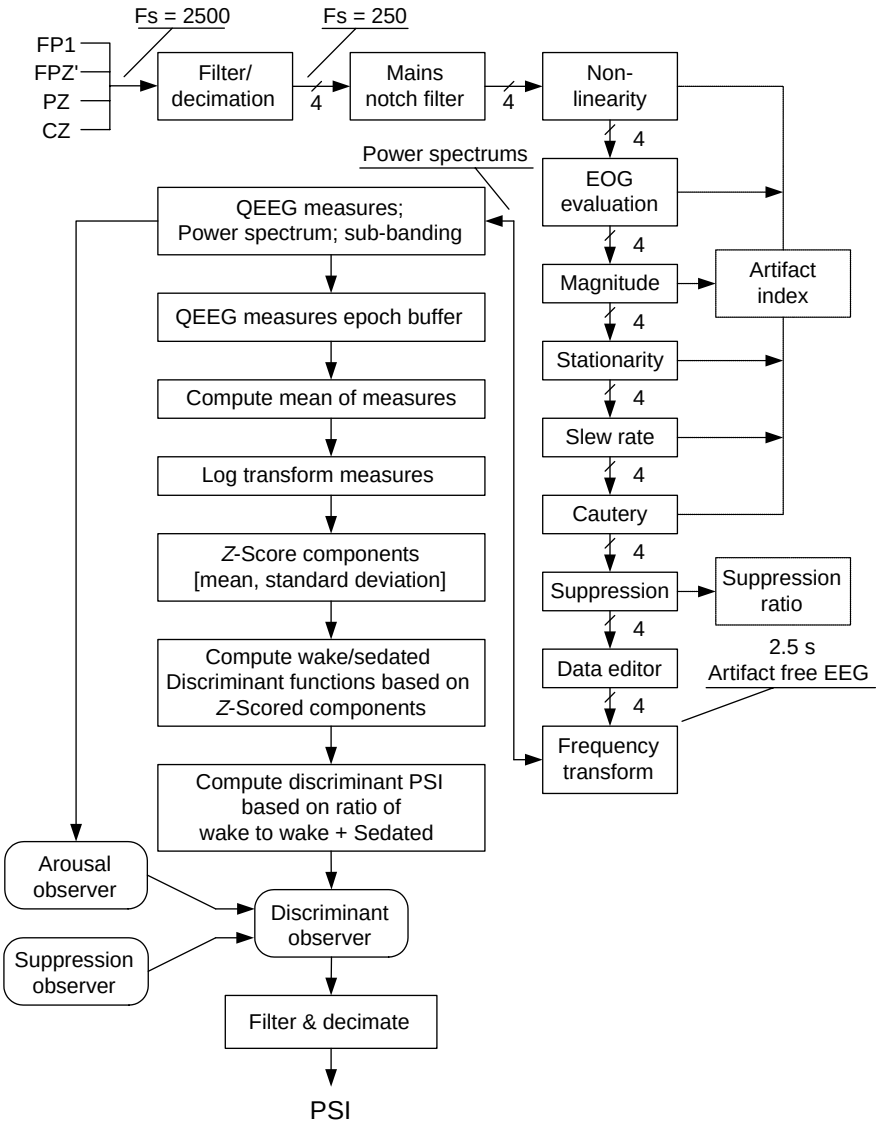


Figure 4. Schematic of the calculation of the PSI index. (Redrawn with permission from D. R. Drover et al., "Patient State Index: Titration of delivery and recovery from propofol, alfentanil, and nitrous oxide anesthesia," Anesthesiology, Vol. 97, Fig. 3 (p. 88), 2002).

where $P_n(f_i)$ is the normalized power spectrum of the signal. $S(f_1, f_2)$ is again normalized by $\log N(f_1, f_2)$, where $N(f_1, f_2)$ is the number of frequency components in the range f_1–f_2, to give a value between 0 and 1. In the original version of the device, the analysis was performed on a single EEG channel measured from the forehead. In this derivation, muscle activity dominates over the EEG at frequencies higher than ~30 Hz. The algorithm of the M-Entropy module, like that of the early ABM-monitor by Datex-Ohmeda, employs these high frequency components to detect the early response of the patient to nociceptive stimuli. This is done by calculating spectral entropy over two frequency ranges: 0.8–32 Hz (called state entropy) and 0.8–47 Hz (called response entropy). The difference between these two entropies indicates the contribution of the EMG component to the response entropy. As in the other described monitors, BS is detected separately. The details of the algorithm (variable window length, obtaining the output value in the case of BS, etc.) are described in (29).

The Danmeter AEP Monitor/2 (further development of the A-Line monitor) employs the composite AAI Index, combining the middle latency auditory evoked potentials in 20–80 ms latency range, calculated from the 25–65 Hz bandpass filtered signal, with spontaneous EEG. The purpose of combining the two modalities is to get a better response to the lightening of hypnosis due to, for example surgical stimuli (achieved by usind AEPs) while retaining sensitivity during deep anesthesia (achieved by using the EEG). The schematic of the cAAI index calculation is presented in Fig. 6. Using evoked potentials poses a problem in on-line monitoring due to the long delay needed for obtaining the averaged response. This problem has been solved in the cAAI calculation by applying the autoregressive model with exogenous input (the ARX model). The ARX model enables to calculate the response to stimuli based on the average of 18 sweeps using the average of 256 sweeps as a reference. The algorithm is described in detail in (30) and compared with conventional evoked potential averaging techniques in Ref. 31. In addition to AEPs, the cAAI index incorporates logarithmic EEG power ratio $[\log(P_{30-47}/P_{10-20})]$ and the burst suppression ratio. The EMG is extracted and monitored

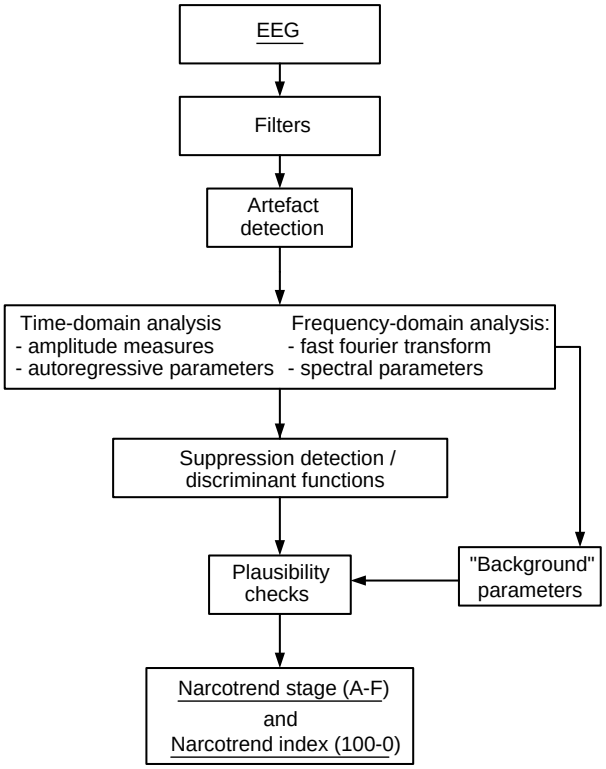


Figure 5. Schematic of the calculation of the Narcotrend index (with permission from B. Schultz).

separately based on the 65–85 Hz bandpass filtered signal.

A somewhat different concept of anesthesia monitoring has been used in the *Cerebral State Monitor* (CSM; Danmeter A/S, Odense, Denmark) and the *SNAP* monitor (Everest Biomedical Instruments Inc). These monitoring devices come in the form of a handheld wireless PDA-type tool, convenient to use in a clinical situation. The CSM displays the Cerebral State Index, calculated based on the 6–42 Hz bandpass filtered EEG, the EMG component calculated from the same signal, but in 75–85 Hz frequency range, as well as the burst suppression ratio. The algorithm of the second version of the SNAP index is described in (32). Two variables, the low frequency variable *LF* (0.1–40 Hz) and the high frequency variable *HF* (80–420 Hz) are derived from a single frontal EEG channel. The HF and LF are scaled to fit into the intervals 0.0–1.0 and 0.0–100, respectively. The SNAP index is expressed as SI = 100—(HF*LF); thus the index can be thought of as the reversed version of HF-modulated LF.

New Parameters Proposed for Monitoring Anesthetic Depth

In spite of the large selection of available methods, new parameters for quantifying depth of hypnosis are being proposed continuously. This is mostly due to the following reasons: the variety of procedures and combinations of drugs in surgical anesthesia is wide. No method performs well in all cases; monitoring in anesthesia is closely related to monitoring brain dysfunction and detection of brain ischemia and hypoxia: important tasks faced in cerebral function monitoring in the ICU and emergency room (see the section Monitoring Outside the Operating Theater). The available depth-of-hypnosis monitors are generally not suitable for these applications; the neurophysiological basis of consciousness is still an unsolved problem: applying modern signal analysis tools to neurophysiological measurements during anesthesia can hopefully offer new insight to the problem.

Several groups have recently published studies on the behavior of various complexity–entropy measures during

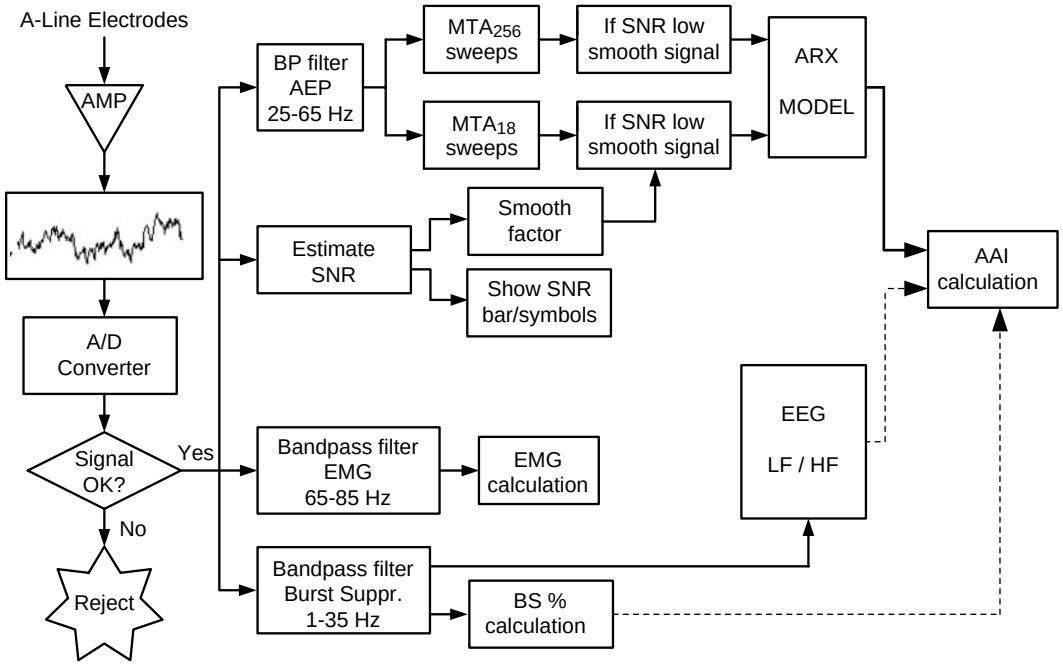


Figure 6. Schematic of the calculation of cAAI index (with permission from E. W. Jensen).

anesthesia and sedation. These measures come from different signal analysis frameworks.

Correlation dimension is a measure for quantifying the behavior of chaotic signals in the phase space (33). The signal s of finite length N is divided into $N-m+1$ time series: $s_m(i)=\{\mathrm{s}(i), \mathrm{s}(i+1), \ldots, \mathrm{s}(i+m-1)\}$, where m is the embedding dimension. After that, for each i the quantity $C_{\mathrm{i}}^{\mathrm{m}}(r)$ is calculated:

$$C_i^m(r) = \frac{\text{number of such } j \text{ that } d[S_m(i), S_m(j)] \le r}{N-m+1} \quad (4)$$

where the distance d between the phase space vectors $s_m(i)$ and $s_m(j)$ is defined as

$$d[S_m(i), S_m(j)] = \max_{k=1,2,\ldots,m} (|s(i+k-1) - s(j+k-1)|) \quad (5)$$

Correlation dimension D can be estimated as

$$D = \frac{d\log(C^m(r))}{d\log(r)} \quad (6)$$

where $C^m(r) = \sum_i C_i^m(r)/N - m + 1$. Although EEG cannot be considered strictly chaotic, except in the case of some abnormal conditions, this measure has, for example, been found to have good correlation with the end-site concentration of sevoflurane (34).

Probably the most intensively studied complexity/entropy measure for the assessment of depth of hypnosis is Approximate entropy (ApEn). In general, entropy measures information-richness, regularity and randomness of a signal. The intuitive idea behind anesthesia monitoring using signal entropy is that with deepening anesthesia EEG becomes more regular and its entropy decreases. Approximate entropy, like correlation dimension, is calculated in the phase space. First, $\phi^m(r)$ is defined based on $C_i^m(r)$ in Eq. 4 as

$$\Phi^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \log C_i^m(r) \quad (7)$$

Approximate entropy is then defined as

$$ApEn(m,r) = \Phi^m(r) - \Phi^{m+1}(r) \quad (8)$$

Approximate entropy has been studied and compared to other methods as an indicator of anesthetic depth in (35–37).

The classical entropy measure, introduced for information theory by Claude Shannon in 1948 (38), the Shannon entropy (ShEn), is calculated as $\mathrm{ShEn} = -\sum_i p_i \log p_i$, where p_i is the probability that signal amplitude obtains the range of values a_i. In practice, ShEn can be estimated based on the histogram of the values of signal samples, and therefore long signal segments are needed to achieve smooth histograms. An important property of Shannon entropy is that signal samples are considered as independent trials of some experiment, taking no notice on the time order of the samples. Signals having equal probability for all possible amplitude values have the highest Shannon entropy. In Ref. 39, it has been found that Shannon entropy of the EEG recorded between frontopolar electrodes increases with increasing concentration of desflurane: A behavior opposite to other entropy measures.

Other measures of the EEG, found to correlate well with depth of hypnosis, include Lempel–Ziv complexity and Higuchi fractal dimension (35,40). Lempel–Ziv complexity is calculated by transforming the signal into symbols and calculating the reoccurrence rate of these symbols (41). Higuchi fractal dimension is calculated as the average rate of increase in the difference of signal amplitude values as the separation between the samples increases in logarithmic scales (42).

These studies demonstrate that although different measures of entropy or complexity quantify different phenomena, many of them may correlate with concentrations of selected anesthetics when electrode positions and signal bandwidth are selected properly.

MONITORING OUTSIDE THE OPERATING THEATER

Development of digital EEG equipment, increase in processing speed and memory capacity, and advancements in telecommunication technology have made cerebral function monitoring feasible in ICU and ER. Brain monitoring in ICU and ER has much in common with monitoring in anesthesia as the changes in the EEG caused by intoxication, metabolic disturbances and brain ischaemia are similar to those induced by general Anesthesia. Also, in the ICU the assessment of depth of sedation is desirable. The advantages offered by EEG monitoring in the ICU are based on the following findings (43): EEG is tightly linked to cerebral metabolism; EEG is sensitive to brain ischemia and hypoxia; EEG detects neuronal dysfunction at a reversible stage; EEG detects neuronal recovery when clinical examination cannot; continuous EEG provides dynamic information; EEG provides useful information about cerebral topography.

However, from the monitoring point of view, the situation in ICU and ER is a lot more complicated compared to that of OR. The patients may need various medication having effect on the EEG signal and misleading automated EEG analysis, the clinical situation of the patients is often complex, and the surrounding is hostile for interference-sensitive equipment. In ICU, recordings often need to last for several days and nights without disturbing the normal care of the patient. In ER, the EEG recording equipment needs to be extremely flexible and easy-to-use. In both situations the interpretation of the recordings poses a problem as no experienced EEG readers are usually around. The solution to the last problem is the usage of telecommunication protocols to transfer the data for interpretation.

Although the above described depth-of-anesthesia monitoring methods are sometimes applied to sedation monitoring and even to the detection of brain dysfunction in ICU, their performance in this situation is questionable. It is difficult to differentiate between the effects of hypoxia, ischemia and sedative drugs. The importance of having the underlying raw EEG signal available for review to confirm the significance of any trends and changes suggested by automatic analysis methods, especially in complex situations like ICU, has been repeatedly emphasized (44,45). A comprehensive brain monitor for ICU, especially for neuroscience ICU, should also be able to detect epileptic

patterns in EEG and desirably have the option for synchronous video recording (46). All this suggests that an adequate brain monitor for ICU or ER should be a much more complex device than today's depth-of-anesthesia monitors.

DISCUSSION

Several considerations are appropriate concerning the available commercial depth-of-anesthesia monitors. First, different modalities should be used to assess the different components of anesthesia (see the section Anesthesia as a Process). Selecting EEG and AEPs as the basis for the assessment, the primary component of anesthesia considered would be hypnosis.

But even in this case there still remain other physiologically separate end-points like subcortically controlled reactions to nociceptive input (e.g., autonomic reactions), increased muscle tone, and movement response to surgery. Adding the fact that there are many anesthetic agents of different cell-level actions and that the interplay of hypnotic and antinociceptive medication modulates the anesthetic state (47), we are left with a complex situation that makes the comparison of the available algorithms for anesthesia monitoring a real challenge.

Another difficulty in comparing the results obtained with different methods is posed by the frequency band used for the calculation. All the commercial methods operate at least partly in the frequency domain although BIS applies third-order spectrum and in the Entropy module a nonlinear transformation follows the calculation of the power spectrum. For anesthesia, monitoring frequency domain can roughly be divided into the following physiologically meaningful areas: δ (and partly θ) frequencies, indicative of pre-BS deep anesthesia (~ 0.5–6 Hz); α and β frequencies; the EMG component, overlapping with the EEG and extending to > 100 Hz.

The devices differ in the usage of δ frequencies and in the way the EMG component is incorporated. While most of the methods employ frequency band starting from 0.1 Hz (SNAP)–0.8 Hz (Entropy), the Cerebral State Index and cAAI by Danmeter do not make use of δ rhythms. Several devices like the A-2000 monitor by Aspect Medical Systems Inc., the AEP Monitor/2 and the Cerebral State Monitor by Danmeter as well as the PSA 4000 monitor by Physiometrix Inc. display the EMG power separately from their corresponding depth-of-anesthesia indices. The frequency band the EMG component is obtained from varies from device to device, falling into the range from 65 to 110 Hz. The SNAP index, the Entropy module and the Narcotrend index incorporate the information on EMG activity into their depth-of anesthesia indexes in different ways. In SNAP, the high frequency band used is 80–420 Hz, while the other two monitors use frequencies up to 47 Hz. The various entropy–complexity measures proposed for the assessment of anesthetic depth are sensitive to the prefilter settings as well (40). Thus it can be concluded that while comparing the performance of various algorithms, the following matters should be considered: the properties of the algorithm itself, the frequency band of the EEG signal it employs, and the location of the EEG electrodes.

In the future, it seems to be inevitable that brain monitoring becomes more common in ICU and emergency room. There is a compromise between the simplicity of the presentation of the output and versatility of the method. Monitoring such complex phenomenon as anesthesia by a single number is clearly an oversimplification. On the other hand, a device requiring sophisticated configuration and displaying a lot of parameters difficult to interpret gets easily rejected by clinicians. Connecting the algorithms to physiological models would certainly help the interpretation of the monitor output. Future will show if any of the new approaches such as measures of signal complexity find their way into the commercial devices. Operating in the frequency domain has the advantage of long-term experience in EEG analysis by means of frequency domain methods. Another advantage is the solid signal processing theory of frequency analysis. On the other hand, the theory of nonlinear systems is developing rapidly having made itsway to physiological signal analysis in various applications.

BIBLIOGRAPHY

1. Myles PS, et al. Patient satisfaction after anesthesia and surgery: Results of a prospective survey of 10811 patients. Br J Anaesth 2002;84:6–10.
2. Committee on Quality of Health Care in America IoM. In: Kohn L, Corrigan J, Donaldson M, editors. To Err Is Human: Building a Safer Health System. Washington: National Academy Press 1999;241.
3. Chopra V, Bovill J, Spierdijk J. Accidents, near accidents and complications. Br J Anaesth 1978;50(10):1041Ü6 36.
4. Lagasse RS. Anesthesia safety: Model or myth? A review of the published literature and analysis of current original data. Anesthesiology 2002;97:1609–1617.
5. Kissin I. General anesthetic action: An obsolete notion? Anesthol Analg 1993;76:215–218.
6. John ER, Prichep LS. The anesthetic cascade: A theory on how anesthesia suppresses consciousness. Anesthesiology 2005; 102:447–471.
7. Rudolph U, Antkowiak B. Molecular and neuronal substrates for general anaesthetics. Nature Rev Neurosci 2004;5:709–720.
8. Flohr H, Glade U, Motzko D. The role of the NMDA synapse in general anesthesia. Toxicol Lett 1998;100-101:23–29.
9. Steyn-Ross DA, Steyn-Ross ML, Wilcocks LC, Sleigh JW. Toward a theory of the general-anesthetic-induced 24 phase transition of the cerebral cortex. II. Numerical simulations, spectral entropy, and correlation times. Phys Rev E 2001; 64:011918 (12 p).
10. Hughes JR, John ER. Conventional and quantitative electroencephalography in psychiatry. J Neuropsychiat Clin Neurosci 1999;11:190–208.
11. Volume 10: Conciousness and Cognition 2001.
12. Mourisse J, et al. Electromyographic assessment of blink and corneal reflexes during midazolam administration: Useful methods for assessing depth of anesthesia? Acta Anaesthesiol Scand 2003;47:593–600.
13. Ramsay M, Savege T, Simpson B. Controlled sedation with alphaxolene/alphadalone. Br J Med 1974;2:656–659.
14. Chernik D, et al. Validity and reliability of the observer's assessment of alertness/sedation scale: Study with intravenous midazolam. J Clin Psychopharmacol 1990;10:244–251.

15. Yli-Hankala A, et al. Epileptiform electroencephaogram during mask induction of anesthesia with sevoflurane. Anesthesiology 1999;91:1596–1603.
16. Vakkuri A. Effects of Sevoflurane Anesthesia on EEG Patterns and Hemodynamics. Ph.D. dessertation, University of Helsinki, Finland, 2000.
17. Thornton C, et al. The auditory evoked response as an indicator of awareness. Br J Anaesthology 1989;63:113–115.
18. Yppärilä H. Depth of Sedation in Intensive Care Patients: A Neurophysiological Study. Ph. D. Dessertation, University of Kuopio, Finland, 2004.
19. Paloheimo M. Quantitative surface electromyography (qEMG): Applications in anaesthesiology and critical care, Ph. D. dissertation Acta Anaesthesiology; Scandinavian. Copenhagen. Munksgaard; 1990; Vol. 34, (Suppl. 93),.
20. Wang DY, Pomfrett CJD, Healy TEJ. Respiratory sinus arrhythmia: A new, objective sedation score. Br J Anaesthology 1993;71:354–358.
21. Maynard D. Development of the CFM: The cerebral function analyzing monitor (CFAM). Ann Anaesthology Francaise 1979;20:253–255.
22. Edmonds HL, Paloheimo M. Computerized monitoring of the EMG and EEG during anesthesia. An evaluation of the anesthesia and brain activity monitor (ABM). Int J Clin Monit Comp 1985;1:201–210.
23. Thomsen CE, Rosenfalck A, Norregaard-Christensen K. Assessment of anaesthetic depth by clustering analysis and autoregressive modeling of electroencephalograms. Comput Methods Progr Biomed 1991;34:125–138.
24. Rampil IJ. A primer for EEG signal processing in anesthesia. Anesthesiology 1998;89:980–1002.
25. Nikias CL, Petropulu AP. Higher Order Spectra Analysis. Englewood Cliffs (NJ): PTR Prentice Hall; 1993.
26. Drover DR, et al. Patient State Index: Titration of delivery and recovery from propofol, alfentanil, and nitrous oxide anesthesia. Anesthesiology 2002;97:82–89.
27. Schultz B, Schultz A, Grouven U. Sleeping stage based systems (Narcotrend). In : Bruch HP. et al., editors. New Aspects of High Technology in Medicine 2000. Bologna: Monduzzi Editore; 2000:285–291.
28. Grouven U, Beger RA, Schultz B, Schultz A. Correlation of Narcotrend Index, entropy measures, and spectral parameters with calculated propofol effect-site concentrations during induction of propofol-remifentanil anesthesia. J Clin Monit Comput 2004;18:231–240.
29. Viertiö-Oja H, et al. Description of the Entropy™ algorithm as applied in the Datex–Ohmeda S/5–Entropy Module. Acta Anaesthesiol Scand 2004;48:154–161.
30. Jensen EW, Lindholm P, Henneberg SW. Autoregressive modeling with exogenous input of middle latency auditory-evoked potentials to measure rapid changes in depth of anesthesia. Methods Inf Med 1996;35:256–260.
31. Litvan H, et al. Comparison of conventional averaged and rapid averaged, autoregressive-based extracted auditory evoked potentials for monitoring the hypnotic level during propofol induction. Anesthesiology 2002;97:351–358.
32. Wong CA, Fragen RJ, Fitzgerald PC, McCarthy RJ. The association between propofol-induced loss of consciousness and the SNAP^{TM} index. Anesthology Analg 2005;100:141–148.
33. Grassberger P, Procaccia I. Measuring the strangeness of strange attractors. Phys D 1983;9:189–208.
34. Widman G, et al. Quantifcation of depth of anesthesia by nonlinear time series analysis of brain electrical activity. Phys Rev E 2000;62:4898–4903.
35. Zhang XS, Roy RJ, Jensen EW. EEG Complexity as a measure of depth of anesthesia for patients. IEEE Trans Biomed Eng Dec 2001;48(12):1424–1433.
36. Bruhn J, Röpcke H, Hoeft A. Approximate entropy as an electroencephalographic measure of anestetic drug effect during desflurane anesthesia. Anesthesiology 2000;92:715–726.
37. Bouillon TW, et al. Pharmacodynamic interaction between propofol and remifentanil regarding hypnosis, tolerance of laryngoscopy, bispectral index, and electroencephalographic approximate entropy. Anesthesiology 2004;100:1353–1372.
38. Shannon CE. A mathematical theory of communication. Bell System Tech J 1948;27:379–423.
39. Bruhn J, et al. Shannon entropy applied to the measurement of the electroencephalographic effects of desflurane. Anesthesiology 2001;95:30–35.
40. Anier A, Lipping T, Melto S, Hovilehto S. Higuchi fractal dimension and spectral entropy as measures of depth of sedation in intensive care unit. Proceedings of the 26-th IEEE EMBS Annual International Conference (EMBC'04), San Francisco; Sept. 2004; pp. 526–529.
41. Lempel A, Ziv J. On the complexity of finite sequences. IEEE Trans Infor Theory 1976;IT–22:75–81.
42. Higuchi T. Approach to an irregular time series on the basis of the fractal theory. Phys D 1998;31:277–283.
43. Jordan KG. Continuous EEG monitoring in the neuroscience intensive care unit and emergence department. J Clin Neurophysiol 1999;16:14–39.
44. Scheuer ML, Wilson SB. Data analysis for continuous EEG monitoring in the ICU: seeing the forest and the trees. J Clin Neurophysiol 2004;21:353–378.
45. Jäntti V, Mustola S, Huotari AM, Koskinen M. The importance of looking at the EEG when presenting uivariate variables to describe it. Br J Anaesth 2002;88:739.
46. Hirsch LJ. Continuous EEG monitoring in the intensive care unit: an overview. J Clin Neurophysiol 2004;21:332–340.
47. Kern SE, X Gie, White JL, Egan TD. Opioid-hypnotic synergy. Anesthesiology 2004;100:1373–1381.
48. Schultz B, et al. Der Narcotrend Monitor: Entwicklung und Interpre-tationsalgorithmus. Anaesthesist 2003;52:1143–1148.

See also ANESTHESIA MACHINES; BLOOD PRESSURE MEASUREMENT; ELECTROENCEPHALOGRAPHY; OXYGEN ANALYZERS; SAFETY PROGRAM, HOSPITAL; TEMPERATURE MONITORING.

MONITORING, AMBULATORY.

See AMBULATORY MONITORING.

MONITORING, FETAL.

See FETAL MONITORING.

MONITORING, HEMODYNAMIC

REED M. GARDNER
LDS Hospital and Utah University
Salt Lake City, Utah

INTRODUCTION

The word monitor has a variety of meanings, depending on the context. A monitor can be any device for checking on or regulating the performance of a machine, aircraft, or a patient. A patient monitor is usually thought of as something that watches, warns, or cautions if there is a life-threatening event. A more rigorous definition of patient monitoring is Repeated or continuous observations or

measurements of the patient, his or her physiological status, and the functions of life support equipment for the purpose of guiding management decisions, including when to make therapeutic interventions and assessment of those interventions (1). As a result, a monitor should not only alert physicians and nurses to potentially life-threatening events, but perhaps should also control devices that maintain life. The primary emphasis of this section deals with hemodynamic monitoring of the critically ill patient who is in an intensive care unit (ICU), but many of the principles apply to all hospitalized patients.

Hemodynamic monitoring relates to monitoring of the blood pressure and blood flow in the cardiovascular system. The cardiovascular system consists of the heart, lungs, and blood vessels, and has a most important function in maintaining life in complex animals, such as humans. Oxygen and fuel must be transported from their source to the individual cells that consume them. The resulting waste products of metabolism must then be disposed of. Thus, the heart and blood vessels transport nutrients to the body and remove the waste products. Clearly, if this system does not function properly, the organism could be compromised. As a consequence clinically applicable methods have been developed to assess the function of the cardiovascular system. Hemodynamic monitoring is one part of this complex monitoring strategy. Typical parameters measured when performing hemodynamic monitoring are heart rate and rhythm, measured through analysis of the electrocardiogram (ECG), blood pressure measurements in various locations in the cardiovascular system, and estimates of blood flow usually using cardiac output as a measure.

THEORY

Hemodynamic monitoring permits minute-to-minute surveillance of the cardiovascular system and provides physiologic data to assist in diagnosis as well as to guide therapy (2–5). The cardiovascular system consists of the heart, lungs, and blood vessels that supply blood to the body and return blood from the peripheral tissue.

It is beyond the scope of this section to describe the detailed anatomy of the cardiovascular system. However, to understand the principles of hemodynamic monitoring knowledge of the functional aspects of the cardiovascular system is essential.

HEART

The heart is made up of four chambers: the right atrium and the right ventricle and the left atrium and the left ventricle (see Fig. 1). The right atrium accepts blood from the systemic circulation (head, arms, and legs) via the superior and inferior vena cava. On atrial contraction the tricuspid valve between the right atrium and right ventricle opens and blood flows into the right ventricle. On ventricular contraction the right ventricle pumps blood through the pulmonic valve into the pulmonary artery and to the lungs where oxygen is added and carbon dioxide is removed. Blood flows from the lungs to the pulmonary veins and then into the left atrium. On atrial contraction

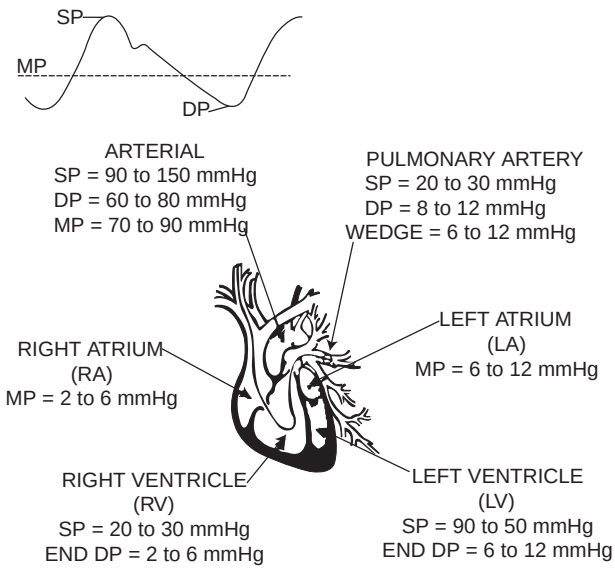


Figure 1. Outline drawing of the heart with its chambers and typical pressures (expressed in mmHg) for each area of the heart. Note the pressures are systolic (SP), diastolic (DP), and mean (MP), as shown on the arterial pressure waveform.

the blood flows into the left ventricle through the mitral valve. On ventricular contraction the left ventricle pumps blood through the aortic valve to the systemic circulation (aorta and the systemic vasculature).

The basic contractile element of the heart is the sarcomere, which is composed of myofilaments, contractile proteins that interdigitate and slide along one another during contraction. Shortening of the sarcomere is the functional unit of heart contraction. Physiologic and pharmacologic agents can change the contractile characteristics of the sarcomeres. Rate and contractility of the heart are controlled by sympathetic and parasympathetic innervation, as well as circulating catecholamines.

Control of Heart Performance

Mechanisms regulating cardiac (heart) output involve not only factors controlling performance of the heart as a pump, but also factors affecting the systemic vascular system and its resistance. Typically, the heart can increase its output to a level of almost five times its resting value. There are two methods by which the heart regulates its cardiac output in response to stress, injury, or disease: by changing heart rate and stroke volume.

Heart Rate Control

Heart rate can be changed rapidly and is thus one of the most effective ways for the heart to change its cardiac output. For a healthy person, an increase in heart rate can more than double the cardiac output when the heart rate increases to near 180 beats $\cdot$ min^{-1}. However, if a patient with heart disease increases their heart rate to $>$120 beats $\cdot$ min^{-1} they may have deleterious responses because of the increased demand for oxygen by the heart muscle. Blood flow in the heart muscle occurs primarily

during diastole (the relaxation phase of heart contraction). Increasing heart rate decreases the time for cardiac circulation during diastole. In normal subjects, decreasing the heart rate to ~ 50 beats $\cdot$ min^{-1} may not decrease cardiac output because there is increased diastolic filling time that increases stroke volume.

Stroke Volume Changes

The stroke volume of an intact ventricle is influenced by (*1*) ventricular end-diastolic volume (called preload), (*2*) ventricular afterload, and (*3*) contractility.

Preload. Preload is the term used to define the end-diastolic stress in the wall of the ventricle. For example, zero preload would result in the ventricle ejecting no blood. However, with increased preload, ventricular ejection generally increases linearly until the capacity of the pump (heart) is exceeded. Since the end-diastolic volume so profoundly influences the myocardial fiber length it has a great influence on the myocardial performance. The Frank–Starling law describes this principle and is illustrated graphically in Fig. 2. The most accessible measure of right ventricular preload is the right atrial pressure. Left atrial pressure is used to estimate left ventricular preload. Since the left ventricle does most of the work of the heart, it is usually the first part of the heart muscle to fail. Consequently, the measurement or estimation of the left atrial pressure is important in assessing a patient's hemodynamic status.

Afterload. Afterload is a measure of the impedance (resistance) against which the right or left ventricles must eject blood. Resistance (R) is calculated by measuring blood flow and pressure and then using Ohm's law {Eq. 1}.

$$R = \frac{\text{mean blood pressure}}{\text{cardiac output}} \tag{1}$$

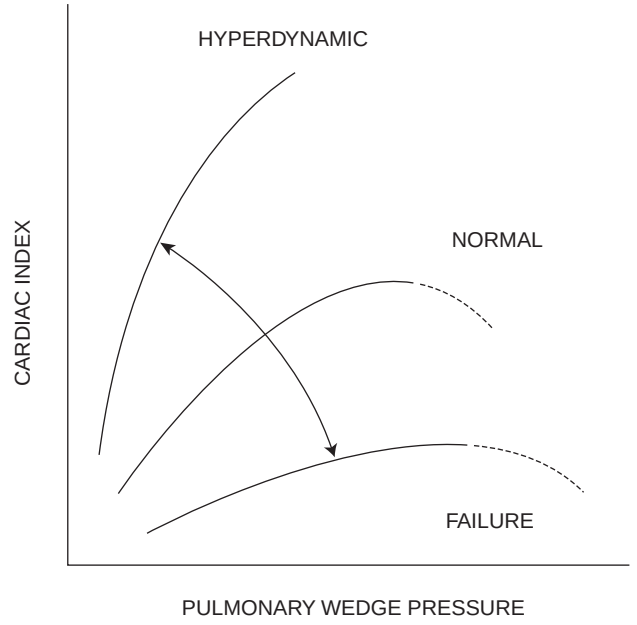


Figure 2. Frank–Starling curve of the heart showing the ventricular performance (cardiac index) plotted against the end-diastolic volume typically estimated by using pulmonary artery wedge pressure. Note to the right of these curves, there is a pulmonary wedge pressure above which the heart in ineffective in producing increased flow.

Systemic Circulation

Blood flow to the periphery of the body is controlled by local autoregulation and by the autonomic nervous system. Local autoregulation of blood flow helps tissue meet its oxygen requirements. For example, with decreased blood flow, metabolic byproducts increase, causing local vasodilatation that tends to increase blood flow. There are baroreceptors, similar to blood pressure transducers, located in the aortic arch and the carotid sinus which sense blood pressure. Via the baroreceptor reflex mechanism, the body regulates the blood pressure. In addition, chemoreceptors in the carotid sinus and other locations regulate respiration by responding to changes in CO_2 and O_2.

Pulmonary Circulation

The pulmonary arterial vessels differ markedly from systemic arterial vessels; they have thinner walls, less muscle, and have a resistance to blood flow about one-sixth that of the systemic circulation.

Contractility. Contractility is a measure of how a healthy heart performs. A healthy heart pumps vigorously and nearly empties its ventricles with each beat and is said to have excellent contractility. On the other hand, a compromised heart may not be able to empty effectively.

HEMODYNAMIC MONITORING

Bedside hemodynamic monitoring makes use of data gathering procedures that were formerly only done in diagnostic cardiac catheterization laboratories. Understanding the relationship between the pressure and blood flow in the cardiovascular system is the primary reason for performing hemodynamic monitoring. The cardiovascular system responds to many and varied stimuli and can be affected by physical conditioning, drugs, disease, blood loss, injury, and myocardial insult such as a heart attack. Because of the complexity of factors controlling the body, it is necessary to make hemodynamic measurements on the system to understand disease processes and provide optimum therapy to the patient.

Electrocardiogram

Electrocardiogram (ECG) monitoring is used to determine heart rate and detect arrhythmias, pacemaker function, and myocardial ischemia (lack of blood flow to the heart muscle). To permit optimum ECG monitoring the signal quality must be excellent (6). Since the ECG electrical signal from the heart is only 0.5–2.0 mV at the skin's surface, it is best measured by properly preparing the skin and optimally placing the electrodes. Skin can be properly prepared by removing oil from the surface and abrading the skin to remove the dry dead outer layer (stratum

granulosum). In 90% of patients, proper skin preparation reduces electrode resistance from as high as 200 to as low as 10 kΩ. Good electrode placement allows the electrodes to receive the maximum ECG signal with minimum noise. By placing the electrodes over bony prominence, such as the sternum or clavicles, muscle artifact (EMG) can be reduced. Motion artifact caused by movement of electrodes can be minimized not only by proper skin preparation, but also by taping a strain-relieving loop in the lead wires to prevent movement artifact. Shielded wire on the ECG leads helps minimizes pickup of alternating current (ac) electrical fields from 60 Hz power sources, electrosurgical units, and other sources like radio transmitters. The two leads that connect the patient form a loop through which magnetic fields pass and can induce unwanted voltages. Pickup from magnetic fields can be minimized by decreasing the loop area, by keeping the lead wires close together (usually twisted pairs), and by avoiding draping the lead wires over motors, lights, or other electrically powered instruments.

Electrocardiogram Arrhythmia Monitoring

Early in the development of monitoring techniques, the application of computer technology to detect patterns of the electrocardiogram caught the attention of those who sought to improve care of the critically ill. The computer appeared to be a logical candidate for relieving the nursing and medical staff of the tedious chore of continuously visually monitoring a multichannel oscilloscope.

Arrhythmia monitoring is one of the most sophisticated of the bedside monitor's tasks. People-based arrhythmia monitoring is expensive and unreliable, and those who do it find the task to be tedious and stressful. Today virtually every bedside monitor has rhythm monitoring built in. These monitors use computers and a variety of algorithms to detect and classify ECG rhythm abnormalities. Classifying these rhythm abnormalities is important to hemodynamic monitoring since irregular rhythms can cause dramatic inefficiencies in how the heart works as a pump. For example, Fig. 3 shows three strip recordings of the ECG and the corresponding pressure waveform from three different arrhythmias (ventricular tachycardia, couplet, which is two beats of abnormal electrical origin, and bigeminy where every other beat is from abnormal electrical origin). Note that several of the abnormal beats are hardly effective at creating any change in the arterial pressure. Those same beats deliver small stroke volumes to the patient's systemic circulation. As a consequence, one cannot assume that the cardiac output remains constant or increases just because the heart rate increases.

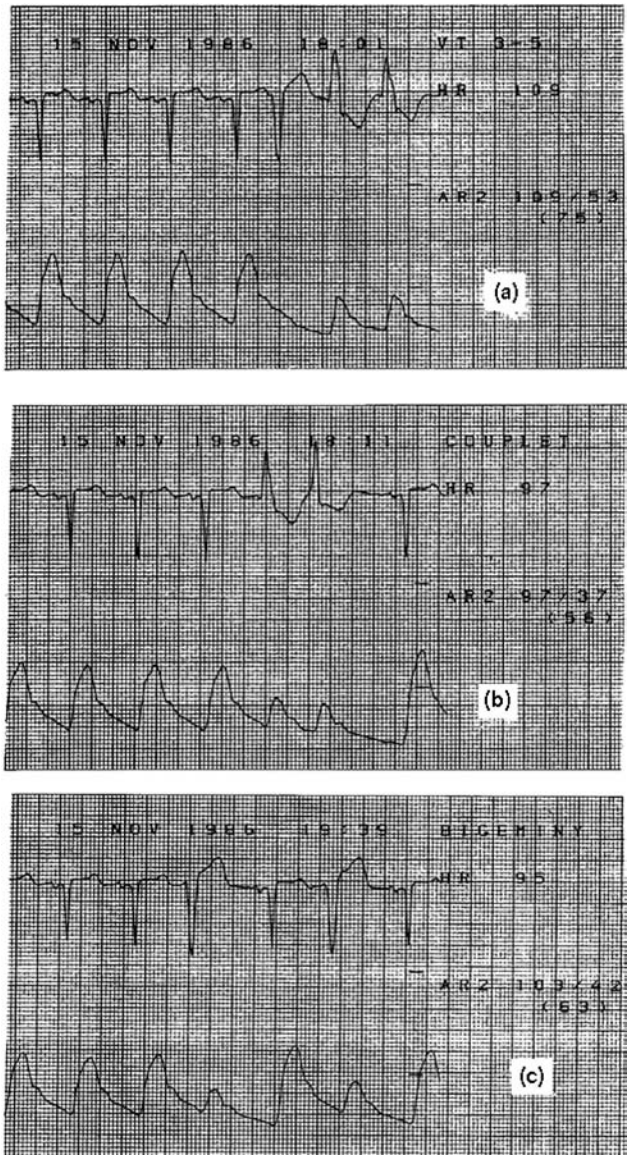


Figure 3. Electrocardiogram and arterial pressure waveforms with three different abnormal rhythms. (a) Ventricular tachycardia (VT), which occurs during the last two beats of the strip. (b) Couplets where two successive beats have an abnormal electrical origin. (c) Bigeminy where every other beat is from an abnormal electrical origin. Pressures are expressed in millimeters of mercury. For example, the patient in (a) has a systolic arterial pressure of 109 mmHg, a diastolic pressure of 53 mmHg, and a mean arterial pressure of 75 mmHg.

MEASUREMENTS

Blood Pressure Monitoring

Arterial blood pressure can be measured by both direct and indirect methods. However, central venous pressure (CVP), pulmonary artery (PA), and pulmonary capillary wedge pressure (PCWP) used to estimate left atrial pressure, at present, can only be measured by direct invasive methods.

Arterial Blood Pressure: Indirect Measurement; Using a Cuff. Recently, the American Heart Association has updated its recommendations on accurate "indirect" measures of blood pressure (7). The update reports that the auscultatory technique with trained observer and mercury manometer continues to be the method of choice for measurement of a patient's blood pressure in a physician's office. The report also suggests that hybrid devices that use electronic transducers rather than mercury have promise. The report indicates that oscillometric devices can also be used, but only after careful validation.

Unfortunately, the indirect measurement of arterial pressure has serious limitations for patients in shock usually signaled by low blood pressure. Also, since virtually all reliable indirect pressure measurement techniques require cuff inflation, such measurements can only be made intermittently.

Direct Blood Pressure Measurements. The direct measurement of blood pressure allows for continuous and accurate assessment of blood pressures. Direct and continuous pressure monitoring allows detection of dangerous hemodynamic events and provides the information necessary to initiate and regulate patient therapy to prevent catastrophic events. However, monitoring of pressures provides valuable information only when it is obtained in a technically satisfactory manner.

To accomplish direct blood pressure measurements, it is necessary to insert a catheter directly into the cardiovascular system (8). This invasive technique has risks that must be weighted against the benefits that can be obtained. These risks include, infection, blood loss, insertion site damage and other factors (9,10). For many patients who are in shock or who have cardiac disease, the benefits far outweigh the risks. Formal methods for assessing these risks have been published by the Coalition for Critical Care Excellence (11).

Blood pressure can be measured on both the pulmonary (right heart) and systemic (left heart) sides of the circulatory system. Measurements of both pulmonary and systemic parameters yield different and important cardiovascular status. The CVP reflects the patient's circulating blood volume or venous tone, and right atrial and ventricular pressures (right ventricular preload). To measure the right atrial pressure accurately a catheter must be placed in a major vein within the chest or directly in the right atrium. The CVP values fluctuate about atmospheric pressure. The level of the right heart is usually taken as the zero reference point from which all other blood pressures are measured. The CVP gives an indication of only the function of the right heart, and not left heart's performance.

To measure the left atrial pressure, it is necessary to place a catheter tip through the atrial septum from the right atrium (usually done only with fluoroscopic control in the cardiac catheterization laboratory) or estimating it by placing a pulmonary artery (Swan–Ganz) catheter in the wedged position by inflating its balloon near the catheter tip.

EQUIPMENT

Components of Direct Pressure Monitoring Systems

The components of a direct blood pressure monitoring system for critically ill patients are shown in Fig. 4 (6,8). The components numbered 1–7 in the figure are known as

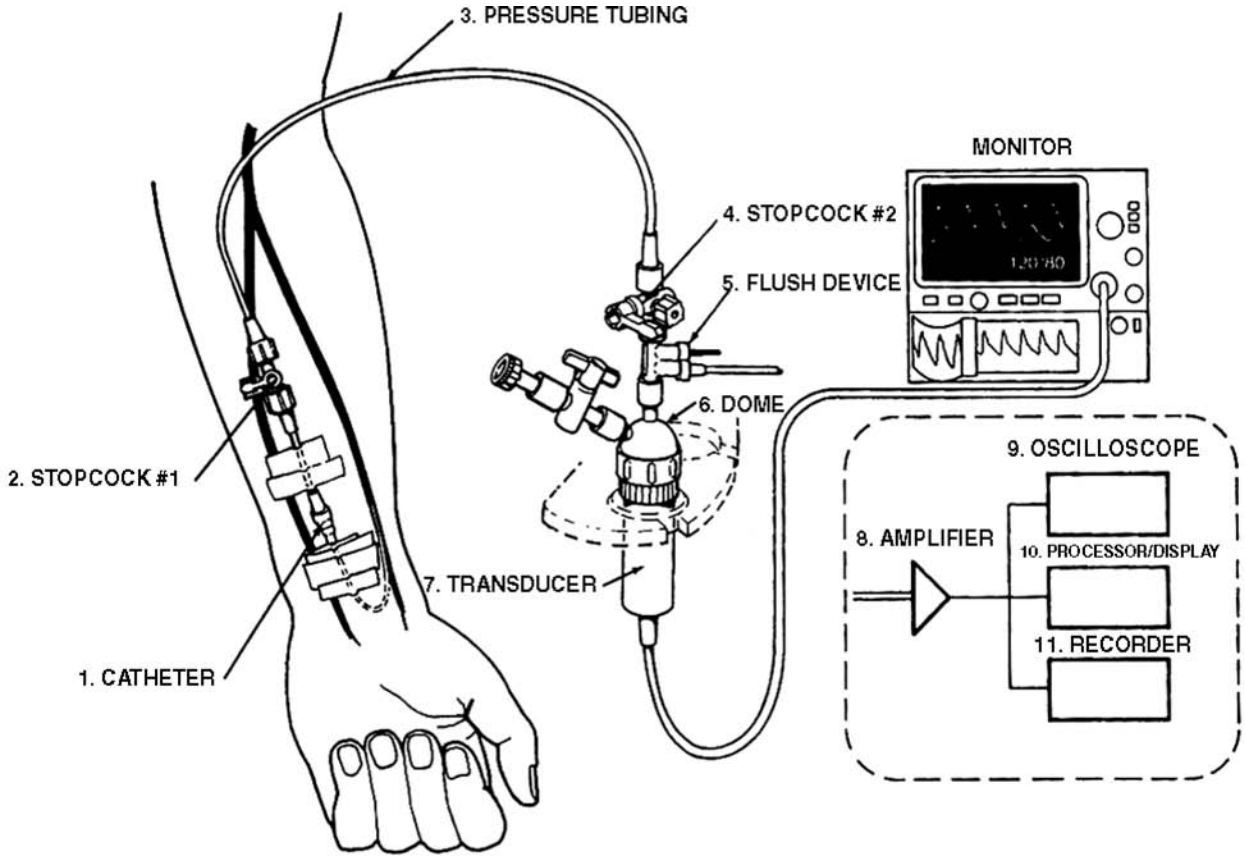


Figure 4. The 10 components used to monitor direct blood pressure. The monitoring components are nearly independent of whether the catheter is in an artery (radial, brachial, or femoral) or in the pulmonary artery. Size of transducer and plumbing components are enlarged for illustration purposes. [Reproduced from Ref. 6, with permission.]

the "plumbing" system and must always be sterile because the fluid contained therein comes in direct contact with the patient's blood. Today virtually all of these components are disposable or single-use items to minimize patient infection. Components 8–11 in Fig. 4 are used for processing and displaying pressure waveforms and derived hemodynamic parameters.

1. Catheter. Arterial and pulmonary artery catheters provide access to the patient's blood vessels to (a) monitor intravascular pressure and (b) provide a site for samples for blood to allow determination of blood gas and other laboratory testing parameters. These catheters are typically placed by the percutaneous method, either by the Seldinger "over-the-needle" technique or by introducing the catheter through a needle (8).
2. Sampling stopcock. Stopcock 1 is used as a site for withdrawing blood for analysis. When filling the catheter-tubing-transducer system with fluid, precautions must be taken to be sure all central switching cavities of the stopcock are filled and that entrapped air bubbles are removed. Because stopcocks are especially vulnerable sources of patient contamination, these devices must be handled with extreme care; ports not in active use should be covered with sterile caps and medical personnel should never touch open ports of the stopcocks.
3. Pressure tubing. The catheter and stopcock are normally attached to a continuous flush device and transducer by noncompliant pressure tubing. To optimize the dynamic response of the catheter-tubing-transducer system, long lengths of tubing must be avoided.
4. Stopcock 2. This stopcock is usually put in place to allow disconnection of the flush device and transducer from the patient when the patient is moved or when initially filling the system with fluid.
5. Continuous flush device. This device is used not only when initially filling the pressure monitoring system, but also to help prevent blood from clotting in the catheter. These devices provide a "continuously flush" of fluid at a rate of from 1 to 3 $mL \cdot h^{-1}$.

6,7. Transducer dome and Pressure transducer. Today virtually all transducers used for monitoring are highly reliable, standardized, disposable devices (12,13).

8. Amplifier system. The output voltage required to drive an oscilloscope or strip-chart recorder is provided by an amplifier system inserted between the transducer and display. Pressure transducer excitation is provided either from a direct current (dc) or ac source at a voltage of 4–8 V revolutions per second (rms). Most amplifier systems include low pass filters that filter out unwanted high frequency signals. Pressure amplifier frequency response should be flat from 0 to 50 Hz to avoid pressure waveform distortion.
9. Oscilloscope. Pressure waveforms are best visualized on a calibrated oscilloscope or other similar display panel.
10. Digital processing and display. Digital displays provide a simple method for presenting quantitative data from the pressure waveform. They are found on most modern pressure monitoring equipment. Systolic, diastolic, and mean pressure are typically derived from the pressure waveforms.
11. Strip-chart recorders. Frequently, strip-chart recorders are used to document dynamic response characteristics, respiratory variations in pulmonary artery pressures, and aberrant rhythms and pressure waveforms.

STATIC CALIBRATION

Zeroing and calibrating the transducer are two important steps in setting up the direct pressure-monitoring system.

Zeroing the Transducer

The accuracy of blood pressure readings depends on establishing an accurate reference point from which all subsequent measurements are made. The patient's midaxillary line (right heart level) is the reference point most commonly used (14). The zeroing process is used to compensate for offset caused by hydrostatic pressure differences, offset in the pressure transducer, amplifier, oscilloscope, recorder, and digital displays. Zeroing is accomplished by opening an appropriate stopcock to the atmosphere and aligning the resulting fluid-air interface with the midaxillary reference point.

Once the system is zeroed the stopcock can be switched to allow the patient's waveform to be displayed. Pulmonary artery and pulmonary artery wedge pressure are especially susceptible to improper zeroing and should be measured only after the zero point has been verified.

Sensitivity Calibration

The sensitivity of most pressure transducers is fixed at 5.0 $\mu V \cdot V^{-1}$ of excitation applied per 1 mmHg (0.13 kpa) and calibrated by the manufacturers to within $\pm 1\%$. This degree of accuracy is adequate for clinical purposes. As a consequence standardized transducers need only to be zeroed to obtain accurate pressure measurements (12,13).

CHECKING DYNAMIC RESPONSE

In the critical care setting, where most hemodynamic monitoring is carried out, the catheter-tubing-transducer systems used can usually be characterized as an underdamped second-order dynamic system analogous, for example, to a bouncing tennis ball. A second-order dynamic system can be expressed mathematically by a second-order differential equation with characteristics determined by three mechanical parameters: elasticity, mass, and friction. These same parameters apply to a catheter-tubing-transducer system where the natural frequency (f_n in Hz) and damping coefficient determine the dynamic characteristics for a catheter-tubing-transducer system. For an underdamped second-order system f_n and define the system's

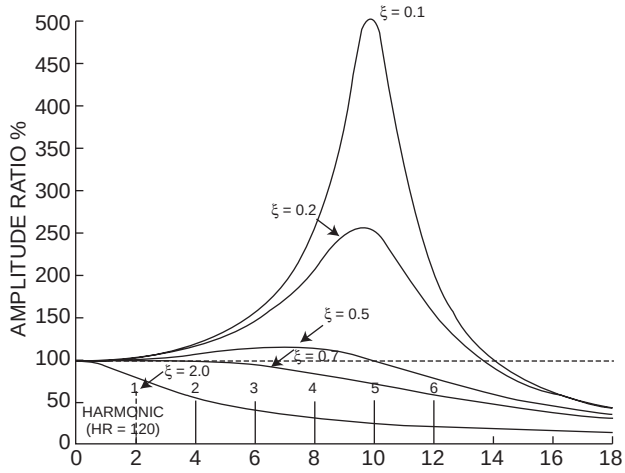


Figure 5. Family of frequency versus amplitude ratio plots for five different damping coefficients ζ and natural frequencies f_n of the plot shown is 10 Hz. When $\zeta = 0.1$, the system is very underdamped, and when $\zeta = 2.0$, it is overdamped. The dashed line shows the frequency versus amplitude characteristic that would occur if the system had a flat frequency response. Along the frequency axis are plotted the harmonics of the pressure wave if the heart rate were 120 beats $\cdot$ min^{-1} (2 beats $\cdot$ s^{-1}). Note that by the fifth harmonic (10 Hz) if $\zeta = 0.1$ the true signal would be amplified five times. If $\zeta = 2.0$ there would be an attenuation to about one-fourth of the amplitude. In both cases there would be a gross waveform distortion because neither situation reflects a high fidelity system dynamic response. Fidelity of the system can be improved by increasing the f_n or adjusting ζ to be in the range of 0.5–0.7. [Reproduced from Ref. 6, with permission.]

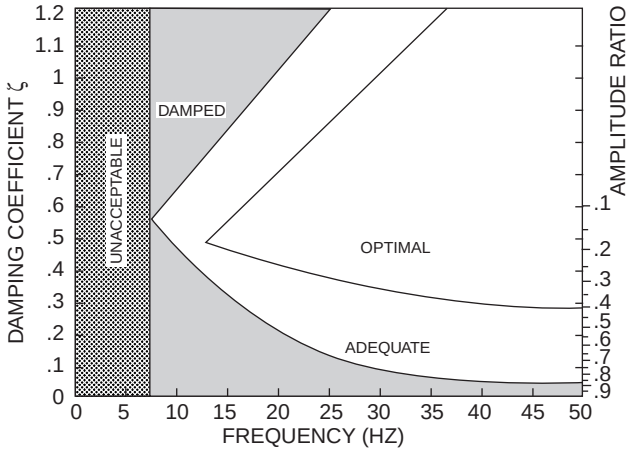


Figure 6. Plot of f_n versus ζ illustrating the five areas into which catheter-tubing-transducer systems fall. Systems in the optimal area will reproduce even the most demanding (fast heart rate and rapid systolic upstroke) arterial or pulmonary artery waveforms without distortion. Systems in the adequate area will reproduce most typical patient waveforms with little or no distortion. All other areas will cause serious and clinically important waveform distortion. Note the scale on the right can be used to estimate ζ from the amplitude ratio determined during fast flush testing (11). See Fig. 8 for an example of waveforms. [Reproduced from Ref. 6, with permission.]

dynamic characteristics. In the clinical setting f_n and can be measured easily and conveniently by using the "fast-flush" method.

Dynamic response characteristics of catheter-tubing-transducer systems have been defined by two interrelated techniques. The first technique specifies the frequency bandwidth and requires that the system frequency response must be flat up to a given frequency so that a specified number of harmonics usually 10 of the original pulse wave can be reproduced without distortion (Fig. 5).

The second technique specifies f_n and The plot of f_n and in Fig. 6 has five areas (6,15). If the characteristics of the catheter-tubing-transducer "plumbing" system fall in the adequate or optimal area of the graph, the pressure waveforms will be adequately reproduced. If the characteristics fall into one of the remaining three areas, there will be pressure waveform distortion. Most catheter-tubing-transducer systems assembled under optimal conditions are underdamped, but a few fall into the unacceptable areas of the chart. Methods for optimizing the catheter-tubing-transducer system components have been outlined (15–20). In the clinical setting, there are dramatic differences between each patient setup; therefore it is mandatory to verify the adequacy of each pressure-monitoring system by testing them. The testing can be done easily using the fast-flush technique.

A fast flush is produced by opening the valve of the continuous flush device, for example, by pulling and quickly releasing the pigtail valve on a continuous flush device. The rapid valve closure generates a near square wave pressure signal from which f_n and of the catheter-tubing-transducer system can be measured.

Once the fast-flush test has been executed two or three times, the dynamic response characteristics (f_n and) can quickly and easily be determined. Natural frequency f_n can be estimated by measuring the period of each full oscillation on a strip-chart recorder following a fast flush (Fig. 7a) and calculating the frequency from the period. Damping coefficient can be determined by using the amplitudes of any two successive peak-to-peak values measured after a fast flush. The amplitude ratio is calculated by dividing the measured height of the smaller peak-to-peak value by that of the amplitude of the larger peak-to-peak value (Fig. 7b). The amplitude ration can then be converted to a damping coefficient by using the scale in the right side of Fig. 6.

Once f_n and have been determined, these data can be plotted on the graph of Fig. 6 to ascertain the adequacy of dynamic response. Some bedside monitors and recorders may compromise the clinical user's ability to use the fast-flush technique because the monitors have built-in low-pass filters. These filters should be expanded to at least 50 Hz or be eliminated.

Several factors lead to poor dynamic responses: (*1*) air bubbles in the system usually caused by a poor initial catheter-tubing-transducer system setup, (*2*) pressure tubing that is too long, too compliant, or a diameter that is too small, and (*3*) pressure transducers that are too compliant. The best way to enhance the system's dynamics is to improve f_n.

Invasive pressure monitoring systems have patient risks, such as a source of infection and air embolism. In addition, great care is required by clinical users to optimize

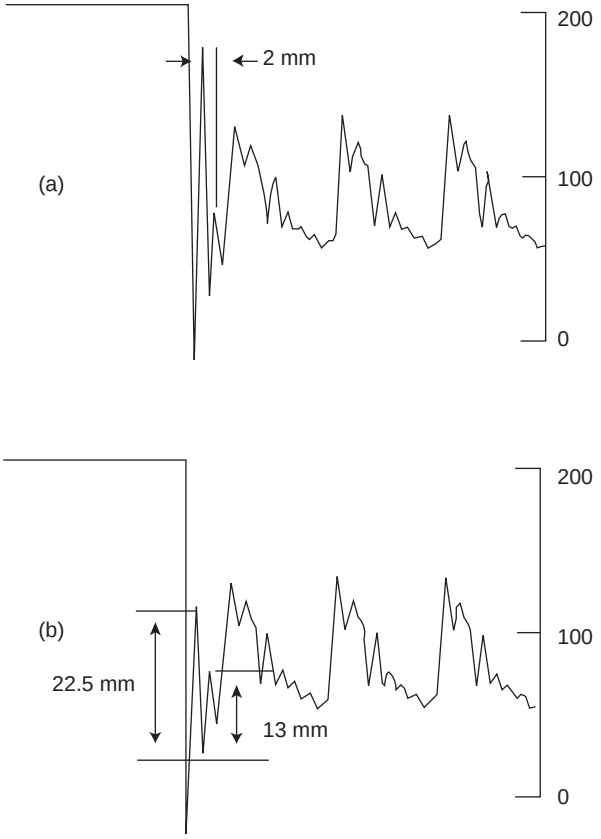


Figure 7. Measuring dynamic response parameters from the fast-flush waveform, (a) The natural frequency f_n can be determined by using a strip-chart recording to measure the period of one full oscillation, as shown. In this example, one full cycle is 2 mm and at a paper speed of 25 mm $\cdot$ s^{-1} this results in $f_n = 12.5$ Hz $=$ 25 mm $\cdot$ s^{-1}/2 mm. (b) Determining the damping coefficient ζ required measuring two successive peak-to-peak values of the resulting oscillations. The amplitude ratio of the two successive peaks is taken, giving a value of $0.58 = 13/22.5$. With use of the amplitude ratio and the scale on the right side of Fig. 6, the damping coefficient $\zeta = 0.17$. Plotting the natural frequency and damping coefficient on Fig. 6 shows that this system is underdamped.

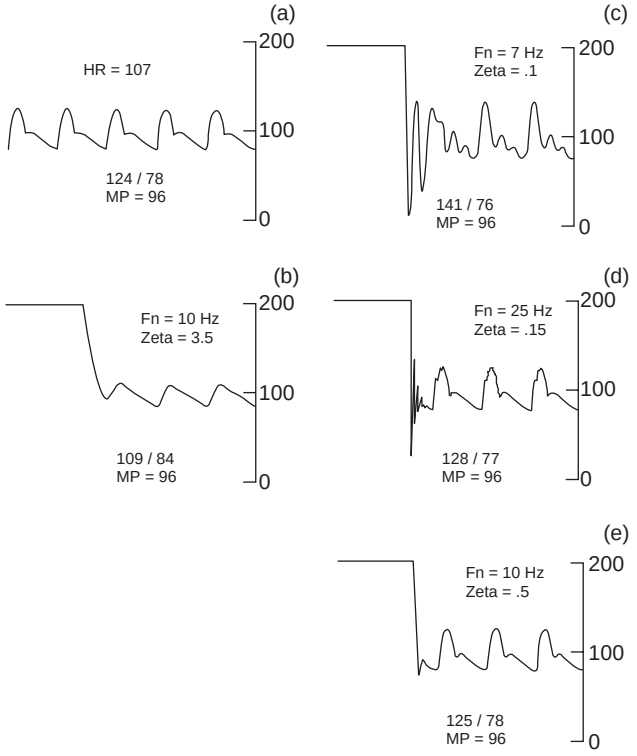


Figure 8. Arterial pressure waveforms were obtained from the same patient. Shown are Systolic/Diastolic and Mean Pressure (*MP*). In panel the (a) Patient's actual arterial pressure waveform as if recorded with a catheter-tipped transducer is shown, (b) shown the same patient's arterial waveform recorded with an overdamped system ($\zeta = 3.5$). Note the fast-flush signal (upper left) returns slowly to the patient waveform. Systolic pressure is underestimated, diastolic pressure is overestimated, and *MP* is unchanged, (c) An underdamped condition ($\zeta = 0.1$) with low $f_n = 7$ Hz. After the fast flush, the pressure signal oscillates rapidly (rings). Systolic pressure is overestimated, diastolic is slightly underestimated, and MP is correct, (d) shows an underdamped condition ($\zeta = 0.15$), but with high $f_n = 25$ Hz. The pressure waveform is slightly distorted and systolic, diastolic, and mean pressures are close to the actual pressures, (e) shown an ideally damped pressure monitoring system ($\zeta = 0.5$). The undershoot after the fast flush is small and the original patient waveform is adequately reproduced. [Reproduced from Ref. 6, with permission.]

dynamic response and proper zeroing to provide accurate and reliable data. Merely looking at pressure waveforms will not provide the information required to determine the adequacy of the system's dynamic response (see Fig. 8). Fast-flush testing to determine these parameters is essential.

SIGNAL AMPLIFICATION, PROCESSING, AND DISPLAY

Once the pressure signal has been transmitted to the transducer, the bedside monitor operates on that signal. Most monitors not only display the heart rate and systolic, diastolic, and mean pressure, but they also display the processed waveform on an oscilloscope and provide an analog output for a recorder or for transmission to a central display.

Placement of the Pulmonary Artery Catheter

The balloon-tipped, flow-directed, pulmonary artery catheter (Swan–Ganz) came into widespread use in 1970 (21). The follow-up development by Ganz of a practical thermal dilution attachment to the pulmonary artery catheter permitted convenient and easy measurement of cardiac output (22). Since these early developments with the Swan–Ganz catheter, the pulmonary artery catheter has been fitted with optical fibers which allow measurement of mixed venous oxygen saturation (23).

The pulmonary artery catheter is inserted into the right side of the circulation using the percutaneous technique typically using entry from either the internal jugular or the subclavian vein. The catheter is floated into the pulmonary artery without use of fluoroscopy, using the hemodynamic pressure waveforms as a guide (Fig. 9).

Accurate Measurement of Pulmonary Artery Pressure. Since it was introduced, the balloon-tipped, flow-directed, pulmonary artery catheter (Swan–Ganz) has been widely used in intensive care units. The ease with which it is

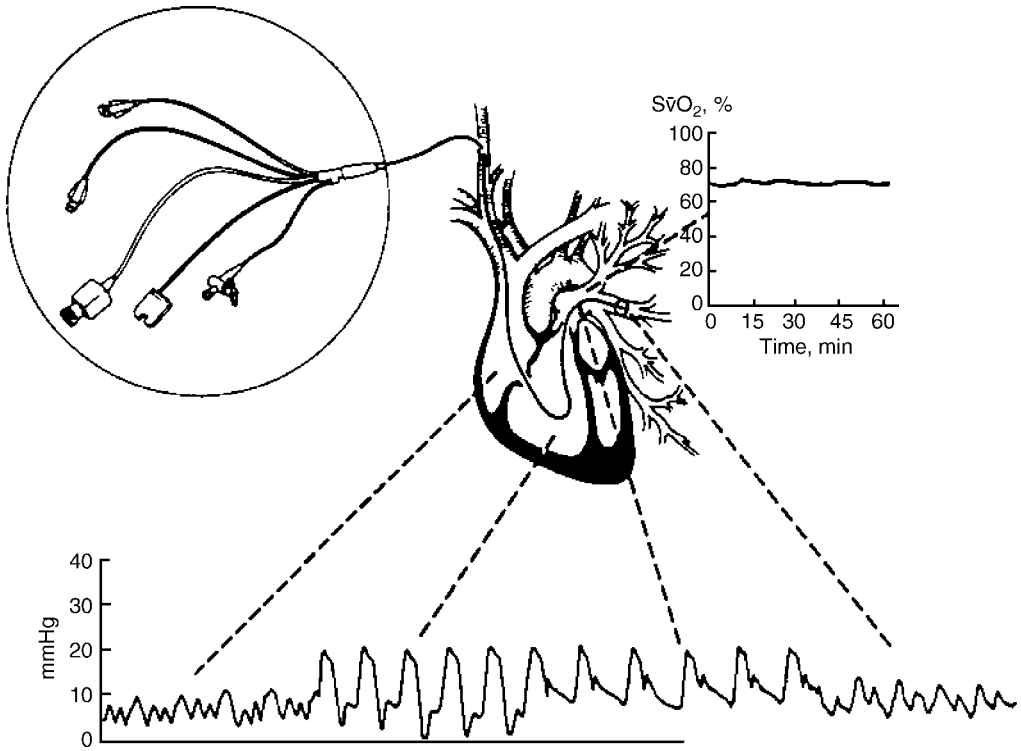


Figure 9. Composite illustration showing normal pressure waveforms obtained as a fiber optic balloon flotation pulmonary artery catheter (Swan-Ganz) is advanced from the right atrium to the pulmonary artery wedge position. [From Daily and Tilkian in Reading List (1986), with permission.]

usually inserted may lead one to conclude that the measurements of pulmonary artery and wedge pressure (PCWP) are easily and reliably measured. However, such is not the case.

Pulmonary artery pressures can be measured accurately only if the following steps are taken (24–27):

1. The monitor is properly zeroed.
2. Strip-chart recordings of all PA pressures for a time period covering at least three respiratory cycles are obtained. Using only the monitor's digital displays is insufficient.
3. Dynamic response testing (fast flush) should be obtained when the catheter is in each position (i.e., wedge and PA). If the dynamic response is not adequate, the problems with the catheter-tubing-transducer system must be resolved before accurate pressures can be measured.
4. Pressures (i.e., systolic, diastolic, and mean pressures) should be assessed from a monitor's display or a strip-chart recording. The pressure measures should be made at the end expiration when the transmural pressure is nearest zero.

CARDIAC OUTPUT DETERMINATION

Cardiac output is the volume of blood ejected by the heart every minute. Cardiac output is a helpful measurement since it can be used to evaluate the overall cardiac status of the critically ill patient, as well as help make the diagnosis of cardiovascular disease. Ideally a cardiac output measurement system would be continuous, automatic, minimally invasive, accurate, fast, inexpensive, and easy to use clinically. The most common method used to measure cardiac output in critically ill patients is still the indicator dilution method. The pulmonary artery catheter (Swan–Ganz) introduced in the 1970s revolutionized the ease with which cardiac output could be measured.

The thermal dilution method requires injection of cold physiological solution, usually normal saline, into the superior vena cava or right atrium. Cardiac output is determined by measuring the area under the time–temperature curve measured in the pulmonary artery that results from the injection of the cold solution.

The thermal dilution method for determining cardiac output relies on several assumptions that are not always correct. First, the exact amount of thermal indicator injected cannot be quantitated precisely. Second, indicator is lost at various stages and this loss of indicator (heat loss) leads to errors.

A block diagram of the thermal dilution measuring system with typical thermal dilution curves and time of injection indicated are shown in Fig. 10. Figure 10c and d show the transit time for the cooled blood moving from the injection site in the right atrium to the pulmonary artery measurement site. Calculation of cardiac output requires measuring the area under the curve. Consequently, a baseline temperature must be established before the injection. In turn, the end point is usually determined by extrapolating to the baseline temperature.

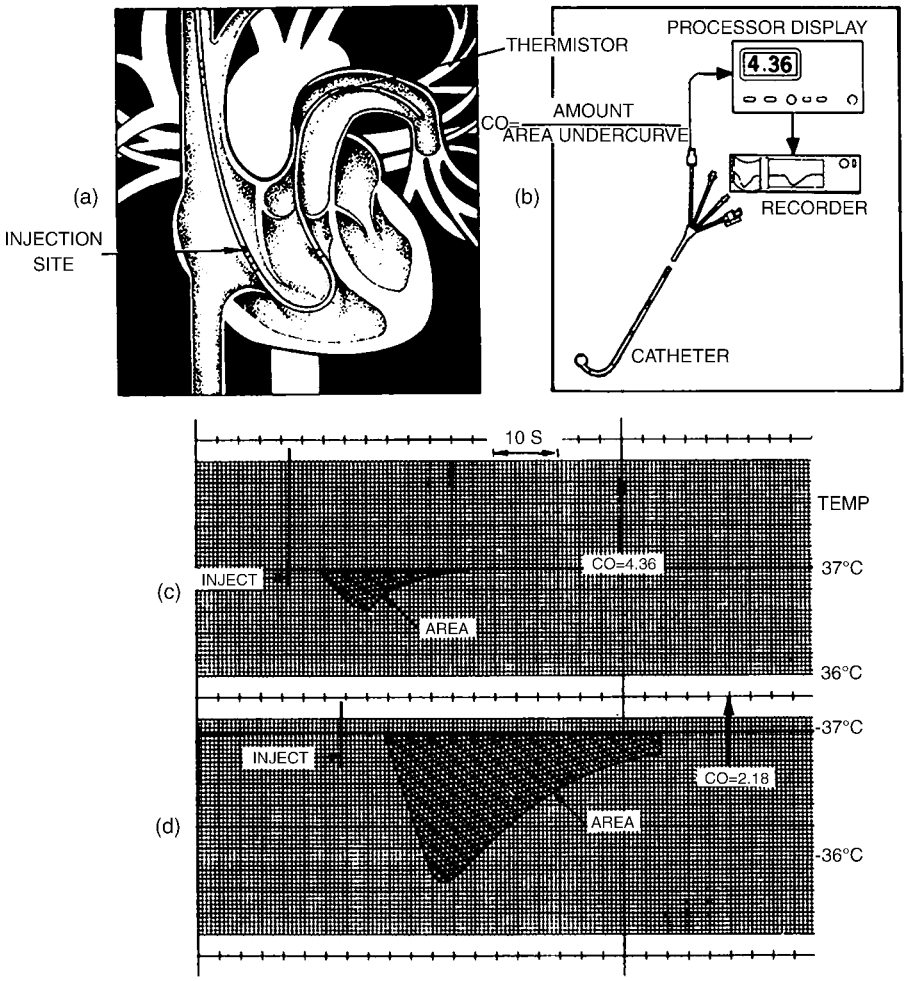


Figure 10. Schematic diagram of the thermal dilution measurement of cardiac output. A recorder of some type should always be used to verify the quality of the thermal dilution curve, (a) shows the thermal dilution catheter placed into the pulmonary artery. Note the location of injection site and thermistor, (b) shows the connection of the thermal dilution catheter connected to a cardiac output processor and recorder, (c) shows a typical temperature-time plot sensed by the thermistor near the catheter tip after an iced saline injection. The cardiac output determined in this case was 4.36 $L \cdot min^{-1}$. (d) Shows a similar temperature–time plot for a patient with low cardiac output (2.18 $L \cdot min^{-1}$). Note the larger area and broader dispersion of the waveform caused by the lower flow. [Reproduced from Ref. 9, with permission.]

To ensure that accurate thermal dilution cardiac output results are obtained, it is recommended that the thermal dilution curves be presented on a monitoring screen or on a strip-chart recorder. Studies have shown that synchronizing the injections with the respiratory cycle improves the technique's reproducibility (28). Since there is considerable variability in cardiac output between measures, at least three reproducible curves are usually obtained. Averaging the findings from these three curves gives a more representative assessment of cardiac output.

In recent times, the complexities of using the Swan–Ganz pulmonary artery catheter have result in controversies. Some clinicians feel that such systems should only be used only when needed and then only sparingly while others have a differing viewpoint (29,30). Still others have questioned the ability of making accurate central venous and pulmonary artery occlusion pressures and whether it matters (26,27). Many of those issues will be resolved in the future when there might be better methods for measuring hemodynamic parameters. Until that time, physicians and nurses caring for critically ill patients who require hemodynamic monitoring should be aware of several how to oriented publications (31–34).

Alarming Based on Hemodynamic Parameters. Clinical hemodynamic monitoring is now several decades old. What started from a simple beginning has since seen many dramatic changes in both the development of new medical devices and skills of the clinicians to use those devices. However, it is my feeling that we are not yet at optimum hemodynamic monitoring. Some recent publications on the topic are illustrative. Sander and colleagues in Germany have recently looked at categories of patients with elevated heart rates who are at higher risk of cardiac complications (35). Their work resulted in an editorial comment Vital are vital signs (36). Additional recent work at Vanderbilt University indicates that volatility in vital signs is a good predictor of death in trauma patients (37). Finally, the problem of false alarms continues to be a huge problem with current bedside monitors. As part of a Master of Science thesis in Medical Informatics at the University of Utah, an investigator found that only about one-third of the standard alarms for patients in a variety of ICU care were true alarms. Thus about two-thirds of the alarms are false. However, if the alarming system used heart rates determined from both the ECG and the Arterial Blood Pressure, the number of alarms decreased by $\sim$ 50% and the false alarm rate was only $\sim$ 25% (38). Having smarter and better hemodynamic monitoring with better and smarter alarming systems will be crucial for to future monitoring systems.

COMPUTERIZED DECISION SUPPORT

Much has been learned about hemodynamic measurements and how to use the data to calculate derived patient parameters. These parameters can then be used to determine patient status and augment patient therapy (39–42).

Using hemodynamic data available from bedside monitors and combining that data, in a structured and coded electronic patient record allows for optimal computerized decision support (42–44). Morris and his colleagues have stated the value of computerized decision support well (45). Only adequately explicit protocols contain enough detail to lead different clinicians to the same decision when faced with the same clinical scenario. Guidelines of care provide only general guidance to patient care and require clinicians considerable latitude in which care decision should be made. Computerized protocols, on the other hand, can be patient-specific and evidence based (46). Using computerized decision-support tools variation in clinical practice can be reduced and favorable effects on improve patient outcomes can be accomplished (45,46).

FUTURE

There are still needs for improvement in hemodynamic monitoring. Being able to make the measurements continuously, less invasively, and more reliably are areas where progress is needed. Clearly, using computer aided decision-support technology to help reduce false alarms and to guide clinicians in making better patient diagnosis and more timely and more optimal and effective treatment decision offer ample opportunity for future research and progress.

BIBLIOGRAPHY

1. Gravenstein JS, Paulus DA. Monitoring Practice in Clinical Anesthesia. Philadelphia (PA): Lippincott; 1982.
2. Bruner JMR. Handbook of Blood Pressure Monitoring. Littleton (MA): PSG Publishing Co.; 1978.
3. Daily EK, Schroeder JS. Techniques in Bedside Hemodynamic Monitoring. 3rd ed. St. Louis (MO): Mosby; 1985.
4. Pinsky MF. Functional hemodynamic monitoring. Intensive Care Med 2002;28:386–388.
5. Pinsky MF. Hemodynamic monitoring in the intensive care unit. Clin Chest Med 2003;24:549–560.
6. Gardner RM, Hollingsworth KW. Optimizing ECG and pressure monitoring. Crit Care Med 1986;14:651–658.
7. Pickering TG, et al. Recommendations for blood pressure measurement in humans and experimental animals: part 1: Blood pressure measurement in humans: A statement for professionals from the subcommittee of professional and public education of the American Heart Association council on high blood pressure research. Circulation 2005;111:697–716.
8. Gardner RM. Hemodynamic monitoring: From catheter to display. Acute Care 1986;12:3–33.
9. Gardner RM, Schwartz R, Wong HC, Burke JP. Percutaneous indwelling radial-artery catheters for monitoring cardiovascular function (Prospective Study of the Risk of Thrombosis and Infection). N Engl J Med 1974;290:1227–1231.
10. Kline AM. Pediatric catheter-related bloodstream infections: Latest strategies to decrease risk. AACN Clin Issues 2005;16: 185–198.
11. Bone RC, et al. Standards of evidence for the safety and effectiveness of critical care monitoring devices and related interventions. Coalition for critical care excellence: Consensus Conference on Physiological Monitoring Devices. Crit Care Med 1995;23:1756–1763.
12. Kutik MH, Gardner RM. Standard for Interchangeability and Performance of Resistive Bridge Blood Pressure Transducers, Arlington (VA): Association for the Advancement of Medical Instrumentation (AAMI); 1986.
13. Gardner RM. Accuracy and reliability of disposable pressure transducers coupled with modern pressure monitors. Crit Care Med 1996;24:879–882.
14. McCann II UG, et al. Invasive arterial BP monitoring in trauma and critical care. Effects of variable transducer level, catheter access, and patient position. Chest 2001; 120:1322–1326.
15. Gardner RM. Direct blood pressure measurement dynamic response requirements. Anesthesiology 1981;54(3):227–236.
16. Gardner RM. Blood pressure monitoring. In: Webb AR, Shapiro MJ, Singer M, Suter PM, editors. Oxford Textbook of Critical Care. Oxford University Press. 1998; p 1087–1090. chapt 16.
17. Gardner RM. Fidelity of recording: Improving the signal-to-noise ratio. In: Martin J. Tobin, editors. Principles and Practice of Intensive Care Monitoring. New York: McGraw-Hill; 1997. p. 123–132. chapt. 8.
18. Kleinman B, Powell S, Kumar P, Gardner RM. The fast flush test measures the dynamic response of the entire blood pressure monitoring system. Anesthesiology 1992;77:1215–1220.
19. Kleinman B, Powell S, Gardner RM. Equivalence of fast flush and square wave testing of blood pressure monitoring systems. J Clin Monit 1996;12(2):149–154.
20. Promonet C, et al. Time-dependent pressure distortion in a catheter-transducer system: Correction by fast flush. Anesthesiology 2000;92:208–218.
21. Swan HJC, et al. Catheterization of the heart in man with the use of a flow directed balloon-tipped catheter. N Engl J Med 1970;283:447–451.
22. Ganz W, et al. A new technique for measurement of cardiac output by thermodilution in man. Am J Cardiol 1971;27:392–396.
23. Cole J, Martin WE, Cheung PW, Johnson CC. Clinical studies with a solid state fiberoptic oximeter. Am J Cardiol 1972;29: 383–388.
24. Morris AH, Chapman RH, Gardner RM. Frequency of wedge pressure errors in the ICU. Crit Care Med 1985;13:705–708.
25. Cengiz M, Crapo RO, Gardner RM. The effect of ventilation on the accuracy of pulmonary artery and wedge pressure measurements. Crit Care Med 1983;11:502–507.
26. Rizvi K, et al. Effect of airway pressure display on interobserver agreement in the assessment of vascular pressure in patients with acute lung injury and acute respiratory distress syndrome. Crit Care Med 2005;33:98–103.
27. Liebowitz AB. More reliable determination of central venous and pulmonary artery occlusion pressures: Does it matter? Editorial Crit Care Med 2005;33:243–244.
28. Stevens JH, et al. Thermodilution cardiac output measurements: Effects of the respiratory cycle on its reproducibility. J Am Med Assoc 1985;253:2240–2242.
29. Pinsky MR, Vincent J-L. Lets use the pulmonary artery catheter correctly and only when we need it. Point of view. Crit Care Med 2005;33:1119–1122.
30. Levin PD, Sprung SL. Another point of view: No Swan song for the pulmonary artery catheter. Crit Care Med 33:1123–1124.

31. Gibbs NC, Gardner RM. Dynamics of invasive pressure monitoring systems: Clinical and laboratory evaluation. Heart Lung 1988;17:43–51.
32. Gardner RM, Hujcs M. Fundamentals of physiologic monitoring. AACN Clinical Issues Crit Care Nursing 1993;4(1):11–24.
33. Daily EK. Hemodynamic waveform analysis. J Cardiovasc Nurs 2001;15(2):6–22.
34. McGhee H, Bridges EJ. Monitoring arterial blood pressure: What you may not know. Crit Care Nurse 2002;22:60–78.
35. Sander O, et al. Impact of prolonged elevated heart rate on incidence of major cardiac events in critically ill patients with a high risk of cardiac complications. Crit Care Med 2003; 33:81–88.
36. Weissman C, Landesberg G. Vital are the vital signs. Comment/Editorial Crit Care Med 2003;33:241–242.
37. Grogan EL, et al. Volatility: A new vital sign identified using a novel bedside monitoring strategy. J Trauma 2005;58:7–14.
38. Poon KB. Fusing Multiple Heart Rate Signals to Reduce Alarms in Adult the Intensive Care Unit. MS dissatation, University of Utah Department of Medical Informatics, Salt Lake City (UT); May 2005.
39. Gardner RM. Information management hemodynamic monitoring. Semin Anesth 1983;2:287–299.
40. Gardner RM. Computerized management of intensive care patients. MD Comput 1986;3(1):36–51.
41. Shabot MM, Carlton PD, Sadoff S, Nolan-Avila L. Graphical reports and displays for complex ICU data: A new, flexible and configurable method. Comput Methods Programs Biomed 1986;22:111–116.
42. Gardner RM, Huff SM. Computers in the ICU: Why? What? So what? Intl J Clin Monit Comput 1992;9:199–205.
43. Gardner RM, Sittig DF, Clemmer TP. Computer in the ICU: A match meant to be! In: Ayers SM, Grenvik A, Holbrook PR, Shoemaker WC. editors. Textbook of Critical Care Medicine. 3rd ed. Philadelphia: W. B. Saunders;1995: p 1757–1770. chapt. 196.
44. Morris AH. Treatment algorithms and protocolized care. Curr Opin Crit Care 2003;9:236–240.
45. Morris AH. Clinical trial of a weaning protocol. Crit Care 2004;8:207-209.
46. Bria II WF, Shabot MM. The electronic medical record, safety and critical care. Crit Care Clin 2005;21:55–79.

Further Reading

Nichols WW, O'Rourke MF. McDonald's Blood Flow in Arteries: Theoretical, Experimental and Clinical Principles. 5th ed. Of special note is Chapter 6: Measuring principles of arterial waves pages 132–135. Hodder Arnold; 2005.

Daily EK, Tilkian AG. Hemodynamic Monitoring. In: Tilkian AG, Daily EK, editors. Cardiovascular Procedures: Diagnostic Techniques and Therapeutic Procedures. St. Louis (MO): Mosby; 1986 chapt. 4.

Geddes LA. The Direct and Indirect Measurement of Blood Pressure. Chicago: Year Book Medical Publisher; 1970.

Geddes LA. Handbook of Blood Pressure Measurement. Clifton (NJ): Humana Press; 1991.

Webster JG, editor. Design of Pulse Oximeters. Institute of Physics; 1997.

Webster JG, editor. Medical Instrumentation: Application and Design. 3rd ed. New York: John Wiley & Sons Inc.; 1998.

Tungjitkusolmun S, Heart and Circulation, In: Webster JG, editor. Bioinstrumentation. John Wiley & Sons, Inc.; 2004. Chapt. 8.

Shabot MM, Gardner RM, editors. Decision Support Systems for Critical Care. New York: Springer-Verlag Inc.; 1994.

Gardner RM, Shabot MM. Patient-monitoring systems. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care. 3rd ed. New York: Springer-Verlag; 2005; in press, Aug. 2005. Chapt. 17.

See also BLOOD PRESSURE MEASUREMENT; CARDIAC OUTPUT, FICK TECHNIQUE FOR; CARDIAC OUTPUT, INDICATOR DILUTION MEASUREMENT; CARDIAC OUTPUT, THERMODILUTION MEASUREMENT; ELECTROCARDIOGRAPHY, COMPUTERS IN; HEMODYNAMICS.

MONITORING, INTRACRANIAL PRESSURE

MICHAEL L. DALEY
IAN PIPER
The University of Memphis
Memphis, Tennessee

INTRODUCTION

Intracranial pressure (ICP) monitoring is a form of pressure monitoring used in patients with pathologies that might give rise to raised ICP. The definition of raised ICP depends on the underlying pathology and, for example in adult patients who have sustained a severe head, injury is defined as pressure greater than 20 mm Hg. Over the past 50 years there has been an active and wide ranging research into the causes and consequences of raised ICP that, to date, has been the subject of 11 international symposia embracing such diverse disciplines as neurosurgery, anesthesia, radiology, biophysics, electrical and mechanical engineering, mathematics, and computer science. This article reviews the underlying physiology pertinent to measurement of ICP as well as gives an overview of some of the current technology used to measure ICP. Then, a brief review follows of the clinical literature underlying the case for ICP measurement, with an emphasis on the main clinical condition of patients with a head injury. The last two sections focus on the use of waveform analysis and mathematical modeling techniques to research into mechanisms underlying raised ICP.

PHYSIOLOGY

ICP is the pressure recorded within the tissue (parenchyma) or fluid-filled spaces and is not uniformly distributed within the craniospinal axis. The craniospinal axis consists of all neural tissue and contiguous fluid-filled spaces within the cranium and spinal sac. The central neural tissue is encapsulated within bone and three-layered tissue coverings or *meninges*. From outside in the meninges are the dura, arachnoid, and pia membranes. The meninges provide a physical barrier between neural tissue and the external environment but also serve both a structural and a physiological function as an architecture for supporting cerebral vessels and maintaining a space for flow of cerebrospinal fluid (CSF). CSF cushions the delicate neural tissues, supports the weight of the brain, and acts as a transport media for nutrients, chemical messengers, and waste products. Except at the choroid plexus, the CSF is freely permeable

to the appendymal lining and is in chemical communication with the interstitial fluid and the CNS. ICP is a reflection of the relationship between alterations in craniospinal volume and the ability of the craniospinal axis to accommodate added volume. The craniospinal axis is essentially a partially closed box with container properties including both viscous and elastic elements. The elastic or, its inverse, the compliant properties of the container will determine what added volume can be absorbed before ICP begins to rise. So an understanding of raised ICP encompasses an analysis of both intracranial volume and craniospinal compliance.

The history of the subject of ICP has been well reviewed (1,2) and starts from the doctrine named after Monro (3) and Kellie (4), which proposed that the brain and its contained blood were incompressible, enclosed in a rigid and inextensible skull, of which the total volume remained constant. In its original form, the Monro–Kellie doctrine did not take into account the CSF as a component of the cranial compartment. The concept of reciprocal volume changes between blood and CSF was introduced in 1846 by Burrows and later extended in the early twentieth century by Weed et al. (5,6) to allow for reciprocal changes in all craniospinal constituents. Kocher (7) in 1901 translated into clinical terms the four stages of cerebral compression proposed almost 25 years earlier by the experimental studies of Duret (8). Kocher described four stages of cerebral compression related to the expansion of intracranial brain tumours. In stage 1, the initial increase in tumor volume is compensated by a reduction in volume of the other intracranial components, chiefly CSF and venous blood. This spatial compensation results in no net increase in intracranial volume or pressure and hence no clinical symptoms. In stage 2, the compensatory mechanisms are exhausted, ICP increases, and the patient becomes drowsy with a headache. Stage 3 is characterized by a considerable increase in ICP, an associated deterioration in conscious level, and intermittent elevations of blood pressure (BP) accompanied by bradycardia. In the fourth and final stage, the patient is unconscious, with bilateral fixed dilated pupils and falling BP usually leading to death.

Cushing (9–11), then a research worker for Kocher, described in both experimental and clinical studies the close relationship between increases in ICP and BP and proposed that the BP rose to maintain adequate blood supply to the hind brain, the stimulus to this vasopressor response believed to be medullary ischemia (12,13). At about this time, a false confidence developed in the lumbar CSF pressure technique (lumbar puncture), which caused Cushing's findings to be challenged. Reports emerged (14–16) that some patients showing clinical signs of brain compression had normal lumbar CSF pressures and that in other patients elevations in BP were found at times when ICP was well below the level of BP.

Partly because of this apparent dissociation between ICP and clinical symptoms, emphasis switched away from ICP measurement toward the relationship between craniospinal volume and pressure, particularly the importance of the elastic properties of the craniospinal system. The relationship between ICP and intracranial volume is often described graphically by an exponential function that is relatively flat at low ICPs but rises rapidly as pressure increases much above 20 mm Hg. Under normal conditions, at the low ICP end of this curve, compliance (the ratio of added volume to pressure) is high. When the patient's volume–pressure status changes to move further up the curve, compliance will fall rapidly. Measurement of craniospinal compliance in brain-injured patients may, therefore, offer the potential for early detection of raised ICP before it rises to levels damaging to brain parenchyma. The most commonly used methods of measuring craniospinal compliance were developed by Marmarou (17,18) and depend on the rapid injection of known volumes of fluid into the CSF space with immediate measurement of the resultant increase in CSF pressure. One of these methods is the pressure volume index (PVI); this value being the volume that when added to the CSF space would produce a tenfold rise in ICP. Miller et al. (19,20) defined a further measure of the craniospinal volume–pressure relationship, the volume pressure response (VPR). The VPR, also calculated from the ICP response resulting from a rapid bolus injection of saline into the CSF space, is a direct measure of the inverse of compliance: elastance. Although both are measures of compliance, some confusion remains as to what exactly is the difference between the VPR and the PVI. The PVI, which assumes a mono-exponential pressure versus volume relationship, is a single index that characterizes the patient's entire volume–pressure relationship and is, numerically, a measure of the slope of the logarithm of the ICP versus intracranial volume relationship. Its strength lies in its ability to define the whole volume–pressure status of the patient with a single index. It was the late J. Douglas Miller who pointed out that if there was only a single volume–pressure curve, then no new information would be gained by measuring compliance or elastance, and a knowledge of absolute ICP alone would suffice in determining the state of a patient's craniospinal volume decompensation. However, several studies (21,22) have shown that the shape of the volume–pressure relationship changes under a variety of conditions both between patients and within patients at different times. Under these circumstances, it is likely that the PVI will provide the more useful information. One weakness of the PVI is that if a patient's pressure–volume relationship remains stable, single measurements will not detect movement along a given pressure–volume curve and thus may not detect slow increases in volume of a space-occupying lesion before it causes significant increases in ICP. It is in this situation where a continuous measure of absolute compliance (or its inverse: elastance as measured by the VPR) provides, potentially, more useful information. It is for this reason that, in head-injured patients, Miller et al. found that the VPR correlated better to the degree of brain mid-line shift, as imaged on CT scan, than it did to absolute ICP alone. They subsequently demonstrated that the VPR could serve as an indicator for surgical decompression, critical levels being between 3 and 5 mm Hg/mL (19,23). However, in clinical practice, both circumstances (movement along the pressure–volume curve and shift of the curve along the pressure axis) may occur within a patient at different times and thus shows the need for ICP monitoring capable of both continuous compliance measurement and derivation of the PVI.

Few clinicians would argue over the theoretical utility of monitoring compliance in brain-injured patients; what has limited its use in practice has been the inherent problems associated with manual volume–pressure testing. It is difficult to inject equal volumes of fluid manually at a constant and rapid rate of injection. As a result, repeated measures are usually required that can result in a lengthy and time-consuming procedure. Also, even with stringent maintenance of the sterile procedure, repeated addition or removal of CSF from patients carries an increased risk of infection to the patient. Despite the reported benefits of compliance measurement, it has been largely these limitations that have prevented widespread adoption of compliance measurement in clinical practice.

It was not until the 1960s when Lundberg (24) published his now classic monograph that interest in clinical ICP measurement was rekindled. Using ventricular fluid pressure recording in brain tumor patients, Lundberg was the first to delineate the frequency with which raised ICP occurs clinically, at times reaching pressures as high as 100 mm Hg. Lundberg also described three types of spontaneous pressure wave fluctuations: "A" waves or plateau waves of large amplitude (50–100 mm Hg) with a variable duration (5–20 min), "B" waves that are smaller (up to 50 mm Hg), and sharper waves with a dominant frequency of 0.5–2 cycles per min.

DEVICES

ICP was first measured experimentally in animals by Baylis (25) in 1897 using an early form of strain gauge. The most common form of strain gauge use specially prepared alloys that change their resistance in proportion to the amount they are elongated or stretched in cross section due to applied strain.

The Wheatstone Bridge and the Strain Gauge

Most common strain gauge transducers are used in a Wheatstone Bridge configuration, where the resistive strain elements are placed on diagonally opposite arms of the bridge. Should the strain gauge change its resistance (for example, due to an applied strain—perhaps caused by a pressure acting on the strain gauge to cause it to elongate), then the bridge will become unbalanced and a potential difference will be generated in proportion to the degree of strain.

Fluid-Filled Catheter Transducer Systems

The standard intraventricular catheter connected to an external strain gauge transducer is called a catheter-transducer system because it behaves, in many ways, like a mechanical system with a *mass* of fluid that acts against the spring-like "elastic" properties of the catheter walls and the transducer diaphragm. A typical catheter-transducer system is one used for the measurement of arterial pressure comprising a silicon strain gauge, fluid-filled catheter, three-way tap, and an arterial cannulae. Modern transducers are semiconductor fabricated and normally disposable. If a flushing device is attached to the transducer, extreme care must be used to avoid accidental flushing into the cranium. The older catheter-transducer system included long tubing and therefore had different frequency characteristics than modern transducers. The older transducer first found widespread use in the 1960s and 1970s following the pioneering work on long-term ICP monitoring by Nils Lundberg (24). Most publications in this field still refer to intraventricular monitoring as the "Gold-Standard" method of measuring ICP for several reasons. First, this method allows checking for zero and sensitivity drift of the measurement system *in vivo*. Second, pressure measurement within the CSF space transduces pressure within a medium that is an incompressible fluid and, provided CSF flow is not blocked, is not subject to the development of intracompartmental pressure gradients. Finally, access to the CSF space provides a method for ICP treatment via CSF drainage. However, concerns are often expressed about the increased risks of infection associated with ventriculostomy. Although a range of infection rates has been reported, some as high as 40% (26), recent reports confirm infection rates to be in the region of 1%, which is not considered a prohibitive risk (27).

The Head Injury Management Guidelines published by the Brain Trauma Foundation in 1994 recommend intraventricular ICP measurement as the first-line approach to monitoring ICP. They state that "A ventricular catheter connected to an external strain gauge transducer or catheter tip pressure transducer device is the most accurate and reliable method of monitoring ICP and enables therapeutic CSF drainage. Clinically significant infections or haemorrhage associated with ICP devices causing patient morbidity are rare and should not deter the decision to monitor ICP" (28). Despite the existence of these guidelines (27,28), catheter-tip intraparenchymal pressure monitoring remains popular, particularly in the United Kingdom, as it does not require catheter placement in the operating theater and thus carries significantly lower resource implications. However, the routine use of fluid-filled catheter-transducer systems is not without difficulties. A catheter-transducer system can be described as a second-order mechanical system and, if under-damped, will oscillate at its own natural frequency producing significant amplitude and phase distortion of the pressure signal. The degree of distortion will depend on the damping factor (β) of the system. For most purposes, a damping factor of 0.64 is optimal as the amplitude error will be less than 2% for up to two thirds of the natural frequency of the system (29). Typically, most pressure catheter-transducer systems used in patients tend to be under-damped with a damping factor (β) less than 0.4 (29,30). Manipulating a catheter-transducer system from under-damped ($\beta < 0.3$) to over-damped ($\beta > 0.8$) can cause a decrease in mean pressure of 7 mm Hg. Commercial devices are available, however, such as the acudynamic adjustable damping device (31) that can alter the damping characteristics of external strain gauge pressure transducers and bring them within the range of optimal damping. Another problem with fluid-filled catheter-transducer systems is correcting for the presence of hydrostatic pressure gradients when measuring cerebral perfusion pressure (CPP). Typically, the external strain gauge transducer is zeroed at the same

level as the arterial pressure (BP) transducer, usually at the level of the right atrium. Although the patient is managed in the horizontal position, there is no column of fluid between the site of ICP and BP measurement. However, when the patient is managed with head-up tilt and if the BP transducer is not moved to the same horizontal level as the head, a hydrostatic fluid column will be created. This can produce a significant error between the observed CPP and the actual CPP, which, in the worst case, can produce an error of as much as 15 mm Hg.

The Spiegelberg ICP monitoring system (Spiegelberg KG, Homburg, Germany) largely overcomes these problems. This system is a special case of a fluid-filled catheter-transducer system. With this device, ICP is measured using a catheter with an air pouch balloon situated at the tip. By maintaining a constant known volume within the air pouch, the pressure within the air pouch balloon is equivalent to the surrounding pressure or ICP. The internal air pouch balloon is transduced by an external strain gauge transducer, and because the fluid used for pressure transduction is air, the pressure error caused by an "air column" is clinically insignificant. The design of this device also allows automatic *in vivo* zeroing of the ICP system and, in laboratory bench tests, showed the least zero drift in comparison with standard catheter-tip ICP devices (32). This system now has versions for use in epidural, subdural, intraparenchymal, and intraventricular sites. The intraventricular catheter is a double lumen catheter that allows access to the CSF space for drainage. The Spiegelberg system, although having many attractive features, is, as yet, a relatively little used system outside of Germany and its long-term clinical utility and robustness require evaluation.

Catheter-Tip Transducer Systems

Now several catheter-tip ICP monitoring systems are available, including the Camino (33–36) systems. The Gaeltec ICP/B solid-state miniature ICP transducers, for use in the epidural space, are reusable, and the zero reference can be checked *in vivo*. However, reports of measurement artifacts (37) and decay in measurement quality associated with repeated use (35) have limited the widespread adoption of this technology.

The InnerSpace OPX 100 system (InnerSpace Medical, Irvine, CA) is, like the Camino system, a fiber optic system. Bench test reports on this system show it to have good zero drift and sensitivity stability (32). However, a recent clinical evaluation of this system in 51 patients reported a high (17%) incidence of hematoma formation around the ICP sensor (36). The authors concluded that improved fixation of the catheter is required to minimize micro-movements.

The two catheter-tip systems most frequently used in the management of head-injured patients are the Codman and Camino systems. Neither allows a pressure calibration to be performed *in vivo*. After these systems are "zeroed," relative to atmospheric pressure during a pre-insertion calibration, their pressure output is dependent on zero drift of the sensor. For this reason, it is critical that these devices exhibit good long-term zero drift characteristics. These devices provide an electrical calibration, to calibrate external monitors, but they cannot be corrected for inherent zero drift of the catheter once placed. For the Camino system, the manufacturers specify the zero drift of the catheter to be ± 2 mm Hg for the first day and ± 1 mm Hg per day thereafter. Czosnyka et al. (32) have confirmed these zero drift findings in bench tests studies although they also reported that the temperature drift of the device was significant (0.3 mm Hg/°C). They reported that if the manufacturers specify the zero drift of the catheter to be ± 2 mm Hg for the first day and ± 1 mm Hg per day thereafter. In clinical practice, the reported zero drift upon removal of the Camino device from the patient has been reported to be greater than the manufacturer's specifications. Munch (38) assessed 136 Camino sensors in a clinical study and found an average *daily* drift rate of 3.2 mm Hg. Chambers (39), in a comparative study of the Camino ventricular catheter with an external fluid-filled catheter-transducer system, reported that only 60% of the readings were within 2 mm Hg of the gold-standard method. There are also reports of Camino probe failure because of technical complications (cable kinking, probe dislocation), with reported failure rates ranging from 10% to 25% (39,40).

The Codman transducer is a micro-miniature strain gauge within a titanium housing side mounted at the tip of a catheter. Similar to other transducers, bench test reports on this technology have been favorable (32,41). However, clinical evaluations have reported the presence of inter-patient and intra-patient biases that are independent on whether the device is compared against the Camino transducer or an intraventricular catheter-transducer system (42). A report by Fernades (43) found that in 24% of the recordings, the Codman sensor over-read the Camino system by 5 mm Hg or more.

The Rehau System

The technology of miniaturization is allowing the development of catheter-tip pressure sensors. Catheter-tip systems, due to their small diameter, are likely to cause less damage to tissue upon placement than larger fluid-filled catheters and are not affected by hydrostatic pressure differences. However, they are potentially more prone to problems of robustness and *in vivo* zero drift. Several similar technologies, although performing well in bench test studies, have been shown to exhibit unacceptable zero drift, fragility, or both during trials conducted under clinical conditions (38,43–45). *In vivo* drift is especially important in catheter-tip strain-gauge technology as it is impossible to check if the calibration has altered after being placed in the patient. Until recently, to reduce the physical size of the catheter-tip strain gauge catheters, often only a partial Wheatstone bridge is employed at the catheter-tip. Recently, a new technology has become available, the Neurovent-P (REHAU AG+CO, REHAU, Germany), in which a full Wheatstone bridge is fabricated in the probe tip, and this solution should, in theory, provide improved zero drift characteristics. One potential advantage of the Rehau NeuroVent system seems to be the incorporation of a full Wheatstone bridge into the catheter-tip electronics. This Wheatstone technological improvement should enhance the zero drift characteristics by reducing temperature sensitivity and the effects of non-pressure-related external

strains. Until recently, this approach was only possible in the much physically larger external strain gauge systems. Through advances in miniaturization technology, it has now been able to incorporate this technology into its catheter-tip systems. In the bench studies performed (46), each catheter was tested for days, from a minimum of 3 to more than 8 days. This timeframe typically covers the period in which an ICP transducer is used in the clinical setting. The results from this bench test study confirm the manufacture's claim that pre-insertion calibration is not required as the drift was within manufacture's specifications (±1 mm Hg). This advantage is very important in the clinical arena because the first pre-insertion zeroing is a crucial step conducted by the surgeons and potentially, if incorrectly performed, could generate erroneous readings during the period of monitoring. Both long-term zero drift and the dynamic pressure test results also confirm that this system performs well in bench test studies and meets manufacturer's specifications. Mean zero drift, after 5 days, was very small and long-term continuous recording of a stable pressure was precise. The response to dynamic tests, i.e., the changes of pressure over a wide range, was excellent. The average bias of the Rehau catheter compared with a hydrostatic pressure column is very small. Despite these promising bench test study results, further work is required to determine the performance of this measurement device in the clinical environment. Following on from these bench tests, the next and most critical step will be to conduct a trial of this promising technology under the more demanding clinical environment. The BrainIT group (http://www.brainit.org) as a multicenter collaborative group of neurointensive care scientists and clinicians are well placed to design and conduct such a trial (47). Should this technology demonstrate, under clinical conditions, the required robustness and low drift as indicated by these bench test studies, it may lead to more precise and reliable measurement of ICP.

CLINICAL LITERATURE

Raised ICP has been found to be associated with a poorer outcome from injury with the higher the level of ICP, particularly the peak ICP level, which has been found to correlate with the expected prognosis for mortality and morbidity (48–51). There has, however, been controversy over the usefulness of monitoring raised ICP with some groups, with a "no ICP monitoring" policy, finding in their studies of head injury mortality and morbidity that outcome is similar to other groups that do monitor ICP (52). Reported differences in the utility of ICP monitoring could be due to variability both in management and in monitoring protocols between different neurosurgical centers. Variation in type of ICP pressure monitor, site of placement, treatment thresholds, patient referral characteristics, and in outcome measures can all combine to produce a large variability both in measured ICP and in outcome irrespective of whether ICP is monitored or how it is treated. Another source of variation in terms of raised ICP is the inherent variability of the head-injured population with outcome being dependant on several other factors. For example, mass lesions are generally accompanied by elevations in ICP of greater than 40 mm Hg and are associated with poorer outcome, whereas diffuse injuries tend to have lower ICP levels associated with a similar poor outcome (50,53). Age is also an important factor with an age-dependent distribution of ICP for both type of injury and outcome. This is particularly so for pediatric cases (54–56). ICP can even be raised in the absence of overt signs of swelling or mass lesions on CT. In a small study of severely head-injured patients, O'Sullivan (57) demonstrated that some comatose head-injured patients whose initial CT scan was normal, with no mass lesion, midline shift, or abnormal basal cisterns, developed raised ICP greater than 20 mm Hg that lasted longer than 5 min. This included a subset of patients showing pronounced raised ICP of greater than 30 mm Hg.

Data from large prospective trials carried out from single centers and from well-controlled multicenter studies have provided the most convincing evidence for a direct relationship between ICP and outcome (58–61). Narayan et al. (58) in a prospective study in 133 severely head-injured patients demonstrated that the outcome prediction rate was increased when the standard clinical data such as age, Glasgow Coma Score on admission (GCS), and pupillary response with extraocular and motor activity was combined with ICP monitoring data. Marmarou et al. (60), reporting on 428 patients' data from the National Institute of Health's Traumatic Coma Data Bank, showed that following the usual clinical signs of age, admission motor score, and abnormal pupils, the proportion of hourly ICP recordings greater than 20 mm Hg was the next most significant predictor of outcome. Outcome was classified by the Glasgow Outcome Score (GOS) at 6 months follow-up. They also found, using step-wise logistic regression, that after ICP, arterial pressure below 80 mm Hg was also a significant predictor of outcome. Jones et al. (61) studied prospectively 124 adult head-injured patients during intensive care using a computerized data collection system capable of minute by minute monitoring of up to 14 clinically indicated physiological variables. They found that ICP, above 30 mm Hg, arterial pressure below 90 mm Hg, and cerebral perfusion pressure below 50 mm Hg significantly affected patient morbidity.

Although differing opinions remain about the contribution of continuous monitoring of ICP to reduction in mortality and morbidity after head injury, there is now sufficient evidence to remove doubt about the value of ICP monitoring toward improving the detection and preventative management of secondary cerebral injury.

Raised Intracranial Pressure: Relationship to Primary and Secondary Injury

Both experimental and clinical studies have clearly shown that after traumatic brain injury, normal physiological mechanisms for maintaining cerebral perfusion can become impaired (62–65). These studies demonstrate that brain injury can cause impairment or loss of autoregulation defined as the ability of the cerebral vessels to respond to changes, in arterial gases or to arterial pressure. As a result of these changes, there can, at times, be a decrease

in cerebrovascular resistance that can lead to raised ICP in both adults and children (66–70). Although brain-injured patients are being managed in intensive care, there are, superimposed on to the primary injury, periods of reduced arterial PO2 or episodes of arterial hypotension often as a result of other injuries or treatment by hypnotic drugs (50,53,60,61,71,72). With an impaired physiological mechanism unable to respond adequately to these adverse changes in physiological parameters (or "secondary insults"), ischemic brain damage can occur. These secondary, chiefly ischemic brain insults are common, with Graham et al. (74) reporting, in a series of 151 fatal cases of severe head injury, a 91% incidence of ischemic brain damage found on autopsy. A second study carried out by the same group over 10 years later found a similar high incidence (>80%) of ischemic brain damage despite subsequent improvements in intensive care of head-injured patients (74).

There has been much interest in the relationship among ICP, CPP, and CBF. The landmark study of Miller and Garibi (75) produced some of the first experimental evidence confirming the concept that changes in ICP affect cerebral blood flow not directly but through changes in CPP, where CPP is defined as the difference between mean arterial pressure and ICP. Strictly speaking, the actual cerebral perfusion outflow pressure would be cerebral venous pressure, although this pressure is, in most situations, impractical to measure routinely. However, it has been established that over a wide range of pressures cerebral venous pressure is well approximated (within 3–4 mm Hg) by ICP (76–78). In an experimental study of CBF as determined by the venous outflow technique in dogs, Miller and Garibi (75) also demonstrated that when MAP and ICP rise in parallel so that CPP remains constant at 60 mm Hg, CBF increases with MAP in animals found to be non-autoregulating. It was further shown that as CPP drops in autoregulating animals, the breakpoint at which CBF starts to decrease is at a higher level if CPP is reduced through hemorrhagic arterial hypotension than through intracranial hypertension. This work suggests that cerebral perfusion is more sensitive to arterial hypotension than to intracranial hypertension.

The clinical significance of this information is that in the management of head injury, it is often necessary to employ therapy to lower raised ICP. Therapeutic agents for reducing raised ICP often do so at the expense of reduced MAP, and as a consequence, CPP may not improve. If autoregulation is preserved, CBF should remain unchanged despite parallel changes in MAP and ICP. However, clinically, autoregulation is likely to be impaired in those conditions in which ICP is increased such as head injury or subarachnoid hemorrhage (68,69,79–82). Under these circumstances, it is important that reduction in ICP should not be achieved at the expense of lowering CBF and provoking brain ischemia.

This earlier work of Miller and Garibi was later extended by Chan et al. (83) to include CPP ranges of 60, 50, and 40 mm Hg. At CPP levels of 50 and 60 mm Hg, when autoregulation was intact, CBF remained unchanged. However, with loss of autoregulation, there was a trend for CBF to increase as MAP and ICP were increased in parallel at a CPP of 50 and 60 mm Hg. Absolute CBF levels were significantly different between the autoregulating and non-autoregulating groups. At a CPP of 40 mm Hg CBF showed a linear correlation with BP. This work demonstrates that when autoregulation is impaired, there is a functional difference between autoregulating and non-autoregulating cerebral vessels despite similar MAP and CPP, and that when autoregulation is impaired, CBF depends more on arterial driving pressure than on cerebral perfusion pressure.

The importance of arterial pressure as the prime factor governing CPP-related secondary insults has been well demonstrated by the work of Jones et al. (61), where they carried out a prospective study over 4 years of the frequency and severity with which secondary insults occur to head-injured patients while being managed in intensive care. They developed a microcomputer-based data collection system that allows the acquisition of data from up to 15 monitored variables minute by minute (84). At each bed space, data collection was under the control of a microcomputer where serial links between the patient monitors and the microcomputer allow the controlled transfer of multiple channels of physiological data once per minute. The controlling software allowed medical staff to add comments to the current active computer file at any time, precisely annotating significant events. The software performs artifact detection, calculates derived data, and highlights valid data that falls outside normal physiological levels. Collected data were stored to disk and could be printed either locally or remotely. Later work by Chambers et al. (85) found that both raised ICP and lowered ABP are major factors in producing secondary insults.

From the valid physiological data produced, manual processing of data was used to identify secondary insults that are defined at one of three grades of severity and that must last for 5 min or longer to be recorded as an insult. This permits calculation of the frequency, severity, and total duration of insults, measured in minutes.

An analysis was made of 124 adult head-injured patients who were monitored during intensive care using the computerized data collection system. Information was logged at 1 min intervals and scanned to identify insults when values fell outside threshold limits for 5 min or longer. Three grades of insult were defined for each variable (Table 1). The duration of insults has been analyzed in relation to the GOS of these patients at 12 months after injury. The monitored patients included 68 with severe head injury (GCS 8 or less with no eye opening), 36 with moderate head injury (GCS 9–12), and 20 with minor head injury (GCS 13–15 but with multiple injuries, scoring 16 or more on the Injury Severity Scale).

Insults were found in 91% of patients at all degrees of severity of head injury. Overall, 10% of patients had insults that were only at the lowest grade 1 level, 31% had insults at both grade 1 and grade 2 levels, and 50% of patients had at least one insult at grade 3 level in addition to grades 1 and 2 insults. Overall, the majority (77%) of all insults detected in the ITU were at grade 1 level, and these represented 85% of the total duration in minutes of insult measured. Differences in the duration of insult between outcome groups 12 months post-injury were compared with

each grade of insult using Kruskall–Wallis one-way analysis of variance and Mann–Whitney U tests. Significant differences in the distribution of hypotensive insults were found between the outcome grades at all levels of severity of insult. Similar results were found for cerebral perfusion pressure insult duration. These data confirm the important adverse effect of even moderate reductions in arterial pressure (systolic BP less than 90 mm Hg or mean BP less than 70 mm Hg).

Although the occurrence and clinical significance of severe and long-lasting secondary insults in head-injured patients is not disputed. The incidence, severity, and duration of shorter acting "minute by minute" cerebral perfusion pressure insults, as defined by the Edinburgh secondary insult detection methodology, has not been defined outside of the Edinburgh study population. In addition, the strong association between the occurrence of specific insult types and the subsequent patient morbidity and mortality found by the Edinburgh study (61) needs to be reproduced in other centers. From the Edinburgh group, Signorini et al. (86,87) developed and validated a model for predicting survival in head-injured patients based on collection of simple demographic features. When the minute by minute secondary insult data were added to the baseline model, they found only ICP insults significantly improved the fit of the model.

Almost all of the evidence for a CPP management is based on single-center cohort studies, often compared with historical controls. For example, Rosner and Becker (26) reported on clinical results of a CPP management protocol where approximately 40% of patients received vasopressor support. They reported a greatly improved incidence of favorable outcome in patients' GCS ≥ 7 compared with historical controls. Despite evidence from single-center studies as described above, a critique of the literature for the purposes of defining head injury management guidelines published by the Brain Trauma Foundation in 1994 states that there is not sufficient evidence to establish either a standard or a guideline for the management of CPP; however they indicate management of CPP greater than 70 mm Hg as a management option (28). Since then, Robertson et al. (88) performed one of the first randomized controlled single-center trials of the CPP management approach. They defined two management cohorts, one based on their normal practice of CPP management and the other CBF, guided practice that included the aggressive management of CPP above 70 mm Hg, together with restricted use of significant hyperventilation. Conversely, their trial has shown that aggressive management of CPP > 70 mm Hg, although reducing the incidence of jugular venous desaturation $< 50\%$, demonstrated no difference in neurological outcome possibly due to the increased incidence of acute respiratory distress syndrome in the CBF management group. Thus, secondary insults are common, result in mainly ischemic brain damage, and are a major contribution to disablement. Moreover Chambers et al. (89) has reported that the critical CPP thresholds for insults of pediatric patients varies with age. For both adult and pediatric patients, insults are important because they are common and yet so potentially avoidable. Clearly a critical challenge facing us is to develop patient monitoring systems and protocols that will lead to rapid detection and resolution of secondary insults. However, detection is not enough; we need also improved and clinically proven methods of treating secondary insults—further evidence is required. In Signorini et al.'s article (87), they conclude that the questions posed by such observational studies can only be answered definitively within the context of a randomized clinical trial. However, to design such a multicenter randomized clinical trial will require improved standards in the monitoring and analysis of secondary insult data. In Europe, improved standards for high-resolution collection and analysis of multicenter data from head-injured data is now being addressed by the *Brain-IT* group (90). The *Brain-IT* group (http://www.brainit.org) is an open consortium of clinicians and basic scientists working toward improving the infrastructure for conducting both observational and controlled trials of medical devices and patient management.

INTRACRANIAL PRESSURE ANALYSIS METHODS

Since the early 1990s, clinical monitoring of ICP has generally become part of the intensive care management of patients with brain injury. Several methods of analysis have been designed to extract pathophysiological information from the ICP and the corresponding ABP recordings. As noted, one particular computation used in clinical practice is the determination of mean CPP, the pressure across the brain. CPP is calculated as the difference between mean ABP and mean ICP and is based on the assumption that cerebral venous pressure is approximately equal to ICP. CPP is a useful parameter because it provides some insight as to whether the blood flow though brain capillaries is regulated. In the uninjured brain, cerebral blood flow is regulated to match the metabolic demand of the brain cells. Regulation of flow is primarily done by active dilation and constriction of cerebral arterioles in response to changes of CPP and/or biochemical vasoconstrictive or vasodilator agents that interact with the vascular endothelium. The steady-state relationship between cerebral blood flow and CPP is termed static pressure regulation and is illustrated by the graphical relationship between cerebral blood flow and CPP shown in Fig. 1.

In the steady state, cerebral blood flow is laminar and computed as the ratio of CPP to hemodynamic resistance, which is inversely proportional to the fourth power of the radius of the vessel. In the autoregulatory range, the arterial–arteriolar bed actively adjusts the resistance of its vessels by dilating when CPP decreases and constricting when CPP increases to maintain a relatively constant cerebral blood flow during changes in CPP. When CPP is below the lower limit of the autoregulatory range, vessels within the arterial–arteriolar bed tend to passively vasoconstrict. When CPP is above the upper limit of autoregulation, passive vasodilation occurs. One proposed intensive care therapeutic procedure designed to prevent secondary complications during recovery is CPP-oriented therapy (91). This therapy requires pressure autoregulation and the ability to manipulate CPP within the autoregulatory range (91). During intact pressure regulation, increases of

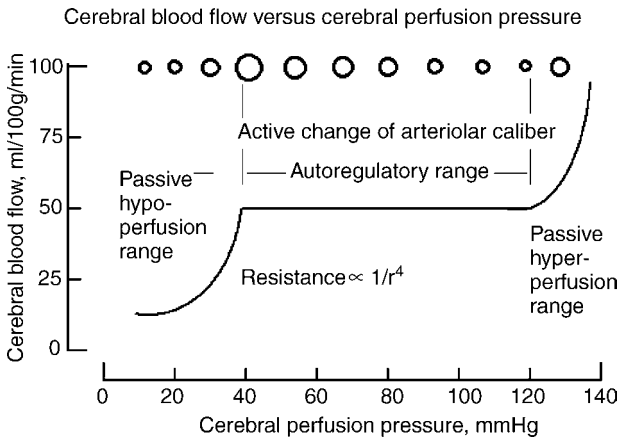


Figure 1. Illustration of static pressure regulation. Within the autoregulatory range, changes in arteriolar diameter, represented here as circles at the top of the graph, are primarily responsible for the regulation of cerebral blood flow. The arterial–arteriolar bed actively constricts with increasing CPP and dilates with decreasing CPP. Hypoperfusion occurs when CPP falls below the lower limit of regulation. As CPP falls below this limit, the vascompression of the arterioles and resistance increases markedly. The lower limit of regulation is not precisely known and varies with age, with 40–50 mm Hg for young pediatric cases (90) and 60–70 mm Hg for an adult (61,86).

CPP cause constriction of the arterial-arteriolar vascular bed and lowering of ICP by a reduction in cerebral blood volume. In addition, the resulting decrease of pre- and postcapillary pressure lessens fluid filtration and increases absorption, thus reducing the effects of edema. The application of CPP-oriented therapy when autoregulation has been lost may result in an imbalance of Starling forces at the capillaries, leading to increased net filtration and further brain injury by increased production of vasogenic edema.

In contrast to clinical cardiology, where the physiological mechanism underlying the dynamic features of the arterial pressure recording are fairly well understood, very little is known about the mechanism that underlies the shape of pulsations of intracranial pressure and how they are influenced by changes in CPP. Past studies have indicated that pulsations in the cerebrospinal fluid develop from either pulsations of the choroids plexus (92,93) or the arterial vasculature (94,95). Depending on the dilatory state of the cerebrovascular bed, ABP, ICP, and whether the autoregulation is intact, at times the primary component of the ICP pulse may be due to pulsations of venous or arterial origin (94–96). Although the origin of the ICP pulsation is not completely understood, possibly useful methodologies involving the pulsation have been developed. Over two decades ago, Avezatt and van Eijnhoven developed a procedure for distinguishing the occurrence of the loss of regulation of cerebral blood flow through numerical analysis of the pulsation of intracranial pressure (96,97) From laboratory studies, they found that within the autoregulatory range, the relationship between the mean amplitude of the pulsation of ICP increased linearly with mean ICP with a high slope value. At mean ICP above 30 mm Hg, the slope of this relationship decreased. They proposed that the flattening of the slope of this relationship was an indication of loss of autoregulation (97). A weakness of this technique is that it is dependent on heart rate (98). A low heart rate reflects an increase in the volume, of pulsatile cerebral blood inflow resulting from increased cardiac stroke volume, and the converse is likely for high heart rates. Thus, the variability in the amplitude of the ICP due to heart-rate variability can complicate the analysis (98). However, one modification of the analysis technique is to examine the correlation value between amplitude-pressure and mean ICP rather than the slope of the respective regression line (99). Positive values of this correlation value, which has been called RAP, indicate the ability of the vasculature to regulate cerebral blood flow, and negative values indicate poor cerebrovascular reserve and impaired cerebrovascular reactivity (99,100). Statistical analysis revealed that patients with a fatal outcome were associated with a negative RAP value (99). Most recently, the RAP index has been used to predict the achievable reduction in intracranial pressure a patient can obtain by the implementation of moderate hyperventilation (101). A modification of the mean amplitude of ICP and mean ICP characteristic has been proposed as a means of predicting which patients with idiopathic adult hydrocephalus will have a good outcome after shunting surgery (102). In this application, the amplitude of B-wave of ICP is used instead of the amplitude of the ICP pulsation (102).

In addition to relatively synchronous pulsations of ICP associated with the cardiac cycle, the ICP recording obtained during mechanical ventilation contains a low-frequency component at the rate of ventilation. Generally during mechanical ventilation, inhalation is produced by positive pressure and expiration occurs during zero pressure. Changes in pulmonary volume produce changes in intrathoracic pressure that mechanically modulate arterial and venous blood flow and produce a cyclic compression of the craniospinal sac. Evidence for this mechanical modulation by intrathoracic pressure can be observed in the spectra of an ICP recording obtained during mechanical ventilation. Because the pulsation in ICP associated with the cardiac cycle produced is quasi-periodic and the rate of ventilation is relatively constant, the spectra of the ICP pressure recording contains salient peaks at the cardiac frequency and its higher harmonics. In addition, each of the spectra associated with cardiac cycle has sidebands with a deviation at the ventilation frequency (103). Such a result is consistent with the premise that intrathoracic pressure changes mechanical modulated arterial and venous blood pressure and volume of the craniospinal sac (103).

Over the last few decades, several clinical and laboratory studies of the correlative relationship between ABP and ICP have been completed. Portnoy et al. reported that during normal vascular tone and intact regulation of cerebral blood flow, the ICP and ABP recordings do not look similar, but during maximum vasodilation of the arterial–arteriolar bed and impaired autoregulation induced by severe hypercapnia, these pressure recordings look remarkably similar (104–106). To numerically quantify the correlation between ABP and ICP, this group examined changes in the coherence function. Specifically, they found

that the frequency domain coherence function approached unity when the ICP and ABP recording became similar. Generally, their observations and the observations of others using a spectral analysis systems approach have suggested that the more similar the spectral components of the ICP recording are to those of the ABP recording, the more likely cerebral autoregulation is impaired (106–108). The physical process describing the ABP as the input to the craniospinal sac and the ICP as the corresponding output has been termed cerebrovascular pressure transmission (106). An observational clinical study determined that four types of frequency descriptions of the transmission characteristic could be identified. Two types were associated with high ICP and the other two with low ICP (47).

More recent studies using time-domain correlation analysis on the ICP and ABP pressure recordings have been completed. As correlation analysis of two signals in the time domain is analytically equivalent to coherence analysis in the frequency domain, it was not unexpected that studies on the normocapnic/hypercapnic piglet model found that as the ICP and ABP recordings became more similar, the maximum value of the correlation function approaches unity, the pial arterioles became more dilated, and cerebral blood flow increased (109,110). Consistent with clinical reports that indicate that unlike adult patients, brain-injured pediatric patients often demonstrate cerebral hyperemia and increased ICP (82,109,111), correlation analysis of pressure recordings obtained from pediatric patients have been found to often approach unity, indicating the occurrence of inappropriate vasodilation and cerebral hyperemia (112). Most recently, Czosnyka et al. have reported the clinical use of a pressure reactivity index (PRx). This index is a moving correlation coefficient between 40 consecutive samples of values for intracranial and arterial pressures averaged for a period of 5 s (113,114). They have concluded that when slow waves in the ABP and ICP recordings are present, the proposed clinical index provides a continuous index of cerebrovascular reactivity (114,115). In particular, the PRx was designed to evaluate the integrity of the cerebrovascular response and estimate cerebrovascular autoregulatory reserve reactivity (114,115). The hypothesis is that because of the sluggish nature of the cerebrovascular system, naturally occurring slow-varying oscillations of ABP can be used to evaluate the autoregulatory reserve reactivity (114,115). Unlike the coherence and correlation indices described above, this index cannot be related to a linear system model. By employing averaging over a 5 s interval, most of the frequency changes above 0.2 Hz in the ABP and ICP recordings are filtered out. In addition, Nyquist's sampling theorem dictates that the highest frequency that can be represented by a signal sampled every 5 s is 0.1 Hz or six oscillations per minute. As a result, aliasing occurs, and the dynamical system relationship between ABP and ICP is not precisely defined by this index. Nevertheless, the PRx has been found to be a very useful tool. Clinical observations demonstrate that the PRx is high both during the occurrence of plateau waves and during refractory ICP hypertension (115). In addition, the PRx has been used to guide proposed therapies (116) and while variable seems to provide a reliable index of autoregulation (117).

Most recently, the regressive relationship between mean ICP and mean CPP has been used to assess whether autoregulation of cerebral blood flow is intact. In these studies, ABP is pharmalogically elevated and the regressive relationship between ICP and CPP during the test period is obtained (118). If pressure regulation of cerebral blood flow is intact, then increases of CPP will cause vasoconstriction, a decrease of ICP, and the regressive relationship should have a negative slope parameter. A marked positive slope parameter of this regressive relationship is an indication of passive pressure regulation; increases of CPP cause dilation and increased ICP (119).

ANALYSIS METHODS BASED ON MODELS OF ICP

Models of ICP dynamics are based on the modified Monroe–Kellie doctrine (see the Physiology section), which assumes that the total volume of intracranial substance, tissue, blood, and CSF, is constant. Initial mathematical models described the relationship between the formation of CSF and absorption of CSF in equilibrium. Specifically, in these models, laminar flow of CSF is assumed during steady-state conditions (120–123) and the parameters of the mathematical model are presented as an electrical circuit (Fig. 2).

Using this circuit and manipulating the volume of CSF by bolus either by withdrawal, injection, or constant infusion, it is possible to estimate R_o, C, the volume–pressure response, and the PVI. The latter two parameters have been discussed previously in the Physiology section. All parameters have been used to guide the management of hydrocephalous and traumatic brain injury.

Higher order mathematical models of intracranial hydrodynamics that incorporate the arterial and venous blood volume compartments into the analog electric circuit model have been developed to simulate the pulsatility, which is present in the ICP recording. To account for the reciprocal relationship between cerebral blood volume and volume of CSF, Agarwal et al. (124) constructed a model that connected vascular compliance to intracranial compliance in a series configuration. This modeling effort was further developed in detail by Ursino's proposed fourth-order analog circuit model of overall human intracranial

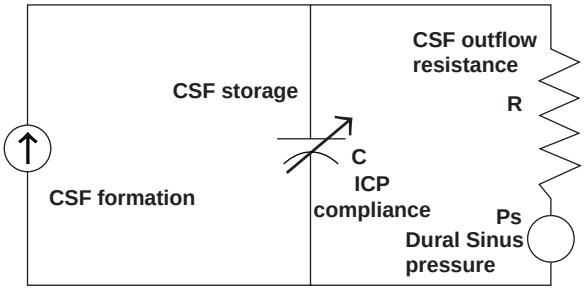


Figure 2. Electric circuit analog model of CSF system. In this model, current represents flow of CSF fluid, and voltage represents pressure. The capacitance C, represents intracranial compliance, and the resistance R represents resistance of CSF fluid flow into the venous system.

hydrodynamics (125,126). A feature of this model is that the arterial–arteriolar vascular bed consists of vasomotor regulating resistance and compliance elements. Dilatory and constrictive responses are primarily simulated by a corresponding decrease or increase of vascular resistance. This initial model has been modified by synchronously modulating the terminal venous bed resistance to account for cyclic variation of ICP produced by positive pressure ventilation (127). The modified model demonstrates that both the depth of modulation and the cerebrovascular venous terminal bed resistance seem to be progressively reduced with increasing levels of vasodilation induced by increasing the levels of partial pressure of arterial blood carbon dioxide. Numerical modeling of cerebrovascular pressure transmission, the relationship between ABP and CPP, has been used to assess changes in the modes of pressure transmission before and after injury. Specifically, a proposed third-order model of ICP dynamics (128) was used to define the mathematical structure to construct a numerical identification model of cerebrovascular pressure transmission from laboratory pressure recordings (112). Consistent with active vasoconstriction before brain injury, during intact regulation of cerebral blood flow, the highest modal frequency of cerebrovascular pressure transmission decreased with increasing CPP. Conversely, consistent with passive vasodilation with loss of autoregulation induced by fluid percussion injury, the highest modal frequency demonstrated a direct relationship with increasing CPP (112).

SUMMARY

Although the subject of intracranial pressure dates back to over 200 ago, both the invasive monitoring procedure required to measure ICP with its accompanying risk of infection and the nonunique nature of the extravascular pressure measurement have retarded the development of knowledge in this area. As a result, the intent of this article is to provide the reader with a broad perspective of ICP monitoring by providing background material on (1) the physiological and anatomical characteristics of the craniospinal, (2) a history of instrumentation techniques including the most recent advances, (3) a brief review of significant clinical ICP monitoring studies with an emphasis on severe head injury, (4) a review of analysis methods, and (5) models of ICP dynamics related to analysis.

BIBLIOGRAPHY

1. Stern WE. Intracranial fluid dynamics. The relationship of intracranial pressure to the Monro-Kellie doctrine and the reliability of pressure assessment. J Royal College Surgeons Edinburgh 1963;9:18–36.
2. Langfitt TW. Increased intracranial pressure. Clin Neurosurg 1969;6:436–471.
3. Monro A. Observations on the Structure and Function of the Nervous System. Edinburgh: Creech and Johnston; 1783.
4. Kellie G. An account of the appearances observed in the dissection of two of three individuals presumed to have perished in the storm of the third and whose bodies were discovered in the vicinity of Leith on the morning of the 4th, November 1821, with some reflections on the pathology of the brain. Trans Med Chir Soc Edinb 1824;1:84–169.
5. Weed LH, McKibben PS. Pressure changes in cerebrospinal fluid following intravenous injection of solutions of various concentrations. Am J Physiol 1919;48:512–530.
6. Weed LH. Some limitations of the Monro–Kellie hypothesis. Arch Surg 1929;18:1049–1068.
7. Kocher T. Hirnerschutterung, Hirndruck und chirurgische Eingriffe bein Hirnerkrankungen Nothnagel: Spezielle Pathologie und Therapie: 1901.
8. Duret H. Etudes experimentales et cliniques sur les traumatismes cerebrau. Paris: 1978.
9. Cushing H. Some experimental and clinical observations concerning states of increased intracranial tension. Am J Med Sci 1902;124:375–400.
10. Cushing H. The blood pressure reaction of acute cerebral compression, illustrated by cases of intracranial haemorrhage. Am J Med Sci 1903;125:1017–1044.
11. Cushing H. Concerning a definite regulatory mechanism of the vasomotor center which controls blood pressure during cerebral compression. Johns Hopk Hosp Bull 1901;12:290–292.
12. Jennet WB. Experimental brain compression. Arch Neurol 1961;4:599–607.
13. Johnston IH, Rowan JO. Raised intracranial pressure and cerebral blood flow: 3. Venous outflow tract pressures and vascular resistances in experimental intracranial hypertension. J Neurol Neurosurg Psych 1974;37:392–402.
14. Browder J, Meyers R. Observations on behaviour of the systemic blood pressure, pulse and spinal fluid pressure following craniocerebral injury. Am J Surg 1936;31:403–427.
15. Smyth CE, Henderson WR. Observations on the cerebrospinal fluid pressure on simultaneous ventricular and lumbar punctures. J Neurol Psychiat 1938;1:226–237.
16. Evans JP, Espey FF, Kristoff FV, Kimball FD, Ryder HW. Experimental and clinical observations on rising intracranial pressure. Arch Surg 1951;63:107–114.
17. Marmarou A, Shulman K, LaMorgese J. Compartmental analysis of compliance and outflow resistance of the cerebrospinal fluid system. J Neurosurg 1975;43:523–534.
18. Miller JD, Stanek AE, Langfitt TW. Concepts of cerebral perfusion pressure and vascular compression during intracranial hypertension. In: Meyer JS, Schade JP, editors. Progress in Brain Research. vol 35: Cerebral Blood Flow. Amsterdam: Elsevier; 1972. pp 411–432.
19. Miller JD, Garibi J, Pickard JD. Induced changes of cerebrospinal fluid volume: Effects during continuous monitoring of ventricular fluid pressure. Arch Neurol 1973;28: 265–269.
20. Miller JD, Pickard JD. Intracranial volume/pressure studies in patients with head injury. Injury 1974;5:265–268.
21. Miller JD, Leech PJ, Pickard JD. Volume pressure response in various experimental and clinical conditions. In: Lundberg N, Ponten U, Brock M, editors. Berlin: Springer; Intracranial Pressure II; 1975. pp 97–99.
22. Miller JD. Volume and pressure in the craniospinal axis. Clin Neurosurg 1975;22:76–105.
23. Hase U, Reulen HJ, Meinig G, Schrmann K. The influence of the decompressive operation on the intracranial pressure and the pressure volume relation in patients with severe head injuries. Acta Neurochir 1978;45:1–13.
24. Lundberg N. Continuous recording and control of ventricular fluid pressure in neurosurgical practice. Acta Psychiat Neurol Scand 1960;36:149.
25. Baylis WM, Hill L, Gulland GL. On intracranial pressure and the cerebral circulation. J Physiol 1897;18:334–362.
26. Rosner MJ, Becker DP. ICP monitoring: Complications and associated factors. Clin Neurosurg 1996;23:494–519.

27. Maas A, Dearden M, Teasdale GM, Braakman R, Cohodan F, Iannotti F, Karimi A, Lapirre F, Murray G, Ohman J, Persson L, Servadei F, Stocchetti N, Unterburg A. EBIC-Guidelines for management of severe head injury in adults. Acta Neurochir 1997;139:286–294.
28. Naryan RK. et al. Guidelines for the management of severe head injury. J Neurotrauma 1997;13:639–734.
29. Gabe IT. Cardiovascular fluid dynamics. Acta Physiol Scand 1972;19:306–322.
30. Piper IR, Dearden NM, Miller JD. Methodology of spectral analysis of the intracranial pressure waveform in a head injury intensive care unit. In: Hoff JT, Betz AL, editors. Intracranial Pressure VII. Berlin: Springer-Verlag; 1989. pp 668–671.
31. Allan MWB. Measurement of arterial pressure using catheter–transducer systems. Br J Anaesth 1988;60:413–418.
32. Czosnyka M, Czosnyka Z, Pickard J. Laboratory testing of the Spiegelberg brain pressure monitor: A technical report. J Neurol Neurosurg Psych 1997;63:732–735.
33. Ostrup RC, Luerssen TG, Marshall LF. Continuous monitoring of intracranial pressure with a miniturized fibreoptic device. J Neurosurg 1987;67:206–209.
34. Marmarou A, Tsuji O, Dunbar JG. Experimental evaluation of a new solid state ICP monitor. In: Nagai H, Kemiya K, Ishiii S, editors. Intracranial Pressure IX. New York: Springer-Verlag; 1994. pp 15–19.
35. Morgalla MH, Cuno M, Mettenleiter H, Will BE, Krasznai L, Skalej M, Bitzer M, Grote EH. ICP monitoring with a reusable transducer: experimental and clinical evaluation of the Gaeltec ICT/b pressure probe. Acta Neurochir (Wien) 1997;139(6): 569–573.
36. Holzschuh M, Woertgen C, Metz C, Brawanski A. Clinical evaluation of the InnerSpace fibreoptic intracranial pressure monitoring device. Brain 1988;12:191–198.
37. Betsch HM, Aschoff A. Measurement artifacts in Gaeltec intracrnial pressure monitors due to radio waves from personal beeper systems. Anaesthesiol Intensivmed Notfallmed Schmerzther 1992;27(1):51–52.
38. Munch E, Weigel R, Schmiedek P, Schurer L. The Camino intracranial pressure device in clinical practice: Reliability, handling characteristics and complications. Acta Neurochir 1998;140:1113–1119.
39. Chambers IR, Kane PJ, Choksey MS, Mendelow AD. An evaluation of the Camino ventricular bolt system in clinical practice. Neurosurgery 1993;33:866–868.
40. Weinstable C, Richling B, Plainer B, Czech T, Spiss CK. Comparative analysis between epidural (Gaeltec) and subdural (Camino) intracranial pressure probes. J Clin Monit 1992;8:116–120.
41. Piper IR, Miller JD. The evaluation of the waveform analysis capability of a new strain-gauge intracranial pressure microsensor. Neurosurgery 1995;36:1142–1145.
42. Signorini DF, Shad A, Piper IR, Statham PFX. A clinical evaluation of the Codman microsensor for intracranial pressure monitoring. Br J Neurosurgery 1988;12:223–227.
43. Fernades HM, Bingham K, Chambers IR, Mendelow AD. Clinical evaluation of the Codman microsensor intracranial pressure monitoring system. Acta Neurochir Suppl (Wien) 1998;71:44–66.
44. Mendelow AD, Rowan JO, Murray L, Kerr AE. A clinical comparison of subdural screw pressure measurements with ventricular pressure. J Neurosurg 1983;58:45–50.
45. Piper IR, Barnes A, Smith DH, Dunn L. The Camino intracranial pressure sensor: Is it optimal technology? An internal audit with a review of current intracranial pressure monitoring technologies. Neurosurgery 2001;49: 1158–1165.
46. Citerio G, Piper I, Cormio M, Galli D, Cazzaniga S, Enblad P, Nilsson P, Contant C, Chambers I, on behalf of the BrainIT Group, Bench test assessment of the new Raumedic.
47. Piper IR, Miller JD, Dearden NM, Leggate JR, Robertson I. Systems analysis of cerebrovascular pressure transmission: An observational study in head injured patients. J Neurosurg 1993;73:871–880.
48. Becker DP, et al. The outcome from severe head injury with early diagnosis and intensive management. J Neurosurg 1977;47:491–502.
49. Marshall LF, Smith RW, Shapiro HM. The outcome with aggressive treatment in severe head injuries. Part 1: The significance of intracranial pressure monitoring. J Neurosurg 1979;50:20–25.
50. Miller JD, et al. Significance of intracranial hypertension in severe head injury. J Neurosurg 1977;47:503–516.
51. Pitts LH, Kaktis JV, Juster R, Heilbrun D. ICP and outcome in patients with severe head injury. In: Shulman K, Marmarou A, Miller JD, Becker DP, Hochwald GM, Brock M, editors. Intracranial Pressure IV. Berlin: Springer; 1980. pp 5–9.
52. Stuart G, Merry G, Smith JA, Yelland JDM. Severe head injury managed without intracranial pressure monitoring. J Neurosurg 1983;59:601–605.
53. Miller JD, et al. Further experience in the management of severe head injury. J Neurosurg 1981;54:289–299.
54. Alberico AM, Ward JD, Choi SC, Marmarou A, Young H. Outcome after severe head injury: Relationship to mass lesions, diffuse injury and ICP course in pediatric and adult patients. J Neurosurg 1987;67:648–656.
55. Choi SC, Muizelaar JP, Barnes TY, Marmarou A, Brooks DM, Young HF. Prediction tree for severely head injured patients. J Neurosurg 1991;75:251–255.
56. Vollmer DG, Torner JC, Jane J, et al. Age and outcome following traumatic coma: Why do older patients fare worse? J Neurosurg 1991;75:s37–s49.
57. O'Sullivan MG, Statham PF, Jones PA, et al. Role of intracranial pressure monitoring in severely head–injured patients without signs of intracranial hyperension on initial computerised tomography. J Neurosurg 1994;80:46–50.
58. Narayan RK, Greenberg RP, Miller JD, Enas GG, Choi SC, Kishore PRS.
59. Saul TG, Ducker TB. Effect of intracranial pressure monitoring and aggressive treatment on mortality in severe head injury. J Neurosurg 1982;56:498–503.
60. Marmarou A, et al. Impact of ICP instability on outcome in patients with severe head trauma. J Neurosurg 1991;75:s59–s66.
61. Jones PA, Andrews PJD, Midgley S, Anderson SI, Piper IR, Tocher JL, Housley AM, Corrie JA, Slattery J, Dearden NM, Miller JD. Measuring the burden of secondary insults in head–injured patients during intensive care. J Neurosurg Anaesth. 6:4–14.
62. Lewelt W, Jenkins LW, Miller JD. Autoregulation of cerebvral blood flow after experimental fluid percussion injury of the brain. J Neurosurg 1980;53:500–511.
63. Povlishock JT, Kontos HA. Continuing axonal and vascular change following experimental brain trauma. Central Nervous System Trauma 1985;2:285–298.
64. Nordstrom CH, et al. Cerebral blood flow, vasoreactivity and oxygen consumption during barbiturate therapy in severe traumatic brain lesions. J Neurosurg 1988;68:424–431.
65. Miller JD, Adams JH. The pathophysiology of raised intracranial pressure. In: Adams JH, Duchen LW, editors. Greenfield's Neuropathology. 5th ed. London: Arnold; 1992: 69–105.
66. DeSalles AF, Muizelaar JP, Young HF. Hyperglycemia, cerebrospinal fluid lactic acidosis and cerebral blood flow in severly head–injured patients. Neurosurgery 1987;21:45–50.

67. Jaggi LJ, Obrist WD, Gennarelli TA, Langfitt TW. Relationship of early cerebral blood flow and metabolism to outcome in acute head injury. J Neurosurg 1990;72:176–182.
68. Muizelaar JP, Ward JD, Marmarou A, Newton PG, Wachi A. Cerebral blood flow and metabolism in severely head–injured children. Part 2: Autoregulation. J Neurosurg 1989;71:72–76.
69. Muizelaar JP, Marmarou A, DeSalles AAF, Ward JD, Zimmerman RS, Zhongchao L, Choi SC, Young HF. Cerebral blood flow and metabolism in severely head–injured children. Part 1: Relationship with GCS score, outcome, ICP and PVI. J Neurosurg 1989;71:63–71.
70. Uzzell BP, Obrist WD, Dolinskas CA, Langfitt TW. Relationship of acute CBF and ICP findings to neuropsychological outcome in severe head injury. J Neurosurg 1986;65:630–635.
71. Rose J, Valtonen S, Jennett B. Avoidable factors contributing to death after head injury. Brit Med J 1977;2:615–618.
72. Gentleman D, Jennett B. Hazards of interhospital transfer of comatose head injured patients. Lancet 1981;ii:835–855.
73. Graham DI, Hume Adams J, Doyle D. Ischemic brain damage in fatal non–missile head injuries. J Neurol Sci 1978;39:213–234.
74. Graham DI, Ford I, Adams JH, Doyle D, Teasdale GM, Lawrence AE, Mclellan DR. Ischemic brain damage is still common fatal non–missile head injury. 1989.
75. Miller JD, Garibi J. Intracranial volume/pressure relationships during continuous monitoring of ventricular fluid pressure. In: Brock M, Dietz H, editors. Intracranial Pressure. Berlin: Springer; 1972. 270–274.
76. Johnston IH, Rowan JO. Raised intracranial pressure and cerebral blood flow: 3. Venous outflow tract pressures and vascular resistances in experimental intracranial hypertension. J Neurol Neurosurg Psychi 1974;37:392–402.
77. Yada K, Nakagawa Y, Tsuru M. Circulatory disturbance of the venous system during experimental intracranial hypertension. J Neurosurg 1973;39:723–729.
78. Nakagawa Y, Tsura M, Yada K. Site and mechanism for compression of the venous system during experimental intracranial hypertension. J Neurosurg 1974;41:427–434.
79. Harper AM. Autoregulation of cerebral blood flow: influence of the arterial pressure on blood flow through the cerebral cortex. J Neurol Neurosurg Psych 1966;29:398–403.
80. Muizelaar JP, Becker DP. Induced hypertension for the treatment of cerebral ischemia after subarachnoid hemorrhage. Direct effect on cerebral blood flow. Surg Neurol 1986;25:317–325.
81. Obrist WD, et al. Cerebral blood flow and metabolism in comatose patients with acute head injury. Relationship to intracranial hypertension. J Neurosurg 1984;61:241–253.
82. Bouma GJ, Muizelaar JP. Relationship between cardiac output and cerebral blood flow in patients with intact and with impaired autoregulation. J Neurosurg 1990;73:368–374.
83. Chan KH, Miller JD, Piper IR. Cerebral blood flow at constant cerebral perfusion pressure but changing arterial and intracranial pressure: Relationship to autoregulation. J Neurosurg Anaesth 1992;4:188–193.
84. Piper IR, Lawson A, Dearden NM, Miller JD. Computerised data collection: a microcomputer data collection system in head injury intensive care. Br J Intensive Care 1991;1:73–78.
85. Chambers IR, Treadwell L, Mendelow AD. The cause and incidence of secondary insults in severely head–injured adults and children. 2000.
86. Signorini DF, Andrews PJD, Jones PAJ, Wardlaw JM, Miller JD. Prediciting survival using simple clinical variables: A case study in traumatic brain injury. J Neurol Neurosurg Psych 1999;66:20–25.
87. Signorini DF, Andrews PJD, Jones PAJ, Wardlaw JM, Miller JD. Adding insult to injury: The prognostic value of early secondary insults for survival after traumatic brain injury. J Neurol Neurosurg Psychi 1999;66:26–31.
88. Robertson CS, Valadka AB, Hannay HJ, Contant CF, Gopinath SP, Cormio M, Uzura M, Grossman RG. Prevention of secondary ischemic insults after severe head injury. Crit Care Med 1999;27(10):2086–2095.
89. Chambers IR, Jones PA, Lo TY, Forsyth RJ, Fulton B, Andrews PJ, Mendelow AD, Minns RA. critical threshold of intracranial pressure and cerebral perfusion pressure related to age in paediatric head injury. 2005.
90. Piper I, et al. The BrainIT Group: Concept and core dataset definition. Acta Neurochir 2003;145:615–629.
91. Rosner MJ, Rosner SD, Johnson AH. Cerebral perfusion pressure: management protocol and clinical results. J Neurosurg 1995;83:949–962.
92. Bearing EA. Choroid plexus and arterial pulsation of cerebrospinal fluid. Arch Neurol Psych 1955;73:165–172.
93. Hamit HF, Beall AC, DeBakey ME. Hemodynamic influences upon brain and cerebrospinal fluid pulsations and pressures. J Trauma 1965;5:174–184.
94. Dardenne G, Dereymaeker A, Lacheron JM. cerebrospinal fluid pressure and pulsatility. An experimental study of circulatory and respiratory influences in normal and hydrocephalic dogs. 1969.
95. Hamer J, Alberti E, Hoyer S, Wiedemann K. Influence of systemic and cerebral vascular factors on the cerebrospinal fluid pulse waves. 1977.
96. Avezatt CJJ, van Eijndhoven JHM. Clinical observations on the relationship between cerebrospinal fluid pulse pressure and intracranial pressure. In: Cerebrospinal Fluid Pulse Pressure and Cranio–spinal Dynamics: A Theoretical, Clinical and Experimental Study. Thesis. Erasmus Univ., Rotterdam, 1984.
97. Avezatt CJJ, van Eijndhoven JHM. Cerebrospinal fluid pulse pressure and intracranial volume–pressure relationships. J Neurol Neurosurg Psych 1979;42:687–700.
98. Daley ML, Gallo A, Mauch W. Analysis of the intracranial pressure pulsation associated with the cardiac cycle. Innovation Et Technology En Biologie Et Medicine 1986;7:537–544.
99. Czosynka M, Guazzo E, Whitehouse M, Smielewski P, Czosynka Z, Kirkpatric P, Piechnik S, Pickard JD. Significance of intracranial pressure waveform analysis after head–injury. Acta Neurochir (Wien) 1996;138(5):531–541.
100. Balestreri M, Czosynka M, Steiner LA, Schmidt E, Smielewski P, Matta B, Pickard JD. Intracranial hypertension: What additional information can be derived from ICP waveform after head injury? Acta Neurochir 2004;146:131–141.
101. Steiner IA, Balestreri M, Johnston AJ, Coles JP, Smielewski P, Pickard JD, Menon DK, Czosnyka M. Predicting the response of intracranial pressure to moderate hyperventilation Acta Neurochirurgica (online) 2005.
102. Lenfeldt N, Anderson N, Agren–Wilsson A, Bergenheim AT, Koskinen LO, Eklund A, Malm J. Cerebrospinal fluid pulse pressure method: A possible substitute for the examination of B waves. J Neurosurg 2004;101(6):944–950.
103. Daley ML, Pasley R, Connolly M, Angel J, Timmons S, Stidham G, Leffler C. Spectral characteristics of B–waves and other low–frequency activity. Acta Neurochurirgica 2002;81:147–150.
104. Chopp M, Portnoy H. System analysis of intracranial pressure. J Neurosurg 1980;53:516–527.
105. Portnoy HD, Chopp M. Cerebrospinal fluid pulse wave form analysis during hypercapnia and hypoxia. Neurosurgery 1981;9:14–27.
106. Portnoy HD, Chopp M, Branch C, Shannon MD. Cerebrospinal fluid pulse waveform as an indicator of cerebral autoregulation. J Neurosurg 1982;56:666–678.

107. Piper IR, Chan KH, Whittle IR, Miller JD. An experimental study of cerebrovascular resistance, pressure transmission, and craniospinal compliance. Neurosurgery 1993;32:805–816.
108. Nichols JS, Beel JA, Munro LG. Detection of impaired cerebral autoregulation using spectral analysis of intracranial pressure waves. J Neurotrauma 1996;13:439–456.
109. Bruce DA, et al. Diffuse cerebral swelling following head injury in children: The syndrome of "malignant brain edema". J Neurosurg 1981;54:170–178.
110. Daley ML, Pasupathy H, Griffith M, Robertson JT, Leffler C. Evaluation of autoregulation of cerebral blood flow by correlation of arterial and intracranial pressure signals. IEEE Trans Biomed Eng 1995;42:420–424.
111. Bruce DA, et al. Pathophysiology, treatment and outcome following severe head injury in children. Child's Brain 1979;5:174–191.
112. Daley ML, Patterson S, Marmarou A, Leffler CW, Stidham G. Pediatric traumatic brain injury: Correlation of intracranial and arterial pressure signals, Proc. 18th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society, Amsterdam, Netherlands, Nov. 1996.
113. Czosnyka M, Smielewski P, Kirkpatrick P, Laing R, Menon D, Pickard J. Continuous assessment of the cerebral vasomotor reactivity in head injury. Neurosurgery 1997;41:11–19.
114. Czosnyka M, Smielewski P, Kirkpatrick P, Laing R, Menon D, Pickard J. Continuous assessment of the cerebral vasomotor reactivity in head injury. Acta Neurochirugica 1998;71:74–77.
115. Czosnyka M, Smielewski P, Piechnik S, Schmidt E, Al–Rawi PG, Kirkpatrick PJ, Pickard JD. Hemodynamic characterization of intracranial pressure plateau waves in head–injured patients. J Neurosurg 1999;92:11–19.
116. Steiner LA, Czosynka M, Piechnik SK, Smielewski P, Chatfield D, Menon DK, Pickard JD. Continuous monitoring of cerebrovascular pressure reactivity allows determination of optimal cerebral perfusion pressure in patients with traumatic brain injury. Critical Care Med 2002;30: 733–738.
117. Steiner LA, Coles JP, Johnston AJ, Chatfield DA, Smielewske P, Fryer TD, Aigvirhio FI, Clark JC, Pickard JD, Menon DK, Czosynka M. Assessment of cerebrovascular autoregulation in head–injured patients. Stroke 2003;34: 2404–2409.
118. Oertel M, Kelly DF, Lee JH, Glenn TC, Vespa M, Martin NA. Is CPP therapy beneficial for all patients with high ICP. Acta Neurochir 2002;81:67–68.
119. Howells T, Elf K, Jones PA, Ronne–Engstrom E, Piper I, Nilsson P, Andrews P, Enbald P. Pressure reactivity as a guide in the treatment of cerebral perfusion pressure in patients with brain trauma. J Neurosurg 2005;102(2):311–317.
120. Marmarou A. A theoretical and experimental evaluation of the cerebrospinal flid system. Ph.D. Dissertation, Drexel University, 1973.
121. Marmarou A, Shulman K, LaMorgese J. Compartmental analysis of compliance and outflow resistance of the cerebrospinal fluid system. J Neurosurg 1975;43:523–534.
122. Marmarou A, Shulman K, Rosende RM. A nonlinear analysis of the cerebrospinal fluid system and intracranial pressure dynamics. J Neurosurg 1978;48(3):332–344.
123. Ekstedt J. CSF hydrodynamic studies in man. 1. Method of constant pressure CSF infusion. J Neurol Neurosurg Psych 1977;40(2):105–119.
124. Agarwal GC, Berman BM, Stark L. A lumped parameter model of the cerebrospinal fluid system. IEEE Trans Biomed Eng 1969;16:45–53.
125. Ursino M. A mathematical study of human intracranial hydrodynamics. Part 1—the cerebrospinal fluid pulse pressure. Ann Biomed Eng 1988;16:379–401.
126. Ursino M. A mathematical study of human intracranial hydrodynamics. Part 2—Simulation of clinical tests. Ann Biomed Eng 1988;16:403–416.
127. Pasley RL, Leffler CW, Daley ML. Modeling modulation of intracranial pressure by variatrion of cerebral venous resistance induced by ventilation. Ann Biomed Eng 2003;31: 1238–1245.
128. Czosnyka M, Piechnik S, Richards HK, Kirkpatrick P, Smielewski P, Pickard JD. Contribution of mathematical modeling to the interpretation of bedside tests of cerebrovascular autoregulation. J Neurol Neurosurg Psych 1997;63: 721–731.

See also BIOTELEMETRY; HYDROCEPHALUS, TOOLS FOR DIAGNOSIS AND TREATMENT OF; NEONATAL MONITORING; NEUROLOGICAL MONITORS.

MONITORING, NEONATAL.

See NEONATAL MONITORING.

MONITORING, UMBILICAL ARTERY AND VEIN

AHMAD ELSHARYDAH
HAIBO WANG
RANDALL C. CORK
Louisiana State University
Health Center
Department of Anesthesiology
Shreveport, Louisiana

INTRODUCTION

In the neonatal intensive care unit (NICU), monitoring is an integral part of patient care. The primary goal of monitoring is to ensure that early and appropriate intervention can be initiated before to the onset of complications. Monitoring is also a means by which the effect of interventions and therapies may be recorded, evaluated, and controlled. The NICU staff have to deal with a full range of conditions that can arise in the preterm or critically ill neonate, including hypoxemia, hypoglycemia, hypotension, acidosis, and other serious problems. This has led to the evolution and development of several monitors and different sensor-based technologies for use in NICU monitoring including umbilical vessel monitoring. These sensors may provide more accurate and reliable monitoring of neonatal physiological and biochemical changes with a rapid response time. The umbilical vessels may be directly accessed in the first few days of life. An umbilical artery catheter (UAC) may be used for blood pressure monitoring, blood sampling, and fluid or drug infusion. An umbilical vein catheter (UVC) may be used for central venous pressure monitoring, blood sampling, and fluid or drug infusion. Different types of commercially available umbilical catheters are used for these purposes. These catheters differ in their length, size, number of ports, and their material (such as silicone and polyurethane). Blood pressure (BP) monitoring is an important part of neonatal intensive care both for the acutely ill and the convalescing neonate. The most accurate method of measuring BP is by direct intra-arterial recordings, which

usually use an umbilical catheter to access the umbilical artery. As blood gas measurement methods and monitors have progressed in adult critical medicine, most of the new techniques and sensors have been used in the NICU by using umbilical artery catheterization. This article addresses the potential benefits of umbilical vessel catheters and associated monitoring devices. It sheds light on the catheters and monitors available on the market and explains the complications and the risks of these catheters. Furthermore, this article looks at the direction of this technology in the future, and it tries to stimulate development of new technology for use in the monitoring of critically ill newborn infants (1–3).

Historical Aspects

In 1946, Louis K. Diamond, a pediatrician from Boston, and F. H. Allen, Jr. developed a technique that allowed blood transfusion to take place through the infant's umbilical cord vein. Regular transfusions were difficult because of the small size of blood vessels in newborns, and there was a further complication due to the use of steel needles and rubber catheters. Diamond used plastic tubing on the umbilical vein, which was larger than average and remained open for several days after birth (4). By the 1960s, electronic monitors came into use and blood gases began to be measured. By the 1970s, the use of umbilical catheters and arterial pressure transducers was routine (5). The first organized NICU opened its doors at Yale-New Haven Hospital in 1960. The first successful use of extracorporeal membrane oxygenation (ECMO) was in 1975. ECMO eventually reduced infant mortality from 80% to 25% for the critically ill infants with acute reversible respiratory and cardiac failure unresponsive to conventional therapy (6).

Anatomical and Physiological Aspects

The umbilical cord is a cordlike structure about 56 cm long, extending from the abdominal wall of the fetus to the placenta. Its chief function is to carry nutrients and oxygen (O_2) from the placenta to the fetus and return waste products and carbon dioxide (CO_2) to the placenta from the fetus. It consists of a continuation of the membrane covering the fetus and encloses a mucoid jelly (Wharton's jelly) with one vein and two arteries (7). Examination of the umbilical cord (after cut) normally reveals two umbilical arteries (UA) and one umbilical vein (UV) (Fig. 1). At skin level, the UV is usually in the 12 o'clock position and has a thinner wall and wider lumen than do the UAs (2). Before birth, blood from the placenta, about 80% saturated with O_2, returns to the fetus by way of the UV. On approaching the liver, most blood flows through the ductus venous directly into the inferior vena cava (IVC), short-circuiting the liver. A smaller amount enters the liver sinusoids and mixes with blood from the portal circulation. After a short course in the IVC, it mixes with deoxygenated blood returning from the lower limbs before it enters the right atrium. The blood leaves the heart to the descending aorta, where it flows toward the placenta by way of the two UAs (7). The O_2 saturation in the umbilical arteries is approximately 58%. Changes in the vascular system at birth are caused by

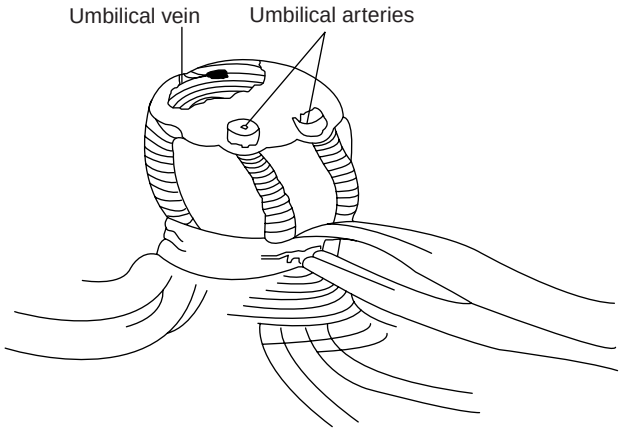


Figure 1. Umbilical cord after it was cut.

cessation of placental blood flow and the beginning of respiration. These changes are summarized in closure of the umbilical arteries, closure of the umbilical vein and ductus venous, significant reduction in the pulmonary vascular resistance and right ventricle and right atrium pressures, and significant increase in the systemic vascular resistance, left ventricle, and left atrium pressures (8). After birth, the blood volume of the neonate is about 300 mL, the cardiac output averages 500 mL/min, and the arterial blood pressure is about 70/50 during the first day, which increases slowly over the next several months. Arterial blood pressure of the neonate has been best correlated with birth weight. Moreover, systolic and diastolic pressures in the neonate are significantly correlated to blood pressure levels in the mother (9).

Indications and Contra-Indications for Umbilical Artery Catheterization

The primary indications for umbilical artery catheterization include frequent or continuous blood gas measurements, continuous monitoring of arterial blood pressure, and resuscitation (umbilical venous line may be the first choice) (10). Secondary indications include infusion of maintenance glucose-electrolyte solutions or medications, exchange transfusions, angiography, and a port for frequent blood sampling, especially in a very low-birth-weight neonate. These catheters should stay in place only as long as a primary indication exists, with the exception of the very low-birth-weight neonate who may need it for vital infusions and frequent blood sampling. Contraindications include evidence of local vascular compromise in lower limbs or buttock areas, peritonitis, necrotizing enterocolitis, omphalitis, gastroischisis, and omphalocele.

Indications and Contra-Indications for Umbilical Vein Catheterization

The most frequent indications for umbilical venous catheterization include emergency medication administration, exchange transfusion, and partial exchange transfusion (10). This catheter is also used for frequent blood sampling and central venous pressure monitoring. Contraindications for the this catheter include routine fluid infusion

Figure 2. Commercially available umbilical catheters: (a) single-lumen catheter and (b) dual-lumen catheter.

(relative contraindication), omphalitis, omphalocele, gastroischisis, necrotizing enterocolitis, peritonitis, and extrophy of the bladder.

Description of Available Catheters (Design and Material)

Catheters must be made of nontoxic materials that are the least injurious to the vascular intima and least likely to cause thromboses and early atherosclerotic lesions (11). Commercially available umbilical catheters are usually latex-free, made of silicon or high-quality aliphatic polyurethane elastomer (Tecoflex). Silicone catheters are soft, not-irritating, usually not reactive to body tissues and body fluids, not supportive of bacterial growth, and less problematic with blood clotting. Tecoflex is an advanced medical formulation of polyurethane. Its physical characteristics are very close to silicone; however, it is slightly stiffer during insertion, which makes it easier to insert and better to conduct arterial pressure. Tecoflex is thermosensitive, softens at body temperature (12), and significantly reduces the trauma to the vascular intima**.** These catheters have a rounded tip, which makes them less likely to perforate through the vessel during insertion. They have depth markings at every centimeter for more accurate placement and encased radiopaque stripes to confirm placement by X ray after placement. Umbilical catheters are available with single-lumen, dual-lumen, and triple-lumen catheters (13) (Fig. 2,a,b) in different sizes ranging from 2.6 to 8.0 Fr. (Table 1) Umbilical artery catheter tips are designed in two different ways: end-hole and side-hole tips. The side-hole catheters use a special electrode to measure blood gases and biochemicals. A Cochrane review by Barrington (14) showed that end-hole catheters are associated with a much decreased risk of aortic thrombosis compared with side-hole catheters. Therefore, umbilical artery catheters designed with a side-hole should not be used routinely for umbilical artery catheterization in the newborn. Furthermore, manufactures have made sterile, ready-to-go umbilical catheterization trays. These trays contain everything needed for umbilical catheterization, including drapes, towels, suture, umbilical tape, skin preparation materials, needles, forceps, syringes, and other instruments.

Table 1. Neonate Weight and Umbilical Catheter Size

Neonate Weight (g)	UAC Size (Fr)	UVC Size (Fr)
<1500 g	3.5 Fr	3.5 Fr
>1500 g	5.0 Fr	5.0 Fr

Umbilical Vessel Catheterization Procedure

The infant must be supine and restrained (2). A field around the umbilicus is sterilized and draped, and a silk suture is looped around the base of the umbilical stump. The distal end of the stump is cut off, leaving 2 cm of stump, and the vessels are occluded to prevent blood loss. For the umbilical artery catheterization (15), the stump is firmly grasped with the gloved fingers of one hand, and one of the two thick-walled umbilical arteries is dilated with a curved iris forceps; then the umbilical artery catheter is inserted into the artery. Some resistance may be encountered when the catheter has been advanced 3 to 5 cm into the vessel, but this resistance can usually be overcome by applying steady downward pressure on the catheter. If the catheter cannot advance, a second catheter can be inserted into the other artery while leaving the first catheter in place. This maneuver often causes one or the other vessel to relax and permits one catheter to be advanced into the aorta. Advancement of an UAC should place the tip above the celiac axis but below the ductus arteriosis. All air should be removed from the system. The accidental injection of small amounts of air (<0.1 mL) may obstruct blood flow to the legs for several hours. The catheter should be attached to a pressure transducer and the arterial pressure measured. For the umbilical vein catheterization, the single, large, thin-walled umbilical vein is grasped with an iris forceps, and the air-free catheter, which is connected to a closed stopcock, is inserted 3 to 5 cm into the vessel with a twisting motion. The UVC tip should lie a few centimeters into the umbilical vein or inferior vena cava. The stopcock must be closed to prevent aspiration of air through the catheter should the patient take a deep breath. It is imperative that no air be injected through venous catheter, because the air may enter the systemic circulation through the foramen ovale and occlude a coronary or cerebral artery. If it does, the neonate may die or suffer central nervous system damage. If the catheter "tickles" the atrial septum, the neonate may suffer arrhythmias. Withdrawal of the catheter a short distance can solve the problem. Plain radiographs should be taken to confirm placement (Fig. 3). "High" placement of the UAC is defined in one major review of the literature as one with "the tip in the descending aorta above the level of the diaphragm and below the left subclavian artery" and "low" placement of the UAC as one with "the tip above the aortic bifurcation and below the renal arteries" (16).

Neonatal Blood Gas and Biochemical Measurement

The blood gas measurement is the most widely used clinical method for assessing pulmonary function in the neonate. It forms the basis for diagnosis and management of neonates with cardiorespiratory disease (17). The physiology of blood gases is discussed in other parts of this Encyclopedia. We will discuss in this section some issues related to the neonatal blood gas measurement. The dissociation curve of fetal hemoglobin (as compared with adult) is shifted to

Figure 3. Anteroposterior roentgenogram shows the position of umbilical artery and vein catheters. Lateral roentgenogram is needed to distinguish the umbilical artery from the umbilical vein catheter and to determine the appropriate level of insertion. A = endotracheal tube; B = umbilical venous catheter. C = umbilical artery catheter passed up the aorta to T12.

the left, and at any arterial O_2 partial pressure (PaO_2) below 100 mm Hg (13332.2 Pa) fetal blood binds more O_2. This shift seems to be the result of the lower affinity of fetal hemoglobin for 2,3-diphosphoglycerate (DPG). Shunting is a common occurrence in the neonate, such as in congenital cyanotic heart disease, persistent fetal circulation, or atelectasis. O_2 supplementation does not prevent the hypoxia produced by such a shunt. Arterial carbon dioxide partial pressure ($PaCO_2$) is an important measure of pulmonary function in neonatal respiratory disease. The initiation of ventilation with the first breath after normal delivery results in a rapid fall in $PaCO_2$ within minutes of birth. PO_2 rises rapidly to levels of 60 to 90 mm Hg (17). Immaturity of the kidney in the newborn affects the basal acid–base status and the response to additional acid and alkali loads (18). The blood bicarbonate (HCO_{3-}) concentration is typically lower than in the adult (18–21 mEq/L). However, the blood pH (7.35 –7.43) is only marginally decreased because of the compensatory increase in the neonatal respiratory rate. Table 2 lists the normal values of pH, $PaCO_2$, and total CO_2 in the adult and in preterm and term neonates (19). The transition from fetal to neonatal life, which is associated with rapid changes in fluid and electrolyte balance, and the neonate's small size make the electrolytes and glucose assessment difficult and complicated (20), especially in the first week of life. A physiologic decrease in extracellular water volume, as well as a transient increase in serum potassium and transient decreases in plasma glucose and total plasma ionized calcium concentrations, must be taken into account when monitoring neonatal electrolytes and glucose. Frequent and even continuous monitoring of physiological parameters is indicated in some cases, including glucose and calcium monitoring in the premature newborn (limited hepatic glycogen storage) and in newborns of diabetic mothers (21). Before birth, fetal glucose is slightly higher than maternal glucose. With cord clamping, neonatal plasma level plummets over the first 60–90 min of life (23), (23). Neonatal hormonal changes later leads to an increase in endogenous glucose production and stabilization of its level.

Table 2. Acid–base Parameters in Neonates and Adults (mean ± SD)

	Preterm	Term	Adult
pH	7.40 ± 0.08	7.40 ± 0.06	7.40 ± 0.03
*P*CO2	34.0 ± 9.0	33.5 ± 3.6	39.0 ± 2.6
Total CO2	21.0 ± 2.0	21.0 ± 1.8	25.2 ± 2.8

Technical Aspects of Neonatal Blood Gas and Biochemical Measurement

Technologic innovations in the development of biosensors and microprocessors have led to development of bedside small and accurate point-of-care (POC) devices (24). POC devices have been used widely in critical care units, including the NICU. These devices are divided into two groups: "Analyzers," which are not attached to the patient blood source and require blood sampling, and "monitors," which are continuous or near-continuous patient-attached POC monitors (25). Neonatal biochemical measurement may include several important blood parameters, such as sodium, potassium, calcium, glucose, and even lactate Table 3.

Intermittent Blood Gas and Biochemical Measurement (Sampling). This is the most common technique used in the NICU for invasive neonatal blood gas and biochemical measurement. Usually a small blood sample is withdrawn from a blood vessel, such as an umbilical vessel through an umbilical catheter. This sample is analyzed by using a bedside point-of-care analyzer or sent to a small satellite laboratory unit in the NICU or to the central laboratory. Blood gas analyzers are discussed in other articles in this Encyclopedia. These analyzers use the principles of Clark's electrode for PO_2 measurement and Severinghaus's electrode for PCO_2 measurement. Some blood-sample collecting systems are commercially available, such as the Edward VAMP Jr. system manufactured by Edward (Fig. 4), which are manufactured specifically for neonate and small children use. This system is latex-free, disposable, closed, with small volume systems. Such systems are designed to the decrease the risks of blood loss, infection, and air bubbles (26).

Continuous Intravascular Neonatal Blood Gas and Biochemical Sensors. There are several drawbacks for using frequent arterial blood gas (ABG) sampling in the neonate (27). This method may result in blood loss that can necessitate blood transfusion. Moreover, in this method, rapid changes in blood gas values may be missed, especially

Table 3. The Main Blood-Chemistry Parameters Monitored in the Neonatal Care Unit, With Typical Sensing Principles and Transducers

Parameter	Sensing Principle(s)	Transducer(s)
Invasive Blood Pressure	Electrical, impedance	Strain gauge, piezoresistor
	Optical, reflection	Photodetector and emitter
PO_2	Optical, fluorescent	Photomultipler tube
	Electrochemical, amperometric	Clark oxygen electrode
PCO_2	Optical, fluorescent	Photomultipler tube
	Electrochemical, potentiometric	Ion-sensitive electrode
Glucose	Optical, colorimetric	Photodetector
	Electrochemical, amperometric	Enzyme modified biosensor
Lactate	Optical, colorimetric	Photodetector
	Electrochemical, amperometric	Enzyme modified biosensor
Electrolytes (K, Na, Ca, Cl)	Optical, colorimetric	Photodetector
	Electrochemical, potentiometric	Ion-selective electrode (ISE)
pH	Optical, colorimetric	Photodetector
	Electrochemical, potentiometric	Ion-sensitive electrode
Hemoglobin	Optical, absorption	Photodetector and emitters

in conditions needing quick and close ABG monitoring, such as after surfactant administration (28) and during high-frequency ventilation (29). These drawbacks dictate the need for a more efficient real-time way to monitor ABGs (30). For the last two decades, intra-arterial PaO_2 monitoring has been available with the use of a Clark electrode (31) or a multiparameter sensor with an umbilical artery catheter (32). New fiber-optic continuous blood gas monitoring sensors have been validated and used in the neonate with an UAC, such as Neotrend. These devices promise to be safe, easy to use, and accurate in newborns. However, the cost-effectiveness of these devices is still not well established (33) (refer to the blood gas measurement article in this Encyclopedia for details about Neotrend). The *ex vivo* in-line VIA Low Volume Mode blood gas and chemistry monitoring system (VIA LVM Monitor; Metracor Technologies, Inc., San Diego, CA) is an in-line, low-volume POC monitor for neonates and children. Studies have shown promising results in using this monitor in the neonate. However, its cost-effectiveness has not been established yet (25,34). This device measures pH, $PaCO_2$, PO_2, Na^+, K^+, and hematocrit (Hct) by automatically drawing blood (almost 1.5 mL) from a patient's arterial catheter, analyzing it, and reinfusing the blood sample back into the patient. Results are usually displayed in 1–2 min. The operator performs an initial calibration, and then the device performs self-calibration after each sample and at least every 30 min. This machine is compatible with all sizes of UACs and peripheral arterial catheters. Figure 5 shows a diagram of the VIA LVM monitor and its components at the neonate bedside.

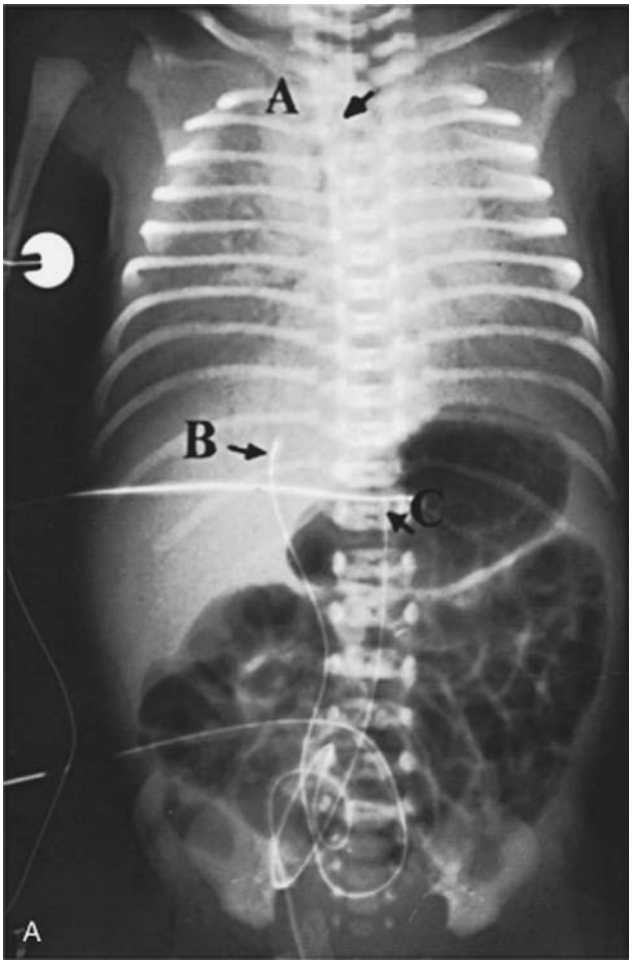


Figure 4. Edward VAMP Jr. blood-sample collecting system.

Figure 5. Diagram of the VIA LVM in-line *ex vivo* monitor and its components at the neonate bedside.

Neonatal Hemodynamic Monitoring

Direct arterial blood pressure monitoring is the most accurate technique for determining arterial pressure in the neonate (9). This method is best done by using umbilical artery catheterization. It is an easy and quick procedure in comparison with other neonatal artery catheterizations, such as radial and femoral artery catheterization. After the umbilical artery catheterization is done as described above, it is connected to a pressure transducer and a continuous flow device, as well as stopcock and manometer tubing.

Basic Concepts. Hemodynamic pressure monitoring requires several basic components to accurately measure the physiologic pressures. These components are: (1) an intravascular catheter, (2) connecting tubing and stopcocks to connect that catheter and the patient's blood vessels to the monitoring system, (3) a pressure transducer to convert the mechanical impulse of a pressure wave into an electrical signal through movement of a displaceable sensing diaphragm, (4) a continuous flush device that fills the pressure tubing with fluid and helps prevent blood from clotting in the catheter, (5) an amplifier that increases the low-voltage signal from the pressure transducer to a signal that can be displayed on a display device, (6) an oscilloscope to display waveforms and a digital readout to display numerical data, and (7) a processor or microcomputer that is used to calculate various hemodynamic parameters based on the measured variables.

Pressure Transducers. Pressure transducers are divided in two groups: (1) External transducers located away from the intravascular catheter and connected to that catheter via fluid-filled pressure tubing, and (2) catheter-tip transducers. The external transducers use three types of sensing elements: (1) strain gauges. These consist of an electrically conductive elastic material that responds reversibly to deformation by a change in electrical resistance. The resistance is converted into a voltage signal by connecting the elements to form a Wheatstone bridge circuit. The output voltage is proportional to the applied pressure and the excitation voltage. Strain gauges are the most common method of pressure transduction. (2) Silicon strain gauges. These are thin slices of silicon crystal bonded onto the back of a diaphragm. The movement of the diaphragm causes a change in the resistance of the crystal, which can be converted into an output signal. Silicon strain gauges are more sensitive than standard strain gauges, but they are affected by temperature and are non-linear. (3) Optical sensors: These are also diaphragms, but in this case, the movement of the diaphragm is sensed by reflecting a beam of light off the silver back of the diaphragm onto a photoelectric cell. The intensity of light sensed by the photoelectric cell changes with the diaphragm position, causing a decrease in its electrical output.

The Pressure Measurement System. The arterial waveform can be characterized as a complex sine wave, which is the summation of a series of simple sine waves of different amplitude and frequencies. The fundamental frequency (or first harmonic) is equal to the heart rate. The first 10 harmonics of the fundamental frequency contribute to the waveform. Any measurement system responds to a restricted range of frequencies only. Within this range, the system may respond more sensitively to some frequencies than to others. The response of the system plotted against the signal frequency is the *frequency response* of the system. However, the measurement system may possess *natural frequencies* or resonances determined by the inertial and compliant elements in a mechanical system. These resonances can distort the output signals. Therefore, it is essential that the natural frequencies do not lie in the operating frequency range of the instrument. Moreover, the output signal of a measurement may differ from the input signal created by the bloodstream because of the inertial components, frictional effects of movement, viscous forces of fluids, and electrical resistance. The property that determines these effects is called the *damping* of the system (35).

Technical Management of Pressure Monitoring System. These are some significant practical points in operating this system:

1. *Removal of all air bubbles from system*: Air is more compressible than fluid, and it tends to act as a "shock absorber" within the pressure monitoring system, leading to an overdamped waveform, which may lead to false readings of the blood pressure. Moreover, air bubbles may cause serious air embolism, especially in neonates and small children.
2. *Zeroing the transducer*: The accuracy of invasive pressure measurements is dependent on the establishment of an accurate reference point (Zeroing). This is done by opening the stopcock to atmospheric pressure and zeroing the measurement system to eliminate the effect of the atmospheric pressure, and by leveling the transducer to the level of the upper portion of the right atrium (the patient's "midaxillary line" or "phlebostatic axis") to eliminate the effect of the blood hydrostatic pressure.
3. *Fast-flush technique*: A "fast-flush" or "square wave test" is performed by opening the valve of the continuous flush device, which leads to an acute increase in the fluid flow rate through the catheter-tubing system from the usual 1–3 mL/h to 30 mL/h. This generates an acute rise in pressure within the system such that a square wave is generated on the monitor. With closure of the valve, a sinusoidal pressure wave of a given frequency and progressively decreasing amplitude is generated. A system with appropriate dynamic response characteristics will return to the baseline pressure waveform within one to two oscillations. If the fast-flush technique produces dynamic response characteristics that are inadequate, the clinician should troubleshoot the system (i.e., remove all air bubbles, minimize tubing length and stopcocks, etc.) until an acceptable dynamic response is achieved. The above-explained basic concepts in hemodynamic monitoring are used in pressure monitoring in neonates as well as in adults. Because of

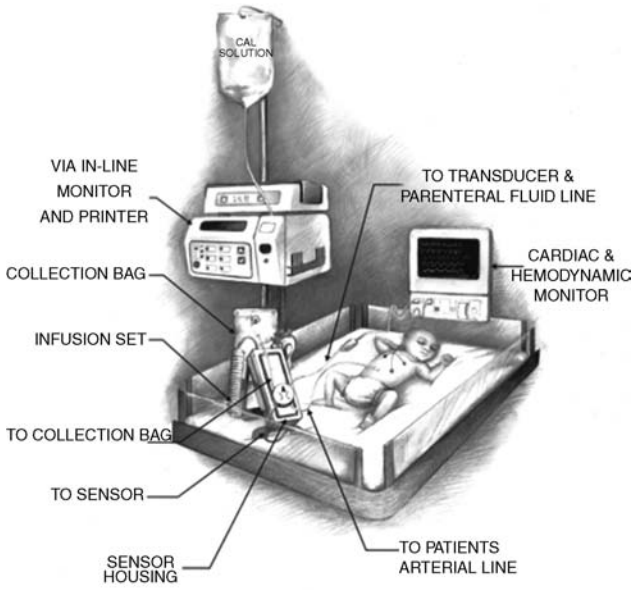


Figure 6. Neonatal/Pediatric Deltran pressure tubing with needleless blood collection system.

the small size of a neonate (or premature newborn) and because of the different indications and expected complications from adults and older children, these monitors have been modified to fit these requirements. These modifications and changes include (1) using a more simple tubing system with fewer stopcocks to reduce the risk of air embolism and infection, (2) using a smaller volume tubing system with small syringes to decrease blood loss and the need for blood transfusion, and (3) using a special constant-low-rate flush system to decrease the risk of fluid overloading. After correct placement of the umbilical artery catheter, a stopcock (free of air bubbles) is connected to its distal end. Then a fluid-filled, well-flushed pediatric/neonatal blood pressure tubing is connected to that stopcock (or directly to the catheter). The transducer is zeroed and leveled to the midaxillary level. There are several commercially available neonatal/pediatric pressure transducers and tubing systems. Figure 6 shows a disposable, latex-free neonatal/pediatric Deltran blood pressure monitoring and needleless blood collection system (36). For central venous pressure (CVP) monitoring, a dual-lumen or triple-lumen umbilical vein catheter may be used. Figs. 2 and 7 show some of the currently available catheters. Neonatal/pediatric pulmonary artery catheters (PACs) have been used through the umbilical vein. However, they are difficult to place and have numerous complications and risks associated with their placement. Other noninvasive cardiac monitors, such as echocardiography, can assess cardiac output and other physiological cardiac parameters. Thus, use of the neonatal PAC has decreased significantly. The most common indications for pediatric PACs are for cardiogenic shock, for severe distributive shock, for the use of very high ventilator pressures to achieve adequate oxygenation, and for the perioperative management of patients who have undergone complex cardiac or other major surgeries (37). The smallest thermodilution catheter available is 5 Fr in size, although a single-lumen 4-Fr catheter exists and is useful for measuring pulmonary artery pressure (PAP) or obtaining mixed venous saturation (MvO_2). Potential complications of pulmonary artery catheterization include pulmonary artery erosion or infarction, dysrhythmia, damage to the pulmonic valve, coiling in the right ventricle, and cardiac perforation (38).

Figure 7. Catheter.

Complications and Risks

Umbilical Artery Catheter. Many complications and risks are associated with UAC placement (39). Therefore, these catheters should not be used solely for fluid and medication administration. If an infant does not require frequent arterial blood sampling or continuous blood pressure monitoring, there is almost no justification for leaving a UAC inserted. The advantages and disadvantages of "high" versus "low" UAC are still debated. Recent studies (40–43) have found that "high" catheters are associated with a decreased incidence of complications without a statistically significant increase in any adverse sequelae. Therefore, one major review of the literature has concluded that "there appears to be no evidence to support the use of low placed umbilical artery catheters. High catheters should be used exclusively" (40). It is clear that UACs that are located between the "high" and "low" positions are never appropriate. Catheters in these positions have been associated with refractory hypoglycemia (infusion into the celiac axis), paraplegia (infusion into the artery of Adamkievicz)

Table 4. Care of Indwelling Umbilical Catheter

- Change all tubings and connections daily.
- Secure and label all tubing, and connections should be secured and labeled appropriately.
- Use only appropriate filters.
- Maintain catheter, connections, and tubing free of blood to prevent clot formation, or inadvertent flushing of preexisting clots into the neonate.
- Flush catheter with 0.5 mL of flush solution each time blood sample is drawn.
- Chart fluids infused in intake/output record.
- Infuse heparinized parenteral solution continuously through catheter, interrupting only to obtain blood samples.

(44), and thromboses that affect the kidneys (infusion into the renal arteries) or the gut (infusion into the mesenteric arteries). A catheter that is found in this intermediate position should be pulled to a "low" position or removed. Similarly, catheters should not be placed below the level of L5 because of the risk of gluteal skin necrosis (45) and sciatic nerve damage (45). Catheters that are placed below the level of L5 should be removed promptly. Other complications may include catheter occlusion, infection, air embolism, breaks or transection of the catheter, electrical hazards, intravascular knots in the catheters, bladder injury, peritoneal perforation, Wharton's jelly embolus, and others.

Umbilical Venous Catheter. Thromboembolic events and infections are common complications of UVC use. These complications are similar to those of other central catheters, although UVCs are associated with an increased risk of localized infections of the liver and heart. Complications that are specific to UVC commonly are the result of malposition of the catheter. Nearly all experts recommend placement of the catheter outside the heart in the IVC (46). Complications occur as a result of placement of the catheter in the right side of the heart or the left side (via the foramen ovale). Cardiac arrhythmias are common complications, but these arrhythmias usually resolve after catheter withdrawal from the heart. Cardiac perforation with subsequent pericardial effusion and cardiac tamponade has been reported (47) Quick diagnosis (high index of suspicion, chest radiograph, and ultrasonography) and prompt treatment with pericardiocentesis decrease mortality significantly in these neonates. Placement of the catheter in the portal system can result in serious hepatic injury. Hepatic necrosis can occur from thrombosis of the hepatic veins or infusion of hypertonic or vasospastic solutions into the liver. Necrotizing enterocolitis and perforation of the colon also have been reported after positioning of the catheter in the portal system. Other complications of umbilical venous catheters have been reported and include perforation of the peritoneum, electrical hazards, and digital ischemia.

Umbilical Catheter Removal and Maintenance

According to the American Academy of Pediatrics (AAP) guidelines (48), umbilical artery or venous catheters should be removed and not replaced if there is any sign of catheter-related bloodstream infections, vascular insufficiency, or thrombosis. Moreover, umbilical catheters must be removed as soon as possible when no longer needed or when any sign of vascular insufficiency to the lower extremities is observed. An umbilical artery catheter should not be left in place more than 5 days. However, an umbilical venous catheter can be used up to 14 days if managed aseptically. Umbilical venous catheters may replaced only if the catheter malfunctions. Table 4 lists the important tasks for daily care of umbilical indwelling catheters.

FUTURE TRENDS IN INVASIVE NEONATAL MONITORING

Providing real-time, accurate, reliable, compact, at the bedside, and safe neonatal monitoring devices is the future goal of researchers and manufactures involved with neonatal critical care (3). However, the cost issue is still outstanding and is going to be a major factor in directing this industry. The market demand for integrated critical care units and the clinical demand for continuous and rapid biochemical monitoring at the bedside, especially in critically ill patients, will lead to a closer integration between the vital signs monitors and POC analyzers. The monitor and POC analyzer manufacturers have been separate companies. This will slowly change through the formation of strategic alliances and mergers. Therefore, new devices will combine continuous vital signs, blood gases, and blood chemistry in the critical care units, including the NICU. The demand for continuous monitoring of biochemical parameters is at the same time bounded by the requirement for low-blood-loss systems, especially in the NICU. This is where the in-line and indwelling analyzers can offer a major advantage over the POC analyzers. Fortunately, technology is also moving in the right direction to minimize iatrogenic blood loss and decrease the risk of infection through the advances made in sampling, preparation, and handling of liquids using microfluidic techniques and closed systems (49). It is likely that in-line monitors that interface to umbilical artery catheters will become more widespread and will require decreasing sample-volumes. The continuing application and refinement of established optical assay methods, such as absorption and fluorescence spectroscopy, onto fibre-optic cables will enable the detection of increasing numbers of analytes by indwelling probes encased within the umbilical catheters (50,51).

SUMMARY

The use of umbilical vessel catheterization and associated monitoring techniques and devices has been advanced dramatically since Dr. Louis K. Diamond of Boston used plastic tubing on the umbilical vein for blood transfusion in 1946. Advances in invasive neonatal monitoring and neonatal intravascular access have led to a significant reduction in the incidence of complications and have increased the number of indications for umbilical catheters. Umbilical vessel catheterization is now a routine and safe procedure

in the NICU. Moreover, with the increase in the survival rate of low- and very low-birth-weight neonates, the need for these catheters has increased. Sometimes these catheters are the only intravascular access that can be established in this new group of patients. The innovation and development of medical applications of silicone and polyurethane enable the manufacturers to make soft, small catheters with adequate lumen size. These catheters cause minimal adverse reaction to the newborn body. To decrease the risk of blood loss and infection, new blood sampling devices have been developed. These systems are closed and have small volume tubing. New technologies for continuous monitoring of blood gases and neonatal chemistries in-line have been integrated with the umbilical catheters by using special sensors encased in the tip of the catheters. The future trend is to use smaller, compact, real-time, bedside monitors combined with pulse oximetry and other vital signs monitors. However, the issue of the cost-effectiveness of these machines has yet to be determined.

BIBLIOGRAPHY

1. Walsh-Sukys MC, Fanaroff AA. Perinatal services and resources. In: Fanaroff AA, Martin RJ, editors. Neonatal-Perinatal Medicine, Diseases of the Fetus and Infant. 6th ed. St. Louis: Mosby; 1999.
2. Stovroff M, Teague WG. Pediatric surgery for the primary care pediatrician, Part II: Intravenous access in infants and children. Pediatric Clin North Am 1998;45(6).
3. Murkovíc I, Steinberg MD, Murkovíc B. Sensors in neonatal monitoring: Current practice and future trends. Technol Health Care 2003;11:399–412.
4. Diamond LK, Allen Jr FH, Thomas Jr WO, Erythroblastosis fetalis. VII. Treatment with exchange transfusion. N Engl J Med 1951;244:39–49.
5. Klaus MH, Fanaroff AA. Care of the High-Risk Neonate, 5th ed. Philadelphia: WB Saunders; 2001
6. Bartlett RH, Roloff DW, Cornell RG, Andrews AF, Dillon PW, Zwischenberger JB. Extracorporeal circulation in neonatal respiratory failure: A prospective randomized study. Pediatrics 1985;76:479–497.
7. Sadler TW. Langman's Medical Embryology: Cardiovascular System, 8th ed. Philadelphia: Lippincott Williams & Wilkins; 2000.
8. Guyton AC, Hall JE. Textbook of Medical Physiology: Fetal and Neonatal Physiology, 10th ed. Philadelphia: W.B. Saunders; 2000.
9. Zahka KG. Principles of neonatal cardiovascular hemodynamics. In: Fanaroff AA, Martin RJ. editors, Neonatal-Perinatal Medicine, Diseases of the Fetus and Infant. 6th ed. St. Louis: Mosby; 1999.
10. Grady M, Procedures. In: Gunn VL. Nechyba C, editors, Harriet Lane Handbook: A Manual for Pediatric House Officers, 16th ed. St. Louis: Mosby; 2002.
11. Brown EG, Krouskop RW. Monitoring, umbilical artery and vein. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation, New York: Wiley; 1988
12. Chidi CC, King DR, Boles Jr ET, An ultrastructural study of the intimal injury induced by an indwelling umbilical artery catheter. J Pediatr Surg 1983;18:109–115.
13. http://www.utahmed.com/umbili-c.htm.
14. Barrington KJ. Umbilical artery catheters: Catheter design (Cochrane Review). In: The Cochrane Library Issue 4. Oxford; 1997.
15. Gregory GA. Resuscitation of the newborn. In: Miller RD, editor. Miller's Anesthesia. 6th ed. New York: Elsevier; 2005.
16. Barrington KJ. Umbilical artery catheters in the newborn: effects of position of the catheter tip. Cochrane Database Syst Rev CD000505, 2000.
17. Carlo WA. Assessment of pulmonary function. In: Fanaroff AA, Martin RJ, editors. Neonatal-Perinatal Medicine, Diseases of the Fetus and Infant. 6th ed. St Louis: Mosby; 1999.
18. Stork JE, Stork EK. Acid-base physiology and disorders in the neonate. In: Fanaroff AA, Martin RJ. editors, Neonatal-Perinatal Medicine, Diseases of the Fetus and Infant. 6th ed. St. Louis: Mosby; 1999.
19. Lorenz JM, Kleiman LI, Kotagal UR, Reller MD. Water balance in very low birth infants: Relationship to water and sodium intake and effect on outcome J Pediatrics 1982;101: 423–432.
20. Lorenz JM. Assessing fluid and electrolyte status in the newborn, Nat Acad of Clin Biochem. Clin Chem 1997;43(1):205–210.
21. Heck LI, Erenberg A. Serum glucose values during the first 48 h of life. J Pediatr 1978;110:119–122.
22. Srinivasan G, Pildes RS, Cattamanchi G, Voora S, Lillian LD. Plasma glucose values in normal neonates: A new look. J Pediatr 1984;105:114–119.
23. Tsang RC, Chen IW, Freidman MA, Chen I. Neonatal parathyroid function: role of gestational and postnatal age. J Pediatr 1973;83:728–730.
24. Ehrmeyer SS, Laessig RH, Leinweber JE, Oryall JJ. Medicare/CLIA final rules for proficiency testing: Minimum intralaboratory performance characteristics (CV and bias) needed to pass. Clin Chem 1990;36:1736–1740.
25. Billman GF, et al. Clinical performance of an in-line, ex vivo point-of-care monitor: A multicenter study. Clin Chem 2002;48(11): 2030–2043.
26. http://www.edwards.com/Products/PressureMonitoring/Vamp Jr.htm
27. Meyers PA, Worwa C, Trusty R, Mammel MC. Clinical validation of a continuous intravascular neonatal blood gas sensor introduced through an umbilical artery catheter. Respir Care 2002;47(6):682–687.
28. Kresch MJ, Lin WH, Thrall RS. Surfactant replacement therapy. Thorax 1996;51(11):1137–1154.
29. Nelle M, Zilow EP, Linderkamp O. Effects of high-frequency oscillatory ventilation on circulation in neonates with pulmonary interstitial emphysema of RDS. Inten Care Med 1997;23(6): 671–676.
30. Goddard P, et al. Use of continuously recording intravascular oxygen electrode in the newborn. Arch Dis Child 1974;49(11): 853–860.
31. Weiss IK, Fink S, Harrison R, Feldman JD, Brill JE. Clinical use of continuous arterial blood gas monitoring in the pediatric intensive care unit. Pediatrics 1999;103:440–445.
32. Morgan C, et al. Continuous neonatal blood gas monitoring using a multiparameter intra-arterial sensor. Arch Dis Child Fetal Neonatal Ed 80(2):F93–F98.
33. Rais-Bahrami K, Rivera O, Mikesell GT, Short BL. Continuous blood gas monitoring using in-dwelling optode method: Comparison to intermittent arterial blood gas sampling in ECMO patients. J Perinatol 2002;22(6):472–474.
34. Widness JA, et al. Clinical performance of an in-line point-of-care monitor in neonates. Pediatrics 2000;106(3): 497– 504.
35. Gardner RM. Invasive pressure monitoring. In: Civetta JM, Taylor RW, Kirby RR. editors Critical Care, 3rd ed. Philadelphia: Lippincott-Raven; 1997. pp 839–845.
36. http://www.utahmed.com/deltran.htm.
37. Ewert P, Nagdyman N, Fischer T, Gortner L, Lange PE. Continuous monitoring of cardiac output in neonates using an intra-aortic Doppler probe. Cardiol Young 1999;9(1): 42–48.

38. Tsai-Goodman B, et al. Development of a system to record cardiac output continuously in the newborn. Pediatr Res. 1999;46(5):621–625.
39. Hermansen MC, Hermansen MG. Intravascular catheter complications in the neonatal intensive care unit. Clin Perinatol; 2005;32(1):141–156.
40. Barrington KJ. Umbilical artery catheters in the newborn: Effects of position of the catheter tip. Cochrane Database Syst Rev CD000505, 2000.
41. Kempley ST, Bennett S, Loftus BG. Randomized trial of umbilical arterial catheter position: Clinical outcome. Acta Paediatr 1993;82:173–176.
42. Mokrohisky ST, Levine RL, Blumhagen JD. Low positioning of umbilical-artery catheters increases associated complications in newborn infants. N Engl J Med 1978;299:561–564.
43. Umbilical Artery Catheter Trial Study Group. Relationship of intraventricular hemorrhage or death with the level of umbilical artery catheter placement: A multicenter randomized clinical trial. Pediatrics 1992;90:881–887.
44. Haldeman S, Fowler GW, Ashwal S. Acute flaccid neonatal paraplegia: A case report. Neurology 1983;33:93–95.
45. Cumming WA, Burchfield DJ. Accidental catheterization of internal iliac artery branches: A serious complication of umbilical artery catheterization. J Perinatol 1994;14:304–309.
46. Klaus MH, Fanaroff AA. Care of the high-risk neonate, 5th ed. Philadelphia: WB Saunders; 2001.
47. Nowlen TT, Rosenthal GL, Johnson GL. Pericardial effusion and tamponade in infants with central catheters. Pediatrics 2002;110:137–142.
48. O'Grady NP, et al. Guidelines for the prevention of intravascular catheter-related infections. Pediatrics 2002;110(5):e51.
49. Tudos AJ, Besselink GAJ, Schasfoort RBM. Trends in miniaturized total analysis systems for point-of-care testing in clinical chemistry. Lab on a Chip: 2001;1(2):83–95.
50. Wolfbeis OS. Fiber-optic chemical sensors and biosensors. Anal Chem 2001;74(12):2663–2677.
51. Zhang XC. Terahertz wave imaging: Horizons and hurdles. Proceedings of the First International Conference on Biomedical Imaging and Sensing Applications of THz Technology, Physics in Medicine Biology 2002;47(21):3667–3677.

See also ARTERIES, ELASTIC PROPERTIES OF; BLOOD GAS MEASUREMENTS; FIBER OPTICS IN MEDICINE; NEONATAL MONITORING; STRAIN GAGE.

MONOCLONAL ANTIBODIES

BRENDA H. LASTER
JACOB GOPAS
Ben Gurion University of the Negev
Beer Sheva, Israel

INTRODUCTION

This article outlines the association of antibodies within the human immune system, the structural and binding characteristics of antibodies, and the development and production of monoclonal antibodies. Recent advancements in recombinant DNA techniques and genetic engineering are described, including the use of plants to increase the production capacity of Mabs. Their usefulness as biological and medical reagents is further elaborated in a description of the various instrumentation, techniques, and assays employed in the diagnosis and treatment of diseases as well as their utility in the research laboratory.

THE IMMUNE SYSTEM AS IT APPLIES TO ANTIBODIES (1)

The immune system is normally directed at foreign molecules borne by pathogenic microorganisms. However, the immune system can also be induced to respond to simple nonliving molecules. Any substance that can elicit an immune response is said to be immunogenic and is called an immunogen. There is a clear operational distinction between an immunogen and an antigen. An antigen is defined as any substance that can bind to a specific antibody (see below), but is not necessarily able to elicit an immune response by itself.

Immunization

The deliberate induction of an immune response is known as immunization. To determine whether an immune response has occurred and to follow its course, the immunized individual is usually monitored for the appearance of antibodies directed at the specific antigen. Monitoring the antibody response usually involves the analysis of relatively crude preparations of sera. The serum is the fluid phase of clotted blood, which contains a variety of specific antibodies against the immunizing antigen as well as other soluble serum proteins.

Cells Participating in an Immune Response

B lymphocytes (or simply B cells) are one of the two major types of lymphocytes that enable the adaptive immune response. When activated, B cells differentiate into plasma cells that secrete antibodies. T lymphocytes or T cells consist of three main classes. One class differentiates upon activation into cytotoxic T cells, which may kill foreign tissues, cancer cells, and cells infected with virus. The second class of T lymphocytes is T helper cells that differentiate into cells that activate and enable the proper function of other cells, such as B cells. The third class is the T suppressor cells that limit the extent of the immune response.

Antigen Recognition

Both T and B lymphocytes bear receptor proteins on their surface that allow them to recognize antigen. Collectively, these receptors are highly diverse in their antigen specificity, but each individual lymphocyte is equipped with membrane-bound receptors that will recognize only one particular antigen. Each lymphocyte therefore recognizes a different antigen. Together, the receptors of all the different lymphocytes are capable of recognizing a very wide diversity of antigens, which encompass most of the different antigens an individual will meet in a lifetime. These include those antigens that are exclusively synthesized in the laboratory. The B cells do not secrete antibody until they have been stimulated by specific antigen. The B-cell antigen receptor (BCR) is a membrane-bound form of the same antibody that they will secrete when activated by antigen. Thus the antigen recognized by both the BCR and

the secreted antibody present in the same B cell, are identical.

Antibodies

Antibody molecules as a class are now generally known as immunoglobulins (Ig), and the antigen receptor of B lymphocytes is known as surface immunoglobulin. The T cell antigen receptor (TCR) is related to immunoglobulins, but is quite distinct from it in structure and function.

Structure and Function of Antibodies

Antibodies are the antigen-specific products (proteins) secreted by B cells. The antibody molecule has two separate functions: one is to bind specifically to molecules from the immunogen (pathogen) that elicited the immune response; the other is to recruit various cells and molecules in order to remove and destroy the pathogen once the antibody is bound to it. These functions are structurally separated in the antibody molecule. One region of the antibody specifically recognizes antigen and the other engages the effector mechanisms that will dispose of it.

The antigen-binding region varies extensively among antibody molecules and is thus known as the variable region or V region, labeled as VH (heavy chain) and VL (light chain). It is this variability that allows each antibody molecule to recognize and bind a particular antigen. The total repertoire of antibodies made by a single individual is large enough to ensure that virtually any structure can be bound. The association between the antibody and the antigen depends on their steric conformation. That is, depending on the size and interatomic distance of these reacting molecules, a tight fit between the antibody combining sites and the antigenic determinant can occur.

The region of the antibody molecule that engages the effector functions of the immune system, but is not associated with antibody binding and does not vary in the same way is known as the constant region or C region. It is typified by the IgG antibody shown in Fig. 1 and is designated CL (light chain) and CH (heavy chain). It has five main forms, or isotypes, that are specialized for activating the different immune effector mechanisms.

The remarkable diversity of antibody molecules is the consequence of a highly specialized mechanism by which

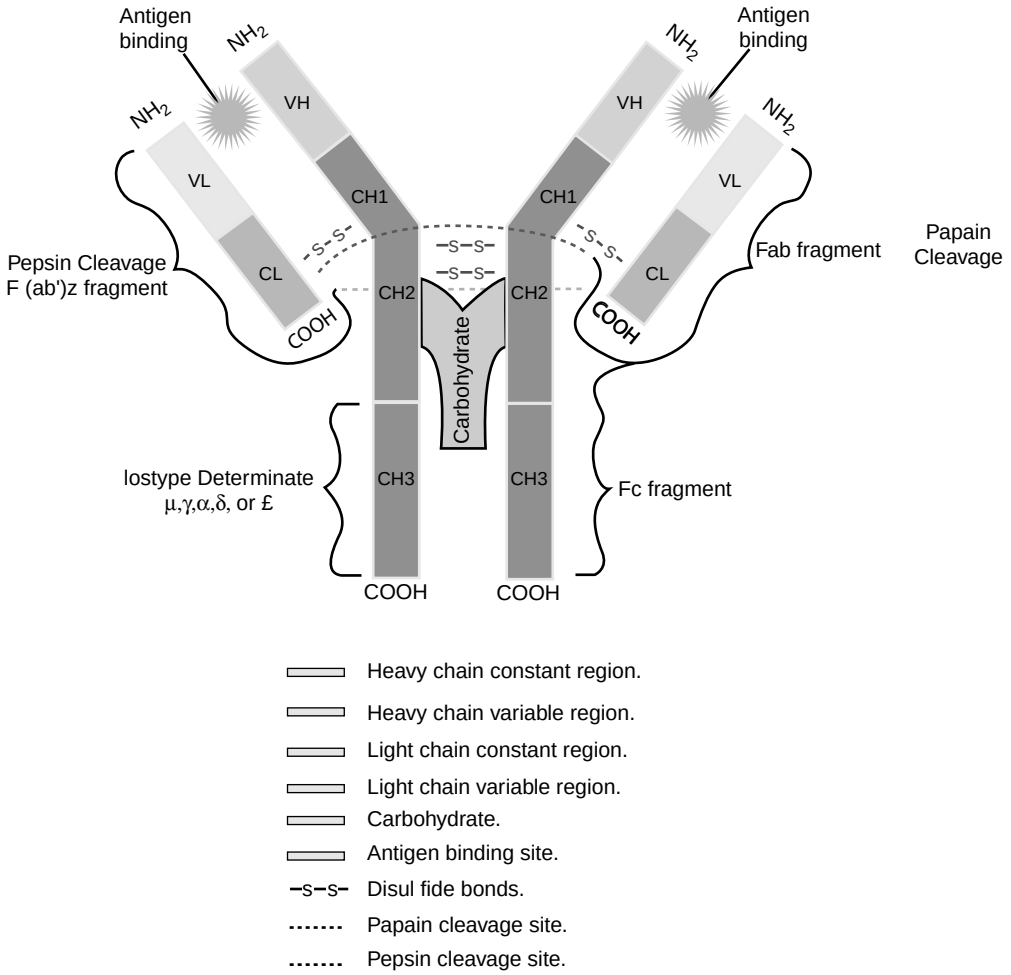


Figure 1. Basic immunoglobulin structure. [Courtesy of Sigma-Aldrich (www.sigmaaldrich.com/img/assets/8181/AntibodyExp).]

the genes that code for antibody production and that are expressed in any given B cell are assembled by DNA rearrangements that join together two or three different segments to form a V region gene during the development of the B cell. Subsequent DNA rearrangement can attach the assembled V-region to any C-region gene and thus produce antibodies of any of the five isotypes.

Antibodies are roughly Y-shaped molecules. All antibodies are constructed in the same way from paired heavy and light polypeptide chains. In Fig. 1, one can observe that the innermost regions of the Y-shaped molecule are the heavy chains; the light chains are the outermost regions. Within this general category, however, five classes (isotypes) of immunoglobulin -IgM, IgD, IgG, IgA, and IgE- can be distinguished biochemically as well as functionally. The five classes are defined by the structure of their heavy chain. Their distinctive functional properties are conferred by differences in the amino acid sequences of the carboxy-terminal part of the heavy chain (COOH in Fig. 1) in the region that is not associated with the light chain. IgG (Fig. 1) will be used to describe the general structural features of immunoglobulin molecules.

The IgG antibodies are large molecules (~150 kDa) composed of two different polypeptide chains. One of these polypeptide chains, ~50 kDa in size, is termed the heavy or H chain. The other, is 25 kDa in size, and is termed the light chain or L chain. The two chains are present in an equimolar ratio, and each IgG molecule contains two heavy chains and two light chains. The two heavy chains are linked to each other by disulfide bonds and each heavy chain is linked to a light chain by a disulfide bond. In any one immunoglobulin molecule, the two heavy chains and the two light chains are identical, enabling them to bind two identical antigenic determinants.

The amino-terminal sequences of both the heavy and light chains vary greatly among different antibodies. The variability in sequence is limited to approximately the first 110 amino acids on the chain, corresponding to the first domain, whereas the carboxy-terminal sequences are constant between immunoglobulin chains, either light or heavy, of the same isotype.

Fragmentation of Antibodies

The antibody molecule can be readily cleaved by different proteases into functionally distinct fragments. For example, Fab fragments, each of which consists of two identical fragments, each containing the antigen binding region. Additionally, an Fc fragment can be extracted that interacts with effector molecules and cells, or one $F(ab')_2$ fragment that contains both arms of the antigen binding region. Figure 1 shows the sites of the derivation of these fragments. Genetic engineering techniques now permit the construction of designed variations of the antibody molecule such as a truncated Fab that comprises only the V region of a heavy chain linked to a V region of a light chain. This is called a single–chain Fv. A broad range of genetically engineered molecules are now becoming valuable therapeutic agents because their smaller size readily permits their penetration into tissue. Useful antibodies from animal sources have been engineered in a process referred to as "humanization". This avoids their recognition as foreign, and prevents their rapid clearance from the body. The process utilizes the variable region of a mouse antibody coupled to the Fc region from human antibodies. Antibody fragments may also be coupled to toxins, radioactive isotopes and protein domains that interact with effector molecules or cells.

MONOCLONAL ANTIBODIES

Antibody Heterogeneity

The antibodies generated in a natural immune response or after immunization in the laboratory are a mixture of molecules of different antigen specificities and affinities. Because of their multiple specificities they are termed polyclonal antibodies. Some of this heterogeneity results from the production of antibodies that bind numerous different antigenic determinants (epitopes) present on the immunizing antigen. However, even antibodies directed at a single antigenic determinant can be markedly heterogeneous. Antisera (serum containing antibodies against specified antigens) are valuable for many biological purposes, but they have certain inherent disadvantages that relate to the heterogeneity of the antibodies they contain. First, each antiserum is different from all other antisera, even when raised in a genetically identical animal while using the identical preparation of antigen and immunization protocol. Second, antisera can be produced in only limited volumes, and thus it is impossible to use the identical serological reagent in a long or complex series of experiments, or in clinical tests or therapy. Finally, even purified antibodies may include minor populations of antibodies that give unexpected cross-reactions that confound the analysis of experiments and can be harmful in therapy. To avoid these problems, and to harness the full potential of antibodies, it became necessary to develop a method for making an unlimited supply of antibody molecules of homogeneous structure and known specificity. This has been achieved through the production of monoclonal antibodies from hybrid antibody forming cells or, more recently, by genetic engineering.

Production of Monoclonal Antibodies

In 1975, using cell culture techniques, Kohler and Milstein (2) found a way to grow an immortal B-cell-like lymphocyte that continuously produced an antibody with a predetermined specificity. The procedure required the immunization of a mouse in order to produce a large population of B-cells in the spleen that would secrete a specific antibody. However, in cell culture, the life span of spleen cells is only a few days. To produce a continuous source of antibody, the B cells had to grow continuously. This was achieved by fusing the spleen cells that contained a particular gene, the hypoxanthine-guanine phosphoribosyl transferase (HGPRT) gene with immortal myeloma cells (cancerous plasma cells). In general, plasma cells are mature-antibody secreting B cells. The myeloma cells were preselected to ensure three specific properties. First, immortality in cell culture; second, that they were sufficiently altered so that

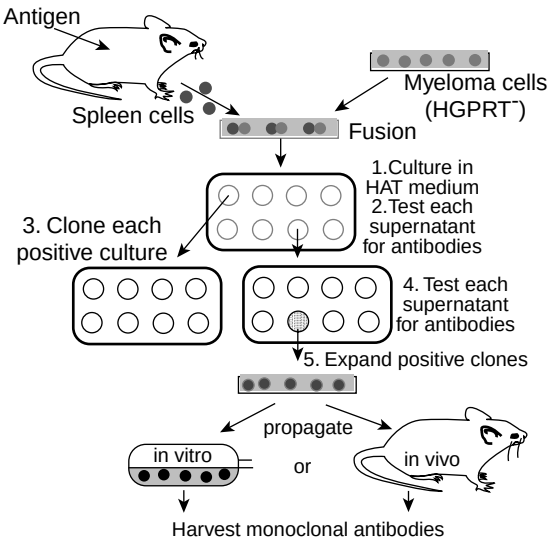


Figure 2. Schematic of hybridoma protocol. [Courtesy of Prof. John Kimball (http://users.rcn.com/jkimball.ma.ultranet/Biology Pages/M/Monoclonals.html).]

they did not secrete antibody, and third, that they would not flourish in a particular growth medium containing hypoxanthine, aminopterin and thymidine (HAT). The HAT medium was previously shown to be a highly selective medium for those specific hybrid cells that lack the gene for the enzyme, HGPRT. Consequently, all unfused myeloma and spleen cells would not survive in the HAT medium. The HGPRT gene that was contributed by the spleen cell permitted only hybrid cells to survive in the HAT medium, because only hybrid cells would be able to grow in the culture due to the conferral of immortality by the myeloma cells. Therefore, the immune spleen cells conferred both antibody specificity and the HGPRT gene to the hybrid cell, while the myeloma cell conferred immortality to the spleen cell and they were able to survive indefinitely in culture. This is the method that is used today to produce individual hybridomas (the hybrid cells) that are then screened for antibody production. Single antibody-producing cells that produce an antibody with the desired specificity are cloned. These cloned hybridoma cells are grown in bulk culture to produce large amounts of antibody that are used in a variety of ways. Since each hybridoma is descended from a single B cell, all cells of a particular hybridoma cell line produce the same hybridoma molecule. This is the monoclonal antibody (Mab). The procedure is shown as a schematic outline in Fig. 2.

Recombinant Genetic Engineering

The field of antibody engineering endeavors to improve the target specificity and effector function of the antibodies by altering their construction, while retaining their binding characteristics. This is achieved by using new recombinant engineering technologies to redesign the molecular architecture of the antibodies. Examples of this are shown in Fig. 3, where in a process called genetic engineering, recombinant DNA (spliced DNA that is formed from two or more sources and then rejoined) is produced by putting genes from one species into the cells of an individual organism of another species. The foreign DNA becomes

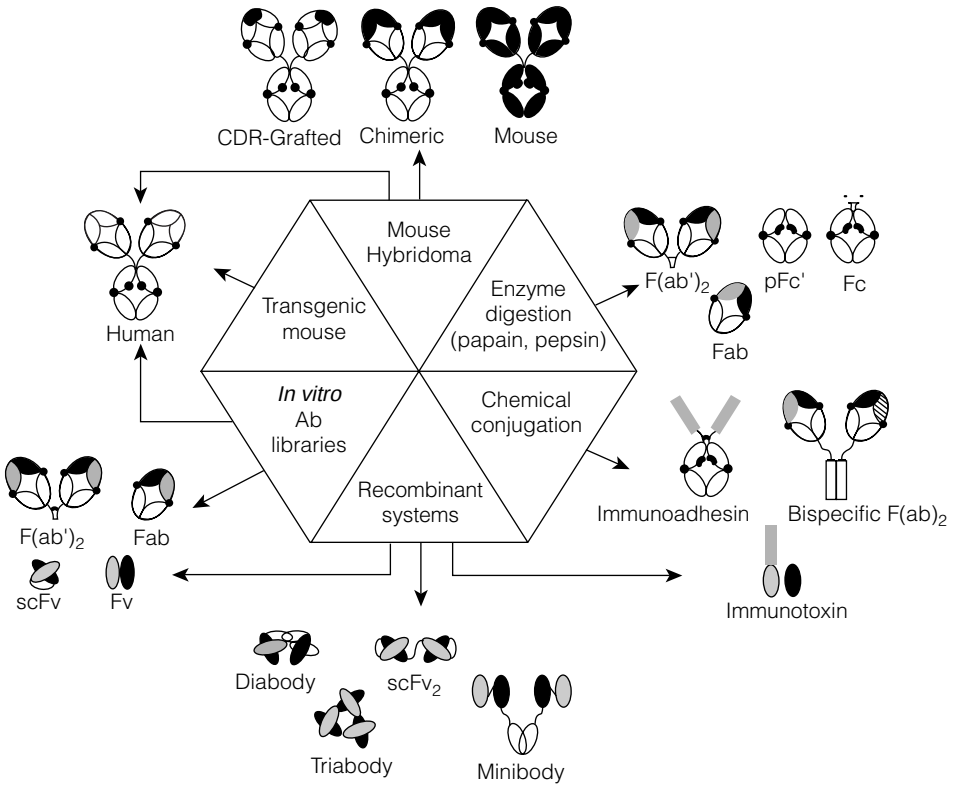


Figure 3. Antibody engineering: examples of antibody-based constructs and respective routes for their generation. Hybridoma technology provides a way of producing mouse monoclonal antibodies. Genetic engineering has encouraged the creation of chimeric, humanized, and human antibodies, antibody fragments (traditionally obtained by partial digestion of immunoglobulins with proteases), multimeric antibody fragments, fusion (immunoadhesins and immunotoxins) and bispecific antibodies. Multimeric antibody fragments (diabody, triabody, and tetrabody) are represented as multivalent structures, although they can also be engineered to be multispecific. The minibody depicted is a dimer that can be linked to the CH_3 fragment via a LD linker or a flexible linker (FlexMinibody). Bispecific F(ab)2 is shown as a Fab dimer linked noncovalently via interaction of amphipathic helices. (Courtesy of American Chemical Society and American Institute of Chemical Engineers, Copyright © 2004.) (See Ref. 3.)

part of the host's genome (its genetic content) and is replicated in subsequent generations descended from that host. An alternative technique for producing antibody-like molecules is the Phage Display Libraries for Antibody V-region Production (4). In this approach, gene segments that encode the antigen-binding variable region of antibodies are fused to genes that encode the coat protein (outside surface) of a bacteriophage (viruses that infect bacteria). In essence, mRNA from primed human B-cells is converted to cDNA. The large variety of diverse antibody genes are expanded by the polymerase chain reaction (PCR) to generate a highly diverse library of antibody genes. Bacteriophage containing such gene fusions are then used to infect bacteria, resulting in phage particles that have outer surfaces that express the both antibody-like fusion protein, and the same antigen-binding domain displayed on the outside of the bacteriophage. The collection of such recombinant phages, each displaying a different antigen-binding domain on its surface, is known as a phage display library. A particular phage can be isolated from the mixture and can be used to infect fresh bacteria. Each phage isolated in this way produces a monoclonal antigen-binding particle analogous to a monoclonal antibody. A complete antibody molecule can then be produced by fusing the V region of a particular phage with the invariant part of the immunoglobulin gene. These reconstructed antibody genes can be introduced (transfected) into a suitable host cell line. These genes will become part of the cell's genome and will secrete antibodies akin to hybridomas.

Monoclonal Antibodies Produced in Plants, Plantibodies

After undergoing genetic engineering techniques, plant cells are capable of assembling and producing unlimited quantities of antibodies, referred to as plantibodies (5). This trademark name for human antibodies manufactured in plants has functionally limitless production capacity and lower costs than those associated with the yeast fermentation process that is currently being used to produce mass quantities of human antibodies. This fairly recent finding might prove to be of benefit in the medical, consumer, and industrial applications of monoclonals. For example, it has been postulated that the development of plantibodies with a capability of sequestering heavy metals or radioactive compounds might have a very positive impact on the environment, particularly because their production is very inexpensive and large supplies are easily produced. Because the corn crop is so readily available worldwide, and its kernel stores natural plantibodies, these can be purified as needed by standard milling procedures. Potato and tomato crops are also being used. The first clinical use of the effectiveness of a plantibody was against the bacterium, *Streptococcus mutans*. This organism produces lactic acid that erodes tooth enamel. The plantibody was brushed onto human teeth for 3 weeks and tooth decay was prevented for up to 4 months. The action of the antibody was to prevent the bacterium from binding to the tooth surface. Plantibody-containing gels are being developed to prevent genital herpes infections and to protect newborn babies during delivery against transmission of the Herpes virus from infected mothers. Plantibodies against human immunodeficiency verus (HIV) and the production of sperm are also being developed. Concerns have been expressed about the use of genetically engineered food crops because of the potential dangers of their getting into the wrong hands, or disturbing the ecological balance.

The aforementioned technologies have revolutionized the use of antibodies by providing a limitless supply of antibodies with single and known specificity. Monoclonal antibodies are now used in most serological assays, as diagnostic probes and as therapeutic agents.

TECHNIQUES FOR USING MONOCLONAL ANTIBODIES AS SEROLOGICAL AND DIAGNOSTIC PROBES

Monoclonal antibodies can serve as tools for diagnosing and treating disease and are valuable agents in the research laboratory. Their utility required the development of procedures that would permit them to be viewed at the particular region of interest. Some of the most widely used techniques are described in the following sections (1).

Immunofluorescence

Since antibodies bind stably and specifically to their corresponding antigen, they are invaluable as probes for identifying a particular molecule in cells, tissues, or biological fluids. Monoclonal antibody molecules (Mabs) can be used to locate their target molecules accurately in single cells or tissue sections by a variety of different labeling techniques. When either the antibody itself, or the anti-Mab that is used to detect it, is labeled with a fluorescent dye the technique is known as immunofluorescence. As in all serological techniques, the antibody binds stably to its antigen, allowing any unbound antibody to be removed by thorough washing. The fluorescent dye can be covalently attached directly to the specific antibody, but more commonly, the bound antibody is detected by a secondary fluorescent anti-immunoglobulin; that is, the first antibody binds to the antigen and a fluorescent secondary antibody (antibody) is targeted to the primary antibody–antigen complex. The technique is known as indirect immunofluorescence, which is demonstrated in Fig. 4, where the binding of the first antibody to the antigen is followed by the binding of the antibody. The dyes chosen for immunofluorescence are excited by light of a particular wavelength, and emit light of a different wavelength in the visible spectrum. By using selective filters that can permit only certain wavelengths of light to pass, only that light coming from the dye or fluorochrome used is detected in the fluorescence microscope. Therefore, the antibody can be located by virtue of its emission of fluorescent light. The recently developed confocal fluorescent microscope considerably enhances the resolution of the technique. If different dyes are attached to different antibodies, the distribution of two or more Mabs can be determined in the same cell or tissue section. Differentiating between the antibodies occurs because either the dye or the fluorochrome will excite at different wavelengths or because they will emit their fluorescence at different wavelengths. An example of the immunofluorescence technique is shown in Fig. 4, whereby, through the use of monoclonal antibodies targeted to

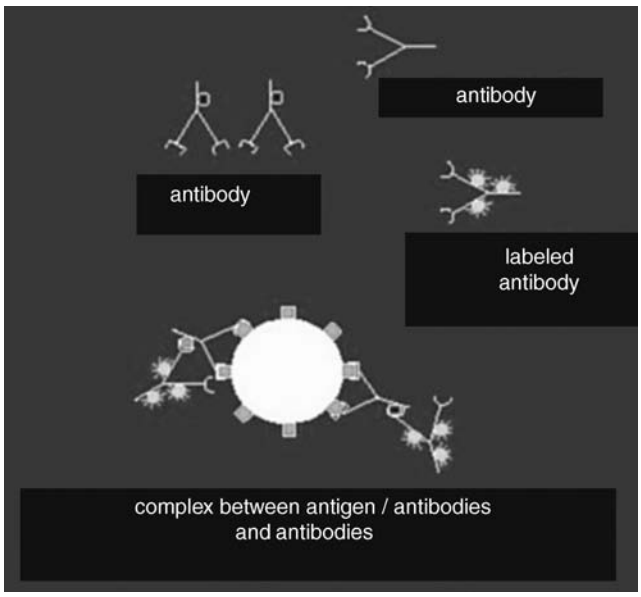


Figure 4. Indirect immunofluorescence. The primary antibody binds to the antigen and the fluorescent antiantibody binds to the primary thereby increasing the signal. [Courtesy of Prof. v. Sengbusch (http://www.biologie.uni-hamburg.de/b-online/d00/copyrig.htm).]

intracellular proteins, certain structures within the cell become visible. The structure shown in Fig. 5 is that of the spindle, which appears during cell division. During certain phases of cell division, the chromosomes arrange themselves in the equatorial plane of the spindle. The spindle is made up of microtubules that, in turn, are composed of proteins. Monoclonal antibodies that would bind to two specific proteins, α and γ-tubulin, of the microtubule were synthesized and labeled with two different fluorochromes. During cell division, the cells were exposed to the fluorescent-labeled Mabs that formed a complex with the proteins and permitted visualization of the spindle.

Immunohistochemistry

An alternative to immunofluorescence for detecting a protein in tissue sections is immunohistochemistry, in which the specific Mab is chemically coupled to an enzyme that converts a colorless substrate into a colored reaction product *in situ*. The Enzyme-Linked ImmunoSorbent Assay (ELISA) is a technique that detects and quantifies specific antigens from a mixture. It is widely used in procedures that screen blood for viral or bacterial contamination, to detect infections, toxins, illegal drugs, or allergens, and in measuring hormone levels, such as in pregnancy or thyroid function. The assay involves the binding of an antibody to a solid surface and exposing it to the antigens. A second complex, consisting of the same antibody, but additionally tagged with a particular enzyme, is exposed to the initial antibody–antigen conjugate and binds. After washing the surface to remove excess unbound antigen, a colorless substrate is added that permits the antigen to be converted into a colored product that can be read and measured using absorption spectrometry. The intensity of color is proportional to the concentration of bound antigen. A schematic of the ELISA assay is shown in Fig. 6. A more detailed description of the Elisa procedure can be found in (Kimball's Biology Pages http://biology-pages.info). Horseradish peroxidase and alkaline phosphatase are also used as enzymes in immunochemistry assays. An example of the utility of these enzymes for protein detection is shown in Fig. 7.

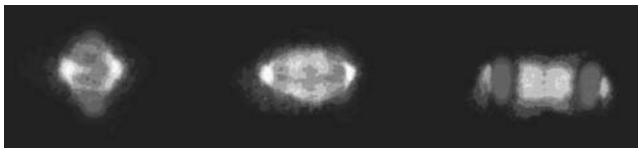


Figure 5. Dividing cells (mitosis: metaphase, anaphase, and telophase) were stained with monoclonal antibodies against two intracellular proteins, α-tubulin in green, and γ-tubulin in red. Because these proteins constitute the spindle, the intracellular structure upon which chromosomes line up during mitosis, the structure is visualized by virtue of the difference in the fluorochromes tagged to the Mabs that were bound to the proteins. Chromosomes were stained with a blue dye. (www.img.cas.cz/dbc/gallery.htm.)

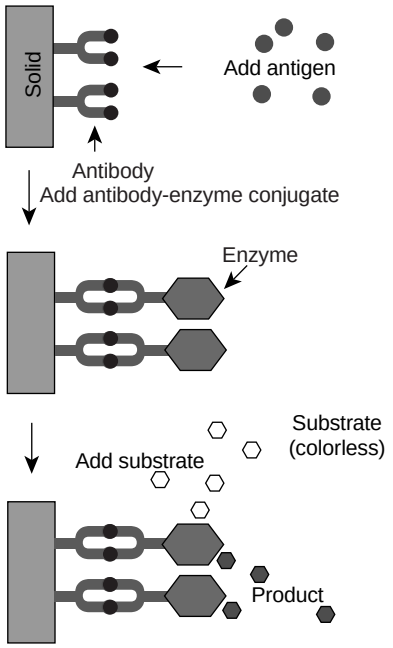


Figure 6. Schematic of ELISA assay protocol. [Courtesy of Prof. John Kimball (http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/E/Elisa.html).]

Immunoelectron Microscopy

Antibodies can be used to detect the intracellular location of structures or particular molecules by electron microscopy, a technique known as immunoelectron microscopy. After labeling Mabs with gold particles and targeting them to samples, they can then be examined in the transmission electron microscope. Since electrons do not penetrate through gold particles, the regions in which the antibodies bind appear as dark dots.

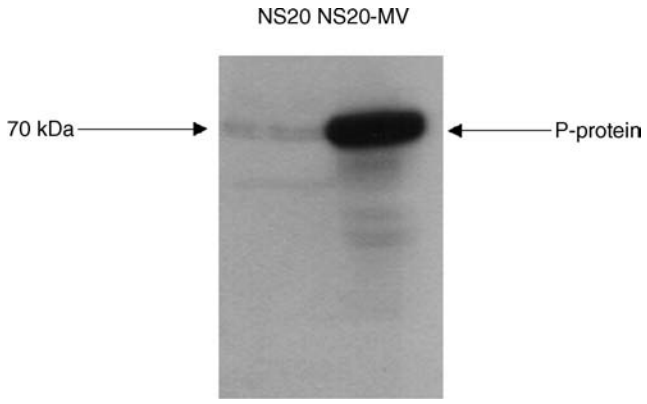


Figure 7. Detection of the measles virus (MV) P-protein by Western Blot in MV infected and noninfected cells. Whole-cell lysates were prepared from MV that was either persistently infected (NS20-MV) or notinfected (NS20) mouse neuroblastoma cells. The proteins in the lysates were separated by SDS–PAGE and blotted onto nitrocellulose paper. The blot was incubated with a Mab against the MV P-protein, followed by a secondary antimouse immunoglobulin antibody linked to horseradish peroxidase. The P-protein band was detected when a substrate was added that was modified by peroxidase on the blot and caused light to be released. Light was detected on a specific band after exposure to film. The results show that only the measles- infected cells express the viral protein. (Courtesy of Jacob Gopas.)

Blotting Techniques

Immunoblotting or Western blotting is used to identify the presence of a given protein in a cell lysate. Cells are placed in detergent to solubilize all cell proteins and the lysate (the material resulting from the ruptured cells) is run on sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS–PAGE), which enables protein migration and separation by size. Further resolution is achieved if the proteins are initially separated by charge according to their isoelectric point and then by size. This technique is referred to as two-dimensional (2D) gel electrophoresis. The proteins are then transferred (blotted) from the gel to a stable support, such as a nitrocellulose membrane for easier handling. Specific proteins of interest in the lysate's mixture are detected by incubating the membrane with a Mab that can react with a defined protein on the membrane. An example of the technique is shown in Fig. 7. The proteins bound to the antibodies are revealed by enzyme-labeled, anti-immunoglobulin antibodies. By this technique the presence or absence, as well as the amounts of specific proteins, can be monitored following a variety cell treatments. Specific DNA labeled with antigen (hapten)-bound nucleotides can be blotted onto a membrane and detected with Mabs against the hapten. This allows the detection of viral or bacterial DNA in tissues or body fluids, as generated by PCR.

Purification Techniques

Affinity chromatography and immunoprecipitation are techniques that enable purification of molecules and their characterization. A mixture of molecules can be incubated with a Mab, which is chemically attached to a solid support. The bound antibody–antigen complex is washed from unbound molecules by centrifugation, and then the molecule of interest is eluted for further characterization. These techniques are useful for protein purification, for determining its molecular weight, its abundance, distribution, and whether it undergoes chemical modifications as a result of processing within the cell.

Immunoelectrophoresis

Two-dimensional electrophoresis is used to separate different antigens that might be present in one solution. The antigens are separated on the basis of their electrophoretic mobility. The currents are run at right angles to each other, driving the antigens into the antiserum (containing Mabs). Peaks are obtained when the antigen forms a complex with the antibody; the area under the peaks gives the concentration of antigen as shown in Fig. 8. Rocket electrophoresis is a similar technique. Here, after a current is applied, the antigens are separated based upon their ionic charge by their differential migration through a gel that contains antibody. As shown in Fig. 9, concentration is determined by the migration distance. In countercurrent electrophoresis, the greater internal osmotic pressure drives the antibody backwards into a gel after a current is applied. An antigen that is negatively charged will form a complex with the antibody in the gel in a pH-dependent process.

Instrumentation

An immensely powerful tool for defining and enumerating and isolating cells is the use of the fluorescence-activated

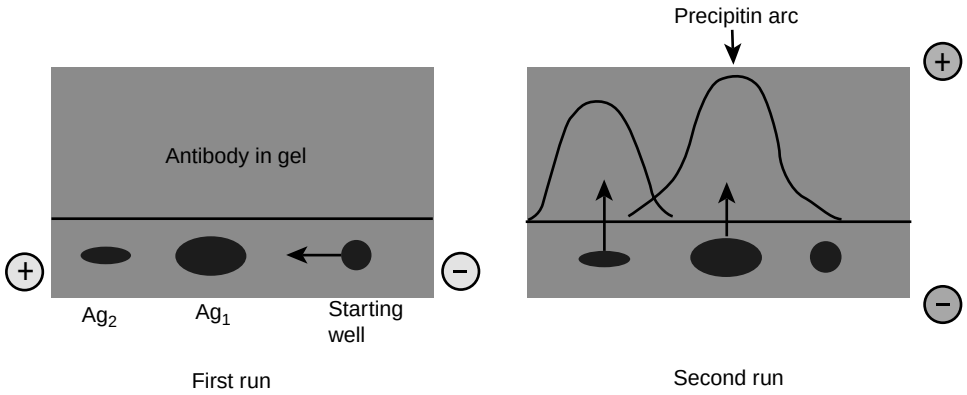


Figure 8. Two-dimensional immunoelectrophoresis. Antigens are separated on the basis of electrophoretic mobility. [Courtesy of the Natural Toxins Research Center at Texas A&M University – Kingsville (http://ntri.tamuk.edu/).]

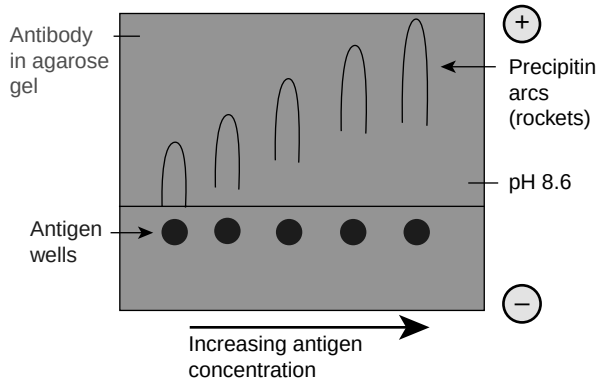


Figure 9. Rocket electrophoresis. Antigen is electrophoresed into gel containing antibody. The distance from the starting well to the front of the rocket shaped arc is related to antigen concentration. [Courtesy of the Natural Toxins Research Center at Texas A&M University – Kingsville (http://ntri.tamuk.edu/).]

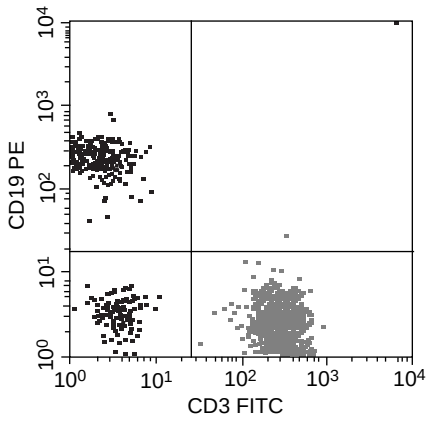


Figure 10. FACS analysis. Characterizing cells at different stages of development through the use of fluorescent labeled monoclonal antibodies against cell surface markers is one of the most common applications of flow cytometry. Changes in the relative numbers, absolute counts, or in the ratio of cell types can provide valuable information as to the status of the immune system in human disorders or animal models. Different cell types can be detected and quantified from a mixed population by the use of monoclonal antibodies labeled with different fluorescent dyes that have nonoverlapping emission spectra. In this example experiment, blood lymphocytes were incubated with two different Mabs, CD19, and CD3. CD19 was labeled with the fluorochrome phycoerythrin (PE) and binds a cell membrane molecule specific for B-lymphocytes. The CD3 was labeled with fluorecein isothyocyanate (FITC) that detects a cell membrane protein specific for T-lymphocytes. Three populations of cells were detected in this experiment according to the antibody bound to the cells. The logarithmic x and y axis represent relative amounts of fluorescence detected on cells labeled with FITC or PE, respectively. The blue dots represent cells unstained by either of the antibodies, the red dots represent B-lymphocytes that were detected by CD19 and the green dots represent T-lymphocytes, CD3 positive cells. No cells were detected that bound both antibodies (top right quadrant).

cell sorter (FACS). This instrument is used to study the properties of cell subsets identified using Mabs to cell surface proteins. Individual cells are first tagged by treatment with specific fluorescent Mabs. The mixture of labeled cells is then forced with a much larger volume of fluid through a nozzle, creating a fine stream of liquid containing cells spaced singly at intervals. As each cell passes through a laser beam it scatters the laser light, and any dye molecules bound to the cell will be excited and fluoresce. Sensitive photomultiplier tubes detect both the scattered light, which gives information on the size and granularity of the cell, and the fluorescence emission, provide quantification of the binding of the labeled Mabs, and on the expression of cell-surface proteins by each cell. In the cell sorter, the signals passed back to the computer are used to generate an electric charge, which is passed from the nozzle through the liquid stream. Droplets containing a charge can then be deflected from the main stream as they pass between plates of opposite charge. In this way a specific population of cells, distinguished by the binding of the labeled antibody and its defined electrical charge, can be extracted and purified from a mixed population of cells. Alternatively, to deplete a population of cells, a labeled antibody directed at marker proteins expressed by undesired cells will direct the cells to a waste channel, retaining only the unlabeled cells. Several Mabs labeled with different fluorochromes can be used simultaneously. FACS analysis can give quantitative data on the percentage of cells bearing different molecules, and the relative abundance of the particular molecules in the cell population, 10,000 cells in a typical experiment demonstrates the retrieval of data after FACS analysis. An example of data output from FACS is shown in Fig. 10.

Mabs as Molecular Probes

Mabs can also be used to determine the function of molecules. Some antibodies are able to act as agonists, when the binding of the Mab to the molecule mimics the binding of the natural ligand (antigen) and activates its function. For example, antibodies to the CD3 antigen present on mature human T cells have been used to stimulate the T cells. This occurs because CD3 is associated with the T-cell receptor and is responsible for signal transduction of the receptor. Conversely, Mabs can function as antagonists, inhibiting the binding of the natural ligand and thus blocking its function. For example, antibodies that block the epidermal growth factor receptor (a growth stimulating protein) function as antagonists.

THE USE OF MONOCLONAL ANTIBODIES AS THERAPEUTIC AGENTS

Mabs against cell-surface molecules have been used to remove specific lymphocyte subsets or to inhibit cell function *in vitro*. Cytotoxic drugs kill proliferating cells indiscriminately. In contrast, antibodies can interfere with immune responses in a nontoxic and much more specific manner. For example, Mabs can be used to remove undesirable lymphocytes from donor bone marrow cells prior to transplantation. This treatment selectively removes lymphocytes that recognize the host tissues as "foreign" and induce a potentially fatal condition known as Graft versus Host reaction (6).

Mabs are being tested experimentally to inhibit transplant rejection, to alleviate and suppress autoimmune disease and in cancer detection and treatment. The major impediment to therapy with monoclonal antibodies in humans is that these antibodies are mostly of mouse origin, and humans rapidly develop antibody responses to mouse antibodies. This not only blocks the actions of the mouse antibodies, but leads to allergic reactions. If this occurs, future treatment of the same patient with any mouse Mab is unacceptable. In principle, the problem can be avoided by producing antibodies that are not recognized as foreign by the human immune system. Several strategies are being explored for their construction. One approach is to clone human V regions into a phage display library (see above) and select for its ability to bind human cells. With this method, Mabs that are entirely human in origin can be obtained. Second, mice that lack endogenous immunoglobulin genes can be made "transgenic" (chimeric). That is, they can have human genes put into their genome through recombinant DNA techniques. When this occurs, they will then express human immunoglobulin heavy and light genes and eventually antibody molecules. A third approach is to graft the variable region of a mouse Mab into the rest of the human immunoglobulin molecule, in a process known as humanization. These recombinant antibodies are far less immunogenic in humans than the Mabs of the parent mouse, therefore, they can be used for more efficient and repeated treatment of humans with far less risks. In some cases even humanized antibodies may evoke an immune response and must be administered with immunosuppressive drugs.

THE USE OF MONOCLONAL ANTIBODIES IN THE DETECTION, FOLLOW-UP, AND TREATMENT OF CANCER

Tumor-Specific Antigens

For the greater part of the twentieth century, it was assumed that any antigens present on the cell surface of tumor cells would also be present in normal cells; therefore, few investigations were undertaken to elicit any autoimmune response against cancer cells. However, once inbred mouse strains bearing transplanted syngeneic (genetically identical) tumors became available, research studies validated that immune reactions against these tumors could be induced with no toxic effects on normal tissues, and scientists began to pursue the identification of "tumor specific antigens". Shared tumor antigens were found in many of the same types of cancers in different patients, and unique antigens were isolated that were specific for a particular cancer in a particular patient. The SEREX database lists the antigens that have been isolated from humans (7). These antigens have the ability to generate an immune response when introduced into a patient.

The advent of monoclonal antibodies suggested the possibility of targeting and destroying tumors by making antibodies against tumor-specific antigens. However, this relies upon the identification of a tumor-specific antigen that is located on the surface of cancer cells. Because of their ability to differentiate between normal and malignant tissues and to exact a variety of antitumor responses, Mabs offer a significant advantage to conventional forms of therapy. Several monoclonal antibodies have already been proven to be relatively well tolerated and effective for the treatment of many different malignant diseases.

Approaches to Cancer Immunotherapy

Approaches to cancer immunotherapy can be either active or passive. For example, in the active category, tumor vaccines that immunize against particularly defined tumor antigens, can be used. In the passive category is the use of monoclonal antibodies that are either conjugated, unconjugated, or radiolabeled. These same approaches can also be categorized as specific, wherein antigens are directly targeted, or nonspecific, where immune cells are used to directly target tumor cells. Other approaches are taken that elicit antitumor effects with different mechanisms, such as using antibodies to block growth factors or receptors on cells; targeting specific tissue components of the tumor or its blood vessels; interfering with cell signals; or with apoptosis (programmed cell death) (8).

Magic Bullets

While such Mab-based therapies offer a high potential to fulfill the promise of "magic bullets" for the treatment of malignant disease, successful application of these therapies is often impaired by several impediments. Factors inhibiting the therapeutic benefit of Mabs may include low or heterogeneous expression of target antigens by tumor cells, high background expression of antigen on normal cells, host antibody immune responses to the Mabs themselves, insufficient anti-tumor response after Mab binding, as well as physical obstructions preventing antibody binding, such as crossing to and from blood vessels as well as tissue barriers en route to the solid tumor mass (9). These factors influence the ability of the Mabs to penetrate to the tumor.

IMAGING TUMORS WITH MONOCLONAL ANTIBODIES

Mabs in Nuclear Medicine

The presence of malignant tumors can be detected through the use of monoclonal antibodies radiolabeled most frequently with the isotopes technetium-99m (^{99m}Tc) or indium-111 (^{111}In). The particular label selected depends upon the size of the antibody. For example, large fragments or whole antibodies require a longer half-life isotope, such as ^{111}In ($T_{1/2} = 2.8$ days), whereas smaller Fab fragments, that are cleared from the body more quickly, can be labeled with ^{99m}Tc ($T_{1/2} = 6$ h). Imaging is performed by a Single Photon Emission Computed Tomography (SPECT) camera whose detectors scan the body and register the radioactive counts. The counts are then mathematically transformed into an image that displays the sites of radioactivity. The nuclear medicine procedure that utilizes this procedure is known as Tumor-Specific Monoclonal Antibody Radioscintigraphy. Because of occasional difficulties with these techniques, such as inadequate tumor perfusion, inadequate amounts of antigen on the surface of the tumor cells, antigen heterogeneity, and nonspecific uptake, new

approaches are being investigated. However, due to limited clinical experience, it is too early to predict whether they will improve imaging performance (10). Among these methods is the use of other imaging techniques, such as bone scans or computed tomography (CT), in conjunction with SPECT. In other approaches, attempts are being made to augment surface tumor cell antigens by prestimulation with growth factors, such as cytokines (11).

TREATMENT OF HEMATOLOGICAL MALIGNANCIES

Blood-Cell Cancers

Surface antigens on B- and T-cell lymphocytes are also useful targets for the treatment of blood cell (hematopoietic) malignancies, such as leukemias and lymphomas. These antigens are also expressed at high levels on the surface of various populations of malignant cells, but not on normal tissues. With few barriers present to impede Mab binding, hematologic malignancies are well suited to Mab-based therapy. In recent years, several promising Mab-based therapies for the treatment of hematologic malignancies have been developed and either have already received U.S. Food and Drug Administration (FDA) approval or are in the advanced phases of clinical testing (12). The chimeric antibody, rituxan (rituximAb, Genentech, San Francisco, CA) was among the first Mabs awarded Food and Drug Administration approval for the treatment of non-Hodgkin's lymphoma (13,14). This chimeric (human–mouse) antibody binds CD20, a cell surface antigen expressed on mature B lymphocytes and over 90% of non-Hodgkin's lymphoma cells, but not on hematopoetic progenitor or stem cells. Rituxan has proven to be well tolerated and effective in the treatment of non-Hodgkin's lymphoma either by itself, or in combination with traditional chemotherapy, particularly in patients who are refractory to other types of therapy (15). Campath-1 (alemtuzumAb, Ilex Oncology, San Antonio, TX) is another antibody that has also received FDA approval for the treatment of patients suffering from chronic lymphocytic leukemia. A third Mab to receive FDA approval for the treatment of hematologic malignancies is the chimeric Mab, mylotarg (gemtuzumAb ozogamicin, Wyeth-Ayerst Laboratories, Philadelphia, PA). This antibody targets the CD33 antigen expressed on myeloid (white cells) precursors and leukemic cells, and is absent from normal tissues and pluripotent hematopoetic (blood-cell producing) stem cells.

TREATMENT OF SOLID TUMORS

In comparison to the management of hematologic malignancies, successful treatment of solid tumors with Mabs has proven more elusive; however, some significant therapeutic benefits have been achieved. Herceptin (trastuzumAb, Genentech) is a humanized antibody that has received FDA approval for the treatment of metastatic breast cancer. This Mab recognizes an extracellular domain of the HER-2 protein. Clinical trials with herceptin have shown it to be well tolerated both as a single agent for second or third line therapy, or in combination with chemotherapeutic agents as a first line of therapy. Combination therapy resulted in a 25% improvement of overall survival in patients with tumors that overexpress HER-2, and that are refractory to other forms of treatment (16).

The antiepithelial cellular adhesion tumors Mab molecule, Panorex (eclrecolomAb, GlaxoSmith-Kline, United Kingdom), is another Mab -based therapy that is currently being used for the treatment of colorectal cancer. Panorex has shown tangible benefit for cancer patients and has received approval in Germany for the treatment of advanced colorectal cancer. Like other Mabs used for the treatment of solid tumors, Panorex has proven more efficacious in the treatment of micrometastatic lesions and minimal residual disease in comparison to bulky tumor masses (17).

The failure of Mabs in the treatment of bulky lesions is primarily attributable to the low level of injected Mabs that actually reaches its target within a sizable solid-tumor mass. Studies using radiolabeled Mabs suggested that only a very small percentage of the original injected antibody dose, ~0.01–0.1/g of tumor tissue, will ever reach target antigens within a solid tumor (18). This low level of binding is due to the series of barriers confronted by an administered Mab en route to antigens expressed on the surface of tumor cells.

ELICITING ANTITUMOR RESPONSES

After successfully negotiating the gauntlet of obstacles obstructing access to the target cells within a tumor, a therapeutic Mab must still be capable of eliciting a potent antitumor response. Although it is often ambiguous as to the exact mechanisms by which a particular Mab may mediate an antitumor response, both direct and indirect mechanisms can potentially be involved.

Antibodies of the IgG_1 and IgG_3 isotypes can support effector functions of both antibody-dependent cell-mediated cytotoxicity and complement-dependent cytotoxicity. Antibody-dependent cell-mediated cytotoxicity is triggered by interaction between the Fc region of a cell-bound antibody and Fc receptors on immune effector cells such as neutrophils, macrophages, and natural killer cells. This mechanism is critical for the antitumor effects of several therapeutic Mabs.

Many early studies showed that murine Mabs had limited potential to elicit a potent antitumor response, because the murine Fc regions are less efficient at recruiting human effector cells than their human counterparts. This problem has been largely alleviated by the use of chimeric and humanized antibodies. Genetic engineering techniques have also been used to improve the immunologic effects of therapeutic Mabs by altering antibody shape and size, increasing the valency (bonds of affinity) of Mabs, and creating bifunctional antibodies with two antigenic receptors, one to a tumor antigen and another to an effector cell to increase efficiency of antibody-dependent cell-mediated cytotoxicity (19).

In addition to immunologic effects, Mabs can induce antitumor effects by a variety of direct mechanisms, including the induction of apoptosis (programmed cell

death) (20), or the prevention of soluble growth factors from binding their cognate receptors, such as epidermal growth factor (EGF-R) (21) and HER-2 (22). Additionally, Mabs can also be engineered to deliver a cytotoxic agent directly to the tumor. This offers the potential to combine the biological effects of Mabs with the additional effect of a targeted cytotoxic response. The anti-CD33 Mabs, mylotarg, is one such antibody. Combined with the cytotoxic agent, calichaemicin, mylotarg has been reported to be relatively well tolerated, and effective in the treatment of chronic lymphocytic leukemia (23). Antibodies can also be engineered to deliver ionizing radiation directly to tumor cells. Mabs have been conjugated to both α- and β-particle emitting radionuclides (24). Clinical trials in humans also portend the promise of radiolabeled Mabs for the treatment of cancer. In a recent phase III randomized study, patients with relapsed or refractory non-Hodgkin's lymphoma were treated with yttrium-90 and iodine-131 labeled Mabs targeting the CD20 antigen (ibrituximomAb, tiuxetan, and tositumomAb, respectively). Patients treated with these radiolabeled Mabs showed a statistically significant increase in overall response compared with those treated with an unlabeled version of the Mab (rituximAb) (25).

OTHER USES FOR MONOCLONAL ANTIBODIES

Proteomics

After having sequenced the entire human genome, the current task is to understand the "proteome" by identification and quantification of all proteins in a given sample. So far, DNA microarrays have been employed to detect the transcription level [production of messenger ribonucleic acid (mRNA)] of genes in cells. However, it has been found that there is no stringent correlation between transcription level and protein abundance. Furthermore, the status of a protein in terms of its modification and structure cannot be determined by DNA microarrays. To solve this problem, antibody microarrays are envisioned to replace DNA microarrays in proteome research. These arrays consist of a multitude of different antibodies that are immobilized on a solid support and allow characterization of the protein repertoire of a given sample. However, the production of such antibody microarrays and its application require the provision of highly specific and stable antibodies, possessing high affinity and showing no cross-reactivity. Protein and antibody microarrays can be made to encompass as many as 10,000 samples on a chip within the dimensions of a microscope slide (26).

Monoclonal Antibodies in the Food Industry

Monoclonal antibodies are being used in the wine industry. Odors sometimes observed in spoiled food or corked wines are often the result of microbes present in the wood packaging materials. However, this phenomenon has also been observed in bottled water, suggesting that there may be secondary contaminants, such as residues of pesticides that can affect the quality of any packaged food or beverage. To further the quality assurance of products in the wine industry, a project is being carried out to raise antibodies against a TCA molecule (2,4,6-trichcholoanisole) that is thought to be present in cork stoppers and is responsible for the musty taste in wine. The ELISA assay will be employed to detect trace amounts of the contaminating molecule. Also an immunosensor will be used to electrochemically detect the antibody levels present (27). Monoclonal antibodies have also been developed against the vegetative cells and spores of *Bacillus cereus* (28). This bacterium seems to be implicated in food poisoning and is also responsible for food spoilage. It is impossible for the food industry to exclude *B. cereus* from its products because *B. cereus* cells can survive heat processing and can grow in foods kept at refrigerated storage conditions. Two different antibodies were developed. One was used as a specific capture antibody to destroy the bacterium; the other as a detector antibody that would simply identify the presence of *B. cereus*. The ELISA assay was used to detect and quantify the vegetative cells of this pathogenic organism.

Potato cyst nematodes are pests that destroy the potato food crops. Monoclonal antibodies are being used to assist in the development of the plant's resistance to the nematode (29). Recombinant plant monoclonal antibodies have been engineered to protect poultry against coccidosis infections (30).

Monoclonal Antibodies and Bioterrorism

The same plant biotechnology described above is being developed to create strategic reserves of vaccines and antibodies for infectious agents that could be used in biowarfare. Multiple genes can be engineered in plants intended to provide prolonged immunity against new strains of pathogens that have different mechanisms of action. With this technology, every plant cell will produce the signature protein of a particular biowarfare agent. That protein, in turn, will trigger an immune response in a person who consumes the plant material in an unprocessed or lightly processed form, but it will not cause the disease. These antibodies can prevent infection on surface areas, including nasal passages; clear infectious organisms from the body; identify foreign organisms for destruction; and neutralize and remove toxins. Among the disease-causing substances are several potential bioterrorism agents, such as the botulism toxin, anthrax, Ebola virus, plague, and ricin, a poisonous protein found in the seeds of the castor oil plant. Vaccines for anthrax (*Bacillus anthracis*) and bubonic and pneumonic plague (*Yersinia pestis*), two potentially deadly diseases that can be delivered as airborne agents, are being developed. Preliminary data predicts success in using these plant-derived vaccines (31).

CONCLUDING REMARKS

Antibodies, monoclonal antibodies and antibody derivatives constitute $\sim$20 % of biopharmaceutical products currently in development. Antibodies represent an important and growing class of biotherapeutics. Progress in antibody engineering has allowed the manipulation of the basic antibody structure into its minimal essential functions, and multiple methodologies have emerged for raising

and tailoring specificity and functionality. The myriad of monoclonal antibody structures that can be designed and obtained in different formats from various production systems (bacterial, mammalian, and plants) represents a challenge for the recovery and purification of novel immunotherapeutics (3). However, the general use in clinical practice of antibody therapeutics is dependent not only on the availability of products with required efficacy but also on the costs of therapy. As a rule, a significant percentage (50–80%) of the total manufacturing cost of a therapeutic antibody is incurred during downstream processing. The critical challenges posed by the production of novel antibody therapeutics include improving process economics and efficiency to reduce costs, and fulfilling increasingly demanding quality criteria for FDA approval.

BIBLIOGRAPHY

1. Janeway CA JR, Travers P, Walport M, Shlomchik M. Immunobiology. The Immune System in Health and Disease. 6th ed. Churchill Livingstone.
2. Kohler G, Milstein C. Continuous cultures of fused cells secreting antibody of predefined specificity. Nature (London) 1975;256:495–497.
3. Roque AC, Lowe CR, Taipa MA. Antibodies and genetically engineered related molecules: Production and purification. Biotechnol Prog 2004;20:639–654.
4. Kretzschmar T, von Ruden T. Antibody discovery: Phage display. Curr Opin Biotechnol 2002;13:598–602.
5. Fischer R, Twyman RM, Schillberg S. Production of antibodies in plants and their use for global health. Vaccine 2003; 21:820–825.
6. Chen HR. et al. Humanized anti-CD25 monoclonal antibody for prophylaxis of graft-vs-host disease (GVHD) in haploidentical bone marrow transplantation without ex vivo T-cell depletion. Exp. Hematol 2003;31:1019–1025.
7. Hartmann TB, Bazhin AV, Schadendorf D, Eichmuller SB. SEREX identification of new tumor antigens linked to melanoma-associated retinopathy. Int. J. Cancer 2005;10;114:88–93. http://www.licr.org/SEREX.html.
8. Davis DW. et al. Regional effects of an antivascular endothelial growth factor receptor monoclonal antibody on receptor phosphorylation and apoptosis in human 253J B-V bladder cancer xenografts. Cancer Res 2004;64:4601–4610.
9. Christiansen J, Rajasekaran AK. Biological impediments to monoclonal antibody–based cancer immunotherapy. Mol Cancer Ther 2004;3:1493–1501.
10. Koral KF. Update on hybrid conjugate-view SPECT tumor dosimetry and response in 131I-tositumomab therapy of previously untreated lymphoma patients. J Nucl Med 2003;44: 457–464.
11. Villa AM, Berman B. Immunomodulators for skin cancer. J Drugs Dermatol 2004;3:533–539.
12. Von Mehren M, Adams GP, Weiner LM. Monoclonal antibody therapy for cancer. Annu Rev Med 2003;54:343–369.
13. Leget GA, Czuczman MS. Use of rituximAb, the new FDA-approved antibody. Curr Opin Oncol 1998;10:548–551.
14. McLaughlin P. et al. RituximAb chimeric anti-CD20 monoclonal antibody therapy for relapsed indolent lymphoma: Half of patients respond to a four-dose treatment program. J Clin Oncol 1998;16:2825–2833.
15. Leyland-Jones B. TrastuzumAb: hopes and realities. Lancet Oncol 2002;3:137–144.
16. Riethmuller G. et al. Monoclonal antibody therapy for resected Dukes' C colorectal cancer: Seven-year outcome of a multicenter randomized trial. J Clin Oncol 1998;16:1788–1794.
17. Mellstedt H, et al. The therapeutic use of monoclonal antibodies in colorectal carcinoma. Semin Oncol 1991;18:462–477.
18. Khawli LA, Miller GK, Epstein AL. Effect of seven new vasoactive immunoconjugates on the enhancement of monoclonal antibody uptake in tumors. Cancer 1994;73:824–831.
19. Zeidler R, et al. The Fc-region of a new class of intact bispecific antibody mediates activation of accessory cells and NK cells and induces direct phagocytosis of tumour cells. Br J Cancer 2000;83:261–266.
20. Trauth BC, et al. Monoclonal antibody-mediated tumor regression by induction of apoptosis. Science 1989;245:301–305.
21. Yang XD, et al. Eradication of established tumors by a fully human monoclonal antibody to the epidermal growth factor receptor without concomitant chemotherapy. Cancer Res 1999;59:1236–1243.
22. Agus DB, et al. Targeting ligand-activated ErbB2 signaling inhibits breast and prostate tumor growth. Cancer Cell 2002;2:127–137.
23. Estey EH, et al. Experience with gemtuzumAb ozogamycin ("mylotarg") and all-*trans* retinoic acid in untreated acute promyelocytic leukemia. Blood 2002;99:4222–4224.
24. Bander NH, et al. Targeted systemic therapy of prostate cancer with a monoclonal antibody to prostate-specific membrane antigen. Semin Oncol 2003;30:667–676.
25. Witzig TE, et al. Randomized controlled trial of yttrium-90–labeled ibritumomAb tiuxetan radioimmunotherapy versus rituximAb immunotherapy for patients with relapsed or refractory low-grade, follicular, or transformed B-cell non-Hodgkin's lymphoma. J Clin Oncol 2002;20:2453–2463.
26. Angenendt P, et al. Seeing better through a MIST: evaluation of monoclonal recombinant antibody fragments on microarrays. Anal Chem 2004;76:2916–2921.
27. Sanvicens N, Varela B, Marco MP. Determination of 2,4,6-trichloroanisole as the responsible agent for the musty odor in foods. 2. Immunoassay evaluation. J Agric Food Chem 2003;51: 3932–3939.
28. Charni N, et al. Production and characterization of monoclonal antibodies against vegetative cells of Bacillus cereus. Appl Environ Microbiol 2000;66:2278–2281.
29. Fioretti L, et al. Monoclonal antibodies reactive with secreted-excreted products from the amphids and the cuticle surface of Globodera pallida affect nematode movement and delay invasion of potato roots. Int J Parasitol 2002;32:1709–1718.
30. Refega S, et al. Production of a functional chicken single-chain variable fragment antibody derived from caecal tonsils B lymphocytes against macrogamonts of Eimeria tenella. Vet Immunol Immunopathol 2004;97:219–230.
31. Petrenko VA, Sorokulova IB. Detection of biological threats. A challenge for directed molecular evolution. J Microbiol Methods 2004;58:147–168.

See also BORON NEUTRON CAPTURE THERAPY; IMMUNOTHERAPY.

MOSFET. See ION-SENSITIVE FIELD-EFFECT TRANSISTORS.

MRI. See MAGNETIC RESONANCE IMAGING.

MUSCLE ELECTRICAL ACTIVITY. See ELECTROMYOGRAPHY.

MUSCLE TESTING, REHABILITATION AND. See REHABILITATION AND MUSCLE TESTING.

MUSCULOSKELETAL DISABILITIES. See REHABILITATION, ORTHOTICS FOR.

ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition

VOLUME 5

Nanoparticles – Radiotherapy Accessories

ENCYCLOPEDIA OF MEDICAL DEVICES AND INSTRUMENTATION, SECOND EDITION

ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition
Volume 5

Nanoparticles – Radiotherapy Accessories

Edited by

John G. Webster
University of Wisconsin–Madison

The *Encyclopedia of Medical Devices and Instrumentation* is available online at http://www.mrw.interscience.wiley.com/emdi

WILEY-INTERSCIENCE

A John Wiley & Sons, Inc., Publication

Published by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Encylopedia of medical devices & instrumentation/by John G. Webster,

editor in chief. – 2nd ed.
p. ; cm.
Rev. ed. of: Encyclopedia of medical devices and instrumentation. 1988.
Includes bibliographical references and index.
ISBN-13 978-0-471-26358-6 (set : cloth)
ISBN-10 0-471-26358-3 (set : cloth)
ISBN-13 978-0-4070-04070-6 (v. 5 : cloth)
ISBN-10 0-471-04070-X (v. 5 : cloth)

1. Medical instruments and apparatus–Encyclopedias. 2. Biomedical engineering–Encyclopedias. 3. Medical physics–Encyclopedias. 4. Medicine–Data processing–Encyclopedias. I. Webster, John G., 1932- . II. Title: Encyclopedia of medical devices and instrumentation.

[DNLM: 1. Equipment and Supplies–Encyclopedias–English. W 13

E555 2006]
R856.A3E53 2006
610.2803–dc22
2005028946

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTRIBUTOR LIST

ABDEL HADY, MAZEN, *McMaster University, Hamilton, Ontario Canada*, Bladder Dysfunction, Neurostimulation of

ABEL, L.A., *University of Melbourne, Melbourne, Australia*, Ocular Motility Recording and Nystagmus

ABREU, BEATRIZ C., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

ALEXANDER, A.L., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

ALI, ABBAS, *University of Illinois, at Urbana-Champaign*, Bioinformatics

ALI, MÜFTÜ, *School of Dental Medicine, Boston, Massachusetts*, Tooth and Jaw, Biomechanics of

ALPERIN, NOAM, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

ANSON, DENIS, *College Misericordia, Dallas, Pennsylvania*, Environmental Control

ARENA, JOHN C., *VA Medical Center and Medical College of Georgia*, Biofeedback

ARIEL, GIDEON, *Ariel Dynamics, Canyon, California*, Biomechanics of Exercise Fitness

ARMSTRONG, STEVE, *University of Iowa, Iowa City, Iowa*, Biomaterials for Dentistry

ASPDEN, R.M., *University of Aberdeen, Aberdeen, United Kingdom*, Ligament and Tendon, Properties of

AUBIN, C.E., *Polytechniquie Montreal, Montreal Quebec, Canada*, Scoliosis, Biomechanics of

AYRES, VIRGINIA M., *Michigan State University, East Lansing, Michigan*, Microscopy, Scanning Tunneling

AZANGWE, G., Ligament and Tendon, Properties of

BACK, LLOYD H., *California Institute of Technology, Pasadena, California*, Coronary Angioplasty and Guidewire Diagnostics

BADYLAK, STEPHEN F., *McGowan Institute for Regenerative Medicine, Pittsburgh, Pennsylvania*, Sterilization of Biologic Scaffold Materials

BANDYOPADHYAY, AMIT, *Washington State University, Pullman, Washington*, Orthopedic Devices, Materials and Design for

BANERJEE, RUPAK K., *University of Cincinnati, Cincinnati, Ohio*, Coronary Angioplasty and Guidewire Diagnostics

BARBOUR, RANDALL L., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

BARKER, STEVEN J., *University of Arizona, Tucson, Arizona*, Oxygen Monitoring

BARTH, ROLF F., *The Ohio State University, Columbus, Ohio*, Boron Neutron Capture Therapy

BECCHETTI, F.D., *University of Michigan, Ann Arbor, Michigan*, Radiotherapy, Heavy Ion

BELFORTE, GUIDO, *Politecnico di Torino – Department of Mechanics*, Laryngeal Prosthetic Devices

BENKESER, PAUL, *Georgia Institute of Technology, Atlanta, Georgia*, Biomedical Engineering Education

BENNETT, JAMES R., *University of Iowa, Iowa City, Iowa*, Digital Angiography

BERSANO-BEGEY, TOMMASO, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

BIGGS, PETER J., *Harvard Medical School, Boston, Massachusetts*, Radiotherapy, Intraoperative

BIYANI, ASHOK, *University of Toledo, and Medical College of Ohio, Toledo, Ohio*, Human Spine, Biomechanics of

BLOCK, W.F., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

BLUE, THOMAS E., *The Ohio State University, Columbus, Ohio*, Boron Neutron Capture Therapy

BLUMSACK, JUDITH T., *Disorders Auburn University, Auburn, Alabama*, Audiometry

BOGAN, RICHARD K., *University of South Carolina, Columbia, South Carolina*, Sleep Laboratory

BOKROS, JACK C., *Medical Carbon Research Institute, Austin, Texas*, Biomaterials, Carbon

BONGIOANNINI, GUIDO, *ENT Division Mauriziano Hospital, Torino, Italy*, Laryngeal Prosthetic Devices

BORAH, JOSHUA, *Applied Science Laboratories, Bedford, Massachusetts*, Eye Movement, Measurement Techniques for

BORDEN, MARK, *Director of Biomaterials Research, Irvine, California*, Biomaterials, Absorbable

BORTON, BETTIE B., *Auburn University Montgomery, Montgomery, Alabama*, Audiometry

BORTON, THOMAS E., *Auburn University Montgomery, Montgomery, Alabama*, Audiometry

BOSE SUSMITA,, *Washington State University, Pullman, Washington*, Orthopedic Devices, Materials and Design for

BOVA, FRANK J., *M. D. Anderson Cancer Center Orlando, Orlando, FL*, Radiosurgery, Stereotactic

BRENNER, DAVID J., *Columbia University Medical Center, New York, New York*, Computed Tomography Screening

BREWER, JOHN M., *University of Georgia*, Electrophoresis

BRIAN, L. DAVIS, *Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, Ohio*, Skin, Biomechanics of

BRITT, L.D., *Eastern Virginia Medical School, Norfolk, Virginia*, Gastrointestinal Hemorrhage

BRITT, R.C., *Eastern Virginia Medical School, Norfolk, Virginia*, Gastrointestinal Hemorrhage

BROZIK, SUSAN M., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

BRUNER, JOSEPH P., *Vanderbilt University Medical Center, Nashville, Tennessee*, Intrauterine Surgical Techniques

BRUNSWIG NEWRING, KIRK A., *University of Nevada, Reno, Nevada*, Sexual Instrumentatio n

BRUYANT, PHILIPPE P., *University of Massachusetts, North Worcester, Massachusetts*, Nuclear Medicine, Computers in

BUNNELL, BERT J., *Bunnell Inc., Salt Lake City, Utah*, High Frequency Ventilation

CALKINS, JERRY M., *Defense Research Technologies, Inc., Rockville, Maryland*, Medical Gas Analyzers

CANNON, MARK, *Northwestern University, Chicago, Illinois*, Resin-Based Composites

CAPPELLERI, JOSEPH C., *Pfizer Inc., Groton, Connecticut*, Quality-of-Life Measures, Clinical Significance of

CARDOSO, JORGE, *University of Madeira, Funchal, Portugal*, Office Automation Systems

CARELLO, MASSIMILIANA, *Politecnicodi Torino – Department of Mechanics*, Laryngeal Prosthetic Devices

CASKEY, THOMAS C., *Cogene Biotech Ventures, Houston, Texas*, Polymerase Chain Reaction

CECCIO, STEVEN, *University of Michigan, Ann Arbor, Michigan*, Heart Valve Prostheses, In Vitro Flow Dynamics of

CHAN, JACKIE K., *Columbia University, New York, New York*, Photography, Medical

CHANDRAN, K.B., *University of Iowa, Iowa City, Iowa*, Heart Valve Prostheses

CHATZANDROULIS, S., *NTUA, Athens, Attiki, Greece*, Capacitive Microsensors for Biomedical Applications

CHAVEZ, ELIANA, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CHEN, HENRY, *Stanford University, Palo Alto, California*, Exercise Stress Testing

CHEN, JIANDE, *University of Texas Medical Branch, Galveston, Texas*, Electrogastrogram

CHEN, YAN, *Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, Ohio*, Skin, Biomechanics of

CHEYNE, DOUGLAS, *Hospital for Sick Children Research Institute*, Biomagnetism

CHUI, CHEN-SHOU, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

CLAXTON, NATHAN S., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

CODERRE, JEFFREY A., *Massachus etts Institute of Technology, Cambridge, Massachusetts*, Boron Neutron Capture Therapy

COLLINS, BETH, *University of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

COLLINS, DIANE, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CONSTANTINOU, C., *Columbia University Radiation Oncology, New York, New York*, Phantom Materials in Radiology

COOK, ALBERT, *University of Alberta*, *Edmonton, Alberta, Canada*, Communication Devices

COOPER, RORY, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CORK, RANDALL C., *Louisiana State University, Shreveport, Louisiana*, Monitoring, Umbilical Artery and Vein, Blood Gas Measurements; Transcuta neous Electrical Nerve Stimulation (TENS); Ambulatory Monitoring

COX, JOSEPHINE H., *Walter Reed Army Institute of Research, Rockville, Maryland*, Blood Collection and Processing

CRAIG, LEONARD, *Feinberg School of Medicine of Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

CRESS, CYNTHIA J., *University of Nebraska, Lincoln, Nebraska*, Communicative Disorders, Computer Applications for

CUMMING, DAVID R.S., *University of Glasgow, Glasgow, United Kingdom*, Ion-Sensitive Field-Effect Transistors

CUNNINGHAM, JOHN R., *Camrose, Alberta, Canada*, Cobalt 60 Units for Radiotherapy

D'ALESSANDRO, DAVID, *Montefiore Medical Center, Bronx, New York*, Heart–Lung Machines

D'AMBRA, MICHAEL N., *Harvard Medical School, Cambridge, Massachusetts*, Cardiac Output, Thermodilution Measurement of

DADSETAN, MAHROKH, *Mayo Clinic, College of Medicine, Rochester, Minnesota*, Microscopy, Electron

DALEY, MICHAEL L., *The University of Memphis, Memphis, Tennessee*, Monitoring, Intracranial Pressure

DAN, LOYD, *Linköping University, Linköping, Sweden*, Thermocouples

DAS, RUPAK, *University of Wisconsin, Madison, Wisconsin*, Brachytherapy, High Dosage Rate

DATTAWADKAR, AMRUTA M., *University of Wisconsin, Madison, Madison, Wisconsin*, Ocular Fundus Reflectometry

DAVIDSON, MICHAEL W., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

DE LUCA, CARLO, *Boston University, Boston, Massachusetts*, Electromyography

DE SALLES, ANTONIO A.F., *UCLA Medical School, Los Angeles, California*, Stereotactic Surgery

DECAU, SABIN, *University of Maryland, School of Medicine*, Shock, Treatment of

DECHOW, PAUL C., *A & M University Health Science Center, Dallas, Texas*, Strain Gages

DELBEKE, JEAN, *Catholique University of Louvain, Brussels, Belgium*, Visual Prostheses

DELL'OSSO, LOUIS F., *Case Western Reserve University, Cleveland, Ohio*, Ocular Motility Recording and Nystagmus

DELORME, ARNAUD, *University of San Diego, La Jolla, California*, Statistical Methods

DEMENKOFF, JOHN, *Mayo Clinic, Scottsdale, Arizona*, Pulmonary Physiology

DEMIR, SEMAHAT S., *The University of Memphis and The University of Tennessee Health Science Center, Memphis, Tennessee*, Electrophysiology

DEMLING, ROBERT H., *Harvard Medical School*, Skin Substitute for Burns, Bioactive

DENNIS, MICHAEL J., *Medical University of Ohio, Toledo, Ohio*, Computed Tomography

DESANTI, LESLIE, *Harvard Medical School*, Skin Substitute for Burns, Bioactive

DEUTSCH, STEVEN, *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

DEVINENI, TRISHUL, *Conemaugh Health System*, Biofeedback

DI BELLA EDWARD, V.R., *University of Utah*, Tracer Kinetics

DIAKIDES, NICHOLAS A., *Advanced Concepts Analysis, Inc., Falls Church, Virginia*, Thermography

DOLAN, PATRICIA L., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

DONOVAN, F.M., *University of South Alabama*, Cardiac Output, Indicator Dilution Measurement of

DOUGLAS, WILSON R., *Children's Hospital of Philadelphia, Philadelphia, Pennsylvania*, Intrauterine Surgical Techniques

DRAPER, CRISSA, *University of Nevada, Reno, Nevada*, Sexual Instrumentation

DRZEWIECKI, TADEUSZ M., *Defense Research Technologies, Inc., Rockville, Maryland*, Medical Gas Analyzers

DURFEE, W.K., *University of Minnesota, Minneapolis, Minnesota*, Rehabilitation and Muscle Testing

DYRO, JOSEPH F., *Setauket, New York*, Safety Program, Hospital

DYSON, MARY, *Herts, United Kingdom*, Heat and Cold, Therapeutic

ECKERLE, JOSEPH S., *SRI International, Menlo Park, California*, Tonometry, Arterial

EDWARDS, BENJAMIN, *University of Wisconsin-Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

EDWARDS, THAYNE L., *University of Washington, Seattle, Washington*, Chromatography

EKLUND, ANDERS, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

EL SOLH, ALI A., *Erie County Medical Center, Buffalo, New York*, Sleep Studies, Computer Analysis of

ELMAYERGI, NADER, *McMaster University, Hamilton, Ontario, Canada*, Bladder Dysfunction, Neurostimulation of

ELSHARYDAH, AHMAD, *Louisiana State University, Baton Rouge, Louisiana*, Ambulatory Monitoring; Monitoring, Umbilical Artery and Vein, Blood Gas Measurements

FADDY, STEVEN C., *St. Vincents Hospital, Sydney, Darlinghurst, Australia*, Cardiac Output, Fick Technique for

FAHEY, FREDERIC H., *Childrens Hospital Boston*, Computed Tomography, Single Photon Emission

FAIN, S.B., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

FELDMAN, JEFFREY, *Childrens Hospital of Philadelphia, Philadelphia, Pennsylvania*, Anesthesia Machines

FELLERS, THOMAS J., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

FERRARA, LISA, *Cleveland Clinic Foundation, Cleveland, Ohio*, Human Spine, Biomechanics of

FERRARI, MAURO, *The Ohio State University, Columbus, Ohio*, Drug Delivery Systems

FONTAINE, ARNOLD A., *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

FOUST, MILTON J., JR, *Medical University of South Carolina Psychiatry and Behavioral Sciences, Charleston, South Carolina*, Electroconvulsive Therapy

FRASCO, PETER, *Mayo Clinic Scottsdale, Scottsdale, Arizona*, Temperature Monitoring

FRAZIER, JAMES, *Louisiana State University, Baton Rouge, Louisiana*, Ambulatory Monitoring

FREIESLEBEN DE BLASIO, BIRGITTE, *University of Oslo, Oslo, Norway*, Impedance Spectroscopy

FRESTA, MASSIMO, *University of Catanzaro Magna Græcia, Germaneto (CZ), Italy*, Drug Delivery Systems

FREYTES, DONALD O., *McGowan Institute for Regenerative Medicine, Pittsburgh Pennsylvania*, Sterilization of Biologic Scaffold Materials

FROEHLICHER, VICTOR, *VA Medical Center, Palo Alto, California*, Exercise Stress Testing

FUNG, EDWARD K., *Columbia University, New York, New York*, Photography, Medical

GAGE, ANDREW A., *State University of New York at Buffalo, Buffalo, New York*, Cryosurgery

GAGLIO, PAUL J., *Columbia University College of Physicians and Surgeons*, Liver Transplantation

GARDNER, REED M., *LDS Hospital and Utah University, Salt Lake City, Utah*, Monitoring, Hemodynamic

GEJERMAN, GLEN, *Hackensack University Medical, Hackensack, New Jersey*, Radiation Therapy, Quality Assurance in

GEORGE, MARK S., *Medical University of South Carolina Psychiatry and Behavioral Sciences, Charleston, South Carolina*, Electroconvulsive Therapy

GHARIEB, R.R., *Infinite Biomedical Technologies, Baltimore, Maryland*, Neurological Monitors

GLASGOW, GLENN P., *Loyola University of Chicago, Maywood, Illinois*, Radiation Protection Instrumentation

GLASGOW, GLENN, *University of Wisconsin-Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

GOEL, VIJAY K., *University of Toledo, and Medical College of Ohio, Toledo, Ohio*, Human Spine, Biomechanics of

GOETSCH, STEVEN J., *San Diego Gamma Knife Center, La Jolla, California*, Gamma Knife

GOLDBERG, JAY R., *Marquette University Milwaukee, Wisconsin*, Minimally Invasive Surgery

GOLDBERG, ZELENNA, *Department of Radiation Oncology, Davis, California*, Ionizing Radiation, Biological Effects of

GOPALAKRISHNAKONE, P., *National University of Singapore, Singapore*, Immunologically Sensitive Field-Effect Transistors

GOPAS, JACOB, *Ben Gurion University of the Negev, Beer Sheva, Israel*, Monoclonal Antibodies

GORGULHO, ALESSANDRA, *UCLA Medical School, Los Angeles, California*, Stereotactic Surgery

GOUGH, DAVID A., *University of California, La Jolla, California*, Glucose Sensors

GOUSTOURIDIS, D., *NTUA, Athens, Attiki, Greece*, Capacitive Microsensors for Biomedical Applications

GRABER, HARRY L., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

GRAÇA, M., *Louisiana State University, Baton Rouge, Louisiana*, Boron Neutron Capture Therapy

GRANT, WALTER III, *Baylor College of Medicine, Houston, Texas*, Radiation Therapy, Intensity Modulated

GRAYDEN, EDWARD, *Mayo Health Center, Albertlea, Minnesota*, Cardiopulmonary Resuscitation

GREEN, JORDAN R., *University of Nebraska, Lincoln, Nebraska*, Communicative Disorders, Computer Applications for

HAEMMERICH, DIETER, *Medical University of South Carolina, Charleston, South Carolina*, Tissue Ablation

HAMAM, HABIB, *Université de Moncton, Moncton New Brunswick, Canada*, Lenses, Intraocular

HAMMOND, PAUL A., *University of Glasgow, Glasgow, United Kingdom*, Ion-Sensitive Field-Effect Transistors

HANLEY, JOSEPH, *Hackensack University Medical, Hackensack, New Jersey*, Radiation Therapy, Quality Assurance in

HARLEY, BRENDAN A., *Massachusetts Institute of Technology*, Skin Tissue Engineering for Regeneration

HARPER, JASON C., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

HASMAN, ARIE, *Maastricht, The Netherlands*, Medical Education, Computers in

HASSOUNA, MAGDY, *Toronto Western Hospital, Toronto, Canada*, Bladder Dysfunction, Neurostimulation of

HAYASHI, KOZABURO, *Okayama University of Science, Okayama, Japan*, Arteries, Elastic Properties of

HENCH, LARRY L., *Imperial College London, London, United Kingdom*, Biomaterials: Bioceramics

HETTRICK, DOUGLAS A., *Sr. Principal Scientist Medtronic, Inc., Minneapolis, Minnesota*, Bioimpedance in Cardiovascular Medicine

HIRSCH-KUCHMA, MELISSA, *University of Central Florida NanoScience Technology Center, Orlando, Florida*, Biosurface Engineering

HOLDER, GRAHAM E., *Moorfields Eye Hospital, London, United Kingdom*, Electroretinography

HOLMES, TIMOTHY, *St. Agnes Cancer Center, Baltimore, Maryland*, Tomotherapy

HONEYMAN-BUCK, JANICE C., *University of Florida, Gainesville, Florida*, Radiology Information Systems

HOOPER, BRETT A., *Areté Associates, Arlington, Virginia*, Endoscopes

HORN, BRUCE, *Kaiser Permanente, Los Angeles, California*, X-Rays Production of

HORNER, PATRICIA I., *Biomedical Engineering Society Landover, Maryland*, Medical Engineering Societies and Organizations

HOROWITZ, PAUL M., *University of Texas, San Antonio, Texas*, Fluorescence Measurements

HOU, XIAOLIN, *Risø National Laboratory, Roskilde, Denmark*, Neutron Activation Analysis

HOVORKA, ROMAN, *University of Cambridge, Cambridge, United Kingdom*, Pancreas, Artificial

HUANG, H.K., *University of Southern California*, Teleradiology

HUNT, ALAN J., *University of Michigan, Ann Arbor, Michigan*, Optical Tweezers

HUTTEN, HELMUT, *University of Technology, Graz, Australia*, Impedance Plethysmography

IAIZZO, P.A., *University of Minnesota, Minneapolis, Minnesota*, Rehabilitation and Muscle Testing

IBBOTT, GEOFFREY S., *Anderson Cancer Center, Houston, Texas*, Radiation Dosimetry, Three-Dimensional

INGHAM, E., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

ISIK, CAN, *Syracuse University, Syracuse, New York*, Blood Pressure Measurement

JAMES, SUSAN P., *Colorado State University, Fort Collins, Colorado*, Biomaterials: Polymers

JENSEN, WINNIE, *Aalborg University, Aalborg, Denmark*, Electroneurography

JIN, CHUNMING, *North Carolina State University, Raleigh, North Carolina*, Biomaterials, Corrosion and Wear of

JIN, Z.M., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

JOHNSON, ARTHUR T., *University of Maryland College Park, Maryland*, Medical Engineering Societies and Organizations

JONES, JULIAN R., *Imperial College London, London, United Kingdom*, Biomaterials: Bioceramics

JOSHI, ABHIJEET, *Abbott Spine, Austin, Texas*, Spinal Implants

JUNG, RANU, *Arizona State University, Tempe, Arizona*, Functional Electrical Stimulation

JURISSON, SILVIA S., *University of Missouri Columbia, Missouri*, Radionuclide Production and Radioactive Decay

KAEDING, PATRICIA J., *Godfrey & Kahn S.C., Madison, Wisconsin*, Codes and Regulations: Medical Devices

KAMATH, CELIA C., *Mayo Clinic, Rochester, Minnesota*, Quality-of-Life Measures, Clinical Significance of

KANE, MOLLIE, *Madison, Wisconsin*, Contraceptive Devices

KATHERINE, ANDRIOLE P., *Harvard Medical School, Boston, Massachusetts*, Picture Archiving and Communication Systems

KATSAGGELOS, AGGELOS K., *Northwestern University, Evanston, Illinois*, DNA Sequencing

KATZ, J. LAWRENCE, *University of Missouri-Kansas City, Kansas City, Missouri*, Bone and Teeth, Properties of

KESAVAN, SUNIL, *Akebono Corporation, Farmington Hills, Michigan*, Linear Variable Differential Transformers

KHANG, GILSON, *Chonbuk National University*, Biomaterials: Tissue Engineering and Scaffolds

KHAODHIAR, LALITA, *Harvard Medical School, Boston, Massachusetts*, Cutaneous Blood Flow, Doppler Measurement of

KIM, MOON SUK, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

KIM, YOUNG KON, *Inje University, Kimhae City, Korea*, Alloys, Shape Memory

KINDWALL, ERIC P., *St. Luke's Medical Center, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

KING, MICHAEL A., *University of Massachusetts, North Worcester, Massachusetts*, Nuclear Medicine, Computers in

KLEBE, ROBERT J., *University of Texas, San Antonio, Texas*, Fluorescence Measurements

KLEIN, BURTON, *Burton Klein Associates, Newton, Massachusetts*, Gas and Vacuum Systems, Centrally Piped Medical

KNOPER, STEVEN R., *University of Arizona College of Medicine*, Ventilatory Monitoring

KONTAXAKIS, GEORGE, *Universidad Politécnica de Madrid, Madrid, Spain*, Positron Emission Tomography

KOTTKE-MARCHANT, KANDICE, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Vascular Graft Prosthesis

KRIPFGANS, OLIVER, *University of Michigan, Ann Arbor, Michigan*, Ultrasonic Imaging

KULKARNI, AMOL D., *University of Wisconsin–Madison, Madison, Wisconsin*, Ocular Fundus Reflectometry, Visual Field Testing

KUMARADAS, J. CARL, *Ryerson University, Toronto, Ontario, Canada*, Hyperthermia, Interstitial

KUNICKA, JOLANTA, *Bayer HealthCare LLC, Tarrytown, New York*, Differential Counts, Automated

KWAK, KWANJ JOO, *University of Miami Miller School of Medicine, Miami, Florida*, Microscopy, Scanning Force

LAKES, RODERIC, *University of Wisconsin-Madison*, Bone and Teeth, Properties of

LAKKIREDDY, DHANUNJAYA, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

LARSEN, COBY, *Case Western Reserve University, Cleveland, Ohio*, Vascular Graft Prosthesis

LASTER, BRENDA H., *Ben Gurion University of the Negev, Beer Sheva, Israel*, Monoclonal Antibodies

LATTA, LOREN, *University of Miami, Coral Gables, Florida*, Rehabilitation, Orthotics in

LEDER, RON S., *Universidad Nacional Autonoma de Mexico Mexico, Distrito Federal*, Continuous Positive Airway Pressure

LEE, CHIN, *Harvard Medical School, Boston, Massachusetts*, Radiotherapy Treatment Planning, Optimization of; Hyperthermia, Interstitial

LEE, HAI BANG, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

LEE, SANG JIN, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

LEI, LIU, *Department of General Engineering, Urbana, Illinois*, Bioinformatics

LEI, XING, *Stanford University, Stanford, California*, Radiation Dose Planning, Computer-Aided

LEWIS, MATTHEW C., *Medical College of Wisconsin, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

LI, CHAODI, *University of Notre Dame, Notre Dame, Indiana*, Bone Cement, Acrylic

LI, JONATHAN G., *University of Florida, Gainesville, Florida*, Radiation Dose Planning, Computer-Aided

LI, QIAO, *University of Michigan, Ann Arbor, Michigan*, Immunotherapy

LI, YANBIN, *University of Arkansas, Fayetteville, Arkansas*, Piezoelectric Sensors

LIBOFF, A.R., *Oakland University, Rochester, Michigan*, Bone Ununited Fracture and Spinal Fusion, Electrical Treatment of

LIGAS, JAMES, *University of Connecticut, Farmington, Connecticut*, Respiratory Mechanics and Gas Exchange

LIMOGE, AIME, *The René Descartes University of Paris, Paris, France*, Electroanalgesia, Systemic

LIN, PEI-JAN PAUL, *Beth Israel Deaconess Medical Center, Boston, Massachusets*, Mammography

LIN, ZHIYUE, *University of Kansas Medical Center, Kansas City, Kansas*, Electrogastrogram

LINEAWEAVER, WILLIAM C., *Unive rsity of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

LIPPING, TARMO, *Tampere University of Technology, Pori, Finland*, Monitoring in Anesthesia

LIU, XIAOHUA, *The University of Michigan, Ann Arbor, Michigan*, Polymeric Materials

LLOYD, J.J., *Regional Medical Physics Department, Newcastle-upon-Tyne, United Kingdom*, Ultraviolet Radiation in Medicine

LOEB, ROBERT, *University of Arizona, Tuscon, Arizona*, Anesthesia Machines

LOPES DE MELO, PEDRO, *State University of Rio de Janeiro, Térreo Salas, Maracanã, Thermistors*

LOUDON, ROBERT G., Lung Sounds

LOW, DANIEL A., *Washington University School of Medicine, St. Louis, Missouri*, Radiation Therapy Simulator

LU, LICHUN, *Mayo Clinic, College of Medicine, Rochester, Minnesota*, Microscopy, Electron

LU, ZHENG FENG, *Columbia University, New York, New York*, Screen-Film Systems

LYON, ANDREW W., *University of Calgary, Calgary, Canada*, Flame Atomic Emission Spectrometry and Atomic Absorption Spectrometry

LYON, MARTHA E., *University of Calgary, Calgary, Canada*, Flame Atomic Emission Spectrometry and Atomic Absorption Spectrometry

MA, C-M CHARLIE, *Fox Chase Cancer Center, Philadelphia, Pennsylvania*, X-Ray Therapy Equipment, Low and Medium Energy

MACIA, NARCISO F., *Arizona State University at the Polytechnic Campus, Mesa, Arizona*, Pneumotachometers

MACKENZIE, COLIN F., *University of Maryland, School of Medicine*, Shock, Treatment of

MACKIE, THOMAS R., *University of Wisconsin, Madison, Wisconsin*, Tomotherapy

MADNANI, ANJU, *LSU Medical Centre, Shreveport, Louisiana*, Transcutaneous Electrical Nerve Stimulation (TENS)

MADNANI, SANJAY, *LSU Medical Centre, Shreveport, Louisiana*, Transcutaneous Electrical Nerve Stimulation (TENS)

MADSEN, MARK T., *University of Iowa, Iowa City, Iowa*, Anger Camera

MAGNANO, MAURO, *ENT Division Mauriziano Hospital, Torino, Italy*, Drug Delivery Systems

MANDEL, RICHARD, *Boston University School of Medicine, Boston, Massachusetts*, Colorimetry

MANNING, KEEFE B., *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

MAO, JEREMY J., *University of Illinois at Chicago, Chicago, Illinois*, Cartilage and Meniscus, Properties of

MARCOLONGO, MICHELE, *Drexel University, Philadelphia, Pennsylvania*, Spinal Implants

MAREK, MIROSLAV, *Georgia Institute of Technology, Atlanta, Georgia*, Biomaterials, Corrosion and Wear of

MARION, NICHOLAS W., *University of Illinois at Chicago, Chicago, Illinois*, Cartilage and Meniscus, Properties of

MASTERS, KRISTYN S., *University of Wisconsin, Madison, Wisconsin*, Tissue Engineering

MAUGHAN, RICHARD L., *Hospital of the University of Pennsylvania*, Neutron Beam Therapy

MCADAMS, ERIC, *University of Ulster at Jordanstown, Newtownabbey, Ireland*, Bioelectrodes

MCARTHUR, SALLY L., *University of Sheffield, Sheffield, United Kingdom*, Biomaterials, Surface Properties of

MCEWEN, MALCOM, *National Research Council of Canada, Ontario, Canada*, Radiation Dosimetry for Oncology

MCGOWAN, EDWARD J., *E.J. McGowan & Associates*, Biofeedback

MCGRATH, SUSAN, *Dartmouth College, Hanover, New Hampshire*, Oxygen Analyzers

MEEKS, SANFORD L., *University of Florida, Gainesville, Florida*, Radiosurgery, Stereotactic

MELISSA, PETER, *University of Central Florida NanoScience Technology Center, Orlando, Florida*, Biosurface Engineering

MENDELSON, YITZHAK, *Worcester Polytechnic Institute*, Optical Sensors

METZKER, MICHAEL L., *Baylor College of Medicine, Houston, Texas*, Polymerase Chain Reaction

MEYEREND, M.E., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

MICHLER, ROBERT, *Montefiore Medical Center, Bronx, New York*, Heart–Lung Machines

MICIC, MIODRAG, *MP Biomedicals LLC, Irvine, California*, Microscopy and Spectroscopy, Near-Field

MILLER, WILLIAM, *University of Missouri Columbia, Missouri*, Radionuclide Production and Radioactive Decay

MITTRA, ERIK, *Stony Brook University, New York*, Bone Density Measurement

MODELL, MARK, *Harvard Medical School, Boston, Massachusetts*, Fiber Optics in Medicine

MORE, ROBERT B., *RBMore Associates, Austin, Texas* Biomaterials Carbon

MORE, ROBERT, *Austin, Texas*, Heart Valves, Prosthetic

MORROW, DARREN, *Royal Adelaide Hospital, Adelaide, Australia*, Intraaortic Balloon Pump

MOURTADA, FIRAS, *MD Anderson Cancer Center, Houston, Texas*, Brachytherapy, Intravascular

MOY, VINCENT T., *University of Miami, Miller School of Medicine, Miami, Florida*, Microscopy, Scanning Force

MÜFTÜ, SINAN, *Northeastern University, Boston, Massachusetts*, Tooth and Jaw, Biomechanics of

MURPHY, RAYMOND L.H., Lung Sounds

MURPHY, WILLIAM L., *University of Wisconsin, Madison, Wisconsin*, Tissue Engineering

MURRAY, ALAN, *Newcastle University Medical Physics, Newcastle upon Tyne, United Kingdom*, Pace makers

MUTIC, SASA, *Washington University School of Medicine, St. Louis, Missouri*, Radiation Therapy Simulator

NARAYAN, ROGER J., *University of North Carolina, Chapel Hill, North Carolina*, Biomaterials, Corrosion and Wear of

NATALE, ANDREA, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

NAZERAN, HOMER, *The University of Texas, El Paso, Texas*, Electrocardiography, Computers in

NEUMAN, MICHAEL R., *Michigan Technological University, Houghton, Houghton, Michigan*, Fetal Monitoring, Neonatal Monitoring

NEUZIL, PAVEL, *Institute of Bioengineering and Nanotechnology, Singapore*, Immunologically Sensitive Field-Effect Transistors

NICKOLOFF, EDWARD L., *Columbia University, New York, New York*, X-Ray Quality Control Program

NIEZGODA, JEFFREY A., *Medical College of Wisconsin, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

NISHIKAWA, ROBERT M., *The University of Chicago, Chicago, Illinois*, Computer-Assisted Detection and Diagnosis

NUTTER, BRIAN, *Texas Tech University, Lubbock, Texas*, Medical Records, Computers in

O'DONOHUE, WILLIAM, *University of Nevada, Reno, Nevada*, Sexual Instrumentation

ORTON, COLIN, *Harper Hospital and Wayne State University, Detroit, Michigan*, Medical Physics Literature

OZCELIK, SELAHATTIN, *Texas A&M University, Kingsville, Texas*, Drug Infusion Systems

PANITCH, ALYSSA, *Arizona State University, Tempe, Arizona*, Biomaterials: An Overview

PAOLINO, DONATELLA, *University of Catanzaro Magna Græcia, Germaneto (CZ), Italy*, Drug Delivery Systems

PAPAIOANNOU, GEORGE, *University of Wisconsin, Milwaukee, Wisconsin*, Joints, Biomechanics of

PARK, GRACE E., *Purdue University, West Lafayette, Indiana*, Porous Materials for Biological Applications

PARMENTER, BRETT A., *State University of New York at Buffalo, Buffalo, New York*, Sleep Studies, Computer Analysis of

PATEL, DIMPI, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

PEARCE, JOHN, *The University of Texas, Austin, Texas*, Electrosurgical Unit (ESU)

PELET, SERGE, *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

PERIASAMY, AMMASI, *University of Virginia, Charlottesville, Virginia*, Cellular Imaging

PERSONS, BARBARA L., *University of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

PIPER, IAN, *The University of Memphis, Memphis, Tennessee*, Monitoring, Intracranial Pressure

POLETTO, CHRISTOPHER J., *National Institutes of Health*, Tactile Stimulation

PREMINGER, GLENN M., *Duke University Medical Center, Durham, North Carolina*, Lithotripsy

PRENDERGAST, PATRICK J., *Trinity College, Dublin, Ireland*, Orthopedics, Prosthesis Fixation for

PREVITE, MICHAEL, *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

PURDY, JAMES A., *UC Davis Medical Center, Sacramento, California*, Radiotherapy Accessories

QI, HAIRONG, *Advanced Concepts Analysis, Inc., Falls Church, Virginia*, Thermography

QIN, YIXIAN, *Stony Brook University, New York*, Bone Density Measurement

QUAN, STUART F., *University of Arizona, Tucson, Arizona*, Ventilatory Monitoring

QUIROGA, RODRIGO QUIAN, *University of Leicester, Leicester, United Kingdom*, Evoked Potentials

RAHAGHI, FARBOD N., *University of California, La Jolla, California*, Glucose Sensors

RAHKO, PETER S., *University of Wisconsin Medical School*, Echocardiography and Doppler Echocardiography

RALPH, LIETO, *University of Wisconsin–Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

RAMANATHAN, LAKSHMI, *Mount Sinai Medical Center*, Analytical Methods, Automated

RAO, SATISH S.C., *University of Iowa College of Medicine, Iowa City, Iowa*, Anorectal Manometry

RAPOPORT, DAVID M., *NYU School of Medicine, New York, New York*, Continuous Positive Airway Pressure

REBELLO, KEITH J., *The Johns Hopkins University Applied Physics Lab, Laurel, Maryland*, Micro surgery

REDDY, NARENDER, *The University of Akron, Akron, Ohio*, Linear Variable Differential Transformers

REN-DIH, SHEU, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

RENGACHARY, SETTI S., *Detroit, Michigan*, Human Spine, Biomechanics of

REPPERGER, DANIEL W., *Wright-Patterson Air Force Base, Dayton, Ohio*, Human Factors in Medical Devices

RITCHEY, ERIC R., *The Ohio State University, Columbus, Ohio*, Contact Lenses

RIVARD, MARK J., *Tufts New England Medical Center, Boston, Massachusetts*, Imaging Devices

ROBERTSON, J. DAVID, *University of Missouri, Columbia, Missouri*, Radionuclide Production and Radioactive Decay

ROTH, BRADLEY J., *Oakland University, Rochester, Michigan*, Defibrillators

ROWE-HORWEGE, R. WANDA, *University of Texas Medical School, Houston, Texas*, Hyperthermia, Systemic

RUMSEY, JOHN W., *University of Central Florida, Orlando, Florida*, Biosurface Engineering

RUTKOWSKI, GREGORY E., *University of Minnesota, Duluth, Minnesota*, Engineered Tissue

SALATA, O.V., *University of Oxford, Oxford, United Kingdom*, Nanoparticles

SAMARAS, THEODOROS, *Aristotle University of Thessaloniki Department of Physics, Thessaloniki, Greece*, Thermometry

SANGOLE, ARCHANA P., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

SARKOZI, LASZLO, *Mount Sinai School of Medicine*, Analytical Methods, Automated

SCHEK, HENRY III, *University of Michigan, Ann Arbor, Michigan*, Optical Tweezers

SCHMITZ, CHRISTOPH H., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

SCHUCKERS, STEPHANIE A.C., *Clarkson University, Potsdam, New York*, Arrhythmia Analysis, Automated

SCOPE, KENNETH, *Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

SCOTT, ADZICK N., *University of Pennsylvania, Philadelphia, Pennsylvania*, Intrauterine Surgical Techniques

SEAL, BRANDON L., *Arizona State University, Tempe, Arizona*, Biomaterials: An Overview

SEALE, GARY, *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

SEGERS, PATRICK, *Ghent University, Belgium*, Hemodynamics

SELIM, MOSTAFA A., *Cleveland Metropolitan General Hospital, Palm Coast, Florida*, Colposcopy

SETHI, ANIL, *Loyola University Medical Center, Maywood, Illinois*, X-Rays: Interaction with Matter

SEVERINGHAUS, JOHN W., *University of California in San Francisco*, CO_2 Electrodes

SHALODI, ABDELWAHAB D., *Cleveland Metropolitan General Hospital, Palm Coast, Florida*, Colposcopy

SHANMUGASUNDARAM, SHOBANA, *New Jersey Institute of Technology, Newark, New Jersey*, Polymeric Materials

SHARD, ALEXANDER G., *University of Sheffield, Sheffield United Kingdom*, Biomaterials, Surface Properties of

SHEN, LI-JIUAN, *National Taiwan University School of Pharmacy, Taipei, Taiwan*, Colorimetry

SHEN, WEI-CHIANG, *University of Southern California School of Pharmacy, Los Angeles, California*, Colorimetry

SHERAR, MICHAEL D., *London Health Sciences Centre and University of Western Ontario, London, Ontario, Canada*, Hyperthermia, Interstitial

SHERMAN, DAVID, *The Johns Hopkins University, Baltimore, Maryland*, Electroencephalography

SHI, DONGLU, *University of Cincinnati, Cincinnati, Ohio*, Biomaterials, Testing and Structural Properties of

SHUCARD, DAVID W.M., *State University of New York at Buffalo, Buffalo, New York*, Sleep Studies, Computer Analysis of

SIEDBAND, MELVIN P., *University of Wisconsin, Madison, Wisconsin*, Image Intensifiers and Fluoroscopy

SILBERMAN, HOWARD, *University of Southern California, Los Angeles, California*, Nutrition, Parenteral

SILVERMAN, GORDON, *Manhattan College*, Computers in the Biomedical Laboratory

SILVERN, DAVID A., *Medical Physics Unit, Rabin Medical Center, Petah Tikva, Israel*, Prostate Seed Implants

SINHA, PIYUSH, *The Ohio State University, Columbus, Ohio*, Drug Delivery Systems

SINHA, ABHIJIT ROY, *University of Cincinnati, Cincinnati, Ohio*, Coronary Angioplasty and Guidewire Diagnostics

SINKJÆR, THOMAS, *Aalborg University, Aalborg, Denmark*, Electroneurography

SLOAN, JEFFREY A., *Mayo Clinic, Rochester, Minnesota*, Quality-of-Life Measures, Clinical Significance of

SO, PETER T.C., *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

SOBOL, WLAD T., *University of Alabama at Birmingham Health System, Birmingham, Alabama*, Nuclear Magnetic Resonance Spectroscopy

SOOD, SANDEEP, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

SPECTOR, MYRON, *Brigham and Women's Hospital, Boston, Massachusetts*, Biocompatibility of Materials

SPELMAN, FRANCIS A., *University of Washington*, Cochlear Prostheses

SRINIVASAN, YESHWANTH, *Texas Tech University, Lubbock, Texas*, Medical Records, Computers in

SRIRAM, NEELAMEGHAM, *University of Buffalo, Buffalo, New York*, Cell Counters, Blood

STARKO, KENTON R., *Point Roberts, Washington*, Physiological Systems Modeling

STARKSCHALL, GEORGE, *The University of Texas*, Radiotherapy, Three-Dimensional Conformal

STAVREV, PAVEL, *Cross Cancer Institute, Edmonton, Alberta, Canada*, Radiotherapy Treatment Planning, Optimization of

STENKEN, JULIE A., *Rensselaer Polytechnic Institute, Troy, New York*, Microdialysis Sampling

STIEFEL, ROBERT, *University of Maryland Medical Center, Baltimore, Maryland*, Equipment Acquisition

STOKES, I.A.F., *Polytechniquie Montreal, Montreal Quebec, Canada*, Scoliosis, Biomechanics of

STONE, M.H., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

SU, XIAO-LI, *BioDetection Instruments LLC, Fayetteville, Arkansas*, Piezoelectric Sensors

SUBHAN, ARIF, *Masterplan Technology Management, Chatsworth, California*, Equipment Maintenance, Biomedical

SWEENEY, JAMES D., *Arizona State University, Tempe, Arizona*, Functional Electrical Stimulation

SZETO, ANDREW Y.J., *San Diego State University, San Diego, California*, Blind and Visually Impaired, Assistive Technology for

TAKAYAMA, SHUICHI, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

TAMUL, PAUL C., *Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

TAMURA, TOSHIYO, *Chiba University School of Engineering, Chiba, Japan*, Home Health Care Devices

TANG, XIANGYANG, *GE Healthcare Technologies, Wankesha, Wisconsin*, Computed Tomography Simulators

TAYLOR, B.C., *The University of Akron, Akron, Ohio*, Cardiac Output, Indicator Dilution Measurement of

TEMPLE, RICHARD O., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

TEN, STANLEY, *Salt Lake City, Utah*, Electroanalgesia, Systemic

TERRY, TERESA M., *Walter Reed Army Institute of Research, Rockville, Maryland*, Blood Collection and Processing

THAKOR, N.V., *Johns Hopkins University, Baltimore, Maryland*, Neurological Monitors

THIERENS, HUBERT M.A., *University of Ghent, Ghent, Belgium*, Radiopharmaceutical Dosimetry

THOMADSEN, BRUCE, *University of Wisconsin–Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

TIPPER, J.L., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

TOGAWA, TATSUO, *Waseda University, Saitama, Japan*, Integrated Circuit Temperature Sensor

TORNAI, MARTIN, *Duke University, Durham, North Carolina*, X-Ray Equipment Design

TRAN-SON-TAY, ROGER, *University of Florida, Gainesville, Florida*, Blood Rheology

YENI, YENER N., *Henry Ford Hospital, Detroit, Michigan*, Joints, Biomechanics of

YLI-HANKALA, ARVI, *Tampere University of Technology, Pori, Finland*, Monitoring in Anesthesia

YOKO, KAMOTANI, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

YOON, KANG JI, *Korea Institute of Science and Technology, Seoul, Korea*, Micropower for Medical Applications

YORKE, ELLEN, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

YOSHIDA, KEN, *Aalborg University, Aalborg, Denmark*, Electroneurography

YOUNGSTEDT, SHAWN D., *University of South Carolina, Columbia, South Carolina*, Sleep Laboratory

YU, YIH-CHOUNG, *Lafayette College, Easton, Pennsylvania*, Blood Pressure, Automatic Control of

ZACHARIAH, EMMANUEL S., *University of Medicine and Dentistry of New Jersey, New Brunswick, New Jersey*, Immunologically Sensitive Field-Effect Transistors

ZAIDER, MARCO, *Memorial Sloan Kettering Cancer Center, New York, New York*, Prostate Seed Implants

ZAPANTA, CONRAD M., *Penn State College of Medicine, Hershey, Pennsylvania*, Heart, Artificial

ZARDENETA, GUSTAVO, *University of Texas, San Antonio, Texas*, Fluorescence Measurements

ZELMANOVIC, DAVID, *Bayer HealthCare LLC, Tarrytown, New York*, Differential Counts, Automated

ZHANG, MIN, *University of Washington, Seattle, Washington*, Biomaterials: Polymers

ZHANG, YI, *University of Buffalo, Buffalo, New York*, Cell Counters, Blood

ZHU, XIAOYUE, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

ZIAIE, BABAK, *Purdue University, W. Lafayette, Indiana*, Biotelemetry

ZIELINSKI, TODD M., *Medtronic, Inc., Minneapolis, Minnesota*, Bioimpedance in Cardiovascular Medicine

ZIESSMAN, HARVEY A., *Johns Hopkins University*, Computed Tomography, Single Photon Emission

PREFACE

This six-volume work is an alphabetically organized compilation of almost 300 articles that describe critical aspects of medical devices and instrumentation.

It is comprehensive. The articles emphasize the contributions of engineering, physics, and computers to each of the general areas of anesthesiology, biomaterials, burns, cardiology, clinical chemistry, clinical engineering, communicative disorders, computers in medicine, critical care medicine, dermatology, dentistry, ear, nose, and throat, emergency medicine, endocrinology, gastroenterology, genetics, geriatrics, gynecology, hematology, heptology, internal medicine, medical physics, microbiology, nephrology, neurology, nutrition, obstetrics, oncology, ophthalmology, orthopedics, pain, pediatrics, peripheral vascular disease, pharmacology, physical therapy, psychiatry, pulmonary medicine, radiology, rehabilitation, surgery, tissue engineering, transducers, and urology.

The discipline is defined through the synthesis of the core knowledge from all the fields encompassed by the application of engineering, physics, and computers to problems in medicine. The articles focus not only on what is now useful but also on what is likely to be useful in future medical applications.

These volumes answer the question, "What are the branches of medicine and how does technology assist each of them?" rather than "What are the branches of technology and how could each be used in medicine?" To keep this work to a manageable length, the practice of medicine that is unassisted by devices, such as the use of drugs to treat disease, has been excluded.

The articles are accessible to the user; each benefits from brevity of condensation instead of what could easily have been a book-length work. The articles are designed not for peers, but rather for workers from related fields who wish to take a first look at what is important in the subject.

The articles are readable. They do not presume a detailed background in the subject, but are designed for any person with a scientific background and an interest in technology. Rather than attempting to teach the basics of physiology or Ohm's law, the articles build on such basic concepts to show how the worlds of life science and physical science meld to produce improved systems. While the ideal reader might be a person with a Master's degree in biomedical engineering or medical physics or an M.D. with a physical science undergraduate degree, much of the material will be of value to others with an interest in this growing field. High school students and hospital patients can skip over more technical areas and still gain much from the descriptive presentations.

The *Encyclopedia of Medical Devices and Instrumentation* is excellent for browsing and searching for those new divergent associations that may advance work in a peripheral field. While it can be used as a reference for facts, the articles are long enough that they can serve as an educational instrument and provide genuine understanding of a subject.

One can use this work just as one would use a dictionary, since the articles are arranged alphabetically by topic. Cross references assist the reader looking for subjects listed under slightly different names. The index at the end leads the reader to all articles containing pertinent information on any subject. Listed on pages xxi to xxx are all the abbreviations and acronyms used in the *Encyclopedia.* Because of the increasing use of SI units in all branches of science, these units are provided throughout the *Encyclopedia* articles as well as on pages xxxi to xxxv in the section on conversion factors and unit symbols.

I owe a great debt to the many people who have contributed to the creation of this work. At John Wiley & Sons, Encyclopedia Editor George Telecki provided the idea and guiding influence to launch the project. Sean Pidgeon was Editorial Director of the project. Assistant Editors Roseann Zappia, Sarah Harrington, and Surlan Murrell handled the myriad details of communication between publisher, editor, authors, and reviewers and stimulated authors and reviewers to meet necessary deadlines.

My own background has been in the electrical aspects of biomedical engineering. I was delighted to have the assistance of the editorial board to develop a comprehensive encyclopedia. David J. Beebe suggested cellular topics such as microfluidics. Jerry M. Calkins assisted in defining the chemically related subjects, such as anesthesiology. Michael R. Neuman suggested subjects related to sensors, such as in his own work—neonatology. Joon B. Park has written extensively on biomaterials and suggested related subjects. Edward S. Sternick provided many suggestions from medical physics. The Editorial Board was instrumental both in defining the list of subjects and in suggesting authors.

This second edition brings the field up to date. It is available on the web at http://www.mrw.interscience.wiley.com/emdi, where articles can be searched simultaneously to provide rapid and comprehensive information on all aspects of medical devices and instrumentation.

JOHN G. WEBSTER
University of Wisconsin, Madison

LIST OF ARTICLES

ABBREVIATIONS AND ACRONYMS

AAMI	Association for the Advancement of Medical Instrumentation
AAPM	American Association of Physicists in Medicine
ABC	Automatic brightness control
ABET	Accreditation board for engineering training
ABG	Arterial blood gases
ABLB	Alternative binaural loudness balance
ABS	Acrylonitrile–butadiene–styrene
ac	Alternating current
AC	Abdominal circumference; Affinity chromatography
ACA	Automated clinical analyzer
ACES	Augmentative communication evaluation system
ACL	Anterior chamber lens
ACLS	Advanced cardiac life support
ACOG	American College of Obstetrics and Gynecology
ACR	American College of Radiology
ACS	American Cancer Society; American College of Surgeons
A/D	Analog-to-digital
ADC	Agar diffusion chambers; Analog-to-digital converter
ADCC	Antibody-dependent cellular cytotoxicity
ADCL	Accredited Dosimetry Calibration Laboratories
ADP	Adenosine diphosphate
A-D-T	Admission, discharge, and transfer
AE	Anion exchange; Auxiliary electrode
AEA	Articulation error analysis
AEB	Activation energy barrier
AEC	Automatic exposure control
AED	Automatic external defibrillator
AEMB	Alliance for Engineering in Medicine and Biology
AES	Auger electron spectroscopy
AESC	American Engineering Standards Committee
AET	Automatic exposure termination
AFO	Ankle-foot orthosis
AGC	Automatic gain control
AHA	American Heart Association
AI	Arterial insufficiency
AICD	Automatic implantable cardiac defibrillator
AID	Agency for International Development
AIDS	Acquired immune deficiency syndrome
AL	Anterior leaflet
ALG	Antilymphocyte globulin
ALS	Advanced life support; Amyotropic lateral sclerosis
ALT	Alanine aminotransferase
ALU	Arithmetic and logic unit
AM	Amplitude modulation
AMA	American Medical Association
amu	Atomic mass units
ANOVA	Analysis of variance
ANSI	American National Standards Institute
AP	Action potential; Alternative pathway; Anteroposterior
APD	Anterioposterior diameter
APL	Adjustable pressure limiting valve; Applied Physics Laboratory
APR	Anatomically programmed radiography
AR	Amplitude reduction; Aortic regurgitation; Autoregressive
Ara-C	Arabinosylcytosine
ARD	Absorption rate density
ARDS	Adult respiratory distress syndrome
ARGUS	Arrhythmia guard system
ARMA	Autoregressive-moving-average model
ARMAX	Autoregressive-moving-average model with external inputs
AS	Aortic stenosis
ASA	American Standards Association
ASCII	American standard code for information interchange
ASD	Antisiphon device
ASHE	American Society for Hospital Engineering
ASTM	American Society for Testing and Materials
AT	Adenosine-thiamide; Anaerobic threshold; Antithrombin
ATA	Atmosphere absolute
ATLS	Advanced trauma life support
ATN	Acute tubular necrosis
ATP	Adenosine triphosphate
ATPD	Ambient temperature pressure dry
ATPS	Ambient temperature pressure saturated
ATR	Attenuated total reflection
AUC	Area under curve
AUMC	Area under moment curve
AV	Atrioventricular
AZT	Azido thymidine
BA	Biliary atresia
BAEP	Brainstem auditory evoked potential
BAPN	Beta-amino-proprionitryl
BAS	Boston anesthesis system
BASO	Basophil
BB	Buffer base
BBT	Basal body temperature

BCC	Body-centered cubic
BCD	Binary-coded decimal
BCG	Ballistocardiogram
BCLS	Basic cardiac life support
BCRU	British Commitee on Radiation Units and Measurements
BDI	Beck depression inventory
BE	Base excess; Binding energy
BET	Brunauer, Emmett, and Teller methods
BH	His bundle
BI	Biological indicators
BIH	Beth Israel Hospital
BIPM	International Bureau of Weights and Measurements
BJT	Bipolar junction transistor
BMDP	Biomedical Programs
BME	Biomedical engineering
BMET	Biomedical equipment technician
BMO	Biomechanically optimized
BMR	Basal metabolic rate
BOL	Beginning of life
BP	Bereitschafts potential; Break point
BR	Polybutadiene
BRM	Biological response modifier
BRS	Bibliographic retrieval services
BSS	Balanced salt solution
BTG	Beta thromboglobulin
BTPS	Body temperature pressure saturated
BUN	Blood urea nitrogen
BW	Body weight
CA	Conductive adhesives
CABG	Coronary artery by-pass grafting
CAD/CAM	Computer-aided design/computer-aided manufacturing
CAD/D	Computer-aided drafting and design
CADD	Central axis depth dose
CAI	Computer assisted instruction; Computer-aided instruction
CAM	Computer-assisted management
cAMP	Cyclic AMP
CAPD	Continuous ambulatory peritoneal dialysis
CAPP	Child amputee prosthetic project
CAT	Computerized axial tomography
CATS	Computer-assisted teaching system; Computerized aphasia treatment system
CAVH	Continuous arteriovenous hemofiltration
CB	Conjugated bilirubin; Coulomb barrier
CBC	Complete blood count
CBF	Cerebral blood flow
CBM	Computer-based management
CBV	Cerebral blood volume
CC	Closing capacity
CCC	Computer Curriculum Company
CCD	Charge-coupled device
CCE	Capacitance contact electrode
CCF	Cross-correlation function
CCL	Cardiac catheterization laboratory
CCM	Critical care medical services
CCPD	Continuous cycling peritoneal dialysis
CCTV	Closed circuit television system
CCU	Coronary care unit; Critical care unit
CD	Current density
CDR	Complimentary determining region
CDRH	Center for Devices and Radiological Health
CEA	Carcinoembryonic antigen
CF	Conversion factor; Cystic fibrosis
CFC	Continuous flow cytometer
CFR	Code of Federal Regulations
CFU	Colony forming units
CGA	Compressed Gas Association
CGPM	General Conference on Weights and Measures
CHO	Carbohydrate
CHO	Chinese hamster ovary
CI	Combination index
CICU	Cardiac intensive care unit
CIF	Contrast improvement factor
CIN	Cervical intraepithelial neoplasia
CK	Creatine kinase
CLAV	Clavicle
CLSA	Computerized language sample analysis
CM	Cardiomyopathy; Code modulation
CMAD	Computer managed articulation diagnosis
CMI	Computer-managed instruction
CMRR	Common mode rejection ratio
CMV	Conventional mechanical ventilation; Cytomegalovirus
CNS	Central nervous system
CNV	Contingent negative variation
CO	Carbon monoxide; Cardiac output
COBAS	Comprehensive Bio-Analysis System
COPD	Chronic obstructive pulmonary disease
COR	Center of rotation
CP	Cerebral palsy; Closing pressure; Creatine phosphate
CPB	Cardiopulmonary bypass
CPET	Cardiac pacemaker electrode tips
CPM	Computerized probe measurements
CPP	Cerebral perfusion pressure; Cryoprecipitated plasma
CPR	Cardiopulmonary resuscitation
cps	Cycles per second
CPU	Central Processing unit
CR	Center of resistance; Conditioned response; Conductive rubber; Creatinine
CRBB	Complete right bundle branch block
CRD	Completely randomized design
CRL	Crown rump length
CRT	Cathode ray tube
CS	Conditioned stimulus; Contrast scale; Crown seat
CSA	Compressed spectral array
CSF	Cerebrospinal fluid
CSI	Chemical shift imaging
CSM	Chemically sensitive membrane
CT	Computed tomography; Computerized tomography
CTI	Cumulative toxicity response index
CV	Closing volume

C.V.	Coefficient of variation
CVA	Cerebral vascular accident
CVP	Central venous pressure
CVR	Cardiovascular resistance
CW	Continuous wave
CWE	Coated wire electrodes
CWRU	Case Western Reserve University
DAC	Digital-to-analog converter
DAS	Data acquisition system
dB	Decibel
DB	Direct body
DBMS	Data base management system
DBS	Deep brain stimulation
dc	Direct current
DCCT	Diabetes control and complications trial
DCP	Distal cavity pressure
DCS	Dorsal column stimulation
DDC	Deck decompression chamber
DDS	Deep diving system
DE	Dispersive electrode
DEN	Device experience network
DERS	Drug exception ordering system
DES	Diffuse esophageal spasm
d.f.	Distribution function
DHCP	Distributed Hospital Computer Program
DHE	Dihematoporphyrin ether
DHEW	Department of Health Education and Welfare
DHHS	Department of Health and Human Services
DHT	Duration of hypothermia
DI	Deionized water
DIC	Displacement current
DIS	Diagnostic interview schedule
DL	Double layer
DLI	Difference lumen for intensity
DM	Delta modulation
DME	Dropping mercury electrode
DN	Donation number
DNA	Deoxyribonucleic acid
DOF	Degree of freedom
DOS	Drug ordering system
DOT-NHTSA	Department of Transportation Highway Traffic Safety Administration
DPB	Differential pencil beam
DPG	Diphosphoglycerate
DQE	Detection quantum efficiency
DRESS	Depth-resolved surface coil spectroscopy
DRG	Diagnosis-related group
DSA	Digital subtraction angiography
DSAR	Differential scatter-air ratio
DSB	Double strand breaks
DSC	Differential scanning calorimetry
D-T	Deuterium-on-tritium
DTA	Differential thermal analysis
d.u.	Density unit
DUR	Duration
DVT	Deep venous thrombosis
EA	Esophageal accelerogram
EB	Electron beam
EBCDIC	Extended binary code decimal interchange code
EBS	Early burn scar
EBV	Epstein–Barr Virus
EC	Ethyl cellulose
ECC	Emergency cardiac care; Extracorporeal circulation
ECCE	Extracapsular cataract extinction
ECD	Electron capture detector
ECG	Electrocardiogram
ECM	Electrochemical machining
ECMO	Extracorporeal membrane oxygenation
ECOD	Extracranial cerebrovascular occlusive disease
ECRI	Emergency Care Research Institute
ECS	Exner's Comprehensive System
ECT	Electroconvulsive shock therapy; Electroconvulsive therapy; Emission computed tomography
EDD	Estimated date of delivery
EDP	Aortic end diastolic pressure
EDTA	Ethylenediaminetetraacetic acid
EDX	Energy dispersive X-ray analysis
EEG	Electroencephalogram
EEI	Electrode electrolyte interface
EELV	End-expiratory lung volume
EER	Electrically evoked response
EF	Ejection fraction
EF	Electric field; Evoked magnetic fields
EFA	Estimated fetal age
EGF	Epidermal growth factor
EGG	Electrogastrogram
EIA	Enzyme immunoassay
EIU	Electrode impedance unbalance
ELF	Extra low frequency
ELGON	Electrical goniometer
ELISA	Enzyme-linked immunosorbent assay
ELS	Energy loss spectroscopy
ELV	Equivalent lung volume
EM	Electromagnetic
EMBS	Engineering in Medicine and Biology Society
emf	Electromotive force
EMG	Electromyogram
EMGE	Integrated electromyogram
EMI	Electromagnetic interference
EMS	Emergency medical services
EMT	Emergency medical technician
ENT	Ear, nose, and throat
EO	Elbow orthosis
EOG	Electrooculography
EOL	End of life
EOS	Eosinophil
EP	Elastoplastic; Evoked potentiate
EPA	Environmental protection agency
ER	Evoked response
ERCP	Endoscopic retrograde cholangiopancreatography
ERG	Electron radiography; Electroretinogram
ERMF	Event-related magnetic field
ERP	Event-related potential
ERV	Expiratory reserve volume

ESCA	Electron spectroscopy for chemical analysis
ESI	Electrode skin impedance
ESRD	End-stage renal disease
esu	Electrostatic unit
ESU	Electrosurgical unit
ESWL	Extracorporeal shock wave lithotripsy
ETO, Eto	Ethylene oxide
ETT	Exercise tolerance testing
EVA	Ethylene vinyl acetate
EVR	Endocardial viability ratio
EW	Extended wear
FAD	Flavin adenine dinucleotide
FARA	Flexible automation random analysis
FBD	Fetal biparietal diameter
FBS	Fetal bovine serum
fcc	Face centered cubic
FCC	Federal Communications Commission
Fct	Fluorocrit
FDA	Food and Drug Administration
FDCA	Food, Drug, and Cosmetic Act
FE	Finite element
FECG	Fetal electrocardiogram
FEF	Forced expiratory flow
FEL	Free electron lasers
FEM	Finite element method
FEP	Fluorinated ethylene propylene
FES	Functional electrical stimulation
FET	Field-effect transistor
FEV	Forced expiratory volume
FFD	Focal spot to film distance
FFT	Fast Fourier transform
FGF	Fresh gas flow
FHR	Fetal heart rate
FIC	Forced inspiratory capacity
FID	Flame ionization detector; Free-induction decay
FIFO	First-in-first-out
FITC	Fluorescent indicator tagged polymer
FL	Femur length
FM	Frequency modulation
FNS	Functional neuromuscular stimulation
FO	Foramen ovale
FO-CRT	Fiber optics cathode ray tube
FP	Fluorescence polarization
FPA	Fibrinopeptide A
FR	Federal Register
FRC	Federal Radiation Council; Functional residual capacity
FSD	Focus-to-surface distance
FTD	Focal spot to tissue-plane distance
FTIR	Fourier transform infrared
FTMS	Fourier transform mass spectrometer
FU	Fluorouracil
FUDR	Floxuridine
FVC	Forced vital capacity
FWHM	Full width at half maximum
FWTM	Full width at tenth maximum
GABA	Gamma amino buteric acid
GAG	Glycosaminoglycan
GBE	Gas-bearing electrodynamometer
GC	Gas chromatography; Guanine-cytosine
GDT	Gas discharge tube
GFR	Glomerular filtration rate
GHb	Glycosylated hemoglobin
GI	Gastrointestinal
GLC	Gas–liquid chromatography
GMV	General minimum variance
GNP	Gross national product
GPC	Giant papillary conjunctivitis
GPH	Gas-permeable hard
GPH-EW	Gas-permeable hard lens extended wear
GPO	Government Printing Office
GSC	Gas-solid chromatography
GSR	Galvanic skin response
GSWD	Generalized spike-wave discharge
HA	Hydroxyapatite
HAM	Helical axis of motion
Hb	Hemoglobin
HBE	His bundle electrogram
HBO	Hyperbaric oxygenation
HC	Head circumference
HCA	Hypothermic circulatory arrest
HCFA	Health care financing administration
HCL	Harvard Cyclotron Laboratory
hcp	Hexagonal close-packed
HCP	Half cell potential
HDPE	High density polyethylene
HECS	Hospital Equipment Control System
HEMS	Hospital Engineering Management System
HEPA	High efficiency particulate air filter
HES	Hydroxyethylstarch
HETP	Height equivalent to a theoretical plate
HF	High-frequency; Heating factor
HFCWO	High-frequency chest wall oscillation
HFER	High-frequency electromagnetic radiation
HFJV	High-frequency jet ventilation
HFO	High-frequency oscillator
HFOV	High-frequency oscillatory ventilation
HFPPV	High-frequency positive pressure ventilation
HFV	High-frequency ventilation
HHS	Department of Health and Human Services
HIBC	Health industry bar code
HIMA	Health Industry Manufacturers Association
HIP	Hydrostatic indifference point
HIS	Hospital information system
HK	Hexokinase
HL	Hearing level
HMBA	Hexamethylene bisacetamide
HMO	Health maintenance organization
HMWPE	High-molecular-weight polyethylene
HOL	Higher-order languages
HP	Heating factor; His-Purkinje
HpD	Hematoporphyrin derivative
HPLC	High-performance liquid chromatography
HPNS	High-pressure neurological syndrome
HPS	His-Purkinje system
HPX	High peroxidase activity

HR	Heart rate; High-resolution
HRNB	Halstead-Reitan Neuropsychological Battery
H/S	Hard/soft
HSA	Human serum albumin
HSG	Hysterosalpingogram
HTCA	Human tumor cloning assay
HTLV	Human T cell lymphotrophic virus
HU	Heat unit; Houndsfield units; Hydroxyurea
HVL	Half value layer
HVR	Hypoxic ventilatory response
HVT	Half-value thickness
IA	Image intensifier assembly; Inominate artery
IABP	Intraaortic balloon pumping
IAEA	International Atomic Energy Agency
IAIMS	Integrated Academic Information Management System
IASP	International Association for the Study of Pain
IC	Inspiratory capacity; Integrated circuit
ICCE	Intracapsular cataract extraction
ICD	Intracervical device
ICDA	International classification of diagnoses
ICL	Ms-clip lens
ICP	Inductively coupled plasma; Intracranial pressure
ICPA	Intracranial pressure amplitude
ICRP	International Commission on Radiological Protection
ICRU	International Commission on Radiological Units and Measurements
ICU	Intensive care unit
ID	Inside diameter
IDDM	Insulin dependent diabetes mellitus
IDE	Investigational device exemption
IDI	Index of inspired gas distribution
I:E	Inspiratory: expiratory
IEC	International Electrotechnical Commission; Ion-exchange chromatography
IEEE	Institute of Electrical and Electronics Engineers
IEP	Individual educational program
BETS	Inelastic electron tunneling spectroscopy
IF	Immunofluorescent
IFIP	International Federation for Information Processing
IFMBE	International Federation for Medical and Biological Engineering
IGFET	Insulated-gate field-effect transistor
IgG	Immunoglobulin G
IgM	Immunoglobulin M
IHP	Inner Helmholtz plane
IHSS	Idiopathic hypertrophic subaortic stenosis
II	Image intensifier
IIIES	Image intensifier input-exposure sensitivity
IM	Intramuscular
IMFET	Immunologically sensitive field-effect transistor
IMIA	International Medical Informatics Association
IMS	Information management system
IMV	Intermittent mandatory ventilation
INF	Interferon
IOL	Intraocular lens
IPC	Ion-pair chromatography
IPD	Intermittent peritoneal dialysis
IPG	Impedance plethysmography
IPI	Interpulse interval
IPPB	Intermittent positive pressure breathing
IPTS	International practical temperature scale
IR	Polyisoprene rubber
IRB	Institutional Review Board
IRBBB	Incomplete right bundle branch block
IRPA	International Radiation Protection Association
IRRAS	Infrared reflection-absorption spectroscopy
IRRS	Infrared reflection spectroscopy
IRS	Internal reflection spectroscopy
IRV	Inspiratory reserve capacity
IS	Image size; Ion-selective
ISC	Infant skin servo control
ISDA	Instantaneous screw displacement axis
ISE	Ion-selective electrode
ISFET	Ion-sensitive field effect transistor
ISIT	Intensified silicon-intensified target tube
ISO	International Organization for Standardization
ISS	Ion scattering spectroscopy
IT	Intrathecal
ITEP	Institute of Theoretical and Experimental Physics
ITEPI	Instantaneous trailing edge pulse impedance
ITLC	Instant thin-layer chromatography
IUD	Intrauterine device
IV	Intravenous
IVC	Inferior vena cava
IVP	Intraventricular pressure
JCAH	Joint Commission on the Accreditation of Hospitals
JND	Just noticeable difference
JRP	Joint replacement prosthesis
KB	Kent bundle
Kerma	Kinetic energy released in unit mass
KO	Knee orthosis
KPM	Kilopond meter
KRPB	Krebs-Ringer physiological buffer
LA	Left arm; Left atrium
LAD	Left anterior descending; Left axis deviation
LAE	Left atrial enlargement
LAK	Lymphokine activated killer
LAL	Limulus amoebocyte lysate
LAN	Local area network
LAP	Left atrial pressure
LAT	Left anterior temporalis
LBBB	Left bundle branch block
LC	Left carotid; Liquid chromatography

LCC	Left coronary cusp
LCD	Liquid crystal display
LDA	Laser Doppler anemometry
LDF	Laser Doppler flowmetry
LDH	Lactate dehydrogenase
LDPE	Low density polyethylene
LEBS	Low-energy brief stimulus
LED	Light-emitting diode
LEED	Low energy electron diffraction
LES	Lower esophageal sphincter
LESP	Lower esophageal sphincter pressure
LET	Linear energy transfer
LF	Low frequency
LH	Luteinizing hormone
LHT	Local hyperthermia
LL	Left leg
LLDPE	Linear low density polyethylene
LLPC	Liquid-liquid partition chromatography
LLW	Low-level waste
LM	Left masseter
LNNB	Luria-Nebraska Neuropsychological Battery
LOS	Length of stay
LP	Late potential; Lumboperitoneal
LPA	Left pulmonary artery
LPC	Linear predictive coding
LPT	Left posterior temporalis
LPV	Left pulmonary veins
LRP	Late receptor potential
LS	Left subclavian
LSC	Liquid-solid adsorption chromatography
LSI	Large scale integrated
LSV	Low-amplitude shear-wave viscoelastometry
LTI	Low temperature isotropic
LUC	Large unstained cells
LV	Left ventricle
LVAD	Left ventricular assist device
LVDT	Linear variable differential transformer
LVEP	Left ventricular ejection period
LVET	Left ventricular ejection time
LVH	Left ventricular hypertrophy
LYMPH	Lymphocyte
MAA	Macroaggregated albumin
MAC	Minimal auditory capabilities
MAN	Manubrium
MAP	Mean airway pressure; Mean arterial pressure
MAST	Military assistance to safety and traffic
MBA	Monoclonal antibody
MBV	Maximum breathing ventilation
MBX	Monitoring branch exchange
MCA	Methyl cryanoacrylate
MCG	Magnetocardiogram
MCI	Motion Control Incorporated
MCMI	Millon Clinical Multiaxial Inventory
MCT	Microcatheter transducer
MCV	Mean corpuscular volume
MDC	Medical diagnostic categories
MDI	Diphenylmethane diisocyanate; Medical Database Informatics
MDP	Mean diastolic aortic pressure
MDR	Medical device reporting
MDS	Multidimensional scaling
ME	Myoelectric
MED	Minimum erythema dose
MEDPAR	Medicare provider analysis and review
MEFV	Maximal expiratory flow volume
MEG	Magnetoencephalography
MeSH	Medline subject heading
METS	Metabolic equivalents
MF	Melamine-formaldehyde
MFP	Magnetic field potential
MGH	Massachusetts General Hospital
MHV	Magnetic heart vector
MI	Myocardial infarction
MIC	Minimum inhibitory concentration
MIFR	Maximum inspiratory flow rate
MINET	Medical Information Network
MIR	Mercury-in-rubber
MIS	Medical information system; Metal-insulator-semiconductor
MIT	Massachusetts Institute of Technology
MIT/BIH	Massachusetts Institute of Technology/ Beth Israel Hospital
MMA	Manual metal arc welding
MMA	Methyl methacrylate
MMECT	Multiple-monitored ECT
MMFR	Maximum midexpiratory flow rate
mm Hg	Millimeters of mercury
MMPI	Minnesota Multiphasic Personality Inventory
MMSE	Minimum mean square error
MO	Membrane oxygenation
MONO	Monocyte
MOSFET	Metal oxide silicon field-effect transistor
MP	Mercaptopurine; Metacarpal-phalangeal
MPD	Maximal permissible dose
MR	Magnetic resonance
MRG	Magnetoretinogram
MRI	Magnetic resonance imaging
MRS	Magnetic resonance spectroscopy
MRT	Mean residence time
MS	Mild steel; Multiple sclerosis
MSR	Magnetically shielded room
MTBF	Mean time between failure
MTF	Modulation transfer function
MTTR	Mean time to repair
MTX	Methotroxate
MUA	Motor unit activity
MUAP	Motor unit action potential
MUAPT	Motor unit action potential train
MUMPI	Missouri University Multi-Plane Imager
MUMPS	Massachusetts General Hospital utility multiuser programming system
MV	Mitral valve
MVO_2	Maximal oxygen uptake
MVTR	Moisture vapor transmission rate
MVV	Maximum voluntary ventilation
MW	Molecular weight

NAA	Neutron activation analysis
NAD	Nicotinamide adenine dinucleotide
NADH	Nicotinamide adenine dinucleotide, reduced form
NADP	Nicotinamide adenine dinucleotide phosphate
NAF	Neutrophil activating factor
NARM	Naturally occurring and accelerator-produced radioactive materials
NBB	Normal buffer base
NBD	Neuromuscular blocking drugs
N-BPC	Normal bonded phase chromatography
NBS	National Bureau of Standards
NCC	Noncoronary cusp
NCCLS	National Committee for Clinical Laboratory Standards; National Committee on Clinical Laboratory Standards
NCRP	National Council on Radiation Protection
NCT	Neutron capture theory
NEEP	Negative end-expiratory pressure
NEMA	National Electrical Manufacturers Association
NEMR	Nonionizing electromagnetic radiation
NEQ	Noise equivalent quanta
NET	Norethisterone
NEUT	Neutrophil
NFPA	National Fire Protection Association
NH	Neonatal hepatitis
NHE	Normal hydrogen electrode
NHLBI	National Heart, Lung, and Blood Institute
NIR	Nonionizing radiation
NIRS	National Institute for Radiologic Science
NK	Natural killer
NMJ	Neuromuscular junction
NMOS	N-type metal oxide silicon
NMR	Nuclear magnetic resonance
NMS	Neuromuscular stimulation
NPH	Normal pressure hydrocephalus
NPL	National Physical Laboratory
NR	Natural rubber
NRC	Nuclear Regulatory Commission
NRZ	Non-return-to-zero
NTC	Negative temperature coefficient
NTIS	National Technical Information Service
NVT	Neutrons versus time
NYHA	New York Heart Association
ob/gyn	Obstetrics and gynecology
OCR	Off-center ratio; Optical character recognition
OCV	Open circuit voltage
OD	Optical density; Outside diameter
ODC	Oxyhemoglobin dissociation curve
ODT	Oxygen delivery truck
ODU	Optical density unit
OER	Oxygen enhancement ratio
OFD	Object to film distance; Occiputo-frontal diameter
OHL	Outer Helmholtz layer
OHP	Outer Helmholtz plane
OIH	Orthoiodohippurate
OPG	Ocular pneumoplethysmography
OR	Operating room
OS	Object of known size; Operating system
OTC	Over the counter
OV	Offset voltage
PA	Posterioanterior; Pulmonary artery; Pulse amplitude
PACS	Picture archiving and communications systems
PAD	Primary afferent depolarization
PAM	Pulse amplitude modulation
PAN	Polyacrylonitrile
PAP	Pulmonary artery pressure
PAR	Photoactivation ratio
PARFR	Program for Applied Research on Fertility Regulation
PARR	Poetanesthesia recovery room
PAS	Photoacoustic spectroscopy
PASG	Pneumatic antishock garment
PBI	Penile brachial index
PBL	Positive beam limitation
PBT	Polybutylene terephthalate
PC	Paper chromatography; Personal computer; Polycarbonate
PCA	Patient controlled analgesia; Principal components factor analysis
PCG	Phonocardiogram
PCI	Physiological cost index
PCL	Polycaprolactone; Posterior chamber lens
PCR	Percent regurgitation
PCRC	Perinatal Clinical Research Center
PCS	Patient care system
PCT	Porphyria cutanea tarda
PCWP	Pulmonary capillary wedge pressure
PD	Peritoneal dialysis; Poly-p-dioxanone; Potential difference; Proportional and derivative
PDD	Percent depth dose; Perinatal Data Directory
PDE	Pregelled disposable electrodes
p.d.f.	Probability density function
PDL	Periodontal ligament
PDM	Pulse duration modulation
PDMSX	Polydimethyl siloxane
PDS	Polydioxanone
PE	Polyethylene
PEEP	Positive end-expiratory pressure
PEFR	Peak expiratory now rate
PEN	Parenteral and enteral nutrition
PEP	Preejection period
PEPPER	Programs examine phonetic find phonological evaluation records
PET	Polyethylene terephthalate; Positron-emission tomography
PEU	Polyetherurethane
PF	Platelet factor
PFA	Phosphonoformic add
PFC	Petrofluorochemical
PFT	Pulmonary function testing
PG	Polyglycolide; Propylene glycol

PGA	Polyglycolic add
PHA	Phytohemagglutinin; Pulse-height analyzer
PHEMA	Poly-2-hydroxyethyl methacrylate
PI	Propidium iodide
PID	Pelvic inflammatory disease; Proportional/integral/derivative
PIP	Peak inspiratory pressure
PL	Posterior leaflet
PLA	Polylactic acid
PLATO	Program Logic for Automated Teaching Operations
PLD	Potentially lethal damage
PLED	Periodic latoralized epileptiform discharge
PLT	Platelet
PM	Papillary muscles; Preventive maintenance
PMA	Polymethyl acrylate
p.m.f.	Probability mass function
PMMA	Polymethyl methacrylate
PMOS	P-type metal oxide silicon
PMP	Patient management problem; Poly(4-methylpentane)
PMT	Photomultiplier tube
PO	Per os
Po_2	Partial pressure of oxygen
POBT	Polyoxybutylene terephthalate
POM	Polyoxymethylene
POMC	Patient order management and communication system
POPRAS	Problem Oriented Perinatal Risk Assessment System
PP	Perfusion pressure; Polyproplyene; Postprandial (after meals)
PPA	Phonemic process analysis
PPF	Plasma protein fraction
PPM	Pulse position modulation
PPSFH	Polymerized phyridoxalated stroma-free hemoglobin
PR	Pattern recognition; Pulse rate
PRBS	Pseudo-random binary signals
PRP	Pulse repetition frequency
PRO	Professional review organization
PROM	Programmable read only memory
PS	Polystyrene
PSA	Pressure-sensitive adhesive
PSF	Point spread function
PSI	Primary skin irritation
PSP	Postsynaptic potential
PSR	Proton spin resonance
PSS	Progressive systemic sclerosis
PT	Plasma thromboplastin
PTB	Patellar tendon bearing orthosis
PTC	Plasma thromboplastin component; Positive temperature coefficient; Pressurized personal transfer capsule
PTCA	Percutaneous transluminal coronary angioplasty
PTFE	Polytetrafluoroethylene
PTT	Partial thromboplastin time
PUL	Percutaneous ultrasonic lithotripsy
PURA	Prolonged ultraviolet-A radiation
PUVA	Psoralens and longwave ultraviolet light photochemotherapy
P/V	Pressure/volume
PVC	Polyvinyl chloride; Premature ventricular contraction
PVI	Pressure–volume index
PW	Pulse wave; Pulse width
PWM	Pulse width modulation
PXE	Pseudo-xanthoma elasticum
QA	Quality assurance
QC	Quality control
R-BPC	Reverse bonded phase chromatography
R/S	Radiopaque-spherical
RA	Respiratory amplitude; Right arm
RAD	Right axis deviation
RAE	Right atrial enlargement
RAM	Random access memory
RAP	Right atrial pressure
RAT	Right anterior temporalis
RB	Right bundle
RBBB	Right bundle branch block
RBC	Red blood cell
RBE	Relative biologic effectiveness
RBF	Rose bengal fecal excretion
RBI	Resting baseline impedance
RCBD	Randomized complete block diagram
rCBF	Regional cerebral blood flow
RCC	Right coronary cusp
RCE	Resistive contact electrode
R&D	Research and development
r.e.	Random experiment
RE	Reference electrode
REM	Rapid eye movement; Return electrode monitor
REMATE	Remote access and telecommunication system
RES	Reticuloendothelial system
RESNA	Rehabilitation Engineering Society of North America
RF	Radio frequency; Radiographic-nuoroscopic
RFI	Radio-frequency interference
RFP	Request for proposal
RFQ	Request for quotation
RH	Relative humidity
RHE	Reversible hydrogen electrode
RIA	Radioimmunoassay
RM	Repetition maximum; Right masseter
RMR	Resting metabolic rate
RMS	Root mean square
RN	Radionuclide
RNCA	Radionuclide cineagiogram
ROI	Regions of interest
ROM	Range of motion; Read only memory
RP	Retinitis pigmentosa
RPA	Right pulmonary artery
RPP	Rate pressure product
RPT	Rapid pull-through technique
RPV	Right pulmonary veins
RQ	Respiratory quotient

RR	Recovery room
RRT	Recovery room time; Right posterior temporalis
RT	Reaction time
RTD	Resistance temperature device
RTT	Revised token test
r.v.	Random variable
RV	Residual volume; Right ventricle
RVH	Right ventricular hypertrophy
RVOT	Right ventricular outflow tract
RZ	Return-to-zero
SA	Sinoatrial; Specific absorption
SACH	Solid-ankle-cushion-heel
SAD	Source-axis distance; Statistical Analysis System
SAINT	System analysis of integrated network of tasks
SAL	Sterility assurance level; Surface averaged lead
SALT	Systematic analysis of language transcripts
SAMI	Socially acceptable monitoring instrument
SAP	Systemic arterial pressure
SAR	Scatter-air ratio; Specific absorption rate
SARA	System for anesthetic and respiratory gas analysis
SBE	Subbacterial endocarditis
SBR	Styrene-butadiene rubbers
SC	Stratum corneum; Subcommittees
SCAP	Right scapula
SCE	Saturated calomel electrode; Sister chromatid exchange
SCI	Spinal cord injury
SCRAD	Sub-Committee on Radiation Dosimetry
SCS	Spinal cord stimulation
SCUBA	Self-contained underwater breathing apparatus
SD	Standard deviation
SDA	Stepwise discriminant analysis
SDS	Sodium dodecyl sulfate
S&E	Safety and effectiveness
SE	Standard error
SEC	Size exclusion chromatography
SEM	Scanning electron microscope; Standard error of the mean
SEP	Somatosensory evoked potential
SEXAFS	Surface extended X-ray absorption fine structure
SF	Surviving fraction
SFD	Source-film distance
SFH	Stroma-free hemoglobin
SFTR	Sagittal frontal transverse rotational
SG	Silica gel
SGF	Silica gel fraction
SGG	Spark gap generator
SGOT	Serum glutamic oxaloacetic transaminase
SGP	Strain gage plethysmography; Stress-generated potential
SHE	Standard hydrogen electrode
SI	Le Système International d'Unités
SEBS	Surgical isolation barrier system
SID	Source to image reception distance
SIMFU	Scanned intensity modulated focused ultrasound
SIMS	Secondary ion mass spectroscopy; System for isometric muscle strength
SISI	Short increment sensitivity index
SL	Surgical lithotomy
SLD	Sublethal damage
SLE	Systemic lupus erythemotodes
SMA	Sequential multiple analyzer
SMAC	Sequential multiple analyzer with computer
SMR	Sensorimotor
S/N	Signal-to-noise
S:N/D	Signal-to-noise ratio per unit dose
SNP	Sodium nitroprusside
SNR	Signal-to-noise ratio
SOA	Sources of artifact
SOAP	Subjective, objective, assessment, plan
SOBP	Spread-out Bragg peak
SP	Skin potential
SPECT	Single photon emission computed tomography
SPL	Sound pressure level
SPRINT	Single photon ring tomograph
SPRT	Standard platinum resistance thermometer
SPSS	Statistical Package for the Social Sciences
SQUID	Superconducting quantum interference device
SQV	Square wave voltammetry
SR	Polysulfide rubbers
SRT	Speech reception threshold
SS	Stainless steel
SSB	Single strand breaks
SSD	Source-to-skin distance; Source-to-surface distance
SSE	Stainless steel electrode
SSEP	Somatosensory evoked potential
SSG	Solid state generator
SSP	Skin stretch potential
SSS	Sick sinus syndrome
STD	Source-tray distance
STI	Systolic time intervals
STP	Standard temperature and pressure
STPD	Standard temperature pressure dry
SV	Stroke volume
SVC	Superior vena cava
SW	Standing wave
TAA	Tumor-associated antigens
TAC	Time-averaged concentration
TAD	Transverse abdominal diameter
TAG	Technical Advisory Group
TAH	Total artificial heart
TAR	Tissue-air ratio
TC	Technical Committees
TCA	Tricarboxylic acid cycle
TCD	Thermal conductivity detector
TCES	Transcutaneous cranial electrical stimulation

TCP	Tricalcium phosphate
TDD	Telecommunication devices for the deaf
TDM	Therapeutic drug monitoring
TE	Test electrode; Thermoplastic elastomers
TEAM	Technology evaluation and acquisition methods
TEM	Transmission electron microscope; Transverse electric and magnetic mode; Transverse electromagnetic mode
TENS	Transcutaneous electrical nerve stimulation
TEP	Tracheoesophageal puncture
TEPA	Triethylenepho-sphoramide
TF	Transmission factor
TFE	Tetrafluorethylene
TI	Totally implantable
TICCIT	Time-shared Interaction Computer-Controlled Information Television
TLC	Thin-layer chromatography; Total lung capacity
TLD	Thermoluminescent dosimetry
TMJ	Temporomandibular joint
TMR	Tissue maximum ratio; Topical magnetic resonance
TNF	Tumor necrosis factor
TOF	Train-of-four
TP	Thermal performance
TPC	Temperature pressure correction
TPD	Triphasic dissociation
TPG	Transvalvular pressure gradient
TPN	Total parenteral nutrition
TR	Temperature rise
tRNA	Transfer RNA
TSH	Thyroid stimulating hormone
TSS	Toxic shock syndrome
TTD	Telephone devices for the deaf
TTI	Tension time index
TTR	Transition temperature range
TTV	Trimming tip version
TTY	Teletypewriter
TUR	Transurethral resection
TURP	Transurethral resections of the prostrate
TV	Television; Tidal volume; Tricuspid valve
TVER	Transscleral visual evoked response
TW	Traveling wave
TxB_2	Thrombozame B^2
TZ	Transformation zone
UES	Upper esophageal sphincter
UP	Urea-formaldehyde
UfflS	University Hospital Information System
UHMW	Ultra high molecular weight
UHMWPE	Ultra high molecular weight polyethylene
UL	Underwriters Laboratory
ULF	Ultralow frequency
ULTI	Ultralow temperature isotropic
UMN	Upper motor neuron
UO	Urinary output
UPTD	Unit pulmonary oxygen toxicity doses
UR	Unconditioned response
US	Ultrasound; Unconditioned stimulus
USNC	United States National Committee
USP	United States Pharmacopeia
UTS	Ultimate tensile strength
UV	Ultraviolet; Umbilical vessel
UVR	Ultraviolet radiation
V/F	Voltage-to-frequency
VA	Veterans Administration
VAS	Visual analog scale
VBA	Vaginal blood volume in arousal
VC	Vital capacity
VCO	Voltage-controlled oscillator
VDT	Video display terminal
VECG	Vectorelectrocardiography
VEP	Visually evoked potential
VF	Ventricular fibrillation
VOP	Venous occlusion plethysmography
VP	Ventriculoperitoneal
VPA	Vaginal pressure pulse in arousal
VPB	Ventricular premature beat
VPR	Volume pressure response
VSD	Ventricular septal defect
VSWR	Voltage standing wave ratio
VT	Ventricular tachycardia
VTG	Vacuum tube generator
VTS	Viewscan text system
VV	Variable version
WAIS-R	Weschler Adult Intelligence Scale-Revised
WAK	Wearable artificial kidney
WAML	Wide-angle mobility light
WBAR	Whole-body autoradiography
WBC	White blood cell
WG	Working Groups
WHO	World Health Organization; Wrist hand orthosis
WLF	Williams-Landel-Ferry
WMR	Work metabolic rate
w/o	Weight percent
WORM	Write once, read many
WPW	Wolff-Parkinson-White
XPS	X-ray photon spectroscopy
XR	Xeroradiograph
YAG	Yttrium aluminum garnet
ZPL	Zero pressure level

CONVERSION FACTORS AND UNIT SYMBOLS

SI UNITS (ADOPTED 1960)

A new system of metric measurement, the International System of Units (abbreviated SI), is being implemented throughout the world. This system is a modernized version of the MKSA (meter, kilogram, second, ampere) system, and its details are published and controlled by an international treaty organization (The International Bureau of Weights and Measures).

SI units are divided into three classes:

Base Units

length	meter[†] (m)
mass[‡]	kilogram (kg)
time	second (s)
electric current	ampere (A)
thermodynamic temperature§	kelvin (K)
amount of substance	mole (mol)
luminous intensity	candela (cd)

Supplementary Units

plane angle	radian (rad)
solid angle	steradian (sr)

Derived Units and Other Acceptable Units

These units are formed by combining base units, supplementary units, and other derived units. Those derived units having special names and symbols are marked with an asterisk (*) in the list below:

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
*absorbed dose	gray	Gy	J/kg
acceleration	meter per second squared	m/s^2	
*activity (of ionizing radiation source)	becquerel	Bq	1/s
area	square kilometer	km^2	
	square hectometer	hm^2	ha (hectare)
	square meter	m^2	

[†]The spellings "metre" and "litre" are preferred by American Society for Testing and Materials (ASTM); however, "−er" will be used in the Encyclopedia.

[‡]"Weight" is the commonly used term for "mass."

§Wide use is made of "Celsius temperature" (t) defined $t = T - T_0$ where T is the thermodynamic temperature, expressed in kelvins, and $T_0 = 273.15\,K$ by definition. A temperature interval may be expressed in degrees Celsius as well as in kelvins.

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
*capacitance	farad	F	C/V
concentration (of amount of substance)	mole per cubic meter	mol/m^3	
*conductance	siemens	S	A/V
current density	ampere per square meter	A/m^2	
density, mass density	kilogram per cubic meter	kg/m^3	g/L; mg/cm^3
dipole moment (quantity)	coulomb meter	C·m	
*electric charge, quantity of electricity	coulomb	C	A·s
electric charge density	coulomb per cubic meter	C/m^3	
electric field strength	volt per meter	V/m	
electric flux density	coulomb per square meter	C/m^2	
*electric potential, potential difference, electromotive force	volt	V	W/A
*electric resistance	ohm	Ω	V/A
*energy, work, quantity of heat	megajoule	MJ	
	kilojoule	kJ	
	joule	J	N·m
	electron volt†	eV†	
	kilowatt hour†	kW·h†	
energy density	joule per cubic meter	J/m^3	
*force	kilonewton	kN	
	newton	N	$kg·m/s^2$
*frequency	megahertz	MHz	
	hertz	Hz	1/s
heat capacity, entropy	joule per kelvin	J/K	
heat capacity (specific), specific entropy	joule per kilogram kelvin	J/(kg·K)	
heat transfer coefficient	watt per square meter kelvin	$W/(m^2·K)$	
*illuminance	lux	lx	lm/m^2
*inductance	henry	H	Wb/A
linear density	kilogram per meter	kg/m	
luminance	candela per square meter	cd/m^2	
*luminous flux	lumen	lm	cd·sr
magnetic field strength	ampere per meter	A/m	
*magnetic flux	weber	Wb	V·s
*magnetic flux density	tesla	T	Wb/m^2
molar energy	joule per mole	J/mol	
molar entropy, molar heat capacity	joule per mole kelvin	J/(mol·K)	
moment of force, torque	newton meter	N·m	
momentum	kilogram meter per second	kg·m/s	
permeability	henry per meter	H/m	
permittivity	farad per meter	F/m	
*power, heat flow rate, radiant flux	kilowatt	kW	
	watt	W	J/s
power density, heat flux density, irradiance	watt per square meter	W/m^2	
*pressure, stress	megapascal	MPa	
	kilopascal	kPa	
	pascal	Pa	N/m^2
sound level	decibel	dB	
specific energy	joule per kilogram	J/kg	
specific volume	cubic meter per kilogram	m^3/kg	
surface tension	newton per meter	N/m	
thermal conductivity	watt per meter kelvin	W/(m·K)	
velocity	meter per second	m/s	
	kilometer per hour	km/h	
viscosity, dynamic	pascal second	Pa·s	
	millipascal second	mPa·s	

†This non-SI unit is recognized as having to be retained because of practical importance or use in specialized fields.

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
viscosity, kinematic	square meter per second	m^2/s	
	square millimeter per second	mm^2/s	
	cubic meter	m^3	
	cubic decimeter	dm^3	L(liter)
	cubic centimeter	cm^3	mL
wave number	1 per meter	m^{-1}	
	1 per centimeter	cm^{-1}	

In addition, there are 16 prefixes used to indicate order of magnitude, as follows:

Multiplication factor	*Prefix*	*Symbol*	*Note*
10^{18}	exa	E	
10^{15}	peta	P	
10^{12}	tera	T	
10^{9}	giga	G	
10^{8}	mega	M	
10^{3}	kilo	k	
10^{2}	hecto	h[a]	[a]Although hecto, deka, deci, and centi are
10	deka	da[a]	SI prefixes, their use should be avoided
10^{-1}	deci	d[a]	except for SI unit-multiples for area and
10^{-2}	centi	c[a]	volume and nontechnical use of
10^{-3}	milli	m	centimeter, as for body and clothing
10^{-6}	micro	μ	measurement.
10^{-9}	nano	n	
10^{-12}	pico	p	
10^{-15}	femto	f	
10^{-18}	atto	a	

For a complete description of SI and its use the reader is referred to ASTM E 380.

CONVERSION FACTORS TO SI UNITS

A representative list of conversion factors from non-SI to SI units is presented herewith. Factors are given to four significant figures. Exact relationships are followed by a dagger (†). A more complete list is given in ASTM E 380-76 and ANSI Z210. 1-1976.

To convert from	*To*	*Multiply by*
acre	square meter (m^2)	4.047×10^{3}
angstrom	meter (m)	$1.0 \times 10^{-10\dagger}$
are	square meter (m^2)	$1.0 \times 10^{2\dagger}$
astronomical unit	meter (m)	1.496×10^{11}
atmosphere	pascal (Pa)	1.013×10^{5}
bar	pascal (Pa)	$1.0 \times 10^{5\dagger}$
barrel (42 U.S. liquid gallons)	cubic meter (m^3)	0.1590
Btu (International Table)	joule (J)	1.055×10^{3}
Btu (mean)	joule (J)	1.056×10^{3}
Bt (thermochemical)	joule (J)	1.054×10^{3}
bushel	cubic meter (m^3)	3.524×10^{-2}
calorie (International Table)	joule (J)	4.187
calorie (mean)	joule (J)	4.190
calorie (thermochemical)	joule (J)	$4.184^{\dagger}$
centimeters of water (39.2 °F)	pascal (Pa)	98.07
centipoise	pascal second (Pa·s)	$1.0 \times 10^{-3\dagger}$
centistokes	square millimeter per second (mm^2/s)	$1.0^{\dagger}$

To convert from	*To*	*Multiply by*
cfm (cubic foot per minute)	cubic meter per second (m^3/s)	4.72×10^{-4}
cubic inch	cubic meter (m^3)	1.639×10^{-4}
cubic foot	cubic meter (m^3)	2.832×10^{-2}
cubic yard	cubic meter (m^3)	0.7646
curie	becquerel (Bq)	$3.70 \times 10^{10\dagger}$
debye	coulomb-meter (C·m)	3.336×10^{-30}
degree (angle)	radian (rad)	1.745×10^{-2}
denier (international)	kilogram per meter (kg/m)	1.111×10^{-7}
	tex	0.1111
dram (apothecaries')	kilogram (kg)	3.888×10^{-3}
dram (avoirdupois)	kilogram (kg)	1.772×10^{-3}
dram (U.S. fluid)	cubic meter (m^3)	3.697×10^{-6}
dyne	newton(N)	$1.0 \times 10^{-6\dagger}$
dyne/cm	newton per meter (N/m)	$1.00 \times 10^{-3\dagger}$
electron volt	joule (J)	1.602×10^{-19}
erg	joule (J)	$1.0 \times 10^{-7\dagger}$
fathom	meter (m)	1.829
fluid ounce (U.S.)	cubic meter (m^3)	2.957×10^{-5}
foot	meter (m)	$0.3048^{\dagger}$
foot-pound force	joule (J)	1.356
foot-pound force	newton meter (N·m)	1.356
foot-pound force per second	watt(W)	1.356
footcandle	lux (lx)	10.76
furlong	meter (m)	2.012×10^2
gal	meter per second squared (m/s^2)	$1.0 \times 10^{-2\dagger}$
gallon (U.S. dry)	cubic meter (m^3)	4.405×10^{-3}
gallon (U.S. liquid)	cubic meter (m^3)	3.785×10^{-3}
gilbert	ampere (A)	0.7958
gill (U.S.)	cubic meter (m^3)	1.183×10^{-4}
grad	radian	1.571×10^{-2}
grain	kilogram (kg)	6.480×10^{-5}
gram force per denier	newton per tex (N/tex)	8.826×10^{-2}
hectare	square meter (m^2)	$1.0 \times 10^{4\dagger}$
horsepower (550 ft·lbf/s)	watt(W)	7.457×10^2
horsepower (boiler)	watt(W)	9.810×10^3
horsepower (electric)	watt(W)	$7.46 \times 10^{2\dagger}$
hundredweight (long)	kilogram (kg)	50.80
hundredweight (short)	kilogram (kg)	45.36
inch	meter (m)	$2.54 \times 10^{-2\dagger}$
inch of mercury (32 °F)	pascal (Pa)	3.386×10^3
inch of water (39.2 °F)	pascal (Pa)	2.491×10^2
kilogram force	newton (N)	9.807
kilopond	newton (N)	9.807
kilopond-meter	newton-meter (N·m)	9.807
kilopond-meter per second	watt (W)	9.807
kilopond-meter per min	watt(W)	0.1635
kilowatt hour	megajoule (MJ)	$3.6^{\dagger}$
kip	newton (N)	4.448×10^2
knot international	meter per second (m/s)	0.5144
lambert	candela per square meter (cd/m^2)	3.183×10^3
league (British nautical)	meter (m)	5.559×10^2
league (statute)	meter (m)	4.828×10^3
light year	meter (m)	9.461×10^{15}
liter (for fluids only)	cubic meter (m^3)	$1.0 \times 10^{-3\dagger}$
maxwell	weber (Wb)	$1.0 \times 10^{-8\dagger}$
micron	meter (m)	$1.0 \times 10^{-6\dagger}$
mil	meter (m)	$2.54 \times 10^{-5\dagger}$
mile (U.S. nautical)	meter (m)	$1.852 \times 10^{3\dagger}$
mile (statute)	meter (m)	1.609×10^3
mile per hour	meter per second (m/s)	0.4470

To convert from	*To*	*Multiply by*
millibar	pascal (Pa)	1.0×10^{2}
millimeter of mercury (0 °C)	pascal (Pa)	1.333×10^{2}†
millimeter of water (39.2 °F)	pascal (Pa)	9.807
minute (angular)	radian	2.909×10^{-4}
myriagram	kilogram (kg)	10
myriameter	kilometer (km)	10
oersted	ampere per meter (A/m)	79.58
ounce (avoirdupois)	kilogram (kg)	2.835×10^{-2}
ounce (troy)	kilogram (kg)	3.110×10^{-2}
ounce (U.S. fluid)	cubic meter (m^3)	2.957×10^{-5}
ounce-force	newton (N)	0.2780
peck (U.S.)	cubic meter (m^3)	8.810×10^{-3}
pennyweight	kilogram (kg)	1.555×10^{-3}
pint (U.S. dry)	cubic meter (m^3)	5.506×10^{-4}
pint (U.S. liquid)	cubic meter (m^3)	4.732×10^{-4}
poise (absolute viscosity)	pascal second (Pa·s)	0.10†
pound (avoirdupois)	kilogram (kg)	0.4536
pound (troy)	kilogram (kg)	0.3732
poundal	newton (N)	0.1383
pound-force	newton (N)	4.448
pound per square inch (psi)	pascal (Pa)	6.895×10^{3}
quart (U.S. dry)	cubic meter (m^3)	1.101×10^{-3}
quart (U.S. liquid)	cubic meter (m^3)	9.464×10^{-4}
quintal	kilogram (kg)	1.0×10^{2}†
rad	gray (Gy)	1.0×10^{-2}†
rod	meter (m)	5.029
roentgen	coulomb per kilogram (C/kg)	2.58×10^{-4}
second (angle)	radian (rad)	4.848×10^{-6}
section	square meter (m^2)	2.590×10^{6}
slug	kilogram (kg)	14.59
spherical candle power	lumen (lm)	12.57
square inch	square meter (m^2)	6.452×10^{-4}
square foot	square meter (m^2)	9.290×10^{-2}
square mile	square meter (m^2)	2.590×10^{6}
square yard	square meter (m^2)	0.8361
store	cubic meter (m^3)	1.0†
stokes (kinematic viscosity)	square meter per second (m^2/s)	1.0×10^{-4}†
tex	kilogram per meter (kg/m)	1.0×10^{-6}†
ton (long, 2240 pounds)	kilogram (kg)	1.016×10^{3}
ton (metric)	kilogram (kg)	1.0×10^{3}†
ton (short, 2000 pounds)	kilogram (kg)	9.072×10^{2}
torr	pascal (Pa)	1.333×10^{2}
unit pole	weber (Wb)	1.257×10^{-7}
yard	meter (m)	0.9144†

N

NANOPARTICLES

O.V. Salata
University of Oxford
Oxford, United Kingdom

INTRODUCTION

Nanoparticle (From a Greek word "ο νάνοζ" meaning "dwarf"), "nano" is a prefix defining a billion times reduction (10^{-9}) in magnitude. Intuitively, a nanoparticle is an object that is a "nano" times smaller than just a particle. The generally accepted size range for the objects to be called "nanoparticles" is between 1 and 100 nm in at least in two dimensions. This broad definition can be narrowed by defining a nanoparticle of a certain material as a particle that is smaller than a critical size for this material (Fig. 1). The critical size is defined as a cross-point size for the transition from bulk to sized-dependent material property or properties, for example, solubility, melting point, magnetization, light absorption, and fluorescence. The actual value of the critical size depends on the strength of the interatomic or intermolecular forces. The weaker the forces, the weaker the overall interactions between the atoms in the particle. It is known from quantum mechanics that an atom or molecule can be characterized by a set of permitted electron energy levels. Those energy levels define physical and chemical properties of atoms and molecules. When two identical atoms are brought into proximity, they would experience both repulsing and attracting forces. An equilibrium distance might exist where the two forces are balanced. A new molecule containing two atoms will have a system of energy levels that is similar to that of the initial atoms, but each level will be split into two sublevels. The energy gap between the two closest sublevels in this molecule will now be smaller than the corresponding smallest energy gap in one atom. Each time a new atom is added, a new set of energy sublevels will be formed. The energy gap will be changing in value with the number of atoms used to build a nanoparticle, and because properties of the nanoparticle will depend on the energy gap, it will also depend on its size. Then millions of the same atoms form a bulk material; these energy sublevels are saturated and broadened into the energy bands that are insensitive to the changes in the number of atoms composing the solid body. Another component affecting the critical size of the nanoparticle is its surface-to-volume ratio. The smaller the size of the particle, the higher the proportion of the total number of atoms or molecules that are positioned at the surface. The surface atoms often have unsaturated or dangling bonds, leading to the high reactivity and catalytic ability of the nanoparticles. It also results in the restructuring and rearrangement of the crystal lattice near the surface. All of this affects the size-dependent properties of nanomaterials.

Nanoparticles can be made out of any material, including metals, ceramics, semiconductors, polymers, and biomolecules. They can possess a complex structure, which might contain a combination of different materials, or have a complex shape. Nanometer-sized objects that include nanoparticles fall in the realm of nanotechnology. Nanotechnology is developing in several directions: nanomaterials, nanodevices, and nanosystems. The nanomaterials/particles level is the most advanced currently, both in scientific knowledge and in commercial applications. Nanobiotechnology is a subfield of nanotechnology that deals with the applications of nanotechnology to biology. Understanding of biological processes on the nanoscale level is a strong driving force behind development of nanobiotechnology. Living organisms are built of cells that are typically more that 10 μm across. However, the cell-forming components are much smaller and are in the submicron size domain. Even smaller are the proteins with a typical size of just 5 nm, which is comparable with the dimensions of the smallest manmade nanoparticles. This simple size comparison gives an idea of using nanoparticles as very small probes (1) that would allow us to spy at the cellular machinery without introducing too much interference. Semiconductor nanoparticles, also known as "quantum dots" (2), show a strong dependence of their physical properties on their size. Just a decade ago, quantum dots were studied because of their size-dependent physical and chemical properties. One of the properties of the semiconductor nanoparticles that are changing with size is the color of their fluorescence, and now they are used as photostable fluorescent probes. As nanoparticles are rapidly taken up by all kinds of cells, they are also used in drug delivery. In pharmacology, the term "nanoparticles" specifically means polymer nanoparticles or, sometimes, submicron particles that carry a drug load (3). This term has been used in drug delivery for more than three decades. At about the same time, magnetic particles with submicron dimensions were employed for the first time to assist with cell separation. However, colloidal gold, which can be alternatively called "a dispersion of gold nanoparticles," has been used in medicine for many decades if not centuries. Colloidal gold tinctures were used by alchemists to treat many illnesses. Colloidal gold was used as a contrast agent by the first optical microscopists as early as the 1600s. In the 1950s, work was started on the use of radioactive colloidal gold as a treatment for cancer (4). When functionalized with antibodies, gold nanoparticles are used to stain cellular organelles or membranes to create markers for the electron microscopy (5). Consequent decoration of the gold markers with silver assists in further signal magnification.

Out of a plethora of size-dependent physical properties available to someone who is interested in the practical side of nanomaterials, optical and magnetic effects are the most used for biological applications. Hybrid bionanomaterials

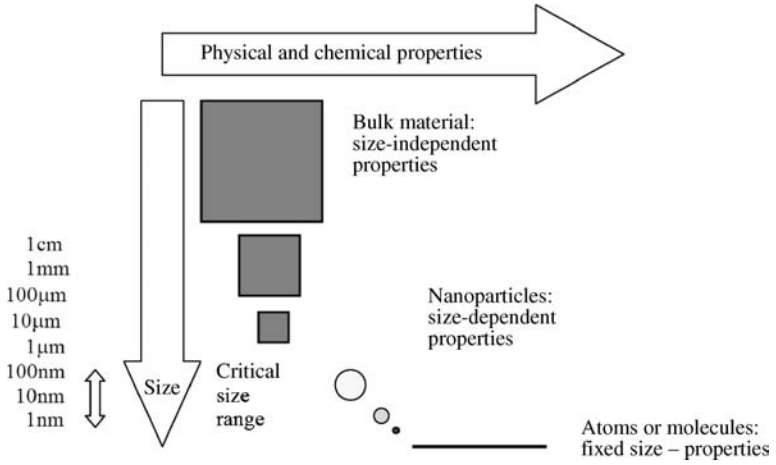


Figure 1. Nanoparticles are a state of matter intermediate between the bulk (size-independent properties) and atomic or molecular (fixed properties) form of the same material with its physical and chemical properties (such as melting temperature, solubility, optical absorbance and fluorescence, magnetization, catalytic activity, and specific chemical reactivity) being dependent on the particle size. The critical size for the transition from the bulk to the nano-regime can vary from tenths to just a few nanometers. A 100 nm threshold is a convenient size to use for a generalized description of the expected bulk-nanoparticle transition.

can also be applied to build novel electronic, optoelectronics, and memory devices.

NANOPARTICLE FABRICATION

Two general ways are available to a manufacturer to produce nanoparticles (Fig. 2). The first way is to start with a bulk material and then break it into smaller pieces using mechanical, chemical, or another form of energy (top-down). An opposite approach is to synthesize the material from atomic or molecular species via physical condensation of atomized materials (energy released), or chemical reactions (exo- or endothermic), allowing the precursor particles to grow in size (bottom-up). Both approaches can be done in gas, liquid, or solid states, or under a vacuum. Both the top-down and the bottom-up processes can be happening during the formation of nanoparticles at the same time, for example, during mechano-chemical ball milling process.

The more detailed classification of the nanoparticle manufacturing techniques relies on the combination of the form of energy with the type of the controlled environment. Each technique has its advantages and disadvantages. Most manufacturers are interested in the ability to control particle size, particle shape, size distribution, particle composition, and degree of particle agglomeration. Absence of contaminants, processing residues or solvents, and sterility are often required in the case of biological and medical applications of nanomaterials. The scale-up of the production volume is also very important. Hence, the discussion of the nanoparticle production techniques is limited to some of those that are currently being pursued by the manufacturers.

Ball Milling

Ball milling is a process where large spheres of the milling media crush substantially smaller powder particles (6). Normally it is used to make fine powder blends or to reduce the degree of powder agglomeration. High-energy ball milling is a more energetic form capable of breaking ceramics into nanoparticles. It is used to create nano-structured composites, highly supersaturated solid solutions, and amorphous phases. The drawbacks of this technique include high energy consumption and poor control over particle sizes. A variation of high-energy ball milling is called mechano-chemical processing. Chemical reactions are mechanically activated during milling, forming nanoparticles via a bottom-up process from suitable precursors. A solid diluent's phase is used to separate the nanoparticles. In the pharmaceutical industry, wet ball milling is often used to produce nano-formulations of the drugs that are poorly soluble in their bulk form but acquire a much improved solubility when turned into nanoparticles.

Electro-Explosion

This process is used to generate 100 nm metal nanoparticles in the form of dry powders. Michael Faraday first observed it in 1773. It involves providing a very high current over a very short time through thin metallic wires, in either an inert or a reactive gas, such that extraordinary temperatures of 20,000 to 30,000 K are achieved. The wire is turned into plasma, contained, and compressed by the

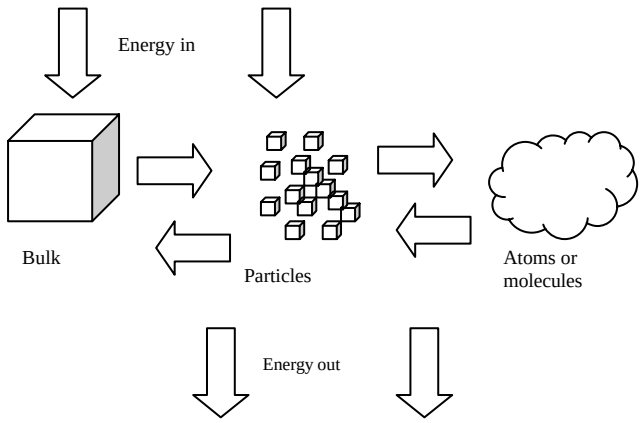


Figure 2. Two basic approaches to nanomaterials fabrication: top-down (shown here from left to the right) and bottom-up (from right to the left). Usually, energy in one of its forms (mechanical, thermal, etc.) is supplied to the bulk matter to create new surfaces. Chemical synthesis of nanomaterials from the atomic or molecular species can be either exothermic or endothermic. Condensation of atoms under vacuum results in cluster formation accompanied by the release of energy.

electromagnetic field. After that, the resistance of the wire becomes so high that the current terminates. The disappearing of the electromagnetic field makes the plasma expand very rapidly. The extremely fast cooling rate results in stabilization of otherwise meta-stable materials. The powders made by electro-explosion have greater purity and reactivity as compared with the ball-milled powders. This technique is used, for example, to produce high surface area nano-porous filtering media benefiting from the enhanced bactericidal properties of silver nanoparticles.

Laser Ablation

In laser ablation, pulsed light from an excimer laser is focused onto a solid target inside a vacuum chamber to supply thermal energy that would "boil off" a plume of energetic atoms of the target material. A substrate positioned to intercept the plume will receive a thin film deposit of the target material. This phenomenon was first observed with a ruby laser in the mid-1960s. Because this process then contaminated the films made with particles, little use was found for such "dirty" samples. The laser ablation method has the following advantages for the fabrication of nanomaterials: The fabrication parameters can be easily changed in a wide range; nanoparticles are naturally produced in the laser ablation plume so that the production rate is relatively high; and virtually all materials can be evaporated by laser ablation.

Colloidal Chemistry

Two general colloidal chemistry approaches are available to control the formation and growth of the nanoparticles. One is called an arrested precipitation; it depends on exhaustion of one of the reactants or on the introduction of the chemical that would block the reaction. Another method relies on a physical restriction of the volume available for the growth of the individual nanoparticles by using various templates (7).

Any material containing regular nanosized pores or voids can be used as a template to form nanoparticles. Examples of such templates include porous alumina, zeolites, di-block copolymers, dendrimers, proteins, and other molecules. The template does not have to be a 3D object. Artificial templates can be created on a plane surface or a gas–liquid interface by forming self-assembled monolayers. The template is usually removed by dissolving it in the solvent that is not affecting the formed nanoparticles. The main advantages of the colloidal chemistry techniques for the preparation of nanomaterials are low temperature of processing, versatility, and flexible rheology. They also offer unique opportunities for preparation of organic–inorganic hybrid materials. The most commonly used precursors for inorganic nanoparticles are oxides and alcoxides.

Aerosols

As an alternative to liquids, chemical reactions can be carried out in a gaseous media, resulting in the formation of nanoparticles aerosols (8). Aerosols can be defined as solid or liquid particles in a gas phase, where the particles can range from molecules up to 100 μm in size. Aerosol generation is driven by the pressure differential created with the help of compressed gases, vacuum, mechanical oscillations, or electrostatic forces acting on liquid. Aerosols were used in industrial manufacturing long before the basic science and engineering of the aerosols were understood. For example, carbon black particles used in pigments and reinforced car tires are produced by hydrocarbon combustion; titania pigment for use in paints and plastics is made by oxidation of titanium tetrachloride; fumed silica and titania formed from respective tetrachlorides by flame pyrolysis; and optical fibers are manufactured by a similar process. Aerosols are also widely used as a drug delivery technique.

Solvent Drying

This technique is frequently used to generate particles of soluble materials (9). Starting materials, for example, a drug and a stabilizing polymer, are dissolved in water-immiscible organic solvent, which is used to prepare an oil-in-water microemulsion. Water can be evaporated by heating under reduced pressure, leaving behind drug-loaded nanoparticles. Both nanospheres (uniform distribution of components) and nanocapsules (polymer encapsulated core) can be created with this method. A monomer can be used instead of the polymer, if the micelle polymerization step is possible. Solvent drying can be achieved via spray-drying step, where a homogeneous solution is fed to an aerosol generator, which produces uniformly sized droplets containing equal amounts of dissolved material. Solvent evaporation from the droplets under the right conditions would result in the formation of nanoparticles with a narrow size distribution.

Electro-Spinning

An emerging technology for the manufacture of thin polymer fibers is based on the principle of spinning dilute polymer solutions in a high-voltage electric field.

Electro-spinning is a process by which a suspended drop of polymer is charged with thousands of volts. At a characteristic voltage, the droplet forms a Taylor cone (the most stable shape with an apex angle of about 57°), and a fine jet of polymer releases from the surface in response to the tensile forces generated by the interaction of an applied electric field with the electrical charge carried by the jet. This produces a bundle of polymer fibers. The jet can be directed to a grounded surface and collected as a continuous web of fibers ranging in size from a few micrometers to less than 100 nm.

Self-Assembly (10)

The appropriate molecular building blocks can act as parts of a jigsaw puzzle that join in a perfect order without obvious driving force present. Various types of chemical bonding can be used to self-assemble nanoparticles. For example, electrostatic interaction between the oppositely charged polymers can be used to build multilayered nanocapsules, the difference in hydrophobicity between the different molecules in the mixture can lead to a formation of a 3D assembly, and proteins can be selected to self-assemble

into virus like nanoparticles. Another example is a use of artificially created oligonucleotides that can be designed to assemble in a variety of shapes and forms.

Nanoparticle Surface Treatment (11)

"Bare" nanoparticles of the same material would rapidly agglomerate with each other, forming bulk material. Encapsulating nanoparticles after production helps to maintain particle size and particle size distribution by inhibiting particle growth that can be caused by evaporation/redeposition, dissolution/reprecipitation, or surface migration and/or flocculation/aggregation/agglomeration. Encapsulation quenches the particle's reactivity and reduces degradation of either the particle or the matrix that surrounds it. The encapsulating coating may be functionalized to facilitate dispersion into organic or aqueous liquid systems. Surface treatment with functional groups enables direct interaction between nanoparticles and resins. Typical functional groups are as follows:

Acrylate
Epoxide
Amine
Vinyl
Isocyanate

The stability of the collected nanoparticle powders against agglomeration, sintering, and compositional changes can be ensured by collecting the nanoparticles in liquid suspension. For semiconductor particles, stabilization of the liquid suspension has been demonstrated by the addition of polar solvent; surfactant molecules have been used to stabilize the liquid suspension of metallic nanoparticles. Alternatively, inert silica encapsulation of nanoparticles by gas-phase reaction and by oxidation in colloidal solution has been shown to be effective for metallic nanoparticles.

When nanosized powders are dispersed in water, they aggregate due to attractive van der Waals forces. By altering the dispersing conditions, repulsive forces can be introduced between the particles to eliminate these aggregates. There are two ways of stabilizing nanoparticles. First, by adjusting the pH of the system, the nanoparticle surface charge can be manipulated such that an electrical double layer is generated around the particle. Overlap of two double layers on different nanoparticles causes repulsion and hence stabilization. The magnitude of this repulsive force can be measured via the zeta potential. The second method involves the adsorption of polymers onto the nanoparticles, such that the particles are physically prevented from coming close enough for the van der Waals attractive force to dominate; this is termed steric stabilization. The combination of two mechanisms is called electrosteric stabilization; it occurs when polyelectrolytes are adsorbed on the nanoparticle surface.

APPLICATIONS (12)

The fact that nanoparticles exist in the same size domain as proteins makes nanomaterials suitable for biotagging and

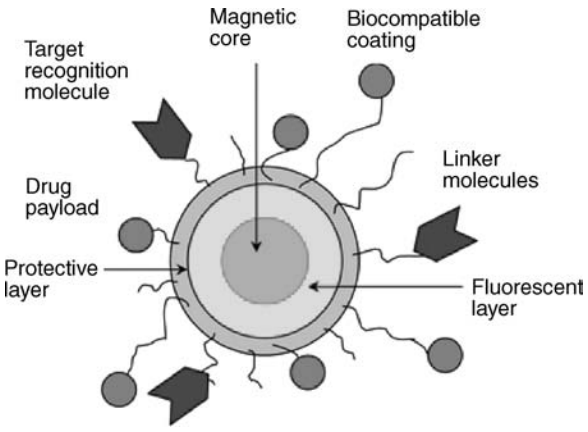


Figure 3. Schematic representation of an example of a biofunctionalized nanoparticle containing a magnetic core coated with a fluorescent layer, which, in turn, is coated by a thin protecting layer (e.g., silica). Linker molecules are attached to the protective layer at one end and to various functional molecules at the other.

payload delivery. However, size is just one of many characteristics of nanoparticles that it is rarely sufficient if one is to use nanoparticles as biological tags. To interact with a biological target, a biological or molecular coating or layer acting as a bioinorganic interface should be attached to the nanoparticle. Examples of biological coatings may include antibodies, biopolymers like collagen, or monolayers of small molecules that make the nanoparticles biocompatible. In addition, as optical detection techniques are widespread in biological research, nanoparticles should either fluoresce or change their optical properties. The approaches used in constructing nano-biomaterials are schematically presented below (Fig. 3). Nanoparticle usually forms the core of nano-biomaterial. It can be used as a convenient surface for molecular assembly and may be composed of inorganic or polymeric materials. It can also be in the form of a nano-vesicle, surrounded by a membrane or a layer. The shape is more often spherical, but cylindrical, plate-like, and other shapes are possible. The size and size distribution might be important in some cases, for example, if penetration through a pore structure of a cellular membrane is required. The size and size distribution are becoming extremely critical when quantum-sized effects are used to control material properties. A tight control of the average particle size and a narrow distribution of sizes allow creation of very efficient fluorescent probes that emit narrow light in a very wide range of wavelengths. This helps with creating biomarkers with many well-distinguished colors. The core might have several layers and be multifunctional. For example, combining magnetic and luminescent layers, one can both detect and manipulate the particles. The core particle is often protected by several monolayers of inert material, for example, silica. Organic molecules that are adsorbed or chemisorbed on the surface of the particle are also used for this purpose. The same layer might act as a biocompatible material. However, more often, an additional layer of linker molecules is required to proceed with further functionalization. This linear linker molecule has reactive

groups at both ends. One group is aimed at attaching the linker to the nanoparticle surface, and the other is used to bind various moieties like biocompatibles (dextran), antibodies, and fluorophores, depending on the function required by the application.

Some current applications of nanomaterials to biology and medicine are listed as follows:

- Biological tags or labels: Mainly gold colloids are used for both electron and light microscopy; also, silver and silver-coated gold nanoparticles are used. A recent addition of semiconductor nanocrystals or quantum dots is finding increasing application as a substitute for the organic fluorophores. Greater photo-stability and the single-wavelength excitation option are among the quoted benefits. Badly designed quantum dots might disintegrate, releasing toxic components (e.g., cadmium or arsenic) that could be lethal at a cellular level. A use of porous silicon nanoparticles as fluorescent tags might be a safer option. Both quantum dots and colloidal gold were used recently for the detection of pathogens, proteins, and for the DNA sequencing. Colloidal gold and, more recently, quantum dots, have also been employed in phagokinetic studies.
- Drug delivery: Mainly use polymeric nanoparticles (13) because of their stability in biological fluids. Flexibility offered by a wide choice of polymers helps to control the rates of drug release and particle biodegradation. Polymeric nanoparticles can be used for all possible administration routes. Surface modifications allow creation of "stealth" as well as targeted drug carriers. Two major approaches are used in their preparation: from polymers (e.g., polyesters) or from monomers (e.g., alkylcyanoacrylate). Either solid nanospheres or liquid core nanocapsules can be produced. The fabrication technologies can be based on the solvent evaporation from oil-in-water microemulsions, created with help of surfactants, or by polymer precipitation caused by the addition of the nonsolvent. Another trend is to produce nanoparticles out of the poorly soluble drugs. Size reduction (nanosizing) (14) can significantly prolong drug bioavailability, increase its dissolution rate and maximum concentration, and dramatically shorten the onset of the drug action.
- MRI contrast enhancement: MRI detects the spatial distribution of the signals from water protons inside the body. These signals depend on the local amount of water and the proton relaxation times T1 and T2. The relaxation times can be affected by several factors; they can also be shortened by the presence of paramagnetic molecules or particles. The T1 shortening would result in the increased signal intensity, whereas the T2 shortening would lead to the opposite effect. These effects are nonlinear functions of the concentration of the contrasting agent.

  Superparamagnetic iron oxide (SPIO), whose average particle size is 50 nm, with dextrane or siloxane coating, is used as a tissue-specific contrast agent (Feridex, Endorem, GastroMARK, Lumirem, Resovist); ultra-small superparamagnetic oxide particles, 10 nm (Sinerem, Combidex), are used to distinguish between the metastatic and inflamed lymph nodes, and to identify arteriosclerosis plaque deposits. Intravenously injected SPIO would pass through the vascular endothelium into the interstitium. After that, the SPIO would be taken up by both healthy and inflamed lymph nodes. The uptake is followed by phagocytosis. Normal lymph nodes show a decrease in signal intensity on T2*- and T2-weighted MR images because of the effects of magnetic susceptibility and T2 shortening on the iron deposits that are a direct result of phagocytosis. However, metastatic lymph nodes are bad at phagocytosis, form no deposits, and do not show such reduction in the signal intensity. This effect can be used to distinguish between the healthy and the benign lymph nodes.
- Separation and purification of biomolecules and cells: Dynabeads are highly uniform in size and superparamagnetic; depending on the antibodies present on the surface, the beads can be applied to different tasks like separation of T-cells, detection of microbes and protozoa, HLA diagnostics for organ transplantation, and various *in vitro* diagnostics assays. They are also used for the isolation of DNA and proteins.
- Tissue engineering: Nanoparticles of hydroxyapatite are used to mimic the mineral particles occurring in the bone structure, whereas collagen is often replaced with a 3D porous scaffolding of a biodegradable polymer. This approach allows for the high mobility of the osteoblasts and, consequently, a uniform growth of the new bone. A similar strategy is used to promote the cellular growth on the surface of prosthetic implants.
- Tumor destruction via heating (hyperthermia): The nanoparticle approach is currently relying on the higher metabolic rates and enhanced blood supply to the tumors. As a result, the cancerous cells are likely to be enriched in a nanoparticulate matter, introduced in the blood circulation, or directly injected into the tumor. An external electromagnetic energy source is directed toward the tumor. The nanoparticles are designed to absorb the electromagnetic energy and convert it into localized heat, which would preferentially cause the apoptosis of the malignant cells. Localized heating might also result in the increased acidosis of the cancer cells. It is also been suggested that the high density of blood vessels in and around the growth stops them from expanding as efficiently as those in the healthy tissue, leading to a higher heat retention. The range of temperatures used is typically within 39 to 43 °C. Alternating magnetic fields can be used to heat up magnetic iron oxide nanoparticles concentrated inside the tumor tissue. More recently developed nanoshells rely on the illumination with a near-infrared laser. Nanoparticles can be designed to actively target the surface receptors on the malignant cells by coating the nanoparticles with the appropriate antibodies.

Recent Developments

Tissue Engineering. Natural bone surface often contains features that are about 100 nm across. If the surface of an artificial bone implant was left smooth, the body would try to reject it. So the smooth surface is likely to cause production of a fibrous tissue covering the surface of the implant. This layer reduces the bone-implant contact, which may result in loosening of the implant and further inflammation. It was demonstrated that by creating nanosized features on the surface of the hip or knee prosthesis, one could reduce the chances of rejection as well as stimulate the production of osteoblasts. The osteoblasts are the cells responsible for the growth of the bone matrix and are found on the advancing surface of the developing bone. The effect was demonstrated with polymeric, ceramic, and more recently, metal materials. More than 90% of the human bone cells from suspension adhered to the nanostructured metal surface, but only 50% did in the control sample. In the end, this finding would allow the design of a more durable and longer lasting hip or knee replacement and reduce the chances of the implant getting loose. Titanium is a well-known bone repairing material widely used in orthopedics and dentistry. It has a high fracture resistance, ductility, and weight-to-strength ratio. Unfortunately, it suffers from the lack of bioactivity, as it does not support cell adhesion and growth well. Apatite coatings are known to be bioactive and to bond to the bone. Hence, several techniques were used to produce an apatite coating on titanium. Those coatings suffer from thickness nonuniformity, poor adhesion, and low mechanical strength. In addition, a stable porous structure is required to support the nutrients' transport through the cell growth. It was shown that using a biomimetic approach — a slow growth of nanostructured apatite film from the simulated body fluid — resulted in the formation of a strongly adherent, uniform nanoporous layer. The layer was found to be built of 60 nm crystallites and possess a stable nanoporous structure and bioactivity. A real bone is a nanocomposite material, composed of hydroxyapatite crystallites in the organic matrix, which is mainly composed of collagen. Thanks to that, the bone is mechanically tough and plastic, so it can recover from mechanical damage. The actual nanoscale mechanism leading to this useful combination of properties is still debated. An artificial hybrid material was prepared from 15 to 18 nm ceramic nanoparticles and poly (methyl methacrylate) copolymer. Using the tribology approach, a viscoelastic behavior (healing) of the human teeth was demonstrated. An investigated hybrid material, deposited as a coating on the tooth surface, improved scratch resistance as well as possessed a healing behavior similar to that of the tooth.

Cancer Therapy. Photodynamic cancer therapy is based on the destruction of the cancer cells by laser-generated atomic oxygen, which is cytotoxic. A greater quantity of a special dye that is used to generate the atomic oxygen is taken in by the cancer cells when compared with a healthy tissue. Hence, only the cancer cells are destroyed and then exposed to a laser radiation. Unfortunately, the remaining dye molecules migrate to the skin and the eyes and make the patient very sensitive to the daylight exposure. This effect can last for up to 6 weeks. To avoid this side effect, the hydrophobic version of the dye molecule was enclosed inside a porous nanoparticle. The dye stayed trapped inside the Ormosil nanoparticle and did not spread to the other parts of the body. At the same time, its oxygen-generating ability has not been affected and the pore size of about 1 nm freely allowed for the oxygen to diffuse out.

Another recently suggested approach is to use gold-coated nanoshells. A surface plasmon resonance effect causes an intense size-dependent absorbance of the near-infrared light by the gold shells. This wavelength of light falls into the optical transparency window of the biological tissue, which can be used to detect cancerous growth up to a certain depth. Compact, powerful, and relatively inexpensive semiconductor lasers are readily available to generate light at this wavelength. The nanoshells are injected into a blood stream and rapidly taken up by the cancerous cells, as they possess a higher metabolism rate. Laser light is absorbed by the gold shells and converted into a local heating, which only kills the cancer cells and spares the healthy tissue. Surface-modified carbon nanotubes can also be used for the same purpose as they would absorb light in the infrared region and convert it into heat.

Multicolor Optical Coding for Biological Assays. The ever increasing research in proteomics and genomic generates an escalating number of sequence data and requires development of high throughput screening technologies. Realistically, various array technologies that are currently used in parallel analysis are likely to reach saturation when several array elements exceed several millions. A three-dimensional approach, based on optical "bar coding" of polymer particles in solution, is limited only by the number of unique tags one can reliably produce and detect. Single quantum dots of compound semiconductors were successfully used as a replacement of organic dyes in various bio-tagging applications. This idea has been taken one step further by combining differently sized and, hence, having different fluorescent colors quantum dots, and combining them in polymeric microbeads. A precise control of quantum dot ratios has been achieved. The selection of nanoparticles used in those experiments had 6 different colors as well as 10 intensities. It is enough to encode over one million combinations. The uniformity and reproducibility of beads was high, allowing for bead identification accuracies of 99.99%.

Manipulation of Cells and Biomolecules. Fictionalized magnetic nanoparticles have found many applications, including cell separation and probing; these and other applications are discussed in a recent review. Most of the magnetic particles studied so far are spherical, which somewhat limits the possibilities to make these nanoparticles multifunctional. Alternative cylindrically shaped nanoparticles can be created by employing metal electrodeposition into a nanoporous alumina template. Depending on the properties of the template, the nanocylinder radius can be selected in the range of 5 to 500 nm while their length can be as big as 60 μm. By sequentially depositing various thicknesses of different metals, the

structure and the magnetic properties of individual cylinders can be tuned widely. As surface chemistry for functionalization of metal surfaces is well developed, different ligands can be selectively attached to different segments. For example, porphyrins with thiol or carboxyl linkers were simultaneously attached to the gold or nickel segments, respectively. Thus, it is possible to produce magnetic nanowires with spatially segregated fluorescent parts. In addition, because of the large aspect ratios, the residual magnetization of these nanowires can be high. Hence, the weaker magnetic field can be used to drive them. It has been shown that a self-assembly of magnetic nanowires in suspension can be controlled by weak external magnetic fields. This would potentially allow controlling cell assembly in different shapes and forms. Moreover, an external magnetic field can be combined with a lithographically defined magnetic pattern ("magnetic trapping").

Protein Detection. Proteins are the important part of the cell's language, machinery, and structure, and understanding their functionalities is extremely important for further progress in human well-being. Gold nanoparticles are widely used in immunohistochemistry to identify the protein–protein interaction. However, the multiple simultaneous detection capabilities of this technique are limited. Surface-enhanced Raman scattering spectroscopy is a well-established technique for detection and identification of single dye molecules. By combining both methods in a single nanoparticle probe, one can drastically improve the multiplexing capabilities of protein probes. The group of Prof. Mirkin has designed a sophisticated multifunctional probe that is built around a 13 nm gold nanoparticle. The nanoparticles are coated with hydrophilic oligonucleotides containing a Raman dye at one end and terminally capped with a small molecule recognition element (e.g., biotin). Moreover, this molecule is catalytically active and will be coated with silver in the solution of Ag(I) and hydroquinone. After the probe is attached to a small molecule or an antigen it is designed to detect, the substrate is exposed to silver and hydroquinone solution. Silver plating is happening close to the Raman dye, which allows for dye signature detection with a standard Raman microscope. Apart from being able to recognize small molecules, this probe can be modified to contain antibodies on the surface to recognize proteins. When tested in the protein array format against both small molecules and proteins, the probe has shown no cross-reactivity.

Commercial Exploration

Some companies involved in the development and commercialization of nanomaterials in biological and medical applications are listed below (Table 1). Most of the companies are small recent spinouts of various research institutions. Although not exhausting, this is a representative selection reflecting current industrial trends. Most companies are developing pharmaceutical applications, mainly for drug delivery. Several companies exploit quantum size effects in semiconductor nanocrystals for tagging biomolecules or use bioconjugated gold nanoparticles for labeling various cellular parts. Many companies are applying nano-ceramic materials to tissue engineering and orthopedics. Most major and established pharmaceutical companies have internal research programs on drug delivery that are on formulations or dispersions containing components down to nanosizes. Colloidal silver is widely used in antimicrobial formulations and dressings. The high reactivity of titania nanoparticles, either on their own or then illuminated with UV light, is also used for bactericidal purposes in filters. Enhanced catalytic properties of surfaces of nano-ceramics or those of noble metals like platinum are used to destruct dangerous toxins and other hazardous organic materials.

Future Directions

As it stands, most commercial nanoparticle applications in medicine are geared toward drug delivery. In the biosciences, nanoparticles are replacing organic dyes in the applications that require high photo-stability as well as high multiplexing capabilities. There are some developments in directing and remotely controlling the functions of nanoprobes, for example, driving magnetic nanoparticles to the tumor and then making them either to release the drug load or just heating them to destroy the surrounding tissue. The major trend in further development of nanomaterials is to make them multifunctional and controllable by external signals or by local environment, thus essentially turning them into nanodevices.

HEALTH ISSUES (15)

It has been shown by several researchers that nanomaterials can enter the human body through several ports. Accidental or involuntary contact during production or use is most likely to happen via the lungs from where a rapid translocation through the blood stream is possible to other vital organs. On the cellular level, an ability to act as a gene vector has been demonstrated for nanoparticles. Carbon black nanoparticles have been implicated in interfering with cell signaling. There is work that demonstrates uses of DNA for the size separation of carbon nanotubes. The DNA strand just wraps around it if the tube diameter is right. Although excellent for separation purposes, it raises some concerns over the consequences of carbon nanotubes entering the human body.

Ports of Entry

Human skin, intestinal tract, and lungs are always in direct contact with the environment. Whereas skin acts as a barrier, lungs and intestinal tract also allow transport (passive and/or active) of various substances like water, nutrients, or oxygen. As a result, they are likely to be a first port of entry for the nanomaterials' journey into the human body. Our knowledge in this field mainly comes from drug delivery (pharmaceutical research) and toxicology (xenobiotics) studies.

Human skin functions as a strict barrier, and no essential elements are taken up through the skin (except radiation necessary to buildup vitamin D). The lungs exchange oxygen and carbon dioxide with the environment, and some water escapes with warm exhaled air. The intestinal tract

Table 1. Examples of Companies Commercializing Nanomaterials for Bio- and Medical Applications

Company	Applications	Technology
Advectus Life Sciences Inc.	Drug delivery	Polymeric nanoparticles engineered to carry anti-tumor drug across the blood-brain barrier
Alnis Biosciences, Inc.	Bio-pharmaceutical	Biodegradable polymeric nanoparticles for drug delivery
Argonide	Membrane filtration, implants	Nanoporous ceramic materials for endotoxin filtration, orthopedic and dental implants, DNA and protein separation
Biophan Technologies, Inc.	MRI shielding, nanomagnetic particles for guided drug delivery and release	Nanomagnetic/carbon composite materials to shield medical devices from RF fields
Capsulution NanoScience AG	Pharmaceutical coatings to improve solubility of drugs	Layer-by-layer poly-electrolyte coatings, 8–50 nm
Dynal Biotech (Invitrogen)	Cell/biomolecule separation	Superparamagnetic beads
Eiffel Technologies	Drug delivery	Reducing size of the drug particles to 50–100 nm
Evident Technologies	Luminescent biomarkers	Semiconductor quantum dots with amine or carboxyl groups on the surface, emission from 350 to 2500 nm
NanoBio Corporation	Pharmaceutical	Antimicrobial nano-emulsions
NanoCarrier Co., Ltd	Drug delivery	Micellar polymer nanoparticles for encapsulation of drugs, proteins, DNA
NanoPharm AG	Drug delivery	Polybutilcyanoacrylate nanoparticles are coated with drugs and then with surfactant, can go across the blood–brain barrier
Nanoprobes, Inc.	Gold nanoparticles for biological markers	Gold nanoparticles bioconjugates for TEM and/or fluorescent microscopy
Nanoshpere, Inc.	Gold biomarkers	DNA barcode attached to each nanoprobe for identification purposes, PCR is used to amplify the signal; also catalytic silver deposition to amplify the signal using surface plasmon resonance
NanoMed Pharmaceutical, Inc.	Drug delivery	Nanoparticles for drug delivery
PSiVida Ltd	Tissue engineering, implants, drugs and gene delivery, biofiltration	Exploiting material properties of nanostructured porous silicone
QuantumDot Corporation	Luminescent biomarkers	Bioconjugated semiconductor quantum dots

is in close contact with all of the materials taken up orally; here all nutrients (except gases) are exchanged between the body and the environment.

The histology of the environmental contact sides of these three organs is significantly different. The skin of an adult human is roughly 1.5 m^2; in area and is at most places covered with a relatively thick first barrier (10 μm), which is built of strongly keratinized dead cells. This first barrier is difficult to pass for ionic compounds as well as for water-soluble molecules.

The lung consists of two different parts: airways (transporting the air in and out the lungs) and alveoli (gas exchange areas). Our two lungs contain about 2,300 km of airways and 300 million alveoli. The surface area of the lungs is 140 m^2 in adults, as big as a tennis court. The airways are a relatively robust barrier, an active epithelium protected with a viscous layer of mucus. In the gas exchange area, the barrier between the alveolar wall and the capillaries is very thin. The air in the lumen of the alveoli is just 0.5 μm away from the blood flow. The large surface area of the alveoli and the intense air–blood contact in this region makes the alveoli less well protected against environmental damage when compared with airways.

The intestinal tract is a more complex barrier and exchange side; it is the most important portal for macromolecules to enter the body. From the stomach, only small molecules can diffuse through the epithelium. The epithelium of the small and large intestines is in close contact with ingested material so that nutrients can be used. A mixture of disaccharides, peptides, fatty acids, and monoglycerides generated by digestion in the small intestine are further transformed and taken in. The area of the gastrointestine tract (G_IT) is estimated as 200 m^2.

Lung. Most nanosized spherical solid materials will easily enter the lungs and reach the alveoli. These particles can be cleared from the lungs, as long as the clearance mechanisms are not affected by the particles or any other cause. Nanosized particles are more likely to hamper the clearance, resulting in a higher burden, possibly amplifying any possible chronic effects caused by these particles. It is also important to note that the specific particle surface area is probably a better indication for maximum tolerated exposure level than total mass. Inhaled nanofibers (diameter smaller than 100 nm) also can enter the alveoli. In addition, their clearing would depend on the length of the specific fiber. Recent publications on the pulmonary effects of carbon nanotubes confirm the intuitive fear that the nanosized fiber can induce a general nonspecific pulmonary response. Passage of solid material from the pulmonary

epithelium to the circulation seems to be restricted to nanoparticles. The issue of particle translocation still needs to be clarified: both the trans-epithelial transport in the alveoli and the transport via nerve cells. Thus, the role of factors governing particle translocation such as the way of exposure, dose, size, surface chemistry, and time course should be investigated. For instance, it would be also very important to know how and to what extent lung inflammation modulates the extrapulmonary translocation of particles. Solid inhaled particles are a risk for those who suffer from cardiovascular disease. Experimental data indicate that probably many inhaled particles can affect cardiovascular parameters, via pulmonary inflammation. Nanosized particles, after passage in the circulation, can also play a direct role in, e.g., thrombogenesis.

Intestinal Tract. Already in 1926, it was recognized by Kumagai that particles could translocate from the lumen of the intestinal tract via aggregations of intestinal lymphatic tissue [Peirel's Patches (PP)] containing M-cells (specialized phagocytic enterocytes). Particulate uptake happens not only via the M-cells in the PP and the isolated follicles of the gut-associated lymphoid tissue, but also via the normal intestinal enterocytes. There have been several excellent reviews on the subject of intestinal uptake of particles. Uptake of inert particles has been shown to occur trans-cellular through normal enterocytes and PP via M-cells and, to a lesser extent, across paracellular pathways. Initially it was assumed that the PP did not discriminate strongly in the type and size of the absorb particles. Later it has been shown that modified characteristics, such as particle size, the surface charge of particles, attachment of ligands, or coating with surfactants, offers possibilities of site-specific targeting to different regions of the GIT including the PP.

The kinetics of particle translocation in the intestine depends on diffusion and accessibility through mucus, initial contact with enterocyte or M-cell, cellular trafficking, and post-translocation events. Cationic nanometer-sized particles became entrapped in the negatively charged mucus, whereas negatively charged nanoparticles can diffuse across this layer. The smaller the particle diameter, the faster they could permutate the mucus to reach the colonic enterocytes. Once in the sub-mucosal tissue, particles can enter both lymphatic and capillaries. Particles entering the lymphatic are probably important in the induction of secretory immune responses, whereas those that enter the capillaries become systemic and can reach different organs. It has been suggested that the disruption of the epithelial barrier function by apoptosis of enterocytes is a possible trigger mechanism for mucosal inflammation. The patho-physiological role of M-cells is unclear; for example, it has been found that in Crohn's disease, M-cells are lost from the epithelium. Diseases other than of gut origin, for example, diabetes, also have marked effects on the ability of the GIT to translocate particles. In general, the intestinal uptake of particles is understood better and studied in more detail than pulmonary and skin uptake. Because of this advantage, it is maybe possible to predict the behavior of some particles in the intestines.

Skin. Skin is an important barrier, protecting against insult from the environment. The skin is structured in three layers: the epidermis, the dermis, and the subcutaneous layer. The outer layer of the epidermis, the stratum corneum (SC), covers the entire outside of the body. In the SC we find only dead cells, which are strongly keratinized. For most chemicals, the SC is the rate-limiting barrier to percutaneous absorption (penetration). The skin of most mammalian species is covered with hair.

At the sites where hair follicles grow, the barrier capacity of the skin differs slightly from the "normal" stratified squamous epidermis. Most studies concerning penetration of materials into the skin have focused on whether drugs penetrate through the skin using different formulations containing chemicals and/or particulate materials as a vehicle. The main types of particulate materials commonly used in contact with skin are liposomes, solid poorly soluble materials such as TiO_2, and polymer particulates and submicron emulsion particles such as solid lipid nanoparticles. TiO_2 particles are often used in sunscreens to absorb UV light and therefore to protect skin against sunburn or genetic damage. It has been reported by Lademann et al. that micrometer-sized particles of TiO_2 get through the human stratum corneum and even into some hair follicles, including their deeper parts. However, the authors did not interpret this observation as penetration into living layers of the skin. In a recent review, it was stated that "very small titanium dioxide particles (e.g. 5–20 nm) penetrate into the skin and can interact with the immune system." Unfortunately, this has not been discussed any further.

Penetration of nonmetallic solid materials such as biodegradable poly(D,L-lactic-co-glycolic acid (PLGA)) microparticles, 1 to 10 μm with a mean diameter of 4.61 ± 0.8 μm, were studied after application on to porcine skin. The number of microparticles in the skin decreased with the depth (measured from the airside toward the subcutaneous layer). At 120 μm depth (where viable dermis is present), a relative high number of particles was found; at 400 μm (dermis), some microparticles were still observed. At a depth of 500 μm, no microparticles were found. In the skin of persons, who had an impaired lymphatic drainage of the lower legs, soil microparticles, frequently 0.4-0.5 μm but as larger particles of 25 μm diameter, were found in the dermis of the foot in a patient with endemic elephantiasis. The particles are observed to be in the phagosomes of macrophages or in the cytoplasm of other cells. The failure to conduct lymph to the node produces a permanent deposit of silica in the dermal tissues (a parallel is drawn with similar deposits in the lung in pneumoconiosis). This indicates that soil particles penetrate through (damaged) skin, most probably in every person, and normally are removed via the lymphatic system.

Liposomes penetrate the skin in a size-dependent manner. Microsized, and even submicron sized, liposomes do not easily penetrate into the viable epidermis, whereas liposomes with an average diameter of 272 nm can reach into the viable epidermis and some are found in the dermis. Smaller sized liposomes of 116 and 71 nm were found in higher concentration in the dermis. Emzaloid particles, a type of submicron emulsion particle such as liposomes and nonionic surfactant vesicles (niosomes), with a diameter of

50 nm to 1 μm, were detected in the epidermis in association with the cell membranes after application to human skin. The authors suggested that single molecules, which make up the particles, might penetrate the intercellular spaces and, at certain regions in the stratum corneum, can accumulate and reform into microspheres. In a subsequent experiment, it was shown that the used formulation allowed penetration of the spheres into melanoma cells, even to the nucleus.

From the limited literature on nanoparticles penetrating the skin, some conclusions can be drawn. First, penetration of the skin barrier is size dependent, and nanosized particles are more likely to enter more deeply into the skin than larger ones. Second, different types of particles are found in the deeper layers of the skin, and currently, it is impossible to predict the behavior of a particle in the skin. Third, materials, which can dissolve or leach from a particle (e.g., metals), or break into smaller parts (e.g., Emzaloid particles), can possibly penetrate into the skin.

Currently, there is no direct indication that particles, that had penetrated the skin also entered the systemic circulation. The observation that particles in the skin can be phagocytized by macrophages, Langerhan cells, or other cells is a possible road toward skin sensitization.

Summary of Health Risks

Particles in the nanosize range can certainly enter the human body via the lungs and the intestines; penetration via the skin is less evident. It is possible that some particles can penetrate deep into the dermis. The chances of penetration would depend on the size and surface properties of the particles and on the point of contact in the lung, intestines, or skin.

After penetration, the distribution of the particles in the body is a strong function of the surface characteristics of the particle. It seems that size can restrict the free movement of particles. The target organ-tissue or cell of a nanoparticle needs to be investigated, particularly in the case of potentially hazardous compounds. Before developing an *in vitro* test, it is essential to know the pharmacokinetic behavior of different types of nanoparticles; therefore, it would be important to compose a database of health risks associated with different nanoparticles.

Beside the study of the health effects of the nanomaterials, investigations should also take into consideration the presence of contaminates, such as metal catalysts present in nanotubes and their role in the observed health effects.

The increased risk of cardiopulmonary diseases requires specific measures to be taken for every newly produced or used nanoparticle. There is no universal "nanoparticle" to fit all cases; each nanomaterial should be treated individually when health risks are expected. The tests currently used to test the safety of materials should be applicable to identify hazardous nanoparticles.

BIBLIOGRAPHY

1. Alivisatos P. The use of nanocrystals in biological detection. Nature Biotechnol 2004;22:47–52.
2. Brus L. Electronic wave function in semiconductor clusters: Experiment and theory. J Phys Chem 1986;90:2555–2560.
3. Kreuter J. Nanoparticles. In: Kreuter J, editor. Colloidal drug delivery systems. New York: Marcel Dekker; 1994.
4. Mikheev NB. Radioactive colloidal solutions and suspensions for medical use. At Energy Rev 1976;14:3–36.
5. Horisberger M. Colloidal gold: a cytochemical marker for light and fluorescent microscopy and for transmission and scanning electron microscopy. Scan Electron Microsc 1981 (Pt 2):9–31.
6. Yavari AR. Mechanically prepared nanocrystalline materials. Mater T JIM 1995;36:228–239.
7. Huczko A. Template-based synthesis of nanomaterials. Appl Phys A-Mater 2000;70:365–376.
8. Gurav A, Kodas T, Pluym T, Xiong Y. Aerosol processing of materials. Aerosol Sci Technol 1993;19:411–452.
9. Jain RA. The manufacturing techniques of various drug loaded biodegradable poly(lactide-*co*-glycolide) (PLGA) devices. Biomaterials 2000;21:2475–2490.
10. Zhang SG. Emerging biological materials through molecular self-assembly. Biotechnol Adv 2002;20:321–339.
11. Caruso F. Nanoengineering of particle surfaces. Adv Mate 2001;13:11–22.
12. Salata OV. Applications of nanoparticles in biology and medicine. J Nanobiotechnol 2004;2:3–8.
13. Soppimatha KS, Aminabhavia TM, Kulkarnia AR, Rudzinski WE. Biodegradable polymeric nanoparticles as drug delivery devices. J Controlled Release 2001;70:1–20.
14. Merisko-Liversidge E, Liversidge GG, Cooper ER. Nanosizing: A formulation approach for poorly-water-soluble compounds. Eur J Pharm Sci 2003;18:113–120.
15. Hoet PH, Bruske-Hohlfeld I, Salata OV. Nanoparticles — known and unknown health risks. J Nanobiotechnol 2004;2:12–26.

FURTHER READING

General

Moriarty P. Nanostructured materials. *Rep Prog Phys* 2001;64: 297–381.

Schmid G, Baumle M, Geerkens M, Heim I, Osemann C, Sawitowski T. Current and future applications of nanoclusters. *Chem Soc Rev* 1999;28:179–185.

Fabrication of Nanoparticles

Bourgeat-Lami E. Organic-inorganic nanostructured colloids. *J Nanosci Nanotechnol* 2002;2:1–24.

Fendler JH. Colloid chemical approach to nanotechnology. *Korean J Chem Eng* 2001;18:1–13.

Gaffet E, Abdellaoui M, Malhourouxgaffet N. Formation of nanostructural materials induced by mechanical processings. *Mater T JIM* 1995;36:198–209.

Meier W. Polymer nanocapsules. *Chem Soc Rev* 2000;29:295–303.

Meisel D. Inorganic small colloidal particles. *Curr Opin Colloid In* 1997;2:188–191.

Shimomura M, Sawadaishi T. Bottom-up strategy of materials fabrication: a new trend in nanotechnology of soft materials. *Curr Opin Colloid In* 2001;6:11–16.

Tomalia DA, Wang ZG, Tirrell M. Experimental self-assembly: the many facets of self-assembly. *Curr Opin Colloid In* 1999;4:3–5.

Ullmann M, Friedlander SK, Schmidt-Ott A. Nanoparticle formation by laser ablation. *J Nanoparticle Res* 2002;4:499–509.

Yu SH. Hydrothermal/solvothermal processing of advanced ceramic materials. *J Ceram Soc Jpn* 2001;109:S65–S75.

Biological and Medical Applications

de la Isla A, Brostow W, Bujard B, Estevez M, Rodriguez JR, Vargas S, Castano VM. Nanohybrid scratch resistant coating

for teeth and bone viscoelasticity manifested in tribology. *Mat Res Innovat* 2003;7:110–114.
Han M-Y, Gao X, Su JZ, Nie S-M. Quantum-dot-tagged microbeads for multiplexed optical coding of biomolecules. *Nature Biotechnol* 2001;19:631–635.
Loo C, Lowery A, Halas N, West J, Drezek R. Immunotargeted nanoshells for integrated cancer imaging and therapy. *Nano Lett* 2005;5:709–711. Ozkan M. Quantum dots and other nanoparticles: What can they offer to drug discovery? *DDT* 2004;9:1065–1071.
Ma J, Wong H, Kong LB, Peng KW. Biomimetic processing of nanocrystallite bioactive apatite coating on titanium. *Nanotechnology* 2003;14:619–623.
Molday RS, MacKenzie D. Immunospecific ferromagnetic iron dextran reagents for the labeling and magnetic separation of cells. *J Immunol Methods* 1982;52:353–367.
Pankhurst QA, Connolly J, Jones SK, Dobson J. Applications of magnetic nanoparticles in biomedicine. *J Phys D: Appl Phys* 2003;36:R167–R181.
Panyam J, Labhasetwar V. Biodegradable nanoparticles for drug and gene delivery to cells and tissue. *Adv Drug Del Rev* 2003;55: 329–347.
Parak WJ, Gerion D, Pellegrino T, Zanchet D, Micheel C, Williams SC, Boudreau R, Le Gros MA, Larabell CA, Alivisatos AP. Biological applications of colloidal nanocrystals. *Nanotechnology* 2003;14:R15–R27.
Pricea RL, Waidb MC, Haberstroha KM, Webster TJ. Selective bone cell adhesion on formulations containing carbon nanofibres. *Biomaterials* 2003;24:1877–1887.
Reich DH, Tanase M, Hultgren A, Bauer LA, Chen CS, Meyer GJ. Biological applications of multifunctional magnetic nanowires. *J Appl Phys* 2003;93:7275–7280.
Roy I, Ohulchanskyy TY, Pudavar HE, Bergey EJ, Oseroff AR, Morgan J, Dougherty TJ, Prasad PN. Ceramic-based nanoparticles entrapping water-insoluble photosensitizing anticancer drugs: A novel drug-carrier system for photodynamic therapy. *J Am Chem Soc* 2003;125:7860–7865.
Sinani VA, Koktysh DS, Yun BG, Matts RL, Pappas TC, Motamedi M, Thomas SN, Kotov NA. Collagen coating promotes biocompatibility of semiconductor nanoparticles in stratified LBL films. *Nano Lett* 2003;3:1177–1182.
Taton TA. Nanostructures as tailored biological probes. *Trends Biotechnol* 2002;20:277–279.
Weissleder R, Elizondo G, Wittenburg J, Rabito CA, Bengele HH, Josephson L. Ultrasmall superparamagnetic iron oxide:characterization of a new class of contrast agents for MR imaging. *Radiology* 1990;175:489–493.
Yoshida J, Kobayashi T. Intracellular hyperthermia for cancer using magnetite cationic liposomes. *J Magn Magn Mater* 1999;194:176–184.
Zhang Y, Kohler N, Zhang M. Surface modification of superparamagnetic magnetite nanoparticles and their intracellular uptake. *Biomaterials* 2002;23:1553–1561.

Commercial Exploration

Mazzola L. Commercializing nanotechnology. *Nature Biotechnol* 2003;21:1137–1143.
Paull R, Wolfe J, Hebert P, Sinkula M. Investing in nanotechnology. *Nature Biotechnol* 2003;21:1144–1147.

Health Risks

Anonymous. Nanotech is not so scary. *Nature* 2003;421:299.
Borm PJ. Particle toxicology: From coal mining to nanotechnology. *Inhal Toxicol* 2000;14:311–324.
Sanfeld A, Steinchen A. Does the size of small objects influence chemical reactivity in living systems? *C R Biol* 2003;326:141–147.
UK Royal Society and Royal Academy of Engineering. Nanoscience and nanotechnologies: opportunities and uncertainties. Final Report. (2004). Available: http://www.nanotec.org.uk/finalReport.htm.

See also DRUG DELIVERY SYSTEMS; MICROSCOPY, ELECTRON; TISSUE ENGINEERING.

NEAR-FIELD MICROSCOPY AND SPECTROSCOPY.

See MICROSCOPY AND SPECTROSCOPY, NEAR-FIELD.

NEONATAL MONITORING

MICHAEL R. NEUMAN
Michigan Technological University
Houghton, Michigan

INTRODUCTION

The care of premature and newborn infants is quite different from other areas of clinical medicine. The infant represents a special patient with special problems not found in other patients. For this reason a subspecialty of pediatrics dealing with these patients has been established. Neonatology is concerned with newly born infants including those prematurely delivered and those delivered at term. The field generally covers infants through the first month of normal newborn life, and so for prematurely born infants this can be several additional months.

Neonatology includes special hospital care for infants who require it. In the case of the prematures, this involves specialized life-support systems, as well as special considerations for nutrition, thermal control, fluid and electrolyte therapy, pulmonary support, and elimination of products of metabolism. While the full-term infant generally only requires routine well-baby care, there are special cases that require intensive hospital care as well. These include treatment of infants of diabetic mothers, some infants delivered by cesarean section, infants with hemolytic diseases, infants who encounter respiratory distress, and other less common problems. Special hospital care is also necessary for infants requiring surgery. These infants are generally born with severe congenital anomalies that would be life threatening if not immediately repaired. These include anomalies of the gastrointestinal system, urinary tract, cardiovascular system, and nervous system. Pediatric surgery has developed to the point where many of these problems can be corrected, and the infant can grow and lead a normal life following the surgery.

The neonatal intensive care unit is a special nursery in major tertiary care hospitals that is devoted to the care of premature or other infants who require critical care. This unit is similar to its adult counterpart in that each patient is surrounded by equipment necessary for life support, diagnosis, and therapy. Often, as indicated in Fig. 1, the patient appears to be insignificant in the large array of equipment, but, of course, this is not the case. Nursing

Figure 1. Typical infant station in a neonatal intensive care unit showing infant warmer, infusion pumps, ventilator, transcutaneous oxygen instrument, cardiopulmonary monitor, bilirubin lights, and other miscellaneous apparatus.

functions in the neonatal intensive care unit are very important. The patient/nurse ratio is small, and the nursing staff must be familiar with the equipment as well as special procedures and precautions in caring for these special patients.

Electronic monitoring of the infant plays an important role in neonatal intensive care. Not only does it allow the clinical caregivers to follow vital signs, such as pulse rate, temperature, blood pressure, and respiration rate, but other critical variables in the care of these special patients can be followed as well. These include blood gas tensions, acid–base balance, bilirubin, and glucose concentrations. Monitoring is especially important in fluid therapy for it can provide precise data for fluid control of these very small patients. Electronic monitoring, however, goes beyond just monitoring the patient and its physiologic functions. A good neonatal intensive care unit also monitors the functioning of life-support systems. These include incubators for maintaining an appropriate thermal environment, ventilators for providing respiratory support, and phototherapy units for the control of bilirubin.

Although electronic monitoring devices for just about all of the areas mentioned in the previous paragraph are used in adult intensive care medicine, their application in neonatology often represents a unique aspect of the technology. The infant should not be viewed as a miniature adult, but rather he/she is a unique physiologic entity. Although similar variables are measured to those measured in adults, they often must be measured in different ways. Frequently, sensors unique for infants must be applied because the sensors used for adults when interfaced to the infant might provide errors or change the variable being measured by their very presence. Size is an important aspect here. If one considers a sensor to be used on an infant and compares it to a sensor for the same variable on an adult, in most cases although the sensor for the infant is smaller than that for the adult, the ratio of sizes of the two sensors is quite different from the ratio of sizes of the different patients. Although sensors for use on infants are reduced in size, they are still quite large when compared to the size of the subject. This is especially true for premature infants and can result in the sensors actually interfering with the care of the patient.

There are also special problems related to the measurement of physiologic variables in infants resulting from the special physiology of newborns and especially premature newborns. One first must realize that a newborn has come to live in a new environment quite different from the uterus. In the case of premature infants, they are not ready for this major change in their lives, and special considerations need to be made to minimize the transitional trauma. In the case of the premature, some of the body systems are immature and not ready for life outside of the uterus. Two notable examples of this are the control of temperature and control of respiration. Both are obviously unnecessary in the uterus, but become crucial in extrauterine life. Instrumentation to assist these control systems or to detect when they are not functioning properly is essential in the care of many premature infants.

One also must realize in applying instrumentation systems for premature infants that the patient in many cases is much more fragile than an adult patient. Fluid and electrolyte balance has already been indicated as an important aspect of neonatal monitoring and control. When one considers some of the very small premature infants that are cared for in neonatal intensive care units today, this can be better appreciated. Infants between 500 and 1,000 g can be successfully cared for and nurtured until they are old enough and grow enough to go home with their parents. These very small babies, however, can easily run into problems if they receive either too much or too little fluid. Since feeding of these very small infants can be done by intravenous hyperalimentation, the possibility of a fluid overload is always present since it takes a certain amount of fluid to transport the nutritional requirements of the infant. Another example of the fragility of these very small patients is the simple problem of attaching devices to the infant's skin. In some infants, the skin is very sensitive and can easily become irritated by the attachment procedure or substance.

This article, looks more closely at electronic monitoring systems for neonatal intensive care and emphasize those aspects of these monitoring systems that differ from similar monitors for adult patients. The reader is encouraged to supplement information contained in the following paragraphs with other articles from this encyclopedia

dealing with the sensors and instrumentation for similar monitoring in adults.

CARDIAC MONITORING

Cardiac monitoring involves the continuous assessment of heart function by electronic measurement of the electrocardiogram and determination of heart rate and rhythm from it by means of electronic signal processing. As such, cardiac monitors for neonatal use are very similar to those for use with adults. There are, however, two major differences. The sensors used with both types of monitors are biopotential electrodes, and in the case of infants the interface between the electrodes and the patients has more stringent requirements than in the adult case. Second, cardiac monitors designed for use with infants frequently are incorporated into cardiorespiratory monitors that include instrumentation for determining breathing rate and apnea as well as cardiac function.

The primary use of cardiac monitors for infants is in determining heart rate. These electronic devices are designed to indicate conditions of bradycardia (low heart rate) and tachycardia (high heart rate) by determining the heart rate from the electrocardiogram. In the case of infants with heart diseases, cardiac monitors are used to detect various arrhythmias as well.

Cardiac monitors for use with infants are organized similarly to their adult counterparts (see MONITORING, HEMODYNAMIC). There are some minor differences due to the fact that infant heart rates are higher than those of adults, and the Q–S interval of the infant electrocardiogram is less than it is in the adult. Thus, heart rate alarm circuits need to be able to respond to higher rates in the infant case than in the adult case. For example, it is not at all unusual to set the bradycardia alarm level at a rate of 90 or 100 beats·min^{-1} for an infant, which is well above the resting heart rate of a normal adult. Filtering circuits in the monitor for infants must be different from those of adult monitors for optimal noise reduction due to the different configuration of the neonatal electrocardiogram. Generally, bandpass filters used for isolating the QRS complex will have a higher center frequency than in the adult case.

Two types of cardiotachometer circuits can be used in cardiac monitors (1). The averaging cardiotachometer determines the mean number of heartbeats per predetermined interval to establish the heart rate. The mean R–R interval over a number, of heartbeats can also be used in average heart rate determination. In such systems the heart rate is calculated by averaging over from as few as three to as many as fifteen or more heartbeats. An instantaneous or beat-to-beat cardiotachometer determines the heart rate for each measured R–R interval. This type of cardiotachometer must be used when one is interested in beat-to-beat variability of the heart rate.

Biopotential electrodes for use with cardiac monitors for infants are usually scaled down versions of skin surface electrodes used for adult cardiac monitoring. As pointed out earlier, the scale factor does not correspond to the body size ratio between the neonate and an adult, and the smallest commercially available skin surface electrodes for neonates only approaches about one-fourth the size of those used in adults. For this reason, electrodes used with neonatal cardiac monitors and their method of attachment can cover a large portion of the neonatal thorax. This is especially true with the small premature infant and can interfere with direct observation of chest wall movements, an important diagnostic method. In addition to size, shape and flexibility of the electrode are important for biopotential electrodes in neonates. Stiff, flat electrode surfaces will not conform well to the curved, compliant surface of the infant. This means that optimal electrical contact is not always possible and it, therefore, becomes more difficult to hold electrodes in place. This problem is further complicated by the fact that the neonatal skin can be sensitive to the electrode adhesive. It is not at all unusual to find skin irritation and ulceration as a result of placement of biopotential electrodes on the infant. Such skin lesions are usually the result of the adhesive and the electrode attachment system, although the electrode itself can in some cases be the problem as well.

Since electrodes are relatively large on the small infant, an additional problem develops. The materials used in many electrode systems are X-ray opaque; hence, it is necessary to remove the electrodes when X rays are taken so that shadows do not appear in the resulting radiograph. Some biopotential electrodes especially developed for neonates have minimized this problem by utilizing special electrode structures that are translucent or transparent to X rays (2). These electrodes are based upon thin films of metals, usually silver, deposited upon polymer films or strips or various fabric materials. These films are sufficiently thin to allow X rays to penetrate with little absorption, and the plastic or polymer substrate is also X-ray transparent. Such electrodes have the advantage of increased flexibility, which helps them to remain in place for longer periods of time. In intensive care units, however, it is a good idea to change electrodes every 48 h to minimize the risk of infection.

Electrode lead wires and patient cables present special problems for cardiac monitors used with infants. Lead wires should be flexible so as not to apply forces to the electrodes that could cause them to become loose, but this increased flexibility makes it easier for them to become ensnarled with themselves and the infant. The potential for strangulation on older, active infants is always present. The connectors between the lead wires and the patient cable also present special problems. They must be capable of maintaining their connection with an active infant and provide a means of connection that will be unique for these components. The possibility of inadvertently connecting the lead wires, and hence the infant, to the power line must be eliminated (3).

RESPIRATORY MONITORING

Respiratory monitoring is the most frequently applied form of electronic monitoring in neonatology. In its most common application, it is used to identify periods of apnea and to set off an alarm when these periods exceed a predetermined limit. There are direct and indirect methods of

sensing alveolar ventilation and breathing effort. The direct methods are those in which the sensor is coupled to the airway and measures the movement or other properties of the air transported into and out of the lungs. In the indirect methods, the sensor looks at variables related to air movement, but not at the air movement itself. Indirect methods involve no contact with the airway or the air being moved into or away from the lungs. Usually, indirect methods are noninvasive and can be mounted on or near the body surface. Some of the most frequently applied methods are described in the following paragraphs.

Direct Methods

Various direct methods of sensing breathing effort and ventilation have been in use in the pulmonary physiology and pülmonary function laboratories for many years. These involve the measurement of volume, flow, and composition of inspired and expired gasses. Table 1 lists some of the principal methods that have been used for the direct measurement of respiration in infants and neonates. Most of these methods are not appropriate for clinical monitoring, since they involve direct connection to the infant airway through the use of a mask over the mouth and nose or an endotrachial canula. In other cases a sensor must be located at the nasal-oral area for signal detection. These methods are, however, useful in some cases for diagnostic studies carried out for periods from several hours to overnight in the hospital setting.

Many of the methods listed in Table 1 are described in detail in the article on pulmonary function testing, and therefore are not repeated here. Others, however, have special application to neonatal monitoring and will be mentioned.

Pneumotachography. Clinicians and researchers involved in neonatal and infant care agree for the most part that the best measurement of ventilation can be obtained using the pneumotachograph (4). Although this involves direct connection to the airway and can add some dead space due to the plumbing, it, with an appropriate electronic integrator, provides good volume and flow measurements that can be used as a standard against which other direct or indirect methods can be calibrated and evaluated. Identical instrumentation as used for adults can be applied in the infant case, but it must be recognized that dead space due to the instrumentation represents a more important problem with the infant than it does with the adult. Lower flows and volumes as well as faster respiration rates will be encountered with infants than with adults.

Table 1. Direct Methods of Sensing Breathing and Ventilation

Method	Primary Sensed Variable
Pneumotachograph	Flow volume
Anemometer	Flow velocity
Expired air temperature	Temperature
Air turbulence sounds	Sound
Spirometer	Volume

Capnography. Special carbon dioxide sensors have been developed for measuring air expired from the lungs, and these are used as the basis of a direct respiration monitor (5). Expired air has a higher percentage of carbon dioxide than inspired air, and this can be sensed by placing an open-ended tube at the nose or mouth so that it samples the air entering and leaving the airway. The sampled gas is transported along the tube to an instrument that contains a rapidly responding carbon dioxide sensor. This is generally a sensor that detects the increased absorption of infrared (IR) radiation by carbon dioxide-containing gas. Thus, when a sample of expired gas reaches the sensor, an increase in carbon dioxide content is indicated, while a decrease in carbon dioxide is seen in samples of air about to be inspired. There is a delay in response of this instrument due to the time it takes the gas to be transported through the tube to the sensor; thus, it is important to have rapid passage through this tube to minimize this delay. This can present some problems since the tube must be thin and flexible and, therefore, offers a relatively high resistance to the flow of gas. While it is generally not necessary to have a quantitative measure of carbon dioxide for respiration monitoring, the system can be refined to the point where it can measure the carbon dioxide content of the end tidal expired air, which is the gas that actually was in the alveoli (6).

Temperature Sensor. Similar sensing systems based upon temperature variations have also been used to monitor respiration (7). These generally can be divided into two types: one that measures temperature differences between inspired and expired air and one that measures the cooling of a heated probe as inspired or expired air is transported past it. In both cases, the temperature sensor of choice is a small, low mass, and therefore fast responding, thermistor. In the first mode of operation, the thermistor changes its resistance proportionally to the change in temperature of the air drawn over it. This can then be electronically detected and processed to determine respiration rate. It is also possible to heat the thermistor by an electrical current. Some of this heat will be dissipated convectively by the air passing over the sensor. As the flow of air over the thermistor increases, more heat will be drawn from the thermistor, and it will cool to a lower temperature. Changes in the thermistor's temperature can be determined by measuring its electrical resistance. Thus, an electrically heated thermistor will cool during both inspiration and expiration, and it will become warmer in the interval between these two phases when air is not passing over it. This type of anemometer gives a respiration pattern that appears to be twice the breathing rate, whereas the unheated thermistor gives a pattern that is the same as the breathing rate. An important consideration in using the nasal thermistor for ventilation measurement is its placement in the flowing air. For young infants, the sensor package can be taped to the nose or face so that the thermistor itself is near the center of one nostril. Another technique is to place a structure containing two thermistors under the nose so that each thermistor is under one nostril and expired air flows over both thermistors.

Nasal temperature sensors, such as thermistors, have been used for monitoring ventilation in research studies and for making physiologic recordings in the hospital and in the laboratory (8). Their advantage is that the electronic circuit for processing the signal is relatively simple and inexpensive compared with other techniques. The major problem of the method is the placement of the thermistor on the infant and maintaining it in place. Thermistors can also become covered with mucus or condensed water, which can greatly reduce their response time. Most investigators who use this technique prefer the temperature sensing rather than the flow-detecting mode. The devices can also be used with radiotelemetry systems to eliminate the wire between the thermistor on the subject's face and the remainder of the monitoring apparatus (9).

Although thermistors have a high sensitivity and can be realized in a form with very low mass, they are fragile when in this low mass form and are relatively expensive components. Low mass, high surface area resistance temperature sensors can also be fabricated using thin- and thick-film temperature sensitive resistors.(10) These can either be fabricated from metal films with relatively high temperature coefficients of resistance or more sensitive films of thermistor materials. Single use disposable sensors have been produced for use in infant and adult sleep studies as shown in Figure 2.

Sound Measurement. Air passing over the end of an open tube generates sound by producing local turbulence. A miniature microphone at the other end of the tube can detect this sound, and the level of sound detected is roughly proportional to the turbulence and, hence, the air flowing past the open end. Nasal air flow can thus be detected by placing the open end of the tube in the stream of inspired or expired air at the nose by taping the tube to the infant's face in much the same way as was done for the carbon dioxide sensor mentioned previously (11). As with the thermistor anemometer, this technique can detect changes for both inspired and expired air and will give a pattern that appears to indicate double the actual respiration rate. The method has been demonstrated to give efficacious monitoring results, but can suffer from sensitivity to extraneous sounds other than the air passing the open ended tube. This can lead to incorrect detection of breaths.

Indirect Sensors of Ventilation

There are a wide variety of indirect sensors of ventilation that can be applied to monitoring in infants. Table 2 lists some of the principal examples of these various types of sensors and sensing systems, and those with aspects unique to neonatal monitoring will be described in the following paragraphs. The main advantage of the indirect methods of sensing ventilation is that attachment to the subject is easier than for the direct measurements and less likely to interfere with breathing patterns. Of the methods described in Table 2 and this section, the transthoracic electrical impedance method is the one used in most presently available respiration-apnea monitors for both hospital and home use. This, therefore, will be described in greatest detail in a separate section.

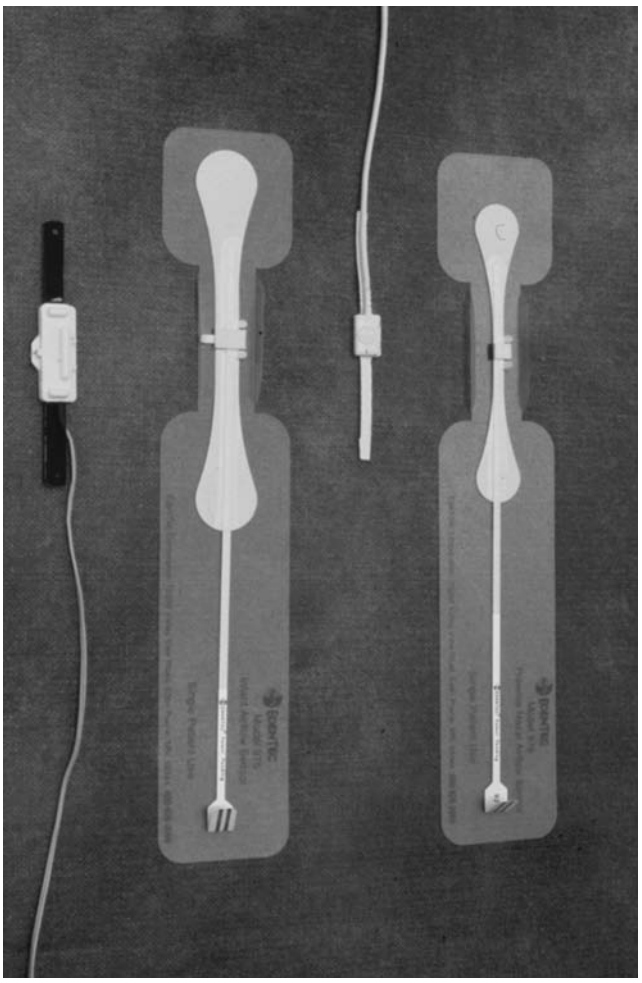


Figure 2. An example of commercially available thick-film nasal temperature sensors for measurement of breathing patterns. The small sensors on the illustration are conventional thermistor sensors.

The Whole-Body Plethysmograph. This method is used primarily in pulmonary function testing, and the reader is referred to the article on this subject for details on the method. Miniature whole-body plethysmographs have been designed for use with neonates and infants, but this

Table 2. Indirect Methods of Sensing Breathing and Ventilation

Transthoracic electrical impedance
Whole-body plethysmograph
Contacting motion sensors
Strain gage
Air-filled capsule or vest
Magnetometer
Inductance respirometry
Noncontacting motion sensors
Motion-sensing pad
Radiation reflection
Variable capacitance sensor
Electromyography
Breath sounds
Intraesophageal pressure

application is strictly for purposes of research or diagnostic studies. The technique is not appropriate for routine clinical monitoring.

Contacting Motion Sensors. Breathing effort involves the movement of different parts of the body for pulmonary ventilation to occur. Sensors can be placed upon and attached to an infant to measure this motion. These contacting motion sensors pick up movements of the chest and/or abdomen, and there are several different types of sensors that fall within this category. These are described in the following sections.

Strain Gage Displacement Sensors. Strain gages measure small displacements or strain in an electrical conductor by measuring changes in its electrical resistance. Most strain gages used for general measurements are made of thin metal foils or wires and are useful for measuring only very small displacements due to their very low mechanical compliance. A special type of strain gage consisting of a compliant, thin-walled, rubber capillary tube filled with mercury was developed by Whitney as a limb plethysmograph (12). This compliant device can be placed on the chest or abdomen of an infant such that breathing movements cause it to stretch and contract without offering significant mechanical constraint to the breathing efforts of the infant. By taping the ends of such a strain gage at different points on the chest or abdomen such that the gage is slightly stretched, the changes in electrical resistance of the gage can then be used to monitor infant breathing movements. Simple electronic resistance measurement circuitry can be used for processing the signal.

This technique is used primarily in research and in some rare cases for in-hospital monitoring and recording of infant respiration patterns. Its limitations are related to use of a toxic substance that could escape from the sensor and put the infant at risk. In addition, the mercury column frequently becomes interrupted after several days of use, thereby limiting the sensor's reliability for infant monitoring. Nevertheless, workers who use this sensor for monitoring purposes are enthusiastic about its reliability in picking up high quality respiration patterns.

Air-Filled Capsule or Vest. Breathing efforts of an infant can also be determined for chest or abdominal movements by a sensor consisting of an air-filled compliant tube, disk, or entire vest attached around an infant. The tube and disk can be taped to the infant's chest or abdomen in a fashion similar to the strain gage, and the structures will be stretched or compressed by the infant's breathing movements. This causes the pressure of the air within to increase or decrease as a result of volume changes, and this pressure variation can be measured by coupling the sensor to a sensitive pressure transducer through a fine-gage flexible tube. The advantage of this system is that the sensors on the infant are simple and inexpensive and thus can be considered disposable devices. Since only air is contained within the sensors, they are not toxic and are much more reliable than the mercury strain gages. They can be produced as inexpensive disposable sensors.

Displacement Magnetometers. The magnetic field from a permanent magnet or an electromagnet decreases as one gets farther from the magnet. By placing such a magnet on an infant's chest or abdomen with a detector located on the back of the subject or underneath the infant, differences in separation between the magnet and the detector can be sensed as the infant breathes (13). It is important that such a system be designed so that it will only respond to breathing movements and will be insensitive to other movements of the infant. Unfortunately, this is not always the case, and sensors of this type can respond to infant limb movement as well as movements between the infant and the pad upon which it is placed.

Inductance Respirometry. The inductance of a loop of wire is proportional to the area enclosed by that loop. If a wire is incorporated in a compliant belt in a zigzag fashion so that the wire does not interfere with the stretching of the belt, such a belt can be wrapped around the chest or abdomen of an infant to form a loop. As the infant inhales or exhales the area enclosed by this loop will change, and so the inductance of the loop will also change. These changes can be measured by appropriate electronic circuits and used to indicate breathing efforts. Investigators have shown that the use of such a loop around the chest and the abdomen of an adult can, when appropriately calibrated, measure tidal volume as well as respiratory effort (14). Although the system is simple in concept, realizing it in practice can involve complicated and therefore costly electronic circuitry (15). Often as the subject moves to a new position, the calibration constant relating inductance and volume will change thereby making the instrument less quantitative, yet still allowing it to be suitable for qualitative measurements. Variations in tidal volume measurements using this technology have been reported by Brooks et al. (16) Since the instrument is sensitive to inductance changes in the wire loop, anything in the vicinity of the wire that affects its inductance also will affect the measurement. Thus, the instrument can also be sensitive to moving electrical conductors or other magnetic materials in the vicinity of the infant.

Noncontacting Motion Sensors. Sensors of infant breathing effort and pulmonary ventilation that detect breathing movements of the infant without direct patient contact fit in this category. These sensors can consist of devices that are placed under the infant or can sense movement of the infant by means of a remotely located sensor. A clinician, in effect, is an indirect motion sensor when he or she determines infant breathing patterns by watching movements of the chest and abdomen. Devices in this category have a special appeal for monitoring systems that are used outside of the hospital, such as instruments for use in the home. With many of the noncontacting sensors, the infant-sensor interface can be created by individuals who do not have specialized training. For example, the motion sensing pad discussed in the next paragraph is attached to the infant by simply placing the infant on top of it in a bassinet or crib.

Motion Sensing Pad. Movements of neonates and infants can be sensed by a flexible pad that responds to

compression by producing an electrical signal when the infant is placed on top of the pad. There are two different forms of this sensor that can be used for motion detection. The first utilizes a piezoelectric polymer film, polyvinylidene fluoride, that has its surfaces metalized to form electrical contacts. Depending on the piezoelectric properties of the film, an electrical signal is produced between the metalized layers when the polymer is either compressed or flexed. In the former case, the polymer film and its metalized electrode need only to be packaged in an appropriate pad structure to be used, while in the latter case the package must be a little more complex with the polymer film positioned between two corrugated, flexible layers so that compression of the structure causes the piezoelectric polymer to be flexed (17). The second form of the pad uses an electret material to generate the electrical signal. The actual pad structure in this case is similar to that for the piezoelectric material.

The sensitive portion of the motion sensing pad structure is usually smaller than the overall size of the infant and is located under the infant's thoracic and/or lumbar regions. Infant breathing efforts result in periodic compression of the pad as the center of mass of the infant shifts cephalad and caudad with respiratory motion. This generates a periodic electrical signal related to the breathing effort.

The major limitation of the motion sensing pad is its sensitivity to movements other than those related to respiratory efforts of the infant. Other body movements can be picked up by the sensor, and the device can even respond to movements that are not associated with the infant at all, such as an adult walking near or bumping the bassinet or crib, a heavy truck, train, or subway passing nearby, or even earthquakes.

Radiation Reflection. Electromagnetic radiation in the microwave range (radar) or ultrasonic radiation (sonar) can be reflected from the surface of an infant. If this surface is moving, as, for example, would be the chest or abdominal wall during breathing efforts, the reflected radiation will be shifted in frequency according to the Doppler effect. In some cases the reflected signal's amplitude will be shifted as well as a result of this motion. These changes can be detected and used to sense breathing efforts without actually contacting the infant. The problem with these methods is that the movement of any surface that reflects the radiation will be detected. Body movements of the infant that are unrelated to respiratory movements can be detected and mistakenly identified as breathing effort, and even in some cases movement of objects in the vicinity of the infant, such as a sheet of paper shifting due to air currents, will also be detected as infant respiration. Thus, this type of monitor has the possibility of indicating apparent breathing activity during periods of apnea if moving objects other than the infant are within the range of the radiation sensor. This technique of noncontacting detection of breathing is not considered to be reliable enough for routine clinical use, and a commercial device based on this principle has been withdrawn from the market.

Variable Capacitance Displacement Sensor. A parallel plate capacitor can be fabricated so that an infant is placed between the parallel conducting planes. For example, such a capacitor could be formed in an incubator by having the base upon which the mattress and infant are placed serving as one plate of the capacitor and having the second plate just inside the top of the incubator (18). To maintain good clinical practice, this second plate should consist of a transparent conductor, such as an indium tin oxide film, so that it does not interfere with a clinician's ability to observe the patient. Since a major component of the infant's tissue is water, and water has a relatively high dielectric constant compared to air, movements of the infant will produce changes in capacitance between plates that can be detected by an electronic circuit. Such changes can be the result of breathing movements by the infant, but they also can result from other infant movement or movement of some other materials in the vicinity of the conducting plates. Therefore, for this system to be effective, adequate electrical shielding of the capacitor is essential. Thus, this indirect motion sensor suffers from some of the same problems as other sensors in this classification: the lack of specificity for breathing movements.

Electromyography. Many different muscles are involved in breathing activity. The diaphragm is the principal muscle for pulmonary ventilation, but the accessory muscles of the chest wall including the intercostal muscles are also involved, Electromyographic activity of the diaphragm and intercostal muscles can be sensed from electrodes on the chest surface. By measuring these signals, one can determine if respiratory efforts are being made, although such measurements cannot be quantitative with regard to the extent of the effort or the volume of gas moved (19). Unfortunately, other muscles in the vicinity of the electrodes that are not involved in breathing also produce electromyographic signals. These signals can severely interfere with those associated with respiration, and this is especially true when the infant is moving. This represents a serious limitation of this method for clinical infant respiration monitoring.

Breath Sounds. Listening to chest sounds through a stethoscope is an important method of physical diagnosis for assessing breathing. The technique can be used for infant monitoring by placing a microphone over the chest or trachea at the base of the neck and processing the electrical signals from this sensor. In addition to the sounds associated with air transport and ventilation, the microphone will pick up other sounds in the body and the environment. Thus, for this type of monitoring to be efficacious, it must be done in a quiet environment. This puts a serious constraint on the practical use of this technique, and it has only been used in limited experimental protocols.

Intraesophageal Pressure. The pressure within the thorax decreases with inspiratory effort and increases with expiratory effort. These changes can be measured by placing a miniature pressure sensor in the thoracic portion of the esophagus or by placing a small balloon at this point and coupling the balloon to an external pressure

transducer through a small diameter flexible tube. While this method is invasive, it is not considered a direct method since there is no contact with the flowing air.

An important aspect of intraesophageal pressure measurement is that it represents a standard method that is accepted by physiologists as a measure of respiratory effort. Thus, by combining intraesophageal pressure measurement and the pneumotachograph, one is able to monitor both gas flow and breathing effort. Although both of these methods are generally too complicated for clinical monitoring, they can be used in conjunction with other monitoring methods described in this article as standards against which to assess the other devices.

Transthoracic Electrical Impedance. The electrical impedance across the chest undergoes small variations that are associated with respiratory effort. The measurement of these variations is the basis of the most frequently used infant respiration and apnea monitoring technique. The following section describes the basic principle of operation, the methods of signal processing, and sources of error for this technique.

RESPIRATION MONITORING BY TRANSTHORACIC ELECTRICAL IMPEDANCE

The chest contains many different materials ranging from bone to air. Each of these materials has its own electrical properties and of its own unique location in the thorax. One can roughly represent a cross section of the infant chest as shown in Fig. 3, where the major components consist of chest wall, lungs, heart, and major blood vessels. The various tissues contained in these structures range in electrical conductivity from blood, which is a relatively good conductor, to air, which is an insulator. Both of these materials in the thoracic cavity show a change in volume with time over the cardiac and breathing cycles. Blood varies in volume over the cardiac cycle due to changes in the amount of blood in the heart and the vascular compartments. Air undergoes wide volume changes in the lungs during normal breathing. Thus, the electrical impedance of the lungs and heart will change as the volume of air and blood in each, respectively, changes. If we want to measure the impedance variation due to these volume changes, this can be done by placing electrodes on the surface of each structure. If it were practical to do this, we would see large changes in impedance as the volumes of the respective structures change. Unfortunately, it is not possible to place electrodes on the structures that are to be measured and so these large impedance differences are not seen in practice.

Electrodes must be placed upon the surface of the skin for practical electrical impedance measurements on infants. Most of the current passing between the electrodes will travel through the chest wall and will not pass through the heart and lungs because of the low resistivity of the tissues in the chest wall. Thus, the changes in impedance of the heart and lungs will only represent a small proportion of the impedance measured between the electrodes. Fig. 3 schematically illustrates the relative distribution of the current through the chest when electrodes are placed on the midclavicular lines at the fourth intercostal space. It is seen that most of the current is conducted along the chest wall, so the chest wall impedance will dominate any measurement.

The actual impedance measured by the monitor consists of more than just the impedance between the electrodes on the chest surface. Since an ac electrical signal is needed to measure the impedance, this signal will affect the measurement as well. Generally, a signal in the frequency range from 20–100 kHz is used. At these frequencies, impedances associated with the electrode, the interface between the electrode and the body, and the lead wires contribute to the measured value along with the actual transthoracic impedance. This is illustrated schematically in Fig. 4. The actual impedances for each block are dependent upon the excitation frequency and the actual structures used, but for most clinical applications the net impedance seen by the monitoring circuit is nominally 500 Ω. Of this, the variation associated with respiration is generally no $> 2\ \Omega$ and frequently even less. The impedance

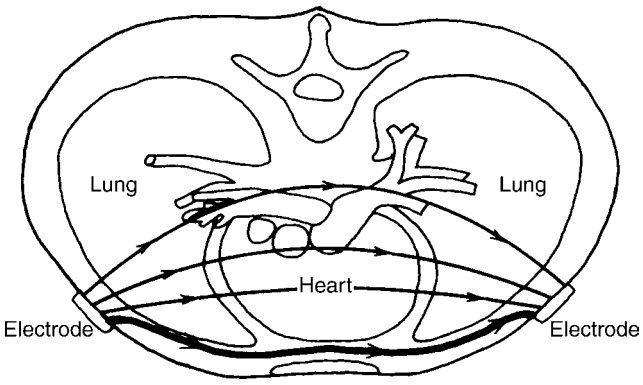


Figure 3. Cross-sectional view of the thorax of an infant showing the current distribution from electrodes placed on the chest wall and excited by a transthoracic impedance type of apnea monitor.

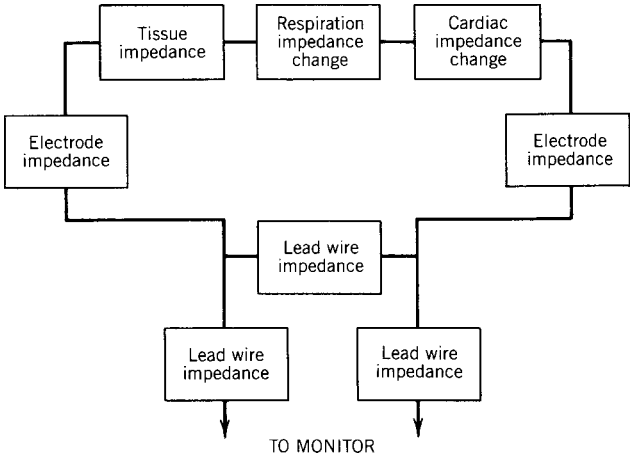


Figure 4. Block diagram of the various impedances seen at the terminals of a transthoracic electrical impedance apnea monitor looking along the lead wires to the patient.

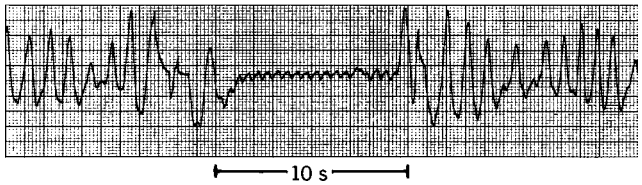


Figure 5. An example of cardiogenic artifact on infant respiration signals from a transthoracic impedance type of monitor illustrating cardiogenic artifact during apnea.

variation associated with the beating heart can be of the same magnitude, although it is generally a little less. Thus, it is seen that a fundamental problem in the indirect measurement of respiration by the transthoracic impedance method is the relatively small changes in impedance associated with the measurement.

To further complicate the situation, each of the nonthoracic impedance components of the circuit illustrated in Fig. 4 can vary in electrical impedance by at least as much if not more than the variation due to respiration. The impedance between the electrode and the infant's skin is strongly dependent on the electrode–skin interface. As electrodes move with respect to the skin, this impedance can vary by amounts much $> 2\ \Omega$. This is also strongly dependent on the type of electrode used and the method that electrically couples it to the skin.

Cardiogenic Artifact

The volume of the heart varies during the cardiac cycle, and so the contribution of the blood to the overall transthoracic impedance will change from systole to diastole. To a lesser extent the vascular component of the chest wall and lungs will also change in blood volume during the cardiac cycle, and this will have some influence on the transthoracic impedance as well, Cardiogenic artifact is illustrated in Fig. 5, which shows a recording of transthoracic impedance from an infant during breathing and during a period of apnea. The cardiogenic artifact is best seen during the apnea, where it appears as a smaller impedance variation occurring at the heart rate. This can be seen by comparing the impedance waveform with a simultaneously recorded electrocardiogram. One notes that the cardiogenic artifact is also present during the breathing activity and appears as a modulation of the respiration waveform.

In the example in Fig. 5 the cardiogenic artifact is relatively small compared to the impedance changes due to breathing, and it is possible to visually differentiate between breathing and apnea by observing this recording. This is not always the case when recording transthoracic impedance as Fig. 6 illustrates. Here one observes periods of breathing and apnea with much stronger cardiogenic artifact. It is difficult to determine what impedance variations are due to breathing and what are due to cardiovascular sources. It is only possible to identify periods of respiration and artifact when the recording is compared with a simultaneous recording of respiration from a recording of abdominal wall movement using a strain gage as shown in Fig. 6. Note that in the case of the impedance signal in this figure, the cardiogenic artifact has two components during each cardiac cycle.

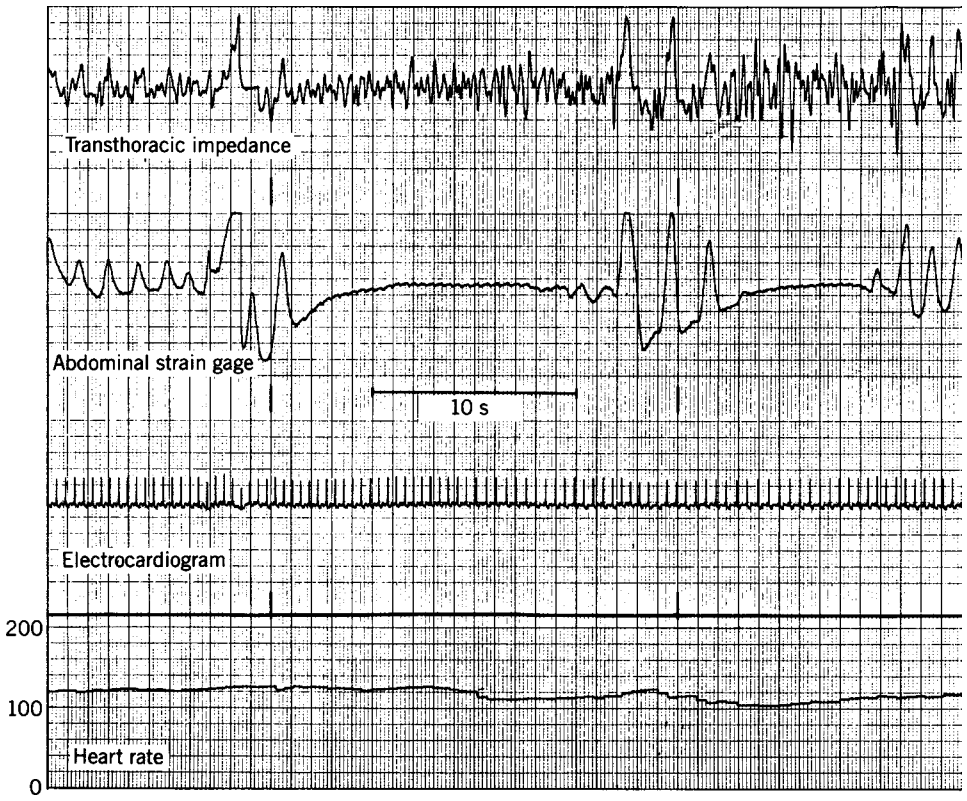


Figure 6. High amplitude cardiogenic artifact is shown on the transthoracic impedance tracing from this recording of multiple signals from a newborn infant. In this case, the transthoracic impedance changes correspond to the electrocardiogram shown on the third trace from the top. Simultaneous recordings from a nasal thermistor and an abdominal strain gage do not show these high frequency variations.

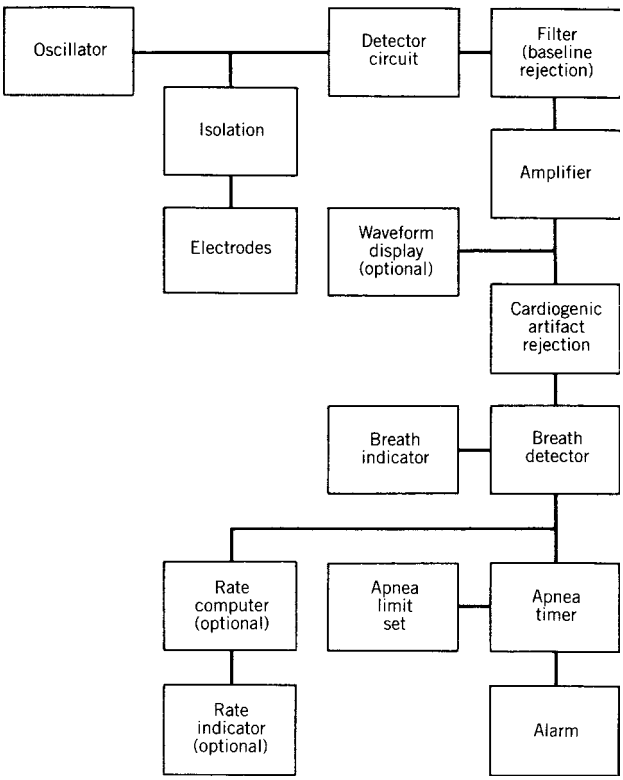


Figure 7. Functional block diagram of a transthoracic impedance infant apnea monitor.

Signal Processing

The electrical signals from the electrical impedance sensor, or any of the other respiration sensors, must be processed to recognize breathing activity and to determine when apnea is present. Different sensors require processors of differing complexity because of different signal characteristics, but the general method of signal processing is the same no matter what sensor is used. The signal processing associated with the electrical impedance method of apnea monitoring will be described in the following paragraphs, since it is one of the most complex as well as most highly developed monitoring systems.

A block diagram of the generalized sections of a transthoracic electrical impedance type of apnea monitor is shown in Fig. 7. The basic functions of the system can be broken down into impedance measurement, breath detection, artifact rejection, apnea identification, and alarm functions. Each of these can be carried out with varying degrees of complexity, and sophisticated signal processing techniques can be used to get the most information out of a less than optimal signal.

A signal generator in the impedance measurement portion of the system produces the excitation signal that is applied to the electrodes. This can either be a sinusoidal or a square wave, and frequently will have a high source impedance so that it behaves as though it was generated by a constant current amplitude source. Passing this current through the lead wire–electrode–body system causes a voltage amplitude proportional to its impedance to appear at the monitor input. Variations in this voltage reflect the variation in impedance. It is therefore important that the current amplitude of the excitation signal remain constant during a measurement. Excitation signal frequency is chosen to be in the range of 20–100 kHz so that electrode–body interface impedances are relatively low, thereby producing less artifact. Detection of individual breaths from a complex breathing signal represents a major task for the respiration monitor. While the design of electronic circuits to carry out such a function on a regular, noise-free, nearly constant amplitude respiration signal such as seen in Fig. 8a presents no problem; very often the respiration waveform is much more complicated and not so easily interpreted, as illustrated in Fig. 8b. Cardiogenic artifact also helps to complicate the signal detection problem since in some cases it can masquerade as a breath. Some of the basic methods of identifying breaths are listed in the following paragraphs. Often individual monitors will use more than one of these in various unique signal processing algorithms.

Fixed Threshold Detection. A breath can be indicated every time the respiration signal crosses a predetermined fixed threshold level. It is important to carefully choose this level so that nearly all breaths cross the threshold, but

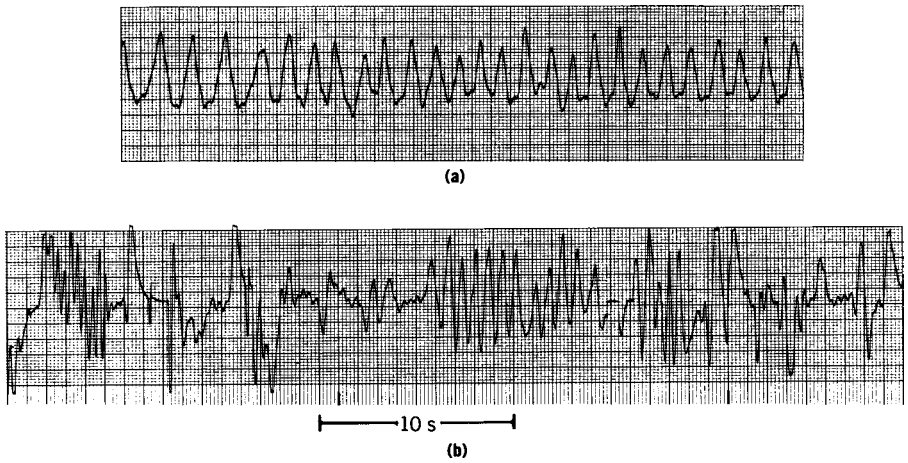


Figure 8. Typical infant respiration signals obtained from infant apnea monitors. (a) The top trace illustrates a relatively quiet signal that can be processed to determine respiration rate and apnea. (b) The bottom trace is a typical example of a noisy signal resulting from infant movement. In this case it would not be easy to determine respiration rate.

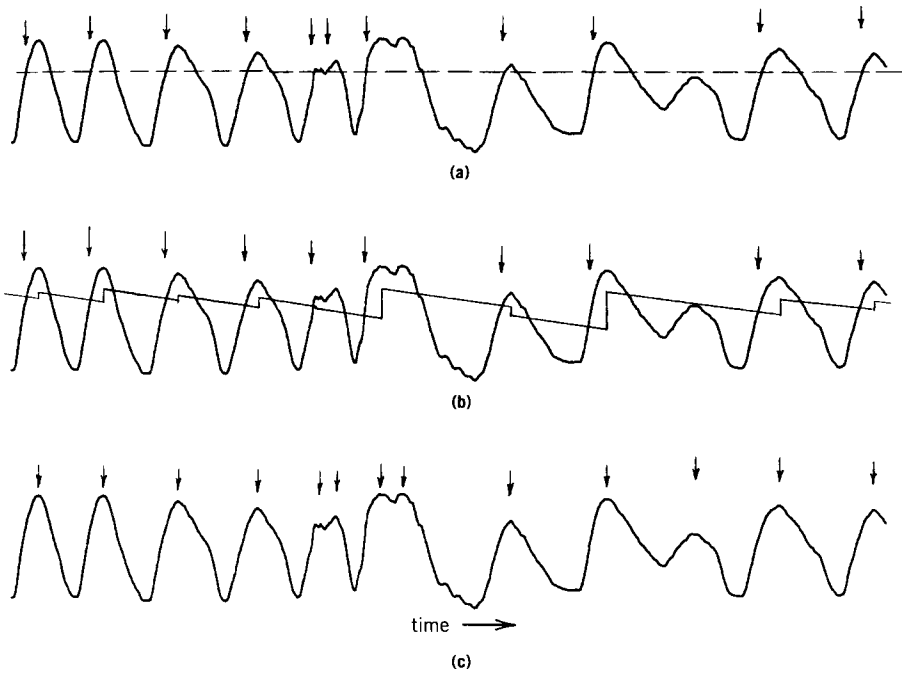


Figure 9. Three identical respiration signals in which different methods of detecting a breath are used. The arrows indicate when the apnea monitor would detect a breath for each method. (a) Fixed threshold detection, note missing breath when signal fails to cross a threshold. (b) Adaptive variable threshold detection, note that breaths can be missed with this method when the amplitude from one breath to the next is significantly different. (c) Peak detector, note that this method can result in double breath detection for signals with multiple peaks.

practically no noise or artifact does. Figure 9a illustrates the basic threshold breath indicator system in which a breath is indicated whenever the signal is greater than the threshold level.

Automatic Gain Control. The fixed threshold method of detection can be improved by preceding the threshold detector with an amplifier that has an automatic gain control. In this way the weaker signals are amplified more than the stronger ones so that all signals appearing at the fixed threshold detector circuit have roughly the same amplitude and will be detected. Although this method makes the fixed threshold detection scheme more reliable, there is also the possibility that noise or cardiogenic artifact will be amplified until it is strong enough to masquerade as a breath during periods of apnea thereby causing the monitor to fail to identify the apnea.

Adaptive Threshold Detection. A variable threshold level can be set by the monitor based upon a preprogrammed algorithm. One common example of this is to have the monitor determine the threshold level based upon the amplitude of the previous breath. This is illustrated in Fig. 9b, where the threshold is set at 80% of the peak amplitude of the previously detected breath. Since this threshold may still be too high if the previous breath had a large amplitude and subsequent breaths were of a relatively low amplitude, this threshold is not fixed, but rather it slowly decreases so that eventually a breath will be detected and the threshold level can be reset. The risk with this type of system is that the threshold will eventually get low enough to detect noise or cardiogenic artifact during an apnea resulting in a breath detected in error. Thus, the algorithm for this adaptive system must have minimum threshold levels that are still well above the noise or cardiogenic artifact level for it to work effectively.

Peak Detector. This circuit recognizes the maximum value of a signal over a short interval of time regardless of the overall amplitude of that signal. The way that a peak detector detects the breaths from a typical respiration waveform is illustrated in Fig. 9c. The basic peak detector can recognize more than one peak in a complex respiration wave. This can give errors if the monitor is used to determine respiration rate. Again, by adding complexity to the signal processing algorithm, this type of error can be greatly reduced.

Filtering. Frequency spectral analysis of infant respiration signals shows that most of the information is contained in the frequency band of 0–6 Hz, and in many cases the band is even narrower (20). Since artifactual signals can exist both within and outside of this frequency range, most apnea monitors filter the respiration signals so that only the frequencies containing information are processed. The type of filtering used depends on the particular monitor design, but any process of filtering can distort the waveforms and may itself introduce artifact. This is especially true when high pass filtering is used to remove the baseline. Thus, filtering can affect the performance of the breath detection method used in the instrument.

Although filtering is an important aspect of the breath detection circuitry, it can in some cases cause motion artifact to begin to look similar to a respiration signal and thus allow the detection circuit to recognize artifact as a breath. Often under the best conditions it is difficult to discriminate between artifact and true breathing signals, and the filtering only further complicates this problem. Nevertheless, without filtering breath detection would be much more difficult.

Pattern Recognition. Computer technology allows algorithms for recognizing various features of the respiration

waveform to be applied for breath detection in infant respiration monitors. Features, such as threshold crossing, peaks and valleys, slopes, amplitudes, width, and interval of a respiration wave can be readily detected. More sophisticated algorithms can be trained to recognize breaths that are similar in appearance to preprogrammed waveforms or based on the appearance of previous breaths for a particular patient. Another important aspect of computer recognition of patterns is that the computer can be programmed to ask various questions: Is the measured value physiologically possible? Does the waveform look more like artifact than information? Is the rate too fast? Does the signal correspond too closely to the cardiac cycle so that it might be cardiogenic artifact? Is there more than one peak per breath? All of these techniques of breath detection have advantages and disadvantages for infant monitoring. Each technique, however, imposes constraints on the signal that determine whether it will also detect artifact or miss some true breaths. Even the most sophisticated computer methods suffer from faults such as these and present limitations in breath detection.

Cardiogenic Artifact Rejection. Although cardiogenic artifact represents a major problem when breathing efforts are measured by the transthoracic electrical impedance method, this interference can be seen at times in the output of other indirect sensors of respiration as well. Usually, for these other sensors this artifact is small and does not pose any problem in breath or apnea recognition. Several methods have been used to reduce the problems associated with cardiogenic artifact in the transthoracic electrical impedance type of apnea monitor. Cardiogenic artifact occurs at the heart frequency and its harmonics, which can be different from the periodicity of the respiration signal. In infants the heart rate is usually higher than the respiration rate, although this is not always the case since infants can breath quite rapidly. If the respiration signal containing cardiogenic artifact is passed through a low pass filter having a cutoff frequency that is higher than the expected respiration rates but lower than the heart rates likely to be encountered, much of the cardiogenic artifact can be removed without seriously distorting the respiration signal. The problem with this approach is the selection of a cutoff frequency for the filter. It is generally not possible to find a frequency that is greater than the maximum respiration rate yet less than the minimum heart rate for small infants. Estimated values of such a frequency have to be changed according to the age of the infant, and since bradycardia can be associated with apnea, it is possible that the heart frequency will drop below the filter cutoff frequency during times of apnea, allowing cardiogenic artifact to get into the respiration channel just at the very time when it should be avoided.

The approach of using a filter, however, has merit if the above limitations can be taken into consideration in the design of the filtering system. Although there is no way that a filter can be useful when the heart rate is less than the respiration rate, the filter can help if its cutoff frequency is based upon the apparent respiration and heart rates of the infant. Such adaptive filtering techniques have been successfully used to minimize the effects of cardiogenic artifact.

Since most transthoracic electrical impedance apnea monitors also determine heart rate from the electrocardiogram, this cardiac signal can be used to help identify when a respiration signal consists primarily of cardiogenic artifact. The temporal relationship between the cardiogenic artifact and the electrocardiogram should be constant since both come from the same source. If the respiration signal consists only of cardiogenic artifact, as would be the case during a period of apnea, it is possible to identify the fixed temporal relationship between the signal and the electrocardiogram and therefore reject the signal from being accidentally detected as a breath. The only limitation with this technique is that in rare cases the infant can breath at the same rate as the heart is beating, and the monitor would indicate that an apnea had occurred when in fact it had not.

COMBINATION TRANSTHORACIC IMPEDANCE AND CARDIAC MONITORS

Most commercially available infant apnea monitors take advantage of the fact that the same sensor system, a set of biopotential electrodes, can be used for both transthoracic electrical impedance respiration monitoring and cardiac monitoring. Since the excitation signal for transthoracic impedance monitoring has a frequency of 20 kHz or greater and the highest frequency component of the infant electrocardiogram is < 200 Hz, the excitation signal can be applied to the same electrodes used for obtaining the electrocardiogram. By connecting a low pass filter between the electrodes and the heart rate monitor circuit, this excitation signal can be kept out of the cardiac monitor, and a bandpass filter in the respiration monitor centered at the excitation signal frequency will keep the electrocardiogram and biopotential motion artifact out of the transthoracic impedance monitor circuit.

The combination of respiration and heart rate monitoring in a single instrument helps to identify life-threatening events. If for some reason the respiration monitor fails to recognize prolonged apneas, bradycardia will often be associated with such episodes, and the heart rate monitor will recognize the reduced heart rate and set off an alarm.

MEASUREMENT OF BLOOD GASES

Blood gases refer to the oxygen and carbon dioxide transported by the blood. Acid–base balance is also included in discussions of blood gases since it is closely related to respiratory and metabolic status. Thus, measurements of blood gases are frequently combined with measurements of blood pH. There are invasive and noninvasive methods of measuring blood gases. Both can be used for hospital monitoring of critically ill infants. The principal methods that are used are described in this section.

Invasive Methods

Invasive blood-gas measurement techniques involve direct contact with the circulatory system so that blood samples can be drawn and measured in a laboratory analyzer or a

miniature sensor can be placed within the blood stream for continuous measurements. Some of these methods are described in the following paragraphs.

Intraarterial Catheter. The newly born infant has an advantage over other medical patients in that the vessels of the umbilical cord stump can accept a catheter for several hours after birth. Thus, it is possible to introduce a fine-gage, flexible, soft catheter into an umbilical artery of a cardiac or respiratory compromised infant and advance the tip into the aorta so that samples of central arterial blood can be obtained for analysis. Blood samples with a volume of only 50 μL can be analyzed for pH, P_{O_2}, and P_{CO_2} by means of specially designed miniaturized versions of standard analytical chemistry sensors of these variables. Such microblood analyzers are also used for analyzing fetal scalp blood samples. It is important in neonatal applications that only microblood analyzers be used since the total blood volume of very small infants is limited. Since an infant's blood gas status can be labile, it is often necessary to draw many blood samples during the clinical course of care, thus significant blood loss can occur unless very small samples are taken.

Microblood analyzers generally use the inverted glass electrode for pH measurement, a miniaturized Clark electrode for P_{O_2} measurement, and a miniaturized version of the Stowe–Severinghaus sensor for P_{CO_2}. The technology of microblood gas analyzers is well developed, and devices perform reliably in the intensive care situation. Instrumentation is frequently located within the neonatal intensive care unit itself, and respiratory therapists for collecting samples and carrying out the analyses as well as calibrating and maintaining the analyzers serve round the clock.

The major limitation of this sampling technique is that the sample only represents the blood gas status at the time it was taken. Thus, frequent samples must be taken during periods when variations can occur to track these variations, and even with microblood analyzers this can sometimes result in significant blood loss for very small infants. If the method for drawing the blood sample from the infant is stressful, such as a painful vascular puncture or heel stick, the blood gases of the sample will probably not reflect the quiescent status of the patient. As a matter of fact the very act of obtaining the blood sample may be of some risk to the infant since it can temporarily increase hypoxia (21).

The umbilical vessel canulation is not without problems itself. In placing the catheters, one must be careful not to damage the lining of the vessels or perforate a vascular wall resulting in severe bleeding or hemorrhage. Catheters must be made of materials that do not promote thrombosis formation. When catheters are not used for drawing blood, they must be filled with a physiological solution containing an anticoagulant such as Heparin so that blood that diffuses into the tip of the catheter does not clot. Any thrombi formed on the catheter wall or within its lumen can break off and cause embolisms further downstream. For arterial catheters this can be in the blood supply to the lower periphery of the infant, and it is possible to see under perfused feet in infants having an umbilical artery catheter. Catheters in the umbilical vein or peripheral veins can also produce emboli. In this case the clots are returned to the right side of the heart and can go on to produce pulmonary emboli.

Peripheral Blood Samples. Although it is frequently possible to introduce a catheter into an umbilical artery in a newly born infant, this is not always the case, or the need for blood gas monitoring may not arise until the infant is sufficiently old that the umbilical vasculature has permanently closed. In this case it is necessary to canulate a peripheral artery to obtain frequent arterial blood samples. On very small infants this is no minor task since these vessels are very small and difficult to canulate transcutaneously.

An alternative to drawing an arterial blood sample is to take a sample of capillary blood from the skin under conditions where the capillary blood flow has been significantly increased so that the capillary blood appears to be similar to peripheral arterial blood. This can be done, for example, in the heel by first warming an infant's lower leg and foot by wrapping it with warm, wet towels. A blood sample of sufficient size for a microblood analyzer can then be obtained by making a small skin incision with a lancet and collecting the blood sample in a capillary tube in a fashion similar to the technique for obtaining a fetal scalp blood sample (see FETAL MONITORING). Although this technique is not as reliable as sampling from an umbilical artery catheter, it can be used when only a single blood sample is desired and an umbilical catheter would be inappropriate or where it is not possible to place such a catheter. An important limitation of the technique is that the infant's heels can become quite bruised when frequent samples are required and suitable locations for additional samples might no longer be available. When frequent samples are required, it is generally better to attempt canulation of a peripheral artery.

Internal Sensors. Blood gases can be continuously monitored from invasive sensors. Generally, these sensors are incorporated into umbilical artery catheters (22), but tissue measurements have also been demonstrated (23). The most frequently applied technique involves the incorporation of an amperometric oxygen sensor into a catheter system. This can be done either by incorporation of the sensor within the wall of the catheter, by using a double lumen catheter with the sensor in one lumen and the second lumen available for blood samples or infusion, or by using a conventional single lumen catheter with a sensor probe that can be introduced through the lumen so that the sensor projects beyond the distal tip of the catheter.

Oximetry, the measurement of hemoglobin oxygen saturation, can be carried out continuously by means of optical sensors coupled to intravascular catheters or probes. Optical fibers can be incorporated in the wall of a catheter or in an intraluminal probe and used to conduct light to the catheter's distal tip. The light illuminates the blood in the vicinity of the catheter tip, and an adjacent fiber or bundle of fibers collects the backscattered light and conducts it to a photo detector where its intensity is measured. By alternately illuminating the blood with light of two or more different wavelengths, one of which is close to

an isosbestic point, and measuring the backscattered light, it is possible to determine the hemoglobin oxygen saturation in the same way as done in laboratory instruments for *in vitro* samples.

Advantages and Disadvantages of Invasive Techniques. The methods described in the previous sections represent direct measurements in that the sensor that is used is in direct contact with the body fluid, usually blood, being measured. This direct contact improves the possibility of accurate measurements. When the sensor is not located in the blood itself but is used to measure samples of blood drawn from the patient, instruments can be frequently calibrated using laboratory standards. Sensors that are used within blood or other tissues have the requirement that they must be small enough to fit in the tissue with minimal damage, either as a part of a catheter or some other probe. The miniaturization process must not compromise accuracy or reproducibility. In cases where microelectronic technology can be used to miniaturize the structures, reproducibility can even be improved in the mass-produced miniature devices as compared to their piece-by-piece-produced larger counterparts. The continuous invasive sensors are also limited in where and when they can be applied. While the umbilical arteries are convenient conduits to the central arterial circulation, they are only patent for a few hours after birth in most newborn infants. Following this time it is very difficult to obtain arterial samples since other vessels must be used. The use of intravascular sensors, and those in tissue as well, also increases the risk of infection and mechanical damage. Care must be taken with intraarterial sensors to avoid serious hemorrhage due to system components becoming disconnected.

Noninvasive Methods

In noninvasive measurement of blood gases, there is no direct contact between the blood or other tissue being measured and the sensor. In this way there is usually less risk to the patient and the technique is easier to apply clinically. The major noninvasive methods used in neonatal monitoring are now described.

Transcutaneous Blood Gas Tension Measurement. One of the major advances in neonatal intensive care monitoring technology was the development of transcutaneous blood gas measurement instrumentation. This allowed the oxygen tension and later the carbon dioxide tension of infants at risk to be continuously monitored without invading the circulatory system (24). These methods make use of a heated sensor placed on the infant's skin that measures the partial pressures of oxygen or carbon dioxide of the blood circulating in the dermal capillary loops under the sensor. The heating of the skin to temperatures of 44 °C arterializes the capillary blood in a manner similar to that used for obtaining capillary blood samples with heel sticks. Although the heating of the blood increases the blood gas tensions in the capillary blood, oxygen consumption by the viable epidermis surrounding the capillaries and diffusional drops through the skin compensate for this increase resulting in good correlations between the transcutaneously measured blood gas tensions and those determined from arterial blood samples in neonates. Sensors can be left in place on neonates for up to four hours, but for longer periods of time it is recommended to move the sensor to a new location to avoid tissue damage due to the elevated temperature. Multiple sensors have been developed in which the heating element is switched between several sensors in the same package periodically so that the overall sensor can be left in place for longer periods of time without producing damage (25).

Although transcutaneous instrumentation can give good correlations between transcutaneous and central arterial blood gas measurements in neonates, it would be misleading to suggest that the transcutaneous instrument is measuring the same thing as is measured from arterial blood samples. Indeed in infants with unimpaired circulatory status, the transcutaneous blood gases and those in the central circulation are similar; however, when there is cardiovascular compromise, heating of the sensor can no longer completely arterialize the capillary blood, and there are significant differences between the transcutaneous and central measurements. Thus, when one makes both transcutaneous and central measurements, differences can be used as a means of identifying shock-related conditions (26).

Transcutaneous Mass Spectrometry. Another noninvasive method for measuring blood gas tensions involves the use of a transcutaneous mass spectrometer (27). A sensor similar to the transcutaneous blood gas sensor in that it contains a heater to arterialize the capillary blood under it is made of a gas-permeable membrane in contact with the skin. This is connected to the mass spectrometer instrument through a fine-bore flexible tube through which an inert carrier gas is circulated to bring the gases that diffuse from the skin into the sensor to the instrument. (For details of this instrument see MASS SPECTROMETERS IN MEDICAL MONITORING). At the present time, mass spectrometry instruments are far more expensive than instruments for electrochemically determining the transcutaneous blood gas tensions. The advantage of the mass spectrometer, however, is that it can simultaneously measure more than a single blood gas component. It can also measure other gases in the blood stream, such as anesthetic agents or special tracers.

Pulse Oximetry. The use of optical techniques to determine the hemoglobin oxygen saturation in blood is well known and is the basis for routine clinical laboratory instrumentation along with the fiber optic catheter oximeter described in the previous section on internal sensors. Oximeters have also been developed for measuring the oxygen saturation transcutaneously. Initial devices measured the continuous steady-state reflection of light of different wavelengths from the surface of the skin. Pigmentation of the skin, unfortunately, limited this technique to qualitative measurements unless the instrument was specifically calibrated to a particular individual at a particular site. Upon examining the backscattered optical signal from the skin, one can notice a small pulsatile

component at the heart rate. This is due to the changing blood volume in the capillary beds reflecting the light, and it can be seen for transmitted light as well. By looking at this pulsatile component of the transmitted or reflected light, it is possible to measure only the effect of each fresh bolus of blood entering the capillary bed at systole. This allows the principle of oximetry to be adapted to the transcutaneous measurement of arterial blood hemoglobin oxygen saturation (28). This technique is used for continuously monitoring tissues that can be transilluminated, such as the hand, foot, fingers, toes, ears, and nasal septum. These pulse oximeters have the added advantage that in most applications it is not necessary to arterialize the capillary blood by heating; thus, sensors can be left in place for longer periods of time without risk of tissue injury.

Pulse oximeters have rapidly achieved a major role in neonatal and adult intensive care medicine. It is important to point out that oximetry differs from oxygen tension measurement in that it tells how much oxygen is carried by the hemoglobin. To know total oxygen transport one needs to know the amount of hemoglobin in the blood as well as the profusion of the tissue in question. Thus, oximetry can with some additional data be quite useful in determining whether adequate amounts of oxygen are being supplied to vital tissues. There is one aspect of neonatal monitoring, however, where oximetry is of little assistance. The condition, known as retinopathy of prematurity, is found in premature infants and thought to be related to the newly formed capillaries in the retina, which are exposed to blood of elevated oxygen tension in infants who are receiving oxygen therapy. An important aspect of oxygen monitoring in premature infants is to determine if the arterial blood oxygen tension becomes elevated, so that the amount of oxygen that the infant breathes can be reduced to protect the eyes. If retinopathy of prematurity occurs as a result of elevated oxygen tensions, blindness can result. Thus, to truly protect the patient from this condition, one must measure oxygen tension not hemoglobin oxygen saturation.

Pulse oximeters in routine clinical use are primarily based on the the transmission mode of operation, although backscatter oximeters have also been developed (29,30). The clinical instruments, therefore, are limited in terms of where they can be attached to the subject. Generally, these positions are found on the periphery and are, unfortunately, the first to experience diminished circulation under shock or preshock conditions. Another limitation of currently available pulse oximeters is their great sensitivity to motion artifact. Signal processing algorithms have been developed to reduce the effect of motion on the pulse oximetry signal and to detect motion artifact and prevent it from being indicated as data (31). Nevertheless, since the oxygen saturation values presented represent averages over several heartbeats, movement can result in an apparent decrease in oxygen saturation that in fact has not occurred.

TEMPERATURE MONITORING

An important aspect of treating premature infants is to the maintenance of their thermal environment. A premature infant is not well adapted to extrauterine life, and its temperature control system is not fully developed since it normally would be in a temperature regulated environment in the uterus. Thus, an artificial environment must be provided to help the neonate control its body temperature. This environment is in the form of convective incubators and radiant warmers. Another reason for providing an elevated temperature environment for premature infants is that very often these infants suffer from problems of the respiratory system that limit the amount of oxygen that can be transported to the blood by the lungs. This oxygen is utilized in the metabolic processes of the infant, and among these are the generation of heat to maintain body temperature. By placing the infant in an environment at a temperature greater than normal room temperature, less energy needs to be expended for thermal regulation. A neutral thermal environment can be found where the temperature and relative humidity are such that the infant utilizes a minimum amount of energy to maintain its temperature, and oxygen and nutritional substrates that would normally go into heat generation can be utilized for metabolic processes related to growth and development. Thus, to maintain this environment, it is necessary to monitor both the temperature of the infant and that of the environment.

Temperature monitoring instrumentation in the nursery is relatively straightforward. The sensor is a thermistor that can be in one of two basic forms, an internal probe or a surface probe. The former consists of a semiflexible lead wire with a thermistor mounted at its distal tip. An electrically insulating polymer with good thermal conductivity covers the thermistor and is contiguous with the lead wire insulation. This probe can be placed rectally to give a neonatal core temperature measurement. The surface probe is a disk-shaped thermistor $\sim$ 6 mm in diameter with lead wires coming out in a radial direction. The sensitive surface of the probe is metallic and is in intimate contact with the thermistor, while the other surface of the probe is covered with a thermally insulating polymer so that the thermistor is well coupled to the infant surface it contacts through the metal but poorly coupled to the environmental air. The surface temperature measured is not necessarily the same as core temperature and is strongly dependent on the infant's environment. Often the surface mounted probe is placed over the liver since this organ is highly perfused and is close to the skin surface in small infants. To aid and maintain a good thermal contact between the surface probe and the infant skin, the lead wires, especially near the probe, should be highly flexible so that the wires do not tend to force the thermistor to come loose from the skin as the infant moves. As was mentioned for surface mounted biopotential electrodes, the skin of premature infants is sensitive to many factors, and strong adhesive can produce severe irritation. Weaker adhesives, however, can allow the thermistor to come off, and the use of flexible lead wires greatly reduces this tendency.

The remainder of the instrumentation in temperature monitoring devices is straightforward. An electronic circuit senses the resistance of the thermistor and converts this to a display of its temperature. In some cases alarm circuits are incorporated in the monitors to indicate when the

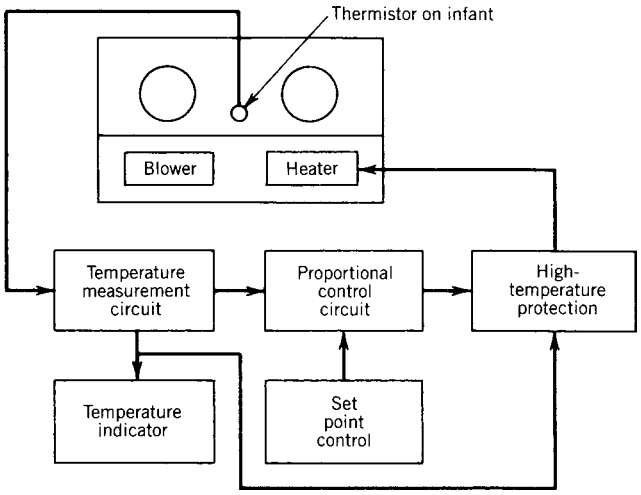


Figure 10. A servo-temperature control system for incubator temperature directed toward maintaining infants at a preset temperature.

temperature lies outside of a preset range. Temperature instruments are used not only for indicating infant surface and core temperatures, but also for the control of incubator or radiant warmer temperature. Although convective incubators have internal control systems to maintain the air temperature at a preset point, the purpose of the incubator is not as an air temperature controller. Instead the incubator is used to maintain the infant's body temperature at a certain point and to minimize thermal losses from the infant. For this reason some incubators have servo systems that control incubator temperature based on infant temperature rather than air temperature. A block diagram of such a system is illustrated in Fig. 10. Here a thermistor is the sensor and is positioned on the infant's skin or internally. If a radiant warmer is used, it is important that any surface mounted thermistor is not in the radiant field and thus directly heated by the warmer. Frequently thermistors are covered with an insulating disk that has a highly reflective outer surface of aluminum foil so that no direct radiant energy from the heater falls on the thermistor. An electronic circuit determines the thermistor resistance, and hence, its temperature, which is assumed to be equivalent to the infant's temperature. This drives a control circuit that provides proportional control to the heating element of the convective or radiant source. In some cases integrating and differential control is added to the system for optimal response. Additional safety circuits are included in the system to prevent it from overheating or underheating, both of which are undesirable for the infant. The final block of the system is the heater itself. The control system must take into account the time response of this element so as to provide optimal control. An indicator is frequently included in the system to show infant temperature and heater status.

PRESSURE MEASUREMENT

The measurement of the pressure in fluids is important in many aspects of medical care. This is especially true in neonatal monitoring, and instrumentation for the intermittent or continuous measurement of blood pressure is frequently used in the intensive care unit. There are also situations where the monitoring of intracranial pressure is important in the care of infants. Instrumentation for measuring these pressures is similar to other monitoring instrumentation described in this article in that measurements can be made by direct and indirect means. These are described in the following paragraphs.

Blood Pressure

The direct measurement of blood pressure consists of coupling the arterial or venous circulations to a pressure transducer that is connected to an electronic instrument for signal processing, display, and recording. The direct methods used are similar to those used in adult intensive care (see BLOOD PRESSURE MEASUREMENT). Generally, measurements are only made on the arterial circulation, and wherever possible an umbilical artery is used for access to this circulation. An umbilical artery catheter, such as described in the section on direct measurement of blood gases, is filled with a physiologic saline solution containing an anticoagulant. The proximal end of this catheter is connected to an external pressure sensor that is positioned in the incubator near the infant. This is usually a disposable semiconductor pressure sensor that is used only on a single infant and then discarded so that there is no risk of cross-contamination from one patient.

Indirect blood pressure monitoring in the neonate presents special problems not seen in the adult. It is generally not possible to measure an infant's blood pressure using a sphygmomanometer and the auscultation technique because Korotkoff sounds cannot be detected. Thus other, more complicated methods of indirectly measuring blood pressure must be used. If a sphygmomanometer cuff around a limb is still employed, it is important to use the correct size of cuff for the infant being studied. The width of the cuff should be from 45 to 70% of the limb circumference (32). Cuffs of several different sizes are, therefore, available for use with infants. These frequently are inexpensive disposable cuffs designed for use with a single infant.

Systolic and diastolic pressures can be sampled noninvasively using the oscillometric or the kinarteriography methods. Both techniques (see BLOOD PRESSURE MEASUREMENT) are based upon blood volume changes in the section of artery under or distal to the sphygmomanometer cuff. In the case of the oscillometric measurement, the actual volume changes are determined, while the kinarteriography method measures the radial velocity of pulsations in the arterial wall. In the former case pressure variations in the cuff itself are sensed, and signal processing allows mean arterial pressures as well as systolic and diastolic pressures to be determined.

The kinarteriographic technique utilizes an ultrasonic transducer under the cuff. Continuous wave ultrasound is beamed at the brachial artery, when the cuff is on an arm, and some ultrasonic energy is reflected from the arterial

wall. This is picked up by an adjacent ultrasonic transducer, and an electronic circuit determines the frequency differences between the transmitted and reflected waves. When the arterial wall is in motion, the reflected ultrasound is shifted in frequency thereby giving a frequency difference between the transmitted and reflected waves. Motion of the arterial wall is greatest when the cuff pressure is between systolic and diastolic pressures; and thus by measuring changes in the frequency shift of the reflected wave, it is possible to determine the systolic and diastolic pressures. Unlike the oscillometric technique, it is not possible to determine the mean arterial pressure with kinarteriography.

Monitoring of Intracranial Pressure

As was the case with blood pressure, intracranial pressure (the pressure of the cerebrospinal fluid and brain within the cranium) can be determined by direct and by indirect methods. The former involves the placement of a tube within the brain such that its distal tip communicates with the intraventricular fluid. The proximal end is connected to a low compliance pressure transducer. Although this technique is highly accurate, a significant risk of infection is associated with its application, and it is only used under extreme circumstances when no other technique is possible.

Noninvasive techniques of monitoring neonatal intracranial pressure are much safer than the direct technique but, unfortunately, are not as accurate. The newborn infant not only has special access available to the central circulation through the umbilical cord, but also has a means of accessing the intracranial contents through the anterior fontanel. This gap in the calvarium means that only soft tissue lies between the scalp surface and the dura mater. Thus, it is possible to assess intracranial pressure through this opening by various techniques.

A skillful clinician can palpate the anterior fontanel and determine to some extent whether the pressure is elevated or not (33). Another clinical method that can be used is to observe the curvature of the scalp over the fontanel as the position of the infant's head with respect to its chest is changed (34). The curvature should flatten when intracranial and atmospheric pressures are equivalent. These techniques are highly subjective and not suitable for neonatal patient monitoring; however, they can be the basis of sensors for more objective measurement.

Various forms of tonometric sensors have been developed for noninvasively measuring and monitoring intracranial pressure (35). These all consist of some sort of probe that is placed over the anterior fontanel in such a way that the curvature is flattened and formed into a plane normal to the surface of the probe itself. guard rings, calibrated springs, and other techniques have been used to achieve this applanation. Ideally once this is achieved the pressure on either side of the membrane consisting of the soft tissue between the dura and the scalp surface should be equal. Thus, by sensing the pressure on the probe side, one can determine the pressure on the other side of the dura. Unfortunately, such a situation only holds in practice when the membrane is thin with respect to the size of its planar portion. In the case of the soft tissue between the dura and the scalp, the membrane thickness can often be close to the size of the planar portion because of the limitations imposed by the opening of the fontanel. Thus, the technique has some definite limitations. The method of attachment of the probe to the infant and the structure of the probe itself are critical to the efficacy of the resulting measurements.

Many investigators have considered different approaches to making an appropriate probe for transfontanel intracranial pressure measurement. These range from strain-gage-based force sensors with guard rings to transducers that attempt to achieve applanation by means of a compliant membrane mounted upon a chamber in which air pressure can be varied. In the case of this latter technique, the position of the membrane is detected and a servo control system is used to adjust the pressure within the chamber so that the membrane presses the tissue of the fontanel into a flat surface. Such a device is shown is schematically is Fig. 11. The position of the membrane is established optically by means of a shutter attached to the membrane such that it varies the amount of light passing from a fiber optic connected to a light source to one connected to a light detector. A servo system controlling the air pressure within the structure adjusts it so that the diaphragm, and hence the tissue of the fontanel, is flat. At this point, the pressure of the air within the sensor should theoretically equal that of the tissue within the fontanel, which in turn should give the intracranial pressure.

Elevated intracranial pressure in infants can be the result of volume occupying lesions, excessive secretion of fluids within the epidural space, the brain tissue itself, or fluid in the ventricles of the brain. A frequent form of lesion is bleeding or hemorrhage within one of these volumes, a condition that is far too often seen in premature infants. Another form of elevated intracranial pressure results from hydrocephalus, a condition in which intraventricular fluid volume and pressure become elevated. By continuous monitoring or serial sampling of intracranial pressure, it

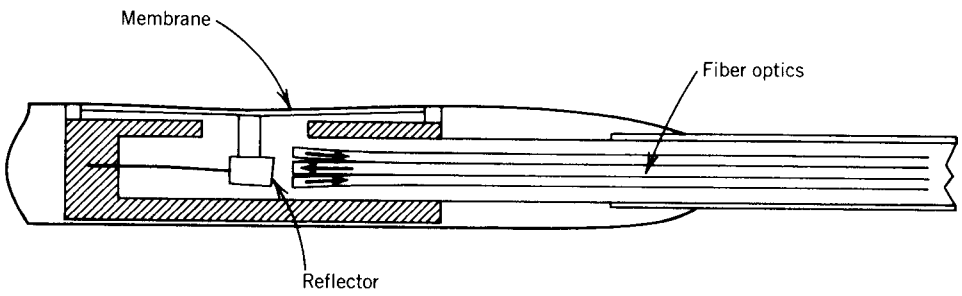


Figure 11. A pressure sensor used far intracranial pressure measurements in the newborn based upon measuring membrane deflection by means of fiber optics.

is possible to provide better control of therapy for these problems.

Monitoring of Intracranial Hemorrhage

As pointed out in the previous section, one cause of elevated intracranial pressure in the newborn is intracranial hemorrhage. It is important to be able to detect when this occurs so that therapeutic measures can be immediately taken to minimize irreversible damage. Although no technique is suitable for continuous monitoring of infants at risk of intracranial hemorrhage, several techniques can be used for surveillance and periodic sampling especially of infants showing possible signs and symptoms of intracranial bleeding.

Devices for the noninvasive measurement of intracranial pressure have been described in the previous section. In addition to these, intracranial hemorrhage can be identified by measurement of transcephalic electrical impedance (36). In this case, it is the baseline value of the impedance that is important. For this reason, a tetrapolar method of measurement must be used to minimize the effects of electrode and lead wire impedances. An alternating current at an excitation frequency of 20–100 kHz is passed between a frontal and occipital electrode placed on the infant's head. A second pair of electrodes are located near the excitation electrodes, but far enough to avoid voltage drops due to the spreading current at the excitation electrodes. The signal is picked up by these electrodes and detected by an electronic circuit to give a voltage proportional to the baseline impedance value. Since the specific impedance of blood is ~ 2.5 times greater than the specific impedance of the cerebral spinal fluid, one should see an elevated transcephalic impedance when the ventricles are filled with blood rather than cerebral spinal fluid. Similarly, if the ventricles have grown in volume because of excess cerebral spinal fluid as in hydrocephalus, the transcephalic impedance should be lower than expected.

In practice, one can only look for changes in transcephalic impedance in infants and not at absolute baseline values because of differences in geometry from one subject to the next. Typically, the technique requires one or more measurements to be taken during the first 24 h of life on each infant and using these measurements to obtain baseline intraventricular, hemorrhage-free data for that particular infant. Subsequent measurements through the infant's hospital course are compared to these initial measurements, and deviations are noted. Significant elevations in impedance have been associated with the occurrence of intraventricular hemorrhage. Studies have been carried out to show that this impedance shift correlates with intraventricular hemorrhage as found using other diagnostic techniques or, in the case of infants who expire, at autopsy (36). The principal advantage of this method is its relative simplicity and ease of measurement on infants in the intensive care unit.

Ultrasonic determination of intraventricular hemorrhage involves making a B scan of the infant's head and locating the ventricular system (37,38). If the ventricles are filled with cerebral spinal fluid alone, the fluid in the ventricles does not reflect ultrasound, and the ventricles appear clear on the image. If there is blood in the ventricular fluid, ultrasonic echoes are produced by the cellular components of the blood. This causes reflections to appear within the ventricle on the image and allows for a definite diagnosis of intracranial hemorrhage.

As with the transcephalic impedance method, the technique of making measurements on infants is straightforward and can be carried out in the neonatal intensive care unit. The equipment necessary for the measurement, however, is more costly than the impedance; but the results are far more specific to identifying intracranial hemorrhage, and thus this is the current method of choice.

An additional method that can be used for detecting bleeding within the ventricles is the use of computerized tomography (CT) scans (39). While this technique is highly efficacious from the standpoint of identifying intracranial bleeding, it is undesirable because it exposes the developing nervous system to significant amounts of X radiation. It also is necessary to transfer infants to the radiology department where the scanning equipment is located to make the measurement. For severely ill, premature infants, this transfer can significantly compromise their care.

MONITORING BILIRUBIN

Bilirubin is a product of the biochemical degradation of the heme moiety that occurs in the hemoglobin molecule as well as other proteins. It is normally found as a result of red blood cell turnover, but it can be elevated in some hemolytic diseases. This form of bilirubin, known as unconjugated bilirubin, enters the circulation and then the skin and other tissues. When it is present in the skin in sufficient concentration, it causes a yellow coloration known as jaundice. It also enters nervous tissue where, if it reaches sufficient concentration, it can cause irreversible damage.

Increased serum bilirubin can occur either as the result of increased production or decreased clearance by the liver. The former situation can occur in normal and premature infants and is referred to as physiologic jaundice of the newborn. It is usually more severe in prematurely born infants and generally peaks about the third day of neonatal life. There are several modes of therapy that can be used to reduce serum bilirubin once elevated values are detected. The simplest way to detect jaundice in the neonate is to observe the infant's skin and sclera for yellow coloration. Quantitative assessment can be carried out by drawing a blood sample and extracting the cellular components from the serum. The absorption of light at a wavelength of 450 nm in such a serum sample is proportional to its bilirubin concentration. Photometric instrumentation for doing this is readily available in the clinical laboratory and some intensive care units (38). The problem with this method is the need for obtaining a blood sample and the time necessary to transport the sample to the laboratory and analyze it in the photometer. A method for the rapid assessment of serum bilirubin in all infants would represent an improvement over these techniques. Fortunately, a relatively simple optical instrument has been developed for assessing serum bilirubin in infants (40). It is illustrated schematically in Fig. 12 and consists of a xenon flash tube

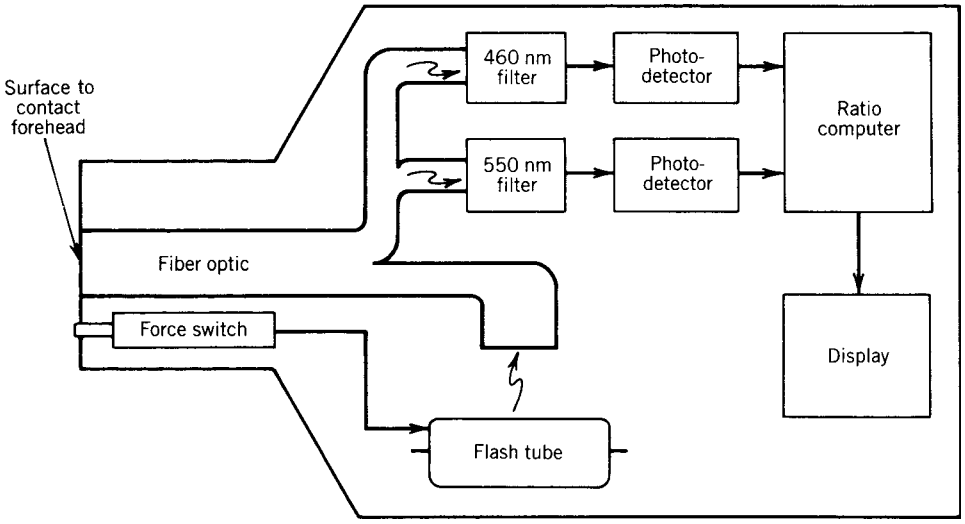


Figure 12. Schematic diagram of a transcutaneous bilirubin instrument.

light source and photometers with filters to measure the reflected light at 460 and 550 nm. A special feature of the instrument is a pressure switch on the portion of the probe that is pressed against the infant's forehead to make the measurement. This probe contains fiber optics that couple the xenon flash tube and the photometers to the skin surface. As the probe is pressed against the skin, the force squeezes the blood from the cutaneous capillaries. Once the force is great enough to sufficiently blanch the skin under the fiber optics so that the blood will not interfere with the measurement, a switch activates the flash tube and readings are taken from the photometers. An internal microprocessor analyzes the reflected light and compensates for any residual hemoglobin. It gives a number proportional to skin bilirubin on a digital display. The proportionality constant, however, differs for different types of neonatal skin; thus, proportionality constants have to be determined for infants of different age, race, and whether they have received phototherapy or not. This instrument involves a sampling technique and thus is only suitable for trend rather than continuous monitoring. Since changes in serum bilirubin are relatively slow, such a technique is entirely appropriate. The principal advantage of this instrument is that it can be readily applied to all infants in the nursery with little effort on the part of the clinical staff. Readings can be made quickly to identify those infants at risk of hyperbilirubinemia.

MONITORING LIFE SUPPORT SYSTEMS

Although one usually associates neonatal monitoring with the measurement of physiologic variables from the newborn or prematurely born infant, an aspect of neonatal monitoring that should not be overlooked is associated with monitoring the patient's environment. Various life support systems are important in neonatal intensive care, and electronic instrumentation for assessing and maintaining the function of these is also important. There should be alarm systems so that when the conditions of life support are inappropriate for neonatal care, care takers are alerted and the problem can be corrected before it causes any harm to the patient. There are many examples of life support system monitoring in the neonatal intensive care unit, and some of the major ones will be described in the following paragraphs.

Maintaining an appropriate thermal environment for the neonate is an important aspect of neonatal intensive care. Incubators and radiant warmers need to have internal temperature instrumentation to ensure that the environment is appropriate for the infant and so that the clinicians providing care can be aware of environmental conditions. Convection incubators frequently have temperature sensors for measuring the environmental air temperature and indicating it on the control panel of the device. These temperature sensors are often a part of the thermal control system in the incubator itself, and as indicated earlier the infant's own temperature can be used as a control signal for some incubators. Built into this incubator temperature monitoring function is an alarm function that can indicate when the incubator temperature becomes too high or too low; potentially life threatening conditions.

It is sometimes necessary to intubate the trachea of a patient and to use a ventilator to control breathing. A gas with elevated oxygen content is often used to help provide additional oxygen to infants who require it. There are many points in a support system such as this where monitoring devices can be useful to assess and in some cases control the function of the device. Where gases of elevated oxygen tension are given, instrumentation to measure the partial pressure of oxygen within the gas to indicate the oxygen fraction of inspired gas is desirable. Various types of oxygen sensors are, therefore, placed in the air circuit to either continuously or intermittently measure and display this quantity. The temperature and humidity of the inspired air are also important, and appropriate sensors and instrumentation for these variables can be included as a part of the respiratory support system. Continuous positive airway pressure is a mode of therapy used in infants requiring ventilatory support. It is necessary to measure and control the positive pressure in such systems to minimize the risk of pneumothorax and to ensure that the desired levels are maintained.

The use of arterial and venous catheters in infants for blood pressure and blood gas monitoring as well as fluid therapy and hyperalimentation represents a safety risk to the infant. Arterial catheters and associated plumbing can become disconnected and cause serious losses of blood that if not quickly checked can result in severe injury or death to the patient. Gas bubbles inadvertently infused along with intravenous fluids can, if sufficiently large, compromise the circulation by producing gas embolisms. Fluid therapy in very small infants must be precisely controlled so that excessive or insufficient amounts of fluid are not provided to the infant. Electronic instrumentation for controlling all of these variables and producing an alarm when dangerous conditions are encountered have been developed (41). Some of these, such as intravenous infusion pumps, are routinely used in neonatal intensive care units. Safety devices for over- or underpressures can be built into the pumps, as can sensors, to indicate when the fluid source has been depleted so that additional fluid can be attached to the pump.

Phototherapy is a technique of illuminating the baby with blue light in the wavelength range of 420–500 nm to oxidize bilirubin to compounds that can be eliminated from the body. Phototherapy units consisting of a group of 20 W fluorescent lamps 30–40 cm above the infant must be used cautiously because there are risks associated with this radiation. Therefore, it is important to determine the amount of radiant energy received by the infant in the 420–500 nm band so that minimal exposure times sufficient to oxidize the bilirubin can be given. Instrumentation consisting of a small probe containing a photosensor that can be held just above the neonate has been developed for this purpose. It is not necessary for this instrumentation to be used to continuously monitor the phototherapy units, but frequent testing of the therapy devices helps not only to determine the appropriate exposure times, but also to indicate when the fluorescent lights become ineffective and need to be changed.

DIAGNOSTIC RECORDINGS

The role of infant monitoring is to determine when events requiring therapeutic intervention occur so that optimal care can be provided. Hard copy recordings from electronic monitoring devices can also be useful in diagnosis of illness and identification of infants who may benefit from electronic monitoring over a longer period of time. Two types of diagnostic recordings are currently used in neonatology: polysomnograms and oxycardiorespirograms although both are still considered experimental and are not routinely used.

Polysomnography. Multiple channel, simultaneous, continuous recordings of biophysical variables related to the pulmonary and cardiovascular systems taken while the newborn or infant sleeps are known as polysomnograms (41–43). These recordings are often made overnight, and 8–12 h is a typical length. The actual variables that are monitored can vary from one study to the next as well as from one institution to the next. However, these usually include the electrocardiogram and/or heart rate. One or more measures of respiratory activity, such as transthoracic electrical impedance or abdominal strain gage; measures of infant activity and movement; measures of infant sleep state, such as eye movements and usually a few leads of the electroencephalogram; and measures of infant blood gas status are also recorded. The number of channels of data in polysomnograms is at least 3 and often is 12 or more. Recordings are made using computer data acquisition systems.

The primary application of polysomnography has been as a tool for research evaluating infant sleep patterns and related physiologic phenomena during sleep. Some investigators feel that polysomnographic recordings are useful in evaluating infants considered to be at risk to sudden infant death syndrome, but at present there is no conclusive evidence that this technique has any value in such screening. There may, however, be specific individual cases where such evaluations may contribute to overall patient assessment.

Oxycardiorespirogram. This technique is used for continuous computer monitoring of infants in the neonatal intensive care unit (44). Four or five physiologic variables are monitored and continuously recorded on a multichannel chart recorder (45). These are the electrocardiogram from which the beat-to-beat or instantaneous heart rate is determined; the respiration waveform or pattern as generally determined by transthoracic impedance monitoring; the respiration rate as determined from the respiration waveform; the oxygen status as determined from a pulse oximeter or transcutaneous blood gas sensors; and at times the relative local skin perfusion as determined by the thermal clearance method from the transcutaneous blood gas sensor.

The importance of the oxycardiorespirogram is that it brings these variables together on a single record, where they can be presented in an organized and systematic fashion. This makes it possible to observe and recognize characteristic patterns between the variables that may be overlooked when all of these quantities are not monitored and recorded together. The oxycardiorespirogram is able to look at variables related to various points along the oxygen transport pathway from the airway to the metabolizing tissue. Thus, it allows a more complete picture of infant cardiopulmonary function than could be obtained by monitoring just one or two of these variables. Typical oxycardiorespirogram patterns have been classified and organized to assist clinicians in providing neonatal intensive care (46).

SUMMARY

Infant monitoring along with adult monitoring under critical care situations involves many individual types of sensors and instruments. Depending on the application, many different types of output data will be produced. In addition, many different conditions for alarms to alert the clinical personnel can occur for each of the different instruments in use. Needless to say that although this provides better assessment of the infant as well as quantitative and

in some cases hard copy data, it also requires that this data be integrated to provide manageable information. By combining data from several different pieces of instrumentation, more specific conditions for alarms can be defined and variables can be more easily compared with one another. This can then lead into trend analysis of the data, and the computer certainly represents an important overall controller, analyzer and recorder of these functions. As technology continues to become more complex, it is important not to lose track of the primary goal of this technology, namely, to provide better neonatal and infant care so that critically ill neonates can look forward to a full, healthy, and normal life.

BIBLIOGRAPHY

Cited References

1. Neuman MR. Biopotential Amplifiers. In: Webster JG, editor. Medical Instrumentation: Application and Design. 3rd ed. New York: John Wiley & Sons, Inc.; 1998. pp. 233–286.
2. Neuman MR. Flexible thin film skin electrodes for use with neonates. Proceedings of the 10th International Conference Medical Biological Engineering. Dresden: DDR; 1973.
3. Accidents with apnea monitors. FDA Drug Bull. Aug. 1985; 15:18.
4. Primiano FP. Measurements of the Respiratory System. In: Webster JG, editor. Medical Instrumentation: Application and Design. 3rd ed. New York: John Wiley & Sons, Inc.; 1998. pp. 372–439.
5. Thomas PE. What's the latest on carbon dioxide monitoring? Neonatal Network 2004 July–Aug; 23(4):70–72.
6. Wu CH, et al. Good estimation of arterial carbon dioxide by end-tidal carbon dioxide monitoring in the neonatal intensive care unit. Pediatr Pulmonol 2003;35(4):292–295.
7. Gordon DH, Thompson WL. A new technique for monitoring spontaneous respiration. Med Instrum 1975;9:21.
8. Kulkarni V, et al. AURA: a new respiratory monitor. Biomed Sci Instrum 1990;26:111–120.
9. Neuman MR. A microelectronic biotelemetry system for monitoring neonatal respiration using thermistors. Proceedings of the 21st Annual Meeting Association Advanced Medical Instrumentation. Chicago: 1986.
10. Neuman MR. Multiple Thin-Film Sensor System. US patent No. 5.394,883. 1995 Mar 7.
11. Werthammer J, Krasner J, DiBenedetto J, Stark AR. Apnea monitoring by acoustic detection of air flow. Pediatrics 1983; 71:53.
12. Whitney RJ. The measurement of volume changes in human limbs. J Physiol (London) 1953;121:1.
13. Rolfe P. A magnetometer respiration monitor for use with premature babies. Biomed Eng 1971;6:402.
14. Sackner JD, et al. Noninvasive measurement of ventilation during exercise using a respiratory inductive plethysmograph. Am Rev Respir Dis 1980;122:867.
15. Cohen KP, et al. Design of an inductive plethysmograph for ventilation measurement. Physiol Meas 1994 May; 15(2): 217–229.
16. Brooks LJ, DiFiore JM, Martin RJ. Assessment of tidal volume over time in preterm infants using respiratory inductance plethysmography, The CHIME Study Group. Collaborative Home Infant Monitoring Evaluation. Pediatr Pulmonol 1997 Jun; 23(6):429–433.
17. Fraden J. Piezo/pyroelectric film as a biomedical transducer. Proc Annu Conf Eng Med Biol 1986;28:221.
18. Barrow RE, Colgan FJ. A noninvasive method for measuring newborn respiration. Respir Care 1973;18:412.
19. Prechtl HFR, van Eykern LA, O'Brien MJ. Respiratory muscle EMG in newborns: A non-intrusive method. Early Hum Dev 1977;1:265.
20. Mendenhall RS, Neuman MR. Efficacy of five noninvasive infant respiration sensors. Proceedings IEEE Fronteuro on Engineering Medical Biology. Columbus, (OH): 1983. p 303–307.
21. Peabody JL, et al. Clinical limitations and advantages of transcutaneous oxygen electrodes. Acta Anaesthesiol Scand Suppl 1978;68:76.
22. Eberhart RC. Indwelling blood compatible chemical sensors. Surg Clin North Am 1985;65:1025.
23. Couch NP, et al. Muscle surface pH as an index of peripheral perfusion in man. Ann Surg 1971;173:173.
24. Huch A, Huch R, Lubbers DW. Transcutaneous PO_2. New York: Thieme-Stratton; 1981.
25. Kimmich HP, Spaan JG, Kreuzer F. Transcutaneous measurement of $PaCO_2$ at 37°C with a triple electrode system. Acta Anesthesiol Scand Suppl 1978;68:28.
26. Tremper KK, Waxman K, Shoemaker WC. Effects of hypoxia and shock on transcutaneous PO_2 values in dogs. Crit Care Med 1979;7:526.
27. Reynolds GJ, Goodwin B, Cowen J, Harris F. Simultaneous transcutaneous measurements of O_2, CO_2, and N_2 in neonates with RDS using a mass spectrometer. In: Rolfe P, editor. Fetal and Neonatal Physiological Measurements. London: Pitman; 1980. p 442.
28. Yoshiya I, Shimada Y, Tanaka K. Spectrophotometric monitoring of arterial oxygen saturation in the fingertip. Med Biol Eng Comput 1980;18:27.
29. Mendelson Y, Lewinsky RM, Wasserman Y. Multi-wavelength reflectance pulse oximetry. Anesth Analg (Suppl.) 2002;94(1): S26–30.
30. Reuss JL, Siker D. The pulse in reflectance pulse oximetry: modeling and experimental studies. J Clin Monit Comput 2004; 18(4):289–299.
31. Workie FA, Rais-Bahrami K, Short BL. Clinical use of new-generation pulse oximeters in the neonatal intensive care unit. Am J Perinato 2005 Oct; 22(7):357–360.
32. Darnall RA. Noninvasive blood pressure measurement in the neonate. Clin Perinatal 1985;12(1):31.
33. Wayenberg JL. Non-invasive measurement of intracranial pressure in neonates and infants: experience with the Rotterdam teletransducer. Acta Neurochir (Suppl.) 1998;71: 70–73.
34. Welch K. The intracranial pressure in infants. J Neurosurg 1980; 52:693.
35. Hill A. Intracranial pressure measurements in the newborn. Clin Perinatal 1985;12(1):161.
36. Lingwood BE, Dunster KR, Colditz PB, Ward LC. Noninvasive measurement of cerebral bioimpedance for detection of cerebral edema in the neonatal piglet. Brain Res 2002; 945(1):97–105.
37. Allan WC, Roveto CA, Sawyer LR, Courtney SE. Sector scan ultrasound imaging through the anterior fontanelle. Am J Dis Child 1980;134:1028.
38. Hintz SR, et al. Bedside imaging of intracranial hemorrhage in the neonate using light: comparison with ultrasound, computed tomography, and magnetic resonance imaging. Pediatr Res 1999;45(1):54–59.
39. Goplerud JM, Delivoria-Papadopoulos M. Nuclear magnetic resonance imaging and spectroscopy following asphyxia. Clin Perinatol 1993;20(2):345–367.
40. Strange M, Cassady G. Neonatal transcutaneous bilirubinometry. Clin Perinatol 1985;12(1):51–62.

41. Phillips BA, Anstead MI, Gottlieb DJ. Monitoring sleep and breathing: methodology. Part I: Monitoring breathing. Clin Chest Med 1998;19(1):203–212.
42. Hoppenbrouwers T. Polysomnography in newborns and young infants: sleep architecture. J Clin Neurophysiol 1992;9(1):32–47.
43. Barbosa GA, Keefe MR, Lobo ML, Henkin R. Adaptation of a cardiac monitor for collection of infant sleep data and development of a computer program to categorize infant sleep state. J Nurs Meas 2003;11(3):241–251.
44. Huch R, Huch A, Rooth G. An Atlas of Oxygen-Cardiorespirograms in Newborn Infants. London: Wolfe Medical Publications, Ltd.; 1983.
45. Neuman MR, Huch R, Huch A. The neonatal oxycardiorespirogram. CRC Crit Rev Biomed Eng 1984;11:77.
46. Neuman MR, Flammer CM, O'Connor E. Safety devices for neonatal intensive care. J Clin Eng 1982;7:51.
47. Neuman MR. Therapeutic and prosthetic devices. In: Webster JG, editor. Medical Instrumentation: Application and Design. 3rd ed. New York: John Wiley & Sons, Inc.; 1998. pp. 577–622.

See also BLOOD GAS MEASUREMENTS; INCUBATORS, INFANT; MONITORING, INTRACRANIAL PRESSURE; MONITORING, UMBILICAL ARTERY AND VEIN; TEMPERATURE MONITORING; VENTILATORY MONITORING.

NERVE CONDUCTION STUDIES. See ELECTRONEUROGRAPHY.

NEUROLOGICAL MONITORS

R. R. GHARIEB
Infinite Biomedical Technologies
Baltimore, Maryland

N. V. THAKOR
Johns Hopkins University
Baltimore, Maryland

INTRODUCTION

The electroencephalogram (EEG) is an electrical activity of the brain that is recorded by using electrodes appropriately placed on the scalp, then amplifying and displaying the electrical signal and its clinical relevant feature using a computer, or other suitable monitors. The EEG signal is a wave that varies in time. This wave contains frequency components that can be measured and analyzed. These frequency components have meaning and valuable properties. Table 1 shows the commonly defined waves or rhythms, their frequency, and their properties. Hans Berger, the discoverer of the EEG in humans, observed in 1924 all of the rhythms known today (except the 40 Hz "gamma" band). The described many of their basic properties. Since then, our definitions and understandings of the rhythms have been refined. However, there still remains some uncertainty, and controversy, in how to define and use these bands, for various purposes. Clinicians view the brainwaves for diagnostic purposes and seek to identify patterns that are associated with specific pathologies or conditions. Psychologists also study them in association with mental states, mental processing, and to test concepts of how the brain processes information (1–6).

The EEG is therefore a noninvasive marker for cortical activity. The EEG in humans and animals is used to monitor alertness; coma and brain death; locate area of damage following head injury, stroke, tumor, and so on; monitor cognitive engagement; control depth of anesthesia; investigate and locate seizure origin; test epilepsy drug effects; monitor human and animal brain development; test drugs for convulsive effects; investigate sleep disorder, monitor and track brain ischemia; and so on. Continuous EEG monitoring is a common routine in the intensive care unit (ICU). However, in digital processed EEG, we study the patterns that emerge during various behavioral, as well as introspective, states, and then see what they are defining in terms of a multidimensional representation of some state space. Research that is focused on understanding specific properties, such as attention, alertness, mental acuity, and so on; has uncovered combinations of rhythms, and other EEG properties, that are relevant to these studies. Generally, derived properties are found, that involve computer processing of the EEG, to produce quantification measurements that are useful for research, monitoring, and so on.

Since high speed computers and sophisticated and efficient digital signal processing methodologies have become available. These properties are significant and new features and properties have been extracted from the EEG signal. These features are combined in a system of multivariable representation to formulate various quantitative EEG (qEEG) measures. The features commonly employed are (7–21).

Amplitude
Subband powers
Spectrogram
Entropy and complexity
Coherence
Biocoherence
Power spectrum
Joint-time frequency
Spectral edge frequencies
Coefficient-based EEG modeling
Bispectrum
Etc.

In the following section, the EEG monitors are classified and the main devices of the monitor are described. This section presents two types of monitors. In the common specification of Optimized Monitor section, the general specifications of the optimized EEG monitor are provided.

CLASSIFICATION OF EEG MONITORS

What is an EEG Monitor?

The neurological monitor is simply a display that shows the ongoing neurological activity recorded as the electrical potential by appropriately placing electrodes on the scalp. The conventional monitor goes back to EEG machine, where the electrical activity of the brain could be detected and plotted on scaled paper. Today, the neurological monitors are based on advanced technologies. They are

Table 1. EEG Rhythms their Frequency Bands and Properties

Rhythm Name	Frequency Band, Hz	Properties
Delta	0.1–3	Distribution: generally broad or diffused, may be bilateral, widespread Subjective feeling states: deep, dreamless sleep, non-REM sleep, trance, unconscious Associated tasks and behaviors: lethargic, not moving, not attentive Physiological correlates: not moving, low level of arousal Effects of Training: can induce drowsiness, trance, deeply relaxed states
Beta	4–7	Distribution: usually regional, may involve many lobes, can be lateralized or diffuse; Subjective feeling states: intuitive, creative, recall, fantasy, imagery, creative, dream-like, switching thoughts, drowsy; oneness, knowing Associated tasks & behaviors: creative, intuitive; but may also be distracted, unfocused Physiological correlates: healing, integration of mind/body Effects of Training: if enhanced, can induce drifting, trance-like state if suppressed, can improve concentration, ability to focus attention
Alpha	8–12	Distribution: regional, usually involves entire lobe; strong occipital w/eyes closed Subjective feeling states: relaxed, not agitated, but not drowsy; tranquil, conscious Associated tasks and behaviors: meditation, no action Physiological correlates: relaxed, healing Effects of Training: can produce relaxation Sub band low alpha: 8–10: inner-awareness of self, mind/body integration, balance Sub band high alpha: 10–12: centering, healing, mind/body connection
Low Beta	12–15	Distribution: localized by side and by lobe (frontal, occipital, etc.) Subjective feeling states: relaxed yet focused, integrated Associated tasks & behaviors: low SMR can reflect "ADD", lack of focused attention Physiological correlates: is inhibited by motion; restraining body may increase SMR Effects of Training: increasing SMR can produce relaxed focus, improved attentive abilities, may remediate Attention Disorders.
Mid-range Beta	15–18	Distribution: localized, over various areas. May be focused on one electrode. Subjective feeling states: thinking, aware of self and surroundings Associated tasks and behaviors: mental activity Physiological correlates: alert, active, but not agitated Effects of Training: can increase mental ability, focus, alertness, IQ
High Beta	15–18	Distribution: localized, may be very focused. Subjective feeling states: alertness, agitation Associated tasks and behaviors: mental activity, for example, math, planning, and so on. Physiological correlates: general activation of mind & body functions. Effects of Training: can induce alertness, but may also produce agitation, etc.
Gamma	40	Distribution: very localized Subjective feeling states: thinking; integrated thought Associated tasks and behaviors: high level information processing,"binding Physiological correlates: associated with information-rich task processing Effects of Training: not known

computer based and display not only the raw EEG, but also various quantitative indexes representing processed EEG. The monitors are EEG processors that have the ability to perform data acquisition, automatic artifact removal, EEG mining and analysis, saving/reading EEG data, and displaying the quantitative EEG (qEEG) measures (indexes) that best describe neurological activity and that are clinically relevant to brain dysfunction.

Neurological Monitor Main Components

As shown in Fig. 1, a typical neurological monitor consists of a few main devices. These devices are connected together through a microcomputer, which supervises and controls the data flow from one device to another. It also receives and executes the user instructions. It implements the EEG methodology routine. The main devices of a typical monitor can be summarized as follows:

ELECTRODES AND ELECTRODE PLACEMENT

Electrodes represent the electrical link between the subject's brain and the monitor. These electrodes are appropriately placed on the scalp for recording the electrical potential changes. Electrodes should not cause distortion to the electrical potential recorded on the scalp and should be made of materials that do not interact chemically with electrolytes on the scalp. The direct current (dc) resistance of each electrode should measure no more than a few ohms. The impedance of each electrode is measured after an electrode has been applied to the recording site to evaluate the contact between the electrode and the scalp. The impedance of each electrode should be measured routinely before every EEG recording and should be between 100 and 5,000 Ω (2).

The international 10–20 system of electrode placement provides for uniform coverage of the entire scalp. It uses

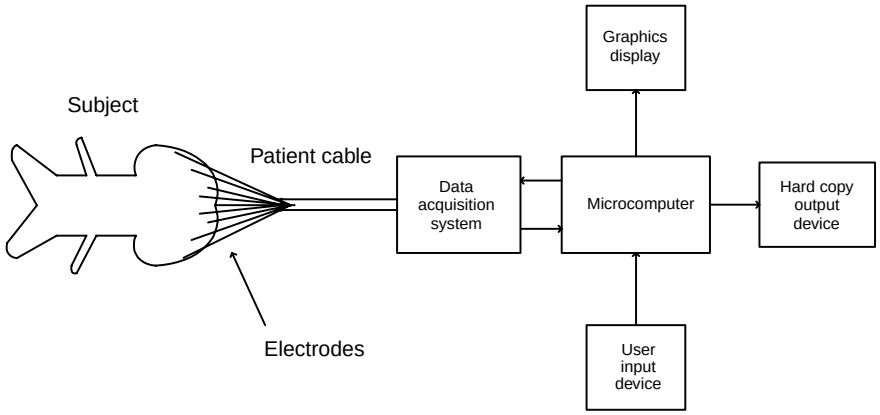


Figure 1. Block diagram of main components of a neurological monitor.

the distances between bony landmarks of the head to generate a system of lines, which run across the head and intersect at intervals of 10 or 20% of their total length. The use of the 10–20 system assures symmetrical, reproducible electrode placement and allows a more accurate comparison of EEG from the same patients, recorded in the same or different laboratories.

Patient Cable

The patient cable assembles the electrode terminals to the recording machine and monitoring instrument. It is preferable that the patient cable be of short length, which assures low impedance and causes no distortion of the electrical potential representing the neurological activity.

Data Acquisition System

It is composed of filters, amplifiers, analog-to-digital converters (ADC), and buffers. Bandpass filters of 0.5–100 Hz band are usually used to enhance the quality of the EEG signal. High gain amplifiers are required since the electrical potentials on the scalp are of microvolt. The input impedance of the amplifiers should be a large value while the output impedance should be a few ohms. The ADC converter digitizes the EEG data by sampling (converts the continuous-time EEG into discrete-time EEG) the data and assign a quantized number for each sample. Figure 2 shows a schematic diagram for the ADC converter while Fig. 3 shows the output–input characteristic of the uniform quantizer. Uniform quantization generates additive white noise to the EEG signal. Portable and wireless units of ADC have been used. The unit is connected to the monitor device through a standard wireless communication routine. This makes the monitor more comfortable and easier to be used.

Microcomputer

The microcomputer represents the master of the EEG monitor. It controls the data flow from one device to another. It reads the EEG data from the ADC buffers. It also hosts the software of the qEEG approaches and the artifact removal programs. Mathematical operations and analysis are carried out in the microcomputer. After processing the EEG data, the microcomputer sends the EEG signal and its qEEG measure (index) to the display. When the microprocessor is instructed to save the EEG session and its qEEG measure, it sends the data to the hard copy device.

Graphics Display

The graphics display displays the contentious EEG signals and online quantitative EEG (qEEG) measure. It helps the neurologists to follow and track in real-time fashion the changes in the brain activity and to monitor the brain development in the intensive care unit (ICU).

Hard Copy Output Device

This device is connected to the microcomputer and stores a version of the EEG data for future use. It could be a hard drive, computer CD, or a printer–plotter for plotting either version of the EEG or the qEEG measure to be investigated by neurologists and to be a part of the patient record.

User Input Device

Through this device, the user can communicate and interact with the monitor. Instructions and various parameters required for the EEG analysis are sent to the microcomputer through this device.

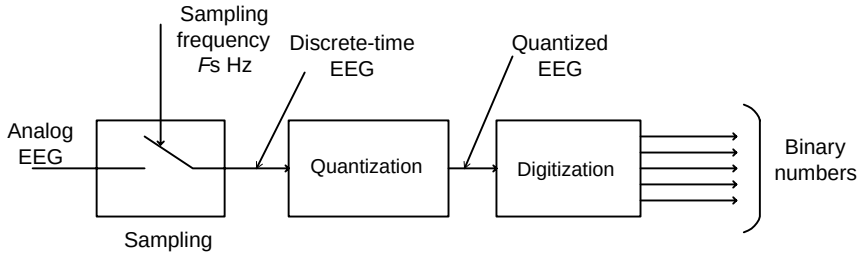


Figure 2. Schematic diagram of the ADC.

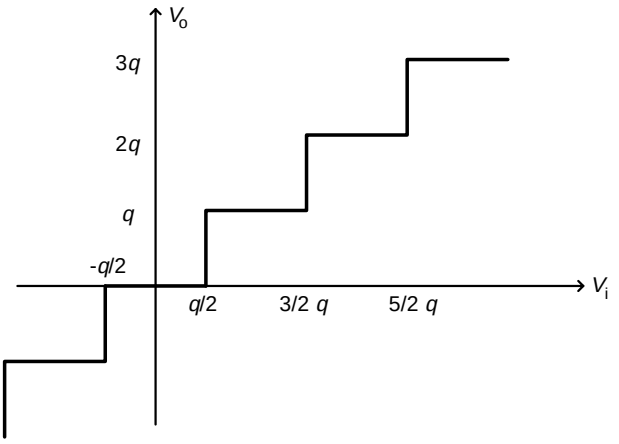


Figure 3. Input–output characteristic of the uniform quantization system.

TYPES OF EEG MONITORS

The EEG monitors can be classified into two main categories based on either their applications or the quantitative EEG index employed for processing and assessment of the brain electrical activity. Accordingly, the most popular monitors can be categorized as follows (7,22):

Application-Based Monitors

Cerebral function monitor (CFM)

Cortical injury monitor (CIM)

Anesthesia monitor

Narcotrend monitor

Patient state analyzer (PSA) with frontal patient state index (PSI)

BrainMaster 2E monitor portable bedside monitor

EEG Index-Based Monitors

Amplitude integrated monitor

Spectral index monitor

Spectral-edge frequency monitor

Bispectral index monitor

Entropy and complexity based monitor

This section presents, a very brief description of both monitor types. This description gives the intuition for the neurological applications of the monitor and the EEG index employed.

Cortical Injury Monitor

The lack of blood and oxygen flow to the brain due to cardiac arrest causes brain ischemia, causing brain cells to die, and consequently affecting (changes) brain activity. It has been demonstrated by many studies that brain ischemia slows the brain electrical activity by suppressing the high frequency and enhances the background activity; the cortical injury monitor has been developed and used for the detection and tracking of brain ischemia. The advantage of the monitor comes from the fact that it provides a quantitative measure extracted from the processed EEG signal for the severity of brain injury after cardiac arrest. It aids neurologists in providing better care for patients with cardiac arrest and provides them with therapeutic intervention, such as hypothermia. The monitor provides assessment of the brain function within the first 4 h after cardiac arrest.

Anesthesia Monitors

Patients receive general anesthesia during surgery. Anesthesia causes reduction of brain activity and concussions. The depth of anesthesia should be evaluated and tracked in real-time fashion to prevent perfect suppression of brain activity. The anesthesia monitor has been developed and used for the assessment of anesthesia and concussions. It provides a quantification measure or index for the depth of anesthesia. The monitor helps patients "rest easy" when they receive general anesthesia for surgery. Of the known anesthesia monitors, the bispectral (BIS) monitor, the narcotrend monitor, and the patient state analyzer (PSA4000) monitor are commonly employed. In the BIS monitor, a qEEG measure based on bispectrum is employed for tracking the depth of anesthesia. The PSA4000 is indicated for use in the operating room (OR), ICU, and clinical research laboratories. The monitor includes the patient state index (PSI), a proprietary computed EEG variable that is related to the effect of anesthetic agent. The narcotrend monitor provides a 6-letter classification from A (awake) to F (general anesthesia with increasing burst suppression). The narcotrend EEG monitor is similar to the BIS monitor positioned on the patient's forehead. The EEG classification made by the narcotrend monitor are 6 letters: A (awake), B (sedate), C (light anesthesia), D (general anesthesia), E (general anesthesia with deep hypnosis), F (general anesthesia with increasing burst suppression) (23).

Cerebral Function Monitor

The cerebral function monitor (CFM) enables continuous monitoring of the cerebral electrical activity over long periods of time due to slow recording speeds. The cerebral electrical signals picked up by the electrodes attached to the scalp are registered in the form of a curve, which fluctuates to a greater or lesser extent depending on the recording speed. Examination of the height of the curve with respect to zero and its amplitude indicates the voltage of cerebral activity and yields information regarding polymorphism. Thus it is possible to monitor variations in cerebral activity over a prolonged period during anesthesia as well as during the revival phase with the monitor of cerebral function. The CFM is common practice in monitoring the cerebral function in intensive care. To bring the CFM into a polygraphy environment the hardware processing and paper write-out have to be implemented in software. The processor comprises a signal shaping filter, a semilogarithmic rectifier, a peak detector, and low pass filter. After taking the absolute value of the filtered EEG signal, the diode characteristic used to compress the signal into a semilogarithmic value was mimicked by adding a small offset to the absolute value before taking the

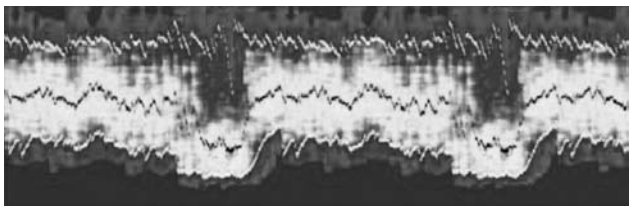


Figure 4. Color CFM of ~ 2.5 h, red is high density, blue low density, and black zero. Vertical scale from 0–5 μV linear, from 5–100 μV logarithmic. The median is given in black and the percentiles in white. As only 1 h of data was available to test the reproducibility of the process, we used a repeated playback mode for this picture. The low median episodes are neonatal State 1 (Quiet Sleep) with the beginning tracé-alternant (high peaks followed by low amplitude EEG) with half way diminishing peak heights, and a neonatal State 2 (REM Sleep) with symmetrical continuous EEG.

logarithm. The envelope of the resulting signal has been made by means of a leaky peak detector and a boxcar averager. Writing the resulting signal on a pixelized computer screen at a speed of 6 cm/h, say 200 pixels per hour gives 18 s per pixel. At a sample rate of 200 Hz, 3600 samples will be written to the same pixel column. Only a line connecting the highest and lowest value of the 18 s period will be seen. All information about local density of the signal between the high and low values will be lost. Therefore there is an amplitude histogram per pixel column and a color plot of this histogram is built. To give even more information, the median and the fifth and ninety-fifth percentile as bottom and peak estimates are shown. The CFM is shown to be useful for seizure detection, neonatal, care in the emergency room, and for the assessment of other brain disorders (18,19,21). The CFM trace may require a specialist for its interpretation. An EEG atlas provides a summary for the interpretation of the EEG based trace. Figure 4 shows an example for neonatal EEG monitoring. Studies have shown that when CFM is used in combination with a standard neurological examination, it enhances the clinician's ability to identify the presence of seizures or to monitor infants EEG and others.

Amplitude-Integrated EEG (aEEG) Monitor

Various brain activities may causes changes in normal EEGs. These changes might be in the amplitude, power, frequency, BIS, entropy or complexity. In fact, since EEG has become available, visual investigation of EEG has been used to asses the neurological function. It is evident that continuous EEG is a sensitive, but nonspecific measure of brain function and its use in cerebrovascular disease is limited. Visual interpretation of EEG is not an easy target and need well-trained expertise, which is not available all the times in the ICU. Besides, information that can be extracted by visual investigation is limited. The EEG amplitude shown in Fig. 5 by the aEEG monitor is the first feature, which has received the attention of neurologists and researchers. It is obvious that there is no clear difference between the aEEG associated with the normal and ischemic injury EEGs. The cerebral function monitor (CFM) uses the aEEG extracted from one channel. The aEEG can show bursts and suppression of the EEG. The

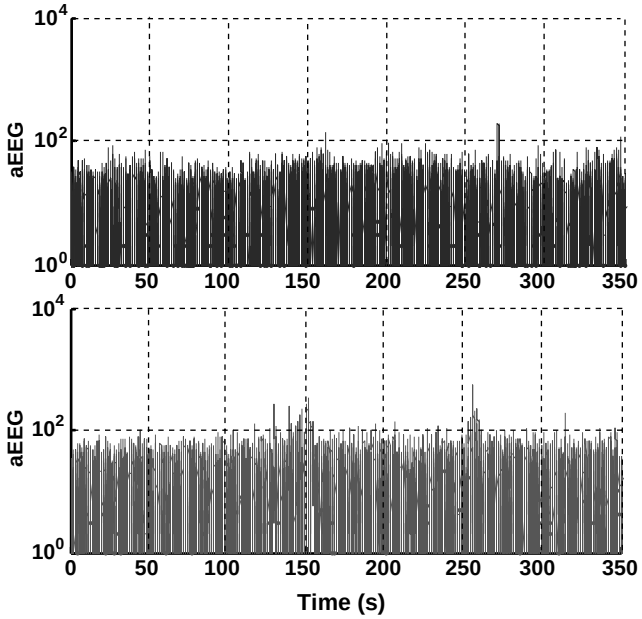


Figure 5. Amplitude-integrated EEG for normal EEG (top) and injury-related EEG of grade CPC 5. It is obvious that no significant differences between the two cases, which implies that amplitude may not be used for injury identification.

CFM is well used for seizure detection and neonatal monitoring (18,19,21). However, the EEG amplitude shows low capability. To clarify this disadvantage, let us ask the question: Does isolectrical EEG mean brain death or even coma? There has been a study seeking the answer of this question (24). In this study, from 15 patients with clinical diagnosis as brain death, EEG was isoelectricity in eight patients while the remaining seven showed persistence of electrical activity. Comatose patients may also show the presence of electrical activity in the alpha band (8–13 Hz). Such diagnosis is referred to as alpha coma. This implies that both investigation and monitoring of EEG amplitude may not be a reliable confirmatory test of brain function and coma. The amplitude assessment of the EEG may then mislead the neurologist's decision.

Spectrogram-Based Monitor

A number of studies have focussed their attention to the prognostication of frequency contents and the power spectrum of EEG (6,20,25,26). The normal EEG of adults often show three spectral peaks in delta, theta, and alpha, as demonstrated in Fig. 5 (top). The most common observation, in ischemic injury, for example, has been slowing background frequencies by increasing the power of delta rhythm and decreasing the powers of theta and alpha rhythms. Numerous approaches have been employed the frequency contents for developing a diagnostic tool or index. Monitoring the real-time spectrum has also been employed. While this approach gives an indication for ischemic injury, it requires a well-trained specialist. In animal studies, the spectral distance between a baseline (i.e., normal) EEG and the underlying injury-related one was employed as a metric for injury evaluation and

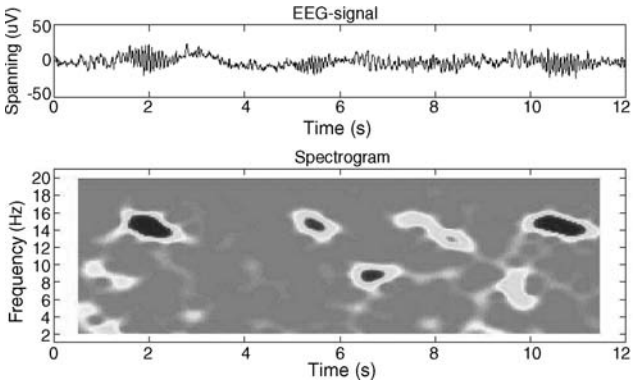


Figure 6. The EEG fragment with isolated 15 Hz spindles, which are clearly visible in the spectrogram. The spectrogram also shows that the spindles are alternated by short periods of 9 Hz activity.

monitoring. However, the spectral distance has the disadvantage of using the whole frequency contents. This is because using the whole frequency contents of the EEG signal increases the likelihood of artifacts-corrupted spectral contents. Time–frequency analysis is a signal analysis technique that provides an image of the frequency contents of a signal as a function of time. Several methods (or time–frequency distributions) can be used, one of which is the spectrogram. The spectrogram is the power spectrum of the investigated signal seen through a time window that slides along the time axis. Figure 6 shows a segment of sleep EEG signal (top) and its spectrogram. It is obvious that the spectrogram shows a sleep spindle at 15 Hz. The spectrogram shows the times where the spindle is activated. The time–frequency analysis can then be a helpful tool to facilitate the EEG interpretation, as is shown in the examples below.

Normalized Separation-Based Monitor

As mentioned, ischemic injury manifests itself in the EEG by slowing the background activity and reducing the high frequency. Such injury-related changes can be used for the separation of normal EEG from injury-related one. Based on this frequency information, a normalized separation was adopted as an qEEG measure. The normalized separation is a spectral-based qEEG measure for assessment of severity of brain injury. It uses the most relevant spectral information related to the normal EEG signal. The normal EEG has a power spectral density showing three fundamental spectral peaks as shown in Fig. 7 (top). It has been demonstrated that employing these three peaks is enough to yield a satisfactory quantitative measure. Moreover, looking selectively at the principal features of the EEG spectrum reduces the sensitivity of the measure to noise and artifacts. This is primarily because a full spectrum-based measure is likely to be susceptible to spectral components related to noise and artifacts. Therefore, the normalized separation employs the principal features of the spectrum and ignores the minor features, which are more sensitive to noise and artifacts. In comparison with amplitude-based measures, such as the aEEG, the aEEG is not a quantitative measure and represents a continuous EEG. This finding implies that well-trained specialist are needed for the interpretation of the aEEG trace. The

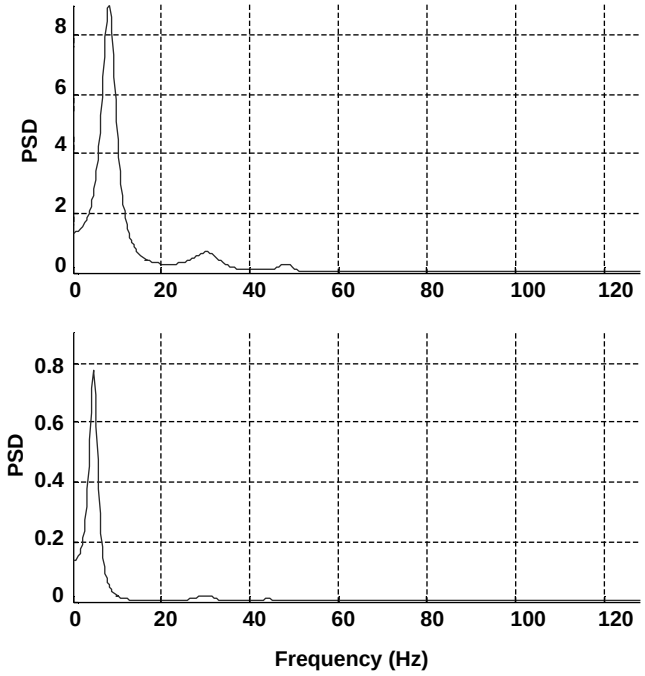


Figure 7. Power spectral density of EEG computed using the AR method applied to 4 s window and averaged >10 windows. (Top) Normal EEG signal and (bottom) abnormal. It is obvious that with abnormal EEG the background frequency gets slower and the high frequencies diminish.

amplitude is also susceptible to noise and artifacts that mislead the interpretation. In comparison with the higher order spectra-based measures, the normalized separation is enough since most information and features of the EEG are described by the power spectrum. A recent study and clinical investigation supports this claim. The EEG is commonly modeled as a stochastic process and for this reason phase is not important. The phase is the only feature retained in higher order spectra. Figure 8 shows three EEG signals and their corresponding normalized separations. The first EEG is very close to normal and provides a normalized separation of 0.2. In the second EEG spectrum, the third peak is diminished and the normalized

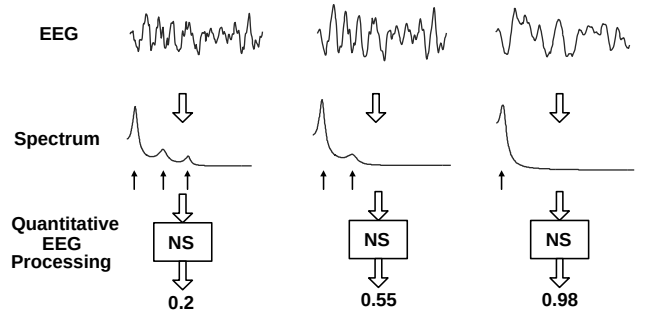


Figure 8. Normalized separation for three EEG cases. The left one is normal EEG where three spectral peaks are shown, the middle one is mildly injury-related, and the right one is a severely injury-related EEG. It is obvious that the spectral-based normalization makes significant separation between these three categories.

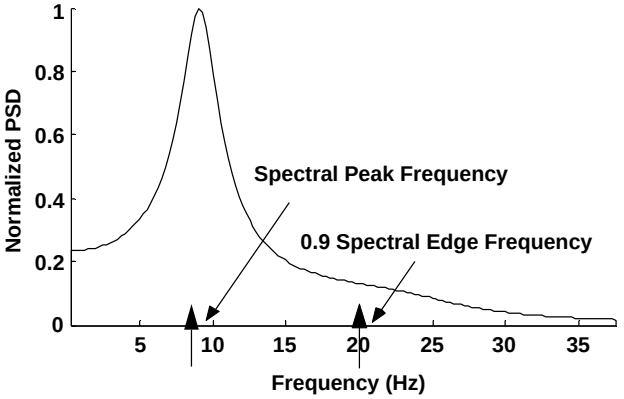


Figure 9. Normalized power spectral density (PSD) showing the spectral peak frequency (SPF) and the 90% spectral edge frequency.

separation is 0.55. In the third EEG spectrum, both second and third spectral peaks are diminished causing the EEG to be separated from the normal EEG by 0.98. The normalized separation ranges from 0 to 1, where the zero value represents and quantifies the normal EEG, while the one value corresponding to the sever abnormal EEG.

Spectral Edge Frequency Monitor

Computers make computation and search for spectral edge frequency of the EEG signal applicable for assessing and monitoring the cortical activity and brain dysfunction. The median frequency and the frequency edges providing 90–95% of the power have been reported to be useful (27,28). The spectral mean and peak frequencies have also been employed (29,30). The success of computing the time-varying spectral edge frequencies depends on the best estimation of the time-varying power spectrum. The fast-Fourier transform (FFT) is a commonly used approach for computing the real-time power spectrum. However, the FFT-based power spectrum provides poor frequency resolution since the resolution is proportional to the reciprocal of the analysis window. The model-based power spectrum estimate, such as the time-varying autoregressive, provides high resolution and low variance estimate of the power spectrum (Fig. 9).

Bispectral Index Monitor

In addition to spectrum, BIS also describes the frequency contents of the EEG. The power spectrum is often used for describing the frequency contents of the EEG modeled as a sum of noncoupled harmonics (17,31–33). In such situations, BIS are identically zero. However, if the focus is on the frequency contents of coupled harmonic (quadratic phase coupling harmonics), BIS is often used. Bispectrum is one of the first successful applications of electroencephalography, which measures the effects of anesthetics on the brain. The BIS index is a number between 0 and 100. It produces a number between 0 and 100 (100 represents the fully awake state, and zero no cortical activity). The BIS correlates with depth of sedation and anesthesia, and can predict the likelihood of response to commands and recall. The BIS values correlate with end-tidal volatile agent concentrations, and with blood and effect-site propofol concentrations. It is not very good at predicting movement in response to painful stimuli. However, there has been a recent study, which shows that BIS information is not necessary and power spectrum is satisfactory to describe an EEG signal. Another fact is that BIS is very sensitive to spike artifacts. The BIS index is a quantitative EEG index developed and employed for measuring the depth of anesthesia. It is based on third-order statistics of the EEG signal, specifically BIS density, and is commercially used for monitoring anesthetic patients. The index quantitatively measures the time-varying BIS changes in the EEG signal acquired from the subject before and during anesthesia. The BIS index will be zero when both the baseline and the underlying signal are either identical or Gaussians. This measure has been demonstrated to be effective for depth of anesthesia measurements. However, in some applications (classification of brain injury due to hypoxic/asphyxic cardiac arrest) the principal information and features of the EEG signal lie in second-order statistics, that is the power spectrum, and only minor information and features are associated with higher order statistics. Therefore, indexes based on higher order statistics may not be best suited to classify brain injury due to hypoxia/asphyxia (33). Besides, that higher order statistics are very sensitive to sparse-like artifacts, which deteriorates the index (34). Therefore, employing higher order statistics-based indexes require an efficient artifact removal approach for preprocessing the EEG signal. Below, a simulated example of BIS is presented. In this example, the BIS density is shown to present information on the coupling harmonics. Let $x(n)$ be a time-series consisting of three sinusoidal components whose frequencies are 64, 128, and 192 Hz. It is obvious that harmonic coupling between the first two sinusoids exists. Fig. 10a

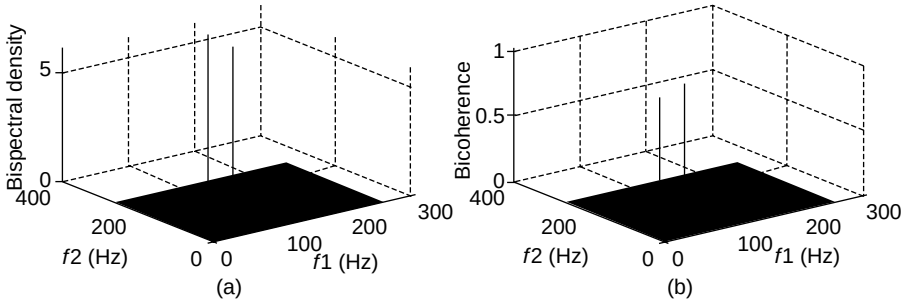


Figure 10. Bispectral density (a) and bicoherence (b) of a simulated sinusoidal signal. Bispectral density shows two lines at the coupling frequencies $(f_1, f_2) = (64, 64)$ and $(f_1, f_2) = (64, 128)$. Bicoherence shows two lines of unity value at the coupling frequencies.

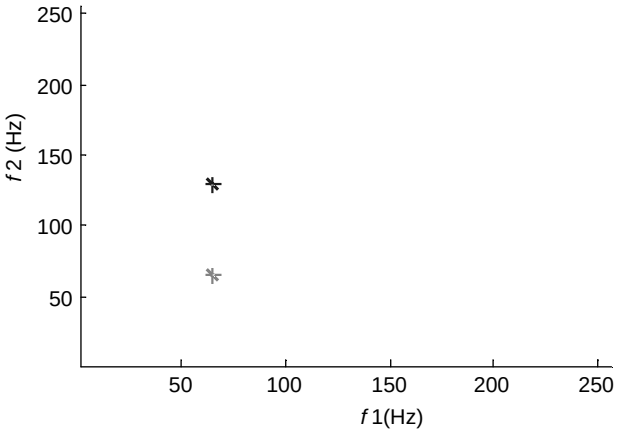


Figure 11. Contour plot of the bicoherence of x(n).

and b shows the mesh plots of the BIS and the two-dimensional bicoherence. Figure 11 shows the contour plot of the bicoherence. It is obvious that the BIS density shows two spectral lines at the frequencies (f_1, f_2) = (64, 64) and (f_1, f_2) = (64,128). The bicoherence is unity (perfect coupling) at these frequencies.

Entropy- and Complexity-Based Monitor

Since the brain processes information, the brain's total electrical activity probably corresponds to information processing in the brain. This assumption was used to study the entropy or self-information in the EEGs of anesthetic patients, postcardiac arrest, in sleep research, and seizure. Entropy as a measure quantifies the disorder of the EEG signal. It represents the complexity and nonlinearity inherent in the EEG signal. It has been shown that normal control subjects provide larger entropy values than those showing ischemic injury postcardiac arrest. The entropy starts to increase with the recovery of brain function. That is, entropy is a relevant indication of the brain order–disorder following cardiac arrest. The subject under anesthesia provides low entropy, while the awake subject shows high entropy since their brain is full of thinking and activity. Numerous approaches for the calculation of entropy have been used, such as Shannon entropy, approximate entropy, Tasllis entropy, and wavelet entropy.

Complexity based on chaotic, state space and correlation dimension has also been employed for assessment and monitoring of brain function (11–13,22,35–41).

COMMON SPECIFICATIONS OF OPTIMIZED MONITOR

The EEG monitor specifications are the hardware and software properties that make the monitor capable of easily and significantly performing assessment and classification of cortical activity. The monitor should satisfy minimum requirements. Common specifications of EEG monitors may include the following: compact design that is rugged and lightweight; automatic classification of EEG; off- and on-line qEEG index; optimized recognition and removal of artifacts; easy operation via friendly touch screen; continues testing of the electrodes to ensure a constant high quality of the EEG signal; variable electrode position; interface to external monitors and documentation systems; wireless communication between various sensors attached to the human; provides a secure way to transmit and store measured data; high-speed data processing; large amount of memory; on-board Ethernet connection.

CONCLUSION

This article presented a descriptive review for commonly known and employed neurological monitors. The typical neurological monitor consists of a few main devices and the software for running these devices. A brief review of the device specifications and their roles have been given. The monitors are classified into two main categories based on their applications and the indexes acquired from the digital EEG signal and employed for monitoring and assessment of cortical dysfunctions. Intuitions of the EEG monitors, with no mathematical details, have been presented. The article concludes by describing the most common specifications for the optimized monitor.

BIBLIOGRAPHY

1. Schneider G. EEG and AEP monitoring during surgery The 9th ESA Annual Meeting, Gothenburg, Swede, April 7–10, 2001.
2. Fisch BJ. EEG Primer- Basic princibles of digital and analog EEG. Fisch & Spehlmann's, Third revised and enlarged edition. New York: Elsevier Science BV; 1999.
3. Collura TF. The Measurement, Interpretation, and Use of EEG Frequency Bands. Report Dec. 7, 1997.
4. Berger H. Uber das elecktroenkephalogram des menchen. Arch Psychiatr Nervenkr 1929;87:527–570.
5. Teplan M. Fundamentals of EEG measurement. Meas Sci Rev 2002;2.
6. Gharieb RR, Cichocki A. Segmentation and tracking of EEG signal using an adaptive recursive bandpass filter. Int Fed Med Biol Eng Comput Jan. 2001;39:237–248.
7. Kong X, et al. Qauntification of injury-related EEG signal-changes using distance measure. IEEE Trans Biomed Eng July 1999;46:899–901.
8. Wendling F, Shamsollahi MB, Badier JM, Bellanger JJ. Time-frequency matching of warped depth-EEG seizure observations. IEEE Trans Biomed Eng May 1999;46:601–605.
9. Mingui Sun, et al. Localizing functional activity in the brain through time-frequency analysis and synthesis of the EEG. Proc IEEE Sept. 1996;84:1302–1311.
10. Ning T, Bronzino JD. Bispectral analysis of the rate EEG during various vigilance states. IEEE Trans Biomed Eng April 1989;36:497–499.
11. Hernero R, et al. Estimating complexity from EEG background activity of epileptic patients-Calculating correlation dimensions of chaotic dynamic attractor to compare EEGs of normal and epileptic subjects. IEEE Eng Med Biol Nov./Dec. 1999; 73–79.
12. Roberts SJ, Penny W, Rezek I. Temporal and spatial complexity measures for electroencephalogram based brain-computer interface. Med Biol Eng Comput 1999;37:93–98.
13. Zhang XS, Roy RJ. Predicting movement during anesthesia by complexity analysis of electroencephalograms. Med Biol Eng Comput 1999;37:327–334.

14. Anderson CW, Stolz EA, Shamsunder S. Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. IEEE Trans Biomed Eng March 1998;45:277–286.
15. Hazarika N, et al. Classification of EEG signals using wavelet transform. Signal Process 1997;59:61–72.
16. Quiroga RQ, et al. Searching for hidden information with Gabor transform in generalized tonic-clonic seizures. Electroenceph Clin Neurophysiol 1997;103:434–439.
17. Gajraj RJ, et al. Analysis of the EEG bispectrum, auditory potentials and the EEG power spectrum during related transitions from consciousness to unconsciousness. Br J Anesthes 1998;80:46–52.
18. Toer MC, et al. Amplitude integrated EEG 3 and 6 hours after birth in full term neonates with hypoxic-ischemic encephalopathy. Rch Dis Child Fetal Neonatal Ed 1999;81:19–23.
19. Toet MC, et al. Comparison between simultaneously recoded amplitude integrated EEG (Cerebral function monitor) and standard EEG in neonates. Pediatrics 2002;109:772–779.
20. Hassanpour H, et al. Time-frequency based newborn EEG seizure detection using low and high frequency signatures. Physiol Meas 2004;25:935–944.
21. Nageeb N, et al. Assessment of neonatal encephalopathy by amplitude-integrated EEG. Pediatrics June 1999;103:1263–1266.
22. Bezerianos A, Tong S, Thakor N. Time-dependent entropy estimation of EEG rhythm changes following brain ischemia. Ann Biomed Eng 2003;31:1–12.
23. Kreuer S, et al. The narcotrend- a new EEG monitor designed to measure the depth of anesthesia. Anethesit 2001;50:921–925.
24. Paolin A, et al. Reliability in diagnosis of brain death. Intensive Care Med Aug. 1995;21:657–662.
25. Jung TP, et al. Estimating alertness from the EEG power spectrum. IEEE Trans Biomed Eng Jan. 1997;44:60–69.
26. Celka P, Colditz P. A computer-aided detection of EEG seizures in infants: A singular spectrum approach and performance comparison. IEEE Trans Biomed Eng May 2002;49:455–462.
27. McDonald T, et al. Median EEG frequency is more sensitive to increase in sympathetic activity than bispectral index. J Neurosurg Anesthesiol Oct. 1999;11:255–264.
28. Inder TE, et al. Lowered EEG spectral edge frequency predicts presence of cerebral white matter injury in premature infants. Pediatrics Jan. 2005;111:27–33.
29. Rampil IJ, Matteo RS. Changes in EEG spectral edge frequency correlated with the hemodynamic response to laryngoscopy and intubation. Anesthesiol 1987;67:139–142.
30. Rampil IJ, Matteo RS. A primer for EEG signal Processing in anesthesia. Anesthesiology 1998;89:980–1002.
31. Akgul T, et al. Characterization of sleep spindles using higher order statistics and spectra. IEEE Trans Biomed Eng Aug. 2000;47:997–1000.
32. Michael T. EEGs, EEG processing and the bispectral index. Anesthesiology 1998;89:815–817.
33. Miller A, et al. Does bispectral analysis of the electroencephalogram add anything bur complexity? B J Anesthesia 2004;92:8–13.
34. Myles PS, et al. Artifact in bispectral index in a patient with severe ischemic brain injury. Case Report Int Anesth Res Assoc 2004;98:706–707.
35. Radhakrishnan N, Gangadhar BN. Estimating regularity in epileptic seizure time-series data- A complexity measure approach. IEEE Eng Med Biol May/June 1998; 98–94.
36. Lemple A, Ziv J. On the complexity of finit sequences. IEEE Trans Inf Theory Jan 1976;22:75–81.
37. Tong S, et al. Parameterized entropy analysis of EEG following hypoxic-ischemic brain injury. Phys Lett A 2003;314:354–361.
38. Lerner DE. Monitoring changing dynamics with correlation integrals: Case study of an epileptic seizure source.
39. Xu-Sheng, et al. EEG complexity as a measure of depth of anesthesia for patients. Yearbook of Medical Informatics. 2003; 491–500.
40. Bhattacharya J. Complexity analysis of spontaneous EEG. Acta Neurobiol Exp 2000;60:495–501.
41. Quiroga RQ, et al. Kullback-Leibler and renormalized entropies application to electroencephalogram of epilepsy patients. Phys Rev E Dec. 2000;62:8380–8386.
42. Mizrahi EM, Kellaway P. Characterization and classification of neonatal seizures. Neurology Dec. 1987;37:1837–1844.

Reading List

Blanco S, et al. Time-frequency analysis of electroencephalogram series. *Phys Rev* 1995;51:2624–2631.

Blanco S, et al. Time-frequency analysis of electroencephalogram series. III wavelet packets and information cost function. *Phys Rev* 1998;57:932–940.

Caton R. The electric currents of the brain. *BMJ* 1875;2–278.

D'attellis CE, et al. Detection of epileptic events in electroencephalograms using wavelet analysis. *Annals of Biomed Eng* 1997;25:286–293.

Franaszczuk PJ, Blinowska KJ, Kowalczyk M. The application of parameteric multichannel spectral estimates in the study of electrical brain activity. *Biol Cybern* 1985;51:239–247.

Gabor AJ, Leach RR, Dowla FU. Automated seizure detection using self-organizing neural network. *Electroenceph Clin Neurophysiol* 1996;99:257–266.

Gath I, et al. On the tracking of rapid dynamic changes in seizure EEG. *IEEE Trans Biomed Eng* Sept. 1992;39:952–958.

Geocadin RG, et al. A novel quantitative EEG injury measure of global cerebral ischemia. *Clin Neurophysiol* 2000;11:1779–1787.

Geocadin RG, et al. Neurological recovery by EEG bursting after resuscitation from cardiac arrest in rates. *Resucitation* 2002;55:193–200.

Gotman J, et al. Evaluation of an automatic seizure detection method for the newborn EEG Electroenceph Clin. *Neurophysiol* 1997;103:363–369.

Hernandez JL, et al. EEG predictability:adequacy of non-linear forcasting methods. *Int J Bio-Medical Comput* 1995;38:197–206.

Holzmann CA, et al. Expert-system classification of sleep/awake states in infants. *Med Biol Eng Comput* 1999;37:466–476.

Liberati D, et al. Total and Partial coherence analysis of spontaneous and evoked EEG by means of multi-variable autoregressive processing. *Med Biol Eng Comput* 1997;35:124–130.

Pardey J, Roberts S, Tarassenko LT. A review of parametric modeling techniques for EEG analysis. *Med Eng Phys* 1996;18:2–11.

Petrosian A, et al. Recurrent neural network based prediction of epileptic seizures in intra- and extracranial EEG. *Neurocomput* 2000;30:201–218.

Popivanov D, Mineva A, Dushanova J. Tracking EEG dynamics during mental tasks-A combined linear/nonlinear approach, IEEE Eng. *Med Biol* 1998; 89–95.

Quiroga RQ, et al. Performance of different synchronization measures in real data: A case study on electroencephalographic signals. *Phys Rev E* 2002;65:1–14

Sadasivan PK, Dutt DN. SVD based technique for noise reduction in electroencephalogram signals. *Signal Process* 1996;55:179–189.

Salant Y, Gath I, Hebriksen O. Prediction of epileptic seizures from two-channel EEG, Med Biol. *Eng Comput* 1998;36:549–556.

Schraag S, et al. Clinical utility of EEG parameters to predict loss of consciousness and response to skin incision during total intervention anesthesia. *Anesthesia* April 1998;53:320–325.

Selvan S, Srinivasan R. Removal of ocular artifacts from EEG using and efficient neural network based adaptive filtering technique. *IEEE Signal Process Lett* Dec. 1999;6:330–332.
Vigario RN. Extraction of ocular artifacts from EEG using independent component analysis. *Electroenceph Clin Neurophysiol* 1997;103:395–404.
Zapata A, et al. Detecting the onset of epileptic seizures. *IEEE Eng Med Biol* May/June 1999; 78–83.
Zygierewicz J, et al. High resolution study of sleep spindles. *Clinic Neurphysiol* 1999;110:2136–2147.

Publisher's note. A revised version of this article was submitted after our print production deadline. The revised version will appear in the online edition of this encyclopedia.

See also ELECTROENCEPHALOGRAPHY; EVOKED POTENTIALS; MONITORING IN ANESTHESIA; MONITORING, INTRACRANIAL PRESSURE.

NEUROMUSCULAR STIMULATION. See FUNCTIONAL ELECTRICAL STIMULATION.

NEUTRON ACTIVATION ANALYSIS

XIAOLIN HOU
Risø National Laboratory
Roskilde, Denmark

INTRODUCTION

Neutron activation analysis (NAA) as an elemental analytical method has been used for a long time to determine trace elements in biomedical research and clinical analysis (1,2). The main problems associated with the determination of trace elements in biomedical research and clinical analysis are the very low concentrations of some elements, and the limited amount of sample materials available. It is necessary that the analytical method be sensitive and free of blank contribution.

As a nuclear analytical technique, NAA is based on the excitation of the atomic nucleus of an isotope of the element with neutrons, and the emission of specific radiation, such as gamma rays, from the excited nucleus by decay. Therefore, only trace elements present in the sample during neutron activation will be excited and are able to be measured in this way. The possible contamination of sample during subsequent handling will not influence the result. Thus a comprehensive radiochemical separation of a particular element from interfering elements can be carried out to significantly improve the detection limit, which will not introduce any blank for the measured element. If no pretreatment of sample is completed, almost no blank value will be introduced in the analytical procedure. Therefore, the method can be free of blank contribution. Neutron activation analysis is very sensitive for many trace elements, while many matrix elements such as carbon, hydrogen, oxygen, and nitrogen are less activated by neutrons and produce almost no activity after neutron activation. Therefore, NAA is suitable for the analysis of biomedical samples due to little or no interference from matrix elements. The interference from activated minor elements such as chlorine, sodium, bromine, and potassium may be eliminated by a few days decay of sample due to the short lifetimes of radioactive nuclides of these elements. It is normally used for *in vitro* analysis, but also can be used for *in vivo* analysis of the whole or part of living bodies. This consists of mostly major and minor elements, such as Ca, Cl, K, N, Na, and P, but also some trace elements, such as iodine in thyroid, Si, Cd, Hg in lung and kidney. The contrast with conventional NAA, isotopic and accelerator neutron sources and small reactor are used and prompt gamma rays are measured (3). However, the problem of radiation to living body limits its application. This article will not describe this problem in detail and is an updated version of the previous article published in the first edition of this encyclopedia (4).

THEORY

Neutron activation is a reaction of the nucleus of an element with neutrons to produce a radioactive species, so-called radionuclide. Neutrons used in NAA can be produced by a nuclear reactor, a radioisotope, and an accelerator, in which the nuclear reactor is the most common neutron source used for NAA due to its high neutron flux and suitable neutron energy. Neutrons can exhibit a wide range of energy, which range from thermal neutrons with an average energy of 0.025 eV in a thermal nuclear reactor to fast neutrons of 14 MeV in an accelerator neutron generator. By bombarding an element with neutrons, a neutron is absorbed by the target nucleus to produce a highly energetic state of the resulting nucleus containing an additional neutron. Some excess energy is immediately lost, usually by emission of a gamma ray, a proton, or an alpha particle. In a sample exposed to neutrons, the type of nuclear reactions depends on the energy of the neutrons and on the elements present. The main reaction occurring with thermal neutrons is the (n, γ) reaction. In this case, the highly energetic level of the produced nucleus is de-excited by emission of a gamma ray, while (n, p) and (n,α) reactions are induced by fast neutrons reaction. In conventional NAA, the element is determined by measurement of the radionuclides formed by de-excitation of the produced nucleus. However, the element can be determined also by the measurement of the gamma rays emitted during the de-excitation of the produced nucleus, which is so-called prompt gamma activation analysis due to the very fast de-excitation process (1 ps). In this article, only conventional NAA is discussed. The probability of a particular reaction is expressed by the activation cross-section, σ, with a unit of barn (10^{-28} m^2). An isotope of an element has its specific activation cross-section. The cross-section of a particular nucleus depends on the energy of the neutron. The (n, γ) reaction normally has a higher activation cross-section than (n, p) and (n, α) reactions. The rate of formation of the radionuclide in the neutron activation is expressed as

$$\frac{dN^*}{dt} = \sigma\phi N \qquad (1)$$

N is the number of the target nuclei in the atom, so $N = (m/M)N_A\,\theta$. Here, m is the mass of the target element (in g); M is the atomic mass; N_A is Avogadro's number; and θ is the isotope abundance of the target nuclide, N^* is the number of the activated nuclide at time t; ϕ is the neutron flux density (in neutron $m^{-2}{\cdot}s^{-1}$), which is used to express the neutron number pass through a unit area in a unit time, which can be considered also as a product of velocity of a neutron and its concentration.

If the nuclide formed is radioactive, it will decay with time and the decay rate of the radionuclide will be

$$\frac{dN^*}{dt} = -\lambda N^* \tag{2}$$

Here λ is the decay constant of activated nuclide, $\lambda = \ln 2/T_{1/2}$ and $T_{1/2}$ is the half-life of activated nuclide. So, the production rate of an activated nuclide is expressed as:

$$\frac{dN^*}{dt} = (\sigma\phi N) - (\lambda N^*) \tag{3}$$

The activity or disintegration rate (A_0) at the end of irradiation time t_i is then:

$$A_0 = \lambda N^* = \sigma\phi N(1 - e^{-\lambda ti}) \tag{4}$$

The saturation activity of the activated nuclides (Am), that is, the activity when the production of activity is equal to the decay of the activity, can be calculated as:

$$\mathrm{Am} = \sigma\phi N = \frac{m}{M} N_A \theta\sigma\phi \tag{5}$$

If considering the decay of activated radionuclides during the decay (t_d) and counting (t_c), the measured activity of radionuclides is calculated by

$$A = \sigma\phi N(1 - e^{-\lambda ti})e^{-\lambda td}(1 - e^{-\lambda tc})/\lambda \tag{6}$$

The actual number of events recorded by a detector for a particular radionuclide is only a fraction f of the number of decays calculated from Eq. 6, because not every decay can emit a characteristic gamma ray, and once a gamma ray is emitted they may not reach the detector. Considering these findings, the simplest and most accurate way to determine an element by NAA is to irradiate and measure a comparator standard with an exact known content of the element together with the sample. In this case, the ratio of the element content in sample m_s to that in comparator standard m_c is equal to the ratio of their activities, $A_s/A_c = m_s/m_c$. Therefore, the content of target element in the sample can be calculated by

$$m_s = \frac{A_s}{A_c} m_c \tag{7}$$

In addition, from such a single comparator, it is also possible to calculate the sensitivity of other elements by means of the k-factor (5). This factor is an experimentally determined ratio of saturation specific activities expressed in counts:

$$k = \frac{f\theta\sigma}{f^*\theta^*\sigma^*}\frac{M^*}{M} \tag{8}$$

Here, the asterisk refers to the element of a single comparator. The factor f is a combination of the emission probability of η and detection efficiency ε. If the relative efficiency function of the detector is known, calibration may be based on k_0 values (5):

$$k = k_0 \frac{\varepsilon}{\varepsilon^*} \tag{9}$$

where

$$k_0 = \frac{\eta\theta\sigma}{\eta^*\theta^*\sigma^*}\frac{M^*}{M} \tag{10}$$

These k_0 values are fundamental constants and may be found in tabulation; methods for taking into account the influence on k_0 values of difference in neutron flux spectrum have been developed.

The analytical sensitivity of NAA for various elements can be predicted from Eq. 6, combined with the emission probability of gamma rays of the radionuclide and the counting efficiency of the characteristic energy of the radionuclide. The calculated interference free detection limits of NAA for various elements are listed in Table 1. Note that this data is only applied to radionuclides completely free from all other radionuclides, that is after a complete radiochemical separation. In actual analysis, the activity from other activated elements will interfere with the detection of the target element by increasing the baseline counts under its peaks, so the detection limit will be poorer than that estimated in Table 1. However, in a sample with known composition, practical detection limits may be predicted in advance (6).

Since NAA is based on the excitation of the atomic nucleus of an isotope instead of the surrounding electrons, no information on the chemical state of the element can be obtained. In addition, the de-excitation of the produced nucleus by the emission of high energy gamma rays or other particles gives the formed radionuclide a nuclear recoil energy of several tens of electron volts. This is more than sufficient to break a chemical bond, and the formed radionuclide may no longer be found in its original chemical state. It is therefore not possible to directly use NAA for chemical speciation of an element. However, a NAA

Table 1. Interference Free Detection Limit for Elements by NAA Based on Irradiation for 5 h at a Neutron Flux Density of 10^{13} n/cm$^2{\cdot}$s^{-1} and Typical Counting Conditions[a]

Detection limit, ng	Element
0.001	Dy, Eu
0.001–0.01	Mn, In, Lu
0.01–0.1	Co, Rh, Ir, Br, Sm, Ho, Re, Au
0.1–1	Ar, V, Cu, Ga, As, Pd, Ag, I, Pr, W, Na, Ge, Sr, Nb, Sb, Cs, La, Er, Yb, U
1–10	Al, Cl, K, Sc, Se, Kr, Y, Ru, Gd, Tm, Hg, Si, Ni, Rb, Cd, Te, Ba, Tb, Hf, Ta, Os, Pt, Th
10–100	P, Ti, Zn, Mo, Sn, Xe, Ce, Nd, Mg, Ca, Tl, Bi
100–1000	F, Cr, Zr, Ne
10,000	Fe

[a]See Ref. (6).

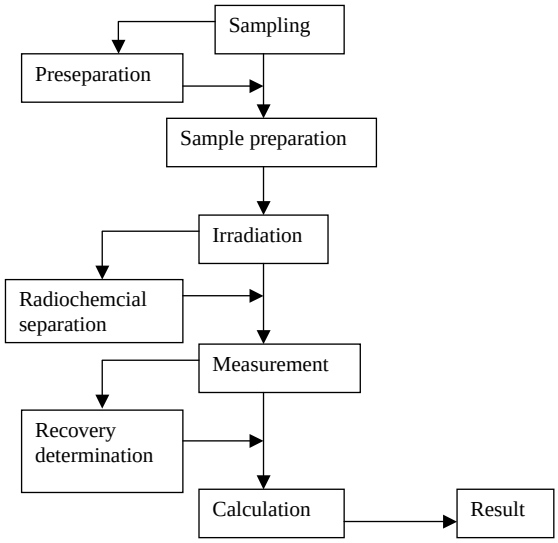


Figure 1. Procedure of neutron activation analysis.

detection coupled with a preseparation of chemical species of elements, so-called molecular activation analysis, can be applied for the chemical speciation of elements for biomedical studies (7). Used in this way, NAA is no longer without a blank problem.

EQUIPMENT AND METHODOLOGY

Optimal utilization of the special qualities of NAA requires an appropriate methodology, which is adapted to the analysis of various biomedical samples for different elements and their species. Figure 1 shows a scheme of NAA, which divides NAA into several steps, some of which are indispensable, while others are supplementary for specific purposes.

Sampling

Sampling is the first step for any analysis, and is the most important step in any meaningful investigation (8,9). There is increasing awareness that the quality of an analytical result may often be influenced more by sample collection than by final measurements. One of the main considerations during sampling should be devoted toward the representativity of the analyzed sample to the object under study. Other considerations should be the avoidance of any contamination during sampling and finding suitable storage of the sample until analysis to prevent the losses of the elements of interest and to avoid contamination. The selection and collection of a representative sample may need to consider two aspects: the type of sample and the size of the selected sample (8). Frequently, valuable data can be obtained from the analysis of readily accessible *in vivo* samples, such as blood and urine. In some cases, the analysis of tissue sample may provide more useful information. A considerable effort has been spent on the analysis of hair and nail (9). They have an advantage of being easily available, however, it may be less representative for most of the elements, and a multitude of factors may affect the observed results. The contamination during the sampling mainly comes from the tools and the container used for samples and the sampling environment. With the exception of excreta, all biomedical materials have to be removed from the organism by means of tools for piercing, cutting, or shearing. The best materials for tools from the point of view of eliminating contamination are polyethylene, Teflon, or other plastic materials, or stainless steel, if Cr, Ni, Co, and Fe are not being considered, but pure titanium is a good material due to its hardness for cutting. For the container, polyethylene, Teflon, and synthetic quartz are the best materials to avoid contamination and absorption of trace elements on the well of the container. The problems of sampling and sample handling have been given more attention in the studies of medical trace elements. A number of sampling protocols were recommended. A detailed discussion of this issue can be found in many articles and books (2,8,9).

Preseparation

Since NAA cannot be directly used for chemical speciation, various chemical species of trace elements have to be separated before irradiation. In recent years, many methods have been developed for the chemical speciation of various trace elements, such as Se, I, Cr, Hg, As, Al, Fe, Zn, Cu, and rare earth elements in biomedical samples (7,10–15). All of these methods are based on preseparation using various chemical and biochemical separation techniques, such as ion exchange, gel chromatography, high-performance liquid chromatography (HPLC), supercritical fluid, electrophoretic, and fractionation techniques. The chemical speciation in biomedical trace elements studies normally includes the following aspects: subcellular distribution of trace elements in various tissues; chemical valence states of trace elements in liquid samples such as urine, blood, and cytosol of tissue homogenate; specific combination of trace elements with various proteins, polypeptides, enzymes, hormones, and small molecular compounds. The separated chemical species of trace elements is then quantitatively determined by NAA. Since the sample is not destroyed and no chemical reagents are used, NAA itself is actually a contamination-free analytical technique. However, in preseparation the utilization of chemical reagents and many types of equipment can cause a high risk of contamination. An enriched stable isotope trace technique was therefore used to overcome this problem (15), which is based on the fact that NAA is actually an analytical technique for isotopes instead of elements. In this case, the contaminations during the separation can be identified and eliminated because of the different isotopic components of the trace elements in the samples with those contaminants from the chemical reagents and other sources.

Due to the very low concentration of most trace elements in biomedical samples, preseparation of elements of interest from main interfering elements and preconcentration is sometimes useful for the improvement of the detection limit of the elements of interest. The simplest preconcentration should be lyophilization and ashing of the samples, which can be directly carried out in an irradiation container, such as a quartz ampoule and aluminum foil to eliminate problems associated with sample transfer. The

techniques used for preseparation of the elements of interest are similar as those for radiochemical separation described below. Special attention should be given to avoid contamination and addition of extra interfering elements, such as Cl and Na.

Sample Preparation

Before neutron irradiation, the sample must be transferred to an irradiation container for transport to and from the irradiation position. These containers should not contaminate the sample, nor should they add significantly to the total amount of radioactivity produced during activation. Suitable materials are low density polyethylene for moderate irradiation and high pure aluminum and synthetic quartz for prolonged irradiation. For short-term irradiation, the sample can be wrapped in a thin polyethylene film and inserted in a polyethylene irradiation capsule. After irradiation, the sample with the polyethylene foil is directly measured, since there is no significant contribution of trace elements and radioactivity from the polyethylene film and difficulties on removal of sample from the irradiated film. In the determination of some elements, such as Se, F, B, and Li (16,17), the irradiated samples have to be measured very quickly ($<$10 s) due to the very short half-lives of the formed radionuclides (0.8 s for ^{8}Li, 17.5 s for ^{77m}Se, 11.0 s for ^{20}F). In this case, the sample with the polyethylene capsule has to be directly transferred to the detector for measurement. For long-term irradiation of sample in a reactor, it is normally required that the sample be dried to avoid any problem of explosion of the container due to the high pressure produced in the container by the radiolysis of water in the samples. In addition, the removal of water from the sample will improve the determination of elements of interest, because an increased amount of sample can be irradiated and the counting geometry can be improved. Lyophilization techniques are normally used for the removal of water, although evaporation by heating can be used. In addition, dry ashing is sometimes used to further improve the determination of elements, but we should watch for loss of elements of interest during this process. The dried sample is normally wrapped in high purity aluminum foil, which is then inserted into an aluminum container for irradiation, since only a short-lived radioisotope of aluminum, ^{28}Al (2.24 min), is formed in the neutron irradiation of Al. This radioisotope quickly disappears by decaying, so it does not interfere with the determination of trace elements that form long-lived radionuclides. For some elements, such as mercury and arsenic, a synthetic quartz ampoule was used because they volatize during irradiation and have very low impurities in quartz materials and are less radioactive from activated SiO_2. To minimize losses and contamination, biomedical samples are often sampled in a high purity quartz ampoule. They are then directly lyophilized and ashed in the quartz ampoule. After irradiation, they can be measured in the same ampoule.

Irradiation

Different neutron sources are available for the activation analysis of samples. Isotopic neutron sources, such as Ra–Be and Am–Be sources that rely on the (α, n) reaction and ^{252}Cf source that is based on spontaneous fission, yield neutron spectra with a limited range of neutron energies. Accelerators can also be used to produce a particular energy of neutrons, such as 14 MeV neutrons produced by (D,T) reaction and 2.5 MeV neutrons by (D,D) reaction. A miniature, sealed-tube neutron generator with a yield of 10^{8}–10^{11} neutron/s has been used widely as a neutron source for NAA (18,19). However, due to low neutron flux produced in these two kinds of neutron sources, they are very seldom used for the NAA of biomedical samples.

Nuclear reactors are much more suitable for NAA of biomedical materials, because they provide much higher neutron flux density (10^{11}–10^{14} n/cm^2·s) with correspondingly higher sensitivities for trace elements. In addition, neutrons in the reactor can be well moderated to thermal energies, which is very suitable for NAA because of the high thermal neutron activation cross-sections of trace elements and less interference from fast neutron reactions. For some elements such as iodine, arsenic, and selenium, which are preferably determined by epithermal NAA, an epithermal neutron irradiation channel was set up in some reactors by shielding thermal neutrons with cadmium (20). Almost all research reactors installed pneumatic transfer systems for easy and automatic transfer of samples to and from the irradiation position. Some of them can also quickly transfer the irradiated sample to the counting position above the detector for the determination of elements by counting short-lived radionuclides. The detection limit and analytical precision of NAA can be improved by repeated irradiation and counting of samples, the so-called cyclic NAA. This is particularly useful for the elements determined by short-lived radionuclides (such as Se, F, I, etc.). Cyclic NAA systems were therefore installed in some reactors (17). As a nuclear facility, many nuclear reactors are located in large research centers, normally outside of the city; it makes them less accessible and consequently makes NAA inconvenient for most researchers and medical doctors. In the 1980s, a miniature neutron source reactor (MNSR) was developed especially for NAA in China. Due to its intrinsic safety in design, this reactor can be installed in hospitals and research institutes in a big city. With a combination of its small size and low price, it makes NAA possible as a clinically analytical tool easily and conveniently accessible for many researchers and medical doctors. Ten such reactors were installed in universities and institutes in China, Pakistan, Iran, Syria, Ghana, and Niagara (20,21). A few similar reactors (SLOWPOKE reactor) were developed and installed in Canada.

Measurement

Trace elements are determined by measurement of the activity of the radionuclides formed after neutron irradiation, which can be carried out by counting the number of events taking place in the detector. The choice of the detector used for NAA depends on the type of decay and energies of the radiation emitted by the radionuclides formed. Some radionuclides produced by neutron activation of uranium and thorium decay by emission of neutrons, which are counted by means of a neutron detector filled with BF_3 or ^{3}He gas. Meanwhile, the radionuclides of

boron and lithium, which decay by emitting very high energy β particles, may be counted by a Cherenkov detector (16). All these analyses require that the detector be installed at the reactor site because of the short half-lives of these radionuclides, and are consequently not commonly used for biomedical materials.

The most common radionuclides measured by NAA decay by emission of characteristic gamma rays. The most widely used detector is a gamma semiconductor detector, of high purity germanium (HpGe). This detector has high energy resolution; it may separate the gamma rays with energy difference of only 2–3 keV. According to the energies of the radionuclides of interest, different types of germanium detectors can be used. Planar germanium detector and HpGe detector with a thin window are used for the measurement of low energy gamma rays down to 10–20 keV. These detectors have the best energy resolution. In addition, a silicon (lithium) detector can be used for the measurement of low energy gamma rays and X rays. A normal coaxial HpGe detector is used for most radionuclides with energies of gamma rays >60 keV, and well-type HpGe detectors can be used for improvement of the counting efficiency if a very low level of radioactivity needs to be measured. In this kind of detector, the sample sits in the middle of the germanium crystal and almost 4π geometry can be obtained. However, only a small sample can be measured in this detector due to the limited well size of the detector. In connection with appropriate electronic equipment, such as high voltage supply, preamplifier and main amplifier, and a pulse height analyzer with 4,096 or 8,192 channels, it is possible to determine simultaneously many different radionuclides if only their gamma-ray energies differ by 2–3 keV. Additional discrimination is achieved between radionuclides with different half-lives by counting the sample at different decay times.

The measurement of gamma rays is based on the interactions of gamma rays with matter, such as the photoelectric effect, Compton scattering, and pair production. In the photoelectric effect, the gamma ray ejects a shell electron from a germanium atom and produces a vacancy; the ejected electron has a kinetic energy equal to the energy of the gamma ray less the binding energy of the electron. It may interact with other electrons, thus causing secondary ionization, and produce more vacancies in the shell of the germanium atoms. The number of produced photoelectrons and corresponding vacancies is determined by the energy of the gamma ray. In this way, the gamma-ray energy is converted to an electric signal in the detector, which is then amplified by preamplifier and main amplifier, the output from the main amplifier is a peak of nearly Gaussian shape with an amplitude proportional to the gamma-ray energy that enters the detector. This electric signal is finally registered by a multichannel analyzer (MCA) as a count, the number of the channel in the MCA corresponds to the gamma-ray energy and the counts in a channel correspond to the numbers of the gamma rays with the same energy. The peak of the gamma spectrum registered in the MCA is normally also a Gaussian distribution. The width of the peak, or the energy resolution, is an important parameter of the detector. Except for the photoelectric effect, pair production process results in a gamma ray with energy less than 0.511 and 1.02 MeV of the original one. Compton scattering results in a continuum of energy being transferred, which increases the baseline counts of the gamma peaks, and therefore worsens the detection limit of the radionuclide of interest. Most gamma emitting radionuclides also emit beta particles. The interaction of the beta particle with the detector results in a continuum of energy under the gamma spectrum. A beta absorber can be used to reduce the interference of the beta particles. The utilization of anti-Compton techniques will reduce Compton interference; a recently developed multiparameter coincidence spectrum technique significantly improved the detection limit of the trace elements of interest (22).

Radiochemical Separation

NAA for very low level of elements is limited by the presence of other elements in the sample. Some minor elements in the biomedical samples such as Na, Cl, Br, and P contribute to high radioactivity of the irradiated sample, and the signals of the radionuclides produced from many trace elements of interest are overlapped by an increased Compton continuum under the gamma peaks. In order to measure the trace elements of interest and improve the detection limit and analytical accuracy of many trace elements, it is required that these interfering elements be eliminated before counting. Since the irradiated sample is radioactive, this procedure is called radiochemical or postirradiation separation. Radiochemical separation is normally carried out by the following steps: addition of carrier, decomposition of irradiated sample, chemical separation, and preparation of separated samples for counting. Detailed development and the present progress of RNAA were reviewed by several authors (22–24). Since the activated elements are radioactive, this procedure does not create any blank problem, but may easily result in losses of trace elements to be determined. In order to minimize these losses, a suitable amount of carrier element is always added to the sample before radiochemical separation. Since the amount of carrier element is much higher than the same element from the original sample, it can then be used to monitor the chemical recovery of the elements determined during the radiochemical separation. In many cases, some carriers of interfering elements are also added to improve the removal of these interferences, the so-called holdback carrier. A complete equilibrium between an inactive carrier and radioactive element is necessary to be sure that both undergo the same physical and chemical procedure, and the same chemical recoveries during the radiochemical separation, which is normally carried out by a comprehensive oxidation–reduction cycle.

Both dry ashing and wet digestion are used to decompose the biomedical samples. Dry ashing at 400–700 °C is more suitable for large samples; then the ashed sample can be easily dissolved by diluted acid. However, a considerable loss of several elements may occur during ashing due to volatilization, which can be partly eliminated by ashing in a closed system or by using low temperature ashing at 200–250 °C under vacuum. Iransova and Kucera (25) recently showed alkali fusion at high temperature (800–850 °C) is

very effective and rapid for decomposition of the biomedical sample and reduction of the losses of many trace elements. Based on the volatilization of some elements, a combustion method was used to directly separate these elements from the matrix and interfering elements. A particular example is the radiochemical NAA of iodine (26). Acid digestion is a more popular method for decomposition of biomedical samples. This method is normally carried out by boiling a sample in concentrated HNO_3 with $HClO_4$ or H_2SO_4. Sometimes H_2O_2 is added to completely decompose the organic components of the sample. Heating in an open system can also result in losses of some volatile elements such as iodine and mercury; utilization of refluxing can significantly improve the recovery of these elements. Recently, a microwave assisted digestion system was also introduced for the decomposition of sample. The main advantages of this method are rapid digestion and less loss of trace elements due to a closed digestion system. Usually, it is used for small sample analysis.

The main activities of irradiated biomedical material are produced from ^{38}Cl, ^{82}Br, ^{24}Na, and ^{32}P. A complete separation of an element from all other radionuclides is rarely necessary. The removal of main radionuclides and several group separations are very useful for improvement of the determination of most trace elements with adequate accuracy and precision (23,24). Precipitation, ion exchange, and extraction are the most commonly used methods for radiochemical separations. Both ^{38}Cl and ^{82}Br can be removed by anion exchange absorption, precipitation as AgCl and AgBr, and evaporation as HCl and Br_2. A simple and effective separation of ^{24}Na can be carried out by absorption on a hydrated antimony pentoxide (HAP) column (27). Various procedures have also been developed for the group separation (23,24,27). However, the separation of a single element sometimes may be necessary due to their very low level in biomedical materials, such as iodine, vanadium, silicon, uranium, selenium, and mercury (26,28,29). For efficient measurement, the separated sample has to be prepared in small amounts for counting on the detector. This process is normally completed by precipitation to convert the separated elements to a solid form or by evaporation to remove most of the water.

Chemical Recovery

In comprehensive radiochemical separation, the addition of carrier cannot entirely prevent losses of the elements of interest, even in a simple separation procedure. For accurate results, a correction for losses should therefore always be made, which can be carried out by measurement of the chemical recovery of the determined element in the radiochemical separation.

Chemical recovery can be measured by carrier and radiotracer. Before decomposition, carriers are added to the irradiated material in macroamount compared to the element originally present in the sample. When the carriers behave as the probed element in the sample, their chemical recoveries should be the same. Thus the chemical recovery of the carrier element is taken as the recovery of the determined element. The carrier element can be easily measured by classic analytical methods, such as gravimetry, calorimetry, and titration method, especially when a single carrier is added and separated, and the contribution from other elements is negligible. When more than one carrier is added and multielements are separated and determined, a reactivation method could be used for determination of the carrier content in the separated sample. After the measurement, the separated sample containing the activated elements and the carriers is reactivated by irradiation with neutrons, and the amount of carrier is determined by NAA. As the amount of the radionuclide originated from the sample is negligible compared to the added carrier, and the amount of added carrier is known, the chemical recovery can be calculated. Radiotracer is a more direct method for monitoring chemical recovery. In this case, radioisotopes are added to the sample with carrier before the radiochemical separation. These radiotracers are not the same as those produced by neutron activation, and can be measured simultaneously by a gamma detector due to the different energies of gamma rays. Since the radiotracer is another indicator of the same element as the radionuclide produced from the element of interest, the chemical recovery of the radiotracer is the same as the indicator. For example, ^{125}I was used as a radiotracer for monitoring the chemical recovery of iodine in RNAA (29), and ^{57}Co for cobalt.

If the radiochemical separation is carried out in a well-controlled manner, a constant chemical recovery can be assumed, and the recovery correction may be carried out without measurement of chemical recovery for every individual sample. But first the constant chemical recovery has to be determined by processing an unirradiated sample to which is added irradiated carriers or other radiotraces using the same chemical separation procedure. In order to evaluate the precision of the correction, at least 10 determinations should be made.

Analysis of Gamma Spectrum and NAA Calculation

The data acquired by a germanium detector are registered in the MCA, and may be transferred directly or via an analyzer buffer to a computer for processing. At present, MCA can even be made as an interface card, which can be directly inserted to a personal computer. Then the computer can control the gamma detector and the gamma spectra can be acquired and analyzed by computer software. Data acquisition software is chiefly concerned with the handling of the MCA system and its components; the programs will connect the hardware of the spectrometry system with the storage memory for the data. Critical physical parameters such as starting time, duration, and dead-time of the acquisition are recorded together with the spectral data. The acquisition software can also take care of controlling sample changers and automatic sample transfer systems for cyclic NAA. The acquired spectrum can be stored separately or summarized.

Many γ-spectrum analysis programs have been developed to analyze gamma spectra, which can search for γ peaks, calculate their energies, and their peak areas. The special software for NAA can even identify nuclides by the energy of γ peaks and decay time, calculate the radioactivity of a radionuclide, and finally calculate the concentrations

of the trace elements in the sample. The energy of a gamma peak can be identified by calibration of the counting system by measurement of some known gamma-ray sources. By comparison with a database of gamma-energy and half-life of radionuclides, the identification of the nuclides can be carried out. The calculation of the peak area is the most critical step in the NAA calculation. Many techniques have been developed to calculate the peak area and to subtract the baseline under the peak, such as Gaussian shape and experimental peak shape fitting. When the peak area and the efficiency of the detector are measured, the activity of the nuclide can be calculated easily by correcting for decay and counting time. If the comparator method is used, the contents of elements of interest can then be calculated using Eq. 7 after the measurement of a comparator element standard. In NAA, overlapping peaks may sometimes occur, when the energy difference between peaks is not large enough, a significant error may result from incorrect separation of peaks and baseline subtraction. However, with proper execution of NAA and correction of results for possible separation losses, NAA is capable of providing unbiased results with known precision for a multitude of trace elements in biomedical materials.

Evaluation and Quality Assurance

NAA has been demonstrated to be reliable and under statistical control due to the absence of unknown sources of variability. This means that it is possible to predict the standard deviation of analytical results so well that the observed and expected variation among replicate data are in agreement. The special qualities of NAA make the method one of the best for the certification of reference materials in the biomedical field. Hardly any certification of trace elements is carried out without considering this method. Its high sensitivity for many elements and the absence of blank values make NAA the preferred method for analysis at an ultratrace level of concentration in many types of biomedical samples. The insensitivity to contamination has proven particularly valuable for the establishment of a normal concentration of a number of elements in human samples. Results by NAA serve as a reference for testing or verifying the reliability of other methods.

The common sources of uncertainty in NAA come from irradiation, counting, spectra acquisition and analysis, blank and interference of nuclear reactions and gamma rays. The uncertainty of irradiation is contributed from the variation of neutron flux during irradiation and inhomogeneous distribution of neutron flux in the irradiation container. The instability of neutron flux may significantly influence the analytical results of elements determined by measuring short-lived radioisotopes, because the sample and standard are not irradiated simultaneously. This problem is normally overcome by on-line monitoring of the neutron flux, and most research reactors, especially a small reactor such as MNSR, installed such a system. In most cases, the uncertainty from this source is low ($<0.5\%$). The neutron flux gradient is sometimes very significant in a big irradiation container, so that a flux monitor foil of Fe, Co, or Au alloy may be used to monitor the different positions to make a neutron flux correction. Uncertainty in counting mainly results from the variation of counting geometry. The uncertainty due to counting geometry can be controlled reliably by well efficiency calibration of the detector for various source shapes; software is available for appropriate calculation. The matrix effects can suppress gamma rays, especially low energy gamma rays, by self-absorption; this effect can be calibrated by measurement of a standard with a similar matrix with sample. Many natural matrix certified reference materials (CRM) have been prepared for this purpose and are commercial available (30).

Uncertainty in the spectra acquisition may come from dead time and pile-up losses of signals. A quick varying and high counting rate often occur in the NAA of biomedical materials due to high concentrations of Cl and Na which may cause a high and varying dead time during the counting. Dead time is normally measured by the gamma spectra system, and live time is used for activity calculation. But when the dead time is quickly changed during counting, the activity may be underestimated by this method. This problem can be solved by using a loss-free counting method, which estimates the number of counts lost during a dead time interval and adds this number to the channel of the just processed pulse, thus presenting a loss-corrected spectrum. The pile-up losses are corrected by electronic or computational mean built in the counting system. Uncertainty in the spectra analysis results mainly from the evaluation of peak area and subtraction of baseline, especially for the analysis of multipeaks. A large attempt has been made to develop effective software to improve the analysis of the gamma spectra.

The blank problem can usually be ignored in NAA. But a blank correction will be necessary, when preseparation and sample preparation are used before irradiation. A given radionuclide can often be produced in more than one way. If the indicator nuclide used in NAA is produced from an element other than the element determined, then nuclear reaction interference occurs, such interference is mainly produced by (n,p) or (n, α) reactions with fast neutrons and elements with an atomic number 1 or 2 above the element to be determined. A typical example of such interference is the formation of ^{28}Al from Al, P, and Si by reactions: ^{27}Al(n, γ)^{28}Al, ^{28}Si(n, p)^{28}Al, and ^{31}P(n, α)^{28}Al. In biomedical materials, the concentrations of Al and Si are very low, but P is high in many kinds of samples. The contribution of P to the activity of ^{28}Al may be very significant, especially when the fraction of fast neutron in the irradiation position is high, it will seriously interfere with the determination of aluminum. This interference can be significantly reduced by using a thermal neutron irradiation channel, where the ratio of thermal/fast neutron flux is very high; in many research reactors, such a thermal neutron channel is available. Since there are no suitable thermal neutron activation products of P and Si, they are determined by using these two fast neutron reactions in NAA, so the interference from Al to P is also very significant due to a much higher cross-section of thermal neutron activation than fast neutrons. The correction can be carried out by the irradiation of the sample with and without thermal neutron shielding. The activity of ^{28}Al in a sample irradiated under thermal neutron shielding (such as in a Cd or B

container) mainly contributes from fast neutron reaction, while under conditions without shielding from both thermal and fast neutron reactions.

Double and triple neutron activation reactions can also result in an interference, for example, the interference from $^{197}Au(2n, \gamma)^{199}Au$ to the determination of Pt by $^{199}Pt(n,\gamma)^{199}Pt(\beta^-)^{199}Au$ reaction, and from $^{127}I(3n, \gamma)^{130}I$ to the determination of ^{129}I by $^{129}(n, \gamma)^{130}I$ reaction (29). But, interference from triple reactions is normally negligible due to their very low probability.

A special type of nuclear reaction interference is caused by the presence of uranium, which yields a large number of radionuclides as a result of fission. The greatest correction is required in the determination of molybdenum and some rare earth elements (REs), but the concentrations of uranium in biomedical materials are usually too low to present any problem.

Gamma-ray spectral interference means that two radionuclides emit gamma rays with the same or nearly the same energy. For example, ^{51}Ti and ^{51}Cr used for NAA of Ti and Cr emitting a single 320.08-keV gamma ray, ^{75}Se and ^{75}Ge used for Se and Ge emitting 264.66 keV gamma ray. For these nuclides, a separation via decay is possible practically due to their different half-lives. However, the interference for the determination of Hg via the 279.19 keV line of ^{203}Hg (46.6 days), which overlapped by the 279.54 keV line of ^{75}Se (119.77 days), cannot be eliminated by decay because their half-lives are not significantly different. This problem can be resolved by measurement of ^{75}Se by other gamma lines and correcting the contribution of ^{75}Se to the peak of 279.54 keV.

As an effective method for analytical quality control, certified reference materials (CRMs) with a matrix similar to the samples, which have appropriate combinations of elements to be determined and with concentration bracketing the range of interest, are normally analyzed to evaluate the analytical accuracy and precision. A large number of CRMs have been prepared by many countries and international organizations (30), which are not only for analytical quality control of NAA, but actually all other analytical methods.

Uncertainty from the sampling procedure and inhomogeneity of the sample should also be considered as a probable contribution to the total uncertainty of the analytical data, although it is not directly related to the NAA method itself. Especially at very low concentration, the trace elements distribution in samples is normally inhomogeneous and a selection of a representative sample is very important for analytical quality. In recent years, the NAA of large samples has been given much attention (31). In this case, a whole sample is analyzed and the uncertainty resulting from the inhomogeneity of elements in the sample can be overcome.

Applications

As an analytical technique for determination of trace elements, the NAA has been widely used in various aspects of biomedical studies and analysis related to trace elements, such as an investigation of the normal trace element level in the human body, trace element related endemic disorder and inheritable diseases, environmental exposure of toxic and nonessential trace elements and their effect on human health, and investigation of trace element nutrition status.

The distribution and normal level of trace elements in blood, urine, and human tissue are basic data for the studies on biomedical trace elements. Some activities have been organized by the International Atomic Energy Agency (IAEA) for establishing the normal level of trace elements in "reference man" with different dietary habits and in different geographic areas. This information was completed by analyzing various normal human tissues collected from different countries in which most of the data was supplied by NAA (32,33). In addition, many NAA were also carried out to analyze a large number of blood, urine, hair, and nail samples for normal trace element level determinations. In combination with the analysis of diets and environmental samples, some trace element related endemic disorders have been investigated, such as Keshan disease found in China, which is related to the deficient intake of selenium, and normally also associated with the deficiency of iodine. Some disorders caused by excessive intake of arsenic from ground water and mercury from contaminated fish and foods have also been investigated by NAA in China and many other countries. Trace element nutrition status by analysis of the trace element level in blood, hair, and diets has also been investigated by NAA in many countries. The results were used as a basis for improved recommendations of safe and adequate, daily average intake of these elements.

Some inheritable diseases have been attributed to the metabolism problem of trace elements, such as Menkes' syndrome, Wilson's disease, and Acrodermatitis enteropathica. Determination of specific trace elements in a small amount of living tissue can be used to diagnose disease and to evaluate the treatment. The ability to analyze very small samples by NAA creates a possibility of following temporal changes in trace element concentrations in a living subject. Menkes' disease was found to be a recessive X-linked disturbance of copper metabolism, resulting in accumulation of copper in several extra-hepatic tissues, including the placenta, by NAA of various human tissues of patients and normal persons (2). The method is being used to diagnose Menkes' disease and identification of female carriers by NAA of the placental tissues (2,34). Wilson's disease was also confirmed to be a copper metabolic defective disease; determination of copper in liver biopsy by NAA was used as a control and verification method for estimation of the adequate treatment of this disease.

A high environmental exposure may cause the increase of some nonessential trace elements in the human body, which may create a potential effect for human health. In recent years, rare earth elements found wide application in industry and agriculture, rare earth trace element fertilizer has widely been used in China for increasing the yield of various agricultural products. It also increases the exposure of humans to these trace elements. Neutron activation analysis was used to investigate the uptake of these elements in the human body via the food chain, their distribution in human tissues, and their combination with different components of tissue, especially brain tissue (7,13).

Since hair and nails are easily obtained compared to blood and tissue samples, they serve as a useful indicator of trace element levels in the body. Neutron activation analysis is very suitable for the analysis of hair and nails, because no decomposition of sample is necessary, therefore they are often used for the analysis of these materials for investigation of the nutrition status of trace elements, the environmental exposure level of toxic elements, and the diagnosis of various diseases related to trace elements (9).

A series international conference entitled Nuclear Analytical Methods in the Life Sciences was held since 1967, and the three most recent conferences in this series were held in 1993, 1998, and 2002. The main achievements in this field can be found in the Proceedings of these conferences (35,36) and other relevant conferences (37,38).

BIBLIOGRAPHY

1. Prasad AS, Oberleas D. Trace elements in human health and disease. New York: Academic Press; 1976.
2. Heydorn K. Neutron activation analysis for clinical trace element research. Boca Raton: CRC Press; 1984.
3. Chettle RD, Fremlin JH. Techniques of in vivo neutron activation analysis. Phys Med Biol 1984;29:1011–1043.
4. Heydorn K. Neutron activation analysis. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation. New York: John Wiley & Sons; 1988.
5. deCorte F, Simnits AS. Recommendation data for use in the $k(0)$ standardization of neutron activation analysis. Atomic data and nuclear data tables 2003;85(1):47–67.
6. Guinn VP, Garzanov E, Cortes E. Further studies in advance prediction of gamma ray spectra and detection limits in instrumental neutron activation analysis. J Radioanal Chem 1978;43(2):599–609.
7. Chai CF, Mao XY, Wang YQ, Sun JX, Qian QF, Hou XL, Zhang PQ, Chen CY, Feng WY, Ding WJ, Li XL, Li CS, Dai XX. Molecular activation analysis for chemical species studies. Fresenius J Anal Chem 1999;363(5–6):477–480.
8. Parr RM. Technical consideration for sampling and sample preparation of biomedical samples for trace element analysis. J Res Nat Bur Stand 1986;91(2):51–57.
9. Chatt A, Katz SA. Hair Analysis, application in the biomedical and environmental sciences. New York: VCH; 1988.
10. Patching SG, Gardiner PHE. Recent developments in selenium metabolism and chemical speciation: A review. J Trace Element Med Biol 1999;13(4):193–214.
11. Shoop DM, Blotcky AJ, Rack EP. Distribution of aluminum species in normal urine as determined by chemical neutron activation analysis. J Radioanal Nucl Chem 1998;236(1–2): 103–106.
12. Hou XL, Chen CY, Ding WJ, Chai CF. Study on chemical species of iodine in human liver. Biological Trace Element Res 1999;69(1):69–76.
13. Chai ZF, Zhang ZY, Feng WY, Chen CY, Xu DD, Hou XL. Study of chemical speciation of trace elements by molecular activation analysis and other nuclear techniques. J Anal Atomic Spectrom 2004;19(1):26–33.
14. Chen CY, Zhang PQ, Hou XL, Chai ZF. Subcellular distribution of selenium and Se-containing proteins in human liver. Biochim Biophy Acta 1999;1427(2):205–215.
15. Feng WY, Li B, Liu J, Chai CF, Zhang PQ, Gao YX, Zhao JJ. Study of chromium-containing proteins in subcellular fractions of rat liver by enriched stable isotopic tracer technique and gel filtration chromatography. Anal Bioanal Chem 2003;375(5):363–368.
16. Heydorn K, Skanborg PZ, Gwozdz R, Schmidt JO, Wacks ME. Determination of lithium by instrumental neutron activation analysis. J Radioanal Chem 1977;37:155–168.
17. Hou XL. Cyclic activation analysis. In: Meyers RA, editor. Encyclopedia of Analytical Chemistry. Applications, theory and instrumentation. Chichester: John Wiley & Sons; 2000.
18. Reijonen J, Leung KN, Firestone RB, English JA, Perry DL, Smith A, Gicquel F, Sun M, Koivunoro H, Lou TP, Bandong B, Garabedian G, Revay Z, Szentmiklosi L, Molnar G. First PGAA and NAA experimental results from a compact high intensity D-D neutron generator. Nucl Instr Meth 2004;A522(3): 598–602.
19. Chichester DL, Simpson JD. Advanced compact accelerator neutron generator technology for FNAA/PGNAA field work, Abstract of 11th International Conference on Modern Trends in Activation Analysis. Guildford, UK; June 2004.
20. Hou XL, Wang K, Chai CF. Epithermal neutron activation analysis and its application in the miniature neutron source reactor. J Radioanal Nucl Chem 1996;210(1):137–148.
21. Khamia I, Al-Somel N, Sarheel A. Neutron flux micro-distribution and self-shielding measurement in the Syrian Miniature Neutron Source Reactor. J Radioanal Nucl Chem 2004;260(2): 413–416.
22. Hatsukawa Y, Toh Y, Oshima M, Hayakawa T, Shinohara N, Kushita K, Ueno T, Toyoda K. New technique for the determination of trace elements using multiparameter coincidence spectrometry. J Radioanal Nucl Chem 2003;255(1):111–113.
23. Heydorn K. Radiochemical neutron activation analysis. In: Meyers RA, editor. Encyclopedia of Analytical Chemistry. Applications, theory and instrumentation. Chichester: John Wiley & Sons; 2000.
24. Alfassi ZB. Determination of trace elements, Balaban Publisher; 1994.
25. Krausov I, Kucera J. Fast decomposition of biological and environmental samples for radiochemical neutron activation analysis, Abstract of 11th International Conference on Modern Trends in Activation Analysis. Guildford, UK; June 2004.
26. Dermelj M, Byrne AR. Simultaneous radiochemical neutron activation analysis of iodine, uranium and mercury in biological and environmental samples. J Radioanal Nucl Chem 1997;216(1):13–18.
27. Girardi F, Sabbioni E. Selective removal of radio-sodium from neutron activated materials by retention on hydrated antimony pentoxide. J Radioanal Chem 1968;1(2):169–178.
28. Heydorn K, Damsgaard E. Simultaneous determination of arsenic, manganese and selenium in biological materials by neutron activation analysis. Talanta 1980;20:1–11.
29. Hou XL, Dahlgaard H, Rietz B, Jacobsen U, Nielsen SP, Aarkrog A. Determination of iodine-129 in seawater and some environmental materials by neutron activation analysis. Analyst 1999;124:1109–1114.
30. Analytical Quality Control Service, International Atomic Energy Agency. www.iaea.org/programmes/aqcs/database/database_search_start.htm, 2004; February 25.
31. Bode P, Overwater RMW, DeGoeij JJM. Large-sample neutron activation analysis: Present status and prospects. J Radioanal Nucl Chem 1997;216(1):5–11.
32. Iyengar G, Kawamura H, Dang HS, Parr RM, Wang JW, Natera ES. Contents of cesium, iodine, strontium, thorium, and uranium in selected human organs of adult Asian population. Health Phys 2004;87(2):151–159.
33. Iyengar GV. Reevaluation of the trace element content in reference man. Rad Phys Chem 1998;51(4–6):545–560.
34. Tonnesen T, Horn N, Sondergaard F, Mikkelsen M, Boue J, Damsgaard E, Heydorn K. Metallothionein expression in placental tissue in Menkes' disease—an immunohistochemical study. APMIS 1995;103(7–8):568–573.

35. Procedings of 7th international conference on nuclear analytical methods in the life science; June 16–22, 2002, Antalya Turkey: J Radioanal Nucl Chem 2004;259:1–539.
36. Proceedings of the International Conference on Nuclear Analytical Methods in the Life Sciences – Sep. 1998, Beijing, China: Bio. Trace Elem; 1999;71(2).
37. Proceedings of the 10th international conference on modern trends in activation analysis, 19–23, April, Maryland: 1999. J Radioanal Nucl Chem 2000;244:1–704 and 245:1–228.
38. Proceedings of the 11th international conference on modern trends in activation analysis; 20–25 June, Guildford, UK: 2004. J Radioanal Nucl Chem 2005; in progress.

Reading List

Afassi ZB. Activation analysis. Boca Raton: CRC; 1990.
Ehmann WD, Vance DE. Radiochemistry and Nuclear Methods of Analysis. New York: John Wiley & Sons; 1991.
Parry SJ. Activation spectrometry in chemical analysis. New York: John Wiley & Sons; 1991.

See also BORON NEUTRON CAPTURE THERAPY; RADIONUCLIDE PRODUCTION AND RADIOACTIVE DECAY; TRACER KINETICS.

NEUTRON BEAM THERAPY

RICHARD L. MAUGHAN
Hospital of the University of Pennsylvania

INTRODUCTION: THE ORIGINS OF FAST NEUTRON THERAPY

The year 1932 was a remarkable one at the Cavendish Laboratory of the University of Cambridge. It represented a pinnacle of achievement for Lord Ernest Rutherford and his collaborators in the nuclear physics laboratories. The results of three experiments performed in this single year resulted in Nobel Prizes in physics for four of the university faculty. Two of these experiments were of great significance for the production of neutrons. Of course, the most significant was the discovery of the neutron itself by James Chadwick in February, 1932 (1). Chadwick needed a source of neutrons to perform this experiment. At that time, the only sources of energetic heavy particles were naturally occurring isotopes that emitted alpha particles. Chadwick's neutron source was comprised of a polonium source (^{210}Po), which emits a 5.3 MeV alpha particle and a block of beryllium housed in a vacuum chamber. The nuclear reaction

$$\alpha + {}^{9}\mathrm{Be} \longrightarrow {}^{12}\mathrm{C} + \mathrm{n}$$

produces a flux of neutrons with energies of several million electronvolts (MeV). To detect these neutrons, Chadwick used an ionization chamber with a thin entrance window. Pulses could be observed on a oscillograph as the source was brought closer to the ionization chamber. The chamber was air filled and the increasing count rate was interpreted as being due to recoil nitrogen nuclei in the chamber. When a sheet of paraffin wax was placed in between the source and the ionization chamber a further increase in the source count rate was observed and interpreted as resulting from recoil protons from n–p elastic scattering. Further measurements confirmed that the particles producing the recoil protons were neutral particles of the same mass as the proton. With this simple and elegant apparatus Chadwick discovered the neutral nuclear particle, the existence of which Rutherford had postulated in 1919.

Sources of this type give very low neutron fluxes and are not suitable for neutron radiation therapy or radiobiology. It was another of the Nobel Prize winning discoveries of 1932 at Cavendish Laboratory that offered the means for producing intense sources of neutrons. This was the discovery of artificial transmutation of nuclei by John Cockroft and Ernest Walton in April 1932. In this experiment, Cockroft and Walton used their specially designed high-voltage apparatus (now commonly known as a Cockroft–Walton accelerator) to produce protons of energy 700 keV. When these protons interacted with a lithium target, the resulting mass eight compound nuclei disintegrated into two alpha particles with the release of ~17 MeV of energy

$$\mathrm{p} + {}^{7}\mathrm{Li} \longrightarrow {}^{4}\mathrm{He} + {}^{4}\mathrm{He}$$

The alpha particles were detected as scintillations on a zinc sulfide screen and, in order to observe these events, Cockroft or Walton had to sit in a small darkened enclosure at the end of the accelerator tube (2).

Although this first artificial splitting of the atom did not involve the production of neutrons, it was not long before another Rutherford's research team, Mark Oliphant, was to apply the new technology to an experiment to demonstrate the fusion of deuterium nuclei. In this reaction, the products are helium 3 and a neutron

$${}^{2}\mathrm{H} + {}^{2}\mathrm{H} \longrightarrow {}^{3}\mathrm{He} + {}^{1}\mathrm{n}$$

The reaction has a positive Q value of 3.27 MeV and a deuteron beam of 400-keV energy produces neutrons with energies up to ~3.5 MeV.

This experiment was performed in 1934 and almost immediately L.H. Gray, yet another of Rutherford's ex-students, working as one of the first British medical physicists at Mount Vernon Hospital in Northwood, England, realized the potential of this reaction as a source of neutrons for radiobiology research. The ability of neutrons to produce ionizing radiation in the form of heavy recoil particles had been realized soon after their discovery and that these particles might be able to produce biological damage similar to that produced by the recoil electrons associated with X ray beams and radium sources had also been recognized. Neutrons are heavy particles and can transfer relatively large amounts of energy to their secondary recoil particles in comparison to the recoil electrons produced when X rays interact with matter. The neutrons are, therefore, capable of high linear energy transfer (LET) to the recoil particles associated with them and for this reason are known as high LET particles. Gray and others observed that the very great difference between the distribution of ions along the track of a recoiling nucleus and a fast electron made it probable that the biological action of neutrons would show interesting differences from that of

the other radiations. It was hoped that a study of these differences would throw light on the mode of action of radiations on biological material. There was also the possibility that eventually neutrons might prove a more potent means of treating cancer (3).

Gray obtained funding to build a 400 keV Cockroft–Walton accelerator in which he used the deuteron–deuteron reaction to produce a neutron beam for radiobiology research (3). The capital cost of the unit was $2400 with maintenance costs of $320/annum for 1937 and 1938. The unit was housed in a wooden shed.

Concurrently, progress was being made in accelerator technology in the United States by another renowned physicist, Ernest O. Lawrence. Lawrence's invention of the cyclotron earned him the 1939 Nobel Prize for physics and provided a means for producing ion beams with energies of several tens of million electronvolts, far in excess of the energies obtained by Cockroft and Walton with their accelerator. Ernest Lawrence's brother, John, was a physician and the two brothers soon realized the potential medical benefits of applying neutron beams produced by accelerated protons or deuterons. In 1938, funding was obtained from the National Cancer Institute (NCI) for the construction of a 60 in. medical cyclotron; this was the first NCI research grant. This cyclotron was used to produce a neutron therapy beam using a 16 MeV deuteron beam incident on a thick beryllium target. With the use of this beam, Dr. Robert Stone performed the first clinical trials of fast neutrons.

Hence, by 1938 accelerated beams of particles were being used to produce high intensity neutron beams for radiobiology and/or external beam radiation therapy research in both the United States and the United Kingdom.

RADIOBIOLOGICAL RATIONALE FOR FAST NEUTRON THERAPY

The first clinical trials of neutron therapy were performed at the University of California in Berkeley, and were interrupted by the Second World War, when the Berkeley cyclotron was utilized for nuclear physics research associated with the Manhattan Project. Patients with advanced inoperable head and neck tumors were treated with neutron therapy using the same number of treatment sessions, or fractions, as was normally used for conventional X-ray therapy. Although remarkable tumor regression was seen in a few cases, this was at the cost of excessive damage to the surrounding irradiated normal tissues. In 1947, Stone concluded that " neutron therapy as administered by us has resulted in bad late sequelae in proportion to the few good results that it should not be continued", (4). The trials had been undertaken to test the hypothesis that neutron radiation may be superior to X-ray radiation in curing human cancers. However, at the time little was known about the radiobiological effects of neutron irradiation in comparison to conventional X-ray irradiation. It was not until later, when some basic radiobiological research on the effects of neutron irradiation on mammalian cells had been completed, that the reasons for the failure of this original clinical trial could be understood. Further neutron radiobiology research enabled a firm rationale for neutron therapy to be established.

When mammalian cells are irradiated in vitro by ionizing radiation (X rays or neutrons) cells are killed in proportion to the radiation dose delivered. A plot of the cell kill as a function of the radiation dose, or survival curve, is markedly different in shape depending on whether the cells are irradiated by X rays or fast neutrons. This finding is illustrated in Fig. 1, taken from the work of McNally et al. (5). On a log-linear plot at high doses the response appears linear (i.e., exponential), but at low doses there is a "shoulder" on the survival curve. This shoulder is more pronounced for X ray irradiations than for neutron irradiations. Another point to notice is that for a given radiation dose the cell-kill by fast neutrons is greater than for X rays. The relative biological effectiveness (RBE) of neutrons relative to X rays is defined as the ratio the dose of X rays required to produce a given level of cell-kill compared to the dose of fast neutrons required to give the same cell kill. Because of the different shapes and sizes of the shoulders on the neutron and X ray cell survival curves it can be seen from Fig. 1 that the RBE at a surviving

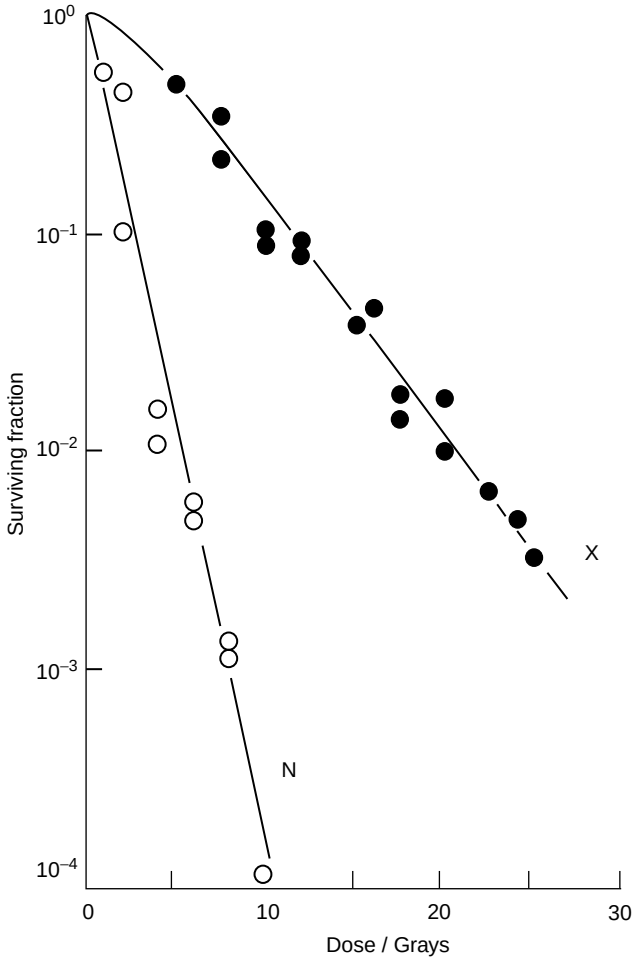


Figure 1. Survival curve for WHFIB mouse tumor cells irradiated with fast neutrons and X rays. Reproduced with permission from McNally, et al. *Rad. Res*. 1982;89:234.

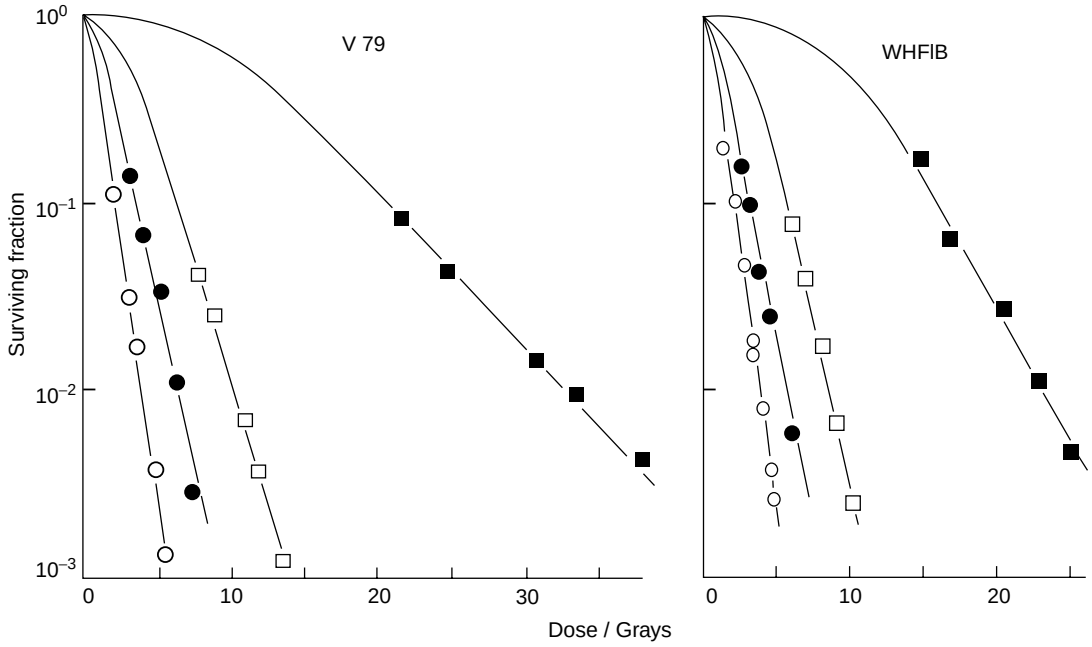


Figure 2. Survival curves for V79 and WHFIB cells, irradiated in air and hypoxia. Squares, X rays; circles neutrons; open symbols, air; solid symbols hypoxia. Reproduced with permission from McNally, et al. *Rad. Res*. 1982;89:232.

fraction of 10^{-2} (i.e., a relatively high dose) is ~3.8, while for a surviving fraction of 3×10^{-1} (i.e., a lower dose) the RBE increases to ~5. John Lawrence measured the RBE in nonhuman biological systems, prior to commencing the Berkeley fast neutron therapy trials with Robert Stone, using large single doses of radiation to produce an observable biological effect. Based on an RBE measured at high single doses, they calculated the required total neutron dose from a knowledge of the total X ray doses delivered to cancer patients at that time. Unfortunately, radiation therapy doses are delivered as a large number of small doses and at these smaller doses the RBE is much larger, hence, the fast neutron dose delivered to patients in the original fast neutron trial was overestimated by a considerable margin. This of course resulted in good tumor control compared to conventional X rays, but also produced an unacceptable level of normal tissue damage. It was not until 1971 that Sheline et al. (6) explained this phenomenon.

In the meantime, L.H. Gray working at Hammersmith Hospital had described a logical rationale for fast neutron therapy. In 1954, in a landmark paper in radiation therapy, Thomlinson and Gray described how, in a poorly vasculated tumor, areas of reduced oxygenation, or hypoxia, can exist as the distance from blood vessels increases (7). If oxygenation drops low enough, the cells become necrotic and die. Gray showed, using diffusion kinetics, that an oxygen concentration gradient can exist within a tumor and that in certain areas of the tumor there may be severely hypoxic, yet viable cells. How this phenomenon can be exploited to advantage in fast neutron therapy is illustrated in Fig. 2. In this figure, cell survival curves for cells irradiated by both X rays and fast neutrons in an oxygen (oxic) and a nitrogen (anoxic) environment are plotted. The anoxic cells are more resistant to radiation than the oxic cells and as tumors are poorly oxygenated and contain anoxic cells this could well result in an inability to deliver sufficient radiation to kill all the tumor cells, leading to tumor recurrence. Normal tissues are well oxygenated and, therefore, are more easily damaged. Hence, it is possible that the doses to the normal tissues surrounding the tumor may reach their acceptable tolerance level before sufficient dose has been delivered to the hypoxic tumor cells to eradicate them. Figure 2 shows that the differential cell killing for oxic and anoxic cells is much greater for X rays than for fast neutrons. Thus for a given level of normal tissue, cell kill or damage neutrons should be more effective at killing hypoxic cells in the tumor than conventional X-ray therapy. It was this hypothesis that lead to the restarting of fast neutron clinical trials at Hammersmith Hospital in London in 1970. The encouraging results obtained at Hammersmith Hospital rekindled interest in fast neutron therapy and by 1980 there were close to 20 centers treating patients.

The clinical results showed that fast neutron therapy appeared to be particularly effective in the treatment of slow growing tumors such as adenocarcinoma of the prostate and bladder. In the meantime, radiobiology research had revealed another important difference between X rays and fast neutrons related to the variation of the radiation sensitivity of mammalian cells at different phases of the mammalian cell cycle. Mammalian cells exhibit well-defined stages in their cycle first described by Howard and Pelc (8). Figure 3 shows a schematic representation of the phases of the cell cycle. Cells spend most of their time in a quiescent phase known as G1 (gap 1), after the G1 phase they move into a phase during which duplicate DNA is synthesized, the S phase. Following DNA synthesis there

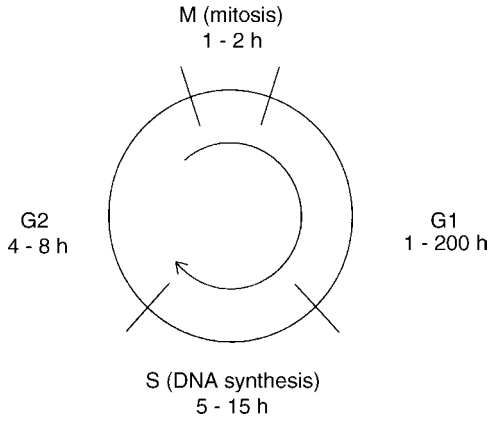


Figure 3. Schematic representation of the phases of the cell cycle.

is another gap phase, G2, which precedes cell division or mitosis (the M phase). The S, G2, and M phase have a similar duration for all mammalians cells, being typically 8, 4, and 1 h, respectively. It is the G1 phase that varies for rapidly and slow growing cells, being as short as 1 h for fast growing cells and as long as 200 h for slow growing cells. The variation in the radiosensitivity of cells irradiated with conventional X rays at different phases of the cell cycle can be considerable (9). The late G1 phase is a relatively radioresistant phase and in cells with a long G1 there is also a period of radioresistance during early G1. As a result, slow growing cells, which spend a larger proportion of the cell cycle in G1 phase than rapidly proliferating cells, would be expected to be more resistant to radiation than the fast growing cells. Withers et al. (10) have shown that the variation of cell sensitivity during the cell cycle is less for fast neutron irradiation than for conventional X-ray radiation. This is a potential advantage for fast neutron therapy, since is means that neutrons are relatively more effective in killing cells in the radioresistant phases of the cell cycle. This would explain the observed efficacy of neutrons in the treatment of slow growing tumors. Another observation, which confirms the efficacy of fast neutrons in treating slow growing tumors, was made by Battermann et al. (11), who showed that the relative biological effectiveness of fast neutrons increases with tumor doubling time.

Thus by the early 1980s it had been established that fast neutron therapy could be justified on radiobiological rationale related to both tumor hypoxia and volume doubling time.

REVIEW OF CLINICAL RESULTS

After the original unsuccessful clinical trials at Berkeley, fast neutron therapy was restarted in 1970 by Dr. Mary Catterall, at Hammersmith Hospital in London. Although Dr. Catterall's studies were not controlled clinical trials they did demonstrate the potential efficacy of fast neutron therapy in the treatment of advanced or recurrent tumors. The success of these studies lead to the establishment of many fast neutron therapy centers around the world and to the instigation of a number of clinical studies of those tumors thought to be the most promising candidates for neutron therapy. These tumors included advanced inoperable tumors of the salivary glands; squamous carcinoma of the head and neck; adenocarcinoma of the prostate, rectum, bladder, and soft-tissue; oesteo- and chondro- sarcomas. The results of these studies demonstrated that fast neutron therapy was particularly promising for the treatment of salivary gland tumors, prostate tumors, and bone tumors. The results for squamous carcinoma of the head and neck, however, were ambiguous. Prompted by these results, the NCI in the United States decided to fund four purpose-built hospital-based neutron therapy facilities to contribute to a definitive series of controlled clinical trials performed at these and other approved centers.

The NCI trials were undertaken in two phases starting with dose searching studies, designated as phase II trials, followed by efficacy studies (phase III trials). The patient accrual into these trials is shown in Table 1. The phase II trials established the general treatment protocol for the phase III studies, that is, a dose of 20.4 Gy would be delivered in 12 fractions over 4 weeks and after 13.6 Gy abdominal and pelvic treatment fields would be reduced in size.

Of the seven phase III trials undertaken, only three were successfully completed; the head and neck, lung, and prostate trials. The salivary gland trial was terminated early; the investigators thought it unethical to continue with randomization because of the excellent results obtained with neutron therapy (12). The remaining three trials were closed because of poor patient accrual. The only other trial to yield a positive result was the prostate trial (13). This trial also highlighted the importance of shaping the beam to conform to the tumor outline. The three centers contributing patients to this trial had facilities with different beam collimation systems of varying degrees of sophistication. The occurrence of normal tissue complications in the bladder and rectum was closely related to the

Table 1. Patient Accrual for NCI–NTCWG Clinical Trials Using the New Generation of Neutron Therapy Facilities in the United States (1984–1991)

Site	No of Patients
Dose searching studies	
Head and neck	59
Thorax	169
Abdomen	78
Pelvis	102
Extremity	92
Subtotal	*500*
Phase III studies	
Salivary gland	9
Head and neck (squamous)	178
Lung	232
Prostate	178
Cervix	28
Rectum	2
Resistant histology	47
Subtotal	*674*
Total accrual	*1174*

sophistication of the beam collimator. The University of Washington, which utilized a sophisticated multileaf collimator for beam shaping, had the least number of treatment related complications, while the M.D. Anderson Hospital in Houston, which used an insert-based collimation with limited shaping capability, had the greatest number of normal tissue complications (13).

An earlier NCI funded Phase III trial for advanced adenocarcinoma of the prostate had shown that a mixed-beam treatment regimen of 60% photons combined with 40% neutrons was superior to conventional photon only therapy (14). With this in mind, Forman et al. (15) undertook a series of in-house neutron therapy trials at Wayne State University using mixed-beam therapy (50% photons and 50% neutrons) for early and some late stage patients. This trial also utilized a sophisticated beam shaping system, a multirod collimator. These trials demonstrated that mixed-beam therapy is as effective as any other state-of-the-art radiation therapy techniques (Intensity Modulated Radiation Therapy or IMRT, proton therapy, brachytherapy) in the treatment of early stage prostate cancer, and that it is probably superior for the treatment of late stage disease (15).

Full reviews of neutron therapy in the treatment of cancer can be found in the work of Wambersie et al. (16) and in a recent International Atomic Energy Agency (IAEA) publication (17). The IAEA report concludes that fast neutrons are superior to photons in the treatment of salivary gland tumors (locally extended, well differentiated), paranasal sinuses (adenocarcinoma, adenoid cystic carcinomas, and possibly other histologies), some tumors of the head and neck (locally extended, metastatic adenopathies), soft tissue sarcomas, osteosarcomas, chondrosarcomas (especially slowly growing/well differentiated), prostatic adenocarcinomas (locally extended), and melanomas (inoperable and recurrent). The IAEA report also identifies tumors for which conflicting or incomplete results have been reported and for which additional studies are necessary, these include inoperable pancreatic tumors, bladder carcinoma, recurrent and inoperable adenocarcinoma of the rectum, tumors of the esophagus, locally advanced tumors of the uterine cervix, and brain tumors for treatment with a neutron boost irradiation before or after X-ray therapy.

NEUTRON SOURCES FOR RADIATION THERAPY

Characteristics of Neutron Sources for Medical Use

In order to be useful for medical applications, fast neutron therapy facilities must meet a set of minimum requirements. As described above, in the early years of the 1970s many neutron therapy centers were set up to carry out clinical trials. Most of these centers made use of existing physics research accelerators (cyclotrons or proton linacs), which were adapted for clinical use. The trials produced results, which were often ambiguous and much of this ambiguity was ascribed to the inadequacies of the equipment. When the NCI in the United States decided to fund a number of therapy facilities, the basic specifications for these accelerators were defined to ensure that the equipment would be adequate to allow meaningful clinical trials to be completed (18); the key requirements are summarized in Table 2. All the facilities funded by the NCI were hospital-based and had rotational isocentric capability, that is, the neutron beam could be rotated around the patient in the treatment position.

Table 2. Summary of Some of the Key Requirements for a Hospital-Based Neutron Therapy Facility as Defined by the NCI[a]

Neutron beams having build-up and depth-dose characteristics equivalent to 4 MV X rays, penumbra not less sharp than ^{60}Co gamma-ray beams, and dose rates not less than 20 cGy/min.
Preferably an isocentric beam delivery system and as a minimum one horizontal and/or vertical beam delivery port.
Access to the neutron beam therapy facility for a minimum of 8 h/day, 4 days/week, 45 weeks/year for patient treatment and 16 h/week additional for physics and biology.
Methodology for providing a variety of square, rectangular, and irregularly shaped fields ranging in size from 5×5 cm^2 to 20×20 cm^2.
Capability to shape treatment fields using a variety of wedges and blocks so that any treatment field normally used for conventional X ray therapy can be reproduced on the neutron beam.

[a]See Ref. (18).

Several of these requirements depend critically on the neutron source. In particular, the attenuation (depth-dose) characteristics of the beam, the neutron dose rate (treatment time), and the reliability of the device may be dependent on the neutron source and the means of neutron production.

In order to fully characterize a neutron source it is necessary to have a detailed knowledge of a number of physical characteristics of the neutron beam. Most importantly, the physical data must be sufficient to allow for the accurate calculation of the physical dose delivered both to the tumor site and the surrounding normal tissues in the patient. In addition, physical data may also be necessary to adequately interpret the biological effects that are observed with high linear energy transfer (LET) beams. Table 3 lists the type of data, which are necessary to fully characterize the neutron source. These data are also necessary to fully assess the relative merits, usefulness, and

Table 3. Physical Data Necessary to Characterize Neutron Sources for Radiation Therapy

Total neutron yield or dose or kerma rate.
Neutron spectrum (i.e., Neutron yield as a function of neutron energy).
Neutron yield as a function of angle relative to the forward direction of the beam.
Neutron dose as a function of depth in a water phantom (i.e, depth-dose data).
Relative neutron dose as a function of the lateral postion (i.e, dose profiles).
The microdosimetric properties of the beam (i.e., the LET distribution of the secondary particles).
Neutron interaction data (cross-sections or kerma) for the interaction for the neutron beam with the constituent nuclei of tissue (C, N, O, H, Ca).

suitability of the various different neutron sources for radiation therapy applications. Much of the data in Table 3 are interconnected. In theory, a detailed knowledge of the neutron spectrum as a function of angle should be sufficient to calculate the other data provided there are comprehensive data on the nuclear cross-sections for the interaction of the neutrons with all the various biological target nuclei involved across the energy range of interest. Although, these data do exist for fast neutron beams they are not comprehensive and performing the necessary calculations with Monte Carlo codes remains a formidable and time consuming task. For this reason, until now it has proved simpler and more efficient to rely on various types of direct measurements to collect the necessary data; this situation may change in the future.

Basic Source Data

Neutron Yield. Neutron yields may be measured using a suitable detector that counts the number of particles arriving at the detector with no regard to the energy of the particle. Such detectors must usually be calibrated in order to determine their absolute efficiency. An excellent account of neutron detectors can be found in the work of Knoll (19). Neutron yields are generally expressed in terms of neutrons per steradian per microcoulomb of incident beam charge ($\mu C^{-1} \cdot sr^{-1}$).

Neutron Spectra. Neutron spectra are measured using a variety of neutron detectors (19,20). For a spectral measurement, the detector must exhibit an energy response in which the magnitude of the detector signal is proportional to the energy of the incident neutrons. Such detectors must also be calibrated so that correction can be made for their counting efficiency.

In a spectrum measurement, the neutron yield is measured as a function of energy ($\mu C^{-1} \cdot MeV^{-1} \cdot sr^{-1}$). The total neutron yield at a particular angle can be calculated from a spectral measurement by integrating over the neutron energy.

Neutron Dose. Neutron dose is most easily and accurately measured using the methods of mixed-field dosimetry (21). A tissue equivalent plastic ionization chamber is used to measure the total neutron plus gamma-ray dose and a Geiger–Muller (GM) tube is used as a neutron insensitive detector. From the two measurements, the neutron and gamma-ray dose can be determined separately. This does not mean that spectral and kerma data are not important. Indeed, the calculation of dose from an ionization chamber reading involves the use of factors, which rely on the exact nature of the neutron spectrum (e.g., kerma ratios and the energy to produce an ion pair). These factors are readily available in the literature in ICRU Report No. 46 (22). In practice, it is sufficient to measure the total dose only, since the percentage of gamma-ray dose is relatively small and its biological effectiveness is ~3 times less than an identical neutron dose.

Microdosimetric Measurements. Microdosimetric data are most commonly measured using a Rossi type A-150 tissue equivalent plastic proportional counter in which the sensitive volume of the counter is filled with a tissue equivalent proportional counter gas at reduced pressure (23). Typically, the internal diameter of these counters is 12.7 mm, and when filled to a pressure of 8.8 kPa with propane-based tissue equivalent gas, the counter simulates a sphere of solid tissue of diameter 2 μm. The counter detects the energy deposited in the gas as recoil particles traverse the gas volume after a neutron interaction has occurred in the plastic wall of the counter. From a knowledge of the counter geometry, it is possible to calculate the energy deposited per unit path length (keV/μm) (24). The microdosimetric spectrum is usually plotted as a single event spectrum in which the event size (y), in units of keV/μm, is plotted on a log scale as the ordinate and the differential dose distribution in event size (y) per unit logarithmic interval, $y \cdot d(y)$, is plotted on a linear scale as the abscissa. A typical microdosimetric spectrum plotted in this form is shown in Fig. 4. In this representation, the area under the curve represents the total dose. The various peaks in the curve can be interpreted as due to the different components of the dose, gamma-rays, recoil protons, alpha particles, and recoil heavy ions. Hence, such plots can be used to distinguish between neutron beams produced by different sources. A detailed account of microdosimetric methods can be found in ICRU Report No. 36 (24).

Beam Characteristics. From a practical radiation therapy physics point of view, the most useful data is that which allows you to calculate the neutron radiation dose distribution within the patient. A detailed knowledge of the neutron fluence and energy is not necessary to make these calculations. In radiation therapy measurements of absorbed dose in a water tank, $\sim 60 \times 60 \times 60$ cm^3 are made using a small volume (typically 0.3 cm^3) ionization chamber; the tank is known as a "water phantom". Neutron dose at a point in the phantom must be determined at a known depth, field size, and source-to-surface distance

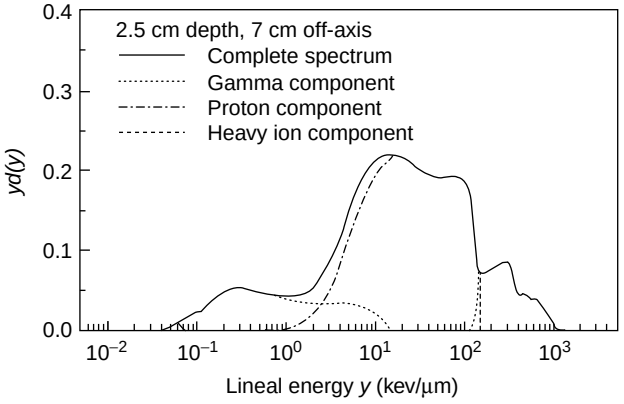


Figure 4. A typical microdosimetric event size spectrum measured with a tissue equivalent plastic Rossi proportional counter. The portions of the spectrum attributable to recoil electrons from gamma-ray interactions and recoil protons, alphas, and heavy ion recoils (C12, N14, and O16) from neutron interactions are identified. Reproduced with permission from Kota and Maughan, Med. Phys. 1996;23:1593.

Table 4. A Comprehensive List of All Neutron Therapy Centers

Facility	Reaction	Type of Machine	Z_{50}, cm	D_{max}, cm	Beam Type	Beam Shaping	First T_x	Status
Berkeley	d(16.0)Be	Cyclotron	8.8	0.2	Fixed	Inserts	1938	Closed
London	d(16.0)Be	Cyclotron	8.8	0.2	Fixed	Inserts/Jaws	1966	Closed
Dresden	d(13.5)Be	Cyclotron	7.9	0.2	Fixed	Inserts	1972	Closed
Houston MDAH	d(50)Be	Cyclotron	13.1	0.8	Fixed	Inserts	1972	Closed
Washington MANTA	d(35)Be	Cyclotron	11.7	0.5	Fixed	Inserts	1973	Closed
Univ. of Washington	d(21)Be	Cyclotron	9.4	0.2–0.3	Fixed	Inserts	1973	Closed
Chiba-Chi	d(30)Be	Cyclotron	10.8	0.5	Fixed	Multileaf	1975	Closed
Fermi Lab	p(66)Be	Proton Linac	16.6	1.6	Fixed	Inserts	1976	Open
Amsterdam	d(0.25)T	D–T	10.3	0.2–0.3	Rotational	Inserts	1976	Closed
Essen	d(14.3)Be	Cyclotron	8.1	0.2	Rotational	Inserts	1976	Open
Glasgow	d(0.25)T	D–T	10.3	0.2–0.3	Rotational	Inserts	1977	Closed
Manchester	d(0.25)T	D–T	10.5	0.2–0.3	Rotational	Inserts	1977	Closed
Heidelburg	d(0.25)T	D–T	10.6	0.3	Rotational	Inserts	1977	Closed
Hamburg	d(0.5)T	D–T	8.8	0.25	Rotational	Inserts	1977	Closed
Cleveland (GLANTA)	p(25)Be	Cyclotron	10.3	0.5	Fixed	Inserts	1977	Closed
Louvain-la-Neuve	p(65)Be	Cyclotron	17.5	1.8	Fixed	Multileaf	1978	Closed
Tokyo	d(14.0)Be	Cyclotron	8.3	0.2	Fixed	Inserts	1978	Closed
Krakow	d(12.5)Be	Cyclotron	7.7	0.2	Fixed	Inserts	1978	Closed
Edinburgh	d(16.0)Be	Cyclotron	8.7	0.2	Rotational	Inserts	1978	Closed
Chicago	d(8.0)D	Cyclotron	9.8	0.15	Fixed	Inserts	1981	Closed
Orleans	p(34)Be	Cyclotron	12.8	0.5	Fixed	Inserts	1981	Open
Cleveland (GLANTA)	p(42)Be	Cyclotron	13.5	2.2	Fixed	Inserts	1982	Closed
Houston (MDA)	p(42)Be	Cyclotron	14	1.2	Rotational	Inserts	1983	Closed
Riyadh	p(26)Be	Cyclotron	10.3	0.5	Rotational	Inserts	1984	Closed
Munster	d(0.25)T	D–T	10.5	0.3	Rotational	Inserts	1984	Closed
Univ. of Washington	p(50)Be	Cyclotron	14.8	1.2	Rotational	Multileaf	1984	Open
Univ. of Pennsylvania	d(0.25)T	D–T	10.3	0.2–0.3	Rotational	Inserts	1985	Closed
Clatterbridge	p(62)Be	Cyclotron	16.2	1.4	Rotational	Jaws	1986	Closed
Seoul	p(50)Be	Cyclotron	14.8	1.2	Rotational	Jaws	1986	Closed
UCLA	p(46)Be	Cyclotron	13.1	1.7	Rotational	Jaws	1986	Closed
Faure S.Africa	p(66)Be	Cyclotron	16.2	1.5	Rotational	Jaws/MLC trim	1988	Open
Detroit	d(48.5)Be	SC Cyclotron	13.6	0.9	Rotational	Multirod	1991	Open
Beijing	p(35)Be	Proton Linac	~13.0	~0.5	Fixed	Inserts	1991	Closed
Nice	p(65)Be	Cyclotron	17.5	1.8	Fixed	Multileaf	1993	Closed

(see the section: Neutron Dose). Measurements are also made along the radiation beam central axis to determine the attenuation of the beam. The parameter Z_{50} is a measure of the beam penetration as defined by the depth in a water phantom at which the neutron dose is reduced to 50% of its maximum value for a 10×10 cm^2 field. The Z_{50} value varies as a function of the energy and type of particle used in the primary beam. As the incident particle energy increases the mean neutron energy increases and the Z_{50} increases; this can be seen from the data in Table 4. This 50% depth-dose point also varies with the dimensions of the irradiation field (normally known as the field size). Some form of collimator is required to set the field size, this may be an attenuating block with a predefined rectangular opening, a pair of attenuating jaws, which provide a variable rectangular opening, or a multileaf collimator, which can define rectangular fields or more complex shapes. Generally, square fields are used for beam data measurements and the field size is defined at the center of the treatment volume (isocenter). The Z_{50} variation with field size arises because the dose at a point depends not only on attenuation of the primary beam in the water phantom, but also on the scattered component of the beam, which is field size dependent. To fully characterize a therapeutic beam it is, therefore, necessary to measure depth dose curves at a variety of field sizes; Fig. 5 illustrates this phenomenon.

Another beam parameter, which varies with the primary beam particle and energy is the beam build-up. When

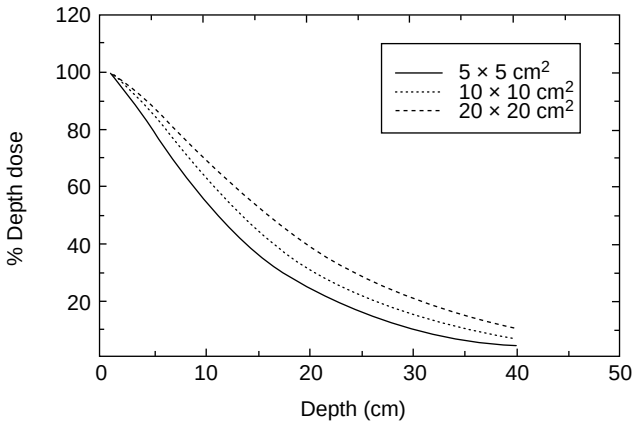


Figure 5. Depth-dose curves for d(48.5)Be neutron beam plotted as a function of field size. Reproduced with permission from Maughan and Yudelev, Med. Phys. 1995;22:1462.

Table 5. The Variation of surface dose and d_{max} as function of the neutron producing nuclear reaction and the energy of the incident particle[a]

Neutron Producing Reaction	Incident Particle Energy, MeV	Surface Dose as a % of the Dose at d_{max}	d_{max}, cm
d + Be	48.5	42	0.9
p + Be	50	38	1.2
p + Be	66	40	1.6

[a]Data taken from Ref. (21) and (25).

an indirectly ionizing radiation beam (neutrons or X rays) passes from air into a solid water medium, the secondary particle fluence in the surface layers is much less than at depth, because in the shallower layers the secondary particles are those that originated in the air and passed into the solid. Since air is much less dense than the solid, there are relatively few secondary particles in the surface layer. As the neutron beam penetrates the solid medium, more and more secondary particles are set in motion in the solid medium, until an equilibrium situation is reached at a depth that is about equal to the average range of the secondary particles in the solid. The energy deposition (dose) reaches a maximum at this depth (known as the depth of maximum dose, d_{max}) and is attenuated beyond this point as shown in Fig. 5. This build-up region of the curve cannot be measured using the instrumentation used to measure the attenuation data in Fig. 5. A specialized thin pill box shaped ionization chamber (known as an extrapolation ionization chamber) is required for these measurements. The build-up region has considerable clinical significance, when treating tumors at depths $>d_{max}$, since the dose in the surface layers of the skin is reduced relative to the tumor dose and, hence, the skin can be spared from excessive radiation damage. Table 5 shows how the surface dose (dose at zero depth), expressed as a percentage of the dose at d_{max}, and d_{max} vary as a function of the neutron producing nuclear reaction and the incident particle energy.

Beam profiles are also measured in the water phantom by scanning the ionization chamber in a direction perpendicular to the radiation beam central axis. A typical beam profile is shown in Fig. 6. As can be seen from this figure, the exact shape of the profiles depends on the depth in the phantom and the radiation field size. The most important feature of the profiles is the sharpness of the beam edges; This parameter degrades with both increase in field size and depth due to increased scattering of the neutrons. The exact sharpness of this penumbra region depends on many factors including, the source size, scattering from the collimator system and beam monitoring components in the beam path, the collimator geometry (i.e., whether the edges of the collimator jaws or leafs are divergent), and finally on neutron scattering in the patient (or phantom). Generally, phantom scatter is the predominating factor.

Depth-dose and profile date are inputted to the computer programs used for calculating the dose distribution in patients. These programs use a variety of different mathematical algorithms to calculate the dose distributions in the patient.

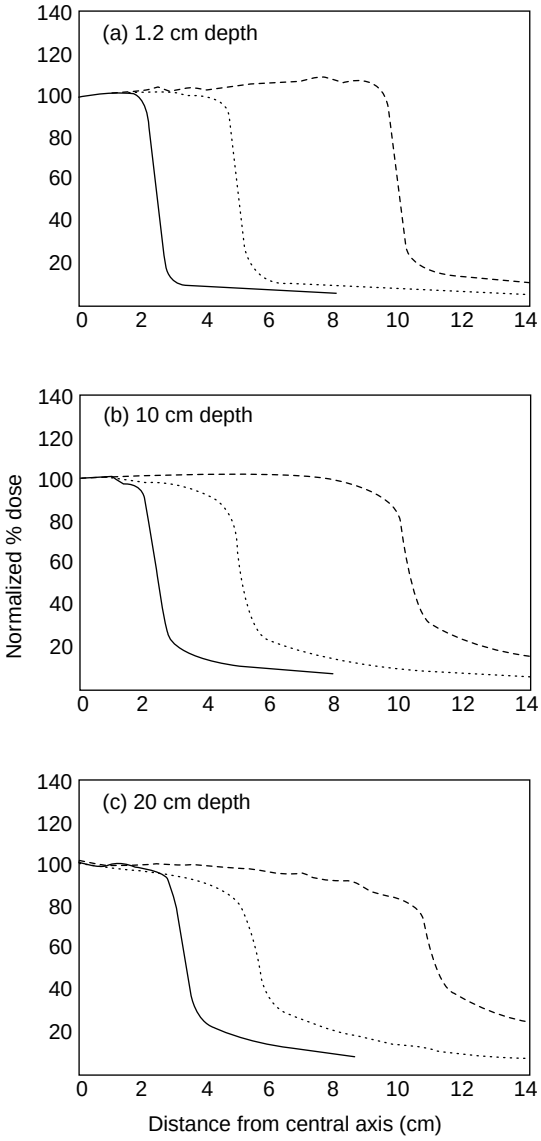


Figure 6. Beam profiles for field sizes of 5×5 cm^2 (solid line), 10×10 cm^2 (dotted line), and 25×25 cm^2 (dashed line) at depths of (*a*) 1.2, (*b*) 10, and (*c*) 20 cm in a water phantom. Reproduced with permission from Maughan and Yudelev, Med. Phys. 1995; 22: 1461.

NEUTRON PRODUCTION

Fast Neutron Therapy Beams

Over the past 65 years, there have been at least 34 centers that have been involved in fast neutron radiation therapy. Table 4 lists these centers and indicates which nuclear reaction has been used to produce the neutron beam. In column 2, the lower case letter indicates the accelerated particle, the number in parenthesis is the energy in megaelectronvolts of this projectile and the final letter(s) represents the target nucleus. Of the facilities, 12 have used the deuteron stripping reaction with a beryllium target, 14 the proton inelastic scattering reaction on a beryllium target, 7 the deuteron–tritium fusion reaction and only 1, the deuteron–deuterium fusion reaction. The relative

merits of these various modes of neutron production will be discussed.

The d-Be Reaction

In practice, the deuteron stripping reaction on beryllium is the most prolific neutron producing reaction since a solid beryllium target capable of stopping the full energy of the beam can easily be constructed. There are basic physical reasons for the deuteron stripping reaction on beryllium being more prolific in producing neutrons than the inelastic scattering of protons from a beryllium target. The deuteron is a loosely bound structure of a neutron and a proton in which the two particles spend most of their time at greater distances from each other than the range of the forces between them. Hence, when an energetic deuteron approaches a beryllium target nucleus it is possible for the proton to be absorbed into the target nucleus, breaking free from the neutron that carries on following its original path at its original velocity (i.e., with half the kinetic energy of the original deuteron). The excited ^{10}B nucleus formed may also decay by neutron emission. When a proton beam interacts with a beryllium target the proton is absorbed into the target nucleus to form a compound nucleus, ^{10}B, in an excited state, which may decay by emitting a neutron. Hence, the stripping reaction is a much more prolific source of neutrons, since many neutrons originate from the break-up of the deuteron. In addition, the stripping reaction is very forward-peaked in the laboratory, while the (p,n) reaction on beryllium produces a more isotopic distribution, since it involves the formation of a compound nucleus. The theoretical aspects of the production of intense neutron beams using the deuteron stripping reaction with beryllium targets has been discussed by August et al. (26). Experimental neutron yield data and spectral data on the characteristics of neutrons from beryllium targets bombarded with protons and deuterons with energies of 16, 33, and 50 MeV are available in the work of Meulders et al. (27). The above experiments measure the total neutron fluence at 0° or the differential fluence as a function of neutron energy at various angles relative to the forward direction. In order to estimate the usefulness of a given reaction as a neutron source for radiation therapy, it is necessary to know the neutron dose rate produced in practice by a given attainable beam current. Such information can be calculated from neutron spectrum data using neutron kerma factors for water or body tissue (22).

Another vitally important parameter is the penetration of the neutron beam in tissue. Although in principle it is possible to calculate this information from the spectral and kerma data, in practice there are insufficient data and the calculations are difficult. Therefore, ionization chamber measurements are often more convenient for measuring both the dose rate and penetration of neutron beams. There is extensive data on neutron dose rates and depth dose characteristics of neutron beams across a wide range of energies for the deuteron stripping reaction on beryllium. Smathers et al. (28) reviewed the available dose rate data in the incident deuteron energy range of 11–50 MeV and concluded that the tissue kerma measured free-in-air at a target-to-detector distance of 1.25 m and for a 5×5 cm^2 field size could be fitted by an equation of the form

$$1\mathrm{n}K = 1\mathrm{n}\ a + b\ \ln E \tag{1}$$

In this equation, E is the energy of the incident beam in million electronvolts, K is the tissue kerma in units of cGy min$^{-1}\cdot\mu$A^{-1} and a and b are constants with numerical values of 1.356×10^{-4} and 2.97, respectively.

Later, Wootton (29) reviewing ionization chamber dose rate data for the d-Be reaction quoted the following expression for the dose rate at 1.25 m

$$D \cdot Q^{-1} = 2.49 \times 10^{-2}\ E_{\mathrm{d}}^{2.95} \tag{2}$$

where $D \cdot Q^{-1}$ (in Gy/C) is the absorbed dose to tissue free-in-air at a 1.25 m target to detector distance per unit charge of deuteron beam, and E_{d} is the incident deuteron energy in million electronvolts. If Eq. 1 is rewritten in this form and normalized to the same units it becomes

$$D \cdot Q^{-1} = 2.26 \times 10^{-2}\ E_{\mathrm{d}}^{2.95} \tag{3}$$

At $E_{\mathrm{d}} = 50$ and 10 MeV, these equations agree to within 2 and 5%, respectively. Of course, such equations can only be used to give a rough estimate of the neutron dose output of a neutron therapy device, since the exact output depends on the details of the target, flattening filter, collimator, and dose monitor design.

The spectral data of Lone et al. (30) gives information on the fluence averaged energy of the neutron beam (E_{n}), they derived the following expression:

$$E_{\mathrm{n}} = 0.4\ E_{\mathrm{d}} - 0.3 \tag{4}$$

relating the mean neutron energy to the incident deuteron energy (E_{d}) in million electronvolts.

Data on the penetration of neutron beams, produced by the d-Be reaction, have been published by Shaw and Kacperek (31). In Fig. 7, the values of Z_{50} from Table 4 for the d-Be reaction is plotted as a function of incident deuteron energy. A power law fit to the curve yields the

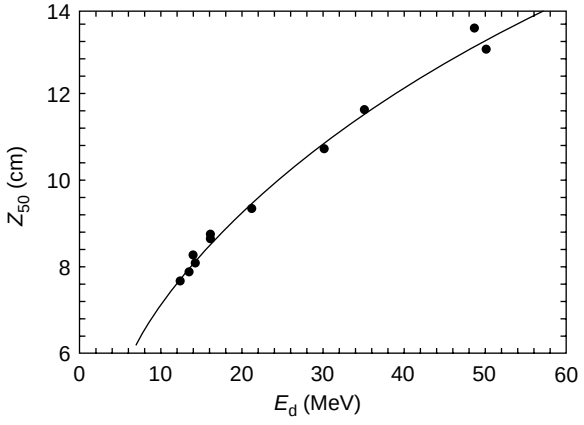


Figure 7. The neutron beam 50% depth-dose value (Z_{50}) for neutrons produced by the d-Be reaction plotted as a function of the incident deuteron energy (E_{d}). The solid line is a power law fit to the data (Eq. 5). The data are for a 10×10 cm^2 field at a range of source-surface distances (SSD) between 91 and 183 cm.

following equation

$$Z_{50} = 2.90\ E_{\mathrm{d}}^{0.39} \tag{5}$$

The exact value of Z_{50} depends on the target structure and the beam filtering effects of the flattening filter and the dose monitor devices.

The clinical relevance of microdosimetric data in neutron therapy planning has been discussed by Pihet et al. (32). The microdosimetric dose distribution shown in Fig. 4 can be analyzed in several ways, the most useful single parameter, which can be used to describe the distribution is the dose mean lineal energy corrected for saturation

$$y_1^* = \int y_{\mathrm{sat}} \cdot d_1(y) \cdot dy \tag{6}$$

This parameter was defined in the dual radiation action theory of Kellerer and Rossi (33). The function y_{sat} is a response function that accounts for the saturation effect that is observed in mammalian cell systems (34); as the LET of the beam is increased the observed RBE decreases due to the overkill effect (35). In Fig. 8, the $y_1{}^*$ values for a variety of different therapy beams are plotted as a function of the mean neutron energy, as defined in Eqs. (4) and (8). The closed circles represent the data for the d-Be reaction.

The p-Be Reaction

From Table 4, it can be seen that most of the early neutron therapy centers (i.e., those operating at lower energies) utilized the deuteron stripping reaction or the fusion reaction as the source of neutrons (see Section: The d-T Reaction). Interest in the p-Be reaction increased when the importance of constructing neutron sources with rotational isocentric capability (i.e., capable of rotating around the patient with the tumor center on the axis of rotation) and with penetration equivalent to 4 MV photon beams was realized (Table 2). Good penetrability requires deuteron or proton beams with energies of 40–50 MeV or greater, and isocentricity requires bending magnet systems capable of bending these beams 180° (a 45° bend followed by a 135° bend). Conventional cyclotrons capable of producing 50 MeV deuterons were too large and expensive as were the magnet systems for bending these beams. Proton cyclotrons of 50 or 60 MeV offered a much less expensive alternative in the late 1970s, when the decision to install a new generation of hospital-based neutron therapy facilities was being made by the NCI in the United States. The problems of switching from the deuteron stripping to the p-Be reaction were soon recognized: the energy spectrum from the reaction of protons on a beryllium target has a significant low-energy tail, which reduces the average neutron beam energy and spoils the penetration. Also the neutron output is much less, therefore, higher beam currents are required with an increase in the problems associated with target cooling and target activation. The penetration problem can be overcome by using nonstopping targets (i.e., beryllium targets in which the incident proton beam does not lose all its energy) in conjunction with polyethylene filters, which filter out the low energy component of the beam. These techniques have been discussed in detail for proton beams with energies between 30 and 60 MeV by Bewley et al. (36) and for a 41 MeV proton beam by Smathers et al. (37). The absorbed dose rate ($D \cdot Q^{-1}$) to tissue at 1.25 m from the target is given by Wootton (29) as

$$D \cdot Q^{-1} = 2.44 \times 10^{-2}\ E_{\mathrm{p}}^{2.37} \tag{7}$$

where $D \cdot Q^{-1}$ is in units of Gy/C and E_{p} is the incident proton energy in million electronvolts.

For the p + Be reaction with a stopping target, the average neutron energy for neutrons with energies >2 MeV (E_{n}) measured at 0° to the incident beam is given by

$$E_{\mathrm{n}} = 0.47\ E_{\mathrm{p}} - 2.2 \tag{8}$$

where E_{p} is the incident proton energy (29).

In Fig. 9, the value of Z_{50} from Table 4 for the p-Be reaction is plotted as a function of incident proton energy, the solid curve is a power law fit to the data that gives the

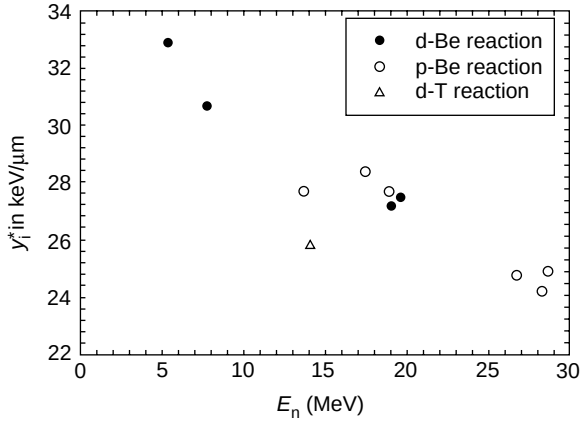


Figure 8. The microdosimetric parameter mean lineal energy corrected for saturation ($y_1{}^*$) plotted as a function of the mean energy of the neutron beam for various neutron producing reactions: d-Be, open circles (E_{n} from Eq. 4); p-Be, open circles (E_{n} form Eq. 8), and d-T reaction, open triangle (E_{n} = 14.1 MeV).

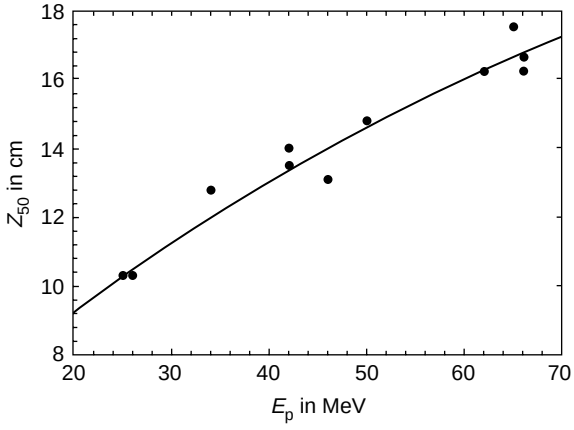


Figure 9. The neutron beam 50% depth-dose value (Z_{50}) for neutrons produced by the p-Be reaction plotted as a function of the incident proton energy (E_{p}). The solid line is a power law fit to the data (Eq. 9). The data are for a 10 × 10 cm^2 field for range of SSD between 125 and 190 cm.

following equation:

$$Z_{50} = 2.06\ E_{\mathrm{p}}^{0.50} \tag{9}$$

The greater spread in the data points, when compared with the similar data plotted in Fig. 7, is a result of the greater variety in the target design (i.e., target thickness and filtration conditions) used at the different facilities.

The microdosimetric data for the p-Be reaction is represented by the open circular data points in Fig. 8. The dose mean lineal energy corrected for saturation correlates with the mean neutron energy for both the p-Be and the d-Be produced neutron beams.

The d-D Reaction

This reaction was used in the neutron radiation therapy facility at the University of Chicago, where a deuteron beam of energy 8.3 MeV was incident on a thick cryogenic deuterium gas target designed by Kuchnir et al. (38). Two reactions predominate when a deuterium target is bombarded with deuterons:

$$\mathrm{d} + \mathrm{D} = {}^{3}\mathrm{He} + \mathrm{n} \qquad Q = +3.27\ \mathrm{MeV}$$

and,

$$\mathrm{d} + \mathrm{D} = \mathrm{d} + \mathrm{n} + \mathrm{p} \qquad Q = -2.22\ \mathrm{MeV}$$

Hence, there are two distinct groups of neutrons produced, the higher energy group resulting from the first of these two reactions. The neutron energy spectrum for bombardment of a thick stopping target exhibits two maxima corresponding to the two groups. The relative magnitude of the two peaks depends on the incident deuteron energy. At an incident deuteron energy of 6.8 MeV the higher energy peak due to the D(d,n)^{3}He reaction predominates, but for an incident energy of 11.1 MeV, the two peaks are comparable (39). Waterman et al. (39) calculated the neutron spectra at 6.8, 8.9, and 11.1 MeV from a knowledge of the mass stopping power of deuterons in deuterium and from the cross-sections of the two reactions as given by Schraube et al. (40).

The dosimetric properties of the d–D neutron beam are summarized in the work of Kuchnir et al. (38). Figure 10 shows the variation in absorbed tissue dose rate (Gy/μC) as a function of the incident deuteron energy for a thick deuterium gas target. The measurements were made at a SSD of 126 cm with a 11.1 × 11.1 cm^2 field size. The data can be fitted by a power law expression.

$$D \cdot Q^{-1} = 2.41 \times 10^{-2}\ E_{\mathrm{d}}^{3.28} \tag{10}$$

where $D \cdot Q^{-1}$ (Gy/C) is the absorbed dose to tissue measured free-in-air per coulomb (C) of incident beam current, and E_{d} is the incident deuteron beam energy. Measurements have been made by Weaver et al. (41) at an incident deuteron energy of 21 MeV, but with a transmission gas target. For a target filled to a pressure of 3.33 MPa (33 atm), equivalent to an energy loss of ~3.5 MeV, the measured dose was 2.25 × 10^{-4} Gy/μC for a 10 × 10-cm field at 1.25 m SSD.

In practice, the University of Chicago neutron therapy facility produced a maximum dose rate of 0.12 Gy/min at an SSD of 91 cm, for a 10 × 10 cm field size. The penetration of the University of Chicago d–D beams in terms Z_{50} is 9.8 cm (Table 4). An interesting feature of the d–D reaction is that as the incident deuteron energy is increased the mean neutron energy produced remains practically constant. This finding is evident in the depth-dose data of Kuchnir et al. (38), where changing the incident deuteron energy has no significant effect on the value of Z_{50}. Even at 21 MeV with a transmission target the Z_{50} remains constant at ~10 cm. Thus, the inherent poor penetration of neutron beams produced by the d–D reaction, combined with the difficulties of producing a cryogenic deuterium gas target discouraged the use of this reaction as a neutron source for radiation therapy.

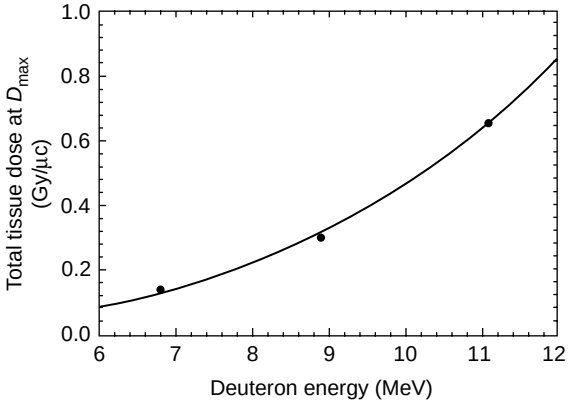


Figure 10. Absorbed dose at the depth of maximum dose as a function of incident deuteron energy for the d–D reaction. The data are from Kuchnir et al.(38). The solid curve is a power law fit to the data (Eq. 10). Measurements were made in a 11.1 × 11.1 cm^2 field at an SSD of 126 cm.

The d–T Reaction

For many years, this reaction was seen as the ideal reaction for producing a relatively inexpensive source of neutrons for radiation therapy. The large positive Q-value for the reaction

$$\mathrm{d} + {}^{3}\mathrm{H} \longrightarrow {}^{4}\mathrm{He} + \mathrm{n} \qquad Q + 16.6\ \mathrm{MeV}$$

results in monoenergetic neutrons of energy ~14 MeV. In principle, a relatively modest deuteron energy of 250–500 keV should be sufficient to produce an intense source of 14-MeV neutrons if sufficient beam current can be obtained. The original intention was to produce the source and target assembly in the form of a sealed tube, which could be easily replaced in the treatment head and would have a lifetime of 1000 h or more. Such a unit would have been similar in this respect to the 250-kVp X-ray units that were in widespread use before the advent of ^{60}Co units and high-energy electron linacs in conventional photon radiation therapy. Initially, the main problem with these devices was that associated with producing a target in the sealed tube configuration that would provide sufficient neutron dose rate. However, many different systems were used in attempts to produce a practical d–T generator and these have been reviewed in detail in ICRU Report No. 45 (21).

Of the five types of commercially available d–T generators, which were used in clinical trials to treat significant numbers of patients, four were of a type that employed some form of sealed tube in which a mixed deuterium–tritium beam was accelerated to an energy of 200–250 keV and used to bombard a tritiated rare earth target (titanium, erbium, or scandium). The characteristics of these four machines are given in Table 4. The Haefely device produced the highest dose rate with the longest average tube life of ~300 h and was installed in Heidelberg and Münster. The operation of the Philips and Elliot tubes are described by Broerse et al. (42). A Philips machine was installed in Amsterdam and the Elliot devices were used in Glasgow and Manchester. An account of the construction of the Haefely machine is given by Schmidt and Rheinhold (43) while a detailed appraisal of its clinical operation can be found in the work of Höver et al. (44). The University of Pennsylvania D–T generator was built by the Cyclotron Corporation (Berkeley, CA).

The fifth commercial unit, installed in Hamburg, was produced as a collaboration between AEG in Germany and Radiation Dynamics Inc. (RCI) in the United States. The machine used a pure deuterium beam accelerated to 500-keV incident on a replaceable rotating tritiated titanium target (45). The source and target design were improved by incorporating an analyzed deuterium beam (to remove molecular D_2^+ beam components) and a larger target (46). With these improvements a dose rate of 0.12–0.13 Gy/min was achieved.

PRACTICAL FAST NEUTRON THERAPY FACILITIES

In fast neutron radiation therapy the need for state-of-the-art neutron facilities, which allow neutron treatments to be delivered with precision and sophistication equivalent to that used in modern conventional X-ray therapy, is well recognized. Modern trends in X-ray therapy are toward conformal therapy with multiple static fields, multileaf collimators, three-dimensional (3D) treatment planning and most recently (IMRT). All these tools must be available for neutron radiation therapy if effective randomized phase III clinical trials are to be completed to compare the two modalities.

An important aspect of this problem is beam penetration. The problem with neutron beams is that it is not possible to increase the mean energy of the neutrons to a point at which the neutron beams have percentage depth-dose characteristics that are equivalent to modern high energy (15–25 MV) photon beams, since as the neutron beam energy increases, the average LET of the beam decreases. If the average LET is decreased too far, the radiobiological advantage of the neutron beam will be significantly diluted (e.g., RBE tends to decrease and neutron beam advantages associated with hypoxia decrease, the radiosensitivity variation within the cell cycle tends to that of low LET radiations). Hence, there is a trade-off between beam penetration and LET effect. This trade-off can be seen in Fig. 8, which illustrates how the effective LET (y_1^*) of the neutron beam decreases as the mean neutron energy (E_n) increases.

The requirement that neutron beams should be at least equivalent to 4 MV photon beams (Table 2) arises in part from this trade-off. Of the 34 facilities listed in this Table 4, only 10 satisfied this penetration requirement. Of the 6 operational facilities 4 satisfy the requirement and the most penetrating beams at the Ithemba Laboratory in South Africa and Fermi Laboratory in the United States, produced by the p(66)Be reaction, are equivalent to an 8-MV photon beam. If all the requirements of Table 1 are considered and in addition a multileaf or multirod collimator for producing irregularly shaped fields is made mandatory, then only three of the operational facilities meet all the requirements. These are at the University of Washington in Seattle, the Ithemba Laboratorty in South Africa and at Harper Hospital, Wayne State University in Detroit. The fact that the neutron beams are less penetrating than the 15–25 MV photon beams that are commonly used for treating deep-seated tumors may not be a problem. In a treatment planning comparison of 3D conformal neutron and photon radiotherapy for locally advanced adenocarcinoma of the prostate, Forman et al. (47) showed that the dose–volume histograms for gross tumor, rectal, and bladder volumes treated with neutrons and photon beams are not significantly different. Wootton (29) suggested that neutron beams with a Z_{50} of >15 cm are required, and that for the d–Be reaction to be useful in this case, an incident deuteron energy of 61 MeV would be required. Forman's data, however, indicate that a Z_{50} of 13.6 cm is adequate for producing acceptable dose distributions for the treatment of pelvic tumors.

In the late 1970s, economic considerations led to the choice of the p–Be reaction as the neutron source for a new generation of hospital-based high-energy proton cyclotrons for clinical trials in the United States, because deuteron producing conventional cyclotrons and the associated bending magnet system required to produce rotational beams were too costly. These machines were installed at the MD Anderson (MDA) Hospital in Houston, at the University of California Los Angeles, and at the University of Washington in Seattle (Table 4). Since this time the development of a compact superconducting deuteron cyclotron for neutron radiation therapy by Henry Blosser and his associates at the National Superconducting Cyclotron Laboratory at Michigan State University has had a significant impact on the technology of neutron therapy. This superconducting facility (25,48) has many innovative features. The accelerator weighs ~25 Mg (25 tons), ~10 times less than a conventional 50 MeV deuteron cyclotron. The unit has an internal beryllium target and is mounted between two large rings (4.3 m outer diameter) in order to provide for 360° rotation around the treatment couch. A 25 Mg counterweight mounted on the rings acts as a primary beam stop, which reduces the required thickness of the shielding walls. The total rotating mass is ~60 Mg (60 tons). Figure 11 is a schematic of the cyclotron and gantry. Figure 12 shows a section through the median plane of the cyclotron indicating its' main components. The unit does not require a separate bending magnet system to produce an isocentric beam and it can be installed in a single shielded room. With no beam extraction or elaborate bending magnet system, the operation is

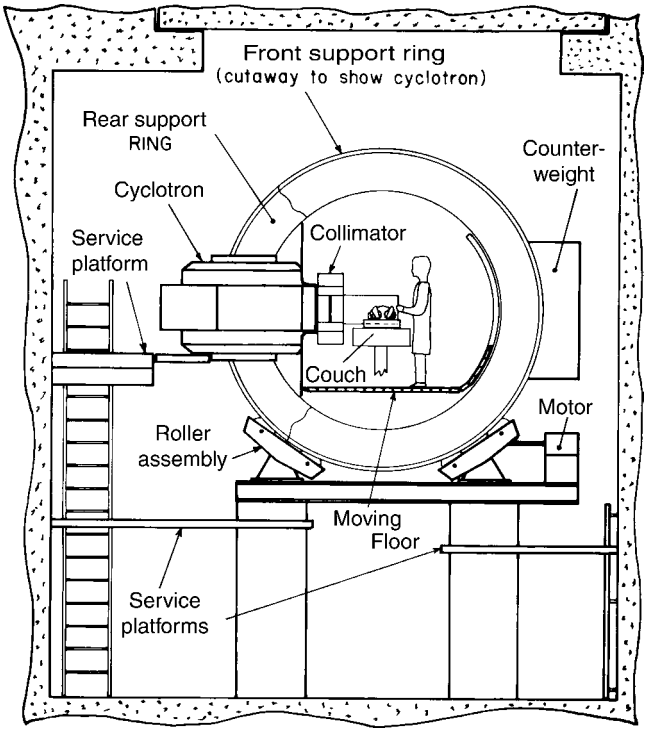


Figure 11. A schematic of the superconducting cyclotron mounted on the rotating gantry at the Wayne State University Facility. Reproduced with permission from Maughan et al., Med. Phys. 1994;21:781.

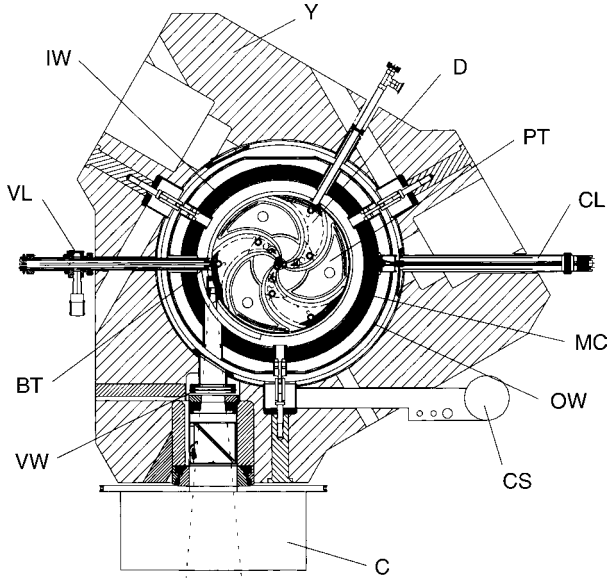


Figure 12. Schematic section through the median plane of the superconducting cyclotron, showing the following features Y = magnet yoke, MC = magnet superconducting coil, PT = magnet hill pole tip, IW = cryostat inner wall, OW = cryostat outer wall, CL = magnet coil electrical leads, CS = cryogen supply and gas return lines, D = radio frequency system dees, BT = internal beryllium target, VL = target vacuum lock, VW = beam chamber vacuum window, and C = neutron beam collimator. Reproduced with permission from Maughan et al., Med. Phys. 1994;21:780.

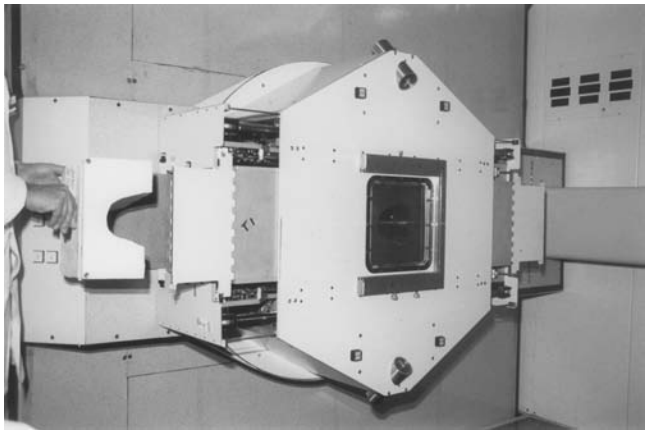


Figure 13. A photograph of the multirod collimator. One-half of the polystyrene foam form used to push the rod array into the desired shape is visible on the left. Reproduced with permission from Maughan et al., Med. Phys. 1994;21:781.

considerably simplified. The unit also incorporates a unique multirod collimator for producing irregularly shaped fields (49), which conform accurately to the tumor volume (Fig. 13). This facility has been in routine clinical use since March of 1992, and up to the end of 2002, ~1800 patients have been treated. Close to 10,000 individual treatment fields have been routinely treated in a single year making this the busiest and most efficient neutron therapy facility in the world.

DISCUSSION AND CONCLUSIONS

Neutron therapy has been demonstrated to be superior to conventional therapy in the treatment of salivary gland tumors, some tumors of the paranasal sinuses and other head and neck sites, soft tissue sarcomas, chondrosarcomas, osteosarcomas, advanced adenocarcinoma of the prostate, and inoperable and recurrent melanoma (17). For a range of other sites, further investigation is necessary to establish the efficacy of neutron therapy; these sites include pancreas, bladder, rectum, esophagus, uterine cervix, and brain.

However, in spite of these successes, neutron therapy appears to be in decline with only six centers actively treating patients (three in the United States and one each in Germany, France, and South Africa). The emphasis on precision radiation therapy has resulted in the development of intensity modulated radiation therapy techniques in conventional X ray therapy. These techniques allow for highly conformal dose delivery, maximizing the dose to the tumor volume and minimizing the dose to the surrounding normal tissues. There is also a considerable increase in the number of proton beam therapy centers, using the unique energy deposition patterns associated with proton beams to achieve even greater conformality than is achievable with IMRT.

In Europe and Asia, there is interest in developing ^{12}C ion beams for radiation therapy. These developments are spurred by the superior results achieved with neutron therapy in the cases outlined above. Heavy ion beams,

such as ^{12}C beams, are high LET beams, which combine the biological advantages of neutrons with the dose distribution advantages of protons. Such beams are extremely expensive to produce. The application of intensity modulated radiation therapy techniques in neutron therapy (IMNT) could improve the conformality of neutron therapy. Compact superconducting cyclotrons with computer controlled multileaf collimators, which allow IMNT to be delivered, could be an attractive and less expensive alternative to ^{12}C therapy.

The superconducting technology could be applied to designing a compact 60–70 MeV gantry mounted proton cyclotron to provide a beam with better depth dose characteristics than the existing Wayne State University cyclotron. The possibility of building a compact conventional 50 MeV proton cyclotron in a similar configuration to the superconducting deuteron cyclotron has been suggested (Jongen, unpublished data). A computer controlled MLC is at present under construction at Wayne State University with the intention of using it to implement IMNT (50). Such advances in neutron therapy technology are important if it is to achieve its full potential and remain competitive with the other radiation therapy modalities (i.e., conventional X rays and electrons, protons, and heavy ions).

In the >80 years since its discovery in Cambridge by James Chadwick, the neutron has found an important place in radiation therapy research, and much has been done to improve the means of neutron production and delivery.

BIBLIOGRAPHY

1. Chadwick J. The existence of a neutron. Proc R Soc London, Ser A 1932;136:629–708.
2. Cockroft JD, Walton ETS. Experiments with high velocity positive ions. I.-The disintegration of elements by high velocity protons. Proc R Soc London, Ser A 1932;137:229–242.
3. Gray LH, Read J, Wyatt JG. A neutron generator for biological research. Br J Radiol 1940;147:82–94.
4. Stone RS. Neutron therapy and specific ionizations. Am J Roentgenol 1948;59:771–779.
5. McNally NJ, Maughan RL, de Ronde J, Roper MJ. Measurements of some radiobiological and physical properties of neutrons produced by a 4-MV Van de Graaff accelerator. Rad Res 1982;89:227–237.
6. Sheline GE, Phillips TL, Field SB, Brennan JT, Raventos A. Effects of fast neutrons on human skin. Am J Roentgenol 1971;111:31–41.
7. Thomlinson RH, Gray LH. The histological structure of some human lung cancers and possible implications for radiotherapy. Br J Cancer 1955: 9:539–549.
8. Howard A, Pelc SR. Synthesis of deoxyribonucleic acid in normal and irradiated cells and its relation to chromosome breakage. Heredity 1953;6 (suppl): 261–273.
9. Sinclair WK, Morton RA. X-ray sensitivity during the cell generation cycle of cultured Chinese hamster cells. Rad Res 1966;29:450–474.
10. Withers HR, Mason K, Reid BO. Response of mouse intestine to neutrons and gamma rays on relation to dose fractionation and division cycle. Cancer 1974;34:39–47.
11. Battermann JJ, Breur K, Hart GAM, VanPeperzeal HA. Observations on pulmonary metastases in patients after single doses and multiple fractions of fast neutrons and cobalt-60 gamma rays. Eur J Cancer 1981;17:539–548.
12. Griffin TW, Pajak TF, Laramore GE, Duncan W, Richter MP, Hendrickson FR. Neutron vs photon irradiation of inoperable salivary gland tumors: results of an RTOG-MRC cooperative study. Int J Rad Oncol Biol Phys 1988;15:1085–1090.
13. Russell KJ, Caplan RJ, Laramore GE, Burnison CM, Maor MH, Taylor ME, Zink S, Davis LW, Griffin TW. Photon versus fast neutron external beam radiotherapy in the treatment of locally advanced prostate cancer: results of a randomized trial. Int J Radiat Oncol Biol Phys 1994;28:47–54.
14. Laramore GE, Krall JM, Thomas FJ, Russell KJ, Maor MH, Hendrickson FR, Martz KL, Griffin TW. Fast neutron radiotherapy for locally advanced prostate cancer: final report of an RTOG randomized clinical trial. Am J Clin Oncol 1993;16: 164–167.
15. Forman JD, Yudelev M, Bolton S, Tekyi-Mensch S, Maughan R. Fast neutron irradiation for prostate cancer. Cancer Metastasis Rev 2002;12:131–135.
16. Wambersie A, Auberger T, Gahbauer RA, Jones DTL, Potter R. A challenge for high-precision radiation therapy: the case for hadrons. Strahlenther Onkolog 1999;175 (Suppl II): 122–128.
17. IAEA Report TECDOC-992, Nuclear data for neutron therapy: Status and future needs, Vienna: International Atomic Energy Agency, 1997.
18. National Cancer Institute Request for Proposal NCI-CM-97282. Clinical Neutron Therapy Program 1979.
19. Knoll GF. Radiation Detection and Measurement. 2nd ed. New York: Wiley; 1989.
20. Cross WG, Ing H. Neutron spectroscopy. In: Kase KR, Bjärngard BE, Attix FH, editors. The Dosimetry of Ionizing Radiation. Volume 2, Orlando: Academic Press; 1987. pp. 91–167.
21. ICRU Report No. 45, Clinical neutron dosimetry part I: determination of absorbed dose in a patient treated with external beams of fast neutrons, Bethesda, International Commission on Radiation Units and Measurements, 1989.
22. ICRU Report No. 46, Photon, electron, proton and neutron interaction data for body tissues, Bethesda, International Commission on Radiation Units and Measurements, 1992.
23. Rossi HH, Rosensweig W. A device for the measurement of dose as a function of specific ionization. Radiology 1955;64: 404–441.
24. ICRU Report No. 36, Microdosimetry. Bethesda, International Commission on Radiation Units and Measurements, 1983.
25. Maughan RL, Yudelev M. Physical characteristics of a clinical d(48.5) + Be neutron therapy beam produced by a superconducting cyclotron. Med Phys 1995;22:1459–1465.
26. August LS, Theus RB, Shapiro R. Stripping theory analysis of thick target neutron production for D + Be. Phys Med Biol 1976;21:931–940.
27. Meulders J-P, Leleux P, Macq PC, Pirart C. Fast neutron yields and spectra from targets of varying atomic numbers bombarded with deuterons from 16 to 50 MeV. Phys Med Biol 1975;20:235–243.
28. Smathers JB, Otte VA, Smith AR, Almond PR. Fast neutron dose rate vs. energy for the d-Be reaction—a reanalysis. Med Phys 1976;3:45–47.
29. Wootton P. Neutron therapy facilities and their specification. Radiat Protection Dosim 1988;23:349–355.
30. Lone MA, Ferguson AJ, Robertson BC. Characteristics of neutrons from Be targets bombarded with protons, deuterons and alpha particles. Nucl Instrum Methods 1981;189:515–525.
31. Shaw JE, Kacperek A. Fast neutron beams, in Central Axis Depth Dose Data for Use in Radiotherapy: 1996. Br J Radiol 1996; (Suppl 17): 97–104.
32. Pihet P, Gueulette J, Menzel HG, Grillmaier RE, Wambersie A. Use of microdosimetric data of clinical relevance in neutron therapy planning. Radiat Protection Dosim 1988;23:471–474.
33. Kellerer AM, Rossi HH. The theory of dual radiation action. Curr Top Radiat Res Q 1972;8:85–158.

34. Barendsen GW. Responses of cultured cells, tumors and normal tissues to radiations of different linear energy transfer. Curr Top Radiat Res Q 1968;4:332–356.
35. Hall EJ. Radiobiology for the Radiologist. 3rd ed. Philadelphia: J.B. Lippincott; 1988. p 170.
36. Bewley DK, Meulders JP, Page BC. New neutron sources for radiotherapy. Phys Med Biol 1984;29:341–349.
37. Smathers JB, Graves RG, Earls L, Otte VA, Almond PR. Modification of the 50% maximum dose depth for 41 MeV (p^+, Be) neutrons by use of filtration and/or transmission targets. Med Phys 1982;9:856–859.
38. Kuchnir FT, Waterman FM, Skaggs LS. A cryogenic deuterium gas target for production of a neutron therapy beam with a small cyclotron. In: Burger G, Ebert EG, editors. Proceedings of the Third Symposium on Neutron Dosimetry. Luxembourg: Commission of the European Communities; 1978. pp. 369–378.
39. Waterman FM, Kuchnir FT, Skaggs LS, Bewley DK. Page BC, Attixs FH. The use of B10 to enhance the tumor dose in fast-neutron therapy. Phys Med Biol 1978;23:592–602.
40. Schraube H, Morhart A, Grünauer F. Neutron and gamma radiation field of a deuterium gas target at a compact cyclotron. In: Burger G, Ebert HG, editors. Proceedings of the Second Symposium of Neutron Dosimetry in Biology and Medicine, EUR 5273. Luxembourg: Commission of the European Communities; 1975. pp. 979–1003.
41. Weaver KA, Eenmaa J, Bichsel H, Wootton P. Dosimetric properties of neutrons from 21 MeV deuteron bombardment of a deuterium gas target. Med Phys 1979;6:193–196.
42. Broerse JJ, Greene D, Lawson RC, Mijnheer BJ. Operational characteristics of two types of sealed-tube fast neutrons radiotherapy installments. Int J Radiat Oncol Biol Phys 1977;3: 361–365.
43. Schmidt KA, Rheinhold G. The Haefely–GFK fast neutron generator. Int J Radiat Oncol Biol Phys 1977;3:373–376.
44. Höver KH, Lorenz WJ, Maier-Borst W. Experience with the fast neutron therapy facility KARIN under clinical conditions. In: Burger G, Ebert HG, editors. Proceedings of the Fourth Symposium on Neutron Dosimetry. EUR 7448EN. Volume II, Luxembourg: Commission of the European Communities; 1981. pp. 31–37.
45. Offerman BP, Cleland MR. AEG/RDI neutron therapy unit. Int J Radiat Oncol Biol Phys 1977;3:377–382.
46. Hess A, Schmidt R, Franke HD. Technical modifications at the DT neutron generator for tumor therapy at Hamburg-Eppendorf. In: Schraube H, Burger G, editors. Proceedings of the Fifth Symposium on Neutron Dosimetry EUR 9762EN. Volume II, Luxembourg: Commission of the European Communities; 1985. pp. 1019–1026.
47. Forman JD, Warmelink C, Sharma R, Yudelev M, Porter AT, Maughan RL. Description and evaluation of 3-dimensional conformal neutron and proton radiotherapy for locally advanced adenocarcinoma of the prostate. Am J Clin Oncol 1995;18:231–238.
48. Maughan RL, Blosser HG, Powers WE. A superconducting cyclotron for neutron radiation therapy. Med Phys 1994;21: 779–785.
49. Maughan RL, Blosser GF, Blosser EJ, Yudelev M, Forman JD, Blosser HG, Powers WE. A multi-rod collimator for neutron therapy. Int J Radiat Oncol Biol Phys 1996;34: 411–420.
50. Farr JB, Maughan RL, Yudelev M, Forman JD, Blosser EJ, Horste T. A multileaf collimator for neutron radiation therapy. In: Marti F, editor. Cyclotrons and Their Applications 2001. Melville, NY: American Institute of Physics; 2001. pp. 154–156.

See also BRACHYTHERAPY, HIGH DOSAGE RATE; IONIZING RADIATION, BIOLOGICAL EFFECTS OF; NUCLEAR MEDICINE INSTRUMENTATION; RADIOTHERAPY, HEAVY ION.

NEUROSTIMULATION. See SPINAL CORD STIMULATION.

NMR. See NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY.

NONIONIZING RADIATION, BIOLOGICAL EFFECTS OF

ANDREW WOOD
Swinburne University of Technology
Melbourne, Australia

INTRODUCTION

Non-ionizing radiation (NIR) refers to that portion of the electromagnetic (EM) spectrum in which the characteristic wavelength is greater than around 180 nm. Radiation of shorter wavelength than this has sufficient quantum energy (given by hc/λ, with h = Planck's constant, c = wavespeed in vacuo, and λ = wavelength) to remove outer electrons from neutral atoms to cause the atom to become ionized, hence, the term "ionizing radiation." NIR consequently does not have the same intrinsic potential for atomic and molecular alteration or the health effects consequent to this. For this reason, damage to DNA and other biomolecules due specifically to the removal of electrons is difficult to envisage. The main groupings of NIR, with increasing wavelength (and decreasing frequency) are ultraviolet (UVR), visible, infrared (IR), radio frequency (RF), and extremely low frequency (ELF). The RF spectrum can be further divided as shown in Table 1 to include microwaves (MW), millimeter waves (MMW), terahertz radiation (THzR), as well as the conventional divisions for broadcast communications. Although not part of the EM spectrum, UVR is normally considered to be part of NIR, as are static (0 Hz) electric, and magnetic fields. The application of the term "radiation" to the ELF portion is also of little consequence, because the wavelength is several thousand kilometers at 50/60-Hz power frequencies.

NON-IONIZING RADIATION PROTECTION

Guidelines on NIR radiation protection are developed by the International Commission on NIR Protection (ICNIRP). In North America, other bodies have developed standards, such as the IEEE International Committee on Electromagnetic Safety and the American National Standards Institute (ANSI), or guidelines, such as the American Conference of Government Industrial Hygienists (ACGIH). Some jurisdictions have chosen to incorporate these (or related) guidelines into legislation.

The mechanism of interaction of NIR with living tissue varies with the groupings just mentioned. These are summarized below, along with effective protection measures against overexposure.

UVR

UVR exposure from the sun outweighs that from all other sources except for a small group of persons in exceptional circumstances. Solar UVR over-exposure is a worldwide problem, leading to increased skin cancer, and by World Health Organization estimates, up to 3 million people are made blind through cataracts. Burning of the skin is a direct indicator of overexposure, at least in the short term. Solar radiation and other UVR sources can initiate photochemical reactions, such as the breakdown of atmospheric oxygen to form oxygen-free radicals and ozone. UVR also has a role in vitamin D control and production. Of greater relevance to adverse health effects, biomolecules (such as DNA components and proteins) can undergo resonant UVR absorption to give rise to dimers (where two similar molecules join to form a single unit). For example, adjacent thymine bases in DNA can fuse to cause an abnormal form. The cell repair mechanisms can sometimes fail to detect this, leading to mutations. The initial response of the skin to UVR within hours of exposure is reddening (erythema or sunburn) due to increased blood flow and edematous changes. The role of photochemical reactions in erythema is unclear. In addition, the immune response can also be suppressed by UVR, increasing risk of infection. On the other hand, the socially attractive tanning of the skin is caused by UVR-induced increase in melanin pigmentation. Chronic exposure leads to skin aging and increased risk of skin cancer. Non-melanoma skin cancers (NMSC) include basal cell carcinoma (BCC: 80%) and squamous cell carcinoma (SCC). The risk of NMSC varies with annual solar UVR dose to the power of between 2 and 3. Melanoma, which has a poor prognosis due it its ability to metastasize, is related to the amount of sun exposure or sunburn during childhood. Chronic eye exposure leads to increased cataract risk. Certain pharmaceutical and other agents lead to photosensitization, in which absorption of longer wavelength UVR can lead to resonant absorption usually associated with shorter wavelengths. The UVR range is usually divided into UV A, B, and C, as indicated in Table 1. The rationale for this is that (1) biological photoreactions are less important above 315–320 nm, and (2) there is virtually no terrestrial solar radiation below 280–290 nm. The boundaries between the ranges are somewhat imprecise. UVA has less capability to cause erythema (by a factor of around 1000) than UV B, but because UVA radiation is the predominant form of solar radiation, it contributes around one sixth of erythemal dose. A minimum erythemal dose (MED) is the UVR exposure (in joule per centimeter squared), which gives rise to just noticeable reddening in the skin of previously unexposed persons. Overexposure is defined as that which leads to erythema within 3 hours or less in a normal population. MEDs have been determined experimentally for narrow bandwidths in the range 180–400 nm, giving a minimum of 30 $J{\cdot}m^{-2}$ at 270 nm. A set of values $S\lambda$, which denote the relative effectiveness of UVR to cause erythema at a specific wavelength λ, are then derived. For example, because at 180 nm, 2500 $J{\cdot}m^{-2}$ is required for the occurrence of erythema compared with 30 $J{\cdot}m^{-2}$ at 270 nm, S_{180} is 30/2500 or 0.012. As exposures are usually to a range of wavelengths (and mainly in the UVA range), a weighted sum for each wavelength component according to its capacity to cause erythema can be obtained. The standard erythemal dose (SED) is then defined such that 1 SED is 100 $J{\cdot}m^{-2}$. This measure is independent of skin type, because MED measurements relate to fair-skinned subjects. Most commonly, overexposure is a result of being outdoors without skin protection, but it can also result from artificial sun-tanning

Table 1. The Non-ionizing Radiation Spectrum

Name of Range	Frequency Range	Wavelength Range	Common Sources
Ultraviolet	UVC 1.07–3 PHz[a]	100–280 nm	Germicidal lamps, Arc welding
	UVB 0.95–1.07 PHz	280–315 nm	Solar radiation, Arc welding
	UVA 750–950 THz[b]	315–400 nm	Solar radiation, Solarium
Visible	430–750 THz	400–770 nm	Solar radiation, indoor and outdoor illumination
Infrared[c]	Near IR (IR A) 214–430 THz	0.7–1.4 μm	Furnaces
	Mid IR (IR B) 100–214 THz	1.4–3 μm	Night photography
	Far IR (IR C) 0.3–100 THz	3 μm–1 mm	Infrared spectroscopy
Terahertz			
Microwave (including millimeter wave)	Extremely High Freq 30–300 GHz	1 mm–1 cm	Satellite, radar, and remote sensing
	Super High Freq 3–30 GHz	1–10 cm	Speed radar guns, Communications
	Ultra High Freq 1–3 GHz	10–30 cm	Mobile telephony
Radio frequency	Ultra High Freq 0.3–1 GHz	30 cm–1 m	Mobile telephony
	Very High Freq 30–300 MHz	1–10 m	TV, FM Radio Broadcasting
	High Freq 3–30 MHz	10–100 m	Electro-welding equipment
	Medium Freq 0.3–3 MHz	100 m–1 km	AM Radio
	Low Freq 30–300 kHz	1–10 km	Long-wave radio
	Very Low Freq 3–30 kHz	10–100 km	Navigation and time signals
Extremely low frequency	< 3 kHz	> 100 km	Electrical power, Electrotherapy
Static	0 Hz		Geomagnetic field, Magnetic Resonance Imaging systems

[a]PHz = peta-Herz, or 10^{15} Hz
[b]THz = tera-Herz, or 10^{12} Hz
[c]The boundaries between near-, mid-, and far-IR are imprecise, as is the terahertz range indicated.

(in a solarium), proximity to tungsten halogen lamps (without filtering glass covers), proximity to UVR lightboxes in scientific and industrial applications, and certain forms of flame welding, with the main possibility of eye damage in these latter sources. Cases of erythema from fluorescent tubes have been reported in extreme cases of photosensitization. The main forms of protection are wearing appropriate clothing, sunblocks (such as zinc oxide cream), sunscreens (based on photo-absorbers, such as para-amino-benzoic acid and cinnamates), and effective sunglasses. Staying out of the sun where this can be avoided is a good behavioral approach for exposure minimization. The "sun protection factor" (SPF) is effectively the ratio of time of exposure before erythema occurs in protected skin to the corresponding time in unprotected skin. A ratio of at least 30 is recommended for effective protection in recreational and occupational exposure to solar radiation. It is important to ensure that sunglasses have sufficient UVR absorption to protect against cataract. Various forms of clothing protect against UVR exposure to differing degrees, ranging from wet open-weave cotton, which offers an ultraviolet protection factor or UPF (which is analogous to SPF) of only around 3–6, to elastane (Lycra) with UPF values of around 100 (99% absorption). It should be noted that these protection factors are computed as the ratio of effective dose (ED) with and without protection (ED/ED_m). The ED is the sum of solar spectral radiance components weighted according to erythemal effectiveness. Here $ED = \sum E_\lambda S_\lambda \Delta\lambda$, where E_λ is the solar spectral irradiance in watt per centimeter squared per nanometer, S_λ is the relative effectiveness of UVR at wavelength λ causing erythema (as mentioned), and $\Delta\lambda$ is a small bandwidth in nanometers. The units of ED are watt per centimeter squared. ED_m is similar, but it contains a factor T_λ to denote the fractional transmission of the test sunscreen (cream, fabric) at a particular wavelength (i.e., $ED_m = \sum E_\lambda S_\lambda T_\lambda \Delta\lambda$). The Global UV Index (UVI) is a dimensionless quantity in which the ED is summed over the range 250—400 nm and multiplied by 40 m^2W^{-1}. In Darwin, Australia, this ranges from 0 to 3 in the early morning and evening to 14 or more at noon on a clear day. At this UVI, erythema will result in fair skin after 6 minutes. See http://www.icnirp.de/documents/solaruvi.pdf for further information.

It is estimated that significant reductions in the incidence of both malignant and benign forms of skin cancer could be achieved by the enforcement of protective measures, particularly in occupational settings involving fair-skinned people in outdoor work in tropical or subtropical regions. Occupational exposures in Australia have recently been measured (1), and UVR safety has been reviewed in several publications (see Reference (2), for example). Indicative exposure limits are given in Table 2. It should be emphasized that for brevity many details are omitted from this table. For full details of limits pertaining to a particular geographical region, local radiation protection authorities should be consulted. The ICNIRP guidelines are readily accessible via downloads from http://www.icnirp.de. These represent reviewed publications originally appearing in *Health Physics*.

Visible Radiation

This is the region of NIR to which the retinal pigments of the eye are sensitive, so understandably, eye injury is the main concern in overexposure. There are two forms of hazard: photochemical and thermal. In addition, if the eye lens has been surgically removed (aphakia), there is an enhanced risk of damage. Photochemical damage becomes more likely with shorter wavelengths and is sometimes referred to as the "blue light hazard." The type of photochemical reaction is bleaching of the visual pigments, leading to temporary loss of vision. Thermal injury can result in permanent impairment of vision, especially if the foveal region, used for fine focus, is involved. Thresholds for these forms of injury have been determined in the wavelength range 400–1400 nm (thus including near infrared, see below) and an assessment of whether these are exceeded, for a particular source takes into consideration the spectral characteristics of the source. For exposures shorter than a few hours, the total radiance should be below 10^6 $W{\cdot}m^{-2}{\cdot}sr^{-1}$, where sr refers to a unit solid angle tended by the source. Lasers represent the sources most likely to cause injury, and because these emit a small number of discrete wavelengths, this assessment can be straightforward. Eye injury is minimized by the blink reflex, but laser wavelengths outside the visible range are less easy to control, because their paths are difficult to track, especially from incidental reflections. Lasers are classified according to the luminous power, their visibility, and their effective aperture, as described further in a separate entry on **Lasers**. High-power lasers are used in machining, welding, and engraving of a variety of materials, including plastics, metals, and fabrics. They also provide the source of beams in communications and photonics research laboratories. During normal operation, a combination of administrative and engineering controls provide adequate protection for workers. On the other hand, high-power lasers used in "light show" entertainment have sometimes given rise to unintentional beams directed at members of the public. The unrestricted distribution of laser pointers, with a capacity of causing eye damage, has also been a concern in several jurisdictions. Apart from laser sources, welding flames represent the next most common form of visible light hazard ("welder's flash"). Hazard can be minimized by the use of appropriate goggles. Recently, high-powered light-emitting diode (LED) sources have been evaluated by the ICNIRP for their potential for visible light hazard, particularly those emitting blue light. Although injury is unlikely, the power density of these devices continues to increase as technology develops.

IR

The major sources of IR radiation that are of concern are furnaces and some high-powered non-visible laser devices (femtosecond lasers). Here there is an increased possibility of local thermal injury, but because there is poor penetration of the lens of the eye, the possibility of retinal damage is reduced compared with the visible range. The IR range is divided into three ranges as shown in Table 1. Above 1–2 μm, water is a strong absorber of IR. Whereas guidelines for optical radiation extend up to 1.4 μm (near

Table 2. Approximate Exposure Limits for NIR: Exact Limits Vary Between Countries and In Some Cases Between Different Contexts of Exposure

Name of Range	Indication of Level Above Which Intervention Is Recommended	Biohazard Forming Basis of Protection	References to *Health Physics* Publications
Ultraviolet	U-shaped over wavelength range: 180 nm–2.5 $kJ \cdot m^{-2}$; 270 nm–30 $J \cdot m^{-2}$ (minimum); 400 nm–1 $MJ \cdot m^{-2}$	Skin reddening due to burn (erythema), also prevention of cataract	Vol 71, p 978 (1996) Vol 84, pp 119–127 (2004)
Visible	Depends on viewing position and spectral content of source	Retinal thermal or photochemical damage	Vol 73, pp 539–554 (1997)
Lasers (includes above and below)	Depends on wavelength, exposure duration, and size of aperture. For long exposures (> 100 s), limits are of the order of 1 $W \cdot m^{-2}$	Retinal (esp. foveal) damage: photochemical or thermal Also skin.	Vol 71, pp 804–819 (1996) Vol. 79, pp 431–440 (2000)
Infrared	100 $W \cdot m^{-2}$ for long exposure* Not well defined	Thermal injury to lens and cornea	Vol 73, pp 539–554 (1997)
Terahertz			
Microwave (including millimeter wave)	6–300 GHz: 50 $W \cdot m^{-2}$ (time averaged) 50 $kW \cdot m^{-2}$ peak[a]	Rise in tissue temperature sufficient to cause protein denaturation	Vol. 74, pp 494–522 (1998)
	10 $mJ \cdot kg^{-1}$ within 50 μs interval[a]	Microwave hearing	As above
Radio frequency	0.1–6,000 MHz: 0.4 $W \cdot kg^{-1}$ for whole-body exposure; 10 $W \cdot kg^{-1}$ for 10 g mass (head and torso)[a].	Rise in tissue temperature sufficient to cause protein denaturation	As above
	3–10,000 kHz: f/100 (f in Hertz) $mA \cdot m^{-2}$ in head and torso[a]	Shocks or burns due to induced current or contact current	As above
Extremely low frequency	Tissue induced field: 18 $mV \cdot m^{-1}$ for f, 20 Hz; 18(f/20) $mV \cdot m^{-1}$ for f between 20 & 800 Hz (IEEE)[b], 10 $mA \cdot m^{-2}$ for range 4–1,000 Hz (ICNIRP)	Magnetophosphenes, micro-shock	As above
Static	0.2 T time weighted average[a], 2 T ceiling, 5 T limbs	Magnetophosphenes associated with movement	Vol 66, pp 100–106 (1994)

[a] These basic restrictions are for occupational exposures: Divide by 5 to get general public limits.
[b] These basic restrictions are for "controlled environment" (i.e., occupational) exposures: Divide by 3 to get general public limits.

infrared), there is some disagreement on the appropriate levels beyond that. Levels of incident radiation above 100 $W \cdot m^{-2}$ are considered as posing an unacceptable thermal hazard. Those at risk of overexposure include foundry workers and welders. Recently, advances have extended telecommunications frequencies into the "terahertz gap," the region between 0.3 and 3 THz, which has been unexploited by technological applications. The health effects are currently unknown, but they are expected to be similar to those of the contiguous frequency ranges. However, there is a current discontinuity between IR and RF standards or guidelines for a 1 mm wavelength (0.3 THz).

RF

Common sources of high-power RF emissions include welding equipment and induction heaters used in industrial drying processes. Radio, TV, and telecommunications transmitters can involve high broadcast powers (400 kW or more for commercial TV stations). There are two types of potential hazard: thermal injury in the range 100 kHz–300 GHz and neural stimulation due to induced currents or contact with metallic surfaces at frequencies below 10 MHz. At 300 GHz, the effective wavelength in tissue is less than 1 mm, so very little will penetrate below the skin. On the other hand, at 80 MHz, the wavelength is comparable with the long axis of the human body, so absorption is enhanced. Protective measures in terms of incident RF power density (W/cm^{-2}) are thus strictest in the range 10–400 MHz. The basic restriction above 100 kHz is on the rate of energy absorption by tissue (specific absorption rate, or SAR, in W/g of tissue). SAR is related to the RF electric field induced in tissue (E_i $V \cdot m^{-1}$) such that

$$\mathrm{SAR} = \sigma E_i / \rho$$

where σ is local conductivity in S/m and ρ is tissue density in kg/cm^3. In unperfused insulated tissue, SAR is related to the rate of rise of temperature dT/dt via

$$\mathrm{SAR} = k \cdot dT/dt$$

where k is the specific heat of tissue, 3480 $J \cdot kg^{-1} \cdot K^{-1}$ approximately.

This basic restriction is limited to values for whole-body or localized exposures such that normal thermoregulation would not be compromised, with a 10-fold safety margin. Although there is some variation between standards in place throughout the world, many countries employ a distinction between occupationally exposed persons ("aware users") and the general public, for whom an extra five-fold level of protection is provided. The ICNIRP value for whole-body SAR for the general public is 0.08 $W \cdot kg^{-1}$, with higher values of 2 $W \cdot kg^{-1}$ in the head and trunk and 4 $W \cdot kg^{-1}$ in the limbs, averaged over 10 g of tissue. The power density of incident plane-wave radiation (in watt per centimeter squared), which would give rise to these levels of SAR (for far-field exposures), has been computed by mathematical modeling and animal studies in a conservative manner, such that if these reference levels are complied with, the basic restrictions will be met. As, for free space, the power density S is related to the electric and magnetic field values (E and H, respectively) by $S = E^2/377 = H^2 0.377$, compliance testing can be accomplished by measuring E-field values alone. Reference levels at particular frequencies can be found by reference to the ICNIRP guideline as indicated in Table 2.

Induced current density restrictions are imposed at 10 MHz and below. Above this frequency, it is considered that the fields vary too quickly to produce neural stimulation. Again, there is a safety factor of 10 between occupational levels and the level at which mild stimulatory effects can be noted in 1% of the population. This ranges from 100 $A \cdot m^{-2}$ at 10 MHz to 10 $mA \cdot m^{-2}$ at 4 Hz–1 kHz, in the ICNIRP guidelines. This will be discussed further in the ELF section.

At frequencies between 0.2 and 6 GHz, a phenomenon of "microwave hearing," due to thermoelastic expansion of brain tissue in response to pulsed radiation, occurs. Additional restrictions are in place in the ICNIRP guidelines to prevent this from occurring.

Overexposure to RF radiation, leading to serious burns, is usually due to the failure of control measures, such as guards on RF seam welding apparatus or work on RF antennas mistakenly thought to be nonoperational.

The safety of communications equipment, including mobile telephony handsets and base stations, is a major community concern. There is little substantive evidence of harm from long-term exposure at so-called "non-thermal" levels, but because there are many young users of handsets, many countries have endorsed a precautionary approach, encouraging use only for necessity. The scientific evidence for the possibility of "non-thermal" effects has been reviewed in the United Kingdom by the Independent Expert Group on Mobile Phones (IEGMP) (3) and by other bodies. The IEGMP concluded that although "the balance of evidence to date suggests that (low levels of RF radiation) do not cause adverse health effects" that "gaps in knowledge are sufficient to justify a precautionary approach." Some national standards (for example, Australia and New Zealand) incorporate a "precautionary" clause; that is, exposures incidental to service delivery should be minimized (but taking other relevant factors into consideration). The limiting of mobile phone use by children was recommended by the IEGMP (3), but the Health Council of the Netherlands sees no convincing scientific argument to support this (4).

ELF and Static

The range of frequencies (0–3 kHz) includes power transmission and distribution systems (50/60 Hz) as well as transportation systems (0, 16.7, 50, and 60 Hz), surveillance systems, and screen-based visual display units. Here the main potential hazard from exposure to fields (rather than direct contact with conductors) seems to be from inappropriate neural stimulation due to induced current (as in the case of RF, above). Consequently, treating ELF as a special case may seem out of place, but because the ELF range is precisely that of biogenic currents due the operation of nerves and muscles, its separate treatment is justified. The susceptibility of cells to the influence of exogenous currents is related to the time constants for the operation of cell membrane channels, which are typically of the order of milliseconds. At lower frequencies, cell membranes tend to adapt to imposed electrical changes, so restrictions need to be strictest in the range 10–1000 Hz. In humans, the retina of the eye represents a complex network of interacting nerve-cells, giving rise to sensations of pinpoints of light when stimulated by external electric and magnetic fields (EMFs). As this gives a guide to the levels at which stimulatory effects could become an annoyance, or could possibly be interpreted as a stressor, a basic restriction for occupational exposure of 10 $mA \cdot m^{-2}$ (which corresponds to an induced field of around 100 $mV \cdot m^{-1}$) has been adopted by the ICNIRP for the range 4–1000 Hz. This restriction rises above and below this range. In particular, at 0 Hz (static fields), levels are restricted to 40 $mA \cdot m^{-2}$. Levels for the general public are less by a factor of 5. Reference levels for magnetic fields are derived from these basic restrictions by considering the body to be simple geometric objects, but more advanced modeling yields similar results. For sinusoidally varying fields, the reference magnetic fields can be derived from basic restrictions via the formulas

$$B = E/(\pi fr) \quad \text{or} \quad B = J/(\sigma \pi fr)$$

where E refers to the basic restriction in terms of induced tissue electric field (in volt per meter), J is the basic restriction in induced current density (A/cm^2), f is the frequency in Hertz, σ is the tissue conductivity (S/m), and r is the radial distance from the center of symmetry (in the same direction as the external magnetic field B).

Electric field reference levels are derived more from considerations of avoiding "microshocks," which may occur, for example, if an arm with finger extended is raised in an intense electric field. Details of these reference levels can be found (for the ICNIRP limits) at http://www.icnirp.de. As it is possible to exceed the electric field reference levels in electrical switchyard work, special precautions need to be taken. Exceeding magnetic field reference levels is rare. Some government and other organizations have advocated a much more prudent approach to limiting exposure, particularly to the general public. This comes from some dozen or so well-conducted

epidemiological studies linking exposure of children to a time-weighted average magnetic field of 0.4 μT or more, to an approximate doubling of leukemia incidence. The possibility of low-level health effects of ELF has been the topic of research for nearly three decades. As there is no agreed mechanism for how elevated leukemia rates could be brought about, nor is there adequate evidence from long-term animal studies, there is doubt that magnetic fields are the causative agent. Nevertheless, time-varying ELF magnetic fields (but not electric fields, nor static fields) have been categorized by the International Agency for Research in Cancer (IARC) as a "possible carcinogen" (category 2B) (5). Essentially, the U.S.-government funded EMF-RAPID (Electric and Magnetic Field Research and Public Information Dissemination) program, whose Working Group reported in 1998 (6), came to a similar conclusion. The final report of the NIEHS Director (7), on the other hand, concluded that "the scientific evidence suggesting that ELF-EMF pose any health risk is weak" but also acknowledged that "exposure cannot be recognized as entirely safe because of weak scientific evidence that exposure may pose a leukemia hazard." The report also advocated "educating both the public and regulated community on means aimed at reducing exposures." There is intense debate on how a policy of prudence should actually be interpreted, because approximately 1% of homes would be in the "over 0.4 μT" category (8,9) (this percentage varies widely between and even within countries). Several moderate cost engineering measures can be employed to reduce field levels from transmission lines, and electric power companies often employ these in new installations.

PERCEIVED ELECTRO-SENSITIVITY

Several persons claim debilitating symptoms associated with proximity to electrical installations or appliances or in association with the use of mobile (cell) phones. Despite several well-conducted, independent, "provocation studies," in which sufferers have been subjected to energized and not energized sources in random order, no association between exposure status and occurrence of symptoms has been established. A recent Dutch study of psychological sequelae of mobile phone use implied that the overall baseline responses in a group of "electro-sensitives" differed from a similarly sized group of "normals," but that the changes associated with mobile phone use were similar in both groups.

ULTRASOUND

Few processes and devices outside of clinical medicine involve the possibility of human exposure to ultrasound if normal protective guarding measures are in place. Airborne ultrasound is used in surveying instruments and in a variety of drilling, mixing, and emulsification industrial processes. Ultrasonic descalers are used in dentistry and to clean jewelry. Reports of injury are rare. For industrial applications, the frequency range of 20–100 kHz is covered by ICNIRP limits and is based on the pressure amplitude of the ultrasound in air (these are of the order of 110 dB, referenced to 2×10^{-5} Pa). In clinical applications, ultrasonic energy is usually delivered across the skin via coupling gel and is in the frequency range 1–25 MHz. Diagnostic ultrasound is designed to prevent tissue temperature rising above 41 °C for sustained periods (10,11). Effectively, beam intensities are capped at 1000 $W \cdot m^{-2}$ (spatial peak, temporal average), except for short periods of insonation. Higher intensities are possible if the energy density is below 500 $kJ \cdot m^{-2}$. This gives a large margin below established hazardous effects. Therapeutic ultrasound exposure is usually limited by patients reporting excessive heat, but use on patients with limited sensation is of concern. Intensities of 10 $kW \cdot m^{-2}$ are common in therapeutic applications. Tissue damage occurs above 10 $MW \cdot m^{-2}$.

SERIOUS INJURY FROM NIR

From above, it would appear that NIR is fairly innocuous. It should be stressed, however, that high-power devices, if inappropriately used or modified, can cause serious injury. UVC is routinely used as in germicidal devices, and the micro-cavitation produced by intense ultrasound beams is used to disrupt tissue. Laser skin burns occasionally occur in research laboratories. Severe injury and fatalities have resulted from surgical uses of lasers in which gas embolisms have become ignited within body cavities. Early unshielded microwave ovens were associated with severe kidney damage. Cases of severe burns are still too common in small businesses using RF heat sealers, often due to the removal of guards. Serious burns result from an accidental or ill-advised approach to broadcast antennas and other communications equipment (12). In addition to burns, severe chronic neurological deficits can also result from overexposure to RF currents (13).

ACHIEVING ADEQUATE PROTECTION AGAINST NIR

Opinion is divided about the need to control NIR exposure by legislation. Communications equipment manufacturers have to comply with rigid requirements related to health guidelines and standards, and many countries have the power to prosecute in instances where equipment is tampered with or altered such that the guidelines would be exceeded. Codes of practice often have provisions for marking "no go" areas where levels could be exceeded, with appropriate signage. In terms of the potential for preventing debilitating illness or early death, the link between solar UVR and skin cancer and cataract represents the area where intervention is most warranted. It is estimated that adequate sun protection could perhaps save tens of lives per million of population per annum with over $5M pa per million in savings in health costs. The costs of ensuring employers of outdoor workers and the workers themselves complying with measures of UVR exposure reduction are hard to estimate, but they are likely to be high. Whereas compliance with a limit of 30 $J \cdot m^{-2}$ equivalent (or MED) is achievable in relation to artificial sources, this level can be exceeded in less than an hour's exposure to intense solar

radiation around noon in low latitudes. Employers can be required to educate their workforce to use appropriate measures to reduce the risk of becoming sunburnt, but it is virtually impossible to eliminate this from actually occurring. It would seem unreasonable to require employers to be responsible for an overexposure to a familiar and essential source of energy to which we have all been exposed since the dawn of time.

As several forms of NIR carry with them an uncertainty of possible harm in the long term, several national radiation protection authorities have espoused the "Precautionary Principle." This entails taking measures to reduce exposure, even where exposures are well within levels set by scientific evaluation of the available research. It is recognized that reducing exposure might itself introduce new hazards or increase other hazards (such as being unable to use a cell-phone in an emergency because of extra power restrictions), so an evaluation of the need to be "Precautionary" with respect to NIR should be in the wider context of overall risk management. In general, the introduction of arbitrary extra margins of safety, in order to appease public outcry, is not warranted.

USES OF NIR IN MEDICAL DIAGNOSIS AND THERAPY

UVR

The UVR-induced photochemical reactions form the basis of an effective treatment of the disease psoriasis, which is marked by widespread red itchy scales on the skin. This is caused by an accelerated cell cycle and DNA synthesis in skin cells. The drug psoralen is preferentially taken up by these dividing cells, which on subsequent exposure to UVA radiation, leads to binding with DNA and subsequent inhibition of synthesis and cell division. A normal course of treatment consists of 25 monthly visits to a clinic, with 8-methoxypsoralen taken orally, followed 2 h later by a UVR exposure of 10–100 $kJ \cdot m^{-2}$ per visit. This is usually delivered via a bank of 48 or so high-intensity fluorescent tubes.

A second use of UVR in biological and clinical analysis and research is in the identification of biomarkers through fluorescence. One technique involves placing electrophoretic gels over a UVR lightbox to localize the fluorescent regions. As mentioned, the possibility of overexposure in those who perform multiple observations is a matter of concern.

Lasers

The high intensity of laser radiation, particularly if it is pulsed, provides a means of tissue ablation, carbonization, coagulation, and desiccation. High-intensity short pulses produce photomechanical disruptions of tissue. At longer pulse lengths (~ 1 s), thermal and photochemical processes become more important. Excimer (= excited dimer) laser radiation has proved to be useful in the surgical treatment of defects in vision. This technique, radial keratotomy or keratectomy, reshapes the corneal surface to alter the effective focal length of the eye and thus do away with the need for spectacles or contact lenses. Laser ablation is also useful in the treatment of ocular melanoma, Barratt's esophagus, removal of "port wine" stains on the skin, and (using an optical fiber delivery system in a cardiac catheter) the removal of atheromatous plaque in coronary arteries. A second property of intense laser light, that of photo-activation, is exploited in a range of treatments known as photodynamic therapy (PDT). In this, several compounds are known to be preferentially taken up by tumor tissue but also have the property of resonant absorption of light to produce free radicals, such as singlet oxygen and oxygen radical, which ultimately lead to endothelial cell membrane damage, blood supply shutdown, and hence necrosis of tumor tissue. These photosensitizing compounds are injected, or in some cases taken by mouth. Intense laser light (of 600–770 nm wavelength) is then directed at the tumor to produce this photo-activation. Energy thresholds are of the order of 1 $MJ \cdot m^{-2}$. Although used mainly on superficial tumors (depth less than 6 mm), optical fiber delivery into deeper tissue (such as the breast) has also been trialled. As the tumor tissue becomes fluorescent on uptake of these compounds, diagnostic techniques (photodynamic diagnosis or PDD) are based on a similar principle. Suitable compounds are related to hemoglobin (hematoporphyrin derivative or HpD), rhodamine, amino levulinic acid, bacteriochlorins, and phthalocyanines. The herb St John's Wort also yields hypericins that have similar properties. The ability to use scanning optics in association with optical fibers has provided ways of making microscopic endoscopy possible.

Incoherent sources of blue light are used in the treatment of neonatal jaundice (hyperbilirubinemia). Bilirubin is decomposed during the exposure of the neonate to fluorescent tubes (filtered to remove wavelengths shorter than 380 nm).

IR

Infrared reflectivity from the skin and from layers immediately below the skin varies with skin temperature. Thermography has been used to identify regions of enhanced or reduced peripheral blood flow, occurring, for example, in mammary tumors. The high false-positive rate has inhibited its use in mass screening for this disease. On the other hand, breast imaging using time-of-flight IR transmission methods shows promise. Blood oxygen saturation is easily measured noninvasively via the ratio of reflectances at two wavelengths, 650 and 805 nm (the wavelengths showing greatest and least sensitivity to the degree of saturation, respectively). This forms the basis of the pulse oximeter, which clips on the finger and gives an indication of pulse rate in addition to oxygen saturation. Laser Doppler blood flow meters give an indication of capillary blood flow via the autocorrelation of reflected light signals. Wavelengths of 780 nm are selected because of the good depth of penetration of skin.

IR spectroscopy has a wide range of industrial and research applications, because of specific molecular stretching, bending, and rotational modes of energy absorption.

Terahertz

Several medical applications have been proposed for terahertz radiation, arising out of differential reflection from

cancerous/normal skin and from its relatively good transmission through bones and teeth. Its use in biosensing is also being investigated.

RF

The tissue heating and consequent protein denaturation has been used in catheter-tip devices for ablating accessory conduction pathways in the atria of the heart, giving rise to arrhythmias. The use of focused RF in cancer hyperthermia treatment has been used in conjunction with conventional radiotherapy to improve the hit rate of the latter, most likely due to the increased available oxygen via thermally induced blood flow increase. Increased blood perfusion is also thought to underlie the use of RF diathermy in physiotherapy, although this has now been almost entirely replaced by therapeutic ultrasonic diathermy (see below). RF exposures are part of magnetic resonance imaging (MRI), where some care has to be taken to avoid "hot spots" during investigation. SARs can exceed 2 $W \cdot kg^{-1}$ at frequencies in the region of 100 MHz. If we can extend the term "radiation" to include the direct application of RF currents, then electrical impedance tomography (EIT) should be included. In this technique, current of approximately 50 kHz is applied via a ring of electrodes to the torso or head, essentially to identify differential conductivity values in different organs and thus track shifts in fluid content, post-trauma, for example.

ELF

In clinical diagnosis, nerve conduction and muscular function studies are performed by examining responses to electrical stimulation (by single pulses or trains of pulses of the order of a few milliseconds in duration) of particular groups of nerve fibers. Electrical stimulation of specific regions of the body are also reported to give rise to beneficial effects. For example, or transcutaneous electrical nerve stimulation (TENS) is of some efficacy in controlling pain by raising the threshold for pain perception. Interferential therapy, which consists of a combined exposure of regions of the skin to low currents at two narrowly separated frequencies (for example, 4 kHz and 3.7 kHz) are claimed to be useful for a range of muscular and joint pain conditions and for circulatory disorders, but the mode of interaction is unclear. The currents are of the order of 50 mA, and the tissue is reportedly performing a demodulation of the 4 kHz carrier to produce a TENS-like deep current of a few hundred Hertz. Similarly, pulsed magnetic fields (PEMFs) are claimed to be effective in speeding healing in bone fractures, despite the small magnitude of induced currents. On the other hand, electroconvulsive therapy (ECT), in which pulses of current of several milliamperes are passed through the head, cause general nerve activation. This therapy is of proven value in cases of severe depression, but the origin of this benefit is an enigma. Transcranial magnetic stimulation (TMS) can be used both in diagnosis by eliciting specific responses and in therapeutic mode, in a manner analogous to ECT. However, the therapeutic efficacy of TMS still awaits clarification.

Static

The application of permanent magnets to painful joints is claimed to have beneficial effects, but the evidence for efficacy is equivocal. It has been suggested that the Lorentz-type forces on flowing electrolytes (such as blood) produce electric fields and currents. However, at typical blood flow velocities of 0.1 $m \cdot s^{-1}$, a 1 mT magnet will only induce 0.1 $mV \cdot m^{-1}$, which is well below levels shown in Table 2.

Ultrasound

The use of ultrasound in the range 1–25 MHz in diagnosis originates from the wavelength (and, hence, resolution) being of the order of a few millimeters. Acoustic mismatch between tissue layers gives radar-type echoes that form the basis of 2D and 3D imaging. The Doppler shift due to flowing fluid forms the basis of its use in blood flow measurements. Differential absorption provides a means for tissue characterization. In therapeutic ultrasound, the warmth is produced by adiabatic expansion and contraction within the tissue, to a depth of several centimeters. At higher intensities, cavitation and mechanical movement of organelles can occur.

BIBLIOGRAPHY

1. Vishvakarman D, Wong JC, Boreham BW. Annual occupational exposure to ultraviolet radiation in central Queensland. Health Phys 2001;81:536–544.
2. Diffey BL. Solar ultraviolet radiation effects on biological systems. Phys Med Biol 1991;36:299–328.
3. Stewart W. Mobile phones and health, Independent Expert Group on Mobile Phones, National Radiation Protection Board, Chilton, Didcot, U.K., 2000.
4. van Rongen E, Roubos EW, van Aernsbergen LM, Brussaard G, Havenaar J, Koops FB, van Leeuwen FE, Leonhard HK, van Rhoon GC, Swaen GM, van de Weerdt RH, Zwamborn AP. Mobile phones and children: is precaution warranted? Bioelectromagnetics 2004;25:142–144.
5. IARC, Non-ionizing radiation, Part 1: static and extremely low-frequency (ELF) electric and magnetic fields. International Agency for Research on Cancer, Lyon, France Monographs, Vol. 80, 2002.
6. Portier CJ, Wolfe MS. Assessment of health effects from exposure to power-line frequency electric and magnetic fields. National Institute of Environmental Health Sciences, Research Triangle Park, NC, NIH Publication 98-3981, 1998.
7. Olden K. NIEHS Report on health effects from exposure to power-line frequency electric and magnetic fields. National Institute of Environmental Health Sciences, Research Triangle Park, NC, NIH Publication No. 99-4493, 1999.
8. Greenland S, Sheppard AR, Kaune WT, Poole C, Kelsh MA. A pooled analysis of magnetic fields, wire codes, and childhood leukemia. Childhood Leukemia-EMF Study Group. Epidemiology 2000;11:624–634.
9. Ahlbom A, Day N, Feychting M, Roman E, Skinner J, Dockerty J, Linet M, McBride M, Michaelis J, Olsen JH, Tynes T, Verkasalo PK. A pooled analysis of magnetic fields and childhood leukaemia. Br J Cancer 2000;83:692–698.
10. Abramowicz JS, Kossoff G, Marsal K, Ter Haar G. Safety Statement, 2000 (reconfirmed 2003). International Society

of Ultrasound in Obstetrics and Gynecology (ISUOG). Ultrasound Obstet Gynecol 2003;21:100.
11. Safety statement, 2000. International Society of Ultrasound in Obstetrics and Gynecology (ISUOG). Ultrasound Obstet Gynecol 2000;16:594–596.
12. Hocking B, Joyner KJ, Newman HH, Aldred RJ. Radiofrequency electric shock and burn. Med J Aust 1994;161:683–685.
13. Schilling CJ. Effects of exposure to very high frequency radiofrequency radiation on six antenna engineers in two separate incidents. Occup Med (Lond) 2000;50:49–56.

Further Reading

Australian Radiation Protection and Nuclear Safety Agency: http://www.arpansa.gov.au
Dennis JA, Stather J, editors. Non-ionizing radiation. Radiation Protection Dosimetry. 1997. 72:161–336.
ICNIRP references from *Health Physics*: http://www.icnirp.de.
National Radiological Protection Board (NRPB) U.K.: http://www.nrpb.co.uk.

See also BIOMATERIALS, SURFACE PROPERTIES OF; IONIZING RADIATION, BIOLOGICAL EFFECTS OF; RADIATION THERAPY SIMULATOR.

NUCLEAR MAGNETIC RESONANCE IMAGING.

See MAGNETIC RESONANCE IMAGING.

NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

WLAD T. SOBOL
University of Alabama
Birmingham, Alabama

INTRODUCTION

When two independent groups of physicists (Bloch, Hansen, and Packard at Stanford and Purcell, Torrey, and Pound at MIT) discovered the phenomenon of nuclear magnetic resonance (NMR) in bulk matter in late 1945, they already knew what they were looking for. Earlier experiments by Rabi on molecular beams, and the attempts of Gorter to detect resonant absorption in solid LiF, seeded the idea of NMR in bulk matter. Fascinating stories describing the trials and tribulations of the early developments of NMR concepts have been told by several authors, but Becker et al. (1) deserve a special citation for the completeness of coverage.

For his achievements, Rabi was awarded a Nobel Prize in 1944, while Bloch and Purcell jointly received theirs in 1952. What was the importance of the Bloch and Purcell discoveries to warrant a Nobel Prize despite an abundance of prior work offering numerous clues? It was not the issue of special properties of elementary particles, such as a spin or magnetic moment: This was first demonstrated by the Stern–Gerlach experiment. It was not the issue of the particle interactions with magnetic field: This was first illustrated by the Zeeman effect. It was not even the magnetic resonance phenomenon itself: This was first demonstrated by Rabi. It was the discovery of a tool that offered a robust, nondestructive way to study the dynamics of interactions in bulk matter at the atomic and molecular level that forms the core of Bloch and Purcell's monumental achievements. However, despite the initial excitement at the time of their discovery, no one could have predicted just how extensive and fruitful the applications of NMR would turn out to be.

What is the NMR? The answer depends on who you ask. For Bloch's group at Stanford, the essence of magnetic resonance was a flip in the orientation of magnetic moments. Bloch conceptual view of the behavior of the nuclear magnetic moments associated with nuclear spins was, in essence, a semiclassical one. When a sample substance containing nuclear spins was kept outside a magnetic field, the magnetic moments of individual spins were randomly oriented in space, undergoing thermal fluctuations (Brownian motion). The moment the sample was placed in a strong, static magnetic field, quantum rules governing the behavior of the spins imposed new order in space: the magnetic moments started precessing around the axis of the main magnetic field. For spin ½ particles (e.g., protons), only two orientations w.r.t. static magnetic field were allowed; thus some spins precessed while oriented somewhat along the direction of the external field, while other spun around while orienting themselves somewhat opposite to the direction of that field. To Bloch, a resonance occurred when externally applied radio frequency (RF) field whose frequency matched the precessional frequency of the magnetic moments, forced a reorientation of precessing spins from parallel to antiparallel (or vice versa). They called this effect a nuclear induction.

As far as the Purcell's group was concerned, NMR was a purely quantum mechanical phenomenon. When a diamagnetic solid containing nuclei of spin I is placed in a static magnetic field, the interactions of nuclear magnetic moments with the external magnetic field cause the energy levels of the spin to split (the anomalous Zeeman effect). When an external RF field is applied, producing quanta of energy that match the energy difference between the Zeeman levels, the spin system would absorb the energy and force spin transitions between lower and upper energy states. Thus, they described the phenomenon as resonance absorption.

It can be proven that these two concepts of NMR phenomenon are scientifically equivalent. However, the two views are psychologically very different, and have been creating a considerable chasm in the accumulated body of knowledge. Some aspects of NMR applications are intuitively easier to understand using Bloch's semiclassical vector model, while other naturally yield themselves to the quantum picture of spin transitions among energy states. The details of this dichotomy and its impact on the field of NMR applications are fascinating by themselves and have been extensively discussed by Ridgen (2).

At the time of the NMR discovery, nobody had any inkling that this phenomenon might have any applications in medicine. To understand how NMR made such a big impact in the medical field, one has to examine how the NMR and its applications evolved in time. Nuclear mag-

netic resonance was discovered by physicists. Thus it is not surprising that the initial focus of the studies that followed was on purely physical problems, such as the structure of materials and dynamics of molecular motions in bulk matter. During a period of frenzied activities that followed the original reports of the discovery, it was very quickly understood that interactions among nuclear spins, as well as the modification of their behavior by the molecular environment, manifest themselves in two different ways. On the one hand, the Zeeman energy levels could shift due to variations in the values of local magnetic field at different sites of nuclear spins within the sample. This causes the resonant absorption curve to acquire a fine structure. Such studies of NMR lineshapes provide valuable insights into the structure and dynamics of molecular interactions, especially in crystals. This branch of NMR research is customarily referred to as radiospectroscopy.

On the other hand, when a sample is placed in the external magnetic field, the polarization of spin orientations causes the sample to become magnetized. When the sample is left alone for some time, an equilibrium magnetization develops. This equilibrium magnetization, $\boldsymbol{M}_0$, is proportional to the strength and aligned in the direction of the external static magnetic field, $\boldsymbol{B}_0$. An application of RF field disturbs the equilibrium and generally produces a magnetization vector, $\boldsymbol{M}$, that is no longer aligned with $\boldsymbol{B}_0$. When the RF field is switched off, the magnetization returns over time to its equilibrium state; this process is called a relaxation. The process of restoring the longitudinal component of the equilibrium magnetization requires that the spins exchange energy with their environment; thus, it is commonly referred to as spin–lattice or longitudinal relaxation. The characteristic time that quantifies the rate of recovery of the longitudinal component of magnetization toward its equilibrium value, $\boldsymbol{M}_0$, is called the spin–lattice relaxation time and denoted T_1 or, in medical applications, *T1*. At equilibrium, the transverse magnetization component is zero. Thus, any nonzero transverse component of nonequilibrium magnetization must decay back to zero over time. This process tends to be dominated by interactions among spins and is thus called a spin–spin or transverse relaxation. The characteristic time that quantifies the rate of decay of the transverse component of magnetization is called the spin–spin relaxation time and denoted T_2 or, in medical applications, *T2*. Both T_1 and T_2 strongly depend on the nature of the molecular environment within which the spins are immersed, thus offering a robust probe of molecular dynamics and structure in a variety of materials (solid, liquid, and gaseous) over a range of conditions (temperature, phase transitions, chemical reactions, translational and rotational diffusion, etc.). Studies of relaxation times are referred to simply as NMR relaxation studies, and sometimes as relaxometry.

In solids, dipole–dipole interactions among spins are dominant, which for proton NMR (^{1}H NMR) studies results in fairly wide lineshapes (with a width of several kHz) with very little fine structure. In most liquids, however, the substantially faster molecular reorientations average the dipole–dipole interactions, effectively suppressing them to produce a vanishing net effect on the NMR absorption curves that become much narrower (typically of the order of Hz). This feature has led to a discovery of chemical shift phenomenon.

The most dramatic demonstration of the chemical shift was the observation made in 1951 by Arnold et al. (3) who showed separate spectral NMR lines from nonequivalent protons in a sample containing a simple organic substance, ethanol. This gave birth to high-resolution NMR spectroscopy or HR NMR, a powerful tool that assists chemists in nondestructive analysis of organic compounds. This *in vitro* technique underwent massive developments over the years and almost overshadowed the NMR applications in physics. An exhaustive overview of HR NMR applications has been published by Shoolery (4). Today, HR NMR spectroscopy plays a major role in studies of biological materials *in vitro* and in drug development research. This research, although not directly used in clinical care, nevertheless is having a major impact on the development of medical arts. A comprehensive review of biological applications of NMR spectroscopy has been provided by Cohen et al. (5).

Standard NMR studies are performed *in vitro*: The sample is placed in the bore of a laboratory magnet, and the signal is collected from the entire volume of the sample. Samples are relatively small: The typical NMR tube vial is $\sim$ 5 mm outside diameter (OD) and holds $\sim$ 0.5 mL of sample material. Nuclear magnetic resonance magnets have relatively small active volumes [typical bore size of modern NMR cryomagnets is $\sim$ 70 mm inside diameter (ID)], but very high magnetic field homogeneity, routinely $> 10^{-9}\,\boldsymbol{B}_0$.

In the early 1970s, a revolutionary concept emerged from the pioneering work of Lauterbur in the United States and Mansfield, Andrew, and Hinshaw in the United Kingdom. They discovered that by using judiciously designed magnetic field gradients it was possible to retrieve an NMR signal from a small localized volume (called a voxel) within a much larger sample (e.g., a human body). This started a new field of NMR applications, called magnetic resonance imaging (MRI) that greatly enhanced the practice of diagnostic medicine (see the section **Magnetic Resonance Imaging**).

One of the frustrating limitations of MRI applications was the ambiguity of lesion characterization. The development of MRI focused on the noninvasive visualization of soft tissues within the living human body; as a result, the technical and engineering trade-offs made in the process of improving the quality of images have essentially rendered the method nonquantitative. In essence, MRI proved to be extremely sensitive in the detection of various lesions within the patient's body, but not very robust in providing information needed to fully identify the characteristics of the tissue within the lesion. Thus, in addition to basic NMR tissue characteristics (proton density, *T1*, and *T2*), the interpreters of medical MR images came to rely on morphology of lesions (size, shape, and location) to draw conclusions about lesion pathology. In this context, the concept of localized NMR spectroscopy experiments, where MRI techniques are used to locate the lesion and localize the volume of interest, while NMR spectroscopic techniques are used to acquire NMR spectra of the tissue within a lesion, becomes intuitively

evident. However, while the concept may appear naturally obvious, implementations have proven to be extremely difficult. Despite first successful experiments in acquiring localized phosphorus ^{31}P NMR *in vivo* spectra from a human forearm, performed in 1980 by a group led by Chance, the true clinical applications of localized NMR spectroscopy have only recently begun to appear. While first attempts focused on ^{31}P NMR spectroscopy using surface coils to localize the signals within the human body, current clinical applications almost exclusively utilize ^{1}H NMR spectra to gain additional, clinically relevant information about the lesion of interest. The methodology used in this approach is referred to as magnetic resonance spectroscopy (MRS), magnetic resonance spectroscopic imaging (MRSI), or chemical shift imaging (CSI). Techniques of this particular application of the NMR phenomena in medical practice will be the subject of further discussion here. While the interest in exploring clinical applications of MRS of nuclei other than protons (e.g., ^{31}P, ^{13}C, ^{19}F, and ^{23}Na) still remains, a vast majority of current clinical applications uses ^{1}H MRS and thus only this particular nucleus will be considered in further discourse. Readers interested in other nuclei are encouraged to explore literature listed in the ***Reading List*** section.

THEORY

In this section, the quantum mechanical approach of formalism is used, since this formalism is most naturally suited to explain the various features of NMR spectra. To begin with, consider an ensemble of noninteracting protons, free in space where a strong, perfectly uniform magnetic field $\boldsymbol{B}_0$ is present. Because all spins are considered identical, the Hamiltonian of the system includes a single spin and all quantum mechanical expectation values are calculated over the entire assembly of spins. Under those conditions, the Hamiltonian ($\mathcal{H}$) describes the Zeeman interaction of the nuclear magnetic moment μ with the external magnetic field and has a form

$$\mathcal{H} = -\boldsymbol{\mu} \times B_0 = -g\hbar B_0 I_z \tag{1}$$

where γ is the gyromagnetic ratio, $\hbar$ is the Planck's constant, and I_z is the z component of the nuclear spin operator $\boldsymbol{I}$, which for protons has an eigenvalue value of ½. Because I_z has only two eigenvalues, $\pm$ ½, the system's ground energy level is split into two sublevels, with the energy gap proportional to $\boldsymbol{B}_0$, as shown in Fig. 1. Now assume that spins are allowed to weakly interact with their molecular environment, which are collectively described as the *lattice* (regardless of the actual state of the sample; e.g., in liquids the lattice refers to thermal diffusion, both rotational and translational, of the atoms or molecules that host the spins). When the system is left undisturbed over time, it will reach a thermal equilibrium, where the spin populations at the higher and lower energy levels are described by a Boltzmann distribution, as shown in Fig. 2. The resultant sample equilibrium magnetization is equal to

$$M_z^{\mathrm{eq}} = M_0 = \frac{g^2\hbar^2 N B_0}{4kT} \tag{2}$$

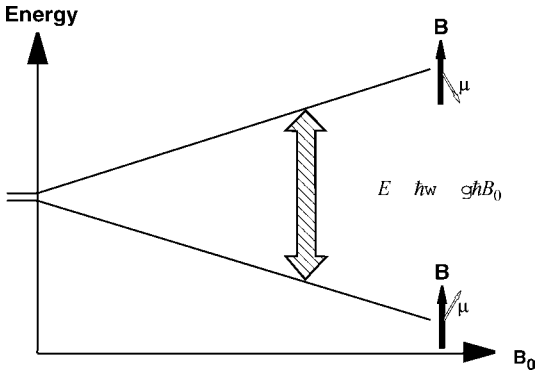


Figure 1. The Zeeman splitting of the ground-state energy levels for the spin ½ system as a function of the external magnetic field strength, B_0.

where T is the absolute temperature of the sample, N is a number of spins in the sample, and k is the Boltzmann constant. At normal conditions, this equilibrium magnetization is too small to be detectable, but when a resonance phenomenon is exploited by applying a short burst of RF energy at resonance frequency ω_0 (called an RF pulse), the magnetization can be flipped onto a transverse plane, perpendicular to the direction of $\boldsymbol{B}_0$. This transverse magnetization will precess at the resonant frequency of the spins and thus will generate an oscillating magnetic field flux in the receiver coils of the NMR apparatus, which will be detected as a time-varying voltage at the coils terminals. This signal is called a free induction decay (FID) and its time evolution contains information about the values of resonant frequency of the spins, ω_0, the spin-spin relaxation time, $T2$, and the distribution of local static magnetic fields at the locations of the spins, $T2^*$. The local static magnetic fields, experienced by spins at different locations in the sample, may vary from spin site to spin site, chiefly due to the inhomogeneity of the main magnetic field B_0. In addition, in heterogeneous samples, commonly encountered in *in vivo* experiments, local susceptibility variations may contribute to $T2^*$ effects. For a variety of reasons, the chief one being the ease of

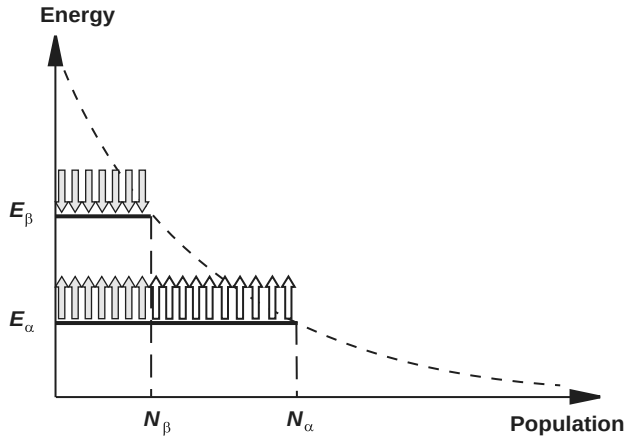


Figure 2. An illustration of Boltzmann distribution of spin populations for an ensemble of identical spins ½, weakly interacting with the lattice, at thermal equilibrium.

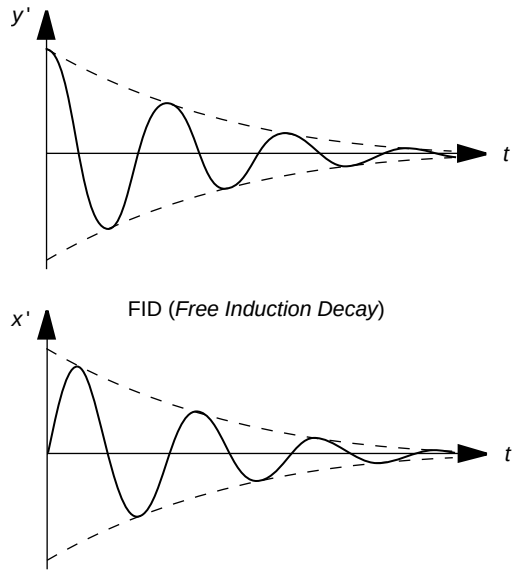


Figure 3. Real and imaginary components of an FID signal.

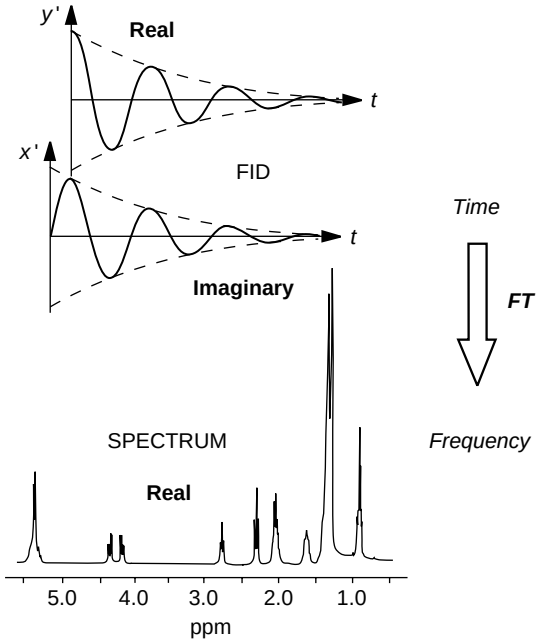


Figure 4. A method of generating an NMR spectrum from the FID signals.

interpretation, the FID signal is always recorded in the reference frame that rotates at the frequency ω close to ω_0, (called "the rotating frame"). From the engineering point of view, recording the signal in the reference frame, rotating at the frequency ω, is equivalent to frequency demodulation of the signal by frequency ω.

The recorded FID has two components: one, called real, or in-phase, is proportional to the value of transverse spin magnetization component aligned along the y' axis of the rotating frame (the [x', y', z] notation is used to denote rotating frame, as opposed to the stationary, laboratory frame [x, y, z]). The other FID component is proportional to the value of transverse spin magnetization projected along the x' axis and is referred to as imaginary, or out-of-phase signal (see Fig. 3). To a human being, the FID signals can be difficult to interpret; thus an additional postprocessing step is routinely employed to facilitate data analysis. A Fourier transform (FT) is applied to the FID data and the signal components having different frequencies are retrieved, producing an NMR spectrum (see Fig. 4).

The chemical shift, mentioned earlier, is responsible for a plethora of spectral lines (peaks) seen in a typical NMR spectrum. Consider a simple organic molecule that contains hydrogen localized at three nonequivalent molecular sites. Chemists call molecular sites equivalent if the structure of chemical bonds around the site creates an identical distribution of electron density at all proton locations. When the sites are nonequivalent, different distributions of electron cloud around the protons will have a different shielding effect on the value of the local magnetic field, experienced by individual protons. The strength of electron shielding effect is proportional to the value of B_0 and is accounted for in the Hamiltonian by using a shielding constant, σ.

$$\mathcal{H} = -g\hbar \sum_{i=1}^{3} (1 - s_i) B_0 I_{zi} \qquad (3)$$

In this example, a molecule with only three nonequivalent proton sites has been considered; in general, the summation index in Eq. (3) must cover all nonequivalent sites, however, many may be present. It is evident from Eq. (3) that individual protons located at different sites will have slightly different resonant frequencies, which will give rise to separate resonant peaks located at different frequencies in the observed NMR spectrum, as shown in Fig. 4. This accounts for a fine structure of NMR spectra that consists of multiple lines at different frequencies, identifying all nonequivalent molecular sites in the sample studied.

The interaction of the nuclear spin with the electron cloud surrounding it has a feedback effect, resulting in a slight distortion of the cloud; the degree of this alteration is different depending on whether the spin is up or down w.r.t. the magnetic field B_0. This distortion has a ripple effect on surrounding nonequivalent spins, and consequently they become coupled together via their interactions with the electron cloud; this phenomenon is called a spin–spin coupling or J coupling, and is accounted for by adding another term to the spin Hamiltonian:

$$\mathcal{H} = -\sum_{i=1}^{n} \left[g\hbar (1 - s_i) B_0 I_{zi} + \sum_{j<i} J_{ij} I_i \times I_j \right] \qquad (4)$$

where J_{ij}, known as a spin–spin coupling constant, describes the strength of this effect for each pair of nonequivalent protons. The presence of spin–spin coupling leads to a hyperfine structure of the NMR spectra, splitting peaks into multiplet structures, as shown in Fig. 5 that contains two fragments of an NMR spectrum of lactic acid. The structure of each multiplet can be explained using simple arrow diagrams, visible next to NMR lines. The signal at 1.31 ppm is generated by 3 equiv protons located in the CH_3 group that are linked to the proton spin in the CH group via J coupling. The spin of the proton in the CH group can have only two orientations: up or down, as indicated by arrows

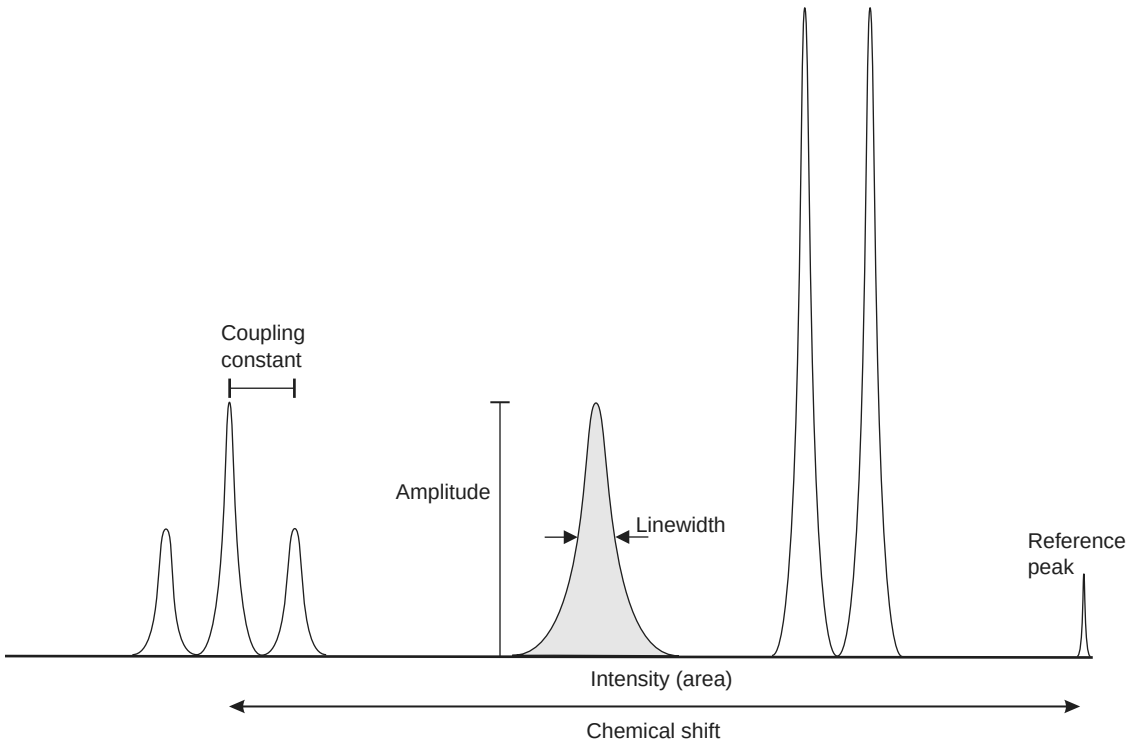


Figure 5. A fragment of experimental spectrum (lactic acid at 500 MHz) showing resonances from a CH_3 group (at 1.31 parts per million, ppm) and a CH group (at 4.10 ppm), respectively. The splitting of CH_3 resonance into a doublet and the CH resonance into a quadruplet is caused by the J coupling. The structure of each multiplet can be derived using simple rule of grouping spins according to their orientations, as shown by groups of arrows.

that follow the bracket next to the CH_3 label. Thus, the signal from CH_3 protons is expected to split into a doublet with relative signal intensities 1:1, as indeed is seen in the recorded spectrum. Similarly, the signal at 4.10 ppm is generated by a single proton in the CH group that is linked via J coupling to 3 equiv protons in the CH_3 group. The spins within this group of three protons can assume eight different configurations, depending on their orientation w.r.t. to the magnetic field $\boldsymbol{B}_0$. Some of those orientations are equivalent (i.e., they have the same energy) and thus can be lumped together, as shown by groups of arrows that follow the bracket next to the CH label. Simple counting leads to a prediction that the signal from the CH proton should split into a quadruplet with relative signal intensities 1:3:3:1. Again, this is clearly visible in the recorded spectrum.

As illustrated in Fig. 6, these simple considerations show that each peak in the spectrum can be fully characterized by specifying its position w.r.t. an established reference peak (chemical shift), amplitude, intensity (intensity, or the area under the peak, is proportional to the concentration of spins contributing to the given peak), linewidth (provides information $\sim T2$ and magnetic field homogeneity), and multiplet structure (singlet, doublet, triplet, etc., carries information about J coupling). Thus, the NMR spectra, like the one in Fig. 4 showing an experimental spectrum of vegetable (maize) oil, contain a wealth of information about the structure and conformation of molecules found within the sample.

Strictly speaking, the linewidths of the peaks in the NMR spectrum are determined by the values of $T2^*$ and thus are sensitive to the homogeneity of the main magnetic field and other factors contributing to the distribution of local static magnetic fields seen by the spins. Wider distributions of local fields, lead to shorter $T2^*$ values and broader corresponding peaks in the NMR spectrum. Broad peaks make spectra harder to interpret due to overlap between peaks located close to each other. This feature puts a premium on shimming skills of the NMR spectrometer operator (shimming includes methods to improve the homogeneity of the static magnetic field). For *in vivo* studies, the shimming tasks are made even more difficult by tissue heterogeneity that causes local variations in the magnetic field (referred to as susceptibility effects within the MRS community).

There is a peculiar feature in the way the NMR spectra are recorded that routinely confounds the NMR newbies. For historical reasons, the NMR spectra are plotted as if the spectrometer frequency was kept constant and the magnetic field B_0 was varied (swept) over a range of values to record all the characteristic peaks (see Fig. 7a). In this approach, the horizontal axis of the plot represents the values of the external magnetic field necessary to reach a resonant condition for a given group of spins. However, all modern NMR spectrometers use an acquisition method in which the magnetic field B_0 is kept constant and differences in resonant frequencies of various nuclei (due to their

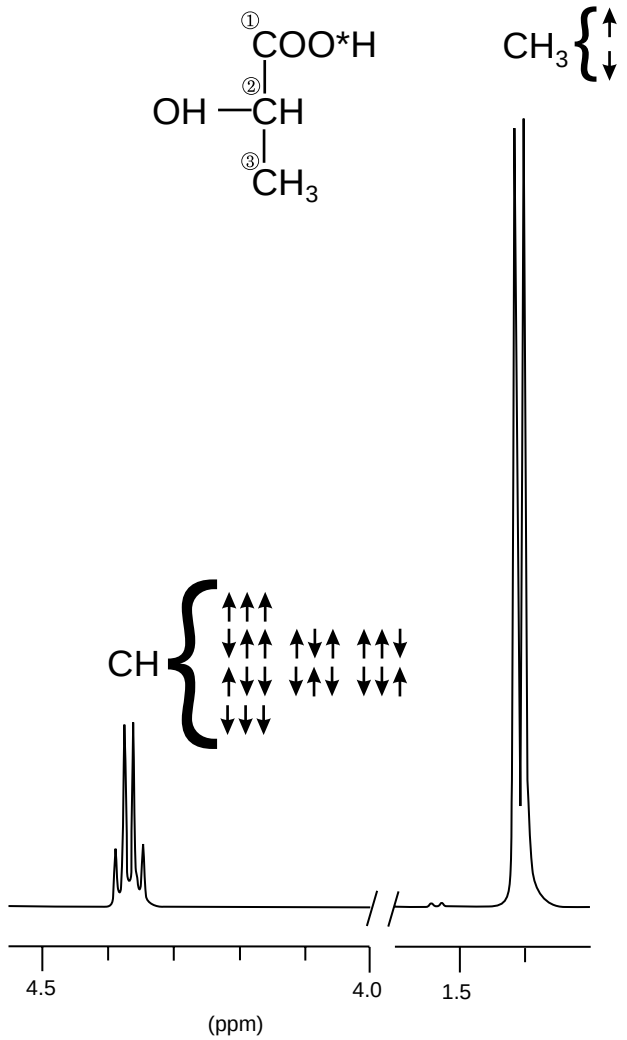


Figure 6. A schematic representation illustrating various characteristic parameters used to describe elements of an NMR spectrum.

diverse chemical shifts) are recorded instead. It is fundamental to understand that if the horizontal axis of an NMR spectrum was meant to represent the resonant frequencies of nuclei with different chemical shift, all residing in the same external magnetic field B_0, then the lines to the right would represent signals with *lower* frequencies, which at best is counterintuitive. This paradox is resolved when one looks at the relationship between magnetic field and NMR resonance frequency for nuclei located at sites with varying electron shielding. This relationship is derived from Eq. (3), leading to the well–known Larmor equation:

$$w = g(1 - s)B_0 \tag{5}$$

This equation is plotted in Fig. 7b for three sites with different values of the chemical shielding constant σ. As σ increases, the slope of the line decreases. Thus, if the RF frequency is kept constant and the magnetic field is swept to reach subsequent resonant conditions, the weakly shielded nuclei will resonate at lower field, and as the strength of the shielding effect increases, the external magnetic field must be increased to compensate for the increased shielding. It must be reiterated that the term magnetic field used in this context refers to the *external* magnetic field B_0, produced by the spectrometer's magnet, and not to the value of the local magnetic field that the spin is actually experiencing. Therefore, the signals from heavily shielded nuclei will appear at the high end of the spectrum, as illustrated by the horizontal line in Fig. 7b. On the other hand, if the external magnetic field B_0 is kept constant and the frequency content of the FID signal is looked at, it will be noticed that nuclei at heavily shielded sites resonate at lower frequency, which reflects the decrease of the *local* magnetic field due to the shielding effects. Therefore, the signals from heavily shielded nuclei will appear at the low end of the spectrum, as illustrated by the vertical line in Fig. 7b. Historically, magnetic field sweeping was used in the early days of NMR spectroscopy and a sizeable volume of reference spectra were all plotted using the fixed-frequency convention. This standard was retained after pulsed NMR technology replaced the earlier continuous wave NMR spectroscopy, despite the fact that the fixed-field approach would have been far more natural.

The size of the chemical shift varies linearly with the strength of the magnetic field B_0, which makes comparison of spectra acquired with NMR spectrometers working at different field strengths a chore. To simplify matters, chemists introduced a concept of relative chemical shift, which is defined as follows: It is realized that the term ω in Eq. 5 represents a resonant frequency of a group of equivalent spins in their local magnetic field, B_L. Thus, Eq. (5) can be used to define a variable τ as

$$\tau = s \times 10^6 = \frac{B_L - B_0}{B_0} \times 10^6 \tag{6}$$

The value of τ is dimensionless and expresses the value of the shielding constant σ in ppm. It is interpreted as a change in the local magnetic field *relative* to the strength of the main magnetic field produced by the spectrometer's magnet. As seen from Eq. (6), the τ scale is directly proportional to σ, that is, heavily shielded nuclei will have a large value of τ (see Fig. 7a). This also makes the τ scale collinear with the B_0 axis, which is inconvenient in modern NMR spectroscopy, which puts a heavy preference on spin resonant frequencies. To address this awkward feature, chemists use a more practical chemical shift scale, called δ, that is defined as

$$d = 10 - \tau \tag{7}$$

This scale is a measure of the change in local resonant frequency relative to the base frequency of the NMR spectrometer, ω_0. The factor 10 in the above relationship arises from the fact that a vast majority of observed proton chemical shifts lie in the range of 10 ppm; by convention, tetrametylsilane (TMS), which exhibits one of the strongest shielding effects, has been assigned a value of $\delta = 0$. This was done after careful practical consideration: all 12 protons in TMS occupy equivalent positions, and thus an NMR spectrum of TMS consists of a strong, single line. The referencing procedure has evoked a considerable amount of

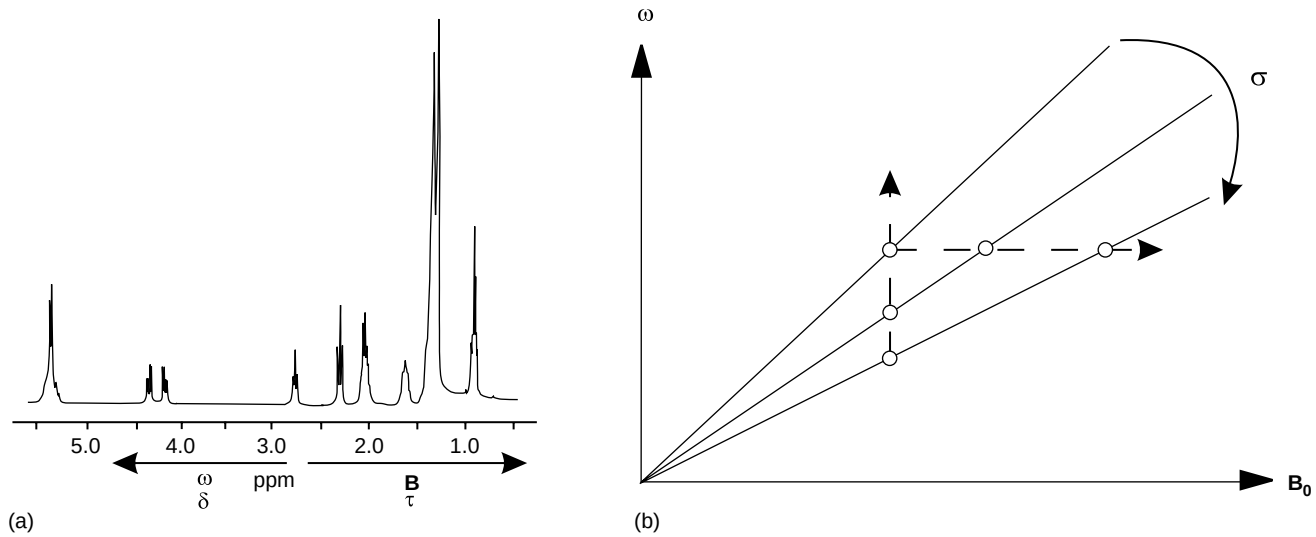


Figure 7. An apparent paradox in the interpretation of the abscissa calibration for an NMR spectrum: (a) the same spectrum can have a value of magnetic field B assigned to the abscissa so that the values of B increase when moving to the right, or have the value of frequency ω assigned to the same abscissa, but increasing when moving to the left. (b) The diagram on the right provides an explanation of this effect (see the main text for details).

debate over the years and currently the International Union of Pure and Applied Chemistry (IUPAC) specifies that shields are to be reported on a scale increasing to high frequencies, using the equation

$$d = \frac{w_x - w_{\text{ref}}}{w_{\text{ref}}} \times 10^6 \tag{8}$$

where ω_x and ω_{ref} are the frequencies of the reported and reference signals, respectively.

Since TMS easily dissolves in most solvents used in NMR spectroscopy, is very inert, and is quite volatile, it is a very convenient reference compound. In practice, a small amount of TMS can be added to the sample, which will produce a well-defined reference peak in the measured spectrum (see Fig. 5). After the experiment is completed, TMS is simply allowed to evaporate, thus reconstituting the original composition of the sample.

In MRS applications, the δ scale is used exclusively to identify the positions of the peaks within the spectra. For example, the water peak is known to have $\delta = 4.75$ ppm; therefore, if an MRS spectrum is acquired on a machine with the main magnetic field of 1.5 T and the base RF frequency of 63.86 MHz, the water line will be shifted by $4.75^{*}63.86 = 303$ Hz toward the higher frequency from the reference TMS peak. It also means that the protons in water experience weaker shielding effects than protons in the TMS. Finally, if the spectrum is plotted according to the accepted conventions, the water line appears to the left of the reference peak. Unfortunately, TMS cannot be used as an internal reference in MRS applications (it cannot be administered to humans). Thus in practice, the signal from NAA is used as a reference and has an assigned value of $\delta = 2.01$ ppm, which is the chemical shift of the acetyl group within the NAA NMR spectrum, acquired *in vitro*.

EQUIPMENT AND EXPERIMENTS

In medical applications, standard MRI equipment is used to perform MRS acquisitions. Thus, in contrast to standard *in vitro* NMR experiments, *in vivo* MRS studies are performed at lower magnetic field strength, using larger RF coils, and with limited shimming effectiveness due to magnetic susceptibility variations in tissue. As a result, the MRS spectra inherently have lower signal-to-noise characteristics than routine *in vitro* spectra; this is further aggravated by the fact that in MRS signal averages are accumulated using fewer scans due to examination time constraints. This creates a new set of challenges related to the fact that when the MRI equipment is used to perform MRS, it is utilized outside its design specifications that focus on the imaging applications of MR technology. Fortunately, many hardware performance characteristics that are absolutely crucial to the successful acquisition of spectroscopy data, such as magnetic field uniformity and stability, or coherence and stability of the collected NMR signals, are appreciated in MRI as well. Thus, steady improvements in MRI technology are contributing to the emergence of clinical applications of MRS. Since this article focuses on MR spectroscopy, the following considerations will describe features of data acquisition schemes that are unique to MRS applications, and disregard those aspects of hardware and pulse sequences design that form the core of the MRI technology (see the section on **Magnetic Resonance Imaging**).

The first challenge of MRS is volume localization. Obviously, an NMR spectrum from the entire patient's body, while rich in features, would be of very little clinical utility. Over the years, many localization techniques have been proposed, but in current clinical practice only two

methods are routinely used. The first one collects NMR spectra from a single localized volume and is thus referred to as single voxel MRS, or simply MRS. The other allows collection of spectra from multiple voxels arranged within a single acquisition slab. It is often referred to as multi voxel MRS, MRS imaging (MRSI), or chemical shift imaging (CSI). With this method, more advanced visualization techniques can be used, such as generation of specific metabolite concentration maps over larger regions of interest (ROI).

Single voxel (SV) MRS is simpler to implement. The volume of interest (VOI, or voxel) is selected using one of the two alternative data acquisition pulse sequences. The first one uses a phenomenon known as a stimulated echo to produce a signal used to generate the spectroscopic FID that is then transformed into the NMR spectrum. To produce a stimulated echo, three RF pulses are used, each rotating (flipping) the magnetization by 90°. The theory of this process is too complex to be discussed in detail here; an interested reader is referred to the original paper by Hahn (6) or to more specific texts on NMR theory, such as those listed in the ***Reading List*** section. The application of the stimulated echo method for *in vivo* MRS was first proposed by Frahm et al. (7) who coined an acronym STEAM (**S**timulated **E**cho **A**cquisition **M**ode) to describe it. A simplified diagram of the stimulated echo MRS acquisition pulse sequence is shown in Fig. 8. The three RF pulses are shown on the first line of the diagram, the echo signal from the localized voxel can be seen on the bottom line of the diagram. How does this sequence allow the selection of a specific VOI as a source of collected signal? The key to successful VOI localization is to use a slice-selective RF excitation. This technology is taken straight from mainstream MRI, and thus the reader is referred to MRI-specific information for further details (see the section **Magnetic Resonance Imaging** or MRI monographs listed in the ***Reading List***). Briefly, the slice selection technique relies on the use of band-limited RF pulses (notice the unusual modulation envelope of RF pulses shown in Fig. 8) applied in the presence of tightly controlled magnetic field gradients (MFG), shown as pulses labeled G_x, G_y, and G_z in Fig. 8. When a band-limited RF pulse is played in the presence of an MFG, only the spins in a narrow range of positions, located within a thin layer of the studied object (a slice) will achieve the resonance conditions and respond accordingly; all the spins outside the slice will be out of resonance and remain unaffected. Thus, the spin magnetization within the selected slice will be flipped by the RF, and magnetization outside the slice will remain unaffected. An analysis of this process shows that the slice orientation is perpendicular to the direction of the MFG, slice thickness is controlled by the amplitude of the MFG pulse, and location (offset from the magnet's isocenter) is determined by the shift in carrier frequency of the RF pulse. Thus, the first pulse in Fig. 8 will excite spins in a slice that is perpendicular to the z axis of the magnet (G_z was used), the two remaining pulses will excite spins in slices perpendicular to the x and y directions, respectively. Since the condition required to produce a stimulated echo is that the spins be subject to all three pulses, only the matter located at the intersection of the three perpendicular slices fulfills the criterion, and thus only the spins located within this volume will generate the stimulated echo signals. The dimensions of the VOI selected in such a way are determined by the thickness of individual slices, and the location of the VOI is determined by the position of individual slices, as illustrated in Fig. 9.

The other single voxel MRS protocol uses a spin echo sequence to achieve volume selection. To produce the localized signal, three RF pulses are used, as before. However, while the first RF pulse still rotates the magnetization by 90°, the remaining two RF pulses flip the magnetization by

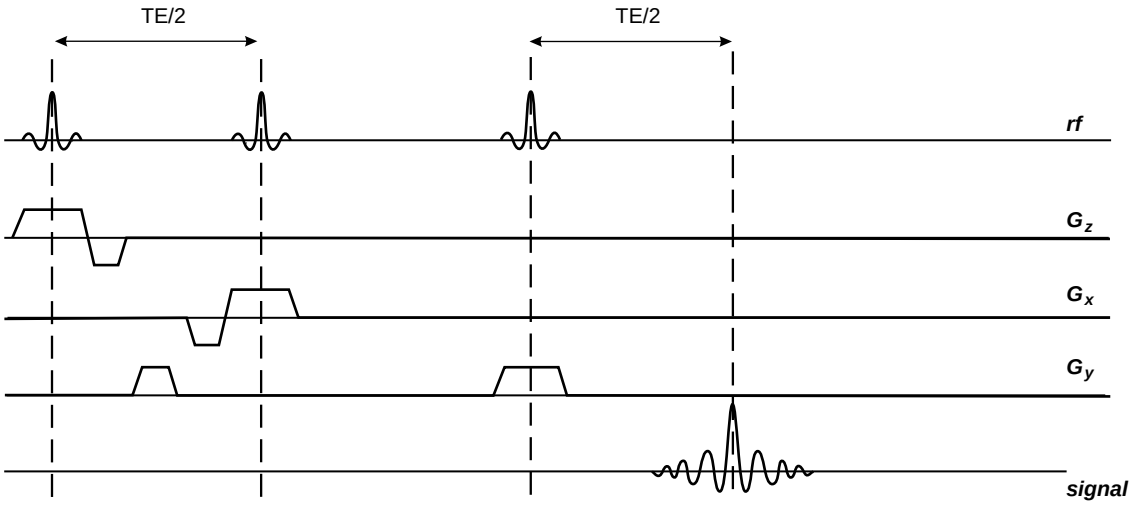


Figure 8. A simplified diagram of a basic STEAM MRS sequence. Time increases to the right, the time interval covered by this diagram is typically ~ 100 ms. On the first line the RF pulses are shown; the next three lines show the timing of the pulses generated by the three orthogonal magnetic field gradient assemblies (one per Cartesian axis in space); the last line, labeled signal, shows the timing of the resulting stimulated echo NMR signal.

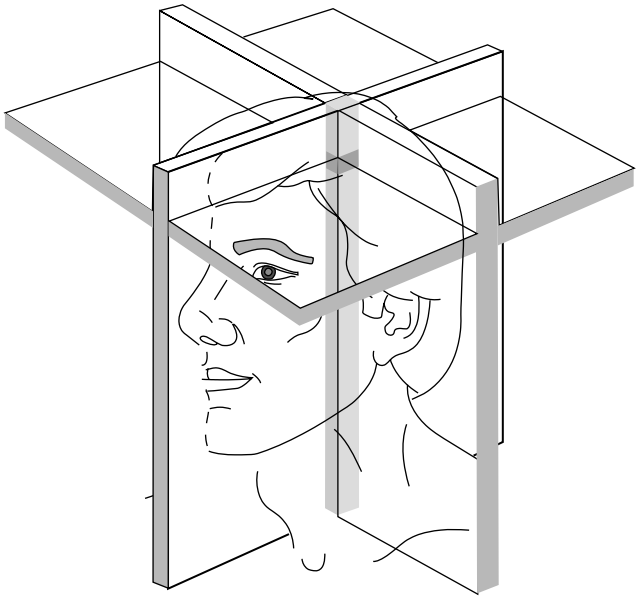


Figure 9. The principle of VOI selection in STEAM protocol (see text for details).

180° each. As a result, two spin echoes are produced, as shown on the bottom line of Fig. 10. The first echo contains a signal from a single column of tissue that lies at the intersection of the slices planes generated by the first and second RF pulses, while the second echo produces a signal from a single localized VOI within this column. As with the stimulated echo, the theory of this process is too complex to be discussed here; an interested reader is referred to the original paper by Hahn (6) or to more specific texts on NMR theory, such as those listed in the ***Reading List*** section.

The application of dual echo spin echo method to *in vivo* MRS has been proposed by Bottomley (8) who coined an acronym **P**oint **R**esolved **S**pectroscopy (PRESS) to describe it. A simplified diagram of the PRESS acquisition pulse sequence is shown in Fig. 10. The three RF pulses are shown on the first line of the diagram; but only the second echo signal seen on the bottom line of the diagram comes from the localized voxel region.

What are the advantages and weaknesses of each volume localization method? First, in both methods an echo is used to carry the spectroscopy data. Since echoes are produced by transverse magnetization, the echo amplitudes are affected by *T2* relaxation and the time delay, *TE*, that separates the center of the echo from the center of the first RF pulse that was used to create the transverse components of magnetization. The longer the *TE*, the stronger the echo amplitude attenuation due to *T2* effects will be. Since the amplitude of the echo signal determines the amplitude of the spectral line associated with it, the SV MRS spectra will have lines whose amplitudes will depend on the selected *TE* in the acquisition sequence. For the STEAM protocol, it is possible to use relatively short TE values; mostly because the magnetization contributing to the stimulated echo signal is oriented along the *z* axis during the period of time between the second and third RF pulses, and thus it is not subject to *T2* decay (in fact, it will grow a little due to *T1* relaxation recovery). Therefore, the time interval between the second and the third RF pulse is not counted toward TE. Furthermore, 90° RF pulse have shorter duration than 180° ones, offering an additional opportunity to reduce TE. Consequently, in routine clinical applications the STEAM protocols use much shorter TE values (∼ 30 ms) than PRESS protocols (routine TE values used are ∼ 140 ms). Therefore, when using the

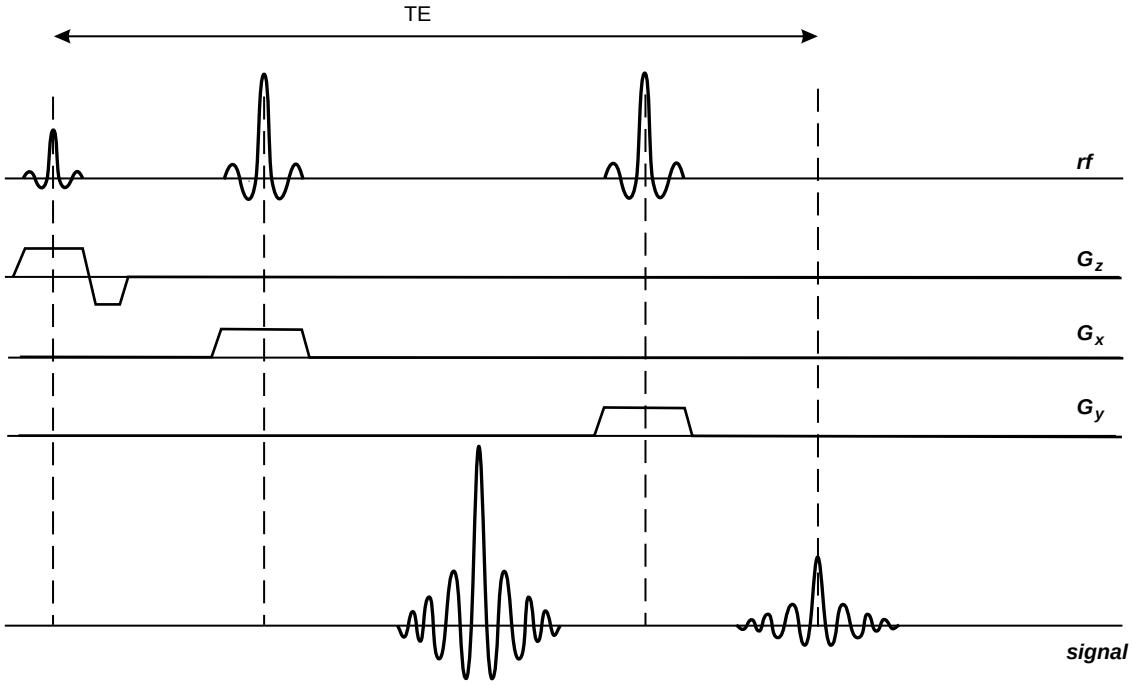


Figure 10. A simplified diagram of a basic PRESS MRS sequence. Notice similarities in VOI selection algorithm within both STEAM and PRESS protocols.

STEAM protocol one is rewarded with a spectrum whose line intensities are closer to the true metabolite concentrations in the studied tissue. However, theoretical calculations show that the signal intensity of a stimulated echo is expected to be 50% less than that of a spin echo generated under identical timing conditions (i.e., *TE*). This is an inherent drawback of the STEAM protocol, since the spectra tend to be noticeably noisier than those produced using PRESS.

PRESS, by design, uses two spin echoes to generate a signal from localized VOI. This limits the minimum *TE* possible for the second echo to $\sim$ 80 ms or so, depending on the MR scanner hardware. Since the magnetization never leaves the transverse plane (except during RF pulses), the *T2* effects are quite strong. As a result, metabolites with shorter *T2* will decay down to the noise levels and vanish from the final spectrum, which has an ambivalent impact on clinical interpretations, simplifying the spectrum on one hand while removing potentially valuable information on the other. This effect can be further amplified by signal intensity oscillations with varying TE, caused by *J* coupling. For further details on this topic the reader is referred to the specific texts on MRS theory and applications, listed in the ***Reading List*** section.

One word of caution is called for now. The numbers quoted here reflect capabilities of MR hardware that represent a snapshot in time. As hardware improves, these parameters rapidly become obsolete.

There are other, finer arguments supporting possible preferences toward either method (STEAM or PRESS). These include such issues as suppression of parasitic signals arising from outer-volume excitation (residual signals coming from outside the VOI), sensitivity to baseline distortion of the spectra due to the use of solvent suppression, accuracy of VOI borders defined by each method, and so on. Thus, in current clinical practice there are strong proponents of both STEAM and PRESS techniques, although lately PRESS seems to be gaining popularity because of simpler technical implementation issues, such as magnetic field homogeneity correction (shimming), or compensation of effects caused by eddy currents induced in the magnet cryostat by magnetic field gradient and RF pulses.

While SV MRS is relatively straightforward, and thus preferred by novices in the practice of clinical MRS, most routine applications today demand spectroscopic data collected from multiple locations within the organ studied. Therefore, a solution that would allow collection of MR spectra from multiple locations at the same time has a natural appeal to physicians. To achieve such a task is no small matter, and many schemes have been proposed before a method that today is considered most practical in daily use has been found. The method was first proposed by Brown et al. (9). Their method is both conceptually simple and revolutionary. It utilizes a method that encodes both space–(localization) and time–dependent (NMR spectra) information using mechanisms that manipulate the phases of signals emitted by individual spins at different locations. The acquisition sequence is schematically shown in Fig. 11, which illustrates the two-dimensional (2D) CSI principle using a PRESS pulse sequence. Of course, this approach is equally applicable to STEAM method as well. An astute reader will immediately recognize that the spatial encoding part of the protocol is virtually identical to dual phase encoding techniques used in 3D MRI acquisitions. At this point, readers less familiar with advanced MRI techniques probably would want to read more about this method elsewhere (see **Magnetic Resonance Imaging** or MRI monographs listed in the ***Reading List***). The enlightenment occurs when one realizes that with this scheme the acquired signal is not frequency encoded at all. Therefore, after 2D FT processing in the two orthogonal spatial directions, one ends up with a collection

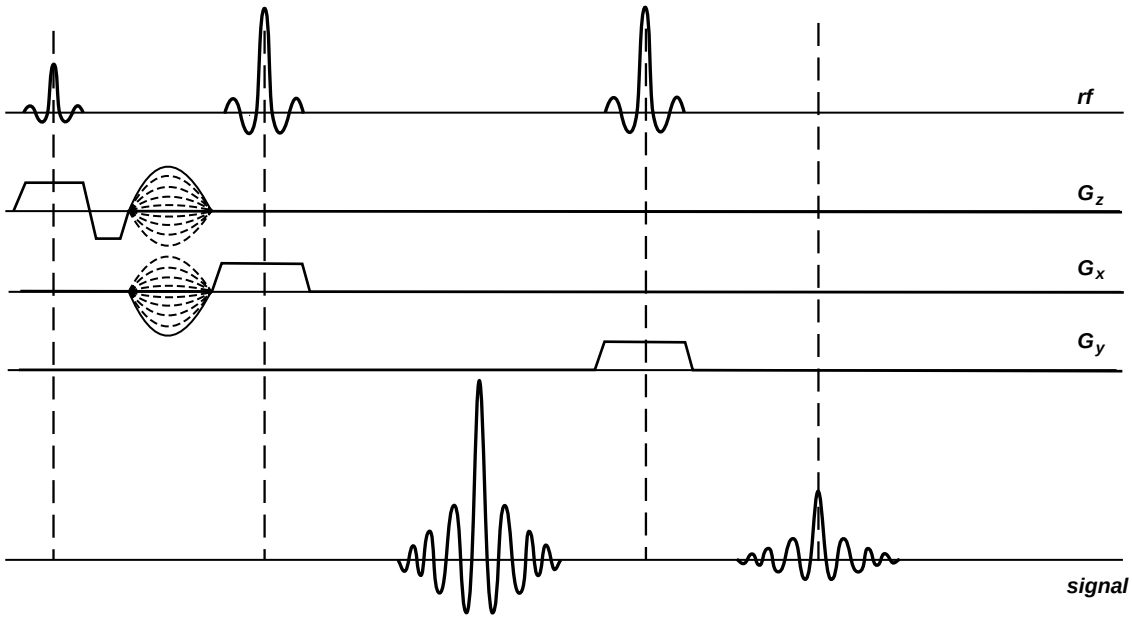


Figure 11. A simplified diagram of a basic 2D CSI MRS sequence. The dotted gradient lobes represent phase encoding gradients responsible for multivoxel localization.

of FIDs that are free from any residual phase errors due to spatial encoding, but nevertheless represent NMR signals from localized VOIs, anchored contiguously on a planar grid. The size and orientation of the grid is determined by the spatial encoding part of the protocol; thus, localization is conveniently decoupled from NMR frequency beats caused by the chemical shifts of studied metabolites.

If this method is so simple, why do people still want to acquire SV spectra? First, the method is quite challenging to implement successfully in practice, despite the seeming simplicity of the conceptual diagram shown in Fig. 11. Both spatial and temporal components affect the phase of collected NMR signals, and keeping them separated requires a great deal of MR sequence design wizardry and the use of advanced MR hardware. Due to peculiarities of the FT algorithm, even slight phase errors (of the order of 1°) are capable of producing noticeable artifacts in the final spectra. Second, the VOI localization scheme is ill suited to a natural way of evaluating lesions spectroscopically; most clinicians like to see a spectrum from the lesion compared to a reference spectrum from a site that is morphologically equivalent, but otherwise appears normal. In the human brain, where most spectroscopic procedures are performed today, this means that the reference spectrum is acquired contralaterally to the lesion location. Such an approach represents a drawback in CSI acquisitions where VOIs are localized contiguously, and typically a few wasted voxels must be sacrificed to ensure the desired anatomic coverage of the exam. Finally, the dual-phase encoding scheme requires that each pulsed view (a single execution of the pulse sequence code with set values of both phase encoding gradients) is repeated many times to collect enough data to localize voxels correctly. As many views must be acquired as there are voxels in the grid, which causes the required number of views to grow very fast. For example, even for a modest number of locations, say 8*8, 64 views must be generated. This leads to acquisition times that appear long by today's imaging standards (typical MRSI acquisition requires 3–8 min).

It is difficult to fully exploit the richness and diversity of technical aspects of localization techniques used in *in vivo* MRS; extended reviews, such as papers by Kwock (10), den Hollander et al. (11), or Decorps and Bourgeois (12), can be used as springboards for further studies.

The second major challenge of *in vivo* ^{1}H MRS arises from the fact that the majority of tissue matter is simply water, which for humans can vary from 28% in the bones to 99% in CSF, with an average of $\sim$ 75–80% in soft tissues (e.g., brain matter, muscle, lung, or liver). Thus, if a proton spectrum from just about any soft tissue (except adipose) is recorded, the result would look like the one presented in Fig. 12a, where an experimental spectrum from a muscle tissue of a young (and lean) rat, collected *ex vivo* on a 300 MHz NMR spectrometer, is shown. At a first glance, the result is boring: Only a single, strong peak from water is visible. Closer inspection, however, uncovers some additional details: First of all, the background of the spectrum is not totally flat, but composed of a broad peak, much wider than the normal range of chemical shifts expected in HR ^{1}H NMR spectra. This is illustrated in Fig. 12b, where the background of the spectrum in Fig. 12a has been enhanced by amplifying the background and cutting off most of the strong, narrow water peak. The broad spectrum is produced by protons in macromolecular components of the tissue: the proteins, DNA, RNA, and thousands of other compounds responsible for function and control of cellular activities. The spectrum broadening is caused by the residual dipolar interactions that were not fully averaged out because large molecules move more slowly than the small ones. *Note:* This component of the NMR signal is normally not visible in MRI and MRS applications; since the line is so broad, the relaxation time *T2* of this component is quite short (on the order of hundreds of μs), thus by the time the MR signals are collected at *TE* that are in the range of milliseconds, this signal has decayed out. One can see some small blips on the surface of the broad line in the Fig. 12b: Those are signals from tissue biochemical compounds whose molecules are either small enough, or have specific chemical groups that are free to move relatively fast due to conformational arrangements. These clusters of protons have *T2s* long enough to produce narrow lines, and their chemical environment is varied enough to produce a range of chemical shifts. In short, those little blimps form an NMR spectral signature of the tissue studied. As such, they are the target of MRS. To enhance their appearance, various solvent suppression techniques are used. In solvent suppression, the goal is to suppress the strong (but usually uninteresting signal) from solvent (in case of tissue, water), thus reserving most of the dynamic range of signal recorder for small peaks whose amplitudes are close to the background of a tissue spectrum. This point is illustrated in Fig. 12c, which shows a spectrum from the same sample as the other spectra in this figure, but acquired using a water suppression technique. Now, the metabolite peaks stand out nicely against the background, in addition to some residual signal from water peak (it is practically impossible to achieve 100% peak suppression).

The most common implementation of water suppression in MRS *in vivo* studies uses a method known as CHESS: **Ch**emical **S**hift **S**elective suppression, first proposed by Haase and colleagues at the annual meeting of the Society of Magnetic Resonance in Medicine in 1984. It consists of a selective 90° pulse followed by a dephasing gradient (homogeneity spoiling gradient, or homospoil). The bandwidth of the RF pulse is quite narrow, close to the width of the water line, and the carrier RF frequency offset is set to the water signal center frequency. When such a pulse is applied, only the water protons will be at resonance and they will flip by 90°, leaving magnetizations of all other protons unchanged. The resultant FID from the water signal is quickly dispersed by using the homospoil gradient. When the CHESS segment is followed by a regular spectroscopy acquisition sequence (STEAM or PRESS), the first RF pulse of those spectroscopic sequences will tip all the magnetizations from metabolites, but will not create any transverse magnetization from water. The reason is that at the time the spectroscopic acquisition routine starts, the longitudinal magnetization for water protons is zero, as prepared by the CHESS routine. The description of technical details of various solvent suppression methods can be found in the paper by van Zijl and Moonen (13). The side effect of solvent suppression is a baseline distortion of the resulting spectrum;

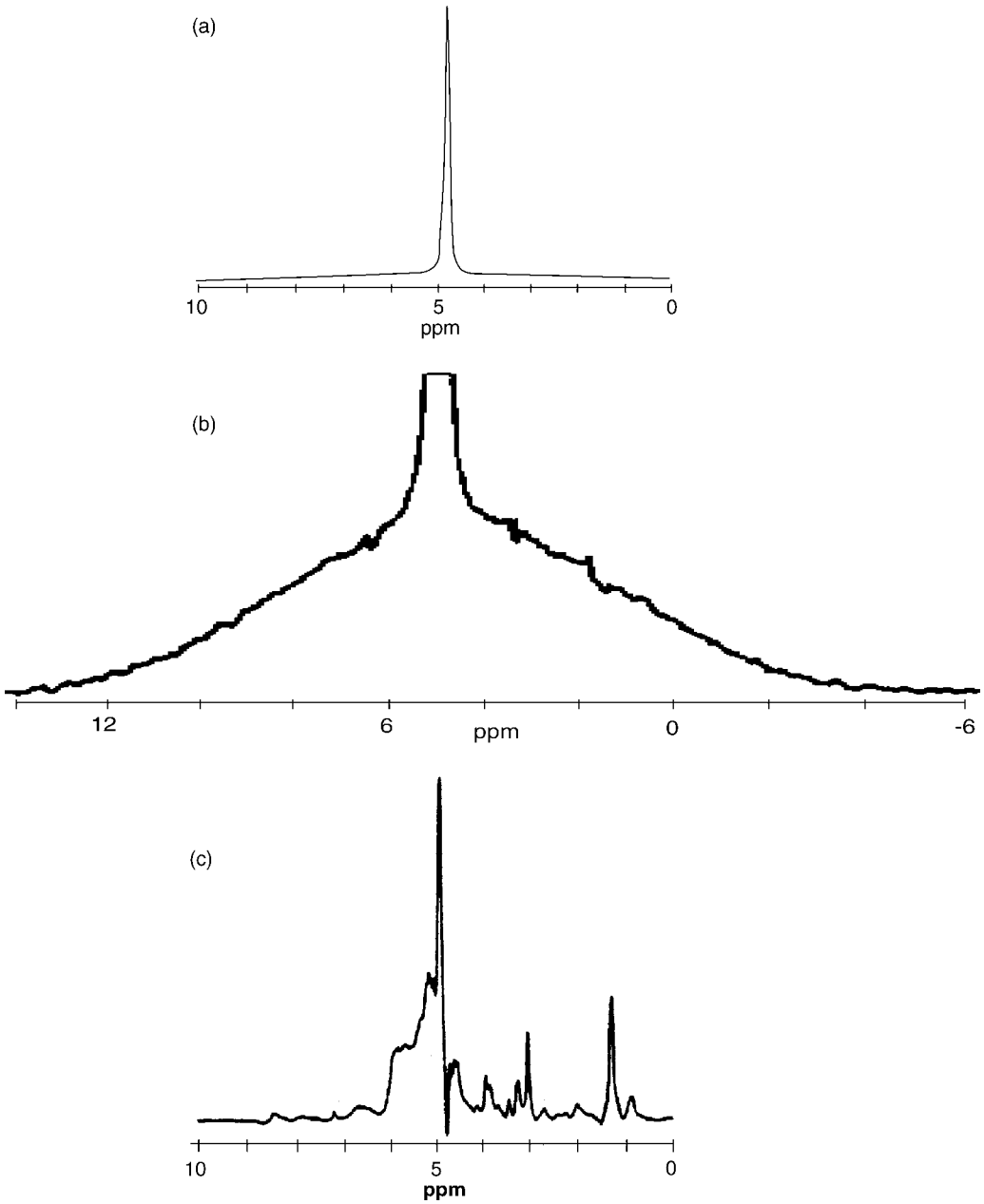


Figure 12. An example of NMR spectra obtained from the same sample (a lean muscle from a young rat's leg) collected *in vitro*: (a) standard spectrum obtained using a single RF pulse and performing an FT on the resulting FID; (b) the same spectrum scaled to reveal a wide, broad line generated by protons within macromolecules, notice small humps on top of this broad line: these are signals from small mobile macromolecules; (c) the high-resolution spectrum of the same sample, obtained using a spin–echo method and applying water suppression, both water and macromolecular peaks are suppressed, revealing small narrow lines that are subject to clinical MRS evaluations.

this distortion can be particularly severe in the vicinity of the original water peak, (i.e., at 4.75 ppm). To avoid difficulties associated with baseline correction, the MRS spectra in routine clinical applications are limited to the chemical shift range from 0 to $\sim$ 4.5 ppm.

In this short description of clinical MRS, the discussion had to be limited to its main features. There are many interesting details that may be of interest to a practitioner in the field, but had to be omitted here because of space constrains; the reader is encouraged to consult comprehensive reviews, such as the work of Kreis (14), to further explore these topics.

APPLICATIONS

Now that there is a tool, it is necessary to find out what can be done with it. The clinical applications of MRS consist of

three steps: first, an MR spectrum is acquired from the VOI; second, specific metabolite peaks are identified within the spectrum; and third, clinical information is derived from the data, mostly by assessing relative intensities of various peaks. The first step has been covered, so let us now look at step two. The easiest way to implement the spectral analysis is to create a database of reference spectra and perform either spectrum fitting or assign individual peaks by comparing the experimental data to the reference spectra. This is easier said than done; the reliable peak assignment in NMR spectra acquired *in vivo* has been one of the major roadblocks in the acceptance of MR spectroscopy in clinical practice.

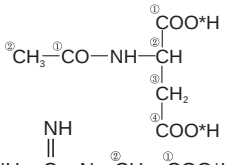
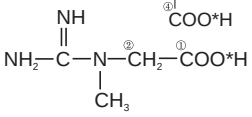
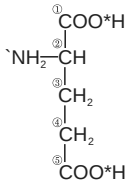
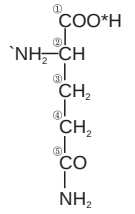
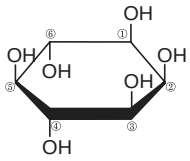
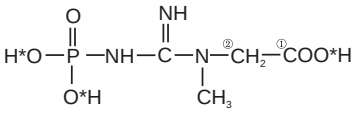
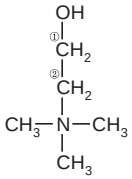
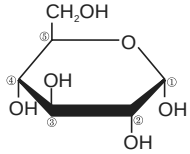
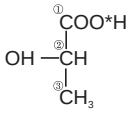
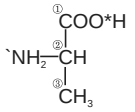
Currently, the majority of clinical MRS studies are performed in the human brain. Over the past several years, a database of brain metabolites detectable by proton MRS *in vivo* has been built and verified. This process is far from finished, as metabolites with lower and lower concentrations in the brain tissue are identified by MRS and their clinical efficacy is explored. A list of components routinely investigated in current clinical practice is shown in Table 1. Even a short glance at this list immediately reveals the first major canon of the MRS method: more than a cursory knowledge of organic and biochemistry is required to fully comprehend the information available. An appreciation of the true extent of this statement can be quickly gained by taking a closer look at the first molecule listed in that table: *N*-acetylaspartate, or NAA. This compound belongs to a class of *N*-alkanoamines; the italic letter *N* represents the so-called locant, or a location of the secondary group attached to the primary molecule. In this case, N is a locant for a group that is attached to a nitrogen atom located within the primary molecule. The primary molecule in this case is an L-aspartic acid, which is a dicarboxylic amino acid. Dicarboxylic means that the molecule contains two carboxylic (COOH) groups. As an amino acid, L-aspartic acid belongs to the group of the so-called nonessential amino acids, which means that under normal physiological conditions sufficient amounts of it are synthesized in the body to meet the demand and no dietary ingestion is needed to maintain the normal function of the body. The *N*-acetyl prefix identifies a molecule as a secondary amine; in such compounds the largest chain of carbon compounds takes the root name (aspartic acid), and the other chain (the acetyl group, CH_3CO–, formed by removal of the OH group from the acetic acid CH_3COOH) becomes a substituent, whose location in the chain (the *N*-locant) identifies it as attached to the nitrogen atom. But what about the L prefix in the L-aspartic acid mentioned above? It has to do with a spatial symmetry of the molecule's configuration. The second carbon in the aspartic chain has four bonds (two links to other carbon atoms, one link to the nitrogen atom, and the final link to a proton). These four bonds are arranged in space to form a tetrahedron with the four atoms just mentioned located at its apexes. Such a configuration is called a chiral center, to indicate a location where symmetry of atom arrangement needs to be documented. There are two ways to distribute four different atoms among four corners of a tetrahedron, one is called levorotatory (and abbreviated by a prefix L-), and the other is called dextrorotatory (and abbreviated by a prefix D-). It turns out that the chirality of the molecular configuration has a major significance in biological applications: Since successful mating of different molecules requires that their bonding sites match, only one chiral moiety is biologically active. In our case, L-aspartic acid is biologically active; the D-aspartic acid is not. Last, but not least, the reader has probably noticed that the name of the molecule is listed as *N*-acetylaspartate, while we keep talking about aspartic acid... well, when an acid is dissolved in water, it undergoes dissociation into anions and cations; the molecule of aspartic acid in the water solution looses two protons from the carboxylic groups COOH (the locations of the cleavages are indicated by asterisks in the structural formulas shown in Table 1), and becomes a negatively charged anion. To reflect this effect, and suffix -ate is used. Thus, the name *N*-acetylaspartate describes an anion form of a secondary amine, whose primary chain is a levorotatory chiral form of aspartic group, with a secondary acetyl group attached to the nitrogen atom. The NAA is so esoteric a molecule that most standard biochemistry books do not mention it at all; its chief claim to prominence comes from the fact that it is detectable by MRS. A recent review of NAA metabolism has been recently published by Barlow (15).

The example discussed above emphasizes that much can be learned just from a name of a biological compound. To gain more literacy in the art of decoding the chemical nomenclature of biologically active compounds, the reader is encouraged to consult the appropriate resources, of which the Introduction to Subject index of CAS (16) is one of the best.

As mentioned earlier, routine clinical MRS studies focus on proton spectra spanning the range from 0 to ~ 4.5 ppm; the NMR properties of compounds listed in Table 1 are presented in Table 2, which identifies each molecular group according to carbon labeling used in structural formulas shown in Table 1. For each molecular group, the chemical shift of the main NMR peak is listed, along with the spectral multiplet structure characterizing this line. A simulated theoretical spectrum shows all lines in the range from 0 to 5 ppm, to give the reader an idea where the molecular signature peaks are located in the spectra acquired *in vivo*. Finally, information is provided whether, for a particular line, the signal acquired in standard MRS *in vivo* studies is strong enough to emerge above the noise levels and become identifiable. This information is further supplemented with comments indicating whether a particular line is expected to be visible on spectra acquired with short or long *TE* values. It is evident that the information provided in Table 2 is absolutely critical to successful interpretation of clinical MRS results; unfortunately, space limitations prevent us from further dwelling into details of spectral characteristics of clinically important metabolites. This information can also be difficult to locate in the literature since most of the data still reside in original research papers; fortunately, a recent review by Govindaraju et al. (17) offers an excellent starting point.

An examination of data shown in Table 2 quickly leads to a realization that only a limited number of metabolite signature lines can be successfully used in the

Table 1. Basic Properties of Metabolites Most Commonly Detected in MRS Spectra of the Human Brain

Metabolite	Full Name	Acronym	Formula	CAS[a] Number	Structure	Molecular Weight	Normal Concentration Range in brain, mmol
N-Acetylaspartate	*N*-Acetyl-L-aspartic acid; amino acid	NAA	$C_6H_9NO_5$	[*997-55-7*]		175.14	8–17
Creatine	(1-Methylguanidino) acetic acid; non-protein amino acid	Cr	$C_4H_9N_3O_2$	[*57-00-1*]		131.14	5–11
Glutamate	L-Glutamic acid; amino acid	Glu	$C_5H_9NO_4$	[*56-86-0*]		147.13	6–13
Glutamine	L-Glutamic acid-5-amide; amino acid	Gln	$C_5H_{10}N_2O_3$	[*56-85-9*]		146.15	3–6
myo-Inositol	1,2,3,5/4,6-Hexahydroxy-cyclohexane	m-Ins	$C_6H_{12}O_6$	[*87-89-8*]		180.2	4–8
Phosphocreatine	Creatine phosphate	PCr	$C_4H_{10}N_3O_5P$	[*67-07-2*]		211.11	3–6
Choline	Choline hydroxide, Choline base, 2-Hydroxy-*N*,*N*,*N*-trimethyl-athanaminium	Cho	$C_5H_{14}NO$	[*62-49-7*]		104.20	0.9–2.5
Glucose	D-Glucose, dextrose anhydrous, corn sugar, grape sugar	Glc	$C_6H_{12}O_6$	[*50-99-7*]		180.15	1.0
Lactate	L-Lactic acid, 2-hydroxypropanoic acid	Lac	C_3H_6O	[*79-33-4*]		90.07	0.4
Alanine	L-Alanine, 2-amino-propanoic acid	Ala	$C_3H_7NO_2$	[*65-41-1*]		89.09	0.2–1.4

[a]CAS = Chemical Abstracts Service Registry Number of the neat, nondissociated compound. In structural formulas, an asterisk * indicates a site where, upon dissociation, a proton is released; apostrophe` indicates a site where, upon dissociation, a proton is attached. Metabolite names refer to dissociated (ionic) forms of the substances since this is the form they are present in the *in vivo* environment.

Table 2. The NMR Properties of Metabolites Most Commonly Detected in MRS Spectra of the Human Brain

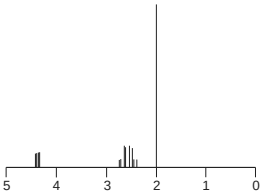
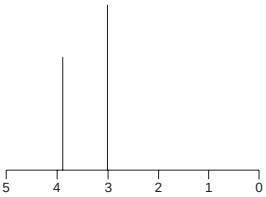
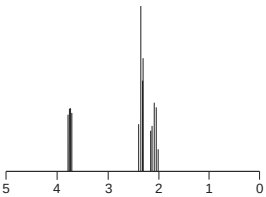
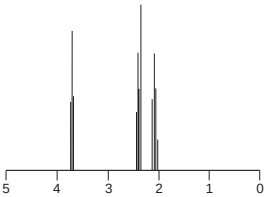
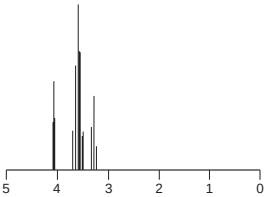
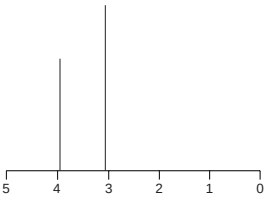
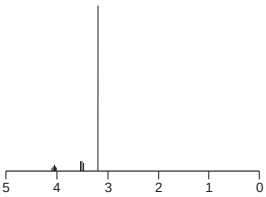
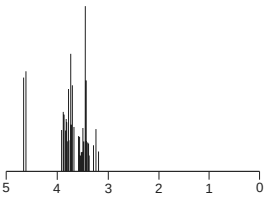
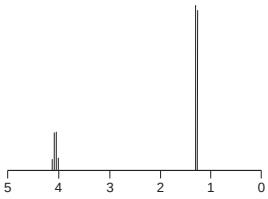
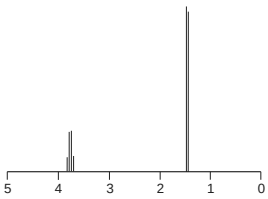
Metabolite	Acronym	Molecular Group	Chemical Shift δ ppm[a]	Multiplet Structure[b]	Visibility *in vivo*[c]	Theoretical Spectrum –Range (0,5) ppm
N-Acetylaspartate	NAA	• CH_3 Acetyl	**2.00**	s	Yes	
		• CH_2 Aspartate	2.49	dd	No	
		• CH_2 Aspartate	2.67	dd	No	
		• CH Aspartate	4.38	dd	No	
Creatine	Cr	N-CH_3	**3.03**	s	Yes	
		• CH_2	**3.91**	s	Yes	
Glutamate	Glu	• CH_2	**2.07**	m	Yes, on short TE	
		• CH_2	**2.34**	m	Yes, on short TE	
		• CH	3.74	dd	No	
Glutamine	Gln	• CH_2	**2.12**	m	Yes	
		• CH_2	**2.44**	m	Yes	
		• CH	**3.75**	t	No	
Myo-inositol	m-Ins	• CH	3.27	t	No	
		• CH, • CH	**3.52**	dd	Yes, on short TE	
		• CH, • CH	**3.61**	t	Yes, on short TE	
		• CH	4.05	t	No	
Phosphocreatine	PCr	N-CH_3	**3.03**	s	Yes	
		• CH_2	**3.93**	s	Yes	
Choline	Cho	N-$(CH_3)_3$	**3.18**	s	Yes	
		• CH_2	3.50	m	No	
		• CH_2	4.05	m	No	
Glucose	Glc	β–• CH	4.63	d	Not visible in normal brain due to low concetration	
		All other CH	**3.23–3.88**	m		

Table 2. (*Continued*)

Metabolite	Acronym	Molecular Group	Chemical Shift δ ppm	Multiplet Structure	Visibility *in vivo*	Theoretical Spectrum –Range (0,5) ppm
Lactate	Lac	• CH_3	**1.31**	d	Not visible in normal brain due to low concetration	
		• CH	4.10	q		
Alanine	Ala	• CH_3	**1.47**	d	Not visible in normal brain due to low concetration	
		• CH	3.77	q		

[a]The bold type indicates a dominant line in the spectrum.
[b]s = singlet, d = doublet, dd = doublet of doublets, t = triplet, q = quadruplet, m = multiplet.
[c]Reference to short TE indicates that those signals have short *T2*s and thus will be suppressed in acquisitions with long TE.

interpretation of clinical proton MRS spectra. In normal volunteers, five major markers can routinely be detected and evaluated:

The *N*-acetylaspartate peak at 2.0 ppm, commonly referred to as NAA.

The combination of creatine (Cr) and phosphocreatine (PCr); reported together as two lines positioned ∼ 3.0 and 3.9 ppm, respectively. Mostly referred to as Cr, but some people use a label tCr (for total creatine). The peak at 3.0 ppm is often identified as Cr, and the peak at 3.9 ppm as Cr2,

The combination of glutamine (Gln) and glutamate (Glu) at 2.2–2.4 ppm; since the peaks from those two compounds strongly overlap and are typically unresolvable, the are routinely reported together and referred to as Glx.

The choline peak at 3.25 ppm, referred to as Cho; The primary contributions to this peak are from bound forms of choline: phosphorylcholine (PC) and glycerophosphorylcholine (GPC), with only minor signal from free choline.

The *myo*-inositol group at 3.6 ppm is visible only in spectra acquired with short *TE* (due to *J* coupling effects that suppress the intensity of lines forming this signal at long *TE*s). *Myo*-inositol name poses evidently a challenge to acronym creators, since it can be found labeled as m-Ins, MI, mI, and In.

In addition, the following markers are routinely evaluated in a variety of diseases:

Lactate (Lac), often visible as a doublet at 1.3 ppm.

Mobile lipds (Lip) whose methyl (CH_3) groups are visible at 0.9 ppm and methylene (CH_2) groups provide signal at 1.3 ppm.

In special cases other metabolites may become visible, and we list here two examples:

Alanine (Ala) with a signature peak at 1.5 ppm.

Glucose (Glc), mostly showing as a broad peak ∼ 3.5 ppm; note that the theoretical spectrum shown in Table 2 is a superposition of α- and β-anomers that occur in 1:2 ratio, respectively, in equilibrium *in vivo* conditions.

Examples of typical MRS spectra obtained from healthy subjects are shown in Fig. 13. The first thing to notice is that the signal noise ratio in those spectra is poor, so indeed, only the strongest lines from metabolites listed in Table 2 have a chance to become visible.

In the context of this discussion, the following questions have been discussed: What is MRS, how to perform it, who is interested in those studies, and why? If an actual clinical MRS examination were being performed, at this stage of MRS study we would have localized the lesion, collected an *in vivo* MR spectrum from the tissue within this lesion, identified the signature metabolite peaks, and analyzed their relative intensities. Now would have come the time to ponder what the results mean and what is their clinical significance. It is not an easy task, since the last analytic step listed above produces results that must be compared to "normal" baseline references. These reference data are obtained by performing statistical analysis of multiple results obtained from healthy people: a challenging task given biological diversity of normal subjects. Thus, MRS results are reported using such imprecise terms as unchanged, elevated, or suppressed. Sometimes, when clinicians want to be particularly precise, they will report ratios, such as the NAA/Cr ratio, which essentially renormalizes all observed signal intensities by assigning an arbitrary intensity of one unit to the Cr peak. In other words, this approach uses the Cr peak as an internal reference. As mentioned in Table 3, the rationale for such

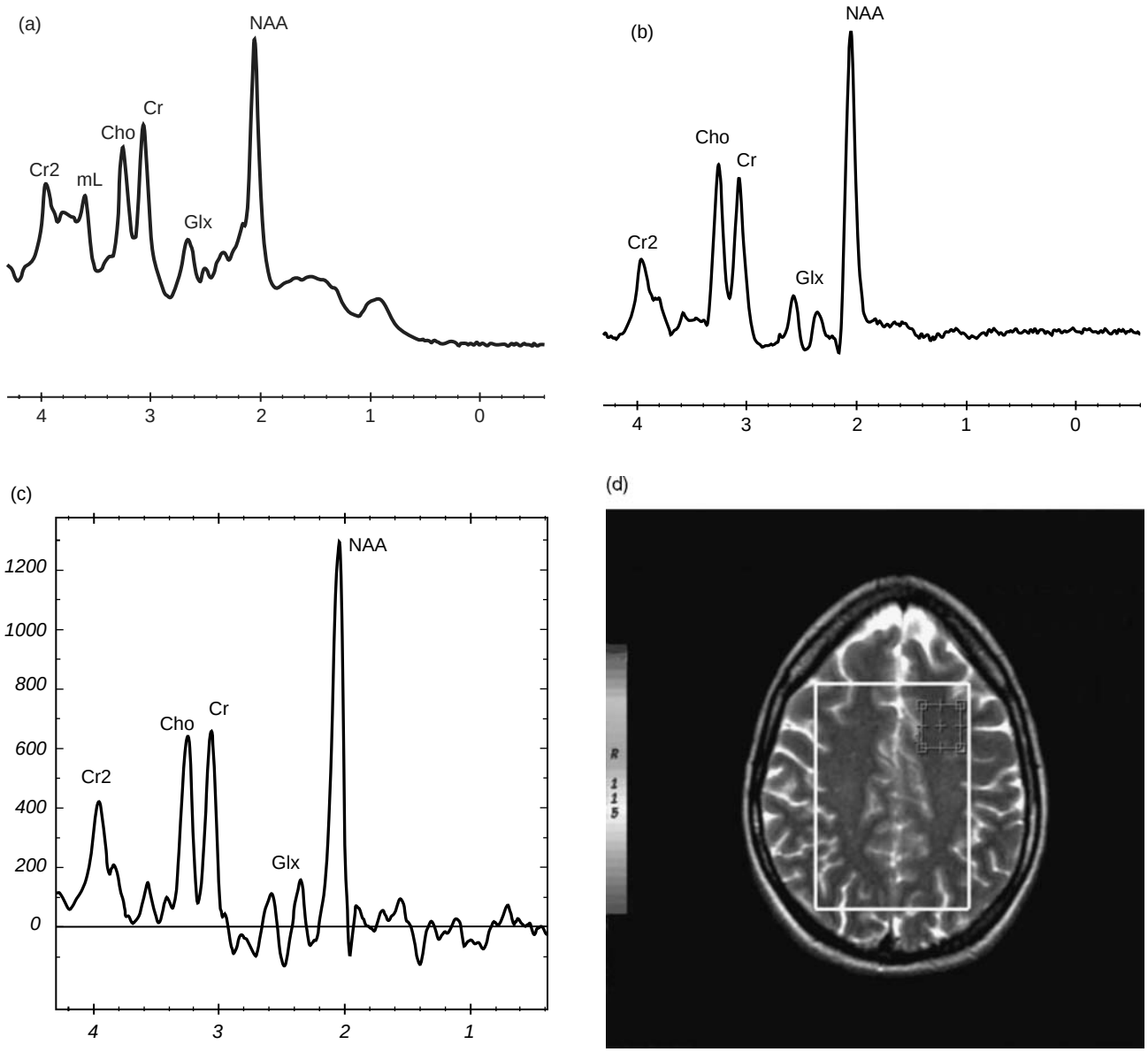


Figure 13. Examples of typical MRS spectra obtained *in vivo* for a normal volunteer: (a) STEAM SV protocol with $TE = 30$ ms and TR of 2000 ms, magnitude mode; (b) PRESS SV protocol with $TE = 144$ ms and TR = 1500 ms, magnitude mode; (c) PRESS 2D CSI with $TE = 135$ ms, TR = 1000 ms, 16*16 voxel locations, real mode; (d) Image reference showing a location of selected VOI associated with the spectrum shown in (c) The VOI is generated by combining voxel locations. All data acquired within a single study, lasting $\sim$ 15 min; spectra (a) and (b) were acquired at the same location using the same VOI size, no signal averaging was used (NEX = 8).

an approach is that the levels of Cr tend to be relatively stable under normal physiologic conditions. Such findings are purely phenomenological and their value is established over time by practicing evidence-based medicine, that is by performing statistical analysis of a large number of findings and looking for correlations between MRS results and the patient's clinical status. The drawback of such an approach is that in the early stages of new methodology, there is no established consensus regarding data interpretation, thus one is forced to read a large number of published clinical reports and develop one's own approach to the inclusion of MRS findings into the clinical decision making process. A short summary of the current understanding of clinical findings related to MRS results is provided in Table 3. However, since these findings are still subject to frequent updates, it would have been pointless to try to provide in-depth coverage of these issues here, knowing full well that the data are likely to be obsolete by the time this publication appears in print. Instead, the reader is strongly encouraged to survey the current literature for reviews of clinical MRS topics; some excellent, previously published papers can serve as a

Table 3. Clinical Properties of Metabolites Most Commonly Detected in MRS Spectra of the Human Brain

Metabolite	Full Name	Acronym	Function	Pathological Variation
N-Acetylaspartate	*N*-Acetyl-L-aspartic acid; amino acid	*N*AA	Plays osmoregulatory function in the intercompartmental system, thought to be responsible for the removal of intracellular water, against the water gradient, from myelinated neurons	Decrease in NAA levels is indicative of reversible axonal injury or neuronal loss. In adults it has been associated with neoplasia, multiple sclerosis, hypoxia, ischemia, stroke, trauma, epilepsy, encephalitis, and neurodegenerative syndromes
Creatine	(1-Methylguanidino)acetic acid	Cr	Synthesized in the liver, it is exported to muscle and brain where it is phosphorylated into phosphocreatine and used as an energy store	Levels of creatine and phosphocreatine are tightly controlled under physiologic conditions, and thus this peak has been suggested as an internal reference for metabolite amplitude or area ratios. Increased levels of Cr and PCr have been observed, however, in hyperosmolar states, as well as in trauma. They also increase with aging
Glutamate	L-Glutamic acid; amino acid	Glu	Most abundant amino acid found in the human brain, acts as an excitatory neurotransmitter	Plays a role in detoxification of ammonia in the hyper ammonemic states
Glutamine	L-Glutamic acid-5-amide; amino acid	Gln	Plays a role in glutamate regulation; astrocytes metabolize glutamate to glutamine, thus preventing excitotoxicity	Elevated levels of glutamate and glutamine have been observed in hepatic encephalopathy, Reyes syndrome, meningiomas, and rare inherited enzyme deficiency. Reduced levels have been associated with Alzheimer's disease
myo-Inositol	1,2,3,5/4,6-Hexahydroxy-cyclo hexane	m-Ins	A component of membrane phospholipids, functions as a cerebral osmolyte. It is also thought to play an essential role in cell growth	Concentration fluctuates more than that of any other major metabolite in the brain. Increased levels of myo-inositol have been observed in neonates, Alzheimer's disease, diabetes, and hypersomolar states. *myo*-Inositol levels are decreased in hepatic encephalopathy, hypoxia, stroke, and some neoplasms
Phosphocreatine	Creatine phosphate	PCr	A major energy storage in the body	See notes for creatine.
Choline	Choline hydroxide, Choline base, 2-Hydroxy-*N*,*N*,*N*-trimethylathanaminium	Cho	Choline is important for normal cellular membrane composition and repair, normal brain function and normal cardiovascular function	Increased levels of choline have been detected in normal infants, and aging adults. They were also associated with neoplasia, gliosis, demyelinating disease, inflammation or infection, trauma, diabetes, chronic hypoxia, and AIDS. Decreased levels of choline have been found in hepatic encephalopathy, stroke, and dementias, including Alzheimer's disease
Glucose	D-Glucose	GLc	A major energy carrier throughout the body.	Increased levels notes in diabetes mellitus, parental feeding, hypoxic encephalopathy
Lactate	L-Lactic acid, 2-hydroxypropanoicacid	Lac	Lactate is a signature byproduct of carbohydrate catabolism and thus when normal cellular oxidative respiration mechanisms are active, its levels in the brain tissues are very low	Visible in a variety of diseases. Increased levels of lactate were observed in some tumors, during the first 24 h after infarction, in hypoxia, anoxia, near-drowning, and hypoventilation
Alanine	L-Alanine, 2-Aminopropanoic acid	ALa	Alanine is a nonessential amino acid of uncertain function in the brain	The alanine peak is difficult to detect since it is easily overshadowed by lactate. Alanine levels might be elevated in meningiomas
Lipids	Fatty acids, glycerides, glycolipids, lipoproteins	Lip	Normally the lipid signals are not visible in the MRS spectra of the brain, but might appear due to fat contamination (voxel bleed)	Mobile protons from lipids (0.9 ppm for CH_3 and 1.3 ppm for CH_2) are not normally visible in brain spectra, but can appear in diseased conditions. Lipid signals are suppressed at long *TE*s. Elevated lipid levels are observed in cellular necrosis, high-grade astrocytoma, and lymphoma

starting point (18–21). The last paper on this list, by Smith and Stewart (21), also offers excellent overview of spectroscopy studies in organs other than the brain.

It is impossible, in one short article, to offer a fully comprehensive coverage of such an advanced and rapidly evolving discipline as clinical magnetic resonance spectroscopy. The author can only hope that this brief overview offers sufficient information and enough reference pointers to let the readers start exploring this new and exciting field on their own.

BIBLIOGRAPHY

Cited References

1. Becker E, Fisk CL, Khetrapal CL. The development of NMR. Vol 1: 2-158. In: Grant DM, Harris RK, editors. Encyclopedia of NMR. Chichester (UK): John Wiley & Sons, Inc.; 1996.
2. Ridgen JS. Quantum states and precession: The two discoveries of NMR. Rev Mod Phys 1986;58:433–488.
3. Arnold JT, Dharmati SS, Packard ME. Chemical effects of nuclear induction signals from organic compounds. J Chem Phys 1951;19:507.
4. Shoolery JN. The development of experimental and analytical high resolution NMR. Progr NMR Spectrosc 1995;28:37–52.
5. Cohen JS, et al. A history of biological applications of NMR spectroscopy. Progr NMR Spectrosc 1995;28:53–85.
6. Hahn EL. Spin echoes. Phys Rev 1950;80:580–594.
7. Frahm J, Merboldt KD, Hänicke W. Localized proton spectroscopy using stimulated echoes. J Magn Reson 1987;72:502–508.
8. Bottomley PA. Spatial localization in NMR spectroscopy *in vivo*. Ann NY Acad Sci 1987;508:333–348.
9. Brown TR, Kincaid BM, Ugurbil K. NMR chemical shift imaging in three dimensions. Proc Natl Acad Sci USA 1982; 79:3523–3526.
10. Kwock L. Localized MR Spectroscopy — Basic Principles. Neuroimaging Clin N Am 1998;8:713–731.
11. den Hollander JA, Luyten PR, Mariën AJH. ^{1}H NMR Spectroscopy and spectroscopic imaging of the human brain. In: Diehl P, et al., editors. NMR — Basic Principles and Progress. Vol. 27. Berlin: Springer-Verlag; 1991.
12. Decorps M, Bourgeois D. Localized spectroscopy using static magnetic field gradients: comparison of techniques. In: Diehl P, et al., editors. NMR — Basic Principles and Progress. Vol. 27. Berlin: Springer-Verlag; 1991.
13. van Zijl PCM Moonen CTW. Solvent suppression strategies for *in vivo* magnetic resonance spectroscopy. In: Diehl P, et al., editors. NMR — Basic Principles and Progress. Vol. 26. Berlin: Springer-Verlag; 1991.
14. Kreis R. Quantitative localized ^{1}H MR spectroscopy for clinical use. Progr NMR Spectr 1997;31:155–195.
15. American Chemical Society, Naming and Indexing of Chemical Substances for Chemical Abstracts, Appendix IV. Chemical Abstracts Service, *Chemical Abstracts Index Guide*, Columbus: American Chemical Society; 2002.
16. Baslow MH. *N*-acetylasparate in the vertebrate brain: metabolism and function. Neurochemical Res 2003;28:941–953.
17. Govindaraju V, Young K, Maudsley AA. Proton NMR chemical shifts and coupling constants for brain metabolites. NMR in Biomed 2000;13:129–153.
18. Blüml S, Ross B. Magnetic resonance spectroscopy of the human brain. In: Windhorst U, Johansson H, editors. Modern Techniques in Neuroscience Research. Berlin: Springer-Verlag; 1999.
19. Ross B, Bluml S. Magnetic resonance spectroscopy of the human brain. Anat Rec (New Anat) 2001;265:54–84.
20. Smith JK, Castillo M, Kwock L. MR spectroscopy of brain tumors. Magn Reson Imaging Clin N Am 2003;11:415–429.
21. Smith ICP, Stewart LC. Magnetic resonance spectroscopy in medicine: clinical impact. Progr NMR Spectrosc 2002;40:1–34.

Reading List

Bernstein MA, King KF, Zhou XJ. Handbook of MRI Pulse Sequences. Burlington (MA): Elsevier Academic Press; 2004. This is a "geek's delight". A detailed, step-by-step presentation and discussion of problems and solutions encountered in routine MRI practice today. A must-read for anyone seriously interested in learning MRI beyond the popular level.

deGraaf R. In Vivo NMR Spectroscopy: Principles and Techniques. Chichester (UK): John Wiley & Sons Inc.; 1999. This book covers both theoretical and practical aspects of MRS and is widely considered to be one of the best textbooks available on the subject. The book is particularly well suited for people involved in MR research outside a clinical medical environment, since it focuses on physics and engineering aspects of the methodology. Students, beware: it is very expensive ($350 for 530 pages), so it is best to seek it out at the library.

Ernst RR, Bodenhausen G, Wokaun A. Principles of NMR in One and Two Dimensions. Oxford (UK): Clarendon Press; 1987. Fundamental monograph on modern theory of NMR spectroscopy. Very comprehensive coverage, but definitely not for beginners.

Fukushima E, Roeder SBW. Experimental Pulse NMR: A Nuts and Bolts Approach. Reading (MA): Addison-Wesley; 1982. The best introduction to practical NMR. The book is out of print, but libraries still carry it and it is relatively easy to purchase second-hand, since it has been hugely popular among graduate students starting up in the NMR field at a graduate level.

Goldman M. Quantum Description of High Resolution NMR in Liquids. Oxford (UK): Clarendon Press; 1988. Very methodical and thorough coverage of HR NMR Spectroscopy in liquids, but sometimes unconventional formalism requires an extra effort on the part of the reader to really understand all aspects of discussed subject matter.

Grant DM, Harris RK, editors. Encyclopedia of NMR. Chichester (UK): John Wiley & Sons Inc.; 1996. This is a monumental piece of work (eight volumes and one update volume thus far) that rightly deserves the title of the most comprehensive review of the field to date.

Martin ML, Martin GJ, Delpuech J-J. Practical NMR Spectroscopy. London (UK): Heyden & Son, Ltd.; 1980. This book is out of print, but copies are available at libraries. It is one of the best "hands-on", practical texts on HR NMR spectroscopy. Covers practical hints on hardware, experiment setup, sample preparation, various techniques of spectral editing, and so on.

Slichter CP. Principles of Magnetic Resonance. Berlin: Springer-Verlag; 1990. Considered by many as the "the Bible" of MR theory. It is an advanced textbook that is meant to provide the beginner a necessary background to get started in the field of MR.

Young IR, editor. Methods in Biomedical MRI and Spectroscopy. Chichester (UK): John Wiley & Sons Inc.; 2000. This two-volume set contains most entries that have been originally included in the Encyclopedia of NMR, but they have been expanded and updated by the original contributors.

See also COMPUTED TOMOGRAPHY; MAGNETIC RESONANCE IMAGING; POSITRON EMISSION TOMOGRAPHY; ULTRASONIC IMAGING.

NUCLEAR MEDICINE INSTRUMENTATION

LAWRENCE E. WILLIAMS
City of Hope
Duarte, California

INTRODUCTION

Nuclear medicine exists as a clinical specialty due to two basic reasons involving signal detection. Of primary importance is the high sensitivity of tissue measurements. In principle, a single labeled molecule or nanostructure may be detected upon the decay of its attached radiolabel. A second reason is the possibility of using radiolabeled

materials of interest to study the physiology of animals and eventually patients. While imaging is the primary application of nuclear techniques, targeting implies an associated therapeutic strategy. All three traditional forms of radioactive emission, alpha (α), beta (β^- and β^+), and gamma radiation (γ) are available to the investigator. Negative betas are identical to the electrons found external to the atomic nucleus and are the antiparticle to β^+ (positron). Penetration distances in soft tissue for α and β rays range from μm and up to several millimeters, respectively, and so limit imaging use to organ samples or perhaps very small intact animals. Both of these particles are, however, employed in radiation therapy.

It is the photon emitter that is most valuable as an imaging label since it can be used *In vivo* on relatively large animals and patients. One exception to this general rule is the application of positron emitters (β^+) in imaging. Notice that a β^+ annihilates with a local atomic electron to form two or three photons of high energy. Thus, the positron emitter is effectively giving off quanta of a detectable type although up to several millimeters away from the site of the original decay. Because of momentum conservation, emission of two annihilation photons is essentially back-to-back; that is, at 180° separation, so as to define a line in space. This fact allows positron emitters to be an almost ideal label for 3D imaging.

Labeling Strategies

Radioactive labels may be used, in principle, to locate and quantitatively measure pharmaceuticals within excised samples, intact animals, and patients. Several strategies of labeling are possible. The radioactive tag may be used directly in the atomic form, such as ^{123}I as a test species replacing the stable isotope ^{127}I for evaluation of the patient's thyroid physiology. A secondary method is to replace a stable atom in a biological molecule by a radioactive isotopic form as ^{14}C in lieu of stable ^{12}C in a sugar. Finally, as is most common, the label is simply attached by chemical means to a molecule or engineered structure of interest. One can tag an antibody with radioactive ^{131}I or use ^{111}In inside a 50 nm phospholipid vesicle to track their respective movements inside the body of a patient. Because of protein engineering and nanotechnology, such radiolabeled manmade structures are of growing importance. Table 1 gives an outline of the three types of labeling and examples of associated clinical studies.

Applications of nuclear tagging can literally go far beyond clinical assays. When the 1976 Viking landers came down on the surface of Mars, a test for living organisms was performed using various ^{14}C labeled nutrients. An assay was then performed on a scoop of Martian soil mixed with the radiotracers using a radiation detector sampling emitted gases. It was thought that ^{14}C-methane would prove metabolism (i.e., life). While a weak positive signal was detected in the reaction chamber, these results have yet to be verified by other test procedures. Methane has, however, been found as an atmospheric gas by more recent exploratory spacecraft.

Limitations of Radioactive Labels

In the last two types (II and III) of labeling, radionuclides can become separated from the molecule or structure of interest. This disassociation may occur during preparation and/or delivery of the pharmaceutical or later *In vivo*. Responsible processes include reversible binding of the radionuclide, enzymatic action, or even competition with stable isotopes of the same element. Nuclear medicine specialists must recognize such limitations in any resultant analyses: a subtlety often overlooked in a report or document.

A second important logical issue associated with nuclear imaging is tissue identification and anatomic localization. Nuclear imaging physicians are very analogous to astronomers in that entities may be observable, but indeterminate as to type or location. Relatively strong (hot) sources appearing against a weak background in a nuclear image may be coming from a number of tissues. The physician may not, in fact, be able to identify what structure or organ is being observed. Hybrid imaging devices combining nuclear and anatomic imagers such as computed tomography, (CT) are being implemented to correct for this ambiguity and are discussed below.

Lack of specific radiopharmaceuticals has been the greatest limitation to the growth of nuclear medicine. Many tracer agents owe their discovery to accidental events or the presence of a traditional metabolic marker for a given tissue type. Yet, these historical entities may target to several different organs *In vivo* and thus lead to ambiguous images. More recently, molecular engineering, computer modeling and the generation of specific antibodies to tissue and tumor antigens have improved production of novel and highly specific agents. The most specific of these entities is the monoclonal antibody binding to a particular sequence of amino acids in the target antigen's structure.

Therapy Applications

Detection and imaging via tracers are not the only clinical tasks performed in nuclear medicine. Of increasing importance is the provision of radiation therapy when there is preexisting imaging evidence of radiopharmaceutical

Table 1. Methods and Examples of the Three Types of Nuclear Medicine Labeling

Method	Label Example	Clinical Study	Detector Device
I. Substitution of radioactive atom for common stable atom	^{123}I for ^{127}I (stable)	Thyroid uptake	Single probe or gamma camera
II. Insertion of radioactive atom in a molecule	^{14}C for ^{12}C in glucose	Glucose metabolism	Liquid Scintillator (LS) detecting exhalation of $^{14}CO_2$
III. Attachment of radioisotope to a structure	^{111}In attached to an liposome	Planar or SPECT image of cancer patient	Gamma camera

targeting to the lesion(s) in question. The oldest such treatment is the use of ^{131}I as a therapy agent for thyroid cancers including both follicular and papillary types. Here, the radionuclide emits imaging photons and moderate energy beta radiation so that localization can be demonstrated simultaneously with the treatment phase of the study. In some applications, the therapy ligand is intentionally a pure beta emitter so as to limit radiation exposure to the medical staff and patient's family. In this case, no gamma photons are available to the imaging devices. The therapist must use the coadministration of a surrogate tracer to track the position of the pure beta therapy agent. An example is the use of ^{111}In-antibodies to cancer antigens to track the eventual location of the same antibody labeled with the pure beta emitter ^{90}Y.

RADIONUCLIDE PRODUCTION

Reactor Production of Radionuclides

Production of radionuclides that are useful in nuclear medicine relies on several different methodologies. The most common nuclear medicine radiolabel, ^{99m}Tc, is produced as a decay product of its parent ^{99}Mo. Production of ^{99}Mo is generally done via nuclear fission occurring inside a nuclear reactor. Radioactive ^{99}Mo is taken into the radiopharmacy where it is attached to an alumina (Al_2O_3) column. By washing physiological saline through this generator device, the user may elute the technetium that is chemically dissimilar from the ^{99}Mo, and so comes free of the column. Possible breakthrough or leakage of Mo is measured upon each so-called "milking" procedure to assure the pharmacist that the eluted material is indeed technetium. While other generator systems are available, obtaining specific radionuclides generally requires provision of the appropriate reaction using a suitable accelerator.

Cyclotron Production of Radionuclides

A more general way to produce radioactive species of a given type is via a designated nuclear reaction. For example, while many isotopes of iodine can be found in fission reactor residues, their chemical identity makes separation a difficult problem. For that reason, ^{123}I has been obtained with the nuclear transmutation:

$$\text{Proton} + {}^{124}\text{Te} \rightarrow {}^{123}\text{I} + 2\,\text{neutrons}$$

More than one reaction can occur given the same initial conditions. In the above case, production of ^{124}I is possible when only one neutron is generated by the bombarding proton in an isotopically pure ^{124}Te target. This contamination is intrinsically present in any ^{123}I product resulting from the bombardment. Since ^{124}I has a 100 h half-life that is much longer than that of ^{123}I (13 h), the relative amount of this impurity increases with time and may become difficult to correct for in resultant gamma camera images.

While a variety of particle accelerators may be used, the most common device to produce a given radionuclide by a specific reaction is the industrial or clinical cyclotron. This is a circular accelerator invented by Lawrence and Livingston in which large electromagnets hold the proton (or other charged particle) beam in a circular orbit of increasing radius as its energy is enhanced twice per cycle with radio frequency (rf) radiation. Circulation of the beam is permitted over extended acceleration times as the volume between the magnetic poles is kept in a relative high vacuum condition.

Straight-line machines, such as tandem Van de Graaff units and linear accelerators (linacs), in which the beam moves in a geometric line from low energy ion source to the reaction site, have some disadvantages compared to a cyclotron design. In linear devices, length is generally proportional to the desired energy so as to make the machine difficult to house: particularly in a clinical setting. The clinical cyclotron is small enough to fit within a medium-sized room as shown in Fig. 1. Second, the high voltage needed to accelerate the proton or other ion may be difficult to maintain over the length of the straight-line device. Electric breakdowns not only interrupt accelerator operation, they may also damage the internal electrodes.

In order that the appropriate nuclear reaction is possible, the proton beam must strike an isotopically purified target. This may occur within the cyclotron or in a separate chamber external to the accelerator. The latter method is preferred as it permits easier access to the resultant product and rapid switching of one target with another as the reactions are varied. External target locations also reduce the radiation level inside the accelerator. In the ^{123}I example shown above, the target is a foil of highly purified Te metal; this is an isotope that is $\sim 5\%$ abundant in natural tellurium.

Unlike linear machines, beam extraction into the target chamber can be problematic for a cyclotron since the ion being accelerated is moving in a stable circular orbit. Traditionally, extraction was done using an electrode. A more effective way to extract protons from the vacuum chamber is to initially attach two electrons to each proton to form an H^- ion. This molecular species is accelerated until it reaches the correct reaction energy and a corresponding outer orbit. At this point, the circulating negative hydrogen ion is allowed to hit a so-called stripper foil that removes both electrons and converts the ion back to an ordinary proton (H^+). The proton is not geometrically stable at that radius and field and is magnetically led out of the cyclotron's vacuum chamber and into the target chamber for the desired reaction.

In addition to longer lived radionuclides, such as ^{123}I, ^{67}Ga, and ^{201}Tl, cyclotrons are conventionally used to manufacture short-lived radionuclides for positron emission tomography (PET) imaging. The latter include ^{11}C (20 min half-life), ^{13}N (10 min), and ^{15}O (2 min). Commercially, the most common product is ^{18}F (110 min) for use in fluorodeoxyglucose (FDG) as described below. Because of the several minute half-lives of the first three of these labels, it is necessary that the cyclotron is available on-site within the nuclear pharmacy. With ^{18}F production, the accelerator may be more remote; perhaps as far as an hour's drive from the clinical site so that fluorine decay does not appreciably reduce the delivered activity.

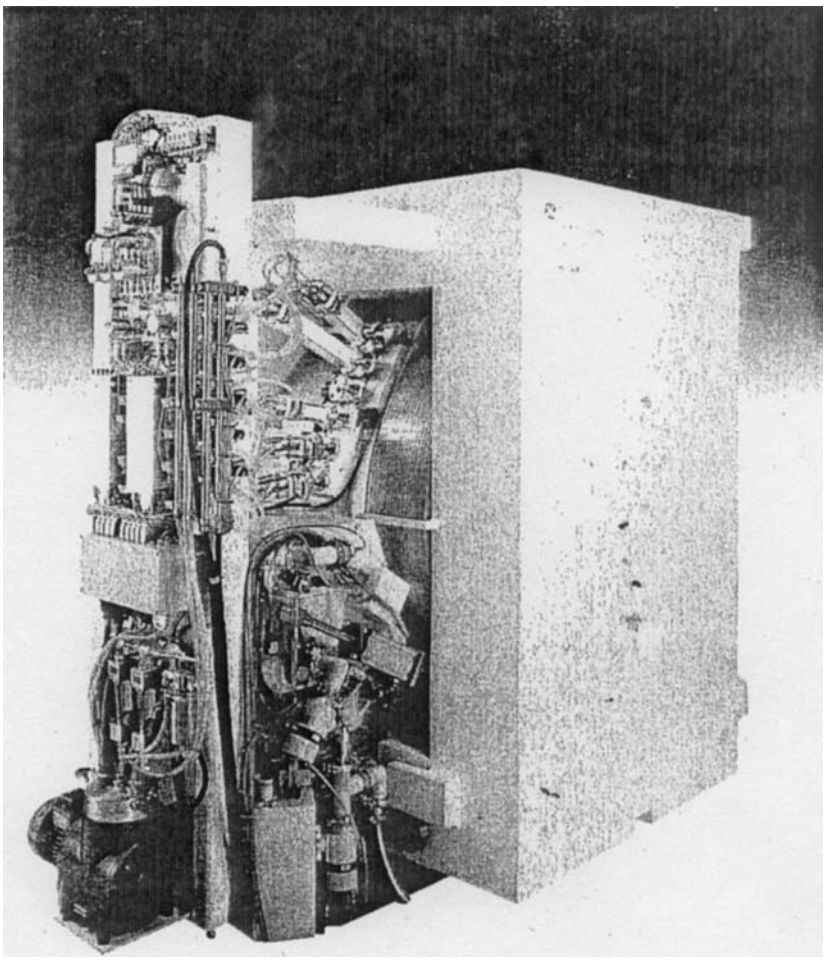


Figure 1. Medical cyclotron.

SYSTEMATICS OF RADIATION DETECTION

Detection Methods for Ionizing Radiation

Ionizing radiation is detected using electrons liberated within a sensitive volume of a detector material. All three classical states of matter, gas, liquid, and solid have been used as an ionization medium. Table 2 lists examples of each state and the devices associated with it. Most materials have ionization energies on the order of 30 eV per electron–ion pair. In solid-state semiconductors, such as Si or Ge, electron–hole pairs can be formed using ~3 eV. This lower value means that semiconductors can provide many more (~10×) ionization events for a given photon or electron energy. Such an increased number of events in turn yields improved statistical certainty that the particle has activated the counter. High thermal noise levels and elevated costs of large arrays of semiconductors have limited their use clinically.

Spectrometry

Signals of various sizes can arise in the detection process. Radionuclide counting depends on selection of the appropriate signal in a milieu of background radiation and other sample decays. For example, the technologist may have to count several beta emitters simultaneously or to detect a given gamma ray energy among many other emissions. Figure 2 shows a gamma spectrum from ^{137}Cs; both

Table 2. Detector Materials used to Measure Ionizing Radiation

State of Matter	Material	Energy per Ionization, eV	Device	Application
Gas	Argon	32	Dose calibrator	Photon activity assay
	Air	32	Ion chamber	Exposure level measurement
Liquid	Scintillation fluid (toluene)	30	Liquid counter	Beta assay in biological samples
Solid	NaI (Tl)	30	Gamma camera or probe	Photon counting
	Si (Li)	3	Solid-state probe	Photon and beta counting
	LiF (Tl)	30	TLD (Thermoluminescent dosimeter)	Radiation safety

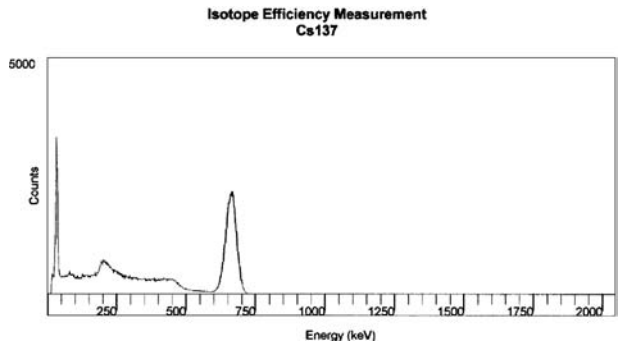


Figure 2. Energy spectrum of ^{137}Cs as measured by a NaI(Tl) probe.

Compton scattering and photoelectric effect (PE) are observed in this probe made of NaI (Tl).

In the PE, all of the photon energy is given over to an electron–ion pair in the absorbing material. Compton scatter may go on inside the patient prior to the photon coming into any detection system. In such cases, the direction and energy of the quantum may be changed so that an unwanted source may contribute to the counting process. Photon energy analysis is used to guard against such events in imaging; if the energy is seen to be reduced from that of the expected value, an electronic discriminator circuit rejects the ionization event. This pulse height analysis (PHA) is common to all nuclear detector systems and is described for imaging devices below.

ONE-DIMENSIONAL NUCLEAR MEDICINE DETECTORS

Well Counters

The most primitive instrument for photon detection is the counter or probe. In this case, a NaI(Tl) crystal is generally used to form a single large scintillation detector. In the scintillation process, the ionization event within the crystal is converted to visible light with a decay time on the order of 2 μs or less. Note that NaI is hygroscopic so that isolation of the crystal from the atmosphere is required. A reflective cap of Al is generally used as part of this hermetic seal. Resultant scintillation light is amplified by photomultiplier (PM) tubes to yield an electric signal proportional to the total amount of visible light. Well counters have the crystal in a hollow (cup) shape with the sample within the cup to maximize geometric sensitivity. Shielding is provided by an external layer of lead so as to reduce background counts. This is particularly important in a laboratory or clinical context. A mobile combination of well counter and probe system is shown in Fig. 3. Applications include sample assay using a standard source to give absolute values to the amount of detected activity. Counting experiments may involve patient tissue specimens obtained from the surgeon or animal organs obtained during measurement of biodistributions. Radiation protection is an additional application, whereby surface swab samples are counted to see if contamination is removable and possibly being spread around a lab or clinical area.

A second type of well counter, using high pressure Argon gas as the detector, is the dose calibrator. This device is

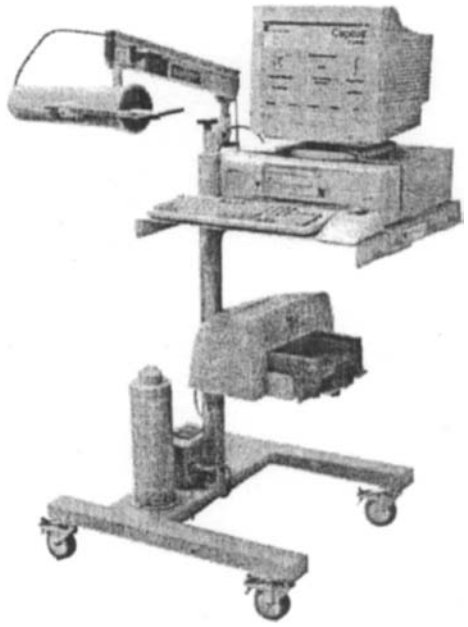


Figure 3. Well counter and probe mounted on a mobile chassis.

used in all nuclear pharmacies and clinics to measure the amount of radioactivity (mCi or MBq) in the syringe prior to administration via injection into a patient's vein. A curie is defined as 3.7×10^{10} decays per second and a bequerel is one decay per second. Standards are used to calibrate the device at the relevant energy of the radiopharmaceutical. Since the walls of the counter stop alpha and beta radiation, a dose calibrator generally may be used only for the photon component of the decay radiation. One exception is the assay of very high energy beta emitters, such as ^{90}Y or ^{32}P. In these cases, the betas give off a continuous spectrum of X rays of appreciable energy while they are decelerated before coming to rest. Such brake radiation (bremsstrahlung in German) may be detected quantitatively to calculate the amount of high energy beta emitter present in the syringe. Lower energy beta emitters, however, present difficulty in quantitative assays and generally require a different strategy for detection.

Liquid scintillation (LS) counters are a third form of well counter. Here, the beta emitter is dissolved into a liquid hydrocarbon that has been doped so as to produce scintillations suitable for PM detection. These devices have wide application in the quantitative assay of low energy beta emitters used for *In vitro* biological research. Radionuclides of interest include ^{3}H (E_{beta} max = 18 keV), ^{14}C (155 keV) and ^{35}S (167 keV). Energies cited refer to the kinetic energy of the betas. These labels are generally used in type II labeling as shown in Table 1. Multiple samples are sequentially measured for a fixed counting interval by lowering the tube containing mixed scintillator and radioactive material into a darkened space viewed by one and probably two PM tubes. The sample is dissolved in a liquid (usually toluene), which is activated with small amounts of fluors, such as PPO (2,5-diphenyloxazole) and POPOP (4-bis-2,5-phenyloxazolyl) benzene as solutes so as to provide visible light upon being struck by the electrons released during decay. Standards are included in the experimental

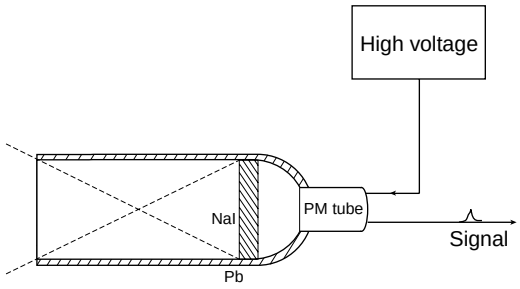


Figure 4. Operating principle of the clinical probe. Note that the observed field of view increases with distance from the opening of the collimator.

run to give absolute values for the activity. Efficiencies may approach 90% or more for moderately energetic betas. Reduction (quenching) of the light output due to solvent impurities and biological molecules within the sample can significantly affect the results and are accounted for by using standards.

The addition of fluors may not be needed to count very high energy beta particles. If the beta speed exceeds that of light in the solvent (E_β max $>$ 0.26 MeV in water), a photon shock wave is produced in the medium. Termed Cerenkov radiation, the emitted light is analogous to that of the acoustic wave or sonic boom produced by an aircraft exceeding the speed of sound. An observer may use Cerenkov light, which includes a continuum of visible and ultraviolet (uv) photons, to directly quantitate beta activity in the sample.

Probes

Clinical probes contain a planar crystal, usually a right circular cylinder, placed at the end of a long, shielded tube called a collimator. The central field of view is typically on the order of a circle 10 cm in diameter. The collimator is another right circular cylinder so that the total field observed increases with distance from the opening. Since the patient has a relative small thickness, on the order of 30 cm or less, this expanding view is not detrimental to the resultant clinical counting experiments. Such static NaI (Tl) devices are routinely used in measurements of thyroid uptake of radioactive ^{123}I as described above. Figure 4 contains a cross-section through a typical probe. With a single probe giving a result of activity for a relatively small, fixed field of view, it is necessary that sets of several probes be employed for measurement in an extended or spatially variable organ.

Conformal arrays have been used to yield information on regional cerebral blood flow (rCBF) in patients. Sets of 10 or more detectors have been arranged around the patient's skull so as to measure regional accumulation of perfusion tracers, such as ^{133}Xe in the brain. Because of the low gamma energy (81 keV) of ^{133}Xe, a given probe essentially views only physically adjacent tissues in rCBF counting. Such arrays led to the discovery that regional brain–blood flow varied with the mental task that the patient was performing during the time of observation. In this application, it is necessary that the intervening scalp blood flow be subtracted from the time–activity curves for each region. Tomographic methods such as PET do not suffer from this limitation. The PET flow measurements have confirmed the probe rCBF results and generalized them to other aspects of brain blood flow and metabolism during conscious and subconscious thought processes.

A more recent probe application is the detection and uptake measurement of so-called sentinel lymph nodes in melanoma, breast, and other cancer patients. These sites are defined as the first draining node associated with the lesion. They are located following a near-primary injection of a ^{99m}Tc-labeled cluster of sulfur colloid particles. Particle sizes up to 1×10^3 nm may be used. Of necessity for spatial resolution, the hand-held probe has a greatly reduced field of view, on the order of a few millimeters, and may be driven by battery power for convenience in the operating room (OR). Because of size limitations at incision sites, such probes may be of the solid-state type, whereby the ionizing event is converted to an electronic signal directly without the necessity of PM tube signal amplification. At present, CdTe and CsI(Tl) detectors have been incorporated into clinical probe systems. In the latter case, a photodiode is used in lieu of a PM tube to provide miniaturization of the device. An example of a surgical probe is shown in Fig. 5. For use in the OR, the device is usually gas sterilized, and then placed into a plastic sleeve before being put into an operating field.

Similar probe applications can involve radiolabeled antibody proteins used to locate small metastatic lesions in cancer patient after removal of their primary tumor. This has been termed radioimmune-guided surgery (RIGS). By measuring the gamma activity per gram of excised tissue, the radiation oncologist may estimate the radiation dose achievable with that patient's disease if radioimmunotherapy (RIT) were eventually utilized. In the case of RIT, a beta label is attached to the antibody in lieu of the gamma label used in localization if the radionuclide label does not emit both types of ionizing radiation. Probe-guided biopsy allows direct treatment planning for the RIT procedure that may follow.

Probes are also available for positron detection in the OR. This measurement assures the surgeon that the resection has taken out all of the suspect tissue that has been previously located using a FDG imaging study and a PET scanner. Because of the presence of both annihilation 511 keV photons and positrons, some correction mechanism is necessary for these instruments. A dedicated microprocessor attached to the detector system will provide this information if the probe has separate sensitive elements for positrons plus photons and for photons alone so that a subtraction may be done in real time.

TWO-DIMENSIONAL DETECTORS

Rectilinear Scanners

Because of the limited field of view of single probes, it was once considered clinically relevant for such devices to be mounted on a motor-driver chassis so as to pass in raster fashion over an entire organ. The trajectory of the probe in this context is the same as a gardener mowing the lawn. A simple thyroid probe in this application would prove problematic since it is focused at infinity; that is, observes all

Figure 5. Operating room probe. Miniaturization is dictated by the need to minimize the incision site at the sentinel lymph node. With robotic developments, even smaller designs will be necessary.

tissues from one side of the patient through to the opposite side. It can be used on the thyroid since no other organ taking up radioiodine usually lies within the neck region. In order to generally restrict the depth of the field of view, focused (converging) collimation was developed for raster-driven rectilinear scanner systems so that only emitters at a fixed distance were detected with relatively high efficiency. Dynamic studies, whereby activity was imaged during its physiological motion within the body, were difficult with this device unless the kinetics were significantly slower than the total raster scan time. Today the rectilinear scanner is a historical artifact that is no longer used in the clinic because of the development of the gamma camera. A camera allows both static and dynamic imaging over a reasonably large field (50 cm) without requiring movement of the detector assembly.

Gamma Cameras

H. Anger, in the late 1950s, avoided most of the scanner problems by inventing a gamma camera. As in the probe example, a right circular cylinder of NaI(Tl) was used to detect the photon. However, instead of a single PM tube, a hexagonal array of such tubes was employed to determine (triangulate) location of a given scintillation within the detector's lateral (x, y) dimensions. This fundamental principle is illustrated in Fig. 6. In order to spread the light somewhat more uniformly over the PM cathode, a light pipe (diffuser) is generally interposed between scintillation crystal and photomultiplier array. Localization was originally done with an analogue computer measuring the relative signal strength from each of a set of PM tubes. A second type of processing occurs with the sum of the PM signals. An energy window is set so that only photons having energy within a prespecified range are recorded as true events. The window is sufficiently wide, for example, $\pm 10\%$, that most signals arising from PE absorption of a monoenergetic gamma are recorded, but other photons, such as those scattered in the patient, are rejected. If the radionuclide emits several different photons, separate energy windows are set to count each energy level. The sum of all counts within all windows is then taken as the clinical result.

The original camera had cylindrical geometry arising from the single-crystal shape. Modern cameras generally have rectangular NaI(Tl) detectors made by combining annealed crystals of relatively large size allowing the

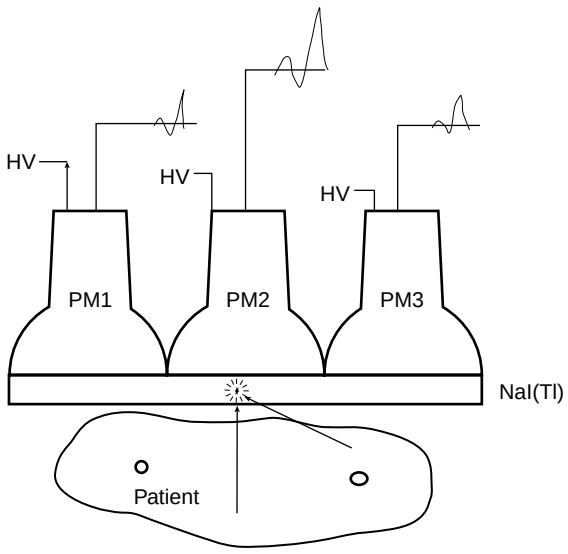


Figure 6. Principle of the Anger gamma camera. If no collimation is included (as shown) there is ambiguity of decay position.

entire width of the patient to appear in one field of view. The ensemble of crystal, multiple PM tubes, and associated computer electronics is referred to as the camera head. It is usually in the form of a rectangular solid and is mounted on a gantry allowing rotation and translation with respect to the patient bed. In the latter case, the motion is one dimensional (ID).

In the absence of directional information, a photon coming from anywhere within the entire hemisphere above the detector may impact the same position on the camera face. To remove the ambiguity, it is necessary that a collimator be provided between the detector crystal and the radioactive object(s). A collimator projects the activity distribution onto the crystal face. Essentially, this is a shadow or projection of the radioactivity distribution. Four standard types of collimators are shown in Fig. 7. The most common of these in clinical use is the parallel-hole type that is focused at infinity; that is, only passes parallel photons (rays) coming from the tissue of interest. Notice that the image and object size are equal in this case (magnification, M, $= 1$). This is essentially the same geometry used in the thyroid probe. Divergent collimators minify ($M < 1$) and convergent collimators magnify ($M > 1$) radioactive objects being imaged. The terms divergent and convergent refer to the point of view of the camera crystal. Convergent collimation is focused at a point in space; this is the same type of system used in the rectilinear scanner described above. However, in the camera case, the focal point is on the other side of the patient where this is no activity. Pinhole collimation may lead to either magnification or minification depending on the location of the object relative to the pinhole aperture.

Efficiencies of all collimators are relatively poor with pinholes becoming the worst at extended distances from the camera face. Typical values are on the order of 1×10^{-4} for commonly used parallel-hole types. Thus, if an experimenter deals with a very flat (essentially 2D) source, such as a thin radioactive tissue sample, it is better to simply remove all collimation and use the intrinsic localizing capability of the bare crystal and attached PM system. A transparent plastic sheet should be placed between source and camera fact to minimize possible contamination.

Every collimator is designed to be effective at a given photon energy. Lead septae in the device are effectively four to five half-value layers for the quantum of interest. A half-value layer is that thickness of material that reduces the intensity of gamma radiation by a factor of 2. Thus, using a collimator designed for high energy photons in the case of a relatively low energy emitter will lead to both lower efficiency as well as poorer image quality. For radionuclides emitting several different photons, the collimation must be appropriate for the highest gamma ray energy being measured. If this is not done, a hazy background of events due to these photons passing through the collimator walls will obscure the image.

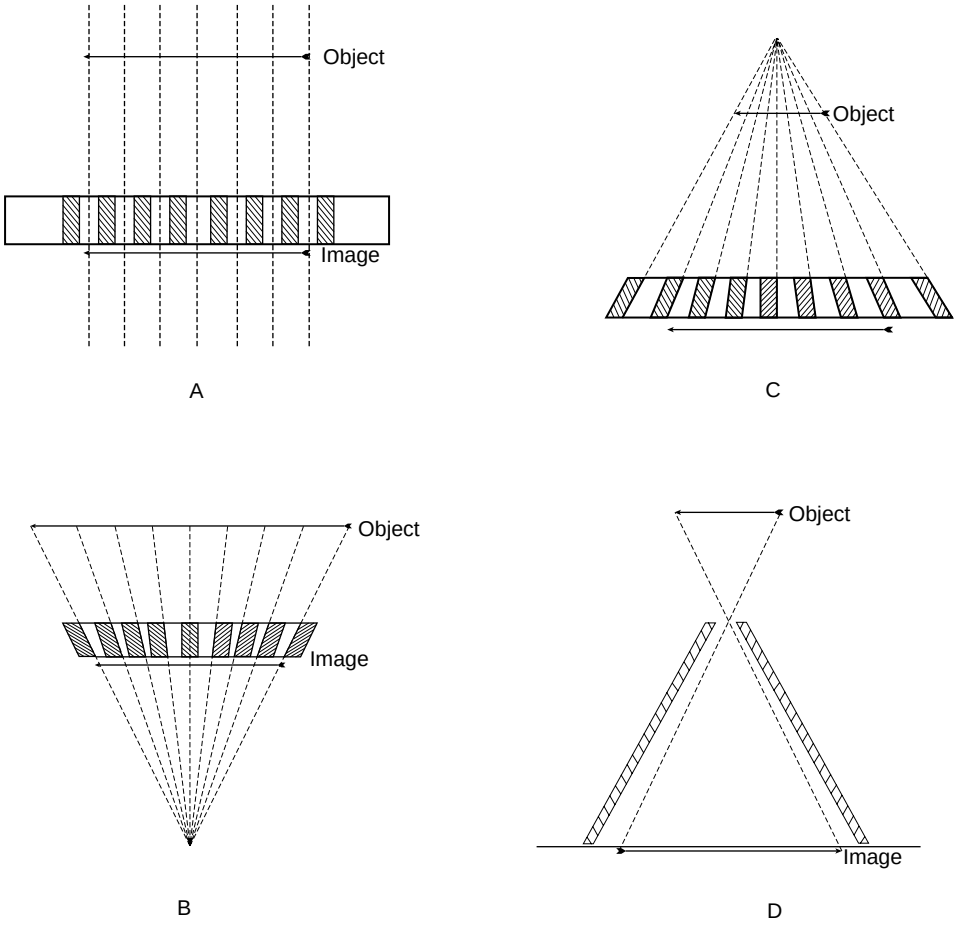


Figure 7. Four standard types of collimators used on the gamma camera.

Spatial resolution of gamma camera systems is on the order of 1 cm near the collimator surface, but generally becomes worse with increasing source depth inside the patient. When the count rate becomes extremely elevated, however, the localizing algorithms of such devices can be confused by multiple simultaneous scintillations with resulting imaging artifacts and reduced resolution. Most clinical protocols recognize this limitation by keeping the count rate at or below 5×10^4 counts per second (cps).

Because absolute measurement of resolution as well as object (organ) size is important, it is useful to image point sources of radiation for testing each camera-collimator system. This test object may be a set of small (1 mm) radiation sources of the imaged radionuclide having a known spacing. Resolution and object size in any resultant film or digital image can be defined directly using such devices. Variation with depth (patient thickness) and distance from the collimator may also be evaluated.

Digital Processor Applications Within the Camera Head

Anger's patented design originally relied on an analog computer to position the scintillation flash within the lateral dimensions of the NaI(Tl) crystal. Each scintillation event was weighted by location and signal amplitude of the several recording PM tubes. One of the original problems of the design has been the non-uniformity of response due to intrinsic and temporal variation in PM tube and other analog circuit components. In modern camera heads, digital processors are used to position the scintillation flash as well as perform spectroscopic analyses in real time on the detected events. Such dedicated processors inside the camera head observing individual PM tubes can greatly improve the uniformity so that the central field of view (CFOV) can have uniformities approaching 2%. Uniformity is particularly important for 3D imaging involving rotation of the camera head as described below. Values for each head are measured regularly with a flat source of radioactivity of an appropriate energy for most of the clinical imaging. Cobalt-57 is the radionuclide of choice for this procedure since it is close in emission energy (120 keV) to the common radiolabel ^{99m}Tc (140 keV) and has an extended half life of 270 d.

Note that communication formats are now available for information transfer between cameras and external computers. The digital imaging and communications in medicine (DICOM) standard is the international format for this transfer of information. This information may be used to produce comparisons of nuclear and other images to improve the diagnostic process.

Types of Acquisition from Gamma Cameras

One very important choice made by an operator prior to any camera study is the method of photon event recording in any external computer or work station memory. It is most common to acquire each scintillation as an event or count at coordinates (x, y). With total time of acquisition fixed at some realistic (patient-derived) limit, these events are added at their spatial positions to form a single digital image. This method of data recording is called frame mode. It is, by far, the most common type of camera data acquisition.

It may be that the timing of the tracer movement is either very rapid or uncertain for the patient–study. In that case, one may *a priori* choose list mode acquisition whereby each event is recorded as a triplet: (x, y, t) with computer clock time (t) included. After all events are list mode recorded, the operator or clinician may reconstruct the study in any sequence of time frames that is desired. For example, the first minute may be assigned to image 1, the second minute to image 2, and minutes 3–10 to image 3. Each of these images would appear to the reader as if they were taken in frame mode over that interval. Such an allotment may be revised subsequently as clinical questions arise. Large memory sizes are clearly useful if list mode imaging is to be pursued. Modern cameras often do not offer the possibility of list mode acquisitions, but instead rely on use of high speed frame-mode data recording.

A special type of frame mode acquisition is the gated study. Here, data are acquired in synchrony with a repeated physiological signal, usually the patient ECG. The R-wave-to-R-wave interval is predivided into a number (n) of equally spaced segments. Data obtained during time segment 1 of the cardiac cycle are placed into image 1, from time segment 2 into image 2, and so on. The result is a closed loop of n images that shows the beating heart when the gating signal is derived from the electrocardiogram (ECG).

External computer processing of camera data has been used to generate an additional type of output referred to as a functional image. For example, the clinician may wish to measure the rate of physiological clearance of a radiotracer from individual pixels within a time sequence of organ images. Using the external computer to calculate regional rate constants and to store this array, the resultant functional image displays the relative magnitudes of the computed kinetic values. Using an arbitrary scale, faster clearing regions are shown as brighter pixels. By looking at the functional image, regions of slower clearance can be readily identified and followed post subsequent therapies such as microsurgery for stroke patients.

Gamma Camera Types

Mobile Cameras. Battery-powered Anger cameras may be mounted on motor-driven chassis for use at the bedside or other remote areas. In such cases, the head is generally smaller than a static camera, on the order of 25 cm in diameter, and the energy range limited to 140 keV (^{99m}Tc) due to shielding weight concerns. Movement up ramps and using elevators would be restricted otherwise. Mobile units are most often utilized in planar heart work and have been involved in the testing of patients under escalating stress such as on a treadmill in cardiology. Patient evaluations in the OR or ICU are other applications of the device. Aside from breast imaging using ^{99}mTc -sestamibi, use of mobile gamma cameras has been limited, however, because of two specific reasons listed below.

Tomographic imaging is generally not possible with the mobile camera due to the difficulty of rotating the device in a rigorous orbit about the patient. In addition, use of high energy gamma labels is not possible for the minimally

shielded detector head. Because of the importance of 3D imaging of the heart (see below), clinical usage has dictated that the more optimal study results if the patient is brought to the nuclear medicine clinic in order that optimal tomographic images be obtained.

Static Single-Head Cameras. The most common camera type, the static single head, is usually a large rectangular device with a NaI(Tl) crystal having a thickness of ~6–9 mm. Larger thicknesses up to 25 mm may be useful for higher energy gammas, but loss of spatial resolution occurs as the PM location of the scintillation becomes more indeterminate. Lateral crystal dimensions are approximately 35×50 cm, although actual external size of the head would be significantly larger due to the necessity of having lead shielding surrounding the detector. This shielding must go both around the detector crystal as well as behind it to prevent radiation coming into the sensitive NaI(Tl) from the direction opposite the patient. Such protection is of importance in a busy clinical situation where more than one study is being conducted simultaneously in a relatively small space. Large rectangular camera heads permit simultaneous imaging over the entire width of a typical patient and allow whole body imaging with a single pass of the detector from the head to the feet. This is essentially an updating of the rectilinear scanner concept although here it is a 1D motion (z).

Images from Single-Head Cameras. Two standard imaging formats are employed with the gamma camera. Regional images, or vignettes, are taken of the organs of interest in the clinical study. A patient complaining of pain in the knee will be placed adjacent to the camera to permit various views of that joint following administration of a bone-seeking radiotracer, such as ^{99m}Tc-MDP. In addition, a whole-body image may be acquired to check for overall symmetry of tracer uptake. An example of the latter is given in Fig. 8. Here, the camera head is driven from the head to the foot of the patient and a series of frame images acquired over a span of 20–30 min. A computer attached to the camera allows these separate images to be seamlessly united to form the whole-body format.

Anger's camera concept has had one of its greatest impacts in cardiac dynamic imaging, whereby the sequential heart images are stored in a repetitive sequence that is correlated to the ECG signal obtained from the patient as described above. Figure 9 includes a continuous loop of 16 images of the left ventricle during a cardiac cycle using a labeled red cell tracer based on ^{99m}Tc. By setting a computer-generated region of interest (ROI) over the ventricle, one can measure the relative amount ejected; that is, the left ventricular ejection fraction (LVEF). Note that absolute amount of the tracer is not needed in the study since it is only a fractional ejection fraction that is of interest to the cardiologist. Irregular heart beats and/or patient motion during the 10–20 min of data taking can make such studies difficult to process.

Other dynamic studies are popular and clinically important. These include the renogram whereby the uptake and clearance of a filtered agent, such as ^{99m}Tc-DTPA is measured over a 1 h period. Both kidneys are followed and characteristic times of tracer accumulation and excretion are estimated by the radiologist: often using external computer software. A partial listing of typical studies involving gamma camera image data is included in Table 3.

Multiple-Head Cameras. It is becoming common to use more than one gamma detector head within a single

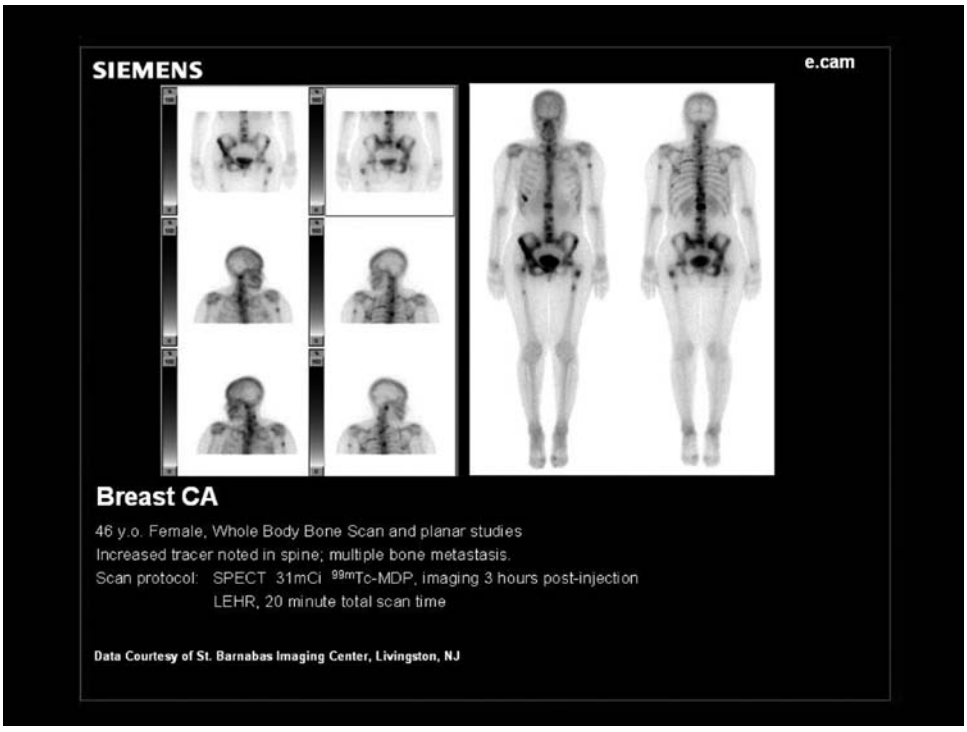


Figure 8. Whole-body image of a bone scan patient using translation of the gamma camera from head to foot. A sample of 20 mCi of ^{99m}Tc-MDP was used as the radiotracer for this image taken at 4 h postinjection.

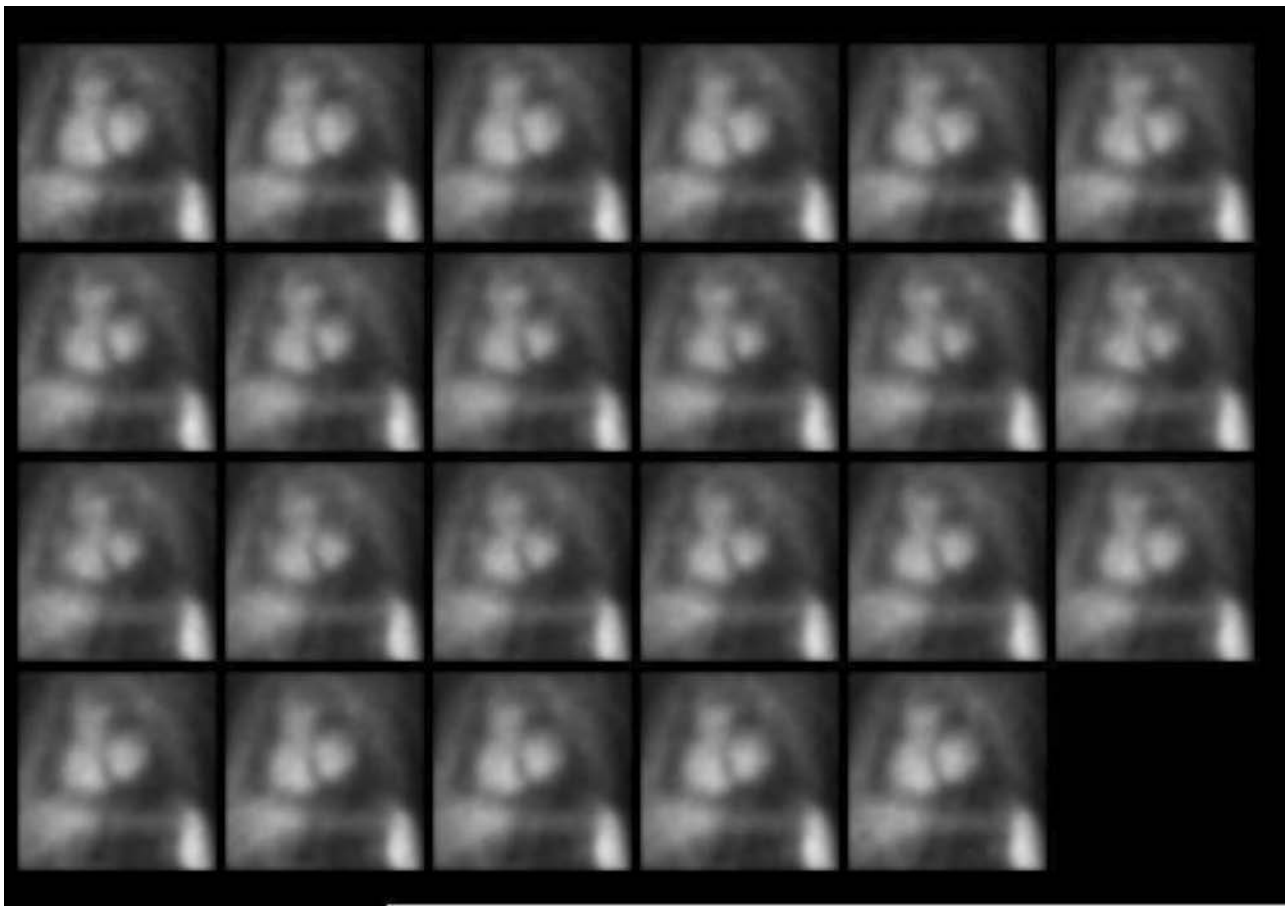


Figure 9. A multiple gated (MUGA) study of the left ventricle. The 16 images acquired over a heartbeat are uniformly assigned in time to the R wave-to-R wave cardiac interval. An ejection fraction of 69% was calculated.

supporting gantry (cf. Fig. 10). Speed of data acquisition, in either 2D or 3D mode, is the most important reason for this augmentation. By using two heads in a 2D study, the patient may be imaged from opposing sides simultaneously. Thus, if the organ of interest or tumor site were closer to the back of the patient, one could obtain information from the posterior head that would be useful even if the anterior head showed no discernible uptake sites. Alternatively, anterior and lateral views of an organ system may be obtained simultaneously and serially in a dynamic study of gastric emptying, for example. A second, and very important, application of multiple head camera systems is in more efficient 3D imaging.

THREE-DIMENSIONAL DETECTORS

There are two quite distinct methods to provide 3D imaging in nuclear medicine. If one uses ordinary (nonpositron) gamma-emitters, the strategy is referred to as single-photon emission computer tomography or SPECT.

SPECT Imaging

Here, the detector head or, more likely the set of two or three heads, is rotated around the patient over an extended arc. This orbit may be a full 360° arc or may be less due to body habitus or tissue location. One uses the rectangular Anger

Table 3. Representative Gamma Camera Imaging Studies Done in Nuclear Medicine

Study	Agent	Label	Device	Results
Renogram	DTPA and MAG3	^{99m}Tc	Camera	Kinetic values
MUGA	Red cells	^{99m}Tc	Camera with EKG gating	Ejection fraction of LV
Myocardium	Sestamibi	^{99m}Tc	Camera	Bulls eye image of LV
Bone scan	MDP	^{99m}Tc	Camera	Fracture location. Tumor location
Lung scan	Aggregated albumin	^{99m}Tc	Camera	Regions of reduced perfusion
Lung scan	Aerosolized albumin	^{99m}Tc	Camera	Regions of reduced ventilation
Lung scan	Xenon gas	^{133}Xe	Camera	Regions of reduced ventilation
Thyroid imaging	Iodine	^{123}I	Camera	Uniformity of uptake in gland

Figure 10. Dual-headed gamma camera. Both detector heads are mounted on the same gantry to allow translation (for whole body) and rotation (for SPECT) of the system. An open geometry permits use of gurneys with this system.

head as described above with the parallel-hole collimation in place. With injected activities on the order of 100–300 MBq, data acquisitions require on the order of 20 min. Patient immobility is necessary. Data may be taken in a shoot-and-step mode at fixed angular intervals or they may be acquired continuously during the rotation. Storage of such vast amounts of information requires a dedicated computer system recording the counts at each spatial position on the head (x,y) and at each angle (θ) during the rotation.

Several reconstruction algorithms are available to the technologist to generate the requisite tomographic images of the patient. Corrections for attenuation and Compton scatter must also be applied for the generation of these images. While pseudo-3D images may appear on the computer monitor as an output of the reconstruction, the radiologist will review and file to the picture archival and communication system (PACS) system the transaxial, sagittal, and coronal projections of the activity. It is important to realize that numerical values usually shown in these various projection images are not absolute, but only relative quantities. Quantitative SPECT, in which the numerical pixel value is equal (or at least proportional) to the activity in Bq, requires, in addition to the above corrections, that a set of standard sources of the same radionuclide be imaged along with the patient. Such calibrations can be done simultaneously with the clinical study, but are usually performed as a separate procedure. Figure 11 shows the three projection sets (axial, sagittal and coronal) in the case of a patient having a ^{99m}Tc sestamibi myocardial scan of the left ventricle.

PET Systems

Back-to-back photon emission (511 keV each) characteristic of positron decay of a labeling radionuclide has led to the development of PET. While paired Anger camera heads have been used as the detectors, it is much more efficient to use a ring of solid-state scintillation detectors arrayed around the patient. Bismuth germinate (BGO) has been the standard material, but LSO (lutetium orthosilicate) is becoming more popular due to its higher light output and shorter pulse length at 511 keV. In the standard situation, each detector block is broken into separate light emitting substructures that act as individual scintillation detectors. By having a few phototubes observing a separate block of such elements, the number of PMs may be reduced using Anger's gamma camera principle. Whole body PET scanners may have $> 10^4$ individual scintillators arrayed in an open circle or set of rings around the patient bed. Multiple rings are conventional so that several axial sections may be acquired simultaneously over a distances of 10–15 cm. Note that no detector rotation is inherently required since the solid-state system completely encircles the patient. If needed, the bed will be driven along the axis of the detector rings in order to perform extended imaging of the subject. The most common study utilizes FDG with ^{18}F as the radiolabel and covers the patient from head to groin. Sites within the body that metabolize glucose are imaged thereby. Brain and possible tumor areas are important applications of PET glucose imaging. Ambiguity with infection sites is a limitation to this protocol; this is particularly the case in the immune-compromised patient.

Because the two emitted photons are coincident in time and define a line in space, the positron detection process does not, in principle, require collimation (Fig. 12). Using contiguous rings of detectors is the most common system design; if the rings act alone or together as a single detector system defines the two types of imaging that are performed on a PET system. Internal (patient) photon attenuation is taken into account in the reconstruction of the PET image set. This is done using a transmission source of positron emitter, usually ^{68}Ge, to evaluate the patient thickness for the various ray directions at each bed position. Typically, the attenuation correction occurs during the scanning procedure with a short time interval given over to use of the source at each bed location.

Two-Dimensional PET Imaging

A clinical PET scanner is shown in Fig. 13. In 2D PET, every ring of detectors is isolated by tungsten collimation from all but single adjacent rings. Thus, each circle of solid-state scintillators is used in isolation to generate a single axial slice through the patient. This approach yields the highest resolution available in positron tomography with systems having spatial resolutions on the order of 5 mm. Reduction in the amount of scatter radiation is also obtained in 2D images. A FDG image is given in Fig. 14. While described as 2D, the result is actually tomographic and gives the usual projections in the three planes intersecting the patient's body. In these planes, the precise estimate of resolution depends on the positron's kinetic energy. One must combine, in quadrature, the positron range in soft tissue with inherent ring resolution to predict the overall spatial distance ambiguity. Higher energy positron emitters will have correspondingly poorer spatial

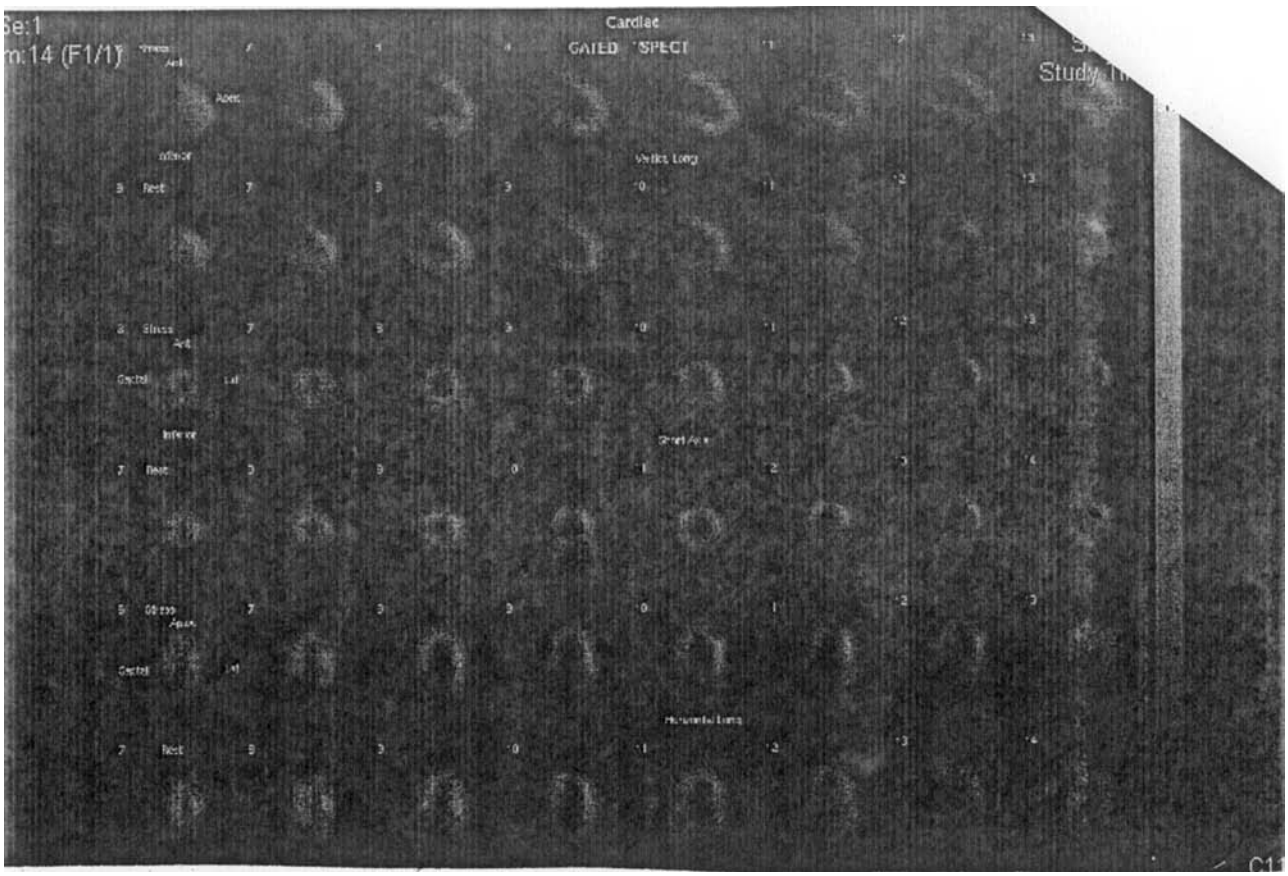


Figure 11. The SPECT image set for a gated myocardial study. In each pair of rows, the upper set of images gives the stress result, the lower set the resting result. The patient received 30 mCi of ^{99m}Tc-sestamibi for the study.

resolution due to the greater range of the positron prior to decay.

Other criteria for the selection of a positron label may be applied; for example, the half-life of the radionuclide. If that lifetime is very short, manufacturing and targeting may take so many physical half-lives that imaging is not possible. Additionally, one should consider the relative probability of β^+ emission in the decay scheme. This likelihood may be reduced because of competition with electron capture from the K shell of the radionuclide. Additionally, there is the possibility that other photons may be emitted along with the positrons so as to cause a background effect in the PET scanner. For example, ^{124}I, along with annihilation radiation at 511 keV, also emits ordinary gamma rays with energy in excess of 2 MeV. Such high energy photons readily penetrate collimators to reduce contrast in the images and make quantitation of the absolute radioiodine activity difficult.

Three-Dimensional PET Imaging

When the collimation between PET scanner rings is removed, each circle of detectors can have coincidences with itself as well as with all other detector rings. This mode of operation is referred to as 3D imaging. Spatial resolution is somewhat worse than that of the collimated (2D) case and may be 1 cm or more. However, the added sensitivity may be very important: particularly if whole body images are to be obtained in a patient with possible multiple sites of interest such as a referral from medical oncology. Sequential PET images of the whole body may be used to evaluate chemotherapy or other interventions. A quantitative method is available for such comparisons.

One feature of PET imaging merits emphasis. In the quality assurance of the positron scanner, the operator will routinely obtain transmission images through a phantom of known size using 511 keV photons from an external source. With this information and calibration using a known activity source, the user may reconstruct radioactivity distributions in the patient with absolute units. Thus, the concentration of positron emitter at a given image voxel can be estimated. Called the specific uptake value (SUV), this parameter is essentially $\%\mathrm{ID}\cdot\mathrm{g}^{-1}$, where ID refers to the injected activity or dose (MBq). The resultant SUV value is a function of time. Two direct consequences result. First, the clinician can make comparisons between organ sites both now and with regard to earlier studies on that patient or relative to normal individuals. Results of therapy may be directly evaluated thereby. The SUV values may even be used to make diagnostic assessments, such as the likelihood of malignancy at the voxel level. In addition, the radiation dose to the entire organ and even to local volumes within the tissue may be directly made with the SUV parameter. This

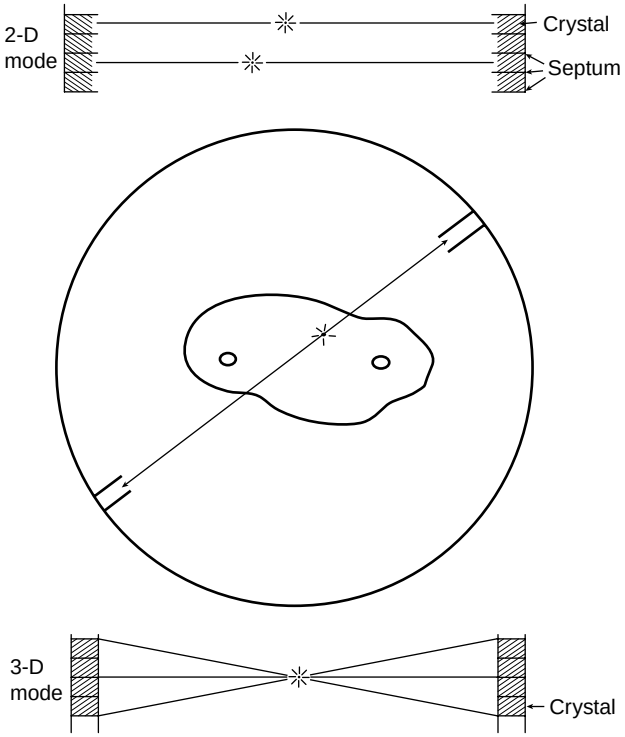


Figure 12. Principle of a PET scanner. Note that the direction of the annihilation radiation defines a line in space. 2D and 3D configurations are accomplished with and without collimation, respectively.

is in contrast to gamma camera planar data whereby the results may be quantified only with associated calculations that depend upon acquiring a set of images from at least two sides of the patient.

HYBRID IMAGING INSTRUMENTS

Nuclear image information, of either gamma camera or PET type, is limited in that regions of elevated (or reduced) activity are not necessarily identifiable as to anatomic location or even organ type. A patient may exhibit a hot spot in a planar gamma camera view that could correspond to uptake in a lobe of a normal organ, such as the liver or perhaps to an adjacent metastatic site. Similar arguments may be made with SPECT or PET images. Clinical decisions and surgical options are difficult to determine in this ambiguous context. Radiologists viewing nuclear medicine images are forced to cloak their patient assessments in correspondingly vague spatial terms.

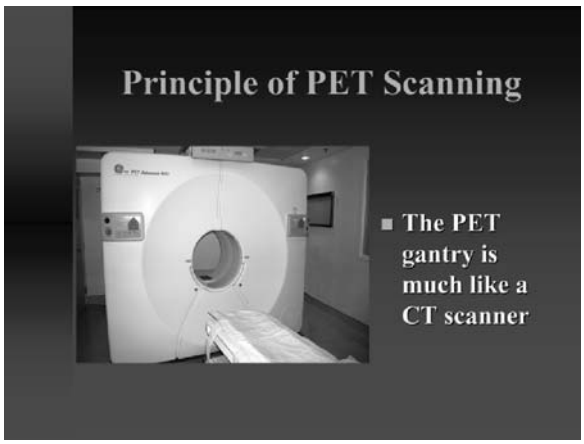
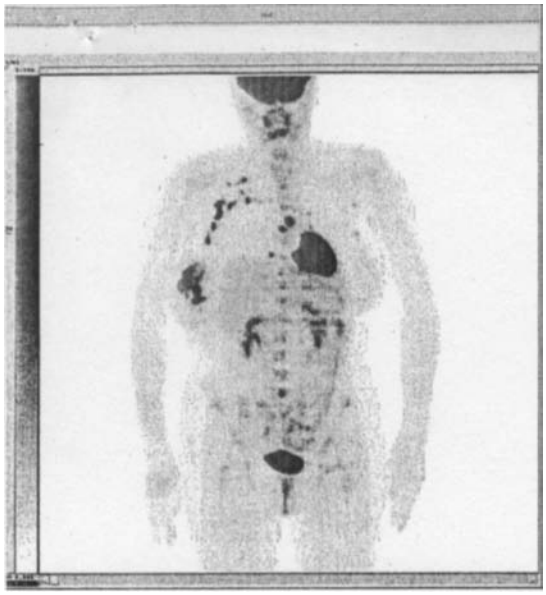


Figure 13. A clinical PET scanner.



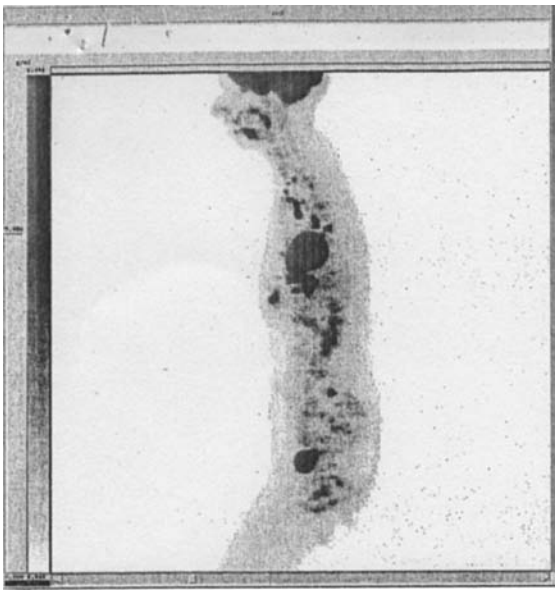


Figure 14. A PET image of a breast cancer patient following injection of 10 mCi of FDG. A MIPS projection is shown with areas of elevated FDG appearing as dark foci. Note accumulation in regional lymph nodes near the breast primary.

Lack of anatomic correlation has been one of the most difficult issues in the history of nuclear imaging. Physiological data determined with nuclear techniques are considered complementary to anatomical information separately obtained by other imaging modalities such as

CT or magnetic resonance imaging (MRI). The radiologist or referring clinician will frequently have to conceptually fuse disparate data sets to help identify the specific organ or tissue where a nuclear tracer uptake zone occurs. Using DICOM and PACS technologies, one may also attempt to digitally overlay nuclear and anatomic images. In this case, however, magnification, rotation, and translation of one image relative to the other must be accounted for with appropriate software and adjustable parameters. Using commercial programs, CT and MRI digital images may be fused to nuclear imaging results using least-squares techniques and an external workstation.

In order to remove this conceptual and computational bottleneck, recent developments in nuclear medicine have included manufacture of hybrid physiologic/anatomic imagers. In this strategy, both devices share a common patient bed so that two types of images are spatially registered and, although successive, nonetheless obtained within a few minutes of each other. Note that the PM tubes of a typical gamma camera or PET system are sensitive to magnetic field effects at the level of the earth's value; that is, at ~0.5 G. Yet clinical MRI scanners operate in the range of 1.5–3.0 T (1.5×10^4 to 3.0×10^4 G) so that hybrids of MRI and nuclear devices would be problematic. Thus, essentially all of the hybrid systems have involved combinations of nuclear and CT imagers.

SPECT/CT Hybrid Imagers

A logical approach to the issue of radionuclide localization is to have two scanners, one nuclear and one based on X-ray attenuation, located on attached gantries. This pair of devices shares the same patient couch. Because the distances of bed movement can be known within 1 mm or less, the user can identify an uptake volume in the nuclear SPECT image with a geometrically corresponding part of the anatomy as seen via CT scan. Additionally, attenuation corrections may be made more effectively using the CT data to improve SPECT sectional images. Some difficulties remain: (*1*) the breathing motion of the patient, and (*2*) possible changes in posture from one sequence to the other during the double imaging procedure. Complementary nature of the two images makes the interpretation of either somewhat clearer.

PET–CT Hybrid Imagers

Analogous to the gamma camera, a PET detector ring imager can be mounted adjacent to a CT scanner to provide registration of images from two modalities. As in the case of SPECT–CT devices, disparities in the speed of the two data acquisitions leads to some remaining ambiguity involving organs that move with respiration such as liver or lungs. While it is possible to hold one's breath for a CT scan, the PET whole body nuclear imaging time remains on the order of 20–30 min to preclude such possibilities for the emission segment of the study. A set of hybrid images and their superimposition are given in Fig. 15.

Radiation therapy treatment planning has been one of the primary beneficiaries of hybrid imaging devices. It may be that some mass lesions visible via CT or other anatomic imagers are necrotic or at least not active metabolically. This result can most clearly be seen in the fused image so that the more physiologically active sites may be treated with higher external beam doses. Likewise, with appropriate resolution, the radiation oncologist may elect to treat part of a lesion that has heterogeneous tracer uptake in an effort to spare contiguous normal (albeit sensitive) sites, such as in the lung, spinal cord, or brain. Those segmental regions of a tumor mass that are metabolically active may be targeted with external beam therapy using a number of linear accelerator strategies including conformal therapy, intensity modulated radiation therapy (IMRT) and tomotherapy using a rotating radiation source.

ANIMAL IMAGING DEVICES

As indicated previously, the growth of nuclear medicine is limited by availability of specific radiopharmaceuticals. Historically, useful agents were often discovered (sometimes by accident) and were almost never invented. This strategy is inefficient and modern molecular biologists and pharmacists attempt to directly engineer improved tracers for a given clinical objective; that is, imaging or therapy of a particular tissue or tumor type. A specific molecule or cellular organelle is generally the target in these efforts. Molecular imaging has become an alternative name for nuclear medicine. After initial protein or nanostructure development is completed, the next task is the determination of the relative usefulness of the prototype in an animal study. Usually, this work involves mouse or rat radiotracer biodistributions involving sacrifice of 5–10 animals at each of a number of serial times. If multiple time points and comparison of various similar radiotracers are involved, numbers of mice may approach thousands for the development of a single radiopharmaceutical.

It is more analogous to clinical procedure if serial images of the same animal are obtained during the course of the research study. Far fewer animals are required and the data are more homogenous internally. Imaging with standard-sized nuclear technology is generally unsatisfactory due to poor spatial resolution associated with typical gamma cameras (1 cm) or PET scanners (0.5 cm). Early investigators had utilized a suitably small pinhole collimator and gamma camera combination on mouse and rat imaging studies. By collimator magnification, the image can be made large enough that the internal structures can be resolved. As noted previously, magnification and sensitivity depend on distance from the pinhole so that quantitative interpretation of these images was difficult. Sensitivity of pinhole imaging was likewise low so that relatively large amounts of activity were required for the study. It is more effective if a dedicated, high efficiency, animal-size imaging device is designed for the experimental species. Such instruments have been developed for planar and SPECT gamma camera as well as PET imager systems.

Animal Gamma Cameras

Imaging a 10 cm mouse is best done with a gamma camera having approximately that sized crystal. Rather than employing a hexagonal array of multiple, miniaturized PM tubes to locate the scintillation, an animal camera

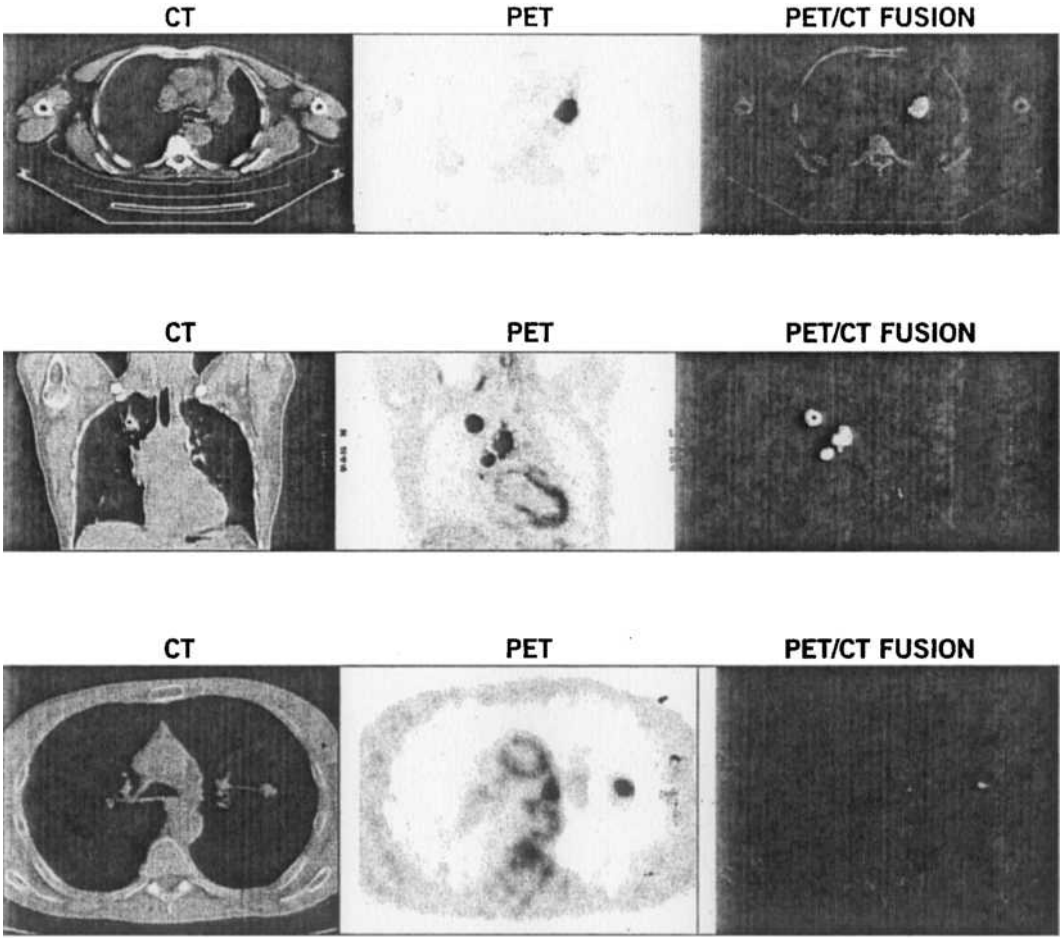


Figure 15. The CT–PET hybrid image showing respective CT, PET, and combined images. Clarity of location follows from the last of these results.

relies on the use of a single spatially sensitive PM tube. This device sends both x and y coordinates and the energy of the scintillation to a dedicated computer. Otherwise, the murine camera is operated essentially identically to the full-size version. Parallel-hole collimation is most common, although pinholes may be used to form highly magnified images of murine organs, such as the liver, kidneys, or even the thyroid. Figure 16 illustrates the last of these targets for a mouse receiving a tracer injection of ^{125}I to enable imaging of the murine thyroid. SPECT imaging is also possible; it is accomplished by rotating a rigorously constrained mouse or other small animal within the field of view of the camera. The usual projections, coronal, sagittal and transaxial are then available.

Animal PET Imagers

Miniature PET scanners have become of importance to the development of new radiopharmaceuticals. Here, a ring of BGO or LSO crystals is installed in a continuous cylinder extending over the entire length of the mouse. Spatial resolution is on the order of 2 mm or less over the 12 cm axial dimension. A sample image is given in Fig. 17 where a number of coronal sections are superimposed to improve the image statistics. Both ^{18}F-FDG and ^{64}Cu labeled to a modified antibody protein called the minibody were the positron emitters used in this study. Again, as in the clinical case, the PET images are intrinsically tomographic unlike the gamma camera results. Therefore, the PET animal imagers have a theoretical advantage in biodistribution

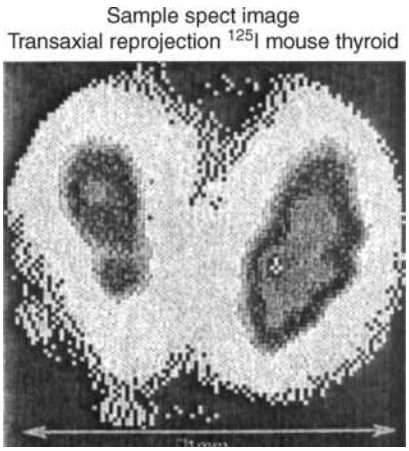


Figure 16. Animal gamma camera image of a mouse thyroid. Iodine-123 was used as the tracer with a pinhole collimator to obtain an image of the normal organ.

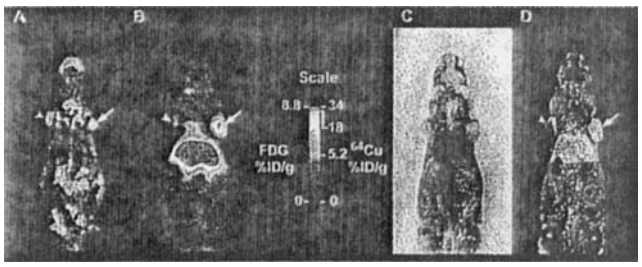


Figure 17. A PET animal scanner murine image. Both FDG (Image a) and ^{64}Cu-minibody (Image b) were used as radiotracers. Images c and d show pathology and autoradiographic results, respectively.

assays. As mentioned there remains the difficulty of finding a suitable positron emitter and method of attachment for a particular imaging experiment.

Hybrid Animal Imaging Devices

Quantitation of radioactivity at sites within the mouse's body is more easily done with an animal gamma camera than in the comparable clinical situation. This follows since attenuation of photons is relatively slight for a creature only a few cm thick in most cross-sections. Because of this simplicity, the output of the small animal gamma camera imaging systems can be modified to yield percent-injected dose (%ID). In order to correct for organ perfusion, it is historically conventional in biodistribution work using sacrificed animals to obtain uptake in %ID/g of tissue. Given the organ %ID, this last parameter may be obtained if the total mass of the target organ can be determined. Two avenues are available; one may employ miniaturized CT or a reference table of organ sizes for the particular strain of animal being imaged. We should note that suitably sized CT scanners are produced commercially and may be used to estimate organ mass. Hybrid SPECT/CT, PET–CT and SPECT–PET–CT animal imagers are now available for mouse-sized test animals.

One caveat regarding the small-scale imaging devices should be added; these systems cannot give entirely comparable results to biodistribution experiments. In animal sacrifice techniques, essentially any tissue may be dissected for radioactivity assay in a well counter. Miniature cameras and PET systems will show preferentially the highest regions of tracer accumulation. Many tissues may not be observable as their activity levels are not above blood pool or other background levels. Hybrid animal scanners can reduce this limitation, but not eliminate it entirely. Those developing new pharmaceuticals may not be concerned about marginal tissues showing relatively low accumulation, but regulatory bodies, such as the U.S. Food and Drug Administration (FDA), may require their measurement by direct biodistribution assays.

BIBLIOGRAPHY

Reading List

Aktolun C, Tauxe WN, editors. Nuclear Oncology, New York: Springer; 1999. Multiple images, many in full color, are presented of clinical studies in oncology using nuclear imaging methods.

Cherry SR, Sorenson JA, Phelps ME. Physics in Nuclear Medicine, 3rd ed. Philadelphia: Saunders; 2003. A standard physics text that describes SPECT and PET aspects in very great detail. This book is most suitable for those with a physical science background; extensive mathematical knowledge is important to the understanding of some sections.

Christian PE, Bernier D, Langan JK, editors. Nuclear Medicine and PET, Technology and Techniques, 5th ed. St. Louis: Mo. Mosby; 2004 The authors present a thorough description of the methodology and physical principles from a technologist's standpoint.

Conti PS, Cham DK, editors. PET-CT, A Case-Based Approach, New York: Springer; 2005. The authors present multiple hybrid (PET/CT) scan case reports on a variety of disease states. The text is structured in terms of organ system and describes the limitations of each paired image set.

Sandler MP, et al. editors. Diagnostic Nuclear Medicine, 4th ed. Baltimore: Williams and Wilkins; 2002. A more recent exposition that is a useful compilation of imaging methods and study types involved in diagnosis. No description of radionuclide therapy is included.

Saha GP, Basics of PET Imaging. Physics, Chemistry and Regulations, New York: Springer; 2005. This text is a useful for technical issues and is written at a general level for technologists. Animal imaging is described in some detail and several of the commercial instruments are described.

Wagner HN, editor. Principles of Nuclear Medicine, Philadelphia: Saunders; 1995. The Father of Nuclear Medicine is the editor of this reasonably recent review of the concepts behind the field. A rather complete but somewhat dated exposition of the entire technology of nuclear medicine operations in a medical context.

Wahl RL, editor. Principles and Practice of Positron Emission Tomography. Philadelphia: Lippincott Williams and Wilkins; 2002. A solid review of PET clinical principles and practical results. Logical flow is evident throughout and the reader is helped to understand the diagnostic process in clinical practice.

See also COMPUTED TOMOGRAPHY, SINGLE PHOTON EMISSION; NUCLEAR MEDICINE, COMPUTERS IN; POSITRON EMISSION TOMOGRAPHY; RADIATION PROTECTION INSTRUMENTATION.

NUCLEAR MEDICINE, COMPUTERS IN

PHILIPPE P. BRUYANT
MICHAEL A. KING
University of Massachusetts
North Worcester, Massachusetts

INTRODUCTION

Nuclear medicine (NM) is a medical specialty where radioactive agents are used to obtain medical images for diagnostic purposes, and to a lesser extent treat diseases (e.g., cancer). Since imaging is where computers find their most significant application in NM, imaging will be the focus of this article.

Radioactive imaging agents employed to probe patient pathophysiology in NM consist of two components. The first is the pharmaceutical that dictates the *in vivo* kinetics

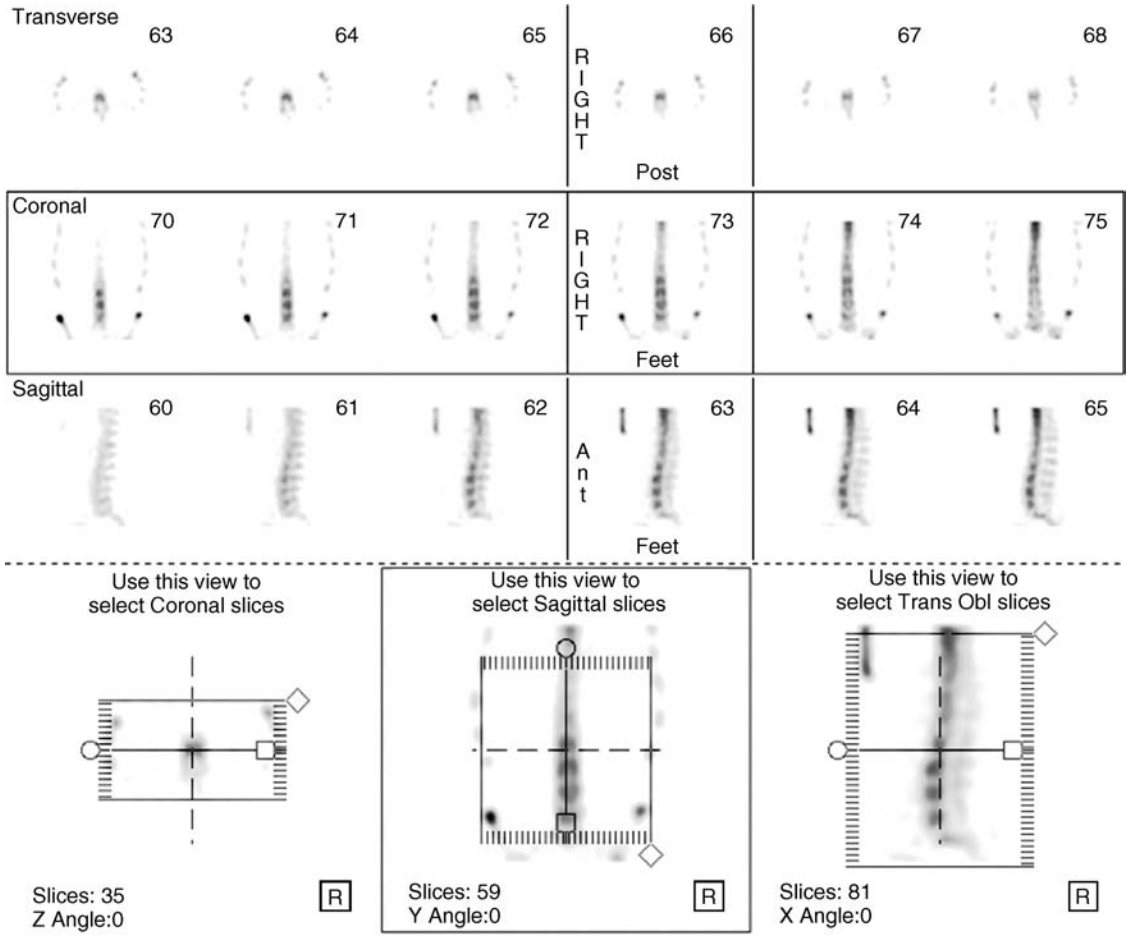


Figure 1. Images after a bone scan.

or distribution of the agent as a function of time. The pharmaceutical is selected based on the physiological function it is desired to image. The second component is the radionuclide that is labeled to the pharmaceutical and emits radiation that allows the site of the disintegration to be imaged by a specifically designed detector system (1). An example of an imaging agent is technetium-99 m labeled diphosphate, which is used to image the skeleton. The diphosphate is localized selectively on bone surfaces by 3 h postinjection and the technetium-99 m is a radionuclide that emits a high energy photon when is decays. A normal set of patient bone images of the mid-section is shown in Fig. 1. Another imaging agent example is thallium-201 chloride, which is localizes in the heart wall in proportion to local blood flow. In this case, thallium-201 is both the radiopharmaceutical and radionuclide. A normal thallium-201 cardiac study is shown in Fig. 2. A final example is the use of an imaging agent called fluorodeoxyglucose (FDG), which is labeled by the positron emitting fluorine-18. As a glucose analog, FDG is concentrated in metabolically active tissue such as tumors. Figure 3 shows a patient study with FDG uptake in a patient with lung cancer. Dozens of tracers are available to study a variety of pathologies for almost all organs (heart, bones, brain, liver, thyroid, lungs, kidneys, etc.).

Because the amount of radioactivity and the imaging duration are kept at a minimum, NM images are typically noisy and lack detail, compared to images obtained with other modalities, such as X-ray computerized tomography (CT), and magnetic resonance imaging (MRI). However, CT and MRI provide mainly anatomical information. They provide less functional information (i.e., information regarding the way organs work) in the part because these techniques are based on physical properties (such as tissue density. . .) that are not strikingly different between normal and abnormal tissues. Actually, after recognizing the differences between the anatomically and physiologically based imaging techniques, the current trend in the diagnostic imaging strategies is, as seen below, to combine anatomical information (especially from CT) and functional information provided by NM techniques (2,3).

As seen below, computers play a number of fundamental roles in nuclear medicine (4). First, they are an integral part of the imaging devices where they perform a crucial role in correcting for imaging system limitations during data acquisition. If the acquired data is to be turned from two-dimensional (2D) pictures into a set of three-dimensional (3D) slices, then it is the computer that runs the reconstruction algorithm whereby this is performed.

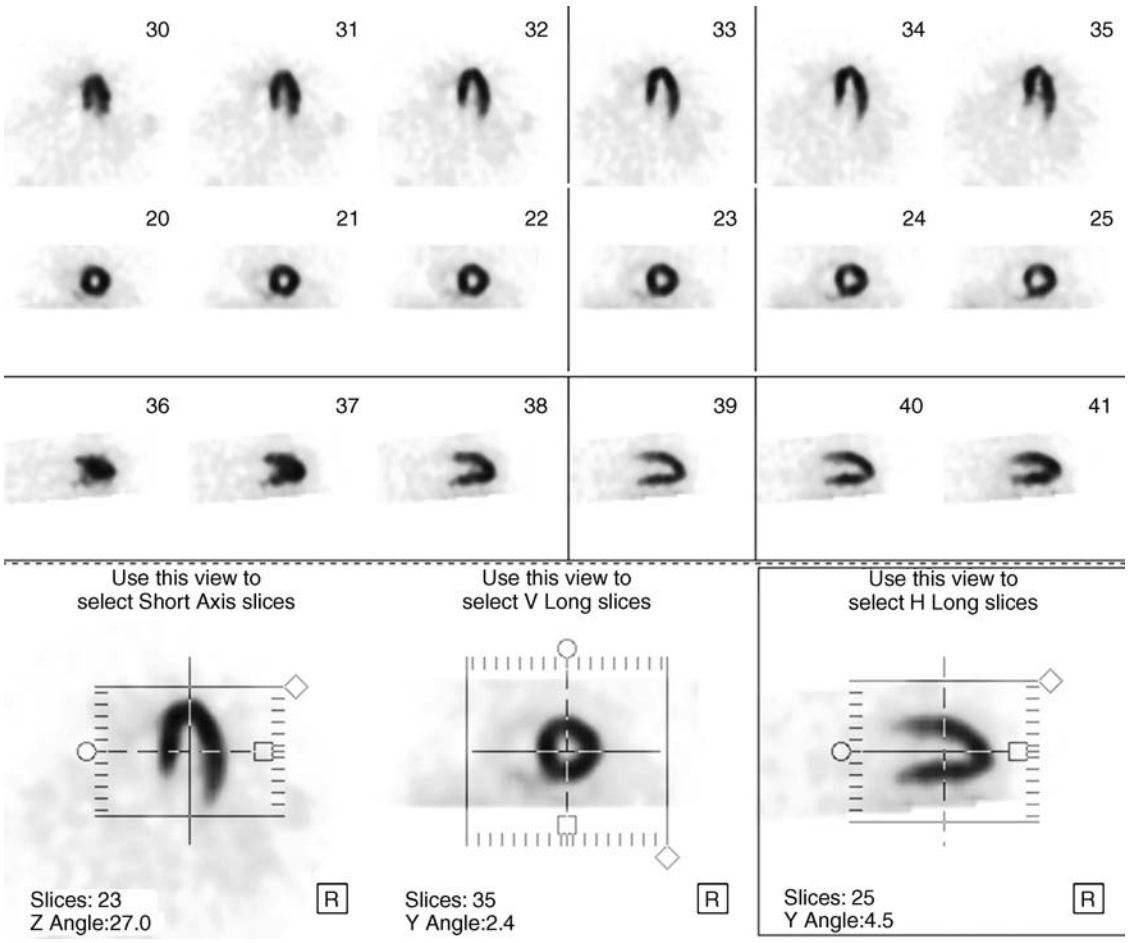


Figure 2. Normal thallium-201 cardiac study. The first three rows show six slices of the left ventricle in three different axes (vertical long axis, short axis, horizontal long axis) of the heart. The fourth row shows how the images can be reoriented along each axis.

Once the final set of pictures are ready for clinical use, then it is the computer that is used for image display and analysis. The computer is also used for storage of the clinical studies and to allow their use by multiple readers at various sites and time points during patient care as required for optimal usage of the diagnostic information they provide. They are also very useful in research aimed at optimizing imaging strategies and systems, and in the education and training of medical personnel.

NM IMAGING

Computers play an essential role in NM as an integral part of the most common imaging device used in NM, which is a gamma camera, and in the obtention of slices through the body made in emission computerized tomography (ECT). Emission CT is the general term referring to the computer-based technique by which the 3D distribution of a radioactive tracer in the human body is obtained and presented as a stack of 2D slices. The acronym ECT should not be confused with CT (for computerized tomography), which refers to imaging using X rays. Historically, the use of two different kinds of radioactive tracers has led to the parallel evolution of two types of ECT techniques: Single-photon emission computerized tomography (SPECT) and positron emission tomography (PET). The SPECT technology is used with gamma emitters, that is, unstable nuclei whose disintegration led to the emission of high energy photons, called γ rays. Gamma rays are just like X rays except that X rays are emitted when electrons loose a good amount of energy and γ rays are emitted when energy is given off as a photon or photons during a nuclear disintegration, or when matter and antimatter annihilate. As the name implies, the gamma camera is used with gamma emitters in planar imaging (scintigraphy) where 2D pictures of the distribution of activity within a patients body are made. An illustration of a three-headed SPECT system is shown in Fig. 4. The evolution of SPECT systems has led to a configuration with one, then two and three detectors that are gamma-camera heads. The positron emission tomography is used with emitters whose disintegration results in the emission of a positron (a particle similar to an electron, but with the opposite charge making it the antiparticle to the electron). When a positron that has lost all of its kinetic energy hits an electron, the two annihilate and two photons are emitted from the annihilation. These two photons have the same energy (511 keV) and opposite directions. To

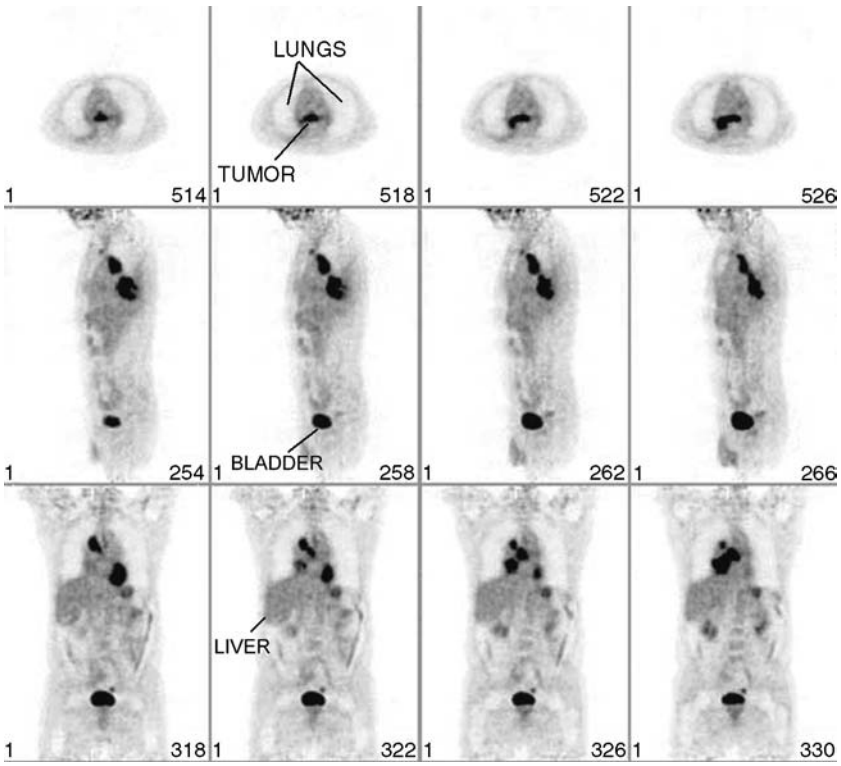


Figure 3. The FDG study for a patient with lung cancer. Upper to lower rows: transverse, sagittal, and coronal slices.

detect these two photons, a natural configuration for a PET system is a set of rings of detectors. The two points of detection of the opposite detectors form a line, called the line of response (LOR). A state-of-the-art PET system combined with an X-ray CT system (PET/CT) is shown in Fig. 5.

A gamma camera has three main parts: the scintillating crystal, the collimator, and the photomultiplier tubes (PMT) (Fig. 6). When the crystal (usually thallium-activated sodium iodide) is struck by a high energy photon (γ or X ray), it emits light (it scintillates). This light is detected by an array of PMT located at the back of the camera. The sum of the currents emitted by all the PMT after one scintillation is proportional to the energy of the incoming photon, so the γ rays can be sorted according to their energy based upon the electrical signal they generate. Because, for geometrical reasons, the PMT closer to the scintillation see more light than the PMT located farther away, the relative amounts of current of the PMT are used to locate the scintillation. This location alone would be of little use if the direction of the incoming photon was not known. The current way to know the direction is by using a collimator. The collimator is a piece of lead with one or more holes, placed in front of the scintillating crystal, facing the patient. Although different kinds of collimators exist, they are all used to determine, for each incoming photon, its direction before its impact on the crystal. To understand the role of the collimator, one can use the analogy of the gamma-camera with a camera that takes photographs. The

Figure 4. A three-headed SPECT system with the third head below the imaging bed the patient lays on.

Figure 5. Illustration of a state-of-the-art PET/CT system (Philips Medical Systems) with patient bed. The CT system is the first ring-shaped gantry and the PET system is the second ring-shaped gantry. (Reproduced with permission of Philips Medical Systems.)

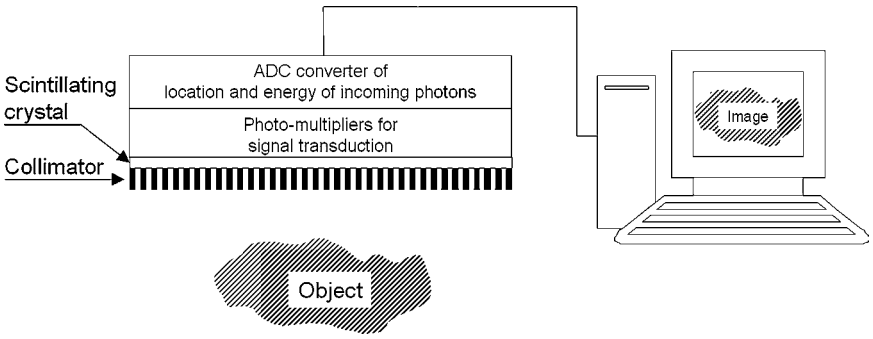


Figure 6. Main parts of a SPECT camera and basic principle. A gamma photon going through the holes of the collimator strikes the crystal. The crystal restitutes the energy of the gamma photon by scintillation, that is, by emitting some transient ultraviolet (UV) light. Some of the UV light is collected by photomultiplier tubes, whose function is to ensure the transduction of the signal (i.e., the conversion of light into electricity). The location of the scintillation and the energy of the photon are estimated, digitized and sent to the computer.

collimator plays the role of the objective lens in a camera. An image acquired without a collimator would be totally blurry, as would be a photograph taken with a camera with no lens. This is because γ rays are emitted in all directions with equal probabilities, and without a collimator, the photons emitted from a radioactive point source would strike the detector almost uniformly. With a collimator, only the photons whose direction is parallel to the axis of the holes may be potentially detected, while others are stopped by the lead. As a result, the image of a source is (ideally) the projection of the source distribution onto the crystal plane (Fig. 7). Gamma rays can be stopped or scattered, but due to their high penetrating power, it is very difficult to bend them like light rays with lenses, and this is the reason why collimators are used instead of lenses. Photons emitted at different distances from the camera, but along the same direction parallel to a hole, are detected at the same location in the crystal. Thus, the image obtained is the projection of the 3D distribution of the tracer onto the 2D plane of the detector. In that sense, a projection is similar to a chest X ray, in which the images of all the organs (ribs, heart, spine, etc.) are overlaid on the film even though organs do not spatially overlap in the body. The overlay might not be a problem for relatively thin parts of the body, such as a hand, or when tracer-avid structures do not overlap, such as the skeleton. In that case, only one projection is obtained from the best angle of view for the gamma-camera head. As stated above, this technique is called planar imaging, or scintigraphy. However, for other thicker organs like the myocardium and the brain, for which one is interested in measuring the 3D tracer inner distribution, more information is gathered by rotating the heads to acquire projections from multiple angles of view (tomographic acquisition, presented later in this article).

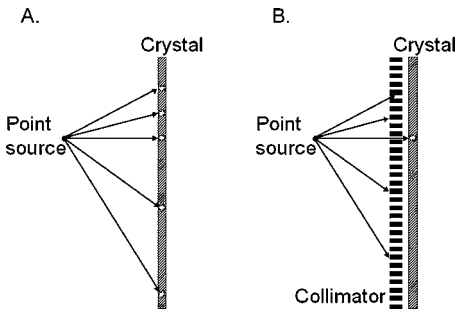


Figure 7. The role of the collimator in a SPECT system.

In PET scanners, hundreds of small crystals arranged in rings are used so that the data can be simultaneously acquired along multiple LOR (Fig. 8). Thousands of photons hit the crystals every second, so how to know which two photons are the result of the same electron-positron annihilation? If two photons are detected almost simultaneously, chances are that they are of the same pair (it is called a true coincidence), so an electronic circuitry checks whether one photon is detected $\sim$ 10 ns (the time window) at most after the previous one. It may happen that, although two photons are detected within that time window, they are not of the same pair, and such an event is called a random coincidence. Because in PET the direction of the photons is known (it is the LOR), collimators are not needed; however, because of the limited counting rate capabilities of older systems, septa made of lead may be used to limit the acquisition to the LORs roughly perpendicular to the axial direction, inside the same ring (2D acquisition). With modern PET systems having a high couting rate capability, a 3D acquisition is possible by detecting LORs even when the two photons hit crystals of different rings.

The computer plays an important role in the formation of the image coming from the gamma camera. As described above, the crystal is viewed by an array of 37 to $>$ 100, depending on the model, of PMT. These are analog devices that can drift with time. Also the positioning in the image of the location of the flash of light when a γ ray is absorbed in the crystal depends to some extent on where the ray interacts relative to the array of PMT. Such local variations lead to nonuniformity (uneven apparent sensitivity) and

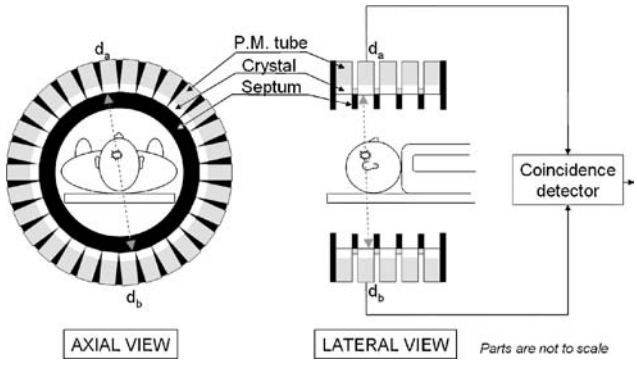


Figure 8. Main parts of a PET system. Pairs of photons are detected in coincidence following the annihilation of a positron with an electron.

nonlinearity (straight lines of activity are bent in the image). Prior to the incorporation of computers into the gamma camera electronics, the impact of such local variations was minimized by allowing the light to spread out before reaching the PMT by passing it through a light guide. This resulted in more PMT receiving enough light to participate in determining the location of the interaction thus improving uniformity and linearity, but at the expense of spatial resolution (i. e., determination of where in the crystal the flash of light originated). Modern gamma cameras incorporate computers to correct for local variations in camera performance so that the light guide is virtually eliminated. This in turn has improved spatial resolution.

Computer correction of the camera image usually takes place in three steps (5). The first is energy correction. As we said, the total magnitude of the signal from all the PMT is related to the energy deposited in the crystal by the γ ray. However, if a large number of γ rays of exactly the same energy interact in the crystal, the magnitude of the electrical signal will vary due to the statistics of turning the light emitted into an electrical signal and local variation in camera performance. By placing a source that will uniformly irradiate the crystal, such as the commercial sheet source shown in Fig. 9, the local variation on average in the magnitude of the signal can be determined on a pixel by pixel basis by computer. The centering of the window employed to select electrical pulses for inclusion in image formation can then be adjusted to give a more uniform response.

Besides varying in the average size of the total electrical pulse detected from the PMT locally, gamma cameras vary in how well they map the true location of the flash of light into its perceived location in the image. Thus, in some regions, detected events are compressed together and in others they are spread apart. Correction of this nonlinear mapping constitutes the second step in computer correction of the gamma-camera image and is called linearity correction. Linearity correction is performed by placing an attenuating sheet with an exactly machined array of very small holes in a precise location between the gamma

Figure 9. Radioactive sheet source in front of the third head of a three-headed SPECT system in position for checking uniformity and loading correction factors.

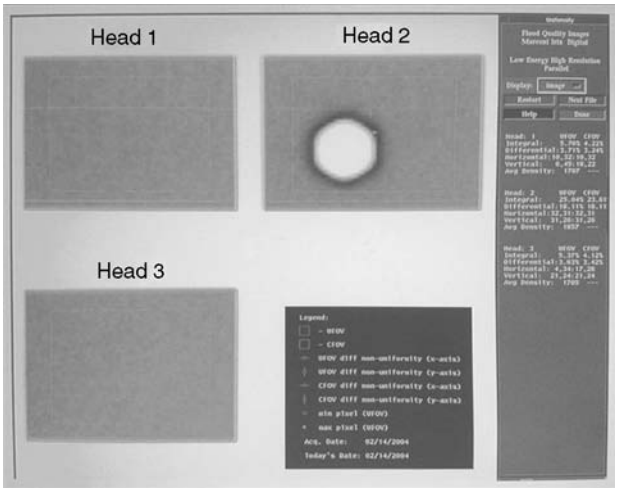


Figure 10. Output from checking camera uniformity when a single PM on head 2 of the three-headed SPECT system of Fig. 9 has failed.

camera and the sheet source of Fig. 9. A high resolution image consisting of a large number of gamma-ray events is then acquired. The images of the holes do not match where they should appear. The vector displacement of the image of the hole back to its true location defines how the mapping from true to detected location is inaccurate at that location. By using the computer to interpolate between the array of measured distortions at the pixel level, a map is generated giving how each event detected at a location in the crystal should be displaced in the resulting image.

The final step in image correction is called flood correction. If an image of a large number of events from a sheet source is acquired with energy and linearity correction enabled, then any residual nonuniformity is corrected by determining with computer a matrix that when multiplied by this image would result in a perfectly uniform image of the same number of counts. This matrix is then saved and used to correct all images acquired by the gamma camera before they are written to disk.

An example of testing camera uniformity is shown in Fig. 10. Here again, a sheet source of radioactivity is placed in front of the camera head as shown in Fig. 9. High count images of the sheet source are inspected numerically by computer and visually by the operator each day before the camera is employed clinically. Heads 1 and 3 in Fig. 10 show both numerically and visually good uniformity. A large defect is seen just below and to the left of center in the image from head 2. This is the result of the failure of a single PMT. A single PMT affects a region much larger than its size because it is the combined output of all the PMT close to the interation location of a gamma-ray that are used to determine its location.

Images can be classified in two types, mutually exclusive: analog or digital. A chest X ray on a film is a typical example of an analog image. Analog images are not divided into a finite number of elements, and the values in the image can vary continuously. An example of digital image is a photograph obtained with a digital camera. Much like roadmaps, a digital image is divided into rows and columns, so that it is possible to find a location given its row

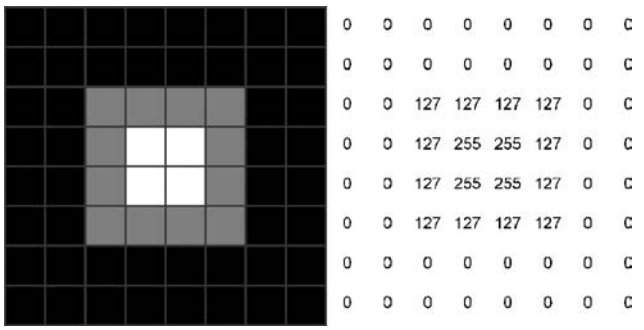


Figure 11. Left: example of an 8×8 image. Each square represents a pixel. The dark gray borders of each square have been added here for sake of clarity, but are not present in the image when stored on the computer. Each pixel has a level of gray attached to it. Right: the values in each pixel. In its simplest form, the image is stored in a computer as a series of lexicographically ordered values. The rank of each value in the series defines the pixel location in the image (e.g., the 10th value of the series refers to the 10th pixel of the image). There is a 1:1 relationship between the brightness and the value. The correspondence between the color and the value is defined in a table called look-up table (LUT) or a color map. An image can be a shade of grays (black and white) or in color.

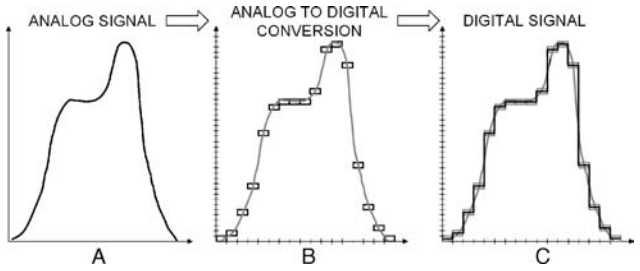


Figure 12. (a) Example of an analog signal, for example the intensity of gray (vertical axis) along a line (horizontal axis) drawn on a photograph taken with a nondigital camera. (b) The process of ADC, for example when the photograph is scanned to be archived as an image file on a computer. The 2D space is divided into a limited number of rectangular cells as indicated by the tick marks on both axes. For sake of clarity, only the cells whose center is close to the analog signal are drawn in this figure, and this set of cells is the result of the ADC. (c) The digital signal is drawn by joining the center of the cells, so that one can compare the two signals.

and column. The intersection of a row and a column defines one picture element, or pixel, of the image. Each pixel has a value that usually defines its brightness, or its color. A digital image can be seen as a rectangular array of values (Fig. 11), and thus be considered, from a mathematical point of view, as a matrix. Computers cannot deal with analog values, so whenever an analog measurement (here, the current pulse generated after the impact of a gamma photon in the crystal) is made, among the first steps is the analog-to-digital conversion (ADC), also called digitization. The ADC is the process during that an infinite number of possible values is reduced to a limited (discrete) number of values, by defining a range (i.e., minimum and maximum values), and dividing the range into intervals, or bins (Fig. 12). The performance of an ADC is defined by its ability to yield a digital signal as close as possible to the analog input. It is clear from Fig. 12 that the digitized data are closer to the analog signal when the cells are smaller. The width of the cells is defined by the sampling rate, that is, the number of measurements the ADC can convert per unit of time. The height of the cells is defined by the resolution of the ADC. A 12-bit resolution means that the ADC sorts the amplitude of the analog values among one of $2^{12} = 4096$ possible values. The ADC has also a range, that is, the minimum and maximum analog amplitudes it can handle. For example, using a 12-bit analog-to-digital converter with a -10 to $+10$ V range (i.e., a 20 V range), the height of each cell is 20/4096 (i.e., ~ 0.005 V). Even if the analog signal is recorded with a 0.001 V accuracy, after ADC the digital signal accuracy will be at best 0.005 V. The point here is that any ADC is characterized by its sampling rate, its resolution and its range. Both the SPECT and PET systems measure the location and the energy of the photons hitting their detectors. The measurements are initially analog, and are digitized as described above before being stored on a computer. In SPECT, prior to the beginning of the acquisition, the operator chooses the width and height (in pixels) of the digital image to be acquired. Common sizes are 64 by 64 pixels (noted 64×64, width by height), 128×128 or 256×256. Dividing the size of the field of view (in cm) by the number of pixels yields the pixel size (in cm). For example, if the usable size for the detector is 40×40 cm, the pixel size is $40/64 = 0.66$ cm for a 64×64 image. Because all devices are imperfect, a point source is seen as a blurry spot on the image. If two radioactive point sources in the field of view are close enough, their spots overlap each other. When the two sources get closer, at some point the spots cannot be visually separated in the image. The smallest distance between the sources that allows us to see one spot for each source is called the spatial resolution. The pixel size is not to be confused with the spatial resolution. While the pixel size is chosen by the operator, the spatial resolution is imposed by the camera characteristics, and most notably by the collimator now that thick light guides are no longer employed. The pixel size is chosen to be smaller than the resolution, so that we can get as much detail as the resolution allows us to get, but it is important to understand that using a pixel size much smaller than the resolution does not increase the image quality. If the pixel size is small (i.e., when the number of pixels in the field of view is large), then the spot spills over many pixels, but with no improvement to image resolution.

The energy resolution (i.e., the smallest change in energy the detector can measure) is limited, and its value has a great impact on image quality, as explained below. Between the points of emission and detection, photons with an energy < 1 MeV frequently interact with electrons by scattering, during which their direction changes and some of their energy is lost. Because their direction changes, an error is made on their origin. However, it is possible to know that a photon is scattered because it has lost some energy, so an energy window, called the photopeak window defined around the energy of nonscattered (primary) photons (the photopeak), is defined prior to the acquisition, and the scattered photons whose energy falls outside the

photopeak window can be identified by energy discrimination and ignored. Unfortunately, photons in the photopeak window can either be scattered photons, or a primary photon whose energy has been underestimated (due to the limited energy resolution of the detectors, an error can be made regarding the actual energy of the photons). If the photopeak window is wide, many scattered photons are accepted, and the image has a lot of scattered activity that reduces the contrast; if the energy window is narrow, many primary photons are rejected, and the image quality is poor because of a lack of signal. As the energy resolution increases, the energy window can be narrowed, so that most scattered photons can be rejected while most primary photons are kept.

As mentioned above, the 3D distribution of the tracer in the field of view is projected onto the 2D plane of the camera heads. As opposed to the list-mode format (presented later in this article), the projection format refers to the process of keeping track of the total number of photons detected (the events, or counts) for each pixel of the projection image. Each time a count is detected for a given pixel, a value of 1 is added to the current number of counts for that pixel. In that sense, a projection represents the accumulation of the counts on the detector for a given period of time. If no event is recorded for any given pixel, which is not uncommon especially in the most peripheral parts of the image, then the value for that pixel is 0. Usually, 16 bits (2 bytes) are allocated to represent the number of counts per pixel, so the range for the number of events is 0 to $2^{16}-1 = 65{,}535$ counts per pixel. In case the maximal value is reached for a pixel (e.g., for a highly active source and a long acquisition time), then the computer possibly stops incrementing the counter, or reinitializes the pixel value to 0 and restarts counting from that point on. This yields images in which the most radioactive areas in the image may paradoxically have a lower number of counts than surrounding, less active, areas.

Different acquisitions are possible with a gamma camera:

Planar (or static): The gamma-camera head is stationary. One projection is obtained by recording the location of the events during a given period from a single angle of view. This is equivalent to taking a photograph with a camera. The image is usually acquired when the tracer uptake in the organ of interest has reached a stable level. One is interested in finding the quantity of radiopharmaceutical that accumulated in the region of interest. Planar images are usually adequate for thin or small structures (relative to the resolution of the images), such as the bones, the kidneys, or the thyroid.

Whole body: This acquisition is similar to the planar acquisition, in the sense that one projection is obtained per detector head, but is designed, as the name implies, to obtain an image of the whole body. Since the human body is taller than the size of the detector ($\sim 40 \times 40$ cm), the detector slowly moves from head to toes. This exam is especially indicated when looking for metastases. When a cancer starts developing at a primary location, it may happen that cancer cells, called metastases, disseminate in the whole body, and end up in various locations, especially bones. There, they may start proliferating and a new cancer may be initiated at that location. A whole-body scintigraphy is extremely useful when the physician wants to know whether one or more secondary tumors start developing, without knowing exactly where to look at.

Dynamic: Many projections are successively taken, and each of them is typically acquired over a short period (a few seconds). This is equivalent to recording of a movie. Analyzing the variations as a function of time allows us to compute parameters, such as the uptake rate, which can be a useful clinical index of normality.

Gated: The typical application of a gated acquisition is the cardiac scintigraphy. Electrodes are placed on the patient's chest to record the electrocardiogram (ECG or EKG). The acquisition starts at the beginning of the cardiac cycle, and a dynamic sequence of 8 or 16 images is acquired over the cardiac cycle ($\sim$ 1 s), so that a movie of the beating heart is obtained. However, the image quality is very poor when the acquisition is so brief. So, the acquisition process is repeated many times (i.e., over many heart beats), and the first image of all cardiac cycles are summed together, the second image of all cardiac cycles are summed together, and so on.

Tomographic: The detector heads are rotating around the patient. Projections are obtained under multiple angles of view. Through a process called tomographic reconstruction (presented in the next section), the set of 2D projections is used to find the 3D distribution of the tracer in the body, as a stack of 2D slices. The set of 1D projections of one slice for all projection angles is called a sinogram.

Tomographic gated: As the name implies, this acquisition is a tomographic one with gating information. The ECG is recorded during the tomographic acquisition, and for each angle of view, projections are acquired over many cardiac cycles, just as with a gated acquisition (see above). Thus, a set of projections is obtained for each point of the cardiac cycle. Each set is reconstructed, and tomographic images are obtained for each point of the cardiac cycle.

In contrast with the types of acquisition above in which the data are accumulated in the projection matrix for several seconds or minutes (frame-mode acquisition), a much less frequent type of acquisition called list-mode acquisition, can also be useful, because more information is available in this mode. As the name implies, the information for each individual event is listed in the list-mode file, and are not arranged in a matrix array. In addition to the coordinates of the scintillations, additional data are stored in the file. Figure 13 illustrates the typical list-mode format for a SPECT system. List-mode information is similar with a PET system, except that the heads location and X–Y coordinates are replaced with the location of the event on the detector rings. The list-mode file ($\sim$ 50 Mb

Gantry Angle 123 degrees
Timestamp: 0 ms
X:1002, Y:1270, Energy:1640
X:1044, Y:1211, Energy:1767
X:1077, Y:741, Energy:1788
X:570, Y:819, Energy:1674
Timestamp: 10 ms
X:1280, Y:1595, Energy:1603
X:576, Y:1181, Energy:1768
Timestamp: 20 ms
X:919, Y:1162, Energy:1828
X:973, Y:1078, Energy:1765
X:1023, Y:1045, Energy:1708
Timestamp: 30 ms
X:955, Y:773, Energy:1717
X:989, Y:702, Energy:1732
X:1060, Y:1145, Energy:1853
...
Timestamp: 19990 ms
X:577, Y:862, Energy:1818
X:556, Y:766, Energy:1682
Gantry Angle 126 degrees
Timestamp: 0 ms
...

Figure 13. Example of data stored on-the-fly in a list-mode file data in a SPECT system. Gantry angle defines the location of the detectors. The X and Y coordinates are given in a 2048×2048 matrix. The energy is a 12-bit value (i.e., between 0 and 4,095), and a calibration is required to convert these values in usual units (i.e., kiloelectron volt, keV).

in size in SPECT) can be quite large relative to a projection file. The list-mode format is far less common than the projection format, because it contains information, such as timing, that would usually not be used for a routine clinical exam. The list-mode data can be transformed into projection data through a process called rebinning (Fig. 14). Since the timing is known, multiple projections can be created as a function of time, thus allowing the creation of "movies" whose rate can be defined postacquisition. A renewed interest in the list-mode format has been fueled these past years by the temporal information it contains, which is adequate for the temporal correlation of the acquisition with patient, cardiac, or respiratory motions through the synchronized acquisition of signal from motion detectors.

TOMOGRAPHIC RECONSTRUCTION

Tomographic reconstruction has played a central role in NM, and has heavily relied on computers (6). In addition to data acquisition control, tomographic reconstruction is the other main reason for which computers have been early introduced in NM. Among all uses of computers in NM, tomographic reconstruction is probably the one that symbolizes most the crunching power of computers. Tomographic reconstruction is the process by which slices of the 3D distribution of tracers are obtained based upon the projections obtained under different angles of view. Because the radioactivity emitted in the 3D space is projected on the 2D detectors, the contrast is usually low. Tomographic reconstruction greatly restores the contrast, by estimating the 3D tracer distribution. Reconstruction is possible using list-mode data (SPECT or PET), but mainly

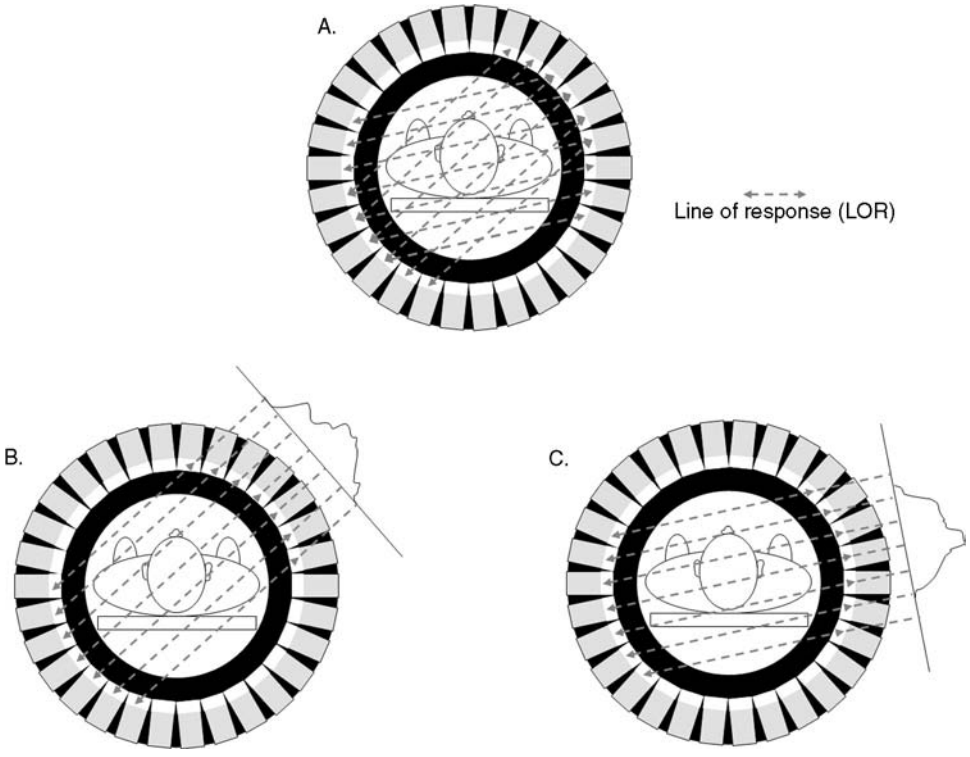


Figure 14. Rebinning process. (a) The line joining pairs of photons detected in coincidence is called a line of response (LOR). (b and c) LOR are sorted so that parallel LOR are grouped, for a given angle.

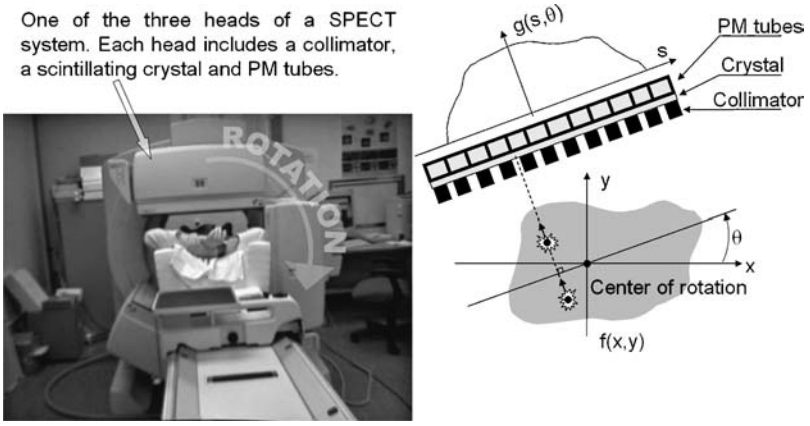


Figure 15. The SPECT acquisition. Left. A three-head IRIX SPECT system (Philips Medical Systems). A subject is in the field of view while the camera heads are slowly rotating around him. Right. Physical model and geometric considerations. The 2D distribution of the radioactivity $f(x,y)$ in one slice of the body is projected and accumulated onto the corresponding 1D line $g(s,\theta)$ of detector bins.

for research purposes. The PET data, although initially acquired in list-mode format, are usually reformatted to form projections, so that the algorithms developped in SPECT can also be used with PET data. There are many different algorithms, mainly the filtered back-projection (FBP) and the iterative algorithms, that shall be summarized below (7–9).

In the following, focuses on tomographic reconstruction when input data are projections, which is almost always the case on SPECT systems. During a SPECT acquisition, the detecting heads rotate around the subject to gather projections from different angles of views (Fig. 15). Figure 16 presents the model used to express the simplified projection process in mathematical terms. Associated with the projection is the backprojection (Fig. 17). With backprojection, the activity in each detector bin g is added to all the voxels which project onto bin g. It can be shown (10) that backprojecting projections filtered with a special filter called a ramp filter (filtered backprojection, or FBP) is a way to reconstruct slices. However, the ramp filter is known to increase the high frequency noise, so it is usually combined with a low pass filter (e.g., Butterworth filter) to form a band-pass filter. Alternatively, reconstruction can be performed with the ramp filter only, and then the reconstructed images can be smoothed with a 3D low pass filter. The FBP technique yields surprisingly good results considering the simplicity of the method and its approximations, and is still widely used today. However, this rather crude approach is more and more frequently replaced by the more sophisticated iterative algorithms, in which many corrections can be easily introduced to yield more accurate results. An example of an iterative reconstruction algorithm includes the following steps:

1. An initial estimate of the reconstructed is created, by attributing to all voxels the same arbitrary value (e.g. 0 or 1).
2. The projections of this initial estimate are computed.
3. The estimated projections are compared to the measured projections, either by computing their difference or their ratio.
4. The difference (resp. the ratio) is added (resp. multiplied) to the initial estimate to get a new estimate.
5. Steps 2–4 are repeated until the projections of the current estimate are close to the measured projections.

Figure 18 illustrates a simplified version of the multiplicative version of the algorithm. This example has been voluntarily oversimplified for sake of clarity. Indeed, image reconstruction in the real world is much more complex for several reasons: (*1*) images are much larger; typically, the 3D volume is made of 128 × 128 × 128 voxels, (*2*) geometric considerations are included to take into account the volume of each volume element (voxel) that effectively project onto each bin at each angle of view, (*3*) camera characteristics, and in particular the spatial resolution, mainly defined by the collimator characteristics, are introduced in the algorithm, and (*4*) corrections presented below are applied during the iterative process. The huge number of operations made iterative reconstruction a slow process and prevented its routine use until recently, and FBP was

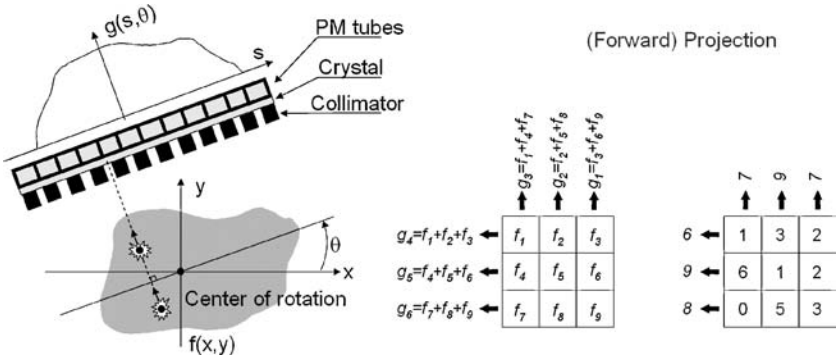


Figure 16. Projection. Each plane in the FOV (left) is seen as a set of values f (center). The collimator is the device that defines the geometry of the projection. The values in the projections are the sum of the values in the slices. An example is presented (right). (Reproduced from Ref. 8 with modifications with permission of the Society of Nuclear Medicine Inc.)

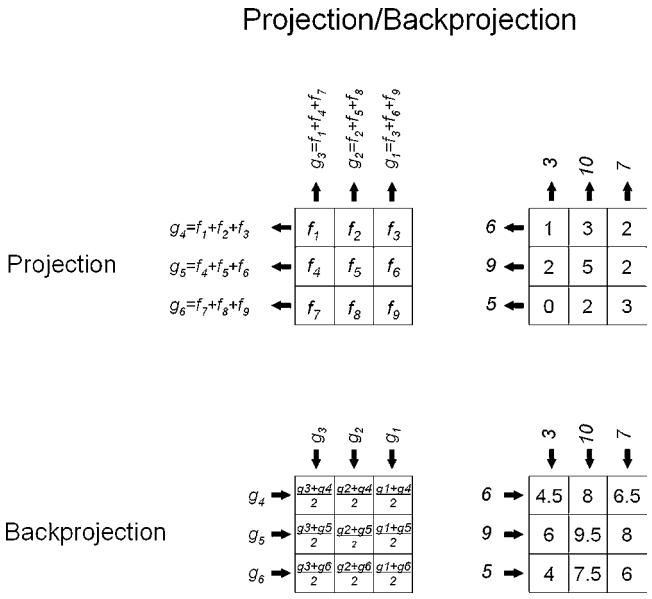


Figure 17. Projection and backprojection. Notice that backprojection is not the invert of projection. (Reproduced from Ref. 8 with modifications with permission of the Society of Nuclear Medicine Inc.)

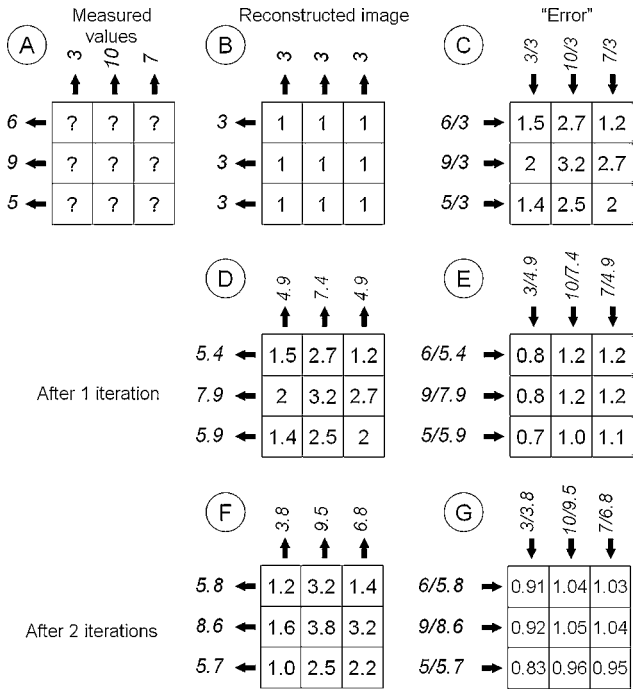


Figure 18. A simplified illustration of tomographic reconstruction with an iterative algorithm. (a) The goal is to find the values in a slice (question marks) given the measured projection values 7, 10, 3, 6, 9, 5. (b) Voxels in the initial estimate have a value of 1, and projections are computed as described in Fig. 17. (c) The error in the projections is estimated by dividing the actual values by the estimated values. The ratios are then backprojected to get a slice of the "error". (d) Multiplying the error obtained in c by the estimate in b yields a second estimate, and projections are computed again (g). After an arbitrary number of iterations (e, f), an image whose projections are close to the measured projections is obtained. This image is the result of the iterative tomographic reconstruction process. Such a process is repeated for the stack of 2D slices.

preferred. Modern computers are now fast enough for iterative algorithms, and since these algorithms have many advantages over the FBP, they are more and more widely used.

A number of corrections usually need to be applied to the data during the iterative reconstruction to correct them for various well-known errors caused by processes associated with the physics of the detection, among which the more important are attenuation (11–13), Compton scattering (11,12), depth-dependent resolution (in SPECT) (11,12), random coincidences (in PET) (14), and partial volume effect (15). These sources of error below are briefly presented:

Attenuation occurs when photons are stopped (mostly in the body), and increases with the thickness and the density of the medium. Thus, the inner parts of the body usually appear less active than the more superficial parts (except the lungs, whose low density makes them almost transparent to gamma photons and appear more active than the surrounding soft tissues). Attenuation can be compensated by multiplying the activity in each voxel by a factor whose value depends upon the length and the density of the tissues encountered along the photons path. The correction factor can be estimated (e.g., by assuming a uniform attenuation map) or measured using an external radioactive source irradiating the subject. A third possibility, which is especially attractive with the advent of SPECT/CT and PET/CT systems (presented below), is to use the CT images to estimate the attenuation maps.

Photons may be scattered when passing through soft tissues and bones, and scattered photons are deflected from their original path. Because of the error in the estimated origin of the scattered photons, images are slightly blurred and contrast decreases. As mentioned in the previous section, the effects of scattering can be better limited by using detectors with a high energy resolution. Scatter can also be estimated by acquiring projection data in several energy windows during the same acquisition. Prior to the acquisition, the user defines usually two or three windows per photopeak (the photopeak window plus two adjacent windows, called the scatter windows). As mentioned, photons can be sorted based upon their energy, so they can be assigned to one of the windows. The amount of scattering is estimated in the photopeak window using projection data acquired in the scatter windows, and assuming a known relationship between the amount of scattering and the energy. Another approach to Compton scattering compensation uses the reconstructed radioactive distribution and attenuation maps to determine the amount of scatter using the principles of scattering interactions.

In SPECT, collimators introduce a blur (i.e., even an infinitely small radioactive source would be seen as a spot of several mm in diameter) for geometrical reasons. In addition, for parallel collimators (the most commonly used, in which the holes are parallel), the blur increases as the distance between the source and the collimator increases. Depth-dependent resolution can be corrected either by fil-

tering the sinogram in the Fourier domain using a filter whose characteristics vary as a function of the distance to the collimator (frequency–distance relationship, FDR) or by modelling the blur in iterative reconstruction algorithms.

In PET, a coincidence is defined as the detection of two photons (by different detectors) in a narrow temporal window of $\sim$ 10 ns. As mentioned, a coincidence is true when the two photons are of the same pair, and random when the photons are from two different annihilations. The amount of random coincidences can be estimated by defining a delayed time window, such that no true coincidence can be detected. The estimation of the random coincidences can then be subtracted from the data including both true and random, to extract the true coincidences.

Partial volume effect (PVE) is directly related to the finite spatial resolution of the images: structures that are small (about the voxel size and smaller) see their concentration underestimated (the tracer in the structure appears as being diluted in the voxel). Spillover is observed at the edges of active structures: some activity spreads outside the voxels, so that although it stems from the structure, it is actually detected in neighboring voxels. Although several techniques exist, the most accurate can be implemented when the anatomical boundaries of the structures are known. Thus, as presented below, anatomical images from CT scanners are especially useful for PVE and spillover corrections, if they can be correctly registered with the SPECT or PET data.

IMAGE PROCESSING, ANALYSIS AND DISPLAY

Computers are essential in NM not only for their ability to control the gamma cameras and to acquire images, but also because of their extreme ability to process, analyze and display the data. Computers are essential in this respect because (*1*) the amount of data can be large (millions of pixel values), and computers are extremely well suited to handle images in their multimegabytes memory, (*2*) repetitive tasks are often needed and central processor units (CPUs) and array processors can repeat tasks quickly, (*3*) efficient algorithms have been implemented as computer programs to carry on complex mathematical processing, and (*4*) computer monitors are extremely convenient to display images in a flexible way.

Both the PET and SPECT computers come with a dedicated, user-friendly graphical environment, for acquisition control, patient database management, and a set of programs for tomographic reconstruction, filtering and image manipulation. These programs are usually written in the C language or in Fortran, and compiled (i.e., translated in a binary form a CPU can understand) for a given processor and a given operating system (OS), usually the Unix OS (The Open Group, San Francisco CA) or the Windows OS (Microsoft Corporation, Redmond WA). As an alternative to these machine-dependent programs, Java (Sun Microsystems Inc.) based programs have been proposed (see next section).

As seen in the first section, an image can be seen as a rectangular array of values, which is called, from a mathematical point of view, a matrix. A large part of image processing in NM is thus based upon linear algebra (16), which is the branch of mathematics that deals with matrices. One of the problems encountered in NM imaging is the noise (random variations due to the probabilistic nature of the radioactive processes and to the limited accuracy of the measurements). A number of methods are available to reduce the noise after the acquisition, by smoothing the minor irregularities or speckles in the images (10). The most common way to filter images is by convolution (a pixel value is replaced by a weighted average of its neighbors) or by Fourier methods. Computers are extremely efficient at computing discrete Fourier transforms thanks to a famous algorithm called the fast Fourier transform (FFT) developed by Cooley and Tukey (17).

The NM images can be displayed or printed in black and white (gray levels) or in color. Pixel values can be visually estimated based on the level of gray or based on the color. Color has no special meaning in NM images, and there is no consensus about the best color map to use. Most often, a pixel value represents a number of counts, or events. However, units can be something else (e.g., flow units), especially after some image processing. So, for proper interpretation, color map and units should always accompany an image.

Regions of interest (ROIs) are defined by line segments drawn to set limits in images and can have any shape or be drawn by hand. The computer is then able to determine which pixels of the image are out and which are in the ROI, and thus computations can be restricted to the inside or to the outside of the ROI. ROIs are usually drawn with the mouse, based on the visual inspection of the image. Drawing a ROI is often a tricky task, due to the low resolution of the images and to the lack of anatomical information regarding the edges of the organs. In an attempt to speed up the process, and to reduce the variability among users, ROIs can also be drawn automatically (18). When a dynamic acquisition is available, a ROI can be drawn on one image and reported on the other images of the series, the counts in the ROI are summed and displayed as a time–activity curve (TAC), so that one gets an idea of the kinetics of the tracer in the ROI. The TAC are useful because with the appropriate model, physiological parameters such as pharmacological constants or blood flow can be determined based on the shape of the curve. An example of dynamic studies with ROI and TAC is the renal scintigraphy, whose goal is to investigate the renal function through the vascularization of the kidneys, their ability to filter the blood, and the excretion in the urine. A tracer is administered with the patient lying on the bed of the camera, and a two-stage dynamic acquisition is initiated: many brief images are acquired (e.g., 1 image per second over the first 60 s). Then, the dynamic images are acquired at a lower rate (e.g., 1 image per minute for 30 min). After the acquisition, ROIs are drawn over the aorta, the cortical part of the kidneys, and the background (around the kidneys), and the corresponding TAC are generated for analysis (Fig. 19). The TAC obtained during the first stage of the acquisition reflect the arrival of the tracer in the renal arteries (vascular phase). The rate at which the tracer invades the renal vascularization, relative to the speed at which it arrives in the aorta above indicates whether the kidneys

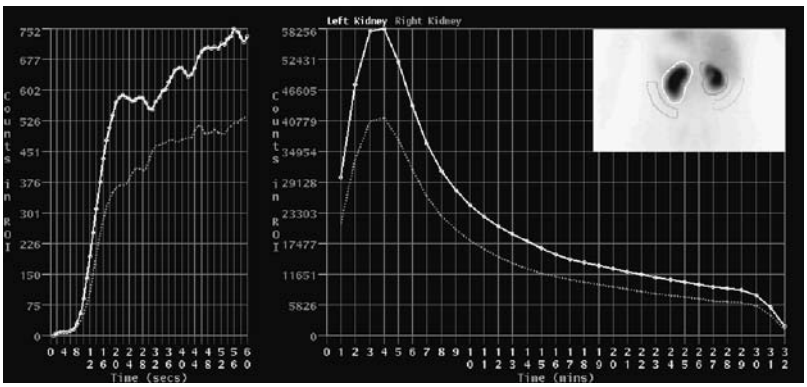


Figure 19. Output of a typical renal scintigraphy. Left: the TAC for both kidneys in the first minute after injection of the tracer. The slope of the TAC gives an indication of the state of the renal vascularization. Center: One-minute images acquired over 32 min after injection. The ascending part evidences the active tracer uptake by the kidneys, while the descending part shows the excretion rate. Right: Insert shows the accumulation of the tracer in the bean-shaped kidneys. The ROI are drawn over the kidneys and the background.

are normally vascularized. The renal TAC obtained in the second stage of the acquisition (filtration phase) show the amount of tracer captured by the kidneys, so that the role of the kidneys as filters can be assessed. When the tracer is no longer delivered to the kidneys, and as it passes down the ureters (excretion phase), the latest part of the renal curves normally displays a declining phase, and the excretion rate can be estimated by computing the time of half-excretion (i.e., the time it takes for the activity to decreases from its peak to half the peak), usually assuming the decrease is an exponential function of time.

The corrections presented at the end of the previous section are required to obtain tomographic images in which pixel values (in counts per pixel) are proportional to the tracer concentration with the same proportion factor (relative quantitation), so that different areas in the same volume can be compared. When the calibration of the SPECT or PET system is available (e.g., after using a phantom whose radioactive concentration is known), the images can be expressed in terms of activity per volume unit, (e.g., in becquerels per milliliters, $Bq \cdot mL^{-1}$; absolute quantitation). Absolute quantitation is required in the estimation of a widely used parameter, the standardized uptake value (SUV) (19), which is an index of the FDG uptake that takes into account the amount of injected activity and the dilution of the tracer in the body. The SUV in the region of interest is computed as SUV = (uptake in the ROI in $Bq \cdot mL^{-1}$)/(injected activity in Bq/body volume in mL). Another example of quantitation is the determination of the blood flow (in $mL \cdot g^{-1} \cdot min^{-1}$), based upon the pixel values and an appropriate model for the tracer kinetics in the area of interest. For example, the absolute regional cerebral blood flow (rCBF) is of interest in a number of neurological pathologies (e.g., ischemia, haemorrage, degenerative diseases, epilepsy). It can be determined with xenon ^{133}Xe, a gas that has the interesting property of being washed out from the brain after its inhalation as a simple exponential function of the rCBF. Thus, the rCBF can be assessed after at least two fast tomographic acquisitions (evidencing the differential decrease in activity in the various parts of the brain), for example, using the Tomomatic 64 SPECT system (Medimatic, Copenhagen, Denmark) (20).

An example of image processing in NM is the equilibrium radionuclide angiography (ERNA) (21), also called multiple gated acquisition (MUGA) scan, for assessment of the left ventricle ejection fraction (LVEF) of the heart. After a blood sample is taken, the erythrocytes are labelled with $^{99\,m}Tc$ and injected to the patient. Because the technetium is retained in the erythrocytes, the blood pool can be visualized in the images. After a planar cardiac gated acquisition, 8 or 16 images of the blood in the cardiac cavities (especially in the left ventricle) are obtained during an average cardiac cycle. A ROI is drawn, manually or automatically, over the left ventricle, in the end-diastolic (ED) and end-systolic (ES) frames, that is, at maximum contraction and at maximum dilatation of the left ventricle respectively. Another ROI is also drawn outside the heart for background activity subtraction. The number of counts nED and nES in ED and ES images, respectively, allows the calculation of the LVEF as LVEF = (nED-nES)/nED. Acquisitions for the LVEF assessment can also be tomographic in order to improve the delineation of the ROI over the left ventricle, and several commercial softwares are available (22) for largely automated processing, among which the most widely used are the Quantitative Gated SPECT (QGS) from the Cedars-Sinai Medical Center, Los Angeles, and the Emory Cardiac Tool box (ECTb) from the Emory University Hospital, Atlanta.

INFORMATION TECHNOLOGY

An image file typically contains, in addition to the image data, information to identify the images (patient name, hospital patient identification, exam date, exam type, etc.), and to know how to read the image data (e.g., the number of bytes used to store one pixel value). File format refers to the way information is stored in computer files. A file format can be seen as a template that tell the computer how and where in the file data are stored. Originally, each gamma-camera manufacturer had its own file format, called proprietary format, and for some manufacturers the proprietary format was confidential and not meant to be widely disclosed. To facilitate the exchange of images between different computers, the Interfile format (23) was proposed in the late 1980s. Specifically designed for NM images, it was intended to be a common file format that anyone could understand and use to share files. At the same period, the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA) developed their standard for NM, radiology, MRI and ultrasound images: the ACR-NEMA file format, version 1.0 (in 1985)

and 2.0 (in 1988). In the early 1990s, local area networks (LANs) connecting NM, radiology and MRI departments started to be installed. Because Interfile was not designed to deal with modalities other than NM, and because ACR-NEMA 2.0 was "only" a file format, and was not able to handle robust network communications to exchange images over a LAN, both became obsolete and a new standard was developed by the ACR and the NEMA, ACR-NEMA 3.0, known as Digital Imaging and Communications in Medicine (DICOM) (24). Although quite complex (the documentation requires literally thousands of pages), DICOM is powerful, general in its scope and designed to be used by virtually any profession using digital medical images. Freely available on the Internet, DICOM has become a standard among the manufacturers, and it is to be noted that DICOM is more than a file format. It also includes methods (programs) for storing and exchanging image information; in particular, DICOM servers are programs designed to process requests for handling DICOM images over a LAN.

DICOM is now an essential part of what is known as Picture Archiving and Communications Systems (PACS). Many modern hospitals use a PACS to manage the images and to integrate them into the hospital information system (HIS). The role of the PACS is to identify, store, protect (from unauthorized access) and retrieve digital images and ancillary information. A web server can be used as an interface between the client and the PACS (25). Access to the images does not necessarily require a dedicated software on the client. A simple connection to the Internet and a web browser can be sufficient, so that the images can be seen from the interpreting or prescribing physician's office. In that case, the web server is responsible for submitting the user's request to the PACS, and for sending the image data provided by the PACS to the client, if the proper authorization is granted. However, in practice, the integration of NM in a DICOM-based PACS is difficult, mainly because PACS evolved for CT and MR images (26,27), that is, as mostly static, 2D, black and white images. The NM is much richer from this point of view, with different kinds of format (list-mode, projections, whole-body, dynamic, gated, tomographic, tomographic gated, etc.) and specific postacquisition processing techniques and dynamic displays. The information regarding the colormaps can also be a problem for a PACS when dealing with PET or SPECT images fused with CT images (see next section) because two different colormaps are used (one color, one grayscale) with different degrees of image blending.

In the spirit of the free availability of programs symbolized by the Linux operating system, programs have been developed for NM image processing and reconstruction as plug-ins to the freely available ImageJ program developed at the U.S. National Institutes of Health (28). ImageJ is a general purpose program, written with Java, for image display and processing. Dedicated Java modules (plugs-in) can be developed by anyone and added as needed to perform specific tasks, and a number of them are available for ImageJ (29). Java is a platform-independent language, so that the same version of a Java program can run on different computers, provided that another program, the Java virtual machine (JVM), which is platform-dependent, has been installed beforehand. In the real world, however, different versions of the JVM may cause the Java programs to crash or to cause instabilities when the programs require capabilities the JVM cannot provide (30). The advantage of the Java programs is that they can be used inside most Internet browsers, so that the user has no program to install (except the JVM). A Java-based program called JaRVis (standing for Java-based remote viewing station) has been proposed in that spirit for viewing and reporting of nuclear medicine images (31).

It is very interesting to observe how, as the time goes by, higher levels of integration have been reached: with the early scintigraphy systems, such as rectilinear scanners (in the 1970s), images were analog, and the outputs were film or paper hard copies. In the 1980s images were largely digital, but computers were mainly stand alone machines. One decade later, computers were commonly interconnected through LANs, and standard formats were available, permitting digital image exchange and image fusion (see next section). Since the mid-1990s, PACS and the worldwide web make images remotely available, thus allowing telemedecine.

HYBRID SPECT/CT AND PET/CT SYSTEMS

Multimodality imaging (SPECT/CT and PET/CT) combines the excellent anatomical resolution of CT with SPECT or PET functional information (2,3). Other advantages of multimodality are (*1*) the use of CT images to estimate attenuation and to correct for PVE in emission images, (*2*) the potential improvement of emission data reconstruction by inserting in the iterative reconstruction program prior information regarding the locations of organ and/or tumor boundaries, and (*3*) the possible comparison of both sets of images for diagnostic purposes, if the CT images are of diagnostic quality. The idea of combining information provided by two imaging modalities is not new, and a lot of work has been devoted to multimodality. Multimodality initially required that the data acquired from the same patient, but on different systems and on different occasions, be grouped on the same computer, usually using tapes to physically transfer the data. This was, $\sim$ 20 years ago, a slow and tedious process. The development of hospital computer network in the 1990s greatly facilitated the transfer of data, and the problem of proprietary image formats to be decoded was eased when a common format (DICOM) began to spread. However, since the exams were still carried out in different times and locations, the data needed to be registered. Registration can be difficult, especially because emission data sometimes contain very little or no anatomical landmarks, and external fiducial markers were often needed. Given the huge potential of dual-modality systems, especially in oncology, a great amount of energy has been devoted in the past few years to make it available in clinical routine. Today, several manufacturers propose combined PET/CT and SPECT/CT hybrid systems (Fig. 20): The scanners are in the same room, and the table on which the patient lies can slide from one scanner to the other (Fig. 21). An illustration of PET/CT images is presented in Fig. 22.

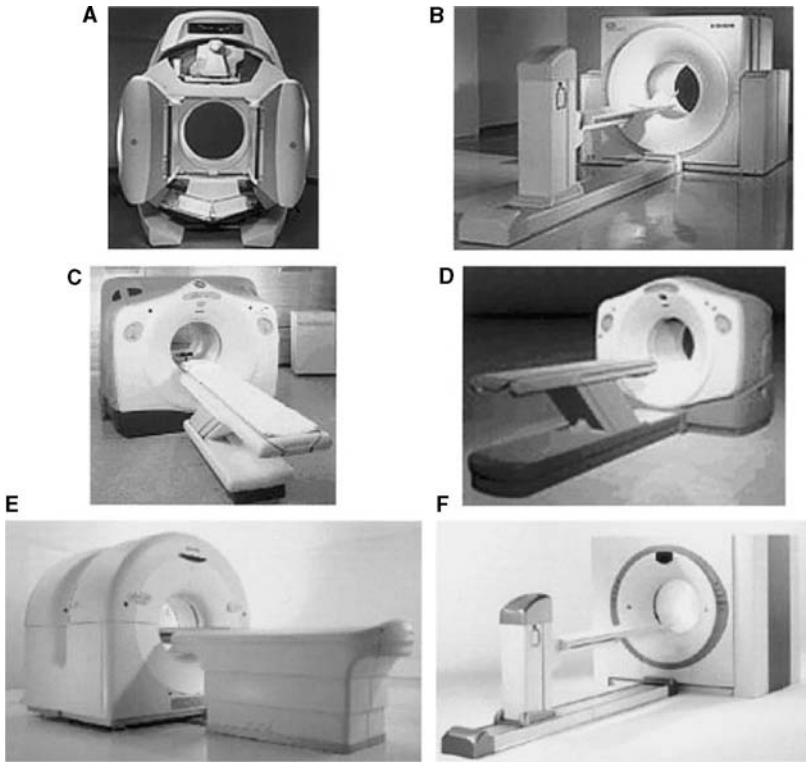


Figure 20. Current commercial PET/CT scanners from 4 major vendors of PET imaging equipment: (a) Hawkeye (GE Medical Systems); (b) Biograph (Siemens Medical Solutions) or Reveal (CTI, Inc); (c) Discovery LS (GE Medical Systems); (d) Discovery ST (GE Medical Systems); (e) Gemini (Philips Medical Systems); (f) Biograph Sensation 16 (Siemens Medical Solutions) or Reveal XVI (CTI, Inc.). (Reproduced from Ref. 2, with permission of the Society of Nuclear Medicine Inc.)

Although patient motion is minimized with hybrid systems, images from both modalities are acquired sequentially, and the patient may move between the acquisitions, so that some sort of registration may be required before PET images can be overlaid over CT images. Again, computer programs play an essential role in finding the best correction to apply to one dataset so that it matches the other dataset. Registration may be not too difficult with relatively rigid structures, such as the brain, but tricky for chest imaging for which nonrigid transformations are needed. Also, respiratory motion introduces in CT images mushroom-like artifacts that can be limited by asking the patients to hold their breath at mid-respiratory cycle during CT acquisition, so that it best matches the average images obtained in emission tomography with no respiratory gating.

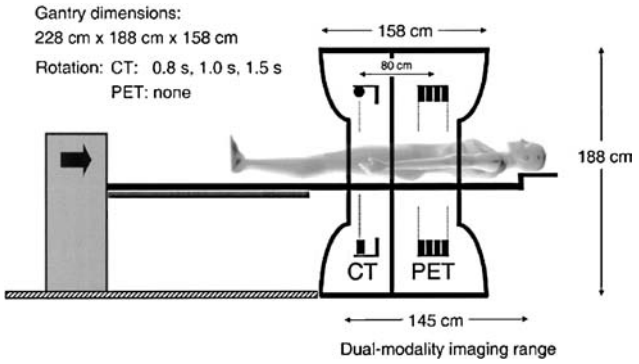


Figure 21. Schematic of PET/CT developped by CPS Innovations. Axial separation of two imaging fields is 80 cm. The coscan range for acquiring both PET and CT has maximum of 145 cm. (Reproduced from Ref. 2, with permission of the Society of Nuclear Medicine Inc.)

Dedicated programs are required for multimodality image display (*1*) to match the images (resolution, size, orientation); (*2*) to display superimposed images from both modalities with different color maps (CT data are

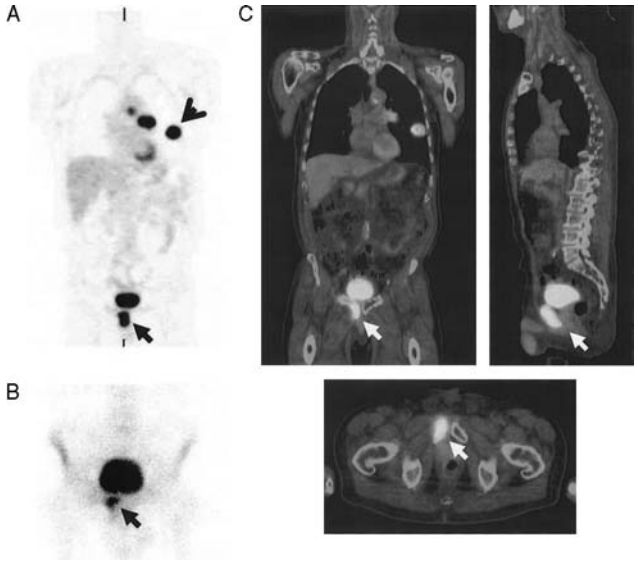


Figure 22. Image of a 66 years old male patient with history of head-and-neck cancer. In addition to ^{18}FDG uptake in lung malignancy, intense uptake is seen on PET scan (a) in midline, anterior and inferior to bladder. Note also presence of lung lesion (arrowhead) due to primary lung cancer. ^{99m}Tc bone scan (b) subsequently confirmed that uptake was due to metastatic bone disease. PET/CT scan (c) directly localized uptake to pubic ramus (arrowed). (Reproduced from Ref. 2, with permission of the Society of Nuclear Medicine Inc.)

typically displayed with a gray scale, while a color map is used to display the tracer uptake); (*3*) to adjust the degree of transparency of each modality relative to the other in the overlaid images; and (*4*) to select the intensity scale that defines the visibility of the bones, soft tissues and lungs in the CT images. For the interpretation of SPECT/CT or PET/CT data, the visualization program has to be optimized, for so much information is available (dozens of slices for each modality, plus the overlaid images, each of them in three perpendicular planes) and several settings (slice selection, shades of gray, color map, the degree of blending of the two modalities in the superimposed images) are to be set. Powerful computers are required to be able to handle all the data in the computer random access memory (RAM) and to display them in real time, especially when the CT images are used at their full quality (512×512 pixel per slice, 16-bit shades of gray). Finally, these new systems significantly increase the amount of data to be archived (one hundred to several hundreds megabytes per study), and some trade-off may have to be found between storing all the information available for later use and minimizing the storage space required. An excellent review of the current software techniques for merging anatomic and functional information is available (32).

COMPUTER SIMULATION

Simulation is a very important application of computers in NM research, as it is in many technical fields today. The advantage of simulation is that a radioactive source and a SPECT or PET system are not required to get images similar to the ones obtained if there were real sources and systems. Simulation is cheaper, faster and more efficient to evaluate acquisition hardware or software before they are manufactured and to change its design for optimization. Thus, one of the applications is the prediction of the performance of a SPECT or PET system by computer simulation. Another application is to test programs on simulated images that have been created in perfectly known conditions, so that the ouput of the programs can be predicted and compared to the actual results to search for possible errors.

Computer simulation is the art of creating information about data or processes without real measurement devices, and the basic idea is the following: if enough information is provided to the computer regarding the object of study (e.g., the 3D distribution of radioactivity in the body, the attenuation), the imaging system (the characteristics of the collimator, the crystal, etc.), and the knowledge we have about the interactions of gamma photons with matter, then it is possible to generate the corresponding projection data (Fig. 23). Although this may seem at first very complex, it is tractable when an acquisition can be modeled with a reasonable accuracy using a limited number of relevant mathematical equations. For example, a radioactive source can be modeled as a material emitting particles in all directions, at a certain decay rate. Once the characteristics of the source: activity, decay scheme, half-life, and spatial distribution of the isotope are given, then basically everything needed for simulation purposes

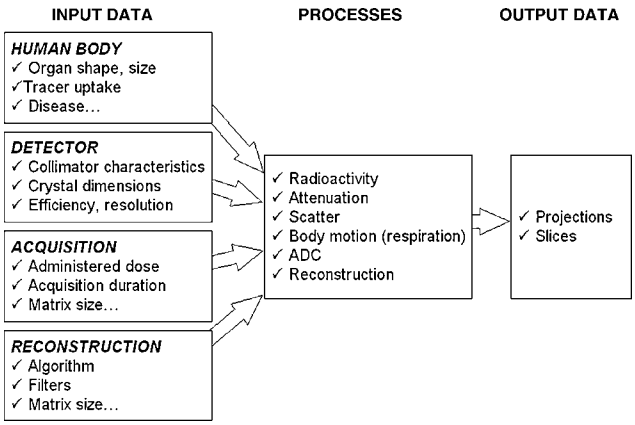


Figure 23. Principle of computer simulation in NM. Parameters and all available information are defined and used by the simulation processes (i.e., computer programs), which in turn generate output data that constitute the result of the simulation.

is known. Radioactive disintegrations and interactions between gamma photons and matter are random processes: one cannot do predictions about a specific photon (its initial direction, where it will be stopped, etc.), but the probabilities of any event can be computed. Random number generators are used to determine the fate of a given photon.

Let us say that we want to evaluate the resolution of a given SPECT system as a function of the distance from the radioactive source to the surface of the detector. The first solution is to do a real experiment: prepare a radioactive source, acquire images with the SPECT system, and analyze the images. This requires (*1*) a radioactive source, whose access is restricted, and (*2*) the SPECT system, which is costly. Because the resolution is mainly defined by the characteristics of the collimator, a second way to evaluate the resolution is by estimation (analytical approach): apply the mathematical formula that yield the resolution as a function of the collimator characteristics (diameter of the holes, collimator thickness, etc.). This approach may become tricky as more complex processes have to be taken into account. A third solution is to simulate the source, the gamma camera, and the physical interactions. It is an intermediate solution, between real acquisition and estimation. Simulation yields more realistic results than estimations, but does not require the use of real source or SPECT system. Simulation is also powerful because when uncertainties are introduced in the model (e.g., some noise), then it is possible to see their impact on the projection data.

Simulation refers either to the simulation of input data (e.g., simulation of the human body characteristics), or to the simulation of processes (e.g., the processes like the interaction between photons and matter). For simulation of input data, a very useful resource in nuclear cardiology is the program for the mathematical cardiac torso (MCAT) digital phantom developed at the University of North Carolina (33). The MCAT program models the shape, size, and location of the organs and structures of the human chest using mathematical functions called non-uniform rational B-splines (NURBS). The input of this program is a list of many parameters, among them: the amount of

radioactivity to be assigned to each organ (heart, liver, spleen, lungs, etc.), the attenuation coefficients of these organs, their size, the heart rate, and so on. This phantom used the data provided by the Visible Human Project and is accurate enough from an anatomical point of view for NM, and it can be easily customized and dynamic sets of slices can be obtained that simulate the effects of respiration and a beating heart. The output is the 3D distribution (as slices) of radioactivity (called emission data) and 3D maps of the attenuation (attenuation data) in the human torso. This output can then be used as an input for a Monte Carlo program to generate projection data. Monte Carlo programs (34,35) use computer programs called random number generators to generate data (for example, the projection data) based upon the probability assigned to any possible event, such as the scattering of a photon, or its annihilation in the collimator. The programs were named Monte Carlo after the city on the French Riviera, famous for its casinos and games based upon probabilities. Among the Monte Carlo simulation programs used in NM are: Geant (36), Simind (37), SimSET (38), and EGS4 (39). Programs, such as Geant and its graphical environment Gate (40), allow the definitions of both the input data (the body attenuation maps, the collimator characteristics) and the interactions to be modeled (photoelectric effect, scatter, attenuation, etc.).

EDUCATION AND TEACHING

As almost any other technical field, NM has benefited from the Internet as a prodigious way to share information. Clinical cases in NM are now available, and one advantage of computers over books is that an image on a computer can be manipulated: the color map can be changed (scale, window, etc.) and, in addition to images, movies (e.g., showing tracer uptake or 3D images) can be displayed. Websites presenting clinical NM images can be more easily updated with more patient cases and some on-line processing is also possible. One disadvantage is the sometimes transitory existence of the web pages, that makes (in the author's opinion) the Internet an unreliable source of information from this point of view. NM professionals can also share their experience and expertise on list servers (see Ref. 41 for a list of servers). The list servers are programs to which e-mails can be sent, and that distribute these e-mails to every registered person.

The Society of Nuclear Medicine (SNM) website hosts its Virtual Library (42), in which > 90 h of videos of presentations given during SNM meetings are available for a fee. The Joint Program in Nuclear Medicine is an example of a program including on-line education by presenting clinical cases (43) for > 10 years. More than 150 cases are included, and new cases are added; each case includes presentation, imaging technique, images, diagnosis, and discussion. Other clinical cases are also available on the Internet (44). Another impressive on-line resource is the Whole Brain Atlas (45) presenting PET images of the brain, coregistered with MRI images. For each case, a set of slices spanning over the brain is available, along with the presentation of the clinical case. The user can interactively select the transverse slices of interest on a sagittal slice. As the last example, a website (46) hosts a presentation of normal and pathologic FDG uptake in PET and PET/CT images.

CONCLUSION

Computers are used in NM for a surprisingly large variety of applications: data acquisition, display, processing, analysis, management, simulation, and teaching/training. As many other fields, NM has benefited these past years from the ever growing power of computers, and from the colossal development of computer networks. Iterative reconstruction algorithms, which have been known for a long time, have tremendously benefited from the increase of computers crunching power. Due to the increased speed of CPUs and to larger amounts of RAM and permanent storage, more and more accurate corrections (attenuation, scatter, patient motion) can be achieved during the reconstruction process in a reasonable amount of time. New computer applications are also being developed to deal with multimodality imaging such as SPECT/CT and PET/CT, and remote image viewing.

ACRONYMS

ACR	American College of Radiology
ADC	Analog to Digital Conversion (or Converter)
CPU	Central Processing Unit
CT	(X-ray) Computerized Tomography
DICOM	Digital Imaging and COmmunications in Medicine
ECT	Emission Computerized Tomography
FBP	Filtered BackProjection
FDG	Fluoro-Deoxy Glucose
FDR	Frequency-Distance Relationship
FFT	Fast Fourier Transform
HIS	Hospital Information System
keV	kiloelectron Volt
LAN	Local Area Network
LOR	Line of Response
MCAT	Mathematical Cardiac Torso
MRI	Magnetic Resonance Imaging
NEMA	National Electrical Manufacturers Association
NM	Nuclear Medicine
NURBS	Nonuniform Rational B-Splines
PACS	Picture Archiving and Communication System
PET	Positron Emission Tomography
PMT	Photomultiplier
PVE	Partial Volume Effect
RAM	Random Access Memory
rCBF	regional Cerebral Blood Flow
ROI	Region of Interest
SPECT	Single-Photon Emission Computerized Tomography
SNM	Society of Nuclear Medicine
SUV	Standardized Uptake Value
TAC	Time–Activity Curve

BIBLIOGRAPHY

Cited References

1. Wagner HN, Szabo Z, Buchanan JW, editors. Principles of Nuclear Medicine. 2nd ed. New York: W.B. Saunders; 1995.
2. Townsend DW, Carney JPJ, Yap JT, Hall NC. PET/CT today and tomorrow. J Nucl Med 2004;45:4S–14S.
3. Ratib O. PET/CT navigation and communication. J Nucl Med 2004;45:46S–55S.
4. Lee K. Computers in NM: a Practical Approach. New York: The Society of Nuclear Medicine Inc.; 1991.
5. Simmons GH. On-line corrections for factors that affect uniformity and linearity. J Nucl Med Tech 1988;2:82–89.
6. Rowland SW. Computer implementation of image reconstruction formulas. In: Herman GT, editor. Topics in Applied Physics: Image Reconstruction from Projections. vol. 32. Heidelberg, Germany: Springer-Verlag; 1979. pp. 29–79.
7. Zeng GL. Image reconstruction: a tutorial. Comput Med Imaging Graph 2001;25:97–103.
8. Bruyant PP. Analytical and iterative algorithms in SPECT. J Nucl Med 2002;43:1343–1358.
9. Brook RA, DiChiro G. Principles of computer assisted tomography in radiographic and radioisotope imaging. Phys Med Biol 1976;21:689–732.
10. Jain AK. Fundamentals of digital image processing. In: Kailath Th, editor. Englewood Cliffs (NJ): Prentice Hall; 1989.
11. King MA, et al. Attenuation, scatter, and spatial resolution compensation in SPECT. Emission Tomography: The Fundamentals of PET and SPECT. Chapt. 22. In: Wernick MN, Aarsvold JN, editors. San Diego: Elsevier Academic Press; 2004.
12. King MA, Tsui BMW, Pretorius PH. Attenuation/scatter/resolution correction: Physics Aspects. In: Zaret BL, Beller GA, editors. Clinical Nuclear Cardiology: State of the Art and Future Directions. 3d ed. Philadelphia: Elsevier Science; 2004.
13. Zaidi H, Hasegawa B. Determination of the attenuation map in emission tomography. J Nucl Med 2003;44(2):291–315.
14. Brasse D, et al. Correction methods for random coincidences in fully 3D whole-body PET: impact on data and image quality. J Nucl Med 2005;46(5):859–867.
15. Rousset OG, Ma Y, Evans AC. Correction for partial volume effect in PET: Principles and validation. J Nucl Med 1998;39:904–911.
16. Anton H. Elementary Linear Algebra. New York: John Wiley & Sons, Inc.; 2000.
17. Cooley JW, Tukey JW. An algorithm for the machine calculations of complex Fourier series. Math Comput 1965;19:297–301.
18. Dai X, et al. Left-ventricle boundary detection from nuclear medicine images. J Digit Imaging 1998;11(1):10–20.
19. Huang SC. Anatomy of SUV. Nucl Med Biol 2000;27:643–646.
20. Celsis P, Goldman T, Henriksen L, Lassen NA. A method for calculating regional cerebral blood flow from emission computed tomography of inert gas concentrations. J Comput Assist Tomogr 1981;5:641–645.
21. Bacharach SL, Green MV, Borer JS. Instrumentation and data processing in cardiovascular nuclear medicine: evaluation of ventricular function. Semin Nucl Med 1979;9(4):257–274.
22. Nakajima K, et al. Accuracy of Ventricular Volume and Ejection Fraction Measured by Gated Myocardial SPECT: Comparison of 4 Software Programs. J Nucl Med 2001;42(10): 1571–1578.
23. Todd-Pokropek A, Cradduck TD. (1998). Interfile resources. [Online]. The Keston Group. Available at http://www.keston.com/Interfile/interfile.htm. Accessed 2005 Sept 27.
24. Clark H. (2004). NEMA-Digital Imaging and Communications in Medicine (DICOM) Part 1 :Introduction and Overview. [Online]. National Electrical Manufacturers Association. Available at http://www.nema.org/stds/ps3-1.cfm. Accessed 2005 Sept 27.
25. Barbaras L, Parker JA, Donohoe KJ, Kolodny GM. (1996) Clinical data on the World Wide Web.
26. Surface D. Making PACS work with nuclear medicine. Radiol Today 2004;12:22.
27. Laβmann M, Reiners C. A DICOM-based PACS for nuclear medicine. Electromedica 2002;70:21–30.
28. Anonymous (2004) Image J. [Online]. National Institutes of Health. Available at http://rsb.info.nih.gov/ij. Accessed 2005 Sept 27.
29. Parker JA. (2004). Parker Plugins. [Online]. The Harvard Medical School. Available at http://www.med.harvard.edu/JPNM/ij/plugins/. Accessed 2005 Sept 27.
30. Wallis JW. Java and teleradiology. J Nucl Med 2000;41(1): 119–122.
31. Slomka PJ, Elliott E, Driedger AA. Java-based remote viewing and processing of nuclear medicine images: towards "the imaging deprtment without walls." J Nucl Med 2000;41(1): 111–118.
32. Slomka PJ. Software approach to merging molecular with anatomic information. J Nucl Med 2004;45(1 Suppl):36S–45S.
33. Lacroix KL. (1998) Home page for MCAT phantomo [Online] University of North Carolina. Available at http://www.bme.unc.edu/mirg/mcat/. Accessed 2005 Sept 27.
34. Buvat I, Castiglioni I. Monte Carlo simulations in SPET and PET. Q J Nucl Med 2002;46:48–61.
35. Ljungberg M, Strand S-E, King MA, editors. Monte Carlo Calculations in Nuclear Medicine. London: Institute of Physics Publishing; 1998.
36. Agostinelli S. (2005). Geant4 home page. [Online]. CERN. Available at wwwasd.web.cern.ch/wwwasd/geant4/geant4. html. Accessed 2005 Sept 27.
37. Ljungberg M. (2003). simind Monte Carlo home page. [Online]. Lund University. Available at http://www.radfys.lu.se/simind/. Accessed 2005 Sept 27.
38. Harrison R. (2004). Simulation system for emission tomography (SimSET) home page. [Online]. University of Washington. Available at http://depts.washington.edu/~simset/html/simset_main.html. Accessed 2005 Sept 27.
39. Liu JC. (2001). Electron gamma shower (EGS) Monte Carlo radiation transport code. [Online]. Stanford Linear Accelerator Center. Available at http://www.slac.stanford.edu/egs/index.html. Accessed 2005 Sept 27.
40. Morel C. (2004). GATE. [Online]. Ecole Federale Polytechnique de Lausanne. Available at http://www-lphe.epfl.ch/~PET/research/gate/index.html. Accessed 2005 Sept 27.
41. Zaidi H. (1997). Medical Physics on the Internet and the world-wide web. [Online]. Geneva University Hospital. Available at http://dmnu-pet5.hcuge.ch/Habib/Medphys.html. Accessed 2005 Sept 27.
42. Anonymous (2005) The Society of Nuclear Medicine Virtual Library. [Online]. The Society of Nuclear Medicine Inc. Available at http://snm.digiscript.com/. Accessed 2005 Sept 27.
43. Anthony Parker J. (2005). Joint Program in Nuclear Medicine: Electronic Learning Resources. [Online]. The Joint Program in Nuclear Medicine. Available at http://www.jpnm.org/elr.html. Accessed 2005 Sept 27.
44. Wallis J. (2005). MIR Nuclear medicine network access page. [Online]. The Mallinckrodt Institute of Radiology. Available at http://gamma.wustl.edu/home.html. Accessed 2005 Sept 27.
45. Johnson KA, Becker JA. (1999). The Whole Brain Atlas. [Online]. Harvard Medical School. Available at http://www.med.harvard.edu/AANLIB/home.html. Accessed 2005 Sept 27.
46. Rajadhyaksha CD, Parker JA, Barbaras L, Gerbaudo VH. (2004). Findings in 18FDG-PET & PET-CT. [Online]. Harvard Medical School and The Joint Program in Nuclear Medicine.

Available at http://www.med.harvard.edu/JPNM/chetan/. Accessed 2005 Sept 27.

Reading List

Lee K. Computers in NM: a Practical Approach. New York: The Society of Nuclear Medicine Inc; 1991.

Wernick MN, Aarsvold JN, editors. Emission Tomography: The Fundamentals of PET and SPECT. San Diego: Elsevier Academic Press; 2004.

See also NUCLEAR MEDICINE INSTRUMENTATION; RADIATION DOSE PLANNING, COMPUTER-AIDED; RADIATION THERAPY SIMULATOR; RADIOPHARMACEUTICAL DOSIMETRY.

NUTRITION, PARENTERAL

HOWARD SILBERMAN
University of Southern California
Los Angeles, California

INTRODUCTION

Parenteral nutrition, total parenteral nutrition (TPN), and hyperalimentation are terms referring to a variety of methods by which all required nutrients are provided intravenously, independent of alimentary tract function. Effective parenteral nutrition represents one of the major advances in medicine in the last 50 years and has led to intense interest in the incidence, consequences, and treatment of malnutrition, an area of study largely heretofore neglected because of a lack of effective therapy. The rationale for prescribing nutrients intravenously for patients unable to eat or receive tube feedings is based on the fact that malnutrition ultimately has an adverse impact on all organs and systems including the heart, lungs, gastrointestinal tract, and the immune system. The degree of morbidity and the rapidity of its onset depend on the magnitude and duration of nutritional deprivation. In addition, there are subtle, but perhaps quantitatively more important, indirect effects of nutritional deficits. Thus, nutritional derangements often have an adverse effect on the prognosis and treatment of concurrent illnesses. For example, malnourished patients who undergo elective surgery have a higher rate of postoperative sepsis and mortality. In addition, significant nutritional deficits may preclude the safe administration of optimal neoplastic chemotherapy because its inherent toxicity may be intolerable in the face of malnutrition.

DEVELOPMENT OF PARENTERAL NUTRITION

The adverse consequences of malnutrition became apparent to gastrointestinal (GI) surgeons during the first half of the twentieth century. Their patients with GI diseases were unable to eat, and the ensuing weight loss was associated with poor wound healing, anastomotic dysfunction and leakage, and poor surgical results. Such observations stimulated experimentation with intravenous nutrition. The early workers identified the nutrients required for human beings and further were able to prepare these nutrients in solutions that could be safely administered intravenously (1,2). However, clinical application was hampered by the fact that available nutrient solutions prepared in isotonic concentration could not meet energy and protein requirements when given in physiologic volumes. Experimentally, massive infusions of isotonic solutions of glucose and protein hydrolysates were indeed capable of supporting the anabolic state, but the intensive care required to monitor fluid balance made it a clinically impractical method. Early attempts at concentrating these solutions in order to provide all the required nutrients in acceptable volumes failed because of the thrombophlebitis that inevitably developed in the peripheral veins through which the solutions were infused. This was the state of the art until the mid-1960s when Rhoads, Dudrick, Wilmore, Vars, and their associates at the University of Pennsylvania undertook experiments in which very concentrated solutions were infused directly into the superior vena cava or the right atrium where instantaneous dilution of the solution occurred (3,4). By using hypertonic solutions of glucose and protein hydrolysates, these investigators demonstrated for the first time, initially in animals and then in humans, that intravenous "hyperalimentation" could support normal growth and development (2).

ESSENTIAL COMPONENTS OF NUTRIENT SOLUTIONS

The essential ingredients of solutions designed to meet all known nutritional requirements include nonprotein calories, utilizable nitrogen for protein synthesis, minerals, essential fatty acids, trace elements, and vitamins (5). In addition, it is likely that other micronutrients are necessary for optimal human nutrition. Normally, these as yet unidentified factors are automatically provided in a well-rounded diet derived from natural foodstuffs. Dudrick et al. (2,3) used glucose exclusively for nonprotein energy. Now fat emulsions are available as an additional clinically useful caloric source. Protein hydrolysates, derived from the enzymatic degradation of fibrin or casein, provided the nitrogen source in the early studies of parenteral nutrition. Currently, synthetic amino acid solutions are used to supply nitrogen.

Meeting the therapeutic goal of homeostasis (nitrogen equilibrium) or growth or nutritional repletion (positive nitrogen balance) is dependent on a variety of factors, most important among which are the levels of energy and nitrogen consumption. At any given level of protein or nitrogen intake, nitrogen balance progressively improves to some maximum level as caloric intake increases from levels below requirements to levels exceeding requirements (6). Maximum protein sparing and optimal utilization of dietary protein are achieved when the energy sources include at least 100–150 g of carbohydrate daily. The remaining energy requirements of most individuals can be met equally effectively by carbohydrate, fat, or a combination of these two. The requirement for this minimum amount of carbohydrate is based on its unique ability to satisfy the energy requirements of the glycolytic tissues, including the central nervous system, erythrocytes, leukocytes, active fibroblasts, and certain phagocytes.

At any given level of energy intake, nitrogen balance improves as nitrogen consumption increases. This dose-response relationship is curvilinear, and the nitrogen balance plateaus at higher dosages of nitrogen intake (6). To avoid the limiting effects of calories on nitrogen or of nitrogen on calories, parenteral nutrition solutions are prepared so that the nitrogen provided bears a fixed relationship to the nonprotein calories provided. In studies of normal, active young men fed orally, optimal efficiency was achieved at a calorie/nitrogen ratio of ~300–350 kcal to 1 g of nitrogen (6,7). However, protein economy decreases during most serious illnesses, and nitrogen losses increase; therefore, dietary protein requirements rise. Nitrogen equilibrium or retention can usually be achieved, however, by approximately doubling the quantity of nitrogen required by a normal man at any given level of caloric intake. Thus, a calorie/nitrogen ratio of 150:1 is thought optimal for seriously ill patients, although the ratio actually may range between 100 : 1 and 200 : 1 (6).

Minerals required in amounts exceeding 200 mg day^{-1} include sodium, potassium, calcium, magnesium, chloride, and phosphate. These macronutrients are essential for the maintenance of water balance; cardiac function; mineralization of the skeleton; function of nerve, muscle, and enzyme systems; and energy transformation. In addition, protein utilization is affected by the availability of sodium, potassium, and phosphorus in the diet; nitrogen accretion is impaired when any of these mineral nutrients is withdrawn from the diet. Nutritional repletion evidently involves the formation of tissue units containing protoplasm and extracellular fluid in fixed proportion and with fixed elemental composition (8). Thus, the retention of 1 g of nitrogen is characteristically associated with the retention of fixed amounts of phosphorus, potassium, sodium, and chloride.

Linoleic acid is the primary essential fatty acid for humans, and consequently it must be provided in order to avoid chemical and clinical evidence of deficiency. Linoleic acid requirements are generally met when a fat emulsion is used to provide at least 4% of calories as linoleic acid.

Micronutrients, or trace elements, presently recognized as essential for humans include iron, iodine, cobalt, zinc, copper, chromium, manganese, and possibly selenium. Cobalt is supplied as vitamin B12, and iron is generally withheld because of poor marrow utilization in critical or chronic illness. The remaining trace elements are routinely supplied in the nutrient solution (Table 1).

The remaining essential components are the water and fat soluble vitamins. Guidelines for parenteral vitamin administration have recently been revised by the Food and Drug Administration to include vitamin K, heretofore withheld, and increased amounts of vitamins B1, B6, C and folic acid (13–15) (Table 2).

Table 1. Trace Element Requirements[a–c]

Chromium	10–15 mcg
Copper	0.5–1.5 mg
Iodine	1–2 mcg kg^{-1}
Manganese	60–100 mcg
Zinc	2.5–4.0 mg

[a]Daily intravenous maintenance allowances for adults with normal initial values.
[b]Based on recommendations of Refs. 9–11.
[c]Selenium, 40–80 mcg day^{-1}, recommended for patients requiring long-term TPN, or those with burns, acquired immunodeficiency syndrome or liver failure. Ref. 12. p 125.

Table 2. Vitamin Requirements[a]

Thiamine (B_1)	6 mg
Riboflavin (B_2)	3.6 mg
Niacin (B_3)	40 mg
Folic acid	600 mcg
Pantothenic acid (B_5)	15 mg
Pyridoxine (B_6)	6 mg
Cyanocobalamin (B_{12})	5 mcg
Biotin	60 mcg
Ascorbic acid (C)	200 mg
Vitamin A	3300 IU
Vitamin D	200 IU
Vitamin E	10 IU
Vitamin K	150 mcg

[a]Daily intravenous allowances for adults. Based on Food and Drug Administration requirements Ref. 15.

The specific nutrient requirements for a given individual depend on the initial nutritional and metabolic status of the patient and his/her underlying disease process. Although the precise requirements for each nutrient can be determined by metabolic balance studies and direct or indirect calorimetry, such techniques are generally not employed in routine clinical practice. Instead, on the basis of data from clinical investigations applying such balance studies to patients with various diseases, injuries, and degrees of stress, requirements can be accurately estimated (16). As a result of these estimates, it is possible to formulate basic nutrient solutions of essentially fixed composition that can be used to meet the needs of most patients by varying only the volume of the basic formulation and by making simple adjustments in the electrolyte content of the solution. Thus, in clinical practice the caloric requirement is usually estimated from simple formulas that are adjusted for activity and stress. Although nitrogen requirements can likewise be determined or estimated, nitrogen needs are usually met as a consequence of the fixed calorie/nitrogen ratio of the nutrient solution. Thus, as the volume of infusate is increased to meet increased caloric demands, the additional nitrogen requirements, which generally parallel the rising caloric needs, are likewise met. Although such standard preparations can be used to satisfy the needs of most individuals, fluid-restricted patients, severely hypermetabolic patients, or those with renal or hepatic failure may require special nutrient formulations.

FORMULATING NUTRIENT SOLUTIONS

In current practice, nutrient solutions designed for parenteral administration are formulated to provide nonprotein calories as carbohydrate or a combination of carbohydrate and lipid. Glucose is the carbohydrate of choice since it is the normal physiologic substrate; it naturally occurs in blood; and it is abundant, inexpensive, and readily purified for intravenous administration. Glucose can be given in high concentrations and in large amounts that are well tolerated by most patients after a period of

adaptation. Other carbohydrates such as fructose, sorbitol, xylitol, and maltose have been evaluated experimentally, but each has disadvantages that preclude clinical application at the present time. Glucose for parenteral infusion is commercially available in concentrations from 5 to 70% and is provided as glucose monohydrate with a caloric density of 3.4 kcal·g^{-1}. Although isotonic (5%) solutions of glucose are available, concentrated glucose solutions are necessary in parenteral nutrition protocols in order to provide required calories in physiologic volumes of fluid.

Lipid is the alternative clinically useful nonprotein caloric source. Fat emulsions derived from soybean and safflower oil were approved for use in the United States in 1975 and 1979, respectively. The soybean oil emulsions had been used in Europe for nearly 20 years prior to their introduction in the United States. Currently available fat emulsions are derived from soybean oil or are mixtures of soybean and safflower oil emulsions. The use of lipid emulsions in intravenous feeding regimens is attractive because of the high caloric density of fat (9 kcal·g^{-1}), and because they are isotonic solutions that can, therefore, provide many calories in relatively small volumes via peripheral veins. Although early experience with lipids using the cottonseed oil emulsion Lipomul was unsatisfactory because of the toxicity of that preparation, fat emulsions derived from soybean and safflower oil have proven safe for clinical use. These newer preparations are purified plant oils emulsified in water. Egg phospholipids are added to regulate the size of the fat particles, stabilize the emulsion, and prevent fusion of the oil drops. Glycerol is added to make the emulsion isotonic, since oil and water emulsions have no osmolal effect. The resulting fat droplets have characteristics that are similar to those of naturally occurring chylomicrons found in the circulation after absorption of dietary lipid from the small intestine. Thus the particle size and the plasma elimination characteristics of these fat emulsions appear comparable to those of chylomicrons (17,18).

Studies of the elimination kinetics of soybean emulsion triglycerides indicate that at very low concentrations the rate of removal from the plasma is dependent on the triglyceride concentration. Above a certain critical concentration representing saturation of binding sites of lipoprotein-lipase enzymes, a maximum elimination capacity is reached that is independent of concentration. This maximum elimination capacity is influenced by the clinical state of the patient. It is increased during periods of starvation, after trauma, and in severely catabolic states (18,19). The infusion of fats is associated with an increase in heat production and oxygen consumption, a decrease in respiratory quotient, and the appearance of carbon-14 (^{14}C) in the expired air of patients receiving ^{14}C-labeled fat. These observations indicate that the infused fats are in fact used for energy. Soybean and soybean–safflower oil emulsions are available in 10, 20, and 30% concentrations and are mixtures of neutral triglycerides of predominantly unsaturated fatty acids. The major component fatty acids are linoleic, oleic, palmitic, and linolenic. The total caloric value of the 10% emulsions, including triglyceride, phospholipid, and glycerol, is 1.1 kcal·mL^{-1}. The corresponding values for the 20 and 30% emulsions are 2 and 3 kcal·mL^{-1}, respectively. In each of these preparations, ~0.1 kcal·mL^{-1} of the total caloric value is derived from the added glycerol.

All nutrient solutions must provide at least 100–150 g of glucose day^{-1} to meet the needs of the glycolytic tissues, as described above. The proportional distribution of glucose and fat to provide the remaining required nonprotein calories apparently can be widely variable with the expectation of achieving the same nutritional goals. Commonly cited guidelines suggest that fat should not provide >30% of nonprotein calories and that the daily dosage of fat not exceed 2.5 g kg^{-1} in adults (11). However, evidence-based reports indicate infusions providing >80% of nonprotein calories as fat and daily dosages of as much a 12 g fat kg^{-1} day^{-1} have been tolerated with clinical benefit and no adverse effects (20–22).

Consequently, protocols for solution preparation vary from institution to institution. Practical considerations in choosing the amount of glucose and the amount of fat relate primarily to the route of administration and to the fluid status of the patient. Parenteral nutrition solutions providing all nonprotein calories as glucose are highly concentrated and require central venous administration (see below). As an increasing proportion of the caloric content of the nutrient solution is provided by an isotonic fat emulsion, the content of glucose is thereby reduced, and, consequently, the concentration of the resultant solution falls. Nutrient solutions with an osmolarity not exceeding approximately three times normal serum levels can be successfully infused through peripheral veins (23); more concentrated solutions must be infused centrally. Parenteral nutrition solutions are formulated in accordance with one of three commonly used protocols: the glucose-based system, the lipid-based system, and a three-in-one system of variable composition.

The nutrient solution prescribed for a 24 h period usually is prepared in single container in a pharmacy under strict aseptic conditions. Manufacturing pharmacies responsible for preparing solutions for many patients often employ automated, computerized compounding apparatus that is programmed to add the specified nutrient components to the infusion container.

THE GLUCOSE SYSTEM

This is the original carbohydrate-based system developed by Dudrick and his associates at the University of

Table 3. The Glucose System[a–d]

50% glucose	500 mL
8.5% Amino acids	500 mL
Sodium (as acetate)	25 meq
Sodium (as chloride)	5 meq
Sodium (as phosphate)	14.8 meq
Potassium (as chloride)	40 meq
Phosphate (as sodium salt)	11.1 m*M*
Magnesium sulfate	8 meq
Calcium gluconate	5 meq

[a]Composition per liter.
[b]Provides 850 nonprotein kilocalories and 7.1 g nitrogen/L.
[c]Trace elements and vitamins are provided daily (see Tables 1 and 2).
[d]Electrolyte additives based on Travasol as the amino acid source.

Pennsylvania (3). The nutrient solution is prepared by the admixture of equal volumes of 50% glucose and 8.5% crystalline amino acids, and the addition of appropriate electrolytes, vitamins, and trace elements (Tables 1–3). A 1 L solution provides 850 nonprotein kilocalories, and ~7 g of nitrogen, equivalent to 44 g protein. Consequently, the solution has a calorie/nitrogen ratio of ~ 120 : 1. The nitrogen content and the calorie/nitrogen ratio will vary slightly depending on the brand of amino acid solution used. Due to the osmolar contribution of each of the constituents, this nutrient solution has a final concentration of ~ 2000 $mOsm \cdot L^{-1}$. Such a solution can never be safely infused through peripheral veins; consequently, the glucose system must be delivered into a central vein where the infusate is immediately diluted. Vascular access is usually through a percutaneously placed subclavian venous catheter or, for short-term use, a peripherally inserted central venous catheter (PICC line). Other routes (eg, via jugular, saphenous, or femoral veins) are used occasionally with variable success. The incidence of morbidity associated with establishing and maintaining central venous access is influenced by the site and technique of insertion of the venous cannula and the diligence with which the apparatus is managed during the course of nutrition therapy.

Because of the high concentration of glucose in this form of parenteral nutrition, therapy should begin gradually to allow adaption and thereby avoid hyperglycemia. Generally, on the first day a patient receives 1 L of nutrient solution, which is infused at a constant rate over the full 24 h period. Blood and urine glucose levels are monitored frequently, and if this initial rate of infusion is well tolerated, the volume prescribed is increased from day-to-day until the volume infused meets the caloric requirement of the individual patient. For the average patient, the nutritional requirements as well as the requirements for fluid and electrolytes are usually met by 2.5 $L \cdot day^{-1}$ of the nutrient solution, providing 2125 nonprotein kilocalories.

Infusion of the glucose system at a constant rate is a critical feature of safe practice since abrupt changes in the rate of delivery may be associated with marked fluctuations in blood sugar levels. The constant rate of infusion is most efficiently achieved by using an infusion pump. At the conclusion of therapy, the rate of infusion should be tapered gradually over several hours to avoid hypoglycemia. When infusion must be abruptly terminated, a solution of 10% glucose is substituted for the nutrient solution.

THE LIPID SYSTEM

The system of total parenteral nutrition based on glucose as the major caloric source is simple in concept and the least expensive, but patients' glucose metabolism must be closely monitored, and administration of the infusate requires technical expertise to achieve and maintain the central venous access necessary for safe treatment. The use of lipid emulsions as the major caloric source is attractive because of the high caloric density and isotonicity of these products. These considerations have logically led to the preparation of nutrient solutions based on fat as the major caloric source, with the goal of providing all required nutrients by peripheral vein. An example of such a lipid-based system of total parenteral nutrition is presented in Table 4. This nutrient solution was devised with the aim of maximizing caloric and amino acid content without producing a solution with a concentration that would preclude safe peripheral venous administration (23). Each liter provides 720-nonprotein kilocalories and 5.0 g of nitrogen, equivalent to 31 g of protein. The calorie/nitrogen ratio is 144: 1. The nitrogen content of the solution will vary slightly, depending on the amino acid product used in its preparation. As with the glucose system, sufficient volume is given to meet measured or estimated caloric requirements. The safety and efficacy of nutrient solutions utilizing lipid as the major caloric source were established by Jeejeebhoy and associates in their landmark investigation of lipid-based TPN in which 83% of nonprotein calories were supplied as fat (22). As many as 5 L daily of the lipid system described here have been infused for periods of weeks to months without apparent adverse effect.

Table 4. The Lipid System[a–e]

10% Fat emulsion	500 mL
50% Glucose	100 mL
8.5% Amino acids	350 mL
Sodium (as acetate)	35 meq
Sodium (as chloride)	5 meq
Sodium (as phosphate)	6 meq
Potassium (as chloride)	40 meq
Phosphate (as sodium salt)[e]	4.5 m*M*
Magnesium sulfate	8 meq
Calcium gluconate	5 meq
Heparin sodium	1000 U
Distilled water	q.s.ad 1000 mL

[a]Composition per liter.
[b]Provides 720 nonprotein kilocalories and 5.0 g nitrogen/L.
[c]Trace elements and vitamins are provided daily (see Tables 1 and 2).
[d]Electrolye additives are based on Travasol as amino acid source.
[e]Aproximately 7 m*M* additional phosphorus derived from fat emulsion.

THE THREE-IN-ONE SYSTEM

Innumerable nutrient solutions can be prepared with a distribution of glucose and lipid calories that differs from the two systems described above. Although there is no established biological advantage of differing proportions of fat and glucose as long as the minimum 100–150 g of carbohydrate are supplied, many clinicians prefer a profile of nutrients that mimics the optimal oral diet. Such a system calls for the admixture of amino acids, glucose, and lipids in which carbohydrate provides 65–85% of nonprotein calories and lipid 15–35% (11,12) (Table 5).

However, altering the lipid system described here by increasing the glucose content would have certain nonnutritional effects. Thus, a solution with a higher proportion of glucose calories could be produced by replacing some of the fat emulsion with isotonic glucose. The concentration of the final solution would remain unchanged, but a much greater total volume would be required to provide the same number of calories. If the substitution were made with hypertonic glucose, as called for in the three-in-one system

Table 5. A Three-in-One System[a–d]

50% Glucose	300 mL
10% Fat emulsion	300 mL
10% Amino acids	400 mL
Sodium (as acetate)	25 meq
Sodium (as chloride)	5 meq
Sodium (as phosphate)	14.8 meq
Potassium (as chloride)	40 meq
Phosphate (as sodium salt)	1.1 m*M*
Magnesium sulfate	8 meq
Calcium gluconate	5 meq

[a]Composition per liter.
[b]Provides 840 nonprotein kilocalories and 6.8 g nitrogen/L.
[c]Trace elements and vitamins are provided daily (Tables 4 and 5).
[d]Electrolyte additives based on Travasol as the amino acid source.

described in Table 5, the final concentration of the nutrient solution would be so increased as to require central venous administration, thereby losing the advantage of peripheral venous delivery.

COMPLICATIONS OF PARENTERAL NUTRITION

Morbidity associated with intravenous feedings may be related to drug toxicity, difficulties with vascular access, sepsis, or metabolic derangements.

Drug Toxicity

Adverse reactions to the components of parenteral nutrition solutions are uncommon. Although glucose is virtually nontoxic, the hypertonic solutions employed in the glucose system of TPN may be associated with potentially serious complications usually related to alterations in blood glucose levels. Currently used solutions of synthetic amino acids provide all of the nitrogen in the form of free l-amino acids and, in contrast to previously used protein hydrolysates, no potentially toxic ammonia or peptide products are present. Toxicity associated with the intravenous infusion of the currently available fat emulsions also has been minimal. The most frequent acute adverse reactions are fever, sensations of warmth, chills, shivering, chest or back pain, anorexia, and vomiting. Similarly, adverse reactions associated with chronic infusions of fat emulsions are also quite uncommon. Anemia and alterations in blood coagulation have been observed during treatment, but the etiologic relationship to lipid infusions has been unconfirmed. The most serious adverse effects have been observed in infants and children. The "fat overload" syndrome associated with the older cottonseed emulsion has rarely been observed with the newer current preparations. Nevertheless, several reports have been published (24) in which children receiving fat emulsions have developed marked hyperlipidemia, GI disturbances, hepatosplenomegaly, impaired hepatic function, anemia, thrombocytopenia, prolonged clotting time, elevated prothrombin time, and spontaneous bleeding. These findings resolved when the fat emulsion was withdrawn.

Complications of Vascular Access

The lipid-based system of parenteral nutrition can be infused through the ordinary peripheral venous cannulae used for the administration of crystalloid solutions. Local phlebitis and inflammation from infiltration and cutaneous extravasation occur with about the same frequency as that associated with the infusion of nonnutrient solutions. In contrast, a central venous catheter is required for infusion of the highly concentrated glucose and three-in-one systems. Insertion and maintenance of such catheters may be associated with a variety of complications. Complications that may occur during the placement of the catheter include improper advancement of the catheter tip into one of the jugular veins or the contralateral innominate vein, instead of into the superior vena cava. In addition, air embolization or cardiac arrhythmias may occur. Percutaneous jugular or subclavian cannulation may rarely result in an injury to an adjacent anatomic structure, such as the brachial plexus, great vessels, or thoracic duct. Pneumothorax, usually resulting from inadvertent entrance into the pleural cavity, is probably the most common complication of attempted subclavian catheterization and has been reported to occur in $\sim$ 2–3% of attempts in large series. Late complications after successful central catheterization may include air embolism, catheter occlusion, central vein thrombophlebitis, and catheter-related sepsis.

Systemic Sepsis

Sepsis attributable primarily to the administration of parenteral nutrition should be an infrequent complication in modern practice. A variety of factors may contribute to the development of this complication. Hyperglycemia, which may be induced or aggravated by nutrient infusions (see below), has been associated with sepsis in critically ill patients. Maintaining blood glucose levels between 80 and 110 $mg \cdot dL^{-1}$ has been shown to significantly reduce the incidence of septicemia (25). In addition, patients requiring TPN are often inordinately susceptible to infection because of serious illness, malnutrition, and chronic debilitation—all conditions associated with impaired immune responses. Patients receiving immunosuppressive therapy, cytotoxic drugs, or corticosteroids are likewise susceptible to infection. These drugs as well as prolonged administration of broad-spectrum antibiotics may subject patients to sepsis from unusual, ordinarily saprophytic, microorganisms.

In addition to these patient-related factors, several specific TPN-related factors contribute to the pathogenesis of sepsis. The various components of the nutrient solution can become contaminated during manufacture or at the time of component admixture in the hospital pharmacy. The ability of the nutrient solution to support microbial growth is well established, but with present techniques of solution preparation sepsis from contamination should be rare. The vascular access apparatus appears to be the most common source of TPN-associated sepsis. Contamination may take place when the infusion catheter is inserted; when containers of the nutrient solution are changed; when intravenous tubing is replaced; when in-line filters are inserted; or when the intravenous cannula is used for measurement of central venous pressure, blood sampling, or the infusion of medication or blood products. In addition, to-and-fro motion of a subclavian catheter due to inadequate fixation will allow exposed portions of the catheter to

enter the subcutaneous tract leading to the vein, which may result in infection. Hematogenous contamination of the infusion catheter may occasionally occur following bacteremia secondary to a distant focus of infection. More commonly, however, catheter-related sepsis is due to contamination of the catheter by organisms colonizing the skin surrounding the catheter insertion site. The incidence of sepsis varies greatly in reported series, but in recent years TPN has been administered with very low rates of infection. This improving trend is evidently due to adherence to rigid protocols of practice, and the employment in many hospitals of a dedicated, multidisciplinary team to manage the nutritional therapy. With this approach, TPN-related sepsis occurs in $\sim 3\%$ of patients receiving the glucose system. This complication is much less common among patients receiving the lipid-based system of parenteral nutrition through a peripheral vein (16).

Metabolic Complications

A variety of metabolic derangements have been observed during the course of total parenteral nutrition. These derangements may reflect preexisting deficiencies, or they may develop during the course of parenteral nutrition as a result of an excess or deficiency of a specific component in the nutrient solution. As would be expected, the standard solutions may not contain the ideal combination of ingredients for a given individual. In fact, adverse effects from an excess or deficiency of nearly every component of nutrient solutions have been described. Consequently, patients must be carefully monitored so that the content of the nutritional solution can be adjusted during the course of therapy. For example, minor alterations in electrolyte content are often necessary.

Abnormalities of blood sugar are the most common metabolic complications observed in patients receiving total parenteral nutrition. Hyperglycemia may be associated with critical illness independent of nutrient infusions. However, patients receiving the glucose-rich glucose and the three-in-one systems are particularly susceptible to elevated blood sugar levels. In addition, hyperglycemia may be manifest when the full caloric dosage of the glucose and three-in-one systems is inappropriately given initially and later if rates of infusion are abruptly increased. In addition, glucose intolerance may be a manifestation of overt or latent diabetes mellitus, or it may reflect reduced pancreatic insulin response to a glucose load, a situation commonly observed during starvation, stress, pain, major trauma, infection, and shock. Hyperglycemia also may be a reflection of the peripheral insulin resistance observed during sepsis, acute stress, or other conditions that are accompanied by high levels of circulating catecholamines and glucocorticoids. Decreased tissue sensitivity to insulin is also associated with hypophosphatemia, and hyperglycemia has been observed in patients with a deficiency of chromium. The latter trace metal probably acts as a cofactor for insulin. The incidence of hyperglycemia can be minimized by initiating therapy gradually with either of the two glucose-rich systems. Full dosage should be achieved over a 3 day period, during which time adaption to the glucose load takes place. In addition, careful metabolic monitoring during this period will disclose any tendency to hyperglycemia. Subsequently, a constant rate of infusion is maintained. An inadvertent decrease in the rate of the infusion should not be compensated by abrupt increases in rate; such "catching up" is not allowed. When hyperglycemia supervenes despite these precautions, the etiology is sought. The common cause of hyperglycemia after a period of stability is emerging sepsis, the overt manifestations of which may not appear for 18–24 h after development of elevated glucose levels.

Recent evidence indicates that maintenance of blood sugars levels between 80 and 110 $mg \cdot dL^{-1}$ in critically ill patients is associated with a significant reduction in mortality and the incidence of septicemia (25). Uncomplicated, moderate hyperglycemia is controlled initially by subcutaneous or intravenous administration of insulin; the TPN infusion is continued at the usual rate. Subsequently, the appropriate amount of insulin is added to the TPN solution during its aseptic preparation in the pharmacy. Providing insulin in the TPN solution has the advantage that inadvertent alterations in the rate of glucose delivery are automatically accompanied by appropriate adjustments in the amount of insulin administered. Patients with hyperglycemia complicated by massive diuresis, dehydration, neurologic manifestations, or the syndrome of hyperosmolar nonketotic coma are managed by immediate termination of the TPN infusion, fluid resuscitation, and insulin administration.

In contrast to the problem of hyperglycemia, blood sugar levels decrease when the rate of infusion of the glucose system is abruptly reduced. Symptomatic hypoglycemia is most likely to occur when the reduction of the infusion rate had been preceded by an increased rate. When the glucose system is to be discontinued electively, the rate of delivery should be tapered gradually over several hours. Patients who are hemodynamically unstable or who are undergoing surgery should not receive TPN, since fluid resuscitation may be inadvertently carried out using the TPN solution. Therefore, the TPN infusion is discontinued abruptly in such patients, and hypoglycemia is averted by infusing a solution of 10% glucose. Hypoglycemia may also reflect an excessive dosage of exogenous insulin. This most commonly occurs as a result of failure to recognize the resolution of peripheral insulin resistance and the associated decreased insulin requirement when the provoking condition responds to therapy.

Serum lipid profiles, which are routinely monitored during treatment with the lipid system, commonly reveal elevations of free fatty acids, cholesterol, and triglycerides. However, adverse clinical effects are uncommon (22,26). Nevertheless, triglyceride levels > 400 $mg \cdot dL^{-1}$ should be avoided since hypertrigylceridemia of this magnitude may be associated with an increased risk of pancreatitis, immunosuppression, and altered pulmonary hemodynamics (11).

Deficiencies of the major intracellular ions may occur in the catabolic state, since the protein structure of cells is metabolized as an energy source, intracellular ions are lost, and the total body concentration of these ions, including potassium, magnesium, and phosphate, are decreased. Furthermore, during nutritional repletion, these ions, derived from the serum, are deposited or incorporated in

newly synthesized cells. When supplementation of these ions in nutrient solutions is insufficient, hypokalemia, hypomagnesemia, and hypophosphatemia ensue. Serum levels of these substances should be measured regularly during TPN since such monitoring will disclose deficiencies before the clinical manifestations develop. Symptoms of hypokalemia are unusual when serum levels of potassium $> 3.0\,\text{meq}\cdot\text{L}^{-1}$. Asymptomatic hypokalemia can be managed by increasing the potassium supplement added to the nutrient solution at the time of preparation. When cardiac arrhythmias or other significant symptoms develop, the rate of TPN infusion should be tapered promptly while serum glucose levels are monitored closely, and an intravenous infusion of potassium chloride is begun.

Intracellular consumption of inorganic phosphate during the synthesis of proteins, membrane phospholipids, DNA, and adenosine triphosphate (ATP) may produce a striking deficit in the serum phosphate level after only several days of intravenous feedings devoid of or deficient in phosphate. Symptoms of hypophosphatemia may occur when serum phosphate levels fall to $2\,\text{mg}\cdot\text{dL}^{-1}$. However, severe manifestations are particularly apt to occur as levels fall to $< 1\,\text{mg}\cdot\text{dL}^{-1}$. These include acute respiratory failure, marked muscle weakness, impaired myocardial contractility, severe congestive cardiomyopathy, acute hemolytic anemia, coma, and death. Hypophosphatemic patients who are asymptomatic can be managed by increasing the phosphate supplement in the nutrient solution. Symptomatic patients or those with serum phosphate levels $< 1\,\text{mg}\ \text{dL}^{-1}$ should be repleted intravenously through a separate infusion line. Parenteral nutrition should be stopped, and a 10% glucose solution should be infused to avert hypoglycemia. Since intracellular phosphate consumption is dependent on caloric intake, withdrawing TPN alone often results in an increased serum phosphate level within 24 h.

A variety of adverse effects comprising the *refeeding syndrome* has been associated with the rapid induction of the anabolic state in severely malnourished, cachectic patients using standard nutrient solutions (16). Cardiac decompensation, the most serious feature of the syndrome, may be due to overhydration and salt retention in the face of starvation-induced low cardiac reserve. Hypophosphatemia, consequent to rapid refeeding, is another important contributing factor to cardiac failure. Rapid nutritional repletion also is implicated in producing deficits of the other major intracellular ions, magnesium and potassium, as well as acute deficiencies of vitamin A (associated with night blindness), thiamine (associated with the high output cardiac failure of beriberi, Wenicke's encephalopathy, and lactic acidosis), and zinc (associated with diarrhea, cerebellar dysfunction, dermatitis, impaired wound healing, and depressed immunity). Refeeding alkalosis also has been described. To avoid the refeeding syndrome in the chronically starved patient, parenteral nutrition should be introduced more gradually than usual, perhaps reaching the full caloric and protein requirements over the course of 5–7 days (16).

Healthy or malnourished individuals who receive a constant parenteral infusion of a fat-free, but otherwise complete diet eventually develop clinical and biochemical manifestations that are completely reversed by the administration of linoleic acid. Thus, the syndrome of essential fatty acid deficiency in humans is due principally, if not exclusively, to a lack of linoleic acid. Exogenous linolenic acid is required by some species, but its essentiality for humans is unproven. The most commonly recognized manifestation of linoleic acid deficiency is an eczematous desquamative dermatitis largely, but not always, confined to the body folds. Other clinical findings may include hepatic dysfuntion, anemia, thrombocytopenia, hair loss, and possibly impaired wound healing. Growth retardation has been observed in infants. Fatty acid deficiency is treated by the administration of linoleic acid, usually by infusing one of the currently available fat emulsions. Patients receiving the glucose-based system of parenteral nutrition should be treated prophylactically by providing 4% of calories as linoleic acid. This requirement is usually met by infusing $1\,\text{L}\cdot\text{week}^{-1}$ of a 10% fat emulsion.

Abnormalities in bone metabolism have been observed in patients receiving parenteral nutrition for prolonged periods, especially in home treatment programs. Such metabolic bone disease includes the common disorders of osteoporosis and osteomalacia and is characterized by hypercalciuria, intermittent hypercalcemia, reduced skeletal calcium, and low circulating parathormone levels. The clinical features have included intense periarticular and lower extremity pain. The pathogenesis of this syndrome is obscure, but hypotheses include an abnormality of vitamin D metabolism and aluminum toxicity (27–29). Most recently, vitamin K deficiency has been considered an etiologic factor since it has been recognized that this condition increases the risk of osteoporosis and fractures and that these risks can be reduced with vitamin K therapy. Vitamin K also appears to be necessary for the synthesis of a diverse group of proteins involved in calcium homeostasis (13,14,30,31). These findings have lead to the recent recommendation to routinely add vitamin K to TPN solutions, as discussed above.

NON-NUTRITIONAL EFFECTS OF PARENTERAL NUTRITION

Effects on the Stomach

Gastric acid secretion is significantly increased during the initial period of treatment with the glucose system, but the duration of this effect is unknown. The acid secretory response observed is due primarily to the infusion of crystalline amino acids, and this effect of amino acids on gastric secretion is virtually abolished by the concurrent intravenous infusion of a fat emulsion. The effect of chronic TPN on gastric secretory function is less clear. Chronic parenteral nutrition in animals has been associated with decreased antral gastrin levels and atrophy of the parietal cell mass. This observation is consistent with anecdotal clinical reports in which gastric hyposecretion has been observed in patients receiving long-term parenteral nutrition at home (32).

Effects on the Intestinal Tract

Morphologic and functional changes occur in the small intestine and the colon when nutrition is maintained

exclusively by vein. A significant reduction in the mass of the small and large intestine occurs, and there is a marked decrease in mucosal enzyme activity. Enzymes affected include maltase, sucrase, lactase, and peroxidase. These changes are not in response to intravenous nutrition per se, but reflect the need for luminal nutrients for maintenance of normal intestinal mass and function. The mechanism by which food exerts a trophic effect is at least in part endocrine in that intraluminal contents stimulate the release of enterotrophic hormones such as gastrin (16,33,34).

Effects on the Pancreas

Similar morphologic and functional atrophy of the pancreas is observed during the course of parenteral nutrition. In contrast to the effect of fat consumed orally, intravenous lipids do not stimulate pancreatic secretion (16,35).

Effects on the Liver

Transient derangements of liver function indexes occur in the majority of patients receiving parenteral nutrition regardless of the proportion of glucose and lipid (36,37). Similar abnormalities also have been observed in patients receiving enteral nutrition (tube feedings) (38,39). The etiology of these changes is uncertain and probably multifactorial. One hypothesis is that glucose and protein infusions in amounts exceeding requirements may contribute to these changes. In addition, an infectious etiology, perhaps related to the underlying condition requiring nutritional support, has been suggested, since oral metronidazole has been reported to reverse the changes in some patients. Administration of ursodesoxycholic acid also has been associated with improvement of TPN-related cholestasis (40). In any case, the clinical course associated with the liver changes is nearly always benign so that TPN need not be discontinued. Nevertheless, patients receiving TPN for several years or more are at greater risk for developing severe or chronic liver disease, but again the etiologic relationship is unclear (36).

Effects on the Respiratory System

Fuel oxidation is associated with oxygen consumption and carbon dioxide production. Oxygenation and carbon dioxide elimination are normal pulmonary functions. Consequently, patients with respiratory failure receiving aggressive nutritional support may not be able to meet these demands of fuel metabolism. It is particularly important to avoid infusing calories in amounts exceeding requirements, since this aggravates the problem, increases tidal volume, respiratory rate, and PCO_2, and offers no nutritional benefit (41).

INDICATIONS FOR PARENTERAL NUTRITION

Although the clinical benefits derived from nutritional substrates infused intravenously appear equivalent to those derived from substrates absorbed from the alimentary tract, feeding through the alimentary tract is preferable when feasible because this route of administration is less expensive, less invasive, and, most importantly, is associated with a significantly lower incidence of infectious complications (42). Nevertheless, many hospitalized patients have conditions in which alimentary tract nutrition either by mouth or tube feeding is inadequate, inadvisable, or would require an operative procedure (e.g., gastrostomy or jejunostomy) to establish access. It is for these patients that parenteral feeding should be considered. Normally nourished patients unable to eat for as long as 7–10 days generally do not require parenteral nutrition. The protein-sparing effect of 100–150 g of glucose provided in a 5% solution is sufficient. Patients in this category include those undergoing GI surgery in whom only several days of ileus are anticipated postoperatively. However, if the resumption of adequate intake is not imminent after 7–10 days, parenteral feedings are recommended. In contrast, normally nourished patients should receive TPN promptly when initial evaluation discloses gastrointestinal dysfunction that is expected to persist beyond 7–10 days. In addition, malnourished or markedly hypercatabolic patients (e.g., those with severe burns, sepsis, or multiple trauma) with GI dysfunction are given parenteral nutrition immediately.

In some patients, parenteral feedings have benefits in addition to improved nutrition. When all nutrients are provided intravenously, a state of bowel rest can be achieved in which the mechanical and secretory activity of the alimentary tract declines to basal levels (see earlier). These nonnutritional effects may be beneficial in the management of GI fistulas and acute inflammatory diseases, such as pancreatitis and regional enteritis. Parenteral nutrition may also be useful as a "medical colostomy". Thus, the reduction or elimination of the fecal stream associated with intravenous feedings may benefit patients with inflammation or decubitis ulcers adjacent to the anus or an intestinal stoma or fistula.

Any preexisting acute metabolic derangement should be treated before parenteral nutrition is begun. In addition, TPN should not be used during periods of acute hemodynamic instability or during surgical operations since the nutrient solution may be used inadvertently for fluid resuscitation. Parenteral nutrition is not indicated for patients with malnutrition due to a rapidly progressive disease that is not amenable to curative or palliative therapy.

COMPARING METHODS OF TOTAL PARENTERAL NUTRITION

Factors to be considered in comparing the glucose, the lipid, and the three-in-one systems of parenteral nutrition include the composition and nutrient value of the three systems, the relative efficacy of glucose and lipid calories, and the ease and safety of administration.

Comparative Composition of Parenteral Nutrition Systems

As outlined in Table 6, the lipid system provides fewer calories and less nitrogen per unit volume than the glucose and three-in-one systems. Thus, greater volumes of the lipid system are required to provide an isocaloric and isonitrogenous regimen. On the other hand, the lower osmolarity of the lipid system permits safe peripheral venous administraton of all required nutrients, whereas

Table 6. Comparison of Parenteral Nutrition Systems

	Glucose System	Lipid System	Three-in-One System
Carbohydrate calorie	850 kcal·L^{-1}	220 kcal·L^{-1}	540 kcal·L^{-1}
Lipid calories	0 kcal·L^{-1}	500 kcal·L^{-1}	300 kcal·L^{-1}
Caloric density	0.85 kcal·mL^{-1}	0.72 kcal·mL^{-1}	0.84 kcal·mL^{-1}
Nitrogen provided	7.1 g·L^{-1}	5.0 g·L^{-1}	6.8 g·L^{-1}
Protein equivalent	44 g·L^{-1}	31 g·L^{-1}	42.5 g·L^{-1}
Calorie/nitrogen ratio	120:1	144:1	124:1
Concentration (approximate)	2000 mOsm·L^{-1}	900 mOsm·L^{-1}	1500 mOsm·L^{-1}

the higher concentration of the other two systems mandates central venous infusion.

Glucose Versus Lipid as a Caloric Source

The relative impact of glucose and lipid calories on nitrogen retention or body composition has been the subject of extensive investigation often with disparate conclusions, depending on the subset of patients studied (16). However, the preponderance of evidence supports the conclusion that the two caloric sources are of comparable value in their effect on nitrogen retention in normal persons or in chronically ill, malnourished patients. The major study supporting this conclusion is that of Jeejeebhoy and associates (22), who observed that optimal nitrogen retention with the lipid system requires a period of $\sim$ 4 days to establish equilibrium, after which nitrogen balance is positive to a comparable degree with both the glucose and lipid systems. More recent data now attest to the equivalent efficacy of lipid as a major caloric source in critically illness and sepsis (43–45).

Ease and Safety of Administration

The glucose and three-in-one systems require central venous administration. Percutaneously inserted central venous catheters must be placed by physicians and peripherally inserted central venous catheters (PICC) by physicians or specially trained nurses under sterile conditions. Insertion and use of central catheters may be associated with certain complications discussed previously that are not seen with the peripherally administered lipid system.

The ordinary venous cannulae used for infusion of the lipid system can be easily inserted at the bedside and maintained by ward personnel. Whereas a central venous catheter requires special care and attention to prevent catheter sepsis, best provided by a dedicated team, the cannulae used in the lipid system require the same simple care as those used in the peripheral venous administration of crystalloid solutions. The peripherally-infused lipid system is rarely associated with systemic sepsis (16,23).

SELECTING THE TPN REGIMEN

For many patients, the nutritional requirements can be met equally well by any of the TPN systems discussed. The selection in these cases is often based on nonnutritional factors, such as the experience of the physician, ease of administration, and anticipated duration of therapy. On the other hand, there are subsets of patients requiring intravenous nutritional support who have associated or concurrent medical conditions that influence the choice of treatment.

Fluid Restriction

The lipid system described here has the lowest caloric and nitrogen content per unit volume of the three standard regimens (Table 6). Thus, a greater volume has to be infused to provide the same nutrients. Fluid restriction is facilitated, therefore, by prescribing the more concentrated glucose or three-in-one system. For patients who must be severely fluid restricted, these two systems may be modified by substituting 70% glucose and 10–15% amino acids for the 50 and 8.5% preparations, respectively, in order to supply equivalent nutrient content in a smaller volume. The recently available 30% fat emulsion, providing 3 kcal·mL^{-1}, may prove useful in designing additional TPN regimens for fluid-restricted patients.

Acute Myocardial Ischemia

In some studies, lipid infusions have been associated with elevated circulating free fatty acid levels. The effect of the latter on patients with acute myocardial ischemia is controversial, but there is evidence that arrhythmias may be precipitated and the area of ischemic damage may be extended in patients with acute myocardial infarctions (46–48). In view of these data, the glucose system is recommended in this group of patients.

Glucose Intolerance

It appears that hyperglycemia due to stress or diabetes mellitus is more easily managed if less glucose is infused, as in the three-in-one and lipid systems (49–51).

HYPERLIPIDEMIA

The lipid infusions are contraindicated in patients with conditions in which the metabolism of endogenous lipids is abnormal. Here the glucose system is prescribed.

Pulmonary Disease

In patients with pulmonary insufficiency it is particularly important that lipogeneis induced by excess glucose be avoided because it results in an increase in total CO_2

production, which may in turn lead to elevated PCO_2 values. In addition, significantly less CO_2 is produced during the metabolism of isocaloric amounts of lipid compared to glucose. Thus, increasing the proportion of lipid calories in the nutrient solution, as in the three-in-one and lipid systems, may facilitate the clinical management of patients with chronic pulmonary insufficiency and hypercarbia (52–54). In contrast, impaired pulmonary function has been observed when patients with acute respiratory distress syndrome receive lipids infusions. The adverse effects reported include decreased PO_2 and compliance and increased pulmonary vascular resistance (55).

HOME PARENTERAL NUTRITION

Methods of TPN have become sufficiently standardized and simplified that such care can now be safely and effectively provided at home on an ambulatory basis. Candidates for such homecare include those in whom the acute underlying medical condition requiring initial hospitalization has resolved, but who still require intravenous nutrition for a prolonged or indefinite period or even permanently. Patients with anorexia nervosa, Crohn's disease, short bowel syndrome, or severe hyperemesis gravidarum are among those who have been successfully managed with ambulatory TPN. Other candidates for home therapy include cancer patients with anorexia associated with chemotherapy or radiation therapy and patients with controlled enterocutaneous fistulas, radiation enteritis, or partial intestinal obstruction.

While the general principles of TPN outlined previously are applicable here, there are certain specific considerations in homecare necessary to make this method safe, convenient, and practical. Home patients should receive their nutrient solution through a tunneled, cuffed, silicone rubber or polyurethane central venous catheter (Hickman-type catheter). Such catheters are of low thrombogenicity, and passing the cuffed catheter through a subcutaneous tunnel reduces the incidence of ascending infection. Central placement, usually through the subclavian vein or internal jugular vein, frees the patient's extremities from any apparatus.

Whereas inpatient TPN is infused around the clock, home TPN is often infused in cyclic fashion, usually during sleeping hours, so that patients may be free of the infusion apparatus for part of the day. Patients must adapt to the more rapid hourly rates of infusion necessary to provide the required volume in a shorter period. Alterations of blood sugar, the commonest acute metabolic abnormalities, are best prevented by gradually increasing the rate of delivery over 1–2 h at the beginning of therapy and tapering the rate over several hours at the conclusion of the daily treatment.

Finally, chronic TPN for months or years appears to be unmasking requirements for additional nutrients that are stored in significant quantities or that are required in minute amounts. For example, further investigation may indicate requirements for selenium, molybdenum, taurine, and probably other micronutrients.

BIBLIOGRAPHY

1. Vinnars E, Wilmore D. Jonathan Roads Symposium Papers. History of parenteral nutrition. JPEN J Parenter Enteral Nutr 2003;27:225–231.
2. Dudrick SJ. Early developments and clinical applications of total parenteral nutrition. JPEN J Parenter Enteral Nutr 2003;27:291–299.
3. Dudrick SJ, Wilmore DW, Vars HM, Rhoads JE. Long-term total parenteral nutrition with growth, development, and positive nitrogen balance. Surgery 1968;64:134–142.
4. Rhoads JE. The history of nutrition. In: Ballinger WF, Collins JA, Drucker WR, Dudrick SJ, Zeppa R, editors. Manual of Surgical Nutrition. Philadelphia: W. B. Saunders; 1975. pp. 1–9.
5. Silberman H. Nutritional requirements. In: Silberman H, editor. Parenteral and Enteral Nutrition. 2nd ed. Norwalk, (CT): Appleton & Lange; 1989. pp. 85–116.
6. Wilmore DW. Energy requirements for maximum nitrogen retention. In: Greene HL, Holliday MA, Munro HN, editors. Clinical Nutrition Update: Amino Acids. Chicago: American Medical Association; 1977. pp. 47–57.
7. Calloway DH, Spector H. Nitrogen balance as related to caloric and protein intake in active young men. Am J Clin Nutr 1954;2:405–412.
8. Rudman D, Millikan WJ, Richardson TJ, Bixler TJ, II, Stackhouse J, McGarrity WC. Elemental balances during intravenous hyperalimentation of underweight adult subjects. J Clin Invest 1975;55:94–104.
9. Guidelines for essential trace element preparations for parenteral use: A statement by an expert panel, AMA Department of Foods and Nutrition. JAMA 1979;241:2051–2054.
10. Shils ME.. Minerals in total parenteral nutrition. Proceedings of the AMA Symposium on Total Parenteral Nutrition Nashville (TN): January 17–19 1972. pp. 92–114.
11. Mirtallo J, Canada T, Johnson D, Kumpf V, Petersen C, Sacks G, Seres D, Guenter P. Safe practices for parenteral nutrition. JPEN J Parenter Enteral Nutr 2004;28:S39–S70.
12. Mirtallo J. Parenteral formulas. In: Rombeau JL, Rolandelli RH, editors. Clinical Nutrition: Parenteral Nutrition. 3rd ed. Philadelphia: W.B. Saunders Company; 2001. pp. 118–139.
13. Bern M. Observations on possible effects of daily vitamin K replacement, especially upon warfarin therapy. JPEN J Parenter Enteral Nutr 2004;28:388–398.
14. Helphingstine CJ, Bistrian BR. New Food and Drug Administration requirements for inclusion of vitamin K in adult parenteral multivitamins. JPEN J Parenter Enteral Nutr 2003;27:220–224.
15. Fed Reg 2000;65:21200–212010.
16. Silberman H. Parenteral and Enteral Nutrition. 2nd ed. Norwalk, (CT): Appleton & Lange; 1989.
17. Hallberg D. Therapy with fat emulsion. Acta Anaesthesiol Scand Suppl 1974;55:131–136.
18. McNiff BL. Clinical use of 10% soybean oil emulsion. Am J Hosp Pharm 1977;34:1080–1086.
19. Hallberg D. Studies on the elimination of exogenous lipids from the blood stream. The effect of fasting and surgical trauma in man on the elimination rate of a fat emulsion injected intravenously. Acta Physiol Scand 1965;65:153–163.
20. Blanchard R, Gillespie D. Some comparisons between fat emulsion and glucose for parenteral nutrition in adults at the Winnipeg Health Sciences Center. In: Meng H, Wilmore D, editors. Fat Emulsions in Parenteral Nutrition. Chicago: American Medical Association; 1976. pp. 63–64.

21. Hadfield J. High calorie intravenous feeding in surgical patients. Clin Med 1966;73:25–30.
22. Jeejeebhoy KN, Anderson GH, Nakhooda AF, Greenberg GR, Sanderson I, Marliss EB. Metabolic studies in total parenteral nutrition with lipid in man. Comparison with glucose. J Clin Invest 1976;57:125–136.
23. Silberman H, Freehauf M, Fong G, Rosenblatt N. Parenteral nutrition with lipids. JAMA 1977;238:1380–1382.
24. Hansen LM, Hardie BS, Hidalgo J. Fat emulsion for intravenous administration: clinical experience with intralipid 10%. Ann Surg 1976;184:80–88.
25. van den Berghe G, Wouters P, Weekers F, Verwaest C, Bruyninckx F, Schetz M, Vlasselaers D, Ferdinande P, Lauwers P, Bouillon R. Intensive insulin therapy in the critically ill patients. N Engl J Med 2001;345:1359–1367.
26. Eisenberg D, Schmidt B, Silberman H. Safety and efficacy of lipid-based TPN: I Effects of 20% fat emulsion on serum lipids and respiratory functions. JPEN J Parenter Enteral Nutr 1982;6: 586.
27. Fuhrman MP. Complication management in parenteral nutrition. In: Matarese LE, Gottschlich MM, editors. Contemporary Nutrition Support Practice: A Clinical Guide. 2nd ed. Philadelphia: W.B. Saunders; 2003.
28. Klein GL, Targoff CM, Ament ME, Sherrard DJ, Bluestone R, Young JH, Norman AW, Coburn JW. Bone disease associated with total parenteral nutrition. Lancet 1980;2:1041–1044.
29. Shike M, Harrison JE, Sturtridge WC, Tam CS, Bobechko PE, Jones G, Murray TM, Jeejeebhoy KN. Metabolic bone disease in patients receiving long-term total parenteral nutrition. Ann Intern Med 1980;92:343–350.
30. Buchman AL, Moukarzel A. Metabolic bone disease associated with total parenteral nutrition. Clin Nutr 2000; 19:217–231.
31. Hamilton C, Seidner DL. Metabolic bone disease and parenteral nutrition. Curr Gastroenterol Rep 2004;6:335–341.
32. Kotler DP, Levine GM. Reversible gastric and pancreatic hyposecretion after long-term total parenteral nutrition. N Engl J Med 1979;300:241–242.
33. Magnotti LJ, Deitch EA. Mechanisms and significance of gut barrier function and failure. In: Rolandelli RH, Bankhead R, Boullata JI, Compher CW, editors. Clinical Nutrition: Enteral and Tube Feeding. 4th ed. Philadelphia: Elsevier-Saunders; 2005. pp. 23–31.
34. Tilson MD. Pathophysiology and treatment of short bowel syndrome. Surg Clin North Am 1980;60:1273–1284.
35. Grundfest S, Steiger E, Selinkoff P, Fletcher J. The effect of intravenous fat emulsions in patients with pancreatic fistula. JPEN J Parenter Enteral Nutr 1980;4:27–31.
36. Shattuck KE, Klein GL. Hepatobiliary complications of parenteral nutrition. In: Rombeau JL, Rolandelli RH, editors. Clinical Nutrition: Parenteral Nutrition. 3rd ed. Philadelphia: W.B. Saunders; 2001. pp. 140–156.
37. Wagner WH, Lowry AC, Silberman H. Similar liver function abnormalities occur in patients receiving glucose-based and lipid-based parenteral nutrition. Am J Gastroenterol 1983;78:199–202.
38. Kwan V, George J. Liver disease due to parenteral and enteral nutrition. Clin Liver Dis 2004;8:ix–x, 893–913.
39. Silk DBA. Nutritional Support in Hospital Practice. Oxford: Blackwell Scientific Publications; 1983.
40. Krawinkel MB. Parenteral nutrition-associated cholestasis—what do we know, what can we do? Eur J Pediatr Surg 2004; 14:230–234.
41. Askanazi J, Rosenbaum SH, Hyman AI, Silverberg PA, Milic-Emili J, Kinney JM. Respiratory changes induced by the large glucose loads of total parenteral nutrition. JAMA 1980;243: 1444–1447.
42. Gramlich L, Kichian K, Pinilla J, Rodych NJ, Dhaliwal R, Heyland DK. Does enteral nutrition compared to parenteral nutrition result in better outcomes in critically ill adult patients? A systematic review of the literature. Nutrition 2004;20:843–848.
43. de Chalain TM, Michell WL, O'Keefe SJ, Ogden JM. The effect of fuel source on amino acid metabolism in critically ill patients. J Surg Res 1992;52:167–176.
44. Druml W, Fischer M, Ratheiser K. Use of intravenous lipids in critically ill patients with sepsis without and with hepatic failure. JPEN J Parenter Enteral Nutr 1998;22:217–223.
45. Garcia-de-Lorenzo A, Lopez-Martinez J, Planas M, Chacon P, Montejo JC, Bonet A, Ortiz-Leyba C, Sanchez-Segura JM, Ordonez J, Acosta J, Grau T, Jimenez FJ. Safety and metabolic tolerance of a concentrated long-chain triglyceride lipid emulsion in critically ill septic and trauma patients. JPEN J Parenter Enteral Nutr 2003;27:208–215.
46. Editorial: Free fatty acids and arrhythmias after acute myocardial infarction. Lancet 1975;1:313–314.
47. Jones JW, Tibbs D, McDonald LK, Lowe RF, Hewitt RL. 10% Soybean oil emulsion as a myocardial energy substrate after ischemic arrest. Surg Forum 1977;28:284–285.
48. Opie LH, Tansey M, Kennelly BM. Proposed metabolic vicious circle in patients with large myocardial infarcts and high plasma-free-fatty-acid concentrations. Lancet 1977;2:890–892.
49. Baker JP, Detsky AS, Stewart S, Whitwell J, Marliss EB, Jeejeebhoy KN. Randomized trial of total parenteral nutrition in critically ill patients: metabolic effects of varying glucose-lipid ratios as the energy source. Gastroenterology 1984;87: 53–59.
50. Meguid MM, Schimmel E, Johnson WC, Meguid V, Lowell BC, Bourinski J, Nabseth DC. Reduced metabolic complications in total parenteral nutrition: pilot study using fat to replace one-third of glucose calories. JPEN J Parenter Enteral Nutr 1982;6:304–307.
51. Watanabe Y, Sato M, Abe Y, Nakata Y, Lee T, Kimura S. Fat emulsions as an ideal nonprotein energy source under surgical stress for diabetic patients. Nutrition 1995;11:734–738.
52. Silberman H, Silberman AW. Parenteral nutrition, biochemistry and respiratory gas exchange. JPEN J Parenter Enteral Nutr 1986;10:151–154.
53. Askanazi J, Nordenstrom J, Rosenbaum SH, Elwyn DH, Hyman AI, Carpentier YA, Kinney JM. Nutrition for the patient with respiratory failure: glucose vs. fat. Anesthesiology 1981;54:373–377.
54. Sherman SM. Parenteral nutrition and cardiopulmonary disease. In: Rombeau JL, Rolandelli RH, editors. Clinical Nutrition: Parenteral Nutrition. 3rd ed. Philadelphia: W.B. Saunders; 2001. pp. 335–352.
55. Lekka ME, Liokatis S, Nathanail C, Galani V, Nakos G. The impact of intravenous fat emulsion administration in acute lung injury. Am J Respir Crit Care Med 2004;169:638–644.

See also DRUG INFUSION SYSTEMS; GLUCOSE SENSORS; HOME HEALTH CARE DEVICES.

NYSTAGMOGRAPHY. See OCULAR MOTILITY RECORDING AND NYSTAGMUS.

O

OCULAR FUNDUS REFLECTOMETRY

Amol D. Kulkarni
Amruta M. Dattawadkar
University of Wisconsin,
Madison, Wisconsin

INTRODUCTION

Ophthalmology involves a study of diagnosis and management of eye diseases. The retina, also called the fundus oculi, constitutes a major component of the posterior segment of the eye. The central area of the fundus that is responsible for vision is called macula. The macula has a large number of photoreceptors that are specialized neurons containing colored light sensitive dyes (visual pigments). The center of macula is known as the fovea. The absorption of light initiates a cascade of events that bleaches these visual pigments and generates an electrochemical signal that is responsible for vision. This sequence of events is known as the visual cycle. Ocular fundus reflectometry is a noninvasive technique for an *in vivo* study of the visual cycle (1). It provides an objective and quantitative assessment of the kinetics of visual pigments.

Fundus reflectometry involves measuring the intensity of light of different wavelengths reflected by the ocular fundus (1). It has been primarily used to study reflectance properties of various structures in fundus, but also can be used for the study of oximetry and blood flow (2,3). However, the interpretation of measurement shows a lot of variation due to the effect of different types of photoreceptors with its own type of pigment, and spectral distortions. It is used in practice to characterize eye disorders with abnormalities in visual pigments, to detect autofluorescence of retinal lesions, and for various research studies in animals and humans (1).

HISTORICAL ASPECT

In 1851, a qualitative method for observation of the light reflected at the fundus was developed by Helmhotz. In 1952, the absorption spectrum of macular pigment was measured by Brindley and Willmer (4,5). In 1954–1971, the density and spectral properties of human rhodopsin was established by Rushton (6,7). He further developed a densitometer in 1971. A spectrophotographic technique that projects the entire spectrum of light was developed by Weale in 1953. This principle was used by Weale (8,9) to measure the density of cone pigments.

PHYSIOLOGIC BASIS AND PRINCIPLES

There are two types of photoreceptors in the human retina, namely, the rods and the cones. The rods are concerned with scotopic vision (dim lighting conditions) and the cones are responsible for photopic (daytime vision) and color vision. The cones are present in large numbers in the macula. Ultrastructure of both the types of photoreceptors as studied by electronmicroscopy consists of an outer and inner segment. The outer segments are made of stack of disks containing the visual pigment. The inner segment is responsible for pigment production and regeneration of outer segments. The pigment in rods is an aldehyde of vitamin A and in combination with protein (opsin) forms a compound known as rhodopsin. Of the various isomeric forms, the 11-cis form is a vital component of visual cycle and is converted to all-trans state on absorption of a photon (8). This sequence of events is responsible for vision and a similar process occurs in the cones (9).

The ability of visual pigments to absorb certain wavelength of the projected light can be determined by spectrophotographic techniques. This forms the basis of fundus reflectometry. In *in vitro* conditions, a monochromatic light can be projected on a sample of pigment and the intensity of the emergent beam is measured. It is possible to deduce the absorbing effect of the pigment by again projecting the light without the pigment. By using various wavelengths, an absorbance spectrum can be calculated. However, this is not possible *in vivo*, and hence an alternative technique has been devised. It consists of measuring the intensity of the emergent beam before and after bleaching of the pigment *in vivo*. This constitutes the physiologic basis of fundus reflectometry (1).

METHODS OF FUNDUS REFLECTOMETRY

The reflectometers can be classified as either spectral or imaging reflectometers. The various types of spectral reflectometers are Utrecht, Boston 1, Jena, Boston 2, and Utrecht 2[1]. These instruments measure the absolute spectral reflectance. The Utrecht reflectometer measures the foveal reflectance and determines the absorption characteristics of the cone visual pigments. The Utrecht 2 is a newer device and measures cone-photoreceptor directionality along with foveal reflectometry. The Boston 1 reflectometer was devised for oximetry and could simultaneously measure the reflectance at six wavelengths between 400 and 800 nm. The Boston 2 consisted of a modified Zeiss fundus camera, which could assess the orientation of foveal photoreceptors, their directionality, and the ratio of directional to diffuse flux (10–13). The Jena was a combination of a xenon lamp with a monochromator, and measured reflectance by photon-counting techniques.

The fundus imaging systems can also be modified to measure reflectance. This is called as imaging densitometry and has inferior resolution as compared to spectral densitometry. The various techniques used include fundus camera, video-based systems, scanning laser ophthalmoscope, and a charge-coupled device (CCD) camera. The

fundus camera and the video-based system generate maps of the visual pigment (14). However, there is an error in measuring reflectance due to stray light. Therefore a scanning laser ophthalmoscope (SLO) was developed in which a laser beam is moved in a raster pattern over the retina. In spite of high contrast and large dynamic range, the SLO did not provide a quantitative determination of fundus reflection. Around the same time, the CCD camera came into vogue and measured fundus reflectance spectra in 400–710 nm wavelength range.

The above-mentioned techniques of reflectometry have been used to develop models to quantify the spectral distribution of light pathways in human fundus. The various structures in the eye with reflectance properties include the cornea, lens, internal limiting membrane, nerve fiber layer, photoreceptors, retinal pigment epithelium, and sclera (15). On the contrary, the various structures that absorb light include lens, macular pigments, visual pigment, lipofuscin, melanin, and hemoglobin. Taking into consideration these various structures and their reflectance properties, numerous attempts have been made to study the visual cycle. These can be as simple as measuring light transmission by the retina to determining foveal fundus reflectance using spectral, directional (16–19) and bleaching effects.

CLINICAL APPLICATIONS OF FUNDUS REFLECTOMETRY

Fundus reflectometry is primarily used to estimate the optical density of various pigments in the eye (19,20). This includes the lens, macular, visual, and melanin pigments. In the lens and macular pigments, optical density measurement helps in determining the effects of aging. Visual pigments are vital component of the light cycle and densitometry can be used to classify photoreceptors on the basis of wavelength sensitive pigments. The extent of melanin pigmentation can be characterized by reflectometry and an index of pigmentation can be established (21).

Apart from measuring the pigment density, reflectometry is also used to study oxygen saturation, and orientation of foveal photoreceptors. These have applications in studying various congenital and acquired disorders of the retina including nutritional deficiency, infections, and degenerations (22–26). They can be used not only to characterize diseases, but also to study the effects of various treatment modalities (25). Moreover, sometimes it also contributes to early detection of particular diseases.

FUTURE DIRECTIONS

Fundus reflectometry has been principally used to measure the optical density of visual pigments and study the function of a normal and diseased retina. However, it is not used in routine patient care. The reason is because it is time consuming and requires complicated equipments. In addition the specificity is low, and so its use is limited only to research purposes. Therefore it has wide applications in epidemiologic studies, such as aging related ocular disorders (27,28).

Due to the ability to measure directionality and spatial distribution, fundus reflectometry is being tested in determining nerve fiber layer thickness, measurement of oxygen saturation, and to monitor the effects of laser therapy. As we move into a new era of prolonged longevity, due to advances in medicine, fundus reflectometry will be used to test new hypothesis and treatments.

BIBLIOGRAPHY

Cited References

1. Killlbride PE, Ripps H. Fundus reflectrometry. In: Martens BR, editor. Noninvasive Diagnostic Techniques in Ophthalmology. New York: Spinger; 1990. pp 479–498.
2. Beach JM, et al. Oximetry of retinal vessels by dual-wavelength imaging: calibration and influence of pigmentation. J Appl Physiol 1999;86:748–758.
3. Delori FC. Noninvasive technique for oximetry of blood in retinal vessels. Appl Opt 1998;27:1113–1125.
4. Killlbride PE, Alexander KR, Fishman GA. Human macular pigment assessed by imaging fundus reflectrometry. Vision Res 1989;29:663–674.
5. Chen SF, Chang Wu JC. The spatial distribution of macular pigment in humans. Curr Eye Res 2001;23:422–434.
6. Rushton WAH, Campbell FW. Measurement of rhodopsin in the living human eye. Nature (London) 1954;174:1096–1097.
7. Rushton WAH. Physical measurement of cone pigment in the living human eye. Nature (London) 1957;179:571–573.
8. Weale RA. Observations on photochemical reactions in living eyes. Br J Ophthalmol 1957;41:461–474.
9. Weale RA. Photochemical reactions in the living cat's retina. J Physiol 1953;121:322–331.
10. Van Blokland GJ. Directionality and alignment of the foveal receptors, assessed with light scattered from the human fundus *in vivo*. Vision Res 1986;26:495–500.
11. Zagers NPA, van de Karrats J, Berendschot TTJM, van Norren D. Simultaneous measurement of foveal spectral reflectance and cone photoreceptor directionality. Appl Opt 2002;41:4686–4696.
12. De Lint PJ, Berendschot TTJM, van Norren D. A Comparison of the optical stiles-crawford effect and retinal densitometry in a clinical setting. Invest Opthalmol Vis Sci 1998;39:1519–1523.
13. Macros S, Burns SA, He JC. Model for cone directionality reflectrometric measurements based on scattering. J Opt Soc Am A 1998;15:2012–2022.
14. Delori FC, Gragoudas ES, Francisco R, Pruett RC. Monochromatic ophthalmoscopy and fundus photography. The normal fundus. Arch Ophthalmol 1977;95:861–868.
15. Hammer M, Roggan A, Schweitzer D, Muller G. Optical properties of ocular fundus tissues -an in vitro study using the double-integrating-sphere technique and inverse Monte Carlo simulation. Phys Med Biol 1995;40:963–978.
16. Delori FC, Burns SA. Fundus reflectance and the measurement of crystalline lens density. J Opt Soc Am A 1996;13: 215–226.
17. Delori FC, Pflibsen KP. Spectral reflectance of the human ocular fundus. Appl Opt 1989;28:1061–1077.
18. Gorrand JM, Delori FC. A reflectometric technique for assessing photoreceptor alignment. Vision Res 1995;35:999–1010.
19. Wooten BR, Hammond BR Jr, Land RI, Snodderly DM. A practical method of measuring macular pigment optical density. Invest Ophthalmol Vis Sci 1999;40:2481–2489.
20. Savage GL, Johnson CA, Howard DL. A comparison of noninvasive objective and subjective measurement of the optical

density of human ocular media. Optom Vis Sci 2001;78:386–395.
21. Hunold W, Malessa P. Spectrophotometric determination of melanin pigmentation of the human fundus oculi. Ophthalmic Res 1974;6:355–362.
22. Highman VN, Weale RA. Rhodopsin density and visual threshold in retinitis pigmentosa. Am J Ophthalmol 1973; 75:822–832.
23. Carr RE, Ripps H, Siegel IM, Weale RA. Rhodopsin and the electrical activity of the retina in congenital night blindness. Invest Ophthalmol Vis Sci 1966;5:497–507.
24. Liem AT, Keunen JE, van Norren D. Clinical applications of fundus reflection densitometry. Surv Ophthalmol 1996;41: 37–50.
25. Augsten R, Konigsdorffer E, Schweitzer D, Strobel J. Non-proliferative diabetic retinopathy-reflection spectra of the macula before and after laser photocoagulation. Ophthalmologica 1998;212:105–111.
26. Landrum JT, Bone RA, Kilburn MD. The macular pigment: a possible role in protection from age-related macular degeneration. Adv Pharmacol 1997;38:537–556.
27. Delori FC, Goger DG, Dorey CK. Age-related accumulation and spatial distribution of lipofuscin in rpe of normal subjects. Invest Ophthalmol Vis Sci 2001;42:1855–1866.
28. Berendschot TTJM, et al. Influence of lutein supplementation on macular pigment, assessed with two objective techniques. Invest Ophthalmol Vis Sci 2000;41:3322–3326.

OCULAR MOTILITY RECORDING AND NYSTAGMUS

Louis F. Dell'Osso
Case Western Reserve
University Cleveland, Ohio

L. A. Abel
University of Melbourne
Melbourne, Australia

INTRODUCTION

This chapter will discuss the different types of eye movements generated by the ocular motor system, the advantages and disadvantages of commonly used recording systems, the requirements for accurate calibration of those systems, and the use of eye-movement recordings in research.

What Can We Record and Why? A Brief Introduction to Types of Eye Movements and Why We Record Them

Humans are highly visually driven animals. Our hearing may be inferior to that of the owl and our sense of smell far poorer than a dog's, but our visual acuity is excelled by few other species. High resolution vision, however, creates a bandwidth problem—if we processed our entire visual field simultaneously at maximal resolution, we would need so many optic nerve fibers to carry visual information back to the brain that our eyes might not fit into our heads. The solution that has evolved is to make the resolution of the retina—the light-sensitive neural layer of the eye—inhomogeneous. Visual acuity in the central 1° of the visual field is maximal, but it falls off rapidly as one moves toward the periphery. What keeps us from ever being aware of this fact is the nearly incessant motion of our eyes, which use a number of interconnected control systems to direct our gaze to an object of interest and to keep it fixated in the face of target and body movement. Considerable processing in the visual areas of the brain is needed to integrate the discontinuous flow of visual images into the clear, stable perception of the world that we usually experience.

EYE MOVEMENTS

The types of eye movements to be discussed here all play a part in the maintenance of vision. There are only 6 muscles per eye, arranged in opposing pairs and moving in a relatively constrained way by virtue of the anatomy of the orbit. Although each type of eye movement serves a specific purpose and is generated by partially distinct brain mechanisms, they nonetheless interact in the course of normal life. Examination and recording of eye movements has a surprisingly long history, going back to the pioneering work of Dodge and Cline (1). Eye-movement recording has enjoyed a number of advantages over the analysis of other motor control mechanisms. The following sections will briefly describe each type of eye movement, what purpose it serves, and why one might wish to record it.

Version and Vergence

The ocular motor system may be divided into two major subsystems: one that controls version (conjugate or conjunctive) eye movements, and one that controls vergence (disconjugate or disjunctive) eye movements. Saccades, pursuit, vestibuloocular, and optokinetic eye movements are types of version movements, and convergent and divergent refixations and pursuits are types of vergence eye movements. Patients may exhibit eye-movement abnormalities stemming from disorders in the version or vergence subsystem and both nystagmus and saccadic intrusions may be disconjugate, even uniocular. Recording systems used for all eye movements should be capable of independently recording data from both eyes, regardless of whether they are presumed to be conjugate, which is especially important when recording patients but is also applicable to normal individuals because conjugacy is not absolute. It is a common misconception that one can record "conjugate" movements from one eye only and presume the other eye is moving in exactly the same manner. In this chapter, only methods that fulfill this requirement are considered, regardless of either the experimental paradigm (version or vergence) or the subject population (normal or patient).

Saccades

Saccades are the fastest eye movements made, with velocities at times approaching 1000°/s. We make them nearly incessantly during our waking hours and during rapid eye movement sleep. Although at the end of each saccade only the most central area around the fixation location is seen with maximal acuity, our brains are able to integrate the

rapidly acquired series of such images into a single, unified perception of the world. Saccades may be horizontal, vertical, or oblique, which has implications for their recording, as will be discussed below. Evaluation of saccades may be grouped broadly into assessment of the saccades themselves and analysis of where they go as an individual views a scene or an image. Some eye trackers are more suitable for one sort of study than another. In particular, some methods are poorly suited to vertical and completely unsuited to torsional eye movements, where as others may have insufficient temporal resolution for assessment of latency or accuracy but excel at mapping sequences of fixations in two dimensions. In this discussion, a somewhat arbitrary distinction will be drawn between the detailed evaluation of individual saccades (as is often done clinically) and the assessment of scanpaths (as is sometimes used in a clinical setting but more often used in studies of man-machine interaction).

Inherent Saccadic Characteristics.

Accuracy. Saccadic accuracy is usually expressed in terms of gain (eye position/target position). Most commonly, if a refixation were comprised of multiple steps toward the target, the gain would be based only on the first step. Gain may be either abnormally high or abnormally low in different neurological conditions.

Latency. Latency is the time between stimulus onset and onset of eye movement. In humans, latency may range from 80 to several hundred ms, depending on the task and the age and health of the patient. Normally, saccades made in anticipation of target motion are excluded, unless stimuli with predictable location and timing are used. To be measured accurately, data must be acquired at a rate permitting the precise resolution of saccade timing (e.g., 500 Hz). Thus, a 25 or 30 Hz video-based system would be useless for this application.

Peak Velocity. Peak velocity can be measured either using analog electronics or, more commonly now, by off-line differentiation using software. Peak velocity is affected by fatigue, sedating drugs, and diseases that affect the cells in the brainstem that generate the fast, phasic component of saccadic innervation. Again, very low frame rates will make accurate calculation of peak velocity impossible, as it would not be possible to measure the rate of change in eye position. Indeed, if the sampling rate is too low, small saccades may be lost altogether, as they could be completed between samples (or video frames).

Scanpaths. Scanpaths can be divided into the descriptions of how individuals view a scene and nystagmus scanpaths that describe the eye trajectories about a fixation point in an individual with nystagmus. The former contain refixation saccades and periods of fixation whereas the latter contain the oscillatory nystagmus movements, braking, and foveating saccades, plus intervals of relatively stable fixation, if present.

Clinical Applications. Demonstration of how individuals (including patients) view a scene is probably the most familiar application of eye movement recording. In these applications, the "fine structure" of each saccade is of less interest than knowledge of where the saccades take the eyes and in what sequence. The classic work of Yarbus demonstrated the stereotyped way in which individuals view faces (2). As these investigations are focused on how cognitive processes control gaze, such work can be used to examine how patients with Alzheimer's disease (3) examine a novel scene or how individuals with schizophrenia attempt to judge the emotions expressed in a face. For scanpath analyses, high temporal resolution is unnecessary and spatial resolution on the order of a degree, not minute of arc, is acceptable. A wide linear range for vertical and horizontal eye movements is essential, however. Unobtrusiveness and minimal obstruction of the visual field are highly desirable when behavior is to be interfered with as little as possible.

Commercial Applications (Usability Studies, Man-Machine Interactions). Commercial applications are probably one of the most rapidly growing areas of eye movement research; it involves evaluating how humans interact with human displays. Here, the goal may be to see how a web page is examined or where a pilot is looking in a cockpit. The technical requirements for the eye tracker are essentially the same as for clinical applications. An exception is the area of gaze-contingent displays, where the endpoint of a saccade is predicted from recording its beginning, and the display is updated in high resolution only at that point. Such applications impose stricter temporal and spatial resolution criteria.

Nystagmus Scanpaths. Plots of the horizontal vs. vertical motion of nystagmus patients' eye movements during fixation of a stationary target provide insight into their ability to foveate the target in a stable (i.e., low retinal-slip velocity) and repeatable (i.e., low variance in the mean positions of target foveation intervals) manner. Nystagmus phase-plane (eye position vs. eye velocity) and scanpath plots were developed to study the foveation periods present in many of the waveforms seen in infantile nystagmus (4–8). They are important methods that provide insight into how individuals with such oscillations can achieve high visual acuity. The recording equipment for nystagmus scanpaths and phase-planes needs to be both accurate and of sufficient bandwidth to record the small saccades imbedded in nystagmus waveforms.

Smooth Pursuit. A correlate of having only a small part of the retina—the fovea—with high spatial resolution is that if a moving object is to be seen clearly, it must be tracked precisely, so that its image remains on the fovea, which is the function of the smooth pursuit system. The brain substrates underlying smooth pursuit are, to a degree, separable from those of the saccadic system, but, as a recent review has noted (9), a high degree of parallelism exists. Given that the two systems must work together for successful tracking, this fact is not surprising. For example, if you hear a bird call in the sky and decide to follow it, you must first locate it with a saccade (and possibly a head movement). Your pursuit system then

keeps your gaze on the bird, but if it moves too swiftly for this system, it can be reacquired by a saccade and tracking can the recommence. If it is lost again, the pattern repeats. Indeed, if pursuit gain (eye velocity/target velocity) is low or even zero, objects can still be tracked by repeated saccades, giving rise to the clinical observation of "cog-wheel pursuit."

In contrast to the many roles that saccades serve, pursuit eye movements are rather specialized for tracking. We all can generate saccades at will, even in the absence of targets, but voluntary generation of smooth pursuit is extremely rare and of poor quality. When recorded, it may be examined qualitatively for the presence of saccades or the smooth tracking segments can be separated out and their gain analyzed. As a result of the bilateral organization of motor control in the brain, it is possible to have a unidirectional pursuit abnormality, which may be of diagnostic value. However, bilaterally reduced smooth pursuit is nonspecific, as it may result from boredom, inattention, alcohol, fatigue, as well as pathology. As the pursuit system cannot track targets moving at greater than approximately 2 Hz, the requirements for its recording are not very demanding. If pursuit velocity is to be derived, however, then a low-noise system with appropriate low-pass filtering is essential to prevent the velocity signal from being swamped by noise. A low-noise system is also crucial in computer analysis of smooth pursuit because the algorithm used to identify saccades must ensure that none of the saccade is included in the data segment being analyzed as pursuit. If the pursuit component of the eye movement is only 5°/s and portions of saccades with velocities $\leq 30^\circ$/s are included, pursuit gain calculations may be highly inaccurate, which is a concern when commercial systems incorporating proprietary algorithms are being used in clinical settings where this possibility has not been anticipated. See Calibration (below) for more information.

Vestibulo-Ocular Response (VOR)

The VOR is a fast reflex whose purpose is to negate the effects of head or body movement on gaze direction. Acceleration sensors in the semicircular canals provide a head-velocity input to the ocular motor system that is used to generate an eye-velocity signal in the opposite direction. The sum of head and eye velocity cancel to maintain steady gaze in space. The VOR is tuned to negate fast head movements and works in concert with the optokinetic reflex (see below), which responds to lower frequency background motion.

Rotational Testing. For vision to be maximally effective, it must continue to work properly as humans move around in the environment. Consider what would happen if the eyes were fixed in the head as one walked about—the image falling on the retina would oscillate with every step. Every turn of the head would cause the point of regard to sweep away from the fovea. Relying on visual input to compensate would be far too slow to generate an accurate compensatory input. Therefore, humans possess the semicircular canals, three approximately (but not precisely) orthogonal transducers of rotational motion, as part of each inner ear. Only three neurons separate the canals from the extraocular muscles that move the eyes. The canals are filled with fluid and, as the head moves, the inertia of the fluid causes it to lag behind, stimulating displacement-sensitive hair cells at the base of each canal. With only two synapses between sensory transducer and motor effector, the core of the VOR pathway can act very rapidly. Note, however, that constant velocity rotation elicits a signal that eventually decays to baseline, as the fluid eventually ceases to lag behind the canals (i.e., it moves with the same rotational velocity as the canals). Of course, prolonged constant velocity rotations are not part of our evolutionary history and are rarely encountered in daily life.

As the function of the VOR is to facilitate the maintenance of stable gaze as we move around in the environment, it makes intuitive sense to assess it in a moving subject. The most common way to make this assessment is to measure the horizontal component of the VOR as the patient is rotated while in the seated position. Spring-loaded Barany chairs were eventually superseded by electrically driven chairs, which could be driven with velocity steps, sinusoidally, or with more complex inputs. Step inputs may be used to quantify the time constant of decay of the VOR, whereas the other inputs can be used to generate gain and phase plots. Directional asymmetries or abnormal gains can be readily detected with such testing. Such tests are also carried out not only under baseline conditions, with the patient in complete darkness, but also with the VOR suppressed (patients fixate a target rotating with them) or enhanced (patients fixate an earth-fixed target).

Rotary chair testing has several shortcomings, particularly for low frequencies (e.g., 0.05 Hz). It takes a long time to obtain several cycles of data, during which time the patient may be lulled to sleep by the slow rotation in the dark. Alerting tasks (e.g., mental arithmetic) can be used to overcome this shortcoming, but the overall testing time may be quite long. Stimuli such as pseudo-random binary sequences have been used, with data analyzed by cross-correlation (10) in order to obtain results across a wide range of frequencies more rapidly. Another limitation, however, is that in order to obtain VOR data with a chair, the entire patient must be rotated, which limits the frequency range of the technique, because rotations of, for example, a 100 kg individual at 2 Hz would require very high forces. In addition, high frequency rotations increase the likelihood that because of inertia, the patient would not rotate precisely in phase or with the same amplitude as the chair. Systems are available that record eye movement and sense head movement during patient-initiated head shaking, which allows for testing at more physiological frequencies, but it requires a cooperative patient.

A fundamental problem with rotary chair testing is that although directional differences can be detected, localizing pathology to one ear is difficult. Obviously, rotating only one side of the head is impossible, and the "push-pull" nature of the vestibular system (due to the juxtaposed semicircular canals in the ears) makes lateralization difficult. For this reason, the next test remains valuable, in spite of its shortcomings.

Caloric Testing. Introduced by Barany in 1903, caloric testing is probably the most widely used of all vestibular tests. When carried out using EOG, it is still often referred to as "electronystagmography" (ENG), a term that is sometimes mistakenly applied to all forms of eye-movement recording. It involves the irrigation of the ear canal with either warm water or cold water, which alters the behavior of the horizontal semicircular canal on the side being irrigated. Cold water simulates reduced ipsilateral activity and warm water simulates an irritative lesion; the temperature of the water thus determines the direction of the resulting induced nystagmus fast phase in the way described by the acronym COWS: cold, opposite; warm, same. Vestibular nystagmus frequency and amplitude can readily be assessed for each ear at various temperature levels, which remains the only practical way to assess each side of the vestibular system independently. However, caloric stimulation has the appreciable shortcoming that it is a dc input to the vestibular system. It thus assesses the function of the system far from its physiological frequency range of several Hertz.

Optokinetic Response (OKR)

The OKR is a slow reflex whose purpose is to negate the effects of retinal image movement on gaze direction. Velocity sensors in the retina provide an input to the ocular motor system that is used to generate an eye-velocity signal in the same direction, maintaining gaze on the moving background. The OKR is tuned to respond to slow retinal image movement and works in concert with the VOR (see above), which responds to high frequency head motion.

Full-Field. The optokinetic nystagmus (OKN) response, like the VOR, may be induced in healthy individuals with appropriate visual stimuli. The fundamental form of the optokinetic response is induced by motion of all (or most) of the visual field, which elicits a slow eye movement in the direction of the stimulus, with a fast phase bringing the eyes back toward their initial position. This response continues as long as the stimulus continues. If one considers how the VOR decays with continuous motion and has low gain at very low frequencies, then it can be seen that the OKR and the VOR are functionally additive. Indeed, the relationship between OKR and VOR can be readily observed by anyone who has sat gazing out of a train window and felt himself moving, only to discover that it was the adjacent train which was pulling out of the station. The optokinetic stimulus evokes activity in the vestibular nuclei of the brain, and this activity elicits a sense of motion—the most common way to activate the vestibular system. This visually-induced motion percept is known as linearvection if the motion is linear and as circularvection if the stimulus is rotational. The nature of OKN differs depending on whether the stimulus is actively followed or passively viewed.

Small-Field (Hand-Held Drum, Tape). Although "train nystagmus" may be relatively easy to induce in the real world, presentation of a full-field OKN stimulus in a clinical setting requires a stimulus that essentially surrounds the patient. For this reason, OKN is more often tested using either a small patterned drum or a striped tape, both of which can be easily held in the examiner's hands. Although the OKN induced in this way looks no different than that deriving from a full-field stimulus, it is primarily a smooth pursuit response, whereas the full-field OKN includes both pursuit components as well as responses deriving from subcortical pathways, including the lateral geniculate body, accessory optic system, nucleus of the optic tract, and the brain stem and cerebellar circuitry governing eye movements.

Spontaneous Nystagmus & Saccadic Intrusions or Oscillations

Diagnostic Classification. In addition to induced nystagmus, some subjects exhibit either spontaneous or gaze-evoked nystagmus or saccadic intrusions or oscillations. The waveforms and other characteristics of these movements often have diagnostic value; accurate calibration of the data is necessary to extract diagnostic information or to deduce the mechanisms underlying the genesis of an intrusion or oscillation (nystagmus or saccadic).

Nystagmus Versus Saccadic Oscillations. The first distinction to be made when spontaneous oscillations are present is to distinguish between the many types of nystagmus and saccadic intrusions and oscillations. Both the slow-phase waveforms and their relationships to target foveation (placement of the target image on the small, high resolution portion of the retina) help in making this determination. Although the details of these determinations are beyond the scope of this chapter, the basic difference is that nystagmus is generated and sustained by the slow phases, whereas saccadic intrusions and oscillations are initiated by saccades that take the eyes off-target.

Congenital Versus Acquired Nystagmus. If nystagmus is present, determination of its origin is necessary (i.e., is it congenital or acquired?). Again, this field is complex and cannot be fully discussed here. Suffice is to say, certain nystagmus waveforms exist that are pathognomonic of congenital nystagmus; they, along with characteristic variations with gaze angle, convergence angle, or fixating eye, help to determine whether a nystagmus is congenital or acquired.

OCULAR MOTOR RECORDING SYSTEMS

Overview of Major Eye-Movement Recording Technologies

The following are descriptions of the more common technologies used to record the eye movements of both normals and patients. Technical descriptions, engineering, and physics of these and other methods may be found elsewhere in this volume (see "Eye Movement Measurement Techniques"). Emphasis in this chapter will be on the abilities of different types of systems and the calibration requirements to provide accurate eye-movement data in the basic and clinical research settings.

Electrooculography. ***Theory of Operation.*** Electrooculography (EOG) is the only eye-movement recording

method that relies on a biopotential, in this case, the field potential generated between the inner retina and the pigment epithelium. This signal may approach 0.5 mV or more in amplitude. If two electrodes are placed on either side of, and two more above and below, the orbit (along with a reference electrode on the forehead or ear), then as the eye rotates in the orbit, a voltage proportional to the eye movement may be recorded, because one electrode becomes more positive and the other more negative with respect to the reference electrode. The technique is one of the oldest and most widespread and has been the standard for assessment of eye movements related to vestibular function. When the term ENG is seen, it is generally EOG that is used.

Characteristics. EOG has the considerable advantage that it requires only a high impedance, low noise instrumentation amplifier for its recording and that the voltage is linearly proportional to eye movement over most of its range. Such amplifiers are relatively inexpensive in comparison with many other eye-tracking technologies. As the electrodes are placed on the skin adjacent to the eye, no contact occurs with the eye itself and no obstruction of any part of the visual field exists. It also is unaffected by head motion, because the electrodes move with the head.

Applications. In theory, the EOG can be used anywhere eye movements are to be recorded. However, as the following section will show, it has a number of inherent limitations that practically eliminate it from many applications. Its widest use remains in the assessment of vestibular function and for the recording of caloric nystagmus and the vestibulo-ocular reflex. It is unsuited for use in environments with changing levels of illumination, as normal physiological processes will change the resting potential of the EOG and thus alter its relationship with amplitude of eye movement. EOG can be used in the assessment of saccades and smooth pursuit, but the low-pass filtering generally required will lead to artificially lowered saccade peak velocities. EOG has occasionally been used in scanpath studies, but its instability and fluctuating gain make it undesirable for this application, because if scenes differing in mean luminance are presented, the EOG will gradually change amplitude.

Limitations. Although conceptually simple and easy to implement, EOG has many shortcomings. One is that because the electrodes are placed on the surface of the facial skin, the EOG is not the only signal they detect. If the patient is nervous or clenches his or her teeth, the resulting electromyographic (EMG) activity in the facial muscles will be recorded as well, with the result that the signal actually recorded is the sum of the desired EOG and the unwanted EMG. As the spectra of the two signals overlap, no amount of filtering can completely separate them.

Another significant problem with EOG is the fact that, like many biopotential recordings, it is prone to drift. Some of this droft may reflect electrochemical changes at the electrode, causing a shift in baseline, which was particularly a problem when polarizable electrodes were used in the early days of the technique. Even nonpolarizable electrodes such as the commonly used Ag-AgCl button electrodes may still yield a varying baseline when first applied. Furthermore, the potential also shifts with changes in illumination. Indeed, assessment of this response to light is itself a clinical tool. This baseline variability can lead to the temptation to use an ac-coupled amplifier in the recording of the EOG, which has frequently been done, particularly in the ENG literature. Although not a problem if the only data required is nystagmus frequency, significant distortion occurs when ac-coupling is used to record saccades. The apparent drift back toward the center closely resembles a saccade whose tonic innervational component is inadequate. Noise and drift limit the resolution of EOG to eye movements of no less than 1°; this threshold may be even higher in a nervous patient or an elderly patient with slack, dry skin. An additional limitation undercuts the EOG's otherwise significant advantage in being able to record vertical eye movements, which is the overshooting seen on vertical saccades. It has long been suggested that the lids, moving somewhat independently of the globe, act as electrodes on the surface of the globe, conducting current in parallel to the other current path between globe and electrodes (11).

Another more practical drawback to the use of EOG when used for recording the movements of both eyes horizontally and vertically is that a total of nine electrodes are required (see Fig. 1). Each must be individually adhered to the patient and must be carefully aligned if spurious crosstalk between horizontal and vertical motion is to be avoided. Even if only horizontal motion is to be recorded, five accurately placed electrodes are still needed. A common but unfortunate clinical shortcut has been to use only three—two at either outer canthus of the eye and one

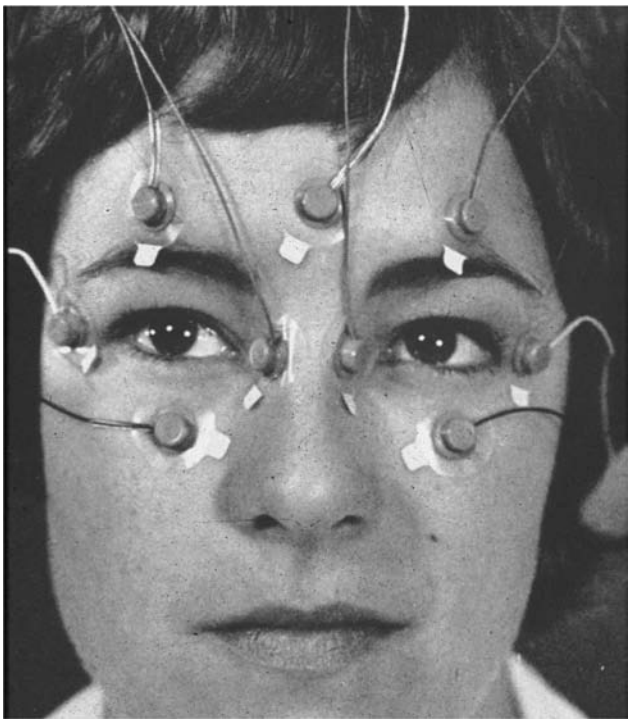


Figure 1. EOG electrodes arranged to record the horizontal and vertical eye movements of both eyes. Reference electrode is in the center of the forehead.

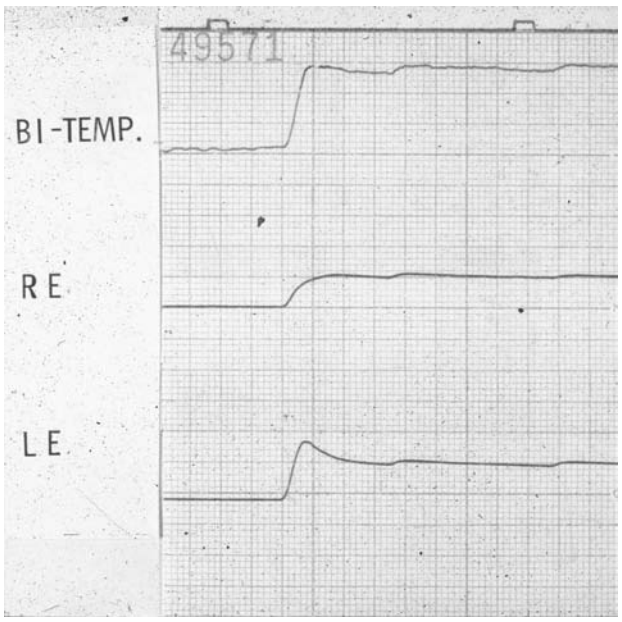


Figure 2. False saccadic trajectory from bitemporal EOG electrodes resulting from the summation of the individual saccadic trajectories shown below.

for reference. This shortcut effectively records a "cyclopean" eye by summing the potentials obtained from each eye. Although eye movements other than vergence are conjugate in normal individuals, it is not generally normal individuals who are seen for clinical evaluation. Figure 2 illustrates how an overshooting and an undershooting eye movement may be combined to give the appearance of a perfect saccade. For this reason, both ac-coupling and bitemporal electrode placement should be avoided when anything other than the crudest information about eye movement is desired.

Infrared Reflectance.

Theory of Operation. Although photographic recording of eye movements dates back to 1901 (1), such methods remained cumbersome to use, especially when they required frame-by-frame analysis of the location of some marker on the eye. Optical levers, where a beam of light was reflected from a mirror attached by a stalk to a scleral contact lens, offered the opportunity for precise registration of eye position, but occluded the view of the eye being recorded. As might be imagined, they were also unpleasant to wear. An alternative recording method that also makes use of reflected light relies on the differential reflectivity of the iris and sclera of the eye to track the limbus—the boundary between these structures. The earliest versions of this system were developed by Torok et al. (12) and refined by several investigators over the years (13–15). Although the number of emitters and detectors vary between designs, they share the same fundamental principle; that is, the eye is illuminated by chopped, low intensity infrared light (to eliminate the effects of variable ambient lighting). Photodetectors are aimed at the limbus on either side of the iris. As the eye moves, the amount of light reflected back onto some detectors increases and onto others decreases. The difference between the two signals provides the output signal. As would be expected, these signals are analog systems, so that the output of the photodetectors is electronically converted into a voltage that corresponds to eye position. Figure 3 shows an IR system mounted on an earth-fixed frame (a), spectacle

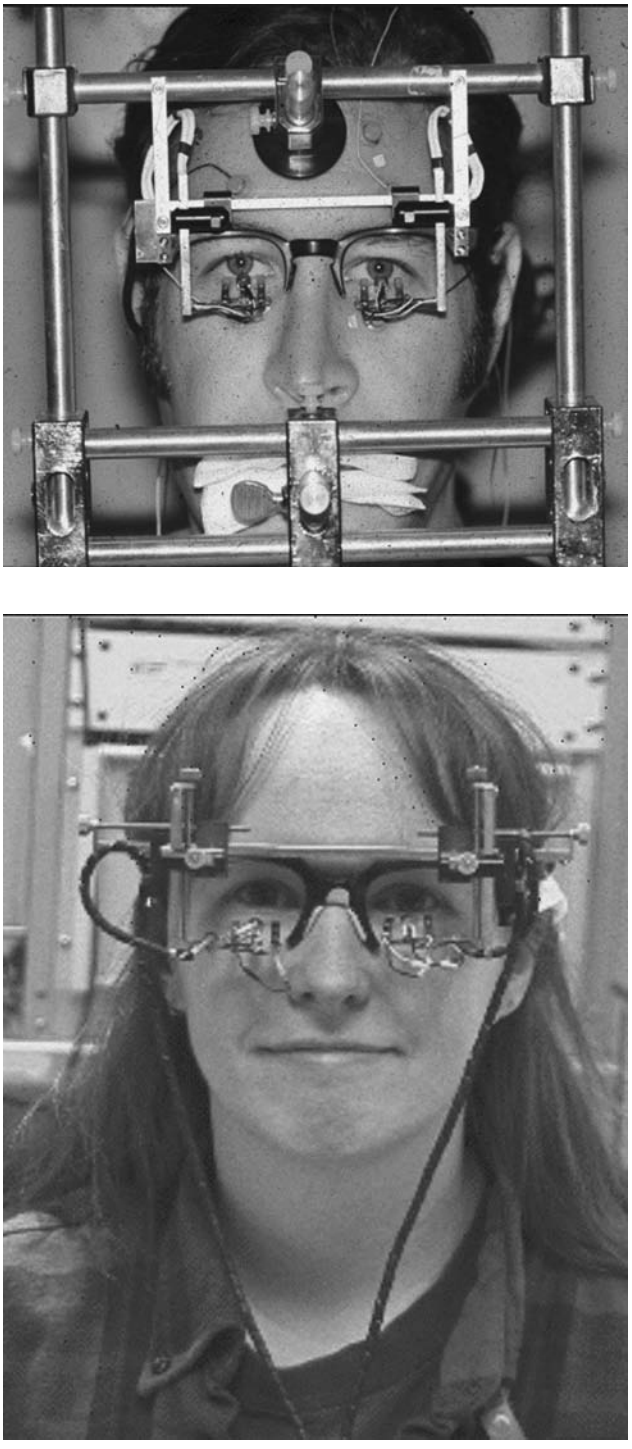


Figure 3. IR system to measure the horizontal eye movements of both eyes shown mounted on an earth-fixed frame (a) and spectacle frame (b) for human subjects and on a spectacle frame for a canine subject (c).

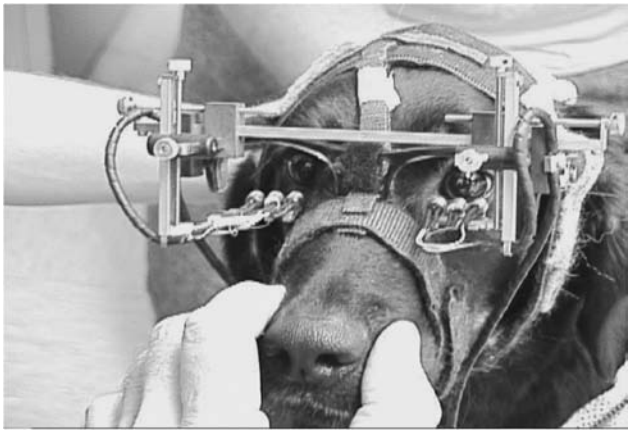


Figure 3. (*Continued*)

frame (b), and spectacle frame on a dog (c). Figure 4 shows an IR system mounted in goggles on a child (a) and a dog (b).

Characteristics. These systems offer a number of advantages over EOG, at least for the examination of horizontal eye movements. As the signal is not a biopotential, it is free of the instability found in the EOG; it is also immune to interference from muscle artifact and changes in electrode potentials. Unlike some earlier photographic methods, the device does not occlude the eyes, as the sensors and emitters are positioned above or below the eye. The field of view is somewhat obstructed by the emitter/detector, in contrast to EOG. Resolution is of the order of minutes of arc. Assuming that nothing disturbs the sensors, a shaken head or a rubbed eye, for example, stability is excellent. Thus, the question of using ac-coupling, as in many electronystagmographic applications of the EOG, never occurs. System bandwidth is generally on the order of 100 Hz, which is sufficient to capture fine details of saccades.

The linear range of these systems generally is between $\pm 15^\circ$ and 20° in the horizontal plane and half this amount or less in the vertical plane (which requires vertical orientation of the detectors or summation of the signals from horizontally-oriented detectors).

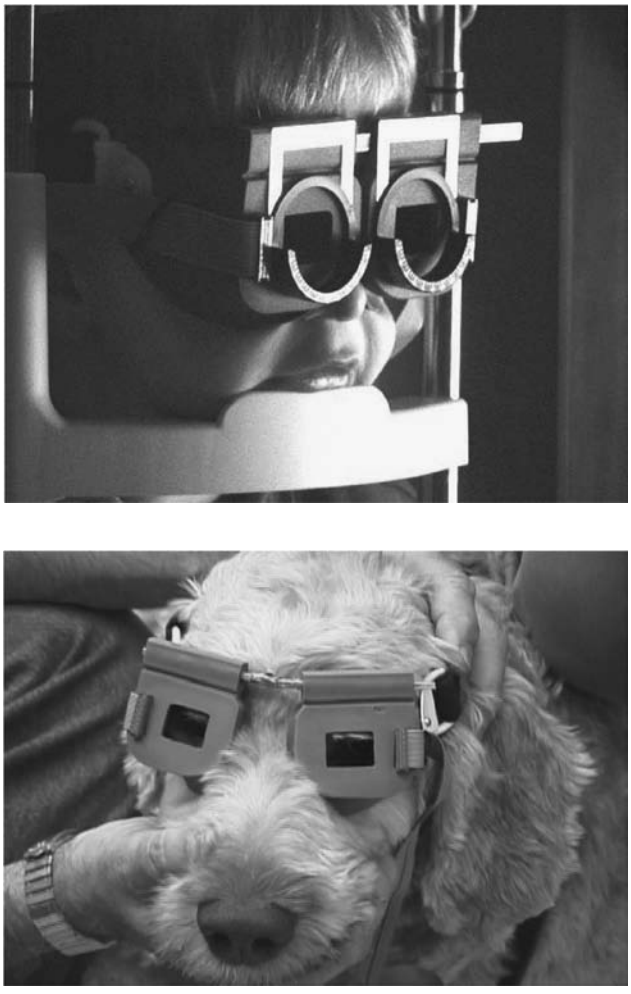


Figure 4. IR system to measure the horizontal and vertical eye movements of both eyes shown mounted in goggles for a human subject (a) and a canine subject (b).

Applications. IR limbus trackers are probably second only to EOG in their range of applications. Their ability to resolve fine detail with low noise makes them excellent for conditions where subtle features of the eye movement are important; examples include analyses of saccadic trajectories or analysis of small corrective saccades within a nystagmus waveform. An important advantage over EOG is that if eye velocity is to be calculated, the resulting signal is far less noisy than the derivative of an EOG recording, especially where broadband EMG noise has contaminated the signal developing from the eye. These systems are well suited to studies of any sort of eye movement that falls within their linear operating range in the horizontal plane. As they are generally head-mounted, they will tolerate modest head movement, but if the stimuli are fixed in the environment, such movement will certainly cause a loss of baseline and may move the tracker outside its linear range, which makes head stabilization highly desirable, especially when stimuli are presented at gaze angles where subjects would normally make both a head movement and an eye movement to acquire the target. Finally, IR systems are noninvasive, a major advantage for many patients and for children.

Limitations. One of the biggest shortcomings of these systems is their poor performance for vertical eye movement, their near-uselessness for oblique eye movements, and their complete lack of value for torsional eye movements. Although the limbus is clearly visible over a wide range of eye positions in the horizontal plane, the eyelids obscure its top and bottom margins. Although a degree of vertical tracking can be obtained by virtue of the differential reflectivity of the iris and pupil, the range over which this is possible is limited, again in part because of occlusion of the lids. Oblique movement suffers from inherent crosstalk because, as eye position changes in one plane, the sensitivity to motion in the other plane will vary, which is a hindrance to using these systems for studies of reading, scanpath analysis, or other applications where 2D eye movements are important. The use of the systems in rotational testing is also limited by the range of allowable gaze

angles and by the possible slippage of the head mounting on the head if accelerations are sufficiently high. Their suitability for small children also varies; some of the systems do not fit small heads well, although if precise calibration is not important, one can generally record patients as young as 3 years. These systems are not generally appropriate for use with infants. The one exception is for diagnosing nystagmus from its waveform by simply holding the sensors in front of the eyes, which can be done for even the smallest infants (e.g., a premature infant still in an incubator).

Scleral Search Coil.

Theory of Operation. Robinson developed the Scleral Search Coil technique in 1963 (16). It relies on the principle that a coil of wire in an alternating magnetic field induces a voltage proportional to the area of the coil, the number of turns, and the number of field lines. This latter measure will vary with the sine of the angle the coil makes with the magnetic field. In the basic configuration, two orthogonal pairs of field coils are used, each modulated by phase-locked square wave sources either operating in quadrature (i.e., one signal 90° phase-shifted relative to the other) (16) or at a different frequency (e.g., 50 and 75 kHz) (17). An annular contact lens with a very fine coil of wire is placed on the eye, so that it surrounds the cornea (or in animals, is surgically implanted under the conjunctiva). Figure 5 shows an annular search-coil contact lens on the eye of a subject. Components of the induced voltage generated by the horizontal and vertical signals can be separated via phase-sensitive detectors. Note that this method of recording horizontal and vertical components of eye movement eliminates the crosstalk present in 2D recordings made by limbus trackers. With an appropriately wound coil added to the lens, torsional eye movements may also be recorded. This technique is the only one able to record torsion with high bandwidth.

Figure 5. An annular search-coil contact lens used to measure the horizontal and vertical eye movements of a human subject. The fine wire from the imbedded coil exits at the nasal canthus.

Characteristics. This technology serves as the "gold standard" for eye-movement recording. Resolution is in seconds of arc and the linear range $\pm$ 20°, with linearization possible outside this range, because the nonlinearity follows the sine function. The signals are extremely stable, because their source is determined by the geometry of coil and magnetic field alone. In the usual configuration, the maximum angle that can be measured is 90°. Although the eyes cannot rotate this far in the head, if the head is also allowed to turn (and its position recorded by a head coil), a net change of eye position $>$ 90° is possible. A solution to this problem was developed whereby all the field coils were oriented vertically, generating a magnetic field whose vector rotates around 360°. Now, the phase of the field coil varies linearly over 360° of rotation (18,19), which is most often used for horizontal eye movements, with vertical and torsional eye movements recorded using the original Robinson design.

Applications. As the search-coil system provides such high quality data, it can be used in nearly any application where stability, bandwidth, and resolution are paramount and free motion by the subject is not essential. However, recent evidence suggests that the coils themselves may alter the eye movements being measured (20). Nonetheless, the low noise level and ability to independently record horizontal, vertical, and torsional movements at high bandwidth and high resolution still make this the gold standard of eye-movement recording techniques.

Limitations. As a result of their size, search-coil systems are clearly not suited for ambulatory studies or those carried out in other real-world settings such as a vehicle. The system also cannot be adapted to use in fMRI scanners, unlike IR limbus trackers or video-based systems. Search coils are invasive, making them unsuitable for some adult patients and for most children. A small risk of corneal abrasion exists when the coil is removed, but this risk is generally minor. Use of the coil in infants or small children would be undesirable, because they could not be instructed not to rub their eyes while the coil was in place. Another practical issue associated with the technology is the cost of the coils, which have a single supplier, have a limited lifetime, and are relatively expensive ($>$ US$100 each). As recommended duration of testing with the coils is 30 minutes or less, long duration studies are also precluded.

Digital Video.

Theory of Operation. Although electronic systems that locate and store the location of the center of the pupil in a video image of the eye were developed in the 1960s, often in combination with pupil diameter measurement (21,22), video-based eye trackers became a major force in eye-tracking technology when digital rather than analog image analysis was implemented. If the camera is rigidly fixed to the head, then simply tracking this centroid is sufficient to identify the location of the eye in its orbit. However, if there is even slight translational movement of the camera with respect to the eye, a large error is introduced: 1 mm of translation equals 10° of angular rotation in the image. For

this reason, video systems also track the specular reflection of a light source in the image in addition to the pupil centroid. As this first Purkinje image does not change with rotation but does change with translation, whereas the pupil center changes with eye rotation as well as translation, their relative positions can be used to compensate for errors induced by relative motion occurring between the head and camera. Figure 6 shows a digital video system in use on a human subject (a) and on dogs (b and c).

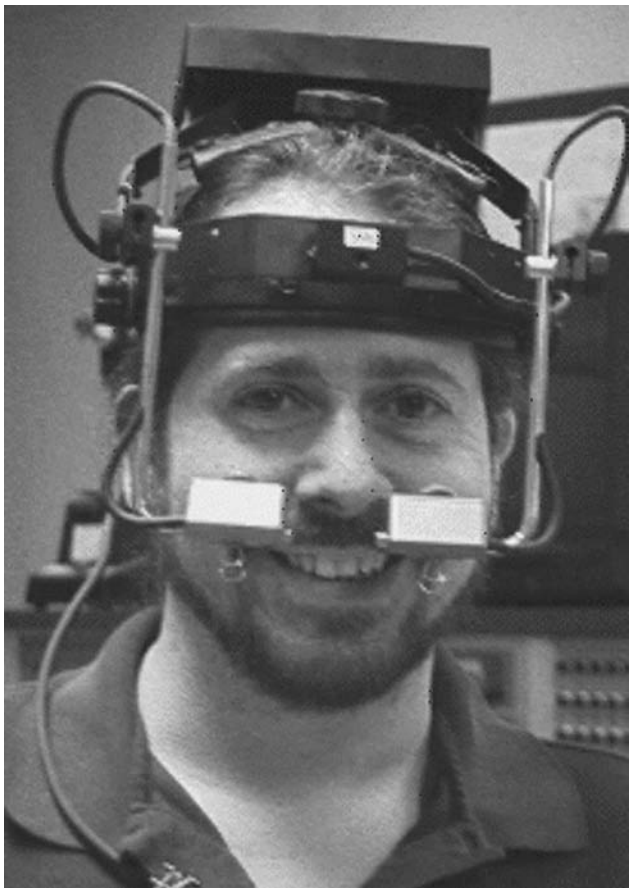


Figure 6. A high-speed digital video system to measure the horizontal and vertical eye movements of both eyes for a human subject (a) and canine subjects (b and c).

Figure 6. (*Continued*)

Characteristics. Assuming that the axes of the head and camera are aligned, then video-based systems are capable of recording both horizontal and vertical eye movements over a relatively wide range (often $\pm 30^\circ$ horizontally, somewhat less vertically). Resolution is better than EOG but generally somewhat less than for IR or search-coil systems, often in the range of 0.5°. As analog video systems use a raster scan to represent an image, spatial resolution is limited by the nature of the video system used (e.g., PAL or NTSC). Bandwidth is limited by the frame rate of the video system. If conventional analog video is used, then frame rates are 50 Hz for PAL and 60 Hz for NTSC. These rates impose a maximum bandwidth of 25 and 30 Hz, respectively. Although adequate for examination of slow eye movements, these frame rates are inadequate for assessment of saccades; indeed, very small saccades could be completed within the inter-frame interval. Systems using digital video are free from the constraints imposed by broadcast TV standards and can make use of higher frame rate cameras—several now operate at 250 or 500 Hz. Generally, a frame rate versus resolution trade-off exists—higher frame rates imply lower image resolution. However, continued improvement in digital video technology and ever faster and cheaper computers continue to improve performance.

Although older video tracking systems often required a good deal of "tweaking" of brightness and contrast settings in an effort to obtain a reliable image of the pupil, many recent systems have more streamlined set-up protocols. In the past, some systems internally monitored fixation on calibration targets and rejected data that were unstable, thereby making the systems unsuitable for use with patients with nystagmus. However, default calibration settings generally permit data to be taken and the nystagmus records can then be retrospectively calibrated.

Applications. In principle, digital video is the most flexible of all eye-movement recording technologies. Some systems use cameras mounted on the head, using either helmets or some other relatively stable mounting system. Other systems use remote cameras, often mounted adjacent to or within a computer stimulus display. Systems used in vehicles may use either remote cameras or

helmet-mount cameras. In addition to conventional clinical eye-movement testing, video systems, especially remote camera models, are increasingly being used in commercial applications such as advertising studies and usability analyses of websites. For such applications, the unobtrusiveness of the technology and the need to only monitor fixations rather than to study saccade dynamics makes even relatively low-frame-rate video ideal. Such systems are also excellent for use with infants and small children, who may be induced to look at some attractive display on a screen but who generally respond poorly to head-mounted apparatus. Remote systems that track more than one first Purkinje image can cope with a wider range of head movements, making the systems even less restrictive for the subjects. Some video systems can also analyze torsional eye movements by identifying some feature on the iris and then tracking changes in its orientation from frame to frame. High-speed (500 Hz) digital video systems are seeing increased use in basic and clinical laboratories, challenging magnetic search coils as the method of choice.

Limitations. The problems associated with calibrating patients whose eyes are never still have already been discussed. As noted before, the other serious limitations of some of these systems are their somewhat limited spatial resolution and bandwidth. Both parameters can be optimized, but doing so leads to marked increases in price. However, unlike other eye-tracking technologies, the limiting factors for high-speed, digital video eye-movement recording systems are the cameras and computing power. As the enormous general consumer market rather than the quite small eye-movement recording market drives improvements in both technologies, improvements can be anticipated to occur much faster than they would otherwise. Even within the eye-tracking field, the development of commercial uses for the technology will facilitate its advance faster than the smaller and less prosperous academic research community.

OCULAR MOTOR RECORDING TECHNIQUES

How Do We Record and Later Calibrate and Analyze Subjects' Eye Movements?

The initial recording and *post-hoc* calibration and analysis of eye movements require following a protocol conducive to both accurate calibration and obtaining the data specific to a particular study. Decades of experience have resulted in the following recording procedures and caveats and in the development of software that allows accurate calibration and linearization of the data.

Real-Time Monitoring. When recording subjects (especially patients), it is necessary to monitor the eye channels in real-time to ensure that the subject is following instructions, which is also imperative when calibrating subjects (see below). Unlike highly dedicated and motivated graduate students, most subjects quickly become bored by the task at hand or distracted and fail to fixate or pursue the stimuli; others may have difficulty doing so. Real-time monitoring via a strip chart or computer display allows the experimenter to detect and correct such failures with a simple verbal instruction encouraging the subject (e.g., "follow the target" or "look at the target").

Monocular Calibration. The key to obtaining accurate eye-movement data that will allow meaningful analysis is monocular calibration; that is, calibration of each eye independently while the other is behind cover. Too often, potentially accurate, commercially available recording systems are seriously compromised by built-in calibration techniques that erroneously presume conjugacy, even for so-called normal subjects. Just as bitemporal EOG makes it impossible to determine the position of either eye individually (see Fig. 2), so do calibration techniques carried out during binocular viewing of the stimuli. Most commercially available software calibration paradigms suffer from this fatal flaw, rendering them totally inappropriate for most clinical research and seriously compromising studies of presumed normal subjects. For methods that depend on subject responses to known target positions (e.g., IR or digital video), both the zero-position adjustment and gains at different gaze amplitudes in each direction must be calibrated for each eye during short intervals of imposed monocular fixation (i.e., the other eye occluded); for methods where precalibration is possible (e.g., magnetic search coils), the zero adjustment for each eye in each plane must also be made during imposed monocular fixation.

Linearization and Crosstalk Minimization. In addition to monocular calibration, linearization is required of most systems, even within the stated "linear" regions of those systems. As a result of different facial geometries and the inability to position the sensors in the precisely optimal positions for linearity, these systems are usually not linear over the range of gaze angles needed for many studies. System responses may be linearized by taking data during short intervals (5 s) of monocular fixation at all gaze angles of interest (e.g., 0°, $\pm 15^\circ$, $\pm 20^\circ$, $\pm 25^\circ$, and $\pm 30^\circ$) and applying post-recording linearization software. Even Robinson-type search coils need an arcsine correction for a linear response. For IR and video-based systems measuring eye motion in both the horizontal and vertical planes, crosstalk is a major problem due to sensor placement. Crosstalk can also be minimized post recording, using software written for that purpose. However, IR systems suffer from the additional problem that, as vertical eye position changes, a change may occur in the sensors' aim regions at the left and right limbal borders, which means that for a diagonal eye movement, the horizontal gain is an unknown function of vertical eye position, making IR systems essentially unsuitable for the recording of oblique eye movements.

All of the problems discussed above are accentuated when recording subjects with ocular motor oscillations, such as nystagmus. In these cases, the experimenter must be familiar with the type of nystagmus the subject has and be able to identify the portions of their waveforms that are used for target foveation. It is the "foveation periods" that are used to set the zero-position and gains at each target position; without them, accurate calibration is impossible.

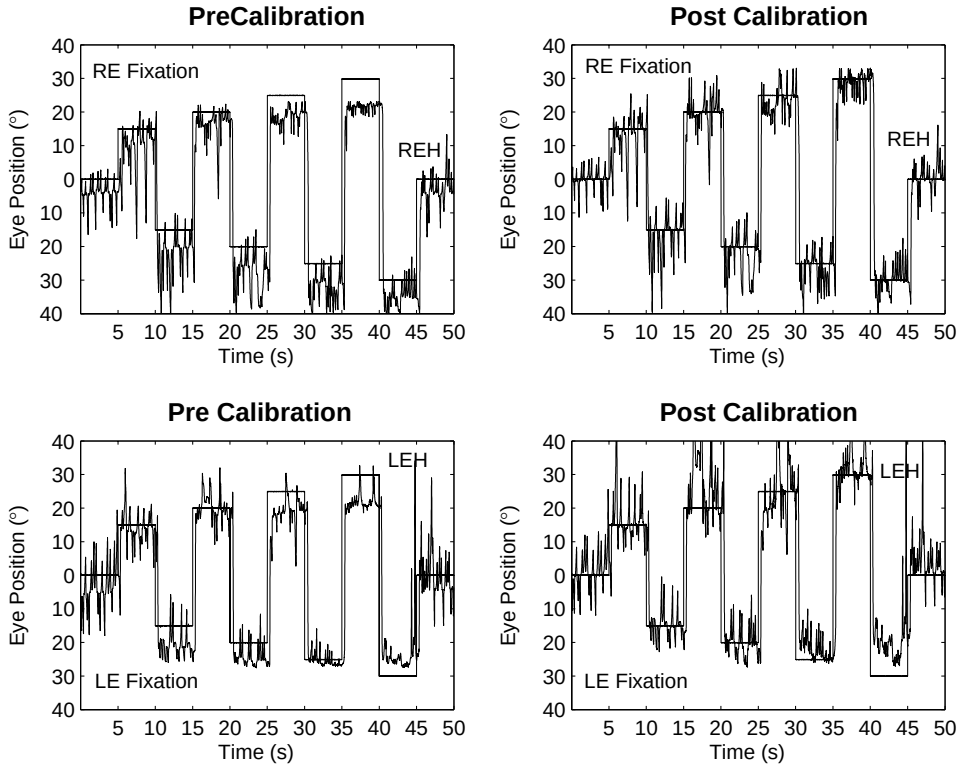


Figure 7. Monocular fixation precalibration and postcalibration (horizontal) records for the right (REH) and left (LEH) eye. Compare the offsets and nonlinear precalibration responses to the bias-adjusted, calibrated, and linearized postcalibration responses. Note the failure of the subject to look at the −30° target during LE fixation. In this figure and Fig. 8, target position is shown by the alternating direction, increasing offset, solid line.

The rest of the nystagmus waveform is irrelevant to target foveation and should be ignored during calibration. With a little practice, investigators can easily determine exactly where the subject with nystagmus is looking, which eye is fixating, and where the other eye is located with respect to the target; they can also determine periods of inattention by the associated waveform changes. Figure 7 demonstrates precalibration and postcalibration (horizontal) records of each eye made under imposed monocular fixation, and Fig. 8 shows the results of applying those calibration factors to a record made during binocular "viewing" of the targets. Note that the fixating eye is easily determined as well as the angle/position of the strabismic eye. Unfortunately, investigators with little or no experience in recording subjects with nystagmus are often reduced to using the average eye position during long periods of presumed binocular fixation to approximate calibration of subjects with nystagmus (and probably strabismus). Averaging anathema to accurate calibration and renders most potentially accurate recording systems (e.g., search coils) no better than bitemporal EOG. Needless to say, the results and conclusions of such studies must be highly suspect and are often incorrect; they exemplify how even the most sophisticated hardware and software can be misused, and prove the old adage, "garbage in, garbage out."

CONCLUSIONS

During the past 40 years, advances in eye-movement recording systems, coupled with the control-systems approach brought to the field by biomedical engineers, have resulted in an explosion of basic and clinical ocular motor research, at the systems as well as single-cell levels. Using the measurement systems and recording and calibration techniques described above, great strides have been made in our understanding of the ocular motor system. Animal studies have provided understanding at the single-cell and cell-network (bottom-up) levels, giving rise to computer models of small portions of the ocular motor system with neuroanatomical correlations. Normal human studies have allowed characterization of ocular motor behavior under a variety of stimulus conditions, giving rise to functional, top-down computer models of ocular motor behavior. Finally, studies of patients with many congenital and acquired ocular motor disorders have provided insights into the functional structure of the ocular motor system, which was not forthcoming from studies of normals (23,24). These latter studies have resulted in robust, behavioral models of the ocular motor system that are able to simulate normal responses and patient responses to a variety of ocular motor stimuli (25–27).

At present, accurate eye-movement recordings are an integral part of the diagnosis of both congenital and acquired forms of nystagmus, and of saccadic intrusions and oscillations. In addition, they provide objective measures of therapeutic efficacy that are related to visual function in patients afflicted with disorders producing ocular motor dysfunction. Indeed, ocular motor studies of the effects of a specific surgical procedure for congenital nystagmus produced an entirely new type of "nystagmus" surgery for both congenital and acquired nystagmus (28–31). This surgery (named "tenotomy") simply requires

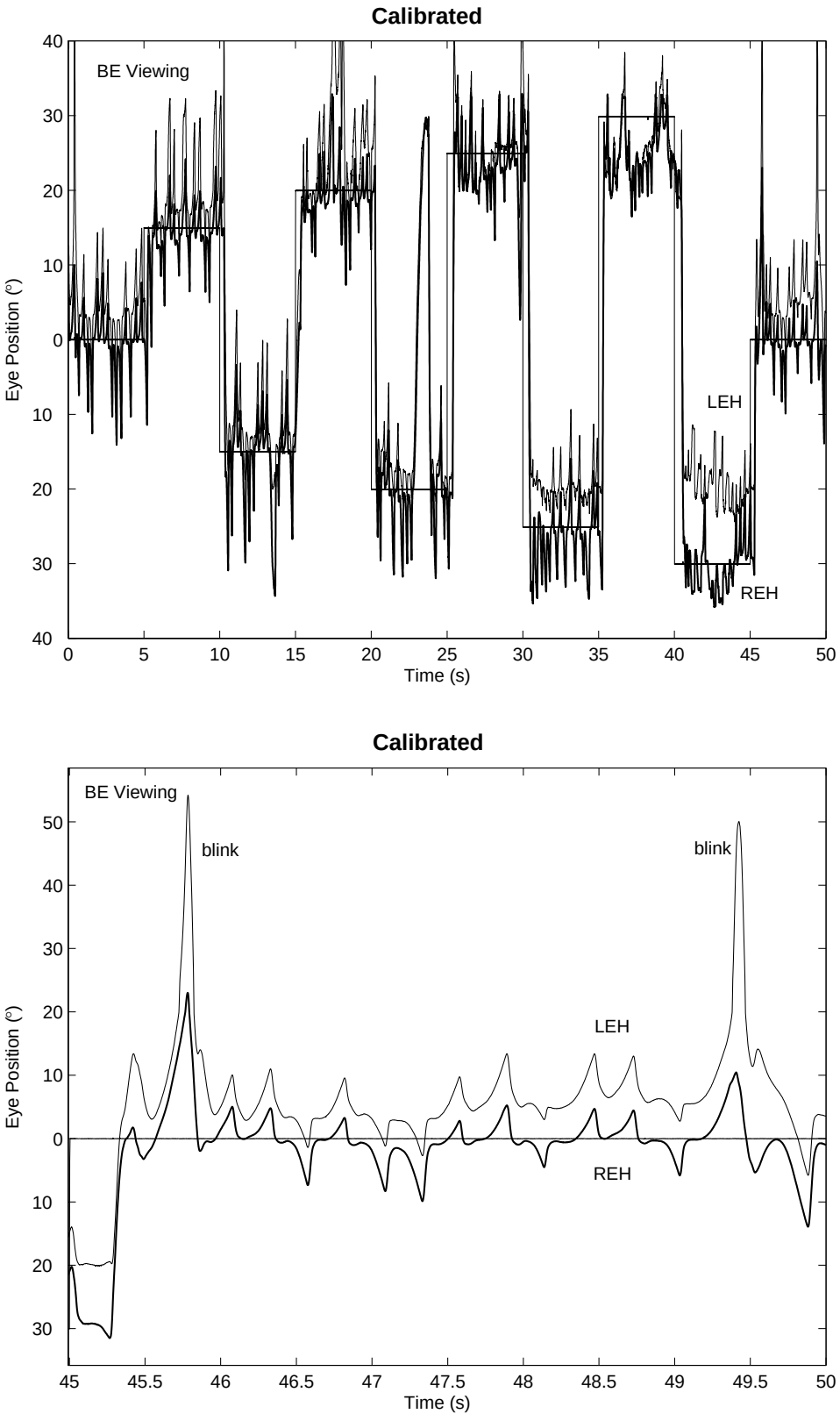


Figure 8. Calibrated binocular viewing records of both eyes (a) and final primary-position segment (b). Note the preference for RE fixation in left gaze and in the final primary-position segment, with the LE 3–5° esotropic. Note also how well the flat foveation periods of the RE line up on the 0° target despite the alternating direction of the nystagmus.

removal and reattaching, at their original insertion points, each of the four extraocular muscles in the plane of the nystagmus. Tenotomy represents a radical paradigm change from the "strabismus" surgeries that preceded it and has resulted in new insights into the anatomic structures responsible for proprioceptive signals from the extraocular muscles and their neurophysiologic role in the control of eye movements (32–34).

BIBLIOGRAPHY

1. Dodge R, Cline TS. The angle velocity of eye movements. Psychol Rev 1901;8:145–157.
2. Yarbus AL. Eye Movements and Vision. New York: Plenum Press; 1967.
3. Daffner KR, Scinto LF, Weintraub S, Guinessey JE, Mesulam MM. Diminished curiosity in patients with probable Alzheimer's disease as measured by exploratory eye movements. Neurology 1992;42:320–328.
4. Dell'Osso LF, Van der Steen J, Steinman RM, Collewijn H. Foveation dynamics in congenital nystagmus I: Fixation. Doc Ophthalmol 1992;79:1–23.
5. Dell'Osso LF, Van der Steen J, Steinman RM, Collewijn H. Foveation dynamics in congenital nystagmus II: Smooth pursuit. Doc Ophthalmol 1992;79:25–49.
6. Dell'Osso LF, Van der Steen J, Steinman RM, Collewijn H. Foveation dynamics in congenital nystagmus III: Vestibuloocular reflex. Doc Ophthalmol 1992;79:51–70.
7. Dell'Osso LF, Leigh RJ. Foveation period stability and oscillopsia suppression in congenital nystagmus. An hypothesis. Neuro Ophthalmol 1992;12:169–183.
8. Dell'Osso LF, Leigh RJ. Ocular motor stability of foveation periods. Required conditions for suppression of oscillopsia. Neuro Ophthalmol 1992;12:303–326.
9. Krauzlis RJ. Recasting the smooth pursuit eye movement system. J Neuriphysiol 2004;91:591–603.
10. Furman JM, O'Leary DP, Wolfe JW. Application of linear system analysis to the horizontal vestibulo-ocular reflex of the alert rhesus monkey using pseudorandom binary sequence frequency sinusoidal stimulation. Biol Cyber 1979;33:159–165.
11. Barry W, Jones GM. Influence of eyelid movement upon electro-oculographic recording of vertical eye movements. Aerospace Med 1965;36:855–858.
12. Torok N, Guillemin VJ, Barnothy JM. Photoelectric nystagmography. Ann Otol Rhinol Laryngol 1951;60:917–926.
13. Young LR. Measuring eye movements. Am J Med Electr 1963;2:300–307.
14. Kumar A, Krol G. Binocular infrared oculography. Laryngoscope 1992;102:367–378.
15. Reulen JP, Marcus JT, Koops D, de Vries FR, Tiesinga G, Boshuizen K, et al. Precise recording of eye movement: The IRIS technique. Part 1. Med Biol Eng Comput 1988;26:20–26.
16. Robinson DA. A method of measuring eye movement using a scleral search coil in a magnetic field. IEEE Trans Bio Med Electron 1963;BME(10):137–145.
17. Remmel RS. An inexpensive eye movement monitor using the scleral search coil technique. IEEE Trans Biomed Eng 1984;31:388–390.
18. Hartmann R, Klinke R. A method for measuring the angle of rotation (movements of body, head, eye in human subjects and experimental animals). Pflügers Archiv (Suppl); 362:R52.
19. Collewijn H. Eye movement recording. In: Carpenter RHS, Robson JG, eds. Vision Research: A Practical Guide to Laboratory Methods. Oxford: Oxford University Press; 1999.
20. Frens MA, van der Geest JN. Scleral search coils influence saccade dynamics. J Neurophysiol 2002;88:676–691.
21. Young LR. Recording eye position. In: Clynes M, Milsum JH, eds. Biomedical Engineering Systems. New York: McGraw-Hill; 1970. pp 1–2.
22. Marchant J. The oculometer: NASA, 1967. Report No. CR-805.
23. Abel LA, Dell'Osso LF, Daroff RB. Analog model for gaze-evoked nystagmus. IEEE Trans Biomed Eng 1978;BME(25):71–75.
24. Abel LA, Dell'Osso LF, Schmidt D, Daroff RB. Myasthenia gravis: Analogue computer model. Exp Neurol 1980;68:378–389.
25. Dell'Osso LF, Jacobs JB. A normal ocular motor system model that simulates the dual-mode fast phases of latent/manifest latent nystagmus. Biolog Cybernet 2001;85:459–471.
26. Dell'Osso LF. Nystagmus basics. Normal models that simulate dysfunction. In: Hung GK, Ciuffreda KJ, eds. Models of the Visual System. New York: Kluwer Academic/Plenum Publishers; 2002. pp 711–739.
27. Jacobs JB, Dell'Osso LF. Congenital nystagmus: Hypothesis for its genesis and complex waveforms within a behavioral ocular motor system model. JOV 2004;4(7):604–625.
28. Dell'Osso LF. Extraocular muscle tenotomy, dissection, and suture: A hypothetical therapy for congenital nystagmus. J Pediatr Ophthalmol Strab 1998;35:232–233.
29. Dell'Osso LF, Hertle RW, Williams RW, Jacobs JB. A new surgery for congenital nystagmus: Effects of tenotomy on an achiasmatic canine and the role of extraocular proprioception. JAAPOS 1999;3:166–182.
30. Hertle RW, Dell'Osso LF, FitzGibbon EJ, Yang D, Mellow SD. Horizontal rectus muscle tenotomy in patients with infantile nystagmus syndrome: A pilot study. JAAPOS 2004;8:539–548.
31. Hertle RW, Dell'Osso LF, FitzGibbon EJ, Thompson D, Yang D, Mellow SD. Horizontal rectus tenotomy in patients with congenital nystagmus. Results in 10 adults. Ophthalmology 2003;110:2097–2105.
32. Büttner-Ennever JA, Horn AKE, Scherberger H, D'Ascanio P. Motoneurons of twitch and non-twitch extraocular fibres in the abducens, trochlear and oculomotor nuclei of monkeys. J Comp Neurol 2001;438:318–335.
33. Büttner-Ennever JA, Horn AKE, Graf W, Ugolini G. Modern concepts of brainstem anatomy. From extraocular motoneurons to proprioceptive pathways. In: Kaminski HJ, Leigh RJ, eds. Neurobiology of Eye Movements. From Molecules to Behavior—Ann NY Acad Sci 956. New York: NYAS; 2002. pp 75–84.
34. Hertle RW, Chan C, Galita DA, Maybodi M, Crawford MA. Neuroanatomy of the extraocular muscle tendon enthesis in macaque, normal human and patients with congenital nystagmus. JAAPOS 2002;6:319–327.

See also ELECTRORETINOGRAPHY; EYE MOVEMENT, MEASUREMENT TECHNIQUES FOR.

OCULOGRAPHY.

See OCULAR MOTILITY RECORDING AND NYSTAGMUS.

OFFICE AUTOMATION SYSTEMS

JORGE CARDOSO
University of Madeira
Funchal, Portugal

INTRODUCTION

The purpose of this article is to help people in fields, such as healthcare, engineering, sales, manufacturing, consulting, and accounting to understand office automation systems

from the viewpoint of a business professional. This is important because personal association with office automation systems is almost unavoidable in today's business world. The widespread adoption of personal computers in conjunction with the development of graphically driven operating systems gave people a more natural and intuitive way of visualizing and manipulating information. The applications that were developed, from word processors to spreadsheets, to take benefit of these new operating systems, led to a growth in the use and acceptance of personal computers that significantly altered the manner organizations conduct their daily business.

Healthcare enterprises involve complex processes that span diverse groups and organizations. These processes involve clinical and administrative tasks, large volumes of data, and large numbers of patients and personnel. The tasks can be performed either by humans or by automated systems. In the latter case, the tasks are supported by a variety of software applications and information systems that are very often heterogeneous, autonomous, and distributed. The development of systems to manage and automate these processes has increasingly played an important role in improving the efficiency of healthcare enterprises.

Office Automation Systems (OAS) are computer-based automated information systems that are used to execute a variety of office operations, such as word processing, electronic spreadsheet, e-mail, and video conferencing. These different office automation systems allow the automation of much of the administrative work in the office and typically focuses on the more repeatable and predictable aspects of individual and group work. They are more and more frequently used by managers, engineers, and clerical employees to increase efficiency and productivity. They support the general activities of workers and underlie the automation of document-centric tasks performed by production office workers.

The OAS encompass a broad set of capabilities, and provide much of the technological basis for the electronic workplace. The focus of OAS have typically been used in supporting the information and communication needs of office workers, and its use by organizations supporting the white-collar work force has revealed itself crucial.

HISTORICAL PERSPECTIVE

In its early days, office automation systems focused on needs generally found in all offices, such as reading and writing. Before the 1950s, electromechanical and electronic devices were used to carry out financial and other numerical record-keeping tasks. During the evolution of OAS solutions, manual typewriters have been replaced by the electric typewriter and the electronic typewriter.

The electronic typewriter, introduced in the early 1970s, was the first of the automated office systems. It could store and retrieve information from memory providing automated functions such as center, bold, underline, and spell check.

The advances in the development of mainframes have caused electromechanical devices to be increasingly replaced by computers. In the 1970s, integrated circuit technology made the production of small and relatively inexpensive personal computers possible. Yet, even with this available technology, many computer companies chose not to adopt personal computers. They could not imagine why anyone would want a computer when typewriters and calculators were sufficient.

In the mid-1970s, computers began to support offices and organizations in more complex ways. The rapid growth of computers furnished the market with sophisticated office automation devices.

In the late 1970s, several researchers started to describe the needs of office automation systems. Computer terminals had replaced electronic typewriters. With the rapid evolution of electronic technology, office information systems were developed to provide for the storage, manipulation, computation, and transmission of large amounts of information. The first sophisticated OAS prototypes included the SCOOP project (1), which was oriented to the automation of office procedures, and Officetalk (2), which provided a visual electronic desktop metaphor, a set of personal productivity tools for manipulating information, and a network environment for sharing information.

In 1981, IBM introduced the IBM PC (Personal Computer). The PC was a milestone and proved that the computer industry was more than a trend, and that the computer was in fact a necessary tool for the business community. Computers, designed solely for word processing and financial tasks, became common. At first, the PC was utilized to replace traditional typewriters and calculators, but persistent technological advances and innovation over the past two decades have put powerful PCs at the center of daily activities for people worldwide.

The growth and widespread adoption of PCs, networks, graphical user interfaces, and communications as allowed the development of complete OAS package suites. For example, in 1985 the Lotus Notes (3) groupware platform was introduced. The term groupware refers to applications that enhance communication, collaboration, and coordination among groups of people. This system included online discussion, e-mail, phone books, and document databases. Throughout the years, continuous improvements were made to Lotus Notes. Nowadays, this system includes new features, such enterprise-class instant messaging, calendaring, and scheduling capabilities with a strong platform for collaborative applications.

In 1992, Microsoft lunched its new operating system (OS), Microsoft for Workgroups (4). This OS allowed the sending of electronic mail and provided advanced networking capabilities to be used as a client on existing networks. This was an important stage in the vast evolution of the world's most popular operating system since it enabled the collaboration of groups of people. Microsoft has also invested in the development of full OAS suites, which are commonly available nowadays. The most well-known and widespread productivity software suite is Microsoft Office (5). Microsoft Office helps workers to complete common business tasks, including word processing, e-mail, presentations, data management and analysis.

ORGANIZATIONAL INFORMATION SYSTEMS AND OAS

While we are interested in studying office automation systems (OAS), it is important to relate this type of systems with other information systems (IS) commonly used inside an organization. An information system can be defined as a set of interrelated components that retrieve, process, store and distribute information to support decision making and control in an organization. The main role of IS is to assist workers in managing resources and information at each level in the organization.

This article, is primarily concerned with OAS and how they can be used in the medical community. For completeness, some other types of information systems commonly used by organizations are also mentioned. There will be no description of how such systems are developed, however, a brief description of their objectives will be given. Organizational information systems (OIS) are systems that support several functions in an organization and can be classified by the activity that they support. The OIS are usually split into six major types: Transaction Processing System, Knowledge Work Systems, Office Automation System, Management Information System, Decision Support System, and Executive Information System. These systems are illustrated in Fig. 1.

It is important to be able to distinguish the objectives and the level in the organization where a particular application or system can be used. For example, Transaction Processing Systems are employed to records daily routine transactions to produce information for other systems, while Office Automation Systems are oriented to increase that productivity of data workers using applications such as word processing and electronic mail applications.

Transaction Processing System (TPS): Is useful for daily transactions that are essential to the organization such as order processing, payroll, accounting, manufacturing, and record keeping.

Office Automation System (OAS): Aids office workers in the handling and management of documents, schedules, e-mails, conferences and communications. Data workers process information rather than create information and are primarily involved in information use, manipulation or dissemination.

Knowledge Work System (KWS): Promotes the creation of new information and knowledge and its dissemination and integration within the organization. In general, knowledge workers hold professional qualifications (e.g., engineers, managers, lawyers, analysts).

Management Information System (MIS): Provides middle-level managers with reports containing the basic operations of the organization which are generated by the underlying TPS. Typically, these systems focus on internal events, providing the information for short-term planning and decision making.

Decision Support System (DSS): Focuses on helping managers to make decisions from semistructured or unstructured information. These systems use internal information from TPS and MIS, but also information from external data sources, providing tools to support 'what-if' scenarios.

Executive Information System (EIS): Supports senior and top-level managers. They incorporate data from internal and external events, such as new legislation, tax laws, and summarized information from the

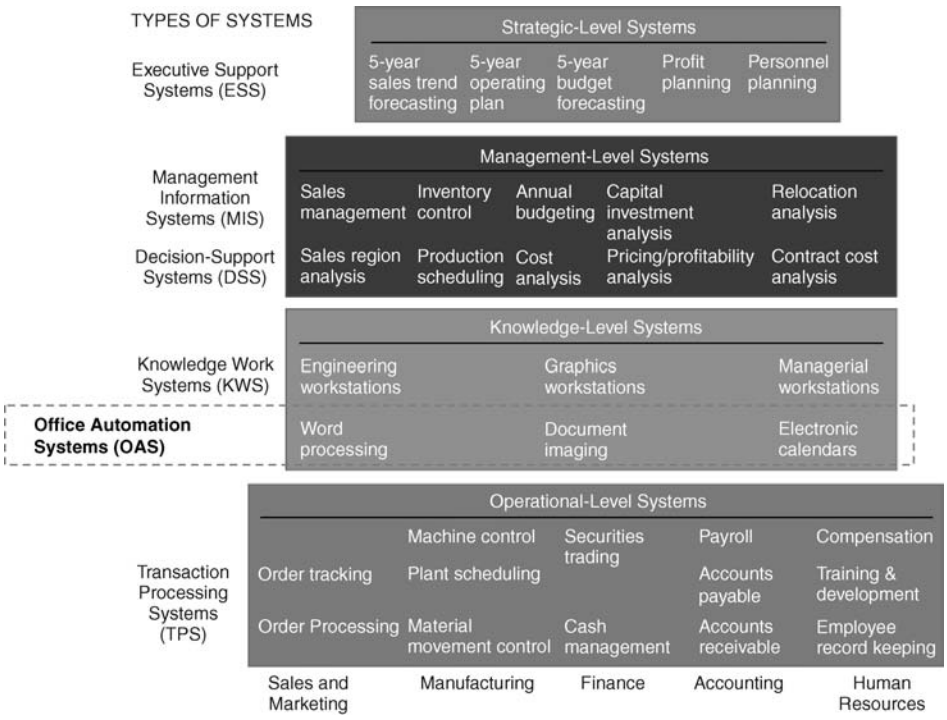


Figure 1. Types of information systems (6).

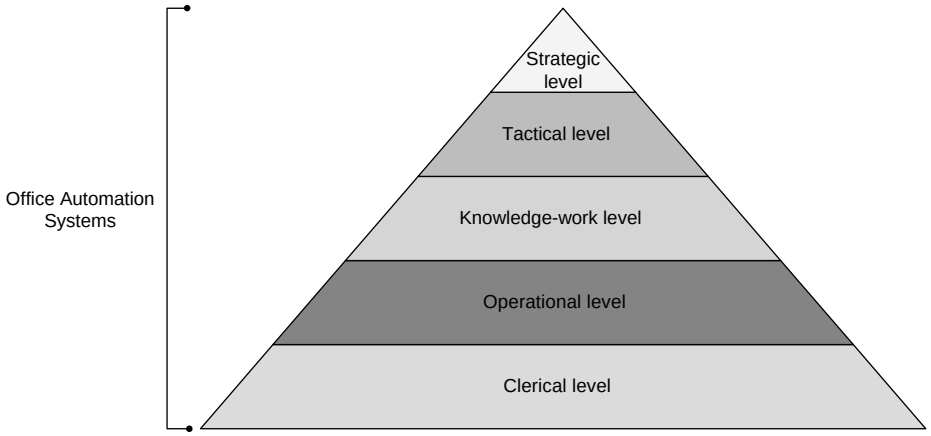


Figure 2. Support of OIS to the different organizational levels.

internal MIS and DSS. EIS software displays data graphically to provide easy-to-use representations of complex information.

Office information systems can also be classified by the organizational level they support. The human resources of an organization work in different areas and levels of operations, are in charge of different functions, and use different OIS. Any organization can be viewed as a multilevel entity with each level representing a different level of control. The levels of an organization can be arranged in a pyramid (Fig. 2).

The pyramid is divided into five horizontal sections:

Clerical level: Employees who support managers at all levels of the organization.

Operational level: First-line managers who make routine decisions and deal with the day-to-day operations of the organization.

Knowledge-work level: Advisors to both top and middle management who are often experts in a particular area.

Tactical level: Middle managers who deal with planning, organizing and the control of the organization.

Strategic level: Strategic managers who make decisions that guide the manner in which business is done.

Each successively lower level has different OIS requirements and a different, and less extensive, view of the organization. Obviously, the higher the level, the more interrelated the business functions become until, at the very top, they are viewed as one homogeneous organization with one continuous data flow.

Office automation systems can be effectively utilized in all the clerical, operational, knowledge-work, tactical, and strategic levels, as illustrated in Fig. 2. They can assist workers who work with word processors, electronic mail, and spreadsheets to use, manipulate, disseminate information, and help managers in planning, organizing, control and taking decisions.

OFFICE AUTOMATION SYSTEMS

Typical office automation systems handle and manage documents through word processing, desktop publishing, document imaging, and digital filing, scheduling through electronic calendars, and communication through electronic mail, voice mail, or video conferencing. In this section, 18 different types of OIS are discussed and described that are classify into four categories: productivity tools, digital communication systems, groupware applications, and teleconferencing systems (Fig. 3).

Productivity Tools

Productivity tools are software programs used to create an end product, such as letters, e-mails, brochures, or images. The most easily recognized tool is a word processing program, such as Microsoft Word (7) or Corel WordPerfect (8).

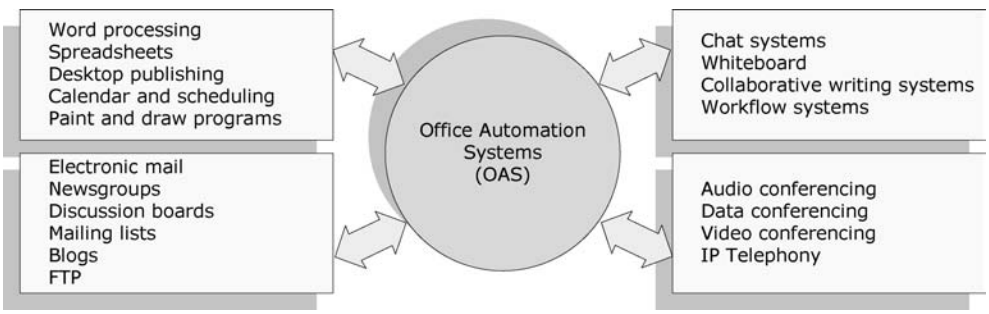


Figure 3. Office information systems.

Other tools help you view, create and modify general office documents such as letters, spreadsheets, memos, presentations, and images.

Word Processing. Of all computer applications, word processing is the most common. Almost every computer has a word processing program of some kind: whether it came free with the operating system or whether it was purchased separately.

In order to perform word processing, it is necessary to acquire a computer, a word processor, and a printer. A word processor enables you to create a document, store it, display it on the computer screen, modify it, and print it using a printer. There are many different word processing programs available, each offering different tools that make it easier to write everything from letters and term papers to theses and Web pages.

Most people use a word processor rather than a typewriter because it allows greater flexibility and control. It is possible to make changes without retyping the entire document. If mistakes are made while typing a text, the cursor can be used to correct errors. Word processors allow text rearranging, changing the layout, formatting the text, and inserting pictures, tables, and charts.

Most word processors available today allow more than just creating and editing documents. They have a wide range of other tools and functions, which are used in formatting documents. The following are the main features of word processors:

Insert, delete, copy, cut, and paste text: Allow to insert, erase, and copy text anywhere in the document. Cut and paste allow removing (cut) a section of text from one place and inserting (paste) it somewhere else in the document.

Search and replace: Allow searching for a particular word and also replacing groups of characters.

Font specifications: Allow to change fonts within a document. For example, you can specify bold, italics, font size and underlining.

Graphics: Allow adding pictures into a document.

Captions and cross-references: Allow placing captions to describe tables and pictures and creating references to them anywhere in the document.

Page setup, headers, and footers: Margins and page length can be adjusted as desired. Allow to specify customized headers and footers that the word processor will put at the top and bottom of every page.

Layout: Allows specifying different margins within a single document and to specify various methods for indenting paragraphs.

Spell checker and thesaurus: Spelling can be checked and modified through the spell check facility. The thesaurus allows the search for synonyms.

Tables of contents and indexes: Allow creating table of contents and indexing.

Print: Allows sending a document to a printer to get a hardcopy.

Spreadsheet. A spreadsheet is a computer program that presents data, such as numbers and text, in a grid of rows and columns. This grid is referred to as a worksheet. You can define what type of data is in each cell and how different cells depend on one another. The relationships between cells are called formulas, and the names of the cells are called labels.

There are a number of spreadsheet applications on the market, Lotus 1-2-3 (9) and Microsoft Excel (10) being among the most famous. In Excel, spreadsheets are referred to as workbooks and a workbook can contain several worksheets. An example of an Excel worksheet is shown in Fig. 4.

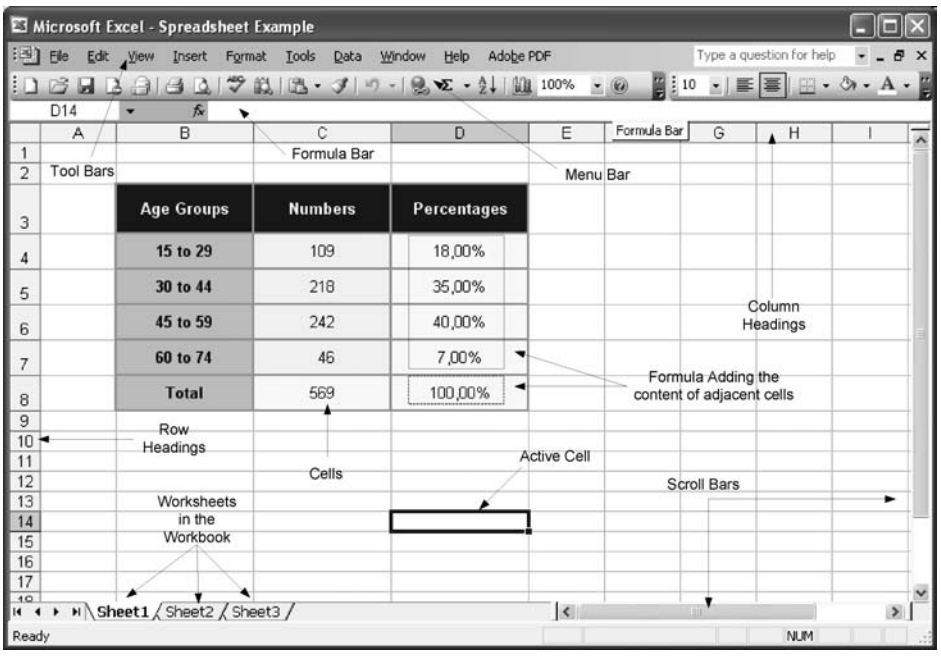


Figure 4. Microsoft Excel spreadsheet program.

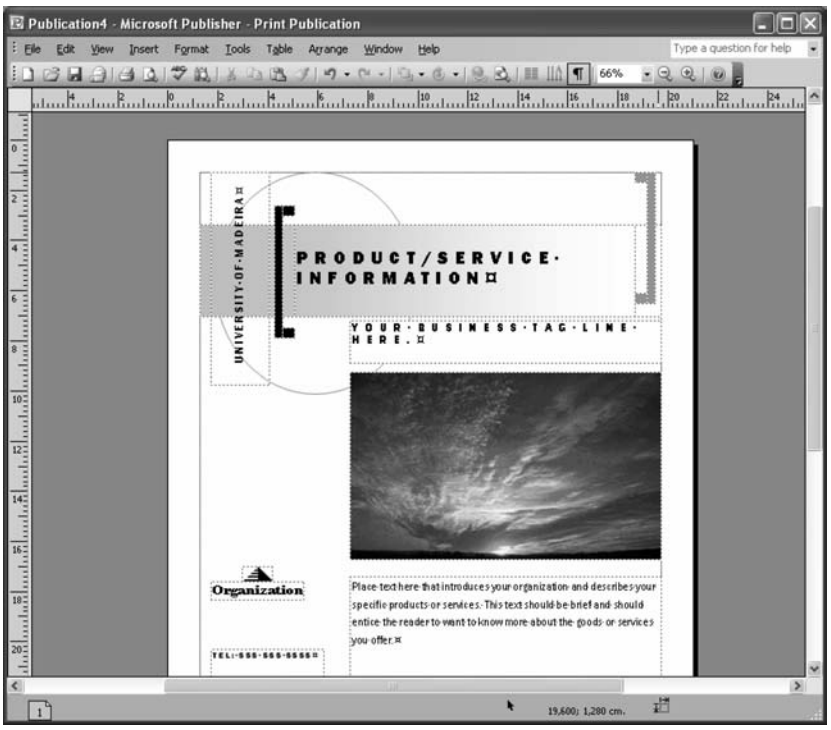


Figure 5. Editing a document with Microsoft Publisher.

Desktop Publishing. Desktop publishing is the use of the computer and specialized software to create high quality documents for desktop or commercial printing. Desktop publishing is the process of editing and layout of printed material intended for publication, such as books, magazines, brochures, and flyers using a personal computer.

Desktop publishing started in 1985, with the commercialization of the software Aldus PageMaker (11) (now from Abode). Nowadays, there are many software programs available for desktop publishing. QuarkXPress (12), Adobe InDesign (13), Abobe PageMaker (11), and Microsoft Publisher (14) are the most widespread. Figure 5 shows a document being created and edited with Microsoft Publisher.

As word processing programs become more sophisticated, the line separating such programs from desktop publishing systems is becoming fuzzy. Cutting-edge word processing programs give you most of the features you could want in a desktop publishing program. Such programs do not generally replace word processors and graphic applications, but are used to aggregate the text and graphic content created in these programs. The most powerful desktop publishing systems enable the creation of illustrations; while less powerful systems let you insert illustrations created by other programs.

Initial desktop publishing solutions were expensive due to the cost of specialized computing systems and accessories, such as printers and scanners. The cost of computers and printers has fallen dramatically in recent years (e.g., inkjet printers are amazingly inexpensive and most can print in color), allowing most personal users to acquire desktop publishing systems.

Calendars and Schedulers. A calendar program enables us to record events and appointments on an electronic calendar. Calendars allow scheduling, project management, and coordination among many people, and may provide support for scheduling equipment as well. Typical features identify conflicts in schedules, find meeting times that will work for everyone, signal upcoming events, and automatically fill in entries for regular events.

A special type of calendar, called a scheduler, is a solution to manage the daily scheduling needs of a business, such as scheduling appointments, equipment, staff (technicians, professionals, healthcare workers, others), vehicles, resources, projects, and meeting rooms. Scheduling software is an important investment for any type of business that wants to improve its scheduling processes. Every employee can have instant access to whom or what is available at any time of the day, week, month, or year and print detailed list reports. It is also possible to export schedules that may be easily opened in a word processor or spreadsheet.

Paint and Draw Program. A paint program or a graphics program enables the creation of pictures, backgrounds, buttons, lines, and other creative art. Paint programs provide easy ways to draw common shapes, such as straight lines, rectangles, circles, and ovals. Some programs also have photoediting capabilities and are optimized for working with specific kinds of images, such as photographs, but most of the smaller paint programs do not have this option. Paint programs are pixel based. They use "raster" images made up of small dots called pixels. As each dot is an individual, it can be difficult to move shapes around the screen.

A draw program is different from a paint program. Draw programs are object based, where an object is a geometrical shape, such as a line, a circle, a curve, a rectangle, a polygon, or a Bezier curve (curves that have hooks along their length so you can alter the angle of the curve at any point.) With draw programs, images are stored as mathematical information in the form of vectors for the lines and curves of each shape. Sophisticated programs often blur the difference between draw and paint, so it is possible to find programs that are able to do both types of work.

Digital Communication Systems

Nowadays, more and more computers are not isolated but, instead, are connected into a computer network that is often connected to other computer networks in much the same way as telephone systems connect telephones. If a computer is connected to such a network, it is possible to communicate with people whose computers are connected to the same network.

Electronic Mail. Electronic mail, or e-mail for short (another common spelling for e-mail is email), is one of the most popular uses of the Internet. It is a simple tool for exchanging brief messages between individuals or among a larger audience. Most mainframes, minicomputers, and computer networks have an e-mail system.

An e-mail address identifies a person and the computer for purposes of exchanging electronic mail messages. It consists of two parts: user name and mail domain or domain name. The user name identifies a particular person. The mail domain identifies the place on the Internet to which the e-mail for that person should be sent. An e-mail address is read from left to right. An example is illustrated in Fig. 6.

With an e-mail account, it is possible to send a message to anyone with an e-mail account. Just as a written letter can be sent to multiple recipients, an electronic mail message can be sent to one or more e-mail addresses. An e-mail can be broken down into several basic fields that include 'From', 'To', and 'Cc'. The 'From' field contains the address of the sender of the message. The 'To' field indicates the addresses of one or more recipients who are the primary audience. All recipients can see every address listed in this field. Finally, the 'Cc' field (Cc - Carbon Copy) contains the addresses of recipients how are not the primary audience for the e-mail.

An electronic mail message is not limited to text. Other types of files can be added to mail messages as attachments. Attachments can be text files or binary files such as word processed documents, spreadsheets, images, files of sound and video, and software. To see if you have any e-mail, you can check your electronic mailbox periodically, although many programs can be configured to alert users automatically when mail is received. After reading an e-mail, it may be stored, deleted, replied to, forwarded to others, or printed.

One of the serious problems with reading e-mail on a PC computer running Windows operating system is that the computer can become infected with an e-mail virus program. It is always advisable to install and use anti-virus software. Such software will offer protection against known malicious programs. A malicious program may be a virus, a worm, a trojan horse, or a spyware. Once it is on your system, a malicious program cause disorder by corrupting, erasing, attaching to, or overwriting other files. In some cases malicious program, such as spyware, have the solely intent of monitoring Internet usage and delivering targeted advertising to the affected system. Unexpected e-mail attachments should not be opened since they are one of the most common ways for computer viruses to spread.

Newsgroups and Discussion Boards. Newsgroups, also known as Usenet, are comparable in essence to e-mail systems except that they are intended to disseminate messages among large groups of people instead of one-to-one communication (Fig. 7).

A newsgroup is a collection of messages posted by individuals to a news server. The concept of newsgroups was started in 1979 at the University of North Carolina and Duke University to create a place, where anyone could post messages.

Although some newsgroups are moderated, most are not. Moderated newsgroups are monitored by an individual (the moderator) who has the authority to block messages considered inappropriate. Therefore, moderated newsgroups have less spam than unmoderated ones. Anyone who has access to the board of a newsgroup can read and reply to a message that, in turn, will be read and replied to by anyone else who accesses it. If you have an interest in a certain topic, chances are it has its own newsgroup. A few examples of newsgroups are shown in Table 1.

Discussion boards (also called message boards) and newsgroups in general both accomplish the same task. They each have general topics, and visitors can post messages about specific topics. Discussion boards are usually read through a web browser, while newsgroups are usually read through a special program called a newsgroup reader. Nowadays, most people prefer discussion boards on the Web to newsgroups because they are easier to use.

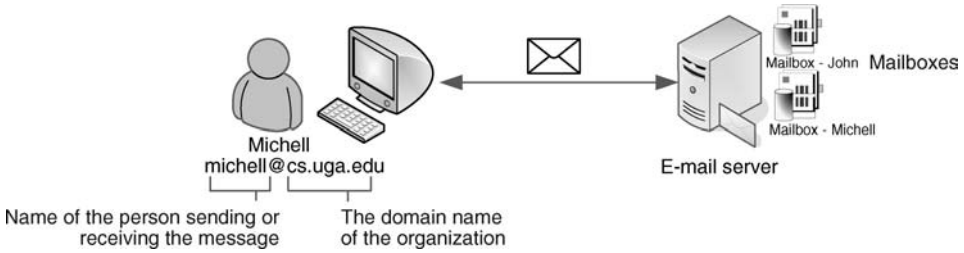


Figure 6. E-mail address structure.

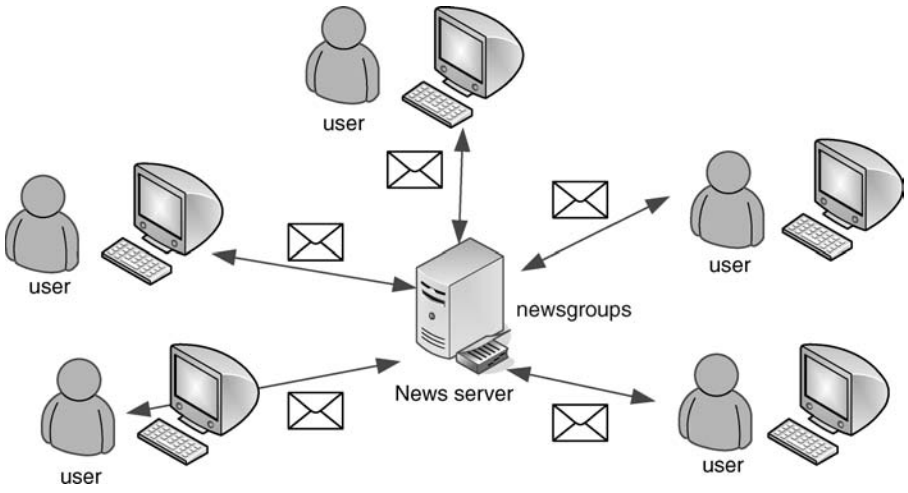


Figure 7. Newsgroup architecture.

Mailing Lists. The main difference between newsgroups and mailing lists is that newsgroups only show messages to a user when they are explicitly requested, while mailing lists deliver messages as they become available. Mailing lists are e-mail addresses that are used to distribute e-mail to many people. Typically, a list subscriber sends a message to the list address, and the message is then distributed to all the list subscribers for everyone to read.

Mailing lists are simple in operation. The first thing to do is to subscribe to a particular list; afterward the user can send messages to the mail server. The following steps are involved: (*1*) send a message (e-mail) to a mail server; (*2*) the mail server sends the message to everyone who is subscribed to the list; and (*3*) if someone replies to the message, then their reply goes to the mail server and is disseminated to everyone on the list.

Blogs. A weblog, or "blog", is a personal journal on the Web, although it can also be owned by a small group. The blog owner periodically writes entries and publishes them onto their blog. Weblogs cover as many different topics and express as many opinions, as there are people writing them.

Table 1. Examples of Newsgroups

Newgroup name	Description
comp.ai	Artificial intelligence discussions
sci.cognitive	Perception, memory, judgment and reasoning
comp.groupware	Hardware & software for facilitating group interaction
comp.multimedia	Interactive multimedia technologies of all kinds
comp.infosystems	Any discussion about information systems
comp.graphics	Computer graphics, art, animation, image processing
alt.comp.blind-users	Discussion of the needs of blind users
comp.windows.misc	General issues regarding the use of windows

A blog is used to show an up-to-date view of the owner's work, ideas, and activities. It provides a continuous record of activities, progress, and development. This type of systems can be effectively used by the healthcare community to discuss specific topics of interest. Examples of blog topics include product reviews, scientific endeavors, and any area of information where people have a deep expertise and a desire to express it. The power of blogs is that they are a fluid and dynamic medium that allow several people to easily publish their ideas and opinions, and allow other people to comment on them.

File Transfer Protocol. The ability to share information throughout organizations is essential in today's business environment. With the explosion of content creation and information in electronic formats, there is simply more electronic data today than ever before. File Transfer Protocol (FTP) is a standard method for sending files from one computer to another over networks, such as the Internet. Applications allow sharing and managing data between multiple remote, local, and home folders. It provides the ability to seamlessly work from a healthcare facility, a remote office, or home and is most commonly used to download a file from a server or to upload a file to a server.

Groupware Systems

Groupware refers to any computer-related tool that improves the effectiveness of person-to-person communication and collaboration. It is intended to create an environment that fosters the communication and coordination among a group of people. Where a traditional user interface generally focuses on the use of only one person, groupware relates to groups and understanding how people work and function in a group.

The groupware concept takes various applications and functionalities under the umbrella of communication and collaboration and integrates them together as a single client application. Groupware systems generally include some of the following systems: chat systems, whiteboarding, collaborative writing, workflow systems, and hypertext linking. Groupware packages are diverse in the

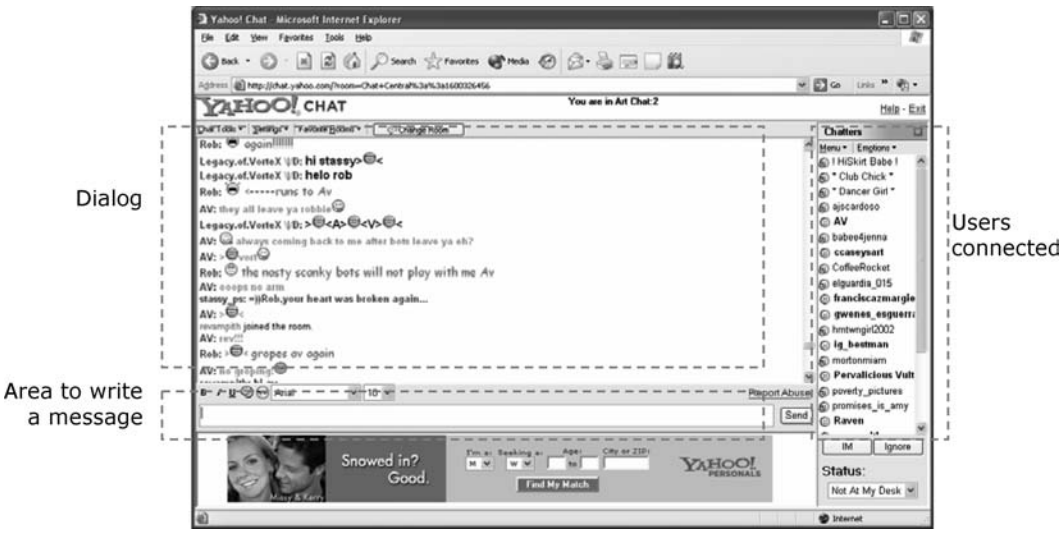


Figure 8. Web-based chat system.

functions they offer. Some include group writing, chat and/or e-mail. Sophisticated workgroup systems allow users to define workflows so that data is automatically forwarded to appropriate people at each stage of a process.

Chat Systems. Chat systems enable a type of group communication in which people located in different geographical locations get together in a virtual room and interact with each other by typing text. Chat systems make it possible for many people to write messages in a public space or virtual room. As each person submits a message, it appears on the screen of the other users located in the same virtual room. Chat groups are usually formed via listing chat rooms by name, location, number of people, topic of discussion, and so on.

Recently, systems accessible on the World Wide Web became widely spread among chat users. These types of chat systems are referred to as Web-based chat because they are accessible using a typical browser. One example of Web-based chat can be found at Yahoo.com (see Fig. 8).

Compared to e-mail, a chat system is a real-time synchronous system, while e-mail is neither real-time nor synchronous. When a user types a comment in a chat system, it is seen almost immediately by the others users present in the same virtual room. All the users are connected to the system at the same time. With e-mail, on the other hand, the two parties involved in the exchange of a message do not need to be connected to the system at the same time. For example, when reading an e-mail message the person who writes it may or may not be sitting in front of their computer at that time.

Whiteboard. A whiteboard provides real-time communication over the Internet and has a visual or graphical component in addition to text-based communication. Using a whiteboard, multiple users can simultaneously review, create, and update documents, images, graphs, equations, text, and information. All changes made by one user to the whiteboard area are displayed to all the other whiteboard users. The whiteboard allows participants to manipulate the contents by clicking and dragging with the mouse. In addition, they can use a remote pointer or highlighting tool to point out specific contents or sections of shared pages.

Most whiteboards are designed for informal conversation, but they may also serve structured communications or more sophisticated drawing tasks, such as collaborative graphic design, publishing, or engineering applications. For example, executives can meet and collaborate on slides for a presentation and architects can revise building plans.

Collaborative Writing Systems. Collaborative writing systems are applications that aim to help the joint editing of text documents by several authors. Coauthors, spread out across different network locations, can work together sharing common documents. When the interactions happen at the same time, they are called synchronous or real-time interactions. Otherwise, they are called asynchronous or non-real-time interactions.

Word processors may provide asynchronous support by showing authorship and by allowing users to track changes and make annotations to documents. It is possible to determine that only certain sections of documents may be modified by specific people to better protect how documents are modified and reduce the number of conflicting comments received. Reviewers can be prevented from making changes unless they turn revision marks on.

Workflow Systems. Workflow management systems (WfMS) appeared in the 1980s, but there is some consensus that the office information systems field is the predecessor of workflow systems (15). Advances in transaction processing and integrated office systems made workflow systems popular in the 1990s. They were innovative and had gained a high level of popularity. Commercial products include IBM MQSeries Workflow, Staffware, TIBCO InConcert, and COSA Workflow. General information on WfMSs can be found at the web sites of the Workflow and Reengineering International Association (16) and the Workflow Management Coalition (17).

A WfMS is implemented in accordance with a business process specification and execution paradigm. Under a WfMS, a workflow model is first created to specify organizational business processes, and then workflow instances are created to carry out the actual steps described in the workflow model. During the workflow execution, the workflow instances can access legacy systems, databases, applications, and can interact with users.

Workflow systems have been installed and deployed successfully in a wide spectrum of organizations. Most workflow management systems, both products and research prototypes, are rather monolithic and aim at providing fully fledged support for the widest possible application spectrum. The same workflow infrastructure can be deployed in various domains, such as bioinformatics, healthcare, telecommunications, military, and school administration.

In Fig. 9, a workflow process from the field of genomics exemplifies how workflow systems can be used to design business processes.

A major task in genomics is determining the complete set of instructions for making an organism. Genome projects are very demanding, and incur high costs of skilled manpower. There are many different types of tasks that must be performed, such as sequencing, sequence finishing, sequence processing, data annotation, and data submission. A single genomic workflow may be spread across multiple research centers, and the individual tasks in a workflow may be carried out at one or more of the participating centers. Many of the challenges of building an information system to manage a physically distributed genome project can be addressed by a workflow system.

The workflow model for such a workflow graphically specifies the control and data flow among tasks. For example, the workflow model in Fig. 9 is composed of several tasks and subworkflows. The tasks illustrated with machine gears represent automatic tasks, while the ones illustrated with boxes represent subworkflows.

At runtime, the workflow system reads the model specifications and transparently schedules task executions, providing the right data at the right time to the right worker. It manages distributed genomic tasks located at different research centers, such as DNA sequencing machines, matching algorithms, and human resources. Further, the workflow system provides a framework to easily reengineer a genomic workflow when new technological, biological, and chemical advances are made.

Teleconferencing

The term teleconferencing refers to a number of technologies that allow communication and collaboration among people located at different sites. At its simplest, a teleconference can be an audio conference with one or both ends of the conference sharing a speakerphone. With considerably more equipment and special arrangements, a teleconference can be a conference, called a videoconference, in which the participants can see still or motion video images of each other. Using teleconferencing systems, organizations can decrease costs and complexity, while increasing efficiency and productivity.

Audio Conferencing. Audio conferencing is the interaction between groups of people in two or more sites in real time using high quality, mobile, hands-free telephone technology. The interaction is possible with an audio connection via a telephone or network connection. It makes use of conventional communication networks such as POTS (Plain Old Telephone Service), ISDN (Integrated Services Digital Network), and the Internet.

Data Conferencing. Data conferencing is the connection of two or more computer systems, allowing remote groups to view, share, and collaborate on prepared documents or information. Data conferencing platforms make it possible to share applications and files with people in other locations. Everyone can see the same document at the same time and instantly view any changes made to it.

A user can share any program running on one computer with other participants in a conference. Participants can watch as the person sharing the program works, or the

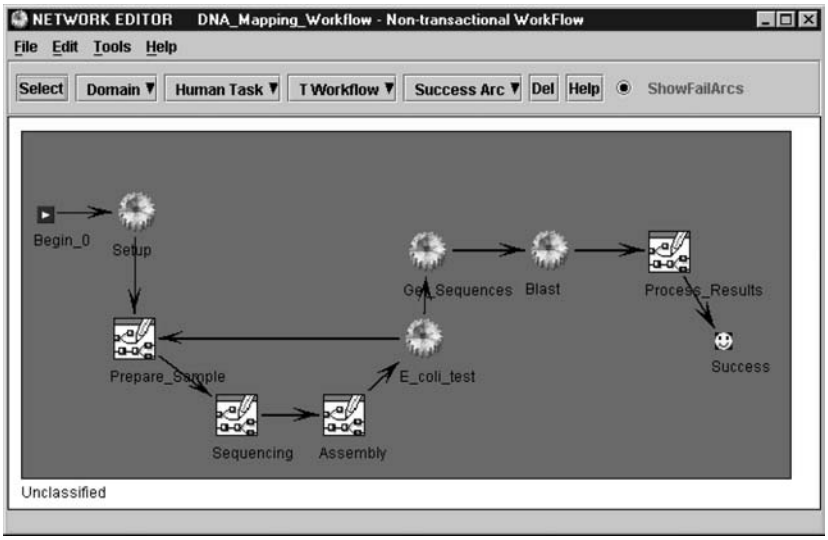


Figure 9. Genomic workflow example.

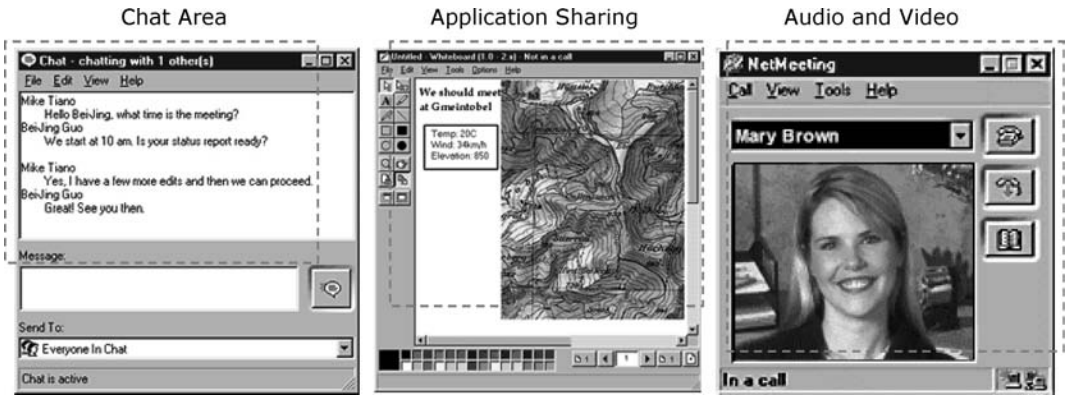


Figure 10. Microsoft NetMeeting with data, audio, and video conferencing (18).

person sharing the program can allow program control to other meeting participants.

Participants in a data conference can use a chat application to communicate in the absence of audio support. Chat can also be used to type text messages to share common ideas or topics with other conference participants or record meeting notes as part of a collaborative process.

Video Conferencing. Video conferencing uses telecommunications of audio and video to bring geographically dispersed people at different sites together for a meeting. Video conferencing is a valuable strategic tool for millions of individuals and small businesses for face-to-face meetings, team collaborations, brainstorming and training. There are two types of video conferencing: point-to-point and multipoint.

Point-to-point. A point-to-point conference is a connection between two video terminals. Each participant has a video camera, microphone, and speakers connected to a computer. As the two participants speak to one another, their voices are carried over the network and delivered to the other speakers, and images that appear in front of the video camera appear in a window on the other participants' monitor. Connecting two locations can be arranged simply by having one location dial the other, just as in a regular telephone call. No outside assistance is necessary.

Multipoint. A multipoint conference involves a connection among several video terminals connecting several sites with more than one person at different sites. This type of connection requires the assistance of a service to bridge the sites together into one conference. Until the mid-1990s, hardware costs made video conferencing prohibitively expensive for most organizations, but that situation is changing rapidly.

A video conference can involve just video, but some systems combine video, audio and data to provide a complete conferencing solution. One of the first and most popular video conferencing systems is NetMeeting (18). A product developed by Microsoft Corporation that enables groups to teleconference using the Internet as the transmission medium. NetMeeting (Fig. 10) supports video and audio conferencing, chat sessions, a whiteboard, and application sharing.

NetMeeting has been updated and extended with significant new capabilities designed to make it the most effective online meeting solution for integrated, interactive, and easy-to-use conferencing. The new version of this powerful application is now called Live Meeting (19).

Another well-known video conferencing program to transmit audio and video signals is CU-SeeMe. Originally developed by Cornell University, CU-SeeMe uses the standard bandwidth available on the Internet. Currently, CU-SeeMe is a low-cost software solution to the problem of electronic communication over the Internet.

IP Telephony. IP Telephony, also called 'Internet Telephony', allows voice and data to be transmitted over the same network using an open standards-based Internet Protocol (IP). It makes possible to exchange voice, fax, and other forms of information that have traditionally been carried over the dedicated circuit-switched connections of the public switched telephone network (PSTN). By combining different types of information on a single network connection, small and medium-sized businesses offices can decrease the costs of their voice and data networks.

IP Telephony is essential not just for its capability to reduce costs by combining voice and data communications, but also for its flexibility in supporting branch offices, mobile workers, and telecommuters that were not effective with PSTN. This technology allows an agile application deployment across the enterprise, increased personal and work group productivity, and permits a rapid return on investment.

CONCLUSIONS

Office Automation Systems specializes in allowing information workers to work, communicate, and collaborate. These systems are interactive and have the ability to allow workers to show and share documents or applications. These systems help workers worldwide to minimize the

costs of business travel and streamline communications with co-workers, business partners, and customers.

Healthcare processes are very complex, involving both clinical and administrative tasks, large volumes of data, and a large number of patients and personnel. For example, an out-patient clinic visit involves administrative tasks performed by an assistant and clinical tasks performed by a doctor or by a nurse. For an in-patient hospital visit, this scenario involves more activities, and the process entails a duration that lasts at least as long as the duration of patient hospitalization. Healthcare processes are also very dynamic. As processes are instantiated, changes in healthcare treatments, drugs, and protocols may invalidate running instances, requiring reparative actions. Common problems reported by healthcare organizations include delays due to the lack of timely communication; time invested in completing and routing paper-based forms; errors due to illegible and incomplete patient information; frustration due to the amount of time spent on administrative tasks instead of patient interactions; long patient wait times caused by slow communication of patient information.

Office automation systems are a major asset to solve many of the problems identified by the healthcare community. For example, using Workflow management systems, paper forms can be easily converted into digital forms for use by caregivers. These electronic forms can be used throughout the patient care process from registration and triage to placing lab orders and charting treatment plans. These forms can be easily modified to accommodate changing business processes. By automating clinical forms processes and eliminating manual systems, caregivers can streamline patient information management and treatment flow. Workflow management systems can connect the data and processes in clinical forms with other systems, such as a lab or patient records system. As another example, using whiteboard technologies, caregivers and administrators can access a central location to view patient information and status including triage category, and lab order status. This level of access can help to quickly determine the next steps in each patient's care. Blogs can also be effectively used by healthcare professionals to discuss specific topics of interest, such as product reviews, scientific endeavors, patient's treatments, and any area of information where people have a deep expertise and a desire to express it.

BIBLIOGRAPHY

1. Zisman M. Representation, Specification and Automation of Office Procedures, Department of Business Administration, Wharton School. Philadelphia: University of Pennsylvania; 1977.
2. Ellis CA. Information Control Nets: A Mathematical Model of Office Information Flow. Conference on Simulation, Measurement and Modelling of Computer Systems. New York: ACM; 1979.
3. Notes. Lotus Notes. 2005. Available at http://www-130.ibm.-com/developerworks/lotus/.
4. WFW. Windows for Workgroups. 2005. Available at http://www.microsoft.com/technet/archive/wfw/4_ch9.mspx. 2005.
5. Office. Microsoft Office. 2005. http://office.microsoft.com/. 2005.
6. Laudon JP, Laudon KC. Management Information Systems. 8th ed. New York: Prentice Hall; 2003.
7. Word. Microsoft Word. 2005. Available at http://office.microsoft.com/en-us/FX010857991033.aspx. 2005.
8. WordPerfect. Corel WordPerfect. 2005. http://www.corel.com/. 2005.
9. Lotus1-2-3. IBM Lotus 1-2-3. 2005. Available at http://lotus.-com/products/product2.nsf/wdocs/123home. 2005.
10. Excel. Microsoft Excel. 2005. Available at http://office.microsoft.com/en-us/ FX010858001033.aspx. 2005.
11. PageMaker. Abode PageMaker. 2005. Available at http://www.adobe.com/products/pagemaker/main.html. 2005.
12. QuarkXPress. Quark. 2005. Available at http://www.quark.com/. 2005.
13. InDesign. Adobe InDesign. 2005. Available at http://www.adobe.com/products/indesign/main.html. 2005.
14. Publisher. Microsoft Publisher. 2005. Available at http://office.microsoft.com/en-us/ FX010857941033.aspx. 2005.
15. Stohr EA, Zhao JL. Workflow Automation: Overview and Research Issues. Information Systems Frontiers 2001;3(3): 281–296.
16. WARIA. Workflow and Reengineering International Association. 2002.
17. WfMC. Workflow Management Coalition. 2002.
18. NetMeeting, Microsoft NetMeeting. 2005. Available at http://www.microsoft.com/windows/netmeeting/. 2005.
19. LiveMeeting. Microsoft Live Meeting. 2005. Available at http://office.microsoft.com/en-us/FX010909711033.aspx 2005.

See also EQUIPMENT ACQUISITION; MEDICAL RECORDS, COMPUTERS IN.

OPTICAL FIBERS IN MEDICINE.

See FIBER OPTICS IN MEDICINE.

OPTICAL SENSORS

YITZHAK MENDELSON
Worcester Polytechnic Institute

INTRODUCTION

Optical sensing techniques have attracted extraordinary interest in recent years because of the key role they play in the development of medical diagnostic devices. Motivated by the expense and time constraints associated with traditional laboratory techniques, there is a growing need to continue and develop more cost-effective, simpler, and rapid methods for real-time clinical diagnostics of vital physiological parameters.

Optical sensors play a pivotal role in the development of highly sensitive and selective methods for biochemical analysis. The number of publications in the field of optical sensors used for biomedical and clinical applications has grown significantly during the past two decades. Numerous books, scientific reviews, historical perspectives, and conference proceedings have been published on biosensors including optical sensors and the reader interested in this

rapidly growing field is advised to consult these excellent sources for additional reading (see Reading List). Some of these references discuss different optical sensors used in research applications and optical-based measurement techniques employed primarily in bench-top clinical analyzers. The emphasis of this article is on the basic concept employed in the development of optical sensors including specific applications highlighting how optical sensors are being utilized for real-time *in vivo* and *ex vivo* measurement of clinically significant biochemical variables, including some examples of optical sensor used for *in vitro* diagnosis. To narrow the scope, this article concentrates on those sensors that have generally progressed beyond the initial feasibility phase and have either reached or have a reasonable good potential of reaching the commercialization stage.

GENERAL PRINCIPLES OF OPTICAL BIOSENSING

The fundamental principle of optical sensors is based on the change in optical properties of a biological or physical medium. The change produced may be the result of the intrinsic changes in absorbance, reflectance, scattering, fluorescence, polarization, or refractive index of the biological medium. Optical sensors are usually based either on a simple light source–photodetector combination, optical fibers, or a planar waveguide. Some types of optical sensors measure changes in the intrinsic optical properties of a biological medium directly and others involve a specific indicator.

Biosensors are typically considered a separate subclassification of biomedical sensors. A biosensor, by definition, is a biomedical sensor consisting of an integrated biological component that provides the selectivity and a physical transducer to provide a solid support structure. Two major optical techniques are commonly available to sense optical changes at optical biosensor interfaces. These are usually based on evanescent wave, which was employed in the development of fiber optic sensors (see the section on Fiber Optic Sensors), and surface plasmon resonance principles, which played a pivotal role in the development and recent popularity of many optical biosensors. The basic principle of each measurement approach will be described first followed by examples arranged according to specific clinical applications.

Evanescent-Wave Spectroscopy

The propagation of light along a waveguide (e.g., a planar optical slab substrate or optical fiber) is not confined to the core region. Instead, when light travels through a waveguide at angles approaching the critical angle for total internal reflection, the light penetrates a characteristic short distance (on the order of one wavelength) beyond the core surface into the less optically dense (known as the cladding) medium as illustrated in Fig. 1. This effect causes the excitation of an electromagnetic field, called the "evanescent" wave, which depends on the angle of incidence and the incident wavelength. The intensity of the evanescent-wave decays exponentially with distance, starting at the interface and extending into the cladding medium.

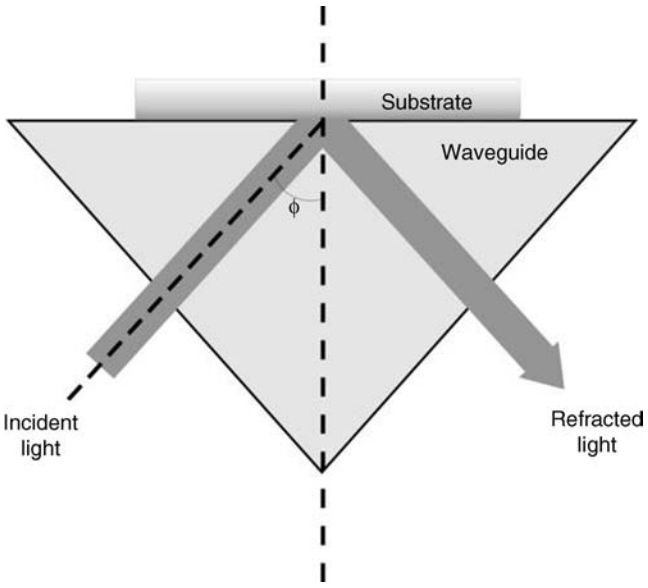


Figure 1. Principal diagram illustrating evanescence-wave spectroscopy sensing. Part of the incident light traveling through the waveguide at the critical angle (ϕ) penetrates a short distance into the substrate to be sensed and the remaining light is refracted.

The evanescent-wave can interact with molecules that are present within the penetration depth distance. This interaction causes attenuation of the incident light intensity and is related to the concentration of the molecules. This phenomenon can be exploited in the development of optical biosensors. For example, if the cladding is stripped and a substrate (such as a ligand) is immobilized on the core, the light will travel through this layer into the sample medium. Reactions close to the interface will perturb the evanescent field and the change in signal can be related to the amount of binding between the target and immobilized ligand at the interface. The measured parameter may be absorbance, fluorescence, or refractive index.

The method was first used as a means to study ultrathin films and coatings, and later was widely exploited to construct different types of optical sensors for biomedical applications. Because of the short penetration depth and the exponentially decaying intensity, the evanescent wave is absorbed by compounds that must be present very close to the surface. The principle can be utilized to characterize interactions between receptors that are attached to the surface of the optical sensor and ligands that are present in the solution probed by the sensor.

The key component in the successful implementation of evanescent-wave spectroscopy is the interface between the sensor surface and the biological medium. Receptors must retain their native conformation and binding activity and sufficient binding sites must be present for multiple interactions with the analyte. In the case of analytes having weakly optical absorbing properties, sensitivity can be enhanced by combining the evanescent-wave principle with multiple internal reflections along the sides of an unclad portion of a fiber optic tip. Alternatively, instead of an absorbing species, a fluorophore can be coated onto

the uncladded fiber. Light propagating along the fiber core is partially absorbed by the fluorophore, emitting detectable fluorescent light at a higher wavelength and thus providing improved sensitivity.

Surface Plasmon Resonance

When monochromatic polarized light (e.g., from a laser source) impinges on a transparent medium having a conducting metalized surface (e.g., Ag or Au), there is a charge density oscillation at the interface. When light at an appropriate wavelength interacts with the dielectric-metal interface at a defined angle, called the resonance angle, there is a match of resonance between the energy of the photons and the electrons at the metal interface. As a result, the photon energy is transferred to the surface of the metal as packets of electrons, called plasmons, and the light reflection from the metal layer will be attenuated. This results in a phenomenon known as surface plasmon resonance (SPR) as illustrated schematically in Fig. 2. The resonance is observed as a sharp dip in the reflected light intensity when the incident angle is varied. The resonance angle depends on the incident wavelength, the type of metal, polarization state of the incident light, and the nature of the medium in contact with the surface. Any change in the refractive index of the medium will produce a shift in the resonance angle, thus providing a highly sensitive means of monitoring surface interactions.

The SPR is generally used for sensitive measurement of variations in the refractive index of the medium immediately surrounding the metal film. For example, if an antibody is bound to or adsorbed into the metal surface, a noticeable change in the resonance angle can be readily observed because of the change of the refraction index at the surface assuming all other parameters are kept constant. The advantage of this concept is the improved ability to detect the direct interaction between antibody and antigen as an interfacial measurement.

Optical Fibers

An optical fiber consists of two parts: a core (typically made of a thin glass rod) with a refractive index n_1, and an outer layer (cladding) with a refractive index n_2, where $n_1 > n_2$.

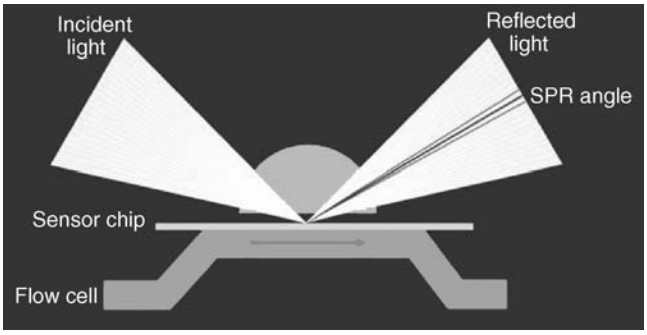


Figure 2. Principle of a SPR detection system. Courtesy of Biacore AB. An increased sample concentration causes a corresponding increase in refractive index that in turn alters the angle of incidence required to create the SPR angle.

The refractive indexes of the core and cladding depend on their material properties. Optical fibers are based on the principle of total internal reflection where incident light is transmitted through the core of the fiber with very little loss. The light strikes the cladding at an angle greater than the so-called critical angle, so that it is totally internally reflected at the core–cladding interface.

Several types of biomedical measurements can be made using either a plain optical fiber as a remote device for detecting changes in the intrinsic spectral properties of tissue or blood, or optical fibers tightly coupled to various indicator-mediated transducers. The measurement relies either on direct illumination of a sample through the end-face of the fiber or by excitation of a coating on the sidewall surface through evanescent wave coupling. In both cases, sensing takes place in a region outside the optical fiber itself. Light emanating from the fiber end is scattered or fluoresced back into the fiber, allowing measurement of the returning light as an indication of the optical absorption or fluorescence of the sample at the fiber tip.

A block diagram of a generic instrument for an optical fiber-based sensor is illustrated in Fig. 3. The basic building blocks of such an instrument are the light source, various optical coupling elements, an optical fiber guide with or without the necessary sensing medium incorporated at the distal tip, and a photodetector.

Probe Configurations

A number of different methods can be used to implement fiber optic sensors. Most fiber optic chemical sensors employ either a single fiber configuration, where light travels to and from the sensing tip in one fiber, or a double-fiber configuration, where separate optical fibers are used for illumination and detection. A single fiber optic

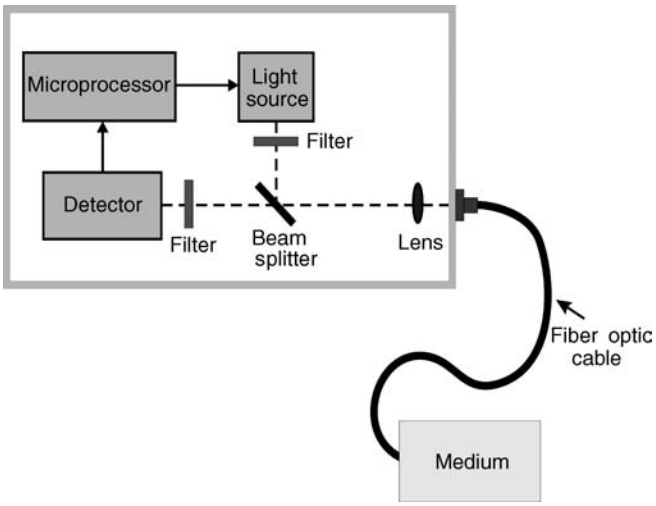


Figure 3. General building blocks of a fiber optic-based instrument for biomedical sensing applications. Typically, a microprocessor is used to control the light intensity and can be used to select either a fixed or a range of wavelengths for sample illumination. The microprocessor is also used to process the output of the photodetector.

configuration offers the most compact and potentially least expensive implementation. However, additional challenges in instrumentation are involved in separating the returning illuminating from the incident light illumination.

Intravascular fiber optic sensors are introduced into a vessel via a catheter. The design of intravascular fiber optic sensors requires additional considerations related to the sterility and biocompatibility of the catheter. For example, intravascular fiber optic sensors must be sterilizable and their material has to be nonthrombogenic and resistant to platelet and protein deposition. Therefore, these catheters are typically made of materials covalently bound with heparin or antiplatelet agents. The catheter is normally introduced via venous or arterial puncture and a slow heparin flush is maintained while the catheter remains in the body for short-term sensing, typically only for a few hours.

Indicator-Mediated Transducers

Indicator-mediated transducers are based on the coupling of light to a specific recognition element so that the sensor can respond selectively and reversibly to a change in the concentration of a particular analyte. The problem is that only a limited number of biochemical analytes have an intrinsic optical absorption that can be measured directly by spectroscopic methods with sufficient selectivity. Other species, particularly hydrogen ions and oxygen, which are of primary interest in diagnostic applications, do not have an intrinsic absorption and thus are not suitable analytes for direct photometry. Therefore, indicator-mediated transducers have been developed using specific reagents that can be immobilized on the surface of an optical sensor. These transducers may include indicators and ionophores (i.e., ion-binding compounds) as well as a wide variety of selective polymeric materials.

Figure 4 illustrates typical indicator-mediated fiber optic sensor configurations. In Fig. 4a, the indicator is immobilized directly on a membrane positioned at the end of a fiber. An indicator can be either physically retained in position at the end of the fiber by a special permeable membrane (Fig. 4b), or a hollow capillary tube (Fig. 4c). Polymers are sometimes used to enclose the indicator and selectively pass the species to be sensed.

Advantages and Disadvantages of Optical Fiber Sensors

Advantages of fiber optic sensors include their small size and low cost. In contrast to electrical measurements, fiber optic are self-contained, and therefore do not require an external reference signal from a second electrode. Because the signal that is transmitted is an optical signal, there is no electrical risk to the patient and the measurement is immune from interference caused by surrounding electric or magnetic fields. This makes fiber optic sensors very attractive for applications involving intense electromagnetic or radiofrequency fields, for example, near a magnetic resonance imaging (MRI) system or electrosurgical equipment. Chemical analysis can be performed in real-time with almost an instantaneous response. Furthermore, versatile sensors can be developed that respond to multiple analytes by utilizing multiwavelength measurements.

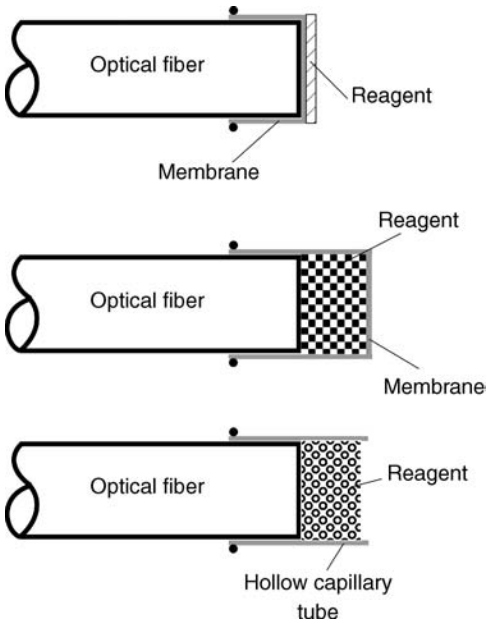


Figure 4. Typical configuration of different indicator-mediated fiber optic sensor tips showing different methods for placement of a reagent. Reprinted with permission from Fiber optic chemical sensors and biosensors, Vol. I, Wolfbeis OS, editor, CRC Press, Boca Raton, FL, c1991.

Despite unique advantages and promising feasibility studies, optical fiber sensors exhibit several shortcomings. Sensors with immobilized dyes and other indicators have limited long-term stability *in vivo* and their shelf-life degrades significantly over time.

INSTRUMENTATION

Instrumentation for optical measurements generally consist of (1) a light source to illuminate the medium, (2) optical components to generate a light beam with specific characteristics and to direct this light to some sensing agent or physical compartment, and (3) a photodetector for converting the optical signal to an electrical signal. The actual implementation of instrumentation designed to construct or interface with optical sensors vary greatly depending on the type of optical sensor used and its intended medical application.

Light Sources

Wide selections of light sources are available commercially for use in optical sensor applications. These include: narrow-band semiconductor diode lasers, broad spectral band incandescent lamps, and solid-state light emitting diodes (LEDs). The important requirement of a light source is obviously good stability. In certain applications, for example, in portable instrumentation, LEDs have significant advantages over other light sources since they are small, inexpensive, consume less power, produce selective wavelengths, and are easy to work with. In contrast, tungsten

lamps produce a broader range of wavelengths, have higher intensities, but require a stable and sizable power supply. Furthermore, incandescent lamps produce significant heat that can degrade or destroy delicate biological samples so special provisions must be made for thermal dissipation of excessive heat.

Optical Elements

Different optical elements can be used to manipulate light in optical instrumentation. These may include mirrors, lenses, light-choppers, beam-splitters, and couplers for directing the light from the light source to the measurement site and back to the photodetector. If special waveguides are involved, additional components may be required to direct the light into the small aperture of an optical fiber or a specific area on a waveguide surface, and collecting the light from the sensor before it is processed by the photodetector. For a narrow wavelength selection when a broad bandwidth incandescent light source is utilized, optical filters, prisms, or diffraction gratings are the most common used components.

Photodetectors

Several specifications must be considered in choosing photodetectors for optical sensors. These include: detector sensitivity, noise level, spectral response, and response time. The most popular photodetectors employed in biomedical sensors are semiconductor quantum photodetectors, such as Silicon photodiodes. The choice, however, is somewhat dependent on the wavelength region of interest. Often, dual photodetectors are used in spectrophotometric instrumentation because it is frequently necessary to include a separate reference photodetector to track fluctuations in source intensity and temperature. By taking a ratio between the reference photodetector, which measures part of the light that is not affected by the measurement, and the primary photodetector, which measures the light intensity that interacts with the analyte, it is possible to obtain a more accurate and stable measurement.

Signal Processing

The signal obtained from a photodetector typically provides a current or voltage proportional to the measured light intensity. Therefore, simple analog circuitry (e.g., a current-to-voltage converter or connection to a programmable gain voltage amplifier) is required. Sometimes, the output from a photodetector is connected directly to a preamplifier before it is applied to a sampling and analog-to-digital conversion circuitry. More recently, advanced Sigma-Delta type analog-to-digital converters became available commercially that can accept input current directly from a photodiode, thus eliminating the need for a separate current-to-voltage converter stage.

Frequently, two different wavelengths of light are utilized to perform a specific measurement. One wavelength is usually sensitive to changes in the analyte being measured, while the other wavelength is unaffected by changes in the analyte concentration. In this manner, the unaffected wavelength is used as a reference to compensate for fluctuations in instrumentation properties over time. In other applications, additional discriminations, such as pulse excitation or electronic background subtraction utilizing a synchronized lock-in amplifier, are useful to gain significant improvement in selectivity and sensitivity.

IN VIVO APPLICATIONS

Glucose Sensors

Research has shown that tightly controlling blood sugar levels can prevent or slow down the development of problems related to diabetes. Presently, the conventional method for measuring blood glucose requires a drop of blood and relies on an electrochemical reaction of glucose with a glucose-specific enzyme such as glucose oxidase. During the past 20 years, numerous attempts have been made to develop optical sensors for continuous invasive and noninvasive measurement of blood glucose. The main driving forces for developing a blood glucose sensor is to enable the development of a closed-loop artificial endocrine pancreas for optimizing the treatment of diabetes. Continuous monitoring of blood glucose would provide the patient more useful information on daily fluctuations in glucose levels, will increase patient's motivation and compliance for daily self-monitoring, and would aid in the optimization of insulin therapy resulting in better metabolic control. If perfected, such a system could provide reliable early warning of large excursions in blood glucose levels that may lead to hypo- and hyperglycemic conditions. Therefore, it would be valuable in preventing long-term complications associated with diabetes, such as coronary artery disease, neuropathy, retinopathy, and hypertension.

A fiber optic sensor for measuring blood glucose *in vivo* utilizing the concept of competitive binding was described by Schultz et al. (1). The idea was based on an analyte (glucose) that competes for binding sites on a substrate (the Lectin Concanavalin A) with a fluorescent indicator-tagged polymer [fluorescein isothiocyanate (FITC)-dextran]. The sensor was arranged so that the substrate is fixed in a position out of the optical path of the fiber end. The substrate is bound to the inner wall of a glucose-permeable hollow fiber tubing and fastened to the end of an optical fiber. The hollow fiber acts as the container and is impermeable to the large molecules of the fluorescent indicator. The light beam that extends from the fiber "sees" only the unbound indicator in solution inside the hollow fiber, but not the indicator bound on the container wall. Excitation light passes through the fiber and into the solution, causing the unbound indicator to fluoresce, and the fluorescent light passes back along the same fiber to a measuring system. The fluorescent indicator and the glucose are in competitive binding equilibrium with the substrate. The interior glucose concentration equilibrates with its concentration exterior to the probe. If the glucose concentration increases, the indicator is driven off the substrate to increase the concentration of the indicator. Thus, fluorescence intensity as seen by the optical fiber follows changes in the glucose concentration. *In vivo* studies

demonstrated fairly close correspondence between the sensor output and actual blood glucose levels. Another novel approach based on fluorescent molecules was suggested by Pickup et al.(2). The idea relied on the covalent binding of a fluorescent dye to glucose that results in a reduction of its fluorescence intensity when excited by light.

Although the measurement of glucose in plasma and whole blood *in vitro* is feasible (3), a more attractive approach for measuring blood glucose involves noninvasive measurement (4–6). The basic premise is to direct a light beam through the skin or laterally through the eye and analyze either the backscattered or transmitted light intensity. Three methods have been commonly attempted based either or changes in light absorption, polarization, or light scattering induced by variations in blood glucose. Although light in the visible and lower part of the near-infrared (NIR) region of the spectrum (700–2400 nm) can penetrate safely down to the vascular layer in the skin without significant attenuation there are major obstacles associated with noninvasive glucose measurement using spectrophotometry. Specifically, the concentration of glucose in tissue and blood is relatively low and light absorption in the NIR region is profoundly dependent on the concentration of water and temperature. Moreover, glucose has no unique absorption peaks in the visible or NIR region of the spectrum. Therefore, physiological variations in blood glucose induce only small and nonspecific changes in backscattered light intensity. Since measurements in the NIR region are due to low energy electronic vibrations, as well as high order overtones of multiple bands, NIR spectroscopy remains purely empirical. Thus, to extract accurate quantitative information related to variations in blood glucose, it is necessary to employ multivariate statistical calibration techniques and extensive chemometric analysis.

In principle, changes in light scattering caused by variations in blood glucose may offer another potential method for measuring glucose noninvasively. The fundamental principle exploited assumes that the refractive index of cellular structures within the skin remains unchanged while an increase in blood glucose leads to a subsequent rise in the refractive index of the blood and interstitial fluid. Limited studies involving glucose tolerance tests in humans (7,8) showed that changes in blood glucose could be measured from changes in light scattering. However, obtaining reliable and accurate measurement of blood glucose noninvasively using NIR spectroscopy remains challenging, mainly because other blood analytes (proteins, urea, cholesterol, etc.), as well as confounding physiological factors such as variations in blood flow, temperature, water content, and physical coupling of the sensor to the skin, are known to influence the measurement.

The implementation of an artificial pancreas would ultimately represent a major technological breakthrough in diabetes therapy. To date, however, inadequate specificity and insufficient accuracy within the clinically relevant range provide major obstacle in achieving this milestone with noninvasive blood glucose sensors using optical means.

Oximetry

Oximetry refers to the colorimetric measurement of the degree of oxygen saturation (SO_2), that is, the relative amount of oxygen carried by the hemoglobin in the erythrocytes. The measurement is based on the variation in the color of deoxyhemoglobin (Hb) and oxyhemoglobin (HbO_2). A quantitative method for measuring blood oxygenation is of great importance in assessing the circulatory and respiratory status of a patient.

Various optical methods for measuring the oxygen saturation in arterial (SaO_2) and mixed-venous (SvO_2) blood have been developed, all based on light transmission through, or reflection from, tissue and blood. The measurement is performed at two specific wavelengths: λ_1, where there is a large difference in light absorbance between Hb and HbO_2 (e.g., 660 nm red light), and a second wavelength, λ_2, which can be an isobestic wavelength (e.g., 805 nm IR light), where the absorbance of light is independent of blood oxygenation, or a higher wavelength in the near-IR region, typically between 805 and 960 nm, where the absorbance of Hb is slightly smaller than that of HbO_2.

The concept of oximetry is based on the simplified assumption that a hemolyzed blood sample consists of a two-component homogeneous mixture of Hb and HbO_2, and that light absorbance by the mixture of these two components is additive. Hence, a simple quantitative relationship can be derived for computing the oxygen saturation of blood based on the relationship:

$$SO_2 = K_1 - K_2[\mathrm{OD}(\lambda_1)/\mathrm{OD}(\lambda_2)]$$

where K_1 and K_2 are empirically derived coefficients that are functions of the specific absorptivities (also called optical extinction) of Hb and HbO_2, and OD (optical density) denotes the corresponding absorbance of the blood at a specific wavelength.

Since the original discovery of this phenomenon > 50 years ago (9), there has been progressive development in instrumentation to measure SO_2 along three different paths. Bench-top oximeters for clinical laboratories, which measure the concentration of Hb and HbO_2 from a small sample of arterial blood, fiber optic catheters for intravascular monitoring, and transcutaneous sensors, which are noninvasive devices placed on the skin.

Mixed-Venous Fiber Optic Catheters

Fiber optic oximeters for measuring mixed-venous oxygen saturation (SvO_2) were first described in the early 1960s by Polanyi and Hehir (10). They demonstrated that in a highly scattering medium, such as blood, it is feasible to use reflectance measurement to determine SO_2 in a flowing blood medium. Accordingly, they showed that a linear relationship exists between SO_2 and the ratio of the infrared-to-red (IR/R) light that is backscattered from blood:

$$SO_2 = A - B\,(\mathrm{IR/R})$$

where, A and B are empirically derived calibration coefficients.

The ability to rely on light reflectance for measurement of SO_2 *in vivo* subsequently led to the commercial development of fiber optic catheters for intravascular monitoring of SvO_2 inside the pulmonary artery.

Under normal conditions, oxygen consumption is less than or equal to the amount of oxygen delivered. However, in critically ill patients, oxygen delivery is often insufficient for the increased tissue demands, because many such patients have compromised compensatory mechanisms. If tissue oxygen demands increase and the body's compensatory mechanisms are overwhelmed, the venous oxygen reserve will be tapped, and that change will be reflected as a decreased SvO_2. Venous oxygen saturation in the pulmonary artery is normally $\sim 75\%$. Although no specific SvO_2 level has been correlated with adverse physiological effects, an SvO_2 level of 53% has been linked to anaerobic metabolism and the production of lactic acid (11), while an SvO_2 level of 50% or less indicates that oxygen delivery is marginal for oxygen demands, and thus venous oxygen reserve is reduced. Thus, continuous SvO_2 monitoring can be used to track the available venous oxygen reserve.

A fiber optic pulmonary artery catheter, with its tip in the pulmonary artery, can be used to sample the outflow from all tissue beds. For this reason, SvO_2 is regarded as a reliable indicator of tissue oxygenation (12) and therefore is used to indicate the effectiveness of the cardiopulmonary system during cardiac surgery and in the ICU.

Fiber optic SvO_2 catheters consist of two separate optical fibers; one fiber is used for transmitting the light to the flowing blood and a second fiber directs the backscattered light to a photodetector. The catheter is introduced into the vena cava and further advanced through the heart into the pulmonary artery by inflating a small balloon located at the distal end. The flow-directed catheter also contains a small thermistor for measuring cardiac output by thermodilution.

Several problems limit the wide clinical application of intravascular fiber optic oximeters. These include the dependence of the optical readings on hematocrit and motion artifacts due to catheter tip "whipping" against the blood vessel wall. Additionally, the introduction of the catheter into the heart requires an invasive procedure and can sometimes cause arrhythmias.

Pulse Oximetry

Noninvasive monitoring of SaO_2 by pulse oximetry is a rapidly growing practice in many fields of clinical medicine (13). The most important advantage of this technique is the capability to provide continuous, safe, and effective monitoring of blood oxygenation.

Pulse oximetry, which was first suggested by Aoyagi et al. (14) and Yoshiya et al. (15), relies on the detection of time-variant photoplethysmographic (PPG) signals, caused by changes in arterial blood volume associated with cardiac contraction. The SaO_2 is derived by analyzing the time-variant changes in absorbance caused by the pulsating arterial blood at the same R and IR wavelength used in conventional invasive-type oximeters. A normalization process is commonly performed by which the pulsatile (ac) component at each wavelength, which results from the expansion and relaxation of the arterial bed, is divided by the corresponding nonpulsatile (dc) component of the PPG, which is composed of the light absorbed by the blood-less tissue and the nonpulsatile portion of the blood compartment. This effective scaling process results in a normalized R/IR ratio, which is dependent on SaO_2, but is largely independent of the incident light intensity, skin pigmentation, tissue thickness, and other nonpulsatile variables.

Figure 5. A disposable finger probe of a noninvasive transmission pulse oximeter. Reprinted by permission of Nellcor Puritan Bennett, Inc., Pleasanton, California. The sensor is wrapped around the fingertip using a self-adhesive tape backing.

Pulse oximeter sensors consist of a pair of small and inexpensive R and IR LEDs and a highly sensitive silicon photodiode. These components are mounted inside a reusable rigid spring-loaded clip, a flexible probe, or a disposable adhesive wrap (Fig. 5). The majority of the commercially available sensors are of the transmittance type in which the pulsatile arterial bed (e.g., ear lobe, fingertip, or toe) is positioned between the LEDs and the photodiode. Other probes are available for reflectance (backscatter) measurement, where both the LEDs and photodetector are mounted side-by-side facing the skin (16,17).

Numerous studies have evaluated and compared the accuracy of different pulse oximeters over a wide range of clinical conditions (18–21). Generally, the accuracy of most noninvasive pulse oximeters is acceptable for a wide range of clinical applications. Most pulse oximeters are accurate to within $\pm$ 2–3% in the SaO_2 range between 70 and 100%.

Besides SaO_2, most pulse oximeters also offer other display features, including pulse rate and analogue or bar graph displays indicating pulse waveform and relative pulse amplitude. These important features allow the user to assess in real time the quality and reliability of the measurement. For example, the shape and stability of the PPG waveform can be used as an indication of possible motion artifacts or low perfusion conditions. Similarly, if the patient's heart rate displayed by the pulse oximeter differs considerably from the actual heart rate, the displayed saturation value should be questioned.

Several locations on the body, such as the ear lobes, fingertips, and toes, are suitable for monitoring SaO_2 with transmission pulse oximeter sensors. The most popular sites are the fingertips since these locations are convenient to use and a good PPG signal can be quickly obtained.

Other locations on the skin that are not accessible to conventional transillumination techniques can be monitored using a reflection (backscatter) SaO_2 sensor. Reflection sensors are usually attached to the forehead or to the temples using a double-sided adhesive tape.

Certain clinical and technical situations may interfere with the proper acquisition of reliable data or the interpretation of pulse oximeter readings. Some of the more common problems are low peripheral perfusion associated, for example, with hypotension, vasoconstriction, or hypothermia conditions. Also, motion artifacts, the presence of significant amounts of dysfunctional hemoglobins (i.e., hemoglobin derivatives that are not capable of reversibly binding with oxygen), such as HbCO and methemoglobin, and intravenous dyes introduced into the blood stream (e.g., methylene blue).

Pulse oximetry is widely used in various clinical applications including anesthesia, surgery, critical care, hypoxemia screening, exercise, during transport from the operating room to the recovery room, in the emergency room, and in the field (22–26). The availability of small and lightweight optical sensors makes SaO_2 monitoring especially applicable for preterm neonates, pediatric, and ambulatory patients. In many applications, pulse oximetry has replaced transcutaneous oxygen tension monitoring in neonatal intensive care. The main utility of pulse oximeters in infants, especially during the administration of supplemental oxygen, is in preventing hyperoxia since high oxygen levels in premature neonates is associated with increased risk of blindness from retrolental fibroplasia.

During birth, knowing the blood oxygenation level of the fetus is of paramount importance to the obstetrician. Lack of oxygen in the baby's blood can result in irreversible brain damage or death. Traditionally, physicians assess the well being of the fetus by monitoring fetal heart rate and uterine contractions through the use of electronic fetal monitoring, which are sensitive, but generally not specific. Approximately one-third of all births in the United States are marked by a period in which a nonreassuring heart rate is present during labor. Without a reliable method to determine how well the fetus is tolerating labor and when dangerous changes in oxygen levels occur, many physicians turn to interventions, such as cesarean deliveries.

Recently, the FDA approved the use of new fetal oxygen monitoring technology, originally developed by Nellcor (OxiFirst, Tyco Healthcare). The pulse oximeter utilizes a disposable shoehorn-shaped sensor (Fig. 6), which is inserted through the birth canal during labor, after the amniotic membranes have ruptured and the cervix is dilated at least 2 cm. The sensor, which comprises the same optical components as other pulse oximeter sensors used for nonfetal applications, rests against the baby's cheek, forehead, or temple, and is held in place by the uterine wall. The fetus must be of at least 36 weeks

Figure 6. Optical probe and a fetal pulse oximeter. Reprinted by permission of Nellcor Puritan Bennett, Inc., Pleasanton, California. The shape of the optical sensor is contoured to enable proper placement of the sensor against the fetus face.

gestation with the head in the normal vertex position for delivery. A controlled, randomized clinical trial, has demonstrated the safety and effectiveness of this new technology (27,28). While some clinical studies found that the rate of caesarean section for nonreassuring heart rate was significantly lower among the group of women monitored by the OxiFirst system, it is not yet clear whether the use of a fetal pulse oximeter may lead to a reduction in the number of Caesarean sections performed. Therefore, the American College of Obstetrics and Gynecology has not yet endorsed the use of this technology in clinical practice.

Noninvasive Cerebral Oximetry

Somanetics, Inc. (Troy, MI) has developed a noninvasive cerebral oximeter (INVOS) for monitoring of changes in brain oxygen saturation that can be used to alert clinicians to changes in the critical balance between arterial oxygen delivery and cerebral consumption (29–31). The method is based on NIR light photons injected by a light source into the skin over the forehead as illustrated in Fig. 7. After being scattered inside the scalp, skull, and brain, some fraction of the incident photons return and exit the skin. By measuring the backscattered light intensity as a function of two wavelengths, 730 and 810 nm, it is possible to

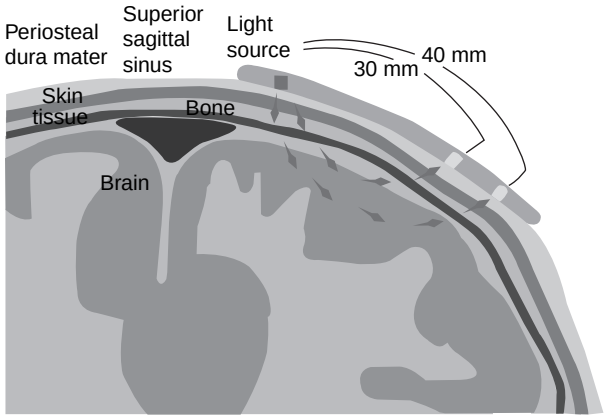


Figure 7. Principle of the INVOS cerebral oximeter. Courtesy of Somanetics, Corp, Troy, MI. Two photodetectors are used to capture the backscattered light from different depths in the brain.

measure changes in regional hemoglobin oxygen saturation (rSO_2 index).

The measurement is intuitively based on the fact that the greater the separation of source and detector, the greater the average depth of penetration. Photons that happen to meander close to the surface are very likely to be lost out of the skin before getting to a distant detector. Large source-detector separation is therefore biased against "shallow" photons except in the tissues directly under the optical sensor. On the other hand, geometry and absorption also make it statistically unlikely that very deeply penetrating photons will find their way back to the detector. Most of the photons reaching the detector tend to take some optimum middle course. This mean photon path is shaped approximately like a "banana" with ends located at the light source and photodetector.

To reduce extraneous spectroscopic interference that is dominated by light scattered by the surrounding bone, muscle, and other tissues, the INVOS SomaSensor® oximeter uses two source-detector separations: a "near" (shallow) spacing and a "far" (deep) spacing. The dual detectors sample about equally the shallow layers in the illuminated tissue volumes positioned directly under the light sources and photodetectors, but the far-spaced detector "sees" deeper than the near-spaced detector. By subtracting the two measurements, the instrument is able to suppress the influence of the tissues outside the brain to provide a measurement of changes in brain oxygen saturation.

Measurement of Blood Gases

Accurate measurement of arterial blood gases, that is, oxygen partial pressure (pO_2), carbon dioxide partial pressure (pCO_2), and pH, is one of the most frequently performed tests in the support of critically ill patients. The measurement is essential for clinical diagnosis and management of respiratory and metabolic acid–base problems in the operating room and the ICU. Traditionally, blood gases have been measured by invasive sampling, either through an indwelling arterial catheter or by arterial puncture, and analyzed in a clinical laboratory by a bench-top blood gas analyzer. However, this practice presents significant drawbacks: Sampling is typically performed after a deleterious event has happened, the measurement is obtained intermittently rather than continuously so physicians can not immediately detect significant changes in a patient's blood gas status, there could be a considerable delay between the time the blood sample is obtained and when the readings become available, there is an increased risk for patient infection, and there is discomfort for the patients associated with arterial blood sampling. Furthermore, frequent arterial blood gas sampling in neonates can also result in blood loss and may necessitate blood transfusions.

In view of the above drawbacks, considerable effort has been devoted over the last three decades to develop either disposable extracorporeal sensors (for *ex vivo* applications) or intraarterial fiber optic sensors that can be placed in the arterial line (for *in vivo* applications) to enable continuous trending that is vital for therapeutic interventions in ICU patients who may experience spontaneous and often unexpected changes in acid–base status. Thus, with the advent of continuous arterial blood gas monitoring, treatment modalities can be proactive rather than reactive. Although tremendous progress has been made in the miniaturization of intravascular blood gas and pH sensors, in order to be acceptable clinically, further progress must be achieved on several fronts. Specifically, there is a need to improve the accuracy and reliability of the measurement especially in reduced blood flow and hypotensive conditions. Additionally, sensors must be biocompatible and nonthrombogenic, the readings must be stable and respond rapidly to changes in physiological conditions, and the disposable sensors need to be inexpensive and more cost effective.

Intravascular Optical Blood Gas Catheters

In the early 1970s, Lubbers and Opitz (32) originated what they called "optodes" (from the Greek word '*optical path*') for measurements of important physiological gases in fluids and in gases. The principle upon which these sensors were designed originally was based on a closed compartment containing a fluorescent indicator in solution, with a membrane permeable to the analyte of interest (either ions or gases) constituting one of the compartment walls. The compartment was coupled by optical fibers to a system that measured the fluorescence inside the closed compartment. The solution equilibrates with the pO_2 or pCO_2 of the medium placed against it, and the fluorescence intensity of an indicator reagent in the solution was calibrated to the partial pressure of the measured gas.

Intraarterial blood gas optodes typically employ a single or a double fiber configuration. Typically, the matrix containing the indicator is attached to the end of the optical fiber. Since the solubility of O_2 and CO_2 gases, as well as the optical properties of the sensing chemistry itself, are affected by temperature variations, fiber optic intravascular sensors include a thermocouple or thermistor wire running alongside the fiber optic cable to monitor and correct for temperature fluctuations near the sensor tip. A nonlinear response is characteristic of most chemical indicator sensors, so they are designed to match the concentration region of the intended application. Also, the response time of optodes is somewhat slower compared to electrochemical sensors.

Intraarterial fiber optic blood gas sensors are normally placed inside a standard 20-gage arterial cannula, which is sufficiently small thus allowing adequate spacing between the sensor and the catheter wall. The resulting lumen is large enough to permit the withdrawal of blood samples, introduction of a continuous heparin flush, and the recording of a blood pressure waveform. In addition, the optical fibers are encased in a protective tubing to contain any fiber fragments in case they break off. The material in contact with the blood is typically treated with a covalently bonded layer of heparin, resulting in low susceptibility to fibrin deposition. Despite excellent accuracy of indwelling intraarterial catheters *in vitro* compared to blood gas analyzers, when these multiparameter probes were first introduced into the vascular system, it quickly became evident that the readings (primarily pO_2) varies frequently

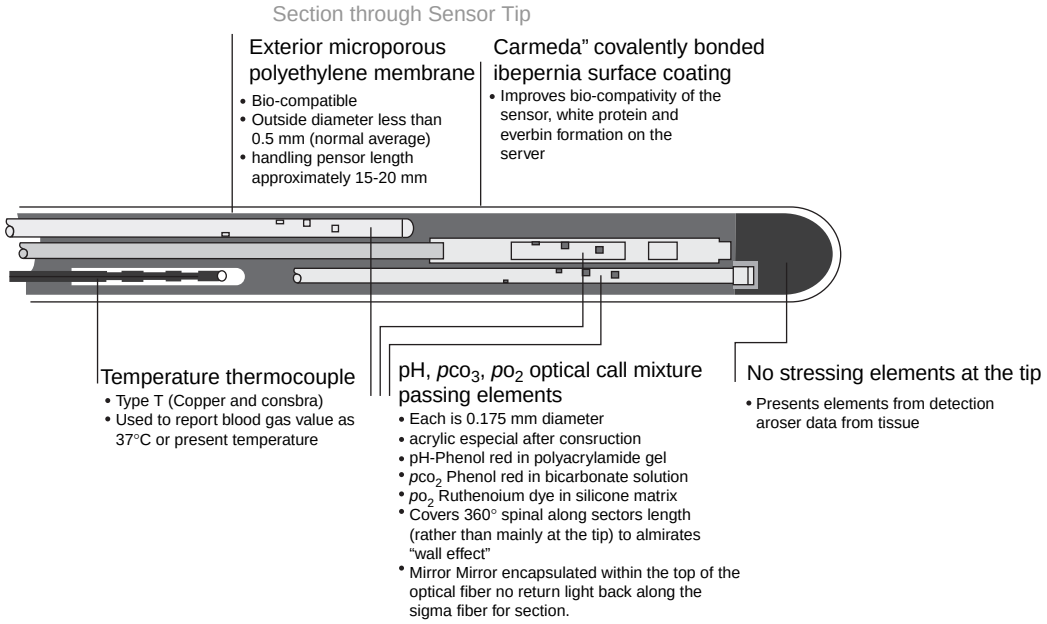


Figure 8. Principle of an indwelling arterial optical blood gas catheter. Courtesy of Diametrics, Inc, St Paul, MN. A heparin-coated porous polyethylene membrane encapsulates the optical fibers and a thermistor.

and unpredictably, mainly due to the sensor tip intermittently coming in contact with the wall of the arterial blood vessel and intermittent reductions in blood flow due to arterial vasospasm (33–35).

Recently, a more advanced multiparameter disposable probe (Fig. 8) consisting of a pO_2, pCO_2, and pH sensors was developed by Diametrics Medical, Inc. (36–38). Several technical issues discovered during clinical evaluations of earlier probes were successfully addressed by this new product, and multiple clinical studies confirmed that the system is adequate for trend monitoring. The device has been evaluated in neurosurgical patients for continuous monitoring in the brain and in critically ill pediatric patients (39–41).

Although significant progress has been made, most of the intravascular probes are adversely affected by patient–probe interfacing problems and high manufacturing cost. These problems must be fully overcome before continuous intravascular blood gas monitoring can be performed reliably and cost effectively in widely divergent patient groups. Continuous arterial blood gas monitoring will not completely replace traditional invasive blood sampling. Nevertheless, it may extend the clinician's ability to recognize and intervene more effectively to correct potentially life-threatening conditions.

pO_2 Sensors

The basic principle of measuring pO_2 optically relies on the fluorescence quenching effect of oxygen. Fluorescence quenching is a general property of aromatic molecules. In brief, when light is absorbed by a molecule, the absorbed energy is held as an excited electronic state of the molecule. It is then lost by coupling to the mechanical movement of the molecule (heat), reradiated from the molecule in a mean time of ~ 10 ns (fluorescence), or converted into another excited state with much longer mean lifetime (phosphorescence). Quenching reduces the intensity of the emitted fluorescence light and is related to the concentration of the quenching molecules. The quenching of luminescence by oxygen was initially described by Kautsky in 1939 (42).

Opitz and Lubbers (43) and Peterson et al. (44) developed a fiber optic sensor for measuring pO_2 using the principle of fluorescence quenching. The dye is excited at ~ 470 nm (blue) and fluoresces at ~ 515 nm (green) with an intensity that depends on the pO_2. The optical information is derived from the ratio of light intensities measured from the green fluorescence and the blue excitation light, which serves as an internal reference signal. The original sensor contained a Perylene Dibutyrate dye contained in an oxygen-permeable porous polystyrene envelope (45). The ratio of green to blue intensity is processed according to the Stern–Volmer equation (46):

$$I_0/I = 1 + K pO_2$$

where I and I_0 are the fluorescence emission intensities in the presence and absence (i.e., $pO_2 = 0$) of the quencher, respectively, and K is the Stern–Volmer quenching coefficient. The method provides a nearly linear readout of pO_2 over the range of 0–150 mmHg (0–20 kPa), with a precision of ~ 1 mmHg (0.13 kPa). The original sensor was 0.5 mm in diameter its 90% response time in an aqueous medium was ~ 1.5 min.

pH Sensors

Peterson and Goldstein et al.(47) developed the first fiber optic chemical sensor for physiological pH measurement by placing a reversible color-changing indicator at the end of a

pair of optical fibers. In the original development, the popular indicator phenol red was used since this dye changes its absorption properties from the green to the blue part of the spectrum as the acidity is increased. The dye was covalently bound to a hydrophilic polymer in the form of water-permeable microbeads that stabilized the indicator concentration. The indicator beads were contained in a sealed hydrogen ion-permeable envelope made out of hollow cellulose tubing. In effect, this formed a miniature spectrophotometric cell at the end of the fibers.

The phenol red dye indicator is a weak organic acid, and its unionized acid and ionized base forms are present in a concentration ratio that is determined according to the familiar Henderson–Hasselbalch equation by the ionization constant of the acid and the pH of the medium. The two forms of the dye have different optical absorption spectra. Hence, the relative concentration of one of the forms, which varies as a function of pH, can be measured optically and related to variations in pH. In the pH sensor, green and red light emerging from the distal end of one fiber passes through the dye where it is backscattered into the other fiber by the light-scattering beads. The base form of the indicator absorbs the green light. The red light is not absorbed by the indicator and is therefore used as an optical reference. The ratio of green to red light is measured and is related to the pH of the medium.

A similar principle can also be used with a reversible fluorescent indicator, in which case the concentration of one of the indicator forms is measured by its fluorescence rather than by the absorbance intensity. Light, typically in the blue or ultraviolet (UV) wavelength region, excites the fluorescent dye to emit longer wavelength light. The concept is based on the fluorescence of weak acid dyes that have different excitation wavelengths for the basic and acidic forms but the same emitted fluorescent wavelength. The dye is encapsulated in a sample chamber that is permeable to hydrogen ions. When the dye is illuminated with the two different excitation wavelengths, the ratio of the emitted fluorescent intensities can be used to calculate the pH of the solution that is in contact with the encapsulated dye.

The prototype device for measuring pH consisted of a tungsten lamp for illuminating the optical fiber, a rotating filter wheel to select the red and green light returning from the fiber, and signal processing to provide a pH output based on the ratio of the green-to-red light intensity. This system was capable of measuring pH in the physiologic range between 7.0 and 7.4 with an accuracy and precision of 0.01 pH units. However, the sensor was susceptible to ionic strength variations.

Further development of the pH probe for practical use was continued by Markle et al. (48). They designed the fiber optic probe in the form of a 25-gage ($\phi = 0.5$ mm) hypodermic needle, with an ion-permeable side window, using 75-mm diameter plastic optical fibers. With improved instrumentation, and with a three-point calibration, the sensor had a 90% response time of 30 s and the range was extended to ± 3 pH units with a precision of 0.001 pH units.

Different methods were suggested for fiber optic pH sensing (49–52). A classic problem with dye indicators is the sensitivity of their equilibrium constant to variations in ionic strength. To circumvent this problem, Wolfbeis and Offenbacher (53) and Opitz Lubber (54) demonstrated a system in which a dual sensor arrangement can measure ionic strength and pH while simultaneously correcting the pH measurement for ionic strength variations.

pCO_2 Sensors

The pCO_2 of a sample is typically determined by measuring changes in the pH of a bicarbonate solution. The bicarbonate solution is isolated from the sample by a CO_2-permeable membrane, but remains in equilibrium with the CO_2 gas. The bicarbonate and CO_2, as carbonic acid, form a pH buffer system and, by the Henderson–Hasselbalch equation,

$$\mathrm{pH} = 6.1 + \log \frac{\mathrm{HCO_3^-}}{p\mathrm{CO_2}}$$

the hydrogen ion concentration is proportional to the pCO_2 of the sample. This measurement is done with either a pH electrode or a dye indicator. Both absorbance (55) and fluorescence (56) type pCO_2 sensors have been developed.

Vurek et al. (57) demonstrated that the same technique could be used also with a fiber optic sensor. In his design, one optical fiber carries light to the transducer, which is made of a silicone rubber tubing ~ 0.6 mm in diameter and 1.0 mm long, filled with a phenol red solution in a 35-mM bicarbonate. Ambient pCO_2 controls the pH of the solution that changes the optical absorption of the phenol red dye. The CO_2 permeates through the rubber to equilibrate with the indicator solution. A second optical fiber carries the transmitted signal to a photodetector for analysis. A different design by Zhujun and Seitz (58) used a pCO_2 sensor based on a pair of membranes separated from a bifurcated optical fiber by a cavity filled with bicarbonate buffer.

Extracorporeal Measurement

Several extracorporeal systems suitable for continuous on-line *ex vivo* monitoring of blood gases, base-access, and HCO_3 during cardiopulmonary bypass operations (Fig. 9) are available commercially (59–61). While this approach circumvents some of the advantages of continuous *in vivo* monitoring, this approach is useful in patients requiring frequent measurements of blood gases. The basic approach is similar to the optical method utilized in intravascular probes, but the sensors are located on a cassette that is placed within the arterial pressure line that is inserted into the patient's arm. The measurement is performed extravascularly by drawing a blood sample past the in-line sensor. After the analysis, which typically takes ~ 1 min, the sample can then be returned to the patient and the line is flushed. The sensor does not disrupt the pressure waveform or interfere with fluid delivery. Several clinical studies showed that in selected patient groups, the performance of these sensors is comparable to that of conventional *in vitro* blood gas analyzers (62–66).

Hematocrit Measurement

In recent years, an optical sensor has been developed by Hemametrics (Kaysville, UT) for monitoring hematocrit,

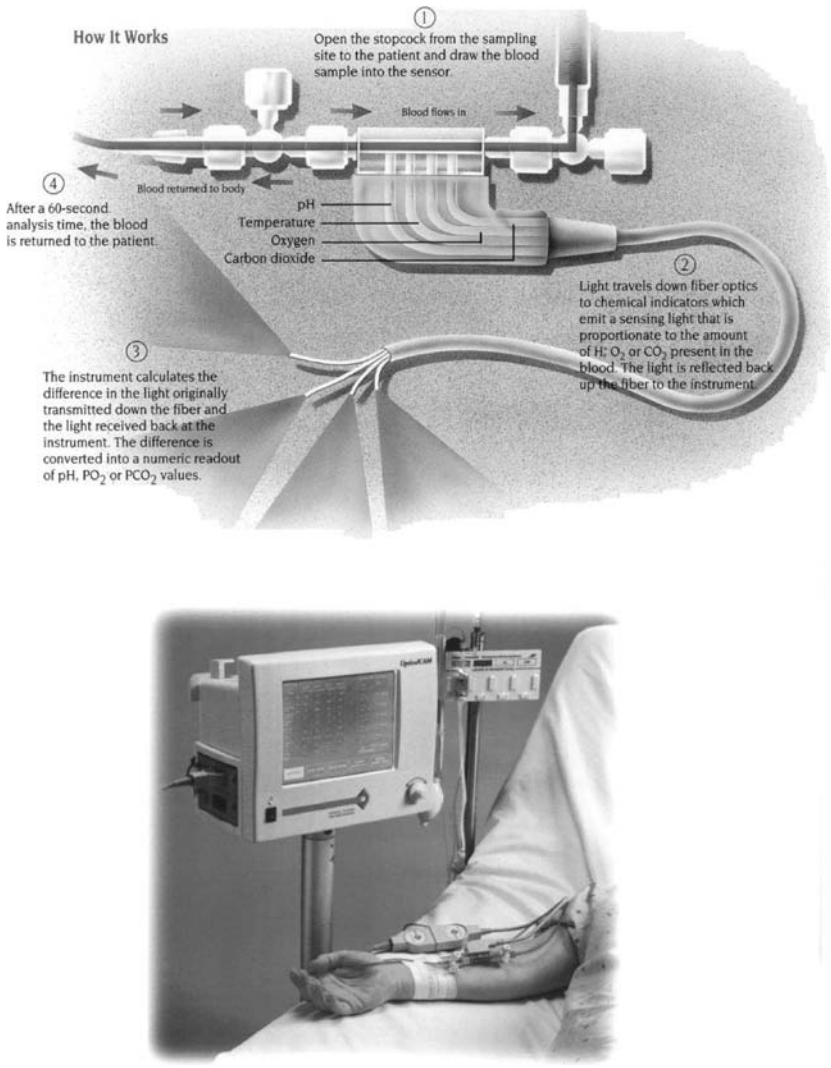


Figure 9. Principle of the GEM SensiCath fiber optic-based blood gas system. Courtesy of Optical Sensors, Inc., Minneapolis, MN.

oxygen saturation and blood volume *ex vivo* during hemodialysis (67–69). It provides vital and real time trending information that can be used to alert clinicians instantaneously to rapid changes during renal therapy (e.g., gastrointestinal bleeding or hemolysis). The optical measurements are based on detecting variations in scattering and absorption by blood flowing through a disposable cuvette that is connected to the arterial side of the dialyzer. A noninvasive version of this technology (CritScan) was recently cleared by the FDA for measuring absolute hematocrit and oxygen saturation by resting a fingertip against an optical sensor array of LEDs and photodetectors. One potential applications of this device, which was developed for spot checking, is for use in blood banking as an anemia screening tool to qualify blood donors, eliminating the need for performing hematocrit measurement based on the traditional and painful needle-stick approach.

Transcutaneous Bilirubin Measurement

Jaundice (hyperbilirubinemia) in newborn babies is evident in $> 50\%$ of newborns during the first week of birth. While hyperbilirubinemia (i.e., total serum bilirubin > 1.0 mg/dL) occurs in nearly all infants, moderate to significant hyperbilirubinemia usually peaks between 3 and 7 days of age and typically requires phototheraphy and/or exchange blood transfusions. If the condition is not recognized and treated properly, it can lead to potentially irreversible bilirubin-induced neurotoxicity and neurologic dysfunction. Visual recognition of jaundice is often inaccurate and unreliable, but excessive hyperbilirubinemia can be diagnosed by laboratory-based assay of total serum bilirubin obtained from a heel stick or an umbilical line.

To achieve a more objective measurement of jaundice, Yamanouchi et al. (70) developed a hand-held transcutaneous bilirubinometer developed by the Minolta Company. The meter used a dual optical filter design to measure a hemoglobin-corrected yellow color that is a distinctive skin color in jaundice patients. The measurement gives a reflectance value that must first be correlated to patient's serum bilirubin. Clinical testing of this device revealed significant inaccuracies and patient dependent variability compared with serum bilirubin determinations due to the presence of melanin pigmentation in the skin and variations in hemoglobin content.

More recently, several companies introduced more advance hand-held devices that uses multiwavelengths to measure bilirubin based on spectral analysis of light reflectance from the skin pioneered by Jacques (71) and others (72–74). The device consists of a light source, a fiberoptic probe, and a photodetector, essentially functioning as a microspectrophotometer. The unit is housed in a hand-held assembly that is positioned against the infant's forehead.

Pressure Sensors

In vivo pressure measurements provide important diagnostic information. For example, pressure measurements inside the heart, cranium, kidneys, and bladder can be used to diagnose abnormal physiological conditions that are otherwise not feasible to ascertain from imaging or other diagnostic modalities. In addition, intracranial hypertension resulting from injury or other causes can be monitored to assess the need for therapy and its efficacy. Likewise, dynamic changes of pressure measured inside the heart, uterus, and bladder cavities can help to assess the efficiency of these organs during muscular contractions.

There has long been an interest in developing fiber optic transducers for measuring pressure inside the cranium, vascular system, or the heart. Several approaches have been considered in the development of minimally invasive miniature pressure sensors. The most common technique involves the use of a fiber optic catheter. Fiber optic pressure sensors have been known and widely investigated since the early 1960s. The major challenge is to develop a small enough sensor with a high sensitivity, high fidelity, and adequate dynamic response that can be inserted either through a hypodermic needle or in the form of a catheter. Additionally, for routine clinical use, the device must be cost effective and disposable.

A variety of ideas have been exploited for varying a light signal in a fiber optic probe with pressure (75–77). Most designs utilize either an interferometer principle or measure changes in light intensity. In general, interferometric-based pressure sensors are known to have a high sensitivity, but involve complex calibration and require complicated fabrication. On the other hand, fiber optic pressure sensors based on light intensity modulation have a lower sensitivity, but involve simpler construction.

The basic operating principle of a fiber optic pressure sensor is based on light intensity modulation. Typically, white light or light produced by a LED is carried by an optical fiber to a flexible mirrored surface located inside a pressure-sensing element. The mirror is part of a movable membrane partition that separates the fiber end from the fluid chamber. Changes in the hydrostatic fluid pressure causes a proportional displacement of the membrane relative to the distal end of the optical fiber. This in turn modulates the amount of light coupled back into the optical fiber. The reflected light is measured by a sensitive photodetector and converted to a pressure reading.

A fiber optic pressure transducer for *in vivo* application based on optical interferometry using white light was recently developed by Fisco Technologies (Fig. 10). The sensing element is based on a Fabry–Perot principle. The Fabry–Perot cavity is defined on one end by a stainless steel diaphragm and on the opposite side by the tip of the optical fiber. When an external pressure is applied to the transducer, the deflection of the diaphragm causes variation of the cavity length that in turn is converted to a pressure reading.

Figure 10. Fiber optic in vivo pressure sensor. Courtesy of Fiso Technologies, Quebec, Canada.

IN VITRO DIAGNOSTIC APPLICATIONS

Immunosensors

The development of immunosensors is based on the observation of ligand-binding reaction products between a target analyte and a highly specific binding reagent (78–80). The key component of an immunosensor is the biological recognition element typically consisting of antibodies or antibody fragments. Immunological techniques offer outstanding selectivity and sensitivity through the process of antibody–antigen interaction. This is the primary recognition mechanism by which the immune system detects and fights foreign matter and has therefore allowed the measurement of many important compounds at micromolar and even picomolar concentrations in complex biological samples.

In principle, it is possible to design competitive binding optical sensors utilizing immobilized antibodies as selective reagents and detecting the displacement of a labeled antigen by the analyte. In practice, however, the strong binding of antigens to antibodies and vice versa causes difficulties in constructing reversible sensors with fast dynamic responses. Other issues relate to the immobilization and specific properties related to the antibody-related reagents on the transducer surface.

Several immunological sensors based on fiber optic waveguides have been demonstrated for monitoring antibody–antigen reactions. Typically, several centimeters of cladding are removed along the fiber's distal end and the recognition antibodies are immobilized on the exposed core surface. These antibodies bind fluorophore–antigen

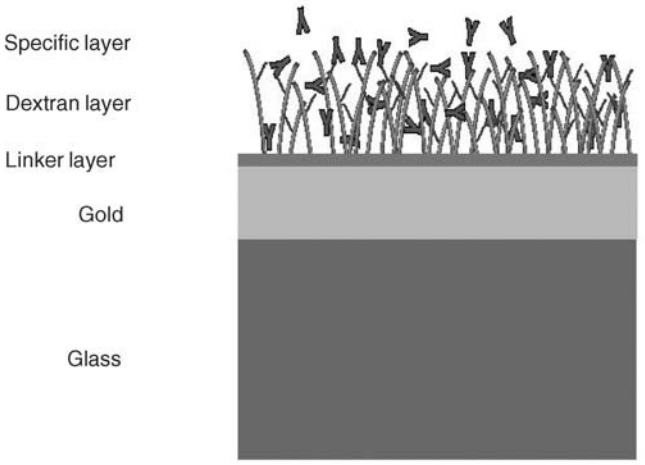


Figure 11. Typical principle of an immunodiagnostic sensor. Courtesy of Biacore AB. The sensor chip consists of a glass surface, coated with a thin layer of gold modified with a Dextran layer for binding of different biomolecules.

complexes within the evanescent wave. The fluorescent signal excited within the evanescent wave is then transmitted through the cladded fiber to a fluorimeter for processing.

Despite an overwhelming number of papers describing immunosensing techniques, there are still only few commercial instruments based upon immunosensors in clinical diagnostics today. Among them is the technology introduced by Biacore (Uppsala, Sweden) (81), which utilizes SPR as the principal probing method (Fig. 11) for real time monitoring of binding events without labeling or often purification of the substances involved. The sensor chip of the Biacore SPR system consists of a glass surface, coated with a thin layer of gold that is modified with a carboxymethylated dextran layer. This dextran hydrogel layer, which provides a hydrophilic environment for attached biomolecules, preserving them in a nondenatured state, forms the basis or a range of specialized surfaces designed to optimize the binding of a variety of molecules that occurs between active biomolecules. The IAsys from Affinity Sensors (Cambridge, UK) (82) also exploits the evanescence field principle, but the method is based on measuring the difference in the phase of a laser beam interacting with a waveguide structure. The method is utilized to study and measure the interactions of proteins binding to nucleic acids, lipids, and carbohydrates and can be used in molecular recognition experiments to study for example the DNA kinetics interactions.

Optical Biosensors in Drug Discovery

Optical biosensors exploit evanescent waves or SPR techniques, where binding of molecules in solution to surface-immobilized receptors causes an alteration of the refractive index of the medium near the surface. This optical change can be used to measure the amount of bound analyte using very low amounts of compound without the need for prior chemical labeling. Hence, evanescent waves and SPR techniques can be used in screening compounds for receptor binding affinity or to study the kinetic analysis of molecular interactions throughout the process of drug development (83). These methods also allow researchers to study *in vitro* the interaction between immobilized receptors and analytes and the general interrogation of intermolecular interactions (84–88). These sensors have become a standard method for characterizing biomolecular interactions and are useful tools to study, for example, if a certain analyte binds to a particular surface, how strong is the binding mechanism, and how much of the sample remains active. The recent development of highly multiplexed optical biosensor arrays has also accelerated the process of assay development and target identification that can speed up the tedious and time-consuming process involved in screening and discovery of new drugs.

BIBLIOGRAPHY

1. Schultz JS, Mansouri S, Goldstein IJ. Affinity sensor: A new technique for developing implantable sensors for glucose and other metabolites. Diabetes Care 1982;5:245.
2. Pickup J, McCartney L, Rolinski L, Birch D. In vivo glucose sensing for diabetes management: progress towards noninvasive monitoring. Br Med J 1999;319:1289–1293.
3. Heise HM. Noninvasive monitoring of metabolites using near infrared spectroscopy: state of the art. Horm Metab Res 1996;28:527–534.
4. Klonoff DC. Noninvasive blood glucose monitoring. Diabetes Care 1997;20:433–437.
5. Cote GL. Noninvasive optical glucose sensing-an overview. J Clin Eng 1997;22(4): 253–259.
6. Heise HM, Marbach R, Bittner A. Clinical chemistry and near infrared spectroscopy: technology for non-invasive glucose monitoring. J Near Infrared Spectrosc 1998;6:349–359.
7. Heinemann L, Schmelzeisen-Redeker G. Noninvasive continuous glucose monitoring in type I diabetic patients with optical glucose sensors. Diabetologia. 1998;41:848–854.
8. Heinemann L et al. Noninvasive glucose measurement by monitoring of scattering coefficient during oral glucose tolerance tests. Diabetes Technol Ther 2000;2:211–220.
9. Severinghaus JW. Historical development of oxygenation monitoring. In: Payne JP, Severinghaus JW, editors. Pulse Oximetry. Berlin: Springer-Verlag; 1986.
10. Polanyi ML, Heir RM. In vivo oximeter with fast dynamic response. Rev Sci Instrum 1962;33:1050.
11. Nelson LD. Mixed venous oximetry. In: Snyder JV, Pinsky JV, editors. Oxygen Transport in the Critically Ill. Chicago: Year Book Medical; 1987. pp. 235–248.
12. Enger EL, Holm K. Perspectives on the interpretation of continuous mixed venous oxygen saturation. Heart Lung 1990;19:578–580.
13. Mendelson Y. Pulse oximetry: theory and applications for noninvasive monitoring. Clin Chem 1992;38(9):1601–1607.
14. Aoyagi T, Kishi M, Yamaguchi K, Watanable S. Improvement of the earpiece oximeter. Jpn Soc Med Electron Biomed Eng 1974; 90–91.
15. Yoshiya I, Shimada Y, Tanaka K. Spectrophotometric monitoring of arterial oxygen saturation in the fingertip. Med Biol Eng Comput 1980;18:27.
16. Mendelson Y, Solomita MV. The feasibility of spectrophotometric measurements of arterial oxygen saturation from the

scalp utilizing noninvasive skin reflectance pulse oximetry. Biomed Instrum Technol 1992;26:215–224.
17. Mendelson Y, McGinn MJ. Skin reflectance pulse oximetry: In vivo measurements from the forearm and calf. J Clin Monit 1991;7:7–12.
18. Iacobelli L, Lucchini A, Asnaghi E, Nesci M. Oxygen saturation monitoring. Minerva Anesthesiol 2002;68(5): 488–491.
19. Sinex JE. Pulse oximetry: principles and limitations. Am J Emerg Med 1999;17(1): 59–67.
20. Wouters PF et al. Accuracy of pulse oximeters: The European multicenter trial. Anesthesiol Analg 2002;94:S13–S16.
21. Severinghaus JW, Naifeh KH. Accuracy of response of six pulse oximeters to profound hypoxia. Anesthesiology 1987;67:551–558.
22. Lee WW, Mayberry K, Crapo R, Jensen RL. The accuracy of pulse oximetry in the emergency department. Am J Emerg Med 2000;18(4):427–431.
23. Kopotic RJ, Lindner W. Assessing high-risk infants in the delivery room with pulse oximetry. Anesthesiol Analg 2002;94:S31–S36.
24. Moller JT et al. Randomized evaluation of pulse oximetry in 20,802 patients: I. Anesthesiology 1993;78:436–444.
25. Jubran A. Pulse Oximetry. In Principles and practice of intensive care monitoring. In: Tobin MJ, editor. New York: McGraw Hill; 1998.
26. Yamaya Y et al. Validity of pulse oximetry during maximal exercise in normoxia, hypoxia, and hyperoxia. J Appl Physiol 2002;92:162–168.
27. Yam J, Chua S, Arulkumaran S. Intrapartum fetal pulse oximetry. Part I: Principles and technical issues. Obstet Gynecol Surv 2000;55(3): 163–172.
28. Luttkus AK, Lubke M, Buscher U, Porath M, Dudenhausen JW. Accuracy of fetal pulse oximetry. Acta Obstet Gynecol Scand 2002;81(5):417–423.
29. Edmonds HL. Detection and treatment of cerebral hypoxia key to avoiding intraoperative brain injuries. APSF Newslett 1999;14(3):25–32.
30. Kim M, Ward D, Cartwright C, Kolano J, Chlebowski S, Henson L. Estimation of jugular venous O_2 saturation from cerebral oximetry or arterial O_2 saturation during isocapnic hypoxia. J Clin Monit 2001;16:191–199.
31. Samra SK, Stanley JC, Zelenock GB, Dorje P. An assessment of contributions made by extracranial tissues during cerebral oximetry. J Neurosurg Anest 1999;11(1):1–5.
32. Lubbers DW, Opitz N. The pCO_2/pO_2-optode: A new probe for measurement of pCO_2 or pO_2 in fluids and gases. Z Naturforsch, C: Biosci 1975;30C:532–533.
33. Hansmann DR, Gehrich JL. Practical perspectives on the in vitro and in vivo evaluation of a fiber optic blood gas sensor. Proc SPIE Opt Fibers Med III 1988;906:4–10.
34. Shapiro BA, Cane RD, Chomka CM, Bandala LE, Peruzzi WT. Preliminary evaluation of an intra-arterial blood gas system in dogs and humans. Crit Care Med 1989;17:455–460.
35. Mahutte CK et al. Progress in the development of a fluorescent intravascular blood gas system in man. J Clin Monit 1990;6:147–157.
36. Venkatesh B, Clutton-Brock TH, Hendry SP. A multiparameter sensor for continuous intraarterial blood gas monitoring: a prospective evaluation. Crit Care Med 1994;22:588–594.
37. Venkatesh B, Clutton-Brock TH, Hendry SP. Continuous measurement of blood gases using a combined electrochemical and spectrophotometric sensor. J Med Eng Technol 1994;18:165–168.
38. Abraham E, Gallagher TJ, Fink S. Clinical evaluation of a multiparameter intra-arterial blood-gas sensor. Intensive Care Med 1996;22:507–513.
39. Zauner A et al. Continuous monitoring of cerebral substrate delivery and clearance: initial experience in 24 patients with severe acute brain injuries. Neurosurgery 1997;40(2):294–300.
40. Coule LW, Truemper EJ, Steinhart CM, Lutin WA. Accuracy and utility of a continuous intra-arterial blood gas monitoring system in pediatric patients. Crit Care Med 2001;29(2):420–426.
41. Meyers PA, Worwa C, Trusty R, Mammel MC. Clinical validation of a continuous intravascular neonatal blood-gas sensor introduced through an umbilical artery catheter. Respir Care 2002 47(6):682–687.
42. Kautsky H. Quenching of luminescence by oxygen. Trans Faraday Soc 1939;35:216–219.
43. Opitz N, Lubbers DW. Theory and development of fluorescence-based optochemical oxygen sensors: oxygen optodes. Int Anaesthesiol Clin 1987;25:177–197.
44. Peterson JI, Fitzgerald RV, Buckhold DK. Fiber-optic probe for in vivo measurement of oxygen partial pressure. Anal Chem 1984;56:62.
45. Vaughan WM, Weber G. Oxygen quenching of pyrenebuteric acid fluorescence in water: a dynamic probe of the microenvironment. Biochemistry 1970;9:464–473.
46. Stern O, Volmer M. Uber die Abklingzeit der Fluorescenz. Z Phys 1919;20:183–188.
47. Peterson JI, Goldstein SR, Fitzgerald RV. Fiber optic pH probe for physiological use. Anal Chem 1980;52:864–869.
48. Markle DR, McGuire DA, Goldstein SR, Patterson RE, Watson RM. A pH measurement system for use in tissue and blood, employing miniature fiber optic probes. In: Viano DC, editor. Advances in Bioengineering. New York: American Society of Mechanical Engineers; 1981 p 123.
49. Wolfbeis OS, Furlinger E, Kroneis H, Marsoner H. Fluorimetric analysis. 1. A study on fluorescent indicators for measuring near neutral (physiological) pH values. Fresenius' Z Anal Chem 1983;314:119.
50. Gehrich JL et al. Optical fluorescence and its application to an intravascular blood-gas system. IEEE Trans Biomed Eng 1986;33:117–132.
51. Seitz WR. Chemical sensors based on fiberoptics. Anal Chem 1984;56:17A–34A.
52. Yafuso M et al. Optical pH measurements in blood. Proc SPIE Opt Fibers Med IV 1989;1067:37–43.
53. Wolfbeis OS, Offenbacher H. Fluorescence sensor for monitoring ionic strength and physiological pH values. Sens Actuators 1986;9:85.
54. Opitz N, Lubbers DW. New fluorescence photomatrical techniques for simultaneous and continuous measurements of ionic strength and hydrogen ion activities. Sens Actuators 1983;4:473.
55. Smith BE, King PH, Schlain L. Clinical evaluation-continuous real-time intra-arterial blood gas monitoring during anesthesia and surgery by fiberoptic sensor. Int J Clin Monit 1992;9:45.
56. Miller WW, Gehrich JL, Hansmann DR, Yafuso M. Continuous in vivo monitoring of blood gases. Lab Med 1988;19:629–635.
57. Vurek GG, Feustel PJ, Severinghaus JW. A fiber optic pCO_2 sensor. Ann Biomed Eng 1983;11:499.
58. Zhujun Z, Seitz WR. A carbon dioxide sensor based on fluorescence. Anal Chim Acta 1984;160:305.
59. Clark CL, O'Brien J, McCulloch J, Webster J, Gehrich J. Early clinical experience with GasStat. J Extra Corporeal Technol. 1986;18:185.
60. Mannebach PC, Sistino JJ. Monitoring aortic root effluent during retrograde cardioplegia delivery. Perfusion 1997; 12(5):317–323.

61. Siggaard-Andersen O, Gothgen IH, Wimberley PD, Rasmussen JP, Fogh-Andersen N. Evaluation of the GasStat fluorescence sensors for continuous measurement of pH, pCO_2 and pO_2 during CPB and hypothermia. Scand J Clin Lab Invest 1988;48 (Suppl. 189):77.
62. Shapiro BA, Mahutte CK, Cane RD, Gilmour IJ. Clinical performance of an arterial blood gas monitor. Crit Care Med 1993;21:487–494.
63. Mahutte CK. Clinical experience with optode-based systems: early in vivo attempts and present on-demand arterial blood gas systems, 12th IFCC Eur. Cong. Clin. Chem. Medlab. 1997;53.
64. Mahutte CK. On-line arterial blood gas analysis with optodes: current status. Clin Biochem 1998;31(3):119–130.
65. Emery RW et al. Clinical evaluation of the on-line Sensicath™ blood gas monitoring system. Am J Respir Crit Care Med 1996;153:A604.
66. Myklejord DJ, Pritzker MR. Clinical evaluation of the on-line Sensicath™ blood gas monitoring system. Heart Surg Forum 1998;1(1):60–64.
67. Steuer R et al. A new optical technique for monitoring hematocrit and circulating blood volume: Its application in renal dialysis. Dial Transplant 1993;22:260–265.
68. Steuer R et al. Reducing symptoms during hemodialysis by continuously monitoring the hematocrit. Am J Kidney Dis 1996;27:525–532.
69. Leypoldt JK et al. Determination of circulating blood volume during hemodialysis by continuously monitoring hematocrit. J Am Soc Nephrol 1995;6:214–219.
70. Yamanouchi I, Yamauchi Y, Igarashi I. Transcutaneous bilirubinometry: preliminary studies of noninvasive transcutaneous bilirubin meter in the Okayama national hospital. Pediatrics 1980;65:195–202.
71. Jacques SL. Reflectance spectroscopy with optimal fiber devices and transcutaneous bilirubinometers. Biomed Opt Instrum Laser Assisted Biotechnol 1996;84–94.
72. Robertson A, Kazmierczak S, Vos P. Improved transcutaneous bilirubinometry: comparison of $SpectR_x$, Bili*Check* and Minolta jaundice meter JM-102 for estimating total serum bilirubin in a normal newborn population. J Perinatol 2002;22: 12–14.
73. Rubaltelli FF et al. Transcutaneous bilirubin measurement: A multicenter evaluation of a new device. Pediatrics 2001;107(6): 1264–1271.
74. Bhutani VK et al. Noninvasive measurement of total serum bilirubin in a multiracial predischarge newborn population to assess the risk of severe hyperbilirubinemia. Pediatrics 2000;106(2):e17.
75. Ivan LP, Choo SH, Ventureyra ECG. Intracranial pressure monitoring with the fiber optic transducer in children. Child's Brain 1980;7:303.
76. Kobayashi K, Miyaji H, Yasuda T, Matsumoto H. Basic study of a catheter tip micromanometer utilizing a single polarization fiber. Jpn J Med Electron Biol Eng 1983;21:256.
77. Hansen TE. A fiberoptic micro-tip pressure transducer for medical applications, Sens. Actuators 1983;4:545.
78. Luppa PB, Sokoll LJ, Chan DW. Immunosensors-Principles and applications to clinical chemistry. Clin Chem Acta 2001;314:1–26.
79. Morgan CL, Newman DJ, Price CP. Immunosensors: technology and opportunities in laboratory medicine. Clin Chem 1996;42:193–209.
80. Leatherbarrow RJ, Edwards PR. Analysis of molecular recognition using optical biosensors. Curr Opin Chem Biol 1999;3:544–547.
81. Malmqvist M. BIACORE: an affinity biosensor system for characterization of biomolecular interactions. Biochem Soc Trans 1999;27:335–340.
82. Lowe P et al. New approaches for the analysis of molecular recognition using IAsys evanescent wave biosensor. J Mol Recogn 1998;11:194–199.
83. Cooper MA. Optical biosensors in drug discovery, Nature Reviews. Drug Discov 2002;1:515–528.
84. Ziegler C, Gopel W. Biosensor development. Curr Opin Chem Biol 1998;2:585–591.
85. Weimar T. Recent trends in the application of evanescent wave biosensors. Angew Chem Int Ed Engl 2000;39:1219–1221.
86. Meadows D. Recent developments with biosensing technology and applications in the pharmaceutical industry. Adv Drug Deliv Rev 1996;21:179–189.
87. Paddle BM. Biosensors for chemical and biological agents of defense interest. Biosens Bioelectro 1996;11:1079–1113.
88. Keusgen M. Biosensors: new approaches in drug discovery. Naturwissenschaften 2002;89:433–444.

Reading List

Barth FG, Humphrey JAC, Secomb TW, editors. Sensors and Sensing in Biology and Engineering. New York: Springer-Verlag; 2003.
Eggins BR, Chemical Sensors and Biosensors for Medical and Biological Applications. Hoboken, NJ: Wiley; 2002.
Fraser D, editor. Biosensors in the Body: Continuous In Vivo Monitoring. New York: Wiley; 1997.
Gauglitz G, Vo-Dinh T. Handbook of Spectroscopy. Hoboken, NJ: Wiley-VCH; 2003.
Kress-Rogers E, editor. Handbook of Biosensors and Electronic Noses: Medicine, Food, and the Environment. Boca Raton, FL: CRC Press; 1996.
Ligler FS, Rowe-Taitt CA, editors. Optical Biosensors: Present and Future. New York: Elsevier Science; 2002.
Mirabella FM, editor. Modern Techniques in Applied Molecular Spectroscopy. New York: Wiley; 1998.
Ramsay G, editor. Commercial Biosensors: Applications to Clinical, Bioprocess and Environmental Samples. Hoboken, NJ: Wiley; 1998.
Rich RL, Myszka DG. Survey of the year 2001 optical biosensor literature. *J Mol Recogn* 2002;15:352–376.
Vo-Dinh T, editor. Biomedical Photonics Handbook. Boca Raton, FL: CRC Press; 2002.
Webster JG, editor. Design of Pulse Oximeters. Bristol UK: IOP Publishing; 1997.
Yang VC, Ngo TT. Biosensors and Their Applications. Hingham, MA: Kluwer Academic Publishing; 2002.

See also BLOOD GAS MEASUREMENTS; CUTANEOUS BLOOD FLOW, DOPPLER MEASUREMENT OF; FIBER OPTICS IN MEDICINE; GLUCOSE SENSORS; MONITORING, INTRACRANIAL PRESSURE.

OPTICAL TWEEZERS

HENRY SCHEK III
ALAN J. HUNT
University of Michigan
Ann Arbor, Michigan

INTRODUCTION

Since the invention of the microscope, scientists have peered down at the intricate workings of cellular machinery, laboring to infer how life is sustained. Doubtlessly, these investigators have frequently pondered What would

happen if I could push on that, or pull on this? Today, these formerly rhetorical thought experiments can be accomplished using a single-beam optical gradient trap, more commonly known as "optical tweezers". This article surveys the history, theory and practical aspects of optical trapping, especially for studying biology. We start by reviewing the early demonstrations of optical force generation, and follow this with a discussion of the theoretical and practical concerns for constructing, calibrating, and applying an optical tweezers device. Finally, examples of important optical tweezers experiments and their results are reviewed.

OPTICAL TWEEZERS SYSTEMS

History

In 1970, Arthur Ashkin published the first demonstration of light pressure forces manipulating microscopic, transparent, uncharged particles, a finding that laid the groundwork for optical trapping (1). Significant application of optical forces to study biology would not occur until almost two decades later, after the first description of a single-beam optical gradient trap was presented in 1986 (2). Soon after, the ability to trap living cells was demonstrated (3,4) and by the late 1980s biophysicists were applying optical tweezers to understand diverse systems, such as bacterial flagella (5), sperm (6), and motor proteins, such as kinesin (7). Today, optical tweezers are a primary tool for studying the mechanics of cellular components and are rapidly being adapted to applications ranging from cell sorting to the construction of nanotechnology (8,9).

Trapping Theory

Ashkin and co-workers 1986 description of a single-beam trap presented the necessary requirements for stable trapping of dielectric particles in three dimensions (2). For laser light interacting with a particle of diameter much larger than the laser wavelength, $d \gg \lambda$, that is the so-called Mie regime, the force on the particle can be calculated using ray optics to determine the momentum transferred from the refracted light to the trapped particle. Figure 1a–c schematically shows the general principle; two representative rays are bent as they pass through the spherical particle producing the forces that push the trapped particle toward the laser focus. A tightly focused beam that is most intense in the center thus pulls the trapped object toward the beam waist. More rigorous treatment and calculations can be found in (2,10).

In the Rayleigh regime, $d \ll \lambda$, the system must be treated in accordance with wave optics, and again the tight focusing results in a net trapping force due to the fact that the particle is a dielectric in a non-uniform electric field. Figure 1d shows a simplified depiction of force generation in this regime. The electric field gradient from the laser light induces a dipole in the dielectric particle; this results in a force on the particle directed up the gradient toward the area of greatest electric field intensity, at the center of the laser focus. Detailed calculations can be found in Ref. 11.

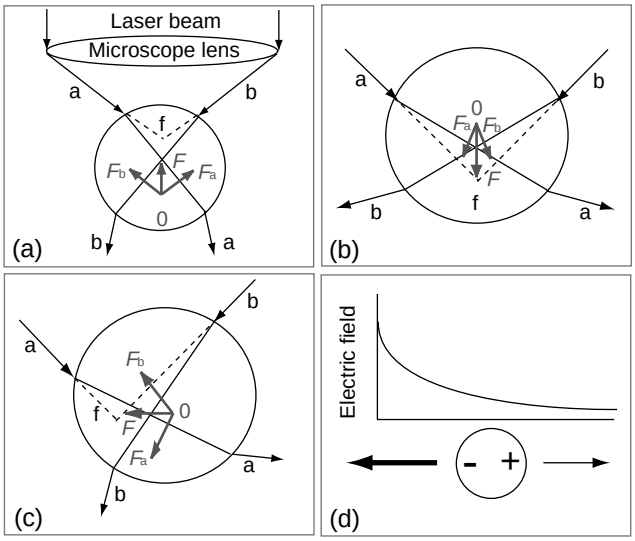


Figure 1. Optical forces on a dielectric sphere. (Adapted from Ref. 2.) Parts a–c show three possible geometries in the ray optics regime: particle above, particle below, and particle to the right of the laser focus, respectively. In each case, two representative rays are shown interacting with the particle. The two rays change direction upon entering and exiting the sphere causing a net transfer of momentum to the particle. Dotted lines show the focus point in the absence of the sphere. The forces resulting from each ray are shown in gray along with the summed force. (Please see online version for color in figure.) In each case the net force pushes the particle toward the focus. A highly simplified mechanism for the generation of force in the Rayleigh regime. An electric field profile is shown above a representative particle. The labels on the sphere show the net positioning of charges due to the formation of the dipole and the arrows show the forces. The induced dipole results in the bead being attracted to the area of most intense electric field, the laser focus.

In either regime, the principal challenge to producing a stable trap is overcoming the force produced by light scattered back in the direction of the oncoming laser, which imparts momentum that pushes the particle in the direction of beam propagation, and potentially out of the trap. When the trapping force that pulls the particle up the laser toward the beam waist is large enough to balance this scattering force, a stable trap results. From examination of Fig. 1a, it is apparent that the most important rays providing the force to balance the scattering force come at steep angles from the periphery of the focusing lens. For this reason, a tightly focusing lens is critical for forming an optical trap; typically oil-immersion lenses with a numerical aperture in excess of 1.2 are used. Furthermore, the beam must be expanded to slightly overfill the back aperture of the focusing lens so that sufficient laser power is carried in the most steeply focused periphery of the beam.

For biological experiments the preferred size of the trapped particle is rarely in the range appropriately treated in either the Mie or Raleigh regime; typically particles are on the order of 1 μm in diameter, or approximately equal to the laser wavelength. Theoretical treatment then requires generalized Lorenz–Mie theory (12,13). In

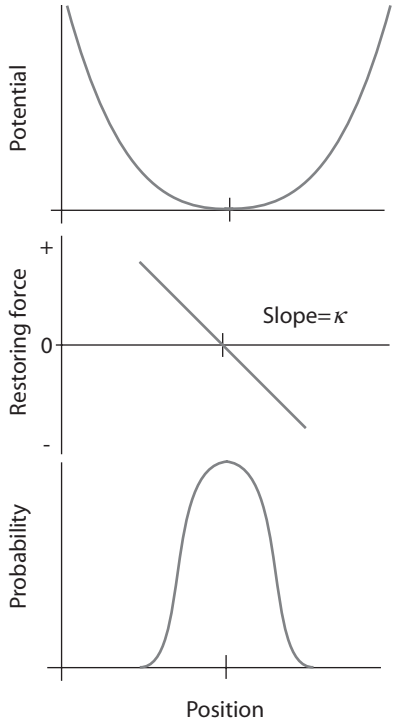


Figure 2. Relationship between potential, force, and bead position probability. The interaction of the tightly focused laser and dielectric results in a parabolic potential well, the consequence of which is a linear restoring force; the trap behaves as a linear radial spring. Here a positive force pushes the particle to the right. When driven by thermal events in solution a particle's position has a Gaussian distribution.

practice, theoretical analysis of specific trapping parameters is not necessary because several methods allow direct calibration of trapping forces. Regardless of particle size, a tightly focused laser with a Gaussian intensity profile (TEM_{00} mode) interacting with a spherical particle, traps the particle in a parabolic potential well. The parabolic potential is convenient because it results in a trap that behaves like a Hookian spring: restoring force that pushes the particle back toward the focus increases linearly with the displacement from the focus. Figure 2 schematically illustrates the potential well, force profile, and distribution of particle positions for a hypothetical trapped sphere. This convenient relationship results in spherical particles being the natural choice for most trapping applications including gradient. These particles are often referred to as beads, microbeads, or microspheres interchangeably.

Optical Tweezers Systems

Implementations of optical tweezers as they are applied to biology are variations on a theme: most systems share several common features although details vary and systems are often optimized for the application. Figures 3 and 4 show a photograph and a diagram of an example system. Generally, the system begins with a collimated, plane-polarized laser with excellent power and pointing stability. The optimal laser light wavelength depends on the application, but near-infrared (IR) lasers with a wavelength $\sim$1 μm are a typical choice for use with biological samples; such lasers are readily available and biological samples generally exhibit minimal absorption in the near-IR. A light attenuator allows for adjustment of laser power and therefore trap stiffness. In the example system this is accomplished with a variable half-wave plate that adjusts the polarity prior to the beam passing through a polarized beam splitting cube that redirects the unwanted laser power. Most systems then contain a device and lenses to actively steer the beam. In the case of the example system, a piezo-actuated mirror creates angular deflections of the beam that the telescope lenses translate into lateral position changes of the laser focus at the focal plane of the objective. This is accomplished by arranging the telescope so that a virtual image of the steering device is formed at the back aperture of the microscope objective lens. This arrangement also assures that the laser is not differentially clipped by the objective aperture as the trap is steered. Following the steering optics, the laser enters a microscope, is focused by a high numerical

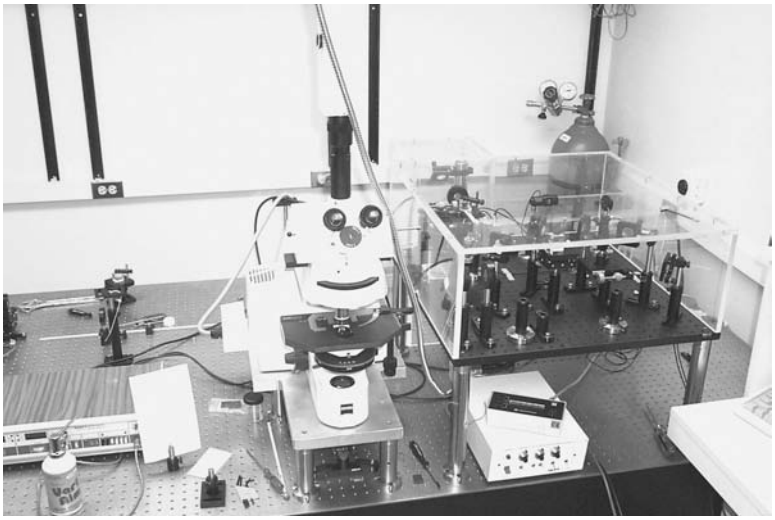


Figure 3. Photograph of an optical tweezers instrument. Note the vibration isolation table that prevents spurious vibrations from affecting experi- ments. The clear plastic cover minimizes ambient air convection that might affect the laser path in addition to limiting access of dust and preventing accidental contact with the optics.

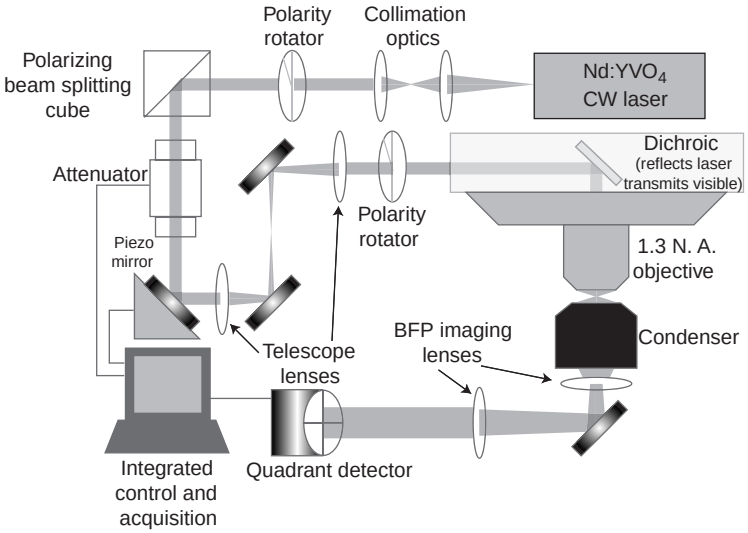


Figure 4. Optical tweezers schematic showing major system components.

aperture objective, and encounters the sample. After passing through the trapping/image plane, the laser usually exits the microscope and enters a detection system, which in this case employs back-focal-plane interferometry to measure the position of the bead relative to the trap (14–16).

There are many optical tweezers designs that can produce stable, reliable trapping. In most cases, the system is integrated into an inverted, high quality research microscope (17,18), although successful systems have been incorporated into upright (e.g., 15) or custom built microscopes (14). Some use a single laser to form the trap (7,15) while a more complicated arrangement uses two counterpropagating lasers (19,20).

Various implementations for detecting the bead position include image analysis (21,22), back focal plane interferometry (BFPI) (14,16), and a host of less frequently applied methods based on measurements of scattered or fluorescent light intensity (23–25) or interference between two beams produced using the Wollaston prisms associated with differential interference contrast microscopy (26). Trap steering can be accomplished with acoustooptic deflectors (15,27,28), steerable mirrors (23), or actuated lenses (14). Figure 5 shows a silica microsphere in a trap being moved in a circle at > 10 revolutions s^{-1} using acoustooptic deflectors, resulting in the comet tail in the image. Alternatively, the sample can be moved with a motorized or piezoactuated nanopositioning stage. In addition, splitting the laser into two orthogonally polarized beams, fast laser steering to effectively multiplex the beam by rapidly jumping the laser between multiple positions (29), or holographic technology (9) can be used to create arrays of multiple traps. Figure 6 shows an image of a 3×3 array of optical traps created using fast beam steering to multiplex a single beam. Five traps are holding beads while four, marked with white circles are empty. Specific optical tweezers designs can also allow incorporation of other advanced optical techniques often used in biology, including, but not limited to, total internal reflection microscopy and confocal microscopy (30).

Figure 5. Micrograph of 1 μm bead being manipulated in a circle with an optical tweezers device. The comet tail is the result of the bead moving faster than the video frame rate.

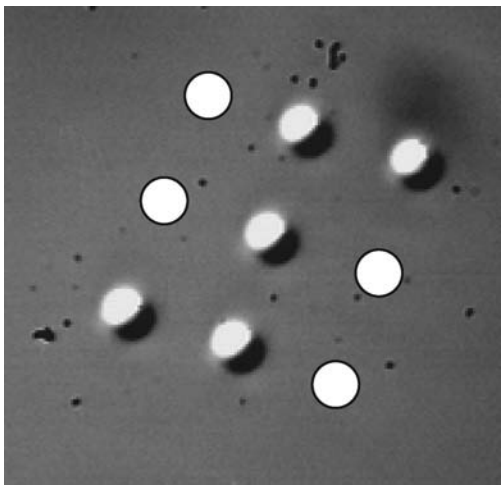


Figure 6. Micrograph of a 3×3 array of traps produced by time sharing a single laser between nine positions. The traps at the four locations marked with white circles are empty while the other five hold a 1 μm diameter silica microsphere.

Detection

Determining the force on a trapped sphere requires measurement of the displacement of the trapped particle from the center of the optical trap. Back focal plane interferometry (BFPI) is the most popular method, and allows high speed detection with nanometer resolution. Rather than following an image of the bead, which is also a viable detection technique, BFPI tracks the bead by examining how an interference pattern formed by the trapping laser is altered by the bead interacting with the laser further up the beam at the microscope focal plane. This interference pattern is in focus at the back focal plane of the condenser lens of the microscope, thus the name of the technique.

Laser light interacting with the bead is either transmitted or scattered. The transmitted and scattered light interfere with one another resulting in an interference pattern that is strongly dependent on the relative position of the bead within the trap. This interference pattern is easily imaged onto a quadrant photodiode (QPD) detector positioned at an image plane formed conjugate to the back focal plane of the condenser lenes by supplemental lenses (e.g., BFPI imaging lens in Fig. 4). A simple divider-amplifier circuit compares the relative intensity in each quadrant according to the equations shown in Fig. 7. This results in voltage signals that follows the position of the bead in the trap.

Back focal plane interferometry has several important advantages compared with the other commonly used bead tracking techniques. Typically, image analysis limits the data collection rate to video frame rate, 30 Hz, or slower if images must be averaged, while BFPI easily achieves sampling frequencies in the tens of kilohertz. Under ideal conditions, image analysis can detect particle positions with ~ 10 nm resolution, while BFPI achieves subnanometer resolution. Speeds and resolution similar to BFPI can be achieved by directly imaging a bead onto a QPD, however, it can be inconvenient in systems where trap steering is employed because the image of the particles will move across the detector as the trap moves in the image plane of the microscope, thus requiring frequent repositioning of the detector. With BFPI, the position of the intereference pattern at the back focal plane of the condenser does not move as the trap is steered, so the detector can remain stationary.

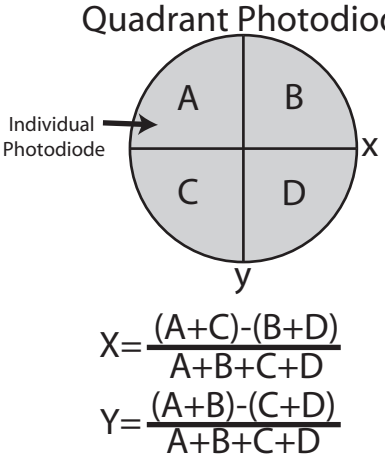


Figure 7. Quadrant photodiode operation. The circular quadrant photodiode detector is composed of four individual detectors (A–D) each making up one quarter of the circle. Each diode measures the light intensity falling on the surface of that quadrant and the associated electronics compares the relative intensities according to the equations shown in the figure. The resulting voltages are calibrated to the intensity shifts caused by the bead as it moves to different positions in the trap.

Calibration

Although BFPI provides a signal corresponding to bead position, two important system parameters must be measured before forces and positions can be inferred. The first is the detection sensitivity, β, relating the detector signal to the particle position. The second is the Hookian spring constant, κ, where

$$F = -\kappa x \tag{1}$$

where x is the displacement of the particle from the center of the trap. Both sensitivity and stiffness depend on the diameter of the particle and the position of the trap relative to nearby surfaces. Stiffness is linear with laser power while sensitivity, when using BFPI, does not depend on laser power as long as the detector is operating in its linear range. Ideally, these parameters are determined in circumstances as near to the experimental conditions as possible. Indeed whenever reasonably possible, it is wise to account for bead and assay variation by determining these parameters for each bead used in an experiment.

Direct Calibration Method

Sensitivity can be measured by two independent methods. The most direct is to move a bead affixed to a coverslip through the laser focus with a nanopositioning piezocontrolled stage, or conversely by scanning the laser over the immobilized bead while measuring the detector signal. An example calibration curve is shown in Fig. 8 for a system using BFPI. The linear region within $\sim$150 nm on either side of zero is fit to a line to arrive at the sensitivity, β, equal to 6.8×10^{-4} V·nm^{-1}. The second method is discussed below under calibration with thermally driven motion.

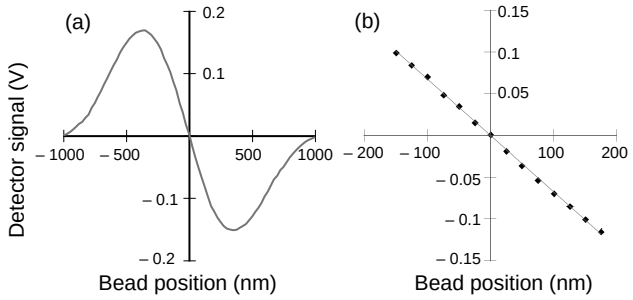


Figure 8. Optical tweezers sensitivity curve acquired with the direct method of moving the laser past an immobilized bead. The full curve is shown in a. The central linear region shown in b is fit to give a sensitivity of 6.8×10^{-4} V·nm^{-1}.

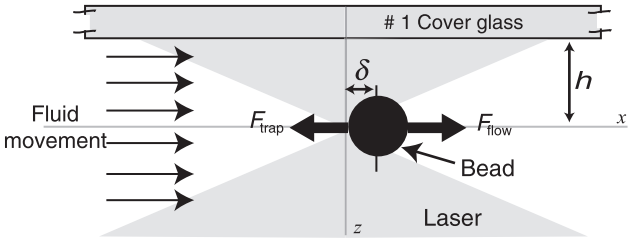


Figure 9. Schematic of drag force calibration. Uniform fluid flow from the left produces a force on the bead that can be calculated with the equations in the text. The bead is displaced to a distance, where the force due to fluid flow is balanced by the force due to the trap.

The direct calibration for sensitivity has the disadvantage of requiring the bead to be immobilized, typically to the coverslip, whereas experiments are typically conducted several microns from the surface. Additionally, with the bead fixed to a surface, its z-axis position is no longer controlled by the trapping forces, resulting in the possibility that the measured sensitivity does not correspond to the actual z axis position of a trapped particle. Estimates of sensitivity may also be rendered inaccurate if attachements to the surface slip or fail. Consequently, thermal force based calibration discussed in the next section are often more reliable.

The trap stiffness, κ, can be measured by applying drag force from fluid flow to the trapped particle. Figure 9 shows a schematic representation of the trap, particle, and fluid movement. Typically, a flow is induced around the bead by moving the entire experimental chamber on a motorized or piezodriven microscope stage. In the low Reynolds number regime where optical traps are typically applied, the drag coefficient, γ, on a sphere in the vicinity of a surface can be accurately calculated:

$$\gamma = 6\pi r c \eta \tag{2}$$

where r is the radius of the microsphere, η is the viscosity of the medium, and c is the correction for the distance from a surface given by

$$c = 1 + \frac{9}{16}\frac{r}{h} \tag{3}$$

where h is the distance from the center of the sphere to the surface. Equation 3 is an approximation and can be carried to higher order for greater accuracy (31).

The force due to fluid flow is readily calculated

$$F = \gamma V \tag{4}$$

where V is the velocity of the particle relative to the fluid medium.

When a flow force is applied to the trapped microsphere, the bead moves away from the center of the trap to an equilibrium position where the flow force and trapping force balance. Applying several drag forces (using a range of fluid flow velocities) and measuring the deflection in each case allows one to plot force of the trap as a function of bead position. The slope of this line is the trap stiffness, κ.

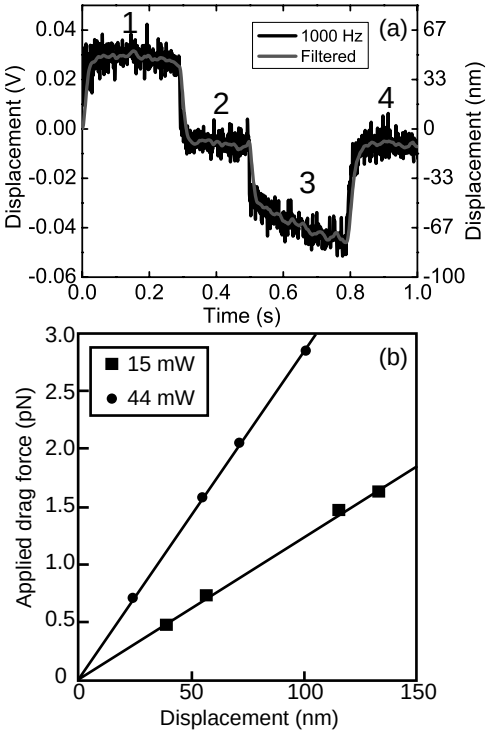


Figure 10. Results of a typical drag force calibration. (a) Typical signal of five averaged applications of uniform drag force on a trapped microsphere. In the region labeled 1, the bead quickly achieves a stable displacement due to the constant velocity. In region 2, the bead briefly returns to the center of the trap as the stage stops. In region 3, the bead is again displaced as the stage moves back to its home position and in region 4 the stage is at rest in its home position. The displaced position () in region 1 is averaged to determine the displacement for a given flow force. (b) Example of data from several different forces applied to the trapped microsphere reveal a linear relation to displacement.

Figure 10a shows a trace of a bead being displaced by fluid drag force provided by a piezoactuated stage moving at a known velocity. In the first region the bead quickly achieves and holds a stable position as long as the stage moves at a constant velocity, making measurement of the displacement relatively simple. When the stage stops the bead returns to the center of the trap as is seen in the second section of the record. The subsequent negative deflection in the third section is the result of the stage returning to its original position at a non-uniform speed. Figure 10b shows force as a function of position for two laser powers; the linear fitting coefficient gives the trap stiffness in each case.

In practice, calibration by viscous drag can be quite labor intensive, and the requirement for the sample to be moved repeatedly and rapidly is too disruptive to be performed "on the fly" during many types of experiments. A more efficient and less disruptive alternative relies on the fact that the thermal motion of the fluid molecules exerts significant, statistically predictable forces on the trapped microsphere; these result in fluctuations in bead position measurable with the optical tweezers.

Thermal Force Based Calibration Methods

At low Reynold's (Re) number the power spectrum of the position of a particle trapped in a harmonic potential well subject only to thermal forces takes the form of a Lorentzian (32):

$$S_x(f) = \frac{B}{f_c^2 + f^2} \tag{5}$$

with

$$B = \frac{k_B T}{\gamma \pi^2} \tag{6}$$

where k_B is the Boltzmann constant, f is the frequency, f_c is a constant called the corner frequency, and

$$f_c = \frac{\kappa}{2\pi\gamma} \tag{7}$$

There are several important practical considerations for applying the power spectrum to calibrate optical tweezers. The mathematical form of the power spectrum does not account for the electrical noise in the signal, so a large signal to noise ratio is required for accurate calibration. In addition, records of bead position must be treated carefully, especially with respect to instrument bandwidth and vibrations, to yield the correct shape and magnitude of the power spectrum.

A typical calibration begins by setting the position of the bead relative to any nearby surface, typically a microscope sample coverglass. With the bead positioned, a record of thermal motion of the bead is taken. For the example system in Figs. 3 and 4 a typical data collection for a calibration is 45 s at maximum bandwidth of the QPD and associated electronics. The data is then low pass (antialias) filtered using a high order Butterworth filter with a cut-off frequency set to one-half of the bandwidth.

Following acquisition and filtering, the power spectrum of the record of bead position is calculated. In practice, many power spectra should be averaged to reduce the inherently large variance of an individual power spectrum. For example, the 45 s of data is broken into 45, 1 s records of bead position, the power spectrum for each is calculated and the spectra are averaged together. To accurately compute the power spectrum the data must be treated as continuous; this is achieved by wrapping the end of the record back to meet the beginning. To avoid a discontinuity where the beginning and end are joined, each record is windowed with a triangle or similar function that forces the ends to meet but maintains the statistical variance in the original record. Figure 11a shows an example of an averaged power spectrum. Note the units on the y axis: often, power spectra are presented in different units according to their intended use. Accurate calibration of optical tweezers requires that the power spectrum be in $nm^2 \cdot Hz^{-1}$ or, provided volts are proportional to displacement, $V^2 \cdot Hz^{-1}$.

The average power spectrum is expected to fit the form of equation 5. To avoid aliased high frequency data corrupting the fit to the power spectrum, the spectrum is cropped well below the cut-off frequency of the antialiasing

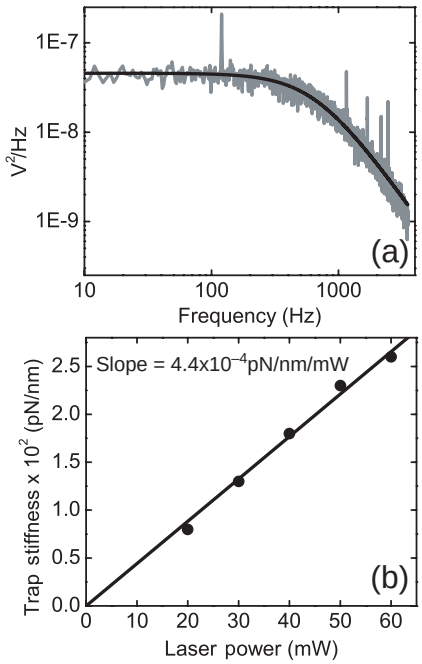


Figure 11. Example power spectrum calibration. (a) An example power spectrum (gray) and Lorentzian fit (black). (b) Power spectrum calibrations for five separate laser powers showing the linear dependence of stiffness on laser power.

filter. The average, cropped power spectrum can be visually inspected for evidence of noise (e.g., spikes or other deviations from the expected Lorentzian form) and fit to equation 5, with f_c and B as fitting parameters. The drag coefficient is calculated using equations 2 and 3 and the stiffness, κ, is then calculated with a simple rearrangement of equation 7. The sensitivity, β, can also be calculated by recognizing that the value of B produced by the fit is proportional to the signal power. The value of B can be calculated from first principles by applying equation 6. In practice, the value of B is determined by fitting equation 5 with B as a parameter. Equation 6 gives B in units of $nm^2 \cdot Hz^{-1}$ while the fitted value of B will typically be $V^2 \cdot Hz^{-1}$. The parameter β is then determined by dividing the fitted value by the theoretical value and taking the square root. Figure 11b shows an example power spectrum used to calculate the trap stiffness and the sensitivity of the detection system with a 0.6 μm bead.

Inspection of the power spectrum in Fig. 11a shows two distinct regions: the low frequency plateau and the high frequency roll off. These two sections can be qualitatively understood as representing two aspects of the motion of the trapped particle. The high frequency roll-off is representative of small, rapid oscillations of the bead in the trap. Since the distance the bead moves over these short time scales is generally small, the bead does not "feel" large changes in the force of the trap pushing it back to the center, and hence this motion has the character of free diffusion. At lower frequency, the spectrum is flat, because large-scale oscillations are attenuated by the trap pushing the bead back to the center. As a result low frequency oscillations do not exhibit the character of free diffusion; the bead feels the trap when it makes a large excursion.

The methods for calculating and then fitting the power spectrum above are adequate in most cases and for most experiments. However, several additional considerations can be added to this relatively simple approach to improve the accuracy in some situations. These include, but are not limited to, accounting for the frequency dependence of the drag coefficient, and theoretically accounting for aliasing that may be present in the signal (54).

Two other calibration methods also take advantage of thermal forces acting on trapped beads. These methods are not conceptually distinct from the power spectrum method, but treat the data differently, are susceptible to different errors, and thus provide a good cross-check with the power spectrum and direct methods above.

As one would expect, the range of bead position should decrease with increasing trap stiffness. For an overdamped harmonic potential, such as a spherical bead in an optical trap, this relation has a specific mathematical form:

$$\kappa = \frac{k_B T}{\mathrm{var}(x)} \tag{8}$$

where the var*(x)* is the variance of the bead position in the trap, k_B is Boltzmann's constant and T is absolute temperature. This equation results from the equipartition theorem: the average thermal energy for a single degree of freedom is $1/2k_BT$, and that the average potential energy stored in a Hookean trap is $1/2\kappa\langle x^2\rangle$. Setting these equal, rearranging and assuming that $\langle x\rangle = 0 =$ the center of the trap, results in equation 8. Thus by simply measuring the variance of the bead position and the temperature, one can determine the trap stiffness.

Alternatively the autocorrelation function of the bead position is used to determine the trap stiffness (33). This method is little different in principal than using the power spectrum, as the spectrum is the Fourier transform of the autocorrelation. The autocorrelation function is expected to exhibit an exponential decay with time constant τ, where

$$\tau = \frac{\gamma}{\kappa} \tag{9}$$

It is often easier to achieve a reliable fit to the autocorrelation than to the power spectrum, making the autocorrelation an attractive alternative to the more common power spectrum methods.

Each calibration method has its distinct advantages over the others. The methods involving direct manipulation are most often used in the course of building an optical tweezers device to verify that the thermal motion calibrations work properly. Once verified the thermal motion calibrations are much less labor intensive, and can be performed "on the fly". This allows sensitivity and stiffness for each bead to be calibrated as it is used in an experiment.

Each method relies on slightly different parameters and thus provides separate means to verify calibration accuracy. For example, the corner frequency of the power spectrum, and hence the stiffness can be determined with no knowledge of the detector sensitivity. However, this does depend on accurately calculating the drag coefficient. By calculating the stiffness from the variance one can avoid considering the drag coefficient altogether, but this method relies on accurate estimation of the sensitivity squared (β^2). Furthermore, the variance method can lead to inaccurate stiffness measurements because system drift will inflate the variance, leading to underestimated trap stiffness.

A safe course is to (1) inspect the power spectrum for evidence of external noise, (2) calculate the stiffness and sensitivity by fitting the power spectrum, (3) recalculate the stiffness with the variance method using the sensitivity determined from the power spectrum, and (4) compare the stiffness results from each method. If the stiffness agrees between the two it indicates that the drag coefficient and sensitivity are both correct. The only possibility for error would be that they are both in error in a manner that is exactly offsetting, which is quite unlikely. Comparison with the direct manipulation methods can alleviate this concern.

Optical Tweezers Compared with Other Approaches to Nanomanipulation

There are several alternatives to optical tweezers for working at the nanometer to micron scales and exerting piconewton forces. Most similar in application are magnetic tweezers, which use paramagnetic beads and an electromagnet to produce forces. The force profile of magnetic tweezers is constant on the size scale of microscopic experiments; the force felt by the bead is not dependent on displacement from a given point as with optical tweezers. This can be convenient in some circumstances, but is less desirable for holding form objects in specific locations.

A second option is the use of glass microneedles. These fine glass whiskers can be biochemically linked to a molecule of interest and their deflection provides a measure of forces and displacements. Stiffness must be measured for each needle with either fluid flow or methods relying on thermal forces similar to those described above. Glass needles can exert a broad range of forces but they only allow for force measurements along one axis, and are difficult for complex manipulations compared to optical tweezers devices.

Atomic force microscopy (AFM) is a third option for measuring small forces and displacements. Originally, this technique was used to image surface roughness by dragging a fine cantilever over a sample. A laser reflecting off the cantilever onto a photodiode detects cantilever deflection. To measure force, a cantilever of known stiffness can be biochemically linked to a structure of interest and deflections and forces measured with similar accuracy to optical tweezers. The AFM has the advantage that reliable AFMs can be purchased, while optical tweezers must still be custom built for most purposes. However, AFMs are unable to perform the complex manipulations that are simple with optical tweezers and typically are limited to forces >20 pN.

OPTICAL TWEEZERS RESEARCH

Within 4 years of the publication of the first demonstration of a single beam, three-dimensional (3D) trap, optical

tweezers were being applied to biological measurements (2,7). In the years that followed, optical tweezers became increasingly sophisticated, and their contributions to biology and biophysics in particular grew rapidly. Some important experimental considerations, and contributions to basic science made possible by optical tweezers, highlighted representative assays, and results are discussed in this section.

Practical Experimental Concerns

Beyond the design and calibration challenges, there are additional concerns that come into play preparing an experimental assay for use with optical tweezers. The most general of these attached is a trappable object to the system being studied, and accounting for series compliances when interpreting displacements of this object.

Probably the most popular attachment scheme currently in use is the biotin-streptavidin linkage. Microbeads are coated with streptavidin, and the structure to be studied, if a protein, can be easily functionalized with biotin via a succinimidyl ester or other chemical crosslinker. When mixed, the streptavidin on the bead tightly binds to biotin on the structure to be studied. Beads coated in this manner tend to stick to the surfaces in the experimental chamber, which is inconvenient for setting up an experiment, and recently developed neutravidin is an attractive alternative to streptavidin with decreased nonspecific binding at close to neutral pH. Alternative attachment schemes usually involve coating the bead with a protein that specifically interacts with the structure to be studied. For example, a microtubule-associated protein might be attached to the bead; subsequently that bead sticks to microtubules, which can then be manipulated with the optical tweezers. Recombinant DNA technology can also be used; in this case the amino acid sequence of a protein is modified to add a particular residue (e.g., reactive cystein), which serves as a target for specific crosslinking to the bead. This approach provides additional control over the binding orientation of the protein, but runs the risk that the recombinant protein many not behave as the wild type.

Interpreting displacements measured with optical tweezers is complicated by compliances in the system under study, or the linkage to the trapped bead; these must be accounted for to determine the actual displacement of specific elements. Figure 12 diagrams the compliances in a sample system. A filament to which the trapped bead is linked is pulled to the right by the force generating process under study (not shown). This causes the bead to be displaced to the right within the trap by an amount, Δ Bead, which is directly measured. However, the other compliance in the system, κ_{linkage}, is also stretched and takes up some amount of the filament displacement, which can only be determined with knowledge of the link stiffness; thus the measured displacement is less than the actual displacement of the filament. In some situations it may be possible to measure the link stiffness directly, but generally this is not possible during an experiment. Furthermore, the link stiffness is usually nonlinear, and may be complicated by

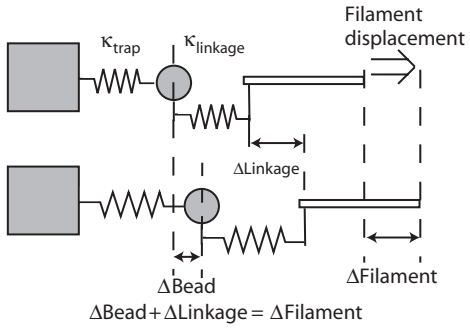


Figure 12. Schematic diagram of series compliances in optical tweezers experiments. Both the optical trap and the linkage between the bead and hypothetical filament are shown as springs. The optical trap is a well characterized spring, having been carefully calibrated prior to beginning experiments. The stiffness of the linkage, however, is unknown. When the filament is pulled to the right by a hypothetical motor or other biological system, both springs are stretched. Therefore the displacement of the trapped bead is not necessarily equal to the displacement of the filament by the motor. Examination of the system shows that the force on the trapped bead is the same as the force on the filament once all elements have reached their equilibrium. Knowledge of κ_{linkage} would allow for displacements at the bead to be used to calculate filament displacements.

the bead rocking about the link under forces applied by the trap. When precision displacement measurements are desired these difficulties can be circumvented with a feedback system that maintains a constant force on the particle (force clamp). By maintaining constant force, the link is held at a fixed strain and the movements of the feedback controlled laser directly follow the positional changes of the filament.

Motor Proteins: Kinesin And Myosin

The kinesins and myosins are two large families of motor proteins that convert chemical energy released by adenosine triphosphate (ATP) hydrolysis into mechanical work. Kinesins move cargo along microtubules while myosins exert forces against actin filaments, most notably in muscle. Many of the mechanical and kinetic aspects of these molecules behaviors have been determined from measurements made with optical tweezers. These include the length of displacements during individual chemomechanical steps, single molecule force generation capabilities, and kinetic information about the enzymatic cycle that converts chemical into mechanical energy.

Due to the differences in the substrate along which myosin and kinesin motors move, and the nature of their movements, assays for studying them have significantly different geometries. In the case of kinesin (Fig. 13), generally a single bead coated with the motor is held in an optical trap (17,26). The motors are sparsely coated on the beads, so that only one motor interacts with the microtubule track, which has been immobilized onto a glass surface. The optical tweezers manipulate the coated bead onto the microtubule; bead movements ensue when a kinesin motor engages the microtubule lattice. If the

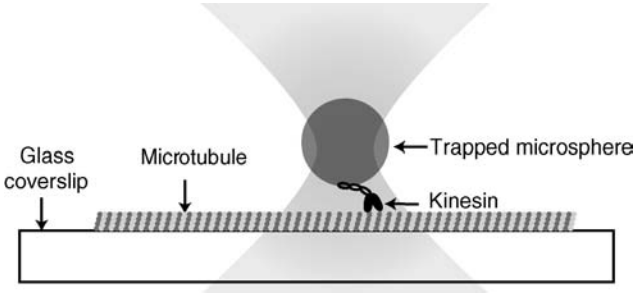


Figure 13. Schematic of an experiment to study kinesin. The motor protein kinesin is processive, meaning that the motor spends a large portion of its force generating cycle attached to the microtubule and is able to maintain movement when only a single motor is present. As a result, a single, sparsely coated bead held in an optical trap can be used to measure the force generating properties of kinesin.

optical tweezers are used in stationary mode, the records will indicate a fraction of the actual motor displacement, the remainder being taken up in the compliant link between the bead and the motor as described above. The actual displacements of the motor must then be inferred using estimates of the bead-motor link compliance, which is estimated in independent experiments (26).

Alternatively, a force clamp is applied to adjust the position of the trap to maintain a constant force on the bead as the kinesin motor travels along the microtubule. The movement of the motor is then inferred from the adjustments to the laser position, which directly reveal 8 nm steps, demonstrating that the kinesin molecule moves along the microtubule with the same periodicity as the microtubule lattice. Additional analysis of this data showed that kinesin stalls under loads of 5–8 pN, dependent on ATP concentration, and exhibits tight coupling between ATP hydrolysis and force generation (17,34).

Assays for studying myosin (Fig. 14) are somewhat more complicated, relying on two traps, each holding a microsphere, and a third large bead sparsely coated with myosin affixed to the surface of the microscope slide (35). An actin filament strung between the two smaller microspheres, each held by separate traps, is lowered into a position where myosin on the large fixed bead can pull against it. Myosin is not processive; it spends only a small fraction of the force generating cycle attached to the actin filament, and an unrestrained filament will diffuse away from the motor during the detached portion of the cycle. Thus the filament is held in close contact with the motor to allow repeated force generating interactions to be observed.

The actin filament and attached beads will move back and forth due to Brownian motion, which is limited by the drag on the particle and the force provided by the trap. This has two important consequences: the myosin molecule on the large bead is exposed to a number of possible binding sites along the actin filament, and attachment of the myosin cross bridge elevates the stiffness of the system sufficiently to greatly reduce the extent of thermally induced bead motion. With bandwidth sufficient to detect the full extent of microsphere Brownian motion, myosin binding events are identified by the reduction in the amplitude of the thermal motion. The distribution of positions of the trapped microspheres and filament at the onset of binding events is expected to be Gaussian of the same width as if the myosin-coated bead was absent. However, shortly after binding, the myosin motor displaces the filament and this shifts the center of the Gaussian by the distance of a single myosin step. Analysis of high resolution, high bandwidth traces of bead position with the above understanding lead to determination of the step length for single myosin subfragment-1 and heavy meromyosin: 3.5 and 5 nm, respectively (28).

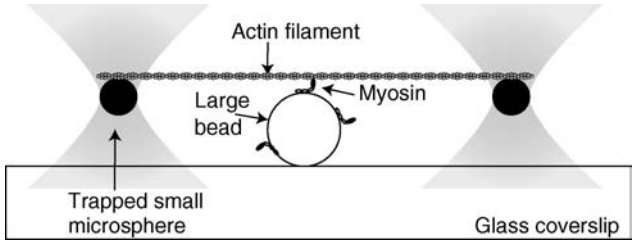


Figure 14. Schematic of an experiment to study muscle myosin. The motor protein myosin is not processive, meaning that a single copy of the motor cannot sustain movement along the actin filament. As a result an experimental arrangement that can keep the myosin in close proximity to the filament is necessary.

Other Motor Proteins

In addition to the classic motor proteins that generate forces against cytoskeletal filaments, proteins may exhibit motor activity, not as their *raison d'etre*, but in order to achieve other enzymatic tasks. One such protein is ribonucleic acid (RNA) polymerase (RNAP), the enzyme responsible for transcription of genetic information from deoxyribonucleic acid (DNA) to RNA. The RNAP uses the energy of ATP hydrolysis to move along the DNA substrate, copying the genetic information at a rate of 10 base pairs per second. The movement is directed, and thus constitutes motor activity. The assay used to study RNAP powered movement along the DNA substrate is shown in Fig. 15. A piece of single-stranded DNA attached to the trapped glass microsphere is allowed to interact with an RNAP molecule affixed to the surface of the slide. Optical tweezers hold the bead so that the progress and force developed by RNAP can be monitored. As the RNAP molecule moves along the

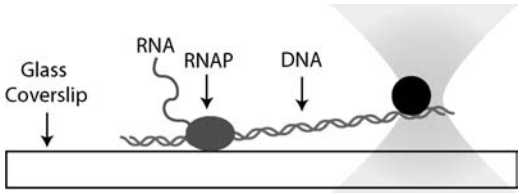


Figure 15. Schematic of an experiment to study RNA polymerase.

DNA, the load opposing the motor movement increases until the molecule stalls (36,37). Details of the kinetics of RNAP can be inferred from the relation between the force and speed of movement.

Nonstandard Trapping

A number of research efforts have studied the trapping of nonspherical objects and/or applied them to study biological processes. For example, the stiffness of a microtubule was measured using optical tweezers to directly trap and bend the rod-like polymer. Image data of the induced microtubule shape were then used in conjunction with mechanical elastic theory to determine the stiffness of the microtubule (38). Recently, there has been significant interest in using optical traps to exert torques. Generally, these techniques rely on nonspherical particles, non-Gaussian beams, or birefringent particles to create the necessary asymmetry to develop torque. One approach relies on using an annular laser profile (donut mode) with an interfering reference beam to create a trapping beam that has rotating arms (39). With such an arrangement torque generation is not dependent on the particle shape, but the implementation is relatively complicated. Two closely located traps can also be used to rotate objects by trapping them in multiple locations and moving the traps relative to one another. Using a spatial light modulator, which splits the beam into an arbitrary number of individual beams, to create and move the two traps allows rotation about an axis of choice (40).

For experimental applications, it is generally desirable to calibrate the torque exerted on the particles. The above techniques are not easily calibrated. A more readily calibrated altenative uses birefringent particles such as quartz microspheres (41). The anisotropic polarizability of the particle causes it to align to the beam polarity vector. The primary advantage of this technique is that the drag coefficient of the spherical particle is easily calculated, allowing torques can be calibrated by following rotational thermal motion of the particle, similarly to the techniques described for thermal motion calibrations above.

Modulation of trap position along with laser power or beam profile can create a laser line trap (42–44). The scheme is similar to time sharing a single laser between several positions to create several traps. However, to form a line trap the laser focus is rapidly scanned through positions along a single line. With simultaneous power modulation a linear trapping region capable of applying a single constant force to a trapped particle along the entire length of the trap is created. This technique can be used to replace the conventional force clamp relying on electronic feedback.

Some Other Optical Tweezers Assays

Optical tweezers have made numerous contributions to other subfields. A number of groups have utilized optical tweezers to study DNA molecules. Stretching assays have produced force extension relations for purified DNA, and stretching of individual nucleosomes revealed sudden drops in force indicative of the opening of the coiled DNA structure (45). Optical trapping has also been used to directly study the forces involved in packaging DNA into a viral capsid. Packaging was able to proceed against forces > 40 pN, and the force necessary to stall packaging was dependent on the how much of the DNA was already packaged into the capsid (46). Double-stranded DNA has also been melted (unzipped) by pulling the strands apart with optical tweezers: this established that lamba phage (a bacterial virus) DNA unzips and rezips in the range of 10–15 pN, dependent on the nucleotide sequence (47). Stretching RNA molecules to unfold loops and other secondary structures with a similar assay has been used to test a general statistical mechanics result known as the Jarzinsky Inequality (48).

Similar stretching experiments have been performed on proteins. A good example is the large muscle protein titin, which mediates muscle elasticity. Optical tweezers were applied to repeatedly stretch the molecule, providing evidence of mechanical fatigue of the titin molecule that could be the source of mechanical fatigue in repeatedly stimulated muscles (49).

Experiments that study interactions in larger, more complex systems are increasing common. Microtubules associated to mitotic chromosome kinetochores have been studied with optical tweezers. Forces of 15 pN were generally found to be insufficient to detach kinetochore bound microtubules and kinetochore attachment was found to modify microtubule growth and shortening (21). A number of studies have also applied optical tweezers to study structures inside intact cells and the force generating ability of highly motile cells such as sperm (6,50).

Beyond biological measurements, optical tweezers have utility in assembling micron scale objects in desired positions. Weakly focused lasers operating similarly to optical tweezers have been used to directly pattern multiple cell types on a surface for tissue engineering (55). Optical trapping has also been used for assembly and organization of nonbiological devices, such as groups of particles (51) and 3D structures, such as a crystal lattice (52).

Additionally, optical tweezers have been applied to great advantage in material and physical sciences. A substantial body of work has applied optical tweezers to study colloidal solutions and microrheology (53).

BIBLIOGRAPHY

1. Ashkin A. Acceleration and trapping of particles by radiation pressure. Phys Rev Lett 1970;24:156–159.
2. Ashkin A, Dziedzic JM, Bjorkholm JE, Chu S. Observation of a single-beam gradient force optical trap for dielectric particles. Opt Lett 1986;11:288–290.
3. Ashkin A, Dziedzic JM, Yamane T. Optical trapping and manipulation of single cells using infrared laser beams. Nature (London) 1987;330:769–771.
4. Ashkin A, Dziedzic JM. Optical trapping and manipulation of viruses and bacteria. Science 1987;235:1517–1520.
5. Block SM, Blair DF, Berg HC. Compliance of bacterial flagella measured with optical tweezers. Nature (London) 1989; 338:514–518.
6. Tadir Y, et al. Micromanipulation of sperm by a laser generated optical trap. Fertility Sterility 1989;52:870–873.
7. Block SM, Goldstein LS, Schnapp BJ. Bead movement by single kinesin molecules studied with optical tweezers. Nature (London) 1990;348:348–352.

8. Grover SC, Skirtach AG, Gauthier RC, Grover CP. Automated single-cell sorting system based on optical trapping. J Biomed Opt 2001;6:14–22.
9. Leach J, et al. 3D manipulation of particles into crystal structures using holographic optical tweezers. Opt Express 2004;12:220–226.
10. Ashkin A. Forces of a single-beam gradient laser trap on a dielectric sphere in the ray optics regime. Biophys J 1992;61: 569–582.
11. Visscher K, Brakenhoff G, Lindmo T, Brevik I. Theroretical study of optically induced forces on spherical particles in a single beam trap I: Rayleigh scatters. Optik 1992;89:174–180.
12. Nahmias YK, Odde DJ. A dimensionless parameter for escape force calculation and general design of radiation force-based systems such as laser trapping and laser guidance. Biophys J 2002;82:166A.
13. Nahmias YK, Gao BZ, Odde DJ. Dimensionless parameters for the design of optical traps and laser guidance systems. Appl Opt 2004;43:3999–4006.
14. Allersma MW, et al. Two-dimensional tracking of ncd motility by back focal plane interferometry. Biophys J 1998;74:1074–1085.
15. Brouhard GJ, Schek HT, Hunt AJ. Advanced optical tweezers for the study of cellular and molecular biomechanics. IEEE Trans Biomed Eng 2003;50:121–125.
16. Gittes F, Schmidt C. Interference model for back-focal-plane displacement detection in optical tweezers. Opt Lett 1998a;23:7–9.
17. Visscher K, Schnitzer MJ, Block SM. Single kinesin molecules studied with a molecular force clamp. Nature (London) 1999;400:184–189.
18. Ruff C, et al. Single-molecule tracking of myosins with genetically engineered amplifier domains. Nature Struct Bio 2001;8: 226–229.
19. Guck J, et al. The optical stretcher: A novel laser tool to micromanipulate cells. Biophys J 2001;81:767–784.
20. Smith SB, Cui YJ, Bustamante C. Overstretching B-DNA: The elastic response of individual double-stranded and single-stranded DNA molecules. Science 1996;271:795–799.
21. Hunt AJ, McIntosh JR. The dynamic behavior of individual microtubules associated with chromosomes in vitro. Mol Biol Cell 1998;9:2857–2871.
22. Kuo SC, Sheetz MP. Force of single kinesin molecules measured with optical tweezers. Science 1993;260:232–234.
23. Florin EL, Pralle A, Horber JKH, Stelzer EHK. Photonic force microscope based on optical tweezers and two-photon excitation for biological applications. J Struct Biol 1997;119:202–211.
24. Friese M, Rubinsztein-Dunlop H, Heckenberg N, Dearden E. Determination of the force constant of a single-beam gradient trap by measurement of backscattered light. Appl Opt 1996;35:7112–7116.
25. Pralle A, et al. Three-dimensional high-resolution particle tracking for optical tweezers by forward scattered light. Microsc Res 1999;44:378–386.
26. Svoboda K, Schmidt CF, Schnapp BJ, Block SM. Direct observation of kinesin stepping by optical trapping interferometry. Nature 1993;365:721–727.
27. Lang MJ, Asbury CL, Shaevitz JW, Block SM. An automated two-dimensional optical force clamp for single molecule studies. Biophys J 2002;83:491–501.
28. Molloy JE, et al. Movement and force produced by a single myosin head. Nature (London) 1995;378:209–212.
29. Visscher K, Brakenhoff GJ, Krol JJ. Micromanipulation by multiple optical traps created by a single fast scanning trap integrated with the bilateral confocal scanning laser microscope. Cytometry 1993;14:105–114.
30. Gensch T, et al. Transmission and confocal fluorescence microscopy and time-resolved fluorescence spectroscopy combined with a laser trap: Investigation of optically trapped block copolymer micelles. J Phys Chem *B* 1998;102:8440–8451.
31. Happel J, Brenner H. Low Reynolds number hydrodynamics. With special applications to particulate media. Leiden: Noordhoff International Publishing; 1973.
32. Gittes F, Schmidt C. Thermal noise limitations on micromechanical experiments. Eur Biophys J with Biophys Lett 1998b;27:75–81.
33. Meiners JC, Quake SR. Direct measurement of hydrodynamic cross correlations between two particles in an external potential. Phy Rev Lett 1999;82:2211–2214.
34. Schnitzer MJ, Block SM. Kinesin hydrolyses one ATP per 8-nm step. Nature (London) 1997;388:386–390.
35. Finer JT, Simmons RM, Spudich JA. Single myosin molecule mechanics - piconewton forces and nanometer steps. Nature (London) 1994;368:113–119.
36. Wang MD, et al. Force and velocity measured for single molecules of RNA polymerase. Science 1998;282:902–907.
37. Yin H, et al. Transcription against an applied force. Science 1995;270:1653–1657.
38. Felgner H, Frank R, Schliwa M. Flexural rigidity of microtubules measured with the use of optical tweezers. J Cell Sci 1996;109(Pt 2):509–516.
39. Paterson L, et al. Controlled rotation of optically trapped microscopic particles. Science 2001;292:912–914.
40. Bingelyte V, Leach J, Courtial J, Padgett MJ. Optically controlled three-dimensional rotation of microscopic objects. Appl Phys Lett 2003;82:829–831.
41. La Porta A, Wang MD. Optical torque wrench: Angular trapping, rotation, and torque detection of quartz microparticles. Phy Rev Lett 2004; 92.
42. Liesfeld B, Nambiar R, Meiners JC. Particle transport in asymmetric scanning-line optical tweezers. Phy Rev 2003; 68.
43. Nambiar R, Meiners JC. Fast position measurements with scanning line optical tweezers. Opt Lett 2002;27:836–838.
44. Nambiar R, Gajraj A, Meiners JC. All-optical constant-force laser tweezers. Bio J 2004;87:1972–1980.
45. Bennink ML, et al. Unfolding individual nucleosomes by stretching single chromatin fibers with optical tweezers. Nature Struct Biol 2001;8:606–610.
46. Smith DE, et al. The bacteriophage phi 29 portal motor can package DNA against a large internal force. Nature (London) 2001;413:748–752.
47. Bockelmann UP, et al. Unzipping DNA with optical tweezers: high sequence sensitivity and force flips. Biophys J 2000; 82:1537–1553.
48. Liphardt J, et al. Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski's equality. Science 2002;296:1832–1835.
49. Kellermayer MS, Smith SB, Granzier HL, Bustamante C. Folding-unfolding transitions in single titin molecules characterized with laser tweezers. Science 1997;276:1112–1116.
50. Aufderheide KJ, Du Q, Fry ES. Directed positioning of micronuclei in paramecium-tetraurelia with laser tweezers—absence of detectable damage after manipulation. J Eukaryotic Microbiol 1993;40:793–796.
51. Misawa H, et al. Multibeam laser manipulation and fixation of microparticles. Appl Phys Lett 1992;60:310–312.
52. Holmlin RE, et al. Light-driven microfabrication: Assembly of multicomponent, three-dimensional structures by using optical tweezers. Angew Chem Int Ed Engl 2000;39:3503–3506.
53. Lang MJ, Block SM. Resource letter: LBOT-1: Laser-based optical tweezers. Am J Phy 2003;71:201–215.

54. Berg-Sorensen, K & Flyvbjerg, H. (2004) Power spectrum analysis for optical tweezers. Review of Scientific Instruments 75, 594–612.
55. Odde, DJ & Renn, M.J. (2000) Laser-guided direct writing of living cells. Biotechnology and Bioengineering 67, 312–318.

See also FIBER OPTICS IN MEDICINE; MICROSURGERY; NANOPARTICLES.

ORAL CONTRACEPTIVES.

See CONTRACEPTIVE DEVICES.

ORTHOPEDIC DEVICES, MATERIALS AND DESIGN.

See MATERIALS AND DESIGN FOR ORTHOPEDIC DEVICES.

ORTHOPEDIC DEVICES MATERIALS AND DESIGN OF

AMIT BANDYOPADHYAY
SUSMITA BOSE
Washington State University
Pullman, Washington

INTRODUCTION

Musculoskeletal disorders are recognized as among the most significant human health problems that exist today, costing society an estimated $254 billion every year, and afflicting one out of seven Americans. Musculoskeletal disorders account for nearly 70 million physician office visits in the United States annually and an estimated 130 million total healthcare encounters including outpatient, hospital, and emergency room visits. In 1999, nearly 1 million people took time away from work to treat and recover from work-related musculoskeletal pain or impairment of function in the low back or upper extremities (1). There is still an ongoing debate on cause, nature and degrees of musculoskeletal disorders particularly related to work and how to reduce it. However, it is agreed unanimously that the number of individuals with musculoskeletal disorders will only increase over the coming years, as our population ages. According to the World Health Organization (WHO), these factors are called "work-related conditions", which may or may not be due to work exposures (1). Some of these factors include: (*1*) physical, organizational, and social aspects of work and the workplace, (*2*) physical and social aspects of life outside the workplace, including physical activities (e.g., household work, sports, exercise programs), economic incentives, and cultural values, and (*3*) the physical and psychological characteristics of the individual. The most important of the latter include age, gender, body mass index, personal habits including smoking, comorbidities, and probably some aspects of genetically determined predispositions (1). Among the various options to treat musculoskeletal disorders, use of orthopedic devices is becoming a routine, with the number of annual procedures approaching five million in the United States alone (2). Some of the common orthopedic devices include joint replacement devices for hip and knee and bone fixation devices such as pins, plates and screws for restoring lost structure and function.

Materials and design issues of orthopedic devices are ongoing challenges for scientists and engineers. Total hip replacement (THR) is a good example to understand some of these challenges. Total hip replacements are being used for almost past 60 years with a basic design concept that was first proposed by Charnley et al. (3). A typical lifetime for a hip replacement orthopedic device is between 10 and 15 years, which remained constant for the last five decades. From design point of view, a total hip prosthesis is composed of two components: the femoral component and the cup component. The femoral component is a metal stem, which is placed into the marrow cavity of the femoral bone, ending up with a neck section to be connected to the ball or head. The neck is attached to the head, a ball component that replaces the damaged femoral head. The implant can be in one piece where the ball and the stem are prefabricated and joined at the manufacturing facility, this is called a monobloc construction. It can also be in multiple pieces, called modular construction, which the surgeon put together during the time of the surgery based on patient needs, such as the size of the ball in the cavity. An acetabular component, a cup, is also implanted into the acetabulum, which is the natural hip socket, in the pelvic bone. The femoral component is typically made of metallic materials such as Ti or its alloys. The balls of the total prostheses are made either from metallic alloys or ceramic materials. The hip cups are typically made from UHMWPE (ultrahigh molecular weight polyethylene). Some part of the stem can be coated with porous metals or ceramics, which is called cementless implant, or used as uncoated in presence of bone cement, which is called cemented implants. Bone cements, which is primarily poly (methyl methacrylate) (PMMA) based, stabilizes the metallic stem in the femoral bone for cemented implants. However, for cementless implants, the porous coating helps in tissue bonding with the implants surface and this biological bonding helps implants to stabilize (Fig. 1).

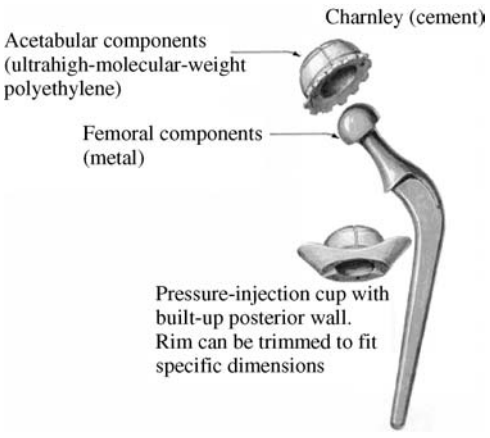


Figure 1. An example of a modern cemented hip prosthesis design and various components.

As it can be seen that THR by itself is a complex device that incorporates multiple materials and designs. However, all of these artificial materials mimic neither the composition nor the properties of natural bone. The inorganic part of natural bone consists of ~ 70 wt%, consists of calcium phosphate. Moreover, patient pool is different in terms of their age, bone density, physiological environment, and postoperative activities. To complicate matters further, surgical procedure and total surgery time can also be different for various patients. Because of these complications it is difficult, if not impossible, to point out the exact reason(s) for the low *in vivo* lifetime for these implants, which remained constant for the past 50 years. However, it is commonly believed that stress shielding is one of the key factors for limiting the lifetime of these implants. Because a metal stem is introduced into the bone, which has a complex architecture including an outer dense surface or cortical bone and an inner porous surface or cancellous bone, during the total hip surgery, the load distribution within the body shifts and the load transfer between tissue and implant does not match with normal physiological system. Typically, more load will be carried by the metal implants due to their high stiffness which will cause excess tissue growth in the neck regions. At the same time, upper part of the femoral bone will carry significantly less load and become weaker, which will make it prone to premature fracture. Both of these factors contribute to the loosening of hip implant that reduces the implant lifetime. This is also called stress-shielding effect, in general, which means the loss of bone that occurs adjacent to a prosthesis when stress is diverted from the area. To reduce stress-shielding, an ideal hip implant needs to be designed in a way that it has similar stiffness as natural femoral bone. However, current biomedical industries mostly use materials for load bearing implants that are typically designed and developed for aerospace or automotive applications, instead of developing new materials tailored specifically for orthopedic devices needs. But the time has come when materials need to be designed for specific biomedical applications to solve long-standing problems like stress-shielding in THR. Though THR is used as an example to show the complexity of materials and design issues in orthopedic devices but THR is not alone. Most orthopedic devices suffer from complex materials and design challenges to satisfy their performance needs.

FACTORS INFLUENCING ORTHOPEDIC DEVICES

There are several factors that need to be considered to design an orthopedic device. From the materials point of view, usually mechanical property requirement, such as strength, toughness, fatigue degradation becomes the most important issues as long as the materials are nontoxic and biocompatible. However, as the body tissue interacts with the surface of the device during *in vivo* lifetime, surface chemistry becomes one of the most important aspects for orthopedic devices. Most complex device functions cannot be accomplished using only one material, and require applications of structures made of multimaterials. As a result, compatibility of multimaterials in design and manufacture becomes another issue. Figure 2 summarizes various parameters that are important toward design and development of an orthopedic implants. Four main areas of orthopedic devices include materials issues, design issues, biocompatibility issues and patient specific information. All four of them are complex in nature and their interactions are even more difficult to appreciate. The following section offers some basic understanding of all of those issues.

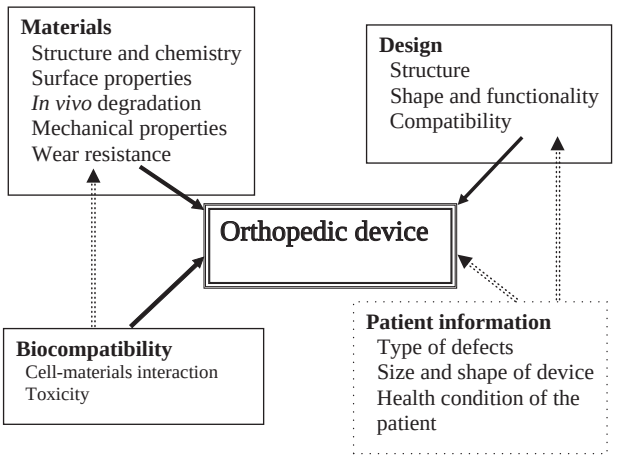


Figure 2. Materials and design parameters for orthopedic devices.

MATERIALS ISSUES IN ORTHOPEDIC DEVICES

Selection of appropriate material is probably the most important issue in successful design and development of orthopedic devices. Among various materials related issues, (*1*) physical and chemical properties, (*2*) mechanical properties, (*3*) surface properties, and (*4*) *in vivo* degradation or corrosion behavior are some of the most important ones. In general, it is most widely accepted to use metallic materials for load bearing, and polymers and ceramic–polymer composites for nonload bearing applications. Some ceramic compositions and glasses are also used for nonload bearing coatings and defect filling applications. Development of biomaterials and materials processing for different orthopedic device applications is currently a very active research area (4–6) and new materials are constantly being developed to meet the current and future needs.

Physical and Chemical Properties

Physical properties include density, porosity, particle size, and surface area type of information. Composition or chemistry is probably the most important chemical property. It is important to realize that orthopedic devices cannot be built with materials that are carcinogenic. Nontoxic materials that do not leach harmful metal ions *in vivo* are ideal. Apart from dense structures, partially or completely porous materials are also used for orthopedic devices. If a porous material is used, then some of the properties, such as pore size, pore volume, and pore–pore interconnectivity

become important. Typically, for most porous materials, an optimum pore size between 100 and 500 μm are used in which cells can grow and stay healthy. Higher pore volume usually adds more space for cells to grow and naturally anchor the device. However, this also exposes higher surface area of the device material that can cause faster degradation or corrosion, which sometimes can be a concern depending on the materials used. For polymeric materials, chemistry, and structure are important because materials with the same chemistry, but different structure can show different *in vivo* response. This is particularly important for biodegradable polymers, such as poly lactic acids (PLA) and polyglycolic acids (PGA) and their copolymers. Trace element is another important factor in materials selection. Sometimes even a small amount of impurities can cause harmful effects *in vivo* (7). However, in calcium phosphate based ceramics, small addition of impurity elements actually proved to be beneficial for mechanical and biological response (8).

Mechanical Properties

Mechanical properties are important in selecting materials for orthopedic devices. Among various mechanical properties, uniaxial and multiaxial strength, elastic modulus, toughness, bending strength, wear resistance, fatigue resistance are some of the most important ones. Mechanical property requirements are tied to specific applications. For example, for the stem in THR, it should have high strength, low modulus, and very high fatigue resistance. As a result, due their low modulus, Ti and its alloys are usually preferred over high modulus metal alloys for the stem part of THR. However, for the acetabular component, high wear resistance requirement is the most important one and high-density polymers are preferred for the acetabular component. Mechanical properties are also linked on how they are processed. For example, casting devices in their near final shape can be a relatively inexpensive way to make complex shapes. However, material selection is critical in casting. The use of cast stainless steel for femoral hip stems is one experience that led to a high failure rate. This result generated significant debate in 1970s regarding processing of load bearing implants using casting and the options were considered by the ASTM Committee F04 on Medical and Surgical Materials and Devices to ban cast load bearing implants (8). For devices made of degradable polymers, strength loss due to degradation is an important factor. For example, materials compositions and structures in resorbable sutures are designed for different degradation times to achieve desired strength loss characteristics.

In vivo Degradation

Some orthopedic devices require materials to be bioresorbale or biodegradable, which will dissolve in body fluid as natural tissue repairs the site. Except for a few polymer and ceramic compositions, most materials are nondegradable in Nature. The degradation behavior is controlled by three basic mechanisms and they are (*1*) physiologic dissolution, which depends on pH and composition of calcium phosphate; (*2*) physical disintegration, which may be due to biochemical attack at the grain boundaries or due to high porosity; and (*3*) biological factors, such as phagocytosis.

In most materials, not just one mechanism but a combination of all three mechanisms control biodegradation behavior. Among them, biological factors are probably the most interesting ones. Though the actual process is quite complex (9), a simplistic action sequence can be viewed as osteoclastic cells slowly eat away the top surface of the foreign material and stimulate osteoblast cells. Osteoblast cells then come and deposit new bone to repair the site. Such dynamic bone remodeling is a continuous process within every human body. The rate at which osteoblastic deposition and osteoclastic resorption are taking place changes with the age of the person. This process controls the overall bone density. In terms of materials, *in vivo* degradation of polymeric materials is probably the most well-characterized field. Numerous products are available in which degradation kinetics has been tailored for specific applications. However, the same is not true for ceramics. Controlled degradation ceramics are not commercially available though degradation behavior of some calcium phosphate based ceramics is well documented (10). For metallic implants, the most serious concern regarding *in vivo* degradation is metal ion leaching or corrosion of the implants, which can cause adverse biological reactions. Corrosion products of nickel, cobalt, and chromium can form metal–protein complexes and lead to allergic reactions (7,11). Early reports of allergic reactions were reported with metal on metal (MOM) total hips (12). The inflammatory response to metallic wear debris from these devices may have been enhanced due to the high corrosion rate of the small wear particles. However, there is a lack of a predictable relationship between corrosion and allergies except in a few cases, such as vitallium implants (13,14). The number of patients with allergic reactions is not large, and it remains to be proven whether corrosion of devices causes the allergy, or the reactions are only manifest in patients with preexisting allergies. Most materials that are currently used in load bearing dental and orthopedic devices are plates and screws and they show minimum long-term degradation and health related concerns such as allergies.

Surface Properties

Surface property of materials is another important parameter for orthopedic device design. Once implanted, it is the surface that the body tissue will see first and interact. As a result, surface chemistry and roughness both are important parameters for device design. Devices that are designed for different joints, where wear is a critical issue, smooth surface is preferred there. For example, in knee joints UHMWPE is used to reduce wear debris. But most other places, where tissue bonding is necessary, rough surface or surface with internal porosity is preferred primarily to enhance physical attachment. However, biomechanical and biochemical bonding to device surfaces are still subject of active scientific investigation (15). Tailoring internal porosity and chemistry of metallic implants is still an active research area. Either metal on metal or ceramic on metal coatings are used to achieve this goal. Different

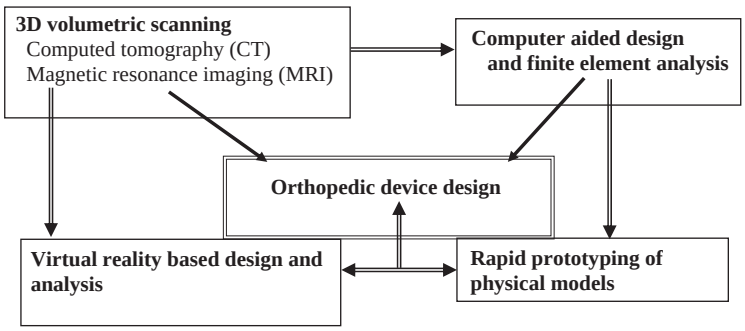


Figure 3. Recent trends in orthopedic device design.

manufacturing techniques are used to modify surfaces of metal implants. Among them, partial sintering of metal powders on metallic implants is one approach (16). For example, Ti powders are sintered on cp-Ti or Ti6Al4V devices, where after sintering the sintered layers leave some porosity in the range of 100–500 μm for osteointegration by bone tissues. Similar techniques are also used for metal fibers to create a mesh with varying porosity for improved osteointegration (17). For ceramic coatings, such as calcium phosphate on Ti or Co–Cr based devices, plasma-spray technique is used. Typically, a direct current (DC) plasma gun is used under a controlled environment to coat metallic device with several hundred microns of calcium phosphate-based ceramics. Because of significantly better bioactivity of calcium phosphates, these coated devices show improved tissue–materials interaction and better long-term tissue bonding (18). In case of THRs, ceramics coated cementless implants are placed without any bone cement during surgery. During healing, body tissue forms strong bonds with the coating and anchors the device. This biological fixation is believed to be equal or better than cemented implants in which bone cement is used during surgery to anchor the device. Though the idea of cementless implants is great, but lack of interfacial strength at the implant metal and ceramic coating interface is still a concern and subject of active research. Formation of amorphous calcium phosphate during processing of plasma-sprayed ceramic coatings increase potential biodegradation rate for the coating material, which is another major concern for these devices. In general, though coated implants are promising, a significant number of uncoated implants are still used in surgery every day that has worked for a long time. In fact there are more research data available today on uncoated implants than on the coated ones.

DESIGN ISSUES IN ORTHOPEDIC DEVICES

Design of orthopedic devices is focused on the needs for that particular problem. As a result, for the same device (spinal grafting cage or THR), different designs can be found from various device makers. These devices can be in single piece or multiple-piece, made from the same or different material(s). As a result challenges are significantly different for multiple piece multimaterial devices like THR than single piece ones like bone screws or plates. For multiple piece multimaterial devices, compatibility among different materials/pieces and overall functionality becomes a more serious design issue.

Current practice in device design usually starts from biomechanical analysis of stress distribution and functionality of a particular device. If it is a joint related device, it is important that the patient can actually move the joint along multiple directions and planes to properly restore and recover functionality of that joint. Figure 3 shows some of the recent trends in orthopedic device designs. During the past 10 years, computer aided design (CAD) and rapid prototyping (RP) based technologies have played a significant role in orthopedic device design. Using this approach, real information from patients can be gathered using a computed tomography (CT) or magnetic resonance imaging (MRI) scans, which then can be visualized in three dimensions (3D). This 3D data can be transformed to a CAD file using different commercially available software.

The CAD file can be used to redesign or modify orthopedic devices that will be suitable to perform patient's need. If necessary, the device can also be tested in a virtual world using finite element analysis (FEA) to optimize its functionality. Optimized device can then be fabricated using mass manufacturing technologies such as machining and casting. If small production volume is needed, then RP technologies can be used. In RP, physical objects can be directly built from a CAD file without using any part specific tooling or dies. Rapid Prototyping is an additive or layer by layer manufacturing process in which each layer will have a small thickness, but the X and Y dimensions will be based on the part geometry. Because no tooling is required, batches as small as 1 or 2 parts can be economically manufactured. Most RP processes are capable of manufacturing polymer parts with thermoset or thermoplastic polymers. Some of the RP techniques can also be used to manufacture metal parts. Figure 4 shows a life-sized human femur made of Ti6Al4V alloy using laser engineered net shaping (LENS) process. The LENS technology uses metal powders to create functional parts that can be used in many demanding applications. The process uses up to 2 kW of Nd:YAG laser power focused onto a metal substrate to create a molten puddle on the substrate surface. Metal powder is then injected into the molten puddle to increase the material volume. The substrate is then scanned relative to the deposition apparatus to write lines of the metal with a finite width and thickness. Laser engineered net shaping is an exciting technology for orthopedic devices because it can directly build functional parts that can be used for different applications instead of

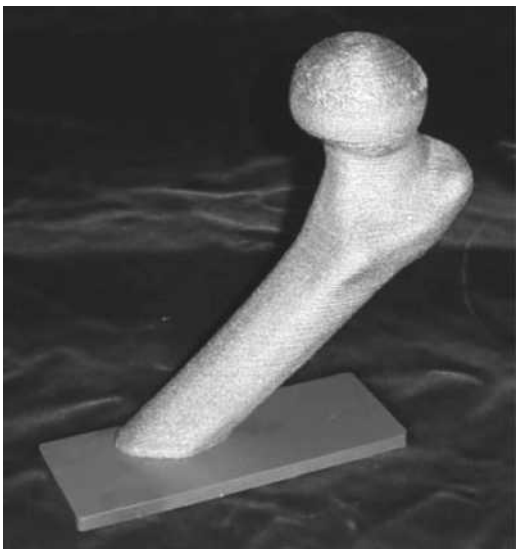


Figure 4. A life-sized human femur made of Ti6Al4V alloy using LENS.

other RP parts that are typically used for "touch and feel" applications. Commercial RP processes have also been modified to make ceramic parts for orthopedic devices. Figure 5 shows one such example in which reverse engineering of horse's-knuckle was used to show how fused deposition modeling (FDM), a commercial RP process, can be used to create tailored porosity bone implants in which pore size and pore volume can be varied simultaneously keeping the outside geometry constant (19,20). The FDM process was used to make polymer molds of to cast porous ceramic structures. The mold was designed from the CAD file of the horse's knuckle. The CAD file was created from the 3D volumetric data received from the CT scan of the bone. Such examples demonstrate the feasibility of patient specific implants through novel design and manufacturing tools.

BIOCOMPATIBILITY ISSUES IN ORTHOPEDIC DEVICES

Biocompatibility issue is an important issue in orthopedic device design and development, but it is usually considered during materials selection and surface modification. From cell materials interaction point of view, materials can be divided into three broad categories: (*1*) toxic; (*2*) nontoxic and bioinert, and (*3*) nontoxic and bioactive. For any application in the physiological environment, a material must be nontoxic. A bioinert material is nontoxic, but biologically inactive such as Ti metal. A bioactive material is the one that elicits a specific biological response at the interface of the biological tissue and the material, which results in formation of bonding between tissue and material. An

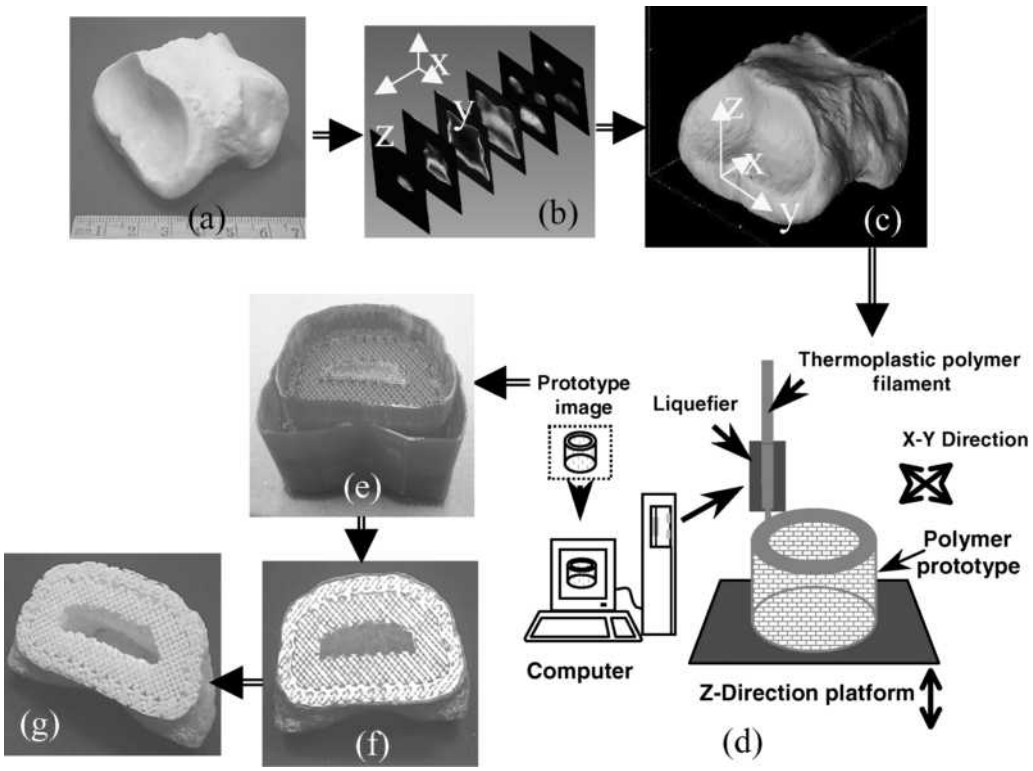


Figure 5. (a) A real bone; (b) CT scans of the bone; (c) a CAD file from the CT scans; (d) FDM process. (e) a polymer mold processed via FDM of the desired bone; (f) alumina ceramic slurry infiltrated polymer mold; (g) controlled porosity alumina ceramic bone graft.

example of bioactive material will be hydroxyapatite, $Ca_{10}(PO_4)_6(OH)_2$, which is similar in composition to the inorganic part of natural bone. The terms biodegradable, bioresorbable, and bioabsorbable are often used interchangeably. For most orthopedic devices, bioactive surfaces are ideal for better cell–materials attachment.

In vivo cell–materials interaction is a fairly complex process. In simple terms, when orthopedic devices are placed inside, our body will try to isolate the device by forming a fibrous tissue around it, which is particularly true for devices made with bioinert materials. If the material is bioactive, bone cells will first attach to the implant surface and then grow or proliferate. A material shows good biocompatibility when cells will attach and grow quickly on the device surface. Growth factors, such as bone morphogenic proteins (BMP), are sometimes used to stimulate cell attachment and growth behavior *in vivo*. After proliferation, cells will differentiate or produce new bones, which will repair the site and anchor the device. All three stages, attachment, proliferation, and differntiation of bone cells are important for repair and reconstruction of musculoskeletal disorders.

PATIENT SPECIFIC ORTHOPEDIC DEVICE: A FUTURE TREND

Patient specific device is the future trend for repair and reconstruction of musculoskeletal disorders. Due to the advancement of CAD and RP based small volume manufacturing technologies, orthopedic devices for people with special needs will be designed to meet the specific requirements. Such activities are currently pursued in academic research and hope to translate into standard industrial practice in the next one or two decades. Three-dimensions (3D) visualization of disorders using computer images and physical models, and follow-up discussion between patients and physicians will help to better educate the patient population about their problems and possible options. These options can then be transformed to physical models for trials before placing them *in vivo*. The goal is to reduce revision surgeries and increase device lifetime while maintaining the quality of life for the patient population. Various innovative scientific and engineering advancements toward novel materials and design options are helping us to make significant developments to achieve this goal.

BIBLIOGRAPHY

Cited References

1. Musculoskeletal Disorders and the Workplace: Low Back and Upper Extremities. New York: National Academic Press; 2001.
2. Materials science. Encyclopædia Britannica. 2005. Encyclopædia Britannica Premium Service.
3. Charnley J, Kamangar A, Longfield MD. The optimum size of prosthetic heads in relation to the wear of plastic sockets in total replacement of the hip. Med & Biol Eng 1969;1:31–39.
4. Hench LL. Bioceramics: From concept to clinic. J Amer Ceram Soc 1991;74(7):1487–510.
5. Burg KJL, Porter S, Kellam JF. Biomaterials development for bone tissue engineering. Biomaterials 2000;21:2347–2359.
6. Berndt CC, Haddad GN, Farmer AJD, Gross KA. Thermal spraying for bioceramic applications. Mater Forum 1990 14:161–173.
7. Mayor MB, Merritt K, Brown SA. Metal allergy and the surgical patient. Am J Surg 1980;139:447–479.
8. Chao EYS, Coventry MB. Fracture of the femoral component after total hip replacement. J Rone Joint Surg 1981;63A:1078–1094.
9. Roodman GD. Mechanisms of bone metastasis. N Eng J Med 2004;350(16):1655–1664.
10. Ravaglioli A, Krajewski A. Bioceramics: Materials, Properties, Application. London: Chapman and Hall, 1992. pp. 156–197.
11. Fontanna MG, Greene ND. Corrision Engineering. New York: McGraw-Hill; 1978.
12. Evans EM, Freeman MAR, Miller AJ, Vemon-Roberts B. Metal sensitivity as a cause of bone necrosis and loosening of prostheses in total joint replacements. J Bone Joint Surg 1974;56B:626–642.
13. Halpin DS. An unusual reaction in muscle in association with Vitallium plate: A report of possible metal hypersensitivity. J Bone Joint Sur Br 1975;57(4):451–453.
14. Garcia DA. Biocompatibility of dental implant materials measured in an animal model. J Dental Res. 1981;60(1):44–49.
15. Roberts WE. Osseous adaptation to continuous loading of rigid endosseous implants. Am J Orthod 1984;86(2):95–111.
16. Pilliar RM. Powder metal-made orthopaedic implants with porous surfaces for fixation by tissue ingrowth. Clin Orthop 1983;176:42–51.
17. Ducheyne P, Martens M. Orderly oriented wire meshes as porous coatings on orthopaedic implants II: The pore size, interfacial bonding and microstructure after pressure sintering of titanium OOWM. Clin Mats 1986;1:91–98
18. Ducheyne P, Qiu Q. Bioactive ceramics: the effect of surface reactivity on bone formation and bone cell function. Biomaterials 1999;20:2287–2303.
19. Darsell J, Bose S, Hosick H, Bandyopadhyay A. From CT Scans to Ceramic Bone Grafts. J Am Ceramic Soc 2003; 86(7):1076–1080.
20. Bose S, et al. Pore Size and Pore Volume Effects on Calcium Phosphate Based Ceramics. Mat Sci Eng 2003; C 23:479–486.

ORTHOPEDICS PROSTHESIS FIXATION FOR

PATRICK J. PRENDERGAST
Trinity Centre for
Bioengineering
Dublin, Ireland

INTRODUCTION

The fixation of an orthopedic implant should secure it rigidly to the underlying bone. The ideal fixation will sustain high forces, pain free, for the remaining lifetime of the patient. Difficulties in achieving this objective arise because (1)

1. The loads are often several times body weight in the lower extremity.The loads are fluctuating, or cyclic, and furthermore extremely high loads can occur occasionally (2).

2. The presence of the implant alters the stress transfer to the underlying bone leading to bone remodelling or fibrous tissue formation at the bone/implant interfaces. This can threaten the long-term mechanical integrity of the prosthetic replacement.
3. The range of materials that can be placed in contact with bone is limited by biocompatibility issues.

The fixation of an orthopedic implant may be catagorized as either cemented fixation or biological fixation.

Cemented fixation involves securing the implant into the bone with a "bone cement." By far the most common bone cement is based on the (polymer polymethylmethacrylate (PMMA)). PMMA bone cement is polymerized *in situ* during the surgery. It contains radiopacificiers in the form of particles of barium sulphate ($BaSO_4$) or zirconia (ZrO_2), which make it visible in radiographs (3). It also contains an inhibitor (hydroquinone) to prevent spontaneous polymerization and an initiator (benzoyl peroxide) to allow polymerization at room temperature. Antibiotics to prevent infection (e.g., gentimacin) may also be added. Table 1 lists typical components of bone cement and their roles. Polymerization begins when a powder of the PMMA polymer is mixed with the MMA monomer liquid. The mixing can either be done by hand in a mixing bowl just before to its use in the surgery or a mechanical mixing system may be used; these have the advantage of reducing the porosity of the bone cement and increasing its fatigue life. The cement is applied in a doughy state to the bone before placement of the implant.

In biological fixation, the implant is secured to the bone by a process known as "osseointegration." Osseointegration occurs by bone ingrowth onto the surface of the implant. The surface of the implant must have a structure so that, when the bone grows in, sufficient tensile and shear strength is created. Bone ingrowth requires a mechanically stable environment and an osteoconductive surface. An osteoconductive surface can be achieved by various treatments, e.g., plasma spraying with hydroxyapatite. Ingrowth occurs over approximately 12 weeks, and during this period, implant stability is required: Initial stability can be achieved by press-fitting the implant into the bone, or by using screws.

Hybrid fixation refers to the use of both cemented and biological techniques for the fixation of a prosthesis. For example, a hip replacement femoral component may be fixated using cement, whereas the acetabular cup may be fixated into the pelvic bone by osseointegration.

Failure of prosthesis fixation is observed as loosening and pain for the patient. If loosening occurs without infection it is called *aseptic* loosening. Loosening is a multifactorial process and does not have just one cause. Loosening of cemented fixation often occurs by fatigue failure of the bone cement, but loosening can have several root causes: fatigue from pores in the cement and stress concentrations at the implant/cement interface, debonding at the prosthesis/cement interface or cement/bone interface, or bone resorption causing stresses to rise in the cement. Loosening of biological fixation occurs if the relative micromotion between the bone and the implant is too high to allow osseointegration, i.e., if the initial stability of the implant is insufficient. Huiskes (4) proposed the concept of *failure scenarios* as a method for better understanding the multifactorial nature of aseptic loosening. The failure scenarios are

1. Damage accumulation failure scenario: the gradual cracking of bone cement, perhaps triggered by interface debonding, pores in the cement, or increased stresses due to peripheral bone loss.
2. Particulate reaction failure scenario: wear particles emanating from the articulating surfaces or from metal/metal interfaces in modular prostheses (fretting wear) can migrate into the interfaces causing bone death (osteolysis).
3. Failed ingrowth failure scenario: High micromotion of the implant relative to the bone can prevent bone ingrowth, as can large gaps (> 3 mm). If the area of ingrowth is insufficient, then the strength of the fixation will not be high enough to sustain loading when weight-bearing commences.
4. Stress shielding failure scenario: Parts of the bone can be "shielded" from the stresses they would normally experience because of the rigidity of the implant. This can lead to resorption of the bone and degeneration of the fixation.
5. Stress bypass failure scenario: In biological fixation, ingrowth can be patchy leading to stress transfer over localized areas. When this happens, some bone tissue is "bypassed," and in these regions, bone atrophy can occur because the stress is low.

Table 1. Components of Bone Cement and Their Roles

Components	Role	Amount
Liquid		20 mL
Methyl methacrylate (monomer)	Wetting PMMA particles	97.4 v/o
N,N,-dimethyl-p-toluidine	Polymerization accelerator	2.6 v/o
Hydroquinone	Polymerization inhibitor	75 + 15 ppm
Solid powder		40 g
Polymethyl methacrylate	Matrix material	15.0 w/o
Methyl methacrylate-styrene-copolymer	Matrix material	75.0 w/o
Barium sulphate ($BaSO_4$), USP	Radiopacifying agent	10.0 w/o
Dibenzoyl peroxide	Polymerization initiator	0.75 w/o

From Park (3).
Note: v/o: % by volume; w/o: % by weight.

6. Destructive wear failure scenario: In some joint replacement prostheses, e.g., hip and knee, wear can occur to such a degree that the component eventually disintegrates.

CEMENTED FIXATION

It is common to classify cementing techniques according to "generation": The first generation involved hand-mixing and finger packing of the cement, and the second generation improved the procedure by using a cement gun and retrograde filling of the canal, with a bone-plug to contain the cement within the medullary canal. This allows pressurization of the cement and therefore better interdigitation of the cement into the bone. Third generation (called modern cementing) uses, in addition, mechanical mixing techniques for the cement to remove pores and pulsative lavage to clean the bone surface of debris. The most common mechanical mixing technique is "vacuum mixing," where the powder and monomer are placed together in a mixing tube and the air is removed under pressure; often the tube can then be placed into an injection gun from which it can be extruded into the bone cavity. Another mechanical mixing technique is centrifugation (i.e., spinning the polymerizing bone cement in a machine), which is found to remove pores and increase the fracture strength (3). Precoating the implant with a PMMA layer or addition of a roughened surface strengthens the implant/bone cement interface.

Fixation strength using bone cement relies on an interdigitation of the bone cement with the bone; i.e., it is a mechanical interlock between the bone and the solidified cement that maintains the strength and not a chemical bond. Good interdigitation requires that the bone bed be rough. Creating a rough surface is done by appropriate broaching during preparation of the bony bed; it also requires lavage to clean the bed of loose debris and marrow tissue. Mann et al. (5) found the strength of the bone cement/bone interface to be positively correlated with the degree of interdigitation (Fig. 1). To achieve superior

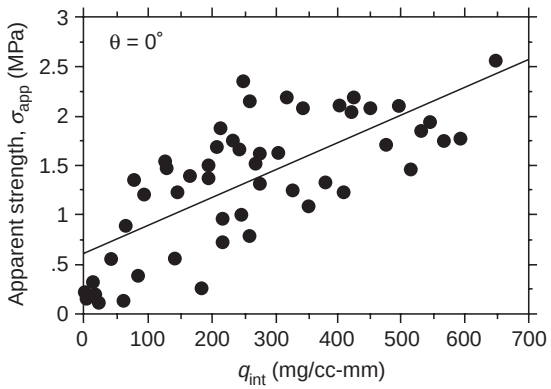


Figure 1. Interdigitation of the bone cement into the bone increases the strength of the bone cement interface. q_{int} is the product of the average value of the thickness of the interdigitated region and the density of the interface region measured using a CT scan. See Mann et al. (5) for details.

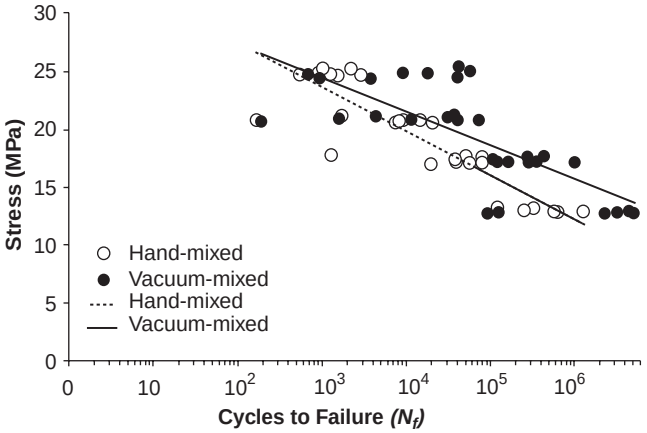


Figure 2. A comparison of the fatigue strength of hand-mixed and vacuum-mixed bone cement. After Murphy and Prendergast (10).

interdigitation, it was thought useful to develop "low viscosity" cements, and although higher penetration was achieved, the clinical outcomes using low viscosity cements in hips were not superior (6).

PMMA bone cement undergoes an exothermic polymerization reaction. This means that heat is produced on polymerization and this can cause necrosis of the surrounding bone tissue. Another consequence of heating is that the cement expands and contracts on cooling. As solidification occurs before to full cooling, residual stresses are generated in the cement (7). This is one reason to minimize the thickness of the cement layer. Also, metallic stems, because they conduct heat, can minimize the peak temperature transmitted to the bone, cooling the metallic implant before implantation has also been suggested. Bioactive cements have also been proposed; see the review by Harper (8). These cements have filler particles added to create a bioactive surface on the cement; fillers can be hydroxyapatite powder or fibers, bone particles, or human growth hormone. Alternatives to PMMA are bisphenol-*a*-glycidyl methacrylate (BIS-GMA) or poly(ethylmethacrylate) (PEMA)/*n*-butylmethacrylate (*n*BMA) cement. However, these cements are not yet widely used.

The mechanical strength depends on the brand of cement used and on the mixing technique (9). To prevent the damage accumulation failure scenario (see above), sufficient fatigue strength is required. This has been measured as a function of mixing technique (Fig. 2) (10). Being a polymer operating close to its melting temperature, bone cement is also subject to creep, i.e., viscoplasticity, and the creep strain as a function of stress has been measured under dynamic loading (11). However, it is clear that the *in vitro* testing conditions may not account for many of the extremely complex *in vivo* conditions, so these results should be interpreted with caution (12).

OSSEOINTEGRATION (CEMENTLESS FIXATION)

There is no simple definition of osseointegration, although Albrektsson (13) advocates the following: Osseointegration means *a relatively soft-tissue-free contact between implant*

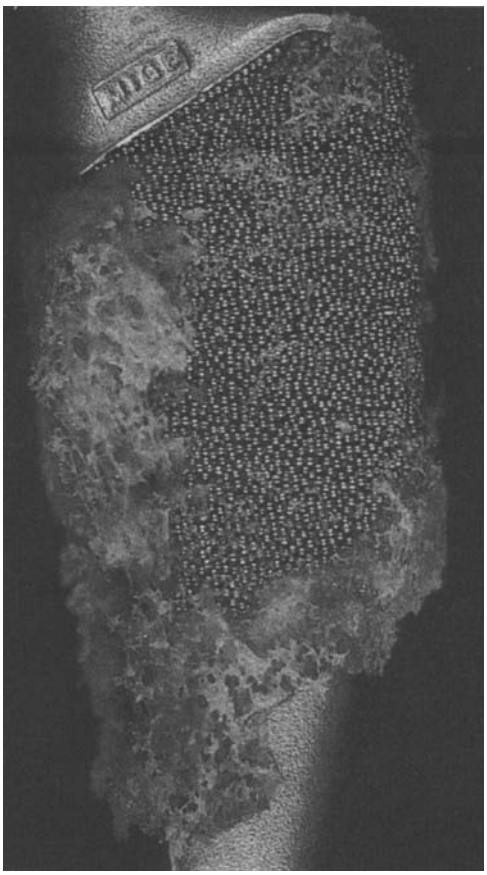


Figure 3. Bone ingrowth into a multilayer of a proximal part of a femoral hip prosthesis. After Eldridge and Learmonth (15).

and bone, leading to a clinically stable implant. Early in the study of the osseointegration concept, Skalak (14) found osseointegration was promoted by a micro-rough surface more so than a smooth one. Since then, many animal experiments investigating the effect of plasma spraying the surface and various methods of creating a porous surface have been reported. For orthopedic fixation, porous surfaces with beads in one or more layers have been used, as have wire meshes attached to surfaces, and plasma spraying the surface with hydroxyapatite.

Figure 3 shows bone ingrowth into a multilayer of a proximal part of a femoral hip prosthesis (15). It can be observed that ingrowth is patchy; this is what is commonly found, even with successful implants retrieved at autopsy (16); it is evident, therefore, that ingrowth is not required everywhere on the prosthesis for a successful fixation. Ingrowth is controlled by a combination of the mechanical environment and the size of the pores; the spacing between the pores should not be greater than the degree of micromotion or else the new bone ingrowth path will be continuously sheared as the tissue attempts to grow in. In experiments in dogs, Søballe (17) studied the relationship between implant coating and micromotion and found that hydroxyapatite coating increased the rate of bone ingrowth, and that a relative motion between implant and bone of 150 μm allowed osseointegration, whereas a relative motion of 500 μm inhibited it. The mechanobiological consequences of these different shearing magnitudes was analyzed by Prendergast et al. (18). The depth of the porosity will also affect the strength, with multilayer beaded surfaces having the potential for greater tensile strength (1).

FIXATION OF PROSTHESIS DESIGNS

Each implant design has specialized fixation features. In the following sections, examples are provided of the fixation approaches used in the main orthopedic implant categories.

Hip Prostheses

Although hip arthroplasty may involve replacement of the femoral side only, *total* hip arthroplasty (THA) involves replacement of the proximal femur and the acetabular socket. Both cemented and cementless fixation is used for both the femoral component (the "stem") and the acetabular component (the "cup"). Selection is a matter of surgeon choice, although there is some agreement that the cementless fixation is preferable in younger patients because cementless implants are easier to revise than cemented where complete removal of the cement mantle may be problematic.

Considering the femoral side first, cemented fixation takes two categories: stem designs in which a bond is encouraged between the stem and the cement (referred here as bonded stems) and designs that discourage a bond (referred here as unbonded stems). Stem bonding can be achieved through roughening of the stem surface to create a mechanical interlock between the metallic stem or cement or through use of a PMMA precoat to create a chemical bond between the precoat/cement interface. Bonded stems usually contain a collar that rests on the bone surface preventing subsidence and often containing ridges, dimples, and undercoats to provide additional interlock with the cement. As long as the bonded stems remain bonded, they have the theoretical benefit of reducing the stress levels in the cement. However, if the bonded stems fail, the roughened surface could generate debris particles and initiate a loosening process. In contrast to the bonded stems, unbonded stems discourage a bond between the stem and the cement through use of a smooth, polished stem surface in combination with a stem design that typically has no collar or macrofeatures to lock into the cement. With the lack of a bond, the polished stems facilitate some stem subsidence within the cement mantle and thereby allow wedging of the implant within the medullary canal. Lennon et al. (19) compared the damage accumulation around polished with matt stems and did not find a difference in the damage accumulated in their cement mantles. Another point of comparison between cemented and cementless fixation is that cementless stems will have a larger cross-sectional area than cemented stems because they must fill the medullary canal; this will make cementless hip prostheses stiffer and predispose them to the stress shielding failure scenario. Recognizing this, it is usual for the osseointegration surface to be on the proximal part of cementless stems to ensure proximal load transfer;

furthermore, patches of osseointegrative surface may be limited to the posterior and anterior faces of the stem.

Considering the acetabular side, the cup is either made from ultra-high-molecular-weight polyethylene (UHMWPE), ceramic, or metal. UHMWPE cups may be metal-backed. As the head can be either ceramic (a modular head can be connected to a metal femoral component using a Morse taper) or metal (modular or monobloc), this means that several combinations of bearing materials are possible. Polyethylene cups and metal heads are the most common, but the others, such as metal-on-metal, are advocated as well. The selection of bearing materials is important for the fixation because a high frictional torque predisposes to loosening of the cup or stem and because the wear particles produced can provoke the particulate reaction failure scenario. Polyethylene cups are cemented into the acetabulum using bone cement. Metal-backing of the cup is designed to decrease stresses in the polyethylene "liner," which should lead to lower wear rates although it is also predicted to increase stress concentrations in the fixation at the periphery of the cup (20). Metal, ceramic, and metal-backed UHMWPE cups may be threaded on the outside so that they can be screwed into the acetabelum, or they may be fixated by osseointegration.

The interrelationship between design factors and fixation of hip implants is complicated and involves maximizing strength of the cement/metal interface, the cement itself, and the bone/cement interface. According to the design philosophy of polished stems, it is better to safeguard the vital bone/cement interface by allowing the cement/metal interface to fail first and facilitating subsidence (21). Not only should the interfaces have the required strength, but the stresses should be minimized to ensure the most durable fixation (22)— the measures to achieve this are listed in Table 2.

Figure 4. A knee replacement prosthesis showing porous-coating for osseointegration and posts for fixation. From Robinson (23).

Knee Prostheses

Total knee replacement involves femoral and tibial components, and a component for patellar resurfacing (a patellar "button") is also often used. Both cemented and cementless fixation is used in knee arthroplasty. The femoral component may be fixated with an intramedullary stem that may be cemented, or it may have a porous surface for osseointegration with medial and lateral "posts" to aid initial stability. The tibial component consists of a metal "tray" and a polyethylene insert; the tray may also be fixated with an intramedullary stem cemented into the tibia, perhaps accompanied by medical and lateral posts/pegs for rotational stability. Figure 4 shows a design fixated by osseointegration (23). Walker (24) gives a thorough description of the options available for knee prostheses.

Upper Extremity and IVD Prostheses

Upper extremity prostheses include the shoulder, elbow, and wrist (1). Total shoulder arthroplasty (TSA) consists of a humeral component with an intermedullary stem and a

Table 2. Measures that Maximize Strength and Minimize Stress in Total Hip Replacement Structures

	Cement/Metal Interface	Cement	Cement/Bone Interface
Maximize strength	Grit-blasted metal PMMA-coated metal	Optimal preparation Pressurization Cement restrictor	Careful reaming Pressurization Minimal polymerization heat Minimal monomer Bone lavage Minimal wear debris
Minimize stress	Reduce patient weight Reduce patient activity Anatomical reconstruction of the femoral head Minimal friction No impingement or subluxation Bonded cement/metal interface Optimal implant and cement mantle design Optimal implant material Optimal (reproducible) placement		

Adapted from Huiskes (22).

glenoid component inserted into the glenoid cavity of the scapula. The glenoid component is either all-polyethylene; in which case, it is cemented; or metal-backed; in which case, it may be fixated by osseointegration. Glenoid components may have several pegs, or they may have one central "keel" for fixation (25). Elbow prostheses consist of humeral, ulnar, and radial components, all which may be fixated with or without cement. Wrist prostheses replaces the radial head and the schapoid and lunate bones of the wrist and may be cemented and uncemented (1). Intervertebral disk (IVD) prostheses replace the degenerated disk with a polymer; there are several strategies for fixation: The endplates may be porous coated and plasma sprayed for osseointegration to the cancellous bone with vertical fins to increase stability. IVD prostheses may also be fixed to adjacent vertebral bodies with screws (26).

EVALUATION OF FIXATION AND FUTURE STRATEGIES

One of the key issues in orthopedic implant fixation is whether to use cemented fixation or biological fixation, with surgeons on both sides of the debate (16,27). Cemented fixation has the advantage of immediate postoperative, stability whereas concerns may be raised about the reliability of bone cement's fatigue strength; furthermore, there is a school of thought that the exothermic polymerization reaction should be avoided if at all possible. Biological fixation by osseointegration has the advantage of avoiding the use of the PMMA cement but runs the risk of the failed ingrowth failure scenario; furthermore, immediate postoperative weight-bearing is not possible. Finally cementless implants are easier to revise if they fail.

Another key issue in orthopedic implant fixation is that of preclinical testing and regulatory approval of new fixation technologies. Considerable challenges exist in achieving consensus around regulatory tests that safeguard patients against ineffective devices while still allowing innovation (4). Preclinical tests can use either (1) finite element models of the direct postoperative situation, e.g., for the hip (28), knee (29), or shoulder (25), or computer simulations of a failure scenario, e.g., damage accumulation (30); (2) physical model "bench" testing with simulators (24,31); or (3) animal testing. Animal testing is not ideal for testing the biomechanical efficacy of orthopedic implant fixation because the implant geometry must be modified to fit the animal skeleton. Furthermore, an important emerging concept is that of patient-specific implants based on computational analysis of a patient's medical images (32).

One useful clinical method to assess implant fixation is through the use of radiostereometric analysis (RSA). With this approach, the migration of the implant relative to the bone can be determined and is used to determine designs that may be at risk of early loosening. Retrospective and prospective clinical studies are also very useful to determine designs or materials that have promising or poor clinical results. On a larger scale, implant registries performed in many countries in Western Europe can provide information on how designs, materials, and surgical techniques rank in terms of risk of failure. All of these clinical tools can aid in understanding the role of implant fixation in success of joint replacements.

A final issue is the degree to which broader technological innovations in surgery and medicine will affect orthopedics. For example, minimally invasive therapy (33) requires special implants and associated instrumentation. Tissue-engineering and regenerative medicine also has the potential to change the nature of orthopedics, not only by reducing the need for joint arthroplasty implants but by integrating tissue engineering concepts with conventional implant technologies, for example, cell seeding into implant surfaces to promote biological fixation.

ACKNOWLEDGMENTS

Research funded by the Programme for Research in Third-Level Institutions, administered by the Higher Education Authority. Dr. A. B. Lennon and Ms. S. Brown are thanked for their comments.

BIBLIOGRAPHY

1. Prendergast PJ. Bone prostheses and implants. In: Cowin SC, editor. Bone Mechanics Handbook. Boca Raton: CRC Press; 2001; 35(1)–35(29).
2. Prendergast PJ, van der Helm FCT, Duda G. Analysis of joint and muscle loading. In: Mow VC, Huiskes R, editors. Basic Orthopaedic Biomechanics and Mechanobiology. Philadelphia: Lippincott Williams & Williams; 2005;29–89.
3. Park JB. Orthopaedic prosthesis fixation. In: Bronzino JD, editor. The Biomedical Engineering Handbook. Boca Raton: CRC Press; 1995. pp. 704–723.
4. Huiskes R. Failed innovation in total hip replacement. Diagnosis and proposals for a cure. Acta Orthopaedica Scandinavica 1993;64:699–715.
5. Mann KA, Mocarski R, Damron LA, Allen MJ, Ayers DC. Mixed-mode failure response of the cement-bone interface. J Orthop Res 2001;19:1153–1161.
6. Balderston RA, Rothman RH, Booth RE, Hozack WJ. The Hip. New York: Lea & Febiger; 1992.
7. Lennon AB, Prendergast PJ. Residual stress due to curing can initiate damage in porous bone cement: experimental and theoretical evidence. J Biomechan 2002;35:311–321.
8. Harper EJ. Bioactive bone cements. Proc Inst Mech Eng Part H J Eng Med 1998;212:113–120.
9. Lewis G. The properties of acrylic bone cement: A state-of-the-art review. J Biomed Mater Res 1997;38:155–182.
10. Murphy BP, Prendergast PJ. On the magnitude and variability of fatigue strength in acrylic bone cement. Int J Fatigue 2000;22:855–864.
11. Verdonschot N, Huiskes R. The dynamic creep behaviour of acrylic bone cement. J Biomed Mater Res 1995;29:575–581.
12. Prendergast PJ, Murphy BP, Taylor D. Discarding specimens for fatigue testing of orthopaedic bone cement: A comment on Cristofolini et al. (2000). Fatigue Fracture Eng Mater Structures 2002;25:315–316.
13. Albrektsson T. Biological factors of importance for bone integration of implanted devices. In: Older J, editor. Implant Bone Interface. New York: Springer; 1990;7–19.
14. Skalak R. Biomechanical considerations in osseointegrated prostheses. J Prosthetic Dentistry 1983;49:843–860.

15. Eldridge JDI, Learmonth ID. Component bone interface in cementless hip arthroplasty. In: Learmonth ID, editor. Interfaces in Total Hip Arthroplasty. London: Springer; 1999: 71–80.
16. Sychterz CJ, Claus AM, Eng CA. What we have learned about cementless fixation from long-term autopsy retrievals. Clin Orthop Related Res 2002;405:79–91.
17. Søballe K. Hydroxyapatite ceramic coating for bone-implant fixation. Mechanical and histological studies in dogs. Acta Orthop Scand 65: (Suppl. 255).
18. Prendergast PJ, Huiskes R, Søballe K. Biophysical stimuli on cells during tissue differentiation at implant interfaces. J Biomechan 1997;30:539–548.
19. Lennon AB, McCormack BAO, Prendergast PJ. The relationship between cement fatigue damage and implant surface finish in proximal femoral prostheses. Med Eng Phys 2003; 25:833–841.
20. Dalstra M Biomechanical aspects of the pelvic bone and design criteria for acetablar prostheses. Ph. D. Thesis, University of Nijmegen, 1993.
21. Lee AJC. Rough or polished surface on femoral anchorage stems. In: Buchhorn GH, Willert HG, editors. Technical Principles, Design and Safety of Joint Implants. Seattle: Hogrefe & Huber Publishers; 1994:209–211.
22. Huiskes R. New approaches to cemented hip-prosthetic design. In: Buchhorn GH, Willert HG, editors. Technical Principles, Design and Safety of Joint Implants. Seattle: Hogrefe & Huber Publishers; 1994:227–236.
23. Robinson RP. The early innovators of today's resurfacing condylar knees. J Arthroplasty 2005;20 (suppl. 1):2–26.
24. Walker PS. Biomechanics of total knee replacement designs. In: Mow VC, Huiskes R, editors. Basic Orthopaedic Biomechanics and Mechanobiology. Philadelphia: Lippincott Williams & Wilkins; 2005:657–702.
25. Lacroix D, Murphy LA, Prendergast PJ. Three-dimensional finite element analysis of glenoid replacement prostheses: A comparison of keeled and pegged anchorage systems. J Biomech Eng 2000;123:430–436.
26. Szpalski M, Gunzburg R, Mayer M. Spine arthroplasty: A historical review. European Spine J 2002;11 (suppl. 2): S65–S84.
27. Harris WH. Options for the primary femoral fixation in total hip arthroplasty—cemented stems for all. Clin Orthop Related Res 1997;344:118–123.
28. Chang PB, Mann KA, Bartel DL. Cemented femoral stem performance—effects of proximal bonding, geometry, and neck length. Clin Orthop Related Res 1998;355:57–69.
29. Taylor M, Barrett DS. Explicit finite element simulation of eccentric loading in total knee replacement. Clin Orthop Related Res 2003;414:162–171.
30. Stolk J, Maher SA, Verdonschot N, Prendergast PJ, Huiskes R. Can finite element models detect clinically inferior cemented hip implants? Clin Orthop Related Res 2003;409:138–160.
31. Britton JR, Prendergast PJ. Pre-clinical testing of femoral hip components: an experimental investigation with four prostheses. J Biomechan Eng. In press.
32. Viceconti M, Davinelli M, Taddei F, Capello A. Automatic generation of accurate subject-specific bone finite element models to be used in clinical studies. J Biomechanics 2004; 37:1597–1605.
33. DiGioia AM, Blendea S, Jaramaz B. Computer-assisted orthopaedic surgery: minimally invasive hip and knee reconstruction. Orthop Clin North Am 2004;35:183–190.

See also BIOCOMPATIBILITY OF MATERIALS; BONE AND TEETH, PROPERTIES OF; BONE CEMENT, ACRYLIC; HIP JOINTS, ARTIFICIAL; MATERIALS AND DESIGN FOR ORTHOPEDIC DEVICES.

ORTHOTICS. See REHABILITATION, ORTHOTICS IN.

OSTEOPOROSIS. See BONE DENSITY MEASUREMENT.

OVULATION, DETECTION OF. See CONTRACEPTIVE DEVICES.

OXYGEN ANALYZERS

SUSAN MCGRATH
SUZANNE WENDELKEN
Dartmouth College
Hanover, New Hampshire

INTRODUCTION

Oxygen is essential for all aerobic life on Earth. It is the most abundant element as it comprises a little more than one-fifth of the weight of air, nine-tenths of the weight of water, and almost one-half of the weight of the earth's crust (1).

Because of its role in supporting and sustaining life, it is often important to monitor the level of oxygen in the atmosphere. Too much oxygen can lead to a toxic atmosphere where as too little oxygen causes asphyxia and eventually death. A relatively constant level of oxygen is required for most aerobic processes.

Oxygen gas monitoring is used for a number of purposes: (*1*) **Medical:** anesthesia (drug delivery, airway monitoring), respiratory oxygen content monitoring (inhaled and exhaled), controlled environments, incubators. (*2*) **Physiological:** exercise (rate of oxygen consumption), aircraft, spacecraft, scuba diving, fire fighting, mountain climbing, spelunking. (*3*) **Biological:** metabolism (oxygen uptake and consumption), fermentation, beverage and food packing. (*4*) **Industrial:** combustion control, fuel and pollution management, safe operation of chemical plants, monitoring gas purity.

This article gives an overview of the analyzers used to measure gaseous oxygen in medicine, physiology, and biology. Measurement of dissolved or bound oxygen is also important in medicine and is discussed in detail elsewhere in this Encyclopedia.

History and Relevance

Oxygen was not known to exist until the 1770s when it was discovered by French scientist Antoine Lavoisier and English clergyman and scientist Joseph Priestly through experiments on combustion. Previously, air was considered to be an element composed of a single substance. Combustible materials were thought to have a substance called phlogiston, from the Greek word meaning to be set on fire, which escaped as a material was burned. Lavoisier, however, believed that combustion resulted from a combination of fuel and air. He conducted experiments where he burned a candle in a sealed jar and observed that only one-fifth of the air was consumed. He named this unconsumed portion of the air oxygen from the Greek word

meaning acid producing. Although his thoughts about oxygen being the corrosive agent in acidic compounds was wrong, the name stuck and the study of oxygen was born (2).

Oxygen is essential for most life on Earth as it plays a key role in aerobic metabolism as a final electron acceptor due to its high electron affinity. Metabolic rate can be indirectly measured by monitoring oxygen consumption as >95% of energy is produced by reactions of oxygen with other food (3). This method is called indirect calorimetry and is a much more cost effective and timely method for measuring metabolic rate as compared to direct calorimetry (the direct measure of heat energy produced).

Oxygen availability is a function of its partial pressure and the total pressure of the gas mixture in which it resides. At sea level, the partial pressure of oxygen is roughly 21%. With decreasing atmospheric pressure, as accompanies increasing altitude, the total amount of available oxygen decreases (Table 1). For example, at 18,000 ft. (5.48 km) above sea level, although the partial pressure of oxygen is still 21%, there is roughly half the amount of available oxygen. At ~29,000 ft. (8.33 km) above sea level on the top of Mt. Everest, there is less than a third the amount of total available oxygen compared to sea level. At such altitudes, most humans require the use of supplemental oxygen. In addition, the body will compensate for the reduced oxygen availability by increasing the heart and respiration rate to keep up with the metabolic demands (3). A climber's resting heart rate at this altitude is double to triple their normal resting heart rate. Long-term exposure to high altitude prompts the body to produce more red blood cells per unit blood volume thus increasing the number of oxygen carriers and making respiration easier. If the body does not properly adapt to such conditions, altitude sickness, pulmonary and cerebral edema, and potentially death may result (3).

Table 1. Atmospheric Pressure, the Fraction of Available Oxygen Compared to Sea Level, and Temperature All Decrease with Increasing Altitude[a].

Altitude, ft.	Barometric Pressure, mmHg	Fraction Available Oxygen	Temperature °C
0	760	1.00	15
1,000	733	0.96	13
5,000	632	0.83	5.1
10,000	523	0.69	−5.4
14,000	447	0.59	−12.7
16,000	412	0.54	−16.7
18,000	380	0.50	−20.7
20,000	349	0.46	−24.6
22,000	321	0.42	−28.6
24,000	295	0.39	−32.6
26,000	270	0.36	−36.5
28,000	247	0.33	−40.5
30,000	228	0.30	−44.4
32,000	206	0.27	−48.4
34,000	188	0.25	−52.4
36,000	171	0.23	−56.3

[a]See Ref. 4.

The partial pressure of oxygen remains a fairly constant 21% until very high altitudes [i.e., >50,000 ft. (15.24 km) (5)]. At these altitudes it is necessary to maintain a pressurized, enclosed environment, such as aircraft, spacecraft, or space suite, in order to sustain human life.

Oxygen availability can be decreased by displacement by other gases, such as nitrogen, carbon dioxide, methane, and anesthetics. Oxygen availability is also easily decreased by combustion and oxidation processes. Thus it is necessary to monitor the atmospheric oxygen level in situations where these gases or combustion is present such as in enclosed environments, closed breathing circuits, and fire fighting.

Each year ~20 deaths occur as a result of asphyxiation due to displacement of oxygen by another gas in air (6). Accidental asphyxia usually occurs in industry as a result of oxygen depletion by carbon dioxide, CO_2, methane (CH_4), or a hydrocarbon gas in a confined space, such as a tunnel, laboratory, sewer, mine, grain silo, storage tank, or well (7). For example, in 1992, a barge operator in Alaska died from asphyxiation and a rescuer lost consciousness during rescue efforts due to a low level of oxygen (6%) in a confined space (8). In anesthesia, accidental asphyxia has resulted from incorrectly connected gas delivery tubes (9).

In choosing an oxygen analyzer for a particular need, it is important to be acquainted with the properties of operation, characteristics, and limitations of these devices. The primary methods for oxygen detection are based on the paramagnetic susceptibility, electrochemical properties, and light absorption properties of oxygen.

PARAMAGNETIC OXYGEN ANALYZERS

All mater exhibits some form of magnetism when placed in a magnetic field. Magnetic susceptibility is the measure of the strength of a material's magnetic field when placed in an external magnetic field. Diamagnetic substances, such as gold and water, align perpendicularly to an external magnetic field causing them to be repelled slightly. This property arises from the orbital motion of electrons that produces a small magnetic moment. In substances with paired valence electrons, these moments cancel out. However, when an external magnetic field is applied it interferes with the motion of the electrons causing the atoms to internally oppose the field and be slightly repelled by it. Diamagnetism is a property of all materials, but is very weak and disappears as soon as the external magnetic field is withdrawn. In materials with unpaired valence electrons (e.g., nickel and iron) an external magnetic field aligns the small magnetic moments in the direction of the field, which increases the magnetic flux density. Materials with this behavior are weakly attracted to magnetic fields and are classified as paramagnetic. Ferromagnetism is a special case of paramagnetism where materials (e.g., iron and cobalt) are strongly attracted to magnetic fields. Paramagnetic materials have a high susceptibility (10).

Oxygen has a relatively high susceptibility when compared to other gases (see Table 2). This property is the key principle behind paramagnetic oxygen analyzers.

Table 2. Relative magnetic susceptibility values on a scale of Oxygen = 100 and Nitrogen ~= 0 at 20° C (11–13).

Gas	Relative Magnetic Susceptibility
Argon	−0.58
Acetylene	−0.38
Air (dry air)	21.00
Ammonia	−0.58
Carbon dioxide	−0.61
Carbon monoxide	0.06
Chlorine	−0.13
Ethane	−0.83
Helium	0.29
Hydrogen	−0.12
Methane	−0.37
Nitrogen	**−0.42**
Nitrous oxide	−0.58
Nitrogen monoxide	43.80
Nitrogen dioxide	28.00
Oxygen	**100.00**

The three main types of paramagnetic oxygen analyzers are (*1*) thermomagnetic (magnetic wind); (*2*) magnetodynamic (dumbbell or autobalance); (*3*) magnetopneumatic (differential pressure).

Paramagnetic analyzers are typically used for monitoring the quality of breathing air in open and enclosed environment, biological laboratory measurements, and in industrial combustion analysis (2,14).

Limitations of Paramagnetic Analyzers. Because the magnetic susceptibility of oxygen depends on temperature, it is necessary to operate at a constant temperature or to have some temperature compensation ability (1,15). The output of the sensor is also proportional to the absolute atmospheric pressure and thus pressure compensation is sometimes necessary (1,15).

Paramagnetic devices are typically delicate instruments with moving parts and are thus adversely influenced by vibrations. They generally are not used as portable devices (13).

Paramagnetic sensors work well for percent oxygen measurement, but are not recommended for trace oxygen measurements. In addition, these sensors should not be used when interference effects cannot be compensated for (i.e., sample gas containing other paramagnetic or diamagnetic gases, or varying background gas composition) (14). The effects of background gases used in anesthesia are small but not always negligible. These effects are summarized in Table 3.

Table 3. Errors in Paramagnetic Analyzer measurements due to anesthesia gases[a].

Gas	Error in Instrument Reading in $\%O_2$ Due to 1% Vapor
Diethyl ether	−0.0068
Halothane	−0.0157
Nitrous oxide	−0.0018
Methoxyflurane	0.0000
Trichloroethylene	−0.0033

[a]See Ref. 15.

Thermomagnetic (Magnetic Wind)

Thermomagnetic analyzers are based on the fact that magnetic susceptibility decreases inversely with the square of temperature.

Principles of Operation. A schematic diagram of a thermomagnetic analyzer can be seen in Fig. 1. A gas sample is admitted into the inlet that branches into equal segments and converges again at the outlet. These tubes are connected by another tube halfway between the inlet and outlet. This cross-tube is heated by a platinum coil that is separated in the center by a thermal resistance bridge. These two heater coil segments form two arms of a Wheatstone bridge with the third arm being the output of the sensor. A magnetic field is applied to one-half of the coil. Any oxygen in the gas is attracted to the magnetic field in the cross-section. These oxygen molecules are subsequently heated by the heater coil and immediately begin to lose their magnetic susceptibility. They are then displaced by cooler oxygen molecules with higher magnetic susceptibility. This flow of gas through the cross-tube, referred to as the magnetic wind, cools the magnetized heating coil and heats the unmagnetized coil causing an imbalance in the bridge resulting from the difference in resistance between the two coils. The bridge output is then calibrated by passing a gas with known oxygen concentration through the chamber (1,11,13).

Limitations of Thermomagnetic Analyzers. Measurement by themomagnetic oxygen analyzers is affected by the magnetic susceptibility and thermal conductivity of the

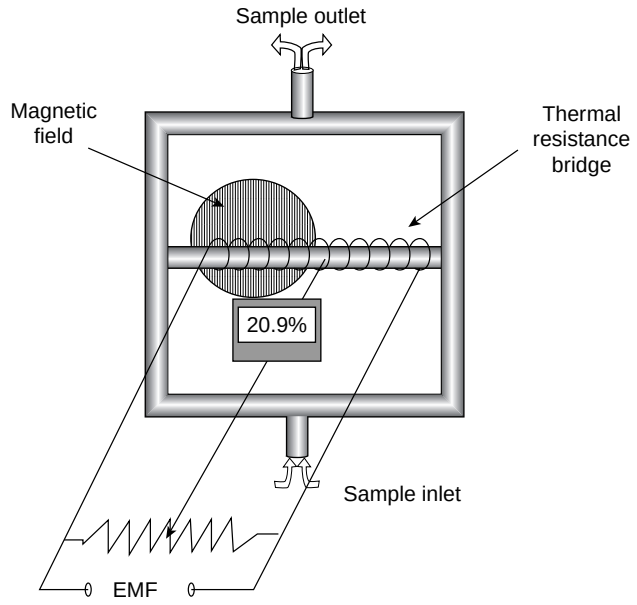


Figure 1. Schematic diagram of a thermomagnetic oxygen analyzer (1).

carrier gas, the sample gas temperature, ambient temperature, tilt, sample flow, and pressure (1,16).

Magnetodynamic (Dumbbell or Autobalance)

Developed by Faraday in 1884, this is the most popular method of paramagnetic oxygen analyzers and the earliest developed oxygen analyzer (13). Magnetodynamic oxygen analyzers are based on property that oxygen will be drawn into a magnetic field because it is paramagnetic. These analyzers essentially measure the magnetic susceptibility of sample gas.

Principles of Operation. A simple form of this device consists of small, dumbbell shaped body made of quartz and filled with nitrogen or some gas with small or negative magnetic susceptibility, an optical lever system, and a nonuniform magnetic field (Fig. 2). The dumbbell is suspended in a closed chamber by a quartz or platinum wire between two permanent magnets that are specially shaped to have a nonuniform magnetic field. The dumbbell is free to rotate. Since the dumbbell is slightly diamagnetic, it will naturally rotate away from the highest magnetic field intensity. Oxygen in a sample gas will be attracted to the region of maximum field intensity and displace the dumbbell even further. This deflection is measured by an optical lever system in which a light source outside the test chamber shines a beam onto a mirror which is mounted in the center of the dumbbell. The beam is then reflected onto a scale outside the chamber. The amount of defection is directly proportional to the partial pressure of oxygen in the sample (1,17).

Modern designs of this sensor are self-nulling and have temperature compensation capabilities. A single turn coil is wound around the dumbbell. The coil produces a magnetic field when current flows through it which will in turn cause the dumbbell to rotate in the external magnetic field. The deflection of the dumbbell is measured by an optical lever system that uses photocells to detect the light reflected from the mirror.

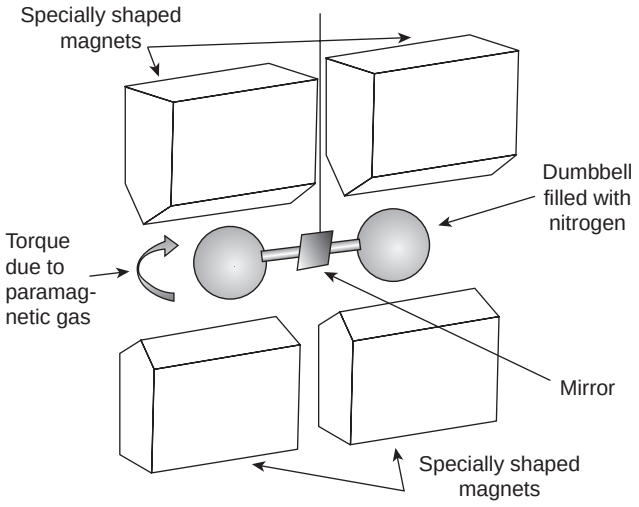


Figure 2. Schematic diagram of paramagnetic "dumbbell" sensor. Adapted from (1).

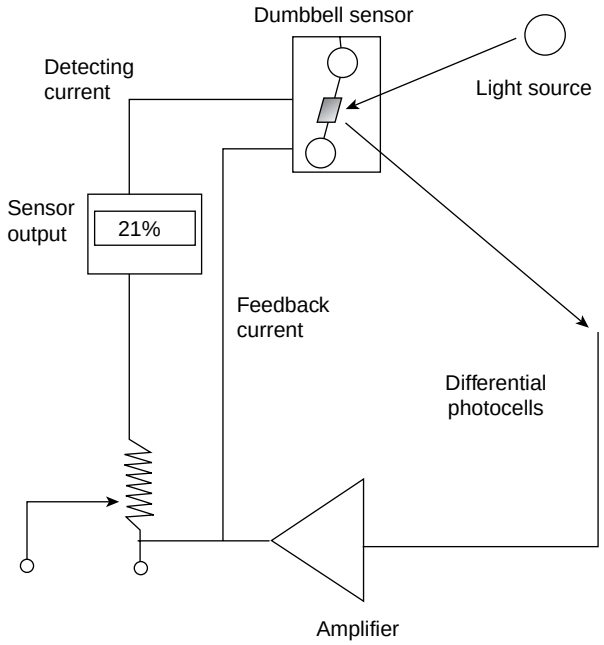


Figure 3. Self-nulling paramagnetic analyzer. Adapted from Refs. 1 and 17.

The photocells are connected to a feedback loop that controls the amount of current through the coil, keeping the dumbbell centered with respect to the photocell detectors. As the paramagnetic components of a gas sample move into the strongest part of the magnetic field, the dumbbell is displaced and begins to rotate. The photocells detect this motion and drive an amplifier to produce the necessary current in the coil to keep the dumbbell in the original zero position. The system is zeroed using a sample of pure nitrogen. In this case, the dumbbell is at an equilibrium position and there is no current flowing through the coil. The current is directly proportional to the magnetic susceptibility of the sample. The system is calibrated using a sample of known oxygen content. See Fig. 3 for a diagram of this design.

Limitations of Dumbbell Analyzers. The main limitation of the dumbbell design is its slow response time ($\sim$10 s) (15). Thus, the dumbbell analyzer is not recommended for uses where real-time oxygen analysis is needed. These analyzers also have moving parts and are extremely sensitive to tilt and vibrations.

Magnetopneumatic (Differential Pressure)

This sensor operates on the principle that a differential pressure will be generated when a sample containing oxygen is drawn into a nonuniform magnetic field with a reference gas of different oxygen content. Differential pressure sensors directly measure the magnetic susceptibility of sample gas and are thus not influenced by thermal properties of background gas.

Principles of Operation. A reference gas is admitted at a constant rate into a chamber like the one seen in Fig. 4. The

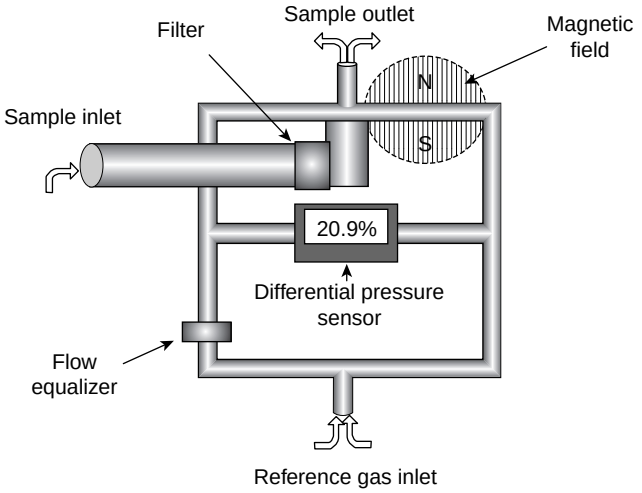


Figure 4. Schematic diagram of a differential pressure sensor. (Adapted from Ref. 1.)

reference gas is split into two paths with a flow equalizer to ensure equal flow in each path. Each path is joined at the midpoint by a connector pipe containing a differential pressure sensor (e.g., a capacitive differential pressure sensor or a microflow sensor). The two paths reconnect at an outlet where the sample gas is admitted. There is a strong nonuniform magnetic field placed over one of the reference gas outlets. The reference gas and the sample gas combine in the outlet. Oxygen or other paramagnetic gases in the sample gas will be drawn into the nonuniform magnetic field and cause a pressure build up on that side of the reference gas path.

The differential pressure is proportional to the magnetic susceptibility of the sample gas only. This imbalance is sensed by the differential pressure sensor in the cross-tube. The output of this sensor is calibrated in terms of oxygen concentration by using a reference and sample gas of known oxygen content. The output is zeroed by using a sample gas that is the same as the reference gas. In this case, the output of the differential pressure sensor will be zero (1,15).

Limitations of Magnetopneumatic Analyzers. Differential pressure sensors are sensitive to tilt and vibrations. An alternating magnetic field reduces the effects of background flow and tilt on the sensor (1).

ELECTROCHEMICAL OXYGEN ANALYZERS

There are two main types of electrochemical oxygen analyzers: those with aqueous electrolytes, and those with solid electrolytes. These sensors use the chemical reactivity of oxygen to produce a measurable current that is proportional to the partial pressure of oxygen in a sample gas.

Aqueous Electrolyte Sensors

Galvanic Oxygen Analyzer. Galvanic oxygen analyzers are commonly called a Hersch cell after the inventor. They are essentially a battery that produces energy when it is exposed to oxygen. Fig. 5. Galvanic sensors are typically insensitive to vibration and tilt. They are usually packaged small and made out of inexpensive and sometimes disposable materials (14). Disposable capsules containing galvanic cells are fairly inexpensive (~$85) and typically last 1–5 years (18). Recently, small, portable galvanic sensors have been manufactured and approved for medical breath analysis purposes (Fig. 6) (19).

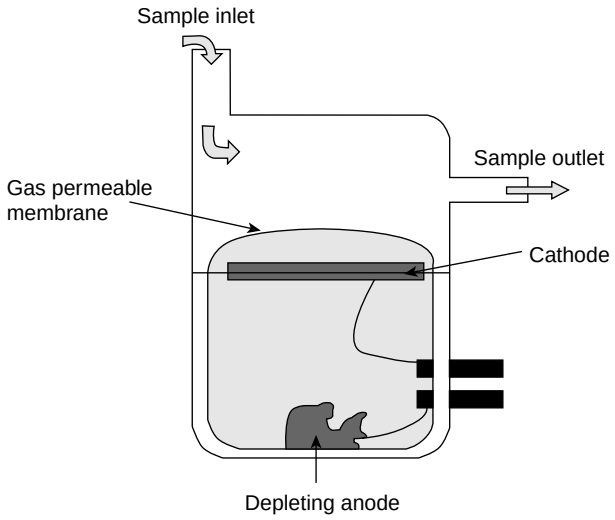


Figure 5. Schematic diagram of a galvanic sensor. (Adapted from Ref. (14).)

Figure 6. A small, handheld galvanic oxygen sensor (model AII 2000, Analytical industries Inc.) (19).

Galvanic sensors are typically used for industrial purposes, such as validating the quality of semiconductor grade gases and for environmental monitoring (e.g., monitoring the quality of breathing air or monitoring the oxygen content in potentially hazardous or explosive environments) (14).

Principles of Operation. Basic cells consist of a cathode made of a precious metal (platinum, gold, silver, etc.) and an anode made of a base metal (lead, cadmium, antimony). These electrodes are in contact with a liquid or semisolid electrolyte, such as potassium hydroxide. A gas sample is admitted into the cell and diffuses through a membrane made of a thin material, such as Teflon or silicone, which is permeable to oxygen but not to the electrolyte. The oxygen in the solution is chemically reduced at the cathode to form hydroxyl ions that flow to the anode where an oxidation reaction occurs. This oxidation–reduction reaction results in an electromotive force (EMF), which is proportional to the oxygen concentration in the solution and the partial pressure of oxygen in the sample gas. The electron flow is measured by an external galvanometer connected to the electrodes.

Reactions at the cathode and anode are as follows 1, 20–22):

1. Cathode reaction:

$$O_2 + 2\,H_2O + 4\,e^- \rightarrow 4\,OH^-$$

2. Anode reaction:

$$2\,Pb + 6\,OH^- \rightarrow 2\,PbO_2H^- + 2\,H_2O + 4\,e^-$$

3. Overall reaction:

$$O_2 + 2\,Pb + 2\,OH^- \rightarrow 2\,PbO_2H^-$$

Designs that allow for every oxygen molecule passing through the cell to react are called coulometers and are suitable for trace (parts per million, ppm) measurements (1,20–22).

Limitations of Galvanic Analyzers. Because the anode is consumed by oxidation, the cell has a limited life. These devices tend to lose sensitivity as they age resulting in falsely low readings (14).

There are some designs that lessen the rate of anode consumption by using a third inert electrode made of platinum that is kept at a constant positive potential. This results in the majority of the current to be conducted through this electrode instead of the consumable anode (1,4).

Acidic gases in the sample (i.e., SO_2 CO_2, Cl_2) react with the electrolyte and must be scrubbed out. There are some coulometric cell designs that overcome this problem by the use of additional electrodes (1,23).

Exposure to very high oxygen concentration can lead to oxygen shock where the sensor is saturated and does not recover for hours (14).

Polarographic Oxygen Analyzers. This sensor responds to changes in the partial pressure of oxygen in a sample gas. Polarographic sensors can be used for measuring oxygen in dissolved liquid (14) or in a gas sample. They are often used in anesthesia gas delivery systems. Polarographic sensors are insensitive to shock, vibration, and tilt. The effects of flow are minimal because diffusion is controlled by a membrane (14).

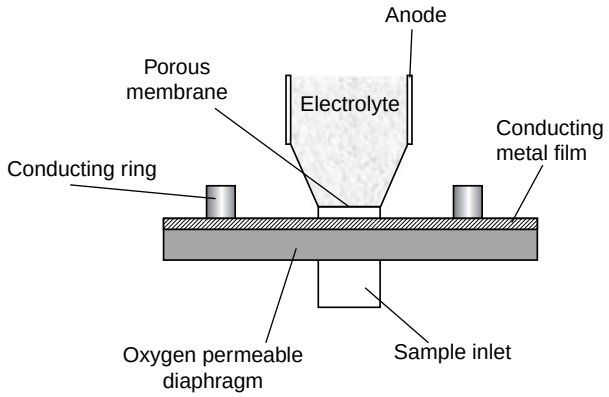


Figure 7. Schematic diagram of a polarographic sensor. (Adapted from Refs. 1,24,25.)

Principles of Operation. A polarographic cell, as seen in Fig. 7, consists of two electrodes, usually a silver anode and a gold cathode, immersed in an electrolyte, such as potassium chloride (1). An EMF is applied across the electrodes inducing oxidation–reduction reactions when a sample containing oxygen is admitted into the cell. Like the galvanic cell, oxygen diffuses through a thin membrane that is preferentially permeable to oxygen and not to the electrolyte. This membrane is usually made from poly(tetrafluoroethylene) (PTFE) and controls the rate of oxygen flux to the cathode. The current flow in the cell is proportional to the applied EMF and the partial pressure of oxygen in the sample. As seen in Fig. 8, there are four main regions in the EMF–current curve of importance (1):

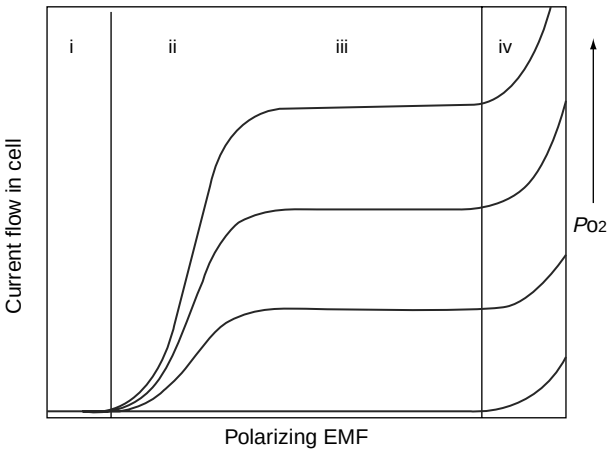


Figure 8. Diagram showing the operating regions of a polarographic sensor. (Adapted from Ref. 10.)

Region i: If the applied EMF is very low, then the presence of oxygen hardly has any effect on the current. There are very little reactions occurring at the electrodes.

Region ii: Oxygen molecules begin to react at the electrodes causing a measurable increase in current. For a given level of oxygen partial pressure, an increase in the EMF produces a sharp increase in the current.

Region iii: This region is the polarized or working region for the polarographic sensor. Here, the current plateaus and an increase in the EMF does not alter the current. In this region, all the oxygen molecules are reduced immediately at the cathode. A calibration curve is used in this region relating oxygen concentration to sensor current.

Region iv: In this region, an increase in EMF leads to a sharp, nonlinear increase in current as a result of the breakdown of the electrolyte solution.

The reactions at the cathode and anode are as follows (1,15):

1. Cathode reaction:

$$O_2 + 2\,H_2O + 4\,e^- \rightarrow 4\,OH$$

2. Anode reaction:

$$4\,Ag + 4\,Cl^- \rightarrow 4\,AgCl + 4\,e^-$$

Limitations of Polarographic Analyzers. Polarographic sensors generally have slow response because oxygen must diffuse through membrane. They are also sensitive to pressure and temperature and compensation for these factors is sometimes required. In addition, these sensors lose sensitivity over time due to degradation of anode and electrolyte solution giving falsely low readings. Polarographic sensors that consume all the injected oxygen change the content of the sample gas and are not good for closed-loop systems, such as closed-circuit anesthesia systems (26).

Capacitive Coulometry. Capacitive coulometry (also referred to as coulometric microrespiometry) is another aqueous electrolytic method for oxygen analysis. This method is based on the replacement of oxygen consumed by an organism in a closed system with electrolytic oxygen produced by discharging a capacitor through a solution of $CuSO_4$ (27). Such analyzers can be used to monitor oxygen consumption and metabolism rate of tissues or microorganisms. However, this type of sensor is not used as frequently as other more common types of aqueous electrolyte sensors.

Solid Electrolyte Cells

Some ceramics conduct electricity at very high temperatures. This conductivity is largely a result of the oxygen mobility in this solid solution. Oxygen mobility is the principle behind solid electrolyte or concentration cells.

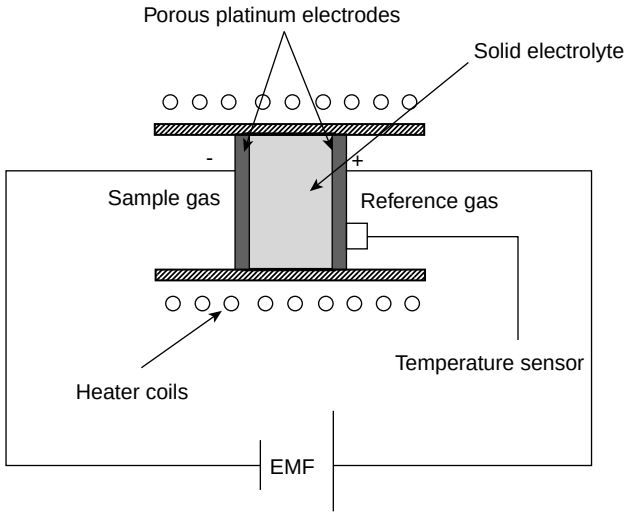


Figure 9. Schematic diagram of a solid electrolyte concentration cell. (Adapted from Refs. 1,15.)

There exists a family of solid electrolytes, such as MgO or CaO in ZrO_2 (ceramic), whose electrical conductivity is mostly due to the mobility of O_2^- as opposed to electrons in the solid. At room temperature, the conductivity is low, but at high temperatures (>600° C) the conductivity is comparable to that of an aqueous electrolyte and electron mobility can be neglected (1).

A solid electrolyte sensor is commonly referred to as a concentration cell. Solid electrolyte oxygen sensors are commonly used in anesthesia and patient monitoring for breath to breath analysis (15). Solid electrolyte cells typically have a fast response (<150 ms) (15) and are good for real-time oxygen analysis.

Principles of Operation. A concentration cell is made by separating a test chamber and a reference chamber by a solid, oxygen conducting electrolyte, such as ZrO_2 or Y_2O_3 with a porous electrode on either side (Fig. 9). When the temperature is increased by an external heater, the solid electrolyte begins to conduct O_2 and an EMF is established between the electrodes. The EMF is related to the partial pressure of oxygen in the test chamber by the Nernst equation:

$$\text{outputEMF} = \frac{RT}{4F}\ln\left(\frac{P'_{O_2}}{P''_{O_2}}\right)$$

where R is the universal gas constant ($8.314\ J \cdot K^{-1} \cdot mol^{-1}$), T is the operating temperature in kelvin, F is Faraday's constant ($9.6485 \times 10^4\ C \cdot mol^{-1}$), P'_{O_2} is the reference partial pressure of oxygen, and P''_{O_2} is the sample partial pressure of oxygen.

Fuel cell: An alternative configuration of a concentration cell is a fuel cell. If a fuel gas such as hydrogen is admitted into one of the chambers, the cell converts

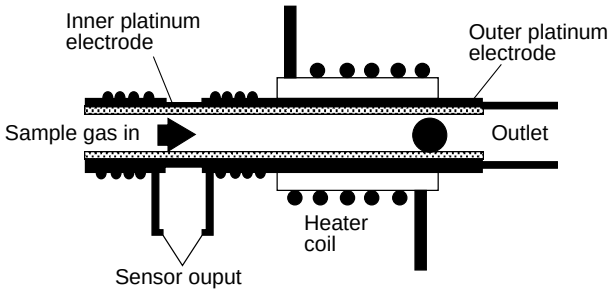


Figure 10. Schematic diagram of a flow through tube sensor. (Adapted from Refs. (1,28,29).)

the chemical energy of the fuel into electrical energy which may be delivered to a load placed across the electrodes (1,15).

Oxygen pump: A concentration cell may also be used as an oxygen pump. An EMF applied across the electrodes will pump oxygen from one chamber to another, the rate and direction depending on the strength and polarity of the applied EMF (1,15).

Sensor design: There are a number of different solid electrolyte sensor designs typically used: a flow through tube sensor, a test tube sensor, and a disk sensor.

Flow-through tube and test tube sensor: The flow-through tube sensor is made from a solid electrolyte tube with a porous platinum electrode on the outside and inside of the tube (Fig. 10). An external heater is used to raise the temperature of the solid electrolyte to its operating temperature where it conducts oxygen. Ambient air outside the tube is used as the reference gas. The sample gas is admitted into the central part of the tube. This simple design was one of the first forms used (15,30,31).

A common variation on this design is the miniature test tube sensor (Fig. 11), which uses a sealed tube containing a reference gas with known oxygen partial pressure (1,32).

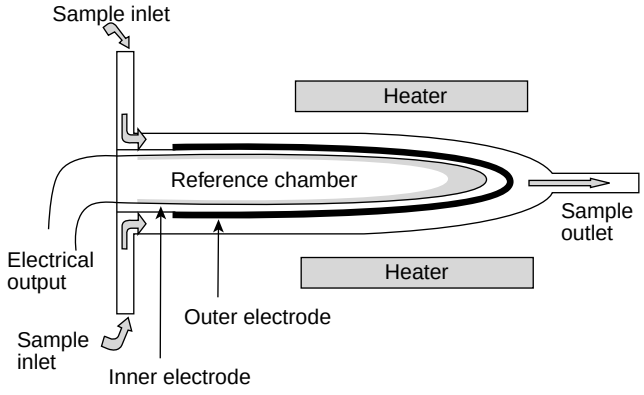


Figure 11. Schematic diagram of a test tube sensor. (Adapted from Refs. 1,33.)

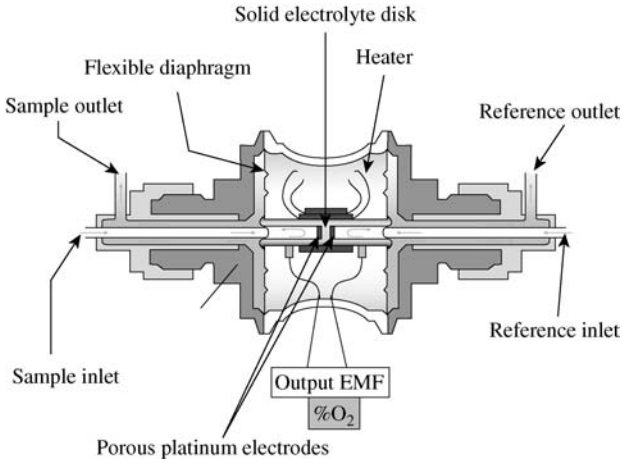


Figure 12. Schematic diagram of a miniature disk sensor. (Adapted from Refs. 1,34,35.)

Like the flow-through sensor, the test tube sensor is made of a solid electrolyte tube with porous electrodes on the inside and outside of the tube, but it is hermetically sealed at one end. The closed end of the tube is place in an external heater. The inside of the tube is used as the reference chamber and ambient air is used for the reference gas. The sample gas is admitted into a chamber surrounding the tube and flows around the outer electrodes. Because of its simple design, it is the most common solid electrolyte cell (1).

Disk sensor: The disk sensor is made with a solid electrolyte disk with porous electrodes on each side. The disk is attached to a metal tube of equal thermal expansion coefficient. Sample gas is admitted into a chamber on one side of the disk and a reference gas on the other. A heater is placed inside the tube on the reference chamber side (1,15,34).

A miniature design of the disk sensor (Fig. 12) has symmetrically placed porous platinum electrodes on either side of a thin solid electrolyte disk. The disk divides a small ceramic cylinder into two hermetically sealed chambers: one for the reference gas and one for the sample gas. An electrical heater brings the cell to the operating temperature. A temperature sensor and feedback loop attached to the heater ensures that the electrodes are also heated to the same temperature as each other and to the disk. If there is a difference in oxygen concentration between the reference and sample chambers, a voltage is generated between the two electrodes. A thin metallic holder is used to suspend the ceramic cylinder and ensures that the ceramic will not crack due to sudden thermal expansion.

Limitations of Ceramic Analyzers. Because the electrode is catalytic, any combustible gas will react with oxygen on the electrode causing it to age and the sensor to give a falsely low reading.

The high operating temperature of these sensors precludes its use around combustible gases.

Ceramics fracture due to thermal shock. Miniaturization helps considerably. These sensors usually require a warm up time (1,15,34).

Pressure on both sides of the disk must be the same or the reading will be inaccurate. This can be accomplished by venting both sides to the ambient atmosphere (15).

OPTICAL SENSORS

Fluorescence Quenching

One of the more recent oxygen analyzer designs uses a technique known as florescence quenching. Fluorescence dyes, such as perylene dibutyrate fluoresce for a certain amount of time in an atmosphere without oxygen. The presence of oxygen quenches this fluorescence. The fluorescence time is thus inversely proportional to the partial pressure of oxygen. In addition to oxygen sensing, fluorescence sensing can be used to detect glucose, lactate, and pH in the laboratory setting (36).

Fiber optic sensors: Fiber optic sensors use florescence quenching to measure the partial pressure of oxygen. A fiber optic strand delivers an optical pulse (usually blue light) to the fluorescent dye. The dye molecules are held in place by small beads of clear plastic. The beads are enclosed by a porous polypropylene membrane that is gas permeable and hydrophobic. The fluorescence is sensed by a photodetector at the end of a second fiber optic strand. The configuration of this sensor can be seen in Fig. 13. The time of fluorescence is calibrated for oxygen concentration. Because the fiber optic is only used to deliver the optical pulse and the fluorescence quenching is used to sense oxygen, the term fiber optic sensor is somewhat of a misnomer.

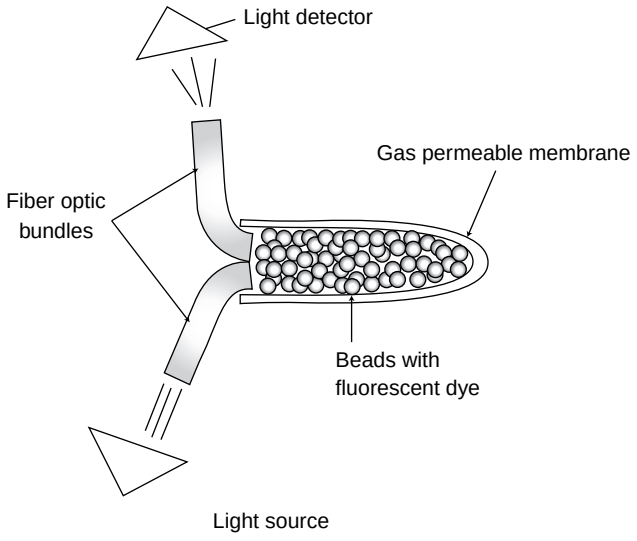


Figure 13. Schematic diagram of a fiber optic sensor. (Adapted from Refs. 1,37.)

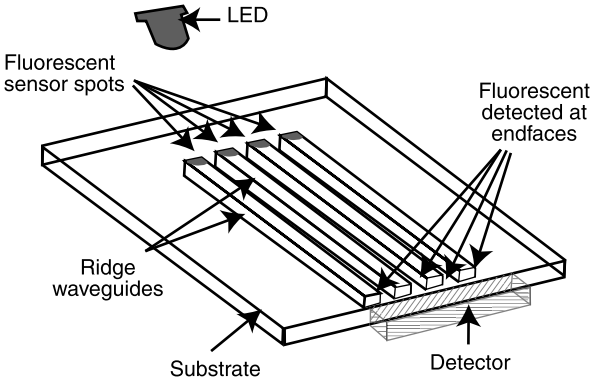


Figure 14. Schematic diagram of an integrated optic oxygen sensor chip. (From Ref. (38).)

Integrated optic oxygen sensor chip: Some sensors that can detect and identify multiple gases in a sample are called multianalyte sensors. One such sensor is an integrated optic multianalyte sensor. This recent development is a miniature optical sensor that has both biomedical and commercial applications. It is based on fluorescence quenching by oxygen.

This sensor, shown in Fig. 14, consists of a multimode ridge waveguide deposited on a dielectric substrate of a higher refractive index than the ridge. Spots of solgel, doped with an oxygen sensitive fluorescent dichlororuthenium dye complex, are deposited at the end of the waveguide and directly excited by a blue LED. Optical detectors at the other end of the waveguide detect emissions from the fluorescent spots. The fluorescence is efficiently coupled to the waveguide as the fluorescent spots are oriented to preferentially emits photons at an angle exceeding the critical angle defined by the two mediums. The theory of fluorescence emission at a dielectric interface is discussed further in (39).

The main limitation of this type of device is the response time. Typically, a 10 s integration period is used for each partial pressure oxygen measurement.

The main advantage of this device is its size and relative ease of fabrication (38). These chips have a very small foot print (<1 cm^2). They can be quickly manufactured using soft lithography.

Polarization-based oxygen sensor: Another sensor based on fluorescence quenching is the polarization based oxygen sensor. This sensor uses an oxygen-sensitive film ($Ru(dpp)_3Cl_2$) and an oxygen-insensitive film (Styrl7). A diagonally polarized source illuminates these films that fluoresce in different ways. The oxygen-insensitive film is stretched so that the molecules preferentially emit vertically polarized photons. The $Ru(dpp)_3Cl_2$ film emits mostly unpolarized photons. Orthogonally oriented polarizers select for the vertical and horizontal components of the combined emitted light. The overall polarization of

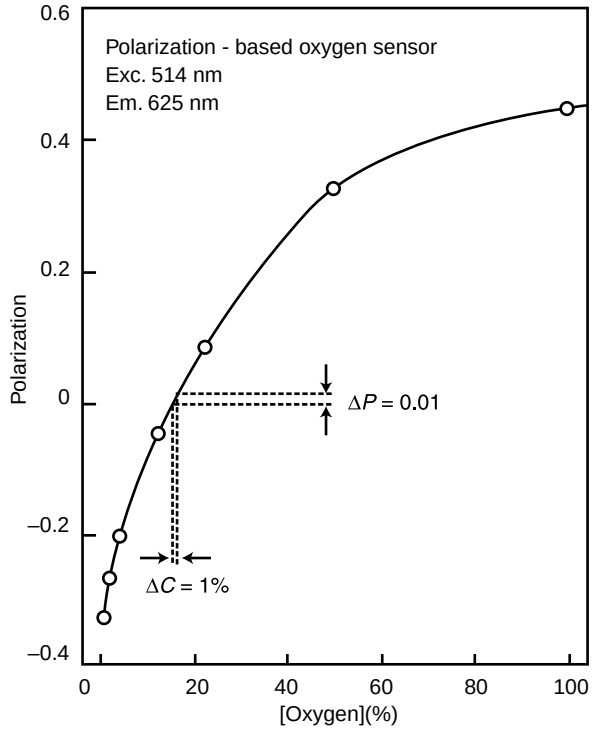


Figure 15. The polarization of the combined film photon emission is proportional to the partial pressure of oxygen. (From Ref. (40).)

the combined emission is sensitive to the partial pressure of oxygen in a sample (Fig. 15). The theory of polarization sensing is discussed further in (40).

UV Absorption

Spectroscopy can also be used to detect oxygen. Oxygen has a maximum absorption coefficient around a wavelength of 0.147 μm, which is in the ultraviolet (UV) range. Most other gases have a much smaller absorption coefficient at this wavelength (1).

A simple device uses an ultraviolet light source, such as a discharge lamp or UV laser as the light source. The beam is split into two paths by a vibrating mirror and directed into either a reference cell filled with nitrogen or a sample cell filled with the gas in question. A photomultiplier tube is used for detection at the end of each path. The ratio of energy received through the sample and reference cell is related to the partial pressure of oxygen in the sample.

Raman Spectroscopy

Raman spectroscopy can be used to monitor multiple gases in a sample. Raman spectroscopy is commonly used in medicine for real-time breath-to-breath analysis or for monitoring respiratory gas mixtures during anesthesia (41–44).

This technique uses the inelastic or Raman scattering, of monochromatic light. The frequencies of the returned light give information about the vibrational, rotational, or other low frequency modes of atoms or molecules in a sample. This information is specific to given elements and compounds and can thus be used to identify and distinguish different gases in a mixture.

In Raman spectroscopy, a sample is illuminated with a laser beam, usually in the visible, near-IR, or near-UV range. Most of the light is scattered elastically or by Rayleigh scattering and is of the same frequency as the incident light. However, a small portion of the light is scattered in-elastically. Phonons, which are quanta of vibrational energy, are absorbed or emitted by the atom or molecule causing the energy of the incident photons to be shifted up or down. This shift in energy corresponds to a shift of frequency. Frequencies close to the laser line are filtered out and frequencies in a certain spectral window are dispersed on to a photomultiplier tube or CCD (charged couple device) camera.

OTHER GASEOUS OXYGEN SENSORS

Gas Chromatography–Mass Spectrometer

A gas chromatography–mass spectrometer (GCMS) combines GC and MS to identify substances in a gas sample. It can be used to detect a variety of compounds in a mixture or it can simply be used to detect the presence of oxygen. The GCMS analyzers are commonly used to measure gas composition in respiratory circuits (42).

Principles of Operation. The gas chromatograph separates compounds into the molecular constituents by retaining some molecules longer than others. These molecules are broken up into ionized fragments that are identified by the mass spectrometer based on the molecules' mass/charge ratio (m/z).

The GC consists of an injector port, an oven, a carrier gas supply, a separation column, and a detector. The injected sample is vaporized in a heated chamber and pushed through the separation column by an inert carrier gas. The separation column is typically a capillary column made of a long, small diameter (usually 1–10 m in length and 0.5 mm in diameter) tube of fused silica (high quality drawn glass) or stainless steel formed into a coil. The components of the sample are separated by two different mediums inside the column that control the speed of travel. These mediums are either coated on the inner surface of the column or packed in the column. Part of the media, known as the stationary phase, absorbs molecules for a certain amount of time before releasing them. The amount of time depends on the chemical properties of the molecule and thus certain molecules are detained longer than others. This, in effect, separates the molecules in the mixture in time. The molecules then travel to the detector. The output of the detector is processed by an integrator. The response of the detector over time is the chromatograph (Fig. 16).

The mass spectrometer separates ions from the gas chromatograph by their charge to mass ratio (m/z) by using an electric or magnetic field. Mass spectrometers usually consist of an ion source, a mass analyzer and a detector.

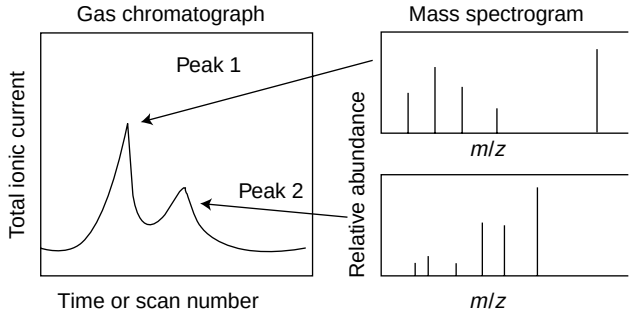


Figure 16. An example gas chromatograph and corresponding mass spectrograms.

The ion source ionizes the sample gas usually with an electron gun. The charged particles are accelerated with an electric field and are steered by the mass analyzer on to the detector by means of a varying electric or magnetic field. The speed and deflection of the particle depends on its mass an charge. When a charged particle comes near or strikes the surface of the detector, a current is induced and recorded. This current is typically amplified by an electron multiplier or Faraday cup. The resulting ionic current plot is the mass spectrum of the ions.

Limitations of Gas Chromatography–Mass Spectrometry. The main limitation of the GCMS is its extremely high price. Mass spectrometers alone run ~$40 k. However, a GCMS may be used for any time of gas measurement.

The Warburg Apparatus

A much older method of oxygen analysis was pioneered by German biochemist Otto Heinrich Warburg (45). The Warburg apparatus was used for measuring cellular respiration and fermentation. This method is based on Boyle's law, which relates the pressure and volume of a gas, and Charle's law, which relates the pressure and temperature of a gas. Combining these laws yields the ideal gas law ($PV = nRT$, where P = pressure, V = volume, n = number moles, R = universal gas constant, T = temperature). At a constant temperature and volume, any change in the amount of gas can be measured as a change in pressure. A typical Warburg apparatus consists of a detachable flask, a waterbath, and a barometer. The sample is placed in a flask and immersed in a bath of water held at a constant temperature. Pressure is measured periodically to determine the amount of gas produced or absorbed by the sample. A variation of this device, used to measure gas production in plants directly from the stem, is the Scholander–Hemmel pressure bomb (46).

CONCLUDING REMARKS

There are three primary types of oxygen sensors available: paramagnetic, electrochemical and spectrographic. Each type of oxygen sensor has limitations and uses dependent on their design, cost and operational environment. Other available oxygen sensors include semiconductor sensors, fluidic sensors, and electron capture oxygen sensors (1). These devices are typically not used in medicine or biology.

BIBLIOGRAPHY

Cited References

1. Kocache R. The measurement of oxygen in gas-mixtures. J Phys E Sci Instrum 1986;19:401–412.
2. O_2 Guide. Electronic Source, Delta-f Corporation. Available at http://www.delta-f.com/O2Guide/o2guide.htm; Accessed 2005.
3. Guyton A, Hall J. Medical Physiology. New York: W.B. Saunders C. 2000.
4. Allsopp PJ. Electrical Cell. UK patent, 969,608. 1964.
5. Johnson T, Rock P. Acute Mountain Sickness. N Eng J Med 1988;319:841–845.
6. Safety_Advisory_Group. Campaign Against Asphyxia. European Industrial Gases Association (EIGA) Safety Newsletter (Special ed.). Brussels.
7. Watanabe T, Morita M. Asphyxia due to oxygen deficiency by gaseous substances. Forensic Sci Int 1998;96:47–59.
8. Operator Dies in Oxygen Defficient Compartment of Ice-Making Barge. Alaska Fatality Assessment & Control Evaluation.
9. Bonsu AK, Stead AL. Accidental cross-connection of oxygen and nitrous oxide in an anesthetic machine. Anaesthesia 1983; 38:767–769.
10. Cheng DK. Field & Wave Electromagnetics. New York: Addison-Wesley; 1984.
11. Instruction Manual: Paramagnetic Oxygen Analyzer: Fuji Electric Co. Ltd.
12. Havens G. The magnetic suceptibilities of some common gases. Phys Rev 1933;43:992–1000.
13. Servomex. Paramagnetic oxygen analysis. Electronic Source, Servomex Corporation. Available at http://www.servomex.com/Servomex.nsf/GB/technology_paramagnetic.html. Accessed 2005.
14. Corporation D-f. O_2 Guide. Electronic Source, Delta-f Corporation. Available at http://www.delta-f.com/O2Guide/o2guide.htm; Accessed 2005.
15. Kocache R. Oxygen analyzers. Encyclopedia of Medical Devices and Instrumentation, Webster JG. editor. New York: Wiley & Sons; 1988. p. 2154–2161.
16. Medlock RS. Oxygen Analysis. Inst Eng 1952;1:3–10.
17. Pauling L. Apparatus for determining the partial pressure of oxygen in a mixture of gases. US patent, 2,416,344, 1947.
18. Meyer RM. Oxygen analyzers: failure rates and life spans of galvanic cells. J Clin Monitoring 1990;6:196–202.
19. Analytical_Industries_Inc. Oxygen analysis pocketed and ABS protected. MD Med. Design, 2001.
20. Hersch P. Electrode assembly for oxygen analysis. UK patent, 913,473. 1962.
21. Hersch P. Improvements relating to the analysis of gases. UK patent, 707,323. 1954.
22. Hersch P. Improvements relating to the analysis of gases. UK patent, 750254. 1956.
23. Gallagher JP. Apparatus and method for maintaining a stable electrode in oxygen analysis. US patent, 3,929,587. 1975.
24. Bergman I. Improvements in or relating to electrical cells. UK patent, 1,344,616. 1974.

25. Bergman I. Improvements in or relating to membrane electrodes and cells. UK patent, 1,200,595. 1970.
26. Li S, Wang Z, Zeng B, Liu J. Multiple respiratory gas monitoring causes changes of inspired oxygen concentration in closed anesthesia system. J Tong Med Univ 1997;17:54–56.
27. Heusner AA, Hurley JP, Arbogast R. Coulometric microrespirometry. Am J Physiol 1982;243:R 185–192.
28. Hickam WM. Device for monitoring oxygen content of gases. US patent, 3,347,767. 1967.
29. Hickam WM. Oxygen control and measuring apparatus. US patent, 3,650,934. 1972.
30. Sodal IE, Micco AJ, Weil JV. An improved fast response oxygen analyzer with high accuracy for respiratory gas analysis. Biomed Sci Instrumentat 1975;11:21–24.
31. Weil JV, Sodal IE, Speck RP. A modified fuel cell for the analysis of oxygen concentration. J Appl Physiol 1967;23:419–422.
32. Deportes CH, Henault MPS, Tasset F, Vitter GRR. Electrochemical gauge for measuring partial pressure of oxygen. US patent, 4,045,319. 1977.
33. Sayles DA. Method and apparatus for continuously seeing the condition of a gas stream. US patent, 3,869,370. 1975.
34. Kocache RMA, Swan J, Holman DF. A miniature rugged and accurate solid electrolyte oxygen sensor. J Phys E Sci Instrum 1984;17:477–482.
35. Servomex. Zirconia oxygen analysis. Electronic Source, Servomex Corporation. Available at http://www.servomex.com/Servomex.nsf/GB/technology_zirconia.html. Accessed 2005.
36. Gryczynski Z, Gryczynski I, Lakowicz JR. Fluorescence-sensing methods. Methods Enzymol 2003;360:44–75.
37. Peterson JI, Fitzgerald RV, Buckhold DK. Fibre-optic probe for *in vivo* measurement of oxygen partial pressure. Anal Chem 1984;56:62–67.
38. Burke CS, et al. Development of an integrated optic oxygen sensor using a novel, generic platform. The Analyst 2005;130:41–45.
39. Polerecki L, Hamrle J, MacCraith BD. Theory of the radiation of dipoles placed within a multilayer system. Appl Op 2000;39:3968–3977.
40. Gryczynski I, Gryczynski Z, Rao G, Lakowicz JR. Polarization-based oxygen sensor. Analyst 1999;124:1041–1044.
41. VanWagenen RA, et al. Dedicated monitoring of anesthetic and respiratory gases by Raman scattering. J Clin Monitoring 1986;2:215–222.
42. Westenskow DR, Coleman DL. Can the Raman scattering analyzer compete with mass spectrometers: an affirmative reply. J Clin Monitoring 1989;5:34–36.
43. Westenskow DR, Coleman DL. Raman scattering for respiratory gas monitoring in the operating room: advantages, specifications, and future advances. Biomed Inst Technol 1989;23:485–489.
44. Westenskow DR, et al. Clinical evaluation of a Raman scattering multiple gas analyzer for the operating room. Anesthesiology 1989;70:350–355.
45. Definition of Warburg Apparatus. Electronic Source, MedicineNet, Inc. Available at http://www.medterms.com/script/main/art.asp?articlekey=7150; Accessed 2005.
46. Scholand Pf, Hammel HT, Bradstre E, Hemmings E. Sap pressure in vascular plants - negative hydrostatic pressure can be measured in plants. Science 1965;148:339.

See also FIBER OPTICS IN MEDICINE; GAS AND VACUUM SYSTEMS, CENTRALLY PIPED MEDICAL; OXYGEN MONITORING.

OXYGEN SENSORS

STEVEN J. BARKER
University of Arizona
Tucson, Arizona

INTRODUCTION

This article reviews recent advances in the monitoring of patient oxygenation. We summarize the transport of oxygen from the atmosphere to the cell, and describe monitors that function at four stages of the O_2 transport process. These four stages include respired gas, arterial blood, tissue, and venous blood. History and recent developments in pulse oximetry will be discussed. Continuous intraarterial blood-gas sensors will be described and contrasted with other oxygen monitors. Finally, tissue oxygen monitoring and mixed-venous oximetry are discussed.

OXYGEN TRANSPORT IN THE HUMAN BODY

At rest, we consume $\sim 10^{23}$ molecules of oxygen per second. Our complex cardiopulmonary system has developed to rapidly transport this large amount of oxygen from the atmosphere to every cell in the body (Fig. 1).

The equation for arterial blood oxygen content (CaO_2) shows that $\sim 99\%$ of the oxygen in arterial blood is in the form of hemoglobin-bound oxygen:

$$CaO_2 = 1.38(SaO_2/100)(\text{Hb}) + 0.003\,PaO_2 \qquad (1)$$

where CaO_2 is in units of milliliters per deciliter of blood (also called vols%); SaO_2 is the arterial hemoglobin saturation in percent; Hb is the total hemoglobin concentration in grams per deciliter; and PaO_2 is the arterial oxygen tension (partial pressure) in millimeters of mercury. Upon inserting typical arterial values [$SaO_2 = 100\%$, Hb $= 15\ \text{g}\cdot\text{dL}^{-1}$, $PaO_2 = 100$ mmHg (13.33 kPa)], we find that normal CaO_2 is $\sim 21\ \text{mL}\cdot\text{dL}^{-1}$. The amount of oxygen delivered to the tissues in the arterial blood is then given by the cardiac output (CO) times the CaO_2 (neglecting the small dissolved oxygen term):

$$DO_2 = 13.8(\text{CO})(\text{Hb})(SaO_2/100) \qquad (2)$$

(The factor 1.38 becomes 13.8 because Hb is normally measured in grams per deciliter, while CO is measured in liters per minute. There are 10 dL in 1 L.)

Finally, the oxygen consumption by the tissues (VO_2) is determined by the difference between arterial oxygen delivery and venous oxygen return:

$$VO_2 = 13.8(\text{Hb})(\text{CO})(SaO_2 - SvO_2)/100 \qquad (3)$$

This Fick equation can be solved for any of the four oxygen variables involved.

If we substitute normal resting values into the equation, we predict a resting VO_2 of

$$VO_2 = 13.8\,(15\,\text{g}\cdot\text{dL}^{-1})(5\,\text{L}\cdot\text{min}^{-1})(99-75)/100$$
$$= 248\ \text{mL}\cdot\text{min}^{-1}$$

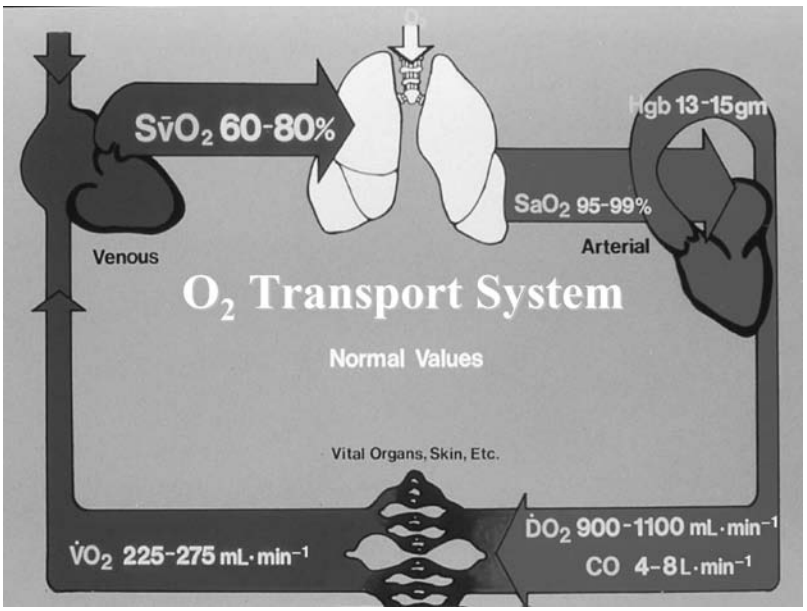


Figure 1. Schematic of the oxygen transport system. Arterial blood leaving the left ventricle (right side of figure) has and oxygen content of 21 mL·dL^{-1}, and the total oxygen delivery (DO_2) is roughly 1000 mL·min^{-1}. At rest, 0.25 of this oxygen delivery is consumed (VO_2), leaving a mixed venous saturation of 75%.

During exercise or stress we can rapidly increase cardiac output to at least 20 L·min^{-1}, and decrease venous saturation to $\sim$ 40%, yielding a VO_2 of

$$VO_2 = 13.8\,(15\,\mathrm{g}\cdot\mathrm{dL}^{-1})(20\,\mathrm{L}\cdot\mathrm{min}^{-1})(99-40)/100$$
$$= 2029\ \mathrm{mL}\cdot\mathrm{min}^{-1}$$

The human ability to rapidly adjust cardiac output (CO) and mixed-venous oxygen saturation (SvO_2) can be used to compensate for disease processes that affect other transport variables, such as anemia (hemoglobin) or hypoxemia (SaO_2). For example, consider a severely anemic patient with Hb value of 2.5 g·dL^{-1}, who compensates by increasing cardiac output to 15 L·min^{-1} and decreasing venous saturation to 50%:

$$VO_2 = 13.8\,(2.5\,\mathrm{g}\cdot\mathrm{dL}^{-1})(15\,\mathrm{L}\cdot\mathrm{min}^{-1})(99-50)/100$$
$$= 254\ \mathrm{mL}\cdot\mathrm{min}^{-1}$$

This extremely anemic patient can thus maintain a normal oxygen consumption by adaptations in CO and SvO_2 that are milder than those we use during normal exercise.

OXYGEN IN THE ARTERIAL BLOOD: PULSE OXIMETRY

Physiologic Considerations

The normal relationship between SaO_2 and PaO_2 is the familiar oxyhemoglobin dissociation curve, shown in Fig. 2. Three convenient points on this curve to remember are $PaO_2 = 27$ mmHg (3.60 kPa), $SaO_2 = 50\%$; $PaO_2 = 40$ mmHg (5.33 kPa), $SaO_2 = 75\%$; and $PaO_2 = 60$ mmHg (8.00 kPa), $SaO_2 = 90\%$. The curve will be shifted toward the right by acidosis, hypercarbia, or increasing 2,3-DPG. At PaO_2, values greater than 80 mmHg (10.66 kPa), SaO_2 is almost 100% and thus becomes virtually independent of PaO_2. It is important to remember this fact during SaO_2 monitoring in the operating room, where elevated inspired oxygen fraction (FIO_2) values will yield PaO_2 values much $>$ 80 mmHg (10.66 kPa) most of the time.

A knowledge of the relationship between SaO_2 and PaO_2 allows us to predict the physiologic limitations of saturation monitoring by pulse oximetry. Specifically, the pulse oximeter will give no indication of downward trends in PaO_2 during anesthesia at elevated FIO_2 until PaO_2 values $<$ 80–90 mmHg (10.66–12.00 kPa) are reached. In an animal study, intentional endobronchial intubations at F_IO_2 values $>$ 30% were not detected by the pulse oximeter (1). This results from the fact that the PaO_2 after endobronchial intubation did not decrease below $\sim$ 80 mmHg (10.66 kPa) when F_IO_2 was elevated.

Technology

Oximetry, a subset of spectrophotometry, determines the concentrations of various hemoglobin species by measuring the absorbances of light at multiple wavelengths. The absorbance spectra of the four most common hemoglobin species are shown in Fig. 3. The number of oximeter light wavelengths used must be equal to or greater than the number of hemoglobin species present in the sample. A laboratory CO-oximeter, which uses four or more wavelengths, can measure the concentrations of reduced hemoglobin, oxyhemoglobin, methemoglobin, and carboxyhemoglobin. If all four hemoglobin species are present in significant concentrations, then an oximeter must have at least four light wavelengths to determine the concentration of *any* of the four species.

The conventional pulse oximeter is a two-wavelength oximeter that functions *in vivo*. Conventional pulse oximetry first determines the fluctuating or alternating current (ac) component of the light absorbance signal. At each of its two wavelengths the oximeter divides the ac signal by the

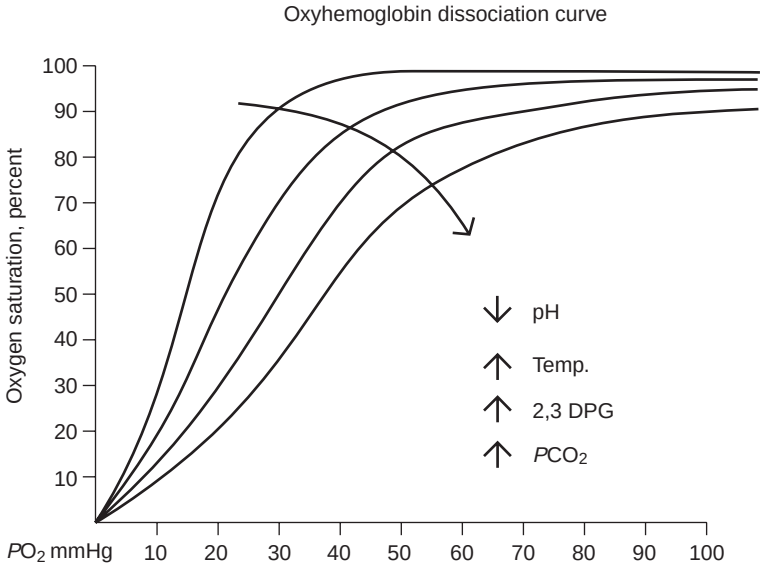


Figure 2. The oxyhemoglobin dissociation curve: a plot of hemoglobin oxygen saturation as a function of oxygen tension (PO_2). The curve is shifted to the right with decreasing pH or increasing temperature, PCO_2, or 2,3-DPG.

corresponding fixed or direct current (dc) absorbance component, to obtain a pulse-added absorbance. It then calculates the ratio of the two pulse-added absorbances (one for each wavelength), and this ratio R is related to arterial hemoglobin saturation by a built-in calibration algorithm. The resulting pulse oximeter saturation estimate is called SpO_2. The calibration curve of the pulse oximeter is empirical; that is, it is based on human volunteer experimental data.

Sources of Error

Dyshemoglobins. As previously noted, the pulse oximeter uses two wavelengths and can distinguish only two hemoglobin species: reduced hemoglobin and oxyhemoglobin. If either carboxyhemoglobin (COHb) or methemoglobin (MetHb) is present, the pulse oximeter effectively has fewer equations than unknowns, and it cannot find any of the hemoglobin concentrations. It is thus unclear *a priori* how the pulse oximeter will behave in the presence of these dyshemoglobins.

Two animal experiments have characterized pulse oximeter behavior during methemoglobinemia and carboxyhemoglobinemia. In one of these, dogs were exposed to carbon monoxide (220 ppm) over a 3–4 h period (2). At a COHb level of 70% (meaning that 70% of the animal's hemoglobin was in the COHb form), the SpO_2 values were $\sim 90\%$, whereas the actual SaO_2 was 30% (Fig. 4). The pulse oximeter thus "sees"

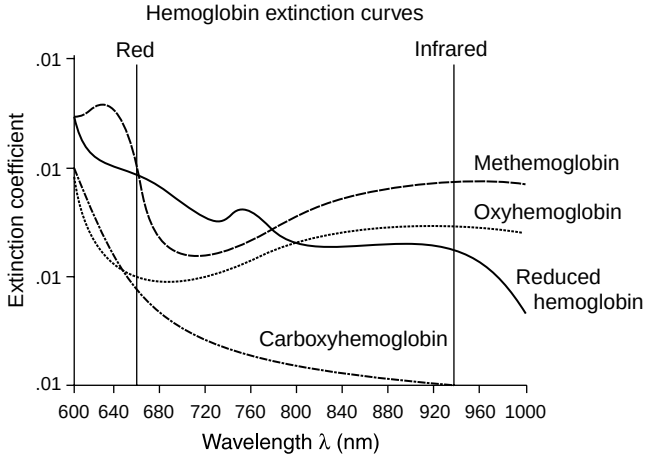


Figure 3. Extinction coefficient (or light absorbance) versus wavelength of light for four different hemoglobins. reduced Hb, O_2Hb, COHb, and MetHb. The two wavelengths used by most pulse oximeters (660 nm, 930 nm) are indicated by vertical lines.

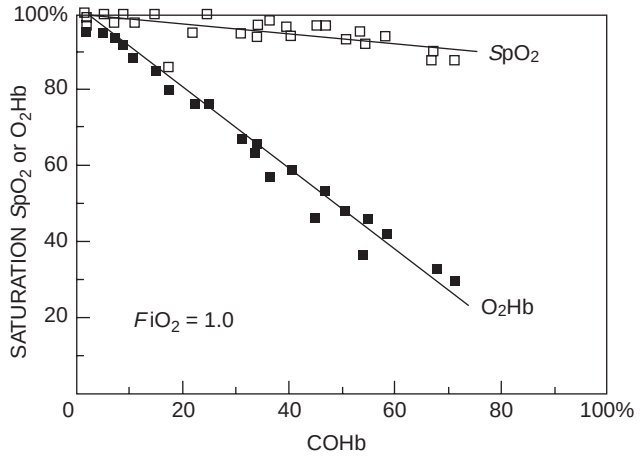


Figure 4. The effect of carbon monoxide on pulse oximetry. Plots of Hb saturation measured by ■ laboratory CO-oximeter, and □ conventional pulse oximeter, as functions of COHb level. As COHb increases, CO-oximeter shows linear decline in saturation, while pulse oximeter remains $> 90\%$ saturation (2).

carboxyhemoglobin as if it were composed mostly of oxyhemoglobin.

In a similar animal experiment, increasing methemoglobin concentrations (up to 60%) produced $S\text{pO}_2$ readings that gradually decreased to ~ 85% (3). As these animals were further desaturated by lowering the $F\text{IO}_2$ during methemoglobinemia, the pulse oximeter $S\text{pO}_2$ reading failed to track either functional or fractional saturation. On the other hand, the presence of fetal hemoglobin has little effect on pulse oximeter accuracy, which is fortunate in the treatment of premature neonates. There are conflicting anecdotal reports on the influence of sickle hemoglobin, and it is impossible to perform volunteer hypoxia studies on patients with sickle-cell disease.

As of this date, no commercially available pulse oximeter can either measure dyshemoglobins or function accurately in the presence of substantial levels of COHb or MetHb. Masimo Inc. has very recently (March, 2005) announced the development of a new Rainbow Technology pulse oximeter that uses multiple light wavelengths to measure COHb and $S\text{aO}_2$ simultaneously. There are as yet no published data regarding the success of this approach, but it is a potentially important new advancement.

Intravenous Dyes. As abnormal hemoglobin species can adversely affect the accuracy of pulse oximetry, so can intravenous dyes injected during surgery or critical care. Two studies found that intravenous methylene blue causes large, rapid decreases in displayed $S\text{pO}_2$ without changes in actual saturation, and that indocyanine green causes smaller false decreases in $S\text{pO}_2$ (4,5). Intravenous fluorescein or indigo carmine appeared to have little effect.

Reductions in Peripheral Pulsation; Ambient Light. Several studies have examined the effects of low perfusion upon $S\text{pO}_2$ (6,7). In a clinical study of critically ill patients during a wide range of hemodynamic conditions, extremes in systemic vascular resistance were associated with loss of pulse oximeter signal or decreased accuracy. During reduced pulse amplitude, pulse oximeters may become more sensitive to external light sources, such as fluorescent room lights (8). Most modern pulse oximeters effectively measure and correct for ambient light intensity.

Motion Artifact. Patient motion, which causes a large fluctuating absorbance signal, is a very challenging artifact for pulse oximetry. Motion artifact rarely causes great difficulty in the operating room, but in the recovery room and intensive care unit it can make the pulse oximeter virtually useless. Design engineers have tried several approaches to this problem, beginning with increasing the signal averaging time. Most current pulse oximeters allow the user to select one of several time averaging modes. However, improving motion performance by simply increasing averaging time is potentially dangerous—it can cause the instrument to miss significant, but short-lived hypoxemic events, which are very common in neonates.

Masimo, Inc. has developed a completely different approach to the analysis of the oximeter light absorbance signals, using adaptive digital filtering. This has led to improved accuracy and reliability during motion artifact, both in laboratory studies (9,10) and in the neonatal intensive care unit (11). The new technology has spurred other manufacturers (e.g., Nellcor, Philips, Datex-Ohmeda) to improve their signal analysis methods, so that today's generation of pulse oximeters has much improved performance during motion.

Venous Pulsations. Conventional pulse oximeter design is predicated on the assumption that the pulsatile component of tissue light absorbance is entirely caused by arterial blood. However, the light absorbance of venous blood can also have a pulsatile component, and this may affect $S\text{pO}_2$ values under some conditions (12). Conventional pulse oximeters may read falsely low values or may fail to give any reading in circumstances leading to venous congestion. This can occur, for example, when using an earlobe sensor on a patient who is undergoing mechanical ventilation, or who is in the Trendelenberg position.

Penumbra Effect. When a pulse oximeter sensor is not properly positioned on the finger or earlobe, the light traveling from the source to the detector may pass through the tissues at only grazing incidence. This penumbra effect reduces the signal/noise ratio, and may result in $S\text{pO}_2$ values in the low 1990s in a normoxemic subject. More importantly, a volunteer study has shown that in hypoxemic subjects, the penumbra effect can cause $S\text{pO}_2$ to either overestimate or underestimate actual $S\text{aO}_2$ values, depending on the instrument used (13). A pulse oximeter with a malpositioned sensor may therefore indicate that a patient is only mildly hypoxemic when in fact he or she is profoundly so.

OXYGEN IN THE ARTERIAL BLOOD: CONTINUOUS INTRAARTERIAL $P\text{O}_2$ MEASUREMENT

There have been a number of efforts to monitor intraarterial oxygen tension directly and continuously, using miniaturized sensors passed through arterial cannulas. The first practical approach to this problem employed the Clark electrode, the same oxygen electrode used in the conventional laboratory blood-gas analyzer. Although miniaturized Clark electrodes have been used in several clinical studies, the technique never achieved popularity because of problems with calibration drift and thrombogenicity (14). More recently, the principle of fluorescence quenching was used to develop fiberoptic "optodes" that can continuously monitor pH, $P\text{aCO}_2$, and $P\text{aO}_2$ through a 20 gauge radial artery cannula (Fig. 5).

Fluorescence quenching is a result of the ability of oxygen (or other substances to be measured) to absorb energy from the excited states of a fluorescent dye, thus preventing this energy from being radiated as light.

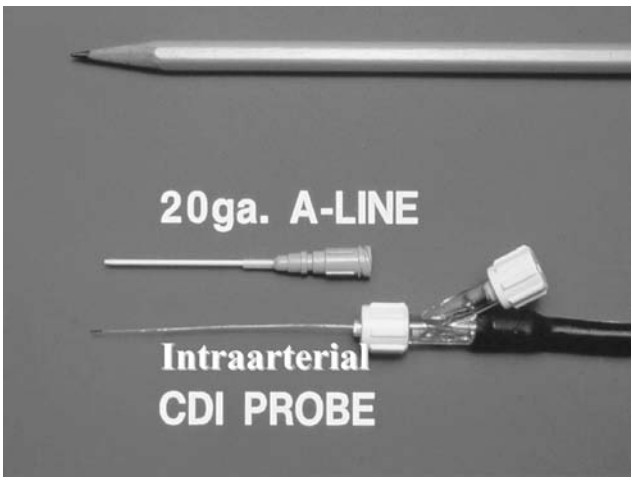


Figure 5. An intraarterial fiber-optic optode sensor. The optode is 0.6 mm in diameter and fits easily through a 20 gauge radial artery cannula, also shown.

Lubbers and Opitz (15) developed the first fluorescence quenching optode that simultaneously measured oxygen and carbon dioxide tensions in gases or liquids. In the 1980s, optodes were successfully miniaturized for intraarterial use, and several studies were reported in both animal and humans (16,17).

Clinical Studies

Several clinical studies suggested the usefulness of intraarterial optodes in the operating room (18). The scatter (random error) of optode oxygen tension values is lowest at low oxygen tensions, a characteristic of these sensors. The accuracy of the optode appeared to be within the clinically acceptable range when 18-gauge arterial cannulas were used. The optode can display complete blood-gas data continuously at the patient's bedside, with a time response measured in seconds. Nevertheless, the high costs of the disposable sensors ($\sim$ \$300 each) and their inconsistent reliability have caused the intraarterial optodes to disappear from the clinical market. These devices have other potential applications in tissues and organs, which may be realized in the future. One manufacturer today is marketing an optode sensor for assessment of the viability of tissue grafts.

OXYGEN IN TISSUE: TRANSCUTANEOUS OXYGEN

Physiologic Considerations

The transcutaneous oxygen ($P_{tc}O_2$) sensor is a Clark electrode that measures oxygen diffusing through the surface of the skin from dermal capillaries (Fig. 6). The sensor must be heated to at least 43 °C (in adults) to facilitate diffusion through the stratum corneum. Surface heating also produces local hyperemia of the dermal capillaries, which tends to "arterialize" the blood and cause a rightward shift in the oxyhemoglobin dissociation curve. The effects above tend to increase $P_{tc}O_2$, and these are counterbalanced by other effects that decrease it, namely

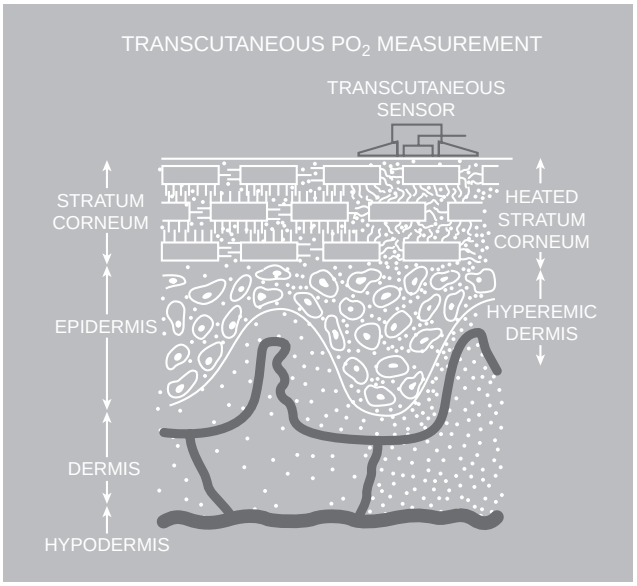


Figure 6. Schematic of transcutaneous PO_2 sensor on skin surface. Heat from sensor "melts" the diffusion barrier of the stratum corneum layer, and "arterializes" the blood in the dermal capillaries beneath.

diffusion gradients and metabolic oxygen consumption by the skin. In neonates, these competing effects nearly cancel and $P_{tc}O_2$ is approximately equal to PaO_2. In adults, the stratum corneum is thicker and hence the $P_{tc}O_2$ is usually lower than PaO_2. The transcutaneous index, $P_{tc}O_2/PaO_2$, has average values of 1.0 in neonates, 0.9 in pediatric patients, 0.8 in adults, and 0.6–0.7 in the elderly.

The most serious challenges with the interpretation of $P_{tc}O_2$ values are their dependence upon cardiac output and skin perfusion. Several studies have shown that the transcutaneous index falls when the cardiac index decreases below its normal range (19). Animal shock studies have shown that $P_{tc}O_2$ decreases when either PaO_2 or cardiac index decreases, and that it closely follows trends in oxygen delivery (i.e., the product of CO and CaO_2). In other words, $P_{tc}O_2$ monitors oxygen delivery to the tissues rather than oxygen content of arterial blood.

Technical Problems

There are several practical problems associated with the use of $P_{tc}O_2$ sensors. The transcutaneous electrode must be gas calibrated before each application to the skin, and then the sensor requires a 10–15 min warm-up period. In children, the warm-up period is usually shorter. The sensor membrane and electrolyte must be replaced at least once a week. The heated $P_{tc}O_2$ electrode can cause small skin burns, particularly at temperatures of 44 °C or greater. Lower probe temperatures (43 or 43.5 °C) should be used on premature infants and neonates, and the sensor location should be changed every 2–3 h. In adults with a sensor temperature of 44 °C, we have used the same location for 6–8 h with no incidence of burns.

Summary

Transcutaneous oxygen sensors provide continuous, noninvasive monitoring of oxygen delivery to tissues. By contrast, pulse oximetry provides continuous monitoring of arterial hemoglobin saturation. The dependence of $P_{tc}O_2$ on blood flow as well as PaO_2 sometimes makes it difficult to interpret changing values. If $P_{tc}O_2$ is normal or high, we know that the tissues are well oxygenated. When $P_{tc}O_2$ is low, we must determine whether this is the result of low PaO_2 or a decrease in skin perfusion.

OXYGEN IN VENOUS BLOOD: PULMONARY ARTERY OXIMETRY

Physiology of Mixed-Venous Saturation

Oxygen saturation in venous blood is related to venous oxygen content CvO_2 by an equation similar to equation 1:

$$CvO_2 = 1.38(\text{Hb})(SvO_2)/100 + 0.003(PvO_2) \qquad (4)$$

The normal CvO_2 value [with $SvO_2 = 75\%$, $PvO_2 = 40$ mmHg (5.33 kPa)] is 15.6 mL·dL^{-1}. If we solve equation 3 (the Fick equation) for the venous saturation (SvO_2), we obtain:

$$SvO_2 = SaO_2 - VO_2/[(13.8)(\text{Hb})(\text{CO})] \qquad (5)$$

Equation 8 shows how SvO_2 depends on the four oxygen transport variables: SaO_2, VO_2, Hb, and CO.

When VO_2 falls behind oxygen demand, lactic acidosis will result, eventually leading to death if the problem is not corrected. When this begins to occur in disease (e.g., anemia), the patient's body will try to maintain normal VO_2 using the same two compensatory mechanisms described above during exercise: increasing CO and/or decreasing SvO_2. In the case of anemia, we saw that such compensation can maintain normal VO_2 values even at hemoglobin levels < 3 g·dL^{-1}. Thus, a decrease in SvO_2 indicates that a patient is using oxygen reserves to compensate for a supply–demand imbalance. Decreasing oxygen supply may result from low CO (shock), low hemoglobin (anemia), abnormal hemoglobin (carboxyhemoglobinemia), or low PaO_2 (hypoxemia). On the other hand, increasing oxygen demand can result from fever, malignant hyperthermia, thyrotoxicosis, or shivering.

The aforementioned are possible clinical causes of a decrease in SvO_2. There are also conditions that can increase SvO_2 above its normal range of 68–77%. High SvO_2 values can result from decreased tissue uptake of oxygen, peripheral arteriovenous shunting, and inappropriate increases in CO. Clinical conditions that produce elevated SvO_2 values include sepsis, Paget's disease of bone, excessive use of inotropes, cyanide poisoning, and hypothermia. A wedged pulmonary artery catheter will also cause a high SvO_2 reading, but this is a measurement artifact. This can actually be a useful artifact, since it warns the clinician of an inadvertently wedged catheter.

Technical Considerations

Pulmonary artery SvO_2 catheters use the technology of reflectance spectrophotometry; that is, they measure the color of the blood in a manner similar to pulse oximetry. The SvO_2 catheters use fiberoptic bundles to transmit and receive light from the catheter tip. Light-emitting diodes provide monochromatic light sources at two or three wave-

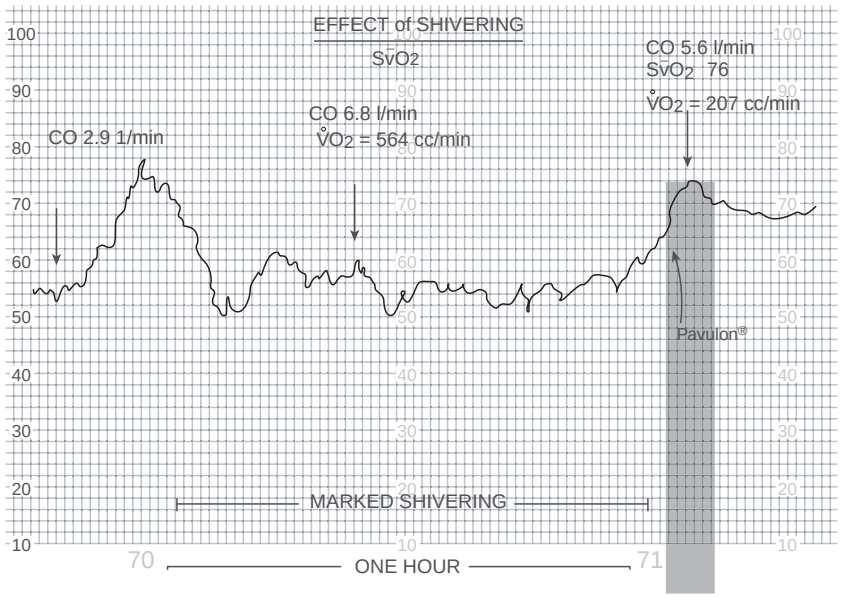


Figure 7. Effect of oxygen consumption (VO_2) upon mixed venous saturation (SvO_2). After weaning from cardiopulmonary bypass, the SvO_2 reaches a normal value of 75%, but then falls to 55%. Measured VO_2 at this time is 564 mL·min^{-1}, or twice normal resting value. Patient noted to be shivering. Administration of muscle relaxant stops shivering, restores VO_2 to normal (207 mL·min^{-1}), and SvO_2 also returns to normal values.

lengths. (Currently, the Edwards system uses two wavelengths, while the Abbott instrument employs three.) A theoretical advantage of a three wavelength system is that its measurements should not depend on the total hemoglobin level (20). Another problem common to all $S\text{v}O_2$ catheters is the so-called wall artifact, whereby reflection from a vessel wall can produce a signal that is interpreted as an $S\text{v}O_2$ of 85–90%. This problem has been reduced by the addition of digital filtering to the processor, which effectively edits out sudden step increases in $S\text{v}O_2$. However, a persistently high value of $S\text{v}O_2$ may alert the user that the catheter is in the wedged condition, as noted above.

Applications and Limitations

When interpreting continuous $S\text{v}O_2$ versus time tracings in the operating room and intensive care unit, we must always consider equation 5, the Fick equation solved for $S\text{v}O_2$. When $S\text{v}O_2$ changes, we should ask which term(s) in equation 5 are responsible. In the operating room, the terms most likely to change significantly are cardiac output (CO) and hemoglobin (Hb). During general anesthesia with mechanical ventilation, $S\text{a}O_2$ and $V\text{O}_2$ are usually constant, with the exception that $V\text{O}_2$ will decrease during hypothermia. On the other hand, this is not the case in the intensive care unit. Patients in respiratory failure will have varying degrees of arterial desaturation (low $S\text{a}O_2$). Note that $S\text{v}O_2$ is directly related to $S\text{a}O_2$; if $S\text{a}O_2$ decreases by 20% and nothing else changes, then $S\text{v}O_2$ will decrease by 20%. Critical care unit patients may also have frequent changes in $V\text{O}_2$, which can be increased by agitation, shivering, coughing, fever, pain, seizures, defecation, or eating, to name just a few (Fig. 7).

Continuous $S\text{v}O_2$ is a valuable adjunct in the treatment of ventilator-dependent patients. As positive end-expiratory pressure (PEEP) is slowly increased to improve oxygenation, $S\text{a}O_2$ will usually increase, but eventually the cardiac output will begin to decrease as venous return is compromised. At this point, oxygen delivery to tissue may begin to decrease (and $S\text{v}O_2$ begins to decrease) even though $S\text{a}O_2$ is still increasing. $S\text{v}O_2$ is a reflection of oxygen delivery in this situation, and can thus provide a means to optimize positive end-expiratory pressure without the need of serial blood gases and CO measurements.

In summary, continuous $S\text{v}O_2$ monitoring is a valuable technology for the operating room and the critical care unit. It reflects the overall health and functional state of the oxygen transport system. To realize the most benefit from this monitor, it is essential to thoroughly understand the physiology of $S\text{v}O_2$ and how it relates to the other oxygen transport variables.

CONCLUSIONS

Monitoring of oxygen in the respired gases and arterial blood is the standard of care during all anesthetics today. None of us would consider administering general anesthesia without both an $F\text{i}O_2$ monitor and a pulse oximeter. New advances in pulse oximetry will make these instruments more reliable in moving or poorly perfused patients, but they will still be subject to the fundamental limitations of saturation monitoring. Further developments will include pulse oximeters that can function in the presence of COHb and MetHb. In the near future, noninvasive monitors of oxygenation in specific organs and tissues (heart, brain) will become available. Finally, mixed venous oxygen saturation indicates how much is "left over" at the end of the oxygen transport process, which gives an indication of the status of the transport system and the degree to which reserves are being used.

BIBLIOGRAPHY

1. Barker SJ, Tremper KK, Hyatt J, Heitzmann H. Comparison of three oxygen monitors in detecting endobronchial intubation. J Clin Monitoring 1988;4:240–243.
2. Barker SJ, Tremper KK. The effect of carbon monoxide inhalation on pulse oximetry and transcutaneous PO_2. Anesthesiology 1987;66:677–679.
3. Barker SJ, Tremper KK, Hyatt J. Effects of methemoglobinemia on pulse oximetry and mixed venous oximetry. Anesthesiology 1989;70:112–117.
4. Sidi A, et al. Effect of fluorescein, indocyanine green, and methylene blue on the measurement of oxygen saturation by pulse oximetry (abstract). Anesthesiology 1986;65(3A): A132.
5. Scheller MS, Unger RJ, Kelner MJ. Effects of intravenously administered dyes on pulse oximetry readings. Anesthesiology 1986;65:550–552.
6. Lawson D, et al. Blood flow limits and pulse oximeter signal detection. Anesthesiology 1987;67:599–603.
7. Narang VPS. Utility of the pulse oximeter during cardio--pulmonary resuscitation. Anesthesiology 1986;65:239–240.
8. Eisele JH, Downs D. Ambient light affects pulse oximeters. Anesthesiology 1987;67:864–865.
9. Barker SJ, Shah NK. The effects of motion on the performance of pulse oximeters in volunteers. Anesthesiology 1997;86:101–108.
10. Barker SJ. Motion Resistant pulse oximetry. A comparison of new and old models. Anesth Analg 2002;95:967–972.
11. Bohnhorst B, Peter C, Poets CF. Pulse oximeters' reliability in detecting hypoxia and bradycardia: Comparison between a conventional and two new generation oximeters. Crit Care Med 2000;28:1565–1568.
12. Kim JM, et al. Pulse oximetry and circulatory kinetics associated with pulse volume amplitude measured by photoelectric plethysmography. Anesth Analg 1986;65:133–139.
13. Barker SJ, et al. The effect of sensor malpositioning on pulse oximeter accuracy during hypoxemia. Anesthesiology 1993;79: 248–254.
14. Rithalia SVS, Bennett PJ, Tinker J. The performance characteristics of an intraarterial oxygen electrode. Intensive Care Med 1981;7:305–307.
15. Lubbers DW, Opitz N. Die pCO_2/pO_2-optode: eine neue pCO_2 bzw. pO_2-Messonde zur Messung des pCO_2 oder pO_2

von Gasen and Flussigkeiten. Z Naturforsch 1975;30:532–533.
16. Barker SJ, et al. Continuous fiberoptic arterial oxygen tension measurements in dogs. J Clin Monitoring 1987;39: 48–52.
17. Barker SJ, et al. A clinical study of fiberoptic arterial oxygen tension. Crit Care Med 1987;15:403.
18. Barker SJ, Hyatt J. Continuous measurement of intraarterial pHa, $PaCO_2$, and PaO_2 in the operating room. Anesth Analg 1991;73:43–48.
19. Tremper KK, Waxman K, Shoemaker WC. Effects of hypoxia and shock on transcutaneous PO_2 values in dogs. Crit Care Med 1979;7:52.
20. Gettinger A, DeTraglia MC, Glass DD. *In vivo* comparison of two mixed venous saturation catheters. Anesthesiology 1987;66:373–375.

See also BLOOD GAS MEASUREMENTS; SAFETY PROGRAM, HOSPITAL.

OXYGEN TOXICITY. See HYPERBARIC OXYGENATION.

PACEMAKERS

ALAN MURRAY
Newcastle University
Medical Physics
Newcastle upon Tyne,
United Kingdom

INTRODUCTION

The primary function of a cardiac pacemaker is for the treatment of bradyarrhythmias, when the heart beat stops or responds too slowly. The clinical condition can be intermittent or permanent. If permanent, the pacemaker will control the heart continuously. If temporary, the pacemaker will respond only when necessary, avoiding competition with the heart's own natural response. As these devices are battery-powered, allowing the pacemaker to pace only when necessary also conserves pacemaker energy, extending its lifetime and reducing the frequency of replacement surgery.

Since their clinical introduction in the late 1950s and early 1960s, pacemakers have significantly improved the ability of many patients to lead normal lives. They also save lives by preventing the heart from suddenly stopping. The small size and long life of pacemakers allow patients to forget that they have one implanted in their chest. The first pacemakers were simple devices designed primarily to keep patients alive. Modern pacemakers respond to patients' needs and can regulate pacing function to enable the heart to optimize cardiac output and blood flow.

CLINICAL USE OF PACEMAKERS

The clinical problem with bradyarrhythmias is often associated with sick sinus syndrome. The heart's own natural pacing function originates from the sinus node in the right atrium. The rate of impulse formation at the sinus node is controlled by nerves feeding the node. Impulses arriving via the vagal nerve act to slow the heart down, as part of the body's parasympathetic response. When a higher heart rate is needed, the vagal nerve impulse rate to the sinus node is slowed down and the sinus node impulse rate increases.

Impulses propagating through the heart tissue are created via a series of action potential changes. Action potential changes can be triggered either naturally from one of the heart's pacing cells or by contact with a neighboring cell, causing the outside of the cell to produce a negative voltage with respect to the inside of the cell. These voltage changes are in the order of only 90 mV, but as we will see, an external voltage from a pacemaker has often to be several volts before depolarization is initiated.

Without pacemaker control, patients with bradyarrhythmias suffer from dizziness and can collapse without warning and, hence, risk injuring themselves. Heart pauses of the order of 10 s will cause unconsciousness. Most patients who collapse will recover their normal heart rhythm. They can then subsequently be examined clinically, and, if necessary, a pacemaker can be implanted to prevent recurrence of a further collapse. Sometimes the heart will stop and not recovers its normal pumping function and the patient will die, but usually there will have been preceding warning events allowing a pacemaker to be fitted to prevent death.

Many good texts exist that explain the clinical background to cardiac pacing, and these texts should be consulted (1–3). This text is primarily a description of the medical device itself.

Practical cardiac pacing started in the 1950s with the first clinical device, which was external to the body and required connection to a main power supply. This device was followed by an implantable pacemaker developed by Elmquist and surgically implanted by Senning in Sweden (4). The device only lasted a short time before failing, but it did show the potential for implanted pacemakers. This work was followed by Greatbatch and Chardack in the United States (5,6), first in an animal and then in a patient two years later. An interesting early review of this period has been given by Elmqvist (7). These first pacemakers were very simple devices and paced only at a fixed rate, taking no account of the heart's natural rhythm. Although this approach was less than ideal, it did provide the necessary spur for both clinical expectations and technical and scientific developments by research bioengineers and industry.

The next major technical development allowed pacemakers to pace on demand, rather than only at a fixed rate. Other pacing functions developed, including pacemakers that could pace more than one heart chamber, and pacemakers that could change their response rate as a function of patient physiological requirements. Three- or four- chamber pacing was an extension of basic pacing. Pacing functions have also been included in implantable defibrillators. More complicated pacing algorithms have been developed for controlling tachyarrhythmias, including ventricular tachycardia and rhythms that can deteriorate to ventricular fibrillation.

With the evolution of smaller devices and leads, their use in pediatrics has grown, including for children with congenital heart problems. Devices as thin as 6 mm are available. Reduction in size has also aided the move from epicardial to endocardial fixation of the lead. When pacemakers are implanted in children, special consideration has to be given to the type of device as children are usually active, the lead length as children continue to grow, and lead fixation as future lead replacement must be considered.

No doubt exists that, with continuing experience, pacing techniques and pacemaker devices will continue to evolve.

PHYSIOLOGICAL FUNCTION

Understanding the physiological function of a pacemaker is more important than knowing the technical details of the pacemaker. The main functional characteristics are the ones that are important for the physician or cardiologist who will want to know how the device will operate when implanted in a patient. A series of clinical questions exist. The first question asks where you want the device to sense the heart's rhythm, in the ventricle, the most common location, in the atrium, or in both. The second question asks in which chamber or chambers you would like the pacemaker to pace. This location is usually the ventricle, but can be the atrium or both. The third question relates to how you want the device to work when it encounters natural heartbeats. It can either be inhibited, which is by far the most common approach, or it can be triggered to enhance the natural beat. It is also possible to switch off the device's ability to sense the heart's natural rhythm, but is rarely done as there could then be competition between the pacemaker output and the heart's natural rhythm.

The answers to these three questions provide the first three codes given to any pacemaker. This code is an international code developed by the Inter-Society Commission on Heart Disease (ICHD) (8). It was subsequently expanded to a five-code system by the North American Society of Pacing and Electrophysiology (NASPE) and the British Pacing and Electrophysiology Group (BPEG) (9,10). The first version of the NASPE/BPEG codes allowed for programmability and communication, but as they became universal functions, the latest version of the codes simplified the codes in the fourth and fifth letter positions for use with rate modulation and multisite pacing only. These codes are used throughout the world. A summary of the coding is given in Table 1.

It is useful to give a few examples to illustrate how the codes are used. VVI pacemakers, which are in common use, allow the pacemaker to sense natural heartbeats in the ventricle (V), and, if they are absent, to pace in the ventricle (V), ensuring that the pacemaker inhibits (I) its output if a natural beat is detected. DDD pacemakers can sense in both the atrium and the ventricle (D), and, if required, pace in the atrium, ventricle, or both (D), with inhibiting and triggering (D). With programming techniques, the pacemaker's mode can be changed after the device is implanted, and so a manufacturer may list a very large number of modes for some pacemakers.

The codes in Table 1 also show the fourth and fifth letters. The fourth tells the user if the device has an internal function for modulating its pacing rate, known as a rate responsive (R) mode. If no code is quoted in the fourth position, it can be assumed that the device is not rate responsive. The fifth letter is for multisite pacing and is used if at least two atrial pacing sites or two ventricular pacing sites exist.

TEMPORARY EXTERNAL PACEMAKERS

This review primarily concerns implantable pacemakers, but the role of temporary external pacemakers must not be forgotten. These devices provide essential support after some cardiac surgery and after some myocardial infarctions, allowing time for the recovery of the heart's own pacing function. The way these devices function is very similar to implantable pacemakers, but they are generally simpler and provide the clinician with access to controls such as for pacing rate and pacing voltage. The pacing leads do not have the tip features required for permanent fixation, and the connector to the temporary pacemaker is simpler. Also, the leads are bipolar with two electrode contacts.

CLINICAL IMPLANTATION

Briefly, pacemakers are implanted most commonly at one of three sites (Fig. 1). Implantation is undertaken by a surgeon or cardiologist using an aseptic technique. The pacemaker pulse generator and leads are delivered in sterile packages with clear use-by dates. Venous insertion of the lead allows it to be pushed through the right atrium and tricuspid valve, and into the right ventricle, where the electrode can be positioned in the apex where it is less likely to move or displace in comparison with other possible positions. If an atrial lead exists, it is positioned in the right atrium. Good contact with the atrial wall is harder to achieve, and active fixation such as with a screw contact can be used, in comparison with ventricular apex passive fixation.

Patients must be followed up at regular intervals to ensure that the device is working correctly, that its output pulse characteristics are appropriate, and that the end-of-life of the internal battery is estimated. This follow-up interval may be over a period of months initially, and then annually, with more frequent follow-up visits toward the end of the device's life.

Most countries have national registration schemes, which enables information on specific patients to be obtained if, for example, a patient develops a problem while away from home. If, however, this information is not available, the device type can be recognized by a unique

Table 1. International Pacemaker Codes

1	2	3	4	5
Chamber sensed	Chamber paced	Response	Rate modulation	Multisite pacing
O = none	O = none	O = none	O = none	O = none
A = atrium	A = atrium	I = inhibited	R = rate modulation	A = artium
V = ventricle	V = ventricle	T = triggered		V = ventricle
D = A+V	D = A+V			D = A+V

The following letters are used sequentially in 5 positions.

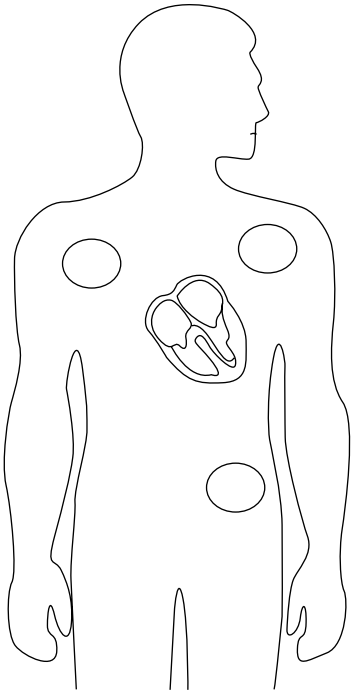


Figure 1. The location of normal pulse generator implantation sites.

radiopaque code that can be obtained by X ray. There has been discussion on whether such codes could be retrieved by a standard external interrogating device without the need for an X ray, but no such device is as yet universally available for all pacemakers.

Another goal of pacemaker registration is to provide useful data on the range of device lifetimes for each pacemaker type and information on sudden pacemaker failures, which enables clinical staff to plan any necessary replacement, and manufactures to act when it appears that failures are not random and may relate to their manufacturing process. This advance has enabled manufacturers to withdraw faulty or potentially faulty devices from the marketplace and correct production faults.

MARKET

Pacemakers have made a remarkable impact on clinical medicine. Over half a million new patients worldwide receive a pacemaker each year (11). In addition, approximately 100,000 patients worldwide receive a replacement pacemaker (11). Most implants are in the United States. When implants are related to the population size, the countries with the greatest new implant rates are Germany, the United States, and Belgium, with between approximately 700 and 800 implants per million population. Many countries with poor economies have very low implant rates. Within Europe, the implant rates are generally high, with, for example, the United Kingdom falling towards the bottom end of the implant rate, at approximately 300 per million population (12), where there are approximately 25,000 implants per year, and, of these, 75% are for new implants and 25% for replacements (12).

Table 2. Example Ranges of Pulse Generator Features

Volume	6–20 ml
Length/Width	30–60 mm
Depth	6–14 mm
Mass	13–50 g
Battery	0.8–2 Ah
Life	5–14 years
Sense threshold	0.1–15 mV
Refractory period	100–800 ms
Lower pulse rate	approximately 20/min
Upper pulse rate	approximately 185/min
Pulse amplitude	0–10 V
Pulse width	0.1–2 ms

FEATURES

Clinically, the most important pacemaker features relate to the device code, discussed above. Next in importance for both the clinician and patient is likely to be pacemaker size and lifetime. As a guide, example ranges of pacemaker features are included in Table 2. With continuous developments, these should be taken only as a guide. The shape and size of some pulse generators are shown in Fig. 2.

For a health-care system, the cost of devices is important. Costs vary significantly in different countries and also relate to the numbers purchased, so no specific figures can be given. It is, however, interesting to note the current relative costs of different types of devices. For guidance, approximate costs relative to a standard ventricular demand pacemaker type VVI are shown in Fig. 3, where the cost of the VVI pulse generator is given as unity. As the proportion of different types used will change, so will the relative costs.

The use of the unique radiopaque code for each pulse generator type is useful when a patient is referred to a medical center away from home, allowing that center to determine the pacemaker pulse generator being used.

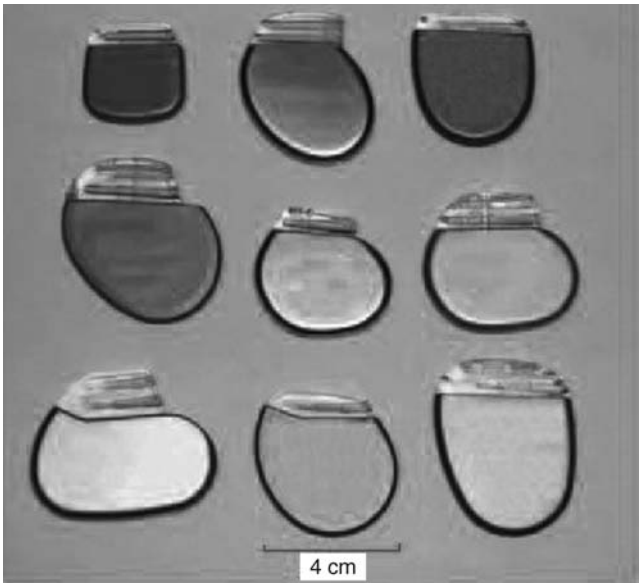


Figure 2. Illustration of some pacemaker shapes and sizes.

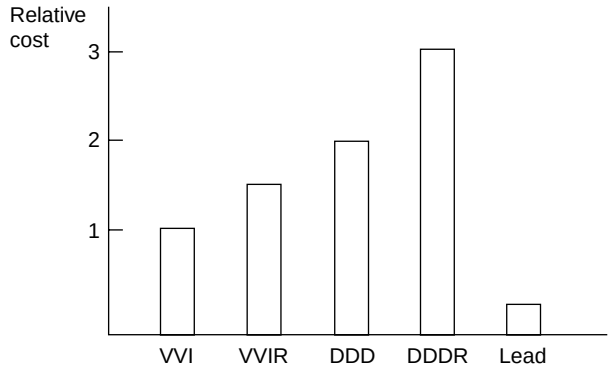


Figure 3. Relative approximate costs of pulse generators and pacing leads. The demand pacemaker (VVI) is taken as the reference.

A unipolar device has only one electrode directly in contact with the heart. In this case, the electrode is at the distal end of the lead. To complete the pacing circuit, another electrode contact is required, and this is on the pacemaker case with current flowing via the muscle in contact with the case electrode to the heart. Bipolar electrodes are also common. Here, both electrode contacts are on the lead, one at the tip and another several centimeters away. The second electrode makes contact with the ventricular wall simply by lying against the wall with the tip firmly located at the apex of the ventricle.

PACEMAKER COMPONENTS

A pacemaker refers to all components necessary for a complete clinical pacing device. At least two components will always exist, the pulse generator and the lead (Fig. 4). More than one lead any exist, such as for dual-chamber pacing, in both the atrium and ventricle. Unusual pulse generator and lead combinations may require an adaptor, but extra components should be avoided whenever possible. Extra components add to the areas where failure might occur.

When a pacemaker has been implanted or is, subsequently, to be checked, external devices will be required. An ECG recorder will confirm correct pacing, and some pacemakers generate impulses that can be visualized on an ECG recorder for relaying pacing information. In addition, a magnet or other technique may be used to put the pacemaker into various test modes, which is now commonly achieved with a programmer that communicates with the pacemaker via an electromagnetic wand using communication technology with coded sequences to prevent external interference such as from a radio telephone accidentally reprogramming the pacemaker. As well as controlling the pacing functions, the programmer can interrogate the pacemaker about the frequency of pacing, provided, of course, these features are available. Telemetry may also be available, where intracardiac waveforms can be relayed as they occur and selected pacing episodes can be recovered from the pacemaker memory.

Further useful technical information can be obtained from the books by Schaldach (13) and Webster (14).

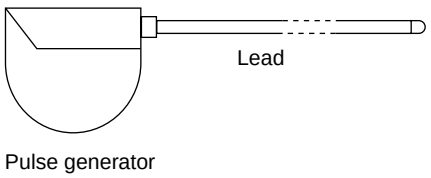


Figure 4. Pulse generator and lead are the two main components of a pacemaker.

PULSE GENERATOR

When pulse generators were first used in the early 1960s, they were simple devices with a battery power supply and a circuit to produce a regular pulse rate with a defined pulse voltage and pulse width output. Modern pulse generators are much more complex, with sensing and output control. Special electronic circuitry has been developed, often with sophisticated microprocessor control. Battery technology has also developed significantly. A block diagram of a complete pulse generator is shown in Fig. 5. Each major part is now described.

Power Supply

The first battery power supplies were made up from separate zinc-mercury cells. They could often be seen through the casing before implantation, after removal, or on the X ray. The voltage of these cells quickly fell to approximately 1.35 V, which was held until the cell reached the end of its life. Unfortunately, these cells could power the early pacemakers for only about two years.

These batteries encouraged the search for other power sources, including, for a short time, nuclear power, but safety concerns discouraged these developments. Rechargeable batteries, where the recharging energy was transmitted to the pulse generator via an external coil, were also employed, and had in fact been used in the first clinical pacemaker. However, reliability and frequency of charging inhibited their use.

Fortunately, a solution was found in the form of lithium-iodide cells. Their use in pacemakers was pioneered by Greatbatch (15) and introduced into clinical use in 1971. The cell has an initial open circuit voltage of 2.8 V, which falls slowly with use until its end of life is approached, when the voltage fall is more rapid. Although other types of lithium cells have been researched, they have not replaced the lithium-iodide cell for pacemakers. Some pacemakers

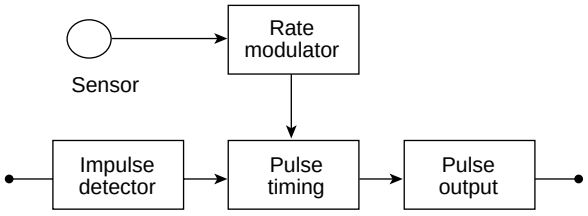


Figure 5. Block diagram of a complete pulse generator.

can now function for ten years, and, for low demand situations, 14 years can be achieved.

Ventricular and Atrial Sensing

All pacemakers sense natural cardiac impulses, which is achieved very reliably. However, the sensing circuitry needs to be able to differentiate between impulses coming from the heart and those due to external interference. The possibility of interference cannot be neglected, especially as the cardiac impulse is, at most, only a few tens of millivolts. In addition, in unipolar pacing, the sensing circuitry cannot differentiate between signals coming from the electrode at the end of the lead or the electrode on the pulse generator case, which is described more fully under leads below. It is not always possible to detect heart impulses reliably during periods of excessive interference, and in these cases, the device needs to be programmed to respond in a clearly defined way, such as by initiating fixed-rate pacing.

Pulse Interval Generator

At the heart of the pacemaker is the interval generator. As with other pulse generator functions, this is achieved with electronic circuitry or microprocessor control that generates the pacing interval measured in seconds or pacing rate measured in beats per minute. Without any means to detect patient activity or the need for a higher heart rate, this interval is fixed, other than for programmed changes that can be made at the clinic. The interval generator is connected to the sensing circuitry so that output from the interval generator can be synchronized or inhibited. Interactions for a VVI pulse generator are illustrated in Fig. 6. If a natural heartbeat is sensed, the pacemaker will time its pulse interval from that beat. If, however, the beat occurs in the refractory period set in the pacemaker, it will not respond to the beat, and the pulse interval timing will not be altered. For devices that can pace in both the atria and ventricle, such as a DDD device, the interval between atrial and ventricular pacing, known as the AV interval, will be set to optimize heart pumping function (Fig. 7).

Interval Modulator

The interval modulator, if it is available, simply attempts to change the response of the interval generator to that required by the patient. Electronic modulation of the interval is easy to achieve, and this function will usually be seen as part of the interval generator. However, determining how to modulate is difficult, and it requires a sensor to detect the patient's need for a higher or lower heart rate, usually in response to patient activity, whether determined directly or indirectly via physiological functions.

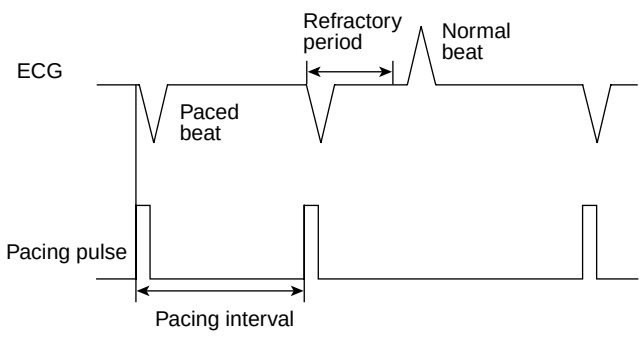


Figure 6. Illustrative example of some pacemaker-heart interactions.

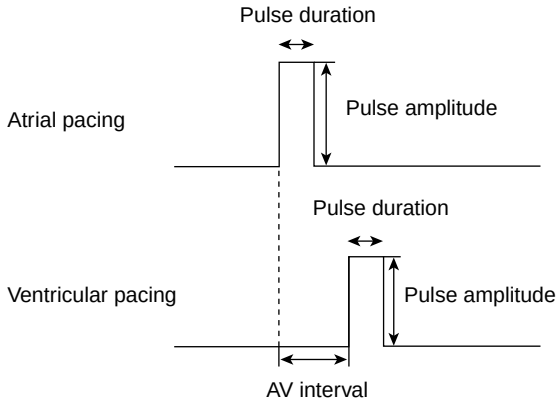


Figure 7. Atrioventricular pacing with the AV interval set within the pacemaker.

Activity Sensor

Many different sensors have been used. Some have been more successful than others. Changes in blood oxygen levels, through oxygen partial pressure or saturation levels, suggest the need for a higher heart rate when these levels fall, but this technique requires a sensor with continuous blood contact that will continue to work over many years, a difficult specification to achieve. Increased respiration rates also suggest the need for a higher rate. To detect respiration, some devices used a special lead to detect respiratory movement or electrical impedance changes but were not always successful. Special leads with inbuilt sensors add to the complexity of the pacemaker. Respiration can now be monitored using electrical impedance changes from a standard lead.

Other successful techniques used the intracardiac electrogram obtained from the sensing electrode or activity sensing from an inbuilt accelerometer. With careful analysis of the electrogram, it is possible to obtain a measure of the repolarization interval, which is known to change systematically with heart rate changes and, for this application, has been shown to shorten even without any increase in paced heart rate when the need for a higher rate is physiologically required. The use of a motion sensor using a piezoelectric sensor has been successful and is most widely applied because the technology is simple. The sensor can be built into the pulse generator housing with better reliability than for techniques requiring additional patient contacts. Movement of the sensor produces an output voltage proportional to acceleration. Algorithms have then been developed to relate changes in sensed activity to changes in pacing interval. They can also, to some extent, be programmed to individual patients. One drawback is that the sensor senses movement of the pacemaker that can be in a patient, for example, driving a car

where an increase in heart rate may not be appropriate. Improvements to deal with such situations are under continuous development, as is research into different sensor techniques.

Without doubt, rate-responsive pacemakers have made a great contribution, and patients welcome the ability of the pacemaker to adapt to their needs, even if the pacemaker response is not physiologically perfect (12).

Lead Connector

Leads and pulse generators are provided separately, which gives greater flexibility in their use and enables different lead lengths and lead types to be selected. However, a connector is then needed. Ideally, this connector should be able to connect any appropriate lead to any pulse generator, and international standards have gone a long way to achieving this. For single-chamber pacemakers, the connector takes one lead that can be either for a unipolar or bipolar lead with one or two electrode contacts.

It is important for the connector to make a good electrical contact, while preventing any body fluids penetrating into the pulse generator, which in the past, have caused pacemakers to fail. Connectors allow the electrical contact to be made easily and provide seals between leads and the pulse generator.

Case

The primary function of the case is to protect the inner electronics from mechanical damage and from penetration of blood or other fluids. It is essential for the case to be biocompatible so that the patient does not attempt to reject the pulse generator as a foreign body. Titanium has been a successful material.

For unipolar devices, the case must contain an electrode, which acts as the reference for the lead electrode. The sensed voltage is that between the case and the lead electrode. Also, when pacing, the stimulation voltage appears between the case and lead electrode. To minimize the possibility of muscle stimulation at the case electrode, this electrode has a large surface area, so reducing the current density in comparison with that at the lead electrode. Also, the output pulse is positive at the case electrode with respect to the negative voltage at the pacing electrode, which confers preference to stimulation at the negative site in the heart.

Telemetry Function and Programming

For programming and telemetry functions, the pacemaker needs to be able to communicate with an external device, usually with coded signals using electromagnetic transmission between the pacemaker and a wand of the programming unit. These devices use standard techniques, with only the coding being specific to pacemakers.

Computer Algorithms

Pulse generators usually can be seen as small microprocessor devices. As such they contain computer code or computer algorithms to control their function, delivering advantages to the patient and clinician, as the pacemaker's mode of operation can be changed, and also to the manufacturer, as it is easier to develop new and improved devices. However, reliable software is notoriously difficult to develop and test. Manufacturers have discovered, to their cost-unusual errors in their software only after devices have been implanted, necessitating a recall of devices not yet implanted and careful follow-up of patients with devices already implanted. High quality software development is taken seriously by the manufacturers and cannot be stressed enough.

Table 3. Example Ranges of Lead Features

Length	20–120 cm
Diameter	1.2–3.5 mm
Tip diameter	0.7–3.3 mm

LEAD

The lead has four main features. It needs a connector to connect it to the pulse generator, a long flexible wire, a biocompatible sheath over the wire, and at least one electrode to make contact with the heart. Table 3 provides illustrative ranges of lead features.

Connector

The connector needs to be compatible with that on the pulse generator. As with the pulse generator, there should be no ingress of fluid, this time into the wire. Also, as the wire can move with each heartbeat, the construction needs to ensure that no extra stress exists on the lead wire near the connector.

Lead Wire

The most important characteristic of the lead wire is that is has to be flexible. Normal wire easily fractures when bent repeatedly. A pacing lead wire has to move with each heartbeat, which averages approximately 100,000 movements each day. Good flexibility is achieved by using a spiral construction. All wires have some impedance, which is taken into account by the pulse generator output.

Insulated Lead Sheath

The sheath covering the lead wire also needs to be flexible and must not become brittle with age. The sheath material must be biocompatible so as not to be rejected by the body. Materials used are silicone rubber or polyurethane.

Electrode

The design of the electrode is very important. In particular, the fixation, contact area, and contact material are essential features. Illustrative examples of basic features of lead-tip electrodes are shown in Fig. 8. Unipolar electrode leads have a single-electrode contact at the tip. Bipolar electrode leads have two contacts, one at the tip and the other a few centimeters distant from the tip.

When the lead is implanted and a suitable electrode site found, the electrode needs to stay in position, which is

Figure 8. Illustrative example of pacing electrodes.

achieved by mechanical features at the tip of the lead, which can be, for example, tines or a helical electrode construction. With time, tissue will grow over the tip holding it in place. One problem is that this fixation can become so good that the lead can be difficult to remove if a problem occurs and it needs to be replaced. Electrodes positioned in the ventricular apex are easy to locate and also tend to stay in position easily, and hence require only passive fixation, such as with tines at the end of the lead. Other tip locations may require active fixation such as with a screw tip.

The electrode area needs to be high enough to ensure good electrical contact. The greater the contact area, the lower the contact impedance, which in turn reduces the electrode-tissue interface impedance and ensures that most of the pulse generator voltage appears at the cardiac tissue.

As with any electrode, the electrode contact material is important. The aim in selecting the material is to reduce polarization effects. Many electrode-coating materials have been studied, including steroid-eluting electrodes to reduce inflammation. Changes in electrode polarization are the cause of the increase in stimulation voltage in the days and early months after implantation, to be subsequently followed by a lowering of the effect and also of the required stimulation voltage.

STIMULATION THRESHOLDS

Stimulation success is a function of both pulse amplitude and pulse width. A minimum voltage and energy is required. The voltage has to be greater than that required to initiate the approximate 90 mV change in action potential. However, because of polarization and other effects, the voltage required is usually in the order of several volts and can reach 10 V soon after implantation. After a few months, this voltage will have reduced to the level of a few volts.

The initial research on stimulation pulse energy was with stimulating nerves, but the results obtained have been shown generally to hold when stimulating or pacing cardiac tissue. The energy used should be the minimum possible to induce stimulation reliably, which is controlled by varying the pulse width. Early work on nerve stimulation showed that no matter how wide the pulse width was, a minimum pulse voltage existed, called the **rheobase** voltage. At about twice this voltage, with a lower pulse width, the minimum energy required is found. If the pulse width is reduced further, the greater voltage required results in increased energy requirements to induce pacing. The pulse width for lowest energy is called the chronaxie time, shown in Fig. 9.

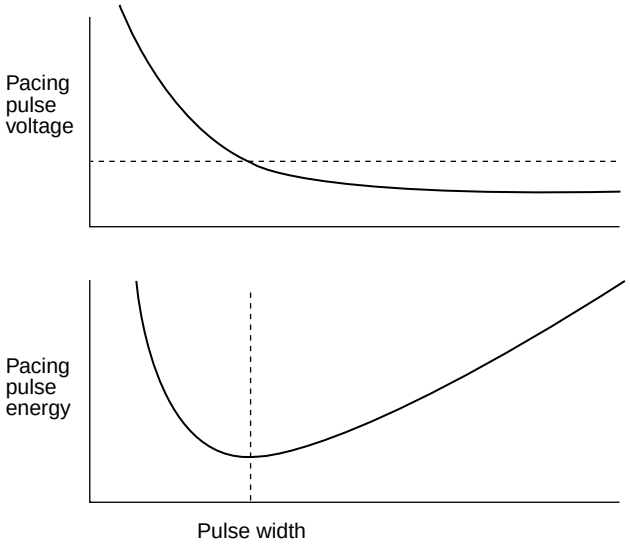


Figure 9. Pulse energy as a function of pulse width.

As the pulse energy is only to initiate cardiac depolarization, and not for providing pumping energy, the energy levels required are low and on the order of only a few microjoules (16).

PROBLEMS IN USE

Interference

Interference is a well-understood problem. Muscle interference at the case electrode in a unipolar system can be a problem in active people, especially when the pulse generator is implanted beside the pectoral muscles. If this particular problem is anticipated, a bipolar system can be used.

Threshold Voltage Changes

Threshold voltages do change. If the pacing voltage is set too high, energy will be wasted, reducing the lifetime of the device. If set too low, changes in threshold voltage may result in the pacing voltage being below the pacing threshold. These factors have to be balanced. Some devices automate the selection of an appropriate voltage output.

Early Failure

Pacemakers are complex devices, and like all devices they can fail. Failure is not a common problem, but because pacemakers are implanted and are life supporting, failure can have fatal consequences. Reporting of individual problems is essential, allowing manufacturers and national health bodies to identify a common problem early and, if necessary, withdraw stocks before they are used in new patients and take action to review patients who already have the device implanted.

SPECIAL DEVICES

This review has concentrated on the main use of pacemakers for treatment of bradyarrhythmias. Other options

are used, but these options are based on the standard pacing approaches.

Implantable defibrillators can have an additional pacing function so that if the heart stops rather than developing ventricular fibrillation, pacing can be initiated. The pacing technology is exactly the same as for pacemakers described above, except that the electrode system will be different.

Devices are used for control of tachyarrhythmias. These devices, rather than using the regular pacing interval, usually use a series of pacing intervals at different rates to terminate the arrhythmia.

Some patients in heart failure were, in the past, often assumed to be untreatable unless by heart transplantation. Much can now be done for these patients, including pacing in all four cardiac chambers, which maximizes the pumping function of the heart by pacing the left as well as the right heart chambers, and pacing the atria and ventricles with an appropriate atrio-ventricular delay. This solution requires complex and multiple leads, and as these leads are used in sick patients, success is not always assured. Many of these patients may also require a defibrillation function.

FUTURE

Cardiac pacing had a small beginning but has grown at a steady rate each decade. With an aging population, the need for pacing will continue to grow. The development and production of pacemakers will remain a major medical device industry.

Of those devices currently available, increased use of physiological or rate responsive devices is likely as clinical studies prove their clinical value to patients, especially those who are active.

Technical advances will, to some extent, be dependent on the production of improved batteries, and then the decision will be either to make them smaller, last longer, or power more microprocessor technology. Improved electrode design to reduce energy requirements could also make a significant impact in reducing pacing pulse energy, and hence overall energy requirements. Improvements in setting the optimum AV delay will help many patients and, in particular, children who are active. Increased ability to store intracardiac data for review will ensure more research into effective use of pacing.

Pacing will continue as an essential therapeutic technique, saving lives and bringing some normality to patients with abnormal physiological heart rate control.

BIBLIOGRAPHY

1. Trohman RG, Kim MH, Pinski SL. Cardiac pacing: The state of the art. Lancet 2004;364:1701–1719.
2. ACC/AHA/NASPE 2002 guideline update for implementation of cardiac pacemakers and antiarrhythmia devices. American College of Cardiology and the American Heart Association; 2002.
3. Gold MR. Permanent pacing: New indications. Heart 2001;86: 355–360.
4. Senning A. Problems in the use of pacemakers. J Cardiovasc Surg 1964;5:651–656.
5. Greatbatch W, Chardack W. A transistorized implantable pacemaker for the long-term correction of complete heart block. Trans Northeast Electron Res Eng Meet Conf 1959; 1:8.
6. Chardack WM, Gage AA, Greatbatch W. A transistorized, self-contained, implantable pacemaker for the long-term correction of complete heart block. Surgery 1960;48:643–654.
7. Elmqvist R. Review of early pacemaker development. PACE 1978;1:535–536.
8. Parsonnet V, Furman S, Smyth NP. Implantable cardiac pacemakers: Status report and resource guideline (ICHD). Circulation 1974;50:A21–35.
9. Bernstein AD, Camm AJ, Fletcher RD, Gold RD, Rickards AF, Smyth NPD, Spielman SR, Sutton R. The NASPE/BPEG generic pacemaker code for antibradyarrhythmia and adaptive-rate pacing and antiachyarrhythmia devices. PACE 1987;10:794–799.
10. Bernstein AD, Daubert J-C, Fletcher RD, Hayes DL, Luderitz B, Reynolds DW, Schoenfeld MH, Sutton R. The revised NASPE/BPEG generic code for antibradycardia, adaptive-rate, and multisite pacing. PACE 2002;25:260–264.
11. Mond HG, Irwin M, Morillo C, Ector H. The world survey of cardiac pacing and cardiovertor defibrillators: Calendar year 2001. PACE 2004;27:955–964.
12. National Institute of Clinical Excellence. Technology Appraisal 88, Dual-chamber pacemakers for symptomatic bradycardia due to sick sinus syndrome and/or artrioventricular block. London, UK: National Institute of Clinical Excellence; 2005.
13. Schaldach M. Electrophysiology of the Heart: Technical Aspects of Cardiac Pacing. Berlin: Springer-Verlag; 1992.
14. Webster JG, ed. Design of Cardiac Pacemakers. Piscataway, NJ: IEEE Press; 1995.
15. Greatbatch W, Lee J, Mathias W, Eldridge M, Moser J, Schneider A. The solid state lithium battery. IEEE Trans Biomed Eng 1971;18:317–323.
16. Hill WE, Murray A, Bourke JP, Howell L, Gold R-G. Minimum energy for cardiac pacing. Clin Phys Physiol Meas 1988;9: 41–46.

See also AMBULATORY MONITORING; BIOELECTRODES; BIOTELEMETRY; DEFIBRILLATORS; MICROPOWER FOR MEDICAL APPLICATIONS.

PAIN CONTROL BY ELECTROSTIMULATION.

See TRANSCUTANEOUS ELECTRICAL NERVE STIMULATION (TENS).

PAIN SYNDROMES.

See BIOFEEDBACK.

PANCREAS, ARTIFICIAL

ROMAN HOVORKA
University of Cambridge
Cambridge, United Kingdom

INTRODUCTION

In 2000, some 171 million people worldwide had diabetes. By 2030, a conservative forecast suggests that this number will increase to 366 million attaining epidemic proportions as the prevalence increases from 2.8 to 4.4% in all age groups (1) due to, primarily, a relative increase in developing countries (2).

Diabetes is a group of heterogeneous chronic disorders characterized by hyperglycemia due to relative

or absolute insulin deficiency. Two major categories of diabetes are recognized according to aetiology and clinical presentation, type 1 diabetes and type 2 diabetes. More than 90% cases are accounted for by type 2 diabetes. Regional and ethnic differences in diabetes incidence and prevalence exist.

Type 1 diabetes is one of the most common chronic childhood disease in developed nations (3), but occurs at all ages. Type 1 diabetes is caused by autoimmune destruction of pancreatic islet beta-cells resulting in the absolute loss of insulin production. Treatment demands the administration of exogenous insulin. Type 1 diabetes is associated with a high rate of complications normally occurring at young ages placing a considerable burden on the individual and the society.

Type 2 diabetes is caused by insulin resistance and relative insulin deficiency, both of which are normally present at the diagnosis of the disease. Environmental and polygenic factors contribute to these abnormalities (4), but specific reasons for their development are not known. A considerable number of subjects with type 2 diabetes progresses to insulin dependency.

The persistent hyperglycemia in diabetes is associated with long-term complications and dysfunction of various organs, especially the eyes, kidneys, nerves, heart, and blood vessels. The Diabetes Control and Complications Trial (DCCT) (5) and the United Kingdom Prospective Diabetes Study (UKPDS) (6) demonstrated that tight glycaemic control reduces the risk of long-term complications of type 1 and type 2 diabetes reducing the cost to the healthcare system (7). There is no threshold for the relationship between blood glucose, that is, glycosylated hemoglobin (HbA_{1C}) and reduced risk. This indicates that glucose levels in subjects with type 1 or 2 diabetes should be as close as possible to those observed in healthy subjects. However, tight glucose control is associated with an increased risk of hypoglycemia (8), which acts as a limiting factor to the effective diabetes management.

In health, insulin is secreted by the pancreas in a highly controlled fashion to maintain the plasma glucose concentration within a narrow physiological range. In type 1 diabetes, insulin is delivered exogenously to mimic the basal and postprandial insulin needs. The standard therapy is based on multiple insulin injections using a combination of short- and long-acting insulin analogs supported by blood glucose self-monitoring (9). Treatment by the continuous subcutaneous insulin infusion (CSII), that is, using insulin pumps, is on the rise (10).

The present review is limited to the artificial electromechanical endocrine pancreas, which has the potential to revolutionize diabetes management. The artificial pancreas replaces the glucose sensing and insulin delivery by beta-cells and is therefore sometimes called an "artificial beta-cell". It consists of three components, a glucose monitor to measure continuously glucose concentration, a titrating algorithm to compute the amount of insulin to be delivered, and an insulin pump to deliver the insulin. Only few prototypes have been developed and tested in a controlled clinical environment. Further progress depends on advancements of continuous glucose monitoring (11).

HISTORICAL BACKGROUND

The hormone insulin was discovered by Banting and Best in 1921. The first patient was treated a year later in 1922.

The first reported continuous *ex vivo* glucose measurement in humans was by Weller *et al.* in 1960 (12). In 1964, Kadish (13) was first to use continuous real-time glucose monitoring in a subject with diabetes to close the loop. The system employed an intravenous (iv) infusion of insulin and glucose, which could be switched on or off, denoted as an "on–off system". At that time, no suitable computational means were available.

In 1974, two groups developed a true "artificial endocrine pancreas". Albisser *et al.* (14,15) in Toronto and Pfeiffer *et al.* (16) in Ulm combined continuous glucose monitors with algorithms implemented on a microcomputer to automate iv delivery of insulin and glucose. The first commercial device, the Biostator (17) (Life Science Instruments, Miles, Elkhart, IN) was put into production in 1977 thanks to the determination by Clemens. The golden age of the Biostator was between late 1970s and early 1980s. It is still used for research purposes.

The last two decades have witnessed a considerable technological progress. Between 1999 and 2005, five continuous or semicontinuous monitors have received regulatory approval in the United States or Europe and further are under development (18). Since the introduction of continuous subcutaneous insulin infusion (CSII) (19), insulin pumps have been miniaturized and their reliability improved (20). Advanced titrating algorithms have been developed.

PHYSIOLOGICAL CONSIDERATIONS

Pancreas

The pancreas has digestive and hormonal functions. It is divided into the endocrine tissues secreting hormones insulin, glucagons, and somatostatin, and the exocrine tissues secreting digestive enzymes.

The endocrine tissues consist of many small clusters of cells called islets of Langerhans. Humans have roughly 1 million islets. Three major cell types are located in islets. Alpha-cells secrete the hormone glucagon. Beta-cells produce insulin and are the most abundant of the islet cells. Delta-cells secrete the hormone somatostatin.

Endogenous Insulin Secretion

Pancreatic beta-cells secrete insulin by sensing the levels of nutrients, such as glucose and certain amino acids. The beta-cells therefore integrate the sensing and secreting functions and are efficient in maintaining glucose homeostasis.

Four different phases of insulin secretion can be identified (21). Basal insulin secretion represents insulin released in the postabsorptive state. The cephalic phase of insulin secretion is evoked by the sight, smell, and taste of food before its digestion or absorption and is mediated by pancreatic innervations. The early-phase relates to the first 0–30 min postmeal insulin secretion, whereas the

late-phase relates to the secretion between 60 and 120 min. During all phases, insulin is secreted in a pulsatile fashion with rapid pulses occurring every 8–15 min superimposed on slower, ultradian oscillations occurring every 80–120 min.

Insulin secretion is induced by other energetic substrates besides glucose, such as amino acids and drugs. Incretin hormones, such as glucagon-like peptide-1 (GLP–1) and to a lesser extent, glucose-dependent insulinotropic polypeptide (GIP), are responsible, in part, for the higher insulin secretory response after oral compared to the intravenous glucose administration.

COMPONENTS

The artificial pancreas consists of three components, a glucose monitor to measure glucose concentration, an algorithm to decide the amount of insulin to be delivered, and a device delivering insulin. This is a minimum setup. Some argue that a safe system should include a device for the delivery of glucose but all existing prototypes, with the exception of the Biostator, avoid the delivery of glucose.

The glucose monitor could be an implantable or extracorporeal device and based on a minimally or noninvasive technology (22). Generally, the implantable sensors are projected to have several months to years lifetime whereas the nonimplantable devices have, at present, lifetime of one-half of a day to several days.

Similarly, the insulin pump can be implanted or extracorporeal. The implantable pump normally delivers insulin intraperitoneally whereas the extracorporeal insulin pump delivers insulin subcutaneously.

The control algorithm can be implemented on a separate device or on the same platform as the insulin pump. The communication between the devices can be achieved using wire or wireless technologies. The latter are becoming prevalent for the transfer of data from insulin pumps onto diabetes management systems. Integrated systems exist which allow wireless transfer of data between glucose meters and insulin pumps such as the "all-in-one" CozMore Insulin Technology System (Smiths Medical MD, Inc. MN).

TYPES OF ARTIFICIAL PANCREAS

Meal Time Insulin Delivery

Artificial pancreas can handle meal delivery in different ways. In a "fully closed-loop" setting, the artificial pancreas delivers insulin without information about the time or size of the meal. Insulin is administered purely by evaluating the glucose excursions and the system works autonomously.

Alternatively, the artificial pancreas is provided with information about the time and size of the meal. The controller generates an advice, in an open-loop manner, on prandial insulin bolus. This can be termed "closed–loop with meal announcement" or "semiclosed-loop" control.

Other ways exist to handle the meal-related insulin delivery, but most systems adopt a fully closed-loop or semiclosed-loop setting.

Body Interface

Depending on body interface, three major types of artificial pancreas are recognized, (i) the subcutaneous (sc) sensing and sc delivery (sc–sc) system, (ii) the iv sensing and intraperitoneal (ip) delivery (iv–ip) system, and (iii) the iv glucose sensing and iv insulin delivery (iv–iv) system. The approaches differ in their invasiveness and associated kinetic delays (11).

Subcutaneous: Subcutaneous Body Interface

As a minimally invasive solution, the sc–sc approach has the potential to achieve a widespread application. However, it is unlikely to be compatible with a fully closed-loop system due to considerable delays disallowing effective compensation of large disturbances, such as meals.

The overall delay from the time of insulin delivery to the peak of its detectable glucose lowering effect is 100 min (11). This consists of a 50 min delay due to insulin absorption with short-acting insulin analogs (23), 30 min and more due to insulin action (24), 10 min due to interstitial glucose kinetics (25), and 10–30 min due to the transport time for *ex vivo* based monitoring system, such as those based on the microdialysis technique (26).

It is likely that users of the sc–sc approach will have to enter nutritional information to assist in the delivery of the prandial insulin dose. Most present prototypes adopt the sc–sc approach.

Intravenous: Intraperitoneal Body Interface

The iv–ip can benefit from existing intraperitoneal insulin pumps. The delays in the system are about 70 min, which comprises a 40 min time-to-peak of plasma insulin following intraperitoneal administration and a 30 min delay due to insulin action (11). Additionally, a delay due to kinetic properties of the glucose sensor applies, such as a 16 min kinetic and transport delay introduced by the long-term sensor system (27). It is unclear whether a fully closed-loop system can be developed under such circumstances.

The drawback of the iv–ip route is considerable invasiveness and relative inexperience with intraperitoneal compared to subcutaneous insulin pumps. Only > 1000 intraperitoneal pumps have been implanted so far (28) compared to $> 200{,}000$ subcutaneous pump users (29). Intraperitoneal insulin can be delivered by an implantable insulin pump Minimed 2007 (28) or via an indwelling intraperitoneal catheter such as DiaPort by Disetronic.

Intravenous: Intravenous Body Interface

The iv–iv approach was the first to have been investigated. It is embodied by the Biostator device. At present, the iv–iv approach is usable at special situations, such as in critically ill patients, surgical operations, or for research investigations. The drawback of the approach is its invasiveness requiring vascular access for both glucose monitoring and insulin delivery and is associated with a high risk of complications arising from, for example, biocompatibility issues.

The benefit of the approach is that the kinetic delays, ~ 30 min due to the delay in insulin action, are minimized enabling the development of a fully closed-loop system.

Figure 1. Biostator is the first commercial artificial endocrine pancreas. (Courtesy of Dr. Freckmann, Institute for Diabetes Technology, Ulm, Germany.)

PROTOTYPES

Biostator

Introduced in 1977, the Glucose-Controlled Insulin Infusion System (GCIIS), trademark name Biostator, is a modular, computerized, feedback control system for control of blood glucose concentrations (17), see Fig. 1. The Biostator is an example of an iv–iv system working in the fully closed-loop mode.

The Biostator was developed to normalise glucose in acute metabolic disturbances such as during diabetic ketoacidosis. However, its primary use has been in research investigating insulin sensitivity by the method of the glucose clamp and assessing insulin requirements and associated inter and intrasubject variability in subjects with type 1 diabetes and other conditions.

The rapid on-line glucose analyzer uses whole blood utilising a glucose oxidase sensor in the measurement process. The analyzer demonstrated both short- and long-range stability based on a two-point calibration.

The nonlinear proportional-derivative controller uses a five-point moving average smoothing and titrates insulin or dextrose intravenous infusion using a multichannel peristaltic infusion system to achieve user-defined glucose concentration. A printer records, on a minute-by-minute basis, the glucose value measured, the insulin and/or dextrose infusion rates, and the cumulative total of the insulin infused. A serial RS232 link allows these data to be downloaded to an external computer. The system response is < 90 s including transport of blood from the patient.

Although a pioneering device, the Biostator suffers from serious limitations. It needs constant technician's supervision. It discards continuously venous blood at a rate of 50 mL per 24 h. The control algorithm is oversimplistic. The original insulin titrating algorithm was linked to the rate of glucose change by Albisser *et al.* (14) with modifications, for example, by Botz (30), Marliss *et al.* (31), and Kraegen *et al.* (32) to reduce postprandial hyperglycemia and hyperinsulinemia. The algorithms require individualization by assigning values to constants. No formal adaptive approach was used to support the assignment, which is based on heuristics. These and similar algorithms were reviewed by Broekhuyse *et al.* (33), who concluded that none of the algorithms was superior and that further work was required to achieved normalization of the glucose concentration.

Over 200 devices have been sold worldwide. The Biostator contributed to the development and acceptance of the present gold standard in the diabetes management by multiple daily injections. At present, it is used for research purposes to evaluate diabetes drugs and technologies. The number of functioning prototypes counts in tens as spare parts run out. The Glucostator (mtb GmbH, Lonsee, Germany) is a CE-marked device recently marketed to replace the aging Biostator devices.

Shichiri's Group

Professor Shichiri and co-workers, Kumamoto, Japan, has developed as early as in 1975, a prototype of an iv–iv artificial endocrine pancreas (34) made later into a compact bedside version, STG-22 (Nikkiso Co. Ltd., Japan) (35) with a similar properties to the Biostator. The device is still marketed. STG-22 uses a glucose sensor for continuous glucose monitoring by combining the immobilised glucose oxidase membrane glucose enzyme sensor measuring hydrogen peroxide.

Following on, the group developed a prototype wearable artificial pancreas using the sc–sc route with the regular (36,37) and short acting insulin (38), and the sc–ip route (39). The latest versions use a microdialysis-type (40) or a ferrocene-mediated needle-type (39) glucose sensor working over a period of 7 days without any *in vivo* calibration (i.e., without using blood glucose measurement to calibrate the glucose sensor) followed by 14 days with one point calibration (41).

The results of the performance of their closed-loop system are even more impressive. With a fully closed sc–sc route using short acting insulin Lispro, the group claimed to have achieved "perfect" normalization of blood glucose over 24 h (38,42).

These results are surprising given that the control algorithm was a simple, nonadaptive PD controller in the form

$$\mathrm{IIR}(t) = K_\mathrm{P} G(t) + K_\mathrm{D} \frac{dG(t)}{dt} + K_\mathrm{C}$$

where IIR*(t)* is insulin infusion rate, *G(t)* is the monitored glucose concentration, and K_P, K_D, and K_C are constants, which are dependent on the type of insulin delivery, subcutaneous versus intravenous, and also on the type of insulin, regular versus short-acting insulin lispro (38,43).

These enviable groundbreaking results, however, failed to be confirmed by other groups. The achievements of the group are summarized in an edited monograph (34).

Minimed: Medtronic

The Continuous Glucose Monitoring System (CGMS; Medtronic MiniMed, Northridge CA) (44) is the first commercial continuous glucose monitor. Approved in 1999, CGMS adopts a Holter-style monitoring to store up to 3-day data for retrospective analysis.

The CGMS employs an electrochemical sensor inserted into the subcutaneous tissue adopting the hydrogen peroxide-based enzyme electrode (45), which provides signal every 10 s. Calibration is achieved using self-monitoring of blood glucose. The new "gold" sensor introduced in November 2002 is more accurate than the original sensor [the mean absolute deviation 0.83 vs 1.11 $mmol \cdot L^{-1}$ (46)].

Employing the CGMS sensor, an external physiologic insulin delivery (ePID) has been developed by Minimed–Medronic. The system uses a PID controller (47), which was designed to reproduce the first phase insulin secretion by linking insulin administration to the rate of change in glucose concentration (the proportional component of the controller) and the second phase by linking insulin administration to the difference between the ambient and target glucose (the integrative component of the controller).

First studies with a fully closed loop were executed in dogs (48). The example presented in (49) shows peak postmeal glucose of 15 $mmol \cdot L^{-1}$ with the set point reached in 11 h indicating a suboptimal performance of a fully closed loop with the sc–sc approach. The adaptation of the PID controller was achieved by assigning the proportional gain K_P a value resulting in a normal daily insulin dose of the dog at euglycemia (48).

An evaluation of the ePID system in six subjects with type 1 diabetes > 27.5 h resulted in preprandial and postprandial (2 h) glucose levels at 5.8 ± 1.2 and 9.8 ± 1.6 $mmol \cdot L^{-1}$ (mean $\pm$ SD) (50). Morning glucose after overnight control was 6.8 ± 1.0 $mmol \cdot L^{-1}$.

Roche Diagnostics

The sc–sc closed-loop prototype with meal announcement (51,52) developed by Roche adopted the subcutaneous continuous glucose monitor (SCGM1; Roche Diagnostics GmbH, Manheim, Germany), which has been designed to monitor glucose in the subcutaneous interstitial fluid for up to 4–5 days (53).

SCGM1 is based on the microdialysis technique with an *ex vivo* glucose measurement. The sensor produces a signal every second. This is reduced to one glucose measurement every 5 min. Calibration is required once every 24 h (26,53). SCGM1 has a low flow rate ($0.3\ \mu L \cdot min^{-1}$), achieves a 100% recovery of the subcutaneous glucose in the dialysate, but has a 30 min technical lag. *In vitro* performance is excellent with a mean absolute difference of 0.2–3.8% in 10 sensor units (53).

An "empirical algorithm" (51) was develop to titrate sc insulin. A set of rules, derived from clinical observations, determine the insulin bolus administered every 10 min.

The closed-loop system with meal announcement was tested in 12 well-controlled ($HbA_{1C} < 8.5\%$) subjects with type 1 diabetes (51). Control lasted over 32 h and included the digestion of breakfast, lunch, dinner, and a snack. The target glucose concentration for the algorithm was 6.7 $mmol \cdot L^{-1}$. Prandial bolus was calculated from the carbohydrate content of the meal.

The algorithm achieved a near-target monitored glucose concentration (6.9 vs. 6.2 $mmol \cdot L^{-1}$; mean, algorithm vs. self-directed therapy) and reduced the number of hypoglycemia interventions from 3.2 to 1.1 per day per subject. During the algorithm therapy, 60% of SCGM1 values were within the 5–8.3 $mmol \cdot L^{-1}$ range compared to 45% with the self-directed therapy.

Adicol Project

The project Advanced Insulin Infusion using a Control Loop (Adicol) (54) was an EC funded project that completed at the end of 2002. The Adicol's sc–sc closed loop with meal announcement consisted of a minimally invasive subcutaneous glucose system, a handheld PocketPC computer, and an insulin pump (Disetronic D-Tron) delivering subcutaneously insulin lispro, see Fig. 2.

As continuous sensor was developed in parallel with the control algorithm and was not sufficiently stable, throughout the Adicol project, the intravenous glucose measurement was used, delayed by 30 min to simulate the lag associated with sc glucose sampling.

Adicol adopted an adaptive nonlinear model predictive controller (MPC) (55), which included a model based on a two compartment representation of glucose kinetics (24)

Figure 2. Components used by the Adicol's biomechanical artificial pancreas. Top left corner shows the microperfusion probe connected to the glucose monitor, which includes microfluidics components, Bluetooth communication, and the sensor. The handheld iPAQ PocketPC maintains wireless communication with the other two components, runs the MPC controller. Disetronic D-Tron insulin pump is equipment with a special sleeve visible on the left hand side of the pump which converts the Bluetooth radiofrequency signal to an infrared signal accepted by the pump (reprinted with permission from (54)).

extended by submodels representing the absorption of short acting insulin lispro, the insulin kinetics, the renal clearance of glucose, and the gut absorption. The MPC approach was combined with an adaptive Bayesian technique to individualize the glucoregulatory model to represent the inter- and intrasubject variability. The individualization was integrated within the control algorithm and was executed at each 15 min control cycle.

The largest clinical study performed in the Adicol project assessed the efficacy of the MPC controller with 30 min delayed glucose sampling $>$ 26 h in 11 subjects with type 1 diabetes. Glucose was normalized from 1400 to 1800. Dinner followed with an individually determined prandial bolus at 1800, and control by the MPC from 1930 to 2200 the following day.

One hypoglycemia event (touchdown at 3.3 mmol·L^{-1}) due to the MPC control was recorded. The highest glucose concentration was 13.3 mmol·L^{-1} following breakfast; 84% of glucose measurements were between 3.5 and 9.5 mmol·L^{-1} (56).

Following the completion of the Adicol project, a viscometric sensor (57) was tested with the MPC algorithm. Five subjects with type 1 diabetes treated by CSII were studied for 24 h (58). No hypoglycemia event ($<$ 3.3 mmol·L^{-1}) due to the MPC control was observed. Overall, 87% sensor values were between 3.5 and 9.5 mmol·L^{-1}. Outside the 3 h postmeal periods, 74% of sensor measurements were in the range 3.5–7.5 mmol·L^{-1}.

Institute for Diabetes Technology, Ulm

Building on foundations laid by Professor Pfeiffer in the early 1970s, the work in Ulm continues (59).

The group used the amperometric–enzymatic approach in combination with the microdialysis technique. The continuous flow method uses a slow continuous flow through the tubing achieving nearly a 100% recovery with a 30 min lag (59). The comparative method does not require calibration (60). Saline with glucose (5.5 mmol·L^{-1}) is pumped through the probe in a stop-flow mode. During the stop mode, a nearly 100% equilibrium between the interstitial plasma glucose and the perfusate is achieved. In the flow mode, the dialysate is pumped rapidly to the sensor chamber. The technique facilitates sensor internal calibration for each measuring cycle and yields five glucose measurements per hour.

The group developed and tested an sc–sc closed-loop approach with meal announcement (61,62), see Fig. 3.

The algorithm uses the basal insulin need, determined from an individual insulin need, and a postprandial insulin need, expressed as an insulin/carbohydrate ratio. A model exploits these values and predicts future glucose excursions. The algorithm was tested in eight subjects with type 1 diabetes over a period of 24 h. The average glucose value was 7.8 $\pm$ 0.7 mmol·L^{-1} (mean $\pm$ SD). The postprandial increases were at 2.9 $\pm$ 1.3 mmol·L^{-1} with largest excursions recorded after breakfast. One hypoglycemia ($<$ 3.3 mmol·L^{-1}) was observed (62).

EVADIAC Group

Exploiting the progress made by the French group on implantable pumps "Evaluation dans le Diabete du Traitement par Implants Actifs" (EVADIAC), the work by Renard *et al.* is at the forefront of the fully closed-loop iv–ip approach. The group has developed the implantable physiologic insulin delivery (iPID) system, which uses a long-term sensor system (LTSS) (63,64).

Figure 3. The system V4-IDT from the ULM group. The system uses a glucose monitor based on microdialysis integrated with a portable computer and an H-Tron pump, Disetronic. (Courtesy of Dr. Freckmann, Institute for Diabetes Technology, Ulm, Germany.)

Long-term sensor system, an intravenous enzymatic oxygen-based sensor developed by Medtronic MiniMed (Northridge CA), is implanted by direct jugular access in the superior vena cava. It is connected by a subcutaneous lead to an insulin pump delivering insulin intraperitoneally and implanted in the abdominal wall, see Fig. 4. The pump implements a PD controller similar to that used by the ePID system.

The system has been investigated in subjects with type 1 diabetes with collected data per sensor of $\sim$ 280 days (65). Most investigations with LTSS have adopted the open-loop approach. The fully closed-loop system was tested $>$ 48 h reducing % time spent at $<$ 3.9 mmol·L^{-1} from 18 to 6%, and % time spent at $>$ 13.3 mmol·L^{-1} from 17 to 2%. The addition of insulin bolus at meal time, all glucose values were inside the range 3.9–13.3 mmol·L^{-1}.

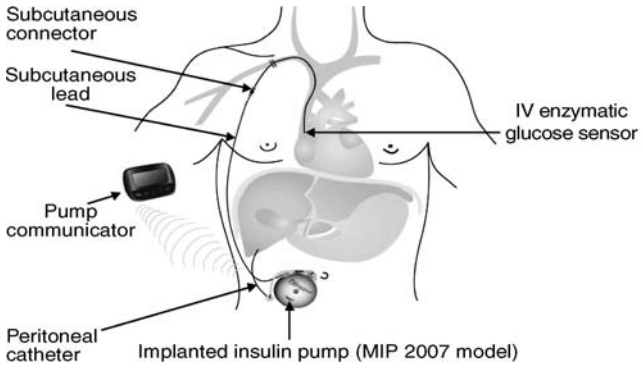


Figure 4. Scheme of human implantation of the Long-Term Sensor System (LTSS, Medtronic-MiniMed), a prototype of artificial pancreas. (Courtesy of Dr. Renard, Lapeyronie Hospital, Montpellier, France.)

Recently, the iPID system was evaluated in four elderly lean subjects with type 1 diabetes over 48 h (66). During the second 24 h control period following empirical tuning of the algorithm, 4 and 7% of time was spent < 4.4 mmol·L^{-1} in the postprandial (0–2 h) and outside meal conditions, respectively, 12 and 32% was spent in the region 4.4–6.7 mmol·L^{-1}, 63 and 60% was spent in the region 6.7–13.3 mmol·L^{-1}, and 20 and 2% was spent > 13.3 mmol·L^{-1}.

CLINICAL STUDIES

Clinical studies performed with prototypes of an artificial pancreas in subjects with type 1 diabetes are summarized in Table 1. All experiments were performed in hospital environment. No prototype has yet been studied in home settings.

Table 1 excludes numerous experiments carried out with the Biostator as the invasiveness and the setup adopted by the device does not permit development into a routinely used system.

Numerous experiments have been carried out in pancreatectomized dogs especially in the early phases of prototype development. This includes the ePID system (48), but some approaches such as that adopted by the Adicol project, did not use testing on animal models, but adopted testing on a simulation environment (82).

INDICATIONS

Artificial pancreas has the potential to be used in various disease conditions. For subjects with type 1 diabetes, the system offers "cure" especially if implemented as a fully closed-loop iv–ip system. Realistically, first prototypes for home setting are expected to adopt the sc-sc route with meal announcement and thus participation of the subjects with type 1 diabetes in the disease management process is required. Fail-safe procedures are needed for such a setup.

An increasing proportion of subjects with type 2 diabetes is treated by insulin. It has been reported that nearly 50% of subjects with type 2 diabetes require insulin treatment at some stage of their disease. The artificial pancreas could provide a solution for a subgroup of subjects with type 2 diabetes, but this has to be justified by a cost-benefit analysis.

The sc–sc route may require a continuing close subject involvement in the disease management. This would restrict the treatment group to those well motivated, who generally have good glucose control. The greatest treatment benefit would be for those with poor glucose control, but other aspects, such as psychological factors might impair or even prevent the system deployment.

A recent study in adult critically ill subjects revealed that glucose control < 6.1 mmol·L^{-1} is associated with reduced mortality by 43%, overall inhospital mortality by 34%, newly developed kidney failure requiring dialysis by 41%, bacteremia by 46%, the number of red blood cell transfusions by 50%, and critical illness polyneuropathy by 44% (83). Although still awaiting confirmation by another prospective study, the results indicate that artificial pancreas for critically ill is likely to bring about major improvements in therapeutic outcomes. Whether these results apply to a broader category of inpatients be it a general ward or pediatric population is yet to be determined. It is also unclear what is the most appropriate setup of a "hospital-based" artificial pancreas.

COST

The cost of the artificial pancreas can only be inferred from the cost of existing technology. The cost of the Biostator was $\sim$ \$70,000 prohibiting its wider use.

Insulin pumps cost from \$5,000 to 6,000. The monthly cost of pump treatment is $\sim$ \$100. The CGMS monitor costs $\sim$ \$4,000 and the single-use 3 day sensors cost \$50 each. It is likely that the sc-sc artificial pancreas will be at least as expensive as the combined cost of the insulin pump and the glucose monitor.

The cost needs to be set against the total cost of diabetes which, in developed countries such as in the United States or the United Kingdom, is 10% of the total health care budget (84). Most of the direct expenditure is on treating diabetes complications, which could be delayed or prevented with tight glucose control.

OUTLOOK

Present technology has made considerable advances toward a truly personal wearable treatment system. The lack of availability of a glucose monitor with adequate properties appears to hinder further progress and the development of a commercially viable system. The algorithms need to be improved and subjected to rigorous clinical testing.

With regard to the existing glucose monitors, it is possible that their potential has not been fully exploited. The regulatory-driven research and development to achieve an equivalence between finger-prick glucose measurements and values provided by continuous glucose monitors in the interstitial fluid hinders the engagement of the industry and the academia in the development and testing of closed-loop systems.

In the first instance, the artificial pancreas is most likely to find its wider use in the supervised hospital environment, such as at the intensive care units. The application in home settings will most likely be gradual starting with a supervised system (by the treated subjects) and increasing its autonomous function following on from wider experience.

The regulatory bodies will play an important role in the introduction of the artificial pancreas into the clinical practice. Until recently, the perception of closed-loop systems was not overly positive by the regulatory authorities. Artificial pancreas with its potential to "cure" diabetes, but also to lead to life-threatening complications, if malfunctioned, will have to pave the way to a new generation of closed-loop home-based biomedical devices if it is ever to succeed.

Table 1. Clinical Studies with sc–sc or iv–ip Closed-Loop Control in Subjects with Type 1 Diabetes[a]

N	Duration of Control, h	Sensing–Infusion	Sensor	Insulin	Algorithm	Control Interval, min	Type of Control	Performance	References
6	72	sc–sc	needle-type Ref. 37	regular	PD[b]	1[c]	f–cl[d]	M-value (67) 15 ± 4; mean glucose 6.1 ± 0.5 mmol·L^{-1}; Mage (68) 73 ± 14 mg·dL^{-1}	37
5	5	sc–sc	needle-type Ref. 69	regular	PD[b]	1[c]	f–cl	postprandial glucose (1.5 h) 10.6 ± 0.9 mmol·L^{-1}; late postprandial glucose (5 h) 2.8 ± 0.4 mmol·L^{-1};	38,70
5	5	sc–sc	needle-type Ref. 71	lispro	PD[b]	1[c]	f–cl	no hypoglycemia (< 2.8 mmol·L^{-1}); postprandial glucose (1 h) 8.5 ± 0.5 mmol·L^{-1};	38,72
5	24	sc–sc	needle-type Ref. 73	regular	PD[b]	1[c]	f–cl	hypoglycemia observed; glucose between peak (b'fast postprandial) 12.5 ± 1.0 mmol·L^{-1} and nadir (before dinner) 2.7 ± 0.3 mmol·L^{-1}	38,74
5	24	sc–sc	needle-type Ref. 75	lispro	PD[b]	1[c]	f–cl	no hypoglycemia; Near normal control	38,76
9	8	iv–sc	offline[d]	lispro	MPC	15	cl–ma[e]	[f]no hypoglycemia (< 3.3 mmol·L^{-1}); [h]6.1 ±0.6 mmol·L^{-1}	77
6	8	simulated[g] sc–sc	offline[d]	lispro	MPC	15	cl–ma	[f]no hypoglycemia (< 3.3 mmol·L^{-1}); [h]6.6 ± 0.8 mmol·L^{-1}	78,79
6	14	simulated[g] sc–sc	offline[d]	lispro	MPC	15	cl–ma	[f]no hypoglycemia (< 3.0 mmol·L^{-1}); preprandial 7.0 ± 1.1 mmol·L^{-1}; [i]6.3 ± 1.6 mmol·L^{-1}	80
11	26.5	simulated[g] sc–sc	offline[d]	lispro	MPC	15	cl–ma	1 hypoglycemia (< 3.3 mmol·L^{-1}); 84% glucose values between 3.5 and 9.5 mmol·L^{-1}	56
5	24	sc–sc	viscometric Ref. 57	lispro	MPC	15	cl–ma	no hypoglycemia (< 3.3. mmol·L^{-1}); 87% sensor values between 3.5 and 9.5 mmol·L^{-1}	58
6	27.5	sc–sc	CGMS/Guardi an Refs. 44,81	lispro	PID	1–5[c]	f–cl	hypoglycemia not reported; preprandial glucose 5.8 ± 1.2 mmol·L^{-1}; postprandial glucose (2 h) 9.8 ± 1.6 mmol·L^{-1}	50
4	48	iv–ip	LTSS Refs. 63,64	U400	PID	1–5[c]	f–cl	hypoglycemia not reported; 84% glucose values between 4.4 and 13.3 mmol·L^{-1}	66
12	32	sc–sc	SCGM1 Ref. 53	lispro	empirical	10	cl–ma	1.1 hypoglycemia per day per subject; 56% glucose values between 5.0 and 8.3 mmol·L^{-1}	51
12	7	sc–sc	comparative microdialysis Ref. 60	lispro	MPC	12	cl–ma	mean glucose 8.0 ± 2.3 mmol·L^{-1}; postprandial glucose increase 3.0 ± 1.6 mmol·L^{-1}	61
8	24	sc–sc	comparative microdialysis Ref. 60	lispro	MPC	12 (day) 36 (night)	cl–ma	one hypoglycemia (< 3.3 mmol·L^{-1}) mean glucose 7.8 ± 0.7 mmol·L^{-1}; postprandial glucose increase 2.9 ± 1.3 mmol·L^{-1}	59,62

[a]Adapted from (11).
[b]Proportional-derivative controller.
[c]Not reported, an estimate from plot(s).
[d]Beckman Glucose Analyzer 2.
[e]f–cl: fully closed–loop; cl–ma: closed-loop with meal announcement.
[f]Number of hypoglycemias due to the controller.
[g]IV glucose measurements delayed by 30 min.
[h]Mean ± SD in the last two hours of control.
[i]Over 4 hours following meal.

ACKNOWLEDGMENT

Support by Disetronic Medical Systems AG, Burgdorf, Switzerland, and the EU Clincip Project (IST-2002-506965) is acknowledged. Dr Malgorzata E Wilinska helped with the preparation of the background material.

BIBLIOGRAPHY

1. Wild S, et al. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care 2004; 27:1047–1053.
2. Diabetes Atlas, International Diabetes Federation, 2003.
3. LaPorte R, Matsushima M, Chang Y. Prevalence and incidence of insulin-dependent diabetes. In NDDG, edition. Diabetes in America, NIH; 1995. p 37–46.
4. Bell GI, Polonsky KS. Diabetes mellitus and genetically programmed defects in beta-cell function. Nature (London) 2001; 414:788–791.
5. Diabetic Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long term complications in insulin-dependent diabetes mellitus. N Engl J Med 1993;329:977–986.
6. Turner RC, et al. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). Lancet 1998;352:837–853.
7. Gilmer TP, O'Connor PJ, Manning WG, Rush WA. The cost to health plans of poor glycemic control. Diabetes Care 1997;20: 1847–1853.
8. Cryer PE, Davis SN, Shamoon H. Hypoglycemia in diabetes. Diabetes Care 2003;26:1902–1912.
9. Standards of Medical Care in Diabetes. Diabetes Care 2005; 28:S4–S36.
10. Pickup J, Keen H. Continuous subcutaneous insulin infusion at 25 years: evidence base for the expanding use of insulin pump therapy in type 1 diabetes. Diabetes Care 2002;25:593–598.
11. Hovorka R. Continuous glucose monitoring and closed-loop systems. Diabetic Med 2005; (in press).
12. Weller C, et al. Continuous *In vivo* determination of blood glucose in human subjects. Ann N Y Acad Sci 1960;87:658–668.
13. Kadish AH. Automation control of blood sugar. I. A. servomechanism for glucose monitoring and control. Am J Med Electron 1964;39:82–86.
14. Albisser AM, et al. An artificial endocrine pancreas. Diabetes 1974;23:389–404.
15. Albisser AM, et al. Clinical control of diabetes by the artificial pancreas. Diabetes 1974;23:397–404.
16. Pfeiffer EF, Thum C, Clemens AH. The artificial beta cell - A continuous control of blood sugar by external regulation of insulin infusion (glucose controlled insulin infusion system). Horm Metab Res 1974;6:339–342.
17. Clemens AH, Chang PH, Myers RW. The development of Biostator, a Glucose Controlled Insulin Infusion System (GCIIS). Horm Metab Res 1977; (Suppl. 7):23–33.
18. Klonoff DC. Continuous glucose monitoring:roadmap for 21st century diabetes therapy. Diabetes Care 2005;28: 1231–1239.
19. Pickup JC, Keen H, Parsons JA, Alberti KG. Continuous subcutaneous insulin infusion: an approach to achieving normoglycaemia. Br Med J 1978;1:204–207.
20. Pickup J, Keen H. Continuous subcutaneous insulin infusion at 25 years: evidence base for the expanding use of insulin pump therapy in type 1 diabetes. Diabetes Care 2002;25:593–598.
21. Caumo A, Luzi L. First-phase insulin secretion: does it exist in real life? Considerations on shape and function. Am J Physiol Endocrinol Metab 2004;287:E371–E385.
22. Klonoff DC. Continuous glucose monitoring: roadmap for 21st century diabetes therapy. Diabetes Care 2005;28: 1231–1239.
23. Plank J, et al. A direct comparison of insulin aspart and insulin lispro in patients with type 1 diabetes. Diabetes Care 2002;25:2053–2057.
24. Hovorka R, et al. Partitioning glucose distribution/transport, disposal, and endogenous production during IVGTT. Am J Physiol 2002;282:E992–E1007.
25. Rebrin K, Steil GM, Van Antwerp WP, Mastrototaro JJ. Subcutaneous glucose predicts plasma glucose independent of insulin: implications for continuous monitoring. Am J Physiol 1999;277:E561–E571.
26. Heinemann L. Continuous glucose monitoring by means of the microdialysis technique: underlying fundamental aspects. Diabetes Technol Ther 2003;5:545–561.
27. Steil GM, Panteleon AE, Rebrin K. Closed-loop insulin delivery-the path to physiological glucose control. Adv Drug Deliv Rev 2004;56:125–144.
28. Selam JL. External and implantable insulin pumps: current place in the treatment of diabetes. Exp Clin Endocr Diab 2001;109:S333–S340.
29. Pickup J, Keen H. Continuous subcutaneous insulin infusion at 25 years: evidence base for the expanding use of insulin pump therapy in type 1 diabetes. Diabetes Care 2002;25:593–598.
30. Botz CK. An improved control algorithm for an artificial beta-cell. IEEE Trans Biomed Eng 1974;23:252–255.
31. Marliss EB, et al. Normalization of glycemia in diabetics during meals with insulin and glucagon delivery by the artificial pancreas. Diabetes 1977;26:663–672.
32. Kraegen EW, et al. Control of blood glucose in diabetics using an artificial pancreas. Aust N Z J Med 1977;7:280–286.
33. Broekhuyse HM, Nelson JD, Zinman B, Albisser AM. Comparison of algorithms for the closed-loop control of blood glucose using the artificial beta cell. IEEE Trans Biomed Eng 1981;28:678–687.
34. Shichiri M. Artificial Endocrine Pancreas: Development and Clinical Applications. Kumamoto: Kamome Press; 2000.
35. Goriya Y, Kawamori R, Shichiri M, Abe H. The development of an artificial beta cell system and its validation in depancreatized dogs: the physiological restoration of blood glucose homeostasis. Med Prog Technol 1979;6:99–108.
36. Shichiri M, et al. Wearable artificial endocrine pancrease with needle-type glucose sensor. Lancet 1982;2:1129–1131.
37. Kawamori R, Shichiri M. Wearable artificial endocrine pancreas with needle-type glucose sensor. In Nose Y, Kjellstrand C, Ivanovich P. editors. Progress in Artificial Organs, Cleveland: ISAO Press; 1986. pp. 647–652.
38. Shimoda S, et al. Closed-loop subcutaneous insulin infusion algorithm with a short-acting insulin analog for long-term clinical application of a wearable artificial endocrine pancreas. Front Med Biol Eng 1997;8:197–211.
39. Shimoda S, et al. Development of closed-loop intraperitoneal insulin infusion algorithm for a implantable artificial endocrine pancreas. Diabetes 2001;50(Suppl. 2):A3.
40. Hashiguchi Y, et al. Development of a miniaturized glucose monitoring system by combining a needle-type glucose sensor with microdialysis sampling method. Long-term subcutaneous tissue glucose monitoring in ambulatory diabetic patients. Diabetes Care 1994;17:387–396.

41. Shichiri M, Sakakida M, Nishida K, Shimoda S. Enhanced, simplified glucose sensors: long-term clinical application of wearable artificial endocrine pancreas. Artif Organs 1998;22:32–42.
42. Shichiri M, Sakakida M, Nishida K, Shimoda S. Enhanced, simplified glucose sensors: long-term clinical application of wearable artificial endocrine pancreas. Artif Organs 1998; 22:32–42.
43. Shichiri M, Sakakida M, Nishida K, Shimoda S. Enhanced, simplified glucose sensors: long-term clinical application of wearable artificial endocrine pancreas. Artif Organs 1998; 22:32–42.
44. Mastrototaro J. The MiniMed continuous glucose monitoring system (CGMS). J Pediatr Endocr Met 1999;12:751–758.
45. Johnson KW, et al. *In vivo* evaluation of an electroenzymatic glucose sensor implanted in subcutaneous tissue. Biosens Bioelectron 1992;7:709–714.
46. Accuracy of the GlucoWatch G2 Biographer and the continuous glucose monitoring system during hypoglycemia: experience of the Diabetes Research in Children Network. Diabetes Care 2004;27:722–726.
47. Steil GM, et al. Modeling beta-cell insulin secretion–implications for closed-loop glucose homeostasis. Diabetes Technol Ther 2003;5:953–964.
48. Steil GM, et al. Tuning closed-loop insulin delivery based on known daily insulin requirements. Diabetes 2002;51 (Suppl. 2):510.
49. Steil GM, Panteleon AE, Rebrin K. Closed-loop insulin delivery-the path to physiological glucose control. Adv Drug Deliv Rev 2004;56:125–144.
50. Steil GM, et al. Continuous automated insulin delivery based on subcutaneous glucose sensing and an external insulin pump. Diabetes 2004;53(Suppl. 2):A2.
51. Galley P, et al. Use of subcutaneous glucose measurements to drive real-time algorithm-directed insulin infusion recommendations. Diabetes Technol Ther 2004;6:245–246.
52. Galley PJ, Thukral A, Chittajallu SK, Weinert S. Diabetes management system. US Pat. 2003. 6,544,212:1–17.
53. Schoemaker M, et al. The SCGM1 System:subcutaneous continuous glucose monitoring based on microdialysis technique. Diabetes Technol Ther 2003;5:599–608.
54. Hovorka R, et al. Closing the loop: The Adicol experience. Diabetes Technol Ther 2004;6:307–318.
55. Hovorka R, et al. Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. Physiol Meas 2004;25:905–920.
56. Schaller HC, et al. Avoidance of hypo- and hyperglycaemia with a control loop system in patients with Type 1 DM under daily life conditions. Diabetes Metab 2003;29:A2225.
57. Beyer U, et al. Recording of subcutaneous glucose dynamics by a viscometric affinity sensor. Diabetologia 2001;44:416–423.
58. Vering T. Minimally invasive control loop system for sc-sc control on patients with type 1 diabetes. Diabetes Technol Ther 2004;6:278.
59. Freckmann G, et al. Recent advances in continuous glucose monitoring. Exp Clin Endocr Diab 2001;109:S347–S357.
60. Hoss U, et al. A novel method for continuous online glucose monitoring in humans: the comparative microdialysis technique. Diabetes Technol Ther 2001;3:237–243.
61. Kalatz B, et al. Development of algorithms for feedback-controlled subcutaneous insulin infusion with insulin lispro. Acta Diabetol 1999;36:215.
62. Kalatz B. Algorithmen zur Glucosegesteuerten Insulininfusion bei Diabetes Mellitus-Entwicklung und Experimentelle Untersuchung, Ulm: Medical Dissertation. 1999.
63. Renard E, Costalat G, Bringer J. From external to implantable insulin pump, can we close the loop? Diabetes Metab 2002;28:S19–S25.
64. Renard E. Implantable closed-loop glucose-sensing and insulin delivery: the future for insulin pump therapy. Curr Opin Pharmacol 2002;2:708–716.
65. Renard E, et al. Sustained safety and accuracy of central IV glucose sensors connected to implanted insulin pumps and short-term closed-loop trials in diabetic patients. Diabetes 2003;52(Suppl. 2):A36.
66. Renard E, et al. Efficacy of closed loop control of blood glucose based on an implantable iv sensor and intraperitoneal pump. Diabetes 2004;53(Suppl. 2):A114.
67. Schlichtkrull J, Munck O, Jersild M. The M-value, an index of blood sugar control in diabetics. Acta Med Scand 1965;177:95–102.
68. Service FJ, et al. Mean amplitude of glycemic excursions, a measure of diabetic instability. Diabetes 1970;19:644–655.
69. Hashiguchi Y, et al. Development of a miniaturized glucose monitoring system by combining a needle-type glucose sensor with microdialysis sampling method. Long-term subcutaneous tissue glucose monitoring in ambulatory diabetic patients. Diabetes Care 1994;17:387–396.
70. Shichiri M, Sakakida M, Nishida K, Shimoda S. Enhanced, simplified glucose sensors: long-term clinical application of wearable artificial endocrine pancreas. Artif Organs 1998;22:32–42.
71. Hashiguchi Y, et al. Development of a miniaturized glucose monitoring system by combining a needle-type glucose sensor with microdialysis sampling method. Long-term subcutaneous tissue glucose monitoring in ambulatory diabetic patients. Diabetes Care 1994;17:387–396.
72. Shichiri M, Sakakida M, Nishida K, Shimoda S. Enhanced, simplified glucose sensors: long-term clinical application of wearable artificial endocrine pancreas. Artif Organs 1998; 22:32–42.
73. Hashiguchi Y, et al. Development of a miniaturized glucose monitoring system by combining a needle-type glucose sensor with microdialysis sampling method. Long-term subcutaneous tissue glucose monitoring in ambulatory diabetic patients. Diabetes Care 1994;17:387–396.
74. Shichiri M, Sakakida M, Nishida K, Shimoda S. Enhanced, simplified glucose sensors: long-term clinical application of wearable artificial endocrine pancreas. Artif Organs 1998; 22:32–42.
75. Hashiguchi Y, et al. Development of a miniaturized glucose monitoring system by combining a needle-type glucose sensor with microdialysis sampling method. Long-term subcutaneous tissue glucose monitoring in ambulatory diabetic patients. Diabetes Care 1994;17:387–396.
76. Shichiri M, Sakakida M, Nishida K, Shimoda S. Enhanced, simplified glucose sensors: long-term clinical application of wearable artificial endocrine pancreas. Artif Organs 1998;22: 32–42.
77. Schaller HC, et al. MPC algorithm controls blood glucose in patients with type 1 diabetes mellitus under fasting conditions using the IV-SC route. Diabetes Technol Ther 2002;4:234.
78. Schaller HC, et al. Feasibility of the SC-SC route for an extracorporeal artificial pancreas. Diabetes 2002;51(Suppl. 2): 462.
79. Schaller HC, Schaupp L, Bodenlenz M, Wilinska ME, Chassin LJ, Wach P, Vering T, Hovorka R, Pieber TR. On-line adaptive algorithm with glucose prediction capacity for subcutaneous closed loop control of glucose: Evaluation under fasting conditions in patients with type 1 diabetes. Diabetic Med 2005; (in press).

80. Canonico V, et al. Evaluation of a feedback model based on simulated interstitial glucose for continuous insulin infusion. Diabetologia 2002;45(Suppl. 2):995.
81. Bode B, et al. Alarms based on real-time sensor glucose values alert patients to hypo- and hyperglycemia: the guardian continuous monitoring system. Diabetes Technol Ther 2004; 6:105–113.
82. Chassin LJ, Wilinska ME, Hovorka R. Evaluation of glucose controllers in virtual environment: Methodology and sample application. Artif Intell Med 2004;32:171–181.
83. Van den Berghe G, et al. Intensive insulin therapy in the surgical intensive care unit. N Engl J Med 2001;345:1359–1367.
84. Hogan P, Dall T, Nikolov P. Economic costs of diabetes in the US in 2002. Diabetes Care 2003;26:917–932.

See also GLUCOSE SENSOR; HEART, ARTIFICIAL.

PARENTERAL NUTRITION. See NUTRITION, PARENTERAL.

PCR. See POLYMERASE CHAIN REACTION.

PERCUTANEOUS TRANSLUMINAL CORONARY ANGIOPLASTY. See CORONARY ANGIOPLASTY AND GUIDEWIRE DIAGNOSTICS.

PERINATAL MONITORING. See FETAL MONITORING.

PERIPHERAL VASCULAR NONINVASIVE MEASUREMENTS

CHRISTOPH H. SCHMITZ
HARRY L. GRABER
RANDALL L. BARBOUR
State University of New York
Brooklyn, New York

INTRODUCTION

The primary task of the peripheral vasculature (PV) is to supply the organs and extremities with blood, which delivers oxygen and nutrients, and to remove metabolic waste products. In addition, peripheral perfusion provides the basis of local immune response, such as wound healing and inflammation, and furthermore plays an important role in the regulation of body temperature. To adequately serve its many purposes, blood flow in the PV needs to be under constant tight regulation, both on a systemic level through nervous and hormonal control, as well as by local factors, such as metabolic tissue demand and hydrodynamic parameters. As a matter of fact, the body does not retain sufficient blood volume to fill the entire vascular space, and only $\sim$ 25% of the capillary bed is in use during resting state. The importance of microvascular control is clearly illustrated by the disastrous effects of uncontrolled blood pooling in the extremities, such as occurring during certain types of shock.

Peripheral vascular disease (PVD) is the general name for a host of pathologic conditions of disturbed PV function. Peripheral vascular disease includes occlusive diseases of the arteries and the veins. An example is peripheral arterial occlusive disease (PAOD), which is the result of a buildup of plaque on the inside of the arterial walls, inhibiting proper blood supply to the organs. Symptoms include pain and cramping in extremities, as well as fatigue; ultimately, PAOD threatens limb vitality. The PAOD is often indicative of atherosclerosis of the heart and brain, and is therefore associated with an increased risk of myocardial infarction or cerebrovascular accident (stroke).

Venous occlusive disease is the forming of blood clots in the veins, usually in the legs. Clots pose a risk of breaking free and traveling toward the lungs, where they can cause pulmonary embolism. In the legs, thromboses interfere with the functioning of the venous valves, causing blood pooling in the leg (postthrombotic syndrome) that leads to swelling and pain.

Other causes of disturbances in peripheral perfusion include pathologies of the autoregulation of the microvasculature, such as in Reynaud's disease or as a result of diabetes.

To monitor vascular function, and to diagnose and monitor PVD, it is important to be able to measure and evaluate basic vascular parameters, such as arterial and venous blood flow, arterial blood pressure, and vascular compliance.

Many peripheral vascular parameters can be assessed with invasive or minimally invasive procedures. Examples are the use of arterial catheters for blood pressure monitoring and the use of contrast agents in vascular X ray imaging for the detection of blood clots. Although they are sensitive and accurate, invasive methods tend to be more cumbersome to use, and they generally bear a greater risk of adverse effects compared to noninvasive techniques. These factors, in combination with their usually higher cost, limit the use of invasive techniques as screening tools. Another drawback is their restricted use in clinical research because of ethical considerations. Although many of the drawbacks of invasive techniques are overcome by noninvasive methods, the latter typically are more challenging because they are indirect measures, that is, they rely on external measurements to deduce internal physiologic parameters. Noninvasive techniques often make use of physical and physiologic models, and one has to be mindful of imperfections in the measurements and the models, and their impact on the accuracy of results. Noninvasive methods therefore require careful validation and comparison to accepted, direct measures, which is the reason why these methods typically undergo long development cycles.

Even though the genesis of many noninvasive techniques reaches back as far as the late nineteenth century, it was the technological advances of the second half of the twentieth century in such fields as micromechanics, microelectronics, and computing technology that led to the development of practical implementations. The field of noninvasive vascular measurements has undergone a developmental explosion over the last two decades, and it is still very much a field of ongoing research and development.

This article describes the most important and most frequently used methods for noninvasive assessment of

the PV; with the exception of ultrasound techniques, these are not imaging-based modalities. The first part of this article, gives a background and introduction for each of these measuring techniques, followed by a technical description of the underlying measuring principles and technical implementation. Each section closes with examples of clinical applications and commercially available systems. The second part of the article briefly discusses applications of modern imaging methods in cardiovascular evaluation. Even though some of these methods are not strictly noninvasive because they require use of radioactive markers or contrast agents, the description is meant to provide the reader with a perspective of methods that are currently available or under development.

NONIMAGING METHODS

Arterial Blood Pressure Measurement

Arterial blood pressure (BP) is one of the most important cardiovascular parameters. Long-term monitoring of BP is used for the detection and management of chronic hypertension, which is a known major risk factor for heart disease. In this case, it is appropriate to obtain the instantaneous BP at certain intervals, such as days, weeks, or months, because of the slow progression of the disease.

In an acute care situation, such as during surgery or in intensive care, continuous BP measurements are desired to monitor heart function of the patients. The following sections describe the most important techniques.

Instantaneous BP Measurements

The most widely used approach is the auscultatory method, or method of Korotkoff, a Russian military physician, who developed the measurement in 1905. A pressure cuff is inflated to $\sim$ 30 mmHg (3.99 k Pa) above systolic pressure on the upper extremity. While subsequently deflating the cuff at a rate of $\sim$ 2 (0.26)–3 mmHg (0.39 kPa) (1), the operator uses a stethoscope to listen to arterial sounds that indicate the points at which cuff pressure equals the systolic and diastolic pressure. The first is indicated by appearance of a "tapping" sound, while the latter is identified by the change from a muffled to vanishing sound.

A second widespread BP measurement technique is the oscillatory method. Here, the cuff contains a pressure sensor that is capable of measuring cuff pressure oscillations induced by the arterial pulse. The cuff is first inflated to achieve arterial occlusion, and then deflated at rate similar to that for the auscultatory method. During deflation, the sensor registers the onset of oscillations followed by a steady amplitude increase, which reaches maximum when the cuff pressure equals the mean ABP. Beyond that, oscillations subside and eventually vanish. Systolic and diastolic pressure are given by the cuff pressure values at which the oscillatory signal amplitude is 55 and 85% of the maximum amplitude, respectively. These objective criteria, based on population studies, make this method superior to the auscultatory method, which relies on the subjective judgment of changes in sounds. Oscillatory measurements are typically used in automated BP monitors.

Continuous BP Monitoring

Currently, the standard of care for obtaining continuous central blood pressure is the insertion of a Swan–Ganz catheter into the pulmonary artery. The device has to be placed by a trained surgeon, and its use is restricted to the intensive care unit. In addition, besides bearing the risk of serious complications, the procedure is costly. There is clearly a need for noninvasive continuous blood pressure monitoring methods, which could be more widely applied, and which would reduce the patient risk. In the following, we describe two such techniques, the vascular unloading method of Peňáz, and arterial tonometry, both of which have been developed into commercial products.

Vascular Unloading. Many noninvasive BP measurements rely on vascular unloading (i.e., the application of distributed external pressure to the exterior of a limb to counter the internal pressure of the blood vessels). Typically, this is achieved with an inflatable pressure cuff under manual or automated control. Because tissue can be assumed essentially incompressible, the applied pressure is transmitted onto the underlying vessels, where it results in altered transmural (i.e., external minus internal) pressure. If the external pressure P_{ext} exceeds the internal pressure P_{int}, the vessel collapses. For the case $P_{\text{ext}} = P_{\text{int}}$ the vessel is said to be unloaded (1).

In 1973, Czech physiologist Jan Peňáz proposed a noninvasive continuous BP monitoring method based on the vascular unloading principle (2). The approach, which was first realized by Wesseling in 1985, employs a servo-controlled finger pressure cuff with integrated photoplethysmography (see below) to measure digital arterial volume changes (3). The device uses a feedback mechanism to counter volume changes in the digital arteries through constant adjustment of cuff pressure, hence establishing a pressure balance that keeps the arteries in a permanently unloaded state. The applied cuff pressure serves as a measure of the internal arterial pressure. The cuff pressure is controlled with a bandwidth of at least 40 Hz to allow adequate reaction to the pulse wave (4). The method was commercialized in the late 1980s under the name Finapres. The instrument has a portable front end, which is worn on the wrist and contains an electropneumatic pressure valve, the cuff, and the PPG sensor. This part connects to a desktop unit containing the control, air pressure system, and data display–output. Two successor products are now available, one of which is a completely portable system.

One problem of this method is that the digital BP can significantly differ from brachial artery pressure (BAP) in shape, because of distortions due to pulse wave reflections, as well as in amplitude because of flow resistance in the small arteries. The former effect is corrected by introducing a digital filter that equalizes pressure wave distortions. The second problem is addressed by introducing a correction factor and calibrating the pressure with an independent return-to-flow BP measurement. It has been demonstrated that the achievable BP accuracy lies well within the American Association for Medical Instrumentation (AAMI) standards (5).

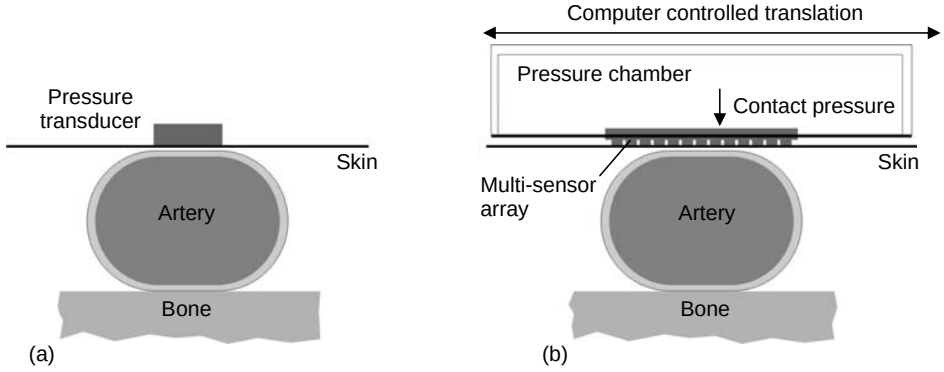


Figure 1. Applanation tonometry principle. (a) Single-element transducer. (b) Sensor array with pneumatic contact pressure control.

Applanation Tonometry. First conceived and implemented by Pressman and Newgard in the early 1960s, applanation tonometry (AT) measures the pulse pressure wave of a superficial artery with an externally applied transducer (6). The method requires the artery to be supported by an underlying rigid (i.e., bone) structure. Therefore, the method has been applied mainly to the temporal and the radial arteries; the last one being by far the most frequently employed measurement site. Figure 1a shows the general principle of the method. A pressure transducer is placed over the artery, and appropriate pressure is applied so as to partially flatten, or applanate, the artery. This ensures that the vessel wall does not exert any elastic forces perpendicular to the sensor face; therefore the sensor receives only internal arterial pressure changes caused by the arterial pulse. To obtain an accurate measurement it is crucial that the transducer is precisely centered over the artery, and that it is has stable support with respect to the surrounding tissue.

The original design used a single transducer that consisted of a rod of 2.5 mm^2 cross-sectional area, which was pressed against the artery, and which transmitted arterial pressure to a strain gauge above it. This early design suffered from practical difficulties in establishing and maintaining adequate sensor position. In addition, Drzewiecki has shown that for accurate pressure readings, the transducer area needs to be small compared to artery diameter (ideally, < 1 mm wide), a requirement that early designs did not meet (1).

The development of miniaturized pressure sensor arrays in the late 1970s has alleviated these difficulties, leading to the development of commercial AT instruments by Colin Medical Instruments Corp., San Antonio, TX. These sensor arrays use piezoresistive elements, which essentially are membranes of doped silicon (Si) that show a change in electrical resistance when subjected to mechanical stress. Piezoresistivity is a quantum mechanical effect rooted in the dependence of charge carrier motility on mechanical changes in the crystal structure. Pressure-induced resistance changes in a monocrytalline semiconductor are substantially greater than in other available strain gauges. This sensitivity together with the possibility of using semiconductor fabrication techniques to create miniaturized structures makes piezoresistive elements ideal candidates for AT applications. The change in resistance is measured with a Wheatstone bridge (e.g., pictured in Fig. 2), which, together with suitable amplification, can be integrated on the same chip as the sensing element. While piezoresistance shows linear change with strain, the devices are strongly influenced by ambient temperature. Therefore, appropriate compensation measures have to be taken.

Figure 1b shows the schematic of a modern AT pressure sensor. Thirty-one piezoeresistive elements form the 4.1×10 mm large sensor array, which is created from a monolithic Si substrate. After placing the sensor roughly over the artery, the signal from each element is measured to determine whether the transducer is appropriately centered with respect to the vessel. If necessary, its lateral position is automatically adjusted with a micromotor drive. The sensor contact pressure is pneumatically adjusted to achieve appropriate artery applanation. To provide probe stability suitable for long-term studies, the device is strapped to the wrist together with a brace that immobilizes the hand in a slightly overextended position so as to achieve better artery exposure to the sensor field.

Because AT can accurately measure the pulse wave shape, not its absolute amplitude, the AT signal is calibrated with a separate oscillatory BP measurement on the ipsilateral arm. Calibration is performed automatically at predetermined intervals.

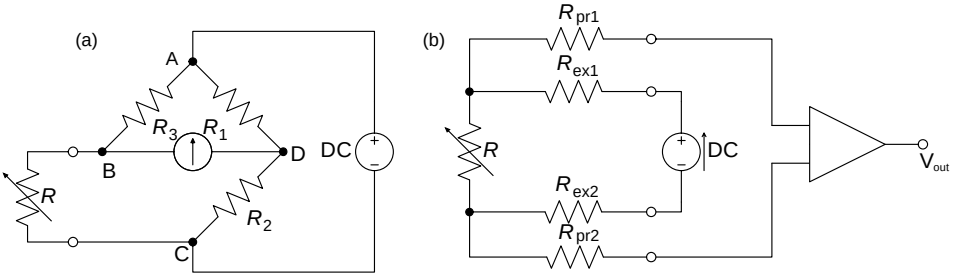


Figure 2. (a) Wheatstone bridge for sensitive detection of resistor changes; gauge lead resistance disturbs measurement. (b) Four-wire strain gauge measurement; influence of lead resistance is eliminated.

In addition to fully automated long-term monitoring devices, simpler single-element transducers are offered for clinical and research applications (PulsePen by DiaTecne, Milan, Italy and SPT-301 by Millar Instruments, Inc., Houston, TX).

Plethysmography

Plethysmography (derived from the Greek words for "inflating" and "recording") is the general name of an investigative method that measures volume change of the body, or a part of it, in response to physiologic stimulation or activity. Specifically, it allows measuring dynamics of blood flow to/from the extremities for the diagnosis of peripheral vascular diseases.

Different approaches have been conceived to measure volume change of a limb, the earliest of which, reaching back to the end of the nineteenth century, were based on measuring the volume displacement in a water filled vessel in which the limb was sealed in (7,8). Even though accurate in principle, the method suffers from practical limitations associated with the need to create a satisfactory watertight seal; it has therefore largely been supplanted by other approaches. The most important methods currently used are strain gauge PG, impedance PG, and photo PG. The working principles and applications of each of these methods will be described in the following sections.

Strain Gauge Plethysmography. Introduced by Whitney in 1953 (9), strain gauge plethysmography (SPG) measures changes in limb circumference to deduce volume variations caused by blood flow activity. The strain gauge is a stretchable device whose electrical resistance depends on its length; for SPG, it is fitted under tension in a loop around the limb of interest. Pulsatile and venous blood volume changes induce stretching/relaxing of the gauge, which is translated into a measurable resistance variation. Measurement interpretation is based on the premise that the local circumference variation is representative of the overall change in limb volume.

Whitney introduced a strain gauge consisting of flexible tubing of length l made from silastic, a silicone-based elastomer, filled with mercury. Stretching of the tubing increases the length and decreases the cross-sectional area a of the mercury-filled space, leading to an increase in its resistance R according to

$$R = \rho_m \frac{l^2}{v} \tag{1}$$

where ρ_m denotes mercury's resistivity (96 $\mu\Omega$ cm), and v is the mercury volume, $v = l\ a$. Differentiation of Eq. 1. shows that the gauge's change in resistance is proportional to its length variation (10):

$$\frac{\Delta R}{R} = 2\frac{\Delta \lambda}{l} \tag{2}$$

Whitney's original strain gauge design is still the most commonly used in SPG. Because of typical gauge dimensions and mercury's resistivity, the devices have rather small resistance values, in the range of 0.5–5 Ω.

The measured limb is often approximated as a cylinder of radius r, circumference C, length L, and volume V. By expressing the cylinder volume in terms of its circumference, $V = C^2L/(4\pi)$, and then differentiating V with respect to C, it is shown that the fractional volume change is proportional to relative variations in circumference:

$$\frac{dV}{V} = 2\frac{dC}{C} \tag{3}$$

Because changes in C are measured by the strain gauge according to Eq. 1, the arm circumference is proportional to the tubing length, and so the following relationship holds:

$$\frac{\Delta V}{V} = \frac{\Delta R}{R} \tag{4}$$

Whitney used a Wheatstone bridge, a measurement circuit of inherently great sensitivity, to detect changes in gauge resistance. In this configuration, shown in Fig. 2a, three resistors of known value together with the strain gauge form a network, which is connected to a voltage source and a sensitive voltmeter in such a manner that zero voltage is detected when $R_1/R_2 = R_3/R$. In this case, the bridge is said to be balanced. Changes in strain, and therefore R, cause the circuit to become unbalanced, and a nonzero voltage develops between points B and D according to

$$\frac{V_{BD}}{V_{AC}} = \frac{R_2}{R_1 + R_2} - \frac{R}{R_3 + R} \tag{5}$$

One disadvantage of the circuit is its nonlinear voltage response with respect to variations in R. For small changes, however, these nonlinearities remain small, and negligible errors are introduced by assuming a linear relationship.

A more significant shortcoming of the Wheatstone setup is that the measurement is influenced by the voltage drop across the lead resistance; especially for small values of R, such as those encountered in SPG applications, this can be a significant source of error. Therefore, modern SPG instruments use a so-called four-wire configuration, which excludes influences of lead resistance entirely. Figure 2b shows the concept. An electronic source is connected to the strain gauge with two excitation leads of resistance R_{ex1}, R_{ex2}, sending a constant current I through the strain gauge. Two probing leads with resistances R_{pr1}, R_{pr2} connect the device to an instrumentation amplifier of high impedance $R_{amp} \gg R_{pr1}, R_{pr2}$, which measures the voltage drop $V_{SG} = I \times R$ across the strain gauge. Because V_{SG} is independent of R_{ex1} and R_{ex2}, and there is negligible voltage drop across R_{pr1} and R_{pr2}, lead resistances do not influence the measurement of R.

Recently, a new type of plethysmography strain gauge has been introduced, which measures circumference variations in a special band that is worn around the limb of interest. The band, which has a flexible zigzag structure to allow longitudinal stretching, supports a nonstretching nylon loop, whose ends are connected to an electromechanical length transducer. Changes in circumference are thus translated into translational motion, which the transducer measures on an inductive basis, with 5 μm

accuracy (11). In evaluation studies, the new design performed comparable to traditional strain gauge designs (12).

Impedance Plethysmography. Electrical conductivity measurements on the human body for the evaluation of cardiac parameters were performed as early as the 1930s and 1940s. Nyboer is widely credited with the development of Impedance Plethysmography (IPG) for the measurement of blood flow to organs, and its introduction into clinical use (13–15).

The frequency-dependent, complex electrical impedance $Z(f)$ of tissue is determined by the resistance of the inter- and intracellular spaces, as well as the capacitance across cell membranes and tissue boundaries. The IPG measurements are performed at a single frequency in the 50–100 kHz range, and only the impedance magnitude Z (not the phase information) is measured. Therefore, Z can be obtained by applying Ohms law

$$Z = \frac{\hat{V}}{\hat{I}} \tag{6}$$

where $\hat{V}$ and $\hat{I}$ denote voltage and current amplitude values, respectively.

In the mentioned frequency range, the resistivity of different tissues varies by about a factor of 100, from $\sim 1.6\ \Omega\cdot$m for blood to $\sim 170\ \Omega\cdot$m for bone. Tissue can be considered a linear, approximately isotropic, piecewise electrically homogeneous volume conductor. Some organs, however, notably the brain, heart and skeletal muscles, show highly anisotropic conductivity (16).

Figure 3 shows a schematic for a typical four-electrode IPG setup. Two electrodes are used to inject a defined current into the body part under investigation, and two separate electrodes between the injection points measure the voltage drop that develops across the section of interest. The impedance magnitude Z is obtained, via Eq. 6, from the known current amplitude and the measured voltage amplitude. The four-electrode arrangement is used to eliminate the influence of the high skin impedance ($Z_{s1} = Z_{s4}$), which is $\sim$ 2–10 times greater than that of the underlying body tissue (17). If the same two electrodes were used to inject the current as well as to pick up the voltage drop, the skin resistance would account for most of the signal and distort the information sought.

The current source generates a sinusoidal output in the described frequency range and maintains a constant amplitude of typically 1 mA. This provides sufficient signal noise ratio (SNR) to detect physiologic activity of interest but is $\sim$ 50 times below the pain threshold for the employed frequency range, and therefore well below potentially hazardous levels.

The voltage difference between the pick-up electrodes is measured with an instrumentation amplifier, whose input impedance is much greater than that of skin or underlying tissue. Therefore, the influence of skin impedance can be neglected, and the measurement yields the voltage drop caused by the tissue of interest. The output of the instrumentation amplifier is a sinusoidal voltage with amplitude proportional to the impedance of interest. The signal needs to be demodulated, that is, stripped of its carrier frequency, to determine the instantaneous amplitude value. This is done with a synchronous rectifier followed by a low pass filter, a technique also known as synchronous, lock-in, or homodyne detection. Figure 3 shows a possible analog circuit implementation; a discriminator, or zero-crossing detector, actuates a switch that, synchronously with the modulation frequency, alternately connects the buffered or inverted measured signal to a low pass filter. If phase delays over transmission lines can be neglected, this will generate a noninverted signal during one-half of a wave, say the positive half, and an inverted signal during the negative half wave. As a result, the carrier frequency is rectified. The low pass filter averages the signal to remove ripple and generates a signal proportional to the carrier frequency amplitude. Inspection shows that frequencies other than the carrier frequency (and its odd harmonics) will produce zero output.

Increasingly, the measured voltage is digitized by an analog-to-digital converter, and demodulation is achieved by a microprocessor through digital signal processing.

The measured impedance is composed of a large direct current (DC), component onto which a small (0.1–1%)

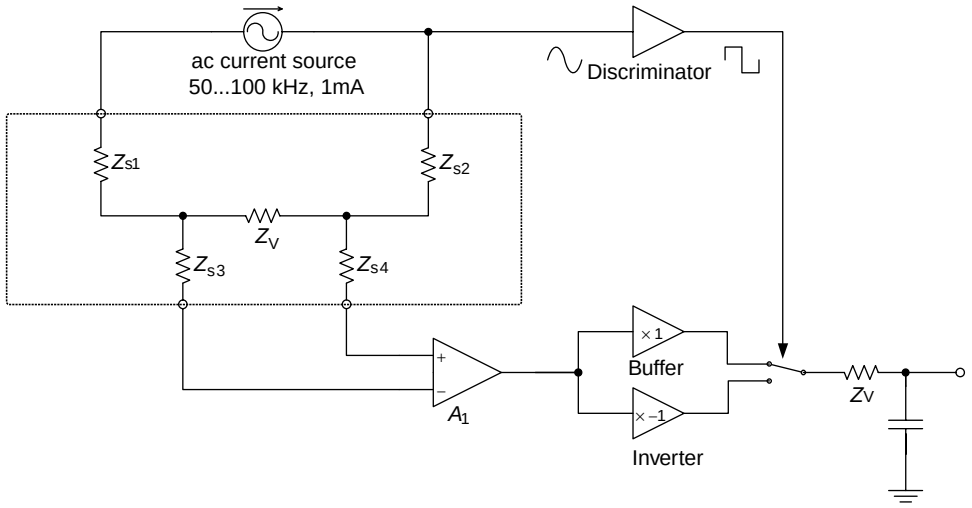


Figure 3. Four-lead tissue impedance measurement; the configuration mitigates influence of skin resistance (A_1, instrumentation amplifier). Synchronous detection removes carrier signal.

time-varying component is superimposed. The former is the constant impedance of the immobile tissue components, such as bone, fat, and muscle, and the latter represents impedance variations induced by volume changes in the fluid tissue components:most significantly, blood volume fluctuations in the vascular component. To obtain a quantitative relationship between measured impedance variations ΔZ and the change in blood volume, a simple electrical tissue equivalent may be considered, consisting of two separate, parallel-connected volume conductors of equal length L. One of these represents the immobile tissue components of impedance Z_0. The other is assigned resistivity ρ and a variable cross-sectional area A, thereby modeling impedance changes caused by variations in its volume V according to the relationship

$$Z_v = \rho \frac{L}{A} = \rho \frac{L^2}{V} \tag{7}$$

Here Z_v denotes the variable compartment's impedance, which is generally much greater than that of the immobile constituents. Therefore, the parallel impedance of both conductors can be approximated $Z \approx Z_0 - Z_v$. To obtain a functional relationship between the measured changes in impedance ΔZ and variations in blood volume ΔV, Eq. 7 is solved for V and differentiated with respect to Z. Making use of $Z \approx Z_0$ and $dZ \approx -dZ_v$ yields

$$\Delta V = -\frac{\rho L^2}{Z_0^2} \Delta Z_v \tag{8}$$

The IPG measurements do not allow independent determination of Z_0 and Z_v; however, the dc component of the IPG signal serves as a good approximation to Z_0, while the ac part closely reflects changes in Z_v. Low pass filtering of the IPG signal extracts the slowly varying dc components. Electronic subtraction of the dc part from the original signal leaves a residual that reflects physiologic impedance variations. The ac/dc separation can be implemented with analog circuitry. Alternatively, digital signal processing may be employed for this task. Digital methods help alleviate some of the shortcomings of analog circuits, especially the occurrence of slow signal drifts. In addition, software-based signal conditioning affords easy adjusting of processing characteristics, such as the frequency response, to specific applications.

Air Plethysmography. Air plethysmography (APG), also referred to as pneumoplethysmography, uses inflatable pressure cuffs with integrated pressure sensors to sense limb volume changes. The cuff is inflated to a preselected volume, at which it has snug fit, but at which interference with blood flow is minimal. Limb volume changes due to arterial pulsations, or in response to occlusion maneuvers with a separate cuff, cause changes in the measurement cuff internal pressure, which the transducer translates into electrical signals. The volume change is given by

$$\Delta V = V \frac{\Delta P}{P} \tag{9}$$

The APG measurements can be calibrated with a bladder that is inserted between the limb and the cuff, which is filled with a defined amount of water. Because temperature changes influence air pressure inside the cuff, and it needs to be worn for a few minutes after inflation before starting the measurement, so the air volume can reach thermal equilibrium.

Photoplethysmography. Optical spectroscopic investigations of human tissue and blood reach back as far as the late nineteenth and early twentieth century, and Hertzman is widely credited with introducing photoplethysmography (PPG) in 1937 (18). The PPG estimates tissue volume changes based on variations in the light intensity transmitted or reflected by tissue.

Light transport in tissue is governed by two principle interaction processes; elastic scattering, that is, the random redirection of photons by the microscopic interfaces of the cellular and subcellular structures; second, photoelectric absorption by molecules. The scattering power of tissue is much greater (at least tenfold, depending on wavelength) than is the absorption, and the combination of both interactions cause strong dampening of the propagating light intensity, which decays exponentially with increasing distance from the illumination point. The greatest penetration depth is achieved for wavelengths in the 700–1000 nm range, where the combined absorption of hemoglobin (Hb) and other molecules show a broad minimum. Figure 4 shows the absorption spectra of Hb in its oxygenated (HbO_2) and reduced, or deoxygenated, forms.

The PPG measurements in transmission are only feasible only for tissue thickness of up to a few centimeters. For thicker structures, the detectable light signal is too faint to produce a satisfactory SNR. Transmission mode measurements are typically performed on digits and the earlobes.

Whenever tissue is illuminated, a large fraction of the light is backscattered and exits the tissue in the vicinity of the illumination point. Backreflected photons that are detected at a distance from the light source are likely to have descended into the tissue and probed deeper lying structures (Fig. 5). As a general rule, the probing depth

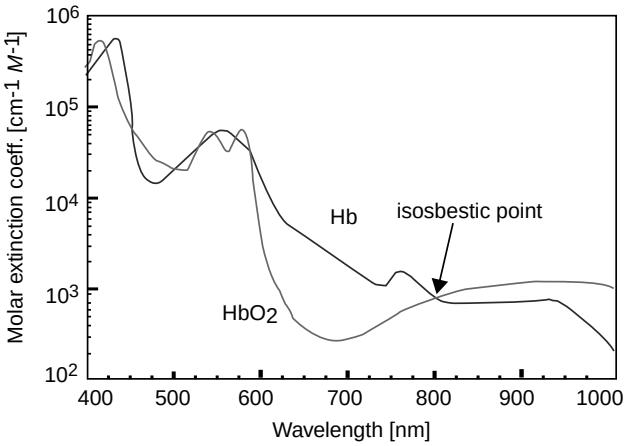


Figure 4. Hemoglobin spectra; typical PPG wavelength for oxygen saturation measurement are 660 and 940 nm.

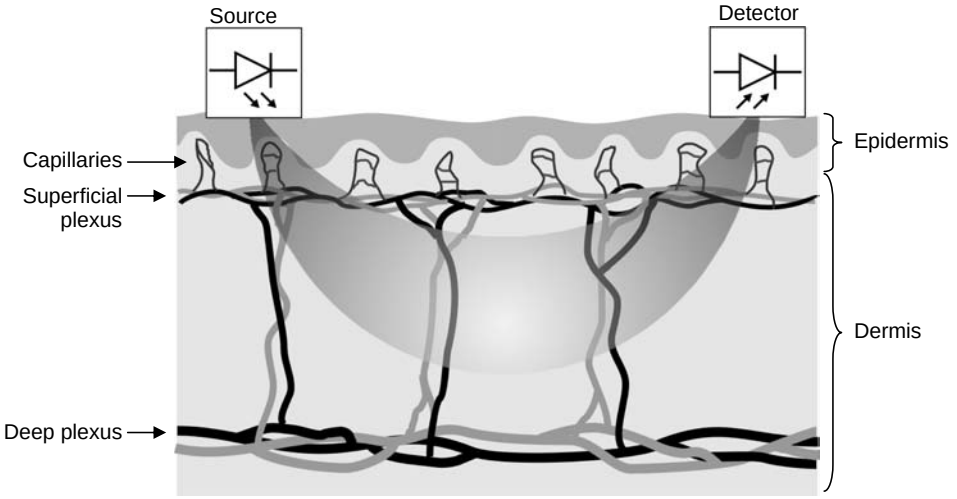


Figure 5. Schematic of the probed tissue volume in PPG reflection geometry.

equals about half the source-detector separation distance. The separation represents a tradeoff between probing volume and SNR; typical values are on the order of a few millimeters to a few centimeters. Reflection-geometry PPG can be applied to any site of the body.

Originally, incandescent light sources were used for PPG. These have been replaced with light emitting diodes (LED), owing to the latter devices' superior lifetime, more efficient operation (less heat produced), and desirable spectral properties. The LED technology has vastly evolved over the last 20 years, with a wide range of wavelengths—from the near-infrared (NIR) to the ultraviolet (UV) —and optical power of up to a tens of milliwatts currently available. These devices have a fairly narrow emission bandwidth, $\sim$ 20–30 nm, which allows spectroscopic evaluation of the tissue. The light-emitting diode (LEDs) are operated in forward biased mode, and the produced light intensity is proportional to the conducted current, which typically is on the order of tens of milliamps. Because in this configuration, the device essentially presents a short circuit to any voltage greater than its forward drop V_{LED} (typically 1–2 V), some form of current control or limiting circuitry is required. In the simplest case, this is a current-limiting resistor R_{lim} in series with the diode (Fig. 6a). The value of R_{lim} is chosen so that $I_{\mathrm{LED}} = (V_{\mathrm{cc}} - V_{\mathrm{LED}})/R_{\mathrm{lim}}$ is sufficient to drive the LED, typically on the order of tens of milliamps, but does not exceed the maximum permissible value. Depending on the voltage source, this results in adequate output stability. Often, there is a requirement to modulate or adjust the diode output intensity. In this case, an active current source is better suited to drive the LED. Figure 6b shows the example of a voltage-controlled current source with a boost field effect transistor. Here, the light output is linearly dependent on the input voltage from zero to maximum current; the latter is determined by LEDs thermal damage threshold.

Either phototransistors (PT) or photodiodes (PD) are used as light sensors for PPG. Both are semiconductor devices with an optically exposed *pn* junction. Photons are absorbed in the semiconductor material, where they create free charges through internal photoelectric effect. The charges are separated by the junction potential, giving rise to a photocurrent I_{p} proportional to the illumination intensity. While a PD has no internal gain mechanism and requires external circuitry to create a suitable signal, the PT (like any transistor) achieves internal current amplification h_{FE}, typically by a factor of $\sim$ 100. Figure 7a shows how a load resistor is used to convert the PT output current to voltage that is approximately proportional to

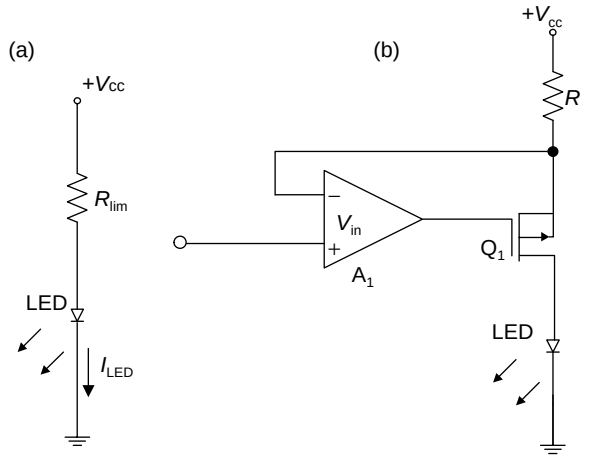


Figure 6. LED drive circuits. (a) Simple current limiting resistor. (b) Voltage controlled current source.

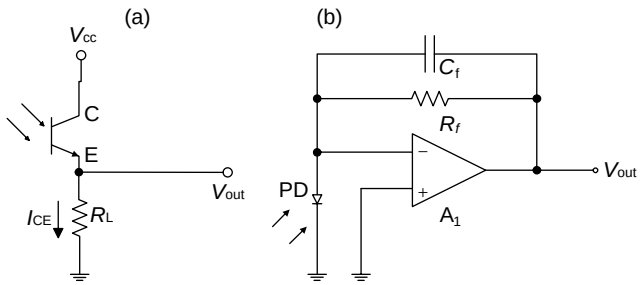


Figure 7. Light detection circuits. (a) Phototransistor operation (V_{CE} = collector-emitter voltage). (b) Use of photodiode with current-to-voltage converter (transimpedance amplifier).

the incident light according to

$$V_L = I_{CE} R_L = h_{FE} I_p R_L \tag{10}$$

h_{FE} unavoidably depends on a number of factors including I_p and V_{CE}, thus introducing a nonlinear circuit response. In addition, the PT response tends to be temperature sensitive. The light sensitivity of a PT is limited by its dark current, that is, the output at zero illumination, and the associated noise. The largest measurable amount of light is determined by PT saturation, that is, when V_L approaches V_{cc}. The device's dynamic range, that is, the ratio between the largest and the smallest detectable signal is on the order of three-to-four orders of magnitude.

Figure 7b shows the use of a PD with photoamplifier. The operational amplifier is configured as a current-to-voltage converter and produces an output voltage

$$V_o = I_p R_f \tag{11}$$

Capacitor C_f is required to reduce circuit instabilities. Even though compared to (a) this circuit is more costly and bulky because of the greater number of components involved, it has considerable advantages in terms of temperature stability, dynamic range, and linearity. The PD current varies linearly with illumination intensity over seven-to-nine orders of magnitude. The circuit's dynamic range is determined by the amplifier's electronic noise and its saturation, respectively, and is on the order of four decades. Proper choice of R determines the circuit's working range; values in the 10–100 MΩ range are not uncommon.

It is clear from considering the strong light scattering in tissue that the PPG signal contains volume-integrated information. Many models of light propagation in tissue based on photon transport theory have been developed that are capable of computing realistic light distributions even in inhomogeneous tissues, such as the finger (19). However, it is, difficult, if not impossible, to exactly identify the sampled PPG volume because this depends critically on the encountered optical properties as well as the boundary conditions given by the exact probe placement, digit size and geometry, and so. These factors vary between individuals, and even on the same subject are difficult to quantify or even to reproduce. Therefore, PPG methods do generally not employ a model-based approach, and the origin of the PPG signal as well as its physiologic interpretation has been an area of active research (20–23).

The PPG signals are analyzed based on their temporal signatures, and how those relate to known physiology and empirical observations. The detected light intensity has a static, or dc component, as well as a time-varying ac part. The former is the consequence of light interacting with static tissues, such as skin, bone, and muscle, while the latter is caused by vascular activity. Different signal components allow extraction of specific anatomical and physiological information. For example, the signal component showing cardiac pulsation, also called the digital volume pulse (DVP), can be assumed to primarily originate in the arterial bed, a premise on which pulse oximetry, by far the most widely used PPG application, is based. This method obtains two PPG signals simultaneously at two wavelengths that are spectrally located on either side of the isosbestic point. The ratio of the DVP peak amplitudes for each wavelength, normalized by their respective dc components, allows assessment of quantitative arterial oxygen saturation.

A number of other applications of the DVP signal have been proposed or are under investigation. For example, the DVP waveform is known to be related to the arterial blood pressure (ABP) pulse. Using empirically determined transfer functions, it is possible to derive the ABP pulse from PPG measurements (24).

Another potential use of arterial PPG is the assessment of arterial occlusive disease (AOD). It is known that the arterial pulse form carries information about mechanical properties of the peripheral arteries, and PPG has been investigated as a means to noninvasively assess arterial stiffness. To better discriminate features in the PPG signal shape, the second derivative of the signal is analyzed (second-derivative plethysmography, SDPTG) (23).

The PPG is furthermore used for noninvasive peripheral BP monitoring in the vascular unloading technique.

Continuous Wave Doppler Ultrasound

Doppler ultrasound (US) methods are capable of measuring blood flow velocity and direction by detecting the Doppler shift in the ultrasound frequency that is reflected by the moving red blood cells. The acoustic Doppler effect is the change in frequency of a sound wave that an observer perceives who is in relative motion with respect to the sound source. The amount of shift Δf in the US Doppler signal is given by (25)

$$\Delta f = \frac{2 f_0 v_b \cos\Theta}{c} \tag{12}$$

where f_0 is the US frequency, v_b is the red blood cell velocity, Θ is the angle between the directions of US wave propagation and blood flow, and c is the speed of sound in tissue. Equation 12 demonstrates a linear relationship between blood flow and US Doppler shift. It is also seen that the shift vanishes if the transducer is perpendicular to the vessel because there is no blood velocity component in the direction of wave propagation. The algebraic sign of the shift depends on the flow direction (toward/away from) with respect to the transducer and is hence influenced by angle Θ. A factor of two appears in Eq. 12 because the Doppler effect takes place twice; the first occurs when the red blood cell perceives a shift in the incoming US wave, and the second shift taks place when this frequency is backreflected toward the transducer. Using typical values for the quantities in Eq. 12 ($f_0 = 5$ MHz, $v_b = 50$ cm·s^{-1}, $\Theta = 45^\circ$) yield frequency shifts of 2.3 kHz, which falls within the audible range (25).

Because the shift is added to the US frequency, electronic signal processing is used to remove the high frequency carrier wave. Analog demodulation has been used for this purpose; by mixing, that is, multiplying, the measured frequency with the original US frequency and low pass filtering the result, the carrier wave is removed, and audio-range shift frequencies are extracted. In so-called quadrature detection, this demodulation process yields two

separate signals, one for flow components toward the detector, and one for flow away from it. Modern instruments typically employ digital signal processing, such as fast Fourier transformation (FFT), to accomplish this.

Continuous wave (CW) Doppler refers to the fact that the measurement is performed with a constant, nonmodulated US wave (i.e., infinite sine-wave). This technology does not create echo pulses, and hence does not allow any form of depth-profiling or imaging. Its advantages are its simplified technology, and that it is not restricted to a limited depth or blood velocity. Because the signal is generated continuously, CW transducers require two separate crystals, one for sound generation, and one for detection. The two elements are separated by a small distance and are inclined toward each other. The distance and angle between the elements determine the overlap of their respective characteristics, and thus determine the sensitive volume. All blood flow velocity components within this volume contribute to the Doppler signal.

In the simplest case, the measured frequency shift is made audible through amplification and a loudspeaker, and the operator judges blood velocity from the pitch of the tone (higher pitch = faster blood movement). These devices are used in cuff-based arterial pressure measurements, to detect the onset of pulsation. They also permit qualitative assessment of arterial stenosis and venous thrombosis. More sophisticated instruments display the changing frequency content of the signal as a waveform (velocity spectral display). Doppler spectra are a powerful tool for the assessment of local blood flow in the major vessels, especially when combined with anatomical US images (Duplex imaging, see section on imaging methods).

Peripheral Vascular Measurements

The following is a description of peripheral vascular measures that can be derived with the nonimaging techniques described in the preceding sections.

Assessment Of Peripheral Arterial Occlusive Disease. Peripheral arterial occlusive disease (PAOD) is among the most common peripheral arterial diseases (26), with a reported prevalence as high as 29% (27). The PAOD is usually caused by atherosclerosis, which is characterized by plaque buildup on the inner-vessel wall, leading to congestion, and hence increased blood flow resistance. Risk factors for PAOD are highly correlated with those for coronary artery disease, and include hypertension, smoking, diabetes, hyperlipidemia, age, and male sex (27–29). A symptom of PAOD is intermittent claudication, that is, the experience of fatigue, pain, and cramping of the legs during exertion, which subsides after rest. Ischemic pain at rest, gangrene, and ulceration frequently occur in advanced stages of the disease (29,30).

Ankle Brachial Index Test. The Ankle Brachial Index (ABI) is a simple and reliable indicator of PAOD (31), with a detection sensitivity of 90% and specificity of 98% for stenoses of >50% in a major leg artery (27,29). The ABI for each leg is obtained by dividing the higher of the posterior and anterior tibial artery systolic pressures for that leg by the greater of the left- and right-brachial artery systolic pressures. In normal individuals both systolic values should be nearly equal, and ABIs of 1.0–1.1 are considered normal. In PAOD, the increased peripheral arterial resistance of the occluded vessels causes a drop in blood pressure, thus diminishing the ABI. Because of measurement uncertainties, ABI values of < 0.9 generally are considered indicative of PAOD (27,29). Table 1 shows a scheme for staging of disease severity according to ABI values, as recommended by the Society of Interventional Radiology (SIR) (32).

Table 1. Ankle-Brachial Index (ABI) staging after ()

AB Ratio Range	Stage
1.0–1.1	Normal
< 1.0	Abnormal, possibly asymptomatic
0.5–0.9	Claudication
0.3–0.5	Claudication, rest pain, severe occlusive disease
< 0.2	Ischemia, gangrenes

The ABI test is performed with the patient in a supine position, and inflatable pressure cuffs are used to occlude the extremities in the aforementioned locations. The systolic pressure is determined by deflating the cuff and noting the pressure at which pulse sounds reappear. The audio signal from a CW Doppler instrument is used to determine the onset of arterial pulse sounds. Although the test can be performed manually with a sphygmomanometer, there are dedicated vascular lab instruments that automatically control the pressure cuff, register the pressure values, and calculate the ABI. Such instruments are commercially available, for example, from BioMedix,MN, Hokanson, WA, and Parks Medical Electronics, OR.

The primary limitation of the ABI test is that arterial calcifications, such as are common in diabetics, can resist cuff pressure thus causing elevated pressure readings. In those patients, the toe pressure as determined by PPG or SPG may be used because the smaller digital arteries do not develop calcifications (29). The normal toe systolic pressure (TSP) ranges from 90 mmHg to 100 mmHg, or 80% to 90% brachial pressure. The TSP < 30 mmHg (3.99 kPa) is indicative of critical ischemia (33). Also, pulse volume recordings (PVR, see below) are not affected by calcifications (26).

Segmental Systolic Pressure. Segmental Systolic Pressure (SSP) testing is an extension of the ABI method, wherein the arterial systolic pressure is measured at multiple points on the leg, and the respective ratios are formed with the higher of the brachial systolic readings. As is the case for ABI testing, a low ratio for a leg segment indicates occlusion proximal to the segment. By comparing pressures in adjacent segments, and with the contralateral side, SSP measurements offer a way to locate pressure drops and thus roughly localize occlusions. Typically, three to four cuffs are placed on each leg; one or two on the thigh, one below the knee, and one at the ankle.

The SSP technique suffers from the same limitation as does the ABI test, that is, it produces falsely elevated readings for calcified arteries. Again, TSP and PVR are

advised for patients where calcifications might be expected because of known predisposition (e.g., diabetics), or unusual high ABI (> 1.5) (26).

Pulse Volume Recording/Pulse Wave Analysis. Pulse volume recording (PVR), a plethysmographic measurement, is usually combined with the ABI or SSP test. Either the pressure cuffs used in these measurements are equipped with a pressure sensor so they are suitable for recordind volume changes due to arterial pulsation, or additional strain gauge sensors are used to obtain segmental volume pulsation. In addition to the leg positions, sensors may be located on the feet, and PPG sensors may be applied to toes. Multiple waveforms are recorded to obtain a representative measurement; for example, the Society for Vascular Ultrasound (SVU) recommends a minimum of three pulse recordings (34). The recorded pulse amplitudes reflect the local arterial pressure, and the mean amplitude, obtained by integrating the pulse, is a measure of the pulse volume. The pulsatility index, the ratio of pulse amplitude over mean pulse volume serves as an index for PAOD, with low values indicating disease (26).

Besides its amplitude and area, the pulse contour can be evaluated for indications of PAOD in a method referred to as pulse wave analysis (PWA). The normal pulse has a steep (anacrotic) rise toward the systolic peak, and a slower downward slope, representing the flow following the end of the heart's left ventricular contraction. The falling slope is interrupted by a short valley (the dicrotic notch), which is caused by the closing of the aortic valve. The details of the pulse shape are the result of the superposition of the incoming arterial pulse and backward-traveling reflections from arterial branchings and from the resistance caused by the small arterioles, as well as from atherosclerotic vessels (35). The pulse shape distal to occlusions tends to be damped and more rounded, and loses the dicrotic wave. Combined SSP and PVR recording has been reported over 90% accurate for predicting the level and extent of PAD (26).

Recently, PWA has increasingly been used to infer central, that is, aortic, arterial parameters from peripheral arterial measures, typically obtained from tonometric measures on the radial artery. These methods, which are commercialized for example by Atcor Medical, Australia and Hypertension Diagnostics, Inc, MN, rely on analytic modeling of transfer functions that describe the shaping of the pulse from the aorta to the periphery. These analytical functions are typically based on models mimicking the mechanical properties of the arterial tree by equivalent circuits (so-called Windkessel models) representing vascular resistances, capacitances, and inductances (36–38). From the derived aortic waveform, parameters are extracted that indicate pathologic states, such as hypertension and central arterial stiffness. One example is the augmentation index, the pressure difference in the first and second systolic peak in the aorta (39,40).

Pulse wave velocity (PWV) depends on arterial stiffness, which is known to correlate with increased cardiovascular risks. The PWV is established by measuring pulse waves on more than one site simultaneously, and dividing their respective separation difference from the heart by the time difference between the arrivals of comparable wave form features. Peripheral PWV measurements on toes and fingers have been conducted with PPG (41).

It was recently shown that arterial stiffness assessed from PWA and PVA on multisite peripheral tonometric recording correlates with endothelial function. The latter was assessed by US measurements of brachial arterial diameter changes in response to a reactive hyperemia maneuver (42). There is increasing clinical evidence that PWA can serve as a diagnostic tool for hypertension and as a cardiac disease predictor (43).

CW Doppler Assessment. Velocity spectral waveform CW Doppler is useful for assessing peripheral arterial stenoses. The Doppler waveform obtained from a normal artery shows a triphasic shape; there is a strong systolic peak, followed by a short period of low amplitude reversed flow, reversing again to a small anterograde value during mid-diastole. Measuring at a stenosis location shows a waveform of increased velocity and bi- or monophasic behavior. Measurements distal to stenosis appear monophasic and dampened (26,33).

Venous Congestion Plethysmography. The described PG methods are most valuable for the assessment of peripheral vascular blood flow parameters for the diagnosis of peripheral vascular diseases (PVD). Venous congestion PG (VCP), also called venous occlusion PG (VOP), allows the measurement of a variety of important vascular parameters including arterial blood flow (Q_a), venous outflow (Q_{vo}), venous pressure (P_v), venous capacitance (C_v), venous compliance (C), and microvascular filtration. In VCP, a pressure cuff is rapidly inflated on the limb under investigation to a pressure, typically between 10 (1.33 kPa) and 50 mmHg (6.66 kPa), sufficient to cause venous occlusion, but below the diastolic pressure so that arterial flow is not affected. Blocking the venous return in this fashion causes blood pooling in the tissue distal to the cuff, an effect that can be quantified by measuring volume swelling over time with PG. Figure 8 sketches a typical

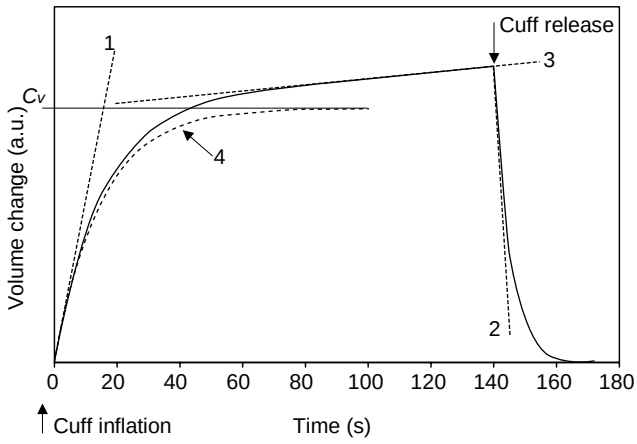


Figure 8. Schematic of typical VCP response curve. (*1*), asymptotic arterial blood flow; (*2*), asymptotic venous outflow; (*3*), swelling due to filtration; (*4*), effect of venous blood pooling only, obtained by subtracting the filtration component (*3*); C_v, venous capacitance.

volume response curve. Upon cuff inflation, an exponential increase in volume is observed, which is caused by the filling of postcapillary compliance vessels, and for which the time constant is $\sim$ 15 s in healthy individuals (44). The initial rate of swelling (indicated by asymptote 1), expressed in (mL·min^{-1}), is a measure of the arterial blood flow. When the venous pressure reaches the cuff pressure, blood can again flow out, and the relative volume increase reaches a plateau, which equals C_v. If C_v is measured for different occlusion pressure values, the slope of the C_v versus pressure curve yields C (in mL·mmHg^{-1} $\times$ 10^{-2}). Upon cuff release, the VCP curve shows fast exponential decay (on the orders of seconds), whose initial rate (indicated by asymptote 2) is a measure for Q_{vo} (45,46). With rising venous pressure an increase in fluid leakage, or filtration, takes place from the veins into the interstitial space, leading to additional tissue volume increase. Therefore, the PG response shows a second slow exponential component with a time constant on the order of 800 s (asymptote 3), from which can be deduced the capillary filtration capacity *CFC* (11).

Another application of VCP is the noninvasive diagnosis of deep vein thrombosis (DVT). The presence of venous blood clots impacts on venous capacitance and outflow characteristics. It has been shown that the combined measurement of C_v and Q_{vo} may serve as a diagnostic discriminator for the presence of DVT (45). A computerized SPG instrument is available (Venometer, Amtec Medical Ltd, Antrim, Northern Ireland), which performs a fully automated VCP measurement and data analysis for the detection of DVT (47).

Laser Doppler Flowmetry

Laser Doppler flowmetry (LDF) is a relatively young method of assessing the perfusion of the superficial microvasculature. It is based on the fact that when light enters tissue it is scattered by the moving red blood cells (RBC) in the superficial vessels (see PPG illustration in Fig. 5), and as a result the backscattered light experiences a frequency (Doppler) shift. For an individual scattering event, the shift magnitude can be described exactly; it depends on the photon scattering angle and the RBCs velocity, and it is furthermore related to the angular orientation of blood flow with respect the path directions of the photon before and after the scattering event. In actual tissue measurements, it is impossible to discern individual scattering events, and the obtained signals reflect stochastic distributions of scattering angles and RBC flow directions and velocities present in the interrogated volume.

In its simplest form, an LDF measurement is obtained by punctual illumination of the tissue of interest with a laser source, and by detecting the light that is backscattered at (or close to) the illumination site with a photodetector, typically a photodiode. The laser is a highly coherent light source, that is, its radiation has a very narrow spectral bandwidth. The spectrum of the backscattered light is broadened because it contains light at Doppler frequencies corresponding to all scattering angles and RBC motions occurring in the illuminated volume. Because of typical flow speeds encountered in vessels, the maximum Doppler shifts encountered are on the order of 20 kHz, which corresponds to a relative frequency variation of $\sim 10^{-10}$. Frequency changes this small can be detected because the shifted light components interfere with the unscattered portion of the light, causing a "beat signal" at the Doppler frequency. This frequency, which falls roughly in the audio range, is extracted from the photodetector signal and further processed. The small amount of frequency shift induced by cell motion is the reason that spectrally narrow, high quality laser sources (e.g., HeNe gas lasers or single-mode laser diodes) are required for LDF measurements.

The basic restriction of LDF is its limited penetration depth. The method relies on interference, an effect that requires photon coherence. Multiple scattering, however, such as experienced by photons in biological tissues, strongly disturbs coherence. For typical tissues, coherence is lost after a few millimeters of tissue. The sensitive volume of single-point LDF measurements is therefore on the order of 1 mm^3.

Single-point LDF measurements are usually performed with a fiberoptic probe, which consist of one transmitting fiber and one adjacent receiving fiber. Typical distances between these are from a few tenths of a millimeter to $>$ 2 mm. According to light propagation theory (see section on PPG), farther separations result in deeper probing depths, however, because of coherence loss, the SNR decreases exponentially with increasing distance, limiting the maximum usable separation.

The use of a probe makes a contact-based single-point measurement very convenient. Fiber-based probe have been developed that implement several receivers at different distances from the source (48). This allows a degree of depth discrimination of the measured signals, within the aforementioned limits.

The LDF signal is analyzed by calculation the frequency power spectrum of the measured detector signal, usually after band-pass filtering to exclude low frequency artifacts and high frequency noise. From this the so-called flux or perfusion value—a quantity proportional to the number of RBCs and their root-mean-squared velocity, stated in arbitrary units—is obtained. It has been shown that the flux is proportional to the width the measured Doppler power spectrum, normalized to optical power fluctuations in the setup (49).

Laser Doppler imaging (LDI) is an extension of the LDF technique that allows the instantaneous interrogation of extended tissue areas up to tens of centimeters on each side. Two approaches exist. In one implementation, the laser beam is scanned across the desired field of view, and a photodetector registers the signal that is backreflected at each scanning step. In another technology, the field of view is broadly illuminated with one expanded laser beam, and a fast CMOS camera is used to measure the intensity fluctuations in each pixel, thus creating an image. This second approach, while currently in a stage of relative infancy, shows the potential for faster frame rates than the scanning imagers, and it has the advantage of avoiding mechanical components, such as optical scanners (50).

The LDF/LDI applications include the assessment of burn wounds, skin flaps, and peripheral vascular perfusion

problems, such as in Raynaud's disease or diabetes. The vascular response to heating (51) and the evaluation of carpal tunnel syndrome also have been studied (52).

IMAGING METHODS

This section provides an overview of existing medical imaging methods, and how they relate to vascular assessment. Included here are imaging modalities that involve the use of contrast agents or of radioactive markers, even though these methods are not considered noninvasive in the strictest sense of the word.

Ultrasound Imaging

Structural US imaging, especially when combined with Doppler US methods, is at present the most important imaging technique employed in the detection of PVD.

Improvements in ultrasound imaging technology and data analysis methods have brought about a state in which ultrasonic images can, in some circumstances, provide a level of anatomical detail comparable to that obtained in structural X-ray CT images or MRIs (53,54). At the same time, dynamic ultrasound imaging modalities can readily produce images of dynamic properties of macroscopic blood vessels (55,56). The physical phenomenon underlying all the blood-flow-sensitive types of ultrasound imaging is the Doppler effect, wherein the frequency of detected ultrasonic energy is different from that of the source, owing to the interactions of the energy with moving fluid and blood cells as it propagates through tissue. Several different varieties have become clinically important.

The earliest, most basic version of dynamic ultrasound imaging is referred to as color flow imaging (CFI) or color Doppler flow imaging (CDFI). Here, the false color value assigned to each image pixel is a function of the average frequency shift in that pixel, and the resulting image usually is superimposed upon a coregistered anatomic image to facilitate its interpretation. Interpretation of a CDFI image is complicated by, among other factors, the dependence of the frequency shift on the angle between the transducer and the blood vessels in its field of view. In addition, signal/noise level considerations make it difficult to measure low but clinically interesting flow rates. Many of these drawbacks were significantly ameliorated by a subsequently developed imaging modality known as either power flow imaging (PFI) or power Doppler imaging (PDI) (57). The key distinction between PDI and CDFI is that the former uses only the amplitude, or power, of ultrasonic energy reflected from erythrocytes as its image-forming information; in consequence, the entire contents of a vessel lumen have a nearly uniform appearance in the resulting displayed image (58).

The greatest value of Doppler imaging lies in the detection and assessment of stenoses of the larger arteries, as well as in the detection of DVT.

An increasingly employed approach to enhancing ultrasound images, at the cost of making the technique minimally invasive, is by introducing a contrast agent. The relevant contrast agents are microbubbles, which are microscopic, hollow, gas-filled polymer shells that remain confined in the vascular space and strongly scatter ultrasonic energy (59). It has been found that use of microbubbles permits a novel type of Doppler-shift-based ultrasound imaging, called contrast harmonic imaging (CHI). The incident ultrasound can, under the correct conditions, itself induce oscillatory motions by the microbubbles, which then return reflections to the transducer at not only the original ultrasonic frequency, but also at its second and higher harmonics. While these signals are not high in amplitude, bandpass filtering produces a high SNR, with the passed signal basically originating exclusively from the vascular compartment of tissue (60). The resulting images can have substantially lower levels of "clutter" arising from nonvascular tissue, in comparison to standard CDFI or PDI methods.

Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is certainly the most versatile of the imaging modalities, and give the user the ability to study and quantify the vasculature at different levels. That is, depending on the details of the measurement, the resulting image may principally confer information about bulk blood flow (i.e., rates and volumes of blood transport in macroscopic arteries and veins), or about perfusion (i.e., amount of blood delivered to capillary beds within a defined time interval). The methods commonly employed to perform each of these functions are given the names magnetic resonance angiography (MRA) and perfusion MRI, respectively. As is the case for the other types of imaging treated here, each can be performed either with or without administration of exogenous contrast agents, that is, in either a noninvasive or a minimally invasive manner (61,62).

Magnetic Resonance Angiography. The MRA methods fall into two broad categories. The basic physical principle for one is flow-related enhancement (FRE), in which the pulse sequence employed has the effect of saturating the spins of ^{1}H nuclei in the nonmoving portion of the selected slice (63,64). Blood flow that has a large component of motion perpendicular to that slice remains significantly less saturated. Consequently, more intense MR signals arise from the blood than from the stationary tissues, and the contents of blood vessels are enhanced in the resulting image. Subtraction of a static image obtained near the start of the data collection sequence permits even greater enhancement of the blood vessel contents. A drawback of this approach is that vessels that lie within the selected slice, with the direction of blood flow basically parallel to it, do not experience the same enhancement.

A phase contrast mechanism is the basic physical principle for the second category of MRA techniques (65,66). Position-selective phase shifts are induced in ^{1}H nuclei of the selected slice via the sequential imposition of at least two transverse gradients that sum to a constant value across the slice. The effect is to induce zero net phase shifts among nuclei that were stationary during the gradient sequence, but a nonzero, velocity-dependent net phase shift on those that were in motion. The net phase shifts associated with the flowing blood are revealed in the

resulting image, again, especially following subtraction of an image derived from data collected before the phase contrast procedure.

Perfusion MRI. Early developments in this field necessarily involved injection of a MR contrast agent that has the effect of inducing a sharp transient drop in signal detected as the contrast bolus passes through the selected slice (67). As with many dynamic MRI techniques, the bulk of the published work is geared toward studies of the brain, where, provided that the blood–brain barrier is intact, the method's implicit assumption that the contrast agent remains exclusively within the vascular space usually is justified. A sufficiently rapid pulse sequence allows the operator to generate curves of signal intensity vs. time, and from these one can deduce physiological parameters of interest, such as the cerebral blood flow (CBF), cerebral blood volume (CBV), and mean transit time (MTT) (67,68).

The minimally invasive approach described in the preceding paragraph remains the most common clinically applied type of perfusion MRI. A noninvasive alternative would have advantages in terms of permissible frequency of use, and would be of particular value in situations where pathology or injury has disrupted the blood–brain barrier. It also would make it possible to obtain perfusion images of other parts of the body than just the brain. Such an alternative exists, in the form of a set of methods known collectively as arterial spin labeling (ASL) (67) or arterial spin tagging (69). The common feature of all these techniques is that the water component of the blood is itself used as a contrast agent, by applying a field gradient that inverts the spins of ^{1}H nuclei of arterial blood before they enter the slice selected for imaging. The signal change induced by this process is smaller than that resulting from injection of a contrast agent, however, so that requirements for high SNRs are more exacting, and subtraction of a control image a more necessary step.

An increasingly popular and important functional MRI technique is blood-oxygen-level-dependent (BOLD) imaging (67,70). This type of imaging produces spatial maps determined by temporal fluctuations in the concentration of deoxygenated hemoglobin, which serves as an endogenous, intravascular, paramagnetic contrast agent. The physiological importance of BOLD images is that they reveal spatial patterns of tissue metabolism, and especially of neuronal activity of the brain. However, careful examination of the BOLD signal indicates that it depends, in a complex way, on many vascular and non-vascular tissue parameters (67,71). As such, it is not (yet) a readily interpretable method for specifically studying peripheral vasculature.

Fast X Ray Computed Tomography

As data acquisition speeds, and consequently repetition rates, for X ray computed tomography (CT) imaging have increased over the last couple of decades, previously unthinkable dynamic imaging applications have become a reality. The most direct approach taken along these lines is to rapidly acquire sequences of images of a slice or volume of interest and then examine and interpret, at any desired level of mathematical sophistication, temporal variations in the appearance of tissue structures in the images. Of course this approach is well suited to studying only organs whose functionality entails changes in shape or volume, such as the heart and lungs, and these have been the subject of many fast CT studies.

The functioning of many other organs, such as the kidneys (72), is related to the flow of blood and/or locally formed fluids, but does not involve gross changes in size or shape. In these cases it is necessary to introduce one or more boluses of X ray contrast agents, and to use fast CT to monitor their transit (73). While these methods violate the strict definition of noninvasive procedure, we include them in this synopsis out of consideration of the fact that the health risks associated with the contrast agents ordinarily are minor in relation to those imposed by the ionizing radiation that forms the CT images. For quantitation of regional blood flow and other dynamic vascular parameters, these techniques invariably employ some version of indicator dilution theory (73).

Indicator Dilution Approach

In one clinically important example, it has been found that the detected X ray CT signal changes in a quantifiable manner following inhalation of a gas mixture containing appreciable levels of nonradioactive isotopes of xenon (typically 33% Xe and 67% O_2, or 30% Xe, 60% O_2, 10% air) (75–77). A variety of techniques based on this phenomenon, and referred to as stable xenon-enhanced CT or ^{s}Xe–CT, were subsequently developed. The common feature is inhalation of a Xe-enriched gas mixture followed by repetitive scanning to monitor the wash-in and/or wash-out of the Xe-affected signal (75). However, while negative side effects are uncommon, they are known to occur (77). Consequently, the ^{s}Xe/CT technique is applied almost exclusively to cerebral vascular imaging studies, in cases where there is a diagnosed injury or pathology (75).

Qualitatively similar approaches, known collectively as perfusion CT, that are based on monitoring the time course of passage of an injected bolus of an iodinated contrast agent, also have been developed (74,78,79). Of course, these are even more invasive than the approaches based on inhalation an inert gas. Conflicting assertions (80) have been made regarding which of two approaches, injection or inhalation based, give superior results.

Positron Emission Tomography with ^{15}O

Strictly speaking, all forms of positron emission tomography (PET) imaging (as do all other nuclear medicine procedures) violate the formal definition of noninvasive measurements. The 2.1 min half-life of the isotope ^{15}O, which is brief relative to those of the other commonly used positron emitters, makes ^{15}O-labeled water a useful indicator of blood flow in dynamic PET measurements (81,82). Of course, as a practical matter radioisotope imaging methods cannot be used on a large scale for research or for clinical screening purposes. Medical literature on ^{15}O-PET-based vascular studies, understandably, also focuses

almost exclusively on studies of circulation in the brain, in subjects with diagnosed vascular pathologies such as arteriovenous malformations (81).

Temporal-Spectral Imaging

A new area of investigation and technology development involving assessment of blood delivery to the periphery includes the use of optical array measurements combined with image formation and analysis capabilities. The basic concept, referred to as temporal-spectral imaging (TSI) (83), considers the functional interactions between peripheral tissues and their blood supply, and the added information that is attainable from a time series of volumetric images of tissue, where the latter are reconstructed from optical measurements of the hemoglobin signal. The information content of the measurement comprises three elements: first, expected differential response properties of tissues as a consequence of their varying blood supply and responses to internal and external stimuli; second, added information available from analysis of a time series, such as measures of rates, time delays, frequency content, and so on; third, further information available from a spatial deconvolution of diffusely scattered optical signals arising from deep tissue structures, to provide a volumetric image. In combination, the method provides for the assignment of multiple characteristics to each tissue type, and these can serve to distinguish one tissue from another, and healthy from diseased tissues. Because the supporting technology can be made compact, it is feasible to consider performing the optical array measurements on multiple sites of the body simultaneously in order to identify regional redistribution of blood, as can occur, for example, in response to shock.

The TSI approach builds upon an expanding investigative field known as diffuse optical tomography (DOT), which was first introduced late 1980s (84–88), and later extended to explore time-varying states (89,90). The DOT technology, unlike laser Doppler techniques, is suitable for exploring deep tissue structures and, similar to PPG, employs near-IR optical sources to assess the hemoglobin signal. In the backreflection mode, penetration depths of 3–4 cm are possible, depending on the tissue type. In the transmission mode, greater penetration is possible, including up to 12 cm of breast tissue and 6–8 cm in the forearm. For larger structures, for example, involving the lower extremities, full transmission measurements are not feasible, although off-axis measures partially inscribing the structure can be made.

The DOT measurements are made using an array of optical emitters and detectors that measure the light field reemitted from tissue over a relatively broad area. Measurements encompassing many tens of square centimeters of tissue surface, with interrogation of subsurface structures to the depths indicated above, are achievable. In practice, the DOT technique can be applied to explore the full volume of a breast, or partial volumes of the head, limbs or portions of the torso.

An example of the DOT approach extended to include TSI is presented below. First, however, it is useful to appreciate how DOT is accomplished. Figure 9 shows a schematic of the measurement strategy applied to time-series DOT. Being a tomographic technique, The DOT employs a multisource, multidetector measurement covering a relative wide area of tissue. Although light migrating in tissue is randomly scattered, predictions can be made as to the volume distribution of paths taken by photons propagating between any one source and detector. These take the shape of a fuzzy banana, and are conceptually similar to the linear trajectories taken by X rays as applied to CT imaging. Collectively, an array measurement can be processed to produce a cross-sectional or volumetric image which, when measured over time, produces a time series of images.

The TSI extends the time-series DOT technique in two ways. First, as noted, it recognizes that a wealth of differential information is available due to the significant

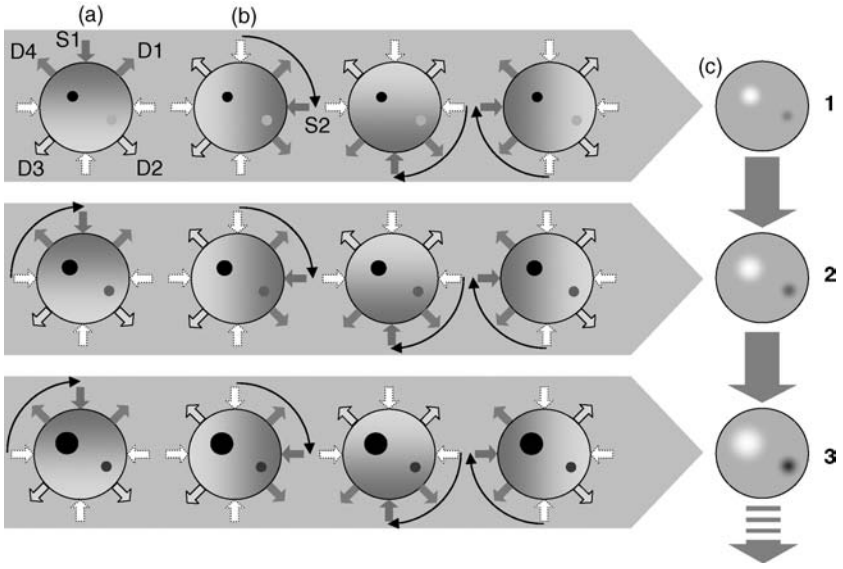


Figure 9. Temporal Spectral Imaging principle. (a) A NIR light source (S1) illuminates the target of interest in a specific location, and the reflected and transmitted light is measured at several locations (D1–D4). The source moves on to the next location (S2), and the measurement is repeated (b). After one full scan, a tomographic image can be reconstructed (c). Steps (a–c) are repeated to obtain successive image frames that show target dynamics, such as local changes in blood volume and oxygenation.

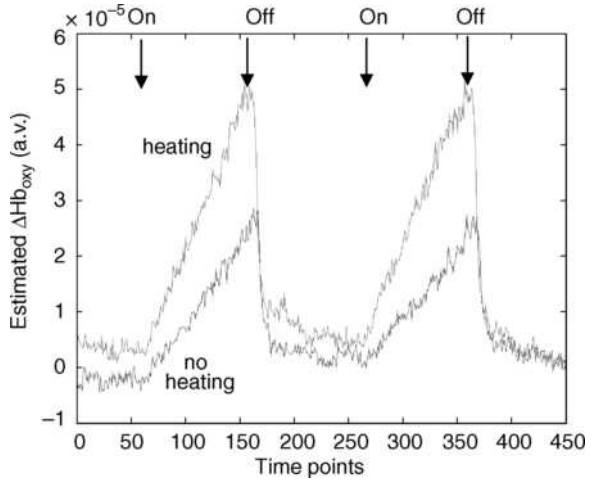


Figure 10. Time course of computed average oxyhemoglobin response to two cycles of mild venous occlusion in the forearm. Dark curve, no heat applied; light curve, with heat. [Reprinted with permission from R.L. Barbour et al., Functional imaging of vascular bed by dynamic optical tomography. Proceedings of SPIE, vol. 5369, in Physiology, Function, and Structure from Medical Images, 2004.]

variations in vascular density among the different tissue types, and in their responses to stimuli. Second, it also recognizes that for any given tissue, multiple response characteristics can be defined that serve to discriminate, as noted, one tissue type from another, or healthy from diseased. An example of this discriminatory capability is the following.

Shown in Fig. 10 is the response to partial cuff occlusion, applied to the arm in order to induce venous engorgement as recorded from a representative site about the forearm. Also shown is an enhanced response to a similar maneuver when the forearm was warmed by a thermal blanket prior to the measurement. We interpret the enhanced response as evidence of a decrease in local peripheral vascular resistance due to vasodilatation. Using the response curve shown in Fig. 10, it is possible to identify where similar behavior is seen in the corresponding cross-sectional image time series. This is shown in Fig. 11, together with an MR map of the forearm for comparison to anatomy. Regions of high correlation to the response curve are seen in many areas, with distinctively lower correlation occurring in regions that correspond to the ulna and radius, a finding consistent with known differences between soft tissue and interosseous hemodynamics (see overlay in e).

Additional discriminatory information about the considered stimulus can be gleaned from other analyses of the image time series. Motivating the approach taken is knowledge that any measure of bulk tissue properties is actually a composite of multiple overlapping features due to known differences, for example, in vascular compliance attributable to the principal elements of the vascular tree (i.e., arteries, veins and microvessels). Separation of the individual components comprising the whole is attainable using a class of methods known as blind source separation techniques.

Data in Figure 12 is an example of use of these methods, applied to four consecutive mild inflation–deflation cycles of the type illustrated in Fig. 10. The dark curve corresponds to the first principal component (PC) and accounts for $\sim$ 80% of the signal variance. The light curve is the second PC, accounting for $\sim$ 10% of total variance. Inspection reveals that the time courses of the two functions differ, and that they change from one application of mild occlusion to another. In the case of the first PC, it is seen that Hb_{oxy} levels increase promptly upon inflation. Also seen is that the magnitude of this response increases modestly following the second challenge. We interpret this to represent blood volume changes in the venous

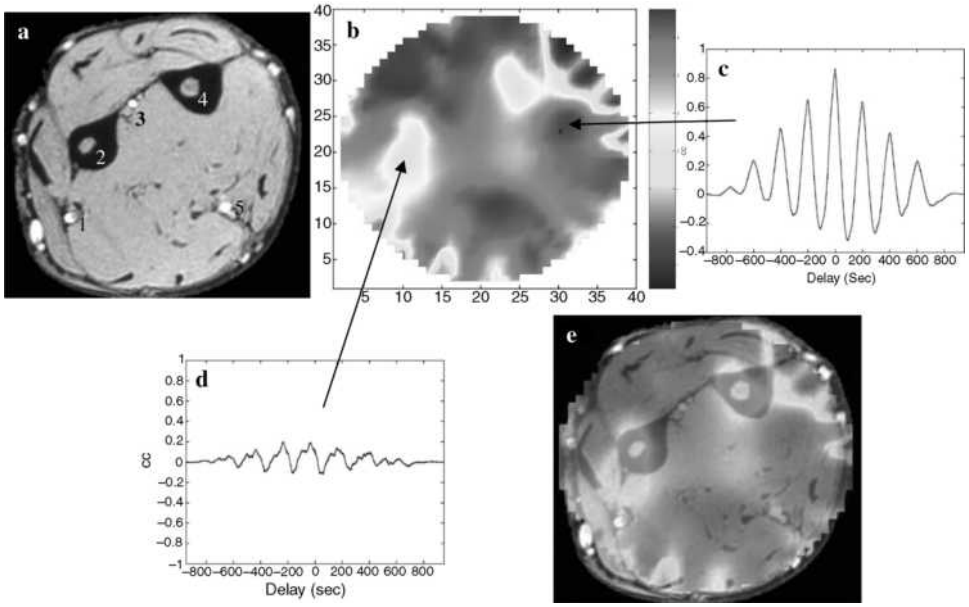


Figure 11. (a) MR cross-section of arm. (b) Cross-correlation (CC) map between (dark) model function in Fig. 10 with Hb image time series. (c) Overlay of a and b. (d and e), Time dependence of CC at indicated locations. (1) Radial artery, (2) radius, (3) interosseous artery, (4) ulna, (5) ulnar artery. [Reprinted with permission from R.L. Barbour et al., Functional imaging of vascular bed by dynamic optical tomography. Proceedings of SPIE, vol. 5369, in Physiology, Function, and Structure from Medical Images, 2004.]

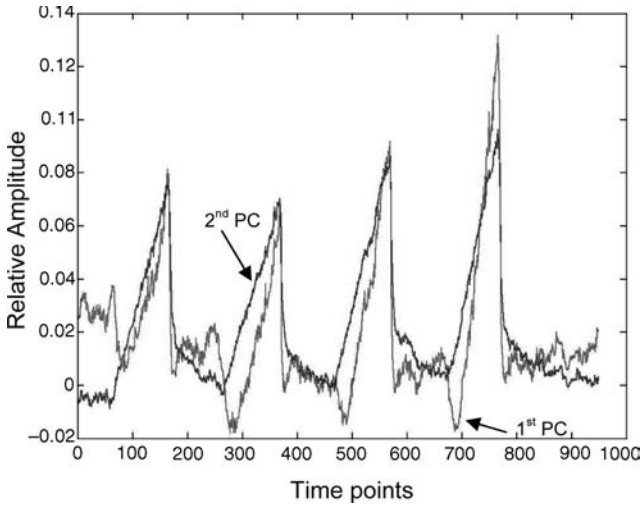


Figure 12. First (light curve) and second (dark curve) principal component (PC) of oxyhemoglobin signal computed for reconstructed image time series for four consecutive cuff inflation cycles. [Reprinted with permission from R.L. Barbour et al., Functional imaging of vascular bed by dynamic optical tomography. Proceedings of SPIE, vol. 5369, in Physiology, Function, and Structure from Medical Images, 2004.]

tree, as these distend most easily relative to the other elements of the vascular tree. The response seen in the second PC is more complex. Early on in the inflation cycle, a decline is observed, followed by an accelerated increase relative to the first PC, with proportionally greater responses upon subsequent challenges. It is believed this signal originates primarily from the microvascular bed. The initial decline in Hb_{oxy} is consistent with blood pooling, allowing for greater oxygen extraction. Following this, dilation occurs, perhaps in response to a buildup of local metabolic factors. The finding that the rate and extent of change in this signal increases in subsequent challenges suggests that insufficient time had elapsed between cycles to allow for complete washout of tissue factors. Support of this assignment is given in Fig. 13. In Fig. 13a it is seen that the signal associated with the first PC is mainly seen in the periphery (red regions), roughly in agreement with the location of near-surface veins. In contrast the signal associated with the second PC in Fig. 13b is found mainly in the ventral aspect of the forearm, which is dominated by well-perfused muscle.

Development of DOT technology is proceeding at a brisk pace, with new system designs being reported on a regular basis (91). As depicted here, TSI can be explored using time-series DOT. This, however, is not the only modality by which optical measures of peripheral vascular dynamics can be studied. Judging from advances made with photoacoustics (92), this approach may also prove useful (93,94). Additionally, the method could be extended further, to include use of injectable fluorescent probes or other forms of optical contrast media (95).

SUMMARY

Knowledge about the physiological and pathological state of the peripheral vasculature is extremely useful in the detection and staging of PVD, which occur with high prevalence and are associated with significant morbidity and mortality rates. Noninvasive measurements of those parameters are desirable because they can be performed more easily, are more cost effective, and are better suited for use as general screening tools than are invasive techniques. Volume-integrated dynamic flow measures can be obtained with a variety of available plethysmography methods. Measuring the flow response to occlusion maneuvers allows extraction of vessel flow and compliance values. In addition, plethysmography allows pulse volume recording in different locations on the extremities, and the interpretation of pulse wave contours and velocities permits drawing of inferences regarding peripheral, and even central, arterial health. Ultrasound methods, both structural and Doppler flow measurements, are valuable tools for diagnosing stenoses and thromboses. Laser Doppler flowmetry is a relatively new method for investigating the superficial microvessels.

Several imaging modalities are available for the assessment of blood flow and the vasculature. Most of these rely on contrast agents, and therefore can not strictly be considered noninvasive. The degree of invasiveness, however, seems small enough to justify their mentioning in the context of this article. Besides the conventional imaging methods, which include X ray CT, various MRI methods, and PET, the new noninvasive imaging technique of temporal spectral imaging, and in particular its

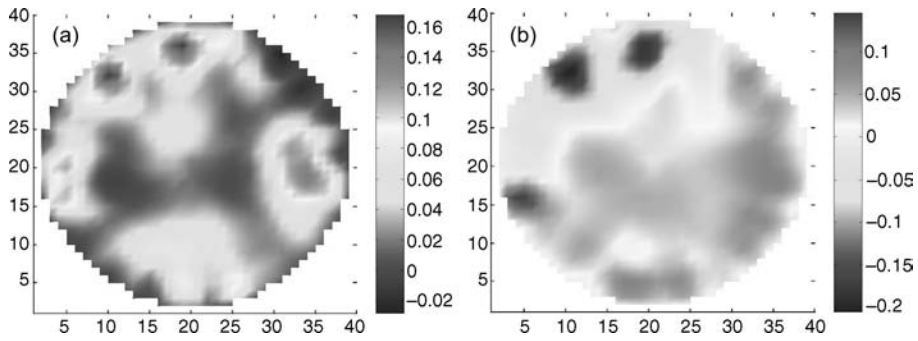


Figure 13. (a) Amplitude map of first principal component (PC) of total Hb (82% of total variance). (b) Amplitude map of second PC of total Hb (10% of total variance). [Reprinted with permission from R.L. Barbour et al., Functional imaging of vascular bed by dynamic optical tomography. Proceedings of SPIE, vol. 5369, in Physiology, Function, and Structure from Medical Images, 2004.]

implementation through diffuse optical tomography is introduced.

BIBLIOGRAPHY

Cited References

1. Drzewiecki G. Noninvasive arterial blood pressure and mechanics. Bronzino X, editor. The Biomedical Handbook. 2nd ed. Boca Raton, (FL): CRC Press and IEEE Press; 2000.
2. Peňáz J. Photoelectric measurement of blood pressure, volume, and flow in the finger. Digest 10th Internation Conference Medical Biological Engineering Dresden. 1973;104.
3. Wesseling KH, Settels JJ, De Wit B. The measurement of continuous finger arterial pressure non-invasively in stationary subjects. In: Schmidt TH, Dembroski TM, Bluemchen G, etitors. Biological and Psychological Factors in Cardiovascular Disease. Berlin: Springer-Verlag; 1986.
4. Parati G, et al. Non-invasive beat-to-beat blood pressure monitoring: new developments. Blood Press Mon 2003;8:31–36.
5. Bos WJW, et al. Reconstruction of brachial artery pressure from noninvasive finger pressure measurements. Circulation 1996;94:1870–1875.
6. Pressman GL, Newgard PM. A transducer for the continuous external measurement of arterial blood pressure. IEEE Trans Biomed Eng 1963;10:73–81.
7. Schäfer EA, Moore B. On the contractility and innervation of the spleen. J Physiol 1896;20:1–5.
8. Hewlett AW, Zwaluwenburg JG. The rate of blood flow in the arm. Heart 1909;1:87–97.
9. Whitney RJ. The measurement of volume changes in human limbs. J Physiol 1953;121:1–27.
10. McGivern RC, et al. Computer aided screening for deep venous thrombosis. Automedica 1991;13:239–244.
11. Schürmann M, et al. Assessment of the peripheral microcirculation using computer-assisted venous congestion plethysmography in post-traumatic complex regional pain syndrome type I. J Vasc Res 2001;38:453–461.
12. Leslie SJ, et al. Comparison of two plethysmography systems in assessment of forearm blood flow. J Appl Physiol 2004;96: 1794–1799.
13. Nyboer J, Bango S, Barnett A, Halsey RH. Radiocardiograms - the electrical impedance changes of the heart in relation to electrocardiograms and heart sounds. J Clin Invest 1940; 19:773.
14. Nyboer J. Regional pulse volume and perfusion flow measurements: Electrical impedance plethysmography. Arch Int Med 1960;105:264–276.
15. Nybor J. Electrical Impedance Plethysmography. 2nd ed, Springfield, (IL): Charles C Thomas; 1970.
16. Malmivuo J, Plonsey R. Bioelectromagnetism—Principles and Applications of Bioelectric and Biomagnetic Fields. New York: Oxford University Press; 1995.
17. Patterson R. Bioelectric impedance measurements. Bronzino X, editor. The Biomedical Handbook. 2nd ed, Boca Raton, (FL): CRC Press and IEEE Press; 2000.
18. Hertzman AB. Photoelectric plethysmography of the fingers and toes. Proc Soc Exp Biol Med 1937;37:529–534.
19. Hielscher AH, et al. Sagittal laser optical tomography for imaging of rheumatoid finger joints. Phys Med Biol 2004;49: 1147–1163.
20. De Trafford J, Lafferty K. What does photoplethysmography measure?. Med Biol Eng Comput 1984;22(5):479–480.
21. Jespersen LT, Pedersen OL. The quantitative aspect of photoplethysmography revised. Heart Vessels 1986;2:186–190.
22. Kamal AA, Harness JB, Irving G, Mearns AJ. Skin photoplethysmography — a review. Comput Methods Programs Biomed 1989;28(4):257–269.
23. Hashimoto J, et al. Pulse wave velocity and the second derivative of the finger photoplethysmogram in treated hypertensive patients: their relationship and associating factors. J Hypertens 2003;20(12):2415–2422.
24. Millasseau SC, et al. Noninvasive assessment of the digital volume pulse — Comparison with the peripheral pressure pulse. Hypertension W 2000;20(12):2415–2422.
25. Webb A. Introduction to Biomedical Imaging. Hoboken, (NJ): John Wiley & Sons Inc.; 2003.
26. Halpern JL. Evaluation of patients with peripheral vascular disease. Thrombosis Res 2002;106:V303–V311.
27. Hirsch AT, et al. Peripheral arterial disease detection, awareness, and treatment in primary care. JAMA 2001;286:317–1324.
28. Hooi JD, et al. Risk factors and cardiovascular diseases associated with asymptomatic peripheral arterial occlusive disease. The Limburg PAOD Study. Peripheral Arterial Occlusive Disease. Scand J Prim Health Care 1998;16(3): 177–182.
29. Weitz JI, et al. Diagnosis and treatment of chronic arterial insufficiency of the lower extremities: A critical review. Circulation 1996;94:3026–3049.
30. Carman TL, Fernandez BB. A primary care approach to the patient with claudication, Phys Am Fam 2000;61: 1027–1032. 1034.
31. Holland T. Utilizing the ankle brachial index in clinical practice. Ostomy Wound Manage 2002;48(1):38–40, 43–46, 48–49.
32. Sacks D, et al. Position statement on the use of the ankle brachial index in the evaluation of patients with peripheral vascular disease — A consensus statement developed by the standards division of the Society of Interventional Radiology. J Vasc Interv Radiol 2003;14:S389.
33. Donnelly R, Hinwood D, London NJM. ABC of arterial and venous disease: Non-invasive methods of arterial and venous assessment. studentBMJ 2000;8:259–302.
34. Vascular Technology Professional Performance Guideline — Lower Extremity Arterial Segmental Physiologic Evaluation, Society for Vascular Ultrasound; 2003.
35. Rietzschel E-R, et al. A comparison between systolic and diastolic pulse. Hypertension 2001;37:e15–e22.
36. Frank O. Die Grundform der Arteriellen Pulses. Acta Biol 1899;37:426–483.
37. Cohn JN, et al. Noninvasive pulse wave analysis for the early detection of vascular disease. Hypertension 1995;26:503–508.
38. Millasseau SC, Kelly RP, Ritter JM, Chowienczyk PJ. Determination of age-related increases in large artery stiffness by digital pulse contour analysis. Clin Sci 2002;103: 371–377.
39. Lekakis JP, et al. Arterial stiffness assessed by pulse wave analysis in essential hypertension: relation to 24-h blood pressure profile. Int J Cardiol 2005;102:391–395.
40. O'Rourke MF, Gallagher DE. Pulse wave analysis. J Hypertens Suppl 1996;14(5):S147–157.
41. Tsai W-C, et al. Association of risk factors with increased pulse wave velocity detected by a novel method using dual-channel photoplethysmography. Am J Hyper 2005;18:1118–1122.
42. Nigam A, Mitchell GF, Lambert J, Tardif J-C. Relation between conduit vessel stiffness (assessed by tonometry) and endothelial function (assessed by flow-mediated dilatation) in patients with and without coronary heart disease. Am J Cardiol 2003;92:395–399.

43. O'Rourke MF, Adji AA. An updated clinical primer on large artery mechanics: implications of pulse waveform analysis and arterial tonometry. Curr Opin Cardiol 2005;20(4): 275–81.
44. Gamble J, Gartside IB, Christ F. A reassessment of mercury in silastic strain gauge plethysmography for microvascular permeability assessment in man. J Physiol 1993;464: 407–422.
45. Maskell NA, et al. The use of automated strain gauge plethysmography in the diagnosis of deep vein thrombosis. Br J Radiol 2002;75:648–651.
46. Vigilance JE, Reid HL. Venodynamic and hemorheological variables in patients with diabetes mellitus. Arc Med Res 2005;36:490–495.
47. Goddard AJP, Chakraverty S, Wright J. Computer assisted strain-gauge plethysmography is a practical method of excluding deep venous thrombosis. Clin Radiol 2001;56: 30–34.
48. Larsson M, Nilsson H, Strömberg T. *In vivo* determination of local skin optical properties and photon path length by use of spatially resolved diffuse reflectance with applications in laser Doppler flowmetry. Appl Opt 2003;42(1):124–134.
49. Bonner R, Nossal R. Model for laser Doppler measurements of blood flow in tissue. Appl Opt 1981;20(20):2097–2107.
50. Serov A, Lasser T. High-speed laser Doppler perfusion imaging using an integrating CMOS image sensor. Opt Expr 2005;13(17):6416–6428.
51. Freccero C, et al. Laser Doppler perfusion monitoring of skin blood flow at different depths in finger and arm upon local heating. Microvasc Res 2003;66:183–189.
52. Eskandary H, Shahabi M, Asadi AR. Evaluation of carpal tunnel syndrome by laser Doppler flowmetry. Iran J Med Sci 2002;27(2):87–89.
53. Scharf A, et al. Evaluation of two-dimensional versus three-dimensional ultrasound in obstetric diagnostics: A prospective study. Fetal Diag Ther 2001;16:333–341.
54. Zotz RJ, Trabold T, Bock A, Kollmann C. *In vitro* measurement accuracy of three-dimensional ultrasound. Echocardiography 2001;18:149–56.
55. Hoksbergen AWJ, et al. Success rate of transcranial color-coded duplex ultrasonography in visualizing the basal cerebral arteries in vascular patients over 60 years of age. Stroke 1999;30:1450–1455.
56. Fürst G, et al. Reliability and validity of noninvasive imaging of internal carotid artery pseudo-occlusion. Stroke 1999;30: 1444–1449.
57. Rubin JM, et al. Power Doppler US: a potentially useful alternative to mean frequency-based color Doppler US. Radiology 1994;190:853–856.
58. Steinke W, et al. Power Doppler imaging of carotid artery stenosis: Comparison with color Doppler flow imaging and angiography. Stroke 1997;28:1981–1987.
59. Blomley MJK, et al. Liver microbubble transit time compared with histology and Child-Pugh score in diffuse liver disease: a cross sectional study. Gut 2003;52:1188–1193.
60. Della Martina A, Meyer-Wiethe K, Allémann E, Seidel G. Ultrasound contrast agents for brain perfusion imaging and ischemic stroke Therapy. J Neuroimaging 2005;15: 217–232.
61. Rofsky NM, Adelman MA. MR angiography in the evaluation of atherosclerotic peripheral vascular disease. Radiology 2000;214:325–338.
62. Baumgartner I, et al. Leg ischemia: Assessment with MR angiography and spectroscopy. Radiology 2005;234:833–841.
63. Kucharczyk W, et al. Intracranial lesions: flow-related enhancement on MR images using time-of-flight effects. Radiology 1986;161:767–772.
64. Whittemore AR, Bradley WG, Jinkins JR. Comparison of cocurrent and countercurrent flow-related enhancement in MR imaging. Radiology 1989;170:265–271.
65. Cebral JR, et al. Blood-flow models of the circle of Willis from magnetic resonance data. J Eng Math 2003;47:369–386.
66. Fatouraee N, Amini AA. Regularization of flow streamlines in multislice phase-contrast MR imaging. IEEE Trans Med Imag 2003;22:699–709.
67. Thomas DL, et al. The measurement of diffusion and perfusion in biological systems using magnetic resonance imaging. Phys Med Biol 2000;45:R97–R138.
68. Sorensen AG, et al. Hyperacute stroke: Simultaneous measurement of relative cerebral blood volume, relative cerebral blood flow, and mean tissue transit time. Radiology 1999;210:519–527.
69. Wang J, et al. Arterial transit time imaging with flow encoding arterial spin tagging (FEAST). Mag Reson Med 2003;50:599–607.
70. Mandeville JB, Marota JJA. Vascular filters of functional MRI: Spatial localization using BOLD and CBV contrast. Mag Reson Med 1999;42:591–598.
71. Ogawa S, Menon RS, Kim S-G, Ugurbil K. On the characteristics of functional magnetic resonance imaging of the brain. Annu Rev Biophys Biomol Struct 1998;27:447–474.
72. Lerman LO, Rodriguez-Porcel M, Romero JC. The development of x-ray imaging to study kidney function. Kidney Inte 1999;55:400–416.
73. Romero JC, Lilach LO. Novel noninvasive techniques for studying renal function in man. Semi Nephrol 2000;20:456–462.
74. Gobbel GT, Cann CE, Fike JR. Measurement of regional cerebral blood flow using ultrafast computed tomography: Theoretical aspects. Stroke 1991;22:768–771.
75. Horn P, et al. Xenon-induced flow activation in patients with cerebral insult who undergo xenon-enhanced CT blood flow studies. AJNR Am J Neuroradiol 2001;22:1543–1549.
76. Matsuda M, et al. Comparative study of regional cerebral blood flow values measured by Xe CT and Xe SPECT. Acta Neurolog Scand 1996;166:13–16.
77. Yonas H, et al. Side effects of xenon inhalation. J Comput Assist Tomogr 1981;5:591–592.
78. Roberts HC, et al. Multisection dynamic CT perfusion for acute cerebral ischemia: The "toggling-table" technique. AJNR Am J Neuroradiol 2001;22:1077–1080.
79. Röther J, et al. Hemodynamic assessment of acute stroke using dynamic single-slice computed tomographic perfusion imaging. Arch Neurol 2000;57:1161–1166.
80. Compare, for example, the corporate web pages http://www.anzai-med.co.jp/eigo/Xe-Per.htm and http://www.ge-healthcare.com/usen/ct/case_studies/products/perfusion.html, which assert superiority for the Xe- and I-based approaches, respectively. One apparent reason for disagrement is the issue of the relative contributions to the image of the blood vessels' walls and of their contents.
81. Bambakidis NC, et al. Functional evaluation of arteriovenous malformations. Neurosurg Focus 2001;11: Article 2.
82. Schaefer WM, et al. Comparison of microsphere-equivalent blood flow (^{15}O-water PET) and relative perfusion (^{99m}Tc-tetrofosmin SPECT) in myocardium showing metabolism-perfusion mismatch. J Nucl Med 2003;44:33–39.
83. Barbour RL, et al. Functional imaging of the vascular bed by dynamic optical tomography. Proc SPIE 2004;5369: 132–149.
84. Barbour RL, Lubowsky J, Graber HL. Use of reflectance spectrophotometry (RS) as a possible 3-dimensional (3D) spectroscopic imaging technique. FASEB J 1988;2:A 1772.

85. Barbour RL, Graber H, Lubowsky J, Aronson R. Monte Carlo modeling of photon transport in tissue (PTT) [I. Significance of source-detector configuration; II. Effects of absorption on 3-D distribution (3DD) of photon paths; III. Calculation of flux through a collimated point detector (CPD); IV. Calculation of 3-D spatial contribution to detector response (DR); V. Model for 3-D optical imaging of tissue]. Biophy J 1990;57:381a–382a.
86. Grünbaum FA. Tomography with diffusion. Inverse methods in Action In: Sabatier PC, Editor. (Proceedings of 1989 Multicennials Meeting on Inverse Problems. New York: Springer-Verlag; 1990; 16–21.
87. Barbour RL, Lubowsky J, Aronson R. Method of Imaging a Random Medium, US pat. 5,137,355; awarded 8/11/92.
88. Wilson BC, Sevick EM, Patterson MS, Chance B. Time-dependent optical spectroscopy and imaging for biomedical applications. Proc IEEE 1992;80:918–930.
89. Barbour RL, et al. Temporal Imaging of Vascular Reactivity by Optical Tomography. In: Gandjbakhche AH, editor. Proceedings of Inter–Institute Workshop on *In Vivo* Optical Imaging at the NIH. (Optical Society of America, Washington, (DC); 1999), pp. 161–166.
90. Barbour RL, et al. Optical tomographic imaging of dynamic features of dense–scattering media. J Opt Soc Am A 2001;18: 3018–3036.
91. Schmitz CH, et al. Instrumentation for fast functional optical tomography. Rev Sci Instr 2002;73:429–439.
92. Wang X, et al. Noninvasive laser-induced photoacoustic tomography for structural and functional in vivo imaging of the brain. Nature Biotechnol 2003;21:803–806.
93. Kruger RA, Stantz KM, Kiser WL. Thermoacoustic CT of the Breast. Proc SPIE 2002;4682:521–525.
94. Ku G, Wang L-H. Deeply penetrating photoacoustic tomography in biological tissues enhanced with an optical contrast agent. Opt Lett 2005;30:507–509.
95. Weissleder R, Ntziachristos V. Shedding light onto live molecular targets. Nature Med 2003;9(1):123–128.

See also CUTANEOUS BLOOD FLOW, DOPPLER MEASUREMENT OF; IMPEDANCE PLETHYSMOGRAPHY.

PET SCAN. See POSITRON EMISSION TOMOGRAPHY.

PHANTOM MATERIALS IN RADIOLOGY

YOICHI WATANABE
C. CONSTANTINOU
Columbia University
New York, New York

INTRODUCTION

What is Phantom?

Radiation has both particle and electromagnetic wave nature. Radiation carries energy; hence, upon striking the human body, radiation deposits the energy in the human tissue. This interaction consequently can damage the tissue by causing strand breaks in genetic molecules called deoxyribonucleic acid (DNA) in nucleus of living cells. Such damages are considered a major cause of cancers.

Radiation, such as X rays, (γ rays, electrons (or β-particles), helium ions (or α-particles), and neutrons were discovered in late nineteenth century to early twentieth century. Since that time scientists and engineers invented and developed beneficial applications of radiation by taking advantages of the penetrating and damaging power of the radiation. Medical uses are the most noticeable applications of radiation. Radiation is used to diagnose and cure human illness.

Radiologists use X rays and other particulate radiation in hospitals and clinics to diagnose disease through imaging diseased sites. Radiation oncologists use X rays, electrons, and other forms of radiation available in radiation oncology centers to cure cancer.

While radiation is useful, careless use of radiation can lead to harmful effects on the health of people. Therefore, it is quite important to carefully evaluate the distribution of radiation energy absorbed by human tissue (or dose) during the radiological procedures. If the potential radiation damage is not well understood, clinical uses of new radiation sources without careful and thorough evaluation must be avoided. Placement of radiation measurement instrumentation in the human body is not easy, thus hampering precise dose measurements. Therefore, radiation scientists developed simulated human bodies or organs, herein called phantoms, to evaluate actual radiation doses. The phantoms are used to estimate radiation dose and transmission (or attenuation) of radiation in the human body for radiological studies.

Phantom materials for radiology should mimic the radiological characteristics of tissues. The homogeneity of radiological characteristics over the phantom is very important. Often the shape of the phantom should mimic the shape of a human body or a part of the body. Hence, the material should be easily made into various shapes and it should be easy to machine the material. The materials should maintain the mechanical integrity and the radiological characteristics for a long time.

Historical Background

To simulate radiation transport processes in human body, scientists developed phantoms made of tissue mimicking materials. The phantom should be made of material that absorbs and scatter radiation in the same way as the real tissue. Spires showed that the phantom material should have the same density as tissue and contain the same number of electrons per gram (1).

Water was the first material to be used as a tissue substitute in radiation measurements by Kienbock (2). Baumeister introduced wax in 1923 (3). The first formulated solid, called Siemen's Wax, composed of paraffin wax and magnesium oxide as a corrective filter, was reported by Ott in 1937 (4). Several similar wax-based products were subsequently introduced in Europe and North America, including MixD (5), Harris Wax (6), and M3 (7). Many phantoms comprised of either simple stacked sheets or more complex body-like structures have been constructed from these latter materials.

Plastics and rubbers have found an increasing application in the specialty of tissue simulation. From the polyethylene-based Markite (8) stemmed the conducting

plastics of Shonka et al. (9) and polyurethane systems of Griffith et al. (10). The last three products have been used in the manufacture of elaborate anthropomorphic body phantoms with airways, simulated lungs, and embedded skeletons. An important elementally equivalent liquid system was introduced by Rossi and Failla (8). A mixture of water, glycerol, urea, and sucrose was used to match an approximate formula for soft tissue. This mixture was simplified by Goodman (11) and extended to more complex elemental formulas (12). Of the tissue substitutes introduced before 1970, only a handful had radiation absorption and scattering characteristics within $\pm$ 5% of those of the corresponding real tissues over extended energy range, and these included most of the above-mentioned phantom materials. The most important of them was water. Fortunately, it was readily available and cheap. An extensive program of research and development was initiated at St. Bartholomew's Hospital in London in 1970. Over 160 tissue substitutes were formulated, simulating a wide range of body tissues. Liquid, gel, solid, and particulate systems were produced for use with photon and particulate radiations (13–16). Other groups also developed tissue equivalent materials. Herman et al. used polyethylene to develop water-equivalent material, as well as fat and muscle materials in 1980s (17–19). Homolka et al. used polymer powders together with suitable additives to adjust photon attenuation (20,21). They created adipose, muscle, bone, and water equivalent materials, which simulate radiological characteristics of tissues for low energy photons, that is, energy $<$ 100 keV for diagnostic X rays. Latest work includes development of tissue equivalent materials for pediatric radiology by Bloch et al. (22). Suess et al. manufactured a phantom material based on polyurethane resin for low contrast resolution measurements of computed tomography (CT) scanners (23). Iwashita used polyurethane resin mixed with $CaCO_3$ to create cortical and cancellous bones (24). Burmeister et al. made brain tissue equivalent conducting plastic for low-energy neutron applications (25).

Physics Background

Medical Radiation. Radiologists and radiation oncologists use radiation in several forms for diagnostics and therapy. The most common radiation is photons, which can be produced by X-ray generators and linear accelerators or are emitted by radioactive source. The photon energy used for medical applications ranges from 10 keV to 20 MeV. Electrons are another common form of radiation for medical uses. Positively charged electron or positrons are used for diagnostic purpose with positron emission tomography (PET) scanners. Heavier particles are also employed for therapeutic radiology. Protons, alpha particles, pi-mesons, neutrons, and heavy charged particles, such as carbonsions were used in the past or are being introduced into clinic.

Interaction of Radiation with Matter. Photons interact with matter in three main physical processes: photoelectric absorption, Compton scattering, and pair production. Depending on the photon energy, one of three interactions play major role. Electrons in the energy range of interest collide with electrons in atoms–molecules of the material. Electrostatic force is the major interaction mechanism. Protons and heavy charged particles go through electrodynamic interactions similar to electrons. Neutrons do not carry electric charge; hence, those interact mostly with the nucleus of atoms.

Photon interaction probability is represented by the attenuation coefficient, which is the loss rate of photon particles per unit length from the original photon flight path. Electron scattering is quantified by stopping power, which represents electron energy loss rate per unit path. Energy absorption of radiation in tissue is considered per unit mass of tissue. The unit of radiation dose is joules per kilogram (J/kg) or gray (Gy). Mass attenuation coefficient and mass stopping power are often used to describe the effectiveness of material to attenuate photons and electrons.

Radiological Equivalence of Material to Tissue. Ideally, a phantom material should have the same mass attenuation coefficient for photons and the same mass stopping power for electrons as the tissue it simulates. If the phantom can have the same atomic composition as tissue, those parameters of the phantom are the same as those of tissue. However, making the atomic composition of phantom exactly the same as tissue is not easily achievable. Hence, as a guideline of tissue equivalent material, physicists expect that the material has a similar mass density, effective atomic number, and electron density as the real tissue. The reasoning for this approach is the following. As mentioned before, photons interact with matter through three physical mechanisms. The magnitude of the photoelectric interaction is approximately proportional to a certain power of the atomic number of the atom. The concept of the effective atomic number is introduced to present how close a material is to another material for photons in the energy range in which the photoelectric effect dominates as the main interaction process. Such energy range is generally $<$ 100 keV. The Compton interaction and pair-production are essentially proportional to the number of electrons in the material. Electron stopping power is also proportional to the number of electrons since electrons directly interact with electrons in the material. Hence, the electron density of the material is another important parameter to represent the radiological characteristics of each material.

Outline

There is concern about the materials used to manufacture phantoms for radiology applications. The next section gives extensive discussion on the materials mainly developed by White, Constantinou, and there co-workers. More detailed discussion can be found in an ICRU report (26) and relevant references by those authors. The third section presents how those materials are used for radiation dosimetry, radiation therapy, and diagnostic radiology. The discussion on applications will be limited to photons and electrons, since most medical applications utilize those

particles. The last section is devoted to speculative discussion on what types of phantom material will be developed in the near future.

PHANTOM MATERIALS: SIMULATED TISSUES AND CRITICAL TISSUE ELEMENTS

The tissue substitutes produced before 1970 were designed to simulate predominantly muscle, bone, lung, and fat. The sources of reliable data on elemental composition and mass densities of real tissues were limited. The main sources included the reports of Woodard (27), giving the elemental composition of cortical bone, and a report by the International Commission on Radiological Units and Measurements (28), giving the elemental composition of striated muscle and compact bone (femur). Unfortunately, there was a disagreement between the above sources on the composition of bone, which made bone simulation work more difficult.

The publication of the *Reference Man* data by the International Commission on Radiological Protection (ICRP) (29) in 1975 and the improvement in available equipment and technology enabled the formulation of new tissue substitutes for 15 different tissues that are described here. The *Reference Man* publication included tabulations of the concentrations of 51 elements in 81 organs, tissues, and tissue components. It also included the mass densities and the ratio of water/fat/protein contents in each one of them. Based on the above information, White and Constantinou developed substitutes for the following categories of tissues and body organs:

1. Principal soft tissues, namely, muscle, blood, adipose tissue, and skin. Adipose tissues are defined by ICRP as composed of 70% fat, 15% water, and 15% protein by mass.
2. Principal skeletal tissues, namely, cortical bone, inner bone, and red marrow. Yellow marrow is very close to adipose tissue and was not included. The formula given by Woodard (27) was adopted as more correct. This reference gives 55.8% Ca plus P, 12.5% water, 25.2% protein, and 6.5% sugar by mass for cortical bone.
3. Body organs, namely brain, kidneys, liver, lung, and thyroid. The elemental data for these organs were obtained from the ICRP Reference Man document.
4. Average tissues, which included average breast, total soft tissue, and total skeleton. The latter two formulas were derived from the ICRP source, while the formulas for average breast were based on 50% fat and 50% water by mass (13). Other formulas for average breast described in the literature (14,30) are based on 25% fat-75% muscle, 50% fat-50% muscle, and 75% fat - 25% muscle, referring to young, middle-aged, and older breast, respectively.

The percentage by mass for the elements H, C, N, O, Na, Mg, P, S, Cl, K, and Ca in each real tissue is given in Table 1 with information on mass densities and additional elements when appropriate. New tissue substitutes presented in this section were formulated so that, whenever possible, they have exactly the same elemental composition and mass density as the corresponding real tissue. In most of the solid substitutes where epoxy resin

Table 1. Elemental Compositions of the Principal Organs and Tissues, healthy adult[a]

	Elemental Composition (percentage by mass)												
Tissue	H	C	N	O	Na	Mg	P	S	Cl	K	Ca	Other Elements	Mass Density, $kg \cdot m^{-3}$
Principal soft tissues													
Adipose tissue	11.2	51.7	1.3	35.5	0.1			0.1	0.1				970
Blood	10.2	11.0	3.3	74.5	0.1		0.1	0.2	0.3	0.2		Fe(0.1)	1060
Muscle	10.2	14.3	3.4	71.0	0.1		0.2	0.3	0.1	0.4			1050
Skin	10.0	20.4	4.2	64.5	0.2		0.1	0.2	0.3	0.1			1090
Principal skeletal tissues													
Cortical bone	3.4	15.5	4.2	43.5	0.1	0.2	10.3	0.3			22.5		1920
Inner bone (Spongiosa)	8.5	40.4	2.8	36.7	0.1	0.1	3.4	0.2	0.2	0.1	7.4	Fe(0.1)	1180
Red marrow	10.5	41.4	3.4	43.9			0.1	0.2	0.2	0.2		Fe(0.1)	1030
Body organs													
Brain	10.7	14.5	2.2	71.2	0.2		0.4	0.2	0.3	0.3			1040
Kidney	10.3	13.2	3.0	72.4	0.2		0.2	0.2	0.2	0.2	0.1		1050
Liver	10.2	13.9	3.0	71.6	0.2		0.3	0.3	0.2	0.3			1040
Lung	10.3	10.5	3.1	74.9	0.2		0.2	0.3	0.3	0.2			260
Thyroid	10.4	11.9	2.4	74.5	0.2		0.1	0.1	0.2	0.1		I(0.1)	1050
Average tissues													
Breast (whole)	11.5	38.7		49.8									960
Average soft tissue (male)	10.5	25.6	2.7	60.2	0.1		0.2	0.3	0.2	0.2			1030
Skelton (sacrum) (whole)	7.4	30.2	3.7	43.8		0.1	4.5	0.2	0.1	0.1	9.8	Fe(0.1)	1290

[a]See Ref. 31

systems, acrylics, or polyethylene were used as major components, a partial replacement of oxygen by carbon and vice versa had to be accepted. For this reason, an effort was made to determine which of the elements present in various tissues play a critical role in the energy deposition process when interacting with various radiation modalities. During this evaluation, basic interaction data have been calculated for photons and electrons from 10 keV up to 100 MeV, protons from 1 up to 1000 MeV, and neutrons from 100 eV up to 30 MeV. Detailed accounts of the computations are given in the literature (13,14,32) and only a summary is given here.

When a photon beam interacts with a tissue, photon energy absorption scattering depends primarily on the atomic number Z of the constituents and the electrons/kilogram of the tissue. Since hydrogen has double the electron density of other elements, hydrogen and the high Z constituents of a tissue are the critical elements. Consequently, their percentage by mass in the substitute must match that of a real tissue as accurately as possible. In order to evaluate the accuracy with which a substitute material simulated the corresponding real tissue, the mass attenuation coefficients (μ/ρ) and energy absorption coefficients (μ_{en}/ρ) were calculated at 33 energy points between 10 keV and 100 MeV, using the mixture rule:

$$\mu/\rho = \sum_i w_i(\mu/\rho)_i$$

where w_i is the proportion by mass of the ith element having a coefficient $(\mu/\rho)_i$.

The irradiation of tissues with beams of charged particles, such as electrons and protons, leads to energy deposition through collisional and radiative processes. Collisional interactions of incident particles with the electrons of the target material are the major cause of energy loss for electrons < 500 keV and protons < 1000 MeV. Radiative (bremsstrahlung) losses become important for higher energies. In order to evaluate the new tissue substitutes for electron and proton interactions, the collision stopping powers $(s/\rho)_{coll}$ and the radiative stopping powers $(s/\rho)_{rad}$ were calculated for both substitutes and the corresponding real tissues, and comparison was made between the total stopping powers $(s/\rho)_{tot}$ of the substitutes and those of the corresponding real tissues. A phantom material was accepted as a useful substitute only if its radiation characteristics were within $\pm$ 5% of those of the real tissue that it was designed to simulate.

In the case of tissues being irradiated with neutrons (10 eV–50 MeV), hydrogen was found to be the most critical element for all energies. Nitrogen was found to be the second most important element of neutron energies < 5 MeV. This is due to the significance of the elastic scattering of neutrons with hydrogen nuclei at higher neutron energies $^1H(n, n)^1H$ and the contribution of the capture process $^1H(n, \gamma)^2H$ and $^{14}N(n, p)^{14}O$ at the low and thermal neutron energies. Oxygen and carbon play a great role than nitrogen above 10 MeV. For neutron energies up to 14 MeV, the interactions with hydrogen account for 70–90% of the total dose in soft tissue (33,34). In view of the above finding, efforts were made to match the hydrogen, oxygen, carbon, and nitrogen contents of all the substitutes to those of the real tissues as accurately as possible.

The relative proportion of carbon and oxygen in tissue was found to be less critical in the attenuation and energy absorption from neutrons and high energy protons. Trace elements, with concentrations of $< 0.5\%$ by mass, were found to play no significant role in the absorption of energy from fast neutrons, high energy protons, and X rays > 100 keV. Detailed calculation and tabulation of the above-mentioned radiation interaction quantities for all the real tissues and their substitute material are available in the literature (13,14,30,35).

FORMULATION PROCEDURES

Three main methods were applied in the formulation of the tissue substitutes described in this article, namely, the elemental equivalence method, the basic data method, and the effective atomic number method. The following criteria formed the basis of the tissue simulation studies.

Criteria for Tissue Equivalence

Two materials will absorb and scatter any type of radiation in the same way, only if the following quantities are identical between them: (*1*) photon mass attenuation and mass absorption coefficients, (*2*) electron mass stopping powers and mass angular scattering powers, (*3*) mass stopping powers for heavy charged particles and heavy ions, (*4*) neutron interaction cross-sections and kerma factors, and (*5*) the mass densities of the two materials must be the same. A brief description of the formulation methods is now presented.

Method of Elemental Equivalence

Based on the above criteria, it is obvious that only material with the same elemental constituents and in the same proportion by mass as the corresponding real tissue can be termed tissue equivalent for all radiation modalities. A number of such materials were formulated, particularly in the liquid and gel phase (14,15,30,36). If a substitute is elementally correct and has the correct bulk density, the only source of error in the absorbed dose calculations from measurements in the phantom material will be phase differences due to differences in chemical binding. Such errors are difficult to evaluate because of lack of extensive data, but they have been found small and rather insignificant in conventional radiation dosimetry.

The method of elemental equivalence was first applied by Rossi and Failla (8) who tried to reproduce an approximate formula for soft tissue $(C_5H_{40}O_{16}N)_n$. They formulated a mixture of water–glycerol–urea and sucrose, which had the formula $C_5H_{37.6}O_{18}N_{0.97}$, but their publication did not explain how they arrived at their formulation. Frigerio et al. (37) used water as base material and then selected compound that could be dissolved in it in such proportions as to satisfy the CHNO molar ratio. They considered each compound as the sum of two components one of which

was water; for example, glycerol $C_3H_2(H_2O)_3$ can be written as $C_3H_8O_3$. Using this approach, they produced a liquid system with elemental composition almost identical to that of muscle tissue.

The method of elemental equivalence was applied later with minor modifications (14,30), and as a result > 35 tissue equivalent liquids and gels were formulated. The following constraints were used during this work.

1. Once a base material was selected, the additives should be chosen from a library of compounds that are neither toxic nor corrosive, explosive, volatile, or carcinogenic.
2. The number of components should be kept to a minimum.
3. The proportion by mass of each constituents of a tissue substitute should be within 0.5% of that of the real tissue, except for hydrogen for which the agreement should be within 0.1%.

The basic steps followed in the formulation of elementally correct tissue substitutes have been reported in the references cited.

Basic Data Method

The second most accurate method of formulating tissue substitutes is the basic data method, which matches basic interaction data, for example, mass attenuation coefficients for photoelectric and Compton scattering, and mass stopping powers of the tissue substitutes to those for the body tissue over the required energy interval. This method was used by White (13,38) to formulate a large group of solid and liquid tissue substitutes for use with photons and electrons. The mathematical procedures developed enable two-component tissue substitutes (base material + filter) to be formulated for a given base material, with the most appropriate filler being selected from a library of compounds. Any degree of matching accuracy (e.g., 1% between μ/ρ values) can be specified. Recently, Homolka et al. (21) developed a computer program, which minimizes the difference between the linear attenuation coefficients of a phantom material and a tissue by considering the energy dependence of the attenuation coefficient. The program optimizes the components of base materials, such as polystyrene, polypropylene, and high density polyethylene together with admixtures of TiO_2, MgO, $CaCO_2$, and graphite. They showed that the measured Hounsfield number of the water equivalent phantom material agrees with those of water within eight Hounsfield units for X ray energy from 80 to 140 kV.

Effective Atomic Number Method

An indirect method of simulation is based on effective atomic number, $\hat{Z}$, which may be used to characterize a partial mass attenuation coefficient (τ/ρ, σ_e/ρ, κ/ρ, etc.) for a given group of elements and specified photon energy. The fundamental assumption for this method is that materials with the same value of the product of electron density and Z^x, where x is the Z exponent derived for a given partial interaction process, as a reference material shows the same photon and electron interaction characteristics as the reference material. A formulation technique similar to the basic data method can be derived, that is, the selection of an appropriate filler for a specific base material and the establishment of the relative proportions to achieve a specified degree of matching accuracy of the electron density and the effective atomic number between two materials. Accounts of this method were given by White et al. (39) and Geske (40).

MATERIALS AND METHOD OF MANUFACTURE OF THE NEW TISSUE SUBSTITUTES

Phantom materials currently in use can be grouped in-to four types depending on the base material. White et al. mainly developed epoxy-resin-based material (13,14,16, 38,41,42). Hermann et al. used polyethylene-based technique. [17]Homolka's group made phantoms based on fine-polymer powers, such as polyethylene, polypropylene, polystyrene or polyurethane (20). Suess, Iwashita and others used polyurethane resin (23,24). Since one of the current authors is very familiar with the epoxy-resin-based method and other methods use similar manufacturing techniques except the base material, more space is devoted to discussing the epoxy-resin-based phantom in this section than other methods.

Epoxy Resin-Based Method

Materials. The base materials used by White and Constantinou for the manufacturing of solid-phantom materials included four epoxy resin systems designated CB1, CB2, CB3, and CB4, respectively. The epoxy resin systems consist of a viscous resin and a lower viscosity liquid hardener (Diluents). The two are mixed in such proportions by mass as determined by the chemical reaction occurring during the curing process. The constitutes and elemental compositions of the epoxy-resin systems used in the manufacture of the new tissue substitutes were described in detail (14,38). These resin systems are rich in hydrogen (7.9–11.3% by mass) and nitrogen (1.60–65.62% by mass), but they are rather low in oxygen (13.15–20.57% by mass), compared to what is needed to match the oxygen content of the real tissue. As a result, in most solid substitutes, part of the oxygen needed was replaced by carbon, but an effort was made to have the sum of (C+O) in the substitute equal to that in the real tissue. Following the addition of the necessary powdered filler, low density ($\sim 200\ kg{\cdot}m^3$) phenolic microspheres (PMS) are also added in small precalculated quantities to make the bulk density of the mixture match that of real tissues. In the case of lung substitutes, the addition of a foaming agent (DC1107) in quantities of 1% by mass or less leads to sample with bulk densities as low as $200\ kg{\cdot}m^{-3}$ (43).

In the case of liquid substitutes, water was selected as the base material because it is an important component of real tissues and it is readily available. Various organic and inorganic compounds can be dissolved in it, in proportions necessary to satisfy the requirements for both the main elements C, H, N, O and the trace elements, such as Na, Mg, P, S, Co, K, and Ca.

The use of gelatin facilitated the formulation of many gel substitutes useful for short-term applications. For the production of elementally equivalent material, gelatin is preferred to other gelling agents such as agar (37) and Laponite used in the production of thyrotrophic gels (44), because it has an elemental composition very close to that of protein. Since real tissues are composed of varying proportions of water, carbohydrates, protein, and fat, it is easier to formulate elementally correct gel substitutes with it. By adding trace quantities of bacteriostatic agent (e.g., sodium azide) and sealing them into polyethylene base, gels can be preserved for longer periods.

Mixing Procedures. The manufacture of a solid substitute starts by adding first the appropriate quantity of the resin into a Pyrex reaction vessel followed by the lower viscosity hardener–diluent. The powder fillers are then added in order of decreasing mass density. Following a short manual mix, a ground glass lid is attached to the reaction vessel. This lid has one central and two peripheral glands (openings). A twin-bladed rotor is passed through the central gland and connected to an electric stirrer. One of the peripheral gland openings is connected to a vacuum pump while the third is used to control the air pressure inside the mixing vessel. During mixing, the system is evacuated to approximately 1.3 Pa (10^{-2} mmHg). The trapped air escapes as the rotor blades break the resulting foam. After $\sim$ 20 min of stirring under reduced pressure, a homogeneous, air-free mix is obtained. The vacuum is then released and the mix is carefully poured into waxed metal, silicon, rubber, or Teflon molds. A more detailed description is found in the references listed above.

When mixing lung substitutes, no vacuum is applied. The components are mixed thoroughly under atmospheric pressure and then a liquid foaming agent is added (activator DC1107) and quickly stirred into the mix. The foaming action starts in $\sim$ 30 s and the mixture must be poured in the mold to foam to the required mass density. The resulting bulk density depends on the volume of the activator added. For example, 170 kg mass of foaming agent will result in a lung substitute with a density of $\sim$ 250 kg·m^{-3}.

The manufacture of water-based liquid and gel substitutes is relatively easy. The required quantity of distilled water is used and the inorganic compound necessary for the introduction of trace elements are stirred into solution one by one, ensuring that each is completely dissolved, before adding the next, thus avoiding the formation of intermediate precipitates. Urea, commonly used to satisfy the nitrogen requirements, is dissolved next, followed by any other organic liquid components. If a gel substitute is required, the water with the dissolved trace elements is heated up to $\sim$ 80 °C before adding and dissolving the necessary quantity of gelatin. Once a clear uniform solution is obtained, it is left to cool to almost room temperature before the remaining components and the bacteriostatic agent are added. The mixture is usually added into polyethylene bags, heat sealed to inhibit water loss, and left to gel before use.

Polyethylene-Based Method

This method uses polyethylene powder and inorganic admixtures, $CaCO_3$ and MgO, in powdered form (17,18). The polyethylene powder has a melting point of 105 °C and density of 0.917 g·cm^{-3}. A processing temperature is 200–240 °C. Dry mixing of the polyethylene powder and inorganic powder is performed in a long Plexiglas drum rotated on a lath, with internal Plexiglas shelves providing mixing. The mixture was then poured on iron plates that carried quadratic iron frames. Plastic plates are formed during melting at 180 °C. Homogeneous and smooth plastics are obtained with inorganic admixtures of up to 10 % of the total mass. Machining is easy to make different thickness of plates. For making thinner foils, a milling machine was provided with a vacuum fixing device.

Polymer Powder-Based Method

Homolka and Nowotny discusses the manufacturing technique of polymer powder-based phantom in a publication (20). The base materials for this method are polymer powders made of polyethylene (PE), polypropylene (PP), polystyrene (PS), and polyurethane (PU). All powders are particles of sizes much < 100 μm. Typical additives were $CaCO_3$, MgO, TiO_2, calcium hydroxyapatite (bone mineral), and high purity graphite. These additives are available in a suitable grain size < 100 μm. A base material is mixed with additives using a ball mill. The material then is sintered in an evacuated vessel at temperature above the melting point of the polymers. The melting temperature (softening temperature) of PE, PP, and PS are 107, 165, and 88 °C, respectively. To remove any air or other gases, a pressure of $\sim$ 1 Pa was applied during the sintering process.

Polyurethane-Based Method

Polyurethane consists of a chain of organic units joined by urethane links. It can be made in a variety of textures and hardness by varying the particular monomers and adding other substances. It is most commonly used to produce foam rubber. Suess and Kalendar manufactured tissue equivalent phantom materials using low density polyurethane-resin (23). The resin has a high viscosity, leading to a homogeneous mixing of fillers. The stirring of the resin, hardening, and additives are done under vacuum conditions. Air bubbles are removed at pressures < 100 Pa. The ingredients are dehumidified since different degrees of humidity interfere with the polymerization and causes variations in the cured resin density. The temperature of the base materials are maintained at 20 °C before mixing and at 40 °C during curing. The density of the material is modified by adding small amounts of low density phenolic microspheres and high density poly(tetrafluorethylene) powder.

Quality Control

Quality control is necessary in order to maintain the quality of the manufactured substitutes. Two simple and effective types of investigations are usually performed, namely, mass density measurements and radiographic imaging. Casing or machining rigid solids into cubes or cylinders and measuring their mass and volume directly

provides mass density data with an error of ± 0.5%. Density bottles are useful for mass density determinations of liquids and gels.

The use of X rays in the 20–50 keV energy range and computed tomography scans are simple and sensitive methods for homogeneity test on the tissue substitutes. With radiographic techniques, the smallest detectable size of high atomic number particulate fillers or trapped air pockets are ~ 100 μm. The high contract resolution of CT scanners is limited to ~ 0.6 mm, but the ability of the scanners to show low contrast differences can help in detecting unacceptably nonuniform macroscopic areas in the samples. Optical transmission microscopy of thin sample scan offers more sophisticated uniformity testing if a high degree of homogeneity is required

The uniformity of the manufactured solid substitutes may be tested by mass density determinations, multiple slice CT scanning, and conventional radiographic techniques as discussed above. The mass densities at different point in a well made sample were found to be within ±0.5% of the average value, except for lung substitutes, which show a density variation of up to ±3% of the average value.

CLASSIFICATION AND TESTING OF THE NEW TISSUE SUBSTITUTES

The available tissue substitutes were classified according to the magnitude of the discrepancy between their radiation characteristics and the radiation characteristics of the corresponding real tissues. A muscle substitute, for example, with mass attenuation and mass energy absorption coefficients within 5% of the same coefficient for real muscle, is considered as Class A substitute for photon interactions. If the discrepancy is between 5 and 20%, the substitute is called Class B material and if the error exceeds 20%, the substitute is termed Class C. In addition, material with discrepancy within 1% are called tissue equivalent and classified as A*. A tissue substitute that is not elementally correct could be Class A for one radiation modality, but may be Class B or even Class C for another. Table 2 shows some of the recommended tissue substitutes and their components by mass, while Table 3 shows their classification for photon, electron, proton, and neutron interactions. The best results are obtained with Class A* materials, which are elementally correct and have mass densities within ±1% of the real tissue densities. The

Table 2. Tissue Substitutes

Tissue Substitutes	Description	$Kg \cdot m^{-3}$	References
	Adipose Tissue		
AP/SF1	Flexible solid based on Epoxy CB3 with fillers of glucose, polyethylene, and phenolic microspheres; a four-component formula is available for trace elements	920	14
AP6	Rigid solid using low exotherm Epoxy CB4; fillers are Teflon, polyethylene, and phenolic microspheres	920	40
AP/LS	Water-based substitutes containing urea, propanol, and phospheoric acid; a four-component formula is available for trace elements	920	14
RF1	Polyethylene-based solid, fat equivalent	930	18
	Blood		
BL/L2	Water-based substitutes containing urea, ethylene glycol, and acetic acid; trace elements are available (five components)	1060	14
	Muscle		
A150	Polymer-based (electrically conducting) substitute comprising polyethylene, nylon, carbon, and calcium fluoride	1120	9, 47
Griffith urethane	Polyurethane-based material having calcium carbonate as filler	1120	10
MS/SR4	Rigid solid using Epoxy CB4 and fillers urea, polyethylene, and phenolic microspheres; five-component formula for trace elements are available	1060	14
MS20	Rigid end-product made up of Epoxy CB2 and fillers magnesium oxide, polyethylene, and phenolic microspheres	1000	40
Figerio liquid	Water-based substitute containing urea, ethylene glycol, and glycerol; a six-component formula for trace elements is available	1080	12
MS/L1	Water-based substitute containing urea, ethylene glycol, urea, and acetic acid; a six-component formula for trace elements is available	1070	14, 30
Water	H_2O	1000	2
MS/G1	A water–gelatin gel containing ethanol and, if required, a six-component formulation for trace elements	1060	14,30
MS/G2	As MS/G1, but urea and propanol replace ethanol.	1050	14,30
RM1	Polyethylene-based solid.	1030	18
	Cortical bone		
B110	A polymer-based electrically conducting, material made up of nylon, polyethylene, carbon, and calcium fluoride	1790	48
HB/SR4	Rigid end-product comprising Epoxy CB2, urea, calcium oxide, calcium hydrogen orthophosphate, magnesium sulfate, and sodium sulfate	1670	14
SB3	Rigid end-product comprising Epoxy CB2 and calcium carbonate	1790	42
Witt liquid	Saturated solution of dipotassium hydrogen orthophosphate in water	1720	49
BTES	Polymer-based material made up of Araldite GY6010, Jeffamine T403, silicon dioxide, and calcium carbonate	1400	22

Table 2. (*Continued*)

Tissue Substitutes	Description	$Kg \cdot m^{-3}$	References
	Inner bone		
IB/SR1	Epoxy resin-based (CB2) solid having filers of calcium orthophosphate, polyethylene, and sodium nitrate	1150	14
IB7	Rigid solid based on Epoxy CB4; fillers are calcium carbonate polyethylene, and phenolic microspheres	1120	40
IB/L1	Water-based substitute comprising dipotassium hydrogen orthophosphate, sodium nitrate, phosphoric acid, urea, and ethylene glycol	1140	14
	Red marrow		
RM/SR4	Rigid solid using Epoxy CB4 and fillers of ammonium nitrate, polyethylene, and phenolic microspheres; a five-component formula for trace elements is available	1030	14
RM/L3	Water-based substitute containing urea and glycerol; a five-component formula for trace elements is available	1040	14
RM/G1	A water–gelatin gel containing glucose; trace elements may be added using a four-component formulation	1070	14
	Brain		
BRN/SR2	Epoxy resin-based (CB2) solid using fillers of acrylics and polyethylene; formula (five-components) for trace element is available	1040	14
BRN/L6	Water-based substitute containing urea, ethanol, and glycerol; a four-component formula for trace elements is available	1040	14, 30
A181	Polyethylene-based solid		25
	Kidney		
KD/L1	Water-based substitute containing sodium chloride, dipotassium hydrogen orthophosphate, urea, and ethylene glycol	1050	14, 30
	Liver		
LV/L1	Water-based substitute containing sodium chloride, dipotassium hydrogen sulfate, sodium chloride, urea, ethanol, and glycol	1060	14, 30
	Lung		
LN/SB4	Foamed rigid epoxy (CB4) system; fillers include urea, polyethylene and the foaming agent DC1107; a five-component formula is available for trace elements	300	14
LN10/75	Foamed rigid epoxy (CV2) system; fillers include polyethylene, magnesium oxide, phenolic, microspheres, surfactant DC200/50, and the foaming agent DC1107	310	42
LTES	Epoxy resin-based system; fillers include phenolic microspheres, surfactant, and foaming agent DC1107,	300	22
	Thyroid		
TH/L2	Water-based substitute containing urea, ethylene glycol, and acetic acid; a three-component formula for trace elements is available	1080	14, 30
	Average breast		
BR12	Rigid solid using low exotherm Epoxy CB4; fillers are calcium carbonate, polyethylene, and phenolic microspheres	970	40
AV.BR/L2	Water-based substitute containing ethanol and pentanediol	960	14
	Total soft tissue		
TST/L3	Water-based substitute containing urea, ethanol, and ethylene glycol; a five-component formula is available for trace elements	1040	14
	Total skeleton		
TSK/SF3	Flexible solid-based on Epoxy CB3; fillers include calcium hydrogen orthophosphate, calcium orthophosphate, and acrylics; a three-component formula is available for trace elements	1360	14
TSK/L1	Water-based substitute containing diammoonium hydrogen orthophosphate, dipotassium hydrogen orthophosphate, and glucose	1360	14

two-part code used indicates the type of tissue being simulated; for example, MS is muscle and BRN is brain, and whether the end product is solid flexible (SF), solid rigid (SR), liquid (L), and so on. Table 4 shows the elemental composition of the recommended substitutes.

As an example for agreement of photon interaction parameters between real tissue and a tissue mimicking material, the mass absorption coefficients for adult adipose tissue and AP6 for photon energies ranging from 10 keV to 10 MeV were calculated. The adult adipose tissue data were obtained from an ICRU report (31). The book compiles photon, electron, proton, and neutron data for body tissues. The AP6 data were calculated using the atomic composition given in Table 3 and the photon interaction data compiled in a NIST report (45). Figure 1 shows an excellent agreement of the interaction parameter between two materials. Table 3 indicates that AP6 is Class A material of the adipose tissue for the entire photon energy range.

Several experiments were carried out to verify the accuracy with which the various substitutes simulate

Table 3. Classification of Tissue Substitute

			Classification					
			Photons		Electrons	Protons	Neutrons	
Tissue being simulated	Substitute	Phase	10–99 keV	100 keV–100 MeV	10 keV–100 MeV	1–1000 MeV	1–99 keV	100 keV–30 MeV
Adipose tissue	AP/SF1	Solid	B	A	A*(A)	A	A*	A
	AP6	Solid	A	A	A	B	C	C
	AP/L2	Liquid	C	B	A(B)	A	A*	A
	AP/RF1	Solid	A					
Blood	BL/L2	Liquid	A	A	A*(a)	A*	A*	A*
Muscle	A150	Solid	C	B	B	B	A	B
	Griffith urethane	Solid	B	B	B	B	B	B
	MS/SR4	Solid	C	B	B	B	A	B
	MS20	Solid	A	A	A	B	B	B
	Figerio liquid	Liquid	A*	A*	A*	A*	A*	A*
	MS/L1	Liquid	A*	A*	A*	A*	A*	A*
	Water	Liquid	A	A(A*)	A*	A	B	B
	MS/G1	Gel	A	A*	A*	A*	A*	A*
	MS/G2	Gel	A	A*	A*	A*	A*	A*
	MS/RM1	Solid	A					
Cortical bone	B110	Solid	A*	A	A	B	B	B
	HB/SR4	Solid	B	A	A	B	C	C
	SB3	Solid	A	A	A	A	C	B
	Witt liquid	Liquid	A(B)	A	A	A	C	C
	BTES	Solid	A*	A*				
Inner bone	IB/SR1	Solid	B	B	B	B	A	B
	IB7	Solid	A	B(A)	A(B)	A	B	C
	IB/L1	Liquid	B	A(B)	A	A	A*	A
Red marrow	RM/SR4	Solid	C	B(A)	A(B)	A	A*	A
	RM/L3	Liquid	C	B	A(B)	B	A*	B
	RM/G1	Gel	C	B	A(B)	B	A*	B
Brain	BRN/SR2	Solid	C	B	B	B	A	B
	BRN/L6	Liquid	A*	A*	A*	A*	A*	A*
	A181	Solid					A	A
Kidney	KD/L1	Liquid	A	A*	A*	A*	A*	A*
Liver	LV/L1	Liquid	A*	A*	A*	A*	A	A*
Lung	LN/SB4	Solid	C	B	B	B	A	B
	LN1	Solid	A	A	A	B	C	C
	LN10	Solid	A	B(A)	A(B)	B	A	B
	LTES	Solid	A*	A*				
Thyroid	TH/L2	Liquid	A(B)	A*(B)	A*	A*	A*	A*
Average breast	BR12	Solid	A	A	A	B	C	C
	AV.BR/L2	Liquid	A*	A*	A*	A*	A*	A*
Total soft tissue	TST/L3	Liquid	A*	A*	A*	A*	A*	A*
Total skeleton	TSK/SF3	Solid	A	A	A	A	A	B
	TSK/L1	Liquid	B	A*(B)	A*(B)	A	B	B

the corresponding real tissues. In one such series of experiments, thin-walled cells with 10×10 cm^2 cross section were filled with real tissues and immersed in to appropriate tissue equivalent liquid to displace equal volume of that liquid. Central axis depth doses and beam profiles were then measured in the liquid behind the cells and the results compared with those obtained in the liquid alone. The material used or the construction of the cells was solid muscle substitute for the comparison with human muscle, beef stake, and pork, Brain substitute was used for the comparison with human brain. Co-60 γ source was used for the irradiation experiments. In no case did the readings at the depth differ by $> 1\%$ in each comparison. Similar tests were made with a 160 MeV synchrocyclotron proton beam and a neutron beam with average energy of 7.5 MeV (14,46). The results with the substitutes in place were generally within 0.5% of the readings for real tissues. In another series of tests, the relationship between the attenuation coefficients and the Hounsfield units measured with a computed tomography (CT) scanner was established first using 120 kVp X rays, and then the CT numbers were measured for a range of the new tissue substitutes. The attenuation coefficients derived from the measured CT number of each material was compared

Table 4. Elemental Compositions of the Tissue Substitute

	Elemental Composition (percentage by weight)											
Tissue Substitutes	H	C	N	O	Na	Mg	P	S	Cl	K	Ca	Other Elements
Adipose tissue												
AP/SF1	11.96	75.50	0.80	11.11	0.05		0.02	0.07	0.45	0.03	0.02	
AP6	8.36	69.14	2.36	16.94					0.14			F(3.07)
AP/L2	12.12	29.29	0.80	57.40	0.05	0.002	0.18		0.12	0.08	0.002	
AP/RF1	14.11	84.07		0.92		0.30					0.60	
Blood												
BL/L2	10.01	9.82	2.91	76.37	0.18	0.002		0.20	0.27	0.14	0.004	
Muscle												
A150	10.10	77.60	3.50	5.20							1.80	F(1.70)
Griffith urethane	9.00	60.20	2.80	26.60							1.72	Sn(0.01)
MS/SR4	9.50	70.28	3.48	15.55	0.08/	0.02	0.18	0.50	0.12	0.30	0.01	
MS20	8.12	58.35	1.78	18.64		13.03			0.09	0.39	0.01	
Figerio liquid	10.20	12.30	3.50	72.89	0.07	0.02	0.20	0.32	0.08	0.39	0.01	
MS/L1	10.20	12.30	3.50	72.90	0.07	0.02	0.20	0.32	0.09	0.39	0.01	
Water	11.19			88.81								
MS/G1	10.20	12.51	3.50	73.00	0.07	0.02	0.20		0.09	0.39	0.01	
MS/G2	10.35	12.31	3.50	73.04	0.07	0.02	0.20		0.09	0.39	0.01	
MS/RM1	12.24	73.36		6.37		6.03					2.00	
Cortical bone												
B110	3.70	37.10	3.20	4.80							26.29	F(24.39)
HB/SR4	4.45	29.09	3.88	31.93	0.06	0.21	10.00	0.32	0.06		19.99	
SB3	3.10	31.26	0.99	37.57					0.05		27.03	
Witt liquid	4.70			56.80			10.90			27.90		
BTES	4.0	37.8	1.5	35.3					0.1		9.4	Si(11.9)
Inner bone												
IB/SR1	8.73	63.19	2.36	17.83	0.06		2.62		0.12		5.09	
IB7	6.86	59.01	2.08	24.12					0.12		7.81	
IB/L1	8.65	17.27	2.58	60.83	0.06		2.49			4.99		
Red marrow												
RM/SR4	10.08	73.57	2.16	13.77	0.01	0.003	0.03	0.14	0.11	0.15		
RM/L3	10.17	12.77	2.22	74.24	0.08		0.03	0.15	0.17	0.17		
RM/G1	10.20	9.38	2.36	78.18	0.08		0.03	0.15	0.17	0.17		
Brain												
BRN/SR2	10.69	72.33	1.28	14.59	0.18	0.01	0.36		0.06	0.30	0.01	
BRN/L6	10.68	15.14	1.29	71.67	0.18		0.34	0.17	0.23	0.30		
A181	10.7	80.3	2.2	3.3							1.8	F(1.7)
Kidney												
KD/L1	10.40	11.35	2.74	74.50	0.18		0.19		0.28	0.25		
Liver												
LV/L1	10.18	14.40	2.83	71.80	0.11			0.24	0.18	0.29		
Lung												
LN/SB4	9.70	70.26	2.80	16.30	0.17	0.01	0.12	0.22	0.11	0.19	0.01	Si(0.50)
LN1	6.00	51.44	4.29	30.72								Al(7.55)
LN10	8.38	60.40	1.68	17.28		11.4			0.15			Si(0.84)
LTES	7.0	57.4	2.1	22.4		9.3	1.7			9.1		
Thyroid												
TH/L2	10.01	13.58	2.20	73.52	0.22		0.08		0.14	0.19		I(0.06)
Average Breast												
BR12	8.68	69.95	2.37	17.91					0.14		0.95	
AV.BR/L2	11.79	37.86		50.41								
Total soft tissue												
TST/L3	10.46	23.33	2.59	62.54	0.11	0.01	0.13	0.20	0.13	0.20	0.02	
Total skeleton												
TSK/SF3	7.16	45.50	3.08	26.12	0.31	0.12	7.02	0.16	0.47	0.15	10.03	
TSK/L1	7.45	4.64	2.94	66.93	0.32		7.00		0.13	10.15		

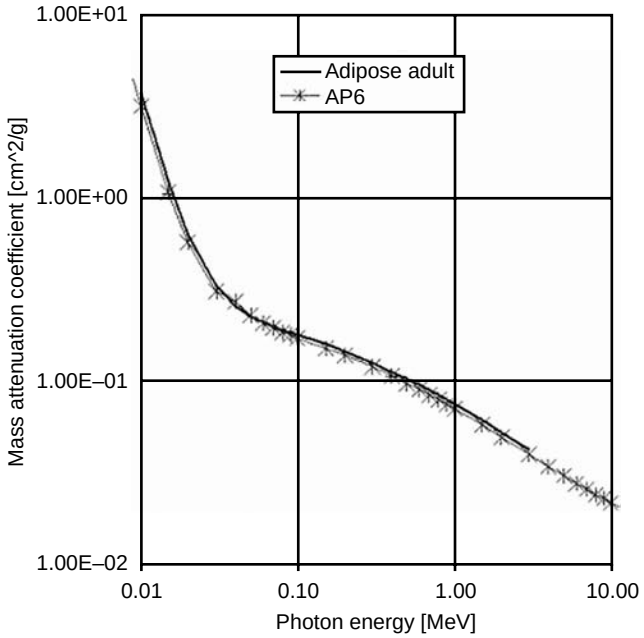


Figure 1. Comparison of mass attenuation coefficient for adult adipose tissue and tissue-mimicking material AP6.

to the calculated μ value of each material. The measured μ values were generally within 2% of the computed ones (14).

APPLICATIONS: RADIATION DOSIMETRY

Radiation exposure is equal to the number of electric charges liberated by interaction of photons with air. The gold standard for exposure measurement is a free-air chamber. The chamber size must be large enough to stop photons and associated secondary electrons. Since the size could be very large for practical uses, physicists developed small cylindrical chambers called thimble chambers by making the chamber wall with air-equivalent material (47–49). Graphite, Bakelite ($C_{43}H_{38}O_7$), or mixture of those is commonly used as the wall material. These have a smaller effective atomic number than that of air; but, it is accepted because the central electrode of the ionization chamber is usually made of aluminum, whose atomic number, 13, is much larger than air (or 7.67).

Radiation dose absorbed in tissue is the most important physical parameter for therapeutic applications of radiation. It can cause fatal effects on a person or may fail to kill the malignant cells of a patient unless the delivered dose is carefully monitored. The mail instrument for dose measurement is the ionization chamber. The most common ionization chambers for dose measurement are cylindrical with an outer diameter of $\sim$ 1 cm and a length of air cavity of $\sim$ 2 cm. When the ionization chamber is used in a solid phantom or water, some corrections are needed to estimate the dose in the medium because of differences in materials of air, the chamber wall, and the phantom material (50).

Absorbed dose in real patients can be measured by placing radiation detectors on or inside the patient during treatment. Thermoluminescent detector (TLD), a solid-state detector, is a well-established instrument for the *in vivo* dose measurements. The TLD is made of thermoluminescent material, which absorbs radiation energy. The electric charges created in the material can be measured by heating the chip after irradiation and used to estimate the absorbed dose. In addition to the thermoluminescent characteristics, TLD should be radiologically equivalent to tissue. The common TLD material is lithium fluoride (LiF) with some impurities such as Magnesium (Mg) or Titanium (Ti) to improve the property. The effective atomic number is kept close to the tissue, or 8.2, for LiF TLD to minimize the fluence disturbance due to a foreign material placed in tissue. Since the TLD material is not identical to tissue, the response to radiation is different for that of tissue. A great concern exists when absorbed dose for low energy photons, or energies $<$ 100 keV, must be measured because the radiation response is very different from the tissue in this energy range (47).

APPLICATIONS: RADIATION THERAPY

Accurate prediction of radiation dose delivered to patients is the most important step to generate effective radiotherapy treatment strategies. Medical physicists, who are responsible for physical aspects of the treatment planning, have to establish the physical data necessary for radiotherapy before anyone can be treated and even during actual treatment. For given radiation sources, medical physicists have to know how much radiation energy or dose is absorbed at any point in the patient's body. Generally, computers are used to predict the dose distribution. The computer can calculate doses using radiation source specific physical data incorporated in the software. The necessary data vary according to the calculation model used by the computer program. In general the absolute dose at any point or at any depth in the human body and the relative dose in the entire body are needed.

Medical physicists have developed many physics concepts since radiation sources were introduced into the clinic. For radiotherapy using electron accelerators, medical physicists introduced three important concepts: output factor, depth dose, and beam profile (48,49). The output factor is the radiation dose at a specific depth along the central beam axis or at a reference point for a given standard field size, (e.g., 10 $\times$ 10 cm). The depth dose gives the dose to a point at a given depth as a fraction of the dose at the reference point along the central beam axis. The beam profile shows the variation of dose along a line on a plane perpendicular to the central beam axis.

Physical models of the radiation source and human body are not perfect. Dose calculations in a real human body are difficult because of the variation in shape and tissue densities. Consequently, necessary data must be measured. The measurements are generally performed in a uniform and large medium, such as a water phantom.

Tissue Equivalent Phantom

Water. Water is a favorite medium for dose measurements for several reasons. Water is nearly tissue equivalent and inexpensive (or readily available). Furthermore, a radiation detector can be scanned through in water for dose

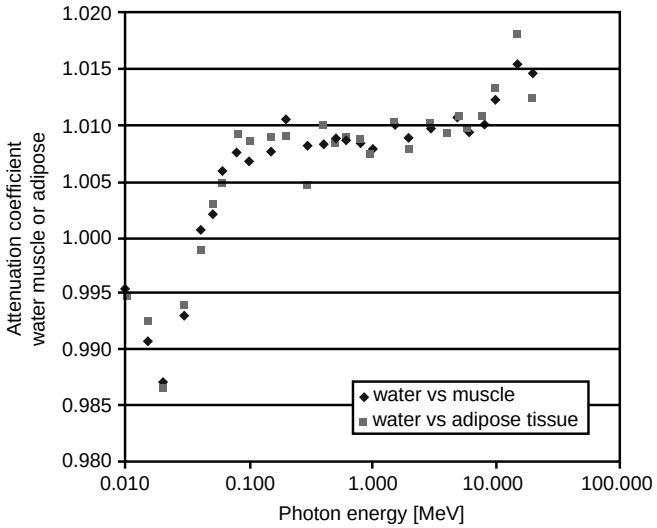


Figure 2. The ratio of mass attenuation coefficients between water and muscle or adipose tissue versus photon energy.

measurements at points spaced very closely. One should remember, however, that the equivalence of water to tissue depends on the photon energy and the tissue to which water is compared. Mass attenuation coefficients of water, muscle, and adipose tissue were evaluated for photons in the energy range between 1 keV and 20 MeV. The ratios of water to muscle and water to adipose tissue are plotted in Fig. 2. The mass attenuation coefficients of water are 1 % larger than those of muscle and adipose tissue for energies > 0.1 MeV. However, in the energy range between 0.01 and 0.1 MeV the mass attenuation coefficients of water are smaller as much as 1.5% with those of muscle and adipose tissue.

Solid Phantom. Liquid water is cumbersome to use in a high electric voltage environment such as in an accelerator room. Hence, medical physicists manufactured water-equivalent solid phantoms. There are a few solid phantom materials currently available commercially. The list of products is given in Table 5. Poly (methyl methacrylate) or acrylic is sold under names of Lucite, Plexiglas, or Perspex. Polystyrene has usually clear color and it is common phantom material for physics quality assurance programs because of relatively low price. Physicists discovered that PMMA and polystyrene leads to as much as 5% error when those are used for the absolute dose measurements of electron beams. Hence, an epoxy resin-based solid water material was developed as a substitute for water (16). It was made to simulate radiological characteristics of water more accurately than PMMA and polystyrene. Both mass attenuation coefficients and electron stopping powers of the solid water agree with those of water within 1% for the energy range between 10 keV and 100 MeV (51).

The new generation of solid water is being manufactured by Gammex-RMI (Middleton, WI). The phantoms are available in slabs of various sizes and thickness as seen in Fig. 3. Plastic Water was developed by CIRS (Computerized Imaging Reference Systems Inc., Norfolk, VA). It is a variation of the solid water. It is flexible. It was shown that measured outputs of electron beams in the Plastic Water agree with water within 0.5% for energy ranging from 4 to 20 MeV (52). Med-Tec, Inc. (Orange City, IA) is manufacturing what they call Virtual Water, which has the same chemical composition and density as solid water, but the manufacturing process is different. Hermann et al. developed polyethylene-based water-equivalent solid material (17). The material is manufactured by PTW (Freiburg, Germany) and it is sold as White Water or RW3. It contains titanium oxide as additive.

The dose at a depth (5 or 7 cm for photon beams and the depth of dose maximum for electron beams) in solid phantoms was measured with ionization chambers for various photon and electron energies. The results were compared with the dose measured in water. Figure 4 shows the ratio of measured doses in a solid phantom and water for plastic water (PW), white water (RW-3), solid water (SW-451 and SW-457), and virtual water (VW) (52,53). The horizontal axis of the figure indicates the photon energy. The larger the ionization ratio, the higher the photon energy. The beam energy ranges from Co-60 γ ray (1.25 MeV) to 24 MV. The solid water and virtual water show the best agreement

Table 5. Elemental Compositions of Common Water-Equivalent Phantom Materials

		Elemental Composition (percentage by mass)												
Tissue		H	B	C	N	O	Mg	Al	Cl	Ca	Other Elements	Mass Density, $kg \cdot m^{-3}$	Z_{eff}	N, $e \cdot kg^{-1} \times 10^{26}$
PMMA[a] (Acrylic)	Lucite, Plexiglas, Perspex	8.0		60.0		32.0						1170	6.24	3.25
Polystyrene		7.7		92.3								1060	5.69	3.24
Solid water 457	GAMMEX-RMI	8.1		67.2	2.4	19.9			0.1	2.3		1042	8.06	3.34
Virtual water	MEDTEC	8.1		67.2	2.4	19.9			0.1	2.3		1070	8.06	3.48
Plastic water	CIRS	7.4	2.26	46.7	1.56	33.52	6.88	1.4	0.24			1030	7.92	3.336
White water RW3	PTW	7.61		91.38		0.14					Ti(0.78)	1045	5.71	3.383
Polymer gel	MGS	10.42		10.45	2.44	76.68						1050	7.37	3.49
Water		11.19				88.81						1000	7.42	3.343

[a]Poly (methyl methacrylate = PMMA.)

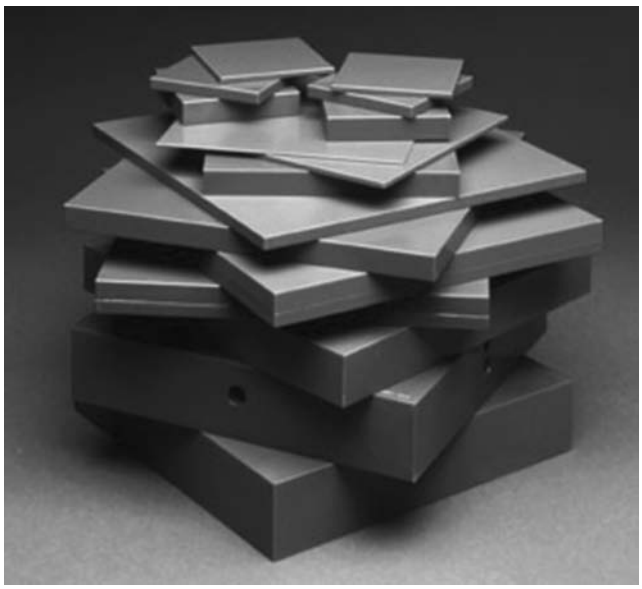


Figure 3. Solid water phantom manufactured by GAMMEX-RMI (Middleton, WI).

among the evaluated materials. The results for electron beams are given in Fig. 5, where the horizontal axis indicates the effective energy, which is an approximation of the beam energy. The solid water and virtual water also show a good agreement for the entire electron energy.

Gel Phantom. A relatively new development is taking place on a tissue-equivalent material in gel form. The new material can record the radiation dose without additional instrument inserted in the phantom. The phantom can be made large enough to simulate the human body. Among many variations of gel phantoms, the most promising is polymer gel manufactured by MGS Research Inc. (Guilford, CT) (54). Originally it was made of acrylamide, *N*, *N*′-methylenebisacrylamide (bis), agarose, and water. The chemical composition was optimized over years. Currently,

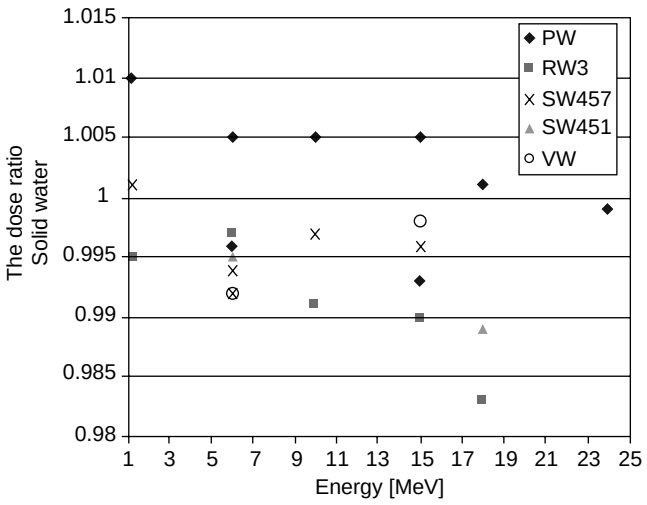


Figure 4. The ratio of measured output in solid phantom and water for photon beams.

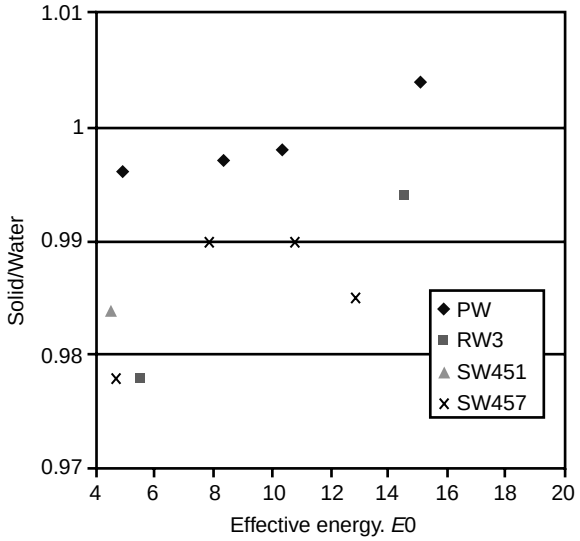


Figure 5. The ratio of measured output in solid phantom and water for electron beams.

the polymer gel is sold as BANG3 with 80% water, 14% of gelatin, and 6% of methacrylic acid by mass. A typical atomic composition of the polymer gel as well as the mass density, the effective atomic number, and the electron density are given in Table 5. The effective atomic number and the electron density of the polymer gel agree with those of water very well. The polymer gel produces highly linked polymers when it is irradiated. The structural change causes a change in color and mass density. The recorded dose distribution can be indirectly read by measuring the photon attenuation of both visible light and X ray. The polymerization also causes a change in the magnetic property of the gel. The most popular method is currently to scan the irradiated phantom with a magnetic resonance imaging (MRI) scanner. It takes advantage of the change in the spin–spin relaxation rate, which increases with increasing absorbed dose.

Differing from more traditional dose measurement tools, polymer gel dosimeter enables medical physicists to obtain full three-dimensional (3D) dose distributions in a geometrically consistent way. Polymer gel phantom was used to measure 3D dose distributions of advanced radiotherapy treatment such as those for intensity modulated radiation therapy (IMRT) (55) and Gamma Knife stereotactic radiosurgery (56). Figure 6 shows an MRI image taken after the polymer gel was irradiated with a Gamma Knife system (Elekta AB, Stockholm, Sweden). The blighter color indicates higher dose.

Geometric Phantom

With the advent of rapid development of highly conformal radiation therapy, modern radiation therapy requires high geometrical precision of radiation delivery. At the same time, the complexity of treatment planning software has substantially increased. This has urged medical physicists to test the geometrical precision of the treatment planning software (57). Special phantoms are being manufactured to assure the quality of geometry created by the software.

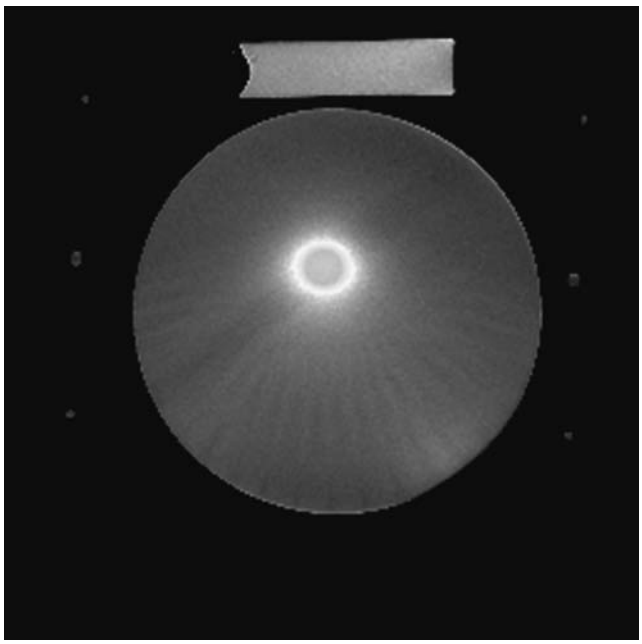


Figure 6. Dose distribution images obtained by polymer gel dosimeter for Gamma Knife irradiation.

Here we discuss two of those phantoms currently on market.

Lucy 3D Precision Phantom. Lucy phantom was developed by Toronto Sunny Brook Regional Cancer Center and Sandstrom Trade and Technology, Inc. (Ontario, Canada) (58). The acrylic phantom of a shape of head (Fig. 7) is used to test the image quality of CT, MRI, and X ray imaging modalities, which are used for stereotactic radiation therapy. It can verify imaging errors, image distortions, and the geometrical accuracy of treatment planning system. It serves also as routine machine QA equipment.

Figure 7. Lucy 3D precision phantom from the Sandstrom Trade and Technology Inc. (Ontario, Canada).

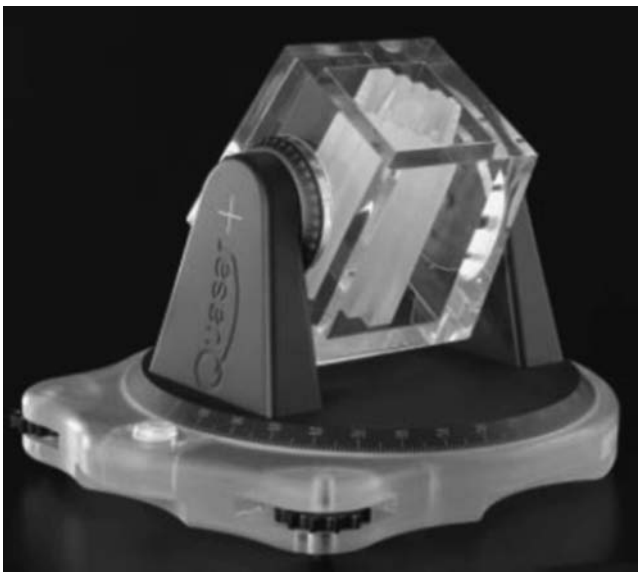


Figure 8. Quasar beam geometry MLC phantom from Modus Medical Devices, Inc. (Ontario, Canada).

Quasar Phantom. Another Canadian company, Modus Medical Device, Inc. (Ontario, Canada), developed phantoms to verify geometrical accuracy generated by radiation therapy treatment planning software (59). The phantoms are made of Lucite, cedar, and polystyrene. One of those, Quasar body phantom, can be used to test the accuracy of the digitally reconstructed radiography (DRR) images. DRR images are generated from a number of axial images taken by a CT scanner and the DRR images are compared with the images taken with an electronic portal imaging system or with radiographic port films before treatment to verify the patient position accuracy. Quasar beam geometry MLC phantom shown in Fig. 8 can verify the geometrical display accuracy of leaf position and the size of the multileaf collimator, which have replaced traditional blocks to define radiation fields.

Humanoid Phantom

Complex high precision radiation delivery techniques, such as IMRT and stereotactic radiation therapy, demand ever increasing accuracy of dose delivery in geometry close to the human body. Body phantoms such as Rando and Alderson were introduced many years ago. Those were used to predict to radiation exposure to humans during radiation therapy and diagnostic radiology procedures. The Rando phantom manufactured by the Phantom Laboratory (Salem, NY) is shown in Fig. 9. The body phantom is sliced into many 5 cm thick slabs. Radiographic films can be inserted between the slabs for dose measurements. The slabs can also hold TLD chips for absolute dose measurements. The phantom is made of soft-tissue equivalent material, which is manufactured with proprietary urethane formulation. The phantom uses natural human skeletons. Lungs and breast closely mimic the real tissues.

Recently, many phantoms which simulate a part of body were developed, for example, Gammex-RMI, Phantom

Figure 9. Rando phantom manufactured by the Phantom Laboratory (Salem, NY).

Laboratory, CIRS (60), and so on. Many of those phantoms are made to measure dose inside the phantom for verification of radiation therapy treatment. There are phantoms in various shapes for IMRT QA. Those phantoms can accommodate ionization chambers, radiographic films, and TLDs for dose measurements. Those may have special inserts for nonsoft tissue materials, such as lung and bone. The shapes of the phantoms are torso, head, thorax, pelvis, and neck.

DIAGNOSTIC IMAGING

X Rays

X rays are used to take images of parts of the body for diagnostic purposes (61). The oldest and most commonly used X ray imaging modality is X ray radiography. Images can be recorded on radiographic films. Recently, digital recording has become more common since the digital technique requires no wet-film processing and allows radiologists to manipulate and store the images more easily than hard-copy films.

Medical physicists use phantoms to test the quality of images. Some of common phantoms are dual-energy X ray absorptiometry phantom, anatomical phantoms (62), digital subtraction angiography (DSA) phantom, contrast detail phantom, and dental image QA phantoms.

Fluoroscopy

A regular X ray device can take static images of patients. A fluoroscopic unit, consisting of an X-ray tube, a camera, an image intensifier, and a TV monitor, can record images of moving parts of the body and of objects placed inside the body. This method of recording images is called fluoroscopy. Interventional radiologists use fluoroscopy to monitor the location of very thin wires and catheters going through blood vessels as those are being inserted during a procedure. Phantoms are an important instrument to assure the quality of fluoroscopy images. There are fluoroscopic phantoms, such as cardiovascular fluoroscopic benchmark phantom, fluoroscopy accreditation phantom, and test phantom.

Mammography

Mammography is a major diagnostic modality to detect breast cancer. The tumor size is often very small, that is, submillimeter diameter. Though small, early discovery of such small tumors is very important and can lead to better therapy outcomes, that is, higher cure rates and longer survival. Hence, the quality of X ray equipment used for mammography is a key for the success of this imaging modality and its performance is tightly controlled by federal and local governments. Consequently, medical physicists developed many phantoms to evaluate the imaging quality of mammography devices accurately and efficiently. Several phantoms are used, known as QA phantom, accreditation phantom, high contrast resolution phantom, contrast detail phantom, digital stereotactic breast biopsy accreditation phantom, phototimer consistency test tool, and collimator assessment test tool. A tissue-equivalent phantom manufactured by CIRS (Norfolk, VA) is shown in Fig. 10. The phantom is 4.5 cm thick and simulates the shape of a breast during the imaging. It is made of CIRS resin material mimicking the breast tissue. Objects with varying size within the phantom simulate calcifications, fibrous tissue in ducts, and tumor masses.

Computed Tomography

Computed tomography scanning systems were developed in the 1970s and rapidly deployed into clinics into the 1980s. Currently, this imaging modality is commonly used for diagnosis and radiation therapy in clinic and hospitals. Compute tomography provides patient's anatomy on planes transverse to body axis (from head to toe). The plane images called axial slices are taken every 0.05–1 cm. Depending on the axial length of the scan, the number of slices can vary from 20 to 200. Computed tomography

Figure 10. Tissue-equivalent mammography phantom from CIRS (Norfolk, VA).

enables radiologists to visualize the location of disease in 3D, leading to potentially more accurate diagnosis. Medical physicists use phantoms to estimate dose to patients during CT scan, to test the precision of scanning geometry, and verify the quality of CT images. Such phantoms are the bone mineral analysis phantom, spiral/helical CT phantom, CT dose phantom, orthopedic calibration phantom, spine phantom, electron density phantom, and CT performance phantom.

Each picture element (pixel) of a CT image is represented by a CT number (or Hounsfield units). The CT number for water is 0, while the CT numbers in the lung are below -200 and those in the bones are $\sim$ 200. Since the CT number is found to depend on the electron density, among other parameters, it can be used to identify the material type and the distribution of the tissue inhomogeneities, which is very useful in radiation therapy treatment planning. One electron density phantom, developed in order to establish the relationship of CT number and tissue electron densities, is made of solid water and has cylindrical shape with a radial size of pelvis and 6 cm thickness. Small cylindrical inserts mimicking different tissues are plugged into the phantom. When it is CT scanned, the measured CT numbers of those materials can be plotted against the predetermined electron densities and the resulting data are used in treatment planning computer. This electron density CT phantom, which is now commercially available, shown in Fig. 11, was designed by Constantinou (63) and manufactured by GAMMEX-RMI (Middleton, WI). Available insert materials are lung (LN-300 and LN-450), adipose (AP6), breast, CT solid water, CB3 resin, brain, liver (LV1), inner bone, bone (B200, CB2-30% mineral, CB2-50% mineral), cortical bone (SB3), Titanium simulating Titanium implants for hip replacements, and true water.

Figure 11. Electron density phantom RMI 467 from GAMMEX-RMI (Middleton, WI).

NUCLEAR MEDICINE

For nuclear medicine procedures, radioactive compound composed from Technetium, Iodine, Fluorine, etc. is injected into blood stream (64). The material is carried to a potential disease site by blood and stays there. Radioactive material emits photons or positively charged electrons called positrons. Since photons can easily escape the patient body, the photons can be detected by a photon detecting device. The photons are used to form images. Positron emission tomography (PET) uses positron emitting radioactive material. When a positron interacts with an electron in the body, both particles are converted into photons. The PET is more advanced nuclear medicine procedure and can generate 3D distributions of radioactive material. Phantoms are used for quality assurance of nuclear medicine systems. A NEMA (National Electrical Manufacturers Association) scatter phantom manufactured by CIRS Inc. (Norfolk, VA) is a circular polystyrene cylinder. Radioactive material with known activity is delivered to line source insert tubes. The image of the phantom is taken to test scatter fraction and count losses.

FUTURE

Tissue-equivalent materials for solid phantoms are well developed. There may be a need of incremental improvement in phantom materials possibly for lower price and more accurate representation of tissue types. Furthermore, phantom materials are needed for newer radiation types such as heavy ions. As new therapy and imaging modalities will be put into practice, new phantoms will be developed for efficient and accurate testing of those new tools. Modern radiology utilizes not only ionizing radiation such as photons, electrons, and heavy particles, but also electromagnetic radiation (or EM-waves) and ultrasound. Phantom materials exist for testing ultrasound devices and MRI scanners. This is an active area for development. For example, D'Souza recently developed a phantom used for quality testing of ultrasound, MRI, and CT (65). Readers interested in those phantoms should consult with the phantom manufactures, such as GAMMEX-RMI (Middleton, WI), CIRS, Inc. (Norfolk, VA) and Fluke/Cardinal Health (Cleveland, OH).

Here, the focus is on two aspects of interest for future development. It may be necessary to develop more biologically tissue equivalent phantom materials. Such materials simulate not only the radiological characteristics of radiation modalities, but also it can closely simulate the radiation effects on the tissues in realistic geometry. Biomedical engineers and scientists are vigorously working on artificial tissue, potentially replacing the real tissue. Such material could be also used as phantom material. Readers should refer to a comprehensive review on the current state of art in tissue engineering authored by Lanza et al. (66).

Computer modeling of human body is an active field of research and development. Noticeable example is the Visual Human Project sponsored by National Library of Medicine, NLM (67). Whole bodies of male and female cadavers were scanned with CT and MRI. The datasets

are available for anyone who is interested in the information through licensing with NLM and with minimal cost. The datasets for the male consist of 12 bits axial MRI images of the head and neck and up to 1871 axial CT slices of the whole body. The female data set is 5000 axial CT images, with which one can reconstruct 3D images with $0.33 \times 0.33 \times 0.33$ mm cubic voxels. The data can be used to construct human body model for quality assurance of radiological systems, in particular, for testing the radiation therapy treatment planning system. In addition to real geometry, the data contain the detailed information of tissue heterogeneities in human body. Such data are indispensable for accurate assessment of new radiological technologies.

BIBLIOGRAPHY

1. Spires FW. Materials for depth dose measurement. Br J Radiol 1943;16:90.
2. Kienbock R. On the quantimetric method. Arch Roentgen Ray 1906;1:17.
3. Baumeister L. Roentgen ray measurements. Acta Radiol 1923;2:418.
4. Ott P. Zur Rontgenstrahlenbehandlung oberblachlich gelagerter Tumoren. Strahlentherapie 1937;59:189.
5. Jones DEA, Raine HC. Bri J Radiol 1949;22.
6. Harris JH, et al. The development of a chest phantom for radiologic technology. Radiology 1956;67:805.
7. Markus B. The concept of tissue equivalence and several water-like phantom substances for energies of 10 KeV to 100 MeV as well as fast electrons. Strahlentherapie 1956; 101:111–131.
8. Rossi HH, Failla G. Tissue equivalent ionization chamber. Nucleonics 1956;14:32.
9. Shonka FR, Rose JE, Failla G. Conducting plastic equivalent to tissue, air and polystyrene. Prog Nucl Energy 1958;12:184.
10. Griffith RV, Anderson AL, Dean PN. Further realistic torso phantom development. University of California Research Laboratory, UCRL-50007-76-1; 1976.
11. Goodman LJ. A modified tissue equivalent liquid. Health Phys 1969;16:763.
12. Frigerio NA, Sampson MJ. Tissue equivalent phantoms for standard man and muscle. Argonne National Laboratory, Argonne, IL. ANL-7635; 1969.
13. White DR. The formulation of substitute materials with predetermined characteristics of radiation absorption and scattering. Ph.D. dissertation University of London, London; 1974.
14. Constantinou C. Tissue substitutes for particulate radiations and their use in radiation dosimetry and radiotherapy. Ph.D. dissertation, University of London, 1978.
15. White DR, Constantinou C. Prog Med Radiat Phys 1982;1:133.
16. Constantinou C, Attix FH, Paliwal BR. A solid water phantom material for radiotherapy x-ray and gamma-ray beam calibrations. Med Phys 1982;9:436–441.
17. Hermann KP, Geworski L, Muth M, Harder D. Polyethylene-based water-equivalent phantom material for x-ray dosimetry at tube voltages from 10 to 100 kV. Phys Med Biol 1985;30: 1195–1200.
18. Hermann KP, et al. Muscle- and fat-equivalent polyethylene-based phantom materials for x-ray dosimetry at tube voltages below 100 kV. Phys Med Biol 1986;31:1041–1046.
19. Kalender WA, Suess C. A new calibration phantom for quantitative computed tomography. Med Phys 1987;14:863–886.
20. Homolka P, Nowotny R. Production of phantom materials using polymer powder sintering under vacuum. Phys Med Biol 2002;47:N47–52.
21. Homolka P, Gahleitner A, Prokop M, Nowotny R. Optimization of the composition of phantom materials for computed tomography. Phys Med Biol 2002;47:2907–2916.
22. Jones AK, Hintenlang DE, Bolch WE. Tissue-equivalent materials for construction of tomographic dosimetry phantoms in pediatric radiology. Med Phys 2003;30:2072–2081.
23. Suess C, Kalender WA, Coman JM. New low-contrast resolution phantoms for computed tomography. Med Phys 1999;26: 296–302.
24. Iwashita Y. Basic study of the measurement of bone mineral content of cortical and cancellous bone of the mandible by computed tomography. Dentomaxillofac Radiol. 2000;29:209–215.
25. Burmeister J, et al. A conducting plastic simulating brain tissue. Med Phys 2000;27:2560–2564.
26. ICRU. Tissue substitutes in radiation dosimetry and measurement. International Commission on Radiological Units and Measurements, Bethesda, MD, ICRU Report 44; 1989.
27. Woodard HQ. The elementary composition of human cortical bone. Health Phys 1962;8:513–517.
28. ICRU. Physical Aspects of Irradiation. International Commission on Radiological Units and Measurements, Bethesda, MD, Report 10b; 1964.
29. ICRP. Reference man: anatomical, physiological and metabolic characteristics. International Commission on Radiological Protection, Stockholm, Sweden, ICRP Publication 23; 1975.
30. Constantinou C. Phantom materials for radiation dosimetry. I. Liquids and gels. Br J Radiol, 1982;55:217–224.
31. ICRU. Photon, electron, proton and neutron interaction data for body tissues, International Commission on Radiological Units and Measurements, Bethesda, (MD): ICRU Report 46; 1992.
32. White DR, Fitzgerald M. Calculated attenuation and energy absorption coefficients for ICRP Reference Man (1975) organs and tissues. Health Phys 1977;33:73–81.
33. Bewley DK. Pre-therapeutic experiments with the fast neutron beam from the Medical Research Council cyclotron. II. Physical aspects of the fast neutron beam. Br J Radiol 1963; 36:81–88.
34. Jones TD. Distributions for the design of dosimetric experiments in a tissue equivalent medium. Health Phys 1974;27: 87–96.
35. White DR. Phantom materials for photons and electrons. Hospital Physicists Association, London, S. Rep. Ser. No. 20; 1977.
36. Frigerio NA, Coley RF, Sampson MJ. Depth dose determinations. I. Tissue-equivalent liquids for standard man and muscle. Phys Med Biol 1972;17:792–802.
37. Frigerio NA. Neutron penetration during neutron capture therapy. Phys Med Biol 1962;6:541–549.
38. White DR. The formulation of tissue substitute materials using basic interaction data. Phys Med Biol 1977;22:889–899.
39. White DR. An analysis of the Z-dependence of photon and electron interactions. Phys Med Biol 1977;22:219–228.
40. Geske G. The concept of effective density of phantom materials for electron dosimetry and a simple method of their measurement. Radiobiol Radiother 1975;16:671–676.
41. White DR, Martin RJ, Darlison R. Epoxy resin based tissue substitutes. Br J Radiol 1977;50:814–821.
42. White DR. Tissue substitutes in experimental radiation physics. Med Phys 1978;5:467–479.
43. White DR, Constantinou C, Martin RJ. Foamed epoxy resin-based lung substitutes. Br J Radiol 1986;59:787–790.

44. White DR, Speller RD, Taylor PM. Evaluating performance characteristics in computed tomography. Br J Radiol 1981;54: 221.
45. Hubbell JH, Seltzer SM. Tables of x-ray mass attenuation coefficients and mass energy-absorption coefficients 1 keV to 20 MeV for elements $Z = 1$ to 92 and 48 additional substances of dosimetric interest. National Institute of Standards and Technology, Gaithersburg, (MD): NISTIR 5632; 1995.
46. Constantinou C, et al. Physical measurements with a high-energy proton beam using liquid and solid tissue substitutes. Phys Med Biol 1980;25:489–499.
47. Attix FH. Introduction to Radiological Physics and Radiation Dosimetry. New York: Wiley-Interscience; 1986.
48. Johns HE, Cunningham JR. The Physics of Radiology. 4th ed. Springfield (IL): Charles C. Thomas Publisher; 1983.
49. Kahn FM. The Physics of Radiation Therapy. 2nd ed. Baltimore: Williams&Wilkins; 1994.
50. AAPM, A protocol for the determination of absorbed dose from high-energy photon and electron beams. Med Phys 1983;10:741–771.
51. Thomadsen B, Constantinou C, Ho A. Evaluation of water-equivalent plastics as phantom material for electron-beam dosimetry. Med Phys 1995;22:291–296.
52. Tello VM, Tailor RC, Hanson WF. How water equivalent are water-equivalent solid materials for output calibration of photon and electron beams? Med Phys 1995;22:1177–1189.
53. Liu L, Prasad SC, Bassano DA. Evaluation of two water-equivalent phantom materials for output calibration of photon and electron beams. Med Dosim 2003;28:267–269.
54. Maryanski MJ, Gore JC, Kennan RP, Schulz RJ. NMR relaxation enhancement in gels polymerized and cross-linked by ionizing radiation: a new approach to 3D dosimetry by MRI. Magn Reson Imaging 1993;11:253–258.
55. Low DA, et al. Evaluation of polymer gels and MRI as a 3-D dosimeter for intensity- modulated radiation therapy. Med Phys 1999;26:1542–1551.
56. Scheib SG, Gianolini S. Three-dimensional dose verification using BANG gel: a clinical example. J Neurosurg 2002;97:582–587.
57. Fraass B, et al. American Association of Physicists in Medicine Radiation Therapy Committee Task Group 53: quality assurance for clinical radiotherapy treatment planning Med Phys 1998;25:1773–1829,.
58. Ramani R, Ketko MG, O'Brien PF, Schwartz ML. A QA phantom for dynamic stereotactic radiosurgery: quantitative measurements. Med Phys 1995;22:1343–1346.
59. Craig T, Brochu D, Van Dyk J. A quality assurance phantom for three-dimensional radiation treatment planning. Int J Radiat Oncol Biol Phys 1999;44:955–966.
60. CIRS. Phantoms, Compurerized Imaging Reference Systems, Inc., Norfolk, (VA); 2005.
61. Curry TS, Dowdey JE, Murry RCJ. Chirstensen's Physics of Diagnostic Radiology. 4th ed. Philadelphia: Lea&Febiger; 1990.
62. Constantinou C, Cameron J, DeWerd L, Liss M. Development of radiographic chest phantoms. Med Phys 1986;13:917–921.
63. Constantinou C, Harrington JC, DeWerd LA. An electron density calibration phantom for CT-based treatment planning computers. Med Phys 1992;19:325–327.
64. Sorenson JA, Phelps ME. Physics in Nuclear Medicine. 2nd ed. Philadelphia: W.B.Saunders; 1987.
65. D'Souza WD, et al. Tissue mimicking materials for a multi-imaging modality prostate phantom. Med Phys 2001;28:688–700.
66. Lanza R, Langer R, Chick W. Principles of Tissue Engineering. New York: Academic Press; 1997.
67. NLM. The Visible Human Project. [Online]. Available at http://www.nlm.nih.gov/research/visible/visible_human.html. Accessed 2003 Sept 11.

See also COBALT 60 UNITS FOR RADIOTHERAPY; RADIATION DOSE PLANNING, COMPUTER-AIDED; RADIATION DOSIMETRY, THREE-DIMENSIONAL.

PHARMACOKINETICS AND PHARMACODYNAMICS

PAOLO VICINI
University of Washington
Seattle, Washington

INTRODUCTION

Drug discovery and development is among the most resource intensive private or public ventures. The Pharmaceutical Research and Manufacturers Association (PhRMA) reported that in 2001, pharmaceutical companies spent $\sim$ \$30.5 billion in R&D, 36% of which was allocated to preclinical functions (1). Given the expense and risk associated with clinical trials, it makes eminent sense to exploit the power of computer models to explore possible scenario of, say, a given dosing regimen or inclusion–exclusion criteria before the trial is actually run. If anything, this goes along the lines of what is already commonly done in the aerospace industry, for example. Thus, computer simulation is a relatively inexpensive way to run plausible scenarios *in silico* and try and select the best course of action before investing time and resources in a (sequence of) clinical trial(s). Ideally, this approach would integrate information from multiple sources, such as *in vitro* experiments and preclinical databases, and that is where the difficulties specific to this field start.

System analysis (in the engineering sense) is at the foundation of computer simulation. A rigorous quantification of the phenomena being simulated is necessary for this technology to be applicable. Against this background, the quantitative study of drugs and their behavior in humans and animals has been characterized as pharmacometrics, the unambiguous quantitation (via data analysis or modeling) of pharmacology (drug action and biodistribution). In a very concrete sense, pharmacometrics is the quantitative study of exposure-response (2), or the systematic relationship between drug dosage (or exposure to an agent) and drug effect (or the consequences of agent exposure on the organism). Historically, there have been two main areas of focus of pharmacometrics (3). Pharmacokinetics (PK) is the study of drug biodistribution, or more specifically of absorption, distribution, metabolism, and elimination of xenobiotics; it is often characterized as "what the body does to the drug". Pharmacodynamics (PD) is concerned with the effect of drugs, which can be construed both in terms of efficacy and toxicity. This aspect is often characterized as "what the drug does to the body".

It has to be kept in mind that both pharmacokinetic and pharmacodynamic systems are "dynamic" systems, in the sense that they can be modeled using differential equations (thus, to an engineer, this distinction may seem

a bit unusual). In addition, a third aspect of drug action will be discussed, disease progression, in the rest of this article.

The joint study of the biodistribution and the efficacy of a compound, or a class of compounds, is termed PK-PD, to signify the inclusion of both aspects. While the PK subsystem is a map from the dosing regimen to the subsequent time course of drug concentration in plasma, the PD subsystem is a map from the concentration magnitudes attained in plasma to the drug effect. This paradigm naturally postulates that the time course in plasma is a mediator for the ultimate drug effect, that is, that there is a causal relationship between the plasma time course and the effect time course. As seen later, this causal relationship does not exclude the presence of intermediate steps between the plasma and effect time courses (e.g., receptor binding and delayed signaling pathways).

Both PK and PD are amenable to quantitative modeling (Fig. 1). In fact, there is a growing realization that the engineering principles of system analysis, convolution, and identification have always been (sometimes implicitly) an integral part of the study of new drugs, and that realization is now giving rise to strategies that aim at accelerating drug development through computer simulation of expected drug biodistribution and efficacy time courses. Interestingly, the intrinsic complexity of PK-PD systems has sometimes motivated the applications of generalizations of standard technology, (e.g., the convolution integral) (4),

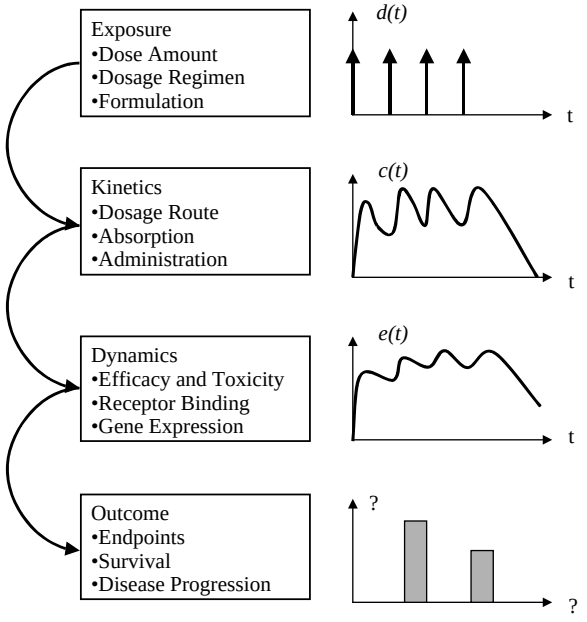


Figure 1. This figure briefly summarizes the functional pathways that exist from dosing (exposure) to outcome through drug biodistribution (pharmacokinetics) and drug effect (pharmacodynamics). it is not unusual to find a disconnect between the resolution available in the dynamic measurements and the actual clinical outcome. Also, while the convolution operator can be used to map the dosing regimen $d(t)$ to the concentration-time profile $c(t)$ and then to the effect $e(t)$, this is not necessarily true when going from the dynamic behavior to clinical measurements, which are often lower resolution and/or discrete.

through sophisticated methodological contributions that have yet to impact the engineering mainstream. Lastly, the PK-PD model is currently viewed as a perfectly adequate approach not only to process, but to extract new, quantitative information from noisy measurements. As such, the model can be construed as a probe, or a measurement tool or medical instrument in its own right, with its own built-in confidence limits and design limitations (5).

Pharmacokinetics: What Shapes the Concentration–Time Curve?

The first application of mathematical models to biology is usually attributed to Teorell (6) and led to the development of the class of compartmental models whose usage is now so widespread. Teorell's work was conducted on exogenous substances, xenobiotics (e.g., drugs). Model-based analysis of exogenous substances is somewhat simpler than endogenous compounds (e.g., glucose, insulin and other hormones, where autoregulation is crucial to understanding), since endogenous fluxes that may be exquisitely sensitive to changes in circulating concentrations are absent.

Teorell's original motivation survives in modern PK-PD modeling. The PK model is typically a useful mathematical simplification of the underlying physicochemical and physiological processes. It is usually a lumped parameter, compartmental model, and describes the absorption, distribution, metabolism, and elimination (ADME) of a drug from the body (7). A practical review of simple PK models can be found, for example, in Ref. 8. As an example, the distribution of a drug in the plasma space following an intravenous injection may be well described by a monoexponential decay:

$$c(t) = C_0 e^{-\alpha t} \tag{1}$$

where $c(t)$ is the drug concentration at time t, C_0 is the concentration at time 0 and α is the decay rate. As it is straightforward to verify, this functional form of $c(t)$ is the solution to a first-order ordinary linear differential equation, or a single compartment model:

$$\begin{aligned} \dot{q}(t) &= -k_{\mathrm{el}} q(t) + D\delta(t) \\ q(0) &= 0 \\ c(t) &= \frac{q(t)}{V} \end{aligned} \tag{2}$$

where $q(t)$ is the amount of drug in the plasma space, D is the injected dose, $\delta(t)$ is the Dirac delta, and V is the drug's volume of distribution. It can be easily verified that

$$\begin{aligned} C_0 &= \frac{D}{V} \\ \alpha &= k_{\mathrm{el}} \end{aligned}$$

so that the elimination rate of the drug is equal to the decay rate of the concentration–time curve. An important pharmacokinetic parameter is the drug's clearance, or the volume cleared per unit time, which for the model above can be expressed as:

$$\mathrm{CL} = k_{\mathrm{el}} V$$

Another common pharmacokinetic parameter is the area under the concentration curve:

$$\mathrm{AUC} = \int_0^\infty c(t)dt = \frac{D}{\mathrm{CL}}$$

The AUC is probably the most common measure of exposure in pharmacokinetics, since it summarizes dosing and systematic information. The model in equation 2 could be extended to accommodate first-order absorption, for example, to model the plasma appearance of an oral or intramuscular dose as opposed to an intravenous dose:

$$\begin{aligned} \dot{q}_1(t) &= -k_\mathrm{a} q_1(t) + D\delta(t) \\ \dot{q}_2(t) &= +k_\mathrm{a} q_1(t) - k_\mathrm{el} q_2(t) \\ q_1(0) &= q_2(0) = 0 \\ c(t) &= \frac{q_2(t)}{V} \end{aligned} \tag{3}$$

where k_a is the absorption rate constant (again in units of inverse time), and $q_1(t)$ and $q_2(t)$ are the amounts in the absorption compartment and in the plasma compartment, respectively. It can be easily verified that this model also describes the convolution of a single-exponential absorption forcing function with the single exponential impulse response of the plasma space.

For the simple single compartmental model (eq. 2), the pharmacokinetic parameters can be readily estimated from the data (9), and the assumption of a mechanistic model does not affect their values. In other words, both compartmental (model-dependent) and noncompartmental (data-dependent) analyses provide the same result. However, this relatively straightforward interpretation of the eigenvalues of the system matrix in relation to the observed rate of decay is correct only for single-compartment systems such as this one. In the case when the drug diffuses in two compartments (e.g., a plasma and extravascular compartment), then its time course is described by a sum of two exponential functions:

$$c(t) = A_1 e^{-\alpha t} + A_2 e^{-\beta t} \tag{4}$$

but the corresponding ordinary linear differential equation is

$$\begin{aligned} \dot{q}_1(t) &= -(k_{10} + k_{12}) q_1(t) + k_{21} q_2(t) + D\delta(t) \\ \dot{q}_2(t) &= +k_{12} q_1(t) - k_{21} q_2(t) \\ q_1(0) &= q_2(0) = 0 \\ c(t) &= \frac{q_1(t)}{V} \end{aligned} \tag{5}$$

where the parameters are the same as before, except that now $q_1(t)$ and $q_2(t)$ are the amount of drug in the plasma and extraplasma compartments, respectively, k_{10} is the rate constant (in units of inverse time) at which the drug leaves the system, and k_{12} and k_{21} are the rate constants at which the drug is transported out of the plasma and extraplasma compartments, respectively. One can relatively easily estimate the α and β parameters describing the observed rates of decay (one slow and the other fast) of concentration; however, the algebraic relationship between those and the fractional rate constants that mechanistically describe plasma-extraplasma exchange is not trivial. The complexity of these expressions increases with increasing number of compartments (10) and rapidly grows to be daunting. The limitations of noncompartmental and compartmental analysis of pharmacokinetic data have been discussed elsewhere, so we will not cover them here (20). There are other sources of complexity in pharmacokinetics and drug metabolism, which turn into more complex expressions for the model equations required to describe the drug's fate. For example, the underlying compartmental model may be not linear in the kinetics. This could happen, for example, when the fractional rate of disappearance changes with the drug level and goes from zero order at high concentrations to first order at low concentrations, as it happens for phenytoin (11) or ethanol (12):

$$\begin{aligned} \dot{q}(t) &= -\frac{V_m}{K_m + q(t)} q(t) + D\delta(t) \\ q(0) &= 0 \\ c(t) &= \frac{q(t)}{V} \end{aligned} \tag{6}$$

where the elimination rate is not constant, rather it exhibits a saturative behavior that resembles the classic Michaelis–Menten expression from enzyme kinetics. In this case, the principle of superposition does not hold and the concentration–time curve cannot be expressed in algebraic (closed) form, and the time course is not exponential, except at values of $q(t)$ that are much smaller than K_m. There are many other types of pharmacokinetic nonlinearities, of which this is just an example. This is why most modern PK analyses directly use the differential equations when building a PK model.

Excellent historical reviews of the properties of compartmental models, especially with reference to tracer kinetics, can be found in Refs. 13–16. More modern viewpoints are available in Refs. 17,18. Perspectives from drug development are available in Refs. 7,19. A succinct and practical review of compartmental and noncompartmental methods can be found in Ref. 8. Lastly, as we mentioned, a comparison of the strengths and weaknesses of compartmental and noncompartmental approaches has been carried out in Ref. 20.

A PK model can be used to make informed predictions about localized drug distribution and dose availability to the target organ, especially with physiologically based pharmacokinetic (PBPK) and toxicokinetic (PBTK) models. An important application has always been to individualize dosing (21,22) and improve therapeutic drug monitoring (23,24), often by borrowing approaches from process engineering and automatic control theory (25,26). Physiologically based models of pharmacokinetics are becoming an integral part of many drug development programs (27), mainly because they provide a mechanistic way to scale dosing regimens between species and between protocols. This affords (at least in theory) a seamless integration between the preclinical (*in vitro* and *in vivo*, animal studies) and clinical (*in vivo*, human studies) aspects of a drug development program. Animal to human scaling is of

increasing importance in several areas of pharmacotherapy (28) and has a long and illustrious history, from the early work by Dedrick and co-workers on methotrexate (29) to a more general application (30,31) to modern experiences (32) motivated by first time in man (FTIM) dose finding (33,34). The approach is, however, not without its critics (35), mainly due to the lack of statistical evaluation that often accompanies the scaling. It is noteworthy that between-species scaling of body mass (36,37) and metabolic rate (38) has been and is currently an object of investigation outside drug development, and mechanistic findings about the origin of allometric scaling (39) have lent scientific support to the empiricism of techniques used in drug development (40,41).

Now is a good time to note that, most often, these projections are accompanied by some statistical evaluation. From the very beginnings of PK-PD, the realization that one needed to account both for variation between subjects (due to underlying biological reasons, e.g., genetic polymorphisms and environmental factors) and variation within subject (due, e.g., to intrinsic measurement error associated with the quantification of concentration time courses and efficacy levels) was particularly acute. Statistical considerations have thus always been a part of PK-PD. The role of *population analysis*, or the explicit modeling of variability sources, in the analysis of clinical trial data is discussed in more detail later. As far as other examples, statistical applications to PK-PD have proven useful in aiding the estimation of rates of disease progression (42,43), determining individually tailored dosing schemes (23,44) and resolving models too complex to identify without prior information (45). As mentioned later, a common framework often exploited in PK-PD has to do with the incorporation of population-level information (e.g., statistical distributions of model parameters) together with individual-specific information (limited concentration–time samples, e.g.).

Pharmacodynamics: The Mechanistic Link Between Drug Exposure and Effect

The PD models (46–49) can link drug effect (characterized by clinical outcomes or intermediate pharmacological response markers) with a PK model (50), through biophase distribution, biosensor process and biosignal flux (51). Formally, the simplest PD model is the so-called $E_{\max}$ model, where effect plateaus at high concentrations:

$$e(t) = \frac{E_{\max} c(t)}{EC_{50} + c(t)} \tag{7}$$

where $E_{\max}$ is maximal effect, and EC_{50} is the effective plasma concentration at which effect $e(t)$ is half-maximal. This model is sometimes modified to account for sigmoidal effect shapes by adding an exponent that varies from $\sim$ 1 to around 2:

$$e(t) = \frac{E_{\max} c^{\mathrm{H}}(t)}{EC_{50}^{\mathrm{H}} + c^{\mathrm{H}}(t)} \tag{8}$$

Often, the driver of pharmacological effect is not plasma concentration $c(t)$, but some concentration remote from plasma and delayed with respect to plasma. This representation is similar to the one used to represent delayed effect of glucose regulatory hormones such as insulin (52). The delayed effect site concentration is them modeled using the differential equation:

$$c_{\mathrm{e}}(t) = k_{\mathrm{eo}}[c(t) - c_{\mathrm{e}}(t)] \tag{9}$$

whereas the PD model now contains effect site concentration, not plasma concentration:

$$e(t) = \frac{E_{\max} c_{\mathrm{e}}(t)}{EC_{50} + c_{\mathrm{e}}(t)} \tag{10}$$

and the interpretation of the parameters is the same as before except that they are defined with reference to effect site concentration as opposed to plasma concentration. Other classes of models are the so-called indirect response models, where the drug modulates either production or degradation of the effect (response) variable. These models are particularly appealing due to their mechanistic interpretation, and have been proposed in Refs. (53) and (54), reviewed in Ref. (55), and applied in many settings, including pharmacogenomics (56).

In many ways, the PD aspect of drug development is more challenging than the PK aspects, for many different reasons. First of all, the process of gathering data to inform about the PD of a drug is more challenging. Moreover, PD measurements may be less sensitive than it would be desirable (pain levels, which are both categorical and subjective, are a good example). Lastly, the mechanism of action of the drug may not be entirely known, and thus the best choice of measurements for the PD time course may be open to debate. How should the effect of a certain drug be quantified? If the drug is a painkiller, are pain levels sufficient, or would the levels of certain chemical(s) in the brain provide a more sensitive and specific correlate of efficacy? These and other questions become very relevant, for example, when drugs are studied that exert their effect in traditionally inaccessible locations (e.g., the brain) (57).

This is the "best biomarker question", as it relates to the now classical classification of biomarkers, surrogate endpoints and clinical endpoints (58). Basically, while a biomarker is any quantitative measure of a biological process (concentration levels, pain scores, test results and the like), a surrogate endpoint is a biomarker that substitutes for a clinical endpoint (e.g., survival or remission). In other words, surrogate endpoints, when unambiguously defined, are predictive of clinical endpoints, with the added advantage of being easier to measure and usually being characterized by a more favorable time frame (59). In the United States, the Food and Drug Administration (FDA) now allows the possibility of accelerated approval based on surrogate endpoints, provided certain conditions are met (60). Against this framework, it makes eminent sense for the PK-PD model to be focused on a relevant biomarker (or surrogate endpoint) as soon as possible in the drug development process. Often, the earlier in the process, the less influence the choice of biomarker will have. In other words, when the selection of lead compounds is just starting, proof of concept may be all that is needed, but the closer one gets to the clinical trial stage, the more crucial an appropriate

choice of surrogate endpoint will be. Another way to think about this is the establishment of a causal link between therapeutic regimen and outcome. Where these concepts come together is in the (relatively novel) idea of disease progression, and how it can be monitored.

Disease Progression: How to Know Whether the Drug is Working

Recently, the contention has been made that an integral part of the understanding of PK and PD cannot prescind from, say, the background signal that is present when the drug is administered. In other words, not all patients will be subjected to therapy at the same point in time, or at the same stage in their individual progression from early to late disease stages. There is thus a growing realization that the therapeutic intervention is made against a constantly changing background of disease state, and that the outcome of the therapy may depend on the particular disease state at a given moment in time. Disease progression modeling (61) is thus the point of contact between PK and PD and the mechanistic modeling of physiology and pathophysiology that is advocated, e.g., by the Physiome Project (62,63), recently taken over by the International Union of Physiological Sciences' Physiome and Bioengineering Committee (64).

Interesting applications are starting to emerge. For example, Ref. 65 has shown that the rate of increase of bloodstream glucose concentration in Type 2 diabetic patients is ~ 0.84 mmol·L^{-1}·year^{-1} (with a sizable variation between patients of 143%), thus providing a quantitative handle on the expected deteriorating trend of overall glycemic levels in this population of patients (together with a measure of its expected patient-to-patient variability). In Alzheimer's disease, another study (66) has demonstrated a natural rate of disease progression of 6.17 ADASC units·year^{-1} (where ADASC is the cognitive component of the Alzheimer disease assessment scale). A word of caution: As often done in this area of application, the models are informed on available data. In other words, the model parameters are quantified (estimated) based on available measurements for the drug PK and PD. This poses the challenge of developing models that are not too detailed nor too simplistic, but that can be reasonably well informed by the data at hand. Clearly, a detailed mechanistic model of the disease system requires substantial detail, but a balance needs to be struck between detail and availability of independently gathered data. Visualization approaches are a recent addition to the arsenal of the drug development expert (67). This kind of mechanistic pharmacodynamics is being applied more and more often to a variety of areas: A good example of rapid development comes from anticancer agents (68), where applications of integrated, mechanistic PK-PD models that take into account the drug mechanism of action are starting to become more and more frequent (69).

Population Variation: Adding Statistical Variation to PK, PD and Disease Models

In drug development, it is of utmost interest to determine the extent of variation of PK-PD and disease progression among members of a population. Basically, this implies determining the statistics of the biomarkers of interest in a population of patients, not just in an individual, and provide these measures together with some degree of confidence. This requirement connects well with analogous epidemiological population studies (often not model-based). The estimation of variability coupled with the evaluation of the relative role of its sources (covariates), for example, demographics, anthropometric variables or genetic polymorphisms, is tackled by the discipline of population kinetics, mainly due to the pioneering work of Lewis Sheiner and Stuart Beal at UCSF (70,71). Often called also population PK-PD, it makes use of two-level hierarchical models characterized by nested variability sources, where the models' individual parameter values are not deterministic, but unknown: They instead arise from population statistical distributions (biological variation, or BSV, between subject variability). On top of this source of variability, measurement noise and other uncertainty sources are added (RUV, residual unknown variation) to the concentration or effect signals (Fig. 2). The pharmaco-statistical models that integrate BSV and RUV are often described in statistical journals as nonlinear mixed effects models.

An example that builds on those we have already presented earlier may suffice here to clarify the fundamental concept of mixed effects models. Let us extend the model just described for intravenous injection (eq. 2) to the situation when the concentration measurements are affected by noise:

$$\begin{aligned} \dot{q}(t) &= -k_{\mathrm{el}} q(t) + D\delta(t) \\ q(0) &= 0 \\ c(t) &= \frac{q(t)}{V} + \varepsilon(t) \end{aligned} \tag{11}$$

where the parameters are the same as in equation 2, except that now $\varepsilon(t)$ is a normally distributed measurement error with mean zero and variance σ^2, that is $\varepsilon(t) \in N[0, \sigma^2]$. If data about $c(t)$ are available and have been gathered in a single individual, the model parameters k_{el} and V (assuming D is known) can then be fitted to the data using weighted (9) or extended (72) least squares or some other variation of maximum likelihood, and thus individualized estimates (with confidence intervals) can be obtained (often the measurement error variance σ^2 is not known but it can also be estimated). This can be done even if only a single subject data are available, provided that the sampling is performed (at an absolute minimum) at three or more time points (since the model has two unknown structural parameters, k_{el} and V, plus σ^2).

The nonlinear mixed-effects modeling approach is an extension of what we have just seen. It takes the viewpoint that the single subject estimates are realizations of an underlying population density for the model parameters. As such, the individual values of k_{el} and V simply become realizations (samples) of this underlying density. For example, the assumption could be made that they are both distributed lognormally: in which case, $\log(k_{\mathrm{el}}) \in N(\mu_{\mathrm{kel}}, \omega^2_{\mathrm{kel}})$ and $\log(V) \in N(\mu_v, \omega^2_v)$, where the μ denote expected logarithmic values and the ω^2 denote population variances.

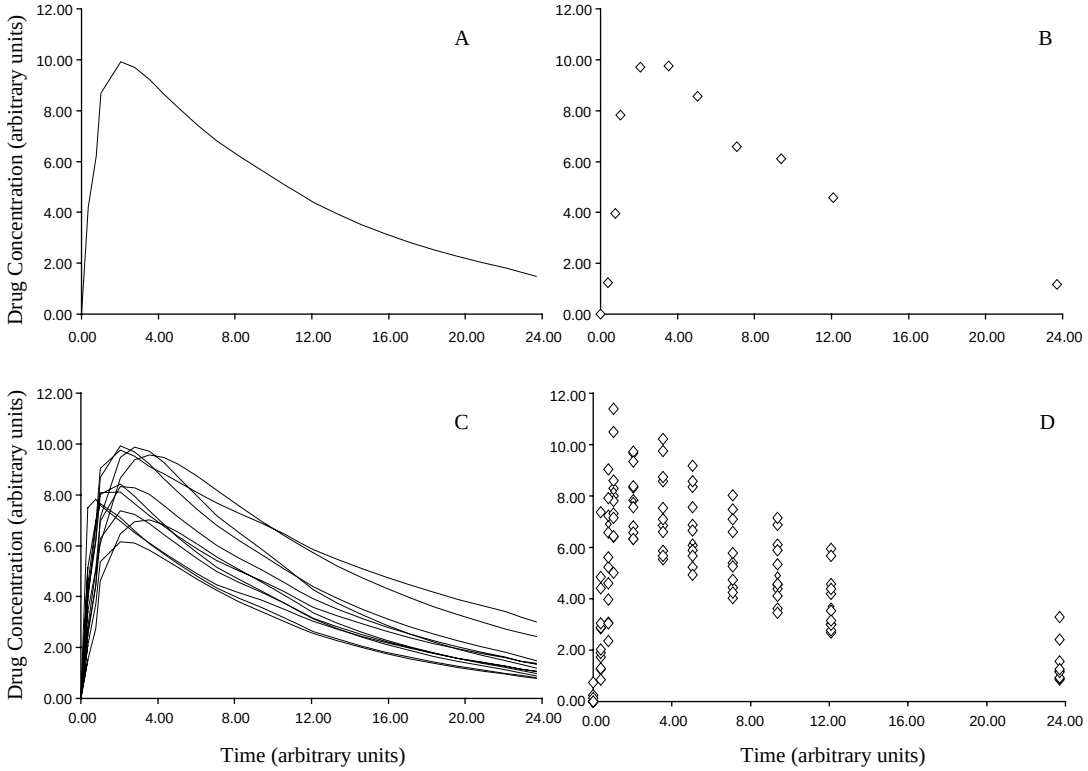


Figure 2. Nested uncertainty in pharmacokinetics. Panel A shows the true, but unknown, time course over 24 time units of a hypothetical drug administered orally to a single experimental subject. The true, but unknown, time course is a smooth function of time. Panel B shows the same smooth function sampled at discrete time points, which in practice may correspond to as many blood draws. Measurement uncertainty (currently termed RUV, residual unknown variability) is now superimposed to the true, but unknown time course in Panel A. Panel C shows the true, but unknown, time course over 24 time units of a hypothetical drug administered orally to several experimental subjects. Clearly, no time course is the same, since each subject will have a different absorption and elimination rate, together with a different volume of distribution. The time courses in Panel C are the result of BSV (between subject variability). Lastly, Panel D shows the same smooth functions in Panel c sampled at discrete time points. The RUV is again superimposed to the true, but unknown time courses, but this time the data spread is due to both BSV and RUV, and distinguishing between BSV and RUV requires a model of the system. See text for details.

The assumption is implicitly made here that k_{el} and V are uncorrelated (which may or may not be the case). In the nonlinear mixed-effects model procedure, the moments of the statistical distributions of model parameters become the new unknowns, and thus μ_{k01}, ω^2_{k01} and μ_ν, ω^2_ν are estimated by optimizing approximations of the maximum likelihood objective function expressed for the whole population of data. The value of σ^2 in the population (a composite value of the measurement error variance across subjects) can also be estimated, as before. The five distribution parameters are the fixed effects of the population (since they do not change between subject), while the individual values attained by the parameters k_{el} and V in separate individuals are random effects, since every subject has a different value (hence, the mixed-effects parlance). The approach requires data on more than one subject, ideally on many more subjects than there are fixed effects (the caveat is that it is easier to estimate expected values than it is to estimate variances or covariances, and the data needs consequently grow). Note also that, if k_{el} and V are both lognormal, clearance CL is *not* lognormal: which statistical model to choose for which parameter will depend on the available information. The main advantage of population kinetics is that, since it is estimating distributional parameters (moments), it can use relatively sparse and/or noisy data at the individual level, provided that there is a large number of population data (in other words, there can be few data for each subject, as long as there are many subjects).

This framework is quite general and powerful, and allows for modeling of complex events (e.g., adherence, or patient compliance to dosing recommendations) (73). Tutorials on this modeling approach can be found in papers (74–76), review articles (70,61,77,78), and textbooks (79). Software is also available, both for population (Aarons, 1999) and individual (Charles and Duffull, 2001) PK-PD analysis.

Mixed-effects models are used both to solve the forward problem (simulation of putative drug dosing scenarios) and the inverse problem (estimation of BSV and RUV statistics conditional on PK-PD models and clinical

measurements). Which one is of interest depends on what is available and what the intent of the study is. If the intent is to analyze data and determine the underlying distribution of PK-PD and disease parameters, then one has an inverse, or estimation, problem (80). If the intent, on the other hand, is to explore possible dosing or recruitment scenarios, then this is a direct, or simulation, problem (81). As mentioned, mixed-effects models can be applied to sparse and noisy data, as often happens in therapeutic monitoring in the clinical setting (a situation that occurs both in drug development and in applied clinical research). Their use is so widespread that the FDA recently issued one of its guidances for industry to deal with their use for population pharmacokinetic analysis (82). Interestingly, a very similar framework is also applied to evolutionary genetics, in the study of "function-valued traits" (also called "infinite-dimensional characters"). The idea is to use mixed models to link genetic information to traits that are not constant, rather are functions of time (83–85). It is important to model population genetics both for polymorphisms of drug-metabolizing enzymes affecting ADME and for polymorphisms affecting the dynamics of response.

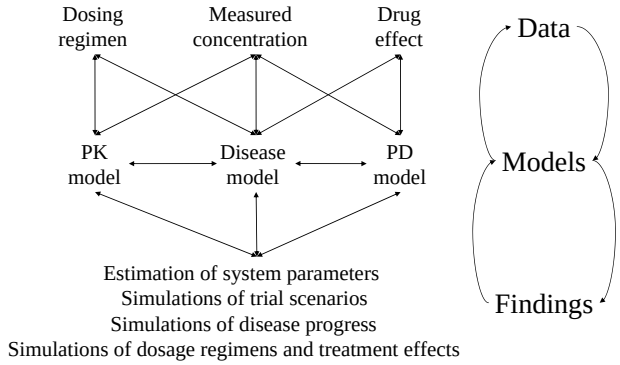


Figure 3. Information flow in a clinical trial simulation or data analysis. The doses, the measured concentration and drug effect are situated at the beginning or the end of the process (Data level). At the core of the simulation (Model level) lie models of PK, PD and disease processes. Trial designs are possible conditional on these other elements, or features like model parameters can be estimated from the trial data and measured doses and PK and PD data (Findings level). The picture is oversimplified and does not include, for example, adherence to dosing recommendations and protocol dropouts, which may be important to ensure realistic trial designs.

A Role for Modeling and Simulation

In summary, the overall goal of integrated PK-PD mathematical models is a better understanding of therapeutic intervention: Their contribution are often reflected in improved clinical trials designs. While evidence of the impact of modeling and simulation of PK-PD on the drug development process is often anectodal, many reviews of PK-PD in drug development have recently appeared (86–89), together with compelling examples of subsequent drug development acceleration (90,91).

While the fundamental algorithmic steps of computer simulation, especially Monte Carlo simulation, are well known, they require some adaptation to be used within the field of pharmacokinetics and pharmacodynamics and in the context of drug development (Fig. 3) A good practices document was issued in 1999 following a workshop held by the Center for Drug Development Science (92). The crafting of this document has contributed to clarifying the steps and tools necessary for carrying out a simulation experiment within the context of drug development. In particular, the report clarifies that simulation in drug development has an untapped potential which extends well beyond the pharmacokinetic aspect and into the pharmacodynamic and clinical domains (92).

Nowadays, pharmacometric technology (93) is used in industry and academia as a way to support and strengthen R&D. As an example, low signal/noise ratio routine clinical data obtained with sparse sampling may often be analyzed with pharmacometric techniques to determine whether a compound is metabolized differently because of phenotypic differences arising from genetic makeup, ethnicity, gender, age group (young, elderly), or concomitant medications causing drug–drug interactions (94). The role of this technology can only increase. A recently issued FDA report that focuses on the recent decrease in applications for novel therapies submitted to the agency is clear in this regard: "The concept of model-based drug development, in which pharmaco-statistical models of drug efficacy and safety are developed from preclinical and available clinical data, offers an important approach to improving drug development knowledge management and development decision making. Model-based drug development involves building mathematical and statistical characterizations of the time course of the disease and drug using available clinical data to design and validate the model. The relationship between drug dose, plasma concentration, biophase concentration (pharmacokinetics), and drug effect or side-effects (pharmacodynamics) is characterized, and relevant patient covariates are included in the model. Systematic application of this concept to drug development has the potential to significantly improve it" (95).

In summary, modern-day drug development displays a need for information integration at the whole-system, cellular, and genomic level (96) similar to that found in integrative physiology (97) and comparative biology (98,99). As mentioned earlier, simulation of clinical trials is a burgeoning discipline well-founded upon engineering and statistics (81): examples have appeared in clinical pharmacology (100,101) and pharmacoeconomics (102). Drug candidate selection is another application, possibly through PK models of varying complexity (103) and high throughput screening coupled with PK (104). Next, integrated models allow to link genomic information with disease biomarkers and phenotypes, such as in the Luo-Rudy model of cardiac excitation (105).

As a concluding remark, progress in the development of plausible, successful, and powerful data analysis methods has already had a substantial payoff, and can be substantially accelerated by encouraging multidisciplinary, multi-institutional collaboration bringing together investigators at multiple facilities and providing the infrastructure to support their research, thus allowing the timely and cost-effective expansion of new technologies. The need is more

and more often voiced for increased training of quantitative scientists in biologic research as well as in statistical methods and modeling to ensure that there will be an adequate workforce to meet future research needs (106).

BIBLIOGRAPHY

1. Marasco C. Surge In Pharmacometrics Demand Leads to New Master's Program. JobSpectrum.org – a service of the American Chemical Society. Available at http://www.cen-chemjobs.org/job_weekly31802.html, 2002.
2. Abdel-Rahman SM Kauffman RE. The integration of pharmacokinetics and pharmacodynamics: understanding dose-response. Annu Rev Pharmacol Toxicol 2004;44:111–136.
3. Rowland M, Tozer TN. Clinical Pharmacokinetics: Concepts and Applications. 3rd Ed. Baltimore: Williams & Wilkins; 1995.
4. Verotta D. Volterra series in pharmacokinetics and pharmacodynamics. J Pharmacokinet Pharmacodyn. 2003;30(5): 337–362.
5. Vicini P, Gastonguay MR, Foster DM. Model-based approaches to biomarker discovery and evaluation: a multidisciplinary integrated review. Crit Rev Biomed Eng 2002;30 (4–6):379–418.
6. Teorell T. Kinetics of distribution of substances administered to the body. I. The extravascular modes of administration. II. The intravascular modes of administration. Arch Int Pharmacodyn Ther. 1937; 57: 205–240.
7. Gibaldi M, Perrier D. Pharmacokinetics, 2nd Ed. New York: Marcel Dekker; 1982.
8. Jang GR, Harris RZ, Lau DT. Pharmacokinetics and its role in small molecule drug discovery research. Med Res Rev 2001;21(5):382–396.
9. Landaw EM, DiStefano JJ 3rd. Multiexponential, multicompartmental, and noncompartmental modeling. II. Data analysis and statistical considerations. Am J Physiol 1984;246(5 Pt. 2):R665–77.
10. DiStefano JJ 3rd, Landaw EM. Multiexponential, multicompartmental, and noncompartmental modeling. I. Methodological limitations and physiological interpretations. Am J Physiol 1984;246(5 Pt. 2):R651–664.
11. Grasela TH, et al. Steady-state pharmacokinetics of phenytoin from routinely collected patient data. Clin Pharmacokinet 1983;8(4):355–364.
12. Holford NH. Clinical pharmacokinetics of ethanol. Clin Pharmacokinet, 1987;13(5):273–292.
13. Anderson DH. Compartmental Modeling and Tracer Kinetics. Lecture Notes in Biomathematics, Vol. 50, Berlin: Springer-Verlag; 1983.
14. Godfrey K. Compartmental Models and Their Application. New York: Academic; 1983.
15. Jacquez JA. Compartmental Analysis in Biology and Medicine, 2nd ed. Michigan: University of Michigan Press; 1985.
16. Carson ER, Cobelli C, Finkelstein L. The Mathematical Modeling of Endocrine-Metabolic Systems. Model Formulation, Identification and Validation. New York: Wiley; 1983.
17. Carson E, Cobelli C. Modelling Methodology for Physiology and Medicine. San Diego: Academic Press; 2000.
18. Cobelli C, Foster D, Toffolo G. Tracer Kinetics in Biomedical Research: From Data to Model. London: Kluwer Academic/Plenum; 2001.
19. Atkinson AJ, et al., eds. Principles of Clinical Pharmacology. San Diego: Academic Press; 2001.
20. DiStefano JJ 3rd. Noncompartmental vs. compartmental analysis: some bases for choice. Am J Physiol 1982;243(1): R1–6.
21. Goicoechea FJ, Jelliffe RW. Computerized dosage regimens for highly toxic drugs. Am J Hosp Pharm 1974;31(1):67–71.
22. Sheiner LB, Beal S, Rosenberg B, Marathe VV. Forecasting individual pharmacokinetics. Clin Pharmacol Ther 1979;26(3):294–305.
23. (a) Jelliffe RW, et al. Individualizing drug dosage regimens: roles of population pharmacokinetic and dynamic models, Bayesian fitting, and adaptive control. Ther Drug Monit 1993;15(5):380–393. (b) Jelliffe RW. Clinical applications of pharmacokinetics and adaptive control. IEEE Trans Biomed Eng 1987;34(8):624–632.
24. (a) Jelliffe RW, et al. Adaptive control of drug dosage regimens: basic foundations, relevant issues, and clinical examples. Int J Biomed Comput 1994;36(1–2):1–23. (b) Jelliffe RW, Schumitzky A. Modeling, adaptive control, and optimal drug therapy. Med Prog Technol 1990;16(1–2):95–110.
25. Jelliffe RW. Clinical applications of pharmacokinetics and adaptive control. IEEE Trans Biomed Eng 1987;34(8):624–632.
26. Jelliffe RW, Schumitzky A. Modeling, adaptive control, and optimal drug therapy. Med Prog Technol 1990;16(1–2):95–110.
27. Parrott N, Jones H, Paquereau N, Lave T. Application of full physiological models for pharmaceutical drug candidate selection and extrapolation of pharmacokinetics to man. Basic Clin Pharmacol Toxicol 2005;96(3):193–199.
28. Gallo JM, et al. Pharmacokinetic model-predicted anticancer drug concentrations in human tumors. Clin Cancer Res 2004;10(23):8048–8058.
29. Dedrick R, Bischoff KB, Zaharko DS. Interspecies correlation of plasma concentration history of methotrexate (NSC-740). Cancer Chemother Rep 1970;54(2):95–101.
30. Dedrick RL. Animal scale-up. J Pharmacokinet Biopharm 1973;1(5):435–461.
31. Dedrick RL, Bischoff KB. Species similarities in pharmacokinetics. Fed Proc 1980;39(1):54–59.
32. Mahmood I, Balian JD. Interspecies scaling: predicting clearance of drugs in humans. Three different approaches. Xenobiotica 1996;26(9):887–895.
33. Mahmood I, Green MD, Fisher JE. Selection of the first-time dose in humans: comparison of different approaches based on interspecies scaling of clearance. J Clin Pharmacol 2003;43(7):692–697.
34. Iavarone L, et al. First time in human for GV196771: interspecies scaling applied on dose selection. J Clin Pharmacol 1999;39(6):560–566.
35. Bonate PL, Howard D. Prospective allometric scaling: does the emperor have clothes? J Clin Pharmacol 2000;40(6):665–670. discussion 671–676.
36. West GB, Brown JH, Enquist BJ. A general model for the origin of allometric scaling laws in biology. Science 1997;276(5309):122–126.
37. (a) Iavarone L, et al. First time in human for GV196771: interspecies scaling applied on dose selection. J Clin Pharmacol 1999;39(6):560–566. (b) West GB, Brown JH, Enquist BJ. The fourth dimension of life: fractal geometry and allometric scaling of organisms. Science 1999;284(5420):1677–1679.
38. Gillooly JF, et al. Effects of size and temperature on metabolic rate. Science 2001;293(5538):2248–2251. Erratum in Science 2001;294(5546):1463.
39. White CR, Seymour RS. Mammalian basal metabolic rate is proportional to body mass2/3. Proc Natl Acad Sci USA 2003;100(7):4046–4049.
40. Anderson BJ, Woollard GA, Holford NH. A model for size and age changes in the pharmacokinetics of paracetamol in neonates, infants and children. Br J Clin Pharmacol 2000;50(2): 125–134.

41. van der Marel CD, et al. Paracetamol and metabolite pharmacokinetics in infants. Eur J Clin Pharmacol 2003;59(3): 243–251.
42. Craig BA, Fryback DG, Klein R, Klein BE. A Bayesian approach to modelling the natural history of a chronic condition from observations with intervention. Stat Med 1999;18(11):1355–1371.
43. Mc Neil AJ. Bayes estimates for immunological progression rates in HIV disease. Stat Med 1997;16(22):2555–2572.
44. Sheiner LB, Beal SL. Bayesian individualization of pharmacokinetics: simple implementation and comparison with non-Bayesian methods. J Pharm Sci 1982;71(12):1344–1348.
45. Cobelli C, Caumo A, Omenetto M. Minimal model SG overestimation and SI underestimation: improved accuracy by a Bayesian two-compartment model. Am J Physiol 1999;277 (3 Pt.1):E481–488.
46. Segre G. Kinetics of interaction between drugs and biological systems. Farmaco [Sci] 1968;23(10):907–918.
47. Dahlstrom BE, Paalzow LK, Segre G, Agren AJ. Relation between morphine pharmacokinetics and analgesia. J Pharmacokinet Biopharm 1978 6(1):41–53.
48. Holford NH, Sheiner LB. Pharmacokinetic and pharmacodynamic modeling in vivo. CRC Crit Rev Bioeng 1981a;5(4): 273–322.
49. Holford NH, Sheiner LB. Understanding the dose-effect relationship: clinical application of pharmacokinetic-pharmacodynamic models. Clin Pharmacokinet 1981b; 6(6):429–453.
50. Holford NH, Sheiner LB. Kinetics of pharmacologic response. Pharmacol Ther 1982;16(2):143–166.
51. Jusko WJ. Pharmacokinetics and receptor-mediated pharmacodynamics of corticosteroids. Toxicology 1995;102(1–2): 189–196.
52. Bergman RN, Phillips LS, Cobelli C. Physiologic evaluation of factors controlling glucose tolerance in man: measurement of insulin sensitivity and beta-cell glucose sensitivity from the response to intravenous glucose. J Clin Invest 1981;68(6): 1456–1467.
53. (a) Dayneka NL, Garg V, Jusko WJ. Comparison of four basic models of indirect pharmacodynamic responses. J Pharmacokinet Biopharm 1993;21(4):457–478. (b) Ramakrishnan R, et al. Pharmacodynamics and pharmacogenomics of methylprednisolone during 7-day infusions in rats. J Pharmacol Exp Ther 2002;300(1):245–256.
54. Jusko WJ, Ko HC. Physiologic indirect response models characterize diverse types of pharmacodynamic effects. Clin Pharmacol Ther 1994;56(4):406–419.
55. Sharma A, Jusko WJ. Characteristics of indirect pharmacodynamic models and applications to clinical drug responses. Br J Clin Pharmacol 1998;45(3):229–239.
56. Ramakrishnan R, et al. Pharmacodynamics and pharmacogenomics of methylprednisolone during 7-day infusions in rats.J Pharmacol Exp Ther 2002;300(1):245–256.
57. Bieck PR, Potter WZ. Biomarkers in psychotropic drug development: integration of data across multiple domains. Annu Rev Pharmacol Toxicol 2005;45:227–246.
58. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 2001;69(3):89–95.
59. Rolan P, Atkinson AJ Jr., Lesko LJ. Use of biomarkers from drug discovery through clinical practice: report of the Ninth European Federation of Pharmaceutical Sciences Conference on Optimizing Drug Development. Clin Pharmacol Ther 2003;73(4):284–291.
60. The Food and Drug Modernization Act of 1997. Title 21 Code of Federal Regulations Part 314 Subpart H Section 314.500.
61. Chan PL, Holford NH. Drug treatment effects on disease progression. Annu Rev Pharmacol Toxicol 2001;41:625–659.
62. Bassingthwaighte JB. The macro-ethics of genomics to health: the physiome project. C R Biol 2003;326(10–11): 1105–1110.
63. Hunter PJ, Borg TK. Integration from proteins to organs: the Physiome Project. Nat Rev Mol Cell Biol 2003;4(3):237–243.
64. Hunter PJ. The IUPS Physiome Project: a framework for computational physiology. Prog Biophys Mol Biol 2004; 85(2–3):551–569.
65. Frey N, et al. Population PKPD modelling of the long-term hypoglycaemic effect of gliclazide given as a once-a-day modified release (MR) formulation. Br J Clin Pharmacol 2003;55(2):147–157.
66. Holford NH, Peace KE. Methodologic aspects of a population pharmacodynamic model for cognitive effects in Alzheimer patients treated with tacrine. Proc Natl Acad Sci USA 1992;89(23):11466–11470.
67. Bhasi K, Zhang L, Zhang A, Ramanathan M. Analysis of pharmacokinetics, pharmacodynamics, and pharmacogenomics data sets using VizStruct, a novel multidimensional visualization technique. Pharm Res 2004;21(5):777–780.
68. Friberg LE, Karlsson MO. Mechanistic models for myelosuppression. Invest New Drugs 2003;21(2):183–194.
69. Karlsson MO, et al. Pharmacokinetic/pharmacodynamic modelling in oncological drug development. Basic Clin Pharmacol Toxicol 2005;96(3):206–211.
70. Beal SL, Sheiner LB. Estimating population kinetics. Crit Rev Biomed Eng 1982;8(3):195–222.
71. Sheiner LB, Ludden TM. Population pharmacokinetics/dynamics. Annu Rev Pharmacol Toxicol 1992;32:185–209.
72. Peck CC, Beal SL, Sheiner LB, Nichols AI. Extended least squares nonlinear regression: a possible solution to the 'choice of weights" problem in analysis of individual pharmacokinetic data. J Pharmacokinet Biopharm 1984;12(5): 545–558.
73. Girard P, et al. A Markov mixed effect regression model for drug compliance. Stat Med 1998;17(20):2313–2333.
74. Sheiner LB. Analysis of pharmacokinetic data using parametric models-1: Regression models. J Pharmacokinet Biopharm 1984;12(1):93–117.
75. (a) Landaw EM, DiStefano JJ 3rd. Multiexponential, multicompartmental, and noncompartmental modeling. II. Data analysis and statistical considerations. Am J Physiol 1984; 246(5 Pt. 2):R665–77. (b) Sheiner LB. Analysis of pharmacokinetic data using parametric models. II. Point estimates of an individual's parameters. J Pharmacokinet Biopharm 1985; 13(5):515–540.
76. Sheiner LB. Analysis of pharmacokinetic data using parametric models. III. Hypothesis tests and confidence intervals. J Pharmacokinet Biopharm 1986;14(5):539–555.
77. Ette EI, Williams PJ. Population pharmacokinetics II: estimation methods. Ann Pharmacother 2004;38(11):1907–1915.
78. Ette EI, Williams PJ. Population pharmacokinetics I: background, concepts, and models. Ann Pharmacother 2004; 38(10):1702–1706.
79. Davidian M,and Giltinan DM. Nonlinear Models for Repeated Measurement Data. Boca Raton: Chapman and Hall/CRC; 1995.
80. Sheiner L, Wakefield J. Population modelling in drug development. Stat Methods Med Res 1999;8(3):183–193.
81. Holford NH, Kimko HC, Monteleone JP, Peck CC. Simulation of clinical trials. Annu Rev Pharmacol Toxicol 2000;40:209–234.

82. United States Food and Drug Administration, Department of Health and Human Services. Guidance for Industry: Population Pharmacokinetics. Availabel at http://www.fda.gov/cder/guidance/1852fnl.pdf, 1999.
83. Kirkpatrick M, Lofsvold D. The evolution of growth trajectories and other complex quantitative characters. Genome 1989;31(2):778–783.
84. Pletcher SD, Geyer CJ. The genetic analysis of age-dependent traits: modeling the character process. Genetics 1999; 153(2):825–835.
85. (a) Ma CX, Casella G, Wu R. Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. Genetics 2002;161(4):1751–1762. (b) Sheiner LB, Beal S, Rosenberg B, Marathe VV. Forecasting individual pharmacokinetics. Clin Pharmacol Ther 1979;26(3): 294–305.
86. Rolan P. The contribution of clinical pharmacology surrogates and models to drug development—a critical appraisal. Br J Clin Pharmacol 1997;44(3):219–225.
87. Sheiner LB, Steimer JL. Pharmacokinetic/pharmacodynamic modeling in drug development. Annu Rev Pharmacol Toxicol 2000;40:67–95.
88. Aarons L, et al. COST B15 Experts. Role of modelling and simulation in Phase I drug development. Eur J Pharm Sci 2001;13(2):115–122.
89. Blesch KS, et al. Clinical pharmacokinetic/pharmacodynamic and physiologically based pharmacokinetic modeling in new drug development: the capecitabine experience. Invest New Drugs 2003;21(2):195–223.
90. Piscitelli SC, Peck CC. Pharmacokinetic and pharmacodynamic methods in biomarker development and application. In: Downing GJ, editor. Biomarkers and Surrogate Endpoints: Clinical Research and Applications. New York: Elsevier; 2000. pp. 27–35.
91. Lesko LJ, Rowland M, Peck CC, Blaschke TF. Optimizing the science of drug development: opportunities for better candidate selection and accelerated evaluation in humans. Pharm Res 2000;17(11):1335–1344.
92. Holford NH, et al. Simulation in Drug Development: Good Practices. Draft Publication of the Center for Drug Development Science (CDDS) Draft version 1.0, July 23, 1999, Available at http://cdds.georgetown.edu/research/sddgp723.html
93. Sun H, et al. Population pharmacokinetics. A regulatory perspective. Clin Pharmacokinet 1999;37(1):41–58.
94. Krecic-Shepard ME et al. Race and sex influence clearance of nifedipine: results of a population study. Clin Pharmacol Ther 2000;68(2):130–142.
95. United States Food and Drug Administration, Department of Health and Human Services. Innovation/Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products. Available at http://www.fda.gov/oc/initiatives/criticalpath/, 2004.
96. Mamiya K, et al. The effects of genetic polymorphisms of CYP2C9 and CYP2C19 on phenytoin metabolism in Japanese adult patients with epilepsy: studies in stereoselective hydroxylation and population pharmacokinetics. Epilepsia 1998; 39(12):1317–1323.
97. Bassingthwaighte JB. Strategies for the physiome project. Ann Biomed Eng 2000;28(8):1043–1058.
98. Davidson EH, et al. A genomic regulatory network for development. Science 2002;295(5560):1669–1678.
99. von Dassow G, Meir E, Munro EM, Odell GM. The segment polarity network is a robust developmental module. Nature (London) 2000;406(6792):188–192.
100. Kimko HC, Reele SS, Holford NH, Peck CC. Prediction of the outcome of a phase 3 clinical trial of an antischizophrenic agent (quetiapine fumarate) by simulation with a population pharmacokinetic and pharmacodynamic model. Clin Pharmacol Ther 2000;68(5):568–577.
101. Nestorov I, et al. Modeling and stimulation for clinical trial design involving a categorical response: a phase II case study with naratriptan. Pharm Res 2001;18(8):1210–1219.
102. Hauber AB, et al. Potential savings in the cost of caring for Alzheimer's disease. Treatment with rivastigmine. Pharmacoeconomics 2000;17(4):351–360.
103. (a) Jang GR, Harris RZ, Lau DT. Pharmacokinetics and its role in small molecule drug discovery research. Med Res Rev 2001;21(5):382–396. (b) Roberts SA. High-throughput screening approaches for investigating drug metabolism and pharmacokinetics. Xenobiotica 2001;31(8–9):557–589.
104. Roberts SA. High-throughput screening approaches for investigating drug metabolism and pharmacokinetics. Xenobiotica 2001;31(8–9):557–589.
105. Rudy Y. From genome to physiome: integrative models of cardiac excitation. Ann Biomed Eng 2000;28(8):945–950.
106. De Gruttola VG, et al. Considerations in the evaluation of surrogate endpoints in clinical trials. Summary of a National Institutes of Health workshop. Control Clin Trials 2001; 22(5):485–502.
107. Charles BG, Duffull SB. Pharmacokinetic software for the health sciences: choosing the right package for teaching purposes. Clin Pharmacokinet 2001;40(6):395–403.

See also DRUG DELIVERY SYSTEMS; DRUG INFUSION SYSTEMS; RADIOPHARMACEUTICAL DOSIMETRY; TRACER KINETICS.

PHONOCARDIOGRAPHY

HERMAN VERMARIEN
Vrije Universiteit Brussel
Brussel, Belgium

INTRODUCTION

Mechanical heart action is accompanied by audible noise phenomena, which are easy to perceive when the ear is placed next to a person's chest wall. These cardiovascular sounds can be designated as being weak in comparison with other physiological sounds, such as speech, stomach and intestine rumbling, and even respiration noises. In fact, the latter can be heard at a certain distance from the subject, which is not true for heart noises (provided one overlooks cases of artificial heart valves). The frequency content of heart sounds is situated between 20 and 1000 Hz, the lower limit being set by the ability of human hearing; Mechanical valve prostheses may largely exceed the upper limit. Examination of cardiovascular sounds for diagnostic purposes through the human hearing sense, that is, auscultation, has been commonly practiced for a long time (1–5). The only technology involved is the stethoscope, establishing a closed air compartment between a part of the person's chest surface and the physician's ear orifice. This investigation method, however, being completely psychophysical and thus subjective, has proved its benefit and continues to be an important tool in cardiovascular diagnosis.

Phonocardiography (PCG) may simply be defined as the method for obtaining recordings of cardiovascular sound, that is, the phenomena perceivable by auscultation. The origins of the method are strongly anchored in auscultation. The recordings of sounds are evaluated, on paper

or computer screen, possibly in the presence of other synchronous signals (e.g., the electrocardiogram) (ECG), partly psychophysically with another human sense, the eye, in examining waveform patterns and their relation with the other signals. Phonocardiographic signals are examined with respect to the occurrence of pathological patterns, relative intensities and intensity variations, timing and duration of events. Evidently more objective evaluation can be performed ranging from simple accurate timing of phenomena to advanced waveform analysis and comparing recorded results with waveforms from data banks. The importance of auscultation can be explained by the simplicity of the technique and by the strong abilities of the ear with respect to pattern recognition in acoustic phenomena. For obtaining equivalent information with phonocardiography, a single recording fails to be sufficient: A set of frequency filtered signals, each of them emphasizing gradually higher frequency components (by using high pass or band-pass filters), is needed. In this way, visual inspection of sound phenomena in different frequency ranges, adapted by a compensating amplification for the intensity falloff of heart sounds toward higher frequencies, is made possible, thus rendering the method equivalent with hearing performance: pattern recognition abilities and increasing sensitivity toward higher frequencies (within the above mentioned frequency range).

Laennec (1781–1826) was the first to listen to the sounds of the heart, not only directly with his ear to the chest, he also invented the stethoscope and provided the basis of contemporary auscultation. As physiological knowledge increased through the following decades, faulty interpretations of heart sounds were progressively eliminated. The first transduction of heart sounds was made by Hürthle (1895), who connected a microphone to a frog nerve-muscle preparation. Einthoven (1907) was the first to record phonocardiograms with the aid of a carbon microphone and a string galvanometer (6). Different investigators were involved in the development of filters to achieve a separation of frequency phenomena, as the vacuum tube, and thus electronic amplification became available. The evolution of PCG is strongly coupled with auscultatory findings and the development was predominantly driven by clinicians. A result of this situation is that a large variety of apparatus has been designed, mostly according to the specific needs of a clinic or the scientific interests of a medical researcher. During the 1960s, the necessity for standardization was strongly felt. Standardization committees made valuable proposals (7–9) but the impact on clinical phonocardiographic apparatus design was limited.

During the 1970s and the beginning of the 1980s, fundamental research on physical aspects of recording, genesis, and transmission of heart sound was performed (10–12) which, together with clinical investigations, improved the understanding of the heart sound phenomena. At the same time, ultrasonic methods for heart investigation became available and gradually improved. Doppler and echocardiography provided information closer related to heart action in terms of heart valve and wall movement, and blood velocity. Moreover, obtaining high quality recordings of heart sound with a high signal-to-noise ratio is difficult. Hampering elements are the inevitable presence of noise (background noise, respiration noise, muscle tremors, stomach rumbling), nonoptimal recording sites, weak sounds (obese patients), and so on. Thus interest in PCG gradually decreased.

In describing the state of the art, PCG is usually compared with ECG, the electrical counterpart, also a noninvasive method. The ECG, being a simple recording of electrical potential differences, was easily standardized, thus independent of apparatus design and completely quantitative with the millivolt scale on its ordinate axis. Phonocardiography has not reached the same level of standardization, remains apparatus dependent, and thus semiquantitative. Nowadays Doppler echocardiography and cardiac imaging techniques largely exceed the possibilities of PCG and make it redundant for clinical diagnosis. Whereas auscultation of cardiac sounds continues to be of use in clinical diagnosis, PCG is now primarily used for teaching and training purposes and for research. As a diagnostic method, conventional PCG has historical value. Nevertheless, the electronic stethoscope (combined with PC and software), as a modern concept for PCG, may gain importance for clinical purposes.

The generation of sounds is one of the many observable mechanical effects caused by heart action: contraction and relaxation of cardiac muscle, pressure rising and falling in the heart cavities, valve opening and closure, blood flowing and discontinuation of flow. Figure 1 shows a schematic representation of typical cardiac variables: the ECG, the logic states of the heart valves, low and high frequency phonocardiograms, a recording of a vessel pulse (carotid artery), and of the heart apex pulse (apexcardiogram). The heart cycle is divided into specific intervals according to the valve states of the left heart. The left ventricular systole is composed of the isovolumic contraction and the ejection period; The left ventricular diastole covers the isovolumic relaxation and the left ventricular filling (successively, the rapid filling, the slow filling, and the atrial contraction). A similar figure could be given for the right heart; Valve phenomena are approximately synchronous with those of the left heart. Small time shifts are typical: Mitral valve closure precedes tricuspid closure and aortic valve closure precedes pulmonary closure. The low frequency PCG shows the four normal heart sounds (I, II, III, and IV); In the high frequency, trace III and IV have disappeared and splitting is visible in I and in II. In the next sections details are given on the physiological significance, the physical aspects and recording methods, processing and physical modeling of heart sounds. Special attention is given to the electronic stethoscope.

HEART SOUNDS AND MURMURS

The sounds of the normal heart can be represented by a simple onomatopoeic simulation: "...lubb-dup..." (1–5). Two sounds can clearly be identified, the first being more dull than the second. A heart sound or a heart sound component is defined as a single audible event preceded and followed by a pause. As such, "splitting of a sound" occurs as one can clearly distinguish two components separated by a small pause. The closest splitting that

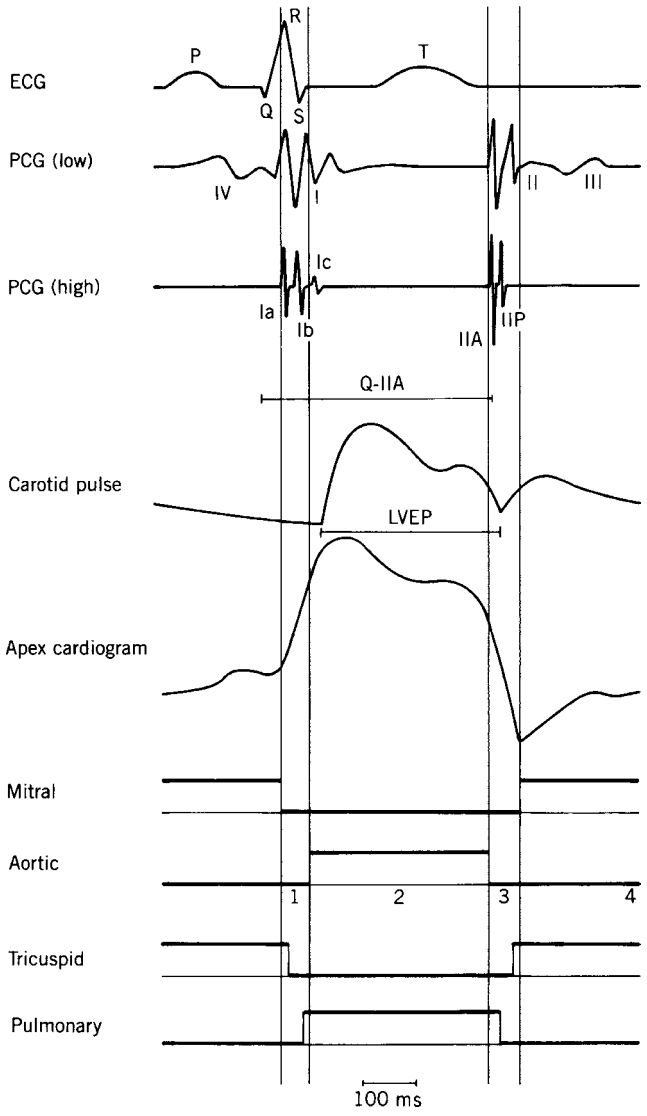


Figure 1. The ECG, PCG (low and high filtered), carotid pulse, apexcardiogram, and logic states (high = open) of left heart valves, mitral and aortic valve, and right heart valves, tricuspid and pulmonary valve. Left heart mechanical intervals are indicated by vertical lines: isovolumic contraction (1), ejection (2), isovolumic relaxation (3), and filling (4) (rapid filling, slow filling, atrial contraction). The low frequency PCG shows the four normal heart sounds (I, II, III, and IV); In the high frequency trace III and IV have disappeared and splitting is visible in I [Ia and Ib (and even a small Ic due to ejection)] and in II [IIA (aortic valve) and IIP (pulmonary valve)]. Systolic intervals LVEP (on carotid curve) and Q-IIA (on ECG and PCG) are indicated.

can be appreciated is ~20–30 ms. Similar guidelines are followed for the identification of phonocardiographic recordings: A sound is a complex of succeeding positive and negative deflections alternating with respect to the baseline, preceded and followed by a pause. A sound is said to be split if a small pause between the components can be perceived. At this point, the effect of frequency filtering may be important: Splitting, being invisible on a low frequency recording, may become recognizable on a high frequency recording (Fig. 1). Summarizing, we can state that in clinical PCG primarily the envelope of the recorded signal is regarded, not the actual waveform as, for example, in ECG, blood pressure, and velocity recordings. As spectral performance of phonocardiography may exceed the possibilities of human hearing inaudible low frequency phenomena can be recorded; They are also indicated as "(inaudible) sounds".

Acoustic phenomena originated by the heart are classified into two categories: heart sounds and heart murmurs (1–5,10–12). Although the distinction between them is not strict, one can state that heart sounds have a more transient, musical character (cf. the touching of a string) and a short duration (Fig. 1), whereas most murmurs have a predominantly noisy character and generally (but not always) a longer duration (e.g., a "blowing" murmur, a "rumbling" murmur) (Fig. 2). It is also believed that the genesis of both types is different: Heart sounds are indicated as types of resonant phenomena of cardiac structures and blood as a consequence of one or more sudden events in the cardiohemic system (such as valve closure), and most heart murmurs are said to be originated by blood flow turbulence. Many aspects of the problem of the genesis of these phenomena are still being discussed, including the relative importance of the valves and of the cardiohemic system in the generation of heart sounds (valvular theory versus cardiohemic theory).

Four normal heart sounds can be described (Fig. 1): I, II, III, and IV (also indicated as S1, S2, S3, S4). The two having the largest intensity, that is, the first (I, S1) and the second (II, S2) sound, are initially related to valve closure. The third (III, S3) and the fourth (IV, S4) sound, appearing extremely weak and dull and observable only in a restricted group of people, are not related to valve effects. The so-called closing sounds (I and II) are not originated by the coaptation of the valve leaflets (as the slamming of a door). On the contrary, it is most probably a matter of resonant-like interaction between two cardiohemic compartments suddenly separated by an elastic interface (the closed valve leaflets) interrupting blood flow: Vibration is generated at the site of the valve with a main direction perpendicular to the valve orifice plane and dependent on the rapid development of a pressure difference over the closed valve. In the case of the first sound, this phenomenon is combined with the effect of a sudden contraction of cardiac ventricular muscle. Pathologies of the cardiohemic system can affect the normal sounds with respect to intensity, frequency content, timing of components (splitting) (1).

The first heart sound (I) occurs following the closing of the mitral valve and of the tricuspid valve, during the isovolumic contraction period, and, furthermore, during the opening of the aortic valve and the beginning of ejection. In a medium or high frequency recording, a splitting of the first sound may be observed. Components related to the closing of the mitral valve (Ia, M1), the closing of the tricuspid valve (Ib, T1) and the opening of the aortic valve may be observed. There is a direct relation between the intensity of I and the heart contractility, expressed in the slope of ventricular pressure rising; with high cardiac output (exercise, emotional stress, etc.) sound I is enhanced. The duration of the PR-interval (electrical conduction time from the physiological pacemaker in the right

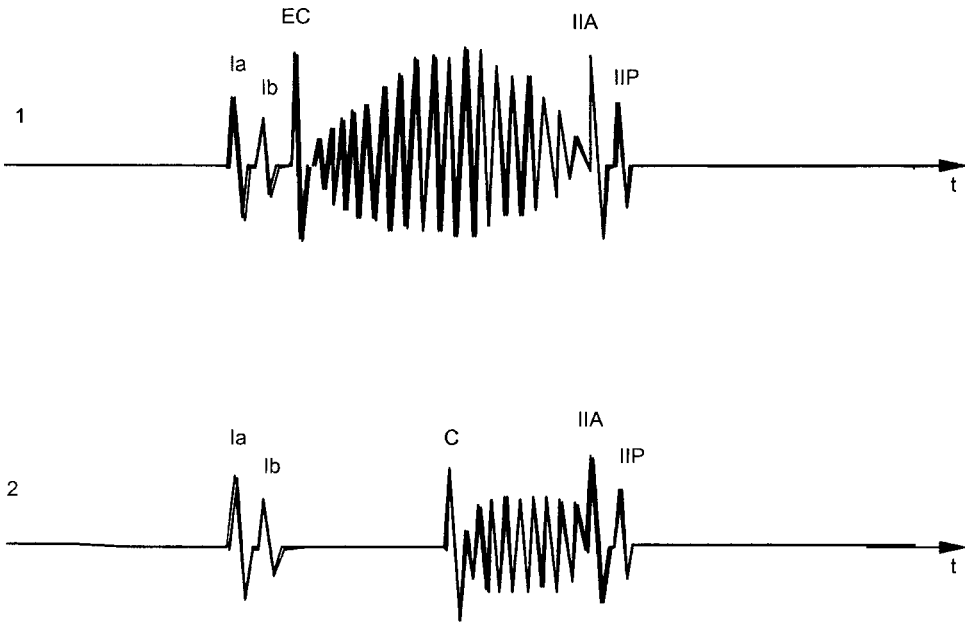


Figure 2. Examples of pathological sounds and murmurs. (1). A systolic murmur (ejection murmur, crescendo, decrescendo) as a consequence of aortic valve stenosis preceded by a clear aortic ejection click (EC). (2). Mid-systolic click (C) as a consequence of mitral valve prolapse followed by a systolic murmur due to mitral valve regurgitation.

atrium to the ventricles) is a determining factor: The shorter the time between the atrial and ventricular contraction and, consequently, the larger the distance between the mitral valve leaflets, the larger the intensity of the first sound appears. With a long PR-interval mitral valve leaflets have evolved from wide open during atrial contraction to a state of partially open to almost closed when ventricular contraction starts; the result is a weak first sound. Cardiovascular pathologies can have an effect on timing and intensities of the first heart sound components. Wide splitting is observed in right bundle branch block, tricuspid stenosis, and atrial septal defect due to a delayed tricuspid component (Ib). In left bundle branch block Ia and Ib can coincide resulting in a single sound I. A diminished sound I is found in cases of diminished contractility (myocardial infarction, cardiomyopathy, heart failure), in left bundle branch block, mitral regurgitation and aortic stenosis; An intensified sound I is found in mitral stenosis with mobile valve leaflets and in atrial septal defect.

The second sound (II) is associated with the closure of the aortic valve and, following, the closure of the pulmonary valve. Splitting of the sound in an aortic (IIA, A2) and a pulmonary (IIP, P2) component is often observed. Splitting increases during inspiration as a consequence of increased difference in duration of left and right ventricular systole caused by increased right and decreased left ventricular filling; both components may fuse together at the end of expiration. Paradoxical splitting (the pulmonary component preceding the aortic one) is pathological. The pulmonary component normally has a lower intensity; an increased intensity with respect to the aortic component is generally abnormal. There is a direct relation between the intensity and the frequency of II and the slope of ventricular pressure falling during isovolumic relaxation. Stiffening of the valve leaflets results in a reduction of II. A higher valve radius or a lowered blood viscosity gives rise to an increased second sound. Cardiovascular pathologies can have an effect on timing and intensities of the second heart sound components. Wide splitting of sound II can be due to delayed pulmonary valve closure or advanced aortic valve closure. Delayed pulmonary valve closure can be caused by right bundle branch block, pulmonary stenosis, pulmonary hypertension, atrial septal defect; advanced aortic valve closure can result from mitral regurgitation and ventricular septal defect. Paradoxical splitting of sound II can be due to delayed aortic valve closure or advanced pulmonary valve closure. Delayed aortic valve closure can be caused by left bundle branch block, aortic stenosis and arteriosclerotic heart disease. Advanced pulmonary valve closure can be caused by tricuspid regurgitation and advanced right ventricular activation. IIA, respectively, IIP, can be absent in severe aortic, respectively, pulmonary, valve stenosis. IIA is decreased in aortic regurgitation and in pathologically diminished left ventricular performance.

The third sound (III) occurs during the rapid passive filling period of the ventricle. It is believed that III is initiated by the sudden deceleration of blood flow when the ventricle reaches its limit of distensability, causing vibrations of the ventricular wall. It can often be heard in normal children and adolescents, but can also be registered in adults (although not heard) in the low frequency channel. It is a weak and low pitched (low frequency) sound. Disappearance of III is a result of aging as a consequence of increasing myocardial mass having a damping effect on vibrations. High filling rate or altered physical properties of the ventricle may cause an increased third sound. If III reappears with aging (beyond the age of 40 years) it is pathological in most cases. A pathological sound III is found in mitral regurgitation, aortic stenosis, ischemic heart disease.

The fourth sound (IV) coincides with the atrial contraction and thus the originated increased blood flow through the mitral valve with consequences as mentioned for the third sound. It is seldom heard in normal cases, sometimes in older people, but is registered more often in the low frequency channel. The sound is increased in cases of

augmented ventricular filling or reduced ventricular distensability. A pathological sound IV is found in mitral regurgitation, aortic stenosis, hypertensive cardiovascular disease, and ischemic heart disease.

Besides these four sounds, some pathological heart sounds may be present (Fig. 2). Among the systolic sounds there is the ejection sound and the nonejection systolic click. The ejection sound can be found in different pathological conditions such as congenital aortic or pulmonary valvular stenosis where opening of the cusps is restricted. A nonejection systolic click may be associated with a sudden mitral valve prolapse into the left atrium. An opening snap, a diastolic sound, may occur at the time of the opening of the mitral valve, for example, in cases with valve stenosis.

Heart murmurs are assumed to be caused by different mechanisms as compared to heart sounds. In fact, most murmurs result from turbulence in blood flow and occur as random signals. In normal blood vessels at normal velocity values blood flow is laminar, that is, in layers, and no turbulence is observed. In a normal resting human, there may be turbulent flow only in the vicinity of the aortic and pulmonary valves. As flow turbulence, a phenomenon that is generally irregular and random, is associated with pressure turbulence and, consequently, vessel wall vibration, acoustic phenomena may be observed. For flow in a smooth straight tube, the value of the Reynolds number, a dimensionless hydrodynamic parameter, determines the occurrence of turbulence. This number is proportional to the flow velocity and the tube diameter, and inversely proportional to the viscosity of the fluid. If this number exceeds a threshold value, laminar flow becomes turbulent. According to this theory, so-called innocent murmurs can be explained: They are produced if cardiac output is raised or when blood viscosity is lowered; they are generally early or midsystolic, have a short duration, and coincide with maximum ventricular outflow. Turbulence and thus intensity of the murmur increase with flow velocity. Pathological murmurs may be originated at normal flow rate through a restricted or irregular valve opening (e.g., in cases of valve stenosis) or by an abnormal flow direction caused by an insufficient (leaking) valve or a communication between the left and the right heart. As such systolic, diastolic, or even continuous murmurs may be observed. Systolic ejection murmurs occur in aortic and in pulmonary stenosis (valvular or non-valvular), diastolic filling murmurs in mitral and tricuspid stenosis. Aortic and pulmonary regurgitation cause diastolic murmurs; mitral and tricuspid regurgitation cause systolic murmurs. A systolic murmur and a diastolic murmur can be observed in ventricular septal defect. Continuous murmurs occur in patent ductus arteriosus (a connection between pulmonary artery and aorta). Musical murmurs occur as deterministic signals and are caused by harmonic vibration of structures (such as a valve leaflet, ruptured chordae tendinae, malfunctioning prosthetic valve) in the absence of flow turbulence; these are seldom observed.

The location of the chest wall where a specific sound or murmur is best observed (in comparison with the other phenomena) may help in discriminating the source of the sound or the murmur (1). These locations are dependent, not only on the distance to the source, but also on the vibration direction. Sounds or murmurs with an aortic valve origin are preferably investigated at the second intercostal space right of the sternum and those of pulmonary origin left of the sternum. The right ventricular area corresponds with the lower part of the sternum at the fourth intercostal space level, the left ventricular area between the sternum and the apex point of the heart (at the fifth intercostal space level). Furthermore, specific physiological maneuvers influencing cardiac hemodynamics may be used for obtaining better evaluation of heart sounds and murmurs.

In conclusion, the existence, timing, location at the chest wall, duration, relative intensity and intensity pattern, and frequency content of murmurs and/or pathological sound complexes form the basis of auscultatory, and/or phonocardiographic diagnosis of cardiac disease.

FUNDAMENTAL ASPECTS OF HEART VIBRATIONS

Mechanical heart action can be indicated by a set of time signals, which can be measured by invasive means: Most important variables are blood pressure in heart cavities and in blood vessels, myocardial and vessel wall tension, ventricular volume, blood flow velocity, heart wall deformation, and movement. At the chest surface only kinematic information is available: the movement of the chest surface as a result of mechanical heart action. As, in general, the movement of a material point can be indicated by a vector and as this vector appears to differ at various chest wall sites, one can state that mechanical information available at the chest wall is described by a spatiotemporal vector function. As far as the effect of the heart is concerned the movement is defined with respect to an equilibrium position; thus one can speak of a vibratory phenomenon. This movement of the chest surface, surrounded with air, gives rise to acoustic pressure in air; the latter is generally so weak that nothing can be perceived by hearing at a distance from the chest wall (except for artificial valve cases). Only if closed air cavities are used is a sound effect observable by the ear: The closed cavity (such as the stethoscope) prevents dispersion of acoustic energy and thus attenuation of acoustic pressure. It is out of the spatiotemporal kinematic vector function that phonocardiography takes a sample in order to evaluate cardiac activity.

As there is a vector function involved, three components should be taken into account. In practice, only the component perpendicular to the chest surface is measured and the two tangential components are disregarded. A kinematic function may be represented by different time representations, for example, displacement (m), velocity (m/s), acceleration (m/s^2), or even higher time derivatives (m/s^n, n representing the order of time derivative). Fundamentally, each representation contains identical information as they are all connected by a simple mathematical operation, that is, time derivation, but for visual inspection or time signal processing they reveal different vibratory patterns. Speaking in terms of the frequency domain, time derivation implies multiplication of the amplitude of an harmonic with its frequency: Time derivation is thus an operation of emphasizing higher frequencies in the signal with respect

to the lower ones. According to linear system theory, a similar effect is obtained with high pass filtering, the effect of filtering being described by the Nth time derivative of the signal in the attenuation band (i.e., for frequencies well below the cutoff frequency). The number N represents the order of the filter and determines the slope in the attenuation band of the amplitude characteristic ($N \times 20$ dB/decade). High pass filtering, order N, is theoretically identical with the corresponding low pass filtering of the Nth time derivative of the signal.

In biomedical signal processing, it is relatively uncommon to consider different time representations. From auscultation, we learned that in case of PCG it is rather beneficial. In a chest wall displacement curve, no sounds can be perceived, for example, at the site of the apex of the heart one can measure the apexcardiogram (Fig. 1), which is essentially a recording of the displacement of the skin surface and the ordinate axis could have a millimeter scale. Nevertheless, at the site of the apex sounds can be recorded if time derivation or high pass filtering is applied. In practice, transients corresponding to heart sounds become clearly visible in the acceleration recording. The ear cannot sense displacements such as the apexcardiogram; this can simply be explained if one regards the ear's sensitivity curve. In the range below < 1000 Hz, the sensitivity increases with increasing frequency, equivalent with the effect of high pass filtering.

The kinematic effect of heart action at the chest surface is not completely covered by phonocardiography. Historically, the frequency spectrum is divided into two parts: the low frequency part (up to 20 Hz) is handled under the title mechanocardiography and the second part beyond 20 Hz under PCG. The reason of separation lies in the nature of auscultation (frequencies beyond 20 Hz, according to hearing performance) and, additionally, palpation (frequencies < 20 Hz, according to tactile sensitivity). According to this, displacement recording belong to the domain of mechanocardiography, which studies arterial and venous pulse tracings, and the apexcardiogram. The carotid pulse (Fig. 1) is a recording of skin surface displacement at the site of the neck where the carotid artery pulsation is best palpated. The curve reflects local volume changes, and consequently pressure changes in the artery at the measurement site. It thus reflects changes in aortic pressure after a time delay determined by the propagation time of the blood pressure wave (10–50 ms). The beginning of the upstroke of the graph corresponds to the opening of the aortic valve and the dicrotic notch corresponds to its closing. As such, the left ventricular ejection period (LVEP) can be derived. The apexcardiogram (Fig. 1) is a recording of skin surface displacement of the chest wall at the site of the heart apex; in this case there is no propagation delay. The abrupt rise of the graph corresponds to the isovolumic contraction, the fall with the isovolumic relaxation. The minimum point occurs at the time of opening of the mitral valve. These displacement curves have been used for identifying heart sounds.

Chest wall kinematics are not exclusively caused by heart action. The most important movement in a resting human is originated by the respiration act. Two phenomena should be mentioned: a low frequency event corresponding with the breathing movement itself, and having its fundamental frequency at the breathing rhythm ($\sim$0.2–0.4 Hz), and a high frequency phenomenon corresponding to breathing noises, "lung sounds", due to the turbulent air stream in airways and lungs. The latter may cause great disturbances in high frequency heart sound recording. To these effects one can add the result of stomach and intestine motility and, moreover, environment noise picked up by the chest wall. From the standpoint of PCG, these effects are merely disturbing and thus to be minimized.

In PCG, one discriminates between heart sounds and murmurs (based on auscultation). It has already been indicated that the source types are different as well as the acoustic impressions they provoke. From the standpoint of signal analysis heart sounds correspond better with transients originated by a sudden impact, whereas murmurs, except for the musical types, have a random character. In fact, if one considers a set of subsequent heart cycles, one may find that heart sounds are more coherent compared to murmurs. For example, averaging of sounds of subsequent heart cycles with respect to an appropriate time reference gives a meaningful result, but the same fails for murmurs as a consequence of their random character.

Conventional heart sound recording is executed at the chest wall. Some exceptions are worth mentioning. Intracardiac phono signals are obtained from cardiac blood pressure curves during catheterization. If a catheter-tip pressure transducer with sufficient bandwidth is used, intracardiac sound recordings can be obtained by submitting the pressure signal to high pass filtering or time derivation. It is also possible to get closer to the heart in a noninvasive way by measuring pressure or kinematics in the esophagus. In this way, the posterior part of the heart, lying close to the esophagus, is better investigated.

RECORDING OF HEART SOUNDS

In auscultation, the physician uses a stethoscope as a more practical alternative for putting the ear in close contact to the chest wall. Recording of heart sounds is a problem of vibration measurement (13), more specifically on soft tissue. It implies the need of a sensor, appropriate amplification and filtering, storage and visualization on paper (14,15), or by using information technology. The useful bandwidth is $\sim$20–1000 Hz. The sensor needs to be a vibration transducer (vibration pickup), in this case also called a heart sound microphone; an alternative is a stethoscope provided with a microphone: the electronic stethoscope. Except for the sensor, virtual instrumentation technology can be used in the measuring chain: This implies a PC with a data acquisition card and signal processing software (such as Labview).

The Transducer

In a normal situation the chest wall vibrates only surrounded with air, which exerts an extremely low loading effect. Consequently, force at the surface may be neglected and only the kinematic variable is important. This ceases to be true when a transducer is connected to the chest wall,

exerting a significant loading effect. This loading influence is described by the mechanical impedance of the pickup (force/velocity) or by the dynamic mass (force/acceleration). The loading effect is also dependent on the chest wall tissue properties; chest wall mechanical impedance is determined by tissue elasticity, damping, and mass. In general, heart sound microphones provoke a large and difficult to quantify loading effect on the chest wall, so that in no case the standard unloaded vibration is recorded. The same is not true in electrocardiography, for example, where the apparatus is designed with a sufficiently high input impedance to record the unaffected electrical potential at the electrode site. The latter is hard to achieve in phonocardiography (16).

Heart sound transducers can be divided into two types: the absolute pickup and the relative pickup. The absolute pickup measures the vibration at the site of application, averaged over the area of application. In general, these are contact pickups that are rigidly connected to the chest wall; the measuring area is thus identical to the contact area. These types are similar to the ones used for industrial measurements on mechanical structures or in seismography. The relative pickup measures the vibration averaged over a certain area with respect to a reference area; it is thus a kind of differential pickup. Air-coupled pickups are differential pickups. Essentially they consist of an air-filled cavity, generally with a circular shape, the edge of the cavity rigidly and air-tight-connected to the chest wall. It is thus the difference of the displacement under the cavity (the measuring area) and the displacement under the edge of the cavity (the reference area) giving rise to an acoustic air pressure within the cavity that is measured. The electronic stethoscope can be considered as a relative pickup. Figure 3 shows the principles.

With the contact vibration pickup the average kinematics of the measuring area in the loaded situation is recorded. The most generally applied transducing principle is the seismic type: A seismic mass is connected via a

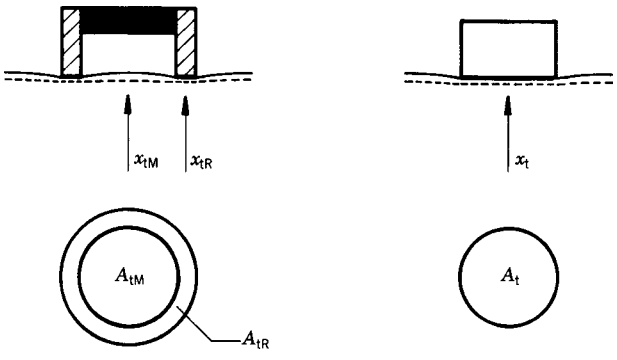


Figure 3. A schematical representation of heart vibration pickups: left, the relative type; right, the absolute type; above, the pickup positioned at the vibrating chest wall; below, the corresponding areas relating the pickup with the chest wall. The absolute pickup measures the kinematics x_t at its contact area A_t. The relative pickup, presented as an air cavity with a pressure-sensing device at the top of the cavity (black part), measures the difference between the kinematics under the cavity x_{tM}, that is, of the measuring area A_{tM}, and the one under the edge of the cavity x_{tR}, that is, the reference area A_{tR}.

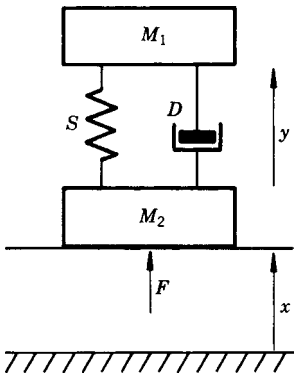


Figure 4. The seismic system, the mechanical model for a contact pickup (i.e., an accelerometer). The system is attached to the vibrating surface (F, x) with its contact mass M_2; The seismic mass M_1 is coupled to the contact mass via a stiffness S and a damping D. The displacement y between the contact and the seismic mass represents the measuring value.

spring-damping system to the contact mass coupled with the vibrating surface (Fig. 4). The relative displacement between the seismic mass and the contact mass is measured with the aid of a mechanoelectric transducing device. The latter can be a piezoelectric crystal that generates an electrical charge proportional to its deformation (Fig. 5); The complete device behaves as an accelerometer for frequencies (f) below its acceleration resonance frequency (f_1); it measures displacement above f_1. Measuring acceleration is generally the normal function. The acceleration charge sensitivity s_Q (charge per acceleration unit, pC/m·s^{-2}) is thus

$$s_Q = B/(2\pi f_1)^2 \tag{1}$$

for $f < f_1$. The parameter B stands for a mechanoelectric transducing efficiency and depends on the crystal type and

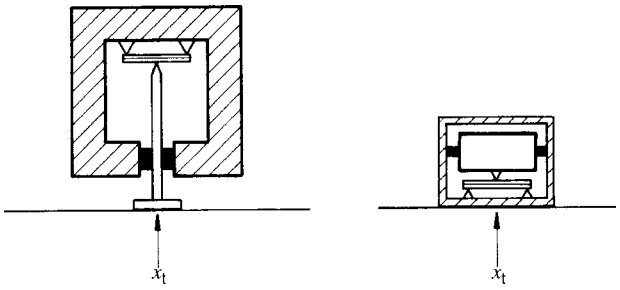


Figure 5. A schematic representation of two types of phonocardiographic contact pickups: at left, the heavy, and at right, the low weight type. The mechanoelectric transducing device (here presented as a bilaminar bending crystal) is built in to measure the displacement between the seismic mass and the material in contact with the vibrating surface. The black parts indicate elastic material. The heavy type (possibly held by hand at the massive case, representing the seismic mass) makes contact with the chest wall via a coin-shaped disk connected to the crystal by an extension rod. For the low weight type, the case makes contact with the vibrating surface and all remaining parts (including the seismic mass attached to the crystal) are at the inside. *Note*: The heavy type is mentioned because of its historical value.

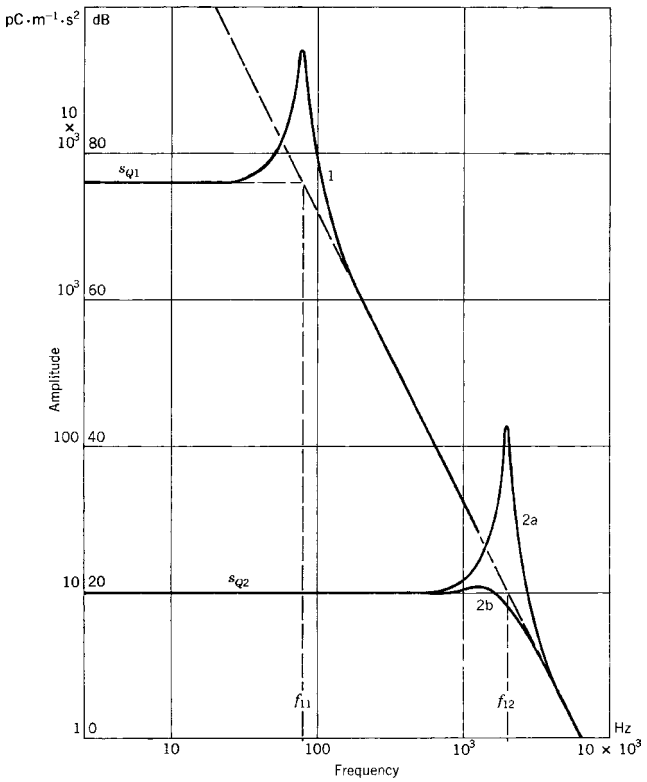


Figure 6. The acceleration amplitude frequency characteristic of a seismic pickup (double logarithmic scale). Curves presented can be obtained using the same mechanoelectric transducing device, but with a different mechanical layout of the seismic system. The oblique broken line represents a sensitivity limit for the specific mechanoelectric transducing device. The characteristic of a seismic system corresponds with a second-order low pass filter. The resonant frequency f_1 determines the bandwidth (flat part), but also the sensitivity (Eq. 1). Two different types are shown; the low resonant frequency f_{11} can be found in heavy phonocardiographic pickups; the high one, f_{12}, in low weight types. For the latter, the effect of damping is shown: Curve 2b corresponds to a higher damping in the system, as can be seen by the decreased height of the resonance peak.

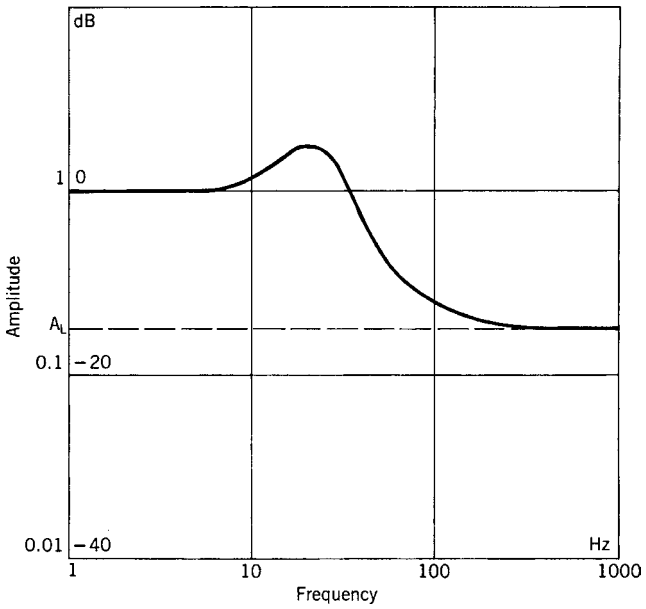


Figure 7. Amplitude frequency characteristic of the loading effect caused by the coupling of a rigid element to the soft tissue of the chest wall (as e.g., a contact pickup) (drawing for a 10 g, 20 mm diameter element). Attenuation is about constant beyond 100 Hz (Eq. 2).

on its mounting; f_1 is a construction parameter determined by the seismic mass and the stiffness incorporated between the seismic and the contact mass. The complete amplitude frequency characteristic corresponds with a low pass second-order filter, with f_1 also representing the cutoff frequency (Fig. 6), thus determining the measuring bandwidth and the sensitivity. Evidently other then piezoelectric elements can be used for measuring displacement: piezoresistance, variable capacitance (both of them needing a polarization voltage), electret elements, and so on.

The loading influence of the pickup can be presented by its dynamic mass. In comparing contact pickups with different masses with an ultralow weight type (Fig. 7) it was found that distortion and attenuation is caused by the mass (16): Beyond 100 Hz, the amplitude ratio (loaded compared to unloaded) approximates a value (A_L):

$$A_L = M_t/(M_t + M_1) \qquad (2)$$

with $M_t = a\,d_t^3$, $a = 280$ kg/m^3 (typical); M_1 is the loading mass (of the pickup), M_t is the thorax wall output mass, and d_t is the contact diameter. According to this formula, a 10-g pickup with a 20 mm diameter would result in an attenuation to 18% of the original unloaded amplitude. For quantitative purposes, ultralow weight pickups can thus be recommended. It must also be emphasized that not the weight per se, but the weight divided by the third power of the contact diameter is the parameter to be minimized (according to Eq. 2).

In the case of the air-coupled vibration pickup (17), the average kinematics of the measuring area (under the cavity) with respect to the reference area (under the edge of the cavity) in the loaded situation is recorded (Fig. 3). Air pressure in the cavity as a result of the relative displacement of the chest wall is registered with a built-in sensor measuring acoustic pressure (a microphone). The movement of the membrane of this microphone is transformed into an electrical signal, for example, by the moving coil principle (dynamic type) and variable capacitance (condenser type with a polarizing voltage, electret type). As such, the measuring characteristics of the air-coupled pickup are determined by the dimensions of the air cavity and by the features of the included microphone. If the microphone membrane is rather stiff as compared with air and the height (l) of the cavity small as compared with the wavelength of heart vibrations in air, the pressure (p) generated at the site of the membrane is simply proportional to the relative displacement of the chest wall ($x_{tMR} = x_{tM} - x_{tR}$):

$$p = (c^2/l)\,x_{tMR} \qquad (3)$$

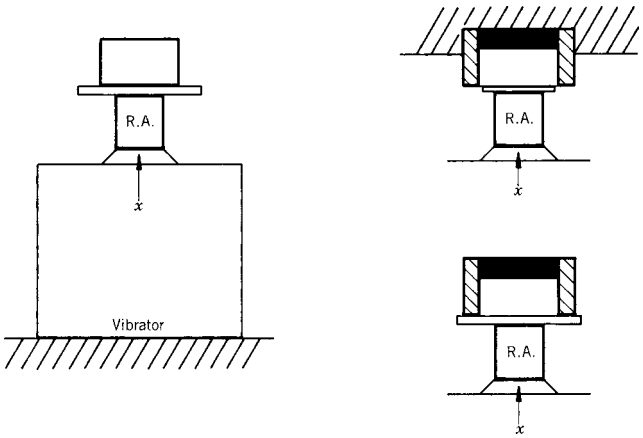


Figure 8. Calibration of heart vibration pickups. The pickup to be tested is compared with a reference accelerometer (RA): left, the contact pickup; right, the relative pickup. The contact pickup is rigidly connected with the reference accelerometer, mounted on a vibrator (left). For the relative pickup two measurements can be performed (right, vibrator not displayed). In a first test, the differential characteristic is determined (right, above): With the housing fixed, a movement is generated at the orifice of the cavity, which is closed airtight with (e.g., an elastic membrane). In a second test the common mode sensitivity is determined, the pickup fixed at a stiff plate. An ideal relative pickup features zero common mode sensitivity.

where ρ is the air density (kg/m^3) and c is the wave propagation velocity in air (m/s). The loading effect under the cavity is given by a stiffness term (S_{tM}, N/m) and

$$S_{tM} = d_{tM}^2 \rho c^2/(4l) \quad (4)$$

where d_{tM} is the diameter of the measuring area. At the reference area, the pickup exerts a mass loading; the combined loading effect is hard to describe.

For calibration of a phonocardiographic pickup one can use a vibrator, a reference accelerometer having an ideal frequency response, and two amplifiers (Fig. 8). The contact pickup to be tested should be rigidly attached to the reference pickup. In case of an air-coupled pickup, one has to investigate the differential and the common mode characteristics, the first one with the housing fixed, a displacement being generated at the input of the cavity (e.g., airtight sealed with an elastic membrane), and the second one with the complete microphone fixed at the rigid vibration table (the relative displacement between inner and edge of the cavity being zero). For an ideal relative pickup, the common mode term should be zero; In practice, it is to be small with respect to the differential sensitivity. Mechanical impedance of a pickup can similarly be measured by applying a force transducer combined with the reference accelerometer (an impedance head). It should be emphasized that in this way only the properties of the pickup can be obtained, but no information is acquired on the ultimate distortion effect due to the loading of the chest wall, as this is also dependent on the tissue parameters. This effect can only be measured *in situ*, that is, on the patient themself; the procedure is rather complicated.

In conclusion, accurate quantitative recording of the unloaded chest wall vibration is extremely difficult. Industrial accelerometers (equivalent to the described contact types) show the same disadvantages with respect to loading on the soft tissue of the chest wall. Moreover, as a rule they are less sensitive and have an unnecessary broad bandwidth. Some noncontact methods have been reported, but only for infraphonocardiographic frequencies.

Preamplification and Signal Preprocessing

Without going into details regarding electronics, we must emphasize the importance of the preamplifier, as this part of the recording unit, in combination with the pickup, determines the electronic noise level of the apparatus. It is adequate to resume possible disturbing influences at this point. First, there are the physiological vibrations generated by organs other than the heart. Respiration sounds may be the most inconvenient disturbances, especially with patients suffering from lung disease. Recording in an expired resting state can thus be advised. Furthermore there are environmental vibrations; Air-transmitted noises, especially the higher frequencies at which cardiac vibration intensity is weak, can be very inconvenient. Therefore, the air-coupled pickup should be attached in an airtight manner to the chest wall. This does not solve the problem completely, as environmental noises seem to be picked up by the chest wall and, in this way, transmitted to the pickup (air-coupled as well as contact type). Theoretically, these disturbing vibrations can be minimized, but for a given apparatus electronic noise cannot be affected and thus sets a limit to noise diminishing and determines a threshold under which no heart sound or murmur can be recognized.

Besides the noise level, the frequency characteristic of the preamplifier connected to the pickup should be regarded. For example, in the case of a piezoelectric pickup, its electrical output property is capacitive and one must bear in mind that the combination of this capacitance with the input resistance of a voltage preamplifier gives rise to a first-order high pass effect. To avoid this problem, charge or current amplifiers can be used. Whereas the charge amplifier measures acceleration, the current amplifier measures its time derivative.

Whether or not signals are digitized after preamplification, a high pass filtering (or band-pass filtering) process is necessary (18,19). High pass filtering and appropriate amplification (analogue or digital) of the filter channels compensate for the fact that visual inspection of a single recorded phonocardiogram (even when optimally chosen) does not reveal the same amount of information as gained from the acoustical impression during auscultation. Furthermore, the amplification has to compensate for the decreasing amplitude of heart vibrations at increasing frequencies. Conventionally, a set of about four high pass filters are used, each characterized by a gradually increasing cutoff frequency and/or increasing slope in the attenuation band. For example, Maass Weber high pass filters are used with a common cutoff frequency of 1 kHz and slopes of 20, 30, 40, and 60 dB/octave (EEC, see the section Hardware and Software). Generally, filter sets have been

determined qualitatively by applying a range of subsequent filters and recording a set of normal subjects and patients with different heart diseases: A filter providing information also perceivable in another one was eliminated from the set. Furthermore, filtering must permit discrimination between phenomena having a different physiological or pathological origin. Clear splitting between slightly asynchronous phenomena (or minimal overlapping) is thus desired for vibrations having similar frequency content. Now discrimination between phenomena having a different frequency content in adjacent filter channels is expected. The chosen set is not uniquely optimal: It depends on the preceding elements in the measuring chain (the vibration pickup, including its loading effect, and the preamplifier). As such, a filter set chosen for a contact pickup is not evidently optimal for an air-coupled type.

Different transducers, the mostly unknown distortion effect due to loading and different filter sets, might seem remarkable to physicists and engineers from the viewpoint of measuring quality. Nevertheless, for (semiquantitative) phonocardiography the use of filtering with adaptable amplification compensates in some degree for microphone, loading, and preamplification characteristics. For example, attenuation due to loading in a specific frequency band may be partly compensated by increased amplification of the corresponding channel.

Storage and Visualization

In older apparatus intended for recording of ECG, PCG, and pulses, a paper recorder (strip chart recorder) was included. The first was an analog type [as the galvanometric pen writer, having a limited bandwidth ($\sim$100 Hz), but equipped with special techniques for recording high frequency sounds], and later it was a digital type as the thermal array recorder. The latter, also available as a general purpose paper recorder, functions completely digitally: It sets a dot at the point corresponding to the instantaneous value of the signal to be recorded: No moving mechanical parts are present except for the paper drive. The recording technique is characterized by a sampling and a writing frequency. The latter may be lower than the first: During the writing interval all points between the maximum and the minimum value of the signal samples are dotted. As such, the recording is a subsequence of vertical lines: For visual inspection of the overall vibration pattern no information is lost.

Furthermore, data can be handled by common information technology: a (portable) personal computer with appropriate data-acquisition possibilities, virtual instrument software for signal conditioning and processing, visualization, archiving, and hard copy generation.

PROCESSING OF HEART SOUNDS AND PHYSICAL MODELING

Physical modeling aims at the localization of a specific sound source in the heart and, by analyzing the externally recorded vibration signals, at the quantification of the constitutive properties of the cardiac structures involved (e.g., stiffness of a valve leaflet, myocardial contractility) and of the driving forces, which set these structures into vibration. The physical situation is extremely complicated. The vibration source is situated within the cardiac structures (having viscoelastic properties) containing and driving blood (a viscous fluid). The transmission medium, the tissues between the heart and the chest wall, is viscoelastic and inhomogeneous.

Transmission in such a viscoelastic medium implies compression and shear waves, which both contribute to the vibrations at the chest wall (20). It is not simply a problem of acoustic pressure as in a perfect fluid. Distortion due to transmission seems obvious. In order to study transmission and to relate chest wall vibrations to properties of cardiac structures and hemodynamic variables, advanced signal processing techniques are used. A broad review is given by Durand et al. (21).

As the chest wall vibratory phenomenon is represented by a spatiotemporal kinematic function, it can principally be approached in two ways: by sampling in time, as a set of images of chest wall movement, or by sampling in space by a set of time signals obtained with multisite recording. Multisite heart sound recording implies a large set of pickups (preferably light weight, thus inducing minimal loading). In this way, spatial distribution of vibration waveforms on the chest wall can be derived. Based on the results of such a method a physical model for heart sound genesis has been presented that can analytically be solved in a viscoelastic medium: a sphere vibrating along the axis of the valve orifice (20). This mechanical dipole model agrees to the idea of sound generation as a resonant-like vibration of the closed elastic valve leaflets and the surrounding blood mass. With this model a typical inversion of vibration waveforms on the chest wall could be explained: The phase reversal is expressed most for the second sound, according to the anatomical position and direction of the aortic orifice. The model has been used to calculate source functions (the inverse problem). Spatial parameters on vibration waveforms have been formulated (22–25).

Physical modeling aims at the quantification of the constitutive properties of cardiac structures (e.g., of the valve leaflets) and the driving forces (e.g., blood pressure). For example, with respect to the second sound, the aortic valve was modeled as a circular elastic membrane, it was allowed to vibrate in interaction with the surrounding blood mass, with as a driving force the slope of the development of the transvalvular pressure difference during isovolumic relaxation (11,26). Typical characteristics of IIA and IIP could thus be explained. For example, the reduction of amplitude and frequency shift (toward higher frequencies) as a consequence of valve stiffening, the diminishing of amplitude in patients with poor ventricular performance (characterized by a slow pressure drop in the ventricle during the isovolumic relaxation), and the augmentation of amplitude in cases of anemia (implying reduced blood viscosity and thus reduced damping in the resonant system). In another model, the ventricle is modeled as a finite thick-walled cylinder and the amplitude spectra of computed vibration waveforms contain information concerning the active elastic state of muscular fibers that is dependent on cardiac contractility (27).

Transmission of vibrations by comparing vibrations at the epicardial surface and at the chest wall has been studied (21). Esophageal PCG proved to be beneficial for recording vibrations originated at the mitral valve (28). The disappearance of the third sound with aging was explained with the ventricle modeled as a viscoelastic oscillating system with increasing mass during growth (29). Spectral analysis of the pulmonary component of the second sound reveals information on the pressure in the pulmonary artery (30).

Frequency content and timing of heart vibrations is of major importance; Time–frequency analysis of signals is thus performed. Classical Fourier analysis uses harmonic signals (sine and cosine waves) as basic signals. The frequencies of the harmonics are multiples of the fundamental frequency and the signal can be composed by summing the sine and cosine waves multiplied with the Fourier coefficients. Sine waves have an infinite duration and the method is thus beneficial for periodic functions. A phonocardiogram can be considered as a periodic function, but it is composed of a number of phenomena shifted in time with specific frequency content (heart sound components and murmurs). When applying classical Fourier analysis, information on timing is lost. Thus Fourier analysis has to be performed on shorter time intervals (by dividing the heart cycle into subsequent small intervals) resulting in time and frequency information. To minimize errors resulting from calculating in these small intervals, mathematical techniques have to be applied. Wavelet analysis calculates wavelet coefficients based on transient-like irregular signals with limited duration, called wavelets. Wavelets are derived from a mother wavelet and obtained by scaling in time (subsequently with a factor 2) and by shifting in time. As in Fourier analysis, the signal can be composed by summing shifted and scaled wavelets multiplied with their wavelet coefficients. The waveform of the mother wavelet can be chosen. As scaling in time corresponds to frequency, this method also gives time and frequency information, but it performs better for analyzing signals of a nonstationary nature, such as heart sounds and murmurs. Sudden changes or discontinuities in the signal can better be identified and located in time. A large number of studies has been executed with respect to time–frequency analysis of heart sounds and murmurs and different calculation methods have been compared (21). Spectral analysis of heart murmurs appeared to be useful to estimate transvalvular pressure difference in patients with aortic valve stenosis (31,32). Spectral analysis was used to monitor the condition of bioprosthetic valves and mechanical valve prostheses (33,34). Wavelet transform (35) and a nonlinear transient chirp signal modeling approach (36) were used to detect the aortic and the pulmonary component of the second sound. The matching pursuit method was used to identify the mitral and the tricuspid component in the first sound (37). A tailored wavelet analysis has been used to automatically detect the third heart sound (38). Time–frequency analysis was applied for classification of heart murmurs produced by bioprosthetic valves (39), for studying the first heart sound (40), and for automated detection of heart sounds (41).

THE ELECTRONIC STETHOSCOPE

Clinical interest in PCG in its classical form has been decreasing during the last decade, but there seems to be an increasing interest in heart sound recording with the aid of electronic stethoscopes (1), combined with information technology allowing easy data-acquisition, visualization, data handling and parameter extraction, playback, telemedicine applications, waveform recognition, and diagnosis with the aid of databanks. Also, virtual instrumentation technology (such as Labview) is used for heart sound acquisition and processing.

The modern acoustic stethoscope comprises a binaural headset and a chest piece connected by elastic tubing (1). The headset is composed of ear tubes and ear tips; the chest piece can consist of a bell and a diaphragm part. The ear tips should fit in the ear orifice, preventing air leakage, with the tube properly aligned to the hearing canal, that is slightly directed forward. The tube connecting the headset and the chest piece should not be to long to restrict attenuation of acoustic pressure, especially of higher frequencies, generated at the chest piece. With the diaphragm part provided with a stiff membrane (diameter $\sim$4 cm), applied firmly to the skin, the high frequency sounds are better observed. With the bell part of the chest piece, applied with low pressure to the skin (enough to prevent air leaks between skin and bell edge) low frequency vibrations are best picked up. The bell diameter should be large enough to span an intercostal space ($\sim$2.5 cm for adults). Firm application of the bell makes the skin act as a membrane thus diminishing its low frequency performance. Some stethoscopes have only one part with a specially mounted membrane, which can function in the "bell mode" or in the "membrane mode" by altering applied pressure for the purposes cited above. As such, in the application of the stethoscope, frequency filtering (as in phonocardiography) is performed by using a specific shape and mechanical coupling to the chest wall.

The electronic stethoscope (e-stethoscope) combines the simplicity of the acoustic stethoscope with the benefits of electronics and information technology. Essentially, the e-stethoscope is an acoustical type provided with a built-in microphone; as such, it can be indicated as an air coupled vibration pick-up. In its simplest form, sounds are transmitted to the ears by tubing as in the acoustical one. The more advanced type has the microphone built within the chest piece, with adjustable amplification and filtering, mode control with easy switching between bell and diaphragm modes, generation of processed sound by miniature speakers to the air tips, cable or wireless connection to a personal computer for further processing. Adjustment of stethoscope performance can be executed during auscultation. Most stethoscopes are intended for observation (and recording) of heart sounds and murmurs, and for lung and airway sounds as well. A special type, the esophageal stethoscope, can be used for monitoring heart and lung sounds during anesthesia (42).

User-friendly software is available for diagnostic and for training purposes. Recorded signals can be printed, visualized, adapted by digital filtering and scaling, improved by elimination of artifacts and disturbances, and combined

with synchronously recorded ECG. Processed sounds can be reproduced and played back with speakers with a sufficient bandwidth (in the low frequency range down to 20 Hz). Spectral analysis is also possible: Frequency content as a function of time can be displayed. Automated cardiac auscultation and interpretation can be useful in supporting diagnosis (43–45). Sounds recorded by a local general physician can be sent via internet to the cardiologist for accurate diagnosis (46).

Educational benefits are obvious. Heart sounds recorded with the e-stethoscope or obtained from a databank can be visually observed and listened to. CD–ROMs with a collection of typical heart sounds and murmurs are available for training purposes. Multimedia tools were found to contribute to the improvement of quality of physical examination skills (47,48).

HARDWARE AND SOFTWARE

In this paragraph, some practical details are given with respect to available hard and software. A conventional form of a phonocardiograph (heart sound transducer, amplifier and filters, audio, output connectable to recorder, also fit for lung sound recording) can be obtained from EEC (http://www.eeconnet.com). ADInstruments provides a heart sound pickup (http://www.adinstruments.com). Colin (http://www.colin-mt.jp/eng) provides a phonocardiograph together with ECG and noninvasive blood pressure measurement for noninvasive assessment of arteriosclerosis. Electronic stethoscopes can be purchased at Cardionics (http://www.cardionics.com), 3M (Litttmann) (http://www.3M.com/product/index.jhtml), Meditron (http://www.meditron.no/products/stethoscope), Philips (http://www.medical.philips.com/main/products/), EEC (http://www.eeconnet.com). Software supporting the physician in the evaluation of heart sounds recorded with an electronic stethoscope is provided by Zargis (http://www.zargis.com), Stethographics (http://www.stethographics.com/index. html). Software intended for training in heart sound auscultation (heart sounds recorded with an electronic stethoscope or from data banks) can be obtained from Biosignetics (http://www.bsignetics.com), Zargis (http://www.zargis.com), Cardionics (http://www.cardionics.com).

EVALUATION

Evaluation of heart sounds and murmurs remains an important method in the diagnosis of abnormalities of cardiac structures. Conventional PCG, however, essentially the graphic recording of sounds for visual inspection, has lost interest as a result of a number of reasons. First, the vibration signals are complex and thus difficult to interpret; they are characterized by a broad frequency range and, as such, different time representations present specific information (low and high frequencies). Obtaining high quality recordings having a high signal-to-noise ratio is difficult. Genesis and transmission of vibrations is difficult to describe and insufficiently known. A variety of waveforms are observable at the chest surface; Multisite recording and mapping are useful with respect to the solving of the genesis and transmission problem but are difficult to execute and result in a large amount of data to be analyzed. The recording technique is not standardized; The ordinate axis of a phonocardiographic waveform does not have a physical unit as, for example, the millivolt in electrocardiography. The latter is due to the different transducer types, unquantified loading effect of the transducer on the chest wall, different frequency filter concepts. Thus, the method remains bound to a specific recording method and is semiquantitative. No guidelines for universal use have been developed and proposed to the clinical users. The most important reason evidently is found in the availability of technologies like echocardiography, Doppler, and cardiac imaging techniques, which provide more direct and accurate information concerning heart functioning. The latter, however, have the disadvantages of being costly and restricted to hospitals. Nevertheless, knowledge of heart sounds and murmurs has been greatly increased with the PCG technique and research is still going on. Signal analysis, more specifically time–frequency analysis, has proven to be very useful in the identification and classification of heart sound components and murmurs and their relation to cardiac structures and hemodynamic variables.

Conventional PCG has lost interest. Nevertheless, the historical value of the method has to be stressed. Auscultation, being simple, cheap, and not restricted to the hospital environment, held its position as a diagnostic tool for the general physician and for the cardiologist as well. However, this technique requires adequate training. Recording and processing of heart sounds remain beneficial for training and for supporting diagnosis. Electronic stethoscopes coupled to a laptop with suitable software and connected to the internet for automated or remote diagnosis by a specialist may grow in importance in the coming years.

BIBLIOGRAPHY

1. Tilkian AG. Understanding heart sounds and murmurs with an introduction to lung sounds. Philadelphia: W.B. Saunders; 2001.
2. Salmon AP. Heart sounds made easy. London: Churchill Livingstone; 2002.
3. Wartak J. Phonocardiology: Integrated Study of Heart Sounds and Murmurs. New York: Harper & Row; 1972.
4. Luisada AA. The Sounds of the Normal Heart. St. Louis, MO: Warren H. Green; 1972.
5. Delman AJ, Stein E. Dynamic Cardiac Auscultation and Phonocardiography. A Graphic Guide. Philadelphia, PA: W.B. Saunders; 1979.
6. Einthoven W. Die Registrierung der Mensclichen Hertztone mittels des Saitengalvanometers. Arch Gesamte Physiol Menschen Tiere 1907;117:461.
7. Mannheimer E. Standardization of phonocardiography. Am Heart J 1957;54:314–315.
8. Holldack K, Luisada AA, Ueda H. Standardization of phonocardiography. Am J Cardiol 1965;15:419–421.
9. Groom D. Standardization of microphones for phonocardiography. Biomed Eng 1970;5:396–398.
10. Rushmer RF. Cardiovascular Dynamics. Philadelphia, PA: Saunders; 1976.
11. Stein PD. Physical and Physiological Basis for the Interpretation of Cardiac Auscultation. Evaluations Based Primarily on the Second Sound and Ejection Murmurs. New York: Futura Publishing Co.; 1981.

12. Luisada AA, Portaluppi F. The Heart Sounds. New Facts and Their Clinical Implications. New York: Praeger; 1982.
13. Harris CM. Shock and Vibration Handbook. 5th ed. New York: McGraw-Hill; 2001.
14. van Vollenhoven E, Suzumura N, Ghista DN, Mazumdar J, Hearn T. Phonocardiography: Analyses of instrumentation and vibration of heart structures to determine their constitutive properties. In: Ghista DN, editor. Advances in Cardiovascular Physics. Vol. 2, Basel: Karger; 1979. pp. 68–118.
15. Verburg J, van Vollenhoven E. Phonocardiography: Physical and technical aspects and clinical uses. In: Rolfe P, editor. Non Invasive Physiological Measurements. London: Academic Press; 1979. pp. 213–259.
16. Vermariën H, van Vollenhoven E. The recording of heart vibrations: A problem of vibration measurement on soft tissue. Med Biol Eng Comput 1984;22:168–178.
17. Suzumura N, Ikegaya K. Characteristics of air cavities of phonocardiographic microphones and the effects of vibration and room noise. Med Biol Eng Comput 1977;15:240–247.
18. Maass H, Weber A. Herzschallregistrierung mittels differenzierende filter. Eine Studie zur Herzschallnormung. Cardiologia 1952;21:773–794.
19. van Vollenhoven E, Beneken JEW, Reuver H, Dorenbos T. Filters for phonocardiography. Med Biol Eng 1967;5:127–138.
20. Verburg J. Transmission of vibrations of the heart to the chest wall. In: Ghista DN, editor. Advances in Cardiovascular Physics. Volume 5, Part III, Basel: Karger; 1983. pp. 84–103.
21. Durand LG, Pibarot P. Digital signal processing of the phonocardiogram: review of the most recent advancements. Crit Rev Biomed Eng 1995: 23(3–4):163–219.
22. Vermariën H. Mapping and vector analysis of heart vibration data obtained by multisite phonocardiography. In: Ghista DN, editor. Advances in Cardiovascular Physics. Volume 6, Basel: Karger; 1989. pp. 133–185.
23. Wood JC, Barry DT. Quantification of first heart sound frequency dynamics across the human chest wall. Med Biol Eng Comput 1994;32(4 Suppl):S71–78.
24. Baykal A, Ider YZ, Koymen H. Distribution of aortic mechanical prosthetic valve closure sound model parameters on the surface of the chest. IEEE Trans Biomed Eng 1995;42(4): 358–370.
25. Cozic M, Durand LG, Guardo R. Development of a cardiac acoustic mapping system. Med Biol Eng Comput 1998;36(4): 431–437.
26. Blick EF, Sabbah HN, Stein PD. One-dimensional model of diastolic semilunar valve vibrations productive of heart sounds. J Biomech 1979;12:223–227.
27. Lewkowicz M, Chadwick RS. Contraction and relaxation-induced oscillations of the left ventricle of the heart during the isovolumic phases. J Acoust Soc Am 1990;87(3):1318–1326.
28. Chin JGJ, van Herpen G, Vermariën H, Wang J, Koops J, Scheerlinck R, van Vollenhoven E. Mitral valve prolapse: a comparative study with two-dimensional and Doppler echocardiography, auscultation, conventional and esophageal phonocardiography. Am J Noninvas Cardiol 1992;6:147–153.
29. Longhini C, Scorzoni D, Baracca E, Brunazzi MC, Chirillo F, Fratti D, Musacci GF. The mechanism of the physiologic disappearance of the third heart sound with aging. Jpn Heart J 1996;37(2):215–226.
30. Chen D, Pibarot P, Honos G, Durand LG. Estimation of pulmonary artery pressure by spectral analysis of the second heart sound. Am J Cardiol 1996;78(7):785–789.
31. Nygaard H, Thuesen L, Hasenkam JM, Pedersen EM, Paulsen PK. Assessing the severity of aortic valve stenosis by spectral analysis of cardiac murmurs (spectral vibrocardiography). Part I: Technical aspects. J Heart Valve Dis 1993; 2(4):454–467.
32. Nygaard H, Thuesen L, Terp K, Hasenkam JM, Paulsen PK. Assessing the severity of aortic valve stenosis by spectral analysis of cardiac murmurs (spectral vibrocardiography). Part II: Clinical aspects. J Heart Valve Dis 1993;2(4): 468–475.
33. Sava HP, Grant PM, Mc Donnell JT. Spectral characterization and classification of Carpentier-Edwards heart valves implanted in the aortic position. IEEE Trans Biomed Eng 1996;43(10):1046–1048.
34. Sava HP, Mc Donnell JT. Spectral composition of heart sounds before and after mechanical heart valve implantation using a modified forward-backward Prony's method. IEEE Trans Biomed Eng 1996;43(7):734–742.
35. Obaidat MS. Phonocardiogram signal analysis: techniques and performance comparison. J Med Eng Technol 1993; 17(6):221–227.
36. Xu J, Durand LG, Pibarot P. Extraction of the aortic and pulmonary components of the second heart sound using a nonlinear transient chirp signal model. IEEE Trans Biomed Eng 2001;48(3):277–283.
37. Wang W, Guo Z, Yang J, Zhang Y, Durand LG, Loew M. Analysis of the first heart sound using the matching pursuit method. Med Biol Eng Comput 2001;39(6):644–648.
38. Hult P, Fjallbrant T, Wranne B, Ask P. Detection of the third heart sound using a tailored wavelet approach. Med Biol Comput 2004;42(2):253–258.
39. Debiais F, Durand LG, Guo Z, Guardo R. Time-frequency analysis of heart murmurs, Part II: Optimisation of time-frequency representations and performance evaluation. Med Biol Eng Comput 1997;35(5):480–485.
40. Chen D, Durand LG, Lee HC, Wieting DW. Time-frequency analysis of the first heart sound. Part 3: Application to dogs with varying cardiac contractility and to patients with mitral mechanical prosthetic heart valves. Med Biol Eng Comput 1997;35(5):455–461.
41. Sava H, Pibarot P, Durand LG. Application of the matching pursuit method for structural decomposition and averaging of phonocardiographic signals. Med Biol Eng Comput 1998;36(3):302–308.
42. Manecke GR, Jr., Poppers PJ. Esophageal stethoscope placement depth: its effect on heart and lung sound monitoring during general anesthesia. Anesth Analg 1998;86(6):1276–1279.
43. Thompson WR, Hayek CS, Tuchinda C, Telford JK, Lombardo JS. Automated cardiac auscultation for detection of pathologic heart murmurs. Pediatr Cardiol 2001;22(5): 373–379.
44. Hayek CS, Thompson WR, Tuchinda C, Wojcik RA, Lombardo JS. Wavelet processing of systolic murmurs to assist with clinical diagnosis of the heart disease. Biomed Instrum Technol 2003;37(4):263–270.
45. Pavlopoulos SA, Stasis AC, Loukis EN. A decision tree—based method for the differential diagnosis of aortic stenosis from mitral regurgitation using heart sounds. Biomed Eng Online 2004;3(1):21.
46. Guo Z, Moulder C, Zou Y, Loew M, Durand LG. A virtual instrument for acquisition and analysis of the phonocardiogram and its internet-based application. Telemed J E Health 2001;7(4):333–339.
47. Stern DT, Mangrulkar RS, Gruppen LD, Lang AL, Grum CM, Judge RD. Using a multimedia tool to improve cardiac auscultation knowledge and skills. J Gen Intern Med 2001; 16(11):763–769.
48. Woywodt A, Herrmann A, Kielstein JT, Haller H, Haubitz M, Purnhagen H. A novel multimedia tool to improve bedside teaching of cardiac auscultation. Postgrad Med J 2004; 80(944):355–357.

See also ELECTROCARDIOGRAPHY, COMPUTERS IN; GRAPHIC RECORDERS.

PHOTOTHERAPY. See ULTRAVIOLET RADIATION IN MEDICINE.

PHOTOGRAPHY, MEDICAL

JACKIE K. CHAN
EDWARD K. FUNG
Columbia University
New York

INTRODUCTION

Photography is widely used in many areas of medicine for the documentation and treatment of diseases. Photography involves making pictures by capturing light reflected from objects onto a sensitive medium (e.g., film or the more recent technique of light-sensitive chips from a digital camera). In ophthalmology, the transparency of the living eye allows photographs to image diseases as far back as the retina. In dermatology, traditional methods of photography are used to document and track skin lesions. Every type of medical imaging comes with its own technical challenges. The medical photographer plays a vital part in promoting and supporting quality healthcare by providing services in photography. Furthermore, imaging has served as an important research tool. Photomicrography involves taking images in the laboratory of tissue or culture specimens in both the gross and cell level. The goals of photography, in general, may include characterizing the basic anatomy and physiology of the body, understanding changes caused by aging or disease, and discovering disease mechanisms.

OPHTHALMIC PHOTOGRAPHY

Photographing the living eye poses some challenging issues despite its transparency. The eye is sensitive to light and can easily be bleached after a certain number of flashes and intensity. The human retina is also designed for capturing light rather than reflecting it. Images may result in poor contrast and may affect the performance of diagnostic procedures. The main absorbing pigments in the eye are blood, hemoglobin, photo pigments, macular pigments, and water. Moreover, research has shown that many sight-threatening diseases are embedded deep in the retina, where conventional tools of photography cannot be used (1,2). Fortunately, specialized instruments and image enhancement processes have been developed to obtain better images (Fig. 1).

The Fundus Camera

The instrument widely used by ophthalmologists to view the posterior segment of the eye is the fundus camera. The fundus is photographed using a white light source to provide high resolution images at the micron range. Fundus photography can also create stereographics images that provide depth. These qualities make fundus photography the established standard for clinical studies of macular diseases. Newer models now take digital images and can be directly stored to a computer database.

Many fundus camera models are currently available in the United States: the German Zeiss, the Topcon fundus camera, the Olympus fundus camera, and the Nikon fundus camera. In choosing an instrument, one should compare the engineering and more importantly, the photo quality. Some instruments will be more expensive than its competitors, but takes excellent photographs while others have a wider view and good quality images (3) (Fig. 2).

General Methodology

A full-time photographer is specialized to operate the equipment in a clinic. While gaining technical skills, the photographer is often familiar with the clinical pathology of the fundus. This is advantageous to the ophthalmologist in situations where photos need to be interpreted. In operating a fundus camera, there are some general guidelines to follow (3):

1. The pupil of the patient must be dilated. Dilation of the pupil may take 20–30 min. An 8 mm diameter pupil is ideal, but even a pupil much smaller may be acceptable.
2. The eyepiece must be carefully focused to avoid getting out-of-focus photographs. Young children will have accommodation problems and may pose extra attention. The settings should be checked before each set of photographs on the patient is taken.
3. Check the shutter speed for the proper electronic flash synchronization. Shutter speeds in the range of 1/30 to 1/60 of a second can be used.
4. Take an identification photograph of the patient's name, date photographed, and other pertinent information. Properly labeling photos will avoid confusion later.

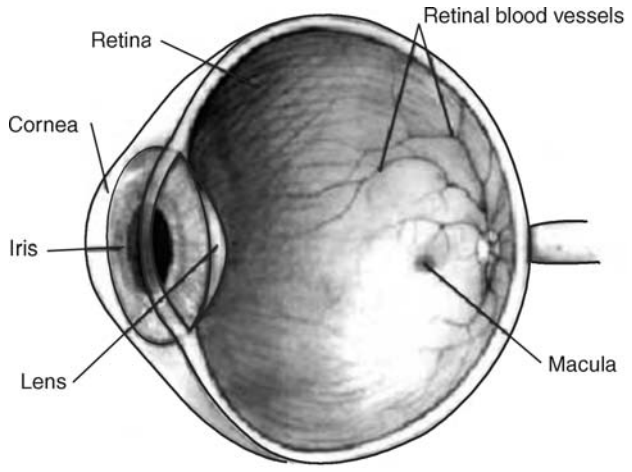


Figure 1. Cross-section of the eye. The transparency of the eye permits the ability to receive light from the external world. Light enters the cornea and goes through the iris and the lens until it reaches the retina. The macula contains many rods and cones and is the area of greatest visual acuity.

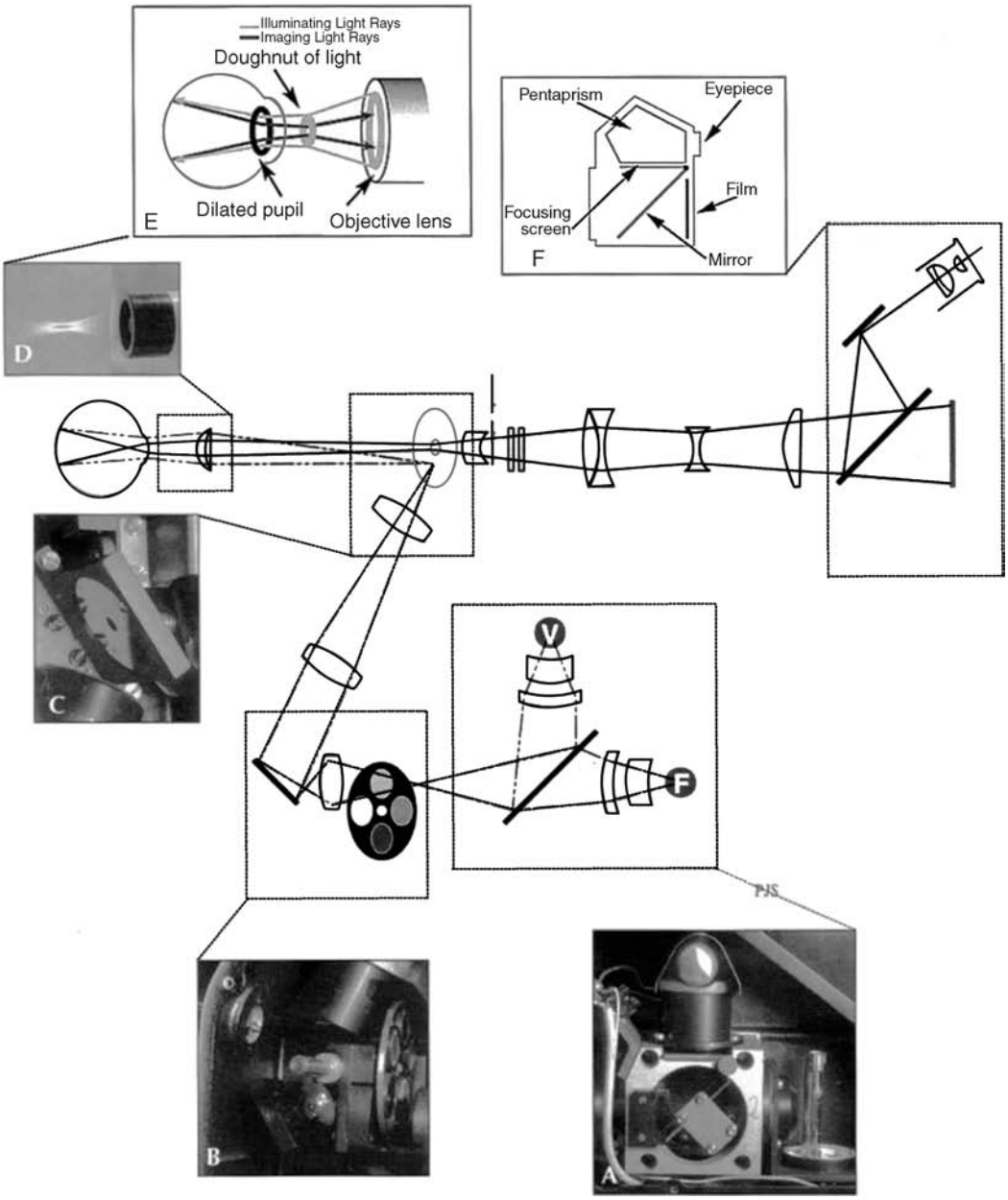


Figure 2. The optical pathway of the fundus camera. Light generated from either the viewing lamp (V) or the electronic flash (F) (A) is projected through a set of filters and onto a round mirror (B). This mirror reflects the light up into a series of lenses that focus the light. A mask on the uppermost lens shapes the light into a doughnut. The doughnut shaped light is reflected onto a round mirror with a central aperture (C), exits the camera through the objective lens, and proceeds into the eye through the cornea. When the illuminating light is seen from the side, one can appreciate the complexity and precision of the optical scheme (D). Assuming that both the illumination system and the image are properly aligned and focused, the retinal image exits the cornea through the central, unilluminated portion of the doughnut (E). The light continues through the central aperture of the previously described mirror, through the astigmatic correction device and the diopter compensation lenses, and then back to the single lens reflex camera system. (From Saine PJ, Taylor ME, Ophthalmic Photography (2nd ed.), Butterworth-Heinemann, 2002 with permission.)

5. When the patient is comfortably seated at the instrument with their chin on the chin rest and their forehead against the headrest, instruct the patient to look at the fixation device.
6. After properly focusing the filament of the viewing lamp, look through the eyepiece and bring the retinal vessels into focus. Release the shutter.

Stereo Fundus Photos

Fundus photos in stereo pair give a perception of depth that greatly improves reading performance. For example, a stereo photo can document blurred disk margins, optic nervehead cupping, and the degree of retinal elevation from conditions (e.g., serous or solid detachments). When evaluating patients with age-related macular degeneration (AMD), graders can view the elevation of drusen under the retina much more effectively.

The most popular method of taking stereo photographs was the cornea-induced parallax method, advocated by Bedell (4). To take the stereo pair, the first photograph is taken through the temporal side of the pupil and the second through the nasal side. A lateral shift of 3.5 mm is recommended for optimum stereopsis, but any lateral shift will create a stereo photograph. The photographer aligns and focuses the camera through the center of the pupil, then uses the joystick to move the camera slightly to the right, takes an image, and then slightly to the left, and takes the second image. The translation of the camera sideways changes the angle of view. The amount of camera shift can vary from image pair to image pair, making apparent depth a variable. Attachments are available for some cameras that allow the photographer to shift the camera through a consistent distance. Specialized cameras are commercially available to take stereo-pair fundus photos simultaneously, though the resolution is reduced.

Digital Imaging

Until recently, fundus photos were developed in film and the photographer had to wait hours or even days to see the results. The resolution was limited by the grain size of the film. Today, fundus cameras take digital photos that can be evaluated instantaneously and stored digitally. A typical image is in 24-bit, red, green, and blue (RGB), true color with a resolution size of 2000×2000 pixels. A digital imaging system ensures that the original quality of the image is preserved and will produce flawless duplication.

Images can be stored in a number of different formats, but with digital files comes the requirement for an efficient digital storage system. Tagged image file format (TIFF) images are not compressed, and therefore require large storage spaces. Recent advances in computer engineering have provided large storage in affordable costs. Nevertheless, it has been shown by Lee (5) that joint photographic experts group (JPEG) formats allow for varying degrees of image compression without compromising the resolution of fundus photos necessary for image analysis. In image compression, images are applied to a lossy algorithm, which permit the image to be reconstructed based on partial data. The result is not an exact restoration of the image, but is sufficient for diagnostic purposes.

In the evaluation of clinical AMD, Lee reported that TIFF images and low compression (30:1) JPEG images were virtually indistinguishable (5). Digital images of AMD patients were analyzed in software for drusen identification and quantification. Drusen detection in the conventional stereo fundus slides using a manual protocol was highly comparable to the digital format (Figure 3).

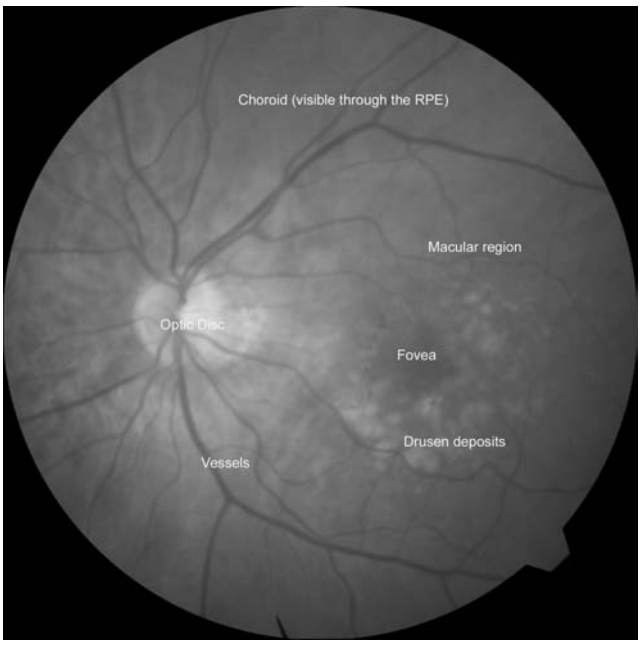


Figure 3. Fundus image of an AMD patient. The fundus camera is used to photograph the retina. The macula corresponds to an area of the retina 5–6 mm in diameter, centered on the fovea. Drusen are the white deposits surrounding the fovea that are characteristic of AMD, a sight-threatening disease. The retina has many layers, including the retinal pigment epithelium (RPE) and the choroid. Drusen are usually found embedded below the RPE layer.

Clinical Evaluation of AMD in Fundus Photos

The fundus camera has been routinely employed for diagnostic purposes (e.g., the clinical study of patients with AMD) (6,7), which is the leading cause of blindness in the developed world (8). The natural history of AMD is hallmarked by a subretinal pathology known as drusen (9–17). The identification and measurement of drusen are central to clinical studies of early AMD. The current standard of grading AMD is through manual viewing of stereographic fundus slides on a light box. However, this method involves time-consuming analysis of drusen size, number, area, and morphology in several subcategories (18).

Despite the movement to digital fundus imaging, a digital and automated method of quantification of macular pathology in AMD has yet to gain widespread use. Computer-assisted image analysis offers the potential for greater accuracy, objectivity, and reproducibility, but designing algorithms for that purpose is nontrivial. Many methods have been attempted in the last two decades with unsatisfactory results (19–25).

One major obstacle stems from the clinical appearance of the normal macula, which is a composite of complex light reflections from and absorption by multiple layers of the retina and associated structures. The application of segmenting any pathology superimposed on the macula in an automated fashion is a difficult task by the nonuniform reflectance of the normal macular background. For example, absorption by luteal pigment in the central macula

(fovea and parafovea) contributes to the darker central appearance. The nerve fiber layer, conversely, is highly reflectant. It is thickest in the arcuate bundles along the arcades and thinnest at the central fovea. This makes the arcade regions relatively brighter, hence also contributes to the macula appearing darker centrally (26). Therefore, simply choosing a global threshold would not be equally effective in segmenting or identifying drusen in the darker central regions as it would in the relatively brighter regions in the periphery, and vice versa.

Shin et al. (19) used adaptive thresholding techniques to handle the nonuniform macular background reflectance. They divided the image into separate windows of variable sizes. Within each window, a local histogram was applied to check for skewness, and to determine if drusen was present. However, the method was often misleading if either a large area of background or a large drusen dominated the region, which resulted in an incorrect threshold. Moreover, windows containing vessels would sometimes be incorrectly interpreted in the bimodal distribution. Thus, operator supervision and some postprocessing steps were added.

Shin's method was improved by Rapantzikos et al. (20), which used morphological operators (e.g., kurtosis and skewness) to predict whether drusen was present in the local window. Their idea gave better results, but was not infallible as a completely automated system of drusen segmentation. For example, many different combinations of image features (drusen and background) can yield the same histogram.

An alternative method was presented by Smith et al. (27,28) that aims to level the macular background reflectance, which can change significantly over distances of 50–100 μm. In previous methods, inadequate segmentation centrally and overinclusive segmentation in the peripheral macula was an indication that the background variability of the macula had not been resolved.

Smith found that the background reflectance of a normal fundus image could be modeled geometrically (29–31). It has been shown that a partial normal background containing drusen provided enough information to model the entire background by a elliptical contour graph (27,32). After leveling the nonhomogeneity of the background reflectance, they overcame the challenges posed in purely histogram-based methods of other researchers. A combined automated histogram technique and the analytic model for macular background presented a completely automatic method of drusen measurement.

Their algorithm is briefly explained. First, an initial correction of the large-scale variation in brightness found in most fundus photographs was applied. This is achieved by calculating a Gaussian blur of the image and subtracting it from the original image in each of the three RGB color channels. Further processing was carried out in this preprocessed image. The main idea was to level the background such that the reflectance was uniform over the entire macula. They proposed a multizone math model to reconstruct the macula. Each zone divided the macula into different annular and grid-like regions. The pixel gray levels were used as input for fitting into the custom software employing least-squares methods. After the geometric model of the macula was created, it was subtracted from the original image to obtain a leveled image. The process was automatically iterated until a sufficient leveling has been achieved. In other words, the range of gray levels in the image is minimized to an acceptable level.

The drusen, which is superimposed on the image, appears brighter than the regular intensity of the background. The final threshold was obtained by applying a histogram analysis approach to the final leveled image (33). An optimal threshold defined the separation of background areas from drusen areas.

Smith's technique has been validated successfully with the current standard of fundus photo grading by stereo pair viewing in the central 1000 and middle 3000 μm diameter subfields (28). One advantage of automation is portability of the software to use at other institutions. The work to create an automated algorithm will provide a useful, cost-effective tool in clinical trials of AMD.

Despite the advances in automated quantification of drusen, some obstacles still remain. Drusen identification may be confounded by other objects present in the image. A computer ultimately must be taught how to differentiate drusen from other lesions (e.g., as retinal pigment epithelial hypopigmentation, exudates, and scars). Implementation with neural networks or morphological criteria may prove effective in eliciting unique features in the lesions. For now, some cases may still require supervision in digital automated segmentation (Fig. 4).

SCANNING LASER OPHTHALMOSCOPE

Photography is often associated with taking a white light source to obtain an image. However, monochromatic sources of light at a specific wavelength are available with the scanning laser ophthalmoscope (SLO). Originally used as a research tool, it is increasingly favored with other clinical imaging standards now. Moreover, pupil dilation is unnecessary, and light levels are dim during acquisition (34). The reflectance and absorbance spectra of the structures of interest can be seen. For example, the SLO provides better penetration and give views of the choroids layer. By the same token, things of less interest on the retina (e.g., drusen) will not be seen (26).

In autofluorescence (AF) imaging, a 488 nm laser source with a 500 nm barrier filter is used (26). This light source reveals structures that intrinsically fluoresce, primarily due to lipofuscin and its main fluorophore, A_2E (35,36). Focal changes in AF or the changes in their spatial distribution are means of studying the health of the RPE. Previous studies have been made on correlations of change in AF distribution with pathological features (35,37–41). Focally increased AF (FIAF) refers to increased fluorescence in an area with respect to the rest of the background. This is abnormally high in patients with Stargardt disease and may be a marker for RPE disease in ARMD. In actual RPE death, as in geographic atrophy, there is focally decreased AF (FDAF), seen as a blackened area in the AF image (26).

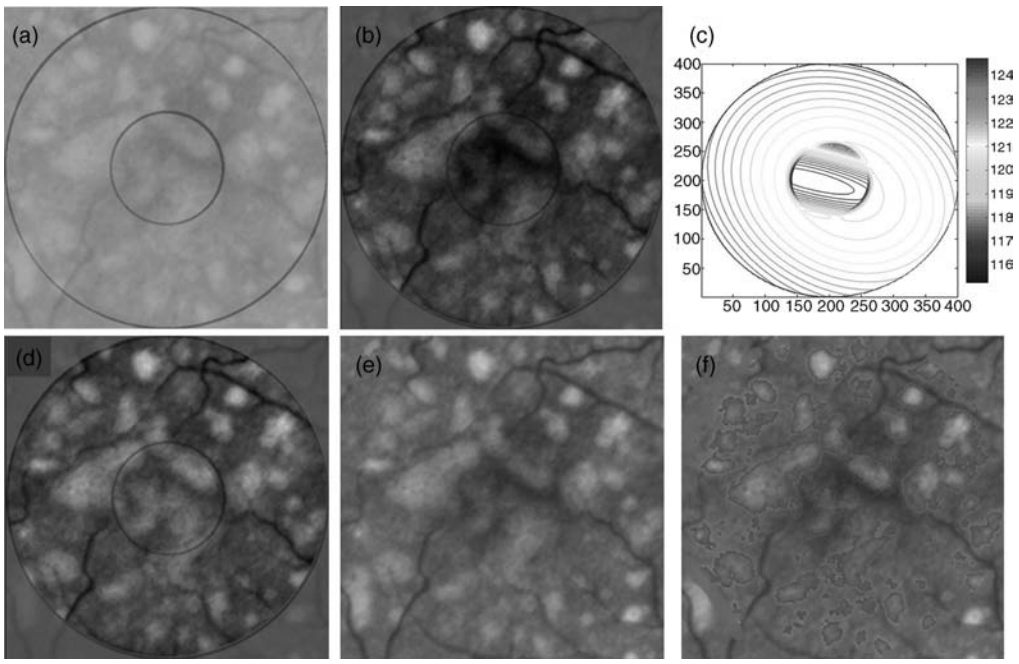


Figure 4. Automated digital photo grading. Color fundus photo with circular grading template overlaid on top (a). Green channel (grayscale) of fundus photo (b). Using an algorithm that finds the lower and upper thresholds, the macula is divided into three sections: vessels and darker perivasculature, normal background, and bright drusen. The normal background was used to calculate the mathematical model, displayed as a contour graph (c). The model was subtracted from the original image to yield a background leveled image of uniform intensity in (d). Contrast-enhanced image of a to show the boundaries of drusen (e). The algorithm then uses the leveled image (d) to calculate a threshold that determines the drusen areas (segmented in green) and overlaid on top of the contrast enhanced layer (f).

TELEMEDICINE

As the average age of populations in developed countries continue to rise, the number of elderly people needing an eye exam will boom. In less populated areas, an ophthalmologist may not be readily accessible. In the United States, less than one-half of the diabetic population received an annual eye examination (42). One of the goals of telemedicine is to bridge the gap between places where patients can be evaluated and where service is rendered. An examination usually involves taking the patient's fundus photos. With the advent of telemedicine and digital photography, the patient and doctor do not even have to be in the same room. Data can be efficiently and securely transferred across different institutions.

Telemedicine Framework

The goal of telemedicine is to establish a screening system with the ability to reach out to millions of unexamined patients for primary prevention. The following guidelines can be used to set up a generalized framework for telemedicine. We will use an ophthalmology setting as an example:

First, a patient will come into a primary care office to have their retina imaged. The camera itself should be operator-friendly, one that is easy to use by either a doctor's assistant or a technician. The disease to be screened should have identifiable pathology on the image. Examples would be diabetic retinopathy or AMD. The fundus photograph is then sent electronically to a reading center, which can be established at a far away university or hospital. Photos should be in a compressed digital format (e.g., JPEG).

In a paper by Sivagnanavel et al. (43), it was suggested that custom software for detecting AMD could be used at two different institutions. A grader at each institution independently ran the software and performed drusen quantification successfully, thus demonstrating portability and potential for automated software examinations. The results of the grading with software are comparable to grading manually by stereo-viewing. The only drawback is that the software may not be suitable in $\sim 20\%$ of the images taken. For the percentage of cases that need supervision, a trained reader can manually examine the images. Images that were disqualified had poorly identified areas or multiple types of lesions that were difficult to distinguish.

At the reading center, a patient report is generated and sent electronically back to the primary care physician (PCP). Finally, a good network referral base must be established such that the PCP can refer an ophthalmologist if the patient screening comes up positive.

Screening Instrument

Built differently from the classic fundus camera, a telemedicine instrument is meant for screening more than

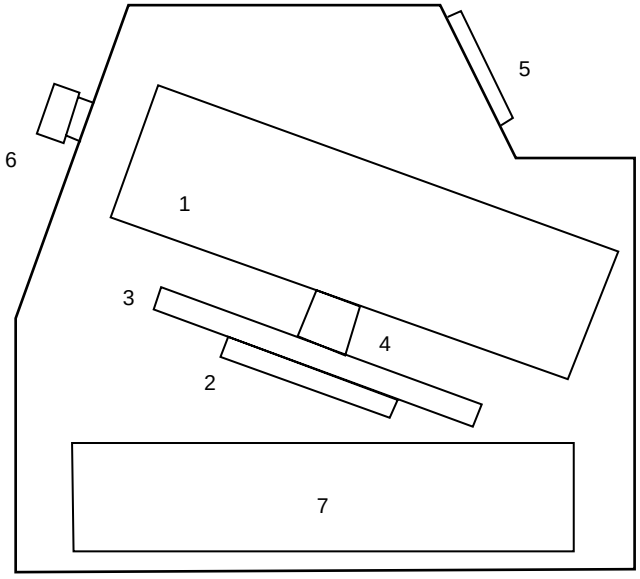


Figure 5. Layout of the DigiScope. The imaging head (1) is mounted on an *XYZ* motorized and computer-controlled stage. The *X*, *Y*, and *Z* components (2, 3, and 4, respectively) allow movements in the superior-inferior direction, to and from the eye, and in the nasal-temporal direction, respectively. The operator interfaces with the camera by a touch screen (5). The subject leans the head and places it against the nosepad (6) to view the imaging system inside. The electronics and computer (7) that control the system and store the images are embedded in the machine and hidden from view. (used with permission for R. Zeimer, IOVS 43:5,2002.)

careful clinical diagnosis. In a telemedicine fundus camera (e.g., the Digiscope) (44), the goal was to create a good model for providing ophthalmic screening to the primary care office.

Keeping the instrument at low cost for production greatly increased its attractiveness for the primary doctor's office. Therefore, a video camera with a resolution of 50 pixels per degree provided enough detail comparable to a fundus photograph. Such cameras typically yield images with 930 pixels diagonally, or a field of 19°.

Figure 5 shows a simplified diagram of the DigiScope. The eye to be imaged is oriented by an internal light-emitting diode. Fixation and imaging covers the entire posterior pole and takes about eight images. The illumination is generated by a halogen bulb. Infrared light is eliminated by the first filter and the visible spectrum is limited to green by the second filter, which passes light between 510 and 570 nm. The beam then expands and illuminates the fundus.

There are two modes of imaging with the DigiScope. The first mode takes four frames per shutter, with the shutter lasting 130 ms. The four frames differ by the fine focus, such that the sharpest image can be chosen out of each set. The second mode is meant for stereoscopic effect. Four frames are acquired while the optical head moves horizontally along the x axis. A pair of images can be selected to generate a stereo effect.

The operator interfaces with the machine by a simple touch-screen. The duties of the operator are limited to basic tasks (e.g., explaining the procedure, encouraging the patient to fixate at the blinking light, and checking the quality of the image).

A non-mydriatic camera is in consideration to provide greater patient comfort. Poor imaging quality and lack of stereo may cause a tendency to include RPE atrophy as drusen. However, newer generation cameras taking higher resolution images, coupled with the ability for digital stereo, may eliminate such drawbacks in the future.

PHOTOMICROGRAPHY

Photomicrography in biomedicine is the creation of images of biomaterial at magnification. Tissue or culture samples can be photographically recorded at both the gross level and at the level of cells. Images can be used for archival purposes or for analysis. Photomicrography replaced earlier camera lucida apparatus that projected microscope images through a beam splitter onto a plane for manual tracing.

Still photographs of biological specimen have traditionally been captured on 35 mm film. The development and popularization of digital camera technology, however, has had a drastic impact on the field of photomicrography. The relative ease of use has spurred the application of photomicrography in many areas, especially pathology and basic research.

Digital Cameras

Digital cameras offer several advantages over their film variants for most uses. Most digital cameras include some type of LCD (liquid-crystal display) screen. This screen can be used to display a photograph right after it is taken, allowing the user to evaluate the result of different camera settings and adjust as needed. This is a great convenience given the unusual light conditions under which photomicrographs are generally taken and the fact that most scientific users will probably not have professional photography training. On consumer models, the LCD can also display a real-time image of the scene to be photographed. Furthermore, digital cameras store images in a digital format, usually on a removable memory card. Images can be directly transferred to a personal computer for analysis or processing without having to use a scanner to digitize as with 35 mm film. The formats come in JPEG or TIFF. Still others cameras have the option of outputting in the RAW format. This is the raw data from the CCD (charge-coupled device) array that forms the core of the camera. The CCD cameras have also proven particularly useful in extremely low light microscopy, such as in certain fluorescence microscopy situations. Quantum efficiency, a measure of light sensitivity, can reach 90% in a back illuminated electron multiplying CCD, compared to $\sim$ 20% for a conventional video camera (45,46) (Fig. 6).

Photomicroscope Description

The attachment of the camera to the microscope is essentially the same for both digital and film cameras though different adapters may be used. A camera can simply be held with the lens pressed flush against the eyepiece of a microscope or mounted with a special adapter for the

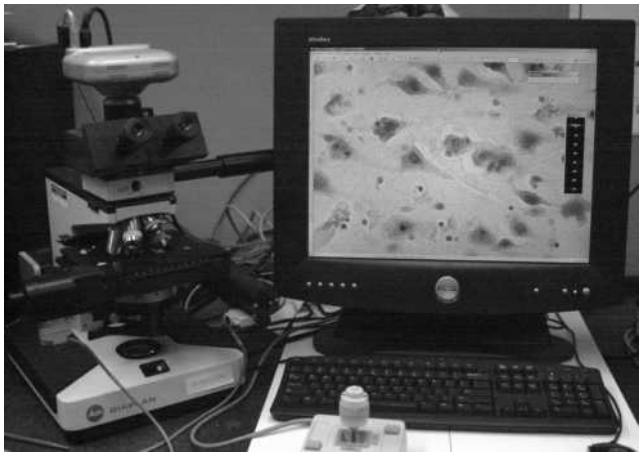


Figure 6. Retrofitted computerized microscope system with motorized stage and CCD digital imager directly connected to computer by Firewire cable. (used with permission from the laboratories of Victoria Arango, Ph.D., J. John Mann, M.D., and Mark Underwood, Ph.D. at NYSPI and Columbia University.)

eyepiece (46,47). This method can be used for many consumer digital cameras. Alternatively, a camera can be attached to the phototube or trinocular port of a microscope or through a special beam splitting adapter (47,48). Typically, SLRs (single lens reflex) and cameras specifically designed for photomicrography are attached in this way by their lens mounts. The tube thread standards commonly used in professional cameras to attach the lens are C-, T-, and F-mounts. Relay lens adapters may be necessary to correctly focus the image onto the film or CCD array in the camera as the normal lens is removed to use the C-, T-, or F-mount. Many of the major camera brands as well as third-party manufacturers produce these adapter kits. The primary concerns when attaching a camera are ensuring correct focus and avoiding vignetting. Vignetting refers to the darkening of the edges of an image (47). Proper Koehler's illumination of the microscope, as well as focusing and zooming should correct this effect.

The CCD technology has also been applied successfully to real-time imaging through the microscope. Dedicated computer microscopy systems with CCD cameras have been made available by the major microscope manufacturers (e.g., Leica and Zeiss). The CCD video cameras can also be attached to existing microscope setups through the phototube as with still cameras. Traditional CCTV cameras can usually output in NTSC (National Television System Committee) or PAL (Phase Alternating Line), the common video standards in the United States and Europe, respectively. An image capture board is necessary to convert the TV signal for use with a computer. The CCD digital imagers specifically designed for scientific or microscopic use often can be connected by means of more conventional computer ports [e.g., Firewire (IEEE 1394) and USB (universal serial bus)]. These are preferred due to their resistance to radio frequency (RF) interference and higher resolutions (49). In both cases, specialized software is needed to manipulate and display the video image. Some of these imagers can be used to capture high resolution still images as well.

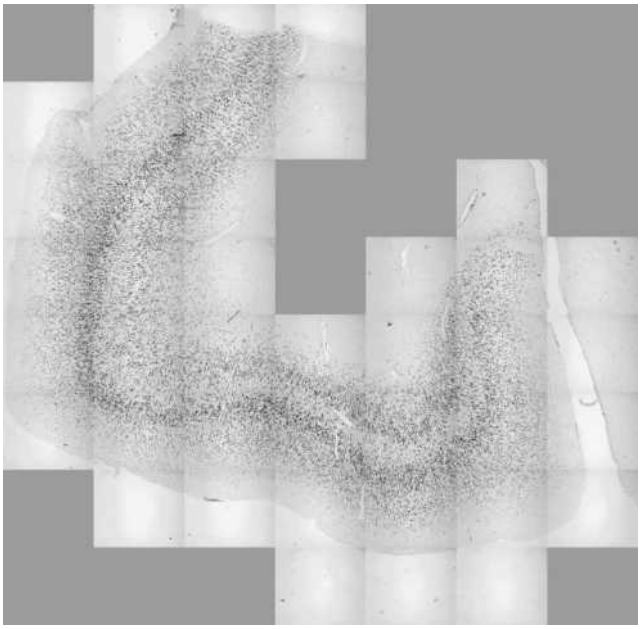


Figure 7. Montage of photomicrographs taken with 40× objective and stitched together with Neurolucida software (Microbrightfield, Inc., Williston, VT). Brodman area 24 of the prefrontal cortex is depicted. (used with permission from the laboratories of Victoria Arango, Ph.D., J. John Mann, M.D., and Mark Underwood, Ph.D. at NYSPI and Columbia University.)

For video applications, a more expensive three-chip, one-shot CCD system is preferred (46). These cameras use an array of photodetectors to record an image. One such array is referred to as a chip. Currently, cameras are available in one-chip and three-chip designs. A one-chip design produces color images by switching three colored filters over the array. This arrangement is referred to as one-chip, three-shot. Due to the switching of RGB (red, green, blue) filters, the framerate is necessarily reduced. A three-chip, one-shot camera splits the individual color channels and directs each to its own CCD array, imaging them simultaneously. Thus a higher framerate is maintained.

Camera computer systems integrated with a motorized microscope stage make possible more complicated photomicrograph sets. A computer controlled stage and camera can be programmed to take many overlapping digital images. With special software, these can be stitched together to form high magnification photomicrographs of large tissue sections (50). This represents an interesting alternative to low magnification macroscopic photography of tissue specimen. Systems have also been developed for 3D reconstruction using stacks of properly registered two-dimensional (2D) photomicrographs (Fig. 7).

BIBLIOGRAPHY

1. Elsner AE. Reflectometry with a Scanning Laser Ophthalmoscope. Appl Opt 1992;31:3697–3710.
2. Figueroa MS, Regueras A, Bertrand J. Laser photocoagulation to treat macular soft drusen in age-related macular degeneration. Retina 1994;14(5):391–396.

3. Yannuzzi LA, Gitter KA, Schatz H. The Macula: A Comprehensive Text and Atlas. Fundus Photography and Angiography. In: Justice JJ, editor. Baltimore: Williams & Wilkins; 1979.
4. Bedell AJ. Photographs of the Fundus Oculi. Philadelphia: F.A. Davis; 1929.
5. Lee MS, Shin DS, Berger JW. Grading, image analysis, and stereopsis of digitally compressed fundus images. Retina 2000;20(3):275–281.
6. Bird AC, et al. An international classification and grading system for age-related maculopathy and age-related macular degeneration. The International ARM Epidemiological Study Group. Survey Ophthal 1995;39(5):367–374.
7. Klein R, et al. The Wisconsin age-related maculopathy grading system. Ophthalmology 1991;98(7):1128–1134.
8. Klaver CC, et al. Age-specific prevalence and causes of blindness and visual impairment in an older population: the Rotterdam Study. Arch Ophthal 1998;116(5):653–658.
9. Smiddy WE, Fine SL. Prognosis of patients with bilateral macular drusen. Ophthalmology 1984;91(3):271–277.
10. Bressler SB, et al. Relationship of drusen and abnormalities of the retinal pigment epithelium to the prognosis of neovascular macular degeneration. The Macular Photocoagulation Study Group. Arch Ophthal 1990;108(10):1442–1447.
11. Bressler NM, et al. Drusen characteristics in patients with exudative versus non-exudative age-related macular degeneration. Retina 1998;8(2):109–114.
12. Holz FG, et al. Bilateral macular drusen in age-related macular degeneration. Prognosis and risk factors. Ophthalmology 1994;101(9):1522–1528.
13. Abdelsalam A, Del Priore L, Zarbin MA. Drusen in age-related macular degeneration: pathogenesis, natural course, and laser photocoagulation-induced regression. Survey Ophthalmol 1999;44(1):1–29.
14. Little HL, Showman JM, Brown BW. A pilot randomized controlled study on the effect of laser photocoagulation of confluent soft macular drusen [see comments]. Ophthalmology 1997;104(4):623–631.
15. Frennesson IC, Nilsson SE. Effects of argon (green) laser treatment of soft drusen in early age-related maculopathy: a 6 month prospective study. Br J Ophthalmol 1995;79(10):905–909.
16. Bressler NM, et al. Five-year incidence and disappearance of drusen and retinal pigment epithelial abnormalities. Waterman study. Arch Ophthalmol 1995;113(3):301–308.
17. Bressler SB, et al. Interobserver and intraobserver reliability in the clinical classification of drusen. Retina 1988;8(2): 102–108.
18. Age-Related Eye Disease Study Research Group, T. The Age-Related Eye Disease Study System for Classifying Age-related Macular Degenration From Stereoscopic Color Fundus Photographs: The Age-Related Eye Disease Study Report Number 6. Am J Ophthalmol 2001;132(5):668–681.
19. Shin DS, Javornik NB, Berger JW. Computer-assisted, interactive fundus image processing for macular drusen quantitation [see comments]. Ophthalmology 1999;106(6):1119–1125.
20. Rapantzikos K, Zervakis M, Balas K. Detection and segmentation of drusen deposits on human retina: Potential in the diagnosis of age-related macular degeneration. Med Image Analysis 2003;7(1):95–108.
21. Sebag M, Peli E, Lahav M. Image analysis of changes in drusen area. Acta Ophthalmol 1991;69(5):603–610.
22. Morgan WH, et al. Automated extraction and quantification of macular drusen from fundal photographs. Aust New Zealand J Ophthalmol 1994;22(1):7–12.
23. Kirkpatrick JN, et al. Quantitative image analysis of macular drusen from fundus photographs and scanning laser ophthalmoscope images. Eye 1995;9(Pt 1):48–55.
24. Goldbaum MH, et al. The discrimination of similarly colored objects in computer images of the ocular fundus. Invest Ophthalmol Vis Sci 1990;31(4):617–623.
25. Peli E, Lahav M. Drusen measurement from fundus photographs using computer image analysis. Ophthalmology 1986;93(12):1575–1580.
26. Smith RT. Retinal Imaging and Angiography, Basic Science Course. New York: Eye Institute, Columbia University; 2005.
27. Smith RT, et al. A method of drusen measurement based on reconstruction of fundus reflectance. Br J Ophthalmol 2005;89(1):87–91.
28. Smith RT, et al. Automated detection of macular drusen using geometric background leveling and threshold selection. Arch Ophthalmol 2005;123:200–207.
29. Smith RT, et al. Patterns of reflectance in macular images: representation by a mathematical model. J Biomed Opt 2004;9(1):162–172.
30. Smith RT, et al. The fine structure of foveal images. Invest Ophthalmol Vis Sci 2001;42(Mar. Suppl.):153.
31. Smith RT, et al. A two-zone mathematical model of normal foveal reflectance in fundus photographs. Invest Ophthalmol Vis Sci 2003;44:E-365.
32. Chan JWK, et al. A Method of Drusen Measurement Based on Reconstruction of Fundus Background Reflectance. Invest Ophthalmol Vis Sci 2004;45(5):E-2415.
33. Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Systems Man Cybernetics 1979;9(1):62–66.
34. Elsner AE, Weiter JJ, Jalkh AE. New Devices for Retinal Imaging and Functional Evaluation. In: Freeman WR, ed. Practical Atlas of Retinal Disease and Therapy. New York: Raven Press; 1993. pp. 19–35.
35. Delori FC, et al. In vivo fluorescence of the ocular fundus exhibits retinal pigment epithelium lipofuscin characteristics. Invest Ophthalmol Vis Sci 1995;36(3):718–729.
36. von Ruckmann A, Fitzke FW, Bird AC. In vivo fundus autofluorescence in macular dystrophies. Arch Ophthalmol 1997;115(5):609–615.
37. Feeney-Burns L, Berman E, Rothman H. Lipofuscin of the human retinal pigment epithelium. Am J Ophthalmol 1980;90:783–791.
38. von Ruckmann A, Fitzke FW, Bird AC. Distribution of fundus autofluorescence with a scanning laser ophthalmoscope [comment]. Br J Ophthalmol 1995;79(5):407–412.
39. von Ruckmann A, Fitzke FW, Bird AC. Fundus autofluorescence in age-related macular disease imaged with a laser scanning ophthalmoscope. Invest Ophthal Vis Sci 1997;38(2):478–486.
40. Feeney-Burns L, Hildebrand E, Eldridge S. Aging human RPE: Morphometric analysis of macular, equatorial, and peripheral cells. Invest Ophthalmol Vis Sci 1984;25:195–200.
41. Wing G, Blanchard G, Weiter J. The topography and age relationship of lipofuscin concentrations in the RPE. Invest Ophthalmol Vis Sci 1978;17:600–607.
42. Mukamel D, et al. Barriers to compliance with screening guidelines for diabetic retinopathy. Ophthal Epidemiol 1999;6:61–72.
43. Sivagnanavel V, et al. An Interinstitutional Comparative Study and Validation of Computer-Aided Drusen Quantification. Br J Ophthalmol 2005;89:554–557.
44. Zeimer R, et al. A fundus camera dedicated to the screening of diabetic retinopathy in the primary-care physician's office. Invest Ophthalmol Vis Sci 2002;43(5):1581–1587.
45. Coates CG, et al. Optimizing low-light microscopy with back-illuminated electron multiplying charge-coupled device: enhanced sensitivity, speed, and resolution. J Biomed Opt 2004;9:1244–1252.

46. Riley RS, et al. Digital photograpy: a primer for pathologists. J Clin Lab Anal 2004;18:91–128.
47. Hamza SH, Reddy VVB. Digital image acquisition using a consumer-type digital camera in the anatomic pathology setting. Adv Anat Pathol 2004;11:94–100.
48. Haynes DS, et al. Digital microphotography: a simple solution. Laryngoscope 2003;113:915–919.
49. Hand WG. Practical guide to digital imaging for microscopy. Biores Online; 2000.
50. Berggren K, et al. Virtual slice: a novel technique for developing high-resolution digital brain atlases. Society for Neuroscience Annual Meeting, Miami; 2002.

See also ENDOSCOPES; FIBER OPTICS IN MEDICINE; IMAGING DEVICES; MEDICAL EDUCATION, COMPUTERS IN; PICTURE ARCHIVING AND COMMUNICATION SYSTEMS; RADIOLOGY INFORMATION SYSTEMS.

PHYSIOLOGICAL SYSTEMS MODELING

N. TY SMITH
University of California,
San Diego, California
KENTON R. STARKO
Point Roberts
Washington

INTRODUCTION

First, physiological systems need to be characterized. The word "system" means an interconnected set of elements that function in some coordinated fashion. Thus, the heart, with its muscles, nerves, blood, and so on, may be regarded as a physiological system. The heart, however, is a subsystem of the circulatory system, which is in turn a subsystem of the body. Dynamic systems are time-varying systems, and most physiological systems fall into this category. This article deals with classical, or macro, physiological systems, usually ignoring genetic, cellular and chemical systems, for example.

It should come as no surprise that the term super system has been used to describe the brain and the immune system (1). The cardiovascular system (CVS) and central nervous system (CNS) are closer to being super systems, however.

Correct terminology is vital. To paraphrase George Bernard Shaw, engineering and medicine are two disciplines separated by the same language. The same word may have two different meanings (system) or two different words may have the same meaning (parameter and variable; model and analogue). The latter is particularly pertinent. Medical researchers working with animals because of their resemblance to humans use the term "animal model", while engineers would use the more descriptive "animal analog".

Unless otherwise stated, model, We means mathematical model. Thus, these models are generally ignore: animal, *In vitro*, chemical, structural (Harvey's observations on the circulation), and qualitative (Starling's law of the heart).

What is a mathematical model? A mathematical model consists of elements each describing in mathematical terms the relation between two or more quantities. If all the descriptions are correct, the model will simulate the behavior of real-life processes. Especially when many different subprocesses are involved, the models are useful in helping understand all the complex interactions of these subsystems. A useful characteristic of a mathematical model is that predictions of the outcome of a process are quantitative. As Bassingthwaighte et al. (2) point out, these models are therefore refutable, thereby facilitating the entire process of science.

The difference between modeling and simulation is important. Modeling attempts to identify the mechanisms responsible for experimental or clinical observations, while with simulation, anything that reproduces experimental clinical, or educational data is acceptable. For this article, when you run a model, you are performing a simulation. Thus, simulation is not necessarily an overworked word. Most of the models described here are not simulators, however. That word should be reserved for those models that allow a nontechnologically oriented user to interact with the model and run a simulation. The words simulation or simulator are used when appropriate.

Having said that, the backgrounds and connotations of simulate and model are strikingly different. Reading the etymology of simulate and simulation is a chastening process, emphasizing as it does the unsavory past of the terms. In contrast, model has a more virtuous past. The Latin simulare can mean to do as if, to cheat, to feign or to counterfeit. Hence, a set of its meanings, from the OED, includes To assume falsely the appearance or signs of (anything); to feign, pretend, counterfeit, imitate; to profess or suggest (anything) falsely. The action or practice of simulating, with intent to deceive; false pretence, deceitful profession.

The Middle French modelle, on the other hand, implies perfect example worthy of imitation. Thus, one definition, again from the OED, is Something which accurately resembles or represents something else, esp. on a small scale; a person or thing that is the likeness of another. Freq. in the (very) model of.... Another definition is "A simplified or idealized description or conception of a particular system, situation, or process, often in mathematical terms, that is put forward as a basis for theoretical or empirical understanding, or for calculations, predictions, etc.; a conceptual or mental representation of something." Our final definition is "A person or thing eminently worthy of imitation; a perfect exemplar of some excellence. Also: a representative specimen of some quality."

This remarkable difference between the two terms can be transmuted to our basic philosophy: A simulation is no better than the model that drives it. Put another way, the model is arguably the most important part of the simulation. It helps prevent the simulation from being something that feigns.

Model Reduction versus Model Simplification

We were asked to consider the introduction of reduction, to distinguish from simplification, of the size of available models by weeding out minor effects. These concepts are indeed essential for the physiome project (at end of this

article). Unfortunately, the literature that we have found usually fails to distinguish between the two terms, and too often uses them interchangeably. The following fragment of a definition was particularly discouraging: "Model reduction is the simplification or reduction of a mathematical model..." Not only is it a partial circular definition, but it also uses the term simplification as a way of achieving reduction of a model.

Dr. George Hutchinson, of GE Healthcare, has provided considerable help, and the remainder of this section comes from his ideas, as well as from the literature. He suggests that the difference between the two terms is in the purpose and the effect rather than in the action. Simplification of a model will eliminate elements, knowing that accuracy or validity may be significantly compromised. This simplification may be necessary because of (1) limitations in the platform running the model, (2) a decision to delay the work to model this portion of the system, or (3) this part of the model's not being germane to the focus of the model's intent. BODY Simulation does not model a real electrocardiogram (it is constructed from a look-up table) and that is a simplification of the model. This does compromise the overall validity of the model, but it was a simplification needed to make the rest of the model available to the public; this currently unnecessary submodel would have overwhelmed the rest of the model.

On the other hand, reduction is the elimination of some elements of a model in a way that should not significantly affect the accuracy or validity of the overall model. This could be done to streamline the code, to improve the faster-than-real-time performance, or to allow the model to run on a lesser platform. Again, with BODY Simulation as an example, deciding to limit the modeling of the arterial tree to large peripheral arteries, but not to extend it to the digits or the capillaries, is a reduction of the overall model, but it does not affect the intended use of BODY. The arterial waveforms are still good enough and the action of the MAP is unaffected.

This suggests that reduction and simplification are relative. Relative morality may be unacceptable, but relative definitions are reasonable. In a model built for one purpose, elimination of elements is reduction. In the same model built for a different purpose, such elimination is simplification. The problem of model reduction is to find a smaller system such that the number of components is much smaller than the original and the transfer function of the new system is close to the original transfer function.

Models, almost by definition, are necessarily simplifications and abstractions, to some degree, of the reality of the system. The modeler attempts to extract the important elements of the system (not all the variables, as that is not possible) and represent them so that they are simple enough to be understood and manipulated, yet realistic enough to portray the essential constructs and parameters of the situation. If either simplification or reduction is used, the techniques must be identified and justified, and the magnitude of their effect quantified.

Uncertainty analysis and sensitivity analysis are prerequisites for model building. The former allows quantifying the precision associated with the model response as a result of problems in the model input. Sensitivity analysis is aimed at establishing how the variation in the model output can be apportioned to different sources of variation, so that the modeler can establish how the given model depends on the information fed into it. Both are essential to assess the impact of simplification or reduction.

Here are some of the techniques that have been proposed for simplification and for reduction. Although these methods do overlap in their intended use, we have chosen to include them only with one or the other term. Also, lack of space precludes our going into detail about any of them.

Simplification

Eliminate short-term changes, when long-term ones only are important.

Make sure that the parameters and variables are important to the model.

Use a coarser (simpler) finite-element model.

Use intermediate variables, that is, reduce the number of input and output variables.

Use the reduced basis technique.

Simplify equations.

Simplify assumptions.

Use algebraic equations, instead of differential equations.

Reduction

Reduce the number of dimensions in a model.

Narrow down the search space among the input parameters.

Considers only significant inputs, as opposed to all inputs, of the full model.

Eliminate redundant information.

Use a numerical model reduction technique that assumes no knowledge of the system involved.

With such a large topic, one must draw the line somewhere, and the line used is rather large. For example, CNS is an enormous topic, and we were have to be content with a small subset: cerebral circulation. In general, pathology, diseases, or their effects are rarely addressed, although normal aging is discussed. Also usually avoided are the following systems and topics: GI, immune, CNS, peripheral nervous, hepatic, enzyme, cellular physiology, receptors and coagulation, as well as the effects of altitude, temperature changes, diving, microgravity, exercise, conditioning, hypertension, or pregnancy, for example. There will be little discussion of pediatrics, obstetric and other specialties. The ANS and acid–base regulation will be included with some extensive models, but not discussed as a separate topic. However, physiological PKPD models, which have been influential in the development of whole-body physiological models are included.

Since many physiological models involve control systems, this feature will be emphasized, including a special kind of control system known as autoregulation. Much of physiological systems modeling involves control systems. The body has many control systems, and only a few of them

can be addressed here. Those discussed will be embedded in larger models, of the CVS, for example.

This is not intended to be a critical review, but rather a review resource for those wishing to use models. A few of the uses for a few of the models are mentioned briefly, in addition to that of incorporating them into one's own model or simulation. An attempted is made to differentiate between where a model is described and where the author presents material, especially equations, that could be used in a model.

Data to construct a model can come from many sources, including the literature and an author's own adhoc experiments. Unless otherwise stated, assume that data generated from the usually performed validating experiments fit a given model reasonably well. Weinstein (3) contends, however, that this form of validation may not be enough. Arguably, one may endorse a standard of model presentation in which the model builder shows not only what works, but also where the model fails, or where it makes novel predictions that have yet to be tested. While Weinstein suggests a good paradigm, published models are offered to the world to test out. Models are hypotheses, and no one expects final proof of an hypothesis in a paper: or series of papers. Some hypotheses, like some of Einstein's, seem to be eternally hypotheses. Modeling is a perpetual process.

The following lists a few uses of models:

1. Education is certainly the primary one.
2. Compression of enormous amounts of data.
3. An hypothesis.
4. Simulation and simulators.
   (a) Manikin-based simulators, such as an operating-room simulator.
   (b) Part-task trainer. The trainer could include a model and a ventilator, to teach students how to use the ventilator.
   (c) Screen-based simulator
5. Construct pharmacological models.
6. Simplify concepts to the user, especially in a simulation or simulator. The user of a model does not need to understand the mathematics or software, no more than one must understand how a car works before one drives it.
7. Suggest counterintuitive concepts for further exploration.
8. Suggest experiments to perform.
   (a) Fill missing gaps of knowledge.
   (b) Answer a posited question.
   (c) Try to explain ostensible modeling anomalies.
   (d) Provide more data for the model.
   (e) Any combination.
9. Substitute for animals or people in experiments.
10. Control part or all of an experiment.
11. Serve as an important component in adaptive control systems. A generic patient model learns about the patient and thereby helps improve the closed-loop control of an infused agent.

The most practical use to the reader, researcher, or educator is to operate a model, that is, run one's own simulation with it. Although most published models can not do that without considerable effort on the user's part, several models or sets of models make the process much easier. Werner et al. (4) state that their excellent model is "available for everybody". Levitt, Don Stanski, and Stephen Shafer all have large amounts of freely available software and data. The physiome project has more resources available than most of us can handle. The BODY Simulation model (see PBPM) is available as asimulator or as a dynamic linked library (dll). Because of the structure of this model, the latter means that the model can be connected to other software, such as simulation interfaces, or to a piece of electronically driven equipment, such as a ventilator, a patient monitor or an anesthesia machine. All of these are described in further detail, below.

RESOURCES

Several books stand out as resources for the reader. The classic is the late Vince Rideout's scholarly monograph *Mathematical and Computer Modeling of Physiological Systems* (5). It covers modeling for the CV, respiratory, and thermal regulatory systems, as well as transport and multiple modeling and parameter estimation. Baan, Noordergraaf, and Raines edited the proceedings of a symposium on cardiovascular system dynamics (6). The participants represented the international leaders in their fields: cardiology, physiology, engineering and physics. The postpresentation discussions alone are worth finding the book. The book comprises 62 papers, and it is difficult to choose which to emphasize. Some of them, however, will be discussed in the appropriate section. Most articles are detailed and worth exploring, for many reasons. A book edited by Papper and Kitz (7), again the result of a symposium, is described below.

TYPES OF MODELS

Mathematical models in our area fall into various categories, including mechanistic, black box; physiological, pharmacological, pharmacokinetic, pharmacodynamic; multiple, transport; analogue, hybrid and digital. Each one except physiological, is briefly described.

Mechanistic models can be understood in contrast to black box models in which only the input–output relations are important and not how these are realized. Though sometimes very useful, black box models have the limitation that they can only be used for descriptive purposes. Mechanistic models, for example, the prediction of flow through a vessel on the basis of diameter and length and pressure fall over it, allow the test of hypotheses. In case of the vessel example, one can test whether the viscosity assumed is the right one. Mechanistic models are important in testing hypotheses and formulating new ones.

Pharmacological is divided into two parts. Pharmacokinetic refers to the uptake, distribution and elimination of an agent, while pharmacodynamic refers to the action of the agent on the body. This is best remembered by the

Table 1. What, where, and how can be transported[a]

What	What (cont.)	Where	Route
Mass	Toxins	Atmosphere	Blood
Momentum	Chemical warfare agents	Anesthesia machine/ventilator	Lungs
Energy	Bacteria	Organs	Nerves
Gases	Viruses	Tissues	Membranes
Drugs	Genes	Blood	
Electrolytes	Molecules	Cells	
Proteins	Atoms	Receptors	
Hormones			
Endocrines			
Heat			
Information			

[a]There is no connection among the terms in each row.

following: pharmacokinetic is what the body does to the agent, while pharmacodynamic refers to what the agent does to the body.

Most of the complex models that are described are modular. Each of these modules can be a model, and the entire model is called a multiple model. A model with two or more modules is called a multiple model. If a model has CVS and respiratory components, each of those is a model. Essentially, anything that can be transported (see Table 1 and the next paragraph) can be a submodel in a multiple model. This includes anything from O_2 to a bacterial species, to a specific atom. In addition, the physiological systems themselves are submodels, for example CV, respiratory, and liver. Thus, the model for BODY Simulation (see PBPM) has almost 100 modules, or submodels. Ideally, each model can be sufficient unto itself, and can therefore be tested before it is incorporated into the main model.

Transport models are accurately named. Essentially, they transport something from one place to another. That something, we call an agent. In the case of the circulation, they, of course, transport agents around and around. The concept of transport modeling was developed by chemical engineers (8), whose models transported mass, momentum, and energy. Transport modeling is incredibly powerful. One has to be optimistic and creative about transport modeling. Assume, until proven otherwise, that a model can transport anything, anywhere and by any route. Table 1 lists a few of the agents that have been or could be transported in a model.

The following describes the various types of transport.

Momentum transport. In the wave equations describing blood flow, the concern is with momentum, blood viscosity, and the elasticity of the vessel walls in determining pressures and flows of the system.

Mass transport. Blood and lungs carry many important substances, such as O_2, CO_2, and pharmacological agents. The diffusion of these substances into or out of tissue is often essential to a model.

Energy transport. The blood in vessels, as well as the air we breathe, carries heat energy. This heat may diffuse through tissues, although in a way different from the mass diffusion mentioned above. Also, energy transformation, as well as transport, occurs in some tissues, including muscle, heart, brain, kidney, and liver.

Information transport. Information is carried throughout the body via nerve fibers, much as messages are transmitted in a communication system. Hormones also carry information, moving mostly with the aid of the bloodstream.

Models can also be categorized by the type of computer that implements them: analogue,hybrid or digital. This is not trivial, because even the most powerful digital computer cannot approach the speed and power of an analogue or hybrid computer, and we have yet to implement the entire original Fukui hybrid model (9) on a digital computer.

Whole-Body Models

A whole-body model is arbitrarily defined as one that includes the circulatory and respiratory systems, plus at least two other major systems. There is no model that includes all the major systems, much less any one system in detail. The good whole-body models have not been used extensively in education, while most simulators have used adhoc models assembled to meet the perceived needs.

The earliest and still archetypal whole-body model is that developed by Guyton and co-workers (4,10–13). To understand Guyton's contributions is to understand his model, and vice versa. One of his most important legacies was his application of principles of engineering and systems analysis to CV regulation. He used mathematical and graphical methods to quantify various aspects of circulatory function before computers were widely available. He built analogue computers and pioneered the application of large-scale systems analysis to modeling the CVS before the advent of digital computers. As digital computers became available, his CV models expanded dramatically in size to include many aspects of cardiac and circulatory functions. His unique approach to physiological research preceded the emergence of biomedical engineering, a field that he helped establish and promote in physiology, leading the discipline into a much more quantitative science.

The Guyton model has five main empirically derived physiological function blocks, with many subcomponents. The model has nearly 400 parameters and is remarkable in

its scope. It comprises the following physiological subsystems, or modules:

Circulatory dynamics.
Nonmuscle oxygen delivery.
Muscle blood control and PO_2.
Vascular stress relaxation.
Kidney dynamics and excretion.
Thirst and drinking.
Antidiuretic hormone control.
Angiotensin control.
Aldosterone control.
Electrolytes and cell water.
Tissue fluids, pressures and gel.
Heart hypertrophy or deterioration.
Red cells and viscosity.
Pulmonary dynamics and fluids.
Heart rate and stroke volume.
Autonomic control.
Nonmuscle local blood flow control.

The Guyton model is alive and healthy. Werner et al. (4) coupled the Guyton model, which does not have a beating heart, to a pulsatile model. The new model comprises the hemodynamics of the four cardiac chambers, including valvular effects, as well as the Hill, Frank–Starling, Laplace, and ANS laws. They combined the two models because, with few exceptions, the extant published models were optimized for either studying short-term mechanics of blood circulation and myocardial performance—pulsatile models, for example—or mid- and long-term regulatory effects, such as exercise, homeostasis, and metabolism. In a gesture of intellectual philanthropy, the authors state that the program is written in the "C" language and is available to everybody (14). Fukui's model (9), actually, could do both short- and long-term studies. This was possible because it was implemented on a hybrid computer, and neither speed nor power was a consideration.

Another whole-body model, the Nottingham Physiology Simulator, uses a combination of CV, acid–base, respiratory, cerebrovascular, and renal models. Hardman et al. (15) partially validated this model for examining pulmonary denitrogenation, followed by apnea, by reproducing the methods and results of four previous clinical studies. They then used the model to simulate the onset and course of hypoxemia during apnea after pulmonary denitrogenation (replacing the nitrogen in the lung with O_2) (16). Several parameters were varied to examine their effects: functional residual capacity, O_2 consumption, respiratory quotient, hemoglobin concentration, ventilatory minute volume, duration of denitrogenation, pulmonary venous admixture, and state of the airway (closed versus open) The Nottingham group used their simulator for two other purposes. First, they assessed the accuracy of the simulator in predicting the effects of a change in mechanical ventilation on patient arterial blood–gas tensions. Second, they compared two methods of venous admixture estimation: one using the simulator and data commonly available in the intensive care unit, and the other using an isoshunt style calculation that incorporated assumed values for the physiological variables (17,18).

To read about the fascinating quest for a true whole-body model, (see Physiome), at the end of this article, plus (http://nsr.bioeng.washington/PLN) Other whole-body models will be discussed in the next section (physiologically based pharmacokinetic/pharmacodynamic models). Also, there are several whole-body models in the section on models of systems.

PHYSIOLOGICALLY BASED PHARMACOLOGICAL MODELS

It would seem outside the remit of this article to discuss pharmacological models, but some of them represent a source of detailed whole-body models. Many whole-body models were developed primarily as pharmacological models, and these models help flesh out what would otherwise be a scanty topic. Whole-body pharmacological models were originally called uptake-and-distribution models, but are now called a more respectable sounding physiological PKPD models. The term is broad, and very few physiological PKPD models are whole-body models, however.

It also turns out that pharmacology can be useful in physiological models, helping create realism. For example, by using epinephrine and norepinephrine, BODY Simulation simulates the CV response to pain or to hypercapnia simply by internally injecting one or both of these agents. Part of the realism relates to the fact that it takes time for circulating epinephrine, and its effects, to fade away: after the stimulus has vanished. The agent must be metabolized and redistributed, as in real life.

The history of these models is briefly explored. The senior author's interest in uptake and distribution was stimulated in the mid-1950s during Avram Goldstein's course on pharmacology, which highlighted the subject in a lucid, prescient way.

The desire for more accurate pharmacological modeling has had a significant impact on the development of better physiological models, especially whole-body models. Part of this improvement came from the need for more compartments, which were often physiological systems. The history of whole-body physiological PKPD models is embedded in the history of uptake and distribution, which goes back to the mid nineteenth century, when John Snow, the first epidemiologist and the first scientific anesthesiologist, made observations on the uptake and elimination of several agents that he was testing in patients, including chloroform (19,20). In the late nineteeth and early twentieth centuries, Frantz (21) and Nicloux (22) observed that among all the tissues, the brain had the highest tissue concentration of diethyl ether after inhalation of that agent. In 1924, Haggard (23) correctly surmised that this phenomenon was related to the brain's small size and relatively large blood flow, as well as ether's brain–blood partition coefficient, which he uncannily estimated to be 1.11, compared with the currently accepted 1.14.

Less than four decades after Haggard's papers, whole-body anesthetic models were appearing in the literature

(7,24–26). Because of the therapeutic and physicalchemical characteristics of anesthetic agents, especially inhaled, the early models usually contained fat, muscle and brain, plus low and high perfused compartments. The seminal book in the area was *Uptake and Distribution of Anesthetic Agents*, edited by Papper and Kitz (7). This multiauthored book summarized the extant modeling knowledge in clear detail. Most of the literature, and the book, were based on the uptake and distribution of inhaled anesthetic agents, with the notable exception of Henry Price (26), whose model on thiopental was apparently the first uptake and distribution model. Modeling pioneers featured in the book include Eger and Severinghaus.

As exciting and useful as the early models were, they suffered from at least four problems. (1) They were not pulsatile, that is, the heart did not beat and the lungs did not breathe. Incorporating pulsatility can solve many physiological and pharmacological problems. (2) They were linear. Neither CO_2, along with its regional distribution, nor ventilation, for example, changed as a function of the concentration of the agent—or vice versa. (3) They usually only dealt with one agent at a time. Thus, for example, one could not determine the interactions among two or more agents. [A brave exception was Rackow et al. (27). (4) They were not true physiological models, partly because of concern No. 1 and partly because none of the physiology of the organs was incorporated: they simply acted as agent capacitors.

Little changed from the 1960s until the early 1970s, when Ashman et al. (28), plus Zwart, Smith and Beneken. (29–31), published the first nonlinear models in this area. In our model, cardiac output, regional circulation, and ventilation changed as a function of agent concentration. These changes, in turn, affected the uptake of the agent. Our model used both mass transport and multiple modeling and was implemented on an analog–hybrid computer.

Another breakthrough also occurred in the 1970s: the incorporation of our uptake and distribution model into the Fukui hybrid-computer pulsatile model (9,32–35). Physiological and pharmacological modeling were united. Among other things, venous return, the effects of arrhythmias, and some drug interactions were now possible. Our original model (29–31) used the inhaled anesthetic halothane for the agent; Fukui's (9) used CO_2 and O_2. The second Fukui-Smith model (35) incorporated halothane, plus CO_2. All of these models were multiple, transport models, implemented on analog–hybrid or hybrid computers.

One major problem remained. Only about half a dozen computers, and people, in the world could run these models, and the wide spread use of simulation was not possible. In 1983 our laboratory translated the hybrid code into digital, on a VAX, in FORTRAN. In 1985, Charles Wakeland, from Rediffusion Simulation Corporation, transferred the code into C on a Sigmagraphics Iris 2300. Two processors were used, one for running the model and one for the graphics display. Finally, there was a portable simulator, portable if one were strong. More importantly, Sleeper could be used for teaching (36).

The next goal was to incorporate the model and simulator into a PC. This was accomplished in 1989 and was a major step. Now it was possible to bring simulation to a larger audience. Unfortunately, the PCs available then were not quite ready. In 1992, Starko completely revised the code, in DOS assembly language, making a simulator available on a laptop. Starko achieved an amazing fourfold increase in efficiency. He used many of the techniques used in flight simulation, his main work, to create stunning graphics and interfaces. In 1998, the code was converted to Windows in C++. The conversion allowed many new features that were not possible in DOS, however, DOS was much faster, and running the simulation in Windows slowed it down noticeably.

The model currently has 37 compartments, >90 agents, >80 parameters to set for each patient, and 45 parameters to set for each agent. The parameters are user settable. The latest additions to the agents have included cyanide and sarin (37,38), the latter a nerve gas. We were able to model cyanide toxicity because each organ and tissue has a changeable O_2 consumption. The nerve gases are actually physiological, involving as they do cholinergic and muscarinic effects. We could implement both agents because BODY Simulation has receptors, with agent concentration-effect curves that have user adjustable slopes (gamma) and amplitudes (IC50). All of the agents involved in the therapy for both toxins have also been incorporated, including the therapeutic byproducts methemoglobin, cyanmethemoglobin, and thiocyanate. Five equations relating to cyanide therapy have been incorporated. In keeping with our definition of a simulator (interactivity) the user can change the rate and, when appropriate, the equilibrium constants of the equations.

Because of the detail in BODY Simulation, it has been possible to incorporate the normal aging process (39). Over 60 user-changeable patient parameters are changed, directly or indirectly, to implement each elderly patient. Four patients were constructed, aged 65, 75, 85, and 95, although patients of any age can be implemented. To assess these patients, three scenarios were run: (1) administration of an anesthetic agent (thiopental) with CV and respiratory depressant properties, (2) hemorrhage of 1000 mL in 10 min, and (3) nonfatal apnea. The results confirm that the elderly generally have a decreased physiological reserve: the older the patient, the less the reserve.

More details on BODY Simulation can be found online (40).

Oliver et al. (41) used a whole-body physiologically based PK model that incorporated dispersion concepts. In whole-body PBPK models, each tissue or organ is often portrayed as a single well-mixed compartment with limited distribution and perfusion rate. However, single-pass profiles from isolated organ studies are more adequately described by models that display an intermediate degree of mixing. One such model is the dispersion model. A salient parameter of this model is the dispersion number, a dimensionless term that characterizes the relative axial spreading of a compound on transit through the organ. The authors described such a model, which has closed boundary conditions.

PKQuest (42–44), as the abbreviation implies, is also a pharmacokinetic model only, with no pharmacodynamics. Thus, the nonlinearities that some PKPD models possess

are lost. Levitt's whole-body models have 12 compartments, including the essential brain and heart. The models, as well as files of tissue and agent parameters, are easily and freely accessible online (45) or from links in the papers, which are published on BioMed Central. The other models described above, including BODY Simulation, use only intravenous or pulmonary administration of an agent, while PKQuest explores the gastrointestinal, intramuscular, and subcutaneous routes. These models have been used to explore many areas that not only elucidate the PK of agents, but also give insight into body compartments, thus enhancing our knowledge of the compartments and suggesting further areas to explore physiologically. These compartments include the interstitial space.

The Stanford group, led by Donald Stanski and Steven Shafer have developed an extensive array of PKPD models, some of them incorporating whole-body models. Details and considerable software, much of it public domain, can be found online (46). These models usually do not incorporate the nonlinearities associated with agent-induced or physiologically induced changes in patient physiology.

Many anesthesia simulators incorporate whole-body models, although the completeness of the model varies considerably. Sometimes, little detail is available to the public concerning these models. A description of some of these models can be found in the section Uses of Models.

Even with the computational resources now available, most PK and many PKPD models contain only two to four compartments. Compare BODY Simulation's 37 compartments. Whole-body models are complex and difficult. Pharmacologists have seriously asked, Why do we need all those compartments? Can't I achieve the same results from an uncluttered three-compartment model? Aside from losing the physiological essence of any simulation that is run, one also loses nonlinearity, the impact of physiology on drug kinetics, the important feature of drug interactions, the influence of patient condition, such as aging, and the effect of stresses, such as hemorrhage or apnea. As just one example, any agent that decreases hepatic blood flow will affect the concentration, and therefore the action, of any agent that is metabolized by the liver.

REGIONAL CIRCULATION AND AUTOREGULATION

Regional Circulation

Only whole-body models can have regional circulation, of course. In fact, regional circulation is crucial in this type of model, to connect the various physiological modules. Many factors control regional circulation, as well as the relation among the circulations, but only some of them are discussed, briefly, primarily because few models take these factors into account. The BODY Simulation model (see PBPM) is one of the few exceptions.

One of the interesting factors controlling regional circulation is the importance to the body of a given organ or tissue. How well is an organ's circulation maintained in the face of an acute stress, rapid hemorrhage, for example? Organs and tissues can be divided, simplistically, into four types/classes.

1. Essential in the short term (a few minutes) (brain and heart).
2. Essential in the medium term (a few hours) (hepatic, renal and splanchnic).
3. Essential in the longer term (several hours) (skeletal muscle).
4. Essentially nonessential (skin, cartilage, bone).

The overall regulation of regional circulation reflects this hierarchy. In addition, some agents can influence it. Few models, except for Guyton's model (see Whole-body Models) and some the whole-body models described above, incorporate detailed regional circulation. BODY Simulation (see PBPM) and one of Ursino's many models (47) take this hierarchy into account. In BODY Simulation (see PBPM), regional circulation is impacted by several factors, including O_2, CO_2, the baroceptors, strength of the heart, blood volume, hemorrhage, extracellular fluid, pharmacological agents, the presence of circulating epinephrine (pain or hypercapnia), and so on.

The regional-circulation hierarchy impacts the various regional circulation control mechanisms, and autoregulation is often subservient to this pecking order (except for cerebral). If severe shock occurs, the hypoxic muscle will not get its allotted supply for example. The cerebral and coronary circulations come first. In some ways, this is a form of regional steal, a term often used in physiology, but not in this context.

General Autoregulation

When sudden alterations in perfusion pressure are imposed in most types of arterial beds, the resulting abrupt changes in blood flow are only transitory, with flow returning quickly to the previous steady-state level. The exception is, of course, a sudden arterial occlusion. The ability to maintain perfusion at constant levels in the face of changing driving pressure is termed autoregulation. Autoregulation only occurs between certain pressure limits (if the pressure drops too low or soars too high, autoregulation fails, and organ perfusion is compromised) at low pressures, perfusion decreases, and at high pressures, excessive flow occurs. Autoregulation keeps tissue or organ flow essentially constant between an MAP of 60–180 mmHg (7.9–23.9 kPa), the limits varying with the tissue or organ.

Much of autoregulation may occur in the microcirculation. Groebe (48) outlined gaps in the understanding of how the microvasculature maintains tissue homeostasis, as of the mid-1990s: (1) integration of the potentially conflicting needs for capillary perfusion and hydrostatic pressure regulation, (2) an understanding of signal transmission pathways for conveying information about tissue energetic status from undersupplied tissue sites to the arterioles, (3) accounting for the interrelations between precapillary and postcapillary resistances, and (4) an explanation of how the body achieves local adjustment of perfusion to metabolic demands. Using mathematical modeling, Groebe argued that precapillary pressure regulation combined with postcapillary adjustment of perfusion to tissue metabolic status helped clarify the understanding of microvascular control.

Mechanisms of autoregulation vary substantially between organs. Coronary autoregulation is, for example, quite different from brain autoregulation. During hemorrhagic shock in pigs, microcirculatory blood flows in the stomach, colon, liver, and kidney decrease in concert with decreasing systemic blood flow; flow in the jejunal mucosa is preserved, and pancreatic blood flow is selectively impaired 180-(49). Differential autoregulation has also been shown in response to increases in blood pressure. For example, in response to infusion of pressor agents, renal autoregulation is almost unimpaired, while regulation in the mesenteric vascular bed is less adequate, and differs with different pressor agents. Muscle autoregulation is mediated partly by the metabolic byproducts of exercise. The main site of autoregulation in the kidney is, however, the afferent glomerular arteriole. There are two main factors that affect vascular tone in the afferent arteriole: stretch-activated constriction of vessels (as for the brain) and tubulo-glomerular feedback. The autoregulatory response and the factors causing it can also vary in the same organ. For example, in exercising dogs, increased O_2 demand of the LV is met primarily by increasing coronary flow, while increased O_2 extraction makes a greater contribution to RV O_2 supply (50).

Needless to say, autoregulation is complex, and difficult to model. The following section gives some examples, in different circulatory beds.

Cerebral Autoregulation

Because the brain resides in a rigid box, very small changes in CBF can lead to catastrophic changes in intercranial pressure (ICP). Rapid CBF autoregulation is therefore vital, and cerebral flow normally remains constant within MAP ranges of 50–150 mmHg (6.6–19.9 kPa). Any sudden changes in MAP are transmitted to the cerebral circulation, inducing similar changes in CBF, but fortunately, under normal conditions the CBF tends to return to its original value within a few seconds. The main regulator of brain blood flow is pressure-dependent activation of smooth muscle in the arterioles of the brain. The more the arteriole is stretched, the more it contracts, and this lasts as long as the stretch occurs.

Paneraiy (51) critically reviewed the concepts of cerebral autoregulation, including the role of mathematical models. The most common approach to evaluating cerebral autoregulation tests the effects of changes in MAP on CBF, and is known as pressure autoregulation. A gold standard for this purpose is not available and the literature shows considerable disparity of methods and criteria. This is understandable because cerebral autoregulation is more a concept than a physically measurable entity. Static methods use steady-state values to test for changes in CBF (or velocity) when MAP is changed significantly. This is usually achieved with the use of drugs, or from shifts in blood volume, or by observing spontaneous changes. The long time interval between measurements is a particular concern. Concomitant changes in other critical variables, such as PCO_2, hematocrit, brain activation, and sympathetic tone, are rarely controlled for. Proposed indies of static autoregulation are based on changes in cerebrovascular resistance, on parameters of the linear regression of flow–velocity versus pressure changes, or only on the absolute changes in flow. Methods of dynamic assessment are based on transient changes in CBF (or velocity) induced by the deflation of thigh cuffs, Valsalva maneuvers (a bearing down, that induces an increase in airway pressure with resultant complex, but easily understood, hemodynamic changes), tilting, and induced or spontaneous oscillations in MAP. Classification of autoregulation performance using dynamic methods has been based on mathematical modeling, coherent averaging, transfer function analysis, cross-correlation function, or impulse response analysis.

Cerebral autoregulation has been modeled independently and in concert with other factors. The most nearly complete model of the combination of factors is by Lu et al. (52). The goal of their work was to study cerebral autoregulation, brain gas exchange, and their interaction. Their large model comprised a model of the human cardiopulmonary system, which included a whole-body circulatory system, lung, and peripheral tissue gas exchange, and the CNS control of arterial pressure and ventilation, and central chemoreceptor control of ventilation, as well as a detailed description of cerebral circulation, CSF dynamics, brain gas exchange, and CBF autoregulation. Two CBF regulatory mechanisms were included: autoregulation and CO_2 reactivity.

Three groups have mathematically examined cerebral autoregulation by itself (53–56). This allows insight into the process, as well as the ability to incorporate the concepts into one's own model. In addition, Czosnyka et al. (57) constructed a model that helps the clinician interpret bedside tests of cerebrovascular autoregulation.

A simple model (authors' term) was described by Ursino and Lodi (58). The model includes the hemodynamics of the arterial–arteriolar cerebrovascular bed, CSF production and reabsorption processes, the nonlinear pressure–volume relationship of the craniospinal compartment, and a Starling resistor mechanism for the cerebral veins. Moreover, arterioles are controlled by cerebral autoregulation mechanisms, which are simulated by a time constant and a sigmoidal static characteristic. The model is used to simulate interactions between ICP, cerebral blood volume, and autoregulation.

Hyder et al. (59) have developed a model that describes the autoregulation of cerebral O_2 delivery *In vivo*. According to the model, the O_2 diffusivity properties of the capillary bed,which can be modified in relation to perfusion, play an important role in regulating cerebral O_2 delivery *In vivo*. Diffusivity of the capillary bed, in turn, may be altered by changes in capillary PO_2, hematocrit, and/or blood volume.

Kirkham et al. (60) presented a mathematical model representing dynamic cerebral autoregulation as a flow-dependent feedback mechanism. They introduced two modeling parameters: the rate of restoration, and a time delay. Velocity profiles were determined for a general MAP, allowing the model to be applied to any experiment that uses changes in MAP to assess dynamic cerebral autoregulation. The comparisons yielded similar estimates for the rate of restoration and the time delay, suggesting

that these parameters are independent of the pressure change stimulus and depend only on the main features of the dynamic cerebral autoregulation process. The modeling also indicated that a small phase difference between pressure and velocity waveforms does not necessarily imply impaired autoregulation. In addition, the ratio between the variation in maximum velocity and pressure variation can be used, along with the phase difference, to characterize the nature of the autoregulatory response.

Coronary Autoregulation

The range of coronary autoregulation is 60–130 mmHg (7.9–17.3 kPa). Demonstration of autoregulation in the coronary bed is difficult in intact animals because modification of coronary perfusion pressure also changes myocardial oxygen demand and the extrinsic compression of the coronary vessels. However, when perfusion pressure is altered, but ventricular pressure, cardiac contractility, and heart rate (the principal determinants of myocardial oxygen demand) are maintained constant, autoregulation is clearly evident.

The coronary circulation has a different set of problems from other systems, as well as many paradoxes. These problems include the following: (1) It is caught in a unique situation: it must feed and cleanse the very organ that generates it. (2) The heart is a high oxidative organ with a high demand for O_2 and a very high O_2 consumption. (3) The av O_2 difference is much wider in the coronary circulation, implying that less reserve is available. (4) As myocardial function increases (increased HR or contractility, e.g.), the demand for O_2 increases dramatically. (5) Intramyocardial flow actually decreases as contractility increases, however, because of compression and shearing effects; in other words, the more the heart works, the more it impedes the flow that it needs. (6) Because of the impairment of myocardial blood flow during systole, nearly 80% of coronary flow occurs during diastole. (7) Although most arterial flow occurs during diastole, most venous flow occurs during systole, both phenomena because of the compression effects. (8) Even though increased HR places an increased energy demand on the myocardium, there is less coronary flow as HR increases, since the valuable diastolic time decreases out of proportion to systolic time during tachycardia.

As in any vascular bed, blood flow in the coronary bed depends on the driving pressure and the resistance offered by this bed. Coronary vascular resistance, in turn, is regulated by several control mechanisms: myocardial metabolism (metabolic control), endothelial (and other humoral) control, autoregulation, myogenic control, extravascular compressive forces, and neural control. Each control mechanism may be impaired in a variety of conditions and each can contribute to the development of myocardial ischemia.

Control of coronary blood flow also differs depending on the type of vessel being considered: arteries, large arterioles, or smaller arterioles. Coronary capillaries also appear to make a significant contribution to coronary vascular resistance!

Several factors play a role in coronary autoregulation itself, including myogenic responses, resistance distribution in various size vessels, O_2 consumption, capillaries, flow-dependent dilation, and direct metabolic control. Regarding the last, for a substance to be considered a mediator of local metabolic control of coronary flow it should (1) be vasoactive, (2) found, in appropriate concentrations, to affect vascular tone, and (3) be of variable concentration in response to changes in metabolism. Myocardial O_2, along with CO_2 and other waste products are likely mediators of local metabolic control. Several models have examined one or more of these factors, although we could find no model that put everything together (61–68). One can, however, begin to separate out the contribution of each factor. For example, the synergistic interaction between PO_2 and PCO_2 accounts for about one-fourth of the change in coronary vascular conductance during autoregulation (62).

Cardiac Autoregulation

Cardiac autoregulation is briefly mentioned, although it is not circulatory autoregulation. The topic was one of the first to be studied, namely, the Frank–Starling mechanism, the dependence of individual stroke volume on end-diastolic volume. This mechanism helps ensure that what comes into the heart goes out, so that the heart does not blow up like a balloon. The functional importance of the Frank–Starling mechanism lies mainly in adapting left to right ventricular output. Just as with vascular autoregulation, many other factors can override the Frank–Starling mechanism, but it can become particularly important in CV disease.

In the past, the study of mechanical and electrical properties of the heart has been disjointed, with minimal overlap and unification. These features, however, are tightly coupled and central to the functioning heart. The maintenance of adequate cardiac output relies upon the highly integrated autoregulatory mechanisms and modulation of cardiac myocyte function. Regional ventricular mechanics and energetics are dependent on muscle fiber stress–strain rate, the passive properties of myocardial collagen matrix, adequate vascular perfusion, transcapillary transport and electrical activation pattern. Intramural hydraulic “loading" is regulated by coronary arterial and venous dynamics. All of these components are under the constant influence of intrinsic cardiac and extracardiac autonomic neurons, as well as circulating hormones.

MODELS OF SYSTEMS

Cardiovascular–Circulation

The CVS is essential to the body as the means for carrying substances, such as O_2, nutrients, and hormones to the tissues where they are needed and for bringing xenobiotics or waste products, such as CO_2 and acids, to the lungs, kidneys, or liver to be eliminated.

A pioneer in CV models was Bram Noordergraaf and his group, starting in the 1950s. His model began with the systemic circulation and later expanded to the pulmonary circulation. Over the period of 15 years, they evolved into

models of great size, complexity, and sophistication. The model has a heart, and the systemic circulation has >115 segments. It included adjustable peripheral resistances, plus viscoelasticity, stiffness, radius and wall thickness of each segment, viscous properties of the blood, tapering in radius and elasticity, and frictional losses. Abnormalities, such as aortic stenosis, aortic insufficiency, idiopathic hypertrophic subaortic stenosis, and atrial septal defect have been simulated. We give two later, easily accessible references (69,70).

Another pioneer, in the Dutch school of modeling, was Jan Beneken. His models have also been sophisticated and rigorous (71,72). His models and the impressive ones of Karel Wesseling will be discussed later.

Two CV models are available for those who want to run one. The first is an interactive tutorial and mathematical model described by Rothe and Gersting (73). In addition, one can play with the 52-compartment CV model of Wesseling, as implemented by Starko, Haddock, and Smith (unpublished data). The inclusion of the two atria has made this model an especially unique and useful tool. Eventually, this model will be incorporated into BODY Simulation (see PBPM). Currently, however, it has no transport model(s), nor any control system, for example, an autonomic nervous system or baroceptors. In essence, it is a heart–lung preparation, without the lungs, but with pulmonary circulation. By itself, however, it is very useful in the study of the basics of the heart and circulation, allowing one to examine in great detail the subtle and not so subtle interactions of the CVS. This is made possible by the detail involved, as well as the ability to change any of hundreds of parameters and to study the time plots of any of hundreds of variables. The model allows the user to adjust the contractility and stiffness of the four chambers, as well as the pressure, volume, compliance and, when appropriate, resistance of each of the 52 compartments. In addition, heart rate, SVR, and total blood volume can be adjusted. Some of these changes involve changes in preload and afterload. Pressure, volume, and flow waveforms can be displayed for each compartment, and pressure–volume plots are available for the four chambers. Thirteen values, such as HR, MAP, and end-diastolic ventricular volumes, are numerically displayed. A small manual describes the model and how to use it.

Most cardiac models incorporate only a freestanding chamber, or chambers. The two ventricles, however, affect each other's dynamics. In addition, the pericardium, the stiff membrane that surrounds almost the entire heart, affects each chamber. The model of Chung et al. (74) describes the dynamic interaction between the LV and the RV over the complete cardiac cycle. The pericardium-bound ventricles are represented as two coupled chambers consisting of the left and right free walls and the interventricular septum. Timevarying pressure–volume relationships characterize the component compliances, and the interaction of these components produces the globally observed ventricular pump properties (total chamber pressure and volume). The model (1) permits the simulation of passive (diastolic) and active (systolic) ventricular interaction, (2) provides temporal profiles of hemodynamic variables (e.g., ventricular pressures, volumes and flows), and (3) can be used to examine the effect of the pericardium on ventricular interaction and ventricular mechanics. The model also yields qualitative predictions of septal and free wall displacements.

Similarly, few models have addressed the mechanical interaction between the CV and pulmonary systems, for example, how the combined cardiopulmonary system responds to large amplitude forcing (change in an important variable). To address this issue, Lu et al. (75), developed a human cardiopulmonary system model that incorporates important components of the cardiopulmonary system and their coupled interaction. Included in the model are descriptions of atrial and ventricular mechanics; hemodynamics of the systemic and pulmonary circulations; baroreflex control of arterial pressure; airway and lung mechanics; and gas transport at the alveolar-capillary membrane. They applied this model to the analysis of the Valsalva maneuver. The model could predict the hemodynamic responses to markedly increased intrathoracic (pleural) pressures during a Valsalva maneuver. In short, this model can help explain how the heart, lung, and autonomic tone interact during the Valsalva maneuver.

BODY Simulation (see PBPM) also incorporates an interaction between intrapleural pressure (as reflected by airway pressure) and venous return. Problems with a patient ventilator, for example, can create disastrous depression of the CVS (76). Combining computational blood flow modeling with 3D medical imaging provides a new approach for studying links between hemodynamic factors and arterial disease. Although this provides patient-specific hemodynamic information, it is subject to several potential errors. A different approach, developed by Moore et al. (77) can quantify some of these errors and identify optimal reconstruction methodologies.

For several reasons, modeling the pulmonary circulation presents challenges. Huang et al. (78) described a mathematical analogue-circuit model of pulsatile flow in cat lung based on existing morphometric and elastic data. In the model, the pulmonary arteries and veins were treated as elastic tubes, whereas the pulmonary capillaries were treated as two-dimensional (2D) sheets. Input impedances of the pulmonary blood vessels of every order were calculated under normal physiological conditions. The pressure-flow relation of the whole lung was predicted theoretically. Comparison of the theoretically predicted input impedance spectra with their experimental results showed that the modulus spectra were well predicted, but significant differences existed in the phase angle spectra between the theoretical predictions and the experimental results. The authors state that the current model cannot explain this latter discrepancy.

Occlusion experiments yield time–pressure and time–flow curves that are related to the longitudinal distribution of compliances and resistances in the pulmonary circulation. The standard approach to the analysis of these curves involves the observation of relevant features of their graphs, which may directly reflect model parameter values. De Gaetano and Cremona (79) considered five possible models of pulmonary vascular pressure dynamics and the relative (nonlinear) least-squares parameter estimation from experimental data, making simultaneous use

of all available information. The five models included two linear models without inductance units, one linear model with inductance units, one nonlinear model with variable resistance, and one nonlinear model with variable compliance. In all cases, parameter estimation for the numerically integrated model was performed by unweighted least squares, using a variable-metric minimization technique.

Potentially, the more detailed the model, the more accurate it is. Karamanoglu et al. (80) simulated the effects of wave travel and wave reflection with a mathematical model of the whole arterial tree, which comprised 142 uniform transmission-line segments. The arterial model was partitioned into three separate segments: upper limbs, trunk, and lower limbs. Aging was simulated by increasing average pulse wave velocities of these segments. Reflection coefficients at the terminal elements were altered to simulate vasodilation and vasoconstriction.

Karamanoglu and Feneley (81) also used a linear mathematical model of the entire human arterial tree to derive realistic impedance patterns by altering (1) Young's modulus of the arterial wall of the individual branches, (2) peripheral reflection coefficients, and (3) distal compliances at the terminations. These calculated impedance patterns were then coupled to realistic LV outflow patterns determined by unique (1) end-diastolic and endsystolic pressure–volume relationships, (2) preload-recruitable stroke work relationships, and (3) shortening paths simulated by altered aortic flow contours. Left ventricular outflow patterns were as important as impedance parameters in determining late systolic pressure augmentation, at least in this model.

Cardiac valvular modeling and simulation are important, especially given the common use of echocardiography. Sun et al. (82) examined this area. The transmitral and pulmonary venous flow velocity patterns were related to the physiological state of the left heart with an electrical analogue model. Filling of the LV through the mitral valve was characterized by a quadratic Bernoulli's resistance in series with an inertance. Filling of the LA through the pulmonary veins was represented by a lumped network of linear resistance, capacitance, and inertance. The LV and LA were each represented by a time-varying elastance. A volume dependency was incorporated into the LV model to produce physiological PV loops and Starling curves. The model accurately reflected the expected effects of aging and decreasing LV compliance, and could serve as a useful theoretical basis for echocardiographic evaluation of LV and LA function.

Yellin et al. (83) examined the mechanisms of mitral valve motion in mid-diastole, diastole and at closure by simultaneously measuring mitral flow (electromagnetic), valve motion (echo), and AV pressures. Large variations in peak flow were accompanied by small variations in valve excursion. They concluded that the valve overshoots its equilibrium position and that the chordae produce tension on the valve during diastole. Their model offered a valve-closure theory unifying chordal tension, flow deceleration, and vortices, with chordal tension as a necessary condition for the proper functioning of the other two.

The Doppler transmitral velocity curve is commonly used to assess LV diastolic function. Thomas et al. (84) developed a mathematical formulation to study the physical and physiological determinants of the transmitral velocity pattern for exponential chamber PV relationships with active ventricular relaxation (2187 combinations investigated). They showed that transmitral velocity is fundamentally affected by three principal physical determinants: the transmitral pressure difference, the net AV compliance, and the impedance characteristics of the mitral valve. These physical determinants in turn are specified by certain compliance and relaxation parameters. They found that the peak mitral velocity is most strongly related to initial LA pressure but decreased by prolonged relaxation, low atria1 and ventricular compliance, and systolic dysfunction. Peak acceleration varies directly with atria1 pressure and inversely with the time constant of isovolumic relaxation, with little influence of compliance, whereas the mitral deceleration rate is approximately valve area divided by AV compliance.

A moderate reduction in coronary blood flow results in decreased myocardial O_2 consumption, accelerated glycolysis, decreased pyruvate oxidation, and lactate accumulation. To quantitatively understand cardiac metabolism during ischemia, Salem et al. (85) demonstrated a mechanistic, mathematical model based on biochemical mass balances and reaction kinetics in cardiac cells. Computer simulations showed the dynamic responses in glucose, fatty acid, glucose-6-phosphate, glycogen, triglyceride, pyruvate, lactate, acetyl-CoA, and free-CoA, as well as CO_2, O_2, phosphocreatine/creatine, nicotinamide adenine dinucleotide (reduced form)/nicotinamide adenine dinucleotide (oxidized form) (NADH/NAD1), and adenosine diphosphate/adenosine triphosphate (ADP/ATP). When myocardial ischemia was simulated by a 60% reduction in coronary blood flow, the model generated myocardial concentrations, uptakes, and fluxes that were consistent with experimental data from *in vivo* pig studies. With the onset of ischemia, myocardial lactate concentration increased and the myocardium switched from a net consumer to a net producer of lactate.

Olufsen et al. (86) modeled blood flow in large systemic arteries by using one-dimensional (1D) equations derived from the axisymmetric Navier–Stokes equations for flow in compliant and tapering vessels. The arterial tree is truncated after the first few generations of large arteries with the remaining small arteries and arterioles providing outflow boundary conditions for the large arteries. By modeling the small arteries and arterioles as a structured tree, a semianalytical approach based on a linearized version of the governing equations can be used to derive an expression for the root impedance of the structured tree in the frequency domain. In the time domain, this provides the proper outflow boundary condition. The structured tree is a binary asymmetric tree in which the radii of the daughter vessels are scaled linearly with the radius of the parent vessel. Blood flow and pressure in the large vessels are computed as functions of time and axial distance within each of the arteries.

The CVS is an internal flow loop with multiple branches circulating a complex liquid. The hallmarks of blood flow in arteries are pulsatility and branches, which cause wall stresses to be cyclical and nonuniform. Normal arterial

flow is laminar, with secondary flows generated at curves and branches. Arteries can adapt to and modify hemodynamic conditions, and unusual hemodynamic conditions may cause an abnormal biological response. Velocity profile skewing can create pockets in which the wall shear stress is low and oscillates in one direction. Atherosclerosis tends to localize to these sites and creates a narrowing of the artery lumen: a stenosis. Plaque rupture or endothelial injury can stimulate thrombosis, which can block blood flow to heart or brain tissues. The small lumen and elevated shear rate in a stenosis create conditions that accelerate platelet accumulation and occlusion. The relationship between thrombosis and fluid mechanics is complex, especially in the poststenotic flow field. New convection models have been developed to predict clinical occlusion from platelet thrombosis in diseased arteries (87).

Cardiovascular Regulation

Vasquez et al. (88) presented a lucid review of the overall coordination of three of the major mechanisms involved in CVS control: the baroreceptors, chemoreceptors, and cardiopulmonary reflexes. The central chemoreceptors are the main body CO_2 (actually, pH acts as a surrogate for CO_2) sensors, the main sensors in the feedback loop. They are located in the medulla, near the ventricle and bathed in brain extracellular fluid, which is close to the composition of CSF. The normal pH of the fluid is 7.32 and is poorly buffered. Thus, CSF pH is regulated more rapidly than in the rest of the body. Although CO_2 diffuses rapidly into all tissues, H^+ does not penetrate into CSF. However, if PCO_2 increases in the blood, it diffuses into CSF and lowers the pH. Ventilation responds to changes in CSF pH.

Peripheral chemoreceptors are located in the carotid and aortic bodies and in clumps along the route of the abdominal vagus. Their role in the regulation of the CVS cannot be understood without knowing the various factors that can change chemoreceptor afferent activity. Furthermore, changes in chemoreceptor activity not only exert primary reflex effects on the CVS, but they evoke changes in the central drive for ventilation that secondarily affect the CVS. For an excellent, comprehensive review of this subject, see Marshall (89).

The arterial baroreflex contributes significantly to the short-term regulation of blood pressure and CV variability. Several factors, including reflex, humoral, behavioral, environmental and age, may influence gain and effectiveness of the baroreflex, as well as CV variability. Many central neural structures are also involved in the regulation of the CVS and contribute to the integrity of the baroreflex. For those who wish a good summary of the subject, read the review by Lanfranchi and Somers (90).

Continuous blood pressure recordings usually show a surprising variability in MAP. Short-term variability includes components with periods of a few seconds to many hours, and can be spontaneous or in response to a maneuver or activity. Blood pressure can increase 50 mmHg (6.6 kPa) during painful stimuli in a matter of a minute, or decrease 20–30% in a few seconds during an orthostatic maneuver. Often, blood pressure shows oscillatory fluctuations in an $\sim$ 10 s rhythm. All this variability has been demonstrated in normal subjects, especially by Wesseling and co-workers (91–93). This short-term blood pressure variability challenges the concept of an effective, stabilizing baroreflex. Wesseling has called this phenomenon the baroreflex paradox and has proposed a baromodulation hypothesis (91–93). The baromodulation hypothesis states that the baroreflex gain can be modulated to have high gain in some situations, low gain in others. A decrease in baroreflex gain by itself causes blood pressure to increase, while an increase in gain causes it to fall (91). A gain change does not change baroceptor function itself. No changing baroceptor sensitivity is postulated, and no baroceptor resetting need occur. Theoretically, the physiological location where modulation occurs could be anywhere in the reflex loop, even in the individual efferent-pathways, separately. However, the logical site would be the vasomotor center in the medulla.

We should point out that this insight was made possible by using a noninvasive monitoring device that is enhanced by several models (see Finapres, under uses for Models), and that yet another model was used to test the hypothesis. Several investigators have based their work on this remarkable insight (references on request).

When blood volume is expanded, CVP increases, stimulating cardiopulmonary receptors in the atria and ventricles and perhaps arterial baroreceptors in the aortic arch and carotid sinus. Volume expansion or stimulation of atrial receptors can inhibit vasopressin release, decrease sympathetic nerve activity, and attenuate drinking. Atrial distension stimulates secretion of atrial natriuretic peptide (ANP) from the atria leading to natriuresis. Thus it is possible that ANP may inhibit vasopressin release, reduce blood pressure, decrease drinking, and lead to natriuresis and diuresis via central pathways.

Rose and Schwaber (94) described a model that included only HR in the baroceptor modulation of arterial pressure, since the vagus HR response is the most rapid responder to changes in arterial pressure. They observed that vagal induced changes in HR do not influence arterial pressure, except when certain initial conditions of the CVS are met. It may be that vagally mediated alterations in an inotropic and dromotropic state, which are not included in this model, play important roles in the fast reflex control of blood pressure or that the vagal limb of the baroreflex is of rather limited effectiveness. Had the authors decreased heart rate >50%, they may have observed profound CV changes arising from a heart rate change.

Ursino and Magosso (95) detailed a mathematical model of the acute CV response to isocapnic hypoxia. The model includes a pulsating heart, the systemic and pulmonary circulation, a separate description of the vascular bed in organs with higher metabolic need, and the local effect of O_2 on these organs. The model also includes the action of several reflex regulatory mechanisms: the peripheral chemoreceptors, the lung stretch receptors, the arterial baroreceptors, and the hypoxic response of the CNS. The early phase of the biphasic response (8–10 s), caused by activation of peripheral chemoreceptors, exhibits a moderate increase in MAP, a decrease in HR, a relatively constant CO, and a redistribution of blood flow to the organs with higher metabolic need, at the expense of other organs (see

the hierarchy scheme, in Regional Circulation and Autoregulation). The later phase (20 s) is characterized by the activation of lung stretch receptors and by the CNS hypoxic response. During this phase, CO_2 and HR increase, and blood flow is restored to normal levels, in organs with lower metabolic need.

These authors performed an extensive validation of this model (96). The role of the different mechanisms involved in the CV response to hypoxia (chemoreceptors, baroreceptors, lung stretch receptors, and CNS hypoxic response) was analyzed in different physiological conditions. The simulation results revealed the following: (1) the model can reproduce the CV response to hypoxia very well between 100 and 28 mmHg (13.3 and 3.7 kPa) PO_2. (2) Sensitivity analysis of the impact of each individual mechanism underlines the role of the baroreflex in avoiding excessive derangement of systemic arterial pressure and CO_2 during severe hypoxia and suggests the existence of significant redundancy among the other regulatory factors. (3) With chronic sinoaortic denervation (i.e., simultaneous exclusion of baroreceptors, chemoreceptors, and lung stretch receptors), the CNS hypoxic response alone is able to maintain reasonably normal CV adjustments to hypoxia, although suppression of the CNS hypoxic response, as might occur during anesthesia, led to a significant arterial hypotension. (4) With controlled ventilation, a significant decrease in HR that can only partly be ascribed to inactivation of lung stretch receptors. (5) When maintaining a constant CO_2 during severe hypoxia, the chemoreflex can produce a significant decrease in systemic blood volume.

As an extension of the group's isocapnic hypoxia model, Magosso and Ursino (97) studied the effect of CO_2 on the CVS. The previous model (95) had already incorporated the main reflex and local mechanisms triggered by O_2 changes. The new features covered by the model were the O_2–CO_2 interaction with the peripheral chemoreceptors, the effect of local CO_2 changes on peripheral resistances, the direct CNS response to CO_2, and the control of central chemoreceptors on minute ventilation and tidal volume. The model could simulate the acute CV response to changes in blood gas content in a variety of conditions (normoxic hypercapnia, hypercapnia during artificial ventilation, hypocapnic hypoxia, and hypercapnic hypoxia). The model ascribes the observed responses to the complex superimposition of many mechanisms simultaneously working (baroreflex, peripheral chemoreflex, CNS response, lung-stretch receptors, local gas tension effect), which may be variably activated depending, on the specific stimulus under study. However, although some experiments can be reproduced using a single basal set of parameters, reproduction of other experiments requires a different combination of the mechanism strengths (particularly, a different strength of the local CO_2 mechanism on peripheral resistances and of the CNS response to CO_2).

Melchior et al. (98) used as an example the short-term response of the human CVS to orthostatic stresses to develop a mathematical model. They reviewed the physiological issues involved and how these issues have been handled in previous CV models for simulation of the orthostatic response. Most extant models were stimulus specific with no apparent ability to simulate the responses to orthostatic stimuli of different types. They suggest that a comprehensive model incorporating all known phenomena related to CV regulation is needed. The paper represents a good start in providing a framework for future efforts in mathematical modeling of the entire CVS, and the review of issues is outstanding.

Lerma et al. (99) combined parts of two systems: the baroceptor reflex in the CVS and the renal system in chronic renal failure (CRF). They developed a model of baroreflex control of MAP, in terms of a delay differential equation, and used it to predict the adaptation of short-term CV control in CRF patients. The model predicts stable and unstable equilibria close to steady-state MAP. Their results suggest that the cardiac pump has a more restricted response in CRF patients. The model quantifies the CV adaptations to CRF, including increased SVR and baroreflex delay, as well as decreased arterial compliance, cardiac period, and stroke volume.

In yet another paper, Ursino and Magosso (100) examined the response to O_2 and CO_2 changes mediated by one CV regulator mechanism, the carotid body chemoreceptor. The model assumes that the static chemoreceptor characteristic depends on O_2 saturation in the arterial blood and on CO_2 arterial concentration. The values of O_2 saturation and of CO_2 concentration are computed, from pressure, using blood dissociation curves, which include both the Bohr and Haldane effects. The dynamic response includes a term depending on the time derivative of CO_2 concentration and a low pass filter, which accounts for the time required to reach the steady-state level. With a suitable choice of parameters, the model reproduced the carotid chemoreceptor response under a variety of combined O_2 and CO_2 stimuli, both in steady-state conditions and in the transient period following acute CO_2 or O_2 pressure changes. During transient conditions, the effect of CO_2 pressure changes prevail over the effect of O_2 changes, due to the intrinsic derivative component of the response to CO_2.

Ursino et al. (101) explored one of the most important regulator effectors of the CVS: venous capacitance. To elucidate the role of venous capacity active changes in short-term CV homeostasis, they developed a mathematical model of the carotid-sinus baroreflex system. In the model, the CVS was represented as a series arrangement of six lumped compartments, which synthesized the fundamental hemodynamic properties of the systemic arterial, systemic venous, pulmonary arterial, and pulmonary venous circulations as well as of the left and right cardiac volumes. Cardiac outputs from the left and right ventricles were computed as a function of both downstream (afterload) and upstream atrial pressure (preload). Four distinct feedback regulatory mechanisms, working on SVR, HR, systemic venous unstressed volume, and systemic venous compliance, were assumed to operate on the CVS in response to carotid sinus pressure changes. The model was used to simulate the pattern of the main hemodynamic quantities in the short time period (1–2 min) after acute carotid sinus activation in vagotomized subjects. Simulations indicated that the model can reproduce experimental data quite well, with reference both to open-loop

experiments and to acute hemorrhage performed in closed-loop conditions. Computer simulations also indicated that active changes in venous unstressed volume are very important in regulating CO_2 and MAP during activation of the carotid sinus baroreflex.

Modeling HR and blood pressure spontaneous variability can contribute to the understanding of both normal and pathologic CVS physiology. The observed fluctuations in HR and blood pressure are meaningful rhythmical fluctuations that reflect useful information about autonomic regulation. These rhythmical fluctuations, known as heart rate variability and blood pressure variation, are normally grouped into three major components: (1) the high frequency component, ~0.25 Hz, in synchrony with respiratory rate; (2) the low frequency component, generally centered ~0.1 Hz, which is attributed to the sympathetic activity and the closed-loop controlling action of cardiovascular regulation; and (3) The very low frequency component, ~0.04 Hz, which is probably due to the vasorhythmicity thermoregulatory system or to humoral regulations. Cohen and Taylor (102) nicely reviewed the subject and constructed their own model. Seydnejad and Kitney (103), as well as Cavalcanti and Belardinelli (104) have also modeled these areas. Of particular interest are the studies of Magosso et al. (105). Their findings include the following: (1) A significant increase in the gains and time delays (>9 s) of all the arterial baroreflex sympathetic mechanisms is required to induce instability (see Baromodulation and Wesseling, below). In this condition, systemic arterial pressure exhibits spontaneous oscillations with a period of ~20 s, similar to Mayer waves. The control of peripheral resistance seems more important than the venous volume control in the genesis of these oscillations. (2) An increase in the gain and time delay (~3 s) of the arterial baroreflex vagal mechanism causes the appearance of unpredictable fluctuations in heart period, with spectral components in the range 0.08–0.12 Hz 3) The cardiopulmonary baroreflex plays a less important role than does the arterial baroreflex in the genesis of these instability phenomena.

Aljuri and Cohen (106) took yet another approach to the problem. They emphasized the analytic algebraic analysis of the systemic circulation composed of arteries, veins, and its underlying physiological regulatory mechanisms of baroreflex and autoregulatory modulation of SVR, where the behavior of the system can be analytically synthesized from an understanding of its minimal elements. As a result of their analysis, they presented a mathematical method to determine short-term SVR fluctuations, which account for observed MAP fluctuations, and proposed a new CVS identification method to delineate the actions of the physiological mechanisms responsible for the dynamic couplings between CO_2, MAP, RA pressure, and SVR.

Ben-Haim et al. (107), using a finite-difference equation to model cardiac mechanics, simulated the stable action of the LV. Their model described the LV end-diastolic volume as a function of the previous end-diastolic volume and several physiological parameters describing the mechanical properties and hemodynamic loading conditions of the heart. Their simulations demonstrated that transitions (bifurcations) can occur between different modes of dynamic organization of the isolated working heart as parameters are changed. Different regions in the parameter space are characterized by different stable limit cycle periodicities. They proposed that mechanical periodicities of the heart action are an inherent part of its nonlinear nature. Although their model predictions and experimental results were compatible with previous experimental data, they may contradict several hypotheses suggested to explain the phenomenon of cardiac periodicities.

Ursino (108) made some interesting observations with his model on short-term carotid baroregulation and the pulsating heart. The model includes an elastance variable description of the left and right heart, the systemic (splanchnic and extrasplanchnic) and pulmonary circulations, the afferent carotid baroreceptor pathway, the sympathetic and vagal efferent activities, and the action of several effector mechanisms. The latter mechanisms work, in response to sympathetic and vagal action, by modifying systemic peripheral resistances, systemic venous unstressed volumes, heart period, and endsystolic elastances. The model is used to simulate the interaction among the carotid baroreflex, the pulsating heart, and the effector responses. Experimental data on HR control can be explained fairly well by assuming that the sympathetic-parasympathetic systems interact linearly on the heart period. The carotid baroreflex can significantly modulate the cardiac function curve. This effect, however, is masked *In vivo* by changes in arterial and atrial pressures. During heart pacing, CO_2 increases with frequency at moderate levels of heart rate and then fails to increase further because of a reduction in stroke volume. Shifting from nonpulsatile to pulsatile perfusion of the carotid sinuses decreases the overall baroreflex gain and significantly modifies operation of the carotid baroreflex. Finally, sensitivity analysis suggests that venous unstressed volume control plays the major role in the early hemodynamic response to acute hemorrhage, whereas systemic resistance and heart rate controls are slightly less important.

Short-term regulation of arterial blood pressure is accomplished by complex interactions between feedback and feed-forward information from the arterial and cardiopulmonary baroreceptors that combine with other local and neural factors to modulate CO_2 (HR and stroke volume) and SVR. Hughson et al. (109) used transfer function analysis and autoregressive moving average analysis to explore the interrelationships between CVP as an input to the cardiopulmonary baroreflex and MAP as an input to the arterial baroreflex in the regulation of SVR.

O'Leary et al. (110) used transfer function analysis to study the HR and vascular response to spontaneous changes in blood pressure from the relationships of systolic blood pressure to heart rate, MAP to SVR, and cerebrovascular resistance index, as well as stroke volume to SVR in healthy subjects in supine and 45° head-up tilt positions. Their data, which showed changes in MAP preceded changes in SVR as well as a possible link between stroke volume and SVR are consistent with complex interactions between the vascular component of the arterial and cardiopulmonary baroreflexes and intrinsic properties such as the myogenic response of the resistance arteries.

Toska et al. (111) used their mathematical model of baroreflexes and a simple circulation to analyze data from a previous study on humans (112). They modeled the heart, vascular bed, baroceptor reflexes and the ANS.

Microcirculation

Ostensibly, the microcirculation is the ultimate mediator of the major purposes of the circulation: delivering O_2 and cleansing the body of waste products, including CO_2. In addition to the mission of mass transport, the microcirculation and its endothelial cells have the role of regulation, signal transduction, proliferation, and repair. Lee (113) suggests, in addition, that the microcirculation is distensible and contains 40–50% of the total blood volume. In his Distinguished Lecture, he emphasized the integrative role of the microcirculation on circulatory control and its therapeutic role on blood volume compensation. He discussed shifts of volume from the microcirculation to the macrocirculation. It is possible that the microcirculation can play a more important role as a reservoir to compensate for blood volume loss than the venous system, and models are needed to investigate this intriguing concept.

Almost as an anomaly, it appears that hematocrit and nodal pressures can oscillate spontaneously in large microvascular networks in the absence of biological control. Carr and Lacoin (114) developed a model that not only explains the phenomenon, but also demonstrates how well-known phenomena explain it.

Schmidt-Schönbein (115) reviewed the mathematics of the microcirculation, including microvascular network topology, growth, and fluid mechanics; viscoelasticity and shape of microvessels; myogenic response; microvascular pressure–flow relationships; microvascular flow during pulsatile pressures; and non-Newtonian properties of blood in the microcirculation.

Pries and Secomb (116), in a paper that is part of the physiome project (see below), suggest a paradigm for attacking the complexity of the microcirculation. Terminal vascular beds exhibit a high degree of heterogeneity. Pertinent parameters are nonlinearly related, and their distributions are not independent. The classical typical vessel approach using averaged values for different vessel classes may not lead to a correct understanding of the physiology and pathophysiology of terminal vascular beds. Such problems can be avoided by studying microcirculatory functions at the network level using a combination of experiments and theoretical models. In this approach, distributions and relationships of pertinent parameters are measured *In vivo*, leading to the development of comprehensive databases. Such databases can be analyzed and complemented by suitable mathematical models, permitting estimation of parameters that are difficult to measure, and critical assessment of quantitative theories and hypotheses for microvascular function. This collaborative process between experimentally and theoretically oriented investigators may be facilitated in the future by the development of Web-based repositories of experimental data and theoretical models.

Beard and Bassingthwaighte (117) used a realistic geometric model for the 3D capillary network geometry as a framework for studying the transport and consumption of oxygen in cardiac tissue. The nontree-like capillary network conforms to the available morphometric statistics and is supplied by a single arterial source and drains into a pair of venular sinks. They explored steady-state O_2 transport and consumption in the tissue using a mathematical model that accounts for advection in the vascular network, nonlinear binding of dissolved oxygen to hemoglobin and myoglobin, passive diffusion of freely dissolved and protein-bound oxygen, and Michaelis–Menten consumption in the parenchymal tissue. The advection velocity field is determined by solving the hemodynamic problem for flow throughout the network. The resulting system is described by a set of coupled nonlinear elliptic equations, which are solved using a finite-difference numerical approximation. They found that coupled advection and diffusion in the 3D system enhance the dispersion of O_2 in the tissue compared with the predictions of simplified axially distributed models, and that no lethal corner, or oxygen-deprived region occurs for physiologically reasonable values for flow and consumption.

Cerebral

Considerable space has been devoted to this physiologically and personally important subject.

Clark and Kufahl (118) described a rigid-vessel model of the Circle of Willis, the complex circulatory system at the base of the brain. The circle is essential for redistributing blood flow after the sudden occlusion of a major cerebral vessel.

Lakin et al. (119) described a whole-body mathematical model for intracranial pressure dynamics. The model does not satisfy our definition of whole-body model, however, and we are including it in this section. Having said that, the model does avoid not simply presenting an isolated model of cerebral circulation. The model incorporates the dynamics of intracranial pressures, volumes, and flows. In addition to vascular connections with the rest of the body, the model incorporates a spinal-subarachnoid CSF compartment that bridges intracranial and extracranial physiology, allowing explicit buffering of ICP fluctuations by the spinal theca. The model contains cerebrovascular autoregulation, regulation of systemic vascular pressures by the sympathetic nervous system, regulation of CSF production in the choroid plexus, a lymphatic system, colloid osmotic pressure effects, and descriptions of CO_2.

Olufsen et al. (120) used a similar approach to study CBF during posture change from sitting to standing. Their model described pulsatile blood flow velocity and pressure in several compartments representing the systemic circulation. The model included compartments representing the trunk and upper extremities, the lower extremities, the brain, and the heart. They used physiologically based control mechanisms to describe the regulation of CBF velocity and arterial pressure in response to orthostatic hypotension resulting from postural change.

Sato et al. (121) first studied dynamic cerebrovascular responses in healthy humans during repetitive stepwise upward tilt (SUT) and stepwise downward tilt (SDT) maneuvers. The tilt maneuvers produced stepwise changes in

both cerebral perfusion pressure and mean CBF velocity. The latter's response to SUT was well characterized by a linear second-order model. However, that to SDT demonstrated a biphasic behavior that was described significantly better by the addition of a slowly responding component to the second-order model. This difference may reflect both different CV responses to SUT or SDT and different cerebrovascular autoregulatory behaviors in response to decreases or increases in cerebral perfusion pressure.

The brain not only needs flow and O_2 regulation, it also needs temperature regulation, because of so many critical chemical and physical-chemical reactions. To study this phenomenon, Zhu (122) modeled selective brain cooling during hyperthermia. They developed a theoretical model to describe the effects of blood flow rate and vascular geometry on the thermal equilibration in the carotid artery based on the blood flow and the anatomical vascular geometry in the human neck. The potential for cooling of blood in the carotid artery on its way to the brain by heat exchange with the jugular vein and by radial heat conduction loss to the cool neck surface was evaluated. They showed that the temperature of the arterial blood can be as much as 1.1 °C lower than the body core temperature, an observation in agreement with the difference between tympanic and body core temperatures. The model also evaluates the relative contributions of countercurrent heat exchange and radial heat conduction to selective brain cooling.

Does O_2 directly regulate CBF, or are there mediators? Ursino et al. (123,124) modeled the production and diffusion of vasoactive chemical factors involved in CBF regulation. Their model comprises two submodels. In the first, transport from capillary blood to cerebral tissue was analyzed to link changes in mean tissue PO_2 with CBF and arterial O_2 concentration changes. The second submodel described the production of vasoactive metabolites by cerebral parenchyma, arising from to a lack of O_2, and their diffusion toward pial perivascular space. They simulated the time dynamics of mean tissue PO_2, perivascular adenosine concentration and perivascular pH with changes in CBF. With their model, they concluded that the time delay introduced by diffusion processes is negligible compared with the other time constants in their system.

A second model (124) incorporated more submodels, each closely related to a physiological event. Thus, they could simulate the role played by the chemical factors described in the paragraph above, in the control of CBF during several different physiological and pathological conditions associated with the O_2 supply to cerebral tissue. These conditions included changes in autoregulation to changes in arterial and venous pressure, reactive hyperemia following cerebral ischemia and hypoxia. Their results suggest that adenosine and pH play a significant, but not exclusive, role in the regulation of the cerebrovascular bed.

Ursino and Magosso (125) presented a mathematical model of cerebrovascular regulation, in which emphasis was given to the role of tissue hypoxia on CBF. In the model, three different mechanisms are assumed to work on smooth muscle tension at the level of large and small pial arteries: CO_2 reactivity, tissue hypoxia, and a third mechanism necessary to provide good reproduction of autoregulation to cerebral perfusion pressure changes. The model is able to reproduce the pattern of pial artery caliber and CBF under a large variety of physiological stimuli, either acting separately (hypoxia, cerebral perfusion pressure changes, CO_2 pressure changes) or in combination (hypercapnia + hypoxia; hypercapnia + hypotension; hypotension). Furthermore, the model can explain the increase in CBF and the vasoconstriction of small pial arteries observed experimentally during hemodilution, ascribing it to the decrease in blood viscosity and to the antagonistic action of the flow-dependent mechanism (responsible for vasoconstriction) and of hypoxia (responsible for vasodilation). The interaction between hypoxia and ICP turns out to be quite complex, leading to different ICP time patterns, depending on the status of the CSF outflow pathways and of intracranial compliance.

Wakeland et al. (126) described a computer model of ICP dynamics that evaluated clinical treatment options for elevated ICP during traumatic brain injury. The model used fluid volumes as primary state variables and explicitly modeled fluid flows as well as the resistance, compliance, and pressure associated with each intra- and extracranial compartment (arteries and arterioles, capillary bed, veins, venous sinus, ventricles, and brain parenchyma). The model evaluated clinical events and therapies, such as intra- and extraparenchymal hemorrhage, cerebral edema, CSF drainage, mannitol administration, head elevation, and mild hyperventilation. The model was able to replicate observed clinical behavior in many cases, including elevated ICP associated with severe cerebral edema following subdural, epidural, or intraparynchemal hematoma. The model also mimics cerebrovascular regulatory mechanisms that are activated during traumatic brain injury.

Lodi and Ursino (127) demonstrated a mathematical model of cerebral hemodynamics during vasospasm. The model divided arterial hemodynamics into two cerebral territories: with and without vasospasm. It also included collateral circulation between the two territories, cerebral venous hemodynamics, CSF circulation, ICP, and craniospinal storage capacity. Moreover, the pial artery circulation in both territories was affected by CBF autoregulation mechanisms. First, the model was used to simulate some clinical results reported in the literature, concerning the patterns of middle cerebral artery flow velocity, CBF and pressure losses during vasospasm. Second, they performed a sensitivity analysis on certain model parameters (severity of caliber reduction, longitudinal extension of the spasm, autoregulation gain, ICP, resistance of the collateral circulation, and MAP) to clarify their influence on hemodynamics in the spastic area. The results suggested that the clinical impact of vasospasm depends on several concomitant factors, which should be simultaneously taken into account to reach a proper diagnosis.

Pasley et al. (128) used a mathematical model to test two hypotheses: (1) cyclic extravascular compressional modulation of the terminal venous bed occurs with positive pressure inhalation; and (2) the degree of modulation is diminished with increasing vascular dilation induced by

increasing the level of $PaCO_2$. They made two modifications of Ursino's model of CSF dynamics (129–131):(1) terminal venous bed resistance was synchronously modulated with the ventilation cycle; and (2) both the depth of modulation and cerebrovascular resistance were progressively reduced with increasing levels of $PaCO_2$. Simulated and experimental correlation values progressively increased monotonically as the level of PCO_2 increased. Their results suggested that dilation of the cerebral vasculature reduces the influence of positive pressure ventilation on ICP by increasing venous pressure and thus diminishing the likelihood of vascular compression.

Increased ICP, which can result from etiologies ranging from tumors to trauma, can produce devastating results, of which death may be one of the more merciful. Modeling the phenomenon is critically important. Ursino and Lodi (132) used a mathematical model to characterize the relationships among CBF, cerebral blood volume, ICP, and the action of cerebrovascular regulatory mechanisms (autoregulation and CO_2 reactivity). The model incorporated CSF circulation, the ICP–volume relationship, and cerebral hemodynamics. The latter is based on three assumptions. (1) The middle cerebral arteries behave passively following transmural pressure changes. (2) The pial arterial circulation includes two segments (large and small pial arteries) subject to different autoregulation mechanisms. (3) The venous cerebrovascular bed behaves as a Starling resistor. A new aspect of this model relates to the description of CO_2 reactivity in the pial arterial circulation and in the analysis of its nonlinear interaction with autoregulation. Simulations obtained at constant ICP using various combinations of MAP and CO_2 support data on CBF and velocity concerning both the separate effects of CO_2 and autoregulation and their nonlinear interaction. Simulations performed in dynamic conditions with varying ICP suggest a significant correlation between ICP dynamics and cerebral hemodynamics in response to CO_2 changes. The authors believe that the model can be used to study ICP and blood velocity time patterns in neurosurgical patients, so that one can gain a deeper insight into the pathophysiological mechanisms leading to intracranial hypertension and resultant brain damage.

Loewe et al. (133) used a mathematical model to simulate the time pattern of ICP and of blood velocity in the middle cerebral artery in response to maneuvers simultaneously affecting MAP and end-tidal CO_2. First, they performed a sensitivity analysis, to clarify the role of some important model parameters (CSF outflow resistance, intracranial elastance coefficient, autoregulation gain, and the position of the regulation curve) during CO_2 alteration maneuvers performed at different MAP levels. Next, the model was applied to the reproduction of real ICP and velocity tracings in neurosurgical patients. They concluded that the model could be used to give reliable estimates of the main factors affecting intracranial dynamics in individual patients, starting from routine measurements performed in neurosurgical intensive care units.

Ursino et al. (134) analyzed changes in cerebral hemodynamics and ICP evoked by MAP and $PaCO_2$ challenges in patients with acute brain damage. The study was performed using a simple mathematical model of intracranial hemodynamics, particularly aimed at routine clinical investigation. The parameters chosen for the identification summarize the main aspects of intracranial dynamics, namely, CSF circulation, intracranial elastance, and cerebrovascular control.

By using a mathematical model, Ursino et al. (135) also studied the time pattern of ICP in response to typical clinical tests, namely, a bolus injection or withdrawal of small amounts of saline in the craniospinal space in patients with acute brain damage. The model included the main biomechanical factors assumed to affect ICP, CSF dynamics, intracranial compliance, and cerebrovascular dynamics. The simulation results demonstrated that the ICP time pattern cannot be explained simply on the basis of CSF dynamics, but also requires consideration of the contribution of cerebral hemodynamics and blood-volume alterations.

Sharan et al. (136) explored an interesting O_2-related phenomenon with a mathematical model. CBF increases as arterial O_2 content falls with hypoxic (low PO_2), anemic (low hemoglobin), and carbon monoxide (CO) (high carboxyhemoglobin) hypoxia. Despite a higher arterial PO_2, CO hypoxia provokes a greater increase in CBF than hypoxic hypoxia. They analyzed published data using a compartmental mathematical model to test the hypothesis that differences in PO_2 in tissue, or a closely related vascular compartment, account for the greater response to CO hypoxia. Calculations showed that tissue, but not arteriolar, PO_2 was lower in CO hypoxia because of the increased oxyhemoglobin affinity with CO hypoxia. Analysis of studies in which oxyhemoglobin affinity was changed independently of CO supports the conclusion that changes in tissue PO_2 (or closely related capillary or venular PO_2) predict alterations in CBF. They then sought to determine the role of tissue PO_2 in anemic hypoxia, with no change in arterial and little, if any, change in venous PO_2. Calculations predicted a small fall in tissue PO_2 as hematocrit decreases from 55 to 20%. However, calculations showed that changes in blood viscosity can account for the increase in CBF in anemic hypoxia over this range of hematocrits. It would have been interesting if the authors had tested hypoxia from cyanide poisoning, which blocks the utilization of O_2 at the enzyme cytochrome oxidase; blood and tissue actually increases.

Exploring well-defined physical phenomena is one thing; exploring fuzzy, undefined concepts, like consciousness, is quite another. Cammarota and Onaral (137) realized that complex physiological systems in which the emergent global (observable) behavior results from the interplay among local processes cannot be studied effectively by conventional mathematical models. In contrast to traditional computational methods, which provide linear or nonlinear input–output data mapping without regard to the internal workings of the system, complexity theory offers scientifically and computationally tractable models that take into account microscopic mechanisms and interactions responsible for the overall input–output behavior. The authors offered a brief introduction to some of the tenets of complexity theory and outlined the process involved in the development and testing of a model that duplicates the global dynamics of the induction of loss of

consciousness in humans due to cerebral ischemia. Under the broad definition of complexity, they viewed the brain of humans as a complex system Successful development of a model for this complex system requires careful combination of basic knowledge of the physiological system both at the local (microscopic) and global (macroscopic) levels with experimental data and the appropriate mathematical tools. It represents an attempt to develop a model that can both replicate human data and provide insights about possible underlying mechanisms. They presented a model for complex physiological systems that undergo state (phase) transitions. The physiological system modeled is the CNS, and the global behavior captured by the model is the state transition from consciousness to unconsciousness. Loss of consciousness can result from many conditions such as ischemia (low blood flow), hypoxia (low oxygen), hypoglycemia, seizure, anesthesia, or a blow to the head, among others. Successful development of a model for this complex system requires careful combination of basic knowledge of the physiological system both at the local (microscopic) and global (macroscopic) levels with experimental data and the appropriate mathematical tools. Due to the wealth of human research and data available, the specific focus of the model is unconsciousness that results from the cerebral ischemia experienced by aircrew during aggressive maneuvering in high-performance aircraft.

Coronary

In the section Coronary Autoregulation, the problems and paradoxes of the coronary circulation are described. One of them was that myocardial perfusion was decreased in some areas by mechanical and shearing effects, and the harder the contraction, the greater the impairment. Smith (138) used an anatomically based computational model of coronary blood flow, coupled to cardiac mechanics to investigate the mechanisms by which myocardial contraction inhibits coronary blood flow. From finite deformation mechanics solutions the model calculates the regional variation in intramyocardial pressure (IMP) exerted on coronary vessels embedded in the ventricular wall. This pressure is then coupled to a hemodynamic model of vascular blood flow to predict the spatial–temporal characteristics of perfusion throughout the myocardium. The calculated IMP was shown to vary approximately linearly between ventricular pressure at the endocardium and atmospheric pressure at the epicardium through the diastolic loading and isovolumic contraction phases. During the ejection and isovolumic relaxation phases, IMP values increased slightly above ventricular pressure. The average radius of small arterial vessels embedded in the myocardium decreased during isovolumic contraction (18% in LV endocardium) before increasing during ejection (10% in LV endocardium) due to a rise in inflow pressure. Embedded venous vessels show a reduction in radius through both phases of contraction (35% at left ventricular endocardium). Calculated blood flows in both the large epicardial and small myocardial vessels show a 180° phase difference between arterial and venous velocity patterns with arterial flow occurring predominantly during diastole and venous flow occurring predominantly during systole. Their results confirm that the transmission of ventricular cavity pressure through the myocardium is the dominant mechanism by which coronary blood flow is reduced during the isovolumic phase of contraction. In the ejection phase of contraction, myocardial stiffening plays a more significant role in inhibiting blood flow.

Also illustrating this problem of impaired coronary flow during systole is a mathematical model that was based on an *In vitro* mechanical model consisting mainly of collapsible tubes (67). The pressure and flow signals obtained from both models were similar to physiological human coronary pressure and flow, both for baseline and hyperemic conditions.

Smith et al. (139) developed a discrete anatomically accurate finite element model of the largest six generations of the coronary arterial network. Using a previously developed anatomically accurate model of ventricular geometry, they defined the boundaries of the coronary mesh from measured epicardial coronaries. Network topology was then generated stochastically from published anatomical data. Spatial information was added to the topological data using an avoidance algorithm accounting for global network geometry and optimal local branch-angle properties. The generated vessel lengths, radii and connectivity were consistent with published data, and a relatively even spatial distribution of vessels within the ventricular mesh was achieved.

Pulmonary–Respiratory

The respiratory system is important as a means of O_2 uptake and CO_2 elimination. It may be compared with the CVS because gases are carried in it by pulsatile fluid flow, somewhat as gases and many other substances are carried by pulsatile blood flow in the CVS. The respiratory system is anatomically simpler, since it has but one branching out of the airflow passages, whereas the CVS fans out to the many body capillaries from the aorta, then fans in to the vena cavae, and repeats this pattern in the pulmonary circulation. However, analysis and modeling are far more complex in the respiratory system because air is a compressible fluid and because flow of air in the lungs is a tidal, or back-and forth, flow, in contrast to the one-way flow with superimposed pulsatility in the CVS. Also, although the respiratory system does not have valves as the CVS does, there are some important and rather difficult nonlinearities.

Good reviews are worth their weight in gold. Grotberg (140) reviewed respiratory fluid mechanics and transport processes. This field has experienced significant research activity for decades. Important contributions to the knowledge base come from pulmonary and critical care medicine, anesthesia, surgery, physiology, environmental health sciences, biophysics, and engineering. Several disciplines within engineering have strong and historical ties to respiration, including mechanical, chemical, civil–environmental, aerospace and, of course, biomedical engineering. Grotberg's review draws from the wide variety of scientific literature that reflects the diverse constituency and audience that respiratory science has developed. The subject areas covered include nasal flow and transport, airway gas

flow, alternative modes of ventilation, nonrespiratory gas transport, aerosol transport, airway stability, mucus transport, pulmonary acoustics, surfactant dynamics and delivery, and pleural liquid flow. Within each area are several subtopics whose exploration can provide the opportunity of both depth and breadth for the interested reader.

The pioneer of computer modeling of respiratory control was Jim Defares (141,142). The one with the most impact, however, has been Fred Grodins (143), and several of the papers in this section acknowledge his significant contributions.

Chiari et al. (144) presented a comprehensive model of O_2 and CO_2 exchange, transport, and storage. The model comprises three compartments (lung, body tissue, and brain tissue) and incorporates a controller that adjusts alveolar ventilation and CO_2 by dynamically integrating stimuli coming from peripheral and central chemoreceptors. A realistic CO_2 dissociation curve based on a two-buffer model of acid–base chemical regulation is included. In addition, the model considers buffer base, the nonlinear interaction between the O_2 and CO_2 chemoreceptor responses, pulmonary shunt, dead space, variable time delays, and Bohr and Haldane effects. Their model fit the experimental data of ventilation and gas partial pressures in a very large range of gas intake fractions. It also provided values of blood concentrations of CO_2, HCO_3^-, and hydrogen ions in good agreement with more complex models characterized by an implicit formulation of the CO_2 dissociation curve.

Good sensitivity analysis can be difficult, at best. The tools in the paper by Hyuan et al. (145) are general and can be applied to a wide class of nonlinear models. The model incorporates a combined theoretical and numerical procedure for sensitivity analyses of lung mechanics models that are nonlinear in both state variables and parameters. They applied the analyses to their own nonlinear lung model, which incorporates a wide range of potential nonlinear identification conditions including nonlinear viscoelastic tissues, airway inhomogeneities via a parallel airway resistance distribution function, and a nonlinear block-structure paradigm. Model nonlinearities motivate sensitivity analyses involving numerical approximation of sensitivity coefficients. Examination of the normalized sensitivity coefficients provides insight into the relative importance of each model parameter, and hence the respective mechanism. More formal quantification of parameter uniqueness requires approximation of the paired and multidimensional parameter confidence regions. Combined with parameter estimation, they used the sensitivity analyses to justify tissue nonlinearities in modeling of lung mechanics for healthy and constricted airway conditions, and to justify both airway inhomogeneities and tissue nonlinearities during broncoconstriction. Some of the variables, parameters and domains included pressures, flows, volumes, resistances, compliances, impedances, pressure-volume and frequency.

A model of breathing mechanics (146) was used to interpret and explain the time course of input respiratory resistance during the breathing cycle, observed in ventilated patients. The authors assumed a flow-dependent resistance for the upper extrathoracic airways and volume-dependent resistance and elastance for the intermediate airways. A volume-dependent resistance described the dissipative pressure loss in the lower airways, while two constant elastances represented lung and chest wall elasticity. Simulated mouth flow and pressure signals obtained in a variety of well-controlled conditions were used to analyze total respiratory resistance and elastance estimated by an on-line algorithm based on a time-varying parameter model. These estimates were compared with those provided by classical estimation algorithms based on time-invariant models with two, three, and four parameters. The results confirmed that the difference between the end-expiration and end-inspiration resistance increases when obstructions shift from the upper to the lower airways.

Ursino et al. (147) presented a mathematical model of the human respiratory control system. It includes three compartments for gas storage and exchange (lungs, brain, tissue, and other body tissues), and various types of feedback mechanisms. These comprise peripheral chemoreceptors in the carotid body, central chemoreceptors in the medulla and a central ventilatory depression. The last acts by reducing the response of the central neural system to the afferent peripheral chemoreceptor activity during prolonged hypoxia of the brain tissue. The model also considers local blood flow adjustments in response to O_2 and CO_2 arterial pressure changes. Sensitivity analysis suggests that the ventilatory response to CO_2 challenges during hyperoxia can be almost completely ascribed to the central chemoreflex, while, during normoxia, the peripheral chemoreceptors also provide a modest contribution. By contrast, the response to hypercapnic stimuli during hypoxia involves a complex superimposition among different factors with disparate dynamics. Results suggest that the ventilatory response to hypercapnia during hypoxia is more complex than that provided by simple empirical models, and that discrimination between the central and peripheral components based on time constants may be misleading.

The phenomena collectively referred to as periodic breathing (including Cheyne Stokes respiration and apneustic breathing) have important medical implications. The hypothesis that periodic breathing is the result of delays in the feedback signals to the respiratory control system has been studied since the work of Grodins et al. (148) in the early 1950s. Batzel's dissertation (149) extended a model developed by Khoo et al. (150), to include variable delay in the feedback control loop and to study the phenomena of periodic breathing and apnea as they occur during quiet sleep in infants. The nonlinear mathematical model consists of a feedback control system of five differential equations with multiple delays. Numerical simulations were performed to study instabilities in the control system, especially the occurrence of periodic breathing and apnea in infants $\sim$4 months of age. This time frame is important, since during it there is a high incidence of Sudden Infant Death Syndrome. Numerical simulations indicate that a shift in the controller ventilatory drive set point during sleep transition is an important factor for instability. Analytical studies show that delay-dependent stability is affected by controller gain, compartment

volumes and the manner in which changes in minute ventilation are produced (i.e., by deeper breathing or faster breathing). Parenthetically, the increased delay resulting from congestive heart failure can induce instability at certain control gain levels.

The dimensions, composition, and stiffness of the airway wall are important determinants of airway cross-sectional area during dynamic collapse in a forced expiration or when airway smooth muscle is constricted. This can occur with asthma or COPD (emphysema). Under these circumstances, airway caliber is determined by an interaction between the forces acting to open the airway (parenchymal tension and wall stiffness) and those acting to close it (smooth-muscle force and surface tension at the inner gasliquid interface). Theoretical models of the airway tube law (relationship between cross-sectional area and transmural pressure) allow simulations of airway constriction in normal and asthmatic airways (151).

An excellent mathematical model of neonatal respiratory control (152) consists of a continuous plant and a discrete controller. Included in the plant are lungs, body tissue, brain tissue, a CSF compartment, and central and peripheral receptors. The effect of shunt is incorporated in the model, and lung volume and dead space are time varying. The controller uses outputs from peripheral and central receptors to adjust the depth and rate of breathing, and the effects of prematurity of peripheral receptors are included in the system. Hering–Breuer-type reflexes are embodied in the controller to accomplish respiratory synchronization. See also the Nottingham Physiology Simulator for a similar approach (Whole-body models).

Lung gas composition affects the development of anesthesia-related atelectasis, by way of differential gas absorption. A mathematical model (153) examines this phenomenon by combining models of gas exchange from an ideal lung compartment, peripheral gas exchange, and gas uptake from a closed collapsible cavity. The rate of absorption is greatest with O_2, less with NO_2 and minimal with N_2. BODY Simulation (see PBPM) achieves the same results.

Most respiratory models are limited to short-term (minutes) control. Those wishing to model longer term (days) control and adjustments should start with the detailed review by Dempsy and Forster (154). They discuss several important areas, including central chemoreception, cerebral fluids and chemoreceptor environment, physiologically important stimuli to medullary chemoreception, medullary chemoreceptor contributions to ventilatory drive, metabolic acid–base derangements, ventilatory response, mediation of ventilatory adaptation, ventilatory acclimatization to chronic hypoxia, ventilation during acute hypoxia, acclimatization during short-term hypoxia, acclimatization during long-term hypoxia, physiological significance of short- and long-term ventilatory acclimatization, ventilatory acclimatization to chronic CO_2 exposure, ventilation during chronic CO_2 inhalation, and whole-body CO_2 and H^+ during CO_2 exposure.

Renal

The kidneys have important physiological functions including maintenance of water and electrolyte balance; synthesis, metabolism and secretion of hormones; and excretion of the waste products from metabolism. In addition, the kidneys play a major role in the excretion of hormones, as well as drugs and other xenobiotics. The story of fluids and solutes is the story of the kidney, and vice versa.

The understanding of renal function has profited greatly from quantitative and modeling approaches for a century (155). One of the most salient examples is the concentration and dilution of the urine: a fundamental characteristic of the mammalian kidney. Only in the last three decades have the necessary components of this and other renal mechanisms been confirmed at the molecular level, but there have also been surprises. In addition, the critical role played by the fine regulation of Na^+ reabsorption in the collecting duct for the maintenance of normal blood pressure presents challenges to our understanding of the integrated interaction among systems. As a first step in placing the kidney in the physiome paradigm (see below), Schafer suggests (1) integrating currently restricted mathematical models, (2) developing accessible databases of critical parameter values together with indices of their degrees of reliability and variability, and (3) describing regulatory mechanisms and their interactions from the molecular to the interorgan level.

By now, you will have guessed our enthusiasm for good reviews. An excellent one to start with in this area is by Russell (156), on Na-K chloride cotransport. Obligatory, coupled cotransport of Na^+, K^+, and Cl^- by cell membranes has been reported in nearly every animal cell type. Russell's review examines the status of the knowledge about this ion transport mechanism.

In another review (the Starling Lecture, actually), DiBona (157) describes the neural control of the kidney. The sympathetic nervous system provides differentiated regulation of the functions of various organs. This differentiated regulation occurs via mechanisms that operate at multiple sites within the classic reflex arc: peripherally at the level of afferent input stimuli to various reflex pathways, centrally at the level of interconnections between various central neuron pools, and peripherally at the level of efferent fibers targeted to various effectors within the organ. In the kidney, increased renal sympathetic nerve activity regulates the functions of the intrarenal effectors: the tubules, the blood vessels, and the juxtaglomerular granular cells. This enables a physiologically appropriate coordination between the circulatory, filtration, reabsorptive, excretory, and renin secretory contributions to overall renal function. Anatomically, each of these effectors has a dual pattern of innervation consisting of a specific and selective innervation, in addition to an innervation that is shared among all the effectors. This arrangement permits maximum flexibility in the coordination of physiologically appropriate responses of the tubules, the blood vessels, and the juxtaglomerular granular cells to a variety of homeostatic requirements.

Physiologists have developed many models for interpreting water and solute exchange data in whole organs, but the models have often neglected key aspects of the underlying physiology to present the simplest possible model for a given experimental situation. Kellen and

Bassingthwaighte (158) developed a model of microcirculatory water and solute exchange and applied it to diverse observations of water and solute exchange in the heart. The key model features that permit this diversity are the use of an axially distributed blood-tissue exchange region, inclusion of a lymphatic drain in the interstitium, and the independent computation of transcapillary solute and solvent fluxes through three different pathways.

Endocrine

The insulin–glucose subsystem will be used as a paradigm of the endocrine system. Because diabetes is such a clinically complex and disabling disease, as well as a major and increasing personal and public health problem, many attempts at modeling have been made. Most of these models have examined various parts of the overall process, and one or two have tried to put it all together.

The normal blood glucose concentration in humans lies in the range of 70–110 $mg \cdot dL^{-1}$. Exogenous factors that affect this concentration include food intake, rate of digestion, exercise, and reproductive state. The pancreatic hormones insulin and glucagon are responsible for keeping glucose concentration within bounds. Insulin and glucagon are secreted from beta-cells and alpha-cells respectively, which are contained in the islets of Langerhans, which are scattered in the pancreas. When blood glucose concentration is high, the beta-cells release insulin, resulting in lowering blood glucose concentration by inducing the uptake of the excess glucose by the liver and other cells (e.g., muscles) and by inhibiting hepatic glucose production. When blood glucose concentration is low, the alpha-cells release glucagon, resulting in increasing blood glucose concentration by acting on liver cells and causing them to release glucose into the blood.

Glucose concentrations outside the range 70–110 $m \cdot dL^{-1}$ are called hyperglycemia or hypoglycemia. Diabetes mellitus is a disease of the glucose–insulin regulatory system that is characterized by hyperglycemia. Diabetes is classified into two main categories: type 1 diabetes, juvenile onset and insulin dependent; and type 2 diabetes, adult onset and insulin independent.

Makroglu et al. (159) presented an extensive overview of some of the available mathematical models on the glucose–insulin regulatory system in relation to diabetes. The review is enhanced with a survey of available software. The models are in the form of ordinary differential, partial differential, delay differential and integrodifferential equations.

Tibell et al. (160) used models to estimate insulin secretion rates in patients who had undergone pancreas-renal transplant procedures.

When the complex physiology goes awry, manmade control systems can be used to understand and perhaps deal with the problem. These control systems require models. Ibbini et al. (161) have recently developed one such system. Parker et al. (162) set up a model-based algorithm for controlling blood glucose concentrations in type I diabetic patients.

Bequette (163) examined the development of an artificial pancreas in the context of the history of the field of feedback control systems, beginning with the water clock of ancient Greece, and including a discussion of current efforts in the control of complex systems. The first generation of artificial pancreas devices included two manipulated variables (insulin and glucose infusion) and nonlinear functions of the error (difference between desired and measured glucose concentration) to minimize hyperglycemia while avoiding hypoglycemia. Dynamic lags between insulin infusion and glucose measurement were relatively small for these intravenous-based systems. Advances in continuous glucose sensing, fast-acting insulin analogues, and a mature insulin-pump market bring closer the commercial realization of a closed-loop artificial pancreas. Model predictive control is discussed in-depth as an approach that is well suited for a closed-loop artificial pancreas. A major remaining challenge is handling an unknown glucose disturbance (meal), and an approach is proposed to base a current insulin infusion action on the predicted effect of a meal on future glucose values. Better meal models are needed, as a limited knowledge of the effect of a meal on the future glucose values limits the performance of any control algorithm.

One of the fascinating features of the insulin–glucose subsystem, and with other endocrine subsystems, is that it either oscillates or releases its hormones in an intermittent fashion. The terms oscillations, rhythms, ultradian, (relating to biologic variations or rhythms occurring in cycles more frequent than every 24 h [cf. circadian, about every 24 h])pulses, biphasic and bursting appear frequently in regards to insulin secretion. The following describes some of the models that study these phenomena.

Insulin is secreted in a sustained oscillatory fashion from isolated islets of Langerhans, and Berman et al. (164) modeled this phenomenon. Straub and Sharp (165) reviewed some models that have tried to explain the well-documented biphasic secretory response of pancreatic beta-cells to abrupt and sustained exposure to glucose.

Lenbury et al. (166) modeled the kinetics of insulin, using a nonlinear mathematical model of the glucose–insulin feedback system. The model has been extended to incorporate the beta-cells' function in maintaining and regulating plasma insulin concentration in humans. Initially, a gastrointestinal absorption term for glucose is used to effect the glucose absorption by the intestine and the subsequent release of glucose into the bloodstream, taking place at a given initial rate and falling off exponentially with time. An analysis of the model was carried out by the singular-perturbation technique to derive boundary conditions on the system parameters that identify, in particular, the existence of limit cycles in the model system consistent with the oscillatory patterns often observed in clinical data. They then used a sinusoidal term to incorporate the temporal absorption of glucose to study the responses in patients during ambulatory-fed conditions. They identified the ranges of parametric values for which chaotic behavior can be expected, leading to interesting biological interpretations.

An interesting electrophysiological phenomenon occurs in the islets of Langerhans: The release of insulin is controlled in these islets by trains of action potentials occurring in rapid bursts followed by periods of quiescence. This

bursting behavior occurs only in intact islets: single cells do not display such bursting activity. Chay and Keizer (167) were the first attempt to model this phenomenon quantitatively. Sherman et al. (168) sought to explain the absence of bursting in single beta-cells using the idea of channel sharing. Keizer (169) modified this model by substituting an ATP- and ADP-dependent K channel instead of the Ca-dependent K channel. This model was then further improved by Keizer and Magnus (170). Sherman et al. (171) constructed a domain model to examine the effect of Ca on Ca-channel inactivation. Further refinements have been made by Keizer and Young (172).

Sherman (173) also reviewed mechanisms of ionic control of insulin secretion. He focused on aspects that have been treated by mathematical models, especially those related to bursting electrical activity. The study of these mechanisms is difficult because of the need to consider ionic fluxes, Ca handling, metabolism, and electrical coupling with other cells in an islet. The data come either from islets, where experimental maneuvers tend to have multiple effects, or from isolated cells, which exhibit degraded electrical activity and secretory sensitivity. Modeling aids in the process by integrating data on individual components such as channels and Ca handling and testing hypotheses for coherence and quantitative plausibility. The study of a variety of models has led to some general mathematical results that have yielded qualitative model-independent insights.

Endocrine systems often secrete hormones in pulses. Examples include the release of growth hormone and gonadotropins,, as well as insulin. These hormones are secreted over intervals of 1–3 h and 80–150 min, respectively. It has been suggested that relative to constant or stochastic signals, oscillatory signals are more effective at producing a sustained response in the target cells. In addition to the slow insulin oscillations, there are more rapid pulses that occur every 8–15 min. The mechanisms underlying both types of oscillations are not fully understood, although it is thought that the more rapid oscillations may arise from an intrapancreatic pacemaker mechanism. One possible explanation of the slow insulin oscillations is an instability in the insulin-glucose feedback system. This hypothesis has been the subject many studies, including some that have developed a mathematical model of the insulin–glucose feedback system. Tolic (174) reviewed several models that investigated oscillations, with glucose and the pancreas. Tolic's model (175) and others are available on the CellML site (176).

Shannahoff-Khalsa et al. (177), in Gene Yates' group, have expanded the purview of these fascinating phenomena by comparing the rhythms of the CV, autonomic, and neuroendocrine systems. Their results, from a time-series analysis using a fast orthogonal search method, suggested that insulin secretion has a common pacemaker (the hypothalamus) or a mutually entrained pacemaker with these three systems.

Fortunately, there is an excellent glucose–insulin model that helps one understand some of these concepts. Erzen et al. (178) have posted GlucoSim, a web-based simulator that runs on almost all computer platforms. GlucoSim is a program for simulating glucose–insulin interaction in a healthy person and in a type 1 diabetes patient. It has a flexible data output structure that can be used as an input into most postprocessing programs such as spreadsheet and graphics programs. Simulations can be performed by changing initial conditions or meal and insulin injection times. Simulation results can be plotted in 2D directly from the simulator by choosing appropriate buttons. The software is freely available to all users. This model is also on the CellML Site (176), but not the running model.

GlucoSim is one of the more nearly complete models that we have reviewed, and it satisfies our criteria for a whole-body model. It is a good example of the usefulness of a whole-body model with many compartments. This model is actually two whole-body models combined, with food ingestion in the glucose model and subcutaneous injection in the insulin model.

Action Potentials

In 1952, Hodgkin and Huxley published a paper showing how a nonlinear empirical model of the membrane processes could be constructed (179). In the five decades since their work, the Hodgkin–Huxley paradigm of modeling cell membranes has been enormously successful. While the concept of ion channels was not established when they performed their work, one of their main contributions was the concept that ion-selective processes existed in the membrane. It is now known that most of the passive transport of ions across cell membranes is accomplished by ion-selective channels. In addition to constructing a nonlinear model, they also established a method to incorporate experimental data into a nonlinear mathematical membrane model. Thus, models of action potentials comprise some of the oldest of nonlinear physiological models.

Action potential is a term used to denote a temporal phenomenon exhibited by every electrically excitable cell. The transmembrane potential difference of most excitable cells rests at some negative potential called the resting potential, appropriately enough. External current or voltage inputs can cause the potential of the cell to deviate in a positive direction, and if the input is large enough, the result is an action potential. An action potential is characterized by a depolarization that typically results in an overshoot >0 mV, followed by repolarization. Some cells may actually hyperpolarize before returning to the resting potential. After the membrane has been excited, it cannot be reexcited until a recovery period, called the refractory period, has passed.

When excitable cells are depolarized from their resting potential beyond a certain level (threshold), they respond with a relatively large, stereotyped potential change. It is the action potential's propagating away from the site of origin that constitutes impulse conduction in nerve, muscle and heart.

Action potentials are everywhere in life; if there's electricity, there are action potentials. This topic deserves an entire book, much less an encyclopedia entry. Accordingly, much of the information in this comes from articles in other books. Unless one defines electricity as a system, there seem to be no models that fit our definition of physiological systems models.

Having said that, electric phenomena are an integral part of every biological system, no matter what one's definition of the latter is. Action potentials keep our heart beating, our mind thinking—and make possible our seeing, hearing, smelling, feeling, tasting, digesting and moving. They are everywhere (plants and animals) and taking place all the time. We tried to estimate the number of action potentials occurring per second in one part of the human body (the brain) and may be underestimating by orders of magnitude. There are a trillion neurons in the human brain alone, and 10 quadrillion synapses, more than there are stars in the universe. The rate of action potentials can be from zero to >1000 action potentials per second. Let us estimate 10^{18} s^{-1} ($10^{16} \times 10^2$). We have not even started with muscles or sensory receptors, for example. We repeat, it's a large topic.

Varghese (180) tells us that the key concept in the modeling of excitable cells is that of ion channel selectivity. A particular type of ion channel will only allow certain ionic species to pass through; most types of ion channels are modeled as being permeant to a single ionic species. In excitable cells at rest, the membrane is most permeable to K. This is because only K channels (i.e., channels selective to K) are open at the resting potential. For a given stimulus to result in an action potential, the cell has to be brought to threshold, that is, the stimulus has to be larger than some critical size. Smaller, subthreshold, stimuli will result in an exponential decay back to the resting potential. The upstroke, or fast initial depolarization of the action potential, is caused by a large influx of Na ions as Na channels open (in some cells, entry of Ca ions though Ca channels is responsible for the upstroke) in response to a stimulus. This is followed by repolarization as K ions start flowing out of the cell in response to the new potential gradient. While responses of most cells to subthreshold inputs are usually linear and passive, the suprathreshold response (the action potential) is a nonlinear phenomenon. Unlike linear circuits where the principle of superposition holds, the nonlinear processes in cell membranes do not allow responses of two stimuli to be added. If an initial stimulus results in an action potential, a subsequent stimulus administered at the peak voltage will not produce an even larger action potential; indeed, it may have no effect at all. Following an action potential, most cells have a refractory period, during which they are unable to respond to stimuli. Nonlinear features, such as these make modeling of excitable cells a nontrivial task. In addition, the molecular behavior of channels is only partially known and, therefore, it is not possible to construct membrane models from first principles.

Most membrane models discussed in Varghese's excellent article (180) involve the time behavior of electrochemical activity in excitable cells. These models are systems of ordinary differential equations where the independent variable is time. While a good understanding of linear circuit theory helps understand these models, most of the phenomena of interest involve nonlinear circuits with time-varying components. Electrical activity in plant and animal cells is caused by two main factors: (1) differences in the concentrations of ions inside and outside the cell; and (2) molecules embedded in the cell membrane that allow these ions to be transported across the membrane. The ion concentration differences and the presence of large membrane-impermeant anions inside the cell result in a polarity: the potential inside a cell is typically 50–100 mV lower than that in the external solution. Almost all of this potential difference occurs across the membrane itself; the bulk solutions both inside and outside the cell are usually at a uniform potential. This transmembrane potential difference is, in turn, sensed by molecules in the membrane, and these molecules control the flow of ions. The lipid bilayer, which constitutes the majority of the cell membrane, acts as a capacitor with a specific capacitance. Because the membrane is thin ($\sim$ 7.5 mm), it has a high capacitance, $\sim 1\ \mu F \cdot cm^{-2}$. The rest of the membrane comprises large protein molecules that act as (1) ion channels, (2) ion pumps, or (3) ion exchangers. The flow of ions across the membrane causes changes in the transmembrane potential, which is typically the main observable quantity in experiments.

Barr (181), in another excellent article, expounds on bioelectricity, which has its origin in the voltage differences present between the inside and outside of cells. These potentials arise from the specialized properties of the cell membrane, which separates the intracellular from the extracellular volume. Much of the membrane surface is made of a phospholipid bilayer, an electrically inert material. Electrically active membranes also include many different kinds of *integral proteins*, which are compact, but complex structures extending across the membrane. Each integral protein comprises a large number of amino acids, often in a single long polypeptide chain, which folds into multiple domains. Each domain may span the membrane several times. The multiple crossings may build the functional structure, for example, a *channel*, through which electrically charged ions may flow. The structure of integral proteins is loosely analogous to that of a length of thread passing back and forth across a fabric to form a buttonhole. Just as a buttonhole allows movement of a button from one side of the fabric to the other, an integral protein may allow passage of ions from the exterior of a cell to the interior, or vice versa. In contrast to a buttonhole, an integral protein has active properties, for example, the ability to open or close. Cell membranes possess several active characteristics important to the cell's bioelectric behavior. (1) Some integral proteins function as *pumps*. These pumps use energy to transport ions across a membrane, working against a concentration gradient, a voltage gradient, or both. The most important pump moves Na ions out of the intracellular volume and K ions into that volume. (2) Other integral proteins function as channels, that is, openings through the membrane that open and close over time. These channels can function selectively so that, for a particular kind of channel, only Na ions may pass through, for example. Other kinds of channel may allow only K ions or Ca ions. (3) The activity of the membrane's integral proteins is modulated by signals specific to its particular function. For example, some channels open or close in response to photons or to odorants; thus they function as sensors for light or smell. Pumps respond to the concentrations of the ions they move. Rapid electrical impulse transmission in nerve and muscle is made possible by changes that respond to the transmembrane potential

itself, forming a feedback mechanism. These active mechanisms provide ion-selective means of current's crossing the membrane, both against the concentration or voltage gradients (pumps) or in response to them (channels). While the pumps build up the concentration differences (and thereby the potential energy) that allow events to occur, channels use this energy actively to create the fast changes in voltage and small intense current loops that constitute nerve signal transmission, initiate muscle contraction, and participate in other essential bioelectric phenomena.

Action potentials, of course, are responsible for the external voltages that are measured all over the body: the electrocardiograph, the electroencephalograph, and the electromyography, for example.

There are two excellent sources for action-potential models, one a book article and the other a Web site. Varghese (180), in his article Membrane Models, (see above) reviewed a large number of models relating to action potentials (we counted 110 models). A listing will give an idea of the magnitude of the scope of this topic. In addition, it points out the fragmentation of the research in this area. Varghese's article nicely demonstrates the disparateness of the topic, as well as the lack of putting the models together into subsystems, much less systems.

Nerve Cells
Sensory Neurons
Efferent Neurons
Skeletal Muscle Cells
Endocrine Cells
Cardiac Cells
Epithelial Cells
Smooth Muscle
Plant Cells
Simplified Models

Under the heading 'Sensory Neurons', for example, are described.

Rabbit Sciatic Nerve Axons
Myelinated Auditory-Nerve Neuron
Retinal Ganglion Cells
Retinal Horizontal Cells,
Rat Nodose Neurons
Muscle Spindle Primary Endings
Vertebrate Retinal Cone Photoreceptors
Primary and Secondary Sensory Neurons of the Enteric Nervous System
Invertebrate Photoreceptor Neurons
Fly Optic Lobe Tangential Cells
Rat Mesencephalic Trigeminal Neurons
Primary Afferents and Related Efferents
Myelinated I Primary Afferent Neurons

Lloyd and the CellML site (176) have assembled an online repository of physiological system models, many of them related to action potentials. The models all conform with the CellML Specification (182). They are based on published mathematical models taken from peer-reviewed journals, from conference proceedings, and from textbook-defined metabolic pathways. The group has remained true to the original publications and has not assumed any reaction kinetics or initial values if they were not included in the original publication. All sources of information have been referenced in the model documentation.

These models represent several types of cellular processes, including models of electrophysiology, metabolism, signal transduction and mechanics. To facilitate the process of finding a particular model of interest, the models are grouped into broad subject categories.

The models on the site have been validated to a certain degree. Current validation processes include comparing the equations in the original paper with a PDF of the equations used in the CellML description. As tool development continues, both by the CellML team and by international collaborators, the groups expect to be able to carry a validation of the intent of the models by running simulations and comparing these results with those of the original publication. Presumably, validation of the model itself will ultimately take place.

With hundreds of models available on this site, it is impossible to go into detail. The topics include the following. Note that many of them deal with action potentials. GlucoSim (see above) is on this site, also: Signal Transduction Pathway Models, Metabolic Pathway Models, Cardiac Electrophysiological Models, Calcium Dynamics Models, Immunology Models, Cell Cycle Models, Simplified Electrophysiological Models, Other Cell Type Electrophysiological Models, Smooth and Skeletal Muscle Models, Mechanical Models and Constitutive Laws.

The following are examples of the formats in which a model may be obtained/downloaded.

1. The raw XML.
2. An HTML version for browsing online.
3. A PDF version suitable for printing.
4. A gzipped tarball with the XML and the documentation on the site.
5. A PDF of the equations described in the model generated directly from the CellML description using the MathML Renderer.

Thermal

Temperature control is one of the more impressive achievements in life. In higher forms of life, internal body temperature is maintained at a rather constant level, despite changes in the internal and external environment, to help provide conditions favorable to metabolism and other necessary processes within blood and tissue. Thus, in the human, an internal temperature close to 37 °C is normally maintained by a thermoregulatory feedback system despite large changes in ambient temperature and other environmental factors. It is also a daunting system to model, for many reasons.

Kuznetz (183) was one of the first to use more than 1D in thermal modeling. Previous models had failed to account for temperature distribution in any spatial direction other

than radially outward from the body centerline. The models therefore could not account for nonuniform environmental conditions or nonuniform heat generation from muscles or organs within the body. However, these nonuniform conditions are commonplace and can lead to disparate skin temperatures and heat loss rates on different sides of the same body compartment. Kuznetz' mathematical model of human thermoregulation could predict transient temperature variations in two spatial dimensions, both radially and angularly, as measured from the body centerline. The model thereby could account for nonuniform environments and internal heat generation rates.

Downey and Seagrave (184) developed a model of the human body that integrates the variables involved in temperature regulation and blood gas transport within the CV and respiratory systems. It describes the competition between skin and muscles when both require increased blood flows during exercise and/or heat stress. After a detailed study of the control relations used to predict skin blood flow, four other control relations used in the model were tweaked. Dehydration and complete water replacement were studied during similar environmental and exercise situations. Control relations for skin blood flow and evaporative heat loss were modified, and water balance was added to study how the loss of water through sweat can be limiting. Runoff from sweating as a function of relative humidity was introduced, along with evaporation.

In two papers, Fiala et al. (185,186) developed a dynamic model predicting human thermal responses in different environmental temperatures. The first paper was concerned with the passive system: (1) modeling the human body, (2) modeling heat-transport mechanisms within the body and at its periphery, and (3) the numerical procedure. Emphasis was given to a detailed modeling of heat exchange with the environment: local variations of surface convection, directional radiation exchange, evaporation and moisture collection at the skin, and the nonuniformity of clothing ensembles. Other thermal effects were also modeled: the impact of activity level on work efficacy and the change of the effective radiant body area with posture. From that passive model, they developed an active mathematical model for predicting human thermal and regulatory responses in cold, cool, neutral, warm, and hot environments (186). The active system simulates the regulatory responses of shivering, sweating, and peripheral vasomotion of unacclimatized subjects. The rate of change of the mean skin temperature, weighted by the skin temperature error signal, was identified as governing the dynamics of thermoregulatory processes in the cold. Good general agreement with measured data was obtained for regulatory responses, internal temperatures, and the mean and local skin temperatures of unacclimatized humans for a wide spectrum of climatic conditions and for different activity levels.

Boregawda (187) used a novel approach based on the second law of thermodynamics to investigate the psychophysiology of and to quantify human stress level. Two types of stresses (thermal and mental) were examined. Using his own thermal and psychological stress indices, he implemented a human thermal model based on a finite element method. With this model, he examined thermal, mental and CV stress in the human body.

Gardner and Martin (188) developed a model of the human thermoregulatory system for normal subjects and burned patients. The human body was split into eleven segments (The Rule of Nines suggests that one can estimate surface areas in the cranial, abdominal, thoracic and extremity regions by multiplying the patient's body surface area by 9%, or a multiple thereof.), each having core, muscle, fat and skin layers. Heat transport through blood flow and conduction were simulated, and surface heat loss was separated into radiative, convective and evaporative components. The model was refined to fit the data through manipulation of heat flow commands and temperature set points controlling sweating and shivering. The model also described the responses of burn patients to skin layer destruction, increased body metabolism and fluid loss. The model shows that the ambient temperature at which sweating occurs increases with the area of burn injury. It has been used to predict optimum environmental temperatures for treatment of patients with burn wounds of varying extent, a critical need.

Aging of Physiological Systems

Geriatric medicine is becoming increasingly important, due to the aging of our population. Physiologists need to understand that physiological function continually changes with age, and that they need to take aging into account in their studies and in their models. Healthcare givers need methods that can teach efficiently and painlessly the complexities involved with aging. Modeling and simulation are ideal for this area. We implemented the anatomy and physiology of aging in BODY Simulation (see PBPM). In the BODY model, we changed >50 patient parameters. The categories changed included anatomical, cardiovascular, and respiratory, as well as hepatic and renal function. Four patients were created: normal, but elderly, patients, aged 65, 75, 85, and 95 years. To evaluate the new patients, we imposed three stresses in the elderly patients and in a young, healthy patient: anesthesia induction, hemorrhage, and apnea. We observed an age-related response to these stresses. In general, we saw a reduced physiological reserve. Again, this model is available as an interactive simulation, no knowledge of mathematics required.

USES FOR MODELS

Please see a list of uses, in the Introduction. One use for models is to substitute as an animal or person in experiments, with the goal of reducing the use of either type of subject in experiments. A control system was developed (189), using three models (190–192). When a control system had been developed with the models, we performed one animal experiment to test the control system. Noting any problems or discrepancies, we returned to the model, tuned up the control system, and repeated an experiment, reiterating between model and animal several times. We estimated that the decrease in animal experiments was 82.5%, from a projected 120–21. Furthermore, the dangerous

experiments were performed on the models, and no animal was sacrificed. Noninvasive methods are less painful and dangerous in patients, and methods that render more accurate the data obtained from these methods are clinically and experimentally valuable. Kjaergaard et al. (193) used a model of pulmonary shunt and ventilationperfusion mismatch to help them quantify those two variables in patients, using noninvasive methods. The equations and diagrams can be found in the supplementary material to the paper, on the Web (click in the appropriate link, below the abstract).

Mesic et al. (194) used a mechanical lung simulator to simulate specific lung pathologies, to test lung-function equipment, and to instruct users about the equipment. A mathematical model of the respiratory system was interfaced with a model of physical equipment (the simulator, actuators, and the interface electronics) so that one can simulate the whole system. The control system, implemented on a personal computer, allows the user to set parameters.

Our noninvasive measurement theme continues, as we explore the work of Wesseling and his group. To the word noninvasive, we can add another important term, continuous, or continual. For areas where a patient's status can change in a matter of seconds (operating room, intensive care unit, and emergency room, e.g.), monitoring data should be available continuously or continually, and without delay. The work began with the Finapres, a device for noninvasively quantifying an arterial pressure waveform (195). They also developed a method for measuring continual cardiac output from an (invasive) aortic pressure waveform (196,197). Using a model, they could combine the two methods to achieve noninvasive, continual cardiac output from the radial artery (invasive) (198) or finger (noninvasive) (199). The pulse contour method has been extensively and carefully tested with aortic pressures, and it seemed prudent to use this pressure, if at all possible. To convert the less invasive radial pressure or the noninvasive finger pressure into an aortic pressure, Wesseling's group uses a three-element model of aortic input impedance that includes nonlinear aortic mechanical properties and a self-adapting systemic vascular resistance (200). Using the model enhances the accuracy of the less invasive methods (198,201). Another model increases the accuracy of the noninvasive technique even further (202). An inverse model of the averaged distortion models corrects for the pulse-wave distortion occurring between brachial and finger arteries.

Education should be one of the primary uses of physiological systems modeling. To adapt a complex whole-body model to a manikin-based simulator, however, requires considerable time and expense. As far as we can tell, two extant model- and manikin-based simulators exist. Another manikin-based simulator, which did not use a model, did not survive. Several disparate models have been published on the educational uses of one of the simulators (METI). The METI manikin-type simulator uses models based on the excellent models of Beneken, although how much, how detailed and to what extent is difficult to determine. The published models include some intracranial dynamics (203,204), pharmacokinetics and pharmacodynamics (205) and obstetric cardiovascular physiology (206). The PK-PD model is not a whole-body model, however, and the obstetric model omits the fetal circulation and hence is not realistic. BODY Simulation (see PBPM) uses a very detailed model for education, both clinical and in the basic sciences, and is a screen-based simulator. The best simulator from a purely educational point of view is GasMan, another screen-based simulator, developed by Jim Philip (207). As the name implies, the simulation deals only with inhaled anesthetic agents, but the manual and the teaching exercises are superb. The simulation is available for a Macintosh or a PC.

THE PHYSIOME PROJECT

What's next in physiological systems modeling? Hundreds of physiological models have been developed, some very small, some surprisingly large and complex. No one, however, is anywhere near describing human physiology in any reasonable completeness and detail. The physiome project may point the way. Essentially, the physiome project is the successor to the genome project, albeit many times more difficult. The following description of the project is adapted from Bassingthwaighte (208); the original was written in 2000. For MUCH more information, go to (209).

The physiome is the quantitative description of the functioning organism in normal and pathophysiological states. The human physiome can be regarded as the virtual human. Think genome, or proteome. The physiome is built upon the morphome, the quantitative description of anatomical structure, chemical and biochemical composition, and material properties of an intact organism, including its genome, proteome, cell, tissue, and organ structures up to those of the whole intact being. (We understand that the morphome, except for gross anatomy, is still in an early stage.) The physiome project is a multicenter integrated program to design, develop, implement, test and document, archive and disseminate quantitative information, and integrative models of the functional behavior of molecules, organelles, cells, tissues, organs, and intact organisms from bacteria to humans. A fundamental and major feature of the project is the databasing of experimental observations for retrieval and evaluation. Technologies that allow many groups towork together are rapidly being developed. Given a project that is so huge and complex, a particular working group can be expert in only a small part of the overall project. The strategies to be worked out must therefore include how to bring models composed of many submodules together, even when the expertise in each is scattered among diverse institutions, departments, talents and constituencies. Developing and implementing code for large-scale systems has many problems. Most of the submodules are complex, requiring consideration of spatial and temporal events and processes. Submodules have to be linked to one another in a way that preserves mass balance and gives an accurate representation of variables in nonlinear complex biochemical networks with many signaling and controlling pathways. Microcompartmentalization

vitiates the use of simplifiedmodel structures. The stiffness of the systems of equations is computationally costly. Faster computation is neededwhen using models as thinking tools and for iterative data analysis. Perhaps the most serious problem is the current lack of definitive information on kinetics and dynamics of systems, due in part to the almost total lack of databased observations, but also because, although we are nearly drowning in new information being published each day, either the information required for the modeling cannot be found, has never been obtained or is totally irrelevant. Simple things like tissue composition, material properties, and mechanical behavior of cells and tissues are not generally available.

Currently, there are several subprojects, the magnitude of each of which boggles the mind. As the *Economist* put it, Computer organs are not for the technologically faint-hearted (210). The subprojects include the Cardiome Project (211), the Microcirculatory Project (212), the Pulmonary/Respiratory Project (213,214), The Kidney Project (215), and the Coagulation Project (210).

How rapidly and completely can all of this take place? This is an enormously ambitious project, many times more difficult than the genome project. The genome project succeeded relatively quickly for at least two reasons. First, to fill well-defined gaps in knowledge, a huge number of devices automatically churned out mountains of carefully and precisely specified data. Second, industry, sensing that there was gold in the genome, invested (and is still investing) enormous sums of money.

We have several questions and comments regarding the physiome project, questions that might apply to any very large project. The questions are restricted mainly to normal physiology. Disease and normal aging will add to the number and difficulty of the questions. These questions are intended to help sort out the potential problems, difficulties and considerations while the project is still at an early stage.

1. The questions in the physiome project are much more difficult to formulate than those of the Genome Project.
   (a) Should these questions be formulated formally?
   (b) If so, who will formulate them?
2. Can or should the physiome project control the data, including the content and structure, back to and including the planning of the experiment that generates the data?
3. Who is going to determine whether the whole-body model works? We worry that it may become like some software (no single person knows every detail about it) what it can do and what it cannot do.
4. How does one handle the flood of new information?
   (a) How often will a large, multicenter submodel be updated: and who decides?
   (b) Do new methods for handling the flood of information need to be developed?
   (c) If so, what?
   (d) What are the methods for validating models and for sensitivity analysis?
   (e) How will validation and sensitivity analysis proceed?
   (f) At what stages should validation and sensitivity analysis proceed, for example, every time models are combined? Every 6 months?
5. There will be an enormous library of physiological normals: equations, parameters, and so on.
   (a) How will one deal with variations from normal? Presumably this will be partly with sensitivity analysis.
   (b) How much variation from normal will be allowed? In other words, what are the limits on normal: upper and lower?
   (c) Does one anticipate large variations in normal, from parameter to parameter and from equation to equation?
   (d) Will clinicians be in on this and other decision processes?
6. How will one deal with the effects of aging and responses to the environment, including exercise, hypoxia, altitude, and temperature?
7. The genome project is relatively simple, and mistakes can be relatively easily corrected.
   (a) Is this the case with the physiome project?
   (b) How will one deal with errors in the physiome project?
   (c) If there is one error in one submodel, how will one find that error?
   (d) How will one determine the effect of errors in a submodel that is a component of the overall model on that larger model: by sensitivity analysis?
8. Can anyone input data, equations, formulae and models into the project database?
   (a) If so, who "edits" the deposits into databases, or are they done just ad hoc?
   (b) If not, is there a gatekeeper?
   (c) Asked another way, what is the quality control and who is in charge of it?
   (d) Who is going to determine what finally winds up in the library/database?
   (e) Who will be in charge of the information/data?
9. Several computer languages are currently mentioned. Will there ultimately be a single language?
10. Is there a freely available Web journal for the project? (The Virtual Journal of the Virtual Human)
    (a) Perhaps the VJVH could be the filter for the information and data, as well as a means for developing standards (see next question).
11. Standards, if done properly, should help in several areas. Some standards are being developed (216).

(a) We gather that part of the significant difference among models arises from their being published in different journals.
(b) Do common standards exist for the project?
(c) Who is responsible for implementing and coordinating generic and specific standards?

12. A very important set of questions involves intellectual philanthropy versus intellectual property.
(a) Are there any tools, including developmental tools, that are freely available, that is, public domain?
(b) The NLM has made the Visible Human data public domain. How much of the physiome database will be public domain and how much private domain?
(c) Would dividing the database into multiple public and private sections inhibit the integrity, the wholeness and the usefulness of the database?
(d) How will one audit, in one huge model, the amalgamated data from government employees, universities, research institutes and industry?
(e) How will one determine credit, cash, brownie points, or whatever for data, equations, models, and so on?
(f) The same questions hold for developmental tools.
13. How will ongoing projects fit into the project? How will future projects fit in?
14. What data are required, for each step?
(a) Who decides what data are required, or do data just appear in the database because they were there?

Excellent reviews on the Physiome Project include (216–219). In addition try an IUPS Web site (220).

EPILOGUE

Over one-half of a century ago, two papers were published in the same year, in the same country. These two papers point the way to the future in physiological systems modeling, given all the enormous amount of models and data pouring in. The authors of one paper went on to win the Nobel Prize. The author of the other committed suicide within two years. Neither knew of the other's existence, and one paper probably went unnoticed, and was certainly misunderstood.

If nothing else, our article has illustrated the need to integrate mathematical biology with experimental biology. To achieve this integration, experimentalists must learn the language of mathematics and dynamical modeling and theorists must learn the language of biology. Hodgkin and Huxley's quantitative model of the nerve action potential and Alan Turing's work on pattern formation in activator–inhibitor systems (221) represent those two languages. These classic studies illustrate two ends of the spectrum in mathematical biology: the detailed-model approach and the minimal-model approach. When combined, they are highly synergistic in analyzing the mechanisms underlying the behavior of complex biological systems. Their effective integration will be essential for unraveling the physical basis of the mysteries of life. For more detail on this important concept, see the fascinating account by Weiss et al. (222).

ACKNOWLEDGMENTS

The authors are grateful to the American Physiological Society and to the American Society for Biological Chemistry and Molecular Biology for making available PDF files of all issues of their journals. The 1924 Haggard articles were required reading for the senior author, during his residency. Classic and macro physiology, much of which is to be found in the older literature have been emphasized.

The senior author is also grateful to all the mentors in his modeling life. There are so many that they can only be listed and thanked, in more or less chronological order. Avram Goldstein, Aldo Corbascio, Bram Noordergraaf, Isaac Starr, Bob Fleming, Jan Beneken, Karel Wesseling, The late Vince Rideout, Yas Fukui, James Martin, Ken Starko. The senior author is an anesthesiologist, and the bias occasionally shows.

GLOSSARY

CVS	Cardiovascular system
CNS	Central nervous system
ANS	Autonomic nervous system
MAP	Mean arterial pressure
CVP	Central venous pressure
CO	Cardiac output
CBF	Cerebral blood flow
CSF	Cerebrospinal fluid
ICP	Intracranial pressure
LV	Left ventricle
LA	Left atrium
RV	Right ventricle
PV	Pressure volume
SVR	Systemic vascular resistance
avu	Arteriovenous
AV	Atrioventricular
PO_2	Partial pressure of oxygen
PKPD	Pharmacokinetic/pharmacodynamic
Physiological PKPD	physiologically based pharmacokinetic/pharmacodynamic (PBPKPD!)
PBPM	Physiologically based pharmacological models

Some systems are so familiar that an abbreviation suffices: CVS for cardiovascular system and CNS for central nervous system. A glossary of abbreviations follows for your reference.

BIBLIOGRAPHY

1. Elenkov IJ, et al. The sympathetic nerve—an integrative interface between two supersystems: The brain and the immune system. Pharmacol Rev 2000;52:595–638.
2. Bassingthwaighte JB. A view of the physiome. Available at http://physiome.org/files/Petrodvoret.1997/abstracts/jbb.html. 1997.
3. Weinstein AM. Mathematical models of renal fluid and electrolyte transport: Acknowledging our uncertainty. Am J Physiol Renal Physiol 2003;284(5):F871–F884.
4. Werner J, Böhringer D, Hexamer M. Simulation and prediction of cardiotherapeutical phenomena from a pulsatile model coupled to the Guyton circulatory model. IEEE Trans Biomed Eng 2002;49(5):430–439.
5. Rideout VC. Mathematical and computer modeling of physiological systems. 1st ed. Biophysics and bioengineering. Noordergraaf A, editor. Englewood Cliffs (NJ): Prentice-Hall; 1991. p 261.
6. Baan J, Noordergraaf A, Raines J. Cardiovascular system dynamics. Cambridge (MA): The MIT Press; 1978. p 618.
7. Papper EM, Kitz RJ. editors. Uptake and distribution of anesthetic agents. McGraw-Hill: New York; 1963. p 321.
8. Bird R, Stewart W, Lightfoot E. Transport phenomena. New York: Wiley; 1960.
9. Fukui Y. A study of the human cardiovascular-respiratory system using hybrid computer modeling. in Department of Engineering. University of Wisconsin: Madison (WI); 1971.
10. Guyton AC, Polizo D, Armstrong GG. Mean circulatory filling pressure measured immediately after cessation of heart pumping. Am J Physiol 1954;179:262–267.
11. Guyton A, Lindsey A, Kaufmann B. Effect of mean circulatory filling pressure and other peripheral circulatory factors on cardiac output. Am J Physiol 1955;180:463–468.
12. Manning RD Jr, et al. Essential role of mean circulatory filling pressure in salt-induced hypertension. Am J Physiol Regulatory Integrative Comp Physiol 1979;5:R40–R47.
13. Guyton A, et al. Systems analysis of arterial pressure regulation and hypertension. Ann Biomed Eng 1972;1(2):254–281.
14. Werner J, Böhringer D, Hexamer M. Simulation and prediction of cardiotherapeutical phenomena from a pulsatile model coupled to the Guyton circulatory model. IEEE Trans Biomed Eng 2002 May;49(5):430–439.
15. Hardman J, Wills J, Aitkenhead A. Investigating hypoxemia during apnea: Validation of a set of physiological models. Anesth Analg 2000;90:614–618.
16. Hardman J, Wills J, Aitkenhead A. Factors determining the onset and course of hypoxemia during apnea: An investigation using physiological modelling. Anesth Analg 2000;90:619–624.
17. Hardman J, Bedforth N. Estimating venous admixture using a physiological simulator. Br J Anaesth 1999;82:346–349.
18. Hardman J, et al. A physiology simulator: Validation of its respiratory components and its ability to predict the patient's response to changes in mechanical ventilation. Br J Anaesth 1998;81:327–332.
19. Snow J. On narcotism by the inhalation of vapours: Part iv. Lond Med Gaz 1848;7:330–334.
20. Snow J. On narcotism by the inhalation of vapours: Part xv. Lond Med Gaz 1850;11:749–754.
21. Frantz R. Ueber das verhalten des aethers im thierischen organismus (cited by Kunkle, a. J., Handbuch der toxikologie, Jena, 1899, I, 434). 1895: Wurzburg.
22. Nicloux M. Les anethesiques generaux au point de vue chemicophysiologique. Bull Acad Med 1908;lx:297.
23. Haggard HW. The absorption, distribution, and elimination of ethyl ether, iii. The relation of the concentration of ether, or any similar volatile substance, in the central nervous system to the concentration in the arterial blood, and, the buffer action of the body. J Biol Chem 1924;59(3):771–781.
24. Severinghaus J. Role of lung factors. In: Papper EM, Kitz RJ, editors. Uptake and distribution of anesthetic agents. 1963. pp. 59–71.
25. Eger EI,Ii. A mathematical model of uptake and distribution. In: Papper EM, Kitz RJ, editors. Uptake and distribution of anesthetic agents. New York: McGraw-Hill 1963. pp. 72–87.
26. Price H. A dynamic concept of the distribution of thiopental in the human body. Anesthesiology 1960;21(1):40–45.
27. Rackow H, et al. Simultaneous uptake of N_2O and cyclopropane in man as a test of compartment model. J Appl Physiol 1965;20:611–620.
28. Ashman M, Blesser W, Epstein R. A nonlinear model for the uptake and distribution of halothane in man. Anesthesiology 1970;33:419–429.
29. Zwart A, Smith NT, Beneken JEW. Multiple model approach to uptake and distribution of halothane. Use of an analog computer. Comp Biol Med Res 1972;55:228–238.
30. Smith NT, Zwart A, Beneken JEW. An analog computer multiple model of the uptake and distribution of halothane. Proc San Diego Biomed Symp 1972;11:235–241.
31. Smith NT, Zwart A, Beneken JEW. Effects of halothane-induced changes in skin, muscle, and renal blood flows on the uptake and distribution of halothane. Use of a multiple model. Proc 5th World Cong Anaesth 1972.
32. Fukui Y, Smith NT. A hybrid computer multiple model for the uptake and distribution of halothane. I. The basic model. Proc San Diego Biomed Symp 1974.
33. Fukui Y, Smith NT. A hybrid computer multiple model for the uptake and distribution of halothane. Ii. Spontaneous vs. Controlled ventilation, and the effects of CO_2. Proc San Diego Biomed Symp 1974.
34. Fukui Y, Smith NT. Interaction among ventilation, the circulation, and the uptake and distribution of halothane. Use of a hybrid computer model I. The basic model. Anesthesiology 1981;54:107–118.
35. Fukui Y, Smith NT. Interaction among ventilation, the circulation, and the uptake and distribution of halothane. Use of a hybrid computer model ii. Spontaneous vs. Controlled ventilation, and the effects of CO_2. Anesthesiology 1981;54:119–124.
36. Schwid HA, Wakeland C, Smith NT. A simulator for general anesthesia. Anesthesiology 1986;65:A475.
37. Smith NT, Starko K. Body simulation enhancements, including chemicalreaction simulation of cyanide therapy. Anaes Analg 2004;98(5s):S38.
38. Smith NT, Starko K. Physiologic and chemical simulation of cyanide and sarin toxicity and therapy. Stud Health Technol Inform 2005;111:492–497.
39. Smith NT, Starko K. The physiology and pharmacology of growing old, as shown in body simulation. Medicine Meets Virtual Reality. The Magical Next Becomes the Medical Now. Long Beach (CA): IOS Press; 2005.
40. Smith NT, Starko K. http://www.advsim.com/biomedical/body_manual/index.htm.
41. Oliver RE, Jones AF, Rowland MA. whole-body physiologically based pharmacokinetic model incorporating dispersion concepts: Short and long time characteristics. J Pharmacokin Pharmacodyna 2001;28(1):27–55.
42. Levitt DG, Quest PK. A general physiologically based pharmacokinetic model. Introduction and application to propranolol. BMC Clin Pharmacol 2002;2(5).
43. Levitt DG, Quest PK. Measurement of intestinal absorption and first pass metabolism—application to human ethanol pharmacokinetics. BMC Clin Pharmacol 2002;2(4).
44. Levitt DG, Quest PK. Volatile solutes—application to enflurane, nitrous oxide, halothane, methoxyflurane

and toluene pharmacokinetics. BMC Anesthesiology 2002;2 (5).
45. Levitt DG: http:/www.pkquest.com/.
46. Shafer S, Stanski DR. Stanford PK/PD software server. 2005.
47. Ursino M, Magosso E. Acute cardiovascular response to isocapnic hypoxia. I. A mathematical model. Am J Physiol Heart Circ Physiol 2000;279:H149–H165.
48. Groebe K. Precapillary servo control of blood pressure and postcapillary adjustment of flow to tissue metabolic status. A new paradigm for local perfusion regulation. Circulation 1996;94:1876–1885.
49. Krejci V, et al. Continuous measurements of microcirculatory blood flow in gastrointestinal organs during acute haemorrhage. Br J Anaesth 2000;84:468–475.
50. Hart BJ, et al. Right ventricular oxygen supply/demand balance in exercising dogs. Am J Physiol Heart Circ Physiol 2001;281:H823–H830.
51. Paneraiy RB. Assessment of cerebral pressure autoregulation in humans—a review of measurement methods. Physiol Meas 1998;19:305–338.
52. Lu K, et al. Cerebral autoregulation and gas exchange studied using a human cardiopulmonary model. Am J Physiol Heart Circ Physiol 2004;286:H584–H601.
53. Panerai RB, Dawson SL, Potter JF. Linear and nonlinear analysis of human dynamic cerebral autoregulation. Am J Physiol 1999;277:H1089–H1099.
54. Gao E, et al. Mathematical considerations for modeling cerebral blood flow autoregulation to systemic arterial pressure. Am J Physiol 1998;274:H1023–H1031.
55. Zhang R, et al. Transfer function analysis of dynamic cerebral autoregulation in humans. Am J Physiol 1998;274:H233–H241.
56. Hughson RL, et al. Critical analysis of cerebrovascular autoregulation during repeated head-up tilt. Stroke 2001;32:2403–2408.
57. Czosnyka M, et al. Contribution of mathematical modelling to the interpretation of bedside tests of cerebrovascular autoregulation. J Neurol Neurosurg Psychiat 1997;63:721–731.
58. Ursino M, Lodi CA. A simple mathematical model of the interaction between intracranial pressure and cerebral hemodynamics. J Appl Physiol 1997;82(4):1256–1269.
59. Hyder F, Shulman RG, Rothman DL. A model for the regulation of cerebral oxygen delivery. J Appl Physio 1998;85:554–564.
60. Kirkham SK, Craine RE, Birch AA. A new mathematical model of dynamic cerebral autoregulation based on a flow dependent feedback mechanism. Physiol Meas 2001;22:461–473.
61. Cornelissen AJM, et al. Myogenic reactivity and resistance distribution in the coronary arterial tree: A model study. Am J Physiol Heart Circ Physiol 2000;278:H1490–H1499.
62. Broten TP, Feigl E. Role of myocardial oxygen and carbon dioxide in coronary autoregulation. Am J Physiol 1992; 262:Hl231–H1237.
63. Cornelissen AJM, et al. Balance between myogenic, flow-dependent, and metabolic flow control in the coronary arterial tree: A model study. Am J Physiol Heart Circ Physiol 2002;282:H2224–H2237.
64. Vergroesen I, et al. Quantification of O_2 consumption and arterial pressure as determinants of coronary flow. Am J Physiol 1987; 252.
65. Guiota C, et al. Model-based assessment of pressure and flow-dependent coronary responses following abrupt pressure drops. Computers Biol Med 2000;30:111–126.
66. Jayaweera AR, et al. Role of capillaries in determining CBF reserve: New insights using myocardial contrast echocardiography. Am J Physiol 1999;277:H2363–H2372.
67. Geven MCF, et al. A physiologically representative in vitro model of the coronary circulation. Physiol Meas 2004;25:891–904.
68. Beyar R, Sideman S. Time-dependent coronary blood flow distribution in the left ventricular wall. Am J Physiol 1987; 252.
69. Pollack GH, Reddy RV, Noordergraaf A. Input impedance wave travel and reflections in the pulmonary arterial tree: Studies using an electrical analog. IEEE Trans Biomed Eng 1968;15:151–164.
70. Jager GN, Westerhof N, Noordergraaf A. Oscillatory flow impedance in electrical analog of arterial system. Circ Res 1965;16:121–133.
71. Beneken JEW, Beneken JEW. Some computer models in cardiovascular research. In: Bergel H, editor. Cardiovascular fluid dynamics. London: Academic Press; 1972. p 173–223.
72. Beneken JEW, DeWit B. A physical approach to hemodynamic aspects of the human cardiovascular system. In: Reeve E, Guyton A, editors. Physical. Basis of circulatory transport. Philadelphia: Saunders; 1967.
73. Rothe CF, Gersting JM. Cardiovascular interactions: An interactive tutorial and mathematical model. Adv Physiol Educ 2002;26:98–109.
74. Chung DC, et al. A dynamic model of ventricular interaction and pericardial influence. Am J Physiol 1997;272:H2942–H2962.
75. Lu K, et al. A human cardiopulmonary system model applied to the analysis of the Valsalva maneuver. Am J Physiol Heart Circ Physiol 2001;281:H2661–H2679.
76. Smith NT, Starko K. Anesthesia circuit. In: Atlee J, editor. Complications in anesthesiology. 1998.
77. Moore JA, et al. Accuracy of computational hemodynamics in complex arterial geometries reconstructed from magnetic resonance imaging. Ann Biomed Eng 1999;27(1):32–41.
78. Huang W, et al. Comparison of theory and experiment in pulsatile flow in cat lung. Ann Biomed Eng 1998;26:812–820.
79. De Gaetano A, Cremona G. Direct estimation of general pulmonary vascular models from occlusion experiments. Cardiovas Eng 2004;4(1).
80. Karamanoglu M, et al. Functional origin of reflected pressure waves in a multibranched model of the human arterial system. Am J Physiol 1994;267:H1681–H1688.
81. Karamanoglu M, Feneley MP. Late systolic pressure augmentation: Role of left ventricular out- flow patterns. Am J Physiol 1999;277:H481–H487.
82. Sun Y, et al. Mathematical model that characterizes transmitral and pulmonary venous flow velocity patterns. Am J Physiol 1995;268:H476–H489.
83. Yellin EL, et al. Mechanisms of mitral valve motion during diastole. Am J Physiol 1981;241:H389–H400.
84. Thomas JD, et al. Physical and physiological determinants of transmitral velocity: Numerical analysis. Am J Physiol 1991;260:H1718–H1730.
85. Salem JE, et al. Mechanistic model of myocardial energy metabolism under normal and ischemic conditions. Ann Biomed Eng 2002;30:202–216.
86. Olufsen M, et al. Numerical simulation and experimental validation of blood flow in arteries with structured-tree outflow conditions. Ann Biomed Eng 2000;28(28):1281–1299.
87. Wootton DM, Ku DN. Fluid mechanics of vascular systems, dis- eases, and thrombosis. Annu Rev Biomed Eng 1999;1:299–329.
88. Vasquez EC, et al. Neural reflex regulation of arterial pressure in pathophysiological conditions: Interplay among the baroreflex, the cardiopulmonary reflexes and the chemoreflex. Braz J Med Biol Res 1997;30:521–532.
89. Marshall JM. Peripheral chemoreceptors and cardiovascular regulation. Physiol Rev 1994;74:543–594.
90. Lanfranchi PA, Somers VK. Arterial baroreflex function and cardiovascular variability: Interactions and implications. Am J Physiol Regul Integr Comp Physiol 2002;283:R815–R826.
91. Wesseling KH, et al. Baromodulation as the cause of short term blood pressure variability? International Conference on

Applications of Physics to Medicine and Biology. Trieste, Italy: World Scientific Publishing Co; 1982.
92. Wesseling KH, Settels J. Baromodulation explains short-term blood pressure variability. In: Orlebeke J, Mulder G, Van Doornen L, editors. Psychophysiology of cardiovascular control: Models, methods and data. New York: Plenum Press; 1983. pp. 69–97.
93. Settels J, Wesseling KH. Explanation of short-term blood pressure responses needs baromodulation. Ann Int Conf IEEE Eng Med Biol Soc 1990;12(2):696–697.
94. Rose WC, Schwaber JS. Analysis of heart rate-based control of arterial blood pressure. Am J Physiol 1996;271:H812–H822.
95. Ursino M, Magosso E. Acute cardiovascular response to isocapnic hypoxia. I. A mathematical model. Am J Physiol Heart Circ Physiol 2000;279:H149–H165.
96. Ursino M, Magosso E. Acute cardiovascular response to isocapnic hypoxia. II. Model validation. Am J Physiol Heart Circ Physiol 2000;279:H166–H175.
97. Magosso E, Ursino M. A mathematical model of CO_2 effect on cardiovascular regulation. Am J Physiol Heart Circ Physiol 2001;281:H2036–H2052.
98. Melchior FM, Srinivasan RS, Charles JB. Mathematical modeling of human cardiovascular system for simulation of orthostatic response. Am J Physiol 1992;262:H1920–H1933.
99. Lerma C, et al. A mathematical analysis for the cardiovascular control adaptations in chronic renal failure. Art Org 2004;28(4):398–409.
100. Ursino M, Magosso E. A theoretical analysis of the carotid body chemoreceptor response to O_2 and CO_2 pressure changes. Resp. Physiol Neurobiol 2002;130:99–110.
101. Ursino M, Antonucci M, Belardinelli E. Role of active changes in venous capacity by the carotid baroreflex: Analysis with a mathematical model. Am J Physiol 1994;267:H253l–H2546.
102. Cohen MA, Taylor JA. Short-term cardiovascular oscillations in man: Measuring and modelling the physiologies. J. Physiol 2002;542(3):669–683.
103. Seydnejad SR, Kitney RI. Modeling of Mayer waves generation mechanisms determining the origin of the low- and very low frequency components of BPV and HRV. IEEE Eng Med Biol 2001;20:92–100.
104. Cavalcanti S, Belardinelli E. Modeling of cardiovascular variability using a differential delay equation. IEEE Trans Biomed Eng 1996;43(10):982–989.
105. Magosso E, Biavati V, Ursino M. Role of the baroreflex in cardiovascular instability: A modeling study. Cardiovas Eng 2001;1(2):101–115.
106. Aljuri N, Cohen RJ. Theoretical considerations in the dynamic closed-loop baroreflex and autoregulatory control of total peripheral resistance. Am J Physiol Heart Circ Physiol 2004;287(5):H2252–H2273.
107. Ben-Haim SA, et al. Periodicities of cardiac mechanics. Am J Physiol 1991;261:H424–H433.
108. Ursino M. Interaction between carotid baroregulation and the pulsating heart: A mathematical model. Am J Physiol 1998;275:H1733–H1747.
109. Hughson RL, et al. Searching for the vascular component of the arterial baroreflex. Cardiovasc Eng 2004;4:155–162.
110. O'Leary DD, et al. Spontaneous beat-by-beat fluctuations of total peripheral and cerebrovascular resistance in response to tilt. Am J Physiol Regul Integr Comp Physiol 2004;287(3):R670–R679.
111. Toska K, Eriksen M, Walloe L. Short-term control of cardiovascular function: Estimation of control parameters in healthy humans. Am J Physiol 1996;270:H651–H660.
112. Toska K, Eriksen M, Walloe L. Short-term cardiovascular responses to a step decrease in peripheral conductance in humans. Am J Physiol 1994;266:H199–H211.
113. Lee J-S. 1998 distinguished lecture: Biomechanics of the microcirculation, an integrative and therapeutic perspective. Ann Biomed Eng 2000;28:1–13.
114. Carr RT, Lacoin M. Nonlinear dynamics of microvascular blood flow. Ann Biomed Eng 2000;28:641–652.
115. Schmid-Schonbein GW. Biomechanics of microcirculatory blood perfusion. Annu Rev Biomed Eng 1999;1:73–102.
116. Pries AR, Secomb TW. Microcirculatory network structures and models. Ann Biomed Eng 2000;28:916–921.
117. Beard DA, Bassingthwaighte JB. Modeling advection and diffusion of oxygen in complex vascular networks. Ann Biomed Eng 2001;29:298–310.
118. Clark ME, Kufahl RH. Simulation of the cerebral macrocirculation. In: Baan J, editors. Cardiovascular system dynamics. Cambridge (MA): The MIT Press; 1978. pp. 380–390.
119. Lakin WD, et al. A whole-body mathematical model for intracranial pressure dynamics. J Math Biol 2003;46:347–383.
120. Olufsen M, Tran H, Ottesen J. Modeling cerebral blood flow control during posture change from sitting to standing. Cardiov Eng: An Int J 2004;4(1).
121. Sato J, et al. Differences in the dynamic cerebrovascular response between stepwise up tilt and down tilt in humans. Am J Physiol Heart Circ Physiol 2001;281:H774–H783.
122. Zhu L. Theoretical evaluation of contributions of heat conduction and countercurrent heat exchange in selective brain cooling in humans. Ann Biomed Eng 2000;28:269–277.
123. Ursino M, Di Giammarco P, Belardinelli E. A mathematical model of cerebral blood flow chemical regulation—part I: Diffusion processes. IEEE Trans Biomed Eng 1989;36(2):183—191.
124. Ursino M, DiGiammarco P, Belardinelli E. A mathematical model of cerebral blood flow chemical regulation—part I: Diffusion processes. IEEE Trans Biomed Eng 1989;36(2):1922–2201.
125. Ursino M, Magosso E. Role of tissue hypoxia in cerebrovascular regulation: A mathematical modeling study. Ann Biomed Eng 2001;29:563–574.
126. Wakeland W, et al. Modeling intracranial fluid flows and volumes during traumatic brain injury to better understand pressure dynamics. IEEE Eng Med Biol 2003;23:402–405.
127. Lodi CA, Ursino M. Hemodynamic effect of cerebral vasospasm in humans: A modeling study. Ann Biomed Eng 1999;27:257–273.
128. Pasley RL, Leffler CW, Daley ML. Modeling modulation of intracranial pressure by variation of cerebral venous resistance induced by ventilation. Ann Biomed Eng 2003;31:1238–1245.
129. Ursino M, Lodi CA. A simple mathematical model of the interaction between intracranial pressure and cerebral hemodynamics. J Appl Physiol 1997;82(4):1256–1269.
130. Ursino M. A mathematical study of human intracranial hydrodynamics. Part 1–the cerebrospinal fluid pulse pressure. Ann Biomed Eng 1988;16(4):379–401.
131. Ursino M. A mathematical study of human intracranial hydrodynamics. Part 2–simulation of clinical tests. Ann Biomed Eng 1988;16(4):403–416.
132. Ursino M, Lodi CA. Interaction among autoregulation, CO_2 reactivity, and intracranial pressure: A mathematical model. Am J Physiol 1988;274:H1715–H1728.
133. Loewe S, et al. Modeling cerebral autoregulation and CO_2 reactivity in patients with severe head injury. Am J Physiol 1998;274:H1729–H1741.
134. Ursino M, et al. Cerebral hemodynamics during arterial and CO_2 pressure changes: In vivo prediction by a mathematical model. Am J Physiol Heart Circ Physiol 2000;279:H2439–H2455.
135. Ursino M, Iezzi M, Stochetti N. Intracranial pressure dynamics in patients with acute brain damage: A critical

analysis with the aid of a mathematical model. IEEE Trans Biomed Eng 1995;42(6):529–540.
136. Sharan M, et al. An analysis of hypoxia in sheep brain using a mathematical model. Ann Biomed Eng 1998;26:48–59.
137. Cammarota JP Jr , Onaral B. State transitions in physiologic systems: A complexity model for loss of consciousness. IEEE Trans Biomed Eng 1998;45(8): 1017–1023.
138. Smith NP. A computational study of the interaction between coronary blood flow and myocardial mechanics. Physiol Meas 2004;25:863–877.
139. Smith NP, Pullan AJ, Hunter PJ. Generation of an anatomically based geometric coronary model. Ann Biomed Eng 2000;28:14–25.
140. Grotberg JB. Respiratory fluid mechanics and transport processes. Annu Rev Biomed Eng 2001;3:421–457.
141. Defares JG, Derksen HE, Duyff JW. Cerebral blood flow in the regulation of respiration. Acta Physiol Pharmacol Neerl 1960;9:327–360.
142. Defares JG, Principles of feedback control and their application to the respiratory control system. Handbook of physiology. Respiration. Washington (DC): American Physiology Society. 1964. pp. 649–680.
143. Grodins FS, Buell J, Bart AJ. Mathematical analysis and digital simulation of the respiratory control system. J Appl Physiol 1967;22(2):260–276.
144. Chiari L, Avanzolini G, Ursino M. A comprehensive simulator of the human respiratory system: Validation with experimental and simulated data. Ann Biomed Eng 1997; 25:985–999.
145. Hyuan U, Suki B, Lutchen KR. Sensitivity analysis for evaluating nonlinear models of lung mechanics. Ann Biomed Eng 1998;26:230–241.
146. Avanzolini G, et al. Role of the mechanical properties of tracheobronchial airways in determining the respiratory resistance time course. Ann Biomed Eng 2001;29:575–586.
147. Ursino M, Magosso E, Avanzolini G. An integrated model of the human ventilatory control system: The response to hypercapnia. Clin Physiol 2001;21(4):447–464.
148. Grodins FS, Yamashiro SM. Respiratory function of the lung and its control. New York: Macmillan; 1978.
149. Batzel JJ. Modeling and stability analysis of the human respiratory control system, in Department of Mathematics. Raleigh (NC): North Carolina State University; 1998. p 195.
150. Khoo MC, Gottschalk A, Pack AI. Sleep-induced periodic breathing and apnea: A theoretical study. J Appl Physiol 1991;70(5):2014–2024.
151. Kamm RD. Airway wall mechanics. Annu Rev Biomed Eng 1999;1:47–72.
152. Tehrani FT. Mathematical analysis and computer simulation of the respiratory system in the newborn infant. IEEE Trans Biomed Eng 1993;40:475–481.
153. Joyce CJ, Williams AB. Kinetics of absorption atelectasis during anesthesia: A mathematical model. J Appl Physiol 1999;86:1116–1125.
154. Dempsey JA, Forster HV. Mediation of ventilatory adaptations. Physiol Rev 1982;62(1):262–346.
155. Schafer JA. Interaction of modeling and experimental approaches to understanding renal salt and water balance. Ann Biomed Eng 2000;28:1002–1009.
156. Russell JM. Sodium-potassium-chloride cotransport. Physiol Rev 2000;80(1):211–276.
157. Dibona GF. Neural control of the kidney: Functionally specific renal sympathetic nerve fibers. Am J Physiol Reg Integ Comp Physiol 2000;279:R1517–R1524.
158. Kellen MR, Bassingthwaighte JB. An integrative model of coupled water and solute exchange in the heart. Am J Physiol 2003;285:H1303–H1316.
159. Makroglou A, Li J, Kuang Y. Mathematical models and software tools for the glucose-insulin regulatory system and diabetes: An overview. 2005.
160. Tibell A, Binder C, Madsbad S. Insulin secretion rates estimated by two mathematical methods in pancreas-kidney transplant recipients. Am J Physiol 1998;274:E716–E725.
161. Ibbini MS, Masadeh MA, Bani MM. Amer, A semiclosed-loop optimal control system for blood glucose level in diabetics. J Med Eng Technol 2004;28(5):189–196.
162. Parker RS, Doyle I, Francis J, Peppas NA. A model-based algorithm for blood glucose control in type 1 diabetic patients. IEEE Trans Biomed Eng 1999;46(2).
163. Bequette B. A critical assessment of algorithms and challenges in the development of a closed-loop artificial pancreas. Diabetes Technol Ther 2005; Feb; 7(1):28–47, 2005;7(1):28–47.
164. Berman N, et al. A mathematical model of oscillatory insulin secretion. Am J Physiol 1993;264:R839–R851.
165. Straub SG, Sharp GWG. Hypothesis: One rate-limiting step controls the magnitude of both phases of glucose-stimulated insulin secretion. Am J Physiol Cell Physiol 2004;287:C565–C571.
166. Lenbury Y, Ruktamatakul S, Amornsamarnkul S. Modeling insulin kinetics: Responses to a single oral glucose administration or ambulatory-fed conditions. BioSystems 2001; 59:15–25.
167. Chay TR, Keizer J. Minimal model for membrane oscillations in the pancreatic beta-cell. Biophys J 1983;42:181–190.
168. Sherman A, Rinzel J, Keizer J. Emergence of organized bursting in clusters of pancreatic beta-cells by channel sharing. Biophys J 1988;54:411–425.
169. Keizer J. Electrical activity and insulin release in pancreatic beta cells. Math Biosci 1988;90:127–138.
170. Keizer J, Magnus G. ATP-sensitive potassium channel and bursting in the pancreatic beta cell. A theoretical study Biophys J 1989;89:229–242.
171. Sherman A, Keizer J, Rinzel J. Domain model for Ca^{2+}-inactivation of Ca^{2+} channels at low channel density. Biophys J 1990;56:985–995.
172. Keizer J, Young GWD. Effect of voltage-gated plasma membrane Ca^{2+} fluxes on ip3-linked Ca^{2+} oscillations. Cell Calcium 1993;14:397–410.
173. Sherman A. Contributions of modeling to understanding stimulus-secretion coupling in pancreatic beta-cells. Am J Physiol 1996;271:E362–E372.
174. Tolic I, Mosekilde E, Sturis J. Modeling the insulin-glucose feedback system: The significance of pulsatile insulin secretion. J Theor Biol 2000;207:361–375.
175. Tolic I, Mosekilde E, Sturis J. Modelling the insulin-glucose feedback system. http://dev.cellml.org/examples/repository/.
176. Nickerson D, Lloyd CM. http://www.cellml.org/examples/repository/.
177. Shannahoff-Khalsa DS, et al. Low-frequency ultradian insulin rhythms are coupled to cardiovascular, autonomic, and neuroendocrine rhythms. Am J Physiol 1997;272:R962–R968.
178. Erzen FC, et al. GlucoSim: A web-based educational simulation package for glucose-insulin levels in the human body. Available at http://216.47.139.198/glucosim/index.html. 2005.
179. Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol (London) 1952;117:500–544.
180. Varghese A. Membrane models, in The biomedical engineering handbook. In: Bronzino JD, editor. Boca Raton (FL): CRC Press LLC; 2000.
181. Barr RC. Basic electrophysiology. 2nd ed. In: Bronzino JD, editor. The biomedical engineering handbook. Boca Raton (FL): CRC Press LLC; 2000. p 18.

182. Cuellar A, et al. CellML specification. Available at http://www.cellml.org/public/specification/index.html. 2005.
183. Kuznetz LH. A two-dimensional transient mathematical model of human thermoregulation. Am J Physiol 1979;237(6):266–277.
184. Downey D, Seagrave RC. Mathematical modeling of the human body during water replacement and dehydration: Body water changes. Ann Biomed Eng 2000;28:278–290.
185. Fiala D, Lomas KJ, Stohrer M. A computer model of human thermoregulation for a wide range of environmental conditions: The passive system. J Appl Physiol 1999;87:1957–1972.
186. Fiala D, Lomas KJ, Stohrer M. Computer prediction of human thermoregulatory and temperature responses to a wide range of environmental conditions. Int J Biometeorol 2001;45:143–159.
187. Boregowda SC. Thermodynamic modeling and analysis of human stress responses Mechanical Engineering. Old Dominion University; 1998. p 175.
188. Gardner GG, Martin CJ. The mathematical modelling of thermal responses of normal subjects and burned patients. Physiol Meas 1994;15:381–400.
189. Quinn ML, et al. The case for designer animals: (use of simulation to reduce animal studies). Anesthesiology 1987;67(3):A215.
190. Martin JF, et al. A new cardiovascular model for real-time applications. Trans-Soc Computer Simulation 1986;3:31–65.
191. Slate JB. Model-based design of a controller for infusing sodium nitroprusside during postsurgical hypertension. Electrical Engineering. Madison (WI): University of Wisconsin; 1980.
192. Wesseling KH, A baroreflex paradox solution. Utrecht: TNO; 1982. pp. 152–164.
193. Kjaergaard S, et al. Non-invasive estimation of shunt and ventilation-perfusion mismatch. Intensive Care Med 2003;29:727–734.
194. Mesic S, et al. Computer-controlled mechanical simulation of the artificially ventilated human respiratory system. IEEE Trans Biomed Eng 2003;50(6):731–743.
195. Smith NT, Wesseling KH, De Wit B. Evaluation of two prototype devices producing noninvasive, pulsatile calibrated blood pressure from a finger. J Clin Monit 1985;1(1):17–29.
196. Wesseling KH, et al. A simple device for the continuous measurement of cardiac output. Adv Cardiovasc Phys 1983;5:16–52.
197. Wesseling KH, et al. Continuous monitoring of cardiac output. Medicamundi 1976;21:78–90.
198. Jansen J, et al. Continuous cardiac output computed from arterial pressure in the operating room. Br J Anaesth 2001;87(2):212–222.
199. Hirschl MM, et al. Noninvasive assessment of cardiac output in critically ill patients by analysis of the finger blood pressure waveform. Crit Care Med 1997;25(11).
200. Wesseling KH, et al. Computation of aortic flow from pressure in humans using a nonlinear, three-element model. J Appl Physiol 1993;74(5):2566–2573.
201. Jansen J, et al. A comparison of cardiac output derived from the arterial pressure wave against thermodilution in cardiac surgery patients. Br J Anaesth 2001;87(3):212–222.
202. Gizdulich P, Prentza A, Wesseling KH. Models of brachial to finger pulse wave distortion and pressure decrement. Cardiovasc Res 1997;33:698–705.
203. Thoman W, et al. A computer model of intracranial dynamics integrated to a full-scale patient simulator. Computers Biomed Res 1998;31:32–46.
204. Thoman W, et al. Autoregulation in a simulator-based educational model of intracranial physiology. J Clin Monit 1999;15:481–491.
205. Van Meurs W, Nikkelen E, Good M. Pharmacokinetic-pharmacodynamic model for educational simulations. IEEE Trans Biomed Eng 1998;45(5):582–590.
206. Euliano T, et al. Modeling obstetric cardiovascular physiology on a full-scale patient simulator. J Clin Monit 1997;13:293–297.
207. Garfield J, Paskin S, Philip J. An evaluation of the effectiveness of a computer simulation of anaesthetic uptake and distribution as a teaching tool. Medi Educ 1989;23:457–462.
208. Bassingthwaighte JB. Strategies for the physiome project. Ann Biomed Eng 2000;28:1043–1058.
209. Anon., The physiome project. http://nsr.bioeng.washington/PLN. 2005.
210. Anon., The heart of the matter, in The Economist. Available at http://www.economist.com/science/tq/PrinterFriendly.cfm?Story_ID=885127. 2005. p 6.
211. Sideman S. Preface: Cardiac engineering—deciphering the cardiome, in Cardiac engineering: From genes and cells to structure and function. In: Sideman S, Beyar R, editors. New York: New York Academy of Sciences; 2004.
212. Popel AS, et al. The microcirculation physiome project. Ann Biomed Eng 1998;26:911–913.
213. Tawhai MH, Burrowes KS. Developing integrative computational models of pulmonary structure. The Anatom Rec (P B: New A) 2003;275B:207–218.
214. Tawhai MH, Ben-Tal A. Multiscale modeling for the lung physiome. Cardiovas Eng 2004;4(1).
215. Lonie A. The kidney simulation project. 2005.
216. Crampin EJ, et al. Computational physiology and the physiome project. Exp Physiol 2004;89(1):1–26.
217. Hunter P, Robbins P, Noble D. The IUPS human physiome project. Pflugers Arch - Eur J Physiol 2002;445:1–9.
218. Crampin EJ, et al. Computational physiology and the physiome project. Exp Physiol 2003;89(1):1–26.
219. Hunter PJ, Borg TK. Integration from proteins to organs-the physiome project. Nat Rev: Mol Cell Biol 2003;4:237–243.
220. Anon., The IUPS physiome project. Available at http://www.physiome.org.nz/anatml/pages/specification.html. 2005.
221. Turing A. The chemical basis of morphogenesis. Philos Trans R Soc (London), Ser A 1952;237:37–72.
222. Weiss JN, Qu Z, Garfinkel X. Understanding biological complexity: Lessons from the past. FASEB J 2003;17.

See also ELECTROCARDIOGRAPHY, COMPUTERS IN; ELECTROPHYSIOLOGY; PULMONARY PHYSIOLOGY; RADIOTHERAPY TREATMENT PLANNING, OPTIMIZATION OF.

PICTURE ARCHIVING AND COMMUNICATION SYSTEMS

KATHERINE ANDRIOLE P.
Harvard Medical School
Boston, Massachusetts

INTRODUCTION AND HISTORICAL OVERVIEW

A picture archiving and communication system (PACS) is a medical image management system, or a collection of electronic technologies that enable the digital or filmless imaging department. In a PACS, images are acquired, stored, and transmitted or distributed digitally, as well as interpreted digitally using what is known as soft copy display.

In the analog world, even images that were created by inherently digital modalities are printed to film for display on a light box or alternator, and for image archival as the legal record where films are stored in large rooms of filed film jackets. Imaging examinations on film must be transported from one location to another by foot for viewing by radiologists and referring clinicians. Films are retrieved and re-archived manually by film library personnel. Using a PACS, images are acquired as digital computer files, stored on computer disks or other digital media, transmitted across computer networks, and viewed and manipulated on computer workstations.

The benefits of PACS are numerous and include rapid and remote data distribution within and between health-care enterprises. Digital archival is more permanent than film with regard to media degradation as well as the problem of lost films. A PACS gives multiple users in distinct locations simultaneous access to the same imaging examinations. Also, the digital nature of the data allows for image manipulation and processing, which may lead to enhanced visualization of radiological features and improved interpretation of imaging studies. The potential impact of using a PACS can be more expedient care, more efficient workflow, and more cost-effective and higher quality care.

PACS and filmless radiology are a better way to practice imaging in medicine today. The number of images per study has grown beyond what is feasible for viewing on film. Still, today, only about 20% of health-care enterprises have implemented PACS.

Picture archiving and communication systems have come about via a convergence of technological and economic factors. The facilitating technologies responsible include a dramatic improvement in computing power, the advancement of network capabilities and storage devices, the development of imaging standards, and systems integration. In the 1990s, PACS applications challenged computer hardware. Today, PACS applications are only a subset of what computers can do.

Early PACS

In 1979, the earliest paper proposing the concept of a PACS was published by Heinz U. Lemke, Ph.D., entitled "Applications of Picture Processing, Image Analysis and Computer Graphics Techniques to Cranial CT Scans" (1). In the early 1970s, M. Paul Capp, M.D., Sol Nudelman, Ph.D., and colleagues at the University of Arizona Health Sciences Center organized a digital imaging group that developed the first digital subtraction angiography (DSA) device that was the precursor to clinical digital imaging. They introduced the notion of a "photoelectronic radiology department" and depicted a system block diagram of the demonstration facility they had built (2).

Samuel J. Dwyer, III, Ph.D. predicted the cost of managing digital diagnostic images in a radiology department (3) and, along with Andre Duerinckx, M.D., Ph.D., organized a landmark conference at which the acronym PACS was coined (4). This meeting, sponsored by the International Society for Photo-Optical Engineering (SPIE), titled "The First International Conference and Workshop on Picture Archiving and Communications Systems (PACS) for Medical Applications," was held in Newport Beach, CA, January 18–21, 1982, and continues today as the Annual SPIE International Symposium on Medical Imaging. Two panel discussions "Equipment Manufacturers' View on PACS" and "The Medical Community's View on PACS" that took place at the first PACS conference were captured in the proceedings (4). Talk occurred of linking imaging modalities into a single digital imaging network and the recognition that, in order for this linking to be practical, standards would be required. Steven C. Horii, M.D., participated in those beginning discussions and has been instrumental in bringing about the creation and implementation of a standard for digital medical imaging, now known as the Digital Imaging and Communications in Medicine or DICOM Standard.

A number of PACS pioneers have contributed to the advancement of digital medical imaging to its current status through efforts in research and development, design, implementation, testing, analysis, standards creation, and education of the technical and medical communities. In 1982–1983, Dwyer oversaw the building of what is often considered the first PACS. In 1983, the first of numerous papers was presented by H. K. Bernie Huang, D.Sc., FRCR, detailing the PACS efforts at the University of California at Los Angeles, which culminated years later in a clinically operational filmless radiology department (5). G. James Blaine, D.SC., and R. Gilbert Jost, M.D., at the Washington University School of Medicine, focused their efforts on the development of utilities enabling PACS research and development (6). In the mid- to late-1980s, several researchers described their prototype PACS hardware and software efforts (7–10).

Similar activities were taking place in Europe and Asia. Hruby opened a completely digital radiology department in the Danube Hospital in Vienna in 1990, setting the tone for the future (11). Several academic radiology departments in the United States began working with major vendor partners to further the technology and its clinical implementation. Such academic–industry collaborations continue the advancement of PACS today.

Standards

The development of standards in medical imaging is one of the facilitating factors that has enabled PACS to mature and become more widely used. DICOM (Digital Imaging and Communications in Medicine) (12) along with several other integration standards, has been one of the most important accomplishments for PACS. DICOM is a standard that was created to promote an open architecture for imaging systems, allowing interoperability between systems for the transfer of medical images and associated information. As an exchange protocol, it was designed to bridge differing hardware devices and software applications.

With the increasing use of computers in clinical applications, and with the introduction of digital subtraction angiography and computed tomography (CT) in the 1970s, followed by other digital diagnostic imaging modalities, the American College of Radiology (ACR) and the

National Electrical Manufacturers Association (NEMA) recognized the emerging need for a standard method for transferring images and associated information between devices manufactured by the various vendors (12). These devices produced a variety of digital image formats. The push by the radiological community for a standard format across imaging devices of different models and makes began in 1982. ACR and NEMA formed a joint committee in 1983 to develop a standard to promote communication of digital image information, regardless of device manufacturer. It was felt that this committee would facilitate the development and expansion of picture archiving and communication systems that could also interface with other hospital information systems and allow the creation of diagnostic information databases that could be interrogated by a wide variety of geographically distributed devices.

The ACR-NEMA Standard version 1.0 was published in 1985. Two revisions followed, one in October 1986 and the second in January 1988 as version 2.0. It included version 1.0, the published revisions, and additional revisions. ACR-NEMA 2.0 consisted of a file header followed by the image data. The file header contained information relevant to the image, such as matrix size or number of rows and columns, pixel size, gray-scale bit depth, and so on, as well as information about the imaging device and technique (i.e., Brand X CT scanner, acquired with contrast). Patient demographic data, such as name and date of birth, were also included in the image header. The ACR-NEMA 2.0 standard specified exactly where in the header each bit of information was to be stored, such that the standard required image information could be read by any device, simply by going to the designated location in the header. Version 2.0 also included new material to provide command support for display devices, to introduce a new hierarchy scheme to identify an image, and to add data elements for increased specificity when describing an image. These standards publications specified a hardware interface, a minimum set of software commands, and a consistent set of data formats. Data included patient demographic information as well as imaging parameters. This standard unified the format of imaging data but functioned only as a point-to-point procedure, not including a networking communications protocol until later versions.

In 1994, at the Radiological Society of North America (RSNA) Meeting, a variety of imaging vendors participated in an impressive demonstration of the new and evolving imaging standard (ACR-NEMA 3.0). Participants attached their devices to a common network and transmitted their images to one another. In addition to the standard image format of ACR-NEMA 2.0, the DICOM standard included a network communications protocol or a common language for sending and receiving images and relevant data over a network.

Today, this standard, which is currently designated Digital Imaging and Communications in Medicine (DICOM), embodies a number of major enhancements to previous versions of the ACR-NEMA Standard, the first that is applicable to a networked environment. The original ACR-NEMA Standard included a well-defined format for the image data but worked only in point-to-point configurations. In order to function in a networked environment, a Network Interface Unit (NIU) was required. Operation in a networked environment is supported today using the industry standard networking protocol TCP-IP (transfer communication protocol - Internet protocol). Thus, in addition to the format of the data being exchanged between medical imaging devices, the DICOM Standard also specifies how the devices themselves should communicate using simple commands such as Find, Get, Move, Send, and Store. These commands operate on objects such as images and text, which are formatted in terms of groups and elements. The hierarchy of data is of the form patient, study, series or sequence, and image.

DICOM specifies, through the concept of service classes, the semantics of commands and associated data, and it also specifies levels of conformance. The DICOM standard language structure is built on information objects (IO), application entities (AE), and service class users (SCU) and providers (SCP). Information objects include, for example, the image types, such as CT, MRI, and CR. The application entities include the devices, such as a scanner, workstation, or printer. The service classes (SC*) define an operation on the information object via service object pairs (SOP) of IO and SCU and SCP. The types of operations performed by an SCU-SCP on an IO include storage, query-retrieve, verification, print, study content notification, study content notification and patient, and study and results management. An example DICOM-formatted message is written in terms of a tag (consisting of a group and an element) followed by the length of the tag, followed by the value: 0008,0020-8-20050402 represents group 8, element 20, which corresponds to the study date given as an 8 character field. DICOM is a continuously evolving standard with significant updates yearly.

The information technologies most familiar to radiology departments are PACS and Radiology Information Systems (RIS). PACS or image management systems typically perform the functions of image acquisition, distribution, display, and archival. Often, separate systems exist for primary interpretation in the radiology department and for use enterprise-wide by nonradiologist clinicians. The RIS typically maintains radiology-specific data, such as imaging examination orders, reports, and billing information.

Although implementation of either one or both of these systems can improve workflow and reduce operating costs, the elimination of film and paper from the practice of radiology is not easily realized without integrating the functions performed by several other information technologies. These systems include hospital information systems (HIS), Computerized Physician Order-Entry (CPOE) Systems, Report Generation Systems, Decision Support Systems, and Case-Based Teaching Files. Together, these systems make up the electronic medical record (EMR).

Several of these systems are more widely known to the health-care enterprise outside of diagnostic imaging departments. They, none-the-less, contain data essential to high quality low cost radiological practice. The value these systems can bring to medical imaging in the clinical enterprise is high, but they must be seamlessly integrated. Including several features, such as single sign-on, electronic

master patient index, and context sensitivity, can help make these information systems tools and technologies most useful to radiology.

An effort known as Integrating the Healthcare Enterprise (IHE) provides a framework using existing standards, such as DICOM and Health Level 7 (HL7), to facilitate intercommunications among these disparate systems and to optimize information efficiency. The IHE is a joint effort of the Radiological Society of North America (RSNA) and the Healthcare Information and Management Systems Society (HIMSS) begun in 1998 to more clearly define how existing standards should be used to resolve common information system communication tasks in radiology.

The IHE technical framework defines a common information model and a common vocabulary for systems to use in communicating medical information. It specifies exactly how the DICOM and HL7 standards are to be used by the various information systems to perform a set of well-defined transactions that accomplish a particular task. The original seven tasks facilitated by the IHE include scheduled workflow, patient information reconciliation, consistent presentation of images, presentation of grouped procedures, access to radiology information, key image notes, and simple image and numeric reports. Profiles continue to be added yearly to the framework, enhancing its value in integrating information systems in a healthcare environment.

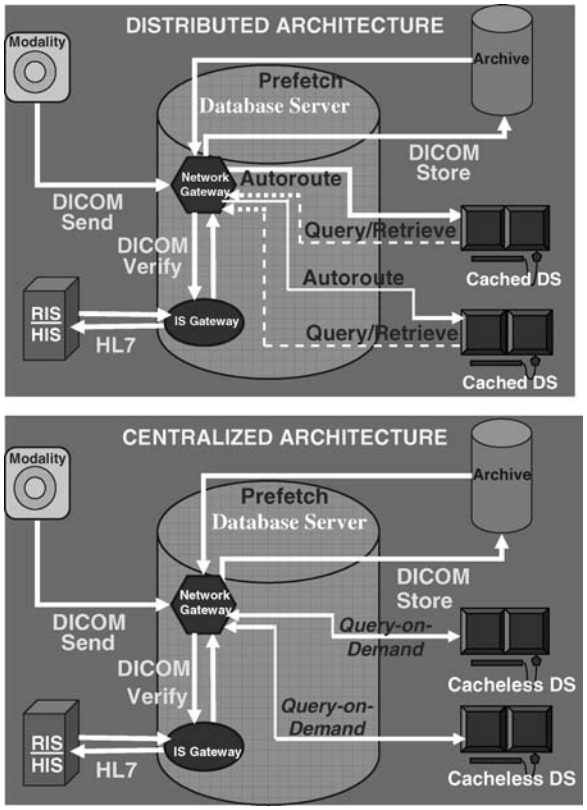


Figure 1. (a) Distributed versus **(b)** centralized PACS architectures.

Architectures

Two basic architectures are used in PACS today, distributed or cached and centralized or cacheless depicted in Figs. 1a and 1b, respectively. Data is acquired into the PACS in the same manner for both architectures, from the imaging modality via a DICOM sent to a network gateway. Demographic data is verified by interfacing to the radiology or hospital information system (RIS-HIS) through an IS gateway. Studies are permanently archived by a DICOM store to an electronic archive device.

In a distributed system, images and other relevant data are automatically routed to the workstation(s) where the studies are expected to be viewed and cached or stored on the local display station disk. The best-distributed PACS also prefetch relevant prior examinations from the long-term archive and automatically route them to the pertinent display for immediate access to comparison images. Studies not automatically routed to a workstation can be queried for and retrieved on request.

In a centralized system, images remain on a large central server and are only sent to the workstation when they are called up for display. In this query-on-demand architecture, data is retrieved instantaneously into memory for viewing and manipulation. Images are never stored on the local display station disk, thus centralized systems are also known as cacheless.

A centralized architecture is easier to implement and maintain and uses a simple on-demand data access model, but also has a single point-of-failure in the central server component. A cacheless system is also bandwith-limited, requiring a fast network connection from display stations to the central server. The distributed architecture requires more complex workflow logic such as autorouting and prefetching of data to implement. However, distributed PACS may have more functionality and may be more easily scalable than centralized systems. Early PACS were predominately distributed systems. With the increasing availability of high bandwidth networks and large amounts of inexpensive storage media, most PACS today follow the centralized architecture. Web-based PACS typical of today's systems are a specialized subset of centralized architectures. Future PACS may evolve to be a combination of both distributed architectures and centralized architectures, encompassing the best of each design.

If a PACS operates with a cached architecture in which data is automatically distributed to and stored at the display station, then the online storage capabilities should include space for maintaining all the pertinent examinations for a given episode of current care, (i.e., three days for outpatients and six days for inpatients). Space for prefetched relevant historical examinations should also be included in the anticipated storage requirements.

If the PACS operates as a cacheless centralized system, then it is important to have adequate capacity to store a patient's clinical encounter on the server. In this case, it is important to have large amounts of online RAID (redundant array of independent disks—see section on RAID below) at the central server instead of large amounts of local storage at each display station. RAID capacity should also encompass relevant prior examinations prefetched to the server.

KEY COMPONENTS AND ESSENTIAL FEATURES

Image Acquisition

Digital acquisition of data from the various imaging modalities for input to a PACS is the first step to eliminating film in medical imaging. Essential features for successful clinical implementation include conformance with the DICOM standard, radiology information system – hospital information system (RIS-HIS) interfacing, and workflow integration Quality assurance (QA) and quality control (QC) and troubleshooting problems occurring specifically at image acquisition are also critical as these problems affect the integrity of data in the archive.

Integration with PACS. Image acquisition is the first point of data entry into a PACS,system and, as such, errors generated here can propagate throughout the system, adversely affecting clinical operations. General predictors for successful incorporation of image acquisition devices into a digital imaging department include the following: ease of device integration into the established daily workflow routine of the clinical environment, high reliability and fault-tolerance of the device, simplicity and intuitiveness of the user interface, and device speed (13). The integration of modalities with PACS and information systems using the DICOM modality worklist feature (see below) can reduce the number of patient demographic errors and the number of cases that are inappropriately or unspecified and therefore not archiveable, which also ensures the correctness of permanently archived information.

DICOM. Imaging modality conformance with the DICOM standard is critical. DICOM consists of a standard image format as well as a network communications protocol. Compliance with this standard enables an open architecture for imaging systems, bridging hardware and software entities, allowing interoperability for the transfer of medical images and associated information between disparate systems.

The DICOM standard is used, for example, to negotiate a transaction between a compliant imaging modality and a compliant PACS workstation. The scanner notifies the workstation, in a language both understand, that it has an image study to send to it. The workstation replies to the modality when it is ready to receive the data. The data is sent in a format known to all, the workstation acknowledges receipt of the image, and then the devices end their negotiation. Data is formatted in terms of groups and elements. Group 8, for example, pertains to image identification parameters (such as study, series, and image number) and Group 10 includes patient demographics (such as patient name, medical record number, and date of birth).

Prior to DICOM, the acquisition of digital image data and relevant information was extremely difficult, often requiring separate hardware devices and software programs for different vendors' products, and even for different models of devices made by the same manufacturer because each vendor used their own proprietary data format and communication's protocol. Most of the major manufacturers of imaging devices currently comply with the DICOM standard, thus greatly facilitating an open systems architecture consisting of multivendor systems. For many legacy devices purchased prior to the establishment of DICOM, an upgrade path to compliance can be performed. For those few devices that do not yet meet the standard, interface boxes consisting of hardware equipment and software programs that convert the image data from the manufacturer's proprietary format to the standard form are available.

RIS-HIS Interfacing for Data Verification. Equally essential, particularly at acquisition, is integrating the radiology information system (RIS) or hospital information system (HIS) with the PACS, which greatly facilitates input of patient demographics (name, date, time, medical record number (MRN) to uniquely identify a patient, accession number (AccNum) to uniquely identify an imaging examination, exam type, imaging parameters, etc.) and enables automatic PACS data verification, correlation, and error correction with the data recorded in the RIS-HIS. Most imaging modalities are now tightly coupled with the RIS, providing automatic downloading of demographic information from the RIS via barcode readers or directly to the scanner console (via modality worklist capability) and, hence, to the DICOM header. This copling eliminates the highly error-prone manual entry of data at acquisition.

HL7 is the RIS-HIS standard, and compliance with HL7 is desirable. RIS-HIS databases are typically patient-centric, enabling query and retrieval of information by the patient and study, series, or image data hierarchy. Integration of RIS-HIS data with the PACS adds intelligence to the system, helping to move data around the system based on "*how, what* data should be delivered *where* and *when*", automating the functions performed traditionally by the film librarian.

Modality Worklist. Many vendors now provide the capability to download RIS-HIS schedules and worklists directly to the imaging modality, such as most computed tomography (CT), magnetic resonance imaging (MRI), digital fluoroscopy (DF), and ultrasound (US) scanners. In these circumstances, the imaging technologist need only choose the appropriate patient's name from a list on the scanner console monitor (i.e., by pointing to it on a touchscreen pad), and the information contained within the RIS-HIS database will be downloaded into the PACS header and associated with the image data for that patient examination.

The general DICOM model for acquisition of image and relevant data from the imaging modality involves the modality device acting as a SCU, which provides the data, and storing it to a SCP, which provides the service: devices such as a PACS acquisition gateway or an image display workstation. In the modality worklist function, however, the image device receives the pertinent patient demographics and image study information from a worklist server, such as a PACS- RIS- or RIS-HIS-interfaced device.

Two modes exist for accomplishing the RIS-HIS data transfer to the imaging modality. The first involves data,

being transferred automatically to the modality based on the occurrence of an event trigger, such as an examination being scheduled or a patient having arrived. The second method involves a query from the modality to the RIS-HIS or some other information system that holds relevant data, such as an electronic order-entry system or even some PACS databases, which may be initiated by entry of some identifier at the modality, such as bar coding of the study accession number or the patient medical record number from the scheduling card. This method then initiates a request for the associated RIS-HIS information (patient name, date of birth) to be sent from the worklist server on demand.

The benefits of the DICOM modality worklist cannot be overstated. Incorrectly (manually) entered patient demographic data, such as all the permutations of patient name (i.e., James Jones, J Jones, Jones J) can result in mislabeled image files and incomplete study information and, as such, is crucial to maintaining the integrity of the PACS database. Furthermore, the improvements in departmental workflow efficiency and device usability are greatly facilitated by modality worklist capabilities. For those few vendors not offering DICOM modality worklist for their imaging devices, several interface or broker boxes are available that interconnect PACS to RIS-HIS databases translating DICOM to HL7 and vice versa. Figure 2 diagrams an example of how RIS, HIS, and PACS systems might interact upon scheduling an examination for image acquisition into a PACS (14).

Acquisition of the Native Digital Cross-Sectional Modalities. Image acquisition from the inherently digital modalities, such as CT, MRI, and US, should be a direct digital DICOM capture. Direct digital interfaces allow capture and transmission of image data from the modality at the full spatial resolution and full bit depth of gray scale inherent to the modality, whereas the currently outdated analog (video) frame grabbers digitize the video signal voltage output going to an image display, such as a scanner console monitor. In the frame-grabbing method, as in printing an image to film, the image quality is limited

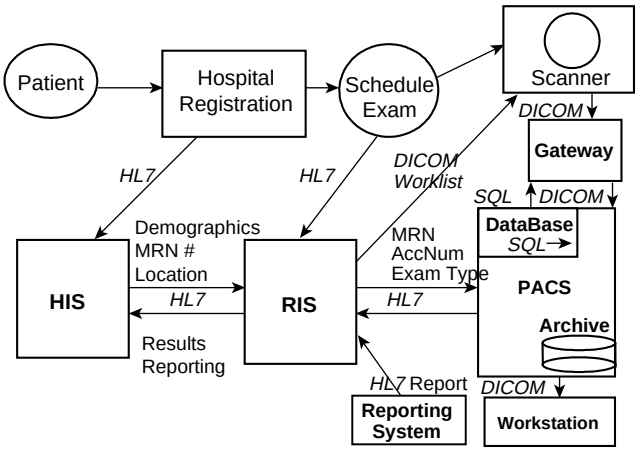


Figure 2. Diagram of how RIS, HIS, and PACS systems might interact on scheduling an examination for image acquisition into a PACS.

Table 1. The Commonly PACS-Interfaced Cross-Sectional Modalities and their Inherent File Sizes

Modality	Image Matrix Size	Grayscale Bit Depth
Computed Tomography (CT)	512 × 512 pixels	12 – 16 bits
Digital Angiography & (RA)	512 × 512 pixels or	8 – 12 bits
Digital Fluoroscopy (DF)	1024 × 1024 pixels or 2048 × 2048 pixels	
Magnetic Resonance Imaging (MRI)	256 × 256 pixels	12 – 16 bits
Nuclear Medicine Images (NUC)	64 × 64 pixels or	8 – 32 bits
	128 × 128 pixels or 256 × 256 pixels	
Ultrasound (US)	64 × 64 pixels or 128 × 128 pixels	16 – 32 bits

by the process to just 8 bits (or 256 gray values), whereas most modalities have the capability to acquire in 12, 16, or even 32 bits for color data. Capture of only 8 bits may not allow viewing in all the appropriate clinical windows and levels or contrast and brightness settings and is, therefore, not optimal.

For example, when viewing a CT of the chest, one may wish to view in lung window and level settings and in mediastinal and bone windows and levels. Direct capture of the digital data will allow the viewer to dynamically window and level through each of these settings on-the-fly (in real time) at the softcopy display station. However, to view all appropriate window and level settings on film, several copies of the study would have to be printed, one at each window and level setting. If one performs the analog acquisition or frame grabbing of the digital data, the viewer can only window and level through the 8 bits captured, which may not be sufficient. Thus, direct capture of digital data from the inherently digital modalities is the preferred method of acquisition. Table 1 lists the cross-sectional modalities commonly interfaced to PACS along with their inherent file sizes and bit depths.

Acquisition of Projection Radiography. Methods for digital image acquisition of the conventional projection X ray include computed radiography (CR) scanners or imaging with photostimulable or storage phosphors, digitization of existing analog film, and digital radiography (DR) devices. Digital acquisition of images already on film can be accomplished using a variety of image digitization devices or film scanners, including the no longer used analog video cameras with analog-to-digital converters (ADC), digital cameras, charge-coupled devices (CCD), and laser scanners. Both CR and DR are replacement methods for capturing conventional screen-film projection radiography.

Film Digitizers. Film digitizers will still be necessary even in the all digital or filmless imaging department, so that film images from outside referrals lacking digital capabilities can be acquired into the system and viewed digitally. Film digitizers convert the continuous optical

density values on film into a digital image by sampling at discrete evenly spaced locations and quantizing the transmitted light from a scan of the film into digital numbers. Several types of film digitizers exist today, with some used more frequently than others in PACS and teleradiology applications.

The analog video camera with ADC, or camera on a stick, was used in low cost, entry-level teleradiology applications but is no longer used in PACS applications today because of its manual operation. The analog video camera requires an illumination source and careful attention to lens settings, focus, f-stop, and so on. In addition, it has a maximum resolution of 1024 by 1024 by 8 bits (256 grays), thus limiting the range of window and level, or contrast and brightness values, the resulting digital image can be displayed in. Digital cameras produce a digital signal output directly from the camera at a maximum resolution of 2048 by 2048 by 12 bits (4096 grays) but are still infrequently used in PACS due to their high cost.

More commonly used are film scanners such as the CCD and laser scanners sometimes called flat-bed scanners. CCD scanners use a row of photocells and uniform bright light illumination to capture the image. A lens focuses the transmitted light from the collimated, diffuse light source onto a linear CCD detector, and the signal is collected and converted to a digital electronic signal via an ADC. CCD scanners have a maximum resolution of 4096 by 4096 by 8–12 bits, but they have a narrow film optical density range to which they can respond. CCD scanners have been used in high-end teleradiology or entry-level in-house film distribution systems, such as image transmission to the intensive care units (ICUs).

The laser scanner or laser film digitizer uses either a helium-neon (HeNe) gas laser or a solid-state diode laser source. The laser beam is focused by lenses and directed by mirror deflection components, and the light transmitted through the film is collected by a light guide, its intensity detected by a photomultiplier tube, converted to a proportional electronic signal, and digitized in an ADC. Laser scanners use a fine laser beam of generally variable or adjustable spot sizes down to 50 μm (producing an image sharpness of approximately 10 line pairs per millimeter). They have a maximum spatial resolution of 4096 by 5120 and a grayscale resolution of 12 bits, and can accommodate the full optical density range of film. They are semi- or fully-automatic in operation and are currently the scanner of choice for PACS applications even though they are often more expensive than CCD scanners.

Computed Radiography (CR). Computed Radiography (CR) refers to projection X-ray imaging using photostimulable or storage phosphors as the detector. In this modality, X rays incident upon a photostimulable phosphor-based image sensor or imaging plate produce a latent image that is stored in the imaging plate until stimulated to luminesce by laser light. This released light energy can be captured and converted to a digital electronic signal for transmission of images to display and archival devices. Unlike conventional screen-film radiography in which the film functions as the imaging sensor, or recording medium, as well as the display device and storage media, CR eliminates film from the image recording step, resulting in a separation of image capture from image display and image storage. This separation of functions potentiates optimization of each of these steps individually. In addition, CR can capitalize on features common to all digital images, namely, electronic transmission, manipulation, display, and storage of radiographs (15).

Technological advances in CR over time have made this modality widely accepted in digital departments. Hardware and software improvements have occurred in the photostimulable phosphor plate, in image reading-scanning devices, and in image processing algorithms. Overall reduced cost of CR devices, as well as a reduction in the cost and increased utility of image display devices, have contributed to the increased acceptance of CR as a viable digital counterpart to conventional screen-film projection radiography.

Review of the Fundamentals

Process Description. ACR system consists of a screen or plate of a stimulable phosphor material that is usually contained in a cassette and is exposed in a manner similar to the traditional screen-film cassette. The photostimulable phosphor in the imaging plate (IP) absorbs X rays that have passed through the patient, "recording" the X-ray image. Like the conventional intensifying screen, CR plates produce light in response to an X ray, at the time of exposure. However, storage phosphor plates have the additional property of being capable of storing some of the absorbed X-ray energy as a latent image. Plates are typically made of an europium-doped barium-fluoro-halide-halide crystallized matrix. Electrons from the dopant ion become trapped just below the conduction band when exposed to X rays. Irradiating the imaging plate at some time after the X ray exposure with red or near-infrared laser light liberates the electrons into the conduction band, stimulating the phosphor to release some of its stored energy in the form of green, blue, or ultraviolet light—the phenomenon of photostimulable luminescence. The intensity of light emitted is proportional to the amount of X ray absorbed by the storage phosphor (16).

The readout process uses a precision laser spot-scanning mechanism in which the laser beam traverses the imaging plate surface in a raster pattern. The stimulated light emitted from the IP is collected and converted into an electrical signal, with optics coupled to a photomultiplier tube (PMT). The PMT converts the collected light from the IP into an electrical signal, which is then amplified, sampled to produce discrete pixels of the digital image, and sent through an ADC to quantize the value of each pixel (i.e., a value between 0 and 1023 for a 10 bit ADC or between 0 and 4095 for a 12 bit ADC).

Not all of the stored energy in the IP is released during the readout process. Thus, to prepare the imaging plate for a new exposure, the IP is briefly flooded with high intensity (typically fluorescent) light. This erasure step ensures removal of any residual latent image.

A diagram of the process steps involved in a CR system is shown in Fig. 3. In principle, CR inserts a digital computer between the imaging plate receptor (photostimulable phosphor screen) and the output film. This digital

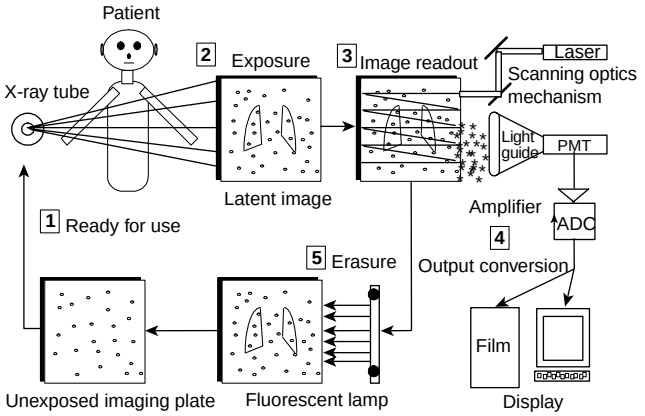


Figure 3. The image production steps involved in CR. The imaging plate is exposed to X rays, read out by a laser scanning mechanism, and erased for reuse. A light guide collects the photostimulated luminescence and feeds it to a photomultiplier tube (PMT) that converts the light signal to an electrical signal. Amplification, logarithmic conversion, and analog-to-digital conversion produce the final digital signal that can be displayed on a cathode ray tube monitor or sent to a laser printer for image reproduction on film.

processor can perform a number of image processing tasks including compensating for exposure errors, applying appropriate contrast characteristics, enhancing image detail, and storing and distributing image information in digital form.

System Characteristics. One of the most important differences between CR and screen-film systems is in exposure latitude. The response of a digital imaging system relates the incident X ray exposure to the resulting pixel value output. System sensitivity is the lowest exposure that will produce a useful pixel value, and the dynamic range is the ratio of the exposures of the highest and lowest useful pixel values (17). Storage phosphor systems have extremely wide exposure latitude. The wide latitude of storage phosphor systems, and the effectively linear detector characteristic curve, allows for a wider range of exposure information to be captured in a single image than is possible with any screen-film system. In addition, the wide dynamic range of CR allows it to be used under a broad range of exposure conditions without the need for changing the basic detector, also making CR an ideal choice for applications in which exposures are highly variable or difficult to control, as in portable or bedside radiography. Through image processing, CR systems can usually create a diagnostic image out of under- or over-exposures via appropriate look-up table correction. In the screen-film environment, such under- or over-exposures might have necessitated retakes and additional exposure to the patient.

Dose requirements of a medical imaging system depend on the system's ability to detect and convert the incoming signal into a usable output signal. It is important to stress that CR systems are not inherently lower dose systems than screen-film. In fact, several studies have demonstrated a higher required exposure for CR to achieve equivalent optical density on screen-film (18,19). However, the wider latitude of storage phosphor systems makes them much more forgiving of under- or over-exposure. As in any digital radiography system, when dose is decreased, the noise due to quantum mottle increases (20). Reader tolerance of this noise tends to be the limiting factor on the lowest acceptable dose.

In some clinical situations, the radiologist may feel comfortable in lowering the exposure technique factor to reduce dose to the patient, such as in pediatric extremity X-ray exams. In others, such as imaging the chest of the newborn, one may wish to increase exposure to reduce the more visible mottle (at lower doses) to avoid mistaking the noise over the lungs as an indication of pulmonary interstitial emphysema, for example. CR systems are signal-to-noise-limited (SNR-limited), whereas screen-film systems are contrast-limited.

Image Quality. DQE: Objective descriptors of digital image quality include detective quantum efficiency (DQE), which is a measure of the fidelity with which a resultant digital image represents the transmitted X-ray fluence pattern (i.e., how efficiently a system converts the X-ray input signal into a useful output image), and includes a measure of the noise added (17). Also taken into account are the input/output characteristics of the system and the resolution response of unsharpness or blur added during the image capture process. The linear, wide-latitude input/output characteristic of CR systems relative to screen-film leads to a wider DQE latitude for CR, which implies that CR has the ability to convert incoming X-ray quanta into "useful" output over a much wider range of exposures than can be accommodated with screen-film systems (20).

Spatial Resolution: The spatial resolution response or sharpness of an image capture process can be expressed in terms of its modulation transfer function (MTF), which, in practice, is determined by taking the Fourier Transform of the line spread function (LSF) and relates input subject contrast to imaged subject contrast as a function of spatial frequency (17). The ideal image receptor adds no blur or broadening to the input LSF, resulting in an MTF response of one at all spatial frequencies. A real image receptor adds blur, typically resulting in a loss of MTF at higher spatial frequencies.

The main factors limiting the spatial resolution in CR, similar to screen-film systems, is X-ray scattering within the phosphor layer. However, it is the scattering of the stimulating beam in CR, rather than the emitted light as in screen-film, that determines system sharpness (20,21). Broadening of the laser light spot within the IP phosphor layer spreads with the depth of the plate. Thus, the spatial resolution response of CR is largely dependent on the initial laser beam diameter and on the thickness of the IP detector. The reproducible spatial frequency of CR is also limited by the sampling used in the digital readout process. The spatial resolution of CR is less than that of screen-film, with CR ranging from 2.5 to 5 line pairs per millimeter (lp/mm) using a 200 μm laser spot size and a digital matrix size of approximately 2,000 by 2,500 pixels versus the 5–10 lp/mm or higher spatial resolution of screen-film.

Finer spatial resolution can technically be achieved today with the ability to tune laser spot sizes down to

50 μm or less. However, the image must be sampled more finely (approximately 4,000 by 5,000 pixels) to achieve 10 lp/mm. Thus, a tradeoff exists between the spatial resolution that can technically be achieved and the file size to practically transmit and store. Most general CR examinations are acquired using a 200 μm laser spot size and a sampling of 2 k by 2.5 k pixels. For examinations requiring very fine detail resolution, such as in mammography, images are acquired with a 50 μm laser spot size and sampled at 4 k by 5 k pixels.

Contrast Resolution: The contrast or gray-scale resolution for CR is much greater than that for screen-film. Note that because overall image quality resolution is a combination of spatial and gray-scale resolution, the superior contrast resolution of CR can often compensate for its lack of inherent spatial resolution. By manipulating the image contrast and brightness, or window and level values, respectively, small features often become more readily apparent in the image, which is analogous to "bright-lighting" or "hot-lighting" a bone film, for example, when looking for a small fracture. The overall impression is that the spatial resolution of the image has been improved when, in fact, it has not changed—only the contrast resolution has been manipulated. More work needs to be done to determine the most appropriate window and level settings with which to initially display a CR image. Lacking the optimum default settings, it is often useful to "dynamically" view CR softcopy images with a variety of window and level settings.

Noise: The types of noise affecting CR images include X-ray dose-dependent noise and fixed noise (independent of X-ray dose). The dose-dependent noise components can be classified into X-ray quantum noise, or mottle, and light photon noise (21). The quantum mottle inherent in the input X-ray beam is the limiting noise factor, and it develops in the process of absorption by the imaging plate, with noise being inversely proportional to the detector X-ray dose absorption. Light photon noise occurs in the process of photoelectric transmission of the photostimulable luminescence light at the surface of the PMT.

Fixed-noise sources in CR systems include IP structural noise (the predominant factor), noise in the electronics chain, laser power fluctuations, quantization noise in the analog-to-digital conversion process, and so on (20,21). IP structural noise develops from the nonuniformity of phosphor particle distribution, with finer particles providing noise improvement. Note that for CR systems, it is the noise sources that limit the DQE system latitude, whereas in conventional X-ray systems, the DQE latitude is limited by the narrower exposure response of screen-film.

Comparison with Screen-Film: The extremely large latitude of CR systems makes CR more forgiving in difficult imaging situations, such as portable examinations, and enables decreased retake rates for improper exposure technique, as compared with screen-film. The superior contrast resolution of CR can compensate in many cases for its lesser spatial resolution. Cost savings and improved radiology departmental workflow can be realized with CR and the elimination of film for projection radiographs.

Available CR Systems.

Historical Perspective. Most of the progress in storage phosphor imaging has been made since World War II (22). In 1975, Eastman Kodak Company (Rochester, NY) patented an apparatus using infrared-stimulable phosphors or thermoluminescent materials to store an image (23). In 1980, Fuji Photo Film (Tokyo, Japan) patented a process in which photostimulable phosphors were used to record and reproduce an image by absorbing radiation and then releasing the stored energy as light when stimulated by a helium-neon laser (24). The emitted phosphor luminescence was detected by a PMT, and the electronic signal produced reconstructed the image.

Fuji was the first to commercialize a storage phosphor-based CR system in 1983 (as the FCR 101) and published the first technical paper (in Radiology) describing CR for acquiring clinical digital X-ray images (25). The central processing type second-generation scanners (FCR 201) were marketed in 1985 (21). Third-generation Fuji systems marketed in 1989 included distributed processing (FCR 7000) and stand-alone (AC-1) types (21). Fuji systems in the FCR 9000 series are improved, higher speed, higher performance third-generation scanners. Current Fuji systems include upright chest units, CR detectors in wall and table buckeyes, multiplate autoloaders, and more compact stand-alone units.

In 1992, Kodak installed its first commercial storage phosphor reader (Model 3110) (16). Later models include autoloader devices. In 1994, Agfa-Gevaert N.V. (Belgium) debuted its own CR system design (the ADC 70) (26). In 1997, Agfa showed its ADC Compact with greatly reduced footprint. Agfa also introduced a low cost, entry-level single-plate reader (the ADC Solo) in 1998, appropriate for distributed CR environments such as clinics, trauma centers, and ICUs. In 1998, Lumisys presented its low cost, desktop CR unit (the ACR 2000) with manual-feed, single-plate reading. Numerous desktop units have been introduced including the Orex CR. Konica Corp. debuted its own device (XPress) in 2002 and, later, the Regius upright unit, both of which have relatively fast scan times (at 40 and 16 s cycle times, respectively).

Many companies have been involved in CR research and development, including N.A. Philips Corp.; E.I. DuPont de Nemours & Co.; 3M Co.; Hitachi, Ltd.; Siemens AG; Toshiba Corp.; General Electric Corp.; Kasei Optonix, Ltd.; Mitsubishi Chemical Industries, Ltd.; Nichia Corp.; GTE Products Co.; and DigiRad Corp. (20).

Technological Advances. Major improvements in the overall CR system design and performance characteristics include a reduction in the physical size of the reading/scanning units, increased plate-reading capacity per unit time, and better image quality. These advances have been achieved through a combination of changes in the imaging plates themselves, in the image reader or scanning devices, and in the application of image processing algorithms to affect image output.

The newer imaging plates developed for the latest CR devices have higher image quality (increased sharpness) and improved fading and residual image characteristics. Higher image quality has resulted from several modifications

in the imaging plate phosphor and layer thickness. Smaller phosphor grain size in the IP (down to approximately 4 μm) diminishes fixed noise of the imaging plate, whereas increased packing density of phosphor particles counteracts a concomitant decrease in photostimulable luminescence (21). A thinner protective layer is used in the plates tending to reduce X-ray quantum noise and, in and of itself, would improve the spatial resolution response characteristics of the plates as a result of diminished beam scattering. However, in the newest IPs, the quantity of phosphor coated onto the plate is increased for durability purposes, resulting in the same response characteristic of previous imaging plates (27).

An historical review of CR scanning units chronicles improved compactness and increased processing speed. The first Fuji unit (FCR 101) from 1983 required roughly 6 m^2 of floor space to house the reader and could only process about 45 plates per hour, whereas today's Fuji models as well as other vendor's devices occupy less than 1 m^2 and can process over 110 plates per hour, which represents a decrease in apparatus size by a factor of approximately one-sixth and an increase in processing capacity of roughly 2.5 times. Desktop models reduce the physical device footprint even further.

CR imaging plate sizes, pixel resolutions, and their associated digital file sizes are roughly the same across manufacturers for the various cassette sizes offered. For example, the 14″ by 17″ (or 35 cm by 43 cm metric equivalent) plates are read with a sampling rate of 5–5.81 pixels per mm, at a digital image matrix size of roughly 2 k by 2 k pixels (1760 by 2140 pixels for Fuji (21) and 2048 by 2508 pixels for Agfa and Kodak (16). Images are typically quantized to 12 bits (for 4096 gray levels). Thus, total image file sizes range from roughly 8 megabytes (MB) to 11.5 MB. The smaller plates are scanned at the same laser spot size (100 μm), and the digitization rate does not change; therefore, the pixel size is smaller (16). The 10″ by 12″ (24 cm by 30 cm) plates are typically read at a sampling rate of 6.7–9 pixels per millimeter (mm) and the 8″ by 10″ (18 cm by 24 cm) plates are read at 10 pixels per mm (16,21).

Cassetteless CR devices have been introduced in which the detector is incorporated into a chest unit, wall, or table buckey to speed throughput and facilitate workflow much like DR devices do. Dual-sided signal collection capability is available by Fuji, increasing overall signal-to-noise. Agfa has shown a product in development (ScanHead CR) that stimulates and reads out the imaging plate line-by-line, as opposed to the point-by-point scanning that occurs in most CR devices today. Increased speed (5 s scan time) and higher DQE have been demonstrated. In addition, needle phosphors have been explored as a possible replacement to powder phosphors, having shown improved spatial resolution and DQE.

Image Processing Algorithms. Image processing is performed to optimize the radiograph for output display. Each manufacturer has a set of proprietary algorithms that can be applied to the image for printing on laser film or display initially only on their own proprietary workstations. Prior to the DICOM standard, only the raw data could be directly acquired digitally. Therefore, to attain the same image appearance on other display stations, the appropriate image processing algorithms (if known) had to be implemented somewhere along the chain from acquisition to display. Now image processing parameters can be passed in the DICOM header and algorithms applied to CR images displayed on generic workstations. Typically, however, advanced real-time manipulation of images can only be done on each manufacturers' specific processing station. In general, the digital image processing applied to CR consists of a recognition or analysis phase, followed by contrast enhancement or frequency processing. Note that the same general types of image processing applied to CR can also be applied to DR images.

Image Segmentation. In the image recognition stage, the region of exposure is detected (i.e., the collimation edges are detected), a histogram analysis of the pixel gray values in the image is performed to assess the actual exposure to the plate, and the appropriate look-up table specific to the region of anatomy imaged and chosen by the X-ray technologist at the time of patient demographic information input is selected. Proper recognition of the exposed region of interest is extremely important as it affects future processing applied to the image data. For example, if the bright-white area of the image caused by collimation at the time of exposure is not detected properly, its very high gray values will be taken into account during histogram analysis, increasing the "window" of values to be accommodated by a given display device (softcopy or hardcopy). The effect would be to decrease the overall contrast in the image.

Some segmentation algorithms, in addition to detection of collimation edges in the image, enable users to blacken the region outside these edges in the final image if so desired (16,28), which tends to improve image contrast appearance by removing this bright-white background in images of small body parts or pediatric patients. The photo in Fig. 4B demonstrates this feature of "blackened surround," as applied to the image in Fig. 4A.

Contrast Enhancement. Conventional contrast enhancement, also called gradation processing, tone scaling, and latitude reduction, is performed next. This processing amounts to choosing the best characteristic curve (usually a nonlinear transformation of X-ray exposure to image density) to apply to the image data. These algorithms are quite flexible and can be tuned to satisfy a particular user's preferences for a given "look" of the image (29). Look-up tables are specific to the region of anatomy imaged. Figure 5 shows an example of the default adult chest look-up table (Fig. 5a) applied to an image and the same image with high contrast processing (Fig. 5b). A reverse-contrast scale or "black bone" technique, in which what was originally black in the image becomes white and what was originally white in the image becomes black, is sometimes felt to be beneficial for identifying and locating tubes and lines. An example is shown in Fig. 6 where the contrast reversal algorithm has been applied to the image in Fig. 6a, resulting in the image in Fig. 6b.

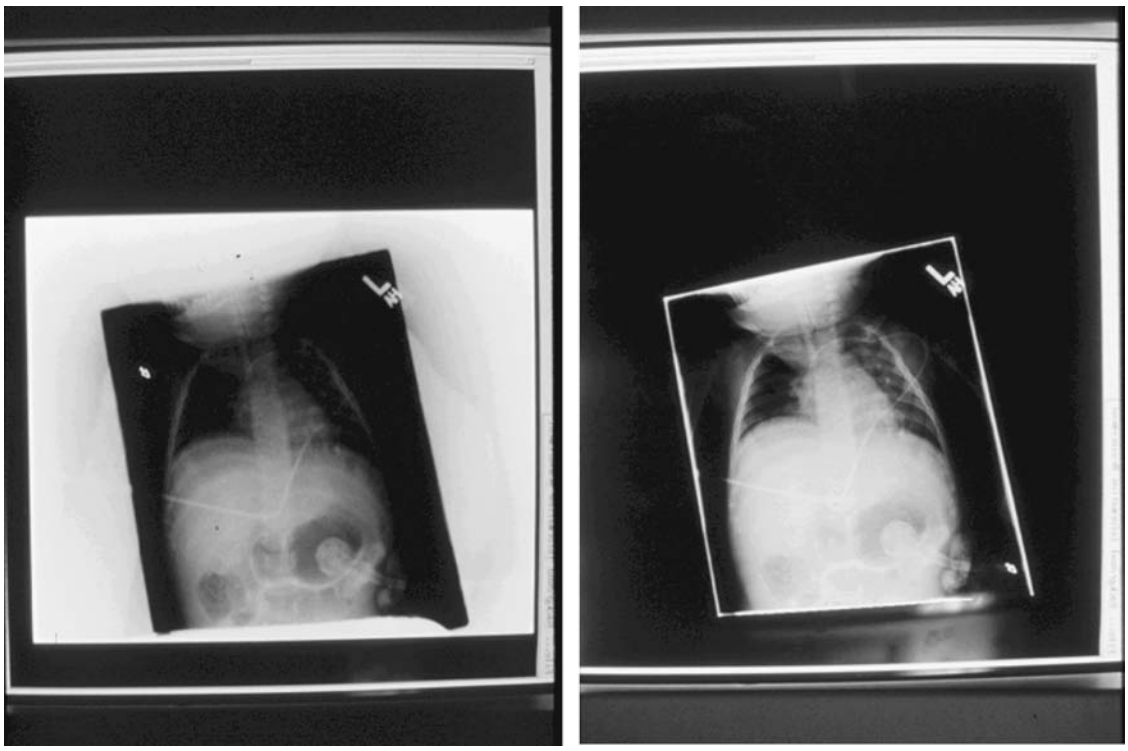


Figure 4. Example image segmentation algorithm detection of (white) collimation edges of exposure region in image **A**, with "blackened surround" applied in image **B**. Note the improved overall contrast in the image in **B**.

Spatial Frequency Processing. The next type of image processing usually performed is spatial frequency processing, sometimes called edge enhancement. These algorithms adjust the frequency response characteristics of the CR systems essentially implementing a high- or band-pass filter operation to enhance the high spatial frequency content contained in edge information. Unfortunately, noise also contains high spatial frequency information and can be exacerbated by edge enhancement techniques. To lessen this problem, a nonlinear unsharp masking technique is typically implemented serving to suppress noise via a smoothing process. Unsharp masking is an averaging technique that, via summation, tends to blur the image. When this result is subtracted from the original image data, the effect is one of noise suppression. Specific spatial frequencies can be preferentially selected and emphasized by changing the mask size and weighting parameters. For example, low spatial frequency information in the image can be augmented by using a relatively large mask, whearas high spatial frequency or edge information can be enhanced by using a small mask size (16).

Dynamic Range Control. An advanced algorithm by Fuji, for selective compression or emphasis of low density

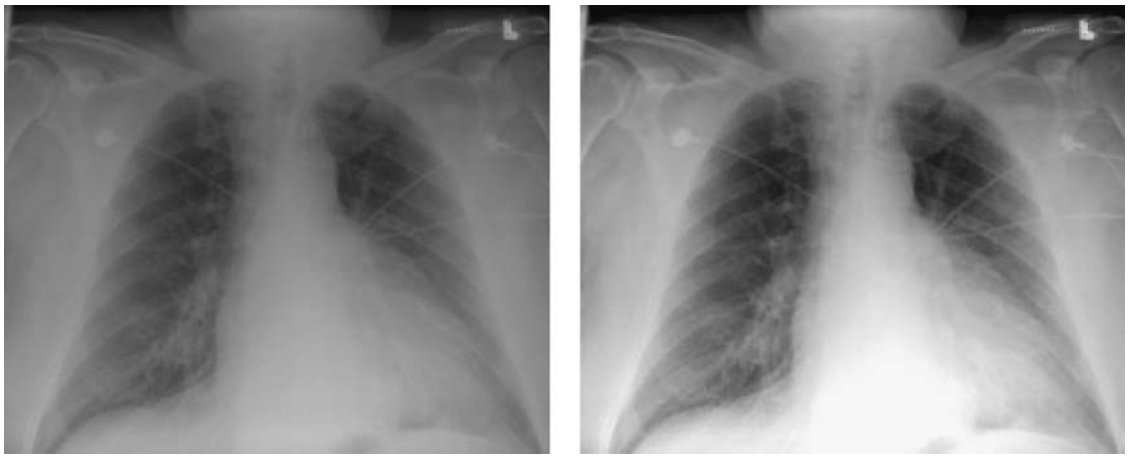


Figure 5. Chest image processed with **A**. default mode and **B**. high contrast algorithm applied.

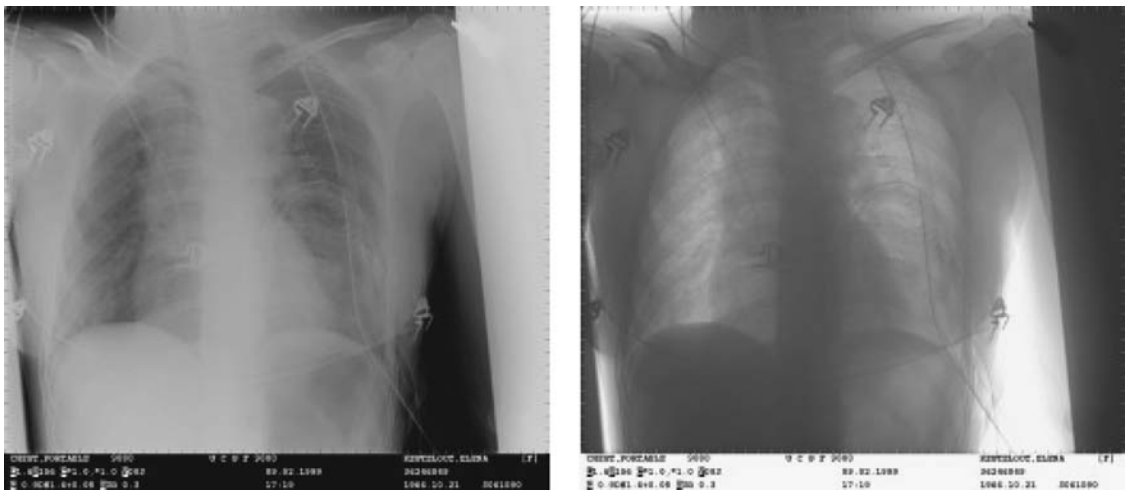


Figure 6. Chest image processed with **A**. default mode and **B**. blackbone or contrast reversal algorithm applied.

regions in an image, independent of contrast and spatial frequency is known as dynamic range control (DRC) processing (30). The algorithm consists of performing an unsharp mask for suppression of high spatial frequency information, then application of a specific look-up table mapping to selected regions (i.e., low density areas). This mask is then added back to the original data with the overall result being improved contrast in poorly penetrated regions, without loss of high frequency and contrast emphasis. In a clinical evaluation of the algorithm for processing of adult portable chest exams, DRC was found to be preferred by five thoracic radiologists in a side-by-side comparison, providing improved visibility of mediastinal details and enhanced subdiaphragmatic regions (31).

Multiscale Image Contrast Amplification. Multiscale image contrast amplification (MUSICA) is a very flexible advanced image processing algorithm developed by Agfa (26,32). MUSICA is a local contrast enhancement technique based on the principle of detail amplitude or strength and the notion that image features can be striking or subtle, large in size or small. MUSICA processing is independent of the size or diameter of the object with the feature to be enhanced. The method is carried out by decomposing the original image into a set of detail images, where each detail image represents an image feature of a specific scale. This set of detail images or basis functions completely describes the original image. Each detail image representation and the image background are contrast equalized separately; some details can be enhanced and others attenuated as desired. All the separate detail images are recombined into a single image, and the result is diminished differences in contrast between features regardless of size, such that all image features become more visible.

Image Artifacts. The appearance and causes of image artifacts that can occur with CR systems should be recognized and corrected. Artifacts can develop from a variety of sources, including those related to the imaging plates themselves, to image readers, and to image processing. Several types of artifacts potentially encountered with CR have been minimized with the latest technology improvements but may still be seen in older systems.

Lead backing added to the aluminum-framed, carbon-fiber cassettes has eliminated the so-called light-bulb effect, darkened outer portions of a film due to backscattered radiation (33). High sensitivity of the CR plates renders them extremely susceptible to scattered radiation or inadvertent exposure, thus routine erasure of all CR plates on the day of use is recommended as is the storing of imaging plates on end, rather than stacking of cassettes one on top of the other (34). The occurrence of persistent latent images after high exposures or after prolonged intervals between plate erasure and reuse (33,35) has been lessened by the improved efficiency of the two-stage erasure procedure used in the latest CR systems (34). Improved recognition of the collimation pattern employed for a given image allows varied (including off-angle) collimation fields and in turn, improves histogram analysis and subsequent processing of the imaged region (34), although these algorithms can fail in some instances. Plate cracking, from wear-and-tear, can create troublesome artifacts as depicted in Volpe (34).

Inadvertent double exposures can occur with the present CR systems, potentially masking low density findings, such as regions of parenchymal consolidation, or leading to errors in interpreting line positions. Such artifacts are more difficult to detect than with screen-film systems because of CR's linear frequency processing response, optimizing image intensity over a wide range of exposures (i.e., due to its wide dynamic range). Figure 7 shows an example double-exposure artifact, and additional examples are included in Volpe (34). Laser scanning artifacts can still occur with current CR readers and are seen as a linear artifact across the image, caused by dust on the light source (34). Proper and frequent cleaning of the laser and light guide apparatus as well as the imaging plates themselves can prevent such artifacts.

The ability of CR to produce clinically diagnostic images over a wide range of exposures is dependent on

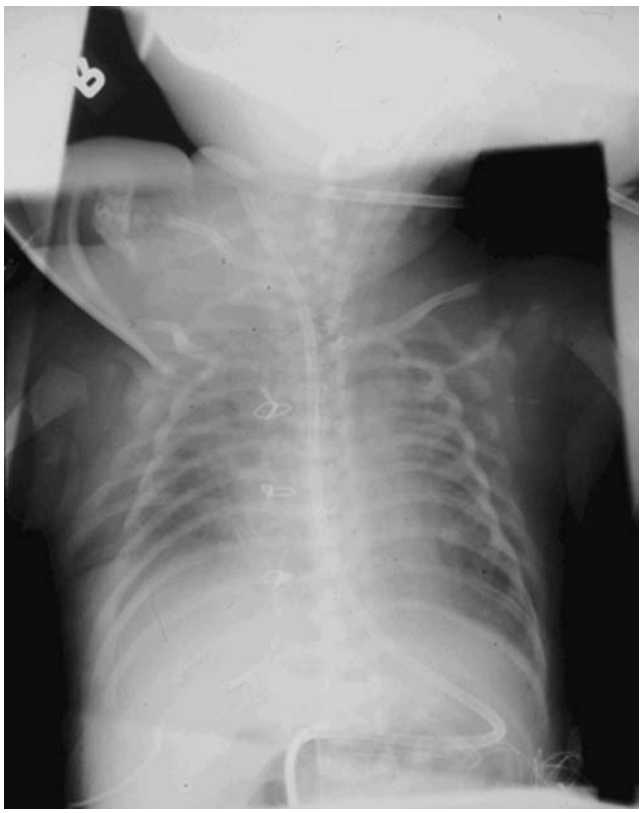


Figure 7. Example inadvertent double exposure.

the effectiveness of the image analysis algorithms applied to each dataset. The specific processing parameters used are based on standards tuned to the anatomic region under examination. Incorrect selection of diagnostic specifier or inappropriate anatomic region can result in an image of unacceptable quality. Understanding the causes of some of these CR imaging artifacts described here, as well as maintaining formal, routine quality assurance procedures, can help to recognize, correct for, and avoid future difficulties.

Summary of CR. CR can be used for the digital image acquisition of projection radiography examinations into a PACS. As a result of its wide exposure latitude and relative forgiveness of exposure technique, CR can improve the quality of images in difficult imaging situations, such as in portable or bedside examinations of critically ill or hospitalized patients. As such, CR systems have been successfully used in the ICU setting, in the emergency room (ER) or trauma center, as well as in the operating room (OR). CR can also be cost-effective for a high volume clinic setting, or in a low volume site as input to a teleradiology service, and have successfully reduced retake rates for portable and other examinations.

Technological advances in CR hardware and software have contributed to the increased acceptance of CR as a counterpart to conventional screen-film projection radiography, making the use of this modality for clinical purposes more widespread. CR is compatible with existing X-ray equipment, yet separates out the functions of image acquisition or capture, image display, and image archival versus traditional screen-film, in which film serves as the image detector, display, and storage medium. This separation in image capture, display, and storage functions by CR enables optimization of each of these steps individually. Potential expected benefits are improved diagnostic capability (via the wide dynamic range of CR and the ability to manipulate the exam through image processing) and enhanced radiology department productivity (via networking capabilities for transmission of images to remotely located digital softcopy displays and for storage and retrieval of the digital data).

Digital Radiography (DR). In addition to CR devices for digital image acquisition of projection X rays are the maturing direct digital detectors falling under the general heading of digital radiography (DR). Note that digital mammography is typically done using DR devices, although CR acquired at much higher sampling matrices has also been tested.

Unlike conventional screen-film radiography in which the film functions as the imaging sensor or recording medium as well as the display and storage media, DR, like CR, eliminates film from the image recording step, resulting in a separation of image capture from image display and image storage. This separation of functions potentiates optimization of each of these steps individually. In addition, DR, like CR, can capitalize on features common to digital or filmless imaging, namely the ability to acquire, transmit, display, manipulate, and archive data electronically, overcoming some of the limitations of conventional screen-film radiography. Digital imaging benefits include remote access to images and clinical information by multiple users simultaneously, permanent storage and subsequent retrieval of image data, expedient information delivery to those who need it, and efficient cost-effective workflow with elimination of film from the equation.

Review of the Fundamentals.

Process Description. Indirect versus Direct Conversion: DR refers to devices for direct digital acquisition of projection radiographs in which the digitization of the X-ray signal takes place within the detector. Compare this method with CR, which uses a photostimulable phosphor imaging plate detector in a cassette design that must be processed in a CR reader following X-ray exposure, for conversion to a digital image. DR devices, also called flat-panel detectors, include two types, indirect conversion devices in which light is first generated using a scintillator or phosphor and then detected by a CCD or a thin-film-transistor (TFT) array in conjunction with photodiodes; and DDR devices, which consist of a top electrode, dielectric layer, selenium X-ray photoconductor, and thin-film pixel array (36). Figure 8 shows a comparison of the direct and indirect energy conversion steps in the production of a digital X-ray image. DDR devices offer direct energy conversion of X ray for immediate readout without the intermediate light-conversion step.

The basis of DR devices is the large area thin-film-transistor (TFT) active matrix array, or flat panel, in which each pixel consists of a signal collection area or charge

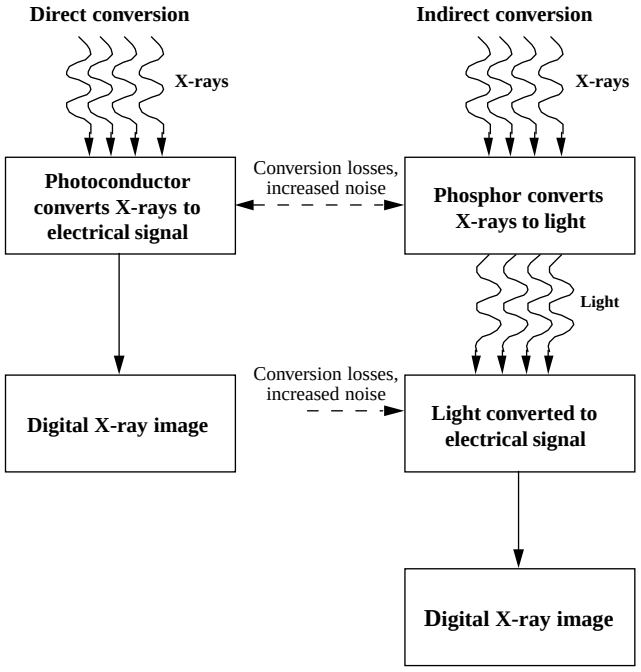


Figure 8. The image production steps involved in direct and indirect digital radiography detectors.

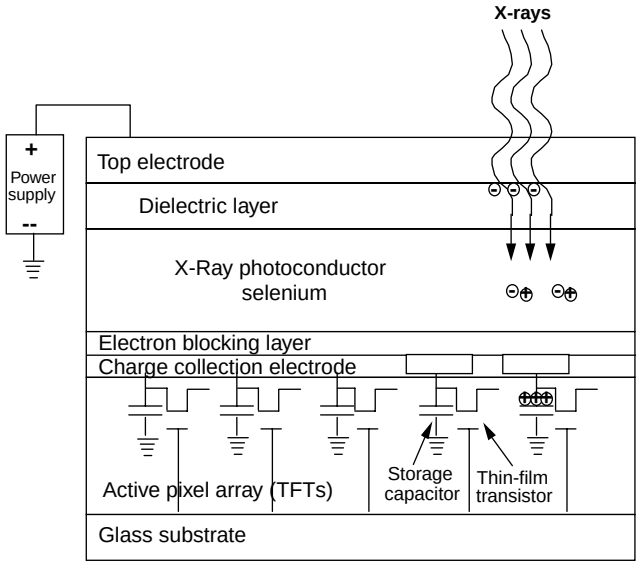


Figure 9. Cross-sectional view of an example direct digital radiography (DDR) detector panel.

collection electrode, a storage capacitor, and an amorphous silicon field-effect transistor (FET) switch that allows the active readout of the charge stored in the capacitor (36). Arrays of individual detector areas are addressed by orthogonally arranged gate switches and data lines to read the signal generated by the absorption of X rays in the detector. The TFT arrays are used in conjunction with a direct X-ray photoconductor layer or an indirect X-ray-sensitive phosphor-coated light-sensitive detector or photodiode array.

An example DDR device, diagramed in cross section in Fig. 9 (36), uses a multilayer detector in a cassette design, in which the X-ray energy is converted directly to electron-hole pairs in an amorphous selenium (Se) photoconductive conversion layer. Charge pairs are separated in a bias field such that the holes are collected in the storage capacitors and the electrons drift toward the Se-dielectric interface. At the end of exposure, the image resides in the pixel matrix in the form of charges, with the charge proportional to the absorbed radiation. At the end of the readout, the charges are erased to prepare for another detection cycle.

An example indirect DR device uses an X-ray-sensitive phosphor coating on top of a light-sensitive flat-panel amorphous silicon (Am-Si) detector TFT array. The X rays are first converted to light and then to a proportional charge in the photodiode [typically a cesium iodide (CsI) scintillator], which is then stored in the TFT array where the image signal is recorded.

System Characteristics. DR detectors have high efficiency, low noise and good spatial resolution, wide latitude, and all the benefits of digital or filmless imaging. Similarly, DR has a very wide dynamic range of quantization to thousands of gray levels. These devices are becoming more widely used clinically and are available in table buckeys as well as chest units. DR units have superior workflow and increased patient throughput due to the elimination of cassette handling (37).

The short imaging cycle time of DR may lend itself to combined static radiographic and dynamic fluoroscopic uses in future applications, which is true especially for the indirect devices. The direct Se detector, for example, has a ghosting problem due to charge trapping, which introduces a lag time at the end of each cycle, lengthening the time between readiness for the next exposure.

The cost of sensor production is still high such that the overall price of devices has not dropped appreciably. DR is sometimes referred to as a one-room-at-a-time technology because the detectors are built into the room and matched to the X-ray source. Detector fragility and poor portability makes DR difficult to use in the bedside X-ray environment, but some portable devices are now being introduced.

Image Quality. DR image quality is comparable with that of CR. However, DR devices have higher DQEs than CR, capturing roughly 80% absorption of the X-ray energy at optimum exposures. Thus, DR is a higher efficiency, low noise detector, converting much of the incoming X-ray signal into useful output. Several recent studies have demonstrated high image quality at lower radiation dose to the patient. The ability to lower exposure would be a significant advantage for DR. A factor limiting DR efficiency involves the packing fraction, or active detector area to dead space taken up by the data readout devices (transistors, data lines, capacitors, etc.). As the physical size of the data readout components is currently fixed, the smaller the pixel size, the smaller the packing fraction, with a larger proportion of dead area overwhelming the active area, in some cases reducing the active area to 30% or less (36). The overall effect is a reduction in geometric and quantum efficiency.

The spatial resolution of DR is comparable with CR, which is still less than that for analog X ray. Typical matrix sizes are on the order of 2000 to 2500 pixels by 2000 to 2500 pixels. The pixel size of the TFT array detector is the limiting factor for spatial resolution, with the direct Se detector yielding a better inherent spatial resolution than indirect detectors, which can lead to better signal modulation and superior contrast.

DR design presents a delicate tradeoff between detector efficiency, inversely proportional to pixel size, and spatial resolution, affected directly by pixel size. Typically, DR devices are specified for higher detection efficiency at a cost of less spatial resolution than screen-film, with compensation by a wide dynamic range or high contrast resolution. Design complexities requiring further development include wiring configurations to minimize dead space and maximize the detector packing fraction, fast and robust signal readout methods, and better error-correction matrices for more accurate signal readout.

Comparison of CR and DR. Table 2 lists the advantages of CR and DR, including all the benefits of digital images that can be electronically processed, manipulated, distributed, displayed, and archived. The superior contrast resolution of the digital modalities can compensate, in many cases, for the lesser spatial resolution as compared with screen-film. Both CR and DR can be used for the digital image acquisition of projection radiography examinations into a PACS.

As for any digital image acquisition device, CR or DR would be the first point of entry into a PACS. Errors may propagate from here, with the quality of the PACS output being directly dependent on the quality of the signal in. In addition to image quality, essential features for successful clinical implementation of CR or DR systems for a PACS include the following. DICOM conformance of the modality is essential and includes compliance with the image data and header format, as well as the DICOM communication protocol. Equally critical is interfacing to the RIS-HIS. Integration of the CR/DR system with the RIS-HIS can reduce human errors on patient demographic information input and improve efficiency. Ease of integration of the device into the daily workflow routine, and simplicity and robustness of the user interface are very important. Reliability, fault-tolerance, and capabilities for error tracking are also major issues to consider, as are device speed and performance.

As a result of CR's convenient workflow and portability, as well as its wide exposure latitude and relative forgiveness of exposure technique, CR can improve the quality of images in difficult imaging situations, such as in portable or bedside examinations of critically ill or hospitalized patients, and enable decreased retake rates for improper exposure technique. As such, CR systems have been successfully used in the ICU setting, in the ER or trauma center, as well as in the OR. CR can also be cost effective for a high volume clinic setting, or in a low volume site as input to a teleradiology service. Cost savings and improved radiology departmental workflow can be realized with CR and the elimination of film (37).

Table 2. Summary of Advantages of CR and DR Systems

- Produce digital images capable of being electronically processed, manipulated, distributed, displayed, and archived.
- Large latitude systems allowing excellent visualization of both soft tissue and bone in the same exposure image.
- Superior contrast resolution can compensate for lack of spatial resolution.
- Decreased retake rates.
- Potential cost savings if film is eliminated.
- Improved radiology department workflow with elimination of film handling routines.

Table 3. Summary of Future Trends in Image Acquisition

Image Matrix Size	↑
Image Quality	↑↑
Spatial Resolution	↑
# Image Slices	↑↑↑
Size of Imaging Examinations	↑↑↑
Size of Devices	↓↓
Portability of Devices	↑
Cost of Devices	↓
% of Image Devices that are Digital	↑↑↑
% of Image Acquisition that is Digital (Elimination of Film)	↑↑

Technological advances in CR hardware and software have contributed to the increased acceptance of CR as the current counterpart to conventional screen-film projection radiography, making the use of this modality for clinical purposes more widespread. CR is compatible with existing X-ray equipment, yet separates out the functions of image acquisition or capture, image display, and image archival versus traditional screen-film, in which film serves as the image detector, display, and storage medium. This separation in image capture, display, and storage functions by CR enables optimization of each of these steps individually. Potential expected benefits are improved diagnostic capability (via the wide dynamic range of CR and the ability to manipulate the data through image processing) and enhanced radiology department productivity (via networking capabilities for transmission of images to remotely located digital softcopy displays and for storage and retrieval of the digital data).

DR devices have more efficient detectors, offering direct energy conversion of X ray for immediate readout. The higher DQE may enable DR to produce high quality images at a lower radiation dose to the patient. These detectors have low noise and good spatial resolution, wide latitude, and all the benefits of digital or filmless imaging. However, cost is still high because detector production is difficult and expensive, and DR is a one-room-at-a-time detector. DR may be cost-effective in high volume settings with constant high patient throughput (37).

However, meeting the cost competitiveness of screen-film systems is difficult unless film printing is eliminated from the cost equation. DR may be preferable for imaging examinations requiring very high quality, such as in mammography, upright chest exams and bone work. DR devices integrated into table and wall buckeys are now making these devices highly efficient for emergency department trauma cases.

Future improvements in image processing algorithms, with a better understanding of optimum display settings for soft copy viewing, have the potential to greatly facilitate and standardize softcopy reading of digital projection radiographs, and further the acceptance of CR and DR in the clinical arena. It is likely that CR and DR devices will coexist for some time.

Future Trends in Image Acquisition. Although the types of imaging modalities will probably not change all that much in the next several years, the anticipated future trends in image acquisition for digital radiology and PACS include changes in the image dataset sizes, changes in the imaging devices themselves, and improvement in image processing for softcopy display of digital images.

Image Data Sets. No new types of imaging modalities are foreseen for the near future. However, it is anticipated, and has to a certain extent already begun, that the image datasets acquired from the existing modalities will increase in overall study file size, in some cases dramatically. For example, many radiology departments have begun installing multiple detector array or multislice CT scanners that tend to generate a greater number of individual images than do the single detector array scanners because the slice thickness in helical acquisition ($\sim$ 0.75 mm) versus the single detector arrays ($\sim$ 7–10 mm), and the clinical imaging protocols used, as well as the increasing clinical utility of 3D image display representations.

Image matrix sizes for the digital projection radiography devices (CR and DR) have gone up from roughly from one and two thousand square matrices to four by five thousand pixels squared for mammography applications. The increased sampling was done to improve the spatial resolution. Most laser film digitizers can now vary their spot sizes from 200 μm down to 50 μm, greatly improving the inherent spatial resolution of the resulting images of the scanned analog film, with a concomitant increase in file size.

The bit depth representation of gray-scale pixel values has also increased from 8 bits to 10, 12, and 16 bits, and color images are stored as 32 bit or 4 byte per pixel data files. Further, the addition of post-processing results or slice reconstructions, and cinegraphic sequences to the image dataset, while improving the overall quality of the image, may greatly increase the amount of data to be acquired into a PACS.

Devices. While image datasets and file sizes are getting larger, the imaging devices themselves will continue to get smaller in physical footprint, which has been seen most dramatically with the CR devices, going from requiring roughly 36 m^2 of floor space and special electrical power and cooling, to desktop devices that can be placed in most any location. CT and MRI devices are also becoming smaller in size, more portable, and more robust. Hopefully, these devices will continue to become less expensive. Terahertz imaging currently used in aerospace applications may become developed for uses in imaging humans for medical purposes. These devices acquire images at 0.25 and 0.3 THz, creating a binary (two-color) picture to contrast between materials with different transmission and reflection properties. The main advantage of a terahertz imager is that it does not emit any radiation and it is a passive camera, capturing pictures of the natural terahertz rays emitted by almost all objects.

Image Processing. An important area of increased attention continues to be image processing capabilities for softcopy image display. Future processing techniques will most likely go above and beyond the simple window and level (or contrast and brightness) manipulation techniques. These post-processing algorithms are currently available and tunable at the imaging modality, or accompanying modality acquisition workstation, but may, in time, be manipulable in real-time at the display station. Stand-alone 3D workstations are becoming more common. Efforts to embed advanced processing and visualization in the PACS workstation will ultimately allow real-time processing to be performed by the radiologist or even the referring clinician.

Image compression is currently being debated, but may, in time, be available at the modality to reduce image transmission time and archival space. Some techniques, such as the wavelet transform, may become more widely used not only as a compression technique, but also for image enhancement at the imaging devices.

In time, it is anticipated that the percentage of all imaging devices used by health-care enterprises that are digital in nature will increase greatly. Further, the percentage of digital image acquisition from the devices that are capable should increase, decreasing the amount of film used as an acquisition, display, and archival medium.

Medical Image Archival

Digital image archives were once thought of as costly inefficient impediments to moving toward PACS and digital imaging departments (38). However, current trends in archival technology have shown the cost of digital storage media decreasing steadily with capacity increasing, whereas analog devices such as paper and film continue to increase in overall cost (39). Improvements in storage devices along with the use of intelligent software have removed digital archives as a major stumbling block to implementing PACS. The following tutorial on electronic archival technologies for medical images includes a discussion of available digital media, PACS system architectures, and storage management strategies.

Digital image archival can be more efficient than the manual data storage of the traditional film file room. A study of image examination retrieval from a PACS versus a film-based system showed statistically significant reduction in times for the digital method, in many cases down from hours to minutes (40). The improved retrieval times with PACS were particularly striking for studies between six months and one year old, and for studies greater than one year (40).

An informal survey of 75 radiologists operating in a traditional film-based radiology department found that 70% experienced delayed access to films, which caused them and their staff money in terms of decreased efficiency

(41). Rarely did this delayed access to films result in repeated or unnecessary studies, or result in longer hospital stays. However, inaccessible or lost films did result in time spent, often by the radiologist or clinician, looking for films.

Digital archives are generally less people-intensive, eliminating the physical handling of films, and are, therefore, less expensive and less subject to the errors in filing and lost films that often plague film stores. Electronic archives can improve the security of stored image data and related records, assuring no loss of exam data while offering simultaneous case availability to many.

Digital archives must have an intelligent patient-centric system database interface to enable easy retrieval of imaging examinations. They should conform to the DICOM standard format and communications protocol by being able to accept and return DICOM format files. Many archive systems reformat the data once inside the storage device to a more efficient schema appropriate for the specific archive architecture.

Medical image data files are large compared with text-based clinical data, and are growing in size as new digital applications prove clinically useful. A single view chest X ray, for example, requires approximately 10 MB of storage space. With the expanding availability of multi-detector CT scanners and increasing use of magnetic resonance angiography examinations, thousand-slice studies are not uncommon. Imaging activity continues to increase significantly as it becomes a key diagnostic triage event, with most diagnostic imaging departments showing an increase in overall volume of cases. A typical 500 bed health-care enterprise performing approximately 200,000 examinations, for example, can generate on the order of 5–6 terabytes (TB) of data per year (42).

Compression can be used to reduce both image transmission time and storage requirements. Note that compression that can be achieved via hardware or software also occurs clinically (i.e., not all images of a study are filmed). Lossless (or bit-preserving) compression at 2:1 is done by most PACS archive systems already. Lossy or non-bit-preserving compression, by definition, does not provide an exact bit-for-bit replica of the original image data on decompression. However, studies have shown that numerically lossy compression can produce visually lossless images at compression ratios of 5:1 to 30:1 depending on modality (43–45). Compression at these levels can achieve much greater space savings and appear to be of adequate image quality for image comparison and review of prior studies. For perspective, without compression only 50 two-view digital projection X-ray examinations at approximately 10 MB per image can be stored on a 1 gigabyte (GB) disk. With compression at 25:1, approximately 1250 examinations can be stored on a 1 GB disk.

Digital Archival Media. The digital storage device media used in PACS today include computer hard drives or magnetic disks, RAID disks (redundant array of inexpensive disks), optical disks (OD), magneto-optical disks (MOD), and tape. Newer technologies such as digital video disks (DVD) and ultra-density optical (UDO) disks are being introduced. The properties and attributes of each storage device, including capacity, performance, and relative cost, are compared and summarized below. Figure 10 graphs the capacity versus performance characteristics of the various digital archive media widely used in PACS. Table 4 summarizes the relative cost of digital archive media per capacity or number of GB that can be stored per dollar, in addition to typical capacities and retrieval times.

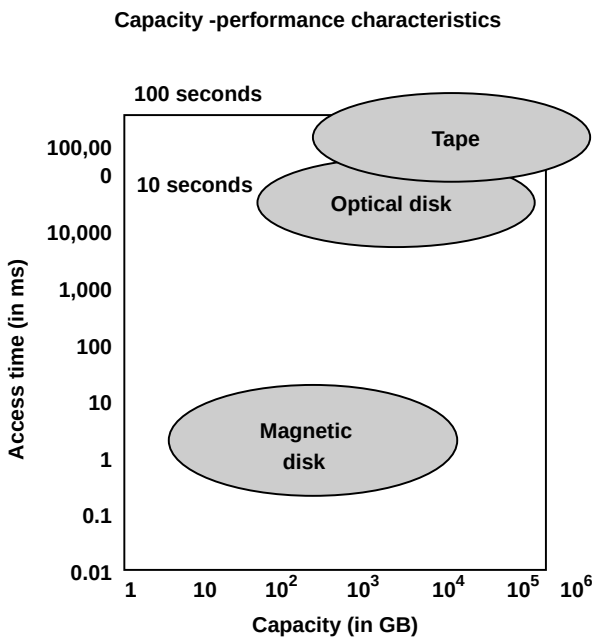


Figure 10. Graph of capacity versus performance characteristics of digital archive media.

Magnetic Disk (MD). The standard computer hard drive or magnetic disk (MD), also known as a direct access storage device (DASD), is the fastest medium from which to retrieve data. Retrieval times are on the order of 1 to 50 milliseconds (ms). However, MDs have the lowest capacity, typically hundreds of MB to hundreds of GB, and the highest cost per amount of data storage of all the archival media used in PACS today, although prices are continuing to decrease rapidly. As a result of the relative high cost in the past, MDs have historically been used for online local storage at the display workstation where fast access was required, yet large capacity was not cost-effective. Today, it is becoming cost-effective to use spinning media for all stages of storage – the trend toward everything-on-line.

Redundant Array of Inexpensive Disks (RAID). RAID devices consist of multiple MDs with high performance and larger capacity (now TB worth per device). These devices can offer redundancy, lessening the concerns with a single point of failure, and have hot-swapable components that can be replaced as needed without bringing the entire archive system down.

RAID has traditionally been used for "near line," intermediate short-term storage, or clinical-operational storage cache to minimize the number of transactions hitting the deep or long-term archive. It is becoming cheap enough per

Table 4. Digital Archive Media Capacity, Retrieval Times, and Relative Costs per GB of Storage

Archive Media Type	Storage Capacity	Performance Retrieval Times	Cost per Capacity (in Order of $ per GB)
Magnetic Disk	100s MB–10s GB	1 to 50 ms	$1.00/GB
Optical Disk	1 – 10s GB for TB devices	s to min	$0.40/GB
Tape	10s – 100s GB for 10s TB devices	10s s to min	$0.25/GB
RAID	10s–100s GB for 10s TB devices	100–300 ms	$10.00/GB
DVD	GB for TB devices	s	$2.50/GB
UDO	30 GB for 10s – 100s TB devices	s to min	$2.00/GB

capacity to consider using RAID in larger configurations for high performance longer-term storage. In these configurations, a higher percentage of studies, perhaps accounting for several years or more, can remain available online for immediate access.

Optical/Magneto-optical Disk (OD/MOD). Optical disks and magneto-optical disks are of the removable spinning storage media class typically stored in an automated media movement device or jukebox giving them total device storage amounts equal to hundreds of times the media capacity. ODs and MODs have higher capacity than MDs, typically GB to tens of GB yielding hundreds of GB to tens of TB total device storage. They are lower cost per capacity than RAID, on the order of a half dollar per GB of storage. Optical disks are a slower medium than RAID, on the order of seconds to minutes for data retrieval in batch and random seek modes.

ODs are also known as WORM or write once, read many disks with data permanently written onto the disks. MODs are erasable reusable platters and are able to hold more data per unit than ODs. As a result of the slower performance and lower cost per capacity, ODs and MODs have traditionally been used for long-term permanent PACS storage.

Tape. Tape is also a removable storage medium typically kept in a jukebox or tape library. The magnetic tape type most often used for PACS is digital linear tape (DLT). It has very high capacity, tens to hundreds of GB for many TB per tape library, and low cost on the order of a quarter dollar or less per GB of storage. It is, however, a slower medium than MOD, for example, in random retrieval times because of its sequential nature. Tape performance is competitive with MOD for retrievals of large files, however, using very high batch read-write rates. Even random retrievals of very large files (on the order of 50 MB) can be transferred faster with DLT than with MODs. Tape has historically and is currently being used for disaster backup as well as for long-term permanent storage.

Newer Technologies

Digital Video Disk (DVD). Newer technologies, such as DVDs, appear promising but have failed to move significantly into the medical arena due to their high cost and slow performance. DVDs use dual-sided storage, thus achieving greater amounts of storage (GB) than MODs fairly inexpensively. However, the cost of the drives still remain quite high and the lack of a universal standard read-write format currently limits the use of DVDs for PACS, although the DICOM standard is currently addressing this issue.

Ultra-Density Optical (UDO). Recently released ultra-density optical (UDO) disks use a blue laser recording technology to achieve much greater data storage densities, on the order of 30 GB capacity per disk, predicted to double within six months. UDO is a WORM disk technology with a 50 year lifespan and a current cost of approximately $2 per GB. Although just a first-generation device release in the medical arena (other fields including the defense industry have used UDOs), it may prove to be a useful technology for PACS.

A summary of capacity, performance, and relative cost of the types of digital archive media available today for PACS is given in Table 4. Figure 10 graphs the capacity versus performance characteristics of the MD, OD, and tape. Note that tape and OD are relatively similar in their tradeoff between capacity and performance.

Archival Strategies

Data Migration. Note that medical images have a life cycle in which, early on, quick access to the data is critical and is often needed by several people in many different locations simultaneously. After a patient has been treated and discharged, however, that same imaging study may rarely need to be accessed again, and if it is, taking minutes or even hours to retrieve it may be acceptable. This pattern of use suggests that hierarchical or staged archival strategies can be implemented for optimum cost-effective use of storage technologies, particularly for the older distributed PACS architectures.

The stages or terms of storage include online or local storage, short- or intermediate-term near-line storage, long-term or offline storage, and disaster recovery or backup storage. Online storage contains information that must be available to the user immediately at the display station and, therefore, requires high-speed access. As this performance is costly, online storage is usually reserved for clinically critical data needed during a single episode of current care (i.e., three days for outpatient clinical encounters and six days on average for a typical inpatient stay). The medium best meeting online local storage needs is the standard computer magnetic disk.

Short-term or near-line storage is used to provide relevant prior or historical imaging studies for comparison during a patient's return visit. This method does not require immediate access, particularly if the data can be

automatically prefetched with advanced notice of scheduled appointments. Note that most patients who do not return for continuing care within 18 months of the acute visit are unlikely to return at all. In other words, a large percentage of imaging examinations performed will never be re-reviewed after the original clinical episode, so slower access may be acceptable. As such, RAID devices work well as short-term or near-line storage, although RAID devices have nearly the high performance of a single magnetic disk but are more costly because of the controller and redundancy capabilities built in.

Long-term or permanent storage provides availability to data with advance notice for retrieval, especially when long-term storage is offline or on-the-shelf. Removable storage media devices such as OD, MOD, or tape jukeboxes are typically used for long-term storage due to their high capacity per cost characteristics. Long-term archives must cover an institution's entire medico-legal storage requirements, which vary from state to state (i.e., 5 years for general adult studies, 21 years for pediatric studies, and life for mammograms and images with orthopedic appliances for Massachusetts).

The requirements for fast retrieval of images initially followed by slower retrieval later, if at all, suggests that different types of storage devices could be used over time to archive images with cost savings. As fast retrieval times grow less important, images could be migrated to less costly, higher capacity, slower storage devices, as diagramed in Fig. 11. Software is used to handle the movement of data from one medium to another, and the strategy makes the actual physical storage device transparent to the end user. Such a strategy is known as a hierarchical storage management (HSM) scheme.

Hierarchical Storage Management and Compression. Note that data compression can be used to maximize the amount of online or near-line storage available to a PACS. Although the full resolution image data can be viewed originally for primary diagnosis, a losslessly compressed version can be sent off site to an inexpensive tape backup archive. The original data can also be wavelet lossy compressed and stored on a large RAID device for maximum cost-effective online storage and retrieval of images for review and comparison (44). This scheme is depicted in Fig. 12.

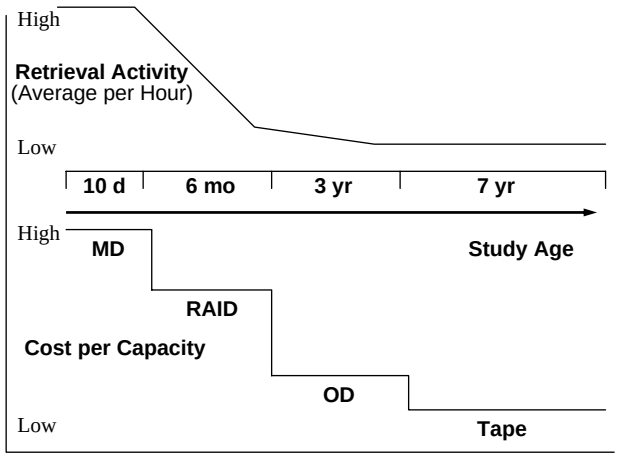


Figure 11. Migration strategy and retrieval requirements versus cost over time.

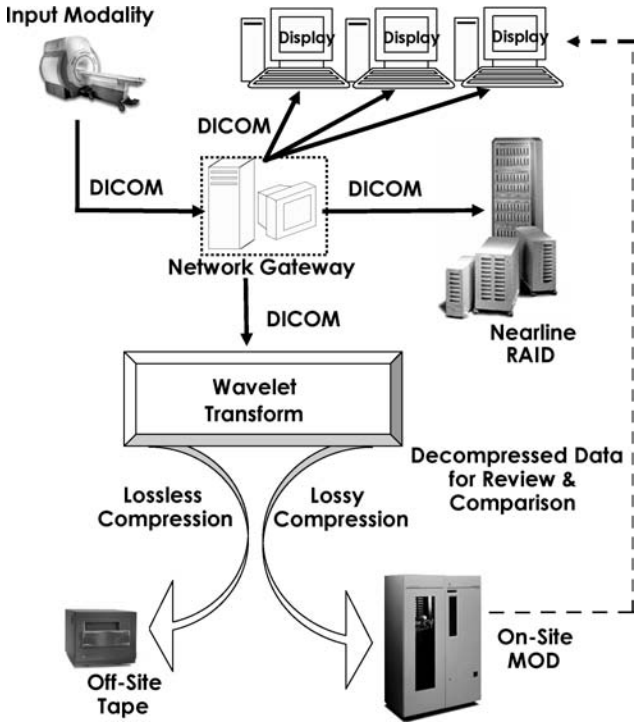


Figure 12. HSM scheme. Image data is viewed at its original content for primary diagnosis, losslessly compressed for the offsite legal record on tape, and wavelet lossy compressed for the onsite near-line storage on MOD for review and historic comparison.

An HSM scheme using short-term archival of uncompressed DICOM data for primary diagnosis, in an onsite RAID, coupled with a very deep long-term archive of diagnostic quality wavelet compressed data in an onsite optical jukebox, cost effectively maximizes online storage for immediate image retrieval. Diagnostically lossy compressed data (at ratios of 25:1 for CR, 10:1 for CT, and 5:1 for MCI) grows the onsite jukebox by 10 times row, depending on the mix of cases, making 20 or more years available online (44), which effectively maintains the entire legal record worth of original plus two relevant prior examinations all online. Note that no large-scale facility has implemented this schema, preferring to accommodate their needs by purchasing more hardware. Also it is unclear what the medico-legal ramifications would be in using lossy compressed images for historical comparison.

A hierarchical storage management scheme such as this provides a solution for maximum intermediate storage and retrieval through the use of onsite lossy compression and offsite tape backup of losslessly compressed data for the legal record and disaster recovery of data. The use of compression in this HSM scheme provides a cost-effective, high performance archive system. This HSM can be tailored to a given health-care enterprise's need to provide clinically and economically beneficial digital archival of medical images.

Other Scalable Solutions: EOL, NAS, SAN, CAS.

Everything-On-Line (EOL). With the dramatic decline in the cost and increase in capacity of RAID devices, it may become feasible to have all studies with their relevant prior examinations accessible online, which is particularly important for centralized or cacheless PACS architectures. On the other hand, imaging volume and study sizes continue to increase and may continue to overburden archive technologies. Thus, perhaps the intelligent use of the hardware and software technologies currently available through data migration schema is a sound strategy.

Networked-Attached Storage (NAS). Networked-Attached Storage (NAS) involves automated storage on a direct-access but separate local drive. NAS uses the existing local area network (LAN) to connect storage devices and the systems requiring data. A NAS server is optimized to perform file-sharing functions without the application processing overhead of typical network file servers, which enables files to be served rapidly to its clients. Performance is affected by the LAN capabilities and system configuration.

Storage Access Networks (SAN). Storage Access Networks (SAN) are dedicated networks that link storage devices and servers, creating an independent directly accessible pool of storage. SANs typically use fiber channel (FC) technology for high speed serial interconnections, usually over optical fiber. This network can provide simultaneous access from one or many servers to one or many storage devices and eliminates potential loading on the LAN.

Content-Addressed Storage (CAS). In Content-Addressed Storage (CAS) systems, data is stored and retrieved based on unique content ID keys. As medical image data is "fixed content" in that its information needs to be stored but cannot (typically) be altered in any way, CAS may prove useful. CAS associates a digital fingerprint, ID, or logical address to a stored element of data providing content security and integrity. The object-oriented nature of CAS could be exploited to improve database searchability.

Application Service Provider (ASP). An Application Service Provider (ASP) approach to medical data archival may be practical for some small entities. This strategy, in which an outside vendor provides services for storage using their hardware and software, has been around for several years. The advantages include less capital requirements for onsite hardware, technology obsolescence protection, maintenance and migration shifted to the ASP vendor, and offsite data backup. Disadvantages include potential vulnerability in performance and long-term viability of the ASP vendor, security issues, and potential high cost of service, particularly for large-volume sites.

Computer Networking

Computer networks enable communication of information between two or more physically distinct devices. They provide a path by which end user radiologists and clinicians sitting at one geographic location, for example, can access radiological images and diagnostic reports from a computer at another location. A private locally owned and controlled network (i.e., within a building or hospital) is called a LAN, whereas a network used outside of a local area is known as a wide area network or WAN. A WAN uses an external service provider and usually has lower bandwidth services than LANs. Intranet communication refers to communication across a private limited-access LAN. Internet communication is across public shared-access WANs.

Signals are transmitted via either bound media such as over cables or unbound broadcast media. Analog communications systems encode information into a continuous wave form of voltage signals, whereas digital systems encode the data into two discrete states or bits, either "0" or "1". The bits are packaged to form bytes, words, packets, blocks, and files based on a specified communication protocol. These communications standards give detailed specifications of the media, the physical connections between devices, the signal levels and timings, the packaging of the signals, and the software needed for the transport of data (46).

Serial data transmission sends digital signals one bit at a time over a single wire; the single bit stream is reassembled at the receiving end of transmission into meaningful byte-word-packet-block-file data (46). Parallel data transmission uses multiple wires to transmit bits simultaneously and, as such, provides increased transmission speeds. Synchronous communication is used in applications that require maximum speed and is carried out between two nodes that share a common clock. Asynchronous communication relies on start and stop signals to identify the beginning and end of data packets (46). An example of this technology is asynchronous transfer mode (ATM) technology.

Hardware. Important networking infrastructure considerations include bandwidth, or how much data can be transferred per period of time; latency, or how long the trip takes; topology or network segmentation, which describes the path data takes; and reliability and redundancy. Table 5 lists different types of network bandwidths or speeds available, in bits per second (bps), along with example transmission times for a single 10 MB CR digital projection radiograph. Note that over a telephone modem it would take approximately 24 minutes at best to transmit a single 10 MB image, whereas over Fast Ethernet it would

Table 5. Network Bandwidths and Example Image Transmission Times

	Maximum Bandwith (bps)	Min Transmission Time for 10 MB CR
Modem	56 kbps	23.8 min
T1 Line	1.54 Mbps	52 s
Ethernet	10 Mbps	8 s
Fast Ethernet	100 Mbps	0.8 s
ATM	155 Mbps	0.52 s
Gigabit Net	1 Gbps	0.08 s

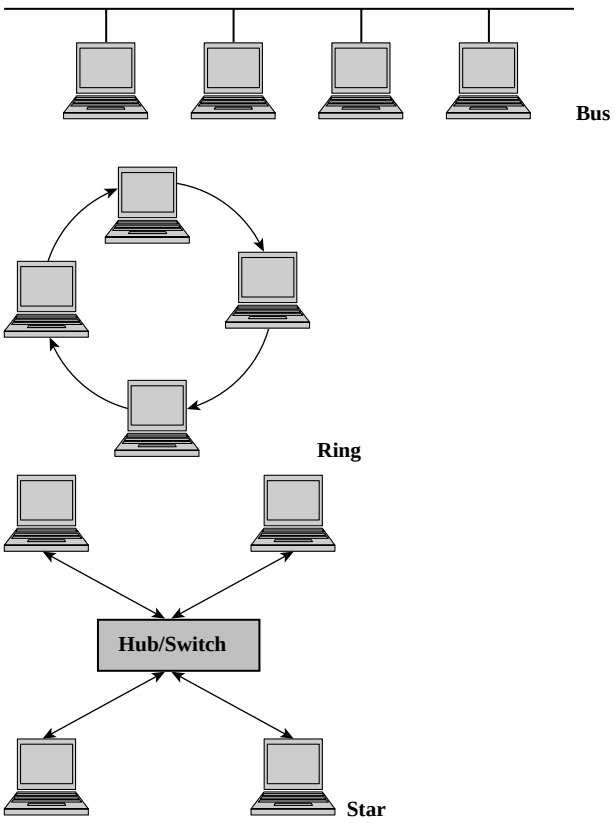


Figure 13. Network topologies typically used in PACS: bus, ring, and star.

be only a fraction of a second. LANs can consist of low speed Ethernet, medium speed Fast Ethernet, or fast speed Gigabit infrastructure. WANs can consist of a range of digital service speeds from slow telephone modem to medium speed T1 lines to fast ATM. Transmission speeds track with cost.

The most common network topologies used in PACS include bus, ring, and star configurations. These are diagramed in Fig. 13. Bus topologies commonly use Ethernet and have the advantage of network simplicity, but the disadvantage of upper-level bottlenecking and difficult to identify channel failure. The ring topology uses fiber distributed data interface (FDDI) or high speed ATM SONET (synchronous optical NETwork) ring technology. Ring topologies offer simplicity and no bottleneck, but, in a single ring, if the channel between two nodes fails, then the network is down. The star or hub topology uses high speed Ethernet or ATM switching and offers network simplicity but a bottleneck as well as a single point of failure at the hub or switch.

The physical media or cabling that makes up PACS networks varies from telephone wires to unshielded twisted pair (UTP) copper cabling, also referred to as CAT5 or CAT3, depending on the quality of the wiring, to coax cable (also known as thinnet or 10Base5), and fiber optic cabling. Fiber optic cabling can transmit more data over longer distances than conventional cabling by using light or lasers instead of electrical signals, but are of relatively high cost. The network interface card or NIC connects a computer to the physical media or cabling of the network. A unique address, the Media Access Control or MAC address is derived from the NIC. This address is used to identify each individual computer on a network.

A hub or multiport repeater connects multiple computers on a network. A bridge isolates network traffic and connects two or more networks together. The bridge listens to all network traffic and keeps track of where individual computers are. A properly located bridge can take a large congested network segment and reduce the number of data collisions, improving performance.

A switch or router can scale a small-bandwidth network to a large bandwidth. Switches tend to be protocol-independent, whereas routers are protocol-dependent. Routers or relays are used in large networks because they can limit the broadcasts necessary to find devices and can more efficiently use the bandwidth. Routers, sometimes referred to as gateways, while beneficial in large complicated environments, can unfortunately slow traffic down because it has to examine data packets in order to make routing decisions. Bridges work much more quickly because they have fewer decisions to make. Switches have revolutionized the networking industry. They look at only as much of the data packet as bridges look at and are, in a sense, bridges with many interfaces. Switching incorporated with routing helps make network bottlenecks easier to remove. Some switches are used in the core of a network whereas others are used to replace hubs. Figure 14 shows example network switch and router devices.

Figure 15 diagrams an example PACS network, the one used at the University of California at San Francisco Medical Center for transmission of radiological images and related data around the hospital LAN and the health-care center WAN. The image acquisition devices, including the various modalities, such as CT and magnetic resonance scanners, are at the top of the diagram. As image acquisition devices are only capable of connecting to a network with Ethernet speeds, a switch or router is used to take scanner outputs over 10 Mbps in and transmit that data to the PACS servers using faster speeds of 100 Mbps. Images are sent to the display stations using the fastest network available. The circle in the upper-right corner of the diagram represents the UCSF Radiology WAN over which images and information are sent to other health-care facilities over 155 Mbps ATM.

Figure 14. Example network switches and routers.

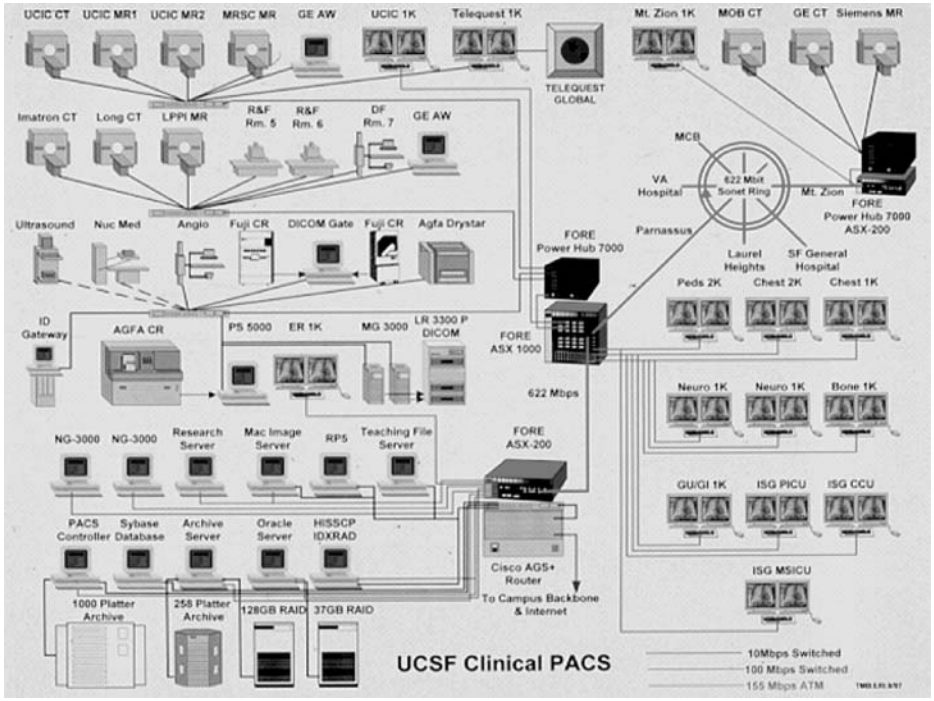


Figure 15. Example PACS network used at the University of California at San Francisco Medical Center for radiological images and related data.

Networking Software. The International Standards Organization (ISO) developed the Open Systems Interconnect (OSI) model as a framework to facilitate interoperation of computer networks from the application layer (i.e., the image viewer) all the way down to the physical layer (i.e., the wires). The ISO/OSI communication protocol stack is shown in Fig. 16. It consists of seven layers (46). Each layer in the stack is interested only in the exchange of information between the layer directly above or directly below, and each layer has different and well-defined tasks.

The top or seventh layer of the ISO/OSI stack is the Application Layer, which provides services to users. The Application Layer knows the data it wants to transmit and which machine it wants to communicate with. The sixth layer is the Presentation Layer, which takes care of data transformation such as encryption, compression, or reformatting. Layer five is the Session Layer, which controls applications running on different workstations, which is followed in the stack by the fourth or Transport Layer, which transfers data between end points and is handled here with error recovery. Layer three is the Network Layer, which establishes, maintains and terminates network connections. The second layer is the Data Link Layer, which handles network access, collision detection, token passing, and so on, and network control of logical links such as sending and receiving data messages or packets. The bottom layer or layer one is the Physical Layer corresponding to the hardware layer or the cable itself.

Also diagramed in Fig. 16 are the stacks for TCP/IP (Transmission Control Protocol/Internet Protocol) widely

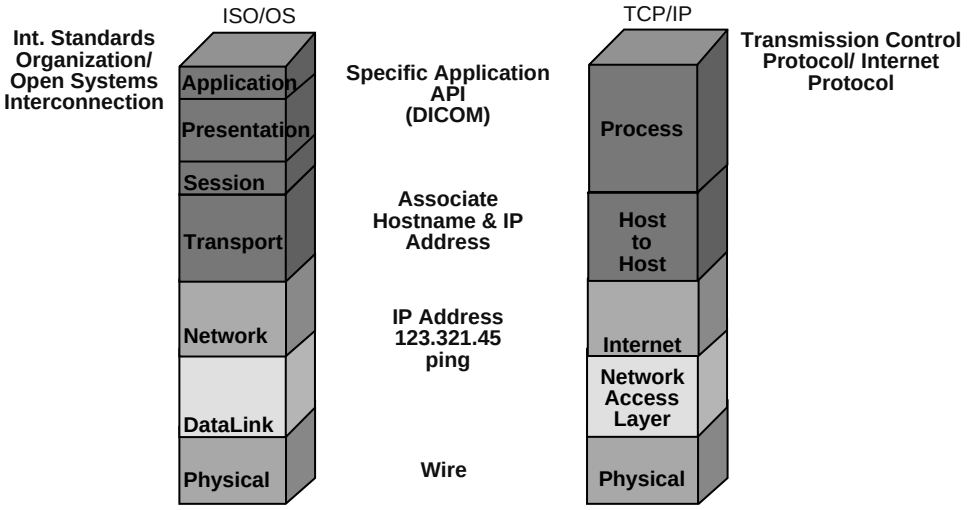


Figure 16. Communication protocol stacks for ISO/OSI and TCP/IP.

used in PACS applications. TCP/IP has four layers but is shown here as five layers, with the lowest level split out into two layers for the network access and physical layers. The top layer is called the Process Layer, and it corresponds to the Application and Presentation Layers in the ISO/OSI model. Such an application in the PACS world might be the DICOM communications protocol application. The layer below is the Host-to-Host or Transport Layer, followed by the Internet Layer and then the Network Access Layer, which encompasses Ethernet, token ring, for example, and the Physical or media Layer. To determine if two devices in a PACS network are communicating, the physical connection is tested. Using the unique IP address of the devices, a "ping" command will validate whether the one computer can reach the other over some network path. The TCP/IP hostname of the computer relates the unique IP address of the device to its DICOM AE (application entity) title so that computers can communicate using this protocol.

Figure 17 demonstrates the path messages take from one subnet, through a router, to another subnet. The process starts at the top of the stack in host A, where a DICOM port is opened. The information travels down the stack through the TCP Host-to-Host Layer to the Internet (IP) Layer and out the Network Access Layer across the Physical Layer. Messages then pass from the Physical Layer up the stack to the Network Access Layer, then the Internet Layer, to the Host-to-Host or Transport Layer, and, finally, a port is opened in the top Processor Application Layer.

Security and Redundancy. Advances in networking and communications devices have greatly facilitated the transfer of information between computers all over the world. For the secure transmission of private information such as clinical images and patient data, additional technologies are put into place to make the Internet a safe medium. The Health Insurance Portability and Accountability Act (HIPAA) of 1996 required the Department of Health and Human Services to establish national standards for electronic health-care transactions and national identifiers for providers, health plans, and employers. It also addresses the security and privacy of health data, such that adopting these standards will improve the efficiency and effectiveness of the nation's health-care system by encouraging the widespread use of electronic data interchange in health care.

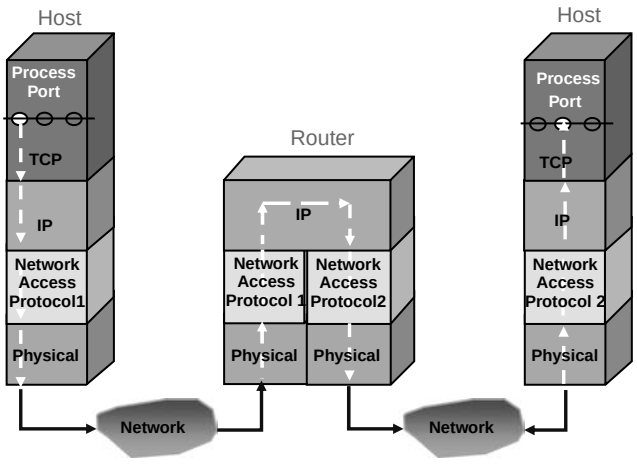


Figure 17. Connecting subnets via a router and DICOM.

Firewalls are protocol-dependent devices often used to secure a network by filtering traffic based on rules and policies. Intrusion detection systems (IDS) are another form of security device that uses policy-based monitoring, event logging, and alarms and alerting messaging to protect a network. Virtual Private Networks (VPNs) protect data by encrypting the information at the source device and decrypting it at the destination device. VPN clients are often used to securely access a hospital network for a location outside its LAN. A path can be created through the firewall or directly to a specific server enabling transmission of data.

PACS networks have become mission-critical devices in the transmission of digital medical images in and among health-care enterprises. As such, the networks must be highly available and have fault-tolerance mechanisms built in. Requirements for high availability networks include having redundant technology with automatic failover when devices are down. Redundant media and multiple paths should exist for the same information to get from one place to another. Redundant power for the devices involved is generally considered routine, as is proactive monitoring and problem mitigation with automated fault-detection processes in place.

Medical Image Display

Hardware – Monitors: CRT versus LCD. The choice of diagnostic display monitors was relatively straightforward for early adopters of PACS. Hardware was of a single type—cathode ray tube (CRT) technology—and was usually oriented in portrait mode emulating the shape of film. Monitors had high brightness, typically 200 to 300 candelas per square meter (cd/m^2), relative to other computer and television monitors, and high refresh rates of greater than 72 Hz to reduce flicker visible to the human eye. The devices themselves were physically large, heavy, and expensive. They generated noticeable quantities of heat while consuming relatively high amounts of power, and their display quality degraded quickly in time, requiring frequent monitor replacement.

Early medical-grade monitors were available in two spatial resolutions (high and low) reflecting their pixel matrix sizes (2 k or 2048 columns by 2500 rows and 1 k or 1024 columns by 1280 columns, respectively). Medium resolution 1.5 k monitors of 1500 columns by 1500 rows were later added to the mix. As a result of the exponentially higher cost of 2 k monitors as compared with 1 k monitors, radiology departments typically had a combination of a few strategically placed high resolution displays and many low or medium resolution displays. The American College of Radiology (ACR) recommended that 2 k monitors be used for primary diagnosis of digital projection radiographs because a single image could be displayed per monitor in its full inherent acquired spatial resolution. The cross-sectional modalities with slice matrix sizes of 512 by 512

for CT and 256 by 256 for MRI were considered adequately displayed on 1 k monitors. As display application software and graphical user interfaces (GUIs) improved, many radiologists became comfortable reading from 1 k monitors even for projection radiography, as long as the images were acquired at their full spatial and contrast resolutions and the GUIs allowed for easy manipulation, magnification, and comparison of images.

Today, a richer array of hardware technologies exist for the purposes of displaying digital medical images. Unfortunately, no formally defined standards or specification guidelines currently exist to clarify choices of monitors for users. The different display devices available today and the specifications to consider when purchasing monitors for use in radiological imaging are described. An explanation of the monitor types including CRTs, active-matrix liquid crystal displays (AM-LCDs), and plasma technologies is given along with a discussion of spatial resolution capabilities and requirements, contrast resolution and monitor luminance, the orientation or shape and number of displays necessary, and a comparison of color versus monochrome or gray-scale monitors. Device calibration and quality assurance practices are also addressed.

Two technology types of hardware displays are currently used in medical imaging, the half-century-old mature cathode ray tubes and what the popular literature refers to as flat-panel technology, of which several types exist (47). Of the two broad categories of flat-panel displays, one filters reflected light or light from a source behind the filter, whereas the second type creates light by exciting a phosphor. Note that the term flat panel is not meant to refer to the face of the monitor as some CRTs have a flat face (48). Rather, it refers to the thin-film transistor array panel that addresses each pixel.

CRTs produce light by exciting a phosphor-luminescent coating with a focused electron beam. Light is generated in an emissive structure, where it diffuses in a controlled manner forming the displayed image. The highest spatial resolution CRT monitors available have a display area of 2048 by 2560 or roughly 5 million (mega) pixels (Mpixels). They come in low- and high-bright versions of 50–60 FL and 100 FL or greater, respectively. High and low resolution (2 k and 1 k) monitors typically come in the portrait mode with a 9:16 or 3:4 aspect ratio emulating the shape of film. Most medium resolution CRTs (1.5 k) are square or in the landscape mode with an aspect ratio of 16:9 or 4:3. Choice of portrait versus landscape monitor shape is a function of personal user preference, with no technical issues bearing on the issue.

The flat-panel display type predominantly used in medical imaging is the active-matrix liquid crystal display (AM-LCD). LCDs use a transistor-driven matrix of organic liquid crystals that filter reflected light. LCDs use a light-modulating as opposed to a light-emitting mechanism for creating the display image. Polarization controls the light intensity such that the maximum intensity is perpendicular to the LCD panel. Consequently, this technology suffers from marked variations in luminance and contrast depending on viewing angle (47), which is the off-axis viewing or angle-of-regard problem in which images can appear quite different if viewed from different angles or heights above and below the center axes of the screen. Newer LCD designs have a more uniform luminance and contrast profile within a larger viewing angle cone (some as high as 170°). Users should inquire about the horizontal and vertical viewing angle capabilities and, better yet, ask the vendor for a demonstration monitor for clinician testing. LCDs typically have the capability to display in portrait and landscape modes.

Plasma display panels (PDPs) are currently being developed largely for high definition television (HDTV) viewing with 37 inch or larger screens. A current passed through ionized gas (Ne-Xe) contained between a pair of glass layers causes emission of ultraviolet light that excites visible light-emitting phosphors to produce the display image. PDPs are very expensive and have roughly the same number of addressable pixels as 17 inch LCDs, can be hung on a wall, have a wide viewing angle with no loss of quality, and have high brightness but relatively slow response time (48). They are not used currently in medical imaging because of their high cost, slow refresh rates, ghosting artifacts, and contrast limitations. Other types of displays, such as field-emissive display (FEDs) and organic light-emitting diodes (OLEDs), are undergoing heavy developmental efforts but are not yet viable for medical imaging display purposes (48).

Although the spatial resolution terminology used for both CRTs and LCDs is based on the device pixel matrix dimensions — 1 k to 2 k for CRTs and 3, 5, and 9 Mpixels for LCDs — not all monitors are created equal. For example, 1 k and 2 k CRT monitors tend to have standard diagonals so that the larger pixel matrix size connotes smaller pixel size and, hence, better spatial resolution capabilities, and all 1 k monitors have had equivalent spatial resolution, as did all 2 k monitors, which is not the case for LCD displays. For example, 3 Mpixel monitors come with different sized diagonals, that is, different physical sizes such that the physically bigger monitor actually has larger pixel size and hence poorer spatial resolution. Users need to understand what the pixel or spot size is, because it directly reflects spatial resolution and perception of fine detail, and not necessarily choose the largest screen. Pixel size can be determined from the physical screen size, typically given as a display area diagonal in inches and total pixel count or horizontal and vertical matrix resolution. Often, vendors give the device pixel density or pixel pitch spacing and, to confuse the issue, it is often given in millimeters. As for comparison between CRT and LCD monitors, the 1 k CRTs at 1024 by 1280 correspond to 1 Mpixel monitors, 1500 by 1500 correspond to 2 Mpixel monitors, and 1760 by 1760 correspond to 3 Mpixel displays. The 2 k or 2048 by 2500 CRTs correspond to the 5 Mpixel LCD. The recently introduced 9 Mpixel LCD display has 200 pixels per inch on a 22 inch diagonal screen.

The brightness of a monitor or its luminance affects perceived contrast or the number of discernable gray levels. Studies have shown that diagnostic accuracy increases as monitor luminance increases. To gain a perspective on luminance values, the typical lightbox or alternator used to display film is on the order of 400 to 600 FL (1360 to 2040 cd/m^2), whereas the standard PC color

monitor is roughly 20 to 40 FL (68 to 136 cd/m^2). An LCD color monitor has 65 to 75 FL (221 to 255 cd/m^2) monitor luminance, whereas the low-bright medical-grade CRT monitors have 50 to 60 FL (170 to 204 cd/m^2) and the high-bright CRTs have 100 FL (340 cd/m^2) or greater monitor luminance. Among the device specifications reflecting monitor brightness and affecting contrast resolution are the monitor and display card bit depth (typically 8 bits for 256 potential gray values) and the monitor dynamic range or contrast ratio reflecting the maximum discernable luminance over the minimum, with typical values of 600:1 or greater.

In comparing CRT versus LCD display technologies, the advantages of LCD over CRT monitors include better stability for longer device lifetime. The change in brightness of standard LCD monitors has been measured at less than 0.5% per month (47). LCDs are not prone to the geometric distortion typical of CRTs, they tend to consume less power, and this have reduced sensitivity and reflection artifacts from ambient room lighting. Disadvantages of LCD monitors versus CRTs include the afore-mentioned off-axis viewing or angle-of-regard distortion of LCDs, backlight instabilities, liquid crystal fluctuations with temperature, and manufacturing defects creating dead or nonresponsive pixel areas.

Receiver Operating Characteristic (ROC) studies are currently the best methodology available to compare monitor quality and associate it with reader performance, that is diagnostic accuracy, sensitivity, and specificity. Numerous clinical studies have been performed, most showing no significant difference between diagnostic performance on CRTs and LCDs. Recent studies representative of CRT versus LCD comparison for radiological diagnosis include one that examined brain CTs for identifying early infarction (49) and the other looked at CRs of the chest for the evaluation of interstitial lung disease (50). The CT ROC study showed no statistically significant differences in diagnostic performance between a 21 inch monochrome CRT monitor with a pixel matrix of 1280 by 1600 and a brightness of 175 FL versus an 18 inch color LCD monitor with a pixel matrix of 1024 by 1280 and a luminance of 55 FL, when 10 radiologists were asked to rate the presence or absence of disease on a 5 point scale. Similarly, an ROC study comparing the efficacy of a 5 Mpixel CRT display versus a 3 Mpixel LCD for the evaluation of interstitial lung disease in digital chest radiography showed no statistically significant change in observer performance sensitivity between the two types of monitors.

Several studies have investigated the comparison of color versus monochrome (technically achromatic) or gray-scale monitors, and a clear consensus does not seem to exist, which is an important issue because color monitors tend to have decreased luminance, contrast, and spatial resolution capabilities than monochrome monitors, and the human visual system has decreased spatial resolution perception in the color channels, but greater dynamic range (500 just-noticeable-differences (JND) versus 60 to 90 JNDs in gray-scale). On the other hand, high performance monochrome monitors are expensive and have a relatively short lifetime of approximately 3 years, and color is becoming increasingly useful in diagnostic imaging with the emergence of 3D display renderings. Although a study comparing monochromatic versus color CRT monitors found no statistically significant differences in display of CR chest images for the detection of subtle pulmonary disease, they did find higher sensitivity rates for specialty chest radiologists on the monochromatic monitor, perhaps due to the lower maximum luminance levels of the color displays (51). Another study comparing pulmonary nodule detection on P45 and P104 monochrome and color 1600 by 1200 pixel monitors found significantly greater false-positive and false-negative responses with the color monitors as well as longer search times (52). So, for primary diagnosis of projection radiographs in particular, monochrome monitors may still be the way to go. Note, however, that users prefer color LCDs when compared with color CRTs. This fact may be related to the Gaussian spot pixel and emissive structure of CRTs and the use of black matrix (shadow mask or aperture grille), which separates the red-green-blue phosphor dots that form an arrangement of color dots or stripes for luminance and chromatic contrast (47). Grille misalignment can degrade color purity and contrast.

Early PACS adopters equipped their radiology reading rooms with the highest quality display monitors, some 2 k, others 1 k, but all high brightness. The software applications were more complex than those targeted for the nonradiologist enterprise user. It was common to provide an intermediate application for use by image-intensive specialists such as orthopedists, neurosurgeons, and radiation oncologists as well as in image-intensive areas such as emergency departments and ICUs. Lesser quality monitors with stripped-down software capabilities were used by enterprise image users. It is interesting to note that as display hardware and software continue to evolve, display application software moves toward melding into one flexible easily configurable GUI, and one monitor type may, in time, meet most needs.

Monitor calibration and QA practices are important to maintaining high performing medical displays. The DICOM 14 Gray-scale Standard Display Function (GSDF) and the AAPM (American Association of Physicists in Medicine) Task Group 18 recommend that monitors be calibrated to a perceptually linearized display function as this is matched to the perceptual capabilities of the human visual system. Monitors forced to follow these standards produce more uniform images with optimum contrast. CRTs are less stable than LCD monitors requiring luminance calibration and matching to be conducted monthly, physically measuring light levels with a photometer. Many LCD displays have embedded luminance meters for automated QA measurements, Although some studies have also recommended doing external luminance measures but less frequently. LCDs must still be manually inspected for nonresponsive pixels. Routine manual viewing of test patterns, such as the SMPTE (Society of Motion Picture and Television Engineers) Test Pattern, are usually sufficient for evaluating overall monitor performance, low contrast, and fine detail detection.

Software Functionality. How many individual monitors does a user need per display workstation — 1, 2, 4, or 8?

Many feel that for primary diagnosis, dual-headed configurations are most efficient for comparison of current and prior relevant studies, particularly for projection radiographs. Note that a good GUI design can reduce the need for multiple monitors. The ability to page through and move images around the screen, the ability to instantaneously switch between tile and stack or cine modes of display, and the ability to view multiple studies on one monitor as well as side-by-side comparison of studies are critical to reducing the amount of hardware and physical display space required. In most cases, the two-monitor setup is sufficient for primary diagnosis and image intensive use with perhaps a third (color) monitor for worklist creation and access to other relevant medical data. The most common configuration for enterprise users is the single-headed or one-monitor display.

First and foremost, a software GUI must be intuitive and easy to use. The software application must be responsive, robust, and reliable. Most display workstations have GUIs to perform two basic functions. The first is to deliver a patient list or worklist of imaging examinations to be read, for example, "today's unread emergency department CTs." The worklist environment allows users to interrogate the entire PACS database for a subset of cases with which they wish to work. Typically, the database can be searched for by a specific patient name or other identifier, for individual imaging modalities, over specified time frames, by imaging or patient location within the hospital or health-care enterprise, and so on. The second basic function workstations perform is study display and image manipulation.

Automated hanging protocols based on examination type, the existence of prior historical examinations, and so on can greatly enhance radiology interpretation workflow. For example, if the current imaging study requiring interpretation is a two-view chest projection radiograph, then the default display might be to place the posterior-anterior (PA) view on the left display monitor and the lateral view on the right. If the patient has had a prior chest X ray, then the current PA should be placed on the left monitor with the lateral view behind and the prior PA on the right monitor with its corresponding lateral view waiting behind. If the current study is an MR of the brain, then automatically hang each sequence as a separate stack and place multiple (i.e., four-on-one) stacks per monitor so that they can be cined through simultaneously.

Basic display manipulation tools include the ability to dynamically change the window and level or contrast and brightness of the displayed image, the ability to magnify a portion of the image or zoom and pan through the entire image, and monitor configuration and image navigation tools such as paging, cine, and linked stack modes. Image mensuration capabilities, including linear, angle, and region-of-interest measurements, are also typical. Some advanced tools include the ability to track on a corresponding perpendicular slice, in MR studies, for example, where the cursor is on the perpendicular view for 3D referencing. Figure 18 depicts several example softcopy image displays on PACS workstations.

Well-designed PACS are tightly integrated with other information systems such as the hospital or radiology information system, enabling access to other relevant data about the patient or imaging examination from within the display application. Sometimes diagnostic radiology report generation systems, such as speech recognition, are embedded in the diagnostic workstation.

Figure 18. Example soft copy image display on PACS workstations.

It is anticipated that the number of specialty display applications available on display stations will continue to increase as more features are added. Examples of the results of several advanced processing techniques are shown in Fig. 19. Some systems also provide algorithms for image processing such as image sharpening, or edge enhancement, and image smoothing. Maximum intensity projection (MIP) displays and multiplanar reformats (MPR) are also appearing on PACS workstations. Real-time 3D reconstruction of image data at the PACS display is beginning to be seen. Multimodality image data fusion such as CT and positron emission tomography (PET) images to depict the functional maps overlaid on the anatomical data maps will also likely be viewable.

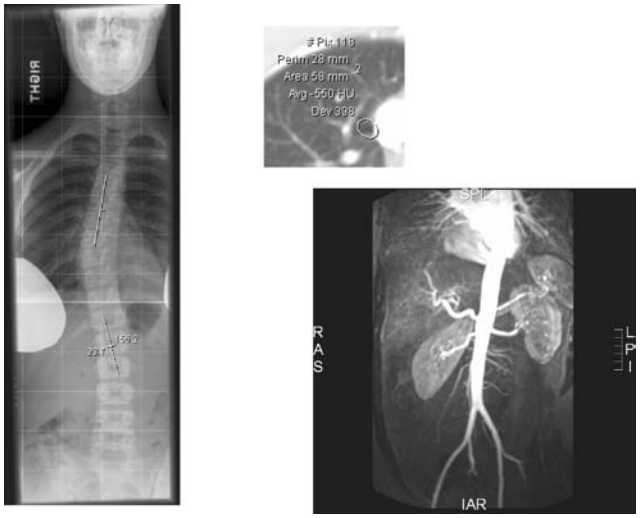


Figure 19. Example advanced processing techniques: calculation of scoliosis Cobb angle, region-of-interest measurements, 3D magnetic resonance angiogram.

CURRENT TRENDS AND FUTURE PROSPECTS

Medical imaging is increasingly the key triage event in a patient's encounter with the health-care system. The ability to eliminate the use of film in radiological imaging is a reality today. In addition to the economic advantages of using PACS for digital medical imaging, rapid access to all clinical information on patients, including imaging studies, anytime and anywhere with security, enhances the quality of patient care.

Hurdles still exist today. PACS are not just a radiological tool. Images are required enterprise-wide by many different types of clinicians and other health-care providers. Images are required for viewing in emergency departments, ICUs, surgical ORs, outpatient clinics and referring physicians' offices as well as for teaching conferences, and at home for viewing by patients and clinical providers. Unless PACS can also deliver images to everyone in the health-care enterprise who requires them, film cannot be turned off, and facilities will have to operate in a painful mixed environment.

Web-based enterprise PACS applications do exist and continue to improve in their performance. Note that web-based PACS for within the radiology department are also becoming more common. Several requirements in the enterprise environment make transition to the all-digital, totally filmless medical imaging facility more difficult than just within the radiology department. The web-PACS enterprise application user interface or GUI must be very intuitive—like the rental car model. Most licensed drivers are able to get into a rental car, orient themselves to where the lights and windshield wipers are, and then go ahead and drive the car. They do not need to know how a car works in order to drive it, and the user interface is very standard across most all cars. The car works robustly and reliably and the user does not need to read a manual before they can drive the car. Unfortunately, the state-of-the-art in computer GUIs is not quite as intuitive, but much progress has been and continues to be made in this area, making GUIs self-training and bullet-proof. Human-computer interfacing and interaction along with GUI design are currently active areas of research and discovery.

Web-PACS applications are often required to operate in mixed-platform environments to accommodate PC, Macintosh, and Unix boxes, which is sometimes problematic. Applications must be improved to be able to handle bottlenecks in the system at both the input to the database and the output to multiple users accessing the system simultaneously. The best applications are purely web-based; that is, they are not dependent on having to download Active-X components to every physical machine that uses the system.

In summarizing the key components and essential features for clinical implementation of a PACS, at acquisition, all image acquisition devices or imaging modalities should be required to conform to the DICOM standard image format and communications protocol. Devices and PACS in general operate best when well interfaced to other clinical information systems such as the HIS-RIS, report generation systems such as speech recognition applications, computerized physician order-entry (CPOE), and decision support systems. The inherently digital imaging modalities should be acquired into a PACS using direct DICOM capture. Film digitizers such as laser scanners can be used to acquire imaging examinations existing only on film into a PACS if required. Acquisition of digital projection radiographs, such as the conventional chest X ray, can be achieved using CR or photostimulable phosphor devices or digital radiography devices, which directly convert images to digital at the time of X ray exposure. CR and DR devices are likely to coexist for some time.

Archival media and devices will continue to advance, with databases becoming more patient-centric and seamlessly searchable. Computer networking will also improve in not only the hardware devices, but also the network management software strategies. Display GUIs must continue to become more intuitive and robust, and the monitors themselves will trend toward LCD devices as opposed to CRTs.

Predictors for success of the introduction of new technologies into the clinical arena include ease of integration into the existing workflow or change management activities to optimize new workflow with new devices. Systems must be reliable, simple to use, and have optimum performance so that processes can be completed more efficiently than in the analog film-based world. Systems must be flexible and configurable on a per-user basis and they must include fault-tolerance, redundancy, and error-tracking capabilities.

In the future, radiology can help to drive the changes that will greatly impact all of health care. Medical image management systems are maturing outside of the radiology department, providing hospital enterprise-wide access to clinical images and information via the Intranet as well as web access from outside the enterprise via a secure Internet connection. The web will change the way people communicate and perform their duties. Web-based PACS applications will become the norm offering ubiquitous distribution of and access to clinical data. Computer workstations will become less costly, more reliable, and have more intuitive GUIs. All relevant medical information systems will become more tightly integrated with each other, sharing and maintaining user, patient, image, and application sensitivity such that multiple distinct applications perform virtually as one.

Future PACS are likely to be on the less expensive off-the-shelf PC platform using industry standards. PACS display stations are likely to be configured with fewer numbers of monitors—two within radiology and image-intensive specialty areas and one out in the enterprise. Small and medium sized community hospitals, private practices, outpatient centers in rural areas, and some indigent care facilities will begin realizing the benefits of PACS and digital medical imaging through better access to high quality diagnostic imaging services.

PACS functionality will be incorporated into handheld devices for some applications, and wireless transmission will mature in the clinical arena. Improved integration with other information technologies into the total electronic medical record (EMR), including automated speech recognition systems, will enable a more efficient filmless environment as well as a paperless workflow.

Advanced image processing utility and translation from the research environment to clinical applications will increase. 3D displays, the use of color and video will increase as will the incorporation of computer-aided detection and decision support through outcomes research and evidence-based medicine will become more prevalent. Multimodality functional and molecular imaging will mature clinically, and value-added applications for specialties outside of diagnostic imaging will increase. Virtual reality imaging presentations and image-guided surgery applications are likely to become more commonly used clinically.

It is likely that the radiological interpretation process will need to transform in order to handle the information and image data overload currently plaguing medical imaging. This image interpretation paradigm shift will be required in order to evaluate, manage, and exploit the massive amounts of data acquired in a more timely, efficient, and accurate manner. Discovery and development of this new paradigm will require research into technological, environmental, and human factors. Interdisciplinary research into several broad areas will be necessary to make progress and ultimately to improve the quality and safety of patient care with respect to medical imaging. These areas are likely to include studies in human perception, image processing and computer-aided detection, visualization, navigation and usability of devices, databases and integration, and evaluation and validation of methods and performance. The result of this transformation will affect several key processes in radiology, including image interpretation, communication of imaging results, workflow and efficiency within health-care enterprises, diagnostic accuracy and a reduction in medical errors, and, ultimately, the overall quality of patient care (53).

Twentieth century medical imaging was film-based in which images were interpreted on analog viewboxes, film was stored as the legal imaging record. Film had to be manually disturbed from one location to another and could be accessed in only one physical location at a time. Twenty-first century medical imaging will be characterized by digital image acquisition, softcopy computer interpretation, digital image archives, and electronic distribution. It is anticipated that the use of PACS and other information technology tools will enable the filmless, paperless, errorless era of imaging in medicine.

BIBLIOGRAPHY

1. Lemke HU, Stiehl HS, Scharnweber H, Jackel D. Applications of picture processing, image analysis and computer graphics techniques to cranial CT scans. Proceedings of the Sixth Conference on Computer Applications in Radiology and Computer Aided Analysis of Radiological Images; Los Alamitos, CA: IEEE Computer Society Press; 1979. 341–354.
2. Capp ML, et al. Photoelectronic radiology department. Proc SPIE–Int Soc Opt Eng 1981;314:2–8.
3. Dwyer III, SJ, et al. Cost of managing digital diagnostic images. Radiology 1982;144:313.
4. Duerinckx A., editor. Picture archiving and communication systems (PACS) for medical applications. First International Conference and Workshop, Proc SPIE; 1982; 318, Parts 1 and 2.
5. Huang HK, et al. Digital radiology at the University of California, Los Angeles: A feasibility study. Proc SPIE 1983; 418:259–265.
6. Blaine GJ, Hill RL, Cox JR, Jost RG. PACS workbench at mallinckrodt Institute of Radiology (MIR). Proc SPIE 1983; 418:80–86.
7. Seshadri SB, et al. Prototype medical image management system (MIMS) at the University of Pennsylvania: Software design considerations. Proc SPIE 1987;767:793–800.
8. Kim Y, Fahy JB, DeSoto LA, Haynor DR, Loop JW. Development of a PC-based radiological imaging workstation. Proc SPIE 1988;914:1257–1264.
9. Horii SC, et al. Environmental designs for reading from imaging workstations: Ergonomic and architectural features. Proc SPIE 1989;1091:172–178.
10. Arenson RL, Chakraborty DP, Seshadri SB, Kundel HL. The digital imaging workstation. Radiology 1990;176:303–315.
11. Hruby W, Maltsidis A. A view to the past of the future – A decade of digital revolution at the Danube Hospital. In: Hruby W, editor. Digital (R)evolution in Radiology. Vienna: Springer Publishers; 2000.
12. DICOM. 2004. Online. Available at http://medical.nema.org/ .http://medical.nema.org/dicom/2004/.
13. Andriole KP. Anatomy of picture archiving and communication systems: nuts and bolts – image acquisition: getting digital images for imaging modalities. Dig Imag 1999;12(2) Suppl 1: 216–217.
14. Andriole KP, Avrin DE, Yin L, Gould RG, Arenson RL. PACS databases and enrichment of the folder manager concept. Dig Imag 2000;13(1):3–12.
15. Andriole KP. Computed radiography overview. In: Seibert JA, Filipow LJ, Andriole KP, editors. Practical Digital Imaging and PACS. Medisson, WI: Medical Physics Publishing; 1999. p 135–155.
16. Bogucki TM, Trauernicht DP, Kocher TE. Characteristics of a storage phosphor system for medical imaging. Kodak Health Sciences Technical and Scientific Monograph. No. 6, New York: Eastman Kodak Co.; July 1995.
17. Barnes GT. Digital X-ray image capture with image intensifier and storage phosphor plates: Imaging principles, performance and limitations. Proceedings of the AAPM 1993 Summer School: Digital Imaging; Charlottesville, VA: University of Virginia; Monograph 22: 23–48.
18. Wilson AJ, West OC. Single-exposure conventional and computed radiography: The hybrid cassette revisited. Invest Radiol 1993;28(5):409–412.
19. Andriole KP, Gooding CA, Gould RG, Huang HK. Analysis of a high-resolution computed radiography imaging plate versus conventional screen-film radiography for neonatal intensive care unit applications. SPIE Phys Med Imag 1994;2163: 80–97.
20. Kodak. Digital radiography using storage phosphors. Kodak Health Sciences Technical and Scientific Monograph. New York: Eastman Kodak Co.; April 1992.
21. Matsuda T, Arakawa S, Kohda K, Torii S, Nakajima N. Fuji Computed Radiography Technical Review. No. 2. Tokyo: Fuji Photo Film Co., Ltd.; 1993.
22. Berg GE, Kaiser HF. The X-ray storage properties of the infrared storage phosphor and application to radiography. Appl Phys 1947;18:343–347.
23. Luckey G. Apparatus and methods for producing images corresponding to patterns of high energy radiation. U.S. Patent 3,859,527. June 7, 1975. Revision No. 31847. March 12, 1985.
24. Kotera N, Eguchi S, Miyahara J, Matsumoto S, Kato H. Method and apparatus for recording and reproducing a radiation image. U.S. Patent 4,236,078. 1980.

25. Sonoda M, Takano M, Miyahara J, Kato H. Computed radiography utilizing scanning laser stimulated luminescence. Radiology 1983;148:833–838.
26. Agfa. The highest productivity in computed radiography. Agfa-Gevaert N.V. Report. Belgium: AGFA; 1994.
27. Ogawa E, Arakawa S, Ishida M, Kato H. Quantitative analysis of imaging performance for computed radiography systems. SPIE Phys Med Imag 1995;2432:421–431.
28. Kodak. Optimizing CR images with image processing: Segmentation, tone scaling, edge enhancement. Kodak Health Sciences Technical and Scientific Manuscript. New York: Eastman Kodak; March 1994.
29. Gringold EL, Tucker DM, Barnes GT. Computed radiography: User-programmable features and capabilities. Dig Imag 1994;7(3):113–122.
30. Ishida M. Fuji computed radiography technical review, No. 1. Tokyo: Fuji Photo Film Co., Ltd.; 1993.
31. Storto ML, Andriole KP, Kee ST, Webb WR, Gamsu G. Portable chest imaging: clinical evaluation of a new processing algorithm in digital radiography. 81st Scientific Assembly and Annual Meeting of the Radiological Society of North America. Chicago, IL: November 26 – December 1, 1995.
32. Vuylsteke P, Dewaele P, Schoeters E. Optimizing Radiography Imaging Performance. Proceedings of the 1997 AAPM Summer School; 1997; 107–151.
33. Solomon SL, Jost RG, Glazer HS, Sagel SS, Anderson DJ, Molina PL. Artifacts in Computed Radiography. AJR 1991;157: 181–185.
34. Volpe JP, Storto ML, Andriole KP, Gamsu G. Artifacts in chest radiography with a third-generation computed radiography system. AJR 1996;166:653–657.
35. Oestman JW, Prokop M, Schaefer CM, Galanski M. Hardware and software artifacts in storage phosphor radiography. RadioGraphics 1991;11:795–805.
36. Lee DL, Cheung LK, Jeromin LS. A new digital detector for projection radiography. Proc SPIE Phys Med Imag 1995;2432: 237–249.
37. Andriole KP. Productivity and cost assessment of CR, DR and screen-film for outpatient chest examinations. Dig Imag 2003;15(3):161–169.
38. Pratt HM, et al. Incremental cost of department-wide implementation of a PACS and computed radiography. Radiology 1998;206:245–252.
39. Chunn T. Tape storage for imaging. Imag World 1996;5(8): 1–3.
40. Horii S, et al. A Comparison of case-retrieval times: Film versus PACS. Dig Imag 1992;5(3):138–143. htpp://www.diagnosticimaging.com.
41. Eisenman, et al.. Diagnost Imag 1996;9:27.
42. Siegel E, Shannon R. Understanding Compression. Great Fall, VA: SCAR Publications; 1997; 11–15.
43. Erickson BJ, Manduca A, Palisson P, Persons KR, Earnest F 4th, Savcenko V, Hangiandreou NJ. Wavelet compression of medicl images. Radiology. 1998;206(3):599–607.
44. Avrin DE, et al. Hierarchical storage management scheme for cost-effective on-line archival using lossy compression. Dig Imag 2001;14(1):18–23.
45. Savcenko V, Erickson BJ, Palisson PM, Persons KR, Manduca A, Hartman TE, Harms GF, Brown LR. Detection of subtle abnormalities on chest radiographs after irreversible compression. Radiology 1998;206(3):609–616.
46. Huang HK. PACS and Imaging Informatics: Basic Principles and Applications. Hoboken, NJ: John Wiley & Sons, Inc.; 2004.
47. Badano A. Principles of cathode-ray tube and liquid crystal display devices. In: Samei E. Flynn MJ, editors. Advances in Digital Radiography: Categorical Course in Diagnostic Radiology Physics. Oak Brook, IL: Syllabus, RSNA; 2003. pp. 91–102.
48. Leachtenauer JC. Electronic Image Display: Equipment Selection and Operation. Bellingham, WA: SPIE Press; 2004.
49. Partan G, et al. Diagnostic performance of liquid crystal and cathode-ray-tube monitors in brain computed tomography. Eur Radiol 2003;13:2397–2401.
50. Langer S, et al. Comparing the efficacy of 5-MP CRT versus 3-MP LCD in the evaluation of interstitial lung disease. Dig Imag, Online publication date, June 29, 2004.
51. Iwano S, et al. Detection of subtle pulmonary disease on CR chest images: Monochromatic CRT monitor vs color CRT monitor. Eur Radiol 2001;11:59–64.
52. Krupinski E, Roehrig H. Pulmonary nodule detection and visual search: P45 and P104 monochrome versus color monitor displays. Academ Radiol 2002;9:638–645.
53. Andriole KP, et al. Addressing the coming radiology crisis: Transforming the radiological interpretation process. Dig Imag Online October 2004.

Further Reading

Andriole KP, Gould RG, Avrin DE, Bazzill TM, Yin L, Arenson RL. Continuing quality improvement procedures for a clinical PACS. *Dig Imag* 1998;11(31):111–114.

Honeyman JC, et al. PACS quality control and automatic problem notifier. *SPIE Med Imag 1997: PACS Design and Evaluation* 1997;3035:396–404.

Honeyman JC, Staab EV. Operational concerns with PACS implementations. *Appl Radiol* 1997; August: 13–16.

Seibert JA. Photostimulable phosphor system acceptance testing. In: Seibert JA, Barnes GT, Gould RG, editors. Specification, Acceptance Testing and Quality Control of Diagnostic X-ray Imaging Equipment. Medical Physics Monograph no. 20. Woodbury, NY: AAPM; 1994. 771–800.

Willis CE, Leckie RG, Carter J, Williamson MP, Scotti SD, Norton G. Objective measures of quality assurance in a computed radiography-based radiology department. *Proc SPIE* 1995; 2432:588–599.

See also PHOTOGRAPHY, MEDICAL; RADIOLOGY INFORMATION SYSTEMS; ULTRASONIC IMAGING.

PIEZOELECTRIC SENSORS

YANBIN LI
Department of Biological and Agricultural Engineering, University of Arkansas
Fayetteville, AR

XIAO-LI SU
BioDetection Instruments LLC
Fayetteville, AR

INTRODUCTION

Piezoelectric sensors are generally referred to as analytical devices for detection of chemical or biological agents with piezoelectric quartz crystals (PQCs) as transducers. The origin of piezoelectric sensors can be traced back to 1880 when Jacques and Pierre Curie discovered normal and converse piezoelectric effects (1). In the former, the application of a mechanic stress to the surface of quartz and some other crystals induces an electric potential across the crystal; in the latter, conversely, the application of a voltage across the crystal results in an internal mechanical

strain. The converse piezoelectric effect is the basis of all piezoelectric sensors.

PQCs have been widely used in oscillators and filter circuits for high-precision frequency control. The use of PQCs as a mass sensor is based on the work of Sauerbrey (2), which established a linear relationship between the decrease in resonant frequency and the increase in surface mass loading of a PQC. Because of its high sensitivity for mass detection, a piezoelectric sensor is usually called quartz crystal microbalance (QCM). QCM was originally and is still used to measure thickness of coatings in vacuum and air. The first example for application of QCM to analytical chemistry was reported by King in 1964 (3). He coated PQCs with various materials and used them as a sorption detector in gas chromatography to detect and measure the composition of vapors and gases. The applications of piezoelectric sensors were limited to the determination of environmental and other gas species for a long time. The liquid-phase measurements began in the 1980s when new oscillator technology emerged and advanced to make PQCs oscillate in solution as stably as in gas. Then, numerous piezoelectric chemical sensors and biosensor have been reported.

Piezoelectric sensors, characterized by their simplicity, low cost, high mass-detection sensitivity, and versatility, have found increasing applications in biomedical analyses. The objective of this article is to briefly present the theory, equipment, and applications of piezoelectric sensors in the biomedical area. Interested readers are referred to some key review articles (4–9) for more theoretical and technical details of piezoelectric sensors.

THEORY

Typically, a piezoelectric sensor is fabricated by modifying a PQC with a layer of sensing material, chemical or biological, that has specific affinity to a target analyte. The specific binding between the sensing material and the target analyte causes a change in the resonant frequency of PQC that is proportional to the amount of target analyte adsorbed or bound on the sensor surface, which can be correlated to the concentration of target analyte in the original sample.

A typical PQC consists of a quartz crystal wafer and two excitation electrodes plated on opposite sides of the crystal. The wafer is cut from a natural or synthetic crystal of quartz. The electromechanical coupling and stresses resulting from an applied electric field depend on the crystal symmetry, cut angle, and electrode configuration (4). Different modes of electromechanical coupling lead to different types of acoustic waves, including thickness shear mode (TSM), surface acoustic wave (SAW), shear horizontal (SH) SAW, SH acoustic plate mode (APM), and flexural plate wave (FPW).

The TSM device, widely referred to as QCM, is the simplest and most popular piezoelectric sensor. Hence, we will focus on TSM piezoelectric sensors in this article. A TSM sensor typically uses AT-cut quartz as a piezoelectric substrate, which has a minimal temperature coefficient and is obtained by cutting quartz crystals at approximately 35° from the z-axis of the crystal. Figure 1 shows the schematic of an AT-cut PQC. When an alternating electric field is applied across an AT-cut quartz crystal through the excitation electrodes, the crystal produces a shear vibration in the x-axis direction parallel to the electric field and propagation of a transverse shear wave through the crystal in the thickness direction. The resonant frequency of the vibration is determined by the properties of the crystal and the contacting medium.

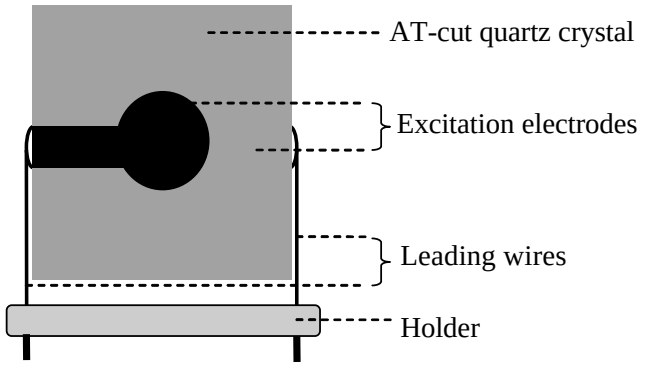


Figure 1. Schematic of an AT-cut piezoelectric quartz crystal.

Mass Response

Most piezoelectric sensors are used as mass sensors that are based on the Sauerbrey equation (2):

$$\Delta F = \frac{-2F_0^2}{A\sqrt{\mu_q \rho_q}}\Delta M \tag{1}$$

where ΔF is the frequency change measured (Hz), F_0 is the resonant frequency of the fundamental mode of the crystal (Hz), A is the area of the gold disk coated onto the crystal (cm^2), ρ_q is the density of the quartz crystal (2.648 $g{\cdot}cm^{-3}$), μ_q is the shear modulus of quartz (2.947×10^{11} $g{\cdot}cm^{-1}{\cdot}s^{-2}$), and ΔM is the mass change (g). The Sauerbrey equation is applicable only to a thin ($\sim$1 μm) and elastic film coupled to the crystal surface, where the mass loading can be up to 0.05% of the crystal mass.

According to the Sauerbrey equation, the magnitude of frequency decrease corresponding to a mass increase is proportional to $F_0{}^2$; i.e. the higher the fundamental resonant frequency, the higher the mass sensitivity. Typical AT-cut PQCs have $F_0 = 5 \sim 30$ MHz with a frequency resolution of $\sim$0.1 Hz and a mass sensitivity of $0.056 \sim 2.04$ $Hz{\cdot}cm^2{\cdot}ng^{-1}$. Thinner crystal wafers have higher F_0 and thus have higher mass sensitivity, but they are also more fragile and thus are more difficult to manufacture and handle. For one of the most commonly used AT-cut PQCs, $F_0 = 9$ MHz, $A = 0.2$ cm^2, and the detectable mass is 1.09 ng per Hz, which is approximately 100 times higher than that of an electronic fine-balance with a sensitivity of 0.1 μg. As AT-cut piezoelectric sensors can sensitively detect mass change at nanogram levels, they are frequently referred to as quartz crystal microbalances (QCMs) or nanobalances.

However, as a mass sensor, the QCM does not have selectivity. To make a selective QCM chemical sensor or biosensor, the QCM must be coated with a film of chemical or biological recognition material that is selective to a target analyte.

Viscosity-Density Effect

When a QCM is employed in liquid phase, in addition to the mass change, it also responds to other factors such as liquid density and viscosity, surface energy, viscoelasticity, roughness, and surface charge density. For a QCM with only one side in contact with a Newtonian liquid, ΔF is linearly proportional to the squared root of the product of viscosity (η_L) and density (ρ_L) of the liquid (10):

$$\Delta F = -F_0^{3/2}\sqrt{\rho_L \eta_L / \pi \mu_q \rho_q} \tag{2}$$

For a QCM with simultaneous mass and liquid loading, ΔF can be expressed as (11)

$$\Delta F = -\frac{2F_0^2}{\sqrt{\mu_q \rho_q}}(\Delta M/A + \sqrt{\rho_L \eta_L / 4\pi F_0}) \tag{3}$$

In equation 3, the first term is equivalent to the Sauerbrey equation, and the second term is equivalent to Kanazawa equation. Equation 3 indicates the additive nature of mass and viscosity-density effects in changing the resonant frequency. It also shows that it is impossible to distinguish the mass effect from the viscosity-density effect when only the resonant frequency is monitored.

Some piezoelectric sensors are based on the density-viscosity change rather than on the elastic pure mass change. For example, a piezoelectric sensor was used to detect *E. coli* based on the gelation of Tachypleus amebocyte lysate (TAL) (12), and the detection range is $2.7 \times 10^4 - 2.7 \times 10^8$ cells·mL^{-1}. Gee et al. (13) used a piezoelectric senor for measuring microbial polymer production and growth of an environmental isolate obtained from river sediment contaminated with petroleum hydrocarbons. The increasing amount of produced polymer corresponded to an increase in the viscosity of the liquid, which was directly measurable as the fluid contacts the surface of the quartz crystal in the sensor system. These methods, although they lack specificity, are advantageous in that coating on the PQC surface is unnecessary.

Equivalent Circuit Analysis

Equivalent circuit analysis can provide more detailed information about the surface/interface changes of a piezoelectric sensor (11,14–17). A PQC can be represented by a Butterworth–Van Dyke (BVD) model (Fig. 2), which is composed of a static capacitance (C_0) in parallel with a motional branch containing a motional inductance (L_m), a motional capacitance (C_m), and a motional resistance (R_m) in series. Each parameter has its distinct physical meaning: C_0 reflects the dielectric properties between the electrodes located on opposite sides of the insulating quartz crystal; C_m represents the energy stored during oscillation, which corresponds to the mechanical elasticity of the vibrating body; L_m is related to the displaced mass; and R_m is the energy dissipation during oscillation, which is closely related to viscoelasticity of the deposited films and viscosity-density of the adjacent liquid.

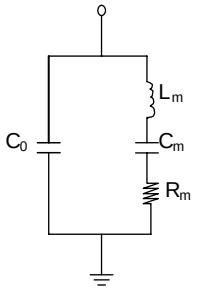


Figure 2. Butterworth–Van Dyke model of a piezoelectric quartz crystal.

The BVD model can be described by the following admittance equations:

$$Y(\omega) = G(\omega) + jB(\omega) \tag{4}$$

$$G(\omega) = \frac{R_\mathrm{m}}{R_\mathrm{m}^2 + (\omega L_\mathrm{m} - 1/\omega C_\mathrm{m})^2} \tag{5}$$

$$B(\omega) = -\frac{(\omega L_\mathrm{m} - 1/\omega C_\mathrm{m})}{R_\mathrm{m}^2 + (\omega L_\mathrm{m} - 1/\omega C_\mathrm{m})^2} + \omega C_0 \tag{6}$$

where Y is admittance, i.e., the reciprocal of impedance. Y is a complex quantity, its real part G is conductance, and its imaginary part B is susceptance. Both G and B are a function of the scanning frequency f ($\omega = 2\pi f$) and the four equivalent circuit parameters. These parameters are determined by physical properties of the quartz crystal, perturbing mass layer and contacting liquid, and can be obtained by fitting the measured impedance/admittance data to the BVD model using the admittance equations. Figure 3 shows typical admittance spectra of an unperturbed 8 MHz AT-cut PQC in air. The fitted results of F_0, R_m, L_m, C_m, and C_0 were 7.99 MHz, 9.6 Ω, 17.9 mH, 22.2 fF, and 8.2 pF (including parasitic capacitance in the test fixture) for the quartz crystal (18).

High-frequency admittance/impedance analysis has been extensively used in surface/interface studies. A simpler way to provide insights into the viscoelastic properties of a bound surface mass is to simultaneously monitor F_0 and R_m or F_0 and the dissipation factor D using a quartz crystal analyzer that is much less expensive than the impedance analyzer. This method has been applied to study the behavior of adherent cells in response to chemical, biological, or physical changes in the environment.

For a QCM in contact with liquid, the change of motional resistance was first derived by Muramatsu et al. (14) as follows:

$$\Delta R = (2\pi F_0 \rho_L \eta_L)^{1/2} A / k^2 \tag{7}$$

where k is the electromechanical coupling factor.

Simultaneous measurements of ΔF and ΔR can differentiate an elastic mass effect from the viscosity-induced effect. ΔR is a good measure of the viscoelastic change. For an elastic mass change, ΔR will be zero and ΔF will be linearly proportional to the mass change in accordance with the Sauerbrey equation. For a QCM with only one side in contact with a Newtonian liquid, both ΔF and ΔR are linearly proportional to the squared root of the product

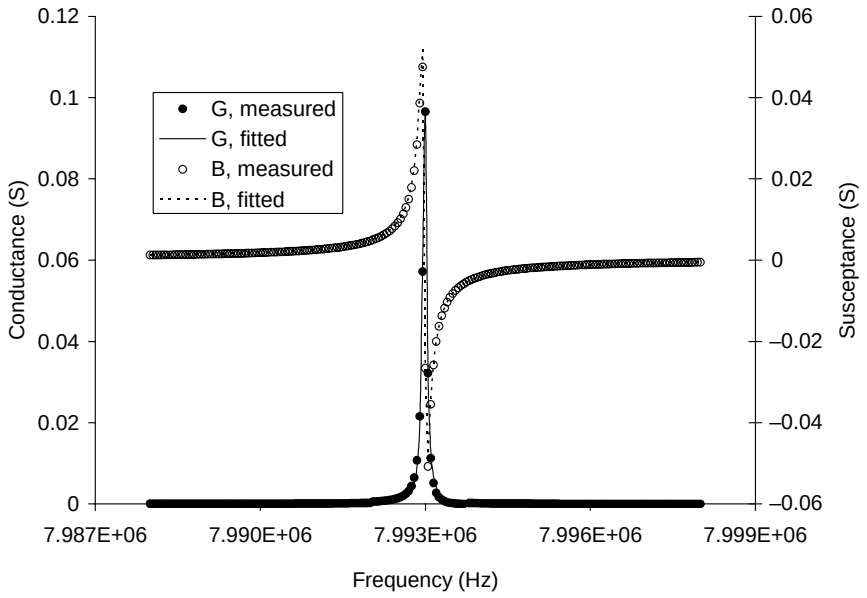


Figure 3. Typical conductance and susceptance spectra of an unperturbed 8 MHz AT-cut PQC in air.

of viscosity and density of the liquid. Therefore, a pure viscosity-density change will result in a linear $\Delta F \sim \Delta R$ plot. In the presence of a viscoelastic change, the $\Delta R \sim \Delta F$ plot will lie between the pure viscosity-density effect line and the elastic mass effect line or even above the former (8,18).

EQUIPMENT AND EXPERIMENTS

Traditionally, a piezoelectric sensor (resonator) is driven by a homemade oscillator, and the oscillation frequency is measured by a frequency counter that is connected to a recorder or computer for data collection. The Pierce oscillator with a fundamental mode AT-cut resonator is the most popular oscillator design type and operates with the piezoelectric sensor as an inductive element. Now such oscillators are commercially available. One of the suppliers is International Crystal Manufacturing (Oklahoma City, OK), which produces a standard (clock) oscillator for gas-phase QCMs and lever oscillator for liquid-phase QCMs.

Highly sophisticated, automatic, and microprocessor-controlled piezoelectric detectors or QCM instruments are commercially available from several manufacturers (19). The main commercial systems include QCA-917- and 922 quartz crystal analyzers (Princeton Applied Research, Oak Ridge, TN), EQCM 400 series electrochemical quartz crystal microbalances (CH Instruments, Austin, TX), EQCN-700 and -900 electrochemical quartz crystal nanobalances (Elchema, Potsdam, NY), the PZ-1000 immuno-biosensor system and PZ-105 gas phase piezoelectric detector (Universal Sensors, Metairie, LA), RQCM research QCM (Maxtek, Santa Fe Springs, CA), and Mark series cryogenic and thermally controlled QCMs (QCM Research, Lake Forest, CA). These systems are designed to reliably measure mass change up to ~100 μg with a resolution of ~1 ng·cm^{-2}. Most of them are programmed and controlled by easy-to-use Windows-based software (Microsoft Corporation, Redmond, WA). The QCA 922, designed for EQCM with a potentiostat or stand-alone operation, can simultaneously measure resonant frequency and resistance of QCM. The RQCM can measure crystal frequency and crystal resistance for up to three crystals simultaneously. Moreover, high-frequency impedance/admittance analyzers such as E4991A (Agilent Technologies, Palo Alto, CA) can be used to obtain the impedance/admittance spectra of the quartz crystal and to acquire the equivalent circuit parameters by fitting the impedance/admittance data to the BVD model as described earlier.

Previously, for liquid-phase applications, the dip-and-dry approach is made in the measurement of piezoelectric sensors, in which the resonant frequency of the same sensor is measured in gas phase before and after the sample solution is dipped and dried. This approach does not need a special fixture to mount the crystal; however, it is unsuitable for automation and has poor reproducibility. By mounting the crystal to a dip or well cell, like those from Princeton Applied Research and International Quartz Manufacturing, and letting only one side of the crystal exposed to the test solution, it is possible to monitor the frequency change in solution in real time.

Flow-through QCM biosensors have drawn increasing attention due to its ease for automation. For example, Su and Li (20) developed a flow-through QCM immunosensor system for automatic detection of *Salmonella typhimurium*. The QCM immunosensor was fabricated by immobilizing anti-*Salmonella* antibodies on the surface of an 8 MHz AT-cut quartz crystal with Protein A method, and then installed into a 70 μL flow-through detection cell. The flow cell was composed of acrylic, with upper and lower pieces held together by two screws with O-rings. One face of the sensor was exposed to the 70 μL chamber that was connected to a peristaltic pump and multiposition switching valve. The flow cell was designed to reduce the potential of air bubbles remaining on the crystal after filling from the dry state and to allow air bubbles in the liquid phase to pass out without sticking to the crystal. The oscillation frequency of the QCM sensor was monitored in real time by a Model 400 EQCM system controlled by a laptop PC

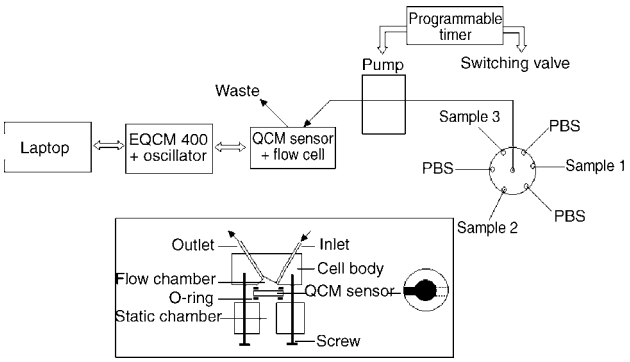


Figure 4. Components of an automatic QCM immunosensor system (top) and the flow cell (bottom).

under the Windows environment. Both the pump and the valve were controlled by a DVSP programmable timer. A schematic diagram of the whole QCM sensor system is illustrated in Fig. 4. The operation of this QCM sensor system was totally automated. As shown in Fig. 5, negative frequency shifts exist between every two neighboring phosphate buffered saline solution (PBS) baselines, which were attributed to the specific adsorption of target bacteria onto the biosensor surface.

APPLICATIONS

Piezoelectric sensors, as simple yet powerful tools, have been extensively employed in detection of chemical and biological agents as well as in the study of chemical, electrochemical, and biological interfacial processes. An online search from SciFinder Scholar (Chemical Abstracts) with the keyword "quartz crystal microbalance" resulted in 4722 references, and 2203 of them are journal articles published during 2000–2005. These studies were conducted to develop (1) antibody/antigen-based biosensors (immunosensors) for detecting biomacromolecules, viruses, cells, as well as small molecules; (2) DNA/RNA probe-based biosensors (genosensors) for *in situ* detection of nucleic acid hybridization; (3) biosensors based on immobilized enzymes, proteins, receptors, lipids, or whole cells; and (4) chemical sensors based on inorganic or organic films for measurements of organic vapors, metal ions, and drugs. Also piezoelectric sensors reported in these studies were used for (1) studies of adsorption of biomolecules and living cells by bare QCM or QCM with functionalized surfaces; and (2) QCM/EQCM investigation/analyses of interfacial phenomena and processes, including self-assembles monolayers (SAMs), films formed using the layer-by-layer assembly technique, molecularly imprinted polymers, biopolymer films, micellar systems, ion transfer at and ion exchange in thin polymer films, doping reactions of conducting polymers, electrodedeposition of metals, and dissolution of metal films.

In the following sections, immunosensors and genosensors, which are most relevant and important to biomedicine, are chosen to discuss the applications of piezoelectric sensors. Some review articles (4–8,21,22) are available for more comprehensive information about applications of piezoelectric sensors.

Immunosensors

One important feature of piezoelectric sensors is that they can be designed as label-free immunosensors. The

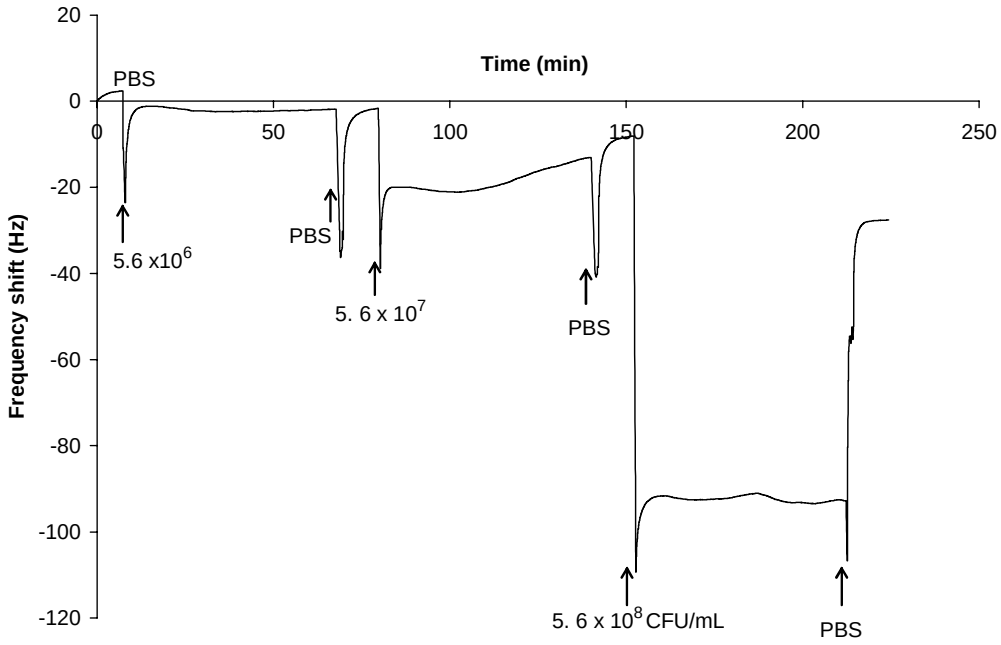


Figure 5. Typical output of an automatic QCM immunosensor system for detection of *S. typhimurium*.

immunosensors taking advantage of antibody-antigen affinity reaction are among the most promising biosensors due to their high specificity and versatility. Conventional immunosensors generally involve the formation of a sandwich immuno-complex consisting of an immobilized primary antibody, captured target analyte, and labeled secondary antibody followed by an optical or electrochemical measurement to detect the label directly or indirectly. Piezoelectric immunosensors do not need a labeled antibody and are thus much simpler and easier in operation than the sandwich immunosensors.

The first piezoelectric immunosensors is reported by Shons et al. (23), who modified a quartz crystal with bovine serum albumin (BSA) and used it to detect anti-BSA antibodies. Since then, numerous piezoelectric immunosensors have been reported for the detection of various analytes from small molecules to biological macromolecules, whole viruses, and cells. In brief, a piezoelectric immunosensor is fabricated by immobilizing a specific antibody/antigen on the surface of an AT-cut PQC. When the immunosensing surface is exposed to a sample solution, a binding reaction occurs between the immobilized antibody/antigen and its complementary part (target analyte). The binding event can be monitored *in situ* by QCM based on the change of surface mass loading and/or other properties such as viscoelasiticity, and thus, the target species is quantitatively detected. Figure 6 illustrates the stepwise assembly and the principle of piezoelectric immunosensor for direct detection of the binding of target analyte and immobilized antibody.

Microbial detection is probably the most common area in which piezoelectric immunosensors are applied. Current practice for effective treatment of infectious diseases without abuse of antibiotics relies on rapid identification of specific pathogens in clinical diagnostics. Nevertheless, conventional methods for microbial detection are inadequate due to being tedious and laborious. Although traditional culture methods hypothetically allow the detection of a single cell, they are extremely time-consuming, typically requiring at least 24 hours and multistep tests to confirm the analysis. Even current rapid methods such as enzyme-linked immunosorbent assay (ELISA) and polymerase chain reaction (PCR) still take several hours to generate only tentative results and require skilled personnel.

Numerous piezoelectric immunosensors have been reported for rapid and specific detection of pathogenic bacteria as alternatives to the conventional methods since the pioneer work of Muramatsu et al. (24), which involved the determination of *Candida albicans* with an AT-cut PQC coated with a specific antibody. Piezoelectric immunosensors have been developed for detection of different bacteria including *S. typhimurium*, *S. paratyphi*, *E. coli*, *E. coli* K12, *Chlamydia trachomatis*, *Yersisinia pestis*, *Candida albicans*, and *Shigella dysenteriae* (25–27). The lower limits of detection typically ranged between 10^5 and 10^7 cells mL^{-1} along with a detection time of tens of minutes to several hours.

QCM has been used to detect various infectious viruses. In the study by König and Gratzel (28), Herpes simplex types 1 and 2, Varicella-zoster virus, Cytomegalovirus, and Epstein–Barr virus were detected using a reusable QCM immunosensor with a detection range from 10^4 to 10^9 cells. They reported that a similar QCM immunosensor could detect Rotavirus and Adenovirus with a linear detection range from 10^6 to 10^{10} cells (25) as well as hepatitis A and B viruses (29). Kosslinger et al. (30) demonstrated the feasibility of detecting HIV viruses using a QCM sensor. Antibodies specific to the HIV were absorbed on the crystal surface, and a serum sample could be detected in 10 min in a flow QCM system.

A piezoelectric immunosensor was developed for the detection of cortisol in a range of 36–3628 ppb (31). Cortisol antibodies were covalently bound onto the Au electrode of a 10 MHz crystal with a water-insoluble polymer and thyroxine antibodies.

Piezoelectric immunosensors have also been frequently reported for determination of biological macromolecules. For example, Kurosawa et al. (32) constructed a high-affinity

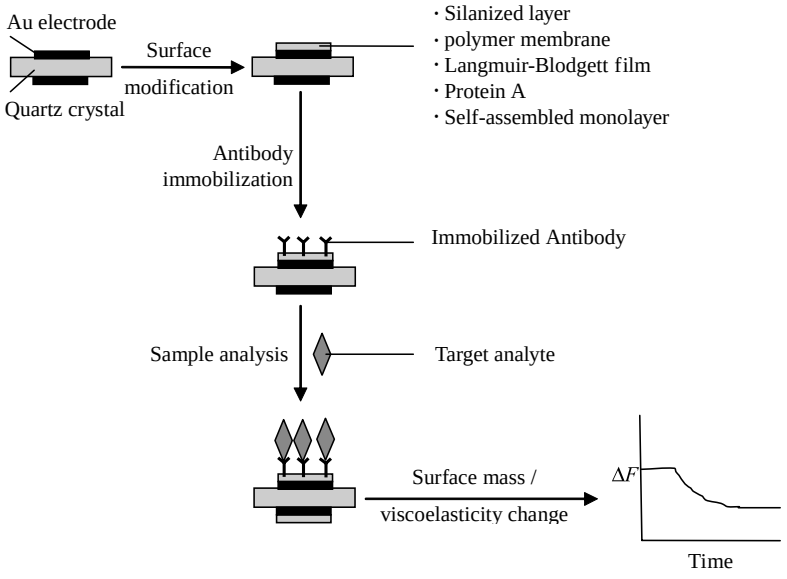


Figure 6. Mechanism of a piezoelectric immunosensor for direct detection of the binding of target analyte and immobilized antibody.

piezoelectric immunosensor using anti-C-reactive protein (CRP) antibody and its fragments for CRP detection. When anti-CRP F(ab')2-IgG antibody was immobilized on the PQC, the detection limit and the linearity of CRP calibration curve were achieved at concentrations from 0.001 to 100 $\mu g \cdot dL^{-1}$ even in serum samples.

Antibody immobilization is vital in successful development of a piezoelectric immunosensor. The current immobilization methods are mainly based on a silanized layer, polymer membrane, Langmuir–Blodgett film, Protein A, and SAM. Protein A, due to its natural affinity toward the Fc region of IgG molecules, has been commonly used to orient antibodies for immunoassays. The orient immobilization has the advantage that it does not block the active sites of the antibodies for binding of target antigens. The procedure for antibody immobilization based on Protein A is simple. Briefly, the Au surface of PQC is coated first with Protein A, and then the antibody is bound to the immobilized Protein A directly. The SAM technique offers one of the simplest ways to provide a reproducible, ultra-thin, and well-ordered layer suitable for further modification with antibodies, which has the potential of improving detection sensitivity, speed, and reproducibility.

Analytical applications of piezoelectric immunosensors based on the direct binding between immobilized antibodies/antigens and target analytes are attractive, owing to the versatility and simplicity of the method. However, the sensitivity of theses approaches is relatively low due to the relatively small numbers of analyte entities that can specifically bind to the limited number of antibody/antigen sites on the surface. Some amplification techniques have been investigated for the sensitivity of piezoelectric immunosensors. Ebersole and Ward (33) reported an amplified mass immunosorbent assay with a QCM for the detection of adenosine 5'-phosphosulfate (APS). The enzymatic amplification led to significant enhancement of the detection sensitivity; levels of approximately 5 $ng \cdot mL^{-1}$ (10^{-14} M) APS reductase could be detected, whereas the direct binding of APS reductase at even more elevated concentrations could not be measured. A sensitive QCM immunosensor was developed by incorporating the Au nanoparticle-amplified sandwiched immunoassay and silver enhancement reaction (34). Au nanoparticle-promoted silver (I) reduction and silver metal deposition resulted in about a two-orders-of-magnitude improvement in human IgG quantification. Su and Li (18) described a piezoelectric immunosensor for the detection of *S. typhimurium* with simultaneous measurements of the changes in resonant frequency and motional resistance (ΔF and ΔR). In the direct detection of *S. typhimurium*, ΔF and ΔR were proportional to the cell concentration in the range of 10^5 to 10^8 and 10^6 to 10^8 $cells \cdot mL^{-1}$, respectively. Using anti-*Salmonella* magnetic microbeads as a separator/concentrator for sample pretreatment as well as a marker for signal amplification, the detection limit was lowered to 10^2 $cells \cdot mL^{-1}$ based on the ΔR measurements.

Genosensors

Piezoelectric genosensors are fabricated by immobilizing a single-stranded (ss) DNA/RNA probe on the PQC surface.

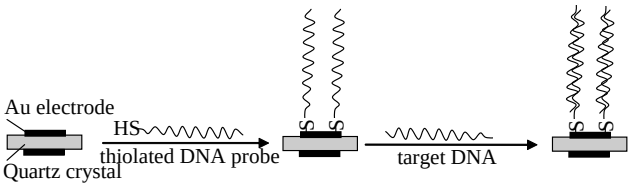


Figure 7. Illustration of piezoelectric genosensor for *in situ* detection of DNA hybridization.

Specific hybridization between the immobilized DNA/RNA probe and its complementary strand in sample causes a change in the resonant frequency of the QCM. Various methods have been used for the immobilization of DNA probes onto the QCM surface. Among these, the SAM method is most commonly used because it offers an ordered, stable, and convenient immobilization. Thiolated oligonucleotides can directly form a SAM on the gold surface of the QCM electrode via the Au-thiolate bond. Figure 7 is an illustration of the piezoelectric genosensor for DNA hybridization detection.

DNA/RNA probes have been applied to detect pathogenic microorganisms in clinical samples. Bacterial and viral pathogens are detectable because of their unique nucleic acid sequences. The DNA/RNA probing process is usually preceded by PCR amplification as target nucleic acid may be present in a sample in very small quantities. Most PCR formats are followed by the detection of amplicons using gel electrophoresis or the membrane-based hybridization method. The former lacks of sensitivity; the latter is more specific, but it requires multistep processing and is thus time-consuming.

Over the past decade, many attempts have been made to develop biosensors for sensitive and reliable detection of DNA hybridization. Fawcett et al. (35) were the first to develop a piezoelectric DNA sensor. Piezoelectric genosensors, due to their simplicity and cost effectiveness, have been recently applied to detect gene mutation, genetically modified organisms, and bacterial pathogens. Su et al. (36) described a piezoelectric DNA sensor for detection of point mutation and insertion mutations in DNA. The method involved the immobilization of ssDNA probes on QCM, the hybridization of target DNA to form homoduplex or heteroduplex DNA, and finally the application of MutS for the mutation recognition. By measuring the MutS binding signal, DNA containing a T:G mismatch or unpaired base was discriminated against perfectly matched DNA at a target concentration ranging from 1 nM to 5 μM.

The sensitivity of piezoelectric genosensors may be improved through optimizations of probe immobilization or by means of signal amplification using anti-dsDNA antibodies, liposomes, enzymes, or nanoparticles. A nanoparticle is an effective marker for mass amplification as it has relatively greater mass in comparison with a DNA molecule. Amplified with nanoparticles, the limit of DNA detection by QCM can be lowered for several orders to as low as 10^{-16} M (37). Mao et al. (38) developed a piezoelectric genosensor for the detection of *E. coli* O157:H7 with nanoparticle amplification, in which thiolated ssDNA probes specific to *E. coli* O157:H7 *eaeA* gene were

immobilized onto the QCM surface through self-assembly and streptavidin conjugated Fe_3O_4 nanoparticles were used as "mass enhancers" to amplify the detection signal. As low as 10^{-12} *M* synthesized oligonucleotides and 2.67×10^2 cells·mL^{-1} of *E. coli* O157:H7 could be detected by the piezoelectric genosensor.

CONCLUSIONS

Piezoelectric sensors, as sensitive mass sensors or QCMs, have been applied to detect and measure a broad variety of biomedical analytes in both gas and liquid phases based on the adsorption/desorption of target analyte(s) on/from the sensor surface, in which the selectivity is controlled by the sensing material. Piezoelectric sensors are more than pure mass sensors as they are also capable of detecting subtle changes in the solution–surface interface that can be due to density-viscosity changes in the solution, viscoelastic changes in the bound interfacial material, and changes in the surface free energy. The most attractive advantage of QCMs is that they are suitable for label-free detection and flow-through, in real-time detection. However, using the direct detection approach, the sensitivity of QCM is inadequate for some applications in biomedical analysis such as detection of low levels of pathogens and other biological agents in clinical samples. The lack of sensitivity may be addressed by employing amplification techniques as introduced earlier, by using piezoelectric films and bulk silicon micromaching techniques to manufacture high-frequency QCMs (21), or by designing devices of other acoustic wave modes such as SAW, APM, and FPW.

BIBLIOGRAPHY

1. Curie J, Curie P. An oscillating quartz crystal mass detector. Comp Rend 1880;91:294–297.
2. Sauerbrey GZ. Use of quartz vibration for weighing thin films on a microbalance. J Phys 1959;155:206–212.
3. King WH Jr. Piezoelectric sorption detector. Anal Chem 1964;36:1735–1739.
4. Guilbault GG, Jordan JM. Analytical uses of piezoelectric crystals: A review. CRC Crit Rev Anal Chem 1988;19:1–28.
5. Ward MD, Buttry DA. In situ interfacial mass detection with piezoelectric transducers. Science 1990;249:1000–1007.
6. Buttry DA, Ward MD. Measurement of interfacial process at electrode surfaces with the electrochemical quartz crystal microbalance. Chem Rev 1992;92:1355–1379.
7. Janshoff A, Galla H-J, Steinem C. Piezoelectric mass-sensing devices as biosensors—An alternative to optical biosensors? Angew Chem Int Ed 2000;39:4004–4032.
8. Marx KA. Quartz crystal microbalance: a useful tool for studying thin polymer films and complex biomolecular systems at the solution-surface interface. Biomacromolecules 2003;4: 1099–1120.
9. Buck RP, Lindner E, Kutner W, Inzelt AG. Piezoelectric chemical sensors. Pure Appl Chem 2004;76:1139–1160.
10. Kanazawa KK, Gordon JG. The oscillation frequency of a quartz resonator in contact with a liquid. Anal Chim Acta 1985;175:99–105.
11. Martin SJ, Granstaff VE, Frye GC. Characterization of a quartz crystal microbalance with simultaneous mass and liquid loading. Anal Chem 1991;63:2272–2281.
12. Qu X, Bao LL, Su X-L, Wei W. Rapid detection of *Escherichia coli* form with a bulk acoustic wave sensor based on the gelation of *Tachypleus* amebocyte Lyste. Talanta 1998;47: 285–290.
13. Gee WA, Ritalahti KM, Hunt WD, Loffler FE. QCM viscometer for bioremediation and microbial activity monitoring. IEEE Sens J 2003;3:304–309.
14. Muramatsu H, Tamiya E, Karbue I. Computation of equivalent circuit parameters of quartz crystals in contact with liquids and study of liquid properties. Anal Chem 1988;60: 2142–2146.
15. Zhou T, Nie L, Yao S. On equivalent circuits of piezoelectric quartz crystals in a liquid and liquid properties, Part I, Theoretical derivation of the equivalent circuit and effects of density and viscosity of liquids. J Electroanal Chem Interf Electrochem 1990;293:1–18.
16. Nöel MAM, Topart PA. High-frequency impedance analysis of quartz crystal microbalance, 1. General considerations. Anal Chem 1994;66:484–491.
17. Xie Q, Wang J, Zhou A, Zhang Y, Liu H, Xu Z, Yuan Y, Deng M, Yao S. A study of depletion layer effects on equivalent circuit parameters using an electrochemical quartz crystal impedance system. Anal Chem 1999;71:4649–4656.
18. Su X-L, Li Y. A QCM immunosensor for *Salmonella* detection with simultaneous measurements of resonant frequency and motional resistance. Biosens Bioelectron. In press.
19. O'Sullivan CK, Guilbault GG. Commercial quartz crystal micobalances theory and applications. Biosens Bioelectron 1999;14:663–670.
20. Su X-L, Li Y. An automatic quartz crystal microbalance immunosensor system for *Salmonella* detection. ASAE Paper No. 047043. St. Joseph, MI: The American Society of Agricultural Engineers; 2004.
21. Martin SJ, Frye GC, Spates JJ, Butler MA. Gas sensing with acoustic devices. Proc-IEEE Ultrasonics Symp 1996;1:423–434.
22. Vaughan RD, Geary E, Pravda M, Guilbault GG. Piezoelectric immunosensors for environmental monitoring. Int J Environ Anal Chem 2003;83:555–571.
23. Shons A, Dorman F, Najarian J. The piezoelectric quartz immunosensor. J Biomed Mater Res 1972;6:565–570.
24. Muramatsu H, Kajiwara K, Tamiya E, Karube I. Piezoelectric immunosensor for detection of *Candida albicans* microbes. Anal Chim Acta 1986;188:257–261.
25. Konig B, Gratzel M. Detection of viruses and bacteria with piezoelectric immunosensor. Anal Lett 1993;26:1567–1575.
26. Ivnitski D, Abel-Hamid I, Atanasov P, Wilkins E. Biosensors for detection of pathogenic bacteria. Biosens Bioelectron 1999;14:599–624.
27. Deisingh AK, Thompson M. Detection of infectious and toxigenic bacteria. Analyst 2002;127:567–581.
28. König B, Gratzel M. A novel immunosensor for Herpes virus. Anal Chem 1994;66:341–348.
29. König B, Gratzel M. Long term stability and improved reusability of piezoelectric immunosensor for human erythrocytes. Anal Chim Acta 1993;280:37–42.
30. Kösslinger C, Crost S, Aberl F. A quartz crystal microbalance for measurements in liquids. Biosens Bioelectron 1992;7: 397–410.
31. Attili BS, Suleiman AA. Peizoelectric immunosensor for detection of cortisol. Anal Lett 1995;28:2149–2159.
32. Kurosawa S, Nakamura M, Park JW, Aizawa H, Yamada K, Hirata M. Evaluation of a high-affinity QCM immunosensor using antibody fragmentation and 2-methacryloyloxyethyl phosphorylcholine (MPC) polymer. Biosens Bioelectron 2004;20:1134–1139.

33. Ebersole RC, Ward MD. Amplified mass immunosorbent assay with a quartz crystal microbalance. J Am Chem Soc 1988;110: 8623–8628.
34. Su X, Li SFY, O'Shea SJ. Au nanoparticle- and silver-enhancement reaction-amplified microgravimetric biosensor. Chem Commun 2001; 755–756.
35. Fawcett NC, Evans JA, Chen LC, Drozda KA, Flowers N. A quartz crystal detector for DNA. Anal Lett 1988;21:1099–1110.
36. Su X, Robelek R, Wu Y, Wang G, Knoll W. Detection of point mutation and insertion mutations in DNA using a quartz crystal microbalance and MutS, a mismatch binding protein. Anal Chem 2004;76:489–494.
37. Liu T, Tang J, Jiang L. The enhancement effect of gold nanoparticles as a surface modifier on DNA sensor sensitivity. Biochem Biophys Res Commun 2004;313:3–7.
38. Mao X, Yang L, Su X-L, Li Y. A nanoparticle-based quartz crystal microbalance DNA sensor for the detection of *Escherichia coli* O157:H7. Biosens Bioelectron. In press.

See also COCHLEAR PROSTHESES.

PLETHYSMOGRAPHY.

See IMPEDANCE PLETHYSMOGRAPHY.

PNEUMATIC ANTISHOCK GARMENT.

See SHOCK, TREATMENT OF.

PNEUMOTACHOMETERS

NARCISO F. MACIA
Arizona State University at the Polytechnic Campus Mesa, Arizona

INTRODUCTION

From the beginning of time, breathing has been important since it has been the most common indicator of life. The Bible indicates that God infused life into man by "blowing into his nostrils the breath of life" (1). Not surprisingly then, the way in which we breath is an important indicator of our health. Consequently, the medical profession has tried to learn our physical condition by focusing on the behavior of the respiratory system. Two main mechanisms take place in the breathing process: (1) Movement of gases from the nose and mouth to the alveoli, and (2) CO_2 and O_2 gas exchange at the alveoli. An interesting historical perspective, standardization of pulmonary function tests (PFTs) and the associated equipment received a great push from the mobile PFT trucks that were part of a campaign to eliminate lung cancer in the United States in the 1950s. Even today, PFTs are often used as a preliminary screen for lung cancer.

This section focuses on the equipment used to make flow and volume measurements in the first category: pneumotachometers, also known as respirometers, spirometers, or simply flowmeters.

There are two clinical areas where flow measurement devices are used. These are (1) the field of spirometry (2–4), dealing with the actual performance of the respiratory system as reflected by the volumes that the lung can realize, and the speed in which these volumes can be moved in and out of the lungs. Indicators such as tidal volume (TV or V_t) and vital capacity (VC) provide a glimpse of the range of motion of the lungs. Similarly, parameters, such as FEV_1 (forced expiratory volume in 1 s) and $FEF_{25-75\%}$ (forced expiratory flow at the mid-portion of forced vital capacity). Notice that these parameters are the result of: (a) the patient's ability to cooperate, (b) condition of the diaphragm, the respiratory system's main workhorse, (c) range of motion of the lungs, and (d) the mechanical components associated with the respiratory pathways (size of the conducting airways). (2) The field of parameter estimation (5–12), dealing specifically with the noninvasive measurement of components descriptive of the mechanical characteristics of the respiratory apparatus. These parameters include resistance, compliance, and inertance. One advantage of this approach is its independence from the patient's ability to cooperate. This approach is particularly useful in unconscious, and very young or very old patient populations. However, this approach requires a much higher level of computation.

Before proceeding with a presentation on pneumotachometer, some definitions and conventions are appropriate.

Open and Closed Systems

In many pulmonary function tests, the test procedure requires that the subject inhales maximally, then place the pneumotachometer in then mouth, and then exhales as fast as possible. This type of set up is referred to as an "open" system since no exhaled air is rebreathed. In other types of pulmonary function testing, the patient exchanges air back and forth with a reservoir. This later system is referred to as a "closed" system.

Variable Used for Flow

The most common variable used to describe flow is $\dot{V}$. The "dot" comes from the mathematical notation of differentiation with respect to time, originally developed by Sir Isaac Newton, or

$$\dot{V} = \frac{dV}{dt}$$

which implies that flow, $\dot{V}$, is the time rate or change of lung volume, V.

Polarity

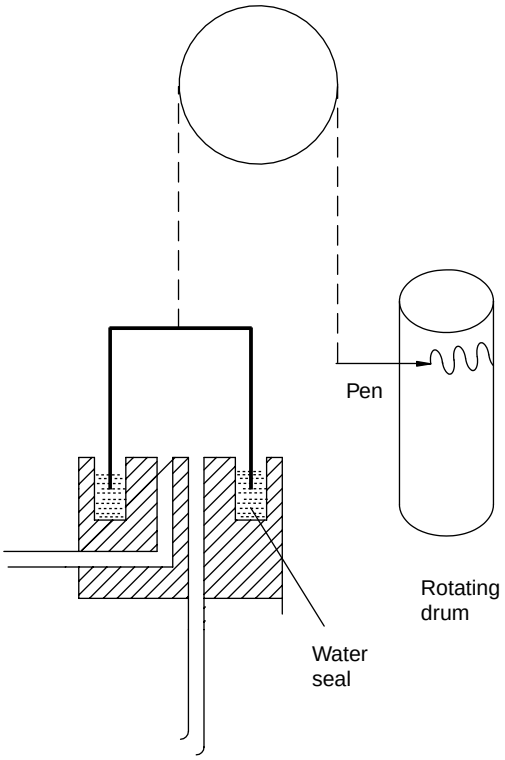
Spirometer tests consider expiratory flow as positive while parameter estimation procedures look at inspiration as positive. Perhaps it is reflective that most spirometer tests are performed during expiration while in parameter estimation procedures, inspiration is the primary arena.

DEVICES FOR MEASURING RESPIRATORY FLOW

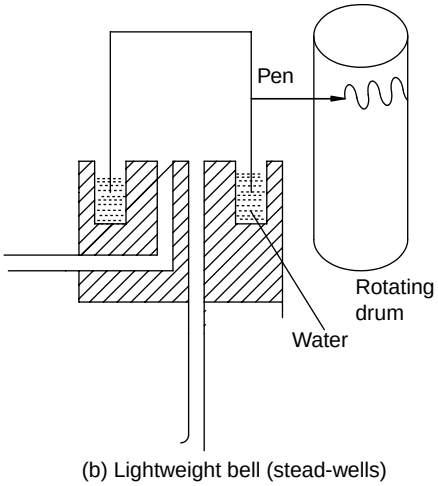
This section presents devices that have and continue to be used for measuring respiratory flow and volume. Even though some of them are not used as often, they are part of the toolbox that clinicians and researchers have used in getting a handle in the respiratory system.

Volume Displacement Type

This type of device, often called volumetric type, captures the volume of gas coming out of the subject's mouth with an expanding reservoir whose degree of expansion can be recorded either mechanically or electronically. Flow can be obtained by differentiating the changing volume. This type of instrument is still used in many pulmonary function facilities and exercise physiology laboratories, since they offer the highest accuracy available. There two types that use a water seal: (1) the counterbalanced bell and (2) the lightweight bell over a water seal (often called the Stead–Wells spirometer after the individuals who requested the device. These are shown in Fig. 1. Even

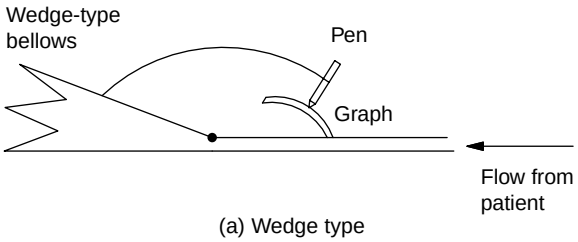


(a) Counterbalanced bell

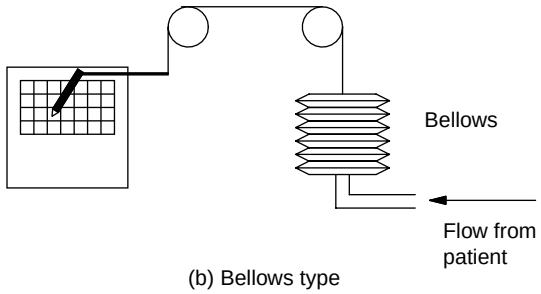


(b) Lightweight bell (stead-wells)

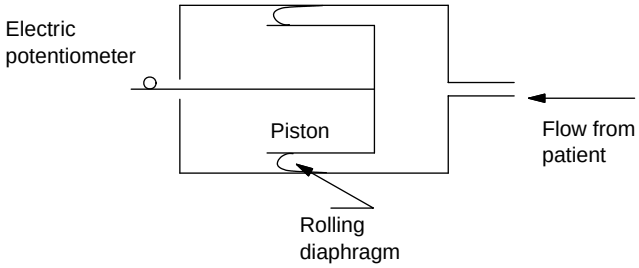
Figure 1. Water seal spirometer.



(a) Wedge type



(b) Bellows type



(c) Rolling diaphragm type

Figure 2. Non-water seal spirometers: (a) wedge type, (b) the bellows, and the (c) rolling diaphragm.

though these devices have served the medical community well, their bulkiness and expense have motivated biomedical equipment developers to design smaller, more portable nonvolumetric units, even though they are not as accurate as the volumetric type. Three other variations of the volume displacement type exist: the wedge type, the bellows, and the rolling diaphragm, as shown in Fig. 2 (13).

Other Applications of Volume Displacement Type. The water seal spirometers have also been used to monitor breathing over longer periods of time. Simply closing the circuit creates some problems since O_2 in the air is being consumed while and the mixture becomes progressively CO_2 rich. To solve this problem, the bell is originally filled with O_2, and a CO_2 scrubber (Baralyme, a trade name for generic BaO) is inserted into the circuit. After a few seconds, additional oxygen is inserted into the circuit. The setup and the resulting the waveform are shown in Fig. 3. This device has found applications to evaluate the metabolic rate (proportional to O_2 consumption).

This device has also been used to measure compliance, the elasticity of the respiratory system (RS). It is carried out by adding a series of weights on top of the spirometer bell, which increases the overall system pressure (14,15). The corresponding change in system volume

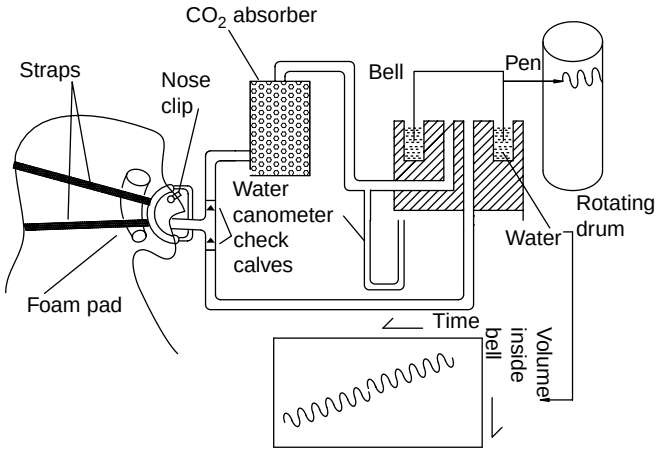


Figure 3. Closed-circuit spirometer.

detected at the bell, which corresponds to the increase in lung volume, is measured. The resulting respiratory compliance is given by

$$C_{\mathrm{RS}} = \frac{\Delta V}{\Delta P}$$

This approach has been automated by Crawford (16). Notice that the resulting compliance, the compliance of the respiratory system, C_{RS}, captures both the compliance of the lungs, C_{L}, and the compliance of the chest wall C_{CW}:

$$\frac{1}{C_{\mathrm{RS}}} = \frac{1}{C_{\mathrm{L}}} + \frac{1}{C_{\mathrm{CW}}}$$

and the lung compliance is made up of the compliance of the left and right lobes.

$$C_{\mathrm{L}} = C_{\mathrm{L-left}} + C_{\mathrm{L-right}}$$

The following type of pneumotachometers measure flow directly, instead of measuring change in volume. These devices electronically integrate the flow signal to obtain volume.

FLUID RESISTANCE TYPE OF PNEUMOTACHOMETERS

Fleish Pneumotachometers

Another device that has been and continues to be used extensively is the Fleish (or Fleisch) pneumotachometer (17). It consists of a fluid resistive device through which the air passes. The pressure drop across a fluid resistive element created by the respiratory flow is applied to a pressure transducer. Since there is a linear correlation between the measured pressure drop and flow, flow can be determined from the pressure drop. This approach has also been the primary workhorse of pulmonary instrumentation. There are two areas of concern with this type of device: (1) the linearity of the correlation between pressure drop and volumetric flow, and (2) the potential condensation of vapor droplets in the resistive element. Both of these factors can affect the device's accuracy. The American Thoracic Society (ATS), the medical section of the American Lung Association (18), has published standards regarding the required accuracy of the equipment (5% in some tests, while 3% in others) They have also established the conditions in which the results should be reported: BTPS. [BTPS stands for body conditions: normal body temperature (37 °C) ambient pressure saturated with water vapor].

Methods for Obtaining a Linear Flow-Pressure Drop Relationship. The design of the fluid resistive element in the Fleish pneumotachometer produces laminar flow. Three general approaches have been implemented to obtain this linearity. The first one uses capillary tubing placed in parallel (bundle); The second method uses a coiled metal strip with capillary tubing-like corrugations; The third one uses a porous medium, for example, a screen or paper similar to what is used in a vacuum cleaner bag.

In the capillary version, if the flow is laminar, the resulting pressure drop is given by

$$\Delta p = \frac{128\,\mu L}{N\pi D^4}\dot{V}$$

where Δp is the pressure drop, L is the length of the capillaries, N is the number of capillaries, D is the diameter of the capillaries, and μ is the absolute viscosity. The above equation can be expressed as

$$\Delta p = R\dot{V}$$

where R is the linear fluid resistive coefficient. Even though effort is taken to make the flow laminar, there are always some turbulent components. This turbulent behavior occurs primarily at the entrance and exit of the capillary tube bundle.

If the pressure drop is created with a square-edge orifice, the flow most likely will be turbulent, and the pressure drop is given by

$$\Delta p = \frac{\rho}{2C_{\mathrm{D}}A^2}\dot{V}^2$$

where ρ is the density, C_{D} is the discharge coefficient, and A is the area of the orifice. Since this function is truly an odd function [$f(-x) = -f(x)$], the even function above [$f(-x) = f(x)$] is modified by means of the absolute value sign is

$$\Delta p = \frac{\rho}{2C_{\mathrm{D}}A^2}|\dot{V}|\dot{V}$$

The above equation can be expressed as

$$\Delta p = k|\dot{V}|\dot{V}$$

where k is the nonlinear (quadratic) fluid resistive coefficient. In actual practice, most fluid resistors can be described as a combination of laminar and turbulent components:

$$\Delta p = R\dot{V} + k|\dot{V}|\dot{V}$$

where R and k are the linear and nonlinear (quadratic) fluid resistive coefficients. The expression for flow in terms of pressure drop is

$$\dot{V} = \frac{-R + \sqrt{R^2 + 4k\Delta p}}{2k}$$

The reader might suggest to themselves: Why not simply use a square-edge orifice and linearize the resulting signal (i.e., take the square root of the resulting pressure to obtain flow)? In some cases this is done, however, it presents some challenges. If the dynamic range of the flow (maximum flow to minimum flow) is 10–1, the resulting range of the corresponding pressure drops would be 100–1, a pressure signal that would be either too small (and consequently being difficult to measure) or too large (and consequently offering some detectable resistance to the patient).

The other concern with any fluid resistance type flowmeter is condensation. If the fluid resistive element of the flowmeter is at a lower temperature than the air exhaled from the subject (37 °C, saturated), there is a likelihood that the water vapor present in the air would condense on the fluid resistive element, changing the flow–pressure drop relationship. To avoid this problem, most Fleish-like pneumotachometers use a heating element to keep the fluid resistive element hotter than the flow. Despite these drawbacks, fluid resistance type pneumotachometers continue to be one of the most popular methods of measuring respiratory flow.

Osborne Pneumotachometer

The last section introduced problems associated with a square edge orifice for producing the necessary pressure drop. An innovative idea that overcomes these problems is the variable area Osborne flowmeter (19,20). In this flowmeter, the resistive element consists of a thin disk with a flap cut into it, as shown in Fig. 4. As the flow attempts to pass through the space between the flap and the disk, the flap bends, increasing the effective area of the orifice. As a result, the square-type relationship between flow and pressure no longer occurs; instead a linear one is obtained. The specific pressure–flow relationship is captured electronically at the time of calibration and later used to obtain flow from the measured pressure drop.

NONFLUID RESISTANCE TYPE OF PNEUMOTACHOMETERS

Respirometers

Another type of flow and volume indicator is called the Wright respirometer (21,22). It operates on the principle that the moving gas imparts movement on a rotating vane. In the newer units, the motion of the vane is sensed electronically. This type of device is often classified as turbine type.

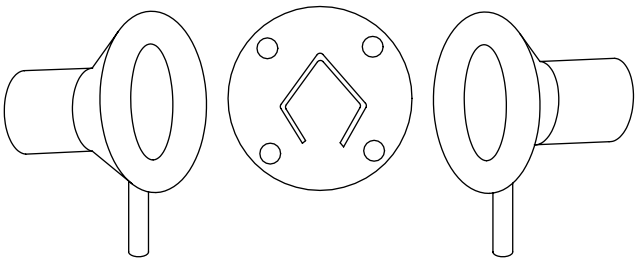


Figure 4. Flow element of the Osborne flowmeter (Adapted from Ref. (20).

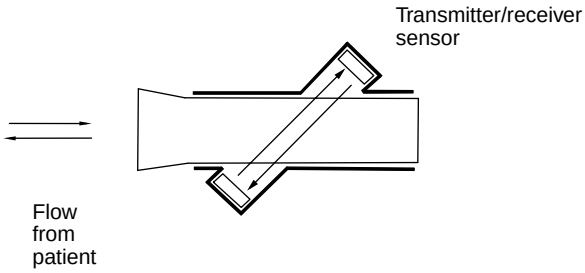


Figure 5. Ultrasound acoustic flowmeter. (Adapted from easy one literature, ndd medical technologies).

Another similar device made by the Wright company (13) records flow instead of expired volume. As flow passes through the device, it simultaneously deflects a vane that as it moves it opens additional passageways through which the flow exits. This mechanical unit is used for measuring peak flow.

Ultrasound–Acoustic Pneumotachometers

This flowmeter uses the Doppler phenomenon, which states that sound travels faster if the medium is also moving. It is very attractive since it provides a property-independent measurement of flow, that is, the measurement is independent of gas composition, pressure, temperature, and humidity. It does not compensate for changes to the air as it enters the respiratory system. (See Calibration section). Until recently, this type of flowmeter was too expensive for regular clinical use. One implementation (23) utilizes a sound pulse along a path that intersects the respiratory flow at an angle, as shown in Fig. 5. A pair of transmitters sends and receives sound in an alternating fashion. The sound pulse gets to the receiver faster if the sound wave is traveling in the same direction as the measured flow. On the other hand, the sound pulse that opposes the direction the respiratory flow takes longer. The difference between these two transit times is used for calculating flow.

Thermal Units

Another class of flowmeter uses thermistors (24,25) to measure flow. Thermistors are electrical resistors made of a material whose resistance decreases with temperature. As the flow passes by the thermistor bead, it attempts to decrease its temperature, which translates to an increase in resistance. An electronic circuit supplies the necessary current to maintain the thermistor at a constant temperature. As a result, the change in current is proportional to gas flow. Often these units are referred to as the hot-wire anemometer type.

Vortex Shedding Flowmeter

Whenever a flow field passes a structure (bluff body), it produces eddies (turbulence) past the structure (26,27). For a particular flow range, the frequency of the shedding is proportional to flow. This principle has been used to measure flow. The flowmeter made by Bourns Medical Systems (28) uses an ultrasonic transducer–receiver pair

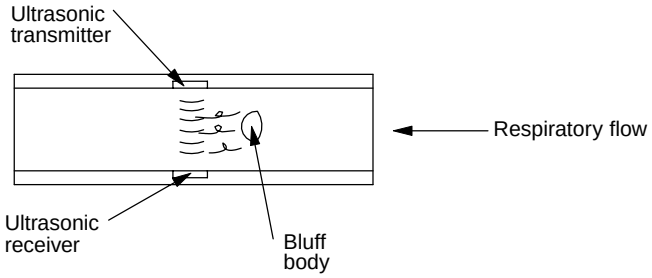


Figure 6. Ohio vortex (Bluff Body) respiration monitor. (Adapted from Ref. 13).

that detects the vortex created by the flow, and converts it to an electric signal. The manufacturer of the device claims that gas composition, temperature, and humidity will not affect the measurement process.

Flutter Flowmeter

Whenever flow passes a flexible structure, there is a possibility of a fluid–structure interaction, which makes the structure vibrate. This phenomenon is often called flutter. In aircraft, this is a major concern because wings that experience this phenomenon can break off; consequently aircraft designers have learned to avoid such a condition. On the other hand, developers of respiratory flowmeters have taken advantage of this principle to measure flow. The Ohio Vortex Respiration Monitor shown in Fig. 6 utilizes a light beam–photoelectric eye to capture the resulting vibration and convert it to flow.

Lift Force Gas Flow Sensor

Airfoils, (the shape of the wing in an aircraft), produce lift when subjected to a flow field. Svedin (29–31) is using this concept to measure respiratory gasses in medical applications. The sensor consists of two plates with polysilicon strain gages connected to a Wheatstone bridge. The plates deflect in response to the resulting lift force, in a direction normal to the flow field. One of the claims made by the developers of this device is the sensor's relative insensibility to inertial forces.

Fluidic Oscillator

Fluidics is a technology that was invented in 1959 and has been used in analogue and digital applications (32,33). It is very similar to pneumatics, but few or no moving parts are used. The devices can also be operated using liquids. The device that gave birth to the technology is the fluid amplifier, a device that with no moving parts can amplify a pressure differential. Figure 7 shows a cross-section of an amplifier, made by staking several perforated stainless steel sheets. Several amplifiers can be staged to achieve gains close to one-half of a million. The most successful fluidics device is the windshield water spray found in most cars. The windshield cleaning fluid or water enters into a cavity that makes the exiting jets oscillate, as shown in Fig. 8. Many applications are possible with this novel technology (34). It must be clarified that prior to the advent of MEMS and microchannels, the word "fluidics" was used

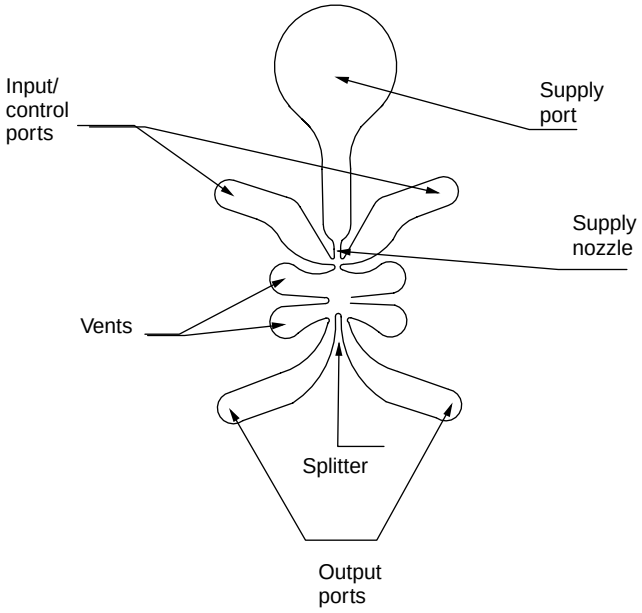


Figure 7. Fluidic fluid amplifier.

only in dealing specifically with this type of devices. Today, any small channel carrying a fluid is called a fluidic device or a microfluidic channel.

The fluidic oscillator flowmeter (unidirectional) offers much promise. It is constructed by configuring a fluid amplifier with feedback: the outputs are connected to the inputs, as shown in Fig. 9. This produces an oscillation whose frequency is proportional to flow. Even though this type of flowmeter has not been specifically used to measure respiratory flow, its performance has been demonstrated successfully as a flow sensor in gasoline delivery systems (35). It aspirates the displaced air in the fuel tank through small holes in the delivery nozzle. It withdraws a volumetric flow equal to the gasoline volumetric flow delivered to the tank. The purpose of course is to minimize the discharge of unburned hydrocarbons into the atmosphere.

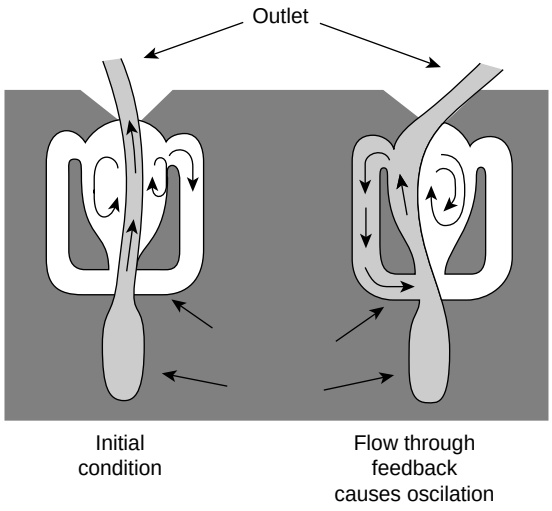


Figure 8. Fluidic windshield water spray.

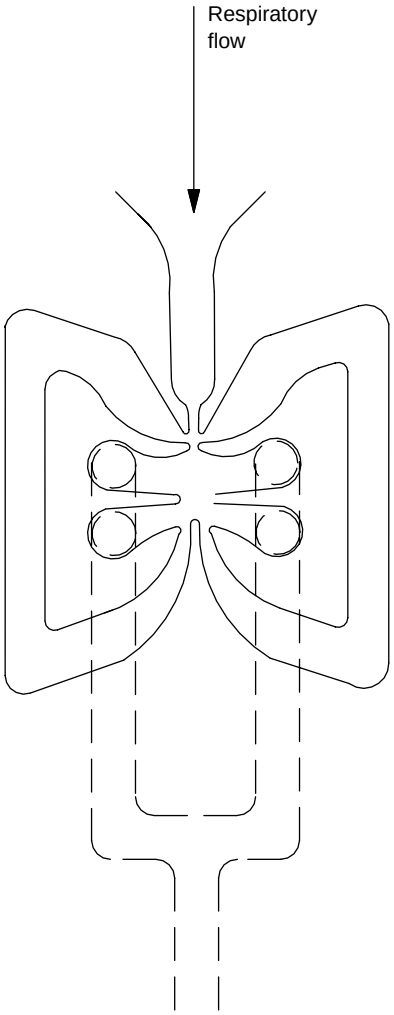


Figure 9. Amplifier with positive feedback: fluidic oscillator.

COMMERCIALLY AVAILABLE SPIROMETERS

For a list of most of the commercially available spirometers, see the online source of information by ADVANCE for Managers of Respiratory Care (36) and American Association of Respiratory Care (37).

CALIBRATION

Calibration of Volume Displacement Type Spirometers

This type of volumetric device (Figs. 1 and 2) rarely loses calibration since the geometry is fixed, especially in the type where the measurements of the volume changes are recorded by means of a pen writing on a revolving graph. Even in systems that have electronic position sensors, due to the robustness of these components, recalibration is seldom necessary. However, recalibration should be done on a regular basis to insure the absence of artifacts such as leaks, rough spots, and so on.

The most common method for performing this procedure is the calibrated syringe: a large piston cylinder device that allows the piston to move within two stops and delivering a fixed volume (typically 3L). The calibration process consists of pulling the piston out to the outer stop, connecting the syringe to the spirometer, and gently pushing the piston to the other stop. The resulting volume indication should be that of the calibrated syringe volume.

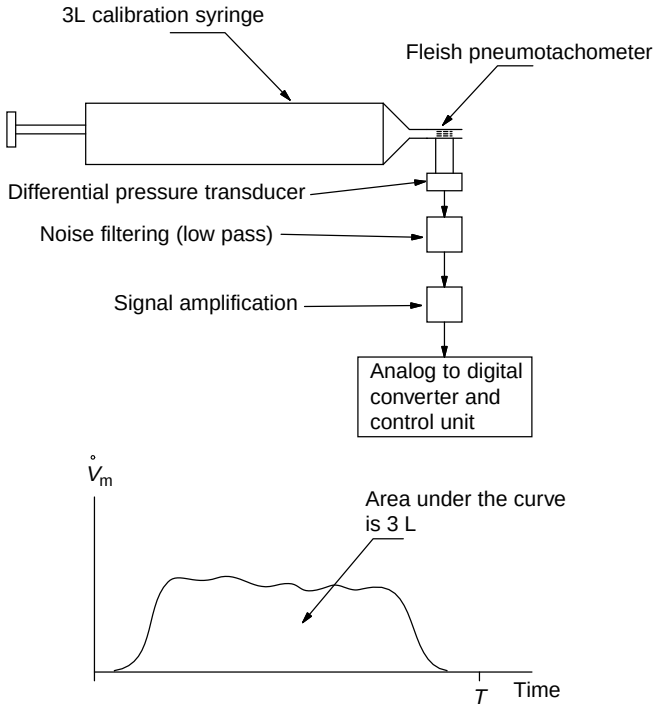


Figure 10. Syringe calibration procedure.

Calibration of Spirometers other than the Volume Displacement Type

There are three approaches for calibrating pneumotachometer that do not utilize volume displacement. They are (1) syringe, (2) wet test gas meter; and (3) comparison with a standard calibration flowmeter.

Syringe Approach. In performing this calibration procedure, it is assumed that the pneumotachometer is integrated into a data acquisition–calibration system. To perform the calibration, a fixed volume is passed through the pneumotachometer using a calibrated syringe, typically 3L. The data acquisition system records the resulting flow signal, $\dot{V}_M$, as shown schematically in Fig. 10. Then it integrates it (i.e., finds the area under the curve) to obtain the measured volume:

$$V_M = \int_0^T \dot{V}_M \, dt$$

Then it determines the correction factor, K, so that:

$$KV_M = 3$$

Afterward it uses K to adjust the measured flow:

$$\dot{V}_{\text{TRUE}} = K\dot{V}_M$$

This approach assumes that the relationship between pressure drop and flow are linear. The user should consult the user's manual, since some of the pneumotachometers automatically adjust for BTPS.

Wet Test Gas Meter. This device is similar to the gas meter found in homes and industrial sites (38). It consists of a series of constant volume chambers that are filled by the incoming flow. They are attached to a rotating structure that enables the chambers to go through the following sequence: (1) fill the chamber with the incoming air, (2) create a water seal at the bottom of the chamber, (3) move the sealed chamber to the exhaust side, and (4) release the volume in the chamber to the outside. As a result, the incoming flow imparts a rotation that is observable from the outside by means of rotating hands. Consequently, there is a 1:1 correlation between the volume that passes through the device and the number of rotations indicated by the hand. To perform a calibration test, a known flow of gas is passed through the Wet Test Gas Meter and the flow recorded:

$$\dot{V} = \frac{(\text{number of revolutions})(\text{volume/revolution})}{\text{time required for above revolutions}}$$

Immediately afterward, the same flow is passed through the pneumotachometer under calibration. The constant flow is produced by applying a source of high pressure to a needle valve. Since the pressure resistance created by the Wet Test Gas Meter or the pneumotachometer is very small (a couple of inches of water), the flow is determined by the upstream pressure, Ps and the size of the orifice in the needle valve. The resulting output signal, *e*, is recorded, as shown in Fig. 11. The procedure is repeated for various needle valve settings to produce a curve similar to the one

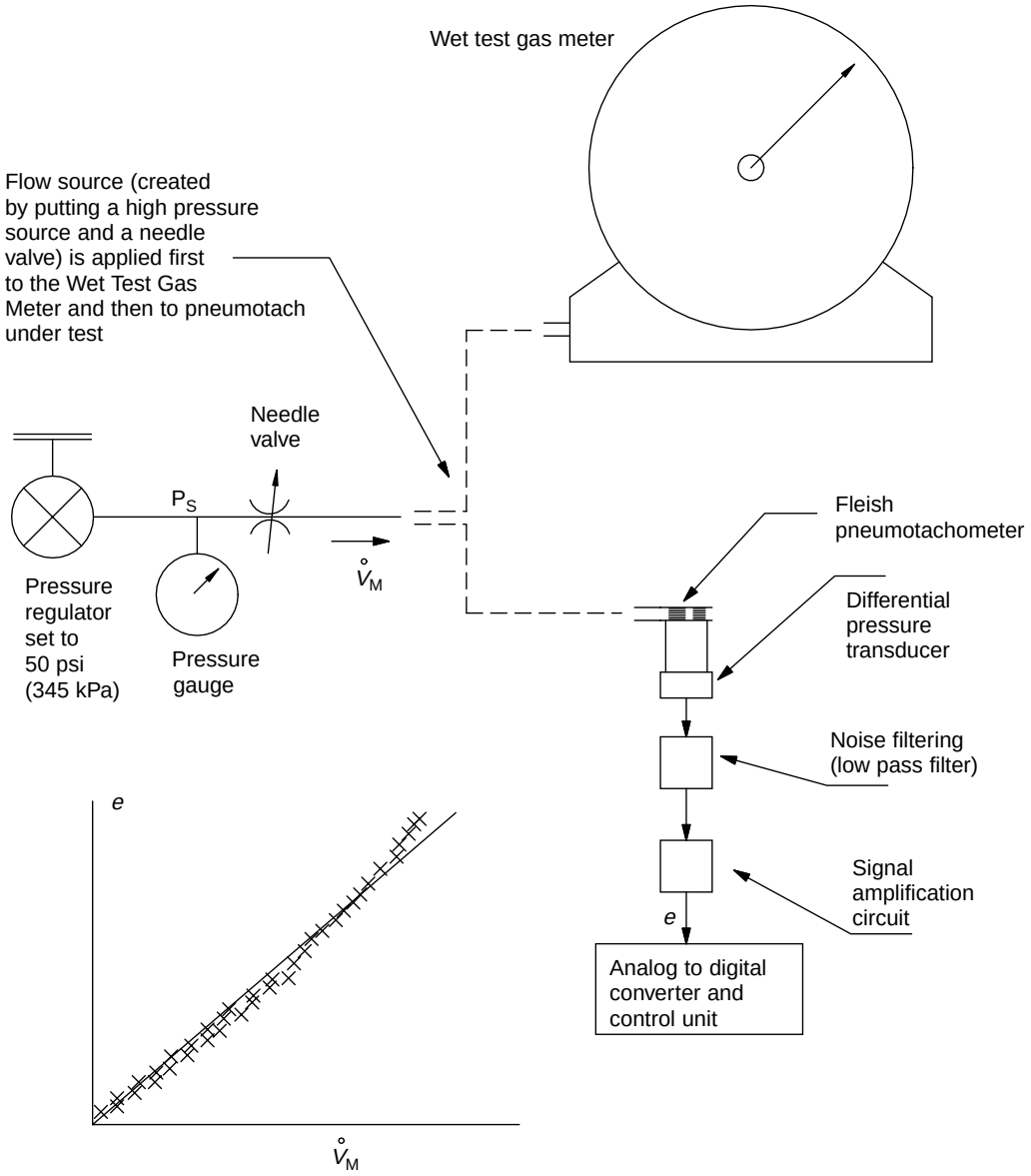


Figure 11. Calibration procedure using the Wet Test Gas Meter.

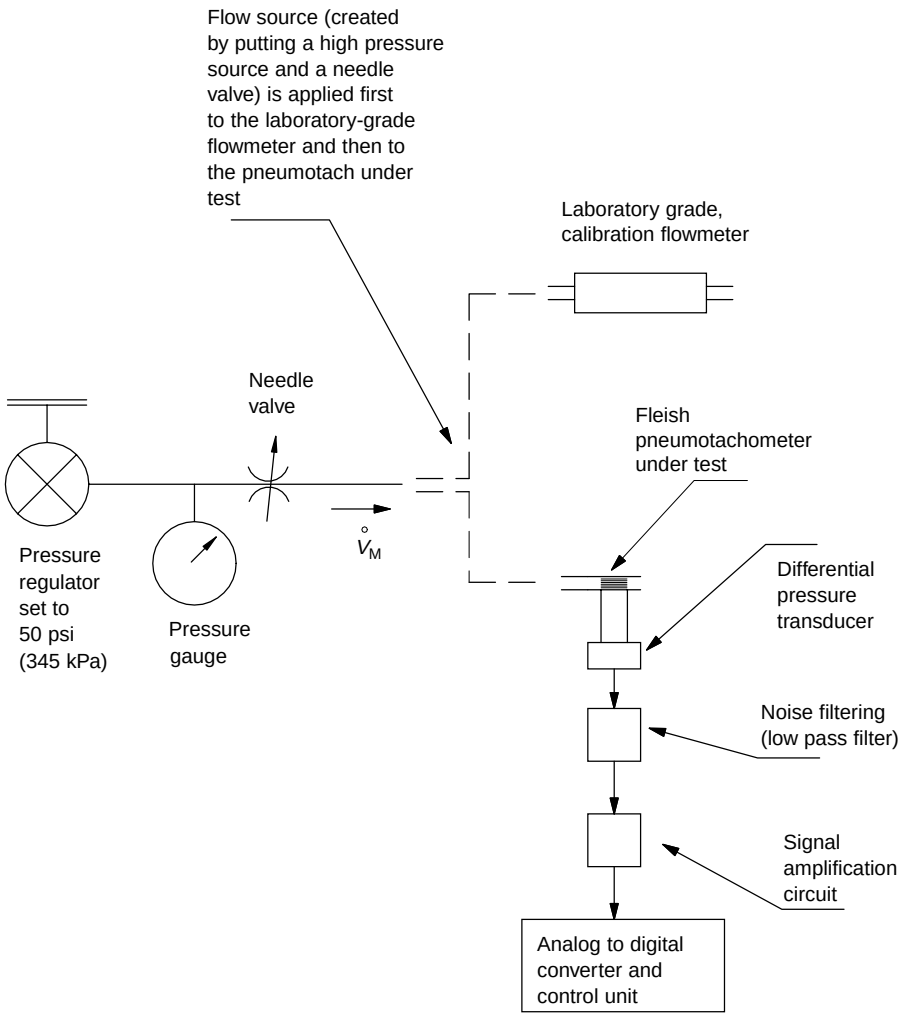


Figure 12. Calibration using a Laboratory grade Flowmeter.

shown in Fig. 11 that can then be used to obtain either a linear approximation or a polynomial fit. Afterward, the calculated flow is adjusted for conversion into standard conditions: BTPS [normal body temperature (36 °C) ambient pressure saturated with water vapor]. This type of calibration device has an accuracy of ±0.5%.

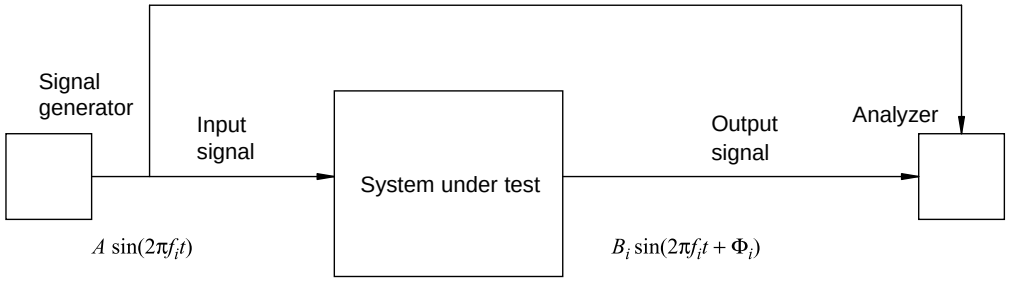
Calibration Using a Laboratory Grade Flowmeter. This method is possible when a laboratory-grade flowmeter is available. As done in the Wet Test Gas Meter, a series of flows are passed through both the laboratory grade flowmeter and the pneumotachometer under test, as shown in Fig. 12. The output of the pneumotachometer under test is made to agree with the flow indicated by the laboratory grade flowmeter.

Piston Prover. Flow meters can also be calibrated using a piston prover (39). This device consists of a precision bore glass tube with a piston, that moves due to the displacement of the gas whose flow rate is being measured. The piston, which is slightly smaller than the bore of the tube, has a groove filled with mercury to create a leak-less, low friction seal. The flow is determined by the time that it takes for the piston to travel between two reference positions, timed using light beams. This device has an accuracy of better than ±0.2%. NIST provides calibration services for flowmeters using this technique and apparatus.

Soap Bubble Technique. From time to time, researchers need to measure a particular flow, but do not have the specialized calibration equipment required. In cases like this, the soap bubble technique can be used. It is implemented by placing a soap film across the open end of a constant, cross-section tube. The other end of the tube is connected to a barb fitting that will easily receive a plastic tubing through which the flow to be measured is passing. Next, the soap film is moved toward the other end (the side with the fitting), by using a source of vacuum that is momentarily connected to the end with the fitting. Then, the tubing with the flow to be measured is connected to fitting, and the time that it takes for the soap film to travel between two reference positions is recorded. This information is then used to calculate flow. This technique resembles the piston prover, except that a soap film is used instead of a piston.

PNEUMOTACHOMETER DYNAMIC CHARACTERISTICS

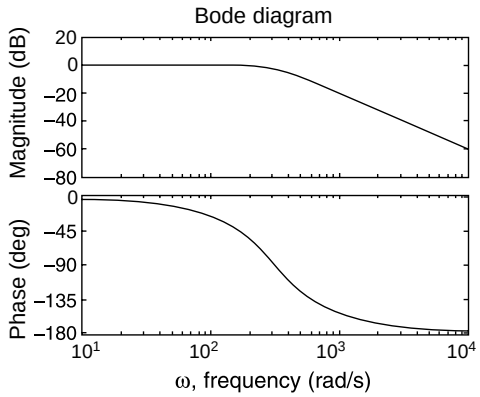
Another important characteristic of a pneumotachometer is its ability to measure rapidly changing flows. If one



(a) General schematic

test no	f_i	B_i	B_i / A	20 log(B_i / A)	ϕ_i
1					
2					
3					
⋮					
n					

(b) Experimental data



(c) Magnitude and phase Bode plot

Figure 13. Block diagram of procedure used for obtaining a frequency response test.

attempts to measure a fast changing flow signal with a flowmeter that does not have the adequate dynamic characteristics or "frequency response", one would obtain inaccurate results. One of the most common vehicles for capturing a device's frequency response is called the Bode Plot (40). This section addresses the general test procedure and presentation format used for obtaining a frequency response or Bode plot. Afterward, it presents the method used for performing a frequency response test on flowmeters.

One of the fundamental properties of a linear system is that if the system is excited with a sinusoidal input signal, after the transients die out, its output will also exhibit a sinusoidal behavior. A frequency response test consists of applying sinusoids of different frequencies and measuring the amplitude of the output and the phase difference between the input and output. Figure 13a shows a sinusoidal generator producing a sine wave with amplitude A and frequency f_i, being applied to a linear system. It also shows the corresponding output, another sinusoid of the same frequency, but with amplitude B_i and phase Φ_i. This procedure is repeated for different frequencies in order to obtain a table as shown Fig. 13b. Notice that the ratio of the output–input amplitude is calculated as well as 20 log of this ratio. This latter quantity has units called dB or deciBels in honor of the communication pioneer, Alexander Graham Bell. The Bode plot, which is actually made up of two plots, presents this information in graphical form. The magnitude plot shows the amplitude ratio in dB (i.e., 20 log B_i/A) on the vertical axis and the frequency [$\omega = 2\pi f$], in a logarithmic scale, on the horizontal axis. This is shown in Fig. 13c. Similarly, the Phase Bode Plot shows the phase against the various input frequencies. This is shown in Fig. 13d. The particular plot shown here is descriptive of a

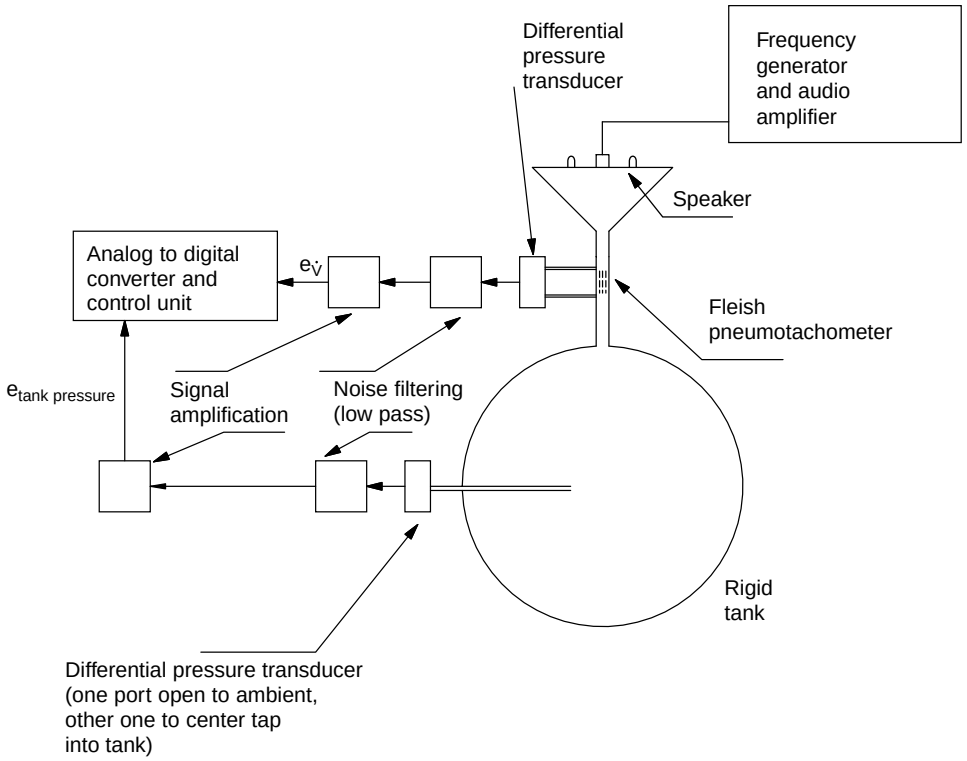


Figure 14. Test setup used for measuring frequency response of a pneumotachometer.

device that has low pass characteristics. This means that it passes signals of low frequencies at a constant gain (flat or horizontal portion), but attenuates ones that are of high frequencies (sloped portion of the graph). For any particular pneumotachometer application, it is desirable to use one that is flat in the range of frequencies that will be present in flow waveform that is to be measured. As an example of this, the ATS Standardization of Spirometry (18) stipulates, "Measuring PEF requires an instrument that has a frequency response that is flat (±5%) up to 12 Hz".

The reader may question the emphasis on sinusoidal inputs when in the real world, very few signals are sinusoidal. Representation in terms of sinusoids provides a general framework for dealing with any type of periodic signals, periodic function can be decomposed into a summation of sinusoids (Fourier Series).

Performing a frequency response test on a pneumotachometer is more challenging than for an electronic device where both the input and output are electronic signals that are easily measurable. Development of this experimental procedure for pneumotachometers has been refined by Jackson and Vinegar (41). The test set up can be represented as shown in Fig. 14. For purposes of this explanation, assume that the pneumotachometer being tested is of the Fleish type, a fluid resistance with a pressure transducer for measurement of the pressure drop. To perform the frequency response test, a small loudspeaker is used to generate the sinusoidal flow that is applied to the pneumotachometer under test. On the other side of the pneumotachometer, a large, rigid tank is connected. Notice that because the tank is filled with ambient air, a compressible fluid, flow enters and exits it. Of course, when the flow enters the tank, the pressure in it increases, and when the flow exits, the pressure decreases. The experimental procedure requires a second pressure transducer that measures the pressure inside the tank. This signal is used to calculate the flow in and out of the tank. Referring to Fig. 14, $e_{\dot{\mathbf{V}}\,\text{measured}}$ is the pneumotachometer output signal (the output voltage of the pressure transducer connected to the pneumotachometer) and $e_{\text{tank pressure}}$ is the tank's pressure transducer output signal. The flow into the tank is given by

$$\dot{V}_{\text{true}} = K_1 \frac{d(e_{\text{tank pressure}})}{dt}$$

A frequency response test is performed by varying the frequency of the loudspeaker, collecting the data and plotting it in Bode format. The response is flat as long as the ratio of $e_{\dot{V}\text{measured}}/\dot{V}_{\text{true}}$ is constant.

CORRECTION FOR STANDARD CONDITIONS

Often the state in which the flow is measured is not reflective of the physiological conditions in which flow becomes meaningful. In other cases the calibration of the flowmeter takes place under a different set of conditions than what the flowmeter will experience in the clinical setting. In both cases, one needs to correct the resulting measurements to accurately capture the true physiological event happening. To illustrate the correction process, an example is provided here. Consider a pneumotachometer that is to be used for measurement of inspiratory flow. Further assume that the pneumotachometer has been calibrated with room air, using the standard 3L syringe. Consequently, there is no need to correct the flowmeter reading since it will measure properly the flow that passes through it during the test. That is, the conditions for calibration and clinical testing are the same. However,

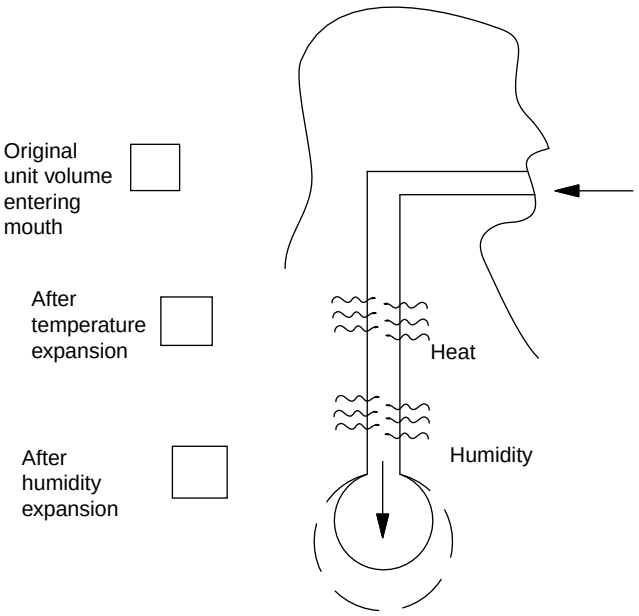


Figure 15. Correction for volume expansion due to increase in temperature and humidity.

when the air enters the respiratory system, it experiences two changes: its temperature increases from room temperature to body temperature, and the humidity from 0% (assuming a very dry day) to 100% RH. Both of these processes need to be taken into consideration if it is desired to know the actual volume that is expanding the alveoli. In reality, both processes are happening simultaneously, but for purposes of this example, we will assume that the air is heated first and then it becomes saturated. Both of these transformations cause the air volume to expand. Figure 15 shows the expansion process.

The increase in temperature causes the air to expand (Dalton's law of gasses). The correction factor for temperature is

$$\mathrm{CF}_{\mathrm{Temperature}} = \frac{T_{\mathrm{body}} + 273.16\,^{\circ}\mathrm{C}}{T_{\mathrm{inlet}} + 273.16\,^{\circ}\mathrm{C}}$$

Assuming that the inlet temperature is 21.1 °C (70 °F) and that the body temperature 37 °C (98.6 °F) this correction factor $\mathrm{CF}_{\mathrm{Temperature}}$ is 1.05.

The increase in humidity also creates an expansion. It is due to the effect of water vapor on the partial pressures. The total pressure of the dry gas is equal to the sum of the partial of its three main constituents N_2, O_2, and CO_2.

$$P_{\mathrm{total}} = P_{\mathrm{N_2}} + P_{\mathrm{O_2}} + P_{\mathrm{CO_2}}$$

However, as the air travels through the airways, it is humidified. We assume that the total pressure remains constant as the air moves through the airways. Now the total pressure becomes:

$$P_{\mathrm{total}} = P_{\mathrm{H_2O}} + P_{\mathrm{N_2}} + P_{\mathrm{O_2}} + P_{\mathrm{CO_2}}$$

The presence of the partial pressure of the water vapor ($P_{\mathrm{H_2O}}$) causes the sum of partial pressures of the last three constituents to decrease. This in turn causes the volume to expand. Thus the correction factor expression for humidity is

$$\mathrm{CF}_{\mathrm{HUMIDITY}} = \frac{p_{\mathrm{ambient}}}{p_{\mathrm{ambient}} - p_{\mathrm{sat}}}$$

Assume that the ambient barometric pressure is 29.92 mmHg (98.3 kPa) and at a body temperature of 36 °C, the water vapor pressure is 1.78 mmHg (6 kPa). Substituting these values into the previous equation produces a correction factor $\mathrm{CF}_{\mathrm{HUMIDITY}}$ of 1.06.

Combining both factors, the total correction factor becomes 1.12. This means that every milliliter that enters the mouth expands by 12% by the time that it reaches the alveoli. Consequently, measured flows need to be increased by 12%.

Comments about Complexities in the Measurement and Correction Process

More accurate (than the upper limits established by the ATS), bidirectional flow measurement is a challenging task when using pneumothachometers other than the volumetric type. This is true since the air volume varies due to density changes (resulting from temperature and composition differences of the incoming and outgoing air). The composition of inspired and expired air is shown in Table 1. The problem is made even more complicated if Fleish-type pneumotachometers are used, since now viscosity (the critical variable in the flow–pressure drop relationship) is influenced by both temperature and gas composition.

OVERVIEW OF DESIGN SPECIFICATIONS

This section presents some of the technical specifications of Fleish-type pneumotachometers, specifically those manufactured by Hans Rudolph a major manufacturer of respiratory components (42). These characteristics are shown in Table 2.

It is difficult to describe accuracy of each device in detail, since it depends not only on the hardware (i.e., resistive fluid sensor element) and the measurement process, but also on the correction factors that are applied (gas composition,

Table 1. Normal Gas Concentrations of Air[a]

Source of sample	Nitrogen (N_2),%	Oxygen (O_2),%	Carbon Dioxide (CO_2),%	Water Vapor (H_2O),%
Inspired air (dry)	78.65	20.9	0.04	0.5
Alveolar air (saturated)	75.45	13.2	5.2	6.2
Expired air (saturated)	74.8	15.3	3.7	6.2

[a]Adapted from Ref. (42).

Table 2. Characteristics of Commercially Available Pneumotachs[a,b]

Application	Flow Range ($L \cdot min^{-1}$)	Fluid Resistance ($mmH_2O/(L \cdot min^{-1})$)	Max. Pressure Drop at End of Range (mmH_2O)
Premature (38 week gestational)	0–10	1.0	10
Neonate (Birth to 1 month)	0–10	0.70	7
Infant (1–12 month)	0–35	0.20	7
Pediatrics	0–100	0.10	10
Pediatrics	0–160	0.10	16
Adults	0–400	0.04	16
Adults	0–800	0.02	16

[a]adapted from Ref. 43.
[b]Note: 1 psi = 1 lbf/in^2 = 6.89 kPa = 704 mmH_2O.

temperature, relative humidity, differences between atmospheric conditions at the time of calibration and those at the time of use, and corrections for quadratic-type pressure drop). As mentioned before, most manufacturers simple indicate that their products meet the ATSs standards.

INDIRECT TECHNIQUES FOR MEASURING FLOW

There are various methods that provide an estimate of respiratory flow, by means of an indirect measurement. These techniques find an application in cases where continuous connection to a pneumotach through a mouthpiece or mark is not feasible.

INDUCTIVE PLETHYSMOGRAPHY

Respiratory inductive plethysmography (RIP) has been used for many years for respiratory monitoring. Used in intensive care units worldwide for monitoring respiratory activity, primarily tidal volume. During inspiration, due to the bucket-handle effect of the ribs and the outward displacement of the abdomen, there is an increase in the cross-sectional area of the rib cage and abdomen, which translates into an increase in circumference (or more accurately, perimeter). This increase in perimeter is measured using elastic bands and correlated to a specific lung volume increase, or often used as a simple, relative measurement of lung volume expansion.

This technique used by the LifeShirt (44) is a vest-like, portable physiological monitoring system. In the LifeShirt, "two parallel, sinusoidal arrays of insulated wires embedded in elastic bands are woven into a flexible garmet. Extremely low voltage electrical current is passed through the wire creating an oscillating circuit. As the body chambers expand and contract, the electrical sensors generate different magnetic fields that are converted into proportional voltage changes over time (i.e., waveforms)". For calibration, which is done immediately after putting on the vest, the user breathes into a fixed-volume, calibration bag. Then the associated tidal volume is correlated to the measured body's expansion.

Respiratory Sounds

There is a correlation between respiratory flow and breath sounds. This method has been pursued for many years and continues to be an ongoing topic for research (45,46). Tracheal breath sounds are preferred by many investigators since they are louder than chest sounds and since there is more correlation of flow at the trachea and less filtering from the chest wall (47).

The U.S. Army (48) has been extremely active in developing physiological monitoring systems based on the sounds detected at the neck. Their motivation for this research is based on the premise that continuous monitoring of the soldier's health "can provide exceptional improvement to survivability, mobility, and lethality" (49). This group has generated a considerable amount of methodology as well as easy-to-build sensors and equipment for this application (50,51).

Some development has been done to obtain an estimate of actual flow from respiratory sounds (instantaneous measurement of the magnitude and direction of respiratory flow) (52). This is possible since inspiration and expiration have different "sounds", due to the asymmetrical nature of the respiratory passageways. This approach is implemented by initially, simultaneous measuring both tracheal sounds (using a small microphone attached to the neck) and actual flow (using a standard pneumotachometer). After recording data for 1–2 min, these two waveforms that are fed into a computer program that produces a correlation algorithm, which can be used subsequently to predict flow (magnitude and direction) from the tracheal sounds. The long-terms aim of this effort is detection of SIDS candidacy. The basic concept is that there might be a flow patterns/signature that could be construe as an early-predictor of SIDS. This test would be done during the first night that a newborn spends at the hospital. If a suspicious pattern is found, then a baby monitor could be sent home with the baby. It could also find application in the analysis the breathing patterns of athletes (in which a mouthpiece-tachometer is not feasible).

SUMMARY

Pneumotachometers have contributed and will continue to contribute significantly to our understanding of the respiratory system. They are an essential tool in the fight against respiratory diseases. In addition new applications are being developed that aim at the prevention of disease.

BIBLIOGRAPHY

1. The Bible, Chapter 2, verse 7.
2. Rupple G. Manual of Pulmonary Function Testing. 2nd ed. St. Louis: C.V. Mosby Company; 1979.
3. Fishman AP. Assessment of Pulmonary Function. New York: McGraw-Hill; 1980.
4. Nunn JF. Applied respiratory Physiology. 4th ed.
5. Pimmel R, Fullon JM. Characterizing Respiratory Mechanics with Excitation Techniques. Ann Biomed Eng 1982;9:475–488.
6. Bakos JH. Estimation of the Total Respiratory Parameters in Paralyzed and Free-breathing Rabbits by the Technique of Forced Oscillation, M.S. dissentation, Pennsylvania State University; 1979.
7. Tsai MJ, et al. Respiratory parameter estimation using forced oscillatory impedance data. J Appl Physiol 1977; 43(2):322–330.
8. Goldman MD. Clinical applications of forced oscillation. Pulmonary Pharmacol Therapeut 2001;14:341–350.
9. Schmid-Schoenbein GW, Fung YC. Forced perturbation of respiratory system: (A) The traditional model. Ann Biomed Eng 1978;6:194–211.
10. Schmid-Schoenbein GW, Fung YC. Forced perturbation of respiratory system: (B) A continuum mechanics analysis. Ann Biomed Eng 1978;6:367–398.
11. Macia NF, Dorson WJ, Higgins WT Jr. Lung-diaphragm Model of the Respiratory System for Parameter Estimation Studies Presented at the 1997 International Mechanical Engineering Congress & Exposition, Dallas, TX, November 1997.
12. Macia NF. Noninvasive, Quick Obstruction Estimation Method for the Measurement of Parameters in the Nonlinear Lung Model, Ph.D. dissertation, Arizona State University, August 1988.
13. McPherson SP. Respiratory Therapy Equipment. 2nd ed. St. Louis: C.V. Mosby Co.; 1981.
14. Cherniack RM, Brown E. A simple method for measuring respiratory compliance: Normal values for males. J Appl Physiol 1965;20(1):
15. Merth IT, Quanjer PH. Respiratory system compliance assessed by the multiple occlusion and weighted spirometer method in non-intubated healthy newborns. Pediatr-Pulmonol 1990;8(4):273–279.
16. Crawford S. Automated System for Measurement of Total Respiratory Compliance, Applied Project Report, Arizona State University East, 2003 (paper also under preparation).
17. Fleish A. Der Pneumotachograph-ein Apparat zur Geshwindigkertsregistrierung der Atemluft Pflugers. Arch Gas Physiol 1925;209:713.
18. The American Thoracic Society, Standardization of Spirometry, 1994 Update. Standardization of Spirometry, 1994 Update. Am J Respir Crit Care Med 1995;152:1107–1136.
19. Osborne JJ. Monitoring respiratory function. Crit Care Med 1974;2:217.
20. Osborne JJ. A flowmeter for respiratory monitoring. Crit Care Med 1978;6:349.
21. Sukes MK, NcNicol MW, Campbell EJM. Respiratory Failure. 2nd ed. Oxford (UK): Blackwell Scientific Publications; 1976.
22. Wright Respiratoy, Harris Calorific; Cleveland, Ohio.
23. ndd Medical technologies, 17 Progress Ave., Chelmsford, (MA), www.nddmed.com.
24. Sulton FD, Nett LM,Petty TL. A new ventilation monitor for the intensive care unit. Care Resp 1974;19:196.
25. Petty TL. Intensive and Rehabilitative Respiratory Care. 2nd ed. Lea and Febiger: Philadelphia; 1974.
26. McShane JL, Geil FG. Measuring flow. Res/Devel February 1975; 30.
27. Frederick G. Application of bluff body vortex shedding and fluidic circuit techniques to control of volumetric flow, M.S.E. dissentation; Arizona State University; 1974.
28. Bourns Medical Systems, Riverside (CA).
29. Svedin N, Kälvesten E, Stemme G. A new edge-detected lift force flow sensor presented at The 10th International Conference on Solid-State Sensors and Actuators (TRANSDUCERS'99) in Sendai, Japan; June 7–10, 1999.
30. Svedin N. A lift force flow sensor designed for acceleration insensitivity. Sensors Actuators 1998;68(1–3): 263–268.
31. Svedin N. A new silicon gas-flow sensor based on lift force. IEEE/ASME J Microelectromech Syst 1998;7(3):303–308.
32. Behrens CW. What is fluidics. Appl Manuf July 1968.
33. Fluidics flown on. The Engineer 28 June 1990.
34. Drzewiecki TM, Macia NF. Fluidic Technology: Adding Control, Computation and Sensing Capability to Microfluidics., Smart Sensors, Actuators, and MEMS conference at SPIE's International Symposium on Microtechnologies for the New Millennium, Gran Canaria, Spain; May 2003.
35. Thurston J. Personal communications, Honeywell, Tempe, (AZ).
36. ADVANCE for Managers of Respiratory Care, www.ADVANCEforMRC.com.
37. American Association of Respiratory Care, www.aarc.org and http://buyersguide.aarc.org/.
38. Varlen Instruments Inc., 2777 Washington Blvd, Bellwood, IL 60104, (800) 648-3954.
39. Wright JD, Matingsly GE. NIST Calibration Services for Gas Flow Meters. NIST special Publication 250–49.
40. Macia NF, Thaler GJ. Modeling and Control of Dynamic Systems. Thomson Delmar Learning; 2004.
41. Jackson AC, Vinegar A. A technique for measuring frequency responses of pressure, volume and flow transducers. J Appl Physio Respirat Environ Excercise Physiol 1979;47(2): 462–467.
42. Martini FH. Fundamentals of Anatomy and Physiology. 4th ed. New York: Prentice Hall; 1998.
43. Hans Rudolph, Inc., Kansas City, Mo., www.rudolphkc.com.
44. Vivo Metrics, Inc. Venturo, CA, www.vivometrics.com.
45. Graviely N. Breath Sounds Methodology. Boca Raton (FL): CRC Press; 1995.
46. Soufflet G, et al. Interaction between tracheal sound and flow rate: A comparison of some flow evaluations for lung sounds. IEEE Trans Biomed Eng 1990;37(4):
47. Mussell MJ, Nakazono Y, Miyamoto Y. Effect of air flow and flow transducer on tracheal breath sounds. Med Biol Eng Comput November 1990.
48. www.arl.army.mil/sed/acoustics.
49. Scanlon MV. Acoustic Sensor Array Element Extracts Physiology During Movement (internal document, available at the above, Army web site)
50. Scanlon MV. Acoustics Sensor for Health Status Monitoring. Proceeding of IRIS Acoustic and Seismic Sensing. 1998. Vol. II, 205–222. (Also available at the above, Army web site.)
51. Bass JD, Scanlon MV, Mills TK, Morgan JJ. Getting Two Birds with one Phone: An acoustic sensor for both speech recognition and medical monitoring. presented in poster format at the 138th Meeting of the Acoustical Society of America. November, 1999. (Also available at the above, Army web site.)
52. Gudala SG. Estimation of Air Flow from Tracheal Breath Sounds, Master of Technology Applied Project, Arizona State University East; May 2003.

See also PULMONARY PHYSIOLOGY; RESPIRATORY MECHANICS AND GAS EXCHANGE; VENTILATORY MONITORING.

POLYMERASE CHAIN REACTION

MICHAEL L. METZKER
Baylor College of Medicine
Houston, Taxas

THOMAS C. CASKEY
Cogene Biotech Ventures
Houston, Taxas

INTRODUCTION

Few techniques rival the impact that the polymerase chain reaction (PCR) has made in the age of molecular biology. Cloning and deoxyribonucleic acid (DNA) sequencing are other such techniques that have become embedded into everyday life on the molecular biologist's bench. Over 60 books alone (not to mention the tens of thousands of research articles) have been devoted to the strategies, methods and applications of PCR for the identification, detection and diagnosis of genetic and infectious diseases. Rightfully so, the inventor of PCR, Kary B. Mullis, was awarded the Nobel Prize in Chemistry for his discovery of the technique in 1993. However, PCR has not been without controversy. In 1989, DuPont challenged the validity of the Cetus PCR patents in federal court and with the Office of Patents and Trade Marks, and by 1991 the Cetus patents were unanimously upheld and later sold to Hoffman La Roche for $300 million. More recently, in 1993, Promega has challenged the validity of the Hoffmann La Roche *Taq* DNA polymerase patent that is currently pending. In this article, we attempt to provide a comprehensive overview for the molecular biologist when applying PCR to his/her application of interest.

DNA POLYMERASE REACTION

The DNA replication is an inherent process for the generation and evolution of future progeny in all living organisms. At the heart of this process is the DNA polymerase that primarily synthesizes new strands of DNA in a $5'\rightarrow3'$ direction from a single-stranded template. Most native DNA polymerases, however, are polyfunctional and show $5'$-exonuclease and/or $3'$-exonuclease activities that are important for cellular DNA repair and proofreading functions. Numerous molecular biology applications have harnessed these activities, such as labelling DNA by nick translation and TaqMan assays (see below), and endrepair of sheared DNA fragments and improving DNA synthesis fidelities, respectively. The PCR is an elegant, but simple, technique for the *In vitro* amplification of target DNA utilizing DNA polymerase and two specific oligonucleotide or primer sequences flanking the region of interest. PCR is a cyclic process of double-strand separation of DNA by heat denaturation, specific hybridization or annealing of short oligonucleotide primers to singlestranded DNA, and synthesis by DNA polymerase (1,2). Each cycle doubles the region marked by the primer sequences. By sequential iteration of the process, PCR exponentially generates up to a billion of copies of the target within just a few hours (Fig. 1).

The specificity of PCR is highly dependent on the careful design of unique primers with respect to the genome under investigation and the nucleotide composition of the primer sequences. Theoretically, a 16-mer (4^{16}) is of sufficient length to represent all unique primer sequences from a completely random genome size of 3 billion base pairs. In the real world, however, all genomes are not random and contain varying degrees of repetitive elements. For the human genome, Alus, LINEs (long interspersed DNA elements) and low complexity repeats are frequently observed and should be avoided in primer design when possible. There are a few simple rules for designing primer sequences that work well in PCR. In practice, PCR primers should be between 18 and 25 nucleotides long, have roughly an equal number of the four nucleotides, and show a G + C composition of 50–60%. Commercially available oligonucleotide synthesizers that show phosphamidite coupling efficiencies $> 98\%$ mean that primers of this size can usually be used in PCR without purification. A variety of computer programs are available for selecting primer sequences from a target region. Many of these programs will reveal internal hairpin structures and self-annealing primer sequences, but manual inspection of the oligonucleotide is still necessary to maximize successful PCR amplifications.

The concentrations of the PCR cocktail ingredients are also important for product specificity, fidelity and yield. In addition to *Taq* DNA polymerase and primers, the PCR mixture contains the cofactor magnesium ion (Mg^{2+}), the four $2'$-deoxyribonucleoside-$5'$-triphosphates (dNTPs) and the buffer. In general, PCR reagent concentrations that are too high from standard conditions result in nonspecific products with high misincorporation errors, and those that are too low result in insufficient product. A typical 50 μL PCR cocktail that contains 0.4 $\mu mol \cdot L^{-1}$ of each primer, 200 $\mu mol \cdot L^{-1}$ of each dNTP, 1.5 $mmol \cdot L^{-1}$ $MgCl_2$, and 1.25 units *Taq* DNA polymerase in 10 $mmol \cdot L^{-1}$ tris-HCl, pH 8.3, 50 $mmol \cdot L^{-1}$ KCl buffer works well for most PCR applications. The optimal Mg^{2+} concentration, however, may need to be determined empirically for difficult target templates. The performance and fidelity of *Taq* DNA polymerase are sensitive to the free Mg^{2+} concentration (3), which ionically interacts with not only the dNTPs but also with the primers, the template DNA, ethylenediaminetetraacetic acid (EDTA), and other chelating agents. In most cases, the Mg^{2+} concentration will range from 1.0 to 4.0 $mmol \cdot L^{-1}$.

The number of cycles and the cycle temperature–length of time for template denaturation and primer annealing and extension are important parameters for high quality PCR results. The optimal number of cycles is dependent on the starting concentration or copy number of the target DNA and typically ranges from 25 to 35 cycles. Too many cycles will significantly increase the amount of nonspecific PCR products. For low copy number targets, such as the integrated provirus of *Human immunodeficiency virus type 1* (HIV-1) from human genomic DNA, two rounds of PCR are employed first using an outer primer pair set followed by an internal (nested) primer pair set flanking the region of interest to yield positive and specific PCR products. Brief, but effective denaturation conditions, that

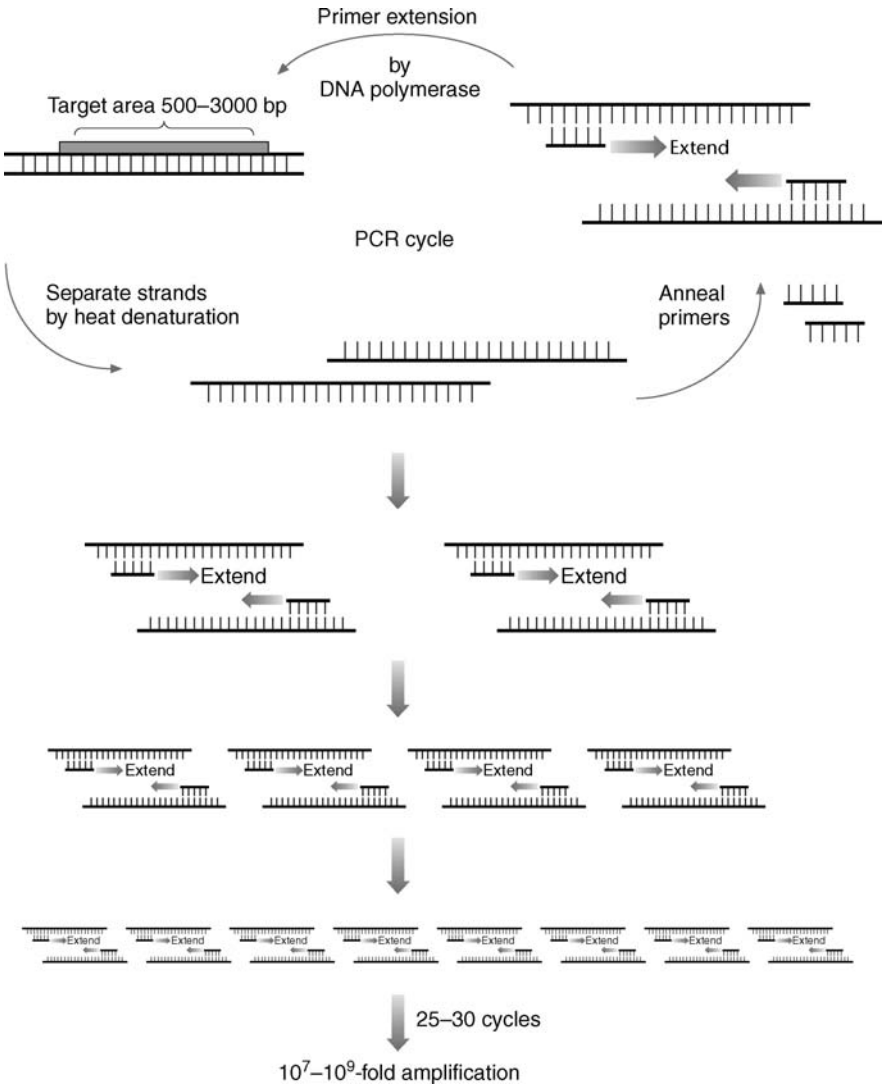


Figure 1. The PCR amplification cycle.

is 94–97 °C for 15–30 s, are necessary as *Taq* DNA polymerase has a half-life of only 40 min at 95 °C. Annealing conditions, on the other hand, are dependent on the concentration, base composition and the length of the oligonucleotide and typically range between 55 and 68 °C for 30–60 s. The length of the amplified target is directly proportional to the primer extension length of time. Primer extension is performed between 68 and 72 °C and, as a rule of thumb, is ∼ 60 s for every 1 kb.

Crude extracts from blood, cerebral spinal fluid, urine, buccal smears, bacterial colonies, yeast spores, and so on are routinely used as sources of DNA for PCR templates. Due to the high sensitivity of PCR, rapid isolation protocols, such as heat and detergent disruptions, and enzymatic digestion of biological samples have been frequently used. Caution should be invoked when using crude extracts as starting materials for PCR amplifications because a number of impurities are known to inhibit *Taq* DNA polymerase. These include red blood cell components, sodium dodecyl sulfate (SDS), high salts, EDTA and too much DNA. Since only a few hundred target molecules are needed for successful PCRs, in most cases, these impurities can be effectively removed by simply diluting the starting material. Each sample should then be tested with control primers that specifically amplify a known target to determine the integrity of the crude extract. Alternatively, the isolation of the desired organism, such as HIV-1, human *Hepatitis A virus*, influenza virus, cytomegalovirus, and so on, or the isolation of specific cell fractions, such as peripheral blood mononuclear cells, can significantly increase the sensitivity and specificity of the PCR amplifications.

SENSITIVITY AND CONTAMINATION OF POLYMERASE CHAIN REACTION

Contamination is the dark side of the PCR force. The exquisite sensitivity of PCR can result in contamination from even a single molecule of foreign or exogenous DNA (4,5). To minimize false positives, standard operating procedures have been described, including the physical isolation of PCR reagents from the preparation of DNA templates and PCR products, using autoclaved solutions, premixing and aliquoting reagents, the use of disposable

gloves, avoiding splashes, the use of positive displacement pipettes, adding DNA last, and carefully choosing positive and negative controls (6). Contamination is likely to surface for DNA samples that are difficult to amplify because of sequence content, or due to poor primer design and chemical impurities in DNA extractions. This is especially true for low copy number targets or degraded samples, as greater numbers of amplification cycles are generally required to achieve the desired product. In these cases, residual amounts of exogenous DNAs can compete and override the amplification process, resulting in spurious data. The best approach to challenge dubious results is to repeat the experiment with scrupulous care to details and controls. Biological samples collected at a single time point should be divided into multiple aliquots such that independent DNA extractions and PCR experiments can be performed to verify and validate initial results. Data should be discarded if inconsistent positive and negative PCR results occur upon repetition of the experiment. While negative controls can rule out reagent contamination, sporadic contamination can go unchecked. The probability of repeating spurious contamination in a consistent manner is extremely low.

There are three sources of contaminating DNA: (*1*) carryover contamination from previously amplified PCR products; (*2*) cross-contamination between multiple source materials; and (*3*) plasmid contamination from a recombinant clone that contains the target sequence. Of the three, carryover contamination is considered to be the major source of contamination because of the relative abundance of amplified target sequences. The substitution of dUTP for dTTP in the PCR cocktail has been routinely used as a method of preventing carryover contamination. Pretreatment of subsequent PCR mixtures prior to thermal cycling with uracil DNA glycosylase results in the removal of dU from any carryover PCR product, but does not affect the template DNA or dUTP. The dU removal creates an abasic site that is heat labile and degrades during thermal cycling, thus preventing carryover amplification. Moreover, ultraviolet (UV) light can reduce work surface and reagent contamination. Cross-contamination between samples is more difficult to diagnose, and suspicious results should be repeated from independent DNA extracts and PCR experiments for samples in question. Plasmid contamination, on the other hand, can be identified by sequence analysis and comparison to all laboratory plasmid sequences.

POLYMERASE CHAIN REACTION INTRODUCES MUTATIONS

The power and ease of PCR, however, were not fully appreciated until the introduction of the thermostable DNA polymerase isolated from *Thermus aquaticus* (*Taq*) (7) and automated instrumentation in 1988. It was here that PCR could be run in fully closed and automated systems. Fresh Klenow DNA polymerase did not have to be added at each cycle and PCR could be performed at higher annealing and extension temperatures, which increased the specificity and yields of the reactions while minimizing the risks of contamination. A hot start PCR further enhances specificity by preventing the formation of nonspecific products that arise during the initial steps of thermal cycling in PCR.

Taq DNA polymerase has been shown to incorporate nucleotides incorrectly at a frequency of 1 in 9000 bases by a mutation reversion assay (8). From sequence analysis of cloned PCR products, a slightly higher error frequency was determined (1 in 4000–5000 bp) for *Taq* DNA polymerase (9). The fidelity of DNA synthesis for *Taq* DNA polymerase, however, can vary significantly with changes in free Mg^{2+} concentration, changes in the pH of the buffer, or an imbalance in the four dNTP concentrations. Polymerase misincorporation errors are minimized when the four dNTPs are equimolar and between 50 and 200 $\mu mol \cdot L^{-1}$ (9). Since *Taq* DNA polymerase lacks a 3′-exonuclease activity, misincorporated bases typically cause chain termination of DNA synthesis that are not propagated in subsequent PCR cycles. In a worst-case scenario, a mutation occurring during the first round of PCR from a single target molecule and propagated thereafter would exist at a frequency of 25% in the final PCR product. Since hundreds of target copies are routinely used as starting DNA in PCR and most misincorporations terminate DNA synthesis, the observed error frequency is $\ll 25\%$.

Cloning of full-length genes from PCR products, however, has been problematic because PCR-induced mutations can cause amino acid substitutions in the wildtype sequence. Thus, significant effort must be employed in the complete sequencing of multiple PCR clones to identify mutation-free clones or ones that contain synonymous substitutions that do not change the protein coding sequence. Accordingly, thermostable DNA polymerases that contain a 3′-exonuclease (3′-exo) activity for proofreading of misincorporated bases have been recently introduced and include DNA polymerases isolated from *Pyrococcus furiosus (Pfu), Thermococcus litoralis* (Vent), *Pyrococcus* species GB-D (Deep Vent) and *Pyrococcus woesei* (*Pwo*). The error frequencies of these DNA polymerases are two- and sixfold < *Taq* DNA polymerase (10), but these polymerases are difficult for routine use, as the 3′-exonuclease activity can easily degrade the single-stranded PCR primers. 3′-Exo DNA polymerases, however, have been successfully used in long PCR in combination with *Taq* DNA polymerase and show an approximately twofold lower error frequency than *Taq* DNA polymerase alone (10).

POLYMERASE CHAIN REACTION LENGTH LIMITATIONS

For most applications, standard PCR conditions can reliably amplify target sizes up to 3–4 kb from a variety of source materials. Target sizes > 5 kb, however, have been described in the literature using standard PCR conditions, but generally yield low quantities of PCR product. The PCR size limitation can be attributed to the misincorporation of nucleotides that occurred 1 in 4000–5000 bp that ultimately reduced the efficiency of amplifying longer target regions. A breakthrough in long PCR came through the combined use of two thermostable DNA

polymerases, one of which contains a 3′-exonuclease activity (11,12). The principle for long PCR is that the *Taq* DNA polymerase performs the high fidelity DNA synthesis part of PCR, coupled with the proofreading activity of *Pfu*, Vent or *Pwo* DNA polymerases. Once the nucleotide error is corrected, *Taq* DNA polymerase can then complete the synthesis of long PCR templates. From empirical studies, only a trace amount of the 3′-exo DNA polymerase, roughly 1% to that of *Taq* DNA polymerase or another DNA polymerase isolated from *Thermus thermophilis (Tth)*, is needed to perform long PCRs > 20 kb. Other important factors for long PCR are the isolation of high quality, high molecular weight DNA and protection against template damage, such as depurination during thermal cycling. The use of the cosolvents glycerol and dimethyl sulfoxide (DMSO) have been shown to protect against DNA damage by efficiently lowering the denaturation temperature by several degrees centigrade. The rule of thumb for primer extensions still applies for long PCRs (60 s·kb^{-1}), although for targets > 20 kb, times extension should not exceed 22 min cycle^{-1}. The complexity and size of the genome under investigation can also affect the size of long PCR products. For example, PCR product lengths of 42 kb have been described for the amplification of λ bacteriophage DNA (11,12), compared with a 22 kb PCR product obtained from the human β-globin gene cluster (12).

CREATION OF NOVEL RECOMBINANT MOLECULES BY POLYMERASE CHAIN REACTION

Polymerase chain reaction can amplify both single- and double-stranded DNA templates as well as complementary DNA (cDNA) from the reverse transcription of messenger ribonucleic acid (mRNA) templates. Because of the flexibility of automated DNA synthesis, *In vitro* mutagenesis experiments can easily be performed by PCR. Recombinant PCR products can be created via the primer sequences by tolerated mismatches between the primer and the template DNA or by 5′-add-on sequences. Primer mediated mutagenesis can accommodate any nucleotide substitutions and small insertions or deletions of genetic material. The desired genetic alteration can be moved to any position within the target region by use of two overlapping PCR products with similar mutagenized ends (Fig. 2, left). This is accomplished by denaturing and reannealing the two overlapping PCR products to form heteroduplexes that have 3′-recessed ends. Following the extension of the 3′-recessed ends by *Taq* DNA polymerase, the full-length recombinant product is reamplified with the outer primers only to enrich selectively the full-length recombinant PCR product. 5′-Add-on adapters can also be used to join two unrelated DNA sequences, such as the splicing of an exogenous promoter sequence with a gene of interest (Fig. 2, right). The promoter–gene sequences are joined at the

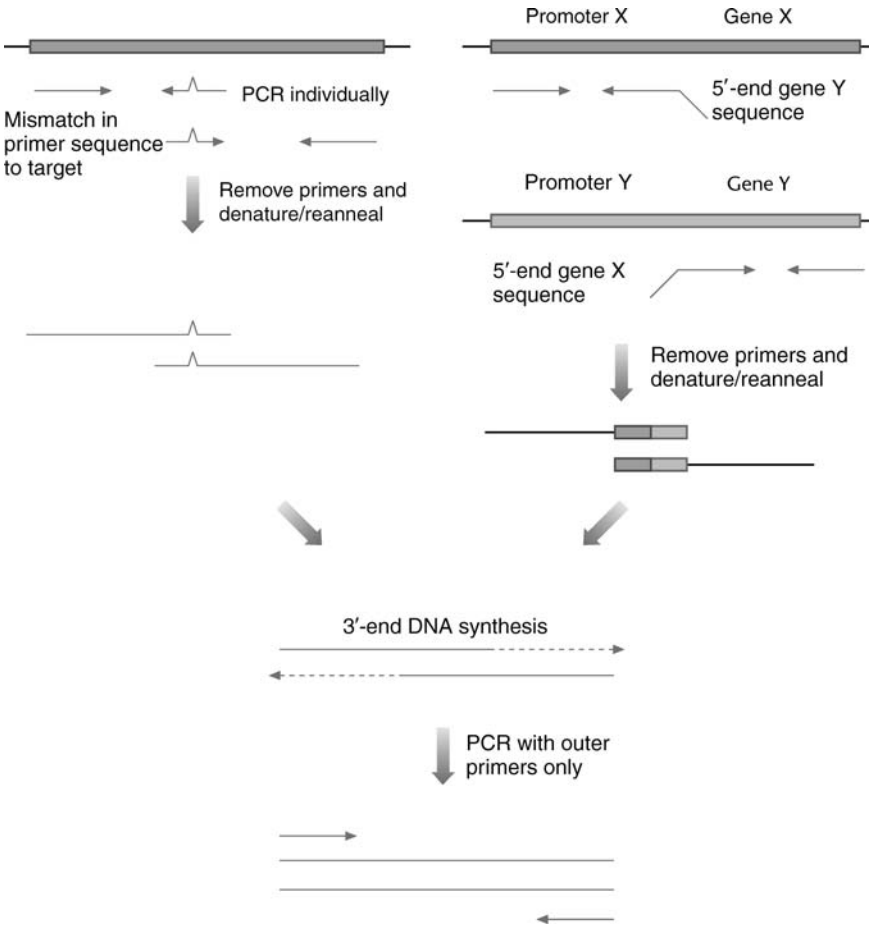


Figure 2. Creation of mutagenized or recombinant PCR products via primer mismatches (left) or 5′-add-on sequences (right).

desired junction by 5′-add-on gene specific and 5′-add-on promoter-specific adapters that can PCR amplify the promoter and the gene targets, respectively. Heteroduplexes can then be formed, as described above, from the two overlapping PCR products, which are then selectively amplified with outer primers to generate the desired full-length recombinant PCR product.

POLYMERASE CHAIN REACTION AS A DETECTION SYSTEM

Polymerase chain reaction is a powerful tool for the detection of human polymorphic variation that has been associated with hereditary diseases. Many PCR techniques have been described that can discriminate between wild-type and mutant alleles, but in this section only a few of the most frequently used techniques are discussed. Of these, DNA sequencing of PCR products is the most widely used and most sensitive method for the detection of both novel and known polymorphic differences between individuals. Complementing scanning technologies, however, have been developed for the rapid detection of allelic differences because of the high costs associated with DNA sequencing and the ability to process large numbers of samples. Singlestrand conformation polymorphism (SSCP) has been commonly used as a technique for the identification of genetic polymorphisms. Following PCR, the product is heat denatured and subjected to native or nondenaturing gel electrophoresis. Allelic differences between samples are detected as mobility band shifts by radioactive and nonradioactive labeling procedures. PCR–SSCP, however, is limited in fragment size to $\sim$ 200 bp because the accuracy in discriminating between different alleles diminishes significantly with an increase in the fragment length.

Multiplex PCR allows for the simultaneous amplification of multiple target regions and has been particularly useful for the detection of exon deletion(s) in X-linked disorders, such as Duchenne muscular dystrophy (13) and Lesch–Nyhan syndrome (14). The multiplex PCR products are resolved by gel electrophoresis and are visualized by ethidium bromide staining. The absence of specific PCR product(s) is diagnostic of exon deletion(s) in affected males, and half-dosage PCR products are diagnostic of carrier mothers (15). Moreover, up to 46 primer pairs have been simultaneously amplified by multiplex PCR with excellent success (90%) for the large-scale identification of human single nucleotide polymorphisms (SNPs) by hybridization to high density DNA chip arrays.

Genetic polymorphisms can also be identified by immobilizing the PCR product on to a nylon membrane in a dot blot format and probing by hybridization with an allele-specific oligonucleotide (ASO) that contains a 5′-biotin group. The ASO hybridization is detected by adding streptavidin-horseradish peroxidase, which binds to the biotinylated probe, followed by a colorimeter assay. The colorimeter ASO assay has been applied to the genotyping of human leucocyte antigen (HLA)–DQA alleles and the detection of β-thalassaemia mutations. More recently, multiplex PCR and colorimeter ASO methodologies have been combined in a reverse fashion, in which ASOs are immobilized on to nylon membrane strips and probed against biotinylated gene-specific multiplex PCR reactions. Allele-specific PCR products are detected by hybridization and conversion of a colourless substrate to a blue precipitate for the simultaneous genotyping of HLA–DQA 1, low density lipoprotein receptor, glycophorin A, hemoglobin G gammaglobin, D7S8, and groupspecific component.

Lastly, *In situ* PCR enables the amplification of target sequences from sections of formalin-fixed, paraffin-embedded tissue specimens to determine the levels of gene expression in specific cell types that otherwise could not be detected by conventional *In situ* hybridization. The PCR is performed directly on glass slides by overlaying the PCR mixture on to the specimen, sealing the slides to prevent evaporation, and temperature cycling using a thermal sensor or modified thermal cycler that holds glass slides.

DEGENERATE POLYMERASE CHAIN REACTION

Degenerate PCR is a powerful strategy for obtaining novel full-length cDNA sequences from limited amino acid sequence information (16). The PCR primer sequences are derived from the reverse translation of 6–9 amino acid codons, which will result in varying levels of degeneracy except for methionine and tryptophan residues. Careful attention should be exercised in the design of degenerate primers because increasing the primer complexity (i.e., using codons that show more than twofold degeneracy) will typically result in an increase in nonspecific PCR products. One approach in reducing the complexity of the degenerate primer is the use of codon bias for the particular organism from which the gene will be cloned. Alternatively, the alignment of orthologous gene sequences from other species can greatly improve the specificity of cloning the gene of interest by revealing evolutionarily conserved domains. Once the optimal primer sequence is determined, the mixture of oligonucleotides can be simultaneously synthesized and will represent all possible amino acid combinations of the degenerate sequence. The specificity of PCR should then selectively amplify the correct primer sequences to generate a gene or gene family-specific probe from which the full-length cDNA can be obtained. Degenerate PCR has been successfully used in the screening of novel gene family members such as G-protein-coupled receptors, nuclear steroid receptors and protein tyrosine kinases.

ANCIENT DNA

Phylogenetics is the study of evolutionary relationships between specimens that are inferred from contemporaneous sequences. The ability to obtain DNA sequences from specimens or even fossils that are millions of years old could equip the phylogeneticist with a powerful means of directly testing an a priori hypothesis. Following death of the tissue or organism, however, DNA is rapidly degraded by, presumably, nuclease activities and hydrolytic processes, resulting in short fragment sizes that are generally no longer than 100–150 bp. Moreover, this old DNA is largely modified by oxidative processes and by intermolecular crosslinks that render it unsuitable for cloning by standard molecular biology procedures. Short PCRs,

however, have been successfully performed from DNA samples isolated from archival and ancient specimens (17).

Museums hold vast collections of archived hospital files of patient specimens and of different species that have been collected over the last century. In a recent study, phylogenetic analyses of DNA sequences were performed from reverse transcriptase PCR (RT-PCR) of formalin-fixed, paraffin-embedded tissue specimens obtained from U.S. servicemen killed in the 1918 Spanish influenza pandemic. Viral sequences from three different gene regions were consistent with a novel H1N1 *Influenza A virus* that was most closely related to influenza strains that infect humans and swine, but not wild waterfowl, considered to be the natural reservoir for the influenza virus (18). Moreover, PCR and DNA sequencing have been performed on DNA extractions of archaeological findings, such as amplifying mitochondrial DNA sequences from a 7000 year old human brain, amplifying both mitochondrial and nuclear DNA sequences from bone specimens from a 14,000 year old saber-toothed cat, and amplifying chloroplast DNA sequences from fossil leaf samples from a 17 million-year-old Miocene *Magnolia* species.

QUANTITATIVE POLYMERASE CHAIN REACTION

Quantitative PCR (QPCR) has been widely used for detecting and diagnosing genetic deletions, for studying gene expression and for estimating the viral load of HIV-1. While DNA quantitation by multiplex PCR has been previously described (15), the quantitation of RNA has been wide reaching for the latter two areas. For many applications, estimating the relative amount of PCR product is sufficient to describe a biological observation. The absolute quantitation of RNA molecules, however, has been more difficult than for DNA because of the difficulty of generating accurate controls. Internal standards derived from synthetic RNA or cRNA have been designed to contain the same primer sequences as the target but yield a different-sized PCR product that can be easily separated by gel electrophoresis. cRNAs are not only coamplified with target sequences, but also coreverse transcribed to account for the variable efficiencies of cDNA syntheses. Moreover, QPCR is typically performed in the exponential or log phase of the amplification process (typically 14–22 cycles) to obtain accurate quantitative results. The absolute amount of target mRNA can be quantitated by serial dilutions of the target/internal control mixture and by extrapolating against the standard curve.

Both the variable range of initial target amounts and the presence of various inhibitors can, however, adversely affect the kinetics and efficiencies of PCR. Alternatively, a strategy based on a quantitative competitive (QC) approach has been used to minimize the effects of these variables. Known quantities of the competitor template, which contains the same primer sequences as the target but differs in size, are introduced into replicate PCRs containing identical quantities of the target. The point at which the intensities of the PCR products derived from the target sequence and the competitor template are equivalent is used to estimate the amount of target sequence in the original sample (19).

Recently, real-time QPCR and QCPCR (20) using a 5′-nuclease fluorogenic or TaqMan assay (21) has been developed to measure accurately the starting amounts of target sequences. Unlike gel electrophoresis, real-time QPCR has the unique advantage of being a closed-tube system, which can significantly reduce carryover contamination. Using this technique, one can easily monitor and quantitate the accumulation of PCR products during log phase amplification. The TaqMan assay utilizes dual reporter and quencher fluorescent dyes that are attached to a nonextendible probe sequence. During the extension phase of PCR, the 5′-nuclease activity of *Taq* DNA polymerase cleaves the hybridized fluorogenic probe, which releases the reporter signal and is measured during each cycle. In addition to real-time QPCR, TaqMan assays have broad utility for the identification of SNPs.

RELATED NUCLEIC ACID AMPLIFICATION PROCEDURES

Other *in vitro* systems can amplify nucleic acid targets such as the transcription-based amplification system (TAS) (6), its more recent version called the self-sustained sequence replication (3SR) (22) and the ligation-dependent Qβ-replication assay (23). These methods are best suited for the detection and semiquantitation of RNA target sequences. The strategy for TAS and 3SR is a continuous series of reverse transcription and transcription reactions that mimic retroviral replication by amplifying specific RNA sequences via cDNA intermediates. The primers contain 5′- add-on sequences for T7, T3, or SP6 promoters that are incorporated into the cDNA intermediates. The rapid kinetics of transcription-based amplifications is an attractive feature of these systems, which can amplify up to 10^7 molecules in 60 min. Short amplify products, however, which are due to incomplete transcription of the target region and incomplete RNase H digestion of the RNA–DNA hybrids, can be problematic in the TAS and 3SR assays.

Unlike PCR, TAS, or 3SR assays, the ligation-dependent Qβ-replication assay results in the amplification of probe, not target, sequences. This assay utilizes a single hybridization to the target sequence, which is embedded within, and divided between, a pair of adjacently positioned midvariant (MDV-1) RNA probes. MDV-1 RNA is the naturally occurring template for the bacteriophage Qβ RNA replicase. Following the isolation of the probe–target hybrids, ligation of the binary probes creates a full-length amplifiable MDV-1 RNA reporter. When Qβ replicase is added, newly synthesized MDV-1 RNA molecules are amplified from ligated binary probes that originally hybridized to the target sequence (23). Similar to TAS and 3SR, the Qβ-replication assay shows rapid kinetics, generating up to 10^9 molecules in 30 min, and all three methods have been successfully used for the detection and quantitation of HIV-1 RNA molecules.

LIGATION CHAIN REACTION

The ligase chain reaction (LCR) can also amplify short DNA regions of interest by iterative cycles of denaturation

and annealing/ligation steps (24). The LCR utilizes four primers: two adjacent ones that specifically hybridize to one strand of target DNA and a complementary set of adjacent primers that hybridize to the opposite strand. LCR primers must contain a 5′-end phosphate group, such that thermostable ligase (24) can join the 3′-end hydroxyl group of the upstream primer to the 5′-end phosphate group of the downstream primer. Successful ligations of adjacent primers can subsequently act as the LCR template, resulting in an exponential amplification of the target region. The LCR is well suited for the detection of SNPs because a single-nucleotide mismatch at the 3′ end of the upstream primer will not ligate and amplify, thus discriminating it from the correct base. Although LCR is generally not quantitative, linear amplifications using one set of adjacent primers, called the ligase detection reaction, can be quantitative. Coupled to PCR, linear ligation assays can also be used as a mutation detection system for the identification of SNPs using both wild-type-specific and mutant-specific primers in separate reactions. The oligonucleotide ligase assay was first reported to detect SNPs from both cloned and clinical materials using a 5′-end biotin group attached to the upstream primer and a nonisotopic label attached to the downstream primer (25). Allele-specific hybridizations and ligations can be separated by immobilization to a streptavidin-coated solid support and directly imaged under appropriate conditions without the need for gel electrophoretic analysis.

SUMMARY

Some of the general concepts and practices of PCR have been reviewed here. Not only has PCR made a major and significant impact on basic and clinical research, but it has also been well accepted and utilized in forensic science. For any scientific methodology to be accepted in the courts as evidence, it must satisfy four criteria: that the method (*1*) be subject to empirical testing, (*2*) be subject to peer review and publication, (*3*) has a known error rate, and (*4*) is generally accepted in the scientific community. The application of PCR has been admitted in the U.S. courts as evidence in criminal cases for the analysis of human DNA sequences, and in January 1997 as evidence for the phylogenetic analysis of HIV DNA sequences (26). Clearly, the scope of applications for PCR seems endless and it is truly a remarkable technique that has been widely used in molecular biology.

BIBLIOGRAPHY

1. Saiki RK, et al. Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anaemia. Science 1985;230:1350–1354.
2. Mullis KB, Faloona FA. Specific synthesis of DNA *In vitro* via a polymerase-catalysed chain reaction. Methods Enzymol 1987;155:335–351.
3. Eckert KA, Kunkel TA. High fidelity DNA synthesis by the *Thermus aquaticus* DNA polymerase. Nucleic Acids Res 1990;18:3739–3744.
4. Gibbs RA, Chamberlain JS. The polymerase chain reaction: a meeting report. Genes Dev 1989;3:1095–1098.
5. Kwoh DY, et al. Transcription-based amplification system and detection of amplified human immunodeficiency virus type 1 with a bead-based sandwich hybridization format. Proc Nat Acad Sci U.S.A. 1989;86:1173–1177.
6. Kwok S, Higuchi R. Avoiding false positives with PCR. Nature (London) 1989;339:237–238.
7. Saiki RK, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 1988;239:487–491.
8. Tindall KR, Kunkel TA. Fidelity of DNA synthesis by the *Thermus aquaticus* DNA polymerase. Biochemistry 1988;27:6008–6013.
9. Innis MA, Myambo KB, Gelfand DH, Brow MAD. DNA sequencing with *Thermus aquaticus* DNA polymerase and direct sequencing of polymerase chain reaction-amplified. Proc Nat Acad Sci U.S.A. 1988;85:9436–9440.
10. Cline J, Braman JC, Hogrefe HH. PCR fidelity of *Pfu* DNA polymerase and other thermostable DNA polymerases. Nucleic Acids Res 1996;24:3546–3551.
11. Barnes WM. PCR amplification of up to 35-kb DNA with high fidelity and high yield from λ bacteriophage templates. Proc Nat Acad Sci U.S.A. 1994;91:2216–2220.
12. Cheng S, Fockler C, Barnes WM, Higuchi R. Effective amplification of long targets from cloned inserts and human genomic DNA. Proc Nat Acad Sci U.S.A. 1994;91:5695–5699.
13. Chamberlain JS, et al. Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. Nucleic Acids Res 1988;23:11141–11156.
14. Gibbs RA et al. Multiplex DNA deletion detection and exon sequencing of the hypoxanthine phosphoribosyltransferase gene in Lesch–Nyhan families. Genomics 1990;7:235–244.
15. Metzker ML, Allain KM, Gibbs RA. Accurate determination of DNA in agarose gels using the novel algorithm GelScann(1.0). Computer App Biosci 1995;11:187–195.
16. Lee CC, et al. Generation of cDNA probes directed by amino acid sequence: cloning of urate oxidase. Science 1988;239:1288–1291.
17. Pääbo S. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. Proc Nat Acad Sci U.S.A. 1989;86:1939–1943.
18. Taubenberger JK, et al. Initial genetic characterization of the 1918 'Spanish' influenza virus. Science 1997; 275:1793–1796.
19. Gilliland G, Perrin S, Blanchard K, Bunn HF. Analysis of cytokine mRNA and DNA: detection and quantitation by competitive polymerase chain reaction. Proc Nat Acad Sci U.S.A. 1990;87:2725–2729.
20. Heid CA, Stevens J, Livak KJ, Williams PM. Real time quantitative PCR. Genome Res 1996;6:986–994.
21. Holland PM, Abramson RD, Watson R, Gelfand DH. Detection of specific polymerase chain reaction product by utilizing the 5′→3′ exonuclease activity of *Thermus aquaticus* DNA polymerase. Proc Nat Acad Sci U.S.A. 1991;88:7276–7280.
22. Guatelli JC, et al. Isothermal, *In vitro* amplification of nucleic acids by a multienzyme reaction modeled after retroviral replication. Proc Nat Acad Sci U.S.A. 1990;87:1874–1878.
23. Tyagi S, Landegren U, Tazi M, Lizardi PM, Kramer FR. Extremely sensitive, background-free gene detection using binary probes and Qβ-replicase. Proc Nat Acad Sci U.S.A. 1996;93:5395–5400.
24. Barany F. Genetic disease detection and DNA amplification using cloned thermostable ligase. Proc Nat Acad Sci U.S.A. 1991;88:189–193.
25. Landegren U, Kaiser R, Sanders J, Hood L. A ligase-mediated gene detection technique. Science 1988;241:1077–1080.
26. State of Louisiana v. Richard J Schmidt. Reasons for ruling of Louisiana State 15th Judicial District Court Judge Durwood

Conque. 15th Judicial District Court, Lafayette Parish, Louisiana. Criminal Docket 73313. 1997.

Further Reading

Erlich HA, Gelfand D, Sninsky JJ. Recent advances in the polymerase chain reaction. Science 1991; 252: 1643–1650.

Innis MA, Gelfand DH, Sninsky JJ, White TJ, editors. PCR Protocols: A Guide to Methods and Applications. Academic; San Diego: 1990.

Mullis KB, Ferré F, Gibbs RA, editors. The Polymerase Chain Reaction. Birkhäuser; Boston: 1994.

See also ANALYTICAL METHODS, AUTOMATED; DNA SEQUENCE; MICROARRAYS.

POLYMERIC MATERIALS

XIAOHUA LIU
The University of Michigan
Ann Arbor, Michigan

SHOBANA SHANMUGASUNDARAM
TREENA LIVINGSTON ARINZEH
New Jersey Institute of Technology
Newark, New Jersey

INTRODUCTION

This article aims to provide basic and contemporary information on polymeric materials used in medical devices and instrumentation. The fundamental concepts and features of polymeric materials are introduced in the first section. In the second section, the major commodity polymers used in medicine are reviewed in terms of their basic chemical and physical properties. The main part of this article, however, is devoted to polymers in biomedical engineering applications, including tissue engineering and drug delivery systems.

Polymers are a very important class of materials. A polymer can be defined as a long-chain molecule that is composed of a large number of repeating units of identical structure. Some polymers, (e.g., proteins, cellulose, and starch) are found in Nature, while many others, including polyethylene, polystyrene, and polycarbonate, are produced only by synthetic routes. Hundreds of thousands of polymers have been synthesized since the birth of polymer science. Today, polymeric materials are used in nearly all areas of daily life.

Polymers can simply be divided into two distinct groups based on their thermal processing behavior: thermoplastics and thermosets. Thermoplastics are linear or branched polymers, and they soften or melt when heated, so that they can be molded and remolded by heating. This property allows for easy processing and recycling. In comparison, thermosets are three-dimensional (3D) network polymers, and cannot be remelted. Once these polymers are formed, reheating will cause the material to scorch.

In addition to classification based on processing characteristics, polymers may also be grouped based on the chemical structure of their backbone. Polymers with one identical repeating unit in their chains are called homopolymers. The term copolymer is often used to describe a polymer with two or more repeating units. The sequence of repeating units along the polymer chain can form different structures, and copolymers can be further classified as random copolymers, alternating copolymers, block copolymers, and graft copolymers. In random copolymers, the sequence distribution of the repeating units is random, while in alternating copolymers the repeating unit are arranged alternately along the polymer chain. A block copolymer is one in which identical repeating units are clustered in blocks along the chain. In graft copolymers, the blocks of one type of repeating unit are attached as side chains to the backbone chains.

Unlike simple pure compounds, most polymers are not composed of identical molecules. A typical synthetic polymer sample contains chains with a wide distribution of chain lengths. Therefore, polymer molecular weights are usually given as averages. The number average molecular weight (M_n), which is calculated from the mole fraction distribution of different sized molecules in a sample, and the weight average molecular weight (M_w), which is calculated from the weight fraction distribution of different sized molecules, are two commonly used values. The statistical nature of polymerization reaction makes it impossible to characterize a polymer by a single molecular weight. A measure of the breadth of the molecular weight distribution is given by the ratios of molecular weight averages. The most commonly used ratio is M_w/M_n. As the weight dispersion of molecules in a sample narrows, M_w approaches M_n, and in the unlikely case that all the polymer molecules have identical weights, the ratio M_w/M_n becomes unity. Most commercial polymers have the molecular weight distribution of 1.5–10. In general, increasing molecular weight corresponds to increasing physical properties and decreasing polymer processability.

In many cases, individual polymer chains are randomly coiled and interviewed with no molecular order or structure. Such a physical state is termed amorphous. Amorphous polymers exhibit two distinctly different types of mechanical behavior. Some, like poly(methyl methacrylate, PMMA) and polystyrene are hard, rigid, glassy plastics at room temperature, while others, like polybutadiene and poly(ethyl acrylate), are soft, flexible, rubbery materials at room temperature. There is a temperature, or range of temperatures, below which an amorphous polymer is in a glassy state, and above which it is rubbery. This temperature is called the glass transition temperature (T_g). The value of T_g for a specific polymer will depend on the structure of the polymer. Side groups attached to the polymer chain will generally hinder rotation in the polymer backbone, necessitating higher temperatures to give enough energy to enable rotation to occur.

For most polymers, the T_g constitutes their most important mechanical properties. At low temperatures ($< T_g$), an amorphous polymer is glass-like, with a value of Young's modulus in the range of 10^9–10^{10} Pa, and it will break or yield at strains greater than a few percent. When the temperature is $> T_g$, the polymer becomes rubber-like, with a modulus in the range of 10^5–10^6 Pa, and it may withstand large extensions with no permanent deformation. At even

higher temperatures, the polymer may undergo permanent deformation under load and behave like a highly viscous liquid. In the T_g range, the polymer is neither glassy nor rubber-like. It has an intermediate modulus and has viscoelastic properties.

CHEMICAL AND PHYSICAL PROPERTIES OF MAJOR COMMODITY POLYMERS

This section reviews the major polymers used in medical applications, with a brief discussion of chemical as well as physical properties and their application. They are grouped as homopolymers or copolymers.

Homopolymers

Polyacrylates (e.g., PMMA) and poly(hydryoxyethyl methacrylate) (PHEMA), are used for hard and soft contact lenses because of their excellent physical, coloring properties, and ease in fabrication. The PMMA polymer is a hydrophobic, linear chain polymer that is glassy at room temperature. It has very good light transmittance, toughness, and stability, making it an excellent material for intraocular lenses and hard contact lenses. The PHEMA polymer is used for soft contact lenses. With the addition of a $-CH_2OH$ group to the methyl methacrylate side group of the PMMA structure, the polymer becomes hydrophilic. Typically, PHEMA is cross-linked with elthylene glycol dimethylacrylate (EGDM) to prevent the polymer from dissolving when hydrated. When fully hydrated, PHEMA is a hydrogel with potential use in advanced technology applications (e.g., biomedical separation and biomedical devices).

Polyolefins, which include polyethylene (PE) and polypropylene (PP), are linear chain polymers. Polyethylene is a highly crystalline polymer that is used in its high density form for biomedical applications because low density forms cannot withstand sterilization temperatures. The high density form is used for drains and catheters. The ultrahigh molecular weight form (UHMWPE) is used in orthopedic implants for load-bearing surfaces in total hip and knee joints. The material has good toughness, creep properties, resistance to environmental attack, and relatively low cost. The PP is related to PE by the addition of a methyl group along the polymer chain. It has similar properties to PE (e.g., high rigidity, good chemical resistance, and tensile strength) and is used for many of the same applications. It also has a high flex life, which is superior to PE and is therefore used for finger joint prostheses.

Polytetrafluoroethylene (PTFE), commonly known as Teflon, is similar in structure to PE except that the hydrogen in PE is substituted with fluorine. This polymer has a high crystallinity ($> 94\%$), high density, low modulus of elasticity, and tensile strength. It is a very stable polymer and difficult to process. The material also has very low surface tension and friction coefficient. It is used for vascular graft applications due to the lack of adherence of blood components.

Poly(vinyl chloride) (PVC) is typically used for tubing for blood transfusions, feeding, and dialysis. Pure PVC is hard and brittle. However, for these applications, the addition of plasticizers makes it soft and flexible. Issues concerning these plasticizers exist because they can be extracted during long-term use, making PVC less flexible over time.

Poly(dimethyl siloxane) (PDMS), or silicone rubber, is a versatile material. Low molecular weight polymers have low viscosity and can be cross-linked to make a higher molecular weight rubber-like material. It has a silicon–oxygen backbone instead of a carbon backbone. The material is less temperature sensitive than other rubbers because of its lower T_g. It also has excellent flexibility and stability. The applications of PDMS are widespread (e.g., catheter and drainage tubing, insulation for pacemaker leads, a component in some vascular graft systems, prostheses for finger joints, blood vessels, breast implants, outer ears, chin and nose implants). Since its oxygen permeability is very high, PDMS is also used in membrane oxygenators.

Polyamides, commonly known as nylons, are linear-chain polymers containing –CONH– groups. With the presence of these groups, the chains attract strongly to one another by hydrogen bonding. Increasing numbers of –CONH– groups and a high degree of crystallinity improves physical properties (e.g., strength and fiber forming ability). They are used for surgical sutures.

Polycarbonates are tough, amorphous, clear materials produced by the polymerization of biphenol A and phosgene. It is used as lenses for eyeglasses and safety glasses, and housings for oxygenators and heart–lung bypass machines.

Copolymers

Poly(glycolide lactide) (PGL) are random copolymers used in resorbable surgical sutures. The PGL is polymerized by a ring-opening reaction of glycolide and lactide and is gradually resorbed in the body due to the ester linkages in the polymer backbone via hydrolysis.

A copolymer of tetrafluoroethylene and hexafluoropropylene (FEP) is used similarly to PTFE. The advantage of FEP is that it is easier to process than PTFE, but still retains excellent chemical inertness and a low coefficient of friction. The FEP has a crystalline melting temperature of 265 °C, whereas PTFE is 375 °C.

Polyurethanes are copolymers, which contain "hard" and "soft" blocks. The "hard" blocks are composed of a diisocyanate and a chain extender, with a T_g above room temperature, and has a glassy or semi-crystalline character. The "soft" blocks are typically polyether or polyester polyols with a T_g below room temperature. Thus, the material also has rubbery characteristics. Polyurethanes are tough elastomers with good fatigue and blood-containing properties. They are typically used for pacemaker lead insulation, vascular grafts, heart assist balloon pumps, and artificial heart bladders.

POLYMERS IN BIOMEDICAL ENGINEERING APPLICATIONS

Polymers Used in Tissue Engineering

Synthetic Polymers. The most widely used synthetic polymers for tissue engineering products, either under

development or on the market, are poly(lactic acid) (PLA), poly(glylic acid) (PGA), and their copolymers PLGA. Both PLA and PGA are linear aliphatic polyesters formed by ring-opening polymerization with a metal catalyst. The PLA can also be obtained from the renewable agricultural source, corn and degrades in two phases: hydrolysis and metabolization. The PLA and PGA polymers have similar chemical structures except that the PLA has a methyl pendant group. Both degrade by simple hydrolysis of their ester linkages. The PGA can also be broken down by nonspecific esterases and carboxypeptidases. The degradation rate is dependent on initial molecular weight, exposed surface area, crystallinity, and, in the case of copolymers, the PLA/PGA ratio present. PGA is highly crystalline, having a high melting point, and a low solubility in organic solvents. It is also hydrophilic in nature, losing its mechanical strength over a period of 2–4 weeks in the body.

The PLGA was developed to achieve a wider range of possible applications for PGA. Due to the extra methyl group in lactic acid, PLA is more hydrophobic and has a slower rate of backbone hydrolysis than PGA. The PLA is also more soluble in organic solvents. The copolymer PLGA degradation depends on the exact ratio of PLA and PGA present in the polymer. The PLGA polymer is less crystalline and tends to degrade more rapidly than either PGA or PLA. Lactic acid is a chiral molecule that exists in two stereoisomeric forms that yield four morphologically distinct polymers. Both *d*-PLA and *l*-PLA are two stereoregular polymers *d,l*-PLA is the racemic polymer and *meso*-PLA can be obtained from *d,l*-lactide.The amorphous polymer is *d,l*-PLA and is used typically for drug delivery applications where it is important to have a homogenous dispersion of the active agents within a monobasic matrix. The *l*-PLA polymer is semi-crystalline and most commonly used because the degradation product of *l*(+)-lactic acid is the naturally occurring steroisomer of lactic acid. It is typically used for high mechanical strength and toughness applications (e.g., orthopaedics).

Some of the other synthetic polymers currently under investigation for tissue engineering applications are described briefly. Polycaprolactone (PCL) is a synthetic aliphatic polyester with a melting point (T_m) of 55–65 °C. Degradation of PCL is a slow process that occurs either by hydrolysis or enzymatic degradation *in vivo*. The slow degradation rate of PCL is particularly interesting for long-term implants and controlled release application. Poly(hydroxy butyrate) (PHB) and its copolymers are semi-crystalline thermoplastic polyester made from renewable natural sources. *In vivo*, PHB degrades into hydroxybutyric acid that is a normal constituent of human blood. The PHB homopolymer is highly crystalline and has a high degradation rate. Its biodegradation and biocompatibility properties have led to research on its prospective use as a material for coronary stents, wound dressings, and drug delivery. Poly(propylene fumarate) (PPF) is an unsaturated linear polyester formed by the copolymerization of fumaric acid and propylene glycol. These polymer networks degrade by hydrolysis of the ester linkage to water-soluble products, namely, propylene glycol, poly(acrylic acid-*co*-fumaric acid), and fumaric acid. Due to its unsaturated sites along the polymer backbone, which are labile and can be cross-linked *in situ*, PPF is currently being evaluated for filling skeletal defects of varying shapes and sizes. Polyphosphoesters (PPE) are biodegradable polymers with physicochemical properties that can be altered by the manipulation of either the backbone or the side-chain structure. This property of PPE makes them potential drug delivery vehicles for low molecular drugs, proteins, deoxyribonucleic acid DNA plasmids, and as tissue engineering scaffolds. Since the phosphoester bond in a PPE backbone is cleaved by water, the more readily water penetrates, with greater bond cleavage and faster degradation rate. The products of hydrolytic breakdown of PPE are phosphate, alcohol, and diol. Polyphosphazenes are inorganic polymers having a phosphorus–nitrogen alternating backbone and each phosphorus atom is attached to two organic or organometallic side groups. They degrade by hydrolysis into phosphate, amino acid, and ammonia. The potential application is in low molecular weight drug release and in formulation of proteins and peptides. Polyanhydrides are a class of degradable polymers synthesized from photopolymerizable multimethacrylate monomers. Many polyanhydrides degrade from the surface by hydrolysis of the anhydride linkages. The rate of hydrolysis is controlled by the polymer backbone chemistry. They are useful for controlled drug delivery as they degrade uniformly into nontoxic metabolites. Polyorthoesters (POE), another class of biodegradable and biocompatible polymers, can be designed to possess a surface-dominant erosion mechanism. Acidic byproducts autocatalyzed the degradation process resulting in increased degradation rates than nonacidic byproducts. The POE, which is susceptible to acid-catalysed hydrolysis, has attracted considerable interest for the controlled delivery of therapeutic agents within biodegradable matrices.

Natural Polymers. Collagen is a widely used natural polymer in tissue engineering. It is a structural protein, being a significant constituent of the natural extracellular matrix. It has a triple-helical molecular structure that arises from the repetitious amino acid (glycine, proline, and hydroxyproline) sequence. *In vivo*, collagen in healthy tissues is resistant to attack by most proteases except specialized enzymes called collagenases that degrade the collagen molecules. Collagen can be used alone or in combination with other extracellular matrix components (e.g., glycosaminoglycan and growth factors) to improve cell attachment and proliferation. It has been tested as a carrier material in tissue engineering applications.

Other natural polymers under investigation for tissue engineering applications are described briefly. Gelatin, denatured collagen, is obtained by the partial hydrolysis of collagen. It can form a specific triple-stranded helical structure. The rate of the formation of a helical structure depends on many factors (e.g., the presence of covalent cross-bonds, gelatin molecular weight, the presence of amino acids, and the gelatin concentration in the solution). Gelatin is used in pharmaceuticals, wound dressings, and bioadhesives due to its good cell viability and lack of antigenicity. It has some potential for use in tissue engineering applications. Silk is a fibrous protein characterized by a highly repetitive primary sequence of glycine and

alanine that leads to significant homogeneity in secondary structure, β-sheets in the case of many of the silks. Silk is biodegradable due to its susceptibility to proteolytic enzymes. Silk studies *in vitro* have demonstrated that protease cocktails and chymotrypsin are capable of enzymatically degrading silk. The mechanical properties of silk provide an important set of material options in the fields of controlled release, biomaterials, and scaffolds for tissue engineering. Alginate is a straight-chain polysaccharide composed of two monomers, mannuronic acid and guluronic acid residues, in varying proportions. Alginate forms stable gels on contact with certain divalent cations, such as calcium, barium, and strontium. Alginate is widely used as an instant gel for bone tissue engineering. Chitosan, a copolymer of glucosamine and *N*-acetylglucosamine is a crystalline polysaccharide. It is synthesized by the deacetylation of chitin. Chitosan degrades mainly through lysozyme-mediated hydrolysis, with the degradation rate being inversely related to the degree of crystallinity. Chitosan has excellent potential as a structural base material for a variety of tissue engineering application, wound dressings, drug delivery systems, and space-filling implants. Hyaluronate, a glycosaminoglycan is a straight-chain polymer composed of glucuronic acid and acetylglucosamine. It contributes to tissue hydrodynamics, movement, and proliferation of cells *in vivo*. Hyaluronan is enzymatically degraded into monosaccharides. It has been used in the treatment of osteoarthritis, dermal implants, and prevention of postsurgical adhesions.

Polymeric Scaffold Fabrication Techniques (6–8). Scaffolds for tissue engineering, in general, are porous to maximize cell attachment, nutrient transport, and tissue growth. A variety of processing technologies have been developed to fabricate porous 3D polymeric scaffolds for tissue engineering. These techniques mainly include solvent casting and particulate leaching, gas-foaming processing, electrospinning technique, rapid prototyping, and thermally induced phase-separation technique, which are described below.

Solvent casting and particulate leaching is a simple, but commonly used method for fabricating scaffolds. This method involves mixing water soluble salt (e.g., sodium chloride, sodium citrate) particles into a biodegradable polymer solution. The mixture is then cast into the desired shape mold. After the solvent is removed by evaporation or lyophilization, the salt particles are leached out and leave a porous structure. This method has advantages of simple operation and adequate control of pore size and porosity by salt/polymer ratio and particle size of the added salt. However, the interconnectivity between pores inside the scaffold is often low, which seems to be problematic for cell seeding and culture.

Gas foaming is marked by the ability to form highly porous polymer scaffold foams without using organic solvents. In this approach, carbon dioxide is usually used as a foaming agent for the formation of polymer foam. This approach allows the incorporation of heat sensitive pharmaceuticals and biological agents. The disadvantage of this method is that it yields mostly a nonporous surface and closed-pore structure.

Electrospinning is a fabrication process for tissue engineering that use an electric field to control the formation and deposition of polymer fibers onto a target substrate. In electrospinning, a polymer solution or melt is injected with an electrical potential to create a charge imbalance. At a critical voltage, the charge imbalance begins to overcome the surface tension of the polymer source, and forms an electrically charged jet. The jet within the electric field is directed toward the ground target, during which time the solvent evaporates and fibers are formed. This electrospinning technique can fabricate fibrous polymer scaffolds composed of fiber diameters ranging from several microns down to several hundred nanometers.

Rapid prototyping is a technology based on the advanced development of computer science and manufacturing industry. The main advantage of these techniques is their ability to produce complex products rapidly from a computer-aided design (CAD) model. The limitation of this method is that the resolution is determined by the jet size, which makes it difficult to design and fabricate scaffolds with fine microstructure. The controlled thermally induced phase-separation process was first used for the preparation of porous polymer membranes. This technique was recently utilized to fabricate biodegradable 3D polymer scaffolds. In this approach, the polymer is first dissolved in a solvent (e.g., dioxane) at a high temperature. Liquid–liquid or solid–liquid phase separation is induced by lowering the solution temperature. Subsequent removal of the solidified solvent-rich phase by sublimation leaves a porous polymer scaffold. The pore morphology and microstructure of the porous scaffolds varies depending on the polymer, solvent, concentration of the polymer solution, and phase separation temperature. One advantage of this method is that scaffolds fabricated with the technique have higher mechanical strength than those of the same porosity made with the well-documented salt-leaching technique.

Polymers for Drug Delivery. Over the past decade, the use of polymeric materials for the administration of pharmaceuticals and as biomedical devices has dramatically increased (9–11). One important medical application of polymeric materials is in the area of drug delivery systems. There are a few polymer molecules having a drug function, however, in most cases when polymers are used in drug delivery systems, they serve as a carrier of drugs. Table 1 lists some of the important biodegradable and nonbiodegradable polymers used in drug delivery systems.

Table 1. Typical Biodegradable and Nonbiodegradable Polymers Used in Controlled Release Systems

Nonbiodegradable Polymers	Biodegradable Polymers
Polyacrylates	Polyglycolides
Polyurethanes	Polylactides
Polyethylenes	Polyanhydrides
Polysiloxanes	Polyorthoesters
	Polycaprolactones
	Poly(β-hydroxybutyrate)
	Polyphosphazenes
	Polysaccharides

Most of the above biodegradable and nonbiodegradable polymers have been discussed in the previous sections; therefore, they are not described further here.

Stimuli-Responsive Hydrogels for Drug Delivery. Hydrogels have been used as carriers for a variety of drug molecules (10). A hydrogel is a network of hydrophilic polymers that are cross-linked by either covalent or physical bonds. It distinguishes itself from other polymer networks in that it swells dramatically in the presence of abundant water. The physicochemical and mechanical properties can be easily controlled, and hydrogels can be made to respond to changes in external factors.

In recent years, temporal control of drug delivery has been of great interest to achieve improved drug therapies. Stimuli-responsive hydrogels exhibit sharp changes in behavior in response to an external stimulus (e.g., temperature, pH, solvents, salts, chemical or biochemical agents, and electrical field). The stimuli-responsive hydrogels have the ability to sense external environmental changes, judge the degree of external signal, and trigger the release of appropriate amounts of drug. Such properties have made it very useful for temporal control of drug delivery (12,13).

Temperature-Sensitive Hydrogels. Temperature is the most widely used stimulus in environmentally responsive polymer systems. Temperature-sensitive hydrogels can respond to the change of environmental temperature. The change of temperature is not only relatively easy to control, but also easily applicable both *in vitro* and *in vivo*. Poly(*N*-isopropylacrylamide) (PNIPA) is representative of the group of temperature-responsive polymers that have a lower critical solution temperature (LCST), defined as the critical temperature at which a polymer solution undergoes phase transition from a soluble to an insoluble state above the critical temperature. The PNIPA exhibits a sharp phase transition in water at $\sim 32\ ^{\circ}C$, which can be shifted to body temperatures by the presence of hydrophilic monomers (e.g., acrylic acid). Reversely, the introduction of a hydrophobic constituent to PNIPA would lower the LCST of the resulting copolymer.

When PNIPA chains are chemically cross-linked by a cross-linker (e.g., *N,N*′-methylenebisacrylamide and ethylene glycol dimethacrylate), the PNIPA hydrogel is formed, which swells, but does not dissolve in water. The PNIPA hydrogel undergoes a sharp swelling-shrinking transition near the LCST, instead of sol–gel phase separation. The sharp volume decrease of the PNIPA hydrogel above the LCST results in the formation of a dense, shrunken layer on the hydrogel surface, which hinders water permeation from inside the gel into the environment. The PNIPA hydrogels have been studied to the delivery of antithrombotic agents (e.g., heparin), at the site of a blood clot, utilizing biological conditions to trigger drug release. Drug release from the PNIPA hydrogels at temperatures below LCST is governed by diffusion, while above this temperature drug release is stopped, due to the dense layer formation on the hydrogel surface.

Some types of block copolymers made of poly(ethylene oxide) (PEO) and poly(propylene oxide) (PPO) also possess an inverse temperature sensitive property. Because of their LCST at around body temperature, they have been widely used in the development of controlled drug delivery systems based on the sol–gel phase transition at the body temperature.

pH-Sensitive Hydrogels. Polymers with a large number of ionizable groups are called polyelectrolytes. The pH-sensitive hydrogels are cross-linked polyelectrolytes containing either acidic or basic pendent groups, which show sudden changes in their swelling behavior as a result of changing the external pH. The pendant groups in the pH-sensitive hydrogels can ionize in aqueous media of appropriate pH value. As the degree of ionization increases (via increasing or decreasing pH value in the aqueous media), the number of fixed charges on the polymer chains increases, resulting in increased electrostatic repulsions between the chains. As a result of the electrostatic repulsions, the uptake of water in the network is increased and thus the hydrogels have higher swelling ratios. The swelling of pH-sensitive hydrogels can also be controlled by ionic strength and copolymerizing neutral comonomers, which provide certain hydrophobicity to the polymer chain. The pH-sensitive hydrogels have been used to develop control release formulations for oral administration. For polycationic hydrogels, the swelling is minimal at neutral pH, thus minimizing drug release from the hydrogels. The drug is released in the stomach as hydrogels swell in the low pH environment. This property has been used to prevent release of foul-tasting drugs into the neutral pH environment of the mouth.

Sometimes, it is desirable that hydrogels with certain compositions can respond to more than one environmental stimulus (e.g., temperature and pH). Hydrogel copolymers of *N*-isopropylacrylamide and acrylic acid with appropriate compositions have been designed to sense small changes in blood stream pH and temperature to deliver antithrombotic agents (e.g., streptokinase or heparin) to the site of a blood clot.

Electrosensitive Hydrogels. The electrosensitive hydrogels, which are capable of reversible swelling and shrinking under a change in electric potential, are usually made of polyelectrolytes. The electric sensitivity of the polyelectrolyte hydrogels occurs in the presence of ions in solution. In the presence of an applied electric field, the ions (both co-ions and counterions) move to the positive or negative electrode, while the polyions of the hydrogels cannot move. This results in a change in the ion concentration-dependent osmotic pressure, and hydrogels either swell or shrink to reach its new equilibrium. The electrosensitive hydrogels exhibit reversible swelling–shrinking behavior in response to on–off switching of an electric stimulus. Thus, drug molecules within the polyelectrolyte hydrogels might be squeezed out from the electric-induced gel contraction along with the solvent flow.

Other Stimuli-Sensitive Hydrogels. Hydrogels that respond to specific molecules found in the body are especially useful for some drug delivery purposes. One such hydrogel is glucose-sensitive hydrogel, which has potential applications in the development of self-regulating insulin delivery systems.

BIBLIOGRAPHY

1. Bower DI. An Introduction to Polymer Physics. Cambridge: Cambridge University Press; 2002.
2. Fried JR. Polymer Science and Technology. 2nd ed. NJ: Pearson Education Inc.; 2003.
3. Lakes R, Park J. Biomaterials: an Introduction. 2nd ed., New York: Plenum; 1992.
4. Ratner B, Hoffman A, Schoen F, Lemons J. Biomaterials Science: An Introduction to Materials in Medicine. 2nd ed. Burlington, (MA): Academic Press; 2004.
5. Lanza R, Langer R, Vacanti J. Principles of Tissue Engineering. 2nd ed. Burlington, (MA): Academic Press; 2000.
6. Agrawal CM, Ray RB. Biodegradable Polymeric Scaffolds for Musculoskeletal Tissue Engineering Hoboken, (NJ): John Wiley & Sons, Inc.; 2001.
7. Liu X, Ma PX. Polymeric scaffolds for bone tissue engineering. Ann Biomed Eng 2004;32:477–486.
8. Smith LA, Ma PX. Nano-fibrous scaffolds for tissue engineering. Colloids and Surfaces B: Biointerf 2004;39:125–131.
9. Langer R. New methods of drug delivery. Science 1990;249:1527–1533.
10. Hoffman AS, Hydrogels for biomedical applications. Adv Drug Deliv Rev 2002;54:3–12.
11. Brannon-Peppas L. Polymers in controlled drug delivery. Med Plast Biomater 1997;4:34–44.
12. Qiu Y, Park K. Environment-sensitive hydrogels for drug delivery. Adv Drug Deliv Rev 2001;53:321–339.
13. Kikuchi A, Okano T. Pulsatile drug release control using hydrogels. Adv Drug Deliv Rev 2002;54:53–77.

See also BIOMATERIALS: POLYMERS; BIOMATERIALS, TESTING AND STRUCTURAL PROPERTIES OF.

POLYMERS. See BIOMATERIALS: POLYMERS.

PRODUCT LIABILITY. See CODES AND REGULATIONS: MEDICAL DEVICES.

PROSTHESES, VISUAL. See VISUAL PROSTHESES.

PROSTHESIS FIXATION, ORTHOPEDIC. See ORTHOPEDICS, PROSTHESIS FIXATION FOR.

POROUS MATERIALS FOR BIOLOGICAL APPLICATIONS

GRACE E. PARK
THOMAS J. WEBSTER
Purdue University
West Lafayette, Indiana

INTRODUCTION

Porous materials have received much attention in the scientific community because of their ability to interact with biological ions and molecules not only at their surfaces, but also throughout their bulk (1). Because of this intrigue, traditional applications of porous materials have involved catalysis, bioseparations, adsorption of select species, and ion exchange (1). As the tissue engineering field has emerged due to the continuous need for better implantable materials, porous materials have also found their niche in regenerative medicine. Specifically, porous materials have been employed as implants for various parts of the body (e.g., bone, cartilage, vasculature, central and peripheral nervous systems, bladder, and skin) either as stand-alone regenerative devices or as drug delivery vehicles to promote tissue growth. Problems associated with current implants and the need for better porous biomaterials in numerous anatomical locations are described below.

Most significantly, estimated annual U.S. healthcare costs related to tissue loss or to organ failure surpassed $400 billion in 1997 (2). An estimated 11 million people in the United States. have received at least one medical implant device; specifically, orthopedic implants (including fracture, fixation, and artificial joint devices) constitute the majority of these and accounted for 51.3% of all implants in 1992 (3). Among joint-replacement procedures, hip and knee surgeries represented 90% of the total and in 1988 were performed 310,000 times in the United States alone (3). Implanting an orthopedic material can be a costly procedure involving considerable patient discomfort, both of which can increase if surgical revisions become necessary after an orthopedic or dental implant is rejected by the host tissue, is insufficiently integrated into juxtaposed bone, and/or fails under physiological loading conditions. Unfortunately, the average lifetime of an orthopedic implant is only 15 years due to many factors including the lack of osseointegration into surrounding bone. Current metallic implants are for the most part nonporous with subsequent poor surface properties to promote new bone ingrowth quickly.

The reason for such a high number of implanted orthopedic–musculoskeletal devices stems from numerous bone diseases. For example, approximately one out of seven Americans suffer from some form of arthritis, which is an inflammatory condition due to wear and tear in the joint (4). The cost of arthritis and rheumatic diseases reaches $86.2 billion a year, according to a study by the Arthritis Foundation and the National Institutes of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) (5). Although a very common disease, repairing damaged articular cartilage is challenging due to its limited ability to self-repair as a result of its avascularity. In fact, one of the most popular surgical techniques to repair cartilage is not through the use of a biomaterial, but rather is a surgical technique that further injures cartilage to induce scar tissue formation. This scar tissue is intended to serve as new cartilage, but since it does not match the mechanical properties of cartilage tissue, patients receiving such treatments usually suffer from additional complications and pain after only 5 years of this procedure.

The story is not any better for vascular diseases requiring biomaterial intervention. Specifically, the leading cause of death in the United States is vascular disease (including atherosclerosis), affecting $\sim$ 58 million people (6). Atherosclerosis, which is hardening of the arteries, is caused by accumulation of cholesterol, fatty molecules, and other substances inside the vessel wall as the lumen becomes gradually narrower. Consequently, complete blockage of the lumen may result, inhibiting the blood flow

through that blood vessel. Treatments for these conditions require the use of a vascular graft, initially seeking autologous (or taken from an individual's own tissue) materials. For those patients receiving a synthetic vascular graft, success rates for vessels < 7 mm approaches only 25% after 5 years. Current biomaterials used as small diameter vascular grafts are usually nonporous and result in the eventual reaccumulation of undesirable substances that clog the vessel lumen to block blood flow.

Neurological problems also necessitate the use of biomaterials. For example, Parkinson's disease, Huntington's disease, Alzheimer's disease, and epilepsy prevail as common central nervous system (CNS) degenerative pathologies, especially targeting the aging population. While most of these diseases may cause a form of dementia (a mental deterioration), Alzheimer's disease involves the loss of nerve cells related to memory and mental functions, whereas Parkinson's and Huntington's disease affect the mind and body. Among these, > 1.5 million Americans have been affected by Parkinson's disease (6) and ~ 24.4 million people are diagnosed with Alzheimer's disease and stroke, costing $>$ \$174 billion annually (7). These diseases, however, account for only those affecting a portion of the CNS: the brain. Equally as troubling are spinal cord disfunctions. Spinal cord injuries can seriously cause damage to a person's quality of life, contributing to $\sim 200{,}000$ Americans with this disability and expenses of up to \$250,000 a year per individual as reported in 1996 (8). Various treatment methods, such as the use of pharmaceutical agents, electrical stimulation probes, and bridges or conduits to physically connect damaged regions of the spinal cord have been developed and improved. However, few clinically approved porous biomaterials are available for treating peripheral and central nerve damage. This is despite the fact that pores in biomaterials could be very useful for guiding nerve fibers through damaged tissues.

Bladder is another organ that could benefit from the use of porous biomaterials. Urinary cancer stands as one of the most common forms of bladder disease, which is the second most common malignancy of the genitourinary tract and the fourth leading cause of cancer among American men (9). Conventional treatment methods include the resection of the cancerous portion of the bladder wall in conjunction with intravesical immunochemotherapy (10). However, these treatments have been less than successful due to local and systemic toxicity of chemotherapy agents (11) and possible reoccurrence of the cancer (12–14). The best approach to resolve these problems is to completely remove the bladder wall, which clearly leads to the need for a replacement porous biomaterial with highly effective designs matching the material and mechanical properties of the native bladder tissue.

The above statistics highlight the current state of diseases in numerous organs and the potential effect porous biomaterials could have in treating these ailments. It is currently believed that porous biomaterials may be the solution to healing these damaged organs if designed appropriately. The next section will emphasize the features a successful porous biomaterial should have for regenerating tissues.

FEATURES OF THE NEXT GENERATION OF SUCCESSFUL POROUS BIOMATERIALS

An ideal porous scaffold for regenerating the tissues–organs mentioned in the previous section should have these characteristics (15,16): a highly porous three-dimensional (3D) interconnected network of pores for cell infiltration and subsequent tissue ingrowth; biodegradable or bioresorbable in a controllable manner with a rate that matches that of tissue growth; appropriate surface chemistry to promote desirable cell adhesion, proliferation, and differentiation; permeable for transporting sufficient amount of nutrients and metabolic waste to and from cells; mechanical properties that match that of the tissues surrounding the biomaterial in situ; and ease of processibility for various desired shapes and sizes to match specific tissue abnormality.

Several studies have confirmed that biomaterial pore size, interconnectivity, and permeability (among other properties) play a crucial role in tissue repair (17–19). Specifically, from the aforementioned list, in the following sections surface, mechanical, degradation, porosity, pore size, and pore interconnectivity properties important for the success of porous biomaterials are elaborated.

Surface Properties

Porous Biomaterial Surface Interactions With the Biological Milieu. Assuming that the porous biomaterial has a clean surface after synthesis (Fig. 1a), the surface will be contaminated with various substances in air (e.g., hydrocarbons, sulfur, and nitrogen compounds) immediately before implantation (Fig. 1a and b) (20). Sterilization and/or introducing coatings can remove or reduce the level of contaminants. The initial interaction between an implant and the biological milieu *In vivo* occurs with water molecules (Fig. 1d) as a mono- or bilayer forms on the surface depending on the porous biomaterial's surface hydrophilicity (or binding strength of water molecules to the implant surface) (21). Water layers form within nanoseconds as other ions contained in body fluids (e.g., Cl^- and Na^+) interact with the adsorbed water molecules depending on the porous biomaterial surface chemistry (Fig. 1e). It is also possible that water interactions containing ions can penetrate the bulk porous biomaterial. Subsequently, proteins adsorb to their surfaces via initial adsorption, and then possible protein conformational changes or denaturation occurs (Fig. 1f). Replacement of these initial proteins with other proteins contained in bodily fluids may occur when biomolecules with stronger binding affinities approach the surface at a later time. Final conformations of the adsorbed proteins may differ from what occurred initially (Fig. 1g). Cells then interact with or bind to the adsorbed proteins on the porous biomaterial surface (Fig. 1h). The type of cells attached to proteins adsorbed on material surfaces and their subsequent activities will determine the tissue formed on the surface (Fig. 1i).

Protein Interactions with Porous Biomaterials

Protein Structure. Clearly, as just mentioned, one of the key events that will determine porous biomaterial success or failure is initial protein adsorption. To further explore

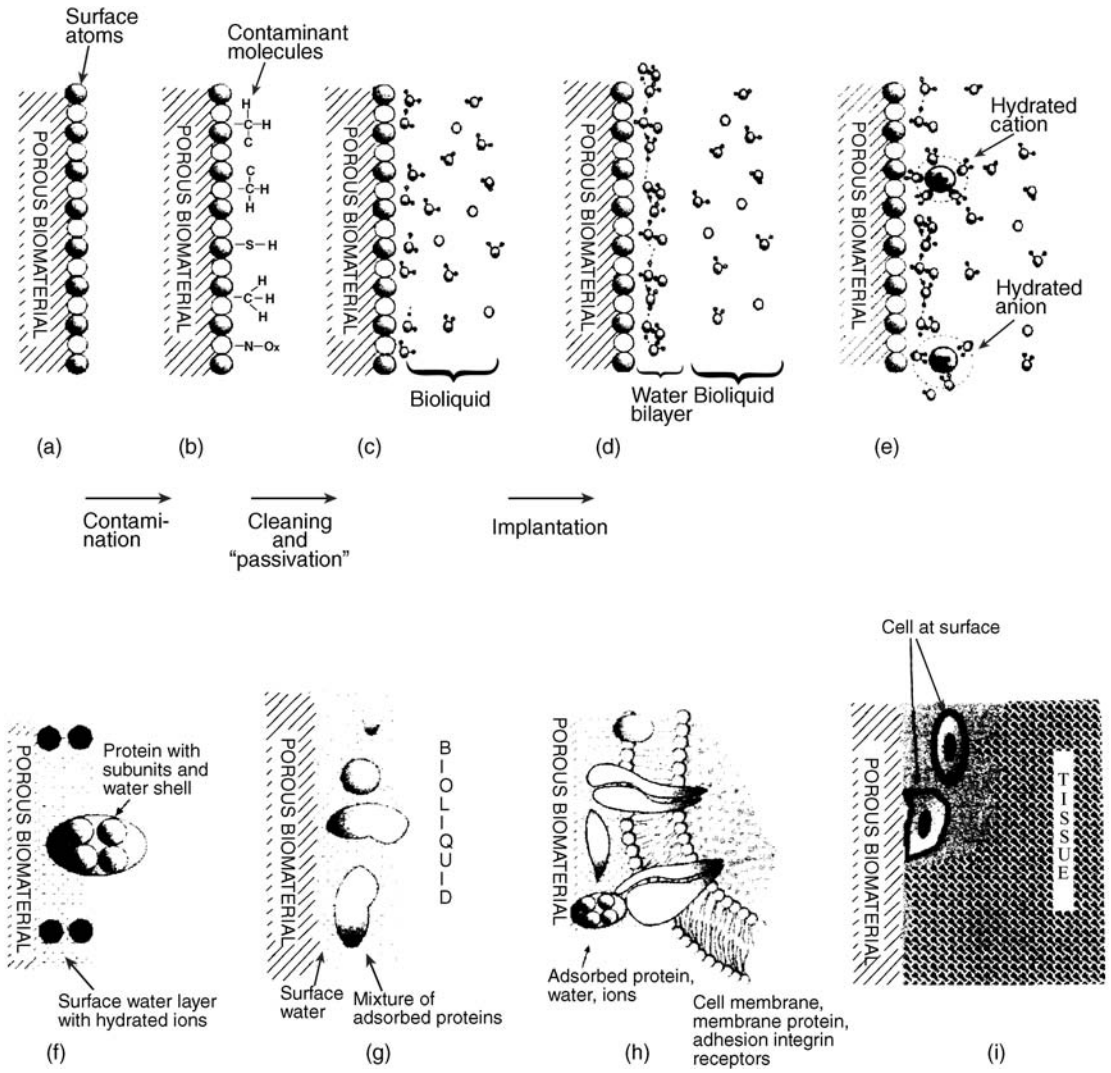


Figure 1. Schematic of the porous biomaterial–tissue interface. (Adapted from Ref. 20) (a) An initially clean porous biomaterial surface possesses surface atoms. (b) A porous biomaterial surface is contaminated with molecules from the ambient environment. (c) The surface is cleaned and passivated by saturation of dangling bonds. (d) A water bilayer forms immediately after implantation. (e) Hydrated ions (e.g., Na^+, and Cl^-, and Ca^{2+}) are incorporated into the water layer. (f) Proteins adsorb onto the surface depending on their concentration and size as well as properties of the porous biomaterial surface. (g) Various types of proteins adsorb to the surface at different conformations. (h) Cells bind to the proteins that adsorbed on the porous biomaterial surface. (i) Activity of the cells at the interface determines the type of tissue formed at that site.

this, first protein structure must be discussed. There are four levels of protein structure: primary, secondary, tertiary, and quaternary structures. It is important to understand how these different types of protein structures influence initial interactions with surfaces and consequently control cellular adhesion. The primary structure of a protein is its linear sequence of amino acids. Each amino acid is linked to another through peptide bonds. Some amino acids have side chains that are charged or are neutral. Those of particular importance in aqueous solutions exhibit polar characteristics. Other amino acids change their properties depending on the pH of the solution they reside in. Therefore, it should not be surprising that proteins exist with a wide range of properties as shown in Table 1. Table 1 describes the diverse nature of proteins in terms of size, shape, stability, and surface activity. To emphasize this diversity in protein properties, note the different interactions of albumin compared to fibrinogen on polyethylene (Table 1). Albumin is a cell nonadhesive protein while fibrinogen adsorption enhances a series of events leading to blood clot formation, a common problem of porous biomaterials in vascular applications.

Secondary protein structure consists of ordered structures in the protein chain. Two main secondary structures of proteins are the α-helix and β-pleated sheet. The degree of these structures may vary in a single protein and they are controlled by hydrogen-bonding mechanisms, which

Table 1. Diverse Properties of Proteins [a]

Protein	Function	Location	Size, kDa	Shape, nm	Stability	Surface Activity
Albumin	Carrier	Blood	65	4.2 × 14.1	Denatures at 60 °C	Low on polyethylene
Fibrinogen	Clotting	Blood	340	46.0 × 6. (trinodular string)	Denatures at 56 °C	High on polyethylene
IgG	Antibody	Blood	165	T-shaped		Low on polyethylene
Lysozyme	Bacterial lysis	Tear; hen egg	14.6	4.5 × 3.0 (globular)	$\Delta G_n = -14$ kcl·mol^{-1}	High on negatively charged surfaces
Hemoglobin	Oxygen carrier	Red blood cells	65	5.5 (spherical)	Normal form	Very high on polyethylene
Hemoglobin S	Oxygen carrier	Sickle red blood cells	65	5.5 (spherical)	Less than hemoglobin	Much higher air–water activity than hemoglobin
Myoglobin	Oxygen carrier	Muscle	16.7	4.5 × 3.5 × 2.5 spherical)	$\Delta G_n = -12$ kcl·mol^{-1}	
Collagen	Matrix factor	Tissue	285	300.0 × 1.5 (triple helical rod)	melts at 39 °C	
Bacteriorhodopsin	Membrane protein		26	3.0–4.0 long	$\Delta G_n = -8.8$ kcl·mol^{-1} denatures at 55 °C	High at cell membrane
Tryptophan Synthase alpha Subunit (wild type)	Enzyme		27		$\Delta G_n = -16.8$ kcl·mol^{-1}	High air–water activity compared to ovalbumin
Tryptophasn Synthase Variant alpha Subunit	Enzyme		27			Much less active at air–water interface than wild type

[a]See Ref. 22.

are electrostatic attractions between oxygen of one chemical group and hydrogen of another chemical group.

Tertiary protein structures are the overall 3D shape of the protein that can be quite ordered or extremely complicated. The tertiary structure of proteins is a consequence of its primary structure as it depends on the spontaneous interactions between different amino acids and, under aqueous conditions, the spontaneous interactions between amino acids and water. There are four main interactions among residues of amino acids that contribute to the tertiary structure of proteins, each with different strengths: covalent, ionic, hydrogen, and van der Waals bonds. Of these interactions, covalent bonds are the strongest, ionic bonds are also strong (occurring between chemical groups with opposite charges), and van der Waals forces resulting from interactions between hydrophobic molecules are the weakest. However, the most influential bonds on protein tertiary structure are the weakest bonds: hydrogen bonds and van der Waals bonds. This is true since, compared to covalent and ionic bonds, these weaker bonds have many more opportunities for interacting in protein tertiary structure. In addition, because proteins exist in aqueous media, residues of amino acids must interact with water, which is a highly polar compound that forms strong hydrogen bonds. Therefore, the most stable structure of proteins in aqueous media is globular, having hydrophobic areas in the center and hydrophilic areas in the outer layer. Thus, although a generalization, it is possible that the adsorption of proteins to a porous biomaterial surface will be influenced by the presence of these hydrophilic amino acids on the outside of proteins in solution. However, when proteins come in contact with solid surfaces (e.g., porous biomaterials), protein structure will drastically change.

Only proteins that posses numerous subunits have quaternary structure. How these subunits interact will determine the quaternary structure of the protein. Interactions between amino acids on the exterior of the tertiary structure (mostly hydrophilic) will influence the quaternary structure, but certainly some hydrophobic interactions will also occur at the surface and impact quaternary structure.

Under certain extreme conditions (e.g., conditions that are outside of the physiological range or outside the range of 0–45 °C, pH 5–8, and in aqueous solutions of ~ 0.15 *M* ionic strength), proteins may loose their normal structure (23). In other words, under such conditions, the spherical or globular tertiary structure most soluble proteins assume in aqueous media will unfold or denature. The structure of denatured proteins has been described as a random coil structure similar to those found in synthetic polymers (23). Since the structure of the protein has changed from that of a hydrophilic–hydrophobic exterior–interior to a more random arrangement, often times denatured proteins loose their solubility, become less dense (folded protein structures have densities of ~ 1.4 g·cm^{-3}), and loose their bioactivity (23). Although there have been many examples of protein denaturation in solution, in general, only few

cases of full protein denaturation on porous biomaterial surfaces have been reported (23). That is, generally, proteins adsorbed at the solid–liquid interface are not fully denatured and retain some degree of structure necessary to mediate cell adhesion.

Protein Interactions Mediated by Surfaces. Soluble proteins present in biological fluids (e.g., blood plasma) are the type of proteins that are involved in immediate adsorption to surfaces (24). In contrast, insoluble proteins that comprise tissues (like collagen and elastin) are not normally free to diffuse to a solid surface; these proteins may, however, appear on solid surfaces of implantable devices due to synthesis and deposition by cells (23). As mentioned, in seconds to minutes, a monolayer of adsorbed protein will form on solid surfaces (23). The concentration of proteins adsorbed on a material surface is often 1000 times greater than in the bulk phase (23). Thus, extreme competition exists for protein adsorption due to a limited space available on the surface. Because of their diverse properties just described, proteins do not absorb indiscriminately to every material surface; that is, complimentary properties of the surface and of the protein as well as the relative bulk concentration of each protein determine the driving forces for adsorption (25,26). Moreover, this initial interaction is extremely important since some proteins are not free to rotate once adsorbed to material surfaces due to multiple bonding mechanisms. Thus, immediately upon adsorption, proteins are somewhat fixed in a preferred orientation or bioactivity to the bulk media that contains cells (23). Some porous biomaterial surface properties that have influenced protein adsorption events include chemistry (i.e., ceramic versus polymer), wettability (i.e., hydrophilicity compared to hydrophobicity), roughness, and charge as will be discussed later.

One of the major differences between a flat two-dimensional (2D) substrate surface and that of a 3D porous material is tortuosity. Clearly, protein interactions are much different on materials due to tortuosity. Specifically, a curved porous surface allows for greater surface area, enhanced interactions between adjacent electrons of the atoms on the surface of the pores, increased localization of point charges, and the potential for greater surface energy due to a larger juxtaposition of localized surface defects. Collectively, all of these differences between a nonporous and porous biomaterial provide for a much more complex environment for interactions between proteins and pore surfaces. It is the challenge of the porous biomaterial community to understand this challenge and thus design scaffolds that control select protein interactions.

Protein-Mediated Cell Adhesion. Interactions of proteins (both their adsorption and orientation or conformation) on porous biomaterials mediate cell adhesion. These interactions lead to extreme consequences for the ultimate function of an implanted device (27,28). An example of the importance of protein orientation for the adhesion of cells is illustrated in Fig. 2. A typical cell is pictured in this figure with integrin receptors that bind to select amino acid sequences exposed once a protein adsorbs to a surface (Fig. 1h). It is the ability of the cell to recognize such exposed amino acids that will determine whether a cell adheres or not. For example, many investigators are designing porous biomaterials to be more cytocompatible. However, it is the adhesion of select cells that must be emphasized. That is, many attempts have been made to immobilize select cell adhesive epitopes in proteins (e.g., the amino acid sequence arginine-glycine-aspartic acid or RGD) onto polymeric tissue engineering scaffolds. But, once implanted into bone, not only do desirable osteoblasts adhere, but so do undesirable fibroblasts (cells that contribute to soft not bony tissue juxtaposition).

Not only will cell adhesion be influenced by the exposure of amino acids in adsorbed proteins, but so will subsequent cell functions (e.g., extracellular matrix deposition). This is true since for anchorage-dependent cells, adhesion is a

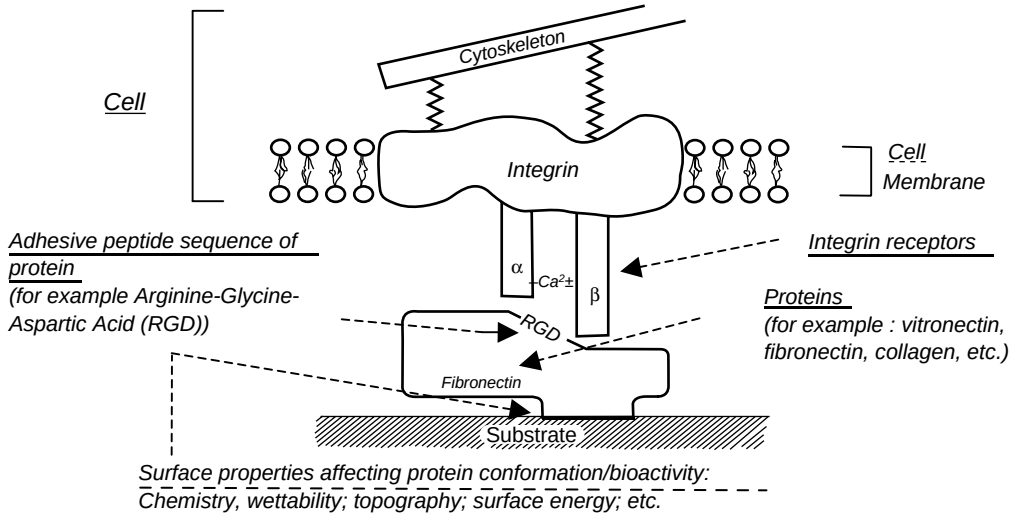


Figure 2. Influence of protein conformation on cell integrin binding. Cell adhesion and its subsequent activity will be determined by the type of integrins that the cell uses to adhere to adsorbed proteins. (Adapted and redrawn from Ref. 35.)

crucial prerequisite for subsequent cell functions. Moreover, specific intracellular messages that control subsequent cell functions are transferred inside the cell depending on which integrin receptors are utilized by the cell to adhere to adsorbed proteins. For example, a recent study by Price et al. (29), demonstrated that new bone growth was promoted when osteoblasts adhere via heparin sulfate proteoglycan binding mechanisms (as opposed to RGD) to vitronectin adsorbed on porous ceramic scaffolds.

In this manner, it is clear that cells interact with their external environment through mechanical, electrical, and chemical signals transmitted through the cell membrane. As mentioned, cell adhesion is established through cell-binding regions of extracellular matrix proteins and respective cell-membrane-intercalated receptors (i.e., integrins) among other mechanisms. Integrins are a family of transmembrane heterodimeric glycoproteins that are receptors for specific epitopes of extracellular matrix proteins and for other cell-surface molecules (30). Integrins exist as a dimer complex composed of an α-subunit (120–180 kDa) noncovalently associated with a β-subunit (90–110 kDa) (31). Several of these integrins have been identified that are concentrated at loci, called focal adhesion sites, of close proximity between cells and extracellular matrices on substrates (31). Focal adhesion sites are points of aggregation of, and are physically associated with, intracellular cytoskeletal molecules that control, direct, and modulate cell function in response to extracellular signals (32).

However, integrin–protein interactions are not the only mechanisms by which cells adhere. Several articles suggested that *In vivo* (6) and *In vitro* (33,34) osteoblasts (bone-forming cells) attach to an implanted material through cell membrane heparin sulfate proteoglycan interactions with, for example, heparin-binding sites on fibronectin and collagen. Moreover, Nakamura and Ozawa (6) immunohistochemically detected heparin sulfate on the membranes of osteoblasts attached to bone matrix.

Whatever the method of cell attachment, protein orientation will alter from surface to surface, since neither proteins nor materials are homogeneous in properties or structure on the exterior. The existence of protein regions that are largely acidic–basic or hydrophilic–hydrophobic or have select amino acids exposed to the media will greatly influence how that protein adsorbs to a surface and, thus, its orientation. Similarly, ceramics, metals, polymers, and composites thereof have vastly different chemistries and atomic bonding mechanisms (i.e., ionic, metallic, and covalent) to influence protein interactions. The initial interactions between proteins important for cell functions and the design of better porous biomaterials is emphasized in the next section.

Design of Better Porous Biomaterial Surfaces. As mentioned, not only do properties of proteins determine the degree of their interactions with surfaces, but properties of the media and surface (specifically, wettability, surface energy, chemistry, roughness, etc.) also influence the degree of protein interactions (35). Clearly, altering surface properties to control such protein events for mediating cell function leading to tissue regeneration is at the heart of the research of many biomaterial scientists and engineers. Surface properties are so important because proteins have relatively large sizes and correspondingly large numbers of charged amino acid residues of different acidity/basicity well distributed on their exteriors. The polyelectrolytic property of proteins provides for exciting design criteria in surfaces to maximize or minimize specific protein interaction. Not surprisingly, at a neutral or slightly charged surface and at a pH in which the net charge on the protein is minimal, most proteins will exhibit maximum adsorption (23). For surfaces with a large net charge, initial protein interactions will be dominated by the degree of the opposite charge on the surface (23,35).

Consideration of the spatial organization of amino acids can be used in the design of surfaces to enhance protein interactions (36). As previously mentioned, for some proteins, hydrophilic and hydrophobic amino acids are present primarily on the exterior and interior, respectively. This spatial arrangement has a direct consequence on the initial interactions of these proteins with surfaces. For example, a surface that initiates interactions with the exterior hydrophilic amino acid residues in that type of a protein may promote its adsorption. In contrast, for the interior hydrophobic amino acid residues to interact with material surfaces, which may contain desirable cell adhesive epitopes (e.g., RGD), the soluble protein may have to unfold or loose tertiary structure. For this reason, one approach to increase the adsorption of a protein whose external amino acids are largely hydrophilic, would be to design a material surface which exhibits polar properties. The same can be said for any type of protein; that is, through an understanding of the amino acids that reside on the protein exterior when in the appropriate biological milieu, a complimentary surface can be designed. It is important to note, though, that this is a generalization as many proteins have a diverse collection of hydrophilic–hydrophobic amino acids externally that must be considered. In addition, as previously mentioned, proteins adsorb to surfaces in a competitive manner in which the adsorption of one protein will influence that of another.

Several studies have confirmed these speculations that properties (chemistry, charge, topography, etc.) of porous biomaterial surfaces dictate select interactions (type, concentration, and conformation–bioactivity, etc.) of proteins (24,37–40). It has been reported in the literature that changes in the type and concentration (up to 2100, 84, and 53% for albumin (40), fibronectin (41), and vitronectin (34), respectively) of protein adsorption on material surfaces depends on material surface properties, such as chemistry (i.e., polymer, metal, or ceramic), hydrophilicity–hydrophobicity, roughness, and surface energy. Consequently, since protein interactions can be controlled on porous biomaterial surfaces, so can cell adhesion. For example, a common porous biomaterial [poly(lactic-*co*-glycolic acid) or PLGA] has been modified to increase the adsorption of vitronectin and fibronectin through NaOH treatments (42–44). Since both vitronectin and fibronectin mediate osteoblast, vascular cell, and bladder cell adhesion, these NaOH treated PLGA scaffolds have found a home in numerous tissue engineering applications.

However, for the field of porous biomaterials to advance even further, instead of broadly speaking of protein adsorption on surfaces, researchers need to investigate and design succinct regions of surfaces to promote protein adsorption considering the complexities of their properties. Only when porous biomaterials are considered from the context of protein interactions necessary for desirable cell interactions, will better tissue engineering materials be formulated.

Mechanical Properties and Degradation Byproducts

Although porous biomaterial surface properties determine cell attachment, mechanical strength of the scaffold and the mechanical environment it provides plays an equally important role in enhancing subsequent cell functions leading to tissue growth (45). Mechanical forces felt by cell membrane molecules are interconnected to the cytoskeleton that can influence messenger ribonucleic acid (mRNA) and subsequent synthesis of intracellular proteins (all the way to the nucleus where gene expression can be changed). It is for these reasons that mechanical properties must also be carefully controlled in porous biomaterials. For example, a study of various mechanical stimuli placed on equine articular chondrocytes within nonwoven polyglycolic acid (PGA) mesh scaffolds indicated that when the stimuli were removed, after a period of 1 week, the mechanical integrity of the resulting tissue construct was lost (46). This result implies that the mechanical stimuli applied to cells within a porous biomaterial may influence the biomechanical functionality of the regenerated tissue.

Although most agree that the mechanical properties of a porous biomaterial should match those of the physiological tissue they are intending to replace, the specific parameters and values desirable in these studies vary. For bone tissue engineering, for example, Yaszemski et al. (47) stated that scaffolds should possess mechanical stiffness matching the low range values of trabecular bone (50–100 MPa), whereas Hutmacher's design principle (15,48) suggests matching the native tissue stiffness (10–1500 Mpa for trabecular bone (49)). Clearly, this wide range in mechanical values can provide for much different porous biomaterial efficacies and a consensus needs to be established.

Once deciding on the optimal mechanical properties needed in scaffold structures, there are numerous design parameters that can be exploited to match such values. For example, for a fibrous mesh, a decrease in fiber diameter increases mechanical strength due to an increase in fiber density (50). Obviously, increasing percent porosity and the diameters of individual pores can also be used to decrease the strength of scaffolds to match desired values. These properties not only influence inherent mechanical properties of scaffolds, but they can also be used to manipulate cell functions.

Specifically, Maroudas postulated that the scaffold surface rigidity or stiffness enhances cell adhesiveness and cell spreading (51). Pelham and Wang (52) have shown that focal adhesion contacts in cells and their migration on acrylamide gels are controlled by scaffold flexibility. They also suggested that tyrosine phosphorylation might be involved, activated by local tension at cell adhesion sites (53). Recently, Ohya et al. (54) studied the effects of hydrogel mechanical properties on cell adhesiveness and found that the higher the strength of the hydrogel formulation, the greater the capability to withstand cell traction forces, thereby resulting in greater cell spreading. These authors also noticed that cells preferred to adhere to stiffer regions within the hydrogel.

Common pore shapes in porous biomaterials include tube-like, spherical, and randomly spaced shapes. Differences in cell attachment, growth, migration, and matrix deposition by cells have all been observed depending on pore structure. Specifically, certain cell types prefer a select pore structure in accordance to their physiological matrix environment. For example, orthopedic tissue engineering scaffolds should have spherical pores with a high porosity to allow for immediate bone ingrowth, while maintaining the mechanical strength and integrity necessary due to their harsh mechanical environments *In vivo* (36). Porous biomaterial pore shapes are critically related to pore interconnectivity. Not only does pore interconnectivity in a porous biomaterial affect nutrient–waste diffusion, but it also influences cell growth. Bignon et al. (55) observed that the density of pore interconnections determines cellular colonization rates; meaning that the larger the macropores (within limits), the fewer pore interconnections that have to be transversed by the cells thus resulting in higher colonization rates. Of course, guided cell growth or migration is possible through deliberate pore shape and interconnectivity. For example, tube-like or fibrous pore shapes may promote neurite extension from neurons in specific directions. Studies have also shown that cells prefer discontinuities within a porous material in terms of growth and migration; clearly pores provide such discontinuities (56–58).

In addition, maintaining mechanical strength and structural integrity of porous biomaterials are crucial because scaffolds may be crushed when implanted or may degrade over time. Mechanical properties are especially important to characterize when they change over time. A thorough knowledge of the degradation process of the porous materials of interest (including degradation byproducts) should be mapped in order to control the mechanical stability and the degradation rate until the native tissue is formed at the site of implantation.

For porous biomaterials, a range of biodegradation choices exist, from nondegradable metals to degradable ceramics and polymers. Importantly as well, degradation rates of porous materials have in some cases been shown to be faster compared to solid block polymers (59,60) because acidic byproducts become trapped inside the bulk as they degrade, therefore causing an autocatalytic effect. Of course, trapping of acidic byproducts in polymeric scaffolds can have detrimental consequences on cell health. Porous degradable polymers, such as PGA, polylactic acid, PLGA, and polycaprolactone (PCL) degrade via nonenzymatic random hydrolytic breaking of ester linkages. Sung et al. (61) studied the degradation of PLGA and PCL scaffolds *In vitro* and *In vivo*. They found a significant decrease in the molecular weight of these polymers within 1 month *In vitro* and, as expected, at a much faster rate *In vivo*.

Specifically, the influence of acidic byproducts from these polymeric scaffolds on cell health was investigated by measuring the pH of the media in which the polymers resided compared to the media in which tissue culture polystyrene (TCPS) was cultured. Changes in the media pH occurred only for PLGA (reducing it by 5) whereas no significant changes were measured during TCPS or PCL culture for up to 28 days (61).

Moreover, an *In vivo* study by Hedberg et al. (62) determined that soluble acidic products from degradable polymers lead to an increased recruitment of inflammatory cells compared to that induced by the scaffold itself. This was evidenced by the fact that a minimal inflammatory response was observed at the site of bone growth juxtaposed to the surface of polymeric scaffolds, whereas a major inflammatory response was observed in the scaffold where there was significant degradation. However, Sung et al. (61) suggested that an inflammatory response can be beneficial towards angiogenesis that is highly desirable to remove harmful degradation products from the interior of a polymer scaffold. This clearly demonstrates the need for controlling polymer degradation products in order to elicit a desirable response from host tissue (63). Collectively, such studies highlight the necessity for a better understanding of material degradation products on cell health.

In addition, according to Wu and Ding (64), the molecular weight of a porous PLGA scaffold decreases during degradation, which not only creates a more acidic local environment, but also leads to other changes. In their study, degradation was divided into three stages, marking distinct characteristics in mechanical properties (Fig. 3). In the first stage (I), the mechanical strength increased as the porous scaffold dimensions decreased while the weight remained constant; this can be interpreted simply as the change of porous biomaterial dimensions resulting in mechanical property increases. Increased elastic modulus of porous PLGA scaffolds with degradation time was also observed in another study by Zhang and Ma (65) who contributed this to decreased porosity of the foams with time. In the second stage (II), a dramatic decrease in mechanical properties were observed, which was correlated with an increased presence of low molecular weight degradation products (64). The third stage (III) was characterized by the breakdown of the scaffold's structural integrity and associated rapid weight loss due to pH decreases from acidic degradation products. Understanding of these three distinct phases of mechanical property degradation for every proposed porous biomaterial is imperative. In addition, more studies are needed that correlate cell function at each stage of mechanical property changes in porous biomaterials as they degrade.

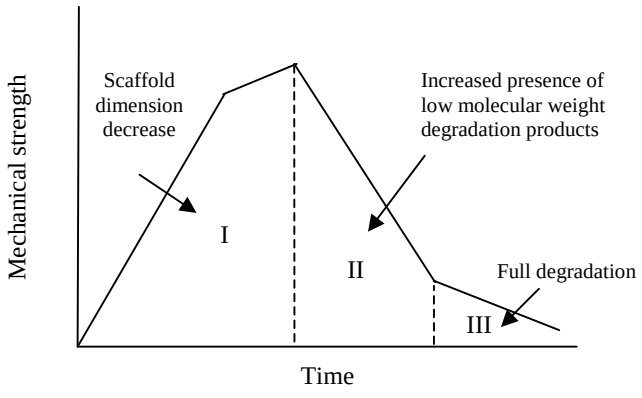


Figure 3. Three stages of mechanical strength degradation in porous biomaterials. (Adapted and redrawn from Ref. 64.)

The Role of Porosity, Pore Size, and Interconnectivity

Among other properties (e.g., the aforementioned mechanical properties), porosity can also influence how cells behave in a scaffold. Open pore structures are desirable in most tissue engineering applications, providing enhanced cell seeding, cell attachment, proliferation, extracellular matrix production, and tissue ingrowth. For example, for orthopedic applications, both *In vitro* and *In vivo* studies demonstrated exceptional osteoblast proliferation and differentiation leading to new bone growth in PLGA foams with 90% porosity (66,67). In addition, a study by Sherwood et al. (68) reported that osteochondral composite scaffolds with 90% porosity at the cartilage portion allowed full incorporation of the chondrocytes (cartilagesynthesizing cells) into the scaffold.

Permeability, or high interconnectivity of pores is a crucial property for a porous biomaterial due to its influence on cellular communication (16), adaptation to mechanical stimulation (45), and prevention of the accumulation of acidic degradation byproducts (69,70). It also allows for uniform cell seeding and distribution, as well as proper diffusion of nutrients and metabolic wastes. Studies have shown that when tissues become thicker than 100–200 μm, the oxygen supply to cells becomes limited in a static environment (71,72). Thus, interconnectivity of pores is an extremely important design consideration to increase tissue growth into porous biomaterials.

In addition, as mentioned in the section above, the increased tortuosity present in porous biomaterials will influence protein interactions and, thus, manipulate cellular functions. Specifically, because of altered initial protein interactions, certain cell types (e.g., chondrocytes) perform much better on porous compared to flat (or nonporous) biomaterials (42). Moreover, macroporosity (pore diameters > 50 μm) influences the type of cells adhering to a polymeric scaffold. For example, large pores (100–200 μm diameter) have been shown to enhance bone ingrowth compared to smaller pores (10–75 μm diameter) in which undesirable fibrous soft tissue formation has been observed (73). Yuan et al. (74) added that pore sizes < 10 μm promotes bone ingrowth due to optimal initial protein adsorption events possibly because of their greater surface areas. Furthermore, Bignon et al. (55) demonstrated greater cell spreading on biomaterials with micro (pore diameters < 10 μm) compared to macroporosity. Importantly as well, pore wall roughness is influenced by pore size that may be providing greater roughness to promote cell functions. Studies are needed to carefully control pore wall roughness to make accurate comparisons between scaffolds of various degrees of pore sizes. Since

very small topographical changes have been shown to alter cell functions (75), surface roughness values in the nanometer regime could also be incorporated into porous biomaterials regardless of their pore sizes to significantly enhance protein–surface and protein–cell interactions (42,76). Fabrication methods that can provide for the manipulation of pore properties is further emphasized here.

POROUS BIOMATERIAL FABRICATION METHODS

Various methods for fabricating porous biomaterials have been explored to date. Examples for forming polymeric porous scaffolds include solvent casting with particulate leaching, gas-foaming processes, emulsion freeze drying, freeze-extraction, electrospinning, rapid-prototyping, and thermally induced phase separation. Although polymers receive the most attention in porous biomaterial applications, porous ceramics, and metals have been recently receiving much attention. This is mostly because ceramics and metals have a long history of implantation, so methods that can improve their cytocompatibility properties (e.g., by creating pores) are highly desirable. For ceramic and metallic porous biomaterials, electrophoretic deposition, salt leaching, microsphere (polymer) melt out, and annodization have been commonly employed. These methods will be briefly described in the following sections.

Cellular Solids

Current methods, such as solvent casting, gas foaming, vacuum drying, and thermally induced phase separation (TIPS) in conjunction with particulate leaching techniques create cellular solids (77). These methods can create porous constructs easily and in an inexpensive manner (78,79). In solvent casting, a pellet or powder form of a polymer is dissolved in a solvent. Then, water-soluble salt particles (e.g., sodium chloride, sodium citrate) or other particulate materials [e.g., gelatin, paraffin (79)] are added to the polymer solution. The solvent is removed through evaporation or lyophilization and then particles are leached out through the use of water or another solvent (depending on the particle chemistry) to create the desired porous structure. The advantages of these methods include simplicity and the ability to control pore size and porosity. However, the pore shape is limited to the shape of the porogen and the pore interconnectivity is poor; thus, the porogen may not be completely removed from the construct (80). Furthermore, uneven dispersion or settling of the particles within the constructs may occur. Lastly, these first generation approaches rarely provided the succinct spatial ability to control protein adsorption necessary for the next generation of more successful biomaterials.

For the gas-foaming process, a gas, usually carbon dioxide (CO_2), is utilized instead of using an organic solvent at high pressures to create a highly porous structure (80–82). Again, these techniques are easy to implement and are inexpensive. However, a polymer with highly amorphous fractions can be processed with this technique even though the interconnectivity of the pores is very low, only $\sim$ 10–30% (18).

Thermally induced phase separation produces a highly porous material using a solvent at elevated temperatures followed by lowering the temperatures to separate the solution into liquid–liquid or solid–liquid phases. Then, the unwanted solvent is removed through sublimation (65,83). Although high mechanical strength may be obtained with this technique, the pore size created with TIPS normally ranges from 10–100 μm, which does not satisfy the permeability requirements for the removal and entry of cellular wastes and nutrients, respectively.

The emulsion freeze-drying method was developed by Whang et al. (84). A porous structure is obtained through homogenization of a polymer, organic solvent, and water mixture; rapidly cooling the mixture to maintain the liquid state structure; and then removing the solvent, and water by freeze-drying (80). In Whang's study, 90% or greater porosity and up to 200 μm diameter pores were created. However, this method is user and technique sensitive, meaning that pore structures and associated interconnectivities greatly depend on the processing method. The freeze-extraction method is a modified version of the freeze-drying technique, in which the solvent in the frozen polymer solution is replaced with a nonsolvent at temperatures below the freezing point of the polymer solution. This procedure removes the solvent before the drying stage (85).

Electrospinning Technique

In electrospinning, an electric field directs polymer fibers to form and deposit onto a substrate of interest (86,87). Specifically, an electric potential is applied as the polymer solution is injected, which ultimately forms an electrically charged jet of polymer landing on the target substrate. The solvent evaporates and porous polymer fibers are formed. Fibrous polymer scaffolds with diameters of several hundred nanometers can be fabricated using this method, thus, simulating the physiological fibrous structure of such proteins like collagen that comprise tissues. Only films and cylindrical shapes of the porous material have been created through this technique, therefore, further investigations are needed. But in addition to creating biologically inspired nanometer fibers an advantage of this process is its ability to coat an existing implant material. Thus, this technique could be used to modify the surface properties of currently used implant materials to promote cell functions.

Rapid Prototyping

Rapid prototyping is a computer-guided manufacturing system that can produce complex designs rapidly. One of the prototyping techniques is called 3D printing (3-DP) and it has been used to fabricate biodegradable polymer scaffolds for tissue engineering purposes (88). This technique produces porous biomaterials by ink-jet printing a binder onto sequential powder layers. Importantly, growth factors, proteins, cells, and other biological factors can be incorporated into the porous biomaterial without risking inactivation because the process is performed at room temperature. However, a disadvantage of this process so far includes porous biomaterial size limitations (due to the size of the ink jet). This can also limit the creation of desirable fine details or nanostructures on the polymer.

Microsphere Burnt Out

The microsphere burnt out method is similar to the previously described salt leaching method except that polymer microspheres are utilized instead of a salt porogen. This method is useful for ceramic materials that require a sintering process at very high temperatures (approaching 1000°C) at which point the polymer melts. As a very simple and easy method, it also has the disadvantages of the need for large amounts of the microspheres to create high pore interconnectivity; this results in poor mechanical properties.

Electrophoretic Deposition

Another attractive method for creating porous ceramics is electrophorectic deposition (or EPD). Due to a relatively simple setup and accommodation of complex designs and shapes, EPD has received much attention for processing fine particles, especially for coating applications (89). For this process, Ma used a graphite cathode and a stainless steel anode in the EPD cell while a current was applied to induce deposition of the particles onto a designated material. In this study, hydroxyapatite 3D porous biomaterials were fabricated. Hydroxyapatite is the main inorganic component of bone and, thus, has experienced wide spread use in orthopedic applications. This simple powder consolidation method requires no additives and high pore interconnectivity can be achieved with sufficient mechanical strength. However, this process can be costly when designing a large sample.

Anodization

Although not many methods exist to create porous metals, anodization is one that is gaining in popularity. Anodization involves the application of a voltage to a metal submerged in an electrolyte solution. Anodization has been used to create various pore sizes (from 10 nm to 1 μm) and shapes on two popular orthopedic metal chemistries: titanium and aluminum. In both cases, compared to respective unanodized metals, increased osteoblast functions have been reported on anodized titanium and aluminum (90,91). In addition, a study by Chang et al. (90) demonstrated that under certain anodization conditions porous nanotubes were created in titanium that further increased osteoblast adhesion. Although more testing is required, these studies highlight the fact that anodization is a fast and inexpensive method for creating pores in metals necessary for promoting bone growth.

Chemical Vapor Deposition

Another technique used to create porous metals is chemical vapor deposition. Chemical vapor deposition has been mostly used to fabricate porous tantalum for orthopedic applications. Tantalum is a new metal to the orthopedic field that possesses exception cytocompatibility properties. Tantalum porous biomaterials have been synthesized using vitreous carbon as the skeleton structure material (92). Tantalum was then coated onto the template and the template was removed by either chemical or heat treatments. Chemical vapor deposition is a common technique used in the coating industry and can easily be utilized for the fabrication of porous materials as long as a template or a mold is provided.

FUTURE DIRECTIONS IN THE DESIGN OF MORE EFFECTIVE POROUS BIOMATERIALS

Although there are numerous avenues, investigators are pursuing to improve the efficacy of porous biomaterials, one approach that involves the incorporation of nanotechnology seems to be working. Nanotechnology embraces a system whose core of materials is in the range of nanometers (1 nm). The application of nanomaterials for medical diagnosis, treatment of failing organ systems, or prevention and cure of human diseases can generally be referred to as nanomedicine. The commonly accepted concept refers to nanomaterials as that material with the basic structural unit in the range of 1–100 nm (nanostructured), crystalline solids with grain sizes 1–100 nm (nanocrystals), extremely fine powders with an average particle size in the range 1–100 nm (nanopowders), and fibers with a diameter in the range 1–100 nm (nanofibers). There have been many attempts to improve health through the use of nanotechnology, but perhaps the closest to clinical applications involves nanostructured biomaterials.

The greatest advantage of nanobiomedical implants in a biological context centers on scientific activities that seek to mimic the nanomorphology that proteins create in natural tissues. As seen in Figs. 4 and 5, bone and vascular tissue possesses numerous nanometer surface features due to the presence of entities like collagen and other proteins (93). Dimensions of some additional proteins found in the extracellular matrix of numerous tissues are found in Table 2 (94). As can be seen, the fundamental dimensions of these proteins (and all proteins) are in the nanometer regime. Clearly, when assembled into an extracellular matrix that comprises a tissue, these proteins provide a diverse surface with numerous nanostructured features for cellular interactions. Since some of these proteins are also soluble and present in bodily fluids, they will initially adsorb to implanted materials to provide for a highly nanostructured surface roughness for cellular interactions. It is for these reasons that cells of our body are accustomed to interacting with nanostructured surfaces. This is in stark contrast to most conventional porous biomaterials that are smooth at the nanoscale.

Aside from mimicking the surface roughness of natural tissues, there are other more scientific reasons to consider porous nanostructured biomaterials for tissue regeneration. Specifically, surface properties (e.g., area, charge, and topography) depend on the surface feature sizes of a material (95,96). In this respect, nanophase materials that, by their very nature, possess higher areas with increased portions of defects [e.g., edge/corner sites and grain or particle boundaries (95,96)] have special advantageous properties that are being exploited by porous biomaterial scientists for applications involving proteins and cells. As mentioned, proteins have complex structures and charges. Thus, surfaces with biologicallyinspired nanometer roughness provide control over protein interactions that were not

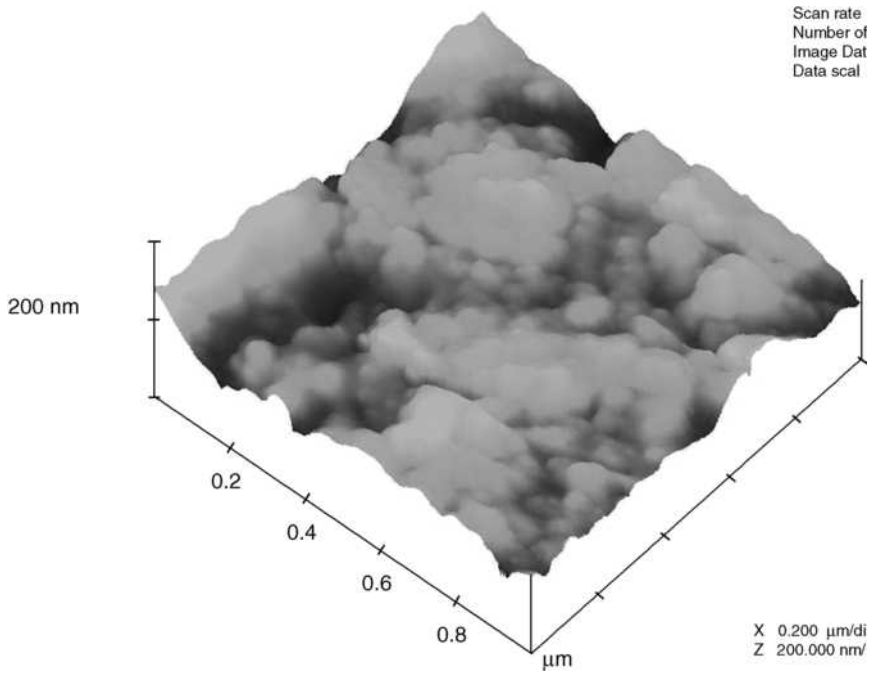


Figure 4. An AFM image of the surface of bovine cortical bone. Numerous nanometer features of bone duplicated in porous biomaterials are showing progress in orthopedic applications.

possible with conventional porous materials. Advances of porous nanostructured biomaterials pertinent to orthopedic, cartilage, vascular, central and peripheral nervous systems, and bladder applications will be briefly discussed in the sections below.

Orthopedic Applications

Nanophase ceramics (including alumina, titania, and hydroxyapatite), metals (e.g., titanium, titanium aluminum alloys, and cobalt chromium alloys), polymers (specifically, PLGA, polyether urethane, and polycaprolactone), and composites thereof have been explored for orthopedic applications (22,97,98). For these studies, nanophase surface features in ceramics and metals were created by using constituent nanometer particles, whereas nanostructured polymers were created using chemical etching techniques. In all of these studies, regardless of the manner in which the materials were synthesized, results indicated that nanophase materials enhanced osteoblast functions (e.g., attachment, proliferation, production of extracellular matrix proteins, and deposition of bone) compared to their respective conventional formulations (Fig. 6).

In addition, other porous biomaterials with nanostructured surface features [e.g., carbon nanotubes in polymer composites (Fig. 7) and porous helical rosette carbon nanotubes (Fig. 8)], increased osteoblast functions over conventionally used PLGA scaffolds (99,100). Interestingly, as opposed to conventional porous biomaterials, helical rosette carbon nanotubes self-assemble into a porous biomaterial, which when heated to temperatures only slightly above body temperature solidify (100); thus, these materials could be formulated immediately before implantation to match the dimensions of any bony defect. These novel porous helical rosette nanotubes also allow for optimal pore interconnectivity for the transfer of nutrients and waste to and from cells (100).

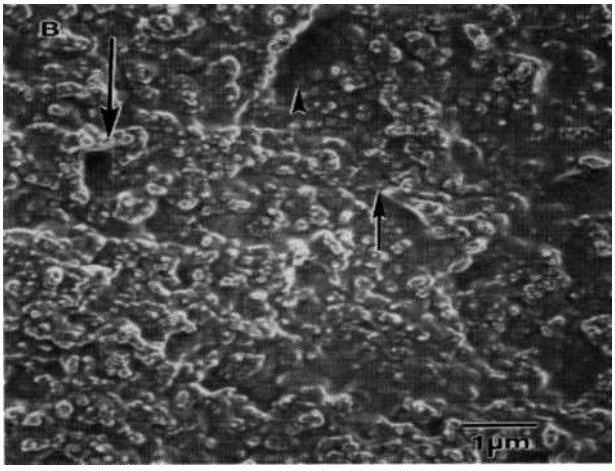


Figure 5. Cast replica of vascular tissue demonstrating nanometer roughness. (Adapted from Ref. 93.) Vascular tissue has numerous irregular nanometer features that when duplicated in porous biomaterials show progress in vascular applications.

Table 2. Nanometer Dimensions of Extracellular Matrix Proteins[a]

Protein	Characteristic Dimensions
Fibronectin	Dimmer of two identical subunits; 60–70 nm long; 2–3 nm wide
Vitronectin	Linear molecule 15 nm long
Laminin	Cruciform configuration with one 50 nm long arm and two 35 nm short arms; total length 50 nm; total width 70 nm
Collagen	Triple helical linear protein consisting of 2 α(1)-chains and one α(2); 300 nm long; 0.5 nm wide; 67 nm periodicity

[a]See Ref. 94.

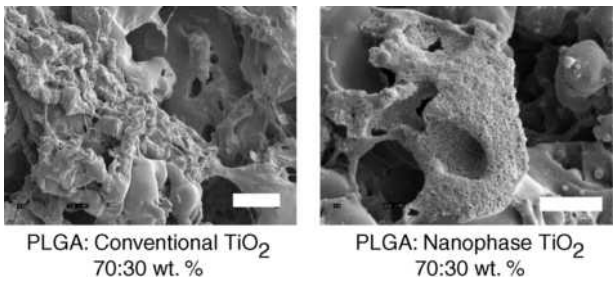


Figure 6. The scanning electron microscopy (SEM) images of PLGA composites containing either conventional or nanophase titania. Increased bone regeneration has been measured in polymer composites containing nanometer compared to conventional ceramics. Bar = 10 μm.

Cartilage Applications

Such pore interconnectivity is also crucial for cartilage forming cells, chondrocytes, since chondrocytes reside far apart from each other and their main communication is through their extracellular matrix. Recently, a porous biomaterial matrix fabricated via solvent casting and particulate leaching to create nanometer surface roughness was tested for cartilage applications (42). The polymer used was PLGA and it was modified to possess nanometer surface features through soaking for 10 min in 10 *N* NaOH (Fig. 9). Compared to conventional PLGA, results showed increased chondrocyte adhesion, proliferation, and synthesis of a cartilage extracellular matrix (as noted by collagen and glycosaminoglycan synthesis) (42).

Vascular and Bladder Applications

Not only do osteoblasts and chondrocytes interact better with nanophase materials, but so do other cells such as vascular (including endothelial and vascular smooth muscle cells) and bladder cells. For example, Miller et al. (43) and Thapa et al. (44) created nanometer surface features on PLGA films by developing novel molds of NaOH treated PLGA (Fig. 10). When compared to PLGA films without nanometer surface features, vascular smooth muscle cell,

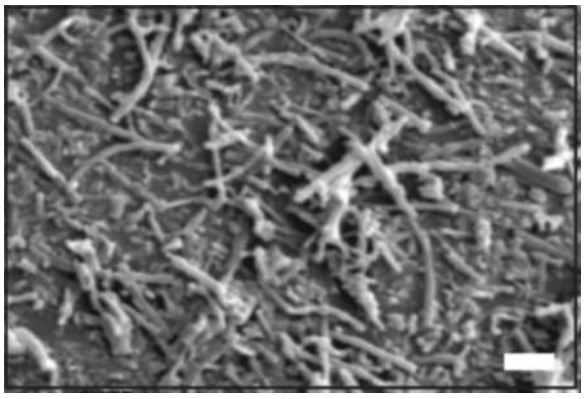


Figure 7. The SEM image of a polyether urethane composite containing nanophase carbon fibers. Increased bone regeneration has been measured in polymer composites containing nanometer compared to conventional carbon fibers. Bar = 1 μm.

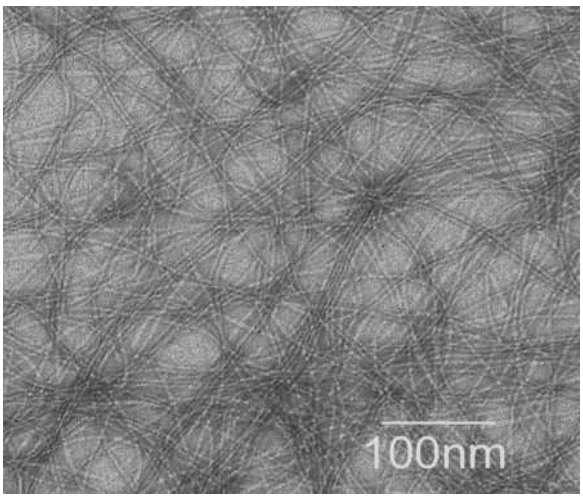


Figure 8. The transmission electron microscopy (TEM) micrograph of porous helical rosette carbon nanotubes. Individual outer-tube diameters are 4.6 ± 0.09 nm. Increased bone regeneration has been measured in helical rosette nanotubes compared to currently used titanium implants.

endothelial cell, and bladder smooth muscle cell functions were enhanced on the nanostructured PLGA. For bladder applications, Pattison et al. (76) created NaOH induced nanofeatures onto 3D PLGA scaffolds and also observed greater bladder smooth muscle cell adhesion, proliferation, and collagen synthesis. Their studies have further demonstrated increased fibronectin and vitronectin adsorption on nanostructured PLGA compared to conventional PLGA, thus, providing a key mechanism for why vascular and bladder cell adhesion is enhanced on nanostructured PLGA surfaces (43). In addition, PCL and polyurethane have been modified to possess nanostructured surface features by NaOH and HNO_3 treatments, respectively; increased vascular and bladder cell functions have been measured on these treated compared to nontreated polymers (43,44). Such studies highlight the versatility of modifying numerous polymers to possess nanostructured features for enhanced vascular and bladder applications.

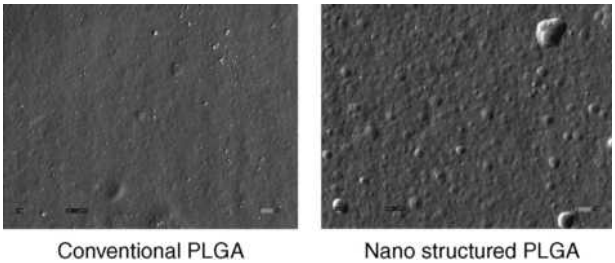


Figure 9. The SEM images of PLGA possessing conventional and nanoscale surface roughness. Nanoscale surface roughness was created by fabricating molds of PLGA etched in 10 *N* NaOH for 10 min. Increased functions of chondrocytes, vascular cells, and bladder cells have been measured on polymers with nanoscale compared to conventional surface features. Bar = 1 μm.

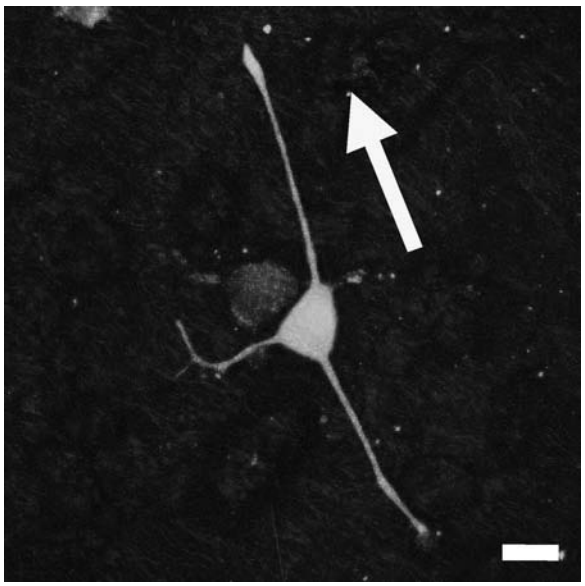


Figure 10. Fluorescent microscopy image of neuron axon alignment corresponding with aligned carbon nanofibers in polyether urethane composites. The arrow indicates carbon nanofiber alignment. Bar = 20 μm.

Central and Peripheral Nervous System Applications

Finally, McKenzie et al. (101) also provided evidence that biomaterials created to have numerous nanometer features can decrease glial scar tissue formation while at the same time increase interactions with neurons. These materials were created by combining carbon nanofibers to polyether urethane. In addition, these investigators have aligned carbon nanofibers in such porous structures to control the direction of axon extension from neurons. In this manner, such porous nanostructured biomaterials could be used to regenerate electrical activity in damaged areas of the brain.

CONCLUSIONS

All these indications point to the conclusion that a successful implantable porous biomaterial should possess properties and structures that simulate the formation of an extracellular matrix similar to that of the target organ it intends to replace. Equally as important for porous biomaterials are appropriate mechanical strength, mechanical structural integrity, degradation rate, permeability, porosity, pore structure, pore interconnectivity, surface energy, surface roughness, and surface chemistry. These all play a role in the function of an optimal porous biomaterial to regenerate tissue. Importantly, to date, to address some of these material properties, several processing techniques have been developed. Although much more needs to be learned concerning the most important aspects of tissue regeneration on porous biomaterials, proper attachment of the appropriate cell is crucial. This is mediated by initial protein interactions that must be the focus of future endeavors to design more effective porous biomaterials. Recent evidence has been provided that porous biomaterials with nanostructured surface roughness might just control initial protein interactions pertinent for enhancing cell functions necessary to improve the efficacy of orthopedic, cartilage, vascular, central and peripheral nervous system, and bladder applications.

BIBLIOGRAPHY

1. Davis ME. Ordered porous materials for emerging applications [Review]. Nature (London) 2002;417(6891):813–821.
2. Niklason LE, Langer R. Advances in tissue engineering of blood vessels and other tissues. Transplant Immunol 1997;5:303–306.
3. Praemer A, Furner S, Rice SD. Musculoskeletal Conditions in the United States, Park Ridge, IL: American Academy of Orthopaedic Surgery; 1992.
4. D' Angelo K, Austin T. The moving target: Understanding why arthritis patients do not consider total joint replacement. American Academy of Orthopaedic Surgeons: News Release July 6, 2004.
5. Arthritis Today. 52 Ways to bite back–5 astounding figures. The Arthritis Foundation. Available at http://www.arthritis.org/resources/arthritistoday/2003_archives/2003_09_10_51_ways_5figures.asp [2005, March 31].
6. American Parkinson's Disease Association. Available at http://www.apdaparkinson.org/user/AboutParkinson.asp. [2005, March 31]. Nakamura H, Ozawa H. Immunohistochemical localization of peharan sulfate proteoglycan in rat tibiae. J Bone Mineral Res 1994;9:1289–1299.
7. National Center for Chronic Disease Prevention. Available at www.cdc.gov/diabetes/pubs/costs/figure.htm figure1, June 1998.
8. National Institute of Health. Spinal cord injury: emerging concepts. NIH Proc Sept 30–Oct. 1, 1996.
9. Greenlee RT, Murray T, Bolden S, Wingo PA. Cancer statistics, 2000. CA Cancer J Clin 2000;50:7–33.
10. Melekos MD, Moutzouris GD. Intravesical therapy of superficial bladder cancer. Curr Pharm Des 2000;6:345–359.
11. Highly MS, Oosterom AT, Maes RA, De Bruijn EA. Intravesical drug delivery. Pharmacokinetic and clinical considerations. Clin Pharmacokinet 1999;37:59–73.
12. Dalbagni G, Herr HW. Current use and questions concerning intravesical bladder cancer group for superficial bladder cancer. Urol Clin North Am 2000;27:137–146.
13. Bianco FJ Jr, et al. Matrix metalloproteinase-9 expression in bladder washes from bladder cancer patients predicts pathological stage and grade. Clin Cancer Res 1998;4:3011–3016.
14. Lebret T, et al. Recurrence, progression and success in stage Ta grade 3 bladder tumors treated with low dose bacillus Calmette-Guerin instillations. J Urology 2000;163:63–67.
15. Hutmacher DW. Scaffold design and fabrication technologies for engineering tissues–state of the art and future perspectives [Review]. J Biomater Sci Polym Ed 2001;12(1):107–124.
16. Sander EA, et al. Solvent effects on the microstructure and properties of 75/25 poly(D,L-lactide-co-glycolide) tissue scaffolds. J Biomed Mater Res Part A 2004;70A(3):506–513.
17. Holy CE, Fialkov JA, Davies JE, Shoichet MS. Use of a biomimetic strategy to engineer bone. J Biomed Mater Res 2003;65A:447–453.
18. Peters MC, Mooney DJ. Synthetic extracellular matrices for cell transplantation. Mater Sci Forum 1997;250:43–52.
19. Gomes ME, et al. Effect of flow perfusion on the osteogenic differentiation of bone marrow stromal cells cultured on starch-based three dimensional scaffolds. J Biomed Mater Res 2003;67A:87–95.
20. Kasemo B, Gold J. Implant surfaces and interface processes [Review]. Adv Dental Res 1999;13:8–20.

21. Vogler EA. Structure and reactivity of water at biomaterial surfaces. Adv Colloid Interface Sci 1998;74:69–117.
22. Webster TJ, Hellenmeyer EL, Price RL. Increased osteoblast functions on theta+delta nanofiber alumina. Biomaterials 2005;26(9):953–960.
23. Horbett TA. Proteins: structure, properties and adsorption to surfaces. In: Ratner BD, Hoffman AS, Schoen AS, Lemmons JE, editors. Biomaterials Science: An Introduction to Materials in Medicine. New York: Academic Press; 1996. p 133.
24. Horbett TA. Chapter 13 Principles underlying the role of adsorbed plasma proteins in blood interactions with foreign materials. Cardiovas Pathol 1993;2(137S):137–148.
25. Hlady V, Buijs J. Protein adsorption on solid surfaces. Curr Opin Biotechnol 1996;7:72–77.
26. Norde W. Driving forces for protein adsorption at solid surfaces. Macromol Symp 1996;103:5–18.
27. Webster TJ, et al. Enhanced functions of osteoblasts on nanophase ceramics. J Biomed Mater Res 2000;51:475.
28. Horbett TA. Techniques for protein adsorption studies. In: Williams DF, editor. Techniques of Biocompatibility Testing, Boca Raton, FL: CRC Press; 1986. p 183.
29. Price RL, Haberstroh KM, Webster TJ. Improved osteoblast viability in the presence of smaller nanometer dimensions carbon fibres. Nanotechnology 2005;15(8):892–900.
30. Kramer RH, Enenstein J, Waleh NS. Integrin structure and ligand specificity in cell matrix interactions. In: Rohrbach DJ, Timpl R, editors. Molecular and Cellular Aspects of Basement Membranes, New York: Academic Press; 1993. p 239–258.
31. Hynes RO. Integrins: versatility, modulation, and signaling in cell adhesion. Cell 1992;69:11–25.
32. Schwartz MA. Transmembrane signaling by integrins. Trends Cell Biol 1992;2:304–308.
33. Puleo DA, Bizios R. Mechanisms of fibronectin-mediated attachment of osteoblasts to substrates *In vitro*. Bone Mineral 1992;18:215–226.
34. Dalton BA, et al. Polymer surface chemistry and bone cell migration. J Biomater Sci Polym Ed 1998;9(8):781–799.
35. Schakenraad JM. Cell: their surfaces and interactions with materials. In: Ratner BD, Hoffman AS, Schoen AS, Lemmons JE, editors. Biomaterials Science: An Introduction to Materials in Medicine, New York: Academic Press; 1996. p 141.
36. Webster TJ. Nanophase ceramics: the future orthopedic and dental implant material. In: Ying JY, editor. Advances in Chemical Engineering, Vol. 27, New York: Academic Press; 2001. p 125.
37. Sinha RK, Tuan RS. Regulation of human osteoblast integrin expression by orthopedic implant materials. Bone 1996; 18(5):451–457.
38. Davies JE. The importance and measurement of surface charge species in cell behavior at the biomaterial interface. In: Ratner BD, editor. Surface Characterization of Biomaterials: Progress in Biomedical Engineering, Vol. 6, New York: Elsevier; 1988. p 219.
39. Brunette PM. The effect of surface topography of cell migration and adhesion. In: Ratner BD, editor. Surface Characterization of Biomaterials: Progress in Biomedical Engineering, Vol. 6, New York: Elsevier; 1988.p 203.
40. Luck M, et al. Analysis of plasma protein adsorption on polymeric nanoparticles with different surface characteristics. J Biomed Mater Res 1998;39:478–485.
41. Degasne I, et al. Effects of roughness, fibronectin and vitronectin on attachment, spreading, and proliferation of human osteoblast-like cells (Saos-2) on titanium surfaces. Calcif Tissue Int 1999;64(6):499–507.
42. Park GE, Pattison MA, Park K, Webster TJ. Accelerated chondrocyte functions on NaOH-treated PLGA scaffolds. Biomaterials 2005;26(16):3075–3082.
43. Miller DC, Haberstroh KM, Webster TJ. Mechanism(s) of increased vascular cell adhesion on nanostructured poly(lactic-co-glycolic acid) films, J Biomed Mat Res 2005;73(4):476–484.
44. Thapa A, Miller DC, Webster TJ, Haberstroh KM. Nanostructured polymers enhance bladder smooth muscle cell function. Biomaterials 2003;24(17):2915–2926.
45. Agrawal CM, Ray RB. Biodegradable polymer scaffolds for musculoskeletal tissue engineering. J Biomed Mater Res 2001;55:141–150.
46. Carver SE, Heath CA. Influence of intermittent pressure, fluid flow, and mixing on the regenerative properties of articular chondrocytes. Biotechnol Bioeng 1999;65:274–281.
47. Yaszemski MJ, et al. Evolution of bone transplantation: molecular, cellular, and tissue strategies to engineer scaffold human bone. Biomaterials 1995;17:175–185.
48. Hutmacher DW. Scaffolds in tissue engineering bone and cartilage. Biomaterials 2000;21:2925–2943.
49. Goulet RW, et al. The relationship between the structural and orthogonal compressive properties of trabecular bone. J Biomech 1994;27:375–389.
50. Kwon IK, Kidoaki S, Matsuda T. Electrospun nano- to microfiber fabrics made of biodegradable copolyesters: structural characteristics, mechanical properties and cell adhesion potential. Biomaterials 2005;26(18):3929–3939.
51. Maroudas NG. Chemical and mechanical requirements for fibroblast adhesion, Nature (London) 1973;244:363–364.
52. Pelham RJ, Wang YL. Cell locomotion and focal adhesions are regulated by substrate flexibility. Proc Natl Acad Sci USA 1997;94:13661–13665.
53. Katz BZ, et al. Physical state of the extracellular matrix regulates the structure and molecular composition of cellmatrix adhesions. Mol Biol Cell 2000;11:1047–1060.
54. Ohya S, Kidoaki S, Matsuda T. Poly(*N*-isopropylacrylamide) (PNIPAM)-grafted gelatin hydrogel surfaces: interrelationship between microscopic structure and mechanical property of surface regions and cell adhesiveness. Biomaterials 2005;26:3105–3111.
55. Bignon A, et al. Effect of micro- and macroporosity of bone substitutes on their mechanical properties and cellular response. J Mater Sci Mat Med 2003;14:1089–1897.
56. Wilkinson CDW, et al. The use of materials patterned on a nano- and micro-metric scale in cellular engineering. Mater Sci Eng 2001;19:263.
57. Clark P, et al. Topographical control of cell behaviour. II. Multiple grooved substrata. Development 1999;108:635.
58. Tranquillo RT. Self-organisation of tissue equivalents: the nature and role of contact guidance. Biochem Soc Symp 1999;65: 27.
59. Li SM, Garreau H, Vert M. Structure-property relationships in the case of the degradation of massive poly(hydroxyl acid) in aqueous media,part 2:degradation of lactide–glycolide copolymers. J Mater Sci Mater Med 1990;1:131–139.
60. Grizzi I, Garreau H, Li S, Vert M. Hydrolytic degradation of devices based on poly(d,l-lactic acid): size dependence. Biomaterials 1995;16:305–311.
61. Sung HJ, Meredith C, Johnson C, Galis ZS. The effect of scaffold degradation rate on three-dimensional cell growth and angiogenesis. Biomaterials 2004;25:5735–5742.
62. Hedberg EL, et al. *In vivo* degradation of porous poly(propylene fumarate)/poly(DL-lactic-co-glycolic acid) composite scaffolds. Biomaterials 2005;26:4616–4623.
63. Perugini P, et al. PLGA microspheres for oral osteopenia treatment: preliminary "*In vitro*"/"*In vivo*" evaluation. Int J Pharm 2003;256:153–160.
64. Wu L, Ding J. *In vitro* degradation of three-dimensional porous poly(D,L-lactide-co-glycolide) scaffolds for tissue engineering, Biomaterials 2004;25:5821–5830.

65. Zhang R, Ma PX. Processing of polymer scaffolds: Phase separation. In: Atala A, Lanza R, editors. Methods of Tissue Engineering, San Diego, CA: Academic Press; 2001. p 715–724.
66. Ishaug SL, et al. Bone formation by three-dimensional stromal osteoblast culture in biodegradable polymer scaffolds. J Biomed Mater Res 1997;36:17–28.
67. Ishaug-Riley SL, et al. Ectopic bone formation by marrow stromal osteoblast transplantation using poly(DL-lactic-co-glycolic acid) foams implanted into the rat mesentery. J Biomed Mater Res 1997;36:1–8.
68. Sherwood JK, et al. A three-dimensional osteochondral composite scaffold for articular cartilage repair. Biomaterials 2002;23:4739–4751.
69. Athanasiou KA, Schmitz JP, Agrawal CM. The effects of porosity on in vitro degradation of polylactic acid-polyglycolic acid implants used in repair of articular cartilage. Tissue Eng 1998;4:53–63.
70. Agrawal CM, McKinney JS, Lanctot D, Athanasiou A. Effects of fluid flow on the in vitro degradation kinetics of biodegradable scaffolds for tissue engineering. Biomaterials 2000;21:2443–2452.
71. Lightfoot EN. Transport phenomena and living systems. New York: John Wiley & Sons, Inc.; 1974.
72. Colton CK. Implantable biohybrid artificial organs. Cell Transplant 1995;4:415–436.
73. Hulbert SF, et al. Potential of ceramic materials as permanently implantable skeletal prostheses. J Biomed Mater Res 1970;4(3):433–456.
74. Yuan H, et al. A preliminary study on osteoinduction of two kinds of calcium phosphate ceramics. Biomaterials 1999;20(19):1799–1806.
75. Turner S, et al. Cell attachment on silicon nanostructures. J Vas Sci Technol B 1997;15:2848–2854.
76. Pattison MA, Wurster S, Webster TJ, Haberstroh KM. Three-dimensional, nano-structured PLGA scaffolds for bladder tissue replacement applications. Biomaterials 2005;26(15): 2491–2500.
77. Gibson LJ, Ashby MF. Cellular Solids: Structure and Properties. 2nd ed. Cambridge University Press; 1997.
78. Ma PX, Langer R. Fabrication of Biodegradable Polymer foams for cell transplantation and tissue engineering. In: Yarmush M, Morgen J, editors. Tissue Engineering Methods and Protocols, Totowa, NJ: Humana Press; 1998. p 47–56.
79. Ma Z, Gao C, Gong Y, Shen J. Paraffin spheres as porogen to fabricate poly(L-lactic acid) scaffolds with improved cytocompatibility for cartilage tissue engineering. J Biomed Mater Res Part B 2003;67(1):610–617. Mikos AG, et al. Preparation and characterization of Poly(L-lactic acid) foams. *Polymer* 1994;35:1068–1077.
80. Liu X, Ma PX. Polymeric scaffolds for bone tissue engineering [Review]. Ann Biomed Eng 2004;32(3):477–486.
81. Harris LD, Kim BS, Mooney DJ. Open pore biodegradable matrices formed with gas foaming. J Biomed Mater Res 1998;42:396–402.
82. Mooney DJ, et al. Novel approach to fabricate porous sponges of poly(D,L-lactic-co-glycolic acid) without the use of organic solvents. Biomaterials 1996;17:1417–1422.
83. Nam YS, Park TG. Porous biodegradable polymeric scaffolds prepared by thermally induced phase separation. J Biomed Mater Res 1999;47:8–17.
84. Whang K, Thomas CH, Healy KE, Nuber G. A novel method to fabricate bioabsorbable scaffolds. Polymer 1995;36(4):837–842.
85. Ho MH, et al. Preparation of porous scaffolds by using freeze-extraction and freeze-gelation methods. Biomaterials 2004;25:129–138.
86. Matthews JA, Wnek GE, Simpson DG, Bowlin GL. Electrospinning of collagen nanofibers. Biomacromolecules 2002;3: 232–238.
87. Reneker DH, Chun I. Nanometre diameter fibres of polymer, produced by electrospinning. Nanotechnology 1996;7:216–223.
88. Giordano RA, et al. Mechanical properties of dense polylactic acid structures fabricated by three dimensional printing. J Biomater Sci-Polym Ed 1996;8:63–75.
89. Ma J, Wang C, Peng KW. Electrophoretic deposition of porous hydroxyapatite scaffold. Biomaterials 2003;24(20):3505–3510.
90. Chang, et al. 2005.
91. Popat KC, et al. Influence of nanoporous alumina membranes on long-term osteoblast response. Biomaterials 2005;26(22): 4516–4522.
92. Shimko DA, et al. Effect of porosity on the fluid flow characteristics and mechanical properties of tantalum scaffolds. J Biomed Mater Res Part B Appl Biomater 2005;73(2):315–324.
93. Goodman SL, Sims PA, Albrecht RM. Related Articles, Links Three-dimensional extracellular matrix textured biomaterials. Biomaterials. 1996;17(21):2087-2095.
94. Ayad S, et al. The extracellular matrix factsbook, San Diego, CA: Academic Press; 1994.
95. Baraton MI, Chen X, Gonsalves KE. FTIR study of a nanostructured aluminum nitride powder surface: Determination of the acidic/basic sites by CO, CO_2 and acetic acid adsorptions. Nanostruct Mater 1997;8(4):435–445.
96. Klabunde KJ, et al. Nanocrystals as stoichiometric reagents with unique surface chemistry. J Phys Chem 1996; 100:12142–12153.
97. Gutwein LG, Webster TJ. Increased viable osteoblast density in the presence of nanophase compared to conventional alumina and titania particles. Biomaterials 2004;25(18):4175–4183.
98. Webster TJ, Ejiofor JU. Increased osteoblast adhesion on nanophase metals: Ti, Ti6Al4V, and CoCrMo. Biomaterials 2004;25(19):4731–4739.
99. Price RL, et al. Osteoblast function on nanophase alumina materials: Influence of chemistry, phase, and topography, J Biomed Mater Res 2004;67(4):1284–1293.
100. Chun AI, Moralez JG, Fenniri H, Webster TJ. Helical rosette nanotubes: a more effective orthopaedic implant material. Nanotechnology 2004;15(4):S234–S239.
101. McKenzie JL, Waid MC, Shi R, Webster TJ. Decreased functions of astrocytes on carbon nanofiber materials. Biomaterials 2004;25(7–8):1309–1317.

See also BIOMATERIALS: TISSUE ENGINEERING AND SCAFFOLDS; ORTHOPEDICS, PROSTHESIS FIXATION FOR; VASCULAR GRAFT PROSTHESIS.

POSITRON EMISSION TOMOGRAPHY

GEORGE KONTAXAKIS
Universidad Politécnica de Madrid
Madrid, Spain

INTRODUCTION: FROM MEDICAL TO MOLECULAR IMAGING

Medical imaging conventionally refers to the non invasive or minimally invasive techniques employed to view internal organs of the body, typically for diagnosing disease. In a broader sense, it refers to a field that enables acquisition, processing, analysis, transmission, storage, display, and

archiving of images of internal body parts for interpretation and patient management (diagnosis, disease staging and evaluation, treatment planning and follow-up). Medical imaging was practically born with the discovery of the X rays by W. C. Roentgen in 1895 and has since based its success on observation and the accumulated experience of the examining physician.

Molecular imaging is a natural out grown of the medical imaging field. Recent advances in molecular biology have resulted in an improved understanding of many disease and natural processes. Consequently, molecular imaging links the empirical diagnostics and experimentally tried treatment management protocols with the fundamental understanding of the underlying processes that generate the observed results. As discoveries of the molecular basis of disease unfold, one top research priority is the development of imaging techniques to assess the molecular basis of cell dysfunction and of novel molecular therapy. Molecular imaging techniques are ideally based on technologies that have an intrinsically high resolution (spatial and temporal) and allow the detection of low concentrations of target biomolecules involved, such as nuclear medicine imaging (Positron Emission Tomography, PET; Single-Photon Emission Tomography, SPET), magnetic resonance imaging (MRI) and spectroscopy (MRS), optical tomography, autoradiography, or acoustical imaging.

The examination of biochemical processes with an imaging technology is of vital importance for modern medicine. As, in most cases, the location and extent of a disease is unknown, the first objective is an efficient means of searching throughout the body to determine its location. Imaging is an extremely efficient process for accomplishing this aim, because data are presented in pictorial form to the most efficient human sensory system for search, identification, and interpretation: the visual system. Recognition depends on the type of information in the image, both in terms of interpreting what it means and how sensitive it is to identifying the presence of disease.

PET stands in the forefront of molecular imaging and allows the quantitative evaluation of the distribution of several pharmaceuticals in a target area *in vivo*. PET is a unique diagnostic imaging technique for measuring the metabolic activity of cells in the human body. It produces images of the body's basic biochemistry and biological activity in a noninvasive way, combining techniques applied in nuclear medicine with the precise localization achieved by computerized image reconstruction. PET is therefore a powerful diagnostic test that is having a major impact on the diagnosis and treatment of disease, as well as on patient management.

PET images can demonstrate pathological changes and detect and stage tumors long before they would be revealed with other conventional imaging modalities. Traditional diagnostic techniques, such as X rays, computerized tomography (CT) scans, or MRI, produce anatomical images of what the internal organs look like. The premise with these techniques is that a visible structural change exists in anatomy caused by disease. However, biochemical processes are also altered with disease and may occur before a change in gross anatomy occurs. Furthermore, PET can provide medical doctors with important early information about very subtle changes of function in the brain and heart, due to disease-related modifications in tissue perfusion, cell metabolic rates heart disease, or neurological disorders (Alzheimer's, Parkinson's, epilepsy, dementia, etc.), allowing physicians to diagnose and treat these diseases earlier and, consequently, more efficiently and accurately, according to the axiom "the earlier the diagnosis, the better chance for treatment." PET can also help physicians monitor a patient's response to treatment, as well as identify distant metastases that can affect treatment, helping curtail ineffective treatments and reduce unnecessary invasive procedures. The field of PET has been emerging today into clinical diagnostic medicine and is approved by many insurance carriers for coverage.

HISTORY OF PET

The positron emission and detection of the radiation produced was a known technique that dates back to the early days of the twentieth century. However, it is only in the last few decades, with the booming development of fast electronic circuits and powerful computer systems, that this knowledge could be used in practice as a valuable diagnostic tool: The electronic circuits used in PET should be able to detect the coincidental arrival of two high energy photons (a timing resolution of the order of few nanoseconds), and the image reconstruction requires modern computer systems in order to produce an accurate image of the activity distribution within a clinically reasonable time.

In the beginning of the 1950s, researchers at the Massachusetts General Hospital (MGH) in Boston and the Duke University in Durham proposed the idea that, in spite of the short half-lives of the, by that time recently discovered, positron-emitting radionuclides, they offered an attractive method for the regional study of metabolism due to their commonality. A single-detector pair brain probe was then developed at MGH and used in experiments. However, it was not until the early 1960s that these positron-emitting radionuclides began to gain popularity, when a number of centers such as the MGH in Boston, the Sloan Kettering Institute in New York, Ohio State University, and the University of California at Berkeley began to use cyclotrons. At the same time, the first image reconstruction techniques were proposed by researchers at MGH, and, in the early 1970s, the concept of computerized tomography (CT) was presented by Hounsfield, who later was awarded with the Nobel Prize.

In the early 1970s, the first PET scanners were developed at the MGH, the Brookhaven National Laboratory, the Washington University, and the Montreal Neurological Institute in Canada, used then as research tools. At the same time, a private company (EG&G OTREC, Oak Ridge, TN, USA) got involved in the developments of the first ring PET scanners, joined in the market a couple of years later by TCC (The Cyclotron Corporation, Berkeley, CA, USA), and in 1976 the first commercial PET scanner was delivered at the University of California, Los Angeles (UCLA). A year later, Scanditronix from Sweden brought Europe into PET. The first PET scanners used single slices when

performing tomographs, with transaxial resolution greater than 2 cm full-width half-maximum (FWHM) and used NaI crystal material. Such systems were installed at several research institutions, apart from the ones mentioned above, like the University of California at Berkeley, the Lawrence Berkeley Laboratory, and the University of Pennsylvania.

By the end of 1970s, PET had shown its potential for application to clinical medicine. The following generation of PET scanners reduced detector size and added additional rings to allow for simultaneous acquisition of multiple slices. The slice resolutions improved from greater than 2 cm FWHM to less than 1 cm FWHM. As time progressed, more detectors and photomultiplier tubes (PMTs) were added to these machines to increase their sensitivity and resolution. In the mid-1980s, the first BGO pixelated detector blocks were presented. At the same time, the first dedicated medical PET cyclotron units with automated radiopharmaceutical delivery systems were commercially available.

At the end of 1980s, the major medical imaging companies (mainly Siemens with CTI PET, Inc., and General Electric with Scanditronix) began investing in PET. The first whole-body PET scanners have been presented and research in new detector materials led to significant discoveries (LSO, etc.) in the beginning of the 1990s. Since then, PET has shown a steady increase in acceptance for clinical application, both medically and administratively, and PET centers are being installed worldwide at an increasing pace. PET is now a well-established medical imaging technique that assists in the diagnosis and management of many diseases.

More details on the history of PET instrumentation and the related developments can be found in References 1 and 2.

PHYSICAL PRINCIPLES OF PET

PET images molecules of substances with a specific biological activity. In order to monitor their distribution, kinetic characteristics, and behavior of (pharmaceuticals) within the body, these substances are tagged with radioactive compounds (with short half-life and at extremely low concentrations) (3). These radiopharmaceuticals are chosen to have a desired biological activity, depending on the metabolic activity of the organ under study, and are introduced to the subject by injection or inhalation.

The most commonly used radionuclides are listed in Table 1 and are compounds that constitute, or are consumed by, the living body, like carbon, nitrogen, and oxygen. They are isotopes of biologically significant chemical elements that exist in all living tissues of the body and in almost all nutrients. Therefore, the above radionuclides are easily incorporated in the metabolic process and serve as tracers of the metabolic behavior of the body part, which can be studied *in vivo*.

Table 1. The Most Commonly Used Radionuclides in PET

Radionuclide	Half-life
Carbon-11	20.3 min
Nitrogen-13	9.97 min
Oxygen-15	2.03 min
Fluorine-18	1.83 h
Gallium-68	1.83 h
Rubidium-82	1.26 min

Table 2. Major PET Radiopharmaceuticals and their Specific Medical Applications

Agent	Images
F-18 fluorodeoxyglucose	Regional glucose metabolism
F-18 sodium fluoride	Bone tumors
C-11 methionine	Amino acid uptake/protein synthesis
C-11 choline	Cell membrane proliferation
C-11 deoxyglucose	Regional brain metabolism
O-15 oxygen	Metabolic rate of oxygen use/OEF
C-11 carbon monoxide	Cerebral blood volume
O-15 carbon monoxide	Cerebral blood volume
O-15 water	Cerebral blood flow
O-15 carbon dioxide (Inhaled)	Cerebral blood flow
C-11 butanol	Cerebral blood flow
C-11 N-methylspiperone	Dopamine D2 and Serotonin S2 receptors
F-18 N-methylspiperone	D2 and S2 receptors
C-11 raclopride	D2 receptors
F-18 spiperone	D2 receptors
Br-76 bromospiperone	D2 receptors
C-11 carfentanil	Opiate mu receptors
C-11 flumazenil	Benzodiazepine (GABA) receptors

Table 2 shows a list of the major radiopharmaceuticals used as PET agents with their specific medical applications. The most common radiopharmaceutical used in PET studies today is fluorodeoxyglucose (FDG) (4), a chemical compound similar to glucose, with the difference that one of the -OH groups has been replaced by F-18. Carbon-11 can also be used as a radiotracer to glucose. The short half-lives of these particles allow the subject and the people handling them to receive only a low radiation dose.

The identification and detection of the presence of the molecules of the radiotracer in a specified location within the source (i.e., the body under study) is performed by a chain of events, based on physical principles and data processing techniques, which are schematically depicted in Fig. 1 and briefly described below.

A positron is emitted during the radioactive decay process, annihilates with an electron, and, as a result, a pair of γ rays is emitted (two high energy photons of 511 keV each). The two γ rays fly off in almost opposite directions (according to the momentum conservation laws), penetrate the surrounding tissues, and can be recorded outside the subject's body by scintillation detectors placed on a circular or polygonal detector arrangement, which forms a PET tomograph. When the γ ray hits a scintillation detector material, it then deposits its energy in that crystal by undergoing photoelectric effect, which is an atomic absorption process where an atom totally absorbs the energy of an incident photon (5). This energy is then used to eject an orbital electron (photoelectron) from the atom and is,

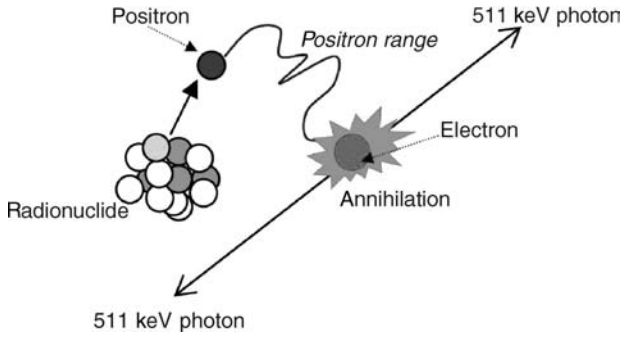


Figure 1. This schematic depicts the chain of events that described the physical properties of high energy gamma pair emission from positron-emitting radioisotopes. All radioisotopes used with PET decay by positron emission. Positrons are positively charged electrons. Positron emission stabilizes the nucleus of unstable radioisotopes by removing a positive charge through the conversion of a proton into a neutron. An emitted positron travels a short distance (*positron range*, which depends on the energy of the positron) and collides with an ordinary electron of a nearby atom in an annihilation reaction. When the two particles annihilate, their mass turns into two 511 keV gamma rays that are emitted at 180° to each other. When detected, the 180° emission of two gamma rays following the disintegration of positronium is called a coincidence line. Coincidence lines provide a unique detection scheme for forming tomographic images with PET.

therefore, transformed in visible light. This light can be detected by specialized devices (photomultiplier tubes, PMT) that capture and transform it into an electronic signal, shaped at a later stage by the electronic circuits of the tomograph to an electronic pulse, which provides information about the timing of the arrival of the incident γ ray and its energy. Figure 2 summarizes the principles of gamma ray event detection in PET described here.

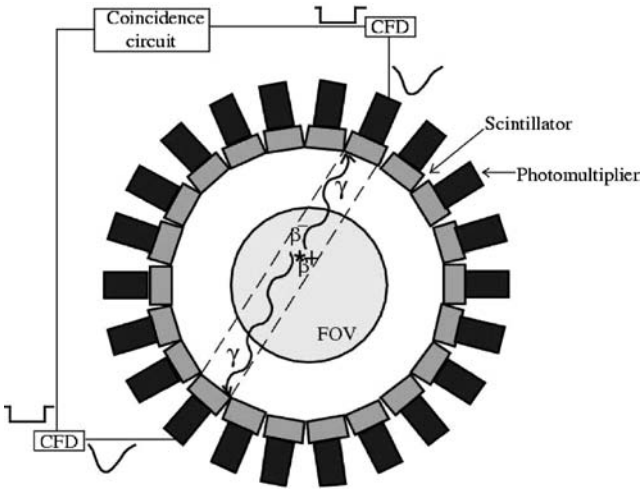


Figure 2. Scintillation detectors coupled to photomultiplier tubes are placed around the detector ring of the scanner. An annihilation event (*) inside the field of view (FOV) produces two γ rays that get detected by a pair of detectors. The event is identified to occur inside a specific detector tube (dashed stripe) by the electronic devices (constant fraction discriminators, CFD, and the coincidence detection circuit) that connect every pair of detectors.

By measuring a coincidence photon, the detector array in a PET system identifies that an annihilation event occurred inside the volume defined between the surfaces of the pair of detectors that registered the coincidence event. At the end of a PET scan, for each pair of detectors, a number of coincidence events that have been identified exist. This information represents the radioactivity in the subject viewed at different angles, when sorted in closely spaced parallel lines. In order to reconstruct the activity density inside the source from its projections (events registered at each detector pair), a mathematical reconstruction algorithm is applied by computer. The collected data are corrected for scatter, attenuation, and accidental coincidences; normalized for the differences in detector efficiencies, and reconstruct the spatial distribution of the radioactivity density inside the organ or the system under study in the form of a 2D or 3D image. The result is a digital image of the source, where the value of each picture element (pixel) or, in modern 3D tomograph systems, volume element (voxel) is proportional to the activity density inside the source at the area (or volume) that corresponds to this pixel/voxel. This image can be directly displayed on a screen. Further analysis of the data and processing of the produced images can be carried out with the use of a computing system.

A high energy photon produced by an annihilation event can deviate from its original trajectory if it gets involved in Compton scattering inside the subject's body, a collision between a photon and a loosely bound outer-shell orbital electron of an atom. In this case, because the incident photon energy greatly exceeds the binding energy of the electron to the atom, the interaction can be considered as a collision between the photon and a "free" electron. The photon does not disappear in Compton scattering, but it is deflected through a scattering angle θ and some of its energy is transferred to the electron (recoil electron) (5). In the case this ray gets detected in coincidence with the second gamma produced at the same event, then this event will be counted to have occurred in a detector tube that will not contain the original annihilation site: This is an erroneous event (scattered event).

It is also possible that this ray will never reach a detector crystal and, therefore, get lost. This type of Compton scattering, along with photoelectric absorption of the produced gamma rays inside the source, where they have been generated, are the major sources of attenuation of the emitted radioactivity.

The physics of positron emission allow for attenuation correction of the collected data, which can produce a quantitatively (but also qualitatively) accurate image that may resolve small lesions, especially when these lie deep within the body. In order to correct for attenuation, two additional measurements are typically performed: the blank scan and the transmission scan. The blank scan is recorded using an external source without the patient, representing the unattenuated case. For the transmission scan, the patient and the bed are placed into the scanner and the attenuated data are measured using the external source. The attenuation correction factors (ACF) can be calculated as the ratio of the measured counts without and with the attenuating object. The disadvantages of attenuation correction are that it

requires more time for image acquisition and the potential exists to add noise to the image if the attenuation measurements become misaligned by patient motion or if inadequate statistics in the transmission scan are collected. As a result of noise, transmission measurements are usually smoothed prior to the division. Otherwise, the noise in the ACF propagates to the corrected emission sinogram. The drawback of smoothing is that the resulting blurring of ACFs propagates to the emission sinogram as well. Techniques for the reduction of noise propagation include, as an example, classification techniques for the main tissue categories observed in the transmission images (segmentation) or the use of iterative methods for the reconstruction of the transmission images (6).

Compton scattering can also occur inside the detector crystal before the ray undergoes (the desirable) photoelectric effect. In that case, it is possible that the ray will escape the detector material and deposit its energy in an adjacent scintillator, causing the detected event to be mispositioned. Another source of erroneously counted events is the coincidental arrival at the detector ring of two single gamma rays coming from two different annihilation events (random or accidental coincidence). When three or more γ rays arrive at the detector ring within the time coincidence window set by the electronic circuitry of the scanner for the coincidence detection, then these gammas must be rejected, because it is not possible to recognize, in that case, the pairs of photons that came from the same annihilation event (7).

The high energy gamma rays have increased penetrating abilities and can be detected coming from deep-lying organs better than α particles or electrons (β particles), which can penetrate only a few millimeters of tissue and, therefore, cannot get outside the body to the radiation detector (5). Imaging system detectors must, therefore, have good detection efficiency for γ rays. It is also desirable that they have energy discrimination capability, so that γ rays that have lost energy by Compton scattering within the body can be rejected and a good timing resolution to accurately measure the time difference of the arrival of two photons. Sodium iodide (NaI), BaF_2 (barium fluoride), and BGO (bismuth germanate oxide) provide both of these features at a reasonable cost (5). Research for new scintillator materials, like LSO (lutetium oxyorthosilicate) (8), GSO (germanate oxide) (9), $PbCO_3$ (lead carbonate) (10), $PbSO_4$ (lead sulfate) (11), CeF_3 (cerium fluoride) (12), YalO (13), and LuAlO (14), is very active in an effort to produce faster detector crystals with good stopping power and light output.

Table 3 summarizes some of the main physical properties of the scintillators used for PET: NaI(Tl), BGO, BaF_2, CsF, GSO, and LSO. In order to interpret this table, assume the following:

- An elevated density guarantees a high stopping power for the high energy 511 keV annihilation photons and consequently assures elevated detection efficiency. High stopping power also allows the use of crystals of small dimensions, which means an improved spatial resolution of the tomograph.
- High scintillation efficiency, due to a good intrinsic energy resolution of the crystal, leads to a good energy resolution of the detection system, which leads to a better discrimination of scatter.
- A fast scintillation (described by a short scintillation constant decay time) translates to a low dead time of the system and, therefore, to good count rate performance. Moreover, this property directly influences the temporal resolution (uncertainty of the moment of detection), on which depends the choice of the length of the time coincidence resolution window and, therefore, the rate of accidental coincidences.

The comparison of the characteristics of scintillation crystals shows that the ideal scintillator for PET must have the temporal characteristics (decay time) of BaF_2, the density (stopping power) of BGO, and the scintillation efficiency (light output) of NaI(Tl) (15). It also reveals that the newest crystals GSO and LSO are very promising for PET applications.

Originally, NaI was the detector of choice for nuclear medicine imaging cameras and is still in use by some manufacturers of gamma cameras, SPET, and even PET systems. NaI is a scintillation crystal discovered in 1949 with very high scintillation efficiency but a stopping power too low for high energy photons; therefore, NaI has very low sensitivity. In the 1980s, BGO emerged as the detector of choice for PET scanners, a material with considerably lower light output than NaI but, on the other side, twice as dense and, therefore, able to detect high energy photons more effectively. LSO was discovered in the early 1990s and exhibits a very fast scintillation time (40 ns), which provides significantly reduced detector dead time and consequently higher count-rate capabilities, which is essential in clinical PET imaging in order to use the injected activity most efficiently and to make the emission scan time as short as possible, meaning the patient spends less time immobile on the tomograph's bed without compromising the image quality.

In the optimization of the design of a PET tomograph, an important aspect is the way crystals are assembled and the

Table 3. Scintillation Crystal Characteristics

	NaI	BGO	BaF_2	CsF	LSO	GSO
Density (g/cm^3)	3.67	7.13	4.87	4.64	7.40	6.71
Relative scintillation efficiency	100	20	16	6	75	30
Decay constant (ns)	250	300	0.6	2.5	40	60
Hygroscopic	Yes	No	No	Yes	No	No

way they are coupled to the photomultiplier tubes. Various strategies have been developed, including:

- one-to-one connection crystal-PMT (5);
- detector blocks, where a crystal array (mainly BGO or LSO) is coupled to a smaller number of PMTs (15,16);
- NaI(Tl) crystals of large dimensions coupled to a grid of PMT (Anger logic, common to gamma cameras) (17);
- the most recent design of a system of GSO crystals coupled to light guides to a PMT grid (18).

Scintillation detectors have been the dominant element in high energy gamma ray detection for PET. However, other technologies have also been applied, explored, and developed for this purpose. One of the oldest alterantive technologies is the High Density Avalanche Chamber (HIDAC) PET system (19), which consists of a Multiwire Proportional Chamber (MWPC) with the provision of laminated cathodes containing interleaved lead and insulating sheets and mechanically drilled with a dense matrix of small holes. Ionization resulting from photons interacting with the lead is trapped by, amplified in, and extracted from, the holes by a strong electric field into the MWPC. On arrival at an anode wire, further avalanching occurs. Coordinate readout may be obtained from orthogonal strips on the cathodes. The result is precise, 2D localization of the incident gamma rays. Every hole on the cathodes acts as an independent counter. By stacking these MWPCs, millions of these counters are integrated to form a large-area radiation camera with a high spatial resolution.

The resolution of a PET scanner primarily depends on the size of the detectors and on the range of positrons in matter (distance traveled by the positron in the tissue before interacting with a free electron, see also Fig. 1). For most of the positron emitters, the maximum range is 2–20 mm. However, the effect on spatial resolution is much smaller, because positrons are emitted with a spectrum of energies and only a small fraction travel the maximum range, and, in addition, in case of 2D acquisitions, the range of the third dimension is compressed. Another limitation in the resolution is that the paired annihilation photons are not emitted precisely 180° from each other, because the e^+–e^- system is not at complete rest. Other components of the system resolution are the sampling scheme used, the interactions between more than one crystal due to intercrystal scatter, the penetration of annihilation photons from off-axis sources to the detector crystals, the reconstruction technique used, the filters applied, and the organ and patient motion during the scan.

Three types of spatial resolution exist in a typical ring PET system, defined by a full-width at half-maximum (FWHM): the radial, tangential, and axial resolutions. The radial, or in-slice, resolution deteriorates as we move from the center of the FOV and is best at the center. The same happens for the tangential resolution, which is measured along a line vertical to a radial line, at different radial distances. In systems with more than one detector ring, the axial resolution, or slice thickness, is measured along the axis of the tomograph.

A major source of error during the coincidence detection is the fact that not all the annihilation events are registered correctly as mentioned earlier. Additional accidental coincidences can result from poor shielding or backscatter and from ordinary γ rays from the radionuclide administered. The random and scattered coincidences are registered together with the true coincidences, obtained when a pair of gammas is correctly identified and classified to the appropriate detector tube, and are sources of background noise and image distortion.

In order to keep the number of scattered coincidences low, a discriminator should be used. A discriminator primarily generates timing pulses upon the arrival of a photon, but also can verify the total energy of the illuminating ray is above a preset energy threshold. Scattered rays have already deposited part of their energy and, therefore, can be identified.

Furthermore, the choice of the appropriate time coincidence (or coincidence resolving time) window is essential: It has to be narrow enough to keep the number of random coincidences as low as possible but also wide enough to include all valid coincidence pulses. In the existing PET units, the timing accuracy is of the order of tenths of nanoseconds.

A PET scanner can be designed to image one single organ, such as the brain or the heart, or can be able to image any organ in the body, including whole-body scans. Whole-body studies with F-18-FDG consist of repeated PET acquisitions at contiguous bed positions in order to provide 3D images (axial, sagittal, coronal, and oblique cut planes) covering one considerable portion of the patient's body (Fig. 3), which facilitates the search for metastases in oncological diagnostics (20).

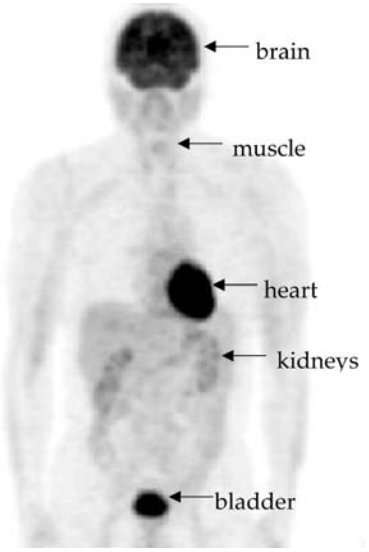


Figure 3. A whole-body F-18-FDG PET image of a normal subject (no pathological situation diagnosed). Areas with high metabolic activity (brain, myocardium) or with high concentration of the radioactive tracer (bladder) are visible. [Courtesy of A. Maldonado and M.A. Pozo from the Centro PET Complutense, Madrid, Spain.]

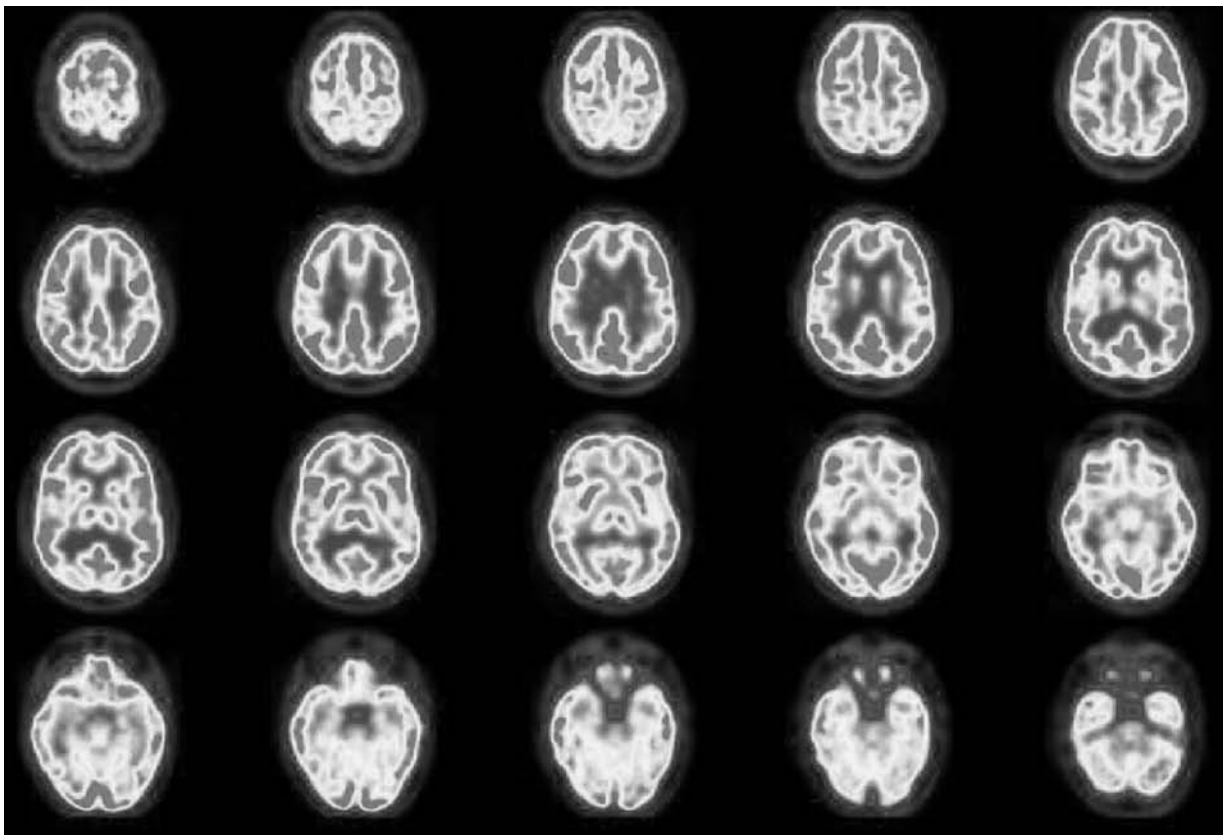


Figure 4. Sequential images from an F-18-FDG PET brain study of a normal individual. Red-yellow areas correspond to the high metabolic activity in the gray matter (cortex). [Courtesy of A. Maldonado and M.A. Pozo from the Centro PET Complutense, Madrid, Spain.]

Most PET systems today are whole-body systems (i.e., they have a typical transaxial FOV of 60 cm). This FOV is adequate to handle most patients. The axial FOV of most PET systems today is limited to approximately 10–15 cm (21). This relatively narrow axial FOV imposes some limitation on the imaging procedures that can be performed clinically. It also requires more accurate positioning of the patient in comparison with conventional nuclear medicine procedures. For a clinical system, it would be desirable to extend the axial FOV to 15–20 cm, which would, for instance, allow full brain (Fig. 4) and heart imaging in a single frame and more efficient whole-body imaging. As the detectors contribute a significant portion of the total cost of the scanner, however, this would bring into question what would be an acceptable cost for the PET scanner.

MANUFACTURING OF RADIOPHARMACEUTICALS

A cyclotron is a particle accelerator that produces positron-emitting elements or short-lived radioisotopes. These radioisotopes can then be incorporated into other chemical compounds that are synthesized into a final product that can be injected into a person. These radioisotopes are used to "label" compounds so it can later be identified where in the body the radiopharmaceutical is being distributed. The compounds that are being labeled are organic molecules normally used in the body, such as sugar, neurotransmitters, and so on (22).

First, the cyclotron bombards nonradioactive elements in the target with accelerated particles, which converts these elements into positron-emitting radioactive isotopes of fluorine, nitrogen, oxygen, or carbon. The major radioactive isotope produced at almost all sites is fluorine-18 (F-18), which has a half-life of 110 min. F-18 thus produced from the cyclotron is delivered to a chemical synthesis unit called the chemical processing unit, which is where F-18 is incorporated into a precursor to produce the final product FDG, the labeled sugar molecule. This entire process is fully automated and performed in the cyclotron lab. When a dose is needed, it is transported to the PET scan room by various means, depending on the distance between the production site and the PET tomograph and ranging from a dedicated pneumatic tube system to long-distance transport via air or road.

APPLICATIONS OF PET

Molecular imaging opens the way for medical doctors to successfully pursue the origin of disease. As long as disease is of unknown origin, more tests and exams are needed, something that means increased health-care costs, in addition to the patient's discomfort and pain. PET can

accurately identify the source of many of the most common cancers, heart diseases, and neurological diseases, eliminating the need for redundant tests, exploratory surgeries, and drug overload of the patient. PET produces powerful images of the body's biological functions and reveals the mysteries of health and disease (23).

PET can be used to obtain information about the tissue perfusion using inert tracers (e.g., O-15 labeled water), the metabolism with metabolically active tracers (e.g., F-18-FDG), or the kinetic of a cytostatic drug (e.g., F-18-Fluorouracil).

In cardiology (22), this imaging technique represents the most accurate test to reveal coronary artery disease or rule out its presence. Traditionally, when a patient shows signs or symptoms of heart disease, his or her physician will prescribe a thallium stress test as the initial diagnostic study. The conventional thallium stress test, however, is often not as accurate as a PET scan. PET images can show inadequate blood flow to the heart during stress that can pass undetected by other noninvasive cardiac tests. A PET study could enable patients to avoid cardiac catheterization when a conventional perfusion or echocardiographic stress test is equivocal. A PET scan shows myocardial viability in addition to perfusion abnormality. More specifically, PET exams for metabolism and perfusion of the heart tissues can determine the need for heart transplant, in case both are absent in a large area of the heart, or confirm with certainty that simple bypass surgery would be enough, when metabolism is maintained even if blood flow is significantly reduced. As metabolism indicates that tissue is still alive, complicated heart transplantation can be avoided and coronary bypass would have great chances to improve cardiac function. Documented studies have shown that thallium stress testing overestimates irreversible myocardial damage in at least 30% of cases, which can result in the patient being placed on the transplant list rather than receiving bypass surgery or angioplasty. No other diagnostic test can more precisely assess myocardial viability than PET. The most recent developments in cardiac PET have been summarized in Reference 3.

PET can reveal abnormal patterns in the brain and is, therefore, a valuable tool for assessing patients with various forms of dementia (3,22). PET images of the brain can detect Parkinson's disease: A labeled aminoacid (F-DOPA) is used as tracer at a PET examination in order to determine if the brain has a deficiency in dopamine synthesis. If it does not, Parkinson's disease can be ruled out and possible tremors in the patient's muscles will be treated in a different manner. Although the only definitive test for Alzheimer's disease (AD) is autopsy, PET can supply important diagnostic information. When comparing a normal brain versus an AD-affected brain on a PET scan, a distinctive and very consistent image pattern appears in the area of the AD-affected brain, where certain brain regions have low metabolism at the early stages of the disease, allowing early detection several years before diagnosis can be confirmed by a physician. PET can also help to differentiate Alzheimer's from other confounding types of dementia or depression (29). Conventionally, the confirmation of AD was a long process of elimination that averaged between two and three years of diagnostic and cognitive testing. PET can help to shorten this process by identifying distinctive patterns earlier in the course of the disease. Furthermore, PET allows the accurate identification of epileptogenic brain tissue (because of its reduced glucose metabolic rates) and can successfully lead the surgical removal of the epileptic foci.

In oncology (3,22), in which the clear majority the total PET examinations refer, this technique inspects all organs and systems of the body to search for cancer in a single examination. PET is very accurate in distinguishing malignant tumors from benign growths. It can help detect recurrent brain tumors and tumors of the lung, breast, lymph nodes, skin, colon, and other organs. The information obtained from PET studies can be used to determine what combination of treatment is most likely to be successful in treating a patient's tumor, as it can efficiently determine the resistance of a specific cancer to the drugs applied and, consequently, can dynamically optimize the treatment management and follow-up of the patient on an individual basis. With this technique, it is possible to evaluate if a tumor has been successfully destroyed after therapy, as anatomical follow-up imaging is often not in the position to assess if a residue is still active or has definitely been eliminated after chemotherapy, radiation, or surgery.

A summary of the current status and future aspects of PET for cancer detection, as it has been recently presented by the Health Technology Advisory Committee is as follows (23):

Brain Cancer: F-18-FDG PET in brain tumor imaging may be useful, but its clinical application has yet to be established. F-18-FDG PET does not appear to be able to define tumor histology. Additional studies are warranted regarding the value of F-18-FDG PET in detecting Central Nervous System (CNS) and nonCNS brain metastasis, differentiating malignant from nonmalignant lesions, detecting disease recurrence in subjects who have undergone intensive radiotherapy, and in pediatric brain tumors. As a result of the paucity of data on radiotracers other then F-18-FDG, further studies will be required to validate the use of PET brain scanning with these radiotracers.

Head and Neck Cancer: Studies suggest that F-18-FDG PET is superior to MRI but comparable with CT in identifying the presence, absence, or recurrence of cancer.

Pituitary Cancer, Thyroid Cancer, Urinary Cancer, Kidney Cancer: The paucity of data on the use of PET in pituitary tumors, thyroid tumors, urinary cancer, and kidney cancer prevents conclusions regarding its value at this time.

Lung Cancer: Numerous studies evaluating PET for lung cancer applications demonstrate that PET, using F-18-FDG as a radiotracer, is effective and may be more effective than other noninvasive techniques, particularly CT, in differentiating benign and malignant pulmonary lesions. Thus, F-18-FDG PET appears to be an effective means of diagnosing lung cancer, whether a primary disease or a secondary metastatic disease, and detecting disease recurrence following lung cancer therapy.

Breast Cancer: Preliminary data suggest that F-18-FDG PET can differentiate benign from malignant breast lesions, when used in breast cancer staging, and can

determine the presence of axillary node involvement. Although data are scarce regarding the use of PET in monitoring the effects of breast cancer therapy, available data suggest that both F-18-FDG PET and C-11-MET PET may be useful for breast cancer and may show response earlier than conventional methods. Regardless, due to the small study samples and limited amount of available data, further studies will be required to confirm the efficacy of PET for breast cancer imaging.

Esophageal Cancer: F-18-FDG PET may be valuable in the staging of esophageal cancer. Evidence is limited by the small number of subjects in each study and the lack of additional trials.

Pancreatic Cancer: Studies indicate that PET may have a role in the imaging of pancreatic tumors, but further study is needed to verify this indication.

Renal Cancer: F-18-FDG PET shows promise for evaluating renal masses, but confirmation is required.

Ovarian Cancer: Preliminary data suggest a potential role for F-18-FDG PET in ovarian cancer; however, further studies are required to confirm these findings.

Prostate Cancer: Although F-18-FDG PET has been used in certain prostate cancer cases, it is possible that the use of radiotracers other than F-18-FDG may be of more value. However, insufficient data exists at this time to draw conclusions regarding the use of PET in prostate cancer.

Testicular Cancer: With limited data, no conclusions can be made at this time.

Malignant Melanoma: Additional studies are needed to determine the role of PET in the imaging of malignant melanoma.

Colorectal Cancer: F-18-FDG PET may be a valuable tool for colorectal cancer in diagnosis, preoperative staging, and monitoring for recurrent disease or treatment response. However, further study is required to confirm these findings.

Neuroendocrine Gastrointestinal Cancer: PET proved superior to CT in detecting, delineating, and visualizing lesions. The study claimed that PET had a superior role, but further study is required to confirm this finding.

Malignant Lymphoma: Studies comparing F-18-FDG PET with alternative techniques found PET to be more accurate than CT, 99mTc-MIBI SPET, and 111In-somatostatin scintigraphy in detecting untreated and treated lymphoma. Supportive evidence is limited to a few trials that are hampered by small study samples. No conclusions can be drawn at this time regarding the efficacy of PET for malignant lymphoma.

A major use of PET is its ability for kinetic imaging analyses. This term refers to the measurement of tracer uptake over time. An image of tracer activity distribution is a good starting point for obtaining more useful information such as regional blood flow or regional glucose metabolism. The process of taking PET images of radioactivity distribution and then using tracer kinetic modeling to extract useful information is termed image analysis. The tracer kinetic method with radiolabeled compounds is a primary and fundamental principle underlying PET and autoradiography. It has also been essential to the investigation of basic chemical and functional processes in biochemistry, biology, physiology, anatomy, molecular biology, and pharmacology. Tracer kinetic methods also form the basis in *in vivo* imaging studies in nuclear medicine (24).

Besides its direct clinical applications, PET imaging is emerging as a powerful tool for use by the pharmaceutical industry in drug discovery and development (25). The role of small animal PET imaging (26) studies in rodents for the discovery of PET tracers for human use is significant, as it has the potential for permitting higher throughput screening of novel tracers in transgenic mice as well as the confounding effects resulting from potential species differences on receptor affinity, blood-brain barrier (BBB) transport, metabolism, and clearance. This setting is expected to allow new and unique experimental laboratory studies to be performed.

Other recent developments include dedicated mammography devices (known as positron emission mammographs, PEM) for breast functional imaging (27). Furthermore, the first PET/CT tomographs have made their way to the market (28). These are devices that house a positron tomograph and a CT scanner in a single device, allowing the acquisition and visualization of registered images detailing both anatomy and biological processes at the molecular level of internal organs and tissue, without the need of multiple examinations and further image processing to achieve similar results.

IMAGE INTERPRETATION

One of the final steps in the processing chain of the PET study is to produce a final layout of the images for the diagnosing physician. The conventional way of presenting the image data is to produce a transparency film (X-ray film) of the images on the computer display. In addition to the image data, the film should also be labeled with demographic data about the study, such as patient name and scan type. As this information is usually stored in the image files together with the image data, the labeling and layout of the images on the display can be automated in software. With the rapid development of local area networks, films may soon no longer be necessary. Instead, the images can be read from a display system located in the reading room, which has access to the PET image data through a computer network. Referring physicians do, in most cases, require a hard copy of the study, which can be accomplished using X ray films. With recent improvements in printer technology, high quality color output may also be a low cost alternative to the traditional film.

PROCEDURE FOR A PET SCAN

Most patients will be in the PET center for 2 or 3 h, depending on the type of study being conducted. The patient is informed as to when to stop eating before the test. Drinking lots of water is recommended before the scan. The patients also need to inform the PET center if they are diabetic or claustrophobic. In general, before the scan is performed, a catheter is placed in the arm so that the radioactive tracer can be injected. A glucose test will also be performed. Depending on the type of study conducted,

scanning may take place before and after the injection is given. After the tracer is given, the patient waits for approximately 40–60 min before the final scan is done.

PET SCAN AND ASSOCIATED RISKS

The radiation exposure of PET is similar to that of having a CT scan or any other standard nuclear medicine procedure involving heart or lung scans. No pain or discomfort results from the scan. The half-life of F-18 is so short that by the time the patient leaves the PET center, almost no activity remains in the body. Patients typically do not experience any reactions as a result of the PET scan, because the tracer material is processed by the body naturally. Therefore, no side effects are expected. Of course, as with any other nuclear medicine procedure, when breast-feeding or pregnant, a PET scan must be performed under special conditions.

CURRENT STATUS AND FUTURE ASPECTS IN PET INSTRUMENTATION

Technological developments and research in the field of PET instrumentation are currently marking a fast evolution (30). New PET systems have been designed and developed with whole-body scanning capabilities. These systems are clearly designed for oncological studies (currently almost entirely performed in the clinical practice with the use of F-18-FDG), which represent maybe more than 80% of the total PET examinations performed worldwide. Therefore, a clear shift has occurred from the earliest systems, which were then mostly oriented to neurological applications.

The main requirement, which drive current R&D activities both in academia and in industry, is to increase the diagnostic accuracy (lesion detectability) of the technique and, at the same time, decrease the cost of a PET system installation, operation, and maintenance, which would allow the widespread use of PET in the clinical practice. In order to achieve this goal, an optimal balance should be found between high performance specifications and cost efficiency for the newly designed tomographs.

In particular, very high resolution 3D PET imaging (with applications in brain imaging, positron emission mammography, as well as small animal imaging) has demanded further advances in scintillation detector development, image reconstruction, and data correction methodology.

Table 4 lists the major performance characteristics of some last-generation tomographs for human whole-body studies, based on different design architectures and operating in 3D acquisition mode. For 2D acquisitions, lead or tungsten septa are placed between the detectors to absorb scattered radiation (out of slice activity). The septa reduce the amount of scatter to 10–15% of the total counts acquired and improve image contrast. For 3D acquisitions, the septa are removed and each individual detector is sensitive to radiation from a much larger area (30). This mode allows a significant increase of the detection efficiency of the order of a factor 5–6 over the 2D mode and therefore, provides an increase of the SNR in the produced images, an aspect of extreme importance in whole-body studies. 3D PET imaging can, in addition, significantly reduce the amount of tracer activity needed for the exam and shorten the acquisition time, thus reducing the time during which the patient must remain immobile on the tomograph's bed.

A limitation of the 3D mode, however, is an increase of the scatter component (almost one out of every two of the detected events has been scattered in the source or even inside the scintillation detectors) as well as of the number of the detected accidental coincidences (randoms) (30). A good energy resolution is therefore imperative in 3D PET systems, in order to reduce the scatter component (by correctly identifying detected γ rays with deposited energy lower than 511 keV). Furthermore, these systems must be able to manage high count rates in order to match the radioactivity present in the FOV. High temporal resolution in PET (high count rate) also permits dynamic imaging (repeated studies at short time intervals). With high count rates, pulses receiving a detector block can "pile-up" and the detector may become paralyzed, which decreases the sensitivity and detection efficiency of the tomograph. In addition, when scanning in a high counting rate environment, the random counting rate increases much more rapidly than does the true counting rate as a function of radioactivity in and near the FOV. In general, in 3D mode, an increased number of random events is detected, which degrades the image quality. Appropriate scatter and randoms corrections must therefore be applied to 3D-acquired data (31). Considering the nature of the scatter correction process and the heterogeneity of the activity distribution in the thoracic and abdominal areas (which are of particular interest for whole-body F-18-FDG PET studies), the use of scatter correction techniques is not yet consolidated and their effectiveness regarding the quality and quantitative accuracy of whole-body PET studies demands more research work.

The performance of a 3D-enabled PET tomograph is, therefore, the result of a compromise between the various physical parameters considered (spatial resolution,

Table 4. Performance Characteristics of PET Scanners in 3D Mode (15)

	Philips C-PET	GE Advance	ECAT HR+	ECAT Accel	Philips Allegro
Crystal	NaI	BGO	BGO	LSO	GSO
Crystal dimensions (mm)	$500 \times 300 \times 25$	$4.0 \times 8.2 \times 30$	$4.0 \times 4.4 \times 30$	$6.8 \times 6.8 \times 20$	$4.0 \times 6.0 \times 20$
Spatial resolution FWHM, mm (10 cm)	6.4	5.4	5.4	6.7	5.9
Efficiency (kpcs/Ci/cc)	450	1060	900	900	> 800
% Scatter fraction	25	35	36	36	25
50% Dead time (kcps/Ci)	0.2	0.9	0.6	-	0.6

detection efficiency, energy resolution, and linearity of count rate). In the modern design of such systems, the primary objective evolves from the optimization of the spatial resolution and the efficiency (typical of tomographs for cerebral applications) to the optimization of the balance between energy resolution and count rate performance. The size of the scintillation detector crystals, which together with the photomultiplier tubes constitute the main elements in the design of a PET system, determines the intrinsic spatial resolution of the tomograph. The volume of each crystal has a minimum, defined by the current technological limitations, but, at the same time, should be large enough to include a sufficient mass of material so that a significant number of the incident high energy γ rays are absorbed and converted to visible (detectable) light. A very small detector crystal could result transparent to γ rays, which would decrease the system's sensitivity.

Spatial resolution, an area of interest in the design of PET systems, refers to the development and implementation of techniques for the correction of the effect of "depth of interaction" (DOI) parallax error, which limits the uniformity of the spatial resolution in the FOV for PET tomographs with rings of detector block arrays (18). In such systems, the length of the detector crystals is about ten times as long as their width in order to improve detection efficiency. Therefore, PET measurements exhibit shift-variant characteristics, such as broadened sensitivity functions of each detector pair from center to edge of FOV and for oblique lines of response. Spatial uniformity can be restored if the DOI of the incident photons is known. A number of techniques for deriving DOI information from PET detectors have been proposed, including the use of a phoswich technique (32) (detector arrangements, composed from scintillation crystal layers; e.g., LSO/GSO phoswich block detector, where the distinct temporal characteristics of the crystals used allow to identify the DOI), extracting the DOI information by controlling the light-sharing between two crystals, coupling of two ends of the detection crystals to separate photodetectors, and extracting DOI information from a 3D matrix detector (33). Other approaches include the application of a light-absorbing band around each crystal, the introduction of a light-absorbing material between sections of the detector or use of a Multipixel Hybrid Photomultiplier (M-HPD) (34). When fully available in commercial tomographs, the implementation of correction techniques for DOI will allow the improvement of the spatial resolution and an ultimate optimization in the design of scintillation detection systems.

In order to draw a full advantage from the increase of the detection efficiency offered by 3D PET, developments in the field of the image formation are equally necessary. The acquired PET data are not an image of the activity distribution in the source but rather projections of it. The unknown image must be estimated from the available data computationally. Great interest is turned nowadays to completely 3D iterative image-reconstruction techniques (21). The more interesting feature of iterative techniques consists of the possibility to incorporate to the reconstruction process the statistical model of the process of acquisition and detection. In spite of their high computational cost, iterative techniques offer greater flexibility in the data processing, particularly of data with elevated statistical noise. The implementation of these reconstruction algorithms on clusters of workstations, grid platforms, or other high performance computing systems is an area of state-of-the-art research (35).

In spite of the fact that the clinical impact of the attenuation correction for whole-body F-18-FDG PET studies is still under discussion and study, iterative image-reconstruction techniques combined with correction for measured attenuation seem to offer various advantages:

- anatomical localization and spatial definition of lesions are improved,
- the geometric distortions observed can be compensated and corrected (requirement for being able to proceed to the co-recording with anatomical images – CT, MRI, and so on),
- the tracer update can be quantified.

An issue of greater technological interest for its major impact on oncological diagnosis is the development of integrated multimodality systems PET/CT (36). A PET/CT system consists of a PET tomograph and a CT tomograph, both of the last generation, assembled in a single gantry, controlled from a single workstation, with one unique patient bed. A PET/CT system allows the acquisition of PET and CT images in a unique examination with significant advantages:

- reduction of the examination time,
- integrated diagnosis by means of combined use of information from PET and CT,
- accurate interpretation of the PET functional images based on anatomical CT images (functional-anatomic correlation),
- improvement of the PET functional image quality using the anatomical information from CT (reconstruction with iterative techniques of the PET data with the use of the anatomical CT information as a priori information, for attenuation correction, and for accurate scatter correction, and for the correction of the partial volume effect),
- elimination of the radionuclide source for transmission scanning and elimination of the need for periodic replacement of decayed transmission sources.

The development of commercial PET/CT systems is quite recent, and the number of such systems installed and operational is still limited. Beyond the evaluation of the clinical effectiveness of such systems, various technical aspects still demand additional studies based on the clinical experience. The techniques of patient positioning must be optimized (arm position, etc.). The correction for attenuation based on CT studies must be validated (calibration of the attenuation-correction coefficients based on CT to the 511 keV energy window). The alignment of CT and PET studies must be verified, in particular regarding the acquisition protocols (conditions of apnea in CT studies and free respiration in PET studies). Furthermore, the

performance of these complex and expensive systems should be compared with the performance of currently available software-based solutions for the co-registration and fusion of multimodality images (PET with CT, but also PET with MRI, ultrasound, etc.), which are shown to produce very accurate results, at least for brain studies.

Apart from whole-body human examinations, a challenging area of state-of-the-art research at the limits of current PET technology is the development of dedicated tomographs for small animal studies (25). In such systems, spatial resolution plays a crucial role as they are applied in the investigation of new pharmaceuticals and the development of new PET probes, as well as in the field of modern molecular biology, a scientific area that is currently focusing its interest toward imaging of laboratory mice and rats. As both the resolution and the sensitivity of small animal PET scanners are still limited by detector technology, image reconstruction algorithms, and scanner geometry, significant improvements may be expected in the performance of small animal PET scanners, whether prototype or commercial systems. In addition, multimodality imaging systems that will provide biological and anatomical information in an integrated setting, according to the model of PET/CT (or even PET/MR, etc.) systems already commercially available for human studies, should soon become available. The role of small animal PET in modern biology and pharmaceutical discovery and evaluation is in the process of being established, and it is likely that in vivo information of great value will be obtained. In addition, it is probable that the demanding requirements that small animal studies place on PET will result in technical advances and new technologies, which will dramatically improve the spatial resolution and image quality of clinical PET scanners for humans.

People today expect quality medical care at a reasonable cost, with accurate diagnosis and treatment, without having to undergo multiple exams and painful surgical exploration, and with fast and reliable results. Molecular imaging techniques, such as PET, display the biological basis of function in the organ systems of the human body unobtainable through any other means (37). PET is changing the way doctors manage care of their patients for some of today's most devastating medical conditions.

BIBLIOGRAPHY

1. Nutt R. The history of positron emission tomography. Mol Imag Biol 2002;4(1):11–26.
2. Brownell GL. A history of positron imaging. Online. 1999. Available at http://www.mit.edu/~glb/.
3. Phelps ME. PET Molecular Imaging and Its Biological Applications. New York: Springer; 2004.
4. Gambhir SS, Czernin J, Schwimmer J, Silverman DHS, Coleman E, Phelps ME. A tabulated summary of the FDG PET literature. J Nucl Med 2001;42:1S–93S. Online. Available at http://www.petscaninfo.com/zportal/portals/phys/clinical/jnmpetlit.
5. Sorenson JA, Phelps ME. Physics in Nuclear Medicine, 2nd ed. Orlando, FL: Grune and Stratton Inc.; 1987.
6. Zaidi H, Hasegawa B. Determination of the attenuation map in emission tomography. J Nucl Med 2003;44(2):291–315.
7. Turkington TG. Introduction to PET instrumentation. J Nucl Med Tech 2001;29(1):4–11.
8. Melcher CL, Schweitzer JS. Cerium-doped lutetium oxyorthosilicate: A fast, efficient new scintillator. IEEE Trans Nucl Sci 1992;39:502–505.
9. Ishibashi H, Kurashige K, Kurata Y, Susa K, Kobayashi M, Tanaka M, Hara K, Ishii M. Scintillation performance of large Ce-doped Gd_2SiO_5 (GSO) single crystal. IEEE Trans Nucl Sci 1998;45(3):518–521.
10. Moses WW, Derenzo SE. Lead carbonate, a new fast, heavy scintillator. IEEE Trans Nucl Sci 1990;37(1):96–100.
11. Moses WW, Derenzo SE, Shlichta PJ. Scintillation properties of lead sulfate. IEEE Trans Nucl Sci 1992;39(5):1190–1194.
12. Moses WW, Derenzo SE. Cerium fluoride, a new fast, heavy scintillator. IEEE Trans Nucl Sci 1989;36(1):173–176.
13. Ziegler SI, Rogers JG, Selivanov V, Sinitzin I. Characteristics of the new $YAlO_3$:Ce compared with BGO and GSO. IEEE Trans Nucl Sci 1993;40(2):194–197.
14. Moses WW, Derenzo SE, Fyodorov A, Korzhik M, Gektin A, Minkov B, Aslanov V. $LuAlO_3$:Ce-a high density, high speed scintillator for gamma detection. IEEE Trans Nucl Sci 1995; 42(4):275–279.
15. Gilardi MC. Tomografi PET: Attualitá e prospettive (in italian). XI Nat. Course on Professional Continuing Education in Nuclear Medicine (Pisa, 29-31/10/2001) Online. Available at http://www.aimn.it/ecm/pisa_01/Gilardi.pdf.
16. Casey ME, Nutt R. A multislice two dimensional BGO detector system for PET. IEEE Trans Nucl Sci 1986;33:460–463.
17. Karp JS, Muehllehner G, Mankoff DA, Ordonez CE, Ollinger JM, Daube-Witherspoon ME, Haigh AT, Beerbohm DJ. Continuous-slice PENN-PET: A positron tomograph with volume imaging capability. J Nucl Med 1990;31:617–627.
18. Surti S, Karp JS, Freifelder R, Liu F. Optimizing the performance of a PET detector using discrete GSO crystals on a continuous lightguide. IEEE Trans Nucl Sci 2000;47:1030–1036.
19. Jeavons A, Parkman C, Donath A, Frey P, Herlin G, Hood K, Magnanini R, Townsend D. The high-density avalanche chamber for Positron Emission Tomography. IEEE Trans Nucl Sci 1983;30:640–645.
20. Phelps ME, Cherry SR . The changing design of positron imaging systems. Clin Positron Imag 1998;1(1):31–45.
21. Tarantola G, Zito F, Gerundini P. PET instrumentation and reconstruction algorithms in whole-body applications. J Nucl Med 2003;44(5):756–768.
22. Let's Play PET. 1995, May 1. Online. Available at http://laxmi.nuc.ucla.edu:8000/lpp/lpphome.html.
23. Health Technology Advisory Committee. 1999, March. Positron emission tomography (PET) for oncologic applications. Online. Available at http://www.health.state.mn.us/htac/pet.htm.
24. Phelps ME. Positron emission tomography provides molecular imaging of biological processes. Proc Nat Acad Sci 2000;97(16): 9226–9233.
25. Fowler JS, Volkow ND, Wang G, Ding Y-S, Dewey SL. PET and drug research and development. J Nucl Med 1999;40: 1154–1163.
26. Chatziioannou AF. Molecular imaging in small animals with dedicated PET tomographs. Eur J Nucl Med 2002; 29(1):98–114.
27. Kontaxakis G, Dimitrakopoulou-Strauss A. New approaches for position emission tomography (PET) in breast carcinoma. In: Limouris GS, Shukla SK, Biersack H-J, eds. Radionuclides for Mammary Gland–Current Status and Future Aspects. Athens, Greece: Mediterra Publishers; 1997. p 21–36.
28. Beyer T, Townsend DW, Brun T, Kinahan PE, Charron M, Roddy R, Jerin J, Young J, Byars L, Nutt R. A combined

PET/CT scanner for clinical oncology. J Nucl Med 2000; 41(8):1369–1379.
29. Reba RC. PET and SPECT: Opportunities and challenges for psychiatry. J Clin Psychiatry 1993;54:26–32.
30. Fahey FH. Data acquisition in PET imaging. J Nucl Med 2002; 30(2):39–49.
31. Castiglioni I, Cremonesi O, Gilardi MC, Bettinardi V, Rizzo G, Savi A, Bellotti E, Fazio F. Scatter correction techniques in 3D PET: A Monte Carlo evaluation. IEEE Trans Nucl Sci 1999; 46(6):2053–2058.
32. Schmand M, Eriksson L, Casey ME, Andreaco MS, Melcher C, Wienhard K, Flugge G, Nutt R. Performance results of a new DOI detector block for a high resolution PET-LSO research tomograph HRRT. IEEE Trans Nucl Sci 1998;45(6):3000–3006.
33. Shao Y, Silverman RW, Farrell R, Cirignano L, Grazioso R, Shah KS, Vissel G, Clajus M, Tumer TO, Cherry SR. Design studies of a high resolution PET detector using APD arrays. IEEE Trans Nucl Sci 2000;47(3):1051–1057.
34. Meng LJ, Ramsden D. Performance results of a prototype depth-encoding PET detector. IEEE Trans Nucl Sci 2000; 47(3):1011–1017.
35. Kontaxakis G, Strauss LG, Thireou T, Ledesma-Carbayo MJ, Santos A, Pavlopoulos S, Dimitrakopoulou-Strauss A. Iterative image reconstruction for clinical PET using ordered subsets, median root prior and a We-based interface. Mol Imag Biol 2002;4(3):219–231.
36. Townsend DW. From 3-D positron emission tomography to 3-D positron emission tomography/computed tomography: What did we learn? Mol Imaging Biol 2004;6(5):275–290.
37. Phelps ME. PET: The merging of biology and imaging into molecular imaging. J Nucl Med 2000;41:661–681.

See also COMPUTED TOMOGRAPHY; IMAGING DEVICES; RADIOPHARMACEUTICAL DOSIMETRY.

PROSTATE SEED IMPLANTS

MARCO ZAIDER
Department of Medical Physics,
Memorial Sloan Kettering
Cancer Center
New York

DAVID A. SILVERN
Medical Physics Unit, Rabin
Medical Center
Petah Tikva, Israel

INTRODUCTION

Prostate seed implantation is a radiation therapy procedure by means of which small radioactive sources (colloquially referred to as "seeds") are permanently implanted in the tumor-bearing tissue. In the absence of more specific information concerning the location within the prostate of tumor deposits, the goal of the implant is to deliver a minimum dose to the entire prostate gland while minimizing the dose to any adjacent healthy tissues, in particular, the urethra and rectum. As cause-specific death in prostate cancer is predominantly the result of distant metastasis (and not local failure), the *raison d'être* of prostate implantation must be the notion of some causal link, as opposed to mere association, between local control and distant disease (1–3). Whether this is indeed the case, remains at this time controversial (4).

The absorbed dose in the target as well as its spatial and temporal configuration is the only treatment tool available to the radiation oncologist. Consequently, patient eligibility for permanent prostate implants is based on the physician's ability to physically deliver the dose to the target, in other words, placing the seeds at relevant locations within or near the gland. Guidelines for patient selection consist of stage (T1–T2 disease) and prostate volume (less than about 50 cm^3). (*Staging* refers to the size and location of the tumor, whether tumor cells have spread to the lymph nodes, whether cancer cells have metastized to other parts of the body and to the abnormality of the tumor cells – this latter is referred to as *grade* and quantified by the Gleason score. Thus, T1 refers to a clinically inapparent tumor (not visible or palpable), and T2 refers to a low-grade tumor confined within the prostate. Pretreatment with androgen-ablation therapy is sometimes used to reduce the prostate volume.) Counter-indications for brachytherapy are short life expectancy (<5 years), the presence of metastatic disease, prior transurethral resection of the prostate (TURP), prostatitis, acute voiding symptoms, and inflammatory bowel disease (5).

The treatment may be delivered as monotherapy (implant alone) or in combination with external beam radiation therapy (EBRT) depending on whether the disease is confined (in which case monotherapy is administered) or extends beyond the prostate. Indications for combined treatment are extracapsular extension (ECE) and/or seminal vesicle invasion (SVI). It has been suggested that the patient's prognosis category as defined by pretreatment stage, Gleason score, and prostate specific antigen (PSA) may be taken as an indication of the likelihood that the disease did not spread outside the prostate; essentially, the lower the risk, the larger the probability of a confined tumor. Two classification schemes are currently in use. According to one proposal, low-risk patients are those with T1–T2b stage, Gleason 2–6, and PSA of less than 10 ng/mL. Intermediate-risk patients have one unfavorable factor: PSA larger than 10 ng/mL, Gleason score 7 or larger, or T2c or greater. High-risk patients have at least two unfavorable risk factors. The other sorting idea considers the risk low when PSA $\leq$ 10 ng/mL, Gleason <7, and T1a or T2a; intermediate when 10.1 $<$ PSA $\leq$ 20, Gleason $=$ 7 or T2b, and high otherwise (6,7).

The implant is a three-step process: First, using an imaging study of the prostate, the treatment planner calculates the number and location of seeds within the prostate volume that will result in a dose distribution in agreement with the prescribed constraints. The implantation is performed as a one-off outpatient surgical procedure. A post-implant evaluation, based on which the dosimetric quality of the implant is assessed, follows the treatment.

As with the other treatment choices (radical prostatectomy, EBRT), the survival benefits of transperineal permanent interstitial implantation among men with early, localized prostate cancer are uncertain (8,9). This state of affairs is compounded by the fact that the treatment of prostate cancer by either modality is accompanied by the

risk of (occasionally permanent) rectal and urinary toxicity, as well as sexual disfunction (10–12). (Brachytherapy may be less likely to result in impotence in urinary incontinence than other forms of treatment.) Thus, for many patients with early, localized prostate cancer (the typical brachytherapy candidate), the decision to undergo a treatment of questionable benefit yet tangibly impacting on their quality of life (QOL) is understandably difficult; as a result, diminishing the risk of complications, and at the same time maintaining good dosimetric coverage of the tumor, remains the overriding concern in prostate brachytherapy.

In this article, we shall provide a step-by-step tutorial on permanent prostate implants, (by necessity) as practiced at the Memorial Sloan Kettering Cancer Center (MSKCC).

PREPLANNING OR INTRAOPERATIVE PLANNING?

Two modalities are currently in use for planning prostate implants. Preplanning refers to the situation where the treatment plan is completed several weeks before the actual implantation. The American Brachytherapy Society (ABS) discourages this approach, essentially because of well-recognized problems that may develop at the time the plan is implemented (5). For instance:

1. It is difficult to duplicate the patient position in the operating room (OR) to match the patient position during the preplanning simulation.
2. Patient geometry can change over time; for instance, urine in the bladder or feces in the rectum may swell the prostate.
3. A pretreatment plan may prove impossible to implement due to the needle site being blocked by the pubic symphysis.

The alternative to preplanning, which we and others strongly advocate, is to perform the plan in the OR using ultrasound (US) images of the treatment volumes acquired with the patient on the operating table in the implantation position (13–19). The ability to obtain a dose-optimized plan within a reasonable time (say, 10 minutes or less) is the key to intraoperative approach. Clearly, a manual (i.e., trial and error) approach to finding optimal seed positions in the OR will not do. The *sine qua non* condition of OR-based planning is then the availability of a computer-optimized planning technique, as described below.

ISOTOPE SELECTION AND DOSE PRESCRIPTION

The two isotopes commonly used for permanent prostate implants are ^{125}I (mean photon energy $E_\gamma = 27$ keV, half life $T_{1/2} = 60$ days) and ^{103}Pd ($E_\gamma = 21$ keV, $T_{1/2} = 17$ days). An important property these isotopes share is their low effective energies. At these energy levels, practically all decay radiation emitted by the implanted sources is absorbed in the patient's body. This enables the patient to be discharged shortly after the implant procedure without fear of posing a radiation hazard to the general public or to the patient's family.

Dose prescription in prostate brachytherapy makes use of the concept of *minimum peripheral dose* (mPD), which is defined as the largest isodose surface that completely surrounds the clinical target (20). The total dose prescription for patients treated at MSKCC is 144 Gy (mPD) for ^{125}I and 140 Gy for ^{103}Pd (21). The initial dose rate $\dot{D}(0)$ corresponding to these values can be calculated from

$$D_{\text{total}} = \frac{T_{1/2}}{\ln(2)}\dot{D}(0) \tag{1}$$

Thus, for ^{125}I, one has $\dot{D}(0) = 7$ cGy/h, whereas for ^{103}Pd, the corresponding value is 24 cGy/h. Based on radiobiological considerations it was hypothesized that the higher initial dose rate of ^{103}Pd may be more appropriate for rapidly proliferating tumor cells. Gleason score is taken as marker for such cells, hence, the notion that ^{103}Pd should preferentially be used for high-grade tumors. Retrospective studies have failed to demonstrate any clinical (22,23) or dosimetric difference between the two isotopes. (One may also note that in the United States, the price of a typical ^{125}I seed is about half the price of a ^{103}Pd seed; as well, a typical implant would use, say, 75 ^{125}I seeds as against some 100–110 ^{103}Pd seeds.)

A second difference between the two isotopes is the potential for somewhat larger relative biological effectiveness (RBE) of the ^{103}Pd compared with ^{125}I (24,25).

THE PHYSICAL CHARACTERISTICS OF ^{125}I AND ^{103}Pd

^{125}I is produced in a reactor by irradiating ^{124}Xe with neutrons to form ^{125}Xe. ^{125}Xe has a 16.9 h half-life and decays to ^{125}I via electron capture. ^{125}I in turn decays to an excited state of ^{125}Te via electron capture. The excited ^{125}Te nucleus immediately releases its excess energy in the form of a 35.5 keV gamma photon in 6.66% of the transformations, with the energy of the remaining 93.34% of the transformations being released as internal conversion electrons (26). The rearrangement of the orbital electrons results in the emission of characteristic X rays and Auger electrons. The Auger electrons are blocked by the encapsulation of the source and do not directly contribute to patient dose. As a result of the lower energy characteristic X-ray emissions, the average photon energy of ^{125}I is 28 keV.

^{103}Pd is produced in a reactor by irradiating stable ^{102}Pd with neutrons. ^{103}Pd decays with a 17 day half-life to excited states of ^{103}Rh via electron capture. The excited ^{103}Rh nuclei lose nearly all of their excess energy via internal conversion (26). The rearrangement of the orbital electrons results in the generation of characteristic X rays and Auger electrons. As is the case with ^{125}I sources, the Auger electrons are blocked by the encapsulation and do not directly contribute to patient dose.

The radioactive sources must be fabricated to high-quality control standards. The sources must remain biocompatible *in vivo* for decades. To prevent the internal contents of the sources from diffusing into the body tissues, the integrity of the source encapsulation must also remain sound for a period of decades. The physical size of the sources must be sufficiently small to allow their interstitial

implantation without causing undue tissue trauma or interfering with physiologic function. In addition to biocompatibility issues, the sources must be physically strong enough to maintain their shape and integrity during sterilization, routine handling, and implantation. As well as to the need for physical ruggedness, the encapsulation must be made of a material that will not absorb an undue fraction of the emitted photons. For the purpose of source localization on radiographs and computed tomography (CT) images, it is desirable for the sources to be radiopaque while causing minimal imaging artifacts.

For meeting the aforementioned requirements, titanium is the material of choice for source encapsulation. Titanium is as strong as steel but 45% lighter. It is only 60% heavier than aluminum but twice as strong. This metal is also highly resistant to sea water. As human tissues have a high degree of salinity, titanium can maintain its integrity *in vivo* for the remainder of the patient's life. Titanium has an atomic number of 22, low enough not to cause serious artifacts on CT images. Due to the high strength of titanium, the encapsulation can be made thin, minimizing absorption of the emitted photons. The main drawback of titanium is that it is expensive, significantly adding to the cost of the radioactive source.

The actual internal structure of the radioactive sources is vendor specific. The thickness of the titanium encapsulation ranges from 0.04 to 0.06 mm (26). In the classic Amersham 6711 source, the radioactive material is adsorbed to the surface of a silver rod. A schematic representation of this source is shown in Fig. 1. The silver serves as a radio-opaque marker. In other source models, the radioactive material is adsorbed onto resin beads, impregnated in ceramic material, or coated by other means onto various substrates. Most ^{125}I and ^{103}Pd radioactive sources include radio-opaque marker material. Gold, silver, tungsten, and lead are the commonly employed marker materials used in ^{125}I and ^{103}Pd sources. The differences in the internal structures of the radioactive sources result in vendor-specific dose distributions. These differences result in differing photon energy spectra, source anisotropy, and self-absorption properties. The dosimetric properties of ^{125}I and ^{103}Pd sources are discussed in the next section.

DOSIMETRY

Before any radioactive source can be employed in a clinical implant procedure, the dose distribution around the source in question must be known. One obvious requirement is that the accuracy of the dose calculations be as high as possible. Another useful requirement is wide acceptance of the dosimetric formalism applied. A universal dose formalism simplifies comparison of treatment outcomes of implants performed by different institutions and helps establish universal treatment protocols.

To assure accurate dosimetric calculations, the following physical properties must be ascertained:

1. Photon interactions with structures inside the source and encapsulation.
2. Geometric distribution of radioactive material within the source.
3. Photon interactions with the tissues encompassing the source.
4. Reduction of radiation intensity as a function of distance from the source.

Photon interactions within the source have a noticeable effect on the shape, intensity, and energy spectrum of the dose distribution surrounding the source. These internal photon interactions affect the anisotropy of the emitted radiation. (Source anisotropy is a measure of angular dependence of the dose distributions surrounding the source.) For a point source with no asymmetric self-absorption, dose would only be a function of distance from the source without any angular dependence. Photon interactions inside the source also influence the energy spectrum of the emitted photons. Depending on the physical construction and materials used to fabricate the source, the emission spectra will vary among different source models. Lower energy photons will be preferentially attenuated compared with higher energy photons. The degree of filtration is dependent on the construction of the source. Another alteration to the intrinsic spectrum of the radioactive material is the generation of characteristic X rays by photoelectric interaction with the materials inside the source. One noteworthy example of spectral variations among different ^{125}I source models is the use of silver as the radio-opaque marker used by certain manufacturers. Silver has a K-edge that occurs at 25 keV (27,28). As the intrinsic photon emissions of ^{125}I are in this energy range, the photoelectric cross section for interaction with the silver marker is high. As a result, I^{125} sources using silver as the radio-opaque marker will have a local peak around 25 keV in their emission spectra. I^{125} sources using another element for radio-opacity will not exhibit a 25 keV peak in their emission spectra. Differences in source construction also influence the degree of Compton scattering, further altering the emission spectra. The geometric distribution of the radioactive material inside the source will affect the angular dependence of the dose distributions around the source as well as the dose reduction as a function of distance from the source. This effect will be investigated when source geometry factors are discussed later in this section.

Another important factor that must be calculated by a dose formalism is tissue attenuation as a function of distance from the source. Tissue attenuation is a function of how the emission photons interact with the tissues. When using I^{125} or ^{103}Pd sources, the predominant interactions

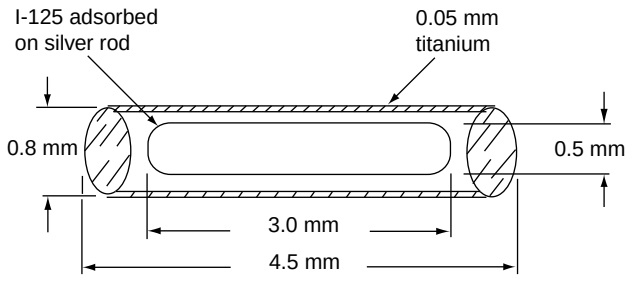


Figure 1. Schematic representation of an Amersham 6711 ^{125}I source.

are Compton scattering and photoelectric effect. In soft tissue, the probabilities of photoelectric and Compton interactions are equal at a photon energy of about 25 keV (28). The dose decreases with distance from the source due to these photon interactions.

The simplest dosimetric formalism is to assume point-source geometry, neglecting angular dependencies on the dose distributions. In this formalism, the dose is strictly a function distance from the source. Two components are assumed to contribute to the resulting dose at a given point. One component is the dose falloff due the geometric shape of the radioactive source. For a point source, this falloff component is the inverse square law, namely $1/r^2$. This dose falloff occurs irrespective of any photon interactions in the medium surrounding the source. It is solely dependent on the photon fluence intensity being geometrically reduced by the inverse square law. The second component of dose falloff results from photon interactions in the surrounding medium. One analytical approach to modeling this phenomenon is to assume that the dose reduction follows a simple exponential function, namely $F(r) = e^{-\mu r}$ where μ is an average linear attenuation coefficient for the energy spectrum of the emitted photons. It is also assumed that dose is directly proportional to the source activity A. The equation for calculating the dose using this formalism is

$$D(r) = \Gamma A \, f_{\text{med}} T_{\text{av}} \frac{e^{-\mu r}}{r^2} \tag{2}$$

where D represents the dose at distance r and A is the source activity. The factors Γ and f_{med} are the exposure rate constant and the tissue f-factor, respectively. The f-factor converts the exposure in air to absorbed dose in a small piece of tissue just large enough to assure electronic equilibrium. T_{av} is the average life an isotope atom exists before undergoing a nuclear transformation and μ is the effective linear attenuation coefficient. This analytical equation suffers from two drawbacks, the first being that this exponential equation is only rigorously correct for narrow-beam geometries. In deriving this equation, it is implicitly assumed that all photons that interact with the medium are removed from the beam and that no further interactions of scattered photons occur in the path of the beam. The geometry of a radioactive source in a medium is clearly not narrow-beam. Compton photons do in fact interact with the medium in the beam and hence contribute to dose. A second drawback of using an exponential is the implicit assumption that the photons are mono-energetic, clearly not the case for either ^{125}I or ^{103}Pd emissions. Although this formalism is not rigorously correct for the reasons stated, it has been used for many years in brachytherapy treatment planning. By empirically determining values for Γ and μ, the discrepancies of calculated and measured doses could be made acceptably small. In the years that this equation was used, modern instrumentation was not available for precise dose measurement, computers were slow, and the standards of conformance were not as stringent as today.

The accuracy of the formalism can be improved by replacing the exponential equation with a data table. The values in the table consist of experimentally measured doses in water at known distances, multiplied by the square of the distance. Photon interactions in water are similar to those in soft tissue. The dose at any arbitrary distance from the source is calculated via linear interpolation of the tabulated data and by dividing by the square of the distance. Using tabulated data solves the two problems associated with the exponential function. As tabulated data were derived from measurements in the true broad-beam geometry of the source, the formalism inherently accounts for the dose occuring from Compton-scattered photons as well as the primary photons. As a table could be created for any source model, the data will inherently account for the energy spectrum of the emitted photons. An added benefit provided by the table is that it accounts for the selective tissue filtration of lower energy photons with increasing depths. As the emission spectra vary among different source models, the tissue filtration will also likewise vary.

A more sophisticated formalism can be developed by accounting for the geometric distribution of the radioactive material inside the source. The most natural extension of the formalism is the assumption that the radioactive material is distributed as a line source. Although the distribution of the radioactive material in many commercially available sources is not a true uniform line, the line source model is still a more accurate and realistic representation than the point source model. Applying a line source model significantly complicates the dosimetric computations. When applying the point source model, only the distance from the source to the calculation point need be known to uniquely determine the dose. By contrast, using a line source model requires that the angular orientation in addition to the distance of each dose calculation point in relation to the source be known. The angular orientation of the calculation point with respect to the source needs to be calculated using analytic geometry.

The angular dependence of the dose distribution around a source is the result of two separate processes. The first results from the distribution of the radioactive material inside the source, and the second results from the angular dependence of attenuation of the photons within the source. To develop a formalism based on a line source, it is easiest to start by calculating the dose to a differential piece of tissue in empty space. It is assumed that there is no attenuation in the tissue and that electronic equilibrium exists. As tissue effects will be introduced at a later stage, this assumption does not compromise the rigor of the development of the formalism. At this stage, the self-absorption and tissue attenuation will not be considered.

The angular dependence of dose resulting from the geometric distribution of the radioactive material can be modeled by a line source. A schematic representation of a line source is shown in Fig. 2.

For any point P one can define a two-dimensional coordinate system as shown in Fig. 2. With such a coordinate system defined, the dose to any point P from the line source can be calculated. It is seen from Fig. 2 that the following formulation holds:

$$X = -r\cos\theta; \quad Y = r\sin\theta; \quad r = Y/\sin\theta; \quad \tan\theta = -Y/X;$$
$$X = -Y/\tan\theta \tag{3}$$

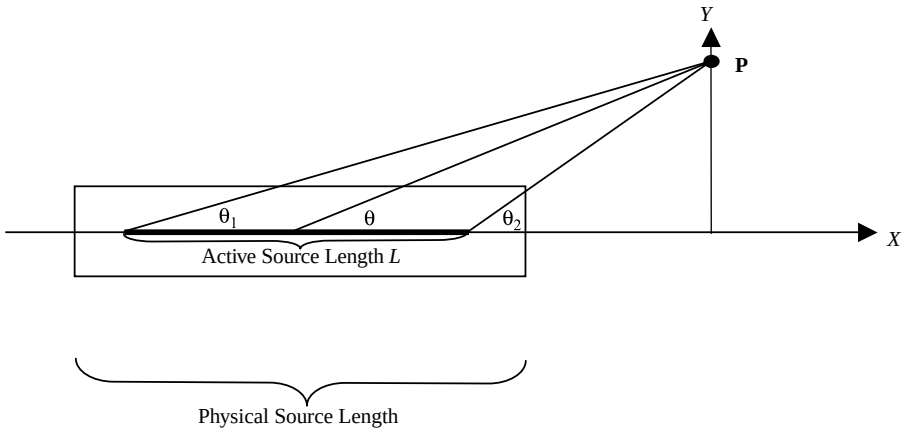


Figure 2. Dose to a point P from a line source. Physical length includes encapsulation. Active length only includes length of radioactive material.

The differential length dX is calculated to be

$$dX = Y(\tan\theta)^{-2}(\sec\theta)^{-2}d\theta = Y(\sin\theta)^{-2}d\theta \qquad (4)$$

The total activity A of the line source is assumed to be uniformly deposited along the active length L of the source. Under ideal conditions, uthe dose to a differential piece of tissue at point P in free space is given by

$$D(r) = A\Gamma\ f_{\text{med}}\frac{T_{\text{av}}}{r^2} \qquad (5)$$

To calculate the dose from the line source shown in Fig. 2, the source can be considered a continuum of differential point sources resulting in

$$dD = \frac{A}{L}\Gamma \quad f_{\text{med}}T_{\text{av}}\frac{dX}{r^2} = \frac{A}{L}\Gamma \quad f_{\text{med}}T_{\text{av}}\frac{d\vartheta}{Y} \qquad (6)$$

The total dose to point P is thus

$$D = \frac{A}{L}\Gamma \quad f_{\text{med}}T_{\text{av}}\int_{\vartheta_1}^{\vartheta_2}\frac{d\vartheta}{Y} = \frac{A}{L}\Gamma \quad f_{\text{med}}T_{\text{av}}\frac{\vartheta_2 - \vartheta_1}{r\sin\vartheta} \qquad (7)$$

By using equation 7, it is now possible to account for the linear distribution of radioactive material inside the source. In clinical brachytherapy, however, one is generally interested in knowing the dose when the source is in the patient and not in free space. This means that the photon interactions with the surrounding tissues must be factored into the analysis. Additionally, the self-absorption of the source must be taken into account. Unlike the linear distribution of the radioactive material, the self-absorption and tissue interactions are not amenable to a rigorous treatment. These data must be derived either from direct measurement or from Monte Carlo simulation, or a combination of the two. The data tables would need to be two-dimensional to account for the distance from the calculation point to the source as well as the angular orientation of the calculation point to the center of the source. As the active length of the source is known, the values of θ_1 and θ_2 are uniquely determined. Once these data tables are available, the component of the dose due to geometric distribution of the radioactive material can be removed from the data by dividing the tabulated doses by the ideal line source dependency. Reviewing equation 7, one can separate the terms involving the linear distribution of the radioactive material. Doing this yields the following:

$$G = \frac{\vartheta_2 - \vartheta_1}{Lr\sin\vartheta} \qquad (8)$$

where G is referred to as the *geometry factor*. Dividing the tabulated doses by G removes the line source contribution. Removing the geometry factor reduces the dependency of the data on distance from the source, enabling the use of smaller data tables for attaining the same degree of accuracy. The data tables still account for the tissue interactions and self-absorption. For permanent implant brachytherapy, the data tables can be in the form of total dose per unit source activity (or air kerma strength) because the implant time is infinite and the isotope half-life is known. The new formalism can be summarized as follows:

1. For each point P in space, define a two-dimensional coordinate system and compute the distances to the centers of each implanted source and compute the respective angles θ, θ_1, and θ_2.
2. Use equation 8 to calculate the geometry factors at point P for each implanted source.
3. Interpolate the data tables based on the distances and angles for each implanted source calculated in step 1.
4. For each implanted source, multiply the respective geometry factor, source activity (or air kerma strength), and interpolated table value together. Each product represents the dose to point P for each implanted source.
5. Sum the individual doses to obtain total dose at point P.

These calculations are tedious and time consuming if performed by hand. The only feasible way forward is to code the formalism in software and have the computer perform the computations. In this way, a finely spaced three-dimensional data grid of calculation points can be rapidly calculated. The dose at any arbitrary point in space can be calculated by interpolating the three-dimensional grid of calculated doses. This will be discussed in a later section devoted to treatment planning.

The dose formalisms discussed thus far are but a small sample of the calculation methods used in brachytherapy

treatment planning. They were discussed to serve as illustrative examples and to provide a brief introduction to the physics involved in the development of dosimetry formalisms. Over the years, many different dosimetry formalisms have been proposed and implemented. By the early 1990s, treatment planning computers have become ubiquitous in radiotherapy departments. With many such systems in use, questions began to arise regarding the accuracy, consistency, and general agreement of the calculated doses generated by these systems. It was obvious that there were differences in the values calculated by these treatment planning systems as each system implemented its own dosimetric algorithm. During this same period, researchers also proposed new formalisms based on physical quantities not widely used in brachytherapy treatment planning. As a result of these issues, the American Association of Physicists in Medicine (AAPM) decided that the time had come to develop a standardized brachytherapy formalism. The adoption of such a formalism would standardize the dosimetric calculations, reducing the differences in values calculated by different treatment planning systems. The doses calculated among different institutions will be in closer agreement, enabling more realistic comparisons of different brachytherapy protocols.

The AAPM Radiation Therapy Committee Task Group 43 established a new recommended formalism. This formalism was published in 1995(26). The Task Group 43 (TG43) formalism is a radical departure from most established methodologies used up to that time. In following the new formalism, changes in dosimetric values in some instances were of sufficient magnitude to mandate changes in prescription doses for maintaining clinical consistency with older formalisms.

The most fundamental change recommended by TG43 is to use air kerma strength S_k in lieu of activity. The activity of a radioactive source is the number of disintegrations per unit time. As a fundamental unit, activity in and of itself does not provide any information regarding the nature of the energy deposition of the decay emissions. Traditional dosimetry formalisms need to use exposure rate constants and *f*-factors to convert from activity to exposure in air to dose deposited in tissue. The exposure rate constant is also isotope-specific.

Air kerma strength is defined as follows. A mass of air, dm, is placed a distance r_{ref} from the source along the perpendicular bisector. The source and air mass are in a vacuum. *Air kerma strength* is the kerma rate in mass dm multiplied by the square of the distance r_{ref}. Unlike activity, which is only applicable to radioactive isotopes, air kerma strength can be applied to any source of uncharged particles. For example, the radiation output from a linear accelerator or X-ray tube can also be quantified in terms of air kerma strength.

Another new quantity used by TG43 is the *dose rate constant* Λ defined as the dose rate to water at a distance of 1 cm from the perpendicular bisector of the source. Using the dose rate constant represents another departure from basing absorbed dose on exposure to air. It is a trend in the medical physics community to move toward basing dosimetric calculations and measurements on dose to water. The dose rate constant replaces the exposure rate constant used in older brachytherapy formalisms.

For modeling the contribution of the geometric distribution of the radioactive material inside the source, TG43 endorses the use of either a point source representation (inverse square law) or a line source representation (equation 8). These geometric factors have been in use before the advent of the TG43 report.

The interactions of the emitted photons in tissues are modeled by a new function defined by TG43, namely the radial dose function $g(r)$. The function $g(r)$ is given by the following equation:

$$g(r) = \frac{dD(r, \vartheta_0)/dt\, G(r_0, \vartheta_0)}{dD(r_0, \vartheta_0)/dt\, G(r, \vartheta_0)} \tag{9}$$

where D is the dose at a point along the perpendicular bisector at a distance r from the source and G is the geometry factor that is equal to either the inverse square law or equation 8. In equation 9, θ_0 is equal to $\pi/2$. The radial distance r_0 is the reference distance to which all TG43 parameters are normalized. In practice, r_0 is taken to be 1 cm. It must be stressed that $g(r)$ is only defined along the perpendicular bisector of the source. It can be seen that when r is equal to r_0, $g(r)$ equals 1. The factors in equation 9 involving G remove the dependency of the geometric distribution of radioactive material inside the source on $g(r)$. The radial dose function models the interactions of the photons with tissues along the transverse axis of the source. In practical implementations of TG43, $g(r)$ is based on tabulated values, which in turn are based on measured data or Monte Carlo simulations. In some implementations, analytic functions are fit to the data, whereas in others, the value of the radial dose function is calculated via interpolation.

Another phenomenon that needs to be modeled is the anisotropic nature of the radiation resulting from photon interactions inside the source. If there were no interactions of the radiation within the source, all of the angular dependence of the radiation pattern would result from the geometry factor, assuming the source is in a homogeneous medium. In reality, however, self-absorption is significant, especially at the low energies of ^{125}I and ^{103}Pd.

In the original TG43 report, three methods of modeling source anisotropy were proposed. The most general of these models is the source anisotropy function $F(r, \theta)$. This function is defined as follows:

$$F(r, \vartheta) = \frac{dD(r, \vartheta)/dt\, G(r, \vartheta_0)}{dD(r, \vartheta_0)/dt\, G(r, \vartheta)} \tag{10}$$

This function in effect is the ratio of the dose rate at an arbitrary distance from the source r and angle θ multiplied by the ratio of the geometry factor at a distance r but on the transverse axis and the geometry factor at the location r and θ. This function quantifies the angular variation of the dose distribution removing the contribution of the geometry function. In most practical implementations, the anisotropy function is calculated via bilinear interpolation of a two-dimensional data table.

A second method for modeling source anisotropy is to represent source anisotropy as a function of only the

distance from the source. In this case, the dose rate is averaged over all values of solid angle from 0 through 4π steradians. The one-dimensional function is referred to as the anisotropy factor $\phi_{an}(r)$. By taking advantage of the cylindrical symmetry of radioactive sources, the solid angle integral for defining this factor reduces to the following:

$$\phi_{an}(r) = \frac{1}{2dD(r,\vartheta)/dt} \int_0^{\pi} \frac{dD(r,\vartheta)}{dt} \sin\vartheta \, d\vartheta \qquad (11)$$

One fundamental difference between defining $\phi_{an}(r)$ and $F(r,\theta)$ should be pointed out. The geometry factor is not removed from $\phi_{an}(r)$ as is the case for $F(r, \theta)$. From a numerical standpoint, the average value of the geometry factor taken over 4π steradians is nearly equal to the nominal value of $G(r,\theta_0)$. This is especially true for distances greater than the active length of the source. Moreover, $\phi_{an}(r)$ is an average value that will result in an inevitable error in the actual dose calculation. Factoring out the geometry factor would not significantly improve the accuracy of the dosimetry. If the geometry factor contribution was in fact removed, two anisotropy factors would need to be calculated, one for point source geometry and the other for line source. The main motivating factor for defining $\phi_{an}(r)$ is to accommodate existing treatment planning systems that do not support two-dimensional dose calculations. Given a choice, using $F(r,\theta)$ is preferred as it is rigorous.

A third anisotropy correction method prescribed by the original TG43 report is the use of an anisotropy constant ϕ_{an} that is an average value independent of distance. The original TG43 report was updated in 2004. In the updated report, use of anisotropy constants was made obsolescent and is no longer considered consistent with current standards of practice. A more detailed discussion of the TG43 update is discussed below.

Application of the TG43 formalism can be summarized by the following equations where all of the terms have been discussed:

$$dD(r,\vartheta) = S_K \Lambda \frac{G(r,\vartheta_0)}{G(r_0,\vartheta_0)} g(r) F(r,\vartheta) \qquad (12)$$

Equation 12 is the rigorous implementation of the original TG43 formalism. A simpler implementation of equation 12 for use in treatment planning systems not supporting two-dimensional dose calculations is the following:

$$dD(r,\vartheta) = S_K \Lambda \frac{G(r,\vartheta_0)}{G(r_0,\vartheta_0)} g(r) \phi_{an}(r) \qquad (13)$$

The third equation using the anisotropy constant is as follows:

$$dD(r,\vartheta) = S_K \Lambda \frac{G(r,\vartheta_0)}{G(r_0,\vartheta_0)} g(r) \phi_{an} \qquad (14)$$

Use of equation 18 is no longer recommended in the updated TG43 protocol. All treatment planning systems need to be capable of performing one-dimensional calculations. It is thus always possible to use the anisotropy factor $\phi_{an}(r)$. In some treatment planning systems, the product of the radial dose function and the anisotropy factor may need to be used if only one distance-dependent term is used for calculating the dosimetry.

To end this section, a brief discussion of the updated TG43 formalism is presented. This update was published in 2004, 9 years after the original TG43 report (29,30). In the original formalism, the same radial dose functions were used irrespective of geometry factor. In the new formalism, two sets of radial dose functions are recommended. One radial dose function is to be used for point source geometry, whereas the other is to be used for line source geometry. Although the updated and original reports define the radial dose function using the same equation, the published values could be used for both geometry factors. To improve the consistency of the dose calculations, two sets of radial dose functions are now recommended. The choice of geometry factor dictates which radial dose function is to be used. Further refinements were also made in the tabulated values presented. In the intervening 9 years between the original report and the update, several new radioactive sources were introduced. The updated report includes published data for use with the newer source models. In the new report, the use of anisotropy constants is no longer recommended.

ULTRASOUND-GUIDED IMPLANTATION TECHNIQUE

The implantation proceeds as follows. The patient is placed in the extended lithotomy position. An ultrasound probe is positioned in the rectum, and needles are inserted along the periphery of the prostate using a perineal template as a guide. Trans-axial images of the prostate are acquired at 0.5 cm spacing (from base to apex), transferred to the treatment planning system using a PC-based video capture system, and calibrated. For each US image, prostate and urethra contours as well as the anterior position of the rectal wall are entered. Needle positions are identified on the ultrasound images and correlated with the US template locations. The contours, dose reference points, needle coordinates, and data describing the isotope/activity available along with predetermined dose constraints and their respective weights serve as input for the dose optimization algorithm.

TREATMENT PLANNING

A commissioned treatment planning system is the minimum equipment required for performing brachytherapy treatment planning. State-of-the-art treatment planning systems are based on modern computer hardware and software. Practically all modern brachytherapy treatment systems implement the TG43 dosimetry formalism for low-energy radioactive sources. These systems provide interfaces to imaging scanners such as CT, US, and magnetic resonance imaging. Having a scanner interface presents the opportunity of superimposing the dosimetric information on the anatomical images. It is a simple matter to superimpose dosimetric and anatomic information on a series of images. In most situations, the anatomical images are parallel to one another. The treatment planning computer calculates the doses to a three-dimensional grid of

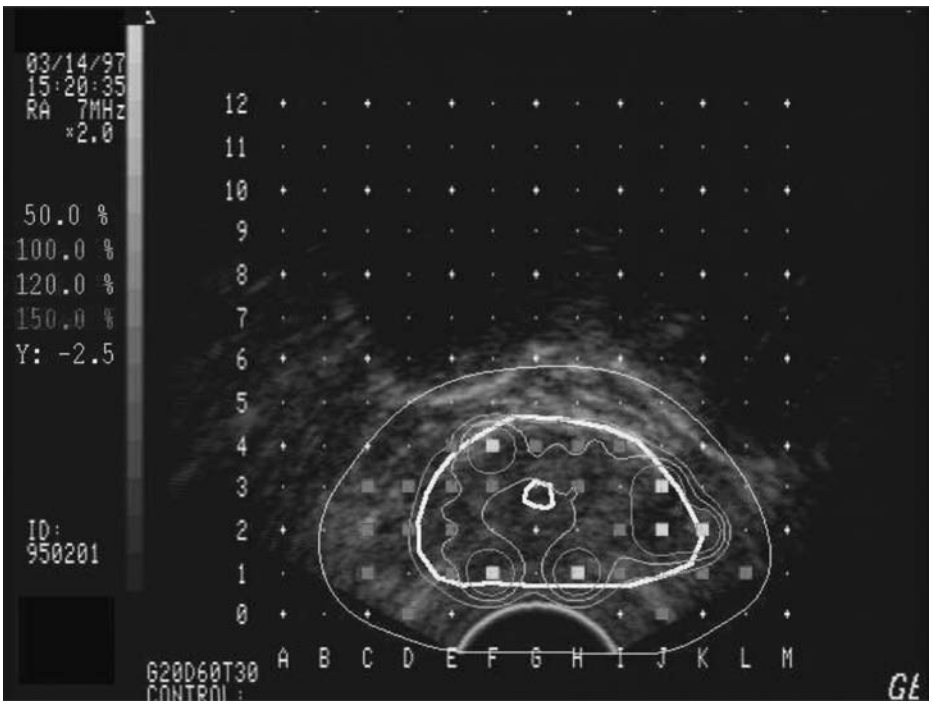


Figure 3. Isodoses superimposed on an ultrasound image of the prostate. The thick white contour line delineates the prostate. The inner isodose lines depict higher dose levels than outer contours. Note the conformity of the green prescription isodose with the shape of the prostate. The inner white contour is the urethra. Note avoidance of the 120% isodose contour with the urethra, a radiosensitive critical structure.

points spatially registered with the anatomic images. After the dosimetric calculations are completed, the dose distributions are represented as a series of colored contour lines commonly known as isodoses. Isodoses are analogous to the contour lines of a mountain on a topographical map. In the case of a mountain, the contour lines represent locations of equal elevation. In a similar manner, isodoses represent the locus of points of equal dose. An example of isodose contour lines superimposed on an axial ultrasound of the prostate is shown in Fig. 3.

In Fig. 3, the outermost yellow contour line represents 50% of the prescribed dose. This is the lowest dose level shown in the figure. Similarly, the innermost lavender contour represents 150% of the prescribed dose. Usually, but not always, the inner isodoses represent higher doses than the outer contours. On a topographical map of a mountain, the inner contour lines represent higher elevations than the outer contours.

Viewing isodose contours superimposed on anatomical images provides a means for visually evaluating the conformity of the dose distributions with the target anatomy. Ideally, the prescription isodose line should exactly conform to the shape of the target volume. This goal is not generally achievable although it can be approached. In Fig. 3, the green contour is the prescription isodose line. There is close conformity of the green isodose line with the prostate. In addition to being able to visually evaluate the conformity of the prescription isodose with the anatomy, it is also possible to ascertain doses received to critical structures. In the case of the prostate, the critical structures where excessive doses should be avoided are the rectum and urethra. Referring to Fig. 3, it is observed that the urethra receives less than 120% of the prescription dose. To properly evaluate a treatment plan, the isodoses on all images must be reviewed. Some slices may manifest excellent conformity and satisfactory critical structure doses, whereas reviewing other images may reveal unsatisfactory dose distributions.

In traditional prostate brachytherapy treatment planning, the physicist manually selects source and needle locations in an attempt to optimize the treatment plan. This entails maximizing the conformity of the prescription isodose with the prostate and minimizing the doses received by the critical structures. This is a tedious, time-consuming process because the conformity and critical structure conditions must be simultaneously met on all slices. Often, improvement on one slice degrades the dosimetry on another. There is no universal standard regarding what constitutes an acceptable treatment plan. There are differing opinions both among individual physicians and treatment centers. These differences generally pertain to acceptable dose values covering the prostate and critical structures. Despite these differences, however, the aim of treatment planning is to maximize the coverage of the prescription dose to the prostate and minimize the doses to the critical structures.

A popular graphical method for evaluating treatment plans is known as a cumulative dose volume histogram, commonly referred to as a DVH. A cumulative DVH is a graph of the volume of a target or critical structure as a function of minimum dose received by the volume. A typical cumulative DVH plot is shown in Fig. 4. Referring to this figure, it is observed that at low doses, the volume covered is essentially 100%. This shows that the target as well as the adjacent critical structures receive a significant dose. In Fig. 4, the blue graph represents the DVH for the urethra. It is observed in this graph that around 95% of the urethra receives doses exceeding 50% of the prescription. As the urethra traverses the prostate, it is physically impossible for the urethra not to receive a significant dose

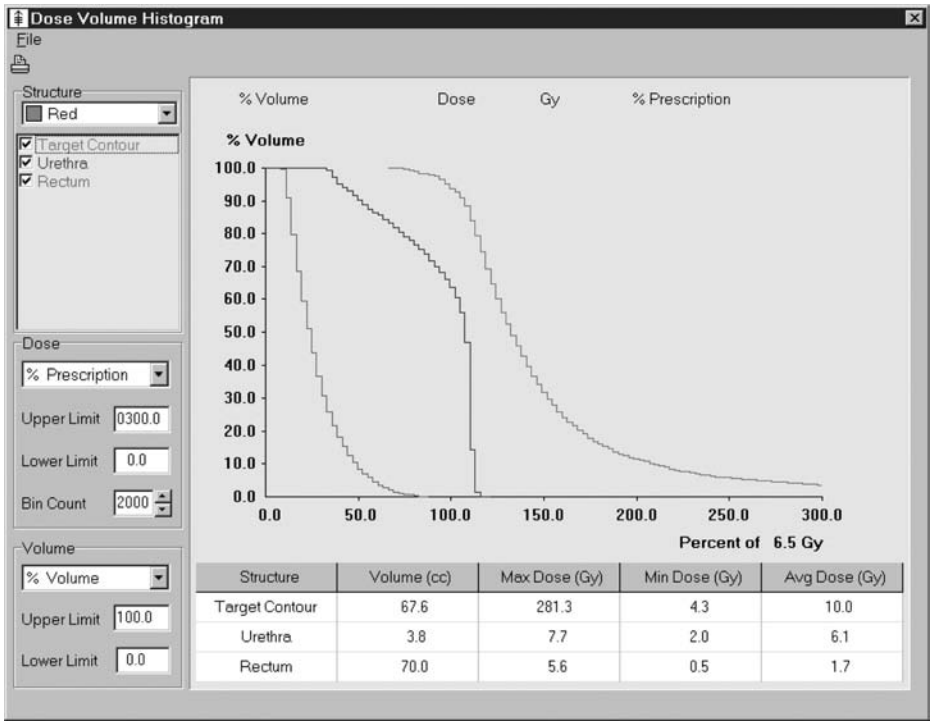


Figure 4. Typical cumulative DVH for a prostate implant.

without sacrificing the dose to the prostate. In well-designed treatment plans, the urethral doses are minimized. The specific criteria depend on the preferences of the physician or institution.

As was briefly indicated, manually optimizing prostate treatment plans is a time-consuming process. To address this issue, the authors developed an automated treatment planning and optimization system. The optimizer used in this system is based on a genetic algorithm (GA) (13,19). GAs are a genre of optimization algorithms that attempt to find the optimal solution of a problem based on evolutionary natural selection (31–34). In the case of the intraoperative optimizer, a population of bit streams is used to represent a population of treatment plans. Each bit stream is a one-dimensional array of numbers whose values are either one or zero. The length of the arrays is equal to the total number of potential source locations in the implant. The three-dimensional coordinates of each potential source location are known and fixed. If a source is present at a potential site, the bit for the corresponding site location is one. Conversely, if a source is not present at a potential site, the corresponding bit in the stream is set to zero. In this way, the treatment planning problem is encoded in a form amenable to genetic optimization.

To determine "how good" is the plan represented by a bit stream, an objective function is defined. The job of the objective function is to calculate a score, representing a figure of merit. The larger the score, the better the treatment plan, as determined by the criteria defined in the objective function. For instance (19), let PD be the prescription dose; the *prostate score* is the number of (uniformity) points that satisfy $\text{PD} \leq D \leq 1.6$ PD, the *urethral score* is the number of points in the urethra for which $D \leq 1.2$ PD, and the *rectal score* sums up all points in the rectum for which $D \leq 0.78$ PD. From this one obtains a *raw score*:

$$\text{Raw score} = 5 \times (\text{prostate score}) + 35 \times (\text{rectal score}) + 50 \times (\text{urethral score})$$

and a *final score* (used in the optimization algorithm), which is a linear function of the raw score [$=A \times$ (raw score) $+ B$].

The objective function is defined to increase the score for greater coverage of the prescription dose to the prostate. On the other hand, as the doses to the rectum and urethra increase, the score is decreased. As can be observed, there are conflicting requirements and the score reflects all of these conditions.

At the beginning of execution, the bit streams are randomly assigned arbitrary bit patterns. During execution of the GA, these bit streams are gradually "evolved" into higher scoring streams, indicating improvements in the treatment plans they represent. In an attempt to improve the scores of a population of treatment plans, the GA mimics biological evolutionary processes. Crossover, mutation, and survival-of-the-fittest are simulated by the GA in an attempt to maximize the scores received by the population. Two bit streams in the population are selected to act as "parents." Crossover is implemented by interchanging corresponding randomly selected bits between the parents. Mutation is affected by inverting a very small ($\sim$1%) number of bits in the two bit streams. The modified bit streams are evaluated and replace the two lowest-scoring chromosomes in the population.

In biological evolution, the fitter species has a higher likelihood of survival compared with the lesser fit species. This means that although the lesser fit species has a nonzero probability of survival, species that are better suited to their environments have higher survival probabilities.

Natural selection is simulated in the GA by favoring higher scoring bit streams to act as parents. In 50% of the parent selections, the two highest scoring bit streams are selected to act as parents. In the other 50% of the selections, two bit streams are randomly selected, irrespective of their score. The parents undergo crossover and mutation and replace the two lowest-scoring bit streams in the population. After this process is repeated several thousands of times, a population of high scoring bit streams is generated. This in turn translates into a population of optimized treatment plans. The highest scoring treatment plan is selected to be used for the implant.

The GA can generally generate a treatment plan in under 5 min. Actual execution times depend on the size of the prostate and the number of needles used for implanting the radioactive sources.

Other optimization algorithms used in prostate brachytherapy are simulated annealing (35,36) and branch-and-bound (13–15).

SECONDARY DOSE VERIFICATION

Once the plan is approved by the clinician a second physicist must independently verify and formally approve the plan. An estimate of the total number of seeds can be obtained with the following equations, which give—for a given average dimension of the volume to be implanted, d_{avg}—the total source strength S_K required (37):

For a ^{125}I *permanent* implant and a prescription dose of 144 Gy:

$$\frac{S_k}{U} = \begin{cases} 5.709 \left(\dfrac{d_{\text{avg}}}{\text{cm}}\right) & d_{\text{avg}} \le 3\,\text{cm} \\ 1.524 \left(\dfrac{d_{\text{avg}}}{\text{cm}}\right)^{2.2} & d_{\text{avg}} > 3\,\text{cm} \end{cases} \tag{15}$$

For a ^{103}Pd *permanent* implant and a prescription dose of 140 Gy:

$$\frac{S_k}{U} = \begin{cases} 29.41 \left(\dfrac{d_{\text{avg}}}{\text{cm}}\right) & d_{\text{avg}} \le 3\,\text{cm} \\ 5.395 \left(\dfrac{d_{\text{avg}}}{\text{cm}}\right)^{2.56} & d_{\text{avg}} > 3\,\text{cm} \end{cases} \tag{16}$$

One evaluates the total number of radioactive sources needed by dividing the total required source strength S_K by the single-seed strength used in that implant. In general, one seeks agreement of 10% or better between the planned and the expected number of seeds.

The plan verification must also determine that the correct prescription dose was used and that input data to the planning software was correctly entered.

DOSE ESCALATION TO PROSTATE SUBVOLUMES

If information is available on tumor burden in specific subvolumes of the prostate, dose escalation to these voxels is recommended. For instance, it has been hypothesized that regions of the prostate where the choline/citrate ratio, as determined by magnetic resonance spectroscopy (MRS), is elevated may contain clinically significant cancer (38–42). In general, "clinically significant" refers to cancer cells that are fast proliferating and/or radioresistant (features associated with local failure) or of high grade and thus with a potentially larger probability of distant dissemination. There is limited evidence of a correlation between choline levels and histological grade (Gleason score). As well, biochemical arguments have been invoked to support the expectation that a larger value of this ratio is expected to reflect an increased rate of cell proliferation although no direct proof exists yet.

As applied at MSKCC, the implementation of an MRS-based dose escalation amounts to increasing the dose to the MRS-positive voxels to 200% of the prescription dose (with no upper limit) while keeping the urethral and rectal dose within the usual range of constraints (14).

POST-IMPLANT ANALYSIS

At MSKCC, post-implant evaluation is performed the day of the implant. The number of seeds implanted is confirmed with a pair of planar radiographic films. A CT study (3 mm slices) is then obtained, and anatomical structures are marked out on each slice. The coordinates of the center of each seed are determined by using appropriate computer software (Interplant Post-implant Analysis System, Version 1.0: Burdette Medical Systems, Inc., Champaign, IL). With the seeds thus identified, (DVHs) are calculated for each structure of interest and compared with the original plan.

CLOSING WORDS: KNOWN PROBLEMS AND POSSIBLE FIXES

Permanent brachytherapy prostate implants are now a well-accepted treatment modality for early stage prostate cancer. The two major limitations of this procedure are higher incidence of urethral complications (when compared with external-beam radiotherapy) and — for some patients — lower than prescribed delivered doses. In this section, we list several unresolved issues that may be responsible for this state of affairs and suggest possible solutions.

Post-implant evaluations of permanent prostate implants often indicate significant differences between the intended plan and its actual implementation. Although an experienced physician can minimize the magnitude of these differences, many factors controlling execution of the plan (e.g., bleeding, tissue swelling) are subject to random fluctuations. This often leads to a higher than intended dose to urethra and rectum and/or lower or higher doses to the prostate, especially at the periphery of the gland. In our view, this discrepancy represents the most important obstacle and challenge that currently needs to be overcome to achieve consistent application of a low urethral and rectal dose range and thereby reduce morbidity after prostate brachytherapy. In a series of recent articles the concept of intraoperative dynamic dosimetric optimization has been proposed (43–46). The idea is to re-optimize the plan several times during the implantation based on the actual

positions of the seeds already implanted. The key problem is obtaining in real time (and within a reasonable time interval — 5 min or less) the coordinates of the implanted seeds in the system of reference used for planning. Visualization of the actual seed positions on the intraoperative ultrasound image is difficult, if not impossible, to achieve because of significant artifacts noted on the ultrasound image from needles and/or hemorrhage within the gland.

One can reconstruct seed coordinates from fluoroscopic images taken at three different angles (43,44) or, equivalently, from a CT study obtained in the OR for instance, using a C-arm with CT capabilities. CT-based systems that perform seed segmentation do exist (e.g., Variseed from Varian Medical Systems, Inc, Charlottesville, VA; Interplant Post-implant Analysis System, Burdette Medical Systems, Inc., Champaign, IL), but at this time they do not seem to have the capability of performing this task on the fly and at the same time maintain the required seed-detection reliability.

A second problem concerns the effect of changes in prostate volume (edema shrinkage) as well as seed migration after implantation and the effect this has on a treatment plan that is based on the geometry of the target at the time of implantation. A method of planning that incorporates temporal changes in the target-seed configuration during dose delivery and makes use of the concept of effective volume has been developed by Lee and Zaider (47).

The preceding enumeration of problems has been brief and (admittedly) selective, but we hope to motivate the reader to take a careful look at these important issues. The desideratum of dosimetric conformality in permanent prostate implants remains a topic of active interest in the brachytherapy community, and no doubt the last word on this subject has not yet been spoken.

BIBLIOGRAPHY

1. Coen JJ, Zietman AL, Thakral H, Shipley WU. Radical radiation for localized prostate cancer: Local persistence of disease results in a late wave of metastases. J Clin Oncol 2002;20:3199–3205.
2. Zagars GK, vonEschenbach AC, Ayala AG, Schultheiss TE, Sherman NE. The influence of local-control on metastatic dissemination of prostate-cancer treated by external beam megavoltage radiation-therapy. Cancer 1991;68:2370–2377.
3. Valicenti R, Lu JD, Pilepich M, Asbell S, Grignon D. Survival advantage from higher-dose radiation therapy for clinically localized prostate cancer treated on the radiation therapy oncology group trials. J Clin Oncol 2000;18:2740–2746.
4. Logothetis C. Challenge of locally persistent prostate cancer: An unresolved clinical dilemma. J Clin Oncol 2000;20:3187.
5. Nag S, Shasha D, Janjan N, Petersen I, Zaider M. The American Brachytherapy Society recommendations for brachytherapy of soft tissue sarcomas. Int Radiat Oncol Biol Phys 2000;49:1033–1043.
6. D'Amico AV, Whittington R, Malkowicz SB, Fondurulia J, Chen MH, Kaplan I, Beard CJ, Tomaszewski JE, Renshaw AA, Wein A, Coleman CN. Pretreatment nomogram for prostate-specific antigen recurrence after radical prostatectomy or external-beam radiation therapy for clinically localized prostate cancer. J Clin Oncol 1999;17:168–172.
7. Moul JW, Connelly RR, Lubeck DP, Bauer JJ, Sun L, Flanders SC, Grossfeld GD, Carroll PR. Predicting risk of prostate specific antigen recurrence after radical prostatectomy with the center for prostate disease research and cancer of the prostate strategic urologic research endeavor databases. J Urol 2001;166:1322–1327.
8. Albertsen PC, Hanley JA, Gleason DF, Barry MJ. Competing risk analysis of men aged 55 to 74 years at diagnosis managed conservatively for clinically localized prostate cancer. JAMA 1998;280:975–980.
9. Chodak GW. Comparing treatments for localized prostate cancer-persisting uncertainty. JAMA 1998;280:1008–1010.
10. Jani AB, Hellman S. Early prostate cancer: Clinical decision-making. Lancet 2003;361:1045–1053.
11. Sandhu AS, Zelefsky MJ, Lee HJ, Lombardi D, Fuks Z, Leibel SA. Long-term urinary toxicity after 3-dimensional conformal radiotherapy for prostate cancer in patients with prior history of transurethral resection. Int J Radiat Oncol Biol Phys 2000;48:643–647.
12. Zelefsky MJ, Hollister T, Raben A, Matthews S, Wallner KE. Five-year biochemical outcome and toxicity with transperineal CT-planned permanent I-125 prostate implantation for patients with localized prostate cancer. Int J Radiat Oncol Biol Phys 2000;47:1261–1266.
13. Lee EK, Gallagher RJ, Silvern D, Wuu CS, Zaider M. Treatment planning for brachytherapy: An integer programming model, two computational approaches and experiments with permanent prostate implant planning. Phys Med Biol 1999;44:145–165.
14. Zaider M, Zelefsky MJ, Lee EK, Zakian KL, Amols HI, Dyke J, Cohen G, Hu Y, Endi AK, Chui C, Koutcher JA. Treatment planning for prostate implants using magnetic-resonance spectroscopy imaging. Int J Radiat Oncol Biol Phys 2000;47:1085–1096.
15. Gallagher RJ, Lee EK. Mixed integer programming optimization models for brachytherapy treatment planning. Proc/AMIA Annu Fall Symp 1997; 278–282.
16. Lee EK, Gallagher RJ, Silvern D, Wuu CS, Zaider M. Treatment planning for brachytherapy: An integer programming model, two computational approaches and experiments with permanent prostate implant planning. Phys Med Biol 1999;44:145–165.
17. Lee EK, Zaider M. Mixed integer programming approaches to treatment planning for brachytherapy. Ann Operat Res Optimizat Med Ann Operat Res 2002;119:147–163.
18. Lee EK, Zaider M. Intraoperative dynamic dose optimization in permanent prostate implants. Int J Radiat Oncol Biol Phys 2003;56:854–861.
19. Silvern DA. Automated OR prostate brachytherapy treatment planning using genetic optimization. 1998.
20. Nag S, Shasha D, Janjan N, Petersen I, Zaider M. The American Brachytherapy Society recommendations for brachytherapy of soft tissue sarcomas. Int J Radiat Oncol Biol Phys 2001;49:1033–1043.
21. Zelefsky MJ, Cohen G, Zakian KL, Dyke J, Koutcher JA, Hricak H, Schwartz L, Zaider M. Intraoperative conformal optimization for transperineal prostate implantation using magnetic resonance spectroscopic imaging. Cancer J 2000;6:249–255.
22. Potters L. Permanent prostate brachytherapy: Lessons learned, lessons to learn. Oncol-New York 2000;14:981–991.
23. Cha CM, Potters L, Ashley R, Freeman K, Wang XH, Waldbaum R, Leibel S. Isotope selection for patients undergoing prostate brachytherapy. Int J Radiat Oncol Biol Phys 1999;45:391–395.
24. Ling CC, Li WX, Anderson LL. The relative biological effectiveness of I-125 and Pd-103. Int J Radiat Oncol Biol Phys 1995;32:373–378.

25. Wuu CS, Zaider M. A calculation of the relative biological effectiveness of 125I and 103Pd brachytherapy sources using the concept of proximity function. Med Phys 1998;25:2186–2189.
26. Nath R, Anderson LL, Luxton G, Weaver KA, Williamson JF, Meigooni AS. Dosimetry of interstitial brachytherapy sources - recommendations of the aapm radiation-therapy committee task group no 43. Med Phys 1995;22:209–234.
27. Handbook of Chemistry and Physics. Boca Raton, FL: CRC Press; 1981.
28. Huda W, Slone R. Review of Radiologic Physics, 1995.
29. Rivard MJ, Butler WM, DeWerd LA, Huq MS, Ibbott GS, Li ZF, Mitch MG, Nath R, Williamson JF. Update of AAPM task group No. 43 report: A revised AAPM protocol for brachytherapy dose calculations. Med Phys 2004;31:3532–3533.
30. Williamson JF, Butler W, DeWerd LA, Huq MS, Ibbott GS, Li Z, Mitch MG, Nath R, Rivard MJ, Todor D. Recommendations of the American Association of Physicists in Medicine regarding the impact of implementing the 2004 task group 43 report on dose specification for Pd-103 and I-125 interstitial brachytherapy. Med Phys 2005;32:1424–1439.
31. Lance Chambers, Practical Handbook of Genetic Algorithms Boca Raton, FL: CRC Press; 1995.
32. Grefenstette JJ. American Association for Artificial Intelligence, Beranek a. N. i. Bolt, Naval Research Laboratory (U.S.), Genetic Algorithms and Their Applications Proceedings of the Second International Conference on Genetic Algorithms, July 28-31, 1987 at the Massachusetts Institute of Technology. Cambridge, MA: Hillsdale, NJ; 1987.
33. Man KF, Tang KS, Kwong S. Genetic Algorithms Concepts and Designs. London: 1999.
34. Zalzala AMS, Fleming PJ. Genetic Algorithms in Engineering Systems. London: 1997.
35. Sloboda RS. Optimization of brachytherapy dose distribution by simulated annealing. Med Phys 1992;19:964.
36. Pouliot J, Tremblay D, Roy J, Filice S. Optimization of permanent I-125 prostate implants using fast simulated annealing. Int J Radiat Oncol Biol Phys 1996;36:711–720.
37. Cohen GN, Amols HI, Zelefsky MJ, Zaider M. The Anderson nomograms for permanent interstitial prostate biplants: A briefing for practitioners. Int J Radiat Oncol Biol Phys 2002;53:504–511.
38. Wefer AE, Hricak H, Vigneron DB, Coakley FV, Lu Y, Wefer J, Mueller-Lisse U, Carroll PR, Kurhanewicz J. Sextant localization of prostate cancer: Comparison of sextant biopsy, magnetic resonance imaging and magnetic resonance spectroscopic imaging with step section histology. J Urol 2000;164: 400–404.
39. Kurhanewicz J, Vigneron DB, Males RG, Swanson MG, Yu KK, Hricak H. The prostate: MR imaging and spectroscopy — Present and future. Radiol Clin North Am 2000;38:115.
40. Scheidler J, Hricak H, Vigneron DB, Yu KK, Sokolov DL, Huang LR, Zaloudek CJ, Nelson SJ, Carroll PR, Kurhanewicz J. Prostate cancer: Localization with three-dimensional proton MR spectroscopic imaging — Clinicopathologic study. Radiology 1999;213:473–480.
41. Kurhanewicz J, Vigneron DB, Hricak H, Narayan P, Carroll P, Nelson SJ. Three-dimensional H-1 MR spectroscopic imaging of the in situ human prostate with high (0.24-0.1-cm(3)) spatial resolution. Radiology 1996;198:795–805.
42. Kurhanewicz J, Vigneron DB, Nelson SJ, Hricak H, MacDonald JM, Konety B, Narayan P. Citrate as an in-vivo marker to discriminate prostate-cancer from benign prostatic hyperplasia and normal prostate peripheral zone — detection via localized proton spectroscopy. Urology 1995;45:459–466.
43. Todor DA, Cohen GN, Amols HI, Zaider M. Operator-free, film-based 3D seed reconstruction in brachytherapy. Phys Med Biol 2002;47:2031–2048.
44. Todor DA, Zaider M, Cohen GN, Worman MF, Zelefsky MJ. Intraoperative dynamic dosimetry for prostate implants. Phys Med Biol 2003;48:1153–1171.
45. Tubic D, Zaccarin A, Pouliot J, Beaulieu L. Automated seed detection and three-dimensional reconstruction. I. Seed localization from fluoroscopic images or radiographs. Med Phys 2001;28:2265–2271.
46. Tubic D, Zaccarin A, Beaulieu L, Pouliot J. Automated seed detection and three-dimensional reconstruction. II. Reconstruction of permanent prostate implants using simulated annealing. Med Phys 2001;28:2272–2279.
47. Lee EK, Zaider M. On the determination of an effective planning volume for permanent prostate implants. Int J Radiat Oncol Biol Phys 2001;49:1197–1206.

See also BRACHYTHERAPY, HIGH DOSE RATE; NUCLEAR MEDICINE INSTRUMENTATION.

PTCA. See CORONARY ANGIOPLASTY AND GUIDEWIRE DIAGNOSTICS.

PULMONARY MECHANICS. See RESPIRATORY MECHANICS AND GAS EXCHANGE.

PULMONARY PHYSIOLOGY

JOHN DEMENKOFF
Mayo Clinic, Dept. of Anesthesia
Scottsdale, Arizona

INTRODUCTION

Present day pulmonary function testing is available in all hospitals and in a less sophisticated form in many physicians' offices. Such was not the case until the 1940s, when the fruits of physiological research dating back 150 years blossomed on the heels of World War II. This so-called golden age of pulmonary physiology spurred many of the currently available lung function tests which are used for diagnosis and treatment of existing lung disease; screening for early pulmonary disease; evaluation of respiratory symptoms such as cough and shortness of breath; performance of disability evaluations; preoperative assessment of thoracic and other surgical patients; determination of level of cardiopulmonary fitness; monitoring of adverse pulmonary effect of certain drug therapies.

Over the years, many have contributed to an understanding of the lung and how it works in health as well as in disease. These discoveries have provided building blocks of knowledge which form the basis of current modern pulmonary function testing.

PRE-1940S

Leonardo DaVinci: This genius drew detailed anatomical illustrations clearly depicting a bellows function of the respiratory muscles.

John Malysed: In 1674, he constructed a model of the chest and lungs with a bladder enclosed inside a

simple bellows with the neck outside. With a glass plate on one side, one could watch the bladder inflate and deflate when the bellows operated.

John Hutchinson: In 1848, he developed a spirometer and measured the vital capacity in thousands of normal subjects (1). He also differentiated normal and abnormal results quantitatively, thus ushering in a diagnostic use for pulmonary testing.

Humphrey Davey: Discoverer of hydrogen gas in the early 1800s. This led the way for measuring various lung volumes and compartments other than Hutchinson's vital capacity. Davey built his own spirometer, filled it part way with hydrogen, and breathed it back and forth "for seven quick breaths", finally exhaling fully into the spirometer. Then by measuring the amount and concentration of hydrogen in the spirometer and assuming an equal concentration in his lungs, he calculated the amount of air in his lungs at the end of full exhalation, known today as the residual volume. Modern day lung volume determinations use the inert gas helium with a slightly different protocol, but the fundamental principles remain the same.

Marie Krogh: Prior to 1915, many eminent physiologists believed that oxygen was actively secreted by the lungs into the blood stream. Marie Krogh challenged this popular notion with her diffusion experiments using carbon monoxide. She devised a single breath test in which a subject first fully exhaled to residual volume, then inspired deeply from a spirometer containing 1% carbon monoxide. After an initial exhalation and a six second breath hold, the subject completed a full exhalation. Krogh measured the alveolar gas before and after the six-second breath hold and calculated the uptake of carbon monoxide by the bloodstream.

The amount of CO transfer was noted to be entirely by the process of diffusion and proportional to the pressure differential across the alveolar capillary membrane ($P_1 - P_2$). Because CO binds so tightly to the hemoglobin molecule, P_2 is small. The driving pressure P_1 can be easily calculated. Krogh took advantage of these factors in devising her test, which confirmed the importance of diffusion, not secretion, in the lung.

For many reasons open to speculation, the importance of Krogh's work was not fully appreciated nor developed clinically until the 1940s. The reader is referred to a delightful discourse on such medical curiosities in Ref. 2.

POST-1940S AND THE GOLD AGE OF PULMONARY PHYSIOLOGY

What occurred in the 1940s was a combination of intellect and pluck driven by military contracts and government funds. The resulting research and understanding of lung physiology paved the way for development of current-day pulmonary function laboratories. Some of the more brilliant, resourceful and ingenious researchers of this time are listed below.

Julius Comroe

Chairman, Department Physiology and Pharmacology, University of Pennsylvania, 1946–1957. Director of Cardiovascular Research Institute, University of California, San Francisco, 1957–1983. At both of these institutions, Dr. Comroe developed and fostered world-renowned faculty who studied multiple facets of pulmonary physiology. While at the University of Pennsylvania, Comroe demonstrated his ingenuity by adapting a used surplus bomber nose cone as a body plesthymograph. He wanted to apply Boyle's Law to the measurement of lung volumes, air flow, and airway resistance. His work ushered in modern-day plethysmography. His text, *Physiology of Respiration* (2) remains a classic.

Herman Rahn, Wallace O. Fenn, Arthur Otis

These remarkable men formed the core of a research effort at the University of Rochester. An account of this creative ground work is found in Ref. 3, and is rich in historical facts. In the 1940s, pneumotachographs had to be fabricated by individual research groups. In the Rochester group's first model, a cluster of soda straws encased in a brass tube served as the flow resistance element. In later versions, they used as resistive elements glass wool enclosed in a lady's hair net. Their contributions are evident today, as many of their postdoctoral fellows and research associates have gone on and taught the next generation of pulmonary specialists.

Andre Frederick Cournand, Dickinson Woodrow Richards

Both rshared shared the 1956 Nobel Prize in medicine and physiology, and formed the famous Bellevue Hospital Cardiopulmonary Lab at Columbia University. Their observations regarding prolonged nitrogen washout in the lungs of emphysematous patients fostered the clinical use of diagnostic pulmonary function tests. They also established normal values and formulated testing protocols. They pioneered catheterization of the right heart, making way for analysis of mixed venous blood and more accurate cardiac output and pulmonary blood flow via the direct Fick technique.

$$\text{Pulmonary blood flow} = \frac{\text{O}_2 \text{ consumption}}{\text{Arterial-mixed venous O}_2 \text{ difference}}$$

Current interventional cardiology, and the understanding of complex interrelatedness of pulmonary diseases on the heart, stem from these studies done in the 1950s at Columbia.

Since the mid-1960s, pulmonary function testing has evolved more slowly. Tests that are reproducible, well tolerated by patients, and offer helpful clinical information have been further refined by advances in instrumentation and computerization.

With the advent of rapidly responding gas analyzers, highly accurate and calibrated pneumotachographs, and sophisticated computer software, the study of lung function during exercise has become possible. The complex interactions of metabolic-cardiopulmonary systems is discussed below, in the section on exercise physiology. While a boon to performance-minded athletes, these tests also shed light on limitation of exercise tolerance due to diseases of the heart and lung.

From the time of DaVinci to the present, great strides have been made in the understanding of lung function and its measurement. Now simple acts, such as blowing out a candle or coughing, are known to be dependent on elastic recoil of the lungs and complex airways dynamics. Both properties of the lung are measured with pulmonary function testing.

Each test discussed in the following text carries with it a rich historical and intellectual story line.

PHYSIOLOGICAL PRINCIPLES UNDERLYING MODERN PULMONARY FUNCTION TESTS

The interpretation and analysis of pulmonary function tests is often conveyed in physiological terms rather than as specific medical diagnoses. As such important underlying physiological concepts are presented that will provide a deeper understanding of pulmonary function test results. Many of these concepts have been developed and refined over time and represent a legacy of scientific achievement.

SINGLE BREATH NITROGEN WASHOUT

Aptly named, this test measures the nitrogen concentration of a normal exhalation after a deep inhalation of 100% oxygen. It was developed by Fowler in 1948 to measure the anatomic dead space $V_{\text{danatomical}}$.

During normal tidal breathing, a part of each breath remains in the conducting airways of the upper airway and tracheobronchial tree. It never reaches the alveoli; therefore it does not participate in gas exchange and is referred to as anatomical dead space. The fraction of total ventilation (O_E) that reaches gas exchanging space of alveoli is called alveolar ventilation or O_A.

$$O_A = O_E - f \times V_{\text{danatomical}}$$
$$\text{where } f = \text{respiratory frequency}$$

In Fowlers method (Fig. 1), a simultaneous recording of nitrogen and exhaled volume is made after a deep inhalation of pure oxygen.

At the start of expiration, the gas comes from the anatomical dead space, which contains no nitrogen. Along the course of the S-shaped N_2 washout, a front between the alveolar air and dead space air can be determined (see Graphical depiction below).

The anatomical dead space is related to body weight and is $\sim$ 150 mL for a normal man. The extrathoracic fraction, mouth and pharynx, contributes 66 mL with a range of 35–105 mL, depending on jaw and neck position. Anatomic dead space represents an inefficiency of the design of the respiratory system. With each breath, a significant volume of air must be moved, requiring work, for which no benefit is derived.

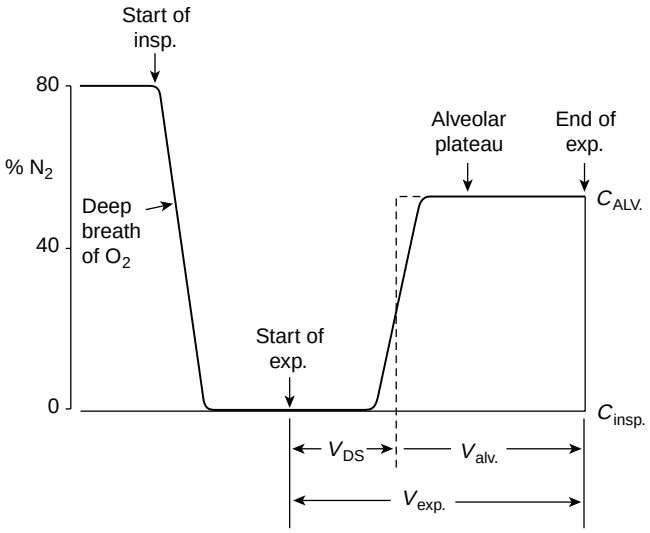


Figure 1. Fowler's method. For determination of anatomic dead space. See text for description. The flat portion of the curve is called the alveolar plateau and represents pure alveolar gas.

PHYSIOLOGICAL DEAD SPACE BOHR EQUATION

Inefficiencies also occur at the level of the alveoli where some air reaches the alveoli, but gas exchange never occurs. For example, the upper lobes in a normal resting upright lung are well-ventilated, but not perfused.

This is wasted ventilation, and when added to the anatomic dead space is designated as the physiological dead space. This can be measured by application of the Bohr equation (4). If one complete expiration is collected in a bag, the amount of carbon dioxide is $[F_ECO_2 \times V_T]$. This volume of CO_2 comes partly from the nonexchanging dead space, which has a volume from the inspired air $[V_D \times F_ICO_2]$, plus the volume from alveolar gas $F_ACO_2 \times [V_T - V_D]$.

$$[F_ECO_2 \times V_T] = F_ICO_2 \times V_D + F_ACO_2[V_T - V_D]$$

$$F_ACO_2 = \frac{V_T \times F_ECO_2 - V_D F_ICO_2}{V_T - V_D}$$

$$V_D = \frac{[F_ACO_2 - F_ECO_2]V_T}{F_ACO_2 - F_ICO_2}$$

If inspired, CO_2 is zero then $F_ICO_2 = 0$.

Hence,

$$V_D = \frac{F_ACO_2 - F_ECO_2}{F_ACO_2} \quad V_T$$

$$\frac{V_D}{V_T} = \frac{F_ACO_2 - F_ECO_2}{F_ACO_2}$$

F_ACO_2 = Alveolar CO_2 fractional concentration measured by obtaining an alveolar sample.

F_ECO_2 = Mixed expired CO_2 fractional concentration measured from a collection of expired air; Douglas bag or mixing chamber.

The parameter V_D/V_T is called the dead space to tidal volume ratio where the dead space is physiologic and includes the anatomic dead space. It is an efficiency rating, typically 0.3 in normals and up to 0.5 or so in patients with emphysema. In the latter case, 50% of the breath is wasted and does not participate in gas exchange.

FORCED EXPIRATION AND DYNAMIC COMPRESSION

The most familiar maneuver that utilizes a maximal expiratory effort is a cough. The resulting dynamic airway compression facilitates clearance of bronchial secretions. Even at maximal exercise, such flow rates are not attained, thus demonstrating an impressive reserve in flow characteristics of the lung.

The complex mechanics of forced exhalations were elucidated by the work of Hyatt, Schilder, and Fry. They described a maximal expiratory flow volume curve, where instantaneous expiratory flow is plotted against volume instead of time (as is done with FEV_1). Flow reaches a maximum at 80% of vital capacity and reaches zero at residual volume. The curve is shown to be effort-dependent >75% of VC and effort-independent <75%.

Once dynamic compression occurs, the lung behaves as a Starling Resistor (Fig. 2). The flow then depends on the elastic recoil of the lung and airway resistance upstream from the compressed lung segment. Under these conditions, an increase in effort produces no increase in flow.

DIFFUSION AND DIFFUSING

The purpose of the lung is to deliver oxygen to the blood stream and remove the byproduct of metabolism, carbon dioxide. This process begins with mass transport of oxygen down conducting tubes of diminishing caliber called bronchi, bronchioles, terminal bronchioles, and finally air sacs or alveoli. Simple diffusion then occurs at the interface between the walls of the alveoli and pulmonary capillaries. The 300 million or so alveoli in the human lung

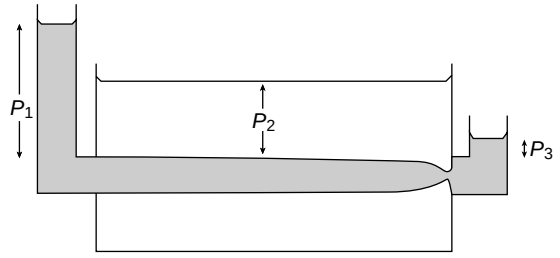


Figure 2. A Starling Resistor that is a mechanical analog for dynamic compression of the airways. The collapsible tubing in the chamber represents small airways. The pressure P_2 is pleural pressure during a forced vital capacity maneuver which collapses the airways at the equal pressure point, that is, $P_2 > P_3$. The pressure P_1 represents the elastic recoil of the lungs. Flow is proportional to $P_1 - P_2$.

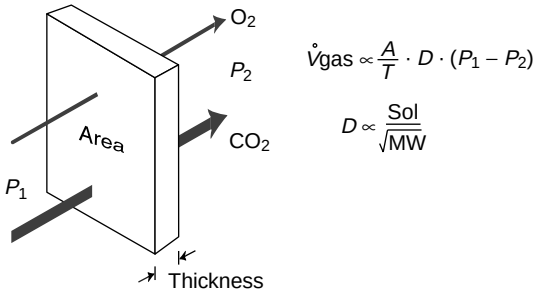


Figure 3. The process of diffusion and Fick's law of diffusion. Within the lung P_1 would represent the alveolar space and P_2 the capillary space.

create a surface area for diffusion of 85 m^2. The rat at which gas, either oxygen or carbon dioxide that transverses this membrane follows Fick's law of diffusion and is proportional to the surface area of the sheet and inversely proportional to its thickness.

The diffusion coefficient is proportional to the solubility of the gas and inversely proportional to the square root of the molecular weight (Fig. 3).

$$\dot{V}_{gas} = \frac{A}{T} \times D(P_1 - P_2)$$

$$D = \frac{Sol}{\sqrt{MW}}$$

where MW = molecular weight

When applying this formula to the special case of oxygen diffusion the partial pressure of oxygen in the alveolus (P_AO_2) or driving pressure (P_1) is 100 mmHg (13.2 kPa) while that in the pulmonary capillary (P_2) is 40 mmHg (5.3 kPa). The amount of oxygen transferred depends in part on this pressure differential ($P_1 - P_2$); 100 mmHg (13.2 kPa) minus 40 mmHg (5.3 kPa), which results in a diffusion gradient of 60 mmHg (7.9 kPa). In addition the thickness of the alveolar-capillary membrane (normally 0.3 μ) appears to be of equal importance. When this membrane is thickened by disease states, such as pneumonia, pulmonary fibrosis, asbestosis, or silicosis, oxygen transfer is considerably impaired.

The complexity of oxygen diffusion becomes apparent when one considers that red blood cells carrying the hemoglobin molecule typically spend only 0.75 s in the pulmonary capillary. Given a normal driving pressure (P_1–P_2) of 60 mmHg (7.9 kPa) and a healthy alveolar-capillary membrane (0.3 μm thick) equilibrium ($P_1 = P_2$) will occur in 0.25 s. In other words, as blood exits the gas exchange space it will have gone from a PO_2 of 40 mmHg (5.3 kPa)–100 mmHg (13.2 kPa) rapidly.

Various lung diseases can compromise the elegant process of diffusion outlined above. In some the alveolar oxygen level is reduced thereby diminishing the driving pressure. In others diffusion is impaired by a thickened alveolar capillary membrane. And finally, with exertion, as blood flow increases, the time available to load oxygen onto the hemoglobin molecule is reduced. Any one or combination of these factors can lead to a reduction in the diffusion of oxygen.

Having said this, the actual measurement of the diffusing capacity of the lung for oxygen as a clinically useful

pulmonary function test has proved difficult to develop. This is due to the fact that the capillary oxygen pressure is ever increasing as P_2 approaches P_1 creating a back pressure thus slowing diffusion as red cells travel along the pulmonary capillary bed. The parameter P_2 can be calculated through a very complex and cumbersome integration method. In the final analysis, oxygen transfer then is actually significantly dependent on total flow of pulmonary capillary blood rather than diffusion alone.

Unlike oxygen, carbon monoxide transfer is diffusion-limited because it binds so tightly to hemoglobin that the partial pressure of CO in pulmonary capillary blood is low. There is little back pressure, so the amount of carbon monoxide transferred is related only to the driving pressure P_1 ($P_2 \approx 0$), which is the alveolar pressure P_A of carbon monoxide.

$$\underline{O_{gas}} = \left(\frac{A}{T} \times D\right)(P_1 - P_2)$$
$$\frac{O_{gas}}{P_1} = \frac{A}{T} \times D$$

The above equation is simplified as follows, where D_L is called the diffused capacity of the lung and includes the area, thickness, and diffusing properties of the sheet and the gas concerned.

$$D_{LCO} = \frac{A}{T} \times D = \frac{O_{CO}}{P_1 CO} = \frac{O_{CO}}{P_A CO}$$
$$P_1 = \text{Driving pressure} = P_A CO = \text{Alveolar CO}$$
$$P_2 = 0$$
$$D_{LCO} = \frac{O_{CO}}{P_1 - P_2} \qquad D_{LCO} = \frac{O_{CO}}{P_A CO}$$

GAS LAWS

By convention, pulmonary function test results are expressed either at body temperature and ambient pressure, saturated (BTPS), ambient temperature and pressure, saturated (ATPS), or standard temperature and pressure, dry (STPD). A working knowledge of gas laws is essential for accurate conversion from one state to another.

Gas inside the lung is at BTPS. The pressure is the barometric pressure (P_B), and saturated refers to the saturated water vapor pressure (P_W), which is a function of temperature. At normal body temperature (37 °C), P_W is 47 mmHg (6.2 kPa).

Gas measured in the equipment is at ambient temperature, dry (ATD) if the expired water vapor is absorbed prior to the measurement or if inspired gas is from a cylinder. Alternately, it is called ATPS if expired gas is collected, but the water vapor was not absorbed. At normal room temperature (25 °C), P_W is 22 mmHg (2.9 kPa).

Inspired gas from the atmosphere is ordinarily between ATPD and ATPS. Since buildings typically are at 50% relative humidity, P_W is 50% of 22 mmHg (1.4 kPa) or 11 mmHg (1.4 kPa).

Boyle first published his ideal gas law in 1662. It states that, for a given mass of gas at constant temperature, the volume varies inversely with the pressure:

$$PV = RT$$

Charles's law states that, for a given mass of gas at constant pressure, the volume varies directly with the absolute temperature. Thus V/T is a constant, where T designates a temperature on the absolute or kelvin scale.

Combining both these gas laws gives an approximation of real gases under various conditions.

$$\frac{P_1 V_1}{T_1} = \frac{P_2 V_2}{T_2}$$

Usually called the ideal gas law.

Suppose that a patient expires into a Douglas bag, which is then transferred to a laboratory at 20 °C and squeezed through a dry gas meter. From a knowledge of the number of expirations collected and the respiratory frequency, the volume of gas at 20 °C corresponding to the minute volume ventilation O_E can be calculated. If the minute volume was 6 L at ATPS then, by use of the combined gas law equation, volumes can be adjusted to BTPS and STPD.

Assume that the patient's body temperature is 37 °C, the saturated water vapor pressure is 18 mmHg (2.4 kPa) at 20 °C and 47 mmHg (6.2 kPa) at 37 ° C and that the barometric pressure is 760 mmHg (101 kPa). Because water vapor does not follow the ideal gas law, its partial pressure is subtracted.

$$\frac{(760 - 18)}{273 + 20} \times 6 = \frac{(760 - 47) \times V_2}{273 + 37}$$

So that

$$V_2 = \frac{742 \times 6 \times 310}{713 \times 293} = 6.61\,\text{L BTPS}$$

The same volume of 6 L measured under atmospheric conditions would represent 5.45 L under STPD, that is, 760 mmHg (101 kPa) and 0 °C.

$$\frac{(760 - 18)}{273 + 20} \times 6 = \frac{760}{273} \times V_3$$
$$V_3 = \frac{742 \times 6 \times 273}{760 \times 293} = 5.45\,\text{L ATPD}$$

BTPS is used for lung volumes and ventilation O_E, ATPS for maximal inspiratory and expiratory flow, and STPD for oxygen consumption and carbon dioxide output.

CALCULATION OF OXYGEN UPTAKE

Oxygen uptake is the difference between oxygen breathed in and the amount in the exhaled air.

$$O_{O_2} = (O_I \cdot F_I O_2) - (O_E \cdot F_E O_2)$$

Where O_{O_2} is the oxygen uptake in liter per minute; O_I is the inspired minute volume ($L \cdot min^{-1}$), $F_E O_2$ is the mixed expired oxygen fraction, and $F_I O_2$ is the inspired oxygen fraction. Because the volume of inspired air is slightly greater than expired air (more O_2 consumed than carbon

dioxide, CO_2, is produced), a correction factor using measured nitrogen is used.

$$O_I = O_E\left(\frac{F_EN_2}{F_IN_2}\right)$$

This has been attributed to the British researcher and is referred to as the Haldane Transformation. It is used to calculate the inspired volume when only O_E is measured, the latter being much easier to measure than the former.

Substituting this correction factor into the original equation,

$$O_{O_2} = \left(O_E \times \left(\frac{F_EN_2}{F_1N_2}\right) \times F_1O_2\right) - (O_E \times F_EO_2)$$

since $F_EN_2 = (1 - F_EO_2 - F_ECO_2)$, this becomes

$$O_{O_2} = \left(O_E \frac{(1 - F_EO_2 - F_ECO_2) \times 0.2093}{0.7904}\right) - O_E \times F_EO_2$$

reducing to

$$O_{O_2} = O_E((1 - F_EO_2 - F_ECO_2) \times 0.265) - (F_EO_2)$$

By convention, O_{O_2} is expressed under standard conditions (STPD).

During a standard cardiopulmonary exercise stress, all the variables on the right side of the equation are measured as follows:

O_E Douglas bag for collection
or
Pneumotachograph interfaced with a computer for exercise testing

F_EO_2 Measured from Douglas bag at rest
or
Mixing chamber for exercise testing

F_ECO_2 Measured from Douglas bag at rest
or
Mixing chamber for exercise testing

Calculation of carbon dioxide output O_{CO_2}

$$O_{CO_2} = O_E \times F_ECO_2$$

Because there is little CO_2 in inspired air this, calculation becomes much simpler. Again by convention, O_{CO_2} is also expressed under STPD.

Respiratory Exchange Ratio (R).

$$R = \frac{O_{CO_2}}{O_{O_2}}$$

This value is typically 0.8 during the steady state of respiration and represents the ratio of CO_2 produced to oxygen consumed by the metabolic pathways of the cell. The value of R is fixed depending on the primary source of fuel being metabolized. Pure carbohydrate gives a ratio of 0.7 and fat burns at a ratio of 1.0. A typical ratio is 0.8 and represents a mixture of the two food groups being metabolically consumed.

In the nonsteady state, the amount of CO_2 exhaled rapidly changes based on the level of hyper or hypoventilation, so R may vary from 0.6 to 1.4. In addition, CO_2 produced by bicarbonate buffering of lactic acid adds to the O_{CO_2} produced by metabolism during peak exercise. This will be discussed further in the section on cardiopulmonary exercise testing.

The measurement of O_{O_2}, O_{CO_2}, and the ratio O_{CO_2}/O_{O_2} provide important information on assessing overall lung function, at rest and especially during exercise testing.

INSTRUMENTATION

Volume Measuring Devises

In order to calculate minute ventilation (O_E), and other derived variables such as O_{O2} and O_{CO_2} the expired volume over time is collected in a Douglas bag or meteorological (Mylar) balloon. So collected, the expired gas is then connected and emitted into a large spirometer, such as the 120 L Tissot spirometer, and the volume is measured by use of a calibration factor. The Tissot spirometer is a typical water-filled spirometer, but due to its size and the considerable inertia of the bell, it is not used for measuring tidal breathing. Smaller water-filled spirometers (9–13.5) liters have a lower airway resistance and an appropriate response time (up to 20 Hz) needed to measure forced exhalation. All water-sealed spirometers, regardless of size, are configured similarly and operate on the same principles (Fig. 4). A bell is sleeved between the inner

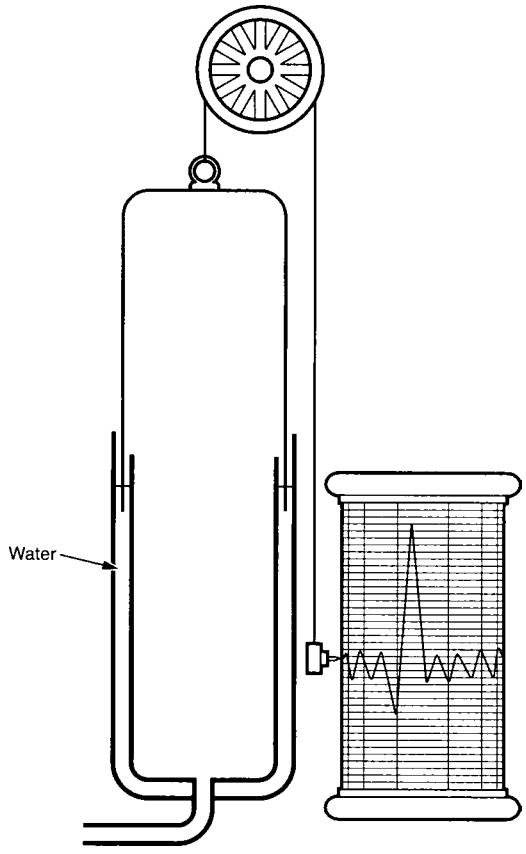


Figure 4. Water filled spirometer connected to a rotating Kymograph.

and outer housing. Water fills the space between the inner cylinder and outer housing, providing an airtight seal for air entering the bell. Rigid tubing connects the inner cylinder to exhaled air from the patient or collection bag. A CO_2 absorbant is placed in circuit when rebreathing maneuvers are carried out, such as resting metabolic rate or FRC determinations. When forced maneuvers, such as MVV, FVC, FEV_1 or PEFR are accomplished, the absorbant is removed, thereby reducing resistance in the expiratory system.

As air enters or leaves the spirometer the chain-suspended bell rises and falls. These movements are recorded by means of pens moving in parallel. A kymograph drum turns at a preselected pace, adding a time dimension to the volume changes. This allows measurement of the based variables such as MVV, FEV_1, FEF_{25-75}, PEF.

Another type of spirometer is the so-called dry rolling seal, also called the Ohio spirometer. A horizontal cylinder is attached to a flexible rolling seal. As air enters, the rolling seal allows the cylinder to move horizontally. Linear transducers attached to the cylinder are interfaced with a computer, allowing measurement of volume over time and flow.

Dry gasmeters typically measure inspired air to avoid accumulation of moisture on a bellows mechanism. The movement of the bellows is transmitted to a circular dial that is labeled with appropriate volumes.

A spirometer that uses a wedge-shaped bellows is called a wedge spirometer (5). The bellows expands and collapses as gas moves in and out. One side is stationary, while the other side moves a pen that records the changes. Pressure activation moves the chart horizontally, giving a time domain to the recording.

The peak flow meter (6) is a spirometer that works on a completely different principle from other spirometers. It is known as a variable orifice meter (Fig. 5), popularized as rotameter gas flow meter on anesthesia machines (7). As air enters the flow meter, a bobbin or light-weight marker is entrained in the vertical column of air. The flow meter has a variable inner orifice dimension that increases with height. The bobbin records the peak flow M, which corresponds to a particular inner orifice (r). The original peak flow meter was developed by F. M. Wright of England in 1959 and is often referred to as the Wright peak flow meter (6). It is based on Poiseuille's equation.

$$M = \frac{\pi P r^4}{8n\ell} \qquad \begin{array}{l} M = \text{flow} \\ r = \text{radius} \end{array}$$

FLOW AND VOLUME TRANSDUCERS

Instantaneous flow signals generated from flow transducers discussed below can be integrated with respect to time, thereby obtaining volume measurements. Harmonic analysis of respiratory flow phenomena has shown significant signals out to 20 C.P.S., requiring all devices to respond with fidelity at this frequency (8).

The Fleish Pneumotach (9) quantifies airflow by measuring the pressure drop across an in-line obstruction, such as mesh or porous membrane (Fig. 6). The pressure drop follows Poiseuille's law and is for laminar or nonturbulent flow. To prevent nonlaminar flow, various size pneumotachographs are used for different settings, such as studying children or exercising adults.

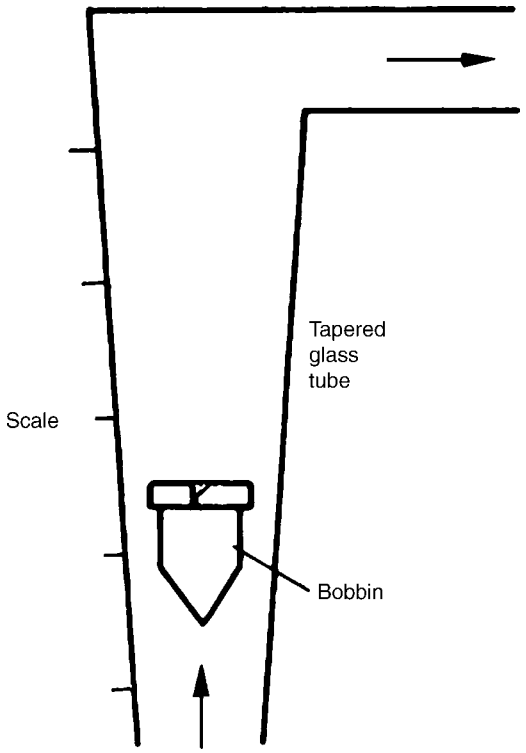


Figure 5. A variable orifice flow meter.

A Pitot tube that utilizes the Venturi effect is another type of flow meter (Fig. 7). The pressures of two tubes, one facing and one perpendicular to the air stream, is measured with a differential pressure transducer. Air flow velocity is proportional to the density of the gas and to the square of the pressure difference. They do not depend on laminar flow, typically are low weight and as opposed to pneumotachograms are low resistance breathing circuits.

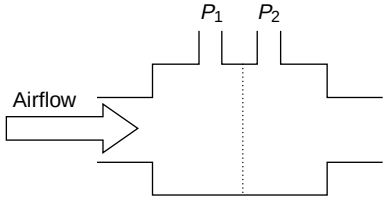


Figure 6. Fleisch pneumotachograph.

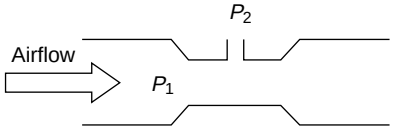


Figure 7. Pitot tube.

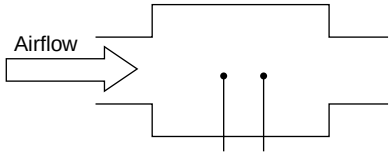


Figure 8. Hot wire anemometer.

Hot-wire anemometers measure mass flow by detecting the increase in current needed to heat a hot wire placed in the air stream as air flows over and cools the wire (Fig. 8).

A turbine transducer uses a low mass helical impeller mounted on bearings. As the impeller blade spins with airflow, an interposed light beam is broken and digital signals proportional to the flow are sent to the pressure sensor (Fig. 9).

The accuracy of each of these flow meters is potentially affected by the temperature, viscosity and density of the gas measured as well as the flow character (laminar or turbulent). When proper calibration is maintained, these devices produce a $\pm$ 3% accuracy, as recommended by the American Thoracic Society Guidelines (10).

GAS ANALYSIS

Rapidly responding (<100 ms) gas analyzers have made breath-by-breath analysis possible. Such measurements of expired oxygen and carbon dioxide give dense data useful in interpretation of cardiopulmonary exercise stress tests. When speed of analysis is not essential, chemical analysis by the Scholander or Haldane methods provide accurate results and are considered the gold standard. Gases measured by this method are used to validate other calibrating gases.

OXYGEN ANALYZERS

Discrete oxygen analyzers commonly used are paramagnetic, fuel cell or zirconium oxide. Each are calibrated over the expected range of measurement (e.g., 12–21%) by validated control gases. Of the three mentioned, the later two respond very quickly and so are used in breath-by-breath analysis. Paramagnetism is a distinctive property of oxygen: The molecules aligning in a magnetic field and thus enhancing it. A typical use of this slower responding analyzer is measuring oxygen concentration in large collecting bags or mixing chambers. Of note, oxygen does not have suitable absorption bands in the ultraviolet (UV), infrared (IR), or visible wavelengths. The following sections discuss gases that do have these properties.

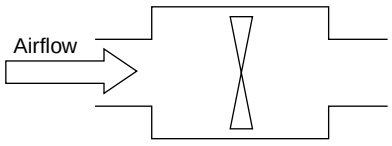


Figure 9. Turbine.

NONDISPERSIBLE INFRARED GAS ANALYZERS

This instrumentation (7,11) is used for multiple polyatomic gases including CO_2 and CO, commonly measured in pulmonary function testing. An IR beam is directed alternately through a reference and measurement cell. By means of a chopper wheel, a detector senses the alternating change in absorption of selected IR wavelengths. This signal is amplified with a high input impedance ac amplifier, rectified and displayed on a meter or digital recorder.

NITROGEN METER

The fact that nitrogen molecules can be excited in a low pressure electric discharge to emit visible light in the purple region forms the basis of the nitrogen meter (7,11). A 1500 V electric potential difference is maintained and optical filters select appropriate wavelengths in the violet range. The resulting light intensity is measured by a photocell with an amplifier.

MASS SPECTROMETER

Used primarily in research labs, mass spectrometry is capable of analyzing any gas with speed, specificity, sensitivity, and accuracy unmatched by any other method. Molecules of the sample are ionized at low pressure by a beam of electrons, and the ions are deflected in a circular path by a magnetic field. The stream of particles splits into beams of different molecular weight, any one of which can be detected by a suitably placed collector. Due to expense, mass spectometry is not a typical part of clinical (hospital or office-based) pulmonary function testing.

PULMONARY FUNCTION TESTS

Pulmonary function tests do not provide a complete diagnostic picture. At best, they support a clinical impression that is formed by a thorough history, physical exam, and X-ray studies. Given the myriad of lung function tests available, an informed decision on the most important ones to order maximized their usefulness.

SPIROMETRY

A forced exhalation maneuver after a deep inhalation is recorded by a moving kymograph on a small water-filled spirometer. A pneumotachograph with computer interfacing could be used with equally acceptable results. A tracing of volume over time is obtained and the following measurements are derived (Fig. 10). This is called a forced vital capacity maneuver.

FVC	Forced vital capacity. This value is effort dependent and depends on the full cooperation of the patient.
FEV_1	From the tracing, the amount of air exhaled in the first second is measured. Patient effort is required to obtain a reliable value.

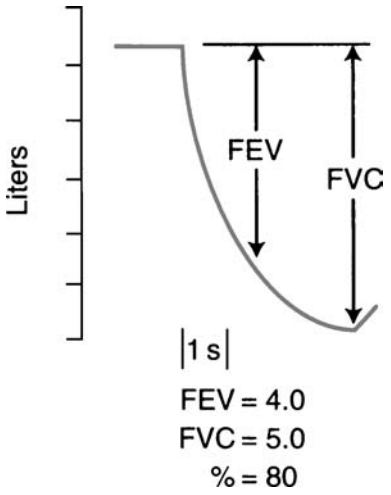


Figure 10. A normal spirographic tracing volume versus time.

FEV_1/FVC This ratio is helpful in ascertaining airflow obstruction and is typically 70% in normal people.

$FEF_{25–75}$ This is a flow rate and represents the slope $\Delta V/\Delta T$ during the mid-section of the spirometer tracing. It is where dynamic compression of airways occurs (Fig. 11).

PEFR Peak expiratory flow rate. The steepest slope on the curve, typically at 80% of the vital capacity maneuver.

MMEF Mid-maximal expiratory flow rate is a slope taken at 50% of the vital capacity maneuver. It reflects smaller airways airflow.

All along the spirometric curve, an infinite number of slopes can be determined, from which a flow–volume curve could be constructed. However, a flow signal from a pneumotachograph plotted against time is the preferred method of generating this data.

SVC Slow Vital Capacity. Instead of a forced maneuver, the vital capacity may be performed with less effort. In normals, the FVC and SVC are equal. Because of air trapping, patients with emphysema more fully exhale with the SVC.

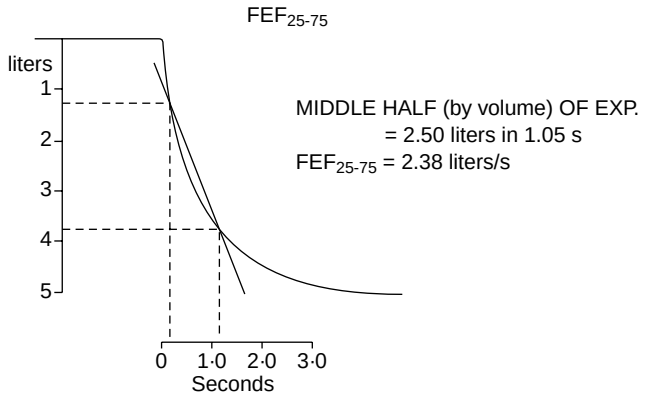


Figure 11. The $FEF_{25–75}$ Slope $\Delta V/\Delta T$ derived from spirogram and expressed in liters per second.

FLOW–VOLUME CURVES

Spirometers and flow-volume curves contain the same information, displayed differently. Below is a typical flow–volume curve generated by a pneumotachograph (Fig. 12). A forced expiratory maneuver is performed. Pneumotachographs can measure peak inspiratory flow volume curves, which are important in diagnosing obstructive supraglottic lesions. Expiratory flow-volume curves can unmask central airway tumors not appreciated on simple spirometry at 50% FVC nor seen on standard chest X ray (12).

This curve reflects phenomena in small airways where dynamic compression occurs. Normally, flow here is dependent on gas density. This fact forms the basis for flow–volume measurements after inhalation of 80% helium–20% oxygen mixtures. When overlayed, the flow–volume curves of smokers show little difference after breathing low density helium–oxygen compared to room air (79% nitrogen–21% oxygen). The density independence reflects increased resistance in diseased small airways and is a very sensitive early indication of smoking-induced lung disease.

LUNG VOLUME

There are three methods for measuring the lung volume FRC or functional residual capacity: body plethysmography, nitrogen washout(13), or helium dilution. Once the FRC is measured, residual volume can be determined by asking a patient to exhale completely from FRC to residual volume (RV). The air left in the lung is the residual volume. The total lung capacity is then determined by measuring a deep inhalation from RV (inspiratory capacity) and adding

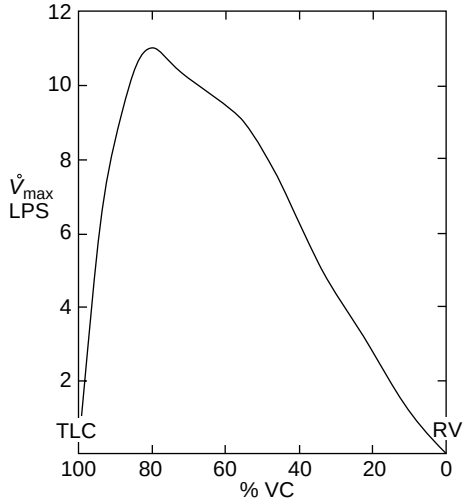


Figure 12. Flow volume curve. Forced exhalation from TLC to RV = FVC recorded by a pneumotachograph.

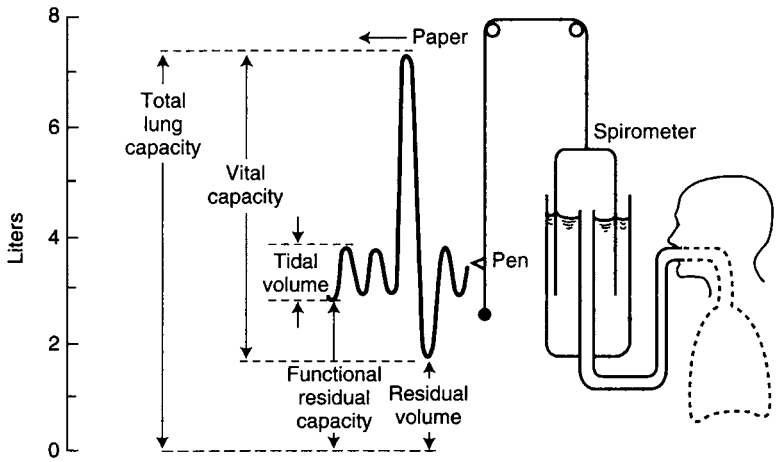


Figure 13. Lung volume. Measurements by a water filled spirometer.

that to the residual volume already measured (Fig. 13).

$$\text{RV} = \text{FRC} - \text{ERC}$$
$$(\text{ERC} = \text{expiratory reserve capacity})$$
$$\text{TLC} = \text{RV} + \text{IC}$$
$$(\text{IC} = \text{inspiratory capacity})$$

FRC BY PLETHYSMOGRAPHY

A plethysmograph is an airtight box of known volume, similar to a telephone booth, in which a patient sits. A mouthpiece connects the patient to air outside the apparatus and pressure sensors are located within the box and within the breathing capacity. At the end of a normal tidal breath, a shutter on the mouthpiece closes and the subject is asked to make respiratory efforts. As the subject tries to inhale, the volume of the lung expands slightly while the pressure drops due to the chest (lung) expansion (Fig. 14). Applying Boyle's law, if the pressures in the box before and after the inspiratory effort are P_1 and P_2, respectively, V_1 is the preinspiratory box volume and ΔV is the change in volume of the box (or lung) ΔV can be obtained from the equation $P_1V_1 = P_2(V_1 + \Delta V)$.

Applying Boyle's law to the gas in the lung, $P_3(V_2) = P_4(V_2 + \Delta V)$, where P_3 and P_4 are the mouth pressures before and after the inspiratory effort and V_2 is the FRC. Thus FRC can be obtained.

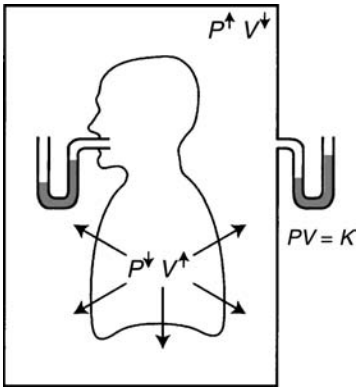


Figure 14. The FRC by whole body plethysmography utilizing Boyle's law.

The body plethysmograph measures the total volume of compressible gas, including any that is trapped behind closed or poorly communicating airways.

The other two methods measure only volumes based on gas communicating with and open to airways. This is not an issue in normal subjects, but in diseased lungs considerable amounts of gas are trapped and do not communicate freely. Therefore the FRC values differ depending on methodology.

NITROGEN WASHOUT

If a subject quietly breathes 100% oxygen for several minutes, all the nitrogen emptied from the lung can be determined by multiplying the exhaled volume by the exhaled nitrogen concentration. Since the initial lung concentration of N_2 is 80%, the measured volume of nitrogen exhaled multiplied by 1/0.8 equals the volume of the lung prior to 100% oxygen breathing. The value of FRC can be underestimated if significant parts of the lung communicate poorly or not at all with the inspired oxygen (13).

HELIUM DILUTION

As a subject breathes from a spirometer with a known concentration of helium, after several normal breaths the helium concentration in the lung and spirometer equilibrate (Fig. 15), Since helium is insoluble in blood, none of it is absorbed, so the final equilibrium concentration is a reflection of dilution only. The amount of helium before equilibration is $(C_1 \times V_1)$ and equals that after equilibration C_2 $(V_1 + V_2)$ solving for V_2, $V_2 = (C_1V_1/C_2) - (V_1)$. During the equilibration period oxygen is added to the spirometer and carbon dioxide is absorbed.

Although many other measurements of lung function are perhaps more useful, knowledge of the lung volumes is essential in other complex measurements such as diffusing capacity.

DIFFUSING CAPACITY

Given the challenges with measuring the diffusing capacity of the lung for oxygen, carbon monoxide as originally

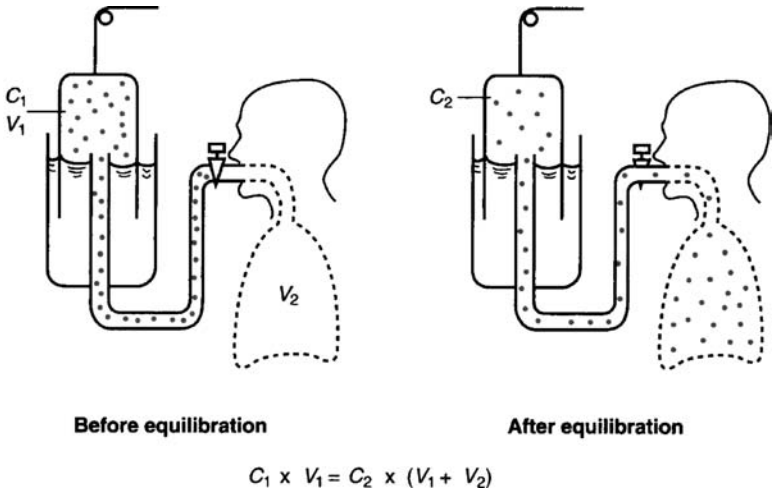


Figure 15. Helium equilibration. Technique for FRC determination.

used by Marie Krogh in 1914 is used for current day measurements. There are at least seven variations of this method that have since been developed.

The most common of these is the single-breath modified Krogh technique attributed to Kety and Fowler. Both helium and carbon monoxide are inhaled. After a period of breath hold (15 s), the alveolar portion of the exhaled gas is collected and the concentration of carbon monoxide and helium is measured.

The initial alveolar carbon monoxide concentration is calculated thus:

$$F_{\mathrm{I}}\mathrm{CO} \times \frac{\mathrm{He\%in\ expired\ alveolar\ sample}}{\mathrm{Inspired\ helium\ percentage}} = F_{\mathrm{A}}\mathrm{CO}$$

$$D_{\mathrm{L}}\mathrm{CO} = \frac{\mathrm{Alveolar\ volume\ STPD} \times 60}{\mathrm{Seconds\ of\ breath\ hold} \times \mathrm{PB} - 47} \times \ln\left(\frac{F_{\mathrm{I}}\mathrm{CO\ alv}}{F_{\mathrm{E}}\mathrm{CO\ alv}}\right)$$

In the above equation, the alveolar volume is measured by a helium dilution technique similar to that in lung volume determinations.

CLOSING VOLUME

This sensitive test detects early changes in lung function and reflects pathology in the small airways. Smokers have an abnormally high closing volume prior to any other pulmonary function test changes. In this test, the subject inhales a breath of 100% O_2 to TLC. During the subsequent exhalation, the nitrogen is measured through the alveolar plateau to an abrupt rise in exhaled nitrogen, so-called phase 4 (Fig. 16). This signals closure of airways in the base of the lungs and preferential emptying of upper airways. Less of the 100% oxygen inhalation is distributed to the upper lung, making it richer in nitrogen. It is this fact that creates phase 4. In some lung diseases, the closing volume is above the FRC. This means that airways close even during normal breathing and is an indication of advanced disease.

MAXIMAL VOLUNTARY VENTILATION

This test measures the volume of air moved during 15 s of repetitive forced deep maximal exhalations. A water filled spirometer is used for measurement with a time kymographic tracing. Pneumotachographs with real time computer graphics may be used. The main requirement for accurate test results is a low resistance breathing circuit and avoidance of resonance in the system. Both problems have been overcome by modern spirometers, valves and tubing. Although this is formally a lung test, nonpulmonary factors such as motivation, muscular strength and endurance are very important and must be taken into consideration when interpreting the test. The results are expressed in liters per minute BTPS.

PEAK FLOW

Peak flow meters have the distinct advantage of being handheld, self-contained and thus very portable. This is an effort-dependent test, yet it is an excellent reflection of airways function. Its main utility is quickness and simplicity, and it is often used for the management of asthma. Much like a home glucose meter in a diabetic, the peak flow meter can give objective assessment of airways function throughout the day to help guide treatment and presage severe attacks.

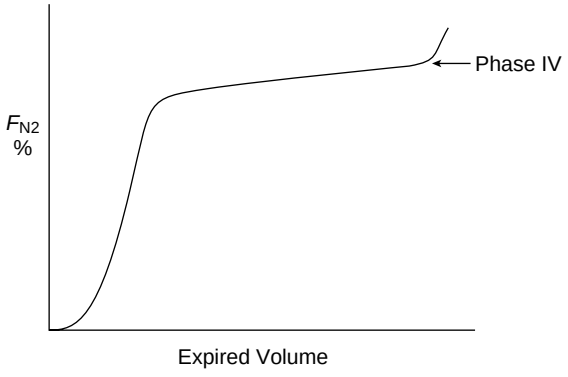


Figure 16. Abrupt rise at the end of exhalation called phase (IV) correlates with the closing volume.

STANDARDIZATION OF PULMONARY FUNCTION TESTS

The American Thoracic Society published standards on spirometry in 1979 at the Snowbird Workshop. An official update was issued in 1994, which details equipment and procedural guidelines to enhance accuracy and reproducibility.

Without meticulous attention to these recommendations, the utility of all pulmonary function tests is compromised.

CARDIOPULMONARY STRESS TESTING

Cardiopulmonary stress testing (14) uncovers disorders of the respiratory system as it functions in the integrated cardiopulmonary response to the metabolic demands of increasing incremental work loads. During the course of a cardiopulmonary stress test, the following variables are measured in real time:

W'	Work
f_c	Heart rate
O_{O_2}	Oxygen consumption
O_{CO_2}	Carbon dioxide production
O_E	Minute Ventilation
R	O_{CO_2}/O_{O_2}
O_E/O_{O_2}	Efficiency of ventilation
O_E/O_{CO_2}	Efficiency of ventilation
$P_{ET}O_2$	End tidal P_{O_2}
$P_{ET}CO_2$	End tidal P_{CO_2}
P_ECO_2	Mixed expired CO_2

The instrumentation required to make these breath-by-breath and time-averaged measurements are rapid responding CO_2 and O_2 meters; pneumotachograph for flows, ventilation; mixing chamber for collection of exhaled gas; ECG; ergometer cycle, arm, treadmill; computer interface with full graphics in real time.

Limits to exercise appear often in clinical medicine with the presenting problem of dyspnea or shortness of breath on exertion. The genesis of this symptom may reside within either the respiratory or cardiac systems or both simultaneously. Analysis of data obtained from cardiopulmonary stress testing can aid in clinical diagnosis.

Testing begins with an incremental ergometric work load that creates a systemic metabolic response measured as oxygen uptake or consumption. There is a strong reproducible linear relationship between the work load and oxygen consumption. It is derived from the coupling of work and mitochondrial metabolic pathways for cellular energy generation.

With increasing work loads and oxygen consumption, the total ventilation of the lung increases. This requires higher airflow which in normals, even at peak work loads, never reaches flows measured in maximal flow–volume curves. Such is not the case in obstructive lung disorders, where reduced flow is a hallmark of the disease. Exercise is limited due to the inability to generate flows capable of sustaining the metabolic demands.

As the exercise test progresses real-time analysis of dead space can be measured. Typically the $P_ECO_2 = P_ACO_2 = P_aCO_2$ remains constant with the mixed expired CO_2 P_E CO_2 increasing. This is reflected in a decrease in the O_E/OCO_2 marking increased efficiency as more CO_2 is exhaled per breath. By making use of the Bohr equation:

$$V_D/V_T = \frac{F_ACO_2 - F_ECO_2}{F_ACO_2}$$

it appears that V_D/V_T or wasted ventilation decreases with increased exercise. This occurs because lung apical units not perfused but ventilated at rest now are fully perfused and participate in gas exchange. If V_D/V_T does not decrease with exercise, then it is likely based on a structural disease such as emphysema. The work of breathing at any given work load is higher in these patients in part because V_D/V_T (wasted ventilation) remains abnormally high.

As the level of incremental work load increases, the delivery of oxygen fails to meet the metabolic demands of tissues and anaerobic metabolism becomes prominent. Lactic acid is dumped into the blood stream and is quickly buffered by bicarbonate, which generates more carbon dioxide. A dramatic upsurge in OCO_2 marks this point and is called the anaerobic or metabolic threshold. The parameter $R(OCO_2/OO_2)$ values that previously were 0.8 are now 1.2–1.4, indicating a combined metabolic and buffer source of carbon dioxide. Ventilation (O_E) is driven by the chemical stimulation of lactic acidosis in addition to the demands of oxygen delivery.

Data from such a stress test yields much clinically useful information and allows one to differentiate a pulmonary from a cardiac cause for exercise limitations. Cardiopulmonary deconditioning has a distinct pattern as does obesity. Low peak OO_2 and low A.T. as a percent of OO_2 max are good indications of these two conditions.

Future of Pulmonary Function Testing

No doubt, advances in instrumentation and computers will continue to refine pulmonary function testing. Miniaturization of testing equipment allows complex measurements not only in the laboratory but also in the wild. The burgeoning science of sleep medicine is an example of this.

Epidemiological studies will explore the relationship of pulmonary function to health and uncover what makes the vital capacity so vital to life. A fundamental role for FEV_1 in total mortality independent of cigarette smoking has been proposed. Whether reduced lung function leaves an individual open to oxidative stress is unknown.

If pulmonary function proves to be a long-term predictor for overall survival rates in both genders, it could be used as a tool in general health assessment.

The search for tests that implicate early potentially reversible lung disease will continue. The benefit to asymptomatic patients and society as a whole is obvious.

The long-term effects of air pollution and impact of air quality on lung health will always be of prime concern, not only to the general public, but to government officials who set air quality standards.

Perhaps pulmonary function testing will ultimately guide and protect us all.

TERMINOLOGY—DEFINITIONS—EQUATIONS

Spirometer	A measuring device for determining lung volume, its subcompartments, and expiratory flow rates.
FRC	Functional residual capacity: The volume in the lung after a normal exhalation. At this volume the recoil pressure of the lungs inward is exactly balanced with the outward recoil pressure of the chest wall.
FEV_1	Forced expiratory volume in the first second: The amount of air expired in the first second of a forced expiratory maneuver.
TLC	Total lung capacity.
RV	Residual volume: The volume of air left in the lungs after a full exhalation.
FVC	Forced vital capacity: The amount of air exhaled during a complete exhalation.
$FEF_{25–75}$	Forced expiratory flow: The mean expiratory flow measured between 75 and 25% of the vital capacity during forced exhalation.
PEFR	Peak expiratory flow rate during forced exhalation.
MVV	Maximal voluntary ventilation expressed in liters per min.
DL_{CO}	Diffusing capacity for carbon monoxide.
Flow–volume curve	A maximal exhalation measuring flow versus volume.
P_AO_2	Alveolar oxygen partial pressure.
P_ACO_2	Alveolar carbon dioxide partial pressure.
$P_{ET}O_2$	End tidal oxygen partial pressure.
$P_{ET}CO_2$	End tidal carbon dioxide partial pressure.
P_EO_2	Mixed expired oxygen partial pressure.
P_ECO_2	Mixed expired carbon dioxide partial pressure.
Oo_2	Volume of oxygen take up per minute.
Oco	Volume of carbon dioxide output per minute.
O_E	Minute ventilation: Total volume of air expressed per minute from the lungs.
V_T	Tidal volume: The volume of a single breath.
V_D	The volume of physiological dead space.
V_D/V_T	The ratio between dead-space volume and tidal volume. This ration indicates the efficiency of ventilation.
V_A	Volume of alveolar gas in the tidal volume.
General gas law	$PV = RT$
Boyle's law	$P_1V_1 = P_2V_2$ (temperature constant)
Charles's law	$\frac{V_1}{V_2} = \frac{T_1}{T_2}$ (pressure constant)
Poiseuille's law	$V' = \frac{P\pi r^4}{8n}$ P = Pressure difference across length ℓ and radius r n = Coefficient of viscosity
Bohr equation	$V_D/V_T = \frac{P_ACO_2 - P_ECO_2}{P_ACO_2}$

BIBLIOGRAPHY

1. Hutchinson J. On the capacity of the lungs, and on the respiratory functions, with a view of establishing a precise and easy method of detecting disease by the spirometer. Med Chir Trans (London) 1846;29:137.
2. Comroe JH Jr. Retrospectoscope. Insights into medical discovery. Menlo Park, (CA): Von Gehr Press; 1977.
3. Otis AB, Rahn H. Developments of Concepts in Rochester, New York, in the 1940's. In: West JB, editor. Pulmonary Gas Exchange Volume 1. New York: Academic; 1980. pp. 33– 65.
4. Bates DV, Macklem DT, Christie RV. Respiratory Function in Disease. 2nd ed. Philadelphia: Saunders; 1971.
5. Horton GE, Phillips S. The expiratory ventilagram: application of total and time vital capacities and maximal expiratory flow rate, as obtained by a bellows apparatus, for bedside and office use. Am Rev Respir Dis 1959 Nov; 80:724–731.
6. Wright BM, McKerrow CB. Maximum forced expiratory flow rate as a measure of ventilatory capacity: with a description of a new portable instrument for measuring it. Br Med J 1951 Nov. 21; 5159:1041–1046.
7. Hill DW. Physics Applied to Anesthesia. New York: Appleton-Century-Crofts; 1972.
8. McCall CB, Hyatt RE, Noble FW, Fry DL. Harmonic content of certain respiratory flow phenomena of normal individuals. J App Physiol 1957 Mar; 10(2):215–218.
9. Bouhuys A. The clinical use of pneumotachography. Acta Med Scand 1957 Nov. 15; 159(2):91–103.
10. Standardization of Spirometry. Official statement of the American Thoracic Society. Am J Respir Crit Care Med. 1995;152:1107–1136.
11. Gaensler EA. Evaluation of pulmonary function: methods. Annu Rev Med 1961;12:385–408.

12. Miller DR, Hyatt RE. Obstructing lesions of the larynx and trachea: clinical and physiologic characteristics. Mayo Clinic Proc. 1966 Mar; 44:145–161.
13. Emmanuel G, Briscoe WA, Cournand A. A method for the determination of the volume of air in the lungs: measurements in chronic pulmonary emphysema. J Clin Invest Feb 40:329–337.
14. Cooper CB, Storer TB. Exercise testing and interpretation. Cambridge (MA): Cambridge University Press; 2001.

Reading List

West JB. Respiratory Physiology the Essentials. 7th ed. Philadelphia: Lippincott Williams & Wilkins; 2004.

See also HEART-LUNG MACHINES; LUNG SOUNDS; RESPIRATORY MECHANICS AND GAS EXCHANGE; VENTILATORY MONITORING.

PUMPS, INFUSION.

See DRUG INFUSION SYSTEMS.

Q

QUALITY CONTROL, X-RAY. See X-RAY QUALITY CONTROL PROGRAM.

QUALITY-OF-LIFE MEASURES, CLINICAL SIGNIFICANCE OF

CELIA C. KAMATH
JEFFREY A. SLOAN
Mayo Clinic
Rochester, Minnesota

JOSEPH C. CAPPELLERI
Pfizer Inc
Groton, Connecticut

INTRODUCTION

The field of patient-reported outcomes, particularly health-related quality of life (QOL), has burgeoned in the last few years (1,2). The importance assigned to the study of these outcomes has been attributed to the aging of the population and consequently higher prevalence of chronic diseases, along with the reality that medical treatment often fails to cure the disease, but may affect QOL (3). Health-related quality of life has gained attention in research and clinical trial settings (3,4).

The increasingly important role assigned by patients and clinicians to QOLs role in medical decision making has resulted in greater attention paid to the interpretation of QOL scores, particularly as it relates to clinical significance (5–7). Clinical significance relates to the clinical meaningfulness of intersubject or intrasubject changes in QOL scores. Clinical significance has been difficult to determine, in part due to the development of a myriad of QOL instruments over the past decade (8,9). Some of these have had little or no psychometric validation (1,2,6,10,11) or clinical validation (9,12). Moreover, relative to traditional clinical endpoints, like survival and systolic blood pressure, QOL as a clinical endpoint is relatively unfamiliar especially in regard to interpretation and relevance of changes in QOL scores (13).

Why is clinical significance of QOL scores important? It aids in the design of studies by helping to determine sample size calculations. Evidence of clinical significance may be used by regulatory agencies for drug approval, by clinicians to decide between treatment alternatives, by patients to make informed decisions about treatment, by the healthcare industry for formulary and reimbursement decisions, and by healthcare policy makers to make policy decisions regarding resource allotment. Early evidence of the clinical implications of QOL is evident in the links between survival and QOL components (e.g., patients' fatigue levels, social support, and group counseling) (14–17). Even a simple, single-item measure of patient global QOL can be related to patient survival (18). Changes in QOL scores can also be linked to positive economic (19,20) and social (21) outcomes.

HISTORICAL BACKGROUND

Statistical significance as measured by a *P*-value is influenced by sample size and data variability. While statistical significance can be considered a prerequisite for clinical significance, only clinical significance assigns meaning to the magnitude of effect observed in any study. Historically, Cohen (22) was responsible for proposing one of the earliest criteria for identifying important change, which can be construed as clinically significant. He suggested that a small effect size (defined later in this article) was 0.2 standard deviation units, a medium effect size was 0.5, and a large effect size was 0.8. Although his intention was to provide guidance for sample size calculations in the social and behavioral science, Cohen's benchmarks have extended to healthcare research to decide whether or not a change in QOL scores is important. Current research suggests that a moderate effect size of one-half a standard deviation unit (effect size = 0.5) is typically important (23). A more recent and popular definition of clinical significance uses an anchor-based approach based on an external standard that is interpretable and appreciably correlated to the target QOL measure, in order to elucidate the meaning of change on the target QOL measure.

Embedded under the rubric of clinical significance is the minimum important difference, a lower bound on clinical significance. One definition of a minimum important difference (MID) is "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side-effects and excessive cost, a change in the patient's management"(24). Some researchers prefer to use the term "minimally detectable difference"(25,26); other approaches have sprouted (e.g., the empirical rule effect size, ERES, method) (27,28).

No single solution to the challenging topic of assessing clinical significance exists. Nevertheless, a series of proposals has engendered understanding and appreciation of the topic. Special issues in *Statistics in Medicine* (1999, Vol. 18) and the *Journal of Consulting and Clinical Psychology* (1999, Vol. 67) have been dedicated to the topic of clinical significance of QOL and other clinical measures. Proceedings from a meeting of an international group of ~30 QOL experts were published recently in a special issue of the *Mayo Clinic Proceedings* (2002, Vol. 77) (29–35), which provides practical guidance regarding the clinical significance of QOL measures.

CHAPTER OUTLINE

This article draws largely from recent scientific literature, including the *Mayo Clinic Proceedings* (29–35) and other

sources (4), to provide an overview on the clinical significance of QOL measures.

The following section on Design and Methodology covers the different perspectives and existing methods to determine clinical significance. The next section on Examples illustrates trials in which researchers attempted to define the concept on specific QOL measures. Then the section on Recent Developments highlights new methods to determine clinical significance. Finally, the section on Concluding Remarks discusses some future directions for research.

DESIGN AND METHODOLOGY

Perspectives for Determining and Interpreting Clinical Significance

Clinical significance involves assigning meaning to study results. The process of establishing such meaning can be conceptualized in two steps: (*1*) understanding what changes in score mean to the concerned stakeholder (e.g., patient, clinician, clinical researcher, policy maker) and (*2*) making results of clinical studies interpretable and comprehensible to such stakeholders or decision makers (30,36). The term clinical in relation to significance has different meaning and implications for different stakeholders such as patients, clinicians, and society.

From the patient's perspective, clinical significance can be defined as the change in QOL scores that patients perceive as beneficial (or detrimental) and important that prompts them to seek health care or request changes in their treatment (33), or that induces patients to determine that the intervention has been successful (24). From the clinician's perspective, it can be defined as the diagnosis of the clinician as to the amount of change in QOL scores that would mandate some form of clinical intervention (37). From the societal or population perspective, clinical significance is based on the values of the group surveyed, where importance is defined by the outcomes that are deemed worthy of society's resources. Any or all of these perspectives for defining clinical significance may be applicable, but they may not always be in agreement (4).

An equally important issue is the different perspectives for interpreting clinical meaningfulness of changes in reported QOL (35). For example, a clinician may use QOL data to explain the treatment alternatives to a patient, while a health policy maker may describe to elected officials the financial impact on a patient population whose QOL has changed. Similarly, a regulatory agency and pharmaceutical company may ascertain the appropriate level of evidence for a successful research study (35). Thus QOL results must be framed, analyzed, and presented in a way that is meaningful to the pertinent audience and its respective needs. Only then will the concept be meaningful and gain greater acceptance and use over time.

METHODS TO EXPLAIN THE CLINICAL SIGNIFICANCE OF HEALTH STATUS MEASURES

Two common approaches used to establish the interpretability of QOL measures are termed anchor and distribution based. The characteristics of each approach are described below. Several examples will be given later in the section on Examples. Interested readers are encouraged to read Lydick and Epstein (1993) [Lydick, 1993 #40] Crosby et al. (4), and Guyatt et al. (30) for an expanded discussion of the concepts presented here.

Anchor-based approaches are used to determine clinically meaningful change via cross-sectional or longitudinal methods involve comparing measures of QOL to measures with clinical relevance (4). Cross-sectional methods include several forms: (*1*) comparing groups that are different in terms of some disease-related criterion (38,39); (*2*) linking QOL to some external benchmarking criteria (40–42); (*3*) eliciting preference-based ratings on a pairwise basis, where one person's ratings state serves as an anchor to evaluate the other person's ratings (43); and (4) using normative information from dysfunctional and functional populations (6). Longitudinal methods involve the comparison of changes in QOL scores across time with the use of (*1*) global ratings of change as "heuristics" to interpret changes in QOL scores (5,24,38,44); (*2*) significant future medical events for establishing difference thresholds (45); and (*3*) comparisons of changes in HRQOL to other disease-related measures of outcome across time (46). Anchor-based methods are cataloged in Table 1.

Anchor-based methods require two properties (30): (*1*) anchors must be interpretable, else they will hold no meaning to clinicians or patients; and (*2*) anchors must share appreciable correlation with the targeted QOL measure. The biggest advantage of anchor-based approaches is the link with a meaningful external anchor (4), akin to establishing the construct validity of the measure (49). Potential problems, however, exist with this approach. These include recall biases (50), low or unknown reliability and validity of the anchor measure (51), low correlation between anchor and actual QOL change score (52–55), and complex relationships between anchors and QOL scores (56) and the challenge of defining a meaningful change in the anchor itself.

Hays and Wooley (57) recommend caution in the indiscriminate dependence and use of a single minimum important difference (MID) measure. They also list several problems in estimating MIDs: the estimated magnitude could vary depending on the distributional index (57,58), the external anchor (59), the direction of change (improvement vs. decline) (60), and the baseline value (61). In general, longitudinal methods are preferable because of their direct link with change (4).

Distribution-based approaches for determining the importance of change are based on the statistical characteristics of the obtained sample, namely, average scores and some measure variability in results. They are categorized as (*1*) those that are based on statistical significance using p-values (i.e., given no real change, the probability of observing this change or a more extreme change), which include the paired t-statistic (62) and growth curve analysis (63); (*2*) those that are based on sample variation (i.e., those that evaluate mean change in relation to average variation around a mean value), which include effect size (22,64), standardized response mean (SRM) (44), and responsiveness statistic (65); and (*3*) those that are based on the measurement precision of the instrument

Table 1. Anchor-Based Methods of Determining Change[a]

Type	Method	Examples	HRQOL evaluated in relation to:	Advantages	Disadvantages
Cross-sectional	Comparison to disease-related criteria	39, 47	Disease severity or diagnosis	Can be standardized Easy to obtain	May not reflect change Groups may differ in other key variables
	Comparison to nondisease-related criteria	40, 41	Impact of life events	Easy to obtain Provides external basis for interpretation	May no reflect change Groups may differ on other key variables Relationship to HRQOL not clear
	Preference rating	43	Pairwise comparisons of health status	All health states are compared	May not reflect change Hypothetical, artificial Time Consuming
	Comparison to known populations	6	Functional or dysfunctional populations	Uses normative information	Normative information not always available Amount of change needed not specified
Longitudinal	Global ratings of change	5, 24, 38, 44	Patients' or clinicians' ratings of improvement	Easy to obtain Best measure from individual perspective Can take into account a variety of information	Does not consider measurement precision Unknown reliability Influenced by specific rating scale and anchors
	Prognosis of future events	45	Those experiencing and not experiencing some future event	Prospective Provides evidence of predictive validity	Does not consider measurement precision Difficult to obtain
	Changes in disease related outcome	48	Changes in clinical outcome	Tied to objective outcome measure Known psychometric properties	Does not consider measurement precision Assumes strong HRQOL-outcome correlation

[a]Reprinted with permission from Ref. 4.

(i.e., evaluate change in relation to variation in the instrument as opposed to variation of the sample), which includes the standard error of the mean (SEM) (7) and the reliable change index (RC) (6). Distributed-based methods are catalogued in Table 2.

An advantage of the distribution-based methods is that they provide a way of establishing change beyond random variation and statistical significance. The effect size version of the distribution-based methods is useful to interpret differences at the group level and has benchmarks of 0.20 standard deviations units as a small effect, 0.50 as a moderate effect, and 0.80 as a large effect (22,64,66). The measure that seems most promising for the purpose of establishing clinical significance at the individual patient level are the SEM and the RC. These measures are based on the measurement precision of the instrument and incorporate the reliability of the instrument (e.g., Cronbach's alpha or test–retest reliability), and the standard deviation of scores. In principle, SEM and RC are sample invariant. Researchers have favored Cronbach's alpha over test–retest reliability to calculate reliability for the SEM (7,30,67), because this is more conveniently available over test–retest data.

Distribution methods are particularly helpful when used together with meaningful anchors, which enhances validity, and hence meaning to the QOL measure. There is some encouragement to know that anchor-based measures appear to coincide with distribution-based methods. Researchers have found a correspondence between SEM and anchor-based determinant of a minimum important difference across difference diseases (7,23,67,68). The 1 SEM benchmark corresponds with an effect size (ES) of ~0.5. Nonetheless, note that the SEM is moderated by the reliability of the measure, where measures with higher reliability are "rewarded" by lowering the effect size (ES) needed to achieve a minimally important difference. Thus the 1 SEM benchmark corresponds with an ES = 0.5, when reliability of the scale = 0.75; the correspondence shifts to 1 SEM is equivalent to an ES = 0.33, when reliability increases to 0.9, which is frequently attainable in focused assessments. A rationale for a SEM as a measure of MID is provided by Norman et al. (23) who assert that Miller's theory (69) of the limits of human discernment is linked to the threshold of 0.5 standard deviation units.

EXAMPLES

This section provides examples of studies used to determine clinical significance and presents general advice for defining and interpreting clinical significance in clinical studies. Table 3 includes several examples on the use of both anchor-based methods and distribution-based methods to establish clinical significance across a wide range of QOL measures. These examples span several disease groups, instruments, and methods for determining clinical significance. Readers are encouraged to review the cited papers for further details on these studies.

Table 2. Distribution-Based Methods of Determining Change [a]

Method	Reference	HRQOL evaluated in relation to:	Calculation	Advantages	Disadvantages
Paired t-statistic	62	Standard error of the mean change	$\frac{x_1 - x_0}{\sqrt{\frac{\sum (d_i - \overline{d})^2}{n(n-1)}}}$	None	Increases with sample size
Growth curve analysis	63	Standard error of the slope	$\frac{B}{\sqrt{V}}$	Not limited to pre-test and post-test scores Uses all of the available data	Increases with sample size Requires large sample sizes Assumes data missing at random
Effect size	22, 64	Pre-test standard deviation	$\frac{x_1 - x_0}{\frac{\sqrt{\sum (x_0 - \overline{x}_0)^2}}{n-1}}$	Standardized units Benchmarks for interpretation Independent of sample size	Decreases with increased baseline variability of sample Does not consider variability of change May vary widely among samples
Standardized response mean	44	Standard deviation of change	$\frac{x_1 - x_0}{\sqrt{\frac{\sum (d_i - \overline{d})^2}{n-1}}}$	Standardized units Independent of sample size Based on variability of change	Varies as a function of effectiveness of treatment
Responsiveness statistic	65	Standard deviation of change in a stable group	$\frac{x_1 - x_0}{\sqrt{\frac{\sum (d_{i\ stable} - \overline{d}_{stable})^2}{n-1}}}$	Standardized units More conservative than effect size Independent of sample size Takes into account spurious change due to measurement error	Data on stable subjects frequently not available
Standard error of measurement	7	Standard error measurement	$\frac{x_1 - x_0}{\sqrt{\frac{\sum (x_0 - \overline{x}_0)^2 (\sqrt{1-r})}{(n-1)}}}$	Relatively stable across populations Takes into account the precision of the measure Cutoffs based on confidence intervals	Assumes measurement error to be constant across the range of possible scores
Reliable change index	6	Standard error of the measurement difference	$\frac{x_1 - x_0}{\sqrt{2(SEM)^2}}$	Relatively stable across populations Takes into account precision of measure Cutoffs based on confidence intervals	Assumes measurement error to be constant across the range of possible scores

[a]Reprinted with permission from Ref. 4.

We begin with a classic paper by Jaeschke et al. (24), one of the first papers on clinically meaningful differences determined through the anchor-based approach. The magnitude of difference considered minimally significant was an average of 0.5 per item on a 7-point scale, which was confirmed by Juniper et al. (5) on the asthma quality of life questionnaire (AQLQ). A third study, by Kazis et al. (64), examined the difference between statistical significance and clinical significance.

Using several pain studies and the Pain Intensity numerical rating scale (PI-NRS) scale, Farrar et al. (70) found a reduction of 2 points or 30% on the 11-point pain scale to be clinically significant. Using the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) scale for osteoarthritis, Angst et al. (71) compared estimates derived from anchor- and distribution-based approaches; they determined sample sizes for specified changes signifying worsening and, separately, improvement. Using the chronic heart failure questionnaire (CHQ), Wyrwich et al. (7) also compared anchor and distribution-based approaches in determining that 1 SEM equals the MCID of 0.5 per item determined by the anchor-based approach. Finally, using the functional assessment of cancer therapy–Lung (FACT-L) questionnaire, Cella et al. (72) showed the convergence of three different complementary approaches on clinical significance.

Taken from Sprangers et al. (34), Table 4 provides a useful checklist of questions to help in interpretation of

Table 3. Examples of Studies for Determining Clinically Significant Change

Authors and Instrument–Anchor Used	Study Description	Study Results	Comments
Jaeschke, Singer and Guyatt (1989) (24) Chronic Respiratory Questionnaire & Chronic Health Failure Questionnaire. QOL dimensions: individual domains of dyspnea, fatigue and emotion. 7 point scale focus: change scores Anchor: Longitudinal within-patient rating of change since last visit A Very great deal worse (−7) to A Very great deal worse (+7)	75 patients QOL questionnaires completed at baseline and at 2,6,12 &24 weeks. Links QOL change scores with classification on global patient-rated anchor. Classification of change on AQLQ Somewhat worse: −3 to −1; +3 to +1. Moderately or good deal worse: −4 or −5; +4 or +5 A great deal worse: −6 or −7; +6 or +7	Global Rating of change None / Small / Moderate / Large Dyspnea 0.10 / 0.43 / 0.96 / 1.47 Fatigue 0.12 / 0.64 / 0.87 / 0.94 Emotion 0.02 / 0.49 / 0.81 / 0.86	Anchor-based approach One of the first studies to determine the magnitude of change that is clinically meaningful
Juniper et al (1994)(5) Asthma Quality of Life Questionnaire (AQLQ) QOL Dimensions: overall score and individual domains of activities, symptoms, emotional 7 point scale focus: change scores Anchor: Longitudinal within-patient rating of change since last visit "Worse": −1 hardly any worse to −7 a very great deal worse "Better": +1 hardly any better to a +7 a very great deal better	39 subjects with asthma treatment AQLQ completed before and after 8 week asthma treatment Links QOL change scores with classification on anchor. Classification of change on AQLQ Small: −2 or −3: +2 or +3 Moderate: −4 or −5; +4 or +5 Large: −6 or −7; +6 or +7	Global Rating of change 0–1 / 2–3 / 4–5 / 6–7 Overall QOL 0.11 / 0.52 / 1.03 / 2.29 Activity 0.12 / 0.47 / 0.87 / 1.83 Symptoms 0.20 / 0.49 / 1.13 / 2.21 Emotional 0.20 / 0.58 / 1.51 / 2.70	Anchor-based approach Based in part on these results, Glaxo-Wellcome obtained a HqoL promotional claim for salmeterol for noctural asthma
Kazis, Anderson & Meenan (1989) (64) Arthritis Impact Measurement Scale (AIMS). 64 items, of which 45 are health status questions. 9 scales 0–10 scale, with higher scores indicating worsehealth states.	299 patients with rheumatoid arthritis AIMS completed at beginning and end of 5 year-period.	Change Score / Effect Size Scale / $\bar{x}_5-\bar{x}_1$ / $SD_{Difference}$ / t / p-value / SRM Mobility 0.07 / 2.80 / 0.43 / 0.6 / +0.02 Physical Activity 0.41 / 2.57 / 2.76 / 0.006 / +0.17 Dexterity 0.42 / 3.63 / 2.00 / 0.046 / +0.12 Activities of Daily Living 0.02 / 1.76 / 0.20 / 0.8 / +0.01 Household Activities 0.05 / 1.61 / 0.54 / 0.6 / −0.03 Anxiety 0.19 / 2.01 / 1.63 / 0.1 / +0.09 Depression 0.25 / 1.71 / 2.53 / 0.012 / +0.14 Pain 0.96 / 2.49 / 6.67 / 0.001 / +0.42 Social Activity 0.36 / 2.11 / 2.95 / 0.003 / +0.17	Combines anchor-based and distributional approaches Compares statistical significance with clinical significance

(*continued*)

Table 3. (*Continued*)

Authors, and Instrument/Anchor Used	Study Description	Study Results	Comments
Farrar JT et al. (2001) (70) Pain Intensity Numerical Rating Scale (PI-NRS) 11-point pain scale: 0 = no pain to 10 = worst baseline score = mean of 7 diary prior to drug endpoint score = mean of last 7 diary entries focus: change scores Anchor: Longitudinal within-patient Global Impression Change (PGIC) Very much improved (1) to Very much worse (7).	10 chronic pain studies with 2724 subjects consisting of several placebo-controlled trials of pregabalin and covering several conditions (e.g., fibromylagia and osteoarthritis). Links clinical improvement on PI-NRS with anchor. Mean change among 'much improved' on PGIC Receiver operating characteristic (ROC) curve Favorable: much or very much improved Not favorable: otherwise	Clinically important difference = Reduction of about 2 points on PI-NRS; reduction of about 30% on PI-NRS. ROC curve analysis: sensitivity = 77% and specificity = 78% area under curve = 78% Consistent relationship between change in PI-NRS and PGIC regardless of study, disease type, age, sex, study results or treatment group. Higher base-line scores required larger raw scores for clinically important differences.	Anchor-based approach
Angst F. et al. (2001) Western Ontario and McMaster Universities Osteoarthritis Index: (WOMAC) QOL Dimensions: global score and individual domains of pain, stiffness and physical function. emotional 10 point scale focus: change scores for worsening and improving patients Anchor: Retrospective within-patient transition rating on health in general related to osteoarthritic joint 3 months ago Much worse Slightly worse Equal Slightly better Much better	122 patients with osteoarthritis of lower extremities Before and after rehabilitation (3 months) Links MCID to sample size for future studies Mean effect = mean difference from baseline to 3 months *within each transition group separately for global WOMAC and each domain* MCID for improvement = *Mean Effect ("slightly better")– Mean Effect ("equal")* MCID for worsening = *Mean Effect ("slightly worse")– Mean Effect ("equal")* Effect size (ES) = MCID/SD (baseline)	MCID and Sample Sizes WOMAC (range 0 to 10): Worsening MCID 1.10, ES 0.49, *n** 66; Improvement MCID 0.75, ES 0.33, *n** 142 Pain (5 items) Stiffness (2 items): Worsening MCID 0.51, ES 0.19, *n** 431; Improvement MCID 0.72, ES 0.27, *n** 216 Physical Function (17 items): Worsening MCID 1.33, ES 0.61, *n** 43; Improvement MCID 0.67, ES 0.31, *n** 167 Global (24 items): Worsening MCID 1.29, ES 0.62, *n** 42; Improvement MCID 0.67, ES 0.32, *n** 153 *Sample size per group, assumes 80% power, 0.05 significance level (two-tailed for two-sample t-test)	Combined anchor and distributional approaches Lower values of MCID for improvement (except stiffness) than worsening; improvement may be subjectively easier to notice Larger sample sizes needed for less responsive sub scale (i.e., stiffness)

Wyrwich et al.(1999) (67) Chronic Heart Failure Questionnaire: CHQ QOL domains *Dyspnea (patient-specific 3–5 items affected by chest pain) scored as one (extreme amount) to seven (none at all)* *fatigue (4 items), emotional function (7 items) scored on a 7 point scale: one (worst), seven (best).* *baseline to follow-up (6,12,18 mths)* *change scores combined* Anchor: *Retrospective within-patient global assessment of change over last 4 weeks.* "Worse": −1 hardly any worse to −7 a very great deal worse "Better": +1 hardly any better to a +7 a very great deal better	605 cardiac patients in an outpatient setting. Secondary analysis of data from a RCT. Anchor standard for MCID *comes from previous research on CHQ of about 0.5 average per item change for each domain* *Classification of change on anchor: improved, stable, declined.* Anchor-based method linked QOL change scores with classification on anchor. Classification of change on anchor Minimal clinically important: 1 to 3/ −3 to−1 Moderate clinically important: 4 to 5/−5 to −4. Large clinically important: 6 to 7/ −7 to −6.	1 SEM based on baseline SD and Cronbach's alpha 1 SEM (Dyspnea) = 2.41 CHQ points per 5 items *equates to 0.48 average per item* 1 SEM (Fatigue) = 2.10 CHQ points per 4 items *equates to 0.53 average per item* 1 SEM (Emot. Func.) = 2.90 CHQ points per 7 items *equates to 0.41 average per item* SEM concords highly with MCID standard of 0.50 per item.weighted kappa (1.0, 0.87, 0.91 for the 3 domains) used to assess degree of association between 1SEM and MCID.	Combines the anchor and distributional approaches. 1 SEM as MCID also found in independent study (7) using the Chronic Respiratory Disease Questionnaire
Cella et al. (2002) (22) Functional Assessment of Cancer Therapy–Lung Questionnaire: FACT-L. *7-item Lung Cancer Sub scale (LCS)* *its Trial Outcomes Index (TOI)* adds scores on physical well-being sub scale and functional well-being sub scale of FACT-L to LCS scores	Randomized trial with 599 patients *advanced non-small cell lung cancer* *3 chemotherapeutic regimens (no difference in FACT-L)* *measurements at baseline and week 12* Three Complementary Approaches to MCID 1) Group means based on baseline differences in LCS and TOI scores on following anchors: *prior 6-month weight loss* (<5% vs.≥5%) *performance status (normal vs. some symptoms)* *primary disease symptoms* (≤1 vs. <1) 2) Group means based on changes in LCS and TOI scores over time on following anchors: *response to treatment (complete/partial vs. stable vs. deterioration)* *time to disease progression (<median time, >median time)* 3) Distribution-based criteria *1/3 and 1/2 standard deviation change standard of baseline scores, at week 12 scores, change scores.* *one standard error of measurement (SEM) at baseline and at week 12. Patients classifies as "declined" or "improved" (if − or + change scores > 1SEM respectively) or "unchanged" (if change score<1 SEM)*	Approximate MCID for 3 approaches converged; Cohen's kappa used to compare between empirically derived categories (i.e., anchor-based) and distribution-based categories Clinically meaningful differences: 2 to 3 points for the LCS (on 0-to-28-point scale) *7.1 points (=2*100/28) on 0-to-100 point sale* 5 to 7 points for the TOI (on 0-to-84-point scale) *6 points (=5*100/84) on 0-to-100 point scale*	Combines and compares anchor and distributional approaches. Use of multiple clinical anchors to validate clinically meaningful difference.

Table 4. Checklist for Assessing Clinical Significance over Time in *QOL*[a]

What are the characteristics of the population for whom changes in QOL are reported?
- What are their disease (e.g., tumor type), treatment (e.g., duration), socio-demographic and cultural (e.g., age, ethnicity), and behavioral (e.g., alcohol use) characteristics?
- To what extent are the QOL data applicable to your patients?
- Is actual QOL status of individual patients reported (e.g., by providing confidence intervals, standard deviations, subgroup data, individual data plots), thus documenting the amount of individual variation in response to treatment?

Is the QOL questionnaire relevant, reliable, valid, and responsive to change?
- Is the questionnaire appropriate given the research objective and the rationale for QOL assessment?
- Is the questionnaire appropriate given the domains included and in light of the disease and population characteristics?
- Is the questionnaire reliable and valid? Is this information reported in the article?
- Is the questionnaire responsive to change? Is this information reported in the article?
- Is the questionnaire appropriate given practical considerations (e.g., regarding respondent burden and the availability of different language versions)?
- Are patients' baseline QOL scores close to the extremes of the response scale? Do the treatment groups differ in baseline QOL?

Are the timing and frequency of assessments adequate?
- Is a baseline assessment included?
- Is QOL assessed at appropriate times for determining minimally important change given the natural course of the disease?
- Is QOL assessed long enough to determine a clinical effect, taking disease stage into account?
- Is QOL assessed at appropriate times to document treatment course, clinical events, and post-treatment effects?
- Are standard research design procedures followed (e.g., avoidance of respondent burden, collection of data prior to treatment or consultation)?
- Is the timing of the QOL assessments similar across treatment arms?

Is the study adequately powered?
- Is the sample size appropriate for the research questions (e.g., by providing a power calculation)?
- Is a rationale and/or source for the anticipated effect size specified?
- Does the power calculation take into account: the scale range of the anticipated effect, the score distribution (i.e., magnitude and form), the number of outcome measures, and research hypothesis (i.e., equivalence versus difference)?

How are multiple QOL outcomes addressed in analyses?
- Is the adopted approach of handling multiplicity explicitly described?
- Which approach is taken: limiting the QOL outcomes, use of summary measures, adjustment of p-values, and/or multivariate statistical analysis and modeling?
- Did the interpretation of the results take the problem of multiple outcomes into account?

How are multiple time-points handled?
- Are the data presented in a meaningful and suitable way enabling an overview of QOL changes over time?
- Do the tabular and graphical presentations take the problems inherent in the data into account (e.g., presence of floor and ceiling effects, patient attrition)?
- Are the data appropriately analyzed (e.g., are all time points included, are missing data taken into account, are pre-treatment co-variates included)?
- Does the article provide sufficient information on the statistical models selected?

Can alternative explanations account for the observed change or lack of observed change?

Are dissimilar baseline characteristics adequately accounted for?
- Is the baseline QOL score used as a co-variate?
- Are missing data handled adequately?
- Does the article indicate how missing items within a questionnaire are handled?
- Does the article report the number of missing questionnaires at each scheduled assessment?
- Does the article report the reasons for missing questionnaires?
- Is there an association between patients' health status and missing QOL data?
- If patients with incomplete data are excluded from the analysis (e.g., by using complete case methods), does the article document that these are non-ignorable missing data?
- In cases of nonignorable missing data, are several analytical approaches presented to address possible bias in conclusions based on this QOL data set?
- Is observed survival difference combined with QOL in evaluating change?
- If patients have died in the course of the study, is mortality accounted for in the evaluation of QOL?
- Are summary indices (e.g., QALYs, Q-TWiST) or imputation techniques used?

Did the patient's QOL perspective change over time?
- Are changes in patient's internal standards, values, and/or the conceptualization of QOL explicitly measured?
- Are insignificant or small changes in QOL reported despite substantial changes in patient's health status (i.e., deterioration or improvement)?
- How likely is it that patients have changed their internal standards, values, and/or their conceptualization of QOL as a result of adaptation to deteriorating or improving health?

How is statistical significance translated into meaningful change?
- Does the article provide some guidance regarding the clinical importance of the observed change in QOL?
- To what extent is the statement of clinical importance appropriate and empirically warranted?

[a]Reprinted with permission from Ref. 34.

longitudinal, patient-derived QOL results presented in clinical trials and the clinical literature. These questions are based on the premise that detecting meaningful change depends on the adequacy of the research design, measurement quality, and data analysis.

RECENT DEVELOPMENTS

The One-Half Standard Deviation Rule

It would be desirable to simply define, at least initially, what a clinical significant result is likely to be. Emerging research comparing anchor- and distribution-based estimates provides an evolving standard as to what to use as an initial estimate (23). The anchor-based estimates averaging 0.5 per item on a 7-point scale appear to converge with an estimate of one-half standard deviation (SD) units. This latter estimate is derived through distribution-based methods, such as the effect size approach (22,64), SEM (7,67,68), and the standardized response mean (73). Potential moderating factors that could impact these estimates upward or downward are the method used to determine minimum difference estimates, the reliability of the measure and whether patients were suffering from acute or chronic conditions (23,74)

Empirical Rule Effect Size

Sloan et al. (27,28) have taken this concept one step further in the form of the ERES by combining Cohen's effect size categorization (22) with the empirical rule from statistical theory (75). The ERES is based on Tchebyschev's theorem and states that the distribution of any QOL tool is contained within six SDs of the observed values. The ERES entails the estimation of QOL change scores in terms of SD estimates, expressed as units on the theoretical range of a QOL instrument. Thus small, moderate, and large effect sizes for comparing QOL treatment groups turn out to be 3, 8, and 13%, respectively, of the theoretical range of any QOL tool. This simple and intuitive rule to identify the magnitude of clinically significant changes is likely to be easy for clinical researchers to comprehend. The rule can facilitate the design of clinical trials in terms of sample size calculations and interim monitoring of clinical trials. The ERES framework for a priori establishment of effect sizes is sample independent and thus an improvement over sample-dependent methods (5,21,76).

However, the simplicity of the ERES method gives rise to some challenges and questions. The theoretical range of the instrument is rarely observed in its entirety, necessitating the modification of the theoretical range to more practical limits before calculating the ERES estimate for one SD as necessarily 16.7% (i.e., one-sixth of distribution of observed values) of the range. Similarly, truncated distributions, where the patient population is homogeneously ill or uniformly healthy, can be accommodated by incorporating this knowledge into the definition of the appropriate range. These guidelines for clinical treatments can be used in the absence of other information, but will need modification in their application to idiosyncratic or unique clinical settings. More research is needed to examine the generalizability of such benchmarks across baseline patient health, severity of illness, and disease groups.

Group Change versus Individual Change

Distinctions should be made in determining the significance of change at the group versus the individual level. Every individual in a group does not experience the same change in outcomes (group level outcomes are assigned a mean change value). There is higher variability in individual responses than those of the group. Depending on the distribution of individual differences, the same group mean can have different implications for an individual (77).

The traversing of group and individual level QOL data entails procedures for moving from one level to the other involving two distinctive scientific traditions: deductive and inductive (31). A deductive approach is employed when one addresses the extent to which group data can be used to estimate clinical significance at the individual level. An inductive approach is used when one evaluates the extent to which individual change data can be brought to the group level to define clinical significance. Related to this is the fact that the lower end of a MID estimate may be useful for powering studies to detect meaningful differences between groups (with ES as low as 0.2), while the higher end of the MID estimate, especially for more sensitive tools, can be used to power studies detecting change at the individual level. Readers are advised to read Cella et al. (31) for a more detailed account.

Quality of Life as a "Soft" Endpoint

The "softness" of QOL as an endpoint, relative to say survival and tumor response, is cited as a particular barrier to implementation and interpretation of results (13). However, methodological and conceptual strides made in defining and measuring QOL, and the growing familiarity with the interpretation and potential utility of QOL data, make those concerns increasingly outdated.

Psychometric advances have been made in QOL assessment tools across disease areas (8,78–81). Funding opportunities to study QOL endpoints have allowed for study designs that are large enough to have power to detect meaningful differences (13). Moreover, accumulated experience with analyzing QOL endpoints have resulted in the recognition that their statistical challenges are no different from those of "hard" endpoints.

CONCLUDING REMARKS

Several suggestion on clinical significance are offered. First, the application of multiple strategies for determining clinical significant is recommended. Doing so would enable better interpretability and validity of clinically significant change, add to existing evidence of the magnitude of change that constitutes clinical significance, and would provide indicators of distributional parameters that create convergence or divergence in estimation of clinical significance. For example, Kolotkin et al. (46) found convergence between anchor- and distribution-based methods at

moderate level of impairment but wide disparities at mild and severe levels of impairment.

Second, more research in needed into the effect of psychometric properties (i.e., reliability, validity and responsiveness of QOL instruments) have in quantifying clinically meaningful change (4,62,82). Similarly, research into the psychometric properties of global rating and health transition scales used in anchor-based methods is also needed. Global ratings tend to be single item measures and may therefore fall short in terms of explaining complex QOL constructs. Anchoring assessment also tends to be positively correlated with post-treatment states but with near-zero correlation with pretreatment states, suggesting a recall bias (83) or response shift (84). More research is needed to address the cognitive process used by patients to retrospectively assess changes in health over time (30).

Third, baseline severity results in regression to the mean (RTM), an error-based artifact describing the statistical tendency of extreme scores to become less extreme at follow-up. Failure to take this into account may lead to false conclusions that patients with severe impairments at baseline have shown clinical significant change, when in fact this was just RTM. The RTM also has a greater impact upon data when the measure is less reliable (4,85). More research is also needed into the effect of baseline QOL impairment on magnitude of clinically meaningful change (4,47,48,66,86,87). Similar research is needed in terms of the generalizability of the standardized benchmarks for determining clinically meaningful change, especially for distribution-based methods (4,66). Specifically, how satisfactory are the evolving benchmarks (effect sizes of 0.2, 0.5, and 0.8 for small, moderate, and large change, respectively) across different dimensions of QOL (e.g., mental versus physical), different disease groups (e.g., arthritis versus cancer), respondents (e.g., patients versus clinicians), measures (e.g., generic versus disease specific) patient populations (e.g., older versus younger), or patient conditions (e.g., improving versus deteriorating)?

Finally, care must be taken in presenting results of studies in a way that is familiar to the user of the information. For example, translating clinical significance into a number needed to treat (NNT) and a proportion of patients achieving various degrees of clinical benefit relative to the control may provide a desirable way to present study results (30).

BIBLIOGRAPHY

Cited References

1. Aaronson N. Methodologic issues in assessing the quality of life of cancer patients. Cancer 1991;67(3 Suppl):844–850.
2. Cella D, Bonomi AE. Measuring quality of life. Oncology 1995;9(11 Suppl):47–60.
3. Berzon R. Understanding and using health-related quality of life instruments within clilnical research studies. Quality of Life assessment in clinical trials: methods and practice. Oxford UK; 2000. p 3–15.
4. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56:397–407.
5. Juniper EF, Guyatt GH, Willan A. Determining a minimal important change in a disease-specific quality of life questionnaire. J Clin Epidemiol 1994;47:81–87.
6. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991;59:12–19.
7. Wyrwich KW, Tiemey WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health related quality of life. J Clin Epidemiol 1999;52:861–873.
8. Cella D. Quality of life outcomes: measurement and validation. Oncology 1996;10(11 Suppl):233–246.
9. Sloan JA, O'Fallon JR, Summan VJ. Incorporating quality of life measurements in oncology clinical trials. Proceeding of the Biometrics Section of the American Statistical Association, 1998. p 282–287.
10. Spilker B. Quality of life and pharmacoeconomics in clinical trials. New York: Lippincott Raven; 1996.
11. Osoba D. What has been learned from measuring health-related quality of life in clinical oncology. Eur J Cancer 1999;35(11):1565–1570.
12. Sloan JA, Symonds T. Health-related quality of life measurement in clinical trials: when does a statistically significant change become relavant?, in Unpublished manuscript. 2003.
13. Frost MHSJ. Quality of Life Measures: A soft outcome—or is it? Am J Managed Care 2002;8(18, Supp.):S574–579.
14. Degner L, Sloan JA. Symptom distress in newly diagnosed ambulatory cancer patients as a preditor of survival in lung cancer. J Pain Symptom Mange 1995;10(6):423–431.
15. Chochinov HM, Kristjanson L. Dying to pay: the cost of end-of-life care. J Palliat Care 1998;14(4):5–15.
16. Silliman RA, Dukes KA, Sullivan LM. Breast cancer care in older women: sources of information, social support, and emotional health outcomes. Cancer 1998;83(4):706–711.
17. Spiegel D, Bloom JR, Kraemer H. Psychological support for cancer patients. Lancet 1989;2(8677):1447.
18. Sloan JA, Loprinzi CL, Kuross SA. Randomized comparison of four tools measureing overall quality of life in patients with advanced cancer. J Clin Oncol 1998;16:3662–3673.
19. Patrick DL, Erickson P. Applications of health status assessment to health policy. Qual Life Pharmacoeco Clin Trials 1996; 717–727.
20. Gold MR, Patrick DL, Torrance GW. Identifying and valuing: Cost Effectiveness in Health and Medicine. 1996; 82–134.
21. Juniper EF. The value and quality of life in Asthma. Eur Resp J 1997;7:333–337.
22. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale (NJ): Lawrence Erlbaum Associates; 1998.
23. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 2003;41(5):582–592.
24. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10(4):407–415.
25. Jones P. Interpreting thresholds for a clinically significant change in health status (quality of life) with treatment for asthma and COPD. Eur Resp J 2002;19:398–404.
26. Wright JG, The minimally important difference:who's to say what is important? J Clin Epidemiol 1996;49:1221–1222.
27. Sloan JA, et al. Detecting worms, ducks, and elephants: a simple approach for defining clinically relevant effects in quality of life measures. J Cancer Integrative Med 2003;1 (1):41–47.

28. Sloan JA. Practical guidelines for assessing the clinical significance of health-related QOL changes within clinical trials. Drug Inf J 2003;37:23–31.
29. Sloan JA, et al. Assessing clinical significance in measuring oncology patient quality of life: introduction to the symposium, content overview, and definition of terms. Mayo Clin Proc 2002;77:367–370.
30. Guyatt GH, et al. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 2002;77:371–383.
31. Cella D et al. Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life. Mayo Clin Proc 2002;77:384–392.
32. Sloan JA, et al. Assessing the clinical significance of single items relative to summated scores. Mayo Clin Proc 2002;77:479–487.
33. Frost MH, et al. Patient, clinician, and population perspectives on determining the clinical significance of quality-of-life scores. Mayo Clin Proc 2002;77:488–494.
34. Sprangers MAG, et al. Assessing meaningful change in quality of life over time: A users' guide for clinicians. Mayo Clin Proc 2002;77:561–571.
35. Symonds T. et al. The clinical significance of quality-of-life results: practical considerations for specific audiences. Mayo Clin Proc 2002;77:572–583.
36. Testa MA, Interpretation of quality-of-life outcomes issues that affect magnitude and meaning. Med Care 2000;38:II166–II174.
37. van Walraven CM, Moher JL, Bohm C, Laupacis A. Surveying physicians to determine the minimal important difference: implications for sample-size calculation. J Clin Epidemiol 1999;52:717–723.
38. Deyo RA, Inui TS. Toward clinical application of health status measures: sensitivity of scales to clinically important changes. Health Serv Res 1984;19:278–289.
39. Johnson PA, Goldman L, EJ O. Comparison of the medical outcomes study short-form 36-item health survey in black patients and white patients with acute chest pain. Med Care 1995;33:145–160.
40. Brook RH, Ware JE, Davies-Avery A. Conceptualization and measurement of health for adults in the health insurance study. 1979.
41. Testa M, Lenderking WR, Interpreting pharmacoeconcomic and quality-of-life clinical trial data for use in therpeutics. Pharmacoeconomics 1992;2:107.
42. Testa M, Simonson DC, Assessment of quality-of-life outcomes. New Engl J Med 1996;28:835–840.
43. Llewellyn-Thomas HA, Williams JI, Levy L. Using a trade-off techniques to assess patients' treatment preferences for benign prostatic hyperplasia. Med Decis Making 1996;16:262–272.
44. Stucki G, Liang MH, Fossel AH. Relative responsiveness of condition specific and health status measures in degenerative lumbar spinal stenosis. J Clin Epidemiol 1995;48:1369–1378.
45. Mossey JM, Shapiro E. Self-rated health: a predictor of mortaility among the elderly. Am J Public Health 1982;72: 800–808.
46. Kolotkin RL, Crosby RD, Kosloski KD. Development of a brief measure to assess quality of life in obesity. Obes Res 2001;9:102–111.
47. Deyo RA, Inui TS, LJ. Physical and psychosocial function in rheumatoid arthritis: clinical use of a self-adminstered health status instrument. Arch Intern Med 1992;142:879.
48. Kolotkin RL, Crosby RD, Williams GR. Integrating anchor-based and distribution-based methods to determine clinically meaningful change in obesity-specific quality of life. Qual Life Res 2002;11:670.
49. Lydick E, Epstein RS. Interpretation of quality of life changes. Qual Life Res 1993;2:221–226.
50. Schwartz N, Sudman S, Autobiographical memory and the validity of retrospective reports. New York: Springer-Verlag.
51. Wyrwich KW, Metz S, Babu AN. The reliability of retrospective change assessments. Qual Life Res 2002;11:636.
52. Mozes B, Maor Y, Shumueli A. Do we know what global ratins of health-related quality of life measure? Qual Life Res 1999;8:269–273.
53. Kirwan JR, Chaput de Sainttonge DM, Joyce CRB. Clinical judgment in rheumatoid arthritis. III. British rheumatologists' judgment of 'change in response to therapy.' Ann Rheum Dis 1984;43:686–694.
54. Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. Qual Life Res 2002;11:207–221.
55. Guyatt GH, Jaeschke R. Reassessing quality of life instruments in the evaluation of new drugs. Pharmacoeconomics 1997;12:616–626.
56. Lydick F, Yawn BP. Clinical interpretation of health-related quality of life data. Quality of Life assessment in clinical trials: methods and practice. Oxford (UK); 1998. p 299–314.
57. Hays RD, Wooley JM. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? Pharmacoeconomics 2000;18(5):419.
58. Wright J, Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol 1997;50:239–246.
59. Barber B, Santanello NC, Epstein RS. Impact of the global on patient perceivable change in an asthma specific QOL questionnaire. Qual Life Res 1996;5:117–122.
60. Ware J, et al. SF-36 health survey: manual and interpretation guide. Boston: The Health Institute; 1993.
61. Baker DW, Hays RD, Brook RH. Understanding changes in health status: is the floor phenomenon merely the last step of the staircase? Medical Care 1997;35:1–15.
62. Husted JA, Cook RJ, Farewll VT. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000;53:459–468.
63. Speer DC, Greenbaum PD. Five methods for computing significant individual client change and improvement rates: support for an individual growth curve approach. J Consult Clin Psychol 1995;63:1044–1048.
64. Kazis L, Anderson JJ, Meenan RS. Effect sizes for interpreting changes in health status. Med Care 1989;27(Suppl 3):S178–189.
65. Guyatt GH, Bombardier C, Tugwell PX. Measuring disease-specific quality of life in clinical trials. CMAJ 1986;134:889–895.
66. Samsa G, Edelman D, Rothman ML. Determining clinically important differences in health status measures: a general approach with illustration to the Health Utilities Index Mark II. Pharmacoeconomics, 1999;15:41–55.
67. Wyrwich KW, Nienaber NA, Tiemey WM. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. Med Care 1999;37:469–478.
68. Wyrwich KW, Tiemey WM, Wolinsky FD. Using the standard error of measurement to identify important changes on the Asthma Quality of Life Questionnaire. Qual Life Res 2002;11:1–7.
69. Miller GG, The magic number seven plus or minus two: some limits on our capacity for processing information. Psychol Rev 1956;63:81–97.
70. Farrar JT, et al. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 2001;94:149–158.
71. Angst F, Aeschlimann A, Stucki G. Smallest Detectable and Minimal Clinically Important Differences of Rehabilitation Intervention With Their Implication for Required Sample Sizes Using WOMAC and SF-36 Quality of Life Measurement

Instruments in Patients With Osteoarthritis of the Lower Extremities. Arthritis Care and Research 2001;45: 384–391.
72. Cella D, et al. What is a clinically meaningful change on the Functional Assessment of Cancer Therapy-Lung (FACT-L) Questionnaire? Results from Eastern Cooperative Oncology Group (ECOG) Study 5592. J Clin Epidemiol 2002;55:285–295.
73. McHorney C, Tarlov A. Individual-patient monitoring in clinical practice: are available health status measures adequate? Qual Life Res 1995;4:293–307.
74. Stewart AL, Greenfield S, Hays RD. Functional status and well-being of patients with chronic conditions: results from the medical outcomes study. JAMA 1989;262:907–913.
75. Pukelsheim F. The three sigma rule. Am Stat 1994;48:88–91.
76. Juniper EF, Guyatt GH, Feeny DH. Measuring quality of life in childhood asthma. Qual Life Res 1996;5:35–46.
77. Guyatt G, et al. Interpreting treatment effects in randomized trials. BMJ 1998;316:690–693.
78. Chassany O, et al. Patient-reported outcomes: the example of health-related quality of life - a European guidance document for the improved integration of health-related quality of life assessment in the drug regulatory process. Drug Inf J 2002;36:209–238.
79. Speilberger C. State-Trait Anxiety Inventory: STAI (Form Y). Palo Alto (CA): Consulting Psychologists Press Inc.; 1983.
80. Radloff L. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas 1977;1:385–481.
81. McNair DM, Lorr M, Droppleman LF. Profile of mood states manual. EdiTS 1992.
82. Hays RD, Anderson R, Revicki D. Psychometric considerations in evaluating health-related quality of life measures. Qual Life Res 1993;2:441–449.
83. Norman GR, Stratford PW, Regehr G. Methodological Problems in the Retrospective Computation of Responsiveness to Change: The Lessons of Cronbach. J Clin Epidemiol 1997;50(8):869–879.
84. Schwartz CE, Sprangers MAG. Methodological approaches for assesing response shift in longitudinal health-related quality-of-life research. Soc Sci Med 1999;48:1531–1548.
85. Moser MT, Weis J, Bartsch HH. How does regression to the mean affect thresholds of reliable change statistics? Simulations and examples for estimation of true change in cancer-related quality of life. Qual Life Res 2002;11:669.
86. McHorney C. Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. Ann Int Med 1997;127:743–750.
87. Stratford PW, Binkley J, Riddle DL. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. Phys Ther 1998;78:1186–1196.

See also HEART, ARTIFICIAL; HOME HEALTH CARE DEVICES; STATISTICAL METHODS.

RADIATION DETECTORS. See RADIATION PROTECTION INSTRUMENTATION.

RADIATION DOSE PLANNING, COMPUTER-AIDED

JONATHAN G. LI
University of Florida
Gainesville, Florida

LEI XING
Stanford University
Stanford, California

INTRODUCTION

The term radiation dose planning in radiation therapy refers to the process of designing treatment strategies for optimally delivering a desired radiation dose to the intended volume while minimizing the dose to healthy tissues as much as possible, and estimating the radiation dose throughout the irradiated volume. In pursuit of better energy deposition characteristics, higher and higher energy photon and electron sources have been developed and used throughout the history of radiation therapy. Ionizing radiation dose can be delivered externally using electron accelerators or radioactive sources, or internally by implanting radioactive sources in the tumor volume. Heavy particles (e.g., protons, neutrons, and carbons) have also been used in radiation therapy, although the high cost of these machines have limited their widespread application. Our discussion will be restrected to radiation dose planning with external photon beams generated with medical megavoltage electron linear accelerators (linacs), which is by far the most widely used method of radiation cancer treatment.

Most medical linacs are isocentrically mounted, that is, they can rotate around a horizontal axis in 360°. Combined with the rotation of the treatment couch, radiation can be directed toward the patient from all possible directions. Two pairs of collimators or jaws moving in the orthogonal direction are usually built in the linac head to collimate the beam into a square or rectangular shape with continuously variable field sizes. Figure 1 shows a typical medical linac. A treatment usually involves several beams from different directions with different beam weights and beam-modifying devices in order to deliver a uniform dose to the target and to limit dose to surrounding normal tissues. Commonly used beam-modifying devices include custom-made blocks, which further collimate the radiation beam into any arbitrary shape, and wedge filters, which are wedge-shaped metal absorbers placed in the path of the beam to cause a tilt of the resulting isodose curves in the patient. To achieve optimal results, computer-aided radiation dose planning has played an increasing role in radiation therapy.

Many steps are involved in planning a radiation treatment. One of the first steps in the process is to establish a three-dimensional (3D) patient anatomy model based on the patient's image information. Toward this goal, one needs to delineate the areas to be treated (targets) and any dose-limiting normal structures. Developments in 3D imaging, digital imaging processing, and multimodality imaging have greatly aided this process. Treatment strategies are then developed where radiation beams are chosen for optimal target coverage without delivering excessive dose to critical structures. Radiation dose throughout the irradiated volume is calculated, and the plan evaluated. Several trial and error efforts are usually required before a clinically acceptable plan is generated. Most of the treatment planning is now performed using computers with dedicated software called a treatment planning system (TPS). With the availability of fast processors and large random access memory (RAM), voluminous patient data depicting accurate 3D geometry and anatomy from a computed tomography (CT) scanner can be manipulated in a TPS, giving radiation oncologists and treatment planners better visualization of the internal structures and greater ability to tailor the treatment to the particular circumstances. Dose display and plan evaluation tools have made it easier to compare different treatment plans. This article concentrates on new developments in radiation dose planning since the first edition of this encyclopedia (1). The widespread clinical implementation of 3D planning techniques (e.g., virtual simulation, image registration, model-based dose calculation algorithms, and treatment plan optimization) have made a major impact on the current practice of radiation therapy. All of the new developments are computationally intensive and require large RAM for image processing. Their increasing role in radiation dose planning has tied closely to the development and availability of fast computers and large RAM.

VIRTUAL SIMULATION

Virtual simulation is a process of delineating target and normal organs and designing treatment field arrangements and portals on a computer, based on a detailed 3D model of a patient built from a sequence of closely spaced transverse images from a CT scanner. Virtual simulation is now widely used and has replaced conventional simulation for most of the treatment planning except in some simple or emergency cases. Conventionally, the patient simulation is done using a simulator with the patient in the treatment position. A conventional simulator duplicates a linac geometry, but uses a diagnostic kilovoltage X-ray tube to enhance image contrast. Two-dimensional (2D) projection radiographs are taken from various gantry positions that have been chosen for treatment. Target volume and normal structures are drawn on the 2D simulation films and correct positioning of the fields and shielding blocks can be obtained in relation to

Figure 1. Photograph of an isocentric medical electron linear accelerator (Elekta Precise, Elekta Inc.).

anatomical and external landmarks. These geometries are transferred to the linac for patient treatment. In contrast to conventional simulation, where the treatment fields are designed on 2D radiographs while the patient stays on the simulation table, virtual simulation relies on the 3D CT data acquired with the patient in the treatment position and using the same immobilization device as will be used for treatment. The volumetric CT data represent the "virtual" or digital patient. Target delineation and field design are then done off-line without the patient's presence using dedicated virtual simulation workstations or treatment planning systems. Figure 2 shows a screen capture of a commercial virtual simulation package with different views that aid in visualizing the patient's internal structures and in the selection of beams and beam portals for treatment.

A typical patient CT data set has > 100 axial slices, each of which contains 512×512 picture elements (pixels). With 16 bits per pixel, a CT data set can easily run over 50 MB. Manipulating, displaying, and storing such voluminous data sets require enormous computer resources and have only been made possible in the past two decades due to the dramatic advancements in computer hardware. Historically, the evolution of radiation therapy has been strongly dependent on the available computer and imaging technologies and this trend is expected to continue in the years to come as radiation therapy proceeds into an era of computer-controlled delivery and real time image guidance and feedback.

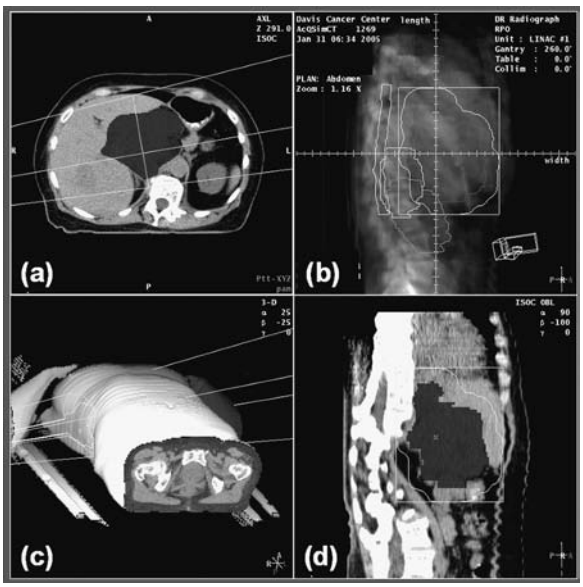


Figure 2. A screen capture of a commercial virtual simulation software (AcQsim 4.9.1, Philips Medical Systems) depicting the various tools available for treatment planning. (a) Axial view through the isocenter of a right posterior oblique (RPO) beam. (b) Digitally reconstructed radiograph of the RPO beam with the beam shape (yellow) and various structures projected onto it. (c) A 3D rendering of the patient's external contour. (d) Sagittal view through the middle of the patient.

An important step in virtual simulation is the generation of digitally reconstructed radiographs (DRR) for treatment planning and verification (2). A DRR is a computer-generated beam's-eye-view image that simulates the X-ray attenuation property and projection geometry of a conventional simulator. It is obtained by tracing the divergent path of X rays from the radiation source through the 3D patient CT data set onto a plane beyond the data set and orthogonal to the central ray. Fast ray-tracing algorithms have been developed to calculate the radiological path through a CT data set (3,4). Compared with radiographs from a conventional simulator, DRRs offer several distinct advantages. Whereas tumors and normal anatomic structures are sometimes difficult to visualize on a conventional radiograph because of the overlapping effect, they can usually be discerned much better on an axial CT image. The contours of the tumor and normal structures drawn on the CT images can be projected onto the DRR. This greatly helps the treatment planner in selecting beam geometries that will irradiate the target while avoiding critical anatomic structures. The DRRs can be generated quickly for any angular projection through the body, whereas using a simulator and film requires many minutes of patient setup and film developing for each projection. The brightness and contrast of a DRR can be digitally manipulated to bring out certain anatomic features. This can be used for patient setup verification by comparing the DRR with a radiographic image of the treatment field (portal image) obtained on the treatment machine. Figure 3 illustrates a DRR with several structures and the treatment field shape projected onto it and the corresponding portal image of the same

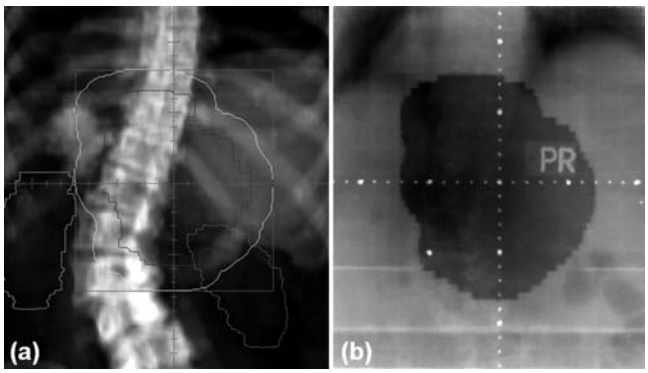


Figure 3. Comparison of a digitally reconstructed radiograph (a) and the corresponding portal film (b). The beam shape (yellow) was projected onto the DRR for treatment verification.

field. Such comparison verifies both the positioning of the patient and the treatment field shape.

IMAGE REGISTRATION

Radiotherapy treatment planning has been based mostly on CT images. Advantages of CT images include high spatial integrity, high spatial resolution, excellent bony structure depiction, and the ability to provide relative electron density information used for radiation dose calculation. However, CT images do not provide good soft tissue contrast. Moreover, CT images are anatomic in nature, that is, they provide snapshots of the patient's anatomy without any functional information of various organs or structures. Other imaging modalities, especially the magnetic resonance imaging (MRI) and positron emission tomography (PET), have been used increasingly in radiation therapy planning in conjunction with CT images. Magnetic resonance imaging provides better soft tissue contrast than CT images, and is the modality of choice when delineating treatment target of brain tumors. Positron emission tomography provides functional information about tumor metabolism and is a useful tool in tumor diagnosis, staging, target volume delineation and assessment of therapeutic response. However, current MRI and PET devices suffer from several drawbacks that make them unsuitable for radiotherapy planning as the sole modality. Imaging artifacts and geometric distortions exist in MR images. PET has a lower resolution than CT and contains no anatomy information of the normal structures. Information derived from MRI or PET needs to be fused or registered with the corresponding CT images for treatment planning.

Three-dimensional image registration aims at finding a geometric transformation that maps the volume elements (voxels) of one tomographic data set onto another tomographic data set of the same or different modality. Since different scans are done at different times and possibly with different patient immobilization devices, image registration is difficult to perform manually, and sophisticated computer algorithms have been developed for various registration applications (5). Some algorithms make use of common geometric features, such as points, lines, or surfaces identified on both image sets to determine the transformation. These features can be extracted either manually or automatically, and the accuracy in the identification of the common features directly affects the registration accuracy. Other image registration algorithms are based on information contained in the whole image sets (e.g., the image intensity values) and seek to maximize the amount of shared information in the two data sets. Image registration between images of different modalities based on shared information methods is especially challenging as different image modalities involve entirely different physical processes, and therefore voxels of the same tissues can appear very differently on different images. For example, high voxel intensity areas on a CT image (e.g., the bone) may correspond to dark areas on a MRI image. Figure 4 shows an axial CT image of a brain with the registered MR image. The registered secondary image (MR image in this case) is interpolated and resampled to give the axial view at the same axial position as the primary CT image. Contours of the tumor and normal structures drawn on one image are transferred and displayed automatically onto the other image.

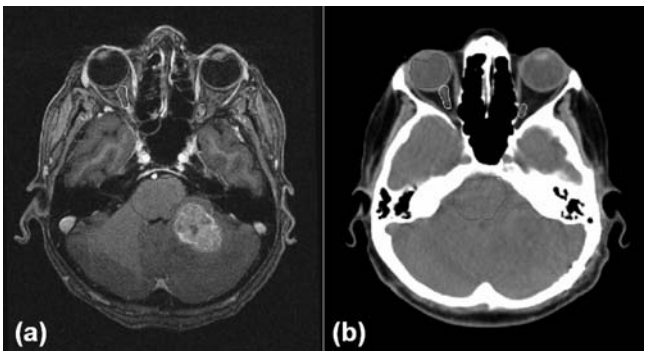


Figure 4. Registered MR (a) and CT (b) images of a brain tumor. The tumor (red contour) is readily seen on the MR image.

Various degrees of simplification have been assumed in the medical image registration. The simplest and widely used registration assumes that the transformation is rigid body, where only image translation and rotation is allowed with six degrees of freedom. Rigid body transformation preserves all distances in an image. The most important and successful application of rigid body registration is in the head, and particularly the brain. Image registration between MRI and CT has been the standard practice for the treatment planning of brain tumors in most cancer centers in the United States. In more general situations, where other parts of the body are involved and images are acquired under different conditions or using different modalities with the patient in different positions, more degrees of freedom are often needed. The simplest of the nonrigid body transformations, the affine transformation, introduces an additional six degrees of freedom with anisotropic scaling and skews. An affine transformation preserves collinearity (i.e., all points lying on a line initially still lie on a line after the transformation), parallelism (parallel lines stay parallel after the transformation), and ratios of distances in an image, but not necessarily angles or distances. Image registration with more degrees of freedom

than the affine transformation is an active area of research and its application in radiation therapy has so far been limited (6–8). This most general type of image registration technique is refereed to as deformable image registration. Commonly used deformable models can be categorized into two categories: free form B-spline and biomechanical finite element methods. Clinically, the need for a robust deformable registration technique is ever increasing because of the recent development in image-guided radiation therapy and much research is being carried out in this area. In four-dimensional (4D) radiation therapy of breast cancer, for example, where time-resolved CT images were used to monitor anatomic changes due to breathing, reconstruction of dose to different organs relied on the registration between the voxels at different phases of the breathing cycle (9). However, a detailed discussion of the subject is clearly beyond the scope of this article and the readers are referred to the references cited above.

The drive for more accurate registration between PET and CT images has led to the development of PET–CT, a new imaging technology that combines high quality PET and CT scanners in a single device (10,11). With PET–CT, patients undergo CT and PET scans sequentially under the same immobilization with a table translation, therefore providing simultaneous anatomic and metabolic information under almost identical conditions. Image registration becomes a simple process of correcting for the known table translation. An added advantage of PET–CT over PET only is the faster scan time, where the CT scans, which only take a few minutes, are used for attenuation correction. With PET only, the attenuation correction is obtained from a transmission scan, which takes on the order of 30 min. Although incorporating functional imaging in radiotherapy treatment planning is relatively new, the interest is increasing steadily and many studies have shown that new functional information would change patient management decisions in many disease sites (12–14).

RADIATION DOSE CALCULATION

Dose calculation plays a pivotal role in radiation therapy treatment planning. To achieve the expected therapeutic results, radiation dose distribution throughout the irradiated volume should be known to a desired degree of accuracy. Generally speaking, there are two major types of dose calculation algorithms: correction and model based. Correction-based methods compute the dose distributions in patients by correcting the dose distributions of similar geometries in a homogeneous water phantom for the beam modifiers, patient contours, tissue heterogeneities, and volume scattering effect. There are several algorithms for heterogeneity corrections. Two simple one-dimensional (1D) methods are the ratio of TPR, in which only densities along primary photon path are considered, and the power-law (or Batho) method, which takes the depth of the heterogeneity with respect to the depth of the point of measurement into account. Sontag and Cunningham (15) implemented the first algorithm, often referred to as the equivalent tissue/air ratio method, to estimate scatter dose in three dimensions and took advantage of the detailed anatomical information derived from CT images. Wong and Purdy (16) examined eight methods of photon inhomogeneity correction for their photon transport approximations and improved correction-based algorithms by introducing more realistic transport models into the calculations. The volume scattering effects (scatter dose as a function of field size and shape) are often computed by using the equivalent square field method and/or Clarkson integration (17). Some pencil beam methods, like the finite-size pencil beam algorithm (18), are also classified as the correction-based methods. Model-based algorithms simulate the treatment situation from first principle and can directly calculate the dose distributions in a patient for a given beam energy, geometry, beam modifiers, patient contour, and tissue heterogeneities. The kernel-based convolution–superposition and Monte Carlo method are representatives of the kind. These commonly used algorithms are briefly summarized below.

Correction-Based Methods

Calculation of radiation dose is conventionally done by interpolating from measured data in a water phantom and correcting for any nonstandard situations. To serve this purpose, large amounts of beam data need to be collected that would allow for data interpolation with reasonable degree of accuracy. Measurement is usually done with a computer-controlled automatic scanning system. Dose as a function of depth along the central axis of the beam and off-axis distance along the transverse directions is measured for a large number of field sizes at a fixed source-to-water surface distance (SSD). These measurements need to be repeated for both open and wedged fields. Dose as a function of depth along the central axis, when normalized to a given depth (usually the depth of maximum dose for a reference field size), is called the percentage depth dose (PDD) and was an important quantity in the early days of radiation therapy when most of the treatment setups were at a fixed SSD. Figure 5a illustrates the definition of PDD and can be measured readily with a scanning system. Modern radiotherapy with isocentric-mounted gantries typically uses fixed SAD setups. As the gantry rotates around the patient with the isocenter near the center of the tumor, the SSD (and the tumor depth) changes. The quantity most useful for dose calculation in these cases is the tissue/phantom ratio (TPR), which is defined as the ratio of dose on the central axis at depth d in a phantom to dose at the same point and field size at a reference depth d_{ref} and is shown in Fig. 5b and c. When the depth of dose maximum is chosen as d_{ref}, TPR is sometimes called the tissue-maximum ratio (TMR). Whereas PDD for the same field size varies significantly with SSD, TPR is essentially independent of SSD. Therefore, a single TPR table can be used for all SSDs. Dose calculations based on TPR or other TPR derivatives have been discussed extensively (1), especially in the textbooks of Johns and Cunningham (19) and Khan (20).

The limitations of correction-based methods are numerous. Patient geometry and internal structures can deviate significantly from a flat and homogeneous water phantom.

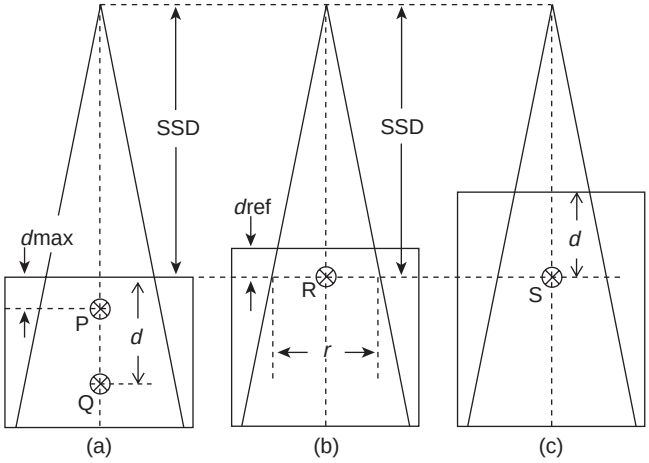


Figure 5. Schematic diagram illustrating the definition of the PDD and the phantom ratio (TPR). The PDD at a depth d is defined as the dose at the point, $D(Q)$, to the dose at point P, at the same SSD. It depends both on the SSD and the field size r. The TPR at depth d is defined as the ratio of dose at point S to dose at point R, at the same source-to-axis distance (SAD).

Although corrections can be made to account for the deviation, they are approximate and do not fully take into account the effect of secondary electron transport. This effect arises because secondary electrons ejected by megavoltage photons deposit a significant fraction of their energy far away from their point of origin. In regions where there is an imbalance between secondary electrons coming in and going out, such as near the interface between tissue–bone or tissue–air or in the beam penumbra, a condition known as charged particle disequilibrium exists. Correction-based methods are likely to produce erroneous results in these regions. Nowadays, correction-based dose calculation algorithms have mostly been replaced by model-based methods in commercial TPSs, where data measured in a water phantom under standard conditions are used to fit a few machine-specific model parameters, which in turn are used to calculate dose under the existing patient geometry. These include the convolution–superposition method and the Monte Carlo method. Nevertheless, correction-based methods are intuitive and have often been used as a quality assurance procedure to double-check the results of computer-generated treatment plans at one or a few calculation points for both conventional and intensity-modulated radiation therapy (21,22).

Convolution–Superposition Method

In the convolution–superposition techniques (23–27), the dose deposition is viewed as a superposition of appropriately weighted kernels of point irradiations and the superposition can be efficiently evaluated by means of convolution if the kernels are considered as spatially invariant. The kernels, representing the energy transport and dose deposition of secondary particles stemming from a point irradiation, can be calculated by Monte Carlo simulation (28). The Monte Carlo method computes dose distributions by simulating particle transport in a patient and will be described in the next section. Model-based algorithms are capable of accounting for electronic disequilibrium, and therefore are more accurate in dealing with tissue inhomogeneity and calculating dose in the electronic disequilibrium regions.

A thorough review of dose calculations in radiation therapy has been given by Ahnesjo and Aspradakis (27). There are a few different convolution–superposition methods, which can be divided into point kernel models and pencil kernel models. Mathematically, the dose at a spatial point, $D(r)$, comprises contributions from the shower of secondary particles resulting from primary interaction sites at r'. Assuming that the direction of all incident particles is parallel to the central axis throughout the beam, the total dose deposited by a monoenergetic beam irradiating a homogeneous medium can be expressed as a convolution operation:

$$D(r) = \int T(r\prime)A(r - r\prime)dr\prime$$

where $A(r - r')$ is the dose spread kernel that describes the mean fraction of energy deposited per unit volume at r per photon interaction at r', and $T(r')$ is the total energy released by primary X-ray interactions per unit mass, or TERMA. The above formula forms the basis for point kernel-based modes. Although the formulation of the point kernel model is simple and appealing, the demand on computer time is enormous due to the need for modeling various second-order beam characteristics. As emphasized by Ahnesjo and Aspradakis (27), there are three major issues that must be addressed for accurate dose calculation: broad primary beam spectral, beam divergence, and tissue density inhomogeneity. After all these factors are considered, the convolution–superposition dose calculation becomes a computationally intensive task. Therefore, their clinical implementation is tied closely to the availability of fast computers. The use of fast transform convolution techniques, such as the fast Fourier transform (FFT) method, to carry out the discrete convolution greatly facilitate the calculation process. Another widely accepted approach is the so-called collapsed cone convolution (26) based on an angular discretization of the kernel.

The pencil beam method is essentially a hybrid algorithm that fully accounts for beam modulations and field shapes, but relies on broad beam scaling/correction methods to handle heterogeneities and patient contour changes (18,27). The poly-energetic pencil beams are generally compiled from a linear combination of monoenergetic pencil beams within the constraints of a spectrum model to reproduce a set of depth-doses. The pencil beam kernels can also be determined by direct Monte Carlo calculation, or derived experimentally based on scatter factor differentiation.

Monte Carlo Method

Monte Carlo is a statistical simulation method that simulates the tracks of individual particles and all subsequently generated particles as they traverse the medium (28–30). The method takes into account all the relevant interactions and physical processes that take place in the medium. For each particle, the probability and types of interaction at a

point, its path length, and its direction are sampled from probability distributions governing the individual physical processes using machine-generated pseudo-random numbers. These particles and all the daughter products are followed until they either are fully absorbed or escape from the region of interest. Dose distribution and other macroscopic quantities can be calculated by simulating a large number of particle histories. Provided that the physical models used in the simulation are accurate, Monte Carlo simulation can accurately predict radiation dose distribution as it simulates particle transport and energy deposition from first principles. In particular, Monte Carlo simulation can calculate dose in regions of charged particle disequilibrium more accurately than any other existing dose calculation algorithms. For a detailed discussion on the Monte Carlo method in radiotherapy dose calculation, see the chapter on "Radiation therapy treatment planning, Monte Carlo calculations".

An intrinsic limitation of the Monte Carlo method is that the results contain statistical noise. The statistical error of Monte Carlo calculation is proportional to $1/\sqrt{N}$, where N is the number of simulated particle histories. To obtain dose distributions with acceptably small statistical uncertainty, a large number of particle histories need to be simulated, which makes the Monte Carlo method computationally intensive. The prohibitively long computation time has been the main obstacle for its routine clinical application. However, the computer speed has been increasing exponentially (Moore's law) since the initial application of the Monte Carlo method in medical physics in the 1970s and this trend is expected to continue. With the rapid increase in computer speed and the development of innovative variance reduction techniques (31), Monte Carlo simulation is fast becoming the next generation dose calculation engine for radiation treatment planning of photon and electron beams. The first commercial TPS that employs Monte Carlo dose calculation engine has already been released (PEREGRINE, North American Scientific) with dose calculation time on the order of minutes on two 2.4 GHz Pentium Xeon processors with a grid size of $0.5 \times 0.5 \times 0.5$ cm^3. The clinical impact of using Monte Carlo dose calculations is a subject of considerable interest and needs to be evaluated carefully (32).

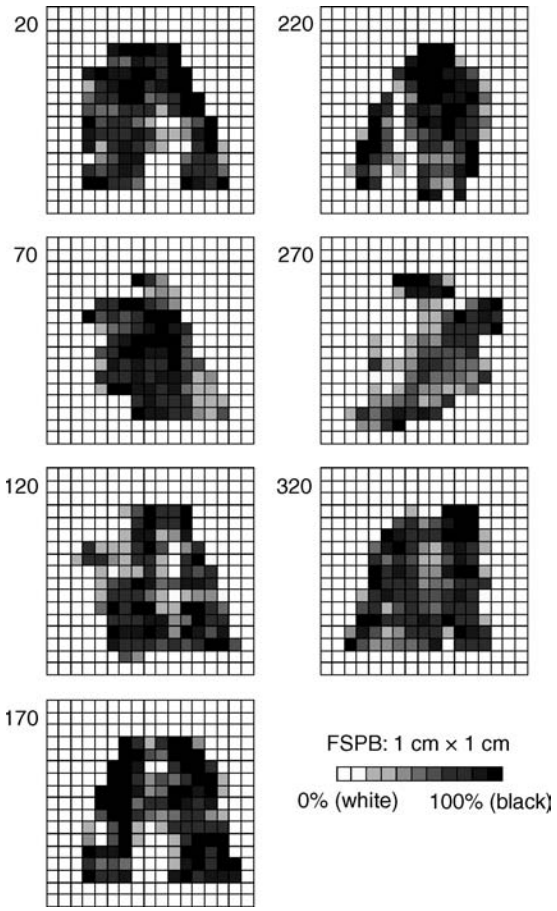


Figure 6. Fluence patterns of a seven-beam IMRT treatment plan of a head-and-neck case from a commercial inverse planning system (CORVUS 5.0, North American Scientific).

TREATMENT PLAN OPTIMIZATION

Many treatment-planning parameters affect the quality of a treatment plan. In conventional 3D conformal radiotherapy (3DCRT), the treatment parameters that are at the treatment planner's disposal include the beam modality and energy, number of beams, beam and treatment couch angles, wedge angles and orientations, radiation-defining blocks, and the weights of each beam. Since the number of beams used in 3DCRT is usually small (3∼5), clinically satisfactory plans can be produced manually in a trial-and-error fashion. Disease- and site-specific treatment techniques developed over the decades help to reduce the number of adjustable parameters significantly. With the development and clinical implementation of intensity-modulated radiation therapy (IMRT), where the fluence of each radiation beam can be modulated arbitrarily in order to achieve a highly conformal radiation dose distribution, manual planning is not practical, and computer optimization algorithms have to be used to design complicated beam fluences (33). Figure 6 shows the fluence maps of a seven-beam IMRT plan from a commercial inverse planning system used for the treatment of a head and neck tumor. Each beam is divided into a grid of 1×1 cm beamlets, and each beamlet can take a fixed number of fluence levels. Such fluence modulation can be achieved with a computer-controlled multileaf collimator (MLC). The treatment planning process where the desired goals (the output) (e.g., the desired dose distribution and dose and/or dose–volume constraints, are specified first, and computer optimization is used to determine the needed beam fluences (the input) is sometimes termed inverse planning.

As with any mathematical optimization problems, inverse planning starts with the construction of an objective function. The objective function, with its value the sole measure of the quality of a plan, guides the optimization algorithm. Additional constraints can also be imposed which limit the solution space. The compromise between

delivering a high dose to the target while limiting the dose to critical structures is implicitly built into the objective function. One of the simplest forms of the objective function that has been used often is the weighted least-squares function, which can be expressed as

$$F_{\text{obj}} = \frac{1}{N}\sum_{n=1}^{N} r_\sigma [d(n) - d_0(n)]^2$$

where $d_0(n)$ and $d(n)$ are the prescribed and calculated dose distributions, respectively, n is the voxel index, N is the total number of voxels, and r_σ is the relative importance factor of structure σ, which controls the tradeoffs between different structures. The calculated dose to voxel n can be obtained from the weighted sum of all the beamlets as

$$d(n) = \sum_{i=1}^{I}\sum_{j=1}^{J(i)} w_{ij} D_{ij}(n)$$

where I is the total number of beams, $J(i)$ is the total number of beamlets of the ith beam, $D_{ij}(n)$ is the relative dose contribution from the jth beamlet of the ith beam to voxel n, and w_{ij} is the weight of the beamlet (i, j). The parameter $D_{ij}(n)$ can be precalculated, so minimization of F_{obj} with respect to w_{ij} produces an optimal plan in the least-squares sense. The choice of the size of the dose calculation grid is an important consideration in IMRT plan optimization. While larger grid sizes reduce the model size and can speed up the computation considerably, too large a grid size will introduce aliasing artifacts. An information theory-based Fourier analysis of a 1×1 cm 6 MV photon beamlet from a medical linac predicted that an isotropic dose grid with < 2.5 mm spacing is sufficient to prevent dose errors larger than a percent (34). The tradeoffs between target coverage and critical structure sparing, which is controlled by the weighting factors r_σ, are usually not known *a priori*, and iterative adjustments are required to tailor the treatment plan to each particular circumstance. Algorithms aiming to automate the selection of the weighting factors have been proposed, which promises to significantly reduce the labor-intensive effort of the trial-and-error determination of the factors. In addition, the concept of intravoxel tradeoff has been introduced and an effective approach to model the tradeoff based on voxel specific weighting factors has been proposed (35,36). To obtain an adequate set of local weighting factors with a manageable amount of computing time, algorithms based on *a priori* dosimetric capability information and *a posteriori* adaptive algorithms were developed. With the introduction of intravoxel tradeoff, the IMRT dose distribution has been remarkably improved in comparison with the conventional plan obtained with structurally uniform weighting factors.

Clinical implementations of IMRT are dominated by dose- and/or dose–volume-based objective functions. Other forms of objective functions, particularly those employing biological indexes, such as the tumor control probability (TCP), normal tissue complication probability (NTCP), and equivalent uniform dose (EUD) (37,38), have also been applied to IMRT plan optimization. The use of biological indexes is especially appealing, as these are the ultimate measures of treatment outcomes. However, there is a lack of clinical and biological data to support these models, and their clinical use at present is not warranted. For example, current TCP models do not contain spatial information, and a cold spot would have the same adverse effect on the TCP irrespective of its location. In reality, the treatment outcome very much depends on the location of the cold spot, that is, whether it is in the periphery or in the middle of the target. Moreover, the TCP-, NTCP-, and EUD-based biological models as they are currently implemented are equivalent to voxel dose-based physical models in a multicriteria framework (39). It is expected that as radiation biology research leads to more robust models, biological based models will become more widely adopted in radiation therapy treatment plan optimization.

A practical approach to bridge the gap between biologically insensible physics-based formalism and a clinically impractical biology-based model is to establish a clinical outcome driven objective function by seamlessly incorporating a clinical endpoint to guide the treatment plan optimization process. Indeed, currently available dose-based objective functions do not truly reflect the nonlinear relationship between the dose and the response of tumors and tissues. On the other hand, biologically based inverse planning involves the use of a number of model parameters whose values are not accurately known and entails a prescription in terms of biological indexes. Recently, Yang and Xing proposed an effective method for formalizing the existing clinical knowledge and integrating clinical endpoint data into inverse planning (40). In their approach, the dose–volume status of a structure was characterized by using the effective volume in the voxel domain. A new objective function was constructed with incorporation of the volumetric information of the system so that the figure of merit of a given IMRT plan depends not only on the dose deviation from the desired dose distribution, but also the dose–volume status of the involved organs. The incorporation of clinical knowledge allows us to obtain better IMRT plans that would otherwise be unattainable.

The considerable interest in developing faster and more robust IMRT optimization algorithms has led to research collaboration between the radiation oncology community and the operations research (OR) community (41). The OR community has long investigated various optimization technologies and has the expertise in addressing large scale, complex optimization problems [see, e.g., the excellent textbooks of Winston (42) and Chong and Zak (43)]. Such collaboration is expected to enhance the IMRT model and algorithm development tremendously due to the large sizes of the problems in IMRT optimization combined with the clinical need to quickly solve each optimization for interactive plan review. For example, to avoid the nonconvex nature of conventional dose–volume constraints (44), Romeijn et al. introduced novel, convex dose–volume constraints that allowed them to formulate the IMRT fluence map optimization as a linear programming (LP) model (45). Using an industrial LP solver (CPLEX 8, ILOG

Inc.), a seven-field head-and-neck case with ~190,000 constraints, ~221,000 variables, and ~1,100,000 nonzero elements in the constraint matrix has been solved to global optimality in ~ 2 min of computation time on a 2.5 GHz Pentium 4 personal computer with 1 GB of RAM. While fluence map optimization has been studied extensively and clinically implemented, other challenging problems, such as finding the optimal number of beams and their angles, and fluence map optimization in the presence of organ motion, have not been solved satisfactorily. The problem of beam angle optimization has an enormous search space (46). The goodness of a chosen set of beam angles is not known until the fluence map optimization is performed. Current computer technology is not fast enough to solve such a nested optimization problem for routine clinical use. It is hoped that the collaboration with the OR community would help to develop novel and efficient algorithms in radiotherapy plan optimization.

DOSE DISPLAY AND PLAN EVALUATION

Treatment plans are evaluated based on the dose coverage and dose uniformity of the target, dose- to-sensitive normal structures, and magnitudes and locations of any hot and cold spots. This is best served by displaying the dose as isodose lines and superimposing them on the corresponding 2D image. Images can be displayed in the axial, sagittal, or coronal planes with the targets and normal structures delineated and dose distribution can be evaluated by going through the entire irradiated volume slice by slice. Figure 7 shows the isodose distribution of a head-and-neck IMRT plan in the three orthogonal planes overlaid on the corresponding CT images. Other graphic visualization tools, such as displaying the isodose in 3D as an iso-surface (47), have also been developed.

For quantitative plan evaluation, dose statistics can be calculated from the 3D dose distribution. Mean, maximum and minimum doses to the targets and mean and maximum doses to the critical structures can aid in plan selection. Dose uniformity throughout the target volume can be assessed from the standard deviations of the target dose distribution. A conformity index, defined as the quotient of the treated volume and the planning target volume (PTV) when the treated volume totally encompasses the PTV, has been introduced (48) to quantitatively evaluate the amount of normal tissue treated. Biological model-based TCP and NTCP are now available on some commercial TPS. However, the large uncertainty in the biological models combined with the lack of clinical experience limit their application in current clinical practice.

A very useful tool in 3D plan evaluation is the calculation of cumulative dose–volume histograms (DVHs) (49). The DVHs summarize 3D dose-distribution data into 2D histograms and are helpful in rapid screening of rival plans. An example of the DVHs of various structures of a head-and-neck IMRT plan is shown in Fig. 8. The DVH of a structure is calculated as

$$\%V(D) = \frac{\sum_{d \geq D} v(d)}{V_0}$$

where $v(d)$ is the volume of a voxel in the structure receiving dose d and V_0 is the total volume of the structure. Therefore, each point $\%V(D)$ on a DVH curve represents the percent volume of the structure receiving doses $\geq D$. From the DVH, dose coverage and dose uniformity of the target(s) can easily be appreciated. The relevant dosimetric parameter to some critical structures that display serial nature such as the spinal cord and the brain stem is the maximum dose (48) and it can be read directly from the DVH. The functionality of many normal structures displays a dose–volume effect (50). For example, in trying to preserve salivary function for head-and-neck patients using IMRT, planning criteria of at least 50% volume of either parotid gland receiving doses < 30 Gy have been established (51,52). The development of $\geq$ Grade 2 pneumonitis in patients after radiation treatment for non-small cell lung cancer was found to be significantly correlated

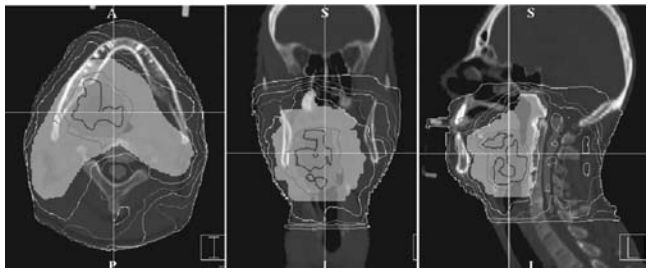


Figure 7. Overlay of isodose lines on an axial, coronal, and sagittal planes of a head-and-neck cancer. The IMRT plan was generated using a commercial inverse treatment planning system (CORVUS 5.0, North American Scientific). The isodose lines are, from inside out, at dose levels of 76, 72, 49.5, 40, 30, 20, and 10 Gy. The gross tumor volume (red) and the subclinical target volume (yellow) are shown as color washes to help evaluate the quality of the plan.

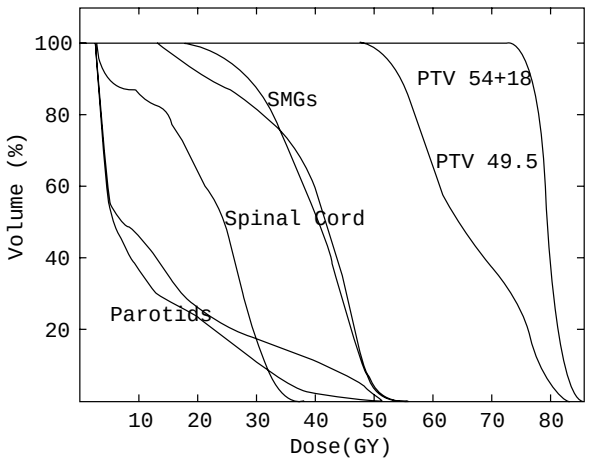


Figure 8. Cumulative DVH of different structures of a head-and-neck IMRT treatment plan. The prescribed doses to the targets are 72 Gy for the gross tumor volume (labeled "PTV 54 + 18") and 49.5 Gy for the subclinical target volume (labeled "PTV 49.5"). Also shown are the DVHs of the left and right parotids, the left and right submandibular glands (SMGs), and the spinal cord.

with the percent volume of the total lung exceeding 20 Gy (53). These dose–volume relations can be obtained directly from the DVH. Dose–volume histograms from two or more competing plans can also be overlaid on top of each other to facilitate plan comparison. The increasing use of DVHs in routine clinical treatment planning over the past decade is a direct result of increased computer power and fast and more accurate dose calculation algorithms which make dose calculation in the entire irradiated volume possible.

The DVH should be used with caution. Since a DVH is a 2D histogram of a 3D dose distribution, it does not provide any spatial information. For example, Fig. 8 indicates that $\sim 2\%$ of the volume of the target labeled "PTV 49.5" did not receive the prescribed dose of 49.5 Gy. However, the locations of the underdosed volume is unknown. While underdosed areas near the periphery of the PTV may be acceptable, they are clearly unacceptable in the middle of the PTV where the gross tumor volume is located. Thus, DVHs must replace isodose distributions. Rather, it is a tool to enhance our ability to choose between different plans. This is especially true when evaluating IMRT plans, as the dose distributions of IMRT plans are spatially independent (54,55). A good DVH is therefore a necessary but not sufficient condition for a clinically acceptable treatment plan.

SUMMARY

In the first edition of this Encyclopedia $>$ 17 years ago (1), Dr. Radhe Mohan envisioned that rapid developments in computer technology would make CT-based 3D treatment planning, including visualization of the anatomic structures and target region in 3D, fast and accurate radiation dose calculation in the entire irradiated volume, and sophisticated graphic and analytic tools for treatment plan display and evaluation, clinically routine. Radiation therapy has certainly gone through a series of revolutionary changes over the past 17 years. Three-dimensional treatment planning is now a standard practice and has mostly replaced 2D treatment planning. Monte Carlo dose calculation, which used to be considered as a luxury research tool, is now being incorporated into treatment planning systems. The fast development and widespread clinical implementation of IMRT over the past decade have provided the radiation oncology discipline a powerful tool to deliver highly conformal doses to the tumor volume while improving sensitive structure sparing. We are just entering an era of image-guided radiation therapy with exciting new developments. Recently, these have spurred efforts toward implementation of time-resolved or 4D imaging techniques, such as 4D CT (56–58) and 4D PET (59), into radiation oncology treatment planning and delivery. Currently, 4D imaging information is mainly being used as a tool for better definition of patient specific margins in the treatment of tumors in the thorax and upper abdomen. The ultimate goal is to establish a new scheme of 4D radiation therapy, where the 4D patient model is used to guide 4D planning and treatment. While there is still a long way to go toward this goal, much progress has been made, especially in the area of 4D inverse planning. Another important area that is under intense investigation is biologically conformal radiation therapy (BCRT) (60–62). Different from the current radiation therapy, which is aimed at producing a homogeneous target dose under the assumption of uniform biology within the target volume, BCRT takes the inhomogeneous biological information into account and produces customized non-uniform dose distributions on a patient specific basis. There are a number of challenges to accomplish the new radiation treatment paradigm, such as the determination of the distribution of biological properties of the tumor and critical structures, the prescription of the desired dose distribution for inverse planning, and the technique for inverse planning to generate most faithfully the prescribed nonuniform dose distribution. The most fundamental issue is, perhaps, how to extract the fundamental biological distribution for a given patient with biological imaging techniques and how to link the imaging information to the radiation dose distribution to maximize tumor cell killing. Hopefully, with the multidisciplinary efforts, the issues related to molecular imaging, image quantitation, planning, and clinical decision making would be resolved in the next decade. This will lead to truly individualized radiation therapy, and eventually individualized medicine when combined with the efforts in molecular medicine.

ACKNOWLEDGMENTS

JGL would like to thank Debbie Louis, CMD, for useful discussions and LX wishes to acknowledge support from the Department of Defense (PC040282) and the National Cancer Institute (5R01 CA98523-01).

BIBLIOGRAPHY

1. Mohan R. Radiation dose planning, computer-aided. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation. New York: John Wiley & Sons, Inc.; 1988. p 2397–2407.
2. Sherouse GW, Novins K, Chaney EL. Computation of digitally reconstructed radiographs for use in radiotherapy treatment design. Int J Radiat Oncol Biol Phys 1990;18:651–658.
3. Siddon RL. Prism representation: a 3D ray-tracing algorithm for radiotherapy applications. Phys Med Biol 1985;30:817–824.
4. Siddon RL. Fast calculation of the exact radiological path for a three-dimensional CT array. Med Phys 1985;12:252–255.
5. Hill DLG, et al. Medical image registration. Phys Med Biol 2001;46:R1–R45.
6. Wang H, et al. Implementation and validation of a three-dimensional deformable registration algorithm for targeted prostate cancer radiotherapy. Int J Radiat Oncol Biol Phys 2005;61:725–735.
7. Wu X, Dibiase SJ, Gullapalli R, Yu CX. Deformable image registration for the use of magnetic resonance spectroscopy in prostate treatment planning. Int J Radiat Oncol Biol Phys 2004;58:1577–1583.
8. Lu W, et al. Fast free-form deformable registration via calculus of variations. Phys Med Biol 2004;49:3067–3087.

9. Ding M, et al. Dose correlation for thoracic motion in radiation therapy of breast cancer. Med Phys 2003;30:2520–2529.
10. Beyer T, et al. A combined PET/CT scanner for clinical oncology. J Nucl Med 2000;41:1369–1379.
11. Schoder H, et al. PET/CT: a new imaging technology in nuclear medicine. Eur J Nucl Med Mol Imaging 2003;30:1419–1437.
12. Esthappan J, et al. Treatment planning guidelines regarding the use of CT/PET-guided IMRT for cervical carcinoma with positive paraaortic lymph nodes. Int J Radiat Oncol Biol Phys 2004;58:1289–1297.
13. van Der Wel A, et al. Increased therapeutic ratio by 18FDG-PET CT planning in patients with clinical CT stage N2-N3M0 non-small-cell lung cancer: a modeling study. Int J Radiat Oncol Biol Phys 2005;61:649–655.
14. Mac Manus MP, et al. F-18 fluorodeoxyglucose positron emission tomography staging in radical radiotherapy candidates with nonsmall cell lung carcinoma: powerful correlation with survival and high impact on treatment. Cancer 2001;92:886–895.
15. Sontag MR, Cunningham JR. The equivalent tissue-air ratio method for making absorbed dose calculations in a heterogeneous medium. Radiology 1978;129:787–794.
16. Wong JW, Purdy JA. On methods of inhomogeneity corrections for photon transport. Med Phys 1990;17:807–814.
17. Clarkson JR. A note on depth doses in fields of irregular shape. Br J Radiol 1941;14:265–268.
18. Bourland JD, Chaney EL. A finite-size pencil beam model for photon dose calculations in three dimensions. Med Phys 1992;19:1401–1412.
19. Johns HE, Cunningham JR. The physics of radiology, 4th ed. Springfield (IL): Charles C. Thomas; 1983.
20. Khan FM. The physics of radiation therapy, 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2003.
21. Kung JH, Chen GT, Kuchnir FK. A monitor unit verification calculation in intensity modulated radiotherapy as a dosimetry quality assurance. Med Phys 2000;27:2226–2230.
22. Xing L, et al. Monitor unit calculation for an intensity modulated photon field by a simple scatter-summation algorithm. Phys Med Biol 2000;45:N1–N7.
23. Mackie TR, Scrimger JW, Battista JJ. A convolution method of calculating dose for 15 MV X-rays. Med Phys 1985;12:188–196.
24. Boyer AL, Mok EC. A photon dose distribution employing convolution calculations. Med Phys 1985;12:169–177.
25. Mackie TR, et al. Generation of photon energy deposition kernels using the EGS Monte Carlo code. Phys Med Biol 1988;33:1–20.
26. Ahnesjo A. Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media. Med Phys 1989;16:577–592.
27. Ahnesjo A, Aspradakis MM. Dose calculations for external photon beams in radiotherapy. Phys Med Biol 1999;44:R99–R155.
28. Nelson WR, Hirayama H, Rogers DWO. The EGS4 code system. Stanford Linear Accelerator Center Report 1985; SLAC-265.
29. Ma C-M, Jiang SB. Monte Carlo modeling of electron beams from medical accelerators. Phys Med Biol 1999;44:R157–R189.
30. Verhaegen F, Seuntjens J. Monte Carlo modeling of external radiotherapy photon beams. Phys Med Biol 2003;48:R107–R164.
31. Bielajew AF, Rogers DWO. Variance Reduction Techniques. In: Jenkins TM, Nelson WR, Rindi A, editors. Monte Carlo Transport of Electrons and Photons. New York: Plenum; 1988. p 407–419.
32. Boudreau C, et al. IMRT head and neck treatment planning with a commercially available Monte Carlo based planning system. Phys Med Biol 2005;50:879–890.
33. Shepard DM, Ferris MC, Olivera GH, Mackie TR. Optimizing the delivery of radiation therapy to cancer patients. SIAM Rev 1999;41:721–744.
34. Dempsey JF, et al. A Fourier analysis of the dose grid resolution required for accurate IMRT fluence map optimization. Med Phys 2005;32:380–388.
35. Yang Y, Xing L. Inverse treatment planning with adaptively evolving voxel-dependent penalty scheme. Med Phys 2004;31: 2839–2844.
36. Shou Z, et al. Quantitation of the a priori dosimetric capabilities of spatial points in inverse planning and its significant implication in defining IMRT solution space. Phys Med Biol 2005;50:1469–1482.
37. Wu QW, Mohan R, Niemierko A, Schmidt-Ullrich R. Optimization of intensity-modulated radiotherapy plans based on the equivalent uniform dose. Int J Radiat Oncol Biol Phys 2002;52:224–235.
38. Thieke C, Bortfeld T, Niemierko A, Nill S. From physical dose constraints to equivalent uniform dose constraints in inverse radiotherapy planning. Med Phys 2002;30:2332–2339.
39. Romeijn HE, Dempsey JF, Li JG. A unifying framework for multi-criteria fluence map optimization models. Phys Med Biol 2004;49:1991–2013.
40. Yang Y, Xing L. Clinical knowledge-based inverse treatment planning. Phys Med Biol 2004;49:5101–5117.
41. Langer M, et al. Operations research applied to radiotherapy, an NCI-NSF-sponsored workshop—February 7–9, 2002. Int J Radiat Oncol Biol Phys 2003;57:762–768.
42. Winston WL. Operations Research: Applications and Algorithms, 4th ed. Belmont: Thomson Learning; 2004.
43. Chong EKP, Zak SH. An Introduction to Optimization, 2nd ed. New York: John Wiley & Sons, Inc.; 2001.
44. Deasy JO. Multiple local minima in radiotherapy optimization problems with dose-volume constraints. Med Phys 1997;24: 1157–1161.
45. Romeijn HE, et al. A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning. Phys Med Biol 2003;48:3521–3542.
46. Stein J, et al. Number and orientations of beams in intensity-modulated radiation treatments. Med Phys 1997;24:149–160.
47. Schreibmann E, Theodorou K, Kappas C, Xing L. A software package for dose visualization in IMRT. XIVth Int Conf Comp Radiat Thera 2004; 700–703.
48. International Commission on Radiation Units and Measurements, Prescribing, recording and reporting photon beam therapy (Supplement to ICRU report 50), ICRU report 62, Bethesda (MD); 1999.
49. Drzymala RE, et al. Dose-volume histograms. Int J Radiat Oncol Biol Phys 1991;21:71–78.
50. Emami B, et al. Tolerance of normal tissue to therapeutic irradiation. Int J Radiat Oncol Biol Phys 1991;21:109–122.
51. Eisbruch A, Chao KSC, Garden A. Phase I/II study of conformal and intensity modulated irradiation for oropharyngeal cancer. Radiation Therapy Oncology Group protocol 0022, 2001.
52. Lee N, Garden A, Kramer A, Xia P. A phase II study of intensity modulated radiation therapy (IMRT) +/− chemotherapy for nasopharyngeal cancer. Radiation Therapy Oncology Group protocol 0225, 2003.
53. Graham MV, et al. Clinical dose-volume histogram analysis for pneumonitis after 3D treatment for non-small cell lung cancer (NSCLC). Int J Radiat Oncol Biol Phys 1999;45:323–329.
54. Xing L, Li JG. Computer verification of fluence map for intensity modulated radiation therapy. Med Phys 2000;27: 2084–2092.
55. Chao KSC, Blanco AI, Dempsey JF. A conceptual model integrating spatial information to assess target volume coverage for IMRT treatment planning. Int J Radiat Oncol Biol Phys 2003;56:1438–1449.

56. Pan T, Lee TY, Rietzel E, Chen GT. 4D-CT imaging of a volume influenced by respiratory motion on multi-slice CT. Med Phys 2004;31:333–340.
57. Keall PJ, et al. Acquiring 4D thoracic CT scans using a multislice helical method. Phys Med Biol 2004;49:2053–2067.
58. Low DA, et al. A method for the reconstruction of four-dimensional synchronized CT scans acquired during free breathing. Med Phys 2003;30:1254–1263.
59. Nehmeh SA, et al. Four-dimensional (4D) PET/CT imaging of the thorax. Med Phys 2004;31:3179–3186.
60. Alber M, Nusslin F. An objective function for radiation treatment optimization based on local biological measures. Phys Med Biol 1999;44:479–493.
61. Xing L, et al. Inverse Planning for Functional Image-Guided IMRT. Phys Med Biol 2002;47:3567–3578.
62. Ling CC, et al. Towards multidimensional radiotherapy (MD-CRT): biological imaging and biological conformality. Int J Radiat Oncol Biol Phys 2000;47:551–560.

See also PHANTOM MATERIALS IN RADIOLOGY; RADIATION DOSIMETRY FOR ONCOLOGY; RADIATION DOSIMETRY, THREE-DIMENSIONAL; RADIATION THERAPY TREATMENT PLANNING, MONTE CARLO CALCULATIONS IN; RADIOTHERAPY TREATMENT PLANNING, OPTIMIZATION OF.

RADIATION DOSIMETRY FOR ONCOLOGY

MALCOM MCEWEN
National Research Council
of Canada
Ontario, Canada

INTRODUCTION

Cancer is a disease that touches everyone, either directly or through a close friend or relative, and radiotherapy is one of the primary modalities for treating cancer. The intent may be a full cure or to relieve pain associated with the cancer and it is either used alone or in conjunction with other techniques, such as surgery or chemotherapy. The aim of radiotherapy is to use ionizing radiation (usually either high energy photons or electrons) to destroy the tumor while at the same time sparing healthy tissues. In the delivery of such treatments the quantity of interest is the absorbed dose (defined as the energy deposited per unit mass) as it can be used to estimate the biological effect of the radiation (i.e., cell killing). Too high a dose will kill all the cancerous cells, but will produce significant side effects due to damage to other organs. Too low a dose will leave some malignant cells alive, which can develop into a new tumor. One of the primary concerns in radiotherapy is therefore delivering the correct dose to destroy the tumor with the minimum of side effects and a fine line exists between under and over dosing. The allowable error in the delivered dose depends on many factors, such as the type and location of the tumor, but for some cancers can be as little as 3–4%.

The output of the machines that produce the radiation for radiotherapy (linear accelerators, X-ray tubes, radioactive sources) must be known to a very high accuracy and a great deal of dosimetry work is carried out in the radiotherapy clinic to monitor dose delivery. For example, before the linear accelerator (linac) can be used for patient treatment, daily checks are carried out to ensure consistency of output. *In vivo* dosimetry may be used to verify the patient dose and detailed measurements are made during weekly and monthly quality assurance sessions checking all aspects of treatment delivery. Audits are used to check that procedures are being followed correctly and to ensure national consistency.

Over recent years the number of treatment modalities has increased significantly as well as the complexities of treatment. The oncologist today can choose from an array of techniques including low energy X-ray tubes (usually used for superficial tumors); low doserate or high doserate brachytherapy, where different radioactive species are inserted or implanted in the body; external beam therapy using a linac (producing either photon or electron beams); protons and heavy ions; and neutrons.

The treatment can be simple, such as a single square field from a ^{60}Co unit, or complex, such as Image Guided Radiotherapy (IGRT) using a modern linac, where the patient is imaged immediately prior to treatment, the tumor volume verified and the dose delivered with a large number of shaped fields.

To cover all possible radiotherapy techniques is beyond the scope of this article, and therefore we will focus on external beam therapy using photon and electron beams, as this is most common and where dosimetry is most advanced. The aim is to give a basic grounding in radiation dosimetry for oncology together with a review of dosimetry techniques and an up-to-date bibliography where the reader can obtain further detail.

RADIATION MEASUREMENT AND QUANTITIES

In radiation oncology, we are interested in the relationship between biological damage to cells and the radiation producing the damage. Various attempts have been made to define biological dosimeters [e.g., deoxyribonucleic acid (DNA) strand breaks, chromosome aberrations], but none have resulted in a quantity that is reproducible and can be transferred from one situation to another: a primary requirement for a measurement quantity. Therefore physical quantities are used as a basis for estimating biological effects.

Fluence

The particle fluence, Φ, is defined (1) as dN/da: the number of particles dN, incident on a sphere of cross-sectional area da. The use of a sphere expresses the fact one considers the area perpendicular to the direction of each particle. The energy fluence Ψ (defined as the energy incident on a sphere of unit area) is generally of more interest for photons as it is more closely related to the dose deposited (see below).

Interaction Coefficients

The stopping power S of a material is defined as the energy lost by the charged particle (electron or positron) dE, along

an increment of path dl. Ignoring energy losses due to nuclear reactions, stopping power has two principal components, namely, that due to collisions and that due to radiative losses. The collision component includes all energy losses in particle collisions that directly produce secondary electrons and atomic excitations. It also includes energy losses due to the production of Cerenkov radiation. The radiative component includes all energy losses of the primary electron that lead to bremsstrahlung production. The collisions of the primary electrons can produce high energy secondary electrons (δ-rays) that then become involved in independent interactions. The concept of restricted mass collisional stopping power L is therefore introduced to calculate the energy transferred to a localized region of interest. This region of interest is defined by a threshold (often denoted as Δ) for the energy transferred to the secondary (charged) delta particles. Highly energetic secondary particles with energy above this threshold escape the region of interest and do not contribute to the local absorbed dose and it is assumed that electrons with energy below Δ have negligible range. The restricted stopping power (L_Δ) is therefore always lower than the unrestricted stopping power and the choice of the energy threshold depends on the problem at hand. For problems involving ionization chambers, a frequently used threshold value is 10 keV since the range of a 10 keV electron in air is approximately 2 mm (the typical dimension of the air cavity of an ionization chamber). The parameter L_Δ is also known as the linear energy transfer (LET).

In practice, mass stopping powers (S/ρ, L/ρ) are generally used so that it is easier to compare the properties of materials with very different densities (e.g., air and water). A complementary quantity is the scattering power T, which describes the increase in the mean square scattering angle of the electron beam as it passes through a material.

For photon beams, there are a much larger number of possible interactions with the medium, the dominant ones in the energy range of interest being the photoelectric effect, Compton effect, pair production, coherent (Rayleigh) scattering, and nuclear photoeffect. The total interaction cross-section is simply the sum of all the individual cross-sections. The attenuation coefficient, μ, tends to be used rather than cross-sections as it describes the probability per unit thickness that a photon will undergo an interaction while traversing a material. As for stopping powers, the effect of density is removed and for dosimetric purposes two further coefficients are defined. The mass energy-transfer coefficient μ_{tr}/ρ relates the energy transferred from the photon to kinetic energy of charged particles and is used in the determination of kerma (see below). The mass energy absorption coefficient μ_{en}/ρ takes account of the fact that some of the energy transferred to charged particles is not deposited locally, but lost as bremsstrahlung.

Kerma

Kerma (kinetic energy released per unit mass), is introduced because neutral particles (photons and neutrons) deposit their energy in two steps: (*1*) interaction of the photon with an atom resulting in the transfer of energy to charged particles (predominantly electrons), and (*2*) deposition of that energy in the medium via Coulomb interactions (excitation and ionization). The dose contributed through direct interactions between photons or neutrons and the absorbing material will generally be negligible compared with this two-step process. Reference 1 gives the definition of kerma as:

$$K = \frac{dE_{tr}}{dm} \tag{1}$$

where dE_{tr} is the kinetic energy transferred from photons to electrons in a volume element of mass dm. Total kerma can be split into two parts: collisional and radiative kerma. Collisional kerma, K_{col}, leads to the production of electrons that dissipate their energy as ionization near electron tracks in the medium. Radiative kerma, K_{rad}, leads to the production of bremsstrahlung as the charged particles are decelerated in the medium.

For a monoenergetic photon spectrum, energy E, with fluence Φ, equation 1 becomes

$$K = \Phi E \frac{\mu_{tr}}{\rho} \tag{2}$$

where (μ_{tr}/ρ) is the mass energy transfer coefficient. For a polyenergetic photon beam, equation 2 becomes an integral over the full photon spectrum. As the photon energy increases, the maximum energy of the secondary electrons increases, the concept of a localized energy transfer begins to break down and kerma is therefore generally limited to photon energies below 3 MeV.

Absorbed Dose

The absorbed dose is defined as the mean energy imparted (absorbed) per unit mass. It is a nonstochastic quantity in that one is not measuring single events-the interaction between an incident photon or electron and a molecule—but the mean energy arising through the interaction of the radiation field with the material it passes through. As the mass of a sample decreases the energy per unit mass will become more random (stochastic). Whereas kerma is only defined for neutral particles, absorbed dose applies both to photon and electron beams.

Reference 2 applies this definition of absorbed dose in the situation where there is a small volume of the medium, which is thermally isolated from the remainder:

$$D_i = \frac{dE}{dm} = \frac{dE_h}{dm} + \frac{dE_s}{dm} \tag{3}$$

where D_i is the mean absorbed dose in the absorber of material i, and mass dm; dE is the mean energy imparted to the absorber by the radiation beam (photons or electrons); dE_h is the energy appearing as heat; and dE_s is the energy absorbed by chemical reactions (which may be positive or negative). The left-hand relation is independent of the measurement technique while the right-hand relation represents one of the most common methods for determining dose: the measurement of heat. The unit of absorbed dose is the gray (Gy); 1 Gy = 1 $J \cdot kg^{-1}$.

It can be inferred from the definitions above that collision kerma and absorbed dose should be related in someway, since they both deal with the deposition of energy in a localized area. If a state of charged particle equilibrium exists (and assuming no energy losses due to bremsstrahlung) then the absorbed dose will be equal to the kerma (conservation of energy). Charged particle equilibrium (CPE) exists at a point in the medium if the number and energy of charged particles entering a volume is equal to that leaving. True CPE only exists in the special case where there is no attenuation of the photon beam. In general there is always some photon attenuation, but there is said to be transient charged-particle equilibrium (TCPE), since the spectrum of charged particles changes very little as the photon beam penetrates the medium. Transient charged-particle equilibrium exists at the center of a broad photon beam at depths away from the surface (the depth at which TCPE is established depends on the incident photon energy and spectrum). For the general case, where TCPE exists and there are bremsstrahlung energy losses, the dose is given by

$$D = K_{\text{col}} = K(1 - g) \tag{4}$$

where g is the fraction of the energy that is lost to bremsstrahlung. For a ^{60}Co beam, g has a value of 0.003.

Absorbed dose is also related to the photon energy fluence at a point in a medium irradiated by a photon beam under conditions of transient charged particle equilibrium by

$$D = \Psi\left(\frac{\mu_{\text{en}}}{\rho}\right)\beta \tag{5}$$

where β is the ratio of absorbed dose to collision kerma at a point. As written, equation 5 is valid for a monoenergetic photon beam; for a realistic (broad) photon spectrum, the mass–energy absorption coefficient must be averaged over the photon fluence.

There is a charged-particle analog of equation 5. Under the restrictive conditions that (*1*) radiative photons escape the volume of interest and (*2*) secondary electrons are absorbed on the spot (or there is charged-particle equilibrium of secondary electrons), the absorbed dose to medium is given by the electron fluence multiplied by the collisional stopping power.

Dose Equivalent

This quantity is useful where the effect produced by the same absorbed dose is dependent on the particle type "delivering" the dose. This is the case in biological damage: the principal pathway for cell killing is a double-strand break of the cell's DNA, which is much more likely for densely ionizing particles, such as protons, neutrons, and α-particles, than it is for electrons or photons. A radiation quality factor, w is therefore introduced to take account of this and the dose equivalent is defined as the absorbed dose multiplied by this quality factor. Values of w vary from 1 for photons and electrons to 20 for α-particles.

PRIMARY METHODS OF DETERMINING ABSORBED DOSE AND AIR KERMA

Due to the complexities of the measurements, the absolute determination of radiation quantities is almost exclusively the domain of national standards laboratories. Two primary International System of Units (SI) quantities are realised by national standards laboratories for radiotherapy dosimetry: air kerma and absorbed dose. Air kerma can only be measured using an air-filled ionization chamber but absorbed dose can be determined in a variety of ways.

The absolute measurement of absorbed dose has a number of problems (some fundamental, others practical) that limit the accuracy of the result and put constraints on the experimental techniques that can be used.

1. *Doses of interest are small*. The definition of absorbed is in terms of the energy absorbed in an amount of material. Radiotherapy dose levels are typically < 10 Gy ($10\ \text{J}\cdot\text{kg}^{-1}$), which represents a very small energy deposition. If one is trying to determine this energy absolutely by measuring the radiation-induced temperature rise (of the order of a few mK) there is a significant challenge in achieving uncertainties $< 0.1\%$.
2. *The quantity required is the dose in an undisturbed phantom*. In radiotherapy, the required end-point is the dose to the tumor. However, since radiation interactions are very material dependent a homogeneous phantom is the chosen medium for reference dosimetry. This immediately presents a problem in that any measuring instrument will perturb the phantom and affect the measurement one wishes to make.
3. *The quantity required is the dose at a point in this phantom*. For radiotherapy dosimetry, one is not interested in the average dose to the whole phantom (although mean dose or integral dose is required for radiation protection, when considering lifetime dose to organs, etc.). Radiotherapy treatments using photon and electron beams produce significant dose variations within a phantom; otherwise, healthy tissue could not be spared. It is therefore important to be able to measure these dose variations, which by implication requires a small detector. Such a detector will generally have a larger uncertainty than a larger detector. Care is required in designing a detector that samples the dose at a point and does not give some unwanted averaging.
4. *Scattered radiation contributes a significant proportion of the absorbed dose*. In a typical radiotherapy radiation field used for cancer treatment (e.g., a 6 MV photon beam), 15% of the dose at the point of interest is due to scattered, rather than primary, radiation. The experimental geometry is therefore very important and care must be taken in designing experiments, especially when comparing or calibrating dosimeters, so that scattered radiation is properly taken into account.

5. *Optimization of the measurement is difficult*. One of the biggest practical constraints is that in the measurement of absorbed dose one is not determining some fundamental constant or characteristic of a material. The dose is the effect of a particular radiation field at a point in a particular material and it is therefore not possible to optimise all aspects of a measurement. There are many "influence quantities" (material, energy spectrum, geometry) so that what may appear to be minor variations from the real measurement problem (dose to a tumor) can result in significant errors being introduced.

These issues also apply to the determination of air kerma.

Ionometry

An ionization chamber measures the ionization produced by the incident radiation beam in a mass of air. Historically, the first quantity to be measured was Exposure (symbol X) and is simply the charge, Q, liberated in a volume of air mass m_{air}. It is not a recommended SI unit but is related to Air Kerma by

$$K_{\text{air}} = X\frac{W}{e}\frac{1}{1-g} \tag{6}$$

where W/e is the average energy required to liberate an ion pair and g has the same definition as in equation 4. The value of W/e has been measured by a large number of experimenters of many years and there is an agreed value of 33.97 eV/ion pair (3), which is constant over widely varying conditions (air pressure, electron energy, etc.).

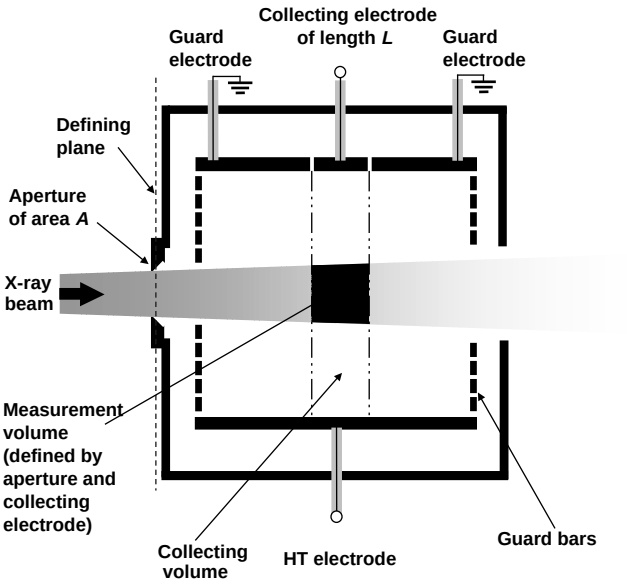
For low energy beams produced by X-ray tubes (< 400 kVp) exposure is measured using a free-air chamber. A typical chamber is shown in Fig. 1. By careful design and precise manufacturing of the electrodes and entrance aperture it is possible to accurately define the volume (and thus mass) of air that is irradiated. The size of the free-air chamber scales with the incident energy and for a ^{60}Co beam a free-air chamber would be impractically large: several meters in each dimension.

At energies of ^{137}Cs (663 keV) and above, one must therefore use a cavity chamber, where the volume of the air cavity is small (typically a few cm^3). One must determine the mass of air in the cavity and various designs been employed: spherical, cylindrical, and "pancake" (Fig. 2).

Until 1990, all absorbed dose measurements in the radiotherapy clinic were based on air-kerma calibrations using protocols as seen in Refs. 4 and 5. In recent years, absorbed dose-based calibrations have become available from national standards laboratories and associated protocols produced (e.g., Refs. 6–8). However, air kerma standards are still required for the dosimetry of kilovolt X-ray beams, brachytherapy, and radiation protection.

The dose to the air volume of a cavity chamber irradiated by an X- or γ-ray beam is given by

$$D_{\text{air}} = \frac{Q}{m_{\text{air}}}\frac{W}{e} \tag{7}$$



Schematic diagram of the National Physics Laboratory (NPL) primary standard free-air ionisation chamber

Figure 1. Diagram of a free-air chamber. (Courtesy NPL.)

Bragg–Gray cavity theory is then used to relate the dose in the air cavity to the dose to the medium. The conditions for application of Bragg–Gray cavity theory are

1. The cavity must be small when compared with the range of charged particles incident on it so that its presence does not perturb the fluence of charged particles in the medium.
2. The absorbed dose in the cavity is deposited solely by charged particles crossing it (i.e., photon interactions in the cavity are assumed negligible and thus ignored).

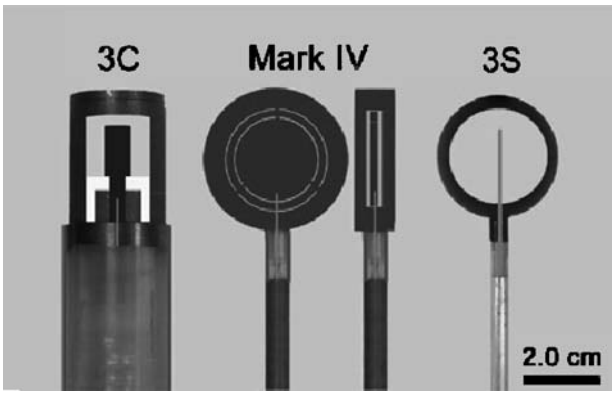


Figure 2. Schematics of three graphite-walled cavity ion chambers designed and operated at the National Research Council (NRC) - cylindrical (3C), parallel-plate (Mark IV), and spherical (3S). The 3S utilizes an aluminum electrode, while the other two chambers utilize graphite electrode.

Condition (2) implies that all electrons depositing the dose inside the cavity are produced outside the cavity and completely cross the cavity. Therefore, no secondary electrons are produced inside the cavity and no electrons stop within the cavity. The dose to the medium is obtained using a ratio of stopping powers:

$$D_{\text{med}} = \frac{Q}{m_{\text{air}}} \frac{W}{e} \left(\frac{S}{\rho}\right)_{\text{med,air}} \tag{8}$$

where $(S/\rho)_{\text{med,air}}$ is the mass stopping power ratio for the medium divided by that for air. The Bragg–Gray cavity theory does not take into account the creation of secondary (delta) electrons generated as a result of the slowing down of the primary electrons in the sensitive volume of the dosimeter. The Spencer–Attix cavity theory is a more general formulation that accounts for the creation of these electrons that have sufficient energy to produce further ionization on their own account.

Equation 8, in principle, gives a possible route to the absolute absorbed dose, if the stopping power ratio is known. The Bureau International des Poids et Mesures (BIPM) maintains such an ionometric standard for absorbed dose to graphite. This is a graphite walled ionization chamber, whose volume has been determined by mechanical means, and is described in detail by Boutillon and Peroche (9). Strictly speaking, however, this is not a primary device, as the value for the product of W/e and the stopping power ratio is taken from calorimeter measurements. Although there are independent measurements of W/e, there is a lack of measured stopping power data (see below) to provide a true measurement of absorbed dose absolutely.

As noted above, one of the difficulties in realising a primary standard cavity ion chamber is defining the effective volume of the chamber. This problem can be overcome to a certain extent by the use of an extrapolation, or gradient, chamber. In such a chamber, the absolute volume of the cavity is not known, but can be changed by a known amount, usually by changing the electrode spacing. Assuming that the chamber does not perturb the medium then the dose is given by

$$D_{\text{med}} = \frac{\Delta Q}{\Delta x} \frac{W}{e} \left(\frac{S}{\rho}\right)_{\text{med,air}} \quad \frac{1}{A\,\rho_{\text{air}}} \tag{9}$$

where ΔQ is the change in the measured ionization charge for a change in the electrode spacing of Δx, A is the area of the electrode and ρ_{air} is the air density. The dose measurement becomes a relatively simple charge measurement and the problem of the determination of volume is reduced to that of determining the area of the collecting electrode. Klevenhagen (10) carried out some of the first work on gradient chambers for the determination of dose in megavoltage photon and electron beams and Zankowski and Podgorsak (11) describe a chamber where the entire device is manufactured from a plastic with similar radiation properties to water (see Fig. 3). It is a parallel plate chamber with a fixed radius and the plate separation is varied by way of a precision micrometer. The charge gradient $\Delta Q/\Delta x$ can be determined at the 0.2% level and the effective

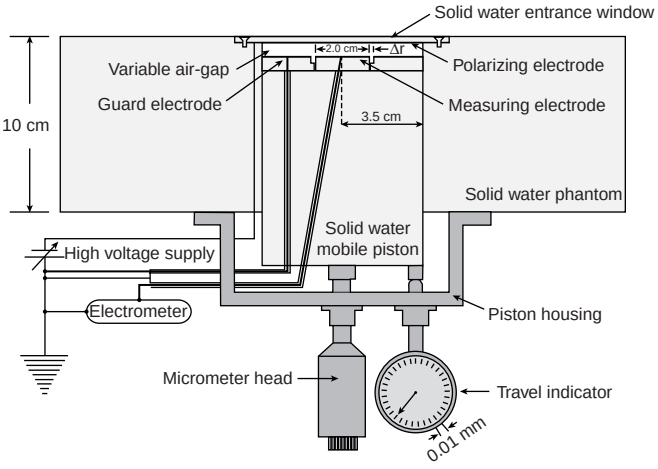


Figure 3. Extrapolation chamber design (11).

area of the collecting electrode (determined by a capacitance measurement) has an uncertainty of 0.1%. The limiting factor for this type of device is the uncertainty in W/e and the stopping power ratio at energies other than ^{60}Co.

An extrapolation chamber is also the device of choice for β-ray brachytherapy sources (such sources are used for the treatment of ophthalmic cancers) and one can derive absorbed dose using the Bragg–Gray principle (12).

Calorimetry

A calorimeter directly measures the absorbed dose as it is defined above. Although care must be taken in the design of a calorimeter, in terms of geometry and material composition, no primary conversion factor (e.g., W/e, G-value) is required. However, the basic operating principle of calorimetry is that all the energy deposited by the radiation is expressed as heat. If this is not the case, then there is said to be a heat defect. The heat defect can be due to crystal dislocations, radiation-driven chemical reactions or some other mechanism and is strongly material dependent. It follows from equation 3 that if there is no heat defect then

$$D_{\text{i}} = c_{\text{p}} \Delta T \tag{10}$$

where c_{p} is the specific heat at constant pressure and ΔT is the temperature rise in the material (absorber). The size of the > element > that defines the measured dose will depend on the specific application as well as the design of the calorimeter and the material used. A material with a high thermal conductivity will require mechanically defined components (e.g., some small absorber thermally isolated from the rest of the material), while it is much easier to measure a point dose in a material with a low thermal conductivity. Inhomogeneities in the phantom will affect the scattering of the radiation beam, and therefore change the absorbed dose measured. A correction will be required to give the dose in a homogeneous medium.

The three challenges in calorimetry are therefore to (*1*) measure the radiation induced temperature; (*2*) measure a material of known specific heat capacity, and (*3*) make sure that what is measured is relevant to the particular application of the radiation beam. These problems have been addressed in a number of (often novel) ways over many years, but currently there are basically two types of calorimeter: graphite (e.g., see Refs. 13 and 14) and water (e.g., see Ref. 15). Graphite has some obvious advantages: it is solid and a graphite calorimeter can be made smaller and more robust than a water device. For example, McEwen and Duane (16) demonstrate a calorimeter designed to be taken routinely into radiotherapy clinics. There is no heat defect or convection to consider for graphite and the temperature rise pure unit absorbed dose is much larger than that of water due to the low specific heat capacity ($c_{\mathrm{p,graphite}} \sim 700$ $c_{\mathrm{p,water}} \sim 4200$ $\mathrm{J{\cdot}kg^{-1}{\cdot}{}^{\circ}C^{-1}}$). However, the high thermal conductivity is an issue and the quantity realized is absorbed dose to graphite so a conversion is required to obtain absorbed dose to water (see below). Since water is the standard reference material for radiation dosimetry the majority of standards laboratories are moving over to water calorimeters as the primary standard and the device operated at the National Research Council in Canada is shown in Fig. 4. The major problems in developing a water calorimeter are controlling convection and obtaining a stable (ideally zero) heat defect for the sample of water irradiated. The present state-of-the-art in calorimetry yields an uncertainty in absorbed dose to water of $\sim 0.3\%$ and for recent reviews of calorimetric development see Ross and Klassen (17), Williams and Rosser (19), and Seuntjens and DuSautoy (20).

Chemical Dosimetry

In chemical dosimetry, the absorbed dose is determined from some chemical change in an appropriate material and any well-characterized chemical reaction where the reaction product(s) can be measured with good precision may serve as the basis for the dosimeter. Chemical dosimeter systems were developed as early as 1927 and a wide range of systems have been studied. Although, in principle a chemical dosimeter is a secondary device, in that it does not directly measure the absorbed energy, it can be regarded as a primary device if the relation between absorbed energy and chemical change can be determined absolutely. This relationship is termed the radiation chemical yield and is expressed as the number of molecules or ions of product X liberated per unit absorbed energy, designated $G(\mathrm{X})$. Assuming that the G-value is known and is constant with dose then the absorbed dose is given by

$$D = \frac{\Delta M}{G(\mathrm{X})\rho} \tag{11}$$

where ΔM is the volume concentration of molecules produced by the radiation absorbed and ρ is the density of the medium.

To act as a primary dosimeter, a chemical dosimeter should be dose, dose-rate, and LET independent. Aqueous systems are preferred as they are basically water equivalent, although this introduces a containment vessel whose effect must be taken into account.

The ferrous sulfate (21) dosimeter is the most widely used and longest established dosimetry system. It demonstrates the advantages and problems of chemical dosimeters. The reaction mechanism is the oxidation of ferrous to ferric ions, in aerated sulfuric acid. The oxidation proceeds via a number of reactions involving hydroxyl radicals, hydroperoxy (HO_2) radicals, and hydrogen peroxide. The ferric ion formation is directly proportional to energy absorbed as long as some oxygen remains in the solution, hence the requirement for aeration. All the reactions are fast (< 1 min), therefore there is no aftereffect under usual γ- or electron irradiations. However, great care must be taken in the preparation of the solutions, particularly with regard to water purity as organic impurities can have a significant effect. Spontaneous oxidation of the ferrous ions occurs that can be corrected for by the use of an unirradiated sample as a control.

The concentration of ferric ions may be determined by titration, but absorption spectroscopy is generally a more convenient technique, using ultraviolet (UV) wavelengths of 304 or 224 nm. The Fricke dosimeter is dose-rate independent for ^{60}Co radiation in the range 0.1–40 $\mathrm{Gy{\cdot}s^{-1}}$ (a range that covers both radiotherapy and industrial dosimetry applications) and for linac irradiations, $G(\mathrm{Fe^{3+}})$ production is linear up to a maximum dose-per-pulse of 2 Gy (significantly greater than radiotherapy linacs). The normal dose range is 5–350 Gy, although this can be extended by suitable modifications of the composition of the system, or of its analysis. With care, Fricke dosimetry is capable of 0.1% precision, but for absolute dosimetry one requires an accurate determination of the G-value. As with ionometry, this factor can be determined from calorimetry, but preferably one would like an independent measurement. Such a measurement is possible if one

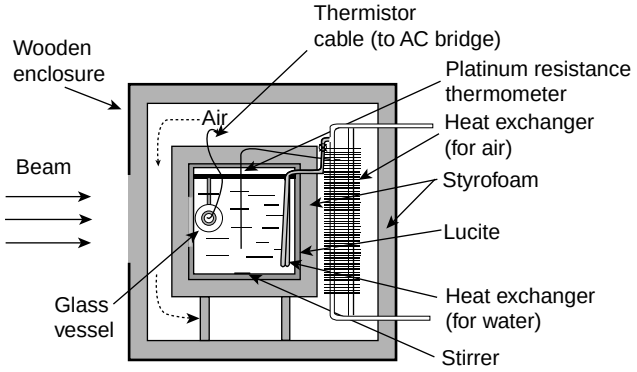


Figure 4. Overview of the NRC sealed water calorimeter. The outer box, which provides thermal insulation, is ~ 80 cm on a side, while the water phantom is ~ 30 cm on a side. The inner-glass vessel provides a stable volume, where the purity of the water can be rigorously maintained (to control the heat defect). The radiation-induced temperature rise (typically a few mK) is measured using thermistor probes and the outer sys- tem controls the temperature at 4°C to minimize convection effects.

knows the total energy in the radiation beam. Roos and Hohlfeld (22) describe the system developed at the PTB (Physikalisch Technische Bundesanstalt) in Germany based on a microtron accelerator with a very well determined electron energy and beam current. With such a system an uncertainty in the *G*-value (the effective response of the Fricke is actually derived in this measurement) of $< 0.4\%$ is achievable and the overall uncertainty in measuring absorbed dose to water is $\sim 0.5\%$. Other chemical dosimeter systems include ceric sulfate, oxalic acid, potassium dichromate, and alanine, which cover higher dose ranges than Fricke. However, *G*-values for these systems are either unknown, or have a much larger uncertainty, and therefore cannot be regarded as primary dosemeters.

Conversion of Dose Between Materials

Since dose is material dependent, the primary device one uses to measure absorbed dose may not yield the quantity required. A conversion procedure is therefore needed. If the uncertainty on this conversion is sufficiently large then the usefulness of the primary device is called into question. The majority of effort in this area has concentrated on water and graphite since graphite is commonly used for primary standard calorimeters and water is the material of interest for radiotherapy dosimetry.

For electron beams, the conversion factor is a product of two factors: a ratio of stopping powers and a fluence correction (the latter takes account the differences in scattering power between the two materials). The most accurate values for stopping powers are those given in Ref. 23, but these are based on calculation alone and a quoted uncertainty of $\sim 1\%$ for each material is given. There have been a number of attempts to measure stopping powers (e.g., Ref. 24), but these did not have the accuracy to validate the calculations. One of the problems in measuring stopping powers is that it is only possible to measure relatively small energy losses and this significantly increases the precision required if the achieved overall uncertainty is to be $< 1\%$. Faddegon et al. (25) presented a new technique using a large sodium iodide detector to directly measure elemental stopping powers and McPherson (26) reports the results of such measurements. The standard uncertainty on these measurements (0.4–0.7%) is significantly lower than the previous attempts and is at a level where the calculated values in Ref. 23 can be tested. Fluence corrections are determined either through direct measurement in phantoms of different materials or using Monte Carlo simulations (see below).

For megavoltage photon beams, there is more than one method available for converting dose from one material to another. Burns and Dale (27) describe two methods: the first making use of the photon fluence scaling theorem (28) and the second based on cavity ionization theory. Nutbrown et al. (29) repeated the experimental work of Burns and Dale and applied a third method based on extensive Monte Carlo simulations. Fricke dosimeters can also be used to transfer the dose between materials as it can be assumed that the *G*-value is independent of the phantom material (27).

Monte Carlo: A Primary Technique for the Future?

There has been a rapid development of Monte Carlo techniques for radiation dosimetry in the last 10–15 years. A Monte Carlo calculation is based on radiation transport physics and tracks individual particles as they interact with the detector and phantom. By averaging over a large number of particles (typically > 10 million), statistical fluctuations can be reduced to an acceptable level. The big advantages of a Monte Carlo simulation are (*1*) there is no reliance on a physical artifact, such as a ion chamber or calorimeter, and (*2*) you are not constrained by many of the problems of physical measurement as outlined above and can derive the exact quantity you require. Calculations initially focused on determining correction factors, such as the ion chamber wall effect for air kerma standards and the effect of inhomogeneities in a medium. More recent Monte Carlo codes have included the accurate simulation of the radiation source (e.g., BEAM (30)) and the detector (e.g., EGSnrc (31)). The sophistication has reached the level where they may be considered as viable alternatives to measurements.

In considering the idea of Monte Carlo as a primary technique one can clearly not escape some absolute measurement for the primary realization of absorbed dose. For example, the absolute beam current produced by a linear accelerator or the total activity of a radioactive source would be required as an input to the simulation, but the dose itself would be calculated. If this measurement can be determined with high accuracy and the absolute uncertainties in the Monte Carlo can be reduced, then this offers a potential alternative to the present primary standards. The major limitation is the accuracy of input data for the physics models: interaction cross-sections, stopping powers, and so on are not known accurately enough. The high accuracy obtained in the determination of correction factors in dosimetry is because in those situations one does not rely in such a direct way on absolute interaction coefficients, but on differences (or ratios) in interaction coefficients, where one benefits from the cancellation of correlated uncertainties. To date, the majority of the effort has been in developing the Monte Carlo codes (improving efficiency and refining the physics modeled), but there are still significant gaps in the input data so it is not clear whether the potential for the absolute application of Monte Carlo techniques can be fulfilled.

REFERENCE OR SECONDARY DOSIMETERS

As with primary devices, there are number of different types of secondary dosimeter that are used in radiotherapy. Secondary dosimeters require calibration against a primary standard and are then used to realise absorbed dose on a more routine basis.

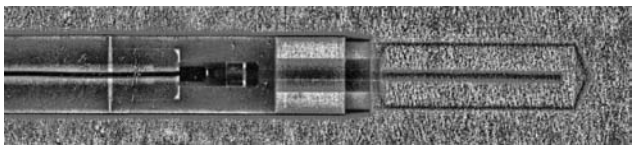


Figure 5. Radiograph of a NE2571 Farmer-type chamber. One can see the graphite outer wall of the cavity, aluminum central electrode, and internal construction of the stem.

Ion Chambers

Ionization chambers (particularly air-filled chambers) are the most widely used instruments in radiotherapy dosimetry. They offer a number of characteristics which are particularly suited to the measurement of therapy radiation beams - sensitivity, long-term stability, and ease-of-use. The most widely used chamber is the Farmer-type (33), an example of which is shown in Fig. 5. Chambers of this type show excellent stability (figures of $\pm$ 0.3% over 25 years are not uncommon) with the only disadvantage being a lack of waterproofing. A waterproof sleeve is therefore required for measurements in water. Such a sleeve should be thin ($<$ 1 mm), close fitting (no air gaps), and made from some low Z material to minimize any additional perturbation effect [PMMA-poly(methyl-methacrylate) is commonly used].

Parallel-plate chambers, which are usually waterproof, are recommended for the dosimetry of electron beams. The NACP design (34) is one of the most widely used (Fig. 6).

Ionization chambers are usually vented to the atmosphere and therefore an air density correction, f_{TP}, is required to normalise for variations in air temperature and pressure (T_{air} and P_{air}, respectively):

$$f_{TP} = \frac{273.15 + T_{air}}{273.15 + T_{ref}} \cdot \frac{P_{ref}}{P_{air}} \qquad (12)$$

P_{ref} is taken to be 101.325 kPa, but there is no agreed value for T_{ref}. In Europe a value of 20 °C is used, while in North America the reference temperature is 22 °C. Care must therefore be taken when comparing results from different laboratories.

A correction for the humidity of the air in the chamber is not generally applied. The presence of water vapor affects the value of W/e (36), as well as stopping power and mass-energy absorption coefficients. However, for relative humidities between 10 and 90% (when comparing to a standard humidity of 50%) the effect is a maximum of 0.1%.

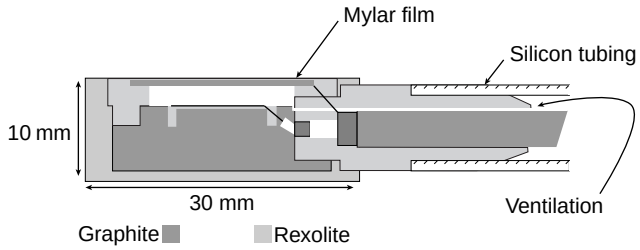


Figure 6. Schematic of the NACP parallel-plate chamber (taken from Ref. 35). Design features include waterproof construction, low Z materials to minimize the perturbation correction, and large guard ring to minimize in-scatter.

A correction is required to account for the incomplete collection of charge due to ion recombination in the chamber volume. The correction for ion recombination is the sum of two components: initial recombination and general or volume recombination. General recombination takes place when oppositely charged ions from different ionization tracks (i.e., created by different incident ionizing particles) recombine while they drift under the influence of the electric field toward their respective electrodes. Initial recombination takes place when oppositely charged ions from the same ionization track recombine; as the name suggests, this process takes place before the electric field is able to pull the track structure apart and therefore precedes general recombination. Initial recombination is independent of dose rate but general recombination depends on the ion density in the cavity. This ion density depends on the dose rate for continuous radiation and on the dose per pulse for pulsed beams. Initial recombination is typically small ($\sim$ 0.1–0.2% for the usual cylindrical and parallel-plate chambers employed in radiotherapy). General recombination is typically a small effect for continuous radiation (e.g., kilovoltage X-ray beams or ^{60}Co γ-ray beams) but for pulsed beams it can often be significant, especially so for modern linear accelerators that employ large dose-per-pulse values (recombination corrections of up to 5% have been reported).

The theoretical aspects of ion recombination for pulsed and continuous radiation have been well discussed in the literature (37–39). However, in recent years a number of authors (40–42) have presented recombination data that do not agree with the standard theory. A number of possible mechanisms have been proposed, including ion multiplication, air volume change, and direct collection of primary electrons, but at present there is no consensus as to which, if any, of these is the reason for these anomalous results. Dosimetry protocols recommend that a full $1/I$ against $1/V$ plot be measured where I is the measured Ionization current and V is the polarizing voltage to establish the range of linearity where the standard theory holds and the chamber then operated at voltages to remain within that range (Fig. 7).

Ion chamber measurements must also be corrected for the effects of polarity. The polarity effect is the difference in readings obtained in the same irradiation conditions, but taken with positive and negative polarizing voltages. Boag (43) identified a number of components of the polarity effect including secondary electron emission (due to the Compton effect) that produces a negative current independent of polarity; uneven distribution of the space charge due to a difference in the drift velocity of negative and positive ions; variation of the active volume due to space charge distortions; stopping of fast electrons in the collecting electrode not balanced by the ejection of recoil electrons; and collection of current outside the chamber volume due to leakage in solid insulators. In practice, it is difficult to identify the mechanisms acting in a particular

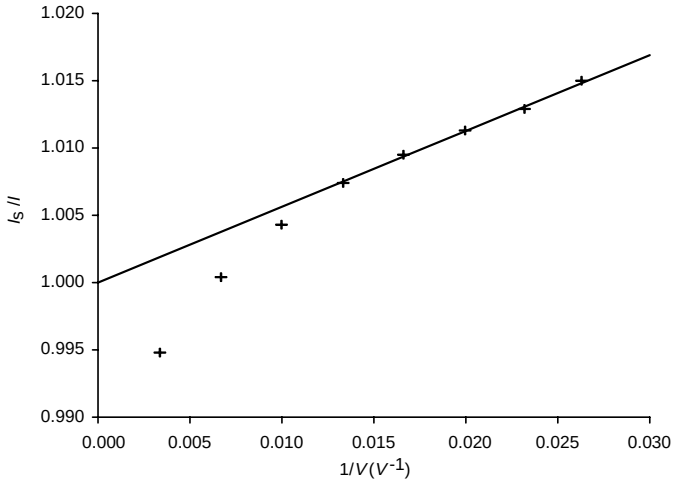


Figure 7. Plot from Burns and McEwen (41) showing the deviation from theory (straight line) of the recombination behavior for a NACP chamber. Without extensive measurements errors of up to 1% are possible.

situation, but measurements show that the polarity effect will vary with chamber type, beam energy and modality, measurement depth and can vary with other factors, such as field size.

The polarity correction is given by

$$f_{\text{pol}} = \frac{|M^+| + |M^-|}{2|M|} \tag{13}$$

where the superscripts + or − indicate the reading (M) with collecting voltage positive and negative, respectively, and M in the denominator is the reading taken with the normal polarity used during measurements. Table 1 summarizes typical polarity corrections for chambers in different beams.

A variation on the air-filled ionization chamber is the liquid ion chamber. In this design, the air is replaced by a liquid, which offers two major advantages: a flat energy response and an increased carrier concentration (and therefore increased spatial resolution). Liquid ion chambers have been developed over many years, but their use as secondary dosimeters has been severely hampered by the volatility of the liquid (usually a short-chain hydrocarbon) resulting in loss of signal. However, recent results (47) show impressive stability and may indicate that liquid ion chambers have a role to play in reference dosimetry.

Table 1. Typical Polarity Corrections

Beam	Cylindrical Chambers	Parallel-Plate Chambers
Megavoltage photons	$< 0.2\%$ beyond $d_{\max}$, more variable in build-up region	Generally $< 0.3\%$, but can show variable behavior.
Megavoltage electrons	Up to 1% at lower end of recommended energy range (44)	$< 0.2\%$ for well-designed chambers (45). Can be significant for other chamber types (46)

Fricke

The Fricke dosimeter was described in detail in the section above on primary dosimeters. As a secondary dosimeter it is used in exactly the same way, except that the G-value is effectively measured for each batch of solution by comparison with a calorimeter (48). The big disadvantages of Fricke are (*1*) the care needed to produce 'good' solutions, and (*2*) the perturbation correction required for the vessel holding the Fricke solution (usually glass or quartz) is generally large. The NRC in Canada has used Fricke to transfer the dose from water calorimeter to ionization chambers (49).

TLD

Another class of systems is thermoluminescent dosimeters (TLDs). One of the obvious advantages of such dosimeters is that they can be made very small, and are therefore ideal for plotting dose distributions. The TLD material can be used as a powder or can be formed in various shapes (chips, rods, pellets, etc.). These materials have a wide dose range, from a few tens of μGy to $\sim$ 1 kGy. The readout (measurement of the glow curve) is destructive, but the dosimeters can be reused. The equipment required is readily available and the production and readout of dosimeters is relatively simple, particularly compared to Fricke or alanine (Fig. 8).

Lithium fluoride is the most widely used system for radiotherapy applications as it has a mean atomic number close to that of tissue ($Z_{\text{eff}} = 8.2$, compared to 7.4 for tissue). It has a fairly flat response with energy (especially in the megavoltage region) and is therefore not particularly sensitive to variations in beam quality. Both CaF_2 and $CaSO_4$ are useful in that they have sensitivities 10–100 times greater than LiF but, because of their high Z values, they show a very rapid change in energy response at low energies. Lithium borate has a better tissue similarity ($Z_{\text{eff}} = 7.4$) but has a sensitivity of only one tenth of that of LiF. As for the other systems based on some chemical change, TLD materials require calibration against a primary dosimeter. It is not possible to determine any thermoluminescent equivalent of a G-value as the dose response depends on the annealing process and tends to be batch dependent. Typical reproducibility at the 1% level is possible routinely with an overall uncertainty of 2–3% (one standard deviation). However, Marre et al. (50) obtained a reproducibility of better than $\pm$ 0.5% and an overall standard uncertainty in measuring absorbed dose to water of $\pm$ 1.6%. These values are approaching those of ion chambers although the delay between irradiation and readout and the care required to achieve this level of precision limit the applications for this dosimeter.

TLD is an attractive dosimeter for the dosimetry of low doserate brachytherapy sources. The source strength is normally too low for small ionization chambers, and large volume chambers have poor spatial resolution. However, for ^{125}I, one of the commonly used isotopes in prostate

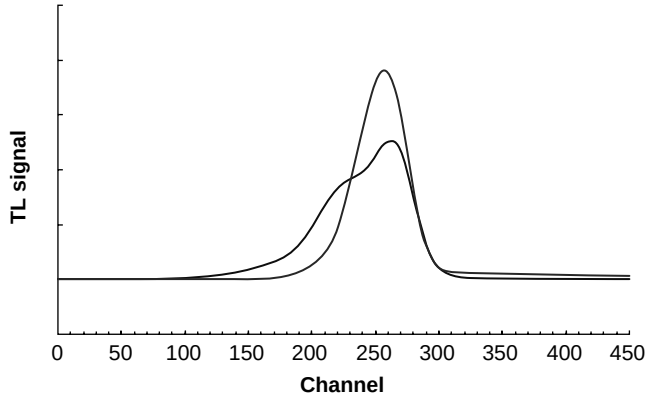


Figure 8. Glow curves from two different TLD materials. The temperature is slowly ramped to a maximum (in this case 240 °C) and the thermoluminescent intensity measured using a photomultiplier. The shape of the glow curve depends both on the material and the thermal pre-treatment (annealing).

treatment, the mean photon energy in only 27 keV and therefore the energy dependence of TLD needs to be known accurately (Fig. 9).

Since LiF is nontoxic, TLD can be used as an *In vivo* dosimeter, placed directly on the patient, to verify treatment delivery. It is a less invasive technique compared to diodes or MOSFET detectors (see below), as there are no trailing wires or associated equipment.

Alanine

Over recent years, alanine has become more widely accepted as a chemical dosimeter for radiotherapy dosimetry. It has a very wide dose range, showing a linear response from 10 Gy to 70 kGy. It is a solid dosimeter, with a density and atomic number close to that of water (close to zero perturbation) and the dosimeters are small: typically disks are 5 mm in diameter and 3 mm thick, but can be made as thin as 0.5 mm for measuring low electron energies. The energy dependence is very small. Zeng et al. (52) showed that any variation in the sensitivity of alanine

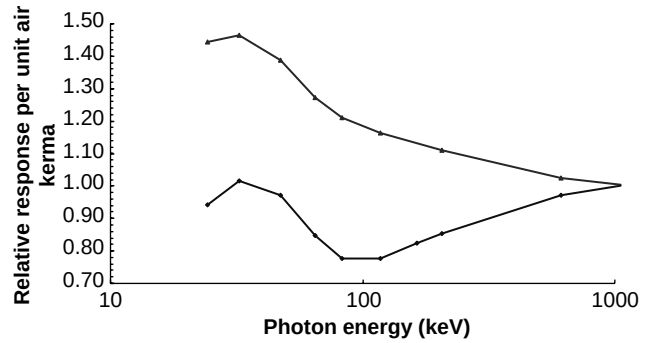


Figure 9. Energy dependence of two types of LiF TLD dosimeters (from Ref. 5). Triangles - TLD-100, diamonds - TLD-100H. LiF:Mg,Ti (TLD-100) has been a widely used dosimeter since the 1960s, LiF:Mg,Cu,P (TLD-100H) was developed in 1976, with 20–30 times greater sensitivity.

is not more than $\pm$ 0.5% over the energy range from ^{60}Co to 25 MV X rays. The dosimeter is read out nondestructively using ESR (electron spin resonance) spectroscopy. This nondestructive read-out, together with the long-term stability of the radiation-induced signal means that alanine has a potential role as a dose record. The National Physical Laboratory in the United Kingdom offers a mailed dosimetry service for radiotherapy using alanine dosimeters.

Summary

For reference dosimetry in radiotherapy clinics the system of choice is the ion chamber. Ion chambers are simple to use, offer high precision and accuracy and give an immediate reading. Integrating dosimeters (Fricke, alanine, and TLD) tend to be used as QA checks, either internally or within a wider framework of national or international comparisons (e.g., TLD is used for both the IAEA international mailed reference dosimetry service (53) and the RPC audit scheme in North America (54)). Generally, ion chambers are calibrated against primary standards and then used to calibrate other dosimetry systems within the clinic.

CALIBRATION OF SECONDARY DOSIMETERS

Basic Formalism

An ideal secondary dosimeter will have a zero energy dependence. Calibration against a primary standard (calorimeter) would then only need to be carried out at one beam quality (e.g., ^{60}Co). Energy independence also implies that the calibration coefficient is the same in photon and electron beams since the dose in a photon beam is dependent on the secondary electron spectrum generated in-phantom. In practice, the majority of secondary dosimeters commonly in use have some energy dependence. Ionization chambers, for example, show a variation of $> 3\%$ over the energy range from ^{60}Co to 25 MV photons, with even larger variations at low X-ray energies.

The obvious method to calibrate an ion chamber in terms of absorbed dose is to compare a chamber with a primary device. However, although accelerators were being used from the 1950s for radiotherapy, there were no absorbed dose standards for megavoltage photon or electron beams until the 1970s. Absorbed dose measurements using ion chambers were therefore based on air-kerma calibrations derived at lower photon energies (either ^{60}Co or 2 MV X rays). Protocols were developed to enable users to obtain a measurement of the absorbed dose delivered by a linac in the clinic. Only in recent years have absorbed dose-based calibrations become available from national standards laboratories and associated protocols produced (e.g., Refs. 6–8). For the purpose of the following discussion, we will only deal with absorbed dose calibrations in megavoltage photon and electron beams, but the principles are basically the same for other situations (kV X rays, protons, etc.).

The basic formalism for the calibration of an ion chamber is simple. The chamber is compared against the

primary device and a calibration coefficient ($N_{D,sec}$) for that beam is derived

$$N_{D,sec} = \frac{D_{std}}{M_{sec}} \tag{14}$$

The parameter D_{std} is the dose measured by the primary device and M_{sec} the chamber reading corrected for influence quantities. This calibration coefficient will be a function of the energy of the photon or electron beam and is given in terms of a beam quality specifier, Q_{ref}. The user then derives the calibration coefficient for the user beam quality, $N_{D,ref}$ (Q_{user}). Some primary laboratories only supply calibration coefficients for ^{60}Co and thus correction factors are required, which are given in dosimetry protocols (e.g., Ref. 7). An alternative approach, as used in the United Kingdom's Code of Practice (6) is to obtain absorbed dose calibration coefficients in linac photon beams. In this case there is no need for the calculated conversion factors and an ion chamber is calibrated in a beam similar to what it will be used to measure.

A measurement is then made in the user's radiation beam to measure the absorbed dose, D_{user}:

$$D_{user} = N_{D,sec}(Q_{user})M_{user} \tag{15}$$

where M_{user} is the chamber reading.

Implied in equations 14 and 15 is the reference depth at which the measurement is carried out. The concept of the reference depth for a calibration is much more important for electrons than photons. In a phantom irradiated by a megavoltage photon beam the secondary electron spectrum (which determines the dose) varies only slowly with depth (for depths greater than the range of incident primary electrons). By contrast, in the situation of a primary electron beam, the electron spectrum seen by the detector constantly changes from the surface to the practical range. The choice of reference depth should be both clinically relevant and reliable in terms of transferring the dose from the primary laboratory to the user's beam. For photon beams there may be only one or two reference depths defined for all energies while for electron beam dosimetry all modern protocols define the reference depth as a function of energy.

Potential Problems with Beam Quality Specifiers

A typical radiotherapy linac accelerates electrons to energies in the range 4–22 MeV and can also produce bremsstrahlung X-ray beams over a similar energy range. In both cases, the detector calibration coefficient will be some function of this spectrum. Since it is not generally possible to measure the energy spectrum directly a beam quality specifier (Q) is used. This is obtained by measuring some property of the radiation beam (e.g., the penetration through a material). A "good" beam quality specifier is one such that a value of Q relates uniquely to the effect of a particular spectrum. A problem arises if Q is not a good beam quality specifier, that is, there is some ambiguity in the relation between Q and the effect of the incident spectrum.

Beam Quality Specifiers for Photon Beams. Typical photon depth–dose curves from a clinical linac are shown in Fig. 10 (it should be noted that MeV tends to be used as a label for electron beams and MV for photon beams). Over the years, a number of beam quality specifiers have been proposed for megavoltage photon beams, but all relate in some way to the penetration of the photons through some material.

The most widely used parameter has been $TPR_{20,10}$ (tissue phantom ratio), which is defined as the ratio of ionization currents at measurement depths of 20 and 10 cm in water with a fixed-field size and source to chamber distance. The 10 and 20 cm points in Fig. 10 are on the downward portion of the curves, and therefore TPR is related to a measurement of the attenuation of the beam.

There has been much debate in recent years over the sufficiency of $TPR_{20,10}$ as a beam quality specifier for

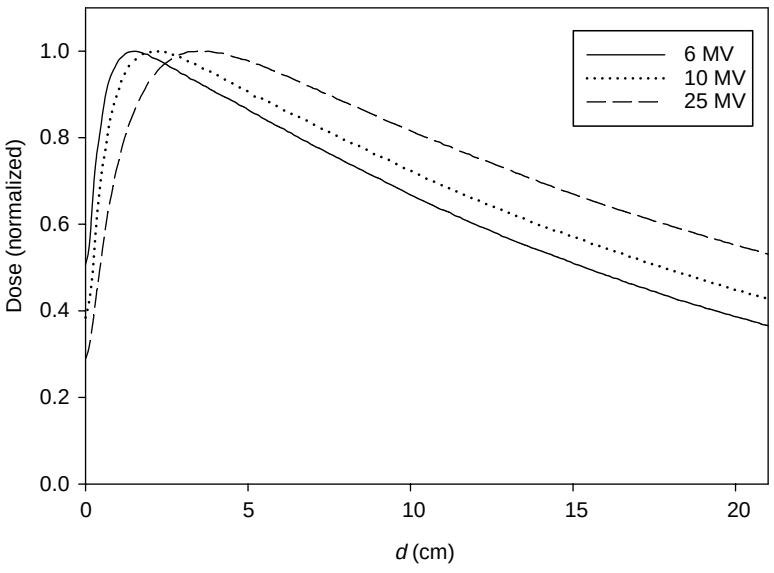


Figure 10. Depth dose curve for three photon beams.

the purpose of ion chamber calibration in terms of absorbed dose to water. Rosser et al. (55) found that an error of up to 0.6% could be introduced by the incorrect application of calibration coefficients using TPR. A number of other beam quality specifiers have been put forward as alternatives to TPR including: d_{80} (the depth at which the dose is 80% of the peak dose); the HVL of water and the percentage depth dose at a depth of 10 cm, $\%dd(10)_X$ (where a 1 mm lead filter is used to correct for electron contamination). There is no consensus on this problem at the moment: the new IAEA Code of Practice (8) uses $TPR_{20,10}$ while the AAPM absorbed dose protocol (7) uses $\%dd(10)_X$. However, in practice there is no real controversy: Kalach and Rogers (56) showed that although $\%dd(10)_X$ gave better agreement for a wide range of accelerators, for the heavily filtered beams produced by modern clinical linacs, $TPR_{20,10}$ was an adequate beam quality specifier.

Beam Quality Specifiers for Electron Beams. It is potentially simpler to measure the electron spectrum from a Linac than a photon spectrum. The most accurate method is to use a calibrated magnetic spectrometer (57,58), but this technique tends to be rather time consuming and the necessary equipment is not always available. The mean energy of the electron beam can be determined via activation analysis (59,60) or the determination of the total charge and energy using a Faraday cup. However, all these systems tend to be rather complex so the actual electron spectrum (or even mean electron energy) is rarely measured.

As with photons, parameters derived from the penetration of electrons in a medium are used as a measure of electron energy. A typical electron depth–dose curve in water is shown in Fig. 11. The two most important parameters obtained from such a curve are R_{50}, defined as the depth at 50% of the peak dose; and R_p (the practical range), defined as the point where the extrapolation from the point of maximum gradient on the downward part of the curve meets the extrapolation of the bremsstrahlung background.

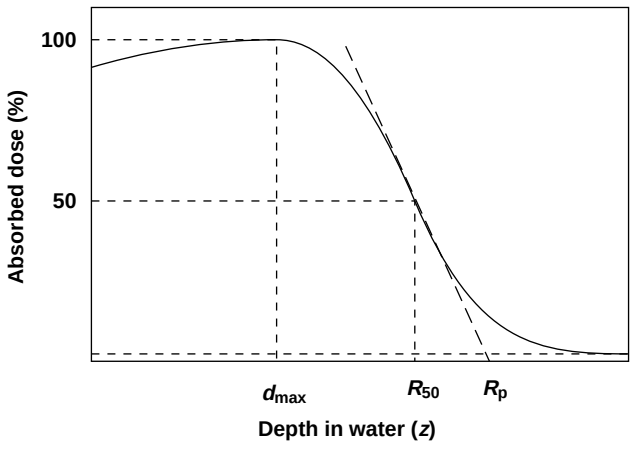


Figure 11. An electron beam depth–dose distribution in water showing the various range parameters. (From Ref. 61).

Considerable work has been done to relate these parameters to beam energy (see Ref. 2), and it is generally understood that R_{50} and R_p described different aspects of the incident electron spectrum. The parameter R_{50} relates to the mean electron energy, while R_p is directly related to the most probable energy. For a symmetrical, single-peaked spectrum the mean and most probable energies will be the same but, as shown by Klevenhagen (62), the spectrum incident on a phantom will be skewed towards lower energies due to scattering in air. Reference 2 shows depth–dose data for two spectra where the most probable energy is the same but with different mean energies (and different energy spreads). In this case, the two curves give the same value for the practical range, but different values for R_{50}. However, Burns et al. (63) collated a large amount of depth–dose data from a wide variety of linacs and showed that there was a direct relation between R_{50} and R_p, indicating that the majority of linacs in use today either generate symmetrical or very similar spectra.

RELATIVE DOSIMETRY AND QUALITY/VERIFICATION

For relative dosimetry or for quality (QA) measurements there are a wide range of dosimetry systems to choose from (including many discussed above as secondary dosimeters). The choice will depend on a number of factors including application (simple external beam therapy, intensity modulated radiotherapy (IMRT), brachytherapy); precision; spatial resolution and/or detector size; type of measurement (i.e., relative, QA, etc.); and immediacy (instant readout required?). However, one of the primary drivers will be practical issues such as cost, availability, complexity and setup time. With so many detectors to choose from it is difficult to give anything other than a very brief overview here.

Solid-State Detectors (1D)

Diodes. Semiconductor diodes offer increased sensitivity over air-filled ionization chambers due to the higher density of charge carriers. This means that the sensitive volume can be made $\sim$ 100–1000 times smaller, giving excellent spatial resolution. The stopping power ratio silicon/water varies much less with energy than the air/water ratio, and therefore diodes are ideally suited for measuring dose distributions. The biggest problem with diodes is that the sensitivity is dose dependent and diodes need recalibrating approximately every few hundred gray. Dose diodes are available in two types: electron and photon. Photon diodes employ shielding around the sensitive volume to correct for the effects of scattered radiation in a photon beam. Uncorrected, an unshielded diode overestimates the dose at depth by as much as 15% for a 6 MV beam and Yin et al. (64) present a method to correct for the response of diodes in photon beams. Due to the potential for confusion as to the construction of a diode, dosimetry protocols usually recommend that diode measurements are validated using an ion chamber.

Diamond. Diamond detectors have been investigated for over 20 years (65,66). The spatial resolution of diamond detectors (1–6 mm^3) is comparable to that of commonly used silicon diode detectors with the added advantage of showing high resistance to radiation damage (0.05% $k{\cdot}Gy^{-1}$, $>$ 100 times lower than typical diode values). As expected for a solid-state dosimeter, diamond detectors have a high sensitivity, but also a good long-term stability and low temperature dependence. Diamond has a reasonable tissue equivalence for both photon and electron beams. The majority of recent work with diamond detectors has focused on their use for small field IMRT and brachytherapy, where the high sensitivity and small size are highly advantageous. Mack et al. (67) compared a number of techniques for the dosimetry of small radiosurgery beams and found that a diamond detector gave very good results down to field sizes of 4 $\times$ 4 mm. However, a significant disadvantage of diamond detectors at present is the very high cost compared to other solid-state detectors. This is because at present natural diamonds are used and have to be carefully selected, since the dose linearity and polarization effects are very sensitive to impurities and defects in the crystal. However, the more recent availability of low cost, polycrystalline diamond specimens produced using chemical vapor deposition (CVD) offers the potential for improved diamond detectors with selectable size and impurities.

MOSFETs. The MOSFET dosimeter (68) is a more recent development. The dosimeter operates by measuring the threshold voltage of a MOSFET field effect transistor, which is permanently changed by the absorption of radiation (radiation damage). As an integrating detector it therefore has similar applications to TLD or alanine and the small detector area (only 0.2 $\times$ 0.2 mm) offers very high spatial resolution. The reproducibility is typically $\pm$ 2%.

Two-Dimensional Detectors

Radiographic Film. Film dosimetry has long been viewed as an attractive alternative to ionometric and thermoluminescent methods as an entire two-dimensional (2D) chart may be extracted from a single film exposure. In addition, film has the highest spatial resolution of any practical dosimeter and is easily set up and exposed. However, radiographic film dosimetry is not without problems - daily film calibrations are essential to obtain absolute dose results and care must be taken with film handling and processing not to introduce artifacts. A more fundamental problem is that the high atomic number of the silver halide film emulsion means that dose response relative to water varies significantly in the low energy photon range ($<$ 200 keV), as can be seen in Fig. 12. This is also an issue for measurements in megavoltage photon beams where the scatter-to-primary ratio can change (e.g. off-axis). Having said this, radiographic film dosimetry is experiencing a renaissance in the radiation therapy community, driven by the need to verify the absorbed dose delivered with IMRT, where both detector resolution and 2D data acquisition are advantages.

As with any other secondary dosimeter, a calibration curve must be derived. The quantity measured using film is

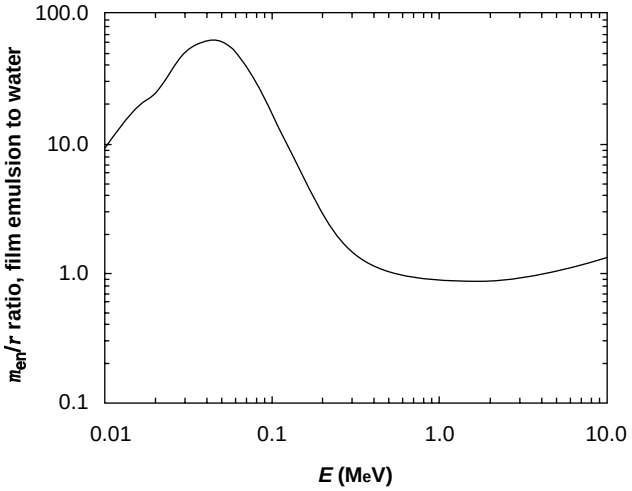


Figure 12. Energy dependence of radiographic film relative to that of water.

the optical density (measured using a scanner or densitometer) and ideally this should show a linear increase with dose (Fig. 13).

Although there are obvious problems with radiographic film, alternatives for 2D dose verification have their own limitations and therefore it is likely that film will continue to play a role in dosimetry. Radiochromic film is considered tissue equivalent and energy independent, but is expensive, limited in size, and prone to large dose response nonuniformities due to the manufacturing process. Commercially available 2D ion chamber and diode arrays provide a fast and accurate evaluation, but are still limited in resolution, and therefore better used to dose verification rather than commissioning IMRT. Electronic portal imagers have the advantage of being available in many modern therapy centres but by design, they measure fluence patterns, not dose distributions in phantoms, and therefore interpretations of delivery errors could be difficult.

Radiochromic Film. A radiation-induced color change is one of the simplest dosimeters one can think of and a number of systems were developed in the early 1900s. Although widely used in the radiation processing industry, where

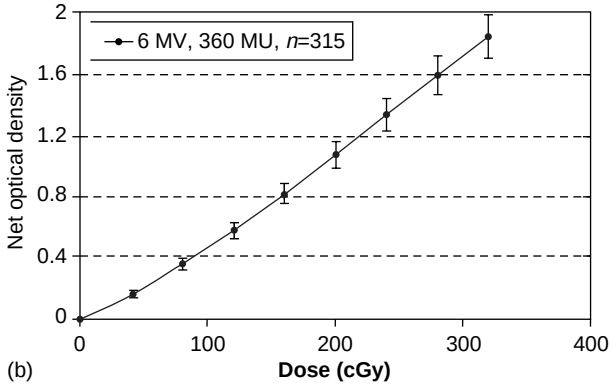


Figure 13. Dose linearity curve for Kodak EDR2 film (from Ref. 69). The response is very repeatable and linear for doses above 150 cGy.

the large doses can induce easy-to-detect color changes, radiochromic films have only recently begun to be used once again for radiotherapy dosimetry. The most promising to date is the GafChromic material. One of the main advantages of radiochromic films over radiographic is that they are essentially tissue equivalent so the energy response relative to water only changes very slowly with energy (Fig. 14).

As with any dosimeter there are problems in obtaining high accuracy dosimetric information. Klassen et al. (71) carried out a detailed investigation and showed that the precision was affected by the readout method, the readout temperature and wavelength as well as the polarization of the light source. However, with care, dosimetry with a relative uncertainty of $< 1\%$ is possible for doses of the order of 6 Gy.

EPIDs. Electronic portal imaging devices (EPIDs) have been gradually replacing conventional radiographic film for geometric verification in radiation therapy. The obvious advantage of using an EPID for dosimetry is that they are now standard equipment on most modern linacs. Early generations, employing liquid ion-chambers or camera-based fluoroscopy, generally produced poorer images compared to film, but it was shown that EPIDs could be used for IMRT quality assurance (e.g., leaf position verification for Multi- Leaf Collimators, or MLCs). The most recent class of EPID uses flat-panel photodiode arrays and with improved spatial resolution and higher detector efficiency are especially well suited for IMRT applications. However, to use any EPID for dosimetric IMRT requires calibration coefficients to relate pixel intensity to either fluence or dose. Calibration of the EPID is more involved than simple cross-calibration of pixel response with dose measurements made with an ion chamber in a homogeneous water phantom, but the ability to verify treatment "as it happens" is a significant advantage over other methods. Warkentin et al. (72) describe the use of a flatpanel detector for accurate pretreatment dosimetric verification of IMRT treatment fields.

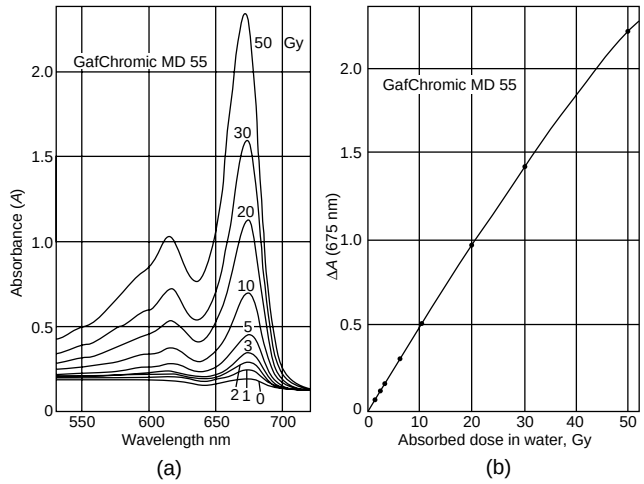


Figure 14. Performance of GafChromic MD-55 film. (From Ref. 70 reprinted with permission from Elsevier.) (a) Absorption spectra as a function of dose. (b) Dose response curve measured at the absorption band peak.

Three-Dimensional Detectors

The adoption of conformal radiotherapy techniques and, in particular, IMRT, where verification of the delivered 3D dose distribution is very important, has been a major driving force in the development of 3D detection systems. Presently, there are basically three options: TLD (either as individual pellets or in powder form), film stacks (radiographic or radiochromic), and gel dosimeters.

Gel Dosimetry. Although the use of radiation sensitive gels for dosimetry measurements was suggested as early as the 1950s, the use and development of this type of dosimeter has only grown significantly in the last decade. Gel dosimeters offer a number of advantages over other 3D techniques, such as TLD or film stacks, including resolution, number of data points, energy dependence, and water equivalence (Fig. 15).

There are currently two main types of gel dosimeter: (*1*) Fricke gel – ferrous sulfate solution is incorporated into aqueous gel matrices of gelatin, agarose or poly (vinyl alcohol) (PVA). As for the Fricke dosimeter, there is a conversion of Fe^{2+} ions to Fe^{3+} and this change in concentration is readout via magnetic resonance imaging (MRI) or optical tomography. One of the main drawbacks of Fricke gels is that there is a rapid diffusion of the ferric ions centers within the matrix, which tends to smooth out the dose distribution. (*2*) Polymer gels: This system is based on the polymerisation of certain materials. Initial work focused on the materials acrylamide (AA) and *N,N*-methylene-bis(acrylamide) (BIS) with readout again via MRI. One of the main problems with these systems is that they are sensitive to atmospheric oxygen contamination. A newer formulation named methacrylic and ascorbic acid in gelatin initiated by copper (MAGIC) is less sensitive to the presence of oxygen and looks promising as a gel dosimeter. Perhaps the biggest problem with gel systems is that they require

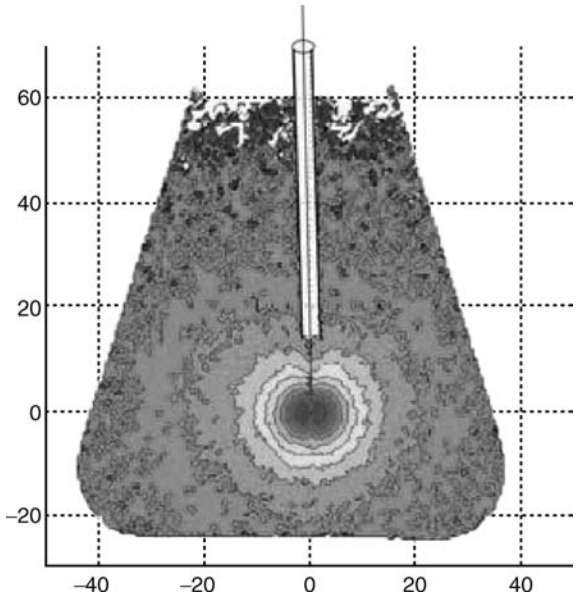


Figure 15. Dose distribution for a high dose rate ^{192}Ir brachytherapy source measured in a flask of polymer gel dosimeter. (From Ref. 73.)

a containment vessel, which can both perturb the dose measurement and introduce imaging artefacts.

There is a very active gel dosimetry community worldwide and development continues both on gel formulations (e.g., reduce diffusion or sensitivity to impurities) and readout (e.g., CT and ultrasound have been suggested as alternative readout methods to MRI). For a recent review of the subject see Baldock (74).

CONCLUSION

This article has outlined the basic theory of radiation dosimetry and the problems involved in measuring absorbed dose. A number of primary and secondary measurement techniques have been described together with the formalism for calibrating dosimeters. Since whole textbooks have been written on this subject, this can be no more than a brief introduction to the field. Readers are referred to the extensive bibliography for further detail.

BIBLIOGRAPHY

1. International Commission on Radiation Units and Measurements (ICRU) 1998 Fundamental Quantities and Units for Ionizing Radiation ICRU Report 60. Bethesda, MD: ICRU; 1998.
2. International Commission on Radiation Units and Measurements (ICRU) 1984 Radiation dosimetry: electron beams with energies between 1 and 50 MeV, ICRU Report 35. Bethesda, MD: ICRU; 1984.
3. Boutillon M, Perroche A M. Re-evaluation of the W value for electrons in dry air. Phys Med Biol 1987;32:213–219.
4. American Association of Physicists in Medicine (AAPM) AAPM TG-21: A protocol for the determination of absorbed dose from high-energy photon and electron beams. Med Phys 1983;10:741–771.
5. International Atomic Energy Agency (IAEA Absorbed dose determination in photon and electron beams: an international code of practice. International Atomic Energy Agency Technical Report 277. Vienna: IAEA; 1987.
6. Institute of Physical Sciences in Medicine (IPSM) Code of Practice for high-energy photon therapy dosimetry based on the NPL absorbed dose calibration service. Phys Med Biol 1990;35:1355–1360.
7. American Association of Physicists in Medicine (AAPM). AAPMs TG-51 protocol for clinical reference dosimetry of high-energy photon and electron beams Report of AAPM Radiation Therapy Committee Task Group No. 51. Med Phys 1999;26:1847–1870.
8. International Atomic Energy Agency (IAEA)Absorbed Dose Determination in External Beam Radiotherapy (IAEA Technical Reports Series No. 398). Vienna: IAEA; 2000.
9. Boutillon M, Peroche A-M. Ionometric determination of absorbed dose to water for cobalt-60 gamma rays. Phys Med Biol 1993;38: 439–454.
10. Klevenhagen SC. Determination of absorbed dose in high-energy electron and photon radiation by means of an uncalibrated ionization chamber. Phys Med Biol 1991;36:239–253.
11. Zankowski CE, Podgorsak EB. Calibration of photon and electron beams with an extrapolation chamber. Med Phys 1997; 24:497–503.
12. van der Marel J and van Dijk E 2003. Development of a Dutch primary standard for beat emitting brachytheraphy sources Standards and Codes of Practice in Medical Radiation Dosimetry (Proc. Int. Symp. Vienna, 2002), IAEA, Vienna.
13. Domen SR, Lamperti PJ. J Res Natl Bur Stand (US) 1974; 78:595.
14. DuSautoy AR. The UK primary standard calorimeter for photon beam absorbed dose measurement. Phys Med Biol 1996;41:137.
15. Ross CK, Seuntjens JP, Klassen NV, Shortt KR. The NRC Sealed Water Calorimeter: Correction Factors and Performance. Proceeding of the Workshop on Recent Advances in Calorimetric Absorbed Dose Standards, Report CIRM 42. Teddington: National Physical Laboratory; 2000.
16. McEwen MR, Duane S. A Portable graphite calorimeter for measuring absorbed dose in the radiotherapy clinic ≅ In: Standards and Codes of Practice in Medical Radiation Dosimetry. Proceeding of the International Symposium Vienna, 2002, Vienna: IAEA; 2003.
17. Ross CK, Klassen NV. Water calorimetry for radiation dosimetry. Phys Med Biol 1996;41:1–29.
18. Williams AJ, Rosser KE, editors. Proceedings of the NPL Workshop on Recent Advances in Calorimetric Absorbed Dose Standards NPL Report CIRM 42. Teddington: National Physical Laboratory, 2000.
19. Seuntjens JP, DuSautoy AR. Review of calorimeter based absorbed dose to water standards. In: Standards and Codes of Practice in Medical Radiation Dosimetry. Proceeding of the International Symposium Vienna, 2002. Vienna: IAEA; 2003.
20. Fricke H, Morse S. The actions of x-rays on ferrous sulfate solutions. Phil Mag 1929;7(7):129.
21. Roos M, Hohlfeld K. Status of the primary standard of water absorbed dose for high energy photon and electron radiation at the PTB. Measurement Assurance in Dosimetry. Vienna: International Atomic Energy Agency; 1994. p 25–33.
22. International Commission on Radiation Units and Measurements (ICRU) Stopping powers for electrons and positrons (ICRU Report 37). Bethesda, MD: 1984.
23. Feist H, Muller U. Measurement of the total stopping power of 5.3 MeV electrons in polystyrene by means of electron beam absorption in ferrous sulphate solution. Phys Med Biol 1989;34:1863.
24. Faddegon BA, Ross CK, Rogers DWO. Measurement of collision stopping powers of graphite, aluminium and copper for 10 and 20 MeV electrons. Phys Med Biol 1992;37:1561–71.
25. MacPherson MS. Accurate measurements of the collision stopping powers for 5 to 30 MeV electrons Ph.D. dissertation. Ottawa: INMS, NRC; 1998. PIRS-0626.
26. Burns JE, Dale JWG. Conversion of absorbed-dose calibration from graphite to water NPL Report RSA(EXT)7. Teddington: NPL; 1990.
27. Pruitt JS, Loevinger R. The photon-fluence scaling theorem for Compton-scattered radiation. Med Phys 1982;9:176–179.
28. Nutbrown RF, Duane S, Shipley DR, Thomas RAS. Evaluation of factors to convert absorbed dose calibrations in graphite to water or mega-voltage photon beams NPL Report CIRM 37. Teddington: NPL; 2000.
29. Rogers DWO, et al. BEAM: A Monte Carlo code to simulate radiotherapy treatment units. Med Phys 1995;22:503–524.
30. Kawrakow I. Accurate condensed history Monte Carlo simulation of electron transport. I. EGSnrc, the new EGS4 version. Med Phys 2000;27:485–498.
31. Aird EGA, Farmer FT. The design of a thimble chamber for the Farmer dosemeter. Phys Med Biol 1972;17:169–174.
32. Mattsson LO, Johansson K-A, Svensson H. Calibration and use of parallel-plate ionization chambers for the determination of absorbed dose in electron beams. Acta Radiol Oncol 1981;20:385–399.
33. Williams AJ, McEwen MR, DuSautoy AR. A calculation of the water to graphite perturbation factor ratios for the NACP type 02 ionisation chamber using Monte Carlo techniques. NPL

Report CIRM(EXT)013. Teddington: National Physical Laboratory; 1998.
34. International Commission on Radiation Units and Measurements (ICRU) Average energy required to produce an ion pair (ICRU Report 31). Bethesda, MD: 1984.
35. Boag JW. Ionization measurements at very high intensities. I. Pulsed radiation beams. Brit J Radiol 1950;23:601–611.
36. Boag JW, Currant J. Current collection and ionic recombination in small cylindrical ionization chambers exposed to pulsed radiation. Brit J Radiol 1980;53:471–478.
37. International Commission on Radiation Units and Measurements (ICRU) The dosimetry of pulsed radiation, ICRU Report 34. Bethesda, MD: ICRU; 1982.
38. Derikum K, Roos M. Measurement of saturation correction factors of thimble-type ionization chambers in pulsed photon beams. Phys Med Biol 1993;38:755–763.
39. Burns DT, McEwen MR. Ion recombination for the NACP parallel-plate chamber in a pulsed electron beam. Phys Med Biol 1998;43:2033–2045.
40. DeBlois F, Zankowski C, Podgorsak EB. Saturation current and collection efficiency for ionization chambers in pulsed beams. Med Phys 2000;27:1146.
41. Boag JW. Ionization Chambers. In: Attix FH, Roesch WC, Tochilin E, editors. Radiation Dosimetry. Vol. II, New York: Academic; 1966. Chapt. 9, p 2–67.
42. Williams JA, Agarwal SK. Energy-dependent polarity correction factors for four commercial ionization chambers used in electron dosimetry. Med Phys 1997;24:785–790.
43. Nisbet A, Thwaites DI. Polarity and ion recombination correction factors for ionization chambers employed in electron beam dosimetry. Phys Med Biol 1998;43:435–443.
44. Pearce JAD. Characterisation of two new ionisation chamber types for use in reference electron dosimetry in the UK, NPL Report DQL-RD001. Teddington: National Physical Laboratory; 2004.
45. Bahar-Gogani J, Grindborg JE, Johansson BE, Wickman G. Long-term stability of liquid ionization chambers with regard to their qualification as local reference dosimeters for low dose-rate absorbed dose measurements in water. Phys Med Biol 2001;46:729–740.
46. Klassen NV, Shortt KV, Seuntjens J, Ross CK. Fricke dosimetry: the difference between $G(Fe^{3+})$ for ^{60}Co gamma-rays and high-energy X-rays. Phys Med Biol 1999;44:1609–1624.
47. Ross CK, Klassen NV, Shortt KR. The development of a standard based on water calorimetry for the absorbed dose to water. Proceeding of the NPL Calorimetry Workshop.Teddington: NPL; 1994.
48. Marre D, et al. Energy correction factors of LiF powder TLDs irradiated in high-energy electron beams and applied to mailed dosimetry for quality assurance networks. Phys Med Biol 2000;45:3657–3674.
49. Davis SD, et al. The response of LiF thermoluminescence dosemeters to photon beams in the energy range from 30 kV X rays to ^{60}Co gamma rays. Radiat Prot Dosimetry 2003;106:33–43.
50. Zeng GG, McEwen MR, Rogers DWO, Klassen NV. An experimental and Monte Carlo investigation of the energy dependence of alanine/EPR dosimetry: I. Clinical X-ray beams. Phys Med Biol 2004;49:257–270.
51. Izewska J, Bera P, Andreo P, Meghzifene A. Thirty Years of the IAEA/WHO TLD Postal Dose Quality Audits for Radiotherapy. Proceeding of the World Congress on Medical Physics. Chicago: AAPM; 2000.
52. Aguirre JF, et al. Thermoluminescence dosimetry as a tool for the remote verification of output for radiotherapy beams: 25 years of experience. In: Standards and Codes of Practice in Medical Radiation Dosimetry. Proceeding of the International Symposium Vienna; 2002, Vienna: IAEA; 2003.
53. Rosser KE, et al. The NPL absorbed dose to water calibration service for high energy photon beams. In: Flitton SP, editor Proceeding of the International Symposium on Measurement Assurance in Dosimetry (IAEA-SM-330/35) Vienna: IAEA; 1994. p 73.
54. Kalach NI, Rogers DWO. Which accelerator photon beams are 'clinic-like' for reference dosimetry purposes? Med Phys 2003;30:1546–1555.
55. Wessels BW, Paliwal BR, Parrot MJ, Choi MC. Characterization of Clinac-18 electron-beam energy using a magnetic analysis method. Med Phys 1979;6:45.
56. Deasy JO, Almond PR, McEllistrem MT. Measured electron energy and angular distributions from clinical accelerators. Med Phys 1996;23:675.
57. Almond PR. The physical measurements of electron beams from 6 to 8 MeV: absorbed dose and electrical calibration. Phys Med Biol 1967;12:13.
58. Klevenhagen SC. Physics and Dosimetry of Therapy Electron Beams. Madison: Medical Physics Publishing; 1993.
59. Institution of Physics and Engineering in Medicine and Biology (IPEMB) The IPEMB code of practice for electron dosimetry for radiotherapy beams of initial energy from 2 to 50 MeV based on an air kerma calibration. Phys Med Biol 1996;41:2557–2603.
60. Klevenhagen SC. Physics of electron beam therapy. Bristol: Adam Hilger; 1985. p 65.
61. Burns DT, Ding GX, Rogers DWO. R_{50} as a beam quality specifier for selecting stopping-power ratios and reference depths for electron dosimetry. Med Phys 1996;23:383.
62. Yin Z, Hugtenberg RP, Beddoe AH. Response corrections for solid-state detectors in megavoltage photon dosimetry. Phys Med Biol 2004;49:3691–3702.
63. Laub WU, Kaulich TW, Fridtjof N. A diamond detector in the dosimetry of high-energy electron and photon beams. Phys Med Biol 1999;44:2183–2192.
64. Planskoy B. Evaluation of diamond radiation dosemeters. Phys Med Biol 1980;25.
65. Mack A, et al. Precision dosimetry for narrow photon beams used in radiosurgery—Determination of Gamma Knife[sup [registered sign]] output factors. Med Phys 2002;29:2080–2089.
66. Ramani R, Russell S, O'Brien P. Clinical Dosimetry Using MOSFETS. Int J Rad Oncol Biol Phys 1997;37:956–964.
67. Childress NL, White RA, Rosen II. Dosimetric accuracy of Kodak EDR2 film for IMRT verifications. Med Phys 2005;32:539–548.
68. McLaughlin WL, Desrosiers MF. Dosimetry systems for radiation processing. Radiat Phys Chem 1995;46:1163–1174.
69. Klassen NV, van der Zwan L, Cygler J. GafChromic MD-55: Investigated as a precision dosimeter. Med Phys 1997;24:1924–1934.
70. Warkentin B, Steciw S, Rathee S, Fallone BG. Dosimetric IMRT verification with a flat-panel EPID. Med Phys 2003;30:3143–3155.
71. De Deene Y, Reynaert N, De Wagter C. On the accuracy of monomer/polymer gel dosimetry in the proximity of a high-dose-rate ^{192}Ir source. Phys Med Biol 2001;46:2801–2825.
72. Baldock C. Radiotherapy gel dosimetry. In: Standards and Codes of Practice in Medical Radiation Dosimetry. Proceeding of the International Symposium Vienna, 2002. Vienna: IAEA; 2003.

Reading List

The following textbooks are recommended for further reading; they provide excellent coverage of the subject of radiation dosimetry.

Attix FH, Roesch WC, Tochilin E, editors. Radiation dosimetry. (Pts I,II,III) 2nd ed. New York: Academic Press; 1966–1969.
Greening JR. Fundamentals of radiation dosimetry. 2nd ed. Bristol [England]: A. Hilger in collaboration with the Hospital Physicists' Association; 1985.
Johns HE, Cunningham JR. The physics of radiology. 4th ed. Springfield, IL: Thomas.
Kase KR, Bjärngard B, Attix FH, editors. The Dosimetry of ionizing radiation. (Pts I, II, III) Orlando, FL: Academic; 1985.
Klevenhagen SC. Physics and Dosimetry of Therapy Electron Beams. Madison, WI: Medical Physics Publishing; 1993.

See also IONIZING RADIATION, BIOLOGICAL EFFECTS OF; RADIATION THERAPY SIMULATOR.

RADIATION DOSIMETRY, THREE-DIMENSIONAL

GEOFFREY S. IBBOTT
Anderson Cancer Center
Houston, Texas

INTRODUCTION

The goal of radiation therapy is to obtain the greatest possible local and regional tumor control, with the fewest complications. The response of many tissues to radiation can be characterized by a sigmoid curve. Relatively little response is seen until the dose reaches some threshold value, after which the response is quite rapid (1,2). In the region of steep response, relatively small variations in dose can yield significant differences in the response of both tumors and normal tissue (3). To minimize the variability of tissue response, the ICRU has recommended that the uncertainty in dose delivery be maintained below $\sim$ 5% (4–6). Delivering a dose to a patient with a tolerance of 5% is not a simple matter (7). It has been estimated that the equipment used by most medical physicists to calibrate therapeutic radiation beams is itself calibrated with an overall uncertainty (expressed at the 95% confidence level) of $\sim$ 1.5% (8). Uncertainties associated with the characterization of radiation beams, patient anatomy, and location of the target volume, as well as reproducibility of the treatment from day to day must be considered (9,10).

A comprehensive radiation therapy quality assurance program must address all sources of variability in the treatment of patients, in an effort to minimize variations. Technical aspects of quality assurance (QA) must address a wide array of issues, including the performance of simulation, treatment, and treatment planning equipment, the stability of measurement and test equipment, the accuracy and appropriateness of treatment planning calculations, and the accuracy and completeness of documentation. Technical quality assurance procedures should also address inventory, calibration, and treatment planning with brachytherapy sources. Recommendations for QA procedures can be found in a number of publications (11–23).

As the equipment used to deliver radiation therapy has evolved, methods of radiation dosimetry have also changed. Multifield, conformal radiation therapy (CRT), intensity-modulated radiation therapy (IMRT), stereotactic radiosurgery (SRS), and stereotactic radiation therapy (SRT) all produce dose distributions that can be highly irregular in three dimensions. Conventional two-dimensional (2D) planning and dosimetry systems are not adequate to simulate and measure such distributions. Instead, new dosimetry systems are required that can record and display these complex distributions (24). This article addresses recent developments in dosimetry systems, and their advantages and complications.

QUALITY ASSURANCE PROCEDURES REQUIRING DOSIMETRY SYSTEMS

External Beam Calibration Consistency-Basic Parameters

Detector systems are required for measurement of accelerator output, for compliance with published recommendations for quality assurance. Most published recommendations suggest that accelerator output constancy be monitored on a daily basis. Consequently, a dosimeter system that is rugged, reliable, and easy to operate is required. Most recommendations for daily output consistency suggest that deviations on the order 2–5% be detectable; therefore the dosimetry system does not need extremely high accuracy.

In addition, measurements of beam flatness and symmetry are recommended on a periodic basis, often weekly. Again, as these measurements are to demonstrate consistency of operation at the 2–5% level, high precision is not required. Several of the available array dosimeters systems are suitable for such frequent QA measurements of treatment unit performance.

External Beam Treatment Delivery, Planning Verification

Several treatment applications require the verification of delivered dose with relatively high accuracy. For example, IMRT requires the precise delivery of relatively small doses through a large number of fields. Even small errors in dose delivery can accumulate and result in a large error in the final dose. Monitoring of dose delivery during IMRT is done best using a real time measuring device, such as online portal imaging. Similarly, CRT delivery demands confirmation that the correct dose has been delivered. As CRT is generally delivered through static fields, point detectors may be used to measure the delivered dose in a suitable phantom. Several of the simpler point dosimeter systems described earlier are suitable for this purpose.

Likewise, SRS and SRT delivered with accelerators may need verification, particularly as SRS is delivered in single large fractions. Again, under most circumstances, point dosimeters are suitable here. However, the characterization of radiation beams for SRS–SRT requires a dosimeter with high spatial accuracy. Several of the detector systems described above would satisfy this requirement, although questions of electronic equilibrium must be addressed (25).

Total body treatments, such as photon TBI for systemic bone marrow ablation, or total skin electron therapy for cutaneous t-cell lymphoma, may require dosimetry to confirm the correct delivery of dose under these conditions of unusual field size and distances.

External Beam Treatment Delivery In Vivo

Modern external treatment delivery requires that doses be delivered with accuracy never before required. Procedures, such as IMRT, are delivered through many field segments, each delivering a small increment of dose. A systematic error in dose delivery can result in a significant error in the final dose received by the patient. Consequently, dosimetry devices for confirming correct dose delivery are necessary. These fall into three broad classes: surface dose measurements, transmission measurements, and true *In vivo* measurements.

Brachytherapy (LDR, HDR, IVBT)

Dosimeter systems are required for at least three purposes related to brachytherapy: source characterization, confirmation of dose distributions from arrangements of multiple sources, and *In vivo* dose measurements (26,27).

Imaging Procedures

Dosimetry measurements are required in cardiology, for procedures, such as cardiac ablation, in which patients can receive significant doses. A detector to be used in imaging must not be intrusive, meaning that it must be virtually transparent to the beam. It must measure dose over a large area, although accommodation needs to be made for the possibility that the beam may be moved during irradiation. Finally, a device attached to the source of radiation, such as a dose area product meter, may be used.

REQUIRED CHARACTERISTICS OF DOSIMETERS FOR QUALITY ASSURANCE

A dosimeter for modern CRT must possess a number of important characteristics. It must be tissue equivalent, as the dosimeter itself must not perturb the dose distribution. It must have a linear dose response over a clinically useful range. Ideally, its response would be independent of dose rate and of beam modality, making it useful for mapping dose distributions from isotope units, linear accelerators, or particle accelerator beams. Some dosimeters must be able to fill a volume, or conform to a surface. This will enable the dosimeter to either mimic any portion of human anatomy, or conform to a section of an anthropomorphic phantom.

The dosimeter must either provide immediate results or be stable for a sufficiently long period to enable irradiation and analysis. Under some circumstances, the delivery of the intended dose distribution may take some time, as is the case with brachytherapy. It is important that the dosimeter remain uniformly sensitive, and unaffected in response over the time required for irradiation. Further, the dosimeter must maintain the dose-deposition information throughout the time required for analysis. For some applications, it may be desirable to transport the dosimeter to another facility for analysis. The dosimeter must remain stable during shipment, unaffected by a variety of environmental conditions, throughout the analysis.

The accuracy and precision required of dosimeters for radiation therapy measurements depend on the intended use of the detector. Devices intended for reference calibration of treatment units should enable the determination of dose with an uncertainty of no more than 0.5%, expressed at the 95% confidence level ($k = 2$) (28,29). Dosimeters intended for verification of dose distributions should provide an uncertainty in dose measurement of no more than 2%, again expressed at the 95% confidence level ($k = 2$).

DETECTORS FOR THREE-DIMENSIONAL DOSIMETRY

Detector Arrays

A number of manufacturers have marketed arrays of conventional detectors using either ion chambers or diodes. These devices are not true three-dimensional (3D) dosimeters, but are included here because they provide 3D information through the use of one, or at most two, manipulations, such as translation across a beam. For example, linear diode arrays are available for the Scanditronix water phantom system, and for stand-alone QA device, such as the Sun Nuclear profiler. An array of ionization chambers has been described for verifying treatment planning for IMRT (30). The ion chambers are arranged in several parallel linear arrays, each one offset from the next. Twenty-four chambers, each 0.03 cm^3 in volume, are arranged in boreholes of a plastic-mounting frame. The assembly is positioned in a water phantom and maybe positioned in different orientations to allow measurements in different plains. Commercial ion chamber devices include the Thebes marketed by Nuclear Associates, an ion chamber array marketed by Wellhofer, the RBA-5 marketed by Gammex, and other devices. These devices range in number of detectors from few (four or five) to many (Fig. 1).

Plastic Scintillator

Some organic plastics fluoresce visible light when irradiated with ionizing radiation. Unlike the fluorescent screens used in imaging, organic scintillators have the

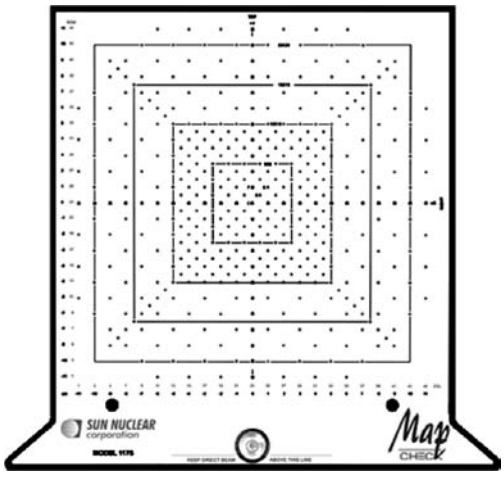


Figure 1. A diode array designed to display the intensity map of a therapy beam in real-time. (Courtesy of Sun Nuclear Corp.)

additional advantage of being approximately tissue-equivalent (31). However, this tissue-equivalence at present only exists at the energies conventionally used for megavoltage treatment. Most of the plastic scintillators currently available exhibit significant differences in the mass energy absorption coefficient relative to that of water. More recently, plastic scintillators have been developed for low energy photon dosimetry that are radiologically water-equivalent, have improved sensitivity over some others scintillators, and offer the potential for high spatial resolution (32,33).

Plastic scintillators may be used as point detectors, in which their potential for manufacturing into very small sizes yields the possibility for improved spatial resolution of measurements. Efforts also have been made to use plastic scintillators as 2D and 3D detectors (34). Two techniques have been used; the first being the use of plastic scintillators as a detector system themselves, using optical coupling through a light pipe assembly to a video detector. This method has been described previously (35–37). Significant difficulties still exist with the spatial resolution of these systems. Light emitted by the scintillator can travel some distance, in any direction, before reaching the light detector. Unless the plastic scintillator is thin, the resolution of the image will be degraded considerably. Some efforts have been to quench the light by adding dyes to the scintillator, to reduce the distance traveled by the light obliquely through the scintillator. Until this problem is resolved, the quality of the imaged dose distribution will not be adequate for radiation dosimetry.

A second technique involves the use of plastic scintillators to enhance the response of another detector, such as radiographic film (38). In this technique, radiographic film is sandwiched between sheets of organic plastic scintillator. Several investigators have noted that radiographic film has a tendency to overrespond to low energy photons (39,40). The use of an organic plastic scintillator has been proposed to enhance the response of radiographic film to higher energy photons, thus making the energy response of the film detector system more uniform.

Film

Radiographic film has long been used as a radiation detector, and as a QA device. Again, film itself is not a 3D dosimeter, but stacks of film have been used to measure dose distributions in 3D. The difficulties with film are well known; energy dependence, requirements for processing, variations from one batch to the next, dose rate dependence, positional dependence, and other issues have been discussed by a number of investigators (41). More recently, use of radiochromic film has been proposed. Radiochromic film requires no processing, has very little energy dependence, no known dose rate dependence, and requires minimal special handling techniques (42,43). The linearity of response of a recently developed model of film is shown in Fig. 2.

The use of film for verification of conformal and IMRT dose distributions has been recommended. At least one manufacturer has marketed a phantom intended for use with IMRT (see Fig. 3, for an example of such a device).

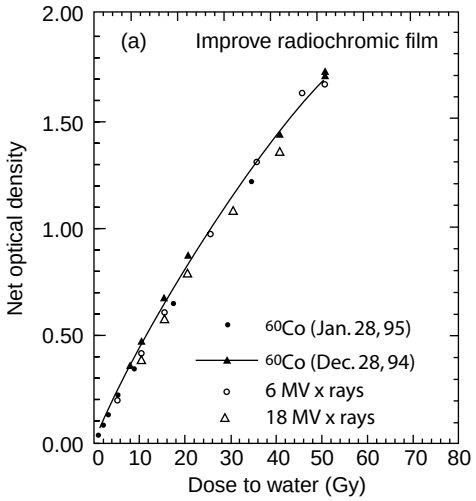


Figure 2. Energy dependence and linearity with dose of an improved radiochromic film. (Reproduced with permission from Ref. 42).

TLD Sheets and Plates

Lithium fluoride, a thermoluminescent material, has been used for many decades as a radiation detector (44). Its use has been limited principally to point measurements, because the dosimeter is provided either as extruded rods or chips, or as a powder that is encapsulated for use. Thermoluminescent dosimetry has a number of limitations, among them energy dependence, but most notably a requirement for delay between irradiation and processing. In addition, an expensive piece of equipment is required for readout of the material. The limitation of the device to point measurements has been addressed recently by the development of TLD sheets. In these, TL material is distributed in an array across a sheet of backing film. The film can be irradiated in much the same manner as conventional radiographic film, and may be immersed in a water phantom as necessary. As with film, 3D measurements can be made only by stacking multiple sheets of TL

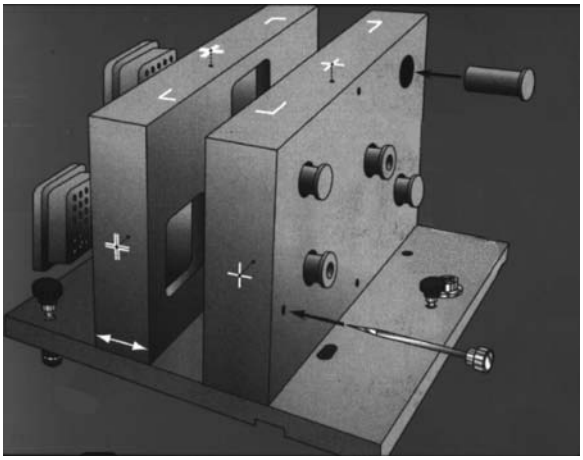


Figure 3. A phantom marketed for evaluating IMRT dose distributions. (Courtesy of Med-Tec Corp.)

material. After irradiation, and following the requisite delay, the film is inserted into a readout device that selectively heats the individual dosimeter regions using a laser. Light is collected from the heated regions using a photomultiplier tube. Through an automated operation, a matrix of data can be obtained quickly and efficiently. However, due to the cost of the reader, this dosimetry system is presently available only as a service (Inovision, Inc.)

Electronic Portal Imaging Devices

An important aspect of quality assurance in radiation therapy involves not just the dose delivered to the patient, but the correct positioning of the patient. For many years, positioning has been verified through the use of conventional radiographic film, or through the use of video imaging techniques (45,46). Video imaging permits only a check of the relative position of external landmarks. Radiographic film permits verification of the patient position through the visualization of internal boning anatomy, but requires a delay while the film is processed. The introduction of electronic portal imaging has brought to the clinic the possibility of immediate verification of patient position. With the introduction to clinical radiation therapy of modern techniques, such as IMRT, immediate verification of correct beam delivery is crucial. The failure or incorrect programming of a multileaf collimator can result in a completely unacceptable dose distribution. With on-line portal imaging, such errors may be detectable promptly, even during treatment (47–52). A further improvement has been the introduction of transmission flat-panel detectors up- and downstream from the patient. These allow the measurement of photon beam fluence entering and exiting the patient, and the estimation of dose within the patient. When combined with images of the patient made at multiple beam angles, as is done for multifield conformal treatment, or IMRT, it may be possible to reconstruct the dose distribution actually delivered to the patient in three dimensions.

GEL DOSIMETRY

Gel dosimetry has been examined as a clinical dosimeter since the 1950s (53,54). During the last two decades, however, the number of investigators has increased rapidly, and the body of knowledge regarding gel dosimetry has expanded considerably (55,56). Gel dosimetry is still considered by some to be a research project, and the introduction of this tool into clinical use is proceeding slowly. However, the interest in, and potential of, gel dosimetry for clinical use is demonstrated by the level of participation in three successful international workshops held to date on this subject (57–59). This section reviews the development of gel dosimetry, several of the formulations that have been investigated intensively, the characteristics of gel dosimetry that make it desirable for clinical use, the postulated and demonstrated applications of gel dosimetry, and some complications, setbacks, and failures that have contributed to the slow introduction into routine clinical use.

Fricke Gels

Nuclear magnetic resonance (NMR)-based gel dosimetry was introduced by Gore who recognized that the ferrous sulfate Fricke dosimeter (60,61) could be examined with magnetic resonance rather than spectrophotometry (55). The Fricke dosimeter is based on the radiation-induced and dose dependent transformation of ferrous (Fe^{2+}) ions into ferric (Fe^{3+}) ions. These two ions have different electron paramagnetic spin states and different ionic radii (60,61). Gore realized that the NMR spin–lattice and spin–spin relaxation rates ($1/T_1$ and $1/T_2$, respectively) of the water protons in the Fricke dosimeter are dependent on the amount of ferric ion present in the solution and that, because changes in these parameters produce the contrast of MR images, radiation induced changes in the solution should be visible by MRI (55). Soon afterward, other researchers began investigating the use of Fricke solutions incorporated into gel matrices (Fricke gels) to provide spatial stability of the dosimeter (62–65). The most common matrices investigated were gelatin, agarose, and sephadex. Each of these systems had its advantages and limitations, but agarose was probably used more than any other detector system. While agarose dosimeters are more sensitive to dose than gelatin-based systems, they are more difficult to produce because they must be bubbled with oxygen to ensure a uniform dose response.

Fricke gel dosimeters have a number of advantages; principle among them is the well-described understanding of the radiation chemistry of this system. In addition, the basic and NMR processes leading to the dosimetry response are well understood (66,67). Fricke gel dosimeters are tissue equivalent over a large range of photon energies. Like other gel dosimeters, they are prepared in a liquid form so that phantoms containing heterogeneities or conforming to anthropomorphic geometries can be constructed.

However, there are a number of significant problems associated with the use of Fricke gels for radiation dosimetry. The dosimeters require high doses, on the order of 10–40 Gy, for the radiation-induced changes to be observed by magnetic resonance imaging (MRI). The ferric ions produced by absorption of radiation diffuse readily through the gel or agarose matrix, leading to a decrease in signal intensity, and a loss of spatial information (64,66–69). Imaging must be performed within ~ 2 h of irradiation to avoid serious degradation of the dosimetric detail (70). The diffusion has been reduced by replacing the gelatin matrix with a poly (vinyl acohol) (PVA) matrix, which is less porous to the ferric ions (71). Other investigators have developed further methods to delay diffusion, although imaging must still be performed quite soon after irradiation (72). Some improvement in the diffusion of ions can be achieved by cooling the gel, but this is rarely practical in a clinical setting. Consequently, Fricke gel dosimetry has seen only limited clinical use.

Several improvements have been reported recently. For example, a Fricke gel dosimeter manufactured using a PVA cryogel technology has been described. The PVA is a common water-soluble polymer that can be cross-linked into its cryogel form by simply freezing and thawing. The cryogel is a rubber like material that holds its shape even at elevated temperatures. Preliminary reports of the PVA

Fricke gel dosimeter indicate that its ($1/T_1$) response has been found to be linear from 0 to 10 Gy, and the ion diffusion constant was found to be only 0.2–0.5 that of traditional preparations in gelatin or agarose (73,74). Representative ion diffusion constants are presented in Table 1 for several gel mixtures (68).

Some preliminary work using Fricke gel dosimetry in anthropomorphic phantoms has been reported (79). Several different gel compositions were investigated, including a lung equivalent gel that was developed with a density of 0.4 $g \cdot cm^{-3}$. This allows measurements of dose within the heterogeneity itself. However, diffusion of ions continues to be a problem with this dosimetry system.

Polymer Gels

Gels that replaced the Fricke solution with acrylic monomers were introduced in 1992 (80–82). Early work was conducted using a polyacrylamide gel based on the radiation-induced polymerization and cross-linking of bis and acrylamide. The formation of acrylic polymer chains largely resolved the problem of diffusion exhibited by Fricke gels, as the long polymer chains were too large to diffuse rapidly. The reciprocal of T_2, or R_2, the relaxation rate, was found to vary proportionally with dose, and MR imaging of polymer gels was shown to yield quantitative dose distributions (81). Subsequently, alternative gel formulations have been developed in which the bis and acrylamide are replaced with acrylic acid or methacrylic acid, which has yielded increased sensitivity of the gels, and reduced toxicity (83,84). However, the polymer gels continued to show another disadvantage; their response was inhibited by the presence of oxygen. This effect was addressed though the recent introduction of a class of polymer gel dosimeters containing oxygen scavengers (85,86). Several variations of these *normoxic* gel dosimeters (so-called because they can be prepared under normoxic conditions) have been characterized (87).

To avoid the disadvantages of the Fricke gel systems, a polymerizing gel dosimetry system was developed (MGS Research, Inc., Guilford, CT). A variety of polymerizing gels have been developed, many of which are based on acrylamide or acrylic acid, and are referred to as polyacrylamide gels (PAG). The dosimeters are based on

Table 1. Summary of Diffusion Measurements in the Literature[a]

Reference	Diffusion Coefficient, 10^{-3} $cm^2 \cdot h^{-1}$	Gel Type and Concentration, %	Other Constituents, m*M*	Temperature, °C
(64)	18.3 ± 1.4	A 1	S 12.5, Fe^{3+} 1	
(64)	15.8 ± 1.1	A 1	S 25, Fe^{3+} 1	
(66)	19.1 ± 1.0	A 1.5	S 50, Fe^{2+} 1	25
(75)	10.9 ± 1.6[b]	A 1	S 50, Fe^{2+} 1, NaCl 1	15–17.5
(68)	9.7 ± 1.1	A 1	S 30, Fe^{2+} 1	22
(68)	11.9 ± 1.8	A 1	S 30, Fe^{2+} 1	22
(69)	12.5 ± 1.1	Agar	S 50, Fe^{2+} 1, NaCl 1	5
(69)	21.3 + 0.5	Agar	S 50, Fe^{2+} 1, NaCl 1	24
(76)	8.2 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5	10
(76)	9.1 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5, Fo 70	20
(76)	10.4 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5, P 0.6	10
(76)	4.4 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5, P 0.6	10
(76)	0.7 ± 0.1	G 8	S 26, Fe^{2+} 0.2, BE 5, Fo 46	20
(76)	1.0 ± 0.1	G 8	S 26, Fe^{2+} 0.2, BE 5, Fo 46, P 0.6	20
(76)	4.4 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5, XO 0.2	10
(76)	6.5 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5, BD 0.6	10
(76)	6.1 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5, Fo 46, XO 0.2	20
(76)	6.3 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5, AC 0.6	20
(76)	8.3 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5	10
(77)	14 ± 3	A 1.5	S 50, Fe^{2+} 0.5	22
(77)	20 ± 5	A 1.5	S 100, Fe^{2+} 0.5	22
(77)	22	A 1.5	S 200, Fe^{2+} 0.5	22
(77)	11	A 1.5	S 50, XO 0.25	22
(77)	5 ± 1	G 10	S 50 and 100, Fe^{2+} 0.5	22
(77)	9	A 1.5, G 3	S 50, Fe^{2+} 0.5	22
(77)	9	A 1, G 2	S 200, Fe^{2+} 0.5, XO 0.2	22
(77)	3 ± 1	A 1.5, G 3	S 50 and 100, Fe^{2+} 0.5, XO 0.1 & 0.25	22
(78)	14.6 ± 0.1	G	S 50, Fe^{2+} 1.5, XO 1.5	
(78)	8.1 ± 0.1	G	S 50, Fe^{2+} 1.5, XO 1.5	
(78)	8.2 ± 0.1	G + BA	S 50, Fe^{2+} 1.5, XO 1.5, BE 5.0	
(78)	17.8 ± 0.2	A 1.5	S 50, Fe^{2+} 1.5, XO 1.5	
(78)	16.3 ± 0.2	A 3	S 50, Fe^{2+} 1.5, XO 1.5	
(71)	1.4	PVA 20	S 50, Fe^{2+} 0.4, XO 0.4	20

[a]A - agarose, Agar= agar, g = gelatin, S= H_2SO_4, XO = xylenol orange, BE = benzoic acid, Fo = formaldehyde, P = phenanthroline, AC = acetylacetone, BD = bathophenanthroline disulfonic acid.

[b]Diffusion coefficient calculated in Ref. 76.

Table 2. Composition of BANG3 Polymer Gel Dosimeter

6% Methacrylic acid
1% Sodium hydroxide
5% Gelatin
88% Water

radiation-induced chain polymerization of acrylic monomers dispersed in a tissue-equivalent gel. The BANG polymer gel system is a proprietary PAG dosimeter made of a mixture of acrylic monomers in a tissue-equivalent gel. Early BANG gels were made from acrylic acid monomers and methylene-bis(acrylamide) cross-linker. More recently, the BANG3 dosimeter was introduce, which contains methacrylic acid monomer (see Table 2, from Ref. 84). Other proprietary response modifiers were added to adjust the dose range and sensitivity. Dissolved oxygen inhibits free radical polymerization reactions and is removed from the mixture by passing nitrogen through it while the gel remains above the gelling temperature, prior to sealing the vessel. Consequently, vessels of glass or other material not permeable to oxygen must be used for irradiating and imaging the gels.

The gelling agent in the BANG dosimeter is gelatin, which is used because the transverse NMR relaxation rate of water ($R_2 = 1/T_2$) in a gelatin gel is nearly an order of magnitude lower than that in agarose gels. Therefore the background R_2 in the gel is substantially reduced, which improves its dynamic range.

MR Imaging of Polymer Gels

Irradiation of the polymer gels induces polymerization and cross-linking of the acrylic monomers. As polymer microparticles are formed, they reduce the NMR relaxation times of neighboring water protons. Magnetic resonance imaging can be used to measure dose distributions in the gel (81,82,88). Water proton NMR (^{1}H NMR) transverse relaxation time T_2 can be τ determined from multiple spin–echo images. Images can be acquired using the Hahn spin–echo pulse sequence: $90^\circ - \tau - 180^\circ - \tau -$ *acquire* for four or more different values of τ. Typical pulse sequence parameters are TR = 2 s, TE = 11, 200, 400, and 600 ms. A field of view of 24 cm and a matrix of 128 × 256 can be used, with one acquisition and a 3 mm slice thickness.

More recently, it has been shown that spin–echo sequences other than the Hahn sequence described above can be used for gel imaging. Improved dose resolution can be achieved through the use of multiple spin–echo pulse sequences (89,90). Optimization of the imaging sequence is necessary, especially with regard to the number of echoes measured. The use of multiplanar imaging can reduce imaging times but can also lead to interference between image planes.

Once MR images have been obtained, they are most conveniently transferred via network to a computer for which a data analysis and display program has been written. One example of such a program has been described previously (82). The program calculates R_2 maps on the basis of multiple TE images, using a monoexponential nonlinear least-squares fit based on the Levenberg–Marquardt algorithm (91). The program also creates a dose-to-R_2 calibration function by fitting a polynomial to a set of dose and R_2 data points, obtained from gels irradiated in test tubes to known doses. This function can then be applied to any other R_2 map so that a dose map can be computed and displayed.

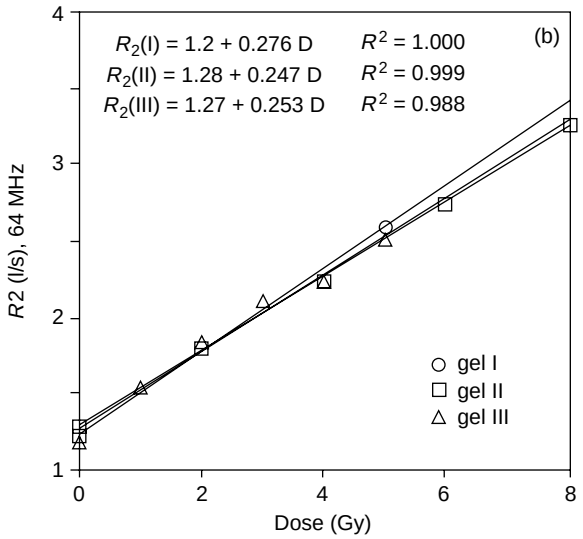


Figure 4. The dose dependence of the transverse relaxation rate (R_2) as a function of dose. Data from several experiments are shown indicating reproducibility over a wide range of doses (92).

Figure 4 shows values of transverse relaxation rates (R_2) for the gels as a function of dose. The pooled data show that the dose response was highly reproducible over a wide range of doses. The dose response is well fitted by a straight line (92).

Additional experiments have shown that the response of the BANG gel can be adjusted by varying the concentration of cross-linker used per total amount of comonomer (93). Figure 5 demonstrates the relationship of R_2 to dose for five different values of the weight fraction of cross-linker per total comonomer. Figure 5b shows that, in the linear region of gel response, the greatest sensitivity of the gel was achieved at 50% cross-linker concentration. Similar data have been shown more recently for several different polymer gel mixtures (94).

The temperature of imaging has a large effect on both the gel sensitivity and its dynamic range (93,95). Dose sensitivity ($s^{-1} \cdot Gy^{-1}$) as a function of concentration of cross-linker is plotted in Fig. 6. Sensitivity is seen to reach a maximum at ~ 50% cross-linker (as described above), but sensitivity at all concentrations increases as the temperature at the time of imaging is reduced.

Figure 7 shows that the maximum R_2 achievable, and therefore the dynamic range of the gel, is dependent on the temperature at the time of imaging. While $R_2{}^{max}$ increases with cross-linking, the dependence is enhanced by cooling the gel during NMR measurement.

For a number of gel compositions presently being evaluated, the fundamental chemistry and physics of response are well understood. Several gel compositions have been characterized in great detail (82,87,92,96–98). In polymer gels, for example, it is understood that the interaction of

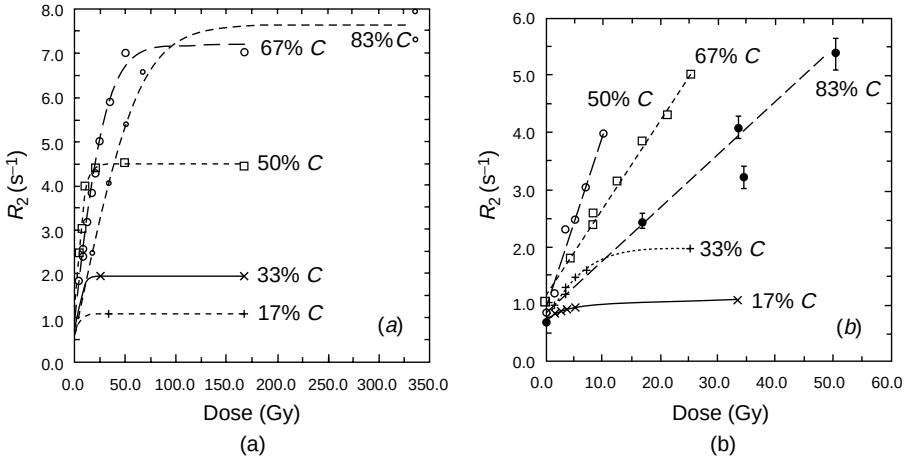


Figure 5. (a): Relationship between R_2 and dose for five different concentrations of cross-linker per total comonomer. (b) Lower dose region of the data from (a). (Reproduced with permission from Ref. 93.)

radiation with water produces free radicals, which trigger the cross-linking of monomers into polymer chains (81,99). The polymer chains bind water protons tightly causing a change in their paramagnetic properties that is detectable by magnetic resonance imaging (92,100). The relationship between dose and relaxation rate can be influenced by several additional factors, including accuracy of the calibration curve (101) and the aging characteristics of the gel (96,102,103).

The quality of the imaging process is affected by the homogeneity of the B_1 field (104) and the presence of eddy currents (105). Some additional complications due to the distortion of MR imaging systems have been identified (106).

Optical Scanning of Polymer Gels

Dosimetric results with MR imaging have been encouraging, but the need to use expensive and often inaccessible imaging systems renders this technique somewhat impractical. In most compositions, polymerization changes the optical characteristics, and measurements of optical density can be related to absorbed dose (85,107–111).

Optical computed tomography (OCT) of polymer gels can be conducted in a similar manner to X-ray CT. To date, OCT has been limited to transmission measurements, although the potential exists for measurements of attenuation, fluorescence, scatter, polarization and refractive index changes (112). Optical computed tomography has been performed by several investigators (107,109–113), but in general, the techniques all require the use of a cylindrical vessel to hold the gel, a tank filled with a medium matching the refractive index of the gel, and a monochromatic light source. Several of these systems use parallel-ray geometry and filtered back projection to reconstruct the image. At least one system uses a diffuse white light and cone-beam geometry (111).

An optical imaging system employing He–Ne laser CT scanning of the gel has been described (107). The scanner operates in a translate-rotate geometry and is capable of producing stacks of planar dose distributions with pixel size and slice thickness as small as 100 μm (114).

Optical scanning of several gels has been conducted using a modified version of a 3D optical CT laser scanner that was developed recently at MGS Research, Inc. (107,108,115,116). (see Fig. 8.)

The scanner, which is PC controlled, operates in a translate-rotate geometry and utilizes a single He–Ne laser

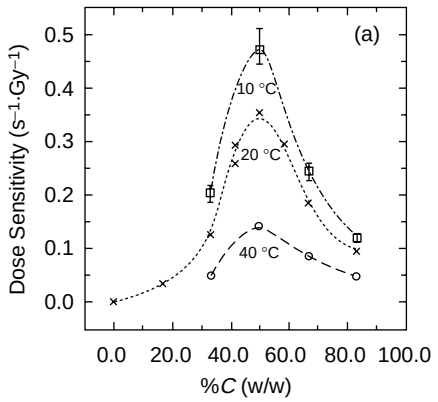


Figure 6. Dependence of the dose sensitivity on the cross-linker content, for three different temperatures. (Reproduced with permission from Ref. 93.)

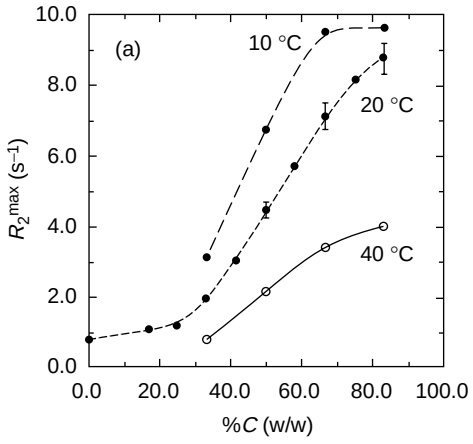


Figure 7. Dependence of the maximum R_2 on cross-linker fraction for three different temperatures. (Reproduced with permission from Ref. 93.)

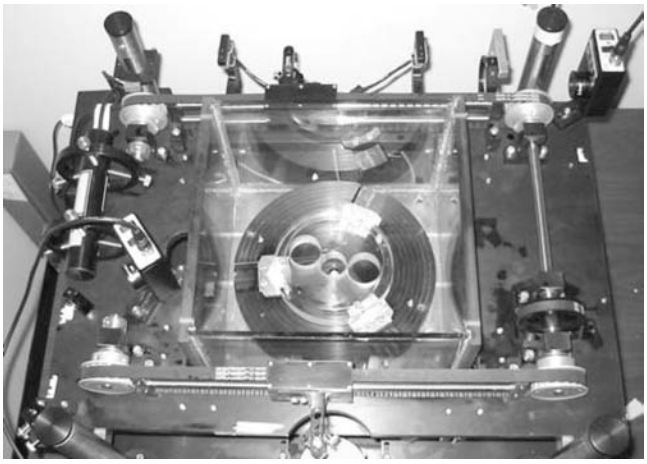


Figure 8. An optical scanner developed for use with polymer gels. (Photograph by M. Heard.)

light source and a photodiode, together with a reference photodiode to account for fluctuations of the laser output intensity. The gel is mounted on a central turntable and is immersed in a liquid that matches the gel's refractive index to minimize the loss of signal from projections at the edges of the gel. The platform on which the light source and the detector are mounted moves vertically. Isotropic resolution of 1 mm is achievable using this scanner, with scan times on the order of 8 min per plane. An image of a gel exposed to an ^{192}Ir high dose-rate (HDR) source appears in Fig. 9.

Further evaluation of an OCT system has been performed, to determine the stability and reproducibility of the system (118). In addition, characterization of gels has been performed to determine the optimum sensitivity consistent with the dynamic range of the scanner (119).

X-Ray CT Scanning of Polymer Gels

The formation of polymer chains increases the physical density of the gel, and the resulting change in attenuation coefficient can be measured by measurements of X-ray transmission, such as by computed tomography (120–125). While the change in density is small, it has been shown to vary proportionally with dose (122,126). This

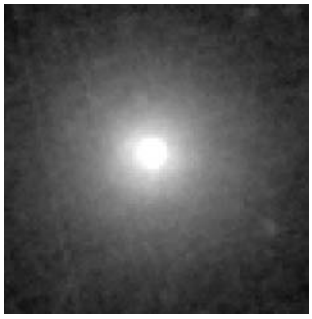


Figure 9. The OCT image of a polymer gel exposed to an ^{192}Ir HDR brachytherapy source. The central region was occupied by the source during irradiation and was replaced with irradiated gel for imaging (117).

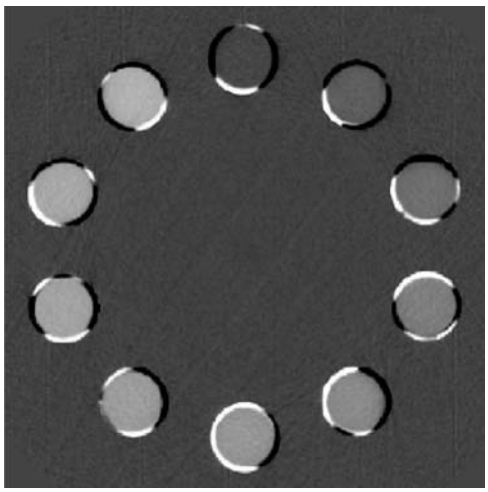


Figure 10. The CT image of several vials of polymer gel irradiated to different doses. (Reproduced with permission from Ref. 128.)

change in density leads to a small change in CT number when irradiated gels are examined with CT. Recent data show that this change can be as much as 0.2 $kg \cdot m^3 \cdot Gy^{-1}$ (127). An image of tubes of gel irradiated to different doses appears in Fig. 10. Methods for improving the quality of X-ray imaging have been developed, and include the acquisition of multiple images, background subtraction, and filtering (126,127).

Ultrasound Imaging of Polymer Gels

Polymerization leads to changes in elasticity of the medium, and the corresponding changes in ultrasound absorption can be exploited (129–132). Ultrasound has been used to evaluate changes in density and elastic constant of a number of materials. Several different ultrasonic parameters can be measured and these can be used to characterize materials. The most commonly measured parameters attenuation and reflection coefficients, and the speed of propagation. A pulse-echo technique using one probe or a transmission technique using two probes is used to measure these parameters. These parameters can be related to structural properties of the sample including bulk density, elastic constants as well as sample inhomogeneities.

Vibrational Spectroscopic Imaging of Polymer Gels

Finally, vibrational spectroscopy can be used to demonstrate the conversion of monomers to polymer chains (133–136). Fourier transform (FT)–Raman vibration spectroscopy of polymer gel dosimeters has been investigated as a means by which the fundamental structure and properties of the dosimeters might be better understood. Raman spectroscopy has also been used to investigate the track structures of proton beams in polymer gel dosimeters (137). This study illustrated the difficulty in using polymer gel dosimeters to extract quantitative dose maps when exposed to proton radiation. Further studies will be required to determine whether Raman microscopy can be used routinely in the evaluation of polymer gel dosimeters.

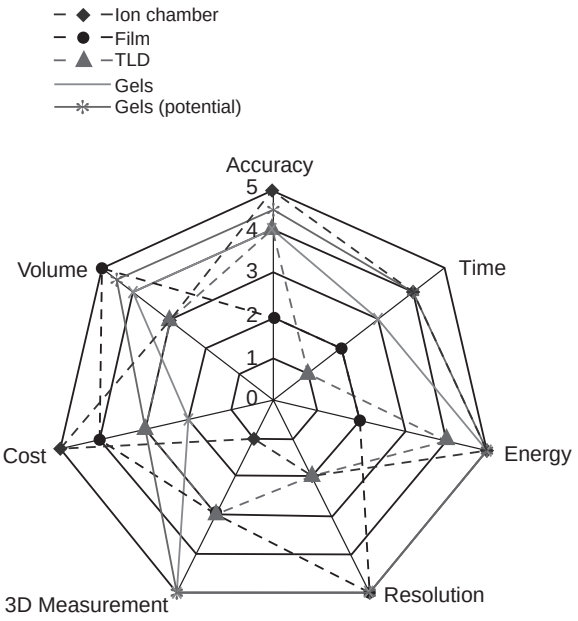


Figure 11. A spider plot, illustrating the capabilities of several common dosimetry systems, as well as gels, and the potential capabilities of gels. (Redrawn with permission from Ref. 110.)

CHARACTERISTICS OF GEL DOSIMETERS

Gel dosimeters have a number of characteristics that make them attractive for radiation dosimetry (138). A novel comparison of gel dosimeters with conventional dosimetry systems has been presented in the form of a spider plot (see Fig. 11, Ref. 110). This graphical presentation illustrates the relative performance of dosimeters, such as ion chambers, film, TLDs and gels by considering such parameters as accuracy, volume measured, cost, three-dimensionality, resolution, energy dependence, and time required for the measurement. Oldham has shown that gels compare favorably with the other detectors in most characteristics, including their relative accuracy, volumetric nature, inherent three-dimensionality, high resolution and lack of energy dependence over much of the important energy range (110). Methods for characterizing the response of gels have been found, and in particular, a technique for characterizing the dose resolution has been described (89,139,140).

However, today gels are still time-consuming and relatively expensive. Several dosimetric aspects have not yet been realized, including the absolute accuracy of measurement, and the ability to render a 3D dose distribution as opposed to multiple planes of data, although progress is being made rapidly on both aspects. In addition, the issues of cost and time required are being addressed. The availability of optical CT scanning and other imaging techniques are likely to drive down the cost of gel analysis, and improve the penetration of this modality into the clinic. At the same time, newer optical CT scanners equipped with more powerful computers are faster and can perform comprehensive imaging of gels in the time previously required for a single slice.

APPLICATIONS OF GEL DOSIMETRY

Potential applications of gel dosimetry have been summarized on several occasions in the recent past (97,138,141–143) although the field is developing rapidly. Today it is considered by many that gel dosimetry has useful characteristics that can facilitate radiation therapy dosimetry, especially in situations that are not handled well by conventional dosimeters. These characteristics include the ability to measure complex 3D dose distributions; to integrate dose accurately without dependence on dose rate, at least over a fairly wide range; tissue-equivalence; high spatial resolution; and lack of energy dependence over most of the kilovoltage and megavoltage range. With most gels, data are stored permanently, making gels suitable for performance of dosimetry at remote locations (144). They also are relatively safe to manufacture and handle, although some components such as acrylamide are toxic and must be handled with appropriate protection until mixed.

Demonstrated applications of gel dosimetry to date include basic dosimetry (depth dose, penumbra, wedge profiles) in photon, electron, neutron, and proton beams; dose distributions from imaging procedures; conformal therapy, stereotactic radiosurgery, and intensity-modulated radiation therapy (IMRT); dose distributions around brachytherapy sources (low and high dose rate, and intravascular sources); internal dosimetry (^{131}iodine doses); and evaluation of tissue heterogeneities. The advances made recently in these areas will be discussed.

Basic Dosimetry

Gel dosimeters have the capability to record and display the dose distribution throughout a 3D volume. This ability affords advantages over conventional dosimeters, even for basic dosimetry parameters such as percent depth dose in photon and electron beams (54,92,145). Gel dosimetry has been shown to be useful to validate simple multiple-field arrangements (146) and more complex anatomical situations including tangential breast treatment (147,148), conformal therapy (149) and scalp treatment with electron beams (150). Dynamic functions, such as a programmable wedge filter are difficult to measure with ionization chambers or diodes, and film is often used to provide data in a single plane. Gels have proven useful for capturing the dose distributions from programmable wedge filters, and allow distributions in multiple planes to be demonstrated from a single exposure (151).

Dose from Imaging Procedures

More recently, the use of gels to demonstrate dose distributions from imaging procedures has been explored (152,153). In a novel experiment, a high sensitivity gel was used to determine the dose from CT imaging. The benefit of this measurement is that the dose distribution throughout a patient volume can be estimated without requiring the use of numerous point dosimeters (e.g., TLD) and without averaging the dose along a line or throughout a volume (e.g., a pencil ionization chamber). These benefits may be most apparent in evaluating the dose distribution from helical CT scanners.

Figure 12. A BANG gel irradiated with a highly conformal dose distribution produced by a Gammaknife treatment unit. The distribution can be appreciated qualitatively without the need of imaging systems or processing. (Photograph by the author. See also Ref. 161.)

Evaluation of Conformal Dose Distributions

Stereotactic Radiosurgery. Gels have been used to demonstrate the dose distributions from stereotactic treatments both from dedicated multisource cobalt units and from linear accelerators (154–161). A clear benefit of gel dosimeters is that they can display a dose distribution, especially a highly conformal one as is produced by stereotactic radiosurgery techniques, so that it can be appreciated qualitatively in three dimensions without need of imaging systems or processing (see Fig. 12, Ref. 161).

In one series of measurements, gels were prepared in glass flasks chosen for their size and shape, which was comparable to that of a human head. Additional polymer gel material from the same batch was prepared in glass test tubes, for irradiation to selected doses, to generate a dose-response curve. The gels were prepared in Guilford, CT, and were shipped to Lexington, Kentucky for irradiation and analysis (161).

A gel prepared in a 16 cm diameter flask was fitted with a radiosurgical head frame (Leksell, Elekta Corporation, Atlanta, GA), as shown in Fig. 13. This flask was also equipped with a glass rod extending to near the center of the flask, to be used as a target. The MR images were obtained and were transferred to a Gammaknife treatment planning computer (Gammaplan; Elekta Corporation), where a complicated dose distribution was planned using multiple target points. Once the plan was completed, the coordinates of the individual target points were determined, and the gel was moved to the Gammaknife irradiation unit. Treatments were delivered to each of the target points, in accordance with the treatment plan. A dose of 10 Gy was delivered to the 100% isodose line.

Dosimetric imaging of the flask and test tubes containing gel was performed between 25 and 36 h after irradiation. The flask was placed in the head coil of the imager and the test tubes irradiated for calibration purposes were placed around the flask. The images were transferred via network to a Macintosh computer, and the DoseMap program was used to compute the maps of transverse relaxation rate (R_2).

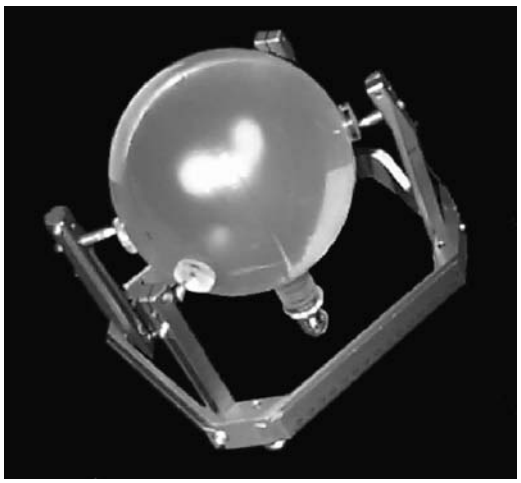


Figure 13. Photograph of a glass flask filled with the BANG Polymer Gel dosimeter. A glass rod was inserted into the gel to provide a target around which to localize the dose distribution. The flask was fitted with a Leksell stereotactic head frame. The gel is shown as it appeared following irradiation. (Photograph by the author).

A dose-response calibration curve was obtained as described earlier. Images of the gel-filled test tubes were obtained, and R_2 determined as a function of dose.

The calibration curve was then applied to R_2 maps of the flask irradiated with the Gammaknife unit. The result yielded an image of the dose distribution, as shown in Fig. 14a and 14b. As all scans were performed with the head ring and localizer box in place, the coordinates of the image plane could be determined. These image planes were located 1 mm from each of the corresponding treatment plans shown in Fig. 14a and 14b. Finally, isodose curves were drawn (by the DoseMap program) by interpolating within the measured dose distribution.

The measured dose distributions were compared with the treatment plans prepared prior to irradiation by superimposing the two data sets. The superimposed data are shown in Fig. 15a and 15b. The calculated and measured dose distributions were registered by aligning the point representing the tip of the glass rod.

The measured dose distributions compare favorably with the calculated dose distributions. In fact, the dose

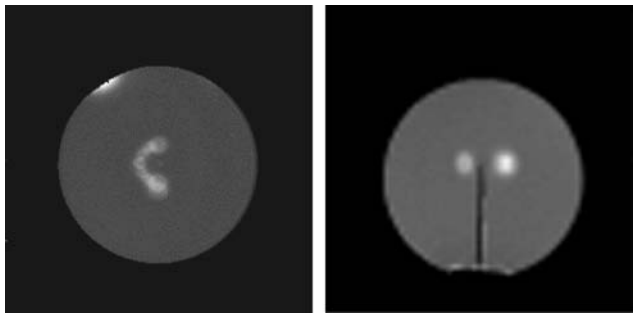


Figure 14. The R_2 maps obtained from the irradiated gel (a) Distribution in the axial plane. (b) Distribution in the sagittal plane. (Reproduced with permission from Ref. 161.)

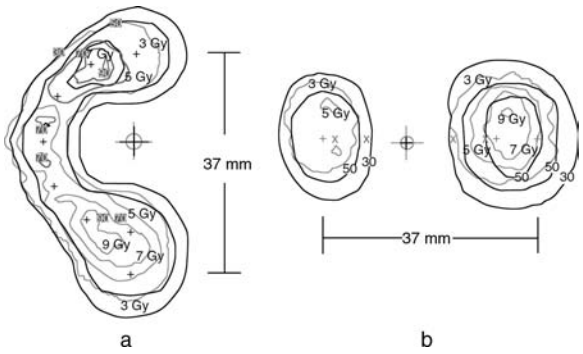


Figure 15. Composite figures showing both the treatment plan prepared using a Gammaplan treatment planning computer (drawn in black, labeled in percent of maximum dose) and isodose curves measured by the technique described in the text (drawn in gray, labeled in Gy). (a) The distribution in the axial plane containing the 8 isocenters. (b) The distribution in a perpendicular sagittal plane. (Reproduced with permission from Ref. 161.

map taken in the plane of the target points (Fig. 14a) indicates regions of overlap not demonstrated by the treatment planning system. As shown in Fig. 15a and 15b, the measured isodose lines conform in shape quite well with the calculated data, but seem to show a shift away from the glass target rod. The dose images were obtained in planes that were shifted 1 mm from the planes of dose calculation, and this shift might account partially for the difference in size and shape of the isodose curves. However, Fig. 15a shows a shift in the lateral (X) direction away from the glass target rod, which cannot be explained by a difference in the axial (Z) coordinates of the planes of calculation and measurement. Instead, it appears more likely that the dose distribution was placed ~ 1 mm further from the glass target rod than intended.

Evaluation of Repeat-Fixation Stereotactic Radiotherapy. In recent years, fractionated stereotactic radiation therapy has been seen as a desirable method of delivering high dose radiation therapy to malignancies of the brain. Techniques developed for immobilizing the patient have also been applied more recently to intensity-modulated radiation therapy, in which conformal dose distributions are delivered through multiple fractions to one or more target volumes. In both techniques, reproducible positioning of the patient is critical, to ensure that the target volume receives the intended dose, and normal tissues are spared to the extent determined by treatment planning techniques. The BANG gel dosimeter has been used in a fractionated regimen to demonstrate the reproducibility of multiple setups under stereotactic position methods (158).

Intensity-Modulated Radiation Therapy (IMRT). Gels dosimeters have proven themselves to be valuable for evaluating and confirming IMRT dose distributions (146,162–169). Most investigations have been conducted in simple geometric phantoms (Fig. 16), but others have employed anthropomorphic phantoms in arrangements that allowed direct comparison with measurements using other techniques such as film and TLD (163,165,166).

Figure 16. A cylindrical flask containing a normoxic gel shortly after irradiation with an IMRT treatment. The dose distribution is clearly visible, demonstrating the change in optical density with dose. (Reproduced from Ref. 167, with permission.)

Beach developed a gel insert for an existing anthropomorphic phantom that had been developed with film and TLD dosimeters (170). The phantom design revision included converting the existing imaging/dosimetry insert from a block-style design to a cylindrical design (Fig. 17). This insert contained embedded structures that simulated a primary and secondary target volume as well as an organ at risk (OAR). An additional insert was then constructed to house the polymer gel dosimeter. This insert was specially designed using Barex plastic. Both the imaging insert and the gel insert had an image registration system incorporated into their construction.

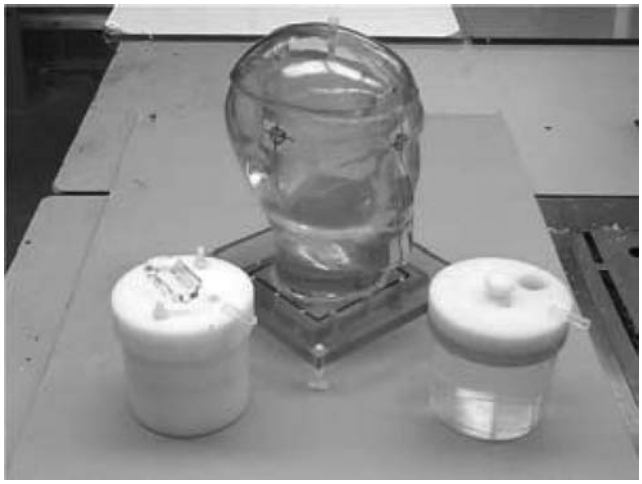


Figure 17. An anthropomorphic head-and-neck phantom developed by the Radiological Physics Center (170) showing the modifications made to accommodate a gel dosimeter.

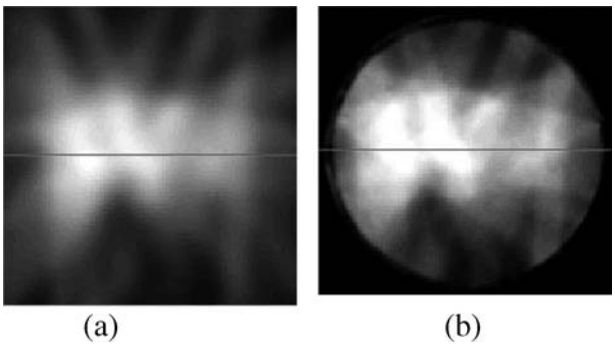


Figure 18. (a) A calculated dose distribution for an IMRT treatment, shown in a gray-scale format. (b) The measured dose distribution obtained from optical CT of a polymer gel, following irradiation with the treatment plan shown in (a). (From Ref. 165, with permission.)

X-ray CT images were obtained of the phantom with the imaging insert in place, and an IMRT treatment plan was developed. The phantom was then taken to the linear accelerator, the imaging insert was replaced with the gel insert, and the IMRT treatment was delivered.

A commercially available optical computed tomography (OCT) scanner (107) was commissioned for this project and future work with polymer gel dosimetry. The OCT scanner was used to image polymer gels before and after being irradiated. The preirradiation images were subtracted from the postirradiation images using a pixel-by-pixel subtraction method. The resultant images had net OD values that were directly proportional to the dose received by each given pixel. A comparison of the calculated dose distribution and the measured distribution is shown in Fig. 18.

Repeated measurements showed that a polymer gel imaged with optical CT was reproducible to within 1% (171). Repeated OCT imaging was shown to be consistent to within 1%. However, the results also showed that the techniques used to calibrate the gel (irradiation of a similar gel container with small-diameter beams delivering doses spanning the expected range) did not provide absolute dose measurements offering better agreement than 10% with the calculated data.

Duthoy compared the dose distribution measured with gels to the calculated distribution, for complex intensity-modulated arc therapy (IMAT) treatments in the abdomen (172). Vergote also examined IMAT with gels and observed a reproducible difference between calculations and measurements in low dose regions near steep dose gradients; a phenomenon also observed by Cadman et. al. and attributed to the failure of treatment planning systems to model the transmission of radiation through the rounded ends of multileaf collimator leaves (169,173).

Brachytherapy. Determining dose distributions and confirming the results of planning for brachytherapy treatment is historically difficult. No suitable methods of dosimetry have existed in the past to enable measurement and display of these 3D and complex distributions. Measurements around single sources have been possible only in a point-by-point fashion, such as with small ionization chambers or with thermoluminescence dosimeters (TLDs), (174) or in planar fashion with film (175). These methods are quite unsatisfactory for anything other than distributions around single sources, or very simple source arrangements. In contrast, the BANG polymer gel dosimetry system has the capability to measure and display complex dose distributions from complicated source arrangements. It is necessary to immerse the applicator containing the sources into the gel, or arrange for its introduction into a catheter already placed in the gel.

The ability of gels to record and display dose distributions around a high dose rate (HDR) source was first demonstrated over a decade ago (92,176,177). Maryanski et al. showed the dose distribution around a single catheter into which a high dose rate (HDR) remote afterloader source had been positioned (178). The HDR unit was programmed to dwell the source at several locations in the catheter, to deliver an elliptical dose distribution. After irradiation, the gel was imaged with MR, and a map of the dose distribution was computed. The map is shown in Fig. 19, where the color intensity is proportional to dose. Isodose lines, determined from the dose map data, are superimposed on the intensity map. Points at which the dose was computed by the treatment planning system also are shown. Excellent agreement between the position of the calculated dose points and the corresponding measured isodose lines indicates the agreement between doses measured by the gel and computed by the treatment planning system.

More recently, measurements have been made in close proximity to HDR ^{192}Ir sources (117,179) (see also Fig. 9). These measurements have shown that complications occur when measurements are made in the steep dose gradients close to an HDR source. Polymerization of the gel causes an increase in the gel density and a corresponding decrease in the volume filled by the gel. The change in density causes shrinkage of the gel in the vicinity of the source, distorting the resulting measured distribution. Changes to the composition of the gel to increase the concentration of gelatin

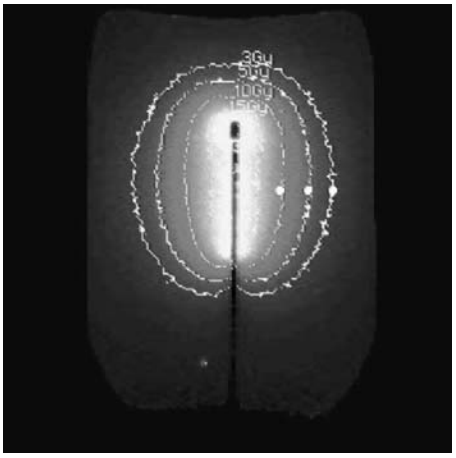


Figure 19. Use of the BANG gel to measure the dose distribution around an HDR source. The source was positioned in a catheter implanted in a BANG polymer gel. The figure illustrates a comparison between the dose distribution determined from a MRI image of the gel and the calculated dose distribution. (From Ref. 178.)

can mitigate the amount and effects of the density changes. Furthermore, there are suggestions that the high dose rates found near brachytherapy sources, particularly those of HDR afterloaders, can introduce temperature gradients that influences the polymerization of acrylamide monomer gels (87,93,180,181).

Efforts also have been made to characterize low dose rate (LDR) sources, such as prostate seeds (182–184), eye plaques (185), ^{137}Cs afterloading sources (186,187) and intravascular sources (188). Studies have indicated that the diffusion of monomers (or ferrous and ferric ions in Fricke gels) across steep dose gradients can introduce errors in measurement (92,189). As the use of gels to measure dose distributions from LDR sources requires long exposure times, diffusion of monomers or ions could introduce significant errors, and gels exhibiting high diffusion rates should be avoided for these measurements.

A further problem with gel dosimetry for LDR brachytherapy has been demonstrated by recent studies indicating energy dependence at lower energies. Data show that a polymer gel dosimeter under responds to radiation in the 20–60 keV range (190). Others have shown differences in gel response from one formulation to another, and suggest that the MAGAT gel is most water-equivalent over a wide range of energies (191). Changes in mass attenuation coefficient of polymer gels during irradiation can also introduce errors in the dose distributions measured around low energy sources.

Internal Dosimetry

Gel dosimetry has shown promise in the determination of dose distributions from administrations of unsealed radioactive sources (192). The authors embedded a vial of ^{131}I into a flask of polymer gel and observed a distance-dependent change in the T_2 signal. They also mixed ^{131}I into the gel and demonstrated a change in T_2 signal that was dependent on distance from the concentration of activity. No more recent investigations have been located.

Measurement of Neutron Dose Distributions

Some developments have been reported in characterizing fast and epithermal neutron beams with gel dosimetry (193–195). Thin layers of Fricke-xylenol orange gels have been irradiated in phantoms composed of insensitive gel. Adding ^{10}B or other nuclides with large cross-sections has increased the sensitivity of the gel dosimeter to neutrons. This technique has been used to determine the profiles of neutron beams used for boron neutron capture therapy. Some benefits of the use of gel dosimetry are the tissue-equivalence of the dosimeter to these energies, and the ability to separate the components of dose.

Measurement of Particle Dose Distributions

Several investigators have demonstrated the use of polymer gel dosimeters to record the dose distributions produced by proton beams (88,137,196–198). However, several authors have noticed disagreements between measurements with gels and conventional dosimeters such as diodes in the peak region of the distribution. Gustavsson has suggested that the response of gels, as they are based on the formation of free radicals, is dependent on the LET of the radiation (197,198). As the LET of the beam increases in the peak region, the local ionization density increases. As the distance between the radicals formed in the gel decreases, the likelihood of recombination of radicals increases. A decrease in the production of radicals with increasing LET has been described previously (199). Consequently, significant differences appear between depth dose measurements with gels and those with detectors such as diodes (see Fig. 20, Ref. 198).

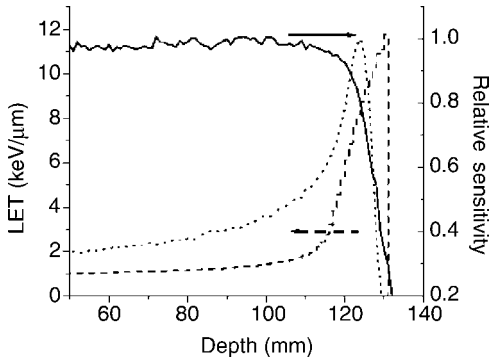


Figure 20. The variation in LET as a function of depth for a monoenergetic proton beam (dashed curve, left-hand scale) and the measured relative sensitivity for the gel dosimeter (full curve, right-hand scale). Also shown is the depth dose curve for the proton beam (dotted curve), normalized to 100% at the Bragg peak. (Reproduced with permission from Ref. 198.

Jirasek et al. performed track energy-deposition calculations and raman spectroscopy and reported agreement between these techniques and gel measurements (137). Their conclusion also was that the high density of delta-ray interactions close to the track of a proton resulted in high doses being delivered to the gel. These doses saturated the response of the gel by consuming the available monomer. This effect was greater near the end of the proton range, consistent with the results of other authors.

Gels have been used also to demonstrate the dose distribution produced by ^{12}C ions (200). Similar effects associated with decreased radical formation at high LET were observed in the carbon beam.

Evaluation of Tissue Heterogeneities

A valuable feature of gel dosimeters is that they are very nearly tissue-equivalent, particularly at photon beam energies above $\sim$ 100 kV. Previous investigations have shown that the BANG gel, the MAGIC and MAGAS normoxic gels, as well as gels based on Fricke or vinyl solutions have electron densities within 1% of soft tissue, and effective atomic numbers in the range of 7.4 (190). However, several investigators have attempted to measure the effects of nonunit density tissues on external beam dose distributions. Early measurements were performed to estimate the dose distribution behind high atomic number heterogeneities, to simulate the presence of bone (201–204). More recently, measurements have been made behind or adjacent to cavities filled with air or with lung-equivalent plastic (168). To attempt a measurement

in lung-equivalent gel, Olberg produced a foam of gel with the approximate density of lung tissue (205). Other investigators have evaluated the promise of gel dosimeters to simulate lung tissue, by introducing polystyrene foam beads into a gel mixture (206). While these measurements showed promise, there were several sources of error. First, the introduction of air, or air-containing polystyrene beads introduced the possibility of oxygen contamination. Purging the polystyrene beads with nitrogen, or using nitrogen rather than air to foam the gel addressed this problem. The introduction of air or polystyrene eliminated the possibility of evaluating the measured dose distribution by optical scanning, and MR imaging must be used. The presence of air may lead to partial volume imaging effects that could introduce errors into the measurement.

COMPLICATIONS TO BE ADDRESSED

As was suggested earlier in this article, there are a number of complications associated with gel dosimetry that remain to be addressed, and that are inhibiting the routine use of gels in the clinic. Some of these are listed below, with short descriptions of the causes of the problems, and possibilities for correcting them.

Imaging Artifacts

This article has discussed several methods of generating images of dose distributions using gels. Principal methods are MRI, OCT, and X-ray CT. Each of these imaging methods is prone to imaging artifacts, although the type of artifact and its causes are different with the different modalities. In MRI, for example, susceptibility artifacts can result from variations in the conductivity of the volume being imaged, and interference is likely when multiplanar imaging of adjacent planes is attempted. The presence of air or low-density structures can lead to partial volume effects or susceptibility artifacts.

In OCT, any structure that blocks the light beam is likely to cause a streak artifact, similar to those produced by high densities in X-ray CT images. In addition, the refraction of the light at interfaces between the gel and other materials can cause ring artifacts or distortion of the image. The artifacts found in OCT images have been described (110). An example of the artifact caused by high optical densities is shown in Fig. 21.

When X-ray CT is used, artifacts can result from the low signal to noise ratio that occurs because of the very small density differences present in the gel. These artifacts have been investigated in some detail previously (121).

Temperature Dependence

The existence of a dependence on temperature during irradiation of polymer gels was not recognized immediately, but it has since been shown that this dependence exists. Furthermore, the temperature dependence can be more pronounced for some polymer gel formulations than others. The polymerization that occurs as a result of irradiation of the gel is exothermic, and consequently can lead to a temperature rise that influences further polymerization of gel in response to continuing exposure. In extreme cases, this temperature rise can exceed several degrees Celcius (207).

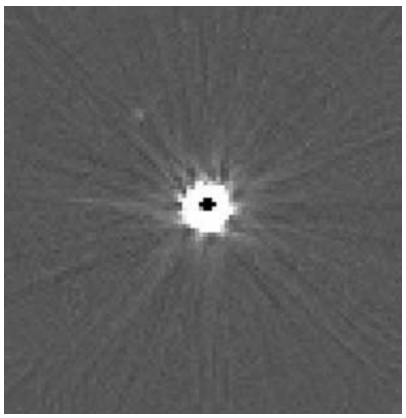


Figure 21. An optical CT scan of a normoxic gel irradiated with a low dose rate ^{125}I brachytherapy source. The high optical densities close to the source completely attenuate the laser, resulting in a streak artifact.

Oxygen Sensitivity

The sensitivity of polymer gels to oxygen has been discussed extensively, and several investigators have responded by developing gels that contain oxygen scavengers, such as the MAGIC gel (86). The oxygen scavenger removes oxygen present in the gel at the time of manufacture, even if this is done in normoxic conditions. It can remove additional small amounts of oxygen, but ultimately will be overwhelmed if exposure to normal atmosphere is ongoing. While this problem has been addressed, it still creates minor inconvenience that might limit the successful introduction of gels into routine clinical use. The characteristics of several normoxic gel dosimeters have been investigated in detail (90,208).

Tissue Equivalence and Energy Dependence

Gels, both Fricke and polymer types, compete well when compared to other dosimeters in terms of their tissue equivalence and energy dependence. In comparison to thermoluminescence dosimeters (TLD), radiographic film, and even ionization chambers, for example, gels are considerably less energy dependent and are much more tissue equivalent (209). However, under extreme conditions of photon energies below 60 keV, and LET values greater than ~ 2.5 keV$\cdot\mu^{-1}$, gels show a dependence that remains to be fully characterized (190).

Simulation of Nonunit Densities

The benefits of gels discussed in the previous paragraph lead to the inability of gels to easily simulate nonunit density tissues. To date, limited efforts have been described to create low density gel mixtures, to simulate lung tissue. No attempts have been described to date to create high density mixtures and are unlikely to be with today's emphasis on the use of gelatin-based formulations.

Diffusion of Monomer in Steep Gradients

The diffusion of monomer, and the shrinkage of gel proportional to the creation of long polymer chains, can be addressed through the development of better gel mixtures. Avoiding small monomers such as acrylamide can reduce the rate of diffusion in regions of steep dose gradient, such as the penumbra of radiation beams (210). Employing different concentrations of gelatin might reduce or eliminate problems associated with shrinkage of gels in high dose regions. Some normoxic gels may demonstrate decreased diffusion in regions of steep dose gradient (90,208).

SUMMARY AND CONCLUSIONS

The importance of quality assurance in radiation therapy is well documented. High quality, safe, and effective radiation therapy is dependent upon the proper operation of equipment, the accuracy of alignment devices, and the dependability of dosimetry procedures. The accurate delivery of the prescribed dose depends on procedures involving dosimetry systems. Properly functioning dosimeters are necessary to assure that the equipment is properly calibrated and that treatment planning procedures are conducted correctly.

A wide variety of dosimetry systems are available to medical physicists today. Choosing the correct dosimetry system for any given application requires an understanding of the operation of the device and its appropriateness for the intended circumstances. This presentation has reviewed a number of dosimetry systems presently available, their design and operation, and some of the uses for which they are valuable.

Gel dosimetry offers the promise of accurate and convenient dosimetry under a variety of circumstances. In most of the examples discussed above, gel dosimeters offer a number of advantages over conventional dosimeters. Chief among these is the ability to measure a complex dose distribution throughout a volume with a single radiation exposure. Additional advantages include tissue equivalence, high spatial accuracy, good dose precision, and reasonable convenience.

However, gel dosimetry continues to experience little acceptance in the clinic, largely because some aspects of promise have not been achieved, and because of a perceived lack of convenience. Members of the radiation physics community are apparently not convinced that the benefits of gels sufficiently outweigh conventional dosimeters such as film and TLD. It is incumbent on those of us working with gels to encourage more widespread use, by taking every opportunity to display the results of measurements with gels.

BIBLIOGRAPHY

1. Cunningham JR. Development of computer algorithms for radiation treatment planning. Int J Radiat Oncol Biol Phys 1989;16:1367.
2. Fischer JJ, Moulder JE. The steepness of the dose-response curve in radiation therapy. Radiology 1975;117:179–184.
3. Hendrickson FR. Precision in radiation oncology. Int J Radiat Oncol Biol Phys 1981;8:311–312.
4. ICRU Rep No. 24: Determination of absorbed dose in a patient irradiated by beams of X or gamma rays in radiotherapy procedures, Washington (DC), 1976, International Commission on Radiation Units and Measurements.
5. ICRU report No. 42, Use of computers in external beam radiotherapy procedure with high energy photons and electrons, Washington (DC), 1988. International Commission on Radiation Units and Measurements.
6. ICRU report No. 50, Prescribing, recording, and reporting photon beam therapy, Washington (DC), 1993, International Commission on Radiation Units and Measurements.
7. Leunens G, et al. Assessment of dose inhomogeneity at target level by in vivo dosimetry; Can the recommended 5% accuracy in the dose delivered to the target volume be fulfilled in daily practice? Radiother Oncol 1992;25:245–250.
8. Ibbott GS, et al. Uncertainty of calibrations at the accredited dosimetry calibration laboratories. Med Phys 1997;24(8):1249–1254.
9. Kartha PKI, Chung-Bin A, Hendrickson FR. Accuracy in clinical dosimetry. Br J Radiol 1973;46:1083–1084.
10. Kartha PKI, et al. Accuracy in patient setup and its consequence in dosimetry. Med Phys 1975;2:331–332.
11. Ahuja SD. Physical and technological aspects of quality assurance in radiation oncology. Radiol Tech 1980;51(6): 759–774.
12. American Association of Physicists in Medicine. Radiation treatment planning dosimetry verification, AAPM Task Group 23 Test Package, 1984. AAPM Rep No. 55, 1995.
13. American College of Radiology: ACR Standard for Radiation Oncology Physics for External Beam Therapy Res. 15-1994, American College of Radiology Standards, adopted 1995 by the American College of Radiology, 1891 Preston White Drive, Reston, VA 22091.
14. American College of Radiology: ACR Standard for Radiation Oncology, Res. 38-1995, American College of Radiology Standards, adopted 1995 by the American College of Radiology, 18941 Preston White Drive, Reston, VA 22091.
15. American College of Radiology, ACR Standards for the performance of Brachytherapy Physics: Manually-Loaded Sources. Res. 25-1995, American College of Radiology Standards, adopted 1995 by the American College of Radiology, 1891 Preston White Drive, Reston, VA 22091.
16. Annett CH. Program for periodic maintenance and calibration of radiation therapy linear accelerators. Appl Radiol 1979;6: 77–80.
17. Earp KA, Gates L. Quality assurance: A model QA program in radiation oncology. Radiol Technol 1990;61(4):297–304.
18. Ibbott GS, et al. Quality Assurance Workshops for Radiation Therapy Technologists. Appl Radiol, March–April 1977.
19. International Electrotechnical Commission Rep No. 976, Medical electrical equipment, Geneva, Switzerland, 1993, Bureau Central de la Commission Electrotechnical Internationale.
20. International Electrotechnical Commission Rep No. 977, Medical electrical equipment: Medical electron accelerators in the range 1 MeV to 50 MeV. Guidelines for functional performance characteristics, Geneva, Switzerland, 1993, Bureau Central de la Commission Electrotechnique Internationale.
21. JCAHO Accreditation Manual for Hospitals, 1995, Oak Brook Terrace (IL): Joint Commission on Accreditation of Healthcare Organizations; 1995.
22. Kutcher GJ, et al. Comprehensive QA for radiation oncology; Report of AAPM Radiation Therapy Committee Task Group 40. Med Phys 1994;21(4):581–618.
23. Van Dyk J, editor. The Modern Technology of Radiation Oncology: A Compendium for Medical Physicists and Radiation Oncologists. Madison (WI): Medical Physics Publishing; 1999.

24. CIRMS 2004 - Council on Ionizing Radiation Measurements and Standards: Fourth Report on Needs in Ionizing Radiation Measurements and Standards, Dec 2004. CIRMS, P. O. Box 1238, Duluth, GA 30096. Available at www.cirms.org. 2004.
25. Rice RK, Hansen JL, Svensson GK, Siddon RL. Measurements of dose distributions in small beams of 6 MV X-rays. Phys Med Biol 1987;32:1087–1099.
26. Alecu R, Alecu M. *In-vivo* rectal dose measurements with diodes to avoid misadministrations during intracavitary high dose rate brachytherapy for carcinoma of the cervix. Med Phys 1999;26(5):768–770.
27. Alecu R, Loomis T, Alecu J, Ochran T. Guidelines on the implentation of diode in vivo dosimetry programs for photon and electron external beam therapy. Med Dosimetry 1999;24(1): 5–12.
28. Mijnheer B, et al. Quality assurance of treatment planning systems: Practical examples for non-IMRT photon beams. 2004 ESTRO, Mounierlaan 83/12 - 1200 Brussels (Belgium)
29. International Atomic Energy Agency Technical Report 430. Commissioning and quality assurance of computerized planning systems for radiation treatment of cancer. Vienna: IAEA; 2004.
30. Karger CP, Heeg P. A system for 3-dimensional dosimetric verification of treatment plans in intensity-modulated radiotherapy with heavy ions. Med Phys Oct 1999;26(10).
31. Knoll GF. Radiation Detection and Measurement, New York: Wiley; 1989, p 216–227.
32. Kirov AS, et al. Towards two-dimensional brachytherapy dosimetry using plastic scintillator: New highly efficient water equivalent plastic scintillator materials. Med Phys August 1999;26(8).
33. Beddar AS, Mackie TR, Attix FH. Water-equivalent plastic scintillation detectors for high energy beam dosimetry: II. Properties and measurements. Phys Med Biol 1992;37: 1901–1913.
34. Olde GL, Brannon E. Three dimensional scintillation dosimeter. Rev Sci Instrum 1959;30:1014–1016.
35. Perera H, et al. Rapid 2-dimensional dose measurement in brachytherapy using plastic scintilaator sheet: Linearity, signal-to-noise ratio, and energy response characteristics. Int J Radiat Oncol Biol Phys 1992;23:1059–1069.
36. Kirov AS, et al. Two-dimensional dosimeter using plastic scintillator: Localization of the scintillation process. Med Phys 1997;24:1005.
37. Kirov AS, et al. New highly efficient water equivalent plastic scintillator materials for radiation dosimetry. Med Phys 1998;25:A153.
38. Yeo IJ, Wang CKC, Burch SE. A new approach to film dosimetry for high-energy photon beams using organic plastic scintillators. Phys Med Biol 1999;44:3055–3069.
39. Yeo IJ, Wang C, Burch SE. A scintillation method for improving X-ray film dosimetry in photon radiotherapy (abstract). Med Phys 1996;23:1161.
40. Burch SE, Yeo IJ, Wang CK. A new approach to film dosimetry for high-energy photon beams: lateral scatter filtering. Med Phys 1997;24:775–783.
41. Seamon JM, Ibbott GS. Errors introduced in electron beam film dosimetry. Med Dosimet 1987;12(2):35–37.
42. Meigooni AS, Sanders MI, Ibbott GS, Szeglin SR. Dosimetric characteristics of an improved radiochromic film. Med Phys 1996;23(11):1883–1888.
43. McLaughlin WL, Miller A, Fiban S, Pejtersen K. Radiochromic plastic film for accurate measurement of radiation absorbed dose and dose distributions. Radiat Phys Chem 1977;9:737–474.
44. Cameron JR, Suntharalingam N, Kenney GN. Thermoluminescent Dosimetry. Madison (WI): The University of Wisconsin Press; 1968.
45. Conner WG, et al. Patient repositioning and motion detection using a video cancellation system. Int J Radiat Oncol Biol Phys 1975;1:147–153.
46. Rogus RD, Stern RL, Kubo HD. Accuracy of a photogrammetry-based patient positioning and monitoring system for radiation therapy. Med Phys 1999;26(5):721–728.
47. Curtin-Savard AJ, Podgorsak EB. Verification of segmented beam delivery using a commercial electronic portal imaging device. Med Phys 1999;26(5):737–742.
48. Petrascu O, et al. Automatic on-line electronic portal image analysis with a wavelet-based edge detector. Med Phys 2000; 27(2):321–329.
49. Gilhuijs KGA, Van Herk M. Automatic on-line inspection of patient setup in radiation therapy using digital portal images. Med Phys 1993;20:667–677.
50. Van Herk M, Bel A, Gilhuijs KGA, Vijlbrief RE. A comprehensive system for the analysis of portal images. Radiother Oncol 1993;29:221–229.
51. Fritsch D, et al. Core-based portal image registration for automatic radiotherapy treatment verification. Int J Radiat Oncol Biol Phys 1995;33:1287–1300.
52. Dong L, Boyer AL. An image correlation procedure for digitally reconstructed radiographs and electronic portal images. Int J Radiat Oncol Biol Phys 1995;33:1053–1060.
53. Day MJ, Stein G. Chemical effects of ionizing radiation in some gels. Nature(London) 1950;166:146–147.
54. Andrews HL, Murphy RE, LeBrun EJ. Gel dosimeter for depth dose measurements. Rev Sci Instr 1957;28:329–332.
55. Gore JC, Kang YS, Schulz RJ. Measurement of radiation dose distributions by nuclear magnetic resonance (NMR) imaging. Phys Med Biol 1984;29:1189–1197.
56. Schreiner LJ. Gel dosimetry: Motivation and historical foundation. In DOSGEL 1999: Proc 1st Int Workshop Radiation Therapy Gel Dosimetry (Canadian Organization of Medical Physicists, Edmonton) Schreiner LJ, Audet C, editors.
57. DOSGEL 1999. Proc 1st Int Workshop Radiation Therapy Gel Dosimetry (Lexington, KY). In: Schreiner L J, Audet C, editors. Ottawa, Ontario, Canada: Canadian Organization of Medical Physicists; 1999.
58. DOSGEL 2001. Proc 2nd Int Conf Radiotherapy Gel Dosimetry. In: Baldock C, De Deene Y editors. Brisbane, Queensland, Australia: Queensland University of Technology; 2001.
59. DOSGEL 2001. Proc 3rd Int Conf Radiotherapy Gel Dosimetry. In: Baldock C, De Deene Y, editors. Gent University, Gent, Belgium. J Phys Conf Ser 2004; **3**.
60. Fricke H, Morse S. The chemical action of Roetgen rays on dilute ferrosulphate solutions as a measure of dose. Am J Roent Radium Ther Nul Med 1927;18:430–432.
61. Fricke H, Hart EJ. Chemical Dosimetry, Vol. 2. In: Attix FH, Roesch WC, editors. Radiation Dosimetry. New York: Academic Press; 1966.
62. Appleby A, Christman EA, Leghrouz A. Imaging of spatial radiation dose distribution in agarose gels using magnetic resonance. Med Phys 1987;14:382–384.
63. Olsson LE, Petersson S, Ahlgren L, Mattsson S. Ferrous sulphate gels for determination of absorbed dose distributions using MRI technique: basic studies. Phys Med Biol 1989;34: 43–52.
64. Schulz RJ, deGuzman AF, Nguyen DB, Gore JC. Dose-response curves for Fricke-infused agarose gels as obtained by nuclear magnetic resonance. Phys Med Biol 1990;35:1611–1622.
65. Olsson LE, Appleby A, Sommer JA. A new dosimeter based on ferrous sulphate solution and agarose gel. Appl Radiat Isot 1991;42:1081.
66. Olsson LE, Westrin BA, Fransson A, Nordell B. Diffusion of ferric ions in agarose dosimeter gel. Phys Med Biol 1992a; 37:2243–2252.

67. Baldock C, Harris PJ, Piercy AR, Healy B. Experimental determination of the diffusion coefficient in two-dimensions in ferrous sulphate gels using the finite element method. Aust Phys Eng Sci Med 2001; 24:19–30.
68. Balcolm BJ, et al. Diffusion in Fe(II/III) radiation dosimetry gels measured by MRI. Phys Med Biol 1995;40:1665–1676.
69. Harris PJ, Piercy A, Baldock C. A method for determining the diffusion coefficient in Fe(II/III) radiation dosimetry gels using finite elements. Phys Med Biol 1996;41:1745–1753. Baldock C, et al. Temperature dependence of diffusion in Fricke gel MRI dosimetry. Med Phys 1995;22:1540.
70. Schreiner LJ. Fricke gel dosimetry. Proc 2nd Int Conf Gel Dosimetry, DOSGEL 2001, 15–22.
71. Chu KC, et al. Polyvinyl alcohol Fricke hydrogel and cryogel: two new gel dosimetry systems with low Fe^{3+} diffusion. Phys Med Biol 2000;45:955–969.
72. Kelly RU, Jordan KJ, Battista J. Optical CT reconstruction of 3D dose distributions using the ferrous benzoic-xylenol (FBX) gel dosimeter. Med Phys 1998;25:1741–1750.
73. Chu KC, et al. A Novel Fricke Dosimeter using PVA Cryogel, DOSGEL'99, Proceedings of the 1st International Workshop on Radiation Therapy Gel Dosimetry. In: Schreiner LJ, Audet C, editors. Ottawa, Ontario, Canada: Canadian Organization of Medical Physicists, 1999.
74. Chu K, Rutt BK. Polyvinyl alcohol cryogel: an ideal phantom material for MR studies of arterial flow and elasticity. Magn Reson Med 1997;37:314–319.
75. Gambarini G, et al. Dose-response curve slope improvement and result reproductibility of ferrous-sulphate-doped gels analyzed by NMR imaging. Phys Med Biol 1994;39:703–717.
76. Rae WID, et al. Chelator effect on ion diffusion in ferrous-sulfate-doped gelatin gel dosimeters as analyzed by MRI. Med Phys 1996;23:15–23.
77. Kron T, Jonas D, Pope JM. Fast T-1 imaging of dual gel samples for diffusion measurements in NMR dosimetry gels. Magn Reson Imaging 1997;15:211–221.
78. Pedersen TV, Olsen DR, Skretting A. Measurement of the ferric diffusion coefficient in agarose and gelating gels by utilizatino of the evolution of a radiation enduced edge as reflected in relaxation rate images. Phys Med Biol 1997;42: 1575–1585.
79. Scherer J, et al. 3D Fricke gel dosimetry in antropomorphic phantoms. Proc 1st Int Workshop Radiation Therapy Gel Dosimetry, Lexington(ky) July 21–23, 1999; p 211–213.
80. Maryanski MJ, Gore JC, Schulz RJ. 3D Radiation Dosimetry by MRI: Solvent Proton Relaxation Enhancement by Radiation-Controlled Polymerization and Crosslinking in Gels. 11th Annu Sci Meet Soc Magnetic Resonance in Medicine, Berlin, (Poster No. 1325). 1992.
81. Maryanski MJ, Gore JC, Kennan RP, Schulz RJ. NMR relaxation enhancement in gels polymerized and cross-linked by ionizing radiation: a new approach to 3D dosimetry by MRI. Magn Reson Imaging 1993;11:253–258.
82. Maryanski MJ, et al. Radiation therapy dosimetry using magnetic resonance imaging of polymer gels. Med Phys 1996;23:699–705.
83. Baldock C, et al. Experimental procedure for the manufacture of polyacrylamide gel (PAG) for magnetic resonance imaging (MRI) radiation dosimetry. Phys Med Biol 1998; 43:695–702.
84. Maryanski MJ. Radiation-sensitive polymer-gels: properties and manufacturing. Proc 1st Int Conf Gel Dosimetry, DOSGEL 1999. Queens University Printing Service, Kingston, Ontario, Canada. 63–73.
85. Maryanski MJ, Gore JC, Schulz RJ. Three-Dimensional Detection, Dosimetry and Imaging of an Energy Field by Formation of a Polymer in a Gel. US Patent 5,321,357.
86. Fong PM, Keil DC, Does MD, Gore JC. Polymer gels for magnetic resonance imaging of radiation dose distributions at normal room atmosphere. Phys Med Biol 2001;46:3105–3113.
87. De Deene Y, et al. A basic study of some normoxic polymer gel dosimeters. Phys Med Biol 2002;47:3441–3463.
88. Maryanski MJ, et al. Three dimensional dose distributions for 160 MeV protons using MRI of the tissue-equivalent BANG Polymer-gel dosimeter. Particles (PTCOG Newsletter) Jan 10–11 1994a.
89. Baldock C, et al. Dose resolution in radiotherapy polymer gel dosimetry: effect of echo spacing in MRI pulse sequence. Phys Med Biol 2001;46:449–460.
90. De Deene Y, Baldock C. Optimization of multiple spin-echo sequences for 3D polymer gel dosimetry. Phys Med Biol 2002;47:3117–3141.
91. Marquardt DW. An algorithm for least-squares estimation of nonlinear parameters. J Soc Ind Appl Math 1963;11:431–441.
92. Maryanski MJ, et al. Magnetic resonance imaging of radiation dose distributions using a polymer-gel dosimeter. Phys Med Biol 1994;39:1437–1455.
93. Maryanski MJ, Audet C, Gore JC. Effects of crosslinking and temperature on the dose response of a BANG polymer gel dosimeter. Phys Med Biol 1997;42:303–311.
94. Hrbacek J, Spevacek V, Novotny J Jr, Cechak T. A comparative study of four polymer gel dosimeters. J Phys Conf Ser 2004;3:150–154.
95. De Deene Y, De Wagter C. Artefacts in multi-echo T2 imaging for high-precision gel dosimetry. III. Effects of temperature drift during scanning. Phys Med Biol 2001;46:2697–2711.
96. De Deene Y, et al. An investigation of the chemical stability of a monomer/polymer gel dosimeter. Phys Med Biol 2000; 45:859–478.
97. MacDougall ND, Pitchford WG, Smith MA. A systematic review of the precision and accuracy of dose measurements in photon radiotherapy using polymer and Fricke MRI gel dosimetry. Phys Med Biol 2002;47:R107–R121.
98. Jirasek AI, Duzenli C, Audet C, Eldridge J. Characterization of monomer/crosslinker consumption and polymer formation observed in FT-Raman spectra of irradiated polyacrylamide gels. Phys Med Biol, 2001;46:151–165.
99. Spinks JWT, Woods RJ. An Introduction to Radiation Chemistry. New York, London, Sydney: Wiley; 1964.
100. Kennan RP, et al. The effects of cross-link density and chemical change on magnetization transfer in polyacrylamide gets. J Magn Res B 1996;100:267–277.
101. Oldham M, et al. Improving calibration accuracy in gel dosimetry. Phys Med Biol 1998; 43:2709–2720.
102. Baldock C, et al. Investigation of polymerisation of radiation dosimetry polymer gels Proceedings. 1st International Workshop on Radiation Therapy Gel Dosimetry (Lexington, KY). Schreiner L J, Audet C, editors. Canadian Organisation of Medical Physics 1999. p 99–105.
103. McJury M, Oldham M, Leach MO, Webb S. Dynamics of polymerization in polyacrylamide gel (PAG) dosimeters I. Ageing and long-term stability Phys Med Biol 1999;44: 1863–1873.
104. De Deene Y, et al. Artefacts in multi-echo T2 imaging for high-precision gel dosimetry. II. Analysis of B1 field inhomogeneity. Phys Med Biol 2000;45:1825–1839.
105. De Deene Y, et al. Artefacts in multi-echo T2 imaging for high-precision gel dosimetry. I. Analysis and compensation of eddy currents. Phys Med Biol 2000;45:1807–1823.
106. Watanabe Y, Perera GM, Mooij RB. Image distortion in MRI-based polymer gel dosimetry of Gamma Knife stereotactic radiosurgery systems. Med Phys 2002;29:797–802.
107. Maryanski MJ, Zastavker YZ, Gore JC. Radiation dose distributions in three dimensions from tomographic optical

density scanning of polymer gels: II. Optical properties of the BANG polymer gel. Phys Med Biol 1996;41:2705–2717.
108. Gore JC, Ranade M, Maryanski MJ, Schulz RJ. Radiation dose distributions in three dimensions from tomographic optical density scanning of polymer gels: I. Development of an optical scanner. Phys Med Biol 1996;41:2695–2704.
109. Oldham M, Siewerdsen JH, Shetty A, Jaffray DA. 1998b; High resolution gel-dosimetry by optical-CT and MR scanning. Med Phys 2001;28:1436–1445.
110. Oldham M et al. Optical-CT gel dosimetry I: Basic investigations. Med Phys 2003;30:623–634.
111. Wolodzko JG, Marsden C, Appleby A. CCD imaging for optical tomography of gel radiation dosimeters. Med Phys 1999;26:2508–2513.
112. Jordan K. Advances in optical CT scanning for gel dosimetry. J Phys Conf Ser 2004;3:115–121.
113. Oldham M. Optical CT scanning of polymer gels. J Phys: Conf Ser 2004;3:122–135.
114. Maryanski MJ. High-resolution 3D dosimetry for endovascular brachytherapy using optical laser CT microimaging of BANG polymer gels. Med Phys 1998;25:A107.
115. Knisely JPS et al. Three-dimensional dosimetry for complex stereotactic radiosurgery using a tomographic optical density scanner and BANG polymer gels. In: Radiosurgery 1997, Dondziolka D, editors Vol 2. Basel : Karger; 1998. p 251–260.
116. Islam KTS, et al. Initial evaluation of commercial optical CT-based 3D gel dosimeter. Med Phys 2003;30:2159–2168.
117. Heard MP and Ibbott GS. Measurement of brachytherapy sources using MAGIC gel. J Phy: Conf Ser 2004;3:221–223.
118. Xu YS, Wuu C-S, Maryanski MJ. Performance of a commercial optical CT scanner and polymer gel dosemeters for 3-D dose verification. Med Phys 2004;31:3024.
119. Xu Y, Wuu C-S, Maryanski MJ. Determining optical gel sensetivity in optima CT scanning of gel dosemeters. Med Phys 2003;30:2257.
120. Hilts M, Audet C, Duzenli C, Jirasek A. Polymer gel dosimetry using X-ray computed tomography: A feasibility study. Phys Med Biol 2000;45:2559–2571.
121. Trapp JV, et al. An experimental study of the dose response of polymer gel dosimeters imaged with x-ray computed tomography. Phys Med Biol 2001;46:2939–2951.
122. Trapp JV, Michael G, De Deene Y, Baldock C. Attenuation of diagnostic energy photons by polymer gel dosimeters. Phys Med Biol 2002;47:4247–4258.
123. Audet C, Hilts M, Jirasek A, Duzenli C. CT gel dosimetry technique: comparison of a planned and measured 3D stereotactic dose volume. J Appl Clin Med Phys 2002;3; 110–118.
124. Brindha S, Venning A, Hill B, Baldock C. Experimental investigation of the attenuation properties of normoxic polymer gel dosimeters. Med Phys 2004;31:1886.
125. Hilts M, Audet C, Duzenli C, Jirasek A. Polymer gel dosimetry using X-ray computer tomography: feasibility and potential application to stereotactic radiosurgery Proc. 1st Int. Workshop on Radiation Therapy Gel Dosimetry (Lexington, KY USA) Schreiner L J Audet C editors. 1999.
126. Hilts M, Duzenli C. Image filtering for improved dose resolution in CT polymer gel dosimetry. Med Phys 2004a;31:39–49.
127. Hilts M, Jirasek A, Duzenli C. Effects of gel composition on the radiation induced density change in PAG polymer gel dosimeters: a model and experimental investigations. Phys Med Biol 2004;49:2477–2490.
128. Hilts M, Jirasek A, Duzenli C. The response of PAG density to dose: a model and experimental investigations. J Phy: Conf Ser. 2004c;3:163–167
129. Mather ML, Whittaker AK, Baldock C. Ultrasound evaluation of polymer gel dosimeters. Phys Med Biol 2002;47: 1449–1458.
130. Mather ML et al. Investigation of ultrasonic properties of PAG and MAGIC polymer gel dosimeters. Phys Med Biol 2002;47:4397–4409.
131. Mather ML, et al. Acoustic evaluation of polymer gel dosimeters. Proc Int Symp Standards and Codes of Practice in Medical Radiation Dosimetry. International Atomic Energy Agency, Vienna; 2002c. p 234–235.
132. Mather ML, Baldock C. Ultrasound tomography imaging of radiation dose distributions in polymer gel dosimeters. Med Phys 2003;30:2140–2148.
133. Baldock C et al. Fourier transform Raman spectroscopy of polyacrylamide gels (PAGs) for radiation dosimetry. Phys Med Biol 1998; 43:3617–3627.
134. Baldock C. X-ray computer tomography, ultrasound and vibrational spectroscopic evaluation techniques of polymer gel dosimeters. J Phy Conf Ser, 2004;3:136–141.
135. Lepage M, Whittaker AK, Rintoul L, Baldock C. ^{13}C-NMR, ^{1}H-NMR and FT-Raman study of radiation-induced modifications in radiation dosimetry polymer gels. J Appl Polym Sci 2001;79:1572–1581.
136. Rintoul L, Lepage M, Baldock C. Radiation dose distributions in polymer gels by Raman spectroscopy. Appl Spectrosc 2003;57:51–57.
137. Jirasek A, Duzenli C. Relative effectiveness of polyacrylamide gel dosimeters applied to proton beams: Fourier transform Raman observations and track structure calculations. Med Phys 2002;29:569–577.
138. McJury M et al. Radiation dosimetry using polymer gels: methods and applications. Br J Radiol 2000;73:919–929.
139. Lepage M, Jayasakera PM, Back SAJ, Baldock C. Dose resolution optimization of polymer gel dosimeters using different monomers. Phys Med Biol 2001;46:2665–2680.
140. Trapp JV et al. Dose resolution in gel dosimetry: effect of uncertainty in the calibration function. Phys Med Biol 2004;49:N139–N146.
141. Day MJ. Radiation dosimetry using nuclear magnetic resonance an introductory review. Phys Med Biol 1990;35:1605.
142. Bonnett D. A review of application of polymer gel dosimetry. DOS GEL 2001. Proc 2nd Int Conf Radiotherapy Gel Dosimetry. In: Baldock C, DeDeene Y, editors. Queensland University of Technology, Brislane, Queensland, Australia. p 40–48
143. Ibbott GS. Applications of Gel Dosimetry. J Phys Conf Ser 2004;3:58–77.
144. Ibbott GS, Maryanski MJ, Avison RG, Gore JC. Investigation of a BANG polymer gel dosimeter for use as a mailed QA device. Med Phys 1995;22:951.
145. Haraldsson P, Back SA, Magnusson P, Olsson LE. Dose response characteristics and basic dose distribution data for a polymerization-based dosimeter gel evaluated using MR. Br J Radiol 2000;73:919–929.
146. Oldham M et al. An investigation into the dosimetry of a nine-field tomotherapy irradiation using BANG-gel dosimetry Phys Med Biol 1998;43:1113–1132.
147. Baldock C et al. A dosimetry phantom for external beam radiation therapy of the breast using radiation-sensitive polymer gels and MRI. Med Phys 1996;23:1490.
148. Love PA, Evans PM, Leach MO, Webb S. Polymer gel measurement of dose homogeneity in the breast: comparing MLC intensity modulation with standard wedged delivery. Phys Med Biol 2003;48:1065–1074.
149. De Deene Y et al. Three-dimensional dosimetry using polymer gel and magnetic resonance imaging applied to the verification of conformal radiation therapy in head-and-neck cancer. Radiother Oncol 1998;48:283–291.
150. Trapp JV, et al. The use of gel dosimetry for verification of electron and photon treatment plans in carcinoma of the scalp. Phys Med Biol 2004;49:1625–1635.

151. Bengtsson M et al. Measurement of dynamic wedge angles and beam profiles by means of MRI ferrous sulphate gel dosimetry. Phys Med Biol 1996;41:269–277.
152. Hill B, Venning C, Baldock C. Acceptance testing of a computer tomography scanner using normoxic polymer gel dosimetry. Med Phys 2004;31:1786.
153. Hill B, Venning C, Baldock C. X-ray computer tomography dose response of normoxic polymer gel dosimeters. Br J Radiol. In press. 2005.
154. Olsson LE, Arndt J, Fransson A, Nordell B. Three-dimensional dose mapping from gamma knife treatment using a dosimeter gel and MR-imaging. Radiother Oncol 1992; 24:82–86.
155. Schulz RJ, Maryanski MJ, Ibbott GS, Bond JE. Assessment of the Accuracy of Stereotactic Radiosurgery Using Fricke-Infused Gels and MRI. Med Phys 1993;20:1731–1735.
156. Ibbott GS et al. Stereotactic radiosurgery simulation using MRI and a polymer-gel dosimeter. Med Phys May/June 1993;20(3) .
157. Ibbott GS, et al. Use of BANG polymer gel dosimeter to evaluate repeat-fixation stereotactic radiation therapy. Med Phys 1996;23:1070.
158. Meeks SL et al. Image registration of BANG gel dose maps for quantitative dosimetry verification. Int J Radiat Oncol Biol Phys 1999;43:1135–1151.
159. Novotny J Jr et al. Quality control of the stereotactic radiosurgery procedure with the polymer-gel dosimetry. Radiother Oncol 2002;63:223–230.
160. Scheib SG, Gianolini S. Three-dimensional dose verification using BANG gel: a clinical example. J Neurosurg 2002;97: 582–587
161. Ibbott GS, et al. Three dimensional visualization and measurement of conformal dose distributions using magnetic resonance imaging of BANG polymer gel dosimeters. Int J Radiat Oncol Biol Phys 1997;38:1097–1103.
162. Low DA et al. Evaluation of polymer gels and MRI as a 3D dosimeter for intensity-modulated radiation therapy. Int J Radiat Oncol Biol Phys 1999;26:154.
163. Beach M et al. Implementation of a Polymer Gel Dosimetry Insert for An Anthropomorphic Phantom Used to Evaluate Head and Neck Intensity-Modulated Radiation Therapy. desseitation, UT M. D. Anderson Cancer Center. 2003.
164. De Neve W. Clinical delivery of intensity modulated conformal radiotherapy for relapsed or second-primary head and neck cancer using a multileaf collimator with dynamic control. Radiother Oncol 1999;50:301–314.
165. Ibbott G, Beach M, Maryanski M. An anthropomorphic head phantom with a BANG® polymer gel insert for dosimetric evaluation of IMRT treatment delivery, Standards and Codes of Practice in Medical Radiation Dosimetry, Proc Int Symp. Vienna 2002;2:361–368.
166. Ibbott GS, Beach ML, Maryanski MJ. IMRT QA with an Anthropomorphic Phantom Employing a Polymer Gel Dosimeter. Int Organization Med Phys Proc. Vol. 1 2003.
167. Gustavsson H et al. MAGIC-type polymer gel for three-dimensional dosimetry: Intensity-modulated radiation therapy verification. Med Phys 2003;30:1264–1271.
168. Vergote K, et al. Application of monomer/polymer gel dosimetry to study the effects of tissue inhomogeneities on intensity-modulated radiation therapy (IMRT) dose distributions. Radiother Oncol 2003;67:119–128.
169. Vergote K, et al. Validation and application of polymer gel dosimetry for the dose verification of an intensity-modulated arc therapy (IMAT) treatment. Phys Med Biol 2004;49:287–305.
170. Molineu A et al. Design and implementation of an anthropomorphic quality assurance phantom for intensity modulated radiation therapy. Int J Radiat Oncol Biol Phys 2005; (In press)
171. Heard MP. Characterizing Dose Distributions of Brachytherapy Sources using Normoxic Gel. MS dissertation M. D. Anderson Cancer Center, Houston (TX) 2005.
172. Duthoy W, et al. Whole abdominopelvic radiotherapy (WAPRT) using intensity-modulated arc therapy (IMAT): first clinical experience. Int J Rad Oncol Biol Phys 2003;57:1019–1032.
173. Cadman P et al. Dosimetric considerations for validation of a sequential IMRT process with a commercial treatment planning system. Phys Med Biol 2002;47; 3001–3010.
174. Nath R, Melillo A. Dosimetric characteristics of a double wall ^{125}I source for interstitial brachytherapy. Med Phys 1993;20:1475–1483.
175. Muench PJ, Meigooni AS, Nath R, McLaughlin WL. Photon energy dependence of the sensitivity of radiochromic film and comparison with silver halide film and LiF TLDs used for brachytherapy dosimetry. Med Phys 1991;18:769–775.
176. Schreiner LJ, et al. Imaging of HDR brachytherapy dose distributions using NRM Fricke-gelatin dosimetry. Magn Reson Imaging 1994;12:901–907.
177. Olsen DR, Hellesnes J. Absorbed dose distribution measurements in brachytherapy using ferrous sulphate gel and magnetic resonance imaging. Br J Radiol 1994;67:1121–1126.
178. Maryanski MJ, et al. Magnetic Resonance Imaging of Dose Distributions from Brachytherapy Sources Embedded in Tissue Equivalent BANG Polymer Gel Dosimeters. Med Phys 1994;21:919 (abstract).
179. De Deene Y et al. On the accuracy of monomer/polymer gel dosimetry in the proximity of a high-dose-rate ^{192}Ir source. Phys Med Biol 2001;46:2801–2825.
180. Gelfi C, Righetti P G. Polymerization kinetics of polyacrylamide gels: II. Effect of temperature. Electrophoresis 1981;2: 220–228.
181. Omidian H, Hashemi SA, Sammes PG, Meldrum IG. Modified acrylic-based superabsorbent polymers. Effect of temperature and initiator concentration. Polymer 1988;39:3459–3466.
182. Ibbott G, et al. Characterization of a New Brachytherapy Source by BANG® Gel Dosimetry. DosGel 99: Proc 1st Int Workshop Radiation Therapy Gel Dosimetry. Canadian Organisation of Medical Physicists and the Canadian College of Physicists in Medicine. 196–198. 1999.
183. Ibbott GS et al. Characteristics of a new brachytherapy source by BANG® gel dosimetry. Int J Rad Oncol Biol Phys. 1999;45(35):417.
184. Heard M, Ibbott G, Followill D. Characterizing Dose Distributions of Brachytherapy Sources Using Normoxic Gel (WIP), AAPM Annual Meeting 2003.
185. Chan M et al. The measurement of three dimensional dose distribution of a ruthenium-106 ophthalmological applicator using magnetic resonance imaging of BANG polymer gels. J Appl Clin Med Phys 2001;2:85–89.
186. Gifford K et al. Verification of Monte Carlo calculations around a fletcher suit delclos ovoid with radiochromic film and normoxic polymer gel dosimetry. Med Phys 2004;31.
187. Gifford K et al. Verification of monte carlo calculations around a Fletcher suit delclos ovoid with normoxic polymer gel dosimetry. J Phys: Conf Ser, 2004;3:217–220.
188. Wuu C-S, et al. Dosimetry study of Re-188 liquid balloon for intravascular brachytherapy using polymer gel dosimeters and laser-beam optical CT scanner. Med Phys 2003;30: 132–137.
189. Vergote K, et al. On the relation between the spatial dose integrity and the temporal instability of polymer gel dosimeters. Phys Med Biol 2004;49:4507–4522.
190. Pantelis E, et al. Polymer gel water equivalence and relative energy response with emphasis on low photon energy dosimetry in brachytherapy. Phys Med Biol 2004;49, 3495–3514.

191. Venning AJ, Brindha S, Hill B, Baldock C. Preliminary study of a nomoxic PAG gel dosemeter with tetrabis (hydroxymethyl phosphonium chloride as an antioxidant. J Phip: Conf Ser 3 2004; 155–158.
192. Courbon F et al. Internal dosimetry using magnetic resonance imaging of polymer gel irradiated with iodine-131. Preliminary results. Proc 1st Int Workshop Radiation Therapy Gel Dosimetry (Lexington, KY). In: Schreiner L J Audet C, editors. Canadian Ottawa, Ontario, Canada: Organization of Medical Physicists, 1999.
193. Gambarini G et al. Three-dimensional determination of absorbed dose by spectrophotometric analysis of Ferrous-Sulphate Agarose gel. Nucl Instrum and Meth 1999; A 422:643–648.
194. Gambarini G. Gel dosimetry in neutron capture therapy. Proc 2nd Int Conf Radiotherapy Gel Dosimetry (Brisbane, Australia). 2001. p 89–91.
195. Gambarini G. et al. Fricke-gel dosimetry in boron neutron capture therapy. Radiat Prot Dosim 2002;101:419–422.
196. Bäck S A et al. Ferrous sulphate gel dosimetry and MRI for proton beam dose measurements. Phys Med Biol 1999; 44:1983–1996.
197. Gustavsson H, Karlsson A, Back S, Olsson LE. Dose response characteristics of a new normoxic polymer gel dosimeter. Ph.D. dissertation Department of Medical Radiation Physics, Lund University, Malmo University Hospital. 2004.
198. Gustavsson H et al. Linear energy transfer dependence of a normoxic polymer gel dosimeter investigated using proton beam absorbed dose measurements. Phys Med Biol 2004;49: 3847–3855.
199. Swallow AJ. Radiation Chemistry: an Introduction London: Longman Group limited; 1973.
200. Ramm U, et al. Three-dimensional BANG™gel dosimetry in conformal carbon ion radiotherapy. Phys Med Biol 2000;45: N95–N102.
201. Vergote Ket al. Comparisons between monomer/polymer gel dosemetry and dose computations for an IMRT treatment of a thorox phantom. In: Baldock- C, De Deene Y editor DOSGEL 2001, Proc 2nd Int Conf Radiotherapy Gel Dosemetry. Queensland University of Technology, Brisbane, Queensland, Australia.
202. Hepworth SJ et al. Dose mapping of inhomogeneities positioned in radiosentive polymer gels. Nucl. Instrum. Methods Phys Res A 1999;422:756–760.
203. Gum F, et al. Preliminary study on the use of an inhomogeneous anthropomorphic Fricke gel phantom and 3D magnetic resonance dosimetry for verification of IMRT treatment plans, Phys Med Biol 2002;47; N67–N77.
204. Watanabe Y, Mooij RB, Perera GM, Maryanski MJ. Heterogeneity phantoms for visualization of 3D dose distributions by MRI-based polymer gel dosimetry. Med Phys 2004;31: 975–984
205. Olberg S, Skretting A, Bruland O, Olsen DR. Dose distribution measurements by MRI of a phantom containing lung tissue equivalent compartments made of ferrous sulphate gel. Phys Med Biol 2000;45:2761–2770.
206. Borges JA, BenComo J, Ibbott GS. A 3 Dimensional Gel Dosimetry Lung Equivalent (WIP), AAPM annual meeting, 10–14 Aug 2003, San Diego(CA); 2003.
207. Salomons GJ, Park YS, McAuley KB, Schreiner LJ. Temperature increases associated with polymerization of irradiated PAG dosimeters. Phys Med Biol 2002;47:1435–1448.
208. De Deene Y et al. Dose-response stability and integrity of the dose distribution of various polymer gel dosimeters. Phys Med Biol 2002;47:2459–2470.
209. Keall PJ, Baldock C. A theoretical study of the radiological properties and water equivalence of Fricke and polymer gels used for radiation dosimetry. Aust Phys Eng Sci Med 1999;22: 85–91.
210. McAuley KB. The chemistry and physics of polyacrylamide gel dosimeters: why they do and don't work. J. Phys Conf Ser 2004;3; 29–33.

See also PHANTOM MATERIALS IN RADIOLOGY; RADIATION DOSIMETRY FOR ONCOLOGY; RADIATION THERAPY, INTENSITY MODULATED; RADIATION THERAPY SIMULATOR; RADIOSURGERY, STEREOTACTIC.

RADIATION, EFFECTS OF.

See IONIZING RADIATION, BIOLOGICAL EFFECTS OF; NONIONIZING RADIATION, BIOLOGICAL EFFECTS OF.

RADIATION PROTECTION INSTRUMENTATION

GLENN P. GLASGOW
Loyola University of Chicago
Maywood, Illinois

RADIATION PROTECTION INSTRUMENTATION

Radioactive materials and equipment that generate radiation are prevalent in industry, military, education, science and medical facilities, and even in the home. Many scientific instruments perform dedicated radiation measurement tasks; the nuclear power industry employs possibly the greatest number of instruments of different designs and degrees of sophistication. This article describes similar instruments commonly used for radiation protection in medicine. Instruments used for radiation dosimetry for medical treatments (e.g., radiotherapy ionization chambers,) and those used for medical treatments (e.g., nuclear medicine well-ionization chambers) are excluded. Included are instruments used for the general tasks of detecting radiation, determining the types of radiation or species of radionuclides present, determining quantities of radionuclides, and measuring radiation levels around materials and equipment. The focus is how instruments detect radiation, not their electronic circuity, which is described only briefly in a few instances. Before choosing an instrument, the user must know about the availability and choice of instruments, types and sources of radiation, special terms that describe quantities of radiation, and measures of biological dose equivalency that individuals receives in the presence of radiation. A science discipline, *Health Physics*, and a scientific society, the *Health Physics Society*, are devoted to these topics (1). Since the1988 first edition of this article, major changes in medical radiation protection instrumentation include the development of the Internet for dissemination, by manufacturers and vendors, of information about instrument designs, operating parameters, and performances; improved performance and electronic circuitry using chips with complementary metal oxide semiconductors (CMOS) microprocessor technology of various types; miniaturization of computer components that reduce the weight and size of instruments; and new

Table 1. Some Distributors and Manufacturers of Radiation Protection Instruments

Company (Product Lines)	Internet Address	Electronic Mail Address
Berkeley Nucleonics	http://www.berkeleynucleonics.com	info@berkeleynucleonics.com
Berthold Technologies GmbH & Co.	http://www.bertholdtech.com	info@BertholdTech.com
Canberra Industries (Packard)	http://www.canberra.com	customersupport@canberra.com
Capintec, Inc	http://www.capintec.com	getinfo@capintec.com
Cardinal Health Nuclear Pharmacy Services (Inovision, Victoreen)	http://www.nps.cardinal.com	npsinfo@cardinal.com
Durridge Company, Inc.	http://www.durridge.com	sales@durridge.com
Far West Technology & Health Physics Instruments	http://www.fwt.com	info@fwt.com
Global Dosimetry Solutions	http://www.dosimetry.com	info@dosimetry.com
International Specialty Products	http://www.ispcorp.com	customerservicecenter@ispcorp.com
Landauer, Inc.	http://www.landauerinc.com	custserv@landauerinc.com
LAURUS Systems, Inc	http://www.laurussystems.com	sales@laurussystems.com
Ludlum Measurements, Inc.	http://www.ludlums.com	ludlum@ludlums.com
Ortec	http://www.ortec-online.com	info@ortec-online.com
Perkin Elmer Life and Analytical Sciences	http://www.las.perkinelmer.com	products@perkinelmer.com
Technical Associates	http://www.tech-associates.com	tagold@nwc.com
Thermo Electron Corporation	http://www.thermo.com	enviromental.radiation@thermo.com

definitions and terms used to describe radiation quantities and units [*Note*: In addition to the common prefixes of kilo- (k), mega- (M), giga- (G), milli- (m), micro- (μ), nano- (n), pico- (p), note the use of the somewhat less familiar femto- (f; 10^{-15}), atto- (a; 10^{-18}), zepto- (z; 10^{-21}), and yocto- (y; 10^{-24})] (2). Manufacturers market smaller, compact survey meters, personnel dosimeters, and specialized detectors and monitors with improved performance. We review common features of instruments, such as ionization chambers, gas-proportional counters, Geiger–Müeller (GM) tubes, scintillation and solid-state detectors, other less-common detectors, and photographic films.

AVAILABILITY OF INSTRUMENTS AND INFORMATION

Table 1 lists some major companies and manufacturers of radiation protection instruments, their worldwide web Internet addresses, and their electronic mail addresses. Commercial product catalogues, usually now available on the Internet, contain a wealth of specific information on the theory and operation of instruments. This company list represents no endorsement by the author; these companies were selected because their worldwide web Internet sites provide details about common radiation protection instruments advertised for research, laboratory, environmental, security, medical, and health physics (safety and protection) applications. Basic instruments require only modest modifications for specific field applications. Table 2 contains a typical product list of medical radiation protection instruments. Instruments are regularly reviewed in Technology Monitor articles in *Health Physics* (3). One general interest group shares information regarding procedures, selection, testing, and standardization of instruments (4). Basic radiation detection principles and instrument designs are described in university level science textbooks in Health Physics curricula; comprehensive descriptions appear in Knoll (5), Shapiro (6), Turner (7), and Gollnick (8).

CHOICE OF INSTRUMENTS

A radiation field often consists of multiple types of radiation. Instruments usually must have the capability to detect particular types of radiation and produce relative or absolute measures of their magnitudes, while discriminating against other types of radiation. Often the radiation energies must be measured. Common medical uses, Table 3, include equipment radiation surveys, area monitoring, area and personnel contamination surveys, personnel dosimetry, finding misplaced radioactive materials (radioactive seeds or sources), surveying radioactive packages, air sampling, and emergency response tasks. Individuals choosing radiation protection instruments for measurements preferably should know about the radiation environment under investigation. Is it predominantly photon

Table 2. A List of Some Medical Radiation Protection Products

Air Monitors	Neutron Meters
Alarm Ratemeters	Package Monitors
Alpha Detectors	Pocket Dosimeters
Alpha/Beta/Gamma Detectors	Pocket Survey Meters
Alpha/Beta Detectors	Portable Accessories
Area Monitors	Portable Scaler/Ratemeters
Beta Detectors	Proportional Probes
Beta/Gamma Detectors	Response Kits
Connectors	Sample Counters
Counters	Sample Holders
Detector Accessories	Scalers and Accessories
Dosimeters	Scintillation Well Counting and Detection Systems
Gamma Detectors	Specialized Monitors
Geiger Counters	Specialized Portable Meters
Geiger-Müller Probes	Survey Meters
Ion Chambers	Test Equipment
MicroR Meters	Wipe Counters
Neutron Detectors	X-ray Monitors

Table 3. Some Typical Radiation Protection Instruments and Their Major Features

Type	Generic Name	Characteristics	Uses
		Portable Survey Meters	
Alpha, Beta, Gamma	Ion Chamber	Air ionization chambers to detect alpha, beta, gamma, and X rays from 50 $nSv \cdot h^{-1}$ to 20 $Sv \cdot h^{-1}$; sliding alpha and beta shield	General purpose survey meters with large range features
Beta, gamma	Geiger Counter Rate Meters	Multiple ranges up to 60 kcpm and 5 $nSv \cdot h^{-1}$; uses halogen quenched GM tube; multiple attachable probes	General purpose survey meters for lower level (5 $nSv \cdot h^{-1}$) surveys
Gamma	Gamma Survey Meters	Multiple ranges to 10 $\mu Sv \cdot h^{-1}$; halogen quenched GM tube with energy compensation to 40 keV	General purpose survey meter for gamma ray surveys
Alpha, (H-3), beta	Gas Proportional Survey Meter	Measures alphas, low energy beta to 500 kcpm	For measurements in presence of volatile vapor, high γ-ray fields, and for surface contamination
Gamma	Micro "R" Meter	Uses 1 × 1 in. NaI(Tl) scintillator to measure 0.1 $\mu Sv \cdot h^{-1}$	For sensitive low level surveys of $\mu Sv \cdot h^{-1}$ levels
Alpha, gamma	Alpha-Gamma Scintillation Counter	Measures alpha to 2×10^6 cpm using scintillator; measures gammas to 20 $mSv \cdot h^{-1}$ using GM tubes	For simultaneous measurements of alpha–gamma contamination
Alpha, beta, gamma, neutrons, X rays	Scaler, Ratemeter, Single Channel Analyzer	Multiple ranges (1 $nSv \cdot h^{-1}$–10 $Sv \cdot h^{-1}$; 1 cpm–500 kcpm) with single-channel analyzer with selected windows	For use with multiple probes of many types; measures identifies type of radiation or radionuclide
Neutrons	Neutron rem Meter	Measures equivalent dose neutrons using BF_3 tube in a cadmium loaded polyethylene moderator	General purpose neutron detection for thermal to high energy neutrons
		Personnel Electronic Dosimeters	
Gamma	Alarming Dose Rate Meter	Scintillation detector sensitive to 0.1 $\mu Sv \cdot h^{-1}$ to 20 $\mu Sv \cdot h^{-1}$ with multiple preset alarm levels	Medical personnel monitoring
Gamma, beta, neutron	Alarming Dose Rate Meter	Silicon semiconductor detectors; 10 μSv–1 Sv	Medical personnel monitoring
		Area Monitors	
Gamma	Area Radiation Monitors	Alarming counter rate meters with adjustable alarm that sounds when exposure rates exceed preset levels. Usually have GM tube detectors	Used to monitor areas where personnel prepare and use sources; used to determine that remote control sources have retracted to a safe.
		Air Monitors and Samplers	
Beta, Gamma	Beta–Gamma Air Particulate Monitor	Measure airborne particulate beta emitting particles using pancake-type GM tubes; ^{133}Xe monitors	Alarm monitor for laboratories using radioactive gasses emitting beta particles
		Well Counting Systems	
Gamma, beta	Liquid Scintillation Counting system	Counting of wipe tests from labs, sources to identify type and amount of radionuclide	General radiation control and containment
		Spectroscopy	
Gamma, beta	Multichannel Analyzer with NaI(Tl) or Ge detector	Identification of radonuclides by characteristic spectral analysis	General radiation control and containment, nuclear medicine labs, research labs, and so on.

(X ray or γ ray) radiation, charged particle (proton, beta particle) radiation, neutron radiation, or mixtures thereof? Spectroscopy measurements can determine the types and energy distributions, but often measurements require simpler detection or measurement devices. The radiation environment may be characterized by the maximum energy of radiation, whether the radiation source is continuous, as with an X-ray unit, rapidly pulsed as with some linear accelerators, or is random decay from a radioisotope. Is the measurement made in the primary direct beam, or in scattered radiation beam filtered by radiation barriers? Choice of instruments depends on why the radiation is being measured. It is desirable to know the approximate magnitude of radiation, expressed in some appropriate units, and the approximate energy of the radiation. It is then possible to estimate personnel equivalent dose rates

in the radiation field. Multiple instruments with different features may be required to properly characterize and measure a radiation environment.

TYPES OF RADIATION

Radiation is a general term used to describe the emission and propagation of energy through space or a material. Texts describing instruments describe types of radiation (5–8). Mohr and Taylor, on behalf of The Committee on Data for Science and Technology, updated, through 2002, the numerical (*Note*: In the interest of better science, we present exact values as they are not widely published nor readily available!) parameters of common radiation types (9). Radiation may be classified as directly ionizing, indirectly ionizing, or nonionizing (e.g., microwaves, laser lights, ultrasound, not further discussed here). Particulate forms of radiation (protons, electrons) possess one unit of electrical charge (160.217653 zC is the unit of electrical charge of an electron) and directly ionize atoms and molecules, as do particulate radiations with multiple charges, such as alpha particles.

Indirecty ionizing forms of radiation (X rays, γ rays, neutrons) lack electrical charge, but interact with matter and produce secondary charged particles (electrons, positrons) that ionize atoms. The magnitude of the energy possessed by the radiation, frequently expressed in millions of electron volts (MeV) and the mass of particulate radiation, expressed in atomic mass units (1 amu is defined as 1/12 of the mass of the carbon atom, and equals 0.00166053886 ykg) are important physical parameters that arise in describing the properties of radiation.

Protons have a mass of 1.007276446 amu, possess one unit of positive charge and are one of the core particles of the nucleus. Protons are heavy charged particles, that lose energy mostly by ionization and excitation of atoms as they exert electromagnetic forces on the orbital electrons surrounding the nucleus. Loosing only a small fraction of energy during each interaction, protons move through matter mostly in a straight-line path leaving in their wake ionized or excited atoms. Heavy charged particles require great energy to penetrate tissue. A 10 MeV proton has a range of ~ 0.11 cm, while a 100 MeV proton has a range of ~ 7.5 cm. Protons with several tens of million electronvolts of energy are used at research facilities with particle accelerators as probes to study nuclear structure. Neutrons with a mass of 1.008664915 amu are slightly more massive than protons, but lack charge and interact in matter primarily by collisions with protons to which they impart a portion of their energy during the collision. Neutrons are generally classified as thermal if their energies are < 0.5 eV, intermediate if their energies are > 0.5 eV, but < 0.2 MeV, and fast if it is > 0.2 MeV, but < 20 MeV. Lacking charge, neutrons are generally more penetrating in tissue than protons of the same energy and interact with atoms by elastic and inelastic collisions. Numerous radioactive sources serve as neutron generators; an alpha particle source, such as ^{241}Am, may be mixed with a light metal, such as beryllium to produce neutrons by a (α, n) reaction. Nuclear reactors are prolific neutron generators and numerous research facilities have accelerators capable of producing neutrons. Alpha particles, usually with several million electronvolts of energy, consist of two protons and two neutrons, and appear when certain nuclides decay into more stable nuclides, such as the decay of ^{238}U to ^{234}Th or the decay of ^{226}Ra and certain daughters. While large mass and double charge prevents even the most energetic alpha particles from penetrating much beyond the most superficial layer of external tissue, alpha particles are hazardous when ingested into the sensitive epithelium of the lungs.

Electrons have a mass of 0.00054857990945 amu, a small fraction of the mass of a proton, but carry an equal quantity of negative charge. Electrons with several tens of million electronvolts of energy can be generated with electron linear accelerators and many other pieces of equipment are capable of generating less energetic, but still hazardous electrons. Electrons interacting in a material can also produce a spectrum of bremsstrahlung X rays with the maximum X-ray energy identical to the maximum energy of the electrons. These X rays are far more penetrating that the electrons. Beta particles with several million electronvolts of energy arise from the nucleus during certain radioactive decay processes. Negative beta particles, negatrons, or ordinary electrons, possess a spectrum of energies below their maximum energy. Many nuclear transformations yield multiple beta particles; a few, for example, the decay of ^{32}P to ^{32}S, yield a single negative beta particle. Positrons have the same mass as electrons, but carrying a positive charge and arise in certain radionuclide transformations, such as the decay of ^{22}Na to ^{22}Ne. Beta particles and positrons with several million electronvolts of energy are more penetrating than alpha particles and even minute quantities of radioisotopes producing these particles, such as ^{32}P, can potentially produce damaging skin burns if spilled on the skin and left unattended.

Gamma rays frequently arise when a daughter radioactive nuclide in an excited state, decays by beta particle decay, or by other modes of decay to form a more stable nuclide, such as the decay of ^{60}Co to ^{60}Ni, yielding 1.17 and 1.33 MeV γ rays. These electromagnetic rays can possess several million electronvolts of energy and, lacking charge and mass, the more energetic γ rays can penetrate deeply into tissue and other materials. Following their interaction in a medium, such as tissue, they generate ionizing secondary electrons that actually produce the damage to cells.

X rays are a form of electromagnetic radiation arising from changes in the arrangements in the orbital electrons surrounding the nucleus, yielding characteristic X rays of several tens of kiloelectronvolts (keV) of energy. Another form of X rays, bremsstrahlung, are produced when energetic electrons with energy of several million electronvolts strike high Z targets yielding X rays with very high energies. Hence, X rays can span a broad energy range, from a few fractions of million electronvolts to several tens of million electronvolts depending on how they are produced. Like γ rays, the most energetic X rays have great potential to deeply penetrate matter. The term photon is used to describe X rays, γ rays, or other form of electromagnetic energy without referring to the method of production or source of the radiation.

Nuclei of atoms, such as a deuterium, ^{2}H, may be accelerated in highly energetic linear accelerators as a probe to study the properties of the nucleus of various elements. Because of the great mass and charge, heavy charged particles must possess several tens of million electronvolts of energy to penetrate tissue.

In addition to this limited list of types of radiation, numerous radioactive isotopes of the elements, such as ^{60}Co, ^{137}Cs, ^{131}I, and ^{125}I, and many more are widely used in medicine for a host of applications. Radioisotopes, through decay, can produce alpha particles, γ rays, beta particles, positrons, and X rays and each radioisotope has a unique spectrum of radiation that allows it to be identified even in the presence of other radioisotopes. While many forms of radioactive materials are encapsulated solids, others are unsealed and as liquids or as gases, are more readily dispersed during accidental releases. Hence, radiation is a term used to refer to many different forms of particulate and nonparticulate radiations with energies from fractions of a million electronvolt to tens and hundreds of million electronvolts. Obviously, the means of detecting radiation must be specific for the types of radiations present in a specific locale.

SOURCES OF RADIATION

The most energetic X rays, γ rays, and heavily charged particles are found almost exclusively in government or university sponsored scientific accelerator research facilities. Photons (X and γ rays) with energies as high as several million electronvolts are the most prevalent forms of radiation as equipment yielding X and γ ray are widely used in medical facilities. X-ray imaging in hospitals is probably the single most common medical use of radiation, with diagnostic use of radiopharmaceuticals next in importance. Beta particle sources and electron producing equipment are the next most prevalent sources of radiation. Neutron producing sources and equipment are frequently found in university research laboratories but are less common in medical facilities.

RADIATION QUANTITIES AND UNITS

Radiation protection definitions and terms often lack clear meanings, as noted by Storm and Watson (10). International commissions make recommendations, but national councils and regulatory bodies in different countries (or even within the same country) adopt or apply the recommendations differently (11). As radiation quantities and units, Table 4, are used on instrument displays, users must understand both historical and newly adopted radiation units. Gollnick offers a useful review (8). We limit this discussion to popular historical quantities and units and provided brief, albeit, limited descriptions of current quantities and units recommended by the International Commission on Radiological Protection, Systeme International

Table 4. International Radiation Concepts and Units[a]

Concept	Quantity	Symbol	SI unit	Numerical Value	Relationship to Other Concepts
Ionization of air by X and γ rays	Exposure	X	None	1 R = 0.000258 C of charge released per kilogram of air	
Kinetic energy released per unit mass of material	kerma (collisional)	K_{air}^{col}	Sv	1 Gy of energy transfered per kilogram of material	$K_{air}^{col} = XW/e$
Absorption of energy in a material	Absorbed dose	D	Gy	1 Gy = 1 J of energy absorbed per kilogram of material	$D = X f_{med}$
Risk of biological energy for different forms of radiation	Radiation weighting factor[b,c]	W_R	None	X rays, γ rays, beta particles = 1; Thermal neutrons, high energy particles = 5; Alpha particles, fast neutrons = 20	
Equivalent biological effect in humans	Equivalent dose[d]	H_T	Sv		$H_T = \Sigma W_R D$
Total (50 y) cummulative dose to an organ for internal radiation	Committed[d] equivalent dose	$H_T(50)$	Sv		
Reduced risk of partial body exposure to radiation	Tissue weighting factor[c]	W_T	None	Skin, bone surface = 0.01; bladder, liver = 0.05 colon; stomach = 0.12 gonads = 0.20	
Sum of weighted equivalent doses of partial body exposures	Effective dose[d]	E	Sv		$E = \Sigma W_T H_T$
Sum of weighted total (50 y) cummulative doses to organs from internal radiation	Equivalent dose	E (50)	Sv		$E(50) = \Sigma W_T H_T(50)$

[a]For complete concepts, definitions, and descriptions, see Refs. 11, 13 and 14.
[b]The equivalent concept, Q, albiet with different values, is used in the United States by the National Council on Radiation Protection Units and the Nuclear Regulatory Commission.
[c]For complete list of W_R, W_T, see Refs. 11–13.
[d]Similar, but different nomenclature is used in the United States by the National Council on Radiation Protection and Units and The Nuclear Regulatory Commission.

d' Unites, and the International Commission on Radiological Units and Measurements (11–13).

Counts (events) or count rates (events per unit time) are denoted on instruments that detect the presence and relative magnitudes of radiation. Count rates of a few counts per minute (cpm) to millions of cpm are possible, depending on the radiation field intensity.

Exposure, denoted by the symbol, X, is the measure of the ability of X and γ rays of energies 10 keV to < 3 MeV to ionize air and is the quotient of $\Delta Q/\Delta m$, where ΔQ is the sum of all charges of one sign produced in air when all of the electrons liberated by photons in a mass Δm of air are completely stopped. Exposure is expressed in a special unit, the Roentgen (R), equal to 258 μC of charge per kilogram of air at standard temperature and pressure. As with count rates, exposure rates generally vary from $\sim 5\ \mu R \cdot h^{-1}$ associated with natural background radiation to nearly $1\ MR \cdot h^{-1}$ in a linear accelerator X-ray beam. Because of its historical use in science and medicine, exposure remains popular even though its continued use is not recommended in the Systeme International d' Unites (14).

Kinetic energy released per unit mass of material, denoted in lower case, by the acronym, kerma, is a measure for indirectly ionizing radiations (photon, neutron) interacting in a material, of the total kinetic energy of all charged particles released per unit mass of material. Kerma possesses a collision component from the kinetic energy imparted by inelastic collisions with electrons and a radiation component (usually much smaller) from interactions with nuclei. For X rays absorbed in air, collision kerma in air is the product of exposure and the average energy, $\overline{W}$, required to produce an ion pair per unit electrical charge. Kerma is measured in gray (Gy), which is one joule (J) of energy released per kilogram (kg) of the medium. Turner offers a complete description of kerma and its relationship to other quantities (7).

Absorbed dose, denoted by the symbol D, is a measure of the amount of the released energy that is absorbed in the medium per unit mass of the medium. Under conditions of charged-particle equilibrium, with negligible energy loss, absorbed dose equals kerma. One joule of energy absorbed per kilogram of the medium, the gray, is widely used in medicine as a measure of absorbed dose, as is the submultiple, the centigray (cGy), 1/100th of a gray. The older special unit (now abandoned) of absorbed dose, rad, an acronym for roentgen absorbed dose, represents 0.01 J of energy absorbed per kilogram of the medium (1 Gy = 100 rad).

The absorbed dose in a medium may be determined from the exposure in air at the same point in the medium by multiplying the exposure by a conversion factor, f_{med} that converts the exposure in air to dose in the medium. The factor, f_{med}, is slightly < 1 for most biological materials, except bone, where values as high as 3 occur for the soft X rays in the diagnostic energy range.

Different types of radiation produce different degrees of biological damage when the same amount of energy per unit mass is deposited in the biological system. The radiation weighting factor, w_{R}, which replaced an older similar concept, quality factor, Q (now abandoned) is a measure of this phenomena (11). Radiation weighting factors of unity, Table 4, are assigned to most electrons, X and γ rays, while factors as high as 20 are assigned to alpha particles and fast neutrons. Hence, the biologically equivalent effect in tissue of the absorption of 1 Gy of alpha particles is 20 times more severe than the absorption of 1 Gy of 1 MeV γ rays.

Equivalent dose, usually denoted by the symbol H_{T}, is the term used in radiation control programs to monitor and record the biological equivalency of exposure to amounts of radiations of different energies that an individual has received. The equivalent dose in sievert (Sv), the product of the absorbed dose, D, in gray and the radiation weighting factor, w_{R}, is commonly used to express the biologically equivalency of absorbed doses of particular types and energies of radiation. Historically, this equivalency was expressed in the special unit, rem (now abandoned) an acronym for roentgen equivalent human. As the f_{med} factor and radiation weighting factors, w_{R}, ~ 1 for most photon energies commonly encountered in many situations, the units for kerma (Gy), absorbed dose (Gy), and equivalent dose (Sv) are nearly numerically equivalent and are often used interchangeably, as were the older abandoned special units, R, rad, and rem.Table 4 summarizes these relationships.

Committed equivalent dose, usually denoted by the symbol $H_{\text{T}}(50)$, in sieverts, is employed to consider exposure within the body from internally deposited radioactivity. It represents the total cumulative dose delivered over 50 years to an organ system by ingested radioactivity.

Effective dose, usually denoted by the symbol E, considers the consequences of partial body radiation exposure (11). Tissue weighting factors, w_{T}, account for the reduced effects that occur when only a portion (or organ system) of the body is irradiated. Values for w_{T} (Table 4) range from 0.01 for bone surfaces to 0.2 for the gonads. The effective dose, E, in sieverts, is the sum, over the body, of the products, $w_{\text{T}} H_{\text{T}}$ for each partially irradiated portion of the body.

Committed effective dose, usually denoted by the symbol, $E(50)$, in sieverts, is employed to consider exposure from internally deposited radioactivity. The committed effective dose, $E(50)$, is the sum, over the body of the products, $w_{\text{T}}\ H_{\text{T}}(50)$, for each partially irradiated portion of the body > 50 years.

For individuals experiencing both external and internal radiation exposure, methodologies exist (omitted here) to combine effective dose and committed effective dose to estimate cumulative risks from both types of exposures (8).

Activity, denoted by the symbol A, describes an amount of radioactivity, expressed in decays per second (dps). One becquerel (Bq) equals one decay per second. The curie (Ci), the original term used to describe an amount of activity, equals 37 Gdps. Trace amounts of radioactivity are generally expressed in the microcuries (μCi) quantities and laboratory cleanliness standards are often expressed in picocuries (pCi) or smaller amounts. One becquerel equals ~ 27 pCi. Activities of millions of curies are commonly found in power reactors while curie and millicurie (mCi) quantities of materials are commonly used in medical applications.

COMMON FEATURES OF INSTRUMENTS

Radiation protection instruments in a facility can be broadly categorized as those (e.g., area monitors, personnel

scanners) used in fixed locations and those (e.g., GM detectors, survey meters) moved for use in multiple locations. Fixed instruments usually are large, heavy, and use permanent electrically power. By design, they may have more features and offer more sensitive detection or measurement features than their portable counterparts. Portable instruments generally are small, lighter, and battery powered. Some feature tripod stands for temporary uses of long duration. Instruments may be categorized as those that detect or measure a quantity of radiation and those that grossly or specifically identify types of radiation or radionuclides. Instruments of all types usually feature multiple signal ranges because of the wide variations in the signal (counts, exposure, dose, etc.) monitored. The number of photons or particulate forms of radiation from a sizable radioactive source or large piece of equipment frequently varies inversely (or approximately so) as the square of the distance of the finite size detector from the radiation source and multiple signal ranges, usually in multiples of 3 or 10, are required to map the spatial variation of the radiation around the source or equipment that will vary from the highest signals measured very near the source to the smallest signal measured at distances far from the source of radiation. Current devices often feature auto-zeroing and auto-ranging of scales. Instrument scales indicate the magnitude or measure of the radiation detected. Often, more than one unit will appear on the scales, or users may electronically select from multiple choices of units. Some instruments feature a rate mode, usually per second, minute, or hour or integrate mode that sums the signals over some predetermined time periods. Instrument scales may only be correct for specific radiation conditions under which the instrument was calibrated. Instruments can yield incorrect readings when used under noncalibrated radiation conditions.

Efficiency of a device is a measure of the number of output parameters (counts, pulses, etc.) produced relative to the number of input parameters (γ rays, particles, etc.) producing the output. Efficiencies (absolute, intrinsic, relative, etc.) have technical definitions beyond the scope of this article. Different applications require instruments with different degrees of efficiency, but, generally, high efficiency is desirable.

Signal accuracy is highly variable and depends on the type of radiation monitored and the design of the instrument. The more intense and hazardous the radiation field the greater the necessity for more accurate measurements. Conversely, measurements in very low level radiation fields need not be as accurate as the risk presented to personnel is roughly proportionately less. For example, ionization chambers used as survey meters will be calibrated to be accurate to 10% at the one-third and two-thirds of full-scale deflection while GM counters may only be 50% accurate over their full-scale range. An instrument should provide precise reproducible readings. Instruments need to reproducibly repeat measurements at the same locations when used repeatedly in the same radiation fields. Portable instruments often feature rugged weatherproof designs with lightweight features, such as ergonomic antifatigue handles to facilitate outdoor use for long times. Individuals with various skill levels frequently use instruments; some personnel use instruments infrequently. In both situations, instruments subsequently suffer some misuse and abuse. Hence, simplicity of use is a highly desirable feature. Instruments should be designed with a minimum of controls or knobs to be adjusted, On, Off, and Battery Check switch positions must be clearly indicated and any scale selection switches should be labeled in an unambiguous manner. Some instruments feature audible signals whose intensity is proportional to the magnitude of the radiation signals. Such a feature often is useful on the most sensitive scale, but may be undesirable on the higher ranges; hence, usually the audible signal is switch selectable and may be turned off when desired. Potentiometer controls necessary during calibration adjustments and voltage setting control should not be so accessible that they can be easily changed. Such controls often are recessed or located on a rear panel so that they can only be changed in a deliberate manner. While every instrument will not possess all of the features discussed here, this discussion has included those found most commonly.

Compact, lightweight designs are easily achieved using microprocessors, liquid-crystal displays, modern CMOS electronics, and phenol, or acrylonitrile–butadiene–styrene (ABS) plastic cases. Some devices feature automatically backlight displays in low ambient light conditions. Current instruments frequently are available with either digital scales or analog scales; some offer both displaying a digital reading and an analogue bar graph emulating analogue meter movement. While digital scales are usually easily read in a constant radiation environment, they are inappropriate in rapidly changing radiation fields as the signal changes rapidly and rapidly changing digital readouts are difficult to read and interpret. For these situations, a freeze mode indicates a peak reading.

The radiation detector may physically be in the instrument with the associated electronics necessary to process the signals, connected to the counting electronics by a cable, or feature a remote reading capability, allowing the observer to stay in an area where radiation exposure is minimal. A cable connection is common in applications where the observer is physically in the radiation field monitored. Cable connectors, with instrument displays of the probe connected, allow the use of multiple probes or detectors (GM tubes, neutron probes, proportional counters, scintillation probes) with different features with a single count rate meter. Instruments with built-in detectors are free of cable problems, such as the loss of charge by poor cable insulation. Many devices feature an RS-232 interface or a universal serial bus (USB) cable that can connect to a computer; data software packages allow data retrieval, time–date stamps, or use parameter selections, such as programmable flashing displays and audible alarms, or measurements for specific applications. Some instruments feature data logging, the sequential capturing of hundreds or thousands of data points under different measurement conditions. Data captured by the instrument can be downloaded, by cable connection or by via infrared (IR) communication, to a personal computer for processing with a numerical spreadsheet (3).

Generally, instruments feature a battery check (a known scale deflection on an analog device or a brief

audible tone or indicator lamp on a digital device) that allows the user to determine that the battery possesses enough charge to operate the instrument successfully. Voltage stability is important; many instruments feature two power supplies, one for the counting electronics and another for the constant voltage required across the detecting volume. Currently, multiple 9 V alkaline batteries are widely used, providing 100–500 h of operation. The voltage across the detecting volume is usually required to be the most stable. A small variation in this voltage can lead to large changes in the observed signal, depending on the design and mode of operation of the instrument. Voltage stability of 0.1% is usually required for the voltage across the detector in most radiation detection instruments.

A zero check allows the zero point on the scale, previously set to zero in a radiation free environment, to be checked in the presence of radiation. Some devices feature auto-zeroing scales. Some instruments have an attached constancy source, a minute quantity of radioactive material, such as 0.06 μCi of ^{238}U or 10 μCi of ^{137}Cs, which, when placed on or near the detector in a predetermined geometry, yields a predetermined signal on the most sensitive range. Proper instrument use requires all three items, battery function, zero point, and known response, to be checked prior to each use. A reduced signal or complete loss of signal from the radiation protection instruments is particularly dangerous because the user falsely concludes that little or no radiation is present.

Radiation detectors generally are designed either to monitor individual events (counts) or pulses or to integrate (sum) counts or pulses that occur in such a short time interval that they cannot be electrically separated. In pulse mode, individual events or signals are resolved in 1 μs, 1 ns, or even smaller time intervals. In integrate mode, the quantity measured is the average of many individual events in some very short time period. Some devices feature signal integration when the device is used in a rate mode.

Response time of an instrument measures how rapidly an instrument responds to the radiation detected. The response time is short (fractions of a second) on the higher multiple scales and becomes longer (several seconds) as the scale multiple decreases with the longest resolving times occurring on the most sensitive scale. The response time (T) is called the time constant, and is proportional to the product of the resistance (R) of the electronic counting circuitry and its capacitance (C). (Fig. 1). Many instruments feature a slow response switch that allows electronic averaging of a rapidly varying scale signals.

Energy independence is desired for most radiation survey instruments, such as ionization chambers and GM counters; the signal is independent of the energy of radiation detected, but is proportional to the magnitude (counts, exposure, etc.) of the radiation field being monitored. However, many instruments exhibited a marked energy dependency at lower X or γ ray energies; the signal varies as the energy of the radiation varies even a constant magnitude radiation field (Fig. 2). Knowledge of the energy dependency of an instrument and of the approximate energy of the radiation field to be monitored is essential in properly using a radiation detector. Whether the signal from the

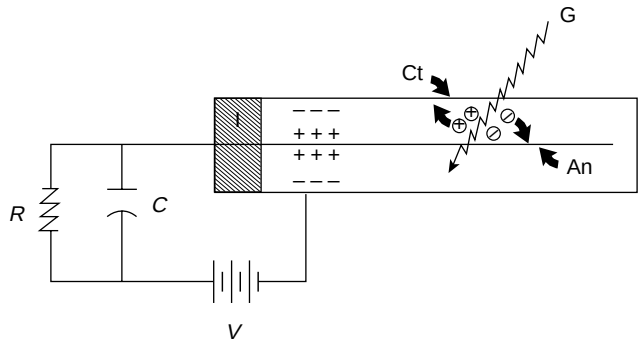


Figure 1. Simple schematic of a gas-filled ion chamber. A voltage (V) is maintained across the central wire anode (An) and chamber wall cathode (Ct). An incident (γ-ray (G) produces ion pairs; they move to the anode and cathode producing a pulse in the circuit containing a resistor (R) and capacitor (C).

instrument is higher or lower than it should be relative to the signal observed at its calibration energy depends on many parameters. Calibration of radiation detectors is required annually for some regulatory agencies; instruments usually display a calibration sticker indicating the most recent calibration date, the calibration source, or sources if several were used to obtain energy response of the instrument, the scale readings obtained (often in $mSv \cdot h^{-1}$, $mR \cdot h^{-1}$, or other multiples, or cpm), and the accuracy and precision of those readings, expressed as a percentage of the scale reading, any necessary scale correction factors to be used specific scales, and instrument response to a reference source containing a minute quantity of radioactive material. By performing a battery check

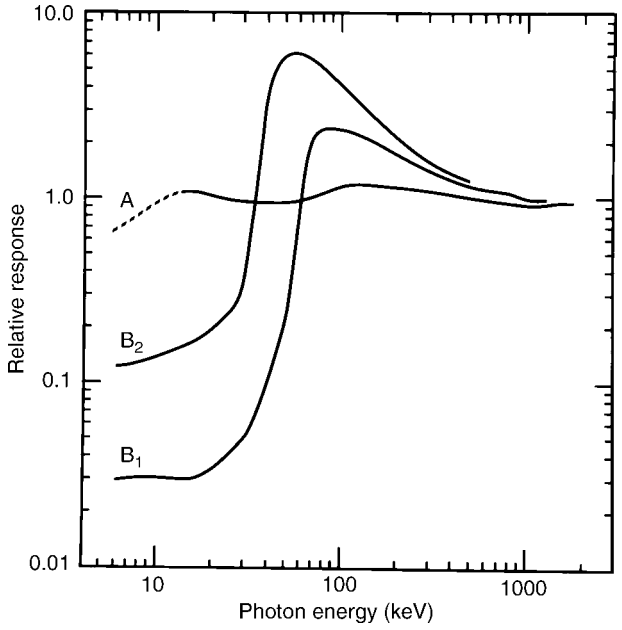


Figure 2. Typical relative response versus incident photon energy (kV): (A) is the ionization chamber; (B_1) is a Geiger counter with thin window shield closed. (B_2) is a Geiger counter with thin window shield open.

or by observing that the battery condition is acceptable by the absence of a low battery indicator, a zero-scale reading check, and meter response check using a reference source, the user can determine that the instrument is operating correctly before use. Moreover, the average energy of radiation at a particular location in an area is highly dependent on the relative amounts of primary and scattered radiation present and frequently varies within an area. Energy dependency is, of course, advantageous when it is desirable to measure the energy spectrum in addition to determine the intensity of radiation.

Some instruments are environmentally sensitive; graphs or tables providing correction factors as a function of temperature, pressure, and humidity indicate the degree to which the signal is altered by environmental conditions. For instruments with the detector volume open to the air, corrections based on thermodynamic gas laws for the mass of air present in the detector are employed.

Excessive humidity can cause incorrect instrument readings. Humidity can cause leakage of current in cables, at electrical contacts, and at other locations in the electronic circuitry. Usually, an instrument will require a warm-up time of 1–2 min or longer. Proper warm-up allows electronic circuitry to stabilize and yields more stable and reproducible signal readings.

Strong radio frequency (RF) fields associated with some equipment generating radiation can cause improper signals in some radiation measurement instruments. The susceptibility of the instrument to strong rf fields will usually be discussed in the user's manual.

Many radiation protection instruments exhibit geotropism, orientation (gravitational) dependency, or angular dependency because radiation incident from the sides and rear of the instrument are attenuated more by metal casing surrounding the counting electronics than radiation incident on the sensitive detecting volume. The proper orientation of the instrument for measurement in a radiation field and the degree of angular response will be indicated in the users manual or on the calibration certificate.

As previously noted, some devices are designed to identify different types of radiation or to identify species of radionuclides. Many detectors will feature a thin detection window of only a few milligrams per square centimeters of thickness protected by a thicker filter, a sliding, or rotating shield, that allows the least energetic forms of radiation, such as soft X rays and alpha and beta particles to be detected through the thin window when the thicker filter is removed. Conversely, with the shield in place alpha and beta particles are discriminated against and only higher energy radiations are detected. Other windowless detectors are designed to detect the low energy radiation by flowing a radioactive gas through the detector.

Radionuclide identification requires spectroscopy, the identification of the characteristic radiation spectra of multiple radionuclides each in the presence of others. Resolution is the measure of the abilities of devices to distinguish a single energy in multiple energy radiation spectra. Different applications require instruments with different degrees of resolution. Spectroscopy formerly was limited to using heavy fixed laboratory-based NaI(Tl) detectors or Ge(Li) detectors often with a multichannel analyzer to quantify and identify radionuclides in test samples. Currently, in-field spectroscopy can be performed with small handheld or portable NaI(Tl)-based detectors or with portable high purity germanium (HPGe) detectors that allow radionuclide identification of the most common radionuclides.

IONIZATION CHAMBERS

The ionization chamber (Fig. 1) consists of a cavity, frequently cylindrical, with a positively charged central electrode (anode) insulated from the chamber walls (cathode) at negative potential. The direct reading pocket dosimeter, with an external dosimeter charger (Fig. 3) is a simple ionization chamber. When fully charged, an internal quartz fiber, visible under a magnification lens, is deflected to a "zero" reading. As the dosimeter is irradiated, the charge is reduced proportionally to the amount of radiation received. Older pocket chambers are being replaced with personnel detectors or monitors with more electronic versatility. As a survey meter, an external power source (Fig. 1) provides the voltage potential; a resistor and capacitor in parallel (or equivalent electronic circuitry) are used to collect the charge produced when ionization occurs in the chamber. The ionization chamber electrode polarity may be reversed for special applications. Incident X or γ rays interact in air or a tissue equivalent gas, producing positive and negative ions in the chamber. If the voltage is sufficiently high to prevent recombination, that is, the positive and negative ions rejoining before they reach the charged surfaces, the negative ions will be attracted to the central electrode and the positive ions will be collected on the chamber wall. The collected charge flows to the capacitor and one electronic pulse is detected in the counting circuitry. In open air chambers, the filling gas is air at ambient temperature and pressure and appropriate corrections to the charge collected may be required as previously discussed.

Historically, ionization chambers were designed to measure exposure (R); newer instruments may offer equivalent dose readings (Sv or their submultiples) (Fig. 4). The walls of the chamber must be sufficiently thick for electronic equilibrium to be established, that is, the number of electrons entering and leaving the cavity is the same and the chamber walls are sufficiently thick to stop any electrons arising from the interaction of the radiations with the gas or in the chamber walls. Moreover, the chamber walls are usually designed of air equivalent materials. Sealed ionization chambers may be filled with a tissue equivalent gas and usually are designed to measure collision air kerma. The chamber size must be small relative to the dimensions of the irradiating beam so the chamber is uniformly irradiated. Typical ionizing voltages required across the sensitive detecting volume are $\sim$ 150–300 V (Fig. 5), sufficiently high to present recombination of the positive and negative ions, but not high enough to cause additional ionizations that amplify the signal. Ion chamber currents are low, usually 1 pA or 1 fA. A 10 mSv·h^{-1} γ ray field yields $\sim$ 1 pA; extraneous currents must be minimized in order to

Figure 3. Personnel dosimeters. Low dose (bottom left; Dosimeter Corp., Model 862;) and high dose (bottom right; Dosimeter Corp., Model 866) γ- and X-ray pocket dosimeters, with charger (top left; Jordan Nuclear Co, Model 750-5). An alarming personal digital dose meter (center; Technical Associates, Model PDA-2) and a miniature pocket digital dosimeter (right; Aloka Co, LTD., Model MYDOSE-mini).

measure such a small current. Electrical leakages can occur across lint, dust, or loose conductive materials between interior conducting surfaces. Cable connections can exhibit greater leakage current in high humidity. A guard ring design in some chambers minimizes leakage and polarization currents that arise after the collection potential is initially turned on. The insulator between the outer and inner electrodes is divided into two segments with the conductive guard ring in between; any leakage current through the insulator is collected and prevented from contributing to the true current. The currents from ionization chambers are normally measured using a potential drop across a high resistor or a rate of charge method. The small currents from the ion chamber are amplified by a vibrating reed electrometer. The amplified

Figure 4. A portable ionization chamber survey meter (Cardinal Health; Inovision, Model 451). The display shows the features during the initial check phase immediately after turning on the instrument.

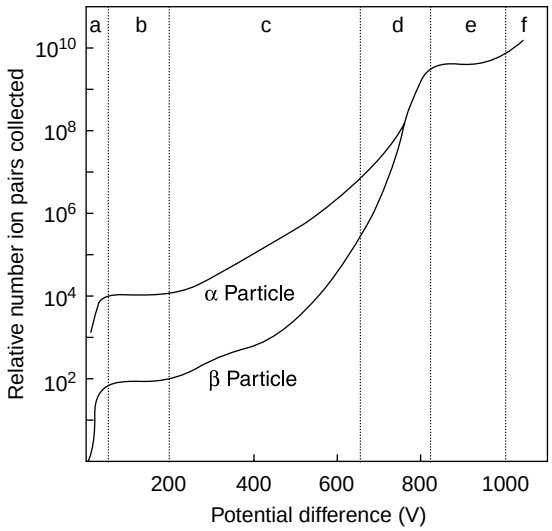


Figure 5. Voltage dependence of a gas-filled cylindrical ionization chamber: (a) Voltage is insufficient to prevent ion recombination. (b) Ionization chamber voltages are sufficient to prevent recombination. (c) Proportional counter voltages, the number of secondary ion pairs is proportional to the number of primary ion pairs. (d) Limited proportionality region. (e) Geiger voltages produce maximum number of ion pairs from a single primary ion pair. (f)Continuous discharge region.

current passes through a precision resister and the voltage drop across the resistor is proportional to the current. If collected on a capacitor, the rate of charge collected on the capacitor is proportional to the current. This latter method is used for smaller current measurements while the former is used in ionization chambers designed for more rugged use.

Ionization chambers normally exhibit good energy independence (Fig. 2) over a large energy range, and this makes them useful for measurement of X or γ rays with energies from $\sim$ 7–1000 keV and for measuring low (1 $\mu Sv \cdot h^{-1}$) to high (10 $mSv \cdot h^{-1}$ or higher multiples) dose rates. Open-air ionization chambers are less useful for low dose rates of < 1 $\mu Sv \cdot h^{-1}$. Pressurized (up to 8 atm) ionization chambers allow accurate measurements < 1 $\mu Sv \cdot h^{-1}$. Properly modified ionization chambers, with sliding shields, may be used to monitor alpha, beta, and neutron radiation. For example, an ion chamber with boron on its interior chamber wall or containing boron gas may utilize the high cross-section of boron for neutrons and the subsequent ^{10}B, ^{7}Li reaction, and the chamber will detect neutrons using the subsequent alpha particles from this reaction.

GAS PROPORTIONAL COUNTERS

Gas proportional counters (Fig. 6) have similar design features as ionization chambers, but employ higher voltages between the central electrode and the chamber walls. The typical operating voltages of 300 up to 1000 V (Fig. 5) are sufficiently high that, following an ionizing event in the chamber, the positive and negative ions generate additional ionizations so that the number of ions from the initial ionizing events are multiplied $\sim$ 1 thousand to 1 million times. The resulting signal is proportional to the energy deposited by the initial number of ionizing events. Proportional chambers can be used in either the pulse or integrate mode, but the pulse mode is used most commonly. They are capable of detecting individual ionizing events. Because of amplification, the current from proportional chambers is much higher than those from ionization chambers. As the signal from a proportional current is dependent on the operating voltage, a highly stable power supply is required. The choice of detector gas in thin-window proportional counters depends on the type of radiation to be detected. For counting alpha particles, helium or argon gas frequently is used. For counting beta particles, a high multiplication gas is required, such as methane (CH_4) or a mixture of a polyatomic gas and a rare gas, such as argon. The gasses also help make the proportionally of the chamber more independent of operating voltage. Proportional counters are generally cylindrical in shape and the central electrode is a very fine wire of uniform diameter as any variation in the electrode's diameter causes small variations in the resulting signal. Gas (often a mixture of 90% argon and 10% methane) flow proportional counters usually have a sample of the radioactive material flow through the chamber. Either 2π (180°) or 4π (360°) solid geometries are used, and these systems are very useful for counting low energy beta particles, alpha particles, or very low energy photons. Proportional counters are useful for spectroscopy (energy determination) measurements.

Proportional counters have the ability to discriminate between alpha and beta particles by discriminating between the magnitudes of the signals produced. Gas proportional counters may be used to measure fluence or absorbed dose.

When neutron spectra are poorly known, neutron rem meters are used to estimate the equivalent dose for fast neutrons. Older style neutron detectors consisted of a proportional counter either lined with boron or filled with boron trifluoride gas; the boron has a high capture cross-section thermal neutron detector. The subsequent charged

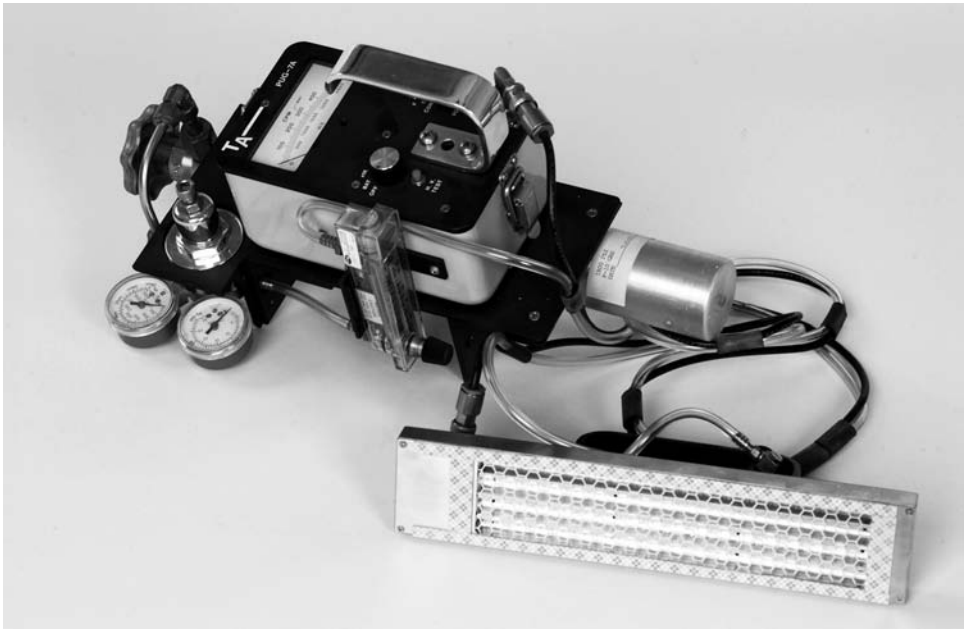


Figure 6. Portable gas proportional counter with alpha probe (Technical Associates, Model PUG-7A).

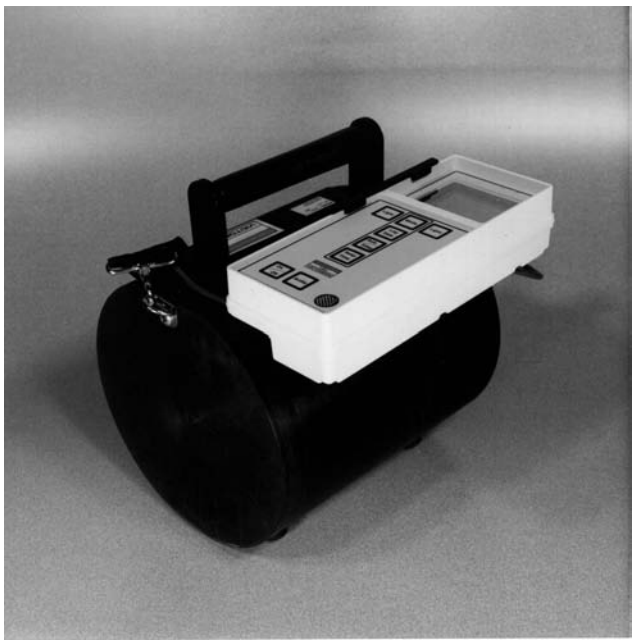


Figure 7. Portable neutron survey meter (Cardinal Health, Inovision, Model190N). (Courtesy Cardinal Health.)

particles (alpha particles) from this reaction are readily counted. Current neutron detectors use ^{3}He as the fill gas and detect both the proton and tritium, ^{3}H, from the subsequent reaction (Fig. 7). Fast neutrons can be moderated in several centimeters of high density polyethylene to thermalize the neutrons for detection by the methods described. Olsher et al. (15) described recent improvements in neutron rem meter instrumentation.

GEIGER–MÜELLER COUNTERS

If the voltage on an ion chamber is increased to $\sim$ 900–1000 V (Fig. 5), the proportionately exhibited at lower voltages is lost. Each initial radiation interaction in the walls or gas of the detector results in complete ionization of the gas in the detector. Interactions in the detector are spatially dependent, but generally, the following sequence occurs. Electrons produced following the initial ionizing event lose energy as they drift toward the anode. They lack enough energy to produce secondary ionization until they approach the anode when secondary ionization begins to occur. This secondary ionization builds up rapidly producing an avalanche of electrical charge in the detector. These processes reduce the potential difference between the central electrode and the chamber walls and the avalanche terminates. Once the necessary ionizing potential is reestablished, the detector is ready again. One undesirable aspect of the movement of the positive ions to the cathode and their resulting collisions with the cathode causes additional electrons to be ejected from the cathode. These additional electrons are undesired and may be controlled by manufacturing a tube containing a quenching and a filling gas. Organic quenching gases, such as ethanol or ethyl formate, are depleted by this process. An inorganic filling gas, such as chlorine, recombines to provide a continuous supply (8). The energy of the undesired electrons dissociates these organic molecules rather than starting new avalanches in the tube. The number of organic molecules available for quenching limits this method of quenching. An alternative method uses halogen gases, usually bromide or chlorine. The extra energy of these electrons is used to disassociate these halogen molecules. As opposed to the organic molecules, halogen reassociates so that the same atoms are available again to continue the process.

Geiger–Müeller tubes generally are used as pulse-type detectors of radiation. Their response is a function of the intensity of the radiation field. The movement of the positive ions to the cathode, previously described, requires from 100 to 200 μs and during this time interval the GM tube is unable to respond to other radiation interactions, have not recovered (reassociated) and are unable to resolve additional events (7). In very intense radiation fields, the relative long resolving times of GM tubes creates periods in which the tube is insensitive to radiation events; the GM tube may not respond to radiation, giving a false zero or low reading when an intense field is present. However, GM tubes are excellent as very sensitive detectors of X and γ rays in low level radiation fields. Commercial manufacturers offer at least three different GM tubes designs (Figs. 8,9) for specific applications. Pancake probes, with covers only a few milligrams per centimeter squared thick, allow the detection of alpha particles $>$ 3.5 MeV, beta particles $>$ 35 keV, and γ-rays $>$ 6 keV, while thin end window probes detect alpha particles $>$ 4 MeV, beta particles $>$ 70 keV, and γ rays $>$ 6 keV (Fig. 8). Some pancake-type probes (Fig. 9) feature removable tin and copper filters $\sim$ 3 mm thick that allow energy compensated exposure rates measurements. Energy-compensated GM probes feature a design that reduces response energy dependency so that it responds more like an ionization chamber (Fig. 5). Geiger–Müeller instruments are basically count rate meters, but may be calibrated in exposure rate for a specified energy of photons. Use of the meter in other energy spectrums different from the calibration spectrum invalidates the meter reading in μSv/h and mSv/h, but still the instrument allows the detection of radiation in the count rate mode.

SCINTILLATION DETECTORS

Luminescence is a physical process in which a substance, a scintillator, absorbs energy and then reemits the energy in the visible or near visible energy range. Prompt scintillators that deexcite in 10 ns following luminescence exhibit many useful properties as radiation detectors. For every photon or particle detected, a single pulse is normally counted and the size of the pulse generated is related to the energy deposited by the radiation interacting in the scintillator. Scintillators exhibit great sensitivity and yield high count rates. They can measure fluence, exposure, or absorbed dose if calibrated for the energy range of interest. Moreover, their exceptional sensitivity allows measurement of radiation rates at or near background levels, such as 1 nSv$\cdot$h^{-1}. Solid inorganic scintillators includes sodium

Figure 8. Portable GM counter (top; Ludlum Measurements, Inc. Model 14C) with an open side-window probe (left; Ludlum Measurements, Inc. Model 44-38), an end thin-window probe (center; Ludlum Measurements, Inc., Model 44-7), and a pancake probe (right; Ludlum Measurements, Inc., Model 44-9).

iodide crystals with trace amounts of thallium, NaI(Tl); cesium iodide with thallium, Csl(Tl); cesium fluoride, CsF; zinc sulfide with silver, ZnS(Ag); and bismuth germanium oxide, BiGeO, also known as BGO. The trace amounts of impurities in these inorganic salt crystals serves as luminescent process activators that promote the efficient conversion of the incident radiation energy into light. The scintillator crystal is connected to a photomultiplier by direct contact or through a light pipe. The crystal and photomultiplier must be encased in a light tight case to prevent light leaks. Typical crystals are cylindrical, ~ 1 in. (2.54 cm) diameter by 1 in. (2.54 cm) thick, but larger sizes (Fig. 10) are available for more sensitive measurements. The resulting photomultiplier signal is amplified by the associated counting electronics. Detectors with thin windows are available for the detection of low energy X rays and energetic beta particles. Inorganic solid crystals are relatively dense and reasonably efficient for detecting higher energy photons (Fig. 11). However, they are also hydroscopic and to protect them from absorbing moisture are encased in light reflecting cases that promote good efficiency. Organic crystal scintillators produce their light by a molecular process. Anthracene and transtiblene are the most widely used organic crystal scintillators. Incoming radiation excites electrons to higher energy levels of vibrational states; the electrons subsequentially decay with a release of energy. Organic liquid scintillators are formed by dissolving organic scintillators in liquid organic

Figure 9. A counter (center; Cardinal Health; Victoreen, Model 190) with an energy-compensated sliding window GM probe (left; Cardinal Health; Inovision, Model 90-12), a pancake detector with filters (center; Cardinal Health; Inovision, Model 489-118FS), and a 1×1 in. NaI(Tl) detector (right; Cardinal Health; Inovision, Model 425-110).

Figure 10. The NaI(Tl) detectors: 1×1 in. (left; Cardinal Health; Inovision, Model 425-110); 2×2 in. (5.08 cm) with center well (center; Nuclear Chicago, Model 321330), and 3×3 in. (2.62 cm) (right; Picker Nuclear Omniprobe, Model 2830-A).

Figure 11. A γ-counter system with a 1×1 in. (2.54 cm) NaI(Tl) detector to identify and measure γ rays. (Canberra; Packard, Model Cobra II Auto-Gamma.)

solvents, such as xylene, toluene, and phenylcyclohexane. A wave shifter fluorescent material shifts the wavelength of the light from the main solute to a longer wavelength and lower energy, so that the wavelength more closely matches the spectral response of the photocathode. Liquid organic scintillators are widely used because the sources of ionizing radiation can be dissolved into the solvent and made a part of the scintillator solution. Low energy beta emitters' tritium, ^{3}H (19 keV), and ^{14}C (156 keV), are counted with high efficiency by these methods (Fig. 12). Modern pulse processing methods allow separation of alpha and beta events.

Plastic scintillators consist of organic scintillators that have been dissolved in a solvent and the existing solvent polymerized to form plastic scintillators. As plastics can be made ultrathin, they can be useful for detecting low energy particles of radiation in the presence of gamma rays or for

Figure 12. A liquid scintillation counter system for beta particles (Canberra; Packard, Model B1500 Tri-Carb).

detecting heavy particles. Plastic scintillators are available in many physical configurations. Neutrons can be defected by incorporating a neutron sensitive material, such as lithium, into the solvent and the subsequent plastic scintillator. Nobel gas scintillators consist of high purity concentration of helium or xenon, which have the property that, following radiation interaction in the gas, both visible and ultraviolet (UV) light is emitted. While these materials exhibit a very short deexcitation time of 1 ns, they yield little light and the conversion of light is reasonably inefficient; nevertheless, they do have numerous applications when a fast response time is required.

As with GM probes, scintillation detectors are available as rectangular pancake probes to detect beta particles > 100 keV and γ rays > 25 keV, thin scintillators to detect γ and X rays > 10 keV, flashlight-like probes to detect alpha particles > 350 keV and beta particles > 14 keV, and conventional thick-crystal cylindrical probes for γ and X rays > 60 keV.

Photocathodes have many applications in devices that measure or detect radiation, such as image intensifiers, vidicon tubes, and other detectors. Usually, the electronics required to amplify the initial signal does not have as short a resolving time as the detector proper, but with modern solid-state electronics, the resolving times have been shortened to less than microseconds. Because of their extreme sensitivity, scintillator detectors are useful when detection and subsequent identification, by spectral analysis, of a type of radioactivity is required (Fig. 13). While the spectral peaks associated with scintillators are reasonably broad, their energy resolution is sufficient to allow rapid identification of minute quantities of various radioisotopes

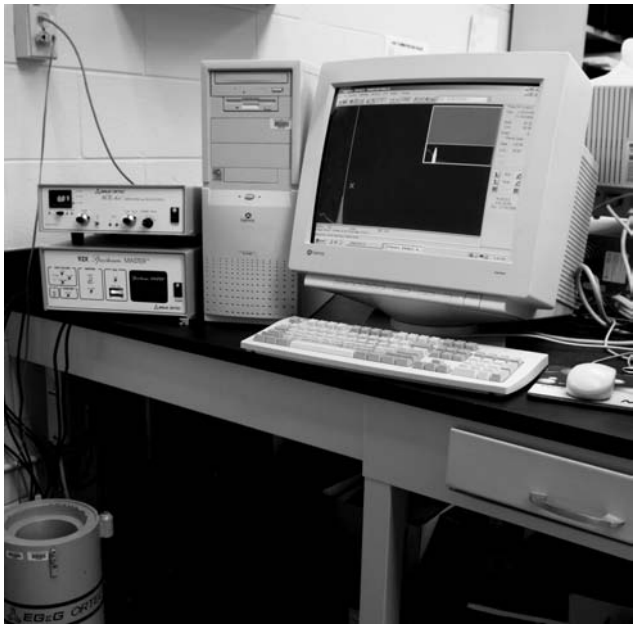


Figure 13. A multiple component NaI(Tl) spectroscopy system. Detector shield (lower left; Ortec) with a NaI(Tl) detector (not shown); amplifier/bias supply(upper center; Ortec, Model Acemate); spectral analyzer (lower center; Ortec, Model 92X Spectrum Master), and computer display (left).

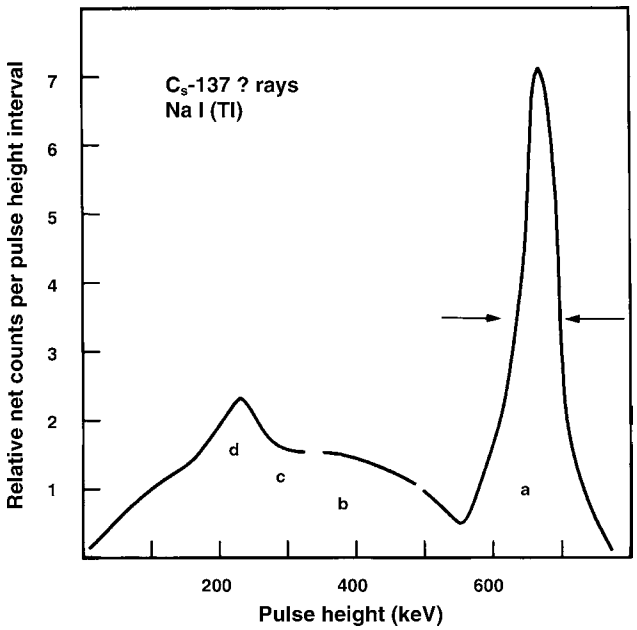


Figure 14. A spectrum measured with a 1×1 in. (2.54 cm) NaI(Tl) dectector: (a) ^{137}Cs full-energy (0.66 MeV) peak; (b) Compton edge (0.48 MeV); (c) Compton distribution; (d) backscatter peak (0.18 MeV).

(Fig. 14). Formerly limited to laboratory analysis, NaI(Tl) spectroscopy is now available in portable handheld units (Fig. 15) that, using quadratic compression conversion (QCC), can identify 128 radionuclides in real-time (1 s intervals). Quadratic compression conversion creates spectral energy peaks whose widths vary proportionally to the NaI(Tl) crystal's energy resolution. All energy peaks are displays with the same peak width that allows radionuclides' distinct spectra to be more readily identified. The electronics associated with scintillators must be extremely stable, but often variations are introduced by environmental factors. High permeability magnetic shielding materials, for example, Mu-metal, often are used to shield against stray magnetic fields. Temperature and humidity variation can produce undesirable electronic noise. As previously noted, some crystals are hydroscopic and the moisture absorbed can reduce the efficiency of the process. Pulse discrimination techniques are often used to distinguish one type of radiation from another.

SOLID-STATE RADIATION DETECTORS

Thermoluminescence (TL) is the emission of visible light released by heating previously irradiated solid-state crystals. The light emitted from a thermoluminescent crystal is proportional to the amount of radiation to which the crystal has been exposed, and this proportionality holds over a large range (10^2–10^5) of exposures. At very high exposures, nonlinearity is exhibited and the amount of visible light released is no longer proportional to the amount of radiation detected. Thermoluminescent dosimetry (TLD) materials (Table 5) commonly used in medicine, include lithium fluoride (LiF), available in three forms (TLD-100, -600, -700,), and lithium borate manganese (Li_2B4O_7:Mn) (TLD-800).

Figure 15. A portable NaI(Tl) surveillance and measurement (spectroscopy) system. (Berkeley Nucleonics Corp., SAM Model 935).

Other TLDs used in environmental dosimetry applications, calcium fluoride manganese (CaF:Mn) (TLD-400), calcium fluoride dysprosium (CaF_2:Dy) (TLD-200), calcium sulfate dysprosium ($CaS0_4$:Dy) (TLD-900), and aluminum oxide (Al_2O_3:C) (TLD-500), are not further described here. X rays, γ rays, and neutrons are easily detected with different TLD materials; the detection of higher energy beta particles is possible, but quantification of the amount of beta radiation and calibration of the solid-state detectors for beta radiation is often more difficult than for X and γ rays. The TL materials exhibit an enhanced response to lower energy (< 200 keV) X or γ rays as compared to higher energy (1 MeV) X or γ rays. For LiF, the over response is only a factor of 1.2, but for lithium borate manganese ($Li_2B_4O_7$:Mn) (TLD-800), there is an under response of ~ 0.9. By adding filters, the energy response of a given type of crystal can be made more uniform and this is commonly done with TLD detectors used as personnel radiation monitors (Table 5).

Table 5. Properties of Some Thermoluminescent Materials

Property/type	LiF:Mg, Ti (TLD-100)	^{6}LiF:Mg, Ti(TLD-600)	^{7}LiF:Mg, Ti (TLD-700)	$LI_2B_4O_7$:Mn (TLD-800)
Applications	Health and medical dosimetry	Neutron dosimetry	Gamma dosimetry	Neutron dosimetry
Relative concentrations	^{6}Li (7.5%) ^{7}Li (92.5%)	^{6}Li (95.6%) ^{7}Li (4.4%)	^{6}Li (0.007%) ^{7}Li (99.993%)	NA[a]
Density (g mL^{-1})	~ 2.6 (ribbons) ~ 1.3 (powder)	2.64	2.64	~ 2.4 (ribbons) ~ 1.2 (powder)
Effective Z for photoelectric absorption	8.2	8.2	8.2	7.4
TL Emission spectra	350–600 nm (400 nm max)	250–600 nm	350–600 nm (400 nm max)	530–630 nm (605 nm max)
Temperature of main TL glow peak	195 °C	195 °C	195 °C	200 °C
Efficiency at ^{60}Co relative to LiF	1.0	1.0	1.0	0.15
Energy response 30 keV/^{60}Co	1.25	1.25	1.25	0.9
Useful range	mR–3 $\times$ 10^5 R	mR–10^5 R	mR–3 $\times$ 10^5 R	50mR–10^6 R
Fading	Negligible* 5%/years at 20 °C	5%/year	5%/year	$< 5\%$ in 3 months
Preirradiation anneal	400 °C at 1 h + (100 °C 2 h or 80 °C at 16 h)	400 °C at 1 h + (100 °C at 2 h or 80 °C at 16 h)	400 °C at 1 h + (100 °C at 2 h or 80 °C at 16 h)	300 °C at 15 min
Postirradiation anneal	100 °C at 10 min	100 °C at 10 min	100 °C at 10 min	
Special feature	Low dose rate dependence	Highly sensitive to thermal neutrons	Insensitive to neutrons	High dose dosimetry

[a]Not available = NA.

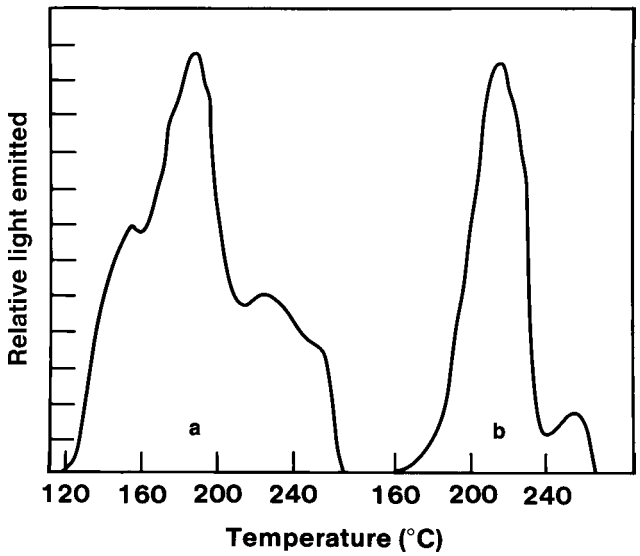


Figure 16. Glow cures for TLD-100 (natural LiF) rods exposed to 10 R. (a) Without proper preparation annealing, multiple natural peaks occur; (b) Using a 1 h annealing at 400 °C and a 2 h annealing at 100 °C a smoother curve is obtained.

Physically, TLD materials consist of loose powder contained or embedded in plastic holders, compressed crystals (chips), extruded rods, and chips on a card in a configuration that allows reproducible heating of the detector materials to selected temperatures. The TL materials exhibit some undesirable features as radiation detectors. At room temperature, fading or loss of signals occurs, from $< 0.5\%$ per month for LiF to 5% in 3 months for $Li_2 B_4 O_7$:Mn. The degree of fading can be controlled, to some extent, by proper preparation (annealing) procedures.

The optical readers generally consist of a heating pan or device that allows the TL material to be uniformly heated in a controlled manner, at a specified temperature for a given period of time. The heating device and material holder are directly below a photomultiplier tube that usually has some filters to remove any IR radiations and transmits light in the blue-green region of the visible spectrum. The signal from the photomultiplier is amplified and used to prepare a glow curve (Fig. 16) a plot of the intensity of light versus the heating cycle of heating elements. Different TL materials exhibit different glow peaks. Numerous peaks occur in the curve and either the area under the curve or the height of the major peak is chosen to be proportional to the amount of radiation to which the material was exposed. Proper preannealing (extended heating at a controlled temperature) and post-annealing of the material will remove some smaller undesirable peaks (Fig. 16) leaving the main peak that is used for measurement purposes.

Current TL readers offer automatic features for glow curve analysis and processing of large numbers of samples (chips, rods, or cards).

Lithium flouride is the most commonly used TL material for personnel dosimetry and consists of natural lithium. Lithium-6 is preferentially sensitive to thermal neutrons and ^{7}Li is insensitive to thermal neutrons, but sensitive to γ-rays. Hence, by using paired ^{6}Li and ^{7}Li materials, it is possible to measure thermal neutrons in the presence of γ rays. Use of natural LiF detectors in radiation environments that contain low levels of thermal neutrons will yield incorrect dose equivalents for personnel, as the thermal neutrons will cause an apparent over response of LiF calibrated only to detect and measure γ rays (16).

Optically stimulated luminescence (OSL) is the release of light by a phosphor following its irradiation by a laser. Aluminum oxide (Al_2O_3) containing carbon impurities exhibit OSL releasing a blue light when excited by a green laser light. Some personnel radiation monitors employ OSL technology that offers some improvement over TLD-based personnel radiation dosimeters (17). The OSL dosimeter offers greater sensitivity, stability, and accuracy than TLD dosimeters. Aluminum oxide is highly linear from 1 mSv to 10 Sv; there is little signal fading > 1 year. It does exhibit an energy dependency below ~ 100 keV. However, unlike TLD chips, the aluminum oxide element can be reread multiple times to confirm an initial reading, an advantage for personnel dosimeter applications. The Luxel badge (Fig. 17; Table 6) contains the Al_2O_3 phosphor element, with 20 mg·cm^{-2} open filter (paper wrapper), 182 mg·cm^{-2} copper filter, and 372 mg·cm^{-2} tin filter in a heat-sealed, light-tight hexagonal plastic badge (18). The combination allows detection of beta particles > 150 keV with a 100 μSv threshold and X and γ rays > 5 keV with a 10 μSv threshold.

Photoluminescence (PL) occurs when the irradiated crystal emits visible light when exposed to UV light instead of heat. Silver activated glass encapsulated PL detectors are available in numerous shapes, sizes, and radiation levels as low as 100 μSv are detectable, but the detectors are commonly used to detect higher exposures. Appropriate filters can be used to make the energy response more linear, but at exposures of 0.1 Sv or higher the response of these detectors is nonlinear. These materials exhibit some signal fading that depends on the composition of the glass. Heating the glass detectors post postirradiation for 30–60 min at 150 °C yields maximum luminescence. Reannealing requires 40 °C for 1 h, which restores the material to its preirradiation state. For low level exposure measurements, care is required to keep glass detectors free of dirt, dust, and other materials that would reduce the amount of light transmitted through the detector.

Semiconductor materials used for radiation detection, Si, Ge, CdTe, HgI_2, and GaAs, have band gaps, the region between the valence and conducting band, of < 2.2 eV. Electrons migrate from the valance to the conduction band, leaving holes, that act like positively charged electrons, in the valence band. In a p-type semiconductor, the current is carried by the positively charged holes; in an n-type semiconductor, current is carried by the electrons. Usually, a potential difference is maintained across the solid-state semiconductor such that the depletion layer is devoid of electrons and holes. The interaction of X rays, γ rays, alpha particles, or beta particles generates additional electrons in the depletion layer; these are then swept away by the potential difference across the material, yielding a small current whose magnitude is proportional to the intensity of the incident radiation Energy required to generate

Figure 17. The Luxel Dosimeter: Holder (top left); front view (top center); rear view (top right); rear view with attached Neutron 144 detector (bottom left); detector element showing Cu filter (right), the Al filter (left), and plastic filter (circle, lower center). (Courtesy Landauer, Inc.)

electron–hole pairs range from ~ 3 to 6.5 eV. Semiconductors exhibit excellent linear response over a large energy range and greater efficiency than gas detectors. (For a full description of semiconductor physics, see Refs. 7 and 8.)

Surface-barrier detectors essentially consist of an n-type and p-type layers that function as anode and cathode, with an intrinsic I layer (depletion region), without electrons, in between, in which radiation interactions occur. Depletion regions range from 0.1 to 1 mm. The composite is commonly called P-I-N structure. With a moderate reverse bias applied, radiation interactions create electron–hole pairs with each electron and hole leaving, or depleting, the intrinsic layer, and, with appropriate circuitry, creating a detection and counting system.

Silicon surface-barrier diode counters are used for charged-particle (alpha and beta particle) detectors. Alpha resolution ranges from 12 to 35 keV; beta resolution ranges from 6 to 30 keV. Passivated Implanted Planar Silicon (PIPS) detectors use implanted contacts more rugged than conventional surface-barrier contacts. They can be costumed-designed for specific applications.

Germanium (Ge) detectors are used for γ- X-ray spectroscopy to identify not only the amount of radiation present, but also to identify the type of radioactivity present. These detectors have excellent energy resolution and are used to identify the individual X and γ rays from a radioisotope, and with peak fitting programs have the ability to resolve closely lying peaks (Fig. 18). Lithium drifted geranium, Ge(Li), detectors have been replaced by high purity pure geranium (HPGe) detectors that only required liquid-nitrogen cooling (Fig. 19) or electrically refrigerated cryostats during measurements or use; otherwise they are kept at room temperature. Lithium drifted silicon, Si(Li), detectors are still used. With microprocessors, these instruments are now used for on-site identification of samples that may contain several different radioisotopes, a process previously limited to the laboratory. They are important tools in research facilities where minute quantities of many

Table 6. Parameters of Some Commercial Personnel Dosimeters[a]

Monitor Designation	Sensitive Element	Primary Filter and Thickness, mg cm^{-2}	Radiation Detected and Detection Threshold, mSv
Gardray	Film or 4-chip TLD	Film wrapper (35) Plastic (325) Aluminum (375) Lead (1600) Cadmium (1600)[b]	β(0.4) X, γ(0.1) Thermal neutron (0.1)
T	2 TLD chips	Plastic (75) Plastic (200)	β(0.4) X, γ(0.1)
Neutrak 144	TLD-600 TLD-700 CR-39	Cadmium (660)	Thermal neutrons (0.1) Fast neutrons (0.2)
Luxel	AI_2O_3	Open (20) Copper (182) Tin (372)	β (0.1) X,γ (0.01)

[a]Courtesy of Landauer, Inc. Parameters quoted are those listed in the company's advertisement.
[b]Cadmium filter provided with film systems only.

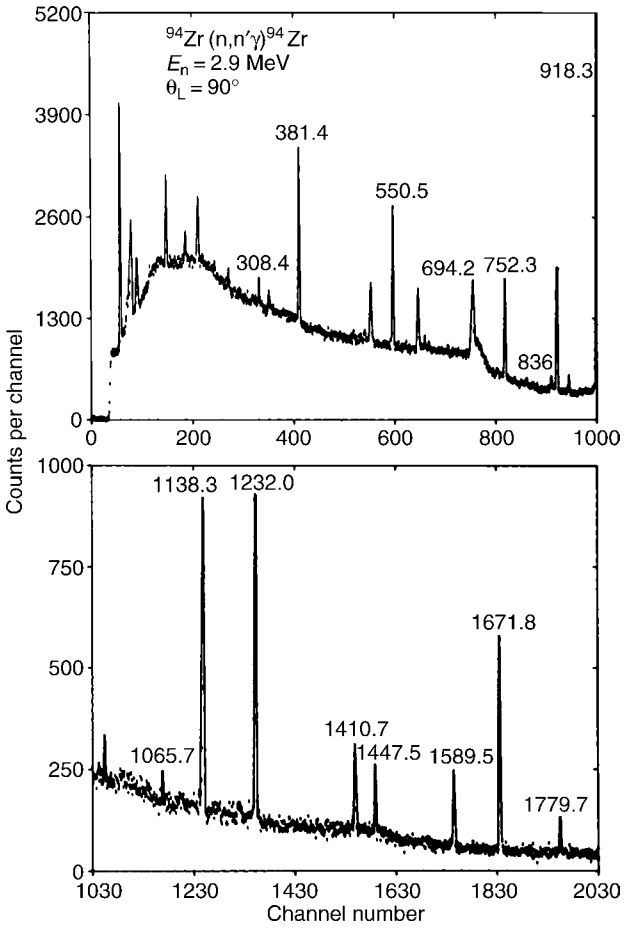


Figure 18. A γ-ray spectrum measured with a 35 cm Ge(Li) detector with good energy resolution.

different radioisotopes may potentially accumulate and radioisotope identification is required.

Cadmium telluridide (CdTe) detectors are popular as they are hygroscopic, do not require a photomultiplier, require only a 50 V bias, and operate at room temperatures. They exhibit high sensitivity, but their energy resolution is intermediate between that of NaI(Tl) and Ge detectors. They are available in a multitude of small sizes for special applications.

OTHER DETECTORS

Other physical changes that arise in materials as a result of irradiation include coloration and nuclear activation. Color changes occurs in some materials following their irradiation. Thin films with a cyanide emulsion (GAFCHROMIC EBT) will exhibit a deep orange color following their irradiation by γ rays to a dose of 10–100 Gy (19). These materials exhibit excellent linearity > 1–800 cGy of radiation dose. While developed for high dose radiation dosimetry studies on linear accelerators, they can be used for high dose radiation protection studies.

Neutrons may be detected by numerous nuclear reactions or by the process of counting the number of recoil proton tracks produced in certain neutron sensitive materials, materials, such as boron (20). A polycarbonate material CR-39 (allyl diglycol carbonate) is insensitive to X rays, γ rays, and beta particles. However, incident neutrons collide with the protons, which produced changed particle tracks; the tracks are enhanced by chemically etching the polycarbonate, so they will be more visible under a microscope. This technology is used as one component in a composite personnel dosimeter used to monitor radiation therapy personnel working around linear accelerators with X-ray energies > 10 MeV (Table 6) (18).

Superheated Drop Detectors (SDD) consists of a small container of gel holding superheated drops ~ 0.1 mm diameter. Neutrons produce recoil protons that strike the drops, causing them to vaporize, generating an audible pop or sound that can be counted. The detector is insensitive to γ-rays, and is independent of neutron energy to ~ 14 MeV. Sensitivity is ~ 80 bubbles per 10 μSv, with a minimum threshold of ~ 1 μSv; there is a linear relationship between the number of bubbles and the neutron dose. The SDD technology has been incorporated into equivalent dose neutron survey meters with replaceable SSD cartridges that must be changed after exposure to certain maximum doses. Alternately, bubbles in samples can be visually counted to determine neutron dose (8).

PHOTOGRAPHIC DETECTORS

Photographic emulsions that darken in proportion to the amount of radiation they receive represent one of the earliest methods of detecting radiation. Most films consist of a thin plastic sheet with ~ 0.2 mm emulsions one or both sides; the emulsion, usually silver halide granules in a gelatin mixture is covered with a thin protective plastic coat. Modern films have emulsions specific for optimal detection of certain energy and intensity of radiation. Radiation incident on the emulsion changes the clear silver halide ions, forming a latent image. During processing additional silver is deposited, the darkness of the film is determined by where silver ions are deposited on the film and the amount of silver deposited. While most films are limited to dynamic ranges of ~ 10^3, multiple film packets can be used in combination to extend the dynamic ranges to 10^5 or higher. Fast films are sensitive to the lowest levels of radiation, while slow films require greater exposure to darken the films. Usually, the film is used in a protective cover, which can be a cassette, commonly used for imaging and including metal screens to enhance the image. Rapid processing film is wrapped in a thin light tight paper wrapping that prevent light leaks and may be used without a cassette.

Personnel monitors frequently use film as the radiation detector (Table 6). The small film packet in its light wrapper is carried in a plastic holder, but the film wrapping is sufficiently thin to allow the transmission of the low energies X rays, γ rays, or beta particles. Films usually exhibit an enhanced sensitivity of several factors of 10, to lowest energies of radiation, below ~ 100–150 keV, relative to their response to γ rays with energies of 150–1.5 MeV. By using filters of copper, aluminum, and lead of varying thickness in the plastic holder in front of the film, estimates

Figure 19. Left Panel: liquid nitrogen dewar (left; Ortec, Model unknown) for cooling a germanium detector (right; Princeton Gamma Tech, Model RG-11B/C;); Right Panel: The assembled detector ready for use.

of the energies of the radiations darkening the film can be made. One distinct advantage of film dosimeters is the permanent record generated. A disadvantage is that radiation incidence at an angle to the filter appears to have passed through a thicker filter than actually available. Film badges or monitors are available in numerous configurations for the body, head, wrist, hand, and fingers. Normally film badges are exchanged monthly if radiation levels are low; individuals working in a higher level radiation environment may be monitored more frequently.

Numerous environmental factors can fog photographic film producing erroneous personnel exposure readings. While films are free of electromagnetic interference, excessive humidity can influence results, as can excessive heat. Images in films can fade with time so prompt collection and processing of personnel monitors is important in obtaining accurate results. Special metal filters, such as cadmium, may be used in a film holder to produce a neutron sensitive film by the (n, α) reaction in cadmium. Film dosimeter are widely used as personnel monitors and the overall accuracy of a film dosimetry system is usually at least 50% or better depending on the energy range of use. Lower energy radiation present in small amounts yields the greater uncertainty in the accuracy of monitor readings. With film dosimeters, the method of film processing is very important, as inconsistent methods of developing film will lead to substantial errors in the final results. Generally, film processors have their developer chemistry optimized for a particular type of film. Monitoring the temperature of the developing chemistry is very important and frequent use of calibration films, films previously given known exposures to radiation, is required to maintain the integrity of a film dosimetry system. Proper care and maintenance and rigorously scheduling of chemical developer replenishers are necessary.

ACKNOWLEDGMENTS

Equipment photographs courtesy of Todd Senglaub, Loyola Radiation Control, and Alvin Hayashi, Loyola Medical Media. I wish to thank Josephine Davis, Loyola Dept. Radiation Oncology, for her patience and excellent assistance preparing this manuscript.

BIBLIOGRAPHY

1. Health Physics Society. (No Date). Home Page. [Online] Health Physics Society. Available at http://www.health-physics.com. [2004, Nov. 16]. 2004.

2. Glasgow GP. Radiation protection instruments. Webster JG, editor. Wiley Encyclopedia of Medical Devices and Instrumentation. New York: Wiley; 1988.
3. Kasper K. Ludlum Model 2360. Health Phys 2003;84:1.
4. Health Physics Instrumentation Committee (1999, Sept. 14) [Online] Department of Energy. Available at http://www.llnl.gov/HPIC/HPICHP.HTML. [2004, Nov. 16]. 2004.
5. Knoll GF. Radiation Detection and Measurement. 3rd ed. New York: Wiley; 2000.
6. Shapiro J. Radiation Protection: A Guide for Scientists, Regulators, and Physicians. 4th ed. Cambridge (MA): Harvard University Press; 2002.
7. Turner JE. Atoms. Radiation, and Radiation Protection. 2nd ed. New York: Wiley; 1995.
8. Gollnick DA. Basic Radiation Protection Technology. 4th ed. Altadena (CA): Pacific Radiation Corporation; 2001.
9. Mohr PJ, Taylor BN. The Fundamental Physical Constants, [2004, Aug] Physics Today. [Online]. Available at http://www.physicstoday.org/guide/fundconst.pdf. [2004, Nov. 16]. 2004.
10. Storm DJ, Watson CR. On being understood: Clarity and jargon in radiation protection. Health Phys 2002;82:373–386.
11. 1990 Recommendations of the ICRP. Publication 60, International Commission on Radiological Protection. New York: Pergamon Press; 1990.
12. International Commission on Radiation Units and Measurements. Quantities and units in radiation protection dosimetry. Bethesda (MD): ICRU Publications; ICRU Report No. 51: 1993.
13. International Commission on Radiation Units and Measurements. Fundamental quantities and units for ionizing radiation. Bethesda (MD): ICRU Publications; 1998.
14. Bureau International des Poids et Measures. Le Systeme International d'Unites (SI). French and English Texts. Servres, France: Bureau International des Poids et Measures; 1991.
15. Olsher RH, et al.. WENDI: An improved neutron rem meter. Health Phys 2000;79:170–181.
16. Glasgow GP, Eichling J, Yoder RC. Observations on personnel dosimetry for radiotherapy personnel operating high energy linacs. Health Phys 1986;50:789.
17. Kasper K. Optically stimulated luminescence dosimeters. Health Phys 2001;81:1108–1109.
18. Products and.Services. (No Date). Home Page. [Online]. Landauer, Inc. Available at http://www.landauerinc.com/luxelosl.htm. [2004, Nov. 16]. 2004.
19. GAFCHROMIC EBT Product Brief. (No Date). Home Page. [Online]. International Specialty Products. Available at http://www.ispcorp.com/products/dosimetry/index.html. [2004, Nov. 19]. 2004.
20. Kumamoto Y, Noda Y. Measurements of effective doses of natural background levels of neutrons with etched-tracked detectors. Health Phys 2002;83:553–557.

See also CODES AND REGULATIONS: RADIATION; EQUIPMENT MAINTENANCE, BIOMEDICAL; SAFETY PROGRAM, HOSPITAL; X-RAY EQUIPMENT DESIGN.

RADIATION THERAPY, INTENSITY MODULATED

WALTER GRANT III
Baylor College of Medicine
Houston, Texas

INTRODUCTION

The past 10 years has seen the rapid emergence of a new radiation therapy process that has become known as intensity modulated radiotherapy, or IMRT. This new process has expanded so rapidly that there are many misconceptions of its origin as well as its identity. The concept of modifying the standard radiation pattern emitted from an external radiotherapy unit has been employed for decades, for example, in the form of physical wedges or compensators. To begin to understand this new technology, it is important to recognize what differentiates the non-uniform intensity patterns associated with IMRT from those accomplished with other methods. For example, The National Cancer Institute Collaborative Working Group for IMRT stated that IMRT is, "An advanced form of image-guided 3dCRT that utilizes variable beam intensities across the irradiated volume that have been determined using computer optimization techniques". (1). While this definition is correct, it is a generalization that adds to the lack of clarity regarding IMRT and a review of the literature should allow a better understanding of the basic technologies required to produce an IMRT treatment.

In 1982, Brahme, Roos, and Lax (the BRL paper) published an article (2) describing the exact solution for the beam intensities required for a new nonlinear wedge shape that could be used to create improved dose uniformity for targets in a cylindrical phantom that were on or near the axes of symmetry and rotation of a problem in arc therapy. The authors also discussed the similarities to the imaging processes used in computed tomography (CT). This technique was used to treat 25 patients.

In 1987, Cormack (3) extended the BRL concept to noncircular symmetry. These results have nonexact solutions, but the author discussed logical approaches to selecting the best intensities.

The major step came just a year later in 1988 when Brahme published an article (4) that proposed the use of inverse treatment planning using filtered backprojections to solve both stationary and rotational problems. He believed that this approach would have a large impact and stated that it, "... largely avoids the trial and error approach often applied in treatment planning of today". His planning scheme was to have the physician place constraints on the doses to tumors and normal tissues and allow a computer to determine the location and intensities for the beams that achieved these results, as opposed to the traditional method where the planner places the beams and then evaluates the resulting dose distribution.

In diagnostic CT, one uses filtered backprojections to eliminate the artifacts that occur during image reconstruction. These filters mathematically convert a uniform beam from an X-ray tube into a nonuniform beam in order to achieve the proper images. They can have fine spatial resolution as well as negative values. It is this reality that creates the fingerprints of the process that is known as IMRT. In order to perform the reverse process in radiation therapy, we would have to use small beams and be able to produce a negative radiation source. This means that we likely will not be able to produce an exact solution, but only a "best" solution based on the clinical constraints of the patient and the source of radiation. For this reason, an optimization algorithm must be employed. We now have the unique markers for IMRT, the ability to plan and

delivery small beamlets of varying intensities that have been determined by computer optimization.

This article also cleared the path for the two delivery approaches used today in IMRT, multiple stationary gantry positions (fixed field) with dynamic intensity map creation and arc therapy delivery (tomotherapy) of dynamic intensity maps. Each of these techniques used different technologies for delivery of the dose pattern and had unique problems to overcome. In order to appreciate the complexities of the technologies, each will be addressed separately, beginning with the fixed field technique.

The limit for the effectiveness of radiation therapy as a treatment modality always has been the volume of normal tissue being irradiated. While early patients had this volume reduced by the use a library of lead blocks that were hand placed, the introduction of low melting point alloys (5) to create patient specific blocking introduced a major improvement in accomplishing the goal of the reduction of dose to normal tissue. The disadvantages of this system include the time to create the blocking as well as the weight of the finished product. To overcome these problems, multileaf collimators (MLC) were introduced (6). These devices were not created to increase blocking effectiveness, but to provide a more efficient blocking system. It took over a decade for these systems to mature to their potential, but today they are a common feature on linear accelerators. Figure 1 shows the exit port of an accelerator with the MLC leaves fully retracted. This would deliver a square radiation field. Using the MLC, one can shape the radiation field easily and Fig. 2 shows a field in the shape of a diamond. This, and any other shape, are set quickly by computer control.

The designs of the commercial MLCs have similar characteristics and some differences. The best place to find particular information is in the American Association of Physicists in Medicine (AAPM) Report No. 72 (7). There are some distinctions that will be made here as they affect the ability of these devices to create the required intensity maps for IMRT, but one should read Report No. 72 for more detail.

Figure 1. The exit port of a linear accelerator with the MLC parked completely under the primary collimating jaws.

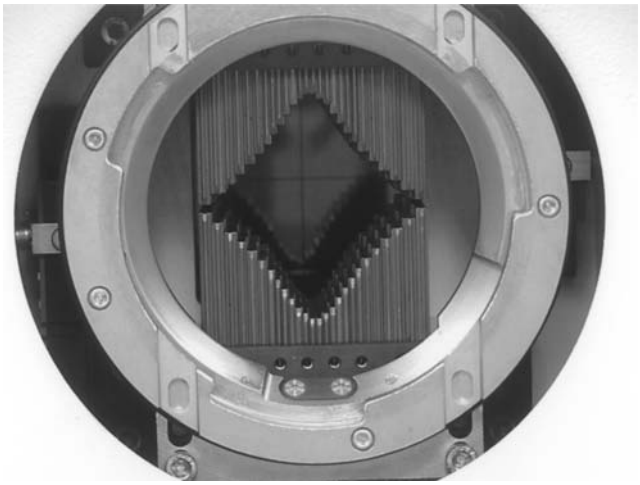


Figure 2. The MLC positioned to deliver a radiation pattern in the shape of a diamond.

Most vendors now have leaves that project to a 1.0 cm width at the treatment distance for the machine, although there are recent advances to 0.5 cm for a general MLC and 0.3 cm for a special purpose MLC. The MLC will either replace one set of primary collimators in the head of the machine or be attached to the machine as a tertiary collimator. The former allows for more clearance between the patient and the exit port of the linear accelerator, but will have greater accuracy requirement for leaf position as inaccuracy is magnified by distance increased distance from the patient. An additional advantage is that these leaves can be double focused, allowing for a beam penumbra that is constant as a leaf moves across the field. Tertiary collimators of today have directly opposite characteristics.

There are additional characteristics that one needs to evaluate for the use of MLCs in general and IMRT in particular. One characteristic is the maximum distance a leaf may travel across the center of the field. This will determine the maximum field that can be treated with IMRT unless some additional facility of the collimator is introduced.

Another is the capability for a leaf moving from one side to pass its two neighboring leaves from the opposite side. The capacity to interdigitate leaves is important for some intensity modulation segmentation techniques, as well as create stand alone or "island" blocks in the central portion of a treatment field. Figure 3 shows an interdigitated set of leaves. The leaves on the edges of the pattern are set to the centerline of the radiation field and one can see alternating leaves extending over the center line from each side.

A final characteristic is that no MLC has adjacent leaves with smooth surfaces. In order to prevent leakage radiation from streaming through the interface of two leaves, leaves are keyed or notched to eliminate a direct path for the radiation. During the creation of an intensity map, adjacent leaves may get far enough apart that there can be a significant increase in unwanted leakage radiation because the leaf edge is not as thick as the leaf center. This is called the tongue and groove effect and will be part

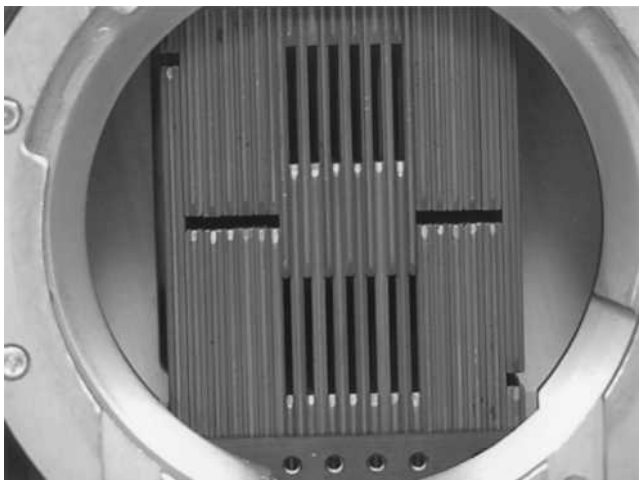


Figure 3. An example of leaves that are interdigitated. The closed leaves are set at the middle of the field.

of a latter discussion regarding the creation of the intensity maps for IMRT.

Although there was a decade of development for the MLC, it was obvious that these devices could be used to create intensity maps and investigators worked on two methods, both of which are clinically available today although not from every vendor. The first involves the creation of a final intensity map by the accumulation of multiple static fields, or segments. This technique was investigated theoretically by Bortfeld et al. (8) and Webb (9) while Bortfeld et al. (10) conducted the first experiment of such delivery in 1994. As a single segment is completed, the beam is turned off and the MLC instructed to change to the next position. Then the beam is turned on and the next segment is delivered. The NCI–CWG–IMRT defines this process as Static MLC IMRT or sMLC. As an example, the MLC in Fig. 4 has the leaves positioned at the boundaries of a radiation field to be treated. If one begins by positioning all the leaves to the left-side boundary and moves them independently across the field at specified dose intervals, one can deliver a non-uniform dose pattern as shown in Fig. 5 with the darker colors indicating higher dose.

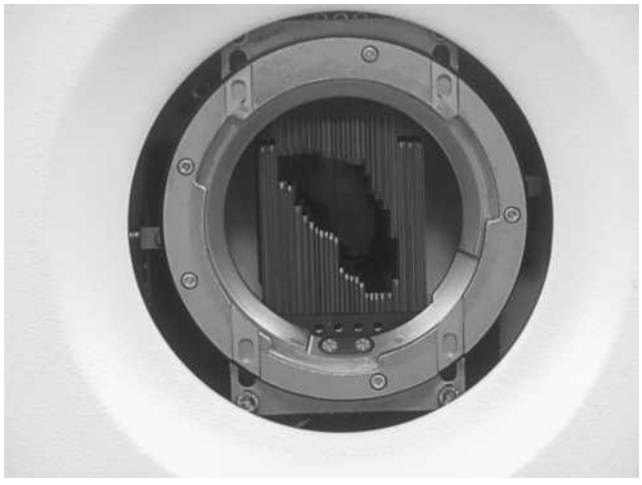


Figure 4. A MLC with the leaves positioned at the boundary of a shaped radiation field. The leaves on the right will move to meet the leaves on the left and then all leaves will move across the radiation field to create a nonuniform intensity map.

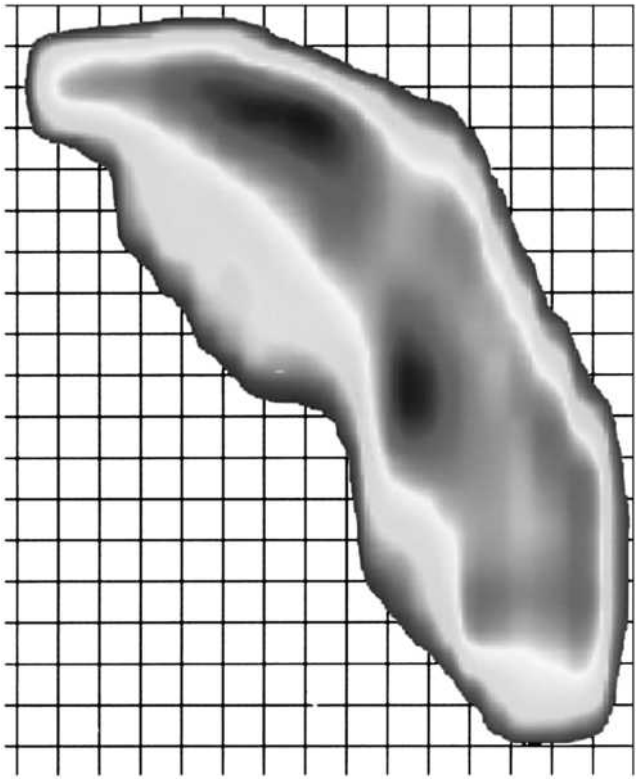


Figure 5. The planar intensity map created by the field from Fig. 4 (black is the highest radiation exposure).

All vendors currently deliver this methodology although with some meaningful differences. As mentioned earlier, the location of the MLC plays a major role in the positional accuracy of any leaf at the treatment distance and there are differences in the time each vendor requires to assure accuracy before initiating the beam on sequence of the accelerator. This time ranges from milliseconds to multiple seconds and will discourage the use of intensity maps with many segments in order to achieve clinically realistic delivery times. The use of this technique may result in segments that have the beam on for a fraction of a second, thereby requiring the accelerator to reach its steady-state performance in that same short time. The accelerator manufacturers have developed this capability on newer accelerators, but one should be careful when attempting IMRT on older accelerators. The second method involves the creation of intensity maps by varying the dose rate while leaves are in motion. The NCI–CWG–IMRT defines this process as Dynamic MLC IMRT or dMLC. Currently, it is available from only one vendor. The technique was postulated by Kallman et al. (11) and the optimal trajectory equations developed by Stein et al. (12), Svennson et al. (13), and Spirou and Chui (14). To successfully deliver such a treatment, the accelerator needs to monitor leaf positions and alter the dose rate as required to allow leaves

to reach the next positions at the proper time. While this is the faster means of creating an intensity map, it also has the potential for more discrepancies in leaf position and dose than the static approach. These discrepancies are likely to be small and, with many ports being delivered, are likely to produce small degradation in the desired pattern. It is also possible to identify potential large errors and modify them to be small. This process is usually done with software programs that optimize the segment delivery. By using this segmental optimization, one can control the magnitude of such things as the number of segments, the tongue and groove effect, and minimize the potential for one beamlet to have a much higher intensity than any of the others,

Delivery of IMRT with fixed field techniques is the most common method used because there are so many MLCs in use and the MLC can be used for conventional beam shaping without intensity modulation of the beam. However, the methodology has some drawbacks. Linear accelerators are expensive ($\sim$\$1.8 M) and it is important economically to treat large numbers of patients on each accelerator. For this reason, the number of ports to be treated must be limited to a few, optimal orientations. For some disease sites, such as prostate, these orientations can be predetermined and applied to all patients. In other disease sites, such as head and neck tumors, the disease presentation may have unique problems that require additional planning time to seek a satisfactory number of ports and orientations. There is no single answer to this problem as situations at each institution vary. For example, a small center might only have one or two machines while large institutions are likely to have more and can identify one that can treat fewer patients.

The alternative to fixed field IMRT is to deliver the treatment with arc therapy or tomotherapy. Since IMRT has such a strong similarity to CT, this is a logical approach. However, with the proliferation of high energy accelerators over the past 20 years, the use of arc therapy has diminished dramatically and is employed mainly in special procedures such as stereotactic radiosurgery (SRS). There are three implementations of tomotherapy and each has an analogy to current CT technology. In 1992 a neurosurgeon, Dr. Mark P. Carol, introduced a commercial system called Peacock (15) to be used as a SRS tool. The system consisted of a unique multileaf collimator (MIMiC) that consisted of two rows of 20 vanes each, with each vane projecting to a nominal $1 \times 1\,\mathrm{cm}^2$ beam at the normal isocenter of an accelerator (100 cm). The MIMiC is operated pneumatically and is a binary collimator meaning that the vane is either open (and permitting radiation to pass) or closed (blocking radiation). Figure 6 shows a patient's view of the MIMiC vanes in an alternating open/close pattern. One can see the two independent rows. The MIMiC is a removable tertiary collimator, so the accelerator could also be used for traditional clinical treatments. Figure 7 shows the MIMiC and it's associated hardware in place on an accelerator. Either a 5 or 10° sector of the arc delivery is considered a fixed field and a beamlet's intensity is based around the center of the arc sector. For a 10° sector, that means a beam with 100% intensity is open for the entire 10°, while a beam with 50% intensity is closed the first 2.5° of the sector, open for the next 5° and closed the last 2.5°.

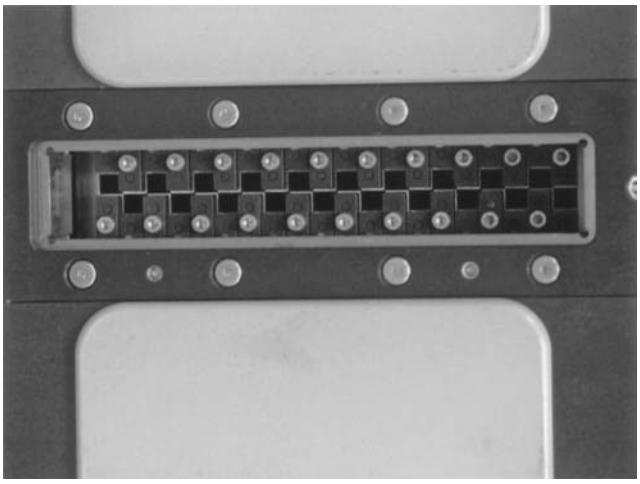


Figure 6. The MIMiC collimator as viewed by the patient. Alternating leaves are open and closed.

Because of the MIMiCs design, one treats a 2 cm length of the patient in one arc. It is then necessary to index the patient precisely and deliver another arc. The process is repeated until the total length of the tumor is treated. A CT analogy is a dual slice axial scanner and has the NCI–CWG–IMRT designation as Sequential Tomotherapy. This system also has historical importance as it was used to treat the first patient with IMRT (16) as defined by the NCI–CWG–IMRT and is still popular in 2004.

In 1995, Yu (17) introduced the concept of intensity modulated arc therapy (IMAT) that uses a traditional MLC to deliver tomotherapy. It basically is a combination of Dynamic Arc, a conformal delivery where the MLC changes its borders as a function of gantry angle but does not modulate the beam in the field, and the creation of intensity maps by using arcs with different shapes multiple times. This technique has limitations in that there are no planning systems that automatically create the fields, so the planner has to develop the skills to do this and that

Figure 7. A sequential tomotherapy system mounted on a linear accelerator.

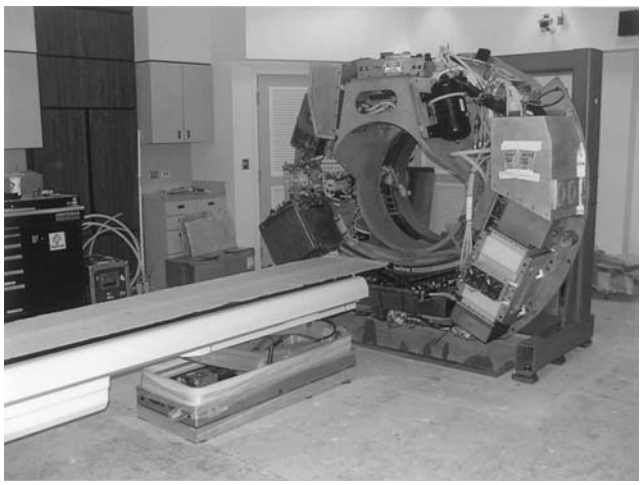


Figure 8. A helical tomotherapy machine at installation. One can see the gantry ring holding the accelerator components.

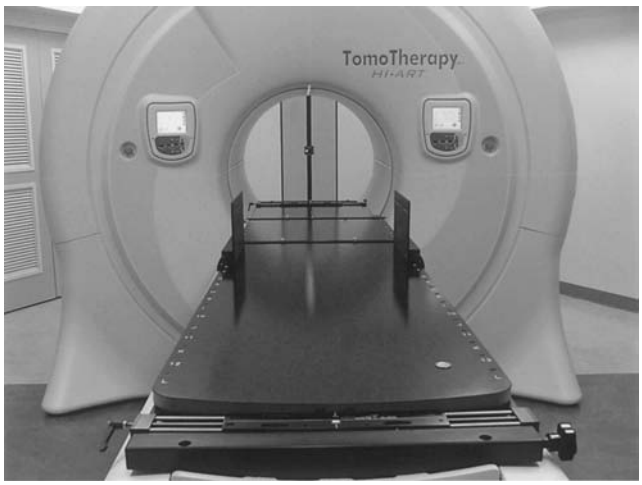


Figure 9. A helical tomotherapy machine in an operational configuration and looking much like a diagnostic CT scanner.

introduces a problem because, given N possible intensities, there are $(N!)^2$ combinations that can produce that pattern. As an example, an intensity map with only 3 intensities can be created using 36 possible patterns. However, with the use of multiple gantry angles, one can create clinically acceptable plans for many simple disease presentations. This technique is extremely useful on machines that would require seconds to verify leaf positions if one were doing sMLC. Since it treats the entire volume in each arc, a CT analogy would be a multislice axial scan. More information on using this technique is found in a more recent article by Yu and Shepard (18).

In 1993, Mackie et al. (19) described a novel machine designed to treat IMRT only with a helical delivery system just like a helical CT scanner. For this reason, the system has been defined as Helical Tomotherapy by the NCI–CWG–IMRT. This machine has the physical appearance of a CT scanner, with a 6 MV waveguide and associated electronics mounted on a rotating annulus. A unit at the time of installation is shown in Fig. 8. The patient is treated with 6 MV X-rays by having the treatment couch moving at a constant speed through the bore of the rotating system. In order to modulate the beam, this machine has a pneumatically driven binary collimator. In addition, this machine had a series of Xenon detectors mounted on the rotating annulus opposite the waveguide so that the radiation exiting from the patient can be captured and analyzed. This machine became available commercially in 2001, with the first clinical installations occurring in 2003. Figure 9 shows a helical tomotherapy machine ready for patient treatment. Because it is the first new design of an external beam treatment machine since the linear accelerator was introduced at Stanford in 1955, it deserves some additional scrutiny.

As with helical CT, the operator can control the pitch (the distance the table travels per revolution) as well as the slice width (the machine has moveable jaws that can be set from 0.5 to 5.0 cm prior to the initiation of treatment). Using a pitch >1 and a large slice width, one can create a megavoltage CT image (the waveguide is tuned for 3.5 MV X-rays) that can be used to verify or correct patient position prior to treatment. During treatment the exit dose can be collected in the xenon detectors and mapped back into the patient to determine the magnitude of any discrepancies between planned and actual delivery, whether they be caused by mechanical errors or patient–organ motion. These types of tools have long been desired to help insure the accuracy of these more conformal dose distributions.

Similar tools are finding their way into conventional linear accelerators as vendors are adding kilovoltage X-ray tubes as well as megavoltage cone CT to their equipment. There are questions as to how one facilitates the use of fixed-field delivery with the need to do arc rotations to gather the necessary information to create the images that are part of the desired schema.

The last issue to address is the use of optimized treatment planning as a necessary part of the IMRT concept. It was stated earlier that perfect shapes in CT often require the application of mathematical filters with negative values and, since we cannot deliver negative radiation, optimization tools are required to achieve the best result possible give the planning constraints. The evaluation of optimization techniques continues today but the application of optimization techniques in use today is very straightforward. They are either a gradient descent or stochastic algorithm. In the late 1980s, Webb (20) as well as Mohan et al. (21) investigated the stochastic algorithm, Simulated Anneling, as a possible algorithm. This was appropriate because the algorithm was powerful and no one was sure just how complicated the optimization need be to create a clinically useful plan. Because of this, Carol designed his system based on Webb's work with this stochastic optimization algorithm.

By their nature, stochastic optimizations are slower than gradient descent algorithms because they permit the solution to get worse for a number of iterations as an attempt to avoid being trapped in a local minimum. Over time, it became clear that gradient descent algorithms are useable and today all planning systems now use gradient descent algorithms and some use both. There is a product that uses no optimization at all and was given

an IMRT moniker by the vendor, but this product does not meet the NCI–CWG–IMRT definition of IMRT. Research interest in the subject continues to grow. A literature search of "inverse treatment planning" found 39 publications in 1998, 48 in 2000, and 108 in 2002, so we can expect optimization algorithms to continue to be tested and improved.

The rapid acceptance of IMRT as a delivery technique is unprecedented in radiation therapy. A mere 6 years after Brahme postulated the concept in 1988, Carol had produced a commercial system that was used clinically. Within a decade after that event, not only do vendors supply additional technology for their accelerators to treat IMRT, but also a new vendor has emerged with a dramatically new machine to treat only IMRT. Image guided systems are being adapted to take advantage of the power to shape radiation fields easily and precisely with IMRT. This technology should continue to expand.

BIBLIOGRAPHY

1. NCI–CWG–IMRT, Intensity-modulated radiotherapy: Current status and issues of interest. Int J Radiat Oncol Biol Phys 2001;51:880–914.
2. Brahme A, Roos JE, Lax I. Solution of an integral equation encountered in rotation therapy. Phys Med Biol 1982;27: 1221–1229.
3. Cormack AM, Cormack RA. A problem in rotation therapy with x-rays: Dose distributions with an axis of symmetry. Int J Radiat Oncol Biol Phys 1987;13:1921–1925.
4. Brahme A. Optimization of stationary and moving beam radiation therapy techniques. Radiother Oncol 1988;12:129–140.
5. Powers WE et al. A new system of field shaping for external-beam radiation therapy. Radiology 1973;108:407–411.
6. Sofia JW. Computer controlled, multileaf collimator for rotational radiation therapy. AJR Am J Roentgenol 1979;133: 956–957.
7. Boyer A et al. Basic applications of multileaf collimator; AAPM Report No. 72. Madison (WI): Medical Physics Publishing; 2001.
8. Bortfeld T, Burkelbach J, Boesecke R, Schlegel W. Methods of image reconstruction from projections applied to conformation radiotherapy. Phys Med Biol 1990;35:1423–1434.
9. Webb S. Optimization by simulated annealing of three-dimensional conformal treatment planning for radiation fields defined by a multileaf collimator. Phys Med Biol 1991;36: 1201–1226.
10. Bortfeld T et al. Realization and verification of three-dimensional conformal radiotherapy with modulated fields. Int J Radiat Oncol Biol Phys 1994;30:899–908.
11. Kallman P, Lind B, Eklof A, Brahme A. Shaping of arbitrary dose distributions by dynamic multileaf collimation. Phys Med Biol 1988;33:1291–1300.
12. Stein J, Bortfeld T, Dorschel B, Schlegel W. Dynamic x-ray compensation for conformal radiotherapy by means of multi-leaf collimation. Radiother Oncol 1994;32:163–173.
13. Svensson R, Kallman P, Brahme A. An analytical solution for the dynamic control of multileaf collimators. Phys Med Biol 1994;39:37–61.
14. Spirou SV, Chui CS. Generation of arbitrary intensity profiles by dynamic jaws or multileaf collimators. Med Phys 1994;21: 1031–1041.
15. Carol MP. Integrated 3-D conformal multivane intensity modulation delivery system for radiotherapy. Hounsell AR, Wilkinson JM, Williams PC, editors. Proceedings of the 11th International Conference on the Use of Computers in Radiation Therapy. Madison (WI): Medical Physics Publishing; 1994.
16. Butler EB, Woo SY, Grant 3rd W, Nizin PS. Clinical realization of 3d conformal intensity modulated radiotherapy. Int J Radiat Oncol Biol Phys 1995;32:1547–1548.
17. Yu CX. Intensity-modulated arc therapy with dynamic multileaf collimation: An alternative to tomotherapy. Phys Med Biol 1995;40:1435–1449.
18. Yu C, Shepard D. Treatment planning for stereotactic radiosurgery with photon beams. Technol Cancer Res Treat 2003;2:93–104.
19. Mackie TR, et al. Tomotherapy: A new concept for the delivery of dynamic conformal radiotherapy. Med Phys 1993;20:1709–1719.
20. Webb S. Optimisation of conformal radiotherapy dose distributions by simulated annealing. Phys Med Biol 1989;34: 1349–1370.
21. Mohan R, et al. Clinically relevant optimization of 3-d conformal treatments. Med Phys 1992;19:933–944.

See also COMPUTED TOMOGRAPHY; RADIOTHERAPY, THREE-DIMENSIONAL CONFORMAL.

RADIATION THERAPY SIMULATOR

DANIEL A. LOW
SASA MUTIC
Washington University School of Medicine
St. Louis, Missouri

INTRODUCTION

Radiation therapy, or radiation oncology, is one of the primary modalities (along with surgery and chemotherapy) for the treatment of cancer patients. Its origins can be traced to the early 1900s and today radiation therapy facilities can be found in most major medical centers and many free standing practices. This medical specialty and its success depend strongly on the technology used for cancer diagnosis and planning, delivery, and verification of patient treatments. Therefore, the amount of efforts and resources invested in the improvement of radiation therapy related technologies is significant. One of the cornerstones of modern radiation therapy practices are volumetric patient images, computed tomography (CT), magnetic resonance (MR) imaging, nuclear medicine imaging (positron emission tomography (PET) and single positron emission tomography (SPECT)), and ultrasound (US). Medical images are used for cancer detection, disease staging, treatment planning, for verification of treatment delivery, and for evaluation of treatment outcomes and patient follow up. Imaging devices that are used to image cancer patients for radiation therapy treatment planning are called Radiation Therapy Simulators. The distinguishing characteristics of radiation therapy simulators, in addition to their basic imaging properties, are that these devices need to have the following characteristics; (*1*) the modality allows patients to be imaged in their treatment position, (*2*) that the acquired images have high spatial accuracy, and (*3*) that the dataset be able to provide image datasets that are of sufficient quality to be used for validating the radiation beam shape and anatomic location. While these may seem like relatively straightforward

requirements, the design and implementation of radiation therapy simulators can be technically challenging. The main source of technical difficulties stems from the fact that the design of the medical imaging devices on which the simulators are based has historically been driven by the needs of diagnostic radiology that have less concern for patient positioning or for the spatial accuracy of the image datasets. For diagnostic scanning, patients often assume a comfortable position with their arms on the side or on abdomen–chest. Diagnostic physicians typically need to determine the presence of anatomic or functional anomalies, so a quantified determination of the size, shape, or location of internal organs or tumors relative to the imaging modality hardware is not a primary consideration.

For radiation therapy imaging, the patient extremities (arms and legs) are often positioned away from the torso to provide access by the radiation treatment beams. Patients are also imaged in immobilization devices that are subsequently used during treatment. Additionally, image spatial accuracy and the geometry of images is extremely important in order to precisely deliver the radiation to the tumors while avoiding radiation sensitive organs. Radiation therapy simulator design is based on an imaging device that was originally developed for diagnostic imaging, and then the device is modified to accommodate patient imaging in the radiotherapy treatment position and to improve image spatial accuracy and geometry to satisfy the needs of radiation therapy treatment planning. This approach is slowly changing and more devices are being designed exclusively for radiation therapy or major features of diagnostic imaging equipment are designed with radiation therapy in mind. This change in manufacturer attitude is reflected in description of radiation therapy simulators in the rest of this article.

The majority of simulation history in radiation therapy is based on conventional simulators (1–6). However, the modern practice of radiation therapy is dominated by CT simulators. Shortly after the introduction of clinical CT scanners in the early 1970s, it was realized that this imaging modality has much to offer in a radiation oncology setting. The CT images provide volumetric information not only about target volumes, but about critical structures as well. Using CT images for radiation therapy treatment, planning has improved dose delivery to target volumes while reducing dose to critical organs. The CT images also provide relative electron density information for heterogeneity-based dose calculations. A major weakness of CT imaging is a relatively limited soft-tissue contrast. This limitation can be overcome by using CT images in conjunction with MR studies for treatment planning. The PET images can be used to add physiological information. Ultrasound has also been useful for imaging in brachytherapy. Multimodality imaging-based treatment planning and target and normal structure delineation offer an opportunity to better define the anatomic extent of target volumes and to define their biologic properties.

Tatcher (7) proposed treatment simulation with CT scanners. This short article described the feasibility of CT simulator and indicated potential economical benefits. In 1983, Goitein and Abrams (8,9) further described multidimensional treatment planning based on CT images. Sherouse et al. (10,11) went on to describe a CT image based virtual simulation process that they referred to as a "software analog to a conventional simulation". This series of manuscripts described software tools and addressed technical issues that affect today's CT-simulation process. The manuscripts pointed out the need for fast computers, specialized software, but also for improved patient immobilization and setup reproducibility.

The radiation oncology community eagerly embraced the concept of virtual simulation and in early 1990s commercial software packages became available. These systems consisted of a diagnostic CT scanner, external laser positioning system, and a virtual simulation software work station. One of the early commercial CT simulation packages is shown in Fig. 1.

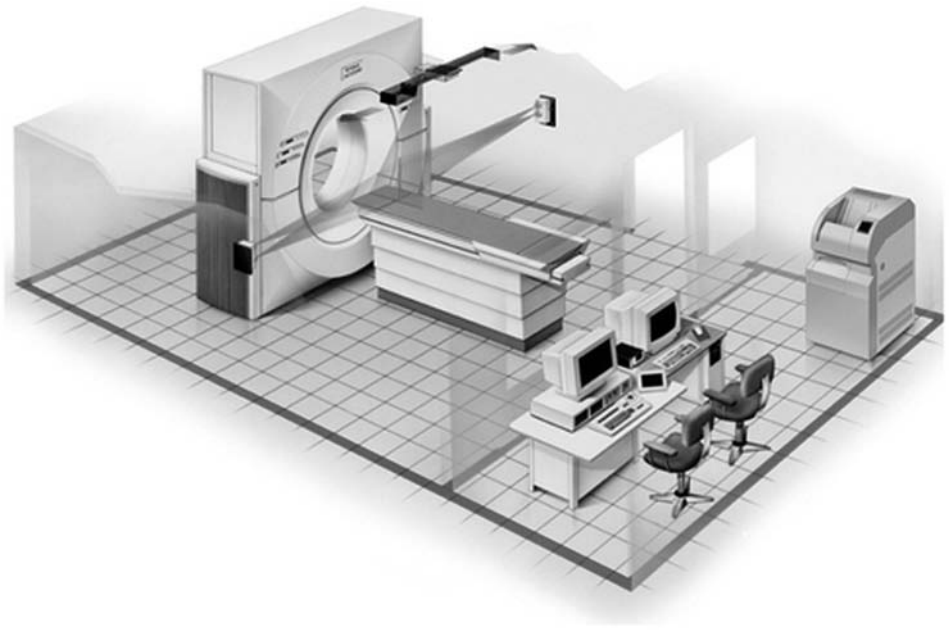


Figure 1. The CT-simulator room layout. (Image courtesy of Philips Medical Systems, Cleveland, Ohio.)

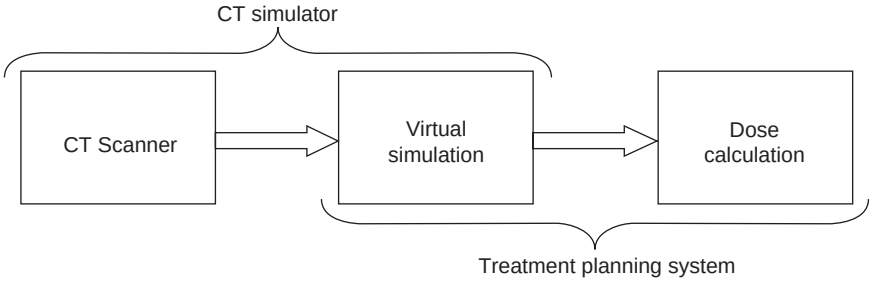


Figure 2. Place of CT simulation in radiotherapy treatment planning process. (Reprinted with permission from Ref. 12.)

The CT simulators have matured to a point where they are one of the cornerstones of modern radiation oncology facilities. Today's systems incorporate specially designed large bore CT scanners, multislice CT scanners, high quality laser positioning systems, and sophisticated virtual simulation packages. Many systems incorporate dose calculation capabilities and treatment plan analysis and evaluation tools.

Additional virtual simulation software features and functions along with increased efficiency and flexibility have enabled CT simulators to replace conventional simulators in many facilities. This trend seems to be further fueled by the increased demand for imaging studies for conformal three-dimensional (3D) and intensity modulated radiation therapy (IMRT) treatment planning where conventional simulators are of limited value. Figure 2 shows the place of CT simulation in the treatment planning process.

Both MR and PET–CT simulators are recent developments in radiation therapy simulation and are designed to complement CT simulation process and shortcomings of CT imaging. Magnetic resonance imaging offers superior soft tissue contrast and PET provides information about biological tissue properties. Computed tomography has relatively poor soft tissue contrast and provides rather limited information about functional tissue characteristics. Both MR and PET imaging in radiation therapy imaging greatly enhance our ability to accurately define anatomical and biological properties of tumors and normal tissues.

The implementation of simulation and treatment planning process varies greatly between radiation oncology departments. This diversity is in part driven by significant technical differences between simulation and treatment planning systems offered by different manufacturers. The discussion of radiotherapy simulators provided here describes general characteristic of processes and technology used for radiation therapy treatment planning. For more specific details, readers are referred to suggested readings list.

TECHNOLOGY OVERVIEW

In the late 1990s, the imaging equipment manufacturers began designing major devices (CT, MR, and PET scanners) specifically for radiation therapy or with radiation therapy needs in mind. This paradigm change resulted in a multitude of imaging devices available for radiation therapy simulation. Not only are there new devices (CT, MR, PET), but conventional simulators are being improved as well in order to be able to compete with other imaging modalities.

CONVENTIONAL SIMULATOR

The radiation therapy simulator has been an integral component of the treatment planning process for > 30 years. Conventional simulators are a combination of diagnostic X-ray machine and certain components of a radiation therapy linear accelerator (1–6). A conventional simulator, as seen in Fig. 3, consists of a diagnostic X-ray unit and fluoroscopic imaging system (X-ray tube, filters, collimation, image intensifier, video camera, generator, etc. (13), patient support assembly (a model of the treatment table), laser patient positioning and marking system, and simulation and connectivity software. The treatment table and the gantry are designed to mimic the geometric functions of a linear accelerator. The gantry head is designed to accommodate the common beam modification devices (blocks, wedges, compensating filters), in a geometry that mimics the linear accelerator. The simulator provides transmission radiographs with radiation portal-defining collimator settings outlined by delineator wires. By using primarily bony landmarks, the physician delineates the radiation portal outlines.

Imaging chain: One of the major improvements in conventional simulator design was the replacement of image intensifiers and video camera systems by amorphous silicon detectors. The new imagers produce high spatial and contrast resolution images which approach film quality, Fig. 4. More importantly, these images are distortion-free, a feature that is important for accurate geometric representation of patient anatomy. The introduction of high quality digital imagers in conventional simulation further facilitates the concept of film-less radiation oncology departments.

Simulation software: Conventional simulation software has also undergone many improvements. Modern simulators use the Digital Image Communications in Medicine (DICOM) standard (14) for data import capabilities. Treatment field parameters can be imported directly from the treatment planning computer for verification on the simulator. The software can then automatically set the simulator parameters according to the treatment plan. This facilitates efficient and accurate verification of patient

Figure 3. Modern version of a conventional simulator. (Image courtesy of Varian Medical Systems, Palo Alto, California, copyright © 2002.)

treatment setup on the conventional simulator. These simulators also have DICOM export capabilities that improves the reliability of treatment setup parameter transfer directly to a record and verify system or to a treatment planning computer. The ability to import and capture digital images enables conventional simulators to have tools for automatic correlation of treatment planning and verification fields.

Another potential improvement for conventional simulators is the capability of providing cone-beam CT (15,16). Because newer simulators are equipped with digital imaging hardware, two-dimensional (2D) projection image datasets are acquired while the gantry is rotating. The projection data is used to reconstruct a high spatial resolution CT image dataset. This capability will significantly improve imaging capabilities and usefulness of these devices. Figure 5 shows a cone beam CT image from a conventional simulator.

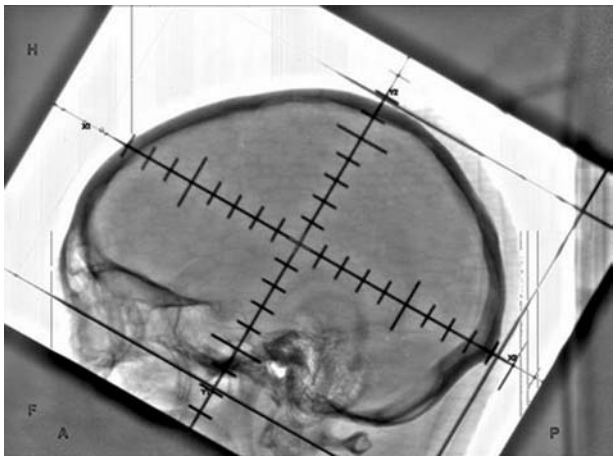


Figure 4. Digital image of a head from a modern conventional simulator equipped with an amorphous silicon imager. (Image courtesy of Varian Medical Systems, Palo Alto, California, copyright © 2002.)

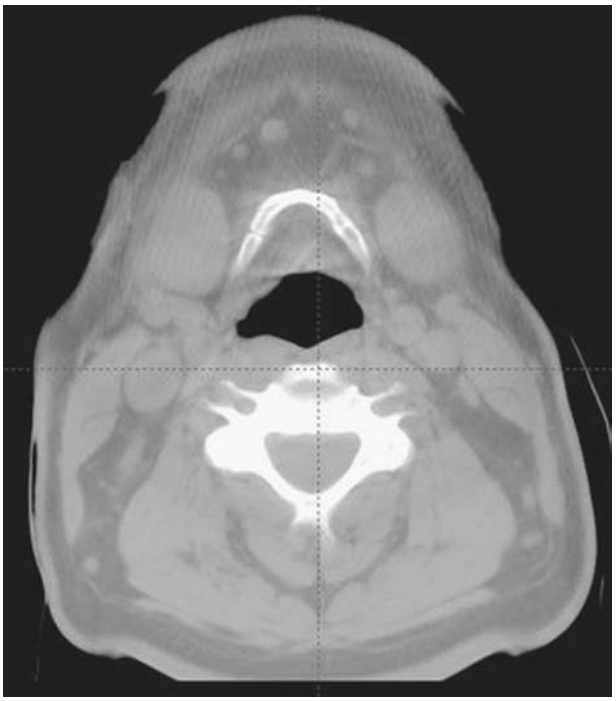


Figure 5. Cone beam CT image of a head acquired on a conventional simulator. (Image courtesy of Varian Medical Systems, Palo Alto, California, copyright © 2002.)

While it is often mentioned that conventional simulators can be completely replaced with CT simulators, new features and usefulness of conventional simulators are slowing down this process. Conventional simulator continues to be an important component of radiotherapy process even though its use for treatment planning of many tumor sites has been significantly reduced.

CT Simulator

Computed tomography simulator (10,11,17–26) consists of a CT scanner, laser patient positioning–marking system, virtual simulation–3D treatment planning software, and different hardcopy output devices, Fig. 1.

The CT scanner is used to acquire volumetric CT scan of a patient that represents the virtual patient and the simulation software recreates the functions of a conventional simulator. In recent years, the three most significant changes in CT-simulation technology have been the introduction of a larger gantry bore opening (Large Bore CT) (27), multislice image acquisition (Multislice CT) (28), and addition of CT-simulation software directly on the CT scanner control console. These innovations improve the efficiency and accuracy of the CT-simulation process. They also improve the patient experience by allowing patients to be positioned in more comfortable positions while reducing the simulation procedure time.

Large Bore CT: Large bore CT scanners (defined here as having $>$ 70 cm diameter bores) were specifically designed with radiation therapy needs in mind. One of the requirements in treatment of several cancer sites (breast, lung, vulva, etc.) is for extremities to be positioned away from the torso. When acquiring a CT scan with a patient in such treatment position, extremities often cannot fit through a conventional 70 cm diameter scanner bore opening. In such situations, patient positioning needs to be modified to acquire the scan. This can result in less than optimal treatment position (patient may be less comfortable and therefore the daily setup reproducibility may be compromised) or in a mismatch between the imaging and treatment positions. The first large bore CT simulator was introduced in 2000, and several additional models with enlarged bore opening have been introduced since then.

Large bore scanners also have increased the available scan field of view (SFOV), which determines the largest dimension of an object that can be fully included in the CT image. It is typically 48–50 cm in diameter on most conventional 70 cm bore opening scanners. For treatment planning purposes it is necessary to have the full extent of the patient's skin on the CT image. Lateral patient separation can often be $>$ 48–50 cm and the skin is then not visible on CT images. Increased SFOV available on large bore scanners solves this problem. There are, however, differences in implementation of extended SFOV and validity of quantitative CT values (quantitative CT) at larger image sizes. The CT numbers for some scanners are accurate only for smaller SFOVs and the values toward the periphery of large SFOV images are not reliable. This can be a concern for some dose calculation algorithms because inaccurate CT numbers can lead to dose calculation errors. The impact of CT number accuracy for increased SFOV images on dose calculation accuracy should be evaluated during scanner commissioning.

Multislice CT: In 1992, Elscint introduced a scanner that had a dual row of detectors and could simultaneously acquire two images (slices). Since then, multislice CT has gained wide spread acceptance and scanners that can acquire up to 64 slices are now available from all major vendors. The basic design behind the multislice CT technology is that multiple rows of detectors are used to create several images for one rotation of the X-ray tube around the patient.

One of the obstacles for radiation therapy scanning with single-slice scanners is the limited tube heat loading capability. Often, fewer images are taken, slice thickness is increased, the image quality is decreased (reduced mAs), or scan pitch is increased to reduce the amount of heat produced during the scan and to allow for the entire scan to be acquired in a single acquisition. Due to the longer length of imaged volume per tube rotation (multiple slices acquired simultaneously), the tube heat loading for a particular patient volume is reduced when using a multislice scanner relative to a single-slice scanner and multislice scanners are generally not associated with tube heat loading concerns. Faster acquisition times and decreased tube loading of multislice scanners, which allow longer volumes to be scanned in a single acquisition, can provide an advantage over single-slice systems for treatment planning purposes. Multislice technology can be especially beneficial for imaging of the thorax where breathing artifacts can be minimized with faster scanning. Multislice technology also facilitates dynamic CT scanning, often referred to as 4D CT (29,30). This application of multislice CT in radiation therapy is yet to be fully explored.

Multislice scanners are also capable of acquiring thinner slices that can result in better quality digitally reconstructed radiographs, used for treatment portal validation, and more accurate target delineation because of the improved spatial resolution with thinner slices, (Fig. 6).

CT simulator tabletop: This section and discussion about simulator tabletops applies equally to all simulators used in radiation therapy (conventional, MRI, CT, and PET) and treatment machines. Tabletops used for patient support in radiation therapy during imaging or treatment should facilitate easy, efficient, reproducible, and accurate patient. It is not only important that a tabletop improves patient positioning on a single device (i.e., treatment machine), but the repositioning of a patient from one imaging or treatment device to another also has to be considered. A great improvement in this process is if all tabletops involved in patient simulation and treatment have a common design. They do not have to be identical, but they should have the same dimensions (primarily width), flex and sag under patient weight, and they should allow registration (indexing) of patient immobilization devices to the tabletop. Figure 7 demonstrates this concept. The CT simulator tabletop has the same width as the linear accelerator used for patient treatment and both allow registration of patient immobilization system to the treatment couch. The ability to register the immobilization device and the patient to a treatment table is extremely important and improves immobilization, set-up

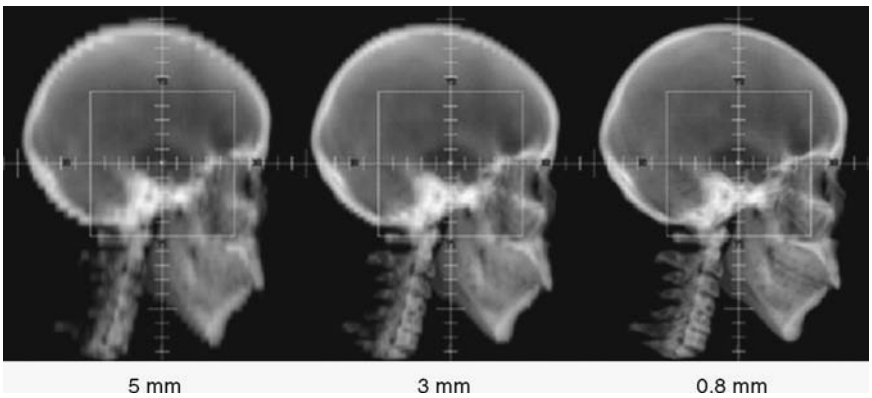


Figure 6. The CT slice thickness DRRs are 5, 3, and 0.8 mm. Thinner slice thickness images reveal much more relevant anatomical detail.

reproducibility, accuracy, and efficiency. The patient is always positioned in the same place on the treatment machine and patient daily setup can be facilitated using the treatment couch positions. If the patient is registered to the treatment couch, the coordinates of the couch used for patient treatment can become a part of parameters that are set and tracked in the linear accelerator record and verify system.

Patient marking lasers: A laser system is necessary to provide reference marks on patient skin or on the immobilization device. Figure 1 shows a laser system for a CT simulator:

Wall lasers: Vertical and horizontal, mounted to the side of the gantry. These lasers can be fixed or movable.

Sagittal laser: Ceiling or wall mounted single laser, preferably movable. Scanner couch can move up/down and in/out, but cannot move left/right, therefore the sagittal laser should move left–right to allow marking away from patient mid line.

Scanner lasers: Internally mounted, vertical and horizontal lasers on either side of the gantry and an overhead sagittal laser.

MR Simulator

The MR images for radiotherapy treatment planning are usually acquired in diagnostic radiology departments because few radiation oncology departments have a dedicated MR scanner. Furthermore, currently the majority of radiotherapy MR studies are limited to brain imaging. The MR scanner has a superior soft tissue contrast compared to CT imaging and there are several benefits that MR can offer for target delineation based on this advantage. There have been several reports describing use of MR scanners for imaging and treatment simulation in radiotherapy (31–35). Some of these reports have suggested that MR studies can be used without a corresponding CT scan for radiotherapy treatment planning. Indeed, if spatial distortions (the geometry of imaged objects is not always reproduced correctly), which is the largest concern with MR imaging, can be removed or minimized MR studies can be used as the primary imaging modality for several treatment sites. Superior soft tissue contrast provided by MR can also be an advantage for treatment planning of certain extracranial tumor sites like prostate, for example (36,37).

Conventional MR scanners are not well suited for extracranial imaging for treatment planning. The main difficulty is placement of patient in treatment position with immobilization device in the scanner. The small diameter and long length of conventional MR scanner openings significantly limits patient positioning options for imaging. Open MR scanners do not share these difficulties and patients can be scanned in conventional

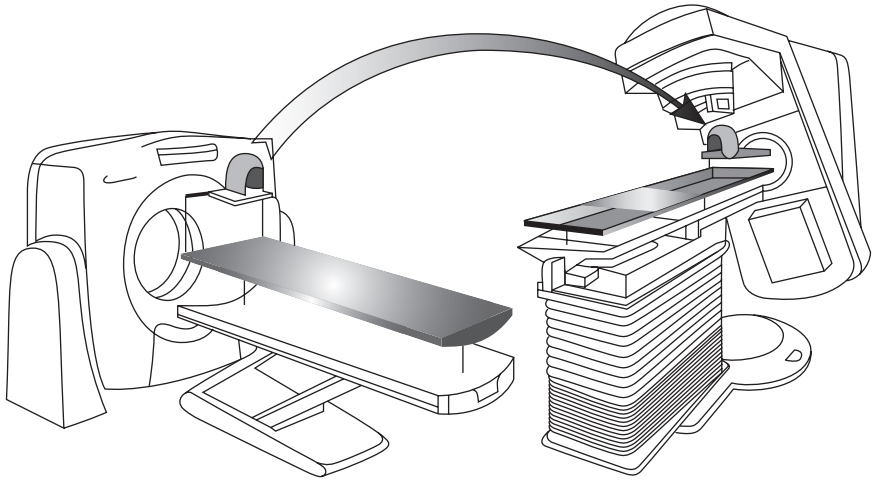


Figure 7. Similarity in design of simulator and treatment machine tabletops allows efficient and accurate reproducibility of patient positioning. (Image courtesy of MED-TEC, Inc, Orange City, Iowa.)

Figure 8. A MR simulator. (Image courtesy of Philips Medical Systems, Cleveland, Ohio.)

treatment positions. At least one manufacturer offers an open MR scanner that has been modified to serve as a radiotherapy simulator, Fig. 8. The scanner table is equipped with a flat top and external patient alignment lasers. The geometry of the scanner is then similar to the CT simulator. Another manufacturer offers a 70 cm diameter gantry opening conventional MRI scanner. The depth of the scanner opening is 125 cm. The dimensions of this scanner are very similar to a conventional CT scanner and in fact the scanner could be mistaken for a CT scanner. The ergonomics of this scanner are also well suited for radiotherapy simulation.

One of the major challenges with MR imaging for radiotherapy treatment planning are geometric distortions in acquired images. The MR scanners are often equipped with correction algorithms that will minimize geometrical distortions. These corrections do not affect the entire image and only the center portion of the image (center 20–35 cm diameter) is adequately correct (within 2 mm). Therefore, the representation of patient's skin and peripheral anatomy for larger body sections may be inaccurate. The effect of these inaccuracies must be evaluated if dose distributions will be calculated directly on MR images.

PET–CT Simulator

The PET images for radiotherapy planning can come from a standalone PET scanner or a combined PET–CT unit. Combined PET–CT scanners are being installed in radiation oncology departments and are used for PET scanning, but also these machines can be used for CT scanning only without PET acquisition. Due to this purpose, these scanners can be classified as CT simulators, though PET–CT simulator term may be more appropriate. Combined PET–CT scanners offer several advantages for radiotherapy imaging and are generally preferred over stand-alone units.

The first combined PET–CT prototype was introduced in 1998 at the University of Pittsburgh (38), since then all major manufacturers have produced several commercial models. The key description of PET–CT scanners is that a PET and a CT scanner are *combined* in the same housing. Meaning that there are two gantries (PET and CT) combined in one housing sharing a common couch. Image reconstruction and scanner operation is increasingly performed from one control console.

Combined PET–CT scanner design varies among different vendors with respect to PET detectors, image quality and resolution, speed, image field of view; number of CT slices, scanner couch design, gantry bore opening, and other considerations. Currently, the commercially available scanners have a 70 cm gantry opening for the CT portion, though large bore CT scanners will likely become part of PET–CT scanners in the future. The PET gantry opening ranges in diameter from 60 to 70 cm, meaning that some of the commercial scanners have a nonuniform gantry opening as the patient travels from the CT portion of the scanner to the PET side. More importantly, the scanners with the smaller gantry opening on the PET side will pose the same difficulties for radiotherapy scanning as stand-alone PET scanners. Again, the size of patient immobilization devices and patient scan–treatment position will have to be adapted to the size of the gantry opening.

The combined PET/CT technology offers two major benefits for radiotherapy planning. First, because the images are acquired on the same scanner, providing that the patient does not move between the two studies, the patient anatomy will have the same coordinates in both studies. These images have been registered using hardware rather than software registration. The second benefit of the

combined PET–CT units is that CT images are used to measure attenuation correction factors (ACFs) for the PET emission data, obviating the need for a time-consuming PET transmission scan (39,40). The use of CT images to generate PET ACFs reduces the scan time up to 40% and also provides essentially noiseless ACFs compared with those from standard PET transmission measurements (41). Shorter scan times can benefit radiotherapy patients who are scanned in treatment position that often can be uncomfortable and difficult to tolerate for prolonged amounts of time. One of the concerns with ACFs generated from CT images is mismatch or misalignment between CT and PET images due to respiration motion. The PET images are acquired during many cycles of free breathing and CT images are acquired as a snapshot in time at full inspiration, partial inspiration, or some form of shallow breathing. The breathing motion will cause mismatch in anatomy between PET and CT images in the base of lung and through the diaphragm region. This mismatch can result in artifacts in these areas that may influence diagnosis and radiotherapy target definition in this region. There are various gating methods that can be used during image acquisition to minimize the motion component and essentially acquire true, motionless, images of patient anatomy. Gated or 4D CT (with time being the fourth dimension) can be used to generate more reliable ACFs and also for radiotherapy treatment planning where gated delivery methods are being used.

Virtual Simulation Software

As with all software programs, user-friendly, fast, and well functioning virtual (CT) simulation software with useful features and tools will be a determining factor for success of a virtual simulation program. Commercially available programs far surpass in-house written software and are the most efficient approach to virtual simulation. Several features are very important when considering virtual simulation/3D treatment planning software:

Contouring and localization of structures: Contouring and localization of structures is often mentioned as one of the most time consuming tasks in the treatment planning process. The virtual simulation software should allow fast user-friendly contouring process with help of semiautomatic or automatic contouring tools. An array of editing tools (erase, rotate, translate, stretch, undo) should be available. An ability to add margins in three dimensions and to automatically draw treatment portals around target volumes should be available. An underlining emphasis should be functionality and efficiency.

Image processing and display: Virtual simulation workstation must be capable of processing large volumetric sets of images and displaying them in different views as quickly as possible (near real-time image manipulation and display is desired). The quality of reconstructed images is just as important as the quality of the original study set. The reconstructed images (DRRs and multiplanar reconstruction) are used for target volume definition and treatment verification and have a direct impact on accuracy of patient treatments.

Simulator geometry: A prerequisite of virtual simulation software is the ability to mimic functions of a conventional simulator and of a medical linear accelerator. The software has to be able to show gantry, couch, collimator, and jaw motion, SSD changes, beam divergence, and so on. The software should facilitate design of treatment portals with blocks and multileaf collimators.

DISCUSSION

As radiotherapy treatment planning and delivery technology and techniques change, so does the treatment simulation. The most significant change in the recent past has been the wide adoption of CT simulation to support conformal radiotherapy and 3D treatment planning. A CT simulation has gone from a concept practiced at few academic centers to several available sophisticated commercial systems located in hundreds of radiation oncology departments around the world. The concept has been embraced by the radiation community as a whole. The acceptance of virtual simulation comes from improved outcomes and increased efficiency associated with conformal radiation therapy. Image-based treatment planning is necessary to properly treat a multitude of cancers and CT simulation is a key component in this process. Due to demand for CT images, CT scanners are commonly found in radiation oncology departments. As CT technology and computer power continue to improve so will the simulation process, and it may no longer be based on CT alone. The PET–CT combined units are commercially available and could prove to be very useful for radiation oncology needs. Several authors have described MR simulators where the MR scanner has taken the place of the CT scanner. It is difficult to predict what will happen over the next 10 years, but it is safe to say that image based treatment planning will continue to evolve.

One great opportunity for an overall improvement of radiation oncology is the better understanding of tumors through biological imaging. Biological imaging has been shown to better characterize the extent of disease than anatomical imaging and also to better characterize individual tumor properties. Enhanced understanding of individual tumors can improve selection of the most appropriate therapy and better definition of target volumes. Improved target volumes can utilize the full potential of IMRT delivery. Biological imaging can also allow evaluation of tumor response and possibly modifications in therapy plan if the initial therapy is deemed not effective.

Future developments in radiotherapy treatment planning simulation process will involve the integration of biological imaging. It is likely that this process will be similar to the way that CT scanning was implemented in radiotherapy. The imaging equipment is initially located in diagnostic radiology facilities and as the demand increases the imaging is gradually moved directly to radiation oncology.

BIBLIOGRAPHY

1. Bomford CK, et al. Treatment Simulators. BJR 1989; (Suppl. 23).

2. Connors SG, Battista JJ, Bertin RJ., On the technical specifications of radiotherapy simulators. Med Phys 1984;11:341–343.
3. Greene D, Nelson KA, Gibb R., The use of a linear accelerator "simulator" in radiotherapy. BJR 1964;37:394–397.
4. McCullough EC., Radiotherapy treatment simulators, in Advances in radiation oncology physics: dosimetry, treatment planning, and brachytherapy. In: Purdy JA, ed. Woodbury (NY): American Institute of Physics; 1992. pp 491–499.
5. McCullough EC, Earle JD., Selection, acceptance testing and quality control of radiotherapy simulators. Radiology 1979; 131:221–230.
6. Van Dyk J, Munro PN. Simulators. In: Van Dyk J, Editor. The modern technology in radiation oncology. Wisconsin Medical Physics Publishing; 1999. pp 95–129.
7. Tatcher M., Treatment simulators and computer assisted tomography. BJR 1977;50:294.
8. Goitein M, Abrams M. Multi-dimensional treatment planning: I. Delineation of anatomy. Int J Rad Oncol, Biol, Phys 1983;9(6): 777–787.
9. Goitein M, et al. Multi-dimensional treatment planning: II. Beam's eye-view, back projection, and projection through CT sections. Inter J Rad Oncol, Biol, Phys 1983;9(6):789–797.
10. Sherouse G, et al. Virtual Simulation: Concept and Implementation. In: Bruinvis IAD, et al. ed. Ninth Int Conf Use of Computers in Radiation Therapy. North-Holland Publishing Co.; 1987. pp 433–436.
11. Sherouse GW, Bourland JD, Reynolds K. Virtual simulation in the clinical setting: some practical considerations. Int J Radiat Oncol Biol Phys 1990;19:1059–1065.
12. Mutic S, et al. Quality assurance for CT simulators and the CT simulation process: Report of the AAPM Radiation Therapy Committee Task Group No. 66. Med Phys 2003;30: 2762–2792.
13. Bushberg JT, et al. Fluoroscopy, in The Essential Physics of Medical-Imaging. 2nd ed. Baltimore: Lippincott Williams & Wilkins; 2002.
14. (NEMA), N.E.M.A. Digital Imaging Communications in Medicine (DICOM). 1998.
15. Jaffray DA, et al. A radiographic and tomographic imaging system integrated into a medical linear accelerator for localization of bone and soft-tissue targets. Int J Rad Oncol, Biol, Phys 1999;45:773–789.
16. Jaffray DA, et al. Flat-panel cone-beam computed tomography for image-guided radiation therapy. Int J Rad Oncol, Biol, Phys 2002;53:1337–1349.
17. Kushima T, Kono M. New development of integrated CT simulation system for radiation therapy planning. Kobe J Med Sci 1993;39(5–6):197–213.
18. Nagata Y, et al. CT simulator: a new 3-D planning and simulating system for radiotherapy: Part 2. Clinical application [see comments]. Int J Rad Oncol, Biol, Phys 1990;18(3): 505–513.
19. Nishidai T, et al. CT simulator: a new 3-D planning and simulating system for radiotherapy: Part 1. Description of system [see comments]. Int J Rad Oncol, Biol, Phys 1990;18(3): 499–504.
20. Butker EK, et al. Practical Implementation of CT-Simulation: The Emory Experience. In: Purdy JA, Starkschall G, eds. A Practical Guide to 3-D Planning and Conformal Radiation Therapy. Middleton (WI): Advanced Medical Publishing; 1999. pp 58–59.
21. Coia LR, Schultheiss TE, Hanks G, eds. A Practical Guide to CT Simulation. Madison (WI): Advanced Medical Publishing; 1995.
22. Conway J, Robinson MH. CT virtual simulation. Br J Radiol 1997;70:S106–S118.
23. Galvin JM. Is CT simulation the wave of the future? [letter; comment]. Med Phys 1993;20(5):1565–1567.
24. Heidtman CM. Clinical applications of a CT-simulator: precision treatment planning and portal marking in breast cancer. Med Dosimetry 1990;15(3):113–117.
25. Jani SK, ed. CT Simulation for Radiotherapy. Madison (WI): Medical Physics Publishing; 1993.
26. Van Dyk J, Taylor JS. CT-Simulators. In: Van Dyk J, ed. The Modern Technology for Radiation Oncology: A Compendium for Medical Physicist and Radiation Oncologists. Madison (WI): Medical Physics Publishing; 1999. pp 131–168.
27. Garcia-Ramirez JL, et al. Performance evaluation of an 85 cm bore X-ray computed tomography scanner designed for radiation oncology and comparison with current diagnostic CT scanners. Int J Rad Oncol, Biol, Phys 2002;52:1123–1131.
28. Klingenbeck_Regn K, et al. Subsecond multi-slice computed tomography: basics and applications. Eur J Radiol 1999;31(2): 110–124.
29. Keall P. 4-dimensional computed tomography imaging and treatment planning. Sem Rad Oncol 2004;14:81–90.
30. Low DA, et al. A method for the reconstruction of 4-dimensional synchronized CT-scans acquired during free breathing. Med Phys 2003;30:1254–1263.
31. Mah D, Steckner M, Palacio E. Characteristics and quality assurance of dedicated open 0.23 T MRI for radiation therapy simulation. Med Phys 2002;29:2541–2547.
32. Potter R, Heil B, Schneider L. Sagittal and coronal planes from MRI for treatment planning in tumors of brain, head and neck: MRI assisted simulation. Radiother Oncol 1992;23: 127–130.
33. Okamoto Y, Kodama A, Kono M. Development and clinical application of MR simulation system for radiotherapy planning with reference to intracranial and head and neck regions. Nippon Igaku hoshasen Gakkai Zasshi 1997;57(4): 203–210.
34. Schubert K, et al. Possibility of an open magnetic resonance scanner integration in therapy simulation and three-dimensional radiotherapy planning. Strahlenther Onkol 1999; 175(5):255–231.
35. Beavis AW, Gibbs P, Dealey RA. Radiotherapy treatment planning of brain tumors using MRI alone. BJR 1998; 71:544–548.
36. Chen L, et al. MRI based treatment planning for radiotherapy: dosimetric verification of prostate IMRT. Int J Rad Oncol, Biol, Phys 2004;60:636–647.
37. Lee YK, Bollet M, Charles-Edwards G. Radiotherapy treatment planning of prostate cancer using magnetic resonance imaging alone. Radiother Oncol 2003;66(2):203–216.
38. Beyer T, et al. A combined PET/CT scanner for clinical oncology. J Nuclear Med 2000;41:1369–1379.
39. Bailey DL. Data acquisition and performance characterization in PET. In: Valk PE, et al. eds. Positron emission tomography: Basic science and clinical practice. London: Springer-Verlag; 2003. pp 69–90.
40. Bailey DL, Karp JS, Surti S. Physics and Instrumentation in PET. In: Valk PE, et al. eds. Positron Emission Tomography: Basic Science and Clinical Practice. London: Springer-Verlag; 2003. pp 41–67.
41. Townsend DW, et al. PET/CT today and tomorrow. J Nucl Med 2004;45(Supl 1):4s–14s.

See also RADIATION DOSIMETRY FOR ONCOLOGY; RADIOTHERAPY TREATMENT PLANNING, OPTIMIZATION OF; X-RAY EQUIPMENT DESIGN; X-RAY QUALITY CONTROL PROGRAM.

RADIATION THERAPY TREATMENT PLANNING, MONTE CARLO CALCULATIONS IN

CHEN-SHOU CHUI
ELLEN YORKE
REN-DIH SHEU
Memorial Sloan-Kettering
Cancer Center
New York City, New York

INTRODUCTION

Boltzmann Transport Equation

The distribution of radiation in a three-dimensional (3D) heterogeneous medium is governed by the Boltzmann transport equation, which is a partial differential integral equation in six dimensions [3D for position, two (2D) for direction, and one (1D) for energy].

Due to the complex nature of the Boltzmann transport equation, analytic solutions are generally not available except for very simple, idealized cases. For realistic problems, numerical methods are needed. The Monte Carlo method is probably the most accurate and widely used method for solving radiation transport problems in 3D, heterogeneous geometry.

Random Sampling, Law of Large Numbers, Central Limit Theorem

The basic idea of the Monte Carlo method is to simulate physical events by random sampling from known probability distributions. For example, the step size a photon particle travels before the next interaction is sampled from the exponential distribution. The particular interaction event is sampled from the relative probabilities of competing interaction types. If a Compton event occurs, the energy and direction of the outgoing photon and electron are sampled from the Klein–Nishina distribution.

There are two mathematical principles underlying the Monte Carlo method: (1) The law of large numbers, and (2) the central limit theorem. The law of large numbers says that as the sample size increases to infinity, the sample average approaches the mean of the probability distribution from which the samples were drawn. Since in practice the sample size is finite, the error of the sample average needs to be estimated. For this, we make use of the central limit theorem, which states that the distribution of the sample averages approaches a normal distribution if the sample size is sufficiently large. Moreover, the standard deviation of the mean is inversely proportional to the square root of the total number of histories. Thus, in order to reduce the standard deviation of the mean by a factor of 2, the number of histories needs to be increased by a factor of 4. To estimate the statistical uncertainty, it is common practice to divide the total number of histories into separate groups, with each group containing sufficiently large number of histories. The sample average of each group, according to the central limit theorem, is normally distributed. The standard deviation of the mean is then calculated from these group averages. Alternatively, a history-by-history method can be used to estimate the standard deviation of the mean. In this method, both the quantity of interest and its square are tallied on the fly for each history. After the simulation is completed, the standard deviation of the mean is then calculated using the sum and the sum of squares of all histories. This method tends to have smaller uncertainty in the uncertainty estimate than the multiple group method.

Applications of Monte Carlo Method in Medical Physics

The Monte Carlo method has been applied to a variety of medical physics problems (1). For radiation therapy, these include the simulation of the machine head (2–22); dose calculation for external photon beams (23–38); electron beams (39–49); proton beams (50–54); and brachytherapy (55–86). For nuclear medicine, it has been used to calculate organ doses due to internal emitters (87–97). For diagnostic radiology, it is used for calculating beam characteristics and dosimetry (98–111). For radiation measurement, various correction factors for ion chambers have been calculated with Monte Carlo (112–129).

A number of Monte Carlo codes have been developed and applied to problems in medical physics (39,43,130–141). For the purpose of radiation therapy treatment planning, the EGS4 system and its user codes (43,130,133,135) are probably the most widely used in North America.

MONTE CARLO SIMULATION

Overview

A Monte Carlo simulation consists of two major components: transport and interaction. Transport moves a particle from one position to another while interaction determines the outcome of a particular interaction event. If an interaction results in multiple outgoing particles, as in the case of Compton scattering or pair production, then each outgoing particle forms its own transport-and-interaction track. This repetitive transport-and-interaction continues until either the particle travels out of the geometry, for example, exiting the patient body, or its energy falls below a cutoff energy.

Geometry Specification

For a given Monte Carlo simulation, the 3D geometry needs to be described by a collection of mathematical objects. One such package, called combinatorial geometry (142) has been used by a number of Monte Carlo codes. The combinatorial geometry package provides a set of primitive objects, such as sphere, cylinder, box, cone, and so on. An object in question is then modeled by logical combinations of these primitives. For example, a hemisphere can be formed as an intersection of a sphere and a box. The physical properties of the object are also assigned, including the material composition and the physical density. The combinatorial geometry package is a powerful tool that can be used to define very complex geometric objects such as the treatment machine head.

If a 3D image set, such as computed tomography (CT), is used for dose calculation, then the entire 3D voxel array defines the geometry. Since the CT image is typically

acquired with kilovoltage X rays, when it is used for dose calculation for megavoltage photons or electrons, the CT Hounsfield number needs to be converted into electron density ratios relative to water. For radiation therapy dose calculations, the typical energy range is up to 20 MeV and the physical properties of the voxel can often be considered as water equivalent but with varying densities. If, however, consideration of material composition is important, such as a metal implant, then different materials can be assigned to the corresponding voxels. This can be done by creating a lookup table for material (air, lung, fat, muscle, water, bone, soft bone, metal) as a function of Hounsfield number and then subdividing each material as a function of density.

Transport

During particle transport, the step size is sampled from the exponential distribution. The exponent in the exponential distribution is related to the mean free path that depends on the particle energy and the material in which the particle travels. For neutral particles (photons or neutrons), the step size is relatively large. For example, the mean free path of a ^{60}Co photon in water is $\sim$ 16 cm. For charged particles (electrons, positrons, protons), the step size is very short due to the Coulomb force. As a result, direct simulation would be very inefficient. To overcome this problem, multiple steps are condensed into a single step, called the condensed-history step (143). Energy deposition due to continuous slowing down is calculated along this step and angular deflection is sampled at the end of the step based on multiple scattering theory.

When transporting a particle, it may cross the boundary of one geometric object to another. When this happens, the original step has to be truncated at the boundary. This is because the next geometric object may have different physical properties from the current one. The remaining step size will have to be adjusted based on the new object and resampled if necessary.

For charged particles, there is continuous energy loss due to ionization and excitation. The energy can be deposited uniformly along the step or at a point randomly selected within the step. For neutral particles, there is no energy deposition during a step. However, if the KERMA (Kinetic Energy Released per unit Mass) approximation is made, then energy can also be deposited using the energy absorption coefficient and the step length, which is an estimate of the photon fluence.

Interaction Types

For radiation therapy dose calculation, we are mostly interested in photons and electrons, so the discussion of interaction types here is limited to these radiation particles only.

For photons, the interaction types are photoelectric, Compton scattering, and pair production. Coherent scattering is relatively unimportant as it involves no energy loss and only small angular deflection. In a photoelectric interaction, the incoming photon collides with an atom and ejects one of the bound electrons (typically K shell). The accompanying fluorescent X rays are low energy photons and are usually ignored in radiation therapy dose calculations. In a Compton scattering event, the incoming photon knocks off a loosely bound electron, considered as a free electron, from the atom. The photon itself is deflected with a lower energy. This is the dominant event for interaction of megavoltage photons with matter. The angular distribution of the outgoing particle is governed by the Klein–Nishina formula, and the angle of the outgoing particle uniquely defines its energy. In a pair production, the incoming photon is absorbed in the field of the nucleus and a positron–electron pair is produced. For this interaction to occur, the photon energy must be > 1.022 MeV, the sum of the rest mass energies of the positron–electron pair.

For electrons and positrons, the discrete interaction types are bremsstrahlung production, delta-ray production, and positron annihilation. Bremsstrahlung production is caused by the deceleration of charged particles (electrons and positrons) passing by the atomic nuclei. This is the mechanism by which photon beams are produced in a linear accelerator. The bremsstrahlung energy spectrum is continuous with the maximum energy equal to the kinetic energy of the incoming electron. The angular distribution is largely forward. A delta-ray is the secondary electron ejected from the atom resulting from a large energy transfer from the incoming electron or positron. If the incident particle is an electron, then the energy of the delta-ray cannot exceed one-half of the incident electron energy, for by definition, the outgoing electron with the lesser energy is the delta-ray. If the incident particle is a positron, then it can give up all its energy to the delta-ray. Positron annihilation is the process that occurs when a positron and an electron collide. If they are approximately at rest relative to each other, they destroy each other upon contact, and produce two photons of 511 keV each that are emitted in opposite directions. If they are moving at different relative speeds, the energies of the photons emitted will be higher.

APPLICATIONS IN RADIATION THERAPY DOSE CALCULATION

Beam Characteristics

In order to perform dose calculation using the Monte Carlo method, it is necessary to have accurate information about the radiation field incident upon the patient, that is, the phase-space data. This data is difficult to obtain by empirical means, therefore in practice it is obtained by Monte Carlo simulation of the machine head. Figure 1 shows a typical configuration of the machine head for a medical linear accelerator that produces clinical photon beams. The components directly in the beam are the target, the flattening filter, and the monitor chamber. The components that collimate the beam are the primary collimator, and the upper and lower collimating jaws. The phase space data can be collected on two scoring planes above and below the collimating jaws, respectively. For clinical electron beams, the machine head is similar to that of photon beams except that the target is removed, the flattening filter is replaced by a scattering foil system, and an additional applicator is used for further collimation of the electron beam (Fig. 2).

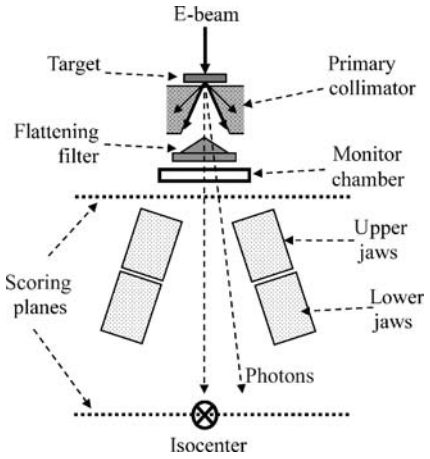


Figure 1. Production of a clinical photon beam (drawing not to scale).

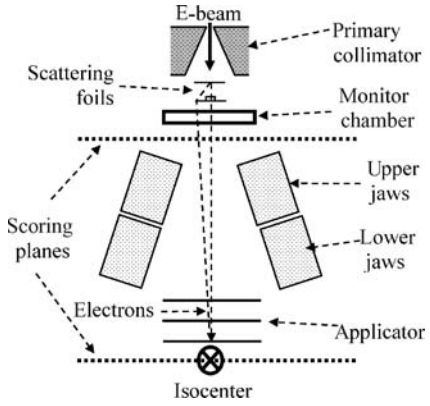


Figure 2. Production of a clinical electron beam (drawing not to scale).

Figure 3 shows the energy spectra of a 15 MV photon beam. Note that the low energy photons have been filtered out by the flattening filter. Moreover, the spectrum near the center of the beam, say, within 3 cm of the central axis, is harder than that away from the axis, say, 10–15 cm from

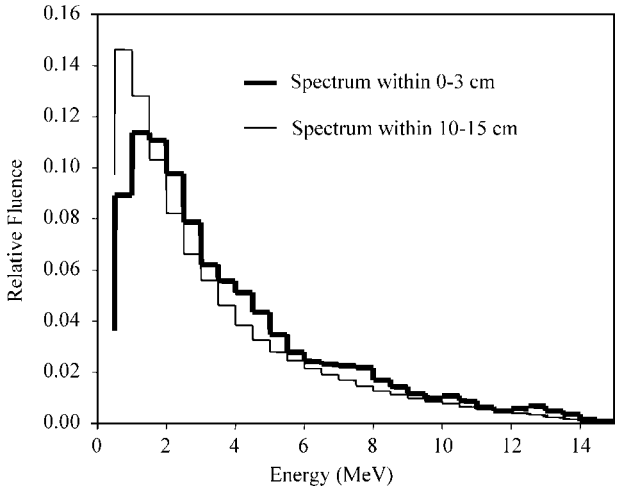


Figure 3. Energy spectra of a 15 MV photon beam.

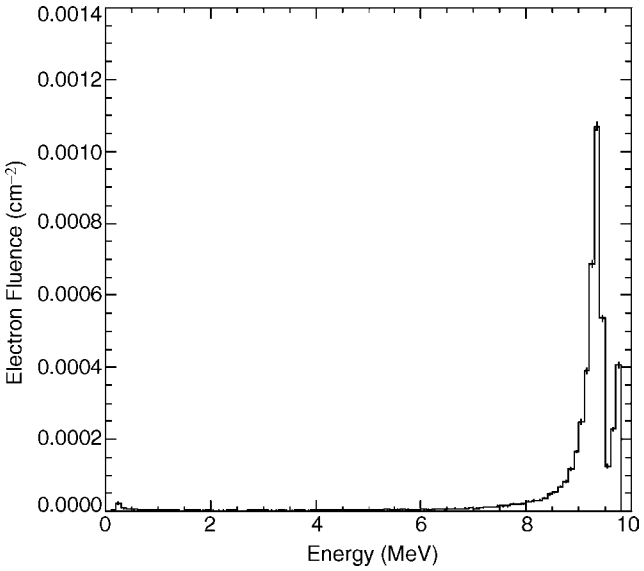


Figure 4. The energy spectrum of a clinical 9 MeV electron beam.

the center. The reason is that the flattening filter is thicker in the middle, thus absorbing more low energy photons. Figure 4 shows the energy spectrum of a clinical 9 MeV electron beam. It is clear there are two peaks corresponding to the thin part and the thick part of the scattering foil system. Figure 5 shows the angular distribution of the electrons at the isocenter plane (~100 cm from the entrance to the primary collimator). Due to the significant scattering of electrons in the scattering foils as well as in the air space above the isocenter, the angular spread is diffused and approximates a normal distribution.

This information about beam characteristics such as the energy and angular distributions is difficult to measure, but can be calculated by Monte Carlo with relative ease.

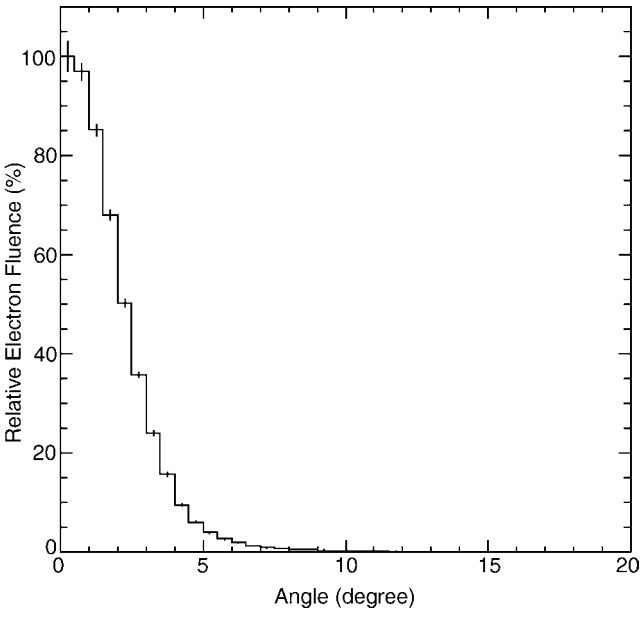


Figure 5. The angular distribution of a clinical 9 MeV electron beam.

Treatment Planning Dose Calculation

Many modern dose calculation algorithms other than Monte Carlo are quite accurate for megavoltage (^{60}Co–20 MV) external photon beam radiation therapy for sites that are composed of soft tissue (density $\sim$ 1 g·mL^{-1}) and bone (brain, pelvis, limbs) (144). However, these algorithms are less accurate when electronic equilibrium is lost due to more severe tissue inhomogeneities. This may be a clinical concern in the lung, where soft tissue tumors are surrounded by low density ($\sim$ 0.2–0.3 g·mL^{-1}) lung and in the head and neck (H&N) due to the presence of air cavities. Although Monte Carlo is currently impractical for routine clinical use, Monte Carlo calculations based on patient CT scans and inhomogeneous phantoms provide clinically valuable information, especially when combined with high resolution phantom measurements (film and/or TLD). For a summary of the status of the field (145).

Lung Cancer

Since lung has lower electron density than soft tissue, there is reduced attenuation of the primary photons of a beam traversing lung compared with the same path length in soft tissue. Most inhomogeneity correction algorithms can account for this effect (144). However, other, more subtle effects are described with reasonable approximation only by superposition-convolution algorithms (146,147) and most accurately by Monte Carlo. The cause of these effects is the long range of the secondary electrons in lung compared to soft tissue (the range is approximately inversely proportional to the ratio of lung to soft tissue density). Energy is thus transported outside the beam's geometric edge, resulting in a broader beam penumbra and reduced dose within the beam. Also, especially for a small soft tissue target embedded in a very low density medium and irradiated with a tight, high energy beam, there is a build-down (low dose region) at the entrance surface and sides of the target. All these effects are more pronounced for higher energy beams and lower density lungs (longer electron ranges) and smaller fields. The clinical concern is that treatment plans developed with algorithms that do not account for these effects can result in target underdose and/or overdose to normal tissues in penumbral regions. Figure 6 shows characteristic differences between the dose distribution of a single 6 MV photon beam predicted by a measurement-based pencil beam calculation that accounts only for changes in primary attenuation and that predicted by a Monte Carlo calculation. Lung radiation treatments usually consist of two or more beams, incident on the tumor in a cross-fire technique. Figure 7 shows that even in this patient's full four-field plan, these characteristic differences between the Monte Carlo and pencil beam calculations persist.

The degree to which the target underdose and broader penumbra in lung may compromise complications-free tumor control has been addressed in several studies (24,33,34,148–155). References 148–151 used measurements only to investigate penumbra broadening and build-down effects. A recent study (154) combined film dosimetry and EGSnrc and DOSXYZnrc Monte Carlo calculations to study the dose distribution in a 2 $\times$ 2 $\times$ 2 cm acrylic ($\sim$ tissue density) cube embedded in cork, simulating a small lesion in lung irradiated with a single and with parallel opposed photon beams from 4 to 18 MV. The parallel opposed geometry is a common field arrangement for treatment of lung tumors. Cork density, field size, and depth of the lesion in cork were varied. For the entire target cube to receive at least 95% of the dose to its center required field edges of the parallel opposed fields to be at least 2 cm from the cube even for the most favorable case (4 MV photons).

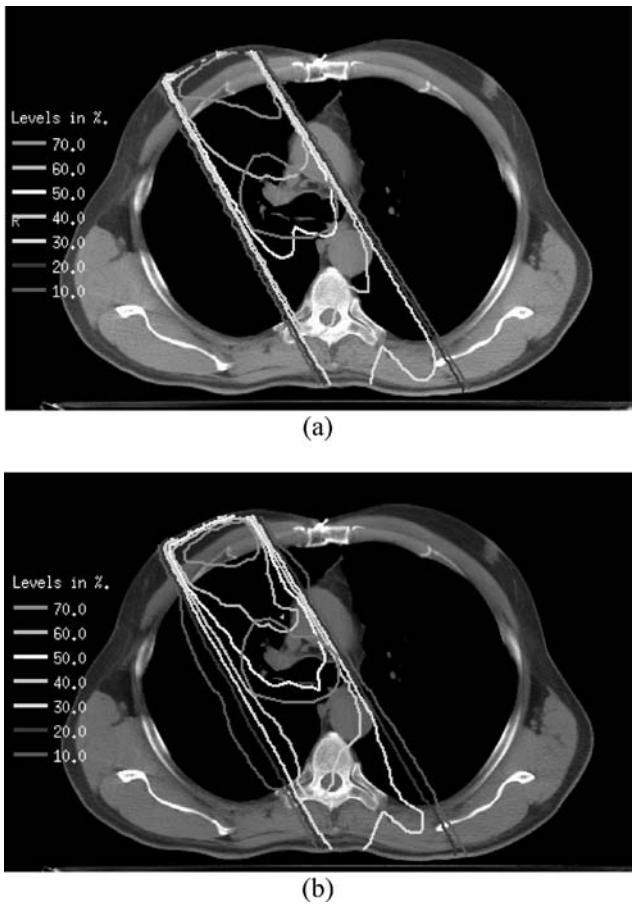


Figure 6. Dose distribution of a single 6 MV photon beam (a) predicted by a measurement-based pencil beam calculation that accounts only for changes in primary attenuation and (b) that predicted by a Monte Carlo calculation. The red contour indicates the target. Please use online version for color figure.

Other recent studies from different institutions have compared more complex treatment plans designed on anatomical phantoms or patient CT image sets and calculated with Monte Carlo versus the local treatment planning system calculation algorithm (24,34,152,153,155). In these studies, as in routine clinical practice, the beam is shaped to cover the planning target volume (PTV), which is larger than the grossly visible tumor gross target volume (GTV). The margin is intended to account for microscopic disease, setup error and breathing motion. Based on these studies, it is expected that (a) Results depend on the treatment planning system algorithm (34,153,155); (b) For the same planning system, results are patient (phantom geometry)

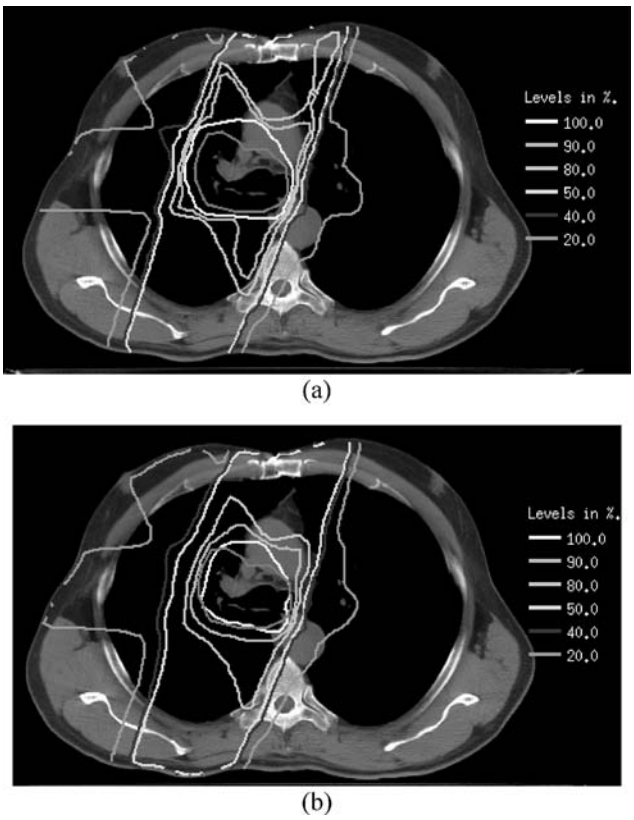


Figure 7. Dose distributions of a four-field plan in the lung. Figure (a) and (b) show dose distributions on a transverse plane predicted by a measurement-based pencil beam calculation and by a Monte Carlo calculation, respectively. The red contour indicates the target. Please see online version for color figure.

dependent as well as dependent on beam energy (24,33,34,153,155); (c) Changes quoted depend on the dosimetric coverage factor being evaluated. Mean target dose and dose encompassing 95% of the target volume are relatively insensitive indices; minimum dose (a single point) and dose-volume points on rapidly changing portions of the dose-volume histogram are more sensitive; (d) The PTV is usually underdosed relative to expectations from the treatment planning system. The degree of underdose varies from 1 to 20%, for 6 MV photons, depending on the dosimetric coverage factor, the lung density and the tumor location. For most of the cases reported, the underdosage is $< 10\%$; (e) The planning system results for coverage of the GTV are more similar to Monte Carlo, as is expected because the margin results in a larger distance from geometric beam edge to the GTV border than to the PTV border; (f) The greatest differences are for tumors surrounded by very low density lungs (33,34); (g) There are greater differences for high (e.g., 15 MV) than low (6 MV) energy beams (153); (h) Normal organ doses (primarily lung and spinal cord) are only slightly affected.

Head and Neck Cancer

H&N cancer radiation therapy usually includes photon irradiation with low megavoltage beams (^{60}Co—6 MV). Target tissue often borders on naturally occurring or surgical air cavities. Experiments demonstrate that build-down accompanying the loss of electronic equilibrium in air cavities in tissue-equivalent phantoms can cause up to a 25% underdose within the first millimeter of tissue (156–159), with particularly pronounced effects for small ($\leq 5 \times 5$ cm^2) fields, such as are used for treatment of larynx cancer. Whether this impacts on local control of larynx cancers treated with 6 MV beams versus ^{60}Co has not been resolved (160). The penumbra broadening and loss of dose within the beam that are noted in lung also occur in air cavities but the small size of these cavities, compared to the size of a lung, prevent these effects from posing a serious clinical problem.

Monte Carlo calculations compared well with parallel plate ion chamber measurements for single field and parallel opposed field irradiation (4, 6, and 8 MV photons) of a $4 \times 4 \times 4$ cm^3 cavity centered in a $30 \times 30 \times 16$ cm^3 phantom (161) though neither method had the spatial resolution to probe the build-down region in detail. A few studies have compared dose distributions on patient CT image sets for clinical beam arrangements as calculated with the local planning system and with Monte Carlo calculations for the same beams (31,33,162). Differences between the two calculation methods are more noticeable for individual beams than when all the beams (from two to seven, depending on the plan) are combined for the overall treatment plan. Monte Carlo predicts inferior target coverage compared to the planning system, but the differences, which depend on dosimetric index and tumor geometry, are less than in lung. Spinal cord maximum dose differences of < 1 Gy were reported in (31) (with the Monte Carlo calculation sometimes higher, sometimes lower) and 3 Gy higher as calculated by Monte Carlo in (162).

DISCUSSION

For treatment planning dose calculations, Monte Carlo is potentially the most accurate method. Monte Carlo dose calculation for electron beams has recently become available on a commercial treatment planning system (48). For photon beams, however, it has not been practical for routine clinical use due to its long running time. To improve the computation efficiency, there are variance reduction techniques available. The most common techniques are splitting and Russian roulette (136). In splitting, a particle is artificially split into multiple particles in important regions to produce more histories. In Russian roulette, particles are artificially terminated in relatively unimportant regions to reduce the number of histories. In both techniques, the particle weight, of course, needs to be adjusted to reflect the artificial increase or decrease of histories.

In addition to dose calculation, perhaps a more important application of Monte Carlo is to provide information that cannot be easily obtained by measurement. For example, in the simulation of the machine head, the phase-space data provide information on the primary and scattered radiation from various components in the machine head. These data provide important information in understanding the beam characteristics and may be used for other dose calculation methods.

BIBLIOGRAPHY

1. Andreo P. Monte Carlo techniques in medical radiation physics. Phys Med Biol 1991;36(7):861–920.
2. Mohan R, Chui C, Lidofsky L. Energy and angular distributions of photons from medical linear accelerators. Med Phys 1985;12(5):592–597.
3. Han K, et al. Monte Carlo simulation of a cobalt-60 beam. Med Phys 1987;14(3):414–419.
4. Chaney EL, Cullip TJ, Gabriel TA. A Monte Carlo study of accelerator head scatter. Med Phys 1994;21(9):1383–1390.
5. Lovelock DM, Chui CS, Mohan R. A Monte Carlo model of photon beams used in radiation therapy. Med Phys 1995 ;22(9):1387–1394.
6. Lee PC. Monte Carlo simulations of the differential beam hardening effect of a flattening filter on a therapeutic X-ray beam. Med Phys 1997;24(9):1485–1489.
7. Bhat M, et al. Off-axis X-ray spectra: a comparison of Monte Carlo simulated and computed X-ray spectra with measured spectra. Med Phys 1999;26(2):303–309.
8. Libby B, Siebers J, Mohan R. Validation of Monte Carlo generated phase-space descriptions of medical linear accelerators. Med Phys 1999;26(8):1476–1483.
9. Ma CM, Jiang SB. Monte Carlo modelling of electron beams from medical accelerators. Phys Med Biol 1999;44(12):R157–R189.
10. Siebers JV, et al. Comparison of EGS4 and MCNP4b Monte Carlo codes for generation of photon phase space distributions for a Varian 2100C. Phys Med Biol 1999; 44(12):3009–3026.
11. van der Zee W, Welleweerd J. Calculating photon beam characteristics with Monte Carlo techniques. Med Phys 1999;26(9): 1883–1892.
12. Deng J, et al. Photon beam characterization and modelling for Monte Carlo treatment planning. Phys Med Biol 2000;45(2): 411–427.
13. Bieda MR, Antolak JA, Hogstrom KR. The effect of scattering foil parameters on electron-beam Monte Carlo calculations. Med Phys 2001;28(12):2527–2534.
14. Antolak JA, Bieda MR, Hogstrom KR. Using Monte Carlo methods to commission electron beams: a feasibility study. Med Phys 2002;29(5):771–786.
15. Ding GX. Energy spectra, angular spread, fluence profiles and dose distributions of 6 and 18 MV photon beams: results of monte carlo simulations for a varian 2100EX accelerator. Phys Med Biol 2002;47(7):1025–1046.
16. Sheikh-Bagheri D, Rogers DW. Monte Carlo calculation of nine megavoltage photon beam spectra using the BEAM code. Med Phys 2002;29(3):391–402.
17. van der Zee W, Welleweerd J. A Monte Carlo study on internal wedges using BEAM. Med Phys 2002;29(5):876–885.
18. Ding GX. Using Monte Carlo simulations to commission photon beam output factors–a feasibility study. Phys Med Biol 2003;48(23):3865–3874.
19. Van de Walle J, et al. Monte Carlo model of the Elekta SLiplus accelerator: validation of a new MLC component module in BEAM for a 6 MV beam. Phys Med Biol 2003; 48(3):371–385.
20. Verhaegen F, Seuntjens J. Monte Carlo modelling of external radiotherapy photon beams. Phys Med Biol 2003;48(21):R107–R164.
21. Fix MK, et al. Monte Carlo source model for photon beam radiotherapy: photon source characteristics. Med Phys 2004; 31(11):3106–3121.
22. Pena J, et al. Commissioning of a medical accelerator photon beam Monte Carlo simulation using wide-field profiles. Phys Med Biol 2004;49(21):4929–4942.
23. DeMarco JJ, Solberg TD, Smathers JB. A CT-based Monte Carlo simulation tool for dosimetry planning and analysis. Med Phys 1998;25(1):1–11.
24. Wang L, Chui CS, Lovelock M. A patient-specific Monte Carlo dose-calculation method for photon beams. Med Phys 1998;25(6):867–878.
25. Jeraj R, Keall P. Monte Carlo-based inverse treatment planning. Phys Med Biol 1999;44(8):1885–1896.
26. Laub W, et al. Monte Carlo dose computation for IMRT optimization. Phys Med Biol 2000;45(7):1741–1754.
27. Lewis RD, et al. Use of Monte Carlo computation in benchmarking radiotherapy treatment planning system algorithms. Phys Med Biol 2000;45(7):1755–1764.
28. Keall PJ, et al. Monte Carlo dose calculations for dynamic IMRT treatments. Phys Med Biol 2001;46(4):929–941.
29. Li XA, et al. Monte Carlo dose verification for intensity-modulated arc therapy. Phys Med Biol 2001;46(9):2269–2282.
30. Shih R, Lj XA, Hsu WL. Dosimetric characteristics of dynamic wedged fields: a Monte Carlo study. Phys Med Biol 2001;46(12):N281–N292.
31. Wang L, Yorke E, Chui CS. Monte Carlo evaluation of tissue inhomogeneity effects in the treatment of the head and neck. Int J Radiat Oncol Biol Phys 2001;50(5):1339–1349.
32. Ma CM, et al. A Monte Carlo dose calculation tool for radiotherapy treatment planning. Phys Med Biol 2002;47(10): 1671–1689.
33. Wang L, Yorke E, Chui CS. Monte Carlo evaluation of 6 MV intensity modulated radiotherapy plans for head and neck and lung treatments. Med Phys 2002;29(11):2705–2717.
34. Yorke ED, et al. Evaluation of deep inspiration breath-hold lung treatment plans with Monte Carlo dose calculation. Int J Radiat Oncol Biol Phys 2002;53(4):1058–1070.
35. Leal A, et al. Routine IMRT verification by means of an automated Monte Carlo simulation system. Int J Radiat Oncol Biol Phys 2003;56(1):58–68.
36. Wieslander E, Knoos T. Dose perturbation in the presence of metallic implants: treatment planning system versus Monte Carlo simulations. Phys Med Biol 2003;48(20):3295–3305.
37. Heath E, Seuntjens J, Sheikh-Bagheri D. Dosimetric evaluation of the clinical implementation of the first commercial IMRT Monte Carlo treatment planning system at 6 MV. Med Phys 2004;31(10):2771–2779.
38. Yang J, et al. Modelling of electron contamination in clinical photon beams for Monte Carlo dose calculation. Phys Med Biol 2004;49(12):2657–2673.
39. Kawrakow I, Fippel M, Friedrich K. 3D electron dose calculation using a Voxel based Monte Carlo algorithm (VMC). Med Phys 1996;23(4):445–457.
40. Keall PJ, Hoban PW. Super-Monte Carlo: a 3-D electron beam dose calculation algorithm. Med Phys 1996;23(12): 2023–2034.
41. Scora D, Faddegon BA. Monte Carlo based phase-space evolution for electron dose calculation. Med Phys 1997;24(2): 177–187.
42. Jiang SB, Kapur A, Ma CM. Electron beam modeling and commissioning for Monte Carlo treatment planning. Med Phys 2000;27(1):180–191.
43. Kawrakow I. Accurate condensed history Monte Carlo simulation of electron transport. I. EGSnrc, the new EGS4 version. Med Phys 2000;27(3):485–498.
44. Lee MC, et al. Monte Carlo based treatment planning for modulated electron beam radiation therapy. Phys Med Biol 2001;46(8):2177–2199.
45. Bjork P, Knoos T, Nilsson P. Influence of initial electron beam characteristics on monte carlo calculated absorbed dose distributions for linear accelerator electron beams. Phys Med Biol 2002;47(22):4019–4041.

46. Deng J, Lee MC, Ma CM. A Monte Carlo investigation of fluence profiles collimated by an electron specific MLC during beam delivery for modulated electron radiation therapy. Med Phys 2002;29(11):2472–2483.
47. Doucet R, et al. Comparison of measured and Monte Carlo calculated dose distributions in inhomogeneous phantoms in clinical electron beams. Phys Med Biol 2003;48(15):2339–2354.
48. Cygler JE, et al. Evaluation of the first commercial Monte Carlo dose calculation engine for electron beam treatment planning. Med Phys 2004;31(1):142–153.
49. Coleman J, et al. A comparison of Monte Carlo and Fermi-Eyges-Hogstrom estimates of heart and lung dose from breast electron boost treatment. Int J Radiat Oncol Biol Phys 2005;61(2):621–628.
50. Carlsson AK, Andreo P, Brahme A. Monte Carlo and analytical calculation of proton pencil beams for computerized treatment plan optimization. Phys Med Biol 1997;42(6):1033–1053.
51. Paganetti H. Monte Carlo method to study the proton fluence for treatment planning. Med Phys 1998;25(12):2370–2375.
52. Fippel M, Soukup M. A Monte Carlo dose calculation algorithm for proton therapy. Med Phys 2004;31(8):2263–2273.
53. Jiang H, Paganetti H. Adaptation of GEANT4 to Monte Carlo dose calculations based on CT data. Med Phys 2004;31(10):2811–2818.
54. Paganetti H. Four-dimensional Monte Carlo simulation of time-dependent geometries. Phys Med Biol 2004;49(6):N75–N81.
55. Williamson JF, Morin RL, Khan FM. Monte Carlo evaluation of the Sievert integral for brachytherapy dosimetry. Phys Med Biol 1983;28(9):1021–1032.
56. Burns GS, Raeside DE. Monte Carlo simulation of the dose distribution around 125I seeds. Med Phys 1987;14(3):420–424.
57. Williamson JF. Monte Carlo evaluation of specific dose constants in water for 125I seeds. Med PhysM 1988;15(5):686–694.
58. Angelopoulos A, et al. Accurate Monte Carlo calculations of the combined attenuation and build-up factors, for energies (20–1500 keV) and distances (0–10 cm) relevant in brachytherapy. Phys Med Biol 1991;36(6):763–778.
59. Williamson JF, Li Z. Monte Carlo aided dosimetry of the microselectron pulsed and high dose-rate 192Ir sources. Med Phys 1995;22(6):809–819.
60. Weaver K, et al. A source model for efficient brachytherapy computations with Monte Carlo. Med Phys 1996;23(12): 2079–2084.
61. Cheung YC, et al. The dose distribution close to an 192Ir wire source: EGS4 Monte Carlo calculations. Phys Med Biol 1997;42(2):401–406.
62. Baltas D, et al. Application of the Monte Carlo integration (MCI) method for calculation of the anisotropy of 192Ir brachytherapy sources. Phys Med Biol 1998;43(6):1783–1801.
63. Daskalov GM, Loffler E, Williamson JF. Monte Carlo-aided dosimetry of a new high dose-rate brachytherapy source. Med Phys 1998;25(11):2200–2208.
64. Mainegra E, Capote R, Lopez E. Dose rate constants for 125I, 103Pd, 192Ir and 169Yb brachytherapy sources: an EGS4 Monte Carlo study. Phys Med Biol 1998;43(6):1557–1566.
65. Wang R, Sloboda RS. Monte Carlo dosimetry of the VariSource high dose rate 192Ir source. Med Phys 1998;25(4):415–423.
66. Karaiskos P, et al. A Monte Carlo investigation of the dosimetric characteristics of the VariSource 192Ir high dose rate brachytherapy source. Med Phys 1999;26(8):1498–1502.
67. Reynaert N, et al. Monte Carlo calculations of dose distributions around 32P and 198Au stents for intravascular brachytherapy. Med Phys 1999;26(8):1484–1491.
68. Casal E, et al. Monte Carlo calculations of dose rate distributions around the Amersham CDCS-M-type 137Cs source. Med Phys 2000;27(1):132–140.
69. Hedtjarn H, Carlsson GA, Williamson JF. Monte Carlo-aided dosimetry of the Symmetra model I25.S06 125I, interstitial brachytherapy seed. Med Phys 2000;27(5):1076–1085.
70. Li Z, Palta JR, Fan JJ. Monte Carlo calculations and experimental measurements of dosimetry parameters of a new 103Pd source. Med Phys 2000;27(5):1108–1112.
71. Mainegra E, Capote R, Lopez E. Radial dose functions for 103Pd, 125I, 169Yb and 192Ir brachytherapy sources: an EGS4 Monte Carlo study. Phys Med Biol 2000;45(3):703–717.
72. Mainegra E, Capote R, Lopez E. Anisotropy functions for 169Yb brachytherapy seed models 5, 8 and X1267. An EGS4 Monte Carlo study. Phys Med Biol 2000;45(12):3693–3705.
73. Williamson JF. Monte Carlo modeling of the transverse-axis dose distribution of the model 200 103Pd interstitial brachytherapy source. Med Phys 2000;27(4):643–654.
74. Ballester F, et al. Technical note: Monte-Carlo dosimetry of the HDR 12i and Plus 192Ir sources. Med Phys 2001;28(12):2586–2591.
75. Capote R, Mainegra E, Lopez E. Anisotropy function for 192Ir low-dose-rate brachytherapy sources: an EGS4 Monte Carlo study. Phys Med Biol 2001;46(5):1487–1499.
76. Rivard MJ. Monte Carlo calculations of AAPM Task Group Report No. 43 dosimetry parameters for the MED3631-A/M125I source. Med Phys 2001;28(4):629–637.
77. Chan GH, Prestwich WV. Monte carlo investigation of the dosimetric properties of the new 103Pd BrachySeedPd-103 Model Pd-1 source. Med Phys 2002;29(9):1984–1990.
78. Hedtjarn H, Carlsson GA, Williamson JF. Accelerated Monte Carlo based dose calculations for brachytherapy planning using correlated sampling. Phys Med Biol 2002;47(3):351–376.
79. Ibbott GS, Meigooni AS, Gearheart DM. Monte Carlo determination of dose rate constant. Med Phys 2002;29(7):1637–1638.
80. Bohm TD, DeLuca Jr PM, DeWerd LA. Brachytherapy dosimetry of 125I and 103Pd sources using an updated cross section library for the MCNP Monte Carlo transport code. Med Phys 2003;30(4):701–711.
81. Medich DC, Munro JJ. 3rd, Monte Carlo calculated TG-43 dosimetry parameters for the SeedLink 125Iodine brachytherapy system. Med Phys 2003;30(9):2503–2508.
82. Anagnostopoulos G, et al. The effect of patient inhomogeneities in oesophageal 192Ir HDR brachytherapy: a Monte Carlo and analytical dosimetry study. Phys Med Biol 2004;49(12):2675–2685.
83. Ballester F, et al. Monte carlo dosimetric study of best industries and Alpha Omega Ir-192 brachytherapy seeds. Med Phys 2004;31(12):3298–3305.
84. Lymperopoulou G, et al. A monte carlo dosimetry study of vaginal 192Ir brachytherapy applications with a shielded cylindrical applicator set. Med Phys 2004;31(11):3080–3086.
85. Reniers B, Verhaegen F, Vynckier S. The radial dose function of low-energy brachytherapy seeds in different solid phantoms: comparison between calculations with the EGSnrc and MCNP4C Monte Carlo codes and measurements. Phys Med Biol 2004;49(8):1569–1582.
86. Perez-Calatayud J, et al. Monte carlo and experimental derivation of TG43 dosimetric parameters for CSM-type Cs-137 sources. Med Phys 2005;32(1):28–36.
87. Furhang EE, Chui CS, Sgouros G. A Monte Carlo approach to patient-specific dosimetry. Med Phys 1996;23(9):1523–1529.

88. Tagesson M, Ljungberg M, Strand SE. A Monte-Carlo program converting activity distributions to absorbed dose distributions in a radionuclide treatment planning system. Acta Oncol 1996;35(3):367–372.
89. Liu A, et al. Monte Carlo-assisted voxel source kernel method (MAVSK) for internal beta dosimetry. Nucl Med Biol 1998; 25(4):423–433.
90. Clairand I, et al. DOSE3D: EGS4 Monte Carlo code-based software for internal radionuclide dosimetry. J Nucl Med 1999;40(9):1517–1523.
91. Zaidi H. Relevance of accurate Monte Carlo modeling in nuclear medical imaging. Med Phys 1999;26(4):574–608.
92. Chao TC, Xu XG. Specific absorbed fractions from the image-based VIP-Man body model and EGS4-VLSI Monte Carlo code: internal electron emitters. Phys Med Biol 2001;46(4): 901–927.
93. Kvinnsland Y, Skretting A, Bruland OS. Radionuclide therapy with bone-seeking compounds: Monte Carlo calculations of dose-volume histograms for bone marrow in trabecular bone. Phys Med Biol 2001;46(4):1149–1161.
94. Yoriyaz H, Stabin MG, dos Santos A. Monte Carlo MCNP-4B-based absorbed dose distribution estimates for patient-specific dosimetry. J Nucl Med 2001;42(4):662–669.
95. Ljungberg M, et al. A 3-dimensional absorbed dose calculation method based on quantitative SPECT for radionuclide therapy: evaluation for (131)I using monte carlo simulation. J Nucl Med 2002;43(8):1101–1109.
96. Kinase S, et al. Evaluation of specific absorbed fractions in voxel phantoms using Monte Carlo simulation. Radiat Prot Dosimet 2003;105(1–4):557–563.
97. Wolf I, et al. Determination of Individual S-Values for (131)I Using Segmented CT Data and the EGS4 Monte Carlo Code. Cancer Biother Radiopharm 2005;20(1):98–102.
98. Chan HP, Doi K. Radiation dose in diagnostic radiology: Monte Carlo simulation studies. Med Phys 1984;11(4):480–490.
99. Dance DR, Day GJ. The computation of scatter in mammography by Monte Carlo methods. Phys Med Biol 1984;29(3): 237–247.
100. Kulkarni RN, Supe SJ. Radiation dose to the breast during mammography: a comprehensive, realistic Monte Carlo calculation. Phys Med Biol 1984;29(10):1257–1264.
101. Kulkarni RN, Supe SJ. Monte Carlo calculations of mammographic X-ray spectra. Phys Med Biol 1984;29(2):185–190.
102. Boone JM, Seibert JA. Monte Carlo simulation of the scattered radiation distribution in diagnostic radiology. Med Phys 1988;15(5):713–720.
103. Papin PJ, Rielly PS. Monte Carlo simulation of diagnostic X-ray scatter. Med Phys 1988;15(6):909–914.
104. Gao W, Raeside DE. Orthovoltage radiation therapy treatment planning using Monte Carlo simulation: treatment of neuroendocrine carcinoma of the maxillary sinus. Phys Med Biol 1997;42(12):2421–2433.
105. Verhaegen F, et al. Monte Carlo modelling of radiotherapy kV X-ray units. Phys Med Biol 1999;44(7):1767–1789.
106. Boone JM, Cooper 3rd VN. Scatter/primary in mammography: Monte Carlo validation. Med Phys 2000;27(8):1818–1831.
107. Boone JM, Buonocore MH, Cooper 3rd VN. Monte Carlo validation in diagnostic radiological imaging. Med Phys 2000;27(6):1294–1304.
108. Ng KP, Kwok CS, Tang FH. Monte Carlo simulation of X-ray spectra in mammography. Phys Med Biol 2000;45(5):1309–1318.
109. Peplow DE, Verghese K. Digital mammography image simulation using Monte Carlo. Med Phys 2000;27(3):568–579.
110. Kramer R, et al. Backscatter factors for mammography calculated with Monte Carlo methods. Phys Med Biol 2001; 46(3):771–781.
111. Ay MR, et al. Monte carlo simulation of X-ray spectra in diagnostic radiology and mammography using MCNP4C. Phys Med Biol 2004;49(21):4897–4917.
112. Andreo P, Nahum A, Brahme A. Chamber-dependent wall correction factors in dosimetry. Phys Med Biol 1986;31(11): 1189–1199.
113. Rogers DW. Calibration of parallel-plate chambers: resolution of several problems by using Monte Carlo calculations. Med Phys 1992;19(4):889–899.
114. Ma CM, Nahum AE. Calculations of ion chamber displacement effect corrections for medium-energy X-ray dosimetry. Phys Med Biol 1995;40(1):45–62.
115. Ma CM, Nahum AE. Monte Carlo calculated stem effect correction for NE2561 and NE2571 chambers in medium-energy X-ray beams. Phys Med Biol 1995;40(1):63–72.
116. Mobit PN, Nahum AE, Mayles P. An EGS4 Monte Carlo examination of general cavity theory. Phys Med Biol 1997; 42(7):1319–1334.
117. Ferreira IH, et al. Perturbation corrections for flat and thimble-type cylindrical standard ionization chambers for 60Co gamma rays: Monte Carlo calculations. Phys Med Biol 1998;43(10):2721–2727.
118. Ferreira IH, et al. Monte Carlo calculations of the ionization chamber wall correction factors for 192Ir and 60Co gamma rays and 250 kV X-rays for use in calibration of 192Ir HDR brachytherapy sources. Phys Med Biol 1999;44(8):1897–1904.
119. Borg J, et al. Monte Carlo study of correction factors for Spencer-Attix cavity theory at photon energies at or above 100 keV. Med Phys 2000;27(8):1804–1813.
120. Seuntjens JP, et al. Absorbed-dose beam quality conversion factors for cylindrical chambers in high energy photon beams. Med Phys 2000;27(12):2763–2779.
121. Mazurier J, et al. Calculation of perturbation correction factors for some reference dosimeters in high-energy photon beams with the Monte Carlo code PENELOPE. Phys Med Biol 2001;46(6):1707–1717.
122. Fu Y, Luo Z. Application of Monte Carlo simulation to cavity theory based on the virtual electron source concept. Phys Med Biol 2002;47(17):3263–3274.
123. Mainegra-Hing E, Kawrakow I, Rogers DW. Calculations for plane-parallel ion chambers in 60Co beams using the EGSnrc Monte Carlo code. Med Phys 2003;30(2):179–189.
124. Piermattei A, et al. The wall correction factor for a spherical ionization chamber used in brachytherapy source calibration. Phys Med Biol 2003;48(24):4091–4103.
125. Rogers DW, Kawrakow I. Monte Carlo calculated correction factors for primary standards of air kerma. Med Phys 2003;30(4):521–532.
126. Siegbahn EA, et al. Calculations of electron fluence correction factors using the Monte Carlo code PENELOPE. Phys Med Biol 2003;48(10):1263–1275.
127. Capote R, et al. An EGSnrc Monte Carlo study of the micro-ionization chamber for reference dosimetry of narrow irregular IMRT beamlets. Med Phys 2004;31(9):2416–2422.
128. McCaffrey JP, et al. Evidence for using Monte Carlo calculated wall attenuation and scatter correction factors for three styles of graphite-walled ion chamber. Phys Med Biol 2004;49(12):2491–2501.
129. Sempau J, et al. Electron beam quality correction factors for plane-parallel ionization chambers: Monte Carlo calculations using the PENELOPE system. Phys Med Biol 2004;49(18): 4427–4444.
130. Nelson WR, Hirayama H, Rogers DWO. The EGS4 Code System. 1985; Stanford Linear Accelerator Center.
131. GEANT team, GEANT version 315. 1992, CERN-data handling division, report DD/EE/84-1 revision.

132. Baro J, et al. PENELOPE: an algorithm for Monte Carlo simulation of the penetration and energy loss of electrons and positrons in matter. Nucl Instrum Methods B 1995;100:31–46.
133. Ma C-M, et al. DOSXYZ users manual. Ottawa: NRCC. 1995.
134. Neuenschwander H, Mackie TR, Reckwerdt PJ. MC–a high-performance Monte Carlo code for electron beam treatment planning. Phys Med Biol 1995;40(4):543–574.
135. Rogers DW, et al. BEAM: a Monte Carlo code to simulate radiotherapy treatment units. Med Phys 1995;22(5):503–524.
136. Briesmeister JF. MCNP-A general Monte Carlo N-Particle transport code, version 4B. 1997; Los Alamos National Laboratory report LA-12625-M.
137. Sempau J, Wilderman SJ, Bielajew AF. DPM, a fast, accurate Monte Carlo code optimized for photon and electron radiotherapy treatment planning dose calculations. Phys Med Biol 2000;45(8):2263–2291.
138. VMC++, electron and photon Monte Carlo calculations optimized for Radiation Treatment Planning, in Advanced Monte Carlo for Radiation Physics, Particle Transport Simulation and Applications: Proceedings of the Monte Carlo 2000. In: Meeting Kling A, et al. editors. Berlin; Lisbon: Springer, 2001; p 229–236.
139. Hartmann Siantar CL, et al. Description and dosimetric verification of the PEREGRINE Monte Carlo dose calculation system for photon beams incident on a water phantom. Med Phys 2001;28(7):1322–1337.
140. Salvat F, Fernandez-Varea JM, DSempau J. PENELOPE-A code system for Monte Carlo simulation of Electron and Photon Transport. 2003; Issy-les-Moulineaux, France: OECD Nuclear Energy Agency.
141. van der Zee W, Hogenbirk A, van der Marck SC. ORANGE: a Monte Carlo dose engine for radiotherapy. Phys Med Biol 2005;50(4):625–641.
142. Guber W, et al. A geometric description technique suitable for computer analysis of both the nuclear and conventional vulnerability of armored military vehicles. Washington (DC): 1967.
143. Berger M. Monte Carlo calculation of the penetration and diffusion of fast charged particles, in Methods in Computational Physics. In: Alder B, Fernbach S, Rotenberg M, editors. New York: Academic; 1963. p 135–215.
144. Papanikolaou N, et al. Tissue inhomogeneity corrections for megavoltage photon beams. Medical Physics Publishing; 2004.
145. Fraass BA, Smathers J, Deye J. Summary and recommendations of a National Cancer Institute workshop on issues limiting the clinical use of Monte Carlo dose calculation algorithms for megavoltage external beam radiation therapy. Med Phys 2003;30(12):3206–3216.
146. Ahnesjo A. Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media. Med Phys 1989;16(4):577–592.
147. Mackie TR, Scrimger JW, Battista JJ. A convolution method of calculating dose for 15-MV X-ray. Med Phys 1985;12(2): 188–196.
148. Ekstrand KE, Barnes WH. Pitfalls in the use of high energy X-ray to treat tumors in the lung. Int J Radiat Oncol Biol Phys 1990;18(1):249–252.
149. White PJ, Zwicker RD, Huang DT. Comparison of dose homogeneity effects due to electron equilibrium loss in lung for 6 MV and 18 MV photons. Int J Radiat Oncol Biol Phys 1996;34(5):1141–1146.
150. Yorke E, et al. Dosimetric considerations in radiation therapy of coin lesions of the lung. Int J Radiat Oncol Biol Phys 1996;34(2):481–487.
151. Klein EE, et al. A volumetric study of measurements and calculations of lung density corrections for 6 and 18 MV photons. Int J Radiat Oncol Biol Phys 1997;37(5):1163–1170.
152. Miften M, et al. Comparison of RTP dose distributions in heterogeneous phantoms with the BEAM Monte Carlo simulation system. J Appl Clin Med Phys 2001;2(1):21–31.
153. Wang L, et al. Dosimetric advantage of using 6 MV over 15 MV photons in conformal therapy of lung cancer: Monte Carlo studies in patient geometries. J Appl Clin Med Phys 2002;3(1):51–59.
154. Osei EK, et al. EGSNRC Monte Carlo study of the effect of photon energy and field margin in phantoms simulating small lung lesions. Med Phys 2003;30(10):2706–2714.
155. Chetty I, et al. The influence of beam model differences in the comparison of dose calculation algorithms for lung cancer treatment planning. Phys Med Biol 2005;50:801–815.
156. Epp ER, Boyer AL, Doppke KP. Underdosing of lesions resulting from lack of electronic equilibrium in upper respiratory air cavities irradiated by 10MV X-ray beams. Int J Radiat Oncol Biol Phys 1977;2(7–8):613–619.
157. Beach JL, Mendiondo MS, Mendiondo OA. A comparison of air-cavity inhomogeneity effects for cobalt-60, 6-, and 10-MV X-ray beams. Med Phys 1987;14(1):140–144.
158. Niroomand-Rad A, et al. Air cavity effects on the radiation dose to the larynx using Co-60, 6 MV, and 10 MV photon beams. Int J Radiat Oncol Biol Phys 1994;29(5):1139–1146.
159. Ostwald PM, Kron T, Hamilton CS. Assessment of mucosal underdosing in larynx irradiation. Int J Radiat Oncol Biol Phys 1996;36(1):181–187.
160. Parsons JT, et al. Treatment of early and moderately advanced vocal cord carcinoma with 6-MV X-rays. Int J Radiat Oncol Biol Phys 2001;50(4):953–959.
161. Kan WK, et al. The effect of the nasopharyngeal air cavity on X-ray interface doses. Phys Med Biol 1998;43(3):529–537.
162. Seco J, et al. Head-and-neck IMRT treatments assessed with a Monte Carlo dose calculation engine. Phys Med Biol 2005;50: 817–830.

See also RADIATION DOSE PLANNING, COMPUTER-AIDED; RADIOTHERAPY TREATMENT PLANNING, OPTIMIZATION OF; STATISTICAL METHODS.

RADIATION THERAPY, QUALITY ASSURANCE IN

GLEN GEJERMAN
JOSEPH HANLEY
Hackensack University Medical
Hackensack, New Jersey

INTRODUCTION

The curative goal in radiation therapy is to deliver sufficient doses of tumoricidal radiation to a target volume while protecting the contiguous normal tissues. As radiation dose and accuracy of treatment delivery correlate with improved disease-free survival and avoidance of toxicity, quality control must be maintained throughout the planning and delivery of radiotherapy. The International Commission on Radiation Units and Measurements (ICRU) has recommended that treatment should be delivered to within 5% of the prescribed dose. Treatment planning and delivery is a multistep process that includes clinical decision making, patient immobilization, simulation, delineation of target and avoidance structures, determination of beam number and orientation, dose calculation, dosimetric scrutiny, patient set up, and treatment administration. In order to meet the ICRU's stringent recommendation, each of these steps must achieve better than 3% accuracy, and yet several studies have shown that geometric uncertainty

and dosimetric inaccuracy impact the clinical reality of radiotherapy. Patient misidentification, set-up variability, organ motion, block or multileaf collimation placement errors, and dosimetric miscalculation can lead to underdosing or overdosing the target volume and unintentional irradiation of normal surrounding tissue. Studies have shown that these types of errors occur at various stages in the treatment planning and delivery process, are often due to inadequate communication and mistakes in data transfer, and quality assurance procedures facilitate early detection of and reduction in their occurrence (1,2). A recent review of radiation therapy errors over a 5 year period at a major tertiary cancer center found that 44% of errors were due to field deviations, 38% due to incorrect use of beam modifiers, and 18% due to deviations from the prescribed dose. Once a quality improvement intervention that addressed several technological issues such as electronic charting was initiated, a significant impact in error reduction was noted (3). The numerous important quality assurance duties coupled with the increasing sophistication of radiation treatment planning and delivery systems, calls for an integrated comprehensive program that validates, verifies, and maintains accuracy throughout the entire process of radiation therapy delivery (4).

Treatment inaccuracies can be divided into systematic or treatment preparation variations (which include positioning errors, organ motion during treatment planning simulation, contouring errors, field shaping errors, and machine calibration errors) and random or execution variations (which include day-to-day patient misalignment and organ motion during treatment). A comprehensive quality assurance program must encompass both systematic and random uncertainty in treatment planning and delivery in order to minimize their occurrence. Although both types of errors lead to geometrical deviations of the target volume, they have different effects on the delivered dose. Systematic errors cause a displacement of the dose distribution away from the intended target and random errors lead to blurring of the dose distribution. The impact of systematic errors on target dose and the tumor control probability is therefore much greater than the impact of random execution variations (5,6). Identifying and correcting planning preparation errors early in the treatment process is critical in order to mitigate their impact on the treatment outcome.

To minimize treatment inaccuracies, it is essential that each radiation oncology department establish a "quality system" or quality assurance program to provide the organizational structure, responsibilities, procedures, processes, and resources for assuring the quality of patient management. A series of checklists to monitor compliance with the objectives of the program should be developed and applied. Due to the ever-changing nature of radiation oncology, this quality assurance program should be reviewed annually.

SIMULATION QA

The foundation of treatment planning is the simulation process. After the radiation prescription has been filled out to include the intended target, the organs to avoid, and the total and fractional doses, the patient undergoes treatment planning simulation, during which anatomical data is acquired, patient topography is measured, and the target volume and avoidance structures are delineated. Radiation therapy simulators use fluoroscopic, X-ray and computed tomography (CT) techniques to visualize internal anatomy in relation to external landmarks and can be divided into two broad categories: conventional or fluoroscopic simulators and CT simulators. These simulators can replicate the treatment machine geometry either physically, in the case of a conventional simulator, or virtually on a computer for CT simulation. The quality assurance program for the physical simulators must be parallel to that of the treatment machines, so that the geometric relationship between the treatment unit and the target volume can be accurately and consistently reproduced. To ensure the same accuracy in the case of virtual simulation, the treatment unit must be precisely modeled in the simulation computer.

To minimize intratreatment movement, and to ensure accurate daily positioning, patients are simulated (conventional or CT) in the treatment position with the use of special immobilization devices. These devices extend beyond the treatment site and rigidly immobilize the patient while providing them with support to enhance relaxation and minimize movement. Studies comparing set-up variations in immobilized versus free set up of patients note a significant reduction in positioning errors. In patients without custom immobilization, the percentage of fractions with set-up errors greater than 5 mm ranged from 17–57% and errors greater than 10 mm occurred in 15% of fractions (7,8). A randomized trial analyzing patients in the prone position receiving pelvic radiotherapy found a statistically significant benefit when using rigid immobilization. In the group treated without immobilization, 31% of port films had isocenter deviations greater than 10 mm compared with 11% in the immobilized patients. Average set-up deviations in the anteroposterior, right-left, and superior-inferior directions were 5.2 mm, 3.2 mm, and 4.3 mm in the patients treated without immobilization versus 2.9 mm, 2.1 mm, and 3.9 mm in those treated with rigid immobilization, respectively (9). Patient-related uncertainties also include organ motion. The patient's treatment position can impact the extent of both inter- and intrafractional movement. A randomized trial analyzing organ motion during treatment demonstrated less prostate motion in the supine treatment position. The mean anterior-posterior organ motion was 0.1 mm for patients treated in the supine position as opposed to 0.7 mm in those treated prone (10). These data demonstrate why proper immobilization is such a vital part of the quality assurance program. It should be noted that the integrity of immobilization devices should be checked on a daily basis during the treatment course.

Quality assurance of the conventional simulator is necessary to avoid inaccuracies that could lead to target-beam misalignment. After installation, and prior to clinical use, a detailed customer acceptance procedure is often performed and can act as a baseline for ongoing testing. A complete QA program for a simulator should follow the guidelines detailed in the American Association of Medical Physicists (AAPM) Task Group 40 (TG40) report (4) and be

augmented by any city, state, or federal guidelines. Table 3 of TG40 specifies the tests to be performed and at what frequency and to what tolerance to perform them. As a minimum, the lasers and the optical distance indicator should be checked daily. On a monthly basis, evaluation of mechanical uncertainties includes a check of field size settings and rotational settings, light-field radiation field congruence, treatment couch movement, laser positioning, and cross-hair centering. An example of a monthly QA checklist is shown in Fig. 1. A more thorough series of mechanical checks should be performed on an annual basis. These checks include determination of true axes of rotation of the simulator and the coincidence of these axes. Annual tests of the X-ray generator are essential to ensure that the exposure to the patient is minimized. It is important to check the imaging chain as image quality can directly affect patient care. Specialized phantoms can be used to determine the spatial and contrast resolution. Conventional simulators are equipped with X-ray and fluoroscopic modes, and both modes should be tested.

During conventional simulation, fluoroscopy is used to determine the treatment portals to cover the appropriate target volume. When using fluoroscopic simulation, one must be certain that the field size and shape adequately encompasses the target volume. Without detailed knowledge of the true extent of the tumor volume, fluoroscopically determined treatment portals may result in inadequate coverage. In an analysis of patients with cervical cancer, reconstruction of CT-defined tumor volumes on the simulation films demonstrated inadequate anterior and posterior borders in a significant proportion of patients (11). When the simulation and target localization is completed, the position of the treatment field's isocenter is marked on the patient's skin or immobilization apparatus. In order to be able to consistently reproduce the treatment set up established at the simulation, three or more laser beams are used to establish fiducial marks often called triangulation points on the patient. A detailed recording of these simulation parameters, including gantry and collimator settings and the source-skin distance (SSD) for each treatment field, in the treatment chart will allow for accurate repositioning in the treatment room (12). The data acquired at the time of simulation can also be directly captured into a Record & Verify system (R&V), obviating the need for manual entry with the potential of transcription error. At completion of the simulation, a set of simulation films that show the field size, field isocenter, and projected anatomy from the chosen beam direction and distance is obtained. These films are used as the standard by which the future port films in the treatment room will be measured for set-up accuracy and to assess patient movement.

Hackensack University Medical Center
Department of Radiation Oncology

Simulator Machine QA Report - ,200

√ = Satisfactory X = Exceeds tolerance / Adjusted
NT = not tested A = Within tolerance / Adjusted

	Test	Tolerance (reference)	HUMC Sim
1	Laser Alignment	2 mm (1)	
2	Gantry Angle indicator	1 degree (1)	
3	Collimator Angle indicator	1 degree (1)	
4	Couch Displacement	1 mm (2)	
5	Couch Rotation	1 degree (2)	
6	FAD Readout	2 mm (1)	NA
7	Optical Distance indicator	2 mm (1)	
8a	Crosshair Centricity - Light Field	2 mm diameter (1)	
8b	Collimator Centricity - (Dllneators)	2 mm (2)	
9	Field Size indicator (Dellneators)	2 mm (1)	
10	Orthogorality of Dellneators	1.5 mm (2)	
11	Crosshair Centricity - Radiation Field	4 mm (2)	
12	Light - Raiation Concidence	2 mm (3)	
13	Safety Check &	Functional (1)	
	Reproting Physicist	NA	

Notes:

References:
(1) Recommended by AAPM TG-40 (4).
(2) Adopted from manufacturer's specification and/or clinical considerations.
(3) Taken from Report No. 13, Physical Aspects of Quality Assurance in Radiotherapy. American Association of Physicists in Medicince, May, 1984, New York.

Varian Acuity S/N 0114
Summary of Monthly QC Reviewed By: _____
Mechanical and Safety Checks Page 5

Figure 1. Monthly checklist for simulator machine QA.

Over the past several years, CT simulators have been replacing or used in conjunction with conventional simulators. The CT simulator acquires CT images of a patient and sends them to a computer workstation on which a virtual simulation can be performed. The patient's anatomy can be reconstructed in 3D allowing for a display in a beam's eye view (BEV) perspective of the target and its relation to the normal surrounding tissues in different treatment angles. The BEV is used to create beam apertures that geometrically conform to the projections of the target and normal anatomy through different treatment angles. The computer software can be used to define the treatment isocenter, and a CT-generated virtual image of the patient is then used to complete the simulation with a digitally reconstructed radiograph (DRR) or a digitally composited radiograph (DCR). DRRs are computed radiographs that use the CT simulation data to provide planar reference images with the target volume, organs at risk, isocenter, and field edges shown. The DCR is created by computer enhancement or suppression of the CT numbers that allow for better visualization of the targeted organs. Although the principles of patient immobilization, treatment field delineation, and patient coordinate marking are similar to those of conventional simulation, the digital nature of CT simulation requires additional quality controls. During the acceptance testing, a CT dataset of a humanoid phantom is used for treatment planning to assess contouring capabilities, isocenter calculation, target localization, and DRR reconstruction and data transfer from the CT to the treatment planning system. The phantom is then used to test field size accuracy, virtual

gantry and collimator rotation, and the ability to accurately shift the isocenter. In addition to standard QA procedures for conventional CT scanners, CT simulators require interval testing of the laser system and of the data link to the virtual simulation computer system that allows tumor contouring, isocenter and field size definition, transfer of coordinates to the patient's skin, as well as construction of the DRR (13). Upon completion of virtual simulation, the patient's images, contours, and treatment beams are electronically sent to the treatment planning system.

TREATMENT PLANNING QA

Modern treatment planning systems consist of complex software run on sophisticated platforms with multiple peripheral devices. Recognizing the challenge of ensuring proper maintenance and use of these increasingly complicated systems, the AAPM Task group 53 (TG53) published a comprehensive set of quality assurance guidelines that can be applied to clinical radiotherapy planning (14). Acceptance testing and commissioning of a treatment planning system provides the benchmark by which the system will be evaluated during the periodic quality assurance testing. Acceptance testing is performed after installation but prior to clinical use of the system. The process entails testing that the system's hardware, software, and peripheral devices function according to manufacturer's specifications. These tests ensure that the system can properly acquire patient data, process anatomical contouring, orient beam direction, perform dose calculation, display the resultant isodose plots, and print hard copies of the approved treatment plan's parameters. The ability to properly transfer imaging data is confirmed by scanning phantoms of known geometry with internal markers and transferring the imaging data to the treatment planning system. The transferred data is then compared with film images to validate orientation, measurement, and fiducial positioning. System commissioning involves extensive testing of the dosimetric algorithms for a variety of clinical scenarios. The physical properties of each treatment unit have to be entered into the system and checked for consistency with manufacturer's specifications. Data such as percent depth-dose tables, off-axis profiles, and output factors are acquired using a computer-controlled water phantom for each treatment beam on each treatment unit to be used in the planning system. Phantoms with known geometric target volumes are used to simulate common clinical scenarios and treatment plans are evaluated to verify calculated dose distributions. Although anthropomorphic phantoms (that are shaped like the human body) are well-suited to test clinical treatment techniques, geometric phantoms (that are cylindrical or cubic) have more reliable ionization chamber positioning (15). If dosimetric calculations that account for inhomogeneities within the patient are to be performed, a CT number to electron density calibration curve must be established, which is performed for each CT acquisition unit that sends images to the treatment planning system. A phantom containing plugs of known electron density is scanned on the CT and the corresponding CT number is determined. These numbers are plotted versus electron density to derive the calibration curve for that scanner. As an incorrect conversion of CT number to electron density can lead to significant dosimetric miscalculations, the American College of Radiology (ACR) recommends testing this calibration curve monthly. Although the most accurate form of dose calculations are Monte-Carlo-based, these calculations are computationally intensive and cannot currently be used for routine planning. All other dose calculation algorithms used in treatment planning systems have limitations, and it is essential to understand where these limitations manifest, for example, in areas where electronic equilibrium does not exist, such as lung-tissue interfaces. Routine periodic quality assurance testing of the planning system consists of daily, monthly, and annual tests. Daily tests validate the performance of input devices such as point digitizers and the accuracy of output devices such as printers. Monthly tests can involve calculating computer checksums for the treatment planning software executables and machine data, to ensure the program and data has not been modified. Annual tests are more involved and should include a subset of standard treatment plans that cover a wide range of clinical scenarios ranging from point-dose calculations, 2D, 3D conformal radiation therapy (3DCRT), and intensity-modulated radiation therapy (IMRT) plans. This set of standard plans are used for testing whenever software upgrades, either patches or version changes, are applied.

Once the imaging data has been transferred to the treatment planning system, the target volume and avoidance structures must be delineated if not already defined at the CT simulation. This delineation can be the major contributor to overall uncertainty in the treatment planning chain, as many factors exist that contribute to this uncertainty. It is imperative that the treating physician and the radiation treatment planner share a common vocabulary regarding the tumor volume and the additional margins necessary to account for organ motion and set-up inaccuracies. Prescribing and designing a treatment plan to a target without correcting for geometric uncertainties will result in a substantially different delivered dose than the intended one. In order to address these issues, the ICRU Report 50 (16) recommended using specific definitions regarding margins and volumes. The gross tumor volume (GTV) represents the visible tumor. The clinical target volume (CTV) denotes the GTV with an additional region encompassing suspected microscopic spread. The planning target volume (PTV) contains the CTV with margins added to account for geometric uncertainties. These margins are determined based on the extent of uncertainty caused by patient and tumor movement as well as the inaccuracies in beam and patient setup. Several margin recipes based on geometrical uncertainties and coverage probabilities have been published; however, their clinical impact remains to be proven (17). The organs at risk (OAR) are the normal tissues that are contiguous with the CTV (such as small bowel, rectum, and spinal cord) whose radiation tolerance can affect the maximum deliverable dose and treatment technique. The ICRU report 62 (18) refined the definition of the PTV with the concepts of

internal margin and set-up margin. Internal margin uncertainty that is caused by physiological changes such as respiratory movement cannot be easily modified without using respiratory gating techniques. In contrast, set-up margin uncertainty can be more readily minimized by proper immobilization and improved machine accuracy. The report also addressed the issue of OAR mobility by introducing the planning organ at risk volume (PRV) in which additional margins are added to account for the geometric uncertainty of these organs. In order to avoid significant radiation toxicity and to maintain post treatment quality of life, the planning physician must be vigilant when considering avoidance structures. In a Radiation Therapy and Oncology Group (RTOG) analysis of the impact of dose escalation in prostate cancer, a lack of physician awareness leading to unnecessary exposure of the penile bulb to high radiation doses lead to treatment-induced impotence (19).

Even with a common terminology and attention to detail when delineating the anatomical structures, several uncertainties exist that are related to the imaging modality used for data acquisition. Proper acquisition of CT data is challenging in that numerous factors, including slice thickness, slice spacing, CT number scale, and organ motion, can affect this information resulting in dosimetric and anatomic inaccuracies. A CT image artifact known as partial volume averaging occurs when two structures of different tissue density occupy the same voxel resulting in an averaging of their CT numbers. Unless the appropriate CT slice thickness is used, accurate target delineation can be compromised, as details of contiguous anatomic structures may not be appreciated. CT imaging of a moving organ can lead to significant distortions, particularly when the organ is small compared with the extent of its displacement. When the scan time is protracted, the artifact can be significant enough to render the reconstructed images unrecognizable in relation to its stationary counterpart (20). TG 53 recommends the use of imaging protocols that standardize scan parameters such as patient position and immobilization, CT slice spacing and thickness, the extent of the patient's anatomy to be scanned, breathing techniques for patients with abdominal or thoracic tumors, and the use of contrast agents (14). Some anatomical structures are better visualized using alternate imaging modalities such as Magnetic Resonance Imaging (MRI) or functional imaging such as Positron Emission Tomography (PET) scans. For example, to improve the accuracy of thoracic GTV recognition, PET scans have been used in conjunction with CT-based simulation. Although in some circumstances, the ability to distinguish between thoracic tumor and atelectasis can result in a smaller GTV (21); at other times subclinical mediastinal adenopathy appreciated on PET will require enlarging the treatment field to encompass all active disease (22). When multiple images sets, acquired with different imaging modalities, are used in the planning process, the images must be accurately correlated in a common frame of reference. Typically, the images sets are "fused" onto the CT frame of reference. In visual fusion, the independent images are studied side by side and are visually fused using data from both to outline the GTV. In software fusion, the independent studies are geometrically registered with each other using an overlay of anatomic reference locations. A recent review found that software fusion reduced intra- and interobserver variability and resulted in a more consistent delineation of tumor volume when compared with visual fusion (23). It is imperative to perform QA on the fusion software. Acquiring datasets of a phantom with known geometrical landmarks on all modalities to be tested and performing the fusion process can accomplish this goal. PET/CT scanners that obtain both images simultaneously allowing for self-registration are becoming more widely available and will further facilitate accuracy in contouring.

In addition to uncertainties associated with various imaging modalities, it is well documented that inter- and intraobserver reproducibility exists in GTV delineation, and significant differences in the size of the GTV are noted depending on the imaging modality used (24,25). When contouring CT images, the correct window level settings must be used to appreciate the extent of the tumor shape and its relation to contiguous organs at risk. The treatment planning CT must be carefully reviewed to assess for positional or anatomic anomalies. For example, data acquired in the thoracic or abdominal region should be carefully examined for any sharp discontinuities in the outer contour that might indicate a change in breathing pattern or physical shift of the patient due to coughing, for example. A retrospective review of prostate cancer patients treated with conformal radiotherapy found an association between rectal distension on the planning CT and decreased probability of biochemical cure. Planning with a distended rectum can result in a systematic error in prostate location and was found to have a greater impact on outcome than disease risk group (26).

Once all the relevant organs have been contoured and the target dose and dose constraints have been unambiguously communicated to the dosimetry team, the appropriate combination of beam number, beam direction, energy, and intensity is determined. These parameters are optimized to deliver maximum dose to the CTV and minimum dose to the OAR. Conventional dosimetric calculations known as forward planning involves an experienced planner choosing multiple beams aimed at the isocenter and altering beam orientation and weighting to achieve an acceptable plan. The dose delivered with the chosen beam arrangement will be affected by the interaction of the radiation beam with the patient's tissue density and is calculated by the planning computer. 3D conformal radiotherapy planning uses CT data to generate tumor and normal organ 3D images and displays them from the perspective of different angles using a BEV technique. Optimization of the treatment plan is performed by iteratively adjusting the beam number and direction, selectively adjusting the field aperture, and applying compensators such as wedges. In contrast, IMRT uses inverse planning to deliver a desired dose to the GTV and PTV with constraints to the OAR. Instead of choosing beam directions and then evaluating the resultant dosimetry, the desired dose distribution is stipulated using dose-volume constraints to the PTV and OAR and then the computer algorithm alters the various beam intensities in an attempt to achieve these planning goals.

After the dosimetry team completes their calculations, the proposed treatment plan must be carefully evaluated to confirm the prescription fractional and total doses and to determine whether it satisfies the prescription goal. For conventional treatments, this determination is performed by inspecting 2D isodose displays through one or more cross sections of the anatomy. For 3DCRT and IMRT, BEV data and dose volume histogram (DVH) analysis is used in addition to the isodose displays to evaluate dose minima, maxima, and means of both target and avoidance structures. The DVH that graphically depicts the percentage of a volume of interest that receives a particular dose does not give spatial information regarding dose distribution. If the DVH indicates underdosing, only by reviewing the plan's isodose display can one locate the area of inadequate coverage. Mathematical models that use DVH statistics to estimate the normal tissue complication probability (NTCP) have been developed. These NTCP models have been found to more accurately predict the likelihood of radiation-induced toxicity than point-dose radiation tolerance data. An important task in a quality assurance program is to calculate the fractional and total doses to the OAR in order to estimate the risk of radiation injury. These doses can be described in terms of minimum, maximum, and mean doses to an entire organ or as the volume of an organ receiving greater than a particular dose. In situations where the PTV anatomically overlaps the OAR, clinical judgment must be used to assign a priority to each goal. The location and volume of dosimetric inhomogeneity (both hot spots and cold spots) must be evaluated, which is particularly true for IMRT where dose homogeneity is often sacrificed for dose conformality.

When the plan has been approved by the dosimetrist and the physician, all documented parameters including patient setup, beam configuration, beam intensity, and monitor units are sent to a R&V system either manually or, preferably, electronically. All the data from the plan, printouts, treatment chart, and R&V undergoes an independent review by a qualified medical physicist. This second check entails review of the prescription, the plan's calculation algorithm, wedge placement, dose distribution, DVH, and beam apertures. Hand calculations of a point dose in each field are analyzed to verify the dosimetry. The patient then undergoes a verification simulation to confirm the accuracy and reproducibility of the proposed plan. During this confirmatory simulation, the isocenter position is radiographically confirmed, block geometry is checked, and measurements such as SSD are validated. Finally, a pretreatment port film verification is obtained on the treatment machine to verify reproducibility of set up and to confirm measurements such as the SSD and distance to tabletop.

The verification of patient-specific dose distributions with water phantoms, ionization chambers, diodes, or film dosimetry is an essential component of the QA process. The standard method of evaluation consists of overlaying hardcopy plots of measured and calculated isodose distributions and qualitatively assessing concordance. As a result of the nonuniform intensity inherent in IMRT and the resulting steep dose gradients throughout the treatment field, IMRT plan verification is more challenging. Radiographic film dosimetry can be used to verify the IMRT leaf sequences and monitor units; however, as film sensitivity varies with beam energy, field size, film positioning, and film processing, great care must be taken to normalize the calculated and measured dose. Computer-assisted registration techniques are now available to determine the relative difference between the planned and delivered individual beam fluence or combined dose distributions on a pixel-by-pixel basis in order to score the plan using a predetermined criterion of acceptability (27).

LINEAR ACCELERATOR QA

The quality assurance protocol for linear accelerators is designed to monitor and correct performance of the equipment so that the physical and dosimetric parameters established during commissioning and acceptance testing can be maintained. TG40 (4) described a thorough QA program for linear accelerators with recommended test frequency. The daily tests include checking the safety features such as the door interlock and audiovisual intercom systems. Mechanical performance such as the localizing lasers and the machine's optical distance indicator and dosimetric output such as the X-ray and electron constancy is also checked daily. Monthly checks of the linear accelerator's mechanical accuracy include light-field coincidence; cross hair centering; gantry, collimator, field size, and couch position indicators; latching of the electron cone, wedge, and blocking trays; electron cone interlocks; and the emergency off switch. An example of a monthly mechanical and safety checklist for a linear accelerator is shown in Fig. 2. Monthly checks of the dosimetric accuracy include constancy of the X-ray and electron output, central axis parameters, and X-ray and electron beam flatness. Annual mechanical tests include checks of the safety locks; the tabletop sag; vertical travel of the treatment couch; the collimator's, gantry's, and couch's rotation isocenter; the coincidence of the radiation and mechanical isocenter; and the coincidence of the collimator gantry and couch axes with the isocenter. The annual dosimetry tests check for monitor chamber linearity; wedge transmission factor constancy; off-axis factor constancy; and X-ray and electron output and off-axis constancy dependence on gantry angle. In addition, a subset of the depth-dose and off-axis profile scans acquired at commissioning are performed and compared with the baseline.

The use of multileaf collimators (MLC) in 3D conformal and intensity-modulated radiotherapy requires additional QA measures. When using MLC for 3DCRT, leaf position inaccuracies will have an effect on the resultant dosimetry; however, because of the PTV margins, the effect is minimal. In contrast, when using MLCs for IMRT, a miscalibration of 0.5 mm causing a 1 mm error in radiation portal size can cause a 10% dose error when delivering IMRT with an average field size of 1 cm (15). As MLC function is critical to dosimetric accuracy, rigorous QA protocols are required. The accuracy of the multileaf collimator (MLC) is verified by using radiographic film to measure radiation dose patterns and by checking for a gap between the leaves when they are programmed to be in the closed position.

Hackensack University Medical Center
Department of Radiation Oncology

Treatment Machine QA Report - ,200

√ = not tested
NT = not tested

X = Exceeds tolerance / Adjusted
A = Within tolerance / Adjusted

	Test	Tolerance (referance)	HUMC 21EX
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			

Notes:

Rerferance:
(1)
(2)
(3)
(4)
(5)
(6)
(7)

VARIAN CLIMAC 21EX — SN2193
Summary of Monthly QC — Month/Year_____
Mechanical and Safety Checks — Page 8

Figure 2. Monthly mechanical and safety checklist for a linear accelerator.

TREATMENT DELIVERY QA

After the linear accelerator's function has been verified with a thorough QA program, it is important to confirm that the treatment parameters established during treatment planning are accurately transferred to the treatment machine and that the daily set up and treatment are accurately executed. The R&V system is an important tool that confirms that proper field size, beam arrangement, multileaf collimator settings, collimator angle, gantry angle, beam energy, wedges, and monitor units are used each day. As an R&V system will give a daily validation of the entered parameters, it is essential to verify the set-up data with an independent check on the first day of treatment. The quality assurance mechanism for daily radiation treatments includes laser alignment of the fiducial tattoos and validation of SSD and tabletop measurements as well as weekly port films. These port films are obtained prior to treatment initiation as well as weekly and are compared with the initial simulation film or with the DRR to evaluate isocenter location and block or MLC position. A recent advance in set-up verification and error detection is the development of electronic portal imaging devices (EPID). As a replacement to port films, the EPID provides faster acquisition time, tools that allow digital image enhancement, and tools that can measure the distance of anatomic landmarks to the isocenter or field edge (28). The use of EPID has allowed for the potential of adaptive radiotherapy that collects geometrical uncertain information during the first few treatment fractions and sends it back to the treatment planning system for further optimization. Once the set-up variations have been characterized, the treatment is adapted to either adjust the field size and treatment couch position or adjust the fluence profiles to correct for the set-up error.

SUMMARY

An overview of the major components of quality assurance for external beam radiation therapy has been given. The need to establish a quality assurance program, which provides the organizational structure, responsibilities, procedures, processes, and resources for assuring the quality of patient management, has been demonstrated. The various components that contribute to treatment inaccuracies have been identified as systematic or random variations, and it was established that the impact of systematic errors on target dose and the tumor control probability is much greater than the impact of random variations. The simulation was described as the foundation of treatment planning process and intra- and interfraction patient motion was addressed with patient immobilization. QA programs for simulators, linear accelerators, and CT scanners should follow the guidelines provided in AAPM TG40 (4). QA for treatment planning systems is outlined in AAPM TG53 (14). The many factors that contribute to target delineation uncertainty, namely, acquisition parameters, organ motion, imaging modality, image fusion, and intra-observer variability, should all be examined closely for their contributions to treatment uncertainties.

As a result of the ever-changing nature of radiation oncology, this quality assurance program should be reviewed annually. Special attention should be given to new devices and treatment protocols. As adaptive radiation therapy evolves, a whole new component of the quality assurance program will need to be developed.

BIBLIOGRAPHY

1. Yeung TK, et al. Quality assurance in radiotherapy: Evaluation of errors and incidents recorded over a 10 year period. Radiother Oncol 2005;74:283–291.
2. Valli MC, et al. Evaluation of most frequent errors in daily compilation and use of a radiation treatment chart. Radiother Oncol 1994;32:87–89.
3. Huang G, et al. Error in the delivery of radiation therapy: Results of a quality assurance review. Int J Radiation Biol Phys 2005;61:1590–1595.
4. Kutcher GJ, et al. Comprehensive QA for radiation oncology: Report of AAPM radiation therapy committee task group 40. Med Phys 1994;21:581–618.

5. van Herk M, Remeijer P, Lebesque JV. Inclusion of geometric uncertainties in treatment plan evaluation. Int J Radiation Biol Phys 2002;52:1407–1422.
6. van Herk M, et al. The probability of correct target dosage: Dose-population histograms for deriving treatment margins in radiotherapy. Int J Radiation Biol Phys 2000;47:1121–1135.
7. Rosenthal SA, et al. Immobilization improves the reproducibility of patient positioning during six-field conformal radiation therapy for prostate carcinoma. Int J Radiation Biol Phys 1993;27:921–926.
8. Catton C, et al. Improvement in total positioning error for lateral prostatic fields using a soft immobilization device. Radiother Oncol 1997;44:265–270.
9. Kneebone A, et al. A randomized trial evaluating rigid immobilization for pelvic irradiation. Int J Radiation Biol Phys 2003;56:1105–1111.
10. Bayley AJ, et al. A randomized trial of supine vs. prone positioning in patients undergoing escalated dose conformal radiotherapy for prostate cancer. Radiother Oncol 2004;70:37–44.
11. Nagar YS, et al. Conventional 4 field box radiotherapy technique for cancer cervix: Potential for geographic miss without CECT scan based planning. Int J Gyn Ca 2004;14:865–870.
12. Reinstein LE. Patient positioning and immobilization. In: Kahn FM, Potish RA, editors. Treatment Planning in Radiation Therapy. Baltimore, MD: Williams and Wilkins Publishing; 1998. p 55–88.
13. McGee KP, Das IJ. Commissioning acceptance testing and quality assurance of a CT simulator. In: Coia LR, Schultheiss TE, Hanks GE, editors. A Practical Guide to CT Simulation. Madison, WI: Advanced Medical Publishing; 1995.p 5–23.
14. Fraass B, et al. American association of physicists in medicine radiation therapy committee task group 53: Quality assurance for clinical radiotherapy treatment planning. Med Phys 1998; 25:1773–1829.
15. Low DA. Quality assurance of intensity modulated radiotherapy. Semin Radiat Oncol 2002;12:219–228.
16. ICRU Report 50 Prescribing, recording, and reporting photon beam therapy. Bethesda, MD: International Commission on Radiation Units and Measurements; 1993.
17. Rasch C, Steenbakkers R, van Herk M. Target definition in prostate, head and neck. Semin Radiat Oncol 2005;15:136–145.
18. ICRU Report 62. Prescribing, recording, and reporting photon beam therapy (Supplement to ICRU Report 50). Bethesda, MD: International Commission on Radiation Units and Measurements; 1999.
19. Roach M, et al. Penile bulb dose and impotence after three dimensional conformal radiotherapy for prostate cancer on RTOG 9406: Findings from a prospective multi-institutional phase I/II dose escalation study. Int J Radiation Biol Phys 2004;60:1351–1356.
20. Gagne IM, Robinson DM. The impact of tumor motion upon CT image integrity and target delineation. Med Phys 2004;31: 3378–3392.
21. Schmuecking M, et al. Image fusion of F-18 FDG pet and CT- is there a role in 3D radiation treatment planning of non small cell lung cancer? Int J Radiation Biol Phys 2000;48(Suppl): 130.
22. Kiffer JD, et al. The contribution of 18F fluoro-2-deoxy-glucose positron emission tomographic imaging to radiotherapy planning in lung cancer. Lung CA 1998;19:167–177.
23. Fox JL, et al. Does registration of PET and planning CT images decrease interobserver and intraobserver variation in delineating tumor volumes for non-small cell lung cancer? IJROBP 2005;62:70–75.
24. Leunens G, et al. Quality assessment of medical decision making in radiation oncology: Variability in target volume delineation for brain tumors. Radiother Oncol 1993;29:169–175.
25. Roach M, et al. Prostate volumes defined by magnetic resonance imaging and computerized tomographic scans for three-dimensional conformal radiotherapy. Int J Radiation Biol Phys 1996;35:1011–1018.
26. De Crevoisier R, et al. Increased risk of biochemical and local failure in patients with distended rectum on the planning CT for prostate cancer radiotherapy. Int J Radiation Biol Phys 2005;62:965–973.
27. Kapulsky A, Gejerman G, Hanley J. A clinical application of an automated phantom film QA procedure for validation of IMRT treatment planning and delivery. Med Dosim 2004;29: 279–284.
28. Herman M. Clinical use of electronic portal imaging. Semin Radiat Oncol 2005;15:157–167.

See also CODES AND REGULATIONS: RADIATION; RADIOTHERAPY TREATMENT PLANNING, OPTIMIZATION OF; X-RAY QUALITY CONTROL PROGRAM.

RADIATION, ULTRAVIOLET. See ULTRAVIOLET RADIATION IN MEDICINE.

RADIOACTIVE DECAY. See RADIONUCLIDE PRODUCTION AND RADIOACTIVE DECAY.

RADIOACTIVE SEED IMPLANTATION. See PROSTATE SEED IMPLANTS.

RADIOIMMUNODETECTION. See MONOCLONAL ANTIBODIES.

RADIOISOTOPE IMAGING EQUIPMENT. See NUCLEAR MEDICINE INSTRUMENTATION.

RADIOLOGY INFORMATION SYSTEMS

JANICE C. HONEYMAN-BUCK
University of Florida
Gainesville, Florida

INTRODUCTION

If asked to define the functionality of a radiology information systems (RIS) 5–10 years ago, one may have listed order entry, film management, charge capture, billing, patient and examination tracking, and possibly inventory management. These systems were in place long before Picture Archiving and Communication Systems (PACS) and the Electronic Medical Record (EMR) were in widespread use. Although PACS, and especially the EMR, are not globally implemented, it is an accepted premise that they will be globally implemented sometime in the future. Now, people often refer to radiology information systems as a suite of computers serving the myriad of functions required for an electronic radiology practice that might include the classic RIS, PACS, speech recognition, modality workflow, and the radiology portion of the EMR and the Hospital Information System (HIS). Generally, in this article RIS will be defined in the more classic sense, it is the computer that manages all

aspects of radiology orders. The RIS must now additionally perform the functions required to automate the workflow in a radiology department including examination ordering and management, modality scheduling, examination tracking, capturing charges, inventory management, electronic signatures, report distribution, and management reporting. Every transaction and interaction with the system must be recorded for auditing purposes and must help maintain the privacy and security of Protected Heath Information (PHI) for a patient. Furthermore, the RIS must interface seamlessly with a PACS, a speech recognition system, an HIS and an EMR. This is a nontrivial task in a multivendor installation.

This article focuses on the RIS as the center of the radiology department workflow management with the interfaces to other systems. Interface standards are introduced with examples and some developing concepts in healthcare as they pertain to radiology are discussed.

INFORMATION SYSTEMS IN RADIOLOGY

Figure 1 shows the typical workflow associated with a radiology study. The referring physician orders the study, typically through the HIS, but often in smaller institutions, through the RIS. The order is then sent to PACS and to the speech recognition system. It is available for the technologists as a virtual worklist and on PACS modalities through DICOM modality worklist. In this case, the modality worklist is supplied by a broker or translation system that creates the correct format from the order. After the examination is completed, it is sent to archives and PACS displays where it is interpreted by the radiologist. The radiologist is working from their own worklist of studies that are available according to their role in the department and which have not been dictated. A role may be defined as a radiologist who interprets chest studies or perhaps a radiologist who interprets CT studies. The radiologist dictates into a speech recognition system that appends the report to the data about the examination that exists in the system. The reports are sent to the RIS, closing the loop, and then are sent to the EMR. The dashed line from the PACS Archive to the EMR indicates that the images may be stored in the PACS archive with links to them in the EMR, so it is unnecessary to store the images in both systems. The PACS displays and databases and the speech recognition system are frequently from different vendors, so an interface between the two must be accomplished to be certain that the radiologist is dictating the report on the study they are viewing. This loose coupling of the report and imaging exam must be carefully developed and tested to be sure the reports are permanently attached to the correct exam. Of course when all the systems are purchased from a single vendor, a tighter coupling of data and images may be easier to accomplish. The magic number in a radiology system that ties all the information, images and reports for a specific study together is the accession number. This number should be unique for a study and should identify the patient, exam, date, time, report, images, contrast used and anything else that goes with that study. An accession number query should only get one accurate result.

THE RIS AND ITS ROLE IN RADIOLOGY WORKFLOW

Although the RIS appears to be a very small actor in radiology workflow, this is far from the truth. Note the number of interfaces, indicated by arrows, with the RIS as opposed to the other systems in Fig. 1. The RIS is the integral part of the total electronic radiology practice and without it there would be no connection with the rest of the healthcare enterprise (1,2).

The other systems in Fig. 1 are also important and must be interfaced carefully. Although at times the lines between the functionality of the various systems blur, each

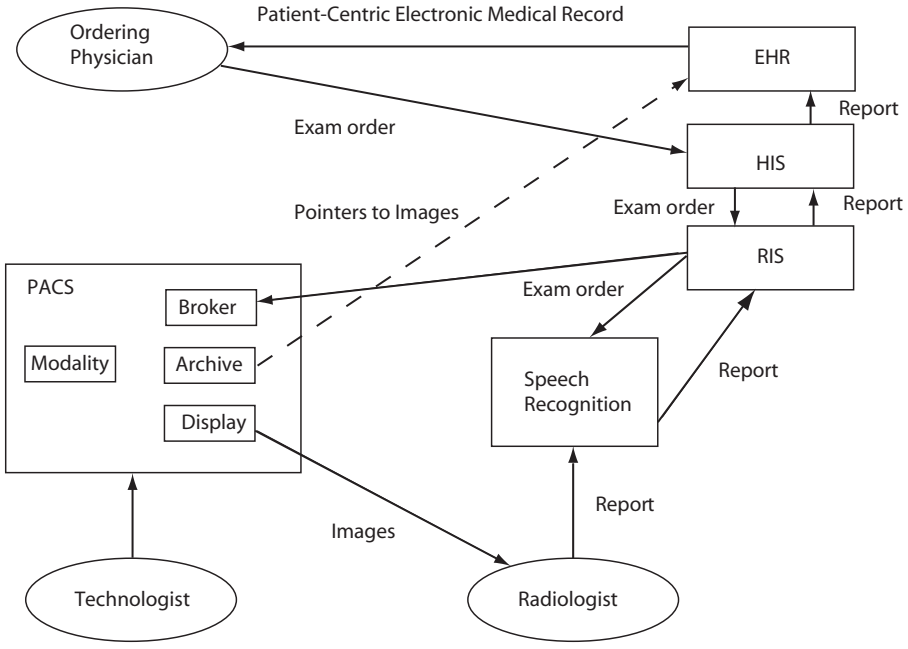


Figure 1. Interactions among information systems during the workflow associated with a radiology study.

system has its purpose in the overall electronic health system. The HIS is generally a system used to capture information about a patient, including but not limited to, their demographics, address, and insurance information (very important). The HIS captures charge information, alerts users when communications with a third-party payer is necessary, and generates bills. Frequently, the HIS is the centralized location for ordering various studies including radiology, lab, speech pathology, physical therapy, and so on. In addition, the HIS can be the center for reporting hospital issues, such as maintenance needs, network failures, and so on. The HIS usually provides extensive administrative reporting tools. The EMR is a patient-centric system containing the electronic record of all a patient's encounters at the institution. The HIS can be used to generate numbers of radiology orders generated in a certain time frame, but when a physician needs to see the total record for his patient with all associated orders, reports, and images to form a diagnosis and treatment plan, the EMR is a better model. The EMR will be discussed with more detail later in the article. The main purpose of PACS is to manage radiology images efficiently. These very large datasets have special needs when it comes to archiving, display, and networks and it makes sense to keep the PACS functionality a little apart from the rest of the usually text-based information systems. Speech recognition is generally developed for specific specialties, such as radiology. The radiology lexicon known by the recognition engine is unique to radiology. Of course, parallel systems exist in other specialties, but the focus here is on radiology. In general the radiologist dictates a report; the speech recognition transforms the voice file into a text file and displays it for confirmation of correctness or for editing.

RIS BASIC FUNCTIONALITY

Most RIS vendors offer a core feature set that captures the information items regarding a patient study in radiology and provides tools for those involved to manage the study. Table 1 is based on the list generated by Aunt Minnie (3) in the section on Radiology Information Systems in their buyer's guide. These functions are probably the minimum required for an RIS purchase. Anyone seeking to purchase an RIS should be aware of the resources available to help them with their decision and should follow a structured Request for Proposal (RFP) or bid format. Buyers should specify how these functions should work in their institution. For example, if the feature "interaction checking between exams" is included, the buyer should write a requirement that matches their workflow. An example of part of the requirement might be, "automatically check patient history for previous exam and warn user at the time of order". If you go back to Fig. 1, it is quite common for the physician or their agent to enter the order on the HIS, then this is transmitted electronically to the RIS, so the RIS would need to alert the HIS, which in turn would alert the physician or their agent. A requirement specification document tells vendors what the buyer expects and forces them to respond to the buyer with respect to their specific workflow. Buyers should be sure to include any special requirements for the institution. Some examples include the length and format of the medical record number or accession number or the requirement that it must be possible to enter a report directly on the RIS in the case of an HIS downtime. The buyer may want to specify their requirements for performance; "a query for a patient record should return results in < 2 s".

Table 1. Core Feature Set for the Typical Current RIS

Function	Comment
Patient registration	This may be performed at an HIS level and transferred to the RIS
Patient tracking	This allows a user to track the patient through the procedure (started, ended)
Order entry	This may be through the HIS, but in the case of HIS downtime, users must be able to enter an order
Merge or reconcile patient information	This is important after a downtime when temporary Medical Record Number (MRN) or accession numbers may have been used or in a trauma situation where a temporary name was used.
Interaction checking between exams	Does a previous exam interfere with the one being ordered?
Single exam code for combined orders	Can you combine Chest/Abd/Pelvis on one exam code with individual accession numbers?
Generate future orders	
Generate recurring orders	Can this also remind the ordering physician that the patient is receiving recurring orders, for example, recurring portable chest exams?
Document imaging	Can you scan paper and include it in the record?
Alerts for pregnancy, diabetes, allergy, and so on	This information will probably be sent from the HIS.
Attaches Prep information to ordered procedures	
Generates an online worklist for technologists	
Supports DICOM modality worklist	This could be a native feature in the RIS or an interface to PACS.
Charge capture	Including examination, supply items
Links CPT and ICD-9 codes to an examination	
Supports distribution of reports	This may be an interface to an HIS or an EHR.
Generates requisitions, labels, and so on	For those still using paper
HIPAA audit capabilities	
Graphic User Interface	

Although an institution may have an HIS and/or an EMR, the RIS will probably also keep an archive of diagnostic reports. From an information sciences storage perspective, a single data repository for specific information is more desirable that multiple, different repositories that have to be synchronized and managed. Since the RIS is rarely the system that is used by the ordering or referring physicians, it serves as an additional archive of diagnostic reports and can be used to track trends, search for diagnoses and impressions, and is used as a backup should other archives fail.

Table 2 contains a (noncomprehensive) list of advanced features for an RIS. Many of these features have become more important since hospitals have added PACS and required integration of all their systems. Bar code support may not be as important in a paperless world, but most of us are not there yet. Bar codes offer a quick and painless way to choose accurate information from a database. One or two bar code entries can locate the right information without the frustration most people experience trying to enter a long string of numbers and letters.

More institutions have merged into one larger entity or have splintered out clinics and imaging centers to locations with easier access. Since it is frequently the case that patients can be seen in various locations in a healthcare system, the RIS needs to support multifacility scheduling, as well as patient tracking. If a group of institutions form an enterprise and each institution can create individual medical record numbers (MRN) identifying patients, it is possible that more than one person will have the same medical record number: from different institutions. The RIS, as well as the EMR and HIS and PACS, must be able to differentiate individuals and track them throughout all the institutions in an enterprise. This can be a difficult and frustrating problem and should be managed carefully. All the interfaces need to be specified in detail. A more comprehensive discussion on the required interfaces is included later in this article. Electronic signatures and addenda need to match the workflow for an institution. In a large teaching hospital, it is common for a resident to dictate a report and for a faculty member to approve and verify it. At this point, two signatures are required. Then, if an addendum is entered on the report, which could happen when comparison studies are received from an outside source, a different resident–faculty member combination could dictate it, resulting in up to four signatures on a report. If multiple addenda are allowed, multiple signatures must also be allowed.

It is important for report distribution, and especially in the case where critical results have been found, that referring physician information is stored somewhere. When an RIS is in a stand-alone installation or the RIS is the distribution entity for reports, then each ordering physician and ordering service must have a unique profile and protocol rules for the communication of reports and especially for critical results. When the RIS is part of a larger enterprise system, the HIS or EMR will distribute the reports, but some triggering mechanism must alert the physician if a critical result or unusual finding is present. This will usually be a function of the RIS. The system must keep an audit of the successful communication with the ordering physician or service as part of the Joint Commission on the Accreditation of Health Care Organizations (JCAHO) recommendations for taking specific steps to improve communications among caregivers (4). This organization requires that each accredited institution have a method for rapid communication of results for both critical results and critical tests in place. In addition, each institution must have a way to monitor and report the efficiency and effectiveness of the communication for a subset of the results considered critical. For radiology, this will most likely be a function of the RIS.

The RIS should be able to generate administrative reports on the numbers of studies performed by date, the modalities used for studies, performance figures for technologists, costs of examinations, and reimbursement trends. In addition, the system should have a query interface so custom reports can be generated. It is very common for this query interface to use the Structured Query Language (SQL) that may have a steep learning curve. Some

Table 2. Advanced Feature Set for the Typical Current RIS

RIS Options	Comments
Bar code support	For quickly entering data in the environment where paper is still used
Supports multifacility scheduling	Can a single person be tracked through multifacility visits?
Interface to PACS	This needs careful specification and configuration
PACS included	
Appointment confirmation / reminders	
Information about referring physicians available	This may be better in the HIS, but will require close coordination
Technologist comments stored	Are they available in the speech recognition system or in PACS? How will the radiologist see them?
Electronic signature	
Proxy signature	
Dual signature	
Different signature on an addendum	
Multiple addenda	
Web based user interface	
Integration with speech recognition	
Interface to HIS	
DICOM modality worklist	This may be provided by the RIS instead of the broker in PACS

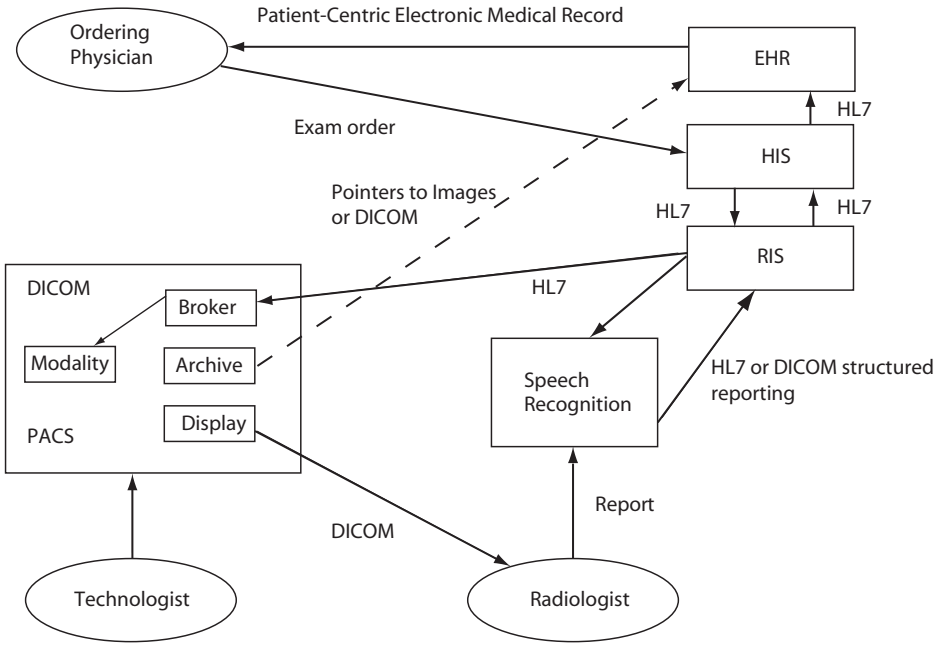


Figure 2. Current standards used as interfaces among information systems associated with a radiology study.

vendors offer a graphical user interface for custom queries that may make queries and reports easier to generate. Many of the RIS systems on the market today have a web-based interface that seems to be intuitive to the current internet-literate generation.

It is increasingly common to find RIS vendors storing images and providing PACS services, such as study display. For some users, this may be a good solution to an electronic radiology practice, for others, the RIS vendors may not have the sophisticated tools for image manipulation that PACS vendors have traditionally supplied. In addition, to complicate matters, PACS vendors are now offering more RIS functionality and once again the borders between the two systems are becoming blurred. It will be up to the institution to decide whether to use one of these hybrid systems or to interface two dedicated systems.

INTERFACING SYSTEMS

In radiology, there are two main interface standards, Health Level 7 (HL7) (5) and Digital Imaging Communications in Medicine (DICOM) (6). The HL7 is a formatted text-based standard that typically specifies the content of messages that are communicated among information systems in a healthcare institution, such as admission information, radiology results, and operating room notes. An RIS typically contains information about radiology schedules, orders, billing and reports, mostly text values. A complete EMR needs reports and images from radiology. Reports can be sent using HL7, but there is no way to encode image data in the current versions of HL7. Because of this limitation, DICOM was developed as the standard to handle transfer of images between systems that acquire, store, and display them.

Figure 2 illustrates the usual standards used for communication in the electronic medical practice first introduced in Fig. 1. Following the information flow, the attending physician places their order for a radiology study on the HIS, which is sent to the RIS via HL7. The order information is sent to PACS and the study is scheduled. The HL7 information is converted to DICOM in PACS using a translation program or broker and the patient demographics are attached to a study produced by PACS acquisition units (CT scanners, MRIs, computed radiology units, etc.). The radiologist using the DICOM viewer views the images and a report is generated using speech recognition. Note that the speech recognition system also receives the ordering information from the RIS, so at this point, there should be a single pointer or index to match the images with the report. The output from speech recognition could be in an HL7 format or a DICOM structured report. The report is usually sent to the RIS and then to the EMR.

The HL7 offers a rich data definition and is widely used to communicate information about the status of the patient in the form of an Admission Discharge Transfer (ADT) packet that informs all relevant systems about the location and identity of a patient, an order packet (ORU), and a report packet (ORP). There are of course many other types of packets that are part of normal transactions of an institution, but these are the ones most often used in the transfer of information between the RIS, PACS, and the speech recognition system. An example of an HL7 message is shown in Table 3. The field entries have been created for this example and this does not reflect an actual patient, referring physician, or radiologist. However, if this were an actual patient, the medical record number would be 0000123456, and the patient's name would Robert Richards, who was born on September 1, 1991 (19910901). The patient's ordering and attending physician is Forest Wood whose phone number is listed in the order. The accession number for this patient is 4445555 and the examination ordered is a bone age radiograph. The reason for the study is included in the text. For this institution, the procedure number for the routine bone age is 3000 and the location where the study will

Table 3. Example HL7 Simulation Showing a Typical Radiology Order

PID||100000123456|**000000123456**|000004567890|**RICHARDS**∧**ROBERT**
||**19910901**|**M**||B|555 NW 2ND ST
PV1|1|O|XRY∧|C|||09999∧WOOD∧FOREST∧(352)555-
1212|**09999**∧**WOOD**∧**FOREST**∧(**352**)**555-1212**||UF||||XRY,
ORC|NW|444444∧0|00003-
001|AUXR3000|SC||1∧∧∧200309151019∧∧R||200309151022|||09999∧WOO
D∧FOREST∧(352)555-1212||||∧||OBR|1|**4445555**∧0|00003-
001|UXR∧3000∧∧BONE
AGE|ROUTINE|200309151022|200309151022|||MIL|||**ORDERED**|2003091
51022||09999∧WOOD∧FOREST∧(352)555-1212|(352)555-
1212||ORT|UEXT||||UXR||∧∧∧∧∧∧∧|1∧∧∧200309151022∧∧R||||BONE AGE
RT WRIST PAIN S/P LT∧DISTAL TIBIA HEMI EPIPHYSIODES IS H/O
MULTIPLE HEREDITARY EX∧STOSES AND LIMB LENGTH
DISCRE∧DISTAL TIBIA HEMI EPIPHYSIODES IS H/O MULTIPLE
HEREDITARY EXO∧STOSES AND LIMB LENGTH
DISCRE|∧∧∧||||200309151022

be performed is AUXR. When two or more vertical lines appear in the message, they represent fields that contain no data. Each field of an HL7 message is carefully specified and each system using HL7 understands that the PID–3 field will contain a medical record number. A report message in HL7 also contains specified fields that include many of those in this order packet as well as the diagnostic report, the reporting radiologist, and the verifying or signing radiologist. The accession number, 4445555 will be the key for connecting the order to the report to the images, and also to identify that this study was performed on Robert Richards. Much more information on HL7 and the standards committee may be found on the HL7 home page (5) and on numerous other web sites.

Since HL7 had no current capacity for images, the American College of Radiology (ACR) and the National Electrical Manufacturer's Association (NEMA) developed a standard for image communication among PACS devices. The earliest incarnations of this standard were hardware based with point-to-point connections between a modality and a computer. Originally named the ACR–NEMA standard, it has evolved to become DICOM and is still evolving to meet the ever-changing demands of radiology. The standard not only specified the content of the messages passed, it specified the communication pathways for carrying these messages. A full description of the DICOM standard is beyond the scope of this article. More information may be found on the DICOM home page (6), where DICOM specifies the transfer of images among devices, printing to film and paper, and the creation of a modality worklist, among many other things. Radiology modalities typically do not have HL7 interfaces, they are strictly DICOM enabled computers, and in order for them to have knowledge of an order placed by the RIS, the HL7 order needs to be translated into DICOM. For historical reasons, this translation device or software is often called a broker. The broker receives the HL7 and builds a DICOM modality worklist that may be queried by modalities, such as computed tormography (CT) or magnetic resonance imaging (MRI) units. The modality then compiles a list of studies that have been ordered and the technologist may pick a study from the list as they are setting up the console for the examination. This alleviates the need to reenter information about the patient and assures accurate and complete information that can be tracked back to the original order. The all-important accession number is thus associated with the study being performed. Table 4 shows a partial DICOM message attached to the image produced for the study ordered previously. This message is a series of groups, elements, element sizes, and element contents that are again very specific. The accession number is always located in group 8, element 50; the patient's name is always in group 10, element 10; and the patient's medical record number is always in group 10, element 20. The groups contain common elements. Group 8, for example, is information about the examination while group 10 contains information about the patient.

When the diagnostic report is generated on the speech recognition system, the accession number is attached and when the study is completed, the image (Fig. 3) and report (Table 5) are matched correctly. An archive query returns the basic information for the study so a correct selection can be made (Fig. 4).

The basis for PACS and speech recognition working together correctly is the RIS acting as an order entry and management system along with the HL7 and DICOM standards. Without the RIS and the standards, there would be no accurate way to capture all the patient information and to attach the report to the image. The PACS images should never be available without accurate and complete information (7). Every PACS must have a way to link pertinent patient information and the diagnostic report with the image. Although the example presented was fairly simple and straightforward with only one image, consider the case where a patient in an intensive care unit has multiple chest radiographs every day. Without a way to associate a report accurately with an image, an ordering physician could read a report for the wrong time. With accurate and complete information everyone can be assured that the report is attached to the correct study.

OPTIONS FOR THE INTEGRATION OF SYSTEMS

Although the standards are crucial to the success of an integration project, there are usually still issues among

Table 4. Example Portion of a DICOM Message Corresponding to the HL7 Demonstrated in Table 3[a]

0008 0020 DA	8	Study Date	20030915
0008 0021 DA	8	Series Date	20030915
0008 0022 DA	8	Acquisition Date	20030915
0008 0030 TM	6	Study Time	111002
0008 0031 TM	6	Series Time	111002
0008 0032 TM	6	Acquisition Time	111002
0008 0050 SH	8	Accession Number	4445555
0008 0060 CS	2	Modality	CR
0008 0070 LO	4	Manufacturer	AGFA
0008 0080 LO	22	Institution Name	Shands Hospital at UF
0008 0090 PN	12	Referring Physician's Name	WOOD∧FOREST∧
0008 1010 SH	10	Station Name	ADCPLUS03
0008 1030 LO	8	Study Description	BONE AGE
0008 103E LO	8	Series Description	hand PA
0008 1040 LO	26	Institutional Department Name	Shands at UF / Orthopedics
0008 1090 LO	8	Manufacturer's Model Name	ADC_5146
0010 0000 UL	4	Group 0010 Length	70
0010 0010 PN	16	Patient's Name	RICHARDS∧ROBERT
0010 0020 LO	8	Patient ID	00123456
0010 0030 DA	8	Patient's Birth Date	19910901
0010 0040 CS	2	Patient's Sex	M
0018 0000 UL	4	Group 0018 Length	198
0018 0015 CS	4	Body Part Examined	HAND
0018 1000 LO	4	Device Serial Number	1581
0018 1004 LO	4	Plate ID	02
0018 1020 LO	8	Software Versions(s)	VIPS1110
0018 1164 DS	30	Imager Pixel Spacing	1.00000000E-01\1.00000000E-01
0018 1260 SH	6	Plate Type	code 15
0018 1401 LO	12	Acquisition Device Processing	10101Ia713Ra
0018 1402 CS	8	Cassette Orientation	PORTRAIT
0018 1403 CS	8	Cassette Size	8INX10IN

[a]Note the inclusion of the accession number.

vendors. The DICOM can be interpreted in different ways: HL7 fields may be required by one vendor and ignored by another, and problems will arise. Buyers have several options to consider. The easiest way to assure that integration will be successful is to purchase all information systems from the same vendor. Of course, that will not completely assure success because not all vendors can supply all the required systems. Most noticeably, they may not all supply the imaging modalities needed. As soon as another vendor's modality is introduced, an integration project is needed.

Another decision that must be made is the actual storage of the images from PACS and the diagnostic reports associated with these images. The RIS or EMR can supply the storage for both, or the PACS can supply the storage for both, or some combination of storage options can be designed. Neither the RIS nor the PACS would be the preferred method for storing and delivering studies and reports to referring and ordering physicians because the information is all radiography-centric and it is far more desirable to see a total picture of the patient with laboratory results, history and physical notes, nursing documentation, and all other information regarding the health record for a patient.

The Health Insurance Portability and Accountability Act (HIPAA), a landmark law that was passed in 1996, specifies mandates in the transactions between healthcare companies, providers, and carriers (8). This act was originally designed to make the healthcare records for a patient available to healthcare institutions and physicians and to insurance companies using standards. Individuals should be able to give a physician permission to access their healthcare record at any location, quickly and efficiently. The law also protects the privacy of the patient and provides security guidelines to assure the information could not be accessed inappropriately. Only an EMR system build on well-accepted standards can meet these requirements (9–12). The radiology information should be either stored in an EMR or the EMR should have pointers to the information and should be able to communicate that information in a standard format.

The Radiological Society of North America (RSNA), the Healthcare Information and Management Systems Society (HIMSS), and the American College of Cardiology (ACC) are working together to coordinate the use of established standards, such as DICOM and HL7, to improve the way computer systems in healthcare share information. The initiative is called "Integrating the Healthcare Enterprise" or IHE (13). Integrating the Healthcare Enterprises promotes the coordinated use of established standards, such as DICOM and HL7, to address specific clinical needs in support of patient care. Healthcare providers envision a day when vital information can be passed seamlessly from system to system within and across departments (10,11,14). In addition, with IHE, the EMR

Figure 3. An image associated with the simulated order shown in Table 3. Note the inclusion of the accession number on this image that is required to associate the image with the report.

will be a standard and will facilitate information communication among healthcare venues. Recent research suggests that the United States could realize a savings potential of $78 billion annually if a seamless, fully interoperable healthcare information exchange could be established among key stakeholders in the healthcare delivery system (9).

SELECTING AN RIS

The KLAS company, founded in 1996, is a research and consulting firm specializing in monitoring and reporting the performance of Healthcare's Information Technology's (HIT) vendors. The comprehensive reports they produced are valuable for comparing the vendors they review. Buyers should be aware that not all vendors are included in the reports, but the major ones are represented. The Comprehensive Radiology Information Systems Report, Serving Large, Community, and Ambulatory Facilities was released in January, 2005 by KLAS (15). In the report, eight RIS vendor products were represented and were reviewed by interviewing users of the systems. In an addendum, seven other vendors were presented whose products did not yet meet the KLAS standards for statistical confidence in order to be compared with other vendors in the main body of the report. Vendors were all allowed to prepare overviews and their perceptions of their products. Of course, these overviews stress the strengths of a vendor's product. Performance measurements of the KLAS traditional 40 indicators (Table 6), technology overviews, client win/loss and pricing provide the bases of the provider experience. The KLAS report should be part of an institution's decision-making process when a report is available, and in this case the report is available and very timely.

This report focused on large ($>$ 200 bed) and ambulatory (free standing clinics and imaging centers) facilities. The large facilities were asked additional questions. The larger institutions were asked why a vendor was selected and why a vendor was NOT selected based on the following six criteria: Functionality, Cost, Relationship with Vendor, References/Site Visits/Technology, and Integration/Interfacing. The most cited reason as to why a vendor was selected was their ability to integrate or interface. The most cited reason as to why a vendor was not selected was their perceived limited functionality.

Survey participants from the large institutions were asked questions regarding their RIS and its; (*1*) benefits; (*2*) functional strength of the reporting module; and (*3*) the RIS–PACS integration. The top benefit, reported by $>$ 50% of the survey respondents, was that of More Efficient/Better Workflow. The number two and three benefits identified were Interface–Integration and Manage Department Better. The functional strength of the reporting module on a scale of 1–5 (1 + weak and 5 = strong) was rated with an average of 3.5, with the highest score of 4, which indicates that there are a lot of users who are not totally satisfied with their reporting module. Forty-seven percent of the survey respondents indicated that they have plans to move to an integrated RIS/PACS solution and 81% of these reported that Radiologists and Clinicians were mostly driving this integration. This indicates that there are a large number of institutions without an integrated solution.

THE RIS OF THE FUTURE

In the future, it is likely that speech recognition systems will be incorporated into an RIS solution. Throughout this article, the two systems have been shown as separate entities and indeed, that is the most common incarnation at this time. Speech recognition has become a more important part of the radiology workflow as researchers demonstrate the potential for improving report turnaround time and decreasing costs when compared to systems with manual transcriptionists (16–19). Unfortunately, the increase in report turnaround time does not guarantee efficient personnel utilization. Although the report can be available for the physician immediately after the radiologist verified the dictation, the responsibility for editing the report falls on the radiologist and may make the time required for the process longer than with a transcriptionist performing the typing and editing. Because of this, speech recognition is not eagerly embraced by many radiologists, and therefore

Table 5. Example Portion of a Radiology Report for the Simulated order in Table 3[a]

SHANDS DIAGNOSTIC	EXAM: BONE AGE
NAME: RICHARDS, ROBERT	
EXAM DATE: 09/15/2003	LOC: XRY- SEX: M
MRN: 00123456 DOB: 09/01/1991	
ORDERING MD: WOOD , FOREST	
ORD. SERVICE: ORT	ORD. LOC: ORT TECH:JJJ
REASON: BONE AGE RT WRIST PAIN S/P LT	
EXAM DATE: 09/15/2003 11:18	
ACCESSION#: 44455555	
UXR 3000 BONE AGE	
ICD9 CODES: 719.43380.8	
FINDINGS:	
Clinical History: 12 year-old male with multiple hereditary exostoses, limb length discrepancies, and left wrist pain.	
Findings: The bone is within 2 standard deviations of the chronological age. There are multiple exostoses in the right hand.	
IMPRESSION:	
1. Normal bone age.	
2. Findings consistent with multiple hereditary exostoses.	
Dictated By: MARK FISCHER, MD	
Dictated Date: SEPT 15, 2003 1:54P	
Interpreted By:	
WILLIAM JENNINGS, MD	MARK FISCHER, MD
This study was personally reviewed by me and I agree with the report.	Resident Radiologist

[a]Note the accession number which associates the report with the study images.

acceptance has been slower than with other information systems used in radiology.

As speech recognition gains more widespread acceptance, researchers are looking into improving the diagnostic reporting methodology by structuring the format of the report. The DICOM Structured Report is a definition for the transmission and storage of clinical documents. A DICOM Structured Report can contain the traditional text report in addition to structured information and links to key images. If you return to Fig. 2, you will notice that the RIS and HIS typically communicate using HL7, not DICOM, so either HL7 will have to change to support these reports or the RIS will have to support DICOM. There is an evolving standard for the HL7 Clinical Document Architecture (CDA) that supports text, images, sounds, and other multimedia content. The DICOM working group 20 (the Integration of Imaging and Information systems Group) and the HL7

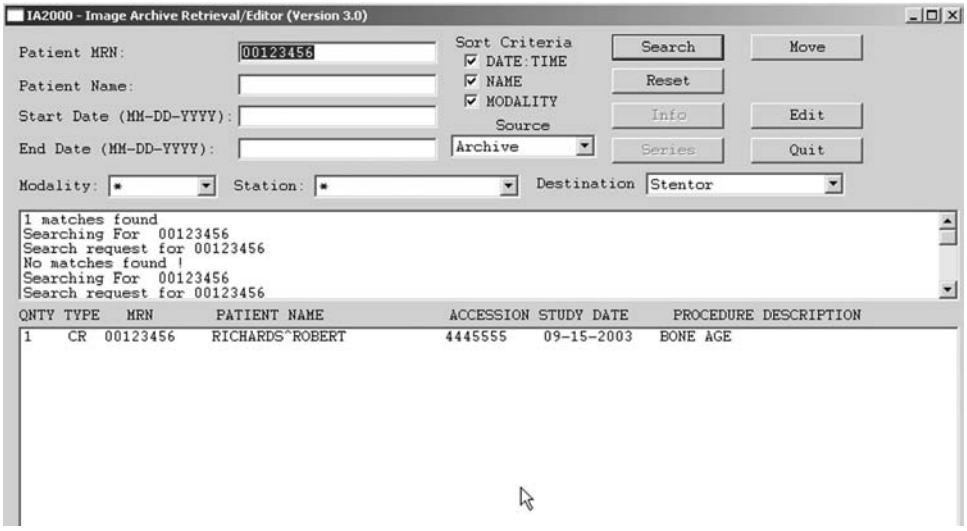


Figure 4. An archive query associated with the simulated order shown in Table 3.

Table 6. The 40 Success Indicators Used by KLAS in Their Information Systems Reports

10 Product/Technology Indicators[a]
Enterprise Commitment to Technology
Product Works as Promoted
Product Quality Rating
Quality of Releases and Updates
Quality of Interface Services
Interfaces Met Deadlines
Quality of Custom Work
Technology Easy to Implement and Support
Response Times
Third-Party Product Works with Vendor Product
10 Service Indicators[a]
Proactive Service
Real Problem Resolution
Quality of Training
Quality of Implementation
Implementation on Time
Implementation within Budget/Cost
Quality of Implementation Staff
Quality of Documentation
Quality of Telephone/Web Support
Product Errors Corrected Quickly
8 Success Indicators[a]
Worth the Effort
Lived Up to Expectations
Vendor is Improving
Money's Worth□□
Vendor Executives Interested in You
Good Job Selling
Contracting Experience
Helps Your Job Performance
12 Business Indicators[b]
Implemented in the Last 3 Years)
Core Part of IS Plan
Would You Buy It Again
Avoids Nickel-and-Diming
Keeps All Promises
A Fair Contract
Contract is Complete (No Omissions
Timely Enhancement Releases
Support Costs as Expected
Would You Recommended to a Friend/Peer
Ranked Client's Best Vendor
Ranked Client's Best or Second Best Vendor

[a]Rating 1–9, where 1=poor and 9=excellent.
[b]Rating is a yes or no.

Imaging Integration Special Interest Group (IISIG) are working on harmonizing the existing standards (20,21). Future RIS implementations will most certainly support the DICOM structured report or its HL7 CDA translation.

In addition to supporting the traditional terminal or remote computer, the RIS of the future will be required to support a web interface as well as supporting handheld devices. A radiologist with a handheld Personal Digital Assistant (PDA) can currently access e-mail, the internet, digital media, and documents. With wireless networking available in many hospitals, these devices can support the exchange of information between ordering physician and radiologist including results, they can support exchange of information between radiologist and technologist including protocol selection, they can access history and lab reports for the patient, they can provide a worklist of current studies, and can even display low resolution images. The ability of the PDA to support these functions is only limited by the ability of the RIS to provide PDA support (22).

The RIS of the future may store PACS images, must be able to interface with the EMR, and should provide structured reporting and support for PDAs. As institutions plan for purchasing new systems or updating the old ones, future functionality of the systems must be considered and be part of the request for proposal or bid process. In the past, institutions were able to select an RIS based on individual preferences without consideration of interface issues. As the national health records structure evolves, all institutions must be in a position to interface to an EMR using well developed standards.

BIBLIOGRAPHY

1. Honeyman-Buck JC. PACS Adoption. Seminars in Roentgenology 2003; July, 38(3):256–269.
2. Thrall JH. Reinventing Radiology in the Digital Age. Part I. The All-Digital Department. Radiology 236(2):382–385.
3. Aunt Minnie Buyer's guide for Radiology Information Systems. Aunt Minnie Web Site. Available at http://www.auntminnie.com. Accessed 2005; Aug 2.
4. 2005 National Patient Safety Goals. JCAHO web site. Available at http://www.jcaho.org. Accessed 2005; Aug 11.
5. HL7. HL7 web site. Available at http://www.hl7.org. Accessed 2005; Aug 11.
6. DICOM. DICOM web site. Available at http://medical.nema.org. Accessed 2005; Aug 11.
7. Carrino JA. Digital Imaging Overview. Seminars Roentgenol 2003;38(3):200–215.
8. HIPAA. HIPAA web site. Available at http://www.hipaa.org. Accessed on 2005; Aug 20.
9. Middleton B, Hammond WE, Brennan PF, Cooper GF. Accelerating U.S. HER Adoption: How to Get there From Here, Recommendations Based on the 2004 ACMI Retreat. JAMIA 2005;12(1):13–19.
10. Makoul G, Curry RH, Tang PC. The Use of Electronic Medical Records: Communication Patterns in Outpatient Encounters. JAMIA 2001;8(6):610–615.
11. Stead WW, Kelly BJ, Kolodner RM. Achievable Steps Toward building a National Health Information Infrastructure in the United States. JAMIA 2005;12(2):113–120.
12. Berner ES, Detmer DE, Simborg D. Will the wave Finally break? A brief View of the Adoption of Electronic Medical Recoreds in the United States. JAMIA 2005;12(1):3–7.
13. IHE. The IHE web site. Available at http://www.ihe.net. Accessed on 2005; Aug 20.
14. Channin DS. Driving Market-driven Engineering. Radiology 2003;229:311–313.
15. Comprehensive Radiology Information Systems Report, KLAS Enterprises. Available at www.healthcomputing.com: Accessed 2005.
16. Langer SG. Impact of Tightly coupled PACS/Speech Recognition in Report Turnaround Time in the Radiology Department. J Digital Imaging 2002;15(Supp 1):234–236.
17. Gutierrez AJ, Mullins ME, Novelline RA. Impact of PACS and Voice-Recognition Reporting on the Education or Radiology Residents. J Digital Imaging 2005;18(2):100–108.

18. Zick RG, Olsen J. Voice Recognition software Versus a Traditional Transcription Service for Physician Charting in the ED. Am J Emerg Med 2001;19:295–298.
19. Bramson RT, Bramson RA. Overcoming Obstacles to Work-Changing Technology Such As PACS and Voice Recognition. AJR 2005;184:1727–1730.
20. Hussein R, Schroeter A, Meinzer H-P. DICOM Structured Reporting.RadioGraphics 2004;24(3):891–896.
21. Dolin RH, et al. The HL7 Clinical Document Architecture. JAMIA 2001;8:552–579.
22. Flanders AE, Wiggins RH, Gozum ME. Handheld Computers Radiol. RadioGraphics 2003;23:1035–1047.

See also EQUIPMENT ACQUISITION; PICTURE ARCHIVING AND COMMUNICATION SYSTEMS; RADIOTHERAPY TREATMENT PLANNING, OPTIMIZATION OF; TELERADIOLOGY.

RADIOLOGY, PHANTOM MATERIALS.

See PHANTOM MATERIALS IN RADIOLOGY.

RADIOMETRY.

See THERMOGRAPHY.

RADIONUCLIDE PRODUCTION AND RADIOACTIVE DECAY

SILVIA S. JURISSON
WILLIAM MILLER
J. DAVID ROBERTSON
University of Missouri
Columbia, Missouri

INTRODUCTION

Radiopharmaceuticals, drugs containing radioactive atoms, are used for diagnostic imaging or therapeutic applications in nuclear medicine, depending on their radioactive emissions. Penetrating radiations (gamma rays or annihilation photons from positron emission) are used for diagnostic applications with gamma cameras, single photon emission computed tomography (SPECT), or positron emission tomography (PET) instrumentation. Diagnostic imaging with gamma emitters became a mainstay of nuclear medicine with the advent of the molybdenum-99/technetium-99m (^{99}Mo/^{99m}Tc) generator (the Brookhaven generator), which was developed by Richards in $\sim$1961 at Brookhaven National Laboratory (1). This generator made the short half-lived ^{99m}Tc (6.01 h; 140 keV γ-ray) readily available to nuclear medicine departments around the world, and not just at the site of ^{99m}Tc production. Today, ^{99m}Tc accounts for >80% of the diagnostic scans performed in nuclear medicine departments in the United States, with a variety of U.S. Food and Drug Administration (FDA) approved agents for functional imaging of the heart, liver, gallbladder, kidneys, brain, and so on (2). Particle emitters, such as alpha, beta, and Auger electron emitters, are used for targeted radiotherapy since their decay energy is deposited over a very short range in tissue. Radiotherapeutic applications in humans began in the late 1930s with the beta emitters radioiodine (^{128}I, ^{131}I), radiophosphorus (^{32}P) and radiostrontium (^{89}Sr) for cancer treatment (3). Radioiodine (^{131}I iodide) was the first, and continues to be the only, true "magic bullet" through its specific and selective uptake in thyroid. It is used to treat hyperthyroidism and thyroid cancer.

Radionuclides for medical applications are produced either at nuclear reactors or accelerators. The "Availability of Radioactive Isotopes" was first announced from the headquarters of the Manhattan Project (Washington, DC) in Science in 1946 (4), and today medical isotope availability remains an important issue for the nuclear medicine community. The selection of radionuclides for use in radiopharmaceuticals is dependent on their decay properties (half-life, emissions, energies of emissions, dose rates) and their availability (production and cost).

Radionuclides suitable for diagnostic imaging are gamma emitters (with no or minimal accompanying particle emissions) such as ^{99m}Tc and positron emitters (annihilation photons) such as ^{18}F. The half-lives should be as short as possible and still allow preparation (synthesis and purification) of the radiopharmaceutical, administration of the agent to the patient, and the diagnostic imaging procedure. Typically, half-lives on the order of minutes to a week are used, with hours to a day considered optimal for most applications. Gamma energies in the range of 80–300 keV are considered good, although higher energy (e.g., ^{131}I at 364.5 keV) are used, and 100–200 keV is considered optimal. The PET instrumentation is designed for the two 511 keV annihilation photons.

The radionuclides used or under investigation for radiotherapeutic applications are particle emitters, with most of the efforts focusing on the beta emitters (e.g., ^{153}Sm, ^{90}Y, ^{186}Re, ^{188}Re, ^{177}Lu, ^{149}Pm, ^{166}Ho, ^{105}Rh, ^{199}Au) and some on alpha emitters (e.g., ^{212}Bi, ^{213}Bi, ^{225}Ac). In the case of radiotherapy, the half-life of the radionuclide should match the biological half-life for the radiopharmaceutical delivery to its *in vivo* target site (i.e., tumor), typically < 1 day – 1 week. The optimum particle and particle energy remains under investigation, and it is not clear that a higher particle energy translates into a more successful treatment (i.e., radiation dose to nontarget organs will limit the dose allowed for administration). Suitable accompanying gamma emissions will allow the *in vivo* tracking of the radiotherapeutic dose.

Two important factors for the use of radionuclides in nuclear medicine are thus radioactive decay and radionuclide production, both of which are discussed in detail below. Radionuclides used in nuclear medicine applications are used as examples in discussing these topics. Radioactive decay includes the types of decay (e.g., alpha, beta, electron capture, positron, gamma) with focus on those modes with nuclear medicine applications and rates of radioactive decay (e.g., units of radioactivity, exponential decay, activity–mass relationships, parent–daughter equilibria, medical generators).

TYPES OF RADIOACTIVE DECAY

A nuclide is considered to be radioactive if its nuclear configuration (number of neutrons and protons in the

nucleus) is unstable. Radioactive decay is then a means for the unstable nucleus to achieve a more stable nuclear configuration although it may or may not form a stable nuclide following a single decay. There are several modes of decay possible for a radioactive nuclide including alpha emission, beta emission, positron emission, electron capture decay and/or gamma emission. An alpha (α) particle is a helium nucleus ($^{4}_{2}He^{2+}$) and is generally a mode of decay available in heavy nuclei (i.e., $Z \geq 83$). Some lanthanides can show α-decay and heavy nuclides can also undergo decay by β^-, EC, and spontaneous fission. The following equation gives an example of alpha decay:

$$^{213}_{83}Bi \xrightarrow{45.59\,min} {}^{209}_{81}Tl + {}^{4}_{2}He^{2+} + \gamma + Q$$

where Q is the energy released during the decay process and γ is the 727 keV gamma photon emitted in 11.8% of the decays. Most of the decay energy in this process is in the form of the kinetic energy of the α-particle, which is rapidly deposited in matter (e.g., tissue) because of its relatively high charge and mass.

Beta (β^-) decay occurs when the nucleus has a proton/neutron ratio that is too low relative to the proton/neutron ratio in the stable nuclei of that element, and during this process a neutron is converted into a proton resulting in a nuclide that is one atomic number (Z) higher than its parent. An example of beta decay is shown below,

$$^{188}_{75}Re \xrightarrow{16.98\,h} {}^{188}_{76}Os + {}^{0}_{-1}\beta + \tilde{\nu}_e + \gamma + Q$$

where Q is the energy released, $\tilde{\nu}_e$ is an antineutrino, and γ is the 155 keV gamma photon emitted in 15% of the decays. These neutron-rich radionuclides are produced in nuclear reactors as will be discussed in the section Radionuclear Production.

When a radioactive nucleus has a proton/neutron ratio that is too high relative to that of the stable nuclei of the same element, two modes of decay are possible (positron emission and/or electron capture decay), with both converting a proton into a neutron resulting in a nuclide that is one atomic number lower than its parent. Electron capture decay arises from the overlap of nucleon orbitals with electron orbitals. It is a result of the "weak force", which is very short ranged. During electron capture decay (ε), essentially an inner-shell electron (usually K-shell) is incorporated into the nucleus converting a proton into a neutron. This process creates an orbital vacancy and results in a cascade of outer-shell electrons filling the lower energy inner-shell vacancies, with the excess energy released as X-rays and/or Auger electrons. The following equation shows the electron capture process:

$$^{201}_{81}Tl + {}^{0}_{-1}e^- \xrightarrow{73.1\,h} {}^{201}_{80}Hg + \nu_e + Q + \text{Auger}\,e^-/\text{X rays}$$

where Q is the energy released and ν_e is a neutrino. The 80 keV X rays emitted are used for myocardial imaging in stress–rest tests performed in nuclear medicine departments. Positron emission is only possible when Q is > 1.022 MeV, the energy equivalent of the mass of two electrons. During positron decay, a proton is converted into a neutron with simultaneous emission of a positron (β^+), which is a positive electron. Since a neutron has greater mass than a proton, the energy equivalent of two electrons (one to convert a proton to a neutron in the nucleus, and one emitted as a positron) must be available for this decay mode to be possible. Practically, positron emission does not occur unless Q is ≥ 2 MeV. An example of the positron decay process is shown below.

$$^{18}_{9}F \xrightarrow{110\,min} {}^{18}_{8}O + {}^{0}_{1}\beta^+ + \nu_e + Q$$
$${}^{0}_{1}e^+ + {}^{0}_{-1}e^- \longrightarrow 2\text{–}511\,\text{keV annihilation photons}$$

where Q is the decay energy (including the maximum positron energy plus 1.022 MeV), ν_e is a neutrino, and the two 0.511 MeV photons result from positron annihilation (often called annihilation photons) and are emitted 180° opposite each other. They are the basis of PET imaging. The ^{18}F radiolabeled fluorodeoxyglucose (^{18}F-FDG) is used in nuclear medicine departments for imaging glucose metabolism, such as found in growing tumors. In some cases, both electron capture and positron emission can occur. These proton-rich radionuclides are produced by accelerators which is discussed below.

Gamma (γ) emission can accompany any other decay process (i.e., alpha, beta, electron capture, positron decay) or it can occur without any particle emission. In the latter case, it is called isomeric transition (IT) and occurs when a metastable radioisotope [higher energy excited state of a nucleus with a measurable lifetime ($\geq$ns)] decays to the ground state (lower energy state of the nucleus). Isomeric transition is accompanied by energy release without any other change occurring in the nucleus (i.e., the number of protons and neutrons in the nucleus does not change during IT). The decay of ^{99m}Tc is an example of an isomeric transition.

$$^{99m}_{43}Tc \xrightarrow{6.01\,h} {}^{99}_{43}Tc + \gamma + Q$$

where Q is the decay energy and γ is the 140 keV photon released from the nucleus. The ^{99m}Tc in various chemical formulations is used routinely for diagnostic imaging of a variety of diseases and/or organ functions.

Whenever a gamma photon is released from the nucleus, the emission of conversion electrons is also possible. A conversion electron is the emission of an electron that has the energy of the gamma photon minus the electronic atomic binding energy, and it is emitted in place of the gamma photon. The probability of conversion electron emission rather than gamma emission increases as the energy of the gamma photon decreases and it increases with increasing nuclear charge. For example, the 140 keV gamma photon of ^{99m}Tc is emitted with 89% abundance; conversion electrons, rather than gamma photons, are emitted 11% of the time when ^{99m}Tc decays to ^{99}Tc.

There are radionuclides in which more than one type of radioactive decay occurs. An example is the decay of ^{64}Cu, which undergoes beta decay (39% of the time), positron emission decay (19% of the time), and electron capture decay (42% of the time) with a weighted average half-life of 12.7 h (the specific half-lives are shown below). This radionuclide has both diagnostic imaging (β^+ emission via the

annihilation photons) and radiotherapy ($\beta^{-/+}$ and Auger electrons) applications.

$$^{64}_{29}\mathrm{Cu} \xrightarrow{33.4\,\mathrm{h}} {}^{64}_{30}\mathrm{Zn} + {}^{0}_{-1}\beta^{-} + \overline{\nu}_{\mathrm{e}} + \gamma + Q$$

$$^{64}_{29}\mathrm{Cu} \xrightarrow{66.8\,\mathrm{h}} {}^{64}_{28}\mathrm{Ni} + {}^{0}_{+1}\beta^{+} + \nu_{\mathrm{e}} + Q + 2 - 0.511\,\mathrm{MeV\ photons}$$

$$^{64}_{29}\mathrm{Cu} + \mathrm{e}^{-} \xrightarrow{30.2\,\mathrm{h}} {}^{64}_{28}\mathrm{Ni} + \nu_{\mathrm{e}} + Q + \mathrm{Auger\ e}^{-}/\mathrm{X\ rays}$$

All of the above modes of radioactive decay have been utilized or are under investigation for utilization in the development of radiopharmaceuticals for nuclear medicine applications (either diagnostic imaging or radiotherapy).

RATES OF RADIOACTIVE DECAY

Radioactive decay is a random process where the number of nuclei in an isotopically pure sample that decay during a given time period is proportional to the number of nuclei (N) present in the sample. If radioactivity is defined as a measure of the rate at which the radioactive nuclei disintegrate, then the number of disintegrations per unit time (dN/dt) from a sample is given by the following expression where λ is the decay constant (the probability of decay per unit time) and negative sign indicates the loss of the radionuclide through the disintegration. This results in a radioactive decay process that follows a first-order rate law:

$$\frac{dN}{dt} = -\lambda N$$

From this definition, the number of disintegrations per unit time (dN/dt) is also known as the activity. For an isotopically pure sample, the number of nuclei N can be calculated knowing Avogadro's number (N_{A}), the mass number of the isotope (A) in daltons or grams per mole and the sample's weight (wt) in grams:

$$\mathrm{Activity} = \lambda N = \lambda \frac{N_{\mathrm{A}}}{A}\mathrm{wt}$$

As with all measurements, several systems of units can be used to define the amount of radioactivity in a particular sample. And, as with many measuring systems, there are both "traditional" units and new SI units (International System of Units). The original units are referenced to the curie (Ci), named in honor of Madam Curie. A curie is 3.7×10^{10} (37 billion) radioactive disintegrations per second, or approximately the amount of radioactivity in a gram of radium, one of the natural radioactive elements with which Madam Curie did her research. For many applications of radiation in medicine or industry, the curie is a relatively large quantity, and so the units of mCi (1/1000th) and microcurie (μCi) (1/1,000,000th) are utilized. On the other hand, a nuclear reactor contains millions of curies of radioactivity and units of kilocurie (kCi) (1000) and megacurie (MCi) (1,000,000) are sometimes used. Although the unit of the curie is being supplemented with the newer SI unit, it is still very much in common use.

The SI unit of radioactivity, the becquerel (Bq), is named after Henri Becquerel, the discoverer of radioactivity. It is defined as one disintegration per second. Thus it is much smaller than the Ci. Multiples of becquerel are the kilobecquerel (kBq) (1000), megabecquerels (MBq) (1,000,000), and so on. These units are summarized in Table 1.

Table 1. Units of Activity

Curies	Becquerels or disintegrations s^{-1}	Becquerels or disintegrations s^{-1}	Curies
1 MCi	3.7×10^{16}	1 Bq	2.7×10^{-11}
1 kCi	3.7×10^{13}	1 kBq	2.7×10^{-8}
1 Ci	3.7×10^{10}	1 MBq	2.7×10^{-5}
1 mCi	3.7×10^{7}	1 GBq	0.027
1 μCi	3.7×10^{4}		

The rate at which radioactive decay occurs can be defined as the half-life, or the amount of time that it takes for one-half of the radiation to decay. Unfortunately, two half-lives do not eliminate it (i.e., one-half decaying in one half-life and then the second one-half decaying in a second half-life), but rather the half-life is always referring to how much is left at any given point in time. Thus, one half-life reduces the radioactivity to one-half or 50%; two half-lives to one-half of one-half, one-fourth or 25%; three half-lives to one-eighth or 12.5%; and so on This leads to a kinetic model that is described by an exponential function (see Fig. 1), which is the solution to the previous equation:

$$N(t) = N_0 e^{-\lambda t}$$

The half-life is related to the physical decay constant (λ) by the simple expression:

$$t_{1/2} = \frac{0.693}{\lambda}$$

As a radionuclide decays, the number of nuclei (N) changes, and the amount of radioactivity present at any given time "t" remaining from an initial amount A_{o} (expressed in either Bq or Ci) can be given by any of the following:

$$A(t) = A_0 e^{-\lambda t}$$

$$A(t) = A_0 e^{-(0.693/t_{1/2})t}$$

$$A(t) = A_0 \left(\frac{1}{2}\right)^{t/t_{1/2}}$$

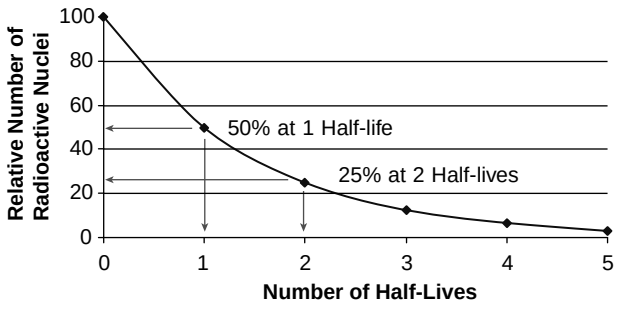


Figure 1. Relative number of radioactive nuclei remaining after various half-lives of decay.

Radioactive nuclei have half-lives that range from fractions of a second to billions of years. The radioactive material introduced into the body for medical purposes would typically have half-lives on the order of a few hours to a few days, so that it decays away to negligible levels in a relatively short amount of time. At the other extreme, some of the naturally radioactive nuclei in our environment have half-lives of billions of years and the reason they are still present is that they have not had enough time to decay to negligible levels since the earth was formed. An example of such a radionuclide is ^{40}K, which makes up 0.0117% of naturally occurring potassium and has a half-life of 1.27×10^9 years.

Another interesting case of radioactive decay involves parent–daughter relationships in which a radioactive parent decays to a radioactive daughter. This process is utilized extensively in the medical profession to provide a long-lived source for a short-lived radioisotope in what is known as a generator. The daughter can be obtained from the generator by "milking" it, taking advantage of the differences in chemistry between the parent and daughter elements to extract the daughter off of an ion column, which holds the parent.

The use of ^{99m}Tc, which has a 6 h half-life, is a case in point. Within a 24 h period it has decayed through four half-lives and is only one-sixteenth of its original value. Thus, ^{99m}Tc would have to be made in a nuclear facility and shipped daily to meet hospital needs. Fortunately, the daughter ^{99m}Tc is produced by the decay of a parent isotope ^{99}Mo, which has a 2.75 day half-life. A supply of ^{99m}Tc can thus be obtained over a period of ~1 week from the more slowly decaying ^{99}Mo.

$$^{99}_{42}\mathrm{Mo} \xrightarrow{2.75\,\mathrm{days}} {}^{99m}_{43}\mathrm{Tc} + \beta^- + \bar{\nu}_e \xrightarrow{6\,\mathrm{h}} {}^{99}_{43}\mathrm{Tc} + \gamma$$

The kinetics of the quantity of a daughter isotope (d) available from a parent isotope (p) can be readily solved using a first-order differential equation resulting in a straightforward algebraic expression:

$$\frac{A_{\mathrm{daughter}}}{A_{\mathrm{parent}}} = \frac{\lambda_d}{\lambda_d - \lambda_p}\left(1 - e^{-(\lambda_d - \lambda_p)t}\right)$$

For most practical cases where the parent is longer lived than the daughter, the daughter activity reaches a value close to the activity of the parent after approximately four half-lives. Thus, in an undisturbed generator that has been allowed to reach equilibrium, the quantity of daughter present (in Ci or Bq) is approximately equal to the activity of the parent at that time. Once the daughter has been extracted (or milked), it immediately begins building up again to a new equilibrium value as shown in Fig. 2.

Also of importance is the time it takes for the daughter to reach its maximum value, which is determined by the decay constants of the two isotopes involved:

$$t_{\max} = \frac{\ln\left(\frac{\lambda_d}{\lambda_p}\right)}{\lambda_d - \lambda_p}$$

Again assuming a typical case where the parent is longer lived than the daughter, this is largely determined by the half-life of the daughter. Thus, the daughter will reach ~ 50% of the equilibrium value in one half-live, 75% of the equilibrium value in two halvelives, and so on. The time between subsequent extractions of a daughter from a generator is set by this regeneration time. For the ^{99}Mo → ^{99m}Tc system, the regeneration time needed to reach the maximum amount of daughter ^{99m}Tc is very close to 24 h, a very convenient amount of time for routine hospital procedures and is just a fortuitous consequence of the half-lives of the two isotopes involved.

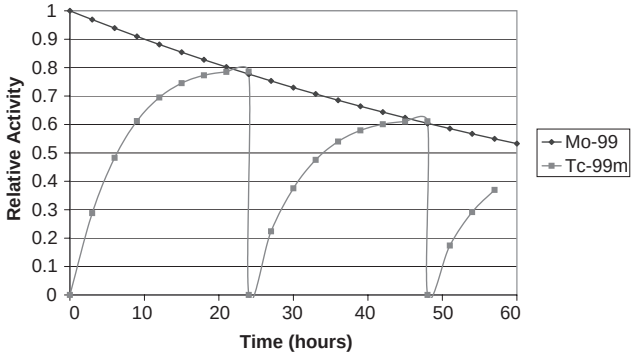


Figure 2. Shows build-up of ^{99m}Tc activity from ^{99}Mo initially and following elution.

Other examples of parent–daughter generators that are currently used or are under development for future use are ^{188}W/^{188}Re generators, ^{90}Sr/^{90}Y generators, ^{82}Sr/^{82}Rb direct infusion generators, ^{62}Zn/^{62}Cu generators, ^{224}Ra/^{212}Bi or ^{224}Ra/^{212}Pb generators, which also are a source of ^{212}Bi, the daughter product of ^{212}Pb, ^{225}Ac/^{213}Bi generators and ^{68}Ge/^{68}Ga generators. (For more information on generators see Ref. 5 and references cited therein.)

Specifically,

$$^{82}_{38}\mathrm{Sr} \xrightarrow{25.4\,\mathrm{days}} {}^{82}_{37}\mathrm{Rb} \xrightarrow{1.6\,\mathrm{min}} {}^{82}_{36}\mathrm{Kr} + \beta^+$$

$$^{68}_{32}\mathrm{Ge} \xrightarrow{271\,\mathrm{days}} {}^{68}_{31}\mathrm{Ga} \xrightarrow{68\,\mathrm{min}} {}^{68}_{30}\mathrm{Zn} + \beta^+$$

are two generators that can produce short lived PET isotopes, with ^{82}Rb being used for PET studies of myocardial function (rubidium acts as a potassium ion mimic).

The following three are examples of generators for beta or alpha emitting radioisotopes as possible therapeutic agents:

$$^{188}_{74}\mathrm{W} \xrightarrow{69.4\,\mathrm{days}} {}^{188}_{75}\mathrm{Re} \xrightarrow{17\,\mathrm{h}} {}^{188}_{76}\mathrm{Zn} + \beta^-$$

$$^{166}_{66}\mathrm{Dy} \xrightarrow{3.4\,\mathrm{days}} {}^{166}_{67}\mathrm{Ho} \xrightarrow{26.8\,\mathrm{h}} {}^{166}_{68}\mathrm{Er} + \beta^-$$

$$^{212}_{82}\mathrm{Pb} \xrightarrow{10.64\,\mathrm{h}} {}^{212}_{83}\mathrm{Bi} \xrightarrow{60.6\,\mathrm{min}} {}^{208}_{81}\mathrm{Tl} + \alpha$$

RADIONUCLIDE PRODUCTION

The radionuclides used in medicine and the life sciences are produced through neutron and charged particle

induced nuclear reactions. As with chemical reactions, the yield of the product (radioisotope) of interest will depend on the nuclear reaction employed, the energetics of the reaction, the probability of competing reaction pathways, and the ability to separate the desired product from the reactants (target) and any additional nuclear reaction products. For a complete overview of the practice and theory of nuclear reactions, the interested reader is referred to an introductory text on nuclear and radiochemistry (6–9).

Proton-rich radionuclides that decay by positron emission and/or electron capture are produced at cyclotrons and accelerators through charged particle induced nuclear reactions. For example, the production of the commonly used PET radionuclide ^{18}F through the bombardment of a target of ^{18}O-enriched water with high energy protons is produced by the following nuclear reaction:

$$^{1}_{1}\mathrm{p} + {}^{18}_{8}\mathrm{O} \rightarrow {}^{18}_{9}\mathrm{F} + {}^{1}_{0}\mathrm{n} + Q$$

that is written in a short-hand notation as ^{18}O(p,n)^{18}F. Note that, as in radioactive decay, charge (number of protons), mass number, and total energy are conserved in the nuclear reaction. The Q value for the reaction is the difference in energy (mass) between the reactants and products and is readily calculated from the measured mass excess (Δ) values for the reactants and products

$$Q = \sum \Delta_{\text{reactants}} - \sum \Delta_{\text{products}}$$

If Q is < 0, then the reaction is endoergic and energy must be supplied to the reaction (through the kinetic energy of the projectile) and if Q is >0, the reaction is exoergic and energy is released in the nuclear reaction. For the ^{18}O(p,n)^{18}F reaction,

$$\begin{aligned} Q &= \Delta^{1}_{1}\mathrm{p} + \Delta^{18}_{8}\mathrm{O} - (\Delta^{1}_{0}\mathrm{n} + \Delta^{18}_{9}\mathrm{F}) \\ &= -0.782 + 7.289 - (8.071 + 0.873) \\ &= -2.44 \text{ MeV} \end{aligned}$$

and the proton must supply 2.44 MeV of energy to the reaction. In practice, the actual proton bombarding energy is higher than the Q value because (1) not all of the kinetic energy of the proton is available for the nuclear reaction because of momentum conservation in the collision and (2) the probability of the reaction is typically very low at the threshold energy (Q). Typical conditions for the ^{18}O(p,n)^{18}F reaction on a water target with an in-house cyclotron are 16.5 MeV at 100 μA yielding 3–4 Ci/h. Even in those cases when the Q value for the reaction is >0, the incoming charged particle must have sufficient kinetic energy to overcome the coulomb or charge barrier between the projectile and the target nucleus and any angular momentum barrier that might exist for the reaction.

The probability that the projectile will strike the target nucleus and produce the radioisotope of interest is quantified with the reaction cross-section (σ) that has the dimensions of area. While quite sophisticated models exist for predicting reaction cross-sections based upon the underlying nuclear physics, the simple physical analogy for the cross-section is the area that the target nucleus presents to the incoming beam of projectiles. The SI unit for cross-section is m^2, but the more common unit is a barn (b); one barn is equal to 10^{-24} cm^2 (or 10^{-28} m^2). The magnitude of a barn can be understood from the fact that a target nucleus with mass number 100 has a radius on the order of 6.5×10^{-15} m and a "cross-sectional area" of 1.3×10^{-28} m^2 or 1.3 b. In the simplest case when a charged particle beam is bombarding a target that is "thin" enough so that the beam does not lose any appreciable energy in passing through the target, then the production rate (R) for the radioisotope of interest is equal to

$$R = n \times I \times t \times \sigma$$

where n is the target nuclei density (nuclei cm^{-3}), I is the number of incident particles per unit time (particles s^{-1}), t is the target thickness (cm), and σ is the reaction cross-section (cm^2). In most cases, radioisotope production is performed using a thick target and an estimate of the reaction production rate takes into account the variation in the reaction cross-section with projectile energy, since the charged-particle beam loses energy as it passes through or stops in the thick target. Charged particle induced reactions used to produce medical radioisotopes have maximum cross-sections on the order of millibarns.

Neutron-rich radioisotopes that decay by negatron or β^- emission are produced primarily through neutron-induced nuclear reactions at nuclear reactors. The most commonly used reactions are direct production through single or double neutron capture, direct production through neutron induced fission, and indirect production through neutron capture followed by radioactive decay. An example of direct neutron capture is the production of the ^{166}Ho through the irradiation of ^{165}Ho in a nuclear reactor by the following reaction:

$$^{1}_{0}\mathrm{n} + {}^{165}_{67}\mathrm{Ho} \rightarrow {}^{166}_{67}\mathrm{Ho} + Q$$

Like all neutron capture reactions used to produce medical radioisotopes, this reaction is exoergic with a reaction Q value 6.24 MeV. In contrast to charged-particle induced reactions, there is no coulomb or charge barrier for neutron induced reactions and the cross-section for the neutron capture reaction increases as the energy of the neutron decreases. This increase in cross-section can be understood in that the wavelength of the neutron, and hence its probability of interacting with the target nucleus, increases as the kinetic energy or velocity of the neutron decreases. The maximum yield for most neutron capture reactions is obtained by irradiating the target material in a region of the nuclear reactor where the high energy neutrons produced by fission have been slowed down or moderated so that they have an average kinetic energy of 0.025 eV. One advantage of producing radioisotopes through neutron capture reactions is that the cross-sections are, at 0.025 eV, on the order of barns, whereas charged particle induced cross-sections have peak values on the order of millibarns. A second advantage is that many different targets can be irradiated at the same time in a nuclear reactor, while most accelerator

facilities can only irradiate one or two targets at a time. The obvious disadvantage of neutron induced reactions is that the isotope production must be performed off-site at a nuclear reactor, whereas compact cyclotrons can be sited at or near the medical facility making it possible to work with quite short-lived proton-rich isotopes.

The production rate (R) of a radionuclide during irradiation with the moderated or "thermal" neutrons in a nuclear reactor is given by

$$R = N \times \Phi \times \sigma$$

where N is the number of nuclei of the target isotopes in the sample, Φ is the flux of thermal neutrons ($n \cdot cm^{-2} \cdot s^{-1}$), and σ is the thermal neutron reaction cross-section (cm^2). Because those radionuclides created during irradiation can decay during the production process, the activity (A) in Bq of a radionuclide with decay constant λ produced by irradiating a sample either in a reactor or with a charged particle beam is given by

$$A = R(1 - e^{-\lambda t_i})e^{-\lambda t_d}$$

where R is the production rate for the reaction, t_i is the irradiation time, and t_d is the amount of time the sample has been allowed to decay following the irradiation. Using again the example of ^{165}Ho(n,γ) ^{166}Ho reaction,^{166}Ho has a half-life of 1.12 days, ^{165}Ho has a thermal cross-section of 58 b and an isotopic abundance of 100%. Irradiation of 87 mg of ^{165}Ho [100 mg target of holmium oxide (Ho_2O_3)] for 5 days in a thermal neutron flux of 1×10^{14} $n \cdot cm^{-2} \cdot s^{-1}$ will produce $\sim 1.8 \times 10^{12}$ Bq or 48 Ci of ^{166}Ho with a specific activity of 0.55 Ci of ^{166}Ho mg^{-1} of Ho in the target. The specific activity that can be achieved in the overall (nuclear and chemical) production process is a critical consideration when the radionuclide is used for therapy because it represents the fraction of the atoms in the sample that will have radio-therapeutic activity. In the example of ^{166}Ho, a specific activity of 0.55 $Ci \cdot mg^{-1}$ means that only 1 out of every 1300 Ho atoms in the sample that will be incorporated into the therapeutic agent are radioactive. A disadvantage of using direct neutron capture reactions is that the specific activity of the product radioisotope cannot be improved through chemical means (the product is an isotope of the target).

The parent material of the most commonly used medical radionuclide (^{99m}Tc) is produced in high specific activity through the neutron induced fission of ^{235}U. On average, every 100 thermal neutron induced fissions of ^{235}U produce six ^{99}Mo atoms. Because the molybdenum produced in the uranium target can be chemically separated from the uranium and other fission products, and because all of the other molybdenum isotopes have much shorter half-lives than ^{99}Mo, the process results in a sample in which nearly every molybdenum atom is ^{99}Mo. While there are a number of advantages from using this "carrier free" (i.e., 100% of the Mo is radioactive ^{99}Mo) ^{99}Mo to create the commercial ^{99}Mo/^{99m}Tc generators, the process does create significant amounts of waste that must be appropriately disposed. Neutron-rich radionuclides can also be produced in high specific activity through an indirect method that utilizes neutron capture followed by radioactive decay. Consider for example the production of ^{105}Rh from the irradiation of ruthenium target. Neutron capture on ^{104}Ru produces ^{105}Ru through the following reaction:

$$^{1}_{0}\mathrm{n} + {}^{104}_{44}\mathrm{Ru} \rightarrow {}^{105}_{44}\mathrm{Ru} + Q$$

^{105}Ru is radioactive ($t_{1/2} = 4.44$ h) and beta decays into ^{105}Rh, which has a half-life of 35.4 h. As in the fission case, the rhodium in the target can be chemically separated from the ruthenium target to produce a sample that is essentially carrier free.

SUMMARY

The use of radionuclides in medicine has been, and continues to be, important to diagnosis of disease and radiotherapy. Many diagnostic agents based on ^{99m}Tc are part of the arsenal of radiopharmaceuticals available to the physician. The last 10 years has seen the FDA approval of three new radiotherapeutic agents, namely, Quadramet (a bone pain palliation agent containing ^{153}Sm) and the first two FDA approved radioimmunoconjugates, Zevalin [a ^{90}Y labeled monoclonal antibody that specifically targets the CD20 antigen expressed on >90% of non-Hodgkin's lymphomas (NHL), and Bexxar (an ^{131}I labeled monoclonal antibody that specifically targets the CD20 antigen expressed on >90% of non-Hodgkin's lymphomas]. The development of new radiodiagnostic and radiotherapeutic agents will continue and will undoubtedly take advantage of the advances occurring in molecular biology and genomics. Radioactive decay and radionuclide production are two important aspects in the design of new radiopharmaceuticals. For more detailed discussions on these topics, the reader is referred to general textbooks on nuclear and radiochemistry (6–8).

BIBLIOGRAPHY

1. Steigman J, Eckelman WC. Nuclear Science Series (NAS-NS-3204). Nuclear Medicine, The Chemistry of Technetium in Medicine. National Academy Press; Washington (D.C.): 1992.
2. Jurisson SS, Lydon JD. Potential Technetium Small Molecule Radiopharmaceuticals. Chem Rev 1999;99:2205–2218.
3. Brucer M. In: Sorenson JA. et al., editors. The Heritage of Nuclear Medicine. The Society of Nuclear Medicine; New York: 1979.
4. Science 1946;103(2685): 698–705.
5. Lever SZ, Lydon JD, Cutler CS, Jurisson SS. Radioactive Metals in Imaging and Therapy. In: Meyer T, McCleverty J, editors. Comprehensive Coordination Chemistry II Volume 9. London: Elsevier Ltd.; 2004. pp 883–911.
6. Friedlander G, Kennedy JW, Macias ES, Miller JM. Nuclear and Radiochemistry. 3rd ed. New York: Wiley; 1981.
7. Choppin G, Rydberg J, Liljenzin JO. Radiochemistry and Nuclear Chemistry. 2nd ed. Oxford: Butterworth-Heinemann Ltd; 1995.
8. Ehmann WD, Vance DE. Radiochemistry and Nuclear Methods of Analysis. New York: Wiley; 1991.
9. Loveland WD, Morrissey D, Seaborg GT. Modern Nuclear Chemistry. New York: Wiley; 2005.

See also NEUTRON ACTIVATION ANALYSIS; TRACER KINETICS.

RADIOPHARMACEUTICAL DOSIMETRY

HUBERT M.A. THIERENS
University of Ghent
Ghent, Belgium

INTRODUCTION

In nuclear medicine, radiopharmaceuticals are administered to patients for diagnosis or treatment purposes. Each pharmaceutical compound has its specific biodistribution over organs and tissues in the body with related retention times. One method of calculating absorbed dose values delivered internally was developed in the 1960s by the medical internal radiation dose (MIRD) committee of the American Society of Nuclear Medicine (1,2). The original aim was to develop a dosimetry methodology for diagnostic nuclear medicine, but the method can be applied to dosimetry for radionuclide therapy, where the need for an accurate dosimetry is more imperative in view of the high activity levels administered to the patient. The MIRD dosimetry protocol is applied by different international organizations. The International Commission on Radiological Protection (ICRP) has published catalogs of absorbed doses to organs and tissues per unit activity administered, calculated using this dosimetric approach, for most diagnostic radiopharmaceuticals commonly applied (3,4). These tables are established for patients with standard biokinetics of the radiopharmaceutical. For application of the MIRD protocol in the nuclear medicine department when taking into account patient-specific biokinetics a user-friendly computer program called MIRDOSE was developed by Stabin (5). This software has been replaced recently by the authors by an U. S. Food and Drug Administration (FDA) approved program OLINDA (Organ Level Internal Dose Assessment) (6).

By combining well-selected β-emitting radionuclides with disease-specific pharmaceuticals, administration of radiolabeled drugs can provide efficient internal radiotherapy for localized disease as well as for metastatic cancer. As a result, an increasing number of radioactive therapeutic agents are being used in nuclear medicine for the treatment of a large variety of diseases (7–12). For these medical applications of radioactive compounds accurate patient-specific internal dosimetry is a prerequisite. Indeed, the basic goal of the majority of these types of metabolic radiotherapy is to ensure a high absorbed dose to the tumoral tissue without causing adverse effects in healthy tissues. In a curative setting an optimized activity has to be calculated and administered to the patient to ensure the delivery of a predetermined absorbed dose to the tumor resulting in complete tumor control, while minimizing the risk of normal tissue complications. The determination of latter activity necessitates a patient specific dosimetry with respect to drug pharmacokinetics and if possible patient-specific anatomical data.

Nowadays, for most applications patient-specific biokinetics are derived from sequential images after administration of a tracer activity and combined with the MIRD methodology to calculate absorbed doses to target and critical tissues (13). For a more complete dosimetric analysis as in the case of clinical trials, the information from imaging is completed by data obtained from blood sampling and urinalysis. In general, patient anatomy is represented by a standard anthropomorphic phantom (14). However, in a complete patient-specific dosimetry approach the individual patient anatomy is also taken into account and derived from computed tomography (CT) or magnetic resonance imaging (MRI). Three dimensional (3D) absorbed dose estimates are then determined from single-photon emission computed tomography (SPECT) or positron emission tomography (PET) activity data using dose-point kernel convolution methods (15,16), or by direct Monte Carlo calculation (17–20). Dose point kernels describe the pattern of energy deposited by the radiation at various radial distances from a point source. Convolution of the dose point kernel of the considered radionuclide with the activity distribution in the patient results in the absorbed dose distribution in the patient. The general idea of Monte Carlo analysis is to create a model as similar as possible to the real physical system and to create interactions within that system based on known probabilities of occurrence, with random sampling of the probability density functions. For dosimetric applications, differential cross-section data of the interactions of the ionizing particles with matter are used and the path of each particle emitted by the radioactive material is simulated until it is completely stopped. The energy deposited in the medium along the path of the ionizing particles results in the absorbed dose distribution. The advantage of direct calculation by Monte Carlo techniques is that this method allows media of inhomogeneous density to be considered. More information on the application of Monte Carlo techniques in dosimetry can be found in Ref. 21. Several software packages have been devised and validated by different groups for patient specific dose calculations. Typical examples are the 3D-ID code from the Memorial Sloan-Kettering Cancer Center (22), the RMDP-MC code from the Royal Marsden hospital (UK) (23) and the VOXEL-DOSE code from Rouen (France) (24). These programs are based on general Monte Carlo codes also used in other medical applications of ionizing radiation as external beam radiotherapy: EGSnrc (25) and GEANT (26).

Radiolabeled pharmaceuticals are also used in the development of new drugs. Before a drug can be applied to patients in phase I and II clinical trials, different steps have to be taken in the investigation of the toxic effects of the new (radio)pharmaceutical compound. This involves firstly a number of animal studies followed by the administration of the pharmaceutical to a restricted number of volunteers. In general, for these animal and volunteer studies, a radiolabeled formulation of the newly elaborated drug is used with ^{3}H or ^{14}C as radionuclide. Sacrifice of the animals at different time points postadministration and quantitative whole-body autoradiography allow the determination of the biodistribution with metabolite profiling, the retention in the different organs and tissues, and the study of excretion pathways of the pharmaceutical in the animals. In the development of radiopharmaceuticals specifically for nuclear medicine imaging purposes, the new compound can be labeled with gamma-emitting radionuclides and the biokinetics are derived from serial

imaging of animals. To this end, dedicated (micro)SPECT and (micro)PET systems were constructed (27–29). The animal activity data are extrapolated to humans to determine the maximal activity of the radiolabeled compound allowed to be administered to healthy volunteers. Criterion is here that the effective dose may not exceed the limits for the considered risk category of the volunteers following the ICRP Publication 62 categories (30). In general, the risk category IIa (risk $\sim 10^{-5}$) with a maximal effective dose of 1 mSv is appropriate for volunteers in testing new drugs. This corresponds to an intermediate-to-moderate level of social benefit for a minor to intermediate risk level for the volunteers. This evaluation procedure necessitates a reliable dosimetry estimate based on the extrapolation of animal activity data to humans. To obtain this dose estimate generally the MIRD formalism is applied.

THE MIRD SCHEMA

Basic Principles and Equations

For patient dosimetry in nuclear medicine diagnostic procedures, the MIRD schema is applied. This formalism is also used systematically for dose calculations in administration of radiolabeled drugs of volunteers in the framework of drugs development.

In the MIRD protocol, organs and tissues in the body with a significant uptake of the radiopharmaceutical are considered as source organs. On the other hand all organs and tissues receiving an absorbed dose are considered as target organs. This is illustrated in Fig. 1 for iodine isotopes as ^{131}I with the thyroid as source organ and the lungs as the target organ on the anthropomorphic phantom developed by Snyder et al. (31). The absorbed dose to a

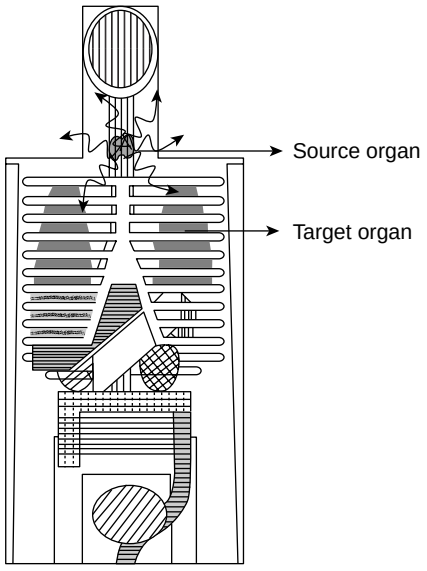


Figure 1. Illustration of the MIRD method for the γ rays emitted in the decay of iodine isotopes as ^{131}I with the thyroid as source organ and the lungs as target organ in the anthropomorphic phantom developed by Snyder et al. (31).

particular target organ t from a source organ s, $D_{t\leftarrow s}$, can be obtained using the following equation (1,2) :

$$D_{t\leftarrow s} = \tilde{A}_s S_{t\leftarrow s}$$

with $\tilde{A}_s$ is the cumulated activity in the source organ and $S_{t\leftarrow s}$ is the mean dose to the target organ per unit cumulated activity in the source organ.

The cumulated activity $\tilde{A}_s$ is the total number of disintegrations of the radioactivity present in the source organ integrated over time and expressed in units Bq.s. It depends on the activity administered, the uptake, retention and excretion from the source organ, and the physical decay of the radionuclide. The $S_{t\leftarrow s}$ values depend on the decay modes of the considered radionuclide and the source-target geometry. The $S_{t\leftarrow s}$ values are tabulated for standard men and children anthropomorphic phantoms (14).

The cumulated activity $\tilde{A}_s$ is the time integral of the activity in the source organ $A_s(t)$:

$$\tilde{A}_s = \int_0^\infty A_s(t)dt$$

The biological retention in the source organ is generally derived from sequential scintigraphies with a gamma camera. For this, opposing planar views or SPECT are used with a calibrated source in the field of view. Corrections are needed for patient attenuation and scatter of the γ radiation. The cumulated activity in the different source organs $\tilde{A}_s$ allows to calculate the residence time τ being the average time the administered activity A_0 spends in the considered source organ:

$$\tau = \frac{\tilde{A}_s}{A_0}$$

In the MIRDOSE software, the cumulated activity in the different source organs is introduced by the values of the residence time (5).

The mean dose to the target organ per unit cumulated activity in the source organ, $S_{t\leftarrow s}$, is given by the expression:

$$S_{t\leftarrow s} = \frac{1}{m_t}\sum_i \Delta_i \phi_i(t\leftarrow s)$$

with m_t mass of the target organ, Δ_i the mean energy emitted per disintegration for radiation of type and energy i, and $\varphi_i(t\leftarrow s)$ the absorbed fraction for radiation of type and energy i. The absorbed fraction $\varphi_i(t\leftarrow s)$ is defined as the fraction of the radiation of type i emitted by the source organ s absorbed in the considered target organ t. The Δi values are obtained from the decay scheme of the considered radionuclide. The values of the specific absorbed fractions $\varphi_i(t\leftarrow s)/m_t$ were calculated by Monte Carlo methods (32). The $S_{t\leftarrow s}$ values for commonly used isotopes in nuclear medicine calculated in this way for a number of standard anthropomorphic phantoms including children of different ages (14) are tabulated and included in the data base of the MIRDOSE package (5). This procedure assumes a uniform distribution of the activity over the source organs and a standard anatomy of the patient.

Table 1. Effective Dose Values Per Unit Activity Administered For A Number Of Radiopharmaceuticals Commonly Applied In Nuclear Diagnostics [a]

Radiopharmaceutical	Effective Dose Per Unit Activity Administered (mSv/MBq)				
	1 year	5 years	10 years	15 years	Adult
^{18}F FDG	0.095	0.050	0.036	0.025	0.019
^{67}Ga citrate	0.64	0.33	0.20	0.13	0.10
^{99m}Tc-DTPA	0.016	0.0090	0.0082	0.0062	0.0049
^{99m}Tc-HMPAO	0.049	0.027	0.017	0.011	0.0093
^{99m}Tc-MIBI	0.053	0.028	0.018	0.012	0.0090
^{99m}Tc-MDP	0.027	0.014	0.011	0.0070	0.0057
^{99m}Tc-pertechnetate	0.079	0.042	0.026	0.017	0.013
^{99m}Tc-leucocytes	0.062	0.034	0.022	0.014	0.011
^{111}In-octreotide	0.28	0.16	0.10	0.071	0.054
^{123}I uptake 35%	2.05	1.08	0.51	0.34	0.22
^{123}I-MIBG	0.068	0.037	0.026	0.017	0.013
^{201}Tl-chloride	2.80	1.70	1.20	0.30	0.22

[a] See Ref. 4.

By using the MIRD working procedure, the absorbed doses of the different target organs for frequently used radiopharmaceuticals per unit activity administered were calculated by the ICRP for an adult and children of 1, 5, 10, and 15 years assuming standard biokinetics and were tabulated (3,4). As measure of the radiation burden patient the effective dose E is calculated by summing up the tissue equivalent doses H_T using the tissue weighting factors w_T as defined in the ICRP 60 publication (33):

$$E = \sum_T w_T H_T$$

For the types of radiation emitted by the radionuclides used in nuclear medicine the radiation weighting factor w_R is one except for alpha particles, where w_R equals 20. In Table 1 the effective dose values per unit activity administered for a number of radiopharmaceuticals commonly applied in nuclear diagnostics under the assumption of standard biokinetics is summarized. Table 1 shows that in diagnostic pediatric nuclear medicine the patient dose is strongly dependent on patient age for the same administered activity. This is mostly due to the change in patient weight. Weight dependent correction factors for the activity to be administered have been calculated to obtain weight independent effective doses (34,35). The concept of effective dose is intended to estimate the risk for late stochastic radiation effects as radioinduced cancer and leukemia in the low dose range, and by this applicable to nuclear medicine investigations for diagnosis. Its value is not representative for the risk for direct deterministic effects as bone marrow depletion in case of therapeutic applications of radiopharmaceuticals.

Microdosimetric Considerations

The S-values commonly applied at the macroscopic level are calculated assuming a uniform distribution of the activity over the source organ and the target being the whole volume of the target tissue. The use of S-values based on these assumptions can lead to erroneous results at the microscopic level in case of self-dose calculation in an organ (target = source) when the isotope distribution is nonuniform at the cellular level and particles with range of the order of cellular dimensions are emitted in the decay. This is particularly the case when the radionuclide used is an Auger electron or an alpha emitter. A typical example is the dosimetry of lymphocytes labeled with ^{99m}Tc. In Fig. 2, the therapeutic range in soft tissue of the low energy electron groups in the decay of ^{99m}Tc is represented and compared to the dimensions of the DNA helix (2 nm) and a lymphocyte (10 μm). Taking into account the range of the large intensity Auger electron groups (2–100 nm) the dose to the DNA in the cell nucleus, which is the biological radiation target in the cell, is strongly dependent if we consider intra- or extracellular distribution of a ^{99m}Tc radiopharmaceutical. In most radiopharmaceuticals, ^{99m}Tc is located extracellularly and the radiation burden from the Auger electrons to the nucleus is very low. In those cases, the cellular dose is due to the 140 keV γ-emission and the macroscopic S-values assuming a uniform distribution of the ^{99m}Tc activity can be used for the dose calculation. However, in case of intracellular labeling as in the case of labeling of lymphocytes with ^{99m}Tc-HMPAO the dose to the lymphocytes due to the Auger electrons is very high, which leads to radiotoxic effects in these cells (36).

For dose assessment in case of intracellular labeling or labeling of the membrane with an Auger electron emitter a microdosimetric approach based on Monte Carlo calculation methods is indicated. In those cases, the MIRD method can be applied at the cellular level for the calculation of the cell nucleus dose from activity uniformly present in the nucleus, the cell cytoplasm or the cell surface using appropriate microscopic $S_{t \leftarrow s}$ values tabulated for all radionuclides (37). To cope with nonuniform activity distributions in source organs, determined by PET and SPECT imaging, radionuclide voxel S-values are tabulated for five radionuclides for cubical voxels of 3 and 6 mm (38).

DOSIMETRY FOR RADIONUCLIDE THERAPY

Methodology

In external beam radiotherapy, there is a long tradition in performing treatment planning calculations for each

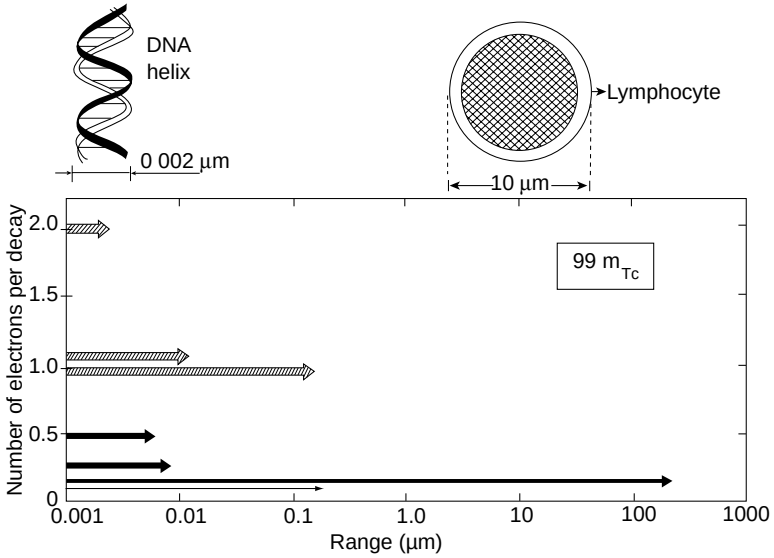


Figure 2. The range in soft tissue of the low energy electron groups in the decay of ^{99m}Tc, compared to the dimensions of the DNA helix and a lymphocyte. The latter comparison is relevant for dosimetry in the intracellular labeling of lymphocytes with ^{99m}Tc-HMPAO.

individual patient. The dosimetry protocols necessary for metabolic radiotherapy, however, are far more complex than those used in external beam therapy. In fact, the *in vivo* activity distribution initially is patient-specific and unknown in both space and time. For the determination of the patient specific drug pharmacokinetics, a tracer activity of the radiopharmaceutical is administered to the patient and quantitative imaging at multiple time points is employed to establish patient-specific biokinetics (13). Here, nuclear medicine imaging with proper correction for photon attenuation, scatter, and collimator resolution is needed to obtain the most accurate activity maps possible. The patient-specific biokinetics can then be combined with the MIRD methodology, described above, to calculate absorbed doses to organs and tissues. The MIRDOSE software allows to calculate the self-absorbed dose to a sphere representing the tumor in case of oncological applications. From these dosimetric calculations, the activity of the radiopharmaceutical to be administered to deliver the prescribed absorbed dose level to the considered tissues is then calculated by extrapolation. This approach does not take into account the patient anatomy. Instead, anatomical data of the average male, female, and children of different ages are introduced by anthropomorphic phantoms (14).

More accurate dosimetric calculations require that the individual patient anatomy derived from CT or MRI images is converted into a 3D voxel representation as in external beam radiotherapy. The 3D absorbed dose estimates from the tracer activity administration are then determined from SPECT or PET activity imaging using dose-point kernel convolution methods, or by direct Monte Carlo calculation (15–24). This approach necessitates image fusion between the different imaging modalities lused. The advent of combined SPECT–CT and PET–CT equipment allows a more general application of this complete patient-specific dosimetry (39–43). In this setting, the CT data may be used as an attenuation map, which is an important improvement for accurate quantification (39).

A dosimetry calculation can be useful not only for assessment of the amount of activity to be administered before radionuclide therapy, but also after the performed radionuclide therapy. First, it is important to verify the predicted absorbed dose distribution. Second, the dosimetry results of a patient population can be combined with the outcome of the therapy to analyse the dose-response of the radionuclide therapy and to make changes in the therapy protocol when necessary (e.g., the predetermined target dose level). As was the case for the pretherapy calculation of the administered activity, posttherapy dosimetry can be performed at different levels of sophistication.

Dosimetry is not only important in the framework of therapy prediction, but also in the dose assessment of organs at risk. In radiopeptide therapy, the kidneys are the dose-limiting organ (44,45). Radiopeptides are cleared physiologically via the kidneys. Most peptides are cleaved to amino acids as metabolites in the kidneys with a high and residualizing uptake in the tubular cells. Damage to the kidneys induced by the radiolabeled metabolites can cause nephropathy after therapeutic application of radiopeptides (46). Application of basic amino acids can reduce the renal accretion of radiolabeled metabolites and the kidney dose (47).

Dosimetry of Radioiodine Therapy for Thyrotoxicosis

The most common application of radionuclide therapy is treatment of hyperthyroidism as observed in Graves' disease or Plummer's disease (toxic nodular goiter) by oral administration of ^{131}I. The rationale behind dosimetry for this kind of treatment is that at long-term hypothyroidism may be the outcome for patients treated with radioiodine and that the incidence of this inverse effect is higher with an earlier onset for patients treated with higher activities (48). A large variation exists in the literature on the value of target dose to be delivered to the hyperthyroid tissue to become euthyroid. Howarth et al. (49) reported that doses of 60 and 90 Gy cured 41 and 59 % of patients after 6 months. Guhlmann et al. (50) cured hyperthyroidism in 83 % of patients at 1 year post-treatment with a dose of 150 Gy. According to Willemsen et al. (51) hyperthyroidism is

eliminated in all patients 1 year post-treatment with a dose of 300 Gy, but at this high dose level 93% of patients became hypothyroid.

For dose calculation in general an adapted version of the Quimby-Marinelli formula (52) has to be used

$$A(\mu\text{Ci}) = \frac{6.67 \times \text{Dose (cGy)} \, \text{mass (g)}}{T_{1/2\text{eff}}(\text{days}) \times \%\,\text{uptake}(24\,\text{h})}$$

Application of this protocol for individual patient dosimetry necessitates the determination of the following important variables: percentage uptake 24 h after administration, effective half-life of the radioiodine, and mass of the thyroid gland. For uptake and kinetics assessment, serial scintigraphies or probe measurements of the patient's thyroid after administration of a tracer dose have to be performed. This approach assumes that the kinetics of a tracer and a therapeutical amount of administered activity are the same. According to some authors, a pretherapeutic tracer dose may induce a stunning effect limiting the uptake of the therapeutic activity in the thyroid afterward (53). The thyroid mass is generally determined by the pretherapeutic scintigraphy, by ultrasonography or by MRI (54). A ^{124}I PET image also allows measurement of the functioning mass of the thyroid (55). Dosimetry protocols exist based on only a late uptake measurement at 96 or 192 h after tracer activity administration (56). A thorough discussion of the activity to be administered and the dosimetry protocol to follow can be found in Refs. 57,58.

Dosimetry of Radioiodine Therapy for Differentiated Thyroid Cancer

Radioiodine is also administered frequently to patients for differentiated thyroid cancer to ablate remnant thyroid tissue in the early postoperative period, for locoregional recurrences, and for distant metastases. Although most centers administer standard activities, typically 2.8–7.4 GBq (75–200 mCi), because of the practical difficulties to determine the target absorbed dose, absorbed dose-based protocols are also applied (59). For the calculation of the activity to be administered to give a predetermined tumor absorbed dose protocols as for thyrotoxicosis treatment described earlier are used. As predetermined absorbed dose-to-remnant thyroid tissue a value of 300 Gy is considered to be sufficient (60). For treatment of metastases lower doses giving a complete response have been reported: 85 Gy (61) and 100-150 Gy (62). This approach necessitates determination of the remnant mass of thyroid tissue or metastases by the methods described earlier, which is now more difficult in practice. This introduces in general a large uncertainty on the activity to be administered to ensure the desired dose to the target tissue. Also, the radioiodine kinetics with the 24 h uptake and the effective half-life has to be determined for the patient by administration of a tracer dose. Because of the relatively high activities necessary for quantitative imaging of the target thyroid tissue for this application (at least 37 MBq-1 mCi ^{131}I) complication of the therapy by stunning introduced by the tracer activity mentioned earlier, is more critical here. Because of this and the inaccuracy in the target mass determination, dosimetry protocols based on target dose levels remain difficult for treatment of differentiated thyroid cancer. The ^{124}I PET imaging allows a more exact *in vivo* determination of iodine concentration and volume determination. By using this method, radiation doses to metastases ranging between 70 and 170 Gy were delivered to the lesions (63).

Instead of target absorbed dose-based protocols, dosimetry protocols based on the largest safe approach are also applied. This approach based on the dose to the critical tissues allows the administration of the maximum possible activity to achieve the maximum therapeutic efficacy. Application of this method necessitates serial total body scintigraphy after the administration of a tracer dose. The dose to the bone marrow, the lungs, and the thyroid tissue or metastases is then calculated by the MIRD formalism. From the absorbed doses obtained by the tracer activity imaging the amount of activity giving the maximal tolerable absorbed dose to the critical tissues is calculated. It has been generally accepted that the activity that delivers 2 Gy whole body dose as a surrogate for the bone marrow dose with a whole body retention < 4.44 GBq (120 mCi) at 48 h postadministration is safe with respect to bone marrow suppression (64). In some departments, the tolerable dose level of the bone marrow in radioiodine treatment of patients with metastatic differentiated thyroid cancer is even increased to 3 Gy based on the $LD_{5/5}$ data of external beam radiotherapy with $LD_{5/5}$ being the dose for the red marrow giving a 5% risk of severe damage to the blood-forming system within 5 years after administration (62). Very high activities of ^{131}I in the range 7.4–37.9 GBq (200–1040 mCi) are then administered for treatment of metastases. In a retrospective study of patients treated with this protocol over a period of 15 years, transient bone marrow depression with thrombopenia and leukopenia was observed recovering after a few weeks (62). No permanent damage was observed. In $\sim$ 10% of the patients the dose-limiting organ were the lungs for which a limit of 30 Gy was adopted from $LD_{5/5}$ data.

A recent review of the evolving role of ^{131}I for the treatment of differentiated thyroid carcinoma can be found in Ref. 9. Preparation of patients by administration of recombinant human thyroid-stimulating hormone (rhTSH) may allow an increase in the therapeutic radioiodine activity while preserving safety and tolerability (65). As side effects of the ^{131}I therapy, impairment of the spermatogenesis in males (66) and earlier onset of menopause in older premenopausal women (67) are reported. With respect to pregnancy, it is recommended that conception be delayed for 1 year after therapeutic administrations of ^{131}I and until control of thyroid hormonal status has been achieved. After this period there is no reason for patients exposed to radioiodine to avoid pregnancy (68).

Dosimetry of ^{131}I-MIBG Therapy

Another radionuclide therapy application for which the importance of patient-specific dosimetry is generally accepted is treatment of pediatric neuroblastoma patients with ^{131}I-MIBG. Neuroblastoma is the most common

extracranial solid tumor of childhood with an incidence of 1/70000 children under the age of 15 (69). Neuroblastoma cells actively take up nor-adrenalin via an uptake-1 system. The molecule *meta*-iodo benzyl guanidine (MIBG), radiolabeled with ^{131}I, has a similar molecular structure, uptake, and storage in the cell as nor-adrenalin. Since 1984, ^{131}I-MIBG has been used therapeutically in neuroblastoma patients (70,71). Aside from the tumor, ^{131}I-MIBG is also taken up in the liver, heart, lungs, and adrenal glands. The bladder is irradiated by the metabolites of ^{131}I-MIBG. For patient dosimetry the largest safe dose approach is applied in ^{131}I-MIBG therapy with bone marrow as dose limiting organ. In practice, the whole body absorbed dose is also used in this setting as an adequate representation or index of bone marrow toxicity. Most treatment regimens consider the maximal activity to be administered limited by rendering a bone marrow dose of 2 Gy. Prediction of whole body doses is based on a pretherapeutic administration of ^{123}I-MIBG. In Fig. 3, predicted whole body doses based on pretherapeutic ^{123}I-MIBG scintigraphies are compared to doses received by patients after ^{131}I-MIBG therapy (72). The received dose values were derived from post-therapy scans. This figure shows also that in the case of repeated therapies pre-therapy scans do not need to be repeated before each therapy except when the biodistribution of ^{131}I-MIBG is expected to change rapidly (e.g., for patients where bone marrow invasion is present). It has also been shown that the accuracy of whole body dosimetry improves when half-life values of tracer and therapy radionuclides are matched (73).

In ^{131}I-MIBG therapy, protocols with administration based on fixed activity per unit mass protocols are also applied. Matthay et al. (74) reported on dosimetry performed in a dose escalation study of patients treated with ^{131}I-MIBG for refractory neuroblastoma with a fixed activity per unit mass ranging from 111 to 666 MBq·kg^{-1} (3–18 mCi·kg^{-1}). Patients treated with a specific activity < 555 MBq·kg^{-1} did not require hematopoietic stem cell support, while this was necessary for one-half of the patients treated with a higher specific activity. The median whole body dose of the group of patients requiring hematopoietic stem cell support was 3.23 Gy (range 1.81–6.50 Gy) while for the other patients the median dose was 2.17 Gy (range 0.57–5.40 Gy).

In order to improve the results of ^{131}I-MIBG therapy for patients refractive of extensive chemotherapy treatment, a high activity ^{131}I-MIBG schedule is now being used in combination with topotecan as radiosensitizer for the therapy of neuroblastoma in a controlled ESIOP (European International Society of Pediatric Oncology for Neuroblastoma) study protocol (75). The aim here is to administer in two fractions the amount of activity needed to reach a combined total body dose of 4 Gy. These kinds of high doses will inevitably invoke severe side effects, thus frequently necessitating hematopoietic stem cell support and even bone marrow transplantation. However, a contemporary oncological department is well equipped to deal with this kind of treatments. The first amount of ^{131}I-MIBG activity is administered based on a fixed activity per unit body mass (444 MBq·kg^{-1}–12 mCi·kg^{-1}) protocol. Total body dosimetry is carried out using serial whole body scintigraphies after the first administration. Radionuclide kinetics are followed by a whole body counter system mounted on the ceiling of the patient's isolation room. These dosimetry results are then used to calculate the activity of the second administration of ^{131}I-MIBG giving a total body dose of 4 Gy over the two administrations. The first results of this study indicate that *in vivo* dosimetry allows for an accurate delivery of the specified total whole body dose and that the treatment schedule is safe and practicable (75). The approach has now to be tested for efficacy in a phase II clinical trial.

Dosimetry in Radioimmunotherapy

In general, red marrow is the dose-limiting tissue in non-myeloablative and lung for myeloablative radioimmunotherapy (RIT). Administration protocols are applied based on absorbed-dose values of the dose-limiting tissue and on an activity per body weight basis. Typical examples

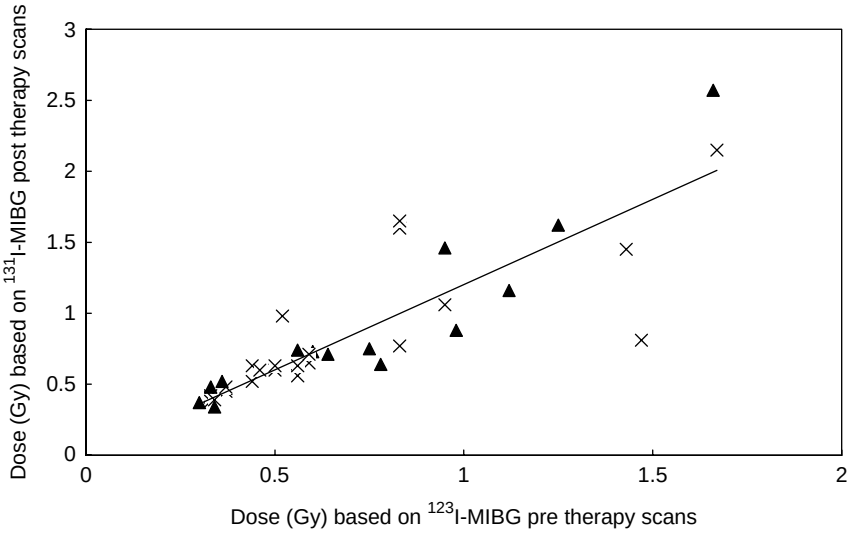


Figure 3. Correlation between the whole body dose estimate based on ^{123}I-MIBG pretherapy scans and the dose derived from ^{131}I-MIBG posttherapy scans in patients treated for neuroendocrine tumors. The triangles represent the data of the first therapies, the crosses the data of retreatments. The straight line is the result of a linear regression to all data ($R^2 = 0.73$).

of these protocols are the ^{131}I-labeled anti-CD20 antibody, tositumomab (Bexxar; Glaxo-SmithKline) (76) and the ^{90}Y-labeled anti-CD20 ibritumomab tiuxetan (Zevalin; Biogen Idec) (77), respectively. These radiolabeled antibodies are used for treatment of non-Hodgkin's lymphoma. The choice for an activity-based protocol for the ^{90}Y-labeled antibody is based on the lack of correlation between absorbed dose and toxicity in the early studies. The explanation for the absence of a dose-response relationship can be found in different sources. In contrast to ^{131}I, ^{90}Y is a pure β-emitter, and ^{90}Y kinetics have to be derived from surrogate ^{111}In imaging. Another point is that prior treatment of these patients and the bone marrow reserve have a strong effect on the bone marrow toxicity in this case. As patients undergoing RIT have been treated previously by chemotherapy, the impact of such prior therapy on the hematopoietic response to the RIT is important.

Although the necessity of patient-specific dosimetry is questionable in some applications of RIT where the dose-response observations for toxicity are poor, there is a general agreement that complete radiation dosimetry is necessary for each new application of a radiolabeled antibody in phase I and most probably also in phase II studies especially for safety reasons (78). An important argument for absorbed dose driven protocols in clinical phase I trials is that many patients are treated below the biologically active level due to the interpatient variability in activity based administration protocols. This implies data difficult to interpret in antitumor response and toxicity.

In view of the central role of red marrow toxicity in RIT methodologies, bone marrow dosimetry got already a lot of attention in the literature (79–85). In general, methods based on imaging as described in the section on ^{131}I-MIBG therapy are used. Also, approaches to calculate the bone marrow dose based on blood activity measurements have been described, but these methods yield only reliable results when the activity does not bind specifically to blood or marrow components including tumor metastases in the marrow (79). By assuming rapid equilibrium of radiolabeled antibodies in the plasma and the extracellular fluid of the red marrow, a red marrow/blood concentration ratio of 0.3–0.4 can be derived. All red marrow dosimetry performed up to now uses a highly stylized representation of the red marrow over the body. More detailed representations are being generated especially for Monte Carlo calculations enhancing accuracy and reliability of the bone marrow doses (86).

Several studies have investigated the relation between the tumor dose and response especially in RIT of non-Hodgkin's lymphoma (87–89) but the results are negative. Possible explanations are the therapeutic effect of the antibody, different confounding biological factors and the accuracy of tumor dosimetry. Here, standardization of data acquisition as presented in MIRD pamphlet No. 16 (13) may help in dose-response investigations. As discussed earlier in the section on ^{131}I-MIBG therapy, a full patient-specific 3D dosimetric approach with imaging data from the combined SPECT–CT systems will improve substantially the accuracy of the tumor dosimetry results.

DOSIMETRY IN THE DEVELOPMENT OF NEW DRUGS

For the study of the absorption, metabolism, and excretion pathways of new drugs a ^{3}H- or ^{14}C-radiolabeled formulation of the drug is administered to healthy volunteers. A dosimetric evaluation of the radiation burden of the volunteers based on animal biodistribution, retention, and excretion data is necessary and presented to an ethical committee before the radiolabeled drug can be administered. This procedure has to ensure that the effective dose will not exceed the limits for the considered risk category of the volunteers according to the ICRP publication 62 categories (30). For testing new drugs mostly a risk category IIa (risk $\sim 10^{-5}$) is adopted corresponding to a maximal effective dose of 1 mSv. Based on this criterion, the activity to be administered is calculated from the dosimetric evaluation.

For the calculation of the dose estimate of the volunteers the MIRD formalism for an administration of a standard activity (37 kBq/1 μCi) of the radiolabeled pharmacon is applied. Animal biodistribution data are used to calculate the residence time in the source organs and tissues based on the maximum uptake f and biological half-life. In general a rat strain is used as animal model. As the organ weights in the rat and man are different an important correction of the animal data is necessary to estimate the f-values in humans. For each organ, dosimetric calculations are performed assuming (*1*) the same fraction of activity is absorbed by the organs in rat and humans irrespective of the difference in relative weight or (*2*) the fraction of activity absorbed by each organ is proportional to the relative organ weight in rat and humans. The latter assumption means that the uptake per kilogram of organ weight normalized to the whole body weight is the same for both species. Table 2 gives an overview of the organ and tissue weights in a male Wistar rat of 250 g reported in the literature (90) and in the standard human of 70 kg (32). For each organ or tissue two dose values are obtained by assuming a species independent organ uptake and an uptake proportional to the relative organ weight in different species. The highest dose estimate of both is restrained for each organ. If the retention for the individual organs is not known the whole body retention is adopted.

As model for the liver and biliary excretion it is generally assumed that a fraction of the radiopharmaceutical is taken up by the liver. Part of this activity goes directly to the small intestine while the resting part goes to the gallbladder, from where it is cleared to the small intestine. For the total fraction of activity excreted in this way by the gastrointestinal tract, the fraction of the activity retrieved in the feces is adopted from animal data. In general, data are available for different species and the maximal value is retained. For the dose calculation of the sections of the gastrointestinal tract, the kinetic model of the ICRP publication 53 is adopted (3). The kidney–bladder model described in this publication is also used to calculate the dose to the urinary bladder. Urine activity measurements in animals are used to estimate the fraction of the activity eliminated through the kidneys and again the maximal value is adopted if data are available for different species.

The dose estimates to organs and tissues of humans extrapolated in this way from animal data are combined

Table 2. Organ Weights of a Male Wistar Rat of 250 g[a] and Human of 70 kg[b]

Organ	RAT Weight, g	RAT Rel. Weight, %	Humans Weight, kg	Humans Rel. Weight, %
Adrenal glands	0.085	0.034	0.014	0.020
Blood	15	6	5.5	7.86
Bone	12.99	5.19	5	7.14
Bone marrow	5.59	2.24	3	4.29
Brain	1.43	0.574	1.4	2.0
Heart	0.835	0.334	0.33	0.47
Kidneys	1.873	0.749	0.31	0.44
Large intestine	2.635	1.054	0.37	0.53
Liver	10.675	4.27	1.8	2.57
Lungs	1.618	0.647	1	1.43
Oesophagus	0.11	0.044	0.04	0.057
Pancreas	0.913	0.365	0.1	0.14
Plasma	10	4	3.1	4.43
Prostate	0.3	0.12	0.016	0.023
Small intestine	7.30	2.92	0.64	0.91
Spleen	0.738	0.295	0.18	0.26
Stomach	1.23	0.492	0.15	0.21
Testes	1.815	0.726	0.035	0.050
Thymus	0.593	0.237	0.02	0.029
Thyroid	0.02	0.008	0.02	0.029

[a]See Ref. 90.
[b]See Ref. 32.

with the tissue weighting factors to obtain the effective dose after the administration of the standard activity (37 kBq/1 μCi) as described earlier (33). Based on the effective dose estimate obtained in this way and the dose limits proposed in the ICRP 62 publication (30) the activity of the radiolabeled drug to be administered to the volunteers is obtained.

CONCLUSIONS

To estimate the risk for late radiation effects as cancer and leukemia in patients after administration of radiopharmaceuticals for diagnosis the MIRD formalism with standard human anatomy and biokinetics is generally applied. This holds also for the estimation of the same risk of volunteers after the administration of a radiolabeled formulation of newly developed drugs. However, therapeutic applications of radiopharmaceuticals necessitate a reliable patient specific approach at least with respect to the biokinetics and if possible also for the patient-specific anatomical data. For curative treatment of malignant diseases there is now a tendency to use the largest safe dose approach with administration of the maximum possible activity based on the dose to the critical tissues. On the other hand, the advent of combined SPECT-CT or PET-CT imaging means an essential step forward toward an accurate 3D tumor dosimetry, the basic need for the administration protocols with the calculated activity based on a tumor dose prescription as used in external beam radiotherapy.

BIBLIOGRAPHY

1. Loevinger R, Berman M. A schema for absorbed-dose calculations for biologically distributed radionuclides. MIRD pamphlet No. 1. J Nucl Med 1968;9(Suppl.1):7–14.
2. Loevinger R, Budinger TF, Watson EE. MIRD primer for absorbed dose calculations, New York: Society of Nuclear Medicine; revised 1991.
3. ICRP publication 53. Radiation dose to patients from radiopharmaceuticals. Annals of the ICRP Vol 18. Oxford: Pergamon; 1987.
4. ICRP publication 80. Radiation dose to patients from radiopharmaceuticals. Annals of the ICRP Vol 28. Oxford: Pergamon press; 1998.
5. Stabin MG. MIRDOSE: personal computer software for internal dose assessment in nuclear medicine. J Nucl Med 1996;37:538–546.
6. Stabin MG, Sparks RB, Crowe E. OLINDA/EXM: The second generation personal computer software for internal dose assessment in nuclear medicine. J Nucl Med 2005;46:1023–1027.
7. McDougall IR. Systemic radiation therapy with unsealed radionuclides. Sem Rad Oncol 2000;10:94–102.
8. Knox SJ, Meredith RF. Clinical radioimmunotherapy. Sem Rad Oncol 2000;10:73–93.
9. Robbins RJ, Schlumberger MJ. The evolving role of 131I for the treatment of differentiated thyroid carcinoma. J Nucl Med 2005;46:28S–37S.
10. Valdés-Olmos RA, Hoefnagel CA. Radionuclide therapy in oncology: the dawning of its concomitant use with other modalities. Eur J Nucl Med Mol Imaging 2004;32:929–931.
11. Larson SM, Krenning EP. A pragmatic perspective on molecular targeted radionuclide therapy. J Nucl Med 2005; 46:1S–3S.
12. Kwekkeboom DJ, et al. Overview of results of peptide receptor radionuclide therapy with 3 radiolabeled somatostatin analogs. J Nucl Med 2005;46:62S–66S
13. Siegel JA, et al. MIRD Pamphlet No. 16: Techniques for quantitative radiopharmaceutical biodistribution data acquisition and analysis for use in human radiation dose estimates. J Nucl Med 1999;40:37S–61S.
14. Cristy M, Eckerman KF. Specific absorbed fractions of energy at various ages from internal photon sources. ORNL Report ORNL/TM-8381. Oak Ridge: Oak Ridge National Loboratory; 1987.

15. Giap HB, Macey DJ, Bayouth JE, Boyer AL. Validation of a dose-point kernel convolution technique for internal dosimetry. Phys Med Biol 1995;40:365–381.
16. Furhang EE, Sgouros G, Chui CS. Radionuclide photon dose kernels for internal emitter dosimetry. Med Phys 1996;23: 759–764.
17. Furhang EE, Chui CS, Sgouros G. A Monte Carlo approach to patient-specific dosimetry. Med Phys 1996;23:1523.
18. Liu A, Wiliams LE, Wong JYC, Raubitscek AA. Monte Carlo assisted voxel source kernal method (MAVSK) for internal dosimetry. J Nucl Med Biol 1998;25:423–433.
19. Yoriyaz H, Stabin MG, dos Santos A. Monte Carlo MCNP-4B-based absorbed dose distribution estimates for patient-specific dosimetry. J Nucl Med 2001;42:662.
20. Zaidi H, Sgouros G, editors. Therapeutic applications of Monte Carlo calculations in Nuclear Medicine. Bristol (UK): Institute of Physics Publishing; 2002.
21. Andreo A. Monte Carlo techniques in medical radiation physics. Phys Med Biol 1991;36:861–920.
22. Sgouros G, et al. Three-dimensional dosimetry for radioimmunotherapy treatment planning. J Nucl Med 1993;34:1595–1601.
23. Guy MJ, Flux GG, Papavasileiou P, Flower MA, Ott RJ. RMDP-MC: a dedicated package for I-131 SPECT quantification, registration, patient-specific dosimetry and Monte Carlo. Seventh International Radiopharmaceutical Dosimetry Symposium. Proceedings of the International Symposium Nashville; (TN): Oak Ridge Associated Universities; 2002.
24. Gardin I. Voxeldose: a computer program for 3D dose calculation in therapeutic nuclear medicine. Seventh International Radiopharmaceutical Dosimetry Symposium. Proceedings of the International Symposium Nashville; (TN): Oak Ridge Associated Universities; 2002.
25. Kawrakow I, Rogers DWO. The EGSnrc code system: Monte Carlo simulation of electron and photon transport. NRC Report PIRS-701. Ottawa: National Research Council of Canada; 2000.
26. Carriers JF, Archembault L, Beaulieu L. Validation of GEANT4, an object-oriented Monte Carlo toolkit for simulations in medical physics.Med Phys 2004;31:484–492.
27. Weber S, Bauer A. Small animal PET: aspects of performance assessment. Eur J Nucl Med Mol Imaging 2004;31:1545–1555.
28. Sossi V, Ruth TJ. Micropet imaging: *in vivo* biochemistry in small animals. J Neural Transmission 2005;112:319–330.
29. Tai YC, et al. Performance evaluation of the microPET focus: a third generation microPET scanner dedicated to animal imaging. J Nucl Med 2005;46:455–463.
30. ICRP Publication 62. Radiological protection in biomedical research. Annals of the ICRP Vol. 22. Oxford: Pergamon; 1991.
31. Snyder WS, Ford MR, Warner GG, Watson SB. MIRD Pamphlet No. 11. "S", Absorbed dose per unit cumulated activity for selected radionuclides and organs. New York: Society of Nuclear Medicine; 1975.
32. ICRP Publication 23. Report of the task group on reference man. Oxford: Pergamon; 1975.
33. ICRP Publication 60. 1990 Recommendations of the International Commission on Radiological Protection. Oxford: Pergamon; 1991.
34. Piepsz A, et al. A radiopharmaceutical schedule for imaging in paediatrics. Eur J Nucl Med 1990;17:127–129.
35. Jacobs F, et al. Optimized tracer-dependent dosage cards to obtain weight-independent effective doses. Eur J Nucl Med Mol Imaging 2005;24:
36. Thierens H, Vral A, Van Haelst JP, Van de Wiele C, Schelstraete K, De Ridder L. Lymphocyte labeling with Technetium-99m-HMPAO: A Radiotoxicity Study using the Micronucleus Assay. J Nucl Med 1992;33:1167–1174.
37. Goddu SM, et al. MIRD Cellular S values. New York: Society of Nuclear Medicine; 1997.
38. Bolch WE, et al. MIRD pamphlet No. 17: the dosimetry of nonuniform activity distributions-radionuclide S values at the voxel level. Medical Internal Radiation Dose Committee. J Nucl Med 1999;40:11S–36S.
39. Seo Y, et al. Correction of photon attenuation and collimator response for a body-contouring SPECT/CT imaging system. J NucL Med 2005;46:868–877.
40. Boucek JA, Turner JH. Validation of prospective whole-body bone marrow dosimetry by SPECT/CT multimodality imaging in I-131-anti-CD20 rituximab radioimmunotherapy of non-Hodgkin's lymphoma. Eur J Nucl Med Mol Imaging 2005;32: 458–469.
41. Coleman RE, et al. Concurrent PET/CT with an intergrated imaging system: Intersociety dialogue from the joint working group of the American College of Radiology, the Society of Nuclear Medicine, and the Society of Computed Body Tomography and Magnetic Resonance. J Nucl Med 2005;46:1225–1239.
42. Mawlawi O, et al. Performance characteristics of a newly developed PET/CT scanner using NEMA standards in 2D and 3D modes. J Nucl Med 2004;45:1734–1742.
43. Keidar Z, Israel O, Krausz Y. SPECT/CT in tumor imaging: technical aspects and clinical applications. Sem Nucl Med 2003;33:205–218.
44. Otte A, et al. Yttrium-90 DOTATOC: first clinical results. Eur J Nucl Med 1999;26:1439–1447.
45. Bodei L, et al. Receptor-mediated radionuclide therapy with 90Y-DOTATOC in association with amino acid infusion: a phase I study. Eur J Nucl Med Mol Imaging 2003;30:207–216.
46. Otte A, Cybulla M, Weiner SM. ^{90}Y-DOTATOC and nephrotoxicity. Eur J Nucl Med Mol Imaging 2002;29:1543.
47. Jamar F, et al. (86Y-DOTAA0)-D-Phe1-Tyr3-octreotide (SMT487): a phase I clinical study-pharmacokinetics, biodistribution and renal protective effect of different regimes of amino acid co-infusion. Eur J Nucl Med Mol Imaging 2003;30: 510–518.
48. Clarke SEM. Radionuclide therapy of the thyroid. Eur J Nucl Med 1991;18:984–991.
49. Howarth D, et al. Determination of the optimal minimum radioiodine dose in patients with Graves'disease: a clinical outcome study. Eur J Nucl Med 2001;28:1489–1495.
50. Guhlmann CA, Rendl J, Borner W. Radioiodine therapy of autonomously functioning thyroid nodules and Graves'disease. Nuklearmedizin 1995;34:20–23.
51. Willemsen UF, et al. Functional results of radioiodine therapy with a 300 Gy absorbed dose in Graves'disease. Eur J Nucl Med 1993;20:1051–1055.
52. Silver S. Radioactive nuclides in medicine and biology. Philadelphia: Lea & Febiger; 1968.
53. Coakley AJ, Thyroid stunning. Eur J Nucl Med 1998;25:203–204.
54. van Isselt JW, et al. Comparison of methods for thyroid volume estimation in patients with Graves'disease. Eur J Nucl Med 2003;30:525–531.
55. Crawford DC, et al. Thyroid volume measurement in thyrotoxic patients: comparison between ultrasonography and iodine-124 positron emission tomography. Eur J Nucl Med 1997;24:1470–1478.
56. Bockisch A, Jamitzky T, Derwanz R, Biersack HJ. Optimized dose planning of radioiodine therapy of benign thyroidal diseases. J Nucl Med 1993;34:1632–1638.
57. Kalinyak JE, McDougall IR. Editorial: How should the dose of iodine-131 be determined in the treatment of Graves' hyperthyroidism ? J Clin Endocrinol Metab 2003;88:975–977.
58. Leslie WD, et al. A randomized comparison of radioiodine doses in Graves'hyperthyroidism. J Clin Endocrinol Metab 2003;88:978–983.

59. Thomas SR. Options for radionuclide therapy: from fixed activity to patient-specific treatment planning. Cancer Bioth Radiopharm 2002;17:71–81.
60. Maxon HR, Thomas SR, Hertzberg VS. Relation between effective radiation dose and outcome of radioiodine therapy for thyroid cancer. N Engl J Med 1983;309:937–941.
61. Maxon HR, Englaro EE, Thomas SR. Radioiodine-131 therapy for well differentiated thyroid cancer- a quantitative radiation dosimetric approach: outcome and validation in 85 patients. J Nucl Med 1992;33:1132–1136
62. Dorn R, et al. Dosimetry guided radioactive iodine treatment in patients with metastatic thyroid cancer: largest safe dose using a risk-adapted approach. J Nucl Med 2003;44:451–456
63. Eschmann SM, et al. Evaluation of dosimetry of radioiodine therapy in benign and malignant thyroid disorders by means of iodine-124 and PET. Eur J Nucl Med 2002;29:760–767.
64. Benua R, Cical N, Sonenberg M, Rawson R. The relation of radioiodine dosimetry to results and complications in the treatment of metastatic thyroid cancer. Am J Roentgenol 1962;87:171–179.
65. de Keizer B, et al. Bone marrow dosimetry and safety of high I-131 activities given after recombinant human thyroid-stimulating hormone to treat metastatic differentiated thyroid cancer. J Nucl Med 2004;45:1549–1554.
66. Pacini F, et al. Testicular function in patients with differentiated thyroid carcinoma treated with radioiodine. J Nucl Med 1994;35:1418–1422.
67. Ceccarelli C, et al. ^{131}I therapy for differentiated thyroid cancer leads to an earlier onset of menopause: results of a retrospective study. J Clin Endocrinol Metab 2001;86:3512–3515.
68. Schlumberger M, et al. Exposure to radioactive iodine-131 for scintigraphy or therapy does not preclude pregnancy in thyroid cancer patients. J Nucl Med 1996;37:606–612.
69. Young JL Jr, Miller RW. Incidence of malignant tumors in US children. J Pediatr. 1975;86:254.
70. Hoefnagel CA, Voute PA, De Kraker J, Marcuse HR. Radionuclide diagnosis and therapy of neural crest tumors using iodine-131-Metaiodobenzylguanidine. J Nucl Med 1987;28:308–314.
71. Hoefnagel CA. Radionuclide therapy revisited. Eur J Nucl Med 1991;18:408–431.
72. Monsieurs M, et al. Patient dosimetry for neuroendocrine tumors based on ^{123}I-MIBG pre-therapy scans and ^{131}I-MIBG post therapy scans. Eur J Nucl Med 2002;29(12):1581–1587.
73. Flux GD, et al. Estimation and implications of random errors in whole-body dosimetry for targeted radionuclide therapy. Phys Med Biol 2002;47:3211–3223.
74. Matthay KK, et al. Correlation of tumor and whole-body dosimetry with tumor response and toxicity in refractory neuroblastoma treated with ^{131}I-MIBG. J Nucl Med 2001;42:1713–1721.
75. Gaze MN, et al. Feasibility of dosimetry-based high dose ^{131}I-meta-iodobenzylguanidine with topotecan as a radiosensitizer in children with metastatic neuroblastoma. Cancer Biother Radiopharm 2005;20:195–199.
76. Wahl RL. The clinical importance of dosimetry in radioimmunotherapy with tositumomab and iodine I-131 tositumomab. Semin Oncol 2003;30:31–38
77. Wiseman GA, et al. Radiation dosimetry results and safety correlations from ^{90}Y-Ibritumomab Tiuxetan radioimmunotherapy for relapsed or refractory non-Hodgkin's lymphoma: combined data from 4 clinical trials. J Nucl Med 2003;44:465–474.
78. DeNardo GL, Hartmann Siantar CL, DeNardo SJ. Radiation Dosimetry for Radionuclide Therapy in a Nonmyeloablative Strategy. Cancer Biother Radiopharm 2002;17(1):107–118.
79. Sgouros G. Bone marrow dosimetry for radioimmunotherapy: theoretical considerations. J Nucl Med 1993;34:689–694.
80. Shen S, Denardo GL, Sgouros G, O'Donnell RT, DeNardo SJ. Practical determination of patient-specific marrow dose using radioactivity concentration in blood and body. J Nucl Med 1999;40:2102–2106.
81. Sgouros G, Stabin M, Erdi Y. Red marrow dosimetry for radiolabeled antibodies that bind to marrow, bone or blood components. Med Phys 2000;27:2150–2164
82. Stabin MG, Siegel JA, Sparks RB. Sensitivity of model-based calculations of red marrow dosimetry to changes in patient-specific parameters. Cancer Biother Radiopharm 2002;17:535–543.
83. Behr TM, Behe M, Sgouros G. Correlation of red marrow radiation dosimetry with myelotoxicity: empirical factors influencing the radiation-induced myelotoxicity of radiolabeled antibodies, fragments and peptides in pre-clinical and clinical settings. Cancer Biother Radiopharm 2002;17:445–464.
84. Shen S, Meredith RF, Duan J, Brezovich I, Khazaeli MB, Lobuglio AF. Comparison of methods for predicting myelotoxicity for non-marrow targeting I-131-antibody therapy. Cancer Biother Radiopharm 2003;18:209–215.
85. Shen S, Meredith RF. Clinically useful marrow dosimetry for targeted radionuclide therapy. Cancer Biother Radiopharm 2005;20:119–122.
86. Stabin MG, et al. Evolution and status of bone and marrow dose models. Cancer Biother Radiopharm 2002;17:427–433.
87. Sgouros G, et al. Patient-specific, 3-dimensional dosimetry in non-Hodgkin's lymphoma patients treated with ^{131}I-anti-B1 antibody: assessment of tumor dose-response. J Nucl Med 2003;44:260-268.
88. Koral KF, et al. Volume reduction versus radiation dose for tumors in previously untreated lymphoma patients who received iodine-131 tositumomab therapy. Conjugate views compared with the hybrid method. Cancer 2002;94:1258–1263.
89. Sharkey RM, et al. Radioimmunotherapy of non-Hodgkin's lymphoma with ^{90}Y-DOTA humanized anti-CD22 IgG (^{90}Y-Epratuzumab): do tumor targeting and dosimetry predict therapeutic response? J Nucl Med 2003;44:2000–2018.
90. Lewi PJ, Marsboom RP. Toxicology reference data Wistar rat. Amsterdam-Holland: Elsevier/North Holland Biomedical Press; 1981.

See also NUCLEAR MEDICINE, COMPUTERS IN; PHARMACOKINETICS AND PHARMACODYNAMICS.

RADIOSURGERY, STEREOTACTIC

THOMAS H. WAGNER
SANFORD L. MEEKS
M. D. Anderson Cancer Center Orlando
Orlando, Florida

FRANK J. BOVA
University of Florida
Gainesville, Florida

INTRODUCTION

Conventional external beam radiotherapy, or teletherapy, involves the administration of radiation absorbed dose to

cure disease. The general teletherapy paradigm is to irradiate the gross lesion plus an additional volume suspected of containing microscopic disease not visible through physical examination or imaging, to a uniform dose level. External photon beams with peak photon energy in excess of 1 MeV are targeted upon the lesion site by registering external anatomy and internal radiographic anatomy to the radiation (beam) source.

Due to uncertainty and errors in positioning the patient, the radiation beam, which is directed at the lesion, may need to be enlarged to ensure that errors and uncertainty in patient positioning do not cause the radiation beam to miss some or all of the target. Unfortunately, enlarging the radiation beam results in a relatively large volume of nondiseased tissue receiving a significant radiation dose in addition to the target. For example, expansion of a 24 mm diameter spherical target volume to 26 mm to ensure that the target is fully irradiated in the presence of a 2 mm positional error will increase the irradiated volume by 60% (1). As a consequence, non-cancerous (normal) tissue in the expansion region will receive the same high dose that the target will receive. To minimize the normal tissue toxicity, the total radiation dose is delivered in many small increments (fractions), a principle first discovered by Bergonie and Tribondeau in the early twentieth century (2) and used routinely for the majority of external radiotherapy treatments.

In contrast to conventional, fractionated radiotherapy, stereotactic radiosurgery (SRS) involves the spatially precise and conformal administration of a relatively large, single dose of radiation (10–20 Gy) to a small volume of disease, thereby abandoning the advantages provided by fractionation. Hence, it is imperative to minimize the amount of normal tissue irradiated to a high dose using such an approach. Radiosurgery is commonly used to treat intracranial lesions including brain metastases, arteriovenous malformations, benign brain tumors (acoustic schwannoma, meningioma), and primary malignant brain tumors (astrocytoma, glioma, glioblastoma).

Leksel, first conceived radiosurgery for intracranial targeting in 1950 (3). His initial goal was to produce a lesion similar to one created by a radiofrequency probe but without the need to physically introduce a probe into the brain. The lesion was to be created by a very concentrated single high dose of radiation. Stereotactic targeting and arc-centered stereotaxis methods were already known to Leksell. In his initial design, Leksell mounted a therapeutic X-ray tube onto an arc-centered frame with the axis of rotation positioned in the target tissues. This approach allowed many different paths of radiation to converge on the target tissues producing a highly concentrated dose at the intersection point. While this concept did provide a concentrated dose, Leksell continued to investigate alternate radiation delivery systems in hopes of finding a better system that could produce a high concentration of radiation at the target tissues while providing more normal tissue sparing. In later designs Leksell attempted to use particle beams to take advantage of the known Bragg peak effect. The physical limitations of the particle beam delivery systems as well as the expense of the device encouraged Leksell to continue his development finally arriving at a design based on 201 pencil thin cobalt-60 gamma radiation sources arranged on a hemisphere and focused at a single point. This device, known as the Gamma Knife (Elekta Oncology Systems), was used to treat both benign and malignant intracranial targets.

In the 1980s, several groups began to develop technology that would adapt more generally available medical linear accelerators (linacs), to deliver radiosurgical style dose distributions, thereby placing radiosurgical capabilities within the reach of many radiation therapy clinics. Betti (4) and Columbo (5) both developed linear accelerator based radiosurgery systems. Although these early linac-based systems did allow the concentration of radiation dose there existed a question as to how accurately the radiation dose could be delivered to the targeted stereotactic coordinates. Winston and Lutz addressed this issue through the use of a stereotactically positioned phantom target system (6,7). It was found that linac-based systems could maintain the accuracy of radiation beam to target coordinates to within a millimeter or two (Fig. 1 and 2). While some felt that this accuracy was adequate, it fell short of the GammaKnife claim of 0.3 mm isocentric beam accuracy. In the late 1980s Friedman and Bova (8) developed an isocentric subsystem that enabled a routine linac to achieve an

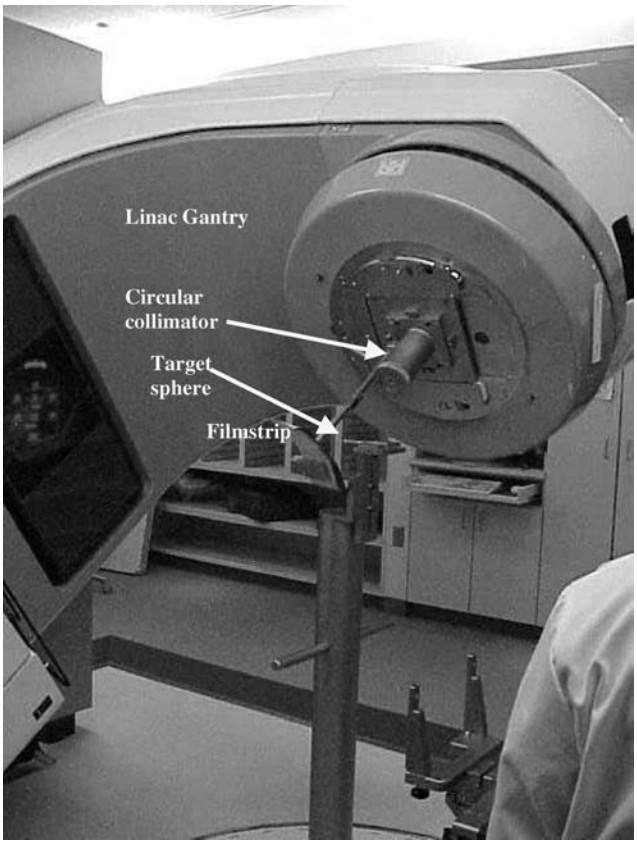


Figure 1. Modified Winston–Lutz test setup, for testing the spatial alignment of the circular radiosurgery X-ray beam with a spherical target ball. The test target sphere should be precisely aligned to the center of the X-ray field defined by the circular collimator. Several film exposures at different linac gantry rotation angles are taken.

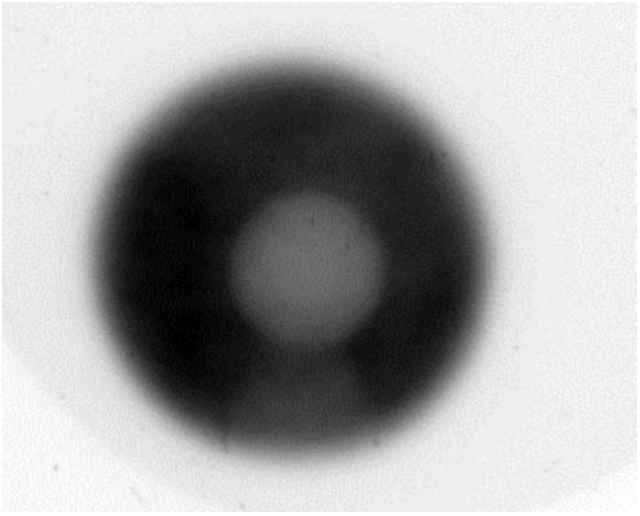


Figure 2. Image of a developed film from the modified Winston–Lutz test, showing the alignment between the circular radiosurgery X-ray beam with a spherical target ball. Analysis of this film image shows that in this case, the target sphere is misaligned from the center of the 20 mm diameter X-ray field by 1.06 mm in one direction, and by 0.97 mm in the other direction. Repeating this test with at least one pair of linac gantry angles gives an estimate of the spatial error inherent in the treatment delivery system.

isocentric beam accuracy of 0.2 mm, thereby matching the accuracy of the GammaKnife.

All early radiosurgery systems used a similar delivery method, namely multiple circular cross-section radiation beams converging to a common point, called the isocenter, located in the center of the target, volume with the directions chosen to minimize the overlap of beams outside the target and hence the normal tissue dose. This scheme worked well for spherical targets using a single isocenter. Non-spherical targets required the use multiple circular collimator diameters focused at multiple isocenters distributed throughout the target volume in an effort to "fill" the volume with dose. The ability to properly select the optimal set of sphere as well as their spacing and weighting were provided by treatment planning systems specially designed to optimize radiosurgical planning.

The next level of advance occurred in the 1990s when new computer controlled collimation devices known as multileaf collimators (MLC) were introduced that were coupled to stereotactic treatment planning software and delivery hardware designed for these new devices. While the early attempts at using MLCs had problems matching the dose conformality and steep dose gradients achieved by multiple isocentric circular collimation techniques, they nevertheless allowed complex targets to be treated more rapidly. Recently, new techniques have been developed that allow both the conformality of multiple isocenters as well as the speed of MLC delivery (9,10).

The majority of medical linear accelerators use microwave radiation in the S band to accelerate electrons and produce X rays. During the 1990s a compact medical X band linear accelerator was developed by Accuray, Inc. Adler placed this X ray source on an industrial robot gantry to create a novel stereotactic radiosurgery system called the CyberKnife (11,12). This system based its stereotactic targeting on an integrated orthogonal X-ray system that performed real-time imaging and correction of beam orientation to compensate for patient motion during treatment. Although this method of targeting was novel in the early 1990s, targeting systems have since been introduced for use on S band medical linacs. Unlike the GammaKnife, which by design is limited to intracranial targets, the CyberKnife can be applied to targets anywhere in the body.

Although radiosurgery based on gamma and X ray sources predominate, the theoretical advantages of proton therapy beams have stimulated great interest in advancing the use of protons to treat intracranial tumors. The physics of proton beam interactions are quite different than those of a photon beam. Unlike X rays, protons have mass and charge that result in a finite range of penetration. Additionally, the density of ionization (linear energy transfer) along the track of a proton beam is greater than that of an X-ray beam, with a region of high ionization density at the end of the track known as a Bragg peak (13,14). The finite range of penetration results in zero radiation dose beyond the Bragg peak, which in theory further allows the concentration of radiation dose to a deep-seated target while sparing underlying radiosensitive structures (15–17). The theoretical advantage of the proton beam Bragg peak is somewhat tempered by practical issue that its width is usually not large enough to encompass an entire radiosurgery target, so that it becomes necessary to superimpose a multitude of proton beams of varying energies to produce a composite depth dose distribution that covers the entire target (16). Historically, proton facilities produced only stationary beams, making it very difficult to bring multiple converging beams upon the patient's lesion. More recently rotating gantry delivery systems for protons have been introduced, which offer more flexibility in selecting beam orientations (18). Nevertheless the high cost (>$50M) of these facilities currently limits their availability to a few large metropolitan centers.

Early stereotactic targeting systems relied upon orthogonal radiographs for target localization, however, stereotactic procedures did not gain wide acceptance until the late 1980s with the development of three-dimensional (3D) treatment planning based on the use of computerized tomography (CT) imaging to obtain a 3D model of the patient. By the 1990s CT based target definition was augmented with magnetic resonance (MR) imaging that provided superior anatomical definition of the central nervous system. Because of difficulties related to MR incompatibility of stereotactic head fixation systems, MR imaging is performed without such hardware and the image set aligned or fused with CT images obtained with head fixation.

Initially, all intracranial stereotactic procedures used a rigid stereotactic head ring, or frame, screwed into the patient's skull to achieve a rigid, reproducible geometry for CT imaging and treatment. While frame-based procedures are still the most precise radiosurgery method they obviously are invasive to the patient and place a time limit on the completion of the procedure within hours of the frame placement. Noninvasive frameless head fixation systems were subsequently developed to address these

issues. While less precise then ring-based approaches, they can be used in certain radiosurgery procedures where extreme accuracy is not required, such as treatment of brain metastases that are not located near critical structures like the brain stem or optical apparatus. Some of these systems are based upon the fitting of patients with thermoplastic face masks (19), while other systems separate the fixation and localization processes through the used of biteplates and thermoplastic masks (20,21).

Extracranial stereotactic targeting of lesions outside of the skull has been made possible through the development of a number of new technologies. One of the first was the use of ultrasound to allow the clinician to obtain a two-dimensional (2D) or 3D image set of the patient in position for radiosurgery (22). The ultrasound probe is tracked during image acquisition and the image voxels are mapped to a reference that allows precise targeting of the radiation beam. To enhance the ability to target tissues these scans are often registered to pretreatment CT and MR scans. Other methods involving fixed stereotactic X-ray tubes with image intensifiers have been developed by Accuray (23) and BrainLab (24). These systems function by obtaining either orthogonal or stereoscopic radiographs that are registered to the projection of a previously obtained 3D CT dataset. While these planar X-ray localization methods work well for bony anatomy the poorer contrast of soft tissues makes them less useful for localizing targets that are not rigidly fixed to bone. More recently the development of large format amorphous silicon detectors has facilitated the development and integration of cone beam CT scanning systems onto medical linear accelerators allowing stereotactic localization and registration of soft tissue anatomy to the linear accelerator's reference coordinate system. These new units have promise of providing unprecedented targeting accuracy to extra cranial targets.

STEREOTACTIC IMAGING AND LOCALIZATION

The uncertainty inherent to the imaging modality used can be the largest source of uncertainty the radiosurgery process. Poor imaging techniques increase this uncertainty and nullify the efforts to improve accuracy in treatment planning and delivery. Therefore, it is important to understand stereotactic imaging techniques, the increased quality assurance demands that are placed on the diagnostic imaging apparatus used, and the inherent limitations associated with each modality. Following are brief explanations of the three stereotactic imaging techniques used in radiosurgery: computed tomography, magnetic resonance imaging, and angiography.

Computed tomography is the primary modality used for radiosurgery treatment planning due to its spatial accuracy and electron density information that are both useful for accurate dose calculation and targeting. Stereotactic CT images can be obtained with a CT-compatible localizer attached to the stereotactic head ring such as the Brown–Roberts–Wells (BRW) (Fig. 3) or other commercially available designs (25,26). Since the geometry of the localizer is known relative to the head ring, stereotactic coordinates of any point in a volumetric CT image set may be

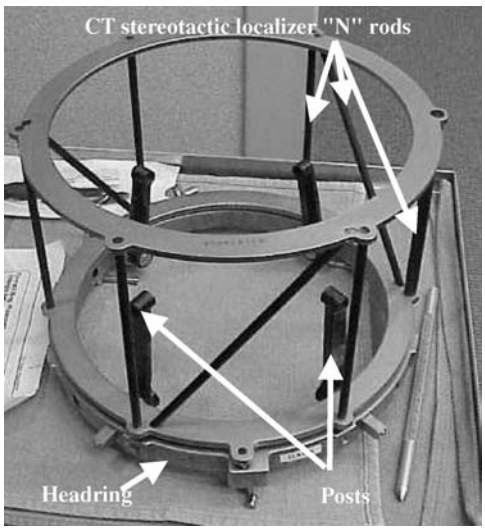


Figure 3. Computed tomography localizer attached to frame.

accurately calculated, using the localizer fiducial markers in each axial image (Fig. 4). For example, the characteristic N-shaped rods of the BRW localizer allow the x–y–z coordinates of any point in space to be mathematically determined relative to the head ring rather than relying on the CT coordinates. This method provides more accurate spatial localization, and minimizes the CT scanner quality assurance requirements. In order to minimize the inaccuracies associated with the stereotactic imaging, it is important to obtain all imaging studies with the best available spatial resolution. Typically this means reducing the uncertainty to < 1 mm by acquiring CT images at an

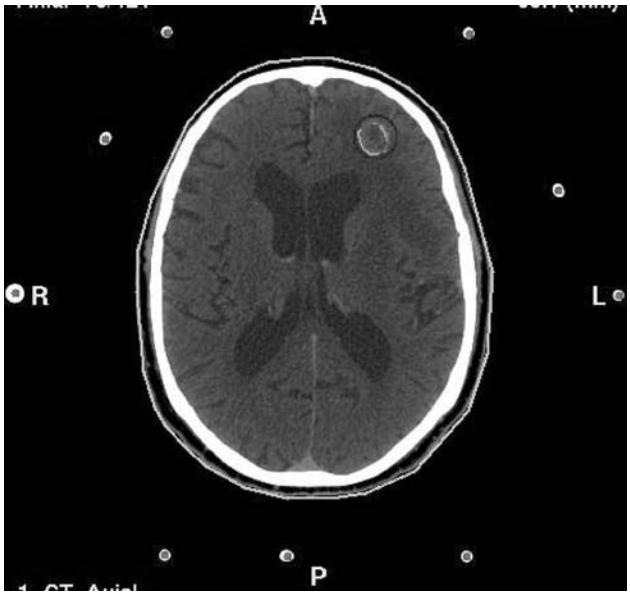


Figure 4. Axial CT image of patient with BRW stereotactic headframe and CT localizer attached. There are three sets of N-shaped rods; the location of any point in a CT image that contains all nine CT localizer rods, can be computed using the interrod distances to precisely define the plane of the image with respect to the BRW headframe and its coordinate system.

image resolution and slice spacing less than this amount. For example, a minimum 34.5 cm field of view is just large enough to image all of the stereotactic fiducials of a BRW localizer. This FOV corresponds to a pixel size of 0.67 mm for a 512 × 512 image matrix. In addition, current multislice diagnostic helical CT scanners can obtain CT images at 0.5–1 mm splice spacing.

Magnetic resonance (MR) imaging often provides superior tumor visualization, but spatial distortion inherent in the MR images due to magnetic field non-uniformities and patient-specific artifacts, and secondarily the lack of electron density information makes the use of MR images less desirable than CT images for radiotherapy dose calculations. Introducing a stereotactic frame and localizer into an MR imager will perturb the magnetic field producing image distortions on the order of 0.7–4 mm in each orthogonal plane (axial, sagittal, coronal) of a stereotactic MRI (27,28). Furthermore the size of the stereotactic head frame may be incompatible with the geometry of standard MRI head coil necessitating the use of a larger MRI coil, such as the standard body coil, with a consequent degradation in image quality due to a reduced signal/noise ratio. These problems are overcome by eliminating the head frame during the MR imaging procedure and using image fusion techniques to register the MR image volume to the CT image volume of the patient in the head frame. For frame-based radiosurgery, the 3D volumetric MR scan is acquired prior to head ring placement using a pulse sequence that allows a fast image acquisition to minimize image distortion due to patient movement. All currently available image correlation routines consider the MR images as rigid bodies, and do not remove local image distortions that can exist in the MR data, hence careful review of the coregistered MRI and CT image sets is essential. This comparison should focus on internal anatomy, such as the ventricles, tentorium, sulci. and avoid the external contour since it can be shifted 3–4 mm due to the fat shift (a distortion resulting in the difference in the resonant frequency of protons in fat relative to their resonant frequency in water).

The third imaging modality important to radiosurgery, angiography, is used for diagnosis and anatomic characterization of cerebral arteriovenous malformations (AVMs). Unlike volumetric CT and MR tomography, stereotactic planar angiography utilizes a set of orthogonal radiographs of a special localizer attached to the stereotactic head ring bearing radio-opaque fiducials. The stereotactic coordinates of any point within the localizer may be calculated very accurately since the geometry of the fiducials is known relative to the head ring. The orthogonal film pair is obtained with contrast injected rapidly at the location of the AVM nidus allowing excellent visualization of fine vasculature and fiducials.

The use of orthogonal images as the sole localization method for treatment planning is inadequate for accurately determining the shape, size and location of an arbitrarily shaped AVM nidus (29–31). Furthermore, overlapping structures, such as feeding or draining blood vessels, may obscure the view of the AVM nidus and will result in unnecessary irradiation of normal tissue if these blood vessels are included in the targeted volume. Because of these issues a volumetric CT angiography image dataset (1 mm slice thickness; intravenous contrast infused at a rate of 1 $cm^3 \cdot s^{-1}$) is always acquired in addition to, or in replacement of, stereotactic angiography. The resultant CT images provide an accurate 3D description of the AVM nidus, along with the feeding and draining vessels.

RADIOSURGERY DELIVERY TECHNIQUES

Numerous radiosurgery techniques have been devised based noncoplanar configurations static beams or arcs. The majority of radiosurgery treatments use circular collimators to create spherical regions of high dose. The classic example of a static beam delivery system is the GammaKnife unit, which consists of 201 narrow-beam cobalt-60 sources arranged on a hemisphere. A collimation helmet containing 201 circular collimators, each of the same diameter, is placed between the hemisphere of sources and the patient's head with the collimator's focal point centered on the intracranial target. This produces a spherical dose distribution or shot in GammaKnife parlance. An irregular volume is treated with multiple shots whose diameters are selected based on the available helmet collimators (32). The CyberKnife robotic radiosurgery unit is also used in a similar manner to deliver treatments from fixed beam orientations using a circular collimator.

Alternatively, a conventional medical linear accelerator can be outfitted with a circular collimator and multiple (5–9) noncoplanar arc delivery used to achieve a spherical dose distribution. When used with linear accelerators, the circularly collimated beam is rotated around the target at isocenter by moving the gantry in arc mode while the patient and treatment couch are stationary, producing a para-sagittal beam path around the target. Betti and Derichinsky developed their linac radiosurgery system with a special chair, the Betti chair, which moved the patient in a side to side arc motion under a stationary linac beam, and which produced a set of para-coronal arcs (4). With modern, computer controlled linear accelerators, more complex motions other than these simple arcs are possible. The Montreal technique, which involves synchronized motion of the patient couch and the gantry while the radiation beam is on, is an example of this, producing a baseball seam type of beam path (33). The rationale of using arcs with circular collimators is to concentrate radiation dose upon the target, while spreading the beam entrance and exit doses over a larger volume of nontarget tissue, theoretically reducing the overall dose and toxicity to nontarget tissue.

Radiosurgery based on circular collimators produces a spherical region of high dose with steep falloff, or gradient, that is adequate for spherical targets. Irregular target volumes require the use of multiple spheres, or isocenters, abutted together to conform the dose more closely to the shape of the target so as to minimize nontarget tissue dose (34). A consequence of the multiisocenter approach is that the shape of the total dose distribution is very sensitive to the abutment of the spherical dose distributions due to their steep falloff. For this reason, it is common practice to accept 30% or greater dose variation over the irregular target volumes using circular collimation.

Figure 5. Beam's eye view showing target shape. Instead of constructing a custom block for the continuous shape of the target, at left a MLC (narrow rectangles) approximates the shape of the conformal beam. Each rectangle represents a tungsten leaf which moves left and right across the field of view shown under computer control. In this example, the MLC leaves would remain stationary while the treatment beam is on, providing a dose distribution very similar to a custom block. At center, the position of the MLC (arrow) on the linac gantry is shown; X rays emerge from the MLC-shaped aperture. At right, close-up photograph of the actual MLC, whose leaves are shaped to the field shown in the left image.

The linear accelerator offers additional flexibility in that tertiary computer-controlled multileaf collimators (MLCs) may be used to produce noncircular beams and beams with nonuniform intensity profiles that conform dose distributions more closely to irregular target volumes with less dose non uniformity. The most common type of MLC consists of two banks of opposed high density metal plates, or leaves that can be moved in a plane perpendicular to the beam's direction. The MLC can be rotated with the treatment machine's collimator in order to align the leaves for the best fit to the target's projected shape. The simplest use of an MLC is simply as a functional replacement for custom made beam shaping blocks, in which the rectangular MLC edges are used to approximate a continuous target outline shape (Fig. 5) (35). This field shaping can be used for either static field treatments or for dynamic arcs in which the MLC shape is continually changed to match the beam's-eye-view projection of the target volume. Moss investigated the efficacy of performing radiosurgery treatments with a dynamically conforming MLC in arc mode, and concluded that dynamic arc MLC treatments offered target coverage and normal tissue sparing comparable to that offered by single and multiple isocenter radiosurgery (36). Nedzi (37) showed that even crude beam shaping devices offered some conformal benefit over single isocenter treatments with circular collimators. Since the mid- to late-1990s, the use of miniature multileaf collimators (MLCs with a leaf width projected to isocenter of 5 mm or less) has become increasingly common.

However, the MLC may be used in a more sophisticated fashion to form many different beam shapes of arbitrary size and intensity (by varying the amount of radiation applied through each beam aperture). In this manner, radiation fields with a similar dose profile as a shaped, wedged field may be delivered using only the computer-controlled MLC, shall can also deliver intensity modulated dose profiles similar to those achievable using custom beam compensators, but without the disadvantages of fabrication time or of needing to manually change a physically mounted beam filter between each treatment field (38). Thus, a computer-controlled MLC and treatment machine offer the potential to deliver more sophisticated radiation treatments to each patient with the same time and cost resources available.

The MLC-based solutions are available for both static multiple beams and arc-based delivery. Radionics initially introduced the use of a mini-MLC for defining static beam shapes that conformed to the projected shape of the target volume in the beam's eye view (35,39,40). The device consisted of multiple thin plates, or MLC leaves, that were mechanically clamped together to form an irregular beam shape defined by a plastic template corresponding to the projected shape of the target. Subsequent developments by other vendors added computer-controlled motorization to the leaves so that treatments could be carried out more efficiently. While most mini-micro-MLC implementations were based on static delivery of few fixed beams, NOMOS, Inc. and 3D Line, Inc. developed specialized arc-based intensity-modulated radiosurgery (IMRS) systems. Most, if not all, tertiary MLC vendors have now developed integrated treatment planning systems designed specifically for their MLCs and treatment applications, including IMRS.

The potential for improvement presented by some of these newer and more sophisticated treatment delivery methods has spurred interest in their evaluation relative to the more traditional linac SRS methods of multiple intersecting arcs and circular collimators. These studies are usually conducted by those who have had difficulty achieving the conformality routinely published by those experienced in multiisocenter planning. These comparisons generally demonstrate that for small to medium (up to $\sim 20\ cm^3$) intracranial targets multiple static beams offer conformity with ratios of normal tissue to target tissue treatments in the range of 1.5–2.0, with target dose homogeneity on the order of 10–20%, while offering a more standard radiation therapy treatment planning interface and process (39,41–43). These studies go on to show that static beam IMRT techniques generally performed comparable to or better than static beam plans, usually increasing the dose homogeneity and possibly conformality (44,45).

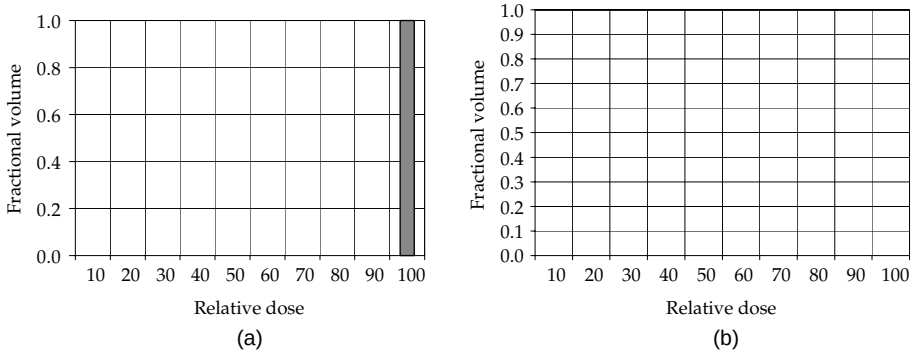


Figure 6. Ideal target (a) and nontarget volume (b) direct DVHs. Note that in the ideal direct DVH of the nontarget volume (right side), the plot is empty, since there is no nontarget volume receiving any dose in the ideal case.

A potential problem with these comparison studies is that they may not equitably compare the full potential of multiple isocenter radiosurgery with circular collimators. A qualitative inspection of the multiple isocenter dosimetric results shown in these comparisons leads one to suspect that in many cases, suboptimal multiple isocenter plans are being compared with reasonably optimized static beam and dynamic MLC arcs–IMRT plans. Although the multiple isocenter treatment plans in these comparisons in the literature may represent a level of plan quality achievable by an average or unfamiliar user, they do not always represent the experience of expert users. Some expert users have reported on the use of multiple isocenter–circular collimator radiosurgery systems to plan and deliver tightly conformal dose distributions to irregularly shaped targets near radiosensitive structures, while maintaining a sharp dose gradient away from the target toward radiosensitive structures (34,46–48).

TOOLS FOR EVALUATING RADIOSURGERY TREATMENT PLANS

The clinical objective of radiosurgery is to deliver a tumorcidal radiation dose to a target volume while minimizing the dose to surrounding tissues. The following tools are available to the treatment planner to evaluate a 3D dose distribution in order to quantify the degree to which this objective is achieved: (*1*) 2D isodose curves and 3D isodose surfaces, (*2*) dose–volume histograms, and (*3*) physical dose–volume figures of merit. The following sections explain the use of each of these tools in radiation therapy and radiosurgery treatment planning.

It is possible to display 3D semitransparent surface renderings of constant dose levels overlaid on 3D renderings of the target volume to determine if the target adequately covered, but these can be difficult to analyze quantitatively. For this reason, 2D cross-sections of the 3D dose distribution are evaluated making it easier to quantitative assess target coverage. The 2D dose cross-sections are displayed as isocontour plots (isodose plots) overlaid on the patient's CT and MR images to allow visual assessment of dose coverage to an accuracy of within one image pixel. Although this implies submillimeter precision, the 1 pixel uncertainty in isodose position can result in a large uncertainty in dose coverage for small intracranial targets. In the case of a 0.67 mm pixel, a 20 mm sphere, equal to 4.2 cm^3, would apparently be equally well covered by an isodose surface ranging in volume from 3.8 to 4.6 cm^3 corresponding to a 10% uncertainty in volume. Hence, although visual inspection of isodose plots on multiple images is commonly performed, it is cumbersome and there is a large uncertainty in assessing the dose coverage that is associated representing small targets using finite size pixels.

One commonly used solution to this problem is to use dose–volume histograms (DVHs). DVHs are a method of condensing 3D dose information into a more manageable form for analysis. The simplest type of DVH is a differential histogram of volume versus dose (49). This is simply a histogram showing the number of occurrences of each dose value within a 3D volume. A second more common representation is the cumulative DVH, which is the integral of the differential DVH as a function of dose. Unfortunately, in either type of DVH, the spatial information of which specific volumes are exposed to each dose level is lost in the process of constructing a DVH. For this reason, DVHs are generally used clinically in conjunction with the evaluation of multiple isodose plots as mentioned earlier.

The ideal treatment planning situation is one in which the target volume receives a uniform dose equal to the maximum dose, and the nontarget volume receives zero dose. This would correspond to ideal differential DVHs for target and nontarget volumes as shown in Fig. 6. Clinically realistic differential DVHs for target and nontarget volumes for a more typical (non-ideal) radiosurgery dose distribution are shown in Fig. 7. Figure 8 shows two differential DVHs from competing radiosurgery plans plotted on a common axes to allow a direct comparison of the plans. Note that it can difficult to evaluate competing plans using such differential histograms (50), as demonstrated in Fig. 8. Above $\sim$ 40 units of dose, both plans appear to be identical, but the two plans expose differing volumes of brainstem at doses less than $\sim$ 40 units. For this reason cumulative DVH analysis is more commonplace. Transforming the differential DVHs into cumulative DVHs, by plotting the volume receiving at least a certain dose versus dose, makes it simpler to evaluate the differences in the dose distributions, as shown in Fig. 9.

Optimal cumulative DVH curves for target structures will be as far toward the upper right hand corner of the plot as possible, while the those for nontarget structures will be as close as possible to the lower left hand corner of the

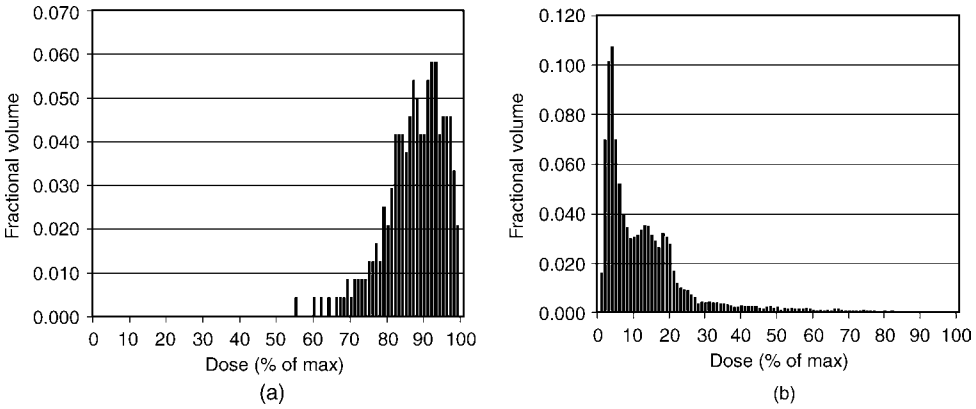


Figure 7. Typical (nonideal) radiosurgery direct DVHs for target volume (a) and nontarget volume (b).

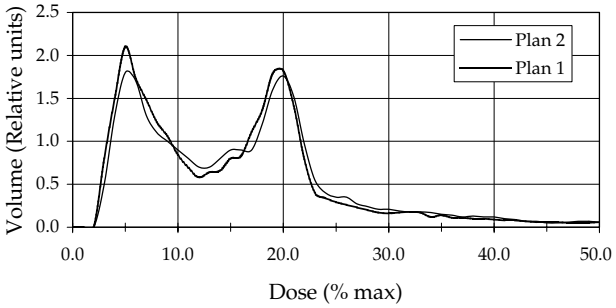


Figure 8. Direct DVHs for a radiosensitive nontarget structure in two hypothetical treatment plans.

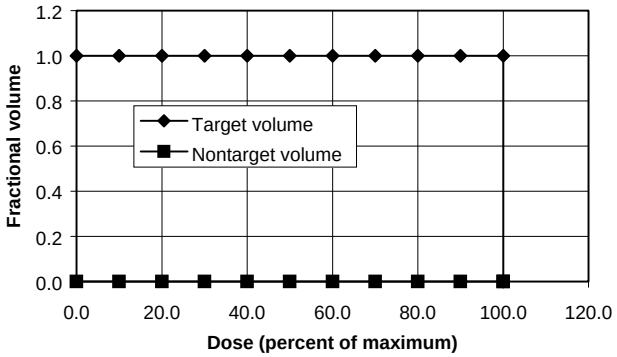


Figure 10. Ideal cumulative DVH curve for target and nontarget volumes.

plot as shown in Fig. 10. Considering the two completing plans shown in Fig. 9, the better plan will have its target cumulative DVH further to the upper right corner and its nontarget cumulative DVH further to the lower left corner than the poorer plan. Plan 1 is the preferred plan, since its nontarget DVH for the brainstem lies below and to the left of that for Plan 2. The relative ease of this comparison underscores the general utility of cumulative DVHs over differential DVHs (49,51). Unfortunately, it is rare for the cumulative DVHs of rival treatment plans to separate themselves from one another so cleanly. Typically, the DVHs cross one another, perhaps more than once as shown in Fig. 11. The simple rules for evaluating DVHs cannot resolve this situation, in which case other means must be used to evaluate the treatment plans by applying a score to each plan derived from a clinically relevant figure of merit.

The three properties of radiosurgery dose distributions that have been correlated with clinical outcome and that lend themselves to clinical figures of merit are (*1*) dose conformity, (*2*) dose gradient, and (*3*) dose homogeneity (34). The conformity of the dose distribution to the target volume may be simply expressed as the ratio of the prescription isodose volume to the target volume, frequently referred to as the PITV ratio (52).

$$\text{PITV} = \text{Prescription isodose volume}/\text{target volume} \quad (1)$$

Perfect conformity of a dose distribution to the target, that is, PITV = 1.00, is typically not achievable, and some

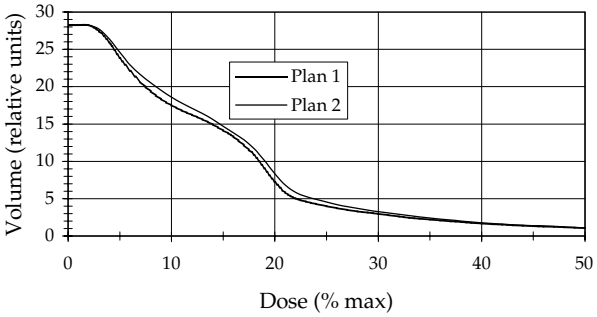


Figure 9. Cumulative DVH plot of the direct DVH data shown in Fig. 8.

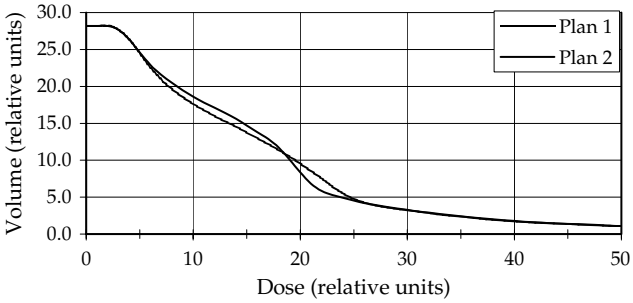


Figure 11. Crossing cumulative DVH curves.

Figure 12. Transaxial, sagittal, and coronal isodose distributions for five arcs of 100° each delivered with a 30 mm collimator. Isodose lines in each plane increase from 10 to 90% in 10% increments, as indicated. The isocenter is marked with cross-hairs.

volume of nontarget tissue must be irradiated to the same dose level as the target, resulting in PITV ratios greater than unity. The most conformal treatment plans are those with the lowest PITVs, if all of the plans under comparison provide equivalent target coverage. This stipulation is necessary because the definition of PITV does not specify how the prescription isodose is determined. It is possible (but undesirable) to lower, and thus improve, the PITV by selecting an isodose level that incompletely covers the target as the prescription isodose, and therefore reduces the numerator of Eq. 1. Many investigators report isodose shells that cover in the neighborhood of at least 95% of the target volume or 99% of the target volume (17,34,46,48,53–57). This ensures a more consistent basis of comparisons for all treatment plans.

A sharp dose gradient (fall off in dose with respect to distance away from the target volume) is an important characteristic of radiosurgery and stereotactic radiotherapy dose distributions. Dose gradient may be characterized by the distance required for the dose to decrease from a therapeutic (prescription) dose level to one at which no ill effects are expected (half prescription dose). For illustrative purposes, a typical dose distribution in a hemispherical water phantom for a single isocenter delivered with five converging arcs and a 30 mm collimator is depicted in Fig. 12. A quantitative measure of gradient is obtained from examining the dose profiles along orthogonal directions in the principal anatomical planes (transaxial, sagittal, and coronal), as shown by cross-plots in Fig. 13. As in this example the steepest dose gradient (4.6 mm) occurs between the 80% and 40% isodose shells, and for this reason single isocenter dose distributions are prescribed to the 80% isodose shell (34). Table 1 lists dose gradient information between the 80 and 40% isodose shells for single isocenter spherically symmetric dose distributions with 10–50 mm diameter collimators.

A method has been proposed that uses easily obtainable DVH information to generate a numerical measure, or score, of the overall dose gradient for evaluating the dose conformality of radiosurgery dose distributions. The Conformity–Gradient Index score, or CGIg, has been proposed as a metric for quantifying dose gradient of a stereotactic treatment plan (58,59). From treatment planning experience at the University of Florida, it has been observed that it is possible to achieve a dose distribution that decreases from the prescription dose level to half of prescription dose in a distance of 3–4 mm away from the target. Taking this as a guide, a gradient score CGIg

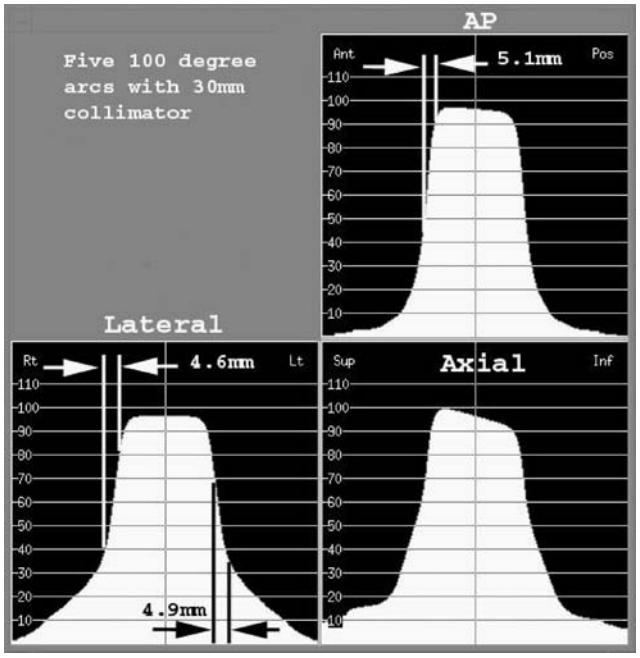


Figure 13. Dose cross-plots through the isocenter, corresponding to the isodose distributions shown in Fig. 12. The sharpest dose fall-off, from dose D to half-dose 0.5D, occurs between dose D of 80% to 0.5D = 40%, which occurs in a distance of 4.6 mm. The D to 0.5D fall-off distance is larger for 90–45% (5.1 mm) and for 70–35% 4.9 mm) doses.

may be computed as

$$\mathrm{CGIg} = 100 - \{100 \times [(R_{\mathrm{eff},50\%\mathrm{Rx}} - R_{\mathrm{eff},\mathrm{Rx}}) - 0.3\,\mathrm{cm}]\} \quad (2)$$

where $R_{\mathrm{eff}50\%\mathrm{Rx}}$ is the effective radius of the half-prescription isodose volume, and $R_{\mathrm{eff}\mathrm{Rx}}$ is the effective radius of the prescription isodose volume. The effective radius of a

Table 1. Single Isocenter (Five Converging Arcs) Dose–Volume and Gradient Information for 10–50 mm Diameter Circular Collimators

Coll.	$V_{80\%}$, cm^3	$R_{\mathrm{eff}80\%}$, mm	$V_{40\%}$, cm^3	$R_{\mathrm{eff}40\%}$, mm	Eff. Gradient, mm	CGIg
10	0.3	4.2	1.2	6.7	2.4	106
20	3.9	9.8	9.7	13.2	3.5	95
30	13.9	14.9	30.8	19.4	4.5	85
50	67.4	25.2	111.6	29.9	4.6	84

volume is the radius of a sphere of the same volume, so that R_{eff} for a volume V is given by

$$R_{\text{eff}} = (3V/4\pi)^{-1/3} \quad (3)$$

The volumes of the prescription isodose shell and the half prescription isodose shell are obtained from a DVH of the total volume (or a sizeable volume that completely encompasses the target volume and a volume that includes all of the half prescription isodose shell) within the patient image dataset. The CGIg score is a dimensionless number that exceeds 100 for dose gradients < 3 mm (steeper falloff from prescription to half-prescription dose level), and which decreases < 100 as a linear function of the effective distance between the prescription and half-prescription isodose shells.

Dose conformity is another important characteristic of a radiosurgery treatment plan that should be considered in plan evaluation. The Conformity-Gradient Index (conformal), or CGIc, is defined as (58):

$$\text{CGIc} = 100 \times (\text{PITV})^{-1} \quad (4)$$

The CGIc converts PITV into a numerical score expressing the degree of conformity of a dose distribution to the target volume. The CGIg score increases as the dose gradient improves, and the CGIc score increases as dose conformity improves. Perfect conformity (assuming the target is adequately covered) of the prescription isodose volume to the target is indicated by a PITV $= 1.00$ and a CGIc $= 100$.

As dose gradient and dose conformity are both important parameters in judging a stereotactic radiosurgery or radiotherapy plan, an overall figure of merit for judging radiosurgery plans should incorporate both of these characteristics. Since clinical data to indicate the relative importance of conformity versus gradient is currently lacking, an index, the Conformity-Gradient Index (CGI) is proposed that assigns equal importance to both of these factors. The overall Conformity-Gradient Index score, or CGI, for a radiosurgery or radiotherapy plan is the average of the CGIc and CGIg scores:

$$\text{CGI} = 0.5 \times (\text{CGIc} + \text{CGIg}) \quad (5)$$

A final measure of plan quality considered by some to be an important factor in evaluating treatment plans is dose homogeneity. While is a homogenous dose is desirable for conventional, fractionated radiotherapy (60), its role is less clear in radiosurgery. Several studies have associated large radiosurgical dose heterogeneity (maximum dose to peripheral dose ratio, or MDPD, > 2.0) with an increased risk of complications (61,62). However, some radiosurgeons have hypothesized that the statistically significant correlation between large dose inhomogeneities and complication risk may be associated with the relatively nonconformal multiple isocenter dose distributions with which some patients in these studies were treated, and not with dose inhomogeneity alone. One theory is that the extreme hot spots associated with large dose heterogenities may be acceptable, if the dose distribution is very conformal to the target volume and the hot spot is contained within the target volume. Nonconformal dose distributions could easily cause the hot spots to occur outside of the target, greatly increasing the risk of a treatment complication. The extensive successful experience of gamma unit treatments administered worldwide (almost all treatments with MDPD ≥ 2.0) lends support to this hypothesis (63). Therefore, as a general principle, one strives for a homogeneous radiosurgery dose distribution, but this is likely not as important a factor as conformity of the high dose region to the target volume, or the dose gradient outside of the target.

SUMMARY

While the use of radiosurgery is now in its fifth decade the basic principles of dose prescription and delivery have changed very little from those first conceived by Leksell. The primary, and still most effective method to treat relatively small target tissues with a high dose and to maintain a very steep dose gradient is through the use of many beams that all converge on the target tissue and diverge along independent paths while approaching and leaving the target region. Other dose targeting and restriction techniques, such as intensity modulation, provide the ability to position beams that geometrically avoid tissue and potentially provide a more powerful tool for dose optimization. These techniques are often combined to allow for the best of both optimized dose planning and efficient dose delivery.

BIBLIOGRAPHY

Cited References

1. Bova FJ, Meeks SL, Friedman WA. Linac Radiosurgery: System Requirements, Procedures and Testing. In: Treatment Planning in Radiation Oncology. Khan FM, Potish RA, editors. Baltimore: Williams and Wilkins; 1998. p 215–241.
2. Hall EJ. Time, dose, and fractionation in radiotherapy. Radiobiology for the Radiologist. Philadelphia: J.B. Lipponcott; 1994. p 211–229.
3. Lindquist C. Gamma Knife Radiosurgery. Semin Radiat Oncol 1995;5(3):197–202.
4. Betti O, Derechinsky V. [Multiple-beam stereotaxic irradiation]. Neurochirurgie 1983;29(4):295–298.
5. Colombo F, et al. External stereotactic irradiation by linear accelerator. Neurosurgery 1985;16(2):154–160.
6. Winston K, Lutz W. Linear Accelerator as a Neurosurgical Tool for Stereotactic Radiosurgery. Neurosurgery 1988;22(3): 454–464.
7. Lutz W, Winston KR, Maleki N. A System for Stereotactic Radiosurgery with a Linear Accelerator. Int J Radiat Oncol Biol Phys 1988;14(2):373–381.
8. Friedman WA, Bova FJ. The University of Florida radiosurgery system. Surg Neurol 1989;32(5):334–342.
9. St John T, et al. Intensity-Modulated Radiosurgery Treatment Planning By Fluence Mapping Multi-isocenter Plans. Med Phys 2001;28(6):1256.
10. St John T, Wagner TH, BFJ, FWA, MSL. A geometrically based method of step and shoot stereotactic radiosurgery with miniature multileaf collimator. Phys Med Biol 2005;50: 3263–3276.
11. Adler JR Jr, et al. The Cyberknife: a frameless robotic system for radiosurgery. Stereotact Funct Neurosurg 1997;69(1–4 Pt. 2): 124–128.

12. Murphy MJ, Cox RS. The accuracy of dose localization for an image-guided frameless radiosurgery system. Med Phys 1996; 23(12):2043–2049.
13. Stanton R, Stinson D. Applied Physics for Radiation Oncology. Madison, (WI): Medical Physics Publishing; 1996. p 366.
14. Moyers MF. Proton Therapy. In The Modern Technology of Radiation Oncology. Van Dyk J, editor. Madison, (WI): Medical Physics Publishing; 1999. p 823–869.
15. Baumert BG, Lomax AJ, Miltchev V, Davis JB. A comparison of dose distributions of proton and photon beams in stereotactic conformal radiotherapy of brain lesions. Int J Radiat Oncol Biol Phys 2001;49(5):1439–1449.
16. Bussiere MR, Adams JA. Treatment planning for conformal proton radiation therapy. Technol Cancer Res Treat 2003;2(5): 389–399.
17. Verhey LJ, Smith V, Serago CF. Comparison of radiosurgery treatment modalities based on physical dose distributions. Int J Radiat Oncol Biol Phys 1998;40(2):497–505.
18. Chapman PH, Loeffler JS. Proton Radiosurgery. In Youman's Neurological Surgery. Winn HR, editor. Philadelphia: Saunders; 2004. p 4123–4130.
19. Willner J, Flentje M, Bratengeier K. CT simulation in stereotactic brain radiotherapy—analysis of isocenter reproducibility with mask fixation. Radiother Oncol 1997;45(1):83–88.
20. Bova FJ, et al. The University of Florida frameless high-precision stereotactic radiotherapy system. Int J Radiat Oncol Biol Phys 1997;38(4):875–882.
21. Meeks SL, et al. IRLED-based patient localization for linac radiosurgery. Int J Radiat Oncol Biol Phys 1998;41(2): 433–439.
22. Meeks SL, et al. Ultrasound-guided extracranial radiosurgery: technique and application. Int J Radiat Oncol Biol Phys 2003; 55(4):1092–1101.
23. Chang SD, et al. An analysis of the accuracy of the CyberKnife: a robotic frameless stereotactic radiosurgical system. Neurosurgery 2003;52(1):140–146; discussion 146–147.
24. Yan H, Yin FF, Kim JH. A phantom study on the positioning accuracy of the Novalis Body system. Med Phys 2003;30(12): 3052–3060.
25. Brown RA. A stereotactic head frame for use with CT body scanners. Invest Radiol 1979;14(4):300–304.
26. Saw CB, Ayyangar K, Suntharalingam N. Coordinate transformations and calculation of the angular and depth parameters for a stereotactic system. Med Phys 1987;14(6):1042–1044.
27. Burchiel KJ, Nguyen TT, Coombs BD, Szumoski J. MRI distortion and stereotactic neurosurgery using the Cosman-Roberts-Wells and Leksell frames. Stereotact Funct Neurosurg 1996;66(1–3):123–136.
28. Kitchen ND, Lemieux L, Thomas DG. Accuracy in frame-based and frameless stereotaxy. Stereotact Funct Neurosurg 1993;61(4):195–206.
29. Spiegelmann R, Friedman WA, Bova FJ. Limitations of angiographic target localization in planning radiosurgical treatment. Neurosurgery 1992;30(4):619–623; discussion 623–624.
30. Bova FJ, Friedman WA. Stereotactic angiography: an inadequate database for radiosurgery? Int J Radiat Oncol Biol Phys 1991;20(4):891–895.
31. Blatt DR, Friedman WA, Bova FJ. Modifications based on computed tomographic imaging in planning the radiosurgical treatment of arteriovenous malformations. Neurosurgery 1993;33(4):588–595; discussion 595–596.
32. Maitz AH, Wu A. Treatment planning of stereotactic convergent gamma-ray irradiation using Co-60 sources. Med Dosim 1998;23(3):169–175.
33. Wasserman TH, Rich KM, Drzymala RE, Simpson JR. Stereotactic irradiation. In: Principles and Practice of Radiation Oncology. Perez CA, Brady LW, editors. Philadelphia: Lippencott-Raven; 1996. p 387–404.
34. Meeks SL, et al. Treatment planning optimization for linear accelerator radiosurgery. Int J Radiat Oncol Biol Phys 1998;41(1): 183–197.
35. Brewster L, et al. Three dimensional conformal treatment planning with multileaf collimators. Int J Radiat Oncol Biol Phys 1995;33(5):1081–1089.
36. Moss DC. Conformal stereotactic radiosurgery with multileaf collimation. In Nuclear Engineering Sciences. Gainesville, (FL): University of Florida; 1992.
37. Nedzi LA, et al. Dynamic field shaping for stereotactic radiosurgery: a modeling study. Int J Radiat Oncol Biol Phys 1993; 25(5):859–869.
38. Sternick ES, Carol MP, Grant W. Intensity-modulated radiotherapy. In: Treatment Planning in Radiation Oncology. Khan FM, Potish RA, editors. Baltimore: Williams and Wilkins; 1998. p 187–213.
39. Shiu AS, et al. Comparison of miniature multileaf collimation (MMLC) with circular collimation for stereotactic treatment. Int J Radiat Oncol Biol Phys 1997;37(3):679–688.
40. Leavitt DD. Beam shaping for SRT/SRS. Med Dosim 1998; 23(3):229–236.
41. Laing RW, et al. Stereotactic radiotherapy of irregular targets: a comparison between static conformal beams and non-coplanar arcs. Radiother Oncol 1993;28(3):241–246.
42. Cardinale RM, et al. A comparison of three stereotactic radiotherapy techniques; ARCS vs. noncoplanar fixed fields vs. intensity modulation. Int J Radiat Oncol Biol Phys 1998; 42(2): 431–436.
43. Hamilton RJ, et al. Comparison of static conformal field with multiple noncoplanar arc techniques for stereotactic radiosurgery or stereotactic radiotherapy. Int J Radiat Oncol Biol Phys 1995;33(5):1221–1228.
44. Woo SY, et al. A comparison of intensity modulated conformal therapy with a conventional external beam stereotactic radiosurgery system for the treatment of single and multiple intracranial lesions. Int J Radiat Oncol Biol Phys 1996;35(3): 593–597.
45. Kramer BA, et al. Dosimetric comparison of stereotactic radiosurgery to intensity modulated radiotherapy. Radiat Oncol Investig 1998;6(1):18–25.
46. Meeks SL, et al. Potential clinical efficacy of intensity-modulated conformal therapy. Int J Radiat Oncol Biol Phys 1998; 40(2):483–495.
47. Meeks SL, et al. Linac scalpel radiosurgery at the University of Florida. Med Dosim 1998;23(3):177–185.
48. Wagner T, et al. A Geometrically Based Method for Automated Radiosurgery Planning. Int J Radiat Oncol Biol Phys 2000; 48(5):1599–1611.
49. Lawrence TS, Kessler ML, Ten Haken RK. Clinical interpretation of dose-volume histograms: the basis for normal tissue preservation and tumor dose escalation. Front Radiat Ther Oncol 1996;29:57–66.
50. Drzymala RE, et al. Dose-volume histograms. Int J Radiat Oncol Biol Phys 1991;21(1):71–78.
51. Kutcher GJ, Jackson A. Treatment plan evaluation. In: Treatment Planning in Radiation Oncology. Khan FM, Potish RA, editors. Baltimore: Williams and Wilkins; 1998 p 281–294.
52. Shaw E, et al. Radiation Therapy Oncology Group: radiosurgery quality assurance guidelines. Int J Radiat Oncol Biol Phys 1993;27(5):1231–1239.
53. Wagner T, et al. Isotropic beam bouquets for shaped beam linear accelerator radiosurgery. Phys Med Biol 2001;46(10): 2571–2586.
54. Wagner TH. Optimal delivery techniques for intracranial stereotactic radiosurgery using circular and multileaf collimators. Nuclear and Radiological Engineering. Gainesville, (FL): University of Florida; 2000. p 306.

55. Tome WA, et al. A high-precision system for conformal intracranial radiotherapy. Int J Radiat Oncol Biol Phys 2000;47(4): 1137–1143.
56. Tome WA, et al. Optically guided intensity modulated radiotherapy. Radiother Oncol 2001;61(1):33–44.
57. Smith V, Verhey L, Serago CF. Comparison of radiosurgery treatment modalities based on complication and control probabilities. Int J Radiat Oncol Biol Phys 1998;40(2):507–513.
58. Wagner TH, et al. A simple and reliable index for scoring rival stereotactic radiosurgery plans. Int J Radiat Oncol Biol Phys 2003;57(4):1141–1149.
59. Bova FJ, Meeks SL, Friedman WA, Buatti JM. Stereotactic plan evaluation tool, the "UF Index". Int J Radiat Oncol Biol Phys 1999;45(3(S)):188.
60. Landberg T, et al.. Prescribing, recording, and reporting photon beam therapy, ICRU Report 50. Bethesda, (MD): International Commission on Radiation Units and Measurements; 1993.
61. Nedzi LA, et al. Variables associated with the development of complications from radiosurgery of intracranial tumors. Int J Radiat Oncol Biol Phys 1991;21(3):591–599.
62. Shaw E, et al. Radiosurgery for the treatment of previously irradiated recurrent primary brain tumors and brain metastases: initial report of radiation therapy oncology group protocol (90-05). Int J Radiat Oncol Biol Phys 1996;34(3):647–654.
63. Flickenger JC, Kondziolka D, Lunsford LD. What is the effect of dose inhomogeneity in radiosurgery?, In: International Stereotactic Radiosurgery Society 3rd Meeting. Kondziolka D, editor. Madrid: Karger; 1997. p 206–213.

See also GAMMA KNIFE; STEREOTACTIC SURGERY.

RADIOTHERAPY ACCESSORIES

JAMES A. PURDY
UC Davis Medical Center
Sacramento, California

INTRODUCTION

The use of radiation to treat cancer is a complex process that involves many trained professionals and a variety of interrelated functions. Radiation oncologists, medical physicists, dosimetrists, and radiation therapists have for many years sought apparatus and devices that aid in this process, particularly in regard to tumor localization, treatment planning, treatment delivery, and verification (1). Precision radiation therapy is necessary as clinical and experimental results show that tumor control and normal tissue response can be a very steep function of radiation dose, and hence, small changes in the dose delivered can result in a dramatic change in the local response of the tumor and/or normal tissues. Moreover, the prescribed curative tumor doses are often, by necessity, close to the doses tolerated by the normal tissues. Thus, for optimum treatment, the radiation dose must be delivered with a high degree of accuracy; a value of $\pm 5\%$ has been recommended by the International Commission on Radiation Units and Measurements (ICRU) (2).

Since the first edition of this encyclopedia, the field of radiation oncology has undergone dramatic changes. At that time, radiation oncologists were trained to plan and treat patients using what has been labeled a two-dimensional (2D) approach. This approach emphasizes the use of a conventional X-ray simulator for designing beam portals that are based on standardized beam arrangement techniques and the use of bony landmarks visualized on planar radiographs. This approach, while still used by some clinics, has been largely replaced by a three-dimensional (3D) approach in modern day radiotherapy clinics. This was made possible by the introduction of commercial 3D treatment planning systems in the early 1990s (3). In contrast to the 2D method, 3D treatment planning emphasizes an image-based virtual simulation approach for defining tumor volumes and critical organs at risk for the individual patient (4). The new 3D process puts new demands on the radiation oncologist to specify target volumes and organs at risk with far greater accuracy than before, and also on the medical physicist to provide effective quality assurance (QA) processes to ensure safe use of the new image-based planning and computer controlled treatment delivery approach (5).

This article presents a review of the devices that have been designed to help achieve the high degree of precision and accuracy needed in the radiation treatment (for both the 2D and 3D approaches) of the cancer patient. These devices have been arranged in the following general categories: tumor localization and treatment simulation devices, patient setup and restraint–repositioning devices, field-shaping devices–dose-modifying devices, and treatment verification and quality assurance devices.

TUMOR LOCALIZATION AND TREATMENT SIMULATION DEVICES

Devices and apparatus in this category are designed to aid in visualizing and determining the extent of the tumor in relation to the treatment geometry (target volume localization) and to obtain measurements of the patient's body contours and thicknesses. In the past, target volume localization was usually accomplished by physical examination and the use of a device called an X-ray simulator, which combines radiographic and fluoroscopic capability in a single machine that mimics the actual treatment unit geometries (Fig. 1). The simulation process itself may be supplemented with other diagnostic imaging studies including computed tomography (CT), magnetic resonance images (MRI), and, more recently, positron emission tomography (PET).

Devices used to aid the conventional simulation process include a radiopaque fiducial grid (Fig. 2) projected on the patient's anatomy, which allows one to determine dimensions of the treatment volume from the simulator plane films. Examples of other devices used in the 2D target volume localization process are magnification rings placed in the irradiated field and lead-tipped rods that can be inserted into body openings, such as the vagina for carcinoma of the cervix or into the rectum for carcinoma of the prostate. The lead tip can be visualized clearly on simulator films or treatment portal films and allows evaluation of treatment field margins.

Note that a new generation of conventional simulators (Fig. 3) has recently been developed in which the image intensifier system has been replaced with an amorphous silicon flat-panel detector. This device produces high resolution, distortion-free digital images, including cone-beam CT.

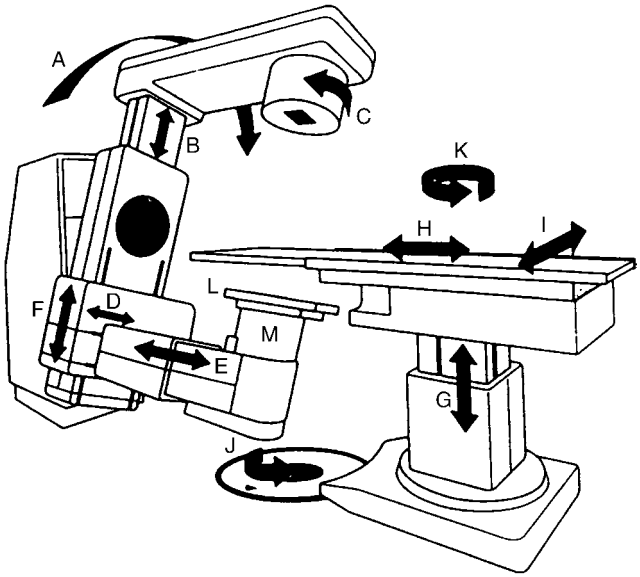


Figure 1. The basic components and motions of a radiation therapy simulator: A, gantry rotation; B, source-axis distance; C, collimator rotation; D, image intensifier (lateral); E, image intensifier (longitudinal); F, image intensifier (radial); G, patient table (vertical); H, patient table (longitudinal); I, patient table (lateral); J, patient table rotation about isocenter; K, patient table rotation about pedestal; L, film cassette; M, image intensifier. Motions not shown include field size delineation, radiation beam diaphragms, and source-tray distance. (See Ref. 6.)

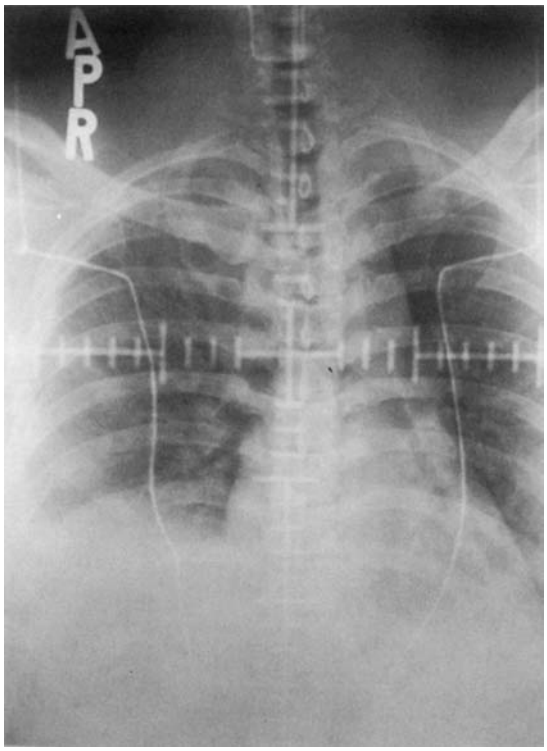


Figure 2. X-ray simulator radiograph showing fiducial grid projected on patient's anatomy. The grid is used for localizing target volume and determining treatment field size.

Figure 3. New generation radiation therapy simulator, in which image intensifier system has been replaced with amorphous silicon flat-panel that produces high resolution, distortion-free images and facilitates a filmless department. (Courtesy of Varian Medical Systems.)

Once the treatment geometry has been determined and the patient is in treatment position, the patient's body thicknesses and contours are measured and recorded for purposes of computing a dose distribution and determining treatment machine settings. Manual methods using calipers, lead solder wire, plaster cast strips, flexible curves, or other devices, such as the contour plotter (see Fig. 4) are the most common methods of obtaining this type of data when using the 2D planning approach.

Fields to be treated are typically delineated in the 2D simulation process using either visible skin markings or marks on the skin visible only under an ultraviolet (UV) light. Some institutions prefer to mark only reference setup points using external tattoos. These skin markings are used to reposition the patient on the treatment machine using the treatment machine's field localization light and optical distance indicator and laser alignment lights mounted in the treatment room that project transverse, coronal, and sagittal light lines on the patient's skin surface (Fig. 5).

In the new 3D era, CT simulators have become the standard of practice; a typical CT simulator facility design is shown in Fig. 6. A volumetric set of CT images is used to define the target volume, critical organs at risk, and skin surface. The CT numbers can be correlated with the electron densities of the tissues imaged to account for heterogeneities when calculating the dose distribution. Numerous studies have documented the improvements in target volume localization and dose distributions achieved with anatomic data obtained from CT scans as compared with the conventional simulation process (7,8). The CT simulators have advanced software features for image manipulation and viewing such as beam's eye view (BEV) display, and virtual simulation tools for setting isocenter, and digital reconstructed radiographs (DRRs)

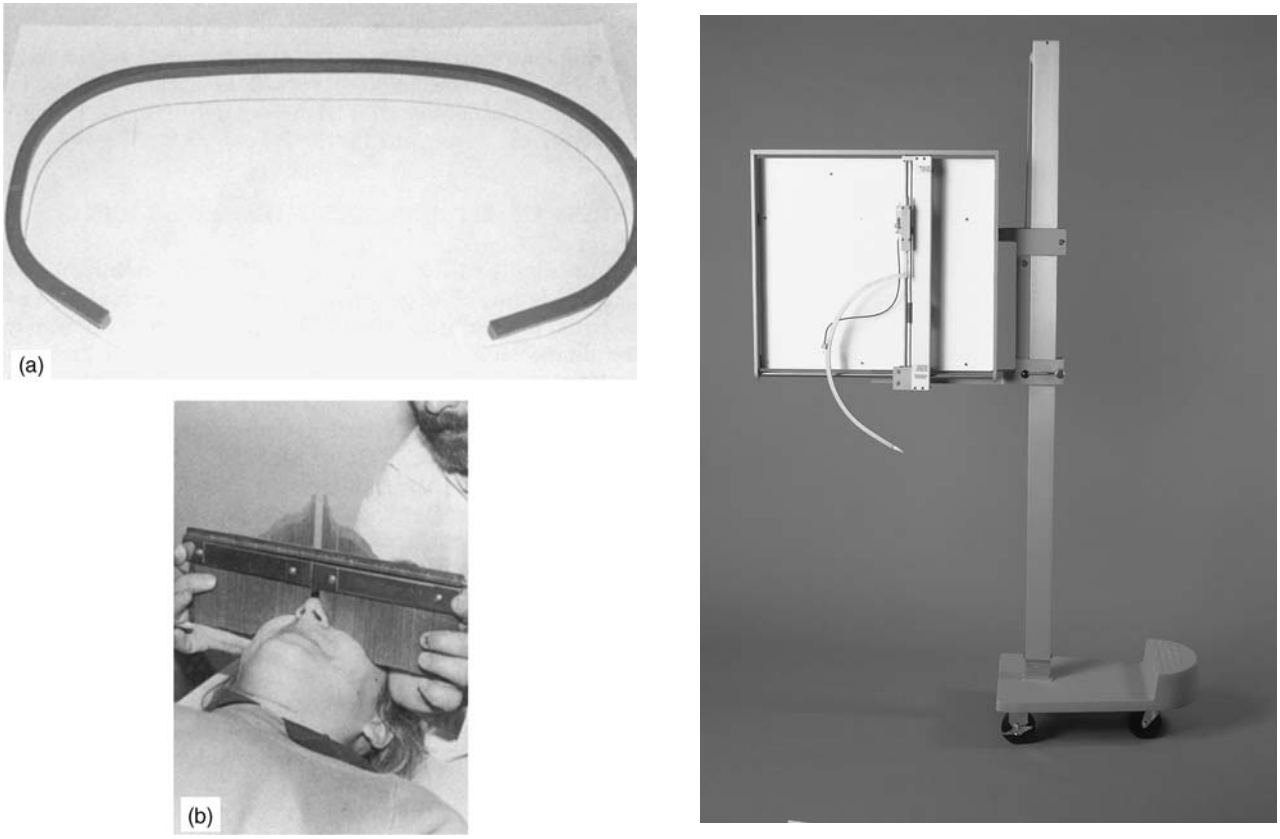


Figure 4. Contour plotter. The device is a simple, easy-to-use precision pantograph that links a drawing pen to a stylus arm and, upon contact with the body, communicates body contours to an overhead drawing board. The contour plotter is suspended on a vertical column and can easily be adjusted and locked securely. A continuous plot is drawn as the operator follows the physical contours of the patient. Marks can be made along the contour to indicate beam entry and laser light locations. (Courtesy of MEDTEC.)

(9,10). Such systems provide all the functionality of a conventional simulator, with the added benefit of increased treatment design options and the availability of software tools to facilitate the understanding and evaluation of treatment options. In addition, the simulation process is more efficient and less traumatic to the patient. Laser alignment lights and repositioning devices registered to the treatment couch are used to facilitate repositioning the patient in the treatment machine coordinate system once the virtual simulation process is complete.

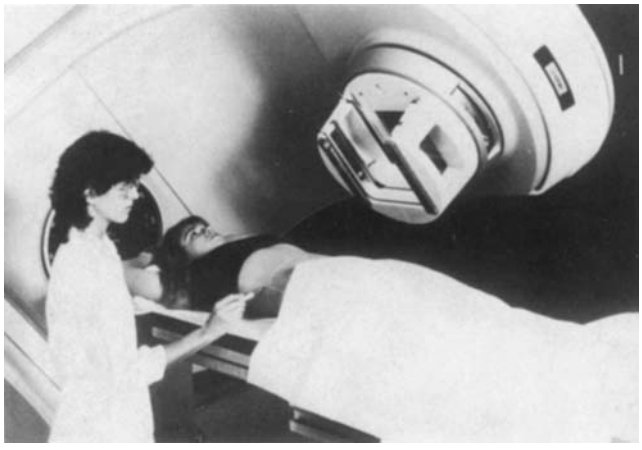


Figure 5. Laser alignment system. Patient in treatment position on treatment couch. Close-up of laser lines imaged on patient skin under typical treatment room lighting conditions. (Courtesy Gammex RMI, Inc.)

PATIENT RESTRAINT AND REPOSITIONING DEVICES

Accurate daily repositioning of the patient in the treatment position and reduction of patient movement during treatment is essential to accurately deliver the prescribed dose and achieve the planned dose distribution. As we will see in this section, modern day immobilization and repositioning systems are designed to be able to be attached to the simulation and treatment couches, so that the immobilization device and the patient are registered to the treatment machine coordinate system. Once the immobilization device has been locked into a specified position, the patient is then aligned to the immobilization system. The end result is that a set of coordinates is obtained from the CT simulator that is used in the virtual simulation process and that can be correlated to the treatment room isocenter.

The anatomic sites most often needing immobilization in radiation therapy are the head and neck, breast,

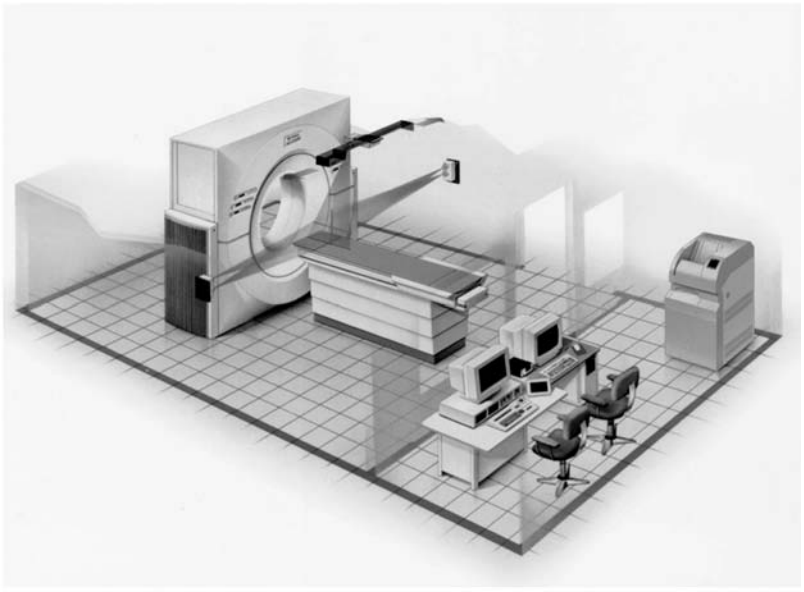


Figure 6. Typical CT simulation suite showing the scanner, flat tabletop, orthogonal laser system, virtual simulation workstation and hardcopy output device. (Courtesy of Philips Medical Systems.)

thorax–esophagus, shoulders and arms, pelvic areas (especially if obese), and limbs rotated to unusual positions. The precision achievable in the daily treatment positions of a patient depends on several factors other than the anatomic site under treatment, such as the patient's age, general health, and weight. In general, obese patients and small children are the most difficult to reposition.

Simple patient restraint and repositioning devices can be used in treating some anatomic sites. For example, the disposable foam plastic head holder shown in Fig. 7 provides stability for the head when the patient is in the supine position. If the patient is to be treated in the prone position, a face-down stabilizer can be used as shown in Fig. 8. This device has a foam rubber lining and a disposable

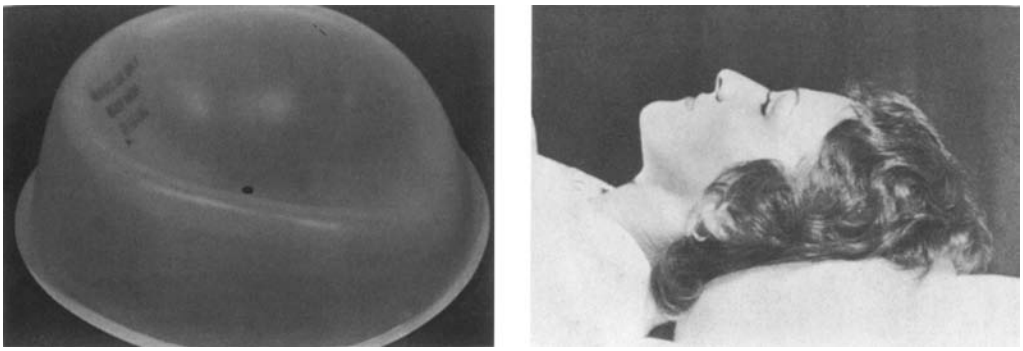


Figure 7. Disposable foam plastic head holder provides stability to the head when the patient is in the supine position.

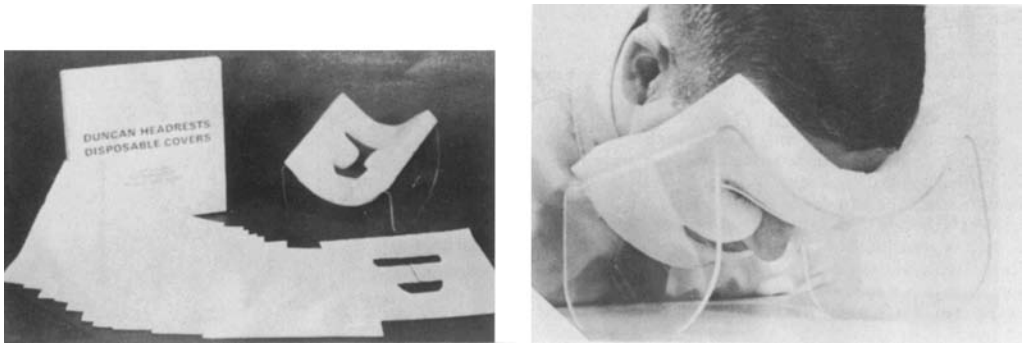


Figure 8. Face-down stabilizer. This formed plastic head holder has a foam rubber lining and disposable paper liner with an opening provided for the eyes, nose, and mouth. It allows comfort and stability as well as air access to the patient in the prone position.

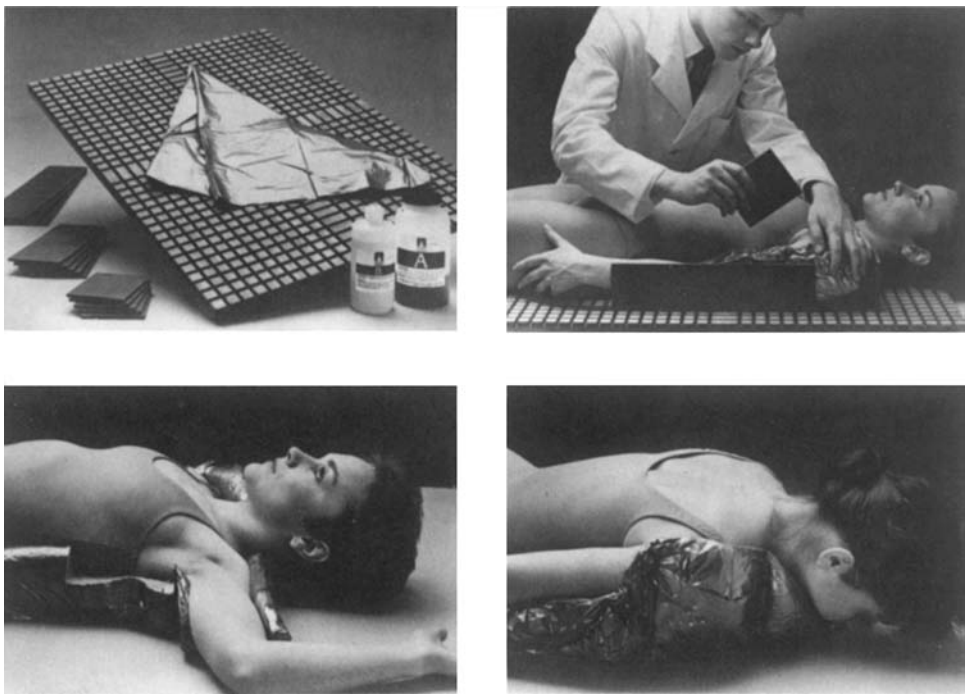


Figure 9. Polyurethane body mold. The chemical mixture is poured into the foam mold under a latex sheet. The patient is positioned in the foam mold as the polyurethane mixture expands to body shape. These body molds are easy to make, save time in patient alignment, and increase patient comfort. (Courtesy Smithers Medical Products, Inc.)

paper lining with an opening provided for the eyes, nose, and mouth of the patient. It allows comfort and stability as well as air access for the patient during treatment in the prone position.

There are now several commercially available body mold systems that are in widespread use as immobilization and repositioning aids. Fig. 9 illustrates one such system that utilizes a foam block cutout of the general anatomic area and polyurethane chemicals, which when mixed expand and solidify to conform to the patient's shape in a matter of minutes. Another widely used system (Fig. 10) consists of a vinyl bag filled with plastic minispheres. The bag is positioned around the patient to support the treatment position and then a vacuum is applied causing the minispheres to come together to form a firm solid support molded to the patient's shape.

Plaster casts are still used in some clinics, but have not gained widespread use in the United States, probably because they are too labor intensive and time consuming. Also, transparent form-fitting plastic shells (Fig. 11) that

Figure 10. Vacuum-form body immobilizer. The system consists of a plastic mattress filled with microspheres connected to a vacuum pump. Under vacuum, the mattress shapes itself to the body contours. (Courtesy of MEDTEC.)

Figure 11. Plastic shell. Transparent form-fitting plastic shells fabricated using a special vacuum device. (See Ref. 11.)

are fabricated using a special vacuum device are also used in some countries (e.g., Great Britain and Canada), but again are very labor intensive and have not gained acceptance in this country. Both methods are described in detail in the book by Watkins (11). In the United States, thermal plastic masks are much more commonly used (Fig. 12). The plastic sheet or mesh is placed in warm water to make it very pliable, and when draped over the patient conforms to the patient's shape and hardens upon cooling.

Fig. 13 illustrates a device called a bite block, which is used as an aid in patient repositioning in the treatment of head and neck cancer. With this device, the patient is

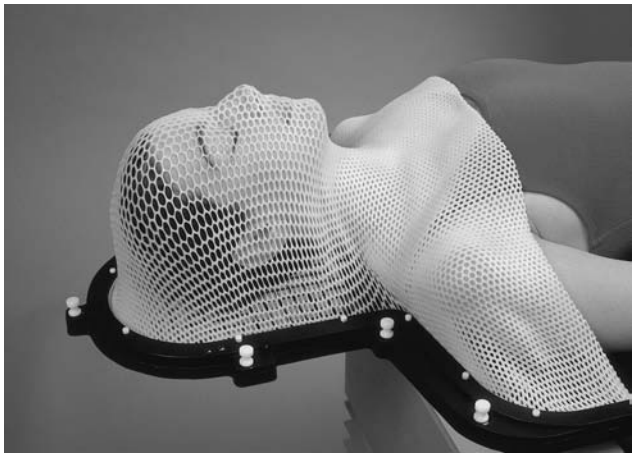


Figure 12. Thermoplastic cast. When placed in a warm (170°F) water bath, thermoplastic material becomes very flexible and can easily be molded to the patient's surface curvature. Immobilization using this material is less labor intensive than the conventional plaster cast or plastic shells and is therefore more readily adaptable on a routine basis for the immobilization of patients during radiation therapy. (Courtesy of MEDTEC.)

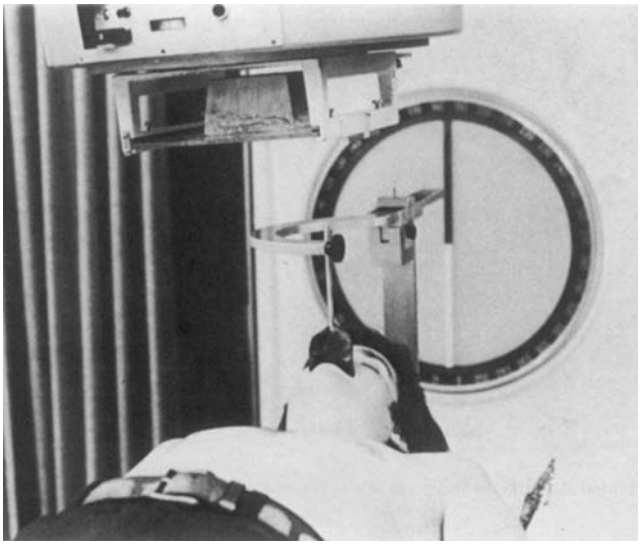


Figure 13. Bite-block system. Placement of the patient in a comfortable supine position and use of bite-block immobilization minimize patient movement for head and neck treatments. Note that the C-arm design allows both lateral and anterior beam arrangements to be used. (Courtesy of Radiation Products Design, Inc.)

placed in the treatment position and instructed to bite into a specially prepared dental impression material layered on a fork which is attached to a supporting device. When the material hardens, the impression of the teeth is recorded. The bite-block fork is connected to a support arm that is attached to the treatment couch and may be used either with or without scales for registration.

There are many other devices used to help in the treatment setup of patients that are site specific. For example, breast patients are usually positioned supine, with the arm on the involved side raised and out of the treatment area. Fig. 14 shows a device called a breast or tilt board that is used to optimize the position of the patient's chest wall (or thorax). The device is constructed with a hinge section that can be positioned and locked into place at various angles to the horizontal treatment table top. Modern breast boards now provide options for head support, arm positioning, and breast support. Sometimes, it is more convenient to use a separate arm board that can be attached directly to the treatment couch (Fig. 15). The perpendicular support provides a hand grasp that can be adjusted to the proper height and assists the patient in holding their arm in a comfortable position away from the treatment field.

Another useful device is the breast bridge (Fig. 16), which can be placed on the patient's chest and adjusted to the skin markings, for determining separation of the tangential fields. Precise angulation of the beam portals is determined using a digital readout level. In addition, a squeeze bridge (Fig. 17) with plastic plates can be used to provide buildup when a higher surface dose is desired, or one with a wire mesh frame can be used when no additional surface dose is warranted. A beam alignment device (Fig. 18) is used to match multiple fields used in tangential breast irradiation (12). The alignment component of the device is a curved piece of aluminum with a row of nylon

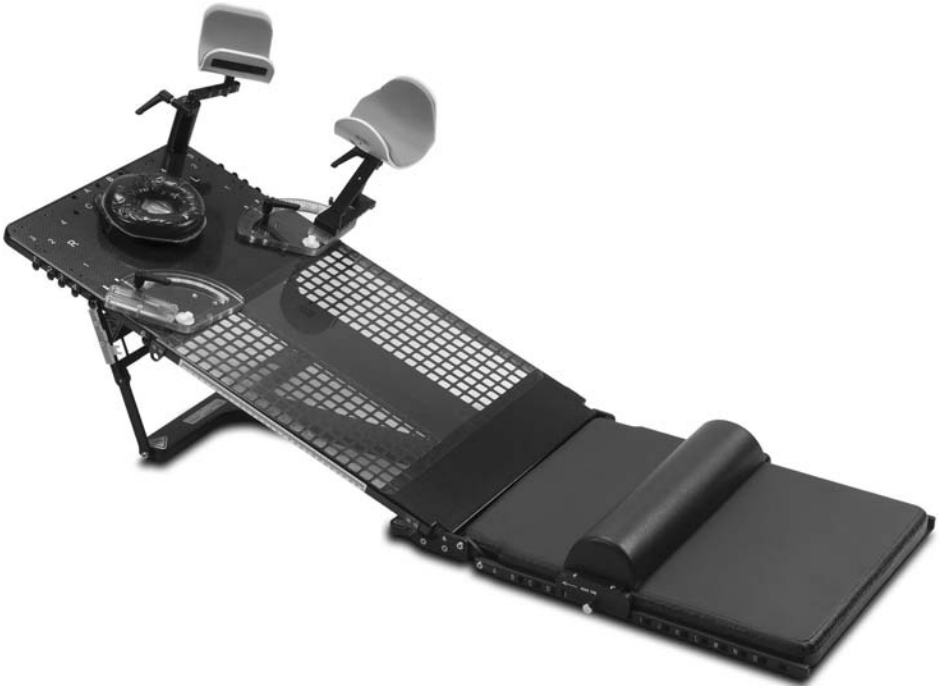


Figure 14. Tilt board or adjustable breast board. The top piece is fabricated with a hinged section that allows the sloping chest wall to be more appositional to a vertical beam. (Courtesy of MEDTEC.)

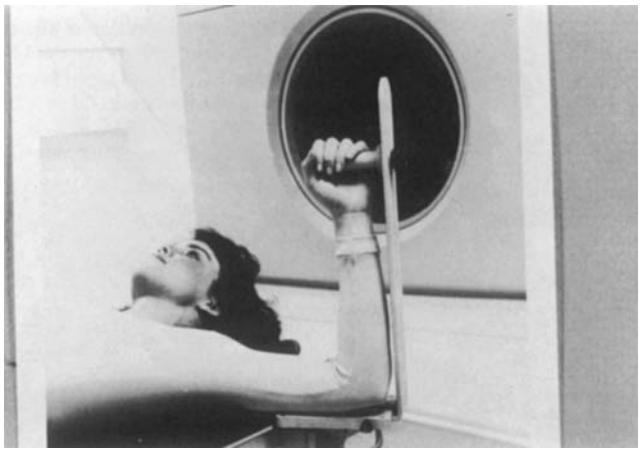


Figure 15. Patient arm support used for breast irradiation assists patients in holding their arm in a comfortable position, away from the treatment field. (Courtesy of Varian Medical Systems.)

Figure 16. A Breast Bridge is used with tangential radiation fields and consists of a pair of plastic plates that can be locked at the appropriate separation determined for the individual patient. After the treatment area has been marked, the bridge is placed on the patient's chest and adjusted to the skin markings, thus determining separation of the fields. Precise angulation of the portals is determined by the digital readout level. Once the portals are set, the bridge may be removed. (Courtesy of MEDTEC.)

pins protruding from its surface. The pins have a thin inscribed line to which the light field is aligned.

In some instances, it may be advantageous to treat the cancer patient in an upright position. The treatment chair (Fig. 19) is a device that facilitates such treatments by allowing a patient to be accurately repositioned in a seated position each day. The chair provides means of stabilizing the patient by the use of hand grips, elbow holders, and a seatbelt. The back of the seat is constructed of carbon fiber and thus the radiation beam can penetrate with minimal effects, and the angle of the seat back is adjustable.

Another device used to optimize patient position is the shoulder retractor. It provides a means by which the patient's shoulders can be pulled down in a reproducible manner, as illustrated in Fig. 20. Such a device is often used in the treatment of head and neck cancer involving lateral fields.

A standard feature on most accelerator treatment couches is a Mylar window or tennis racket-type table insert. This device consists of a thin sheet of Mylar stretched over a tennis racket-type webbing material and mounted in a frame that fits into a treatment table that has removable sections. Newer table insert devices made of carbon fiber (Fig. 21) eliminate the need to "restring" such panels and minimize the "sag" that can occur with nylon string panels. Such inserts provide excellent patient support

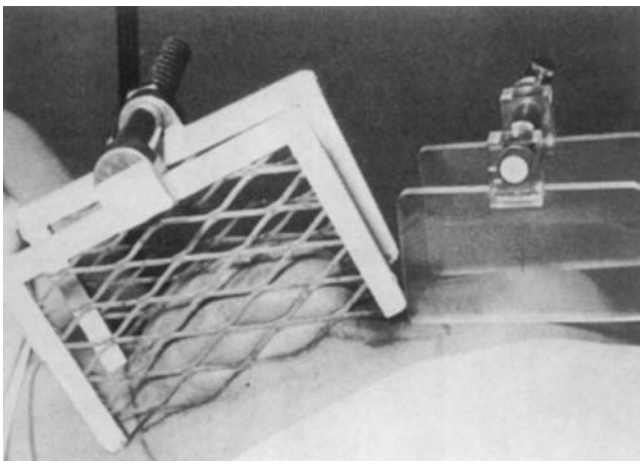


Figure 17. A squeeze bridge used for breast treatments with and without bolus. The wire mesh device is used where no additional surface dose is desired and the plastic frame device is used when increased surface dose is needed. (See Ref. 1.)

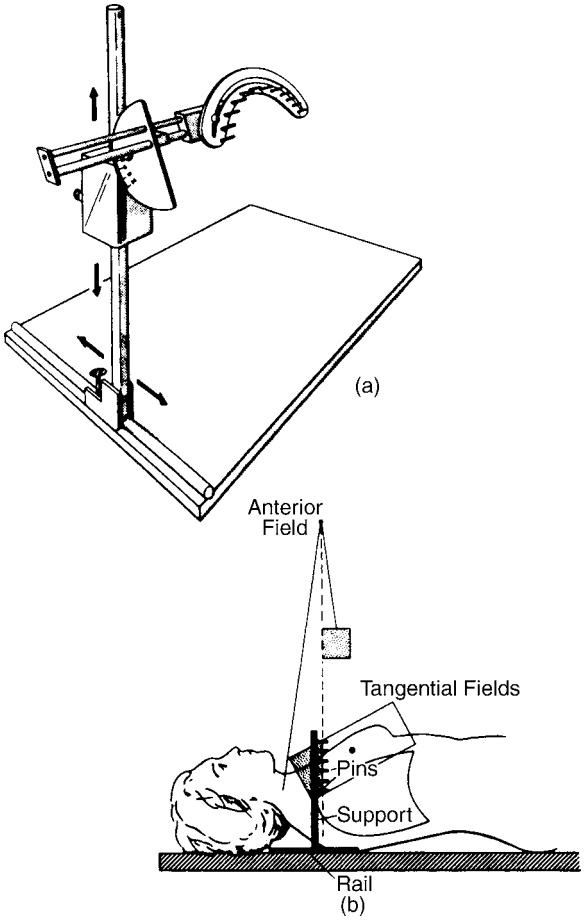


Figure 18. A beam alignment device used to implement field matching techniques. (a) Schematic of beam alignment device. (b) Example of use of device with three-field technique for breast treatment. Superior edges of tangential breast fields are coplanar and abutted to the vertical inferior edge of the anterior supraclavicular field. (See Ref. 12.)

Figure 19. Treatment chair. Provides positioning and fixation for breast, lung, and thorax patients who require vertical–upright positioning; adjusts to different locking positions and can accommodates a thermoplastic mask for head fixation. (Courtesy of MEDTEC.)

with a minimum of surface buildup effect for opposing beam portals, minimum reduction of beam intensity, and good visual access to the treatment surface. In addition, most medical linacs come with clamps that can be attached to

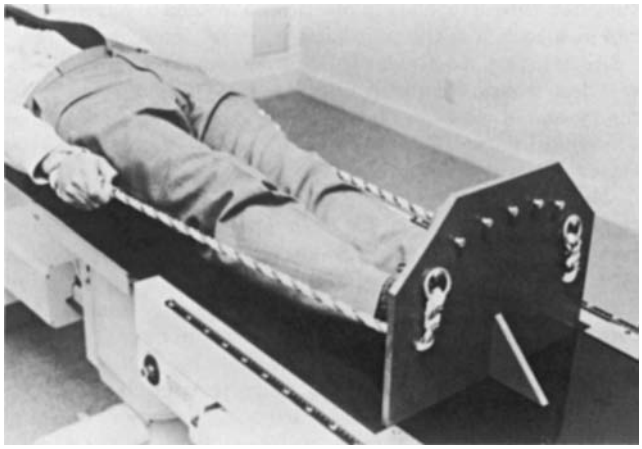


Figure 20. Shoulder retractor. Device used to pull the patient's shoulders down in a reproducible manner for treatment of the head and neck with lateral radiation fields. (Courtesy MED-TEC, Inc.)

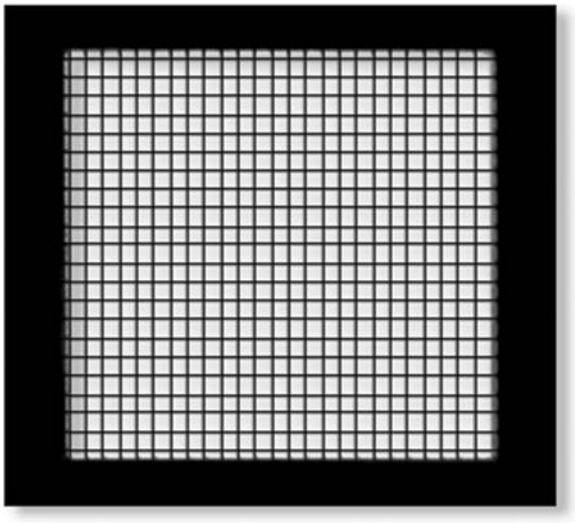


Figure 21. Treatment couch insert. Carbon fiber table and spine inserts for simulator and treatment couches. Rigid carbon fiber minimizes the "sag" that can occur with nylon string panels. (Courtesy of MEDTEC.)

the treatment couch and can be used with several different accessories, such as hand grips.

The development of 3D conformal radiation therapy (3DCRT) and more recently intensity-modulated radiation therapy (IMRT) has greatly enhanced the radiation oncologist's ability to plan and deliver very high doses that conform closely to the target volume, and falls off sharply, thus avoiding high dose to the nearby organs at risk (13,14). Both 3DCRT and IMRT invite the use of tighter margin to achieve higher dose escalation, and thus have spurred the development of new accessories and processes to better account for setup variation and organ motion that can occur during one (intra-) fraction, and between (inter-) fractions. Efforts thus far have focused on accounting for internal movement of the prostate gland and internal motion cause by respiratory function.

For example, for prostate cancer, the use of daily ultrasound imaging (Fig. 22), or daily electronic portal imaging of implanted radiopaque markers, has now become standard practice in many clinics (15,16).

Devices–methodologies used to address the problem of breathing motion in radiation treatment include: (1) gating and/or tracking and (2) breathhold devices–strategies. In gating and tracking, the state of the treatment machine is adjusted in response to a signal that is representative of a patient's breathing motion. With breathholding, the lung volume of the patient is directly immobilized prior to beam-on, and released after the beam is off. The basic components of a gating or tracking system consist of a respiration sensor whose signal is processed and evaluated by a computer for suitability to trigger, or gate, the radiation. An example of a respiratory gating system is the Real-time Position Management (RPM) system (Fig. 23) commercially available from Varian Medical Systems (Palo Alto, CA) (17,18). An example of a breath control device is the Elekta Inc. (Norcross, GA) Active Breathing Coordinator (ABC) (Fig. 24) (19). The ABC apparatus is used to suspend breathing at any predetermined position along the normal breathing cycle, or at active inspiration, and consists of a digital spirometer to measure the respiratory trace, which is in turn connected to a balloon valve. Another example is the ExacTrac system (BrainLAB AG) that combines X-ray imaging and infrared tracking that permits correlation of internal 3D tumor motion with the patient's breathing cycle (Fig. 25) (20). Automatic fusion of digitally reconstructed radio-graphic (DRR) images computed from the treatment planning CT data to the live X-rays allows any set-up error or target shift and rotation to be identified and any discrepancy compensated for via robotic table movement.

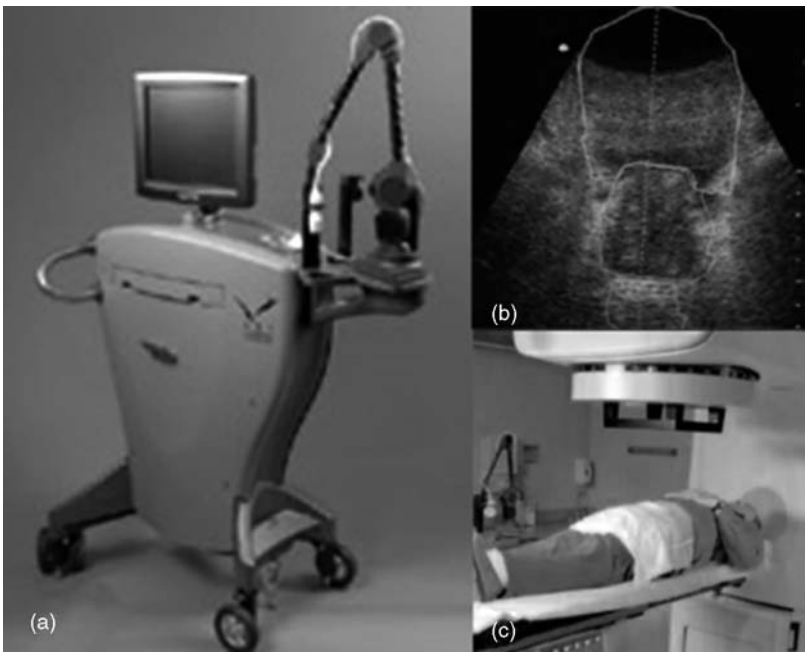


Figure 22. BAT (B-mode Acquisition and Targeting) SXi system. Targeting device that provides fast ultrasound localization of a treatment target on a daily basis. (Courtesy North American Scientific)

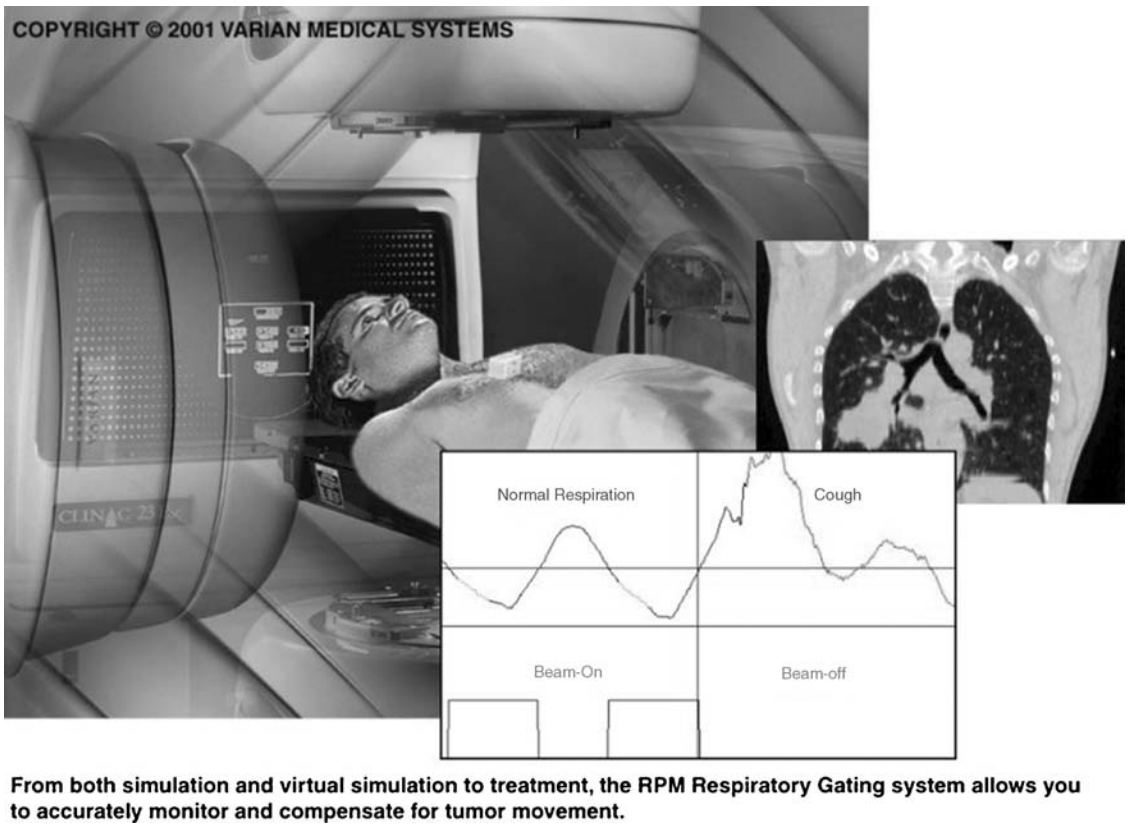


Figure 23. Real-time Position Management system. Device used to allow respiratory gating treatment delivery. (Courtesy of Varian Medical Systems)

For stereotactic radiosurgery or fractionated stereotactic radiotherapy, a suite of accessories is now available. Most important is a head stereotactic localizer such as the Gill–Thomas–Cosman (GTC) relocatable head ring (Fig. 26), which enables precise fixation and localization and repositioning of targets in the cranium (21). Another important device when using the newly emerging radiotherapy treatment, stereotactic body radiation therapy (SBRT), in which a high dose is delivered in either a single fraction or just a few fractions, is the frame-based body stereotactic immobilization system (22). There are several now commercially available; an example is the Elekta Stereotactic Body Frame shown in Fig. 27. This device provides a reference stereotactic coordinate system that is external to the patient's body, so that the coordinates of a target volume can be reproducibly localized during simulation and treatment. This frame has built-in reference indicators for CT or MR determination of target volume coordinates. In addition, a diaphragm control attached to the frame can be used to minimize respiratory movements. Horizontal positioning of the frame, on the CT simulator or treatment couch, is achieved using an adjustable base on the frame.

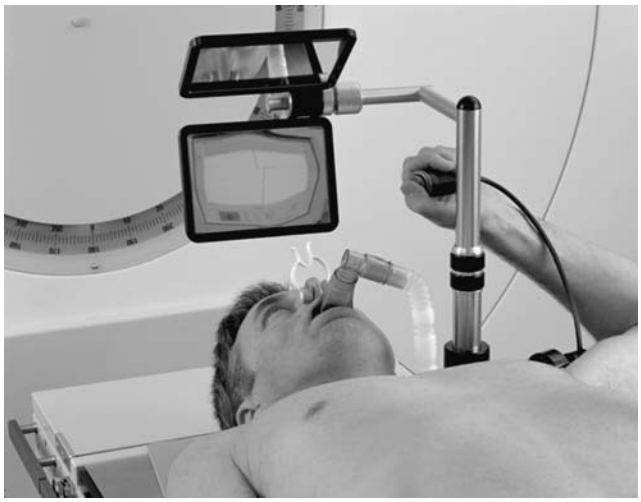


Figure 24. Active Breathing Coordinator^{TM}system. Device allows the radiation oncologist to pause a patient's breathing at a precisely indicated tidal volume and coordinate delivery with this pause. (Courtesy of Elekta AB)

FIELD-SHAPING, SHIELDING, AND DOSE MODIFYING DEVICES

The Lipowitz metal (Cerrobend) shielding block system introduced by Powers et al. (23) is in widespread use throughout the world. The block fabrication procedure is illustrated in Fig. 28 and more details using this form of field shaping can be found in the review article by Leavitt and Gibbs (24). Lipowitz metal consists of 13.3% tin, 50.0% bismuth, 26.7% lead, and 10.0% cadmium, and has a

Figure 25. ExacTrac X-Ray 6D automated image-guided radiation therapy system. System consists of 2 kV X-ray units recessed into the linac floor and two ceiling-mounted amorphous silicon Flat Panel detectors integrated with a real-time infrared tracking device to enable imaging of internal structures or implanted markers for extremely accurate set-up of the target volume's planned isocenter position. (Courtesy Brain LAB AG.)

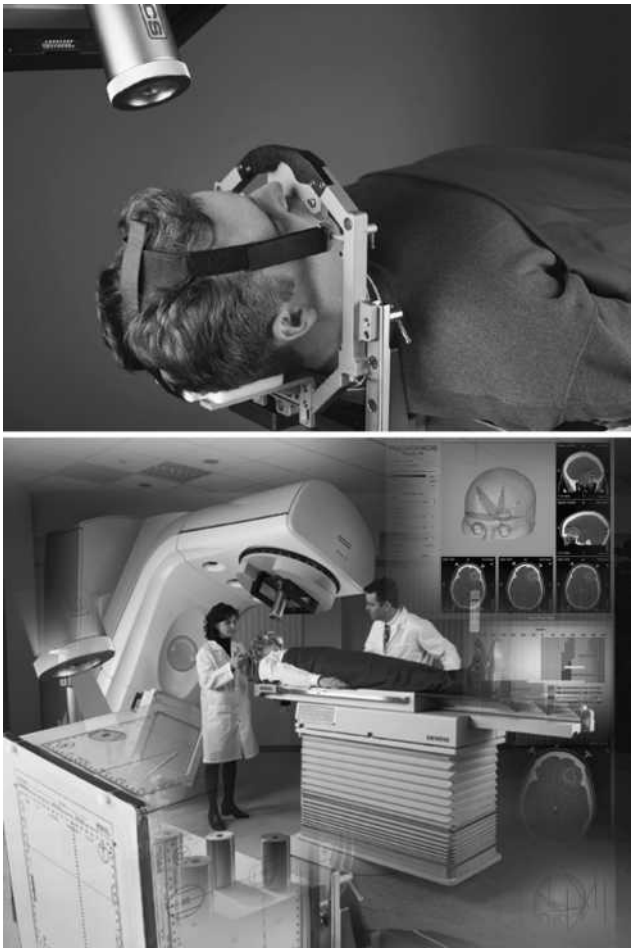


Figure 26. Gill–Thomas–Cosman relocatable head ring. Device used for stereotactic radiosurgery or fractionated stereotactic radiotherapy that enables precise fixation and localization and repositioning of targets in the cranium. Copyright © 2005 Radionics. All rights reserved. Reprinted with the permission of Radionics, a division of Tyco Healthcare Group LP.

physical density at 20 °C of 9.4 $g \cdot cm^{-3}$ as compared with 11.3 $g \cdot cm^3$ for lead. A simulation radiograph is obtained with the patient in the treatment position and the desired treatment field aperture is drawn on the radiograph by the radiation oncologist. The marked radiograph is then used as a template for cutting a foam mold with a "hot-wire" cutting device in which molten Lipowitz alloy is poured and then cooled to form a shielding block. Computer-controlled adaptations of the hot-wire cutting technique have evolved as an adjunct to 3D treatment planning in which the treatment field shape is defined based on beam's eye-view displays. The shaped field coordinates are transferred directly to the computer-controlled blockcutting system, thereby eliminating potential errors in manual tracing, magnification, or image reversal. The other steps in the block forming and verification process remain similar to the manual procedure.

Figure 27. Elekta Stereotactic Body Frame®. Used for stereotactic body radiation therapy (SBRT), in which a high dose is delivered in either a single fraction or just a few fractions. (Courtesy of Elekta AB.)

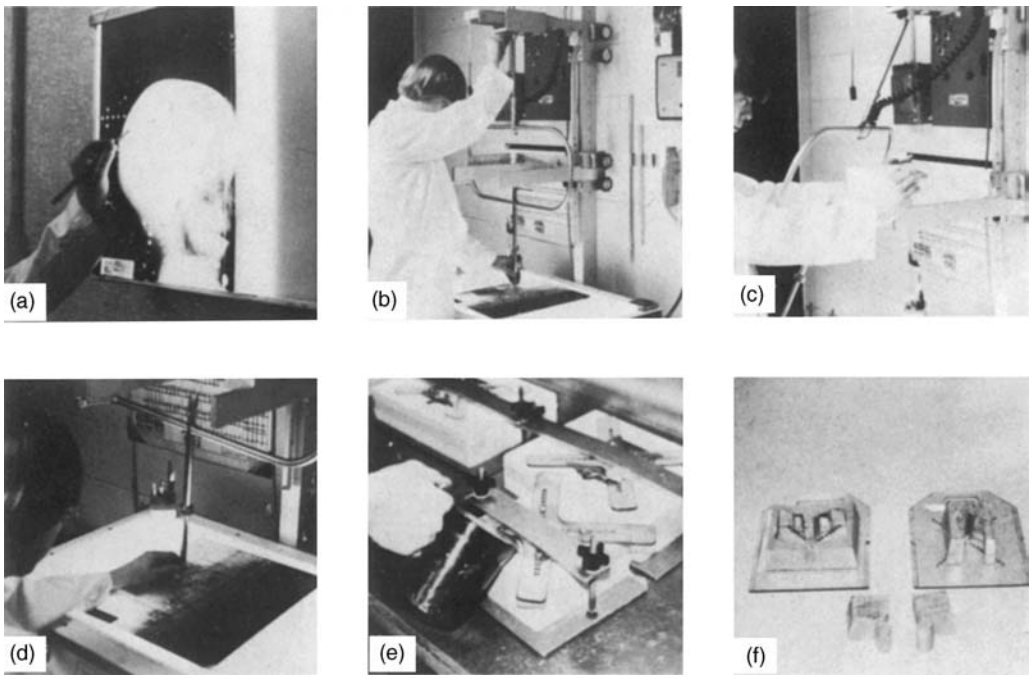


Figure 28. Shielding block for external photon beam irradiation. The design and fabrication technique is described as follows. (a) Physician defines the treatment volume on the X-ray simulator radiograph. (b) Physics technician adjusts SSD and STD of hot-wire cutter to emulate simulator geometry. (c) Proper-thickness foam block is aligned to central axis of cutter. (d) Foam mold is cut using hot-wire cutter. Courtesy of Huestis Machine Corp. (e) Foam pieces are aligned and held in place with a special clamping device. Molten alloy is poured into the mold and allowed to harden. (f) Examples of typical shielding blocks cast using this system.

Returning to accessories used for field shaping, Fig. 29 shows a Multileaf Collimator (MLC) system. The MLCs were first introduced in Japan in the 1960's (25) and have now gained widespread acceptance, and have replaced alloy blocking as the standard-of-practice for field shaping in modern radiation therapy clinics. All medical linac manufacturers now provide MLC systems. The leaves are typically carried on two opposed carriages that transport the leaves in unison. The leaves (under computer control) can be individually positioned to serve either as a block replacement, or as a means to provide IMRT treatment delivery.

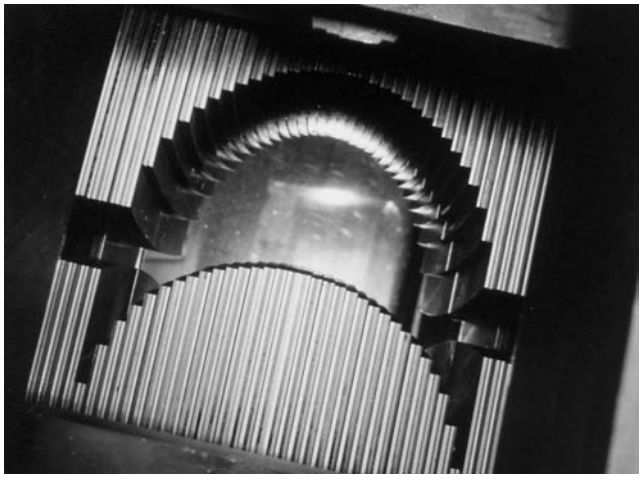


Figure 29. Multileaf Collimation (MLC) system. Computer controlled system used to shape beam aperture and also to deliver IMRT. (Courtesy of Elekta AB.)

Fig. 30 shows a commercial binary multi-leaf collimator system (called "MIMiC") designed and built by the NOMOS Corporation and incorporated into their serial tomotherapy system, known as Peacock, for planning and rotational delivery of IMRT treatments (26). This device can be attached to a conventional linac to deliver IMRT by rapidly moving leaves in or out of a slit field. Like a CT unit, the radiation source and the collimator continuously revolve around the patient.

Other more conventional devices used to achieve a desired dose distribution and to correct for perturbing influences, such as patient shape, include bolus, wedges, and compensating filters. Bolus is a tissue-equivalent material used to smooth an irregular surface, increase the dose to the patient's surface region, or sometimes to fill external air cavities, such as the ear canal or nasal passage. Bolus should have an electron density and atomic number similar to that of tissue or water. Examples of bolus materials include slabs of paraffin wax, rice bags filled with soda ash, gauze coated with Vaseline, and synthetic-based substances, such as Superflab (Fig. 31).

Wedge filters (Fig. 32) are typically made of a dense material, such as brass or steel, and are mounted on a frame that can be inserted into the beam at a specified

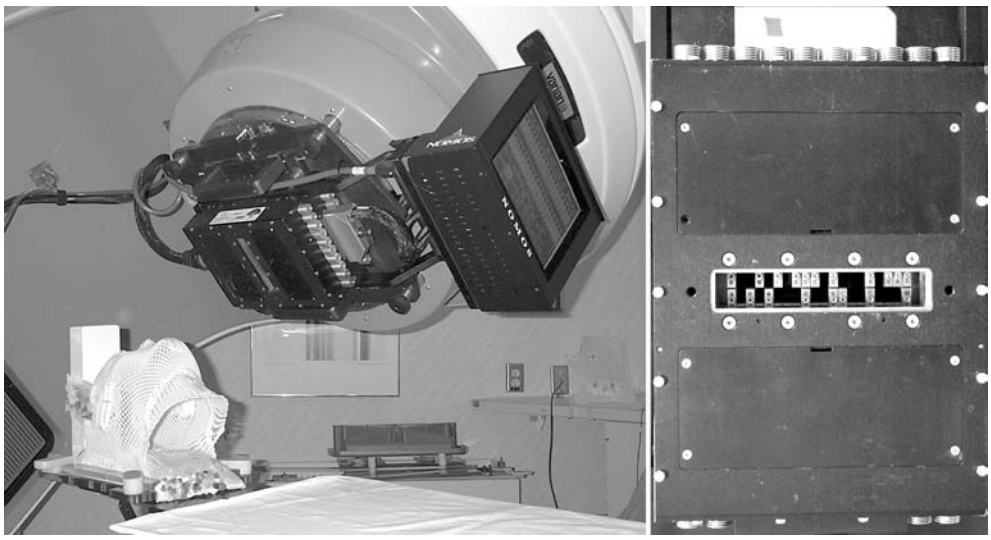


Figure 30. Binary multileaf collimator system (called MIMiC) designed and built by the NOMOS Corporation and incorporated into their serial tomotherapy system, known as Peacock for planning and rotational delivery of IMRT treatments. This device can be attached to a conventional linac to deliver IMRT by rapidly moving leaves in or out of a slit field. Like a CT unit, the radiation source and the collimator continuously revolve around the patient.

distance from the source and cause the dose distribution at a specified depth to be angled to a desired amount relative to the incident beam direction. Modern linacs have replaced physical wedges with software control of independent jaws that provide what is called dynamic wedging (27, 28). Currently, there are two versions of dynamic wedging, the Varian Enhanced Dynamic Wedge (EDW), and the Siemens Virtual Wedge (VW). The desired wedge angle is achieved by moving a collimating jaw while the beam is on, thereby shrinking the field during treatment.

A compensating filter (Fig. 33) is a beam modifier that is used to counteract the effect of air gaps caused by the patient's topography while still preserving the skin-sparing characteristic of megavoltage photon beams (29). To do this, the compensating filter is placed in the beam some distance away from the patient's skin surface. This requires that the lateral dimensions of the filter be reduced and causes the scatter radiation conditions to be altered, complicating the relationship between the thickness of the compensator along a ray and the amount of tissue deficit to be compensated.

Other accessories include shields designed specifically to protect certain organs at risk. For example, in the treatment of Hodgkin's disease and other malignant lymphomas, in which the inguinal and femoral lymph nodes are frequently irradiated, a testicular shield (Fig. 34) is frequently used. This device is typically constructed with lead and is designed to reduce scatter radiation to the testicles. Another example is the eye shields shown in

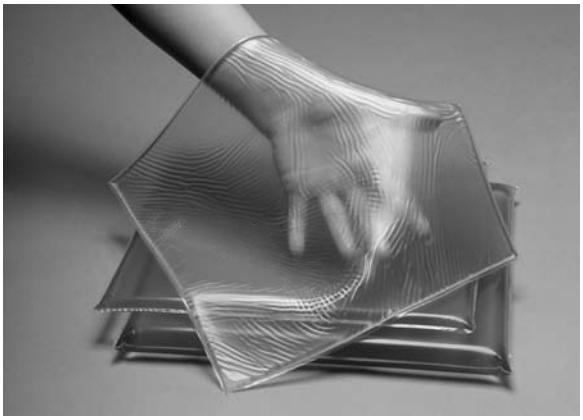


Figure 31. Example of bolus material used in radiation treatment to smooth the patient's irregular surface or to increase the dose to the surface region. (Courtesy of MEDTEC.)

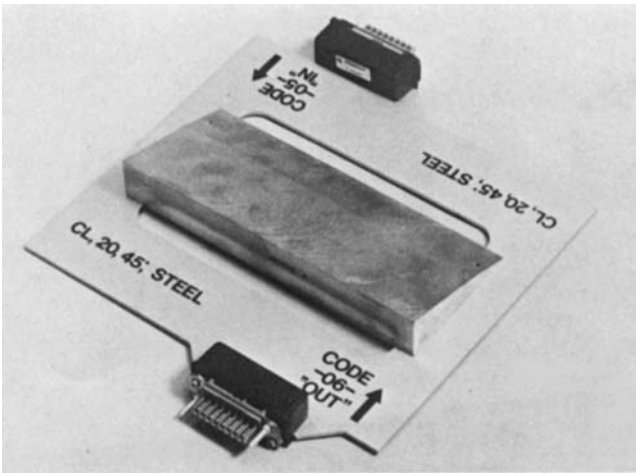


Figure 32. Wedge filter used on medical linear accelerator to shape dose distribution. (Courtesy Varian Medical Systems.)

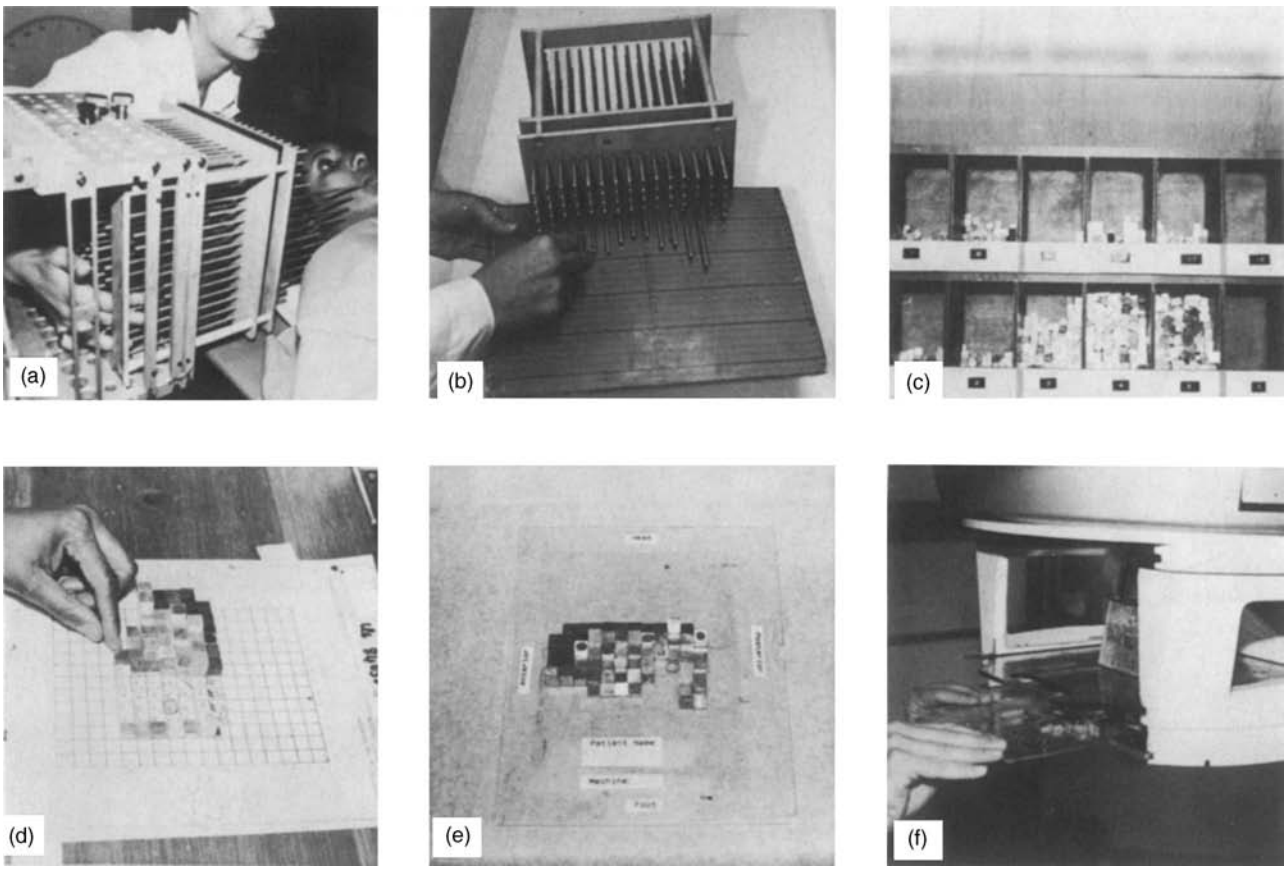


Figure 33. Compensating filter. Composite photograph illustrating design and fabrication procedure: (a) Push rod device is attached to the head of the simulator directly over the patient in the treatment position. The rods are lowered one by one, until they are in contact with the patient's surface. (b) Tissue deficits are determined from measurement of each rod length. (c) Aluminum and brass blocks, which are numbered to correspond to the tissue deficits determined, are selected from storage bins. (d) Blocks are attached on the plastic tray according to the pattern specified by the calculations. (e) Completed filter. (f) Compensating filter positioned on treatment machine.

Figure 34. Testicular Shield. Device used to position and shield the gonadal area from scattered radiation. (Courtesy of MEDTEC.)

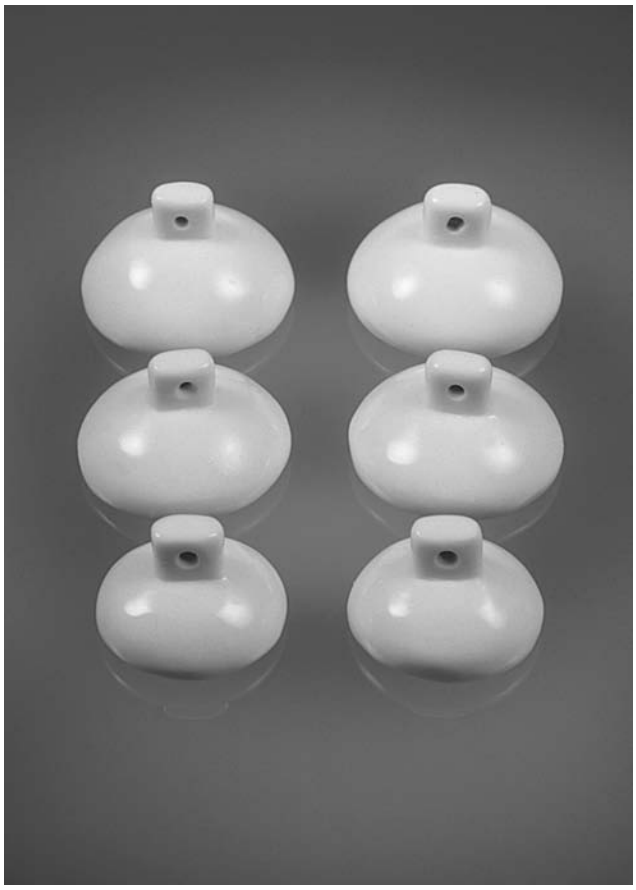


Figure 35. Tungsten eye shields provide protection of the ocular structure for electrons up to 9 MeV. Each eye shield is coated with a 2 mm minimum thickness of dental acrylic on the beam entrance of the shield to reduce electron backscatter to an acteptable level. (Courtesy of MEDTEC.)

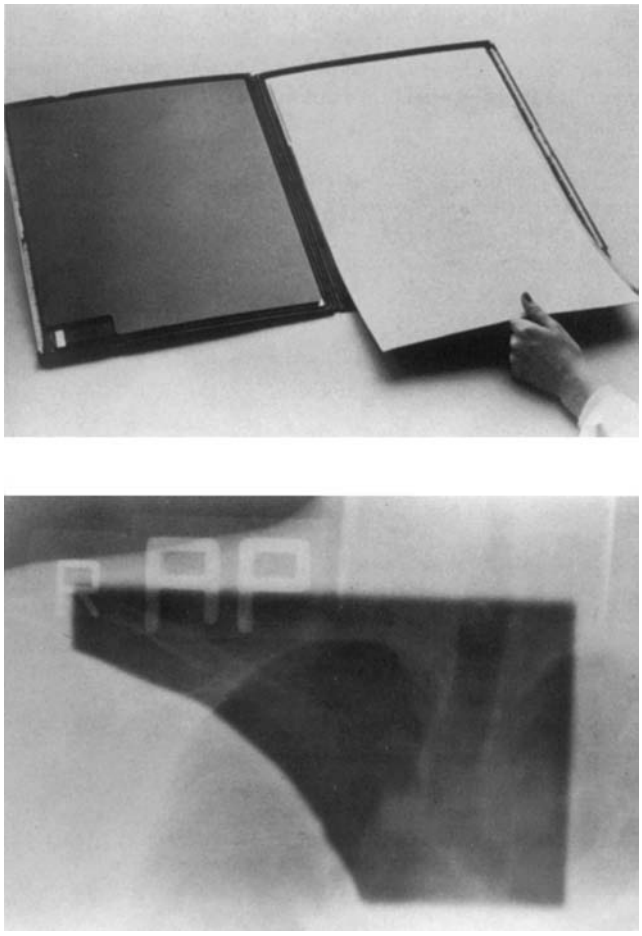


Figure 36. Radiotherapy port film cassette. Copper screens are used to improve subject contrast and improve resolution as shown in the portal localization radiograph. (Courtesy Eastman Kodak Co.)

Fig. 35 that provide protection of the ocular structure for electrons up to 9 MeV. Each eye shield is made of tungsten and coated with a 2 mm minimum thickness of dental acrylic on the beam entrance of the shield to reduce electron backscatter to an acceptable level.

TREATMENT VERIFICATION AND QUALITY ASSURANCE DEVICES

Radiographic film (port films) using film cassettes with lead or copper filters which improve the radiographic contrast of the port films are typically used to verify patient isocenter position and for portal shape (Fig. 36). Devices to support the port film cassettes and to help insure that the film plane is orthogonal to the beam direction and close to the patient's surface from which the beam exits are also important accessories (Fig. 37). In addition, radiopaque graticules (Fig. 38) that can be inserted into the treatment beam are invaluable for the evaluation of port films (30). Typically, such devices consist of platinum or tungsten wire embedded in plastic and molded in a frame that can be attached to the treatment machine accessory mount.

The generally poor quality of the film images and the inconveniences of the film processing and physician reviewing procedure have spurred development of computer-aided enhancement and digital imaging techniques in radiation therapy. Computed radiography (CR) systems, such as the Kodak 2000RT CR system shown in Fig. 39, is an example of such a system that allows digital DICOM images to be distributed electronically throughout the department.

In addition, electronic portal imagers (EPID) have made great strides this past decade and such devices are poised to replace film over the next few years (31,32). More recently, linac manufacturers have integrated EPID and tomographic imaging systems on their linacs (Fig. 40) for localization of bone and soft-tissue targets and have set the stage for image-guided radiation therapy to move into routine practice (33,34). Such systems will make quantitative evaluation of immobilization and repositioning of the patient much more achievable by allowing daily imaging of the patient's treatment.

In addition to imaging verification, there is sometimes a need to verify actual dose delivered to the patient. Simple point dose verification can be achieved using TLDs, diodes,

Figure 37. Port film cassette holder. Devices used to support a port film cassette behind the patient in any orientation device are especially useful for oblique treatment angles as the plane of film can be adjusted so as to produce normal incidence of the radiation field. (Courtesy of MEDTEC.)

Figure 39. Kodak 2000RT CR system. Computed radiography system that allows digital DICOM images to be distributed electronically throughout the department. (Courtesy Eastman Kodak Co.)

or MOSFET dosimeters placed on the patient's skin surface or in body cavities. The most recent development in these types of devices is the OneDose patient dosimetry system (Fig. 41) developed by Sicel Technologies, Inc. It consists of a wireless handheld reader which interacts with self-adhesive external MOSFET dosimeters placed on the patient. To use, the therapist simply places the precalibrated dosimeters on the patient's skin in the treatment field, treats the patient, and then slides the dosimeter into the reader

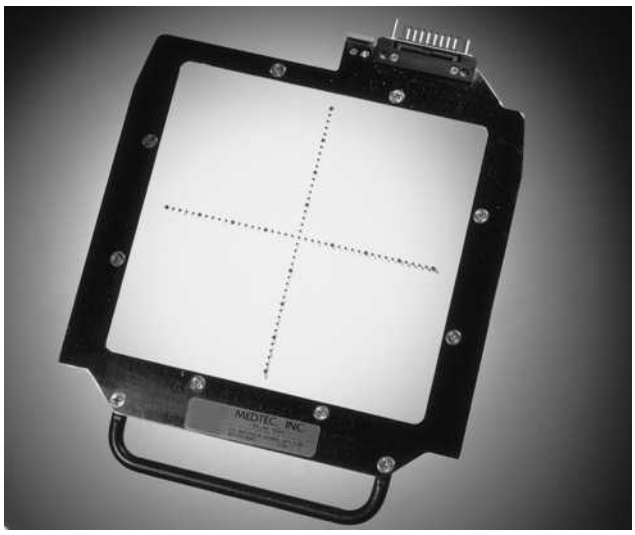


Figure 38. Example of a port film fiducial grid. Device identifies the central axis of the radiation beam and provides a scale at the calibration distance, thus providing a magnification factor for the port film. (Courtesy of MEDTEC.)

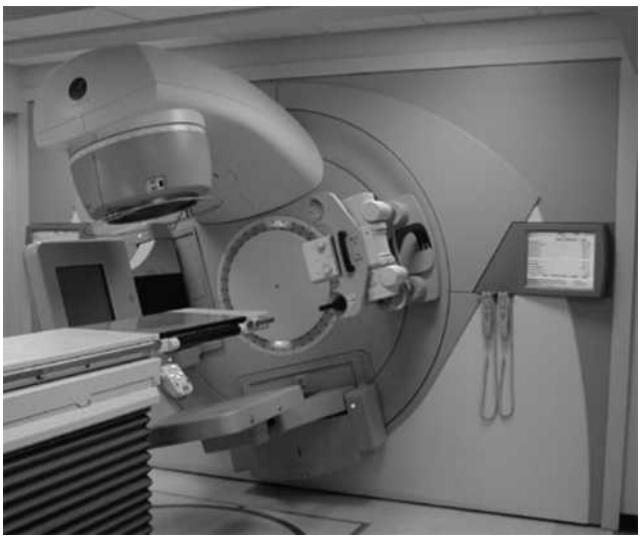


Figure 40. Elekta Synergy® system. Linac with electronic portal imager (EPID) and conebeam tomographic imaging system for localization of bone and soft-tissue targets. (Courtesy of Elekta AB.)

Figure 41. Patient dosimetry monitoring system. Semiconductor detectors are typically used and applied on the patient's surface using surgical tape. (Courtesy of MEDTEC.)

for an immediate display. The reader automatically provides a permanent record of dosage, time, and date with minimal data entry. Another dose verification device is shown in Fig. 42; because IMRT treatments presently require verification of the dose delivered and pattern for each patient, special phantoms and/or check devices have been developed to facilitate the IMRT verification measurements.

Also, there are numerous QA devices available to check the constancy of the linac's beam calibration, symmetry, and radiation-light field alignment. Generally, the QA radiation detection devices consist of several ionization chambers or semiconductors positioned in a plastic phantom that can be placed in the radiation beam (Fig. 43).

Finally, one of the latest advances is a system (still under development) that will be capable of performing continuous objective, real-time tracking of the target volume during treatment (Calypso Medical Technologies, Inc.) (35). The system is based on alternating current (ac) magnetic fields utilizing permanently implantable wireless transponders that do not require additional ionizing radiation and do not depend on subjective interpretation of images. The system is currently undergoing clinical evaluation and is not yet available for clinical use.

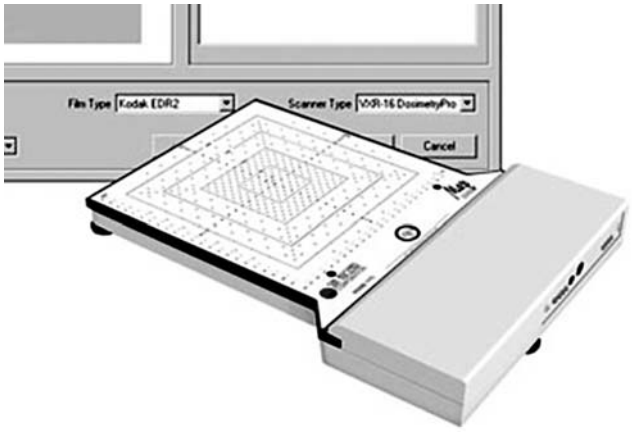


Figure 42. MapCHECK device provides 2D therapy beam measurements intended for quick and precise verification of the dose distribution resulting from an IMRT plan. (Courtesy of Sun Nuclear.)

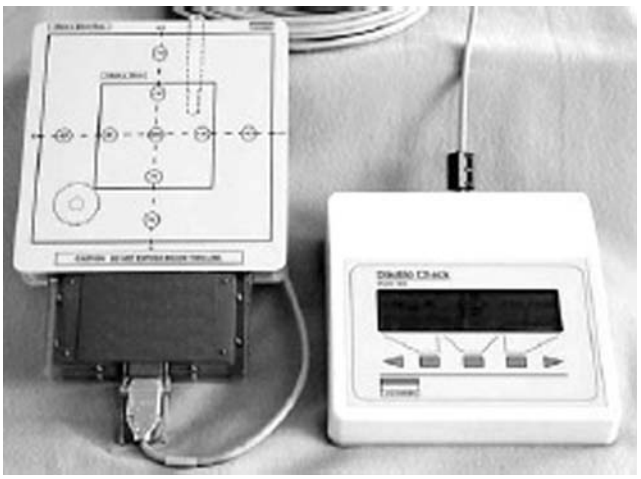


Figure 43. DoubleCheck A matrix detector, consisting of multiple ionization chambers embedded in a plastic phantom, used to check the constancy of radiation beam output , symmetry, and flatness. (Courtesy MED-TEC, Inc.)

SUMMARY

In summary, numerous radiotherapy accessories are used to aid in planning, delivering, and verifying radiation treatments. New systems continue to be developed that give the ability to more accurately position the patient and account for the internal target volume relative to the treatment machine's isocenter. Such devices enable significant reductions in the amount of normal tissue included in the irradiated volume. Finally, it should be recognized that the preferences of the individual radiation oncologist, radiation therapist, and clinical physicist still play a major role in the acceptance and use of all of the types of devices discussed in this article. The major considerations for any specific application are typically cost, speed and ease of preparation, ease of use, and effectiveness.

BIBLIOGRAPHY

1. Levitt SH, Khan FM, Potish RA, Perez CA, editors. Technological Basis of Radiation Therapy: Clinical Applications. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 1999.
2. ICRU, Report No. 24, Determination of Absorbed Dose in a Patient Irradiated by Beams of X or Gamma Rays in Radiotherapy Procedures. Washington, D.C.: International Commission on Radiation Units and Measurements. 1976.
3. Purdy JA. 3-D radiation treatment planning: a new era, in Frontiers of Radiation Therapy and Oncology. 3-D Conformal Radiotherapy: A New Era in the Irradiation of Cancer. In: Meyer JL, Purdy JA, editors. Basel: Karger; 1996. p 1–16.
4. Purdy JA. Defining our goals: volume and dose specification for 3-D conformal radiation therapy, in Frontiers of Radiation Therapy and Oncology. 3-D Conformal Radiotherapy: A New Era in the Irradiation of Cancer. In: Meyer JL, Purdy JA, editors. Basel: Karger; 1996. p 24–30.

5. Purdy JA, Klein EE, Low DA. Quality assurance and safety of new technologies for radiation oncology. Sem Radiat Oncol 1995;5(2):156–165.
6. Van Dyk J, Mah K. Simulators and CT scanners. In: Williams JR, Thwaites DI, editors. Radiotherapy Physics. New York: Oxford Medical Publications; 1993.
7. Prasad S, Pilepich MV, Perez CA. Contribution of CT to quantitative radiation therapy planning. Am J Radiol 1981;136:123.
8. Goitein M. Applications of computed tomography in radiotherapy treatment planning. Progress in Medical Radiation Physics. 1 In: Orton CG, editor. New York: Plenum; 1982. p 195–293.
9. Perez CA, et al. Design of a fully integrated three-dimensional computed tomography simulator and preliminary clinical evaluation. Init J Radiat Oncol Biol Phys 1994;30(4):887–897.
10. Mutic S, et al. Quality assurance for computed-tomography simulators and the computed-tomography-simulation process: Report of the AAPM Radiation Therapy Committee Task Group No. 68. Med Phys 2003;30(10):2762–2792.
11. Watkins DMB. Radiation Therapy Mold Technology. Toronto, Canada: Pergamon Press; 1981.
12. Buck BA, Siddon RL, Svensson GK. A beam alignment device for matching fields. Int J Radiat Oncol Biol Phys 1985;11:1939.
13. Purdy JA, et al. 3-D Conformal and Intensity Modulated Radiation Therapy: Physics and Clinical Applications. Madison, (WI): Advanced Medical Publishing, Inc; 2001. p 612.
14. Sternick ES, editor. The Theory and Practice of Intensity Modulated Radiation Therapy. Madison (WI): Advanced Medical Publishing; 1997. p 254.
15. Balter JM, et al. Measurement of prostate movement over the course of routine radiotherapy using implanted markers. Int J Radiat Oncol Biol Phys 1995;31:113–118.
16. Lattanzi J. A comparison of daily CT localization to a daily ultrasound-based system in prostate cancer. Int J Radiat Oncol Biol Phys 1999;43(4):719–725.
17. Mageras GS. Flouroscopic evaluation of diaphragmatic motion reduction with a respiratory gated radiotherapy system. J Appl Clin Med Phys 2001;2(4):191–200.
18. Vedam SS, Keall PJ, Kini VR, Mohan R. Determining parameters for respiration-gated radiotherapy. Med Phys 2001;28(10): 2139–2146.
19. Wong J, et al. The use of active breathing control (ABC) to reduce margin for breathing control. Int J Radiat Oncol Biol Phys 1999;44:911–919.
20. Verellen D, et al. Quality assurance of a system for improved target localization and patient set-up combines real-time infrared tracking and stereoscopic X-ray imaging. Radio Oncol 2003;67(1):129–141.
21. Schlegel W, et al. Computer systems and mechanical tools for stereotactically guided conformal therapy with linear accelerators. Int J Radiat Oncol Biol Phys 1992;24:781.
22. Lax I, et al. Stereotactic radiotherapy of malignancies in the abdomen: methodological aspects. Acta Oncol 1994;33:677–683.
23. Powers WE, et al. A new system of field shaping for external-beam radiation therapy. Radiology 1973;108:407–411.
24. Leavitt DD, FA Gibbs Jr. Field shaping, in Advances in Radiation Oncology Physics: Dosimetry, Treatment Planning, and Brachytherapy. In: Purdy JA, editor. New York: American Institute of Physics; 1992. pp. 500–523.
25. Takahashi S. Conformation radiotherapy: rotation techniques as applied to radiography and radiotherapy of cancer. Acta Radiol (Suppl) 1965;242:1–42.
26. Carol MP. Integrated 3-D conformal multivane intensity modulation delivery system for radiotherapy. In Proceedings of the 11th International Conference on the Use of Computers in Radiation Therapy. Madison (WI): Medical Physics Publishing; 1994.
27. Leavitt DD, et al. Dynamic wedge field techniques through computer-controlled collimator motion and dose delivery. Med Phys 1990;17:87–91.
28. Kijewski PK, Chin LM, Bjarngard BE. Wedge-shaped dose distributions by computer-controlled collimator motion. Med Phys 1978;5(5):426–429.
29. Ellis F, Hall EJ, Oliver R. A compensator for variations in tissue thickness for high energy beams. Br J Radiol 1959;32:421–422.
30. van de Geijn J, Harrington FS, Fraass B. A graticule for evaluation of megavoltage X-ray port films. Int J Radiat Oncol Biol Phys 1982;8:1999.
31. Herman MG, et al. Clinical use of electronic portal imaging: Report of AAPM RadiationTherapy Committee Task Group 58. Med Phys 2001;28(5):712–737.
32. Antonuk LE. Electronic portal imaging devices: a review and historical perspective of contemporary technologies and research. Phys Med Biol 2002;47(6):R31–R65.
33. Jaffray DA, et al. A radiographic and tomographic imaging system integrated into a medical linear accelerator for localization of bone and soft-tissue targets. Int J Radia Oncol Biol Phys 1999;45:773–789.
34. Jaffray DA, Siewerdsen JH, Wong JW, Martinez AA. Flat-panel cone-beam computed tomography for image-guided radiation therapy. Int J Radia Oncol Biol Phys 2002;53(5): 1337–1349.
35. Mate TP, et al. Principles of AC magnetic fields for objective and continuous target localization in radiation therapy (abstract). Init J Radiat Oncol Biol Phys 2004;60(1):S455.

See also RADIATION THERAPY SIMULATOR; RADIATION PROTECTION INSTRUMENTATION.

ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition

VOLUME 6

Radiotherapy, Heavy Ion – X-Rays, Production of

ENCYCLOPEDIA OF MEDICAL DEVICES AND INSTRUMENTATION, SECOND EDITION

ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition
Volume 6

Radiotherapy, Heavy Ion – X-Rays, Production of

Edited by

John G. Webster
University of Wisconsin–Madison

The *Encyclopedia of Medical Devices and Instrumentation* is available online at http://www.mrw.interscience.wiley.com/emdi

WILEY-INTERSCIENCE

A John Wiley & Sons, Inc., Publication

Published by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Encylopedia of medical devices & instrumentation/by John G. Webster,

editor in chief. – 2nd ed.
p. ; cm.
Rev. ed. of: Encyclopedia of medical devices and instrumentation. 1988.
Includes bibliographical references and index.
ISBN-13 978-0-471-26358-6 (set : cloth)
ISBN-10 0-471-26358-3 (set : cloth)
ISBN-13 978-0-470-04071-3 (v. 6 : cloth)
ISBN-10 0-470-04071-8 (v. 6 : cloth)

1. Medical instruments and apparatus–Encyclopedias. 2. Biomedical engineering–Encyclopedias. 3. Medical physics–Encyclopedias. 4. Medicine–Data processing–Encyclopedias. I. Webster, John G., 1932- . II. Title: Encyclopedia of medical devices and instrumentation.

[DNLM: 1. Equipment and Supplies–Encyclopedias–English. W 13

E555 2006]
R856.A3E53 2006
610.2803–dc22
2005028946

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTRIBUTOR LIST

ABDEL HADY, MAZEN, *McMaster University, Hamilton, Ontario Canada*, Bladder Dysfunction, Neurostimulation of

ABEL, L.A., *University of Melbourne, Melbourne, Australia*, Ocular Motility Recording and Nystagmus

ABREU, BEATRIZ C., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

ALEXANDER, A.L., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

ALI, ABBAS, *University of Illinois, at Urbana-Champaign*, Bioinformatics

ALI, MÜFTÜ, *School of Dental Medicine, Boston, Massachusetts*, Tooth and Jaw, Biomechanics of

ALPERIN, NOAM, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

ANSON, DENIS, *College Misericordia, Dallas, Pennsylvania*, Environmental Control

ARENA, JOHN C., *VA Medical Center and Medical College of Georgia*, Biofeedback

ARIEL, GIDEON, *Ariel Dynamics, Canyon, California*, Biomechanics of Exercise Fitness

ARMSTRONG, STEVE, *University of Iowa, Iowa City, Iowa*, Biomaterials for Dentistry

ASPDEN, R.M., *University of Aberdeen, Aberdeen, United Kingdom*, Ligament and Tendon, Properties of

AUBIN, C.E., *Polytechniquie Montreal, Montreal Quebec, Canada*, Scoliosis, Biomechanics of

AYRES, VIRGINIA M., *Michigan State University, East Lansing, Michigan*, Microscopy, Scanning Tunneling

AZANGWE, G., Ligament and Tendon, Properties of

BACK, LLOYD H., *California Institute of Technology, Pasadena, California*, Coronary Angioplasty and Guidewire Diagnostics

BADYLAK, STEPHEN F., *McGowan Institute for Regenerative Medicine, Pittsburgh, Pennsylvania*, Sterilization of Biologic Scaffold Materials

BANDYOPADHYAY, AMIT, *Washington State University, Pullman, Washington*, Orthopedic Devices, Materials and Design for

BANERJEE, RUPAK K., *University of Cincinnati, Cincinnati, Ohio*, Coronary Angioplasty and Guidewire Diagnostics

BARBOUR, RANDALL L., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

BARKER, STEVEN J., *University of Arizona, Tucson, Arizona*, Oxygen Monitoring

BARTH, ROLF F., *The Ohio State University, Columbus, Ohio*, Boron Neutron Capture Therapy

BECCHETTI, F.D., *University of Michigan, Ann Arbor, Michigan*, Radiotherapy, Heavy Ion

BELFORTE, GUIDO, *Politecnico di Torino – Department of Mechanics*, Laryngeal Prosthetic Devices

BENKESER, PAUL, *Georgia Institute of Technology, Atlanta, Georgia*, Biomedical Engineering Education

BENNETT, JAMES R., *University of Iowa, Iowa City, Iowa*, Digital Angiography

BERSANO-BEGEY, TOMMASO, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

BIGGS, PETER J., *Harvard Medical School, Boston, Massachusetts*, Radiotherapy, Intraoperative

BIYANI, ASHOK, *University of Toledo, and Medical College of Ohio, Toledo, Ohio*, Human Spine, Biomechanics of

BLOCK, W.F., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

BLUE, THOMAS E., *The Ohio State University, Columbus, Ohio*, Boron Neutron Capture Therapy

BLUMSACK, JUDITH T., *Disorders Auburn University, Auburn, Alabama*, Audiometry

BOGAN, RICHARD K., *University of South Carolina, Columbia, South Carolina*, Sleep Laboratory

BOKROS, JACK C., *Medical Carbon Research Institute, Austin, Texas*, Biomaterials, Carbon

BONGIOANNINI, GUIDO, *ENT Division Mauriziano Hospital, Torino, Italy*, Laryngeal Prosthetic Devices

BORAH, JOSHUA, *Applied Science Laboratories, Bedford, Massachusetts*, Eye Movement, Measurement Techniques for

BORDEN, MARK, *Director of Biomaterials Research, Irvine, California*, Biomaterials, Absorbable

BORTON, BETTIE B., *Auburn University Montgomery, Montgomery, Alabama*, Audiometry

BORTON, THOMAS E., *Auburn University Montgomery, Montgomery, Alabama*, Audiometry

BOSE SUSMITA,, *Washington State University, Pullman, Washington*, Orthopedic Devices, Materials and Design for

BOVA, FRANK J., *M. D. Anderson Cancer Center Orlando, Orlando, FL*, Radiosurgery, Stereotactic

BRENNER, DAVID J., *Columbia University Medical Center, New York, New York*, Computed Tomography Screening

BREWER, JOHN M., *University of Georgia*, Electrophoresis

BRIAN, L. DAVIS, *Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, Ohio*, Skin, Biomechanics of

BRITT, L.D., *Eastern Virginia Medical School, Norfolk, Virginia*, Gastrointestinal Hemorrhage

BRITT, R.C., *Eastern Virginia Medical School, Norfolk, Virginia*, Gastrointestinal Hemorrhage

BROZIK, SUSAN M., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

BRUNER, JOSEPH P., *Vanderbilt University Medical Center, Nashville, Tennessee*, Intrauterine Surgical Techniques

BRUNSWIG NEWRING, KIRK A., *University of Nevada, Reno, Nevada*, Sexual Instrumentatio n

BRUYANT, PHILIPPE P., *University of Massachusetts, North Worcester, Massachusetts*, Nuclear Medicine, Computers in

BUNNELL, BERT J., *Bunnell Inc., Salt Lake City, Utah*, High Frequency Ventilation

CALKINS, JERRY M., *Defense Research Technologies, Inc., Rockville, Maryland*, Medical Gas Analyzers

CANNON, MARK, *Northwestern University, Chicago, Illinois*, Resin-Based Composites

CAPPELLERI, JOSEPH C., *Pfizer Inc., Groton, Connecticut*, Quality-of-Life Measures, Clinical Significance of

CARDOSO, JORGE, *University of Madeira, Funchal, Portugal*, Office Automation Systems

CARELLO, MASSIMILIANA, *Politecnicodi Torino – Department of Mechanics*, Laryngeal Prosthetic Devices

CASKEY, THOMAS C., *Cogene Biotech Ventures, Houston, Texas*, Polymerase Chain Reaction

CECCIO, STEVEN, *University of Michigan, Ann Arbor, Michigan*, Heart Valve Prostheses, In Vitro Flow Dynamics of

CHAN, JACKIE K., *Columbia University, New York, New York*, Photography, Medical

CHANDRAN, K.B., *University of Iowa, Iowa City, Iowa*, Heart Valve Prostheses

CHATZANDROULIS, S., *NTUA, Athens, Attiki, Greece*, Capacitive Microsensors for Biomedical Applications

CHAVEZ, ELIANA, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CHEN, HENRY, *Stanford University, Palo Alto, California*, Exercise Stress Testing

CHEN, JIANDE, *University of Texas Medical Branch, Galveston, Texas*, Electrogastrogram

CHEN, YAN, *Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, Ohio*, Skin, Biomechanics of

CHEYNE, DOUGLAS, *Hospital for Sick Children Research Institute*, Biomagnetism

CHUI, CHEN-SHOU, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

CLAXTON, NATHAN S., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

CODERRE, JEFFREY A., *Massachus etts Institute of Technology, Cambridge, Massachusetts*, Boron Neutron Capture Therapy

COLLINS, BETH, *University of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

COLLINS, DIANE, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CONSTANTINOU, C., *Columbia University Radiation Oncology, New York, New York*, Phantom Materials in Radiology

COOK, ALBERT, *University of Alberta*, *Edmonton, Alberta, Canada*, Communication Devices

COOPER, RORY, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CORK, RANDALL C., *Louisiana State University, Shreveport, Louisiana*, Monitoring, Umbilical Artery and Vein, Blood Gas Measurements; Transcuta neous Electrical Nerve Stimulation (TENS); Ambulatory Monitoring

COX, JOSEPHINE H., *Walter Reed Army Institute of Research, Rockville, Maryland*, Blood Collection and Processing

CRAIG, LEONARD, *Feinberg School of Medicine of Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

CRESS, CYNTHIA J., *University of Nebraska, Lincoln, Nebraska*, Communicative Disorders, Computer Applications for

CUMMING, DAVID R.S., *University of Glasgow, Glasgow, United Kingdom*, Ion-Sensitive Field-Effect Transistors

CUNNINGHAM, JOHN R., *Camrose, Alberta, Canada*, Cobalt 60 Units for Radiotherapy

D'ALESSANDRO, DAVID, *Montefiore Medical Center, Bronx, New York*, Heart–Lung Machines

D'AMBRA, MICHAEL N., *Harvard Medical School, Cambridge, Massachusetts*, Cardiac Output, Thermodilution Measurement of

DADSETAN, MAHROKH, *Mayo Clinic, College of Medicine, Rochester, Minnesota*, Microscopy, Electron

DALEY, MICHAEL L., *The University of Memphis, Memphis, Tennessee*, Monitoring, Intracranial Pressure

DAN, LOYD, *Linköping University, Linköping, Sweden*, Thermocouples

DAS, RUPAK, *University of Wisconsin, Madison, Wisconsin*, Brachytherapy, High Dosage Rate

DATTAWADKAR, AMRUTA M., *University of Wisconsin, Madison, Madison, Wisconsin*, Ocular Fundus Reflectometry

DAVIDSON, MICHAEL W., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

DE LUCA, CARLO, *Boston University, Boston, Massachusetts*, Electromyography

DE SALLES, ANTONIO A.F., *UCLA Medical School, Los Angeles, California*, Stereotactic Surgery

DECAU, SABIN, *University of Maryland, School of Medicine*, Shock, Treatment of

DECHOW, PAUL C., *A & M University Health Science Center, Dallas, Texas*, Strain Gages

DELBEKE, JEAN, *Catholique University of Louvain, Brussels, Belgium*, Visual Prostheses

DELL'OSSO, LOUIS F., *Case Western Reserve University, Cleveland, Ohio*, Ocular Motility Recording and Nystagmus

DELORME, ARNAUD, *University of San Diego, La Jolla, California*, Statistical Methods

DEMENKOFF, JOHN, *Mayo Clinic, Scottsdale, Arizona*, Pulmonary Physiology

DEMIR, SEMAHAT S., *The University of Memphis and The University of Tennessee Health Science Center, Memphis, Tennessee*, Electrophysiology

DEMLING, ROBERT H., *Harvard Medical School*, Skin Substitute for Burns, Bioactive

DENNIS, MICHAEL J., *Medical University of Ohio, Toledo, Ohio*, Computed Tomography

DESANTI, LESLIE, *Harvard Medical School*, Skin Substitute for Burns, Bioactive

DEUTSCH, STEVEN, *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

DEVINENI, TRISHUL, *Conemaugh Health System*, Biofeedback

DI BELLA EDWARD, V.R., *University of Utah*, Tracer Kinetics

DIAKIDES, NICHOLAS A., *Advanced Concepts Analysis, Inc., Falls Church, Virginia*, Thermography

DOLAN, PATRICIA L., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

DONOVAN, F.M., *University of South Alabama*, Cardiac Output, Indicator Dilution Measurement of

DOUGLAS, WILSON R., *Children's Hospital of Philadelphia, Philadelphia, Pennsylvania*, Intrauterine Surgical Techniques

DRAPER, CRISSA, *University of Nevada, Reno, Nevada*, Sexual Instrumentation

DRZEWIECKI, TADEUSZ M., *Defense Research Technologies, Inc., Rockville, Maryland*, Medical Gas Analyzers

DURFEE, W.K., *University of Minnesota, Minneapolis, Minnesota*, Rehabilitation and Muscle Testing

DYRO, JOSEPH F., *Setauket, New York*, Safety Program, Hospital

DYSON, MARY, *Herts, United Kingdom*, Heat and Cold, Therapeutic

ECKERLE, JOSEPH S., *SRI International, Menlo Park, California*, Tonometry, Arterial

EDWARDS, BENJAMIN, *University of Wisconsin-Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

EDWARDS, THAYNE L., *University of Washington, Seattle, Washington*, Chromatography

EKLUND, ANDERS, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

EL SOLH, ALI A., *Erie County Medical Center, Buffalo, New York*, Sleep Studies, Computer Analysis of

ELMAYERGI, NADER, *McMaster University, Hamilton, Ontario, Canada*, Bladder Dysfunction, Neurostimulation of

ELSHARYDAH, AHMAD, *Louisiana State University, Baton Rouge, Louisiana*, Ambulatory Monitoring; Monitoring, Umbilical Artery and Vein, Blood Gas Measurements

FADDY, STEVEN C., *St. Vincents Hospital, Sydney, Darlinghurst, Australia*, Cardiac Output, Fick Technique for

FAHEY, FREDERIC H., *Childrens Hospital Boston*, Computed Tomography, Single Photon Emission

FAIN, S.B., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

FELDMAN, JEFFREY, *Childrens Hospital of Philadelphia, Philadelphia, Pennsylvania*, Anesthesia Machines

FELLERS, THOMAS J., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

FERRARA, LISA, *Cleveland Clinic Foundation, Cleveland, Ohio*, Human Spine, Biomechanics of

FERRARI, MAURO, *The Ohio State University, Columbus, Ohio*, Drug Delivery Systems

FONTAINE, ARNOLD A., *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

FOUST, MILTON J., JR, *Medical University of South Carolina Psychiatry and Behavioral Sciences, Charleston, South Carolina*, Electroconvulsive Therapy

FRASCO, PETER, *Mayo Clinic Scottsdale, Scottsdale, Arizona*, Temperature Monitoring

FRAZIER, JAMES, *Louisiana State University, Baton Rouge, Louisiana*, Ambulatory Monitoring

FREIESLEBEN DE BLASIO, BIRGITTE, *University of Oslo, Oslo, Norway*, Impedance Spectroscopy

FRESTA, MASSIMO, *University of Catanzaro Magna Græcia, Germaneto (CZ), Italy*, Drug Delivery Systems

FREYTES, DONALD O., *McGowan Institute for Regenerative Medicine, Pittsburgh Pennsylvania*, Sterilization of Biologic Scaffold Materials

FROEHLICHER, VICTOR, *VA Medical Center, Palo Alto, California*, Exercise Stress Testing

FUNG, EDWARD K., *Columbia University, New York, New York*, Photography, Medical

GAGE, ANDREW A., *State University of New York at Buffalo, Buffalo, New York*, Cryosurgery

GAGLIO, PAUL J., *Columbia University College of Physicians and Surgeons*, Liver Transplantation

GARDNER, REED M., *LDS Hospital and Utah University, Salt Lake City, Utah*, Monitoring, Hemodynamic

GEJERMAN, GLEN, *Hackensack University Medical, Hackensack, New Jersey*, Radiation Therapy, Quality Assurance in

GEORGE, MARK S., *Medical University of South Carolina Psychiatry and Behavioral Sciences, Charleston, South Carolina*, Electroconvulsive Therapy

GHARIEB, R.R., *Infinite Biomedical Technologies, Baltimore, Maryland*, Neurological Monitors

GLASGOW, GLENN P., *Loyola University of Chicago, Maywood, Illinois*, Radiation Protection Instrumentation

GLASGOW, GLENN, *University of Wisconsin-Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

GOEL, VIJAY K., *University of Toledo, and Medical College of Ohio, Toledo, Ohio*, Human Spine, Biomechanics of

GOETSCH, STEVEN J., *San Diego Gamma Knife Center, La Jolla, California*, Gamma Knife

GOLDBERG, JAY R., *Marquette University Milwaukee, Wisconsin*, Minimally Invasive Surgery

GOLDBERG, ZELENNA, *Department of Radiation Oncology, Davis, California*, Ionizing Radiation, Biological Effects of

GOPALAKRISHNAKONE, P., *National University of Singapore, Singapore*, Immunologically Sensitive Field-Effect Transistors

GOPAS, JACOB, *Ben Gurion University of the Negev, Beer Sheva, Israel*, Monoclonal Antibodies

GORGULHO, ALESSANDRA, *UCLA Medical School, Los Angeles, California*, Stereotactic Surgery

GOUGH, DAVID A., *University of California, La Jolla, California*, Glucose Sensors

GOUSTOURIDIS, D., *NTUA, Athens, Attiki, Greece*, Capacitive Microsensors for Biomedical Applications

GRABER, HARRY L., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

GRAÇA, M., *Louisiana State University, Baton Rouge, Louisiana*, Boron Neutron Capture Therapy

GRANT, WALTER III, *Baylor College of Medicine, Houston, Texas*, Radiation Therapy, Intensity Modulated

GRAYDEN, EDWARD, *Mayo Health Center, Albertlea, Minnesota*, Cardiopulmonary Resuscitation

GREEN, JORDAN R., *University of Nebraska, Lincoln, Nebraska*, Communicative Disorders, Computer Applications for

HAEMMERICH, DIETER, *Medical University of South Carolina, Charleston, South Carolina*, Tissue Ablation

HAMAM, HABIB, *Université de Moncton, Moncton New Brunswick, Canada*, Lenses, Intraocular

HAMMOND, PAUL A., *University of Glasgow, Glasgow, United Kingdom*, Ion-Sensitive Field-Effect Transistors

HANLEY, JOSEPH, *Hackensack University Medical, Hackensack, New Jersey*, Radiation Therapy, Quality Assurance in

HARLEY, BRENDAN A., *Massachusetts Institute of Technology*, Skin Tissue Engineering for Regeneration

HARPER, JASON C., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

HASMAN, ARIE, *Maastricht, The Netherlands*, Medical Education, Computers in

HASSOUNA, MAGDY, *Toronto Western Hospital, Toronto, Canada*, Bladder Dysfunction, Neurostimulation of

HAYASHI, KOZABURO, *Okayama University of Science, Okayama, Japan*, Arteries, Elastic Properties of

HENCH, LARRY L., *Imperial College London, London, United Kingdom*, Biomaterials: Bioceramics

HETTRICK, DOUGLAS A., *Sr. Principal Scientist Medtronic, Inc., Minneapolis, Minnesota*, Bioimpedance in Cardiovascular Medicine

HIRSCH-KUCHMA, MELISSA, *University of Central Florida NanoScience Technology Center, Orlando, Florida*, Biosurface Engineering

HOLDER, GRAHAM E., *Moorfields Eye Hospital, London, United Kingdom*, Electroretinography

HOLMES, TIMOTHY, *St. Agnes Cancer Center, Baltimore, Maryland*, Tomotherapy

HONEYMAN-BUCK, JANICE C., *University of Florida, Gainesville, Florida*, Radiology Information Systems

HOOPER, BRETT A., *Areté Associates, Arlington, Virginia*, Endoscopes

HORN, BRUCE, *Kaiser Permanente, Los Angeles, California*, X-Rays Production of

HORNER, PATRICIA I., *Biomedical Engineering Society Landover, Maryland*, Medical Engineering Societies and Organizations

HOROWITZ, PAUL M., *University of Texas, San Antonio, Texas*, Fluorescence Measurements

HOU, XIAOLIN, *Risø National Laboratory, Roskilde, Denmark*, Neutron Activation Analysis

HOVORKA, ROMAN, *University of Cambridge, Cambridge, United Kingdom*, Pancreas, Artificial

HUANG, H.K., *University of Southern California*, Teleradiology

HUNT, ALAN J., *University of Michigan, Ann Arbor, Michigan*, Optical Tweezers

HUTTEN, HELMUT, *University of Technology, Graz, Australia*, Impedance Plethysmography

IAIZZO, P.A., *University of Minnesota, Minneapolis, Minnesota*, Rehabilitation and Muscle Testing

IBBOTT, GEOFFREY S., *Anderson Cancer Center, Houston, Texas*, Radiation Dosimetry, Three-Dimensional

INGHAM, E., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

ISIK, CAN, *Syracuse University, Syracuse, New York*, Blood Pressure Measurement

JAMES, SUSAN P., *Colorado State University, Fort Collins, Colorado*, Biomaterials: Polymers

JENSEN, WINNIE, *Aalborg University, Aalborg, Denmark*, Electroneurography

JIN, CHUNMING, *North Carolina State University, Raleigh, North Carolina*, Biomaterials, Corrosion and Wear of

JIN, Z.M., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

JOHNSON, ARTHUR T., *University of Maryland College Park, Maryland*, Medical Engineering Societies and Organizations

JONES, JULIAN R., *Imperial College London, London, United Kingdom*, Biomaterials: Bioceramics

JOSHI, ABHIJEET, *Abbott Spine, Austin, Texas*, Spinal Implants

JUNG, RANU, *Arizona State University, Tempe, Arizona*, Functional Electrical Stimulation

JURISSON, SILVIA S., *University of Missouri Columbia, Missouri*, Radionuclide Production and Radioactive Decay

KAEDING, PATRICIA J., *Godfrey & Kahn S.C., Madison, Wisconsin*, Codes and Regulations: Medical Devices

KAMATH, CELIA C., *Mayo Clinic, Rochester, Minnesota*, Quality-of-Life Measures, Clinical Significance of

KANE, MOLLIE, *Madison, Wisconsin*, Contraceptive Devices

KATHERINE, ANDRIOLE P., *Harvard Medical School, Boston, Massachusetts*, Picture Archiving and Communication Systems

KATSAGGELOS, AGGELOS K., *Northwestern University, Evanston, Illinois*, DNA Sequencing

KATZ, J. LAWRENCE, *University of Missouri-Kansas City, Kansas City, Missouri*, Bone and Teeth, Properties of

KESAVAN, SUNIL, *Akebono Corporation, Farmington Hills, Michigan*, Linear Variable Differential Transformers

KHANG, GILSON, *Chonbuk National University*, Biomaterials: Tissue Engineering and Scaffolds

KHAODHIAR, LALITA, *Harvard Medical School, Boston, Massachusetts*, Cutaneous Blood Flow, Doppler Measurement of

KIM, MOON SUK, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

KIM, YOUNG KON, *Inje University, Kimhae City, Korea*, Alloys, Shape Memory

KINDWALL, ERIC P., *St. Luke's Medical Center, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

KING, MICHAEL A., *University of Massachusetts, North Worcester, Massachusetts*, Nuclear Medicine, Computers in

KLEBE, ROBERT J., *University of Texas, San Antonio, Texas*, Fluorescence Measurements

KLEIN, BURTON, *Burton Klein Associates, Newton, Massachusetts*, Gas and Vacuum Systems, Centrally Piped Medical

KNOPER, STEVEN R., *University of Arizona College of Medicine*, Ventilatory Monitoring

KONTAXAKIS, GEORGE, *Universidad Politécnica de Madrid, Madrid, Spain*, Positron Emission Tomography

KOTTKE-MARCHANT, KANDICE, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Vascular Graft Prosthesis

KRIPFGANS, OLIVER, *University of Michigan, Ann Arbor, Michigan*, Ultrasonic Imaging

KULKARNI, AMOL D., *University of Wisconsin–Madison, Madison, Wisconsin*, Ocular Fundus Reflectometry, Visual Field Testing

KUMARADAS, J. CARL, *Ryerson University, Toronto, Ontario, Canada*, Hyperthermia, Interstitial

KUNICKA, JOLANTA, *Bayer HealthCare LLC, Tarrytown, New York*, Differential Counts, Automated

KWAK, KWANJ JOO, *University of Miami Miller School of Medicine, Miami, Florida*, Microscopy, Scanning Force

LAKES, RODERIC, *University of Wisconsin-Madison*, Bone and Teeth, Properties of

LAKKIREDDY, DHANUNJAYA, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

LARSEN, COBY, *Case Western Reserve University, Cleveland, Ohio*, Vascular Graft Prosthesis

LASTER, BRENDA H., *Ben Gurion University of the Negev, Beer Sheva, Israel*, Monoclonal Antibodies

LATTA, LOREN, *University of Miami, Coral Gables, Florida*, Rehabilitation, Orthotics in

LEDER, RON S., *Universidad Nacional Autonoma de Mexico Mexico, Distrito Federal*, Continuous Positive Airway Pressure

LEE, CHIN, *Harvard Medical School, Boston, Massachusetts*, Radiotherapy Treatment Planning, Optimization of; Hyperthermia, Interstitial

LEE, HAI BANG, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

LEE, SANG JIN, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

LEI, LIU, *Department of General Engineering, Urbana, Illinois*, Bioinformatics

LEI, XING, *Stanford University, Stanford, California*, Radiation Dose Planning, Computer-Aided

LEWIS, MATTHEW C., *Medical College of Wisconsin, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

LI, CHAODI, *University of Notre Dame, Notre Dame, Indiana*, Bone Cement, Acrylic

LI, JONATHAN G., *University of Florida, Gainesville, Florida*, Radiation Dose Planning, Computer-Aided

LI, QIAO, *University of Michigan, Ann Arbor, Michigan*, Immunotherapy

LI, YANBIN, *University of Arkansas, Fayetteville, Arkansas*, Piezoelectric Sensors

LIBOFF, A.R., *Oakland University, Rochester, Michigan*, Bone Ununited Fracture and Spinal Fusion, Electrical Treatment of

LIGAS, JAMES, *University of Connecticut, Farmington, Connecticut*, Respiratory Mechanics and Gas Exchange

LIMOGE, AIME, *The René Descartes University of Paris, Paris, France*, Electroanalgesia, Systemic

LIN, PEI-JAN PAUL, *Beth Israel Deaconess Medical Center, Boston, Massachusets*, Mammography

LIN, ZHIYUE, *University of Kansas Medical Center, Kansas City, Kansas*, Electrogastrogram

LINEAWEAVER, WILLIAM C., *Unive rsity of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

LIPPING, TARMO, *Tampere University of Technology, Pori, Finland*, Monitoring in Anesthesia

LIU, XIAOHUA, *The University of Michigan, Ann Arbor, Michigan*, Polymeric Materials

LLOYD, J.J., *Regional Medical Physics Department, Newcastle-upon-Tyne, United Kingdom*, Ultraviolet Radiation in Medicine

LOEB, ROBERT, *University of Arizona, Tuscon, Arizona*, Anesthesia Machines

LOPES DE MELO, PEDRO, *State University of Rio de Janeiro, Térreo Salas, Maracanã, Thermistors*

LOUDON, ROBERT G., Lung Sounds

LOW, DANIEL A., *Washington University School of Medicine, St. Louis, Missouri*, Radiation Therapy Simulator

LU, LICHUN, *Mayo Clinic, College of Medicine, Rochester, Minnesota*, Microscopy, Electron

LU, ZHENG FENG, *Columbia University, New York, New York*, Screen-Film Systems

LYON, ANDREW W., *University of Calgary, Calgary, Canada*, Flame Atomic Emission Spectrometry and Atomic Absorption Spectrometry

LYON, MARTHA E., *University of Calgary, Calgary, Canada*, Flame Atomic Emission Spectrometry and Atomic Absorption Spectrometry

MA, C-M CHARLIE, *Fox Chase Cancer Center, Philadelphia, Pennsylvania*, X-Ray Therapy Equipment, Low and Medium Energy

MACIA, NARCISO F., *Arizona State University at the Polytechnic Campus, Mesa, Arizona*, Pneumotachometers

MACKENZIE, COLIN F., *University of Maryland, School of Medicine*, Shock, Treatment of

MACKIE, THOMAS R., *University of Wisconsin, Madison, Wisconsin*, Tomotherapy

MADNANI, ANJU, *LSU Medical Centre, Shreveport, Louisiana*, Transcutaneous Electrical Nerve Stimulation (TENS)

MADNANI, SANJAY, *LSU Medical Centre, Shreveport, Louisiana*, Transcutaneous Electrical Nerve Stimulation (TENS)

MADSEN, MARK T., *University of Iowa, Iowa City, Iowa*, Anger Camera

MAGNANO, MAURO, *ENT Division Mauriziano Hospital, Torino, Italy*, Drug Delivery Systems

MANDEL, RICHARD, *Boston University School of Medicine, Boston, Massachusetts*, Colorimetry

MANNING, KEEFE B., *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

MAO, JEREMY J., *University of Illinois at Chicago, Chicago, Illinois*, Cartilage and Meniscus, Properties of

MARCOLONGO, MICHELE, *Drexel University, Philadelphia, Pennsylvania*, Spinal Implants

MAREK, MIROSLAV, *Georgia Institute of Technology, Atlanta, Georgia*, Biomaterials, Corrosion and Wear of

MARION, NICHOLAS W., *University of Illinois at Chicago, Chicago, Illinois*, Cartilage and Meniscus, Properties of

MASTERS, KRISTYN S., *University of Wisconsin, Madison, Wisconsin*, Tissue Engineering

MAUGHAN, RICHARD L., *Hospital of the University of Pennsylvania*, Neutron Beam Therapy

MCADAMS, ERIC, *University of Ulster at Jordanstown, Newtownabbey, Ireland*, Bioelectrodes

MCARTHUR, SALLY L., *University of Sheffield, Sheffield, United Kingdom*, Biomaterials, Surface Properties of

MCEWEN, MALCOM, *National Research Council of Canada, Ontario, Canada*, Radiation Dosimetry for Oncology

MCGOWAN, EDWARD J., *E.J. McGowan & Associates*, Biofeedback

MCGRATH, SUSAN, *Dartmouth College, Hanover, New Hampshire*, Oxygen Analyzers

MEEKS, SANFORD L., *University of Florida, Gainesville, Florida*, Radiosurgery, Stereotactic

MELISSA, PETER, *University of Central Florida NanoScience Technology Center, Orlando, Florida*, Biosurface Engineering

MENDELSON, YITZHAK, *Worcester Polytechnic Institute*, Optical Sensors

METZKER, MICHAEL L., *Baylor College of Medicine, Houston, Texas*, Polymerase Chain Reaction

MEYEREND, M.E., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

MICHLER, ROBERT, *Montefiore Medical Center, Bronx, New York*, Heart–Lung Machines

MICIC, MIODRAG, *MP Biomedicals LLC, Irvine, California*, Microscopy and Spectroscopy, Near-Field

MILLER, WILLIAM, *University of Missouri Columbia, Missouri*, Radionuclide Production and Radioactive Decay

MITTRA, ERIK, *Stony Brook University, New York*, Bone Density Measurement

MODELL, MARK, *Harvard Medical School, Boston, Massachusetts*, Fiber Optics in Medicine

MORE, ROBERT B., *RBMore Associates, Austin, Texas* Biomaterials Carbon

MORE, ROBERT, *Austin, Texas*, Heart Valves, Prosthetic

MORROW, DARREN, *Royal Adelaide Hospital, Adelaide, Australia*, Intraaortic Balloon Pump

MOURTADA, FIRAS, *MD Anderson Cancer Center, Houston, Texas*, Brachytherapy, Intravascular

MOY, VINCENT T., *University of Miami, Miller School of Medicine, Miami, Florida*, Microscopy, Scanning Force

MÜFTÜ, SINAN, *Northeastern University, Boston, Massachusetts*, Tooth and Jaw, Biomechanics of

MURPHY, RAYMOND L.H., Lung Sounds

MURPHY, WILLIAM L., *University of Wisconsin, Madison, Wisconsin*, Tissue Engineering

MURRAY, ALAN, *Newcastle University Medical Physics, Newcastle upon Tyne, United Kingdom*, Pace makers

MUTIC, SASA, *Washington University School of Medicine, St. Louis, Missouri*, Radiation Therapy Simulator

NARAYAN, ROGER J., *University of North Carolina, Chapel Hill, North Carolina*, Biomaterials, Corrosion and Wear of

NATALE, ANDREA, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

NAZERAN, HOMER, *The University of Texas, El Paso, Texas*, Electrocardiography, Computers in

NEUMAN, MICHAEL R., *Michigan Technological University, Houghton, Houghton, Michigan*, Fetal Monitoring, Neonatal Monitoring

NEUZIL, PAVEL, *Institute of Bioengineering and Nanotechnology, Singapore*, Immunologically Sensitive Field-Effect Transistors

NICKOLOFF, EDWARD L., *Columbia University, New York, New York*, X-Ray Quality Control Program

NIEZGODA, JEFFREY A., *Medical College of Wisconsin, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

NISHIKAWA, ROBERT M., *The University of Chicago, Chicago, Illinois*, Computer-Assisted Detection and Diagnosis

NUTTER, BRIAN, *Texas Tech University, Lubbock, Texas*, Medical Records, Computers in

O'DONOHUE, WILLIAM, *University of Nevada, Reno, Nevada*, Sexual Instrumentation

ORTON, COLIN, *Harper Hospital and Wayne State University, Detroit, Michigan*, Medical Physics Literature

OZCELIK, SELAHATTIN, *Texas A&M University, Kingsville, Texas*, Drug Infusion Systems

PANITCH, ALYSSA, *Arizona State University, Tempe, Arizona*, Biomaterials: An Overview

PAOLINO, DONATELLA, *University of Catanzaro Magna Græcia, Germaneto (CZ), Italy*, Drug Delivery Systems

PAPAIOANNOU, GEORGE, *University of Wisconsin, Milwaukee, Wisconsin*, Joints, Biomechanics of

PARK, GRACE E., *Purdue University, West Lafayette, Indiana*, Porous Materials for Biological Applications

PARMENTER, BRETT A., *State University of New York at Buffalo, Buffalo, New York*, Sleep Studies, Computer Analysis of

PATEL, DIMPI, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

PEARCE, JOHN, *The University of Texas, Austin, Texas*, Electrosurgical Unit (ESU)

PELET, SERGE, *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

PERIASAMY, AMMASI, *University of Virginia, Charlottesville, Virginia*, Cellular Imaging

PERSONS, BARBARA L., *University of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

PIPER, IAN, *The University of Memphis, Memphis, Tennessee*, Monitoring, Intracranial Pressure

POLETTO, CHRISTOPHER J., *National Institutes of Health*, Tactile Stimulation

PREMINGER, GLENN M., *Duke University Medical Center, Durham, North Carolina*, Lithotripsy

PRENDERGAST, PATRICK J., *Trinity College, Dublin, Ireland*, Orthopedics, Prosthesis Fixation for

PREVITE, MICHAEL, *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

PURDY, JAMES A., *UC Davis Medical Center, Sacramento, California*, Radiotherapy Accessories

QI, HAIRONG, *Advanced Concepts Analysis, Inc., Falls Church, Virginia*, Thermography

QIN, YIXIAN, *Stony Brook University, New York*, Bone Density Measurement

QUAN, STUART F., *University of Arizona, Tucson, Arizona*, Ventilatory Monitoring

QUIROGA, RODRIGO QUIAN, *University of Leicester, Leicester, United Kingdom*, Evoked Potentials

RAHAGHI, FARBOD N., *University of California, La Jolla, California*, Glucose Sensors

RAHKO, PETER S., *University of Wisconsin Medical School*, Echocardiography and Doppler Echocardiography

RALPH, LIETO, *University of Wisconsin–Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

RAMANATHAN, LAKSHMI, *Mount Sinai Medical Center*, Analytical Methods, Automated

RAO, SATISH S.C., *University of Iowa College of Medicine, Iowa City, Iowa*, Anorectal Manometry

RAPOPORT, DAVID M., *NYU School of Medicine, New York, New York*, Continuous Positive Airway Pressure

REBELLO, KEITH J., *The Johns Hopkins University Applied Physics Lab, Laurel, Maryland*, Micro surgery

REDDY, NARENDER, *The University of Akron, Akron, Ohio*, Linear Variable Differential Transformers

REN-DIH, SHEU, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

RENGACHARY, SETTI S., *Detroit, Michigan*, Human Spine, Biomechanics of

REPPERGER, DANIEL W., *Wright-Patterson Air Force Base, Dayton, Ohio*, Human Factors in Medical Devices

RITCHEY, ERIC R., *The Ohio State University, Columbus, Ohio*, Contact Lenses

RIVARD, MARK J., *Tufts New England Medical Center, Boston, Massachusetts*, Imaging Devices

ROBERTSON, J. DAVID, *University of Missouri, Columbia, Missouri*, Radionuclide Production and Radioactive Decay

ROTH, BRADLEY J., *Oakland University, Rochester, Michigan*, Defibrillators

ROWE-HORWEGE, R. WANDA, *University of Texas Medical School, Houston, Texas*, Hyperthermia, Systemic

RUMSEY, JOHN W., *University of Central Florida, Orlando, Florida*, Biosurface Engineering

RUTKOWSKI, GREGORY E., *University of Minnesota, Duluth, Minnesota*, Engineered Tissue

SALATA, O.V., *University of Oxford, Oxford, United Kingdom*, Nanoparticles

SAMARAS, THEODOROS, *Aristotle University of Thessaloniki Department of Physics, Thessaloniki, Greece*, Thermometry

SANGOLE, ARCHANA P., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

SARKOZI, LASZLO, *Mount Sinai School of Medicine*, Analytical Methods, Automated

SCHEK, HENRY III, *University of Michigan, Ann Arbor, Michigan*, Optical Tweezers

SCHMITZ, CHRISTOPH H., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

SCHUCKERS, STEPHANIE A.C., *Clarkson University, Potsdam, New York*, Arrhythmia Analysis, Automated

SCOPE, KENNETH, *Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

SCOTT, ADZICK N., *University of Pennsylvania, Philadelphia, Pennsylvania*, Intrauterine Surgical Techniques

SEAL, BRANDON L., *Arizona State University, Tempe, Arizona*, Biomaterials: An Overview

SEALE, GARY, *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

SEGERS, PATRICK, *Ghent University, Belgium*, Hemodynamics

SELIM, MOSTAFA A., *Cleveland Metropolitan General Hospital, Palm Coast, Florida*, Colposcopy

SETHI, ANIL, *Loyola University Medical Center, Maywood, Illinois*, X-Rays: Interaction with Matter

SEVERINGHAUS, JOHN W., *University of California in San Francisco*, CO_2 Electrodes

SHALODI, ABDELWAHAB D., *Cleveland Metropolitan General Hospital, Palm Coast, Florida*, Colposcopy

SHANMUGASUNDARAM, SHOBANA, *New Jersey Institute of Technology, Newark, New Jersey*, Polymeric Materials

SHARD, ALEXANDER G., *University of Sheffield, Sheffield United Kingdom*, Biomaterials, Surface Properties of

SHEN, LI-JIUAN, *National Taiwan University School of Pharmacy, Taipei, Taiwan*, Colorimetry

SHEN, WEI-CHIANG, *University of Southern California School of Pharmacy, Los Angeles, California*, Colorimetry

SHERAR, MICHAEL D., *London Health Sciences Centre and University of Western Ontario, London, Ontario, Canada*, Hyperthermia, Interstitial

SHERMAN, DAVID, *The Johns Hopkins University, Baltimore, Maryland*, Electroencephalography

SHI, DONGLU, *University of Cincinnati, Cincinnati, Ohio*, Biomaterials, Testing and Structural Properties of

SHUCARD, DAVID W.M., *State University of New York at Buffalo, Buffalo, New York*, Sleep Studies, Computer Analysis of

SIEDBAND, MELVIN P., *University of Wisconsin, Madison, Wisconsin*, Image Intensifiers and Fluoroscopy

SILBERMAN, HOWARD, *University of Southern California, Los Angeles, California*, Nutrition, Parenteral

SILVERMAN, GORDON, *Manhattan College*, Computers in the Biomedical Laboratory

SILVERN, DAVID A., *Medical Physics Unit, Rabin Medical Center, Petah Tikva, Israel*, Prostate Seed Implants

SINHA, PIYUSH, *The Ohio State University, Columbus, Ohio*, Drug Delivery Systems

SINHA, ABHIJIT ROY, *University of Cincinnati, Cincinnati, Ohio*, Coronary Angioplasty and Guidewire Diagnostics

SINKJÆR, THOMAS, *Aalborg University, Aalborg, Denmark*, Electroneurography

SLOAN, JEFFREY A., *Mayo Clinic, Rochester, Minnesota*, Quality-of-Life Measures, Clinical Significance of

SO, PETER T.C., *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

SOBOL, WLAD T., *University of Alabama at Birmingham Health System, Birmingham, Alabama*, Nuclear Magnetic Resonance Spectroscopy

SOOD, SANDEEP, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

SPECTOR, MYRON, *Brigham and Women's Hospital, Boston, Massachusetts*, Biocompatibility of Materials

SPELMAN, FRANCIS A., *University of Washington*, Cochlear Prostheses

SRINIVASAN, YESHWANTH, *Texas Tech University, Lubbock, Texas*, Medical Records, Computers in

SRIRAM, NEELAMEGHAM, *University of Buffalo, Buffalo, New York*, Cell Counters, Blood

STARKO, KENTON R., *Point Roberts, Washington*, Physiological Systems Modeling

STARKSCHALL, GEORGE, *The University of Texas*, Radiotherapy, Three-Dimensional Conformal

STAVREV, PAVEL, *Cross Cancer Institute, Edmonton, Alberta, Canada*, Radiotherapy Treatment Planning, Optimization of

STENKEN, JULIE A., *Rensselaer Polytechnic Institute, Troy, New York*, Microdialysis Sampling

STIEFEL, ROBERT, *University of Maryland Medical Center, Baltimore, Maryland*, Equipment Acquisition

STOKES, I.A.F., *Polytechniquie Montreal, Montreal Quebec, Canada*, Scoliosis, Biomechanics of

STONE, M.H., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

SU, XIAO-LI, *BioDetection Instruments LLC, Fayetteville, Arkansas*, Piezoelectric Sensors

SUBHAN, ARIF, *Masterplan Technology Management, Chatsworth, California*, Equipment Maintenance, Biomedical

SWEENEY, JAMES D., *Arizona State University, Tempe, Arizona*, Functional Electrical Stimulation

SZETO, ANDREW Y.J., *San Diego State University, San Diego, California*, Blind and Visually Impaired, Assistive Technology for

TAKAYAMA, SHUICHI, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

TAMUL, PAUL C., *Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

TAMURA, TOSHIYO, *Chiba University School of Engineering, Chiba, Japan*, Home Health Care Devices

TANG, XIANGYANG, *GE Healthcare Technologies, Wankesha, Wisconsin*, Computed Tomography Simulators

TAYLOR, B.C., *The University of Akron, Akron, Ohio*, Cardiac Output, Indicator Dilution Measurement of

TEMPLE, RICHARD O., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

TEN, STANLEY, *Salt Lake City, Utah*, Electroanalgesia, Systemic

TERRY, TERESA M., *Walter Reed Army Institute of Research, Rockville, Maryland*, Blood Collection and Processing

THAKOR, N.V., *Johns Hopkins University, Baltimore, Maryland*, Neurological Monitors

THIERENS, HUBERT M.A., *University of Ghent, Ghent, Belgium*, Radiopharmaceutical Dosimetry

THOMADSEN, BRUCE, *University of Wisconsin–Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

TIPPER, J.L., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

TOGAWA, TATSUO, *Waseda University, Saitama, Japan*, Integrated Circuit Temperature Sensor

TORNAI, MARTIN, *Duke University, Durham, North Carolina*, X-Ray Equipment Design

TRAN-SON-TAY, ROGER, *University of Florida, Gainesville, Florida*, Blood Rheology

YENI, YENER N., *Henry Ford Hospital, Detroit, Michigan*, Joints, Biomechanics of

YLI-HANKALA, ARVI, *Tampere University of Technology, Pori, Finland*, Monitoring in Anesthesia

YOKO, KAMOTANI, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

YOON, KANG JI, *Korea Institute of Science and Technology, Seoul, Korea*, Micropower for Medical Applications

YORKE, ELLEN, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

YOSHIDA, KEN, *Aalborg University, Aalborg, Denmark*, Electroneurography

YOUNGSTEDT, SHAWN D., *University of South Carolina, Columbia, South Carolina*, Sleep Laboratory

YU, YIH-CHOUNG, *Lafayette College, Easton, Pennsylvania*, Blood Pressure, Automatic Control of

ZACHARIAH, EMMANUEL S., *University of Medicine and Dentistry of New Jersey, New Brunswick, New Jersey*, Immunologically Sensitive Field-Effect Transistors

ZAIDER, MARCO, *Memorial Sloan Kettering Cancer Center, New York, New York*, Prostate Seed Implants

ZAPANTA, CONRAD M., *Penn State College of Medicine, Hershey, Pennsylvania*, Heart, Artificial

ZARDENETA, GUSTAVO, *University of Texas, San Antonio, Texas*, Fluorescence Measurements

ZELMANOVIC, DAVID, *Bayer HealthCare LLC, Tarrytown, New York*, Differential Counts, Automated

ZHANG, MIN, *University of Washington, Seattle, Washington*, Biomaterials: Polymers

ZHANG, YI, *University of Buffalo, Buffalo, New York*, Cell Counters, Blood

ZHU, XIAOYUE, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

ZIAIE, BABAK, *Purdue University, W. Lafayette, Indiana*, Biotelemetry

ZIELINSKI, TODD M., *Medtronic, Inc., Minneapolis, Minnesota*, Bioimpedance in Cardiovascular Medicine

ZIESSMAN, HARVEY A., *Johns Hopkins University*, Computed Tomography, Single Photon Emission

PREFACE

This six-volume work is an alphabetically organized compilation of almost 300 articles that describe critical aspects of medical devices and instrumentation.

It is comprehensive. The articles emphasize the contributions of engineering, physics, and computers to each of the general areas of anesthesiology, biomaterials, burns, cardiology, clinical chemistry, clinical engineering, communicative disorders, computers in medicine, critical care medicine, dermatology, dentistry, ear, nose, and throat, emergency medicine, endocrinology, gastroenterology, genetics, geriatrics, gynecology, hematology, heptology, internal medicine, medical physics, microbiology, nephrology, neurology, nutrition, obstetrics, oncology, ophthalmology, orthopedics, pain, pediatrics, peripheral vascular disease, pharmacology, physical therapy, psychiatry, pulmonary medicine, radiology, rehabilitation, surgery, tissue engineering, transducers, and urology.

The discipline is defined through the synthesis of the core knowledge from all the fields encompassed by the application of engineering, physics, and computers to problems in medicine. The articles focus not only on what is now useful but also on what is likely to be useful in future medical applications.

These volumes answer the question, "What are the branches of medicine and how does technology assist each of them?" rather than "What are the branches of technology and how could each be used in medicine?" To keep this work to a manageable length, the practice of medicine that is unassisted by devices, such as the use of drugs to treat disease, has been excluded.

The articles are accessible to the user; each benefits from brevity of condensation instead of what could easily have been a book-length work. The articles are designed not for peers, but rather for workers from related fields who wish to take a first look at what is important in the subject.

The articles are readable. They do not presume a detailed background in the subject, but are designed for any person with a scientific background and an interest in technology. Rather than attempting to teach the basics of physiology or Ohm's law, the articles build on such basic concepts to show how the worlds of life science and physical science meld to produce improved systems. While the ideal reader might be a person with a Master's degree in biomedical engineering or medical physics or an M.D. with a physical science undergraduate degree, much of the material will be of value to others with an interest in this growing field. High school students and hospital patients can skip over more technical areas and still gain much from the descriptive presentations.

The *Encyclopedia of Medical Devices and Instrumentation* is excellent for browsing and searching for those new divergent associations that may advance work in a peripheral field. While it can be used as a reference for facts, the articles are long enough that they can serve as an educational instrument and provide genuine understanding of a subject.

One can use this work just as one would use a dictionary, since the articles are arranged alphabetically by topic. Cross references assist the reader looking for subjects listed under slightly different names. The index at the end leads the reader to all articles containing pertinent information on any subject. Listed on pages xxi to xxx are all the abbreviations and acronyms used in the *Encyclopedia.* Because of the increasing use of SI units in all branches of science, these units are provided throughout the *Encyclopedia* articles as well as on pages xxxi to xxxv in the section on conversion factors and unit symbols.

I owe a great debt to the many people who have contributed to the creation of this work. At John Wiley & Sons, Encyclopedia Editor George Telecki provided the idea and guiding influence to launch the project. Sean Pidgeon was Editorial Director of the project. Assistant Editors Roseann Zappia, Sarah Harrington, and Surlan Murrell handled the myriad details of communication between publisher, editor, authors, and reviewers and stimulated authors and reviewers to meet necessary deadlines.

My own background has been in the electrical aspects of biomedical engineering. I was delighted to have the assistance of the editorial board to develop a comprehensive encyclopedia. David J. Beebe suggested cellular topics such as microfluidics. Jerry M. Calkins assisted in defining the chemically related subjects, such as anesthesiology. Michael R. Neuman suggested subjects related to sensors, such as in his own work—neonatology. Joon B. Park has written extensively on biomaterials and suggested related subjects. Edward S. Sternick provided many suggestions from medical physics. The Editorial Board was instrumental both in defining the list of subjects and in suggesting authors.

This second edition brings the field up to date. It is available on the web at http://www.mrw.interscience.wiley.com/emdi, where articles can be searched simultaneously to provide rapid and comprehensive information on all aspects of medical devices and instrumentation.

JOHN G. WEBSTER
University of Wisconsin, Madison

LIST OF ARTICLES

ABBREVIATIONS AND ACRONYMS

AAMI	Association for the Advancement of Medical Instrumentation
AAPM	American Association of Physicists in Medicine
ABC	Automatic brightness control
ABET	Accreditation board for engineering training
ABG	Arterial blood gases
ABLB	Alternative binaural loudness balance
ABS	Acrylonitrile–butadiene–styrene
ac	Alternating current
AC	Abdominal circumference; Affinity chromatography
ACA	Automated clinical analyzer
ACES	Augmentative communication evaluation system
ACL	Anterior chamber lens
ACLS	Advanced cardiac life support
ACOG	American College of Obstetrics and Gynecology
ACR	American College of Radiology
ACS	American Cancer Society; American College of Surgeons
A/D	Analog-to-digital
ADC	Agar diffusion chambers; Analog-to-digital converter
ADCC	Antibody-dependent cellular cytotoxicity
ADCL	Accredited Dosimetry Calibration Laboratories
ADP	Adenosine diphosphate
A-D-T	Admission, discharge, and transfer
AE	Anion exchange; Auxiliary electrode
AEA	Articulation error analysis
AEB	Activation energy barrier
AEC	Automatic exposure control
AED	Automatic external defibrillator
AEMB	Alliance for Engineering in Medicine and Biology
AES	Auger electron spectroscopy
AESC	American Engineering Standards Committee
AET	Automatic exposure termination
AFO	Ankle-foot orthosis
AGC	Automatic gain control
AHA	American Heart Association
AI	Arterial insufficiency
AICD	Automatic implantable cardiac defibrillator
AID	Agency for International Development
AIDS	Acquired immune deficiency syndrome
AL	Anterior leaflet
ALG	Antilymphocyte globulin
ALS	Advanced life support; Amyotropic lateral sclerosis
ALT	Alanine aminotransferase
ALU	Arithmetic and logic unit
AM	Amplitude modulation
AMA	American Medical Association
amu	Atomic mass units
ANOVA	Analysis of variance
ANSI	American National Standards Institute
AP	Action potential; Alternative pathway; Anteroposterior
APD	Anterioposterior diameter
APL	Adjustable pressure limiting valve; Applied Physics Laboratory
APR	Anatomically programmed radiography
AR	Amplitude reduction; Aortic regurgitation; Autoregressive
Ara-C	Arabinosylcytosine
ARD	Absorption rate density
ARDS	Adult respiratory distress syndrome
ARGUS	Arrhythmia guard system
ARMA	Autoregressive-moving-average model
ARMAX	Autoregressive-moving-average model with external inputs
AS	Aortic stenosis
ASA	American Standards Association
ASCII	American standard code for information interchange
ASD	Antisiphon device
ASHE	American Society for Hospital Engineering
ASTM	American Society for Testing and Materials
AT	Adenosine-thiamide; Anaerobic threshold; Antithrombin
ATA	Atmosphere absolute
ATLS	Advanced trauma life support
ATN	Acute tubular necrosis
ATP	Adenosine triphosphate
ATPD	Ambient temperature pressure dry
ATPS	Ambient temperature pressure saturated
ATR	Attenuated total reflection
AUC	Area under curve
AUMC	Area under moment curve
AV	Atrioventricular
AZT	Azido thymidine
BA	Biliary atresia
BAEP	Brainstem auditory evoked potential
BAPN	Beta-amino-proprionitryl
BAS	Boston anesthesis system
BASO	Basophil
BB	Buffer base
BBT	Basal body temperature

BCC	Body-centered cubic
BCD	Binary-coded decimal
BCG	Ballistocardiogram
BCLS	Basic cardiac life support
BCRU	British Commitee on Radiation Units and Measurements
BDI	Beck depression inventory
BE	Base excess; Binding energy
BET	Brunauer, Emmett, and Teller methods
BH	His bundle
BI	Biological indicators
BIH	Beth Israel Hospital
BIPM	International Bureau of Weights and Measurements
BJT	Bipolar junction transistor
BMDP	Biomedical Programs
BME	Biomedical engineering
BMET	Biomedical equipment technician
BMO	Biomechanically optimized
BMR	Basal metabolic rate
BOL	Beginning of life
BP	Bereitschafts potential; Break point
BR	Polybutadiene
BRM	Biological response modifier
BRS	Bibliographic retrieval services
BSS	Balanced salt solution
BTG	Beta thromboglobulin
BTPS	Body temperature pressure saturated
BUN	Blood urea nitrogen
BW	Body weight
CA	Conductive adhesives
CABG	Coronary artery by-pass grafting
CAD/CAM	Computer-aided design/computer-aided manufacturing
CAD/D	Computer-aided drafting and design
CADD	Central axis depth dose
CAI	Computer assisted instruction; Computer-aided instruction
CAM	Computer-assisted management
cAMP	Cyclic AMP
CAPD	Continuous ambulatory peritoneal dialysis
CAPP	Child amputee prosthetic project
CAT	Computerized axial tomography
CATS	Computer-assisted teaching system; Computerized aphasia treatment system
CAVH	Continuous arteriovenous hemofiltration
CB	Conjugated bilirubin; Coulomb barrier
CBC	Complete blood count
CBF	Cerebral blood flow
CBM	Computer-based management
CBV	Cerebral blood volume
CC	Closing capacity
CCC	Computer Curriculum Company
CCD	Charge-coupled device
CCE	Capacitance contact electrode
CCF	Cross-correlation function
CCL	Cardiac catheterization laboratory
CCM	Critical care medical services
CCPD	Continuous cycling peritoneal dialysis
CCTV	Closed circuit television system
CCU	Coronary care unit; Critical care unit
CD	Current density
CDR	Complimentary determining region
CDRH	Center for Devices and Radiological Health
CEA	Carcinoembryonic antigen
CF	Conversion factor; Cystic fibrosis
CFC	Continuous flow cytometer
CFR	Code of Federal Regulations
CFU	Colony forming units
CGA	Compressed Gas Association
CGPM	General Conference on Weights and Measures
CHO	Carbohydrate
CHO	Chinese hamster ovary
CI	Combination index
CICU	Cardiac intensive care unit
CIF	Contrast improvement factor
CIN	Cervical intraepithelial neoplasia
CK	Creatine kinase
CLAV	Clavicle
CLSA	Computerized language sample analysis
CM	Cardiomyopathy; Code modulation
CMAD	Computer managed articulation diagnosis
CMI	Computer-managed instruction
CMRR	Common mode rejection ratio
CMV	Conventional mechanical ventilation; Cytomegalovirus
CNS	Central nervous system
CNV	Contingent negative variation
CO	Carbon monoxide; Cardiac output
COBAS	Comprehensive Bio-Analysis System
COPD	Chronic obstructive pulmonary disease
COR	Center of rotation
CP	Cerebral palsy; Closing pressure; Creatine phosphate
CPB	Cardiopulmonary bypass
CPET	Cardiac pacemaker electrode tips
CPM	Computerized probe measurements
CPP	Cerebral perfusion pressure; Cryoprecipitated plasma
CPR	Cardiopulmonary resuscitation
cps	Cycles per second
CPU	Central Processing unit
CR	Center of resistance; Conditioned response; Conductive rubber; Creatinine
CRBB	Complete right bundle branch block
CRD	Completely randomized design
CRL	Crown rump length
CRT	Cathode ray tube
CS	Conditioned stimulus; Contrast scale; Crown seat
CSA	Compressed spectral array
CSF	Cerebrospinal fluid
CSI	Chemical shift imaging
CSM	Chemically sensitive membrane
CT	Computed tomography; Computerized tomography
CTI	Cumulative toxicity response index
CV	Closing volume

C.V.	Coefficient of variation
CVA	Cerebral vascular accident
CVP	Central venous pressure
CVR	Cardiovascular resistance
CW	Continuous wave
CWE	Coated wire electrodes
CWRU	Case Western Reserve University
DAC	Digital-to-analog converter
DAS	Data acquisition system
dB	Decibel
DB	Direct body
DBMS	Data base management system
DBS	Deep brain stimulation
dc	Direct current
DCCT	Diabetes control and complications trial
DCP	Distal cavity pressure
DCS	Dorsal column stimulation
DDC	Deck decompression chamber
DDS	Deep diving system
DE	Dispersive electrode
DEN	Device experience network
DERS	Drug exception ordering system
DES	Diffuse esophageal spasm
d.f.	Distribution function
DHCP	Distributed Hospital Computer Program
DHE	Dihematoporphyrin ether
DHEW	Department of Health Education and Welfare
DHHS	Department of Health and Human Services
DHT	Duration of hypothermia
DI	Deionized water
DIC	Displacement current
DIS	Diagnostic interview schedule
DL	Double layer
DLI	Difference lumen for intensity
DM	Delta modulation
DME	Dropping mercury electrode
DN	Donation number
DNA	Deoxyribonucleic acid
DOF	Degree of freedom
DOS	Drug ordering system
DOT-NHTSA	Department of Transportation Highway Traffic Safety Administration
DPB	Differential pencil beam
DPG	Diphosphoglycerate
DQE	Detection quantum efficiency
DRESS	Depth-resolved surface coil spectroscopy
DRG	Diagnosis-related group
DSA	Digital subtraction angiography
DSAR	Differential scatter-air ratio
DSB	Double strand breaks
DSC	Differential scanning calorimetry
D-T	Deuterium-on-tritium
DTA	Differential thermal analysis
d.u.	Density unit
DUR	Duration
DVT	Deep venous thrombosis
EA	Esophageal accelerogram
EB	Electron beam
EBCDIC	Extended binary code decimal interchange code
EBS	Early burn scar
EBV	Epstein–Barr Virus
EC	Ethyl cellulose
ECC	Emergency cardiac care; Extracorporeal circulation
ECCE	Extracapsular cataract extinction
ECD	Electron capture detector
ECG	Electrocardiogram
ECM	Electrochemical machining
ECMO	Extracorporeal membrane oxygenation
ECOD	Extracranial cerebrovascular occlusive disease
ECRI	Emergency Care Research Institute
ECS	Exner's Comprehensive System
ECT	Electroconvulsive shock therapy; Electroconvulsive therapy; Emission computed tomography
EDD	Estimated date of delivery
EDP	Aortic end diastolic pressure
EDTA	Ethylenediaminetetraacetic acid
EDX	Energy dispersive X-ray analysis
EEG	Electroencephalogram
EEI	Electrode electrolyte interface
EELV	End-expiratory lung volume
EER	Electrically evoked response
EF	Ejection fraction
EF	Electric field; Evoked magnetic fields
EFA	Estimated fetal age
EGF	Epidermal growth factor
EGG	Electrogastrogram
EIA	Enzyme immunoassay
EIU	Electrode impedance unbalance
ELF	Extra low frequency
ELGON	Electrical goniometer
ELISA	Enzyme-linked immunosorbent assay
ELS	Energy loss spectroscopy
ELV	Equivalent lung volume
EM	Electromagnetic
EMBS	Engineering in Medicine and Biology Society
emf	Electromotive force
EMG	Electromyogram
EMGE	Integrated electromyogram
EMI	Electromagnetic interference
EMS	Emergency medical services
EMT	Emergency medical technician
ENT	Ear, nose, and throat
EO	Elbow orthosis
EOG	Electrooculography
EOL	End of life
EOS	Eosinophil
EP	Elastoplastic; Evoked potentiate
EPA	Environmental protection agency
ER	Evoked response
ERCP	Endoscopic retrograde cholangiopancreatography
ERG	Electron radiography; Electroretinogram
ERMF	Event-related magnetic field
ERP	Event-related potential
ERV	Expiratory reserve volume

ESCA	Electron spectroscopy for chemical analysis
ESI	Electrode skin impedance
ESRD	End-stage renal disease
esu	Electrostatic unit
ESU	Electrosurgical unit
ESWL	Extracorporeal shock wave lithotripsy
ETO, Eto	Ethylene oxide
ETT	Exercise tolerance testing
EVA	Ethylene vinyl acetate
EVR	Endocardial viability ratio
EW	Extended wear
FAD	Flavin adenine dinucleotide
FARA	Flexible automation random analysis
FBD	Fetal biparietal diameter
FBS	Fetal bovine serum
fcc	Face centered cubic
FCC	Federal Communications Commission
Fct	Fluorocrit
FDA	Food and Drug Administration
FDCA	Food, Drug, and Cosmetic Act
FE	Finite element
FECG	Fetal electrocardiogram
FEF	Forced expiratory flow
FEL	Free electron lasers
FEM	Finite element method
FEP	Fluorinated ethylene propylene
FES	Functional electrical stimulation
FET	Field-effect transistor
FEV	Forced expiratory volume
FFD	Focal spot to film distance
FFT	Fast Fourier transform
FGF	Fresh gas flow
FHR	Fetal heart rate
FIC	Forced inspiratory capacity
FID	Flame ionization detector; Free-induction decay
FIFO	First-in-first-out
FITC	Fluorescent indicator tagged polymer
FL	Femur length
FM	Frequency modulation
FNS	Functional neuromuscular stimulation
FO	Foramen ovale
FO-CRT	Fiber optics cathode ray tube
FP	Fluorescence polarization
FPA	Fibrinopeptide A
FR	Federal Register
FRC	Federal Radiation Council; Functional residual capacity
FSD	Focus-to-surface distance
FTD	Focal spot to tissue-plane distance
FTIR	Fourier transform infrared
FTMS	Fourier transform mass spectrometer
FU	Fluorouracil
FUDR	Floxuridine
FVC	Forced vital capacity
FWHM	Full width at half maximum
FWTM	Full width at tenth maximum
GABA	Gamma amino buteric acid
GAG	Glycosaminoglycan
GBE	Gas-bearing electrodynamometer
GC	Gas chromatography; Guanine-cytosine
GDT	Gas discharge tube
GFR	Glomerular filtration rate
GHb	Glycosylated hemoglobin
GI	Gastrointestinal
GLC	Gas–liquid chromatography
GMV	General minimum variance
GNP	Gross national product
GPC	Giant papillary conjunctivitis
GPH	Gas-permeable hard
GPH-EW	Gas-permeable hard lens extended wear
GPO	Government Printing Office
GSC	Gas-solid chromatography
GSR	Galvanic skin response
GSWD	Generalized spike-wave discharge
HA	Hydroxyapatite
HAM	Helical axis of motion
Hb	Hemoglobin
HBE	His bundle electrogram
HBO	Hyperbaric oxygenation
HC	Head circumference
HCA	Hypothermic circulatory arrest
HCFA	Health care financing administration
HCL	Harvard Cyclotron Laboratory
hcp	Hexagonal close-packed
HCP	Half cell potential
HDPE	High density polyethylene
HECS	Hospital Equipment Control System
HEMS	Hospital Engineering Management System
HEPA	High efficiency particulate air filter
HES	Hydroxyethylstarch
HETP	Height equivalent to a theoretical plate
HF	High-frequency; Heating factor
HFCWO	High-frequency chest wall oscillation
HFER	High-frequency electromagnetic radiation
HFJV	High-frequency jet ventilation
HFO	High-frequency oscillator
HFOV	High-frequency oscillatory ventilation
HFPPV	High-frequency positive pressure ventilation
HFV	High-frequency ventilation
HHS	Department of Health and Human Services
HIBC	Health industry bar code
HIMA	Health Industry Manufacturers Association
HIP	Hydrostatic indifference point
HIS	Hospital information system
HK	Hexokinase
HL	Hearing level
HMBA	Hexamethylene bisacetamide
HMO	Health maintenance organization
HMWPE	High-molecular-weight polyethylene
HOL	Higher-order languages
HP	Heating factor; His-Purkinje
HpD	Hematoporphyrin derivative
HPLC	High-performance liquid chromatography
HPNS	High-pressure neurological syndrome
HPS	His-Purkinje system
HPX	High peroxidase activity

HR	Heart rate; High-resolution
HRNB	Halstead-Reitan Neuropsychological Battery
H/S	Hard/soft
HSA	Human serum albumin
HSG	Hysterosalpingogram
HTCA	Human tumor cloning assay
HTLV	Human T cell lymphotrophic virus
HU	Heat unit; Houndsfield units; Hydroxyurea
HVL	Half value layer
HVR	Hypoxic ventilatory response
HVT	Half-value thickness
IA	Image intensifier assembly; Inominate artery
IABP	Intraaortic balloon pumping
IAEA	International Atomic Energy Agency
IAIMS	Integrated Academic Information Management System
IASP	International Association for the Study of Pain
IC	Inspiratory capacity; Integrated circuit
ICCE	Intracapsular cataract extraction
ICD	Intracervical device
ICDA	International classification of diagnoses
ICL	Ms-clip lens
ICP	Inductively coupled plasma; Intracranial pressure
ICPA	Intracranial pressure amplitude
ICRP	International Commission on Radiological Protection
ICRU	International Commission on Radiological Units and Measurements
ICU	Intensive care unit
ID	Inside diameter
IDDM	Insulin dependent diabetes mellitus
IDE	Investigational device exemption
IDI	Index of inspired gas distribution
I:E	Inspiratory: expiratory
IEC	International Electrotechnical Commission; Ion-exchange chromatography
IEEE	Institute of Electrical and Electronics Engineers
IEP	Individual educational program
BETS	Inelastic electron tunneling spectroscopy
IF	Immunofluorescent
IFIP	International Federation for Information Processing
IFMBE	International Federation for Medical and Biological Engineering
IGFET	Insulated-gate field-effect transistor
IgG	Immunoglobulin G
IgM	Immunoglobulin M
IHP	Inner Helmholtz plane
IHSS	Idiopathic hypertrophic subaortic stenosis
II	Image intensifier
IIIES	Image intensifier input-exposure sensitivity
IM	Intramuscular
IMFET	Immunologically sensitive field-effect transistor
IMIA	International Medical Informatics Association
IMS	Information management system
IMV	Intermittent mandatory ventilation
INF	Interferon
IOL	Intraocular lens
IPC	Ion-pair chromatography
IPD	Intermittent peritoneal dialysis
IPG	Impedance plethysmography
IPI	Interpulse interval
IPPB	Intermittent positive pressure breathing
IPTS	International practical temperature scale
IR	Polyisoprene rubber
IRB	Institutional Review Board
IRBBB	Incomplete right bundle branch block
IRPA	International Radiation Protection Association
IRRAS	Infrared reflection-absorption spectroscopy
IRRS	Infrared reflection spectroscopy
IRS	Internal reflection spectroscopy
IRV	Inspiratory reserve capacity
IS	Image size; Ion-selective
ISC	Infant skin servo control
ISDA	Instantaneous screw displacement axis
ISE	Ion-selective electrode
ISFET	Ion-sensitive field effect transistor
ISIT	Intensified silicon-intensified target tube
ISO	International Organization for Standardization
ISS	Ion scattering spectroscopy
IT	Intrathecal
ITEP	Institute of Theoretical and Experimental Physics
ITEPI	Instantaneous trailing edge pulse impedance
ITLC	Instant thin-layer chromatography
IUD	Intrauterine device
IV	Intravenous
IVC	Inferior vena cava
IVP	Intraventricular pressure
JCAH	Joint Commission on the Accreditation of Hospitals
JND	Just noticeable difference
JRP	Joint replacement prosthesis
KB	Kent bundle
Kerma	Kinetic energy released in unit mass
KO	Knee orthosis
KPM	Kilopond meter
KRPB	Krebs-Ringer physiological buffer
LA	Left arm; Left atrium
LAD	Left anterior descending; Left axis deviation
LAE	Left atrial enlargement
LAK	Lymphokine activated killer
LAL	Limulus amoebocyte lysate
LAN	Local area network
LAP	Left atrial pressure
LAT	Left anterior temporalis
LBBB	Left bundle branch block
LC	Left carotid; Liquid chromatography

LCC	Left coronary cusp
LCD	Liquid crystal display
LDA	Laser Doppler anemometry
LDF	Laser Doppler flowmetry
LDH	Lactate dehydrogenase
LDPE	Low density polyethylene
LEBS	Low-energy brief stimulus
LED	Light-emitting diode
LEED	Low energy electron diffraction
LES	Lower esophageal sphincter
LESP	Lower esophageal sphincter pressure
LET	Linear energy transfer
LF	Low frequency
LH	Luteinizing hormone
LHT	Local hyperthermia
LL	Left leg
LLDPE	Linear low density polyethylene
LLPC	Liquid-liquid partition chromatography
LLW	Low-level waste
LM	Left masseter
LNNB	Luria-Nebraska Neuropsychological Battery
LOS	Length of stay
LP	Late potential; Lumboperitoneal
LPA	Left pulmonary artery
LPC	Linear predictive coding
LPT	Left posterior temporalis
LPV	Left pulmonary veins
LRP	Late receptor potential
LS	Left subclavian
LSC	Liquid-solid adsorption chromatography
LSI	Large scale integrated
LSV	Low-amplitude shear-wave viscoelastometry
LTI	Low temperature isotropic
LUC	Large unstained cells
LV	Left ventricle
LVAD	Left ventricular assist device
LVDT	Linear variable differential transformer
LVEP	Left ventricular ejection period
LVET	Left ventricular ejection time
LVH	Left ventricular hypertrophy
LYMPH	Lymphocyte
MAA	Macroaggregated albumin
MAC	Minimal auditory capabilities
MAN	Manubrium
MAP	Mean airway pressure; Mean arterial pressure
MAST	Military assistance to safety and traffic
MBA	Monoclonal antibody
MBV	Maximum breathing ventilation
MBX	Monitoring branch exchange
MCA	Methyl cryanoacrylate
MCG	Magnetocardiogram
MCI	Motion Control Incorporated
MCMI	Millon Clinical Multiaxial Inventory
MCT	Microcatheter transducer
MCV	Mean corpuscular volume
MDC	Medical diagnostic categories
MDI	Diphenylmethane diisocyanate; Medical Database Informatics
MDP	Mean diastolic aortic pressure
MDR	Medical device reporting
MDS	Multidimensional scaling
ME	Myoelectric
MED	Minimum erythema dose
MEDPAR	Medicare provider analysis and review
MEFV	Maximal expiratory flow volume
MEG	Magnetoencephalography
MeSH	Medline subject heading
METS	Metabolic equivalents
MF	Melamine-formaldehyde
MFP	Magnetic field potential
MGH	Massachusetts General Hospital
MHV	Magnetic heart vector
MI	Myocardial infarction
MIC	Minimum inhibitory concentration
MIFR	Maximum inspiratory flow rate
MINET	Medical Information Network
MIR	Mercury-in-rubber
MIS	Medical information system; Metal-insulator-semiconductor
MIT	Massachusetts Institute of Technology
MIT/BIH	Massachusetts Institute of Technology/ Beth Israel Hospital
MMA	Manual metal arc welding
MMA	Methyl methacrylate
MMECT	Multiple-monitored ECT
MMFR	Maximum midexpiratory flow rate
mm Hg	Millimeters of mercury
MMPI	Minnesota Multiphasic Personality Inventory
MMSE	Minimum mean square error
MO	Membrane oxygenation
MONO	Monocyte
MOSFET	Metal oxide silicon field-effect transistor
MP	Mercaptopurine; Metacarpal-phalangeal
MPD	Maximal permissible dose
MR	Magnetic resonance
MRG	Magnetoretinogram
MRI	Magnetic resonance imaging
MRS	Magnetic resonance spectroscopy
MRT	Mean residence time
MS	Mild steel; Multiple sclerosis
MSR	Magnetically shielded room
MTBF	Mean time between failure
MTF	Modulation transfer function
MTTR	Mean time to repair
MTX	Methotroxate
MUA	Motor unit activity
MUAP	Motor unit action potential
MUAPT	Motor unit action potential train
MUMPI	Missouri University Multi-Plane Imager
MUMPS	Massachusetts General Hospital utility multiuser programming system
MV	Mitral valve
MVO_2	Maximal oxygen uptake
MVTR	Moisture vapor transmission rate
MVV	Maximum voluntary ventilation
MW	Molecular weight

NAA	Neutron activation analysis
NAD	Nicotinamide adenine dinucleotide
NADH	Nicotinamide adenine dinucleotide, reduced form
NADP	Nicotinamide adenine dinucleotide phosphate
NAF	Neutrophil activating factor
NARM	Naturally occurring and accelerator-produced radioactive materials
NBB	Normal buffer base
NBD	Neuromuscular blocking drugs
N-BPC	Normal bonded phase chromatography
NBS	National Bureau of Standards
NCC	Noncoronary cusp
NCCLS	National Committee for Clinical Laboratory Standards; National Committee on Clinical Laboratory Standards
NCRP	National Council on Radiation Protection
NCT	Neutron capture theory
NEEP	Negative end-expiratory pressure
NEMA	National Electrical Manufacturers Association
NEMR	Nonionizing electromagnetic radiation
NEQ	Noise equivalent quanta
NET	Norethisterone
NEUT	Neutrophil
NFPA	National Fire Protection Association
NH	Neonatal hepatitis
NHE	Normal hydrogen electrode
NHLBI	National Heart, Lung, and Blood Institute
NIR	Nonionizing radiation
NIRS	National Institute for Radiologic Science
NK	Natural killer
NMJ	Neuromuscular junction
NMOS	N-type metal oxide silicon
NMR	Nuclear magnetic resonance
NMS	Neuromuscular stimulation
NPH	Normal pressure hydrocephalus
NPL	National Physical Laboratory
NR	Natural rubber
NRC	Nuclear Regulatory Commission
NRZ	Non-return-to-zero
NTC	Negative temperature coefficient
NTIS	National Technical Information Service
NVT	Neutrons versus time
NYHA	New York Heart Association
ob/gyn	Obstetrics and gynecology
OCR	Off-center ratio; Optical character recognition
OCV	Open circuit voltage
OD	Optical density; Outside diameter
ODC	Oxyhemoglobin dissociation curve
ODT	Oxygen delivery truck
ODU	Optical density unit
OER	Oxygen enhancement ratio
OFD	Object to film distance; Occiputo-frontal diameter
OHL	Outer Helmholtz layer
OHP	Outer Helmholtz plane
OIH	Orthoiodohippurate
OPG	Ocular pneumoplethysmography
OR	Operating room
OS	Object of known size; Operating system
OTC	Over the counter
OV	Offset voltage
PA	Posterioanterior; Pulmonary artery; Pulse amplitude
PACS	Picture archiving and communications systems
PAD	Primary afferent depolarization
PAM	Pulse amplitude modulation
PAN	Polyacrylonitrile
PAP	Pulmonary artery pressure
PAR	Photoactivation ratio
PARFR	Program for Applied Research on Fertility Regulation
PARR	Poetanesthesia recovery room
PAS	Photoacoustic spectroscopy
PASG	Pneumatic antishock garment
PBI	Penile brachial index
PBL	Positive beam limitation
PBT	Polybutylene terephthalate
PC	Paper chromatography; Personal computer; Polycarbonate
PCA	Patient controlled analgesia; Principal components factor analysis
PCG	Phonocardiogram
PCI	Physiological cost index
PCL	Polycaprolactone; Posterior chamber lens
PCR	Percent regurgitation
PCRC	Perinatal Clinical Research Center
PCS	Patient care system
PCT	Porphyria cutanea tarda
PCWP	Pulmonary capillary wedge pressure
PD	Peritoneal dialysis; Poly-p-dioxanone; Potential difference; Proportional and derivative
PDD	Percent depth dose; Perinatal Data Directory
PDE	Pregelled disposable electrodes
p.d.f.	Probability density function
PDL	Periodontal ligament
PDM	Pulse duration modulation
PDMSX	Polydimethyl siloxane
PDS	Polydioxanone
PE	Polyethylene
PEEP	Positive end-expiratory pressure
PEFR	Peak expiratory now rate
PEN	Parenteral and enteral nutrition
PEP	Preejection period
PEPPER	Programs examine phonetic find phonological evaluation records
PET	Polyethylene terephthalate; Positron-emission tomography
PEU	Polyetherurethane
PF	Platelet factor
PFA	Phosphonoformic add
PFC	Petrofluorochemical
PFT	Pulmonary function testing
PG	Polyglycolide; Propylene glycol

PGA	Polyglycolic add
PHA	Phytohemagglutinin; Pulse-height analyzer
PHEMA	Poly-2-hydroxyethyl methacrylate
PI	Propidium iodide
PID	Pelvic inflammatory disease; Proportional/integral/derivative
PIP	Peak inspiratory pressure
PL	Posterior leaflet
PLA	Polylactic acid
PLATO	Program Logic for Automated Teaching Operations
PLD	Potentially lethal damage
PLED	Periodic latoralized epileptiform discharge
PLT	Platelet
PM	Papillary muscles; Preventive maintenance
PMA	Polymethyl acrylate
p.m.f.	Probability mass function
PMMA	Polymethyl methacrylate
PMOS	P-type metal oxide silicon
PMP	Patient management problem; Poly(4-methylpentane)
PMT	Photomultiplier tube
PO	Per os
Po_2	Partial pressure of oxygen
POBT	Polyoxybutylene terephthalate
POM	Polyoxymethylene
POMC	Patient order management and communication system
POPRAS	Problem Oriented Perinatal Risk Assessment System
PP	Perfusion pressure; Polyproplyene; Postprandial (after meals)
PPA	Phonemic process analysis
PPF	Plasma protein fraction
PPM	Pulse position modulation
PPSFH	Polymerized phyridoxalated stroma-free hemoglobin
PR	Pattern recognition; Pulse rate
PRBS	Pseudo-random binary signals
PRP	Pulse repetition frequency
PRO	Professional review organization
PROM	Programmable read only memory
PS	Polystyrene
PSA	Pressure-sensitive adhesive
PSF	Point spread function
PSI	Primary skin irritation
PSP	Postsynaptic potential
PSR	Proton spin resonance
PSS	Progressive systemic sclerosis
PT	Plasma thromboplastin
PTB	Patellar tendon bearing orthosis
PTC	Plasma thromboplastin component; Positive temperature coefficient; Pressurized personal transfer capsule
PTCA	Percutaneous transluminal coronary angioplasty
PTFE	Polytetrafluoroethylene
PTT	Partial thromboplastin time
PUL	Percutaneous ultrasonic lithotripsy
PURA	Prolonged ultraviolet-A radiation
PUVA	Psoralens and longwave ultraviolet light photochemotherapy
P/V	Pressure/volume
PVC	Polyvinyl chloride; Premature ventricular contraction
PVI	Pressure–volume index
PW	Pulse wave; Pulse width
PWM	Pulse width modulation
PXE	Pseudo-xanthoma elasticum
QA	Quality assurance
QC	Quality control
R-BPC	Reverse bonded phase chromatography
R/S	Radiopaque-spherical
RA	Respiratory amplitude; Right arm
RAD	Right axis deviation
RAE	Right atrial enlargement
RAM	Random access memory
RAP	Right atrial pressure
RAT	Right anterior temporalis
RB	Right bundle
RBBB	Right bundle branch block
RBC	Red blood cell
RBE	Relative biologic effectiveness
RBF	Rose bengal fecal excretion
RBI	Resting baseline impedance
RCBD	Randomized complete block diagram
rCBF	Regional cerebral blood flow
RCC	Right coronary cusp
RCE	Resistive contact electrode
R&D	Research and development
r.e.	Random experiment
RE	Reference electrode
REM	Rapid eye movement; Return electrode monitor
REMATE	Remote access and telecommunication system
RES	Reticuloendothelial system
RESNA	Rehabilitation Engineering Society of North America
RF	Radio frequency; Radiographic-nuoroscopic
RFI	Radio-frequency interference
RFP	Request for proposal
RFQ	Request for quotation
RH	Relative humidity
RHE	Reversible hydrogen electrode
RIA	Radioimmunoassay
RM	Repetition maximum; Right masseter
RMR	Resting metabolic rate
RMS	Root mean square
RN	Radionuclide
RNCA	Radionuclide cineagiogram
ROI	Regions of interest
ROM	Range of motion; Read only memory
RP	Retinitis pigmentosa
RPA	Right pulmonary artery
RPP	Rate pressure product
RPT	Rapid pull-through technique
RPV	Right pulmonary veins
RQ	Respiratory quotient

RR	Recovery room
RRT	Recovery room time; Right posterior temporalis
RT	Reaction time
RTD	Resistance temperature device
RTT	Revised token test
r.v.	Random variable
RV	Residual volume; Right ventricle
RVH	Right ventricular hypertrophy
RVOT	Right ventricular outflow tract
RZ	Return-to-zero
SA	Sinoatrial; Specific absorption
SACH	Solid-ankle-cushion-heel
SAD	Source-axis distance; Statistical Analysis System
SAINT	System analysis of integrated network of tasks
SAL	Sterility assurance level; Surface averaged lead
SALT	Systematic analysis of language transcripts
SAMI	Socially acceptable monitoring instrument
SAP	Systemic arterial pressure
SAR	Scatter-air ratio; Specific absorption rate
SARA	System for anesthetic and respiratory gas analysis
SBE	Subbacterial endocarditis
SBR	Styrene-butadiene rubbers
SC	Stratum corneum; Subcommittees
SCAP	Right scapula
SCE	Saturated calomel electrode; Sister chromatid exchange
SCI	Spinal cord injury
SCRAD	Sub-Committee on Radiation Dosimetry
SCS	Spinal cord stimulation
SCUBA	Self-contained underwater breathing apparatus
SD	Standard deviation
SDA	Stepwise discriminant analysis
SDS	Sodium dodecyl sulfate
S&E	Safety and effectiveness
SE	Standard error
SEC	Size exclusion chromatography
SEM	Scanning electron microscope; Standard error of the mean
SEP	Somatosensory evoked potential
SEXAFS	Surface extended X-ray absorption fine structure
SF	Surviving fraction
SFD	Source-film distance
SFH	Stroma-free hemoglobin
SFTR	Sagittal frontal transverse rotational
SG	Silica gel
SGF	Silica gel fraction
SGG	Spark gap generator
SGOT	Serum glutamic oxaloacetic transaminase
SGP	Strain gage plethysmography; Stress-generated potential
SHE	Standard hydrogen electrode
SI	Le Système International d'Unités
SEBS	Surgical isolation barrier system
SID	Source to image reception distance
SIMFU	Scanned intensity modulated focused ultrasound
SIMS	Secondary ion mass spectroscopy; System for isometric muscle strength
SISI	Short increment sensitivity index
SL	Surgical lithotomy
SLD	Sublethal damage
SLE	Systemic lupus erythemotodes
SMA	Sequential multiple analyzer
SMAC	Sequential multiple analyzer with computer
SMR	Sensorimotor
S/N	Signal-to-noise
S:N/D	Signal-to-noise ratio per unit dose
SNP	Sodium nitroprusside
SNR	Signal-to-noise ratio
SOA	Sources of artifact
SOAP	Subjective, objective, assessment, plan
SOBP	Spread-out Bragg peak
SP	Skin potential
SPECT	Single photon emission computed tomography
SPL	Sound pressure level
SPRINT	Single photon ring tomograph
SPRT	Standard platinum resistance thermometer
SPSS	Statistical Package for the Social Sciences
SQUID	Superconducting quantum interference device
SQV	Square wave voltammetry
SR	Polysulfide rubbers
SRT	Speech reception threshold
SS	Stainless steel
SSB	Single strand breaks
SSD	Source-to-skin distance; Source-to-surface distance
SSE	Stainless steel electrode
SSEP	Somatosensory evoked potential
SSG	Solid state generator
SSP	Skin stretch potential
SSS	Sick sinus syndrome
STD	Source-tray distance
STI	Systolic time intervals
STP	Standard temperature and pressure
STPD	Standard temperature pressure dry
SV	Stroke volume
SVC	Superior vena cava
SW	Standing wave
TAA	Tumor-associated antigens
TAC	Time-averaged concentration
TAD	Transverse abdominal diameter
TAG	Technical Advisory Group
TAH	Total artificial heart
TAR	Tissue-air ratio
TC	Technical Committees
TCA	Tricarboxylic acid cycle
TCD	Thermal conductivity detector
TCES	Transcutaneous cranial electrical stimulation

TCP	Tricalcium phosphate
TDD	Telecommunication devices for the deaf
TDM	Therapeutic drug monitoring
TE	Test electrode; Thermoplastic elastomers
TEAM	Technology evaluation and acquisition methods
TEM	Transmission electron microscope; Transverse electric and magnetic mode; Transverse electromagnetic mode
TENS	Transcutaneous electrical nerve stimulation
TEP	Tracheoesophageal puncture
TEPA	Triethylenepho-sphoramide
TF	Transmission factor
TFE	Tetrafluorethylene
TI	Totally implantable
TICCIT	Time-shared Interaction Computer-Controlled Information Television
TLC	Thin-layer chromatography; Total lung capacity
TLD	Thermoluminescent dosimetry
TMJ	Temporomandibular joint
TMR	Tissue maximum ratio; Topical magnetic resonance
TNF	Tumor necrosis factor
TOF	Train-of-four
TP	Thermal performance
TPC	Temperature pressure correction
TPD	Triphasic dissociation
TPG	Transvalvular pressure gradient
TPN	Total parenteral nutrition
TR	Temperature rise
tRNA	Transfer RNA
TSH	Thyroid stimulating hormone
TSS	Toxic shock syndrome
TTD	Telephone devices for the deaf
TTI	Tension time index
TTR	Transition temperature range
TTV	Trimming tip version
TTY	Teletypewriter
TUR	Transurethral resection
TURP	Transurethral resections of the prostrate
TV	Television; Tidal volume; Tricuspid valve
TVER	Transscleral visual evoked response
TW	Traveling wave
TxB_2	Thrombozame B^2
TZ	Transformation zone
UES	Upper esophageal sphincter
UP	Urea-formaldehyde
UfflS	University Hospital Information System
UHMW	Ultra high molecular weight
UHMWPE	Ultra high molecular weight polyethylene
UL	Underwriters Laboratory
ULF	Ultralow frequency
ULTI	Ultralow temperature isotropic
UMN	Upper motor neuron
UO	Urinary output
UPTD	Unit pulmonary oxygen toxicity doses
UR	Unconditioned response
US	Ultrasound; Unconditioned stimulus
USNC	United States National Committee
USP	United States Pharmacopeia
UTS	Ultimate tensile strength
UV	Ultraviolet; Umbilical vessel
UVR	Ultraviolet radiation
V/F	Voltage-to-frequency
VA	Veterans Administration
VAS	Visual analog scale
VBA	Vaginal blood volume in arousal
VC	Vital capacity
VCO	Voltage-controlled oscillator
VDT	Video display terminal
VECG	Vectorelectrocardiography
VEP	Visually evoked potential
VF	Ventricular fibrillation
VOP	Venous occlusion plethysmography
VP	Ventriculoperitoneal
VPA	Vaginal pressure pulse in arousal
VPB	Ventricular premature beat
VPR	Volume pressure response
VSD	Ventricular septal defect
VSWR	Voltage standing wave ratio
VT	Ventricular tachycardia
VTG	Vacuum tube generator
VTS	Viewscan text system
VV	Variable version
WAIS-R	Weschler Adult Intelligence Scale-Revised
WAK	Wearable artificial kidney
WAML	Wide-angle mobility light
WBAR	Whole-body autoradiography
WBC	White blood cell
WG	Working Groups
WHO	World Health Organization; Wrist hand orthosis
WLF	Williams-Landel-Ferry
WMR	Work metabolic rate
w/o	Weight percent
WORM	Write once, read many
WPW	Wolff-Parkinson-White
XPS	X-ray photon spectroscopy
XR	Xeroradiograph
YAG	Yttrium aluminum garnet
ZPL	Zero pressure level

CONVERSION FACTORS AND UNIT SYMBOLS

SI UNITS (ADOPTED 1960)

A new system of metric measurement, the International System of Units (abbreviated SI), is being implemented throughout the world. This system is a modernized version of the MKSA (meter, kilogram, second, ampere) system, and its details are published and controlled by an international treaty organization (The International Bureau of Weights and Measures).

SI units are divided into three classes:

Base Units

length	meter[†] (m)
mass[‡]	kilogram (kg)
time	second (s)
electric current	ampere (A)
thermodynamic temperature§	kelvin (K)
amount of substance	mole (mol)
luminous intensity	candela (cd)

Supplementary Units

plane angle	radian (rad)
solid angle	steradian (sr)

Derived Units and Other Acceptable Units

These units are formed by combining base units, supplementary units, and other derived units. Those derived units having special names and symbols are marked with an asterisk (*) in the list below:

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
*absorbed dose	gray	Gy	J/kg
acceleration	meter per second squared	m/s^2	
*activity (of ionizing radiation source)	becquerel	Bq	1/s
area	square kilometer	km^2	
	square hectometer	hm^2	ha (hectare)
	square meter	m^2	

[†]The spellings "metre" and "litre" are preferred by American Society for Testing and Materials (ASTM); however, "−er" will be used in the Encyclopedia.

[‡]"Weight" is the commonly used term for "mass."

§Wide use is made of "Celsius temperature" (t) defined $t = T - T_0$ where T is the thermodynamic temperature, expressed in kelvins, and $T_0 = 273.15\,K$ by definition. A temperature interval may be expressed in degrees Celsius as well as in kelvins.

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
*capacitance	farad	F	C/V
concentration (of amount of substance)	mole per cubic meter	mol/m^3	
*conductance	siemens	S	A/V
current density	ampere per square meter	A/m^2	
density, mass density	kilogram per cubic meter	kg/m^3	g/L; mg/cm^3
dipole moment (quantity)	coulomb meter	C·m	
*electric charge, quantity of electricity	coulomb	C	A·s
electric charge density	coulomb per cubic meter	C/m^3	
electric field strength	volt per meter	V/m	
electric flux density	coulomb per square meter	C/m^2	
*electric potential, potential difference, electromotive force	volt	V	W/A
*electric resistance	ohm	Ω	V/A
*energy, work, quantity of heat	megajoule	MJ	
	kilojoule	kJ	
	joule	J	N·m
	electron volt†	eV†	
	kilowatt hour†	kW·h†	
energy density	joule per cubic meter	J/m^3	
*force	kilonewton	kN	
	newton	N	$kg·m/s^2$
*frequency	megahertz	MHz	
	hertz	Hz	1/s
heat capacity, entropy	joule per kelvin	J/K	
heat capacity (specific), specific entropy	joule per kilogram kelvin	J/(kg·K)	
heat transfer coefficient	watt per square meter kelvin	$W/(m^2·K)$	
*illuminance	lux	lx	lm/m^2
*inductance	henry	H	Wb/A
linear density	kilogram per meter	kg/m	
luminance	candela per square meter	cd/m^2	
*luminous flux	lumen	lm	cd·sr
magnetic field strength	ampere per meter	A/m	
*magnetic flux	weber	Wb	V·s
*magnetic flux density	tesla	T	Wb/m^2
molar energy	joule per mole	J/mol	
molar entropy, molar heat capacity	joule per mole kelvin	J/(mol·K)	
moment of force, torque	newton meter	N·m	
momentum	kilogram meter per second	kg·m/s	
permeability	henry per meter	H/m	
permittivity	farad per meter	F/m	
*power, heat flow rate, radiant flux	kilowatt	kW	
	watt	W	J/s
power density, heat flux density, irradiance	watt per square meter	W/m^2	
*pressure, stress	megapascal	MPa	
	kilopascal	kPa	
	pascal	Pa	N/m^2
sound level	decibel	dB	
specific energy	joule per kilogram	J/kg	
specific volume	cubic meter per kilogram	m^3/kg	
surface tension	newton per meter	N/m	
thermal conductivity	watt per meter kelvin	W/(m·K)	
velocity	meter per second	m/s	
	kilometer per hour	km/h	
viscosity, dynamic	pascal second	Pa·s	
	millipascal second	mPa·s	

†This non-SI unit is recognized as having to be retained because of practical importance or use in specialized fields.

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
viscosity, kinematic	square meter per second	m^2/s	
	square millimeter per second	mm^2/s	
	cubic meter	m^3	
	cubic decimeter	dm^3	L(liter)
	cubic centimeter	cm^3	mL
wave number	1 per meter	m^{-1}	
	1 per centimeter	cm^{-1}	

In addition, there are 16 prefixes used to indicate order of magnitude, as follows:

Multiplication factor	*Prefix*	*Symbol*	*Note*
10^{18}	exa	E	
10^{15}	peta	P	
10^{12}	tera	T	
10^{9}	giga	G	
10^{8}	mega	M	
10^{3}	kilo	k	
10^{2}	hecto	h[a]	[a]Although hecto, deka, deci, and centi are
10	deka	da[a]	SI prefixes, their use should be avoided
10^{-1}	deci	d[a]	except for SI unit-multiples for area and
10^{-2}	centi	c[a]	volume and nontechnical use of
10^{-3}	milli	m	centimeter, as for body and clothing
10^{-6}	micro	μ	measurement.
10^{-9}	nano	n	
10^{-12}	pico	p	
10^{-15}	femto	f	
10^{-18}	atto	a	

For a complete description of SI and its use the reader is referred to ASTM E 380.

CONVERSION FACTORS TO SI UNITS

A representative list of conversion factors from non-SI to SI units is presented herewith. Factors are given to four significant figures. Exact relationships are followed by a dagger (†). A more complete list is given in ASTM E 380-76 and ANSI Z210. 1-1976.

To convert from	*To*	*Multiply by*
acre	square meter (m^2)	4.047×10^{3}
angstrom	meter (m)	$1.0 \times 10^{-10\dagger}$
are	square meter (m^2)	$1.0 \times 10^{2\dagger}$
astronomical unit	meter (m)	1.496×10^{11}
atmosphere	pascal (Pa)	1.013×10^{5}
bar	pascal (Pa)	$1.0 \times 10^{5\dagger}$
barrel (42 U.S. liquid gallons)	cubic meter (m^3)	0.1590
Btu (International Table)	joule (J)	1.055×10^{3}
Btu (mean)	joule (J)	1.056×10^{3}
Bt (thermochemical)	joule (J)	1.054×10^{3}
bushel	cubic meter (m^3)	3.524×10^{-2}
calorie (International Table)	joule (J)	4.187
calorie (mean)	joule (J)	4.190
calorie (thermochemical)	joule (J)	$4.184^{\dagger}$
centimeters of water (39.2 °F)	pascal (Pa)	98.07
centipoise	pascal second (Pa·s)	$1.0 \times 10^{-3\dagger}$
centistokes	square millimeter per second (mm^2/s)	$1.0^{\dagger}$

To convert from	*To*	*Multiply by*
cfm (cubic foot per minute)	cubic meter per second (m^3/s)	4.72×10^{-4}
cubic inch	cubic meter (m^3)	1.639×10^{-4}
cubic foot	cubic meter (m^3)	2.832×10^{-2}
cubic yard	cubic meter (m^3)	0.7646
curie	becquerel (Bq)	$3.70 \times 10^{10\dagger}$
debye	coulomb-meter (C·m)	3.336×10^{-30}
degree (angle)	radian (rad)	1.745×10^{-2}
denier (international)	kilogram per meter (kg/m)	1.111×10^{-7}
	tex	0.1111
dram (apothecaries')	kilogram (kg)	3.888×10^{-3}
dram (avoirdupois)	kilogram (kg)	1.772×10^{-3}
dram (U.S. fluid)	cubic meter (m^3)	3.697×10^{-6}
dyne	newton(N)	$1.0 \times 10^{-6\dagger}$
dyne/cm	newton per meter (N/m)	$1.00 \times 10^{-3\dagger}$
electron volt	joule (J)	1.602×10^{-19}
erg	joule (J)	$1.0 \times 10^{-7\dagger}$
fathom	meter (m)	1.829
fluid ounce (U.S.)	cubic meter (m^3)	2.957×10^{-5}
foot	meter (m)	$0.3048^{\dagger}$
foot-pound force	joule (J)	1.356
foot-pound force	newton meter (N·m)	1.356
foot-pound force per second	watt(W)	1.356
footcandle	lux (lx)	10.76
furlong	meter (m)	2.012×10^2
gal	meter per second squared (m/s^2)	$1.0 \times 10^{-2\dagger}$
gallon (U.S. dry)	cubic meter (m^3)	4.405×10^{-3}
gallon (U.S. liquid)	cubic meter (m^3)	3.785×10^{-3}
gilbert	ampere (A)	0.7958
gill (U.S.)	cubic meter (m^3)	1.183×10^{-4}
grad	radian	1.571×10^{-2}
grain	kilogram (kg)	6.480×10^{-5}
gram force per denier	newton per tex (N/tex)	8.826×10^{-2}
hectare	square meter (m^2)	$1.0 \times 10^{4\dagger}$
horsepower (550 ft·lbf/s)	watt(W)	7.457×10^2
horsepower (boiler)	watt(W)	9.810×10^3
horsepower (electric)	watt(W)	$7.46 \times 10^{2\dagger}$
hundredweight (long)	kilogram (kg)	50.80
hundredweight (short)	kilogram (kg)	45.36
inch	meter (m)	$2.54 \times 10^{-2\dagger}$
inch of mercury (32 °F)	pascal (Pa)	3.386×10^3
inch of water (39.2 °F)	pascal (Pa)	2.491×10^2
kilogram force	newton (N)	9.807
kilopond	newton (N)	9.807
kilopond-meter	newton-meter (N·m)	9.807
kilopond-meter per second	watt (W)	9.807
kilopond-meter per min	watt(W)	0.1635
kilowatt hour	megajoule (MJ)	$3.6^{\dagger}$
kip	newton (N)	4.448×10^2
knot international	meter per second (m/s)	0.5144
lambert	candela per square meter (cd/m^2)	3.183×10^3
league (British nautical)	meter (m)	5.559×10^2
league (statute)	meter (m)	4.828×10^3
light year	meter (m)	9.461×10^{15}
liter (for fluids only)	cubic meter (m^3)	$1.0 \times 10^{-3\dagger}$
maxwell	weber (Wb)	$1.0 \times 10^{-8\dagger}$
micron	meter (m)	$1.0 \times 10^{-6\dagger}$
mil	meter (m)	$2.54 \times 10^{-5\dagger}$
mile (U.S. nautical)	meter (m)	$1.852 \times 10^{3\dagger}$
mile (statute)	meter (m)	1.609×10^3
mile per hour	meter per second (m/s)	0.4470

To convert from	*To*	*Multiply by*
millibar	pascal (Pa)	1.0×10^{2}
millimeter of mercury (0 °C)	pascal (Pa)	1.333×10^{2}†
millimeter of water (39.2 °F)	pascal (Pa)	9.807
minute (angular)	radian	2.909×10^{-4}
myriagram	kilogram (kg)	10
myriameter	kilometer (km)	10
oersted	ampere per meter (A/m)	79.58
ounce (avoirdupois)	kilogram (kg)	2.835×10^{-2}
ounce (troy)	kilogram (kg)	3.110×10^{-2}
ounce (U.S. fluid)	cubic meter (m^3)	2.957×10^{-5}
ounce-force	newton (N)	0.2780
peck (U.S.)	cubic meter (m^3)	8.810×10^{-3}
pennyweight	kilogram (kg)	1.555×10^{-3}
pint (U.S. dry)	cubic meter (m^3)	5.506×10^{-4}
pint (U.S. liquid)	cubic meter (m^3)	4.732×10^{-4}
poise (absolute viscosity)	pascal second (Pa·s)	0.10†
pound (avoirdupois)	kilogram (kg)	0.4536
pound (troy)	kilogram (kg)	0.3732
poundal	newton (N)	0.1383
pound-force	newton (N)	4.448
pound per square inch (psi)	pascal (Pa)	6.895×10^{3}
quart (U.S. dry)	cubic meter (m^3)	1.101×10^{-3}
quart (U.S. liquid)	cubic meter (m^3)	9.464×10^{-4}
quintal	kilogram (kg)	1.0×10^{2}†
rad	gray (Gy)	1.0×10^{-2}†
rod	meter (m)	5.029
roentgen	coulomb per kilogram (C/kg)	2.58×10^{-4}
second (angle)	radian (rad)	4.848×10^{-6}
section	square meter (m^2)	2.590×10^{6}
slug	kilogram (kg)	14.59
spherical candle power	lumen (lm)	12.57
square inch	square meter (m^2)	6.452×10^{-4}
square foot	square meter (m^2)	9.290×10^{-2}
square mile	square meter (m^2)	2.590×10^{6}
square yard	square meter (m^2)	0.8361
store	cubic meter (m^3)	1.0†
stokes (kinematic viscosity)	square meter per second (m^2/s)	1.0×10^{-4}†
tex	kilogram per meter (kg/m)	1.0×10^{-6}†
ton (long, 2240 pounds)	kilogram (kg)	1.016×10^{3}
ton (metric)	kilogram (kg)	1.0×10^{3}†
ton (short, 2000 pounds)	kilogram (kg)	9.072×10^{2}
torr	pascal (Pa)	1.333×10^{2}
unit pole	weber (Wb)	1.257×10^{-7}
yard	meter (m)	0.9144†

continued

RADIOTHERAPHY, COBALT 60 UNITS FOR.

See COBALT 60 UNITS FOR RADIOTHERAPY.

RADIOTHERAPY, HEAVY IONS AND ELECTRONS

F. D. BECCHETTI
University of Michigan
Ann Arbor, Michigan

INTRODUCTION

The use of subatomic particles to treat cancer and other medical conditions can be traced back (1) to the discovery of natural radioactivity by Bequerel in 1896. Marie Sklodowska Curie and her husband Pierre Curie quickly identified the primary radiation emitted by radioactive materials such as uranium and thorium as consisting of three principal types: α, β, and γ rays. Among these rays, only γ rays, being high energy photons, are a form of electromagnetic radiation. Therefore, they are similar to lower energy photons and, in particular, X rays with regard to their interaction in matter (2–4).

In contrast, the α and β rays were observed to be much more ionizing than γ rays, and unlike γ rays had a finite range in materials. The latter feature is characteristic of energetic subatomic particles and, indeed, the α ray was later identified as an energetic 4 He ion. It is emitted via nuclear α decay from heavy radioactive elements. Likewise, β rays were identified as energetic electrons emitted via nuclear β decay from light and heavy radioactive elements. Subsequently, the Curies were able to identify one particular highly radioactive element, radium. With considerable time and effort, they were able to extract small but usable pure samples of this from large amounts of uranium ore. As is also the case with X rays, workers using strong radioactive sources, including the Curies, often developed skin rashes and other symptoms related to exposure to natural radiation from radioactive materials. Medical doctors quickly realized that this radiation also could be used in medical applications.

Thus, following closely on the work by Roentgen et al. on the application of manmade radiation (i.e., X rays) to treat cancer, the Curies and others used radium and other natural radioactive sources to treat cancer tumors (Fig. 1) (1). Although the α, β, and γ rays emitted from nuclear decay were emitted with a much higher energy (MeV vs keV for X rays), the α and β particles had very short ranges in tissue (e.g., a mm or less for α particles and a few cm for β particles). Therefore, these sources, particularly including the associated MeV gamma rays, primarily were used to treat surface tumors (Fig. 1). An early form of bracyatherapy using radioactive needles also was developed to treat deeper tumors (1).

Figure 1. An early form of radiation treatment (*ca*. 1920) using radioactive sources (1).

It would await the development of high energy particle accelerators after WWII to provide α and β particles (helium ions and electrons) with sufficient energy and intensity to be effective in treating deep tumors (5). However, belatedly, in 1932, Chadwick discovered another basic subatomic particle, the neutron. Being uncharged, neutrons unlike α and β particles behave more like X rays and γ rays in tissue. The discovery of the neutron coincided with the invention of the cyclotron by E. Lawrence, and an early use of the cyclotron was to produce energetic neutrons (5–7). Stone et al. proposed to use neutrons for the treatment of cancer and many treatments were performed in the late 1930s.

The modern era of particle-beam radiotherapy begins after WWII with the development of linear accelerators (LINACs), betatrons, the synchro-cyclotron, and the synchrotron (5–8) to provide electron, proton, α and heavy-ion beams at energies sufficient to penetrate many centimeters of tissue. Much of this work was pioneered at the University of California-Berkeley by E. Lawrence, his brother

Table 1. Worldwide Charged Particle Patient Totals

January 2005 WHO	WHERE	WHAT	DATE FIRST RX	DATE LAST RX	RECENT PATIENT TOTAL	DATE OF TOTAL
Berkeley 184	CA. USA	p	1954	–1957	30	
Berkeley	CA. USA	He	1957	–1992	2054	
Uppsala (1)	Sweden	p	1957	–1976	73	
Harvard	MA. USA	p	1961	–2002	9116	
Dubna (1)	Russia	p	1967	–1996	124	
ITEP, Moscow	Russia	p	1969		3785	Dec-04
Los Alamos	NM. USA	π^-	1974	–1982	230	
St. Petersburg	Russia	p	1975		1145	April-04
Berkeley	CA. USA	ion	1975	–1992	433	
Chiba	Japan	p	1979		145	Apr-02
TRIUMF	Canada	π^-	1979	–1994	367	
PSl (SIN)	Switzerland	π^-	1980	–1993	503	
PMRC (1), Tsukuba	Japan	p	1983	–2000	700	
PSI (72 MeV)	Switzerland	p	1984		4182	Dec-04
Uppsala (2)	Sweden	p	1989		418	Jan-04
Clatterbridge	England	p	1989		1372	Dec-04
Loma Linda	CA, USA	p	1990		9585	Nov-04
Louvain-la-Neuve	Belgium	p	1991	–1993	21	
Nice	France	p	1991		2555	April-04
Orsay	France	p	1991		2805	Dec-03
iThemba LABS	South Africa	p	1993		468	Nov-04
MPRI (1)	IN USA	p	1993	–1999	34	
UCSF - CNL	CA USA	p	1994		632	June-04
HIMAC, Chiba	Japan	C ion	1994		1796	Feb-04
TRIUMF	Canada	p	1995		89	Dec-03
PSI (200 MeV)	Switzerland	p	1996		209	Dec-04
G.S.I Darmstadt	Germany	C ion	1997		198	Dec-03
H.M.I, Berlin	Germany	p	1998		546	Dec-04
NCC, Kashiwa	Japan	p	1998		300	Oct-04
Dubna (2)	Russia	p	1999		296	Dec-04
HIBMC, Hyogo	Japan	p	2001		483	Dec-04
PMRC (2), Tsukuba	Japan	p	2001		492	July 04
NPTC, MGH	MA USA	p	2001		973	Dec-04
HIBMC, Hyogo	Japan	C ion	2002		30	Dec-02
INFN-LNS, Catania	Italy	p	2002		82	Oct-04
WERC	Japan	p	2002		19	Oct-04
Shizuoka	Japan	p	2003		100	Dec-04
MPRI (2)	IN USA	p	2004		21	July-04
Wanjie, Zibo	China	p	2004		1	Dec-04
					1100	pions
					4511	ions
					40801	protons
				TOTAL	46412	all particles

Adopted from PTOG 35 Newsletter (Jan. 2005) (12).

John, a medical doctor, and C. Tobias. Using the 184 inch synchro-cyclotron to accelerate high energy α particles, an active radiation therapy program, as an adjunct to the primary nuclear physics research program, was carried out from 1954 to 1986 (Table 1) (9–12).

The justification for using a heavy, charged subatomic particle such as a proton, α or heavier ion lies in the fact that the main kinetic energy loss per path length in tissue ($\Delta E/\Delta X$) occurs primarily via collisional losses i.e., via many collisions with atomic electrons (2). Each collision absorbs a small amount of energy with only minimal scattering of the incident particle, which produces a well-defined energy loss vs depth curve (linear-energy transfer or LET) and hence range of the particle in tissue or its near equaivalent, water (Fig. 2) (13). In medical physics, LET is usually specified in keV/μm or, alternatively, keV/g/cm^2 for a specific medium. Specifically, in a given material, the LET for a charged, nonrelativistic heavy particle of mass m and atomic number z moving at a velocity v [hence, kinetic energy $E = (1/2)mv^2$] has the behavior (2–5)

$$\begin{aligned} \text{LET} &= k z^2/v^2 \\ &= K z^2/(E/m) \end{aligned} \tag{1}$$

where k and K are constants that depend on the atomic composition of the medium. Suitable integration of LET from $E =$ incident E $(= E_0)$ to $E = 0$ then gives the range, R of the particle in the medium. Based on the above, we would expect R to then be proportional to ${E_0}^n$ with $n = 2$, which is approximately true but, because of various

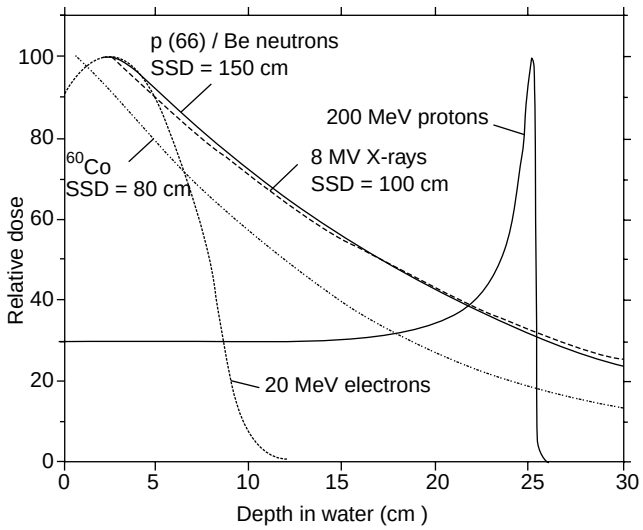


Figure 2. Depth-dose curves in water (which is similar to soft tissue) for various types of photons and particles commonly used in radiation therapy (arbitrary normalization; adapted from Ref. 13).

atomic effects (2–5), one typically has $1.4 < n < 2$ depending on z and E_0.

Owing to the $1/E$ dependence of LET (Eq. 1, above), protons, alphas, and heavier particles such as 12 C ions exhibit a very high LET at the end of their range, resulting in a sharp "Bragg peak" (Fig. 2). In cancer treatment, this sharp LET peak is often then spread out using energy-loss absorbers or other means to produce a spread-out Bragg peak (SOBP) suitable for treating an extended tumor (Fig. 3).

In the case of heavy particles, the biological dose will depend on the LET as well as the relative biological effectiveness (RBE) of the incident particle (2). The RBE may

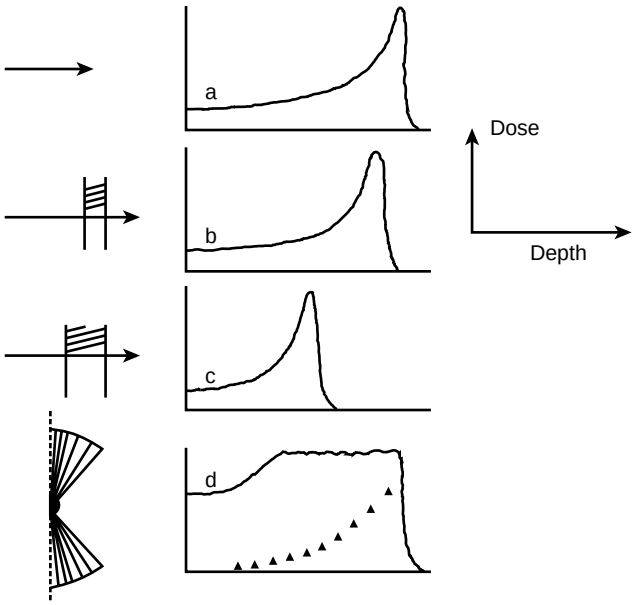


Figure 3. Illustration showing modification of a particle depth-dose curve by means of a spread-out Bragg peak (SOBP; adapted from Ref. 5).

depend on m, z, E, and several other factors including the LET. Although electrons, X rays, and γ rays generally have an RBE near unity, heavy particles (including neutrons) can have much larger values *viz*. RBE = 4 or greater (5–11,14,15). In addition, energetic heavy particles can lose energy and even transmute into a different particle via nuclear reactions and fragmentation (9–11,14–16), which can extend the net LET beyond the range of the incident particle due to atomic collisional losses alone and also introduce scattering of the incident beam. These features can become significant for heavier particles such as α particles and 12 C ions.

In contrast, photons such as X rays and γ rays lose energy through several different mechanisms (2), which normally only involve a few collisions for a particular photon. Each collision usually produces a scattered, secondary electron that can carry away (and hence absorb) significant energy from the photon as well as scatter the photon. The energy loss of all the secondary electrons and their RBE (near unity) yields the biological radiation dose attributed to the photon. One can sum this dose over many incident photons to generate a dose-distance curve for a beam of photons (Fig. 2). However, no single X-ray or γ-ray photon has a well-defined energy-loss curve or range per se. Instead, like visible-light photons, a beam of such radiation primarily will be attenuated or scattered as it passes through tissue and thus yield the dose–distance curves shown in Fig. 2.

In contrast to charged particles, most of the dose for a single photon beam thus occurs near the surface and continues through the treatment area including the exit region (Fig. 2). Treating deep-seated tumors while sparing adjacent healthy tissue from a lethal dose requires multiaxis beam treatment planning together with fractionation of the dose (15–20), which can present a problem for tumors near critical organs, such as the spinal cord, and optical nerve, and, for such tumors, certain particle beams can be advantageous.

Energetic electron beams exhibit characteristics of both particle beams as well as X rays and γ rays. Although the primary energy loss is via collisions with atomic electrons, owing to the incident mass also being that of an electron, this energy loss is accompanied with significant scattering of the primary electrons (2), which results in a complicated dose-distance profile (Figs. 2 and 4) (21). In addition, the electrons used in radiation therapy are at relativistic energies (i.e., $E_0 > 1$ MeV) and their range, unlike that of heavier particles, increases more slowly (2) with incident energy ($n \approx 1$), but, due to scattering, the range along the incident beam axis is not well defined. Hence, the large penumbra exhibited (Fig. 4) can limit the usefulness of electron beams in treating deep tumors. The application of a magnetic field to reduce the penumbra has been proposed and is under active study by several research groups (see below).

Subatomic antimatter particles such as antiprotons and positrons (antielectrons) have also been proposed for cancer therapy, which is based on past work (14–22) done at the Los Alamos Meson Factory (LAMPF) using another form of antimatter, the negative pi meson (negative pion). Antimatter particles like the negative pion undergo a

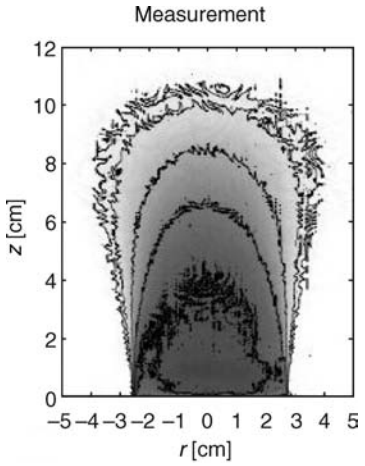


Figure 4. A 2D profile of a *ca*. 20 MeV electron beam incident on a tissue-equivalent plastic phantom (21).

nuclear annihilation reaction when they stop near the end of their range, which converts the mass of the antiparticle and a related nuclear constituent into photon energy, which produces additional radiation to enhance the dose near the end of the range (and enhanced dose peak, Fig. 5). The LAMPF, TRIUMF, and SIN nuclear research laboratories had active negative pion radiation oncology treatment programs from the early 1970s to the early 1990s (Table 1). However, owing to practical considerations associated with the large accelerator needed, this type of treatment is no longer in active use.

A characterization of the dose and LET of various particle beams used in radiation oncology is displayed in Fig. 6 (23). The features noted can help determine the type of beam and treatment plan that may be optimal to treat a specific type of tumor. Table 2 lists the various types of particle-beam modalities, their characteristics, and the type of cancers generally treated with each modality (24).

Until recently, excluding electrons, most particle-beam cancer treatment facilities used accelerators primarily designed and operated for nuclear physics research. Such accelerators were generally not optimized for radiation

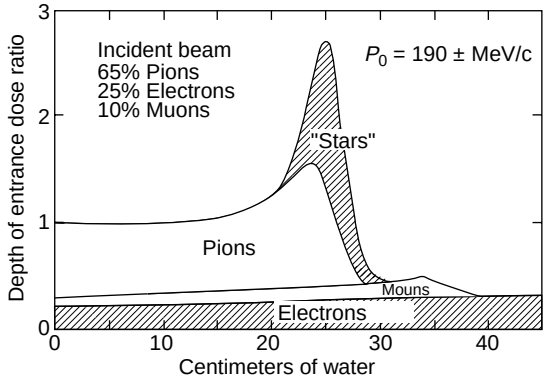


Figure 5. Dose-depth curve for negative pions (a form of antimatter) in water. Note the enhanced dose at the end of the range due to annihilation radiation (adopted from Ref. 5).

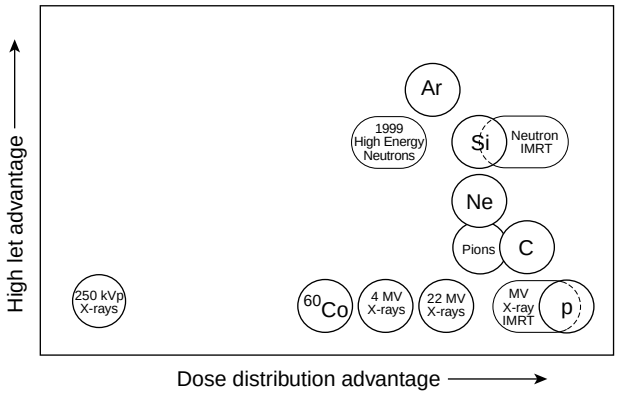


Figure 6. Characteristics of dose and LET for various forms of radiotherapy beams (23).

therapy and were costly to operate and maintain. As an example, nuclear research accelerators are often designed for much higher beam intensities and beam energy than needed for tumor treatment. Beam time also may be limited, which complicates treatment by requiring fractionated doses, and the patient-related aspects in a research laboratory are not ideal in most cases. Nonetheless, these facilities, many still in operation, demonstrated the effectiveness of particle-beam therapy for certain types of tumor treatment, which has lead to the construction (or planning) of several dedicated particle-beam cancer treatment facilities throughout the world. A listing of these, together with some of the pioneering, older charged-particle facilities is given in Table 1.

In the following sections, we will describe in more detail specific types of particle-beam treatments and related apparatus. Additional information may be found in other sections of this volume as well as in several review articles (9–11,14).

In addition, several groups such as the Proton Therapy Co-operative Group (PTCOG) (25) and the TERA Foundation (26) issue newsletters and hold regular conferences with proceedings covering recent advances in particle-beam therapy.

Many of the techniques used in particle-beam radiation therapy are similar to those used in more conventional radiation therapy using photons, and extensive literature for the latter exists (17–20, 27–30). Likewise, several detailed references discuss the effect of radiation on cell survival and, hence, RBE for particular types of radiation including particle beams (2–5,7,15,31–35).

ELECTRON-BEAM RADIOTHERAPY

The most common type of particle-beam radiotherapy involves the use of energetic electrons, either directly from an electron accelerator (8,17,27,36) or as the ionizing radiation in radio-isotope brachytherapy. Most hospital-based radiation-oncology departments use an electron LINAC, typically with electron beam energies from 5 MeV to 25 MeV. This beam produces, via bremstrahlung, on a tungsten or similar target, the megavolt X rays (i.e., high energy photons used for standard radiotherapy)

Table 2. Summary of Present External Beam Radiotherapy Options for Malignant Tumors[a]

Particle	Tumor Characteristics	Energy deposition	Bragg peak	Radiation source	Accelerator cost[b]
Photons	Rapidly growing, oxygenated	Low LET	No	Cobalt 60; electron linac;microtron	1
Electrons	Superficial	Low LET	No	Electron Linac; microtron	1–2
Protons	Early stage, near-critical structures	Low LET	Yes	Synchroton; cyclotron	10–15
Fast neutrons	Slow growing, hypoxic	High LET	No	Proton linac; cyclotron;	8–10
Heavy ions	Same as fast neutrons	High LET	Yes	Synchrotron	40
Pions	Same as fast neutrons	High LET	Yes	Proton linac; cyclotron	35–40
Slow neutrons	Glioblastoma; Some melanomas	Very high LET with BNCT	No	Low energy accelerator; nuclear reactor	1–2

[a]Adopted from Table 3 of Ref. 5.
[b]Relative cost not including building and clinical equipment costs. Assumes room-temperature magnets. Superconducting magnets can reduce accelerator cost in some cases (e.g., see Ref. 24).

(Fig. 7) (8). The electron LINAC usually is used in conjunction with an isocentric gantry (Fig. 8) (8) to permit stereotactic treatment with minimal movement of the patient.

In principle, with an electron beam of sufficient intensity and energy, the direct beam also can be used for radiotherapy. However, as noted, the large collisional scattering of the electrons in tissue results in a large angular spread (i.e., penumbra) (2) of the beam (Fig. 4), which limits the usefulness of the direct beam for therapy other than for treatment of skin cancer, other shallow tumors, or for noncancerous skin diseases (Table 2). However, owing to their compactness, a gantry-mounted electron-beam system (Fig. 8) can also be used in an operating room to directly irradiate an internal tumor (or surrounding area) made accessible via surgery. This technique, interoperative electron-beam radiation therapy (IOERT), may be particularly advantageous for treatment of certain types of cancers when combined with more conventional treatment (37).

It was once thought that higher energy electrons ($E > 25$ MeV) and the corresponding high energy X rays produced could have advantages in radiotherapy relative to the more typical energies used ($E < 25$ MeV). These higher energy electrons were typically produced using a special LINAC or a compact race-track microtron (36) accelerator (Fig. 9) with a separate gantry for the beam (Fig. 8). High energy electrons possibly suitable for radiation therapy also recently have been produced using high power compact pulsed lasers (38).

At the higher electron energies, nuclear reactions and significant bremstrahlung can take place and contribute to the patient dose. Although the higher energy may have some benefit, it is generally offset by the larger dose imparted to surrounding healthy tissue because of the longer range of the secondary electrons involved. Likewise, improvements in radiotherapy techniques using lower energy photons, such as intensity-modulated radiotherapy (IMRT) and Monte-Carlo treatment planning (17,18), have offset most of the advantages of higher energies. Thus, few

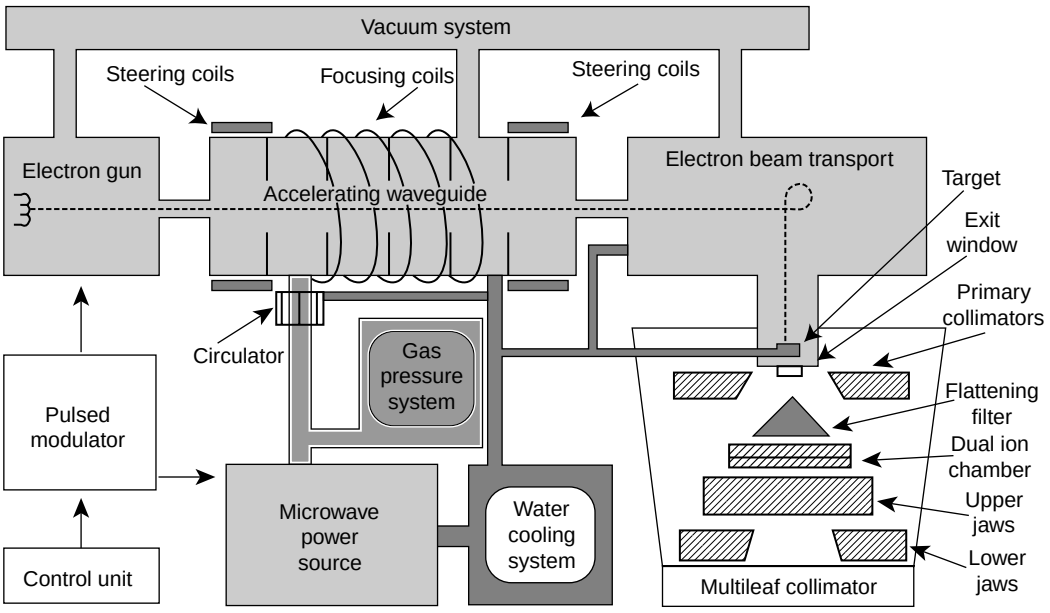


Figure 7. Schematic diagram of an electron LINAC of the type used for radiation therapy (8).

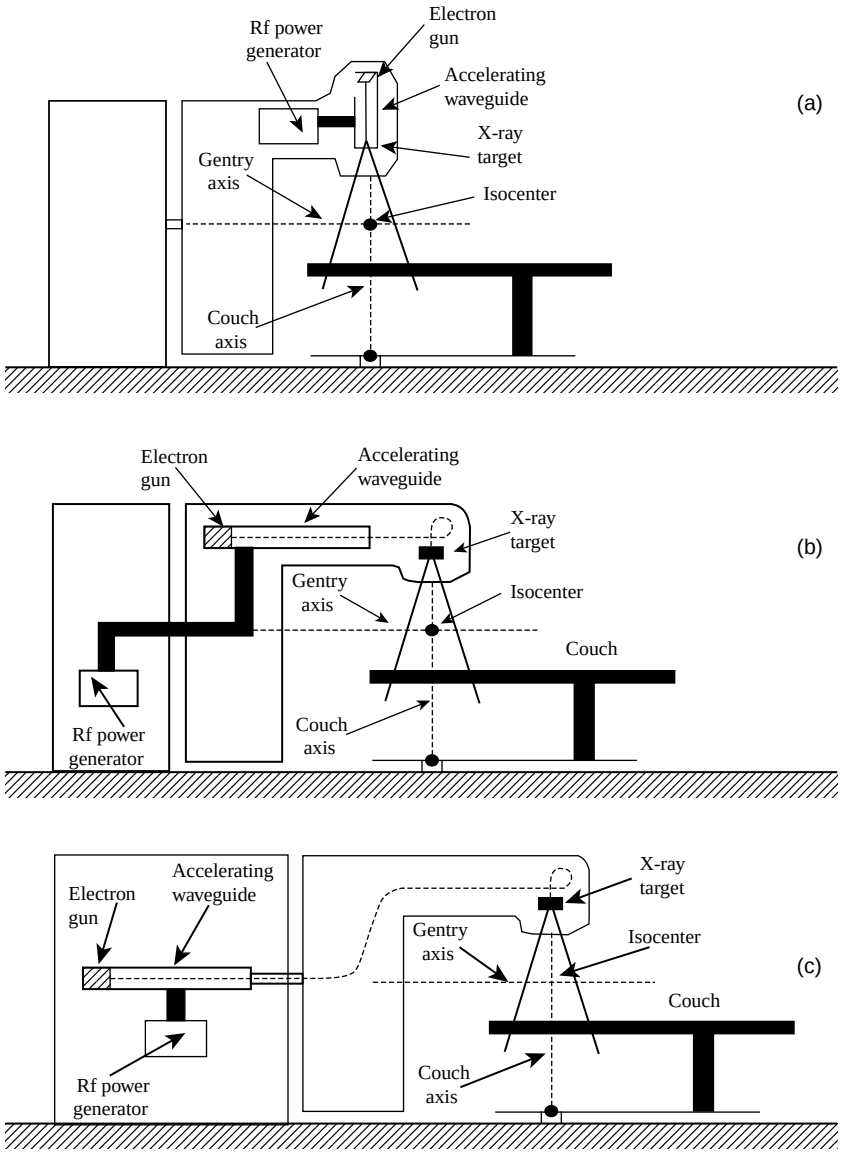


Figure 8. Various isocentric gentry arrangements used with electron accelerators for radiation treatment using secondary X rays or the direct electron beam (8).

facilities use electron beams or X rays with energies greater than 25 MeV. However, it has been indicated by theoretical calculations and by recent experiments with phantoms (21,39–44) that high energy direct electron beams confined by a high magnetic field potentially could be useful for some cancer treatments (see below).

FAST AND SLOW NEUTRON RADIOTHERAPY

Fast Neutrons

As noted previously, it was realized shortly after their discovery in 1932 by Chadwick that fast neutrons could have advantages over X rays and γ rays in cancer therapy (5–7,14) Unlike protons and α particles, neutrons being massive, yet uncharged, can penetrate more easily into tissue even at modest energies (e.g., 20–40 MeV in kinetic energy). Such neutron energies are readily available via nuclear reactions such as 9Be (d, n) using conventional cyclotrons, including many of those available during the 1930s (6). Hence, fast neutrons were used in cancer treatment well before high energy protons, α, and other heavy particles became available.

As a neutron therapy beam is produced as a secondary beam from a nuclear reaction on a production target with a primary charged-particle beam, the associated accelerator generally must have a high primary beam current (e.g., a proton or deuteron beam at the μA level). The accelerator facility also must then have the necessary massive shielding for fast neutrons. In some cases, the primary accelerator also can then be used to produce radioisotopes (such as 18F and 11C) that can be used for positron emission tomography (PET). Like other early cancer-therapy accelerators, most of the first-generation fast-neutron treatment facilities were adjuncts to nuclear research facilities. Recently, several new dedicated fast-neutron treatment facilities have become available (6,23,45).

Neutrons in tissue (and other material) behave quite differently than protons, α particles, or other heavy charged particles. Specifically, being uncharged, the dose

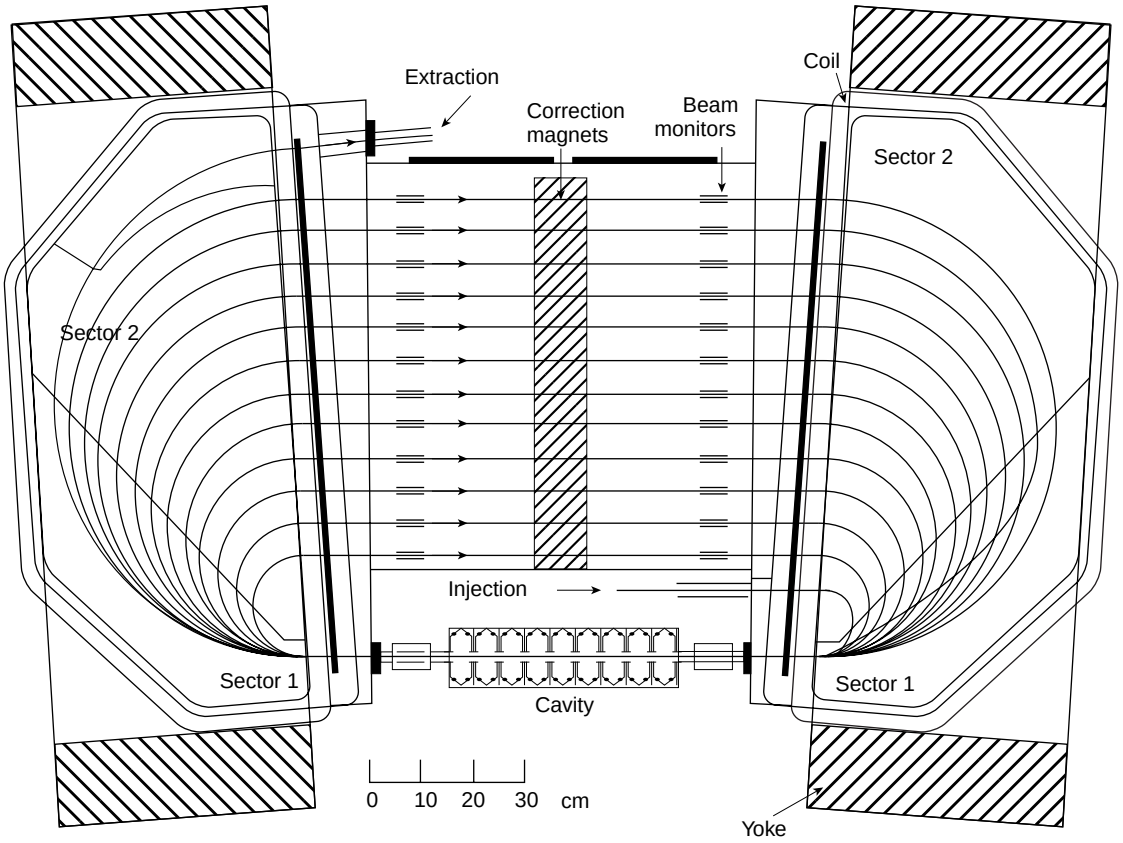


Figure 9. Compact race-track electron microtron used to produce high energy electrons; $E > 25$ MeV (36).

deposited along the path is *not* due to a large number of low LET collisions with atomic electrons. Instead, large-LET nuclear collisions play a dominant role with neutron-proton and neutron-nucleus collisions important in tissue (2,6,7). The latter produce secondary ionization and biological dose due to the recoil protons and other recoiling tissue nuclei. In this sense, recoil protons (and other recoil nuclei) play the role that the secondary electrons play in X-ray and γ-ray therapy. It, then, is not surprising that the deposited dose curve due to fast neutrons looks similar to that from X rays and γ rays (Fig. 2). However, the RBE for neutrons is generally greater than that of the latter e.g., $\times 4$ or larger (2,5–7,23,45).

Like X rays and γ rays, and unlike charged particles, neutrons are attenuated and scattered in tissue and do not have a well-defined range. Treatment of a localized tumor requires either highly fractionated doses or stereotactic treatment to spare healthy tissue, which, together with uncertainties in the RBE for neutrons, was a major problem in early studies of fast-neutron cancer therapy. Many patients were found to suffer long-term complications from the treatment (7), and many of the early neutron treatment facilities ceased operation until the latent effects of the therapy were better understood.

Today, fast-neutron therapy is usually restricted to treatment of special types of tumors (Table 2) where this type of therapy has been shown to be advantageous, yet without a high probability of long-term complications, or where such complications may be justified (e.g., for older patients), which includes cancers of the salivary glands, prostate cancer, and several types of soft-tissue and inoperable cancers. Neutrons have been shown to be especially advantageous (23,45) in treating radiation-resistant cancer cells.

A recent state-of-the-art dedicated fast-neutron treatment facility is the one located at the Harper-Grace Hospital in Detroit, Michigan (23,45), which uses an innovative gantry-mounted, compact super-conducting cyclotron (45) to produce a high intensity 48 MeV deuteron beam (Fig. 10) (23). This beam then impinges on a beryllium target to produce a range of MeV-energy neutrons for treatment. Having the cyclotron mounted on the treatment gantry minimizes the size of the facility. A special multipin collimator is used to collimate the 2D profile of the treatment beam. Treatments are generally fractionated and done in a special prescribed sequence with X-ray therapy to enhance destruction of radiation-resistant tumor cells that could later metastasize. This type of treatment appears to significantly enhance the long-term survival rate for certain stages of prostate cancer (23,45).

Slow Neutrons and BNCT

As suggested by G. Locher in 1936, one method to increase the localized dose from fast and slow neutrons is to tag tumors with certain elements such as boron, which preferentially capture neutrons (46). Boron, specifically the isotope 10B, which is about 20% of natural boron, has a

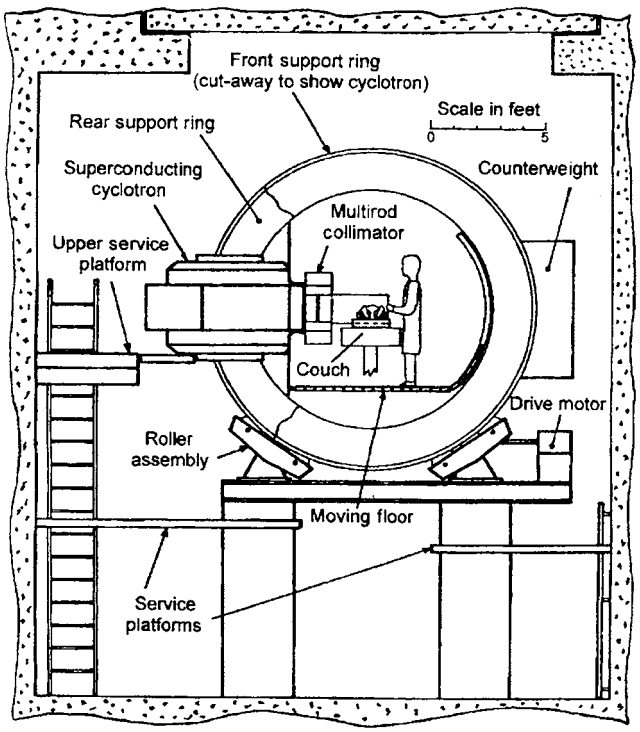


Figure 10. Schematic of the Harper-Grace Hospital (Detroit) gantry-mounted compact super-conducting cyclotron and treatment area used for fast-neutron radiotherapy (23).

high cross section for the reaction 10B (n, α) and produces (Fig. 11) (23) highly localized, low energy α particles (1.47 MeV) and 7Li recoil ions (0.8 MeV), which quickly stop, yielding a highly localized LET that greatly enhances the local dose to a tumor. The boron is preferentially attached to the tumor site using a tumor-specific boron-loaded pharmaceutical (46,47). As this technique works particularly well with slow neutrons (keV to MeV), it can be used with slow or low energy neutrons produced from small accelerators or from nuclear reactors, which is the basis for boron-neutron capture therapy (BNCT), and a number of clinical trials using BNCT are underway, primarily outside of the United States.

Related to BNCT, the manmade isotope 252Cf, which produces both energetic α particles ($E\alpha = 6.1$ MeV) together with fission fragments and associated fission neutrons ($En \approx 2$–3 MeV), has been proposed (46,47), together with boron-loaded tumor-specific compounds, as a special form of BNCT brychatherapy.

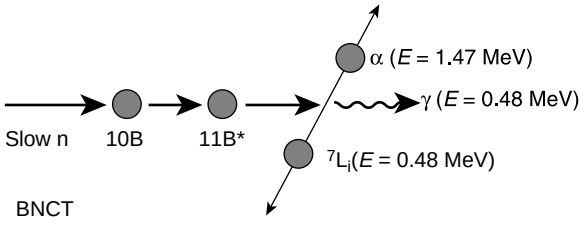


Figure 11. Illustration of the nuclear processes involved in BNCT (adopted from Ref. 23).

Proton-Beam Radiotherapy

Excluding electron beams, which, as noted, have limitations for treatment of deep tumors, protons are presently the primary beams used in particle-beam therapy. A number of cyclotrons, originally operated as nuclear research facilities, have since been converted for use as medical treatment facilities (Table 1). A conventional cyclotron i.e., one that uses a fixed radio frequency (RF) accelerating voltage and a single large conventional magnet (6) is typically limited in proton beam energy to less than 80 MeV, which limits the penetration depth to less than a few cm in tissue (Fig. 2). Hence, these facilities (about one-third of all proton-beam therapy facilities) are generally limited to treatment of shallow tumors, especially eye tumors, or for treatment of certain noncancerous conditions such as age-induced macular degeneration (AMD) (10,12,48,49). As many of these facilities are located in a nuclear research laboratory rather than in or near a hospital, certain treatments requiring fractionated doses or treatment done in combination with other modalities can be problematic.

Treatment of deep-seated tumors requires a proton beam energy of 200 MeV or more (Fig. 2). At this energy, the protons' relativistic increase of mass with increasing velocity requires the use of a large separated-sector cyclotron or high field cyclotron, such as those in use at Indiana University, Massachusetts General Hospital (Fig. 12) (49), and elsewhere, or a "race track" synchrotron adapted from high energy physics. An example of the latter is the synchrotron at the Loma Linda proton-beam treatment facility (49). Synchrotrons are also generally used to produce heavier particles [e.g., 12C ions used for radiation treatment (see below)].

A cyclotron produces a beam with a 100% macroscopic duty cycle, although it still has a beam modulated by the accelerating voltage RF, typically tens of MHz (6). The synchrotron produces a beam modulated by the ramping time of the variable-field accelerator magnets. The latter can be on the order of seconds (6), which generally is not a problem in radiation therapy and can be used as an advantage in some treatments. A third type of accelerator, the synchro-cyclotron, developed after WWII at LBL and used at LBL for α-beam radiotherapy for many years, is presently only in limited use (Table 1).

As the direct proton beam in these accelerators is used for therapy, only a modest beam intensity is required relative to the beam intensities needed for a nuclear research accelerator or one used to produce neutrons or pions, *viz*. namps vs μamps, which can simplify the accelerator design, the building, and shielding required . However, in some cases, a high intensity beam is desirable in order to produce radioisotopes used in PET and other imaging methods.

All types of proton facilities, owing to the high "magnetic-rigidity" of high energy protons, require a set of large (and costly) beam-switching magnets and patient-treatment gantries (Fig. 13) (49). Likewise, raster-scanning the beam and varying the dose-depth required to treat a particular tumor is not trivial, which can be done (Fig. 14) (49) with electronic elements (active scanning) or with shaped absorbers (passive modulation).

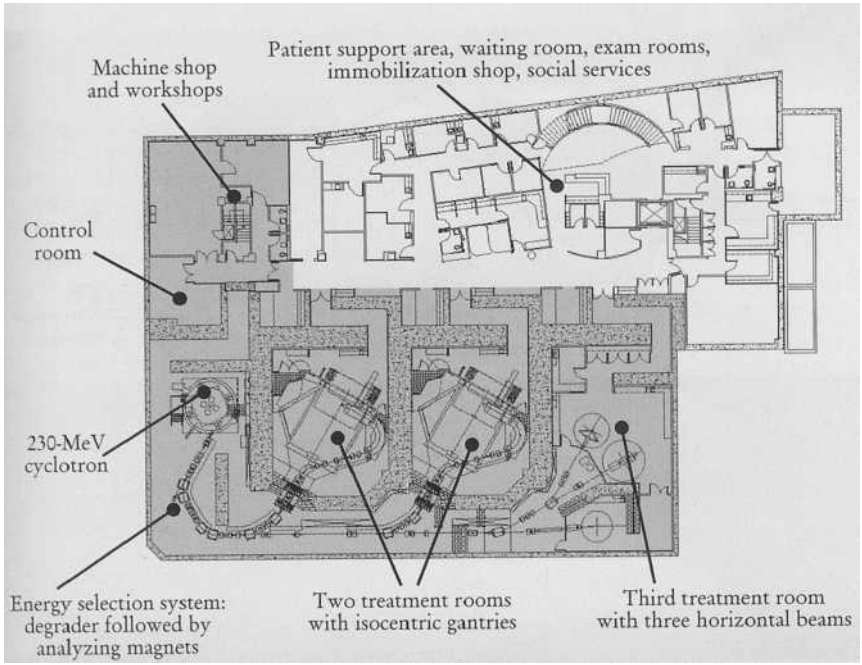


Figure 12. High field spiral-ridge 230 MeV proton cyclotron facility and treatment room layout of the Massachosetts General Hospital Northeast regional proton-beam radiotherapy facility. (Reprinted with permission from Ref. 49, copyright 2002 American Physical Society.)

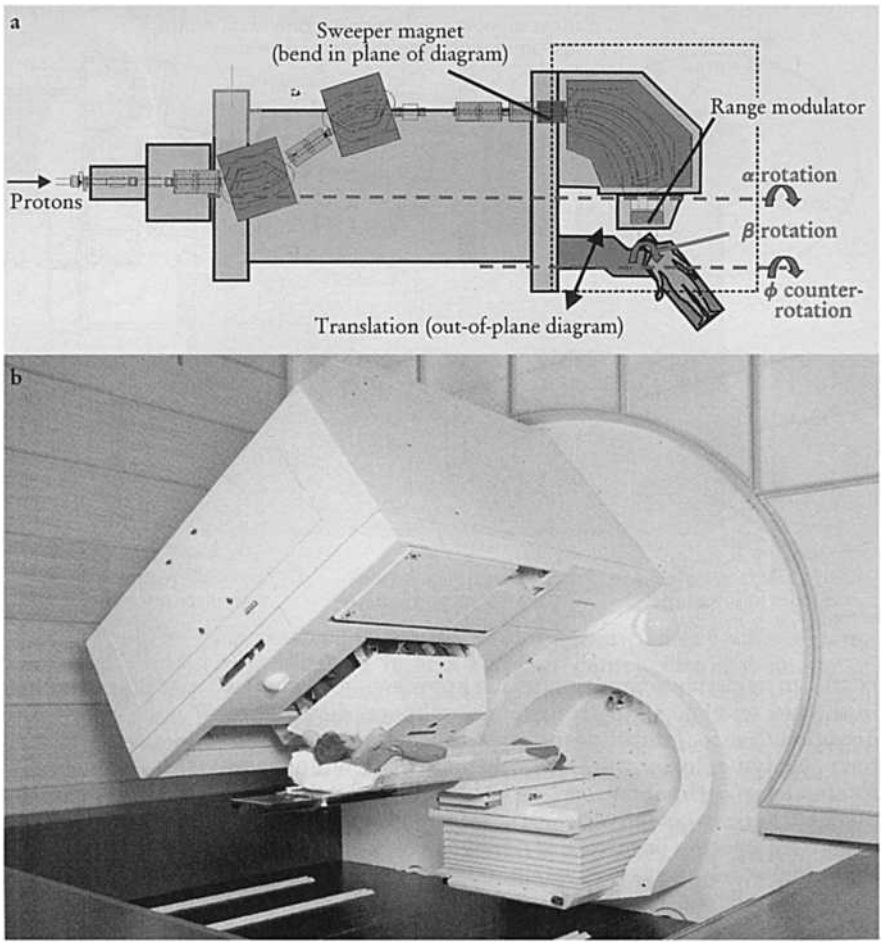


Figure 13. The isocentric proton-beam gantry used at the PSI (Switzerland) proton-beam cancer treatment facility. (Reprinted with permission from Ref. 49, copyright 2002 American Physical Society.)

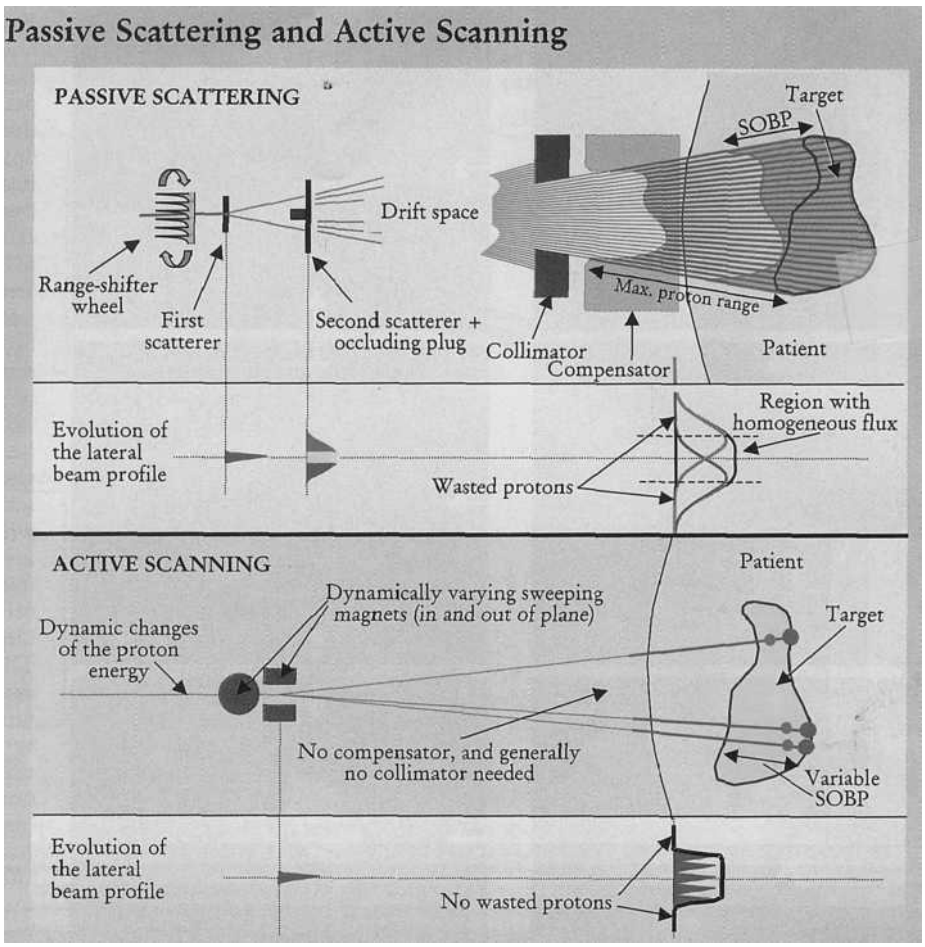


Figure 14. Schematics of typical passive and active ion-beam raster-scanning apparatus used to generate suitable dose profiles for treatment of specific tumors. (Reprinted with permission from Ref. 49, copyright 2002 American Physical Society.)

Although the primary energy loss mechanism and, hence, dose from high energy protons is still due to collisions with atomic electrons (2), a significant fraction of the incident beam can induce nuclear reactions in beam collimators and in the patient, which, in the patient, can lead to the production of neutrons or other reaction products complicating the calculation of the biological dose. Some of these reactions can be detected by observing the annihilation radiation (i.e., emission of back-to-back 511 keV γ rays) from proton-rich positron-emitting nuclear reaction products (9,10,50). This technique (based on PET) has been adapted to image the dose profile of heavy-ion beams used in radio therapy (see below) where the beam itself may fragment (10,11,15,16) into positron-emitting nuclei.

Related to this technique, collimating an energetic proton beam and stopping it (including in a patient) can produce copious amounts of fast neutrons. Although the direct internal dose to the patient due to the neutrons may be small, the facility itself may require extensive neutron shielding for protection of the workers and surrounding environment (Fig. 12), which results in a very large "footprint" for such a facility and a building cost that often exceeds the cost of the accelerator and beam line themselves. As a result, only a few hospital-based proton-therapy facilities presently exist. These facilities are generally dedicated to the treatment of certain tumors (Table 2) and for conditions that are otherwise inoperable or difficult to treat with conventional radiotherapy. However, it is estimated that, in the United States alone, over 10,000 patients/year could benefit from proton-beam radiotherapy treatment (12).

As the advantages of proton therapy becomes better documented for certain types of tumors, more medical insurance companies will likely approve treatment, which would justify the construction of more facilities. Nonetheless, the number of patients treated worldwide with protons has increased steadily, with over 40,000 treated through 2004 (Table 1).

Heavy-Ion Radiotherapy

Excluding antimatter particles, the particle beams with the highest LET and, hence, dose rate per unit path length (i.e., dose-depth profile) are heavy ions (HI) with $z \geq 2$. (Eq. 1), which leads to an extremely sharp Bragg peak and, hence, localization of the dose near the end of the HIs' range (2,10,11,14). Also, the biological dose is enhanced over the HI LET alone owing to the large values of RBE determined for HIs. The latter can be on the order of 10 or more near the end of the HIs' range (15).

As with proton therapy, in a typical treatment, the sharp Bragg peak in the dose curve is spread out using absorbers (or other means) to provide a suitable overlap with the tumor (SOBP) (9–11,14). Like protons, energetic HIs, like those needed for therapy (typically a few hundred MeV/nucleon), can also induce nuclear reactions. In the case of HI therapy beams, these reactions, which are primarily beam fragmentation, can transmute the incident beam into other nuclear species (9–11,14–16), which usually include lighter, lower z fragments, which can extend the radiation dose well beyond the range of the primary HI beam (16). Fortunately, many of the fragments and residual nuclei produced in HI (or, as noted, proton) beam-induced nuclear reactions are positron emitters and their intensity and location can be imaged via the back-to-back 511 keV γ rays emitted (PET) (9,10,50–52), which is now being actively exploited in treatment planning at GSI and elsewhere (see below).

It also has been suggested to specifically produce a short-lived positron-emitting secondary beam such as 9C for radiation treatment and, thus, facilitate direct imaging of the treatment beam itself (51), however, while feasible, it is not yet a practical option. Among other problems, such beams are easily fragmented, and as a secondary beam, production requires a high intensity primary HI beam accelerator. Instead, like proton therapy, a direct, low intensity HI beam is more practical for therapy (53–55).

Excluding the use of natural α-emitting radioactive sources, the use of HIs in cancer therapy was pioneered at the University of California, Berkeley, laboratory now known as the Lawrence Berkeley Laboratory (LBL). The Bevatron at LBL, which originally was contructed to discover the antiproton, was converted to accelerate HIs such as 12C, 16O, and 20Ne at energies up to several hundred MeV/nucleon (9–11,14). These energies are those required for cancer treatment of deep tumors. Over 400 patients were treated from 1975 to 1992 at LBL with HIs.

As part of this program, many techniques were developed for controlling and monitoring the HI dose-depth profiles, and providing a suitable SOBP when needed, which included imaging positron-emitting HI beam fragments (9–11,14) and reaction products, which, as noted, is a technique later adapted for proton and HI therapy at other facilities.

More recently, the HI nuclear research facility at Darmstadt, Germany (GSI) has run a HI prototype cancer therapy facility primarily using 12C beams at energies of several hundred MeV/nucleon. They have implemented the online PET imaging technique to deduce dose-depth profiles (52) for clinical treatment planning and verification (Fig. 15) (52). They, likewise, have done extensive measurements and modeling to determine the RBE appropriate for HIs in tissue (15), which, as noted previously, can be relatively large (x10 or more) and thus must be included in treatment planning (51–56). (The latter also has implications for space travel and other activities involving radiation from heavy particles in cosmic rays, etc.).

Many of the tumors treated at the GSI facility cannot be optimally treated with conventional radiotherapy or surgery due to the close proximity of the tumor to a critical area (Fig. 15). As with other particle-beam therapies, most

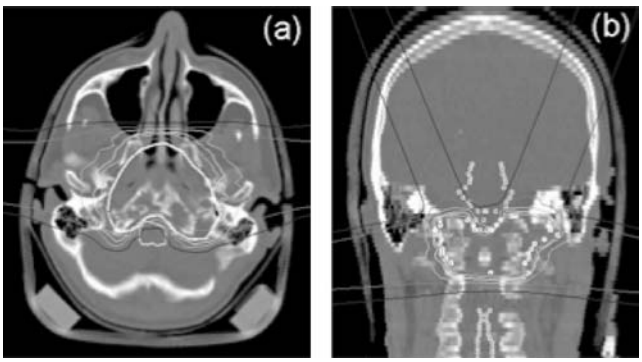


Figure 15. Example of a tumor treatment plan used for HI radiotherapy at GSI (52).

HI treatments involve fractionated doses, and the HI therapy is often used in conjunction with other treatment modalities such as chemotherapy or photon therapy. Like other pioneering particle-beam facilities, GSI is a nuclear research facility and not *a priori* part of a hospital facility, which can often limit patient throughput. Nonetheless, several hundred patients have been treated since 1997 (Table 1). Based on the results demonstrated, the GSI nuclear and biomedical research group together with a hospital facility in Heidelberg, Germany is finishing the constructing of a dedicated HI accelerator and HI treatment facility. This facility is expected to become fully operational in mid-2005 and will serve as a German national treatment facility for the 3000 or more patients identified as optimal for HI therapy each year in Germany. A similar facility is under contruction in Italy (12).

At present (2005), the primary facilities built and dedicated to HI cancer therapy are the HIBMC and HIMAC facilities (53) in Japan (Fig. 16) (54). As at GSI, these facilities primarily use 12C ions at energies of several hundred MeV/nucleon. (HIBMC also has the capability to use protons in radiotherapy). Several thousand patients have been treated at HIBMC and HIMAC (Table 1). Like proton therapy facilities, the large footprint (Fig. 16) and costs associated with an HI accelerator and the associated treatment facilities will generally limit their availability. In countries with national health services, one, two, or at most three HI facilities may accommodate those patients who might benefit from HI therapy. The situation becomes complicated in countries like the United States where private insurers must approve treatment.

FUTURE DEVELOPMENTS

Magnetically-Confined Electron Beams

It has been suggested, and recently demonstrated with experiments, that one might use high energy electrons ($E \leq 100$ MeV) for radiotherapy with suitable magnetic fields applied to reduce (Fig. 17)(21) the pneumbra from scattering (21, 31–44). Again, if the direct electron beam is used, relatively low intensity beams can be used, which simplifies the accelerator, beam handling, and shielding. Most hospitals operate and maintain electron accelerators ($E = 10$–25 MeV), mostly LINACs, for radiation therapy

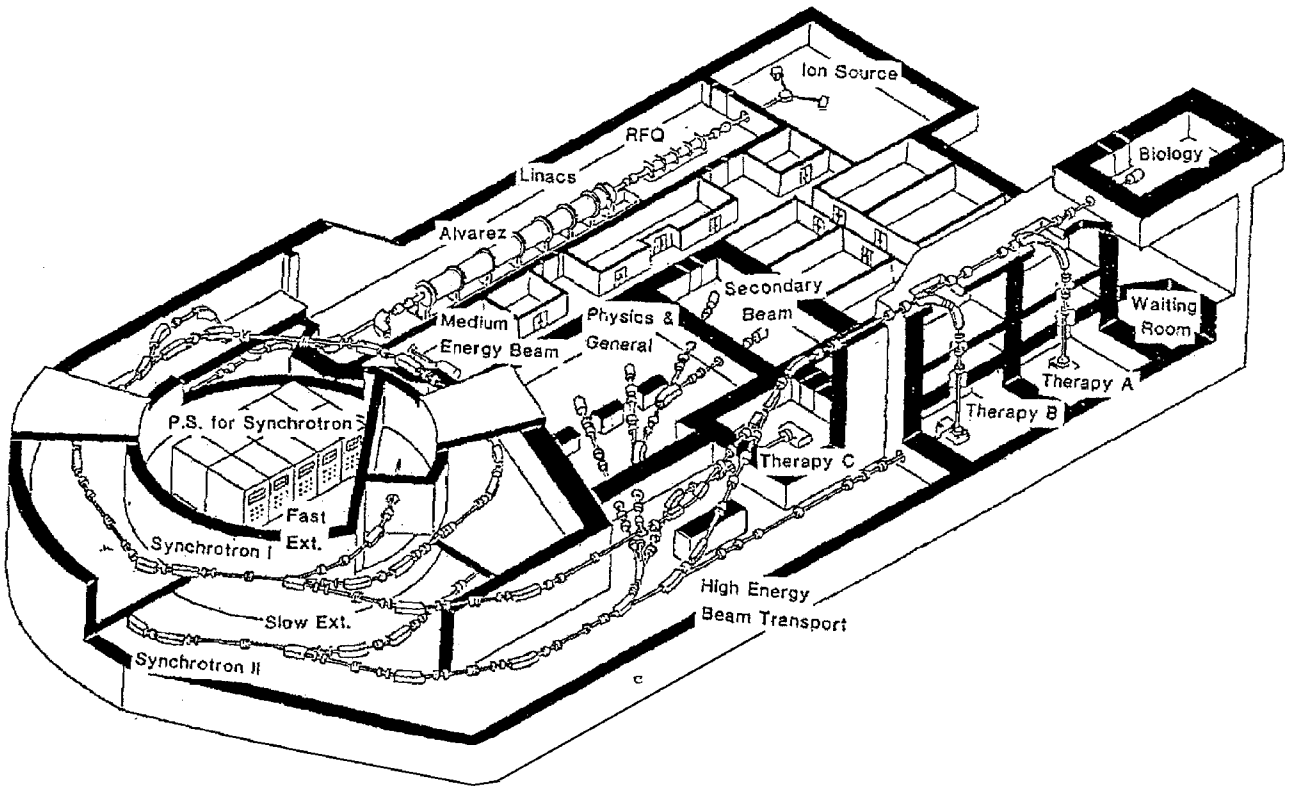


Figure 16. Layout of the HI-beam cancer treatment facility HIMAC in Chiba, Japan (adopted from Ref. 54).

and higher energy electrons (50–100 MeV) are readily produced using expanded versions of LINACS (6) or table-top race-track microtrons (36). Both are quite feasible for operation at most hospital clinical oncology facilities.

Recent developments in super-conducting magnet technology including gantry-mounted systems and LHe-free systems make such magnets technically feasible. These magnets could be used in conjunction with a suitable electron accelerator for cancer therapy for certain types of tumors (soft tissue, etc.). A sample of a calculated multibeam dose profile in a skull-tissue phantom using 35 MeV electrons confined with an axial (solenoidal) $B = 6$ T magnetic field (44) is shown in Fig. 18.

Superconducting Accelerators and Beam Gantries

One method to possibly reduce the footprint (and cost) of proton and HI radiotherapy facilities is to use, more widely, superconducting magnet technology for both the accelerator and beam line components (24). However, extensive radiation shielding may still be required.

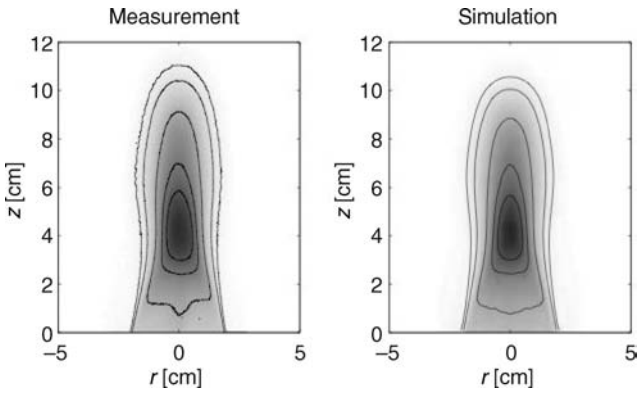


Figure 17. Measured and calculated 2D isodose profiles for a *ca*. 20 MeV electron beam incident on a tissue-equivalent phantom with an applied 2 T longitudinal magnetic field. Compare with Fig. 4 (adopted from Ref. 21).

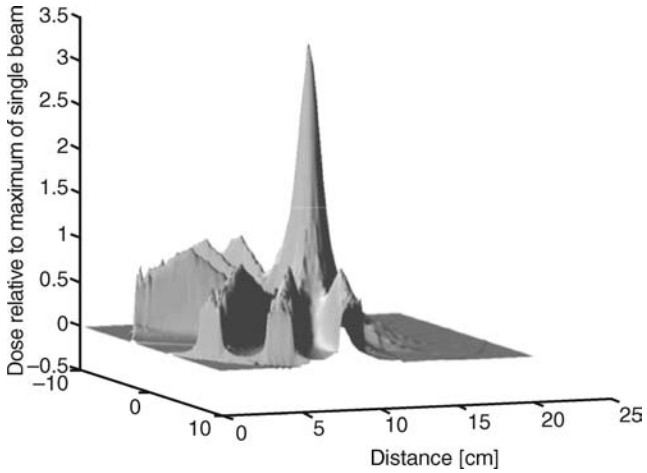


Figure 18. A simulated stereotatic dose profile in a head phantom for a set of 35 MeV electron beams confined by a $B = 6$ T longitudinal magnetic field (44).

Pulsed-Laser Accelerators

It is now possible to produce very high electric fields in plasmas using compact ultra-fast pulsed lasers, which can be used to accelerate electrons, protons, and other ions to MeV energies (38). Although the particle-beam intensities and, in some cases, energies demonstrated so far are less than those needed for radiotherapy, such "table-top" accelerators may prove viable in the future.

ACKNOWLEDGMENTS

The author thanks J. Sisterson, Ph.D., Yu Chen, Ph.D., D. Litzenberg, Ph.D., Prof. L. Jones, and Hao Jiang for their assistance.

References are on page 609.

See also IONIZING RADIATION, BIOLOGICAL EFFECTS OF; RADIATION DOSIMETRY FOR ONCOLOGY; RADIOTHERAPY TREATMENT PLANNING, OPTIMIZATION OF.

RADIOTHERAPY, INTRAOPERATIVE

PETER J. BIGGS
Harvard Medical School
Boston, Massachusetts

INTRODUCTION

What Is Intraoperative Radiotherapy?

Intraoperative radiotherapy (IORT) is a technique that combines radiation therapy with surgery to irradiate tumors *in situ*, without delivering a significant dose to surrounding normal, critical structures and is usually performed using electron beams from linear accelerators. This contrasts with external beam therapy using X-ray beams where the dose that can safely be delivered to most tumors is limited by the dose that is given consequentially to normal, critical structures.

The practice of IORT began soon after X rays were used therapeutically, almost 100 years ago. However, although several investigators used this technique over the ensuing years with various X-ray modalities, it was not until the mid-1960s that intraoperative radiotherapy made a serious mark on the field of radiotherapy with the work of Abe (1–4) in Japan using electron beams from linear accelerators. This was rapidly followed in the mid-1970s by investigations in the United States, first at Howard University (5), then at the Massachusetts General Hospital (MGH) (6), followed by the National Cancer Institute (7), and then the Mayo Clinic (8). The reason for this dramatic change was due to the introduction of linear accelerators capable of generating high energy electron beams.

In 1992, Coia and Hanks (9) reported on patterns of care study, which indicated that of 1293 radiation oncology facilities in the United States, 108 reported doing IORT, of which 29 have two or more residents. Since there were $\sim$88 training programs in existence at that time, roughly one-third of hospitals or medical centers with residency training programs were performing IORT. They did not indicate whether or not this list included only electron beam IORT or other modalities.

Initially, IORT flourished in both the academic and community hospital setting, but it is clear from informal surveys and anecdotal evidence that fewer centers are now performing IORT compared with 1992. The reasons for this decline in interest are twofold. First, establishing the usefulness of IORT as a beneficial adjunctive therapy has been difficult. Second, IORT as practiced by the majority of centers, those that do not have the luxury of a dedicated or conventional mobile linear accelerator in the operating room (OR), is technically difficult and demands time on the part of a clinical staff that is under great time constraints, brought on by the present reimbursement climate. Thus, this method taxes the interests of all the parties after a number of years. The uphill battle faced by proponents of intraoperative radiation therapy is the high cost of a dedicated facility in the operating room. A dedicated linear accelerator in an operating room is no longer a cost-effective option for any hospital (10), due to the cost of the machine as well as the radiation shielding. The entry into the field of IORT of mobile linear accelerators that can be used in existing OR rooms without requiring additional shielding makes the cost and logistics of setting up an IORT program much easier and therefore provides a stimulus to the field. There are now three manufacturers of such equipment and $>$30 units have been installed in the United States and Europe.

In addition to using electron beams from linear accelerators, many other radiotherapy modalities can be classified as IORT. They all share the same principle that the dose is delivered only locally, so that dose to the skin uninvolved adjacent tissues and organs is minimized. These include high dose rate brachytherapy that is delivered in an operative setting and stereotactic radiosurgery using a 50 kV X-ray device. The Papillon technique (11) was a technique whereby 50 kVp X rays were used to irradiate lesions on the rectal wall by inserting the X-ray tube into the dilated rectum. Orthovoltage X rays are still used in one or two places (12), based on the issue of lower cost, but there is a clinical price to pay since the dose to any bone in the field is much greater than the prescribed dose to tissue, and this can result in osteoradionecrosis. However, this can be obviated to some extent by heavy filtration of the beam at the cost of a lower dose rate.

General Description of the Treatment Apparatus

Historically, adapters were made to fit existing linear accelerators so that the electron beam could be directed onto the tumor while at the same time protecting normal tissue. This was first achieved by having an applicator, $\sim$30 cm long and generally circular in cross-section that mates with another cylinder attached to the head of the linear accelerator. This mating piece is centered with the radiation beam and only slightly larger than the treatment applicator. Thus, by adjusting couch movements very carefully, the treatment applicator can be slid into the mating piece, even when the gantry is angled far from the vertical. The greater the gap between the two pieces, the easier it is to achieve docking, but the greater the possible degree of

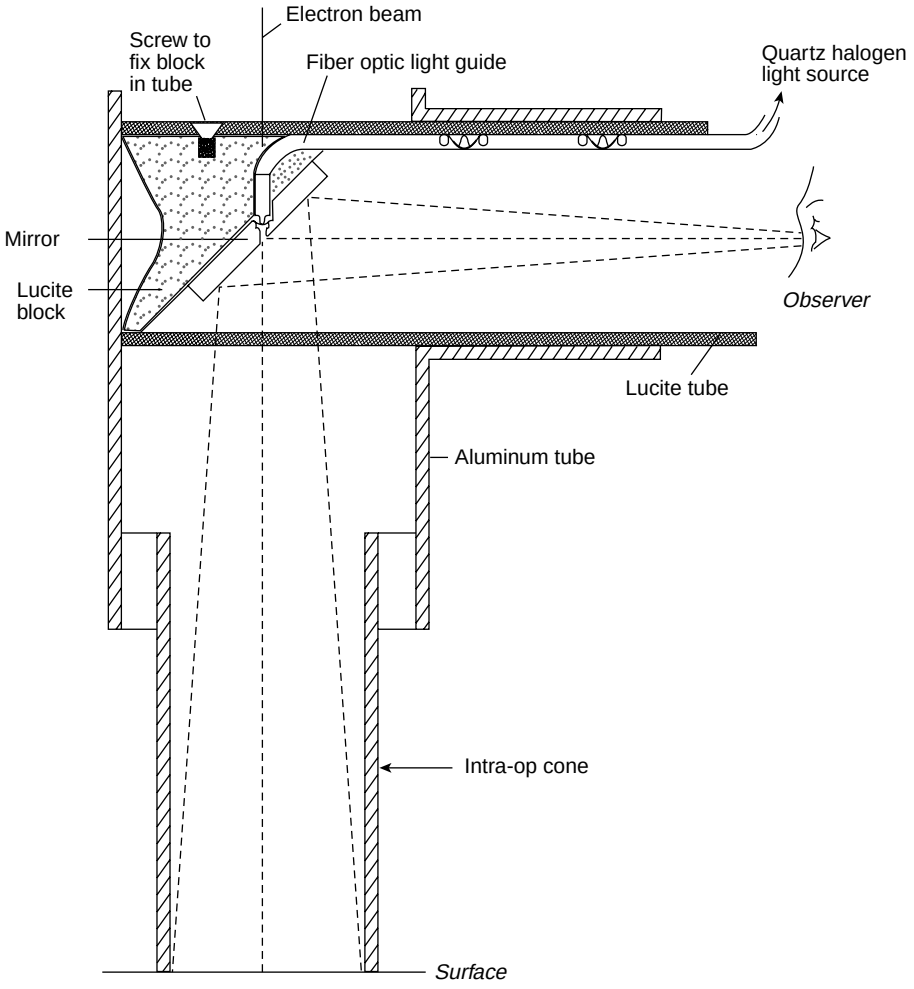


Figure 1. Diagram of optical system used for a hard dock system. Note that the unit slides in and out of the field and that the reflector is metallic.

misalignment, meaning the beam is not aligned with the applicator in the patient, so the beam, as delivered, will have different dose characteristics from a perfectly aligned system. Typically, the gap between the two cylindrical pieces is ~0.2–0.3 mm. This method is known as the hard dock technique (13). The advantage of this method is that the system is self-aligning, within limits defined by the gap between the two mating pieces, but the disadvantages are that (1) the docking process can be lengthy, depending on the gantry angle, (2) there is a potential safety issue if either the couch or the gantry moves during the procedure, and (3) the treatment field is no longer visible once the docking process is complete. Hence, the hard dock procedure requires a lock-out system for the drive motors during the treatment when there is no one in the treatment room and an optical system to view the field enclosed by the applicator after the applicator has been docked. A cut-away view of such an optical system is shown in Fig. 1. A photograph of a completed hard dock process is shown in Fig. 2. This photo shows that the docking or alignment process can be a complex and time-consuming task in some situations.

The alternative method that avoids the problem of the hard dock procedure is the soft dock procedure. In this method, the treatment applicator is separated from the head of the machine and thus the potential for patient injury is greatly lessened and the treatment field can be

Figure 2. Example of electron beam alignment using the hard docking process. The linear accelerator is a conventional machine in the therapy department and the applicator was fabricated in-house.

viewed even after the docking is complete. The disadvantage of this method is that an alternative to the simple mechanical alignment is needed. This need has been answered by a variety of optical systems as explained below. A description of two commercial alignment systems is given below. A description of several noncommercial soft-docking systems has been published by several authors (14–17).

The Clinical Rationale for IORT

To understand the rationale for intraoperative radiotherapy, it is necessary to understand some of the basic principles of radiotherapy. Patients are routinely treated for cancer using fractionated radiation. This means that the radiation is delivered in increments on a daily basis, 5 days $week^{-1}$ for up to 8 weeks, depending on the lesion under treatment. The reason for doing this is that if one were to deliver a tumoricidal amount of radiation to the tumor in one or a few fractions, serious long-term side effects to normal tissue and organs would result. By fractionating the radiation, an equivalent tumoricidal dose can be given to the tumor without serious long-term side effects. Fractionation schemes currently in use have developed empirically over the history of radiotherapy going back to the use of early X-ray tubes. The daily fraction dose is limited by acute radiation effects. These effects, such as reddening of the skin in the era of low energy X rays, bowel problems, and so on, appear during the course of treatment, but will generally resolve themselves without long-term consequences after the radiation treatment has been completed. This maximum radiation dose may produce acute effects in a few patients, due to biological variation between individuals and, hence, their response to radiation. The prescription dose or the maximum overall dose, which is the number of daily fractions times the daily dose, is also limited. This limit is due to normal tissue and organ tolerance. Thus, for example, abdominal radiation for rectal or colon cancer is limited to ~50 Gy because of small bowel complications (Gy is the unit of absorbed dose and is expressed in $J \cdot kg^{-1}$). The kidneys can tolerate a dose of no more than ~15 Gy and treatment for lung cancer requires keeping the dose to the spinal cord below ~45 Gy to avoid transverse myelitis. Some organs, such as the liver, can tolerate varying amounts of radiation, depending on the fraction of liver that is irradiated. The smaller the fraction of organ treated, the larger the dose that can be tolerated. Excess dose to the whole lung can produce fibrosis, resulting in a nonfunctional lung. However, even with these dose limits, local control rates, or control of the primary tumor that is being irradiated is not 100%, so it would be highly desirable to have a method for increasing the dose to the tumor without increasing the dose to the surrounding normal tissue. This is known as improving the therapeutic ratio, which, for a given dose fractionation, is defined as

$$\text{Therapeutic ratio} = \text{tumor control probability/normal tissue complication probability}$$

Ideally, this ratio should be as high as possible. One method to improve this ratio is through the use of intraoperative irradiation. In this technique, the area to be treated, whether a tumor or tumor bed, is exposed during surgery and normal tissue (e.g., small bowel) is moved out of the way by using an applicator though which the radiation is delivered. By irradiating with electrons rather than photons (see section below for comparison between therapeutic photon and electron beams), the radiation can be safely and effectively limited to the area of the tumor. It has been firmly established using experiments on dogs (18,19) that a safe upper limit for single fraction doses is ~20 Gy. Doses delivered through IORT generally range between 7.5 and 20 Gy.

While IORT can increase the therapeutic ratio by excluding normal tissue and organs from the radiation field, it suffers from the fact that it is a single fraction procedure, which, as noted above is limited and therefore, by itself may not provide a curative dose of radiation for all tumors. The reason for this is that such processes as repair of sublethal damage, repopulation, redistribution, and reoxygenation (20), which contribute to enhancing the therapeutic ratio for fractionated radiation, are limited in single-dose therapy. While the initial Japanese study used IORT as the only source of treatment radiation, it has been customary in the United States to use IORT as a boost therapy. This means that the patient is treated using the standard fractionation scheme using external radiation and is given the IORT as an additional boost dose. Thus the tumor dose has been increased with only a small increase in dose to a fraction of the surrounding tissue. Table 1 (21) shows the radiobiological equivalent fractionated dose of single doses of radiation between 10 and 25 Gy. Assuming a conventional fractionation scheme for external therapy of 2 Gy per fraction, a single dose of 10 Gy is equivalent to 17 Gy for tumors. For late effects in normal tissue, this figure is much higher at 26 Gy, but is considerably lower if the dose per fraction is only 50%. Thus normal tissue toxicity determines the maximum dose that can be delivered intraoperatively.

Historical Review of IORT

The first cases of IORT were reported by Beck as far back as 1907 (22) and again in 1909 (23). He used X rays (probably 50 kVp, see below) to treat several cases of stomach and colon cancer. Several years later, Finsterer, in 1915 (24), reported on the treatment of stomach and colon cancer. He used doses between 2500 and 4500 R using X rays with filtration that varied between none, variable thicknesses of

Table 1. Equivalent Dose of a Single Dose of Radiation in Terms of the Fractionated Dose for Both Acute and Late Normal Tissue Reactions

IORT Single Dose, Gy	Equivalent Dose for Tumor and Acute Normal Tissue Reactions[a]	Equivalent Dose for Late Normal Tissue Reactions[a]
10	17	26
15	31	54
20	50	92
25	73	140

[a]2 Gy fractions.

aluminum, or 0.25–0.45 mm Cu. The "R" stands for the roentgen, which is a unit of radiation exposure; 1 Gy is roughly equivalent to 87*R in air. Note that no external radiation was given to these patients. The choice of filtration was dictated by the thickness of the lesion. It was noted by Abe (25), in his historical review of the topic, that the practice of IORT then was different from that of today; however, it is different only in the sense that kilovoltage radiation is exponentially attenuated and delivers its maximum dose at the surface, whereas electron beams have a maximum dose below the surface and the dose falls sharply beyond a given depth, depending on energy. Thus, the intent then was the same as it is now, namely, to give additional dose to the tumor and spare normal tissues, even though there were more practical problems in dose delivery and differences in beam quality. However, there are photon beam modalities used in IORT that get around the problems mentioned above with X rays. High dose rate brachytherapy using radioactive ^{192}Ir sources, for example, is able to deliver a high tumor dose with a low dose to nearby critical structures.

Barth (26) gives a long account of the technique he used for many anatomical sites where he opened the skin and treated the underlying tumor with 50 kVp X rays at 2 cm SSD for very small field ($\sim$2.5 cm diameter) sizes. Many of these sites were in the head and neck region, just below the skin; more deep-seated tumors would have been hard to treat with this technique. Interestingly, Goin and Hoffman (27) describe instances, where IORT was delivered on more than one occasion. He reports on 13 patients, all but 1 of whom received IORT from 2 to 12 times, with overall doses ranging from 500 to 30,672 R. Thus surprisingly, fractionated IORT was practiced then, something that would not be countenanced today.

Current Status of IORT Application (User Surveys)

A survey of current institutions practicing IORT in the United States (28), whether using electron beams, orthovoltage X rays, the Photon RadioSurgery System (PRS) device or by High Dose Rate (HDR) was conducted in 2003. It was found that the number of institutions practicing IORT by each technique is shown in Table 2.

It can be seen that $\sim$55% of the institutions perform IORT in the OR and 75% perform IORT with a dedicated unit. One of the centers using orthovoltage X rays and another using a dedicated linac in a shielded OR are converting to mobile linear accelerators. Approximately 72% of the institutions above responded to the survey, including all six with mobile units and all four dedicated units in the OR. The remaining eight responders consisted of six sites with an OR in the therapy department and two that used patient transport. By comparison with the 1992 survey, it was found that the greatest decline in the number of centers involved with IORT was the group performing IORT by patient transport, and these were primarily community centers. Thus a higher proportion of those sites still practicing IORT are academic centers. Of the institutions using nonmobile units, the average date inception of the program was 1986 ($\pm$4 years), whereas all the mobile units were installed after 1998. The most commonly used energy was 9 MeV, followed by 12 MeV. This certainly justifies the choice of maximum energy of the mobile units. Slightly more institutions use the soft-docking (61%) rather than the hard-docking technique. The most commonly used field size is a 7 cm diameter applicator, followed by a 6 cm diameter applicator. Finally, the average number of IORT treatments performed by the reporting centers (72% of institutions responded) is 500, whereas the number of treatments performed in the last 12 months is 428. The respective numbers per institution are 31.2 $\pm$ 23.5 (range, 10–90) and 26.8 $\pm$ 22.7 (range, 0–70).

A similar survey was recently carried (29) out for European institutions and there are some similarities and differences. Table 3 shows the number in institutions performing each type of IORT.

The chief difference is that there are 40% more sites in Europe than in the United States and that more than one-half the European sites have mobile linear accelerators compared with 16% in the United States. Moreover, >60% of the treatments carried out are for breast cancer compared with almost no cases in the United States. Also, the average number of patients treated is in excess of 1000, compared with 500 in the United States, noted above. Of great importance is the fact that there are a number of clinical trials underway in IORT in Europe, whereas there are no trials in progress in the United States.

What the two continents have in common is the typical applicator sizes and energies. This result is what one would expect if each were treating the same distribution of disease sites. However, since those distributions are not the same, this commonality is quite remarkable. Finally, one other similarity is that a large percentage of institutions are performing a few cases and a few institutions

Table 2. Number of Institutions in the United States Practicing the Various IORT Modalities (2003)

Type of IORT Practiced	Number of Institutions
Mobile linear accelerator	6 (15.8%)
Dedicated linac in shielded OR	4 (10.5%)
OR in radiotherapy department	8 (21.1%)
IORT by patient transport	9 (23.7%)
Orthovoltage X-rays	2 (5.3%)
Intrabeam (50 kV X ray)	5 (13.2)
HDR	4 (10.5)

Table 3. Number of Institutions in Europe Practicing the Various IORT Modalities (2005)

Type of IORT Practiced	Number of Institutions
Linear accelerator in radiotherapy department	11 (31.6%)
Dedicated linac in shielded OR	5 (13.2%)
Mobile linear accelerator	21 (55.3)
Intraoperative interstitial brachytherapy	6 (15.8%)
Intraoperative HDR flaps	7 (18.4%)
Intrabeam (50 kV X ray)	2[a] (5.3%)

[a]This figure may be substantially higher.

(primarily academic centers in the case of the United States) are performing a large number of cases.

IORT TECHNOLOGY

Early Technology

Radiotherapy prior to the 1960s used primarily X-ray machines for the treatment of cancer patients, although in the years just before that Cobalt-60 teletherapy units were coming into widespread use. Schultz (30) describes in great detail the history of the development of X-ray machines of increasing energy. Thus, in the early days, X-ray machines were the only modality to carry out IORT. In the earliest applications of IORT, Practitioners used 50 kV X rays with a focus-to-skin (FSD) distance of ~2 cm with a field size (diameter) of about the same dimension, varying the filtration to achieve sufficient penetration. Henschke and Henschke (31) provide considerable detail on the practical application of this technique to the treatment of large fields. They show that this can be accomplished either by increasing the FSD, primarily to increase the percent depth dose, or by use of multiple, overlapping fields at short FSD. Thus, until X-ray machines operating at higher energies with adequate dose rates at larger distances and covering larger fields became available, the role of IORT was relatively limited. Beginning in the late 1930s through the 1940s, such high energy units became available. Eloesser (32) described the use of an X-ray machine with filtrations between 0.25 and 0.5 mm Cu. He notes that the filtration depends on the thickness of the tumor being treated; the FSD was 30 cm. In 1947, Fairchild and Shorter (33) describe a technique using 250 kV X rays with a half value layer (HVL) of 1.7 mm Cu. At a FSD of 21.7 cm, a dose rate of 100 $R \cdot min^{-1}$ could be delivered over a 13 cm field diameter. Note that since these high energy X-ray beams were not collimated, lead sheets had to be placed around the surgical opening and internal viscera to provide adequate radiation protection.

Whereas the difficulty of using low energy X-ray beams is one of insufficient penetration, the problem with high energy X-ray beams is that, although their intensity is exponentially attenuated, tissues or organs beneath the tumor to be treated can receive a considerable dose of radiation. For this reason and many technical reasons associated with using an X-ray machine in an OR, it is clear that IORT could not have been anything other than an experimental technique pursued by a few investigators.

Recent IORT Technology: Photons versus Electrons

In the 1960s, the first linear accelerators appeared in clinical use for the treatment of cancer. One of the first units in the United States was a 6 MV machine (Varian Medical Associates, Palo Alto, CA), which produced a bremsstrahlung X-ray beam from a beam of 6 MeV electrons. However, it was not until the 1970s that linear accelerators were capable of producing clinical electron beams. Omitted from this historical review is a discussion of the betatron. Designed and built in the early 1940s by Donald Kerst, the first patient was treated with X rays in 1948; electron beams were a natural by product of this machine, although the dose rates were quite low. However, the betatron played no role in the field of IORT, and there are no betatrons currently in clinical use in the United States. Clinical electron beams are produced by passing the electron beam exiting the linear accelerator's waveguide through high atomic number scatterers (34). The great advantage of electron over X-ray beams is in the depth dose curve. Figure 3 shows a comparison of the percent depth dose curves, normalized to 100%, between electron beams of various energies and a 1.5 mm Cu X-ray beam. What is immediately apparent is that while the curve for the X-ray beams fall off more or less exponentially with distance beyond the depth of maximum dose, the curves for electron beams show a very sharply falling dose and have a finite range. This range is a function of the electron beam energy, so the higher the energy the greater the range. A rule of thumb is that for every 3 MeV of electron energy, the depth of the 80% dose changes by 1 cm. There are currently only one or two active IORT programs in the United States using orthovoltage X rays.

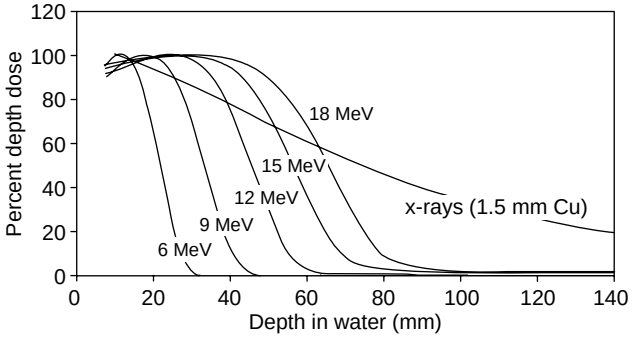


Figure 3. Comparison between electron (6, 9, 12, 15, and 18 MeV) and X ray (1.5 mm Cu X rays) percent depth dose curves.

Dedicated IORT Units

The IORT programs can use one of two approaches. In the first approach, patients are transported from the OR to the radiation therapy department and the IORT treatment is delivered on one of the department's linear accelerators that normally treats outpatients with external beam therapy. Thus IORT is blended in with the other treatments. In the second approach, IORT treatments are delivered on a dedicated machine. Dedicated linear accelerators are those accelerators that are used exclusively for the treatment of patients through the intraoperative technique. The word dedicated is used here to mean that the linear accelerator, conventional (X rays plus electrons or electron-only) or mobile, is used exclusively for IORT. It is recognized here that no more electron-only linear accelerators of the conventional, non-mobile type, will be built in the future. Dedicated linear accelerators may be located in the OR itself or in a room in the radiotherapy department. If the former, the room is basically an OR that also contains a linear accelerator. If the latter, it is a radiotherapy room that is equipped as an operating room. The former is preferred since it is part of the total OR complex and the patient does not have to be moved outside the room. In the

Figure 4. Mevatron ME by Siemens Medical Systems located in an OR. Note that the machine is a conventional accelerator whose gantry (C-arm) support system is mounted in the end wall. Since the room is slightly oversize for an OR room at this hospital, this still leaves adequate room for surgery.

latter case, the patient undergoes surgery far from the hospital's main OR and issues of maintaining sterility of the operating area have to be addressed as well as the problem of what to do if a surgical emergency arises for which the satellite OR is not equipped to cope. There are currently very few dedicated units in the OR in the United States; there are slightly more located in the therapy department. For the non-dedicated linear accelerator, the need to transport the patient for each case from the OR to the radiotherapy area is a disincentive to the IORT program. For this reason, programs using nondedicated linear accelerators have seen a decline in numbers (see Fig. 4). It is highly unlikely that any more dedicated units of the conventional or electron-only type will be installed in the United States in the future for the reasons outlined above. An example of a dedicated electron-only linear accelerator located in the OR is shown in Fig. 4. The room measures $\sim 6.1 \times 8.0$ m and the unit is mounted into the wall at one end, ensuring an adequate space for surgery.

Experience has shown that centers with dedicated OR suites or mobile systems perform more IORT procedures than centers that use the patient transport technique. Details of the shielding aspects of an OR-based IORT machine have been published (13,35).

Mobile IORT Units

The cost of installing a dedicated linear accelerator in an operating room along with the required shielding is very high. A linear accelerator, even though it is run only in the electron mode, has a price tag of ~$1.5–2 M and the shielding costs for a room that is not located in a basement, which is often the case, can be a substantial fraction of the linear accelerator cost. Given that the number of patients to be treated with an intraoperative machine is unlikely to be >5 per week, it is clear that the economics for such a unit are not altogether favorable (10). Since 1996, several mobile linear accelerators have become commercially available. These are the Mobetron, produced by IntraOp Medical (Santa Clara, CA), the Novac7 produced by Hitesys (Aprilia, Italy) and the Liac produced by Info&Tech (Udine, Italy). These are linear accelerators that have special features, in addition to being mobile, that reduce the need for extensive shielding in the adjacent walls, ceiling, and floor. This means that they can be used for treating patients in any OR room with little more than a portable lead shield to protect personnel in the surrounding areas, as well as above and below. (However, note that for commissioning and annual quality assurance purposes, a well-shielded room is required because of the high beam-on time needed for these tests.) This reduction in radiation leakage has been achieved in several ways. The first is to limit the maximum energy of the electrons that can be accelerated so that photoneutrons are not generated; the Mobetron has a maximum energy of 12 MeV, the Novac7 has a maximum energy of 9 MeV and the Liac has a maximum energy of 10 MeV (Note that according to European law, energies >10 MeV cannot be used in an OR without special protective shielding in the walls and floor). The second is to use an in-line beam, that is, the direction of the electron beam in the waveguide is the same as the direction of the beam that treats the patient. This avoids the use of a heavy magnet. Thus avoids the use of a heavy magnet. Conventional medium-to-high energy linear accelerators generally use a 270° bending magnet to bend the electron beam from the horizontal direction in the waveguide toward the patient at isocenter (34). The reason is that with the need to have a fully isocentric C-arm machine with an acceptable isocenter height, (Fig. 4), typically ~125 cm above the floor, the waveguide cannot be pointing toward the isocenter, given the length of the guide. This reduction of leakage also means that the shielding around the waveguide (usually lead or a tungsten alloy) is lower than for a conventional therapy linear accelerator, so the weight of the unit is reduced—essential for good mobility. Thus the major advantage of a mobile electron linear

accelerator is that minimal, in-room shielding is required for the unit so that it can be moved to any OR room, provided mechanical access is possible. A comparison of some of the other properties of these three mobile linear accelerators is in order since their method of beam generation and delivery are different. Major differences are that the Mobetron is a gantry-mounted unit that uses soft docking for alignments, whereas the two Italian models have the waveguide mounted on a robotic arm. Furthermore, the Mobetron waveguide operates in the X-band mode (8–12 GHz) whereas the Italian models operate in the conventional S-band mode (3 GHz). Note that the length of the accelerator's waveguide depends inversely on the frequency, so, for the same final electron energy, the X-band waveguide will be shorter.

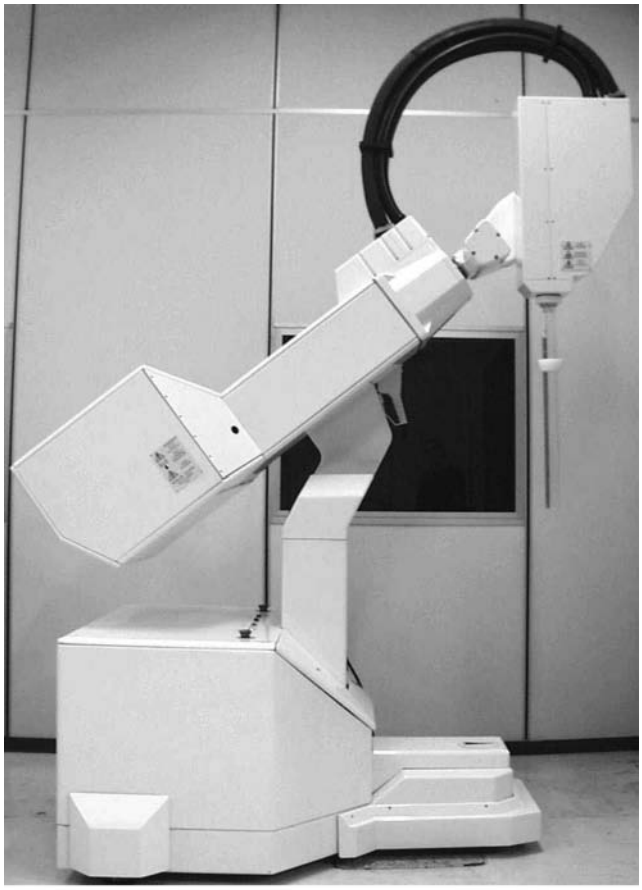
Mobetron. The first Mobetron was installed and treated its first patient in 1997. There are currently 12 Mobetron units installed worldwide, 7 operating in the United States, 1 in Japan, and 4 in Europe. This unit operates with electron energies of 4, 6, 9, and 12 MeV. A view of this unit is shown in Fig. 5. The unit is mounted on a gantry, cantilevered with a beam stopper to intercept the X-ray component of the primary beam after it has passed through the patient. This ensures that the radiation exposure in the room below is sufficiently low as to be within the limits set for the general public and is the principal feature that allows this unit to be used in any OR without additional shielding. Although the unit is gantry mounted, the beam direction is not limited to the plane of gantry rotation, as with external beam linear accelerators, since the head can tilt in and out of this plane. This increases its versatility in setting up patients for treatment and reduces the amount of time needed to align the radiation field with the applicator. The beam stopper also moves in synchrony with this gantry motion to ensure that the primary beam is always intercepted.

A review of the salient mechanical properties and operating parameters of the machine is in order. The weights of the treatment module and modulator are 1250 and 432 kg, respectively. The Mobetron is moved around through the use of a modified pallet jack. Thus the two units can be moved with reasonable ease between different ORs. The gantry can rotate off vertical by 45° in each direction, while the head tilt has a range of ±30°. Since the distance between the machine target and the patient's skin is 50 cm, half that of a conventional linear accelerator, the maximum dose rate is 10 $Gy \cdot min^{-1}$. Thus the maximum treatment time is ~2 min. For setting up the machine to treat the patient, the unit has five degrees of freedom. There are two orthogonal translational motions whereby the stand can be moved relative to the base by up to ±5 cm. There is also a translation motion of the head along the axis of the waveguide and, finally, there are 2 rotational degrees of freedom of the head, one the gantry rotation and the other tilt in and out of the gantry plane.

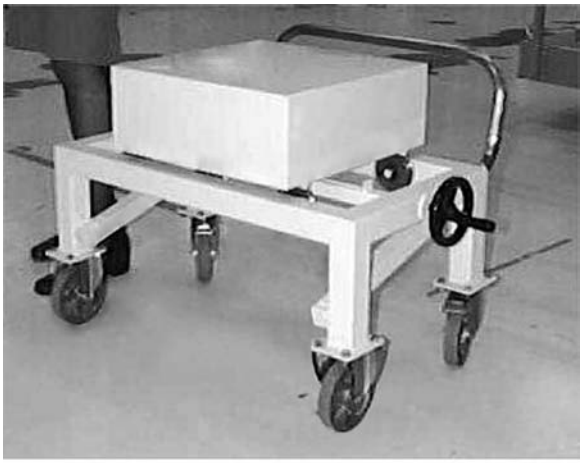
Meurk et al. (36) reviewed the physical properties of this device and Mills et al. (37) and Daves and Mills (38) provide a comprehensive review of the commissioning and shielding requirements of a Mobetron accelerator.

Figure 5. Mobetron mobile linear accelerator by IntraOp Medical Inc. Note that this unit has a gantry that also allows the linac head to pivot in a direction orthogonal to the gantry rotation plane.

Novac7. The first Novac7 unit was installed in Rome in 1997. To date, there are 20 Novac7 systems operating in Italy, Germany, and Greece. The Novac7 operates with electron energies of 3, 5, 7, and 9 MeV. Unlike most conventional linear accelerators and the Mobetron, the Novac7 is mounted on the end of a cantilevered robotic arm, not a C-arm, has a total weight of 500 kg. Figure 6 shows the Novac7 unit as well as the beamstopper to attenuate the forward directed X rays. The beamstopper operates independently of the machine and must be put in place by the medical physicist responsible for the IORT procedure. In many situations, however, movable wall barriers must also be used. The applicators have diameters of 4, 5, 6, 7, 8, and 10 cm with available bevel angles of 0, 15, 22.5, 30, and 45°. Note, however, that while the unit can be adjusted mm by mm, the cylinder attached to the linac cannot be moved coaxially with the applicator; instead, movements are made through a combination of rotational movements. By using a robotic arm, the unit has more degrees of freedom available for aligning the electron beam with the cone set up in the patient than a conventional accelerator or the Mobetron. In addition to its movement across the floor, it possesses four rotational degrees of freedom, which provides flexibility in setting up the device for treatment. Interestingly, the machine does not use scattering foils to produce a broad, uniform beam, but, instead, relies on electrons scattering in the air within the tubes and from the walls of the tubes to produce the desired, flattened fields. As a result, the length of the applicator used depends on the applicator size selected.



(a)



(b)

Figure 6. (a) Novac7 mobile linear accelerator by Hitesys; (b) beam stopper for Novac7. Note that this unit is a robotic arm with several degrees of freedom. It also uses the hard dock procedure.

Partly because of this, the dose rate varies between 6 and 26 Gy·min^{-1}, significantly higher than most linear accelerators. At these dose rates, treatments typically last $<$1 min. The lower dose rate corresponds to the largest applicator and lowest energy while the higher dose rate corresponds to the smallest applicator and highest energy.

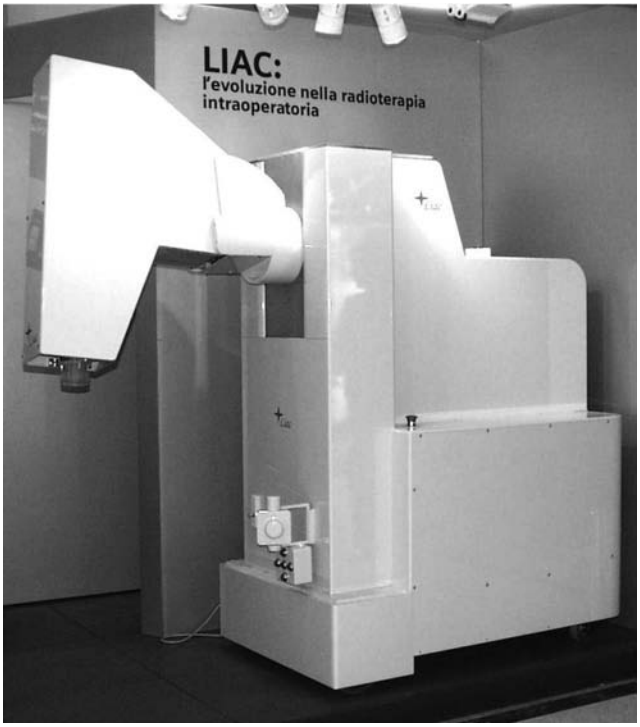


Figure 7. Liac mobile linear accelerator by Info&Tech. This unit looks similar to the Novac7 in terms of its movements and clinical set up, but there are differences in the design (see text for details).

These high dose rates result in substantial dosimetric problems, which have been studied and quantified by Piermattei et al. (39). A full technical description of the Novac7 has been given (40) and a review of the physical and dosimetric properties of the machine has been described by Tosi and Ciocca (41).

Liac. The first Liac unit (prototype) was installed in Milan in 2003. In its mechanical operation it is very similar to the Novac7 unit (see Fig. 7), with the accelerator guide mounted on a robotic arm. However, there are several major differences between the two units. First of all, the highest electron energy is now either 10 MeV (four energies of 4, 6, 8, and 10 MeV) or, potentially, 12 MeV (four energies of 6, 8, 10, and 12 MeV). This allows a greater penetration of tissue by the radiation, so that more deeply seated tumors can be treated. Second, the weight and dimensions of the machine are both lower, so the unit can be moved between rooms in the OR more readily, pass through doors more easily and fit into an elevator. The applicator system is the same as that of the Novac7, namely, it uses a hard docking procedure with no scattering foils. However, the applicator that defines the extent of the radiation field in the patient is now 30 cm in length. This length is standard for most commercial and noncommercial systems and allows the radiation oncologist and surgeon to view the radiation field directly, just before docking the applicator with the machine.

The Liac has the same high instantaneous dose rates as the Novac7, in this case 1.5–13 Gy·min^{-1} for the 10 MeV version and 3–22 Gy·min^{-1} for the 12 MeV version.

Table 4. Variation in the Depth of the 90% Dose with Electron Beam Energy

Electron Beam Energy, MeV	Depth of 90% Dose, cm[a]
6	1.7
9	2.6
12	3.7
15	4.5
18	5.1

[a] For a 7 cm diameter circular applicator that is commonly used in IORT; note that in IORT, the dose is normally prescribed to 90% level, taking into account that the surface dose is at or about that level.

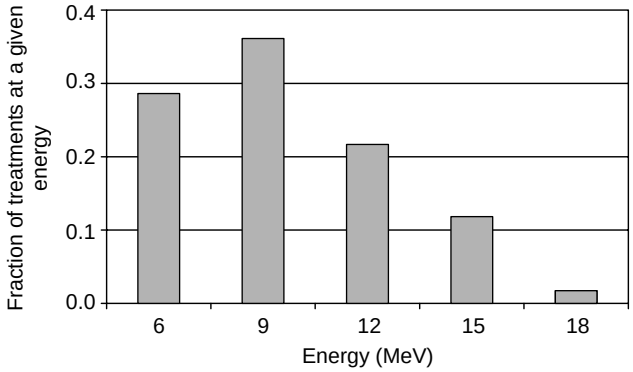


Figure 8. Fraction of treatments delivered at a given energy as a function of the energies available. From this graph, one can see that the majority of treatments are at 9 MeV or less and that <15% or patients are treated with an energy >12 MeV.

Comparison between Conventional and Mobile Units

It is worthwhile comparing the pro's and con's of conventional linear accelerators in the OR versus mobile linear accelerators. The conventional machine has the advantage that since it is fully shielded for the highest energy available, ~18 MeV or higher, tumors with a greater depth can be treated. As a general rule, the depth of treatment increases by ~1 cm for every 3 MeV increase in energy at the 80% dose. Table 4 shows the depth of the 90% dose for one conventional unit.

However, the question arises as to how many treatments are delivered at high energy versus low energy. Figure 8 shows data from the MGH on the use of the various electron energies. It can be seen that all but 15% of the treatments were at 12 MeV or less, indicating that high energies are infrequently used. In contrast, the disadvantages are the cost of the machine, generally greater than for a mobile unit, and the shielding for neutrons as well as photons. At MGH, this shielding amounts to ~50 metric tons, which makes retrofitting existing ORs all but impossible. Moreover, should the machine malfunction, the OR becomes unavailable for the duration of the repair. Another fundamental difference is that conventional machines use a 270° bending magnet with momentum slits to focus the electrons onto the scattering foil, whereas the mobile units have no energy or momentum selection. While both units change energy by adjusting radio frequency (rf) power, the conventional units have the advantage of better energy selection. This leads to a higher surface dose and slower fall-off in the depth dose curves for mobile units (48).

Mobile linear accelerators, on the other hand, have distinct advantages over the conventional unit. First of all, the units are mobile so that they can be moved to any OR that has sufficient space. This is possible because the units require almost no additional shielding to operate safely. However, a beam stopper is required for the primary beam in all mobile units. Whereas a dedicated unit is limited to isocentric movement, all mobile units have more degrees of freedom, allowing for greater flexibility in setting up the patient.

A summary of the chief differences between the three different modes of IORT is given in Table 5.

Treatment Applicators

There have been two types of applicators used in electron beam IORT, those made from poly methyl methacrylate (PMMA) and those made of metal, usually chrome-plated brass. The supposed advantage of the PMMA applicators was that one could see the tissues to be irradiated through the walls, although this turned out to be largely not the case. The disadvantage of using PMMA applicators is that they have to be sufficiently thick so as to prevent radiation penetrating through the walls and damaging normal tissue; this minimum thickness is ~6 mm. Metal applicators, on the other hand, do not have to be so thick because of their greater density. Since the pelvis is a tight area, anatomically, having a thick-walled applicator can restrict the treatment area for a pelvic side wall lesion. Thus, metal applicators are preferred, also because they can be flash (steam) sterilized at the time of the procedure, whereas applicators made from PMMA must be gas sterilized to avoid heat-related deformities and this requires ~12 h. Figure 9 shows examples of plastic applicators. Examples of metal cones are shown in Fig. 10.

Table 5. Comparison among the Three Methods of IORT Using Linear Accelerators

Mode of IORT	Advantages	Disadvantages
Linac in oncology department (patient transport)	Inexpensive; needs minimal additional equipment	For a busy outpatient machine, can do only 1–2 cases per week. Enthusiasm wanes due to effort required
Dedicated conventional linac in OR	Available on full-time basis. Maximizes convenience for surgical staff	Requires expensive shielded suite with low use factor
Mobile accelerator	Can be used in almost any OR; minimal additional shielding required	Limit on the maximum energy due to leakage X rays and neutrons

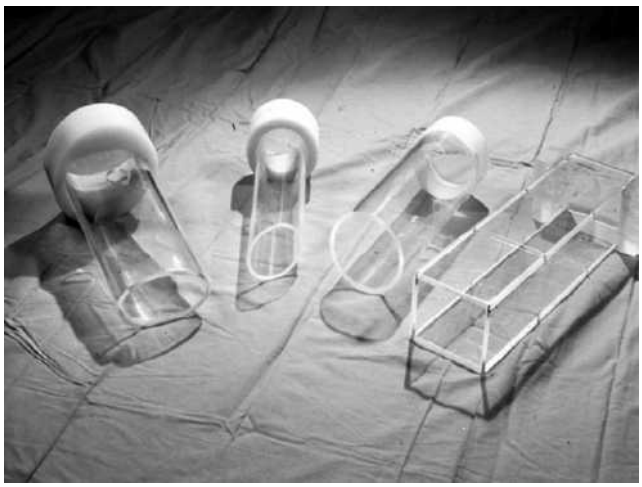


Figure 9. Examples of plastic applicators used in electron IORT. From left to right are shown an elliptical, a circular with beveled end, a circular, and a rectangular applicator.

Most manufacturers of dedicated linear accelerators provide applicators with a range of dimensions, usually circular in cross-section. A description of the design of treatment applicators for an OR-based IORT unit has been given by Hogstrom et al. (42)

Beam Alignment Devices

For nondocking machines, a system is required to align the electron beam from the waveguide with the axis of the treatment applicator in the patient (this is unnecessary for hard-docking systems since alignment is guaranteed with the tolerance between the two mating pieces). This is usually achieved optically. In the case of one dedicated, conventional linac (Siemens Mevatron ME, Concord, CA; this unit is no longer manufactured), beams from two lasers are split into four point sources and four line sources. When the points converge with the lines on a particular radius, then the distance is correct and the beam axis is correctly angled with respect to the treatment applicator. A metal disk with this radius drawn on it is placed on top of the treatment applicator for this purpose. A view of how

Figure 10. Examples of metal cones used in electron IORT. From left to right are shown a circular with bevel, a circular, an elliptical, and a rectangular applicator.

Figure 11. Example of the optical alignment used on the Mevatron ME linear accelerator. This test jig is used before each procedure to check the alignment of the lasers. Alignment is correct when the dots and lines (not easily seen here, but overlapping the radial lines) cross on the circle.

the correct alignment should look is shown in Fig. 11 for the test jig. Clearly shown is the circle with four laser dots at 90° intervals; the laser lines are in alignment with the radial lines shown, but on a black/white image, they are not visible. This system requires training to fully understand what movements should be made with the couch and gantry to align the treatment applicator with the beam when the dots and lines are not aligned. A more sophisticated system, that is simpler to use, is adopted with one of the mobile linear accelerators (Mobetron by IntraOp Medical, Inc.). This system also uses optical alignment, but provides visual indicators as to how to adjust the gantry to complete the alignment. As with the other system, laser light is reflected off a mirror on top of the applicator and is sensed by detectors in the head. This display is shown in Fig. 12, which shows the status of the alignment that is complete

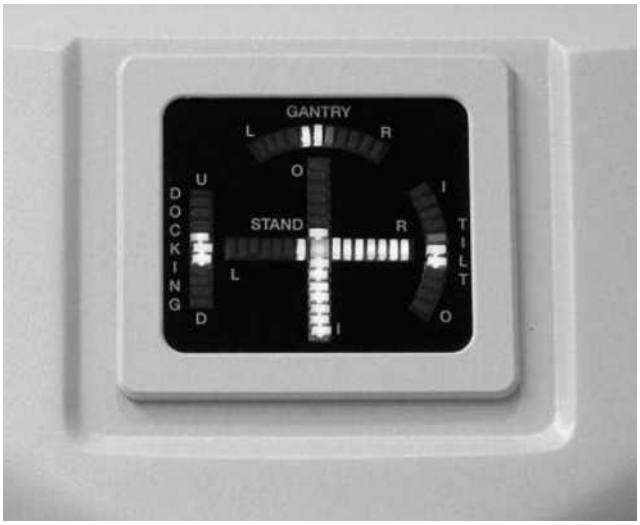


Figure 12. Display of the alignment system used with the Mobetron mobile unit. (Photo courtesy of G. Ezzel, Mayo Clinic.)

when each of four parameters is indicated by a single yellow bar either side of a green bar (colors not indicated). Parameters on the display are the gantry rotation, the docking or distance between the head and the applicator, the tilt of the head and the position of the stand (in-out and left–right). The display shown indicates that the stand has to be moved from in to out and from right to left to complete the docking. When the alignment is complete, the alignment is accurate to 2 mm along any one axis and 0.1° in angle.

IORT Treatment Logistics

Since the radiation is delivered through linear accelerators, there are several options for performing this technique. The first is that the patient is transported from the operating room (OR) to the radiation therapy department and set up and treated on one of the linear accelerators usually used for treating cancer patients with external beams. The second is to have a dedicated treatment room in the OR. In this case, the patient has to be moved only from one side of the room to the other where the linear accelerator is located. The third option is a mobile linear accelerator. A mobile linear accelerator is one that can be moved from one OR to another. A fourth option is a hybrid of the first and second methods in that a linac in the radiation therapy suite serves also as an operating room so that both the surgery and IORT are carried out in the same room. The problem with this option is that if this room is far from the other operating rooms and if difficulties are encountered during the surgery, assistance might be problematic, particularly if specialized equipment is required.

There is a significant difference between the first option and the other three options since the patient has to be transported from the OR to the radiation therapy area. Once there, the machine used for these treatments is sequestered for the duration of the set-up and treatment. Additional time is required before and after the IORT procedure to prepare and clean up the room. Thus there are two negative effects, one, the surgical procedure is lengthened because of the transport and reset-up of the patient, and two, the therapy department loses significant time on one of its linear accelerators. Clearly, a department having only one or two machines with a busy outpatient workload could not readily perform these procedures, or, at least, not more than once a week. Finally, while the patient is being transported or in the therapy area, the OR has to be held open for surgical closure or additional surgery and this has an impact on a busy surgical department and its resources.

IORT Dosimetry, Calibration, QA, and Radiation Safety Electron Beam Calibration, Dosimetry, and Quality Assurance. Important components of IORT dose delivery are the calibration of the of the electron beam, dosimetry and routine quality assurance. These three important areas of IORT will be addressed separately. Electron beams are calibrated using a recommended methodology (43) and according to national protocols. In the United States and Canada, the appropriate protocol has been established by the American Association of Physicists in Medicine (AAPM) (44). This protocol dictates that measurements be made using an ionization chamber in a water phantom under specific conditions. The ionization chamber is typically a cylindrical chamber having an air volume of 0.6 cm^3, although the protocol also recommends the use of plane parallel chambers for low energies. The specific conditions relate to the depth of measurement as a function of electron beam quality and chamber dimensions. The protocol provides factors to calculate the absolute dose from the measured chamber exposure. The chamber is calibrated at an AAPM certified laboratory whose calibration standards are, in turn, referenced to the National Institute for Standards and Technology (NIST). Thus users' chambers around the country have calibration factors that are traceable to NIST so that 1 cGy in one institute is equal to 1 cGy in any other institute. The 2σ uncertainty in the users' calibration is stated to be 1%.

Having calibrated the IORT electron beam under standard conditions, it is necessary to determine the dose delivered to the patient under the most general conditions. For this purpose, medical physicists perform a series of measurements on the machine, including the variation of dose with depth and applicator size. Using these measurements, medical physicists can recommend the optimal energy for treating a specific lesion given the depth of the lesion and will then set the parameters on the accelerator to deliver the dose prescribed by the radiation oncologist.

Quality assurance is a very important aspect of radiation therapy, both external and IORT. In general, quality assurance is mandated by state and federal agencies and recommendations are made by the AAPM (45). These laws and recommendations specify the types and frequency of tests that must be carried out on linear accelerators used in radiation therapy. For IORT, quality assurance is of particular importance since, unlike external radiation therapy where up to 42 fractional doses may be given over an 8 week period, the dose is given in a single fraction. Therefore, quality assurance tests must be particularly probing to ensure that the probability of any error is as low as possible (46). The AAPM has made specific recommendations for IORT (47,48) at the daily, monthly and annual level. Table 6 shows the data taken from the latest AAPM report (48).

In addition, additional, independent quality assurance on electron beam output and percent depth dose is provided by the Radiological Physics Center in Houston, using mailable TLDs (49).

Radiation Safety Issues

All radiation-producing equipment is subject to state and federal regulations. These regulations restrict the dose that a member of the general public can receive to 1 mSv $year^{-1}$ or 0.02 mSv·$week^{-1}$. There is also a further restriction that such a person may receive no more than 0.02 mSv in any 1 h. For controlled areas (not accessible to the general public), the allowed limits are higher. Since many of the linear accelerators are located/used in the operating room environment, regulations for the general public apply.

As noted above, radiation safety requirements for electron beam IORT depend on the approach taken. For treatment

Table 6. Quality Assurance Tests for Mobile Linear Accelerators

Frequency	Parameter
Daily	Output constancy
	Energy constancy
	Door interlocks
	Mechanical motions
	Docking system
Monthly	Output constancy
	Energy constancy
	Flatness and symmetry constancy
	Docking system
	Emergency off buttons
Annual	Output calibration for reference conditions
	Percent depth dose for standard applicator
	Percent depth dose for selected applicators
	Flatness and symmetry for standard applicator
	Flatness and symmetry for selected applicators
	Applicator output factors
	Monitor chamber linearity
	Output, percent depth dose and profile constancy over the range of machine orientations
	Inspection of all devices normally kept sterile

[a] AAPM task group 72.

with a linac in a radiotherapy department, the room is already well shielded for any number of IORT treatments. For a dedicated, conventional linear accelerator in the OR, the shielding is usually designed for a specific number of cases per week. A maximum number would be 10 cases per week, based on the length of an average surgical procedure. However, patient demographics and past IORT experience indicate that this limit is never reached on a continuing basis. However, the machine has to undergo acceptance testing and commissioning when first installed and extensive checks every year by the medical physicists that require extensive beam-on time (there are also daily and monthly checks, but these require much less beam-on time). Therefore, this work has to be carried out at night and over the weekend when personnel are generally not present. For corridors in the OR that are still in use, this may require installing temporary barriers with appropriate radiation warning signs; rooms above and below have to be checked to ensure that they are not occupied. For mobile linear accelerators, it was noted above that, due to their low leakage, considerably less shielding is required for their use in the OR, apart from a primary beam stopper. Lead shields placed strategically around the patient provide secondary shielding. However, before any OR can be used for treatment with a mobile linear accelerator, extensive surveys have to be carried out to determine radiation levels in the immediate vicinity with and without the secondary shielding. Based on these readings, the number of cases that can safely be performed on a weekly basis without exceeding the maximum permissible dose can readily be calculated. Clearly, this exercise has to be performed for every OR in which IORT cases may be performed. As noted above for the dedicated, conventional linear accelerator, there are occasions when the medical physicist has to perform extensive testing of the equipment. This can also only be done at night or on weekends, provided the occupancy of nearby areas can be controlled. If this cannot be done, the unit has to be taken to a shielded area for testing, such as the radiotherapy department. Issues relating to the radiation safety aspects of mobile linear accelerators are fully covered in the AAPMs report on mobile linear accelerators (48). In all cases, consultation with the hospital's radiation safety officer is highly recommended.

OTHER IORT TECHNIQUES

INTRABEAM System for Intracranial Lesions

Another device that has come into use for the intraoperative treatment of intracranial and other lesions is the INTRABEAM System. This device was originally manufactured by Photoelectron Corporation, but is now marketed by Care Zeiss Surgical GmbH, Germany. This system is a 50 kV device that fits into a standard neurosurgical stereotactic frame. Figure 13 shows the INTRABEAM device; electrons are accelerated down a 10 cm long 3.2 mm diameter evacuated tube, striking a gold target at the tip. The thicker section between this tube and the body of the device contains coils to steer the electron beam. Built into the body of the device is a scintillation detector that detects backward emitted photons in a fixed geometry. This

Figure 13. INTRABEAM X-ray tube. The body of the device contains the high voltage electronics, power for the steering coils and the electron gun. It also contains the internal radiation monitor for integrating the radiation output of the device, similar to the internal ionization chambers in a linear accelerator. Electrons are accelerated to the tip of the probe where they strike a gold target and produce the X rays. The large diameter section contains the steering coils.

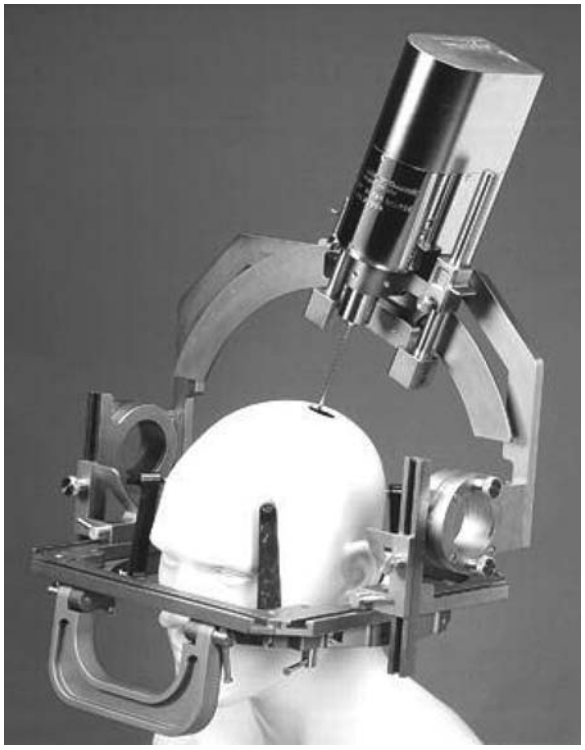


Figure 14. The INTRABEAM device inserted into a stereotactic head frame demonstrating how the device is used for the treatment of intracranial lesions.

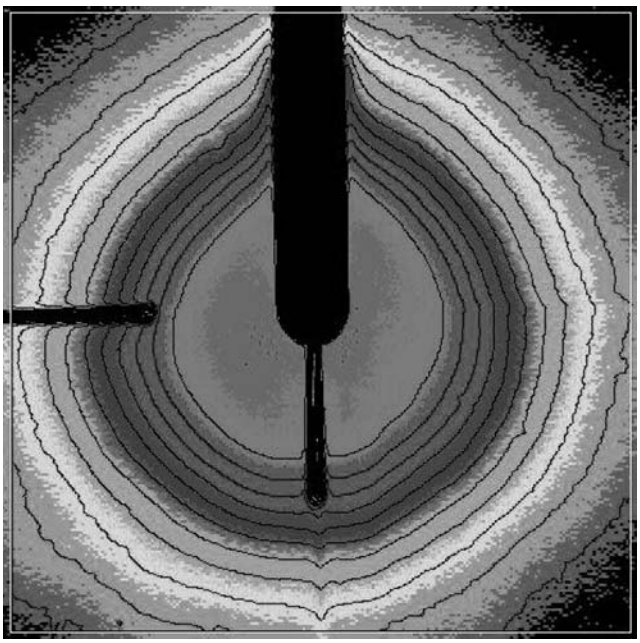


Figure 15. Isodose curves for the INTRABEAM X-ray device. It can be seen that the tip of the probe is the center of the X-ray source and that the source is nearly isotropic in intensity except in the backward direction close to the tube.

is known as the internal radiation monitor. Figure 14 shows the X-ray device inserted into a CRW neurosurgical frame so that the probe tip can be placed at the location of the stereotactic site to within an accuracy of better than 1 mm. Thus an intracranial lesion is first imaged using Computerized tomography (CT) and the coordinates of the center of the lesion determined so that, in addition to the neurosurgeon performing a stereotactic biopsy, the tip of the X-ray source, where the X rays are generated, can be accurately located at this point and a tumoricidal dose of radiation delivered. Figure 15 shows a series of isodose curves generated using radiochromic film (Gafchromic film, type MD-55, ISP Technologies Inc., Wayne NY). The outline of the probe tip in the north–south direction can easily be identified as the source of the radiation. The horizontal line and the line extending from the probe tip represent cuts in the film. The dose depth curve for 50 kV X rays in water is shown in Fig. 16. The dose rate from this device is about 2 $Gy \cdot min^{-1}$ at 1 cm from the probe tip in water. The dose falls off as the inverse third power of the distance and, hence, the dose delivered outside the tumor drops very rapidly. At this dose rate and for this dose fall-off, typical treatment times last from 15 to 30 min, depending on the treatment radius. A fuller description of the apparatus and the clinical testing of the device has been given by Beatty et al. (50). A Monte Carlo has been performed (51) to predict both the spectrum and output of the X-ray beam with good accuracy.

Although initially developed for intracranial lesions, the INTRABEAM system is approved by the U.S. Food and

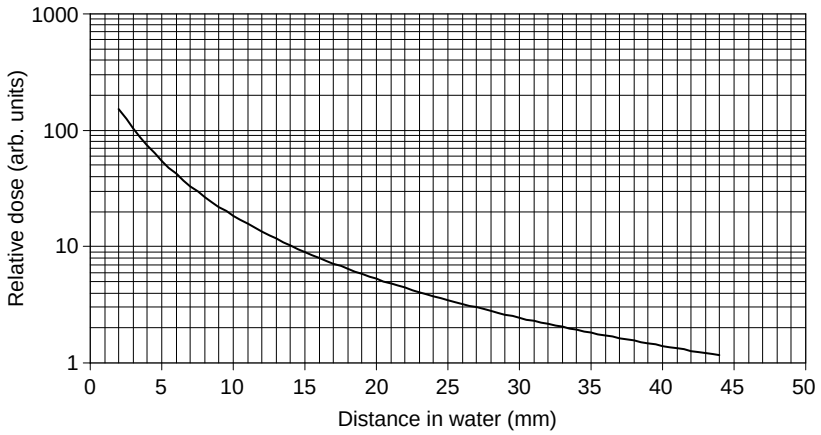


Figure 16. Distance–dose curve for the INTRABEAM device for 50 kVp X rays. The output falls off as the inverse of the third power of the distance from the probe tip. In addition to inverse square, there is an additional factor due to the attenuation of the X-ray beam, which is significant at these energies.

Drug Administration (FDA) for application to any location in the body. To that end, a set of spherical applicators was developed that could be used to treat meningiomas or breast cancer, after surgical resection. For breast and other extracranial treatments, the stereotactic frame is not used. Instead, a stand developed by Zeiss is used to support the INTRABEAM device. It can be set and locked into any position so that it remains stationary during the procedure. In Europe, the treatment of breast cancer is one of the principle uses of this device. There are currently ~30 units operating in the United States and Europe. Of note is the fact that since this device delivers X rays at very low energies, the shielding requirements are minimal. All personnel, except for the operator and perhaps the anesthesiologist, leave the room. A portable diagnostic X-ray shield is used to protect the operator and anesthesiologist. All entry doors except one, beyond which all the other OR personnel wait during the treatment, are locked.

High Dose Rate (HDR) Brachytherapy

Brachytherapy is a technique whereby sealed radioactive sources are placed inside the body or in body cavities, usually for a short duration of time or, in some cases, permanently. Most of the sources for temporary implants have activities such that the dose rate to an isodose line surrounding the tumor is ~50 $cGy \cdot h^{-1}$. With the typical doses given in conventional, low dose brachytherapy, this means that the treatment time is often several days and the patient has to be hospitalized. However, by using a HDR source (typically 10 Ci of ^{192}Ir), the treatment time can be shortened to minutes. Figure 17 shows an HDR (Nucletron Microselectron HDR) with one catheter inserted. The other end of the catheter is place in a well counter, a chamber used by medical physicists to measure the strength of the radioactive source. A total of 18 catheters can be attached to the unit and a computer control system (not shown) guides the single source down each catheter to a given position for a predetermined dwell time. By varying the dwell time of the source at each location, the dose distribution can be optimized for each treatment. In this procedure, as many catheters as are required to cover the tumor or tumor bed are inserted into the patient or placed on the outside for skin treatments; a single source is used and this is mechanically driven along each catheter, programmed to dwell at preselected positions for lengths of time determined by the treatment planning program. The HDR treatments took some time to be accepted by the medical community because it was initially thought that dose rate effects in brachytherapy were crucial and that HDR treatments would be less effective than those using conventional, low dose rate techniques. Because of this shortened treatment time, it became possible to perform exploratory surgery in the OR, lay out the catheters in the area of residual disease, treat the patient while still on the operating room table and then close up the patient. Figure 18 shows a Harrison–Anderson–Mick (H.A.M.) applicator for HDR–IORT. This applicator consists of a number of catheters embedded in a 1 cm thick silicone rubber mat. The catheters are located mid-plane and therefore 0.5 cm from each surface. Applicators are available with the edges curved in a half circle (not shown here) to help maintain a more uniform dose at the surface. Applicators are available in sizes from 3–24 channels in groups of 3, namely, 3, 6, 9, and so on. Because this is a planar application of ^{192}Ir sources, the depth of treatment is almost always taken to be 1 cm, that is, 0.5 cm from the surface of the applicator, to avoid an excessively high dose at the skin–applicator interface. This restricts the number of applications for which this approach can be used, typically lesions that have been resected with positive margins, compared with electron beams, which, as shown above, can penetrate much greater depths. Care must be exercised in the planning dosimetry for these applications since the dose depends on the curvature of the applicator (52) and the lack of backscatter material (53). Figure 19 shows an applicator in place in a patient prior to treatment. Note the curvature of the applicator.

Shielding, in the form of concrete or lead is required in the walls and floor it a significant number of patients are

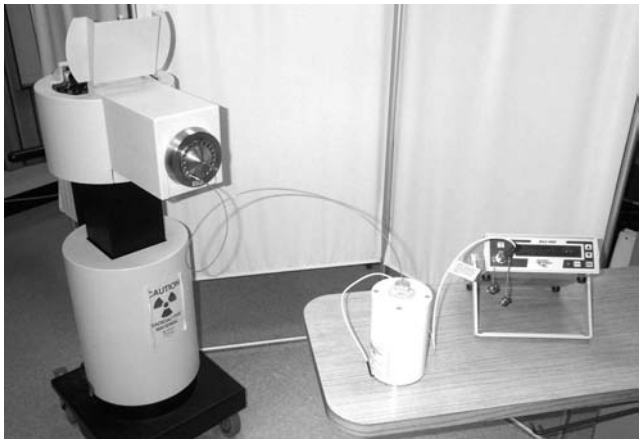


Figure 17. View of the HDR system with one catheter hooked up to a well chamber for calibration of the source by a medical physicist.

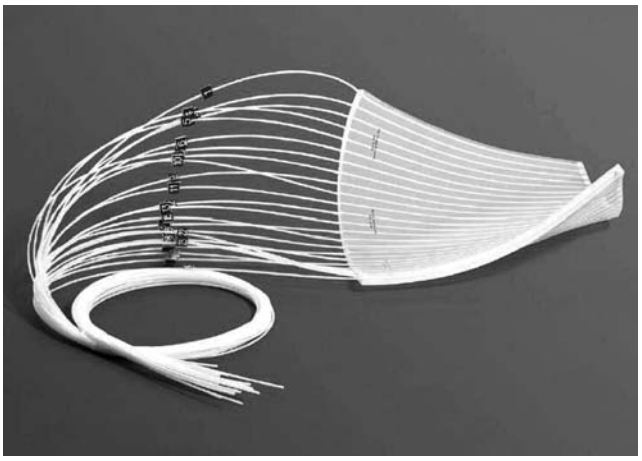


Figure 18. Sample H.A.M. applicator. The catheters are embedded in a 1 cm thick silicone rubber mat. Note the clear identification numbers on each catheter. The manufacturer supplies applicators with varying numbers of catheters. (Photo courtesy of Felix Mick.)

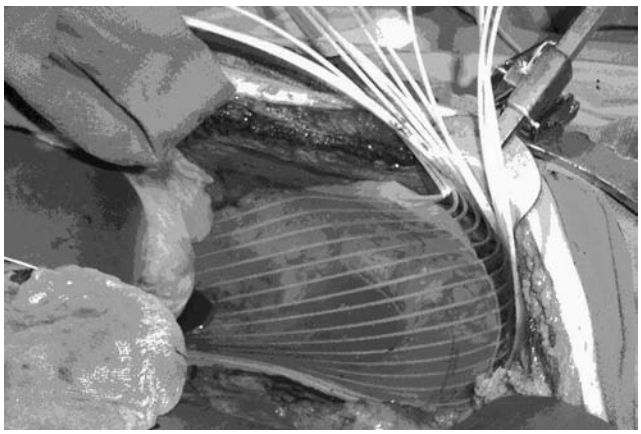


Figure 19. View of a H.A.M. applicator inserted in a patient and ready for treatment. (Photo courtesy of Dr. C. Willett.)

Table 7. Clinical Results for IORT Treatment of Primary and Recurrent Rectal Cancer Using Electrons

Primary Colorectal Cancer					
			Survival, years		
	No. Pts	Median (month)	2	3	5
EBRT	17	18	35%	24%	24%
EBRT+IORT	56	40	70%	55%	46%
Recurrent Colorectal Cancer					
			Survival, years		
	No. Pts	Median (month)	2	3	5
No IOERT	64	17	26%	18%	7%
IORT+/-EBRT	42	30	62%	43%	19%

treated. The exact thicknesses will depend on the number of cases treated per week.

At present, there are four institutions in the United States performing HDR in the OR. Excellent technical descriptions of the HDR technique applied in the operative setting have been given by Harrison et al. (54) and Nag et al. (21).

CLINICAL RESULTS

Disease sites that have been treated with IORT include pelvic lesions such as pancreas, rectum, colon, retroperitoneal sarcoma, bladder, kidney, sacrum, liver, endometrium, cervix, and ovary. Extra-pelvic sites include head and neck, breast, and distal limbs. At the MGH, the most commonly treated disease sites include the pancreas, rectum, colon, and retroperitoneal sarcoma. UCSF reported in 2003 that a majority of treatments had been in the area of head and neck and thorax. A considerable number of pediatric patients had also been treated in a few centers. Thus the type of sites treated in different institutions reflects the interests of the surgical and oncology departments of those institutions.

One of the problems encountered when comparing the efficacy of IORT combined with external beam radiotherapy (EBRT) with EBRT alone in terms of overall survival is that distant metastases play a major role in many disease sites. While it is true that IORT combined with EBRT provides a result as good as any EBRT alone series, the difference in survival is often not significant. Moreover, due to the rapid progression of distant metastases, it is often not possible to know the exact pathologic stage of each patient under treatment, and so comparisons among different series is difficult. Data from MGH (unpublished) shows that overall survival for patients with pancreatic cancer depends on the field size of the treated field, the smaller the field, the greater the survival. What this says is that patients with bulkier disease will experience more rapid development of distant metastases.

One of the success stories of IORT, however, is colorectal cancer since the effect is statistically significant. Table 7 shows the results from MGH for primary and recurrent colorectal cancer. This was not a randomized trial, but a comparison between EBRT with electron beam IORT and historical controls, EBRT alone. In each case, the survival is greater when IORT is used, even up to 5 years. While this does not have the power of a randomized trial, the difference in the results in each case is both statistically significant and impressive. Moreover, similar results have been observed at other institutions, for example, the Mayo Clinic (55). Similar results have been obtained with HDR–IORT.

Current results in the treatment of gastric cancer indicate that while some benefits may accrue from IORT, further studies are needed to demonstrate a significant benefit. Interestingly, gastric cancer was one of the sites first chosen by the Japanese for IORT, which is quite understandable, given the high rate of incidence of that disease in Japan. For locally advanced primary and recurrent gynecological malignancies IORT has shown that results are comparable to historical results with standard salvage therapy. In one of the rare randomized trials involving IORT, a study of retroperitoneal sarcoma accumulated 35 patients between two arms, EBRT and EBRT+IORT. With a minimum follow-up of 5 years and a median follow-up of 8 years, a significant difference in local control was found between the two groups (56). However, as with studies for other diseases, there was no difference in median survival between the two groups. Two major centers, the Mayo Clinic and the Denver Children's Hospital, have been involved with pediatric malignancies. They have shown that IORT is effective in local control for locally advanced pediatric malignancies. Interestingly, the Denver Children's Hospital used only IORT whereas the Mayo Clinic used combined IORT and EBRT. In bladder cancer, IORT has been shown to be effective in bladder preservation. Investigators found few complications related to these treatments.

Perhaps the greatest thrust at the moment in the field of IORT is the treatment of breast cancer. This is actively being pursued in Europe, particularly Italy, followed by Germany and the United Kingdom (UK), and, to a lesser extent, in the United States. The drive for this is related to the need to shorten the overall duration of the conventional

external therapy treatment time. In this area, IORT is competing with several other techniques, including partial breast irradiation and temporary implants using ^{192}Ir brachytherapy seeds and the INTRABEAM system. The IORT and the INTRABEAM systems would, of course, require only a single fraction treatment. Although a large number of breast cancer patients have been treated intraoperatively to date, the follow-up time is too short to draw any conclusions.

At MGH, interstitial radiosurgery with the INTRABEAM has been performed for intracerebral metastases, primary malignant gliomas and as adjunct to resective surgery in meningiomas. Local control of metastases was possible in 80–85% of cases ($n = 73$), which is consistent with other forms of external radiosurgery. In a handful of cases with malignant gliomas, tumor recurred as expected beyond the areas of high dose irradiation. In 12 parasagittal meningiomas where residual tumor was left along the sagittal sinus, there has been no evidence of local recurrence with mean follow-up of 3 years, which is too early to determine efficacy.

FUTURE DIRECTIONS

If one looks at the history of IORT, one sees that the procedure has been practiced almost as long as there have been X rays. One of the driving forces for IORT in the early days was undoubtedly the need to overcome the limitation in dose delivery due to skin desquamation, a breakdown of the skin, resulting from the highest dose being at the surface. Certainly, the conditions of using X-ray equipment available at that time could not have been conducive to ease of use in an OR setting. It is clear, however, that the history of its practice has been punctuated by the introduction of new technology, first the change from low kilovoltage X rays (50 kV) through orthovoltage X rays to electrons. However, it was clearly the introduction of linear accelerators with electron beams that provided the greatest impetus over the last century, although the technique was not universally adopted because of cost, time, and many other factors. Indeed, there was a decline after the 1990s associated with these factors. It was not until the introduction of mobile linear accelerators that the field regained its strength. Part of the reason for this is the desire, for logistical reasons, particularly in Europe, to speed up the radiation therapy treatment process. Granted that this is also possible using external beam techniques, such as partial breast irradiation, but IORT certainly has a role to play in this change in therapeutic practice. The increase in the number of units attests to this: In 1997, the first unit was installed and at the time of this writing, 33 have been installed worldwide with units available from three manufacturers. This technology has enabled more centers to participate in IORT procedures, allowing both surgeons and radiation oncologists more convenient ways to treat patients. Moreover, several photon modalities have come into increasing use in IORT, including HDR–IORT and the Zeiss INTRABEAM device. These are playing a greater role in IORT, in part because the capital investment in these technologies is considerably less than a dedicated mobile linear accelerator.

However, before the widespread acceptance of IORT is possible, proponents need to demonstrate clearly that IORT provides a therapeutic advantage for their patients. While this has been shown for a subset of patients with rectal cancer, based on retrospective studies, it is clear that randomized trials are needed for other disease sites to provide unequivocal evidence for IORTs benefit. This is currently being done for breast cancer in Europe.

Thus, while the field is still small, it continues to grow, thanks to better technology. More widespread use of IORT will determine whether the procedure confers a benefit to patients across a broad class of diseases.

BIBLIOGRAPHY

1. Abe M, et al. Intra-operative radiation in abdominal and cerebral tumors. Acta Radiol 1971;10:408.
2. Abe M, et al. Intraoperative radiotherapy of gastric cancer. Cancer 1974;34:2034.
3. Abe M, et al. Techniques, indications and results of intraoperative radiotherapy of advanced cancers. Radiology 1975;116:693.
4. Abe M, Takahashi M. Intraoperative radiotherapy, The Japanese experience. Int J Radiat Oncol Biol Phys 1981;7: 863–868.
5. Goldson A. Preliminary clinical experience with intraoperative radiotherapy (IORT). Semin Oncol 1978;8:59–65.
6. Gunderson LL et al. Intraoperative irradiation. A pilot study combining external beam photons with "boost" dose intraoperative electrons. Cancer 1982;49:2259–2266.
7. Tepper J, Sindelar WF, Glatstein E. Phase I study of intraoperative radiation therapy combined with radical surgery for intra-abdominal malignancies. ASCO Proc 1980;21:395.
8. Gunderson LL et al. Intraoperative and external irradiation with or without resection: Mayo pilot experience. Mayo Clin Proc 1984;59:691.
9. Coia LR, Hanks GE. The need for subspecialization, intraoperative radiation therapy. Int J Radiat Oncol Biol Phys 1992;24:891–893.
10. Wolkow H. The economics of intraoperative radiation therapy. In: Dobelbower RR, Abe M, editors. Intraoperative Radiation Therapy. Boca Raton, (FL): CRC Press; 1989.
11. Papillon J. Conservative treatment by irradiation as an alternative to surgery. Rectal and Anal Cancers. New York: Springer Verlag; 1982.
12. Rich TA, Piontek RW, Kase KR. The role of orthovoltage equipment for intraoperative radiation therapy. In: Dobelbower RR, Abe M, editors. Intraoperative Radiation Therapy. Boca Raton, (FL): CRC Press; 1989.
13. Biggs PJ, et al. Dosimetry, field shaping and other considerations for intra-operative electron therapy. Int J Radiat Oncol Biol Phys 1981;7:875–884.
14. Palta JR, Suntharalingam N. A nondocking intraoperative electron beam applicator system. Int J Radiat Oncol Biol Phys 1989;17:411–417.
15. Kharrati H, Aletti P, Guillemin F. Design of a non-docking intraoperative electron beam applicator system. Rad Oncol 1999;33:80–83.
16. Jones D, Taylor E, Travaglini J, Vermeulen S. A noncontacting intraoperative electron cone apparatus. Int J Radiat Oncol Biol Phys 1989;16:1643–1647.
17. Björk P, Knöös T, Nilsson P, Larsson K. Design and dosimetry characteristics of a soft-docking system for intraoperative radiation therapy. Int J Radiat Oncol Biol Phys 2000;47: 527–533.

18. Sindelar WF, Johnstone PAS, Hoekstra HJ, Kinsella TJ. Normal tissue tolerance to intraoperative irradiation. In: Gunderson LL, Willett CG, Harrison LB, Calvo FA, editors. Intraoperative Irradiation: Techniques and Results. Totowa, (NJ): Humana Press; 1999.
19. Gillette EL, Gillette SM, Powers BE. Studies at Colorado State University of normal tissue tolerance of beagles to IOERT, EBRT or a combination. In: Gunderson LL, Willett CG, Harrison LB, Calvo FA, editors. Intraoperative Irradiation: Techniques and Results. Totowa, (NJ): Humana Press; 1999.
20. Hall EJ. Radiology for the radiologists. Hagerstown, (MD): Harper and Row; 1978.
21. Nag S et al. Intraoperative irradiation with electron-beam or high-dose-rate brachytherapy: Methodological comparisons. In: Gunderson LL, Willett CG, Harrison LB, Calvo FA, editors. Intraoperative Irradiation: Techniques and results. Totowa, (NJ): Humana Press; 1999.
22. Beck C. Über Kombinationsbehandlung bei bösartigen Neubildungen. Berl Klin Wochenstr 1907;44:1335.
23. Beck C. On external Roentgen treatment of internal structures (eventration treatment). N Y Med J 1909;89:621–622.
24. Finsterer H. Zur Therapie inoperabler Magen- und Darmkarzinome mit Freilegung und nachfolgender Röntgenbestrahlung. Strahlentherapie 1915;6:205–218.
25. Abe M. History of intraoperative radiation therapy. In: Dobelbower RR, Abe M, editors. Intraoperative Radiation Therapy. Boca Raton, (FL): CRC Press; 1989.
26. Barth G. Erfahrungen und Ergebnisse mit der Nahbestrahlung operativ freigelegten Tumoren. Strahlentherapie 1953;91: 481–527.
27. Goin LS, Hoffman EF. The use of intravesical low voltage contact Roentgen irradiation in cancer of the bladder. Radiology 1941;37:545.
28. Biggs PJ, Noyes CG, Willett CG. Clinical physics, applicator choice, technique and equipment for electron intraoperative radiation therapy. In: Petrelli N, Merrick III HW, Thomas Jr CR, editors. Surgical clinics of North America. Philadelphia, PA: W.B. Saunders; 2003.
29. Hensley FW, Ciocca M, Petrucci A, Biggs PJ. Survey of IORT activities in Europe. IVth International meeting of the International Society of Intraoperative Radiation Therapy. Miami, (FL): Mar 17–19, 2005.
30. Schultz MD. The supervoltage story. Amer J Roentgenol Rad Therapy Nucl Med 1975;124:541–559.
31. Henschke G, Henschke U. Zur Technik der Operationsbestrahlung. Strahlentherapie 1944;74:228–239.
32. Eloesser L. The treatment of some abdominal cancers by irradiation through the open abdomen combined with cautery excision. Ann Surg 1937;106:645–652.
33. Fairchild GC, Shorter A. Irradiation of gastric cancer. Br J Radiol 1947;20:511.
34. Podgorsak E, Metcalfe P, Van Dyk J. Medical Accelerators. In: Van Dyck J, editor. The modern technology of radiation oncology—a compendium for medical physicists and radiation oncologists. Madison, (WI): Medical Physics Publishing; 1999.
35. Mills MD et al. Shielding consideration for an operating room based intraoperative electron radiotherapy unit. Int J Radiat Oncol Biol Phys 1990;18:1215–1221.
36. Meurk ML, Schonberg RG, Haynes G, Vaeth JM. The development of a small, economic mobile unit for intraoperative electron beam therapy. Am J Clin Oncol 1993;16: 459–464.
37. Mills MD et al. Commissioning of a mobile electron accelerator for intraoperative radiotherapy. J App Clin Med Phys 2001;2: 121–130.
38. Daves JL, Mills MD. Shielding assessment of a mobile electron accelerator for intraoperative radiotherapy. J Appl Clin Med Phys 2001;2:165–173.
39. Piermattei A et al. The saturation loss for plane parallel ionization chambers at high dose per pulse values. Phys Med Biol 2000;45:1869–1883.
40. Fantini M et al. IORT Novac7: A new linear accelerator for electron beam therapy. In: Vaeth JM, editor. Intraoperative radiation therapy in the treatment of cancer. Vol. 31. Front. Radiation Therapy Oncology Basel: Karger; 1997.
41. Tosi G, Ciocca M. IORT with mobile linacs: The Italian experience. Oncologia 2004;27:350–354.
42. Hogstrom KR et al. Design of metallic electron beam cones for an intraoperative therapy linear accelerator. Int J Radiat Oncol Biol Phys 1990;18:1223–1232.
43. AAPM, American Association of Physicists in Medicine, Task Group 25, Radiation Therapy Committee, Clinical electron beam dosimetry, Report of AAPM radiation therapy committee Task Group 25. Med Phys 1991;18:73–109.
44. Almond PR et al. AAPM's TG-51 protocol for clinical reference dosimetry of high-energy photon and electron beams. Med Phys 1999;26:1847–1870.
45. Kutcher GJ et al. Comprehensive QA for radiation oncology, Report of AAPM Radiation Therapy Committee Task Group 40. Med Phys 1994;21:581–618.
46. Davis MG, Nyerick CE, Horton JL, Hogstrom KR. Use of routine quality assurance procedures to detect the loss of a linear accelerator primary scattering foil. Med Phys 1996;23:521–522.
47. Palta JR, et al. Intraoperative electron beam radiation therapy, technique, dosimetry, and dose specification, report of Task Group 48 of the radiation therapy committee, American Association of Physicists in Medicine. Int J Radiat Oncol Biol Phys 1995;33:725–746.
48. Beddar AS et al. Intraoperative radiation therapy using mobile electron linear accelerators: Report of the AAPM Radiation Therapy Committee Task Group 72. To be published.
49. Hazle JD, Chu JCH, Kennedy P. Quality assurance for intraoperative electron radiotherapy clinical trials: Ionization chamber and mailable thermoluminescent dosimeter results. Int J Radiat Oncol Biol Phys 1992;24:559–563.
50. Beatty J, et al. A new minature x-ray device for interstitial radiosurgery: dosimetry. Med Phys 1996;23:53–62.
51. Yanch JC, Harte KJ. Monte Carlo simulation of a minature radiosurgery x-ray tube using the ITS 3.0 coupled electron-photon transport code. Med Phys 1996;23:1551–1558.
52. Beddar SA. Dose delivery for HDR-IORT using curved HAM applicators. IVth International meeting of the International Society of Intraoperative Radiation Therapy. Miami, (FL): Mar 17–19, 2005.
53. Raina S et al. Quantifying IOHDR brachytherapy underdosage resulting from an incomoplete scatter environment. Int J Radiat Oncol Biol Phys 2005;61:1582–1586.
54. Harrison LB, Cohen AM, Enker WE. High-dose-rate intraoperative irradiation (HDR-IORT): Technical factors. In: Gunderson LL, Willett CG, Harrison LB, Calvo FA, editors. Intraoperative Irradiation: Techniques and results. Totowa, (NJ): Humana Press; 1999.
55. Gunderson LL et al. Locally advanced primary colorectal cancer: Intraoperative electron and external beam irradiation ± 5-FU. Int J Radiat Oncol Biol Phys 1997; 37:601–614.
56. Gieschen HL et al. Electron or Orthovoltage IORT for retroperitoneal sarcomas. In: Gunderson LL, Willett CG, Harrison LB, Calvo FA, editors. Intraoperative Irradiation: techniques and results. Totowa, (NJ): Humana Press; 1999.

See also RADIOSURGERY, STEREOTACTIC; X-RAY THERAPY EQUIPMENT, LOW AND MEDIUM ENERGY.

RADIOTHERAPY, THREE-DIMENSIONAL CONFORMAL

GEORGE STARKSCHALL
The University of Texas
Austin, Texas

INTRODUCTION

The goal of radiation therapy is to deliver a sufficient dose of radiation to a tumor site to control the disease while keeping doses to uninvolved tissue within tolerable limits. In order to achieve this goal, radiation beams are typically directed to the tumor target from several directions. Consequently, multiple beams irradiate the target, but only a few beams irradiate the uninvolved tissue surrounding the target, thus reducing the dose to the uninvolved tissue. The portals used to define the radiation fields are carefully shaped to fit the target in order to further reduce the amount of radiation reaching the tissue surrounding the target. In three-dimensional conformal radiotherapy (3DCRT), the radiation dose distribution is designed to conform to the contours of a target volume. The method by which the beam geometries and treatment portals of 3DCRT are designed differentiates 3DCRT from earlier methods of radiation treatment planning and delivery.

Prior to the 3DCRT era (mid-1990s), radiation treatment planning was based on a small number of computed tomography (CT) images of the patient and radiographic images from a simulator, a diagnostic X-ray machine whose geometries simulated the geometries of the radiation treatment machine. Beam configurations were determined based on a CT image of the patient in a single transverse plane, typically the plane containing the central axes of the radiation beams. Treatment portals were determined from the simulation radiographs; the design of these portals was based primarily on bony landmarks that were visible on the radiographs. Radiation doses were computed in the plane containing the beam central axes, and plans were evaluated on the basis of information displayed in that plane (1).

With the development of fast CT scanners and fast treatment planning computers with inexpensive memory, radiation therapy moved into the 3DCRT era. Whereas in the past, radiation treatments were planned and evaluated on the basis of information in a single or limited number of planes, radiation treatment planning is now based on patient volume. Patient information is acquired over a volume of the patient, treatment portals are designed to irradiate a target volume, doses are calculated in volumes of tissue, and treatment plans are evaluated based on dose–volume considerations. The 3DCRT process currently used in contemporary radiation therapy consists of the following steps: (1) accurately imaging the tumor and normal tissue in three dimensions, (2) precisely defining the target volumes, (3) optimizing beam geometries and beam weights, (4) translating the treatment plan to produce accurate delivery of radiation according to the treatment plan, and (5) verifying the accuracy of the delivery of radiation.

IMAGING FOR RADIATION ONCOLOGY

In the current state of practice in 3DCRT, treatment planning is based on virtual (CT) simulation. A 3D CT image data set of the patient is acquired with the patient in the treatment position. Patients are typically positioned in an immobilization device, and marks are placed on the patient's surface to assist in maintaining setup reproducibility. The desired setup accuracy may vary from submillimeter accuracy for treatments in the head, where the tumor may lie very close to highly critical, radiosensitive, normal anatomic structures, to accuracies of 0.5–1 cm in areas of the thorax or abdomen.

Various types of CT scanners are used to acquire patient information. The third generation CT scanner consists of a single radiation source emitting X rays in a fan-shaped radiation pattern, which is detected by a single array of detectors. The gantry containing the radiation source makes a single (whole or partial) rotation around the patient with the detectors on the opposite side of the patient. Following acquisition of projection data from a single transverse slice, the table is indexed and another slice is acquired. This procedure continues until the entire region of interest is scanned.

Technologic developments in CT image acquisition include helical, or spiral, CT. This technology combines gantry rotation with table translation so that the path of the radiation source with respect to the patient forms a spiral trajectory. Image acquisition times using helical CT are significantly faster than times using third generation CT. Even faster is a new technology in which a multislice X-ray detector is used. Typically, such detectors can acquire from 4 to 16 image slices simultaneously. Multislice helical CT has the potential for increased axial resolution as well as scan times that can be as short as 3–5 s (2).

The ability to increase the axial resolution of CT image data sets brings up the issue as to the desired axial resolution for planning 3DCRT. Prior to the 3DCRT era, CT scans were typically acquired at planar separations of $\sim$10 mm. This axial resolution should be compared with the typical picture element (pixel) dimensions in each plane of <1 mm^2. The ideal axial resolution would be comparable with the planar resolution; however, a submillimeter axial resolution would result in a very large number of CT images. Because the contours of tumors as well as those of some normal anatomic structures have to be manually delineated on the CT images for treatment-planning purposes, requiring such delineation on so many CT images would result in an extraordinarily time-consuming, labor-intensive task. Consequently, axial resolutions presently used in 3DCRT represent a compromise and are typically in the vicinity of 3 mm.

A key component in CT simulation is the production and display of digitally reconstructed radiographs (DRR). A DRR is a radiographic image that models the transmission of an X-ray beam through the patient (3). A DRR is obtained by tracing the path of X rays from the location of the radiation source through the 3D CT patient image data set to a plane beyond the data set and calculating the X-ray transmission through the CT data set. Generating a DRR offers several advantages over obtaining radiographic images from a conventional simulator. Whereas tumors are sometime difficult to visualize on a conventional radiograph, they can often be visualized more easily on a CT image. The contours delineating the tumor can be drawn on

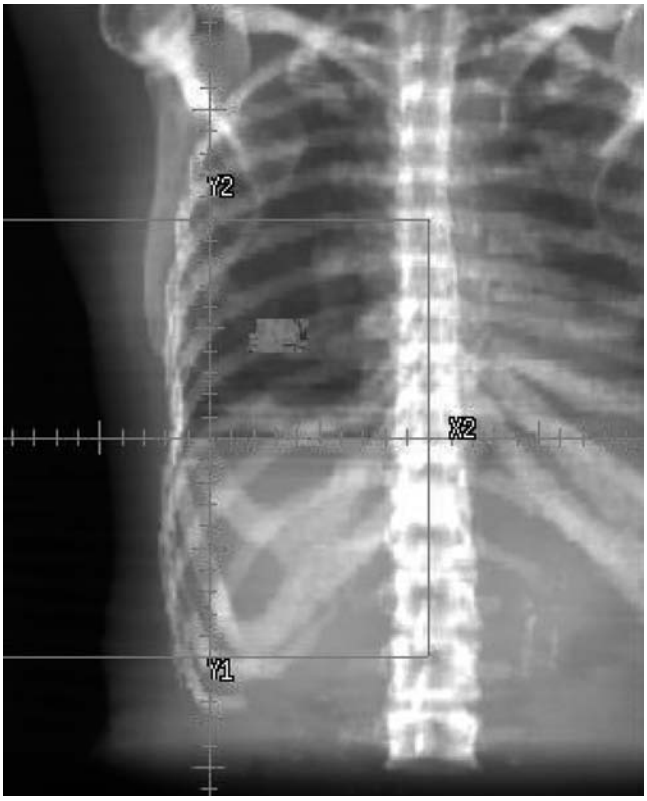


Figure 1. A digitally reconstructed radiograph illustrating a tumor in a patient's right lung.

the CT image data set and projected onto the DRR. Treatment portals are then designed based on the projected contours. The design of a treatment portal based on the volumetric extent of the tumor is a key component of the 3DCRT process. In addition, many soft-tissue anatomic structures may also be difficult to visualize on a conventional radiograph. Contours delineating these structures can be drawn on the CT image data set and projected onto the DRR as well. Display of these projected contours can assist the treatment planner in selecting radiation-beam geometries that will irradiate the tumor while avoiding critical anatomic structures. Finally, the DRR can be compared with a radiographic image of the treatment field acquired on the treatment machine to verify that the patient is correctly set up on the treatment machine. Figure 1 illustrates a DRR with a target volume drawn on it.

Computed tomography image data sets are not the only images used in planning the 3DCRT. They are the images of choice for performing actual radiation-dose calculations because the values of the CT volume elements (voxels) can best be correlated to the extent of interactions of the X-ray beam with the patient. However, other imaging modalities may be more effective than CT in imaging the tumor. For example, magnetic resonance (MR) imaging is often more reliable for imaging soft tissue, especially for planning radiation treatment of brain tumors. Positron emission tomography (PET) is a very useful modality for imaging tumor metabolism and often provides additional information about the presence and extent of lung tumors. Figure 2 illustrates a CT image and a PET image illustrating the differences in the information provided by the two imaging modalities.

A major issue encountered in the use of multimodality imaging for 3DCRT planning is that of image registration, that is, the geometric correlation of information from the various imaging modalities with that of the CT image data set. The coordinate systems used in the acquisition of images from various modalities are typically different, requiring some sort of registration so that information delineated on one image data set is correctly displayed on the treatment-planning CT data set. Moreover, MR images are sometimes subject to geometric distortion,

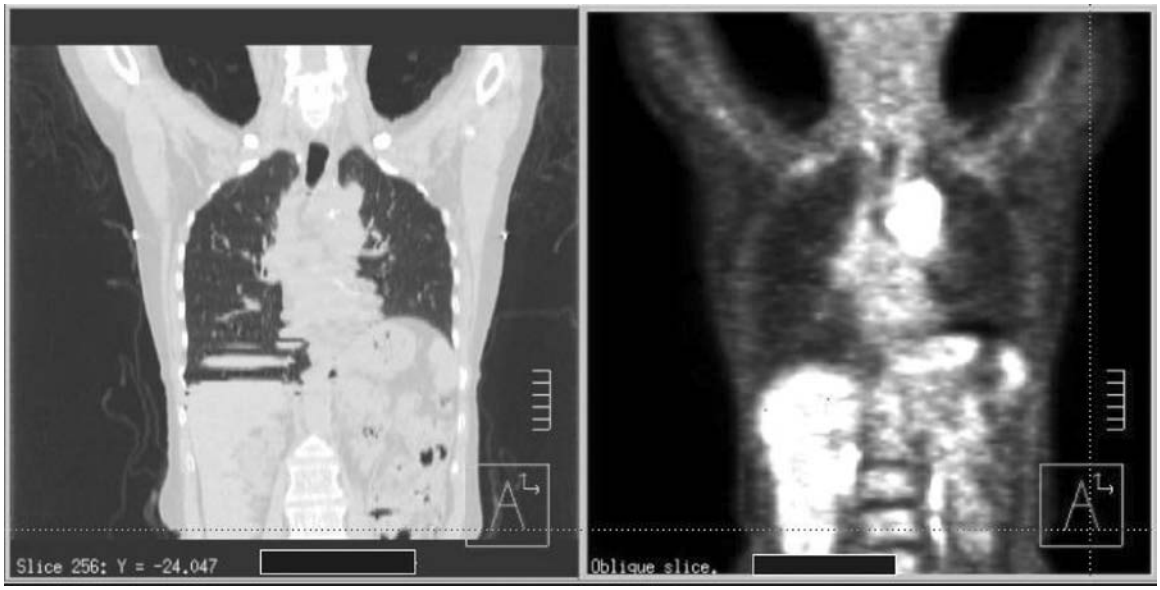


Figure 2. Coronal CT and PET images of a patient with lung cancer.

Figure 3. A combination PET/CT scanner.

especially near the periphery of such images, making the registration problem more complex. The PET and CT images are very difficult to register manually, as the two modalities image very different properties of the patient. Registration of PET and CT image data sets is sometimes accomplished by registering the CT image data set with a transmission scan taken on the PET scanner and used to make corrections in the PET scan for patient attenuation. A new device, in which a PET and a CT scanner are mounted together, has recently been introduced. An example of such a device is shown in Fig. 3. Using this device, the patient first undergoes a CT scan. With the patient immobilized, the support table is translated into the PET scanner, and a PET scan is obtained. In this manner, the geometries of the PET scan and CT scan differ only by a known table translation, making image registration a simple process.

DEFINITION OF TARGET VOLUMES

Because dose prescriptions used in 3DCRT are based on irradiating volumes of tissue, more uniformity in terminology has been required in the specification of dose prescriptions. The International Commission on Radiation Units (ICRU) has developed a series of documents that describe a methodology for specification of target volumes (4,5). The ICRU definitions constitute a methodology for reporting radiation treatments in an unambiguous manner.

The ICRU defines the gross tumor volume (GTV) to be the "gross demonstrable extent and location of the malignant growth" (4). The GTV consists of the primary tumor, involved lymph nodes, and metastatic disease that can be visualized. The specific method of visualization is not specified; the GTV might be identified and delineated on clinical examination, radiographs, CT images, MR images, or any other visualization method. Moreover, the extent of the GTV may be different for different examinations, depending on the method of visualization. One of the aims of the radiation therapy is to deliver a tumoricidal dose to the entire GTV. The patient may have several GTVs depending on the nature of the malignant disease, and each GTV may have a different therapeutic aim. In some cases, in particular after surgical intervention, the GTV may not exist.

In addition to clinically demonstrable disease, the patient is likely to have subclinical disease that can only be visualized under pathologic examination. This subclinical disease must also be eliminated for the tumor to be controlled. The combination of GTV and subclinical disease constitutes the clinical target volume (CTV). The radiation oncologist establishes the margin that defines the CTV on the basis of clinical experience. In a few cases, such as nonsmall cell lung tumors, pathology studies have helped the radiation oncologist define the CTV margin. In the case of surgical intervention, a CTV might exist without a GTV. The existence of GTVs and CTVs are based on general oncologic principles and are independent of any therapeutic approach. The GTVs and CTVs can be identified and delineated prior to the selection of the treatment modality and the treatment-planning procedures.

Treatment planning for 3DCRT is typically based on information obtained from a CT image data set acquired a few days before treatment is scheduled to begin. An inherent assumption in basing the treatment plan on the CT data set is that the data set accurately models the patient during the entire course of radiation therapy. In reality, the patient moves, exhibiting intrafractional motion such as respiratory and cardiac motion, as well as interfractional motion such as bladder and rectum filling and tumor regression. In order to account for motion, an internal target volume (ITV) is defined. The ITV is the CTV with an internal margin (IM). In many cases, the ITV is assumed to be the CTV with a uniform, isotropic IM, established from clinical experience; but in some cases, the ITV is explicitly defined from multiple CT image acquisition.

A final margin needs to be added to the ITV to account for the fact that there is some degree of nonreproducibility in the day-to-day setup of the patient for radiation treatment. Consequently, a planning target volume (PTV) is defined to include the ITV along with a setup margin (SM) to account for these setup uncertainties. The PTV is thus defined to be a region that, if irradiated to the prescription dose, makes sure that the entire CTV will be irradiated to the prescription dose, regardless of internal motion or setup uncertainty. The extent of the SM is based on the device used to immobilize the patient. When invasive devices are used, the SM may be reduced to <1 mm; typically, however, the SM is in the range 0.5–1.0 cm. Figure 4 illustrates an axial CT image of a patient's lung with the GTV, CTV, and PTV identified.

In 3DCRT, a DRR with the delineated PTV is normally displayed to the treatment planner. A radiation-treatment portal is established to include the entire PTV along with an additional 5–7 mm margin that accounts for the fact that the edge of the radiation beam is fuzzy due to the finite size of the radiation source as well as effects of scattered radiation. Thus, a GTV is expanded by a margin to account for microscopic disease to generate a CTV, which is expanded by a margin to account for motion to generate an ITV, which, in turn, is expanded by a margin for setup uncertainty to generate a PTV.

In addition to identifying and delineating target volumes, the 3DCRT planning process must also identify

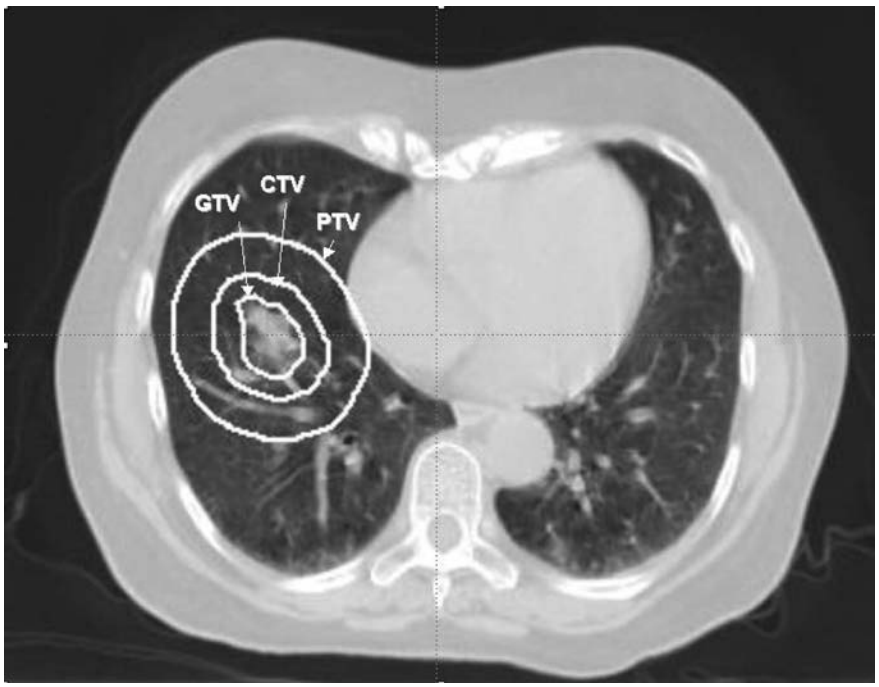


Figure 4. An axial CT image of a patient's thorax, illustrating a GTV, CTV, and PTV. The CTV is obtained by expanding the GTV by a uniform 0.8 cm margin, and the PTV is obtained by expanding the CTV by a uniform 1.5 cm margin.

and delineate organs at risk (OAR), which are defined to be "Normal tissues whose radiation sensitivity may significantly influence treatment planning and/or prescribed dose" (5). For example, in planning the radiation treatment of lung tumors, the OARs typically include the right and left lung, heart, esophagus, and spinal cord, whereas in planning the radiation treatment of prostate tumors, the OARs typically include the bladder, rectum, and femoral heads. Setup uncertainty and OAR motion are accounted for in defining the planning organ at risk volume (PRV), which contains the OAR along with an internal margin and a setup margin (5). Thus, in the planning of 3DCRT, the planner must determine one or more beam geometries and treatment portals that will adequately irradiate the PTV, while keeping doses to the various PRVs within tolerable limits.

DOSE PRESCRIPTION

Prior to the development of 3DCRT, radiation-dose prescriptions were generally based on point-dose methodology. A dose prescription might have read, "66 Gy to the isocenter", where the isocenter was the point around which the gantry of the radiation machine rotated. Another prescription might have read, "72 Gy to the 95% isodose", where the 95% isodose represented the locus of those points receiving 95% of the dose to isocenter. In both cases, the dose prescriptions described doses to a single point, typically the point at isocenter.

In 3DCRT, one is concerned with doses delivered to volumes, in particular, the target volumes. The goal of a 3DCRT treatment plan is to irradiate the entire CTV to the prescription dose. Because the CTV may move anywhere within the PTV, it would also be desirable for the entire PTV to be irradiated to the prescription dose as well. However, because the CTV does not lie in the entire PTV for the entire treatment, one may allow a small amount of the PTV (typically 5%) to receive a dose less than the prescription dose without adversely compromising the goals of the radiation treatment, especially if doing so would significantly improve organ sparing. If $<100\%$ of the CTV receives the prescription dose, the radiation oncologist may have to address the potential consequences that could occur if tumor cells do not receive a tumoricidal dose. However, compromises in CTV dose may be necessary to avoid excessive morbidity. In many cases, the 3DCRT planning process represents a compromise between the dose needed for adequate tumor control and the dose that would cause unacceptable side effects of the radiation.

BEAM DETERMINATION

Determination of radiation beam geometries for 3DCRT is, to an initial approximation, a trial-and-error process based on class solutions. For example, a pair of parallel-opposed beams delivers approximately a uniform dose to the entire volume irradiated by the beams. Tumors, however, rarely extend completely through the patient; so in 3DCRT, more than two beams are likely to be used. Placing two more beams at right angles to a parallel-opposed pair of beams generates what traditionally has been called a "four-field box". This configuration delivers $\sim$50% of the target dose to regions that are irradiated by only two of the four beams. More complicated beam geometries are sometimes used based on the need to avoid irradiation of critical structures. Figure 5 illustrates a multiple-field beam configuration used to treat prostate cancer. Incorporating couch rotations resulting in beams whose axes are noncoplanar can sometimes be used to improve avoidance of critical structures. Wedge-shaped metallic filters placed in the beam to generate dose gradients can be used to compensate for dose nonuniformities generated from certain beam configurations.

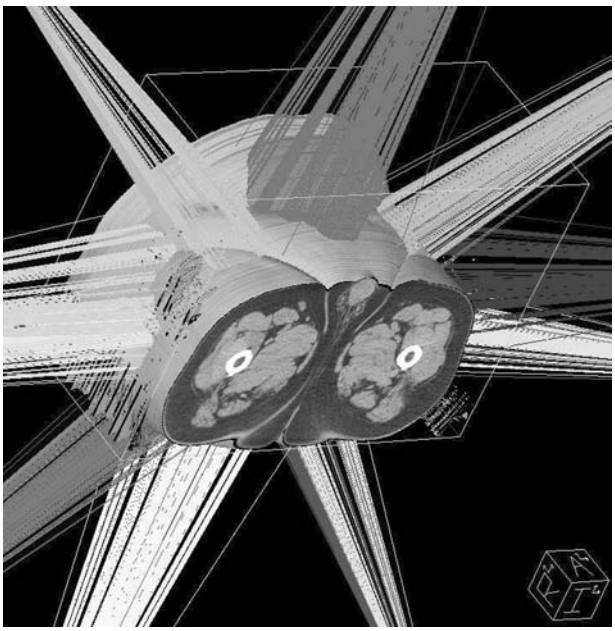


Figure 5. A multiple field beam configuration used to treat prostate cancer.

Treatment portals are determined for each beam geometry and designed to encompass the PTV along with a margin to allow for the fact that the beam does not end abruptly at the geometric edge but rather the dose distribution has a finite slope at the beam edge. Caused both by the fact that the radiation source is not a point source and that radiation can scatter from the irradiated region to the blocked region, this penumbra requires that the treatment portal surround the PTV with a margin of from 0.5 to 1 cm in order to deliver a high dose to the PTV. The treatment planning typically designs the treatment portal based on viewing the BEV projection of the PTV on a DRR.

In addition to optimizing the beam geometries, the planning often must also optimize the beam weights. Typically, this optimization consists of manual fine tuning of the beam weights to deliver a more uniform dose across the target volume, decrease the dose to specific critical structures, or both of these outcomes. However, if specific prescription goals are established, such as desired doses to target volumes and maximum allowed doses to critical structures, various computer algorithms can be used to optimize the beam weights. If additional conformality of the dose distribution to the target volume is desired, techniques exist for the planning and delivery of intensity-modulated radiation therapy (IMRT). This technique is particularly useful if the target volume is concave.

DOSE CALCULATION

In most current radiation treatment planning systems, X-ray dose distributions are calculated using a convolution algorithm (6,7). In this algorithm, the radiation dose is obtained by convolving an incident fluence distribution with a dose-spread array. The dose-spread array represents the distribution of radiation dose around a point in which a unit amount of X-ray fluence is seen. Because the dose-spread array is dependent solely on the X-ray energy and distance from the point of fluence deposition, dose-spread arrays, one for each energy, need only be computed once, and then stored in the treatment-planning computer. The in-air fluence exiting the linear accelerator is determined once for each linear accelerator and is modeled by a set of parameters. The X-ray fluence is calculated by propagating the in-air fluence through the CT model of the patient, attenuating the fluence based on exponential attenuation, characteristic of X-ray propagation through matter.

More accurate calculations of X-ray dose distributions in the patient can be achieved by more accurate modeling of the interactions that X rays undergo with the patient using Monte Carlo techniques (8). Originally limited in applicability by limitations in computing power, newer high speed processors are making such calculations clinically feasible.

CONFORMAL PARTICLE THERAPY

The 3DCRT can also be achieved using beams of charged-particle radiation, such as electrons or protons. In some respect, achieving 3DCRT is easier with charged-particle beams than it is with X-ray beams. One feature that differentiates the dose distribution from charged-particle beams from that of X-ray beams is that the dose distribution from charged-particle beams, at least to a first approximation, is constant from the patient surface to a depth equal to the range of the charged particle, and then goes rapidly to zero. This distribution results from the charged particle's property of depositing a fixed amount of energy per interaction until the particle energy is exhausted; beyond that depth, no energy is deposited. Consequently, the lateral extent of a charged-particle beam can be shaped based on the BEV projection of the PTV, just as are X-ray beams, but the depth of penetration of a charged-particle beam can be made to track the distal surface of the PTV by spatially modulating the energy of the beam, which shifts the range of the charged particles. A simple method for range-shifting involves placing a sufficient amount of attenuating material in the path of the beam to reduce the local energy of the beam so that its range is slightly greater than the depth of the distal surface of the PTV. Alternatively, one could modify the spatial distribution of the energy of the beam by modifications in the beam transport in the head of the particle accelerator.

Particle therapy, on the other hand, has some limitations. The margins of electron beam dose distributions are not as sharp as those for X-ray beams because of multiple scattering. In addition, heterogeneities and surface irregularities cause significant perturbations in the electron beam dose distributions. Widespread use of proton beams may be limited because of the significantly higher cost of production of therapeutically useful beams.

RADIATION DELIVERY

Once a 3DCRT treatment plan has been developed, it is necessary to translate the treatment plan accurately to the

radiation-delivery system. Although this can be done manually, the complexity of modern treatment plans demands that plan information be transferred automatically from the treatment-planning computer to a linear accelerator. Because of the variety of treatment-planning systems and linear accelerators, significant effort has been expended to standardize the exchange of information among the imaging systems, treatment planning computers, and treatment delivery systems.

VERIFICATION OF DELIVERY

The final step in 3DCRT is that of quality assurance, verification that the radiation beams delivered to the patient are consistent with those planned for delivery. All components of the planning and delivery process need to be reviewed as part of a quality assurance process, including the validity of the treatment plan, the accuracy of the machine settings, the validity of the transfer of data from the planning system to the radiation machine, the accuracy of the portal for delivery of the radiation beam, the accuracy of the patient setup, and the accuracy of the radiation delivery.

The first step in verification is that of verifying that the treatment plan accurately reflects the radiation dose that is actually designed to be delivered to the patient. This is achieved by implementation of a quality assurance program for the treatment planning system. In such a program, one verifies that the treatment-planning system accurately depicts the patient images, contours, and beams that are provided as input into the dose calculations, and that the beam models are accurate, so that the doses calculated by the treatment planning system are accurate (9).

The next component of verification is that the machine setting determined by the treatment-planning system will deliver the correct dose to the target. Accuracy of machine setting is verified initially when the beam model is first commissioned and introduced into the treatment-planning system and is verified for each dose calculation by and independent calculation of radiation doses to individual points within the patient.

Next, it is necessary to verify that the machine settings, such as beam geometries, and monitor units are accurately transferred from the treatment-planning system to the machine. In many radiation oncology clinics, a record and verify (R&V) system is used to record the machine settings used to deliver the radiation beams to the patient, ensuring that machine settings lie within clinically accepted tolerances of the values established on the treatment plan. A component of a comprehensive quality assurance program verifies that the machine settings determined by the treatment-planning system are accurately transferred into the R&V system.

Another key step is to verify that the treatment field actually delivered to the patient is identical to the treatment portal as determined on the treatment plan. In order to accomplish this task, a radiographic image of the delivered treatment portal is acquired. The portal image is then compared to a DRR of the treatment field extracted from the treatment-planning system. Ideally, the geometric relationship between the tumor target and the treatment portal would be compared, but tumors are rarely visible on portal images. Consequently, bony landmarks in the vicinity of the tumors are typically used as surrogates for the tumors. Originally, the portal image was recorded on radiographic film, but many radiation oncology clinics are acquiring this information using an electronic portal imaging device (EPID). The use of digital images produced by EPIDs allows for rapid display of the portal images as well as off-site review of the portal images. Figure 6 illustrates a DRR and a portal image used to verify the accuracy of the treatment portal.

One important issue regarding the use of portal images for treatment portal verification is the substantial difference in image quality between simulation and portal images. Simulation images are typically acquired in the energy range of 50–100 keV. In this energy range, the primary interaction between the incident X rays and the patient results in absorption of the incident X rays. Consequently, the only X rays that reach the detector are those that are not absorbed in the patient. Moreover, in this energy range, bone absorbs significantly more radiation than soft tissue; hence, the contrast between bony anatomy and the surrounding soft tissue anatomy results in clear, sharp radiographic images. Portal images are acquired at much higher X-ray energies, typically several megaelectronvolts, the energies that are used in radiation therapy. In this energy range, the difference in absorption of radiation between bone and soft tissue is significantly less, resulting in lower contrast. Moreover, the presence of scattered radiation resulting from X-ray interactions with the patient at these energies results in a significant amount of radiation reaching the detector that gives no indication of the point of origin, resulting in a noisy image.

In addition to verifying the geometry of the treatment portal, acquisition of a portal image is one of several techniques used to verify that the patient has been set up in a reproducible manner for each treatment. Because the treatment field is often positioned adjacent to critical uninvolved anatomic structures, accurate and reproducible patient positioning is essential so that these uninvolved anatomic structures do not get unnecessarily irradiated. Perhaps the simplest method of position verification is through the use of external markings. Lasers in the walls and ceilings are all directed to a particular point in space, the machine isocenter, where the axis of gantry rotation coincides with the axis of collimator rotation. In 3DCRT, the patient is often positioned so that the isocenter lies in the approximate center of the tumor volume. The points at which the lasers intersect the patient surface are marked and used on a daily basis to assist in ensuring that the patient is set up reproducibly.

In conjunction with external markings, patients are often placed in immobilization devices to assist in reproducible positioning. Immobilization devices have many forms. For example, invasive devices, such as stereotactic head frames, which are screwed into the patient's skull, can achieve submillimeter reproducibility and are necessary when the geometric tolerances are very small. When

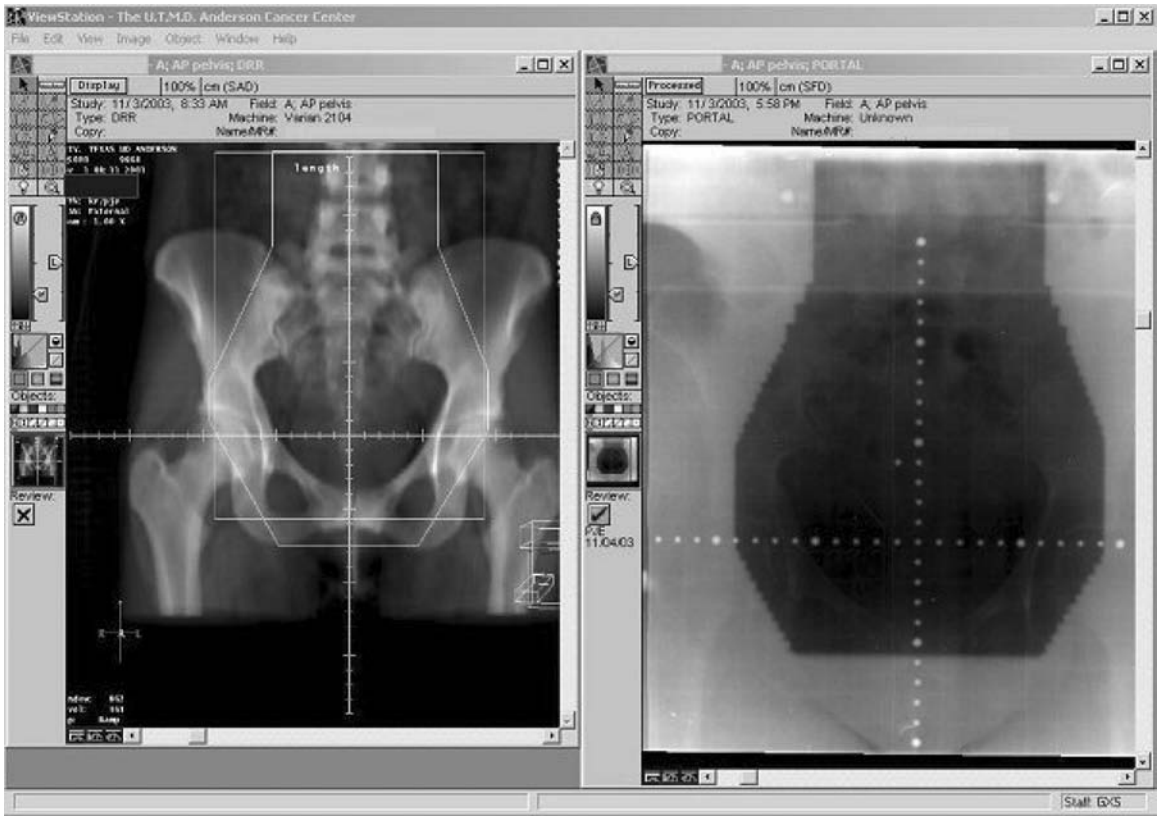


Figure 6. A DRR of an anterior pelvic treatment field compared with a portal image of the same field used to verify that the patient is set up accurately and that the location and extent of the irradiated region is as per plan.

submillimeter tolerance is not as crucial, less invasive devices can be used, including thermoplastic masks, molds, and vacuum bags.

Portal-imaging studies have demonstrated differences between true setup errors and random uncertainties in patient setup, which are a consequence of the degree of patient immobilization, and have indicated guidelines as to when and how to correct for setup inaccuracies. In the treatment of some cancer sites, for example, the prostate, daily variations in patient anatomy have led to the development of more sophisticated methods of setup verification. Differences in bladder and rectal filling on a daily basis combined with the tight treatment margins typical of 3DCRT may cause one or more radiation beams to miss part of the target volume. One technique that enables correction for daily anatomic variations is to scan the patient just prior to treatment using an ultrasound technique. Figure 7 illustrates an example of the unit used in such an ultrasound process. The outlines of the target volumes and anatomic structures, which are extracted from the radiation treatment plan, are superimposed on the ultrasound scan, and the patient is moved so that the image of the target volume on the scan is superimposed on the outline of the target volume on the treatment plan, thus ensuring accurate delivery of radiation to the target volume. Figure 8 illustrates the use of the ultrasound images in realigning a patient.

Another promising technique involves placing a CT scanner in the treatment room in a configuration that allows both imaging and treatment on the same patient couch. The patient is set up in the treatment position and then scanned. It is then possible to compare the CT image data set thus acquired with the data set from which the treatment plan is based, allowing for the patient to be repositioned or even replanned. Rather than allowing the patient table to move through the CT scanner, as is conventionally done, this device moves the CT scanner while the patient remains stationary, hence the device is referred to as "CT-on-rails".

In addition to verifying that the radiation beams accurately irradiate the target volume and spare surrounding normal-tissue anatomy, it is essential to verify that the radiation dose actually delivered to the target is identical to the radiation dose prescribed on the treatment plan. This is achieved by means of a thorough quality assurance system, including a regular schedule of well-defined daily, monthly, and annual evaluations of the output of the radiation machine, including accurate measurements of the magnitude of the radiation emitted from the linear accelerator as well as the energy of the radiation (10). In addition, techniques exist for measurement of doses to accessible parts of the patient during treatment, using small radiation detectors such as thermoluminescent dosimeters (TLD) or diode detectors, which can be placed either on external surfaces of the patient or in accessible cavities within the patient. Current studies are underway to assess the safety and accuracy of radiation detectors that can be implanted into the tumor inside the patient.

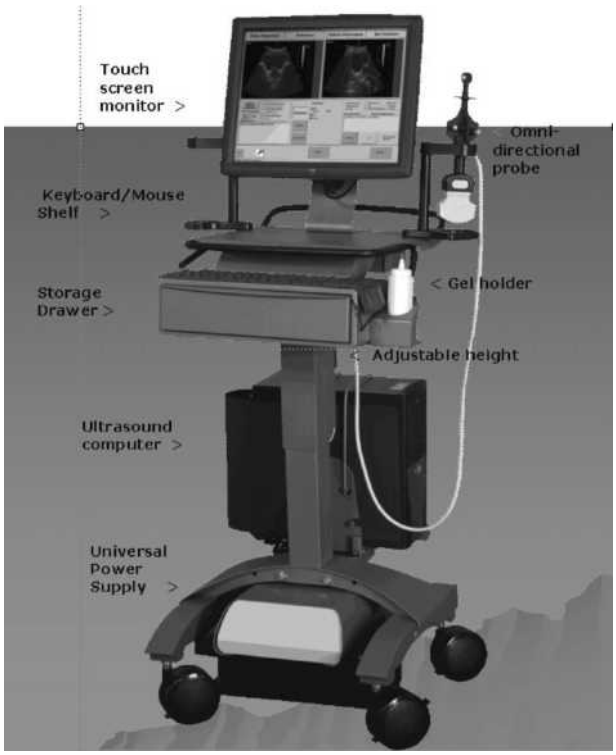


Figure 7. An ultrasound imaging device used to verify reproducibility of patient position for radiation treatment of the prostate. (Figure courtesy of NOMOS Radiation Oncology division of North American Scientific.)

CLINICAL CONSEQUENCES OF 3DCRT

Finally, it is necessary to determine whether 3DCRT represents an improvement in the treatment of cancer. Treatment-planning studies can readily demonstrate that 3DCRT allows higher doses to be delivered to target volumes than was possible with conventional radiation therapy. Higher radiation doses have been shown to lead to increased rates of tumor control. An early study comparing 3DCRT to conventional radiation therapy in the treatment of prostate cancer was conducted by Pollack (11) in the mid-1990s. In this study, patients were randomly selected to receive radiation therapy based on conventional treatment planning or a dose ~10% higher based on 3DCRT. Pollack demonstrated a significant increase in freedom from failure for intermediate-to-high risk patients, but with some increase in rectal toxicity.

CONCLUSIONS

Three-dimensional conformal radiotherapy is a radiation treatment planning and delivery technique in which the design of the radiation-treatment portals is based on images of the tumor target. Planning is based on high resolution CT images that explicitly display the tumor as well as provide a mathematical model for the patient. Additional imaging modalities are often used to define the extent of tumor involvement more clearly. Sophisticated algorithms are used to calculate the magnitude of radiation dose deposited in the patient, allowing the assessment of the potential for tumor control as well as toxicity of the treatment plan. In order to make certain that the radiation is delivered to the tumor and not the surrounding uninvolved tissue, as indicated in the treatment plan, extensive quality assurance procedures are required that verify the accuracy of the geometry as well as the radiation dosimetry.

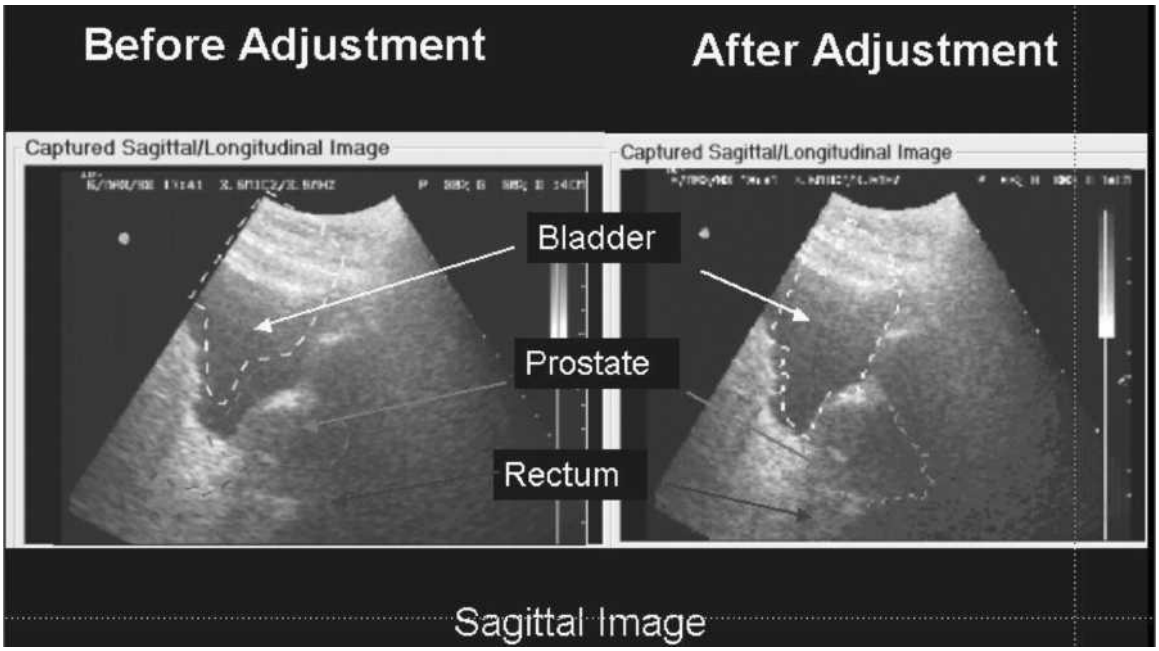


Figure 8. Sagittal ultrasound images of a patient before and after realignment of the patient. (Figure courtesy of Lei Dong, Ph.D., U. T. M. D. Anderson Cancer Center.)

BIOLIOGRAPHY

1. Starkschall G. What is 3-D radiotherapy treatment planning. In: Purdy JA, Starkschall G, editors. A Practical Guide to 3-D Planning and Conformal Radiation Therapy. Madison, WI: Advanced Medical Publishing; 1999. p 1–16.
2. Bushberg JT, Seibert JA, Leidholdt EM, Boone JM. The Essential Physics of Medical Imaging. Philadelphia, PA: Lippincott Williams & Wilkins; 2003. Chapt. 13.
3. Goitein M, Abrams M, Rowell D, et al. Multi-dimensional treatment planning: II Beam's eye-view, back projection, and projection through CT sections. Int J Radiat Oncol Biol Phys 1983;9:789–797.
4. International Commission on Radiation Units and Measurements (ICRU) Report 50: "Prescribing, Recording, and Reporting Photon Beam Therapy". ICRU (1993) Bethesda MD.
5. International Commission on Radiation Units and Measurements (ICRU) Report 62: "Prescribing, Recording, and Reporting Photon Beam Therapy (Supplement to ICRU Report 50)". ICRU (1999) Bethesda MD.
6. Mackie TR, Scrimger JW, Battista JJ. A convolution method of calculating dose for 15 MV x-rays. Med Phys 1985;12:188–196.
7. Boyer AL, Mok EC. A photon dose distribution employing convolution calculations. Med Phys 1985;12:169–177.
8. Ma CM, Mok EC, Kapur A, Pawlicki T, Findley D, Brain S, Forster K, Boyer AL. Clinical implementation of a Monte Carlo treatment planning system. Med Phys 1999;26:2133–2143.
9. Fraass BA, Doppke KP, Hunt MA, et al. Task Group 53: Quality assurance for clinical radiotherapy treatment planning. Med Phys 1998;25:1773–1829.
10. Kutcher GJ, Coia L, Gillin M, et al. Comprehensive QA for radiation oncology: Report of AAPM Radiation Therapy Committee Task Group 40. Med Phys 1994;21:581–618.
11. Pollack A, Zagars GK, Starkschall G, et al. Prostate cancer radiation dose response: Results of the M.D. Anderson phase III randomized trial. Int J Radiat Oncol Biol Phys 2002;53: 1097–1105.

Further Reading

Khan F. The Physics of Radiation Therapy. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2003. especially Chapt. 19.

See also RADIATION DOSIMETRY FOR ONCOLOGY.

RADIOTHERAPY TREATMENT PLANNING, OPTIMIZATION OF

ANTHONY B. WOLBARST
Georgetown Medical School
Washington, DC

LEE CHIN
Harvard Medical School
Boston, Massachusetts

PAVEL STAVREV
Cross Cancer Institute
Edmonton, Alberta, Canada

INTRODUCTION

Treatment of solid tumors typically involves some combination of surgery, chemotherapy, and radiation therapy. With any of these approaches—indeed, as with so much of medicine—the objective is to eradicate or control the disease without producing unacceptable side effects. Radiation therapy, in particular, is used for one-half of all cancer patients in the United States, and can frequently contribute to the cure of a malignant tumor, or at least to a significant improvement in quality of life.

Intense localized irradiation of the region of a solid tumor can both incapacitate or kill cancer cells directly, and strangle them through damage to tumor-bed microvasculature. High doses of ionizing radiation will kill not only the cells of tumors, however, but also those of healthy tissues. So a successful treatment must target the tumor accurately, with little radiation ending up elsewhere; it would be of questionable benefit to eradicate a tumor of the esophagus if this also led to severe functional complications in the spinal cord.

Radiation therapy is therefore preceded by a careful, and sometimes extensive, planning process, and a search for a best treatment plan.

This article is based on material found in Ref. 1.

ASSESSING BENEFITS AND RISKS

Deciding upon a medical strategy, like most other activities in life, involves an optimization process. In some situations, the proper approach is so obvious (Take two aspirin and call me in the morning!) that one hardly gives thought to the matter. In the treatment of cancer with radiation, however, the decisions are usually much more difficult. It may not even be clear, in fact, how the various possible criteria should be balanced in choosing the best treatment.

Ultimately, the desired result of radiotherapy is the complete disappearance of the disease, or at least long-term palliation, without the onset of unacceptable side effects. For some types of cancer, such as carcinomas of the skin, the malignant cells can be exposed directly to radiation, with limited risk to any critical organs. Also, they may be inherently more susceptible to radiation damage than are the adjacent healthy tissues. The reasons for this, while not completely understood, involve the tendency of tumor cells to divide faster than healthy ones—if both kinds are irradiated, a greater fraction of the cancer cells will be undergoing division, during which time they are more vulnerable to radiation damage, and a greater fraction of them will die. They may also have a diminished ability to repair radiation damage. In any case, complete cure without complication is frequently achievable.

For diseases in which the neoplastic (cancerous) cells are relatively radioresistant or less accessible to irradiation, however, the odds are greater that a curative dose of radiation would cause extensive damage to surrounding healthy tissues. The amount of radiation to be delivered to the tumor region and the means of physically delivering it must then be determined by assessing the likelihood of effecting a cure or prolonged survival against that of inducing unacceptable complications.

Thus any new patient presents the radiation oncologist, in theory, with a fundamental, three-part problem. For

every reasonable treatment strategy (including the various reasonable dose distributions and their timing), the physician has to

1. Estimate the probabilities of eradicating the disease, on the one hand, and of inducing complications of a range of severities, including death, on the other; such numbers are, in general, not well known.
2. Assuming that they have somehow arrived at plausible values for these probabilities, the physician and staff must then assign a weight, or measure of the seriousness or undesirableness, perhaps relative to death by the disease, to all the possible complications; such an assignation cannot help but be highly subjective.
3. Finally, they must combine the probability and undesirability information for every strategy in such a way as to arrive at a medically significant overall figure of its merit; the treatment plan with the best such score would then indicate the regimen of choice. Unfortunately, there is no obvious or simple way of doing this.

Thus the physician and patient must together decide how much they would be willing to compromise the quality of life in an attempt to preserve it. Unfortunately, the currently available data on the probable outcomes of the different therapeutic approaches are both sparse and crude. The response of liver to non-uniform irradiation has been studied, for example, but this kind of work is still in an early stage (2,3). And even if the essential statistical information were available, there is no unequivocal way of weighting it with the necessary value judgments on harm in a systematic treatment-selection process. In other words, there is considerable art in treatment planning, in addition to the science.

A PRELIMINARY EXAMPLE

Let us begin by considering some of the difficulties to be found in even the simplest of idealized cases.

A hypothetical patient presents with pockets of lethal disease within a single vital organ, and will die if either the disease spreads or the organ fails. Suppose that it is possible to irradiate this tumor-bearing organ uniformly and without risk to other tissues. And finally, assume that the dose-response characteristics both of the tumor and of the organ itself have separately been determined, Fig. 1.

The probability that the patient will survive the ravages of the disease, and that the tumor will be controlled, $S_{\text{tumor}}(D)$, will improve with increasing dose, D, until every viable tumor stem cell has been eliminated. On the other hand, the odds that the patient will escape unacceptable (e.g., life-threatening) complications to the organ, $S_{\text{organ}}(D)$, will be a decreasing sigmoidal function of D. (Either way, S is used to denote the probability of patient **S**urvival, or well being.) The parameters $S_{\text{tumor}}(D)$ and $S_{\text{organ}}(D)$ are also known as the tumor control probability (TCP) and normal tissue complication probability (NTCP), respectively, and by other names as well.

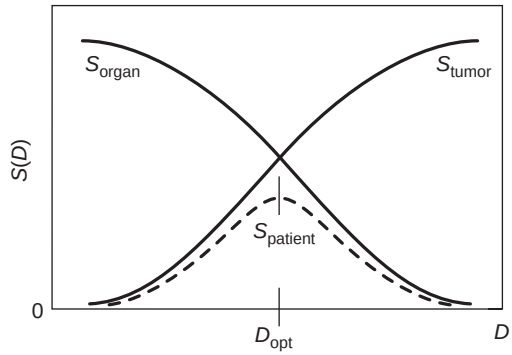


Figure 1. For this idealized, hypothetical example, the tumor is spread throughout a single critical organ, and only that organ is irradiated. The probability of the patient surviving the ravages of the disease, $S_{\text{tumor}}(D)$, increases with dose, D, to the tumor volume. Unfortunately, the likelihood of escaping an unacceptable complication that might arise from the treatment, $S_{\text{organ}}(D)$, decreases with D. The compound odds of both eradicating or controlling the disease and avoiding the complications is then $S_{\text{tumor}}(D) \times S_{\text{organ}}(D)$. This overall survival probability, $S_{\text{patient}}(D)$, reaches its maximum for some optimal value of the dose, D_{opt}.

At any level of uniform dose, D, the overall likelihood of uncomplicated survival, $S_{\text{patient}}(D)$, is then given by the law of multiplication of probabilities for independent events,

$$S_{\text{patient}}(D) = S_{\text{tumor}}(D) \times S_{\text{organ}}(D) \tag{1}$$

shown as the dashed line. At low doses, failure of the treatment through continuing tumor growth is likely. At high doses, an unacceptable complication in the organ is liable to occur. But between these two extremes, the patient may well survive both the disease and the effects of the treatment. At some particular optimal value of the dose, D_{opt}, the probability of that happening is greatest, and $S_{\text{patient}}(D)$ passes through a maximum.

Figure 1 illustrates a case in which there is a fairly broad range of dose over which the probability of uncomplicated cure is high, a situation said to be of good therapeutic ratio. When this is not the case, radiation treatment is a less promising option.

If the sole consideration in selecting D were the maximization of the probability of complication-free survival, then the appropriate value would be D_{opt}. This theoretical criterion, however, is often difficult to apply in practice, since it is generally accepted that the incidence of severe complications must be kept to a "tolerable" level. Hence, the dose actually chosen for treatment would normally be somewhat $< D_{\text{opt}}$, and the odds of its failure would likely be correspondingly greater.

This example is highly idealized, of course, but it can serve as a backdrop against which to contrast some of the issues that arise in real cases:

1. It is frequently difficult to determine in 3D, even with computerized tomography (CT), positron emission tomography (PET), or magnetic resonance imaging (MRI), the location and extent of a tumor; and it is virtually impossible to track down nearby microscopic clusters of neoplastic cells which, if untreated,

might themselves grow into tumors. The temptation is therefore to treat the region thought to harbor disease with generous spatial margins of healthy tissue; the risk of inducing complications, on the other hand, will increase with the volume of healthy tissue taken to high dose.

2. Commonly in clinical practice, one standard (but not necessarily optimal) treatment objective is to irradiate the tumor as uniformly as possible (4–10). But it would be better to sculpt a unique, patient-specific distribution of dose, non-uniform within the region of suspected disease, that takes into account the spatial variations in the density and responsiveness of tumor cells, the uncertainty in the physician's estimate of their whereabouts, and the sensitivity and criticality of the infiltrated and surrounding healthy tissues (4,5,9). The kinds of information necessary are not readily obtainable.
3. Equation 1 is meaningful only when any possible complication is of a seriousness comparable to that of death. The various complications that can arise in practice, however, differ greatly in severity. (And who is best able to assess the severity? The patient? The doctor? The insurance companies?) In any case, the gravity of any one acute or chronic complication (not only its probability of occurrence) is likely to vary with dose.
4. Equation 1 is valid only so long as $S_{\text{tumor}}(D)$ and $S_{\text{organ}}(D)$ are completely independent of one another; if the tumor reacted to intense irradiation by emitting a toxin or other biochemical that compromises the resistance of healthy organ tissue (or of the entire person) to radiation damage, for example, then equation 1 would begin to break down. Likewise, a number of organs doubtless will be at least partially irradiated in any real treatment, and some of these might behave as coupled systems, affecting one another's radiation responses; here, too, the generalization of equation 1 becomes nontrivial.
5. Even if one could apply equation 1 directly in some situations, information on $S_{\text{organ}}(D)$ is difficult to obtain from patient or radiobiological data, and little of it is yet available (11,12). Most healthy organs in a patient undergoing treatment will be irradiated nonuniformly, moreover, making the prediction of their responses considerably more complex.
6. The above points have concerned the spatial distribution of dose. The timing of treatments is equally important, adding yet another facet to the problem. It has been found that healthy tissues fare better than do tumor cells if the dose is fractionated, or spread out over an extended period; they seem to retain more ability to repair damage to DNA over time. Since the normal tissue is irradiated heterogeneously, there may be, in addition to the cellular repair between the irradiations, organ repair on a macrolevel; that is, tissue rescuing units (cells migrating from the healthier parts of the organ, which received much lower doses) may help the organ to reestablish some of its damaged functionality. Treatment is therefore delivered typically in 20–30 smaller fractions, rather with a single shot.

These and other issues, together with patient-specific factors, leave the radiation oncologist with a formidable, multidimensional treatment-planning optimization problem. The usual solution in real clinical situations is to sublimate most aspects of the probabilistic issues described above, and employ empirically established, time-tested protocols. It is in modifying these procedures, as warranted by the condition and response of the individual patient, that the acquired skills and informed judgment of a radiation oncologist are essential.

THE TREATMENT PLANNING PROCESS

The objective of curative radiotherapy is to deliver a tumoricidal dose to a clinically determined target volume (i.e., the volume of suspected gross disease plus a suitable margin), while depositing safe doses in the neighboring healthy tissues. The optimal strategy depends on the sensitivity of the tumor to radiation and its accessibility to treatment, and on the sensitivity and criticality of adjacent tissues that will be unavoidably dosed in the treatment.

For a superficial, radiosensitive lesion, it is often appropriate to employ an electron beam from a linear accelerator, which deposits most of the dose near the surface and leaves the underlying tissues unaffected.

The most common means of treatment of a deep-seated tumor in a modern medical setting is by means of highly penetrating, high-energy photons from a linac. (This article will use external photon-beam examples, but exactly the same issues are of concern for any form of radiotherapy.) Typically two to five X-ray beams cover a wide enough area (tumor region plus a *margin*) in their cross-fire region to eliminate any small clusters of cancer cells that may have extended into tissues adjacent to the primary target. The beams are arranged so as to deposit an adequate and uniform dose to the lesion, but without exceeding the tolerance levels in healthy tissues elsewhere. (It may often be helpful later to add a local *boost*, with a smaller-field photon or electron beam, or with brachytherapy, to increase the dose locally in the immediate vicinity of the primary tumor.) The resulting Compton (which predominate at high energies in soft tissue) and photoelectric interactions ionize the tissues they traverse. The production of free radicals and other molecular instabilities leads to damage to DNA and to the tissue microenvironment. That, in turn, is intended to kill the tumor cells.

In the first step in the treatment planning process itself, the radiation oncologist and treatment planning staff identify the tumor, through some combination of CT, MRI, PET, or by other means, and also the healthy critical structures to be avoided; multimodality imaging, in particular CT–PET fusion, is playing an increasingly important role in this. The oncologist then determines, based on experience and generally accepted treatment protocols, the dose that should be delivered to the tumor, and limits (to the known tolerance levels) the doses to be allowed at the normal tissues and critical organs. The radiation oncologist also

decides upon the *fractionation* schedule (the timing of the treatments.)

After the position, size, and shape of the lesion are determined, an *isocenter* (the single point at which the central axes of all treatment fields intersect) is chosen within it. Information on the patient's external shape, variations in tissue density, tumor geometry, and so on, is entered into the treatment planning computer, and one or more transverse planes are selected for the future display of the organs and superimposed topographical *isodose map*, with its contour lines of constant dose deposition.

The radiotherapy physicist and/or dosimetrist (a professional especially trained in treatment planning) then make an educated guess as to the energy, number, orientations, sizes, and weightings (the relative strengths, or relative contributions to the total dose at isocenter) of the beams, which can either be stationary or rotate through an arc, taking into account ways to block or otherwise modify parts of each one.

The computer then draws upon previously stored data to generate a representation of the spatial distribution of dose imparted by each of these beams, correcting for the patient's body shape, the lower density of lung tissue, and so on. Finally it sums, at each point of interest, the doses delivered by all the beams together, thereby producing a 3D isodose map. Although some work remains to be done for tissue inhomogeneities (especially lung and bone) for both photon and electron beams, most of the necessary dose calculation software has been developed. With our current abilities to model the interaction of high energy photon and charged particle beams with matter, and to localize healthy and pathological body structures, the computer can generate such maps with uncertainties in dose of less than a few percent.

In all likelihood, the plan can be improved by changing some of the beam parameters, so the process is repeated iteratively several times with different treatment configurations until a good (or at least clinically acceptable) dose distribution is obtained, as determined visually. (Numerical information, such as that from dose-volume histograms and the other approaches discussed below, is also being used increasingly as a guide.) This may involve the manipulation of a fair number of variables, and the plan ultimately selected may be only locally, rather than globally, optimal. Indeed, the limiting factors in the selection process may be the experience, creative skills, patience, and available time of the staff. This process is usually referred to as *forward treatment planning*.

The radiotherapy community has established a few empirical ground rules for judging treatment plans:

1. The entire tumor and a small margin should all receive at least the prescribed dose, and irradiation of the target volume should be reasonably uniform, to ensure the absence of any cold spots.
2. The highest dose isodose lines or surfaces should, when superimposed upon a set of CT or MRI scans, conform closely to the target volume.
3. The dose should decrease rapidly away from the target. Although a fair amount of healthy tissue may have to be irradiated (and may thereafter become nonfunctional), the volume of it taken to high dose should be as small as possible, and dose falloff within it should be as fast as possible.
4. Dose to most of the lung, to all of the spinal cord, and so on, must be kept below their presumed tolerance levels; the tolerance dose of a healthy organ, however, may depend on the fraction of it that is being irradiated—although for some organs, such as the spinal cord or some of the brain, the dose limit applies to any small volume of tissue—which clearly complicates matters.

In summary, dose throughout the target should achieve the prescribed level, the volume of healthy tissues should be minimized, and doses to critical structures should be below their tolerance levels.

EXAMPLE OF PLAN OPTIMIZATION

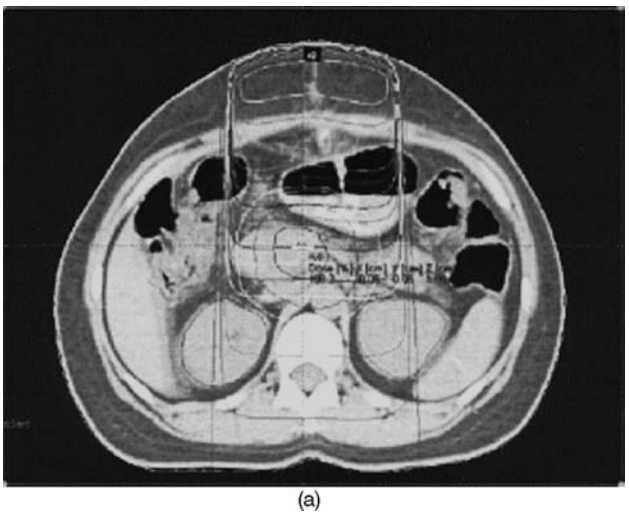
A unique treatment plan of beam configurations must be designed for each patient, based upon the specific clinical situation and requirements. Figure 2, for example, presents three possible plans and dose distributions for treating an oblong tumor of the pancreas, following surgery. The primary objectives are to take the entire tumor to a curative dose, typically 50–68 Gy, while making certain that the normal tissues of concern, the kidneys, spinal cord, and small bowel, remain functional. We shall examine a single-field plan first, then a pair of parallel-opposed fields, and finally a four-field plan.

For the single anterior field configuration, the dose across the target region ranges from 125% to 80% of the value at isocenter, taken to be at the center of the tumor, Fig. 2a. The spinal cord is at $\sim$ 65% of the treatment value, which it can tolerate, and the maximum dose of 160% is situated in the anterior subcutaneous tissue layer. The lateral aspects of the kidneys are clear of any substantial dose. But this plan is not satisfactory because of both the cold area within the tumor and the anterior hot region, even though it has good cord- and kidney-sparing characteristics.

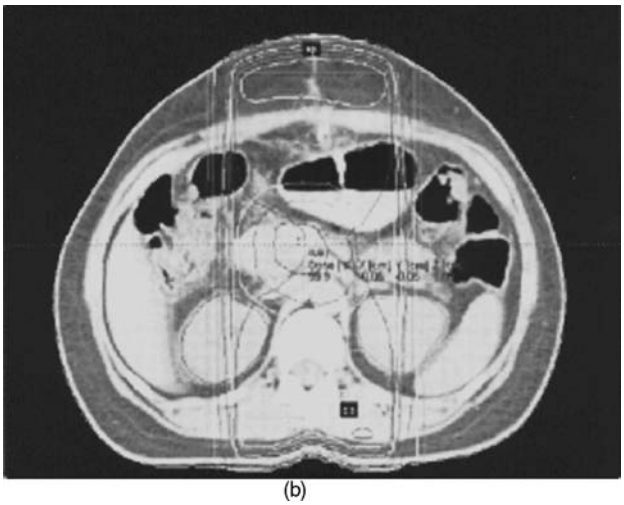
When a posterior beam is added to the anterior beam, the dose uniformity improves greatly, Fig. 2b. The dose gradient across the target region for this parallel-opposed plan is of the order of 5%, and the kidneys are still okay. The spinal cord now receives 105% of the target dose, however, which is too high.

Figure 2c is a four-field "box" plan that shows dose uniformity across the target and a high dose encompassing it. The spinal cord gets 60% of the tumor dose, and the kidneys and anterior bowel are spared reasonably well.

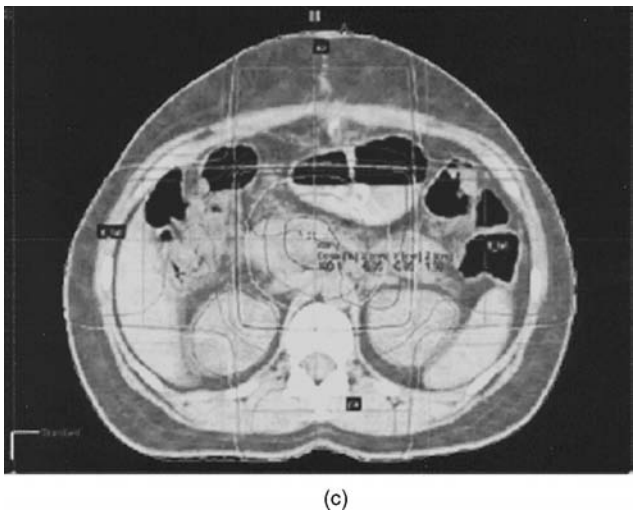
After going through such an analysis for a number of potential plans, it is usually apparent which ones work best for the patient. In the case of Fig. 2, an experienced physicist or dosimetrist would know immediately to begin with the box, but still would have to play with it to find the best weights, or relative contributions, of the four fields. For some more difficult cases, it takes considerable effort to find a set of fields, and their settings, for which the ratio of



(a)



(b)



(c)

Figure 2. As possible plans for treatment of the pancreas with 10 MV X-ray beams, the physicist and/or dosimetrist consider the isodose maps for (a) a single AP field; (b) a pair of parallel-opposed fields; (c) a standard four-field box plan that covers the tumor relatively uniformly, spares most healthy tissue, and does not exceed the tolerance levels of either the kidney or the spinal cord.

dose in tumor to dose in the surrounding tissues is high; tumor dose is reasonably uniform; relatively little normal tissue is taken to high dose; and critical organs are not in serious jeopardy.

THE NEED FOR QUANTITATIVE OPTIMIZATION TOOLS

At present, visual inspection is still the standard process for optimizing isodose plans, as is done in this example, but this can be highly time and energy consuming. Meanwhile, the patient population is growing rapidly as the baby boomers reach older ages. Also, treatment strategies are becoming significantly more complex and demanding, especially so with the advent of intensity-modulated radiation therapy (IMRT), in which the beam direction, shape, intensity, and so on, may vary continuously over time during irradiation. So efficiencies in treatment planning that once were desirable are now becoming essential.

A means of assigning reliable numerical scores to competing plans would be of great value, not only in speeding up the through-put of routine cases, but also in facilitating decisions on the more complicated ones. Such scoring capability, moreover, forms the backbone of any computer-based expert system developed for automated treatment plan decision making. Much of the heavy mathematical machinery to find a maximum in an *objective function*, or *Figure of Merit*, is already available; the major problem, rather, is to find clinically meaningful criteria by which plans may be judged, and quantitative forms in which these criteria can be expressed.

Two fundamentally different approaches have arisen to the use of computers in judging treatment plans.

The first draws upon some of the physical and geometrical rules of thumb that are commonly adopted in visual plan assessment. How constant is the dose throughout the target volume, and does the prescribed-dose isodose surface conform closely to its surface? How much dose is being deposited in healthy tissues, and is the tolerance level of any critical organ exceeded? Such criteria have the benefits of being easily understood and clinically widely accepted, and of being easy to implement in an automated optimization system. On the downside, it is not clear that the plan with the most uniform tumor dose is necessarily most likely to effect a cure, or that the one with the lowest average dose to healthy tissue will most probably avoid complications. The obvious physically-based criteria, either individually or in combination, do not necessarily reveal the strategy most likely to work.

The second is somewhat more abstract, and more removed from traditional clinical practice. It attempts to draw upon various kinds of available dose-response information on normal tissues to estimate the probabilities that any particular dose distribution in an organ will lead to complications (13–15). This biologically and probabilistically based mathematical modeling work is still in early stages of development, and obtaining and interpreting the necessary data on radiation damage is difficult. Nonetheless, some researchers feel that the approach has perhaps greater potential for eventual success.

Either way, three aspects to the optimization process must be addressed:

1. First, it is necessary to decide what would constitute an ideal treatment. A plan judged to be nearly perfect according to one criterion might well be far from it according to another; the best overall optimization criterion to be found might, indeed, be some balanced amalgam of several others.
2. One must then find a quantitative measure of the deviation of any computer-generated plan from the ideal.
3. Finally, an algorithm is needed to search out quickly the plan that deviates least from the ideal, as beam weights, field sizes, and other parameters are varied.

OPTIMIZATION USING PHYSICAL/GEOMETRICAL CRITERIA

Several traditional criteria for the visual optimization of plans are in common use. One would like the prescription-dose isodose surface to wrap as closely as possible around the target volume. Also, some have argued, the dose throughout the tumor should be uniform. It is desirable to keep the amount of dose to healthy tissues to a minimum, and doses to all critical structures must be maintained below their tolerance levels. All of these ideas, it turns out, have been easy to incorporate into computer-based optimization programs.

For all of what follows, we shall partition the region of interest into J small voxels, all of volume Δv, and assume that a particular configuration of beams delivers the dose D_j to the jth such voxel. For simplicity, we shall also consider situations in which all beam parameters have already been chosen by the treatment planner, with the exception of the best values of the relative weightings, w_k, of the K beams (w_k refers to the relative dose at isocenter produced by the kth field).

For any possible treatment plan, the dose in the jth voxel is now fixed completely by

$$D_j = \sum_{k}^{K} d_{jk}\, w_k \tag{2}$$

where $w_k \geq 0$. For each particular configuration of beams, the matrix of elements d_{jk} determines the dose deposited at the jth voxel by the kth field. A standard treatment planning program is designed to calculate such matrices. Changing the energy, orientation, blocking, and so on, of the beams, however, will result in a revised matrix of d_{jk} values.

Healthy Tissue Integral Dose

The simplest single measure of the irradiation of a block of healthy tissue is the *integral dose* (ID). Integral dose is defined through the function

$$f_{ID} = \sum_{j}^{J} D_j\, \Delta v \quad \text{(healthy tissue)} \tag{3a}$$

where the sum is confined to healthy tissues outside the target volume. Equation 3a serves as the objective function that is to be minimized, by manipulating various beam parameters, when integral dose is selected as the optimization criterion. Note that apart from a factor of $J\Delta v$, the integral dose is just the average voxel dose. By equation 2, the integral dose may be expressed as

$$f_{ID} = \sum_{j}^{J}\sum_{k}^{K} d_{jk}\, w_k\, \Delta v \quad \text{(healthy tissue)} \tag{3b}$$

It is intended, moreover, that the distribution deliver at least a tumoricidal dose, $D_{\text{prescribed}}$,

$$D_l \geq D_{\text{prescribed}} \quad \text{(tumor)} \tag{4a}$$

at previously selected points within the target volume. Similar constraints,

$$D_{l'} \leq D_{\text{tolerance}} \quad (\text{point } l' \text{ in critical organ}) \tag{4b}$$

protect any critical structure, such as the spinal cord.

The task now is to find those weights that deliver the prescribed dose to the target volume, but minimize the integral dose objective function, while keeping irradiation of certain voxels below specified limits. The objective function is linear in the weights w_k, the variables of interest, and equations 3 and 4 define a routine linear programming problem whose solution can be obtained in seconds with the Simplex method.

Despite the mathematical attractiveness of the integral dose objective function, its application does not necessarily lead to results in accord with clinical experience. Taking an entire organ to 30 Gy, for example, yields the same f_{ID} as does taking half of it to 60 Gy; but the two irradiation schemes could lead to vastly different results if the organ's tissues exhibit a response threshold at 45 Gy. This may serve as a reminder of a fundamental, but sometimes overlooked truth: The availability of powerful mathematical optimization algorithms does not ensure the clinical value of an objective function upon which they can be employed.

Optimal Point Doses

In perhaps the most widely explored and successful of the physical–geometric approaches, the physician prescribes the doses to be delivered at a designated set of points within the patient; some of these, in the tumor and in critical organs, act as rigid constraints. The treatment planning system then uses available mathematical techniques to find that plan (i.e., set of beam sizes, angles, relative weightings, etc.) for which the calculated doses at these points come closest, on the whole, to the prescribed values. Points might be selected, for example, on the tumor border, to force the prescribed-dose isodose surface to envelope it closely, while also satisfying the constraint requirements. Alternatively, they might be placed throughout the target volume, to maximize tumor dose uniformity.

Again, let the variable parameters be the relative beam weightings, and now suppose that the objective is to maximize tumor dose uniformity (TDU). The M prescription

points will be scattered throughout the tumor, and, by our criterion, doses there should all be nearly the same. The deviation of any plan from the ideal may be expressed as

$$f_{TDU} = \sum_{m}^{M} (D_m - D_{\text{av}})^2 \tag{5}$$

where the D_m provided by equation 2 are functions of the w_j values, and D_{av} is the average dose for the selected points.

Similar objective functions can be devised to force the prescribed isodose line to lie as near as possible to the target volume boundary, or to cause certain discrete anatomic regions to receive prescribed doses. Methods such as quadratic programming and constrained least-squares algorithms are available for rapidly finding extrema in functions of the form of equation 5, subject to constraints like equations 4. While optimization with such objective functions may not alone yield clinically appropriate plans, they can often provide starting points for further manual searching.

Developments in computer-controlled radiation treatment have stimulated renewed interest in such approaches. Computers have long been employed to monitor a linear accelerator's treatment parameters, verifying and recording every field's gantry angle, jaw setting, dose delivered, and so on. More recently, with the growing adoption of IMRT systems, dedicated computers are controlling variables such as collimator configuration, dose rate, gantry angle, table movements in real time. While such flexibility allows the implementation of highly refined treatment plans, the associated planning process can be very cumbersome and time consuming. The problem would be totally unmanageable were it not for the availability of various objective functions and search programs (known as inverse planning algorithms) to help in finding suitable time-dependent linac output parameters (16,17).

References that describe optimization by way of physical–geometric criteria may be found in Ref. 18.

BIOLOGICAL–PROBABILISTIC APPROACHES TO TREATMENT PLAN OPTIMIZATION

The physical–geometric approaches just discussed could perhaps best be called pragmatic. They reflect mathematically some of the conventional clinical criteria for plan optimization; as such, they can discriminate among plans only as effectively as do the standard clinical guidelines that they mimic. They cannot choose, for example, between a treatment that irradiates a sizable portion of lung to a dose near its tolerance level and another treatment that takes even more lung to a somewhat lower dose, other aspects of the case being equal.

The biological–probabilistic methods, admittedly in their infancy, attempt to address directly the fundamental question: How probable is it that some given spatial and temporal distribution of dose will eradicate or control a tumor, and do so without resulting in severe complications? Asking this returns us to the thinking illustrated in Fig. 1.

Limited theoretical progress has been made on several fronts. Among them, we can derive some dose–response curves of the general form of S_{tumor} and S_{organ} from more basic radiobiological principles. We have elementary ideas on how to use such information to estimate the probability of complications arising in a healthy organ irradiated nonuniformly, and we can crudely model the radiation response of a tumor.

Shape of a Tumor Dose-Response Curve, S_{tumor}

Of an initial cluster of $n(0)$ of any kind of cells, only $n(D)$ of them will remain viable after receiving an initial dose of D Gy; the rest will die before or while attempting to divide. If an incremental dose, ΔD, soon follows (before repair or repopulation occurs), the number of surviving cells diminishes by an additional

$$\Delta n = -c(D)n(D)\Delta D \tag{6a}$$

equivalently,

$$c(D) = -\Delta n / n(D)\Delta D \tag{6b}$$

is the probability per unit dose of inactivating a cell that has already been given D. There is no physics or real biology in the picture yet—everything so far is just bookkeeping.

For the purposes of this article, the discussion will be greatly simplified by restricting the biophysics being considered. Although low Linear Energy Transfer (LET) radiations, namely high-energy photons and electrons, are almost always employed in practice, here the treatment is by way of a beam of energetic, heavy charged particles, such as cyclotron-produced protons.Every proton leaves a dense trail of ionization in a cell it traverses, and damage to any DNA molecule it passes sufficiently close to is likely to involve irreparable breaks in both sugar–phosphate strands (19). Cells in which this occurs will probably be completely inactivated by a single hit of an incident particle; most cells more distant from its track, on the other hand, will be oblivious to its passage. Cell killing in this situation is thus an all-or-nothing affair. Surviving cells retain no memory of earlier irradiations, and $c(D)$ becomes a constant, commonly called $1/D_0$, independent of dose.

In addition, it is assumed that the tumor cells of concern do not communicate with one another in any significant way; that is, radiation damage in one voxel does not notably affect (by releasing toxins, cutting off the influx of nutrients, etc., as in "bystander" phenomena) the dose-response characteristics of an adjacent voxel.

After replacing $c(D)$ in equation 6 with $1/D_0$, the associated differential equation

$$dn/dD = -(1/D_0)n(D) \tag{7a}$$

can be integrated to yield

$$n(D)/n(0) = e^{-D/D_0} \tag{7b}$$

This is the fraction of cells, either in the tumor or in any small part of it, that survives irradiation to dose D. It may be viewed equally well as the survival probability for an individual tumor cell. Expressing $n(0)$ as the product (ρV), where ρ is the tumor cell density and V is the tumor volume, this becomes

$$n(D) = n(0)e^{-D/D_0} = (\rho V)e^{-D/D_0} \tag{7c}$$

This is the number of cells statistically expected to survive when the entire tumor receives a dose of D. All the relevant radiobiology and radiation physics reside within D_0, which can be a complex function of any number of interesting parameters that may, or may not, be understood. (With X rays or electrons, single-strand DNA breaks occur; some degree of repair can take place, and $c(D)$ is not independent of D. Integration of equation 7a then leads to a cell survival curve with a shoulder, rather than equation 7c.)

It is believed that a tumor can repopulate from a single clonogenic cell. The desired outcome of a curative treatment must therefore be the total eradication of all tumor cells. That is, the probability that *no* cells survive must be high. Poisson (or binomial) statistics provides a way of assessing that probability: if the average number of times some event takes place in a situation of interest is μ, then the probability of exactly zero such events occurring, $P_\mu(0)$, is $e^{-\mu}$,

$$P_\mu(0) = e^{-\mu} \quad (\mu = \text{average number of events}) \tag{8a}$$

In the present case, the average number of cells expected still to be clonogenic after an irradiation of D is given by equation 7c. From equation 8a, then, the probability, $P_D(0)$, that there will be *no* tumor cells left viable is

$$S_{\text{tumor}}(D) = P_D(0) = e^{-(\rho V)e^{-D/D_0}} \tag{8b}$$

This can be rewritten in terms of the geometrical characteristics of the curve, namely D_{50} the dose resulting in 50% probability of complication, S_{organ} 0.5 (see also Ref. 20), and slope, γ_{50}:

$$S_{\text{tumor}}(D) = (\tfrac{1}{2})^{e^{[2\gamma_{50}(1-D/D_{50})/\ln 2]}} \tag{8c}$$

This fits clinical data sufficiently well (21,22); indeed, a number of tumor cell populations characterized by different parameter values result in almost indistinguishable sets of dose response curves, in part because of strong correlations between the parameters. The D_{50} and γ_{50} values for a number of sites have been published (21).

Equations 8b and 8c provide, and Fig. 1 displays, one particular form for the probability, $S_{\text{tumor}}(D)$ of equation 1, that the patient will survive the disease. It increases sigmoidally with dose, as certainly expected. It also decreases exponentially with tumor size; equation 8b indicates that the doses required to eradicate (with equal probabilities of success) two otherwise identical tumors that differ only in pre-irradiation volume are related as

$$D_2 = D_1 + D_0 \ln (V_2/V_1) \tag{8d}$$

This model is simple but, not too surprisingly, it agrees with the clinical finding that larger tumors require more dose to achieve a cure than do smaller tumors of the same histological type.

Equation 8 describe the irradiation characteristics of a tumor exposed uniformly. The result is of legitimate clinical interest since it is common practice to attempt to impart a fairly flat dose across the target volume. In the case of non-uniform irradiation, the tumor can be viewed as consisting of J voxels that receive doses D_j, with control probabilities of $s(D_j)$, again assuming their statistical independence. The parameter S_{tumor} is then given by

$$S_{\text{tumor}} = \prod_j^J s(D_j) \tag{8e}$$

One of the main characteristics of a tumor is its comparatively fast repopulation rate. To account for this in conjunction with the Poisson (or binomial) distribution, several authors in the 1980s proposed setting the initial number of clonogens to $n(0)$, after which $n(t) = n(0)e^{+\lambda t}$, where λ and t are the rate constant and repopulation time, respectively (23–26). This approach predicts a $S_{\text{tumor}}(t)$ that always tends to zero for large post-treatment times, which is incorrect. Later, taking clonogen repopulation between fractions (27–30) into account lead to the Zaider–Minerbo model (31) of $S_{\text{tumor}}(t)$, which is applicable for any temporal treatment protocol. An expression for $S_{\text{tumor}}(t)$ with different time intervals between consecutive irradiation fractions and with varying cell survival probability per fraction was obtained by Stavreva et al. (32) based on the Zaider–Minerbo approach. Animal experiments support the validity of the Zaider–Minerbo approach (32).

Other approaches to the determination of S_{tumor} are discussed in the literature (33–36).

Integral Response for a Healthy Organ

Several ways have been devised to address the non-uniform irradiation of healthy tissues. One of these is an extension of the integral dose idea, introduced in equations 3 and 4. It does not focus on the spatially varying dose distribution within a tissue *per se*, as do the physical–geometric approaches, but rather on the local biological response that the dose elicits. Variation on the approach, of a range of levels of complexity, have been discussed by a number of researchers (18,37–41).

Imagine an organ or biological compartment that produces some physiologically important substance (such as a critical enzyme or hormone) or that performs an important task (like phagocytosis or gas exchange). If too many of its cells are inactivated by irradiation, then the organ cannot do its job adequately, and the organism runs into trouble. As before, mathematically partition the organ into J small volume elements of size Δv, and assume that the radiation response of any such voxel (or of the cells within it) is nearly independent of its neighbor's response. If, moreover, high LET radiation is again involved, the dose–response relationship is of the form of equation 7b, where D_0 contains all the important biophysics.

Of the entire organ, only the fraction

$$v = (1/J\,\Delta v)\sum_j^J e^{-D_j/D_0}\,\Delta v \tag{9a}$$

of its tissue will remain functional, where $(J\Delta v)$ is its volume. (The parameter $(1 - v)$ thus provides a direct measure of the amount of radiation damage to the organ.) For the case of nearly uniform irradiation to the level D,

v reduces to the

$$n(D)/n(0) = e^{-D/D_0} \tag{9b}$$

of equation 7b. This observation, along with the form of equation 9a, suggests that v be viewed as a generalized, or spatially averaged, dose–response parameter, prompting adoption of the name "integral response" (18,34).

The probability that the patient will escape serious complications, S_{organ}, is a nondecreasing function of the relative amount of organ that remains intact,

$$S_{\text{organ}} = S_{\text{organ}}(v) \tag{9c}$$

the shape of which must be obtained experimentally or by other means (39,40). The parameter $S_{\text{organ}}(v)$ and equations 2 and 9a together define the integral response (IR) objective function, f_{IR}, and the optimization problem is bound by the constraints of equations 3 and 4.

If Q radiobiologically independent organs are irradiated, and if the possible complications are all of comparable severity, then the overall survival probability may be given by the product

$$S_{Q\,\text{organs}} = \prod_{q}^{Q} S_{\text{organ}\,q} \tag{9d}$$

Automated treatment plan optimization may then be carried out, in principle, with a combination of equations 2, 4, and 9.

Another possible method of handling non-uniform irradiation of such an organ is based on its N-step dose–volume histogram. The idea is to reduce it in such a fashion as to yield a revised histogram that corresponds to the same complication probability, but that contains only an $(N - 1)$ steps; this calculation is repeated $(N - 1)$ times, until there is left a single-step histogram, the S_{organ} of which can be obtained from experiment or clinical observation.

Several algorithms have been proposed for carrying out the histogram-reduction process (42–47).

Healthy Organ Composed of Separately Critical Voxels

The integral response objective function is based on the effectiveness of operation of a healthy organ taken as a unit (48). A radically different approach is needed for an organ, such as the spinal cord or certain regions of the brain, that behaves like a chain or computer program, in which serious complications in any single small part can spell disaster for the whole.

Once again, consider an organ made up of J radiobiologically independent voxels. Let $s(D_j)$ refer to the probability that the organ will suffer no serious complications when the jth voxel is taken to dose level D_j and the rest is left unirradiated. If the entire organ is to escape damage, each of its J parts must do so separately, and

$$S_{\text{organ}} = \prod_{j}^{J} s(D_j) \tag{10a}$$

If the small-volume tolerance dose is exceeded in even a single voxel, then the objective function S_{organ} can become perilously low. If the organ is irradiated uniformly, incidentally, equation 10a reduces to

$$S_{\text{organ}} = [s(D)]^J \tag{10b}$$

The methods just described for dealing with nonuniform irradiation make use of the assumption of the radiobiological independence of adjacent voxels. While this assumption clearly is not valid for most organs, nor perhaps for tumors, it may apply to some (e.g., the blood pool). More importantly, models built upon it may serve as jumping-off points for the development of more realistic pictures.

Other approaches to the determination of S_{organ} are discussed in the literature (49–52).

CONCLUSION

The radiation response of a tissue depends in an extremely complex way on a number of parameters, some of which the radiation oncologist can control directly, some indirectly, and some not at all. One can choose the modality (X rays, electrons, gamma-rays, protons, neutrons); the volume to be irradiated; the dose per fraction and number of fractions; the administration of response-modifying drugs; and the spatial dose distribution in healthy tissues. Some of the generally uncontrollable (or weakly controllable) parameters are radiosensitivity differences of different constituent parts of a tissue; tissue concentrations of oxygen, drugs, toxins, and other compounds; numbers of cells in each portion of the mitotic cycle; differential cell population kinetics; and repair of radiation-induced cellular injury. Manipulation of directly controllable variables may lead to changes in others: reoxygenation; differential cell cycle phase redistribution; differential recruitment of proliferative cells; and differential repair.

In view of this complexity, it is not clear how effective empirical or *ab initio* mathematical modeling can be in providing clinically useful descriptions of processes as intricate as the kinetics of irradiated tissues. At the present time, the subject is largely of academic interest. But itis to be expected that over the next several decades, computer-based expert systems will continue to spread throughout all of medicine. This development, together with the increasing automation of radiation delivery systems, will doubtless cause the search for quantitative methods of treatment plan optimization to expand.

BIBLIOGRAPHY

1. Wolbarst AB. Physics of Radiology. 2nd ed. Madison (WI): Medical Physics Publishing; 2005.
2. Lawrence TS, Ten Haken RK, Kessler ML, Robertson JM, Lyman JT, Lavigne ML, Brown MB, Duross DJ, Andrews JC, Ensminger Wd, Lichter AS. The use of 3-D dose volume analysis to predict radiation hepatitis. Int J Radiat Oncol Biol Phys 1992;23:781–788.
3. Jackson A, Ten Haken RK, Robertson JM, Kessler ML, Kutcher GJ, Lawrence TS. Analysis of clinical complication data for radiation hepatitis using a parallel architecture model. Int J Radiat Oncol Biol Phys 1995;31:883–891.

4. Fischer JJ. Theoretical considerations in the optimisation of dose distribution in radiation therapy. Br J Radiol 1969;42: 925–930.
5. Brahme A, Agren AK. Optimal dose distribution for eradication of heterogeneous tumours. Acta Oncol 1987;26:377–385.
6. Webb S, Nahum AE. A model for calculating tumour control probability in radiotherapy including the effects of inhomogeneous distributions of dose and clonogenic cell density. Phys Med Biol 1993;38:653–666.
7. Webb S, Evans PM, Swindell W, Deasy JO. A proof that uniform dose gives the greatest TCP for fixed integral dose in the planning target volume. Phys Med Biol 1994;39:2091–2098.
8. Stavreva NA, Stavrev PV, Round WH. A variational approach to the problem of optimizing the radiation dose distribution in tumours. Australas Phys Eng Sci Med 1996;19:9–11.
9. Stavreva NA, Stavrev PV, Round WH. A mathematical approach to optimizing the radiation dose distribution in heterogeneous tumours. Acta Oncol 1996;35:727–732.
10. Ebert MA, Hoban PW. Some characteristics of tumour control probability for heterogeneous tumours. Phys Med Biol 1996;41:2125–2133.
11. Sanchez-Nieto B, Nahum AE. Bioplan. Software for the biological evaluation of radiotherapy treatment plans. Med Dosim 2000;25:71–76.
12. Warkentin B, Stavreva N, Stavrev P, Field C, Fallone BG. A TCP-NTCP estimation module using DVHs and known radiobiological models and parameter sets. J Appl Clin Med Phys 2004;5:1–14.
13. Brahme A. Biologically based treatment planning. Acta Oncol 1999;38(13 Suppl): 61–68.
14. De Gersem WR, Derycke S, De Wagter C, De Neve WC. Optimization of beam weights in conformal radiotherapy planning of stage III non-small cell lung cancer: Effects on therapeutic ratio. Int J Radiat Oncol Biol Phys 2000;47:255–260.
15. Stavrev P, Hristov D, Warkentin B, Sham E, Stavreva N, Fallone BG. Inverse treatment planning by physically constrained minimization of a biological objective function. Med Phys 2003;30:2948–2958.
16. Brahme A, Roos JE, Lax I. Solution of an integral equation encountered in rotation therapy. Phys Med Biol 1982;27:1221–1229.
17. Webb S, Convery D, Evans PM. Inverse planning with constraints to generate smoothed intensity-modulated beams. Phys Med Biol 1998;43:2785–2794.
18. Wolbarst AB, Chin LM, Svensson GK. Optimization of radiation therapy: Integral-response of a model biological system. Int J Radiat Oncol Biol Phys 1982;8:1761–1769.
19. Hall E. Radiobiology for the Radiologist. 5th ed. Baltimore: Lippincott Williams & Wilkins; 2000.
20. Stavreva N, Stavrev P, Warkentin B, Fallone BG. Derivation of the expressions for gamma50 and D50 for different individual TCP and NTCP models. Phys Med Biol 2002;7:3591–3604.
21. Okunieff P, Morgan D, Niemierko A, Suit HD. Radiation dose-response of human tumors. Int J Radiat Oncol Biol Phys 1995;32:1227–1237.
22. Goitein M, Niemierko A, Okunieff P. The probability of controlling an inhomogeneously irradiated tumor: A stratagem for improving tumor control through partial tumor boosting. 19th L H Gray Conference: Quantitative Imaging in Oncology. Newcastle (UK): 1995; p 25–39.
23. Thames HD, Rozell ME, Tucker SL, Ang KK, Fisher DR, Travis EL. Direct analysis of quantal radiation response data. Int J Radiat Oncol Biol Phys 1986;49:999–1009.
24. Tucker SL, Travis EL. Comments on a time-dependent version of the linear-quadratic model. Radiother Oncol 1990;18: 155–163.
25. van de Geijn J. Incorporating the time factor into the linear-quadratic model. Br J Radiol 1989;62:296–298.
26. Yaes RJ. Linear-quadratic model isoeffect relations for proliferating tumor-cells for treatment with multiple fractions per day. Int J Radiat Oncol Biol Phys 1989;17:901–905.
27. Tucker SL, Thames HD, Taylor JM. How well is the probability of tumor cure after fractionated irradiation described by Poisson statistics? Radiat Res 1990;124:273–282.
28. Yakovlev AY. Comments on the distribution of clonogens in irradiated tumors. Radiat Res 1993;134:117–120.
29. Kendal WS. A closed-form description of tumour control with fractionated radiotherapy and repopulation. Int J Radiat Oncol Biol Phys 1998;73:207–210.
30. Tucker SL, Taylor JM. Improved models of tumour cure. Int J Radiat Biol 1996;70:539–553.
31. Zaider M, Minerbo GN. Tumour control probability: A formulation applicable to any temporal protocol of dose delivery. Phys Med Biol 2000;45:279–293.
32. Stavreva N, Stavrev P, Warkentin B, Fallone BG. Investigating the effect of cell repopulation on the tumor response to fractionated external radiotherapy. Med Phys 2003;30:735–742.
33. Wolbarst AB, Sternick ES, Curran BH, Dritschilo A. Optimized radiotherapy treatment planning using the complication probability factor (CPF). Int J Radiat Oncol Biol Phys 1980;6:723–728.
34. Roberts SA, Hendry JH. A realistic closed-form radiobiological model of clinical tumor-control data incorporating intertumor heterogeneity. Int J Radiat Oncol Biol Phys 1998;41:689–699.
35. Fenwick JD. Predicting the radiation control probability of heterogeneous tumour ensembles: Data analysis and parameter estimation using a closed-form expression. Phys Med Biol 1998;43:2159–2178.
36. Brenner DJ. Dose, volume, and tumor-control predictions in radiotherapy. Int J Radiat Oncol Biol Phys 1993;26:171–179.
37. Kallman P, Agren A, Brahme A. Tumour and normal tissue responses to fractionated non-uniform dose delivery. Int J Radiat Biol 1992;62:249–262.
38. Withers HR, Taylor JM, Maciejewski B. Treatment volume and tissue tolerance. Int J Radiat Oncol Biol Phys 1988; 14:751–759.
39. Jackson A, Kutcher GJ, Yorke ED. Probability of radiation-induced complications for normal tissues with parallel architecture subject to non-uniform irradiation. Med Phys 1993; 20:613–625.
40. Niemierko A, Goitein M. Modeling of normal tissue response to radiation: The critical volume model. Int J Radiat Oncol Biol Phys 1993;25:135–145.
41. Olsen DR, Kambestad BK, Kristoffersen DT. Calculation of radiation induced complication probabilities for brain, liver and kidney, and the use of a reliability model to estimate critical volume fractions. Br J Radiol 1994;67:1218–1225.
42. Lyman JT, Wolbarst AB. Optimization of radiation therapy, III: A method of assessing complication probabilities from dose-volume histograms. Int J Radiat Oncol Biol Phys 1987;13:103–109.
43. Lyman JT, Wolbarst AB. Optimization of radiation therapy, IV: A dose-volume histogram reduction algorithm. Int J Radiat Oncol Biol Phys 1989;17:433–436.
44. Kutcher GJ, Burman C. Calculation of complication probability factors for non-uniform normal tissue irradiation: The effective volume method. Int J Radiat Oncol Biol Phys 1989;16:1623–1630.
45. Kutcher GJ, Burman C, Brewster L, Goitein M, Mohan R. Histogram reduction method for calculating complication probabilities for three-dimensional treatment planning evaluations. Int J Radiat Oncol Biol Phys 1991;21:137–146.
46. Niemierko A, Goitein M. Calculation of normal tissue complication probability and dose-volume histogram reduction

schemes for tissues with a critical element architecture. Radiother Oncol 1991;20:166–176.
47. Niemierko A. A generalized concept of equivalent uniform dose. 41th AAPM Annual Meeting, Nashville, 24–29 July, 1999. Med Phys 1999;26:1100.
48. Wolbarst AB. Optimization of radiation therapy II: The critical-voxel model. Int J Radiat Oncol Biol Phys 1984;10:741–745.
49. Lyman JT. Complication probability as assessed from dose-volume histograms. Radiat Res (Suppl) 1985;8:S13–S19.
50. Yaes RJ, Kalend A. Local stem cell depletion model for radiation myelitis. Int J Radiat Oncol Biol Phys 1988;14: 1247–1259.
51. Stavrev P, Stavreva N, Niemierko A, Goitein M. Generalization of a model of tissue response to radiation based on the idea of functional subunits and binomial statistics. Phys Med Biol 2001;46:1501–1518.
52. Stavreva N, Niemierko A, Stavrev P, Goitein M. Modeling the dose-volume response of the spinal cord, based on the idea of damage to contiguous functional subunits. Int J Radiat Biol 2001;77:695–702.

See also RADIATION DOSE PLANNING, COMPUTER-AIDED; RADIATION THERAPY TREATMENT PLANNING, MONTE CARLO CALCULATIONS IN; RADIATION THERAPY, QUALITY ASSURANCE IN; RADIOLOGY INFORMATION SYSTEMS.

RANGE OF MOTION.

See REHABILITATION AND MUSCLE TESTING.

RECORDERS, GRAPHIC

HERMAN VERMARIEN
Vrije Universiteit Brussel
Brussels, Belgium

INTRODUCTION

Graphic recorders, as considered here, are essentially measuring instruments that produce in real-time graphic representations of biomedical signals, in the form of a permanent document intended for visual inspection. Recorded data are thus fixed on a two-dimensional (2D) medium which can simply be called "paper" (although the material applied may differ from ordinary writing paper); A so-called hard copy of the information is generated. Equivalent names are paper recorder, direct writing recorder, plotter, chart recorder, and strip chart recorder (if long strips of paper are used); In some cases, the more general term hard-copy unit is also applied. The nomenclature oscillograph is sometimes used (in correspondence with the other display instrument, the oscilloscope). By the aspect of visual inspection the graphic recorder is distinguished from other data storage devices (e.g., magnetic or optical disk, solid-state memory), whereas in the property of permanence graphic recording differs from visual monitoring as realized by the oscilloscope or the computer screen. Real-time paper recordings have the benefit of direct visual access to signal information, allow immediate examination (and re-examination) of trends (as long strips of paper can be used), present better graphic quality than most screens and can be used as a document for scientific evidence.

The graphic records are inspected through the human visual waveform recognition abilities; Moreover, discrete parameter values can be derived that are further evaluated. For the latter purpose, the measuring ruler still continues to be an intensively used instrument. Values for the physical variables presented, including time, can be derived, usually by comparing different related biomedical signals that were simultaneously recorded. For this purpose most biomedical recorders are multichannel recorders, which are able to process a set of signals. As the information has to be stored, the content of the records cannot be limited to the naked signal tracings; It is obviously essential that additional information concerning the recorded signals (identification, calibration, etc.) and the experimental circumstances (date, subject, experiment description, stimulus type, etc.) be kept in file, by preference directly on the recorded charts.

Two primary aspects describing the performance of the graphic recorder are (1) the properties of the instrument as a recorder of information, that is, measuring accuracy and the ability to display additional information, and (2) the quality of graphics, implying the clearness of the tracings and the overall graphic layout (e.g., including identification of the curves by color difference) in relation to ergonomics in visual examination. Also, the stability in time (permanence) of these graphic qualities can be catalogued under item 2. Secondary aspects (but not necessarily less important to the user) are ease of control of the apparatus (adjustments, calibration, adding alphanumeric information, remote control, computer connection and communication possibilities, paper loading), input amplification and signal conditioning facilities, unavoidable maintenance (ink, pens, mechanical parts), costs (the cost not only of the apparatus, but also of maintenance and, not negligibly, paper), and service life.

In early physiological experiments on mechanical functioning, such as muscle contraction, recording was performed with the aid of a directly coupled writing stylus; Suspended with minimal weight and friction, the stylus arm was applied as a cantilever, one end connected to the contracting muscle, the other end provided with a tip writing on a rotating drum. A directly coupled method is found in the spirograph: The low weight air cavity of the spirometer, which moves up and down as the subject expires and inspires, is mechanically linked to a pen writing on calibrated paper that moves at constant velocity, thus generating a trace of the course of lung volume versus time. Nevertheless, most recorders use a transducer that converts electrical information (signal voltage or current) into mechanical data, more specifically, position on the paper. This can be achieved by moving the writing element to the specific position (analog transducers) or by activating the correct point in a large array of stationary, equally spaced writing elements (digital transducers).

At present, sophisticated graphic recorders that offer a broad variety of possibilities to the operator are available. The measuring quality of transducers has continuously been improved and, as digital techniques have been applied, the possibilities of automatic control and addition

of alphanumeric information have been largely extended. A typical example is automatic recording of standard ECG derivations (frontal bipolar and unipolar, and precordial derivations) with calibration, identification of the curves, and determination of typical parameters (such as the duration of specific electrocardiographic intervals) and generation of diagnostics. Nowadays, the borderline between graphic recorders, digital oscilloscopes, data-acquisition systems, and PC-based virtual measuring instruments become less clear. There is a decreasing interest in analog recorders except for the multipen recorders having the benefit of simplicity, low price, and excellent identifiability of curves by the use of different colors. They are used in laboratory applications (and process monitoring) and handle slow varying signals. Digital recorders are generally provided with an LCD screen allowing monitoring with different colored tracings. Moreover, signal processing and data extraction, storage of data (and replay of original or processed signals), computer connection for handling data and control of recording settings are possible. Measurements where immediate visualization of the tracings on a long strip of paper is not required and data are to be digitally analyzed and stored, can be performed by PC with virtual instrument software and printer; Nowadays the majority of data collection is accomplished using digital PC approaches.

In giving an overview of graphic recorder function, one inevitably refers to technology of transducers. Especially in analog recorders the capacities and the limitations of the transducer is of major importance to the quality of the complete recorder. A number of mechanisms were applied in commercially available devices, but in many applications the analog types have been replaced by the digital type or by the PC system. Recorders can be called special purpose when built within a complete measuring system [e.g., an electrocardiograph (ECG), an electroencephalograph (EEG)], or they can be general purpose. The latter may be provided with specific signal conditioning modules to be connected to biophysical sensors (e.g., manometers for blood pressure measurements, Doppler probes for blood velocity assessment, thermal sensors for temperature recording) or, even more generally, to standard amplifiers allowing amplification levels, direct current (dc) adjustment, filtering, and so on. In most applications the graphic transducer, not the electronic signal conditioning hardware, is the crucial stage in the measuring chain. This article is intended to cover primarily the essentials of graphic recording, focusing on the principal aspects mentioned before: recording of information and graphic quality. For extensive practical details on recorders and signal conditioning modules, waveform monitoring, digital storage, processing and communication facilities, the manufacturers' data sheets should be consulted.

FUNDAMENTAL ASPECTS OF GRAPHIC RECORDING

Graphics: Images or Tracings

As the paper used for graphic recording is a 2D medium, two coordinates, x and y, can be defined: y represents the ordinate corresponding to paper width; x is the abscissa, corresponds to paper length. If an exception is made for the strip chart recorder, the distinction between x and y is merely a matter of definition. In the most general case an image can be presented on the paper: At each point (x, y) a gray scale or color scale value is displayed. For example, in speech analysis, the so-called spectrograph displays amplitude in the form of a gray scale value versus frequency (y) and time (x). Such types are image recorders. Nevertheless, in most biomedical recordings the content of z is limited to some discrete values, sometimes two (black and white or, more generally, marked and not marked), or more, in case different colors or marking intensities are applied. As such, the image is reduced to a set of tracings. This implies that for each value of x a set of y values is marked. Marking intensity (gray or color scale) and line thickness are insignificant with respect to signal representation and can be beneficially used for trace distinction and identification in relation to the quality of graphic layout.

In most applications, x corresponds to time and y to the physical magnitudes recorded (a t–y recorder or y–t recorder). Most attention will be paid to these types. The number of y magnitudes then stands for the number of channels (signal inputs) of the recorder. If the recorder is designed for an arbitrary abscissa input, the indication x–y recorder is used. The ideal multichannel t–y recorder produces a graphic representation of a set of time signals (Fig. 1). The effect of time is generally originated by pulling a strip of paper at constant velocity; A site has to be marked to indicate the time reference. Sensitivities and reference (e.g., zero) levels have to be known for each channel. Paper velocity, time reference, sensitivities, and reference levels are important scale factors that, apart from a few exceptions, are absolutely necessary when tracings are examined. Indeed, in specific applications some of these parameters are not required as they carry no information. A typical example is the zero level in electrophysiological measurements using skin electrodes, such as ECG and EEG. In these signals, the dc level is not significant as it is not generated by the physiological source (the heart in ECG, the brain in EEG); Moreover, they are usually high pass filtered to eliminate baseline disturbance from electrode-tissue interface potentials. Evidently, the same does not apply for such information as blood pressure and blood velocity, where the zero level is indispensable for evaluation. An accurate time reference is required only if the recorded signal is of the evoked type, that is, a response to a specific stimulus. For the ideal apparatus the recording parameters are constants; Obviously, this is but an approximation for the real recorder and its quality as a measuring instrument is determined by the constraints imposed on the deviations of the parameters with respect to their nominal values.

Analog and Digital Recorders

Depending on the transducing device applied, two categories of recorders can be considered: analog and digital. In analog recorders, a physical displacement occurs: Positioning (y) of the pen (or ink jet, light beam or other), toward the site on the paper to be marked. In digital recorders, as

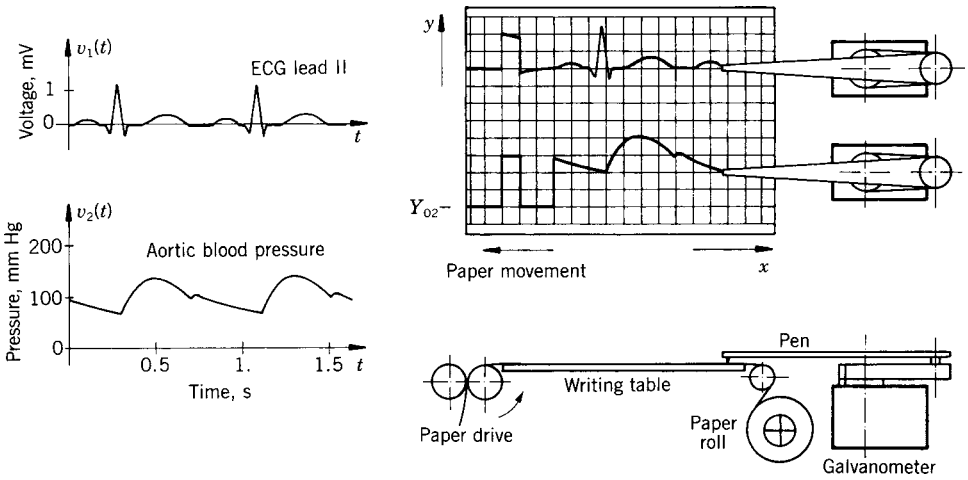


Figure 1. Time signals and graphic representation. On the left, two time signal examples are given: ECG (bipolar lead II) and aortic blood pressure. On the right, the *t–y* recording is sketched as executed by galvanometric rectilinear pen-writing devices on calibrated paper. For ECG and blood pressure the recording sensitivities have to be known. This is accomplished by recording a calibration pulse: 1 mV for ECG, 100 mmHg (0.133 kPa) for pressure. For pressure the zero level is obviously necessary. The time effect is generated by pulling a paper supplied from a paper roll over the writing table. Remark: Nowadays these types of signals are mostly recorded with a digital recorder (thermal array) or a PC application.

considered here, no moving parts are present. The writing device consists of a linear array of stationary writing styli, positioned at equal distances. The array is directed according to the *y*-axis and covers the complete width of the paper. The location *y* to be marked is identified by giving the address of the point in the array situated at the corresponding location.

Three elementary functions can be considered: *x* positioning, *y* positioning, and intensity control of the writing process. The most striking difference between analog and digital recorders is that in digital transducers the *y* positioning does not exist, as the function of indicating the *y* value to be marked is carried out by the intensity control on the stationary styli in the writing array. Analog recorders are capable of drawing continuous lines; digital recorders essentially put separate dots on the paper at discrete *y* levels corresponding to the positions of styli in the array and at discrete *x* values according to the incremental steps at the *x* input. Besides the continuous mode, analog transducers can be used in a discontinuous mode (scanning mode). In this case, the *x* drive is essentially similar to the one used for digital systems (a progression with equal steps) and at each step the transducer scans the paper width at high speed, putting dots (or dashes) at locations corresponding to the different signal values. As such, the digital transducers construct tracings by setting dots, the analog transducers used in the scanning mode generate tracings by setting dots or dashes directed according to the *y* axis. Only the analog transducer in continuous mode behaves as a real curve tracer, implying that the device is unable to generate images.

Digital transducers as well as analog transducers in the scanning mode can produce images that are composed as indicated (dots or lines). Applied as waveform recorders, both instruments have the special advantage that a single transducer can handle a set of signals, the number being limited only by properties of resolution (dependent on the number of writing points of the digital transducer and on the recording width) and of identifiability of possibly overlapping curves. In the continuous mode a separate analog transducer is required for each channel to be recorded. To complete the discussion of analog and digital transducers, note that analog types can be used in a digital way, when positioning is bound to equal incremental steps, as applied in printers and plotters for computer graphic output. Moreover, analog recorders may be used in connection with digital systems, provided the necessary digital-to-analog conversion facilities are present. In such cases, simple digital transducers may be added to the apparatus design. A typical example is the analog apparatus equipped with printing heads for adding alphanumeric information to each tracing.

With these technological possibilities in mind, the problem of reporting the recording parameters (sensitivities, reference levels, paper velocity, timing reference) can be further discussed. Transducers capable of handling more than one signal (analog transducers in scanning mode and digital transducers) may draw calibration lines as well (ordinate and abscissa). The same is not possible for analog transducers in continuous mode. For example, for drawing a zero-level line an additional transducer would be necessary. It is obvious that this solution would be expensive. Moreover, it does not feature the benefits of accuracy as in the method where signals and calibration lines are generated by the same transducer. The problem is solved by using calibrated paper. Nevertheless, in some apparatus a separate timing marker is provided, writing small dashes that indicate time divisions (e.g., each second), or an event marker can be used to identify the start of an experiment (e.g., the stimulus if applied). Such timing devices are

evidently low cost, low performance elements and cannot be used as measuring instruments.

In the case of analog transducers and calibrated paper a zero line and a level corresponding to a standard value of the physical magnitude can be drawn before recording signal information (Fig. 1). For example, in the ECG the 1 mV pulse is commonly used. Ideally, this kind of calibration should cover the complete measuring chain: The zero line and standard level are to be applied at the biophysical sensor input. This can easily be performed in electrophysiological measurements, but for other physical magnitudes it is quite problematic, as it requires the continuous availability of a calibration setup (e.g., blood pressure measurements). In these cases one uses previously calibrated transducers and electrical calibrations corresponding to zero and standard levels of the physical magnitude are then simply provided at the recorder input (Fig. 1). Alternatively, one can draw a piece of zero line and keep the sensitivity values (physical unit/paper division) in file (written on the record). Evidently, both methods, with and without self-generation of coordinate lines, will feature different recording accuracy properties.

In the technique of analog transducing there are essentially two methods. In the direct method, the transducer generates a positioning directly determined by the electrical input. In the second method, the actual position of the writing device is measured by a sensor system and the result is compared with the recorder input value. The difference between both values is (following amplification and signal conditioning) fed to the transducer, which originates a movement tending to zero the position deviation. This second type is called the feedback, compensation, or servo method and corresponds to the well-known null detection technique in general measurement theory. Both methods, direct and feedback, are applied in analog recorders.

Chart Abscissa Generation

An essential feature of the graphic recorder is the ability of making recordings in real time. The time effect is generated by pulling a strip of paper at constant velocity and the name t–y recorder is applicable (Fig. 1). In digital recorders chart abscissa generation is performed essentially in small identical increments with the aid of a stepping motor that is controlled by a stable stepping frequency. In analog recorders, a continuous movement is envisaged and different types of motors are applied in the apparatus design (synchronous, direct current and also stepping motors). The motor can be combined with a tachometer: In this case, the actual velocity is measured through the tachometer, and with the result the motor velocity is corrected to the desired reference velocity value via a feedback circuit (null detection technique).

Paper can be fed in Z-fold or on a roll. The pulling of the paper can be achieved by sprocket wheels operating within equally spaced perforations at both sides of the paper. This method is seen mostly in low paper speed applications. At higher speeds, the paper is pressed between two rollers driven by the motor system (Fig. 1). In most cases as a result of the pulling force, the paper is pulled to the writing table (e.g., pen writing) or to the writing head (e.g., digital systems), which is essential for thorough graphic recording. If necessary, additional pressing facilities are provided to ensure optimal contact with the writing device. Evidently the use of graphic recorders is not limited to real-time applications of the time signals. Delayed and time-expanded or compressed recordings can be realized if computing and memory facilities are provided (for digital as well as for analog recorders). For example, signals with a higher frequency spectrum in comparison with the bandwidth of the transducer, which thus fail to be plotted in real time without considerable distortion, can be recorded in this way: Data are stored at high speed in memory and are released afterward (off-line) at low speed to the recorder input.

Recorders equipped with identical input hardware for both x and y coordinates are called x–y recorders. In this case, the chart is stationary during recording. It can be supplied as separate sheets or from a roll. Different methods are used to stabilize the paper during recording (e.g., mechanical clamps, magnetic parts on an iron plate writing table), but electrostatic attraction to the writing table seems to be the most elegant method of fixation. An x–y recorder can also be used in the t–y mode by applying a linearly increasing voltage to the x input.

Recording Accuracy

Recording accuracy is the first principal quality of the graphic recorder to be discussed. This quality comprises the recorder's performance as a measuring instrument and its ability to display additional information concerning the measured tracings or images. With respect to measuring performance, the larger part of the discussion can be formulated as for other measuring instruments (1–3). Such aspects as accuracy, precision, resolution, static linearity, noise content (including drift), dead zone and hysteresis, dynamic behavior comprising frequency domain and time domain responses, and sampling and digitization effects are typical performance indicators. Parameters for both x and y axes have to be considered. It is known that energy transducing devices usually represent the most delicate functions in the chain. This is valid for biophysical sensors and it applies equally to electromechanical recording transducers. Moreover, the power amplifier driving the transducer may be critical (e.g., with respect to saturation effects). Signal amplification and conditioning modules generally play no limiting role in the overall performance. Properties of the biophysical sensors are not discussed, as this subject falls outside the scope of this article. Analog recording transducers suffer from the limitations and errors typical of analog systems, which are excluded from digital systems. In the latter case, limitations are determined by sampling (sampling frequency) and digitization (number of bits).

Accuracy is an overall parameter defined as the difference between the recorded and the true value, divided by the true value, regardless of the sources of error involved. In digital systems a number of error sources are excluded. Moreover, accuracy is largely dependent on how calibration is performed and how the recording parameters are

reported. It is evident that the recording method in which the transducer itself generates the coordinate lines (x and y) as well as the signal tracings is less subject to error effects than the method using precalibrated paper. In the first type, accuracy is simply related to the correctness of the coordinate values generated; In the second type, there are found a number of additional error sources caused by the mechanical positioning of the writing device (inherent to the analog electrical-to-mechanical conversion) and also the paper positioning. Two methods of analog transducing have been mentioned: It is known that higher accuracy can be achieved with the feedback type (null detection) than with the direct type. The same applies for the accuracy of paper velocity.

Precision of the graphic recording is related to the preciseness with which a value can be read on a tracing. It is thus dependent on the line thickness (a sharp line implies a high precision) and the paper width covered by the tracing (maximal deflection in the y direction) and, additionally, paper velocity (with respect to time readings, in the x direction).

Resolution, being the smallest incremental quantity that can be measured, is determined by the digitizing step in digital systems and the dead zone in analog systems. The dead zone usually originates as a consequence of static friction (e.g., in the case of moving, paper-contacting devices, such as pens, the friction between the paper and the writing element) or backlash in mechanical moving parts. The phenomenon also gives rise to hysteresis: A curve recorded in continuously increasing coordinates will not exactly fit the same curve recorded in continuously decreasing coordinates.

Error sources can be found in the instability of the recording parameters (sensitivity and reference level). There can be small alterations comparable to electronic noise. If the alterations occur very slowly they are called "drift". Drift can result from temperature variations. Digital transducers are not subject to drift. Nevertheless, a mechanical source of drift on the reference levels can be the shifting of the paper along the y axis. Also, the paper velocity might not be stable as a consequence of motor speed variations or paper jitter from the mechanical pulling system. It is obvious that such errors have minimal effect if the transducer itself generates the coordinate lines. This applies for drifts of the sensitivities as well as the reference levels. Noise due to electric mains interference can occur with poor channel isolation and/or poor grounding techniques. Gain and phase distortion due to impedance loading can occur if transducers are driving multiple data collection systems (i.e., chart recorder, medical monitor, VHS tape recorder, computer). As for any instrumentation system the use of appropriate preamplification may help to avoid these inconveniences.

Furthermore, the recording sensitivities may be slightly dependent on the values of the signals processed, implying that there is a deviation from strict static linearity. A specific nonlinearity problem arises in the case of galvanometric transducers where rotation has to be converted into translation. In digital transducers, linearity is determined by the accuracy of the construction of the array of stationary writing styli. Moreover, nonlinear effects are less inconvenient if the transducer generates its coordinate lines, as these are consequently subject to the same nonlinearity as the signals. Note that in transducers with moving parts a specific nonlinearity is introduced for safety purposes. By electrical means (saturation levels) or by mechanical stops the deflection of the moving part is limited (e.g., in multichannel galvanometric pen recorders).

The dynamic behavior of the recorder refers to its response to sine waves or to transients (pulses or step functions). For analog transducers, one describes the frequency dependence of the sensitivities; A general discussion can be found in linear system theory (1–3). Some typical properties of linear systems can be mentioned. For example, a sine wave finds itself, in any case, reproduced as a sine wave at the output of the system, possibly with altered amplitude and phase. Another typical feature is that the spectral bandwidth of the system is independent of the signal amplitude; likewise the sensitivity in the bandwidth. Most linear analog recording transducers act as low pass systems: Frequencies from 0 Hz (dc) up to a certain cutoff frequency (the bandwidth) are about equally recorded. Beyond the cutoff frequency, amplitudes are progressively attenuated as frequency increases. When a step input is applied, the recorder will not exactly follow: There will be some delay, a limited rise time, and possibly an overshoot with respect to the steady-state level. These parameters (cutoff frequency, delay, rise time, overshoot) are significant in characterizing the dynamic behavior of the analog recorder. Real analog transducers behave as a linear system only within restricted limits (of deflection, slew rate, and also acceleration). Nevertheless, their performance is characterized by the same parameters. Limitations and errors are due mostly to the electromechanical transducer itself and possibly to the driving power amplifier. As for any measurement system, insufficient bandwidth will give rise to gain and phase distortion, delay and decreased slew rate. In digital transducers, where no moving parts are present, there are no errors connected to electromechanical positioning, moreover, coordinate lines are easily reproduced. In this case, limitations are determined primarily by the act of sampling and digitization. The phenomenon of overshoot is nonexistent. Delay and rise time correspond to the interval between two writing (printing) actions (the writing (printing) period). The bandwidth in analog systems determined by the –3 dB frequency (the frequency at which the sensitivity is reduced to 70% of the static sensitivity), is seen in another way: It depends on the number of samples one finds necessary to represent a complete sine wave period. If 10 is the approved number of samples for a complete period, the bandwidth is restricted to one-tenth of the writing frequency (in real-time applications). With respect to bandwidth considerations one must keep in mind that the result, the graphic record, is intended for visual inspection and that paper velocity is limited. For example, with a 100 Hz sine wave to be recorded, at high paper velocity, such as $100\,\mathrm{mm\,s^{-1}}$, a full sine wave period covers only 1 mm, implying a poor recognizability. If the latter is essential, memory recording has to be used.

The ability to display additional information is complementary to the measuring performance. Recording

parameters (sensitivities, reference levels, time references, paper velocity) are, as already indicated in relation to accuracy, best reported in the form of calibration lines. Alphanumeric information for identification of signals and calibration lines, on the one hand, and data concerning the circumstances of the experiment (patient or subject name, date, experiment identification, stimulus type, etc.), on the other hand, are indispensable for off-line examination. They either must be manually added to the record or, preferably, must be added through the recorder itself. Digitally controlled recorders (using digital as well as analog transducers) may allow the operator to introduce the alphanumeric data together with tracings, provided the device is able to do it. In the case of digital transducers this is only a matter of appropriate hardware and software (to be supplied in the recorder) and/or interfacing with a computer. Whereas in digital transducers recording is essentially based on intensity control, in analog transducers this aspect is not necessarily provided.

Reading stored paper tracings by using waveform tracing techniques in order to be able to process data in digital format may be interesting for a number of studies. Nevertheless, attention has to be paid to the recording accuracy of older tracings and the experimental procedures used.

Graphic Quality

As mentioned in the introduction, the second of the two principal recording properties is the quality of graphics. The correctness of locating the point to be marked is part of the performance of a measuring instrument. In this section, the quality of the apparatus in producing a graphic hard copy is discussed. The clearness of the individual tracings is the first quality. Sharp and clean tracings on a bright background give the best impression. Vague or blurred lines and dirty background are not wanted. Furthermore, curves should be easily identified: Ease and speed in examination are thus improved and the risk of misinterpretation is considerably decreased. Although the use of different colors seems most appropriate, other techniques may be applied if color differences are technologically excluded, for example, writing intensity (gray scale) or even line thickness. Also, the impression of continuous lines can assist in curve examination; Adequate interpolation between sample points is thus a specific problem to be handled in the case of digital and discontinuous analog recorders (Fig. 2).

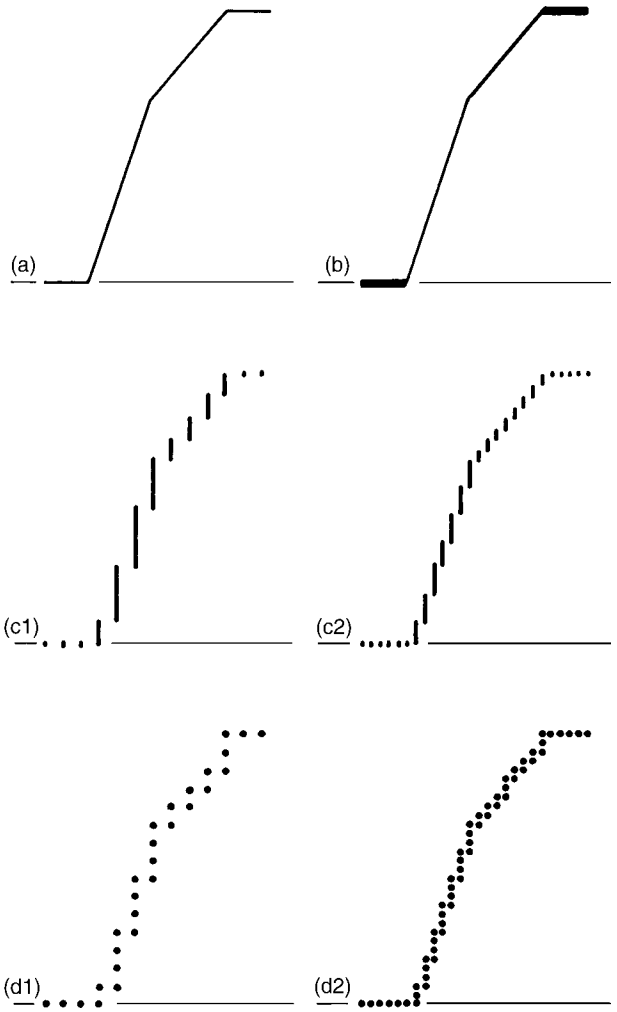


Figure 2. Graphic quality. (a) Artificial signal to be recorded. (b) Record from a pen-writing system, showing the effect of deflection velocity on line thickness. (c) Record from an analog scanning system. Interpolation is performed by vertical dashes drawn between previous and present signal values at each writing act. The effect of the writing frequency is visible in comparing (c1) and (c2) (doubled writing frequency). (d) Record from a digital system with similar interpolation. The effect of resolution is visible in comparing (d1) and (d2) (halved distance between writing points, i.e., one added to the number of bits, and doubled writing frequency).

In the graphic process, two steps are involved. First, the act of writing (putting a visual mark on the paper at the located point). Second, the act of fixation (to maintain the mark for a long time). This evidently has to do with the aspects of visual inspection and permanence in the definition of graphic recorders. Graphic quality also relates to the aspect of permanence. For example, depending on the fixation process, photographic records may be affected by environmental light in the form of a darkened background.

Attention should be paid to the fact that the flow of marking matter (ink, heat, light, etc.) is to be adapted to the writing velocity in order to have optimal line thickness and grayness. In analog transducers the writing velocity is determined by the paper velocity (in the x direction) and the deflection velocity of the writing device (in the y direction). In digital recorders, only the paper velocity is involved. Adapting the marking flow is evidently a matter of intensity control (ink pressure, heating power, light intensity, etc.). In low cost apparatus, manual intervention is necessary. Adaptation might not even be possible. In more sophisticated apparatus, writing matter flow is automatically adjusted in relation to paper velocity. In the steady state, then only alterations in intensity and line thickness are discernible in the case of variations of deflection velocity (Fig. 2). This explains why, in typical ECG pen recordings, the baseline appears much heavier compared with the slopes in the QRS complex. One could say that this side phenomenon has a beneficial effect on waveform recognizability since information on the signal

time derivative is also included in the graph. Evidently, this statement may be subjective. Some apparatus are even equipped with an intensity control that continuously adapts the writing flow to the total writing velocity. The latter is possible only if the response of the writing system is fast enough. A good example is the thermal system with a pen-tip writing device. Because of the small size of the heating resistance in the tip of the pen, the thermal time constant can be reduced to a few milliseconds and fast heating flow adaptation is possible. In the thermal edge writing device, the resistance producing the heat is much larger. The thermal time constant is then $\sim$1 s so that the latter may even result in a nonnegligible waste of paper when the instrument is started at high speed paper velocity.

When ink is used, the paper can be marked by a contacting device, a pen, or by a noncontacting device, an ink jet. Pens can be designed as capillary tubes connected to a pressurized ink reservoir. They can be fiber-tip pens or ballpoint pens (disposable pens) and ink can be supplied from a common container or from (disposable) cartridges connected to each pen. In the case of ink jets, there is no contact with the paper. Ink is squirted as a very thin jet from a fine nozzle directed toward the paper. Fixation occurs as the ink is absorbed and dries. Ordinary paper can be used, satisfying the needs of absorbency quality and surface smoothness (in the case of contacting pens). A disadvantage with ink systems is the possibility of ink stains and smears, as fixation (absorbency and drying) does not occur instantaneously. A practical problem is that pens may clog or "bleed."

The following methods require special paper types. These can be composed of an appropriate paper base coated with a special layer functional with respect to the writing process (e.g., thermochemical). In the case of the burning method, an electric current from a stylus tip in contact with the paper is passed through the electrically conductive paper to the metal support on which the paper is mounted. The current gives rise to paper burn and thus blackening. The density of the burned mark depends on the current magnitude. Thus, a gray scale effect can be obtained. As burning is irreversible, the graphics remain fixed. This method finds application in the previously mentioned sound spectrograph. Furthermore, marking can be executed by heat on thermosensitive paper. The writing effect can be originated by thermally melting off a white coating from the chart paper and thus exposing the black underlayer. In another process, the base paper layer is covered with a thermochemical layer that irreversibly changes color after exposure to heat. In thermal writing, the color is usually black, but other colors can be generated. The shelf life of thermal chart recordings can vary on paper, transducer, and user storage techniques; Thermal recordings may be degraded by exposure to heat, light, friction, solvents, and so on.

In both electrostatic and photographic methods, the writing act comprises two phases. First, a latent invisible image is transferred to the paper. Second, the image is developed and thus made visible. In the electrostatic method, paper coated with a special dielectric layer is locally charged (the latent image) with electrical charges. These charges attract, in a second step, ink particles from a toner supply (fluid or dry) after the excess toner is removed only ink is left at the electrically charged sites. Fixation is then achieved by drying (in the case of toner fluid with ink particles in suspension) or heating. In the photographic method, which uses photosensitive paper (the most expensive type), a light beam (generally ultraviolet, UV) activates the photolayer, leaving a latent image. Graphics are then visualized and fixed according to the photographic process used by exposure to visible light (environmental light or an additional light source specially provided) or by heating. Dependent on the process used, long exposure to environmental light may degrade the quality of the record after the examination by darkening the background. A darkened background may also appear in the electrostatic writing process if the apparatus is not optimally adjusted (e.g., if not all of the excess toner is removed). The photographic method is best fit for the production of images, but it has also been used for tracings.

Some commonly used writing techniques have thus been described. Acts of marking and fixing are typical technological problems and this area continues to evolve. The method used in a specific apparatus depends on different factors. With respect to the application envisaged and the type of transducer used, specific writing techniques may appear advantageous, but the manufacturers' patents also play an important role in apparatus design. The paper cost is certainly not to be neglected (e.g., in long-term recordings). Electrostatic and photographic writing have lost interest in modern chart recorder design.

A remark has to be added with respect to graphic quality as well as precision. As already mentioned, recorded tracings should be optimally identifiable, especially when they cross each other on the record. The use of different colors is obviously beneficial. In this respect, the paper width and the usable range for one channel are important. Traces can be limited within separate ranges or can cover the complete paper width. Precision is maximal in the latter case, but there are problems involved. Some transducers, more specifically galvanometric devices, produce only limited deflections. Furthermore, there is a problem with contacting devices (pens). Since it is physically impossible that they cross each other, they are spatially shifted with respect to the abscissa of the paper, which corresponds to a desynchronization on the records. If the complete paper width is covered, precision is maximal, but identifiability is decreased, especially when the use of different colors is excluded as in those devices where the color is not determined by the writing element, but by the paper properties. The latter applies to all devices other than ink-writing devices.

Another remark concerns immediate visibility of real-time recorded tracings. Although the recording may be preferably in real time, the visibility may not have the same benefit: More specifically, tracings can be seen after a small time delay. This is the case if the writing device is positioned within the recorders housing; The time delay is then dependent on the distance to the visual part of the paper and, evidently, of the paper velocity. Digital transducers have this shortcoming: pen recorders do not, except in the multipen recorder with dislocated pens (allowing

overlapping graphs). In this case, a pen offset compensation may be used to synchronize recorded tracings, also causing a time delay.

In this Paragraph a number of technological aspects were just mentioned. Some of these concepts have lost interest with respect to new design, but may still be found in older apparatus.

ANALOG RECORDERS

An analog recorder has been defined as a recorder that applies an analog electromechanical transducing principle. The electrical magnitude is transduced into a translational value, that is, positioning of a movable part, such as a pen, toward the site to be marked. Although analog recorders are used mostly for signals in analog form, they may also be applied to digital signals and consequently to computer output, provided digital-to-analog conversion facilities and appropriate control access (paper velocity, intensity control) are available.

For the description of the transducer, analog system theory is applicable (1–3). Two typical measurement principles are commonly used in transducer design. In the direct method, the electrical magnitude is directly transduced into a displacement magnitude. In the feedback method, corresponding to the null detection technique, the actual position of the marking device is measured with an accurate sensor. The difference (the error) between the measured and the input value is, after appropriate conditioning, fed back to the transducer, which generates a movement tending to zero the position error. In the latter case, the accuracy of the transducing system depends on the quality of the position sensor. It is known that with this method higher accuracies and more stable transducing properties can be achieved. Furthermore, an analog transducer can be applied in the continuous mode or the discontinuous mode. In the first case, a single continuous tracing is generated by each transducer; The number of channels is equal to the number of transducers provided in the multichannel recorder. Production of an image is excluded. In the second case, the writing device is repeatedly swept over the complete paper width, setting dots or dashes at sites corresponding to the signal values (the scanning mode). A single transducer can thus handle a set of signals. It is obvious that what is gained by the ability to process different signals is lost with respect to the writing speed of the system.

Positioning of the writing device can be caused by a transducer that causes a rotation (the galvanometric transducer) or a translation (the translational servosystem). The translational servotransducer evidently makes use of a feedback mechanism. The galvanometric transducer, originally strictly built as a direct device, has also been designed according to the feedback principle.

Whereas two decades ago a number of analog transducer techniques were applied in apparatus design for biomedical signal recording, mostly for obtaining different bandwidths and image recordings (as for echocardiography), only a few of them are still used. The galvanometric recorder and, more importantly, the translational servorecorder.

Galvanometric Recorders

Galvanometric recorders make use of transducers that are essentially rotational transducers, making use of the d'Arsonval movement as applied in ordinary galvanometers (4). A positioning, more specifically a rotation over a certain angle, is obtained as a result of an equilibrium between two torques, the first an electromagnetic torque, proportional to current, the second a mechanical torque, proportional to positioning (a rotation angle). A coil wound in a rectangular form is suspended within the air gap between the two poles of a permanent magnet and a stationary iron core. A current flowing through the coil gives rise to an electromagnetic torque that tends to rotate the coil. To position at a specific angle θ, there must be a restoring torque, essentially proportional to the angle, for example, a torsional spring characterized by its rotational stiffness. The signal to be recorded is fed to a current amplifier that drives the coil. As such, there is direct proportionality between the angle and the signal. This principle corresponds to the direct transducing method.

In a discussion of dynamic behavior, not only stiffness, but also damping (viscous friction) and inertial moment (of the coil and the attached mechanical parts, e.g., the pen) must be taken into account. As a first approximation the device acts as a second-order linear system, characterized by a resonant frequency and a damping factor. The resonant frequency depends on the ratio of stiffness divided by inertia. The resonant frequency and the damping factor determine the bandwidth of the system (with respect to sine wave response) and the time delay, rise time, and overshoot (with respect to step input transient response). For frequencies sufficiently above the resonant frequency, the corresponding amplitudes are attenuated proportional to the square of the frequency. A second-order system can show resonant phenomena, depending on the value of the damping factor. If this factor is <1, the response of the transducer to a step input shows a ringing effect, implying that an overshoot exceeding the steady-state value has occurred. The smaller the damping, the smaller the rise time, but the higher the overshoot. At critical damping (damping factor $= 1$) no overshoot is present. Such an overshoot is considered inconvenient as it gives an erroneous impression of the waveform (especially if sudden alterations in the signal occur). With the damping factor in the proximity of 0.7 a good compromise is obtained between the overshoot (4%) and the rise time (3) (Fig. 3). The useful bandwidth of the recorder is thus determined by the resonant frequency and thus by the ratio of stiffness to inertia. To obtain a sufficient damping factor, extra damping must be applied within the galvanometer (mechanical or electromagnetic). In the case of paper-contacting devices, the damping should exceed the effect of the static friction of the writing element (e.g., the pen) on the paper (as this effect is unreliable and consequently is not allowed to affect the measuring performance).

The feedback technique has also been used. According to the general idea, the actual angle is measured by a position-sensing device, delivering an output proportional to the angle, which, after amplification and waveform conditioning, is negatively fed back to the input of the

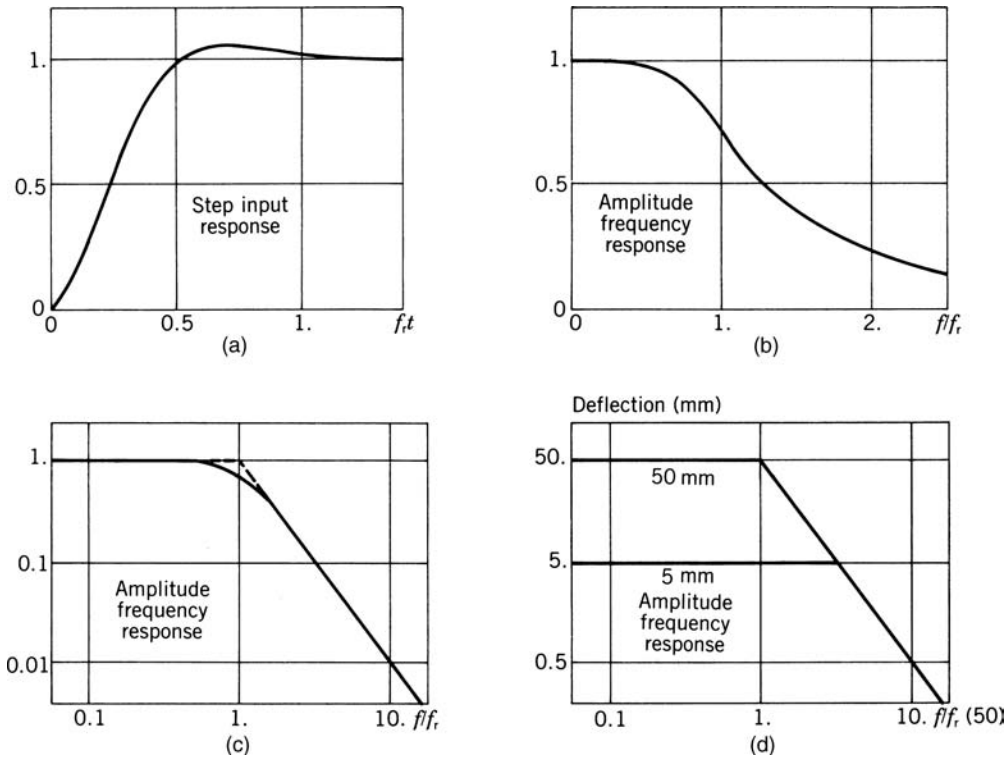


Figure 3. Characteristics describing dynamic galvanometric transducer responses. (a–c) Linear system behavior; Damping factor 0.7. (d) Nonlinear behavior. Coordinates are given in generalized form (time = $f_r t$; frequency = f/f_r) (f_r = the transducer resonant frequency). (a) Time response to a unit step input. Delay, rise time, and overshoot are visible. (b) Amplitude frequency response (sine waves) in linear coordinates. (c) Amplitude frequency response in logarithmic coordinates. (d) Typical amplitude frequency responses (50 and 5 mm deflection) for a nonlinearly behaving system. The bandwidth appears to depend on the tracing amplitude (displayed: inversely proportional to the square root of the amplitude).

current amplifier driving the galvanometer. In case the position signal simply undergoes amplification, the effect of the negative feedback is a restoring torque proportional to the angle, thus resembling the effect of a rotational spring in an apparent stiffness. An equilibrium is reached when the coil angle (as measured) corresponds to the value at the input of the system. At equilibrium, the current through the coil equals zero (provided one does not take into account the effect of an additional mechanical spring). This is advantageous with respect to the current amplifier.

A real galvanometer usually does not feature constant linear system properties, such as stiffness (mechanical or apparent, as generated by feedback) and damping, independent on deflection, deflection velocity, and acceleration. Particularly with the feedback system, nonlinear effects can appear as a result of current saturation. In a linear system, the bandwidth is independent of the signal amplitude; In a real galvanometer, the same usually does not apply. For larger sine wave amplitudes, the useful bandwidth appears decreased and sine wave distortion occurs at frequencies in the vicinity of the apparent resonant frequency. A bandwidth inversely proportional to the sine wave amplitude to the power of m, with the exponent m approximately between 0.5 and 1, is obtained (Fig. 3).

Apart from nonlinearity in dynamic behavior, a problem with respect to static linearity arises. The galvanometric recorder essentially generates a rotation, not a translation, as one would expect for a graphic recorder. If a pen is connected to the coil of the galvanometer, the tip of the pen rotates with a radius equal to the pen arm length. Such a recording is curvilinear. It has been used on paper with curvilinear coordinate lines. Rectilinear recording is obtained with special techniques. Whereas in curvilinear recording mostly ink pens are used (a low cost solution), there are different methods for obtaining rectilinear records: pen methods (long-arm pens, knife edge recording, mechanical rectilinearization), ink jet method and light beam method) (Fig. 4). Curvilinear errors can be kept small if the radius is large, which is the case with long-arm pens. Knife edge recording is the simplest real rectilinear recording. The chart paper is pulled over a sharp edge, accurately directed according to the y-axis, and the writing stylus moves over and presses on it. The impression is thus made at the site of the edge and thus rectilinear. In this case, the marking method applied is thermal. It should be remarked that in this method a large part of the stylus has to be heated (as compared to the thermal point writing), giving rise to large thermal time constants (order of magnitude 1 s). This is inconvenient for high paper speeds as, when

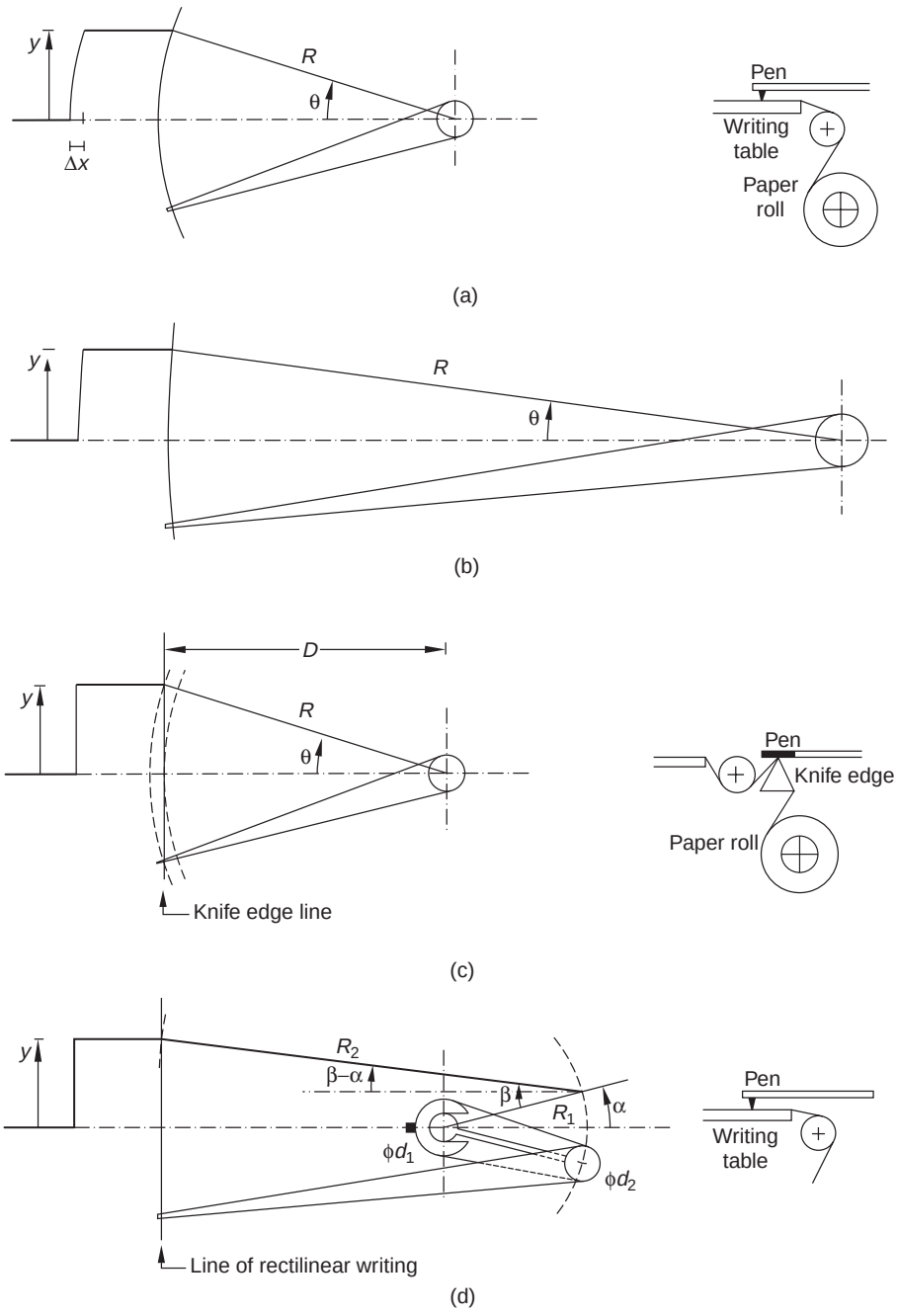


Figure 4. Galvanometric pen-recording assemblies. An artificial recording is shown corresponding to a zero level, abrupt transition to maximal positive, holding, and finally abrupt transition to maximal negative deflection. Errors resulting from dynamic behavior are not considered in this figure. (a) Curvilinear recording. (b) The long-arm pen recording is curvilinear but, as a consequence of restricted angle deflection, approximates rectilinear recording. (c) Knife edge recording. Rectilinear recording as the writing essentially occurs at the site of the knife edge, where the pen presses on the paper. For thermal writing, the black end of the pen represents the part to be heated (minimal length indicated by the dashed lines). (d) Mechanical linkage for rectilinearization. The pen arm (length R_2) is connected to a pulley (diameter d_2), which is allowed to rotate at an axis at the end of the galvanometer arm (length R_1). A metallic belt attached at a point of the pulley is also fixed at a stationary circular part (diameter d_1) at the galvanometer frame. As the galvanometer arm rotates, the pen arm is subjected to a corresponding rotation dependent on the diameter ratio: $\beta = (d_1/d_2)\alpha$. When $\beta = \alpha$, the pen arm is moved parallel to the zero position and the pen point moves in the same curvilinear direction as the end of the galvanometer arm. Thus β must be larger than α, and more specifically so that the rotation through the angle $(\beta - \alpha)$ the curvilinear abscissa error of the galvanometer arm end is corrected for. A power series expansion shows that this is obtained when $R_1/R_2 = [(d_1 - d_2)/d_2]^2$.

the instrument is started, a significant amount of thermosensitive paper may be wasted. Instead of a hot stylus, a carbon ribbon with an ordinary stylus pressing on it has been used. In this case, ordinary paper can be used, but an additional mechanical device linked with the paper velocity equipment for driving the carbon ribbon is necessary. Rectilinear recordings can also be generated via mechanical linkages that compensate for the normal curvilinear movement of the pen and convert the rotary motion of the coil into an (approximately) straight-line motion of the pen tip. In the case of ink jet recorders, the writing device does not make physical contact with the paper. In this case, a fine ink jet is produced by pumping ink through a nozzle connected to and thus rotated by the galvanometer toward the paper. The recording is rectilinear at the intersecting line of the paper plane and the plane of the rotating ink jet. The same holds for the optical method, where a sharp light beam is reflected by a small mirror connected to the rotating coil of the galvanometer. In this case (expensive), photosensitive paper has to be used.

The principal difference in the abilities of these types of galvanometers is in the achievable bandwidth. Galvanometric pen-writing systems have to produce a considerable torque, as the pen, being pressed to the paper for graph production, must be moved easily without interference from the static friction in the recorder performance. This implies a large driving current and, for a defined angle, a large stiffness. This stiffness, together with the amount of

inertia, determines the resonant frequency and thus the bandwidth. A typical bandwidth value for a galvanometric pen system is 100 Hz. Theoretically, a higher bandwidth could be obtained for the same inertia, but this would imply a larger stiffness and thus a larger current, the latter being limited by its heating effect on the coil. As this paper-contacting problem is nonexistent in ink jet and light beam galvanometric recorders, current and stiffness can be lower. Moreover, the coil assembly can be made with such low inertia that, not withstanding the lower stiffness, a much higher resonant frequency can be obtained. An order of magnitude for ink jets is 1000 Hz. For optical systems, 10 kHz has been reached.

The static friction in the case of paper-contacting writers give rise to a dead zone: A zone in which input voltage can be altered without causing any pen movement. It thus determines the resolution of the pen recorder. Hysteresis phenomena can consequently be observed. A curve recorded in continuously increasing coordinates will not exactly fit the same curve recorded in continuously decreasing coordinates.

The paper width covered by galvanometric pen writers is usually small (40–80 mm) as a result of the angle limitation (with respect to linearity) and pen arm length restriction (with respect to inertia and consequently bandwidth). Moreover, as overlap of tracings is physically impossible (the pens might strike each other) in multichannel recorders, a limited part of the paper is assigned to each channel.

As already mentioned, galvanometric recorders have lost interest. Pen recorders (e.g., edge recorders) can still be found on the market, but the types with higher bandwidth (ink jet and light beam) have been replaced by digital (memory) recorders.

Translational Servorecorders

In the translational servorecorder, the writing device (usually an ink pen) undergoes a real translation as it finds itself bound to mechanical straight-line guidance. The translation is generated by a motor and an appropriate mechanical linkage composed of a wheel and closed-loop wire system (Fig. 5). It is essentially a feedback method. A position sensor supplies a voltage directly proportional to the position of the pen, that is, the distance with respect to its zero position. In most apparatus, this sensor is a rectilinear potentiometer; This explains the use of the alternative nomenclature (potentiometric recorder) for this type. A further description of its function agrees with the general feedback principle. The input signal is compared with the pen position value and the amplified and conditioned difference voltage drives the servomotor, causing the pen to move (rectilinear) until the position value equilibrates with the input signal. The ink pens used are capillaries, fiber tips, or ballpoint types, generally supplied with ink cartridges. Evidently different colors can be used for optimal signal discrimination. In modern designs, digital servodevices are being used, with position reading by optical or ultrasonic means.

The static linearity of the recording is determined by the linearity of the position sensor. Noise can result from problems with the sliding contact on the potentiometer

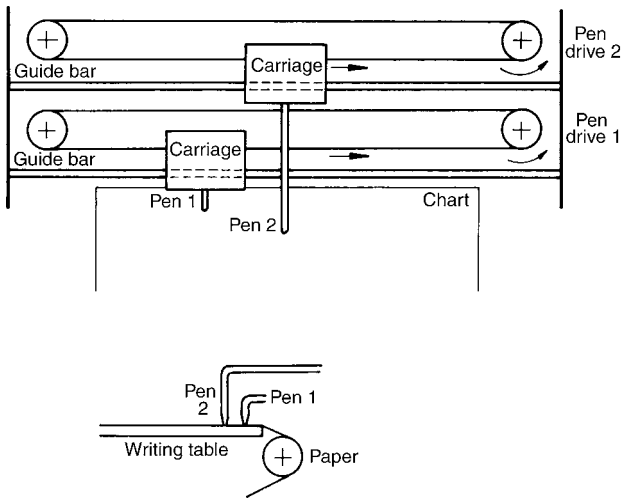


Figure 5. Translational servo *t–y* recorder. Two channels are shown. Pens (1 and 2) undergo real translations as they are connected to carriages that slide along guide bars. To be able to cover the complete paper width the pens are mechanically staggered with respect to the paper abscissa.

resistance. Hysteresis can be caused by the properties of the servosystem (e.g., static friction) and the mechanical backlash between the pen tip (the real location) and the wiper on the potentiometer (the location processed in the feedback loop). As for dynamic behavior, one can state that the servorecorder generally does not behave as a linear system (except for small amplitudes). It is characterized by a limited slewing velocity (order of magnitude: 0.5–2 $m \cdot s^{-1}$) dependent on the type of motor and mechanical linkage. Also, the acceleration of the moving parts is subject to limitations (order of magnitude: 50 $m \cdot s^{-2}$). If one of these limits is reached, the device ceases to act as a linear system and, as described for galvanometric recorders in the feedback mode, the bandwidth depends on the tracing amplitude (order of magnitude of the cutoff frequency: 1–5 Hz). As such, these recorders are fit for slowly varying parameters (temperature, heart rate, mean blood pressure, respiration, etc.). They are generally applied for higher precision: The common paper width is 200 or 250 mm. A writing control is normally provided in the form of a pen lift (on-off functioning) by electromagnetic and/or by manual means.

Servorecorders are used as strip chart recorders (*t–y* recorders) and also as *x–y* recorders. In the latter case, two servosystems are assembled. The *x* system drives a long carriage, directed parallel to the *y* axis and covering the complete width of the paper in the *y* direction. Hereon, the second servosystem *y*, which eventually positions the pen, is mounted. Paper is fixed on the writing table (preferably by electrostatic means). Driven by computer output these *x–y* recorders can plot arbitrary graphs (plotters). When addition, alphanumeric information and coordinate lines can be added if the pen lift control is used. Some *x–y* recorders also have the built-in facility of *t–y* recording (with a ramp signal at the *x* input).

In the case of the *t–y* recorder, several channels (1, 2, 4, 8, 12 channels) can be assembled, all of them covering the

complete paper width, provided the pens are mechanically shifted with respect to their abscissa position (Fig. 5). This mechanical offset, evidently corresponding to a shift in time on the graph, may be inconvenient if values of different signals at a specific time instant have to be compared. To overcome this disadvantage, most servorecorders can be supplied with a pen position compensation unit. All channels, except one corresponding to the first positioned pen along the time axis, are digitized and stored in memory. Data are released, converted to analog, and supplied to their corresponding channels with a time delay equal to the distance from the pen to the first pen divided by the paper velocity. This compensation is evidently beneficial for examination afterward, but can be inconvenient at the time the experiment is executed as the information is plotted only after a time lag dependent on the paper velocity (except for the first pen).

Some modern recorders have data-acquisition facilities. Data (signals and instrument settings) are digitally processed and can be stored, for example, on floppy disk or on a memory card. Communication with computers can be done via standard interfacing (GP-IB, RS-232C, IEEE-488). In some apparatus, a display is provided allowing visualization of recorder settings. As slowly varying signals are envisaged, sampling frequencies used range from 100 to 400 Hz.

DIGITAL RECORDERS

A digital recorder has been defined as a recorder that uses a digital transducer. In this case, no moving parts are present and the transducer consists of a stationary straight-line array of equally spaced writing styli covering the complete chart width (Fig. 6). As such, the nomenclature array recorders can also be used. Progression along the x axis is essentially discontinuous. The paper is held stationary during the writing act. At a given x position, a set of writing points is activated. The paper is marked with dots at the sites in intimate contact with these points. The resolution with respect to the y axis depends on the density of writing points. The resolution in time (x axis) depends on the writing (printing) frequency (the inverse of the writing interval) and on the paper speed. Signals are thus reproduced as discrete values at discrete times. Analog as well as digital signals can be processed. In the analog case, sampling and digitization are part of the recording process.

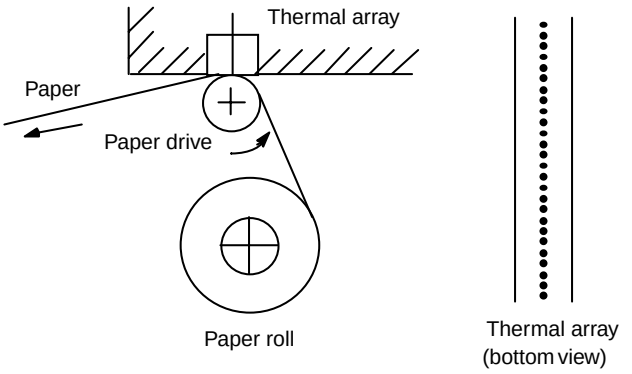


Figure 6. Thermal array recorder. The thermosensitive paper is pressed to and pulled over the thermal array by the paper driving roll.

At present, the borderline between digital graphic recorders, digital oscilloscopes, and data-acquisition systems becomes unclear as a number of the latter instruments are also provided with graphic recording facilities with strip charts. Some digital recorder designs primarily focus on fast acquisition of data and storage in memory for reproduction afterward. These are indicated as memory recorders and generally their real-time properties do not match the fast graphical recording types. Digital graphic recorders have a number of properties regarding data-acquisition, monitoring, storage, and communication with computers.

The act of writing of the digital recorder is generally thermal on thermosensitive paper. Also, an electrostatic method has been used. The writing transducer is essentially a rectilinear array of equidistal writing points covering the total width of the paper. These writing styli consist of miniature electrical resistances that are activated, that is heated, by a current pulse of short duration. Typical resolutions are 8 or 12 dots·mm^{-1} according to the y direction. Static linearity of the recorder thus depends on the quality of the construction of the array. Inconveniences resulting from mechanical moving parts are not present. There is no overshoot. Delay and rise time are equal to one writing interval. All processing occurs in digital form and recording is a matter of generating the correct addresses of writing points to be activated.

Evidently, the dynamic thermal properties of the resistances in close contact with the chart paper and the shape of the electric current pulse set a limit to the writing frequency. A typical value for the writing frequency is 1.6 kHz. In a memory recorder, it may be lower. Local heating of the thermosensitive paper results in a black dot with a good long-term stability, dependent on the paper quality. Evidently, thermal paper is sensitive to heat, but also care has to be taken with pressure, light, solvents, glue, and so on, in order to prevent detoriation of the graph. Generally, the heating pulse is controlled with respect to chart velocity in order to obtain an appropriate blackness and line thickness at different velocities. Nevertheless, at the highest velocities the print may become less black and sharp. Although possible within a limited range, blackness and thickness of tracings are seldom used for the purpose of trace identification.

Whereas the resolution according to the y axis is determined by the thermal dot array, the x-resolution is limited by the dynamic properties of the array, and thus by the paper velocity. At low velocities, generally 32 dots·mm^{-1} (exceptionally 64 dots·mm^{-1}) are applied. At the highest velocities used (100 mm·s^{-1}, in some apparatus 200 and 500 mm·s^{-1}), the number of dots printed is determined by the writing frequency (e.g., 1600 Hz results in 8 dots·mm^{-1} at 200 mm·s^{-1}).

Data acquisition is determined by the sampling frequency, expressed per channel recorded. According to the sampling theorema this should be at least twice the highest signal frequency in order to prevent aliasing

effects, but for obtaining a graph with minimal graphic quality at least 10 points are needed for a sine wave period. On the other hand, sampling frequency should be limited in order to prevent an excess of data stored in memory (if applicable). Sampling frequencies used are typical 100 or 200 kHz (exceptionally 500 kHz and 1 MHz). A basic component of the digital recorder is RAM memory in order to create high quality graphs. The sampling frequency is generally higher then the writing frequency. If the sample frequency equals the writing frequency, the recorder prints all dots between the former and the present value. In fact, curves composed only of points are difficult to examine and interpolation improves the impression of smoothness of the sampled curves. If the sample frequency is higher than the writing frequency, the recorder prints all the dots between the smallest and the largest values (including the former value within the writing interval). As such at each writing act vertical lines are drawn. In this way, fast transients, disregarded by the ordinary interpolation between former and present value, are also captured in the graph. Although the form of the transient cannot be interpreted, the graph provides evidence of its presence. A signal having a frequency larger than the writing frequency is consequently displayed as a continuous black band.

Typical digitization levels are 8 or 12 bits, exceptionally 16 bits. Widths of thermal arrays range from 100 to 400 mm. With a resolution of $12\,\text{dots}\cdot\text{mm}^{-1}$, the largest number of dots full scale is then 4800. In case 16 bits are used, a signal with a large dc offset can be reproduced with excellent graphic quality as the offset can digitally be removed and the scale adapted.

Coordinate lines (with identification) can be generated and precalibrated paper is unnecessary. Errors resulting from paper shifting are thus eliminated. Tracings can cover the full width of the paper. Tracing identification in alphanumerical form and comment on experiments performed can be easily added through an internal or external keyboard. Signals can be calibrated and real physical values and units can be printed at the calibration lines. Simple mathematical calculations can be performed on signals in real time and results recorded in virtual channels. In most performant recorders a screen, mostly color LCD, is provided, allowing display of tracings in different colors and settings of recording parameters. Monitoring signals certainly has advantages. In real-time recording signals are not directly visible as a result of a delay corresponding to the distance between the internal thermal array and the visible part of the paper, but can be observed without delay on the monitor. Recording parameters (gain, offset, ...) can be set using the monitor and thus avoiding paper spoiling. Signals stored in memory can be viewed before recording on paper. Cursors on the display can be used for reproduction of selected parts of signals. Some recorders have a built-in display (up to 18 in., 45.72 cm). Others can be connected with an external display.

Digital recorders, especially memory recorders, are able to store recorded data and apparatus setting. Besides RAM for fast storage, recorders may have a built-in hard disk, floppy, magnetooptical or ZIP disk drive and a slot for a memory card. Connections may be provided for external memory media. Signals can thus be reproduced from memory, processed and recorded on chart. Time compression or expansion is obviously possible. As data can be sampled at a high rate, stored in memory, and then recorded as they are released at a reduced rate, a bandwidth higher than that determined by the writing frequency can be achieved. In this way, fast transients that cannot be handled on-line can be accurately reproduced. Data capture and/or graphic recording (with an appropriate speed) can be controlled by trigger functions (external or derived from recorded signals); Pre- and posttrigger may be chosen. Via memory x–y plots can be obtained. When appropriate software is available, FFT plots can be made.

Digital recorders can have analog and logic channels. Typical numbers of analog channels are 4, 8, 16 (and exceptionally 32 and 64). General monopolar or differential amplifiers or signal conditioners with analog antialiasing filters can be plugged-in, as well as special purpose amplifiers for biomedical purposes. Furthermore interfaces can be provided for computer connection (RS-232C, IEEE-488, GPIB, Ethernet, USB, SCSI) or external hardware. Software can be provided for control, data transfer and conversion of data to formats for popular data analysis software programs.

EVALUATION

Table 1 shows a list of manufacturers of recorders, website addresses and typical products. Analogue recorders with a rotating pen arm, analog recorders with a translating pen and digital thermal array recorders. Evidently, the list is incomplete and continuously changing. For extensive practical details on recorders, signal conditioning modules, displays, data storage and processing, computer interfacing, and so on, the manufacturers' data sheets should be consulted.

Analog recorders are characterized by their specific technological limitations: the bandwidth, restricted mostly by the inertia of the moving parts, imposing constraints on the frequency content of the signal; the paper width covered by a trace, the number of channels, and the trace overlap (including pen dislocation); the stability of the recording parameters (including shifting of calibrated paper) and the difficulties of adding coordinate lines and alphanumeric identification to the records. The use of auxiliary devices (timing marker, printing head, pen position compensation, etc.) has been shown to overcome some of the limitations. The bandwidth is considered one of the most important limitations and thus the signal's spectral content determines the choice of the recorder type. Galvanometric pen recorder types (100 Hz bandwidth) have lost interest but some types can still be purchased. A number of mechanisms have been applied in commercially available devices, but in many applications analog types have been replaced by digital types or PC set-ups. Translational servorecorders have kept their position in their specific domain; They can be used for slowly varying signals with a signal spectrum up to 5 Hz (body temperature, heart rate, mean blood pressure, laboratory applications, etc.).

Table 1. List of Manufacturers of Graphic Recorders with Websites and Products[a]

Manufacturer	Rotating Pen	Translating Pen	Thermal Array
Astro-Med, Inc. http://www.astromed.com	Model 7 (ink)		Dash 18 Dash 2EZ Dash 8X Everest
Western Graphtec, Inc http://www.westerngraphtec.com	WR3310 (tp) WR7200 (tp)	WX3000/ WX4000 (dig) (*xy*/*ty*)	WR300 WR1000 WR8500 DMS1000
Hugo Sachs Elektronik – Harvard Apparatus GmbH http://www.hugo-sachs.de	Mark VII-c (tp)	R-60	
LDS Test and Measurement LLC http://www.gouldmedical.com			TA11 TA240 TA6000 WindoGraf
Yokogawa Electric Corporation http://www.yokogawa.com/daq/daq-products.htm		3057 LR4100E/ LR4200E/ LR8100E/ LR12000E (dig) 3023/3024 (*xy*/*ty*) 3025 (*xy*/*ty*)	OR100E/ OR300E DL708E/ DL716 (ds) DL750 (ds)
Hioki USA Corporation http://www.hiokiusa.com			8826 (m) 8835-01 (m) 8841 (m) 8842 (m)
Kipp&Zonen http://www.kippzonen.com		BD11/12 SE 102/122 SE 110/111/112 SE 124 BD300 (dig) SE790 (*xy*/*ty*)	SE 520/540 SE 570
Soltec http://www.solteccorp.com		MCR 560 DCR 520 (dig) DCR 540 (dig)	TA200-938 TA200-3304 TA220-1200 TA220-3216/3208 TA220-3424 TA220-3608
Omega Engineering, Inc http://www.omega.com		142 156 555/585/595 640 RD45A/46A RD1101 RD1201/1202 RD2000 RD3720 (dig) RD6100 RD6110 RD6112 600A (*xy*/*ty*) 790/791(*xy*/*ty*) RD3020 (*xy*/*ty*)	
Linseis http://www.linseis.net		L120/200/250 L6514II L7005II LY14100II (*xy*/*ty*) LY15100II (*xy*/*ty*)	

[a]Recorders with analogue transducers with rotating pen and with translating pen (servo) and recorders with digital transducer (thermal dot array). ink: inkpen; tp: thermal pen; ds: digitals scope with chart recorder; m: memory recorder; *xy*: *x*–*y* recorder; *xy*/*ty*: *x*–*y* and *t*–*y* recorder; dig: digital signal processing.

Digital recorders are characterized by the typical features of sampling and digitization in waveform reproduction. Although the writing frequency is limited by the thermal time constant of the thermal dots, with respect to bandwidth, it does not impose a constraint on the signal frequency spectrum, because, by the use of facilities inherent to digital apparatus (including a high sampling frequency), the problem can be solved by off-line recording. Signals that vary too fast for on-line recording can be stored in memory and reproduced at a speed the recorder is able to handle. Moreover by the specific writing method where vertical dotted lines are printed varying from the minimal to the maximal value sampled within the writing interval, fast transients, disregarded by ordinary interpolation, are also captured in the graph. Although the form of the transient cannot be interpreted, the graph provides evidence of its presence. In digital recording, the addition of coordinate lines and alphanumeric information to the record presents no difficulty for the transducer, as there is no essential difference compared to recording signal tracings. As connections to computing devices are possible, processed data can also be recorded, reducing the work of visual inspection of the record. Also, in some translational servorecorders signals are processed completely digitally, but in this equipment typical transducer limitations still exist and the domain of applications (i.e., the slow varying signals) remains the same.

As digital recorders are generally sophisticated, one has to pay the price for the flexibility provided. In addition to low price, analog recording has the advantage that different-colored inks can be used, which is important in multichannel recorders with overlapping (full range) curves. Available digital recorders allow only one color, thereby reducing identifiability. As the digital writing array is inside the apparatus there is a small latency between writing and appearing of the tracings. In overlapping multipen devices pens are physically dislocated and application of the pen offset compensation also creates a time delay. In digital recorders, displays are added to provide immediate visibility and easy setting of the recorder parameters. Control of the analog recorder is generally limited and accordingly simple to perform. Digital recorders provide more facilities, implying the need of training and experience to install instrument settings for the application envisaged.

Analog and digital systems thus have their specific limitations and benefits. The choice of the equipment for a specific application is a matter of performance, operating flexibility, and price. Measurements where immediate visualization of the tracings on a long strip of paper is not required and data are to be digitally analyzed and stored, can be performed by PC with virtual instrument software and printer. Nowadays the majority of data collection is accomplished using digital PC approaches. As previously stated, besides the measuring properties, the graphic quality is extremely important, as it assists visual examination of the records. In this case the general rule is applicable. Before one chooses an apparatus, one should see it in operation, that is, making graphic representations of the data.

BIBLIOGRAPHY

1. Olsen WH. Basic concepts in instrumentation. In: Webster JG, editor. Medical Instrumentation: Application and Design. 3rd edition. New York: John Wiley & Sons; 1998. p 1–43.
2. Bentley JP. Principles of measurement systems. 3rd ed. Longman House: Longman Group Limited; 1995.
3. Sydenham PH. Static and dynamic characteristics of instrumentation. In: Webster JG, editor. The Measurement, Instrumentation and Sensors Handbook. Boca Raton (FL): CRC Press; 1999. p 3/1–3/22.
4. Miller A et al., editors. Electronic Measurements and Instrumentation. New York: Mc Graw-Hill; 1975. p 427–479.
5. Bell DA. Electronic Instrumentation and Measurements. 2nd ed. Englewood Cliffs (NJ): Prentice-Hall; 1994.

See also ELECTROCARDIOGRAPHY, COMPUTERS IN; ELECTROENCEPHALOGRAPHY; ELECTROMYOGRAPHY; PHONOCARDIOGRAPHY.

RECORDS MANAGEMENT SYSTEM. See MEDICAL RECORDS, COMPUTERS IN.

REGULATIONS FOR MEDICAL DEVICES. See CODES AND REGULATIONS: MEDICAL DEVICES.

REGULATIONS FOR RADIATION. See CODES AND REGULATIONS: RADIATION.

REHABILITATION AND MUSCLE TESTING

W.K. DURFEE
P.A. IAIZZO
University of Minnesota
Minneapolis, Minnesota

INTRODUCTION

There is a growing need in clinical medicine to validate the quantitative outcomes of an applied therapy. In addition, the measurement of muscle function is an essential component of many neurological and physical exams. Muscle strength is correlated to function, work productivity, and general quality of life. Muscle function becomes compromised: (1) as we age, (2) when associated with a skeletal impairment, and/or (3) as a secondary consequence of many disease processes. Therefore, assessing muscle function is an important clinical skill that is routinely used by neurologists, orthopedists, general practitioners, anesthesiologists, and occupational and physical therapists. Evaluation of muscle strength is used for differential diagnosis, to determine if an impairment or disability is present, to decide if a patient qualifies for treatment, and or to track the effectiveness of a treatment.

In a research setting, the measurement of muscle function is used to further our understanding of the normal and potentially impaired neuromuscular system in human and/or animal experiments. In such research, muscle function can be assessed at the intact individual level (*in vivo*), in chronic and acute animal models (*in situ*), within isolated

muscle strips or even within single myofibrils (*in vitro*), and/or at the molecular–biochemical level. In this article, only whole muscle testing (*in vivo* and *in situ*) is discussed.

There are several components of muscle performance. The American Physical Therapy Association uses various definitions to explain the characteristics of muscle function (1). Muscle performance is the capacity of a muscle to do work. Muscle strength is the force exerted by a muscle or group of muscles to overcome a resistance in one maximal effort. Instantaneous muscle power is the mechanical power produced by the muscle (muscle force times muscle velocity). Muscle endurance is the ability to contract a muscle repeatedly over time. Of these performance indicators, muscle strength is the one most commonly measured when assessing the muscle function of intact humans.

In assessing muscle strength, the conditions under which the muscle contracts must be specified so that the muscle test data can be interpreted properly. The following conditions are relevant: "Isometric contraction": the muscle contracts while at a fixed length; "Isotonic contraction": the muscle contracts while working against a fixed load, for example, a hanging weight; "Isokinetic contraction": the muscle contracts while moving at a constant velocity; (generally, isokinetic contractions are only possible with the limb strapped into a special machine that imposes the constant velocity condition); "Eccentric contraction": the muscle contracts against a load that is greater than the force produced by the muscle so that the muscle lengthens while contracting; and "Concentric contraction": the muscle contracts against a load that is less than the force produced by the muscle so that the muscle shortens while contracting.

Isometric muscle tests are the most common as they are the simplest to perform and reproduce and, because the test conditions are well defined, they are the most appropriate for comparing results within a population. Two considerations are important when testing muscle under isometric conditions. First, because muscle force varies with muscle length, the length of the muscle must be specified when planning and reporting a muscle test. For example the manual muscle test has strict and well-defined rules for the subject's posture and joint positions that must be followed if one is to make clinical decisions based on the test (2).

Second, all isometric muscle tests of intact human muscle are conducted with the limb either held in a fixed position by the examiner, or with the limb fixed to a brace or jig (see the Stimulated Muscle Force Assessment section). While these methods hold the limb in a fixed position, the muscle will not be strictly isometric because of tendon stretch. The mismatch between limb condition and muscle condition only causes problems when trying to infer details about muscle dynamics, such as rise time or contraction speed from externally measured forces. Even if the whole muscle could be fixed at proximal and distal ends, during a twitch, the distance between z lines in the myofibril will shorten, which means the sarcomeres are shortening due to internal muscle elasticity. This is why the length tension and dynamic properties of whole muscle deviate somewhat from those of the isolated sarcomere. Nevertheless, length tension or ankle–angle/isometric–torque analyses can be done *in vivo* (3,4).

Figure 1. Muscles wrap around joints. Muscle force is related to external force produced by a limb through skeletal geometry of joints and attachment points.

Testing of intact human muscle requires that muscle output be measured external to the body and, as a result, muscle force is never measured directly. As shown in Fig. 1, muscles wrap around joints and attach to limbs at the proximal and distal ends. There is a kinematic relationship between the measured force and the actual muscle force that depends on the details of muscle attachment and varies with joint angle. Therefore, to ultimately solve the kinematic relationship, one will require information about muscle attachment location, the geometry of the joint, and the joint angle. Such geometric information can be readily obtained from a magnetic resonance imaging (MRI) scan or a more generic geometry can be assumed, for example, obtained dimensions gathered from cadaver studies (5,6).

While often reported as a force, external testing of muscles more correctly should be reported as a torque. Figure 2 illustrates how force varies with location of the resistive load along the limb, while torque does not. Reporting muscle strength as torque about a joint eliminates this difficulty. If force is reported, the distance between the

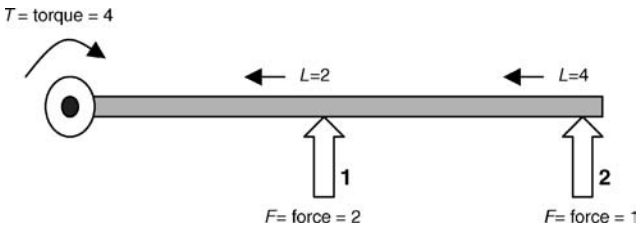


Figure 2. A torque of 4 is produced about a joint. It takes an opposing force of 2 to balance the torque if the opposing force is applied at arrow 1. If the opposing force is applied at arrow 2 it only takes a force of 1 to balance the torque. Thus, the perceived torque, and therefore the scoring of muscle strength, depends on where along the limb the examiner places his/her hand.

joint and the resistive load point should be measured to permit conversion to torque.

External measurement of torque about a limb joint means that all of the forces acting on that joint are measured, and that the contribution of the muscle or muscle group under study cannot be easily separated out. In other words, there are confounding forces generated by synergistic muscles. For example, when testing foot plantar flexion to determine gastrocnemius strength, the soleus may also be contributing to the measured torque about the ankle. Yet, another complicating factor may be undesired activations of antagonist muscles. One example is when you "flex your arm muscles". In general, the resulting torque from the biceps and triceps are in balance, the arm does not move, and no external torque will be measured even though the muscles are contracting actively.

MANUAL MUSCLE TEST

The simplest and most common method of assessing muscle strength is the manual muscle test (MMT). Manual muscle testing is a procedure for evaluating strength and function of an individual muscle or a muscle group in which the patient voluntarily contracts the muscle against gravity load or manual resistance (2,7). It is quick, efficient, and easy to learn, however, it requires total cooperation from the patient and learned response levels by the assessor.

The procedures for conducting the MMT have been standardized to assure, as much as possible, that results from the test will be reliable (2,7,8). The specific muscle or muscle group must be determined and the examiner must be aware of, and control for, common substitution patterns where the patient voluntarily or involuntarily uses a different muscle to compensate for a weak muscle being tested.

To conduct a MMT, the patient is positioned in a posture appropriate for the muscle being tested, which generally entails isolating the muscle and positioning so that the muscle works against gravity (Fig. 3). The body part proximal to the joint acted on by the muscle is stabilized. A screening test is performed by asking the patient to move the body part through the full available range of motion (ROM). The main test is then performed either unloaded, against a gravity load, or against manual resistance, and a grade is assigned to indicate the relative muscle strength.

Figure 3. Manual muscle test of the iliopsoas.

Table 1. Manual Muscle Test Scores[a]

Score	Description
0	No palpable or observable muscle contraction
1	Palpable or observable contraction, but no motion
1+	Moves limb without gravity loading less than one-half available ROM[b]
2−	Moves without gravity loading more than one-half ROM[b]
2	Moves without gravity loading over the full ROM[b]
2+	Moves against gravity less than one-half ROM[b]
3−	Moves against gravity greater than one-half ROM[b]
3	Moves against gravity less over the full ROM[b]
3+	Moves against gravity and moderate resistance less than one-half ROM[b]
4−	Moves against gravity and moderate resistance more than one-half ROM[b]
4	Moves against gravity and moderate resistance over the full ROM[b]
5	Moves against gravity and maximal resistance over the full ROM[b]

[a]Adapted from Ref. 2
[b]ROM = range of motion.

Manual grading of muscle strength is based on palpation or observation of muscle contraction, ability to move the limb through its available ROM against or without gravity, and ability to move the limb through its ROM against manual resistance by the examiner. Manual resistance is applied by the examiner using one hand with the other hand stabilizing the joint. Exact locations for applying resistive force are specified and must be followed exactly to obtain accurate MMT results (2). A slow, repeatable velocity is used to take the limb through its ROM, applying a resistive force just under the force that stops motion. The instructions to the patient are, use all of your strength to move the limb as far as possible against the resistance. For weaker muscles that can move the limb, but not against gravity, the patient is repositioned so that the motion is done in the horizontal plane with no gravity.

Grades are assigned on a 0–5 scale with ± modifiers (1 = trace score, 2 = poor, 3 = fair, 4 = good, 5 = normal) (Table 1). Grades > 1 demonstrate motion, and grades >3 are against manual resistance. Other comparable scoring scales exist (7).

As noted above, importantly, the assignment of scores is based on clinical judgment and the experience of the examiner. The amount of resistance (moderate, maximal) applied by the examiner is also based on clinical experience and is adjusted to match the muscle being tested as well as the patient's age, gender and/or body type.

A common alternative to motion-based MMT is the isometric MMT in which the limb is held in a fixed position while the examiner gradually applies an increasing resistance force. The instructions to the patient are, Don't let me move you. The amount of force it takes to "break" the patient is used to assign a score. Scoring norms for isometric MMT are provided in Table 2. While the MMT is the most widely used method to assess muscle function, its reliability and accuracy can be questionable (9,10). The MMT

Table 2. Grading of Isometric Manual Muscle Test[a]

Score	Description
3	Maintains position against gravity
3+	Maintains position against gravity and minimal resistance
4–	Maintains position against gravity and less than moderate resistance
4	Maintains position against gravity and moderate resistance
5	Maintains position against gravity and maximal resistance

[a]Adapted from Ref. 2.

scores are least accurate for higher force levels (9,11,12). Interrater reliability for MMT is not high, suggesting that the same examiner should perform multiple tests on one subject or across subjects (2). While not entirely accurate, MMT scores do correlate well with results from handheld dynamometers (13), implying that both are valid measures of muscle strength. However, as explained in a later section, all tests based on voluntary activation of a muscle are prone to artifact because of patient motivation and examiner encouragement.

APPARATUS

The appeal of the MMT is that it can be performed simply with the patient, an examiner, and a bench or table. This makes it ideal for the routine clinical environment where specialized equipment is unavailable and time is short. It is also suited for situations in which testing must be performed away from the clinic, for example, in nursing homes, rural areas, or remote emergency settings.

When greater accuracy of results is needed, instruments are available that provide precise readouts of the resistive force the muscle works against (14). One example is a handheld dynamometer, such as the one shown in Fig. 4. This instrument can be sandwiched between the examiner's hand and the patient's limb, and provides a "readout" of force. The interrater reliability for handheld dynamometers is good when used with a standard procedure (15–17), as is the test-retest reliability (13).

Other products have been developed for specific tests of muscle strength, for example, the hand dynamometer and pinch grip devices shown in Fig. 5. These are easy to use and common for diagnostic tests of the hand. Despite their quantitative nature, readings between different types and brands of dynamometers can vary (18,19).

Computer-controlled dynamometers offer a variety of loading conditions for muscle testing and for strengthening treatments (20,21) (Fig. 6). Along with isometric and isotonic loading, dynamometer machines provide isokinetic conditions in which the muscle group acts against a computer-controlled resistance that moves the limb at a constant angular velocity.

ADVANCED MUSCLE ASSESSMENT METHODS

Measuring Muscle Dynamics

Muscle is a complex actuator whose external properties of force and motion result from the action of thousands of muscle fibers which, in turn, result from the action of

Figure 4. Handheld dynamometer. Pictured is the Lafayette manual muscle test system from Lafayette instrument company (Lafayette, IN).

Figure 5. The Jamar hand dynamometer is pictured on the left (NexGen Ergonomics, Quebec, Canada), and the B & L Pinch Gauge is shown on the right (B & L Engineering, Tustin, CA).

Figure 6. Biodex dynamometer for computer-controlled muscle testing (Biodex Medical Systems, Shirley, NY).

millions of structural and active proteins whose interaction is triggered by biochemical events. While most muscle testing focuses on the overall strength of a muscle or muscle group, more sophisticated assessment can be useful for in-depth examination of muscle function, including its dynamic, kinematic and fatigue properties.

The approach used to measure muscle function in more detail involves developing a mathematical model of muscle activity and then using experiments to identify the parameters of the model. Overviews of these methods are provided in Zajac and Winters (22), Durfee (23), Zahalak (24), Crago (25), and Kearney and Kirsch (26). The modeler must first choose the appropriate complexity of the mathematical model. The optimum choice is a model that is sufficiently complex to reveal the behavior of interest, but not so complex that parameters cannot be identified. Generally, Hill-type input–output models (27,28) are a good balance, as they capture key force–velocity, force–length, and activation dynamics at a whole muscle level (Fig. 7).

Model parameters can be identified one at a time, using the approach followed by Hill (27), or all at once using modern system identification techniques (23,26). Electrical activation of the muscle is a particularly convenient means for excitation because, unlike voluntary activation, there is control over the input, an essential component for an effective system identification method. Testing can be done under isometric conditions for determining recruitment and twitch dynamic characteristics, or under arbitrary loading to find active and passive force-length and force-velocity properties.

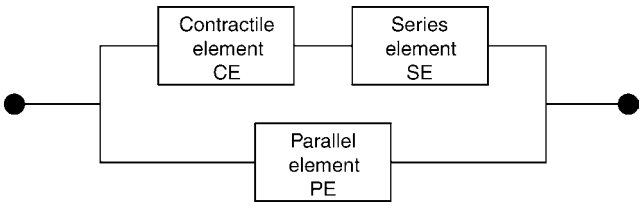


Figure 7. Hill muscle model. The contractile element (CE) contains the active element with dynamics, force–velocity and force–length properties. The series element (SE) is the inherent internal elastic elements, and the parallel element (PE) represents passive connective tissue.

Identification of muscle properties is most easily accomplished using isolated muscle in acute animal model studies. Here the muscle is unencumbered by joint attachments and extraneous passive tissue. Muscle tendon can be directly attached to a force sensor and placed in a computer-controlled servo mechanism to apply known length and velocity trajectories, all while being stimulated. For example, the isometric recruitment curve, the relationship between muscle force and stimulus strength, can be identified using either point-at-a-time or swept amplitude methods, the latter being efficient in implementation (29). Using the model shown in Fig. 8, active and passive force–length and force–velocity properties can be estimated using brief bouts of controlled, random length perturbations, and then verified through additional trials where both stimulation and length are varied randomly (23,30) (Fig. 9). Simultaneous identification of active and passive muscle properties for intact human muscles is more challenging and represents an ongoing area of research (26).

Electromyogram

Contracting skeletal muscle emits an electrical signal, the electromyogram (EMG). Electrical recording of the EMG using needle or surface electrodes is an important diagnostic indicator used in clinical neurology to diagnose neuromuscular disorders including peripheral neuropathies, neuromuscular junction diseases, and muscular dystrophies. The EMG is also used in research as an estimator of muscle activity for biomechanics and motor control experiments. The reader is referred to Merletti and Parker (31) and Basmajian and DeLuca (32) for a comprehensive discussion of surface and needle EMG used in research applications, and to Preston and Shapiro (33), Kimura (34), and Gnatz (35) for an introduction to clinical EMG. Nevertheless, these assessment approaches require volitional activation of the patient's musculature under investigation.

STIMULATED MUSCLE FORCE ASSESSMENT

As described above, most devices used clinically to quantitate force and increase objectivity still rely on voluntary effort,which can be problematic. Pain, corticospinal tract lesions, systemic illness, and inconsistent motivation can significantly affect voluntarily activated muscle force. In addition, some neurologically impaired patients have difficulty maintaining a constant velocity of limb movement, and some very weak patients are unable to complete a full range of motion in voluntary force assessment tasks (36,37). As a specific example, monitoring muscle function in patients confined to the intensive care unit is a difficult challenge. Often such patients are on potent pain medications (e.g., morphine) and/or are sedated, or may have significant alterations in levels of consciousness due to critical illness (38,39). Thus, it can be extremely difficult to ask such patients to provide reproducible voluntary

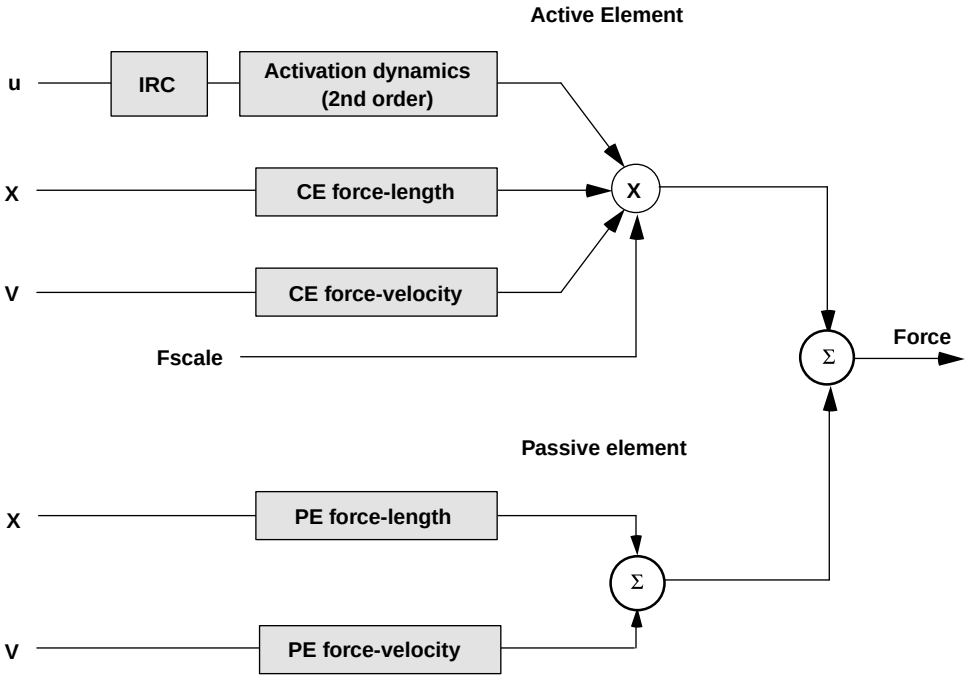


Figure 8. A model that can be used for muscle property identification. The active element has recruitment and twitch dynamics that multiplicatively combine with active force–length and force–velocity properties and sum with passive force–length and force–velocity properties to produce overall muscle force. (For details, see Refs. 23 and 30). IRC = isometric recruitment curve; CE = contractile element; PE = parallel element.

efforts. Even when cooperation is good, most measures of force assessment are qualitative, similar to using a hand-held unit for testing neuromuscular blockade.

Stimulated muscle force assessment is a versatile approach for quantitative involuntary muscle torque in humans. A muscle is activated by noninvasive nerve or motor point stimulation. A rigid apparatus is used to secure the appropriate portion of the subject's body in a predetermined position that confines movement to a specific direction, for example, ankle dorsiflexion, thumb adduction, arm flexion, or neck flexion (3,4,40,41) (Figs. 10 and 11). The innervating nerves or the motor points of the muscle are stimulated using surface electrodes, with either a single stimulus to generate a twitch contraction or with short trains of stimuli to produce tetanic contractions (e.g., 5 ms interpulse intervals) (3,4). Incorporated strain gauges are used to measure isometric torque and, via acquisition software, all data are immediately displayed and on-line analyses are performed. Various parameters of the obtained isometric contractions are measured, for example, time between stimulus and torque onset, peak rate of torque development, time to peak torque, half-relaxation time, and other observed changes (Fig. 12; Table 3).

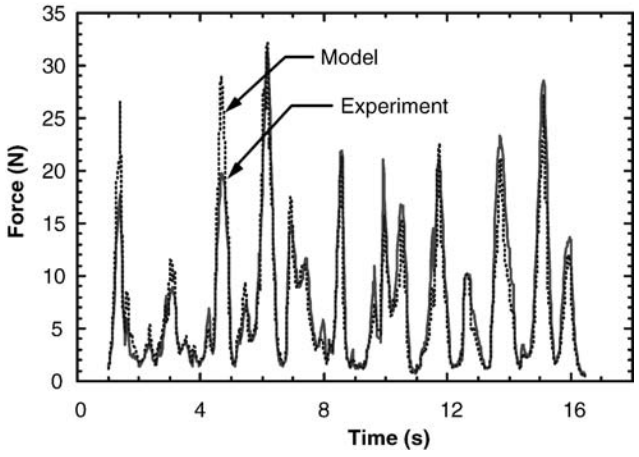


Figure 9. Results from an isolated muscle experiment where muscle active and passive properties were identified, then the resulting model was verified against experiment data. Data were generated while the muscle underwent simultaneous, random, computercontrolled stimulation and length perturbations (30).

Such information is predicted to correlate with underlying physiological conditions and/or the presence of a myopathic or neuropathic disorder. To date, the average torque generated by healthy control subjects varies by $<5\%$ with repeated testing for contractions elicited from the various muscle groups studied (4,40,41). Thus, this assessment approach has potential utility in a number of research arenas, both clinical and nonclinical. Specifically, it has added clinical value in diagnosing a neuromuscular disorder, tracking weakness due to disease progression, and/or quantitatively evaluating the efficacy of a therapy (37,42–45). Compared to current assessment methods, we consider that monitoring isometric muscle torque generated by stimulation improves objectivity, reliability, and quantitative capabilities, and increases the type of patients that can be studied, including those under sedation (39) (Fig. 10). Stimulated muscle force assessment may be of particular utility in studying patients with a known underlying genetic disorder, for it could then provide important information as to genotype–phenotype associations (37).

The general configuration of the measurement system consists of the following main components: a stabilizing device that holds either the subject's arm or leg in a defined position; a force transducer that detects the evoked torque produced by a specific muscle group; hardware devices for nerve stimulation, signal amplification,

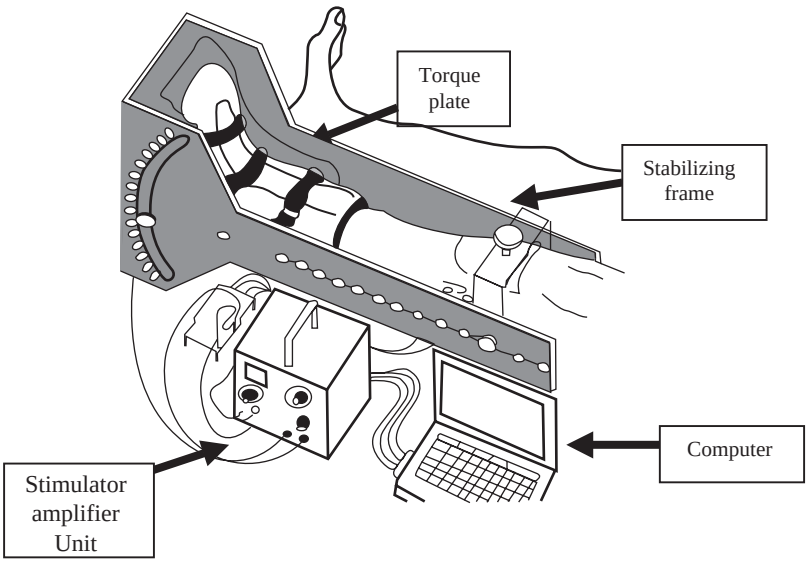


Figure 10. Muscle force assessment system to determine involuntary isometric torque of the human dorsiflexor muscles. It is comprised of the following main components: (1) a stabilizing frame with knee supports; (2) the torque plate with mounted boot to fix the foot which can be rotated between -40° and 40°; (3) a strain gauge system (Wheatstone bridge circuit) that detects the evoked torque; (4) a stimulator–amplifier unit that can supply variable stimulus pulse amplitudes and pulse durations and can amplify the voltage changes from the Wheatstone bridge circuit; and (5) a computer with data acquisition hardware and software for recording, analyzing and displaying all signals. (Modified from Ref. 39.)

and signal conditioning; a computer for stimulus delivery; and data acquisition software for recording, analyzing, and displaying all signals simultaneously (torque, EMG, applied stimulus) (Figs. 10–12).

The stabilizing device system currently used to study the dorsiflexors is a modification of a previously described apparatus (3). This device can be configured to maintain the subject's leg in a stable position while allowing access

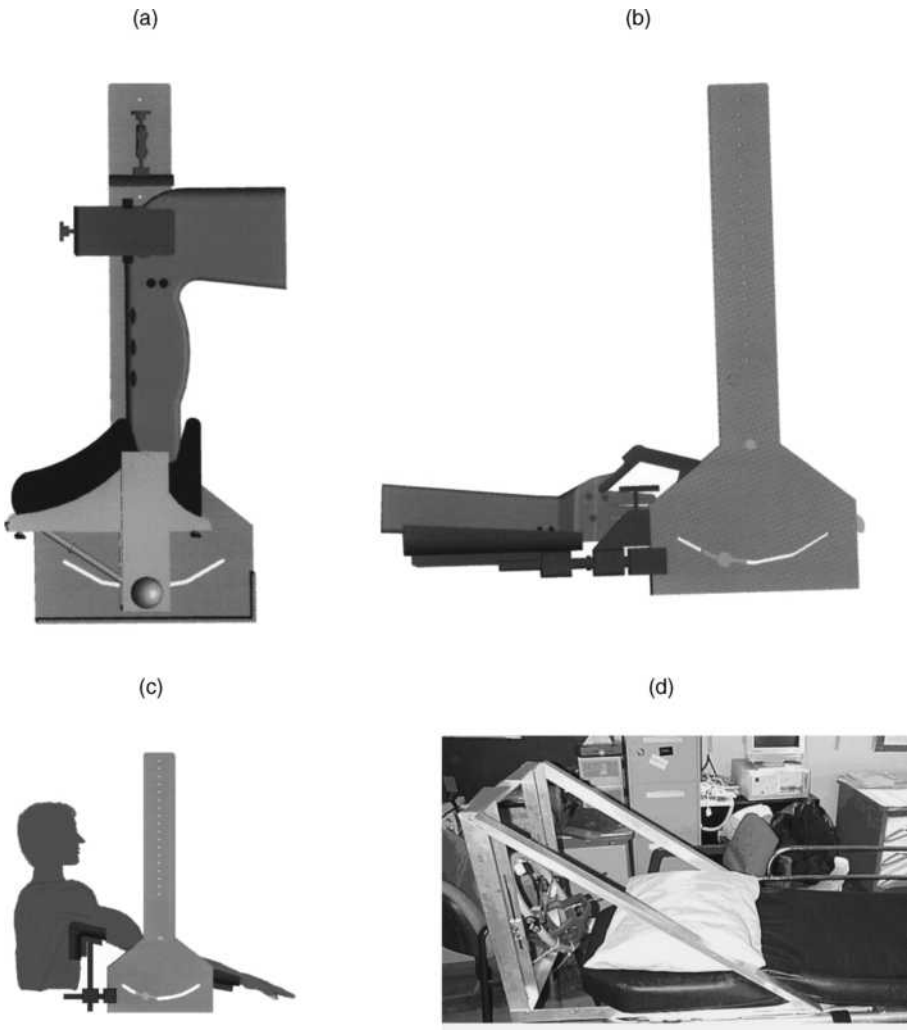


Figure 11. Various applications of stimulated muscle force assessment: (a) dorsiflexor muscles in a seated individual with stimulation of the common peroneal nerve lateral to the fibular head; (b) adductor pollicus muscle following ulnar nerve stimulation; (c) activated biceps force with motor point stimulation; and (d) head stabilizing/force system to study sternocleidomastoid muscle function following motor point stimulation. (See also Refs. 4, 40, and 41.)

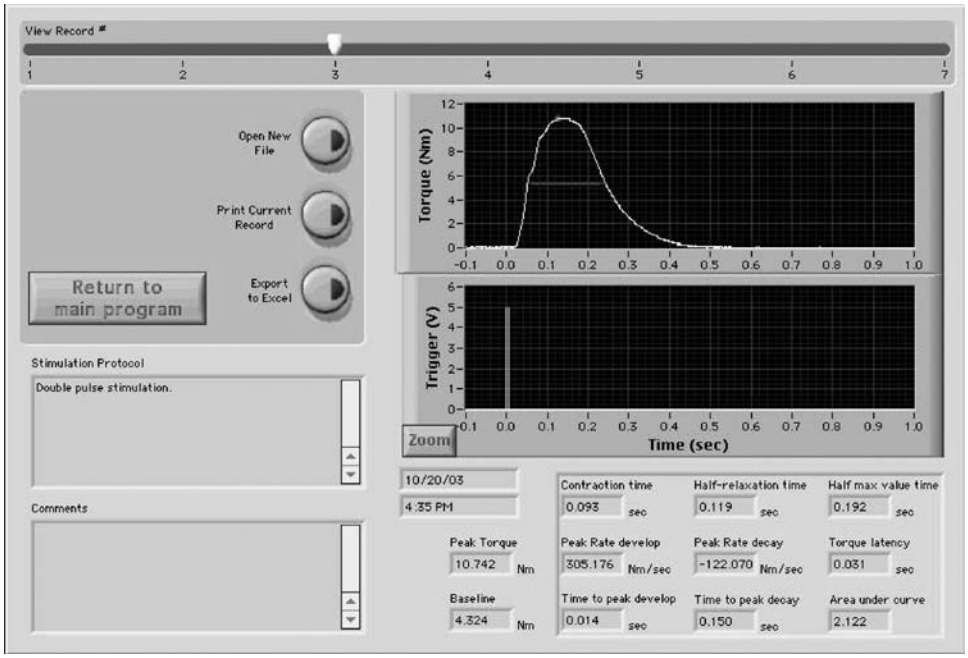


Figure 12. An example of a typical data display available to the investigator during subsequent off-line analyses. Graphically displayed are the muscle torque waveform and the stimulus administered (a double pulse with a 5 ms interpulse interval). Numerically displayed are various contractile parameters. Time 0 is the time of stimulation. The horizontal line indicates the time when half of the peak torque has been generated. The display shown is the torque generated by the dorsiflexor muscles of a normal, healthy subject.

for stimulation of the common peroneal nerve lateral to the fibular head (Fig. 11a). The torque about the ankle joint, produced by the dorsiflexor muscles (i.e., primarily generated by the tibialis anterior with contributions from the peroneus tertius and extensor digitorum muscles), is then quantified. One or two padded adjustable clamps can be used to maintain stability of the leg (knee slightly flexed in a supine position or flexed at 90° while seated). Modified in-line skate boots of varying sizes are affixed to the torque plate and adapted for either the right or left foot. The foot and ankle can be rotated within a 40° range while secured in the skate boot. This device can also be used for subjects in a supine position, in which case the support frame is secured to the upper leg proximal to the knee (Fig. 10) (39,45). To emphasize the extreme versatility of this methodology, note that a specialized version of this device was constructed and used to study dorsiflexor torques in hibernating black bears, Ursus Americanus, in the Rocky Mountains (46).

Briefly, the arm and hand stabilizing apparatus, used to measure muscle torque of the adductor pollicis, is easily attached to the main stabilizing frame (Fig. 11b). Using straps, the forearm can be secured to the arm stabilizing unit, which can be adjusted for varying arm lengths. The digits (2–5) are placed in the hand well, and the thumb is secured to the constructed thumb bar attached to the torque plate. Shown in Fig. 11c is the configuration that is used to record force generated by the biceps muscle. As for the aforementioned muscles, force can be produced by peripheral nerve stimulation or by voluntary effort. In addition, we have successfully employed motor point stimulation of the muscle itself with large surface electrodes (40). The forces of the isometric muscle contractions are obtained as changes in torque applied to the instrument torque plate. Finally, we recently reported the optimization of an approach to study forces in the sternocleidomastoid muscle in the anterior neck (Fig. 11d), and plan to use these methodologies to study the effect of therapy in patients with cervical dystonia.

To date, this assessment approach has been used to study patients with a wide variety of disorders including: amyotrophic lateral scoliosis, Brody's disease, chronic inflammatory demyelinating polyneuropathy, malignant hyperthermia, muscular dystrophy, myotonia, periodic

Table 3. Contractile Parameters that Can Easily be Quantified

Parameter	Units[a]	Definition
Peak torque	N·m	Maximum amount of torque developed
Contraction time	s	Time from onset of torque to time of peak torque (e.g., calculated at 90% of peak)
Half-relaxation time	s	Time from peak torque to time when torque decays to half of peak torque
Peak rate of development	N·m/s	Maximum rate of torque development
Peak rate of decay	N·m/s	Maximum rate of torque decay
Time to peak development	s	Time from onset of torque to the peak rate of development
Time to peak decay	s	Time from peak rate of development to peak rate of decay
Half-maximal duration	s	Time when the generated torque is maintained at a level of half of the peak torque
Latency to onset	s	Time from the stimulus to the onset of torque development

[a]N·m = newton meters.

paralysis, and nerve conduction blocks. The new insights to be gained by employing this approach in a variety of healthcare situations will further our clinical understanding of the underlying pathophysiologies, and provide us an accurate means to determine clinical outcomes. Recently, we employed this approach to evaluate athletes with potential overtraining syndrome (47), thus the applications for these methodologies could be considered quite limitless.

SUMMARY

The assessment of a patient's muscle strength is one of the most important vital functions that is typically monitored. Specifically, strength assessment is necessary for determining distribution of weakness, disease progression, and/or treatment efficacy. The particular assessment approach will be, in part, dictated by the clinical circumstance or severity of illness. Several assessment techniques and tools are currently available to the healthcare provider and/or researcher, yet each has its unique attributes. Nevertheless, as outcomes-based medical practice becomes the norm, the need for quantitative outcomes assessment of muscle strength will become even more important.

BIBLIOGRAPHY

1. American Physical Therapy Association. Guide To Physical Therapy Practice. Alexandria, Virginia: American Physical Therapy Association; 1997.
2. Clarkson HM. Musculoskeletal Assessment: Joint Range of Motion and Manual Muscle Strength. Philadelphia: Lippincott Williams & Wilkins; 2000.
3. Quinlan JG, Iaizzo PA, Gronert GA, Lambert EH. Ankle dorsiflexor twitch properties in malignant hyperthermia. Muscle Nerve 1989;12:119–125.
4. Brass TJ, Loushin MK, Day JW, Iaizzo PA. An improved method for muscle force assessment in neuromuscular disease. J Med Eng Technol 1996;20:67–74.
5. Winter DA. Biomechanics of Human Movement. New York: John Wiley & Sons; 1979.
6. Yamaguchi GA. A survey of human musculotendon actuator parameters. Winters JM, Woo S, editors. Multiple Muscle Systems: Biomechanics and Movement Organization 1990; New York: Springer-Verlag.
7. Neistadt M, Crepeau E. Willard and Spackman's Occupational Therapy. Philadelphia: Lippincott Williams & Wilkins; 1998.
8. Daniels L, Worthingham C. Muscle Testing: Techniques of Manual Examination. Philadelphia: W.B. Saunders; 1995.
9. Iddings DM, Smith LK, Spencer WA. Muscle testing. 2. Reliability in clinical use. Phys Ther Rev 1961;41:249–256.
10. Schwartz SM, Cohen ME, Herbison GJ, Shah A. Relationship between two measures of upper extremity strength: Manual muscle test compared to hand-held myometry. Arch Phys Med Rehabil 1992;73:1063–1068.
11. Bohannon RW. Manual muscle test scores and dynamometer test scores of knee extension strength. Arch Phys Med Rehabil 1986;67:390–392.
12. Knepler C, Bohannon RW. Subjectivity of forces associated with manual-muscle test grades of 3 +, 4 −, and 4. Percept Mot Skills 1998;87:1123–1128.
13. Bohannon RW. Test-retest reliability of hand-held dynamometry during a single session of strength assessment. Phys Ther 1986;66:206–209.
14. Amundsen L. Muscle Strength Testing: Instrumented and Non-Instrumented Systems. New York: Churchill Livingstone; 1990.
15. Bohannon RW, Andrews AW. Interrater reliability of hand-held dynamometry. Phys Ther 1987;67:931–933.
16. Horvat M, Croce R, Roswal G. Intratester reliability of the Nicholas Manual Muscle Tester on individuals with intellectual disabilities by a tester having minimal experience. Arch Phys Med Rehabil 1994;75:808–811.
17. Dunn J, Iversen M. Interrater reliability of knee muscle forces obtained by hand-held dynamometer from elderly subjects with degenerative back pain. J Geriat Phys Ther 2003; 26:23–29.
18. Click Fenter P, Bellew JW, Pitts T, Kay R. A comparison of 3 hand-held dynamometers used to measure hip abduction strength. J Strength Cond Res 2003;17:531–535.
19. Massy-Westropp N, et al. Measuring grip strength in normal adults: reference ranges and a comparison of electronic and hydraulic instruments. J Hand Surg [Am] 2004;29:514–519.
20. Cabri JM. Isokinetic strength aspects of human joints and muscles. Crit Rev Biomed Eng 1991;19:231–259.
21. Dvir Z. Isokinetics: Muscle Testing, Interpretation and Clinical Applications. New York: Churchill Livingstone; 1995.
22. Zajac F, Winters JM. Modeling musculoskeletal movement systems. Winters JM, Woo S, editors. Multiple Muscle Systems: Biomechanics and Movement Organizations. New York: Springer-Verlag; 1990.
23. Durfee WK. Model identification in neural prosthesis systems. In: Stein RB, Peckham PH, Popovic D, editors. Neural Prostheses: Replacing Motor Function After Disease or Disability. New York: Oxford University Press; 1992.
24. Zahalak G. An overview of muscle modeling. In: Stein RB, Peckham PH, Popovic D, editors. Neural Prostheses: Replacing Motor Function After Disease or Disability. New York: Oxford University Press; 1992.
25. Crago P. Creating neuromusculoskeletal models. In: Winters JM, Crago P, editors. Biomechanics and Neural Control of Posture and Movement. New York: Springer-Verlag; 2000.
26. Kearney R, Kirsch R. System identification and neuromuscular modeling. In: Winters JM, Crago P, editors. Biomechanics and Neural Control of Posture and Movement. New York: Springer-Verlag; 2000.
27. Hill AV. The heat of shortening and the dynamic constants of muscle. Proc R Soc Lond [Biol] 1938;126:136–195.
28. Winters JM. Hill-based muscle models: A systems engineering perspective. In: Winters JM, Woo S, editors. Multiple Muscle Systems: Biomechanics and Movement Organizations. New York: Springer-Verlag; 1990.
29. Durfee WK, MacLean KE. Methods for estimating isometric recruitment curves of electrically stimulated muscle. IEEE Trans Biomed Eng 1989;36:654–667.
30. Durfee WK, Palmer KI. Estimation of force-activation, force-length, and forcevelocity properties in isolated, electrically stimulated muscle. IEEE Trans Biomed Eng 1994;41:205–216.
31. Merletti R, Parker P. Electromyography: Physiology, Engineering, and Noninvasive Applications. Hoboken: John Wiley & Sons; 2004.
32. Basmajian JV, DeLuca CJ. Muscles Alive: Their Function Revealed by Electromyography. Baltimore, MD: Williams & Wilkins; 1985.
33. Preston DC, Shapiro BE. Electromyography and Neuromuscular Disorders: Clinical-Electrophysiologic Correlations. Boston: Butterworth-Heinemann; 1998.
34. Kimura J. Electrodiagnosis in Diseases of Nerve and Muscle: Principles and Practice. New York: Oxford University Press; 2001.
35. Gnatz SM. EMG Basics. Austin, TX: Greenleaf Book Group; 2001.

36. Fillyaw MJ, Tandan R, Bradley WG. Serial evaluation of neuromuscular function in management of chronic inflammatory demyelinating polyneuropathy. Phys Ther 1987;67: 1708–1711.
37. Day JW, et al. Force assessment in periodic paralysis after electrical muscle stimulation. Mayo Clin Proc 2002;77: 232–240.
38. Jackson AC, Gilbert JJ, Young GB, Bolton CF. The encephalopathy of sepsis. Can J Neurol Sci 1985;12:303–307.
39. Ginz HF, Zorzato F, Iaizzo PA, Urwyler A. Effect of three anaesthetic techniques on isometric skeletal muscle strength. Br J Anaesth 2004;92:367–372.
40. Hong J, Iaizzo PA. Force assessment of the stimulated arm flexors: quantification of contractile properties. J Med Eng Technol 2002;26:28–35.
41. Hong J, Falkenberg JH, Iaizzo PA. Stimulated muscle force assessment of the sternocleidomastoid muscle in humans. J Med Eng Technol 2005;29:82–89.
42. Quinlan JG, Iaizzo PA, Gronert GA, Lambert EH. Twitch responses in a myopathy with impaired relaxation but no myotonia. Muscle Nerve 1990;13:326–329.
43. Quinlan JG, Wedel DJ, Iaizzo PA. Multiple-pulse stimulation and dantrolene in malignant hyperthermia. Muscle Nerve 1990;13:904–908.
44. Schulte-Mattler WJ, et al. Increased metabolic muscle fatigue is caused by some but not all mitochondrial mutations. Arch Neurol 2003;60:50–58.
45. Ginz HF, et al. Decreased isometric skeletal muscle force in critically ill patients. Swiss Med Weekly 2005; (in press).
46. Harlow HJ, Lohuis T, Beck TD, Iaizzo PA. Muscle strength in overwintering bears. Nature (London) 2001;409:997.
47. Nelson MF, Day SL, Sandell EN, Iaizzo PA. Quantitative analyses of dorsiflexor forces in marathon runners—A pilot study. J Orthop Sports Phys Ther (in review).

See also BIOMECHANICS OF EXERCISE FITNESS; ELECTROMYOGRAPHY; JOINTS, BIOMECHANICS OF; STRAIN GAGE.

REHABILITATION, COMPUTERS IN COGNITIVE

BEATRIZ C. ABREU
GARY SEALE
RICHARD O. TEMPLE
ARCHANA P. SANGOLE
Transitional Learning Center
at Galveston
Galveston, Texas

INTRODUCTION

The aim of this chapter is to provide an overview of current issues involving the use of computers in cognitive rehabilitation. The chapter begins with a brief historical review of computer use with a variety of disabilities including brain injury, learning disability, psychiatric disorders, and dementias. It continues to address selected research findings on the use of virtual reality for rehabilitation of impairments in attention, memory, and functional daily living activities. Finally, the chapter ends with conclusions and ethical reflections on using computers in research and direct care practice.

Impairments in cognitive function frequently occur as a result of acquired brain injury (i.e., trauma, cerebrovascular accidents, anoxic encephalopathy, meningitis), specific learning disabilities, mental illness, Alzheimer's disease and other causes of dementia, and as a result of the natural aging process. These cognitive impairments can include, but are not necessarily limited to, decreased attention/concentration, memory, problem-solving and decision-making, planning, and sequencing. These impairments can negatively impact learning and skill acquisition, interfere with the ability to engage in everyday activities, preclude participation in social engagements, and hinder quality of life.

Cicerone et al. (1) and Giaquinto and Fioro (2) have defined cognitive rehabilitation as the systematic application of interventions to remediate or compensate for cognitive deficits and improve abilities in daily living skills and problem-solving. Cognitive rehabilitation interventions teach an individual to appropriately attend to, select, understand, remember relevant information, and apply the information appropriately in order to engage in meaningful daily activities and solve problems that occur in our complex society. Successfully completing daily activities and solving novel problems supports participation in meaningful societal roles such as breadwinner, husband, wife, and parent. Cognitive rehabilitation is generally carried out as part of a service delivery system that is interdisciplinary in nature (3). Services delivered are tailored to individual needs, are relevant to the person receiving the services, and bring about change that impacts daily functioning (1,2).

Over the past three decades, the personal computer has been employed as a tool to deliver interventions to remediate or compensate for cognitive deficits. Computers offer a number of advantages over traditional methods of cognitive remediation (i.e., flexibility, data collection, accessibility, portability, and cost), and, while not a treatment approach in and of themselves, computers can be a powerful tool to enhance the efforts of educators and clinicians.

COGNITIVE TRAINING FOR PERSONS WITH BRAIN INJURY

Computer-assisted cognitive retraining (CACR) for persons with acquired brain injury (ABI) began with the use of standard video games as an adjunct to traditional approaches to address deficits in attention/concentration, visual scanning, information processing speed, and divided attention (4). Lynch (4) reported that the initial use of video games for cognitive retraining was appealing to both clinicians and ABI survivors, as the games were inexpensive and widely available, and were interesting and motivating to the user. However, despite improvements on neuropsychological measures of basic cognitive skills reported in early single subject and small group pre-post design experiments using CACR (4,5), little carry over into everyday activities occurred. Other limitations existed as well, such as the inability to modify computer games for individual use and score user performance in a consistent and meaningful way (6). For these reasons, computer games gave way to educational programs that were developed for drills and practice of basic academic skills (i.e., vocabulary, math skills, and simple problem-solving/decision-making). However, the commercially produced educational software was

not without limitations, primarily the inability to easily modify the program to meet specific needs of an individual user or clinician. By the mid-to-late 1980s, specific computer software was developed for CACR with the ABI population. Some software programs addressed a number of cognitive skills (such as attention, memory, and sequencing) in a "package" (2,7). These programs allowed the clinician to vary levels of task complexity and individualize treatment by adjusting the speed of stimulus presentation, the delivery of cues/prompts, establishing reinforcement schedules, and so on. Results of studies using CACR to address specific deficits in attention/concentration (8,9), memory (10), visual processing (11), and visual scanning (12) also appeared in the literature.

Today, computers are used as assistive devices to help persons with cognitive deficits complete essential daily activities such a remembering appointments and items on a to-do list, and to overcome specific limitations such as difficulty speaking (voice synthesizer). Virtual environments also allow persons to practice skills in a safe, simulated environment (13,14). Weiss et al. (13) used a PC-based VR system to train stroke patients with unilateral spatial neglect to practice crossing a typical city street safely. Zhang et al. (14) developed a PC-based virtual kitchen allowing persons with acquired brain injury to practice cooking a meal. The computer program provided prompts as needed to assist the user to sequence and complete the task correctly. The use of virtual reality in cognitive rehabilitation is discussed in greater depth later on in the chapter.

COGNITIVE TRAINING FOR STUDENTS WITH LEARNING DISABILITIES

Personal computers appeared in the classroom in the late 1970s and early 1980s (15). Computer-assisted instruction (CAI) presented information in the three primary modalities: drill and practice, simulation, and tutorials. Drill and practice programs provided a single question or problem and elicited a response from the student. The student's answers were met with feedback from the computer program, followed by another question or problem. Simulation programs were more complex, requiring the student to process information on more than one level simultaneously and engage in a decision-making process. The computer program provided cues and feedback to assist the student in reaching the correct answer. Tutorial programs simulated the traditional style of classroom education delivered by most teachers. The information or content material was presented to the student. The computer program asked questions of the student, provided cues, prompts, and feedback as needed until the material was mastered. For some teachers and administrators, computers were thought to be the answer to problems associated with traditional approaches to instruction, particularly for educating challenging students. Teachers struggled with providing appropriate levels of instruction for students with a variety of learning styles, aptitudes, and in some cases, disabilities. With PCs in the classroom, gifted and talented students would be able to receive additional or more challenging assignments, while ensuring additional drill and practice and self-paced learning for students who required more individualized assistance. Those who embraced early computer technology in the classroom experienced the PC as inexpensive, portable, and a way to provide challenging but nonthreatening instruction. Some viewed the computer as a "fashion statement," whereas others were afraid of the technology and resisted its use in the classroom. The greatest limitation of early computer technology for the classroom was memory capacity. Also, with pressure from parents and the booming PC industry, CAI programs for the classroom were introduced before being adequately assessed. Finally, it was difficult to measure the effectiveness of CAI as compared with traditional methods of instruction due to the number of variables that must be controlled in the classroom setting.

Over the past two decades, computer usage has increased, primarily due to the increased memory capacity available in today's computers. Special applications have been developed for the special education populations and for students with specific learning disabilities. Multimedia and hypermedia (i.e., presentation of information in text, graphics, sound, animation, and video) are now used with special populations to enhance writing skills (16) and mathematics and problem-solving skills (17). PCs are now used to overcome physical disabilities and language difficulties (i.e., problems understanding or using spoken or written language) of students participating in special education curriculums and continue to be used for drill and practice of basic academic skills. Research is mixed with regard to the impact of computer-assisted instruction on student's academic performance, primarily due to research design flaws in two critical areas: (1) inclusion of adequate controls and (2) holding instructional variables constant. However, a common theme emerging from most published studies is that technology cannot take the place of good teacher instruction. Use of computers in educational settings must include effective instruction from teachers, proper social organization of the classroom, and meaningful assignments.

COGNITIVE TRAINING FOR PERSONS WITH PSYCHIATRIC DISORDERS

Early use of computers in psychiatry and psychotherapy borrowed successes of the technology in educational settings and rehabilitation of persons with acquired brain injury (18,19). Mentally ill patients often demonstrate cognitive deficits similar to individuals with learning disabilities and traumatic brain injury, including decreased attention/concentration, memory, planning and problem-solving, and judgment/decision-making (18,19).

In the area of psychiatry, reports of computer-assisted interventions began to appear in the literature in the late 1980s. Computers were used as interviewing and assessment devices (e.g., diagnostic interviews) and for self-administered rating scales for depression and other mental illnesses (20–22). Computers were thought to have some advantage over a professional conducting a clinical interview, as some psychiatric patients were more willing to

disclose sensitive information to a computer rather than to a person (22).

Later, studies appeared in the psychiatric literature using computers to treat specific cognitive deficits, such as decreased attention and psychomotor speed in persons with schizophrenia (23,24).

Greater memory capacity and improved graphics allowed the development of multimedia presentations for patient education and specific data collection. In one study, Morss et al. (25) used a multimedia computer program to assess and evaluate the side effects of antipsychotic medication in persons with schizophrenia.

Finally, computers have been used in long-term psychiatric settings to teach high level vocational skills and to remediate educational disabilities. Brieff (26) reported use of computers to teach advanced computer applications (such as database development for accounting and inventory, desktop printing and publishing, installing and upgrading software, and teaching staff word processing and spreadsheet skills), and to remediate deficits in mathematical abilities, reading comprehension, and vocabulary in persons with chronic mental illness. Brieff (26) and Perr et al. (19) have cited numerous advantages of computer use with this population. Persons with chronic mental illness can learn to use computers and appear motivated to use the technology. Computers can be more engaging and provide for self-paced learning, making computer use more attractive than traditional classroom settings. Computer programs can be easily modified or individualized. Many patients demonstrate enhanced self-esteem as they master specific skills or become productive. Some other benefits of computer use with this population included increased attention/concentration and decreased frustration.

Computer technology has also been applied to the use of psychotherapy. Like psychiatry, computers have been used to aid in diagnostics, patient education and computer-assisted instruction, and cognitive rehabilitation (27). Computers have also been used in some forms of psychotherapy. With the advent of the PC, reports of the application of computer technology immediately began to appear in the literature in the late 1970s and early 1980s. Computers were used to desensitize anxious test-takers (28), in the treatment of agoraphobia (29), in the treatment of obesity (30), and in the treatment of sexual dysfunction in individuals and couples (31). Later, applications appeared in the field of behavioral health using computers for promoting smoking cessation and substance use/abuse (32).

In addition to numerous self-help applications, computers have been used in brief psychotherapy for presenting straightforward information or feedback (33).

Studies using computers in psychotherapy have shown that the technology is widely accepted and most people find computers to be a reliable source of information (databases, Internet, etc.). Similar to reports in the psychiatric literature, some persons find it easier to share sensitive information with a computer as opposed to a person. Rialle et al. (27) cite a number of advantages of computer-mediated psychotherapy services. First, regarding ethical considerations, it is highly unlikely that exploitation, abuse, or boundary issues will occur in a relationship between a computer and persons receiving computer-assisted psychotherapy services. With continued increases in health-care costs, computers can be cost-effective and provide greater access to growing demands for mental health services. It is unlikely that computers, at least in the near future, will replace human psychotherapists given the complexity of the interaction that takes place in the course of intensive and long-term psychotherapy. The relationship between the person and the therapist is where the work of psychotherapy occurs, and computers cannot replace the warmth, empathy, and genuineness responsible for change in the therapeutic relationship.

COGNITIVE TRAINING FOR PERSONS WITH DEMENTIA

Reports of computer-assisted instruction and cognitive rehabilitation with the elderly have appeared in the nursing (34), geriatric (35), and psychology (36) literature over the past 15 years. PCs have been used in the treatment of age-related cognitive decline (37), Alzheimer's disease (AD), and other dementias (38). Computers have also been used for instruction, entertainment, and socialization of otherwise healthy elderly people without self-reported cognitive decline (39).

In the treatment of age-related cognitive decline, Gunther et al. (37) demonstrated that a 14 week computer-assisted cognitive rehabilitation program resulted in improved memory, information processing speed, learning, and interference tendency for 19 elderly participants who showed age-related cognitive decline without dementia. Follow-up five months after the completion of the cognitive rehabilitation program showed that information processing speed, learning, and interference tendency were maintained. The study also listed a number of advantages of computer use with this population, including the computer's value to motivate the elderly to learn, the computer's ability to directly measure success, it's flexibility, and the ability to provide immediate and totally value-free feedback.

Computer-based interactive programs have been used to treat mild to moderate AD. Hoffman et al. (38) reported results from a pilot study of 10 AD patients. Although no evidence of general cognitive improvement or transfer of skills to real-life settings was noted, most participants in the study showed increased speed of computer use, required less assistance to use the computer program, and 8 of 10 made fewer mistakes. Mahendra (40) pointed out that many of the principles that facilitate learning in patients with AD or vascular dementia are easily incorporated into computer programs. For instance, repetition, active involvement in learning (i.e., interactive programs), cueing, feedback, and reinforcement of correct responses can easily be built into computer-assisted cognitive rehabilitation programs.

In a review article, Hendrix (39) reported on a number of studies that revealed computer usage by otherwise healthy elderly individuals resulted in improved self-esteem (i.e., from a sense of accomplishment or productivity), increased attention to task, and greater social interaction. Computer usage also provided entertainment and mental stimulation

in the form of games, puzzles, and the like. Many of the sensory (i.e., visual and hearing) and motor deficits associated with aging were overcome by the use of a PC. For example, increasing font size to at least 18 made text more readable. External speakers (for amplification) or visual indicators on the screen compensated for poor hearing. Touch screens, voice-activated typing programs, and keyboard guards were among other modifications that allowed the elderly to use computers for entertainment, research, and contact with friends and family.

Finally, computers and computer programs have recently been developed as a screening tool to identify mild cognitive impairment in early dementia patients (41). Future trends in the use of computers with the elderly will be aimed at preventing cognitive loss. "Geroprophylaxis" (37) or preventive therapies for the elderly may incorporate the use of computers in everyday activities, and probably at a younger age (for example, just after retirement) in an effort to prevent cognitive decline.

Although all these implementations present the conventional use of computers in cognitive rehabilitation, which have followed ever since the advent of computers in rehabilitation, VR technology, a more recent computer-based intervention in cognitive training, is increasing gaining attention. The technology offers a safe appendage to clinical interventions and is therefore just beginning to gain a therapeutic appeal in the area of rehabilitation. The following section discusses a few studies that have implemented the VR technology in cognitive rehabilitation.

VIRTUAL REALITY AS A COMPUTER-GENERATED TECHNOLOGY FOR REHABILITATION

VR provides a natural and intuitive interaction with a simulated environment because of the enhanced kinesthetic feedback gained by using various haptic devices. It provides the capability to display a 3D computer-generated environment, which allow individuals to interact and become immersed in the simulation as if they were in a naturalistic setting (42,43). Immersion or presence is achieved by a variety of projection systems ranging from basic flat-screen systems to projection walls and rooms known as CAVES (www.evl.uic.edu/pape/CAVE/). These projection systems produce virtual or mixed environments where real and simulated representations of objects and people can be used for evaluation and training of individual skills (44).

Specialized devices such as head-mounted displays (HMDs) combined with tracking systems, earphones, gesture-sensing gloves, and haptic feedback also facilitate the sense of immersion in the virtual environment (45). However, the VR headsets may produce deficits of binocular function after a period as short as 10 min (46,47). HMDs contribute to ocular discomfort, headaches, and motion sickness (48). Many factors may contribute to the symptoms when using HMDs, including the weight and fitting of the HMDs, the postural demands of the equipment, low illumination and spatial resolution, as well as the sensory conflict between the visual, vestibular, and nonvestibular proprioceptive system (48). Although some studies report minimal risk or rapid dissipation of side effects when using VR technology, additional research is needed to determine the duration and severity of the symptoms (45,48).

Computer graphics techniques are used to create a virtual setting, complete with images and sound, that corresponds to what the user would experience in a real environment. The VR headset and a tracking system sense the position of the user's head and communicate the information to the computer that uses this spatial information to immerse and orient the user in the virtual setting.

Therefore, the user can navigate and interact with objects in the virtual environment using other VR devices such as data gloves, joy sticks, or even the natural hand (45,49). In other words, these collections of 3D computer-generated images can generate a continuum of high fidelity virtual environments. The VR devices and displays range from "virtual world objects" to "mixed reality" in which the real world and the virtual world objects are presented together within a single display (49,50).

The VR technology seems to offer many opportunities for evaluation and training of healthy and disabled populations. Successful reports include the U.S. National Aeronautics and Space Administration's use of VR training for astronauts to repair the Hubble telescope in space (51) and use of VR in the rehabilitation of people with intellectual disabilities (52).

VR FOR COGNITIVE TRAINING

This section will focus on the potential of VR for cognitive training. Persons who suffer from central nervous system damage may experience profound and pervasive difficulties with cognition. The beneficial aspects and limitations of VR for cognitive impairments will be discussed from the impairment and functional disability perspective. This model represents the continuum of the health-care services at the body and functional levels. VR training in cognitive rehabilitation has been divided into attention training and memory training.

VR FOR ATTENTION ASSESSMENT AND TRAINING

The brain allows individuals to constantly scan the environment for stimuli. Arousal, orientation, and focus are regulated in the brain. The reticular activating systems, the superior colliculus and the parietal cortex, and the lateral pulvinar nucleus in the thalamus are active processors of attention. All these brain structures in connection with the frontal lobes allow the individual to establish and maintain stimulus control and observe features in the environment. Attention is a multicomponent behavior and can be impaired in many neurological conditions. VR has been used for attention training. Some of the theories about the perception of reality in the virtual experience are, in fact, attributed to the three dimensions of attention (53). The three attention dimensions are (1) the focus of attention between presence and absence of reality; (2) the locus of attention between the virtual and physical world; and (3) the "sensus" of attention between arousal and the users internal physiological responses (53).

In general, attention is the ability to focus on critical aspects of the stimulus in the environment. VR technology can provide a controlled stimulus environment in which cognitive distractions can be presented, monitored, manipulated, and recorded using various levels of attention, including (1) **focused attention** (to perceive and respond to specific information/stimuli), (2) **sustained attention** (to maintain consistent concentration or vigilance on performing a task), (3) **selective attention** (to avoid distractions or competing stimuli), (4) **alternating attention** (to shift or alternate the focus of attention between tasks), and (5) **divided attention** (to respond to multiple stimuli or to give two or more responses simultaneously) (45,54). Several studies have supported the potential use of VR for the assessment and training of attention skills. In 2001, Rizzo et al. (55) reported on a virtual classroom environment, and in 2002, Lengenfelder et al. (42) reported on a driving course.

Divided attention specifically requires the ability to respond to multiple tasks at the same time or give two responses simultaneously. Lengenfelder et al. (42) used a VR driving course environment displayed on a computer screen to study divided attention of VR drivers with and without traumatic brain injury (TBI). The task required driving while identifying a four-digit number appearing in the same or random locations on the vehicle's windshield. The preliminary results showed no differences in relative speed between VR drivers with and without TBI on any of the four attention conditions used but rather the rate of the stimulus presentation seemed to influence the driving performance. In addition, the VR drivers with TBI showed a greater number of errors on the secondary task (number recall) performed while driving. Spearman's correlations between the VR performance of divided attention and neuropsychological measures of divided attention indicated that the more errors that were made during the VR divided attention task, the lower the number of correct responses on neuropsychological measures of divided attention. The findings suggest that VR may provide a way to measure divided attention and its impact on driving.

Other researchers have designed VR classrooms to examine children with attention deficit hyperactivity disorder (ADHD) during the control and manipulation of visual and auditory distracters (41,56). The preliminary results showed that children with ADHD had significantly more omission errors in the distracting conditions than the children without ADHD (56). Cho et al. (41) developed and studied two virtual cognitive training courses, VR classroom environment and a comparison computer cognitive training. The VR head-mounted display was used to validate the possibility of attention enhancement on 30 teenagers who had been isolated in a reformatory facility. The participants were assigned into three groups: VR group, nonVR (cognitive training) group, and control group (no special treatment). The interventions took eight sessions over two weeks. The results showed that the VR group was the most improved in attention training. These studies support the use of VR for attention impairment training. VR also has potential for memory impairment training.

VR FOR MEMORY ASSESSMENT AND TRAINING

An important feature of cognitive rehabilitation is neuropsychological assessment in order to determine the areas of impairment and function to use for training and compensation (45). Conventional standardized memory assessments have been criticized for lacking in ecological validity (57). VR may be able to increase the ecological validity that many conventional standardized memory assessments are lacking (57). VR can be designed for assessment and training of memory impairments after disability. Memory is a multifaceted process of brain function. Although many types of memory exist that activate a complex network of structures in the nervous system, memory requires attention to the information, encoding and maintaining information in short-term memory, followed by storing information in long-term memory, and finally, consolidating the information for retrieval as needed (54). At the cellular level, memory can be conceived as a specific neuronal association pattern that remains as a permanent pattern in the brain after the original information stimulus has ceased to exist (58).

Many brain structures are involved in memory function. Many areas are located anatomically beneath the cingulated cortex, and include the thalamus, fornix, mammillary bodies, hippocampus, amygdala, basal forebrain, and prefrontal cortex (59).

Some exploratory studies indicate that VR has potential for memory remediation in people with memory impairments. It has been found that VR can promote procedural learning and transfer to improved real-world performance (57). Burgess et al. (60) investigated four types of recognition memory in a VR town environment. Thirteen healthy young male volunteers explored the VR town until they felt confident they could find their way around. They were then asked four types of forced choice recognition memory questions about person, place, object, and the width of the object. Event-related functional magnetic resonance imaging (efMRI) was performed while participants answered the questions. The results revealed that no significant difference in performance existed between the memory for person and the memory for place. The performance on the memory for object was significantly better. By combining fMRI with VR technology, the investigators were able to investigate spatial memory in simulated life-like events. The results suggested that the retrieval of VR spatial events (place, location) activated medial temporal, parietal, and prefrontal systems in the brain.

In another study, Hoffman et al. (61) explored remembering real, virtual, and false memories using a Virtual-Real Memory Characteristic Questionnaire (VRMCQ). As people can differentiate between memories of real and imagined events, VR environments offer a new source of memories for events. The authors explored how accurately people can distinguish reality from VR in memory via a source identification test. Participants were exposed to real and virtual objects and, one week later, took an identification test in which they determined whether the object had been real, virtual, or new. They then rated the qualities and associated with each memory using the (VRMCQ). Clear differences in the qualities associated with real

and VR objects were found. The real objects were more likely than VR objects to be associated with perceptual cues (61,62).

A wide variety of exploratory studies have been reviewed by Brooks and Rose (57) indicating that VR can enable a more comprehensive and controlled assessment of prospective memory than is possible with paper-pencil standardized tests. Brooks and colleagues further investigated the differences between active and passive participation in the nonimpaired participants.

Phobias have been treated using VR to desensitize patients to heights (acrophobias) (63), public speaking (agoraphobia) (64), small spaces (claustrophobia) (65), and flying (66).

VR FOR ACTIVITIES OF DAILY LIVING ASSESSMENT AND TRAINING

VR Kitchen Environments

Effective VR assessment and training of activities of daily living have been reported in a variety of areas including a virtual kitchen (14,67–70). Our research team developed an HMD virtual kitchen environment designed to assess persons with brain injury on their ability to perform a 30 step meal preparation task (68). The prototype was tested using a sample of 30 persons with traumatic brain injury. The pilot results showed the stability of performance estimated using intraclass correlation coefficient (ICCs) of 0.73. When three items with low variance were removed, the ICC improved to 0.81. Little evidence of vestibular optical side effects was noted in the subjects tested. In 2001, our research team used the same virtual kitchen environment to assess selected cognitive functions of 30 patients with brain injury and 30 volunteers without brain injury to process and sequence information (70). The results showed that persons with brain injuries consistently demonstrated a significant decrease in the ability to process information identify logical sequencing and complete the overall assessment compared with volunteers without brain injury. The response speed was also significantly different. Our team's VR third pilot was designed to test the stability and validity of the information collected in the VR environment from 54 consecutive patients with TBI (14). The subjects completed meal preparation both in a virtual kitchen and in an actual kitchen twice over a three week period. The results showed an ICC value of 0.76. Construct validity of the VR environment was demonstrated. Multiple regression analysis revealed that the VR kitchen test was a good predictor for the actual kitchen assessment. Our investigations demonstrated adequate reliability and validity of the VR system as a method of assessment in persons with brain injury.

Exercising in VR environments offers the potential for gains in cognitive and motor functions (71–73). Refer to Fig. 1 for examples of a VR software program that has been used for cognition and motor skills training. Grealy et al. (71) supported the impact of exercise and VR on the cognitive rehabilitation of persons with traumatic brain injury in a study that provided a four week VR intervention and compared it with control patients of similar age, severity, and time post injury. They found the patients performed significantly better than controls on the tests of psychomotor speed, and on verbal and visual learning. Significant improvements were observed following a single session of VR exercise.

VR Mobility Environments

VR environments have been developed to train aspects of mobility including street crossing, wheelchair mobility, and use of public transportation and driving. Strickland et al. (74) found that children with autism were able to accept and wear VR helmets. McComas et al. (75) also investigated the effectiveness of VR for pedestrian safety in healthy children. Their results showed that significant change in performance occurred after three trials with VR intervention. Children learned safe street crossing in a desktop VR program. Their crossing skills were transferred to real behaviors in the suburban school group but not in the urban school group. The investigators provided no explanation for the difference noted.

VR environments have been developed for detection of driving impairment in persons with cognitive disability in persons with TBI and persons with driving and flying phobias (76,77). As stated before, impairments in divided attention has an impact on driving skills; therefore, attention training is promoted for driving skills training.

CONCLUSION

The fast data recording and processing capabilities have stimulated the application of computers in cognitive rehabilitation since the mid-1970s. It was first used in the treatment of persons with traumatic brain injury and learning disability. The technology was later applied to persons with mental illness and, most recently, in the treatment of adults with Alzheimer's and other dementias. In all applications, computers have been found to be interesting and motivating to the users, an efficient stimulus presentation, and an excellent data collection system. Today's computers are portable, fast, and software can be customized to meet specific needs of an individual user, teacher, or therapist.

A growing body of research continues to accumulate investigating the potential of computer technology to improve impairments, functional performance, and societal participation in persons with disabilities. It is clear that some evidence exists to support computer-based interventions as an adjunct to clinician-guided treatment (78). However, sole reliance on repeated exposure and practice on computer-based tasks without some involvement and intervention by the therapists is not recommended for cognitive rehabilitation (78).

Computer technology can provide valuable opportunities to health-care providers, educators, and researchers. Computer technology training is not a simple or single solution, and although it can promote knowledge and solutions, we also need to address other implications for cognitive rehabilitation. For example, access to computer technology is needed for personal assistance and to provide

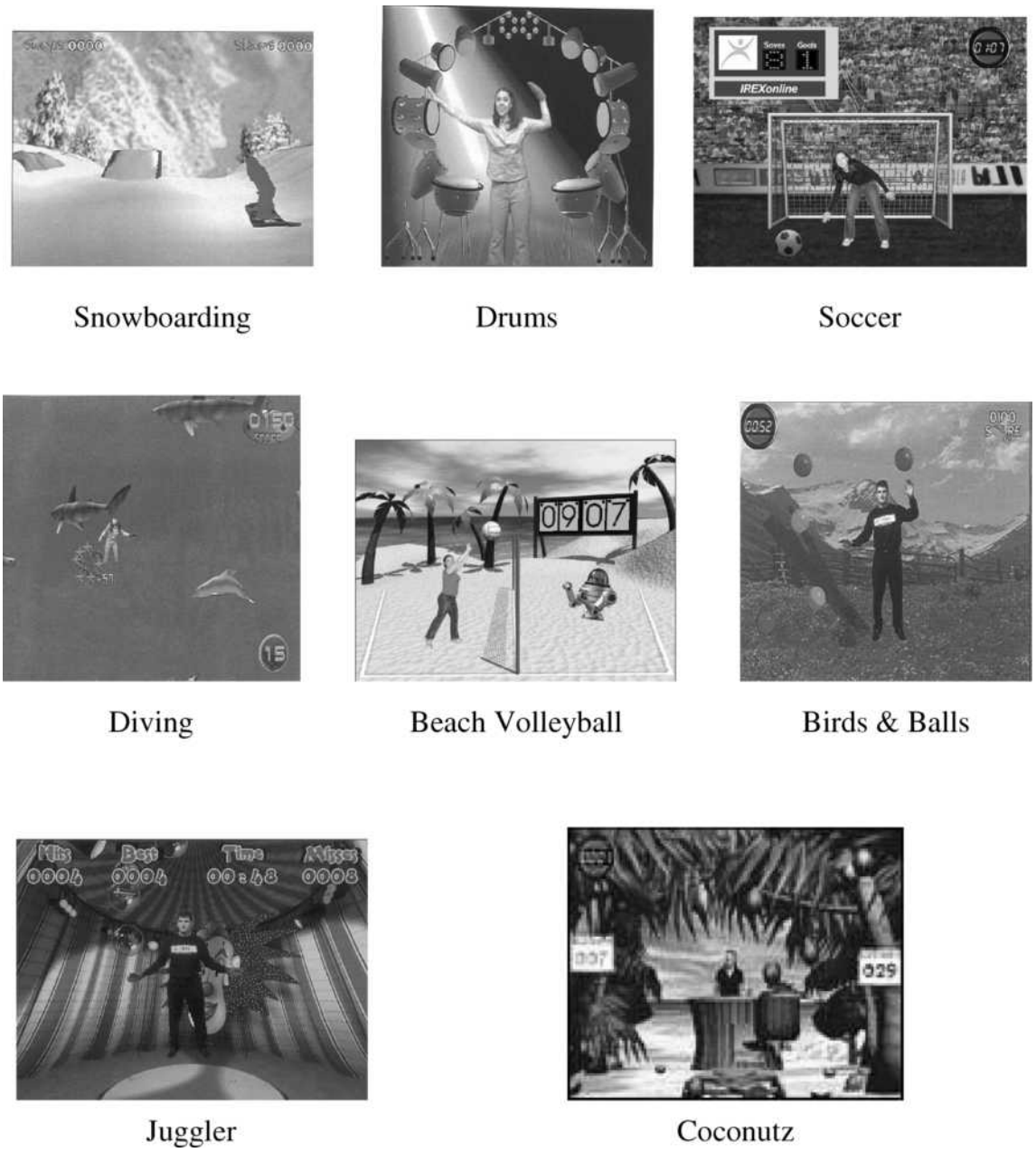


Figure 1. Examples of virtual reality environments. Image of the IREX™ system used with permission of GestureTek Inc™ - World Leaders in Gesture Recognition Technology www. gesturetek.com.

access to education, employment, and social opportunities. However, because of the expense that limits technology access because of constraints in institutions budgets, personal income limits, and funding inadequacies and constraints, only a limited number of persons with disabilities can access computer technology (79).

Finally, ethical considerations of computer technology, including VR, require mentioning. Several areas need to be considered, including (1) the patient or user must share the control and responsibility of the computer technology experience with the therapist, educator, or researcher; (2) careful care and monitoring is required for patients with certain psychopathology for potential user difficulties that may be encountered through the computer technology experience; (3) research participants must not be deceived into believing that they are experiencing real-life events; and (4) the patient or user must not be deprived of real-life experiences.

The application of computer technology is promising. However, substantial work needs to be conducted to identify and improve best practices through this medium. An important question is whether the improvement in the individual who used computer technology training is transient or sustained. Cognitive rehabilitation needs to promote generalization to everyday functioning and everyday activities. We need more rigorous research activities geared toward establishing a body of evidence that supports the effectiveness of computer-based technology.

BIBLIOGRAPHY

1. Cicerone KD, Dahlberg C, Kalmar K, Langenbahn DM, Malec JF, Berquist TF, Felicetti T, Giacino JT, Harley JP, Harrington DE, Herzog J, Kneipp S, Laatsch L, Morse PA. Evidence-based cognitive rehabilitation: Recommendations for clinical practice. Arch Phys Med Rehabil 2000;81:1596–1615.
2. Giaquinto S, Fioro M. THINKable, a computerized cognitive remediation: First results. Acta Neurol (Napoli) 1992;14(4–6):547–560.
3. Harley JP, Allen C, Braciszewski TL, Cicerone KD, Dahlberg C, Evans S, Foto M, Gordon WA, Harrington D, Levin W, Malec JF, Millis S, Morris J, Muir C, Richert J, Salazar E, Schiavone DA, Smigelski JS. Guidelines for cognitive rehabilitation. NeuroRehabilitation 1992;2(3):62–67.
4. Lynch B. Historical review of computer-assisted cognitive retraining. J Head Trauma Rehabil 2002;17(5):446–457.
5. Larose S, Gagnon S, Ferland C, Pepin M. Psychology of computers: XIV. Cognitive rehabilitation through computer games. Percept Mot Skills 1989;69(3 Pt. 1):851–858.
6. Kurlychek RT, Levin W. Computers in the cognitive rehabilitation of brain-injured persons. Crit Rev Med Inform 1987; 1(3):241–257.
7. Harley JP. Software review: "COGREHAB" cognitive rehabilitation programs. Comput Psychiatr Psychol 1984;6:15.
8. Niemann H, Ruff RM, Baser CA. Computer-assisted attention retraining in head-injured individuals: A controlled efficacy study of an outpatient program. J Consult Clin Psychol 1990; 58(6):811–817.
9. Gray JM, Robertson I. Remediation of attentional difficulties following brain injury: Three experimental case studies. Brain Inj 1989;3(2):163–170.
10. Glisky E, Schacter D. Long-term retention of computer learning by patients with memory disorders. Neuropsychologia 1988;26:173–178.
11. Dirette DK, Hinojosa J. The effects of a compensatory intervention on processing deficits of adults with acquired brain injuries. Occup Ther J Res 1999;19(4):224–239.
12. Ross FL. The use of computers in occupational therapy for visual-scanning training. Am J Occup Ther 1992;46(4):314–322.
13. Weiss PL, Naveh Y, Katz N. Design and testing of a virtual environment to train stroke patients with unilateral spatial neglect to cross a street safely. Occup Ther Int 2003;10(1):39–55.
14. Zhang L, Abreu BA, Seale GS, Masel BE, Christiansen CH, Ottenbacher KJ. A virtual reality environment for evaluation of a daily living skill in brain injury rehabilitation: Reliability and validity. Arch Phys Med Rehabil 2003;84:1118–1124.
15. Kolich EM. Microcomputer technology with the learning disabled: A review of the literature. J Learn Disabil 1985;18(7):428–431.
16. MacArthur CA. Using technology to enhance the writing process of students with learning disabilities. J Learn Disabil 1996;29(4):344–354.
17. Babbitt BC, Miller SP. Using hypermedia to improve mathematics problem-solving skills of students with learning disabilities. J Learn Disabil 1996;29(4):391–401.
18. Burda PC, Starkey TW, Dominguez F, Vera V. Computer-assisted cognitive rehabilitation of chronic psychiatric patients. Comput Hum Behav 1994;10(3):359–368.
19. Perr A, White S, Rekoutis PA. Assistive technology and computer-based intervention in psychiatric settings. Technol Special Interest Section Quart 2002;12(2):1–4.
20. Erdman HP, Greist JH, Gustafson DH, Taves JE, Klein MH. Suicide risk prediction by computer interview: A prospective study. J Clin Psychiatry 1987;48:464–467.
21. Ancil RJ, Rogers D, Carr AC. Comparison of computerized self-rating scales for depression with conventional observer ratings. Acta Psychiatr Scand 1983;71:315–317.
22. Erdman HP, Klein MH, Greist JH. Direct patient computer interviewing. J Consult Clin Psychol 1985;53:760–773.
23. Hermanutz M, Gestrich J. Computer-assisted attention training in schizophrenics: A comparative study. Eur Arch Psychiatry Clin Neurosci 1991;240(4–5):282–287.
24. Benedict RHB, Harris AE. Remediation of attention deficits in chronic schizophrenic patients: A preliminary study. Br J Clin Psychol 1989;28:187–188.
25. Morss SE, Lenert LA, Faustman WO. The side effects of antipsychotic drugs and patients' quality of life: Patient education and preference assessment with computers and multimedia. Proc Annu Symp Comput Appl Med Care 1993; 17–21.
26. Brieff R. Personal computers in psychiatric rehabilitation: A new approach to skills training. Hosp Commun Psychiatry 1994;45(3):257–260.
27. Rialle V, Stip E, O'Connor K. Computer-mediated psychotherapy: Ethical issues and difficulty with implementation. Humane Med 1994;10(3):185–192.
28. Biglan A, Willwock C, Wilk S. The feasibility of a computer-controlled program for the treatment of test anxiety. J Ther Exp Psychiatry 1979;10:47–49.
29. Ghosh A, Marks IM. Self-treatment of agoraphobia by exposure. Behav Ther 1987;18:3–16.
30. Wylie-Rosett J, Swenchionis C, Ginsberg M, Cimino C, Wassertheil-Smoller S, Caban A, Segal-Isaacson CJ, Martin T, Lewis J. Computerized weight loss intervention optimizes staff time: The clinical and cost results of a controlled clinical trial conducted in a managed care setting. J Am Diet Assoc 2001;101(10):1155–1164.
31. Reitman R. The use of small computers in self-help sex therapy. In: Schwartz M, editor. Using Computers in Clinical Practice. New York: Haworth Press; 1984. p 363–380.
32. McDaniel AM, Hutchison S, Casper GR, Ford RT, Stratton R, Rembusch M. Usability testing and outcomes of an interactive computer program to promote smoking cessation in low-income women. AMIA 2002 Annual Symposium Proceedings. 2002. 509–513.
33. Gould RL. The Therapeutic Learning Program (TLP): A computer-assisted short-term treatment program. Comput Psychiatry Psychol 1986;8(3):7–12.
34. Taira F. Computer use by adults with disabilities. Rehabil Nurs 1994;19(2):84–86, 95.
35. Weisman S. Computer games for the frail elderly. Gerontologist 1983;23(4):361–363.
36. Schreiber M, Schweizer A, Lutz K, Kalveram KT, Jancke L. Potential of an interactive computer-based training in the rehabilitation of dementia: An initial study. Neuropsycholog Rehab 1999;9(2):155–167.
37. Gunther VK, Schafer P, Holzner BJ, Kemmler W. Long-term improvements in cognitive performance through computer-assisted cognitive training: A pilot study in a residential home for older people. Aging Ment Health 2003;7(3):200–206.
38. Hofmann M, Hock C, Kuhler A, Muller-Spahn F. Interactive computer-based cognitive retraining in patients with Alzheimer's disease. J Psychiat Res 1996;30(6):493–501.
39. Hendrix C. Computer use among elderly people. Comput Nurs 2000;18(2):62–71.
40. Mahendra N. Direct interventions for improving the performance of individuals with Alzheimer's disease. Semin Speech Lang 2001;22(4):291–303.
41. Cho, Yang J, Kim SY, Yang DW, Park M, Chey J. The validity and reliability of a computerized dementia screening test developed in Korea. J Neurol Sci 2002;203–204:109–114.
42. Lengenfelder J, Schultheis MT, Al-Shihabi T, DeLuca J, Mourant R. Divided attention and driving: A pilot study using virtual reality. J Head Trauma Rehabil 2002;17(1):26–37.

43. Rizzo AA, Buckwalter JG. Virtual reality and cognitive assessment and rehabilitation: The state of the art. In: Riva G, editor. Virtual Reality in Neuro-Psycho-Physiology: Cognitive, Clinical And Methodological Issues In Assessment And Rehabilitation, vol. 44. Amsterdam: ISO Press; 1997. p 123–145.
44. Cosman PH, Cregan PC, Martin CJ, Cartmill JA. Virtual reality simulators: Current status in acquisition and assessment of surgical skills. ANZ J Surg 2002;72(1):30–34.
45. Schultheis MT, Himelstein J, Rizzo AA. Virtual reality and neuropsychology: Upgrading the current tools. J Head Trauma Rehabil 2002;17(5):378–394.
46. Mon-Williams M, Plooy A, Burgess-Limerick R, Wann J. Gaze angle: A position mechanism of visual stress in virtual reality. Ergonomics 1998;41(3):280–285.
47. Wann JP, Rushton S, Mon-Williams M. Natural problems for stereoscopic depth perception in virtual environments. Vision Res 1995;35(19):2731–2736.
48. Ames SL, Wolffsohn JS, McBrien NA. The development of a symptom questionnaire for assessing virtual reality viewing using a head-mounted display. Optom Vis Sci 2005;82(3): 168–176.
49. Nothhelfer U. Landscape Architecture in the Reality-Virtuality. Paper presented at the Trends in GIS and Virtualization in Environmental Planning and Design, Anhalt University of Applied Sciences, 2002.
50. Milgram P, Kishino F. A taxonomy of mixed reality visual displays. IEICE Trans Inform Syst (Special Issue on Networked Reality) 1994;E77-D(12):1321–1329.
51. Loftin RB, Kenney PJ. Training the Hubble-space telescope flight team. IEEE Comput Graph Appl 1995;15(5):317.
52. Standen PJ, Brown DJ. Virtual reality in the rehabilitation of people with intellectual disabilities: Review. CyberPsychol Behav 2005;8(3):272–282.
53. Waterworth L, Waterworth JA. Focus, locus, and sensus: The three dimensions of virtual experience. CyberPsychol Behav 2001;4(2):203–213.
54. Malia KB, Bewick KC, Raymond MJ, Bennett TL. Brainwave-R: Cognition Strategies and Techniques for Brain Injury Rehabilitation: User's Guide and Introduction to Brain Injury. Austin, TX: Pro-Ed; 2002.
55. Rizzo AA, Buckwalter JG, McGee JS, Bowerly T, Van der Zaag C, Neumann U, Thiebaux M, Kim L, Pair J, Chua C. Virtual environments for targeting cognitive processes: An overview of projects at the University of Southern California Integrated Media Systems Center. Presence: Teleoperators Virtual Environ 2001;10:359–374.
56. Rizzo A, Buckwalter JG, Bowerly T, Van der Zaag C, Humphrey L, Neumann U, Chua C, Kyriakakis C, van Rooyen A, Sisemore D. The virtual classroom: A virtual reality environment for the assessment and rehabilitation of attention deficits. CyberPsychol Behav 2000;3(3):483–499.
57. Brooks M, Rose FD. The use of virtual reality in memory rehabilitation: Current findings and future directions. NeuroRehabilitation 2003;18(2):147–157.
58. Carter R. Mapping the Mind. Los Angeles, CA: University of California Press; 1998.
59. Parenté R, Herrmann D, editors. Retraining Cognition: Techniques and Applications. Gaithersburg, MD: Aspen; 1996.
60. Burgess N, Maguire EA, Spiers HJ, O'Keefe J. A temporoparietal and prefrontal network for retrieving the spatial context of lifelike events. Neuroimage 2001;14(2):439–453.
61. Hoffman HG, Garcia-Palacios A, Thomas AK, Schmidt A. Virtual reality monitoring: Phenomenal characteristics of real, virtual, and false memories. CyberPsychol Behav 2001;4(5): 565–572.
62. Johnson MK, Raye CL. Reality monitoring. Psychol Rev 1981;88:67–85.
63. Rothbaum O, Hodges LF, Kooper R, Opdyke D, Williford JS, North M. Effectiveness of computer-generated (virtual reality) graded exposure in the treatment of acrophobia. Am J Psychiatry 1995;152(4):626–628.
64. North MM, North SM, Coble JR. Virtual reality therapy: An effective treatment for psychological disorders. In: Riva G, editor. Virtual Reality in Neuro-Psycho-Physiology: Cognitive, Clinical and Methodological Issues in Assessment and Rehabilitation, vol 44. Amsterdam: ISO Press; 1997. p 59–70.
65. Bullinger AH, Roessler A, Muller-Spahn F. Three-dimensional virtual reality as a tool in cognitive-behavioral therapy of claustrophobic patients. CyberPsychol Behav 2000;3:387–392.
66. Kahan M, Tanzer J, Darvin D, Borer F. Virtual reality-assisted cognitive-behavioral treatment for fear of flying: Acute treatment and follow-up. CyberPsychol Behav 2000; 3(3):387–392.
67. Gourlay, Lun KC, Lee YN, Tay J. Virtual reality for relearning daily living skills. Int J Med Inform 2000;60(3):255–261.
68. Christiansen C, Abreu B, Ottenbacher K, Huffman K, Masel B, Culpepper R. Task performance in virtual environments used for cognitive rehabilitation after traumatic brain injury. Arch Phys Med Rehabil 1998;79(8):888–892.
69. Lee JH, Ku J, Cho W, Hahn WY, Kim IY, Lee S-M, Kang Y, Kim DY, Yu T, Wiederhold BK, Wiederhold MD, Kim SI. A virtual reality system for the assessment and rehabilitation of activities of daily living. CyberPsychol Behav 2003;6(4): 383–388.
70. Zhang L, Abreu BC, Masel B, Scheibel RS, Christiansen C, Huddleston N, Ottenbacher KJ. Virtual reality in the assessment of selected cognitive function after brain injury. Am J Phys Med Rehabil 2001;80(8):597–604.
71. Grealy MA, Johnson DA, Rushton SK. Improving cognitive function after brain injury: The use of exercise and virtual reality. Arch Phys Med Rehabil 1999;80(6):661–667.
72. You SH, Jang SH, Kim Y-H, Hallett M, Ahn SH, Kwon Y-H, Kim JH, Lee MY. Virtual reality-induced cortical reorganization and associated locomotor recovery in chronic stroke: An experimenter-blind randomized study. Stroke 2005;36: 1166–1171.
73. Gaggioli A, Morganti F, Walker BA, Meneghini A, Alcañiz M, Lozano JA, Montesa J, Gil JA, Riva G. Training with computer-supported motor imagery in post-stroke rehabilitation. CyberPsychol Behav 2004;7(3):327–332.
74. Strickland D, Marcus LM, Mesibov GB, Hogan K. Brief report: Two case studies using virtual reality as a learning tool for autistic children. J Autism Dev Disord 1996;26(6):651–659.
75. McComas J, Mackay M, Pivik J. Effectiveness of virtual reality for teaching pedestrian safety. CyberPsychol Behav 2002;5(3): 185–190.
76. Wald J, Liu L, Hirsekorn L, Taylor S. The use of virtual reality in the assessment of driving performance in persons with brain injury. Stud Health Technol Inform 2000;70:365–367.
77. Wald J, Taylor S. Efficacy of virtual reality exposure therapy to treat driving phobia: A case report. J Behav Ther Exp Psychiatry 2000;31(3–4):249–257.
78. Cicerone KD, Dahlberg C, Malec JF, Langenbahn DM, Felicetti T, Kneipp S, Ellmo W, Kalmar K, Giacino JT, Harley P. Evidence-based cognitive rehabilitation: Updated review of the literature from 1998 through 2002. Arch Phys Med Rehabil 2005;86(8):1681–1692.
79. U.S. Department of Health and Human Services. Delivering on the Promise: Report of Federal Agencies' Actions to Eliminate Barriers and Promote Community Integration. Executive Order No. 13217, Washington, DC: U.S. Department of Health and Human Services; 2002.

See also COMMUNICATION DEVICES; COMMUNICATIVE DISORDERS, COMPUTER APPLICATIONS FOR; ENVIRONMENTAL CONTROL; HOME HEALTH CARE DEVICES.

REHABILITATION, ORTHOTICS IN

LOREN LATTA
University of Miami
Coral Gables, Florida

INTRODUCTION

Paralysis

Muscle weakness resulting from various forms of paralysis is the major type of disability treated with orthotic devices. Paralysis may be acute or chronic to varying degrees depending on the state of the disease causing paralysis. The degree of paralysis or paresis can change with time as recovery is accomplished through therapy or healing of the physiological cause. Thus, the orthotic need for support, control of joint motion, and so forth, may change with time for many of these patients. Orthotic devices in general are designed for long-term use because many of the designs were developed for chronic applications to polio patients in the late 1940s and 1950s. With many new applications and new materials, a variety of devices have been developed that are ideal for short-term applications to more acute conditions, for example, stroke, head injury, spinal cord injury, and fractures.

Neuromuscular Control

Patients who suffer loss of neuromuscular control in the form of spasticity, paralysis of isolated muscle groups, and/or recovery of neuromuscular function resulting from regeneration of neural tissue, which requires relearning of coordination, are often helped with orthotic devices. Although most of these applications are short term, a few are chronic, such as in cerebral palsy, established spinal cord injury, and head injury.

Deformity

Another common use of orthotics in a growing number of cases is for musculoskeletal deformity. The first type of deformity is related to mechanical instability of the limbs or spine. Mechanical instability can result from soft tissue or skeletal injury or from degenerative joint diseases that cause chronic and progressive instability of the musculoskeletal system. One form of soft-tissue injury relating to the instability of the skeleton is caused by surgical reconstruction of the joints. Therefore, many orthotic devices are used for stabilization postoperatively when the spine or upper or lower limb joints have been reconstructed. The surgery often causes necessary damage to the stabilizing soft-tissue structures that often adds to instability of prosthetic components that require bone healing for final stabilization. The second type of deformity is due to the growth or remodeling disturbances in the skeleton. Orthotic applications are both acute and chronic in problems relating to musculoskeletal deformity. Most chronic applications or orthotics are in progressive deformities resulting from degenerative joint diseases such as rheumatoid arthritis, osteoarthritis, hemophilia, and diabetes. Orthoses generally are applied in these instances to prevent progressive deformity and any resultant' mechanical instability in the skeleton that it may cause. In realigning limb mechanics, orthoses may also prevent some of the disabling pain of degenerative joint diseases. Acute applications of orthoses for deformities include the treatment of fractures and soft-tissue injuries about the joints and postoperative protection for fracture stabilization or joint reconstructive or arthrodesis procedures.

DEVICES

Orthoses have an almost infinite variety of materials, designs, and constructions. Many factors contribute to this variety: (1) performance criteria, (2) available materials, (3) skills of the orthotist, and (4) desires of the patient, surgeon, and therapist.

Performance Criteria

As mentioned, many devices are used in chronic applications with significant loading and require great

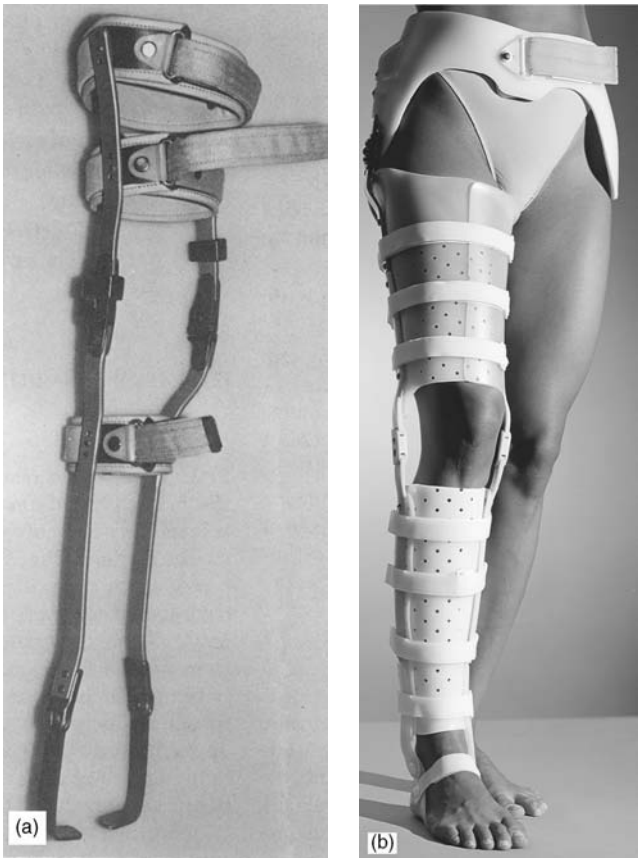


Figure 1. (a) This KAFO is designed for chronic use. It has metal uprights connected to metal bands covered with leather and padding and a "caliper" type of stirrup that attaches to the shank of an orthopedic shoe. All parts are custom fit from mostly prefabricated components. (b) The other KAFO is designed for short-term use and has a thermoplastic thigh and leg interface with the soft tissue, connected to a plastic knee hinge suspended with a plastic posterior shoe insert. All parts are prefabricated in standard sizes. Reprinted with permission from Maramed Orthopaedic Systems.

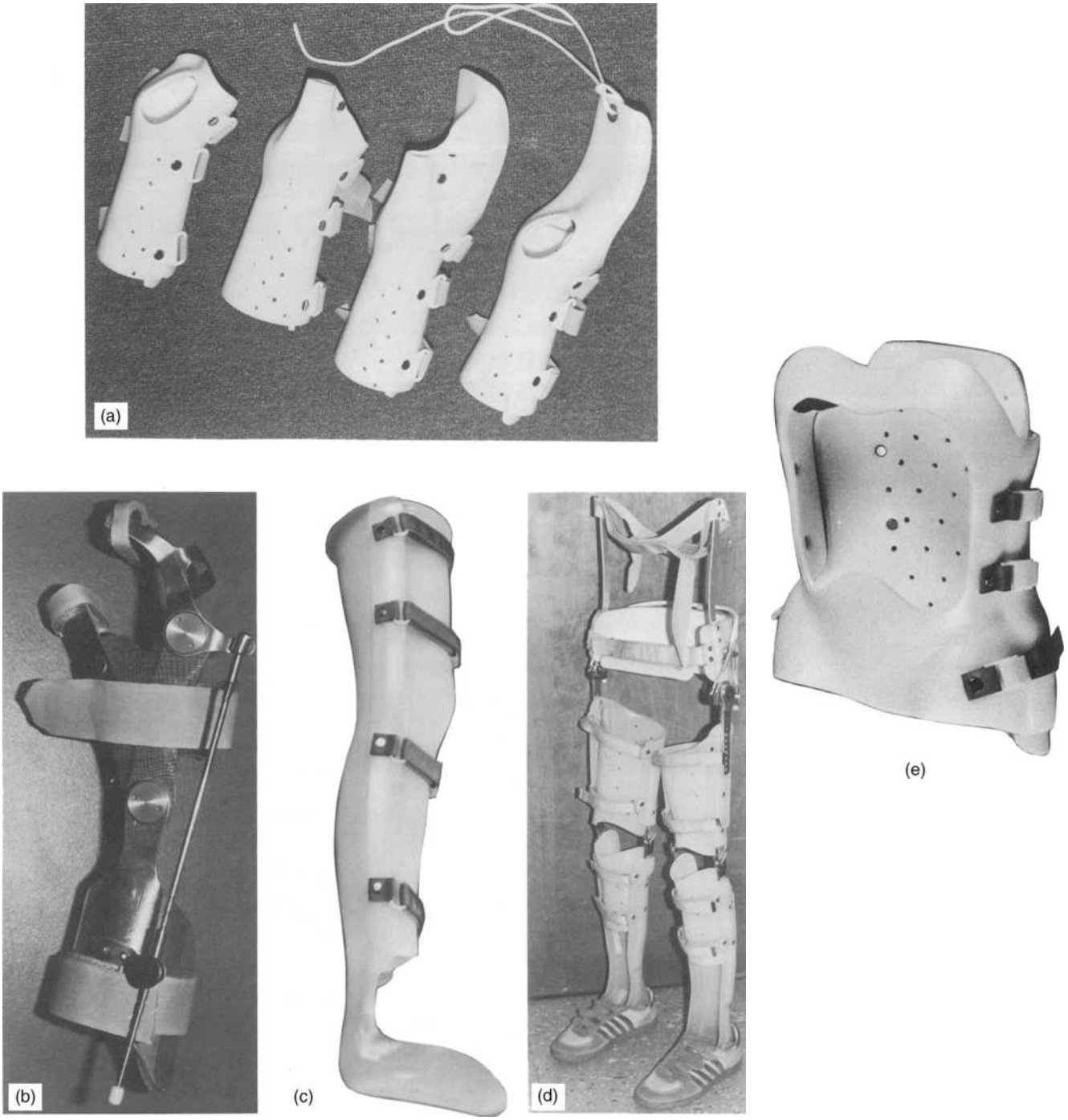


Figure 2. Splints that are used for short-term or long-term immobilization of joints are typically of a simple design and can be prefabricated as in (a) WHOs (reprinted with permission from Maramed Orthopaedic Systems) or custom fabricated as in (c) KAFO or (e) TLSO. Orthoses for similar parts of the anatomy with much more complex control criteria have typically more complex designs as in the WHO in (b), which provides a tenodesis action to close the fingers with wrist extension, and the complex reciprocator, bilateral HKAFO in (d), which provides stability for the knee and ankle while providing assistance to hip flexion.

fatigue resistance. In other instances, the loading conditions are slight, but the application is long term; thus, the primary requirement is good compatibility with the skin. Applications for very short-term use under either heavy or light loading conditions are much less demanding on the materials and design of the device (Figs. 1–3).

Many devices simply act as a splint to immobilize musculoskeletal structures for a short time (Figs. 2 and 7); the design for such a device is thus relatively simple. Other

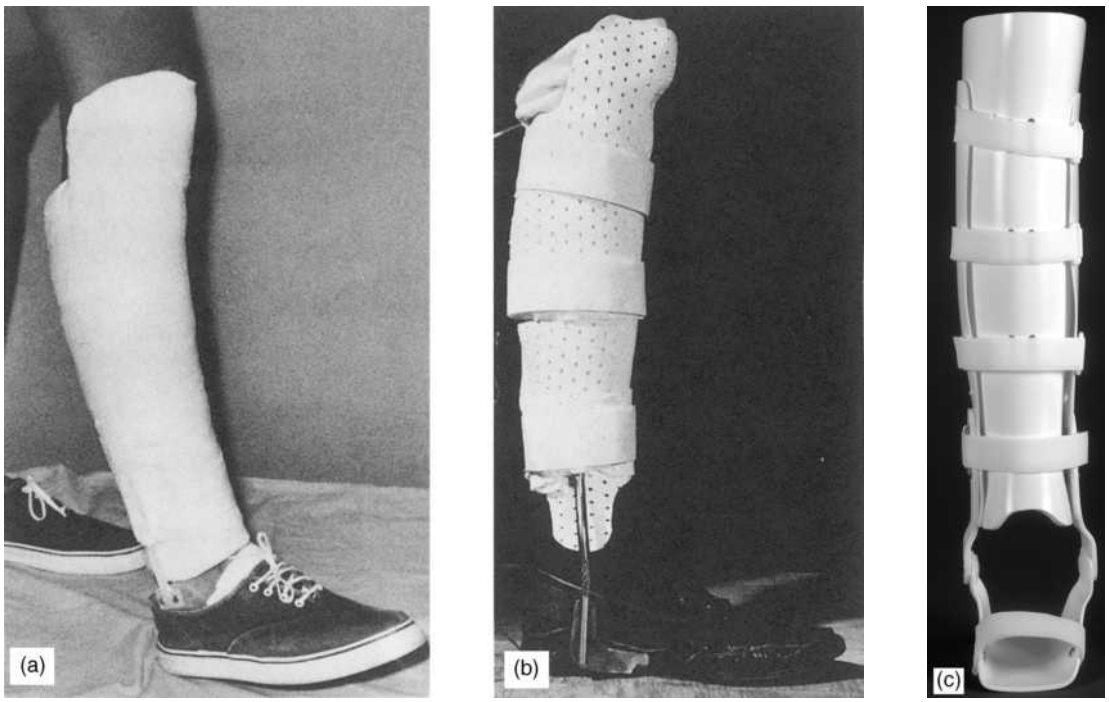


Figure 3. Many material options can accomplish the same performance criteria for a particular orthotic prescription. In this example, a tibial fracture orthosis (AFO) is constructed by three different techniques to accomplish the same end result. Plaster can be molded directly to the patient's limb and attached to a prefabricated ankle joint (a), an isoprene thermoplastic material can be molded directly to the patient's limb attached to a metal ankle joint incorporated into a stirrup permanently attached to the shoe (b), or prefabricated components can be applied with hand trimming and assembly (c). Reprinted with permission from Maramed Orthopaedic Systems.

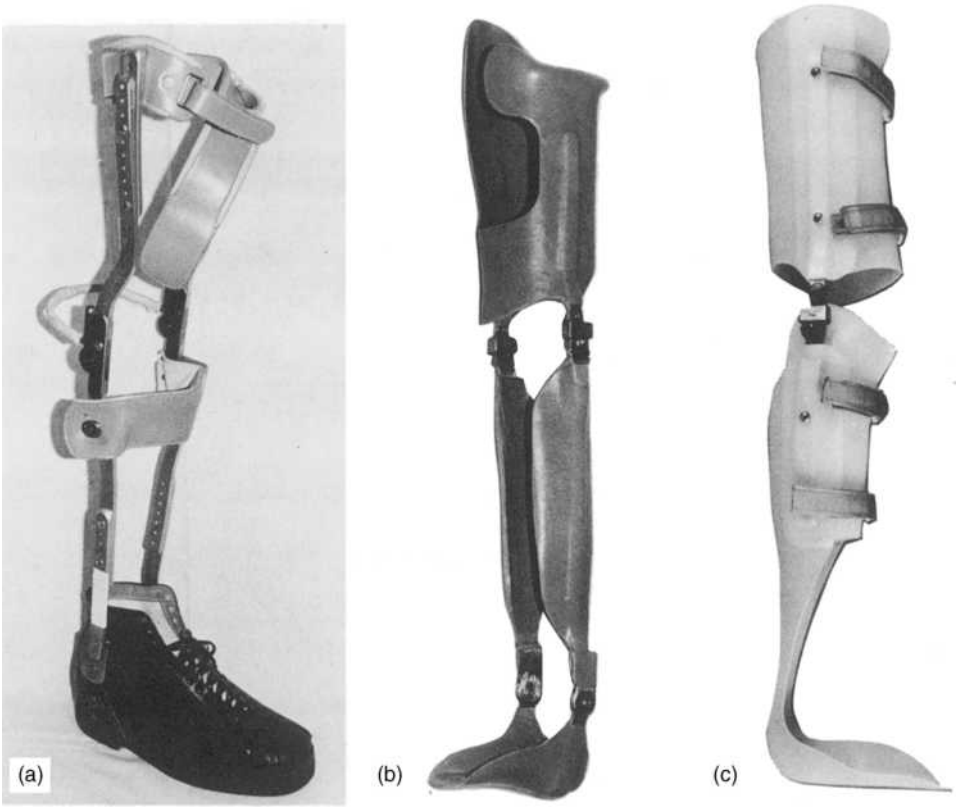


Figure 4. Typical fabrication techniques for orthoses include the use of metal uprights connected by metal bands with leather and padding for covering and skin interface (a), or custom laminated thermosetting plastic with fiber-glass or fabric reinforcement incorporating prefabricated joints (b), or custom molded thermoplastic sleeves incorporating prefabricated joints (c).

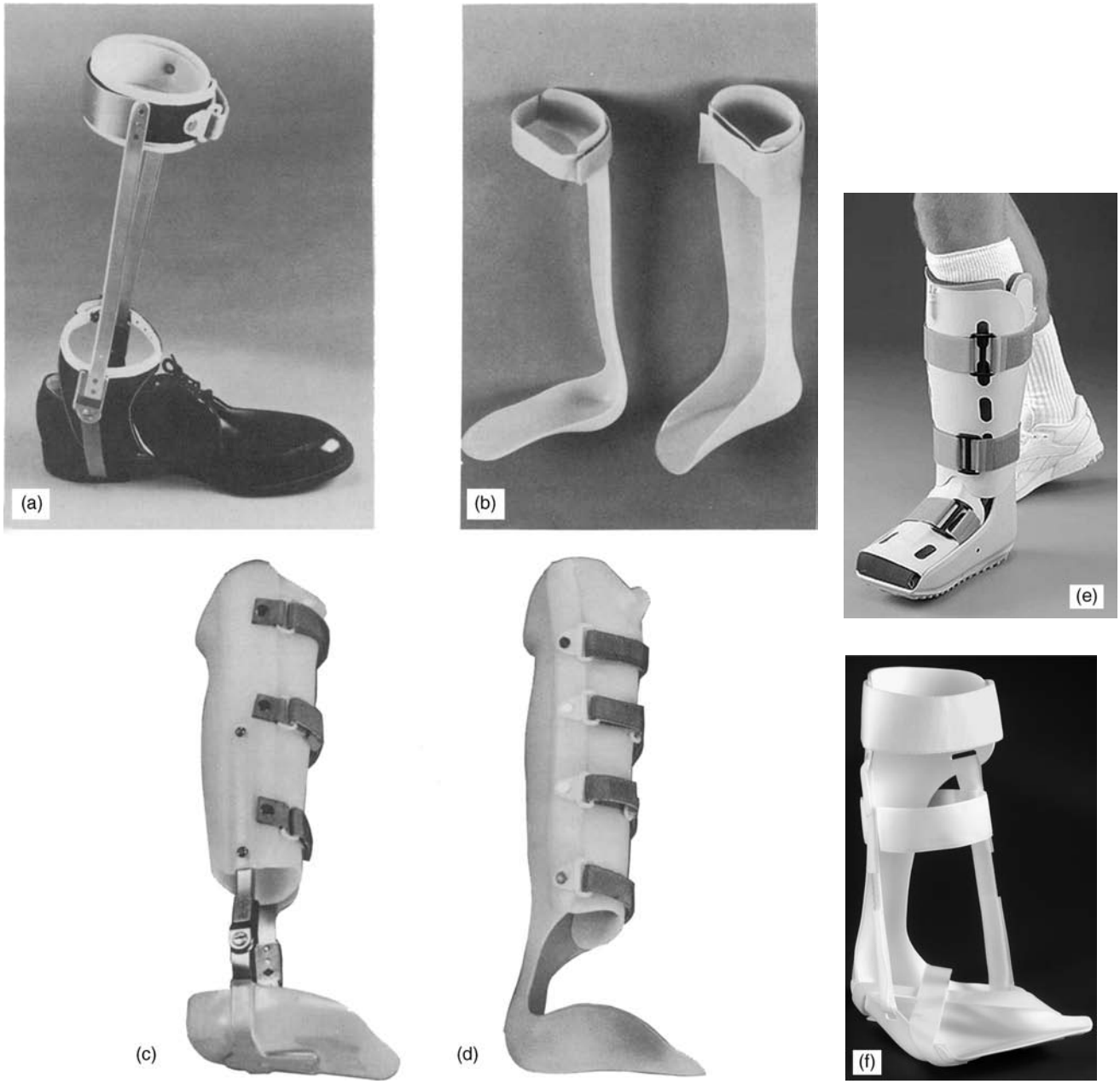


Figure 5. Various combinations of prefabricated and custom-fabricated components can be used to accomplish a particular prescription criterion as in this series of examples of AFOs. Prefabricated components are used for the stirrup connection to the shoe and the modular dorsiflexion assist ankle joint that are attached to custom-formed metal bands, and custom-fabricated leather cuffs and medial T-strap (a). These thermoplastic, posterior leaf AFOs (b) demonstrate totally prefabricated design (left, reprinted with permission from Maramed Orthopaedic Systems) for dorsiflexion assist, custom fabrication with complete limitation, or ankle motion (right). The patellar tendon bearing orthosis (PTB) is designed to transmit axial loads as well as to control ankle and subtalar motion. This example uses a custom-molded thermoplastic soft-tissue interface with the calf, incorporating metal uprights into modular, prefabricated limited motion-ankle joints connected to a stirrup custom formed from prefabricated components incorporated into a "UCB"-type shoe insert custom molded from a thermoplastic (c). A similar type of PTB orthosis uses totally custom-fabricated thermoplastic sleeves with a posterior leaf-limited-motion ankle control attached to a shoe insert (d). The CAM walker, which is totally prefabricated, provides a "roll over" sole to accommodate ambulation with an immobilized ankle, (e, reprinted with permission from Aircast, Inc.). This plantarfasciitis AFO is also completely prefabricated (f, reprinted with permission from Sky Medical, Inc.).

devices have complex prescription criteria if used for assisting, resisting, or holding motion of joints under a variety of temporal and spatial conditions; such designs might be complex (Fig. 2d).

Materials

The choice of material varies according to durability, strength, stiffness, skin compatibility, and fabrication requirements. For short-term applications, prefabricated devices that can be produced by sophisticated manufacturing techniques are often used. Thus, most any commonly available material may be used for short-term custom applications; materials that can be applied directly to the patient will often be used (Figs. 3 and 9–11). Four basic forms of such material exist: plaster, fabric reinforced, fiberglass reinforced, and solid sheets. Some are thermosetting materials activated by promotors or moisture; some are thermoplastics. Devices for long-term use tend to be custom fabricated from materials that are formed to measurements or molds of the patient. The materials are typically thermoplastic, laminated thermosetting materials with carbon fiber, fiberglass or fabric reinforcement, leather, fabric, aluminum, stainless steel, steel, and foam plastics and elastomers (Figs. 1–13).

Description of Conventional Devices

Tables 1–3 describe in a general nature the types of components that make up conventional orthoses and the functional controls and typical materials used. In general, the soft-tissue interfacing components tend to be custom fabricated, the structural members tend to be prefabricated, and the joint mechanisms tend to be prefabricated and modular. It is possible, however, to find prefabricated components and preassembled prefabricated devices in each of these areas. It is also common to find orthotic devices constructed completely from custom-fabricated components integrated together in a totally custom device for a given patient (Figs. 1–3, 5, 7, and 10).

Under functional controls, the word "feedback" is used to indicate that certain proprioceptive feedback signals may be incorporated into these components to give information to the patient regarding the load on the device and/or the position of the device in space. This is sometimes helpful to supplement loss of normal proprioception in the limb. It is possible for feedback mechanisms; EMG-, load-, OR microprocessor-activated locks; and/or resistive or assistive mechanisms at all of these joints to be applied; however, the tables simply reflect those applications that are used in relatively common practice in the field of orthotics. Applications not listed here, to the knowledge of the authors, are simply research

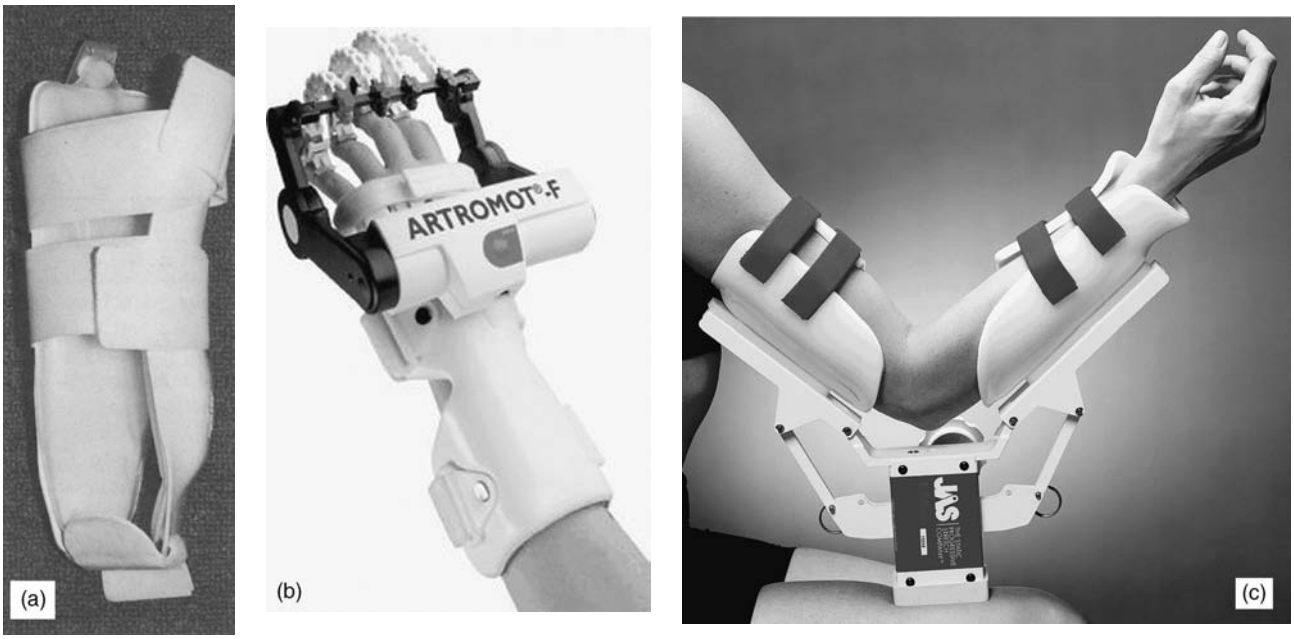


Figure 6. Many types of devices use components that are similar to classic orthoses, but at this time are considered by the authors not to be "classic" orthoses because they are not fabricated or applied by orthotists. An example is this acutely applied air splint for first aid or emergency, designed to control edema and limit motion at the ankle joint for short-term use (a). Orthotic-like devices are used to supplement various therapeutic regimens in the rehabilitation of patients post-injury and post-surgery. This example provides continuous passive motion to the joints through powered systems attached to orthotic-like components (b, reprinted with permission from Orthomed Medizintechnik, GmbH), and the other one provides a corrective force to gradually regain motion in joints with contractures, (c, reprinted with permission from Joint Active Systems, Inc.).

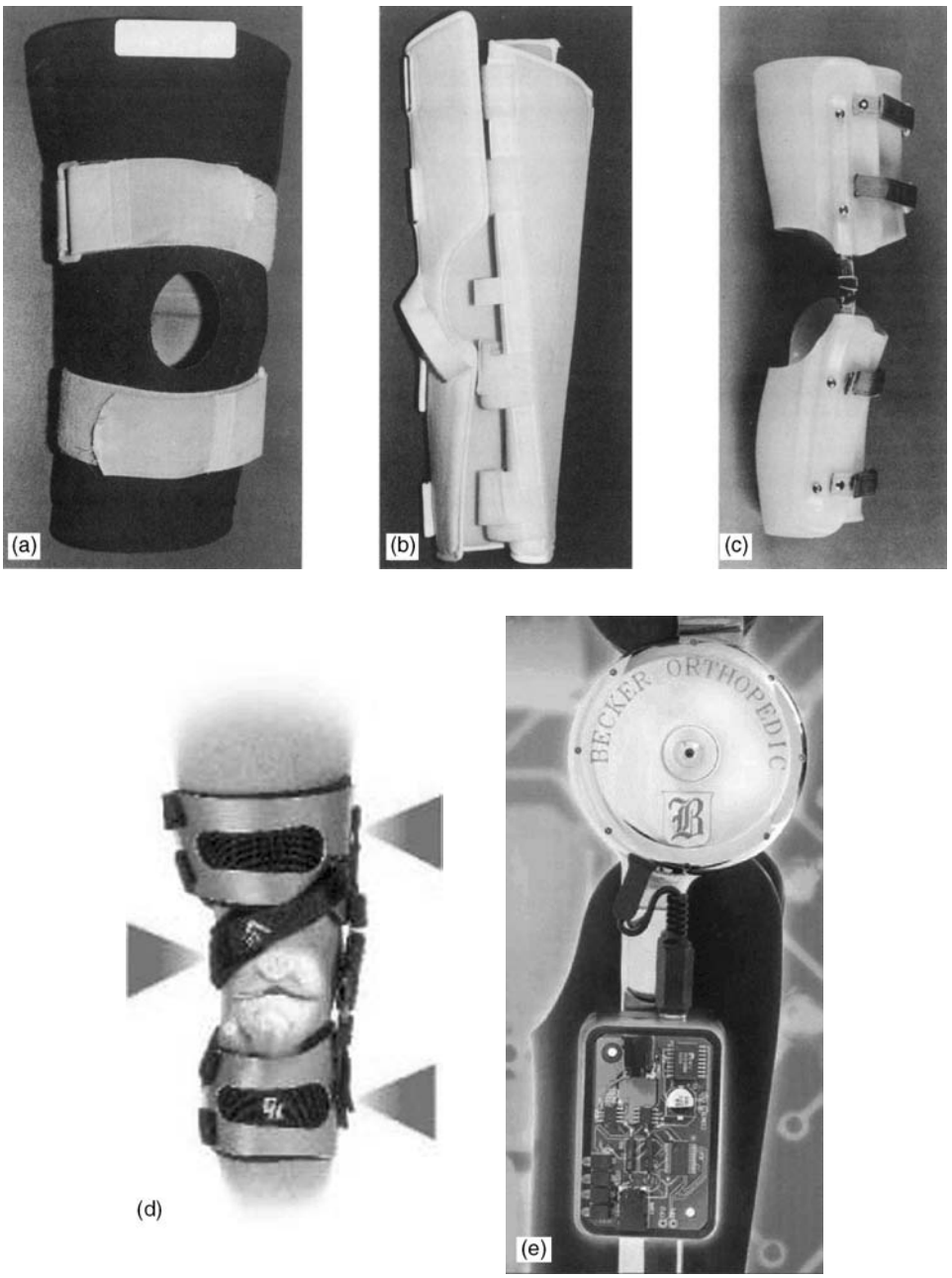


Figure 7. Most KOs are designed for the prevention or protection of knee injuries. Simple soft devices to limit knee motion (a), splints to immobilize the knee temporarily (b), custom orthoses with knee mechanisms to control motion and provide support to the knee (c), and orthoses that provide adjustable corrective forces or moments to the knee to alter knee mechanics (d), reprinted with permission from Generation II USA, Inc.), and even powered knee mechanisms with microprocessor controls, (e), reprinted with permission from Becker Orthopedic, Inc.).

applications not in common use. The description of typical design types and materials also reflects the opinion of the authors on the systems commonly used in orthotics today. Many sophisticated designs and materials are being used in research and may be commonly applied in the future.

Some conventional devices do not strictly meet the nomenclature system because they do not cross a joint or control a joint. They simply are a sleeve that compresses soft tissue to stabilize a fracture or protect a limb with soft-tissue injuries. These devices are only for short-time use until in injury can heal. Examples are shown in Fig. 14.

FUNCTIONAL EVALUATION

Evaluation of musculoskeletal function is the key to adequate prescription criteria or performance criteria, for conventional orthoses. Communication of this evaluation by the orthopedic surgeon to the orthotist is an important step in obtaining agreement on, and optimum use of, orthoses in the rehabilitation of the patient. To this end, the Committee on Prosthetics and Orthotics of the American Academy of Orthopedic Surgeons has devised the technical analysis form that is used for recording the functional evaluation of the patient, documenting the abilities and disabilities of the patient, and forming a

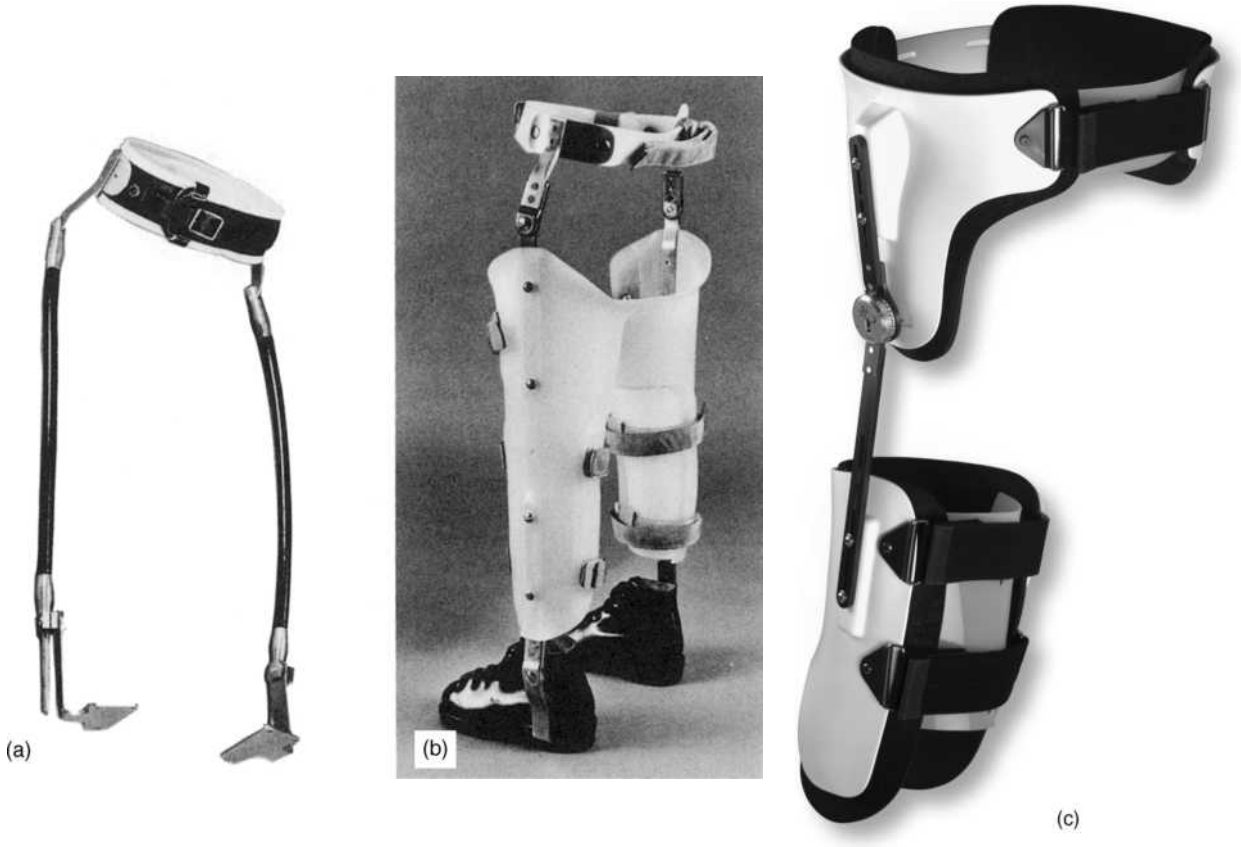


Figure 8. HKAFOs are generally used for the control of deformities. Axial rotation alignment of the lower limbs can be controlled with an HKAFO commonly called a twister brace (a), which uses a flexible cable to provide free motion at the knee and ankle in all degrees of freedom while applying a constant axial torque for rotation correction. An HKAFO for control of valgus at the knee provides three-point support through the soft-tissue interfacing sleeves with as little restriction to ankle and hip flexion and extension as possible (b). An HO is often used after total hip replacement to provide an abduction resistance to the hip during flexion to help prevent hip dislocation until the soft-tissue healing is completed (c, reprinted with permission from Sky Medical, Inc.).

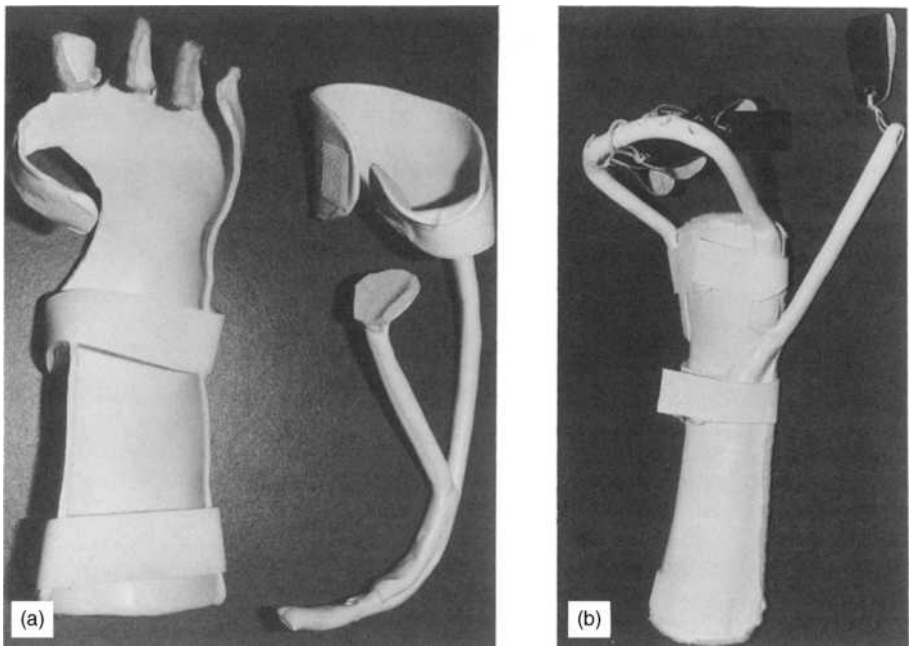


Figure 9. WHOs can be of static (a) or dynamic (b) variety to control, support, resist, or assist motion of the fingers, hand, and/or wrist. These examples are made of low-temperature thermoplastics that can be applied directly to the patient.

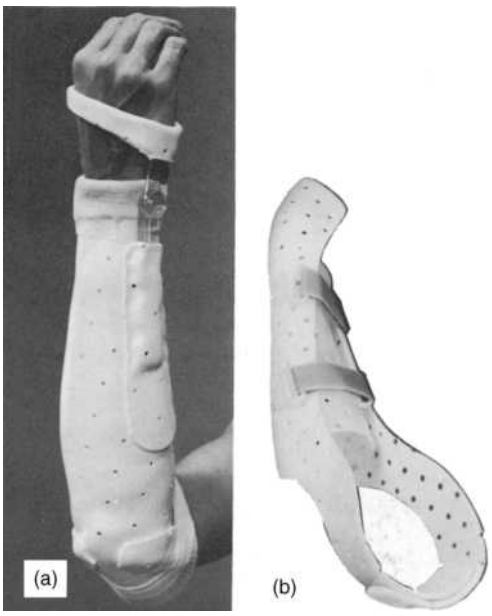


Figure 10. EWHOs are most typically of the custom-fabricated type (a), but they are also available in prefabricated forms (b, reprinted with permission from Maramed Orthopaedic Systems), as shown in these examples of devices for the treatment of Colles' fractures.

recommendation for the orthosis. In a uniform, simplified system, these details can be communicated to the orthotist for construction and fitting of the device and rehabilitation of the patient. Evaluation of each segment of the musculoskeletal system is divided into parts as described in Fig. 14, which shows a typical evaluation form. Note that the range of motion of each joint is indicated on the background of the normal range of motion for that joint and there are places for recording the static position of each joint and part of the skeleton and the deformities that exist in the skeletal structures. The first page of the technical analysis form contains didactic information on the impairments and

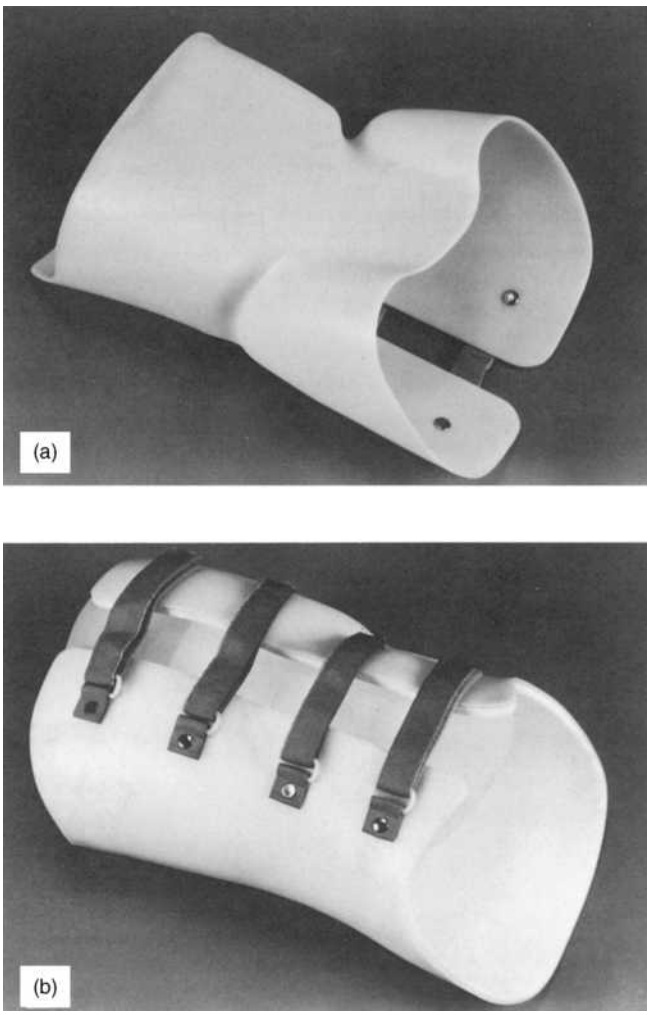


Figure 12. TLSOs are used in a wide variety of forms, most typically a custom-fabricated thermoplastic design for control of spinal deformities, such as this low-profile scoliosis TLSO for control of spinal curvature (a). This anterior closing TLSO is used for instability of the low back or low back pain (b).

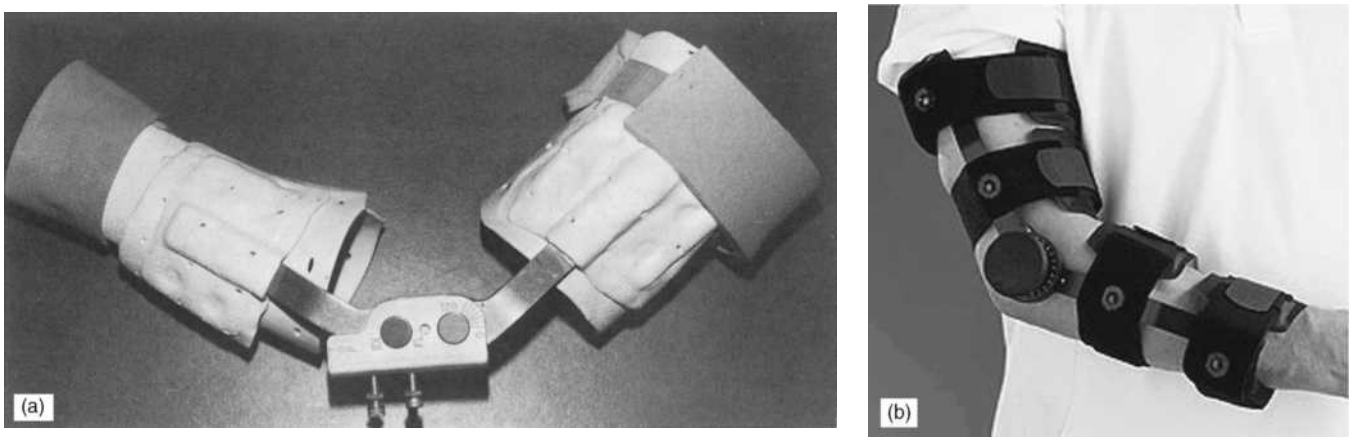


Figure 11. EOs typically are of the custom-fabricated type with prefabricated components for the control of elbow flexion and extension during rehabilitation post-injury (a) or elbow reconstruction, but many are totally prefabricated designs (b, reprinted with permission from Aircast, Inc.).

Figure 13. This CTLSO is designed to provide corrective forces for spinal curvature with thermoplastic skin interfacing components for the pelvis and chin and prefabricated, metal uprights (a) This halo orthosis is designed to protect the cervical spine after acute injury and after spinal fusion (b). The skull pins in the proximal portion of the orthosis are percutaneous components that anchor into the skull through the skin for positive control of head position.

a nomenclature system for their description on the physical evaluation form. A legend for description of fixed deformities, alterations of range of motion, and description of skeletal as well as muscular abnormalities is provided in Fig. 16 for use in the boxes in Fig. 15 for description of the physical examination. On the reverse side of this form, the treatment objectives are outlined and the orthotic recommendation includes a description of the controls for each part of the musculoskeletal system that are recommended as performance criteria for the orthosis. The orthotist is then given the choice of how to accomplish these recommendations in orthotic design and materials. The recommendation form is completed with standard nomenclature describing the control of the joints and musculoskeletal structures: F designates free motion, A assistance, R resistance, S stop, H hold, and L lock controls on motion. A "v" accompanying any of these designations will recommend a variable control of the type described for that degree of freedom (Fig. 17). The degrees of freedom are described across the top of the table for flexion, extension, abduction, adduction, internal and external rotation, and axial loading on the skeletal parts. Also note that a standard nomenclature system is developed to describe the joints that are encompassed by the orthosis. Thus, FO designates an orthosis that controls the foot; AFO describes an orthosis that controls the ankle and the foot, and so forth. In addition to the designations shown on the orthotic recommendation table, it is possible to have a device that simply

Table 1. Lower Limb Orthotic Devices

	Functional control			
Component	Passive	Active	Typical Design Types	Typical Materials
Soft-tissue interface	Alignment to structural members and joints	Feedback load/position	Wraparound or interlocking shells; adjustable pads and straps	Fabric, leather, hand laminates, thermoplastics, foam or elastomer polymers
Structural members	Alignment to joints and soft tissue interfaces, extension blocks	Feedback load/position	Modular uprights; reinforced shell; extension of joints and bands; single axis; I or 2 DOF, cam or drop lock	Metals, hand laminates, prepregs, thermoplastics
Hip joint	Lock, stop, free, assistive, resistive motions	Manual-activated lock	Single axis; 1 or 2 DOF; cam or drop lock	Metals, reinforced polymers, thermoplastics
Knee joint	Lock, stop, free, assistive, resistive motions, adjustable varus/valgus corrective force	Feedback position, load-, manual-, or EMG-activated lock or resistance	Single- or multiaxis; 1 DOF; variable stop; cam, link, or drop lock; friction resistance; elastic assistance; single lateral or posterior upright; medial-lateral upright	Metals, reinforced polymers, thermoplastics
Ankle joint	Lock, stop, free, assistive, resistive motions	Feedback position, load-, manual-, or EMG-activated lock or resistance	Single- or multiaxis; 1 DOF variable stop; cam; friction resistance; elastic assistance; single lateral or posterior upright; medial-lateral upright	Metals, reinforced polymers, thermoplastics
Foot	Alignment		Shoe insert: caliper or stirrup attached to shoe with shank	Metals, reinforced polymers, thermoplastics
FES orstimulation	Electrode placement	Microprocessor, load- or EMG-activated stimulus	Implantable: electrodes, self-contained power and control module, stimulator module RF to external power External: electrodes, stimulator, and controls	Silicone-coated silver-braided steel, silver-impregnated silicone, SS electrodes, epoxy- or silicone-encapsulated electronic components

[a]For examples, see Figs. 1–5, 7, and 8. DOF, degree(s) of freedom; FES, functional electrical stimulation.

Table 2. Upper Limb Orthotic Devices

Component	Functional control: Passive	Functional control: Active	Typical Design Types	Typical Materials
Soft-tissue interface	Alignment to joints and structural members	F feedback load/position	Wraparound or interlocking shells; adjustable padding and straps	Fabric, leather, hand laminates, thermoplastics, foam or elastomer polymers
Structural-members	Alignment to soft tissue interfaces and joints, extension blocks	F feedback load/position	Modular reinforced shells; extension of joints and bands	Metals, hand laminates, prepregs, thermoplastics
Shoulder joint	Lock, stop, free, assistive, motions	Manual-or EMG-activated lock or assistance	Single axis; 1 or 2 DOF; electrical or pneumatic motor; manual or elastic assistance	Metals and reinforced polymers
Elbow joint	Lock, stop, free, assistive, motions	Manual-or EMG-activated lock or assistance	Single- or multiaxis; 1 DOF; electrical or pneumatic motor; manual or elastic assistance; drop or cam lock, extension or mechanical joint	Metals and reinforced polymers
Wrist joint	Lock, stop, free, assistive, motions	Manual-or EMG-activated lock or assistance	Single- or multiaxis; 1 or 2 DOF; electrical or pneumatic motor; manual or elastic assistance; drop or cam lock, extension or mechanical joint	Metals and reinforced polymers
Hand	Lock, stop, free, assistive, motions	Manual-or EMG-activated lock or assistance	Single- or multiaxis; 1 or more joints; 1 DOF; electrical or pneumatic motor; manual or elastic assistance; drop or cam lock, extension or mechanical joint stop in 1 or more DOF	Metals and reinforced polymers
FES orstimulation	Electrode placement	Microprocessor-, load- or EMG-activated stimulus	Implantable: electrodes, self contained power and control module, stimulator module RF to external power External: electrodes, stimulator, and controls	Silicone-coated silver-braided steel, silver-impregnated silicone, SS electrodes, epoxy- or silicone-encapsulated electronic components

[a]For examples, see Figs. 2 and 9–11. DOF, degree(s) of freedom.

Table 3. Spinal Orthotic Devices

Component	Functional control: Passive	Functional control: Active	Typical design types	Typical materials
Soft-tissue interface	Alignment to structural members	F feedback load	Wraparound or interlocking shells; adjustable pads and straps	Fabric, leather, hand laminates, thermoplastics, foam or elastomer polymers
Structural members	Alignment to soft tissue interfaces	F feedback load	Upright adjustable superstructures; reinforced shells; percutaneous pins and belts	Metals, hand laminates, thermoplastics
Electrical stimulation	Electrode placement	Microprocessor-activated stimulus	Implantable: electrodes, self contained power and control module, stimulator module RF to external power External: electrodes, stimulator, and controls	Silicone-coated silver-braided steel, silver-impregnated silicone, SS electrodes, epoxy- or silicone-encapsulated electronic components

[a]For examples, see Figs. 2, 12, and 13.

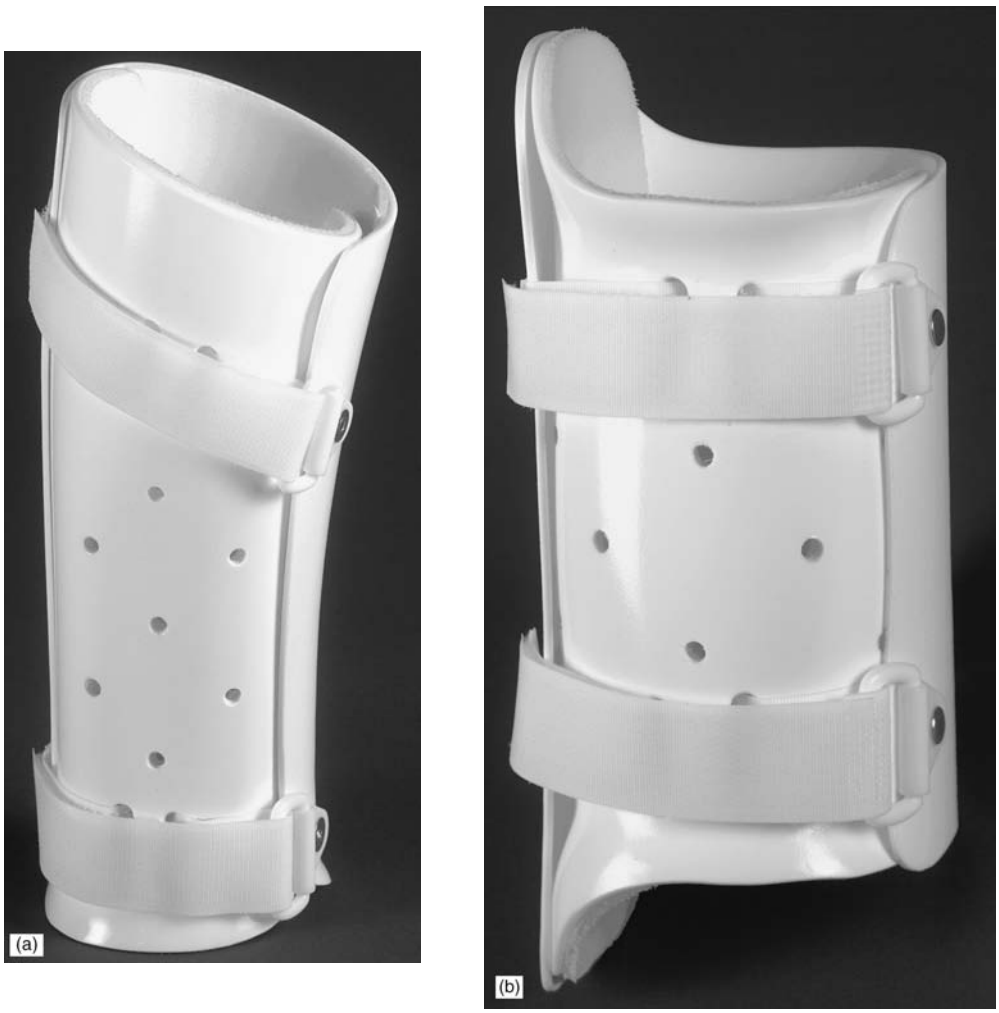


Figure 14. Fracture orthoses that do not span a joint are used for selected humeral fractures (a) and selected isolated ulnar fractures and some soft-tissue injuries (b, reprinted with permission from Sky Medical, Inc.).

controls one joint and does not include all joints distal to it. For example, an orthosis that encompasses and controls only the knee joint is termed a "KO". A similar nomenclature system is shown for the upper limb and for the spine in Figs. 18 and 19, which show the orthotic recommendation and treatment objective portions of the technical analysis form for each of those applications. For examples of orthoses fitting each nomenclature descriptor, see the figures associated with each table. Many orthotic facilities and surgeons have created their own forms for orthotic

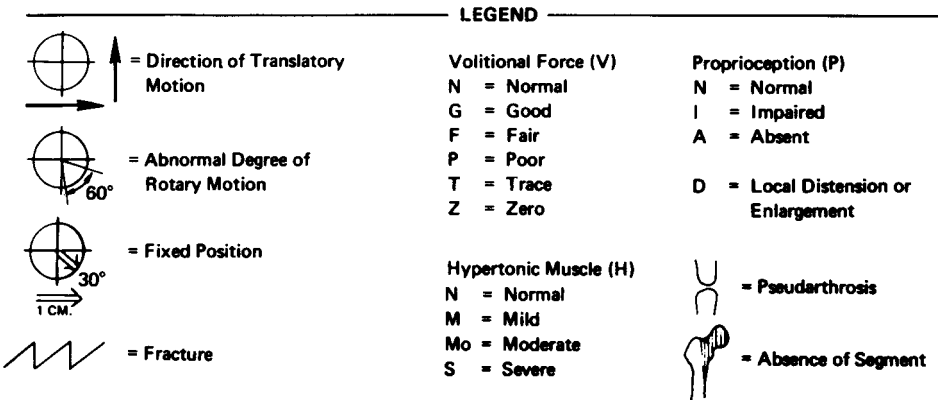


Figure 15. This portion of the technical analysis form for the lower limb demonstrates the graphic manner in which a complete passive and active evaluation of the extremity is recorded to document both normal and abnormal behavior Reproduced by permission from the American Academy of Orthopaedic Surgeons, *Atlas of Orthotics*, 2nd ed., St. Louis, C. V. Mosby, 1985.

Treatment Objectives:

Prevent/Correct Deformity ☐ Improve Ambulation ☐
Reduce Axial Load ☐ Fracture Treatment ☐
Protect Joint ☐ Other _____

ORTHOTIC RECOMMENDATION

LOWER LIMB		FLEX	EXT	ABD	ADD	ROTATION Int.	ROTATION Ext.	AXIAL LOAD
HKAO	Hip							
KAO	*Thigh*							
	Knee							
AFO	*Leg*							
	Ankle	(Dorsi)	(Plantar)					
FO	Foot – Subtalar					(Inver.)	(Ever.)	
	Foot – Midtarsal							
	Foot – Met.-phal.							

Figure 16. The legend for the technical analysis form uses standard descriptors for qualitative and quantitative documentation of the physical examination. Reproduced by permission from the American Academy of Orthopaedic Surgeons, *Atlas of Orthotics*, 2nd ed., St. Louis, C. V. Mosby, 1985.

Treatment Objectives: Prevent/Correct Deformity ☐ Improve Function ☐
Relieve Pain ☐ Other _____

ORTHOTIC RECOMMENDATION

UPPER LIMB		FLEX	EXT	ABD	ADD	ROTATION Int.	ROTATION Ext.	AXIAL LOAD
SEWHO	Shoulder							
EWHO	*Humerus*							
	Elbow							
	Forearm					(Pron.)	(Sup.)	
WHO	Wrist			(RD)	(UD)			
HO	Hand							
Fingers 2-5	MP							
	PIP							
	DIP							
Thumb	CM					(Opposition)		
	MP							
	IP							

Figure 17. The orthotic recommendation portion of the lower limb technical analysis form provides for description of the treatment objective as well as a prescription recommendation for control of musculoskeletal system by the orthosis. Reproduced by permission from the American Academy of Orthopaedic Surgeons, *Atlas of Orthotics*, 2nd ed., St. Louis, C. V. Mosby, 1985.

Treatment Objectives:

Spinal Alignment ☐ Motion Control ☐
Axial Unloading ☐ Other _____

ORTHOTIC RECOMMENDATION

SPINE		FLEX	EXT	LATERAL FLEXION R	LATERAL FLEXION L	ROTATION R	ROTATION L	AXIAL LOAD
CTLSO	Cervical							
TLSO	Thoracic							
LSO	Lumbar							
	(Lumbo sacral							
SIO	Sacroiliac							

Figure 18. The upper limb technical analysis form provides for description of treatment objectives and orthotic prescription with the standard nomenclature system. Reproduced by permission from the American Academy of Orthopaedic Surgeons, *Atlas of Orthotics*, 2nd ed., St. Louis, C. V. Mosby, 1985.

prescription, but they include similar features to those in the accepted standards shown here.

OUTCOME

Orthotic devices are designed to improve the function of persons with musculoskeletal disabilities. The goals of treatment are outlined on the technical analysis form for each application. The optimal outcome is obviously achievement of these goals. In many instances, achievement of these goals is related to the provision of functional independence for many persons who would otherwise be partially or totally dependent. Where total or partial functional independence is not feasible, the goal is to provide a better quality of life. When orthoses are applied acutely for temporary stabilization or temporary protection until healing or recovery from a disabling injury can occur, the goal is often an uneventful recovery. In many instances, the patient returns to employment or functional independence, or function is completely restored before the injury heals. After sufficient healing, the orthosis is usually discontinued. In some cases, orthotic care allows the patient to be discharged from the hospital or transferred to a less expensive support system earlier, reducing the cost of medical care. In most short-term applications of orthotic devices, patients are relatively compliant and cooperative and use the orthoses well. Most long-term applications of orthoses are well accepted by patients, but in a higher percentage of chronic (compared with acute) applications, patients use orthoses for limited activities or periods of time and choose to alter the original treatment goals. Such behavior is also not unusual for the users of many external prosthetic devices. Such patients often develop compensatory means of function (i.e., retraining of the contralateral limb) or use other assistive devices (like wheelchairs) to accomplish their personal goals.

For acute applications of orthoses, there is a strong trend toward the greater use of totally prefabricated systems. For chronic applications, very few prefabricated systems have proved to be adequate; the trend here is to use more thermoplastic materials. Patients find these materials more lightweight, cosmetic, and comfortable.

Another trend helping to facilitate these changes is the increased use of central fabrication facilities. Traditionally, orthotists have measured the patient, designed the orthosis, and completely fabricated the orthosis before fitting and training the patient in its use. Today, orthotists are being trained to devote more attention to the measurement, fitting, and training of patients and less attention to the fabrication of devices. This is because most types of devices can be fabricated in a factory with highly skilled technicians using the prescription criteria and the measurements of the orthotist. This trend helps to reduce costs; if the orthotist spends most of his or her time evaluating, fitting, and training the patients, and the technicians fabricate the devices, the orthotist can care for a much greater number of patients and the consistency of fabrication of the devices is improved considerably. Fabrication of a replacement device for a patient already under the care of an orthotist is often simplified with the availability of a central fabrication facility.

FUTURE

Historically, developments in orthotics have followed developments in the field of external prosthetics. If this continues to hold true, one can anticipate that in the immediate future, orthotics will make increased use of proprioceptive feedback systems, composite materials, microprocessor controls for the dynamics of the orthosis, automation of production facilities and improvements in the design of skin interfacing components, and temporary use of standardized devices for patient training before fabrication and fitting. Also, because orthotic devices are increasingly used for many new applications and means for increasing the number of skilled orthotists needed to meet the demands are not available, there will probably be major increases in the use of prefabricated systems, central fabrication facilities, and/or simplified systems for specific applications that can be handled by other paramedical personnel. There is a growing trend toward greater use of functional electrical stimulation for even the most complex neuromuscular disabilities using microprocessor-controlled multichannel systems to control coordinated muscular activity in limbs with paralysis or severe paresis. Recent advances in electrode design, miniaturized microprocessor systems, and knowledge of musculoskeletal functions are continually producing breakthroughs in research. CAD/CAM methods for automated measurement, modification, and fabrication of custom devices are developing also. The orthotist will find continuing advances in orthotic systems for preventive medicine in sports and the work environment. The rehabilitation team, of which the orthotist is an important member, will include new specialists for biofeedback, rehabilitation engineering, and new branches of therapy and physicians and surgeons from specialties that previously were not involved in rehabilitation.

DEFINITIONS

1. *Ortho*. From the Greed *orthos*, meaning straight or to correct.

2. *Orthosis*. Orthopedic appliance used to straighten, correct, protect, support, or prevent musculoskeletal deformities.

3. *Orthotics (orthetics)*. Field of knowledge relating to the use of orthoses to protect, restore, or improve musculoskeletal function. (Note: This field overlaps the field of mobility devices, including wheelchairs, crutches, and special vehicles, for transportation of persons with musculoskeletal disabilities. Mobility aids will not be discussed in this article.)

4. *Orthotist (Orthetist)*. A person practicing or applying orthotics to individual patients.

5. *Custom versus prefabricated orthoses*. Custom-fabricated orthoses are devices that have components that are molded specifically to the contours of the musculoskeletal structures of a patient. This fabrication can be accomplished by making molds, tracings, or careful measurements of the patient's anatomy for custom shaping in the fabrication of the device. Prefabricated orthoses are made completely from prefabricated components that are fit to the patient in standard sizes. Some systems are preassembled, and some are not. Prefabricated components may be used in custom-fabricated devices along with custom-fabricated components to provide a custom orthosis.

6. *Modular component*. A component that can be assembled from prefabricated parts and can be disassembled and reassembled in different combinations.

7. *Nomenclature*. A standard set of abbreviations for an orthosis related to the anatomic parts that are to be controlled by the orthosis (see Figs. 14–18); for example, AFO is ankle-foot orthosis.

8. *Free motion*. No alteration or obstruction to normal range of motion of an anatomic joint.

9. *Assistance*. Application of an external force for the purpose of increasing the range, velocity, or force of motion of an anatomic joint.
10. *Resistance*. Application of an external force for the purpose of decreasing the velocity or force of motion of an anatomic joint.
11. *Stop*. Inclusion of a static unit to deter an undesired motion in a particular degree of freedom of an anatomic joint. Variable control parameter that can have multiple adjustments without making a structural change.

12. *Hold*. Elimination of all motion in a prescribed degree of freedom of an anatomic joint or anatomic structure.

13. *Lock*. A device that has an optional mode of holding a particular anatomic joint from motion and that can be released to allow motion when desired.

14. *Paralysis*. Loss or impairment of motor function (paresis, incomplete paralysis).

15. *Volitional force*. Voluntarily controlled muscle activity.

16. *Hypertonicity*. High resistance for muscle to passive stretching.

17. *EMG*. Electromyography is a measure of electrical potential changes caused by muscle contraction/relaxation.

See also CARTILAGE AND MENISCUS, PROPERTIES OF; JOINTS, BIOMECHANICS OF; LIGAMENT AND TENDON, PROPERTIES OF.

RESIN-BASED COMPOSITES

MARK CANNON
Northwestern University
Chicago, Illinois

INTRODUCTION

Composite dental restorative filling materials are synthetic resins that have had a revolutionary impact on dental practice. No other dental materials have stimulated such rapid change in an inarguably short period of time. The advent of resin-based composite occurred during the "Golden Age of Dentistry" and mirrored the overall effect that high technology had on society. Virtually everyone has been made aware of "bonding" or "cosmetic dentistry" by the mass media. The success of "bonding" or "cosmetic dentistry" depended on the development of esthetic dental restorative (filling) materials such as resin-based composites and to the advent of dental adhesives, which allowed for the successful placement of the new esthetic materials.

Composite resins have replaced the previous dental restorative materials for anterior teeth, silicate cements, and unfilled acrylic resins, because of superior physical properties. The overall improvement in physical properties has encouraged a great increase in the clinical use of composite resins. Composite resins are now used in virtually all aspects of dentistry. Their application was first limited to simple cavities in anterior teeth (incisors) or the repair of a broken tooth. However, composite resins are now used to cement orthodontic brackets onto teeth, seal pits and fissures in molars and premolars, splint periodontally or traumatically loosened teeth together for stability, repair not just broken teeth, but also porcelain restorations, revitalize and enhance the esthetic quality of discolored or misshapen teeth. Certainly, no other dental material systems offers such a broad range of applications or greater opportunities for the improvement of dental care. Researchers are working diligently to develop newer resins that possess all the necessary qualities required of an ideal dental restorative material, esthetic by nature and resistant to all the deleterious effects of the oral environment.

FABRICATION

Unfilled Acrylic Resins

The autopolymerizable, unfilled acrylic resin was one of the precursors to the development of composite resin materials. The composite resins were introduced to overcome the problems associated with the clinical use of the unfilled resins. The unfilled acrylic resins were supplied as powder and liquid preparations. The powder consisted of the polymer, polymethylmethacrylate and an initiator, benzoylperoxide (1). The monomer consisted mostly of methyl methacrylate, a cross- linking agent (ethylene dimethacrylate), a tertiary amine, and an inhibitor (methylhydroquinone). Although the unfilled resins were considered to be

Acronym	Organic compounds
Bis-GMA	Bisphenol-α-glycidyl methacrylate

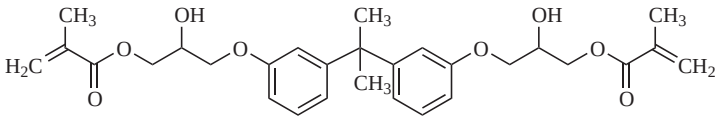


Figure 1. Bis-GMA.

polishable and esthetically acceptable, numerous problems were reported clinically. Pulpal irritation and sensitivity were noted and probably related to microleakage. The microleakage was undoubtably due to the high degree of polymerization shrinkage and the higher coefficient of thermal expansion of unfilled resins to tooth structure. The expansion and contraction of the unfilled resin restoration would "percolate" deleterious salivary contents leading to pulpal sensitivity, marginal discoloration and secondary caries.

Resin Based Composite

Dentistry has long recognized the need for an esthetic anterior restorative material. The unfilled acrylic resins and the silicate cements were generally considered to be clinical failures and inadequate for many situations commonly seen in dental patients (veneering of discolored teeth, repair of badly fractured teeth, etc.) The composite principle, using filler particles with a proper index of refraction, thermal expansion coefficient similar to enamel and a resin adhesion capability was advocated in 1953 by Paffenbarger et al. (2) The term composite may be defined as a three-dimensional (3D) combination of at least two different materials with a distinct interface (3). A composite resin restorative material consists of three phases: the organic phase (matrix), the interfacial phase (coupling agents), and the dispersed phase (filler particles).

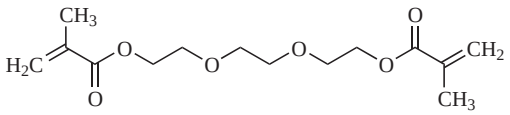
Matrix Phase

One of the main components of all composites is the addition reaction product of bis(4-hydroxyphenol), dimethylmethane, and glycidylmethacrylate known as "Bis-GMA" (4). Bis(4-hydroxyphenol) is an oligimer with a fairly high molecular weight. Figure 1 shows its generalized structural formula.

Other aromatic dimethacrylates are also used in composite resin materials (5). These oligimers include Bis-MA (2,2-bis[4-(2-methacryloyloxyphenol] propane), Bis-EMA (2,2-bis [4-(3-methacryloyloxyphenol] propane) and Bis-PMA (2,2-bis[4-(3-methyacryloyloxypropoxy) phenol] propane. To decrease the high viscosity of the Bis-GMA resin systems, low viscosity liquids, such as TEGDMA (triethyleneglycol dimethacrylate) and EGDMA (ethyleneglycol dimethacrylate) are used. Figure 2 shows the structural formula for TEGDMA.

Inhibitors are necessary to prevent the premature polymerization of the dimethacrylate oligimers and viscosity controllers. An example of an inhibitor would be BHT (2,4,6- tritertiarybutylphenol). Autopolymerizable composite resins (paste mixed with paste or powder mixed with liquid two component systems) utilized a thermochemical initiator, most commonly, benzoylperoxide. The decomposition of benzoyl peroxide results in free radicals that initiate polymerization. The initiator would be present in only one portion of the two-component system. The other portion would contain the accelerator, such as, a tertiary aromatic amine. When the two components of an autopolymerizable composite resin are mixed, the tertiary aromatic amine (e.g., *N*,*N*-dimethyl *p*-toluidine) interacts with benzoyl peroxide to produce free radicals necessary to initiate the matrix polymerization (6). The main disadvantage of auto-cure resin-based composites was the inability of the clinician to control the setting time. Once mixed, the material would irreversibly begin to polymerize whether the clinician was totally prepared or not.

The ability to initiate the polymerization reaction when most desired has been achieved by the introduction of photochemically initiated composites. The first photochemically initiated composite resins used in dentistry required ultraviolet (UV) radiation (7). The UV radiation source employed a high pressure mercury arc and provided a simple mechanical shutter to mask off the radiation when not in use. The UV radiation was emitted through a light guide to expose the composite resin. The effective wavelength was between 364 and to 367 nm. An organic compound that generates free radicals when exposed to 365-nm wavelength electromagnetic radiation, such as, benzoin alkyl ether, was added to composite resins in place of the thermochemical initiator (benzoyl peroxide). Visible light initiated composite resins depend on a diketone (e.g., camphoroquinone) and an organic amine (e.g., *N*,*N*- dimethylaminoethylmethacrylate) to produce free radicals that result with polymerization initiation. The diketone absorbs electromagnetic radiation in the 420–450 nm wavelength range. The unit that provides the electromagnetic radiation (dental curing light unit) usually consists of a light source, a filter that selects the range of transmitted wavelength and a light tube that directs the light beam to the composite. The units generally emit wavelength in the 400–550 nm range. There is some concern over the potential of damage to the eyes of dental operators (8). Indeed, health concerns



TEGDMA Triethyleneglycol dimethacrylate

Figure 2. TEGDMA.

over the use of UV radiation for polymerization initiation initially encouraged the investigation into using "visible" light instead.

Dispersed Phase

Inorganic filler particles were added to dental resins as early as 1951. It was a number of years, however, before composite materials were generally utilized by dentists. The research of Bowen (4) improved the mechanical properties of filled resins. Bowen treated silica powder with 1.0% aqueous solution of tris(2-methoxyethoxy) vinyl silane to which sodium hydroxide was added and the resultant slurry dried at 125°C. Peroxide was added in an acetone solution and the solvent evaporated. This treatment resulted with an organophilic silica powder. The first dental composites utilized fillers, such as E-glass fibers, synthetic calcium phosphate, fused silica, soda-lime glass beads, and other glass formulations (9). The commonly used filler particles currently used as reinforcing materials are quartz, colloidal silica, lithium aluminum silicate, and silica glasses containing either strontium or barium (10). Quartz was the most successful of the commercial fillers due to its index of refraction and inert nature, however, its hardness and large particle size made polishing difficult resulting with the transition to softer glass fillers (11). In addition to being radiopaque, the softer glass fillers facilitate easier polishing of the set composite resin and the production of fine filler particles for incorporation into the matrix. There is, however, a potential toxicity problem with barium glasses and the radiopaque glasses may be susceptible to degradation in water (12). Chemical processes may also be used to synthesize filler particles, such as the colloidal silicon dioxide pyrogenic particles (13). The colloidal silicon dioxide particles may be fabricated by burning silicon tetrachloride in the presence of an hydrogen and oxygen gas mixture. Particles ranging from 0.007 to 0.14 μm result. The small size of the colloidal silicon dioxide filler particle, when incorporated in the dispersed phase, allows the polishing of a resin-based composite to a smooth finish. However, the small filler particle size prevents high filler loading. This tends to decrease the filler fraction of resin-based composites manufactured with a colloidal silicon dioxide dispersed phase (Table 1).

Hydroxyapatite and amorphous calcium phosphate filler particles have been advocated for resin-based composites (15,16). The reported advantage of these fillers is the potential for reminerializing adjacent tooth structure.

Filler-Matrix Interface

Transfer of stress from the dispersed-phase filler particles through the ductile matrix phase should occur to improve mechanical properties. The bond between the dispersed filler phase and the organic matrix may be achieved by two different methods, chemically or mechanically. The mechanical retention of the dispersed phase is achieved by sintering particles or fibers together, or by etching away a phase of a glass filler particle leaving a porous surface. Monomer may then flow into the porous surface creating a mechanical interlocking. Chemical bonding to the dispersed phase may be achieved by using coupling agents, such as an epoxy silane (17). Two of the silane agents are γ-glycidoxypropyltrimethoxysilane and γ-methacryloxypropyltrimethoxysilane.

Silane coupling may involve hydrolysis of the methoxy groups with bound surface water on the dispersed phase-filler particle or with aluminol or silanol groups of the dispersed phase filler particle. During polymerization of the composite resin, the unsaturated carbon double bonds are available to react with the matrix. *In vitro* tests suggest, that a general rule, as filler volume is increased, wear is reduced regardless of filler treatment (18).

PHYSICAL AND MECHANICAL PROPERTIES

Physical Properties

Wettability partially determines the marginal microleakage and surface staining of a composite. Wettability may be determined by the contact angle formed by a drop of water on the resin-based composite surface. Composite resins are considered hydrophilic because the advancing contact angle of water on composite surface is $\sim 65^\circ$. The water sorption value is 0.6 mg/cm^2 and water uptake of composites is a function of polar groups in the polymer structure. The thermal conductivity of composites closely matches that of dentin and enamel. The coefficient of thermal expansion usually averages $26-40 \times 10^{-6}$cm/cm per°C for the range of 0–60°C, and recently marketed composite resins even more closely match the values normally obtained for tooth structure. A composite resin with a thermal coefficient of expansion similar to tooth structure (10×10^{-6}cm/cm per°C) would theoretically suffer from less margin microleakage. When there is a difference in the values, composite resin restorations may expand or contract more than tooth structure during temperature changes resulting with gap formation between the two substances. Margin microleakage may then occur. Polymerization contraction (% by volume) for the typical composite resin averages 3.2–3.8%. The amount of contraction due to polymerization is effected by the types of oligimers used in the matrix phase and the filler volume. As the composite resin shrinks, it pulls away from the walls of the cavity preparation in the tooth. This polymerization contraction encourages the ingress of salivary contaminants and bacteria (microleakage). Not all studies have shown a statistical correlation between polymerization shrinkage and microleakage (19).

Methods have been employed to reduce the stresses associated with polymerization shrinkage. Pulse or two-step light activation for polymerization initiation has been recommended (20,21). Other methods include incremental filling, directed shrinkage, void incorporation, filler and matrix alteration.

Mechanical Properties

Virtually all composite resin manufacturer's spend considerable resources on testing their own and competitor's products for mechanical properties. Dentists are continually

Table 1. List of Materials and Percentage of Fillers by Weight Determined by Ashing in Air[a]

Material	Classification[b]	Manufacturer	Batch and Shade	% Fillers by Weight
Aeliteflo	VLC Hyb Flow CS	Bisco, Inc Itasca, IL, USA	039317 (A3)	54.9
Amelogen	VLC Hyb Univ CS	Ultradent products, UT, USA	2CPM (A2)	72.9
Arabesk	VLC Hyb Univ CS	Voco, Cuxhaven, Germany	70500 (A3)	71.6
Arabesk-Flow	VLC Hyb Flow CS	Voco, Cuxhaven, Germany	82777 (A3)	61.8
Arabesk-Top	VLC Hyb Univ CS	Voco, Cuxhaven, Germany	81594 (A3)	71.5
Ariston-pHc	VLC 'Smart Material'	Vivadent, Schaan, Liechtenstein	A00001 (–)	74.8
Brilliant-Dentin	VLC Hyb Univ CS	Colténe, Whaledent, Switzerland	GE931 (A3)	75.6
Brilliant-Enamel	VLC Hyb Univ CS	Colténe, Whaledent, Switzerland	GE902 (A3)	75.4
Charisma-F	VLC Hyb Univ CS	Heraeus Kulzer, Wehrheim, Germany	23 (A20)	76.4
Charisma-PPF	CC Hyb Univ CS	Heraeus Kulzer, Wehrheim, Germany	2 (A10)	68.3
Clearfil Photo Post	VLC Hyb Univ CS	Kuraray, Osaka, Japan	0035A (UL)	84.7
Clearfil Photo Ant.	VLC Hyb Univ CS	Kuraray, Osaka, Japan	0024C (A3)	59.9
Colténe-SE	VLC Hyb Univ CS	Colténe, Whaledent, Switzerland	FBJOl (A3)	71.3
Concise	CC Conventional CS	3M, St. Paul, MN, USA	19970303(U)	80.2
Dyract-Flow	VLC Flow CM	Dentsply De Trey, Konstanz, Germany	9809000103 (A2)	55.4
Elan	VLC CM	Sybron/Kerr, Orange, USA	805872 (A3,5)	71.2
EXI-119 (Z-250)[c]	VLC Hyb Univ CS	3M, St. Paul, MN, USA	030998 (A3.5)	77.4
EXI-120 (P-60)[c]	VLC Hyb Pack CS	3M, St. Paul, MN, USA	030998 (A3,5)	78.9
F-2000	VLC CM	3M, St. Paul, MN, USA	19970905 (A3)	80.5
Glacier	VLC Hyb Univ CS	Southern Dental Industries, Australia	60506 (B3)	78.2
Metafil-CX	VLC Microfine CS	Sun Medical, Shiga, Japan	71201 (A3,5)	41.7
Pertac-II	VLC Hyb Univ CS	Espe, Seefeld, Germany	00634764 (A3)	70.0
Polofil-Molar	VLC Hyb Univ CS	Voco, Cuxhaven, Germany	63596(U)	78.5
Quadrant Anterior	VLC Microfine CS	Cavex Haarlem, Holland	22C (A2)	58.6
Quadrant Posterior	VLC Hyb Pack CS	Cavex Haarlem, Holland	30C (A2)	65.2
Revolution	VLC Hyb Flow CS	Sybron/Kerr, Orange, USA	710669 (A3)	53.9
Silux-Plus	VLC Microfine CS	3M, St. Paul, MN, USA	6DH (U)	54.8
Solitaire	VLC Hyb Pack CS	Heraeus Kulzer, Wehrheim, Germany	26 (A30)	64.3
Spectrum	VLC Hyb Univ CS	Dentsply De Trey, Konstanz, Germany	9608244 (A3)	75.3
Surefil	VLC Hyb Pack CS	Dentsply De Trey, Konstanz, Germany	980818 (A2)	79.4
Tetic-Ceram	VLC Hyb Univ CS	Vivadent, Schaan, Liechtenstein	900513 (A3)	75.7
Tetric-Flow	VLC Hyb Flow CS	Vivadent, Schaan, Liechtenstein	901232 (A3)	64.0
Wave	VLC Hyb Flow CS	Southern Dental Industries, Australia	80608 (A3)	60.7
Z-100	VLC Hyb Univ CS	3M, St. Paul, MN, USA	19960229 (UD)	79.6

[a]Reprinted with permission from Ref. 14.
[b]Visible light cured = VLC, chemically cured = CC, composite = CS, compomer = CM, flowable = Flow, packable = Pack, universal = Univ, hybrid = Hyb.
[c]The P-60 (Posterior composite) and the Z-250 (Antero-Posterior composite), marketed by 3M were included in this study as experimental composites EXI-120 and EXI-119.

bombarded with advertisements touting a composite resin's "compressive strength". It cannot be denied that mechanical properties are important for consideration. On the other hand, it is quite evident that clinical success may not be solely correlated with laboratory results or easily predictable. The oral environment is rather hostile to inorganic or foreign materials. Corrosion or degradation occurs with most metals and resins used by dentists. Composite resins demonstrate a change in roughness of the surface with laboratory aging suggesting degradation (22). Mechanical properties of the composite resins have improved considerably the last decade. Unfortunately, the exact relationship between mechanical properties and clinical success has yet to be determined.

The compressive strength of composite resins is greatly superior to their tensile strength. The clinical significance of this is unclear. It may be assumed that tensile strength is of major concern when placing composite resins in posterior teeth as a sliding tooth cusp will cause shear stresses (23). Many manufacturers claim that the compressive strength of their composite resins equal silver amalgam (silver filling material). However, the clinical relevance of this high compressive strength is not known due to a lack of data demonstrating the minimum strength required to resist the forces of occlusion and mastication. The modulus of elasticity is an important factor to consider when a composite resin is used on a stress bearing area. A composite resin with a very low modulus will deform under occlusal stress and increase microleakage or transfer the stress to the supporting tooth. This does not mean that low modulus composites cannot be used when well supported by adjacent tooth structure (14).

Composite resins may be classified by the component materials that comprise each of the three phases. The matrix phase may be described as being either hydrophilic or, as in the case of some experimental resins, hydrophobic. The interfacial phase is not as amenable as the matrix phase to classification. It is possible to describe the coupling mechanism as being chemical (either polymeric or silane treated) or mechanical. However, the most appropriate basis for classification would appear to be the size and chemical composition of the dispersed phase (24). Composite resins may be classified into five main categories: traditional hybrid, microfill (pyrogenic silica),

microhybrid, packable, and flowable (25). Hybrid resin composites have an average particle size of 1–3 μ and are 70–77% filled by volume. They present with good physical properties compared to the microfilled resin composites. Some composite resins used softer, radiopaque glass fillers averaging 1–5 um (26). Microhybrid resin composites have an average particle size of 0.4–0.8 μ and are 56–66% filled by volume. Microhybrid resin-based composites generally have good physical properties, improved wear resistance, and relatively high polishability. Microfill composite resins utilize pyrogenic silicon dioxide with an average particle size of 0.04–0.1 μ, but due to the increase in viscosity associated with an unacceptably low filler fraction (35–50% by volume), a modified method of manufacture was developed. Microfiller-based complexes consists of prepolymerized particles (pyrogenic silicon dioxide mixed with resin matrix and cured) and resin matrix with a equal concentration of dispersed microfiller (pyrogenic silicon dioxide) as was in the prepolymerized particles. This formulation allows for an increased filler fraction without the difficulty in manipulation due to high viscosity. The microfiller composite resins exhibit high polishability (27). The smoother the surface of the composite resin, the less that surface wear occurs and plaque accumulates (28,29). Packable composites have an average particle size of 0.7–20 μ and are 48–65% filled by volume. Their higher viscosity is considered desirable in establishing interproximal contacts and is the result of a higher percentage of irregular, fibrous, or porous filler in the resin matrix. Flowable composites have an average filler particle size of 0.04–1 μ and are 44–54% filled by volume. Their reduced filler volume decreases viscosity to ease placement in thin increments or small cavity preparations. They are also more flexible, with a decreased modulus of elasticity, to absorb stress in certain applications (Table 2).

CLINICAL APPLICATION

Bonding is a general term that is used to describe the joining, uniting, or attaching of adhesives to an adherend (6). Bonding describes the attachment of two materials, but not the mechanism by which the bonding occurred. A bonded assembly may be attached together by mechanical means or by physical and chemical forces. The bond that occurs between composite resin and tooth enamel is mechanical and is achieved by the process of acid etching. The acid-etch technique was developed by Buonocore in 1955 for use in dentistry (31). An 85% phosphoric acid solution was applied to the teeth of volunteer subjects and greatly enhanced the bond strength of an unfilled acrylic resin to the enamel surface. Retief demonstrated the effectiveness of a 50% phosphoric acid pretreatment of enamel for bonding and also the reversibility of the process by the saliva's remineralization of the enamel (32–35). Bonding is now universally accepted by the dental profession.

Enamel bonding ushered in the age of cosmetic dentistry and popularized the use of composite resins. It was also desirable to develop materials that would adhere to the second layer of tooth structure, the dentin. This is especially applicable in situations where the available enamel for bonding is minimal. Commercially available dentin bonding systems have recently been developed. Bonding may occur by two different means, micromechanically or chemically. Adhesion may be chemically established to either the inorganic or organic portions of dentin. The dentin bonding systems have improved considerably in a short period of time showing great promise for reducing margin microleakage and increasing restoration retention. Composite resins have rapidly established themselves as an important segment of a dentist's restorative armamentarium. Composite resins were initially used solely as a replacement for the unfilled acrylic resins and silicate cements. Use was limited to the incisors where esthetics were of paramount importance. Even then, many dentists considered all resins to be strictly temporary at best.

The patient was often advised to consider a permanent restoration that was metallic, such as a gold foil (placed by compaction of overlapping gold increments directly into the prepared cavity of a tooth). The failure rate of the early resins was considered unacceptable by many dentists and composite resins had to prove themselves "worthy".

The introduction of "bonding" reduced the severity and occurrence of microleakage with the composite resin restorations. Bonding also provided a conservative means to retain a composite resin restoration without substantial sacrifice of tooth structure (by cutting undercut areas into the tooth). Composite resins soon became useful in restoring traumatically fractured teeth (34–38). Malformed or hypoplastic teeth were reconstructed using composite resins and bonding (39,40). The pigments incorporated

Table 2. Classification and Physical Properties of Resin-Based Composites[a]

			Physical Properties[b]		
Composite Type	Average Particle Size (Micrometers)	Filler Percentage (Vol %)[b]	Wear Resistance	Fracture Toughness	Polishability
Microfill	0.04–0.1	35–50	E	F	E
Hybrid	1–3	70–77	F↔C[c]	E	G
Microhybrid	0.4–0.8	56–66	E	E	G
Packable	0.7–20	48–65	P↔G[c]	P↔E[d]	P
Flowable	0.04–1	44–54	P	P	F↔G[d]

[a]Reprinted with permission from Ref. 30.
[b]Sources: Kugel,[47] Wakefield and Kofford[50] and Leinfelder and colleagues[53].
[c]E: Excellent; G: good; F: fair; P: poor.
[d]Varying among the same type of resin-based composite.

into the composite resin matrix provide a wide range of natural shades. Additional tinted and opaquing resins are available that enable the dentist to achieve a life-like result closely mimicking enamel. The bonding of composite resins as a thin veneer to enamel was advocated for an aesthetic technique of restoring discolored teeth. Young discolored or malformed teeth could be given a natural appearance by covering the enamel surface with a thin "veneer" of the appropriately shaded composite resin (41,42). The public eventually became aware of dentistry's newest advancement and "cosmetic bonding" was "born".

Preventive dentistry benefited by the development and introduction of pit and fissure sealants. A pit and fissure sealant material is a BIS-GMA based resin that is introduced into the caries susceptible pits and fissures of teeth forming a barrier against the action of decay producing bacteria. Pit and fissure sealant resin is retained or bonded to the enamel surface of the teeth by the acid-etch technique. In the 1980s, a conservative cavity preparation that utilizes composite resins and bonding had been proposed by Simonsen (43). This technique involves the removal of only decayed tooth structure with the composite resin restoration bonded to the enamel surface sealing the pits and fissures from future decay. Simonsen termed this technique the "Preventive Resin Restoration" (43). Composite resins are also utilized as cementing or as luting agents.

Acrylic preformed laminate veneers were bonded with composite resins to discolored teeth as aesthetic restorations (44). Porcelain veneers have been bonded to discolored teeth with composite resins since 1983 (45). Composite resins have also been used to bond orthodontic brackets, and fixed partial dentures (Maryland Bridges) to teeth (46). Splints to stabilize periodontally or traumatically loosened teeth are constructed of composite resin. Restorations of decayed posterior teeth (molars) are more and more done with composite resins and bonding (47). A number of dentists and patients are reportedly concerned over the potential toxic effect of mercury, one of the constituents of silver amalgam filling material (48). Studies have shown that dentists with the highest exposure to silver amalgam also have the highest urinary mercury levels (49). The concern about mercury toxicity may yet be the necessary impetus to finally develop an ideal composite resin (50).

New low shrink composite resins consist of a bis-GMA matrix with a strontium glass filler of an average size of 1.1 μ. The composite is 86.5% filled by weight and has a compressive strength of 267.5 MPa. The composite has a volumetric shrinkage of 1.4%, which reduces the shrinkage stresses induced by polymerization. Also recently introduced have been the no-rinse conditioners (self -etching primers). No rinse conditioners are very convenient for the dentist and are very useful in patients with severe "gag" reflex (51).

The self-etching primer consist of a penetrating, polymerizable monomer and an acidic component. The pH of the self-etching primer is between 0.6 and 2, which is acidic enough to etch cut or prepared enamel (52). Several self-etching primers currently available will not etch unprepared enamel leaving resin extensions improperly bonded (53). Self-etching primers that leave an acidic intermixed zone will demonstrate osmotic blistering (54). Osmotic blistering is the separation of the intermixed zone due to water penetrating the hydrophilic primers. Water is pulled from the dentin because of the high concentration of ionic species retained in the primer when no-rinse conditioners are used. The osmotic gradient from the ionic species and acid monomers creates pockets or blisters of water between the adhesive and composite resin. In addition, both the filler and the resin matrix may suffer degradation from the free acid radicals if the intermixed zone is left unbuffered. The shear bond strength of resin-based composite or sealants is reportedly less with the self-etching primers than the "total etch" technique, a pretreatment of both enamel and dentin with 32–40% phosphoric acid solution or gel (55). Immobilization of an antibacterial component into the resin matrix has also been achieved creating antibacterial composite resins. A new monomer, MDPB, has been produced by combining a quaternary ammonium with a methacryloyl group and incorporating it into the resin matrix for copolymerization with other monomers. The antibacterial effect is on contact only and does not dissipate by dissolution into the saliva (56).

BIBLIOGRAPHY

1. Anusavice K., editors. Phillip's Science of Dental Materials. 11th ed. Philadelphia: W.B. Saunders; 2003.
2. Paffenbarger G, Nelson R, Sweeney W. Direct and indirect filling resins: A review of some physical and chemical properties. J Am Dent Assoc 1953;47:516–521.
3. Stanford J. The current status of restorative resins. Dent Clin N Am 1971;15:57–66.
4. Bowen R. Dental filling Material comprising vinyl silane treated fused silica and a binder consisting of a reaction product of bisphenol and glycidyl acrylate. US Patent 3, 066, 112. Nov. 1962.
5. Ruyter I, Sjouik I. Monomer Composition of dental composites and sealants. J Dent Res (Spec Iss A) Jan. 1978; 57:249, Abstr. 700,
6. Craig R. Chemistry, composition, and properties of composite resins. In: Horn H, editor. Dental Clinics of North America. Philadelphia: W. B. Saunders; 1981. p 223.
7. Rock W. The use of ultra-violet radiation in dentistry. Br Dent J 1974;136:455–458.
8. Fan P, et al. Visible light-cured composites and activating units. J Amer Dent Assoc 1985;110:100–103.
9. Bowen R. Compatibility of various materials with oral tissues, I: The components in composite restorations. J Dent Res 1979;58:1493–1503.
10. Zinck J, Norling B, Buchanan R. Composite resins systems: A comparison. Dent Stud 1982;61:51–55.
11. Bowen R, Cleek G. A new series of X-ray opaque reinforcing fillers for composite materials. J Dent Res 1972;51:177–82.
12. McKinney J, Wu W. Chemical softening and wear of dental composites. J Dent Res 1985;64(11):1326–331.
13. Jorgensen K, Asmussen E. Occlusal abrasion of a composite restorative resin with ultra-fine filler—an initial study. Quintessence Int 1978;6:73.
14. Sabbagh J, Vreven J, Leloup G. Dynamic and static moduli of elasticity of resin based materials. Dent Mater 2002;18:64–71.
15. Arcis R, et al. Mechanical properties of visible light-cured resins reinforced with hydroxyapatite for dental restoration. Dent Mater 2002;18:49–57.

16. Skrtic D, Antonucci J, Eanes E. Improved properties of amorphous phosphate fillers in remineralizing resin composites. Dent Mater 1996;12:295–301.
17. Sterman S, Marsden J. Silane coupling agents as integral blends in resin-filler systems. Mod Plastics 1963;49(11):125.
18. Lim B, Ferrocane J, Condon J, Adey J. Effect of filler fraction and filler surface treatment on wear of micro-filled composites. Dent Mater 2002;19:1–11.
19. Rosin M, et al. Polymerization shrinkage-strain and microleakage in dentin-bordered cavitites of chemically and light cured restorative materials. Dent Mater 2002;18:521–528.
20. Yoshikawa T, Burrow M, Tagami J. A light curing method for improving marginal sealing and cavity wall adaptation of resin composite restorations. Dent Mater 2001;17:359–366.
21. Lim B, et al. Reduction of polymerization contraction stress for dental composites by two-step light-activation. Dent Mater 2002;18:436–444.
22. Powers J, Fan P. Surface degradation of composite resins. J Dent Res (Spec. Iss. A) 1979;58:328.
23. Soderholm K. Filler systems and resin interface. In: Vanherle G, Smith D, editors. Posterior Composite Resin Dental Restorative Materials. The Netherlands: Peter Szule Publishing; 1985. p 149.
24. Lutz F, Phillips R. A classification and evaluation of composite resin systems. J Pros Dent 1983;50:480–488.
25. Deliperi S, Bardwell D. An alternative method to reduce polymerization shrinkage in direct posterior composite restorations. JADA 2002;133:1387–1398. October.
26. Dogon I, Cross M, Douglas W. Clinical and laboratory studies on a fine grind composite. J Dent Res 1982;61:214 (Abstr. No. 320).
27. Christensen R, Christensen G. In vivo comparison of a microfilled and a composite resin: A three year report. J Pros Dent 1982;48(6):657–663.
28. Shampanier A. A comparative Study of the surface resistance of various composite filling materials to toothbrushing abrasion. Masters dissertation. Chicago, (IL):Northwestern University Dental School; 1978.
29. Sotrez S, Van Huysen G, Gilmore N. A histologic study of gingival tissue response to amalgams, silicate, and resin restorations. J Perio 1969;40:543–546.
30. An alternative to reduce polymerization shrinkage in direct posterior composite restorations. J Am Dent Assoc 2002; 133.
31. Buonocore M. A simple method of increasing adhesion of acrylic filling materials to enamel surface. J Dent Res 1955;34:849–853.
32. Retief D. Effect of conditioning the enamel surface with phosphoric acid. J Dent Res 1973;52:333–341.
33. Retief D. A comparative study of three etching solutions. J Oral Rehabil 1974;1:381–390.
34. Jordan R, Suzuki M, Charles D, Gwinnett A. Esthetic and conservative restoration of the fractured incisor by means of microfilled composite materials. A O 1981;74:51–59.
35. Black J, Retief D, Lemons J. Effect of cavity design on retention of Class IV composite resins restorations. J Am Dent Assoc 1981;103:42–46.
36. Roberts M, Moffa J. Restoration of fractured incisal angles with an ultraviolet activated sealant and a composite resin: a case report. J Dent Child 1972;39:364–365.
37. Buonocore M, Davila J. Restoration of fractured anterior teeth with ultraviolet—light polymerized bonding materials: a new technique. J Am Dent Assoc 1973;86(6):1349–1354.
38. Hill F, Soetopo A simplified acid-etch technique for the restoration of fractured incisors. J Dent 1977;5:207–212.
39. Jordan R, Suzuki M, Gwinnett A. Restoration of fractured and hypoplastic incisors by the acid etch resin technique: A three year report. J Am Dent Assoc 1977;95:795–803.
40. Black J. Esthetic restoration of tetracycline-stained teeth. J Am Dent Assoc 1982;104:846.
41. Cooley R. Laminate preformed resin veneer. Proceedings of the American Dental Association Meeting; Anaheim, CA, Oct. 1978.
42. Faunce F, Myers D. Laminate veneer restoration of permanent incisors. J Am Dent Assoc 1976;93:790–792.
43. Simonsen R. Prevention resin restoration: Three year results. J Am Dent Assoc 1980;100:535–539.
44. Cooley R. Status report on enamel bonding of composite, preformed laminate, and laboratory fabricated resin veneers. J Am Dent Assoc 1984;109:762–764.
45. Horn L. A new lamination: Porcelain bonded to enamel. Clin North Am 1983;27:671–684.
46. Williams V, Dehehy G, Thayer K, Boyer D. Acid-etch retained cast metal prostheses: A seven year retrospective study. J Am Dent Assoc 1984;108:629–631.
47. Ernst C, Buhtz C, Rissing C, Willershausen B. Clinical performance of resin composite after 2 years. Compendium 2002;23(8):711–724.
48. Abraham J. The effect of dental amalgam restorations of blood mercury levels. J Dent Res 1984;63(1):71–73.
49. Naleway C, Sakaguchi R, Mitcheell E, Muller T, Ayer W. Hefferren J. Urinary mercury levels in U.S. dentists, 1976–1983: Review of Health Assessment Program. J Am. Dent Assoc 1985;111:37–42.
50. Lobner D, Asrari M. Neurotoxicity of dental amalgam is mediated by zinc. J Dent Res 2003;82(3):243–246.
51. Cannon M. Advances in pediatric esthetic dentistry. Compendium 2003;24(8):34–39.
52. Wang Y, Sharp L, Suh B. The morphology study of several self etching adhesive systems. J Dent Res 2002; Abstr. 1898: (Spec. Iss. A).
53. Tay F, et al. Single step adhesives are permeable membranes. J Dent Sep–Nov 2002;30(7–8):371–382.
54. Tay F, et al. Osmotic blistering in enamel bonded with one-step self-etch adhesives. J Dent Res 2004;83(4):290–295.
55. Fuks A, Eidelman E, Lewinstein I. Shear strength of sealants placed with non-rinse conditioning compared to a conventional acid etch rinse technique. J Dent Child Sept–Dec 2002; 239–242.
56. Imazato S. Antibacterial properties of resin composites and dentin bonding systems. Dent Mater 2003;19:449–457.

See also BIOMATERIALS FOR DENTISTRY; BIOMATERIALS, TESTING AND STRUCTURAL PROPERTIES OF; ULTRAVIOLET RADIATION IN MEDICINE.

RESPIRATOR.

See VENTILATORS, ACUTE MEDICAL CARE.

RESPIRATORY MECHANICS AND GAS EXCHANGE

JAMES LIGAS
University of Connecticut
Farmington, Connecticout

INTRODUCTION

Respiratory mechanics applies the principles of solid and fluid mechanics to pressure, flow, and volume measurements obtained from the respiratory system. The utility of the resulting mathematical models depends on how well they guide clinical decisions and prove consistent with the results of new experiments. Although very sophisticated

models of respiratory system structure and function exist, our focus here is on simple concepts used for pulmonary function testing and mechanical ventilation.

RESPIRATORY SYSTEM STRUCTURE

The airways, or bronchial tree, conduct air to the small air sacs, or alveoli, which comprise the lung parenchyma. The upper, larger airways are lined with ciliated epithelium and mucus secreting cells to warm, humidify, and filter small particles from the inhaled gas. The heated and humidified air is ultimately exhaled, resulting in a water loss from the body of $\sim 1\ \mathrm{L} \cdot \mathrm{day}^{-1}$. When the upper airways are bypassed by tracheostomy or by intubation for mechanical ventilation, drying of the respiratory tract can result in inspissation of secretions and obstruction of the airways. Extracorporeal humidification of the compressed, zero-humidity gases used in mechanical ventilation is a necessity (1).

The cross-sectional area of the trachea is roughly the size of a United States quarter dollar (0.25¢). Alveolar surface area is about the size of a tennis court, so that over a very short distance the airways branch repeatedly to bring inhaled gases into contact with a large surface area for diffusion into the blood. The branching is generally dichotomous and asymmetric. The number of branches from the trachea to an alveolus varies from 6 to 30. Small nerve fibers coursing through the walls of airways cause glandular secretion and determine muscle tone. The blood vessels bringing venous blood to the alveoli for gas exchange follow a similar branching structure so that air and blood flow are closely matched for efficient exchange of gases (2).

The last generation of bronchioles ends in sprays of alveolar ducts and sacs. There are some 300 million alveoli, each $\sim 300\ \mu$m in diameter at full inflation, forming a network of interconnecting membranes. Within those membranes pass the smallest vessels, the alveolar capillaries. Microscopic studies show that the membrane consists of two parallel tissue sheets separated by a series of tissue posts, much like the deck of a parking garage. Red blood cells spend <1 s in this structure, and diffusion of respired gases occurs across the 0.3–3.0 μm barrier between air space and red cells in the capillaries. At low intravascular pressures, some parts of this network remain collapsed, yet can be recruited if the pressure in the pulmonary vessels increases. In addition, lymphatics also drain the interstitial space and follow the structure of the bronchial tree. They are capable of removing large volumes of extravasated fluid if necessary.

A thin membrane, the pleura, covers the outer surface of the lungs and is reflected to line the inside surface of the thoracic cavity. Between these two pleural surfaces is a space 6–30 μm in thickness. A very small amount of fluid is normally present, providing lubrication that allows the lungs to slide freely over the interior of the chest wall (3).

The thoracic cage itself consists of the spine, sternum, ribs, and associated muscles of the chest wall. The diaphragm separates the abdominal cavity from the thoracic cavity and is the major muscle effecting resting ventilation (4). As muscular contraction expands the thoracic cavity, the lungs follow, filling with air. When the subject is at rest exhalation is passive. The muscles relax, the thoracic cavity decreases in volume, and air flows out of the lung.

In addition to the solid mechanics of these structures, analyses of fluid flows are important. The behavior of fluids spans many regimes: convection-dominated flows in the large airways, gaseous diffusion at the alveolar level, interphase diffusion through the capillary walls into the blood, and viscous flows in the alveolar ducts, capillary spaces, and intrapleural space.

STATIC MECHANICS OF THE RESPIRATORY SYSTEM

The simplest test of respiratory system behavior is to ask a subject to inhale or exhale to different lung volumes (Fig. 1). The tidal volume, V_T, is the volume of air that moves in and out of the lungs in a normal breath. At the end of a normal exhalation all the muscles of respiration are relaxed. With the glottis open, no forces are exerted on the respiratory system and atmospheric pressure exists both in the alveolar spaces and on the body surface. The lung volume under those conditions is called the functional residual capacity, or FRC. One can exhale beyond FRC by forcing more air out of the lung. The amount that can be forced out is called the expiratory reserve volume (ERV). Even at that point, there is still residual air in the lungs: the residual volume (RV). If one inhales maximally, the volume of air in the lung is the total lung capacity, or TLC. Similar to ERV, there is an inspiratory reserve available to us beyond our usual tidal volume: the inspiratory reserve volume, or IRV. The total amount of air that could be inhaled from FRC is called the inspiratory capacity, and the total amount that could be exhaled from TLC is called the vital capacity (VC). Figure 1 shows that quantities labeled "capacity" are the sum of two or more quantities labeled "volume". Together with statistical tables relating normal values to gender, age, and size, this data provides one measure of any muscular weakness or structural impairment that might limit the ability to move air into or out of the respiratory system.

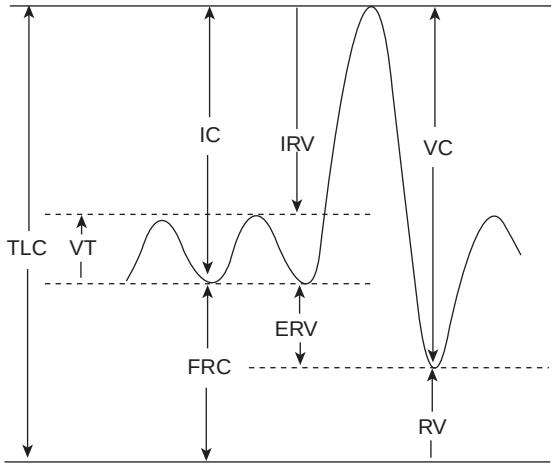


Figure 1. Lung volumes and capacities.

Beyond simply measuring these volumes, one can ask how much force is required to change the volume by a certain amount. These experiments have been performed upon excised lungs and intact animals and subjects.

Parenchyma

Experiments that applied static pressure to an excised, air-filled lung supported in air showed that the relationship between lung volume and the applied transpulmonary pressure difference, that is, airway pressure minus pressure at the pleural surface, is dependent on the volume history of recent expansion (Fig. 2). To achieve any given volume, the required pressure difference is greater for inflation than for deflation. A large part of this behavior is due to surface tension at the gas–liquid interface in the alveoli and smallest airways (5,6). When the lung is degassed and then filled with and surrounded by saline, inflation pressures are reduced and much of the hysteresis is eliminated (Fig. 2). The surface-active agent, or surfactant, responsible for most of the hysteresis consists of a phospholipid protein complex secreted by certain cells found in the alveolar epithelium. Surface tension–area data from experimental systems utilizing films of surfactant indicate that surface tension varies over the range of 2–50 dyn $\cdot$ cm^{-1}, decreasing markedly as the surface area of the film decreases, and is independent of cycling frequency. Surfactant lowers the surface tension below that for a pure water–air interface, and therefore decreases the pressure necessary to inflate the lungs and keeps alveoli from collapsing during exhalation. Infants born prematurely can lack surfactant, making respiration difficult. The development of an artificial surface-active agent delivered by aerosol was an important advance in neonatal care (7).

Because one often deals with changes about a specified lung volume, the elastic behavior of the lung is routinely described in terms of the tangent to the pressure–volume curve at that point, referred to as the *local compliance* (C_L) of the lung,

$$C_L = dV/dP \quad (1)$$

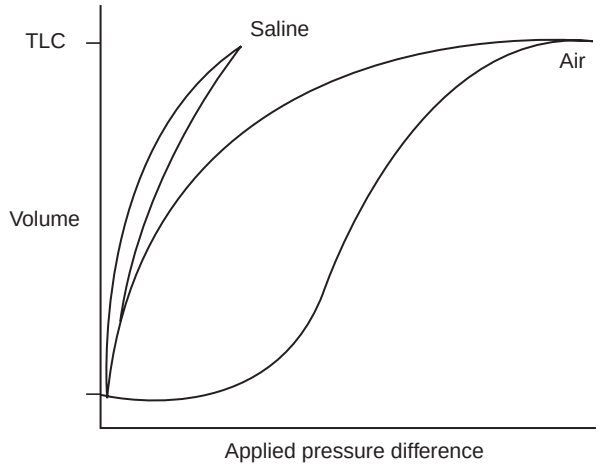


Figure 2. Pressure–volume curves for excised lungs filled with air and saline.

where P is the transpulmonary pressure (alveolar minus pleural surface pressure). Because of the nonlinear nature of the pressure–volume relationship, the compliance varies with lung volume. Thus the simplest model of lung elasticity is a single parameter derived from the transpulmonary pressure–volume curve of the liquid- or air-filled lung. The compliance for an air-filled lung will reflect both the mechanical properties of the tissue and the effects of surfactant.

Intrapleural Space

For spontaneously breathing subjects, the stresses generated by muscle contraction cause thoracic cavity expansion and are transmitted to the lungs through the intrapleural space. The thin fluid layer present between the membrane lining the lung (the visceral pleura) and the membrane lining the inside of the thoracic cavity (the parietal pleura) must in some way transmit those forces. Early models of this coupling postulated the concept of an intrapleural pressure (8). Because the minimal volume attainable at zero distending pressure for an excised lung (V_0) is below the residual volume of the intact subject, and because functional residual capacity is less than the volume of the thoracic cavity when the lungs are removed, the simplest model was that there must be a negative pressure, that is, a pressure less than atmospheric pressure in the pleural space (Fig. 3). In mechanical terms, any shear stresses in the minute amount of pleural fluid were neglected and the mechanics were modeled as a normal stress, the intrapleural pressure (P_{pl}).

Measurements of intrapleural pressure in intact animals or human subjects were traditionally made by placing

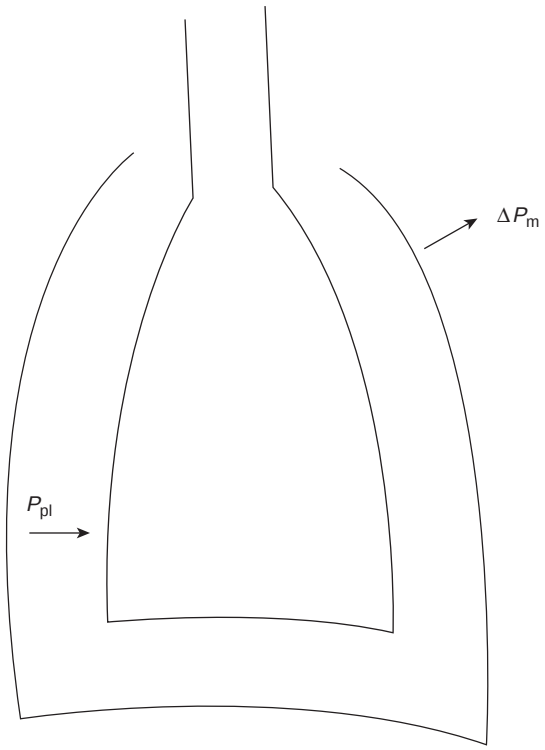


Figure 3. Lung–chest wall model. See text for details.

a catheter tipped with a small latex balloon into the thoracic esophagus. The balloon was partially inflated, but not stretched. The theory is that both the balloon and the esophageal wall are flaccid structures, and so pressure changes measured by this manometry system would reflect changes in intrathoracic pressure (9). These measurements were undertaken in intact animals and subjects to try to separate the mechanics of the chest wall from the properties of the lung itself.

Thoracic Cage

In a manner similar to the parenchyma, the chest wall was conceived of as an elastic structure and the slope of the pressure–volume curve for the relaxed chest wall was called the chest wall compliance. The active forces exerted by the muscles (ΔP_m) are conceived of as a normal force applied to change the intrapleural pressure. During inhalation, expansive chest wall forces decrease the intrapleural pressure, causing expansion of the lung. If the subject forcefully exhaled, or used the muscles to exhale to a volume below functional residual capacity, the intrapleural pressure would increase.

These were the simplest models for the static mechanics of the respiratory system. Pressure differences at various points in the system are related to volume changes, and the tissue properties were modeled as a simple elasticity although the compliance could be dependent on lung volume.

DYNAMIC EVENTS IN THE RESPIRATORY SYSTEM

Respiration requires the flow of gases into and out of the lungs. Many studies investigated the nature of the fluid flow, but early attention focused on a very simple model: that of a "resistance" to air flow. If a subject inhales to total lung capacity and then exhales to residual volume, the volume exhaled is the vital capacity. However, if performed with maximal expiratory effort, this experiment is called the forced vital capacity (FVC) maneuver (10). Mathematical analysis of the resulting data showed that the volume exhaled was almost exponentially related to time:

$$V = \mathrm{FVC}(1 - e^{-kt}) \tag{2}$$

Where FVC was the volume eventually exhaled, t is time, and k was the constant in the exponential. Investigators acquainted with electrical analogues were quick to point out that this resembled the discharge of a capacitor C through a resistor R, with the time constant

$$k = 1/RC \tag{3}$$

The obvious analog was to equate C with the compliance of the respiratory system and R with the resistance to air flow. The pressure drop between the airway opening and the pleural surface was the driving force, so that the simple model for exhalation became:

$$P_{ao} - P_{pl} = R\,dV/dt + 1/CV \tag{4}$$

Where d/dt represents the time derivative, V the volume change of the lung, and C the compliance. The convention is that $dV/dt < 0$ represents exhalation. This equation fit conveniently with the classification of respiratory diseases into those that were restrictive and those that were obstructive. Restrictive diseases, such as muscle weakness, deformities of the chest wall, neurologic illnesses, and fibrosis of the lungs, were those that altered lung volumes and/or respiratory system compliance. Obstructive diseases such as asthma and bronchitis were those that interfered with the ability to quickly move air into, or especially out of, the lungs, leading to a high resistance.

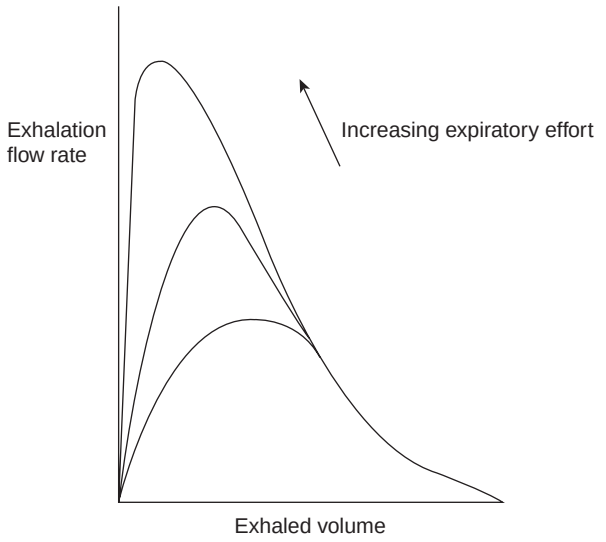


Figure 4. Maximum expiratory flow-volume maneuver at different degrees of effort.

Further support for these concepts came from the Maximal Expiratory Flow–Volume (MEFV) maneuver in which a subject performs the FVC maneuver while exhaling through a pneumotachograph. A plot of flow versus its integral (Fig. 4) has a straight-line portion, consistent with an exponential relationship, because if the pressure difference is constant the time derivative of equation 2 gives

$$dV/dt = kV \tag{5}$$

that is, a linear relationship between volume and flow rate. Once again, normal values are a function of race, gender, age, and size (11).

Equal Pressure Point Concept. Equation 4 would suggest that as the applied pressure difference is increased, the flow rate will increase indefinitely. However, the linear portion of the MEFV curve is relatively independent of effort. Beyond a certain point, trying to exhale more vigorously has no effect on increasing the expiratory flow. The idea of an "equal pressure point" was proposed to explain this flow limitation (12) (Fig. 5). If the subject closes the glottis, the pressure at the glottis will be equal to that in the alveoli because no flow occurs. Equation 4 shows that the pressure measured at the airway opening, or in this case just below the closed glottis, will be equal to the intrapleural pressure plus V/C. The quantity V/C is the elastic recoil force due to stretch in the alveolar walls. If the subject then performs a forced exhalation, intrapleural pressure

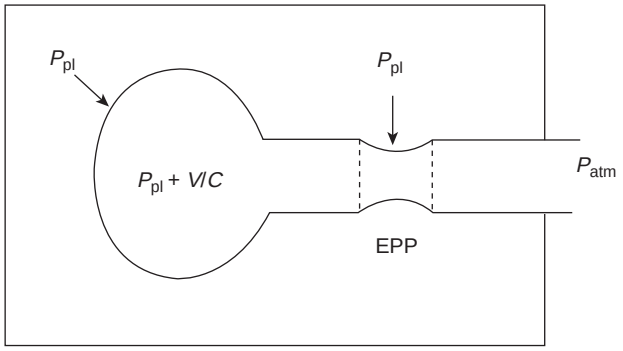


Figure 5. Equal pressure point concept. See text for details.

will be greater than atmospheric pressure and alveolar pressure is greater than this by V/C. Yet atmospheric pressure exists at the airway opening, so that at some point between the alveoli and the airway opening the pressure in the airway is equal to intrapleural pressure. This is the equal pressure point: The transmural pressure across the airway wall is zero, and if the EPP occurs in smaller airways unsupported by cartilage, they can collapse, impeding the flow. The model postulates that this occurs in such a fashion that the harder one tries to exhale, the more the airways are compressed, so that flow is limited. At a constant intrapleural pressure, as lung volume decreases during exhalation the equal pressure point moves toward the alveoli. Although the exact mechanism for flow limitation is undoubtedly much more complex (13,14), this model served as a simple explanation for the observed phenomenon. It also explained why flows would be reduced in patients with diseases such as emphysema, where airway obstruction from mucous and inflammation is not a major factor. In emphysema, fewer alveolar walls mean less elastic recoil, and so the equal pressure point occurs in smaller airways earlier in the exhalation.

MORE COMPLEX MODELS

These simple models underlie much of pulmonary function testing and mechanical ventilation. However, there is an extensive literature presenting more sophisticated analyses. Some of these are multicompartment versions of the simple models in that they postulate differing compliances and resistances in different regions of the lungs. Others propose nonlinear resistance and compliance elements, for example, a variable resistance representing the smaller, collapsible airways and a constant resistance for the more rigid upper airways (15). Some models are more consistent with physics in that they apply the principles of continuum mechanics, namely, conservation of mass, momentum, and energy, to the respiratory system. To do so, investigators have often had to restrict both the scope of the experiments and the portions of the respiratory system they model. Others have chosen to formulate input–output analyses without attempting to model the physical system (16). To date, none of these models has achieved wide acceptance or clinical utility.

THE PULMONARY VASCULATURE

Analyses of blood flow and mechanical changes in the pulmonary vasculature began with a concept analogous to airway resistance: the vascular resistance. The right ventricle pumps blood to the lungs, and the pulmonary veins return the blood to the left atrium. Measurements of mean pulmonary artery pressure (MPAP) and estimates of pressure in the left atrium (LAP) together with the determination of cardiac output (CO) allow the calculation of resistance as pressure drop is divided by flow:

$$\mathrm{PVR} = (\mathrm{MPAP} - \mathrm{LAP})/\mathrm{CO} \tag{6}$$

This is the clinically used model. Analyses of the "sheet flow" concept (17) have been used to explain the mechanics of blood flow through the lungs, but once again clinical applicability has been limited.

GAS TRANSPORT AND EXCHANGE

Although understanding the mechanical behavior of the respiratory system is important, ultimately one must both understand normal gas transfer, and then account for the effects of diseases on gas exchange. These concepts are among the most difficult to master in all of physiology. They bear detailed discussion because in practice physicians use the results of gas-exchange measurements to infer changes in mechanics rather than measure such changes directly. Once again, simpler concepts are used clinically although more complex and physically correct models do exist.

Respiratory control centers in the brain sense the partial pressures of carbon dioxide and oxygen in the blood and drive the respiratory muscles to move air into and out of the lungs. The amount of air that must move to the gas-exchanging alveolar surfaces each minute is known as alveolar minute ventilation, V_A. Some of the inhaled gas resides in the larger airways, where no gas exchange can occur. The volume of these airways is known as the anatomic dead space, V_D. If the gas were to flow in and out of the airways with no mixing, a concept known as "plug flow", the total amount of air that is inhaled is composed of that which reaches the alveoli and that which is wasted in that it never reaches a gas-exchanging surface,

$$V_{\min} = V_A + f\,V_D \tag{7}$$

Where f is the respiratory rate. Unless oxygen tensions are very low, carbon dioxide is the major driving force for respiratory effort. The partial pressure of carbon dioxide in the blood for normal people is $\sim$ 40 Torr (5.320 kPa) and is closely regulated near that level.

The structure of the airways and blood vessels is such that local air flow and blood flow are closely matched: for the normal lung the ratio of ventilation to perfusion for any portion of the lung is nearly uniform even though different areas of the lung receive markedly different amounts of air and blood. The pulmonary vessels are capable of constriction in response to low oxygen tension, a process known as "hypoxic pulmonary vasoconstriction".

This process tends to keep local blood flow well matched to local air flow (18). So, for the purposes of estimating gas transfer, one can lump all 300 million alveoli together and treat the normal lungs as if there were only one large alveolus receiving all the blood flow and all the air flow. One can then estimate (1) what level of alveolar ventilation would be required to maintain a normal partial pressure of carbon dioxide, and (2) what level of blood oxygenation would be expected for a given inhaled oxygen concentration.

Partial Pressures of Gases Related to Volume and Composition

It is molecules of gas that flow in the respiratory system, not volumes of gas. Pressure, volume, and the number of molecules are related by the equation of state:

$$PV = nRT \tag{8}$$

Where P is the pressure, V is the volume, n is the number of moles of gas, T is the temperature, and R is the gas constant. When physiologists speak of volume rather than moles, they must specify the pressure, temperature, and humidity at which that volume is measured. Two sets of conditions are used in respiratory physiology. The first is BTPS, or body temperature and pressure, saturated (with water vapor). Volumes given as BTPS are those that the gases would occupy if they were at 37 °C, standard atmospheric pressure (760 Torr or 101.080 kPa), and fully saturated with water vapor. The partial pressure of water vapor at 37 °C, the normal human body temperature, is 47 Torr or 6.266 kPa. Volumes given as STPD are standard temperature and pressure, dry, that is, at 0 °C, standard atmospheric pressure, with all the water vapor removed. The equation of state can be used to convert between conditions.

As an example, consider what would occur if an 80 kg human were to use glucose as the only fuel. A moderately ill person requires ~25 cal · kg^{-1} · day^{-1} (104.6 J). The metabolism of glucose generates roughly 4 cal · g^{-1} (16.75 J) and consumes an amount of oxygen equal to the carbon dioxide produced:

$$C_6H_{12}O_6 + 6O_2 \leftrightarrow 6CO_2 + 6H_2O \tag{9}$$

To generate 2000 cal (8373.6 J) would require use of 500 g, or 2.8 mol, of glucose/day. This would in turn use 16.8 mol of oxygen and generate 16.8 mol of carbon dioxide.

Using the equation of state, or remembering Boyle's law, we know that 1 mol of gas at STPD occupies 22.4 L. During a day, our patient would use 376 L of oxygen and generate an equal amount of carbon dioxide. On a per-minute basis, oxygen consumption and carbon dioxide production would be

$$V_{O_2} = V_{CO_2} = 260\,\text{mL} \cdot \text{min}^{-1}$$

These are volumes at STPD. What would the volumes be at BTPS? At STPD,

$$n_{O_2} = (760\,\text{Torr})(260\,\text{mL})/R(273\,^\circ\text{C}) \tag{10}$$

These same molecules heated to body temperature and humidified would occupy a larger volume. Under these new conditions, the pressure due to the dry gases is lower and the temperature is higher.

$$n_{O_2} = (760 - 47)V/R(310\,^\circ\text{C}) \tag{11}$$

The number of molecules is the same, so equating 9 and 10 yields

$$V = (760 * 260/273) * (310)/(760 - 47) = 314\,\text{mL} \tag{12}$$

For these calculations, glucose was the sole fuel used to meet the daily energy requirement. However, normally fat forms part of the fuel supply and because the units of a fatty acid chain lack the oxygen atom:

$$\text{— } CH_2 \text{ — or, if there is a double bond, = CH —}$$

fat as an energy source consumes more oxygen per carbon dioxide produced, and therefore lowers the respiratory quotient (RQ). We generally estimate

$$\text{RQ} = V_{CO_2}/V_{O_2} = 0.8$$

when performing calculations, so that if

$$V_{O_2} = 310\,\text{mL} \cdot \text{min}^{-1}\text{BTPS}$$

Then,

$$V_{CO_2} = 250\,\text{mL} \cdot \text{min}^{-1}\text{BTPS}$$

Alveolar Minute Ventilation

With each breath, we inhale surrounding air with 21% oxygen. We do not use all of it (only about a quarter) and exhale the remainder. The amount of air we need each minute is usually determined by the need to rid ourselves of carbon dioxide. What volume of gas would have to reach the alveoli each minute to remove 250 mL · min^{-1} of carbon dioxide while keeping blood, and hence alveolar, carbon dioxide tension at 40 Torr?

The fraction of alveolar gas that is carbon dioxide must be

$$\text{FA}_{CO_2} = V_{CO_2}/V_A \tag{13}$$

But partial pressures are related to the proportion of the gas in the total mixture:

$$P\text{A}_{CO_2} = F\text{A}_{CO_2}(P_{atm} - P_{H_2O}) \tag{14}$$

Combining 12 and 13,

$$P\text{A}_{CO_2} = V_{CO_2}/V_A(P_{atm} - 47) \tag{15}$$

Some textbooks show $P\text{A}_{CO_2}$ multiplied by a constant, 0.863, in this equation. In those books, CO_2 production is given at STPD and V_A is at BTPS. The constant results from using the equation of state to convert V_{CO_2} to BTPS.

Substituting standard atmospheric pressure of 760 Torr or 101.325 kPa, a desired alveolar CO_2 tension of 40 Torr (5.332 kPa), and CO_2 production of 250 mL · min^{-1} leads to the conclusion that alveolar ventilation must be

$$V_A = 4.45\,\text{L} \cdot \text{min}^{-1}\text{BTPS}$$

If the inhaled oxygen is 21% of the respired gases and oxygen utilization is 310 mL min^{-1}, less than a one-third of the inhaled oxygen is consumed and the rest is exhaled. Alveolar minute ventilation is therefore usually determined by the need to exhale the carbon dioxide produced.

Expected Oxygen Tension

We next ask what degree of blood oxygenation we would expect if all the alveoli had the same ventilation/perfusion ratio, that is, if our assumption that the normal lung can be modeled as a single alveolus is correct. Consider the steps involved in inhalation:

1. Gas is brought into the nose and upper airways, heated to body temperature, and humidified.

   The gases we inhale are almost completely composed of nitrogen and oxygen at the local barometric pressure. When water vapor is introduced, the partial pressure of oxygen is the fraction of oxygen inhaled times the sum of the partial pressures of the dry gases, which is barometric pressure minus the water vapor pressure:

   $$P\mathrm{I}_{\mathrm{O}_2} = (P_{\mathrm{atm}} - 47)F\mathrm{I}_{\mathrm{O}_2} \quad (16)$$

   Where the subscript I denotes inhaled.
2. This gas reaches the alveoli, where a certain amount of oxygen is taken up and carbon dioxide is added. Although not strictly true (the correction is relatively minor), the simplest approach is to consider how much oxygen is used compared to carbon dioxide delivered. The respiratory quotient can be used

   $$V_{\mathrm{O}_2} = V_{\mathrm{CO}_2}/\mathrm{RQ} \quad (17)$$

   and an estimate of how much the alveolar P_{O_2} falls compared to the inhaled P_{O_2} becomes

   $$P_{\mathrm{O}_2} = P\mathrm{A}_{\mathrm{CO}_2}/\mathrm{RQ}$$

   giving

   $$P\mathrm{A}_{\mathrm{O}_2} = P\mathrm{I}_{\mathrm{O}_2} - P\mathrm{A}_{\mathrm{CO}_2}/\mathrm{RQ}$$

   or

   $$P\mathrm{A}_{\mathrm{O}_2} = (P_{\mathrm{atm}} - 47) * F\mathrm{I}_{\mathrm{O}_2} - P\mathrm{A}_{\mathrm{CO}_2}/\mathrm{RQ} \quad (18)$$

   For an RQ of 0.8, normal CO_2 tension, and breathing room air (21% O_2) this gives

   $$P\mathrm{A}_{\mathrm{O}_2} = 100\,\mathrm{Torr} \text{ or } 13.300\,\mathrm{kPa}$$

Note that the calculations of alveolar minute ventilation and expected alveolar oxygen tension were based upon the hypothesis that for a normal lung, ventilation and perfusion are closely matched. Some 300 million alveoli were modeled as a single unit in terms of gas exchange. Miraculously, for the normal lung this turns out to work quite well. The difference between measured arterial oxygen tension ($P\mathrm{A}_{\mathrm{CO}_2}$) and calculated alveolar oxygen tension ($P\mathrm{A}_{\mathrm{CO}_2}$) is known as the alveolar–arterial oxygen gradient, or A–a gradient. In young normal subjects it is usually $<$10 Torr (1.330 kPa).

The Role of the Cardiovascular System

Blood delivered by the heart brings carbon dioxide to the lungs for elimination and oxygen to the cells of the body. Figure 6 shows the relationship between the partial pressures of the gases and the amounts of those gases in the blood. For simplicity's sake, these are shown as single curves although the amount of one gas present does affect the carrying capacity for the other somewhat (the Bohr and Haldane effects). The curves are very different. That for carbon dioxide is almost linear, while that for oxygen is sigmoid shaped. This difference is crucially important. Together with the fact that the major regulator of alveolar ventilation, is carbon dioxide, mismatches in the distribution of ventilation and perfusion caused by acute lung injuries will have less effect on carbon dioxide tension in the blood than they will on oxygenation, as we shall see below.

The cardiac output is $\sim$5 L $\cdot$ min^{-1}. Blood with a hemoglobin content of 14 g $\cdot$ dL^{-1} exposed in the lungs to 100 Torr (13.330 kPa) partial pressure of oxygen will contain $\sim$195 mL of O_2 per liter of blood. Therefore, 965 mL of oxygen will be pumped to the body every minute. If oxygen consumption is 310 mL $\cdot$ min^{-1}, 655 mL of oxygen will return through the venous circulation unused. This translates to a venous hemoglobin saturation of $\sim$ 70% with a venous blood oxygen tension of 40 Torr (5.320 kPa).

The carbon dioxide tension of arterial blood is regulated near 40 Torr (5.320 kPa). The arterial blood content of carbon dioxide is therefore $\sim$350 mL $\cdot$ L^{-1}. If carbon dioxide production is 250 mL $\cdot$ min^{-1} and cardiac output is 5 L $\cdot$ min^{-1}, $\sim$50 mL of carbon dioxide will be added to each liter of blood, so that venous partial pressure of carbon dioxide is $\sim$46 Torr (6.118 kPa).

During light to moderate exercise, or with a moderate illness, the normal response to increased oxygen utilization and carbon dioxide production is an increase in alveolar minute ventilation and an increase in cardiac output rather than an increase in oxygen extraction at constant

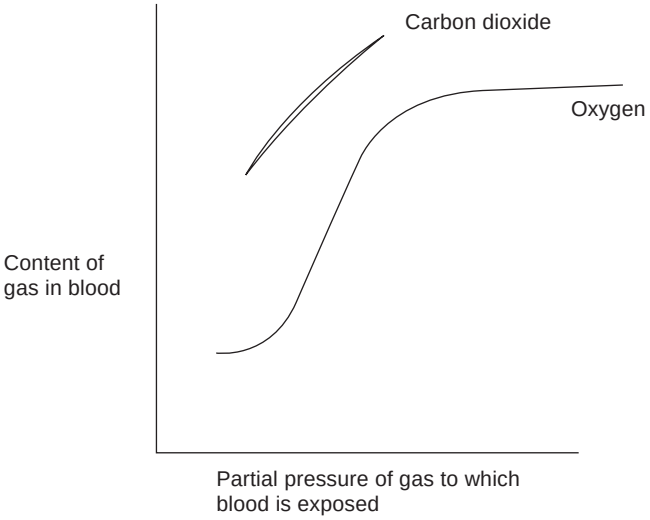


Figure 6. Carrying capacity of blood for carbon dioxide and oxygen.

cardiac output. If the cardiovascular system is impaired, oxygen extraction will increase.

RESPIRATORY MECHANICS IN DISEASE STATES

Several observations of mechanical behavior of the lung show that when diseases occur, the simple resistance-compliance model of the lung may cease to apply. The first concept relating to disease is that of dynamic compliance.

If a simple resistance-capacitor circuit is driven by a sinusoidal voltage, an analysis of voltage and current waveforms will give the same value for resistance and capacitance independent of the frequency of the sinusoid. However, attempts to calculate compliance from a diseased subject while varying the breathing frequency appeared to give a value for compliance that was dependent on the respiratory rate. This became known as "dynamic compliance". The reason for this is that disease processes are seldom homogeneous. The inhomogeneity leads to varying mechanical properties, and a more appropriate model is then a sinusoidal driving force applied to several resistors and capacitors in parallel. In this case,

$$P \sin \omega t = R_1 \, dV_1/dt + 1/C_1 \, V_1 = R_2 dV_2/dt + 1/C_2 \, V_2 = \cdots \tag{19}$$

Adding the equations above would give

$$aP \sin \omega t = (R_1 dV_1/dt + R_2 dV_2/dt + \cdots) + (V_1/C_1 + V_2/C_2 + \cdots)$$

where a is the number of parallel components. However, one measures only the total flow and total volume exhaled. Analyzing data obtained from an inhomogeneous lung using equation 4 would make the resulting R and C functions of all the resistances and compliances in addition to how the total volume and flow were distributed within the system, which itself would be a function of frequency. The term dynamic compliance is a misnomer: The existence of dynamic compliance is an indicator that a simple one-compartment model does not apply (19).

And as one would then expect, the MEFV curves from many subjects with obstructive lung diseases show that not only are the magnitudes of flows reduced, but the portion of the curve that was almost linear in normal subjects now has a definite curvature. This could result from having several compartments of different mechanical properties emptying at different rates.

These concepts fit well with observations of patients' breathing patterns. If increased airway resistance seems to be the issue, the patients breathe more slowly with a greater tidal volume, which might allow enough time for the slow compartments to empty. Equation 1 shows that complete emptying will require a time more than three time constants (RC). If retention of air is a problem because of loss of elasticity, such as in patients with emphysema and high FRC or with structural abnormalities lowering TLC, patients may breathe more rapidly with a smaller tidal volume. In addition, at times of increased airway resistance, some patients purse their lips while exhaling. According to the Equal Pressure Point concept, creating a sharp pressure drop at the lips would raise the pressure throughout the airway system, moving the equal pressure point into larger airways that might not be so collapsible, and possibly allowing an increased expiratory flow rate. These patterns seem to be set by the respiratory controller in the brain in a way that may minimize the work and discomfort of breathing (20,21).

The simple dynamic models applicable to the normal lung have been extended in many ways to attempt to model the effects of disease processes. For example, investigators realized that at higher frequencies, such as those employed with jet ventilators or forced oscillatory ventilation in infants, the inertia of the fluid would play a role. They introduced the electrical analog of inertia, an inductance I, into their models. The models became more complex as investigators attempted to assign various components of the respiratory system their own electrical analog properties, and to allow these properties to vary in different areas of the lungs. However, measurements of lung volumes, compliance, and flows remain the mainstay of clinical pulmonary function testing and are based upon the simple concepts presented above.

EFFECTS OF DISEASE ON GAS TRANSPORT

In acute illnesses or acute exacerbations of chronic lung diseases, formal mechanical testing is usually not performed except in cases such as asthma, where abrupt and dramatic increases in airway resistance markedly reduce the expiratory flows. Clinicians draw conclusions about changes in mechanics from changes in arterial oxygen or carbon dioxide tensions. Understanding these changes is important for both assessing acute decompensations in ambulatory patients and for providing effective mechanical ventilation. These changes are perhaps the most misunderstood aspect of respiratory pathophysiology.

For the normal lung, ventilation and perfusion are matched to assure efficient gas transport. The elimination of carbon dioxide is the usual determinant of alveolar minute ventilation. Diseases affecting the respiratory system can decrease the total amount of air or blood delivered or disturb this matching. Although there are many different disease, the mechanisms are few. First, the membranes responsible for gas transport can be destroyed leaving fewer alveoli to do the job of gas transport, but without significantly affecting the matching. Second, the airways and/or the small blood vessels can become inflamed or blocked, either partially or completely, in a way that interferes with the normal matching of air and blood flows. This results in areas of the lungs with widely varying ventilation/perfusion ratios. Third, problems with the nervous system, with the musculature, or with the structure of the chest wall itself can limit alveolar ventilation. Interpreting the changes in blood gas tension using the simple model for the normal lung may lead to erroneous interpretations.

For example, consider what would happen if 50% of the blood flow went to alveoli, which were fluid filled or

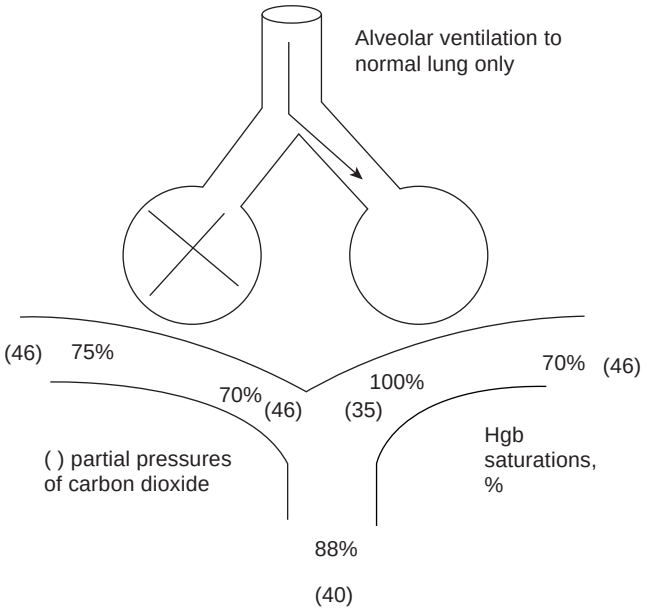


Figure 7. Effect of a 50% shunt on oxygen delivery and partial pressures of carbon dioxide. See text for details.

collapsed (Fig. 7). In the normal person, cardiac output is $\sim 5\ L \cdot min^{-1}$. Venous oxygen tension is $\sim$ 40 Torr (5.320 kPa), and at that pressure the hemoglobin is $\sim$ 70% saturated. Venous carbon dioxide tension is $\sim$ 46 Torr (6.118 kPa). Normally, the lungs would deliver oxygen, remove carbon dioxide, and the arterial blood would have a Pa_{O_2} of 100 Torr (13.330 kPa) and a P_{CO_2} of 40 Torr (5.333 kPa). Because respiration is driven by the carbon dioxide tension, the brain would try to keep arterial P_{CO_2} at 40 Torr (5.320 kPa). But one-half of the blood passing through the lung would maintain its venous content of carbon dioxide, which has a venous partial pressure of 46 Torr (6.118 kPa). This means that the respiratory controller would increase alveolar minute ventilation to the functional alveoli, and alveolar partial pressure of carbon dioxide in those alveoli would fall to something like 32 Torr (4.256 kPa). If one measured the increase in alveolar minute ventilation, it would amount to some 10–15% increase. If one attempted to interpret this according to the one-compartment model, which applied to normal lungs, one might conclude that some of the alveolar ventilation was wasted, that is, that it went to areas of the lung that did not participate in gas exchange (that there was an increase in physiologic dead space). However, that is not what the figure above shows: It shows that blood flow goes to unventilated areas of the lung, not that ventilation goes to unperfused areas of the lung. The true situation is a shunt, whereas the model would predict an increase in dead space.

What would happen to oxygenation? The venous blood would not pick up any oxygen in the unventilated alveoli. Because of the shape of the oxyhemoglobin dissociation curve, very little extra oxygen could be delivered to the blood flowing through the functional alveoli even if the inspired oxygen concentration were increased to 100%.

Note the implications of this. Any condition that causes a mismatch between the amount of blood flow to given alveoli and the amount of air flow will affect oxygenation much more profoundly than it will affect carbon dioxide elimination. There are two major reasons for this: (1) chemosensors are very sensitive to increases in CO_2 tension and will increase alveolar ventilation to compensate; and (2) this compensation is possible for CO_2 and not for O_2 because of the different shapes of the gas content: partial pressure curves for the two gases.

MECHANICAL VENTILATION

The mechanisms by which diseases affect gas transport can be used to understand mechanical ventilation. Because the primary determinant of carbon dioxide tension is alveolar ventilation, adjustments in respiratory rate and tidal volume will allow control of carbon dioxide levels. Because oxygenation is dependent on the fraction of inspired oxygen AND upon how ventilation and perfusion are distributed, the FI_{O_2} is just one of the controls affecting oxygenation. Positive end-expiratory pressure (PEEP) can be applied to try to recruit alveoli that might be collapsed and responsible for the shunting of blood. That is, end-expiratory pressure at the mouth is held above atmospheric pressure to hold alveoli open. The price paid is that normal alveoli can become overdistended. Therefore the exact level of PEEP that should be used is still controversial (22,23).

Although there are many types of ventilators that perform many sophisticated functions, there are two basic ways to ventilate a human being. For adults, volume-cycled ventilators are usually used. A tidal volume is set, and pressures measured for safety reasons. For infants, tidal volumes are too small to accurately measure, so that pressure-cycled ventilators are used. In this case, a peak inspiratory pressure and PEEP are chosen, and the volume actually delivered depends on the mechanics of the ventilator and of the infant's respiratory system (24). In this case, tidal volume becomes a variable so that safe pressure limits are never exceeded.

Although there is much written about "barotrauma" from mechanical ventilation, that is actually a misnomer. Experiments on cells from the lungs show that when a cell is stretched beyond a certain point, inflammatory mediators are generated and the cell can die (25). The correct term is "volutrauma", but because we cannot measure cell stretch adequately, pressure limits are generally respected (26).

CONCLUSION

This article does not attempt to summarize the literature on continuum mechanics, statistical models, or tissue mechanics of the respiratory system. However, the simple concepts described above are essential for understanding pulmonary function testing, respiration, and patient support with mechanical ventilators. Once these concepts are mastered, the complex and often conflicting literature can be critically read and understood.

BIBLIOGRAPHY

1. Kollef MH, Shapiro SD, Boyd V, Silver P, Von Harz B, Trovillion E, Prentice D. A randomized clinical trial comparing an extended-use hygroscopic condenser humidifier with heater-water humidification in mechanically ventilated patients. Chest 1998;113:759.
2. Huang W, Yen RT, McLaurine M, Bledsoe G. Morphometry of the human pulmonary vasculature. J Appl Physiol 1996;81: 2123.
3. Wang NS. Anatomy and physiology of the pleural space. Clin Chest Med 1985;6:3.
4. Lichtenstein O, Ben-Haim SA, Saidel GM, Dinnar U. Role of the diaphragm in chest wall mechanics. J Appl Physiol 1992;72:568.
5. Scarpelli EM. Physiology of the alveolar surface network. Compar Biochem Physiol Part A, Mol Int Physiol 2003;135: 39.
6. Escolar JD, Escolar A. Lung hysteresis: A morphological view. Histol Histopathol 2004;19:159.
7. Ainsworth SB, Milligen DWA. Surfactant therapy for respiratory distress syndrome in premature neonates: A comparative review. Am J Resp Med 2002;1:417.
8. Lai-Fook SJ. Pleural mechanics and fluid exchange. Physiol Rev 2004;84:385.
9. Milic-Emili J, Mead J, Turner JM, Glauser EM. Improved technique for estimating pleural pressure from esophageal balloons. J Appl Physiol 1964;19:207.
10. Pride NB. Tests of forced expiration and inspiration. Clinics Chest Med 2001;22:599.
11. Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general U.S. population. Am J Respir Crit Care Med 1999;158:179.
12. Hyatt RE, Wilson TA, Bar-Yishay E. Prediction of maximal expiratory flow in excised human lungs. J Appl Physiol 1980;48:991.
13. Elliott EA, Dawson SV. Test of wave-speed theory of flow limitation in elastic tubes. J Appl Physiol 1977;43:516.
14. Kamm D. Airway wall mechanics. Ann Rev Biomed Eng 1999;1:47.
15. Polak AG, Lutchen KR. Computational model for forced expiration from asymmetric normal lungs. Ann Biomed Eng 2003;31:891.
16. Zhang Q, Suki B, Lutchen KR. Harmonic distortion from nonlinear systems with broadband inputs: Applications to lung mechanics. Ann Biomed Eng 1995;23:672.
17. Fung YC, Sobin SS. Theory of sheet flow in lung alveoli. J Appl Physiol 1969;26:472.
18. Brimioulle S, Lejeune P, Naeije R. Effects of hypoxic pulmonary vasoconstriction on pulmonary gas exchange. J Appl Physiol 1996;81:1535.
19. Otis AB, McKeerow CB, Bartlett RA, Mead J, McIlroy MB, Silverstone NJ, Radford EP Jr. Mechanical factors in the distribution of pulmonary ventilation. J Appl Physiol 1956;8:427.
20. Oku Y, Saidel GM, Altose MD, Cherniack NS. Perceptual contributions to optimization of breathing. Ann Biomed Eng 1993;21:509.
21. Guz A. Brain, breathing, and breathlessness. Respir Physiol 1997;109:197.
22. Schmitt J-M, Vieillard-Baron A, Augarde R, Prin S, Page B, Jardin F. Positive end-expiratory pressure titration in acute respiratory distress syndrome patients: Impact on right ventricular outflow impedance evaluated by pulmonary artery Doppler flow velocity measurements. Crit Care Med 2001;29: 1154.
23. Rouby J-J, Puybasset L, Nieszkowska A, Lee Q. Acute respiratory distress syndrome: Lessons from computed tomography of the whole lung. Crit Care Med 2003;31(Suppl): S285.
24. Ligas JR, Moslehi F, Epstein MAF. Occult positive end-expiratory pressure with different types of mechanical ventilators. J Crit Care 1990;5:95.
25. Pugin J. Molecular mechanisms of lung cell activation induced by cyclic stretch. Crit Care Med 2003;31(Suppl): S200.
26. Acute Respiratory Distress Syndrome Network (ARDS). Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. NEJM 2000;342:1301.

See also CONTINUOUS POSITIVE AIRWAY PRESSURE; HIGH-FREQUENCY VENTILATION; PNEUMOTACHOMETERS; PULMONARY PHYSIOLOGY.

RESPIRATORY MONITORING. See VENTILATORY MONITORING.

RESPIRATORY SOUNDS. See LUNG SOUNDS.

RESUSCITATION, CARDIOPULMONARY. See CARDIOPULMONARY RESUSCITATION.

RHEOLOGY, BLOOD. See BLOOD RHEOLOGY.

S

SAFETY PROGRAM, HOSPITAL

JOSEPH F. DYRO
Setauket, New York

INTRODUCTION

This article describes the broad scope of safety issues in the hospital and recommends ways in which hospitals should address these issues. Particular emphasis is placed on medical devices and instrumentation and how they interrelate with the hospital's environment, facility, patients, and device users and other personnel. Medical device safety depends in large part upon a comprehensive medical device technology management program, which includes elements ranging from technology assessment, evaluation, and procurement to device replacement planning and includes such components as user training, medical device incident investigation, and device quality assurance. The clinical engineer by education, training, and experience is shown to be ideally suited to implement and execute a technology management program and to lead in the effort to assure hospital-wide patient safety. Available resources are described including web-based training, distance learning, standards, publications, and professional organizations.

Innovative surgical techniques, improved invasive and noninvasive diagnostic procedures, advanced diagnostic and therapeutic medical devices, pharmacological progress, and an overall better understanding of disease processes continue to benefit the treatment of the sick; however, advanced technology has been a mixed blessing to recipients of these advances. Its positive impact on healthcare has been countered by the creation of problems engendered by an increasingly complex medical device and instrumentation environment. In the use of devices and techniques, misuse or procedural error can and does occur. Some level of risk is associated with everything that is done to care for the patient. Healthcare professionals must keep abreast of the advancing technologies, be aware of their limitations and risks, and manage them. The hospital safety program is an integral part of the system of risk reduction and can and should be broad in scope, not limited by undue attention to only one or two risk categories such as slips and falls or needle sticks. It should encompass means to address and control the risks of injury to patient and staff, risks that arise from all sources, not only those specifically related to medical devices, but those related to the performance of the individual healthcare giver.

THE PROBLEM WITH HEALTHCARE

The complexity and proliferation of the three components of technology (i.e., drugs, devices, and procedures) continues unabated. In the hospital complex, interrelationships exist among environment; facilities and utilities; medical and nonmedical policies; and procedures, economics, ethics, and human performance. Inappropriate drugs, devices, and procedures; defectively designed, manufactured, and maintained medical devices; inappropriate staffing; inability to follow procedures, neglect of duties, inattention, and ignorance; and deficient education and training are among the factors that increase the risk of injury in the hospital.

The Institute of Medicine (IOM) of the National Academy of Sciences produced a report containing estimates as high as 98,000 deaths occurring annually in U.S. hospitals caused by medical errors (1). About 1 million medical injuries occur annually in the United States, with 200,000 involving some negligence (2). Such iatrogenic injury (injury caused by the physician, now more broadly defined as injury caused by giving care in a hospital) has been well documented (3). Medical devices used in delivering care have a finite probability of doing harm to the patient or to the person using the device (4) and are included as one source of iatrogenic injury. Estimates of the percentage of device-related incidents over all incidents vary from 1 to 11% (2).

Patient Safety Movement

The IOM report spawned a patient safety movement. Organizations, actively working to identify the weaknesses in the healthcare arena, are unified in their support of patient safety initiatives (5–7). The Institute of Medicine (9), in its sequel report, *Crossing the Quality Chasm: A New Health System for the 21st Century*, calls for a health care delivery system that is safe, effective, patient-center, timely, efficient, and equitable. Patient safety can be improved by implementing and executing effective hospital safety programs.

Safety Programs

Programs exist that address the overall safety concerns within a hospital. Under the umbrella of a *total hospital safety program*, there are various components, such as fire safety, infection control, medical device safety, and radiation safety. Guidelines for a comprehensive total hospital safety program incorporating all the various components are given below. Following this is presented a program that addresses medical device safety, in particular. Before elaborating upon these two safety programs, to obtain a clearer sense of the problem with health care the injurious forces and mechanisms present in a hospital are described and their interrelation with patients, staff, facilities, environment, medical technology are described.

Injurious Forces and Mechanisms

The hospital is a place where injury mechanisms abound. The nature of a hospital's business concentrates the injury mechanisms. For example, life-threatening situations demand rapid response by medical personnel often

utilizing a host of complex technologies. All components of the system must work for successful outcome. One element failure, for example, an error in device use, administration of the wrong drug, or inappropriate surgical technique, can produce disastrous results even though all other elements are working. Injury to patients, personnel, and visitors can arise from many different sources within a hospital. The following is a list of some of the many safety concerns and examples of hospital occurrences:

- Fire: Electrosurgical energy in an oxygen-enriched atmosphere has ignited a fuel source such as the hair of a patient resulting in a rapidly spreading, deadly conflagration.
- Air, medical gases, and vacuum: Crossed-medical gas pipelines have killed patients mistakenly given nitrous oxide instead of oxygen.
- Water: Water with inappropriate chemical content used in dialysis has injured patients.
- Chemicals, solvents, sterilizing agents, skin preparation solutions, anesthetic gases: Anesthetic gases can cause birth defects; ethylene oxide used in sterilization has caused severe skin lesions.
- Drugs: Administration of wrong or inappropriate amounts of medications have killed patients.
- Filth, microorganisms, vermin: Inadequately sterilized bronchoscopes have cross-contaminated hundreds of patients.
- Waste, bodily fluids, sharps (needles, scalpels): Improperly discarded needles have punctured hospital workers infecting them with human immunodeficiency virus (HIV).
- Sound, noise: Ambient noise levels have obscured audible alarms.
- Ionizing and nonionizing radiation [X-rays, laser, ultraviolet (UV), visible light, infrared (IR), microwave]: Interventional cardiologists have burned patients by using excessive X-ray exposure times during fluoroscopically guided catheter placement.
- Electricity: Lethal electric shock was delivered to a patient by improper connection of ECG leads to a power cord.
- Natural and unnatural disasters: Hurricanes have disrupted electrical power and back-up generators have failed to function.
- Mechanical stress: Static and dynamic forces on the body can result in injury. For example, excessive and prolonged mechanical pressure to parts of the body during surgery has resulted in pressure necrosis of tissue. Among the highest causes of injury are slips and falls and back strains.
- People: Human error, for example, administering the wrong medication or cutting-off the wrong leg, is the largest single cause of injury in the hospital. Abduction and elopement are also people-related safety concerns.
- Devices: A defectively designed check valve has failed preventing ventilation of a patient resulting in brain death.

From the above, it is clear that a hospital patient is exposed to a variety of risks many of which involve, in some way, medical devices and instrumentation. Human error resulting in patient injury often involves the use of a medical device in an inappropriate fashion. Such use may signify that the device was designed with inadequate attention paid to human factors considerations (9). A hospital safety program must pay due attention to the relationships among devices, people, places and things, and take a systems perspective on safety.

Systems Perspective on Safety

The hospital is a complex environment where medical devices and instrumentation utilized for diagnostic, therapeutic, and rehabilitative purposes interact with each other and with patients, staff, facilities, and the environment. A systems approach to understanding the mechanisms contributing to incidents and accidents aids in the development of programs to control and minimize safety risks (10). The efficacy and safety of the delivery of patient care depends, in general, on five main components: medical device; operator; facility; environment; and patient. The following is a description of these components along with examples of the contributions they can make toward creating hazardous situations.

Medical Devices

The device itself can be the sole cause of an iatrogenic injury. A medical device can be rendered injurious by the actions or inactions of the device inventor, designer, manufacturer, shipper, inspector, maintainer, or user. For example, the manufacturer may fail to properly assemble or construct an otherwise efficaciously designed medical device. The shipper who transports the device from the manufacturing plant to the hospital may damage a properly manufactured device. The inspector at the hospital may fail to properly inspect the device upon receipt, allowing a damaged or defective device to be placed into service. The maintainer may fail to keep the device operating within manufacturer's specifications by not adequately performing preventive maintenance and calibration procedures. The manufacturer may fail to provide adequate warnings and instructions. The user may abuse or misuse the device, rendering it hazardous to the patient. Finally, the device may not be retired when it has reached the end of its useful life and is subject to more failure modes and is not up to date with current medical practice.

Injury can result from the absence or failure to utilize medical devices that are accepted as the standard of care in the community. Physicians and nurses have an obligation to utilize technology that has been accepted as the standard of care, and liability is incurred if available technology is not used.

Definition

For an all-encompassing safety program, the definition of medical device adopted by the U.S. Food and Drug Administration (FDA) is recommended (11):

"The term device means an instrument, apparatus, implement, machine, contrivance, implant, in vitro

reagent, or other similar or related article, including any component part, or accessory, which is

1. Recognized in the official National Formulary, or the United States Pharmacopoeia, or as supplements to them.
2. Intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease, in humans or other animals.
3. Intended to affect the structure or any function of the body of humans or other animals, and that does not achieve any of its principal intended purposes through chemical action within or on the body of humans or other animals and that is not dependent on being metabolized for the achievement of any of its principal intended purposes."

All devices used in the hospital or provided by the hospital must be addressed whether they are purchased, donated, leased, or owned by a physician or are under evaluation. Technically, a safety program must include devices ranging from tongue depressors to magnetic resonance imaging (MRI) units. Both reusable and disposable devices must be considered. As medical device technology becomes more complex and new modalities are employed for diagnostic or therapeutic intervention, the hospital safety program must adapt to these changes. It was not that long ago that such issues as laser safety, magnetic field strength hazards, and nonionizing radiation hazards were unknown concepts. A hospital safety program must also include a way of dealing with devices used outside the hospital, such as devices provided by or through the hospital to enable ambulatory and home care or devices used in the transport of patients to or between hospitals. A hospital department such as clinical engineering that extends its services to other entities such as a doctor's office or clinic must adhere to the same principles that guide its in-house safety program such as regular preventive maintenance schedules.

Electricity

Medical devices that are electrically operated can pose risk to patients and staff of electric shock. Incoming inspection and periodic inspections thereafter can assess the safety of such devices. As with all electrically operated devices, care should be taken to ensure that surfaces that may be contacted by patients and operators are not energized because, for example, of poor design, manufacturing defect, or deterioration and inadequate maintenance and repair. Electrical safety of medical devices is but one aspect of safety as is shown below. Safety programs should not be obsessed with electrical hazards, shown to be a relatively minor contributor to death and injury in hospitals since first introduced several decades ago as a major problem (12).

Chemicals

Many potentially injurious chemical compounds are or have been associated with the operation of medical devices and instrumentation. For example, ethylene oxide gas used in hospital sterilizers poses a risk to operators and to patients if improperly utilized (13). Iatrogenic complications can be introduced by infusion pumps, widely used for intravenous and intraarterial fluid and drug administration. Incompatibility of certain drugs with the plastics from which infusion pump administration sets are made is a source of risk. Nitroglycerin in solutions administered by infusion pumps has been shown to interact with certain formulations of poly(vinyl chloride) (PVC) tubing, ultimately decreasing the potency of the drug and making the administration of a known amount impossible to gauge. Other examples include allergic reactions to latex gloves, skin preparation agents, and contrast media and adverse reactions to implantable devices such as artificial joints and nerve stimulators.

Sound and Electromagnetic Radiation

A neonate's auditory system is particularly susceptible to injury from high sound levels as can occur in infant incubators. Diagnostic ultrasound units must be properly designed and maintained to ensure that output levels remain within acceptable limits recommended in safety standards.

The hazards of ionizing radiation became well known not long after the development and use of the X-ray machine. A good radiation safety program is essential to ensure that radiographic equipment and protective measures in radiographic suites meet acceptable performance and safety standards and that operators are appropriately trained and follow established safety procedures.

Nonionizing radiation is a significant health hazard in all hospitals and includes UV, microwave, and laser radiation. Ultraviolet therapy is employed in the treatment of some skin disorders. However, UV radiation can adversely affect the eye (keratitis) and the skin (erythema and carcinogenesis). Microwave radiation is commonly used in physical therapy for diathermy treatment of patients. Microwave effects are largely thermal, with the eye being the most susceptible organ (cataractogenesis). Surgical lasers are particularly hazardous since the beam can travel large distances with little attenuation and can reflect off surfaces in the room. The intensity of the beam is of necessity sufficient to burn body tissue as is required in surgical procedures. Momentary eye contact with the beam can cause severe eye damage. A laser safety program is recommended wherever lasers are used. Eye protection and restriction of the area to trained personnel are essential steps in the program.

Alarms

The Joint Commission on Accreditation of Healthcare Organizations identified appropriate attention to alarms as one of six patient safety goals of 2003 (14). Alarms are sometimes ignored, unheard, unnoticed, misinterpreted, or disabled. A life-threatening alarm that goes unattended can have disastrous consequences for the patient. Alarms are rendered ineffective for a host of reasons, some of which are listed here. Frequent false alarms cause the care givers to relax their guard to the point of ignoring a real alarm condition when it occurs. Doors to patient rooms, closed to give the patient some respite from the noise in the corridor

and in other rooms, can attenuate a bedside monitor alarm such that it is not heard by nurses on the floor. Medical staff members, either through inattention, lack of knowledge, or intent, can disable the alarm function on a device. Mechanical or electrical failures could occur that render the alarm circuitry ineffective. Artifact present can be interpreted by a monitor as a real physiological signal and can cause a monitor to fail to alarm when the patient's physiological signal exceeds limits.

Mechanics

Device design should account for motion, maneuverability, and impact resistance. A transport stretcher that is difficult to maneuver can strain the muscles of the transporter. Devices should be made of materials with strength adequate for their intended purpose. Device mountings and supports must have adequate load-bearing characteristics. Inadequately secured physiological monitors have fallen onto patients from overhead shelves. Latches, locks, protective covers, and restraints must be designed with appropriate attention to human factors and conditions of use to prevent inadvertent opening and loss of protection from falls or tampering. Patients of all ages have died from asphyxiation after becoming entrapped by bed rails. Connectors and couplings abound in the hospital environment. Blood line disconnections have resulted in deadly air embolisms; airway disconnections have resulted in death and brain damage from oxygen deprivation; and reversed connections to anesthesia vaporizers have resulted in deadly anesthetic overdoses.

Operator

Lack of knowledge concerning the operation of a device is the most common failing of the operator. A safety program must recognize that the rapid introduction and wide proliferation of complex medical device technology taxes the operator's ability to remain competent and must assure that adequate educational programs are in place to address this need. Educational programs and credentialing requirements minimize user error caused by ignorance of device operation.

Failure to correctly diagnose is a claim made increasingly more frequently in malpractice actions. Risks associated with diagnostic medical devices such as ECG telemetry systems or clinical laboratory analyzers often relate to device set-up, calibration, and operation and data gathering, manipulation, display, storage, and retrieval. The harm that people (e.g., doctors, nurses, technologists), can do to the patient can relate directly to the presence, quality, and limitations of diagnostic data obtained by the device (15).

Errors in the operator's use of devices are more likely when manufacturers do not provide instructions or provide ambiguous or misleading instructions, labels, indicators, or controls on the device. Manufacturers have a duty to communicate word to their customers of upgrades and product improvements that address design deficiencies in the products that they have sold.

The manufacturer may contribute substantially to the operator's failure by not understanding the limitations of the operator and designing a device without adequately addressing human factors engineering. The FDA recognizing the important role sound human factors engineering plays in safe medical device design has spearheaded efforts to develop standards and guidelines in this area (16). The degree to which human factor problems contribute to safety risks and incidents is indeterminate largely because those reporting device problems often lack the understanding of how faulty medical devices and instrumentation design contributes to user errors (17).

Drug and medication errors constitute one of the major sources of hospital incident reports. Safety problems include adverse drug reactions and inappropriate dosage, drug, frequency of administration, and route of administration.

Economic pressures influencing staffing patterns and work requirements result in understaffed and overworked operators who do not adhere to policies and procedures, but rather take more expedient, albeit more risky, avenues. For example, an ECG telemetry central station, which by the manufacturer's instructions and hospital policy should be monitored constantly, may go unattended because of competing demands for personnel resources. A study of anesthesia claims over an 8-year period revealed that 14% of the claims alleged failure to monitor (18). Personnel who do not adhere to hospital or departmental policies and procedures pose safety problems for the patient. Policies and procedures also address such issues as patient care procedures or surgical procedures.

Facility

The hospital's physical plant and its facilities can have a substantial effect upon safety. Some of the major facilities of a typical hospital are listed below followed by some examples of how these facilities can affect patient and personnel safety (19):

- Heating, ventilation, and air conditioning (HVAC)
- Medical gases
- Water
- Electric power
- Sanitation systems
- Transport and space

Heating, Ventilation, and Air Conditioning

Air pollution can adversely affect the compressed air supply that is needed to provide respiratory support and to power pneumatic devices. Inadequate ventilation has resulted in hazardous concentrations of ethylene oxide gas used in the sterilization of heat-labile devices. Failure to control relative humidity has resulted in water condensing on wrappings of sterile surgical instruments and thus breaching the sterility barrier.

Medical Gases

Switched oxygen and nitrous oxide supply pipelines have caused patient deaths. Poorly designed and maintained medical gas and vacuum distribution systems can fail to

provide adequate pressures and flows for the proper operation of such devices as suction machines, pneumatically powered surgical instruments, ventilators, and anesthesia machines.

Water

A hospital must have an adequate supply of water (20). Because it is in direct contact (through a membrane) with a patient's blood, the water used in hemodialysis must be monitored and its components controlled. Numerous adverse reactions from untreated water have occurred (21). Components that can affect the dialysis treatment have been identified and include insoluble particles, soluble organics and inorganics, heavy metals, bacteria, and pyrogens.

Electrical Power

The performance of sensitive electronic medical devices can be adversely affected by irregularities in electrical power distribution systems. Line voltage variations, line transients, and interruption of power can all result in harm to the patient by adversely affecting the performance of such devices as physiological monitors, ventilators, electrosurgical units, and clinical laboratory analyzers.

Sanitation Systems

The safe disposal of solid, liquid, and gas waste constitutes a significant environmental concern to the hospital, its patients and personnel, and the community in which the hospital is situated (22). It is common for clinical personnel to be stuck by needles in the course of either injecting a patient or disposing of a used needle. Housekeeping personnel are victims of improperly discarded needles, scalpels, broken glass, and biohazardous and radioactive waste. Chronic low level exposure to inhalation anesthetics is associated with spontaneous miscarriage, liver disease, cancer, and other physiological disorders.

Hepatitis C, HIV, SARS, influenza, and other infection agents pose threats to patients and personnel who come in contact with infected patients or their bodily fluids and waste. To add to the problem, infectious agents have developed into virulent forms that are resistant to standard antibiotics.

Transport and Space

Elevators, escalators, doors, stairwells, and staff and patient care areas affect the level of hospital safety. A patient room that is too small results in crowding of medical devices, personnel, and patient. Crowding restricts rapid access to necessary devices and instruments and interferes with access to the patient. Poorly maintained and designed elevators delay a caregiver's access to a patient. Fire doors propped open or emergency exit doors propped shut pose serious hazards in the event of fire. Fire is particularly hazardous in an environment in which an individual's ability to evacuate is compromised by illness or disability. A closed door to a patient room may afford the patient some quiet from the noise from the outside, but it can also attenuate a bedside alarm to make it inaudible to the staff.

Environment

Those elements within a hospital environment that can influence the safe operation of a medical device include such things as medical and nonmedical devices that are not directly involved in a patient's care, people, electromagnetic interference, cleanliness, psychological factors, biomechanics, and natural and unnatural disasters. The environment in which medical devices are used has grown beyond the hospital to now include emergency vehicles, ambulatory care facilities, long-term care facilities, and the home. The hospital, to the extent that it furnishes devices for patient use, must assume the responsibility for ensuring that these devices are safe and effective.

Device–Device Interaction

The hospital has become a complex electrical environment where incompatibility can exist between medical devices (24). For example, electrosurgical units have obscured electrocardiographic traces on cardiac monitors, defibrillator discharges have damaged internal pacemakers, and electrical transients have modified patient ventilator rates.

Nonmedical Devices

An ever-increasing variety and number of nonmedical devices are often requested to be brought into hospitals by patients. Some of these devices include cell phones, space heaters, electric heating pads, televisions, electric shavers, electric guitars, and video games. A hospital safety program must specify what is not allowed to be brought into the hospital or into specific areas of the hospital, must detail the means of enforcement, and give guidance on making exceptions in unusual circumstances. Policies and procedures will vary with the institution, medical condition of the patient, and inspection capabilities of the hospital's clinical engineering resources.

Hospital staff often desire to bring in personal-use devices such as fans, electric space heaters, radios, and hot plates. Office equipment such as personal computers and radios also may be a source of risk and should be considered in a comprehensive safety program. Equipment used by housekeeping and plant engineers such as floor buffers, vacuum cleaners, and power tools are a source of risk.

People

Besides the operator described above (i.e., the patient's nurses, doctors, and technicians), other people in the patient's environment can poses safety risks. These include other medical staff not directly involved in the patient's care; other patients; nonmedical hospital personnel (e.g., housekeeping, maintenance, administrators, and security personnel); and visitors, salespeople, and interlopers.

Visitors have altered controls on life support equipment to the detriment of the patient. Housekeepers have disconnected life-support ventilators from wall outlets in order to plug in floor sweepers. Premeditated or nonpremeditated attack either on a patient or by a patient is a possibility. Abduction of infants or elopement of mentally deranged patients occurs.

Poor communication has an adverse effect upon a patient's treatment. For example, a care giver's order for a test or request for a report that goes unheeded can be detrimental. Not communicating what has already been done for the patient may at best result in unnecessary and costly repeat procedures or at worse the repeat administration of medications and resulting overdose.

Electromagnetic Interference

Electromagnetic interference (EMI) is the disruption of the performance of one device by electrical energy emitted by another device or by another source of radiation such as a radio or television transmitter. EMI is becoming more of a problem as the number of electronic devices in the hospital increases (24,25). The utilization of an ever increasing amount of the electromagnetic spectrum and the proliferation of devices that are intended and unintended emitters and receivers of electromagnetic energy require that hospitals pay due attention to compatibility and interference (26).

The high strength magnetic fields about MRI units have drawn in from the surrounding environment ferrous metallic objects resulting in death and injury when these objects collide with great force against patients and personnel.

Cleanliness

Infection is a major complication of invasive arterial, central venous, and pulmonary artery monitoring. Nosocomial (hospital acquired) infection can be spread by way of medical devices that have not been properly handled, cleaned, disinfected, or sterilized (27). Vermin, animals, fomites (bedding, clothes), food, and people are other vehicles for the spread of nosocomial infection. Dirt and dust can contaminate wound sites, can damage delicate electronic instrumentation and computer data storage devices, and can harbor pathogenic organisms.

Psychological Factors

Periontogenic (having its genesis in the surroundings) illness afflicts patients and staff. Such illness has its roots in the stresses engendered by the high technology surroundings and includes such syndromes as intensive care unit (ICU) psychosis, the adverse psychological impact of the ICU setting. Dehumanization of patients can lead to such neuroses and psychoses as depression, denial, and dependency.

Biomechanics

Thousands of patients each year are injured in slips and falls while attempting to get out of bed to make use of bathroom facilities. Back injuries to staff from lifting heavy equipment or positioning patients are not uncommon occupational injuries.

Natural and Unnatural Disasters

Major safety problems are associated with a wide range of natural disasters such as tornado, hurricane, and earthquake and unnatural disasters such as terrorism, and nuclear power plant meltdown. Disaster planning will minimize loss in these circumstances (28).

Patient

Failure of patients to exercise their duties and responsibilities can lead to their own injuries. Patients can contribute to their own injury by failing to use reasonable care in attending to themselves. Such examples include deliberate removal or tampering with support equipment, failure to follow instructions to remain in bed, tripping or falling over obvious obstacles in hospitals, and introduction of unauthorized items into the environment of the hospital.

Systems Analysis

The interaction of the above five components (patient, device, facility, operator, and environment) can be demonstrated by considering the example shown in Fig. 1. A neonate (the patient) is in an infant radiant warmer (the device) electrically powered from the wall receptacle (the facility). The nurse (the operator) prepares to connect the skin temperature sensor to the warmer's control panel. A nearby examination lamp and external influences of a bright sun and electromagnetic waves emitted by a local station are shown (the environment). All five elements of the system can interact and give rise to a situation resulting in patient injury. For example, electromagnetic radiation can interfere with the operation of the radiant warmer's heater control. The nearly examination lamp can add to the heat transfer to the neonate leading to hyperthermia. The skin temperature sensor could be incompatible with the temperature monitoring system, but its connector could permit it to be connected to the system only to give false reading and result in excessive radiant heat output. Interruption of electric power because of a damaged electrical receptacle could result in loss of heater output and subsequent hypothermia.

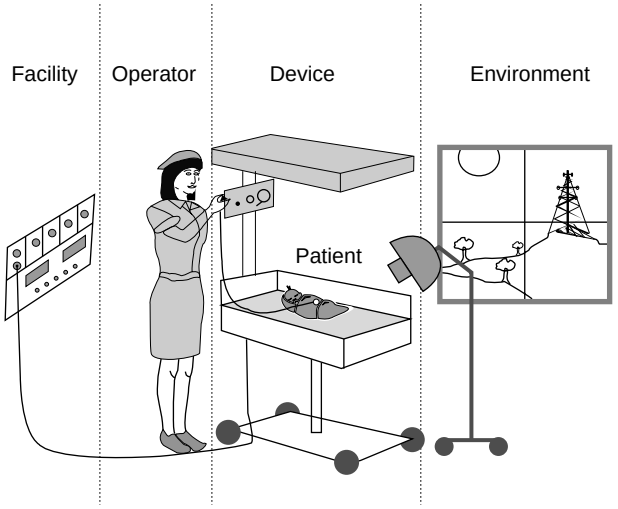


Figure 1. Systems approach to medical device safety.

TOTAL HOSPITAL SAFETY PROGRAM

As seen from the preceding section, the issue of total hospital safety encompasses human elements, machine elements, and environmental elements, all of which must be considered in a total hospital safety program (30).

JCAHO Recommendations

The *Comprehensive Accreditation Manual for Hospitals* (CAMH) (30) published by the Joint Commission on Accreditation of Healthcare Organizations (JCAHO) includes a chapter entitled, "*Management of the Environment of Care*". The standards contained in this article require hospitals to manage seven areas by the development and implantation of appropriate plans:

- Safety
- Security
- Hazardous materials and waste
- Emergency
- Life safety
- Medical equipment
- Utility systems

Safety Committee

The safety management plan requires a hospital to identify "*processes for examining safety issues by appropriate hospital representatives*", that is, the safety committee, now often called the environment of care committee. The JCAHO refers to the safety committee as "*a multidisciplinary improvement team*" and defines its activities and responsibilities as follows:

- Safety issues are analyzed in a timely manner.
- Recommendations are developed and approved.
- Safety issues are communicated to the leaders of the hospital, individuals responsible for performance-improvement activities, and when appropriate, to relevant components of the hospital wide patient safety program.
- Recommendations for one or more performance improvement activities are communicated at least annually to the hospital's leaders.

The safety committee serves as a forum for reports from the seven environment of care areas. For example, the safety committee will receive reports as specified in the hospital's medical equipment management plan, including performance monitoring data and incident investigation reports. Serving as a reporting forum is an essential part of the safety committee's role. However, the great value of the safety committee is found in its ability to subject these reports to multidisciplinary analysis. Careful analysis from a representative range of perspectives will produce communications, recommendations, and performance improvement activities that enhance safety throughout the hospital. Hospitals may assign managers to oversee the implementation of the management plans in the above environment of care areas.

Cooperation and Communication

The hospital safety program requires the conscientious participation and work of a wide range of committees, departments, and individuals with full cooperation and support of the hospital's administration. It involves a set of policies and procedures that cut through departmental barriers. Interdepartmental cooperation, involving virtually every clinical and support service, is imperative to adequately address the wide range of issues that bear upon the safety of patients and personnel. Hospital safety programs are inherently difficult to implement as a consequence of the diversity of their component parts.

A total hospital safety program should include patient and employee, environmental, and medical device safety. Hospital-operated and physician-directed safety programs should be linked in a single system. The hospital and medical staff should pool their resources and direct all patient safety information to a central point. At that point, data analysis can identify problem areas before compensable events occur. The components of a total hospital safety program that are actively involved in hospital safety issues are committees, departments, and administration.

Committees

The committee is a means by which formal lines of communication can be exercised. A hospital has various committees that relate in some fashion to an overall hospital safety committee (environment of care committee). Each committee necessarily deals for the most part with rather narrow issues. Those committees that have the strongest influence on overall hospital safety are listed in Table 1. No two hospitals are alike and the names of the committees may vary from one institution to the next.

Participation of key individuals, representing the entire spectrum of services provided within the hospital, in the work of the safety committee as well as in the work of the other committees listed is an integral component of the hospital safety program. Formal communication channels must be created between these committees and the safety committee. Committee attendance is an aid to interdepartmental communication. Transfer of information in an informal setting in a face-to-face meeting before or after a committee meeting can often be more effective, rapid, and accurate than written or telephone communications.

Table 1. Committees Concerned with Hospital Safety

Radiation safety	Specialty medical services	Infection control
Quality assurance	Emergency resuscitation	Utilization review
Standards and evaluation	Medical board	Capital equipment
Continuing medical education	Operating room	Credentials and privileges

Table 2. Departments, Offices, and Authorities Concerned with Hospital Safety

Risk management	Staff development	Materials management
Plant and facilities engineering	Environmental health and safety	Fire marshall
Medical departments	Quality assurance	Information services
Radiation safety	Infection control	Clinical engineering

Departments

The departments, offices, and authorities that play important roles in the hospital safety program are shown in Table 2.

Risk management is a key element in any hospital safety program (31). It has become not only a moral, but a financial necessity to provide a safe environment for the staff and patients. It is the risk manager who must make recommendations, in the face of cost containment efforts, on the investment of capital in monitoring systems and modern equipment as a means of reducing risk. Risk management can determine a course of action aimed at reducing risk through, for example, increased surveillance of the patient by appropriate monitoring devices such as electrocardiographs, capnometers, blood pressure monitors, and pulse oximeters. Should an analysis show that the main problems lie in human error, clinical engineering and staff development might implement an educational program as a response. For example, human error, rather than equipment failure, appears to be the primary factor in most preventable anesthesia mishaps. Increased risk management initiatives can reduce these adverse incidents. Cost-effective anesthetic practices actually require using the most up-to-date and well-maintained equipment. The risk management component is central to controlling the costs of safety. Care must be taken to deploy risk management resources judiciously. A disproportionately large amount of attention has been focused on some relatively minor safety issues, such as electrical safety, while by comparison few resources have been commanded to address issues that pose greater risks to patient safety, such as slips and falls (32).

Materials management is generally responsible for the acquisition and distribution to patient floors of a wide variety of device accessories and disposable products. Prepurchase evaluation of these items should not be overlooked. Clinical engineering is often in a position to lend its expertise in device testing to the evaluation of new products proposed for addition to the hospital's standing inventory.

Environmental health and safety is concerned with such issues as fire safety; air quality; spillage, handling, and storage of chemicals and unidentified substances; and noise abatement.

Radiation safety is concerned with the protection of patients and personnel from the effects of ionizing radiation (33). The primary aim of radiological protection is to provide an appropriate standard of protection for humans without unduly limiting the beneficial practices giving rise to radiation exposure according to the International Commission on Radiological Protection (ICRP) (34). The ICRP system of protection is based on the following three objectives:

1. To prohibit radiation exposure to individuals unless a benefit to the exposed individuals and to the society can be demonstrated, and be sufficient to balance any detriment.
2. To provide adequate protection to maximize the net benefit, taking into consideration economic and social factors.
3. To limit the dose (other than from medical exposures) received by individuals as a result of all uses of radiation.

Infection control identifies the means of growth, reproduction, and transmission of pathogenic organisms. Infection control advises on methods for eliminating or controlling the reservoirs of microorganisms, on appropriate sterilization, disinfection, and sanitation techniques, and on disturbing or eliminating the means of transmission. Where this means of transmission involves the medical device, close cooperation with clinical engineering is advisable.

Staff development has a substantial contribution to make in training staff and in periodically reviewing competence levels in device use and patient treatment techniques. Staff development typically develops suitable curricula with the assistance of clinical engineering.

Plant and facilities engineering maintains the facility, ensures satisfactory heating, ventilation and air conditioning, electric power, water supply, medical gases and vacuum, and sanitation. Facilities engineering keeps abreast of the building codes pertaining to construction and renovation (35).

The *medical departments* constantly review the quality of the care administered and remain alert to indicators of inappropriate diagnoses or therapeutic intervention. The medical staff's quality assurance programs are coordinated with each other and integrated with hospital programs under the office of the safety director.

Information technology departments are responsible for the computerized acquisition, storage, and retrieval of patient information of both an administrative and clinical nature. Timely, accurate, and conveniently displayed information reduces medical errors and the risks associated with providing treatment on the basis of clinical laboratory and other diagnostic data.

Administration

The hospital administration must support the safety program by allocating adequate personnel and financial resources and by enforcing compliance with policies and procedures and assigning responsibility to program implementation to a hospital safety officer.

Safety Officer

The key individual in a hospital safety program is the hospital safety officer. This position should be held by a

full-time hospital employee, especially in large, tertiary care, teaching hospitals. For small hospitals, community hospitals, and surgical centers the responsibilities of the safety officer's position can be exercised by an administrative staff member. The safety officer develops the administrative policies and procedures to address all aspects of safety within the hospital. Each department will have its own particular policies and procedures governing safety as will multidisciplinary committees.

Performance Improvement

A fundamental concept underlying all JCAHO standards is *performance improvement*, which is a continuous process of monitoring and evaluating performance to identify and realize opportunities for improvement. The performance improvement cycle links the safety committee, the safety officer, and the managers of the seven environment of care management plans into a framework for coordinating the many separate activities that constitute a hospital-wide safety program.

SAFETY PROGRAM: MEDICAL DEVICES AND INSTRUMENTATION

The focus in the past of most hospital safety programs on employee and device safety has changed to include a broader range of patient safety issues. The hospital safety program must address all the risks to which patients and staff are exposed. The duties and responsibilities of the hospital safety officer should reflect this broader safety perspective. Clinical engineering contributes to the hospital safety program principally through its management of medical device technology. A broadening of the clinical engineering department's perspective could be required in some cases, particularly in a hospital in which undue attention is placed on electrical safety issues.

Clinical Engineering

A medical device safety program can be administered in several ways: through an in-house clinical engineering (CE) department, either centralized and full service or departmentalized and partial service, through outside service provided by manufacturers' programs, or through an independent service organization (ISO) (36). The hospital personnel required to support such a program will range from one part-time worker when outside services are employed to 30–40 full-time workers when, for example, an in-house program is utilized in a major medical center (38).

Clinical engineering departments are concerned primarily with medical device performance and safety. The responsibility for servicing nonmedical devices and medical and nonmedical facilities may rest within or outside of clinical engineering; however, of overriding importance are the concepts that the hospital safety program should include all medical devices, medical device systems, utilities and facilities that may directly affect the safety of hospital patients or staff, and that the hospital should have in place a system that addresses all of these safety issues. The clinical engineering department's responsibilities with regard to the hospital safety program should include the implementation and execution of a medical device technology management program.

Technology Management

Medical device technology management is the sound foundation upon which any safety program must be built. Technology management is the management of medical devices throughout their useful life and includes selection and acquisition and eventual replacement (38). Elements of technology management follow:

- Strategic planning (technology assessment)
- Acquisition (specification, evaluation, purchasing)
- Utilization and asset management (impact analysis)
- Medical device service (inspection, maintenance, repair and modification)
- Replacement planning

When a technology management program is run effectively, a hospital can be assured that the need for all medical devices has been ascertained; that medical devices have been properly evaluated, selected, purchased, inspected, and calibrated for optimal performance and safety; that all users of the devices are appropriately trained; that the devices will be maintained on a regular basis; that service will be provided promptly so that patient care will not be disrupted; that a system of tracking and identifying every medical device will be in place; that the device is retired from service at the appropriate time; and that device replacement planning ensures continuity of technological support. Documentation of all elements of the system is essential. A deficiency in any one of the above elements can increase the exposure of a hospital patient or employee to injury. From technology management flow all other aspects of a safety program.

Strategic Planning

The first step in minimizing the risk of medical device technology is strategic planning and technology assessment, determining what services will be provided and what technologies are appropriate for delivering those services.

Hospitals need to monitor changes in technology to enable sound strategic planning. For example, examination of the effects of new therapeutic or diagnostic modalities, such as magnetism in the case of MRI, would allow the hospital to circumvent incompatible or health-threatening situations. The hospital should maintain ready access to technological assessment reports such as those published by the National Center for Technology Assessment. A *chief technology officer* may be warranted for large, university-affiliated teaching hospitals. The acquisition of appropriate medical device technologies can be aided by the use of such tools as the Essential Health Technology Package developed by the World Health Organization (39).

Acquisition

Acquisition of adequate medical device technologies starts with a statement of clinical needs. Detailed specifications

are then written, accurately reflecting the needs expressed by the clinical staff. Prospective vendors are then selected and invited to demonstrate their products. Clinical and laboratory evaluation of these products will determine suitability to the hospital's needs and adherence to safety and performance standards. The evaluation will often involve a multidisciplinary approach so that all aspects of the device (e.g., life cycle cost analysis, clinical efficacy, safety, and engineering design) can be evaluated. A hospital standards and evaluation committee is a good mechanism to control the acquisition of a broad range of medical devices from infusion pumps to gauze pads. Purchasing documents should contain device performance specifications and conditions of sale such as withholding payment until all in-coming acceptance inspections are complete and the device is installed and operational. Other conditions include provision of operator and service manuals, vendor training, and warranty periods.

Utilization and Asset Management

At the initial (in-coming) inspection of a device entering the hospital, a record of that device (asset) is entered into the management system. Generally, affixing an asset management control number to the device enables the identification of that particular device and allows a computerized medical device management system to store and retrieve device histories. Computerized Medical Device Management Systems greatly enhance the acquisition and analysis of data and thus improve the management process (40). The unique control number points to a file containing information about the device that is essential to management of that resource. Such information as generic name of the device, manufacturer, location, and phone number of the manufacturer or authorized service representative, applicable inspection procedure to be used during scheduled preventive maintenance, and cost of the device are generally included in the medical device data base. The management system linked to a *medical device recall system* (see below) enables rapid determination of whether or not the hospital possesses devices that are subject to a recall.

Asset management data of the acquired medical device technology can support utilization and impact analysis enabling the hospital to reap maximum economic benefit (41). Improvements in utilization can positively affect a patient's care by maximizing availability of diagnostic and therapeutic services.

Medical Device Service

Periodic inspection, maintenance and repair are essential elements of medical device safety. Whether these functions are carried out by an in-house department or by an outside service organization, it is crucial that the service be performed in a timely manner so that there is negligible impact on the delivery of care for want of the necessary device or instrument. Several guides to developing maintenance and repair programs are available (42). Techniques for obtaining maximum support from vendors and outside service providers have been described (43).

Replacement Planning

All medical devices reach the point in their lives where replacement becomes necessary because of decreased reliability, increased downtime, safety issues, compromised care, increased operating costs, changing regulations or, simply, obsolescence. Healthcare organizations have limited funding for capital purchases with many under strict budgeting guidelines. Replacement based on anecdotal and subjective statements and politics create havoc related to finances, morale, and operations. Clinical engineering departments are impacted if they do not have a replacement plan when major repairs occur in older equipment.

The ideal healthcare technology replacement planning system would be facility-wide covering all clinical equipment, utilize accurate, objective data for analysis, be flexible enough to incorporate nonequipment factors, and be futuristic by including strategic planning related to clinical and marketplace trends and hospital strategic initiatives related to technology. The plan should encompass many factors related to cost benefit, safety, support, standardization, and clinical benefit (44).

Medical Device Recall System

A medical device recall system should be in place in all hospitals to receive recall notices, hazard bulletins, and safety alerts and to disseminate them to the appropriate authorities (45). The system should have a means for documenting any corrective actions required and implemented. The pharmacy typically operates a separate drug recall system. The *medical device safety officer* (see below) should manage the recall system. Close interdepartmental communication is required. All clinical departments, ancillaries, and support or service departments within the hospital are typically involved in medical device recall system. Virtually all types of medical devices are subject to recall: reusables or disposables, accessories and supplies, and capital equipment. Every medical device from tongue depressor to computed tomography (CT) scanner are included in the system.

Several publications contain hazard and recall notices including *FDA Enforcement Report, Biomedical Safety Standards Newsletter, and Health Devices Alerts* (see Reading List).

Incident–Accident Investigation and Reporting

The hospital safety program should encompass the mechanism to respond to an accident or incident in which a patient or staff member is injured (46). Incidents and accidents associated with medical devices and instrumentation fall within the purview of the clinical engineering department. The Safe Medical Devices Act of 1990 requires that the manufacturer of a medical device be notified within 10 business days if one of its devices contributed to death or serious injury. The manufacturer, under FDAs Mandatory Device Reporting (MDR) regulation, is required to report these incidents to the FDAs Center for Devices and Radiological Health. Manufacturers are required to report to the FDA when they receive information from any source which reasonably suggests that a

device they manufacture contributed to death or serious injury (47).

The clinical engineering department has the most familiarity with medical device performance and safety features and would be best qualified to serve as incident investigation coordinator of all medical device-related incidents. Many clinical engineers are ideally qualified by education, training, and experience to investigate incidents. When in-house clinical engineering expertise is unavailable or when an outside, unbiased investigation is required, hospitals make use of independent third-party investigators.

Medical Device Accident and Incident Investigation Methodology

Impoundment of devices after incidents is recommended. Control settings, dials, knobs, and switches should not be changed. Photographic evidence should be obtained. Disposable devices must be saved along with packaging material that can often be the only indication of lot number. Cleaning, disinfection, or sterilization of the item(s) involved could hinder further investigation and should be done only under the guidance of the incident coordinator. Policies should be written advising clinical staff to preserve all equipment involved in an incident, including disposable devices and associated packaging. Relative positions of devices in the room should be noted. Staff should record all relevant identifying information, such as the manufacturer of the device, date of use, location, serial or lot number, and hospital equipment control number if applicable.

Incidents and accidents often go unreported for fear of legal repercussions and from reluctance to admit that a procedural error may have been committed. An error in the use of a device may be related more to the human factors engineering failings of the device than to the carelessness of the operator. Only through aggressive reporting of medical device problems can the need for technological improvements be identified and the likelihood of similar incidents occurring be eliminated.

Quality Assurance

A means by which the CE department can measure the quality of its service is recommended (48). A departmental quality assurance plan should aim to identify problems or potential problems related to the safe, effective, and appropriate use of patient care equipment. Quality assurance indicators would include such items as the number of in-service education sessions, the number of device prepurchase evaluations, the average length of time required to restore damaged or defective equipment to working condition, the scope of department services, productivity, and proficiency. A computerized medical device management system is an invaluable aid in developing the necessary statistics for measuring departmental performance.

Information Technology

Clinical engineering and information technology are converging as an increasing number of medical devices become integrated into hospital information systems (49). Sensitive patient information is contained in many medical devices and thus the security provisions of the Health Insurance Portability and Accountability Act (HIPAA) apply (50). Inappropriate use of medical information is a patient safety hazard. For example, reporting a negative HIV result to a person who is actually HIV positive could result in that person not taking appropriate steps to reduce the spread of the disease.

Education and Training

Developing a comprehensive education program in the use of medical devices is one of the principal goals of the hospital safety program (51). Changing legal attitudes in response to technological advances have resulted in additional responsibilities for nurses. Training to enhance the recognition of the day-to-day condition of devices, the need for in-service, and alertness to the signs of incipient device failures are essential components of a sound medical device safety program.

Tools

Professional organizations provide a good forum for the exchange of safety information. Keeping current with developments in medical device technology, patient and hospital safety, and regulatory issues by attendance at meetings of these organizations is recommended. The American College of Clinical Engineering (ACCE) is the only international professional organization dedicated exclusively to representing the interest of clinical engineers (53). Other biomedical engineering, facilities engineering, and medical instrumentation organizations support clinical engineering efforts at the national, regional, and local levels. These include the Engineering in Medicine and Biology Society of the Institute of Electrical and Electronics Engineers, the American Society of Hospital Engineers of the American Hospital Association, and the Association for the Advancement of Medical Instrumentation.

Numerous resource materials are available to assist the clinical engineering department in administering a safety and technology management program. The Reading List section includes texts that contain instructional material relating to technology management and patient safety. Teleconferences, workshops, and Internet also serve as good sources of information (53–55). Manufacturers, recognizing the need to provide informative material to the user to promote a better understanding of safety issues, often provide beneficial educational material.

Medical device standards, technology management recommendations, and safety guidelines are promulgated by several professional and trade organizations. Clinical engineers are encouraged to participate in standards-setting activities to contribute their informed user opinion on medical device performance and safety issues.

ENHANCING PATIENT SAFETY

The clinical engineer has many opportunities to design means to reduce the occurrence of medical errors and to

enhance patient safety through the application of engineering skills to solve problems of medical device selection, application, utilization, and safety (56,57). The proactive engineer will implement solutions to problems before they occur. Root-cause-analyses and failure modes and effects analysis (FMEA) will aid the engineer in determining what is wrong with the system and what needs to be fixed (58,59).

Medical Device Safety Officer

The medical device safety officer (MDSO) will one day be commonplace in hospitals. The clinical engineer is well suited to serve in this role and to manage the medical device safety program.

Postmarket Surveillance

Improved postmarket surveillance of medical device performance and safety is needed to provide more device problem feedback to the manufacturer, regulatory authorities, and to the hospital itself (60). Such feedback enables enhanced device design, revisions in standards, and improvements in educational programs with the ultimate goal of reduction in hazards. Improved methods are required to determine the prevalence, nature, and causes of medical device-related errors.

BIBLIOGRAPHY

1. Kohn LT, Corrigan JM, Donaldson MS, editors. To Err is Human: Building a Safer Health System. Washington, DC: National Academy Press; 2000.
2. Trandel-Korenchuk DM, Trandel-Korenchuk KM. Legal forum, malpractice and preventive risk management. Nurs Admin Q 1983;7:75–80.
3. Steel K, Gertman PM, Crescenzi C, Anderson J. Iatrogenic illness on a general medical service at a university hospital. N Engl J Med 1981;304:638–641.
4. Abramson N. et al., Adverse occurrences in intensive care units. JAMA, J Am Med Assoc 1980;244:1582–1584.
5. Enhancing Patient Safety: The Role of Clinical Engineering. Plymouth Meeting, PA: American College of Clinical Engineering; 2001.
6. JCAHO. 2001. Revisions to Joint Commission standards in support of patient safety and medical/health care error reduction. Joint Commission on Accreditation of Healthcare Organizations. Oakbrook Terrace, IL.
7. Shojania KG, Duncan BW, McDonald KM, Wachter RM, editors. Making Health Care Safer: A Critical Analysis of Patient Safety Practice. Evidence Report/Technology Assessment No. 43 (Prepared by the University of California at San Francisco—Stanford Evidence-Based Practice Center under Contract No. 290-97-0013), AHRQ Publication No. 01-E058, Rockville, MD. Agency for Healthcare Research and Quality; 2001. p 668.
8. Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: National Academy Press; 2001.
9. Bogner MS, editor. Human Error in Medicine. Hillsdale, NJ: Lawrence Erlbaum Associates; 1994.
10. Shepherd M. Systems approach to medical device safety. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
11. Federal Food, Drug, and Cosmetic Act, As Amended October 1976. Washington, DC: U.S. Government Printing Office; 1976.
12. Ridgway MG. The great debate on electrical safety—in retrospect. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
13. Dyro JF, Tai S. Evaluation of ethylene oxide sterilizers. Proc Ann Conf Eng Med Biol 1976;18:499.
14. Special Report! 2003 JCAHO National Patient Safety Goals: Practical Strategies and Helpful Solutions for Meeting These Goals. Joint Commission Resources 2003;3(1):1–12.
15. Hyman WA. Risks associated with diagnostic devices. J Clin Eng 1986;11:273–278.
16. ANSI/AAMI Human factors design process for medical devices; ANSI/AAMI HE74:2001. Arlington, VA: Association for the Advancement of Instrumentation; 2001.
17. Hyman WA, Wangler V. Human factors: Environment. In: Dyro JF editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
18. Bowyer EA. Anesthesia claims study identifies recurring areas of loss. Forum 1985;6(2):3–5.
19. Hyndman B. Physical plant. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
20. Hernández D, Hernández A. Water. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
21. Klein E, Evaluation of Hemodialyzers and Dialysis Membranes, DHEW/NIH 77-1294. Bethesda, MD: U.S. Department of Health, Education, and Welfare, Public Health Service, National Institutes of Health; 1977.
22. Brito LF, Magagna D. Sanitation. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
23. Miodownik S. Interactions between medical devices. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
24. Paperman WD, David Y, Hibbetts J. Electromagnetic interference in the hospital. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
25. Tan K-S, Hinberg I. Electromagnetic interference with medical devices: In vitro laboratory studies and electromagnetic compatibility standards. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
26. Witters D, Campbell CA. Wireless medical telemetry: Addressing the interference issue and the new wireless medical telemetry service (WMTS). In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
27. Ben-Zvi S, Gottlieb W. Medical instrumentation and nosocomial infection. J Clin Eng 1979;4:135–145.
28. Epstein A, Harding GH. Disaster planning. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
29. Baretich MF. Hospital safety programs. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
30. Comprehensive Accreditation Manual for Hospitals. Oakbrook Terrace, Illinois: Joint Commission on Accreditation of Healthcare Organizations; 2001.
31. Harding GH, Epstein A. Risk management. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
32. Ridgway MG. Hospital environmental safety and the safety codes. J Clin Eng 1977;2:211–215.

33. Strzelczyk J. Radiation safety. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
34. International Commission on Radiological Protection. Radiological Protection and Safety in Medicine. Oxford: Pergamon Press; ICRP Publication 73; Ann ICRP 26(2); 1996.
35. Goodman G. Hospital facilities safety standards. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
36. Smithson P, Dickey D. Outsourcing clinical engineering service. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
37. Soller I. Clinical engineering in an academic medical center. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
38. David Y, Judd TM, Zambuto RP. Introduction to medical technology management practices. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
39. Heimann P, Issakov A, Kwankam Y. The essential healthcare technology package. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
40. Cohen T, Cram N. Computerized maintenance management systems. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
41. Seaman G. Industrial/management engineering in healthcare. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
42. Hertz E. Medical equipment management program and ANSI/AAMI EQ56. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
43. Dyro JF. Vendor and service management. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
44. Clark JT. Healthcare technology replacement planning. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
45. Hyman WA, Schlain LA. User problems and medical device recalls. Med Instrum 1986;20:14–16.
46. Dyro JF. Accident investigation. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
47. Winston FB. United States Food and Drug Administration. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
48. Autio D. Clinical engineering program indicators. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
49. Cohen T, Ward C. The Integration and Convergence of Medical and Information Technologies.
50. Grimes SL. Health Insurance Portability and Accountability Act (HIPAA). In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
51. Bauld TJ, Dyro JF, Grimes SL. Clinical engineering and nursing. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
52. Ott JC, Dyro JF. American College of Clinical Engineering. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
53. Wear JO, Levenson A. Distance education. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
54. Dyro JF, Judd TM, Wear JO. Advanced clinical engineering workshops. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
55. Lozano-Nieto A. Emerging technologies: Internet and interactive videoconferencing. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
56. Patail B. Patient safety and the clinical engineer. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
57. Cram N, Stephens J-P, Lessard C. The role of clinical engineers in reducing medical error. J Clin Eng 2004;29(1):33–35.
58. Root Cause Analysis in Health Care: Tools and Techniques. Oakbrook Terrace, IL: Joint Commission on Accreditation of Healthcare Organizations; 2000.
59. Stalhandske E, DeRosier J, Patail B, Gosbee J. How to make the most of failure mode and effects analysis. Biomed Instr Tech 2000;34(2):96–102.
60. Cheng M. Post-Market Surveillance and vigilance on medical devices. In: Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.

Further Reading

Wildavsky A. Searching for Safety. New Brunswick, NJ: Transaction Publishers; 1989.
ACCE News. Plymouth Meeting, PA: American College of Clinical Engineering (bimonthly).
Accreditation Manual for Hospitals. Chicago: Joint Commission on Accreditation of Healthcare Organizations (annual).
Biomedical Instrumentation & Technology. Arlington, VA: Association for the Advancement of Medical Technology (bimonthly).
Biomedical Safety & Standards. Hagerstown, MD: Lippincott Williams & Wilkins (semimonthly).
Gendron FG. Unexplained Patient Burns: Investigating Iatrogenic Injuries. Brea, CA: Quest Publishing Co.; 1988.
FDA Enforcement Report. Rockville, MD: U.S. Food and Drug Administration (weekly).
Health Devices Alerts. Plymouth Meeting, PA: ECRI (semimonthly).
Health Facilities Management. Chicago: Health Forum, Inc.
Bronzino JD. Management of Medical Technology. Boston: Butterworth-Heinemann; 1992.
Kolka JW, Link DM, Scott GG. European Community Medical Device Directives:Certification, Quality Assurance and Liability. Fairfax, VA: CEEM Information Services; 1992.
Joint Commission Perspectives on Patient Safety, Oakbrook Terrace, IL: Joint Commission Resources (monthly).
Journal of Clinical Engineering. Lippincott Williams & Wilkins, Hagerstown, MA (quarterly).
Shojania KG, Duncan BW, McDonald KM, Wachter RM. editors, Making Health Care Safer: A Critical Analysis of Patient Safety Practice. Evidence Report/Technology Assessment No. 43, AHRQ Publication No. 01-E058, Rockville, MD: Agency for Healthcare Research and Quality, 2001.
Geddes LA. Medical Device Accidents and Illustrative Cases. 2nd ed. Tucson: Lawyers & Judges Publishing Company; 2002.
MD&DI/Medical Device & Diagnostic Industry, Los Angeles, CA: Canon Communications (monthly).
Medical device user facility and manufacturer reporting, certification and registration. 21 CFR Part 803, July 31, 1996.
NCPS Triage Cards for Root Cause Analysis. National Center for Patient Safety, Department of Veterans Affairs, 2001.
National Electrical Code: NFPA 70. Quincy, MA: National Fire Protection Association, 2002.
Fish RM, Geddes L. Medical and Bioengineering Aspects of Electrical Injuries. Tucson: Lawyers & Judges Publishing Company; 2003.

Zambuto RP, When worlds collide, Health Facilities Management Magazine. April 2004.
Reiser SJ, Anbar M, editors. The Machine at the Bedside: Strategies for Using Technology in Patient Care. New York: Cambridge University Press; 1984.
Shepherd's System for Medical Device Incident Investigation and Reporting. Shepherd & Baretich. Baretich Engineering, 2003.
Standard for Health Care Facilities: NFPA 99, Quincy, MA: National Fire Protection Association, 2002.
Dyro JF, editor. The Clinical Engineering Handbook. Burlington, MA: Elsevier; 2004.
The Healthcare Failure Mode Effect Analysis Process. National Center for Patient Safety, Department of Veterans Affairs, 2002.

See also CODES AND REGULATIONS: MEDICAL DEVICES; EQUIPMENT MAINTENANCE, BIOMEDICAL; GAS AND VACUUM SYSTEMS, CENTRALLY PIPED MEDICAL; IONIZING RADIATION, BIOLOGICAL EFFECTS OF; RADIATION PROTECTION INSTRUMENTATION; X-RAY QUALITY CONTROL PROGRAM.

SCAFFOLD MATERIALS, STERILIZATION OF.

See STERILIZATION OF BIOLOGIC SCAFFOLD MATERIALS.

SCAFFOLDS.

See BIOMATERIALS: TISSUE ENGINEERING AND SCAFFOLDS.

SCANNING ELECTRON MICROSCOPY.

See MICROSCOPY, ELECTRON.

SCANNING FORCE MICROSCOPY.

See MICROSCOPY, SCANNING FORCE.

SCANNING TUNNELING MICROSCOPY.

See MICROSCOPY, SCANNING TUNNELING.

SCOLIOSIS, BIOMECHANICS OF

I.A.F. STOKES
C.E. AUBIN
Polytechnique Montreal
Montreal, Quebec, Canada

INTRODUCTION

Scoliosis deformity occurs in the spine quite frequently, especially in growing children. In spite of the attention given to this deformity over the past century, still very little is known about its etiology or progression mechanisms.

There are a number of health related consequences of a progressive scoliosis. In large scoliosis deformities there is impairment of respiratory and cardiovascular function. If progressive, the deformity produces thoracic compromise, pain, and degenerative changes in the spine. In large curves that have been surgically fused, there may be iatrogenic problems secondary to multisegmental arthrodesis. Although scoliosis is often perceived as having a minimal cosmetic impact, it has psychosocial consequences (1–3).

TERMINOLOGY AND GENERAL MORPHOLOGY OF SCOLIOSIS

Geometry of Scoliosis

Scoliosis literally means a lateral curvature of the spine, but as a deformity it includes overall asymmetry and lateral deviation of the trunk, as well as axial rotation of the spine and trunk. The total geometric description of scoliosis is of paramount importance and merits more detailed characterization than can be given by a single-plane X ray. Three-dimensional (3D) measurements of the complete spinal and thoracic geometry require complex techniques of radiographic and surface measurement. Some measurement techniques currently used in research may achieve clinical application in the future.

Frontal, Sagittal and Transverse Planes Components of Spinal Deformity

Frontal Plane. Scoliosis is strictly defined as a lateral curvature of the spine although the scoliosis deformity is a 3D abnormality of the axial skeleton. Scoliotic curves have been described as being primary or secondary, major or minor, or structural or compensatory. The structural curve is defined as having asymmetry not reversible in lateral bending. Usually, a compensatory curve is present to maintain the alignment of the body so that the head is over the pelvis. When two curves of the same spine are defined as structural, the resultant S-shaped deformity is called a double structural curve. The apex of a scoliosis curve is at the vertebra or disk that is the most laterally deviated from the vertical axis passing through the sacrum (base of the spine) (4).

The Sagittal Plane. The mid-sagittal plane is the plane that divides the left side of the body from the right. The terms kyphosis and lordosis (Fig. 1) are used to describe curvatures that exist normally in the sagittal plane, with the prefixes hyper- or hypo- describing abnormal curvatures. Curvatures in this plane interact with the abnormal lateral curvature of the spine in scoliosis. This interaction is termed coupling (see below).

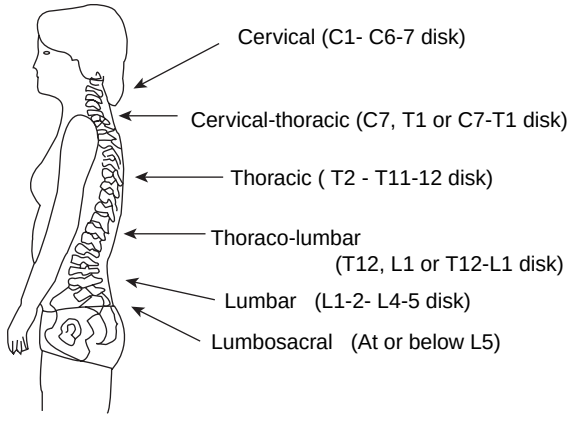


Figure 1. Lateral view of the spine defining regions of curvature.

Transverse Plane. Axial rotation of vertebras accompanies the lateral deviation. The vertebral rotation can be measured from a frontal plane radiograph by grading the positions of the vertebral pedicles relative to the center of the vertebral body. Also, templates and formulas have been developed to convert these pedicle offsets into estimates of the rotation in degrees. The rotation can be measured more directly from transverse plane images obtained by computed tomography (CT) scanning or magnetic resonance imaging (MRI). The usual rotation of the vertebral bodies is to the convex side (5). Asymmetries of the vertebras also develop, probably as a result of altered stress on these bones (6). The back surface rotates in the same sense, producing a "rib hump" (gibbosity). Back surface rotation correlates weakly with the scoliosis magnitude as measured by the Cobb angle (7), leading to the idea that curve magnitude could be measured topographically, without X rays. However, the back surface rotation is of lesser magnitude than the vertebral rotation (8), which in turn is of lesser magnitude than the Cobb angle, so this attenuation of the asymmetry as seen at the surface produces technical measurement difficulties.

Stereoradiography and stereophotogrammetry have been used to document the transverse plane deformity in patients with idiopathic scoliosis with simultaneous radiography to record the skeletal shape (9). The back surface rotation was reported to be greatest at the level of the apex of a lumbar scoliotic curve, but located as much as four vertebras lower than the apex in thoracic and thoracolumbar curves (9). Thus the surface deformity is apparently augmented by the effects of the rib cage where the ribs point downward. In thoracic curves the radiographic rotation of vertebras is reported to be greatest on average at the apex (defined as the most laterally deviated vertebra), but with a range of three vertebras above to three below (9,10). In most of the lumbar curves the greatest surface rotation, the maximal vertebral rotation, and the apex coincides. The axial rotation of the surface had a variable relationship with the Cobb angle and in all cases was less than the Cobb angle (11). The amount of back surface rotation is less than the amount of axial rotation of the underlying vertebras (9).

Coupling of Rotations. Coupling is the term used to describe the tendency of the spine while moving intentionally in one plane to produce secondary or coupled motion in another plane. Two kinds of coupling can be defined, orientation and kinematic. Orientation coupling describes the relationships between the linear or angular displacements of a single vertebra from its reference or nonscoliotic position. Kinematic coupling is defined as the association of two variables describing the relative motion of two vertebras without regard to forces. Kinematic coupling has described from *in vivo* measurements of normally curved spines (12–14). The relationship may be altered by adaptation of bone and soft tissue resulting from the scoliosis deformity.

Lateral bending of the healthy lumbar spine is accompanied by a rotation in the same direction as that seen in scoliosis (15); however, the coupling of axial rotation with lateral bending is in the opposite sense in the thoracic spine (16). In normal spines, lumbar curves produced by lateral bending are associated with a vertebral body rotation toward the concavity of the curve (15). This is not the direction of the static rotation in a spine with scoliosis. The orientation coupling in scoliosis has the same direction as was noted in lateral bending of normal spines in extended positions.

In scoliosis, for each vertebra there are five orientation variables of interest: flexion–extension angle, lateral tilt angle, axial rotation angle, and lateral and anteroposterior deviations from the spinal axis. This leads to 10 possible coupling coefficients, which are ratios between pairs of these geometric variables. There is no reason *a priori* why orientation coupling in scoliosis and/or kinematic coupling coefficients in healthy spines should be related especially when ongoing growth complicates the relationship between different components of the deformity. However, after the end of growth, if a kinematic process governed by such coupling occurred during curve progression, then these relationships would be apparent.

Rib Cage and Back Surface Shape

The scoliosis deformity of the spine (curvatures and associated rotation) is associated with a complex pattern of rib cage deformity, evident as a thoracic rotation (rib hump). The rib hump becomes more subjectively evident in the back with scoliosis on forward bending of the trunk but its magnitude is in fact not increased (17). Thoracic rotation or rib hump can be measured on clinical examination in forward bending and recorded as the depression from the horizontal plane, or as the angular rotation from the horizontal (18) (Fig. 2). The balance of the torso or the extent to which the top of the spine deviates from being vertically above its base (sacrum). Several contacting and noncontacting (optical) methods are available to document the surface trunk deformity (reviewed below).

Classification of Scoliosis

Scoliosis is considered to exist if the Cobb angle is $>10^{\circ}$. The pattern of scoliosis can be classified by different criteria.

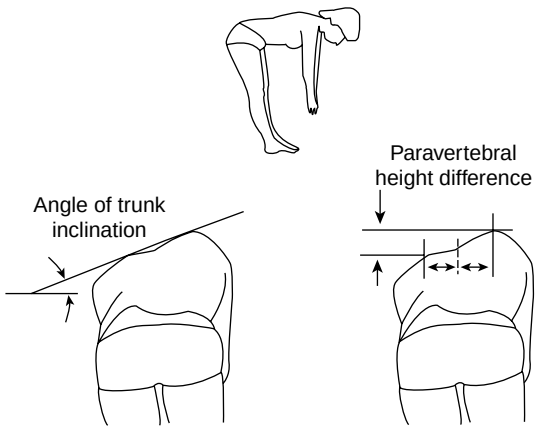


Figure 2. Rib hump seen in forward bend position.

Classification by Anatomical Region of the Curve. In classification of a curve by the location of the apex: the apex of a cervical scoliosis is between C1 and the C6–7 disk; cervicothoracic scoliosis between C7 and T1; thoracic scoliosis between T2 and the T11–T12 disk; thoracolumbar scoliosis between T12 and L1, lumbar scoliosis between the L1–L2 and L4–L5 disk spaces, and lumbosacral scoliosis below L5 (19).

Classification by Age. Scoliosis can be classified according to the skeletal age of the patient at diagnosis (not necessarily the same as the age of onset). Infantile scoliosis is diagnosed during the first 3 years of life, juvenile scoliosis between the ages of 4 and 10 years; adolescent scoliosis is a scoliosis diagnosed between ages 10 and 18 years, and adult scoliosis is diagnosed after age 18 years (19). A large percentage of infantile cases resolve spontaneously. These cases may be related to the position of the body in utero, but generally the likelihood of progression of the deformity is worse the earlier the age at which it develops, consistent with progression being associated with continued skeletal growth. Usually, skeletal maturation is complete after the age of 17 in females and after the age of 19 in males. For reasons that are unclear, in idiopathic scoliosis (81% of all scoliosis) the convexity is more common on the left side in lumbar curves and on the right side in thoracic curves.

Classification by Causation. In most instances, scoliosis can also be separated into one of the four major subgroups depending on the probable initial cause: idiopathic (spontaneous or unknown cause), neuromuscular (associated with disease of the neuromuscular system), congenital (associated with a prenatal developmental anomaly) (20), and iatrogenic (from radiation or thoracic surgery). Idiopathic scoliosis (for which no obvious cause is ascertained) represents 81% of the cases. There are biomechanical factors associated with scoliosis of any cause, since the deformity alters the biomechanics of the spine and trunk, and this in turn alters the processes of musculoskeletal growth, degeneration, and so on.

Classification by Curve Pattern. Classification systems are used to guide the management of scoliosis when a curve type can be consistently related to a different prognosis or management strategy. In surgical planning for the surgical management of adolescent idiopathic scoliosis, classification by radiographic measures is used to help decide on the extent of the arthrodesis. The King et al. classification (21) is probably still the most widely used in planning of spinal fusions, although it was originally developed to aid planning for Harrington instrumentation, that is seldom used now. It defines five thoracic scoliosis curve types, and an additional group called miscellaneous. The King et al. classification relies on subjective identification of the radiographic features that provide the measures used for classification. It may also require individual interpretation and memory of the classification criteria. As a result, there are numerous opportunities for variable implementation that produce inter- and intraobserver variability. A recent empirical study (22) of repeat curve-type classification has demonstrated poor reliability. An algorithm (23) aims to overcome these problems by defining an objective classification procedure.

The Lenke (24) classification system was developed to provide a comprehensive and reliable means to categorize all surgical AIS curves, using posteo anterior (PA), lateral, and side bending X-ray films. The three classification components are curve type (1–6), a lumbar spine modifier (A, B, C), and a sagittal thoracic modifier (−, N, +). The largest curve is considered to be the major curve, and minor curves are subdivided into structural and nonstructural types.

Axes and Coordinate Systems

The global coordinate system describes the positions and displacements of vertebras relative to the whole body; regional coordinates systems employ a reference within a spinal curve or the entire spine, and local coordinate systems are fixed in a particular vertebra and aligned with its anatomic features are needed to describe motion between vertebrae (Fig. 3). The Scoliosis Research Society (SRS) (4,25) placed the origin at the center of the superior endplate of S1 for both global and spinal axis systems of patients with scoliosis. This is a commonly accepted convention. The ISO 2631 (VDI 2057) right-handed axis convention has x signifying anterior, y signifying left, and z signifying the cephalad direction. The global and spinal axis systems have the origin and sagittal plane defined by the pelvis, with the anterior superior iliac spines (ASIS) defining the transverse global (Y) direction. (This would normally be achieved by positioning the ASIS parallel to the X-ray film plane.) The other principal directions are aligned either with gravity (global system), or with spinal landmarks.

Plane of Maximum Curvature. The geometry of the spine can also be considered as a curved line in space with vertebras positioned along this vertebral body line (4).

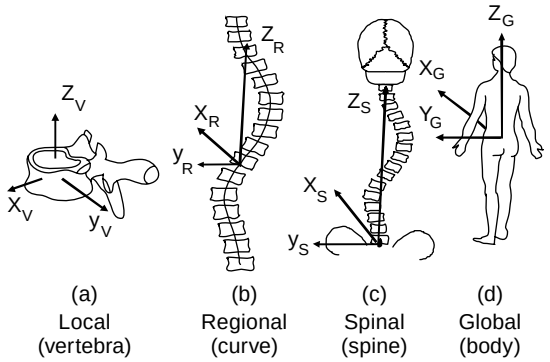


Figure 3. The hierarchy of four coordinate systems defining spinal geometry: (a) Local coordinates based on a vertebra. (b) Regional, curve-based coordinates defined by the end vertebras of a curve or other spinal region. (c) Spinal coordinates, defined by the Z axis passing through the centers of the most caudal and cephalad vertebrae of the entire spine. (d) Global, whole-body-based coordinate system, with its origin at the base of the spine (S1), and the Z axis vertical (gravity line).

The vertebral body line in turn defines a plane of maximum curvature and best-fit plane (4,26), that lie in the mid-sagittal plane for an undeformed (symmetrical) spine, but are rotated from that plane when a scoliosis is present. In a straight, healthy spine the plane of maximal curvature is the sagittal plane, but in a scoliosis the maximum curvature is seen by viewing the spine from an intermediate angle that may be rotated from the sagittal plane by approximately the same number of degrees as the rotation of the apical vertebrae (27). In a purely geometrical sense, the trihedron axis system (28) provides the basis for 3D measurements, including the local curvature and the geometrical torsion of the vertebral body line. It also defines the osculating plane (local transverse plane), the tangent and normal directions, and the binormal. While providing a pure geometrical approach to defining spinal geometry (29), this approach has not been much used either in clinical or many research contexts.

Measurement of Curve Magnitude. The degree of the scoliosis deformity is most commonly measured clinically by the Cobb angle, which is the angle between lines drawn on the end plates of the most tilted vertebras in the radiographic curve as shown in Fig. 4. The Cobb (30) measurement of a scoliosis is the sum of the angles of inclination in the frontal plane of the two end vertebrae. The end vertebras are defined as those with the largest inclinations to the horizontal, and are typically close to the points of inflection of the curve. End vertebras normally have no axial rotation. Reliability studies estimate the interobserver error of the Cobb measurement as 5° (23). The Cobb angle measurement summarizes curve magnitude, but ignores numerous important structural characteristics including the length of the curve (numbers of vertebras in it), its flexibility, and the relative contributions of vertebral and discal wedging. While it is presently the major objective measurement used in planning management of patients with scoliosis, these shortcomings suggest that there is scope for improvements in clinical measurement.

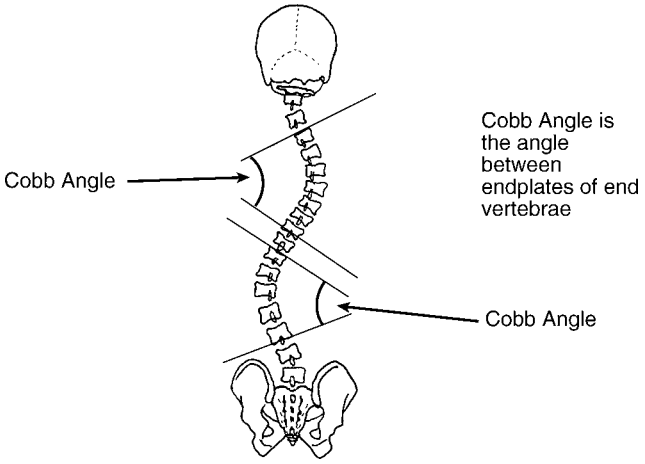


Figure 4. The Cobb angle as used to measure scoliosis curve magnitude. It is defined as the angle between the lines drawn on upper and lower end-vertebral endplates, where the end-vertebras are those having the greatest inclination to the horizontal.

FUNCTIONAL ANATOMY

Spine: Vertebras, Disks, Facet Joints

The structure and function of the spinal column depends on a complex interaction between the vertebral bodies, the articulations between their posterior elements (facet, or zygoapophyseal joints), interconnecting ligaments and intervertebral disks, the rib cage and the trunk musculature. In a scoliosis the vertebras become wedged and twisted, the intervertebral disks become wedged, and there is a variable degree of adaptive changes in the soft tissues (ligaments, muscles, etc.) that probably increase with age. The ribs are more sharply angled on the convex side, and this asymmetry, together with the axial rotation component of the deformity produces the rib hump or gibbosity that is the most evident outward indication of the deformity. Surgical treatment of adult scoliosis is more difficult in part because of the adaptive changes and the typically greater curve magnitude and cardiothoracic compromise.

The Spinal Motion Segment

The spinal motion segment is a structural unit of the spine consisting of two vertebras and the intervening soft tissues. It is a valuable tool in spinal biomechanics because it is the basic element of the spinal column. Once motion segment behavior is defined, the behavior of the whole spine can be represented by a series of fundamentally similar components. In the absence of scoliosis, the motion segment can be considered as being symmetrical about the sagittal plane, thus right and left lateral bends, axial rotation and shear are theoretically symmetrical. The presence of the posterior elements (zygoapophyseal or facet joints and the ligaments) introduces differences between the flexion and extension behaviors. The presence of these posterior structures also produces a posterior displacement (relative to the disk center) of the effective structural axis of the motion segment. The stiffness of the lumbar motion segments has been described by a stiffness matrix, and by an equivalent beam structure (31).

Ribcage and Costovertebral Articulations

Motion of the thoracic spine is connected to that of the rib cage by the costovertebral joints. The ribs articulate with the transverse processes, as well as with the vertebral bodies of the corresponding vertebra and that immediately cephalad. Motion of the ribs has been described as consisting of both pump-handle and bucket-handle motion. Although not well defined, these terms imply rotations about the horizontal (mediolateral axis) and an axis joining the ends of the ribs, respectively. The complex interactions in chest wall mechanics during breathing are difficult to explain in terms of muscular recruitments that involve the diaphragm, as well as the intercostal musculature and abdominal musculature. Rib deformation is part of the scoliosis deformity, thus complicating the progression of thoracic curves as well as their surgical treatment (32) where costoplasty is sometimes considered. The biomechanical consequences of costoplasty were analyzed biomechanically by Aubin and co-workers (33,34).

KINEMATICS OF THE SPINE AS IT RELATES TO SCOLIOSIS

Very little is known about how a scoliosis deformity specifically affects spinal kinematics, with little information either from cadaveric specimens or intraoperative measurements. Studies of gait (35) suggest that asymmetrical spinal axial rotation motion is the main kinematic consequence of a scoliosis. After spinal fusion, the overall motion of the spine is reduced as expected, but with reduced motion also in unfused segments (36).

Range of Motion, Coupled Motion, Intervertebral Stiffness Matrix

The intervertebral articulations have six degrees of freedom (three translations and three rotations) each of which has a measurable stiffness. Traditionally, the load-displacement characteristics of these joints have been described by a stiffness matrix (37). This stiffness matrix has off-diagonal terms, as well as diagonal terms. The off-diagonal terms resulting from this tendency of certain degrees of freedom to be associated with each other (especially axial rotation and lateral bending) has been referred to as coupling. The pattern of motion that occurs between two vertebras depends on the combination of forces applied, and the axis of rotation is not fixed. It is only possible to define an instantaneous axis of rotation.

Most experimental motion segment stiffness data are limited as they do not include all six degrees of freedom, were obtained by inverting flexibility data, and were obtained without physiological levels of axial compressive force. Physiological axial compression has been observed to increase lumbar motion segment stiffness by a factor of 2 or more (31,38). Janevic et al. (38) reported that the stiffening of the motion segment with preload was approximately linear with preload magnitude. With 2200 N preload, rotational flexibility decreased on average 2.6 times, and shear flexibility 6.16 times, the effects being even greater at 4400 N preload. The stiffening of the motion segment with preload appears to result from nonlinearities in the intrinsic material properties of the annulus fibers, as well as the more easily visualized engagement of the facet joints (31).

METHODS TO DOCUMENT TRUNK DEFORMITY

Plane Radiography, and Derived Measures

The PA radiograph is normally used for documentation of scoliosis to show the lateral curvature and some indication of the axial rotation. The anteroposterior (AP) radiograph delivers higher radiation dose to breast tissue, sternum, thyroid, and ovaries (39), whereas the PA projection gives a higher dose to posterior structures, including the marrow of the spine. Radiographs used to monitor a scoliosis deformity can involve a cumulative skin entry dose of 2–3 rems per year during the growth years, or $\sim$10 times the annual background dose; however, most of the body receives a much lower dose. Radiation exposure can be minimized by such techniques as the Ardran method (40), by using experienced technicians to minimize retakes, by proper collimation, and by proper monitoring of X-ray tube performance and filtration. Also, digital radiography with highly efficient X-ray detector arrays hold promise for further dose reductions relative to conventional passive systems with photographic films and intensifying screens.

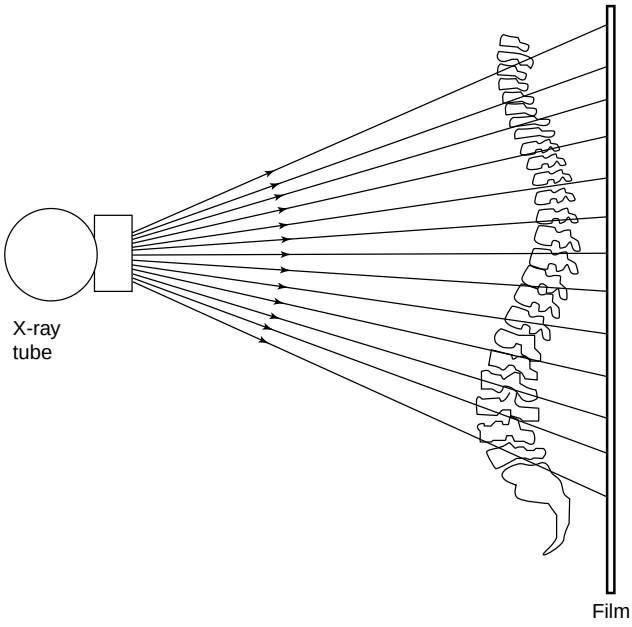


Figure 5. The distortion inherent in an X-ray image, due to the divergence of the X-ray beam relative to the object in the image, and the flat film plane.

Both PA and AP radiographs give a magnified distorted image because of divergence of the X-ray beam (Fig. 5). DeSmet et al. (41) compared the measurements of the Cobb angle using both AP and PA projections. There were small differences ($\sim 2^\circ$) in the lordotic lumbar region and kyphotic thoracic region. In the thoracolumbar region where the spine is more nearly vertical and the central X-ray beam is more perpendicular to the film, there was no difference between measurements in the PA and AP projections.

The frontal plane projections may also be used to estimate axial rotation of the vertebras, based on the positions of the pedicles relative to the apparent center of the vertebral body (42). However, the exact relationship between rotation and this pedicle offset depends on the shape of vertebras, which is a factor that can be taken into account by using statistical data for vertebral shape (43,44).

Multiview Radiography and 3D Reconstruction

In idiopathic scoliosis the standard frontal and lateral radiographs do not supply all the needed information to understand the 3D aspects of the deformities (e.g., intrinsic vertebral deformities, intervertebral disk wedging, spine torsion, rib cage and pelvis deformation, etc.). Computerized tomography and MRI can be used for 3D imaging, but these modalities have limitations. The image acquisition is normally done in the recumbent position (therefore not

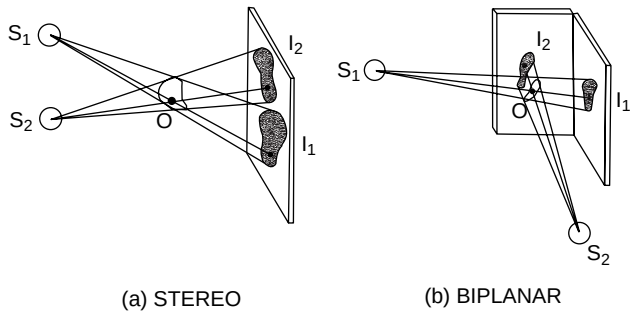


Figure 6. Principles of stereo and biplanar radiogaphy. Stereophotogrammetry principles are used to identify the 3-D positions of object points. The location of the object points can be visualized as the intersections of lines joining the X-ray sources to the corresponding image points.

documenting the standing, functional position). The CT scans can require high X-ray dose of radiation, and both CT and MRI may be incompatible postoperatively with surgical metallic implants. The cost and the image acquisition time are considerable.

Stereo or biplanar techniques (with PA and lateral views, or PA and oblique views) permit 3D reconstruction of selected spinal landmarks using photogrammetric principles (Fig. 6). In the reconstruction, the position of a selected anatomical point visible in both views can be calculated in 3D space (45), using an algorithm such as direct linear transformation (DLT) (46). Dansereau and Stokes (46,47) developed an iterative method for line reconstruction that they applied to the 3D reconstruction of the rib cage based on the images of ribs as seen in PA and oblique radiographs. The average error of these reconstruction methods is ~ 3 mm, depending on the nature of the landmark (some are more precisely identified than others)(48,49).

These 3D reconstruction methods require calibration to take into account the relative positions of the X-ray sources and film planes. A calibration object with object points in known positions is normally employed. This process was substantially facilitated by the DLT method (46,47), and a recently described self-calibration algorithm (50) relies on point-correspondences between two views of the same scene, without using a calibration object. *A priori* knowledge of typically shaped vertebras and the pelvis can be used to enhance the anatomical detail, employing a database of geometrical templates in an atlas derived from statistical anthropometrical data from the literature (51,52).

Tomography

Computerized tomography provides several measures of vertebral and thoracic rotation in the transverse plane (53). However, this method has technical difficulties because of the need to select the orientation of the image plane correctly: usually a global transverse plane is used that is not coincident with the local or regional planes in the deformed spine. While this technique can obtain rotation both at the vertebral level and in the trunk cross-section, it is an expensive procedure, and the radiation dose can be high. It may be used in planning complex surgical procedures for spinal reconstruction after trauma or in cases of severe congenital spinal abnormalities.

Back Surface Topography

The topographical reconstruction of the human trunk has been developed in an attempt to improve documentation of the scoliotic deformity, to eliminate radiation exposure, and to document the external cosmetic aspects of the deformity. These methods may have potential clinical application in detection of scoliosis progression in serial back shape recordings using an automated surface shape measurement system (54). These noncontacting, optical techniques include Moiré fringe photography (55), laser scanning (56,57), and Raster photography (9,58). The principle behind the Moiré fringe technique is shown in diagrammatic form in Fig. 7. The interference patterns, or fringes, represent the surface height or topography of the surface upon which a least one set of lines is projected. Raster techniques involve projection of a structured light pattern (usually lines or dots) on to the body surface, and viewing from an oblique direction (Fig. 8). With suitable calibration, the distorted pattern of the projected lines or points can be converted into the three dimensional coordinates of surface points.

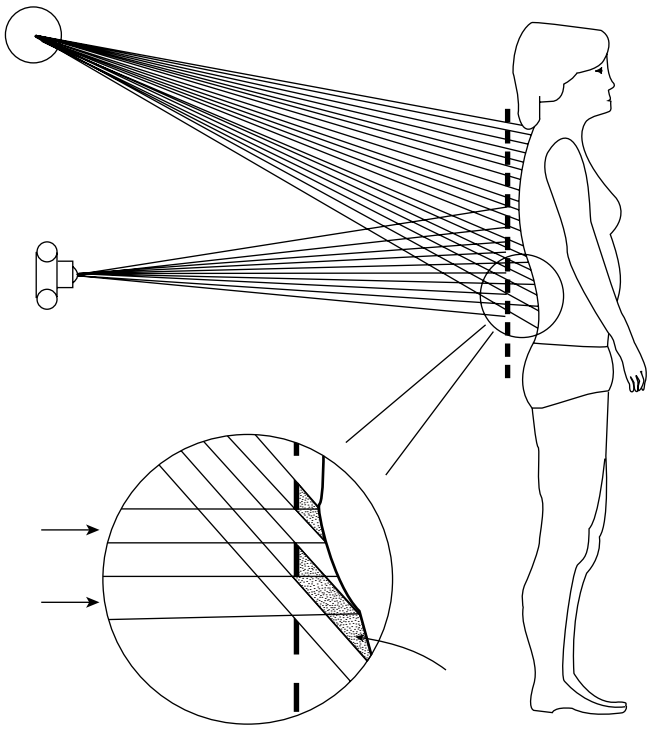


Figure 7. Principles of Moiré fringe photography for recording the shape of the back surface. The surface is viewed through a grid of parallel lines, through which also the surface is illuminated from a different angle. The Moiré fringe pattern results from the interference between the curved line shadows, viewed through the straight line grid.

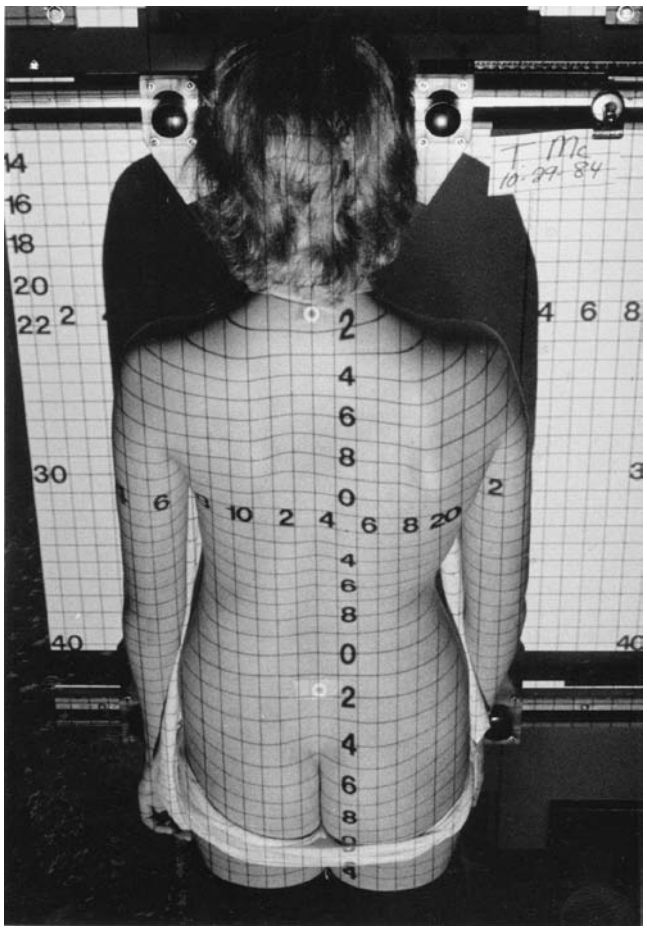


Figure 8. Example of raster stereophotograph of the back surface of a person with scoliosis. The back surface is illuminated by a square grating pattern of lines. These lines appear curved when viewed from a different angle. This system can be calibrated (by a priori knowledge of the geometry, or empirically by measuring the distortion of the grid when projected on to a surface of known shape. This measurement technique is suitable for automatic measurement, for example, by computer analysis of a digital image.

Recent developments in surface topography include the 360° reconstruction of the trunk to analyze the overall external asymmetry with laser scanning or range sensing (56,57,59) (Fig. 9), and its combination with a neural network approach to identifying the optimal relationship between body surface shape and spinal asymmetry.

THEORIES OF IDIOPATHIC SCOLIOSIS ETIOLOGY

The exact biomechanical contributions to etiology and progression of idiopathic scoliosis deformities are not well understood. Scoliosis consists of a lateral spinal curvature, together with transverse plane rotations of the spine and rib cage, and any complete theory concerning its etiology should be able to explain the originals of this complex geometry. In the absence of clear mechanism explaining the etiology, it is considered that the causation is probably multifactorial (59,60).

Biomechanical Aspects

Several biomechanical factors have been invoked to explain the biomechanics of scoliosis. These include intervertebral motion coupling, spinal tethering and buckling, and mechanical influences on growth. While in some forms of scoliosis, the initiating causes are relatively clear (e.g., muscle imbalance or weakness in neuromuscular scoliosis), the etiology of idiopathic scoliosis remains obscure. Somerville (62) emphasized the importance of sagittal plane curvatures and demonstrated with an articulated model how a spine with tight posterior ligaments, which is then forced into flexion will become unstable and develop a rotated, scoliotic deformity.

Coupling. Note that both lateral bending of the spine and scoliotic deformities involve transverse plane rotation, but it appears unlikely that coupling of rotational motion in intervertebral segments controls the development of vertebral rotation in scoliosis. The normal kinematic relationships between lateral bending and axial rotation produce a different spinal shape than that seen in scoliosis (16,63,64). In particular these changes do not explain the pattern of deformity that develops in scoliosis with maximal vertebral rotation at the curve apex. Spinal tethering by posterior structures of the spine (62) has also been invoked to explain the spinal shape in scoliosis and its etiology. There are several parts to this theory. Tethering is thought to prevent flexion of the spine and lead to a hypokyphotic or lordotic shape, which then has a greater tendency to instability. Secondly, the tether is thought to maintain a straighter alignment of the posterior elements than of the vertebral bodies, thus the anterior part of the spine (the vertebral bodies) become more laterally deviated, producing the rotation of the vertebras.

Buckling. The fact that sagittal plane curvature of the spine is flattest in the early teen years supports the idea that this shape places the spine at risk for development of scoliosis. The shape of the spine in scoliosis is reminiscent of a buckled beam, but buckling may not explain the development of lateral curvatures since buckling of the ligamentous spine first exaggerates the sagittal curvatures (64).

Nonbiomechanical Factors

Despite several recent promising developments concerning genetic (65–67) and systemic (68,69) anomalies in patients with AIS, the origin of the initial spinal asymmetry remains unknown. While the regulation of growth and development produces random distributions of right–left asymmetry that are commonly centered around a mean of perfect symmetry (called fluctuating asymmetry), the spine is apparently subject to a higher tendency to spontaneous development of right convex thoracic asymmetry, possibly linked to developmental or mechanical instability. There is some evidence of subtle disturbance

Figure 9. Range sensing topographic scanner (two of the six cameras that surrounds the subject are shown on left panel) and biplanar X-ray system (middle panel) for concurrent surface and spine imaging, showing typical superimposed 3D torso spine model (right panel).

of systemic growth factors (70), abnormal collagen synthesis (71), and abnormalities of muscles and neuromuscular control (72). Recently, Moreau et al. (73) reported that bone cells of patients having surgery for idiopathic scoliosis had an abnormal response to melatonin, implicating a dysfunction melatonin signaling in these patients. However, all these empirical findings have not yet been incorporated into a coordinated theory of the cause of idiopathic scoliosis.

BIOMECHANICS OF SCOLIOSIS PROGRESSION DURING GROWTH

Scoliosis progression is associated with the rate of skeletal growth (74,75). In the progression of a small deformity to a large one, it has been proposed that a small lateral curvature of the spine would load the vertebras asymmetrically (76), leading to asymmetrical growth in the vertebral growth plates. This acceleration of the deformity associated with mechanically modulated growth has been termed a vicious cycle (77) starting after a certain threshold of spinal deformity has been reached. This theory of spinal growth sensitivity to loading asymmetry must, however, explain why the normal spinal curvatures (kyphosis and lordosis) in the sagittal plane do not progress into hyperkyphosis and hyperlordosis by the same mechanism, although progressive deformity in Scheuermann's kyphosis has been attributed to a similar mechanism (78). This concept of a biomechanical mechanism of progression of deformity is attractive intuitively, and has been incorporated into the rationale for brace and surgical treatments (79). However, it cannot be quantified without better knowledge of the normal spinal loading and the alteration of loading in scoliosis, and better understanding of the sensitivity of growth to the time course of mechanical load and its magnitude. It is known that bone growth, including that of vertebras, is modulated by sustained altered stress (increased stress slows growth). However, the mechanism of growth of intervertebral disks, and its possible modulation by altered mechanical conditions is unknown (80). Notably, there does not appear to be any sudden change in disk matrix synthesis at the time of skeletal maturity (81), although scoliosis progression slows considerably after cessation of growth (82).

Scoliosis progression by alteration of bone growth has been investigated both experimentally (83) and analytically. By using heuristic (84) or empirical (85) estimates of spinal loading and growth plate response to altered stress it has been possible to demonstrate the plausibility of the vicious cycle theory of mechanically modulated vertebral growth in progressive scoliosis. In these incremental analyses the asymmetrical forces acting on vertebrae are estimated, and a small increment of spinal growth is applied to the vertebras, modulated by the applied load. Then the spinal geometry is updated, loading asymmetries recalculated, and the process is repeated until the simulated adolescent growth is achieved. Results are sensitive to numerous assumptions about the initial geometry, the prevailing loading of the spine, as well as the extent to which growth is modulated by chronically altered stress.

BIOMECHANICS OF CONSERVATIVE TREATMENT

Treatment for patients with scoliosis aims to reduce not only the lateral deviation of the axial skeleton, but also the cosmetic deformity that is associated with the axial torsion of the trunk. All the common treatment methods are mechanical in nature, yet the mechanical aspects of the treatment have only recently been studied. The magnitude, area of application and balance of the forces, and their

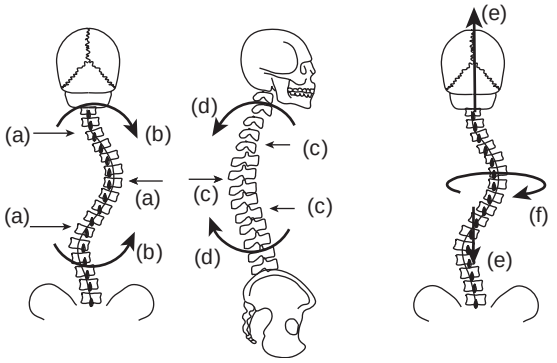


Figure 10. Possible correction force components applied by surgical instrumentation to the spine. Lateral force (a); coronal moment (b); sagittal force (c); sagittal moment (d); distraction force (e); axial moment (f).

moments about the spine are all important to the end result. The types of force applied to the spine for correction are shown in Fig. 10. These forces may either come from an external passive system (i.e., brace) or from an internal active mechanism (i.e. muscles control to shift the trunk away from the pressure areas of the brace).

Braces, Exercises, Electrical Stimulation

A brace is an orthosis that is used to prevent an increase in, or even to reduce, the curvature while waiting for the patient to reach skeletal maturity. The Boston brace, introduced in 1971 by Hall et al. (86) is widely used to treat moderate scoliosis. It has a plastic body envelope with pressure cushions adjusted in the back with straps. Its predecessor, the Milwaukee brace, is made of a neck ring extended by two metal bars and corrective pads. Braces are tailored specifically for each patient and are adjusted by the orthotist. Appropriate positions of the brace's pressure pads are determined from radiographs and clinical examination. Braces use a three- or four-point pressure principle (87), with pads placed generally over the posterolateral part of those ribs that connect to the apex of the thoracic spine deformity, and in the lumbar region, at the level of the scoliosis apex. The strap tension is determined empirically by the orthotist or the orthopedic surgeon (88).

Braces can improve the natural history of scoliosis according to case studies (87), a meta- analysis of published literature (89) and multicenter randomized and controlled studies (90). However, in 81% of the cases the correction is partial or the brace is not sufficient to stop the progression, which cast some doubt on the usefulness and efficacy of braces (91). Clinical studies may be biased by including patients having scoliosis curves that are not likely to progress, with the brace thereby erroneously deemed as effective. Also, it is apparent that many patients do not wear the brace as much as is prescribed (poor compliance), which may reduce the documented effectiveness of bracing. Physical exercises can also be recommended to strengthen back muscles and to improve the spine's stability as well as to reduce back pain. However, these exercises have not been proven to correct nor prevent the progression of scoliosis deformity and should be used along with a more effective treatment like bracing.

Electrical stimulation of the back muscles has been used in the 1970s for the conservative treatment of scoliosis, but has not been found to be effective (90).

Biomechanical Evaluation of Brace Function

The mode of action of a brace is indirect, since it does not apply forces directly to the spine (92). Many factors influence brace efficiency, including the flexibility of the spine, the shape and stiffness of the brace shell, the location, size and thickness of the brace pads (or of the voids inside the brace), the strap tension adjustment, the biomechanical properties of trunk and thorax tissues, and the duration of the brace forces (93,94).

Analyses made using 3D reconstruction of the spine and rib cage of patients with scoliosis who were treated with the Boston brace have revealed that the brace produced immediate significant curve correction of the spinal deformity in the frontal plane at the expense of a significant reduction of thoracic kyphosis in the sagittal plane and without significant effect on the rib hump, vertebral rotation or frontal balance (95). Coupling mechanisms between the applied forces on the rib cage and the coronal deformity of the spine may explain partially the lack of effect in these planes (32).

The forces generated by braces have been evaluated by measurement of pressure distribution between the brace and the patient's torso (Fig. 11) (94,96) High force zones (20–113 N) are mostly located on both sides of the pelvis, on the lower anterior part of the abdomen, on the posterolateral part of the right thoracic zone, and on the lateral part of the left lumbar zone. However, in some cases unfavorable forces were measured on the left thoracic or on the right lumbar parts of the torso, which can explain some of the negative results of bracing. Correction of curves was not solely dependent on the level of force applied by the brace (96) and some patients with the greatest curves achieved little correction despite significant levels of applied force. Jiang (68) found that although high strap forces are necessary to ensure lateral and derotational forces on the spine, they also cause undesirable forces that induce lordosis. Large variability (8–81 N) was measured in the prescribed tension in the thoracic and pelvic straps of the brace (93), and significant relaxation of strap tension also was found a few minutes after adjustments had been made.

Computer Modeling of Bracing

Biomechanical finite element models simulating the orthotic treatment have been developed to analyze brace biomechanics. In the analyses, forces representing brace pads were applied to models representing specific patients' trunk geometry. The traction force applied by the mandible support of the Milwaukee brace was found to be of relatively minor importance in the correction of scoliotic curves, whereas the lumbar pad significantly affects correction in a lumbar curve and thoracic pad often completely dominates the nature of the correction produced (98). The best correction of scoliotic curves could be obtained using

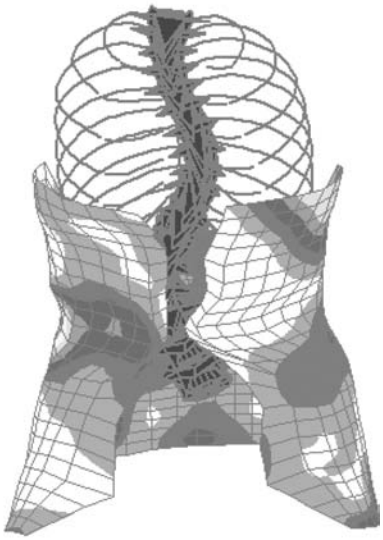


Figure 11. Analysis of brace effectiveness using a flexible matrix of 192 pressure sensors (at the brace–torso interface) and force cells (to measure strap tension). The right side panel shows a typical pressure distribution generated by a brace that is mapped onto a surfacic model of the brace envelope, and superimposed over the three-dimensional reconstruction of the spine and rib cage.

a thoracic pad without lumbar or subaxillary counterpad for thoracic curves, and a lumbar pad with a thoracic counterpad for lumbar curves (99). Application of passive primary forces on the convex side and counter forces on both thoracic and lumbar concave sides of a thoracic scoliotic curve had a substantial corrective effect on the Cobb angle and lateral alignment, as well as do active muscle forces (100). The Boston brace system also was modeled and the brace mechanisms investigated (96,101). An optimization approach of the Boston brace treatment of scoliosis showed that the optimal (most effective) brace forces were mostly located on the convex side of the spinal curves (102). However, the optimal configuration only achieved overall correction of 50% on average.

In previous models of bracing, the action of the brace was represented as force generators instead of passive deformable systems that interact with the flexible torso. Also the complex mechanical action of the brace on the entire torso was not completely addressed. Current modeling efforts are oriented toward the detailed explicit representation of the brace–torso interface (103) (Fig. 12) to improve their simulation of the complete brace system's interaction with the patient, and to optimize the brace design parameters.

BIOMECHANICS OF SURGICAL TREATMENT

In the case of severe spinal deformities (Cobb angle $> \sim 50^{\circ}$), surgical spinal fusion with metallic implants is often performed. The goals of surgical treatment are to prevent further worsening in the scoliosis, and if possible to straighten the spine. This is normally achieved through arthrodesis (bony fusion) of a region of the spine that spans one or more scoliosis curves. Implants used in surgical treatment of spinal deformities have two complementary mechanical roles. The first is to apply forces to correct or reduce the spinal deformity intraoperatively and to maintain correction subsequently. The second is to create the correct mechanical environment for spinal fusion to occur by immobilizing the spine until bony fusion has occurred.

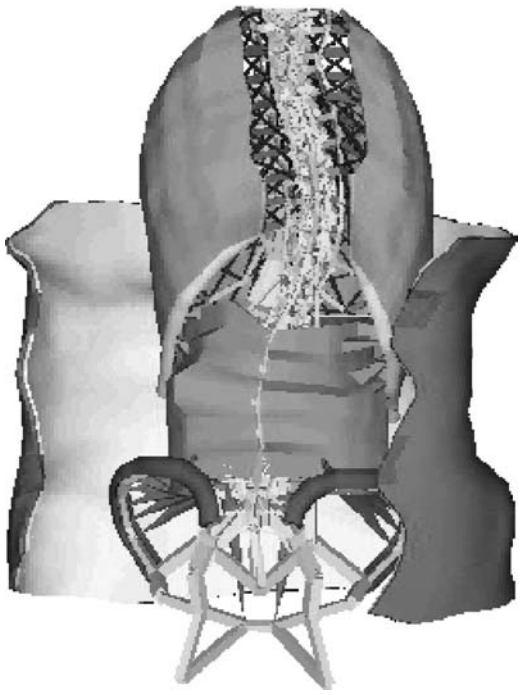


Figure 12. detailed finite element modeling of the brace–torso interface (pelvis not shown). The opened brace is first introduced onto the spine and rib cage model. When the brace is closing (when the opening forces are released), contacts are established at the brace-torso interface and reaction forces are generated. Strap tightening is further generating additional forces.

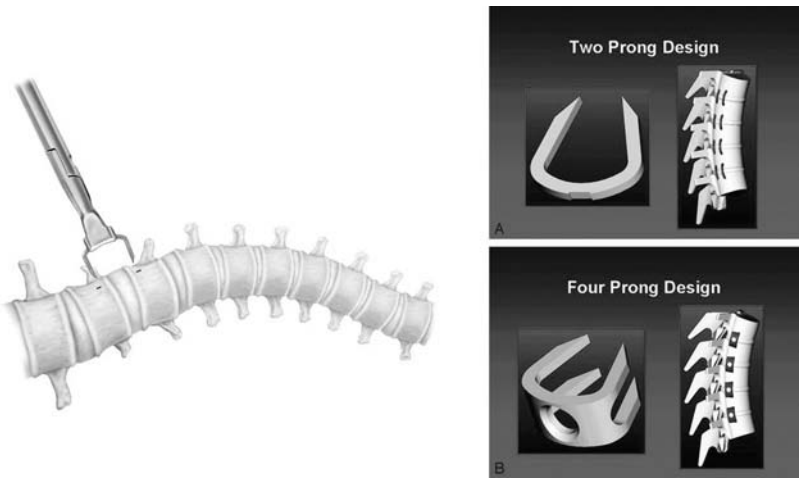


Figure 13. Stapling of intervertebral disk interspace. (Reproduced from Ref. 109, with permission.)

Types of Spinal Instrumentation and Surgical Planning

The Harrington instrumentation system was introduced in the 1960s, with many subsequent advances, including segmental instrumentation (attached to numerous vertebras) and anterior spinal instrumentation systems (attached to the front of the spine). The main components of the Harrington instrumentation are hooks, rod with ratchets and cross members. An outrigger could be used to provide some predistraction and stretching of ligaments before the actual distraction rod is implanted on the concave side of the spine deformity. A second rod in slight compression can be installed on the convex side of the spine curve. The main disadvantages of this instrumentation are its limited number of attachments to the spine; consequently, it mainly corrected the deformity in the frontal plane and occasionally resulted in back flattening (104). In the mid-1981s, the 3D nature of scoliosis deformity was becoming more widely recognized and surgical systems were developed that addressed correction of the deformity in 3D. Luque (105) advanced the use of wires passed around the neural arch of vertebras (sublaminar wiring) to provide multiple connections between a profiled (curved) rod and the vertebras. The first 3D system was developed by Dubousset and Cotrel (106). It had a curved rod system that was rotated during surgery. A precontoured rod, to account for the deformity and a final kyphosis, is loosely attached to the hooks and then rotated $\sim 90^\circ$ and tightened. The belief is that the deformity in the frontal plane is rotated 90° to provide a natural kyphosis in the sagittal plane. Subsequent developments included pedicular screws that provide stronger anchorage to the vertebrae than simple hooks. With screws, direct derotation of the apical vertebra can be applied before securing the screws on the rods. Instead of additional hooks, sublaminar wires may be used to attach the spine to the rod.

Surgical release of anterior structures and fusion of adjacent vertebral bodies are sometimes necessary in conjunction with posterior fusion to prevent the crankshaft phenomenon. This occurs due to the continued anterior growth in skeletally immature patients. A completely anterior approach and fusion to correct scoliosis was first introduced by Dwyer and Schafer (107) and was popularized with the Zielke instrumentation in the 1981s (108). Anterior fusion is considered controversial (e.g., because of risks to anterior structures including nerves and major blood vessels), so posterior instrumentation with fusion is still considered the gold standard.

Staples applied between adjacent vertebras are a promising minimal invasive technology to correct spinal deformities (109) (Fig. 13). Compression staples are used on the convex side of a spine curve to reduce growth while distraction staples are placed on the concave side to accelerate growth.

Testing of Spinal Instrumentation (Construct Design)

Much of the understanding about results of spinal instrumentation and fusion is empirical, based on follow-up studies. Biomechanical principles should be able to provide additional information. The use of surgical instrumentation has a direct mechanical effect on the fused part of the spine, and an indirect effect on the unfused region. It remains as a challenge to biomechanics to further our understanding of the interactions between the spine, the instrumentation, and the muscular and other forces. The muscles, and their control presents the greatest difficulties, since CNS (central nervous system) control of trunk balance is so poorly understood.

There are three basic types of spinal implant construct testing: tests of individual implant components (hooks, screws, rods etc.), tests of the connections between an implant, and the spine and tests of complete instrumentation assemblies. All can be tested in simulations of *in vivo* conditions to evaluate their performance. Laboratory testing of a spinal construct (consisting of a test instrumentation applied to a standardized spinal specimen) can provide information about its flexibility, strength, and fatigue life. Components, such as pedicle screws, may be tested in pullout (110) or bending. The challenge in all these tests is to establish conditions that are representative of the in vivo situation. Both strength and stiffness are desirable properties of instrumentation, and fatigue testing is important to establish whether instrumentation will mechanically fail

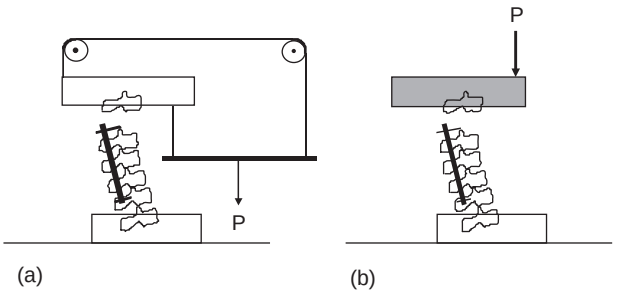


Figure 14. Mechanical test of spinal instrumentation construct, (a) with a pure moment imposing flexion-extension by means of a couple (equal parallel forces). (b) with offset compression loading.

or the conditions will lead to pseudoarthroses (failure to fuse this spine segment completely) with repetitive loading over a period of time.

In a typical construct test, instrumentation is applied to a spine specimen, which is then loaded. The resulting deformations are measured. Alternatively (but rarely), a construct can be tested by applying a displacement, and measuring the necessary forces: a flexibility test, which is the inverse of the stiffness test. In a typical stiffness test, a pure moment is applied via a couple consisting of two equal parallel forces. Alternatively, offset compression loading (Fig. 14) is intended to be more physiologically realistic by compressing the spine axially at the same time as creating a bending moment. Most reports then give the resulting rotations about the same axis as the applied torque, or collinear with the applied load. If additional rotations or displacements about other axes are recorded, these provide information about the coupled (as opposed to direct) stiffnesses. The motion is typically recorded by mechanical or opto-electronic methods, or by magnetic sensors of position and orientation.

Some information about *in vivo* loading of internal fixators has been provided by means of telemetry of signals from force transducers built into the instrumentation (111). These show forces on the order of 250 N and moments on the order of 6 Nm transmitted through the instrumentation during walking, but this constitutes an unknown proportion of the total load carried by the spine and the instrumentation.

Many instrumentation tests are performed with animal spines because human spines representative of the target population (of people requiring implants) are hard to obtain. Furthermore, human spinal specimens are more variable in their mechanical properties than animal spines of similar ages, taken from animals of the same breed. Bovine calf and pig spines have been shown to be a reasonable choice based on similar vertebral dimensions and flexibility properties.

Standardized instrumentation testing protocols (112–114) should be advantageous to facilitate comparisons between tests performed in different labs. However, standardization is difficult. In addition to defining the exact conditions for the tests, standardization has to take into account the difficulty of obtaining suitable human cadaveric material or the appropriateness of using animal spines. Standardization has been elusive, and to date only the testing of instrumentation systems attached to plastic blocks (corpectomy model) has been provisionally standardized by the ASTM (111).

It remains unclear how to interpret the information in reports of stiffness and strength testing of spinal implants, in terms of how to select an instrumentation system and adapt it to the individual needs of a particular patient. It appears that stiffer instrumentation is better, providing the instrumentation size is not excessive. Given that there are six degrees of freedom possible in a stiffness test, it is not known which components of rotation and translational stiffness are the most important. An important aspect of the instrumentation is how easy it is to apply, with minimum blood loss and risk of injury, for example, to neural and vascular structures. Also, the ability to achieve the desired realignment of the spinal column is important. Biomechanical information can be used in conjunction with other information in making surgical decisions, and biomechanical testing ought to be used with the goal of reducing the need for empiricism in the development of knowledge about the outcomes of different surgical strategies.

Analytical Simulation of Surgical Maneuvers

Segmental instrumentation offers surgeons many variables and multistep maneuvers to adapt to individual patients' needs. Biomechanical modeling offers the possibility to explore these options in advance, and to assist in decision making. In theory, if the mechanical properties of both the spine and the instrumentation were understood completely, then biomechanical models could be built and be used to predict the outcome of surgery. However, the complexity of the surgical choices creates many unknown inputs for the biomechanical analyses, and difficulties in validation of model predictions. To simulate the procedures of the Cotrel–Dubousset instrumentation (Medtronic Sofamor-Danek, Memphis, TN), for example, the surgical maneuvers for the concave-side rod of a segmental instrumentation can be represented as four steps: (*1*) install the rod passively to the end hooks, then approximate the intermediate hooks to the rod; (*2*) displace the hooks along the rod to their final positions (hook distraction); (*3*) rotate the concave-side rod; (*4*) lock the hooks to the rod and relax the applied torque (spring-back). It has proven very difficult to quantify the required magnitude and direction of all these displacement inputs in the simulation. Another difficulty in using a mechanical deterministic model to predict the outcome of surgery for an individual patient is the unknown flexibility of each motion segment and of the rib cage.

Another possibility offered by analytical modeling is the ability to calculate stresses at selected sites in the spine and instrumentation. Finite element analyses were used to estimate stresses in internal fixation devices (115,116). These models have provided estimates of the change in bone stress for different vertebral injury situations, with inclusion of instrumentation components such as plastic washers in the construct, and after incorporation of bone graft. Biomechanical analyses have also been used to investigate the consequences of surgical variables such

as angulation of pedicle screws on the rigidity of the construct (110,117).

Most biomechanical models simulating surgical procedures have employed finite element models. A finite element model of the spine and rib cage (118) was adapted to investigate the biomechanics of Harrington instrumentation (119). The hooks were connected to the instrumented vertebrae (with rotation allowed along vertical axis), while the rod distraction maneuver was modeled as a thermal expansion of the rod to reach its final shape. The biomechanics of Cotrel–Dubousset instrumentation also was studied with this model, in an idealized geometry (120). This study was extended by simulating the surgical maneuvers on fifteen surgical cases (121,122). Each model's geometry was personalized to each patient's 3D skeletal geometry, built from preoperative stereo X rays, and results were compared with documented post-surgical geometry. The simulations of surgical maneuvers showed generally good agreement (on average) with measured effects of surgery in the frontal plane. However, in the sagittal and transverse planes, the response was dependent on model's boundary conditions, motion segments' mechanical properties and instrumentation variables.

A kinematic model including flexible elements to represent each motion segment and kinematic joints and sets to model the instrumentation (flexible multibody approach) was developed by Aubin et al. (123,124). This was intended to overcome limitations of finite element structural modeling that occur when there are large differences in the structure's stiffness (the surgical construct has properties of a mechanism as well as of a structure). Fig. 15 shows preliminary results of this modeling approach for 1 of 20 cases. In the simulations, geometric indices such as Cobb angles showed trends similar to those observed during surgery: gradual correction during surgical procedures and partial loss of correction after hooks/screws lock-up, when the strains induced by the correction maneuvers and stored in the noninstrumented part of the spine were released. To complement this kind of surgical simulation, a spine surgery simulator and a virtual reality training simulator (Fig. 16) currently are under development (125).

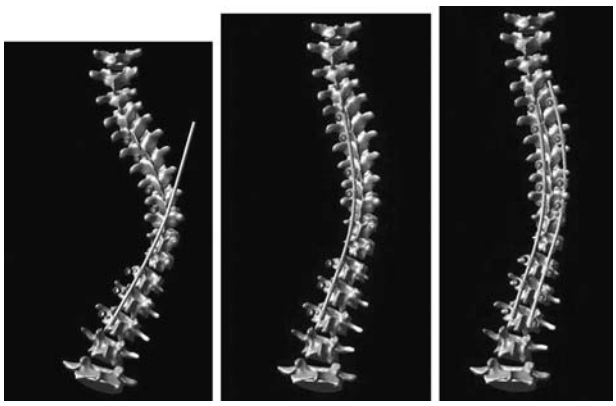


Figure 15. Biomechanical simulation of a spinal instrumentation procedure. The left and central panels show the attachment of the first rod, while the right panel shows the installation of the second rod.

Figure 16. Virtual reality spine surgery simulator. The left panel shows one part of the virtual surgery room with the operating table, the patient and the exposed spine, as well as the radiographs and a magnification of the selected vertebra. The right panel shows an operator that is performing a virtual reality surgery (using stereo stereoscopic glasses and a 3D mouse).

Once validated, these combined tools will allow surgeons to test various surgical options before the real surgery and determine the best scenario for each patient's spinal deformity.

Simulation of the mechanical action of various designs of costoplasties (operations to change the shape of the ribs) by Grealou et al. (33) and Carrier et al. (34) is another recent development. They analyzed different procedures (side, location, length, and number of ribs to resect or to graft), to investigate the biomechanical effect on the spinal and thoracic shape, and to test the idea that spinal deformity can be treated by operations on the rib cage alone.

CONCLUSION

Scoliosis is a costly problem in terms of both healthcare resources and human suffering. Many cases have unknown etiology, and there may be a genetic component in their causation that is modified by environmental factors.

Mechanical analyses of scoliosis as a buckling phenomenon have not been very illuminating, possibly because they are single plane models and do not incorporate rotation or other coupling. The interactions between biomechanical factors and spinal stability during growth are probably important in most scoliosis deformities, irrespective of their initial cause. Biomechanical factors including motion segment coupling, spinal length and slenderness, muscle asymmetry, and postural sway, are apparently relevant in the development and progression of these deformities.

Geometrical aspects of scoliosis can be better documented and understood with 3D measurement including low dose 3D X-ray techniques. These provide a more complete picture of the deformity than the relatively inaccurate single plane Cobb angle measurement. Noninvasive surface measurement techniques show promise, but there is still a need to establish their relationship to radiographic measurements, especially during progression of a curve.

Although many experimental and biomechanical models of scoliosis have been developed, the mechanisms behind the spinal deformity process and its treatment

are not yet clear. Developing a better understanding of the mechanical aspects of the deforming scoliotic spine will lead to a more selective and appropriate means of treatment, especially if the mechanisms that produce the deformity are confirmed to be reversible. Certainly, the empirical evidence from successful brace or stapling treatments suggests that this is the case. Early detection and treatment by mechanical methods show promise for prevention of progression of the deformity.

BIBLIOGRAPHY

1. Danielsson AJ, Nachemson AL. Back pain and function 22 years after brace treatment for adolescent idiopathic scoliosis: A case-control study-part I. Spine 2003;28:2079–2085.
2. Danielsson AJ, Nachemson AL. Childbearing, curve progression, and sexual function in women 22 years after treatment for adolescent idiopathic scoliosis: A case-control study. Spine 2001;26:1449–1456.
3. Noonan KJ, Dolan LA, Jacobson WC, Weinstein SL. Long-term psychosocial characteristics of patients treated for idiopathic scoliosis. J Pediatr Orthop 1997;17:712–717.
4. Scoliosis Research Society. Three-dimensional terminology of Spinal Deformity. Available at http://www.srs.org/professionals/glossary/3-d.asp. Accessed 2005.
5. Armstrong GW, Livermore NB 3rd, Suzuki N, Armstrong JG. Nonstandard vertebral rotation in scoliosis screening patients. Its prevalence and relation to the clinical deformity. Spine 1982;7:50–54.
6. Parent S, et al. Morphometric analysis of anatomic scoliotic specimens. Spine 2002;27:2305–2311.
7. Thometz JG, Lamdan R, Liu XC, Lyon R. Relationship between Quantec measurement and Cobb angle in patients with idiopathic scoliosis. J Pediatr Orthop 2000;20:512–516.
8. Stokes IA. Axial rotation component of thoracic scoliosis. J Orthop Res 1989;7:702–708.
9. Stokes IA, Armstrong JG, Moreland MS. Spinal deformity and back surface asymmetry in idiopathic scoliosis. J Orthop Res 1988;6:129–137.
10. Acaroglu E, et al. Does transverse apex coincide with coronal apex levels (regional or global) in adolescent idiopathic scoliosis? Spine 2001;26:1143–1146.
11. Thulbourne T, Gillespie R. The rib hump in idiopathic scoliosis. Measurement, analysis and response to treatment. J Bone Joint Surg Br 1976;58:64–71.
12. Pearcy MJ, Whittle MW. Movements of the lumbar spine measured by three-dimensional X-ray analysis. J Biomed Eng 1982;4:107–112.
13. Pope MH, Wilder DG, Matteri RE, Frymoyer JW. Experimental measurements of vertebral motion under load. Orthop Clin N Am 1977;8:155–167.
14. Stokes IA, Wilder DG, Frymoyer JW, Pope MH. 1981 Volvo award in clinical sciences. Assessment of patients with low-back pain by biplanar radiographic measurement of intervertebral motion. Spine 1981;6:233–240.
15. Lovett RW. The mechanics of the normal spine and its relation to scoliosis. Boston Med Surg J 1905; 153.
16. White AA 3rd. Kinematics of the normal spine as related to scoliosis. J Biomech 1971;4:405–411.
17. Stokes IA, Moreland MS. Measurement of the shape of the surface of the back in patients with scoliosis. The standing and forward-bending positions. J Bone Joint Surg Am 1987;69:203–211.
18. Bunnell WP. An objective criterion for scoliosis screening. J Bone Joint Surg Am 1984;66:1381–1387.
19. Scoliosis Research Society Terminology Committee. Revised Glossary of Terms. http://www.srs.org/professionals/glossary/glossary.asp. Accessed 2005.
20. Goldstein LA, Waugh TR. Classification and terminology of scoliosis. Clin Orthop Relat Res 1973;93:10–22.
21. King HA, Moe JH, Bradford DS, Winter RB. The selection of fusion levels in thoracic idiopathic scoliosis. J Bone Joint Surg Am 1983;65:1302–1313.
22. Richards BS, Sucato DJ, Konigsberg DE, Ouellet JA. Comparison of reliability between the Lenke and King classification systems for adolescent idiopathic scoliosis using radiographs that were not premeasured. Spine 2003;28: 1148–1156. discussion 1156–1157.
23. Stokes IA, Aronsson DD. Computer-assisted algorithms improve reliability of King classification and Cobb measurement of scoliosis. Spine (in press).
24. Lenke LG, et al. Curve prevalence of a new classification of operative adolescent idiopathic scoliosis: Does classification correlate with treatment? Spine 2002;27:604–611.
25. Stokes IA. Three-dimensional terminology of spinal deformity. A report presented to the Scoliosis Research Society by the Scoliosis Research Society Working Group on 3-D terminology of spinal deformity. Spine 1994;19:236–248.
26. Peloux J du, Fauchet R, Faucon B, Stagnara P. Le plan délection pour l'examen radiologique des cyphoscolioses Rev Chir Orthop 1965;51:517–524.
27. Raso J, Gillespie R, McNiece G. Determination of the plane of maximum deformity in idiopathic scoliosis. Orthop Trans 1981; 4.
28. Kreyzig E. Advanced Engineering Mathematics 7th ed. and Mathematics Manual to Accompany Advanced Engineering Mathematics 7th ed. New York: John Wiley & Sons Inc; 1995.
29. Poncet P, Dansereau J, Labelle H. Geometric torsion in idiopathic scoliosis: Three-dimensional analysis and proposal for a new classification. Spine 2001;26:2235–2243.
30. Cobb JR. Outline for the Study of Scoliosis. The American Academy of Orthopaedic Surgeons: Instructional Course Lectures 5 48: p 261–275.
31. Gardner-Morse MG, Stokes IA. Structural behavior of human lumbar spinal motion segments. J Biomech 2004;37:205–212.
32. Aubin CE, Dansereau J, de Guise JA, Labelle H. Rib cage-spine coupling patterns involved in brace treatment of adolescent idiopathic scoliosis. Spine 1997;22:639–635.
33. Grealou L, Aubin CE, Labelle H. Rib cage surgery for the treatment of scoliosis: A biomechanical study of correction mechanisms. J Orthop Res 2002;20:1121–1128.
34. Carrier J, Aubin CE, Villemure I, Labelle H. Biomechanical modelling of growth modulation following rib shortening or lengthening in adolescent idiopathic scoliosis. Med Biol Eng Comput 2004;42:541–548.
35. Kramers-de Quervain IA, Muller R, Stacoff A, Grob D, and Stussi E. Gait analysis in patients with idiopathic scoliosis. Eur Spine J 2004;13:449–456.
36. Lenke LG, et al. Prospective dynamic functional evaluation of gait and spinal balance following spinal fusion in adolescent idiopathic scoliosis. Spine 2001;26:E330–E337.
37. Panjabi MM, Brand RA. Jr, White AA 3rd. Three-dimensional flexibility and stiffness properties of the human thoracic. Spine. J Biomech 1976;9:185–192.
38. Janevic J, Ashton-Miller JA, Schultz AB. Large compressive preloads decrease lumbar motion segment flexibility. J Orthop Res 1991;9:228–236.

39. Nash CL. Jr, Gregg EC, Brown RH, Pillai K. Risks of exposure to X-rays in patients undergoing long-term treatment for scoliosis. J Bone Joint Surg Am 1979;61: 371–374.
40. Ardran GM, et al. Assessment of scoliosis in children: low dose radiographic technique. Br J Radiol 1981;53:146–147.
41. DeSmet AA, Goin JE, Asher MA, Scheuch HG. A clinical study of the differences between the scoliotic angles measured on posteroanterior and anteroposterior radiographs. J Bone Joint Surg Am 1982;64:489–493.
42. Perdriolle R, Vidal J. [A study of scoliotic curve. The importance of extension and vertebral rotation (author's transl)]. Rev Chir Orthop Reparatrice Appar Mot 1981;67: 25–34.
43. Drerup B. Principles of measurement of vertebral rotation from frontal projections of the pedicles. J Biomech 1984;17: 923–935.
44. Stokes IA, Bigalow LC, Moreland MS. Measurement of axial rotation of vertebrae in scoliosis. Spine 1986;11:213–218.
45. Brown RH, Burstein AH, Nash CL, Schock CC. Spinal analysis using a three-dimensional radiographic technique. J Biomech 1976;9:355–365.
46. Marzan GT. Rational design for close-range photogrammetry Ph.D. dissertation, University of Illinois, 1976 Xerox University, Microfilms, Ann Arbor MI 1976; p. 46.
47. Dansereau J, Stokes IA. Measurements of the three–dimensional shape of the rib cage. J Biomech 1988;21:893–901.
48. Aubin CE, et al. Morphometric evaluations of personalised 3D reconstructions and geometric models of the human . Spine. Med Biol Eng Comput 1997;35:611–618.
49. Delorme S, Petit Y, de Guise JA, Aubin CÉ, Dansereau J. Assessment of the 3-D reconstruction and high-resolution geometrical modeling of the human skeletal trunk from 2-D radiographic images. IEEE Trans Biomed Eng 2003; 50:989–998.
50. Cheriet F, et al. Towards the self-calibration of a multi-view radiographic imaging system for the 3D reconstruction of the human spine and rib cage. Int J Pattern Recog Arti Intell 1999;13:761–779.
51. Aubin CE, et al. [Geometrical modeling of the spine and the thorax for the biomechanical analysis of scoliotic deformities using the finite element method]. Ann Chir 1995;49:749–761.
52. Dansereau J, Labelle H, Aubin CE. 3-D personalized parametric modelling of reconstructed scoliotic spines. IVth International Symposium on Computer Simulation in Biomechanics. 1.6–1.9 p. 93.
53. Aaro S, Dahlborn M. The longitudinal axis rotation of the apical vertebra, the vertebral, spinal, and rib cage deformity in idiopathic scoliosis studied by computer tomography. Spine 1981;6:567–572.
54. Theologis TN, Fairbank JC, Turner-Smith AR, Pantazopoulos T. Early detection of progression in adolescent idiopathic scoliosis by measurement of changes in back shape with the Integrated Shape Imaging System scanner. Spine 1997;22: 1223–1227; discussion 1228.
55. Stokes IA, Moreland MS. Concordance of back surface asymmetry and spine shape in idiopathic scoliosis. Spine 1989;14:73–78.
56. Jaremko JL, et al. Genetic algorithm-neural network estimation of Cobb angle from torso asymmetry in scoliosis. J Biomech Eng 2002;124:496–503.
57. Jaremko JL, et al. Estimation of spinal deformity in scoliosis from torso surface cross sections. Spine 2001;26(14):1583–1591.
58. Frobin W, Hierholzer E. Analysis of human back shape using surface curvatures. J Biomech 1982;15:379–390.
59. Pazos V, et al. Accuracy assessment of human trunk surface 3D reconstructions from an optical digitising system. Med Biol Eng Comput 2005;43:11–15.
60. Nachemson A, Sahlstrand T. Etiologic factors in adolescent idiopathic scoliosis. Spine 1977;2:176–184.
61. Robin GC. The Aetiology of Scoliosis. A Review of a Century of Research. Boca Raton (FL): CRC Press: 1990.
62. Somerville EW. Rotational lordosis; the development of single curve. J Bone Joint Surg Br 1952;34-B:421–427.
63. Stokes IA, Gardner-Morse M. Analysis of the interaction between vertebral lateral deviation and axial rotation in scoliosis. J Biomech 1991;24:753–759.
64. Veldhuizen AG, Scholten PJ. Kinematics of the scoliotic spine as related to the normal spine. Spine 1987;12:852–858.
65. Axenovich TI, et al. Segregation analysis of idiopathic scoliosis: demonstration of a major gene effect. Am J Med Genet 1999;86:389–394.
66. Bashiardes S, et al. SNTG1, the gene encoding gamma1-syntrophin: A candidate gene for idiopathic scoliosis. Hum Genet 2004;115:81–89.
67. Justice CM, et al. Familial idiopathic scoliosis: Evidence of an X-linked susceptibility locus. Spine 2003;28:589–594.
68. Bagnall KM, et al. Melatonin levels in idiopathic scoliosis. Diurnal and nocturnal serum melatonin levels in girls with adolescent idiopathic scoliosis. Spine 1996;21:1974–1978.
69. Lowe T, et al. Platelet calmodulin levels in adolescent idiopathic scoliosis: Do the levels correlate with curve progression and severity? Spine 2002;27:768–775.
70. Skogland LB, Miller JA. Growth related hormones in idiopathic scoliosis. An endocrine basis for accelerated growth. Acta Orthop Scand 1981;51:779–781.
71. Aigner T. Variation with age in the pattern of type X collagen expression in normal and scoliotic human intervertebral discs. Calcif Tissue Int 1998;63:263–268.
72. Veldhuizen AG, Wever DJ, Webb PJ. The aetiology of idiopathic scoliosis: biomechanical and neuromuscular factors. Eur Spine J 2000;9:178–184.
73. Moreau A, et al. Melatonin signaling dysfunction in adolescent idiopathic scoliosis. Spine 2004;29:1772–1781.
74. Little DG, Song KM, Katz D, Herring JA. Relationship of peak height velocity to other maturity indicators in idiopathic scoliosis in girls. J Bone Joint Surg Am 2000;82:685–693.
75. Lonstein JE, Carlson JM. The prediction of curve progression in untreated idiopathic scoliosis during growth. J Bone Joint Surg Am 1984;66:1061–1071.
76. Stokes IA, Gardner-Morse M. Muscle activation strategies and symmetry of spinal loading in the lumbar spine with scoliosis. Spine 2004;29:2103–2107.
77. Stokes IA, Spence H, Aronsson DD, Kilmer N. Mechanical modulation of vertebral body growth. Implications for scoliosis progression. Spine 1996;21:1162–1167.
78. Scoles PV, et al. Vertebral alterations in Scheuermann's kyphosis. Spine 1991;16:509–515.
79. Roaf R. The treatment of progressive scoliosis by unilateral growth-arrest. J Bone Joint Surg Br 1963;45:637–651.
80. Urban JP, Roberts S. Development and degeneration of the intervertebral discs. Mol Med Today 1995;1:329–335.
81. Antoniou J, et al. Elevated synthetic activity in the convex side of scoliotic intervertebral discs and endplates compared with normal tissues. Spine 2001;26:E198–E206.
82. Weinstein SL, Ponseti IV. Curve progression in idiopathic scoliosis. J Bone Joint Surg Am 1983;65:447–455.

83. Mente PL, Aronsson DD, Stokes IA, Iatridis JC. Mechanical modulation of growth for the correction of vertebral wedge deformities. J Orthop Res 1999;17:518–524.
84. Villemure I, Aubin CE, Dansereau J, Labelle H. Simulation of progressive deformities in adolescent idiopathic scoliosis using a biomechanical model integrating vertebral growth modulation. J Biomech Eng 2002;124:784–790.
85. Stokes IA. Biomechanical spinal growth modulation and progressive adolescent scoliosis—a test of the 'vicious cycle'. Electronic Focus Group report. Eur Spine J (in press).
86. Hall JE, Miller ME, Schumann W, Stanish W. A refined concept in the orthotic treatment management of scoliosis. Orthot Prosthet 1975;4:7–13.
87. Emans JB, et al. The Boston bracing system for idiopathic scoliosis. Follow-up results in 295 patients. Spine 1986;11:792–811.
88. Watts HG. Bracing in spinal deformities. Orthop Clin N Am 1979;10:769–785.
89. Rowe DE, et al. A meta-analysis of the efficacy of nonoperative treatments for idiopathic scoliosis. J Bone Joint Surg Am 1997;79, 664–674.
90. Nachemson AL, Peterson LE. Effectiveness of treatment with a brace in girls who have adolescent idiopathic scoliosis. A prospective, controlled study based on data from the Brace Study of the Scoliosis Research Society. J Bone Joint Sur Am 1995;77:815–822.
91. Edgar MA. To brace or not to brace? J Bone Joint Surg Br 1985;67:173–174.
92. Ogilvie J. Spinal Orthotics. An overview. The pediatric spine: principles and practice 1787–1793. New York, Raven Press Ltd.; 1994.
93. Aubin CE, et al. Variability of strap tension in brace treatment for adolescent idiopathic scoliosis. Spine 1999;24:349–354.
94. Mac-Thiong JM, et al. Biomechanical evaluation of the Boston brace system for the treatment of adolescent idiopathic scoliosis: Relationship between strap tension and brace interface forces. Spine 2004;29:26–32.
95. Labelle H, Dansereau J, Bellefleur C, Poitras B. Three-dimensional effect of the Boston brace on the thoracic spine and rib cage. Spine 1996;21:59–64.
96. Perie D, et al. Boston brace correction in idiopathic scoliosis: A biomechanical study. Spine 2003;28:1672–1677.
97. Chase AP, Bader DL, Houghton GR. The biomechanical effectiveness of the Boston brace in the management of adolescent idiopathic scoliosis.Spine 1989;14:636–642.
98. Andriacchi TP, Schultz AB, Belytschko TB, Dewald R. Milwaukee brace correction of idiopathic scoliosis. A biomechanical analysis and a restrospective study. J Bone Joint Surg Am 1976;58:816–815.
99. Patwardhan AG, et al. A biomechanical analog of curve progression and orthotic stabilization in idiopathic scoliosis. J Biomech 1986;19:103–117.
100. Wynarsky GT, Schultz AB. Optimization of skeletal configuration: studies of scoliosis correction biomechanics. J Biomech 1991;24:721–732.
101. Perie D, et al. Personalized biomechanical simulations of orthotic treatment in idiopathic scoliosis. Clin Biomech (Bristol, Avon) 2004;19:190–195.
102. Gignac D, Aubin CE, Dansereau J, Labelle H. Optimization method for 3D bracing correction of scoliosis using a finite element model. Eur Spine J 2000;9:185–190.
103. Perie D, et al. Biomechanical modelling of orthotic treatment of the scoliotic spine including a detailed representation of the brace-torso interface. Med Biol Eng Comput 2004;42:339–344.
104. Humke T, Grob D, Scheier H, Siegrist H. Cotrel-Dubousset and Harrington Instrumentation in idiopathic scoliosis: a comparison of long–term results. Eur Spine J 1995;4:281–283.
105. Luque ER. The anatomic basis and development of segmental spinal instrumentation. Spine 1982; May–Jun; 7(3):256–259.
106. Dubousset J, Cotrel Y. Application technique of Cotrel-Dubousset instrumentation for scoliosis deformities. Clin Orthop Relat Res 1991; 103–110.
107. Dwyer AF, Schafer MF. Anterior approach to scoliosis. Results of treatment in fifty-one cases. J Bone Joint Surg Br 1974;56:218–224.
108. Metz P, Zielke K. [First results of the Luque operation (author's transl)]. Z Orthop Ihre Grenzgeb 1982;120:333–337.
109. Betz RR, et al. An innovative technique of vertebral body stapling for the treatment of patients with adolescent idiopathic scoliosis: A feasibility, safety, and utility study. Spine 2003;28:S255–S265.
110. Krag MH. Biomechanics of thoracolumbar spinal fixation. A review. Spine 1991;16:S84–S99.
111. Rohlmann A, Bergmann G, Graichen F. Loads on an internal spinal fixation device during walking. J Biomech 1997;30: 41–47.
112. ASTM. Standard test method for PS5-94 static and dynamic spinal implants assembly in a corpectomy model. American Society of Testing and Materials; 1997.
113. Ashman RB, et al. *In vitro* spinal arthrodesis implant mechanical testing protocols. J Spinal Disord 1989;2:274–281.
114. Panjabi MM. Biomechanical evaluation of spinal fixation devices: I. A conceptual framework. Spine 1988;13:1129–1134.
115. Goel VK, Pope MH. Biomechanics of fusion and stabilization. Spine 1995;20:85S–99S.
116. Skalli W, Lavaste F, Robin S, Dubousset J. A biomechanical analysis of short segment spinal fixation using a 3D geometrical and mechanical model. Spine 1993;18:536–545.
117. Ruland CM, et al. Triangulation of pedicular instrumentation. A biomechanical analysis. Spine 1991;16:S270–S276.
118. Stokes IA, Laible JP. Three-dimensional osseo-ligamentous model of the thorax representing initiation of scoliosis by asymmetric growth. J Biomech 1990;23:589–595.
119. Stokes IA, Gardner-Morse M. Three-dimensional simulation of Harrington distraction instrumentation for surgical correction of scoliosis. Spine 1993;18:2457–2464.
120. Gardner-Morse M, Stokes IA. Three-dimensional simulations of the scoliosis derotation maneuver with Cotrel-Dubousset instrumentation. J Biomech 1994;27:177–181.
121. Stokes IA, et al. Biomechanical simulations for planning of scoliosis surgery. Res Spinal Deformities II, IOS Press 1998;59:343–346.
122. Grealou L, Aubin CE, Labelle H. Biomechanical modeling of the CD instrumentation in scoliosis: A study of correction mechanisms. Arch. Physiol Bioche 2000;108:194.
123. Aubin CE, et al. Biomechanical modeling of posterior instrumentation of the scoliotic spine. Comput Methods Biomech Biomed Eng 2003;6:27–32.
124. Poulin F, et al. [Biomechanical modeling of instrumentation for the scoliotic spine using flexible elements: A feasibility study]. Ann Chir 1998;52:761–767.
125. Plouznikoff A, Aubin CÉ, Ozell B, Labelle H. Virtual reality scoliosis surgery simulator. Inter Res Soci Spinal Deformities 2004; 139–142.

See also HUMAN SPINE, BIOMECHANICS OF; REHABILITATION, ORTHOTICS IN; SPINAL IMPLANTS.

SCREEN-FILM SYSTEMS

ZHENG FENG LU
Columbia University
New York, New York

INTRODUCTION

Projection radiography generates the majority of the imaging volume in a typical hospital-based radiology department. Figure 1 illustrates the basic geometry for projection radiography. As shown, an X-ray beam incidents upon and transmits through the patient. Differential attenuation occurs as X-ray photons interact with tissues in the patient. Consequently, an altered X-ray distribution pattern is generated behind the patient. This altered X-ray pattern can be recorded by an image receptor placed underneath the patient.

Film has been utilized as an image receptor since the very beginning of radiographic imaging. The first radiograph of a living hand was exposed by Roentgen in 1895 using a photographic plate (1). Film is sensitive to both X rays and light photons. However, the sensitivity of film to direct X-ray exposure is low. In order to overcome this problem, the fluorescent screen was developed shortly after the discovery of X rays to be combined with the film as an image receptor. Fluorescence, as used in radiography, refers to the phenomenon that the absorption of X rays by crystals of certain inorganic salts (called phosphor) is followed by the emission of light photons (2). The emitted light photons are then utilized more efficiently to expose the film that is sandwiched between the screens (see Fig. 2). These fluorescent screens are also called intensifying screens. Initially, calcium tungstate was the phosphor of choice for screens in radiography (3). Over the years, especially since early 1970s, more efficient phosphor materials have been developed, such as rare earth phosphors (2,5–7). The "rare earth" elements are found in a row of the periodic table of the elements with the atomic numbers 57–71. The most common rare earth phosphor currently available is godalinium oxysulfide (4). Phosphors, other than rare earth types, are used as well, such as yttrium tantalate (5–7).

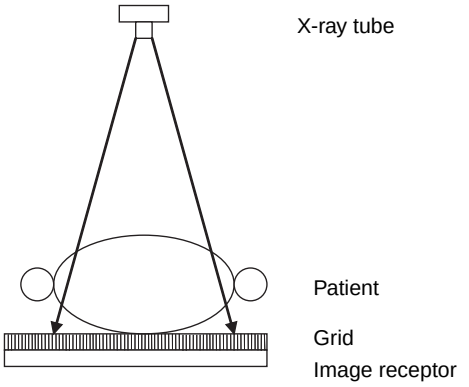


Figure 1. Illustration of the basic geometry for projection radiography.

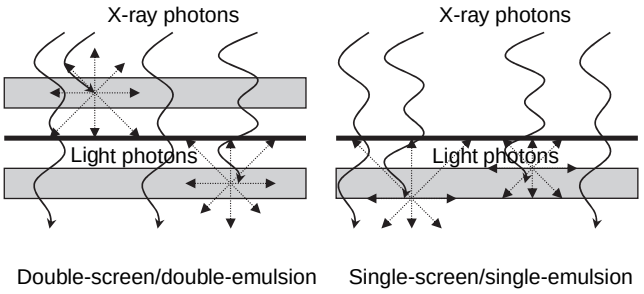


Figure 2. X-ray photons are absorbed by the intensifying screens and the emitted light photons from the screens expose the film to form the image.

Despite numerous improvements made in technology since the discovery of X rays, screen-film systems had remained, for a long period of time, as the main recording image receptors for radiography until recently. Since the 1980s, the essential role of the screen-film system has been challenged by the new invention of computed radiography (CR) and digital radiography (DR) (8). In comparison to screen-film systems, the CR and DR system has a much larger dynamic range. It has been reported that a linear response over more than four orders of magnitude of radiation exposure is achievable (2). This allows CR and DR systems to have a high tolerance for variations in radiation exposures. Therefore, the repeat rate due to underexposure or overexposure is significantly reduced. While the CR and DR systems provide tremendous potentials in image manipulation associated with digital imaging, their current spatial resolution is usually inferior to that of the screen-film systems (9,10). This explains the fact that the last stronghold of the screen-film system is in mammography, where the spatial resolution is on high demand. However, the digital technology continues to evolve, and the results of several large clinical trials of digital versus screen-film mammography suggest that the diagnostic accuracy of the digital mammography "is similar to" that of the screen–film mammography (11). In recent years, more and more radiology departments moved toward "filmless" with picture archiving and communication systems (PACS) (8). It is expected for DR and CR to become more widely available, and even replace screen-film systems completely in the near future (12).

In this article, we will outline the physical and photographic properties of intensifying screens and film. The steps for film processing will be delineated. Performance evaluation of the screen-film system will be discussed in relation to image quality and patient radiation dose. Finally, quality assurance and quality control testing will be described for the applications of screen-film systems in radiology.

INTENSIFYING SCREEN

Since only a small percentage (2–6%) of the X rays can be absorbed by the emulsion of the film, the sensitivity of the film alone is very low (2). With the screen–film combination,

X rays are absorbed more efficiently by the intensifying screen. After absorption, the screen furnishes a light image converted from the X ray image. The light image is then recorded on the film with much higher sensitivity. The main purpose of using the intensifying screen with a combination of film, instead of film alone, was to raise the sensitivity of this image receptor in order to reduce the radiation dose to patients. A factor of 50–100 dose reduction is achievable. This dose reduction is often depicted by the intensification factor that is defined as the ratio of the X-ray exposure required to produce a film density (e.g., 1.0 o.d. net density) without screens versus the exposure required to produce the same film density with screens (2):

$$\text{Intensification factor} = \frac{\text{exposure required without screen}}{\text{exposure required with screen}} \quad (1)$$

Also, lower exposure required by screen-film systems can be achieved by using a shorter exposure time that has the added benefits of minimizing patient motion artifacts and reducing X-ray tube heat capacity. However, these advantages are gained at the cost of a decrease in spatial resolution and an increase in noise on the image.

Physical Properties of Intensifying Screen

As shown in Fig. 3, an intensifying screen is composed of four layers (2): a base, a reflecting coat, a phosphor layer, and a plastic protective coat. The total thickness of a typical intensifying screen is $\sim$0.3–0.6 mm.

The base is the screen support. The materials to make the screen base may be high grade cardboard or polyester plastic (2). The thickness of the base varies among manufacturers. The approximate range is 0.2–0.3 mm.

As described by Curry et al. (2), the reflecting coat is made of a substance, such as titanium dioxide (TiO_2), and is spread in a thin layer ($\sim$ 30 μm in thickness) between the phosphor layer and the base. Because light from the phosphor is emitted in all directions, including the direction away from the film, this reflecting coat acts to reflect those light photons back toward the film so that they may also contribute to the exposure on film. Although this increases the sensitivity of the screen, it also increases the blurring effect due to light spreading (shown in Fig. 3). Therefore, some manufacturers choose not to include a reflecting coat in the screen in order to improve image sharpness. This is normally the case for those screens that emphasize spatial resolution, such as those for bone imaging. Consequently, more X-ray exposure is needed to compensate for the loss in the reduced intensification factor of the screen if the reflecting coat is not added.

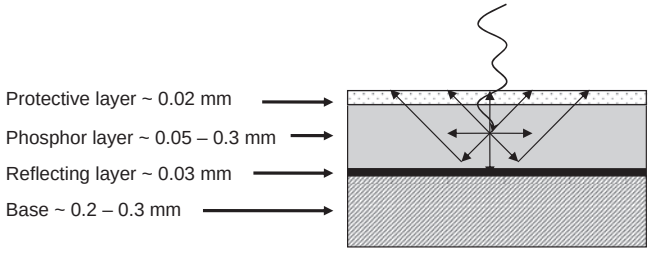


Figure 3. A cross-sectional view of an intensifying screen.

The phosphor layer, the key component of the screen, contains phosphor crystals that generate fluorescence. It is used to absorb X-ray photons and convert the absorbed X-ray energy into light photons that expose the film. The thickness of the phosphor layer varies in a range 0.05–0.3 mm, depending on the desired screen intensification factor. Usually, a thicker phosphor layer leads to a faster screen and a thinner phosphor layer leads to a slower screen.

The outmost layer consists of a protective coat that is applied over the phosphor crystals. This protective coat serves three major functions as outlined by Curry et al. (2): (1) it protects the delicate phosphor crystals; (2) it provides a surface that can be cleaned; (3) it helps to prevent static electricity that may generate image artifacts. This layer is usually made of very thin plastic.

Absorption Efficiency and Conversion Efficiency

The process of screen intensification is twofold. First, X-ray photons are absorbed by the phosphor, quantified by the absorption efficiency. Second, the absorbed X-ray energy is converted into light photons that expose the film, quantified by the conversion efficiency. The overall efficiency of a screen is the product of the absorption efficiency and the conversion efficiency, which leads to the intensification factor of the screen mentioned in equation 1. Raising either the absorption efficiency or the conversion efficiency or both can increase the screen intensification factor; thus raising the speed of the screen-film system.

The absorption efficiency, describing how efficiently the phosphor crystals absorb the incident X-ray photons, is a function of the X-ray photon energy and varies with the phosphor type (2–7,13–14). As we learn from the X-ray interactions with matter, X-ray photons are most likely to undergo the photoelectric effect when the X-ray photon energy is just slightly greater than the K-shell binding energy. This forms an abrupt increase in the X-ray photon absorption above the K-shell binding energy, called the K-edge. Obviously, it is desirable to have the K-edge of the phosphor coincide with the effective energy of the X-ray beam so that a high percentage of X-ray photons may interact with K-shell photoelectric effect. The effective energy of an X-ray beam is usually one-third to one-half of the peak in kilovolts. Most X-ray beams commonly employed in diagnostic radiology are operated in a kVp range of 50–120 kVp. Therefore, majority photons incident upon the image receptor are $<$60 keV. As a result, screen phosphors with K-edges $<$60 keV may have improved absorption efficiencies. As shown in Table 1 (7,13,14), the K-edge of the conventional calcium tungstate is 69.5 keV, above the energy of majority photons in a typical diagnostic X-ray beam spectrum, especially those X-ray beams with low kVp settings. In contrast, the K-edges of the others are $<$60 keV. Consequently, the calcium tungstate screen does not absorb X rays as efficiently as those screens with lower K-edges. Other factors that change the absorption efficiency are the thickness of the phosphor layer, the composition of the phosphor crystal, and

Table 1. Physical Properties of Some Common Phosphors[a]

Phosphor	Atomic Number of Heaviest Element	K-Edge, keV	Conversion Efficiency, %	Light Emission Spectrum
Calcium tungstate, $CaWO_4$	74	69.5	3.5	Blue (340–540 nm)
Barium strontium sulfate, $BaSO_4$:Eu	56	37.4	6	Blue (330–430 nm)
Barium fluorochloride, BaFCl:Eu	56	37.4	13	Blue (350–450 nm)
Gadolinium oxysulfide, Gd_2O_2S:Tb	64	50.2	15	Green (400–650 nm)
Lanthanum oxybromide, LaOBr	57	38.9	13	Blue (360–620 nm)
Lanthanum oxysulfide, La_2O_2S:Tb	57	38.9	12	Green (480–650 nm)
Yttrium oxysulfide, Y_2O_2S:Tb	39	17.0	18	Blue (370–630 nm)

[a]See Refs. 13 and 14.

grain size. A thicker phosphor layer will result in greater absorption efficiency. However, the disadvantage of doing so is the reduction in spatial resolution because light diffuses laterally as it propagates through the screen phosphor layer. The thicker the phosphor layer, the wider spread the light diffusion will be. For a single-screen single-emulsion system, the film is placed closer to the patient prior to the screen (shown in Fig. 2). This arrangement is based upon the fact that X-ray interaction with the phosphor reduces exponentially with the depth. Therefore, the majority of the X rays are absorbed in the top most region of the phosphor layer. Placing the intensifying screen underneath the film ensures a shorter light diffusion path. Consequently, a better spatial resolution is achieved.

The conversion efficiency is also increased from a calcium tungstate screen to a rare earth screen. As defined, the conversion efficiency is the fraction of absorbed X-ray energy converted to light energy for exposing the film. As shown in Table 1, the conversion efficiency varies according to the composition of the phosphor material.

An increase in conversion efficiency results in a greater intensification factor. However, since fewer X-ray photons are involved in image formation, an increase in conversion efficiency also results in an increase in quantum mottle in the image. Quantum mottle is the image noise caused by statistical fluctuation of the finite number of X-ray photons that form the image on film (2). As the conversion efficiency increases, the number of the X-ray photons detected by the intensifying screen reduces; thus the quantum mottle increases due to the reduced number of X-ray quanta that form the image. Unfortunately, an increased quantum mottle is associated with a decrease in low contrast detectability.

FILM

Physical Characteristics of Film

Figure 4 illustrates the cross-section of a film. The film base is to support the fragile photosensitive emulsion. Several conditions for film base materials have been outlined by Curry et al. (2) and include the following: (1) It must be transparent to visible light so that when a developed film is viewed (e.g., on a light box) the base will not interfere with the visual pattern (e.g., the image) recorded in the emulsion layer. (2) It must be strong enough to endure the film developing procedure, and, yet at the same time, be flexible and easy to handle. (3) It must be physically stable, that is, the shape of the film base must not distort during the developing process and over the long period of film storage. Historically, glass was utilized for film base. Although it was transparent, glass was too fragile and difficult to handle. In 1914, cellular nitrate was adapted to replace glass for X-ray film. But cellular nitrate was flammable and caused possible fire hazards. Later in 1924, less flammable cellulose triacetate base was developed to replace cellular nitrate (2). In 1960, polyester plastic was first introduced to make the X-ray film base. Thereafter, film base has been made of a thin, transparent sheet of plastic; the thickness of the base being ~0.2 mm.

Firmly attached to the base through adhesive substances are photosensitive emulsion layers. Because the emulsion materials are delicate in nature, a supercoat layer is used to prevent damages to the emulsion. If emulsion layers are coated on both sides of the base, the film is called double-emulsion film. Sometimes, such as in mammography, only one side of the base is coated with emulsion. Thus, the film is called single-emulsion film. Intuitively thinking, the single-emulsion film would be less sensitive than the double-emulsion film because it has only one emulsion layer. However, it gains in spatial resolution. The thickness of the emulsion layer is usually very thin, ~10 μm.

The emulsion, the key component of the film, consists of two major ingredients: gelatin and grains made of silver halide. The gelatin is used as the binder for the silver halide grains in order to keep them well dispersed. Because the gelatin is stable in nature, it provides protection and stability for the film emulsion layer before and after the film development process that we will discuss in detail in the section Film Processing. Also, the gelatin allows easy penetration for the film-processing solutions (13).

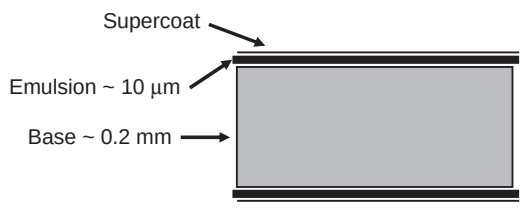


Figure 4. A cross-sectional view of an X-ray film.

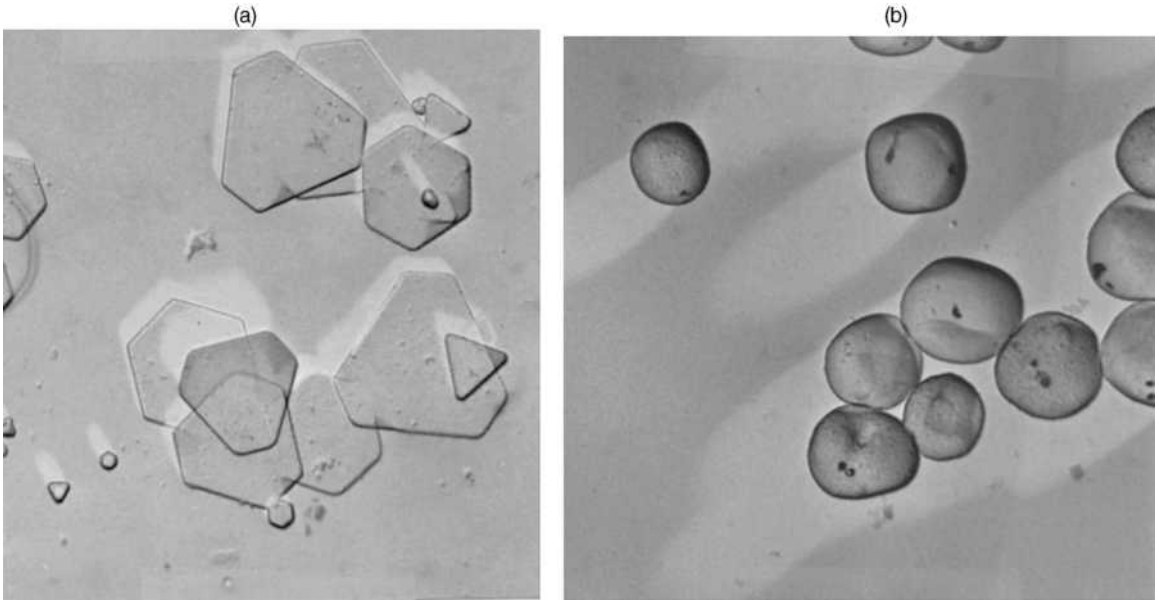


Figure 5. Photomicrographs of flat, tabular-shaped grains with the crystals orientated with the flat side parallel to the film base (a) and three-dimensional (3D) silver halide grains (b). (Courtesy of Dr. R. Dickerson of Eastman Kodak Company.)

The light sensitive silver halide is in the form of small crystals whose sizes vary in the order of 0.1–1.0 μm in diameter (15). The grain sizes, size distribution, and their shapes play a major role in determining the film speed and contrast. Generally, the larger the grain size, the greater the film sensitivity will be. This may be changed if the grain shape varies at the same time. Figure 5 shows photomicrographs of two types of crystals in the film emulsion. One contains flat, tabular-shaped grains, and the other contains the conventional, 3D silver halide crystals. As explained by Dickerson (16), the use of tabular grains increases the ratio of the surface area to volume, which leads to a significant improvement in the film sensitivity. Also, the tabular grains provide better image sharpness by reducing the light "crossover", which refers to the light emitted not by the screen in contact with the film emulsion, but by the screen opposite the film emulsion and passed through the film base. Newer technologies in 1990s developed a zero-crossover system that the film emulsion on one side of the film base was isolated from the film emulsion on the other side of the film base by adding a light-stopping dye to the film base (16,17). Obviously, the added dye must be removable during the film processing so that it will not be visible on the developed film.

Silver halide crystals can be grown in a variety of sizes and shapes by choice of emulsion precipitation conditions. Details of emulsion making were described by Wayrynen (15) and Dickerson (16). The film designers vary the emulsion grain morphology to meet various needs in film speed, resolution, contrast, and latitude.

The silver halide consists of, predominantly, silver bromide (AgBr) and a very small fraction of silver iodide (AgI) or silver chloride (AgCl), added as a sensitizer. All ions are bonded to form a cubic crystalline lattice.

Latent Image Formation

The silver halide crystals in the film emulsion also contain silver sulfide (AgS) molecules randomly distributed on the crystal surface, which tend to trap free electrons from the bromide ion during X-ray exposure. The trapping site is called the sensitivity speck. This is the location where the latent image formation begins. The mechanism of latent image formation was initially described by the Gurney-Mott theory and has remained a topic of research by many over the years (18). The Gurney-Mott theory is accepted as incomplete, but basically correct. The theory describes the latent image formation as two steps: an electronic excitation and an ionic migration. First, a visible light photon interacts with a bromide ion, forming a bromide atom and releasing an excited mobile electron:

$$Br^- + \text{photon} \rightarrow Br + \text{free electron} \qquad (2)$$

The bromide atom, Br, then migrates out of the crystal into the gelatin while the free electron becomes subsequently trapped at a sensitivity speck. The trapped electron (negatively charged) at the sensitivity speck attracts a mobile silver ion, Ag^+, (positively charged), and the two combine to form a silver atom, Ag, at the sensitivity speck:

$$Ag^+ + \text{electron} \rightarrow Ag \qquad (3)$$

This single silver atom then acts as an electron trap for subsequent freed electrons. There is a continued accumulation of silver atoms at the sensitivity speck following repeated trapping of electrons and their neutralization with silver ions. At least 3–5 silver atoms must be accumulated at the sensitivity speck in order to form a valid latent image center that can become a clump of the silver atoms after the film is developed. The film darkening is produced by the accumulated silver atoms.

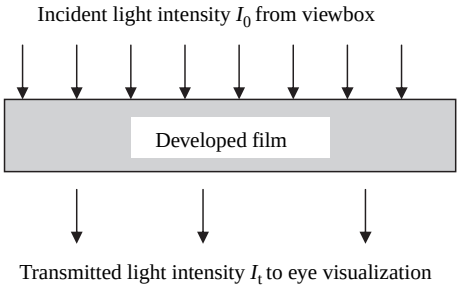


Figure 6. Illustration for the definition of the optical density of a developed X-ray film.

Usually, the film is developed immediately after the X-ray exposure is made. If the film is delayed in developing, the film speed and contrast will be reduced. This phenomenon is called latent image fading that can also be explained by the mechanism mentioned above.

Optical Density

In photographic print film, the more exposure the film receives, the brighter the print film will be. In screen-film radiography, it is the opposite because X-ray film is a negative recorder. An increased X-ray exposure produces darker film. The degree of darkness of the developed film is defined by a term called optical density or, simply, density (4). Shown in Fig. 6, if the incident light intensity from a viewbox, where the film is read, is given by I_0, and the transmitted light intensity through the film for one particular spot is given by I_t, then the optical density (O. D.) at that location is given as the 10 based logarithm of the ratio of the incident light intensity versus the transmitted light intensity:

$$\text{O.D.} = \log_{10}\left(\frac{I_0}{I_t}\right) \tag{4}$$

For example, if the transmittance at one film location is 10%, then the optical density of that location is 1.0. Notice the relationship of the transmittance and the optical density is logarithmically compressed. Hence, if the optical density is 3.0, then the transmittance is only 0.1%. The optical density is often measured by a device called a densitometer. A densitometer has a small sensitive area (aperture), where the light intensity without film in the light beam—I_0 is measured; then with the film inserted and the region of interest covering the aperture, the densitometer, measures the light transmitted through the film—I_t. The optical density is read accordingly when the read-out button is pressed on the densitometer.

Table 2 lists a range of optical densities. Optical density 0.0 is reached only when the light transmission through the film is 100%, which is not practically possible. Even when a fresh unexposed film is processed, it will still have a low film density of ~0.15 O.D., which is known as the film base plus fog. The base density is caused by the absorption of light through the film base material. The current polyester film base is usually slightly blue-tinted in order to ease the human eye visualization and help cancel distractions from faint chemical residue resulted from film processing. The fog density is mainly due to unintended exposure or the development of unexposed silver halide grains in the film emulsion. The fog may be affected by conditions related to the film storage and the film processing. It is desirable to keep the film base plus fog <0.20 O.D. (19).

Table 2. Optical Density Range of an X-Ray Film

Optical Density	Comments
0.0	100% transmission (practically impossible)
0.15–0.20	Typical base + fog
0.45–2.20	Useful density range
3.0	Very dark; needs a hot lamp for high luminance
>3.5	Typical maximum darkness

Table 2 lists the useful density ranges between 0.45 and 2.20 O.D. In this range, the human eye can appreciate the shades of contrast with a standard viewbox of which the luminance level is ~1500 nits. For mammography, in which films are generally darker and with more demands of low contrast perception, the viewbox luminance level should be at least 3500 nits (19). A brighter viewbox improves the ability to appreciate the contrast in the dark density regions. Sometimes, hot lamps are utilized to provide more intense incident light to view the film. It is typically utilized to check the dark regions on film such as those specific regions with optical densities >3.0 O.D. The maximum darkness a film can reach is called D_{max}. It varies from film to film. A mammography film may reach a D_{max} of 3.5 O.D. or greater; while a regular film may barely reach 3.0 O.D. For comparison, print films are normally saturated at a density of 1.7 or 2.0 O.D. at best (2).

The Hurter and Driffield Curve (H&D Curve)

The relationship between the exposure to a film and the film density corresponding to that exposure can be plotted as the characteristic curve or H&D curve (named after Hurter and Driffield who were the first to use such a curve) (3). The exposure is plotted along the horizontal axis in

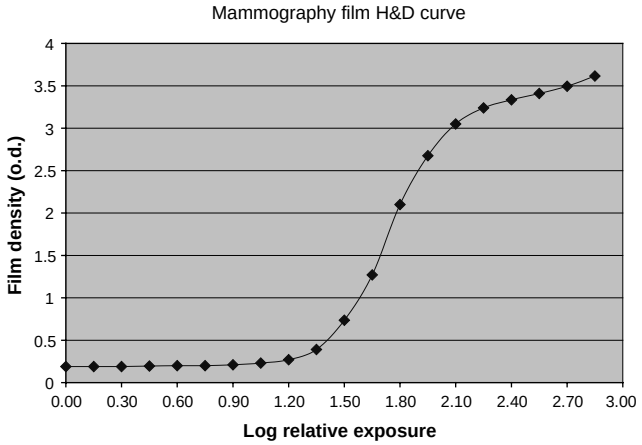


Figure 7. An example of H&D curve of an X-ray film.

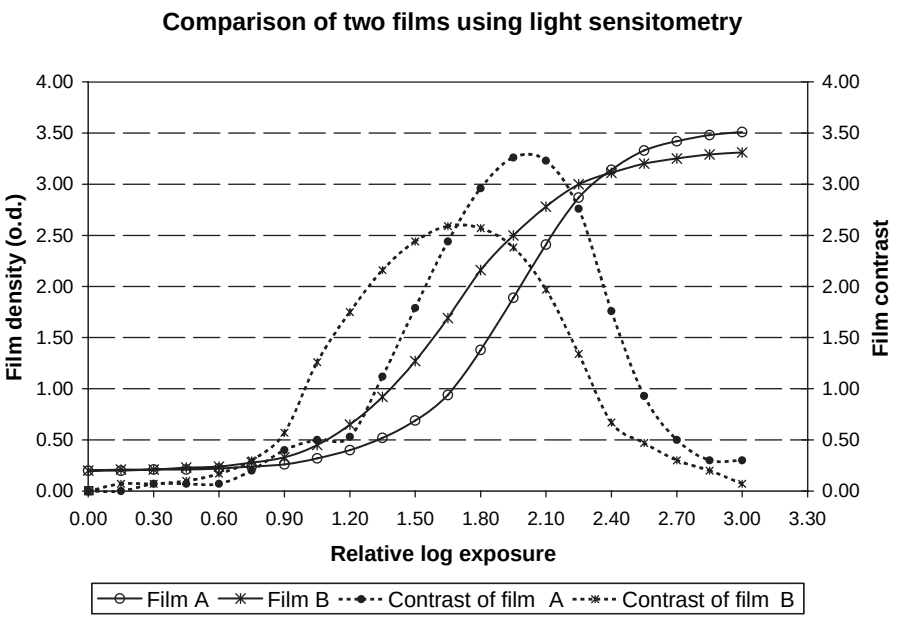


Figure 8. A comparison of two films using their H&D curves obtained by light sensitometry.

logarithmic scale, and the film density is plotted along the vertical axis. Figure 7 shows an example of the H&D curve of an X-ray film using light sensitometry, where relative exposures are utilized instead of actual exposures, because, in practice, the relationship between two exposures is often more important than the actual exposure. The H&D curve illustrates how much the optical density may change corresponding to the exposure change. As outlined by Bushberg et al. (4), the curve has a sigmoid shape, beginning with the base plus fog for no exposure, which cannot actually be plotted on a logarithmic scale, and the "toe" for the low exposure region. In the "toe", the film density rises very slowly as the exposure increases. This region corresponds to "underexposed" film regions, such as the mediastinum on a chest radiograph. Beyond the "toe", the film density escalates linearly to the increase of logarithmic exposure. This is called the "linear region" or the "straight-line portion" of the H&D curve. In this region, the film contrast is at its best and the film density fits in the useful density range as shown in Table 2. Ideally, a radiograph image should be exposed with proper technique to have densities within this region. Further beyond the "linear region", the curve flattens out. This region is called the "shoulder". It corresponds to the "overexposed" region where the film contrast is compromised.

Film Contrast and Latitude

The image contrast in film radiography is shown as the difference in film densities at various locations on the image. It is contributed by both the subject and film contrast. The subject contrast is determined by the difference in X-ray intensity after transmitting through various regions of the patient. It is primarily determined by the differential attenuation of tissues. For example, a solid tumor may attenuate more X rays than its surrounding tissue; thus leaving a lower X-ray intensity behind the tumor. The subject contrast relies on factors such as kVp, differences in patient part thickness, atomic number, and density of tissues. Film contrast represents the ability of the film to manifest the subject contrast into the image contrast on film. Choosing the correct film will enhance the subject contrast on the final image. The film contrast is determined by the slope of the H&D curve: a steeper slope leads to a higher film contrast. For a given H&D curve, the slope is a function of film density, being low in the toe region, peaking in the straight-line portion, then being low again in the shoulder region (shown in Fig. 8). Based upon the H&D curve, the film contrast is often quantified by two commonly used parameters: gamma and average gradient. Gamma is defined in the straight-line portion by the maximum slope of the H&D curve:

$$\gamma = \frac{D_2 - D_1}{\log E_2 - \log E_1} \tag{5}$$

where D_2 and D_1 outline the straight-line portion of the H&D curve and E_2 and E_1 are exposures needed to produce the film densities of D_2 and D_1. The average gradient is often calculated over a density range of 0.25–2.0 O.D. both above the base plus fog, because such a density range is considered the useful density region:

$$\text{Average Gradient} = \frac{D_S - D_B}{\log E_S - \log E_B} = \frac{1.75}{\log E_S - \log E_B} \tag{6}$$

where D_S is 2.0 O.D. above the base plus fog and D_B is 0.25 O.D. above the base plus fog. The parameters E_S and E_B are exposures needed to produce the film densities of D_S and D_B.

Very often, a special term called "latitude" is also cited for film photographic characteristics. The latitude is usually defined as the exposure range that produces a certain density range over which a usable image can be made on film (20). If this density range is between the upper and lower optical density limits on an image, then the corresponding exposure range is also called the

"dynamic range". Obviously, the film latitude changes inversely with the film contrast. Films with wide latitude exhibit lower film contrast than films with narrow latitude. Sometimes, a wide latitude is preferred in spite of compromising the film contrast. For example, in chest imaging, wide latitude films are desirable because of the large dynamic range needed to cover areas behind lungs as well as mediastinum. In other words, a wide latitude film is capable of keeping the lung regions below the shoulder and the mediastinal region above the toe. With the new inventions of digital imaging (e.g., CR, DR), the wide linear dynamic range of these technologies has proved to be advantageous and ultimately solve the conflict between the film latitude and contrast.

The characteristic curves are often employed to compare films regarding various properties including speed, contrast, base plus fog, and maximum density. Figure 8 shows the characteristic curves of two films. The H&D curves are shown in solid lines and the film contrasts as functions of the film density are shown in dashed lines. The graph has dual vertical axes with the left axis for the film density and the right axis for the film contrast. The graph demonstrates Film B has a faster speed, but a lower contrast than Film A. Both have similar base plus fog. Film A can reach a higher D_{max} than Film B. Note that the characteristic curves shown in Fig. 8 are from light sensitometry. A device called a sensitometer is utilized to expose the film in a "step" fashion with a range of constant light intensities that each differs by a factor of the square root of 2 (i.e., 0.15 increment on the logarithmic exposure scale of the H&D curve). The sensitometer simulates the light emitted from the intensifying screens. Nevertheless, because the simulated light is not the same as the emitted light from screens, the response of film to the sensitometer may differ from that to the actual light from the intensifying screens caused by X-ray exposure. Special cautions are needed when comparisons are made based upon the light sensitometry (21).

The contrast of the image receptor of the screen–film system, as a whole, is primarily dependent on its film contrast; although there are exceptions, such as Kodak InSight asymmetric film–screen systems, of which the contrast can be varied by changing the intensifying screens (16,17).

Spectral Emission and Spectral Sensitivity

Obviously, the wavelength of light emitted by an intensifying screen should correspond closely with the spectral sensitivity of the film used with the screen. Otherwise, the total photographic effect is decreased, and the patient has to suffer more radiation exposure to get the proper film density. Usually, the sensitive spectrum of the film is matched to cover all the wavelengths emitted by the screen in order to maximize the speed of the screen–film combination. Shown in Fig. 9 is a spectral match of blue sensitive X-ray film with the calcium tungstate spectral emission that is continuous with a peak in the blue region (16). The invention of rare earth phosphors made the intensifying screen more efficient in both absorption and conversion of X-ray photons. However, the luminance from the majority

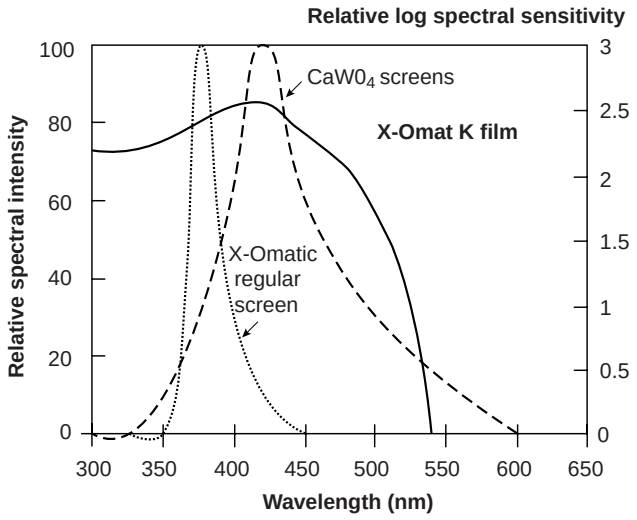


Figure 9. Relative spectral match of Kodak X-Omat K film with ultraviolet (UV) and blue-emitting screens. (Courtesy of Dr. R. Dickerson of Eastman Kodak Company.)

of the rare earth screens was changed to the color green, which made it necessary to change the spectral sensitivity of X-ray films accordingly. The extension of the film spectral sensitivity to longer wavelengths was achieved by adding spectral sensitizing dyes absorbed by silver halide (13). Figure 10 shows a spectral match of green sensitive X-ray film with the spectral emission of a rare earth intensifying screen. Note that, in Fig. 10, despite the fact that the screen emission spectrum was beyond the green region, the film spectral sensitivity was topped at the green region in order to have a safelight zone in the color red. The safelight in the darkroom is for the convenience of handling the film without film fogging. For many years, amber safelights were utilized to handle blue sensitive X-ray films. As the upper limit of the film spectral sensitivity moved up from the blue region to the green region, the

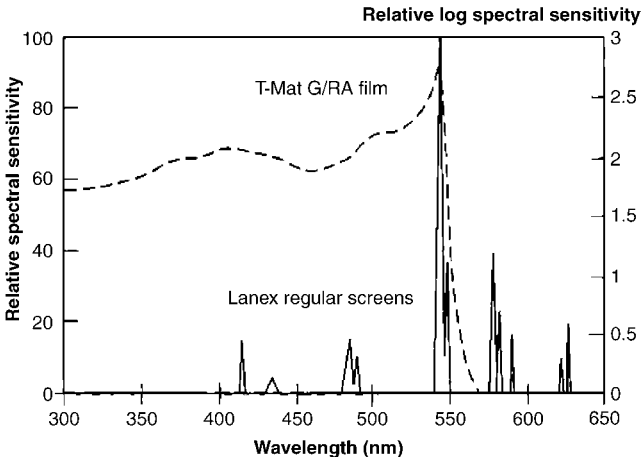


Figure 10. Relative spectral match of Kodak T-Mat G film with Kodak Lanex regular screens. (Courtesy of Dr. R. Dickerson of Eastman Kodak Company.)

darkroom safelight was necessary to change from the amber to the red safelight.

Speed and Resolving Power

The speed of a film, also known as the system sensitivity, is defined in the unit of $1/R$ as the inverse of the exposure that produces a net film density of 1.0 O.D. above the base plus fog. Because film is utilized in combination with intensifying screens, the speed commonly quoted is actually the speed of the screen–film system. Although with a faster film, the speed of screen–film system may be increased accordingly, the determining factor for the speed of a screen–film system is in the screen not in the film. For a screen–film system, the speed can be compared by examining the location of the H&D curve along the horizontal logarithmic exposure scale at a net density of 1.0 O.D. above base plus fog. In general, when a horizontal line is drawn at 1.0 O.D. above the base plus fog, the curve that appears toward the left on the exposure scale has faster speed than those that appear toward the right (see Fig. 8). Apparently, this speed depends on a number of variables, especially kVp and film processing conditions. The American National Standards Institute (ANSI) has published guidelines to standardize the method for measuring the speed of screen-film systems (22). The standard method is rigid in defining the exposure and the controlled film-processing conditions. Unfortunately, such absolute measured results are not widely available. In practice, a concept of "speed class system" has been commonly accepted for comparing the relative speeds among various screen–film systems (2). The specified speed class differs in value by ~25%, which is similar to the camera speed class adapted in photography. For example, the sequence of the numbers in speed class starting from 100 is 100, 125, 160, 200, 250, 320, 400, 500. A great number in speed class means a fast screen–film system.

A faster screen–film system is mainly due to a greater intensification factor in the screen, which is often achieved by utilizing a thicker screen phosphor layer that leads to a reduction in resolving power. A general comparison of screen speed and resolving power is shown in Table 3.

Reciprocity Law

The relationship between the optical density and the X-ray exposure depicted in the H&D curve is usually not dependent on the exposure rate. The film density remains the same as long as the amount of the X-ray exposure received by the screen–film system is the same for a wide range of exposure rates. This phenomenon is known as reciprocity law. It allows the X-ray technique to vary in tube current milliamps (mA) and exposure time as long as the product of milliamps and exposure time is constant.

However, the reciprocity law has been known to fail at very short (i.e., the exposure rate extremely high) or very long exposure (i.e., the exposure rate extremely low) with screen–film systems. For example, the film speed is observed to be reduced in mammography where very long exposure time (>2 s) has to be utilized for large or dense breasts (23). The reciprocity law failure can be explained by the latent image formation mechanism described earlier as the Gurney and Mott theory (18). In the case of extremely low exposure rate, the rate of photon absorption is too low to allow the gathering of enough silver atoms at the sensitivity speck to make it stable. Only one or two silver atoms at a sensitivity speck may disintegrate; thus the latent image may disappear before enough silver atoms can be gathered. Therefore, the speed is reduced at the extremely low exposure rate. In the opposite case, where the exposure rate is extremely high, the production of free electrons is too fast for all the silver ions to migrate to the proper sensitivity specks. Therefore, a fraction of the silver ions may not be utilized in forming silver atoms. This also leads to the reduction in film speed. For X-ray units, the automatic brightness control (ABC) on the system, or sometimes referred to as the phototimer, is a feature designed to maintain the consistency in film density regardless of the patient thickness. It is also designed to compensate for the reciprocity law failure.

The reciprocity law failure applies only to screen–film systems, but not to direct exposed film systems.

Table 3. Physical Characteristics of Screen–Film Systems.

Type of screen-film system	Speed Class	Resolving Power, lp·mm^{-1}
Film alone		~100
Fine detail	<100	~10–15
Par-speed	100	~8–10
Regular	200–400	~6–8
High speed	~800	~4–6

THE SCREEN–FILM CASSETTE

Figure 11 shows images of two screen–film cassettes. The double emulsion film is sandwiched between the two intensifying screens. In some applications, such as mammography, a single emulsion film is utilized with only one intensifying screen in the cassette. The front of the cassette is X-ray transparent usually made of materials with low atomic numbers, while the back of the cassette often includes a sheet of lead to reduce backscattering. The cassette is light tight so that the film, once being shut into the cassette, will not be exposed to ambient light. Opening or closing the cassette has to take place in a darkroom or other dark environment. The screens are usually mounted on layers of compressible foams so that when the two cassette halves close, air is expelled from the space between the screens and the film. This is to ensure good image quality by having a firm film screen contact without air trapping in between.

FILM PROCESSING

After a latent image is formed on the film by X-ray exposure, the film is subjected to a sequence of chemical processes in order to make the latent image visible to human eyes. This is film processing or film development (21).

Films had been developed manually in the darkroom until the first automatic film processor was introduced in 1956 (4). An automatic film processor (see Fig. 12) consists

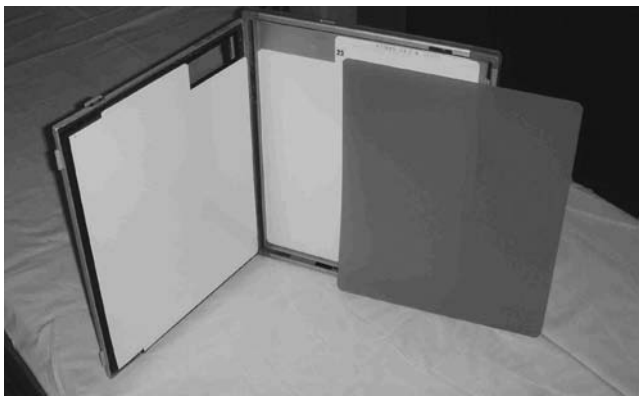


Figure 11. Pictures of the screen–film cassettes: (a) a double-screen double-emulsion cassette; (b) a single-screen single-emulsion cassette.

of three tanks: developer, fixer, and water to wash off the chemicals from the film. The film is then dried in a drying chamber and ready for reading at the drop slot. A standard cycle takes ~90 s for film processing. There are rapid cycles that take less time and extended cycles that need longer time. Rollers are utilized to transport the film along the circuitous path through each of these tanks. These rollers need to be serviced periodically to prevent roller transport processing artifacts (24). Film artifacts refer to those features on the image that do not reflect the true subject being imaged. The film artifacts are highly undesirable because they distract or even mislead readers. Although artifacts may be generated by factors other than the film processing such as X-ray equipment, patient positioning, film storage and handling, and so on, the film-processing artifacts are among the most commonly seen artifacts in routine clinical images. Details of the film artifacts, and how to battle against these artifacts, can be found in Ref. 21 Chapt. 6.

As the film enters the developer tank, the developer acts as the reducing agent that selectively reduces those

Figure 12. A picture of an automated daylight film processor.

exposed grains to silver. For those silver halide crystals that have been exposed to photons, latent image centers have been formed, each center having at least 3–5 silver atoms, and typically having 10 or more silver atoms. These silver atoms act to catalyze the further reduction of the silver in the presence of the reducing agent which, in this case, is the developer. After development, those exposed silver halide crystals become dark silver grains. One sensitivity speck can be converted to $\sim 10^9$ silver atoms, which gains an enormous amplification (16). They contribute to the O.D. of the film. A great O.D. region has a high concentration of such silver grains; a light O.D. region has a low concentration of such silver grains. An ideal developer should have only the exposed grain developed and unexposed grain washed out. But realistically, unexposed silver halide grains that do not contain latent images may also be developed. This causes "fog".

The film developing time and temperature of the developer must be optimized because the film speed and the film contrast are directly affected by both factors. Usually, the speed and the contrast increase, up to a certain degree, by the increase of the development time. Then the speed and the contrast reach a plateau or may even decrease with further increase of the development time. Similar effect is observed for the development temperature on the speed and the contrast. Also, note that film fog increases with the increasing development time or temperature, which is not desirable. Therefore, it is important to optimize these film-processing conditions in order to achieve a reasonable compromise among film speed, contrast, and fog.

A fixer is used to terminate the development process and dissolves unexposed silver halide grains without damaging the image formed by metallic silver already developed by

the developer. An incompletely fixed film is easily recognized because it has a "cloudy" appearance. This is a result of the dispersion of transmitted light by those very small silver bromide crystals that have not been removed by the fixer.

Washing the film thoroughly with water is necessary. Otherwise the chemical residues left on the film will turn the film into brown as it ages. This is the function of the water tank after the developer and the fixer.

The last step for film processing is drying. Warm air is blown over both surfaces of the film when it is transported by rollers through the drying chamber. Finally, the dried film is delivered at the drop slot of the automated film processor.

QUALITY ASSURANCE AND QUALITY CONTROL (QA/QC) FOR SCREEN–FILM SYSTEMS

There are many aspects of QA/QC protocols for screen–film system maintenance and film processing. We will briefly discuss several important tests for routine QA/QC in the clinical environment.

The film processor needs day-to-day quality control in order to ensure consistent performance. The major concern in film processing is the instability due to wet chemistry variation and temperature fluctuation. The film sensitometry is a method to evaluate and monitor the performance of a film processor on a daily basis. As an important step in film processor QC, a light sensitometry strip (see Fig. 13) is made by a sensitometer and charted in the morning before any patient images are processed. Needed for film sensitometry are a sensitometer, a box of control films that are identical so that variations among films are negligible, and a densitometer to measure the film density. In addition, a thermometer is utilized to measure the developer temperature to ensure the optimal temperature according to the manufacturer's guideline. Three basic concerns for film sensitometry are (1) the base plus fog; (2) the speed index; (3) the contrast index. The base plus fog is often measured at an unexposed area of the sensitometry film. For the speed index, a step is often designated with a mid-gray film density such as a film density of no less than, but close to 1.2 O.D. for mammography (19). For the contrast index, two steps are designated with one step of high film density, such as a film density ~2.20 O.D. and the other of low film density such as a film density of no less than, but close to 0.45 O.D. (19). The difference in density between these two steps is called the contrast index. The aim values of these three indexes are established over a period such as 5 consecutive days to smooth out normal day-to-day variation at the initial film processor installation or after a thorough preventative maintenance and calibration on the film processor. Day-to-day observation compares the measured indexes with the aim values. Figure 14 demonstrates sensitometric data for a film processor within one month. Records like these allow the tracking of the performance trends. The purpose is to eliminate problems in processor performance before those problems affect patient images. In mammography, 0.10 O.D. variation from the aim value triggers the action level for further monitoring and correction, and 0.15 O.D. variation demands immediate corrective actions before patient films can be processed (19).

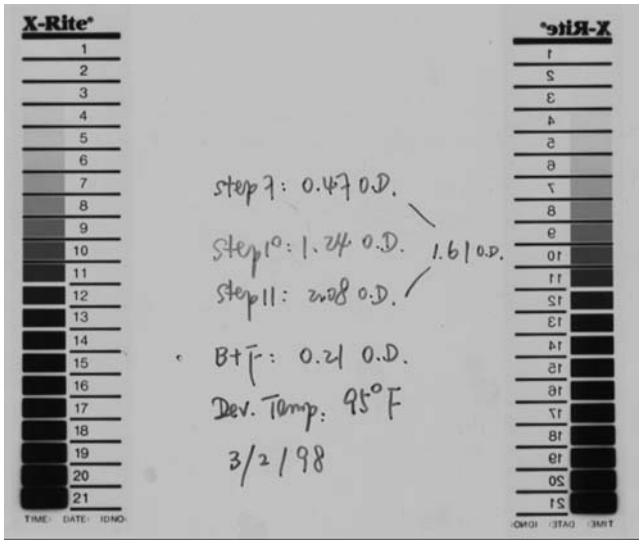


Figure 13. An example of a film sensitometry strip made by a sensitometer.

There are certain screen–film cassette maintenance issues. First, the intensifying screens need to be kept clean. Artifacts may appear on images in the presence of dust on the screens. In mammography, screen clean-up is required at least once a week (19). In general radiography, screens are recommended to be cleaned at least on a quarterly basis. The cleaning procedure involves a solution containing an antistatic compound and a detergent. With a soft lint-free cloth, the screen surface should be wiped very gently with the solution. Then the cassette should be left open to air-dry. Another important issue of the screen–film cassette maintenance regards screen–film contact. A good screen–film contact is essential to prevent loss of spatial resolution. A simple method for testing the screen–film contact is to image a piece of wire pattern such as a mesh phantom placed on the top of the cassette. The sharpness of the wire pattern in all regions of the filmed image should be examined. If the screen–film contact is poor in certain areas, fuzziness or slightly increased density will be observed at those locations.

Finally, an important routine QA/QC test in a film-based imaging environment is related to darkroom fog. The darkroom fog degrades the image quality unnecessarily and should be prevented. Remedies to eliminate darkroom fog are straightforward, involving sealing the darkroom from light leaks (e.g., the door frame, the film processor, the film pass-box and any openings into the room) and following instructions on safelight type, bulb wattage and safelight positioning (19,25,26). Such easily preventable problems are often unnoticed or neglected (26). Therefore, It needs to be emphasized to have the darkroom fog tested during the initial film processor installation and monitored thereafter periodically (e.g., semiannually). To measure the darkroom fog, we may expose a film with a phantom that would produce a mid-gray film density when the film is developed (19). In the darkroom before

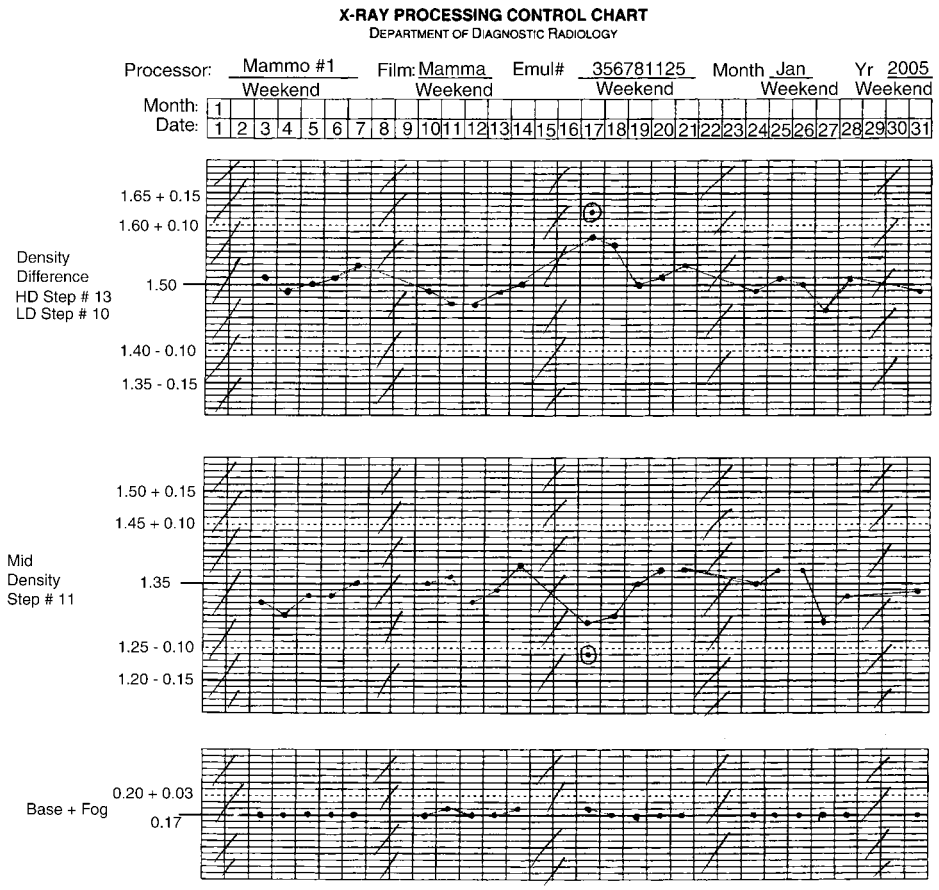


Figure 14. A chart of a film processor through a period of 1 month using the sensitometry method.

processing the film, place the film on the darkroom counter and bisect the latent image by using an opaque paper so that one-half of the film is protected from and the other exposed to any possible fogging sources in the darkroom. The covering line should be parallel to the anode–cathode direction of the X-ray tube in order to prevent any density difference due to the heel effect. Turn on the safelight in the darkroom. After 2 min of such fogging exposure, develop the film. The density difference of the two portions of the film near the covering line determines the darkroom fog. Usually, the darkroom fog should not exceed 0.05 O.D. (19).

ACKNOWLEDGMENTS

The author is grateful to Dr. Robert Dickerson at Eastman Kodak Company for many insightful suggestions and to Ms. Maryellen Peinelt for assistance in preparation of the article.

BIBLIOGRAPHY

1. Glasser O. Wilhelm Conrad Röntgen and the Early History of the Roentgen Rays. San Francisco: Norman Publishers; 1989.
2. Curry III TS, Dowdey JE, Murry Jr RC. Christensen's Physics of Diagnostic Radiology. 4th ed. Philadelphia: Lea & Febiger; 1990. Chapt. 9–11. p 118–164.
3. Ter-Pogossian MM. The Physical Aspects of Diagnostic Radiology. New York: Harper & Row; 1967. Chapt. 6. p 185–240.
4. Bushberg JT, Seibert JA, Leidholdt Jr EM, Boone JM. The Essential Physics of Medical Imaging. 2nd ed. Philadelphia: Lippincott Williams & Wilkins; 2002. Chapt. 6–7. p 145–189.
5. Buchanan RA. An improved X-ray intensifying screen. IEEE Trans Nucl Sci 1972;NS-19:81–86.
6. Buchanan RA, Finkelstein SI, Wickersheim KA. X-ray exposure reduction using rare earth oxysulfide intensifying screens. Radiology 1972;105:185–190.
7. Stevels AN. New phosphors for X-ray screens. Medica Mundi 1975;20:12.
8. Huang HK. PACS Basic Principles and Applications. New York: John Wiley & Sons Inc.; 1999. Chapt. 4. p 63–90.
9. Sanada S, Doi K, Xu X, Yin F, Giger ML, MacMahon H. Comparison of imaging properties of a computed radiography system and screen-film systems. Med Phys 1991;18: 414–420.
10. Samei E, Flynn MJ. An experimental comparison of detector performance for direct and indirect digital radiography systems. Med Phys 2003;30:608–622.
11. Pisano ED, Yaffe MJ. Digital mammography. Radiology 2005;234:353–362.
12. Samei E, Seibert JA, Andriole K, Badano A, Crawford J, Reiner B, Flynn MJ, Chang P. AAPM/RSNA tutorial on equipment selection: PACS equipment overview: general guidelines for purchasing and acceptance testing of PACS equipment. Radiographics 2004;24(1):313–334.
13. Arnold BA. Physical characteristics of screen-film combinations. In: Haus AG, editor. The Physics of Medical Imaging, Recording System Measurements and Techniques. New York: American Institute of Physics; 1979. p 30–71.
14. Thomason C. Screen-film systems. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation. New York: John Wiley & Sons Inc.; 1988. p 2599–2609.

15. Wayrynen RE. The photographic process. In: Haus AG, editor. The Physics of Medical Imaging, Recording System Measurements and Techniques. New York: American Institute of Physics; 1979. p 1–15.
16. Dickerson R. Fundamental of silver halide film design. In: Haus AG, editor. Advances in Film Processing Systems Technology and Quality Control in Medical Imaging. Medical Physics Publishing; 2000. p 73–84.
17. Pizzutiello Jr RJ, Cullinan JE. Introduction to Medical Radiographic Imaging. Rochester (NY): Kodak Publication M1-18; 1993. Chapt. 4–5. p 71–122.
18. Mees CEK, James TH. The Theory of the Photographic Process. New York: Macmillan; 1966. Chapt. 5. p 87–119.
19. Mammography Quality Control Manual, by ACR Committee on Quality Assurance in Mammography chaired by Hendrick RE, ACR; 1999.
20. Meeson S, Young KC, Rust A, Wallis MG, Cooke J, Ramsdale ML. Implications of using high contrast mammography X-ray film-screen combinations. Br J Radiol 2001;74:825–835.
21. Haus AG, Jaskulski SM. The Basics of Film Processing in Medical Imaging. Medical Physics Publishing; 1997. Chapts. 3, 5, 6.
22. American National Standards Institute: Method for the Sensitometry of Medical X-Ray Screen-Film-Processing Systems. New York: (ANSI PH2.43-1982); 1982.
23. Almeida A, Sobol WT, Barnes GT. Characterization of the reciprocity law failure in three mammography screen-film systems. Med Phys 1999;26:682–688.
24. Widmer JH, Lillie RF. Roller transport processing artifact diagnosis. In: Haus AG, editor. Film Processing in Medical Imaging. Medical Physics Publishing; 1993. p 115–129.
25. Suleiman OH, Showalter CK, Gross RE, Bunge RE. Radiographic film fog in the darkroom. Radiology 1984;151(1): 237–238.
26. Gray JE. Mammography (and radiology?) is still plagued with poor quality in photographic processing and darkroom fog. Radiology 1994;191:318–319.

See also X-RAYS: INTERACTION WITH MATTER.

SENSORS, GLUCOSE. See GLUCOSE SENSORS.

SENSORS, OPTICAL. See OPTICAL SENSORS.

SENSORS, PIEZOELECTRIC. See PIEZOELECTRIC SENSORS.

SENSORY AIDS. See BLIND AND VISUALLY IMPAIRED, ASSISTIVE TECHNOLOGY FOR.

SEXUAL INSTRUMENTATION

KIRK A. BRUNSWIG NEWRING
CRISSA DRAPER
WILLIAM O'DONOHUE
University of Nevada
Reno, Nevada

INTRODUCTION

From the annals of measurement comes an item of lore on the origins of measurement, occurring when two male hominids were urinating side by side arguing over whose was bigger. More recently, it is told in the pubs of Ireland that the yardstick was invented in Dublin to settle such disputes.

As Semmlow (1) stated in the first edition of this Encyclopedia, sexual instrumentation is a rapidly evolving science. (*Note*: This article does not endorse an instrument or instruments. None of the authors have received compensation from a manufacturer of any of the products described in this work.) Many early developments in the arena of sexual instrumentation relied on "nonspecific responses associated with increased activity in the sympathetic nervous system", such as heart rate and blood pressure. However, more recent investigations have turned to more specific physiological responses, generally focusing on the physiological responses of the genitals. While the scope of this entry prohibits a full description of the history and developments in the assessment and interventions related to the human sexual response, interested readers are welcome to refer to the works of Kinsey and Masters and Johnson for a more detailed review of the history of human sexuality research.

The need for and sophistication of measurement of human sexual behavior has improved dramatically over the years; however, there are still several scientific, methodological, and statistical challenges facing those researching the measurement of human sexual behavior. The current entry will review these measurements along with instruments and procedures related to female and male sexual behavior. Within each sex, the measures related to assessment and diagnosis of sexual concerns are reviewed first, followed by a review of devices used in treatment.

INSTRUMENTS AND MEASUREMENT OF FEMALE HUMAN SEXUAL BEHAVIOR

Formal research into physiological aspects of female sexual behavior was spurred in a large part by the work of Masters and Johnson in the 1960s. They highlighted the importance of genital vasocongestion as an indicator of sexual arousal. Later researchers explored the multidimensional nature of female arousal, often ascribing three necessary aspects: physiological, cognitive, and emotional. That is, physiological arousal is a necessary condition for sexual arousal, though it alone is often insufficient; however, early research into women's sexual behavior focused almost exclusively on the physiological aspects; detailing the physiological changes across the human female sexual response of arousal, plateau, orgasm, and resolution. The absence of an integrative research approach toward the assessment and treatment of female sexual response is discouraging, as this is a necessary condition for sexual arousal, and it seems to be a common problem among women. Estimates range from 1 in 3 to 1 in 5 women between 18 and 59 years of age complain of a lack of interest in sex.

A series of studies in the 1980s evaluated the normative aspects of human sexuality research volunteers. Through

these studies, it became clear that both male and female university-based research participants were decidedly different than their college peers. Typically, participants were more likely to participate in human sexuality research if they were to remain clothed. Also, those who were willing to participate were more likely to masturbate, to have had early and more frequent exposure to commercial erotica, to have less sexual fear, and (for female participants) to have a history of being sexually abused. Thus, for the early research studies listed, restraint should be used when extrapolating the findings discussed to dissimilar populations. More recent research has employed intent-to-treat and wait list control approaches in community-based research. These approaches have increased the ecological validity of this line of research, and this trend should continue as researchers increasingly use more ecologically valid methodology.

FEMALE SEXUAL BEHAVIOR ASSESSMENT

Internal Devices

In the late 1960s and 1970s, several devices were constructed to assess physiological components of female sexual response. One early device measured vaginal blood flow using two thermistors mounted on a diaphragm ring (please refer to later section on thermistor clip for the use of temperature measurement in female sexual behavior for more information). Other early devices included those intended to measure vaginal pH, temperature, and genital engorgement and blood flow properties via photoplethysmography. Common to all of these early instruments was a lack of standardization. Of these instruments, photoplethysmography received the most empirical attention.

Vaginal Photoplethysmography. Hatch's 1979 (2) review of vaginal photoplethysmography (VPPG) provides a description of the device as well as the issues and controversies surrounding its use at the time. The VPPG is described as a cylindrical probe approximately the size and shape of a menstrual tampon which is easily inserted into the vagina by the subject. A light source contained within the clear acrylic probe is directed at the wall of the vaginal lumen. A photoelectric transducer is also situated on the probe in such a position that it detects only that fraction of the incident light which is reflected from the vaginal tissue and the blood circulating within it. Changes in blood volume within the vagina produce changes in the amount of the incident light backscattered to the light detector, because of the large difference in transparency between blood and bloodless tissue. Changes in blood volume can therefore be easily recorded as changes in the output of photoelectric transducer (p. 358) (Figs. 1 and 2).

Hoon et al. (5) provide a description of a VPPG (including construction, instruction, recommended calibration procedures, and diagrams) and note there is a lack of standardization among VPPG, indicating that differences in probe size may have impacted cross-study comparisons, as larger probes are thought to reduce movement at the expense of possible tissue stretching, blood supply occlusions, reduced sensitivity, and decreased acceptability by women.

Figure 1. Vaginal photoplethysmograph similar to that originally designed by Sintchak and Geer (3) using a single light source and photodetector. (Courtesy of Farrall Instruments, Inc., Grand Island, NE.)

One of the challenges with VPPG is the choice of current: alternating current (ac) or direct current (dc). Direct current coupling is used to assess slowly developing changes in vaginal blood volume (VBV), which is suggested as reflecting the pooling of blood in the vagina. Alternating current coupling is used to assess vaginal pulse amplitude (VPA), which is the observation of the arrival of the pulse wave at the observation point on each cardiac cycle. While some researchers approach the ac/dc issue as an either/or issue (2), more recent research (5) has measured both VBV and VPA. In reviewing the controversy surrounding the sensitivity of VBV and VPA, Hatch concludes, "there is some evidence suggesting that VPA is the more sensitive variable in the sense that it has been shown to significantly discriminate among responses in various types of erotica in some cases where VBV has not" (Ref. 2 p. 359).

Geer et al. (6) presented one of the earlier reports on the use of VPPG to demonstrate differential vaginal states following exposure to erotic and nonerotic visual stimuli. In their study, VPPG assessed VBV and VPA. However, they were one of many research groups to find that physiological arousal did not necessarily correlate with subjective ratings of arousal. Similarly, van Dam et al. (7) used VPPG to compare women with and without amenorrhea. The authors theorized and found that women with

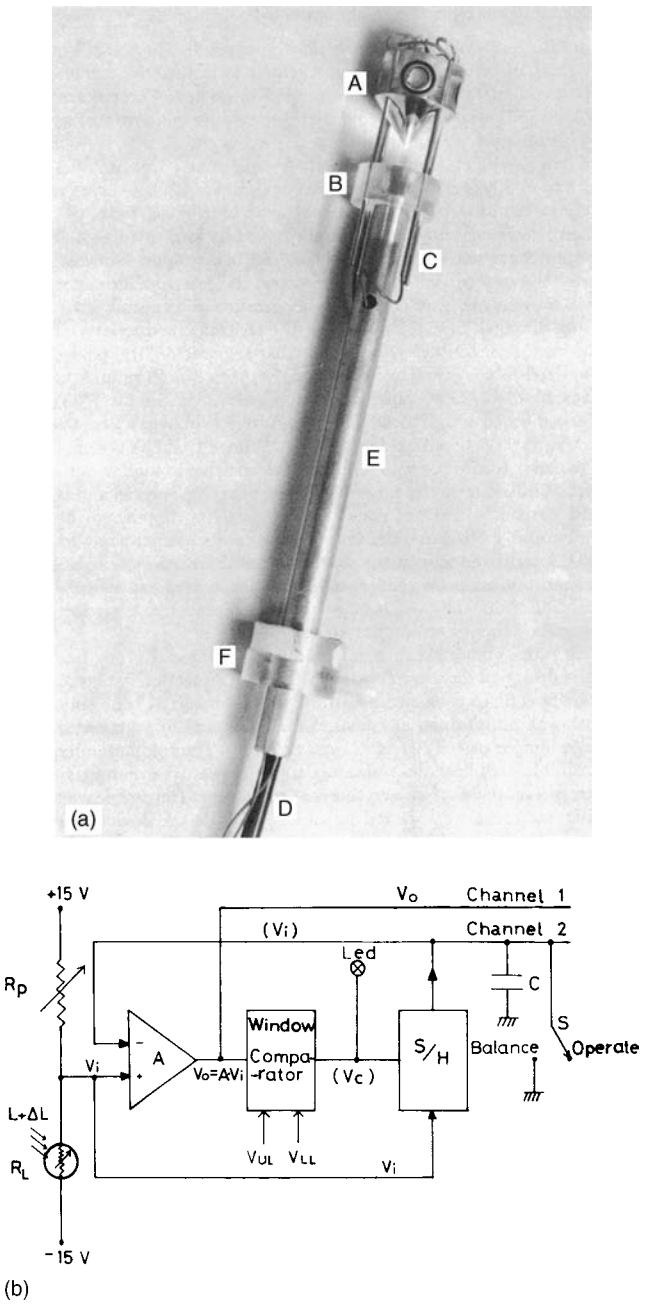


Figure 2. (a) Vaginal plethysmograph featuring a circularly symmetrical source/detector pattern. (b) Automatic baseline offset circuit. (From Ref. 4 © 1978, IEEE. Figures courtesy of the authors.)

amenorrhea had decreased vaginal blood flow as assessed via VPPG in contrast with "normals" under equivalently rated conditions of sexual arousal. Again, there was a negligible correlation between physiological and subjective ratings of arousal. Vaginal photoplethysmography VPPG has also been used to assess hematological function pre- and postoperatively for a variety of procedures.

Notably, both VBV and VPA are measured via polygraph. This results in the use of millimeters of deflection or change scores as the dependent measure in such measurements. This often results in within-study consistency, with cross-study inconsistency. Furthermore, some researchers use deviation from baseline (average over the length of preintervention assessment), while others use deviations from time-specific baselines (e.g., average of 10 s prior to next condition, which can be erotic or nonerotic).

Other problems with the use of VPPG include the reliance of the measurement on blood oxygen saturation, temperature, and response time of the light indicator. Hoon et al. (5) also discuss the relative difference of pre- and postmenopausal status, as well as sleep cycle status, as it relates to basal blood flow. Each of these affects the ability of the VPPG to reliably detect a difference within and across subjects over time. Furthermore, genital engorgement during the time of ovulation has been reported as much greater than during the 10th day of the cycle (2), suggesting hormonal levels may influence baseline genital hematology through the menstrual cycle. Notably, variations in hormone levels through both menstrual cycles and entire lifespans are related to vaginal blood flow. Similarly, increases or decreases in naturally occurring agents, such as prolactin and androgens, can change the basal levels of vaginal blood flow and alter the rapidity and duration of genital engorgement. Finally, external agents, such as SSRIs and oral contraceptives can interfere with genital blood flow. However, arousal responses appear to be minimally reliant upon the woman's hormonal levels relative to other variables, such as type of stimulation (e.g., fantasy, film, audiotape) or activity (e.g., watching a film versus masturbating).

Beck et al. (8) highlighted some of the above-mentioned problems with VPPG, emphasizing the measurement concerns related to temperature, response consistency, and drift. They assert that, at the time of publication, no devices seemed sufficient in measuring female sexual arousal.

Several recent studies have explored the relationship among parasympathetic nervous system arousal, genital arousal, and subjective arousal. Early returns suggest parasympathetic arousal may facilitate sympathetic activation in female sexual arousal. Vaginal photoplethysmography has been used to distinguish between vaginal blood flow related to sexual arousal versus sexual anxiety. The ability to differentiate anxiety-induced blood flow versus arousal-based blood flow can be quite helpful, depending on the issue being assessed. Further, sexually dysfunctional women often report significantly less autonomic arousal. One possible extension of these data would be to include treadmills and exercise bikes as potential sexual response enhancement devices, as physical exercise often leads to parasympathetic arousal.

Perhaps surprisingly, the gross differences within and across vaginas has not been shown to introduce confounds in experimental research. Given the changing shape of the vagina during sexual response cycle, researchers have found a fair degree of reliability is possible, even when the VPPG shifts or is repositioned within the vagina. However, researchers and practitioners need to heed the inherent measurement artifact introduced due to movement within subjects and differences across subjects.

Another problem in comparing VPPG study is the use of arousing stimuli, and the induction of arousal in a

laboratory setting. Several researchers have commented on the challenge to ecological validity when having volunteer undergraduates insert an acrylic tampon in a university research laboratory to assess sexual response. One improvement on this variable has been the development of a portable data collection augment to the VPPG to enhance validity. This provides for a greater sense of ecological validity, in that the research volunteers are able to pursue their sexuality in more naturalistic settings, such as their bedroom or dormitory room.

Further confounding is the type of activity employed to induce arousal in these settings. Some researchers use visual material, such as video depictions of hetero- and homosexual congress, including oral, anal, manual, and vaginal. There is empirical, marketing, and anecdotal evidence that women are less responsive to visual erotica than to other media. Some researchers have instead employed audiotaped erotica, while others direct the subjects to engage in fantasy.

Some have used the VPPG to assess responding to male- versus female-produced sexually explicit videos. In addition to these variations, other studies have included the request to digitally masturbate, use a vibrator on the vulva and clitoris (without penetration), and to attempt or achieve orgasm, or some combination of all of these. Thus, it is difficult to compare the findings across studies using the VPPG to assess arousal, as there has been so little uniformity in the experimental procedures.

Several researchers have noted that physiological arousal and subjective arousal are not synonymous One of the few studies to demonstrate a link among physiological and subjective arousal involved the use of VPPG in postoperative male-to-female transsexuals (9). Notably, they found postoperative male-to-female transsexuals demonstrated male-typical category-specific sexual arousal patterns following sex reassignment surgery. Their study included both genital and physiological responding, and discussed some of the physiological differences among natal women and sex-reassigned women, as the later may also include penile erectile tissue.

While relatively few in comparison to male sex offenders, female sex offenders appear to be receiving more attention in judicial and forensic settings. Due to their small numbers, there are few studies assessing female sexual arousal patterns using the VPPG. However, there are case studies in this area, with results likening the utility of the VPPG with the penile plethysmograph (described below).

The VPPG has its strengths and weaknesses relative to other assessment devices. Compared to self-report measures, it is more invasive, cumbersome, and difficult to use. However, it typically provides reliable and valid data that is otherwise unknown. It appears to have fallen out of favor in large scale randomized clinical trials, though is still popular for use in smaller scale and individually based studies.

Vaginal Electromyography. Engman et al. (10) used pelvic floor muscle surface electromyography (EMG) to assess for partial vaginismus (involuntary contractions of the muscles of the outer third of the vagina such that intromission is painful or impossible) and vulvar vestibulitis (a specific form of vulvar pain). They conclude that EMG is not a useful method to distinguish between asymptomatic women and women with partial vaginismus and vulvar vestibulitis. However, other researchers have found EMG to differentiate women with vaginismus from those without during basal and induced vaginismus conditions. Others have used EMG to demonstrate differences between ejaculatory and nonejaculatory orgasmic women. Electromyography has been used successfully in several studies researching etiology, treatment, and biofeedback for urinary incontinence. While this approach appears promising, there does not appear to be a consensus on the utility of this approach relative to other approaches (e.g., Vulvalgesiometer, discussed below) (Fig. 3).

Figure 3. The EMG transducer with associated detector and recording instrumentation for monitoring the contractile activity of the pubococcygens muscle. (Courtesy of Farrall Instruments, Ins., Grand Island, NE.)

Thermistor Clip and Vaginal Temperature. In a series of studies, Henson et al. (11) tested for a relationship between labial temperature and subjective ratings of arousal. They began by developing a thermistor-clip to measure changes in labial temperature, and found that labial temperature rose in response to erotic stimuli, while other measures of body temperature were unaffected. In assessments throughout the sexual arousal cycle, VPPG data and labial temperature variation were found to show corresponding variations up to the point of orgasm. The temperature remained relatively constant during orgasm, but began to rapidly decrease during the resolution phase. A follow-up study by another group using a portable vaginal temperature gauge found variation in vaginal temperature over the course of the day. These researchers hypothesized that the variations may be concomitant with circadian temperature changes, with temperature decrease in arousal being perhaps related to vaginal wall edema or a change in the position of the uterus. While Henson and colleagues made strong arguments for the use of temperature as a meaningful measure, later researchers have not followed-up on their early works.

External Measurement Devices

Vulvalgesiometer. Pukall et al. (12) introduced a new instrument for the assessment of pain in vulvar vestibulitis syndrome (VVS), which is a common form of dyspareunia in premenopausal women. Given the lack of standardization in the cotton-swab test for diagnosing VVS, Pukall et al. (12) developed a simple mechanical device, the Vulvalgesiometer, to standardize genital pain assessment. The device consists of a set of pen-like devices, each holding springs with varying resistance. The bottom of the device holds a Q-tip-type cotton swab, which is pushed down on the area being tested for pain. The pain threshold is measured by how much pressure can be exerted (or by how far the spring is compressed). The researchers reported that subjects endorsed similar pain experiences with the Vulvalgesiometer in comparison to intercourse pain. Furthermore, the authors contend that practitioners will be able to assess change over time more reliably with the Vulvalgesiometer. While still relatively new to the field, this approach shows promise as it appears relatively easy to use for both practitioner and client, as well as both reliable and valid for the assessment of vulvar pain (Fig. 4).

Vaginal Fluid Production. Levin (13) reviews the early work in the assessment of human female genital function. In his review, he discusses the measurement of vaginal pH, $p0_2$, blood flow, motility, fluid and its ionic concentrations, electrical activity, and amino acid concentrations in both arousal and basal states. Other researchers have evaluated the prevalence and elements of female ejaculate. The measurement of vaginal lubrication has been assessed through the weight gain of preweighed filter papers. Research in this area has demonstrated the correlation between vaginal lubrication and other physiological measures of arousal. Thus, while not a device *per se*, it is a measurement approach to an aspect of female sexuality. However, the lack of uniform methods of assessment and relative difficulty in obtaining the data appear to have impeded additional research in this area.

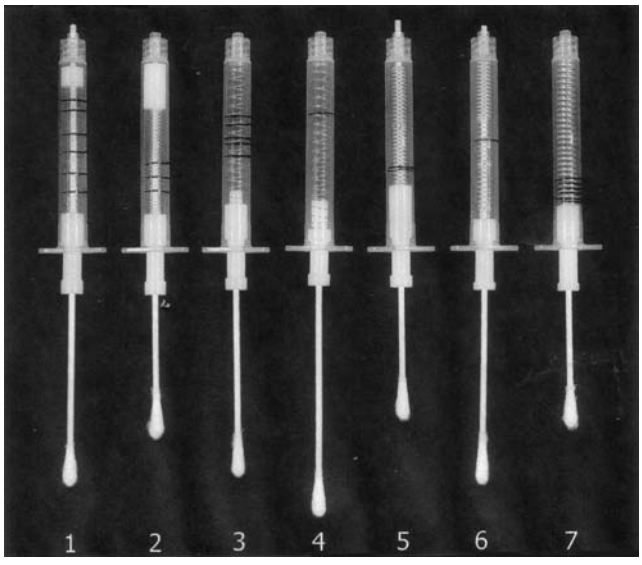


Figure 4. Figure of vulvalgesiometer.

Imaging

Duplex Doppler Ultrasonography. Munarriz et al. (14) reported on the use of duplex doppler ultrasonography (DDU) to assess genital engorgement. They used ultrasonography combined with the Doppler effect to assess clitoral and corpus spongiosum diameter, as well as clitoral and corpus spongiosum peak systolic and end-diastolic velocity. The Doppler effect can show whether blood is flowing toward or away from the device, while the ultrasound helps for visualization. They concluded the use of ultrasonography appears to be a less invasive and a consumer-friendly approach to assess genital engorgement in comparison with VPPG. Others have used DDU to assess changes in clitoral engorgement in a study of Alprostadil. Relative to other approaches, such as self-report, DDU is more time and cost-intensive and less invasive than VPPG.

Magnetic Resonance Imaging. Rosen (15) reports on the promising developments in the use of magnetic resonance imaging (MRI) to assess genital engorgement and response in human sexuality. However, there are few published studies in this promising area in the assessment of human female sexuality. Like the DDU, it is anticipated that with additional research, the MRI will be rated as more consumer-friendly than VPPG, and will likely be more cost-intensive than self-report measures of sexual arousal and genital engorgement.

FEMALE SEXUAL BEHAVIOR TREATMENT

Devices and Instruments

Vaginal Dilators. Vaginal dilators are generally a set of graduated glass tubes or cylinders, 2–5 cm in diameter and 15 cm in length. Typically, vaginal dilators are used for the treatment of vaginismus, which is the involuntary muscle contraction of the outer third of the vagina.

With the advent of plastics, nonbreakable dilators are preferable. Furthermore, while vaginal dilation had historically been conducted under general anesthesia, it is now being offered as an outpatient referral for self-administration at home. A recent study using vaginal dilators for in-home use lists their instructions as: These dilators are to relax the muscles around the entrance to the vagina and to gently stretch the area. You should start using the smallest of the dilators, which should be inserted for 10–15 min·day^{-1} and preferably for two episodes per day of 10–15 min, passing the dilator downward and backward into your vagina. Light lubrication with K-Y jelly or similar is advised. After 1 week you should increase the size of the dilator to the next biggest and increase the size by one every week until you are using the largest of the dilators or you feel comfortable to resume intercourse, whichever happens first. The time that the vaginal dilator should spend in the vagina is the same on each occasion (i.e., 10–15 min).

They also note: "Sometimes it was necessary to help the patient insert the first dilator in the clinic explaining the need to relax her pubococcygeal muscles whiles advancing the tube during muscle relaxation" (16) (Fig. 5).

Figure 5. Vaginal dilator (17). (Courtesy of www.vagenesmus.com.)

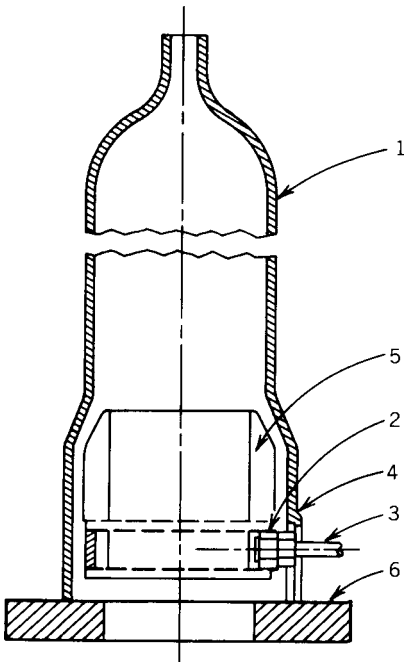


Figure 6. The Freund phalloplethysmograph: (1) glass cylinder, (2) plastic ring, (3) metal tube with threads, (4) lock nut, (5) rubber cuff, (6) flat, soft sponge rubber ring. (From Freund et al. (22). Copyright 1965 by the Society for the Experimental Analysis of Behavior, Inc. Reprinted by permission.)

Several researchers have noted the benefit of pairing muscle relaxation exercises, and specifically pelvic floor muscle relaxation, as a beneficial adjunct to vaginal dilator therapy. Weiss (18) describes the use of dilators in the treatment of vaginismus. She presents a case study in which, due to economic circumstances, an assortment of vegetables was substituted successfully in the treatment protocol, with positive long-term results. Dilators appear to be one of the most commonly used treatment approaches for vaginismus. For more on the etiology, diagnosis, and interventions related to vaginismus, please refer to Koehler (19). Dilators have also been used in several cases of vaginal construction or reconstruction. However, there is some variability in practice using dilators in this way.

Clitoral Vacuum. Billups et al. (20) introduced the Eros clitoral vacuum for the treatment of female sexual dysfunction. The device is a small battery-powered device designed to enhance clitoral engorgement, increase blood flow to the clitoris, and ultimately to work toward increased sexual arousal and response. It looks a bit like a computer mouse and has a small suction cup to be placed over the clitoris. In follow-up studies, this device was shown to increase genital and clitoral engorgement. Participants self-reported improved libido, arousal, lubrication, orgasm, and satisfaction, as well as decreased pain. Gynecological exams revealed improved mucosal color and moisture, vaginal elasticity, and decreased ulceration and bleeding (Fig. 6, Ref. 21).

Clitstim and Vibrotactile Stimulation. Riley and Riley (2003) (23) present a small study on the use of a finger-cot shaped device (sometimes called a vibrotactile genital stimulator) designed to increase digital clitoral stimulation when worn on a finger. The device is similar in nature to a nonmedical clitoral vibrator, but its unobtrusive design was created with the hopes that it would be more socially acceptable. Based on their results, the researchers conclude that the device may be of assistance in anorgasmic or slow-to-orgasm women.

Over-the-counter vibrotactile devices (e.g., vibrators) are the most commonly used sexual instruments for anorgasmic and slow-to-orgasm women. Vibrotactile devices have also been developed for use with men, though the literature in this area is sparse.

IUD with Danazol. Cobellis et al. (24) report on the successful use of an IUD loaded with danazol for the treatment of dysmenorrhea, pelvic pain, and dyspareunia associated with endometriosis. Danazol had previously been used as an oral treatment for these ailments, however, this route of administration led to side effects. Cobellis et al. (24) found that the release of the danazol through the IUD eliminated these side effects (only adding a common IUD side-effect of first-month spotting). They reported 6 month duration of benefit with generally favorable consumer satisfaction.

Treatment Articles. There are several articles reviewing the treatment of sexual dysfunction in the human female without the use of specific instruments or devices. These include the use of botulin toxin for vaginismus, the role of fantasy training, androgen therapy, and cognitive-behavioral therapy including progressive muscle

Table 1. Female Sexual Behavior Assessment

Device	Function	References
VPPG	Measures changes in blood volume	2,5
EMG	Tests for vaginismus and vulvar vestibulitis	10
Thermistor clip	Measures labial temperature	11
Vulvalgesiometer	Measures pain associated with vulvar vestibulitis syndrome	12
DDU	Measures genital engorgement	14
MRI	Measures genital engorgement	15
Vaginal dilators	Treats vaginismus	16
Clitoral Vacuum (EROS therapy)	Increase genital engorgement	20
IUD with danazol	Treats dysmenorrhea	24

relaxation and meditation. Many of the more recent treatment outcome studies limit their dependent measures to self-report, often ignoring the physiological assessment that some researchers see as necessary and vital. Beck (25) reviewed the theories of etiology, prevalence estimates, and theories related to hypoactive sexual desire disorder. She concluded that a lack of valid, reliable, and consumer-satisfactory dependent measures impede scientific progress in this area.

Self-Report. For a review of self-report measures of female sexual function and behavior, please refer to Althof et al. (26) (Table 1).

INSTRUMENTS AND MEASUREMENT OF MALE HUMAN SEXUAL BEHAVIOR

Erectile Dysfunction and Sexual Deviance

The measurements of and devices related to male sexual behavior typically fall into one of two categories: erectile dysfunction and deviant sexual arousal. Coleman (27) reviews the scientific literature on the etiology and intervention for Erectile Dysfunction (ED), which is defined as the persistent inability to achieve or maintain an erection sufficient for satisfactory sexual activity. It is estimated that $> 50\%$ of men of 65 years-of-age experience ED. Meuleman and Van Lankveld (28) note that with the increasing availability of pharmacological interventions for ED, male hypoactive sexual desire disorder may be commonly misdiagnosed. They conclude, "HSDD is more common in men than in women. In public opinion and in medical practice, HSDD is often misinterpreted as ED, and treated as such. There is a need for physicians and patients to be educated, and for the development of reliable clinical tools to assess this aspect of male sexual function (294). The tools used for the assessment and treatment of ED are described below.

The second area of assessment and intervention is toward deviant sexual interest and arousal: most notably, pedophilia. For the assessment of sexual interest, researchers have often turned away from face-valid self-reports and looked to physiological data. However, as with the women discussed above, it has been shown that physiological arousal is not uniformly correlated to emotional or cognitive arousal.

An area of concern in this topic is the limits to ecological validity inherent in university- or medical-center-based research on male sexuality. Just as the women above are presumed to rarely masturbate with photoplethysmographs inserted, it is presumed that most men do not wear penile Strain Gauges or plethysmographs when engaging in sexual activity (see Ref. 29 for more in this area). Again, more recent researchers have adopted intent-to-treat or wait-list controls to enhance the ecological validity of their research.

MALE ERECTILE DYSFUNCTION

Assessment

Circumferential versus Volumetric Assessment. The early research on male erectile response typically used volumetric assessment. Volumetric assessment usually involves the placement of a flaccid penis in an airtight cylinder with a monitored release valve. Engorgement leads to displacement of air out the valve, leading to an indication of volume displacement due to erection.

Circumferential assessment often involves the placement of a mercury-filled rubberized band around the shaft of the penis near the base. Engorgement typically leads to expanding girth, resulting in millimeter displacement as measured by an accompanying graphing device (Figs. 7 and 8).

In a comparison of volumetric and circumferential measures of penile erection, Wheeler and Rubin (30) found that the circumferential method was preferable for several reasons. While both measures showed a correlation, the authors contend the volumetric approach produced more artifacts, was more difficult to use, and was no more sensitive than the easier-to-use circumferential method. However, a rapid change in length may result in a net decrease in girth, at least temporarily. Taken together, information on the length and width of penile tumescence are consistent with recommendations to take measures over time to assess the level, course, and trend of the observed response.

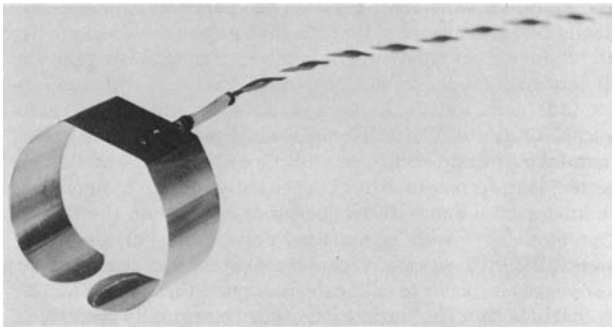


Figure 7. Barlow-type strain gage for monitoring penile tumescence. (Courtesy of Farrall Instruments, Inc., Grand Island, NE.)

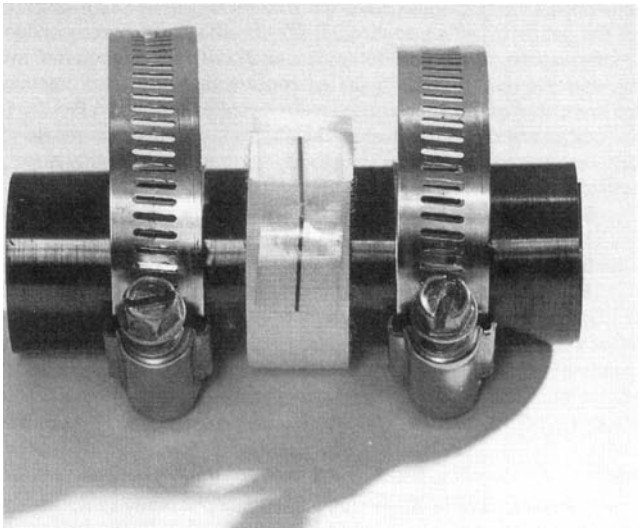


Figure 8. Measuring graphing device for expanding girth.

The assessment of ED often involves the use of a snap gauge, which is an application of multiple circumferential measures (there are several brands and types available, e.g., Dacomed's Snap Gauge, the RigiScan). In essence, this measures a sample of volumetric expansion, which allows for the assessment of both rigidity and circumference. Snap gauges often involve the placement of a band or sleeve around the shaft of the penis, with two or more expansion-specific snaps placed along the length of the gauge. Thus, as the penis expands from the base to tip, the snaps are presumed to snap in sequence dependent on the length of turgidity.

A common assessment avenue for ED is the use of a snap gauge during sleep. Typically, adult males enter REM stage three to four times per night, achieving at least one erection during each REM stage. The sleep-related erections are termed nocturnal penile tumescence (NPT). Presence of NPT is suggestive of psychogenic ED, while an absence of NPT is often inferred as organic ED.

Libman et al. (31) discussed the issues and problems with the reliance on NPT and snap gauges to assess for erectile dysfunction. These authors emphasize the growing body of research indicating waking erectile capacity is not directly related to NPT, concluding there may be different processes underlying NPT and erotically induced erection. Further, Libman noted several subjects balked when presented with the notion of "snap" and "penis" in the same sentence. They altered the description of the procedure to impotence testing and expansion tape. In their study, they confirmed that NPT, as assessed by a snap gauge, did not constitute a valid measure of daytime erectile ability in their sample. However, they provide recommendations to enhance the utility of the assessment, including assessment during waking and sleeping states, as well as a multimodal approach.

The Snap Gauge (Dacomed Inc.) is described by Diedrich et al. (32): A commercial device, which, according to the manufacturer, can be used for the determination of organic impotence. The Snap Gauge consists of nonstretchable fabric band that is fastened around midshaft of the penis by Velcro straps. Three plastic snap elements are attached to the device such that they break in sequence if an erection of sufficient circumferential expansion and local hardness is obtained. A male using the device may break none, one, two, or three of these elements corresponding to the sufficiency of his erection.

It is possible to judge a person's ability to reach sufficient tumescence for intercourse by looking at how many snaps remained intact. With all three intact, no intercourse would be possible. With all three unsnapped, the erection would be insufficient for intercourse. If two snaps remain, it is questionable whether the subject has reached sufficient rigidity (Fig. 9, Ref. 32).

Nobre et al. (33) found a similar discordance between physiological and subjective ratings of arousal. Generally, they found men in their sample to underreport their arousal in relation to physiological measures of tumescence. Notably, in their study, participants were unable to manually or visually assess their erections. The researchers noted that positive affect facilitated arousal and was predictive of subjective and physiological arousal estimate concordance.

Strain Gauge. The measurement of human penile tumescence was one of the first review articles describing the construction and measurement issues associated with the armamentarium for the assessment of male sexual arousal. Rosen and Keefe (34) recommend volumetric assessment when precision is important (e.g., specific research question) and circumferential assessment for efficiency and ease of use. They cite the mercury-in-rubber as an optimal tool for the measurement of circumference.

Strain Gauge measurement (a component in PPG) is often conducted via a method involving millimeter displacement as a percentile of maximum erection. In this procedure, the participant places the gauge on the flaccid penis. They are then directed to self-stimulate to a full erection. The point of maximum tumescence is then taken as 100% erection, with later erections being taken as a proportion. Notably, this procedure allows for erections of $>100\%$. There is also an inherent floor effect, in that the lab setting allows for a measurement of flaccidity that does not measure any aversive or derousing responding (33).

The Dacomed Snap Gauge was evaluated by Diedrich et al. (32) using both mechanical and human subject tumescence. In the biomechanical procedure, a device was manufactured that allowed for the incremental increase in circumference (1 mm intervals). The Snap Gauge was placed on the mechanical device, as well as a mercury Strain Gauge. With the human participants, a mercury gauge was placed near the base of the penis, with the Snap Gauge at mid-shaft. Participants were able to manually and visually assess their erections, and were asked to report as the Snap Gauge snaps broke, as well as to report their sufficiency for penetration following the presentation of sexually explicit videotape. The authors noted that the Snap Gauge appears to be lacking in validity based on three problematic deficiencies: the lack of erection uniformity, lack of placement uniformity, and lack of uniform rigidity. The authors direct practitioners to proceed

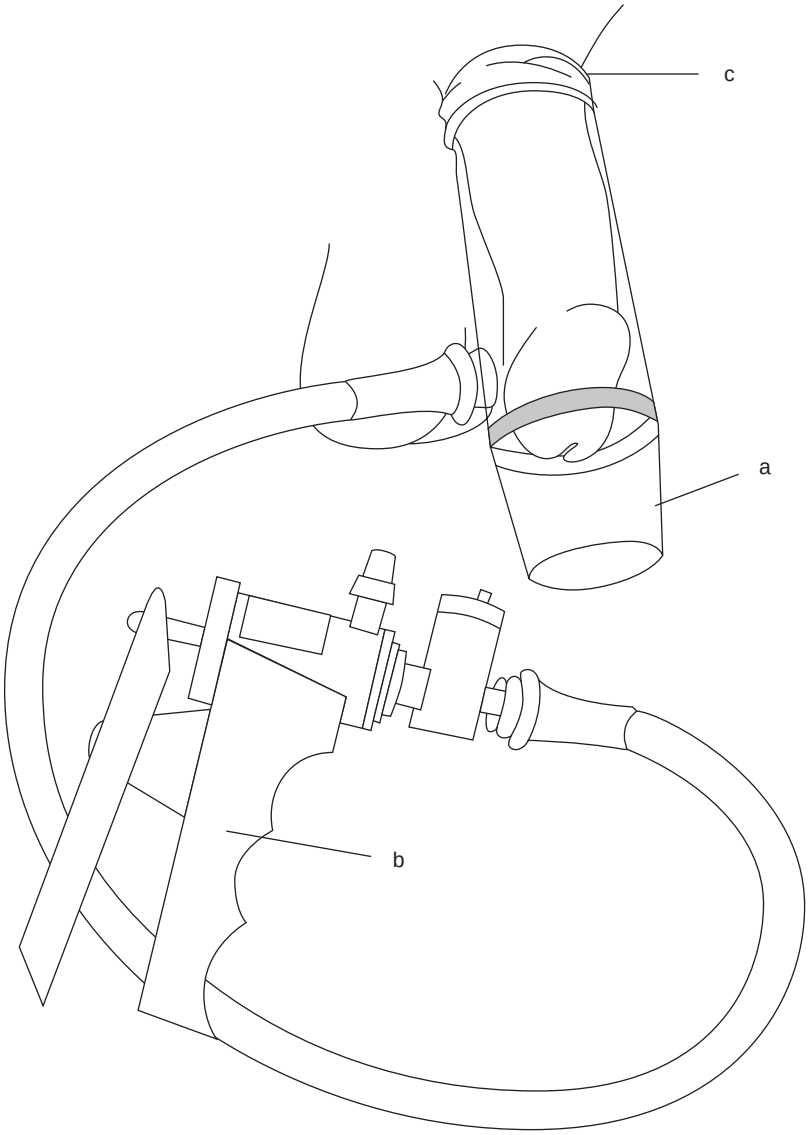


Figure 9. Snap Gauge (Dacomed Inc.).

with caution when diagnosing ED based upon Snap Gauge data, as in their analysis, the Snap Gauge was subject to both false positives and negatives.

In comparison studies of the RigiScan and Snap Gauge in the assessment of penile rigidity, researchers have concluded the Snap Gauge is more cost-effective than and as reliable as the more complicated RigiScan. They recommend the use of The Snap Gauge primarily, with the RigiScan being recommended for clients when The Snap Gauge is inconclusive or when assessment that is more detailed is necessary (e.g., per research protocol).

Stamp Test. One of the simplest, least expensive, and perhaps oldest tests for ED is known as the stamp test. The stamp test is used to assess NPT. In this procedure, the subject fastens postage stamps around the base of the penis, similarly to where the snaps would be found on a snap gauge. The NPT is likely to displace the stamps. The stamp test was developed before the widespread use of self-adhesive stamps and instead used the moisture-catalyzed gumming adhesive. Moreover, while this is relatively inexpensive, the published reports provide two false positives and one false negative in a rather small sample.

Electromyography. Da Silva et al. (35) found that EMG could be useful in the differential diagnosis of ED. In this clinician-administered procedure, EMG inserts electrode needles into the muscle, and when the subject is asked to move that muscle, the EMG gives a general picture of the muscle activity by showing the action potentials occurring in the specific surrounding muscle cells. This makes it possible to see whether the dysfunction is indeed physiological. However, few researchers have completed follow-up studies on this promising work.

Imaging

Near-Infrared Spectroscopy. Burnett et al. (36) compared near-infrared spectroscopy (NIS) with color duplex ultrasonography, Strain Gauge circumference measure,

penile tonometry, and clinical assessments. They found NIS to be a safe, inexpensive, and easy-to-use device that provides for quantitative measurements of vascular physiology in erectile assessment, by measuring the percentage of blood volume reaching the penis, and monitoring the circulation. Again, few researchers have continued in this line of research.

Self-Report. A review of self-report measures for the assessment of male sexual behavior can be found at Berman et al. (37) and Kalmus and Beech (38).

THE TREATMENT OF ED

Devices and Instruments

With the advent and marketing of several pharmacological agents for the treatment of ED, some authors have questioned the utility of mechanical devices and surgical interventions. There have been several cases of adverse responses from pharmacological interventions, and there are several conditions for which these medications are contraindicated. For these individuals, surgical or mechanical intervention may be their only hope for a reliable and useful erection.

Vacuum Constrictive Device. The vacuum constrictive device (VCD) typically includes an acrylic cylinder into which the flaccid penis is inserted. The VCD has a manually squeezed bulb pump for the forced suction of air out of the cylinder (up to 250 mmHg, 33.33 kPa), which in turn leads to increased engorgement. A constrictive band is then placed at the base of the penis, with the intent of maintaining the engorgement through the completion of the sexual act by reducing, not preventing, penile venous outflow. Some pumps allow for the placement of the constrictive band (similar to what is sometimes called a "cockring") while the pump is in place; others involve the removal of the pump and the placement of the band prior to detumescence. A device similar to the VCD (albeit satirized) played a supporting role in the first of the Austin Powers films. Wylie and Steward describe a homemade device for ED, for use when pharmacotherapy is contraindicated (39). In their case study, a 65 year old client fashioned his own VCD for the effective creation of erections (Fig. 10, Ref. 40).

Several researchers have commented on the utility of the VCD, with successes reported with diabetic and neuropathic populations, as well as with psychogenic ED.

Researchers have also found that the device is generally more effective when it is combined with couples' therapy. However, several complications have also been discussed. Some men have complained of feelings of demoralization, discouragement, and confusion with the device, while others reported frustration when their spouses were not supportive or helpful when using VCD. Furthermore, some men report that even though they can create an erection, they are not necessarily emotionally aroused. Lastly, as the VCD involves the restriction of blood flow out of the penis, possible problems include hemotoma, skin irritation, and in at least one reported case, penile gangrene. This complication underscores the need for proper instruction and compliance with the device.

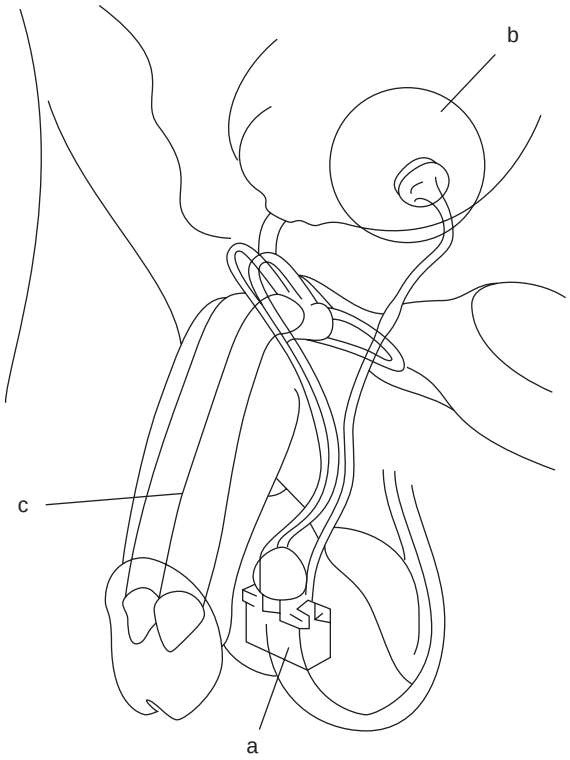


Figure 10. The vacuum constrictive device (VCD) sucks air out of a cylinder to increase penile engorgement.

Penile Prosthetics. Penile implants were some of the earliest treatments for ED. While practitioners now recommend nonsurgical interventions (e.g., pharmacological or the VCD treatment), there are still some clients for whom a prosthetic implant is preferable. Penile prosthetics have also been used successfully in the creation of neophallus in sex reassignment surgery, and in reconstructive surgery following injury.

Penile prosthetics are typically either malleable or inflatable. Malleable implants are basically permanent, semirigid tubes placed within the penis. When an erection is desired, the man (or partner) positions the penis into the desired direction. Inflatable prosthetics often involve two or more components. A fluid-filled pump reservoir is implanted in the scrotal sac, with hollow tubes (initially one, now two or more in connection) placed within the penile shaft. When an erection is desired, the man or partner squeezes the pump, leading to an outflow of fluid from the reservoir into the tubes, resulting in an erection. Surgically, the procedure can involve the removal of cavernous tissue. With the placement of prosthetic, it may result in fibrous involvement, scarring, or both. These complications can result in difficulty changing implants if needed, and preventing future "natural" erections. A fenestrated implant has been successful in the sparing of cavernous tissue and spontaneous erections.

Follow-up studies on penile prosthetics show high rates of consumer satisfaction, with occasional problems with

Table 2. Erectile Dysfunction

Device	Function	Reference
Snap Gauge	Measures penile rigidity and circumference	32
Strain Gauge	Measures penile circumference via displacement	34
EMG	Tests for physiological penile dysfunction	35
NIS	Measures penile blood volume	36
VCD	Treats ED by increasing engorgement	39
Penile Prosthetics	Surgical treatment for ED	

leakage, breakage, infection, and stretching due to overuse. A common partner complaint is the lack of girth; other complaints have included spontaneous deflation, a cold penis, and that intercourse felt unnatural. Malleable implants are reputed to be easier to install, with inflatable prosthetics typically rated higher by consumers and their partners (Table 2).

Summary. Several researchers have explored the effectiveness, advantages and disadvantages of pharmacological, psychotherapeutic, surgical, and mechanical treatment of ED, and in combinations of these. Taken together, there is no clear gold standard for the treatment of ED; rather, there are several tools available to the clinician and consumer that may be appropriate given the client's physiology, psychology, and context.

DEVIANT PHYSIOLOGICAL AROUSAL

Assessment

The second major area of inquiry into male sexual behavior relates to physiological changes related to deviant sexual arousal. The following section describes the use of the penile plethysmograph (PPG), visual tracking (VT), and pupillometry (PM) to assess differential arousal in response to presentation of varied stimuli. The bulk of this work centers on deviant sexual arousal, typically with adult males being shown or presented videotapes, still images, or audio recordings. The stimuli typically depict age-appropriate as well as age-inappropriate stimuli, and may include other challenges, such as sadistic themes, adult rape themes, fetishes, bondage or sadomasochistic themes, or both, and control conditions with no intended sexual content.

While in Freund's early phallometry work (41) PPG typically assessed volume, the bulk of the last half-century's research using PPG seems to be reliant upon circumferential assessment. Bancroft et al. (42) describe the construction, provide schematics, and a photograph of a simple transducer for measuring penile tumescence. The device is a mercury and rubber strain gauge, which they describe is inexpensive, portable, easy to apply, and unobtrusive in use. They contrast the simple strain gauge with the more complex volumetric PPG used by Freund (described earlier, please refer to earlier figures of penile plethysmography devices).

Penile Plethysmograph. A PPG is a strain gauge (a stretchable band filled with mercury that is fitted around a subject's penis, discussed earlier in this article), connected to a video screen and data recorder, which records changes in the penis with different stimulus. While PPG can assess both volumetric and circumferential changes, PPG using measures of circumference changes appears to be more widely used. Advances in technology have led to the advent of computer-based assessment and scoring in penile plethysmography. O'Donohue and Letourneau (43) concluded, "there does not appear to be a standardized penile tumescence assessment, but rather there is a family of procedures that share some common aims and features (p. 126)." The 17 potential sources of variation in the assessment that O'Donohue and Letourneau described have not been satisfactorily addressed in most recent reviews. Still, some researchers use an average of pen deflection; some use a sampled interval average, while some use maximum response and percentage of maximum response. Further, as measurement is often less reliable at the ends of the metric, there was some discussion about the reliability and validity of PPG at large expansions.

Kalmus and Beech (38) reviewed the self-report and physiological assessment of male sexual response. They noted that PPG is still the most common method for such assessment, though to acknowledge it is prone to a number of possible artifacts, including faking or intentional suppression. Other clinicians have noted a possible retest habituation effect, which may be an artifact of attempts at standardization. However, some authors contended the PPG is useful in repeated assessments, provided the stimuli are varied and content matched to reduce habituation.

Golde et al. (44) assessed participants' ability to suppress penile responding under audio-only and audiovisual stimulus conditions. They also included measures common to the polygraph (discussed in a later section polygraph), galvanic skin response (GSR), and finger pulse amplitude (FPA). The authors found participants were able to suppress penile arousal while not providing indicators of deception as measured through GSR and FPA. Both naïve and experienced participants were able to suppress, and arousal was less pronounced in audio-only conditions. These data suggest the PPG is susceptible to faking through suppression, especially in the audio-only presentation.

Other evidence of altered penile tumescence comes in treatment studies using the PPG as a dependent measure. Rosen and Kopel (45) demonstrated that with multiple sessions of biofeedback, penile tumescence to undesired sexual stimuli was reduced with lasting effects. The authors contended the result was not due to habituation or generalized suppression. Follow-up data indicated the case study subject had been deceptive in his self-report and had renewed the targeted sexual behavior. This research vignette highlights the oft-noted weakness of deception when using self-report data.

In a study comparing adult males alleged to have sexually offended against children and "normal controls",

Haywood et al. (46) assessed the relationship between physiological arousal as measured by PPG and subjective arousal assessed through self-report. The alleged sex offenders showed more subjective arousal to children, with nonincest offenders reporting higher subjective arousal than incest offenders. Notably, both alleged offenders and normals reported subjective arousal without physiological responding, and subjective repulsion, despite showing penile responding. The authors recommend for the continued used of subjective and physiological measures in the assessment of sexual response.

Other concerns with the PPG include the ethical use of such a device in correctional or forensic settings, including the use of PPG as a treatment device for the challenging of sexual offenders reported responses and arousal patterns. There are also cautions against the use of PPG for predictive purposes, or for reasons beyond which it has been demonstrated to be reliable and valid.

Recent work has explored the use of portable circumferential PPG. Rea et al. (47) found the data taken inside the lab to generally correspond with data taken outside the lab. Not surprisingly, there appeared to be an inverse relationship between penile tumescence and proximity of researcher.

Finally, Mathews et al. (48) reviewed the arguments and problems with the use of PPG in legal settings. They discussed the Frye and Daubert standards of evidence as it relates to the PPG. These standards describe requirements of a measure before it is considered acceptable in evidentiary purposes. A description of these standards can be found in the O'Donohue and Levensky edited collection, *The Handbook of Forensic Psychology* (49). Given the above findings indicating the PPG can be suppressed, and be subject to false positives and negatives, Mathews et al. (48) contended the PPG should not be admitted as evidence for forensic purposes.

Pupillometry. Pupil dilation is indicative of arousal, including sexual arousal. Pupillometry uses a magnified video recorder to track a subject's pupil changes when presented to different stimuli. While the results using pupillometry originally seemed promising, the research has been criticized for a variety of methodological problems, and appears to have lost favor as an avenue of inquiry.

Viewing Time. In viewing time studies, participants are presented visual stimulus materials and allowed to view the materials. The instructions vary across studies as to how long and how much liberty the participant has to view the materials (e.g., in some, the participants are allowed to choose display times, while in others, visual field is tracked). The principle behind this approach is that participants will view images that hold more sexual interest to them for longer periods than those without such appeal.

Abel et al. (50) reported on the development of the "Abel Screen", also known as the Abel Assessment for Sexual Interest (AASI). This device is reported to correspond well with the more intrusive volumetric assessments, and to be more efficient. Laws and Gress (51) review the multiple criticisms of this specific instrument, although they also acknowledge the benefit of VT-based assessments. The Affinity [Glasgow et al. (52); Kalmus and Beech (38)] is a VT device designed for use with learning-disabled offenders. While there are minimal data to support its use, the existing data are promising.

The Abel Screen [cf. Gaither (53)] was created in response to some of the criticisms of plethysmography. The Abel Screen is different in that it measures attention (as opposed to tumescence or vasocongestion) as measured by tracking visual focus on stimulus objects. The Abel Screen was developed in part to eliminate the need for nude slides as stimulus materials. By using the Abel Screen, clinicians are able to assess the focus of the respondent's attention, and the duration of that attention. Further, the Abel Screen is less time intensive, and less intrusive than plethysmography. While Abel and colleagues report that visual reaction time is a reliable (alphas 0.84–0.90 across stimuli) and valid means of assessing interest, Gaither's data was mixed in an analogue study using undergraduate students. Other researchers have criticized VT studies for their clear face validity, which makes the VT vulnerable to faking. The device consists of a questionnaire, which tests for things such as deviant and inappropriate sexual behaviors, and a program on a computer, in which the participant is to rate the slides on a seven-point Likert scale from "highly sexually disgusting" to "highly sexually arousing". The computer tracks how long it takes for the subject to rate the slide and advance to the next slide, thus tracking how long the subject looks at the slide. Abel and colleagues also state that measures have been taken to identify offenders attempting to conceal their offenses (abelscreen.com) (54), and even with deniers Abel and colleagues contend device is able to detect 88% of sexual abusers.

Laws and Gress (51) reported on the development of computer-generated images in a standardized assessment using PPG, VT, and Choice Reaction Time. They posited this approach as a stopgap measure in the advancement of the inclusion of new technologies to assist in the assessment and monitoring of deviant sexual arousal. They argued that developments in virtual reality might provide avenues of worthy inquiry by posing more realistic challenges to relapse prevention plans while addressing community safety and privacy. With increases in technology and privacy regulations, many researchers are turning to computer-generated composite images to further enhance and clarify sexual arousal patterns. In doing so, researchers may be able to provide for individually based stimulus materials that would maximize the likelihood of deviant arousal and more realistic virtual reality scenarios to provide therapeutic challenges for sex offender treatment skills application.

In a comparison of Abel's visual reaction time (VRT) and PPG with audiostimuli, Letourneau (55) revealed interesting data: both identified offenders against young boys, and neither reliably identified offenders against adult women. Visual reaction time identified offenders against young girls (although not reliably), and surprisingly, PPG data indicated men with female child victims produced significantly lower levels of arousal than men in other offence categories.

Thermistor, Temperature, and Photoelectric Surface Blood Volume. Temperature measurements, indicative of penile and groin blood flow, have not been supported as valid and reliable measures of arousal due to the latency involved in tumescence and detumescence (38). Previous reviews discussed the possibility of telemetric temperature dissipation assessment as an avenue of physiological arousal assessment, however, there are few publications promoting its use in this way. While photoelectric blood volume research is the basis for VPPG in women, it has not born fruit in assessment of male sexual response (38).

Electroencephalogram Movement. Cohen (56) presented one of the few research projects in which the sex organ between the ears was assessed. Their preliminary results suggested an electroencephalogram (EEG) can be used to assess sexual arousal, and that responding was different based on stimulus modality. Although limited, a promising area of inquiry is the EEG measurement of contingent negative variation. Researchers in this area noted a presumed difficulty in faking EEG response. However, given the limited data in this area, caution is warranted in using this approach in the assessment of sexual preference (38).

Galvanic Skin Response. The relatively few studies in this area suggest galvanic skin response (GSR) may be a useful augment to PPG in assessment in which the respondent is showing minimal response or is hypothesized to be faking or suppressing a response. Typically, research in this area is based on the concurrent use of PPG and polygraph.

Polygraphy. Oksol and O'Donohue (57) provided a thorough review of the use of the polygraph in a forensic setting. In their review, they described polygraphy as the evaluation of physiologic reactions that purportedly occur in response to the emotions of fear or conflict, or are in some other way associated with lying. The polygraph measures a number of subtle and involuntary changes in physiological functions, such as heart rate, skin resistance, and blood pressure. By amplifying and recording autonomic functions on a multichannel instrument, the polygraph detects the changes in these functions. It is so named because it has many (poly) pens, with each pen measuring and recording (graphing) a different physiological response. Some of these physiologic responses are recorded on a polygraph chart and the polygrapher interprets these changes to be indicative of truthfulness or deception.

Over the years, there have been many settings in which polygraph examinations have played an important role. These settings include criminal investigations of suspects accused of theft, rape, murder, or lesser crimes. Recently, a number of criminal justice and treatment professionals have been advocates of the increased use of polygraph tests to assess child sexual abuse allegations and have implemented polygraph testing in their practices or treatment programs. For example, polygraphers have been called to assess the verity of a subject's self-reported masturbation patterns, fantasy content, and truthfulness in treatment.

Contrary to lay opinion, there is no such thing as a "typical polygraph test". In reality, the polygraph examination varies immeasurably from operator to operator. From formulating the questions that will be asked of the suspect to scoring the results of the physiological responses, polygraph techniques are not standardized. Rather, each polygraph administration is a complex and highly variable conglomerate of choice points and interview situations. Polygraphers choose from at least eight different question formats of lie detection, with each having its own psychometric properties (e.g., accuracy rates, reliability).

Oksol and O'Donohue (57) highlighted the multitude of problems facing the forensic use of the polygraph. Most problematic are the lack of uniformity in the procedures, the lack of reliability, and the lack of validity, in that the polygraph has not been able to reliably demonstrate that it assesses what its proponents argue the polygraph assess.

TREATMENT OF SEXUAL DEVIANCE

There are not any accepted devices or instruments for the treatment of sexual deviance, other than the assessment devices described above. For a review of the issues, controversies, and methodology in the treatment of sexual deviance, please refer to Sbraga (58) (Table 3).

Summary. There are several devices and instruments available to the clinician assessing and treating human sexual function. A common theme throughout the measures of human sexuality is the lack of concurrence among physiological and self-report data. It seems as though self-report measures are more satisfactory to the consumer, and perhaps have better psychometric properties. As a result, more and more outcome studies are relying on self-report data for their ease of use and psychometric superiority to physiological devices and instruments.

While there are several reliable, valid, and useful devices and instruments for the assessment and treatment of human sexuality issues, there are some shortcomings that limit the use of these tools. A common theme in review articles is a lack of uniformity in the psychophysiological assessment of human sexual behavior. This lack of uniformity is likely one component in the relatively lower reliability in comparison with self-report data. In addition, many of the research studies in this area have used

Table 3. Deviant Sexual Arousal

Device	Function	Reference
PPG	Measures circumference changes with different stimuli	43
Pupillometry	Tracks pupil dilation with different stimuli	
Abel Screen	Tracks visual tracking with different stimuli	59
EEG Movement	Assesses sexual arousal stimuli	56
GSR	Measures skin response with different stimuli	
Polygraphy	Measures subtle physiological changes associated with lying	57

volunteers, which is problematic due to the demonstrated differences among human sexuality research volunteers and normals. However, some gains have been made in increasing the ecological validity of human sexuality research. Taken together, the above data are consistent with the recommendation for a multimodal assessment, addressing the physiological and psychological (cognitive and emotional) components of human sexuality. Consumer satisfaction with medical devices and instruments appears to be driven by ease of use, partner support, and efficaciousness. Technological advances may increase the precision with which the assessment and treatment of the above concerns can be made; however, a lack of cohesiveness in the field may serve to perpetuate the aforementioned problems.

BIBLIOGRAPHY

1. Semmlow JL. Sexual instrumentation. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation. New York: John Wiley & Sons, Inc.; 1988.
2. Hatch JP. Vaginal photoplethysmography: Methodological considerations. Arch Sexual Behav 1979;8(4):357–374.
3. Sintchak G, Geer JH. A vaginal plethysmography system. Psychophysiology 1975;12:113.
4. Ormon H, Weinnian J, Weinstcin D. A vaginal photoplethysmography transdueer. IEEE Trans Biomed Eng BME 1978; 25:434.
5. Hoon PW, William D, Laughter JS. Infrared vaginal photoplethysmography: Construction, calibration, and sources of artifact. Behav Assess 1984;6(2):141–152.
6. Geer JH, Morokoff P, Greenwood P. Sexual arousal in women: The development of a measurement device for vaginal blood volume. Arch Sexual Behav 1974;3(6):559–564.
7. Van Dam FS, Honnebier WJ, van Zalinge EA, Barendregt JT. Sexual arousal measured by photoplethysmography. Behav Eng 1976;3(4):97–101.
8. Beck JG, Sakheim DK, Barlow DH. Operating characteristics of the vaginal photoplethysmograph: Some implications for its use. Arch Sexual Behav 1983;12(1):43–58.
9. Lawrence A, Latty E, Chivers M, Bailey J. Measurement of sexual arousal in postoperative male-to-female transsexuals using vaginal photoplethysmography. Arch Sexual Behav 2005;34(2):135–145.
10. Engman M, Lindehammer H, Wijma B. Surface electromyography diagnostics in women with partial vaginismus with or without vulvar vestibulitis and in asymptomatic women. J Psychosomat Obs Gyn 2004;25(3–4):281–294.
11. Henson DE, Rubin HB, Henson C. Labial and vaginal blood volume responses to visual and tactile stimuli. Arch Sexual Behav 1982;11(1):23–31.
12. Pukall C, Binik YM, Khalif S. A new instrument for pain assessment in vulvar vestibulitis syndrome. J Sex Marital Ther 2004;30(2):69–78.
13. Levin RJ. A journey through two lumens. Inter J Impot Res 2003;15(1):2–9.
14. Munarriz R, et al. A prospective duplex doppler ultrasonographic study in women with sexual arousal disorder to objectively assess genital engorgement induced by EROS Therapy. J Sex Marital Ther 2003;29(Special Issue): The Annual Meeting of the Female Sexual Function Forum: 85–94.
15. Rosen RC. Assessment of female sexual dysfunction: Review of validated methods. Fertility Sterility 2002;77(4):89–93.
16. Idama TO, Pring DW. Vaginal dilator therapy—an outpatient gynaecological option in the management of dyspareunia. J Obst Gyn 2000;20(3):303–305.
17. Vaginismus.com (No date). Dilator set. [Online]. Available at http://www.vaginismus.com/products/dilatorset/. Accessed 2005, August.
18. Weiss JC. Treating Vaginismus: Patient without partner. J Sex Educ Ther 2001;26(1):28–33.
19. Koehler JD. Vaginismus: Diagnosis, etiology and intervention. Cont Sexuality 2002;36(9):9–17.
20. Billups KL, et al. A new non-pharmacological vacuum therapy for female sexual dysfunction. J Sex Marital Ther 2001;27(5).
21. UroMetrics (No date). Photo Downloads. [Online]. Available at http://urometrics.com/pressroom/photodownloads.cfm. Accessed 2005, August.
22. Freund K, Sedlacek F, Kanob K. A simple transducer for mechanical plethysmography of the mole genital. J Exp Anal Behav 1965;8:169.
23. Riley A, Riley E. The effect of Clitstim (Vielle™) on sexual response induced by masturbation in female volunteers. Sexual Relationship Ther 2003;18(2):45–52.
24. Cobellis L, et al. A donazol-loaded intrauterine device decreases dysmenorrheal, pelvic pain, and dyspareunia associated with endometriosis. Fertility Sterility 2004;82(1):239–240.
25. Beck JG, Sakheim DK, Barlow DH. Operating characteristics of the vaginal photoplethysmograph: Some implications for its use. Arch Sexual Behav 1983;12(1):43–58.
26. Althof SE, et al. Outcome measurement in female sexual dysfunction clinical trials: Review and recommendations. J Sex Marital Ther 2005;31(2):153–166.
27. Coleman E. Erectile dysfunction: A review of current medical treatments. Can J Human Sexuality 1998;7(3):231–244.
28. Meuleman EJH, van Lankveld JJDM. Hypoactive sexual desire disorder: An underestimated condition in men. BJU Inter 2005;95(3):191–196.
29. Rowland DL. Issues in the laboratory study of human sexual response: A synthesis for the nontechnical sexologist. J Sex Res 1999;36(1):3–15.
30. Wheeler D, Rubin HB. A comparison of volumetric and circumferential measures of penile erection. Arch Sexual Behav 1987;16(4):289–299.
31. Libman E, et al. Sleeping and waking-state measurement of erectile function in an aging male population. Psycholog Assess 1989;1(4):284–291.
32. Diedrich GK, Stock W, LoPiccolo J. A study on the mechanical reliability of the Dacomed Snap Gauge: Implications for the differentiation between organic psychogenic impotence. Arch Sexual Behav 1992;21(6):509–523.
33. Nobre PJ, et al. Determinants of sexual arousal and accuracy of its self-estimation in sexually functional males. J Sex Res 2004;41(4):363–371.
34. Rosen RC, Keefe FJ. The measurement of human penile tumescence. Psychophysiology 1978;15(4):366–376.
35. Da Silva JP, Santiago L, Goncalves JC. The corpus cavernous electromyography in the erectile dysfunction diagnosis. Acta Chirur Hung 1994;34(3–4):243–252.
36. Burnett AL, et al. Near infrared spectrophotometry for the diagnosis of vasculogenic erectile dysfunction. Inter J Impot Res 2000;12(5):247–254.
37. Berman L, Berman J, Zierak M, Marley C. Outcome measurement in sexual disorders. In: IsHak WW, Burt T, editors. Outcome Measurement in Psychiatry: A Critical Review. Washington (DC): American Psychiatric Publishing; 2002.

38. Kalmus E, Beech AR. Forensic assessment of sexual interest: A review. Aggression Violent Behav 2005;10(2):193–217.
39. Wylie KR, Steward D, A Homemade Device for Erectile Dysfunction. J Sex Marital Ther 2000;26(4):335–339.
40. Brosman SA. (Dec. 14, 2004). Erectile Dysfunction Resource Center. [Online]. eMedicine. Available at http://www.emedicine.com/rc/rc/pimages/i8/s11/ed.htm. Accessed August 2005.
41. Freund K, Charles U. A laboratory method for diagnosing predominance of homo- or hetero- erotic interest in the male. Behav Res Ther 1963;1(1):85–93.
42. Bancroft JJJ, Jones HG, Pullan BR. A simple transducer for measuring penile erection, with comments on its use in the treatment of sexual disorders. Behav Res Ther 1966;4: 239–241.
43. O'Donohue W, Letourneau EY. The psychometric properties of the penile tumescence assessment of child molesters. J Psychopathol Behav Assess 1992;15(3):259–274.
44. Golde JA, Strassberg DS, Turner CM, Psychophysiologic assessment of erectile response and its suppression as a function of stimulus media and previous experience with plethysmography. J Sex Res 2000;37(1):53–59.
45. Rosen RC, Kopel SA. Penile plethysmography and biofeedback in the treatment of a transvestite-exhibitionist. J Consulting Clin Psychol 1977;45(5):908–916.
46. Haywood TW, Grossman LS, Cavanaugh JL. Subjective versus objective measurements of deviant sexual arousal in clinical evaluations of alleged molesters. Psycholog Assess 1990;2(3):269–275.
47. Rea JA, DeBriere T, Butler K, Saunders KJ. An analysis of four sexual offenders' arousal in the natural environment through the use of a portable penile plethysmograph. Sexual Abuse: J Res Treatment 1998;10(3):239–255.
48. Mathews C, Hartsell JE, Kohn M. Debunking penile plethysmograph evidence. Reporter 2001;28(2):11–14.
49. O'Donohue W, Levensky E. Handbook of Forensic Psychology. New York: Academic Press; 2004.
50. Abel GG, et al. Screening tests for pedophilia. Criminal Justice Behav 1994;21(1): Special issue: The assessment and treatment of sex offenders 115–131.
51. Laws RD, Gress CLZ. Seeing things differently: The viewing time alternative to penile plethysmography. Legal Criminolog Psychol 2004;9(2):183–196.
52. Glasgow DV, Osborne A, Croxen J. An assessment tool for investigating paedophile sexual interest using viewing time: An application of single case methodology. Br J Learning Disab 2003;31(2):96–102.
53. Gaither GA. The reliability and validity of three new measures of male sexual preferences. Dissertation Abs Inter: Sec B: Sci Eng 2001;61(9-B):4981.
54. Web brochure. (No date) Abel Assessment for Sexual Interest Brochure. [Online]. Abel Screening, Inc. Available at http://www.abelscreen.com/asipdfs/webbrochure.pdf. Accessed August 2005.
55. Letourneau EJ. A comparison of objective measures of sexual arousal and interest: Visual reaction time and penile plethysmography. Sexual Abuse: J Res Treat 2002;14(3): 207–223.
56. Cohen AS, Rosen RC, Goldstein L. EEG hemispheric asymmetry during sexual arousal: Psychophysiological patterns in responsive, unresponsive, and dysfunctional men. J Abnormal Psyschol 1985;94(4):580–590.
57. Oksol EM, O'Donohue W. A critical analysis of the polygraph. In: O'Donohue W, Levensky ER, editors. Handbook of Forensic Psychology New York: Elsevier; 2004.
58. Sbraga TP. Sexual deviance and forensic psychology: A Primer. In: O'Donohue W, Levensky ER, editors. Handbook of Forensic Psychology New York: Elsevier; 2004.
59. Abel, Saether 2001.

See also ANALYTICAL METHODS, AUTOMATED; PERIPHERAL VASCULAR NONINVASIVE MEASUREMENTS; STRAIN GAGE; TEMPERATURE MONITORING.

SHAPE MEMORY ALLOYS.

See ALLOYS, SHAPE MEMORY.

SHOCK, TREATMENT OF

SABIN DECAU
COLIN F. MACKENZIE
University of Maryland,
School of Medicine
Baltimore, Maryland

INTRODUCTION

Assessment and treatment of shock is based on understanding of circulatory system physiology and cellular metabolism. Shock is defined as inadequate supply of oxygen and nutrients due to inadequate capillary perfusion to the cells. This definition is not always correct, as in severe shock some cells cannot metabolize oxygen even with adequate perfusion. In addition, the removal of metabolites is equally important or even more crucial over time, since accumulated metabolites will cause cell injury.

Deficient capillary perfusion triggers a host of metabolic changes in every organ and tissue, which affects whole body homeostasis and circulation. Ideally, one should evaluate the cellular metabolism, but this assessment can be done only indirectly by measuring the acid–base balance in the form of blood gas and pH. This arterial or venous blood test does not give any information about the regional metabolic alterations. For evaluation of the circulatory function, indirect measurements are used including heart rate, blood pressure, and urinary output. Invasive hemodynamic monitoring devices, such as the pulmonary artery catheter are also used for cardiac function assessment. To effectively reverse shock requires monitoring of the pathophysiologic state of hypoperfusion. Various devices are used for this state to be monitored directly or indirectly, continuously and intermittently, so that the patient response to therapy can be determined and treatment adjustments made.

ETIOLOGY

There are three common types of shock: loss of blood volume (hypovolemic shock), abnormal blood distribution (e.g., septic shock), and cardiac pump failure (e.g., cardiogenic shock). In practice, these three types of shock do not occur independently, but are frequently mixed with a similar final pathway, irrespective of the circulatory or cardiac state that may have been the precipitant (1).

Heart

The critical event in the development of cardiogenic shock is severe impairment of heart muscle contractile performance. Cardiac performance is primarily determined by four factors: preload, afterload, strength of contraction, and heart rate. Preload is defined as the force exerted on myocardium at the end of ventricular filling. If the filling is inadequate (low preload), cardiac output will be reduced. This finding is best exemplified by extracardiac obstruction, such as pericardial tamponade (blood in the pericardial sac that surrounds the heart) or by the reduction of venous return to the heart caused by high thoracic pressure (such as occurs with a pneumothorax, which is air under tension between the ribs and lung). Afterload is the resistance to emptying of the ventricles with myocardial contraction after the opening of the pulmonary and aortic valves. A rapid increase in afterload as in valvular stenosis leads to decreased volume of blood ejected from ventricles and decreased cardiac output. Contractility is most affected by loss of heart muscle as a result of myocardial infarction (heart attack). In the acute setting of myocardial ischemia, loss of at least 40% of left ventricular heart muscle results in severe depression of cardiac performance and shock.

A similar picture may also result from myocarditis (an inflammation of the heart muscle) and with prolonged cardiopulmonary bypass during cardiac surgery. In addition, myocardial stunning may occur following reversible myocardial ischemia. If the heart rate is too fast as in ventricular dysrhythmias, this may compromise ventricular filling and decrease cardiac output. If the rate is too slow, cardiac output may be insufficient, and shock may ensue. Both high and low heart rates are common complications of myocardial infarction (1–3).

Loss of Volume

Hypovolemic shock is caused by a reduction in intravascular circulating volume to a point where compensation is no longer possible, by constriction of venous capacitance vessels, to maintain cardiac filling. The loss of circulating volume may result from hemorrhage, dehydration, or leakage from the circulation into the body tissues. Trauma and gastrointestinal bleeding are common causes of rapid blood loss and lead to a reduction in preload and cardiac output. However, the oxygen (O_2) carrying capacity of blood is not severely impaired except in massive blood loss causing a decrease in hemoglobin concentration. Dehydration results in intravascular volume reduction with an increase in O_2 carrying capacity, since the number of red blood cells per unit of volume will increase, but the cardiac output is decreased.

Abnormal Distribution

Maldistribution of blood flow occurs with widespread vasodilation usually caused by infectious agents (septic shock) but vasodilation with low systemic vascular resistance may be caused by other mechanisms including endocrine disease, anaphylaxis (severe systemic allergic reaction), and neurogenic shock.

PATHOPHYSIOLOGY

The underlying result of shock of any etiology is hypoperfusion at the cellular level due to an inability to provide the cell with adequate oxygen, glucose, and other nutrients necessary to maintain normal functions. Consumption is calculated by the Fick equation, where cardiac output = O_2 consumption % (arterial-mixed venous O_2 content). If the cardiac output is low, the tissue blood flow is reduced.

Hypoperfusion and hypovolemia due to blood loss decrease the blood pressure and cardiac output. With the loss of up to 15% of total blood volume, no detectable changes in heart rate and blood pressure may be found because venous capacitance vessel constriction compensates for hypovolemia and maintains cardiac filling pressures. If fluid loss is 15–30% of blood volume, that is, 750–1500 mL in a 70 kg male, heart rate will increase tending to maintain cardiac output. Pulse pressure (the difference between systolic and diastolic blood pressure) decreases due to a rise in vascular resistance mediated by circulating catecholamines. The systemic vascular resistance (SVR) increases by constriction of arterioles, which tends to maintain an adequate arterial pressure. However, an increase in SVR raises arterial pressure, but unless it is accompanied by an increase in cardiac output, it has no effect on tissue blood flow. The relationship between flow and pressure is as follows: flow = pressure/resistance, and can be applied to the entire circulatory system, or to a single organ or even an electric circuit (Ohms law).

If the heart produces a constant flow per unit time (cardiac output), then the tissue blood flow (perfusion) will be inversely proportional to the vascular resistance. Therefore, in the case of shock, capillary perfusion should be globally reduced. However, this is not the case, as different organs have variable blood flow and oxygen utilization. The kidneys have high blood flow and little O_2 extraction, whereas the heart has relatively low blood flow (because it is contracting) and high O_2 extraction (Fig. 1) (4). In shock, vasoconstriction occurs preferentially in certain regions (skin, muscle, viscera), diverting the blood flow from these high vascular resistance organs to more vital organs with low resistance, such as heart and brain. If blood loss continues beyond 30–40% of total blood volume, the compensatory mechanisms fail and the blood pressure will fall, accompanied by marked tachycardia, changes in mental status, and decreased urinary output (1).

In the first stage of shock, arterioles are constricted and the amount of oxygen available to tissues may be insufficient for their metabolic needs. The global oxygen consumption will decrease after O_2 extraction from hemoglobin has been maximized (from a normal of 25% of the O_2 carriage to up to 70%). When O_2 consumption falls below a critical level, local metabolism becomes anaerobic, which leads to an increased production of lactic acid by the cells. Later on, this metabolic acidosis causes relaxation of the precapillary sphincters while the postcapillary venules are still constricted. Therefore, the capillaries become overfilled with blood and the hydrostatic pressure

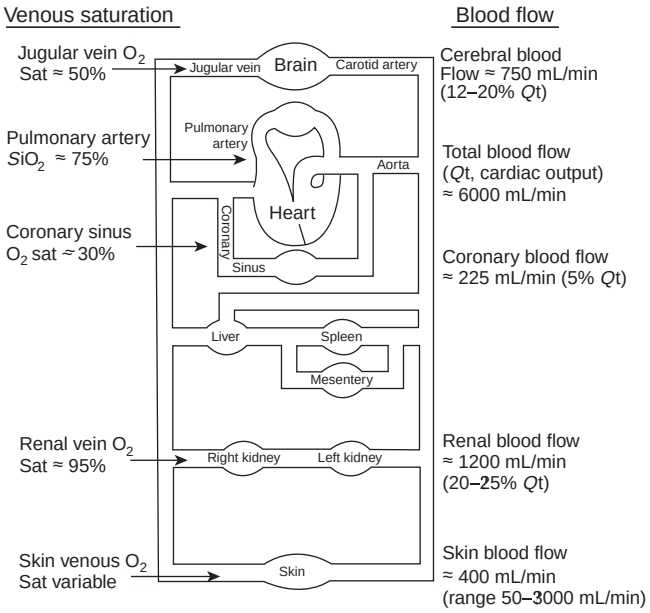


Figure 1. Venous oxygen saturations of blood leaving various organs are shown on the left side and blood flow expressed in mL/min and as a % of total cardiac output (Qt) are shown on the right side of this figure. Note that the heart and brain with low blood flow relative to their oxygen requirements (e.g., coronary blood flow and carotid artery) have low venous oxygen saturations. In comparison, an organ such as the kidney has high blood flow, but little oxygen requirement and contributes relatively more to the final mixed-venous O_2 saturation of 75% in the pulmonary artery than does the much smaller venous blood flow from the heart and brain. For this reason, a normal pulmonary artery oxygen saturation is not a good indicator of adequate shock resuscitation of the brain or heart.

increases such that there will be loss of plasma into the interstitial space, further depleting the circulating volume. Hemoconcentration occurs and the blood viscosity is increased, causing slowing of the blood flow through the microcirculation. A vicious circle is established, where slow flow causes an increase in viscosity, which in turn leads to a decrease in flow, and so on.

In the latter stages, capillaries are filled with a sludge of red blood cells, which impairs the local flow. In addition, there is further reduction in tissue exchanges by shunting of blood through arteriovenous anastomoses so the functional flow is nearly zero. Therefore, cells lack O_2 and anaerobic metabolism will proceed to a point where cells are no longer able to survive. If a significant number of cells die, the organ function will be compromised. The lung is the first organ to fail, approximately 3–7 days after the primary shock event. The condition that results is called adult respiratory distress syndrome (ARDS) and is characterized by an increase in physiologic shunting and refractory hypoxemia. Kidney involvement is apparent within 2–5 days with acute renal failure due to ischemic tubular necrosis, followed by electrolyte disturbances and metabolic acidosis. Clearance of creatinine by the kidney approximates glomerular filtration rate and when this falls $<$25 mL/min, there is early onset of renal failure that is still potentially reversible (5). If liver failure occurs, the first sign is jaundice, but the most significant evidence of failure is abnormal hepatic metabolism, with impaired protein synthesis and an inability to process available energy substrates. Gastric hemorrhage may occur as a late manifestation and is usually precipitated by coagulopathy, a common event in shock. The immunologic response is depressed leading to an increased susceptibility to infections. The syndrome in which all these events occur—multiple organ system failure—is a terminal event common to all types of shock (3).

SHOCK ASSESSMENT

Assessment and management of a patient in shock are accomplished simultaneously. Since evaluation of cellular metabolism cannot be done directly, the physician must rely on surrogate clinical findings such as blood pressure (BP), heart rate, skin temperature, urinary output and mental status, and on data obtained by using various monitoring devices.

Measurement of BP is routine in shock patients. The systolic blood pressure measure (SBP) is not a good indicator of blood loss, as up to 30% of circulating volume may be lost without any change in SBP. Instead the diastolic blood pressure (DBP) is more sensitive and is usually elevated in shock due to peripheral vasoconstriction. Therefore, mean arterial pressure defined as MAP $=$ DBP $+$ 1/3 pulse pressure (SBP $-$ DBP) seems to be more useful for BP monitoring. Blood pressure measured by auscultation is inaccurate in patients with low peripheral blood flow. Invasive measurement using an intra-arterial catheter inserted into radial, brachial, axillary, or femoral arteries is more precise and provides continuous monitoring and easy access for blood gas and pH analysis.

Change in heart rate occurs with increasing blood loss, values $>$100/min in adults are detectable before any change in SBP. Pulse pressure is usually low due to increased DBP. Another clinical monitor is capillary refill time, which is the time for return of blood flow after compression of the nail beds until blanching occurs. It is prolonged $>$2 s in severe peripheral vasoconstriction. Skin temperature correlates well with peripheral blood flow and a difference of 3–6 $^\circ$C between toe and rectal temperature reflects severe peripheral vasoconstriction (2). The ability of commonly used clinical parameters to quantify acute hemorrhage is shown in Fig. 2 (6). Base deficit, mean arterial pressure, serial hemoglobin, and serum lactate are related to blood loss.

Preload is assessed by measuring central venous pressure (CVP), which correlates with right atrial pressure, with a catheter inserted in the superior vena cava via the jugular or subclavian veins. Normal CVP is 5–12 cm of H_2O and values $<$5 are generally found in hypovolemic states and indicate the need for assessment of reserve cardiac function by rapid fluid administration together with assessment of changes in CVP. However, CVP does not always correlate with fluid requirement because an increase in pulmonary vascular resistance (hypoxia, acidosis, increased intrathoracic pressure) may be associated

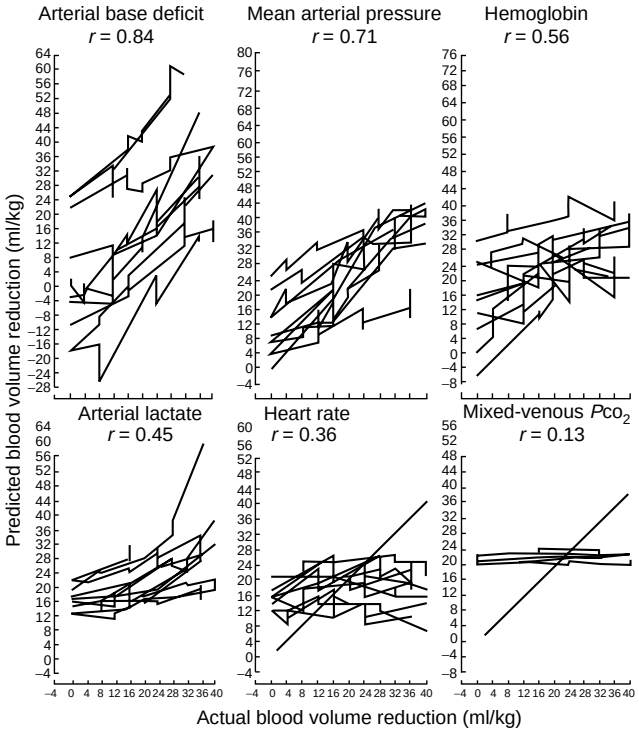


Figure 2. Commonly used clinical and laboratory values compared to quantity of acute hemorrhage. Graphs show predicted versus actual blood volume reductions for six representative parameters. Solid black lines represent an individual animal ($n = 10$). Gray bar represents the ideal in which predictions equal actual blood volume reductions. For mean arterial pressure, predictions at large volume hemorrhage were more accurate than predictions with small volume bleeds. Models such as heart rate and lactate both showed significance variability before hemorrhage among animals and flat slopes (e.g., mixed-venous PCO_2) indicated fixed-volume predictions regardless of actual degree of hemorrhage. (Reproduced with permission from Waisman Y et al. J. Appl. Physiol. 1993;74:410–519.)

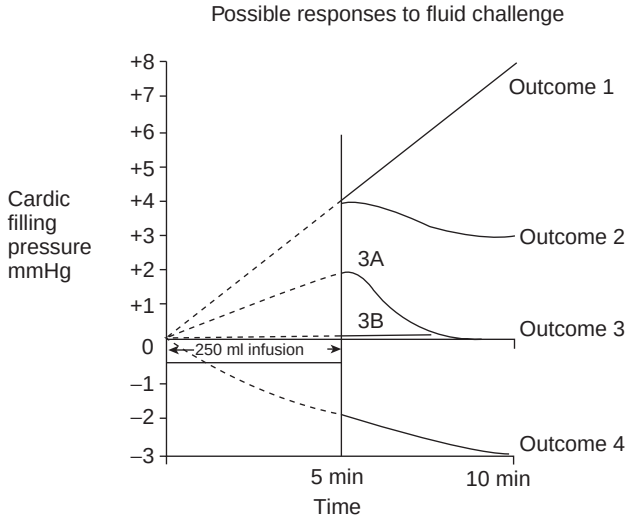


Figure 3. The possible responses to rapid (50 mL/min for 5 min) administration of a fluid is shown schematically. The vertical axis shows the change in cardiac filling pressure (could be either CVP or PAQP) in response to fluid. Time is on the horizontal axis. Fluid infusion ceases after 5 min and the filling pressures are remeasured (and cardiac output measured if available). Fluid challenge in outcomes 3A and 3B should be repeated until filling pressures increase and remain elevated for 5 min after ceasing fluid infusion (outcome No. 2). (Reproduced with permission from Mackenzie CF et al. J. Neurosurg. 1985;62:843–849.)

with high CVP, reflecting right ventricular failure, even when there is considerable blood loss.

Fluid Challenge

Reserve cardiac function is assessed by means of a fluid challenge until CVP pressures are elevated at least 2 mmHg above the baseline for 10 min after fluid infusion ceases. There are four possible outcomes when 250 mL boluses of fluid are given over 5 min (Fig. 3). Outcome No. 1: Filling pressures rise with the challenge and continue to rise even after fluid infusion ceases. If cardiac output is measured (see below), there is no increase with elevation of filling pressures and the heart has limited reserve function and is not able to deal with the increased fluid load by increasing contractility by the Frank–Starling mechanism (this mechanism describes the property of heart muscle increasing its contraction in proportion to fiber length, up to a maximum point when contractility decreases and cardiac failure occurs). Further fluid infusion (when this response occurs) is expected to produce cardiac failure. The therapeutic indication this response dictates is to restrict fluid infusion and reduce myocardial depressant agents. If the trend continues, inotropic agents are required to increase cardiac contractility and reverse myocardial depression. Outcome No. 2: Central venous pressure or pulmonary capillary wedge pressure (PCWP) rise 3–4 mmHg, but then falls to a level 2 mmHg above baseline, indicating myocardial contractility is adequate for the increase in cardiac preload. Management should be to infuse fluids and maintain cardiac filling pressures within this range for optimum cardiac output and oxygen transport for the prevailing vascular tone. In outcome No. 3, filling pressures either rise briefly (3A) or do not rise at all (3B) with the 250 mL fluid challenge indicating that the patient has considerable reserve cardiac function and a greater fluid load could be tolerated. If the patient has low urine flow, has evidence of poor tissue perfusion, such as acidosis or low mixed venous oxygen tension, inotropic agents, or diuretics should not be given. Rather, fluid infusion should continue at the same rate until outcome No. 2 is obtained. In fact, because filling pressures have not increased in outcome No. 3, it is unlikely that myocardial fiber length would increase and, therefore, cardiac contractility would not have changed by the Frank–Starling mechanism. Outcome No. 4 is seen in ~5% of cases and results in a fall in filling pressures when fluid is infused rapidly. The two most likely explanations are either that rapid fluid infusion has a vasodilator effect on peripheral vasculature and reduces left ventricular afterload, increasing cardiac output, or alternatively, it may be related to a reduction in heart rate seen when fluid is infused in the hypovolemic patient. The decreased heart rate allows more

time for myocardial perfusion, which occurs mostly during diastole or cardiac relaxation. Cardiac output increases and the Frank–Starling function curve shifts to the left due to improve myocardial oxygenation (7).

Cardiac Catheterization

Catheterization of pulmonary artery (PA) is a method of hemodynamic monitoring that can be used to assess right and left ventricular function as well as quantitate the proportion of blood shunting and to calculate tissue delivery and O_2 extraction. Catheterization is usually reserved for chronically ill patients with heart disease or shock refractory to conventional therapy, because it requires an invasive flow directed, balloon tipped catheter to be floated in the blood stream through the right side of the heart into the pulmonary artery. Such heart catheterization has complications including infection and cardiac rhythm irregularities. It provides data on PA pressure, PCWP, vascular resistances, cardiac output (commonly by a thermodilution technique), and also allows blood sampling from PA, for measurement of mixed-venous saturation (SvO_2) and calculation of intrapulmonary shunting of blood.

Equations

Normal SvO_2 is ~75% (Fig. 1), and it reflects the ratio between oxygen delivery [arterial O_2 content (CaO_2) × cardiac output (Qt)] and oxygen extraction [ratio of O_2 consumption (VO_2) and O_2 delivery]. Consumption is calculated by the Fick equation, where cardiac output = O_2 consumption % (arterial mixed-venous O_2 content). If the cardiac output is low, the tissue blood flow is reduced.

Blood gas and pH analysis give a rough estimate of oxygen utilization and cellular metabolism by calculation of bicarbonate and base deficit. A low arterial pO_2 (hypoxemia) may accompany shock and be found before any clinical sign due to ventilation–perfusion inequality and increased venous admixture by shunting (see above). The causes of hypoxemia include hypoxic hypoxemia (low inspired O_2), anemic hypoxemia (low hemoglobin for carrying O_2) stagnant hypoxemia (low cardiac output for delivery of O_2), and histotoxic hypoxemia (poisoning of the enzyme systems used to offload O_2 from hemoglobin at the tissues). Blood lactate is an indicator of tissue hypoperfusion and anaerobic metabolism and is elevated in patients with low cardiac output. Lactate is a clinically useful marker of the amount and duration of shock. Sustained reduction in elevated lactate is an important clinical marker of recovery. Arterial pH can be high or normal despite metabolic acidosis, because of low pCO_2 due to increased respiratory rate and alveolar ventilation, common in patients with low cardiac output (1,3,8).

Real-Time Noninvasive Measures of Systemic Perfusion

Sublingual capnometry (measurement of sublingual $pCO_2 - PslCO_2$) is a new technique for assessment of systemic perfusion failure. It is based on elevated pCO_2 in tissues with low blood flow due to intracellular buffering of hydrogen ions by bicarbonate. Elevated $PslCO_2$ correlates well with increased blood lactate and low mean arterial pressure (MAP), markers of tissue hypoperfusion. $PslCO_2$ has the advantage of prompt indication and continuous monitoring of tissue flow reversal, unlike lactate whose clearance presents significant delay. A threshold value of 70 mmHg for $PslCO_2$ has been identified that is predictive of both the severity state and survival. A $PslCO_2$ >70 mmHg is highly predictive of circulatory failure whereas readings <70 mmHg are highly predictive of survival (9). A similar technique to sublingual capnometry is gastrointestinal tonometry, which measures gut mucosal pCO_2. Calculation of intramucosal gut pH is possible (pHi) using the Henderson–Hasselbach equation: $pHi = 6.1 + \log (HCO_3^- / \alpha^* \text{tonometer } pCO_2)$, where α is the solubility of CO_2 in plasma ($\alpha = 0.03$) Values of pHi <7.32 define mucosal hypoperfusion and are associated with high mortality rates (10).

Brain Perfusion

Brain perfusion monitoring is technically difficult and inaccurate. Clinically, signs of confusion, altered sensorium, agitation, and decreased consciousness give a rough idea about cerebral hypoperfusion. Jugular venous oxygen saturation ($SjvO_2$), transcranial cerebral oximetry, and brain tissue oxygen tension ($PbtO_2$) monitoring are the methods of monitoring brain oxygenation. Measurement of $SjvO_2$ is performed using a catheter inserted in the jugular bulb, the upper part of the internal jugular vein. Continuous monitoring of venous saturation without blood sampling is possible by using intravenous oximetry. This type of device has been used in patients with head injury and during anesthesia for neurosurgery. It provides only global brain oxygenation monitoring and is susceptible to errors. Cerebral oximeters using near-infrared spectroscopy seem to be a promising alternative. They can evaluate regional ischemia, hemoglobin saturation, and even concentration of oxygenated and reduced hemoglobin. However, these monitors can only assess trends, where each patient is their own control. There are no normative data for comparison and the boundaries of monitored brain tissue cannot be precisely defined. Brain tissue oxygen tension is measured with small catheters inserted directly into the brain tissue during a craniotomy or via a burr hole. These catheters measure pO_2, pCO_2, pH, and temperature. Some studies suggest a normal value for $PbtO_2$ of ~35 mmHg, whereas cerebral ischemia is usually defined as a $PbtO_2 < 8$ mmHg. These data were obtained in patients with traumatic brain injury, but their usefulness should be confirmed by further studies (11,12). Extra cellular glutamate and aspartate measures (obtained by microdialysis) are closely related to outcome after head injury (13). These markers were also related to the type of head injury and suggest that excitatory amino acids play a role in the evolution of brain injury.

SHOCK MANAGEMENT

Treatment of shock should be directed to its underlying cause. However, establishing an exact diagnosis can be

time consuming, so management is focused on simultaneously stabilizing the patient and proceeding with diagnostic tests to identify the cause of shock.

For cardiogenic shock, the goal is an increase in cardiac output by acting to change preload, afterload, contractility, or heart rate. Therapy is tailored using information provided by a PA catheter. Various drugs are given to increase contractility and to relieve pulmonary congestion. In unresponsive cases, additional measures may be considered, such as urgent myocardial revascularization in acute myocardial infarction, intraaortic balloon counterpulsation, and anatomic defects repair (ruptured valves). In extracardiac compression, relieving the pressure of pericardial tamponade by pericardial puncture or insertion of a chest tube for increased intrathoracic pressure due to pneumothorax is the treatment of choice, when these are the causes of impaired cardiac filling and decreased cardiac output.

In case of septic shock (the most common form of distributive shock), large quantities of fluids are administered to fill the vascular bed and maintain perfusion pressure. Cardioactive drugs are used only if cardiac output declines. At the same time steps are taken to identify and control the source of infection (3).

Hypovolemic shock requires initial rapid expansion of circulating volume. Fluid resuscitation is initiated with administration of crystalloid or colloid solutions through large-bore intravenous lines. Rapid infusion devices can be used to pump large amounts of fluids in <10 min. A potential adverse effect resulting from resuscitation with large amounts of fluid, when using rapid infusion devices, is a drop in body core temperature. Levels $<35\,^{\circ}C$ are associated with impaired coagulation and depressed cardiac contractility (14). Covering the patient with inflatable warming blankets can prevent this complication, but the most effective method for rewarming is an extracorporeal countercurrent warmer through femoral artery and vein cannulation, which can elevate temperature $\sim 6\,^{\circ}C$ in <30 min. During fluid infusion, hemodynamic parameters are continuously monitored and signs of instability (persistent SBP <90 mmHg) imply there is ongoing blood loss or shock has not been reversed. Classically, a hemoglobin (Hb) level <10 g/dL with continuous loss requires blood transfusion, but recent studies have demonstrated that this level can be as low as 7 g/dL without adverse effects in the majority of the population (15). Those with cardiac or cerebrovascular disease should be transfused at higher hemoglobin concentrations.

THE FUTURE OF SHOCK DIAGNOSIS AND MANAGEMENT

Future trends in shock diagnosis and management include identification of mediator's released in shock states and blockage of the release of harmful mediator factors while facilitating release of those with benefits. The field of proteomics, defining protein expression with shock, will provide many future treatment and diagnostic opportunities. Genomics may identify some individuals or disease states susceptible to adverse outcomes from shock. These future findings could lead to improved outcome, particularly from septic shock, which has a high mortality and morbidity.

BIBLIOGRAPHY

1. Kelman GR. Applied Cardiovascular Physiology. London: Butterworths; 1977.
2. Hardaway G. Shock. The Reversible Stage of Dying. Littleton: KPGS Publishing Company; 1988.
3. Shoemaker WC, Appel PL, Kram HB. Role of oxygen transport in the pathophysiology, prediction of outcome and therapy of shock. In: Bryan-Brown CW, Ayres SM, editors. Oxygen Transport and Utilization. The Society of Critical Care Medicine 1987. p 65–92.
4. Mackenzie CF. Anesthesia in the shocked patient. J Eur Emer 1994;4:163–172.
5. Shin B, Mackenzie CF, Helrich M. Creatinine clearance for early detection of post-traumatic dysfunction. Anesthesiology 1986;64:605–609.
6. Waisman Y, Eichacker PO, Banks SM, Hoffman WD, MacVittie TJ, Natanson C. Acute hemorrhage in dogs: construction and validation of models to quantify blood loss. J Appl Physiol 1993;74(2):510–9.
7. Mackenzie CF, Shin B, Krishnaprasad D, Illingworth W. Assessment of cardiac and repiratory function during surgery on patients with acute quadriplegia. J Neurosurg 1985;62: 843–849.
8. Darovic GO. Hemodynamic Monitoring: Invasive and Noninvasive Clinical Application. Philadelphia: WB Saunders; 1995.
9. Weil MH, Nakagawa Y, Tang W, Sato Y, Ercoli F, Finegan R, Grayman G, Bisera J. Sublingual. capnometry: a new noninvasive measurement of diagnosis and quantitation of severity of circulatory shock. Crit Care Med 1999;27(7):1225–1230.
10. Gutierrez G. Gastrointestinal Tonometry. Oxford Textbook of Critical Care. Oxford: Oxford University Press; 1999.
11. Symthe PR, Samra SK. Monitors of cerebral oxygenation. Anesth Clin N Am 2002;20(2):293–313.
12. Valadka AB, Gopinath SP, Contant CF. Relationships of brain tissue PO_2 to outcome after severe head injury. Crit Care Med 1998;26:1576–1581.
13. Gopinath SP, Valadka AB, Goodman JC, Robertson CS. Extracellular glutamate and aspartate in head injuried patients. Acta Neurochir Suppl 2000;76:437–438.
14. Dunham CM, Belzberg H, Lyles R, Weireter L, Skurdal D, Sullivan G, Esposito T, Hamini M. Resuscitation of hypovolemic traumic patients. Resuscitation 1991;Apr 21(2–3), 207–229.
15. Herbert PC, Wells G, Blajchman MA. A multicenter randomized controlled clinical trial of transfusion requirements in critical care. N Eng J Med 1999;340:409–417.

See also BLOOD PRESSURE MEASUREMENT; CARDIOPULMONARY RESUSCITATION; PERIPHERAL VASCULAR NONINVASIVE MEASUREMENTS.

SHUNT, FOR HYDROCEPHALUS. See HYDROCEPHALUS, TOOLS FOR DIAGNOSIS AND TREATMENT OF.

SIMULATION. See PHYSIOLOGICAL SYSTEMS MODELING.

SINGLE PHOTON EMISSION COMPUTED TOMOGRAPHY. See COMPUTED TOMOGRAPHY, SINGLE PHOTON EMISSION.

SKIN SUBSTITUTE FOR BURNS, BIOACTIVE

ROBERT H. DEMLING
LESLIE DESANTI
Harvard Medical School
Boston, Massachusetts

INTRODUCTION

A body burn is a complex injury process resulting in local changes in skin integrity as well as profound systemic changes in fluid and electrolyte balance, metabolism, and immune defenses. Severe psychosocial changes also occur in addition to long-term, often permanent changes in skin function. Major advances in care have resulted in a marked decrease in mortality and also morbidity, especially with massive burns. In addition to survival, the current focus in burn care is on improving the long-term function and appearance of the healed or replaced skin cover as well as quality of life.

This focus on quality has generated a significant interest in the use of skin substitutes to be used to improve wound healing, to control pain, to more rapidly close a burn wound, to improve functional and cosmetic outcome, and, in the case of massive burns, to increase survival.

To more effectively address these new roles, the new generation of skin substitutes are developed to be biologically active. The concept behind providing bioactivity is that the wound healing process can be substantially improved as compared with a simple synthetic barrier-type dressing. It remains to be seen just how much better this new generation of products will impact the burn wound. To date, the new products to be discussed have not displaced the more inert standard burn wound dressings, but rather are used in conjunction and for quite specific indications.

The skin substitutes are initially classified according to whether they are to be used as a temporary wound covering to decrease pain and augment healing or a permanent skin substitute to add to or replace the remaining skin components.

The ideal properties and indications for these products will be better clarified after a discussion of the function of normal skin and the effect of a burn on skin integrity.

THE NORMAL PROPERTIES OF SKIN

Normal skin is a very complex bilayer organ with a wide variety of properties mainly protective barriers, which are critical to survival (1–5). Loss of these protective barrier functions occurs with a skin burn. Restoration of skin structure and function become the necessary properties of temporary and permanent skin substitutes. The skin is a bilayer organ with each layer having specific functions. But both layers are needed for proper skin function. An outer thin epidermal layer covers an inner thicker dermal layer.

Epidermis

The outer thinner layer known as the epidermis is composed mainly of epithelial cells or keratinocytes (1–3). The deepest epidermal cells are immature cells and are continually dividing and migrating toward the surface, to replace lost surface cells; keratinocytes. The same types of regenerating epidermal cells are found in hair follicles and other skin appendages, which are anchored in the dermis. As the cells mature and migrate to the surface, they form keratin, which becomes an effective barrier to environmental hazards such as infection and excess water evaporation.

Stratum Corneum. The otnatum corneum is the "outermost" layer of the epidermis consisting of several flattened layers of dead keratinocytes as well as keratin. This layer protects against entry of bacteria and toxins. The epidermal layer regenerates every 2–3 weeks, but regeneration requires the structure and functional components of the dermis.

Skin Functions: Epidermis (Outer Layer).

- Protection from drying.
- Protection from bacterial entry (infection).
- Protection from toxin absorption, like chemicals on the skin.
- Fluid balance: avoiding excess evaporative water loss that would cause dehydration.
- Neurosensory (touch, pain, pressure, sensation).
- Social-interactive (visible portion of the body covering).

The epidermis is firmly anchored to the dermis by the rete pegs, which are ingrowth of epithelial cells interdigitalizing into the upper dermis like the teeth on a saw.

Dermis

The dermis (1,4,5) is the deeper skin layer responsible for skin durability, and the barrier functions of controlling body temperature, and flexibility and barrier function. The nerves for touch and pain, blood vessels, and hair follicles are present in the dermis. The dermis is responsible for orchestrating the formation of the epidermis.

Skin Functions: Dermis (Inner Layer).

- Regulation of body temperature avoiding hypothermia and hyperthermia.
- Properties of elasticity and durability necessary for movement.
- Epidermal regeneration.

The imbedded epidermal cells can multiply and re-form an epidermal structure under the direction of growth

factors and cell signals found in the dermis (6). The dermal signals are richest in the upper third of the dermis known as the papillary dermis. The deeper dermal layer is less able to regenerate epidermis and itself, therefore the thinner the thickness of remaining dermis in a wound, the less likely it is that the skin can regenerate. More scar will develop to replace the lost skin. If the dermis is totally destroyed, a burn cannot heal by itself and a skin graft or permanent skin substitute is required.

Of extreme importance is the psychological impact in the quality of healed or replaced skin, as this organ has a major role in human communication and helps define the individual. One of the key objectives, of the newer skin substitutes, is to restore some normalcy to the new skin cover.

BURN INJURY

A skin burn is the damage to the structure and function of skin, caused by heat or other caustic materials (7,8). Severity is based on the degree to which the outer epidermis and the inner dermis are destroyed or damaged. The most immediate and obvious injury is one due to heat. Excess heat causes rapid protein denaturation and cell damage (9,10). The depth of heat injury is dependent on the depth of heat penetration. In addition, the body's response to the burn in the form of inflammation and inflammatory mediator release, especially oxidants, results in further cell damage. The damage to skin caused by a burn is therefore a very dynamic process starting with a heat or chemical insult, and then evolving with time, especially in the deeper higher risk burn insult. The initial thickness of the skin is also a major factor as to severity as the thinner the initial skin the more severe will be the burn for the same heat insult. Children and the elderly have very thin skin (7,8).

Burn severity is determined by the depth of the burn, the burn size relative to the percent of total skin burned, and the location. The greater the function (e.g., hands) or cosmetic importance (e.g., face) the more severe the burn. Other factors include age and status of overall health. The very young and very old are at a greater risk with a burn due to an impaired ability to tolerate severe bodily trauma as well as thinner skin. The body response and postburn complications, especially infection, add to severity (7–10).

Burn Size

Burn size is defined as the percent of the persons body skin burned. In the adult, the "Rule of Nines" assessment is commonly used. Each area is considered 9% of total body surface (TBS), each leg 18%, back 18%, front torso 18%, and head being 9% of total. The palm of the patients' hand is considered to be 1% of that persons body skin surface area.

Burn Depth

Burn depth is defined by how much of the two skin layers are destroyed (Fig. 1). Burns can be categorized by degree:

First degree: confined to the outer layer only.
Second degree: also involves part of the dermis.
Third degree: destruction of both layers of skin.

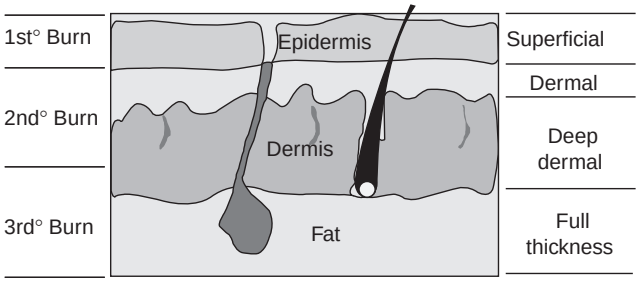


Figure 1. Schema describing burn depth with terminology.

More recent terminology is

Partial thickness: a second degree burn consisting of injury to part of the dermis.
Full thickness: a third degree burn consisting of necrosis to both layers.

Partial thickness or second degree is further divided into superficial-confined to epidermis and upper dermis and deep: when most of the dermis is destroyed.

Full thickness is further categorized as burns involving destruction of the skin alone and burn extending below skin into (e.g., muscle). The latter is often referred to as a fourth degree burn.

Superficial Second Degree

Involves the entire epidermis and no more than the upper third of the dermis is heat destroyed. Rapid healing occurs in 1–2 weeks because of the large amount of remaining skin and good blood supply. Scar is uncommon. However, in young children and the elderly, with an already thin dermis, a burn of this type actually extends to mid-dermis increasing the healing rate with an increasing risk of scarring, especially if re-epithelialization takes >3 weeks.

Initial pain is the most severe of any burn, as the nerve endings of the skin are now exposed to the air. Protection of the burn wound surface from desiccation, inflammation, and infection is necessary to optimize healing. Temporary skin substitutes are used, especially in children, to protect the wound and relieve the pain while it re-epithelializes beneath (11,12).

Deep Second Degree (Deep Partial Thickness) Burn

Most of the skin is destroyed except for small amount of remaining dermis. The wound looks white or charred indicating dead tissue. Pain is much less as the nerves are actually destroyed by the heat. Usually, one cannot distinguish a deep dermal from a full thickness (third degree) by visualization. The presence of sensation to touch usually indicates the burn is a deep partial injury rather than full thickness. There is a high risk of infection due to decreased blood flow and impaired local immune defenses (1,2). This burn typically takes several months to re-epithelialize due to few remaining epidermal cells. The quality of the epithelial covering is typically poor, being thin and friable. Wound scar is usually severe with healing (13).

Table 1. Mean Survival Rate after Burns in Burn Centers,[a] age versus Burn Size

	Age (years)						
Burn size, % of total skin[b]	0–1	2–4	5–34	35–49	50–59	60–74	>75
0–10	>95	>95	>95	>95	>95	>95	90
10–20	>95	>90	>95	>90	>90	>70	>60
20–30	>90	>90	>90	>90	>75	50	35
30–40	75	80	90	80	70	40	<20
40–50	50	65	80	60	40	10	<10
50–60	50	60	70	60	40	<25	<10
60–70	40	50	60	40	25	<10	0
70–80	35	40	45	30	25	0	0
80–90	30	35	30	30	<20	0	0
90–100	20	20	20	15	0	0	0

[a]See Refs. (10) and (14).
[b]The body surface burn is the combined second and third degree area relative to total body surface (TBS) (adapted from current literature).

Typically this depth of burn is managed by early surgical excision of the dead tissue and coverage with a skin graft (or permanent skin substitute) (11,12,14).

Third Degree (Full Thickness) Burn

Both layers of skin are completely destroyed leaving no cells to heal. Any significant burn will require skin grafting. Small burns will heal with scar. There is typically no initial pain as nerve endings are destroyed. Because of the high risk of infection and inability to heal primarily third degree burns are typically managed by early excision of the dead burn tissue and skin grafting (or permanent skin substitute) (11–14).

Survival Rate

As can be seen from Table 1, many young patients with massive burns now survive in burn centers, increasing the need for both temporary and permanent skin substitutes.

Note that survival is very high for even massive burns (<75% of the body) in older children and in the young adult. Survival decreases in the elderly due to the presence of preexisting disease and inability to withstand severe stress. Survival is also less for babies and toddlers for the same size (14).

Burn Scar and Pain Relative to Depth

The initial problems of pain and later problems of scar relate to burn depth and healing time (Table 2) (13,15,16).

Pain is a major problem with superficial burns. Temporary skin substitutes markedly decrease initial pain. Scarring is a result of loss of a large amount of dermis and a resulting prolonged wound healing period (13–17). The wound healing process includes an initial inflammation followed by an increase in wound fibroblasts, new vessels, and epithelial cell proliferation (if the burn is superficial). By 7 days, the fibroblasts are producing increased amounts of collagen, which persists until the wound has healed or is closed. Wound inflammation persists as long as the wound is open.

In superficial burns, epidermal regeneration will be relatively rapid, if the wound environment is optimized. The injured dermal elements are usually covered by new epithelium within 2 weeks if protected from environmental insults. Only modest amounts of collagen are deposited. The wound usually becomes relatively pliable with time and minimal to no wound contraction is seen. Cosmetically, the superficial second degree burn, which heals in 2 weeks, results in very minimal to no long-term scarring (15–17).

The histology of the wound bed, however, changes dramatically if it has not been re-epithelialized by 3 weeks, as would be the case with a deeper dermal burn (13–17). Fibroblasts and macrophages become the predominant cells. Large numbers of myofibroblasts also enter the wound. Besides leading to contraction, these cells continue to deposit large quantities of collagen and glycosaminoglycans. Later, closure of this wound by re-epithelialization or

Table 2. Etiology and Prognosis Relative to Burn Depth[a]

Second degree (partial thickness)	Cause	Appearance	Pain	Healing	Scar
Superficial	Hot liquid, short exposure	Wet, pink blisters	Severe	10–14 days	Minimal
Mid-dermal	Hot liquid, hot grease, or flash flame with longer exposure	Less wet, red ± blisters	Moderate	2–4 weeks	Moderate
Deep dermal	Chemicals, direct contact with flames	Dry, white	Minimal	8–16 weeks	Severe, usually needs skin graft
Third degree (full thickness)	Chemicals, flames, explosion with very high temperature	Dry, white or char	None	Needs skin graft	Mild to severe, depending on timing and type of graft

[a]The short- and long-term impact of a burn is described relative to depth.

grafting does not eliminate the stimulus for ongoing scar formation. The components produce a harder, less pliable wound shortening the scar and causing fusion of the collagen fibers in the contracted site. The process of wound contraction will lead to joint contractures.

Also, there are typically no finger like structures called rete pegs produced if healing time exceeds 4–6 weeks and the epidermis is not well anchored such that blistering is common (13,16).

The risk of hypertrophic scar formation increases with the healing time of deeper burns. Hypertrophic scar is an excess scar formation leading to a red, raised, itchy, nonpliable skin cover (13,15,16).

ROLE OF BIOACTIVE SKIN SUBSTITUTES

Outcome is defined both in terms of survival, the quality of the healed or grafted skin, and the degree of morbidity from the burn. The final objective of burn management is achieving survival as well as minimizing the morbidity such as loss of muscle and strength, maintaining quality of life, minimizing scarring, and optimizing the healing process (7,8).

The major stimulus for advances in skin substitutes is to improve the quality of the closed burn wound and avoid poor quality skin (18–20). As can be seen in the table describing burn survival, there are now many massive burns that survive necessitating the use of skin substitutes. Remaining nonburn skin is not sufficient to close these massive burns (11,12,14).

A superficial burn involves the epidermis and very little of the dermis, and the dead tissue peels off, in the form of blisters, leaving a viable wound bed that must be protected. Temporary skin substitutes can improve the healing while decreasing pain (16,17). With deeper burns involving some or all of the dermis, the dead tissue, remains adherent to the wound. The dead tissue known as the burn eschar, will then cause an inflammatory response producing both local and systemic effects. The systemic effects include a profound increase in metabolic rate with a marked increase in muscle wasting, and impairment in immune defenses (1,2). Controlling this systemic response, by earlier removal of the dead burn tissue and closure of the wound, has markedly decreased overall mortality morbidity (11–13). Skin substitutes are used to temporarily or permanently close the excised burn, especially in large burns where there isn't enough remaining skin to use for skin grafts (18–21).

The addition of biological activity to the skin substitute, is intended to improve the healing process with the intention of more rapidly healing a superficial burn and restoring valuable dermal components in a deeper burn wound bed thereby minimizing scarring and optimizing function (18–21). As stated before, the impact of the added bioactivity in burns is yet to be firmly established. A list of the noncellular components of dermis, used in available skin substitute, is shown below (22–26) (Table 3). Epidermal and dermal cells are usually also added to dermal elements in permanent skin substitutes in addition to these dermal components.

Table 3. Components of Dermis that Are Involved in Healing and Used in Skin Substitutes

Dermal Components Stimulating Healing[a]
Structural component or scaffolding
Biologically active component stimulating all phases of healing
Collagen (protein)
Scaffold for cell migration and matrix deposition
Cell guidance
Elastin (protein)
Tissue elasticity
Fibronectin (protin)
Cell-to-cell adherence
Contact orientation for cells
Increases epithelial cell division, migration
Chemo attractant for fibroblasts, macrophages
Growth factors (proteins)
Stimulate all phases of wound healing
Glycosaminoglycan (glycosylated protein)
Cell adherence properties
Conduit for healing factors
Deactivator or proteases
Scaffold or foundation for dermal elements
Hyaluronic acid (complex carbohydrate)
Maintaining matrix hydrated
Decreases inflammation
Stimulates healing
Proper cell alignment

[a]See Refs. (22–26).

AVAILABLE BIOACTIVE SKIN SUBSTITUTE

A list of skin substitutes, categorized by biologic make-up, is presented below (Table 4). All have some degree of biologic activity for improving the wound-healing environment. A disadvantage of all of these skin substitutes is the absence of active antimicrobial activity. However, early effective wound closure does decrease the risk of infection.

Skin substitutes can also be categorized as to use and indication into temporary or permanent.

Table 4. Available Skin Substitutes Are Categorized Based on Composition

Available Biologically Active Skin Substitutes
Naturally occurring tissues
Cutaneous allografts
Cutaneous xenografts
Amniotic membranes
Porcine small intestinal submucosa
Composite Synthetic-Biological
Collagen based dermal analogs
Integra
Culture-derived tissue
Bilayer human tissue
Cultured autologous keratinocytes
Fibroblast seeded dermal analogues
Epithelial seeded dermal analogue

Skin substitutes can also be categorized as to use and indication into temporary or permanent.

Table 5. Ideal Properties of a Temporary Skin Substitute[a]

Ideal Properties of a Temporary Skin Substitute
Rapid and firm adherence properties for closure of the wound
Relieves pain
Easily applied and secured
Does not incite inflammation
Stimulates wound healing
Barrier to microorganisms
Avoids wound desiccation
Optimizes healing environment
Does not cause hypertrophic tissue response
Hemostatic
Prevents evaporative water loss
Flexible yet durable
Easy to remove when
Wound has re-epithelialized
Wound ready for grafting
Cannot transmit disease
Inexpensive
Long shelf life
Does not require refrigeration

[a]These properties are then sought when developing new skin substitutes.

Temporary skin substitutes are used to help heal the partial thickness burn (or donor site) and close the clean excised wound until skin is available for grafting. There are typically no living cells present.

Permanent skin substitutes are used to replace lost skin providing either epidermis or dermis, or both and to provide a higher quality of skin than a thin skin graft. Most permanent skin substitutes contain viable skin cells as well as components of the dermal matrix.

Temporary Bioactive Skin Substitutes

The purpose of a temporary skin substitute is twofold (Table 5). Temporary skin substitutes are typically a bilayer structure. There is an outer epidermal analog and a more biologically active inner dermal analogue (18–21). The first objective is to close the wound, thereby protecting the wound from environmental insults (18,19,27–29). The second objective is to provide an optimal wound healing environment by adding dermal factors that activate and stimulate wound healing (18–29). Biologically active dermal components naturally are typically provided to the inner layer, which is then applied to the remaining dermis in a partial thickness burn or to an excised wound. Below is a list of the commonly available dermal matrix elements present in these products, and their actions.

The currently available products are listed in Table 6.

Human Allograft (Cadaver Skin)

Human allograft is generally used as a split-thickness graft after being procured from organ donors (30–32). When used in a viable fresh or cryopreserved state, it vascularizes and remains the "gold standard" of temporary wound closures. It can be refrigerated for up to 7 days, but must be stored frozen for extended periods. It is also used in a

Table 6. Available Bioactive Temporary Skin Substitutes[a]

Product	Company	Tissue of Origin	Layers	Category	Uses	How Supplied
Human allograft	Skin bank	Human cadaver	Epidermis and dermis	Split thickness skin	Temporary coverage of large excised burns	Frozen in rolls of varying size
Pig skin xenograft	Brennan Medical St. Louis, MO	Pig dermis	Dermis	Dermis	Temporary coverage of partial thickness and excised burns	Frozen or refrigerated in rolls
Human amnion	On site procurement	Placenta	Amniotic membrane	Epidermis Dermis	Same as above	Refrigerator
Oasis	Healthpoint, LTD San Antonio, TX	Xenograft	Extracellular wound matrix from small intestine submucosa	Bioactive Dermal like Matrix	Superficial burns Skin graft donor sites Chronic wounds	Room temperature storage Multiple sizes 3×3.5 cm 7×20 cm
Biobrane	Dow Hickam/Bertek Pharmaceuticals	Synthetic with added denatured bovine collagen	Bilayer product outer silicone Inner nylon mesh with added collagen	Synthetic epidermis and dermis	Superficial partial thickness burns, Temporary cover of excised burns	Room temperature storage 15×20 in. 10×15 cm 5×15 in. 5×5 in.
Transcyte	Smith and Nephew Wound Management Largo, FL	Allogenic dermis	Bilayer product Outer silicone Inner nylon seeded with neonatal fibroblasts	Bioactive dermal matrix components on synthetic dermis and epidermis	Superficial to mid-partial thickness burns Temporary coverage of excised burns	Frozen in 5×7.5 in. sheets

[a]The names and properties of available temporary skin substitutes, with some biological activity are described. Also listed are the indications of the various products.

Table 7. Advantages and Disadvantages of Allograft Skin as a Skin Substitute

Allograft Skin
Advantages
A bilayer skin providing epidermal and dermal properties
Revascularizes maintaining viability for weeks
Dermis incorporates into the wound
Disadvantages
Epidermis will reject
Difficult to obtain and store
Risk of disease transfer
ExpensiveNeed to cryopreserve

nonviable state after preservation in glycerol or after lyophilization: however, most existing data describe best results when it is used in a viable state. The epidermal component provides a barrier until rejected by the host in 3–4 weeks. The dermis revascularizes and incorporates.

Homograft, another term for human allograft, can only be obtained from a tissue bank as strict protocols are required for harvesting and storage. Donors must be rigidly screened for potential viral and bacterial disease to avoid any transmission of disease. The product is in limited supply and very expensive.

The primary indication for use is to cover a large excised burn wound until an autogenous skin or a permanent skin substitute becomes available. Allograft is also used to cover a wide meshed skin graft, sealing the interstices during the healing process (Table 7).

Xenografts

Although various animal skins have been used for many years to provide temporary coverage of wounds, only porcine xenograft is widely used today (33,34) (Table 8). The epidermis of the porcine xenograft is removed and the split thickness dermis is provided in rolls. Split-thickness porcine dermis can be used after cryopreservation, or after glycerol preservation. It effectively provides temporary coverage of clean wounds such as superficial second degree burns and donor sites (33,34). Porcine xenograft does not vascularize, but it will adhere to a clean superficial wound and can provide excellent pain control while the underlying wound heals. In general, xenograft is not as effective as homograft but is more readily available and less expensive. Primary indications are for coverage of partial thickness burns during healing and used burn wounds prior to skin grafting.

Table 8. Advantages and Disadvantages of Xenograft as a Skin Substitute

Xenografts
Advantages
Good adherence
Decreases pain
More readily available compared to allograft
Bioactive (collagen) inner surface with fresh product
Less expensive than allograft
Disadvantages
Does not revascularize and will slough
Short term use
Need to keep the fresh product frozen

Table 9. Advantages and Disadvantages of Human Amnion as a Skin Substitute

Human Amnionic Membrane
Advantages
Acts like biologic barrier of skin
Decreases pain
Easy to apply, remove
Transparent
Disadvantages
Difficult to obtain, prepare and store
Need to change every 2 days
Disintegrates easily
Risk of disease transfer

Human Amnion

Human amniotic membrane is used in many parts of the world as a temporary dressing for clean superficial wounds such as partial-thickness burns, donor sites, and freshly excised burns (35,36). Amniotic membrane is generally procured fresh and used after brief refrigerated storage. It can also be used in a nonviable state after preservation with glycerol. Amnion does not vascularize but still can provide effective temporary wound closure. The principal concern with amnion is the difficulty in screening the material for viral diseases. The risks of disease transmission must be balanced against the clinical need and the known characteristics of the donor (Table 9). The primary indications are the superficial burn and the excised wound.

Oasis Wound Matrix

This product is made of the submucosa of the porcine small intestine found between the mucosa and muscularis, in the wall of the porcine small intestine (37,38). The freeze dried acellular natural matrix retains its natural collagen and matrix structure and contains most of the bioactive matrix proteins found in the human dermis (Table 10).

Table 10. Advantages and Disadvantages of OASIS as a Skin Substitute

Oasis Wound Matrix
Advantages
Excellent adherence
Decreased pain
Provides bioactive dermal like properties
Long shelf life, store at room temprature
Relatively inexpensive
Disadvantages
Mainly a dermal analog
Incorporates and may need to be reapplied

The submucosal layer is ~0.2 mm in thickness, but is quite durable. The product is freeze-dried removing the cells. The product is sterile, porous, biocompatible and nonimmunogenic. It has a long shelf life and can be stored at room temperature. The OASIS is incorporated into the wound bed over ~7 days and needs to be reapplied if the wound has not yet healed. The outer-barrier function is diminished with incorporation.

The primary indication is for use in difficult to heal nonburn wounds. It's use in burns is for the partial thickness burn and the skin graft donor site.

Biobrane

This product is a two-layer membrane (39,40). The outer epidermal analog is constructed of a thin silicone film with barrier functions comparable to skin. Small pores present in silicone allow for exudates removal and has permeability to topical antibiotics.

The inner dermal analogue is composed of a three-dimensional (3D) irregular nylon filament weave upon which is bonded type I collagen peptides. The surface binding of inner membrane is potentiated by collagen–fibrin bonds as well as fibrin deposition between the nylon weave. A thin water layer is maintained at the wound surface for epidermal cell migration maintaining moist wound healing. Excellent adherence to the wound significantly decreases pain in the superficial partial thickness burns. The silicone and nylon weave provides flexibility. The biobrane is removed once the partial thickness wound has re-epithelialized or the covered excised burn wound is ready for grafting. However, if left in place for >2 weeks the product is difficult to remove as tissue grows into the inner layer. Biobrane L contains a nylon fabric woven from monofilament threads that provide a less dense matrix and less adherence, preferred (e.g., on a donor site). There is likely very little direct bioactivity from the collagen peptides (40). The product has a long shelf life and can be stored at room temperature. It is also relatively inexpensive (Table 11).

The primary indication is for closure of the clean superficial burn or the excised burn wound.

Table 11. Advantages and Disadvantages of the Use of BIOBRANE as a Skin Substitute

Biobrane
Advantages
Bilayer analog
Excellent adherence to a superficial burn
Decreases pain
Maintains flexibility
Easy to store with long shelf life
Relatively inexpensive
Disadvantages
Has very little direct bioactivity
Difficult to remove if left in place >2 weeks

Transcyte

This product is also a bilayer skin substitute (41,42). The outer epidermal analogue is a thin nonporous silicone film with barrier functions comparable to skin. The inner dermal analog is layered with human neonatal foreskin fibroblasts that produce products, mainly collagen type I, fibronectin, and glycosaminoglycans.

Table 12. Advantages and Disadvantages of TransCyte as a Skin Substitute

Transcyte
Advantages
Bilayer analog
Excellent adherence to a superficial to middermal burn
Decreases painProvides bioactive dermal components
Maintains flexibility
Good outer-barrier function
Disadvantages
Need to store frozen till use
Relatively expensive

A subsequent cryopreservation destroys the fibroblasts, but preserves the activity of fibroblast derived products on the inner surface. These products are then anticipated to stimulate the wound healing process (Table 12). A thin water layer is maintained at the wound surface for epidermal cell migration.

The nylon mesh provides flexibility and excellent adherence properties significantly decrease pain in the partial thickness burn. The product is peeled off after the wound has re-epithelialized.

The Transcyte must be stored at $-70\,^{\circ}C$ in order to preserve the bioactivity of the dermal matrix products. The primary indication is for closure of the clean superficial to mid-dermal burn, especially useful in children. Transcyte is also indicated for the temporary closure of the excised wound prior to grafting. Tissue ingrowth tends to be less of a problem even if the product is kept in place for >2 weeks.

Permanent Skin Substitutes

The purpose of these products is to replace full thickness skin loss as well as to improve the quality of the skin, which has been replaced after a severe burn (20–23).

As opposed to the bilayer concept of the ideal temporary skin substitute, permanent skin replacement is much more complex.

This area can be arbitrarily divided into two approaches (21–23). The first approach is the use of a bilayer skin substitute, with the inner layer being incorporated into the wound as a neodermis, rather than removed like a temporary product. The outer layer is either a synthetic to be replaced by autograft (epidermis) or actual human epithelial cells. If the outer layer is composed of epithelial cells that will form epidermis barrier function is not sufficiently developed at placement to act immediately as an epidermal barrier.

The second approach is the provision of either just an epidermal or a dermal analog or simply a coculture of cells containing elements of both. These products are technically not permanent skin substitutes (Table 13) upon initial placement as there is no bilayer structure until the product

Table 13. Components of Permanent Skin Substitutes

Permanent Skin Replacement
Bilayer structures with biologic dermal analog and either synthetic or biologic epidermal analog
Skin components
Epidermal cells alone
Dermis alone
Coculture of epidermal cells and fibroblasts

Table 14. Ideal Properties of a Permanent Skin Substitute[a]

Rapid and excellent adherence properties
Easily applied and secured to an excised wound
Minimum wait period from time of burn to availability of skin substitute
Bilayer tissue containing both epidermal and dermal eliminates to best replicate normal skin
Rapid incorporation
Cannot transmit disease
Good functional and cosmetic result
Inexpensive

[a]As yet, the ideal skin does not exist.

evolves once placed on the wound. Both approaches will be discussed as will the ideal properties (Table 14).

The ideal property would be that of a bilayer structure.

The currently available clinical products are listed below (Table 15). There are a number of permanent skin substitutes in the development stage, which will not be listed.

Table 16. Advantages and Disadvantages of the Use of Apligraf As a Skin Substitute

Apligraf
Advantages
Bilayer skin containing human neonatal cells
Not requiring patients skin biopsy
No lag time for production
Contains epidermal and dermal functions (eventually)
Does not need to be frozen
Disadvantages
Made in small pieces and not practical for large burn
Cannot be stored for over 24 h
Relatively expensive

Apligraf

The dermal analog is made of fibroblasts from neonatal foreskin populated in a bovine type I collagen matrix. This layer incorporates into the excised full thickness wound adding a dermal component. Epithelial cells (keratinocytes) are also obtained from neonatal foreskin. No Langerhans cells at present so rejection does not occur. The epithelial cells divide, migrate and form an epidermis, which will eventually provide biologic barrier function. Donor foreskin is screened for viruses, which could be transmitted. The product is indicated mainly for chronic wounds. It's use in burns is currently off label (44,45). Apligraf is typically provided as a 7.5 cm diameter disk that is 0.75 mm thick. The advantages are that the product is already made and does not depend on obtaining the patients own cells (Table 16). However, the product must be shipped the night before use in a polyethylene bag in agar nutrient medium and a 10% CO_2 content and stored at room temperature until used (within 24 h).

Table 15. Properties and Uses of the Currently Available Permanent Skin Substitutes

Available Permanent Skin Substitutes						
Product	Company	Tissue of Origin	Layers	Category	Uses	How supplied
Apligraf	Organogenesis, Inc and Novartis Pharmaceuticals Corp	Allogenic Composite	Collagen matrix seeded with human neonatal keratinocytes and fibroblasts	Composite: Epidermis and dermis	Chronic wounds, often used with thin STSG Excised deep burn	7.5 cm diameter disk 1/pack
OrCel	Ortec International Inc.	Allogenic Composite	Collagen sponge seeded with human neonatal keratinocytes and fibroblasts	Composite: Epidermis and Dermis	Skin graft donor site, chronic wounds	6 × 6-cm sheets
Epicel	Genzyme Tissue Repair Corp	Autogenous keratinocytes	Cultured autologous keratinocytes	Epidermis Only	Deep partial and full thickness burns >30% TBSA	50 cm^2 sheets in culture medium
Alloderm	Life Cell	Allogenic dermis	Acellular Dermis (processed allograft)	Dermis only	Deep partial and full thickness burns, Soft tissue replacement, Tissue patches	1 × 2–4 × 12 cm
Integra	Integra Life Science Corp	Synthetic	Silicone outer layer on collagen GAG dermal matrix	BioSynthetic Dermis	Full thickness soft tissue defects definitive "closure" requires skin graft	2×2 in. 4×10 in. 8×10 in. 5/pack

Table 17. Advantages and Disadvantages of Orcel as a Skin Substitute

Orcel
Advantages
Eventual bilayer skin
Not requiring patient's cells
Does not need to be frozen
Disadvantages
Bilayer structure requires 14 days to develop once applied
Not indicated for excised burns
Short storage time
Relatively expensive

Orcel

The product is produced as a coculture of keratinocytes and fibroblasts (neonatal foreskin) in a cross-linked bovine collagen sponge (46). The donor tissue is screened for viruses. The nonporous side of the sponge is seeded with the keratinocytes and the porous side with fibroblasts. After application, a neodermis forms. Once applied the epidermis and an epidermal barrier requires 14 days to develop as the keratinocytes migrate and divide on the surface (Table 17).

The product is indicated mainly for chronic wounds but is also indicated for coverage of split thickness skin graft (STSG) donor sites. This product is not currently indicated for use on excised burn. The purpose for the STSG is to provide a better functional and cosmetic outcome compared to a healed donor site. It is shipped in a package filled with it's culture medium which is stored at room temperature until used, usually within a few days.

Epicel

This product, used mainly for very large burns is composed of the patients skin epithelial cells and referred to as cultured epithelial autograft (CEA). Therefore, only the epithelial layer is provided (47–49). The product is made from a small biopsy of normal skin (2×2 cm) from the burn patient. The epithelial cells are extracted and cultured. Use of a cell culture technique allows the keratinocytes to be grown in a thin sheet 10,000 times larger than the initial biopsy. This process does require 2–3 weeks from the time of biopsy. Often the burn wound is excised and covered with homograft (allograft) until the cells are ready to be transplanted. The CEA is then applied to the clean excised (or allograft covered) wound.

Table 18. Advantages and Disadvantages of Epicel as a Skin Substitute

Epicel
Advantages
Patients own keratinocytes expanded several thousand fold
Small skin biopsy required
Can cover very large surfaces with reasonable graft take
Used in large burns
Disadvantages
2–3 week lag time for production
Provides only the epidermal layer
Epithelial layer can be quite fragile for some time
Needs to be used immediately on delivery
Very expensive

The CEA is supplied in sheets 2–6 cells thick on small pieces of petroleum gauze (50 cm^2), which are bathed in culture medium. Immediate application is necessary. The CEA grafts are very fragile and easily rubbed off for at least several weeks. The backing is removed in several weeks as the CEA thickens and adheres. Graft take ranges from 30 to 75% of total epithelium applied. The epithelium gradually thickens but has a low resistance to sheer forces for some time. Application of allograft dermis, prior to CEA grafting appears to improve skin quality.

The primary indication is for very large burns.

Alloderm

This product is basically treated human allograft with the epidermis removed (50–52). The dermis is treated to produce a copreserved lyophilized allodermis, which incorporates. The product is used as a dermal implant. Therefore application of a thin epithelial autograft is required.

Primary indication is for use in the replacement of soft tissue defects. This product is not commonly used in large burns. Typically the alloderm is applied to an excised wound and then a split thickness skin graft is placed on top of the alloderm (Table 19). The product has a long shelf life in its lyophilized form. It requires rehydration prior to use.

Integra

This product is composed of a dermal analog made of a biodegradable bovine collagen-glycosaminoglycan copolymer matrix The collagen and glycosaminoglycan is cross-linked to attempt to maximize ingrowth of the patients own cells (53–55) (Table 20).

The epidermal analog is a thin silicone elastomer providing temporary barrier protection. After the dermal analog incorporates and the surface revascularizes, at ~2–3 weeks, the silicone layer is removed and replaced with a very thin skin graft from the patient (or CEA cells). The Integra needs to be carefully immobilized for the first 2 weeks as movement will cause devascularization and loss of the product. The primary indication is the treatment of large deep burns as well as reconstruction procedures. The

Table 19. Advantages and Disadvantages of Alloderm as a Skin Substitute

ALLODERM
Advantages
Easy to store, an off the shelf product
Does not require skin bank
Comes in large and small pieces
Disadvantages
Requires thin skin graft to provide epidermis
Two procedures required to achieve bilayer skin
Relatively expensive

Table 20. Advantages and Disadvantages of Integra as a Skin Substitute

Integra
Advantages
Provides thick dermal analog
Reasonable shelf life
No risk of transmitting viruses
Relatively inexpensive
Used in large burns
Disadvantages
Need to provide epidermis from the patient
Dermal cells must come from the patient requiring product incorporation
Two procedures required to achieve bilayer skin

incorporated neodermis appears to improve the function of the final skin once the epithelial graft is applied. The product is provided in a number of sizes and sheets stored in 70% isopropyl alcohol. Shelf life is very good.

SUMMARY

The scientific principles and practical approaches, to replacing skin either temporarily or permanently are advancing at a rapid rate. Much of these advances can be attributed to both advances in the field of bioengineering as well as increasing interest in optimizing the outcome of the burned skin.

The ideal properties of a bioactive temporary and a permanent skin substitute have been well defined. As expected, the properties of temporary skin substitutes are more concrete, easier to categorize and determine efficacy. A bilayer structures is the current standard with the dermal component being bioactive. Permanent skin replacement on the other hand is much more complex. A variety of approaches are being used which can be loosely categorized as either use of bilayer products (usually the outer layer to be replaced by epidermal autograft) or replacement of either dermal or epidermal elements separately. The terminology of the latter approach is difficult because these component products are really not permanent skin substitutes on initial application but become so only when all the elements are in place.

An understanding of the properties of each product is essential for the user to optimize outcome.

BIBLIOGRAPHY

1. Mast B. The skin. In: Cohen K, Diegelmann I, editors. Wound Healing. Philadelphia: WB Saunders; 1992. p 344–355.
2. Wright N, Allison M. The Biology of Epithelial Cell Populations. Clarendon Press; 1984. p 283–345.
3. Stenn S, Malhotra R. Epithelialization. In: Cohen C, editor. Wound Healing. Biochemical and Clinical Aspects. Philadelphia: WB Saunders; 1992. p 115–127.
4. Grillo H. Origin of fibroblasts in wound healing. Ann Surg 1963;157:453–467.
5. Karasck M. Mechanism of angiogenesis in normal and diseased skin.
6. Raghow R. The role of extracellular matrix in post-inflammatory wound healing and fibrosis. FASEB J 1994;8:823–850.
7. Demling R. Burn care. In: Wilmore D, editor. ACS Surgery. New York: Web MD; 2002. p 479.
8. Herndon D. Total Burn Care. Philadelphia: WB Saunders; 2002.
9. Neely A, Brown R, Chendening C, et al.Proteolytic activity in human burn wounds. Wound Rep Regen 1997;5:302–309.
10. Muller M, Pegg S, Rule R. Determinants of death following burn surgery. Br J Surg 2001;88:583–587.
11. Komgova R. Burn wound coverage and burn wound closure. Acta Chir Plast 2000;42:64–68.
12. Sheriden R. Management of burn wounds with prompt excision and immediate closure. J Inten Care Med 1994;9:6–17.
13. Demling R, DeSanti L. Scar management strategies in wound care. Rehab Manage 2001;14:26–32.
14. Spres M, Herndon D, et al. Prediction of mortality from catastrophic burns in children. Lancet 2003;361:989–994.
15. Ladin D, Garner W, Smith D. Excessive scarring as a consequence of healing. Wound Repair Reg 1994;3:6–14.
16. Scott P, Ghabary A, Chambers M, et al. Biologic basis of hypertrophic scarring. Adv Struct Biol 1994;3:157–165.
17. Erlich HP, Krummell T. Regulation of wound healing from a connective tissue perspective. Wound Repair Reg 1995;4: 203–210.
18. Badylak S. The extracellular matrix as a scaffold for tissue reconstruction. Cell Develop Biol 2002;13:377–383.
19. Jones L, Currie L, Martin R. A guide to biological skin substitutes. Br J Plast Surg 2002;55(3): 185–193.
20. Gallico GG. Biologic skin substitutes. Clin Plast Surg 1990; 512–520.
21. Sheridan R, Tompkins R. Alternative wound coverings. In: Herndon D, editor. Total Burn Care. Philadelphia: WB Saunders; 2003. p 712.
22. Clark R, Folkvard J, Wortz R. Fibronectins, as well as other extracellular matrix proteins mediate human keratinocytes adherence. J Invest Dermatol 1985;84:378–383.
23. Clore J, Cohan I, Diegelmann R. Quantitation of collagen types I and III during wound healing. Proc Soc Exper Biol Med 1979;161:337–340.
24. Doillon C, Dunn M, Bender E, et al. Collagen fiber formation in repair tissue: Development of strength and toughness. Collagen Relat Res 1985; 481–485.
25. Takashima A, Grinnell F. Human keratinocytes adhesion and phagocytosis. Prompted by fibronectin. J Invest Derm 1984;83:352–358.
26. Miller E, Gay S. Collagen structures and function in wound healing: biochemical and clinical aspects. In: Cohen K, editor. Philadelphia: Saunders; 1992. p 130.
27. Nowicki CR, Sprenger C. Temporary skin substitutes for burn patients: a nursing perspective. J Burn Care Rehab 1988;9(2):209–215.
28. Shakespeare P. Survey: use of skin substitute materials in UK burn treatment centers. Burns 2002;28(4):295–297.
29. Smith K, Rennie MJ. Management of burn injuries: A rationale for the use of temporary synthetic substitutes? Prof Nurse 1991;571–574.
30. Bondoc CC, Burke JF. Clinical experience with viable frozen human skin and a frozen skin bank. Ann Surg 1971;174:371–382.
31. Herndon DN. Perspectives in the use of allograft. J Burn Care Rehab 1997;18:56.
32. May SR, Still JM Jr., Atkinson WB. Recent developments in skin banking and the clinical uses of cryopreserved skin. (Review) J Med Assoc GA 1957;73:233–236.

33. Song IC, Bromberg BE, Mohn MP, Koehnlein E. Heterografts as biological dressings for large skin wounds. Surgery 1966;59:576–583.
34. Elliott RA Jr., Hoehn JG. Use of commercial porcine skin for wound dressings. Plast Reconstr Surg 1973;52:401–405.
35. Ramakrishnan KM, Jayaraman V. Management of partial thickness burn wounds by amniotic membrane: A cost-effective treatment in developing countries. Burns 1997;23 (Suppl. 1): 533–536.
36. Ganatra MA, Durrani KM. Method of obtaining and preparation of fresh human amniotic membrane for clinical use. J Pakistan Med Assoc 1996;46:126–128.
37. Brown-Estris M, Cutshall W, Hiles M. A new biomaterial derived from small intestinal submucosa and developed into a wound matrix device. Wounds 2002;14:150–166.
38. Demling R, Niezgoda J, Haraway G, Mostow E. Small intestinal submucosa wound matrix and full thickness venous ulcers. Wounds 2004;16:18–23.
39. Smith DJ Jr. Use of biobrane in wound management. J Burn Care Rehab 1995;16:317–320.
40. Yang J, Tsai Y. Clinical comparison of commercially available Biobrane preparations. Burns 1989;15:197–203.
41. Purdue G, Hunt J, Still M, et al. A multicenter clinical trial of a biosynthetic skin replacement, Dermagraft-T compared with cryopreserved human cadaver skin for temporary coverage of excised burn wounds. J Burn Care Rehab 1997;18:52–57.
42. Demling RH, DeSanti L. Management of partial thickness facial burns (comparison of topical antibiotics and bio-engineered skin substitutes). Burns 1999;25:256–261.
43. Bell YM, Falabella AF, Eaglstein WH. Tissue engineered skin. Current status in wound healing. Am J Clin Dermatol 2001;2:305–313.
44. Folanga V, Sabolinski M. A bilayered living skin construct (APLIGFAF) accelerates complete closure of hard-to-heal venous ulcers. Wound Repair Regen 2000;7:201–207.
45. Fivenson DP, Scherschun L, Choucair M, Kukuruga D, Young J, Shwayder T. Graftskin therapy in epidermolysis bullosa. J Am Acad Dermatol 2003;48:886–892.
46. Still J, Glat P, Silverstein P, Griswold J, Mozingo D. The use of a collagen sponge/living cell composite material to treat donor site burn patients. Burns 2003;29(8):837–841.
47. Sheridan RL, Tompkins RG. Cultured autologous epithelium in patients with burns of ninety percent or more of the body surface. J Trauma 1995;38:48–50.
48. Rue LW III, Cioffi WG, McManus WF, Pruitt BA Jr. Wound closure and outcome in extensively burned patients treated with cultured autologous keratinocytes. J Trauma 1993; 34:662–667.
49. Loss M, Wedler V, Kunzi W, Meuli-Simmen C, Meyer VE. Artificial skin, split-thickness autograft and cultured autologous keratinocytes combined to treat a severe burn injury of 983% of TBSA. Burns 2000;26:644–652.
50. Buinewicz B, Rosen B. Acellular cadaveric dermis (AlloDerm): a new alternative for abdominal hernia repair. Ann Plast Surg 2004;52:188–194.
51. Wax MK, Winslow CP, Andersen PE. Use of allogenic dermis for radial forearm free flap donor site coverage. J Otolaryngol 2002;31:341–345.
52. Druecke D, Steinstraesser L, Homann HH, Steinau HU, Vogt PM. Current indications for glycerol-preserved allografts in the treatment of burn injuries. Burns 2002;28 (Suppl. 1): S26–30.
53. Wisser D, Rennekampff HO, Schaller HE. Skin assessment of burn wounds covered with a collagen based dermal substance in a 2-year follow-up. Burns 2004;30:399–401.
54. Navsaria HA, Ojeh NO, Moiemen N, Griffiths MA, Frame JD. Re-epithelialization of a full-thickness burn from stem cells of hair follicles micrografted into a tissue-engineered dermal template (Integra). Plast Reconstr Surg 2004;113:978–981.
55. Frame JD, Still J, Lakhel-LeCoadou A, Carstens MH, Lorenz C, Orlet H, Spence R, Berger AC, Dantzer E, Burd A. Use of dermal regeneration template in contracture release procedures: A multicenter evaluation. Plast Reconstr Surg 2004;113:1330–1338.

See also BIOCOMPATIBILITY OF MATERIALS; ENGINEERED TISSUE; SKIN, BIOMECHANICS OF; SKIN TISSUE ENGINEERING FOR REGENERATION.

SKIN TISSUE ENGINEERING FOR REGENERATION

BRENDAN A. HARLEY
IOANNIS V. YANNAS
Massachusetts Institute of Technology
Cambridge, Massachusetts

MAMMALIAN RESPONSE TO INJURY

Scale of Functional Deficit

Medical treatment options for the loss of normal organ or tissue function depend heavily on the scale of the defect, either macromolecular- or organ-scale. Since antiquity, macromolecular-scale defects have been treated with chemical therapeutics such as herbs and potions. More recently, pharmaceuticals, vitamins, hormones, and antibiotics have been used to treat a vast array of medical maladies that are caused by a macromolecular defect; these treatment regimens have been used successfully to replace or correct a missing function on the molecular scale. Organ-scale defects present a significantly larger wound site and require considerably different treatment practices.

Large-scale failures of a tissue or organ are created primarily by disease or by an acute or chronic insult; injury or damage of this type typically results in wound sites on the scale of a millimeter or centimeter, as opposed to the nanometer scale of a molecular defect. This type of damage cannot be treated with drugs because the problem is the failure of a mass of tissue, including a large number of cells, soluble proteins and cytokines, and insoluble extracellular matrix. The typical organismic response to an injury at the organ-scale is cell-mediated wound contraction and synthesis of nonphysiologic tissue (scar); this process is termed repair. Regeneration of lost or damaged tissue describes a process marked by synthesis of physiologic (normal, functional) tissue in the wound site. There are a few notable exceptions to the rule that the mammalian response to organ-scale injuries is repair. The epithelial tissues of the skin, genitals, cornea, and gastrointestinal tract all regenerate spontaneously, and the liver has shown the ability to synthesize a substantial organ mass without recovery of the original organ structure, but also without contraction or scar synthesis (1). Despite these few notable exceptions, the mammalian response to the loss or damage of a tissue or organ is almost exclusively repair, an irreversible

response resulting in the formation of scar tissue that lacks the structure and function of the damaged organ or tissue.

Regeneration Versus Repair

As previously noted, there are two possible outcomes of the mammalian healing process following acute or chronic injuries: regeneration or repair. Regeneration is characterized by synthesis of a replacement tissue in the anatomical wound site that is structurally and functionally similar to the original tissue. Repair is characterized by synthesis of scar tissue (nonphysiological tissue) to replace the normal tissue lost due to injury. In addition to synthesis of new, nonphysiologic tissue, contraction of the wound site is also observed during repair. Cell-mediated contraction of the wound site has been observed in many different species to varying degrees at different wound sites (1).

Wound closure following severe injury occurs by one or more of three distinct processes. The initial wound area can close by contraction (C), scar formation (S), and/or regeneration (R). The process of wound closure can be represented quantitatively by the defect closure rule: $A_C + A_S + A_R = 100$ ($A_X \equiv$ percentage wound area closed by process X); these three processes are the only mechanisms by which wound closure takes place. In adult mammals, chronic and acute injuries have a common clinical outcome because the repair processes responsible for wound healing close the wound through contraction and scar formation only ($A_R = 0$). With the exception of a certain class of injuries, such as injuries to epithelial tissues in select organs, the ability to regenerate tissues and organs is lost in mammalian adults. Spontaneous regeneration (regeneration without external stimulation) is observed in very specific tissues in adult mammals following minor injuries such as a small skin scrape or a first or second degree burn, while more severe injuries such as a deep skin wound or third degree burn result in contraction and scar formation. A review of the available data comparing cases of regeneration with those of repair has led to the proposal of antagonistic relationship between contraction and regeneration; in cases of adult healing where contraction occurs, regeneration is not observed, and regeneration is observed when wound contraction is blocked (1–3).

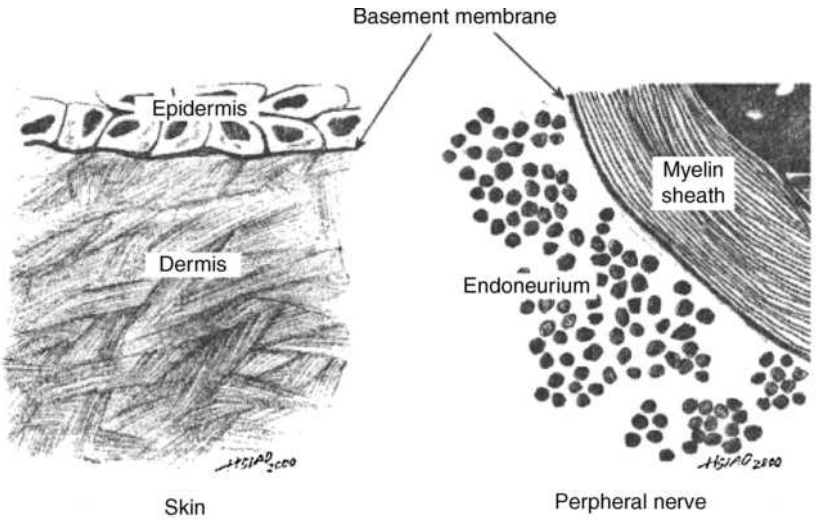
Mammalian Response to Injury

While the mammalian adult responds to severe injury by repair processes resulting in the formation of scar ($A_C + A_S = 100$; $A_R = 0$), the mammalian fetus is able to regenerate the lost tissue spontaneously ($A_C, A_S \ll A_R$) (4,5). For the sake of this article, the fetal classification refers to all mammalian fetuses that respond to injury with regeneration processes while the adult classification refers to all mammals (adult as well as juvenile) that respond to injury with repair processes. Modifying the adult mammalian wound healing response to more closely resemble that observed with the fetus has been an area of extensive study. In addition to understanding the differences between the fetal and adult healing processes, analogs of the extracellular matrix (ECM) have been used as templates for a variety of tissue engineering related disciplines toward the goal of inducing regeneration of tissues damaged by severe injury where the organism would normally respond to injury with repair processes.

The extensive study of the mammalian response to injury has focused on understanding the mechanism and timing of the transition from regeneration to repair processes during the fetal–adult transition of development and whether regeneration can be induced in the adult mammals once the initial transition has taken place. For mammals, the transition from wound closure by regeneration to wound closure by scar synthesis and contraction takes place during the third trimester of gestation (1,4,5). While certain classes of amphibians have been studied extensively throughout metamorphosis from tadpole to young adult (6), there is for the most part much less information available about higher mammals. A detailed study of wound closure mechanisms of the amphibian frog during larval development was based on measurement of the fractional extent of wound contraction, scar formation, and regeneration during development. With increasing fetal development, wound closure depended increasingly less on regeneration and correspondingly more on contraction. Once metamorphosis to the young adult frog was complete, regeneration was not observed while contraction and synthesis became the predominant mechanisms for wound closure (7). These and several other observations support the conclusion that there is an antagonistic relationship between regeneration and wound contraction. To date, while the causes for the transformation in the mode of mammalian response to injury from the fetus to the adult are not known, adult mammals are known to be unable to regenerate tissue lost due to severe injury and close severe wounds by contraction and scar formation.

There are three tissue layers that are grouped together in sequence in all organs, namely, epithelia, the basement membrane, and the stroma (1,8). This sequence has been termed the tissue triad. A pattern has been observed as a result of examinations of the structure and tissue-specific response to injury of the three tissues that to date have been induced to regenerate: skin, peripheral nerves, and conjunctiva. Understanding of the response of each member of the tissue triad to injury aids in shedding light on the process of regeneration and has suggested methodology for inducing regeneration in tissues that do not regenerate spontaneously. A layer of epithelial cells (epithelia) covers all surfaces, tubes, and cavities of the body; this layer is totally cellular with no ECM component. The epithelia is completely cellular, avascular, and is the only member of the tissue triad that does not contain an extracellular matrix. Developmental and functional similarities between epithelial cells and tissues from a variety of different tissues and organs such as the skin and peripheral nerves have been observed; these observations have suggested that the tissue triad can be used as a more general tool to understand organismic response to injury (1).

The basement membrane (also termed basal lamina) is a continuous layer of tissue separating the epithelial layer from the stroma. In all tissues, the basement membrane is acellular, and no blood vessels pass through the basement membrane layer. The stroma contains connective tissues as well as the blood supply, and provides a reservoir for



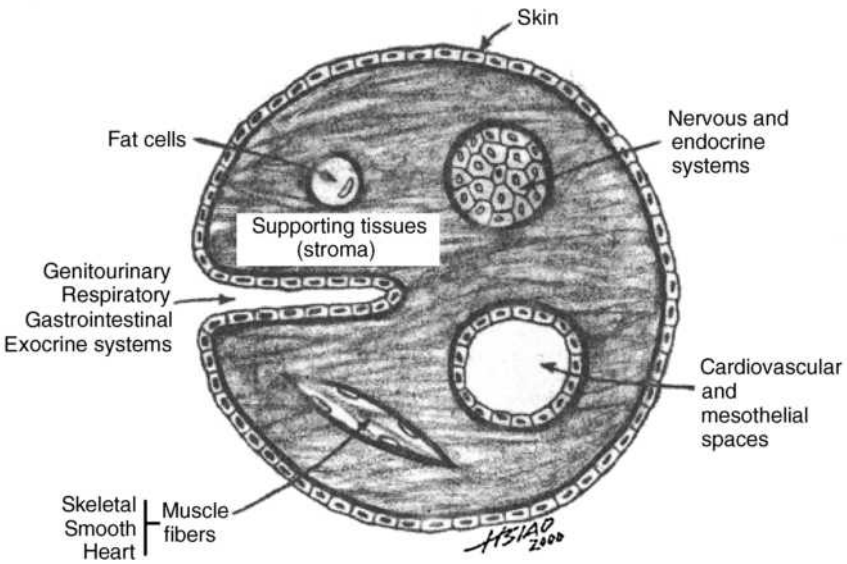


Figure 1. Schematics of the tissue triad structure observed in mammalian tissue. Top: Tissue triad of skin and peripheral nerves. The basement membrane is a thin extracellular matrix tissue located between the cellular and nonvascular epithelia (epidermis, myelin sheath) and the cellular, vascularized stroma (dermis, endoneurium). Bottom: Diagram of the distribution of epithelial, basement membrane, and stomal tissues in the mammalian system. Examples of stromal tissues are bone, cartilage, and their associated cell types as well as elastin and collagen. Examples of epithelial tissues are those covering the surface of the genitourinary, respiratory, and gastrointestinal tracts as well as surfaces of the mesothelial cells in body cavities, muscle fibers, fat cells, and endothelial cells in the cardiovascular system (1).

nutrient uptake to and waste removal from the basement membrane and epithelia. Figure 1 provides a basic diagram of the organization of the tissue triad in the adult mammalian system.

Using the tissue triad as a guide, we can identify similarities among the three tissues. In skin, peripheral nerves, and the conjunctiva, there are tissues that spontaneously regenerate (the epithelia and basement membrane) and a tissue that does not (the stroma). The regenerative capacity of epithelial tissue and the basement membrane as well as the irreversible (repair) nature of stromal wound healing have been extensively reported (1,9–13). Specifically, the stroma has been repeatedly observed as nonregenerative in skin, peripheral nerves, blood vessels, lung, kidney, and pancreas (1,3,14–18). It has been suggested that the mechanism for the irreversibility of injury (nonregenerative response) is entirely dependent on disruption of the stromal architecture, and that proper replacement of the stromal layer is critical for any regeneration to occur.

Methods to Treat Loss of Organ Function

While a molecular scale defect can often be treated with the use of pharmaceuticals, an organ-scale defect requires more extensive treatments. Significant loss of function in the affected tissue or organ, also referred to as the "missing organ" (19), can lead to a number of significant consequences such as lack of social acceptance in the case of severe burns and facial scars, loss of mobility and sensory function in the case of neuroma, or life-threatening symptoms in the case of a cirrhotic liver, large-scale severe burns, and ischemic heart muscle.

Six basic approaches have been used to treat the problem of the missing organ: transplantation, autografting, implantation of a permanent prosthetic device, use of stem

cells, *in vitro* synthesis of organs, and induced regeneration. The last three techniques have been grouped together and are known by the moniker "tissue engineering" (20). All six techniques will be discussed in the following sections.

Transplantation. Transplantation, the transfer of an organ or a fraction thereof from a donor to a host, is widely utilized as a therapeutic strategy to replace complex tissues and organs. Since the introduction of transplantation in the early twentieth century (21), it has been used in increasingly complex organ systems such as the skin, cornea, kidney, liver, lungs, and heart. Patients have exhibited extraordinary survival characteristics even with the simultaneous transplantation of multiple organs (22–27).

While transplantation of tissues at a few immune-privileged sites such as the eye and testis can occur without rejection, the most significant challenge facing modern transplantation is the immunological barrier for transplantation of tissues from donor to host, where the donor organ is attacked and rejected by the host's immune system upon transplantation; this response is termed host-versus-graft disease where the host's (patient's) immune system recognizes the antigens expressed on the graft tissue as foreign and attacks and destroys the tissue. The primary clinical method for preventing rejection of the transplanted tissue is the use of drugs to suppress the immune system of the host. Immunosuppression therapy often is necessary for the remainder of the host's life to prevent transplant rejection; however, immunosuppression also presents a significant hazard to the host, as the immunosuppressed host becomes vulnerable to infections (28). Significant efforts have been mounted to develop technologies that allow for local rather than whole-body immunosuppression, such as the development of cells that can express a protein (i.e., Fas ligand) to induce immune cell apoptosis and the use of natural and synthetic polymers to encapsulate heterologous cells to mask their antigens (24,28–31).

The major obstacle in using human donors for transplantation has been the difficulty in finding immunocompatible donors and the shortness of supply of suitable organs because the supply is greatly exceeded by the demand (32). A more recent area of research has focused on developing a transgenic pig model to be used as an immunocompatible donor for humans for a procedure known as xenotransplantation. Development of this area of transplantation has been slowed by evidence that pig viruses are capable of infecting human cells and producing unique viral infections in the host (33–35).

Autografting. With autografting, the donor and the recipient are the same individual; a fraction of the tissue or organ is harvested from an uninjured site and grafted at the nonfunctional site (25). Autografting offers a technology that removes the danger of organ or tissue rejection due to a host-versus-graft response, but is relatively limited in its scope of application. Autografting necessitates the creation of a second wound site (donor site), subjecting the patient to a second severe trauma. Therefore, autografting is utilized only when the loss of functionality at the secondary wound site is outweighed by the loss of functionality or morbidity at the primary wound site. This procedure is obviously limited by the availability of functional tissue that can be transplanted without additionally harming the patient. Major clinical uses of autografting have been associated with skin grafting in massively burned patients (36), the use of the sural nerve to bridge a severe peripheral nerve injury, primarily in the case of hand injuries (37–41), and the use of autologous vein graft to bypass a restricted artery (42).

Permanent Prosthetic Device. The implantation of a permanent prosthetic device to replace the functionality of lost or damaged tissues offers a number of advantages and disadvantages. Typical examples of prosthetic devices are artificial hips and knees (43), cardiac pacemakers (44), heart valves (45), stents (46), cochlear implants (47), and contact lenses (48). Prosthetic devices are typically fabricated from biologically inert materials such as metals, ceramics, and synthetic polymers. Hence, these devices do not provoke the immune response problems inherent to transplanted tissues and organs and can also be manufactured on a mass scale. Even though these devices are fabricated from bioinert materials, interactions with the biological environment surrounding the prosthesis lead to a number of unfavorable physical and biological manifestations. Specific examples of negative biomaterial–tissue interactions are the formation of a thick, fibrous scar tissue capsule around a silicone breast implant (49,50), stress-shielding due to the implantation of a relatively stiff (compared to the host bone) hip prosthesis that eventually leads to a loss of bone mass (51), platelet aggregation to implanted surfaces, also known as biofouling (52–55), and the accumulation of polyethylene particles in the lymph system as a result of wear of an orthopedic implant (56,57). The spontaneous remodeling process of the surrounding tissues can also be significantly altered, negatively or positively, by the presence of the prosthetic device (58). The often-serious side effects that appear as a result of interactions between nonbiological materials and the surrounding tissues illustrate the difficulty of replacing bioactive tissues with bioinert materials with drastically different material and mechanical properties.

Stem Cells. Stem cells present an exciting possibility for replacement of lost or damaged organs and tissues. The pluripotential nature of stem cells offers the possibility for the synthesis of tissues from the least differentiated cells in the body (59,60). Current efforts in stem cell research have focused on identifying protocols to harvest stem cells, expand them in culture, and reimplant them at a site of injury, as well as to develop technologies to introduce genes into stem cells so that when reintroduced to the patient, they will synthesize the required proteins *in vivo*. Currently, mesenchymal stem cells (61), epithelial stem cells (62), and neural stem cells (63) have been grown *in vitro* and studied. While few advances in the use of stems cells for replacement of damaged tissues have been made to this point, stem cell technologies present a new area of scientific study for future exploration with a great deal of promise.

***In Vitro* Synthesis.** *In vitro* synthesis requires the growth of a functional replacement tissue using an *in vitro* culture environment to replace a lost or damaged organ or tissue. Traditional *in vitro* synthesis techniques have utilized both cell culture systems and culture systems based on interactions between cells and an extracellular matrix analogue. *In vitro* synthesis allows for total control over the culture environment, specifically the inclusion or exclusion of soluble regulators (i.e., growth factors, cytokines), insoluble regulators (i.e., extracellular matrix proteins), and a variety of cell culture media and conditions; the complexity of biological systems, specifically their cytokine and growth factor needs, and the necessity for developing an efficient method for providing nutrients and cell–cell signals to the site of the developing tissue have to date precluded the formation of complex tissues *in vitro* (1,64).

Early successes with *in vitro* synthesis were encountered using cultured epithelial cells to produce a physiological epidermis (65). In these studies, keratinocyte sheets were grown *in vitro* from skin explants; these keratinocyte sheets were implanted into skin wounds, and induced the formation of a fully mature, stratified epidermis, the uppermost tissue layer that makes up skin (66,67). Later study found that keratinocyte sheets could be grown *in vitro* starting from disaggregated epidermal cells and could then be implanted into a skin wound, inducing the development of a mature, fully stratified epidermis as well (68). This technique has been used to prepare autologous sheets of keratinocytes to treat skin wounds in severely burned patients (69–77). Known as a cultured epithelial autograft (CEA), this technology will be discussed in greater depth later in this article as one of the five major techniques that lead to at least partial regeneration following severe skin injuries. *In vitro* synthesis of more complex tissues and organs began with work aimed at developing an epithelial–mesenchymal bilayer in order to produce a material that could be implanted to replace damaged skin. The fabrication of a "living skin equivalent" (LSE) involved culture of fibroblasts within a collagen gel, followed by introduction of keratinocytes in order to produce an immature skin equivalent; this bilayer was implanted into a skin wound and was observed to lead to the formation of a mature, stratified epidermis as well as an immature neodermis (78–82). This technology will also be covered in greater detail later in this article.

In vitro synthesis of increasingly complex tissues has necessitated the development of technologies for culturing cells in three-dimensional (3D) scaffolds known as ECM analogs and modifying the surface chemistry of these scaffolds to control cell–substrate interactions (20,48,83–86). Synthetic polymer meshes have been used as an ECM analog for culturing keratinocytes and fibroblasts as a potential skin replacement that has been used clinically in the treatment of severe burns and ulcers (87–94). This technology (Living Dermal Replacement, LDR) will be described in greater detail later in the article as one of the major techniques utilized to treat severe skin wounds. *In vitro* synthesis using ECM analogs as a culture environment has also been studied using hepatocytes in an attempt to synthesize a functioning liver (95–97), and chondrocytes in an attempt to synthesize a cartilaginous network (98–101). Continued exploration of *in vitro* techniques to synthesize tissue replacements has been hampered by the complexity of biological systems and an inability to provide the proper nutrient cocktails (i.e., cytokines, growth factors), the structures necessary to deliver these nutrients (i.e., arteries, veins, and capillary systems), and the correct mechanical environments (ECM analog structures) for complex tissue and organ growth outside of the organism.

Induced Organ Synthesis *In Vivo* (Regeneration). Induced organ synthesis *in vivo* relies on the healing processes that are inherently active in a wound site to regenerate lost or damaged tissue. In this method, an analog of the ECM is implanted in the wound site and combined with the biological processes in the wound induce synthesis of a physiologic replacement tissue. Induced organ synthesis was first identified following the development of fabrication techniques that allowed for synthesis of ECM analogs with a well-defined macromolecular structure, specifically controlling the ECM specific surface, chemical composition, and degradation rate (102–106).

The first application of induced organ synthesis was in the fabrication of an ECM analog able to induce skin regeneration, the dermal regeneration template (DRT). The DRT showed very high biological activity when implanted into a full-thickness skin wound and was capable of inducing regeneration of the underlying dermal layer of skin as well as the epidermal and basement membrane layers (3,103,107–113). The speed of this regeneration process was significantly increased when the DRT was further seeded with autologous keratinocytes prior to implantation (3,111,112,114–116). This technology will be covered in greater detail later in the article as one of the major tissue engineering techniques utilized to treat severe skin injuries.

Induction of organ regeneration, first observed in the study of skin regeneration, has also been observed in other tissues and organs with the use of other specialized ECM analogs. Regeneration of peripheral nerves has been achieved using a tubular device that incorporates an ECM analog as a filling, known as the nerve regeneration template (NRT) (1,117–120). The long-term morphological structure and electrophysiological function of nerves regenerated using the NRT has been observed to be at the level of an autografted nerve, the current gold-standard for peripheral nerve injury treatment (121,122). In addition, the DRT was implanted into a conjunctiva stromal wound model, where induced regeneration of the conjunctival tissue structure was also observed (123).

In vitro synthesis and induced organ regeneration (*in vivo* synthesis) currently constitute an area of study termed tissue engineering (20). For the remainder of this article, we will focus on studies on the structure and function of ECM analogs for use in tissue engineering and on an overview of the tissue engineering approaches that have been utilized to treat skin wounds.

Basic Parameters of the Living Environment During *In Vivo* Synthesis

The process of *in vivo* regeneration of lost or damaged tissue can be modeled as if the entire process was taking

place within a "bioreactor" that is surrounded by a reservoir with constant properties, representing the entire organism with its complex homeostatic mechanisms. The "bioreactor" itself has a defined anatomical and physicochemical environment (environment of the wound site); parameters include the temperature, the pH, the structure of a template within the "bioreactor", and the flow rate and composition of the exudate. The flow of exudate entering the "bioreactor" starts almost immediately following the creation of the wound site, while the structure in the wound bed is provided by implantation of an analog of the ECM. The ECM analog, and any cells that may be seeded within, constitutes the exogenous stimulus provided to the wound bed. During the remainder of the healing process, the reservoir (surrounding organism) is considered to maintain the "bioreactor" environment: temperature, pH, as well as cytokines, growth factors, and cells present in the exudate. While these factors are considered standardized in a wound bed, it is the ECM analog that provides the variable structure and ECM components (proteins) that are critical for inducing regeneration. If no active ECM analog is present or if its structural features are changed, the reaction sequence within the "bioreactor" is observed to follow almost completely normal repair processes that result in wound contraction and scar formation. In contrast, when the appropriate ECM analog is introduced into the "bioreactor" normal repair mechanisms are replaced by induced regeneration, also referred to as *in vivo* or *in situ* regeneration (1).

Experimental study of *in vivo* regeneration is complicated by a lack of reproducibility between different reaction sites (anatomical sites); unless the wound site is standardized, it will be impossible to accurately study differences between ECM analogs when implanted into wound models and results obtained in independent laboratories will not be statistically significant. Billingham and Medawar (124,125) introduced the concept of an anatomically constant wound for the study of massive skin injuries. For skin injury models, the entire thickness of the skin (epidermis, basement membrane, and dermis) was consistently excised down to the layer of the underlying muscle and fascia. Except for edge effects, this model standardized the wound environment from one animal to another, making it possible to obtain statistically significant results from a study with several animals. For the remainder of this article, all animal results that are discussed will be from this wound model. In clinical cases, the nature of the wounds is different; such issues will be discussed in greater detail later in this article.

Mammals possess a small, quite limited, ability to spontaneously regenerate damage inflicted to most of their tissues and organ systems. The epidermis and basement membrane in the skin regenerates spontaneously, provided that there remains an intact, underlying dermal layer (1). In another example, a small, cylindrical defect (<1 cm in diameter) in mammalian long bones is spontaneously refilled with normal bone (126), and a gap <5 mm in a transected rat sciatic nerve can be spontaneously regenerated to a moderate extent (127–131). In these cases, regeneration is observed without the implantation of any active ECM analog. More specifically in the case of skin injuries, the epidermis and basement membrane have been observed to spontaneously regenerate to form fully mature tissue, but the adult mammal is unable to spontaneously regenerate the dermis (1,124,125,132). In studies of induced regeneration using grafting techniques, it is important to first understand the fundamentals of behavior of the tissues involved, specifically the type of tissue that can be spontaneously regenerated, in order to understand the effectiveness of the graft. Regeneration of tissue that has resulted from the exogenous graft will be referred to as induced regeneration. In the case of studies of skin regeneration, the presence of new dermis following grafting of a full thickness skin wound with an appropriate device will be considered induced regeneration.

CHARACTERISTICS OF EXTRACELLULAR MATRIX ANALOGS THAT DEFINE BIOACTIVITY

Fundamental Design Principles for Tissue Regeneration Scaffolds

Porous scaffolds are utilized in the study of tissue regeneration; the term active extracellular matrix analogs (bioactive analogs) refers to scaffolds that induce regeneration of normally nonregenerative tissues following severe injury. Bioactivity is observed in only a limited number of scaffold variants, and is measured by the scaffold's ability to prevent irreversible repair behavior while inducing regeneration. With a bioactive scaffold, the cells, cytokines, and biological exudate in the wound site interact with the scaffold such that the mechanisms and kinetics normally associated with spontaneous wound closure by wound contraction and scar synthesis (repair) are modified leading to the induced regeneration response. Bioactivity of tissue regeneration scaffolds has been observed to depend on the structural characteristics of the scaffolds, notably the chemical composition, the template pore structure, and the template residence time (1). These characteristics, and any governing models to describe cellular behavior in regeneration scaffolds, will be discussed in the following sections.

Template Residence Time

The residence time of an implanted scaffold is a critical variable that helps to define the bioactivity of the scaffold. For physiologic tissue to be synthesized in the wound bed, the scaffold must degrade in such a way that it does not interfere with the production of physiologic tissue. Empirical evidence supports a requirement for the implanted active ECM analog to be capable of isomorphous tissue replacement, that is, degradation of the active ECM analog at a rate of the same order as the rate of synthesis of new tissue (1).

These considerations are consistent with a model that defines a scaffold residence time with both an upper and a lower bound. Using the isomorphous tissue replacement model, the appropriate time period for scaffold residence is approximately equal to the time period required to synthesize a mature tissue at the specific site by regeneration. A reasonable approximation of the time for regeneration is the time period observed for the conventional healing

process of a wound that involves wound contraction and scar formation at the anatomic site of interest. In the case of a full-thickness skin wound, the healing time is ~25 days (1,3,133). As the intact scaffold cannot diffuse away from the wound bed, the simplest method for achieving isomorphous tissue replacement requires the macromolecular scaffold structure to be degraded by enzymes in the wound bed into low molecular weight fragments that are able to diffuse away. In a model of template degradation characteristics, the lifetime of the scaffold in the wound bed can be defined by the time constant for degradation (t_d) and can be compared to the time constant for a normal healing process of a wound at the anatomic site of interest (t_h). For isomorphic tissue replacement:

$$\frac{t_d}{t_h} = O(1)$$

The isomorphous tissue replacement hypothesis has been supported by observations made with different implanted devices that were studied in the context of *in vivo* synthetic processes. When the ratio of t_d/t_h was much >1, the scaffold remained in the wound bed virtually as a nondegradable implant, and dense fibrotic tissue similar to scar was observed to be synthesized underneath the scaffold (1,133). As the ratio became considerably <1, the initially insoluble scaffold became rapidly degraded and did not induce a regenerative healing process in the wound bed; instead, wound bed healing was marked by scar synthesis similar to the healing process observed in an ungrafted wound (109).

Satisfying the principle of isomorphous tissue replacement requires adjustment of the structure of the ECM analog such that the biodegradation time constant of the scaffold (t_d) closely matches the healing time constant (t_h). In the case of collagen-based ECM analogs, degradation of the scaffold in the wound bed is accomplished primarily by collagenases present in the wound site. Reduction of the biodegradation rate of collagen scaffolds has been achieved both by introducing glycosaminoglycans (GAGs) into the collagen mixture and by cross-linking collagen fibers to one another: scaffold resistance to degradation is increased with an increase in the density of cross-links (103,105,108,117,134). For collagen–GAG scaffolds, precipitating the two polymers together under an acidic pH mixes the glycosaminoglycan and collagen components, and the scaffold is fabricated from the coprecipitate. Cross-linking of the collagen fibers can be accomplished by a variety of different techniques, both physical and chemical.

Dehydrothermal (DHT) cross-linking is a physical technique that has been often utilized for cross-linking collagen and collagen–copolymer scaffolds, and allows for the creation of a variety of cross-linking densities. Dehydrothermal cross-linking involves exposure of the scaffold to a high temperature (90–120 °C) under vacuum (i.e., 95 kPa), leading to removal of water from the scaffold. When the water content in the ECM analog is removed below ~1%, drastic dehydration of the scaffold leads to the formation of interchain amide bonds through condensation (135). This cross-linking reaction is a condensation reaction involving carboxylic groups from glutamyl/aspartyl residues on collagen polypeptide chain P_1 and the amino groups of lysyl residues on an adjacent P_2 chain to yield covalently bonded collagen fibers. In addition, graft copolymers of collagen and GAG are formed by dehydration, leading to cross-link formation by condensation of amine groups on collagen chains (denoted P_1 and P_2 below) with carboxylic groups of glucuronic acid residues on GAG chains (104).

$$\text{Collagen–Collagen} \quad P_1 - COOH + P_2 - NH_2 \rightarrow P_2 - NHCO - P_1 + H_2O$$

$$\text{Collagen–GAG} \quad GAG - COOH + P_2 - NH_2 \rightarrow P_2 - NHCO - GAG + H_2O$$

The density of cross-links formed through DHT cross-linking depends on the temperature as well as the length of exposure, with higher temperatures and longer exposure lengths producing a higher density of cross-links (104,105). Ultraviolet (UV) treatment is a second physical cross-linking technique that can create cross-links between collagen fibers (136–138). Additionally, cross-linking can be induced chemically by introduction of glutaraldehyde (GT) (104) or with 1-ethyl 3-(3dimenthylaminopropyl)carbodiimide (EDAC) (139). These chemical cross-linking techniques are considerably more powerful than the previously noted physical methods and lead to a much higher cross-link density and a much longer time constant of degradation (t_d). Chemically cross-linked scaffolds must be extensively washed to remove all trace of the cytotoxic cross-linking chemicals prior to use in the wound bed; in addition to concerns regarding washing away the excess chemical, some chemical cross-linkers function by having all or a portion of the chemical compound becoming part of the cross-link. In this case, slow degradation of the scaffold could result in the release of potentially cytotoxic agents into the wound site (1,3).

By using both physical and chemical methods, the cross-linking density of a particular collagen or collagen–GAG scaffold can be effectively adjusted to create a wide range of enzymatic degradation rates. These cross-linking tools can also be applied to a multitude of other scaffolds, fabricated from both natural and synthetic materials. As different wound sites and the same wound sites in different species have been observed to exhibit very different time constants for healing (t_h), it is necessary to adjust the degradation rate of the ECM analog to the characteristics of the specific wound bed site and species in order to satisfy the principle of isomorphic tissue replacement and to induce regeneration.

Critical Cell Path Length

Migration of cells into the active ECM analog (regeneration template) is critical for the synthesis of new tissue. The use of porous templates allows for more rapid incorporation of cells into the template. While the effect of the structural characteristics of the template will be discussed in the following sections, there is another important characteristic to consider: an adequate source of metabolites (i.e., oxygen, nutrients) available to the cells. There are two mechanisms available for transport of metabolites to and

waste products from the migrating cells: diffusion to and from the surrounding wound bed or transport along capillaries that have sprouted into the scaffold as a result of angiogenesis. While angiogenesis becomes the limiting factor for long-term cell survival and growth, significant angiogenesis is not observed for the first few days after implantation of the template. Therefore, early cell survival inside the scaffold is completely controlled by the diffusional mode of transport of metabolites from the wound bed.

A simplified model of cellular metabolic requirements and nutrient diffusion characteristics can be utilized to describe the critical cell path length (L_c) for cellular migration into a scaffold; the model assumes a totally diffusion-based mode of metabolite transport. Beyond the critical cell path length, diffusional transport will not provide for cellular survival; this model is important for predicting the initial "take" of a graft, as diffusional transport, along with prevention of graft rejection or infection, will be responsible for initial cell survival and graft "take". Such a model simplifies the cell's metabolic requirements by defining a "critical" nutrient that is metabolized by the cell at a rate of R mol/cm^3/s. This nutrient is transported to the scaffold from the wound bed, where there is a constant concentration of nutrient, C_0, that remains essentially constant due to resupply from the organism that acts as an infinite source. The transport processes are modeled by the diffusivity of the nutrient through the scaffold, D, with units of square centimeters per second (cm^2/s) and a length of diffusion, L, with units of centimeters. The *cell lifeline number*, S, is a dimensionless number that is derived from these variables by dimensional analysis to express the relative importance of the consumption of an essential nutrient by the cell and the diffusion of the nutrient from the wound bed into the scaffold:

$$S = \frac{RL^2}{DC_0}$$

The magnitude of the cell lifeline number describes three distinct cases of cellular migration into the scaffold in the absence of angiogenesis. For $S \gg 1$, the rate of consumption of the critical nutrient exceeds the supply provided through diffusion processes, resulting in the death of cells that have migrated that distance L into the scaffold. For $S \ll 1$, the supply of the critical nutrient by diffusion exceeds the rate of consumption, resulting in cell proliferation and migration further into the scaffold. The steady-state condition, $S = O(1)$, describes the state where the value of L becomes the critical cell path length, L_c. L_c describes the longest distance from the wound bed edge that a cell can migrate while depending solely on diffusional processes as the source of metabolites before angiogenesis begins. In the case of many common nutrients with a low molecular weight, L_c can be predicted using this model on the order of 100 μm (109). This estimation of L_c indicates that there needs to be close contact between the wound bed and the scaffold immediately after grafting in order for the initial migration of cells into the wound site to take place prior to angiogenesis and suggests that very stiff ECM analog grafts will show inferior results. A stiffer scaffold will be unable to conform to the wound bed, thus reducing the effective distance that cells can migrate into the scaffold and be metabolically supported prior to angiogenesis, eventually reducing overall graft "take". Unfavorable values of the scaffold-tissue surface tension can also prevent close contact.

Chemical Composition

The chemical composition of an ECM analog plays a critical role in defining the bioactivity of the device. In order to design a suitable device for use in wound healing that will induce regeneration rather than repair mechanisms, it is important for the device to keep the wound edges apart and to drastically modify the healing processes in the wound bed to yield physiologic tissue rather than scar. At the outset, the chemical composition defines the ligands that are displayed on the scaffold surface. Cellular activities such as binding, migration, and contraction are all mediated by interactions between the integrins expressed by a specific cell type and the ligands available on the scaffold surface. The design of a scaffold to be used for tissue engineering purposes must accordingly be fabricated in such a manner and using such specific materials as to allow for appropriate binding between cell and scaffold. The specific cell–matrix binding that appears to be required in cases of induced regeneration is that which blocks contractile cells from implementing their program of wound contraction (1).

A number of different materials have been used for the production of scaffolds for tissue engineering. Several synthetic nondegradable polymers, such as poly(dimethyl siloxane), have been utilized; these polymers, which parenthetically violate the principle of isomorphous tissue replacement, have not exhibited the ability to induce regeneration. Degradable synthetic polymers, such as poly(lactic acid), can be fabricated to satisfy isomorphous tissue replacement and the surfaces can be modified to properly induce cellular binding, but have not been shown to prevent contraction and scar formation. A chemical composition that has been used successfully to induce regeneration has been a graft copolymer of type I collagen and a sulfated GAG (1). These natural polymers are capable of facilitating cell binding in part due to their expression of natural ligand-binding sites.

Template Pore Structure

Although the chemical composition is a critical component in defining the biological activity of the scaffold, it is not the sole characteristic. The biological activity of a particular ECM analog also depends significantly on the architecture of the three-dimensional (3D) network. Having migrated into the scaffold, the cell interacts with the structure that defines the porous scaffold, making use of its cell surface receptors (integrins) to bind to specific ligands on the scaffold surface. The first critical components of the scaffold pore structure to consider is the open- or closed-cell nature of the scaffold and it's relative density: A tissue engineering scaffold must possess an open-cell pore structure with a relative density below a critical value. Three-dimensional porous structures can possess an open- or a closed-cell pore structure; in a closed-cell structure, each

individual pore is separated from adjacent pores by membrane like faces while open-cell pore structures exhibit interconnectivity between adjacent pores. Interconnectivity is critical for cells to be able to migrate through the 3D structure. The most important structural feature of porous scaffolds is the relative density (R_d): the density of the scaffold divided by the density of the solid from which it is made. The porosity of a scaffold, or pore volume fraction, is defined as $(1 - R_d)$. The relative density defines the amount of solid material available for cells to bind to; when the pores are closed or when the relative density is too large, cells are not able to migrate through the scaffold, a significant impediment for using such a scaffold for a tissue engineering application. Such structural aspects also suggest that in designing scaffolds for tissue engineering applications, there is a critical number of cells required for scaffolds to appear bioactive. There must be a large enough area available for cells to bind to in order to support a large enough population of cells within the scaffold; the existence of a critical density of cells has been hypothesized as a result of a number of experiments studying cell–scaffold interactions (1).

The development of a highly detailed model, describing the number of receptors utilized per bound cell and the nature of the binding and receptor sites, is required to describe even a simple interaction between a cell and a generic scaffold surface. However, a more generic explanation can be used to indicate both the complexity of the cell–scaffold interaction and the significant influence the scaffold pore structure has on scaffold bioactivity. In particular, we will examine the affect of another critical factor on scaffold bioactivity: mean scaffold pore size.

The structure of a porous scaffold is defined by the pore volume fraction, mean pore size, and pore orientation in the scaffold. All of these characteristics have been shown to significantly affect the bioactivity of the scaffold. The pore volume fraction and the mean pore size together define the specific surface area of the scaffold, the total surface area of pore walls available for cellular binding. Increasing the mean pore size while keeping the pore volume fraction constant decreases the specific surface area of a scaffold. Decreasing the pore volume fraction and keeping the mean pore size constant increases the specific surface area (140). It has been estimated that a 30-fold increase in pore diameter leads to a 27-fold decrease in specific surface (2). A change in the specific surface area of the scaffold significantly changes the area available for cells to bind to. More specifically, the surface density of bound cells (Φ) in a 3D dimensional porous scaffold is a function of both the density of bound cells in the scaffold (ρ) and the specific surface of the scaffold (σ):

$$\Phi = \frac{\rho}{\sigma}$$

This calculation suggests the significance of the specific surface of the scaffold in defining the scaffold bioactivity. If the specific surface is too small due to large pores, an insufficient number of cells will be able to bind to the scaffold and the cells that remain free will contribute to the spontaneous repair mechanism. There is a minimum pore size as well, defined by the characteristic dimension of the cell, $\sim$10–50 μm for most cell types. When the scaffold pore size is smaller than this critical dimension, cells will be unable to migrate through the porous structure, and will be unable to infiltrate and bind to the template. These upper and lower bounds of the scaffold pore diameter, mediated by cell size and specific surface requirements, have been determined experimentally for each cell type for tissues where regenerative templates have been used (3,141); however, future work is necessary to develop a better understanding of cell adhesion and it's relation of scaffold structure.

The shape of the pores that make up the porous scaffold must also be considered; slight changes in the mean shape of the pores can result in significant variations in the mechanical properties of the scaffold (140). In addition, changes in mean pore shape may also play a role in defining the areas of the scaffold available or unavailable for binding and in defining available directions for cell migration. The template pore structure plays a very significant role in defining the overall bioactivity of the scaffold and the open- or closed-cell nature, the mean pore size, relative density, and pore shape and orientation all are critical components to consider.

TISSUE ENGINEERING OF THE SKIN

Structure and Function of Skin

Mammalian skin is a stratified tissue made up of three distinct layers of tissue: the epidermis, the basement membrane, and the dermis. Each tissue displays unique structural and functional properties as well as distinctive responses to injury or damage. These tissues, an epithelial layer (epidermis in skin), a basement membrane layer, and a stromal layer (dermis in skin) make up the previously described tissue triad for the skin (1,8,142). Using the tissue triad to model a generic tissue, the epithelia covers all body surfaces, tubes, and cavities, and is separated from the underlying stroma by a continuous basement membrane layer. As the basement membrane is totally acellular and is not penetrated by the vascular system, the survival of the epithelia depends on diffusion of metabolites, nutrients, oxygen, and waste products across the basement membrane to and from the stroma. The stromal layer contains the vascular system and other supporting tissues (connective tissues) that serve to nourish and anchor the basement membrane and the epithelia. In addition, while both the stromal and the basement membrane layers contain an extracellular matrix structure, the epithelial layer does not. In the following sections, the structure and function as well as the response to injury of each of the layers of the tissue triad that constitutes adult skin will be described in detail.

Morphology and Function of the Epidermis

The epidermis is the exterior layer of tissue that makes up the skin. The epidermal layers act as a physical barrier to protect the organism against microorganisms; prevents organismic dehydration; and protects the organism from mechanical, thermal, chemical, and UV insults. It is a

stratified tissue, consisting of five distinct tissue layers (strata) that form a tissue ~0.1 mm thick. These layers represent a cell–maturation gradient along which cells move, from the interior to the most exterior layers; cells become increasingly mature and keratinized during this migration process. The most interior layer is the basal cell layer, known as the *stratum malpighii* or *stratum germinativum*. The next layer is the prickle cell layer, also known as the *stratum spinosum*, describing the prickly morphology of the cells. The granular layer, also known as the *stratum granulosum*, is the next layer consisting of keratohyalin granules (intracellular granules) that contribute to the process of keratinization. The *stratum lucidum* is the fourth epidermal layer, found only in the very thick skin associated with the fingertips, palms, and soles of the feet. The most exterior layer is the cornified or horny layer, the *stratum corneum*, made up of flattened cell remnants that are fused together forming a compact layer of keratin, the fibrous protein that makes up the external armor of the epidermis (142). The cell–maturation gradient observed in the epidermis starts with immature cells (keratinocytes) in the basal layer; these cells undergo mitosis and migrate through the cell layers toward the cornified layer over a period of 25–50 days. Along this path, the cells become increasingly keratinized, until they reach the cornified layer where the dead cells are desquamated. This stratified tissue is avascular, relying on the underlying dermis for a nutrient supply.

The five cell layers and the individual keratinocytes within each layer are bound together by desmosomes; keratin filaments, a meshwork of filaments inside the keratinocyte cytoplasm (also known as tonofilaments), anchor neighboring cells to one another. Additional mechanical stability is provided to the epidermis by its attachment to the underlying dermis; bonding at the epidermal–dermal junction is mediated by the basement membrane. Hemidesmosomes, located inside the cell membrane of basal cells, attach to the epidermis and to the basement membrane by means of tonofilaments via junctions on the subbasal plates (142). This construct forms a stratified, mechanically stable keratinized epithelium able to withstand the thermal, mechanical, chemical, and UV insults to which the body is continuously exposed.

Morphology of the Basement Membrane

The basement membrane, also known in the literature as the basal lamina, is found in many different tissues as an acellular, avascular layer between the avascular, cellular epidermis (no ECM) and the cellular, vascularized dermis (developed ECM). The basement membrane performs a number of significant duties; notably, it provides a secure and adhesive layer to facilitate a strong connection between the epithelia and stroma, serves as a boundary that can regulate cell and molecular movement, provides a scaffold to facilitate repair following injury, and facilitates differentiation and growth of the epithelial and stromal layers (9,143–146).

The basement membrane is an acellular, avascular stratified tissue made up of three distinct strata that are, in total, ~100 nm in thickness. The basement membrane structure is often observed under light microscopy as having only a single layer, termed the *lamina densa*; electron microscopy of the basement membrane reveals the *lamina lucida* and the *fibroreticularis* (142). The first layer of the basement membrane, sitting adjacent to the basal cell layer of the epidermis, is termed the *lamina lucida*, an electron-lucent membrane ~20–40 nm in thickness that consists primarily of the glycoprotein laminin. The middle layer of the basement membrane, termed the *lamina densa*, consists predominantly of type IV collagen, is ~40–50 nm in thickness, is significantly more electron-dense than the *lamina lucida*, and is the region visible under light microscopy. The final layer of the basement membrane is located adjacent to the underlying dermis and is termed the reticular layer, also known as *fibroreticularis*. This electron-lucent layer is composed primarily of type VII collagen fibers and is responsible for fixing the basement membrane to the underlying dermis by anchoring fibrils attached to specific anchoring plaques that are embedded in the underlying dermis (147,148). While the primary components of the basement membrane are type IV and type VII collagen, the basement membrane also contains significant amounts of chondroitin sulfate, heparin sulfate, fibronectin, tenascin, nidogen, enactin, thrombospondin, and 1-microglobulin (149). When viewed from a more macroscopic scale, the topography of the basement membrane surface appears as an undulating line between the dermis and the epidermis, significantly increasing the surface area between these two structures. This undulating structural feature, termed rete ridges, will be discussed in greater depth in the next section describing the dermis.

Morphology and Function of the Dermis

The dermis is the final component of the skin tissue triad, lying below the basement membrane and above the underlying muscle and fascia. The dermis, considered anatomically to be a single layer, actually consists of two zones: the papillary dermis and the reticular dermis. The papillary dermis is the upper zone adjacent to the basement membrane and consists primarily of loosely packed collagen fibers. The papillary dermis forms the upward projections of the dermis into the epidermis that define the rete ridges; these projections are filled with capillary loops responsible for providing metabolites, nutrients, and oxygen to the epidermis. In addition, the papillary dermis contains fine axonal connections of unmyelinated sensory nerves that extend up to the basement membrane. The bulk of the dermis, termed the reticular zone, is found below the papillary dermis. The reticular zone is comprised of thicker and more closely packed collagen fibers as well as a significant content of elastin fibers that are interlaced with the collagen fibers to form an isotropic, collagen–elastin network. While collagen fibers are highly crystalline microfibers that have a limited stretching ability and provide the strength to a tissue, elastin fibers are considerably thinner and amorphous (noncrystalline), providing the ductile strength (stretching without yielding) of a tissue (1). While the strength of the dermis is defined by the collagen content and the ability of the dermis to bend and deform without

permanent damage is defined by the elastin content, it is the combination of collagen and elastin fibers that is responsible for the robust nature of the dermis.

The dermis has two major functional roles: providing mechanical stabilization for and metabolic support to the epidermis. The combination of mechanical strength and deformability gives the dermis the ability to provide a stable base for the epidermis that is able to withstand the substantial shear, tensile, and compressive forces associated with ordinary activities that would cause an unsupported epidermis to fail. In addition, the undulatory nature of the dermo–epidermal junction allows for the intimate presence of an extensive dermal vascular system that provides metabolic support (providing nutrients and oxygen while removing waste products) to the avascular epidermis. The rete ridges also provide increased surface area for attachment of the epidermis to the basement membrane and the basement membrane to the dermis, increasing the strength of the dermo–epidermal connection and enhancing the surface area available for the capillary loops to provide metabolic support to the epidermis. In addition to the two zones that make up the dermal layer, hair follicles, sweat glands, and oil-secreting glands originate in the dermis and extend through the basement membrane and epidermis to the skin surface. The dermis also provides tactile sensation through the unmyelinated sensory nerves that extend through the dermis up to the dermo–basement membrane junction, and allows for thermoregulatory control (1,142).

Current Treatment of Massive Skin Loss

Traditional treatments of a severe skin wound have focused on developing a temporary technique or product that serves to close the wound, preventing infection and dehydration (important for large skin wounds) (1). Historically, attempts have been made to treat severe wounds and burns dating back almost 3500 years, and have included a wide variety of temporary devices such as membranes of organic and synthetic polymers, skin grafts from animals (heterografts or xenografts), skin grafts from human cadavers (homografts or allografts), and skin grafts from the patient (autografts) (133). Allografts are used as a temporary covering for excised (cleaned) wounds prior to autograft, where the allograft is removed and the permanent autograft is placed into the wound. Xenografts are typically taken from the pig due to the great affinity between human and pig skin. Like the allograft, xenograft skin is a temporary wound dressing used until autograft. Temporary dressings immediately reestablish the skin barrier, decrease inflammation and risk of infection, decrease fluid loss, and reduce patient mortality. More recently, the need to develop technologies suitable for treating severe skin injuries over large areas has increased, increasing the requirement for a material to rapidly close a severe skin wound. The fundamental reasons for such a change are that an increasing percentage of patients survive the acute phase of the injury due to improved medical care and that the widespread use of early escharectomy (complete debridement of the wound immediately after injury) requires immediate coverage of large wound areas. It is often not possible to harvest enough autograft tissue to cover a large wound. In these cases, the graft is perforated and then stretched to cover much more space than the original tissue; this meshing process decreases the quality of regeneration, but increases the area that can be treated and will be discussed in greater detail later in this section. There are also a number of problems related to the use of temporary dressings such as allografts and xenografts. Especially in cases of severe burns over large areas of the body, the xenograft or allograft may need to remain in place for a significant period of time until autografting or other treatments are possible. In transplantation of donor tissue, histocompatibility becomes an issue; typically, xenografts and allografts are rapidly rejected, usually within a month of transplantation. This phenomenon illustrates the concept of host-versus-graft disease, where the patient's (host's) body mounts a host defense that ultimately destroys the implanted tissue (graft). Histocompatibility antigens expressed in the transplanted tissue are identified as foreign by the patient's immune system, leading to an inflammatory and immune response that destroys the grafted skin; such problems are also seen in transplantation of almost all other organs and tissue in the body (i.e., liver, kidney, heart, lungs, bone marrow). Patients with allografts or xenografts can be immunosuppressed using a variety of drugs to prevent host-versus-graft disease and prolong the viability of the transplanted tissue (150); however immunosuppression introduces a variety of complications such as a decreased ability to fight infection, a prime concern for people with severe skin injuries.

Polymeric membranes used as temporary dressings often lack biological activity due to the chemical composition and structure of the membrane; these membranes often have to be removed after only a few days due to lack of formation of physiological structures and incidence of infection (133). Often a temporary graft, such as a synthetic or organic polymeric membrane, xenograft, or allograft, is useful in early management of a wound while an autograft site is prepared. Permanent treatments for massive skin loss have traditionally been focused on the autograft technique.

Despite the presence of other grafting techniques, the autograft is the current clinical standard; it addresses both the urgent need to cover an exposed wound and results in an adequate long-term result. Under ideal conditions and in the case of small wounds, when full-thickness skin wounds are treated with an autograft, an almost fully functional skin has been observed to regenerate. However, the skin replaced via the autografts has been observed to lack hair follicles and other adnexa. Despite these missing components, this skin replacement remains functional for the remainder of the patient's lifetime. Two major problems complicate the use of autografts to treat full-thickness skin wounds: the creation of a second wound (donor site), and the requirement for large autografts in the case of massive skin injury. The removal of the autograft results in a secondary full-thickness skin wound that eventually becomes reepithelialized, but considerable scar formation and contraction are observed at the donor site. This factor coupled with the usual need for large amounts of autograft

tissue due to the typical size of severe skin defects in humans has resulted in the surgical meshing procedure, where a small amount of autograft tissue is harvested, then passed through a device to cut a pattern of slits in the autograft tissue; this tissue is then stretched, greatly increasing the area of coverage and thus decreasing the amount of harvested tissue needed to cover the wound. This technique is not without problems as scar synthesis is observed in the areas of the wound not covered by the stretched autograft mesh, resulting in a pattern of scar that greatly reduces the value of the autografting procedure. Due to clinical attempts to minimize the size of the autograft wound, meshing is used almost exclusively to treat skin wounds. This results in adequate coverage and closure of the wound, but the coverage is marked by considerable scar synthesis and contraction, reducing the aesthetic and functional value of the treatment.

It is this inability to utilize the autograft without considerable scar formation as well as the requirement for the creation of a secondary wound site that have provided the stimulus to investigate alternative dressing options that could potentially lead to regeneration of physiologic tissue rather than healing by contraction and scar formation. The resulting technologies will now be discussed in detail in the following sections.

TECHNOLOGIES FOR REPLACEMENT OF THE SKIN

A number of technologies have been developed in an attempt to induce regeneration of skin following injury, both in conjunction with or without the use of an autograft. These technologies have met with differing levels of success. There are five main technologies for treating massive skin wounds by grafting that will be discussed in this article: sheets of epidermis cultured *in vitro* (Cultured Epithelial Autograft, CEA), cell-seeded nylon scaffolds (Living Dermal Replacement, LDR), a 3D living bilayer first cultured *in vitro* with dermal and epidermal cells (Living Skin Equivalent, LSE), a collagen scaffold that was either seeded with keratinocytes or implanted as an acellular construct (DRT), and a naturally derived collagen matrix (NDCM, Alloderm). These distinct procedures will now be discussed in detail, describing the design and manufacture of each device as well as the attendant experimental and clinical results.

Cultured Epithelial Autograft

Cultured epithelial autografts have been studied in both experimental and clinical settings as a possible treatment for massive skin injuries. This technique uses an epidermal graft that is grown *in vitro* and then implanted into the wound site to cover the skin defect. While initially used to provide immediate coverage of the wound site to prevent excessive fluid loss and infection, the CEA has also been studied in models to assess its potential as a permanent skin replacement graft. In the literature, the CEA technology has also been referred to as a keratinocyte sheet, cultured epithelia sheets, or cultured autologous keratinocyte sheets. This technique to culture keratinocyte cells to form and epithelial sheet was utilized because keratinocytes make up ~90% of mammalian epidermal cells (1). There are three major sources for keratinocytes that have been used in cultivating keratinocytes sheets: keratinocytes from disassociated cells, keratinocytes from epidermal explants, and suspensions of pellets of disaggregated keratinocytes (65).

The CEA technology relies on culturing keratinocytes isolated from the patient to produce a graft of autologous tissue, removing any immune complications observed in the case of xenografts or allografts; additionally, since the epidermis can spontaneously regenerate, epidermal tissue harvested from the patient will regenerate without further scarring. Therefore, development of a successful CEA relies upon the development of *in vitro* methods for rapid, serial cultivation of human keratinocytes from a disaggregated cell suspension; these techniques allow for expansion of the (small) harvested cell population by >10,000-fold in 3–4 weeks, a rate necessary to culture the volume of cells required to rapidly produce a keratinocyte sheet large enough to cover a wound site in a clinically acceptable time period (76,77,151,152). In a clinical setting, keratinocytes are typically isolated from skin biopsies; the biopsy tissue is then treated enzymatically to allow removal of the dermal tissue and to dissociate the remaining epidermis. This sequential process prevents contamination of the keratinocyte cell line with dermal cells (mainly fibroblasts). The keratinocytes can then be cultured using a defined *in vitro* process (73,76,77). All of these techniques can be utilized to culture the requisite cell expansion.

Using these established cell culture techniques, an intact, coherent sheet of stratified epithelium can be produced *in vitro* that is on the order of four to six cell layers thick and is bound together by the desmosomes seen in the normal epidermis. Similar to normal epidermis, sparse keratin fibers are observed running parallel to the long axis of the flattened keratinocyte cells in this new epithelial layer. While keratinization is not always observed, the maturity of this newly formed neoepidermis is moderately high. The epithelial maturity can be acutely affected by the identity of the substrate on which the CEA sheet is grown. When the stratified neoepithelium is grown on collagen gels, hemidesmosomes are not synthesized; however, the use of a surface formed from reconstituted basement membrane led to synthesis of hemidesmosomes and a more mature epidermis (1,148,153).

Studies of CEA development have indicated that a partially mature epidermis can be synthesized *in vitro* starting from disaggregated keratinocytes. There is no requirement for the presence of any dermal component or for fibroblasts in the synthesis of a neoepidermis. However, there is a temporary requirement for a nondiffusible substrate onto which the cells are grown in order to develop stratified and keratinized cell layers. While contact with specific connective tissue surfaces can induce formation of the mature neoepidermis, it is still possible to develop a less mature epidermal layer with culture on plastic or glass surfaces (1).

In the clinical setting, keratinocytes are harvested from the patient via a biopsy; the cells are then dissociated, cultured, and expanded *in vitro* for ~3 weeks to form a neoepidermis. At the end of this period, the mature,

keratinizing epidermal layer that forms *in vitro* is then implanted directly into the wound site. One major drawback to the use of the CEA is that due to its extreme friability, handling and grafting the CEA into a wound bed require extreme care. Additionally, the grafted site must be kept immobile so that the CEA can remain in place and not break apart. After implantation, the epidermal cells continue to multiply and spread, covering the entire wound. Clinically, the success of the CEA treatment depends significantly on the condition of the skin wound. The adhesion (take) of the CEA was very different depending on whether the CEA was grafted onto a full-thickness or a partial-thickness skin wound. In the case of partial-thickness wounds, the *take* of the graft has been very good and the CEA has been used to cover significant areas as large as half of the total body surface, making the CEA a life-saving, although temporary graft (133). Take was considerably inferior in the case of full-thickness skin wounds, where there was no underlying dermis to support the neoepidermis; in particular, one persistent problem was the formation of blisters under large areas of the graft (avulsion). Regardless of whether the graft was placed upon a partial- or full-thickness skin wound, the resulting CEA graft exhibited mechanical fragility due to a lack of three specific structural features present in normal, adult epidermal and basement membrane that serve to tether these layers onto the underlying dermis: the 7-S domain of type IV collagen, anchoring fibrils, and rete ridges (154). These structures are required for the formation of a physiological dermo–epidermal junction, and CEA grafts have failed to induce formation of these structures, a collagen fiber architecture, or the elastin fiber network that are all observed in the normal, adult dermis (133). Without these structures, the CEA cannot be used as a permanent skin replacement; instead, the CEA is often used as a temporary coverage as part of a larger treatment regimen. The CEA also exhibits high vulnerability to cytotoxins and bacterial proteases, as the CEA does not behave as a full-thickness graft in preventing infection. During the initial period after grafting, the CEA is extremely sensitive to the effects of bacterial or fungal infections of the wound bed: a full-thickness graft such as the autograft can tolerate infections that result in a near or complete loss of the CEA.

Living Dermal Replacement

The LDR was developed to be a more permanent treatment for severe skin injuries. While the CEA relied upon developing a stratified epidermal layer *in vitro* that could be used to permanently treat injuries to the epidermis and basement membrane and to temporarily treat full-thickness skin injuries, the LDR was developed in an attempt to utilize a structure that could permanently treat full-thickness skin injuries. The LDR technology used an acellular scaffold cultured with both fibroblasts and keratinocytes *in vitro* prior to implantation. In addition to introducing a cell population to the wound site, the scaffold structure was included to provide an immediate 3D architecture for both structural and cellular support that could be synthesized and implanted rapidly into the wound site. The acellular LDR scaffold consists of a copolymer of glycolic acid (90 wt%) and lactic acid (10 wt%), termed polyglactin-910 surgical mesh (PGL). The PGL fibers, $\sim$100 μm in thickness, were knitted into a mesh that exhibited a pore structure with a characteristic dimension of 280×400 μm. This mesh structure presented a large-weave structure to the cells, relative to the characteristic cell dimension of $\sim$10 μm, allowing rapid cell incorporation into the mesh as well as diffusion of an ample supply of nutrients to support cell activity. The polyglactin mesh was cultured with fibroblasts until all of the pores in the mesh were covered with cells; the confluent cells were observed to have begun to synthesize several important ECM components *in vitro*. Immediately prior to grafting, the upper surface of the polyglactin scaffold that was confluent with fibroblasts was seeded with keratinocytes in an attempt to form a bilayer graft that would mimic the structure of skin. Once the keratinocytes reached confluence on the surface of the PGL mesh, the entire structure was grafted into the wound site (88,92).

The LDR was studied as a stand-alone graft to be used temporarily prior to eventual, permanent closure by an autograft, primarily in a mouse model where the device was grafted into full-thickness skin wounds (1,133). A thin, fragile epidermal layer developed initially by 10 days postgrafting and it became cornified as early as 20 days postgrafting. In addition, by 20 days the nylon scaffold degraded completely with minimal inflammatory response. While the interface between the graft and the wound bed stained positive for laminin, consistent with the synthesis of a lamina lucida layer, no other component of the basement membrane was synthesized. In addition, rete ridges were not synthesized and a thick fibrotic tissue layer characterized by a large fibroblast population and vascular in-growth was observed below the newly synthesized epidermal layer. Additionally, the cellular component of the LDR was critical in achieving these results. When fibroblasts were not seeded into the scaffold, the mesh rapidly separated from the wound bed and fibrovascular in-growth did not occur; the presence of keratinocytes in the graft was required to prevent contraction of the wound site (88,92). While the LDR showed the ability to induce regeneration of a neoepidermis, the LDR did not exhibit the ability to induce regeneration of a complete basement membrane or a dermal layer.

Living Skin Equivalent

Preparation of the LSE graft utilized a novel approach for producing a full thickness graft with 3D architecture. While the LDR consisted of an acellular, synthetic scaffold that was cultured *in vitro* with dermal cell then seeded with epidermal cells immediately prior to implantation, the LSE approach utilized dermal cells to create a cellular, organic structure *in vitro* that could be then seeded with epidermal cells and implanted into the wound site as a cellular, bilayer neotissue (78–80,82). The skin equivalent was formed by populating a collagen lattice with heterologous fibroblasts that *in vitro* contracted the lattice and synthesized additional extracellular matrix proteins that were incorporated into the base lattice. After this contraction period, the upper surface of the neodermal

layer was then seeded with a suspension of epidermal cells, primarily keratinocytes. Once seeded, the epidermal cells attached to the collagen scaffold, proliferated, and differentiated to form a multilayered, epidermal structure within 1–2 weeks. At this point, the collagen scaffold populated by fibroblasts with a neomature epidermis upper structure was then grafted into the wound site. This technology requires a significant culture period, necessitating temporary treatment of the wound for patients during the *in vitro* culture period if this technology were to be used in the clinical setting.

Detailed studies have been made of the structure of the LSE following *in vitro* culture but prior to grafting. The keratinocytes that were seeded onto the contracted collagen scaffold formed a multilayered, partially keratinized epidermis *in vitro* that included tonofilaments, keratohyalin granules, and desmosomes (82,155). The intercorneocyte lipid lamellae that were synthesized in the *stratum corneum* of the neoepidermis did not have the same repeating pattern of narrow and broad electron lucent bands that are responsible for the epidermal barrier properties in the normal, adult epidermis. As a result of this structural abnormality, the LSE has an increased water permeability compared with normal skin, and hence a greater chance for wound site dehydration (156). Although short segments of the lamina densa were observed along the dermo–epidermal convergence, the LSE did not exhibit a complete basement membrane layer at the end of the *in vitro* culture period (157,158). Rete ridges and skin appendages were also consistently absent. Continued structural and biological changes were observed in the LSE following grafting, indicating that remodeling was taking place (1,133).

The LSE was tested experimentally in full-thickness skin wounds in a rat model; the bilayer graft exhibited remodeling in both the neodermal and neoepidermal layers. A functional, fully differentiated epidermis was observed as early as 7 days following grafting, and by 14 days a vascularized subepidermal layer with many of the structural characteristics of normal dermis, such as a "basketweave" collagen fiber pattern, was present. While a pattern of collagen fibers was observed in the subepidermal layer, the fibers were much thinner and much more tightly packed than those fibers observed in physiologic dermis, and no rete ridge structure was synthesized (155). When the LSE was implanted into a mouse model, similar remodeling effects were observed in both the epidermal and subepidermal layers. A mature epidermal layer and basement membrane were synthesized *in vivo*, with only anchoring fibrils missing from these tissue structures. The subepidermal layer again exhibited densely packed collagen fibers, but a physiologic dermis was not observed, and rete ridges again failed to form (159). Experimental results across these animal models and a number of experimental trials were consistent in indicating that the LSE was able to induce regeneration of a mature epidermis and basement membrane, but a mature dermal layer was not observed and the dermo–epidermal junction remained flat, without any sign of rete ridges (1).

The results observed in experimental trials of the LSE led to a series of clinical trials for the LSE graft that focused on its potential use for treatment of severe burn patients. Full-thickness skin wounds that covered >15% of their body surface area were grafted with the LSE. Extensive lysing of the graft was observed as early as 2 days postgrafting, and after 2 weeks only one patient showed any significant amount of *take* ($\sim$40%). The investigators concluded that the LSE was not appropriate to use as a permanent treatment to replace the autograft for burn patients (160). Additional clinical studies were performed on patients who had skin tumors removed where these acute wounds were not full-thickness skin wounds. While acceptable *takes* and no evidence of graft toxicity or rejection were observed in these trials, wound contraction by 10–15% was also observed; wound contraction to this extent was significantly larger than what is observed after grafting with an autograft. A biopsy taken from the graft site showed evidence of scar formation, and the authors hypothesized that the LSE was replaced by host tissue through more traditional modes of injury response (i.e., contraction and scar synthesis) (161).

A final series of clinical trials involving the LSE used the graft to treat chronic wounds due to venous ulcers, a chronic skin defect that has been observed to be of variable depth. Patients were diagnosed as having chronic ulcers when their wounds remained open for at least 1 month prior to LSE implantation, although the median duration of the ulcers for these patients was $\sim$1 year. No information on the initial depth of each of the ulcers was reported, so it is not clear whether these were full- or partial-thickness skin wounds. In these studies, the use of the LSE was compared to a standard clinical treatment for such ulcers: bandaging with a compression regimen. Following a 6 month study, it was concluded that the time to wound closure was significantly shorter for patients treated with the LSE compared to those treated with standard bandages. However, observations from both experimental (animal models) and clinical trials indicated that while the LSE displayed a significant ability to regenerate both mature epidermal and basement membrane layers, dermal regeneration was not observed and normal skin architecture (i.e., collagen fiber architecture, rete ridges) was lacking in both animal and human models (1,162,163). More recent animal experiments utilizing the LSE have indicated that once implanted, the graft is rapidly incorporated into the host tissue and began to undergo remodeling; basic dermal organization, the appearance of specific ECM constituents such as type I, III, V, and VI collagen as well as elastin, and the gradual disappearance of contractile myofibroblasts occurred by 1-year postimplantation. However, complete regeneration of the tissue triad (epidermis, basement membrane, dermis) was not observed. The LSE possesses the adequate environment to remain bioactive >1 year postimplantation and provide a platform for long-term *in situ* therapy studying wound healing (164). While the LSE is used clinically to treat chronic (nonclosing) venous leg and diabetic foot ulcers and significantly increases the rate of wound closure and the likelihood of wound closure compared to conventional treatments (debridement followed by the application of a synthetic dressings), the LSE has yet to be able to replace the wound with normal, physiological skin.

Dermal Regeneration Template

The DRT is the fourth major paradigm investigated to treat massive skin injury. Like the LDR and the LSE, the DRT utilized a 3D scaffold as the basis for an implant to be grafted into a severe skin wound. But unlike the cell-seeded, synthetic, acellular scaffold of the LDR and the cell-seeded, organic, cellular LSE where the focus was not on developing a specific scaffold structure, the DRT was an acellular scaffold fabricated from primary components of the extracellular matrix (i.e., collagen, proteoglycans). The focus was on fabricating a bioactive ECM analog that presented a 3D architecture that induced the endogenous cell population in the wound site to regenerate the lost tissue (1,3). The DRT device sequentially performs two separate tasks during the wound healing process: prevent wound dehydration and infection while modifying the healing response to induce regeneration. The first task addressed by the DRT is the management of the acute phase of the clinical healing process; the graft must protect against severe fluid loss and prevent massive infection as a result of the open skin wound. The second task for the DRT to modify is the chronic phase of the wound healing process; in the chronic phase, wound contraction and scar synthesis take place, leaving a mechanically, functionally, and aesthetically inferior tissue compared to normal skin (repair mechanism).

The DRT is a two-stage device with a top layer of poly(dimethyl siloxane), a silicone elastomer, bonded to an active ECM analog beneath. The active ECM analog was designed to induce synthesis of new, physiological dermal tissue while the silicone elastomer was designed to prevent flow of exudate outside the defect, acting as a barrier to control moisture permeability and to shield the wound site from bacteria (109). The top silicone layer was designed so it would be easily removed after an initial healing period (Stage 1 of the healing response) so that a keratinocyte sheet could be grafted on top of the newly synthesized dermal bed, producing a stratified epithelial tissue on top of the neodermis. This grafting procedure was developed based on observations that the epidermis regenerates spontaneously provided that there is an underlying dermal structure, while the dermis does not. The graft design is also compatible with a view of the active ECM analog (template) as a structure that prevents contraction and scar formation, while the silicone elastomer layer protects the wound site from dehydration and infection. In the clinical setting (107), the silicone layer was replaced by a thin autoepidermal graft harvested from the patient once initial regeneration of the dermis occurred ($\sim$10–15 days). Following the introduction of a keratinocyte sheet, a mature epidermal and basement membrane layer forms on top of the regenerated dermis.

The DRT structure was optimized in studies utilizing animal models where it was observed that skin regeneration did not occur unless the ECM analog effectively delayed wound contraction. Preliminary optimization of the structural features of the ECM analog was quantitatively based on comparing the delay in the onset of contraction. The structural characteristics of an active ECM analog were individually studied to determine the optimal chemical composition, template residence time, and template pore structure needed to create a bioactive scaffold that prevented contraction and induced regeneration (1,3,134).

The ECM analogs that were used successfully to induce skin regeneration are graft copolymers of type I collagen and chondroitin-6-sulfate, a GAG, with a collagen:GAG weight ratio of 98:2 (3). The porous structure of the DRT was produced using a freeze–drying process. The collagen–GAG (CG) copolymer was produced in slurry form from microfibrillar collagen, an aqueous glycosaminoglycan solution, and acetic acid. The slurry was frozen to a final temperature of $-40\,^{\circ}$C and then sublimated [pressure $<$100 mtorr (13.3 Pa), temperature = $0\,^{\circ}$C]; after freezing, an interpenetrating network of ice crystals that is surrounded by collagen and GAG fibers has formed. The sublimation process converts all of the ice crystals formed in the frozen slurry into vapor and generates empty pores, creating a collagen–GAG scaffold formed from interconnected sublimated pores (3,165). The scaffold structure (pore size, specific surface) can be adjusted by varying the temperature of freezing of the CG slurry prior to sublimation; a lower temperature of freezing increases the frequency of nucleation of ice crystals in the slurry and decreases the rate of material transport within the slurry. These two processes result in the formation of smaller ice crystals, and therefore smaller pores after sublimation, for lower temperatures of freezing (3,166,167). In addition to being able to adjust the composition and the pore structure, the template residence time can be adjusted by changing the cross-link density, allowing the fabrication of scaffolds with specific degradation rates (3,105). Both chemically (i.e., glutaraldehyde or carbodiimide based cross-linking) and physically induced (i.e., dehydrothermal crosslinking) covalent bonds between collagen fibers (cross-links) can be introduced; as the scaffold becomes more highly cross-linked, it is increasingly difficult to degrade through enzymatic digestion. Adjustment of the severity of cross-linking (increasing the chemical exposure time or increasing the heat and/or exposure time to a dehydrating environment increases the severity of cross-linking) allows for adjustment of the scaffold crosslink density and therefore degradation rate (3,120). Previous research has determined that crosslinked collagen scaffolds degrade at a rate that monotonically decreases with increasing crosslink density (105). Systematic use of the contraction inhibition criterion was used to select the average pore diameter and biodegradation rate of the DRT that results in an optimal regeneration result. The speed of contraction of a wound site was measured using the wound half-life criterion, the time it took for the wound to decrease in area by 50% from its original size.

The kinetics of contraction of full-thickness skin wounds in the guinea pig were used to separate CG copolymer grafts into three classes: Class 0, I, or II (Fig. 2). The guinea pig model was employed during early studies because of the vigorous contraction observed in skin wounds; this rapid and significant contraction was used to identify templates that were active in preventing contraction and inducing regeneration (108). While the skin of most mammals is securely tethered to the underlying fascia and skin wounds

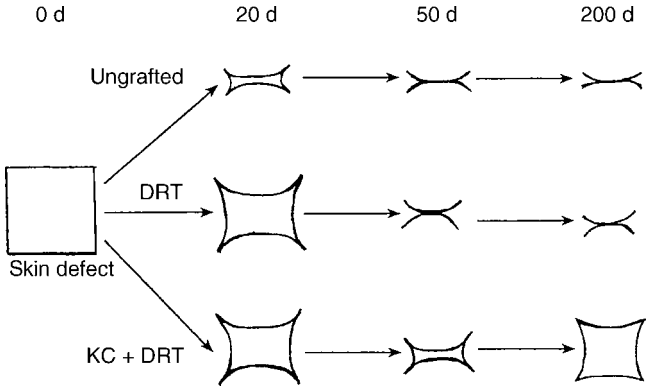


Figure 2. Schematic representation of wound healing kinetics in the adult guinea pig. A full-thickness skin wound was either ungrafted, grafted with the DRT, or grafted with a keratinocyte-seeded DRT, leading to three classes of wound healing: class 0, class I, or class II, respectively. *Top*: The ungrafted defect closed rapidly by contraction and formation of a characteristic "stellate" scar (class 0 healing). The ungrafted defect exhibited a half-life of 8 ± 1 days, and the wound contracted to $<10\%$ of the original area within 20 days. *Middle*: Grafting with a cell-free DRT led to a significant delay in wound contraction, and synthesis of a small mass of dermis and epidermis was observed. In the guinea pig wound model, however, the wound eventually closed by wound contraction (class I healing), but the resulting scar was more rounded than the "stellate scar". The unseeded DRT exhibited a half-life of 27 ± 2 days, but the wound contracted to $<10\%$ of the original area within 40 days. *Bottom*: Grafting with a keratinocyte-seeded DRT led to delay in wound contraction, followed by the complete arrest of contraction and expansion of the defect parameter due to synthesis of new, physiologic skin (class II healing). The keratinocyte-seeded DRT exhibited a half-life of 22 ± 2 days, and as a result of synthesis of new skin the wound edges reached a steady-state condition by 200 days after grafting that was $72 \pm 5\%$ of the original wound area (1).

close by contraction, but mainly scar synthesis, the skin of the guinea pig is not nearly as well tethered and skin wounds close almost entirely by contraction with very little scar synthesis and no regeneration (Fig. 3) (1,3,133). Class 0 healing, attained by grafting either no scaffold or a biologically inactive scaffold in the guinea pig model is characterized by a wound half-life of $\sim$1 week (8 ± 1 day); the wound was observed to contract to $<10\%$ of the original area within 20 days. Class I healing, attained by grafting a bioactive, cell-free scaffold into the full-thickness wound, is characterized by a significantly longer wound half-life (27 ± 2 days); however, the wound is still observed to contract to $<10\%$ of its original area after $\sim$40 days. While this implant in the guinea pig still shows significant contraction of the wound site, in other animal models that more closely mimic human skin contraction kinetics, class I devices with the guinea pig prevented significant wound contraction (1). The final mode of guinea pig wound healing (class II) was observed when the wound was grafted with a bioactive scaffold seeded with cells (keratinocytes); this healing was characterized by an extended wound half life (22 ± 2 days) compared to class 0 healing. After $\sim$200 days, the wound was observed to have stabilized at $\sim 72 \pm 5\%$ of the original wound area. Due to the more secure attachment of the skin to the underlying fascia with humans,

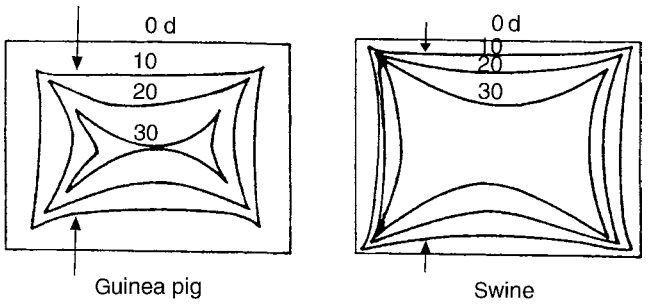


Figure 3. Kinetics of wound closure of full-thickness skin wounds (mediated primarily by contraction) in guinea pig and swine animal models. Each time point represents the change in defect area relative to the original wound dimensions on the indicated day after wounding. The guinea pig exhibits much more robust wound contraction compared to the swine model, arriving at a significantly smaller asymptotic wound area (1).

both class I and class II healing devices are useful clinically. The scaffold that led to these two healing modes was identified as having maximal biological activity, termed the DRT.

A homologous series of ECM analogs, varying in scaffold pore diameter from $\sim$10 to 1000 μm, was used in the above studies to determine the optimal pore diameter necessary to prevent class 0 contraction (Fig. 4). Maximum delay of wound contraction half-life, up to 27 ± 3 days, was observed when the average pore diameter was in

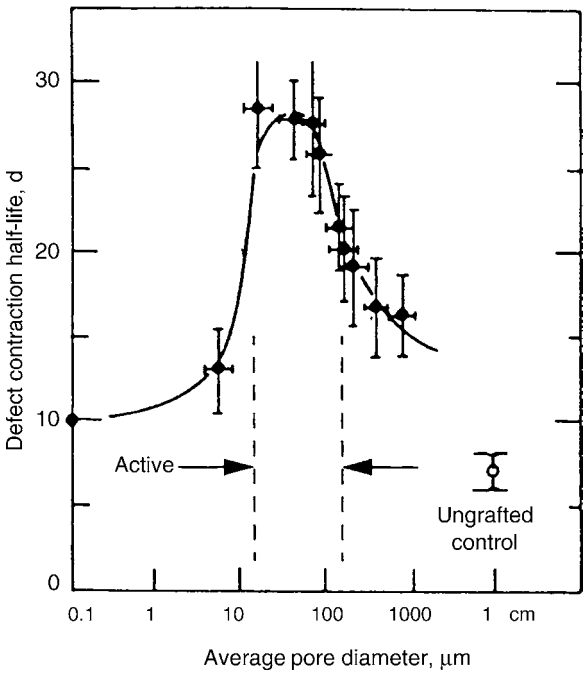


Figure 4. Identification of the optimal pore diameter for a bioactive extracellular matrix analog designed to induce skin regeneration. The range of maximal contraction-delaying activity for collagen–GAG scaffolds was observed when the average pore diameter of the scaffold was between 20 and 120 μm. The limits to the area of maximal activity are indicated by broken lines and correspond to the range where contraction was most delayed (13).

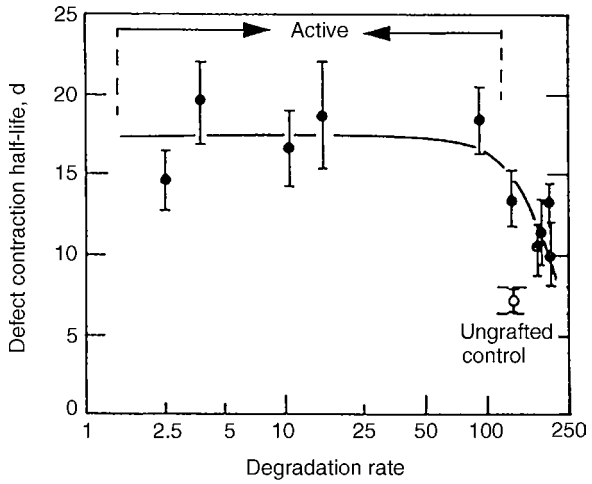


Figure 5. Identification of the optimal degradation rate for a bioactive extracellular matrix analog designed to induce skin regeneration. Contraction delaying activity of the collagen–GAG scaffold was observed when the device degradation rate was maintained between the two levels indicated in the figure. The degradation rate is expressed in units of standardized collagenase solution (*in vitro* assay). When the degradation rate was <2.5 enzyme units or >100 units, scar synthesis and contraction were observed. The limits to the area of maximal activity are indicated by broken lines, and correspond to the range where contraction was delayed the most (13).

the range between a lower limit of 20 ± 4 μm and an upper limit of 125 ± 35 μm (3). The lower limit in pore diameter is on the order of two cell diameters, suggesting that the scaffold pore size must be large enough to maintain space for dermal fibroblasts, the major cellular constituent of the dermis, to migrate from the wound bed into the DRT. The upper limit has been hypothesized to be as a result of an apparent requirement for attachment of a minimal number of cells onto the surface of the scaffold (167). Another homologous series of ECM analogs, where the scaffold degradation rate was varied, was also studied to match the degradation rate of the scaffold with synthesis of new physiologic tissue, as outlined in the section discussing the isomorphous tissue replacement rule (Fig. 5). A significant delay in wound contraction was observed when the degradation rate was $<115 \pm 25$ enzyme units; the enzyme units used here are arbitrary units based on an *in vitro* assay using bacterial collagenase to degrade the scaffolds and then correlating these results with the ability of each device to prevent wound contraction in the guinea pig model (3,103,105). The upper limit in degradation rate is consistent with a lower limit for the time of biodegradation of the scaffold, indicating that the scaffold needs to be present for a specific amount of time in the wound site when the contractile response is active in order to prevent contraction (~2–3 weeks in humans) (1).

The knowledge obtained from the series of experiments used to determine the optimal chemical composition, degradation kinetics, and pore structure that each prevented wound contraction were required in order to interpret the unusual biological activity of this specific ECM analog, termed the DRT. These characteristics are apparently required to block contraction for the entire period that contraction was active in the healing wound bed. Host fibroblasts were observed to migrate into the DRT from the wound edges at a speed of ~0.2 mm/d, making the endogenous cells able to cross the entire 0.5-mm thickness of the DRT within a few days, provided the appropriate contact with the wound bed was available. With a calculated critical cell path on the order of 100 μm, a distance that is filled largely with wound exudate, growth factors and serum nutrients, it was expected that fibroblasts from the host (surrounding) tissue were able to migrate readily into the DRT, multiply, and differentiate, leading to the synthesis of a physiologic dermis (1,3).

In addition to inducing the synthesis of a physiological dermis using the silicone-covered DRT, it was necessary to induce formation of a mature epithelial layer if the DRT was to be used as a permanent treatment. In the case of small wounds, keratinocytes can migrate from the wound edges across the top of the regenerating dermis. Once the keratinocytes migrate over the entire surface of the new dermis, the keratinocytes multiply and differentiate, creating a mature, stratified epidermis and basement membrane (3,133). This process was termed sequential regeneration because regeneration of the dermis had to occur prior to keratinocyte migration from the wound edges across the top of the regenerating dermis to form the neoepidermis. For larger wounds, the cells can only migrate into the edges of the wound due to their average migratory speeds of 0.2–0.5 mm/d, and are unable to cover the surface of the DRT rapidly enough to create a functional epidermal layer over the entire wound surface. In this case, it is necessary to supply an exogenous source of keratinocytes. Two distinct techniques have been utilized to overcome this shortcoming. The first utilizes uncultured autologous epidermal cells that are harvested using a skin biopsy from the patient and are then seeded into the DRT prior to implantation. This procedure results in the formation of a confluent epidermis after ~2 weeks provided that a large enough amount of cells were seeded originally ($>5 \times 10^4$ epithelial cells/cm^2 area of DRT); this modified process has been termed simultaneous regeneration (133). A second procedural option for introducing keratinocytes to the wound can be performed when the silicone elastomer is removed (after ~2–3 weeks). When the silicone is removed, a thin autoepidermal graft is applied to the surface of the regenerating dermis (107). Approximately equivalent final results have been observed when utilizing either of these techniques for creating a mature epidermal and basement membrane layer over the regenerating dermis (1,3,133).

The quality of regeneration using the DRT was determined through immunohistochemical analysis of the regenerated tissue and comparing these results to those seen in normal skin (Fig. 6). The regenerated skin was observed to have the three tissue layers present in normal skin (tissue triad): an underlying dermis, the stratified epidermis, and a basement membrane layer between the two. For all of the following images, observations, and analyses discussed in this paragraph, the DRT treatment utilized was the DRT seeded with autologous keratinocytes prior to implantation into a full-thickness skin wound in a

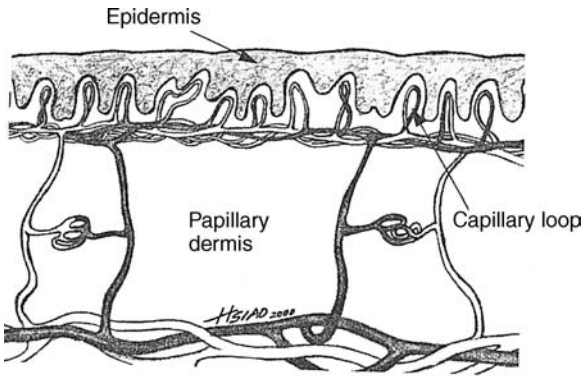


Figure 6. Diagram of normal skin showing the characteristic rete ridges at the dermo–epidermal junction and the vascular network (capillary loops) that populates the subepidermal region (1).

porcine model. As early as day 12 postgrafting, anchoring fibers were observed in the regenerated basement membrane (Fig. 7). By day 35 postgrafting, a rete ridge structure had formed, complete with vascular loops (Fig. 9), and a confluent hemidesmosomal staining pattern was observed at the dermo–epidermal junction (Fig. 8). All of these immunohistochemical results indicated that the regenerating skin was taking on all of the structural properties observed in the epidermis, dermis, and basement membrane of normal skin, except for skin appendages such as sweat glands and hair follicles (115).

Clinical studies of the DRT initially emphasized treatment of massively burned patients while later studies have focused on patients who elected to have plastic and reconstructive surgery. In both patient populations, the primary defect is severe, traumatic skin loss down to the underlying muscle and fascia. In studies with burn patients, burns were excised down to muscle fascia prior to grafting with DRT. Wound closure was accelerated greatly by removing the silicone cover after ~2–4 weeks and grafting the newly synthesized dermis with an ultrathin autoepidermal graft, nominally 100 μm in thickness, that was free from dermal components (107,168). In a detailed histological study of the resulting new organ the endogenous cell population degraded the DRT structure; remodeling then occurred and the newly synthesized collagen fibers became coarser and a distinction between papillary and reticular layers of the dermis appeared. Scar synthesis was not observed either at a gross or at a histological level at any time during the course of healing. In contrast to studies with the swine in which rete ridges had clearly formed (115), rete ridges were not reported in this study with humans and skin appendages were absent from the human as they were from studies with swine and rodents (169). A related immunological study showed a very small rise in immunological activity in patients' sera for the components of DRT: bovine skin collagen and chondroitin 6-sulfate. The overall conclusion from the clinical studies was that DRT presents few, if any, immunological problems to patients (170). In other clinical studies of DRT, the focus was on follow-up of massively burned pediatric patients over a 6- or 10-year period (171–173). It was reported that the new integument did not restrict joint function, suggesting the absence of crippling wound contraction and that the new skin had the ability to grow and mature long after grafting even when children were treated with the DRT (171,173). The DRT has also been used to treat patients with purpura fulminans (174), to release skin contractures (175), and to resurface scarred areas resulting from full-thickness burns (176,177); in all cases, regeneration of a functional epidermis, basement membrane, and dermis was observed.

Naturally Derived Collagen Matrices

Separate from the development of a series of scaffold materials for use to treat severe skin injuries, a separate technology has been developed to utilize naturally derived collagen matrices (NDCM) to treat skin injuries. Instead of relying on technologies to fabricate a 3D scaffold structure

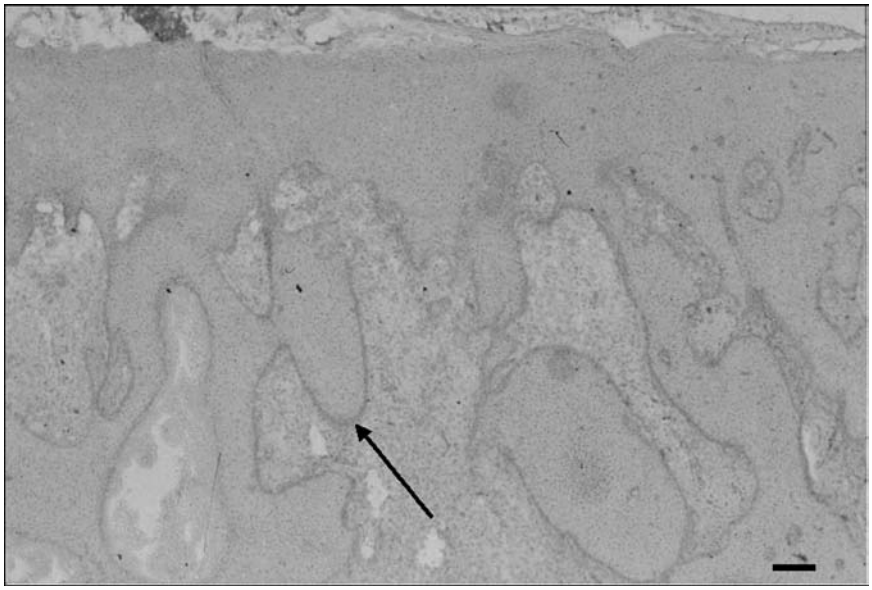


Figure 7. As early as 12 days after grafting a full-thickness skin wound with a keratinocyte-seeded DRT, anchoring fibrils were observed in the regenerating basement membrane (arrow). The basal surface epithelium and the periphery of the epithelial cords are labeled with type VII collagen immunostaining, identifying the anchorage structures at the dermo-epidermal interface. Bar: 150 μm. Reprinted with permission from Reference (115).

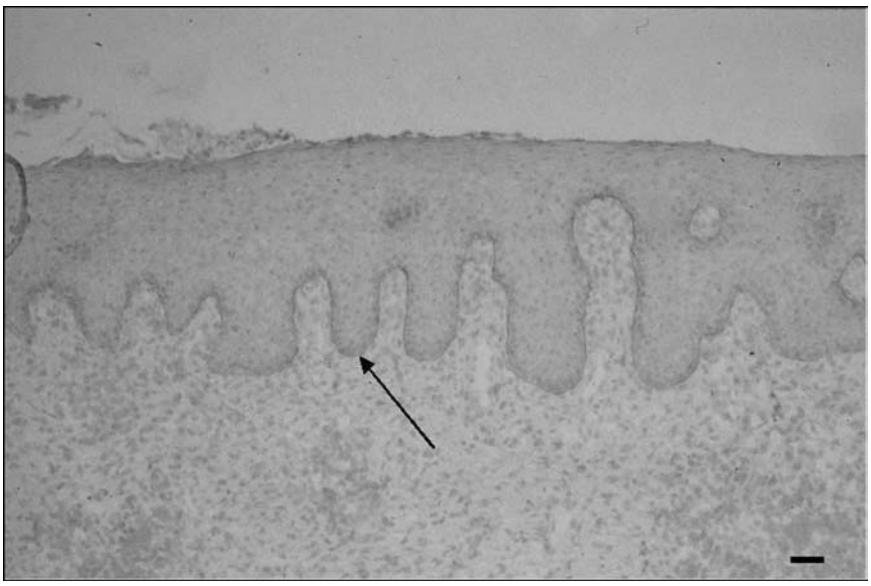


Figure 8. As early as 35 days after grafting a full-thickness skin wound with a keratinocyte-seeded DRT, a confluent hemidesmosomal staining pattern is observed at the dermo-epidermal junction (arrow) by immunostaining for the $\alpha_6\beta_4$ integrin. The pattern observed in the regenerating skin is identical to that observed in physiologic skin. Bar: 100 μm. Reprinted with permission from Reference (115).

from either synthetic or natural materials (i.e., CEA, LSE, LDR, or DRT), this technology uses decellularized dermal tissue as a scaffold structure. The most analogous treatment to the NDCM that has already been discussed in this article is the allograft and xenograft; the NDCM is a decellularized version of an allograft. The NDCM is designed to serve as a scaffold to support normal tissue remodeling following severe injury, thereby inducing regeneration.

The NDCM is produced through a three-step process. The epidermal tissue is completely removed from full-thickness autograft tissue, leaving both the dermal tissue and the basement membrane. The dermal cells are then removed using detergent washes. The decellularized tissue is then freeze-dried to preserve the NDCM structure and to maintain the bioactivity of the dermal matrix. The main advantage of the NDCM over the homograft and xenograft are that owing to decellularization, the antigenicity of the scaffold is significantly reduced. Reducing the antigenicity reduces the immunological response of the patient to the graft and reduces the chance of implant rejection. To treat a severe skin wound, the NDCM can be rehydrated in saline solution then be implanted directly into a wound site in the same manner as a tissue autograft. The NDCM has been used primarily to treat full-thickness burns and burns to areas of the skin where contraction and scar formation would inhibit functionality (i.e., feet and hands). AlloDerm, a product of the LifeCell Corp., is a common NDCM available for experimental and clinical trials.

The NDCM is implanted into full-thickness skin wounds and is often covered by a thin autograft of the patient's own epidermis to speed the healing process. This treatment typically results in a high percentage of graft take; additionally, the thin autograft of epidermal tissue significantly decreases the time for complete reepithelialization of the graft and reduces the number of complications due to infection. Patients showed normal range of motion, grip strength, motor control, and functionality following

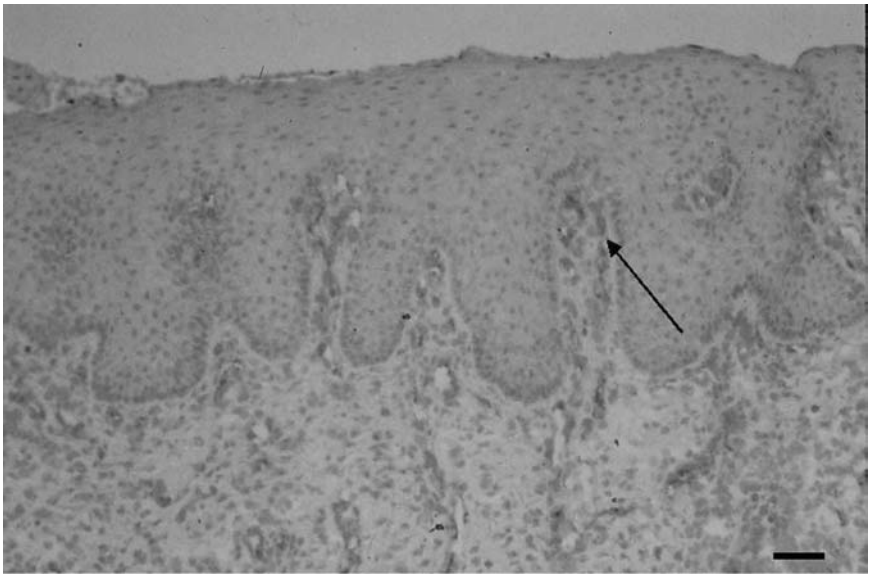


Figure 9. A full-thickness skin wound grafted with a keratinocyte-seeded DRT was observed to regenerate many of the structure observed in normal skin. Immunostaining for Factor VIII 35 days after grafting revealed that capillary loops had formed in the rete ridges of the regenerated dermis (arrow) similar to those observed in physiologic skin. Bar: 75 μm. Reproduced with permission from Reference (115).

treatment (178,179). The NDCM technology demonstrates the use of a naturally derived material in tissue engineering applications. A major restriction on the use of an NDCM such as Alloderm is the same as that faced by allografting: a limited supply of donor tissue available for grafting.

DISCUSSION AND CONCLUSIONS

Future improvements in treating injuries to various tissues and organs, including skin, via tissue engineering protocols will eventually depend critically on theoretical developments that explain the mechanism of tissue and organ replacement, thereby suggesting novel experimental studies and new paradigms for inducing regeneration. The goal of this article was to discuss the process of wound healing following injury, specifically the response of the skin to severe injuries and to discuss a number of paradigms that have been investigated in an attempt to alter the organismic (contraction-mediated) response to severe injuries. Evidence has been introduced to support the conclusion that contraction is the main engine for wound closure both in skin and peripheral nerve wounds, that scar formation is a byproduct of contraction, and that there is an antagonistic relationship between contraction and regeneration. Regeneration of skin and peripheral nerve requires selective blocking of contraction, but not blocking of other aspects of the healing process. The structural requirements for bioactive regeneration templates were also discussed in the light of this theory, and a number of distinct paradigms using very different scaffold structures were studied to determine the experimental and clinical outcomes of their use. This article conclude with a detailed discussion of the mechanism of regeneration after implantation of the DRT into severe wound sites.

The empirical evidence described here is consistent with the conclusion that synthesis of an organ *in vivo* (induced regeneration) requires grafting of an appropriate scaffold that has or has not been seeded with epithelial cells from the desired organ. In this approach, there is no requirement for addition of exogenous reactants such as mesenchymal cells (i.e., fibroblasts) or for addition of cytokines. These empirical findings appear to have direct relevance for the future selection of tissue engineering protocol based on both *in vitro* and *in vivo* environments.

ACKNOWLEDGMENTS

The authors acknowledge partial support from the NIH/NIBIB, grant DE13053, as well as a grant from the Cambridge University-MIT Institute.

BIBLIOGRAPHY

1. Yannas IV. Tissue and Organ Regeneration in Adults. New York: Springer-Verlag; 2001.
2. Yannas IV. In vitro synthesis of tissues and organs. In: Lanza RP, Langer R, Vacanti J, editors. Principles of Tissue Engineering. 2nd ed. San Diego: Academic Press; 2000.
3. Yannas IV, Lee E, Orgill DP, Skrabut EM, Murphy GF. Synthesis and characterization of a model extracellular matrix that induces partial regeneration of adult mammalian skin. Proc Natl Acad Sci USA 1989;86:933–937.
4. Mast BA, Diegelmann RF, Krummel TM, Cohen IK. Scarless wound healing in the mammalian fetus. Surg Gynecol Obstet 1992;174:441–451.
5. Mast BA, Nelson JM, Krummel TM. Tissue repair in the mammalian fetus. In: Cohen IK, Diegelmann RF, Lindblad WJ, editors. Wound Healing. Philadelphia: W.B. Saunders; 1992.
6. Wallace H. Vertebrate Limb Regeneration. New York: Wiley; 1981.
7. Yannas IV, Colt J, Wai YC. Wound contraction and scar synthesis during development of the amphibian Rana catesbeiana. Wound Rep Reg 1996;4:31–41.
8. Martinez-Hernandez A. Repair, regeneration, and fibrosis. In: Rubin E, Farber JL, editors. Pathology. Philadelphia: J.B. Lippincott; 1988.
9. Uitto J, Mauviel A, McGrath J. The dermal-epidermal basement membrane zone in cutaneous wound healing. In: Clark RAF, editor. The Molecular and Cellular Biology of Wound Repair. New York: Plenum Press; 1996.
10. Fu SY, Gordon T. The cellular and molecular basis of peripheral nerve regeneration. Mol Neurobiol 1997;14:67–116.
11. Haber RM, Hanna W, Ramsay CA, Boxall LB. Cicatricial junctional epidermolysis bullosa. J Am Acad Dermatol 1985;12:836–844.
12. Ikeda K, Oda Y, Tomita K, Nomura S, Nakanishi I. Isolated Schwann cells can synthesize the basement membrane in vitro. J Electron Microsc (Tokyo) 1989;38:230–234.
13. Stenn KS, Malhotra R. Epithelialization. In: Cohen IK, Diegelmann RF, Lindblad WJ, editors. Wound Healing. Philadelphia: W. B. Saunders; 1992.
14. Ferdman AG, Yannas IV. Scattering of light from histologic section: A new method for the analysis of connective tissue. J Invest Dermatol 1993;100:710–716.
15. Lehv M, Fitzgerald PJ. Pancreatic acinar cell regeneration IV: Regeneration after surgical resection. Am J Pathol 1968;53:513–535.
16. Oliver J. Correlations of structure and function and mechanisms of recovery in acute tubular necrosis. Am J Med 1953;15:535–557.
17. Stemerman MB, Spaet TH, Pitlick F, Cintron J, Lejnieks I, Tiell ML. Intimal healing. The patterns of reendothelialization and intimal thickening. Am J Pathol 1977;87:125–142.
18. Vracko R. Significance of basal lamina for regeneration of injured lung. Virchows Arch (Pathol Anat) 1972;355:264–274.
19. Yannas IV. Regeneration of skin and nerves by use of collagen templates. In: Nimni N, editor. Collagen: Biotechnology. Boca Raton: CRC Press; 1988.
20. Lanza RP, Langer R, Chick WL, editors. Principles of Tissue Engineering. San Diego: Academic Press; 1997.
21. Brown H. Wound healing research through the ages. In: Cohen IK, Diegelmann RF, Lindblad WJ, editors. Wound Healing. Philadelphia: W.B. Saunders; 1992.
22. Bach FH, Robson SC, Winker H, Ferran C, Stuhlmeiser KM, Wrighton CJ, Hancock WW. Barriers to xenotransplation. Nature Med 1995;1:869–873.
23. Cooper DKC, Kemp E, Platt JL, White DJG, editors. Xenotransplantation. New York: Springer-Verlag; 1997.
24. Lanza RP, Chick WL. Endocrinology: Pancreas. Lanza RP, Langer R, Chick WL, editors. Principles of Tissue Engineering. San Diego: Academic Press; 1997.

25. Medawar PB. The behavior and fate of skin autografts and skin homografts in rabbits. J Anat 1944;78:176–199.
26. Medawar PB. The storage of living skin. Proc R Soc Med 1954;47:62–64.
27. Murray JE, Merrill JP, Harrison JH. Renal homotransplantations in identical twins. Surg For 1955;6:432–436.
28. Wickelgren I. Muscling transplants in mice. Science 1996;273:33.
29. Lau HT, Fontana A, Stoeckert CJ. Prevention of islet allograft rejection with engineered myoblasts expressing FasL in mice. Science 1996;273:109–112.
30. Avgoustiniatos ES, Colton CK. Design considerations for immunoisolation. In: Lanza RP, Langer R, Chick WL, editors. Principles of Tissue Engineering. San Diego: Academic Press; 1997.
31. Lim F, Sun AM. Microencapsulated islets as bioartificial endocrine pancreas. Science 1980;210:908–910.
32. Lanza RP, Cooper DKC, Chick WL. Xenotransplantation. Sci Am 1997;July: 54–59.
33. Kaiser J. IOM backs cautious experimentation. Science 1996;273:305–306.
34. Patience C, Takeuchi Y, Weiss RA. Infection of human cells by an endogenous retrovirus of pigs. Nature Med 1997;3: 282–286.
35. Sikorski R, Peters R. Xenotransplanters turn xenovirologists. Science 1997;276:1893.
36. Burke JF, Bondoc CC, Quinby WC. Primary burn excision and immediate grafting: A method of shortening illness. J Trauma 1974;14:389–395.
37. Millesi H. Erfahrungen mit der Mikrochirurgie peripherer Nerven. Chir Plast Reconstr 1967;3:47–55.
38. Sunderland S. Nerves and Nerve Injuries. New York: Churchill Livingstone; 1978.
39. Terzis JK. Microreconstruction of Nerve Injuries. Philadelphia: W.B. Saunders; 1987.
40. Millesi H, Meissl G, Berger G. The interfasicular nerve grafting of the median and ulnar nerves. J Bone Joint Surg 1972;54-A:727–750.
41. Millesi H, Meissl G, Berger G. Further experience with interfascicular grafting of the median, ulnar, radial nerves. J Bone Joint Surg 1976;58-A:209–216.
42. Grondin CM, Campeau I, Thorton JC, Engle JC, Cross FS, Schreiber H. Coronary artery bypass grafting with saphenous vein. Circ 1989;79(I): 24–29.
43. Kohn DH, Ducheyne P. Materials for bone and joint replacement. In: Williams DF, editor. Materials Science and Technology. New York: VCH Publishers; 1992.
44. Neuman MR. Therapeutic and prosthetic devices. In: Webster JG, editor. Medical Instrumentation. 4th ed. New York: Wiley; 1998.
45. Schoen FJ, Hobson CE. Anatomic analysis of removed prosthetic heart valves: Causes of failure of 33 mechanical valves and 58 bioprostheses, 1980 to 1983. Hum Pathol 1985;16(6): 549–559.
46. Caldarone CA, McCrindle BW, Van Arsdell GS, Coles JG, Webb G, Freedom RM, Williams WG. Independent factors associated with longevity of prosthetic pulmonary valves and valved conduits. J Thorac Cardiovasc Surg 2000;120(6): 1022–1030.
47. Facer GW, Peterson A, Brey RH, Marion M, Cevette M, Balko K, Green JD, Rose D, Pool A. The mayo clinic experience with the cochlear implant. Ear Nose Throat J 1994;73(3): 149–152.
48. Peppas NA, Langer R. New challenges in biomaterials. Science 1994;263:1715–1720.
49. Ginsbach G, Busch LC, Kuhnel W. The nature of the collagenous capsules around breast implants. Plast Reconst Surg 1979;64:456–464.
50. Rudolph R, Van de Berg J, Ehrlich P. Wound contraction and scar contracture. In: Cohen IK, Diegelmann RF, Lindblad WJ, editors. Wound Healing. Philadelphia: W.B. Saunders; 1992.
51. Spector M, Heyligers I, Roberson JR. Porous polymers for biological fixation. Clin Orth 1993;235:207–219.
52. Chandy T, Das GS, Wilson RF, Rao GH. Surface-immobilized biomolecules on albumin modified porcine pericardium for preventing thrombosis and calcification. Int J Artif Organs 1999;22:547–558.
53. Park JY, Davies JE. Red blood cell and platelet interactions with titanium implant surfaces. Clin Oral Implants Res 2000;11:530–539.
54. Snyder TA, Watach MJ, Litwak KN, Wagner WR. Platelet activation, aggregation, life span in calves implanted with axial flow ventricular assist devices. Ann Thorac Surg 2002;73:1933–1938.
55. Suggs LJ, West JL, Mikos AG. Platelet adhesion on a bioresorbable poly(propylene fumarate-co-ethylene glycol) copolymer. Biomaterials 1999;20:683–690.
56. Basle MF, Bertrand G, Guyetant S, Chappard D, Leonard M. Migration of metal and polyethylene particles from articular prostheses may generate lymphadenopathy with histiocytosis. J Biomed Mater Res 1996;30(2): 157–163.
57. Urban RM, Jacobs JJ, Tomlinson MJ, Gavrilovic J, Black J, Peoc'h M. Dissemination of wear particles to the liver, spleen, and abdominal lymph nodes of patients with hip and knee replacement. J Bone Joint Surg Am 2000;82(4): 457–476.
58. Willert HG. Reactions of the articular capsule to wear products of artificial joint prostheses. J Biomed Mater Res 1977;11(2): 157–164.
59. Prockop DJ. Marrow stromal cells as stem cells for nonhematopoietic tissues. Science 1997;276:71–74.
60. Solter D, Gearhart J. Putting stem cells to work. Science 1999;283:1468–1470.
61. Pittenger MF, Mackay AM, Berk SC, Jaiswal RK, Douglas R, Mosca JD, Morman MA, Simonetti DW, Craig S, Marshak DR. Multilineage potential of adult human mesenchymal stem cells. Science 1999;284:143–147.
62. Slack JMW. Stem cells in epithelial tissues. Science 2000; 287:1431–1433.
63. Gage FH. Mammalian neural stem cells. Science 2000; 287:1433–1438.
64. Yannas IV. Synthesis of organs: In vitro or in vivo? Proc Natl Acad Sci USA 2000;97:9354–9356.
65. Compton CC. Keratinocyte grafting models. In: Lane EB, Leigh I, Watt F, editors. The Keratinocyte Handbook. London: Cambridge University Press; 1994.
66. Karasek MA. In vitro culture of human skin epithelial cells. J Invest Dermatol 1966;47:533–540.
67. Karasek MA. Growth and differentiation of transplanted epithelial cell cultures. J Invest Dermatol 1968;51: 247–252.
68. Worst PKM, Valentine EA, Fusenig NE. Formation of epidermis after reimplantation of pure primary epidermal cell cultures from perinatal mouse skin. J Natl Cancer Inst 1974;53:1061–1064.
69. Compton CC, Gill JM, Bradford DA, Regauer S, Gallico GG, O'Connor NE. Skin regenerated from cultured epithelial autografts on full-thickness burn wounds from 6 days to 5 years after grafting. Lab Invest 1989;60:600–612.
70. Eldad A, Burt A, Clarke JA, Gusterson B. Cultured epithelium as a skin substitute. Burns 1987;13:173–180.

71. Gallico GG, O'Connor NE, Compton CC, Kehinde O, Green H. Permanent coverage of large burn wounds with autologous cultures human epithelium. New Engl J Med 1984;311:448–451.
72. Green H, Kehinde O, Thomas J. Growth of cultured human epidermal cells into multiple epithelia suitable for grafting. Proc Natl Acad Sci USA 1979;76:5665–5668.
73. Green H, Rheinwald JG. Process for serially culturing keratinocytes, US patent 4,016,036, 1977.
74. Munster AM. Use of cultured epithelial autograft in ten patients. J Burn Care Rehab 1992;13:124–126.
75. Munster AM. Cultured skin for massive burns. Ann Surg 1996;224:372–377.
76. Rheinwald JG, Green H. Formation of a keratinizing epithelium in culture by a cloned cell line derived from a tetroma. Cell 1975;6:317–330.
77. Rheinwald JG, Green H. Serial cultivation of strains of human epidermal keratinocytes: The formation of keratinizing colonies from single cells. Cell 1975;6:331–343.
78. Bell E, Ehrlich HP, Buttle DJ, Nakatsuji T. Living skin formed in vitro and accepted as skin-equivalent tissue of full thickness. Science 1981;211:1052–1054.
79. Bell E, Ehrlich HP, Sher S, Merrill C, Sarber R, Hull B, Nakatsuji T, Church D, Buttle DJ. Development and use of a living skin equivalent. Plast Reconstr Surg 1981;67:386–392.
80. Bell E, Ivarsson B, Merrill C. Production of a tissue-like structure by contraction of collagen lattices by human fibroblasts of different proliferative potential in vitro. Proc Natl Acad Sci USA 1979;76:1274–1278.
81. Bell E, Sher S, Hull B. The living skin equivalent as a structural and immunological model in skin grafting. Scan Electr Micr 1984;4:1957–1962.
82. Bell E, Sher S, Hull B, Merrill C, Rosen S, Chamson A, Asselinea D, Dubertret L, Coulomb B, Lapiere C. The reconstitution of living skin. J Invest Dermatol 1983;81:2s–10s.
83. Drumheller PD, Hubbell JA. Surface immobilization of adhesion ligands for investigations of cell-substrate interactions. In: Bronzio JD, editor. The Biomedical Engineering Handbook. Boca Raton: CRC Press; 1997.
84. Griffith Cima L. Polymeric biomaterials. Acta Mater 1994;48:263–277.
85. Langer R, Vacanti JP. Tissue engineering. Science 1993;260:920–928.
86. Lanza RP, Langer R, Vacanti J, editors. Principles of Tissue Engineering. 2nd ed. San Diego: Academic Press; 2000.
87. Cooper ML, Hansbrough JF. Use of a composite skin graft composed of cultured human keratinocytes and fibroblasts and a collagen-GAG matrix to cover full-thickness wounds on athymic mice. Surgery 1991;109:198–207.
88. Cooper ML, Hansbrough JF, Spielvogel RL, Cohen R, Bartel RL, Naughton G. In vivo optimization of a living dermal substitute employing cultured human fibroblasts on a biodegradable polyglycolic acid or polyglactin mesh. Biomaterials 1991;12:243–248.
89. Dore C, Noordenbos J, Hansbrough JF. Management of partial thickness burns with Dermagraft-TC. J Burn Care Rehabil 1998;19:S172.
90. Hansbrough JF, Cooper ML, Cohen R, Spielvogel R, Greenleaf G, Bartel RL, Naughton G. Evaluation of biodegradable matrix containing cultured human fibroblasts as a dermal replacement beneath meshed skin grafts on athymic mice. Surgery 1992;111:438–446.
91. Hansbrough JF, Dore C, Hansbrough WB. Clinical trials of a living dermal tissue replacement placed beneath meshed, split-thickness skin grafts on excised burn wounds. J Burn Care Rehab 1992;13:519–529.
92. Hansbrough JF, Morgan JL, Greenleaf GE, Bartel R. Composite grafts of human keratinocytes grown on a polyglactin mesh-cultured fibroblast dermal substitute function as a bilayer skin replacement in full-thickness wounds on athymic mice. J Burn Care Rehab 1993;14:485–494.
93. Naughton G, Mansbridge J, Gentzkow G. A metabolically active human dermal replacement for the treatment of diabetic foot ulcers. Artif Org 1997;21:1–7.
94. Purdue GF, Hunt JL, Still JM, Law EJ, Herndon DN, Goldfarb IW, Schiller WR, Hansbrough JF, Hickerson WL, Himel HN. A multicenter clinical trial of a biosynthetic skin replacement, Dermagraft-TC, compared with cryopreserved human cadaver skin for temporary coverage of excised burn wounds. J Burn Care Rehab 1997;18:52–57.
95. Park A, Wu B, Griffith LG. Integration of surface modification and 3D fabrication techniques to prepare patterned poly(L-lactide) substrates allowing regionally selective cell adhesion. J Biomater Sci Polym Ed 1998;9(2): 89–110.
96. Powers MJ, Domansky K, Kaazempur-Mofrad MR, Kalezi A, Capitano A, Upadhyaya A, Kurzawski P, Wack KE, Stolz DB, Kamm R, Griffith LG. A microfabricated array bioreactor for perfused 3D liver culture. Biotechnol Bioeng 2002;78(3): 257–269.
97. Xu J, Clark RAF. Integrin regulation in wound repair. In: Garg HG, Longaker MT, editors. Scarless Wound Healing. New York: Marcel Dekker; 2000.
98. Freed LE, Vunjak-Novakovic G. Cultivation of cell-polymer tissue constructs in simulated microgravity. Biotechnol Bioeng 1995;46:306–313.
99. Freed LE, Vunjak-Novakovic G, Briton RJ, Eagles DB, Lesnoy DC, Barlow SK, Langer R. Biodegradable polymer scaffolds for tissue engineering. Biotechnology 1994;12:689–693.
100. Lee CR, Breinan HA, Nehrer S, Spector M. Articular cartilage chondrocytes in type I and type II collagen-GAG matrices exhibit contractile behavior in vitro. Tissue Eng 2000;6(5): 555–565.
101. Nehrer S, Breinan HA, Ramappa A, Shortkroff S, Young G, Minas T, Sledge CB, Yannas IV, Spector M. Canine chondrocytes seeded in type I and type II collagen implants investigated in vitro. J Biomed Mater Res 1997;38(2): 95–104.
102. Dagalakis N, Flink J, Stasikelis P, Burke JF, Yannas IV. Design of an artificial skin. Part III. Control of pore structure. J Biomed Mater Res 1980;14:511–528.
103. Yannas IV, Burke JF, Huang C, Gordon PL. Suppression of in vivo degradability and of immunogenicity of collagen by reaction with glycosaminoglycans. Polym Prepr Am Chem Soc 1975;16(2): 209–214.
104. Yannas IV, Burke JF, Gordon PL, Huang C, Rubinstein RH. Design of an artificial skin II: Control of chemical composition. J Biomed Mater Res 1980;14:107–131.
105. Yannas IV, Burke JF, Huang C, Gordon PL. Correlation of in vivo collagen degradation rate with in vitro measurements. J Biomed Mater Res 1975;6:623–625.
106. Yannas IV, Burke JF, Umbreit M, Stasikelis P. Progress in design of an artificial skin. Fed Proc Am Soc Exp Biol 1979;38:988.
107. Burke JF, Yannas IV, Quincy WC, Bondoc CC, Jung WK. Successful use of a physiologically acceptable artificial skin in the treatment of extensive burn injury. Ann Surg 1981;194:413–428.
108. Yannas IV. Use of artificial skin in wound management. In: Dineen P editor. The Surgical Wound. Philadelphia: Lea & Febiger; 1981. p 171–190.
109. Yannas IV, Burke JF. Design of an artificial skin I. Basic design principles. J Biomed Mater Res 1980;14:65–81.

110. Yannas IV, Burke JF, Huang C, Gordon PL. Multilayer membrane useful as synthetic skin. US patent 4,060,081. 1977.
111. Yannas IV, Burke JF, Warpehoski M, Stasikelis P, Skrabut EM, Orgill DP, Giard DJ. Prompt, long-term functional replacement of skin. Trans Am Soc Artif Intern Organs 1981;27:19–22.
112. Yannas IV, Burke JF, Orgill DP, Skrabut EM. Wound tissue can utilize a polymeric template to synthesize a functional extension of skin. Science 1982;215:174–176.
113. Yannas IV, Orgill DP, Skrabut EM, Burke JF. Skin regeneration with a bioreplaceable polymeric template. In: Gebelein CC, editor. Polymeric Materials and Artificial Organs. Washington, D.C.: American Chemical Society; 1984. p 191–197.
114. Butler CE, Yannas IV, Compton CC, Correia CA, Orgill DP. Comparison of cultured and uncultured keratinocytes seeded into a collagen-GAG matrix for skin replacements. Br J Plast Surg 1999;52:127–132.
115. Compton CC, Butler CE, Yannas IV, Warland G, Orgill DP. Organized skin structure is regenerated in vivo from collagen-GAG matrices seeded with autologous keratinocytes. J Invest Dermatol 1998;110:908–916.
116. Murphy GF, Orgill DP, Yannas IV. Partial dermal regeneration is induced by biodegradable collagen-glycosaminoglycan grafts. Lab Invest 1990;62:305–313.
117. Chang AS, Yannas IV. Peripheral nerve regeneration. In: Smith B, Adelman G, editors. Encyclopedia of Neuroscience. Boston: Birkhauser; 1992. p 125–126.
118. Chang AS, Yannas IV, Perutz S, Loree H, Sethi RR, Krarup C, Norregaard TV, Zervas NT, Silver J. Electrophysiological study of recovery of peripheral nerves regenerated by a collagen-glycosaminoglycan copolymer matrix. In: Gebelein CC, Dunn RL, editors. Progress in Biomedical Polymers. New York: Plenum Press; 1990. p 107–119.
119. Yannas IV, Orgill DP, Silver J, Norregaard TV, Zervas NT, Schoene WC. Regeneration of sciatic nerve across 15 mm gap by use of a polymeric template. In: Gebelein CC, editor. Advances in Biomedical Polymers. New York: Plenum Publishing Corporation; 1987. p 1–9.
120. Harley BA, Spilker MH, Wu JW, Asano K, Hsu H-P, Spector M, Yannas IV. Optimal degradation rate for collagen chambers used for regeneration of peripheral nerves over long gaps. Cells Tissues Organs 2004;176:153–165.
121. Chamberlain LJ, Yannas IV, Hsu H-P, Strichartz G, Spector M. Collagen-GAG substrate enhances the quality of nerve regeneration through collagen tubes up to level of autograft. Exper Neurol 1998;154:315–329.
122. Chamberlain LJ, Yannas IV, Hsu H-P, Strichartz GR, Spector M. Near terminus axonal structure and function following rat sciatic nerve regeneration through a collagen-GAG matrix in a 10-mm gap. J Neurosci Res 2000;60:666–677.
123. Hsu WC, Spilker MH, Yannas IV, Rubin PA. Inhibition of conjunctival scarring and contraction by a porous collagen-glycosaminoglycan implant. Invest. Ophthalmol Vis Sci 2000;41:2404–2411.
124. Billingham RE, Medawar PB. The technique of free skin grafting in mammals. J Exp Biol 1951;28:385–394.
125. Billingham RE, Medawar PB. Contracture and intussusceptive growth in the healing of expensive wounds in mammalian skin. J Anat 1955;89:114–123.
126. Shapiro F. Cortical bone repair. J Bone Joint Surg 1988;70-A:1067–1081.
127. Lundborg G. Nerve regeneration and repair. Acta Orthop Scand 1987;58:145–169.
128. Lundborg G, Dahlin LB, Danielsen N, Gelberman RH, Longo FM, Powell HC, Varon S. Nerve regeneration in silicone chambers: Influence of gap length and of distal stump components. Exp Neurol 1982;76:361–375.
129. Lundborg G, Dahlin LB, Danielsen N, Johannesson A, Hansson HA, Longo F, Varon S. Nerve regeneration across an extended gap: A neuronobiological view of nerve repair and the possible involvement of neuronotrophic factors. J Hand Surg 1982;7:580–587.
130. Lundborg G, Gelberman RH, Longo FM, Powell HC, Varon S. In vivo regeneration of cut nerves encased in silicone tubes: Growth across a six-millimeter gap. J Neuropathol Exp Neurol 1982;41:412–422.
131. Lundborg G, Longo FM, Varon S. Nerve regeneration model and trophic factors in vivo. Brain Res 1982;232:157–161.
132. Peacock EE. Wound healing and wound care. In: Schwartz SI, Shires GT, Spencer FC, Storer EH, editors. Principles of Surgery. New York: McGraw-Hill; 1984.
133. Yannas IV. Artificial skin and dermal equivalents. In: Bronzio JD, editor. The Biomedical Engineering Handbook. Boca Raton: CRC Press; 2000. p 138-1–138-15.
134. Yannas IV. Regeneration Templates. In: Bronzio JD, editor. The Biomedical Engineering Handbook. Boca Raton: CRC Press; 2000. p 113-1–113-18.
135. Yannas IV, Tobolsky AV. Crosslinking of gelatine by dehydration. Nature (London) 1967;215:509–510.
136. Lee JE, Park JC, Hwang YS, Kim JK, Kim JG, Sub H. Characterization of UV-irradiated dense/porous collagen membranes: morphology, enzymatic degradation, and mechanical properties. Yonsei Med J 2001;42(2): 172–179.
137. Weadock KS, Miller EJ, Bellincampi LD, Zawadsky JP, Dunn MG. Physical crosslinking of collagen fibers: Comparison of ultraviolet irradiation and dehydrothermal treatment. J Biomed Mater Res 1995;29(11): 1373–1379.
138. Weadock KS, Miller EJ, Keuffel EL, Dunn MG. Effect of physical crosslinking methods on collagen-fiber durability in proteolytic solutions. J Biomed Mater Res 1996;32(2): 221–226.
139. Lee CR, Grodzinsky AJ, Spector M. The effects of crosslinking of collagen-glycosaminoglycan scaffolds on compressive stiffness, chondrocyte-mediated contraction, proliferation, biosynthesis. Biomaterials 2001;22:3145–3154.
140. Gibson LJ, Ashby MF. Cellular solids: Structure and properties. Cambridge, U.K.: Cambridge University Press; 1997.
141. Yannas IV. Studies on the biological activity of the dermal regeneration template. Wound Rep Reg 1998;6:518–524.
142. Burkitt HG, Young B, Heath JW. Wheater's Functional Histology. 3rd ed. Edinburgh: Churchill Livingstone; 1993.
143. Farquhar MG. The glomerular basement membrane: A selective macromolecular filter. In: Hay ED, editor. Cell Biology of Extracellular Matrix. New York: 1981. Plenum Press; p 335–378.
144. Vracko R. Basal lamina scaffold: Anatomy and significance for maintenance of orderly tissue structure. Am J Pathol 1974;77:313–346.
145. Woodley DT, Briggaman RA. Re-formation of the epidermal-dermal junction during wound healing. In: Clark RAF, Henson PM, editors. The Molecular and Cellular Biology of Wound Repair. New York: Plenum Press; 1988.
146. Hay ED. Collagen and embryonic development. In: Hay ED, editor. Cell Biology of the Extracellular Matrix. New York: Plenum Press; 1981.
147. Briggaman RA, Wheeler CE. The epidermal-dermal junction. J Invest Dermatol 1975;65:71–84.
148. Carver N, Navsaria HA, Fryer P, Green CJ, Leigh IM. Restoration of basement membrane structure in pigs following keratiocyte autografting. J Plast Surg 1993;46:384–392.

149. Rigal C, Pieraggi M.-T, Vincent C, Prost C, Bouissou H, Serre G. Healing of full-thickness cutaneous wounds in the pig I: Immunohistochemical study of epidermodermal junction regeneration. J Invest Dermatol 1991;96:777–785.
150. Gallico GG. Biological skin substitutes. Clin Plast Surg 1990;17:519–526.
151. Eisinger M, Lee JS, Hefton JM, Darzynkiewicz Z, Chiao JW, DeHarven E. Human epidermal cell cultures: Growth and differentiation in the absence of dermal components or medium supplements. Proc Natl Acad Sci USA 1979;76: 5340–5344.
152. Freeman AE, Igel HJ, Herrman BJ, Kleinfeld KL. Growth and characterization of human skin epithelial cultures. In Vitro 1976;12:352–362.
153. Lillie JH, MacCallum DK, Jepsen A. Growth of stratified squamous epithelium on reconstituted extracellular matrices: Long-term culture. J Invest Dermatol 1988; 90:100–109.
154. Woodley DT, Peterson HD, Herzog SR, Stricklin GP, Burgeson RE, Briggaman RA, Cronce DJ, O'Keefe EJ. Burn wounds resurfaced by cultured epidermal autografts show abnormal reconstitution of anchoring fibrils. JAMA 1988; 259:2566–2571.
155. Hull BE, Sher SE, Rosen S, Church D, Bell E. Structural integration of skin equivalents grafted to Lewis and Sprague-Dawley rats. J Invest Dermatol 1983;81:429–436.
156. Cumpstone MB, Kennedy AH, Harmon CS, Potts RO. The water permeability of primary mouse keratinocyte cultures grown at the air-liquid interface. J Invest Dermatol 1989;92:598–600.
157. Nolte CJM, Oleson MA, Bilbo PR, Parenteau NL. Development of a stratum corneum and barrier function in an organotypic skin culture. Arch Dermatol Res 1993;285: 466–474.
158. Nolte CJM, Oleson MA, Hansbrough JF, Morgan J, Greenleaf G, Wilkins L. Ultrastructural features of composite skin cultures grafted onto athymic mice. J Anat 1994;185:325–333.
159. Bosca AR, Tinois E, Faure M, Kanitakis J, Roche P, Thivolet J. Epithelial differentiation of human skin equivalents after grafting onto nude mice. J Invest Dermatol 1988;91:136–141.
160. Wassermann D, Sclotterer M, Toulon A, Cazalet C, Marien M, Cherruau B, Jaffray P. Preliminary clinical studies of a biological skin equivalent in burned patients. Burns 1988;14:326–330.
161. Eaglstein WH, Iriondo M, Laszlo K. A composite skin substitute (Graft-skin) for surgical wounds. Dermatol Surg 1995;21:839–843.
162. Falanga V, Margolis D, Alvarez O, Auletta M, Maggiacomo F, Altman M, Jensen J, Saboinski M, Hardin-Young J. Rapid healing of venous ulcers and lack of clinical rejection with an alogeneic cultured human skin equivalent. Arch Dermatol 1998;134:293–300.
163. Sabolinski ML, Alvarez O, Aulettea M, Mulder G, Parenteau NL. Cultured skin as a 'smart material' for healing wounds experience in venous ulcers. Biomaterials 1996; 17:311–320.
164. Guerret S, Govignon E, Hartmann DJ, Ronfard V. Long-term remodeling of a bilayered living human skin equivalent (Apligraf) grafted onto nude mice: Immunolocalization of human cells and characterization of extracellular matrix. Wound Rep Reg 2003;11:35–45.
165. O'Brien FJ, Harley BA, Yannas IV, Gibson LJ. Influence of freezing rate on pore structure in freeze-dried collagen-GAG scaffolds. Biomaterials 2004;25:1077–1086.
166. Kurz W, Fisher DJ. Fundamentals of Solidification. Switzerland: Transtech Publications; 1989.
167. O'Brien FJ, Harley BA, Yannas IV, Gibson LJ. The effect of pore size on cell adhesion in collagen-GAG scaffolds. Biomaterials In press, 2004.
168. Heimbach D, Luterman A, Burke J, Cram A, Herndon D, Hunt J, Jordan M, McManus W, Solem L, Warden G. Artificial dermis for major burns. Ann Surg 1988;208:313–320.
169. Stern R, McPherson M, Longaker MT. Histologic study of artificial skin used in the treatment of full-thickness thermal injury. J Burn Care Rehab 1990;11:7–13.
170. Michaeli D, McPherson M. Immunologic study of artificial skin used in the treatment of thermal injuries. J Burn Care Rehab 1990;11:21–26.
171. Burke JF. Observations on the development and clinical use of artificial skin: An attempt to employ regeneration rather than scar formation in wound healing. Jpn J Surg 1987;17:431–438.
172. Tompkins RG, Hilton JF, Burke JF, Scoenfeld DA, Hegarty MT, Bondoc CC, Quinby WC, Jr., Behringer GE, Ackroyd FW. Increased survival after massive thermal injuries in adults: Preliminary report using artificial skin. Child Care Med 1989;17:734–740.
173. Sheridan RL, Hegarty M, Tompkins RG, Burke JF. Artificial skin in massive burns: Results to ten years. Eur J Plast Surg 1994;17:91–93.
174. Besner GE, Klamar JE. Integra Artificial Skin as a useful adjunct in the treatment of purpura fulminans. J Burn Care Rehab 1998;19:324–329.
175. Spence RJ. The use of Integra for contracture release. Burn Care Rehabil. 1998;19:S173.
176. Lorenz C, Petracic A, Hohl H-P, Wessel L, Waag K-L. Early wound closure and early reconstruction: Experience with a dermal substitute in a child with 60 percent surface area burn. Burns 1997;23:505–508.
177. Pandya AN, Woodward B, Parkhouse N. The use of cultured autologous keratinocytes with Integra in the resurfacing of acute burns. Plast Reconstr Surg 1998;102:825–828.
178. Lattari V, Jones LM, Varcelotti JR, Latenser BA, Sherman HF, Barrette RR. The use of permanent dermal allograft in full-thickness burns of the hand and foot: a report of three cases. J Burn Care Rehabil 1997;18:147–155.
179. Tsai CC, Lin SD, Lai CS, Lin TM. The use of composite acellular allodermis-ultrathin autograft on joint area in major burn patients—one year follow-up. Kaohsiung J Med Sci 1999;15:651–658.

See also ENGINEERED TISSUE; SKIN, BIOMECHANICS OF; SKIN SUBSTITUTE FOR BURNS, BIOACTIVE.

SKIN, BIOMECHANICS OF

YAN CHEN
BRIAN L. DAVIS
Department of Biomedical Engineering/ND20 Lerner Research Institute
The Cleveland Clinic Foundation
Cleveland, Ohio

INTRODUCTION

Skin, the body's largest organ, accounts for about 16% of an average adult's body weight and occupies a surface area of

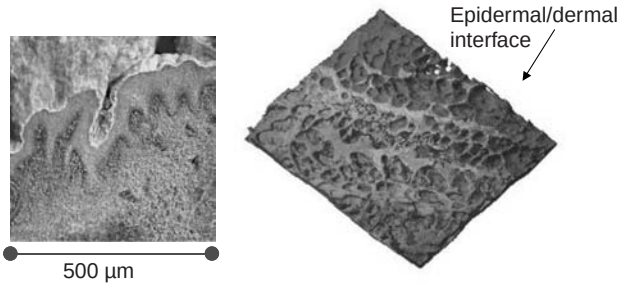


Figure 1. Layers of skin (left) and view of rete ridges (right) (1).

approximately 2 m^2. It has multiple essential functions: It provides the inner organs and vasculature with a protective barrier against injury or invasion by microorganisms; it prevents water loss; its sensory nerves can sense contact, pain, pressure, heat, and cold; and it plays an important role in maintaining normal body temperature via the function of sweat glands.

Structurally, the skin consists of two principal layers—epidermis and dermis—and various sublayers (Fig. 1). The epidermis is the thinner outer layer, composed of stratified squamous epithelium organized in four or five layers. The main cells in the epidermis are keratinocytes, which are generated by continuous divisions of stem cells that form the basal layer of the epithelium. As keratinocytes migrate to the surface of the epidermis, they grow and differentiate, synthesizing large amounts of a cytoskeletal protein called keratin. This protein builds into 10 nm filaments that gradually come to occupy 80% of the cell volume.

The epidermis is 0.07–0.12 mm in thickness over most of the body surface. It has no blood vessels and thus relies on the dermis for nutrients. The dermis is a supportive layer beneath the epidermis, composed of fibroblasts, fibrocytes, collagen, elastic fibers, glycosaminoglycan (GAG) matrix, blood vessels, and nerves. It is about 0.5–2.5 mm thick. The dermis has two distinct layers: papillary dermis (about 20% the overall thickness) and reticular dermis. The papillary layer is a loose connective tissue, containing large blood vessels, interlacing elastin fibers and bundles of collagen fibers. In contrast, the reticular dermis is a dense, irregularly arranged connective tissue containing interlacing bundles of type I collagen fibers and coarse elastic fibers.

The interface between the dermis and the epidermis varies from region to region. In thick skin, such as that found on the plantar aspect of the foot, the epidermis–dermis interface has the configuration of rete ridges (downgrowths of epidermis and dermis). The upper dermis exhibits a pattern of primary ridges separated by deep primary grooves (Fig. 1). Rows of conical dermal papillae project upward into the conforming concavities between the ridges in the deep surface of the epidermis. Tethering fibers connecting dermis and epidermis to the base membrane keep the dermis and epidermis from separating. In thin skin (e.g., facial skin), the dermo-epidermal junction is much simpler. The dermal papillae in thin skin are shorter, wider, and fewer and not arranged in the pattern of ridges and grooves as observed in thick skin.

Beneath the dermis is a subcutaneous layer (also called the hypodermis) consisting of areolar and adipose tissues. This layer is sometimes considered a third layer of skin. Collagen fibers from the dermis extend down into the superficial fascia and anchor the skin to the subcutaneous layer, firmly attaching it to underlying tissues and organs. The superficial fascia provides the skin's loose flexible connection with the other internal soft tissues, whereas the upper layers (epidermis and dermis) protect skin from injuries.

MECHANICAL PROPERTIES OF SKIN

Microstructure of Skin

The mechanical properties of skin have long interested dermatologists and bioengineers (2–11). Studies of the mechanical properties of skin could provide objective information related to skin and the changes it undergoes with age and disease. For instance, knowledge of the skin's mechanical properties is of potential use in assessing conditions of connective tissue disease, skin aging, wound healing, and scarring.

Like any other material, the mechanical properties of skin are determined by those of its structural constituents. Major structural components of skin are elastin fibers, collagen fibers, and the ground substance. The individual mechanical and structural properties of collagen fibers, elastin fibers, and their interaction with the ground substances in skin form the basis of mechanical properties of skin.

Collagen. Collagen, the body's most abundant and common protein, represents approximately 72% of the dry weight of dermal tissue. There are now more than 19 known collagen types, each of which is a genetically distinct product. The collagen of adult dermis is mainly type I (85–90%), with 8–11% of type III and 2–4% of type V collagen. In all connective tissues, collagen exists as fibers made up of crimped fibrils. Depending on the tissue, a collagen fibril varies from 50 nm to 300 nm in diameter. Collagen fibers in skin are randomly oriented in layers or lamellae, which give skin large extensibility and resiliency under stress. This arrangement contrasts with that found in the tendon and ligament, where the parallel alignment of collagen fibers gives these tissues higher values of tensile strength requirement.

Collagen fibrils are stabilized and strengthened by the formation of covalent cross-links. These cross-links are formed by lysine and hydroxylysine residues. The intermolecular cross-links are formed by the joining of two hydroxylysine residues and one lysine residue. The cross-links are formed between residues near the amino terminus of one collagen molecule and the carboxyl terminus of another.

There is evidence that in the presence of hyperglycemia (such as in diabetes), some proteins, such as collagen, undergo nonenzymatic glycation (12–15). This process modifies the structure of collagen and has a direct effect on the mechanical properties of collagen, resulting in increased mechanical stiffness and decreased flexibility

of these collagen-containing tissues (16–18). Increased collagen cross-linking due to a buildup of advanced glycation endproducts (AGEs) is also believed to be a major contributor to many complications of diabetes (18,19). AGE content has been shown to be four times higher in the collagen of diabetic versus nondiabetic patients (20). Besides being more rigid, highly crosslinked tissues are more resistant to enzymatic digestion by collagenase (18,21).

The strength of collagen fiber is similar to that of skin, which suggests that collagen fibers are the dominant load-bearing material at high strains. However, the extension of collagen at failure is found to be only 10% that of skin. Studies of the geometry of collagen fibers show that the apparently randomly coiled collagen fibers do not carry any part of the applied stress during the initial elastic deformation of skin. As fibers become oriented and straightened out in the stress direction, they start to carry stress, and the high stiffness of the collagen prevents any further large strain.

Elastin. Elastin accounts for 4% of the dry weight of dermal tissue. Elastin polypeptide molecules are cross-linked together to form rubberlike, elastic fibers. Each elastin molecule uncoils into a more extended conformation when the fiber is stretched and will recoil spontaneously as soon as the stretching force is relaxed. Elastin fibers are highly extensible, and their extension is reversible even at high strains. They exhibit linear elasticity with low stiffness up to about 200% elongation followed by a short region where the stiffness increases rapidly until failure. The loading and unloading paths of elastin fibers do not show significant hysteresis. Elastin fibers are characterized as low-modulus elastic material. Analyzing skin disks from cadavers mechanically and histologically, Dick (22) concluded that the loss of elastin fibers in older persons resulted in a loss of skin resistance to deformation at low stress. King and Lawton (23) investigated the behavior of elastin-rich tissues in terms of elastomer theory and found good agreement with experimental data. When other structural components such as collagen and ground substance are present, however, the elastomer theory is no longer applicable. Daly (24) found that the elastic recovery of skin is completely lost when elastin is removed. Thus, he concluded that the elastic behavior of skin during the initial low strain region is due to the small amount of elastin in dermis.

Ground Substance. Ground substance is the intercellular, non-fibril material in connective tissue. It is composed of tissue (extracellular) fluid, amorphous component proteoglycans, and GAGs. Ground substance accounts for 20% of the dry weight of dermal tissue.

Ground substance contributes to the time-dependent properties of skin. Minns et al. (25) performed stress-strain and relaxation studies on human tendon, aorta, and bovine ligamentum nuchae after removing the ground substance with an enzyme or chelating agents. They noted a decrease in stress level, stiffness, stress relaxation, hysteresis, and other time-dependent effects in all three tissues.

Much is known about the structure of collagen, elastin, and proteoglycans at the molecular level. However, less is known about the higher level of structure. For example, how the collagen fibers and elastin fibers are connected to each other in the network of skin and how this might be affected by disease, such as diabetes, are not known.

Skin Mechanics

Most engineering materials are elastic material, which can be described, for small strains, by Hooke's law of linear elasticity: Stress σ is linearly proportional to strain ε. Elastic materials show time-independent material behavior, that is, they deform instantly when external loading is applied and return back to their original state immediately when the applied loads are removed.

Unlike engineering materials, the responses of most biological tissues to mechanical loading are complex. Biological tissues have self-adapting and self-repairing characteristics. That is, they can alter their mechanical properties to adapt to changing mechanical demands and can repair themselves. The mechanical properties of biological tissues tend to change with age and some connective tissue diseases. Most biological tissues are composite materials with nonhomogeneous and anisotropic properties, and almost all biological tissues are viscoelastic in nature.

Viscoelasticity. The response of viscoelastic materials not only depends on the strain level applied but also on the time and rate at which the external loading are applied. Skin is such a composite material, made up of collagen fibers, elastin fibers, keratin, and ground substance. Collagen fibers and elastin fibers explain the solid behavior of skin; ground substance contributes to the fluid behavior of skin. The mechanical properties of skin are determined not only by the individual component properties, but also by the geometry configuration and interaction of these components.

As a viscoelastic material, skin has the characteristic of creep, stress relaxation, hysteresis, and precondition.

- Creep: When skin is suddenly loaded with a constant tension, its lengthening velocity decreases against time until equilibrium.
- Stress relaxation: When skin is suddenly extended and maintained at its new length, the stress gradually decreases over time.
- Under cyclic loading, the stress-strain curve shows two distinct paths corresponding to the loading and unloading trajectories. This phenomenon is called hysteresis. With repeated loading cycles, the load-deformation curves shift to the right in a load-elongation diagram and the hysteretic effects decrease. By repeated cycling, eventually a steady state is reached at which no further internal changes in tissue structure will occur unless the cycling routine is changed. At this point, tissue is defined as preconditioned.

Nonlinear Elasticity. The elasticity of skin is strongly nonlinear (Fig. 2). In the initial low-strain region, elongation of skin occurs without appreciable force (a). In the

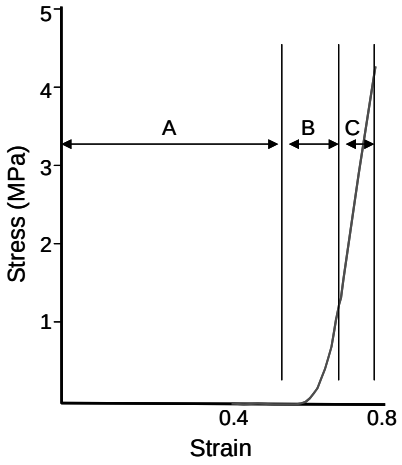


Figure 2. Typical stress-strain curve of skin under incremental loading (4).

mid-strain region, the curve increases in a roughly exponential manner (b) until, at the high-strain region, it becomes almost a straight line (c) immediately before rupture occurs. The accepted explanation of the curve is that the initial deformation (Fig. 2a) is due to deformation of the delicate elastin network; the second part of the curve (b) is due to a gradual straightening of the randomly oriented collagen fibers; the third part of the curve (c) is because most collagen fibers are elongated in the direction of the stress. If the initial part of this curve is magnified, a straight-line approximation will allow the calculation of elastic properties such as Young's modulus. From a functional part of view, the first parts of the curve are more important because they correspond to the physiological range in which the skin normally functions.

Anisotropy. Like most composite materials, skin is anisotropic. This mechanical behavior of human skin can be demonstrated by stretching skin in different directions or by a biaxial loading test (26). Results from tests such as these show directional variance related to lines of tension or cleavage (known as Langer's lines) within the skin that are characteristic for each part of the body. They are named after Austro-Hungarian anatomist Karl Langer (1819–1887), the first to systematically investigate the configuration of the tension on cuts in the human body. Microscopic examination of sections of skin along and across Langer's lines shows a preferential orientation of the collagen fibers of skin. The orientation seems to be at an angle slightly less than 45° from the direction of the lines. Ridge and Wright (27) performed a series of experiments with an extensometer in which they stretched pieces of skin with and against the normal skin tension lines. They could show a clear difference in the directional mechanical properties in the skin, which corresponded to Langer's lines.

Measurement of Skin Mechanical Properties

Information about the mechanical properties of skin falls under three broad categories: strength values (e.g., breaking strain), time-dependent values (e.g., creep, relaxation), and non-time-dependent values (e.g., elasticity). The measurement of mechanical properties of skin can be performed either *in vitro* or *in vivo*. Results from *in vitro* tests are estimates of properties that are assumed to have an important influence on function *in vivo* and can provide information about the basic mechanical properties of strength, elasticity, or density inherent in the architecture of skin. However, *in vitro* studies are limited in that the characteristics of isolated skin are modified. Dissected skin samples no longer have the same physiological properties as *in vivo* skin, such as those of internal tension, hydration, and vascularization. *In vivo* tests may provide information on the function and the kinetics of change in mechanical properties.

In Vitro Studies. In testing the strength characteristics of biological material, Yamada (28) suggested that the material should be in a mechanically stabilized state in which constant values are obtained. Such a state can be achieved by putting the biological material in a physiological saline solution and storing it in a refrigerator overnight or longer. Yamada noted that the duration of the mechanically stabilized state for skin is 3 days.

There have been many investigations of the mechanical properties of both animal and human skin. Most have used uniaxial test procedures. Many of these early experiments were flawed due to unsophisticated equipment and poor control of such variables as temperature and humidity. Ridge and Wright (1964) developed a skin extensometer, which stretched 1 cm × 0.4 cm strips of skin at a constant rate of 0.2 in. per min. They characterized the resulting stress-strain curves with the equation:

$$e = c + kL^b$$

where e is extension; L is load; and c, b, and k are constants. The authors felt that constant b reflected a specific property of the collagen fibers and constant k represented the condition of the fiber network, which was related to the length and area of the fibers.

Daly (3) further refined uniaxial testing. He first appreciated that for consistent results the specimens must be tested at a constant temperature, humidity, and pH and with a constant strain rate. In addition to measuring stress-strain curves and Poisson's ratio, he performed extensive experiments documenting stress relaxation and creep in human cadaver skin.

An accurate biomechanical description of skin requires constitutive equations that characterize soft tissues in three dimensions. For practical analysis, skin can be considered an incompressible solid material. It is possible to determine three-dimensional mechanical properties from two-dimensional tests for an incompressible solid. Lanir and Fung (29) developed a two-dimensional experimental system for biaxial testing. Each specimen of skin is hooked along its edges by many small staples. Each hook connects by means of a small thread to a force-distributing platform, where tension can be individually adjusted. The investigators used this device to measure the biomechanics of rabbit skin. In their study, the deformed skin always returned to its predeformed state so long as no surface dimension had been allowed to decrease below its initial

unstressed value. This study is obviously different from uniaxial tests, in which the skin essentially always diminishes in lateral directions and usually does not return to its prestress state. Biaxial tests showed that:

1. The stress-strain relationship was extremely nonlinear.
2. Hysteresis was present at all strain levels.
3. Stress-strain relationships were minimally dependent on the strain rate.
4. The relaxation curve did not terminate at the origin and returned to it only after a long period of relaxation.

Reihsner et al. (9) studied the two-dimensional biomechanical behavior of human skin samples from different anatomical sites. Using the *in vivo* geometry of the specimen as reference, a set of incremental strains was applied to the skin. After stress relaxation was completed, the final values of stresses were recorded and related to the incremental strains. Six independent elastic constants were determined. The average deviation of the orientation of maximum principal stress from the direction of the Langer cleavage lines was in the range of angles of -10° to $+10^\circ$. The effect of a directional dependency was most pronounced in skin samples from patella and abdomen. Across a range of ages, Reihsner et al. observed no uniform trend in the principal stresses except that the oldest subject showed the highest principal stresses. The strain necessary to restore the *in vivo* configuration decreased with age in both principal axes of strain. Reihsner et al. suggested that the regional differences in anisotropic behavior could be explained by the different interdigitation between the epidermis and the dermis, the polydispersity of fiber diameter distribution, and the differences in fiber bundle orientation.

Daly (4) conducted tensile tests on specimens oriented at right angles to each other and found that anisotropy is related only to the magnitude of the initial large extension region and the initial low stiffness and the final high stiffness is not orientation dependent. He concluded that the geometry of elastin is not entirely random, as the elastin network determines the initial configuration of the collagen.

***In Vivo* Studies.** *In vivo* tests can be classified as static or dynamic. In static tests, a single modulating stimulus is applied and some resulting change is measured. In dynamic tests, a cyclical stimulus is applied and the initial adapting and final steady-state reactions are monitored. Dynamic tests can provide more information than static tests. However, there is a potentially unlimited combination of frequencies, magnitude, and attachment area. The effects of these parameters on the skin may not be fully understood and may result in difficulties in interpretation of data.

Several techniques have been used to study the mechanical properties of human skin *in vivo*:

1. Tonometric measurements that evaluate the ability of the skin to withstand vertical forces of extension (8).
2. Traction, which applies a linear displacement in the horizontal plane of the skin.
3. Indentation.
4. Torsion, which applies a torque to the skin. The response of skin to shear force can be measured with this device. A device using this technique, the "dermal torque meter," has been commercialized.
5. Suction, which involves placing a suction cup or cylinder on the skin surface and applying a negative pressure to raise a dome of skin. The pressure and the height of the dome are used to calculate elasticity parameters. Two devices based on suction, the Cutometer SEM 474 (Courage and Khazaka, Cologne, Germany) and the Dermaflex A (Cortex Technology, Copenhagen, Denmark), have been commercialized.
6. Elastic wave propagation systems involve measurements of shear wave velocities and the rate at which they are dissipated during their passage in the skin. One study using this technique suggests a probable reduction in skin water content with age (10).
7. Mechanical impedance method. Skin is made to vibrate by a probe over a rectangular frequency spectrum of 20 to 500 Hz after applying a standard preload. One device using this method (DPE, Cotas Computer Technology, Denmark) has been commercialized.

When studying the mechanical properties of skin *in vivo* by such methods, consideration should be given to the contribution of different tissue layers (i.e., dermis and subcutaneous fat). Diridollou et al. (11) reported a device called the "echorheometer," comprising a suction system with an ultrasound scanner. Using this device, the behavior of the dermis and subcutaneous fat under suction was investigated. They observed that there is a certain amount of infiltration of fluid into the tissue under suction. Upon returning to atmospheric pressure, excessive fluid remained in the tissue, which might explain the hysteresis of skin. The authors found that resistance of skin to the applied vertical stress is essentially due to the dermis rather than the subcutaneous fat. The relative contribution of each is not easily evaluated. With high suction pressures applied to large surface areas or low suction pressures to smaller surface areas, the effects on skin cannot be isolated from those of the subcutaneous fat.

Biomechanics of Diabetic Skin Ulceration

The mechanics of skin breakdown under the foot incorporate several aspects described in this article: biaxial skin properties and applied stresses, differing stiffness values within each skin layer, effects of subcutaneous fat, collagen cross-linking and nonlinear, and viscoelastic properties of connective tissue. Although it is recognized that neuropathy and ischemia are primary predisposing factors in the formation of diabetic foot ulcers, an initiating factor, such as physical or mechanical stress, is required for an ulcer to develop (30). For patients with neuropathy, this

mechanical stress need not be excessive, only repetitive, for ulcers to develop.

Brand (31) conducted a series of experiments on rat footpads in which lightly anesthetized rats had one foot placed in a machine that applied repeated intermittent stress to the footpad to simulate a human walking 7 miles. The footpads became swollen after 2 to 3 days, blistered by 1 week, and ulcerated and necrotic after 10 days. In neuropathic animals, the number of cycles that could be tolerated was reduced. Histologic studies on the rats showed inflammation of the footpads after a few days, with small foci of necrosis in areas with large numbers of inflammatory cells. Brand postulated that a similar mechanism occurs in the formation of neuropathic foot ulcers, with a sterile inflammatory autolysis occurring in the subcutaneous tissue due to the repetitive moderate stresses imposed by walking.

The stresses that the foot experiences during walking are vastly different from those encountered by any other skin surface. At the foot–ground interface, a complex combination of shear and pressure exists (32) that could explain why some patients get foot ulcers and others do not. Mathematical modeling of (1) the effects of stresses at the skin/fat interface and (2) the influence of more cross-linked collagen (33) showed that small increases in shear stresses and/or increased skin stiffness had a profound effect on the stresses at the interface between dermis and underlying soft tissue (Fig. 3).

The effects of increased dermal interface stresses have also been assessed in an experimental study (1). In total, 21 diabetic (68.0 ± 2.4 years) and 17 nondiabetic (74.6 ± 3.8 years) amputated foot specimens were tested. The procedure involved peeling strips of skin off the plantar surface of each foot specimen. Each strip was approximately 8 mm wide (in the antero-posterior direction) and 20 mm long (in the medio-lateral direction). While tension was applied via a Materials Testing System (MTS), force and displacement of the loading head were recorded at a sampling rate of 60 Hz. Data across all 38 specimens are shown in Fig. 4.

Within the nondiabetic group, analysis of variance showed that the required forces for removing skin at different sites were significantly different ($p < 0.001$), with the skin–fat interface being significantly stronger

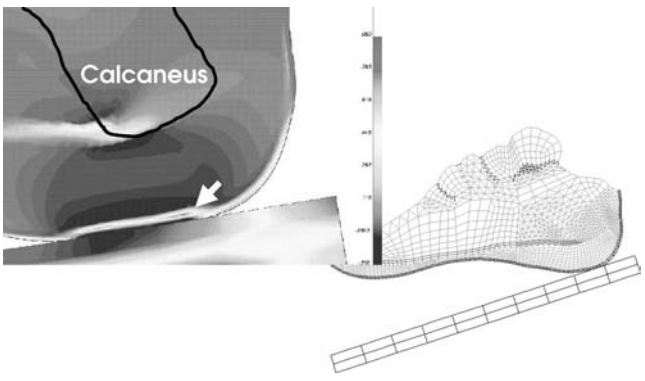


Figure 3. Through finite element modeling of the hindfoot, the predicted effect of diabetes-induced skin stiffening is to increase the stresses at the skin–fat interface by 100% (33).

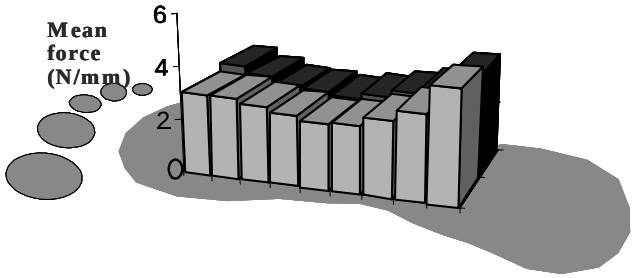


Figure 4. Combined diabetic and control data showing the forces required to remove skin strips off the plantar surface of amputated foot specimens. These forces are significantly higher ($p < 0.05$) in the forefoot and hindfoot regions compared with the midfoot area (1).

in both the metatarsal head and heel regions than in the midfoot area. These regions also withstand the largest weight-bearing loads during normal locomotion. Neuropathic diabetic patients (who typically have higher plantar pressures) also showed higher strengths at the skin–fat interface than age-matched nondiabetic subjects (1). Every strip (e.g., first strip on medial side, or third strip on lateral side, etc.) for the diabetic group required larger forces to pull it off the underlying soft tissue ($p < 0.05$). Across all skin strips, diabetic specimens required a 20% greater removal force.

SUMMARY

Skin has multiple essential functions ranging from protection against injury or invasion by microorganisms, to a barrier to water loss, to maintaining normal body temperature. It consists primarily of two principal layers—epidermis and dermis—and various sublayers. From a mechanical point of view, these layers have stiffness values that can differ by three orders of magnitude! The area of skin that withstands the greatest daily history of loading—that under the foot—has an interface between the dermis and the epidermis characterized by rete ridges. The precise mechanical function of these ridges is not yet clear. What is known is that this interface is profoundly affected by the magnitudes of shear and pressure loading that plantar soft tissue is required to withstand. Areas that have higher loading (e.g., heel or metatarsal head region) have a dermis that is more tightly bonded to the subcutaneous soft tissue than other regions. In this respect, skin seems to follow the "form follows function" law (or "Wolff's Law) that is usually used to describe bone's adaptation to applied loading. When these loads become extreme (e.g., greater than 1 MPa applied pressure), even the short weight-bearing intervals during gait are sufficient to lead to skin breakdown. The precise site for the initiation of failure (e.g., which layer of skin fails first) has yet to be determined. As skin breakdown under the foot is a clinical problem with significant risks for the patient, it is likely that research into this problem will continue until skin failure mechanisms are better understood. These may then serve as the basis for better prevention or treatment strategies.

BIBLIOGRAPHY

1. Chen Y. The influence of diabetes on mechanical properties of skin. Unpublished doctoral dissertation. Cleveland, OH: Cleveland State University; April 2003.
2. Ridge MD, Wright V. The description of skin stiffness. J Biorheol 1964;10:139–155.
3. Daly CH. The biomechanical characteristics of human skin. Ph.D. Thesis, University of Strathcylde, Glasgow, Scotland, 1966.
4. Daly CH. Biomechanical properties of dermis. J Invest Dermatol 1982;79(Suppl 1):17s–20s.
5. Danielson DA. Human skin as an elastic membrane. J Biomech 1973;6:539–546.
6. Lanir Y, Fung YC. Two-dimensional mechanical properties of rabbit skin. II. Experimental Results. J Biomech 1974;7: 171–182.
7. Lanir Y. Structural theory for the homogeneous biaxial stress-strain relationships in flat collagenous tissues. J Biomech 1979;12:423.
8. Warren R, Gartstein V, Kligman AM, Montagna W, Allendorf RA, Ridder GM. Age, sunlight, and facial skin: A histologic and quantitative study. J Am Acad Dermatol 1991;25:751–760.
9. Reihsner R, Balogh B, Menzel EJ. Two-dimensional elastic properties of human skin in terms of an incremental model at the in vivo configuration. Med Eng Phys 1995;17:304–313.
10. Potts RO, Buras EM Jr, Chrisman DA Jr. Changes with age in the moisture content of human skin. J Invest Dermatol 1984;82:97–100.
11. Diridollou S, Berson M, Vabre V, Black D, Karlsson B, Auriol F, Gregoire JM, Yvon C, Vaillant L, Gall Y, Patat F. An in vivo method for measuring the mechanical properties of the skin using ultrasound. Ultrasound Med Biol 1998;24:215–224.
12. Brownlee M, Vlassara H, Cerami A. Nonenzymatic glycosylation and the pathogenesis of diabetic complications. Ann Intern Med 1984;101(4):527–537.
13. Buckingham B, Reiser KM. Relationship between the content of lysyl oxidase-dependent cross-links in skin collagen, nonenzymatic glycosylation, and long-term complications in Type I diabetes mellitus. J Clin Invest 1990;86:1046–1054.
14. Delbridge L, Ellis CS, Lequesne LP. Non-enzymatic glycosylation of keratin from the diabetic foot. Br J Surg 1983;70: 305.
15. Monnier VM, Kohn RR, Cerami A. Accelerated age-related browning of human collagen in diabetes mellitus. Proc Natl Acad Sci USA 1984;81:583–587.
16. Cerami A, Vlassara H, Brownlee M. Role of advanced glycosylation products in complications of diabetes. Diabetes Care 1988;11(Suppl 1):73–79.
17. Hamlin CR, Kohn RR, Luschin JJ. Apparent accelerated aging of human collagen in diabetes mellitus. Diabetes 1975;24:902–904.
18. Reiser KM. Nonenzymatic glycation of collagen in aging and diabetes. Proc Soc Exp Biol Med 1991;196:17–29.
19. Kennedy L, Baynes JW. Non-enzymatic glycosylation and the chronic complications of diabetes: An overview. Diabetologia 1984;26:93–98.
20. Makita Z, Radoff S, Rayfield EJ, Yang Z, Skolnik E, Delaney V, Friedman EA, Cerami A, Vlassara H. Advanced glycosylation end products in patients with diabetic nephropathy. N Engl J Med 1991;325(12):836–842.
21. Elkeles RS, Wolfe JHN. ABC of vascular diseases. The diabetic foot. BMJ 1991;303:1053–1055.
22. Dick JC. The tension and resistance to stretching of human skin and other membranes. J Physiol 1951;112:102–113.
23. King AL, Lawton RW. Elasticity of body tissues. In: Glasser O, editor. Medical Physics, Vol. 3. Chicago: Yearbook Publishers; 1960. 234–247.
24. Daly CH. The role of elastin in the mechanical behavior of human skin. Proc. 8th Intl. Conf. Med. Biol. Engr.. Chicago, IL: 1969: 18–27.
25. Minns RJ, Soden PD, Jackson DS. The role of the fibrous components and the ground substance in the mechanical properties of biologic tissues: A preliminary investigation. J Biomech 1973;6 153–165.
26. Tong P, Fung YC. The stress-strain relationships for the skin. J Biomech 1976;9:649–657.
27. Ridge MD, Wright V. The directional effects of skin. A bioengineering study of skin with particular reference to Langer's lines. J Invest Dermatol 1966;64:341–346.
28. Yamada H. Strength of Biological Materials. Baltimore, MD: Williams & Wilkins; 1970.
29. Lanir Y, Fung YC. Two-dimensional mechanical properties of rabbit skin. I. Experimental System. J Biomech 1974;7 29–34.
30. Laing P. The development and complications of diabetic foot ulcers. Am J Surg 1998;176(Suppl 2A):11S–19S.
31. Brand PW. Pathomechanics of diabetic (neurotrophic) ulcer and its conservative management. In: Bergan JJ, Yao JST, editors. Gangrene and Severe Ischaemia of the Lower Extremities. New York: Grune & Stratton; 1978. p 185–189.
32. Davis BL. Foot ulceration: Hypotheses concerning shear and vertical forces acting on adjacent regions of skin. Med Hypoth 1993;40:44–47.
33. Thompson DL. Finite element modeling of the diabetic foot. Unpublished Masters Dissertation, Ohio State University, 1998.

See also SKIN SUBSTITUTE FOR BURNS, BIOACTIVE; SKIN TISSUE ENGINEERING FOR REGENERATION; TACTILE STIMULATION; ULTRAVIOLET RADIATION IN MEDICINE.

SLEEP LABORATORY

RICHARD K. BOGAN
SHAWN D. YOUNGSTEDT
University of South Carolina
Columbia, South Carolina

INTRODUCTION

Sleep is a fundamental, homeostatic process necessary for human existence and quality of life. Sleep disorders account for impairment of alertness (1); cognitive function (2), especially short-term memory and divided-attention tasks (3); mood (4); work productivity (5); and driving ability (and hence, an increase in automobile accidents). Obstructive sleep apnea has been clearly linked to cardiovascular disease.

There are > 80 sleep disorders, which generally increase in prevalence with age. Indeed, > 50% of adults over age 60 years suffer from some sleep-related complaint (6), including daytime fatigue, difficulty initiating or maintaining sleep, snoring, obstructive apnea, insomnia, restless legs syndrome, narcolepsy, or circadian rhythm disorders. Most disorders have attendant morbidity issues. For example, patients with sleep apnea have an increased risk of cardiovascular disease, stroke, and diabetes. Patients with protracted insomnia have

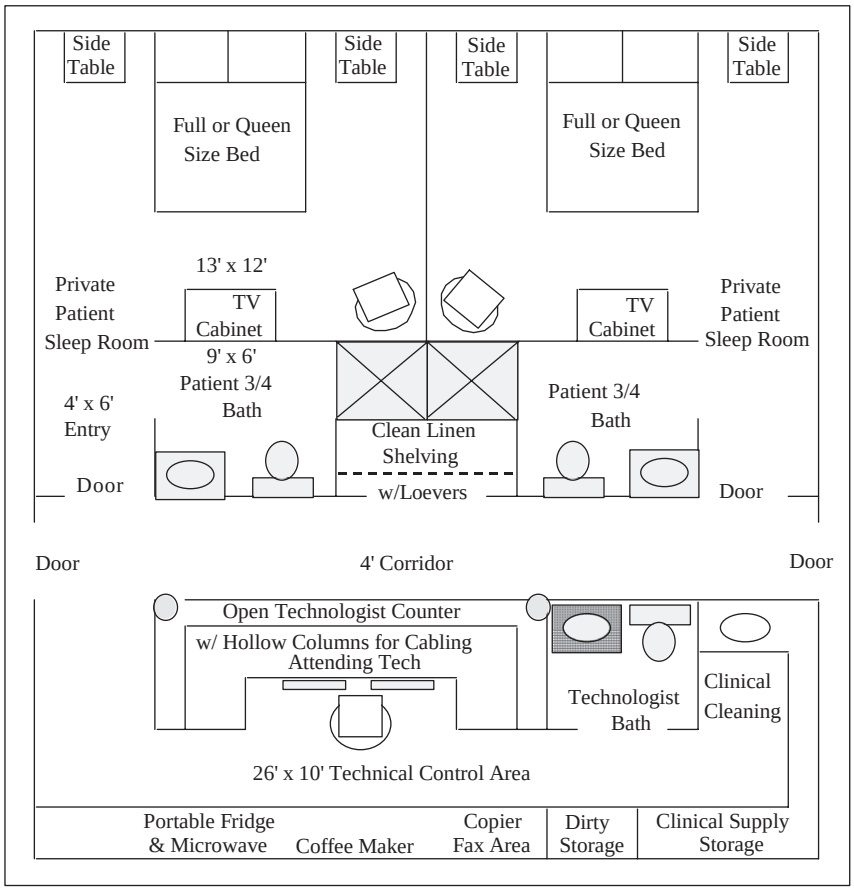


Figure 1. Sample sleep lab layout (26 ft × 32 ft 832 ft^2 for a two-bed sleep lab).

an increased risk of depression, anxiety, and substance abuse. Many sleep disturbances are secondary to other medical problems, such as nocturia, pain associated with arthritis, and cardiopulmonary disorders.

POLYSOMNOGRAPHIC RECORDING

The traditional evaluation of normal sleep and sleep disorders incorporates the gold standard, nocturnal polysomnography, performed in a sleep clinic or laboratory (Fig. 1). Growth of laboratories has occurred in the last 30 years to 2515 estimated labs, generally located in hospitals. Most current laboratories require a night tech, a scoring tech, a clinician, or a scientist to acquire and process sleep data. Some laboratories are freestanding or exist in private clinics. Contract service entities also provide outsourcing of sleep diagnostics.

The sleep laboratory classically consists of diagnostic bedrooms, accompanied by a central technical area for collection of data, observation of patients, and data processing. Ancillary areas include a business office, exam rooms, a lounge area, a break area, and file storage, much like a clinical practice. Sleep diagnostics for the sleep bedroom are focused primarily on a monitoring EEG to determine sleep states, with expanded measurements to allow quantitative assessment of EEG (see Table 1). The sleep laboratory environment attempts to achieve a bedroom environment, with the patient interface allowing minimally invasive techniques.

Electroencephalography

The EEG is the fundamental measurement of polysomnography and consists of application of electrodes according to the International 10–20 electrode placement system (7) (Fig. 2). The skin interface is established through cleansing and removal of dead skin layers, with electrode application, typically using an electrode cup with a conductive medium. There are many derivations of electrode placement, but for sleep, typical placements are at occipital lobe sites (O1 or

Table 1. Sleep Laboratory Measures (Possible)

Eye movements	Inductance plethysmography
Pupillometry	Intercostal EMG
Arms, legs and chin electromyography	Esophageal pressure
Nasal and oral airflow	Esophageal pH
End tidal CO_2	Pulse oximetry
Nasal pressure-flow	Transcutaneous oximetry
Pneumotach	Transcutaneous CO_2
Snoring microphone	Electrocardiogram
Chest wall movement	Blood pressure
Abdominal movement	Penile tumescence
	Body position

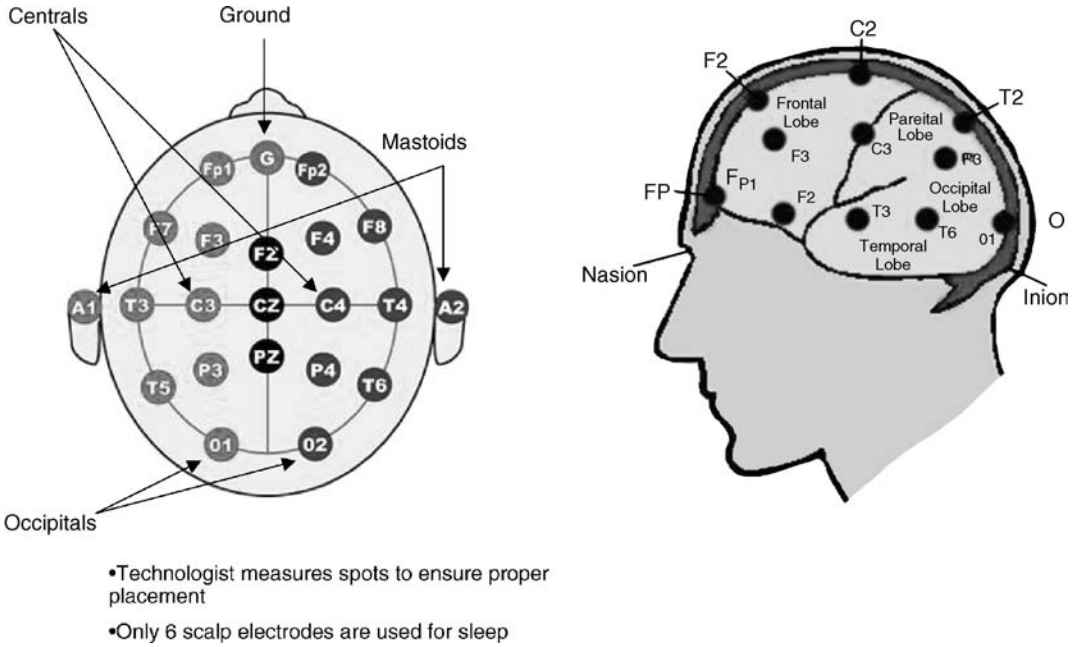


Figure 2. Electrode placement.

O2) with contralateral placement of a reference electrode on the mastoid process or earlobe. This gives the standard O1/A2 or O2/A1 electrode placement. Many labs also incorporate central electrode placement (usually C3/A2 or C4/A1). Occasionally, frontal electrodes are placed. The contralateral reference of electrodes typically allows a high amplitude of EEG signal, enabling adequate measurement of frequency and amplitude information. Normal sleep scoring rules have been based on central electrode placement. Occipital electrodes enhance alpha frequency (8–12 Hz) measurements and frontal electrodes enhance delta frequency (2–4 Hz) measurements.

Full montage EEG electrode placement is sometimes done to enhance the capture of nocturnal seizure activity. This replicates bipolar studies done for seizure monitoring and is especially useful for nocturnal movement disorder evaluation.

Electrooculography

The electrooculogram (EOG) is a measurement of eye activity. Phasic bursts of rapid eye movement occur during rapid eye movement sleep (REM sleep). The EOG is important for recognition of REM sleep, as well as for detecting the transition from wakefulness to Stage 1 sleep, which is characterized by slow, rolling eye movements.

The EOG recordings are possible due to a small electropotential difference from the cornea to the retina. The electrodes are positioned at the right outer canthus (ROC) and the left outer canthus (LOC), eferenced to auricular electrodes. Thus, ROC/A1 and LOC/A2 will register as out-of-phase pen deflections. This facilitates artifact recognition, as well as EEG activity recorded in the eye electrodes. EOG placements incorporate a somewhat superior placement of the electrode on one eye and inferior on the other eye, which allows recognition of vertical eye movements.

Electromyography

Three electromyographic (EMG) electrodes placed in the mentalis–submentalis area allow detection of EMG activity. During sleep, there is a gradual decline in EMG activity, and the dramatic reduction of muscle tone during REM, makes the EMG an important confirmation of REM sleep. Again, adequate removal of oils and dead skin cells with a conductive electrode placement enhances signal processing.

Electrodes or strain gauge sensors are also commonly placed on the tibialis muscles (Fig. 3). The EMG and movement signals from the lower extremities, and sometimes the arms, allow measures of isolated movements, as well as periodic limb movements that may be seen in restless legs syndrome, periodic limb movements of sleep, REM behavior disorder, seizures, parasomnias, and other disorders.

Figure 3. Electrodes applied to legs.

Figure 4. Technologist monitors through the night.

Other Measures and Filtering of Signals

Other biological measures have been incorporated into the polysomnographic measurements. These include body position, body movement, transient arousals, respiratory abnormalities, heart rate, oxygen saturation, esophageal pH, and esophageal pressure.

Video monitoring, as well as audio monitoring, are desirable features and occur in most laboratories. Video recordings with infrared or low light allow correlations of behavioral states with the physiologic measures.

The multiple channels of biological data are gathered in a bundle of electrical wires (the ponytail) at the back of the scalp. The electrical signals require signal processing, amplification, filtering, monitoring, and subsequent collation. The impact of signal output, sampling rates, filtering, and video recording are all factors in the design and in the monitoring of the patient (Fig. 4), as well as scoring information that is processed according to peer-reviewed criteria and event definitions.

The essential measurements in a sleep laboratory involve sleep staging (described below) and this requires calibration, with close attention to signal amplitude and filtering. Frequency measurements ranging from 2–50 Hz are common for measurement of EEG signals. The EMG signals have much higher frequency and sample filtering usually up to 75 Hz, with a notch filter at 60 Hz used to reduce alternating current noise. Signals <10 Hz are usually filtered for an EMG signal.

SCORING AND INTERPRETING PSG

Sleep Staging

Standard sleep stage scoring is guided by Rechtschaffen and Kale's scoring rules (8). The standard EEG recording is negative-up with amplitude measured from peak to valley of the waveform. Delta rhythm is 2–4 Hz; theta is 4–8 Hz; alpha rhythm is 8–12 Hz; beta rhythm is 13–30 Hz; and gamma is >30 Hz. Sleep is scored in 30 s epochs, with a chart speed equivalent to 10 $mm \cdot s^{-1}$. The EEG, EOG, and EMG signals are necessary for sleep state assessment and are used to study clinical and research physiology, as well as pathologic processes. The epoch is scored based on the majority population of EEG/EOG/EMG activity during the 30 s epoch.

Sleep is thus staged as non-REM sleep Stages 1, 2, 3, and 4 and REM sleep. The data are collated in both graphic, as well as tabulated, form to present information from lights-out to sleep onset, as well as quantifying total sleep time, time in bed, sleep efficiency (sleep time divided by time in bed), wake after sleep onset, and sleep stage distribution. A sleep histogram of the night of sleep summarizes the stage distribution pictorially.

Special rules are established for movement, arousals, and periodic limb movements in sleep, as well as consideration for specific disease processes. For example, narcolepsy may be associated with excessive motor activity during sleep, with an elevated EMG amplitude, particularly during phasic rapid eye movement sleep. Obstructive sleep apnea patients may have <15 s of sleep during a 30 s epoch due to arousals from obstructive events, forcing the scoring of the epoch as awake. Special rules to allow for these variances are in place.

Sleep-Disordered Breathing

Abnormal breathing during sleep is one of the most common disorders and accounts for as much as 70% of patients presenting to a sleep laboratory. Some laboratories specialize in sleep-disordered breathing alone. Snoring with increase in upper airway resistance due to functional relaxation of dilator muscles in the pharynx is a common disorder, ranging from 20% to 40% of the adult population and 13% of children. Anatomic changes, particularly tonsillar hypertrophy, add to the increase in upper airway resistance. This increase in the work of breathing may cause unstable breathing, with episodes of apnea or hypopnea, with associated changes in oxygenation, autonomic tone, and sleep state.

Obstructive apnea is defined as 10 s of cessation of airflow, despite a continued effort to breathe, in the adult population. Hypopnea is defined as a discernable reduction in flow or effort that produces a 4% desaturation. Central apnea is cessation of flow associated with cessation of effort for at least 10 s in the adult population. In the pediatric population, the duration for defining apneas is shorter. In some laboratories, hypopneas are also defined as a discernable reduction in flow or effort, which terminates in a 3% desaturation or an arousal. Mass loading of the diaphragm due to obesity and intrinsic central nervous system regulation of breathing are other variables to be considered clinically, as well as in research.

Current laboratories monitor flow, effort, and oxygen saturation as surrogate measures of minute ventilation and gas exchange. The intent is to assess the effort to breathe, the results of the output of ventilation, and, ideally, gas exchange, especially O_2 content or oxygen saturation. These measurements along with clinical correlation provide risk assessment and guide treatment plans.

Methods to detect airflow include pneumotachography, nasal airway pressure, thermistors, and thermocouples, as well as expired CO_2. Nasal airway pressures resemble the signals from a pneumotachograph, and therefore give a

Figure 5. Sleep setup with respiratory belts.

qualitative measure of increase in upper airway resistance. The pressure gradient intranasally compared to ambient atmospheric pressure is used to calculate the flow. Effort signals may be generated by flexible bands around the chest and the abdomen to detect movement. This is the most common technique (Fig. 5). Strain gauges, impedance pneumography, inductance plethysmography, and intercostal muscle electromyography have been used. Snoring sensors include microphones and piezoelectric assessment of vibration.

Oxygen content is most commonly measured by pulse oximetry. Decrease in oxygen saturation during sleep is most commonly due to changes in ventilation. Oxygen desaturation is an important criterion for scoring respiratory events, and thus critical for diagnostic studies. Intrinsic lung disease or cardiac disease may produce ventilation-perfusion mismatch and may also produce decreases in O_2 saturation. In a stable state, however, most decreases in oxygen saturation reflect changes in minute ventilation.

Pulse oximetry uses a two wavelength light transmitter, using spectrophotoelectrical techniques based on oxyhemoglobin absorption. Pulsatile tissues are necessary for pulse oximetry, and therefore these are usually applied to the finger, earlobe, or nasal sites. Reduction in pulsation, skin pigment, and dyshemoglobinemias may interfere with signal processing. The sampling rate, filtering, and proprietary algorithms for signal processing may affect the resolution of the signal.

Transcutaneous oxygen and carbon dioxide are other techniques available, but they are not commonly used, except in neonates. Transcutaneous CO_2 may be useful to assess chronic hypoventilation, but the resolution time for these measures limits their use in routine polysomnography.

Esophageal pressure measurements using an esophageal balloon provide a highly accurate measure of intrapleural pressure and, therefore, work of breathing. Some sleep centers use esophageal pressure measurements; however, this is not routine, as the invasive technique disturbs sleep. However, esophageal pressure measurements enhance the ability to measure upper airway resistance syndrome and central alveolar hypoventilation. Cardiac data primarily consists of heart rate and rhythm. Limited EKG electrodes, equivalent to lead II (right arm–left arm) or precordial leads (V5 or V6) are used, depending on the electrical axis and lab protocol. Clinical studies focus on high, mean, and low heart rates, as well as semiquantitative data on arrhythmias as to frequency and type. These are correlated with respiratory events or sleep state.

In patients with obstructive sleep apnea, efficacy of treatment is measured by applying positive airway pressure (PAP) or bilevel positive airway pressure (BiPAP). Flow signals are generated from the hardware device, allowing quantitative measures of flow and upper airway resistance based on waveform characteristics. Sleep state, body position, oxygen saturation, electrocardiogram, flow, and effort data are recorded to assure proper titration. The goal is to reduce the respiratory disturbance index, ideally < 5, and to minimize oxygen desaturations.

Daytime Sleepiness

Objective assessment of daytime sleepiness is another important laboratory diagnostic measure. The multiple sleep latency test involves a series of nap opportunities at 2 h intervals during the day. The standardized technique and normative data facilitate objective measures of excessive daytime sleepiness. This test is primarily used in the diagnosis of narcolepsy. The maintenance of wakefulness test is similar to the multiple sleep latency test, but requires the individual to attempt to remain awake during 20 or 40 min nap opportunities in a dark room.

Treatment Guidelines. Practice-based guidelines by the American Academy of Sleep Medicine and medical specialty societies have established guidelines for diagnosing and treating sleep disorders. These can be accessed on their website (www.aasmnet.org).

ADDITIONAL AMBULATORY MEASURES

Alternative or ancillary devices have been developed, primarily in the area of sleep-disordered breathing and brain state (sleep–wake estimate). Screening devices have been developed for obstructive sleep apnea primarily. Portable outpatient systems for unsupervised polysomnography are available. The unattended study is considered a level II evaluation. Level III is portable sleep apnea testing, monitoring airflow, effort, electrocardiogram, and oxygen saturation. Level IV is arterial oxygen saturation measurement alone.

Actigraphy incorporates an accelerometer used to measure wrist movement. Complex algorithms have been developed to estimate sleep/wake state, periodic limb movements, and circadian rhythm abnormalities. Actigraphy is particularly useful for long-term monitoring of patients in their homes, and for detection of daytime napping, which is usually not possible for EEG, as it is restricted to the night.

DIGITAL PGS AND FUTURE TRENDS

The EEG, EOG, and EMG signals were traditionally recorded in a sleep laboratory with limited channels on

analog systems, with pen deflections recorded on paper. Digital recording systems are more often used now, replicating a chart-paper speed of 10 $mm \cdot s^{-1}$. Digitized signals with current hardware and software systems allow enhanced collation of data, increased numbers of channels, enhanced filtering, and the potential for automated analysis. Archiving, data compression, and feature extraction are also possible.

Digitized signals of the ECG with high sampling rates, preferably at 200 Hz or higher, provide an opportunity for high resolution analysis. High resolution analysis might provide clearer insights into autonomic tone and vascular resistance and, the influence of sleep states, pathological conditions, and transient arousals. Further development of quantitative measures of ventilatory response, changes in gas exchange, as well as central nervous system and cardiovascular changes, could be helpful for diagnosing and treating sleep-disordered breathing.

Software processing collates data, presenting information in a tabular, as well as in a graphic, format. Clinicians and scientists review raw data, as well as human-supervised or scored changes. These changes can produce an audit trail referenced to the raw data. The fields of data can populate an electronic medical record, as well as a database, including demographics, comorbidities, medications, and fields of data from the polysomnogram, as well as the final diagnosis, treatment, and outcome measures.

Conceivably, advanced EEG analysis might reduce interscorer variability, thus better defining state and process. Enhanced resolution of EEG with improved consistency might allow better feature extraction, such as transient arousals and continuity measures that may improve research and clinical care. This could significantly improve the efficiency, cost and safety in drug development.

Opportunities for the future in sleep diagnostics are numerous. The duration of a polysomnogram, as well as long-term EEG monitoring, create challenges in the quality of the signal. Improvement of electrode–skin interface and conductivity are desirable. Application of wireless technologies to minimize the ponytail bundle of wires would be extremely useful. Audiovisual monitoring quality, advanced respiratory analysis, high resolution cardiac assessment, technical acquisition, collation, and data management are all current challenges.

BIBLIOGRAPHY

1. Schnieder C, Fulda S, Schulz H. Daytime variation in performances and tiredness/sleepiness ratings in patients with insomnia, narcoplepsy, sleep apnea and normal controls. J Sleep Res 2004;13:373–383.
2. Ancoli-Israel S, Cooke JR. Prevalence and comorbidity of insomnia and effect on functioning in elderly populations. J Am Geriatr Soc 2005;53(Suppl. 7):S264–S271.
3. Thomas RJ, et al. Functional imaging of working memory in obstructive sleep-disordered breathing. J Appl Phyiol 2005; 98:2226–2234.
4. Drake CL, Roehrs T, Roth T. Insomnia causes, consequences, and therapeutics: An overview. Depress Anxiety 2003;18:163–176.
5. Metlaine A, Leger D, Choudat D. Socioeconomic impact of insomnia in working populations. Indus Health 2005;43:11–19.
6. Foley DJ, et al. Sleep complaints among elderly persons: An epidemiologic study of three communities. Sleep 1995;18:425–432.
7. Jasper HH. The ten-twenty electrode system of the International Federation. Electroencephalogr Clin Neurophysiol 10:370–375.
8. Rechtschaffen A, Kales A. A Manual of Standardized Terminology, Techniques, and Scoring Systems for Sleep Stages of Human Subjects. Los Angeles: Brain Information/Brain Research Institute UCLA; 1968.

Further Reading

Bassetti CL. Sleep and stroke. Semin Neurol 2005;25:19–32.

Drummond SP, Gillin JC, Smith TL, DeModena A. The sleep of abstinent and pure primary alcoholic patients: Natural course and relationship to relapse. Alcohol Clin Exp Res 1998;22:1796–1802.

George CF. Sleep. 5: Driving and automobile crashes in patients with obstructive sleep apnoea/hypopnoea syndrome. Thorax 2004;59:804–807.

Phillips B. Sleep disordered breathing and cardiovascular disease. Sleep Med Rev 2005;9:131–140.

Spira AP, Friedman L, Flint A, Sheikh JI. Interaction of sleep disturbances and anxiety in later life: Perspectives and recommendations for future research. J Geriatr Psychiatry Neurol 2005;18:109–115.

Vgontzas AN, Bixler EO, Chrousos GP. Sleep apnea is a manifestation of the metabolic syndrome. Sleep Med Rev 2005;9:211–224.

See also CONTINUOUS POSITIVE AIRWAY PRESSURE; ELECTROENCEPHALOGRAPHY; SLEEP STUDIES, COMPUTER ANALYSIS OF; VENTILATORY MONITORING.

SLEEP STUDIES, COMPUTER ANALYSIS OF

DAVID WM. SHUCARD
BRETT A. PARMENTER
State University of New York
Buffalo, New York

ALI A. EL SOLH
Erie County Medical Center
Buffalo, New York

INTRODUCTION

Human beings spend approximately one-third of their lives asleep. Prior to the advent of physiological records, technology, and digital computers, our understanding of this ubiquitous phenomenon was based on observations of people or animals sleeping. Yet, this observational approach placed severe limitations on our understanding of the functions of sleep and the neurological processes that underlie this apparently quiescent period. It was because of several major scientific developments, all of which occurred during the twentieth century, that the concept of sleep as an active process (rather than a passive one) became appreciated, making the advent of modern sleep medicine possible.

Sleep was generally regarded as a passive process, "...the intermediate state between wakefulness and death..." (1). In 1939, University of Chicago physiologist Nathaniel Kleitman published a monograph "Sleep and Wakefulness" based on observations of human subjects who were sleep deprived. He determined that sleep may be viewed as a "let down of waking activity" and that "there may be different kinds of wakefulness" (2). In the 1950s, Kleitman and his students, Aserinsky and Dement, went on to describe rapid eye movement sleep (REM) in humans and the relationship between REM and dreaming (3,4). To follow up on the earlier work of, Dement and Kleitmen (5) undertook a particularly ambitious project, considering the available technology. They recorded the electroencephalogram (EEG) and electrooculogram (EOG) continuously from 33 subjects for a total of 126 nights. Recordings were all done on paper and manually scored. This work revealed that sleep had an architecture and that the EEG showed cyclical variation throughout the night. REM periods occupied ~20–25% of sleep. Dement and Kleitman characterized sleep stages as they are known today and determined that sleep consisted of two distinct states (REM and non-REM), neither of which could be considered as a time of brain inactivity.

Importantly, the foundation for this work was the recording of the human EEG in 1928 by Berger who showed that there were differences in the rhythm (or EEG frequency) when subjects were awake or asleep. Berger's work allowed, for the first time, the EEG to be monitored continuously during sleep without disturbing the individual. Hobson (6) suggested that Berger's work was the turning point of sleep research. Much of what is known about the EEG characteristics of the different stages of sleep was described in the 1930s. Further, it was determined that sleep was an active process that involves the synchronization of a number of neurological systems.

Perhaps one of the most important events in the history of sleep medicine was the discovery of sleep apnea by two separate groups in France and Germany in 1965. Narcolepsy and complaints of insomnia by patients further spurred the development of sleep centers that monitored patients' physiology as they slept. According to Dement (7), the "consolidation and formalization of the practice of sleep disorders medicine was largely completed" (p. 13) at the end of the 1970s. This labor intensive all night monitoring of patients was done without the benefit of digital computers, which did not become generally available for over a decade.

The purpose of this article is to discuss the use of computers in sleep disorders medicine and how this technology has enhanced the diagnostic and scientific capabilities of the sleep laboratory. The focus of the remainder of this article is on the computer analysis of the data obtained during polysomnography.

INTEGRATION OF COMPUTERS INTO POLYSOMNOGRAPHY

Over the years, and in particular with the development of the digital desktop computer, tools for studying sleep have advanced rapidly. Polysomnography (PSG), a term coined in 1974 by J Holland at Stanford University, allows for the assessment of sleep by simultaneously measuring multiple physiologic parameters. These parameters generally include EEG, EOG, electromyogram (EMG), electrocardiogram (ECG), respiration (both effort and air flow or pressure), blood oxygenation, and body–limb movement (8).

The results of an 8 h PSG can consist of >1000 pages (depending on the paper speed, or epoch length (e.g., 10, 15, 30 s) of data from multiple recording channels, making manual and visual interpretation a difficult and overwhelming task (9). Typically, each page (e.g., 30 s) of the recording was examined and sleep stage was determined based on the Rechtschaffen and Kales scoring system (10). Depending on the reason for the study, respiratory events such as nasal–oral air flow and respiratory effort, as well as blood oxygen saturation, were frequently recorded. Paper speeds were used so that each page of recording represented a 15–30 s epoch over the full night. So, for example, for a 7 h study and a paper speed set at 30 s epochs, there would be 840 pages of data to review. Scoring was done during the day after the study was completed and, as noted, each page was examined not only for sleep stage, but for changes in other measures including respiration, oxygen saturation, and limb movements. These events, along with stages, were manually tabulated and summarized. For clinical studies, each record was reviewed by the polysomnographer or sleep specialist and a report was composed that summarized the major diagnostic aspects of the study. For research purposes, variables of interest were entered into a data base for later analysis.

The development of digital computers (in particular the PC), their general availability, and the evolution of software for these systems specifically designed for the collection, storage, and manipulation of sleep data have played a significant role in the growth of the field of sleep medicine, and insights into sleep, in general. The vast amount of data collected from a full night sleep study can now be rapidly consolidated, allowing for comprehensive visualization of the measures obtained. Summary statistics pertaining to the sleep data recorded during the night are also obtained and easily accessed with the commercial computer data acquisition and analysis systems available for sleep studies. For example, once data are collected in an all-night study, the study can be examined over a variety of epoch window sizes from 10 s up to the thousands of seconds required to represent the entire study as one epoch. The ability to change the epoch window width allows for evaluation of different waveforms (e.g., EEG, respiration) using an optimal time scale. Further examples of sleep analysis software capabilities are presented below. By using such sleep systems, many laboratories that focus on the diagnosis of sleep disorders are now able to study as many as 6–10 patients in a single night.

COMPUTER ANALYSIS OF POLYSOMNOGRAPHIC DATA

Stages of Sleep

Sleep can be dichotomized into REM sleep and nonrapid eye movement (NREM) sleep (11). The NREM sleep can be

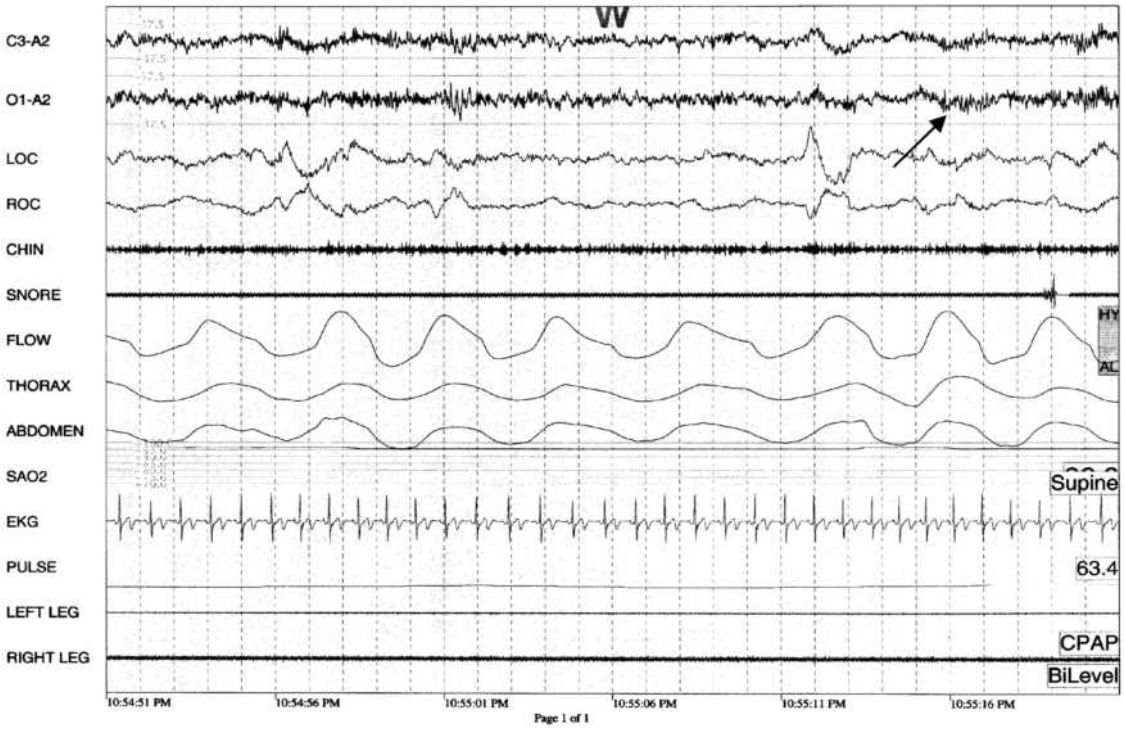


Figure 1. Illustration of Stage 1 sleep in Patient C. Epoch width and recording channels are the same as those in Fig. 1. Note the slowed eye movements (arrow).

further divided into four stages, according to rules established by Rechtschaffen and Kales (R&K) (10). The following outlines these stages and summarizes the major criteria that characterize them. *Stage 1 Sleep, illustrated in* Fig. 1, consists of a relatively low voltage EEG, with activity prominently from 2 to 7 cycles per second (Hz). Waves at this frequency are referred to as theta waves (12). Rapid eye movements are not present and eye movements are generally slow and rolling. This stage typically occurs during the transition from wakefulness (seen in Fig. 2) to the other sleep stages, and is relatively short. *Stage 2 Sleep, illustrated in* Fig. 3, is marked by K complexes and sleep spindles on a background of relatively low voltage EEG. The K complexes are negative, sharp, high voltage waves that are usually followed by a positive component. The complex duration, according to the R&K rules, should not exceed 0.5 s. The K complexes can occur spontaneously or in response to sudden stimuli and they are highest in amplitude over the vertex scalp regions (10,11). Sleep spindles are defined as "bursts of waves having a frequency of 12 to 15 [Hz]" (11, p. 16). In addition to the K complexes and sleep spindles, Stage 2 sleep lacks waves of high amplitude and slow activity seen in later stages (10). *Stage 3 Sleep, illustrated in* Fig. 4, is defined by delta waves, or slow waves of 2 Hz or less and amplitudes $>75\ \mu V$ (11,12). The EEG activity in Stage 3 should account for 20–50% of the time interval (epoch) studied (10). Sleep spindles may or may not be present. *Stage 4 Sleep, illustrated in* Fig. 5, is defined when >50% of the epoch consists of high amplitude waves that are 2 Hz or slower.

Stage REM, illustrated in Fig. 6, typically occurs after the first 70–100 min of sleep. Rapid eye movements characterize this stage of sleep. The EEG during REM sleep is similar to a waking or Stage 1 EEG, with low amplitude mixed frequency waves. If eye movements and EEG activity were the only available data by which to characterize this sleep stage, one would suspect that the person was likely awake. It is because of the similarity to wakefulness that REM sleep has often been called paradoxical sleep. Other significant features of REM are as follows: saw tooth waves that may or may not be present in conjunction with bursts of REM activity, the lack of sleep spindles, and mental–submental EMG that is almost always lowest during REM than that seen for other stages of sleep. After REM sleep, there is generally a cycling through Stages 2, 3, and back to 4. These cycles continue throughout the night (11).

Computer Identification of Sleep Stages

Manual scoring of the various sleep stages is a tedious and time-consuming task, with interrater reliability between 67 and 91% (13). As such, both research and clinical laboratories have explored the use of computers to objectively classify these stages. Initial attempts at automating sleep analysis concentrated on programming computers to identify sleep stages according to the R&K rules. However, newer methods have focused on pattern recognition of distinct sleep stages (13,14).

One method, the polysomnogram assay (PSGA), was developed by Bartolo et al. (9), and condenses the large

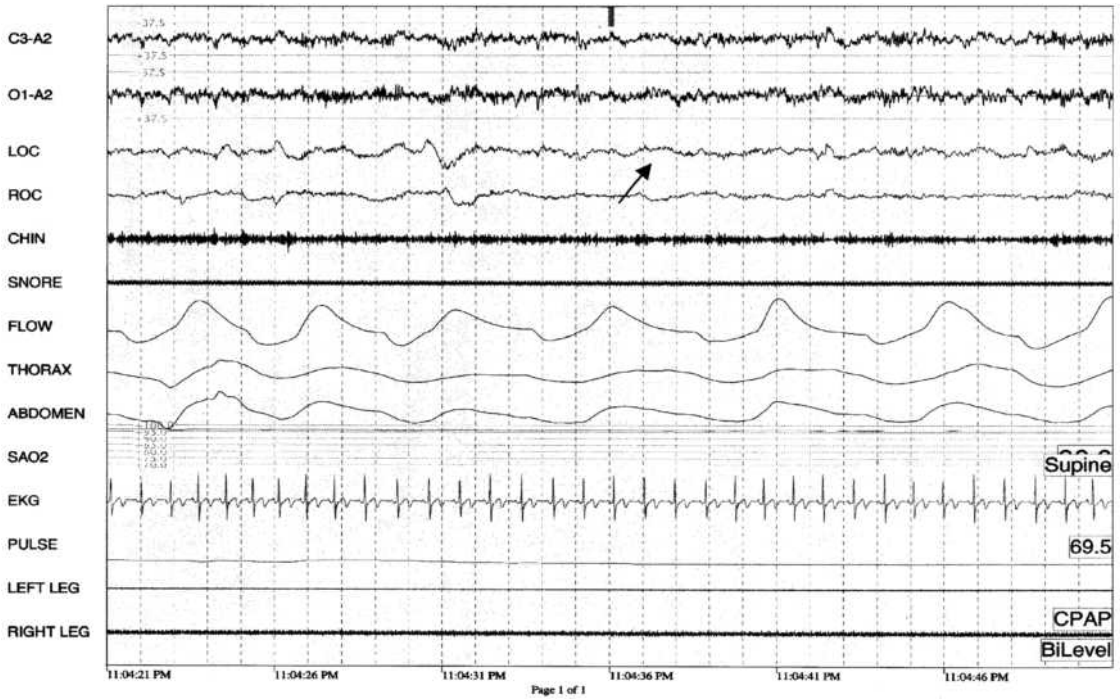


Figure 2. Example of a 30 s epoch of wakefulness from patient C being evaluated for a sleep disorder. The labels on the left indicate the measures at each channel of recording. The first two channels are EEG from the left central and left occipital scalp sites. LOC and ROC = left and right eye movements followed by chin EMG, nasal/oral air flow, thorax and abdominal respiratory effort, oxygen saturation, electrocardiogram, average pulse, left and right leg movement. Labels on the right indicate the patient's body position, average oxygen saturation (not shown), average pulse rate, level of continuous positive or bilevel air pressure (CPAP/BIPAP), if administered.

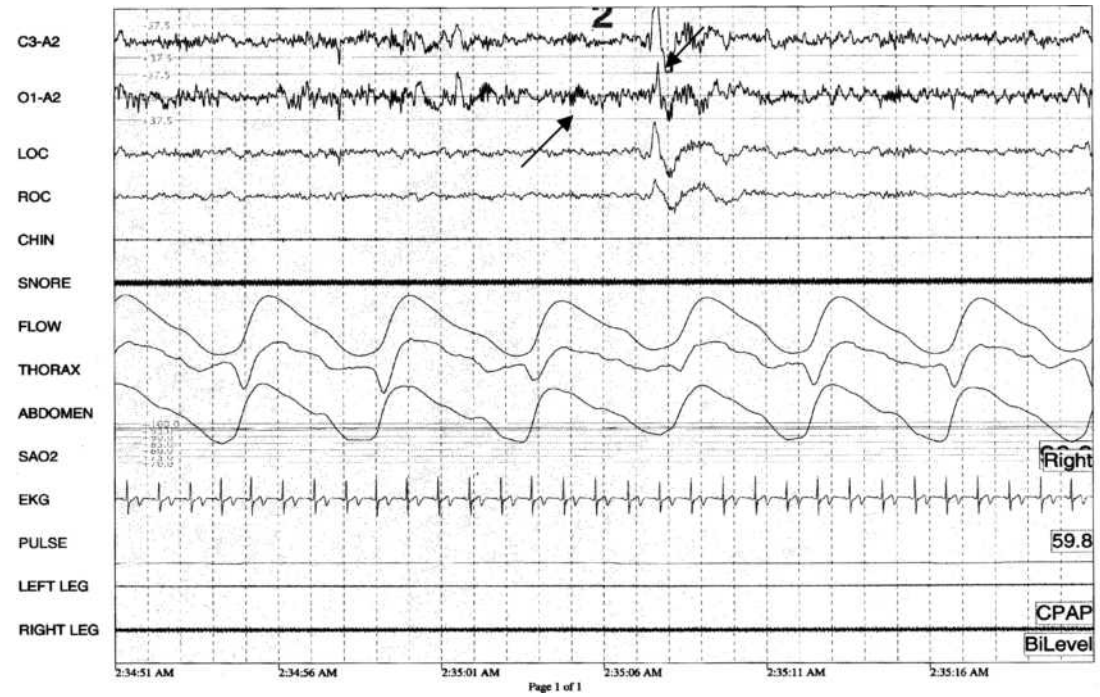


Figure 3. Illustration of Stage 2 sleep in patient C. Epoch width and recording channels are the same as those in Fig. 2. Note the K complexes (arrow Channel 1) and sleep spindles (arrow Channel 2).

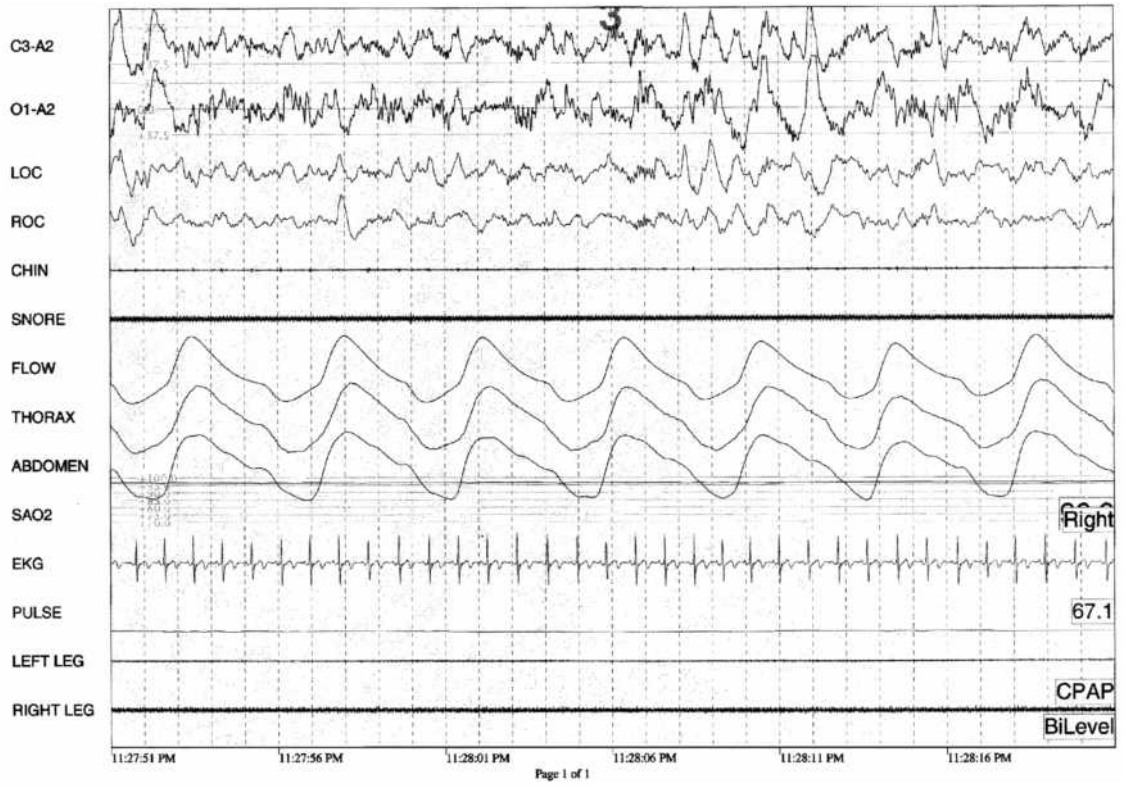


Figure 4. Illustration of Stage 3 sleep in patient C. Epoch width and recording channels are the same as those in Fig. 2. Note the high amplitude slow EEG waves and the sleep spindles in EEG channels 1 and 2 (arrow).

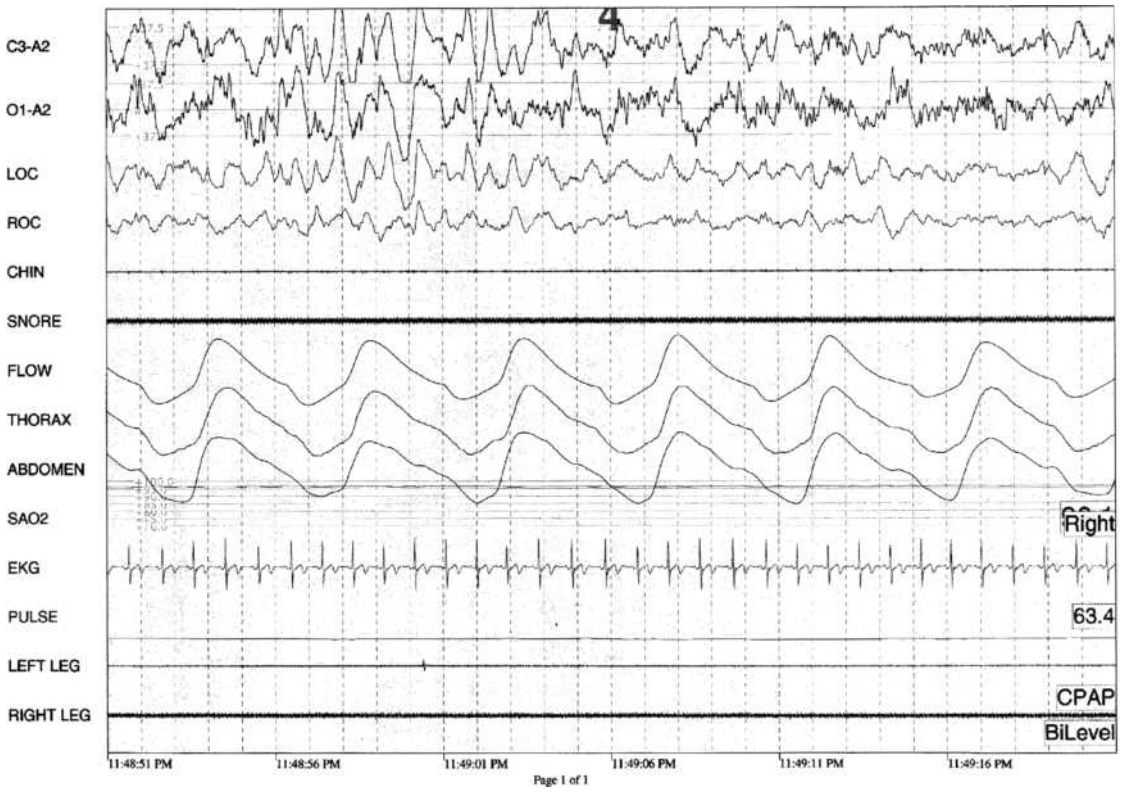


Figure 5. Illustration of Stage 4 sleep in patient C. Epoch width and recording channels are the same as in Fig. 2. Note the high amplitude slow EEG waves in EEG channels 1 and 2 (arrow). This slow activity occupies >50% of the epoch.

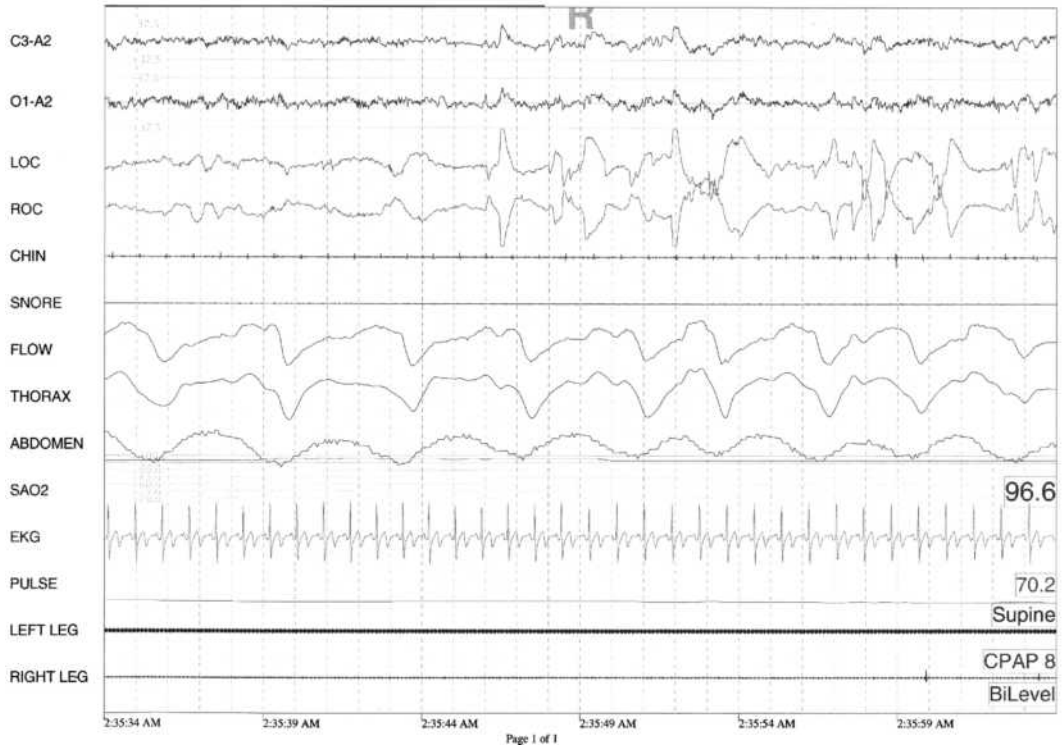


Figure 6. Illustration of Stage REM sleep in patient G. Epoch window and recording channels are the same as in Fig. 2. Note the low voltage EEG in EEG channels 1 and 2, the rapid eye movements in eye channels 3 and 4 (arrow), and the low amplitude chin EMG in channel 5.

amount of PSG data into a more comprehensible format while retaining events of diagnostic significance. The program allows gross distinction among sleep stages without manually scored data. However, because features associated with each sleep stage vary among individuals, the investigators noted that "it is unlikely that adequate sleep staging in the traditional sense can be accomplished by the PSGA or any other method except by visual analysis of the PSG itself" (p. 119). Nonetheless, the PSGA is able to present consistent patterns related to different sleep states, and with refinement may be useful for sleep staging. Two examples of how the data are consolidated are seen from the descriptions of leg and eye movements. Leg jerks appear as "abrupt peaks" (p. 121) in the EMG window assigned to leg movements. The EMG variables are represented by signal intensity or power. According to these investigators, periodic leg movements are more visible in the PSGA than in the PSG record. Additionally, the PSGA replaces the PSG EOG signals by markers that indicate the presence and intensity of eye movements. Thus, the PSGA transforms the data into events or markers appropriate for each type of measurement and displays these measures in a highly condensed format, allowing for up to 15 min of PSG data to be displayed on a single page. This method reduces the output from nearly 1000 pages to ~30. The purpose for the development of the PSGA was to "reduce the considerable human effort involved in scoring and evaluating PSG studies" (p. 124). This approach exemplifies how computers can assist in analyzing sleep, maximizing the amount of time polysomnographers can have for diagnosis and understanding unusual aspects of the sleep study that computers can only identify.

Agarwal and Gotman (13,15) described a computer assisted sleep staging method, which searches for clusters of wave patterns that have been identified and programmed as prototypes of waves or sleep stages. In this method, the PSG is first divided into multiple segments, containing at least 3 s of data. Second, prespecified features, such as sleep spindles and eye movements, are extracted by the computer from the PSG. Also, each epoch is examined for the different types of waves, including alpha and slow waves (delta, theta). After the data have been segmented and extracted, they are then "clustered" into groups with similar properties. These clusters can then be classified as particular sleep stages according to R&K or any other system preferred by the user. It appears that the approach taken by Agarwal and Gotman (13,14) relies on a self-organizational scheme that is based on relative differences in activity within a subject rather than fixed rules. So, for example, according to these investigators, stage REM sleep may be lost if hard thresholds for detection of atonia are used. This is the first study that has applied "self-organization[al] techniques to generate a hypnogram for an all night PSG in the context of R&K classification" (p. 1419). Agreement between automated staging and manual scoring according to R&K standards was 80.6% for 12 sleep records. This process uses the computer to initially analyze the data and identify various sleep

stages before it is inspected manually by a polysomnographer. Since this method is not fully computer automated, but is computer assisted, the investigators referred to this approach as "computer-assisted sleep staging (CASS)."

Because sleep stages are not distinct entities, computer errors are made. Empirical studies confirm that, despite our best effort to demarcate sleep stages for computer detection, errors occur, especially for the less clearly defined stages (13). For example, Agarwal and Gotman (13) reported that many errors occur when attempting to define Stage 1 sleep, perhaps the most ambiguous, which was often misclassified as either wakefulness or Stage 2 sleep. For computers to identify components of sleep, these components must be clearly defined. At present, the best defined features of sleep are the various stages according to the R&K rules. Some of these rules are specific enough for an algorithm to be programmed so that a computer can identify it on a polysomnogram. For example, Stage 3 sleep is defined by delta waves with amplitudes $>75\,\mu V$, present between 20 and 50% of the time. Other definitions are less precise, such as the definition of Stage 2 sleep, which is marked by sleep spindles, K-complexes, and waves that lack the amplitude and slower activity seen in sleep Stages 3 and 4. Stage 2 sleep is more difficult to quantify and, therefore, less amenable to detection by computers.

Computer Analysis of Sleep Microstructure

From a more academic, less clinical, perspective, computers have been used to assist with analysis of the microstructure of sleep, such as parsing out the various components of specific waves (e.g., sleep spindles, K complexes) and exploring their relationships to various sleep problems. Although the presence of sleep spindles in Stage 2 sleep is well known, the different types of spindles and their subsequent spatial distribution are less well studied. With the aid of computers, two types of spindles have been identified: 12 spindles per second located primarily in the frontal region and 14 spindles per second located primarily in the centroparietal region (14). Our understanding of K complexes has also been improved with the aid of computers. Specifically, a classification of K complexes has been recommended, that includes K0 complexes (without sleep spindles), K1 complexes (spindles preceding the K complexes), K2 complexes (spindles occurring simultaneously with the K complexes), and K3 complexes (spindles following the K complexes (14)).

Computers have also been used to assess new components within the stages of sleep (e.g., slow wave sleep and REM sleep). Initially, slow wave sleep was defined as sleep marked by delta waves, occurring during stages 3 and 4. Delta waves were believed to be uniform, with no variable components. However, with the aid of computers, two types of these waves have been identified with scalp topographical differences. Sinusoidal $1–2\,s^{-1}$ delta waves tend to be located frontally, and have been compared to K complexes that are not precipitated by sensory stimuli. Polymorphic $<1\,s^{-1}$ delta waves, on the other hand, are located parietotemporally. The role of these components is presently unclear. Our understanding of the microelements of REM sleep is also being explored. For instance, although sawtooth waves have been linked to REM sleep, their function is poorly understood (14). Only since the advent of computers have such microelements of the various stages of sleep been able to be identified, and is it only possible to study these components with computers in order to understand their functions and their possible contribution to sleep disorders.

A PROTOTYPICAL COMMERCIAL SLEEP ACQUISITION, ANALYSIS, AND MANAGEMENT SYSTEM

Although, computers do not eliminate errors in scoring sleep stages or various waves, they do substantially increase our ability to measure more aspects of sleep and to organize and reduce the data so that they are more readily interpretable. Digital polysomnography has become an important means of sleep analysis today. An important advantage is probably that the tracings can be zoomed in and out. While the EEG is best displayed on a high resolution monitor with 10–15 s per page (s/p) and 20–30 s/p is sufficient for sleep staging, respiratory-related signals and body movements can best be recognized at 2–10 min/p. The correlations between slight respiratory changes and EEG arousals are sometimes best observed with a 2 min page. Thus events that often remain undetected by the conventional paper method can be visualized. Consequently, the paper method is not an optimal method or a good alternative for a "gold standard". The possibility of adjusting the gain off-line is also valuable. However, the low dynamic range provided by many manufacturers set unnecessary limits. Eight-bit A/D conversion is still used in many systems, which gives a resolution of only 256 points. The introduction of 24 bit resolution in modern equipment has allowed for same amplifiers and gains to be used for practically all signals.

Computers also can provide summary statistics from data obtained over the entire night of recording. One frequently used commercial computerized sleep system allows up to 64 channels of recording. Computer algorithms are available that will detect significant changes in EMG, EEG, respiration, and heart rate. These programs are dependent on people to set criteria based on specific standards of what constitutes an event, such as an arousal, an obstructive apnea, hypopnea, a bradycardia, and a tachycardia.

In order to provide the reader with an appreciation of the state of the art in sleep study computer technology, one of the systems used by many sleep laboratories, including our own will be described next. The software runs on most PCs and is designed to collect, store, analyze, summarize, and manage the large amount of physiological data obtained during an all night sleep study. This system, with its appropriate amplifiers, digitizes the electrical voltages obtained from the various sensing devices, such as electrodes applied to the scalp, face, and chest. After these data are collected and saved, the software provides a number of analytic tools and modules that allow the polysomnography technician/polysomnographer to reduce and quantify the data. As described by the manufacture (Sandman Sleep Diagnostic Systems), the software consists of four main

components: *(1) Data Collection*, for recording and saving the signals obtained from the patient. *(2) Analysis*, for scoring and analyzing the data saved from in the patient file. Computer assisted scoring modules are part of this component. *(3) Data Management*, used for file manipulation such as copying, backing up, and deleting patient data. *(4) Configuration*, provides the user with the ability to manipulate some of the features of the software, change the storage media, and so on.

Figures 2–6 illustrating the stages of sleep are examples of the data obtained in our laboratory with this system. These figures represent epochs or time periods of the actual, data collected during an all night polysomnogram. Figure 7 presents sleep data from a patient with obstructive sleep apnea. Apnea is characterized by a cessation of airflow for 10 s or longer. Obstructive sleep apnea in contrast with central sleep apnea is caused by an actual obstruction of the airway and is characterized by a cessation of air flow, arousal, and increased effort to breathe (Fig. 7). Hypopnea is also an obstructive event, but the obstruction is not complete. It is generally defined by a reduction, without complete cessation, in airflow or effort (16). There are several criteria that laboratories have used to define an event as hypopnea. For example, with respect to air flow, both a 50% and 30% decrease from baseline air flow have been used. The EEG arousals and amount of oxygen desaturation have also played a role in the definitions of obstructive respiratory events. These criteria, used to define respiratory events, are all entered into the sleep scoring program so that the computer using the appropriate definition can select these events automatically. In addition, a polysomnographic technician reviews the entire sleep study and manually stages and confirms the relevant events. Once stages and events are reliably identified, a number of calculations are made via computer algorithms that summarize the information obtained over the sleep study. Figure 8 is an illustration of how data obtained over an entire night of sleep are summarized.

Software programs can now provide a report format that consolidates the information from the entire 6–8 h study once criteria are specified for relevant events by a sleep technician or polysomnographer. Such information can include measures of sleep efficiency, percentage of time in various sleep stages, the number of respiratory disturbances, the number of arousals, and other aspects of sleep, in order to accurately diagnose sleep disturbances. Figures 9 and 10 present a numerical summary and graphs of the data obtained for a patient being evaluated for sleep apnea. The system allows each laboratory to customize their own summaries and graphs with those variables they wish to send to the clinician along with the diagnostic interpretation. Note in Fig. 9 the patient's respiratory disturbance index (RDI) during REM sleep is $19.6\,h^{-1}$. Thus, this patient, during their total time in REM sleep, had an average of 19.6 respiratory events in one hour of REM.

As indicated, all of these processes are semi-automated and under the watchful eyes of trained polysomnography technicians and sleep specialists. The degree of automation is dependent on the reliability of the measure. As would be expected, sleep staging is less automated than the detection of awakenings or arousals. The most automated

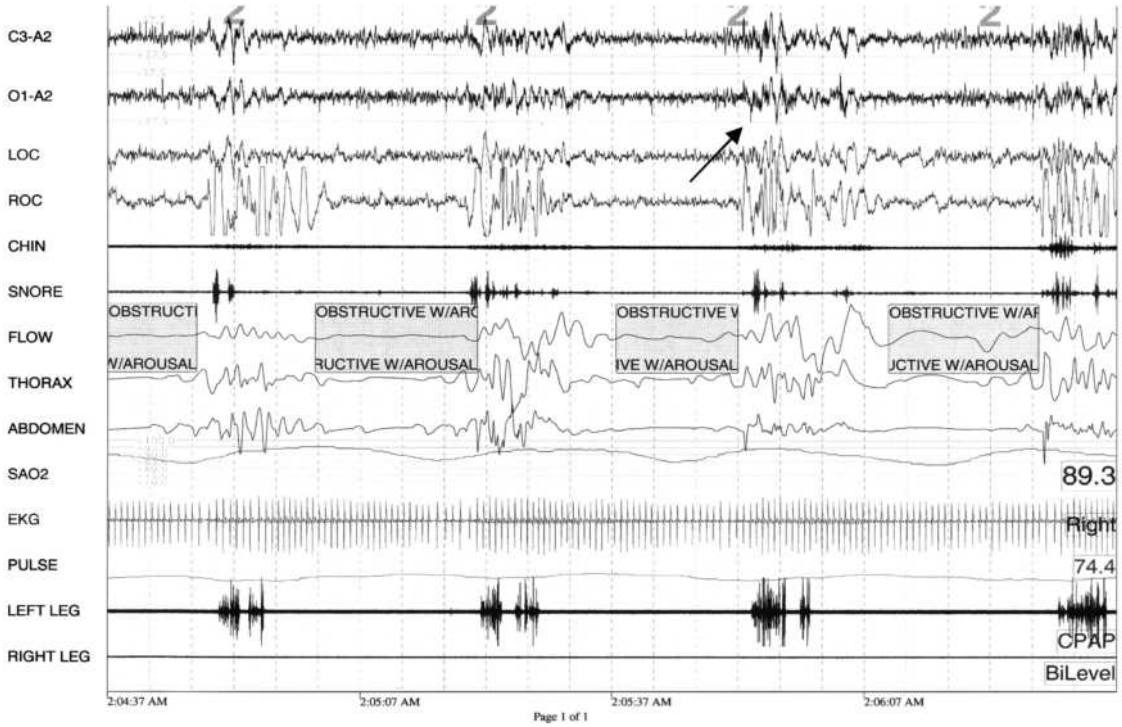


Figure 7. Illustration of obstructive sleep apnea in patient M. Epoch window is 120 s. Recording channels are the same as in Fig. 1–6. Arrow indicates EEG arousal following the event.

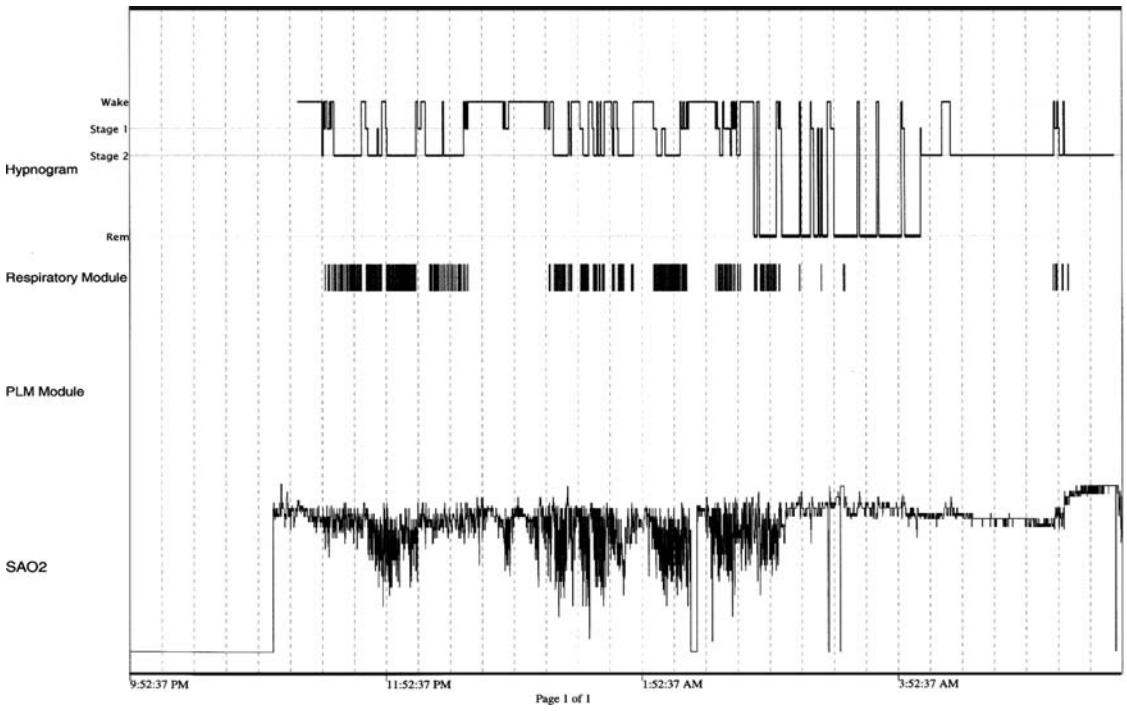


Figure 8. Illustration of computer summary of nocturnal polysomnogram for patient M. Note the oxygen desaturations (SaO_2) associated with respiratory events (Respiratory Module). Periodic limb movements (PLM Module) were not present.

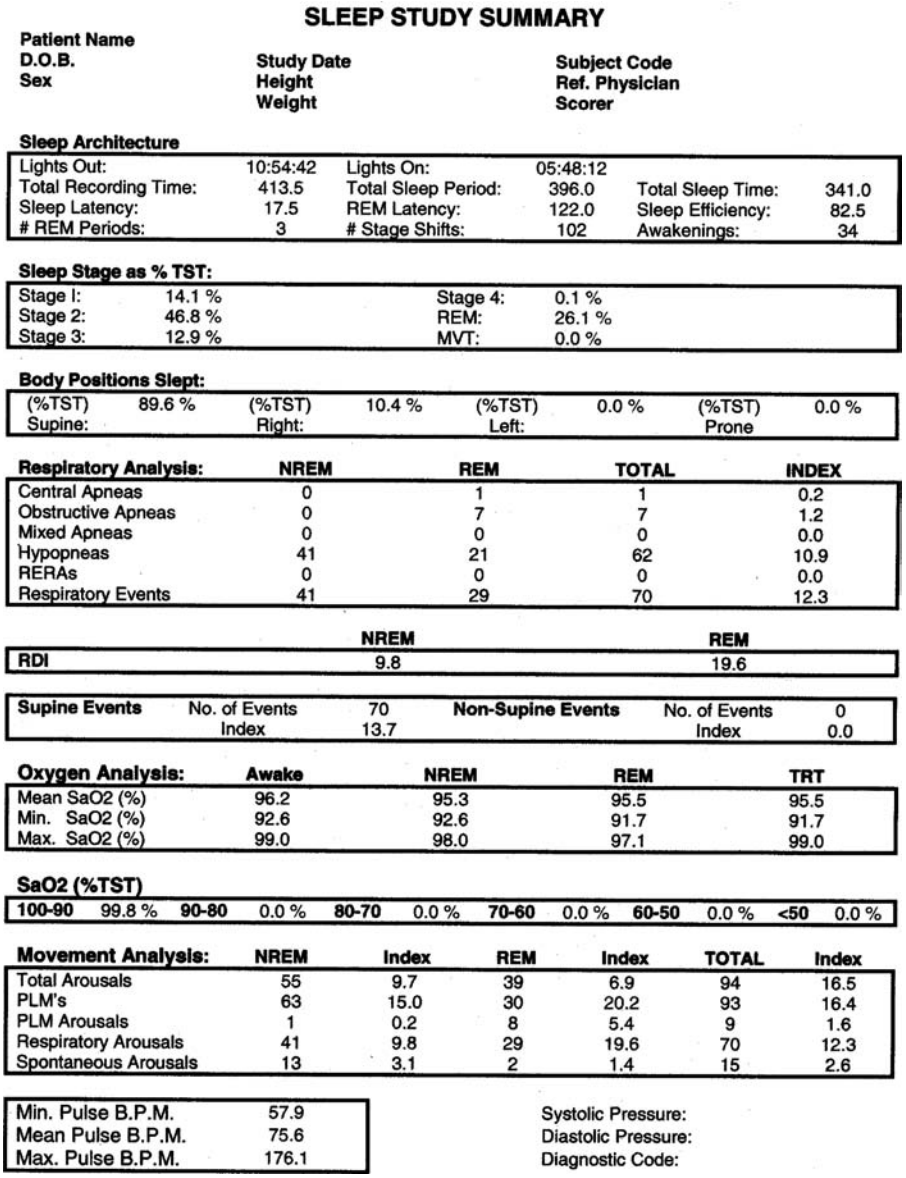
processes are those that involve specific calculations derived from the data collected during the sleep study. For example, the calculation of the overall RDI is the total number of respiratory events that occurred over the entire sleep study divided by the number of hours of sleep.

A tool available on the system, more for research than clinical evaluation, is the fast fourier transform (FFT). FFT is a mathematical analysis of a wave form that derives a frequency spectrum over time. The software displays a signal as a function of amplitude on the y axis versus time on the x axis. The module will perform this transformation on data of interest and write the results to a specified file. Frequency spectral analysis is also available for data reduction of EEG activity. For example, sleep spindles and EEG activity, such as alpha, beta, delta, and theta, can be summarized by plotting the signal amplitude on the y axis versus the frequency of the signal on the x axis. However, the utilization of fixed frequency bands can give misleading results. The bands can be too wide, too narrow, or displaced. Alpha activity of drowsiness and sleep is often on the border between alpha and theta (17) and the frequency of the sigma spindles can have a frequency in both the alpha and the beta range (18). Ideally bands should not be preselected. Rather, bands should be adjusted according to the signal. Martens et al. (19) applied matched filtering based on information obtained by FFT for this purpose. Another attempt to solve the problem is to study very narrow frequency bands and their relationships (20).

The accuracy of frequency analysis is also dependent on sampling rate. Often frequency is understood as an equivalent to the inverse of the baseline-crossing interval of a wave. However, in spectral analysis it is defined by the frequency content of the signal. With FFT the Nyquist theorem states that the sampling rate has to be at least twice the frequency of the fastest sine waves. However, EEG rhythms are often not sine waves and thus include frequency components that are both faster and slower than the frequencies calculated on the basis of their baseline-crossing intervals. Thus, in order to preserve the form of the waves, higher sampling rates have to be used. Also, the accuracy of baseline-crossing is dependent on the sampling rate. Thus, if the maximum frequency of a sigma spindle is 16 Hz, then a sampling rate of 1024 Hz is required in order to obtain an accuracy of at least 0.25 Hz.

LIMITATIONS OF THE USE OF COMPUTERS TO ANALYZE SLEEP

Despite the overwhelming contribution of computers in analyzing and furthering our understanding of sleep, limitations exist. Such limitations are expected when computers are relied on to explain a complicated, biological process. In fact, both intra- and interindividual variability contribute to the challenge of quantifying components of sleep, such as sleep stages or waveforms. The separation of wakefulness and stage REM is, in some systems, very dependent on the quality of the EMG recording. Because separation of REM and wakefulness is often difficult due to the considerable variation in the EMG activity level obtained for each stage, human supervision and adjustment of detection levels are necessary in practice.



SLEEP STUDY SUMMARY

Patient Name
D.O.B. **Study Date** **Subject Code**
Sex **Height** **Ref. Physician**
Weight **Scorer**

Sleep Architecture

Lights Out:	10:54:42	Lights On:	05:48:12		
Total Recording Time:	413.5	Total Sleep Period:	396.0	Total Sleep Time:	341.0
Sleep Latency:	17.5	REM Latency:	122.0	Sleep Efficiency:	82.5
# REM Periods:	3	# Stage Shifts:	102	Awakenings:	34

Sleep Stage as % TST:

Stage I:	14.1 %	Stage 4:	0.1 %
Stage 2:	46.8 %	REM:	26.1 %
Stage 3:	12.9 %	MVT:	0.0 %

Body Positions Slept:

(%TST) Supine:	89.6 %	(%TST) Right:	10.4 %	(%TST) Left:	0.0 %	(%TST) Prone	0.0 %

Respiratory Analysis:	**NREM**	**REM**	**TOTAL**	**INDEX**
Central Apneas	0	1	1	0.2
Obstructive Apneas	0	7	7	1.2
Mixed Apneas	0	0	0	0.0
Hypopneas	41	21	62	10.9
RERAs	0	0	0	0.0
Respiratory Events	41	29	70	12.3

	NREM	**REM**
RDI	9.8	19.6

Supine Events	No. of Events	70	**Non-Supine Events**	No. of Events	0
	Index	13.7		Index	0.0

Oxygen Analysis:	**Awake**	**NREM**	**REM**	**TRT**
Mean SaO2 (%)	96.2	95.3	95.5	95.5
Min. SaO2 (%)	92.6	92.6	91.7	91.7
Max. SaO2 (%)	99.0	98.0	97.1	99.0

SaO2 (%TST)

100-90	99.8 %	**90-80**	0.0 %	**80-70**	0.0 %	**70-60**	0.0 %	**60-50**	0.0 %	**<50**	0.0 %

Movement Analysis:	**NREM**	**Index**	**REM**	**Index**	**TOTAL**	**Index**
Total Arousals	55	9.7	39	6.9	94	16.5
PLM's	63	15.0	30	20.2	93	16.4
PLM Arousals	1	0.2	8	5.4	9	1.6
Respiratory Arousals	41	9.8	29	19.6	70	12.3
Spontaneous Arousals	13	3.1	2	1.4	15	2.6

Min. Pulse B.P.M.	57.9
Mean Pulse B.P.M.	75.6
Max. Pulse B.P.M.	176.1

Systolic Pressure:
Diastolic Pressure:
Diagnostic Code:

Figure 9. Illustration of a sleep study summary. The values are calculated by specific computer algorithms and represent, in this case, data obtained over ~7 h of recording. This summary accompanies the diagnostic interpretation of the study that is forwarded to the patient's clinician.

Computer analysis is also often dependent on measurement of the wakefulness alpha activity, which causes problems with low alpha subjects. In this case, it is an advantage to rely on theta activity and non-EEG waveforms (21,22). Also, the distinction between REM and NREM sleep is difficult in patients with poor sigma spindles or if the spindles are of a frequency outside the usual range (18). Toussaint et al. (23) discussed the first-night effect in sleep studies, noted by lower sleep efficiency, increased wakefulness, and longer REM periods in patients. Computer algorithms have been difficult to derive that can take into account the first-night effect in individuals because it is relative to the person's normal sleep rather than referring to objective, specific changes in sleep across all individuals.

Interindividual differences in sleep measures have also been demonstrated. For example, variability in polysomnograms has been associated with both gender (24–26) and age (27,28). For example, Stage 4 sleep presents differently in a young child than it does in an adult (28). Older adults have lower amplitude delta waves than children. Although this is a broader issue regarding the definition of Stage 4 sleep across the age range, it is also illustrative of the difficulties in standardizing automated computerized sleep staging systems. A computer must be programmed to specifically identify such sleep variability for fully automated scoring to be accurate. Furthermore, studies also show marked variability between various patient populations, with different polysomnograms seen for patients with schizophrenia (29,30), autism (31), and attention deficit/hyperactivity disorder (32). Thus, one issue with computer automated sleep scoring is how automated can or should the process be? Computers can only accomplish what they are programmed to do (27), and errors will occur due to artifacts or unexpected events (15). This issue was highlighted in a study by Cirignotla et al. (33) in which they

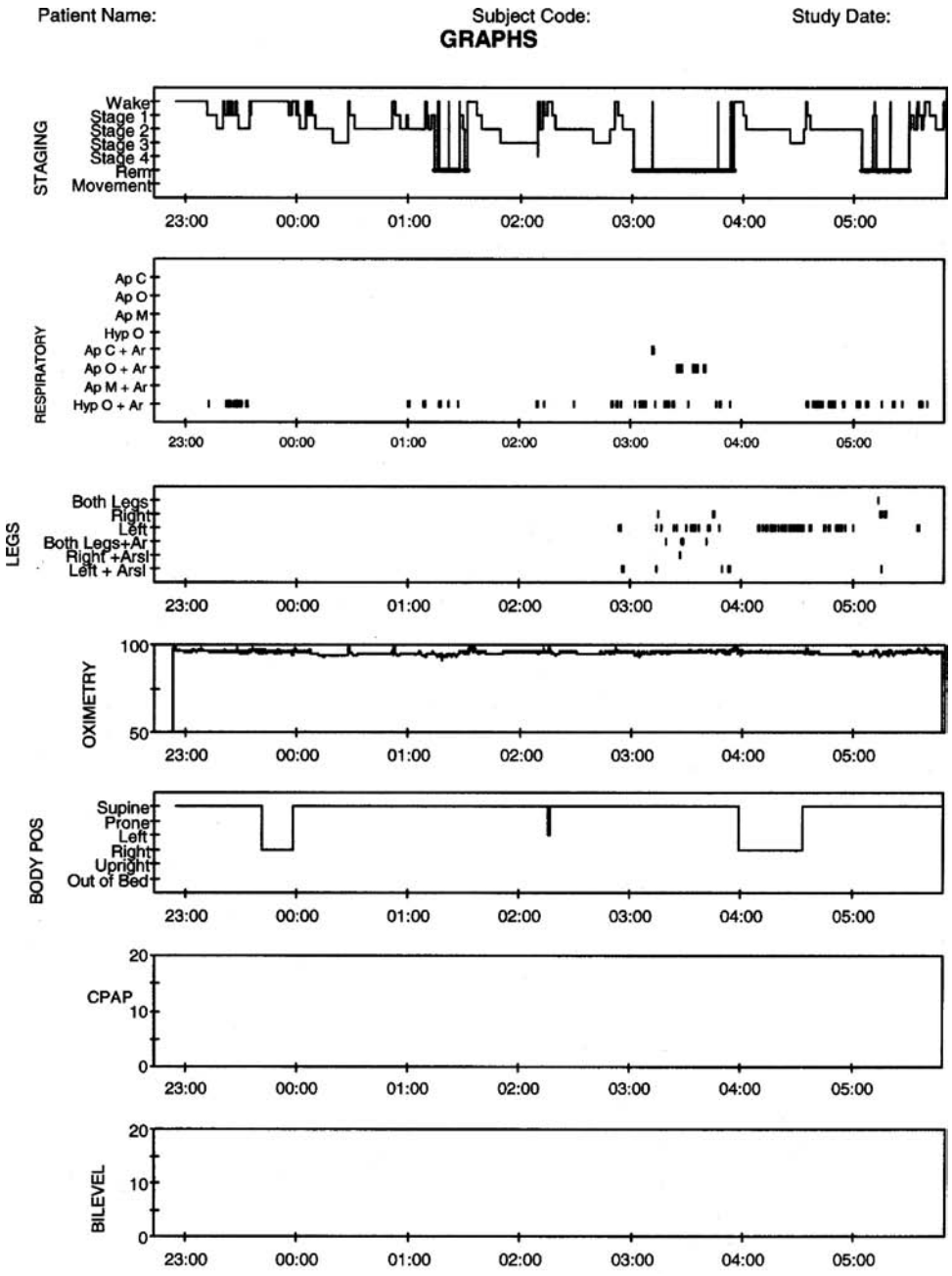


Figure 10. Illustration of graphs representing the numerical data presented in Fig. 9. These graphs are automatically plotted by the computer. The specific format is designed by the laboratory. Note that this patient, for example, has a significant positional component to their sleep apnea. As indicated in Fig. 9, the Supine RDI is $13.7\,h^{-1}$ compared to a nonsupine RDI of $0.0\,h^{-1}$.

investigated the MESAM 4, a computerized, ambulatory cardiorespiratory monitor that can help diagnosis sleep apnea. However, the investigators found that this computerized system significantly underestimated the oxygen desaturation index in patients with complicated obstructive sleep apnea, or patients who suffered from sleep apnea along with another respiratory syndrome (e.g., chronic obstructive pulmonary disease). In other words, the MESAM 4 was unable to identify patients that did not fit the programmed definition of sleep apnea. This study highlights the current limitations of computers and the need for humans to be intimately involved in setting criteria for event identification and evaluating the reliability of the computerized measures obtained.

CONCLUSIONS

Advances in technology have contributed to major insights into the mechanisms and processes of sleep. However, computer programs can be written to recognize only what we understand and can clearly define (14). At present, reliable completely automated sleep scoring systems are not available. The EEG variables present particular difficulties for automated sleep staging systems. However, the ability of computers to store large amounts of data, to organize and consolidate these data, and to perform summary statistics has allowed the polysomnogram to become a significant clinical and research tool. It provides clinicians and scientists with increased power for the

evaluation and understanding of sleep pathology, as well as the basic mechanisms underlying sleep itself.

BIBLIOGRAPHY

1. McNish R. The Philosophy of Sleep, 1st ed. New York: D. Appelton & Company; 1934.
2. Kleitman N. Sleep and Wakefulness. Chicago: The University of Chicago Press; 1939.
3. Aserinsky E, Kleitman N. Regularly occurring periods of eye motility, and concomitant phenomena, during sleep. Science 1953;118:273–274.
4. Aserinsky E, Kleitman N. Two types of ocular motility occurring in sleep. J Appl Physiol, 1955;8:11–18.
5. Dement W, Kleitman N. Cyclic variations in EEG during sleep and their relation to eye movements, body motility, and dreaming. Electroencephalogr Clin Neurophysiol 1957;9:673–690.
6. Hobson JA. Sleep. New York: Scientific American Library; 1989.
7. Dement WC. Normal sleep and its variations. In: Kryger MH, Roth T, Dement WC, editors. Principles and Practice of Sleep Medicine, 2nd ed. Philadelphia: W.B. Saunders; 1994; p 3–15.
8. American Electroencephalographic Society. Guideline fifteen: Guidelines for polygraphic assessment of sleep-related disorders (polysomnography). J Clin Neurophysiol 1994;11:116–124.
9. Bartolo A, Clymer BD, Golish JA, Burgess RC. The polysomnogram assay: a method to represent the overnight polysomnogram in a condensed format. Comput Biomed Res 2000; 33:110–125.
10. Rechtschaffen A, Kales A. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. University of California: Brain Information Service; 1968.
11. Siegel J. The neural control of sleep and waking. New York: Springer-Verlag; 2002.
12. Hauri PJ. The Sleep Disorders. Kalamazoo (MI): The Upjohn Company; 1982.
13. Agarwal R, Gotman J. Computer-assisted sleep staging. IEEE Transactions on biomedical engineering 2001;48: 1412–1423.
14. Kubicki S, Herrmann WM. The future of computer-assisted investigation of the polysomnogram: Sleep microstructure. J Clini Neurophysiol 1996;13:285–294.
15. Agarwal R, Gotman J. Digital tools in polysomnography. J. Clin Neurophysiol, 2002;19(2):136–143.
16. Moser NJ, Phillips B, Berry DT, Harbison L. What is hypopnea anyway? Chest 1994;105(2):426–428.
17. Broughton RJ, Hasan J. Quantitative topographic electroencephalic mapping during drowsiness and sleep onset. J Clin Neurophysiol 1995;12:372–386.
18. Jankel WR, Niedermeyer E. Sleep spindles. J Clin Neurophysilo 1985;2:1–35.
19. Martens WLJ, Declerck AC, Kums DJThM, Wauquier A. Considerations on a computerized analysis of long-term polygraphic recordings. In: Stefan H, Burr W, editors. EEG monitoring. Stuttgart: Gustav Fisher; 1982; p 265—274.
20. Badia P, Wright KP, Wauquier A. Fluctuations in Single-Herz EEG activity during the transition to sleep. In: Ogilvie RD, Harsh JR, editors. Sleep Onset: Normal and Abnormal Processes. Washington (DC): American Psychological Association; 1995. p 201–218.
21. Hasan J. Differentiation of normal and disturbed sleep by automatic analysis. Acta Phys Scand(Suppl) 1983;526:1–103.
22. Hasan J, Hirvonen K, Värri A, Häkkinen V, Loula P. Validation of computer analysed polygraphic patterns during drowsiness and sleep onset. Electroencephalogr Clin Neurophyso 1993;87:117–127.
23. Toussaint M, et al. Changes in EEG power density during sleep laboratory adaptation. Sleep 1997;20(12):1201–1207.
24. Carrier J, et al. The effects of age and gender of sleep EEG power spectral density in the middle years of life (ages 20–60 years old). Psychophysiology 2001;38:232–242.
25. Ehlers CL. Kupfer DJ. Slow-wave sleep: Do young adult men and women age differently? J Sleep Res 1997;6(3):211–215.
26. Huupponen E, et al. A study on gender and age differences in sleep spindles. Neuropsychobiology 2002;45(2):99–105.
27. Hirshkowitz M, Moore CA. Issues in computerized polysomnography. Sleep 1994;17(2):105–112.
28. Tan X, Campbell IG, Feinberg I. Internight reliability and benchmark values for computer analyses of non-rapid eye movement (NREM) and REM EEG in normal young adult and elderly subjects. Clin. Neurophysiol 2001;112:1540–1552.
29. Keshavan MS, et al. Delta sleep deficits in schizophrenia: evidence from automated analyses of sleep data. Arch Gen Psychia 1998;55(5):443–448.
30. Keshavan MS, Reynolds CF, Miewald JM, Montrose DM. A longitudinal study of EEG sleep in schizophrenia. Psychiat Res 1996;59:203–211.
31. Limoges E, et al. Atypical sleep architecture and the autism phenotype. Brain 2005;128(Pt5):1049–1061.
32. Barry RJ, et al. Age and gender effects in coherence: II. Boys with attention deficit/hyperactivity disorder. Clin Neurophysiol 2005;116(4):977–984.
33. Cirignotta F, et al. Unreliability of automatic scoring of MESAM 4 in assessing patients with complicated obstructive sleep apnea syndromexd Chest. 2001;119(5):1387–1392.

See also ANALYTICAL METHODS, AUTOMATED; ANESTHESIA, COMPUTERS IN; SLEEP LABORATORY.

SPECT. See COMPUTED TOMOGRAPHY, SINGLE PHOTON EMISSION.

SPECTROFLUORIMETRY. See FLUORESCENCE MEASUREMENTS.

SPECTROPHOTOMETRY. See COLORIMETRY.

SPEECH REHABILITATION. See LARYNGEAL PROSTHETIC DEVICES.

SPINAL CORD STIMULATION

TERRENCE L. TRENTMAN
KENT P. WEINMEISTER
Mayo Clinic Scottsdale
Scottsdale, Arizona

INTRODUCTION

Spinal cord stimulation (previously known as dorsal column stimulation) is a minimally invasive technique used primarily to treat chronic, refractory neuropathic pain. It is

based upon Melzack and Wall's gate control theory (1), and was first introduced by Shealy in 1967 (2). Neuropathic (nerve injury) pain has many etiologies, including trauma, stroke, diabetes, infection [e.g., human immunodeficiency virus (HIV), or shingles], and cancer. Unfortunately, nerve injury pain can be extremely difficult to manage. Many types of therapy have been used for neuropathic pain including medications such as antiinflammatories, opiates, and antiepilepsy drugs. Physical therapy and psychologically based approaches have also been tried with variable success. Spinal cord stimulation (SCS), transcutaneous electrical nerve stimulation (TENS), and peripheral nerve stimulation (PNS) are all forms of neuromodulation that are used for nerve injury pain.

Spinal cord stimulation is typically reserved for patients with refractory neuropathic pain, whereas deep brain stimulation is currently used for patients with movement and some pain disorders. In SCS, a lead is percutaneously inserted into the epidural space, and an electric field is applied in the vicinity of the spinal cord. The electric field depolarizes neural elements or in some way modifies the function of the nervous system. The goal is for the patient to experience a pleasant paresthesia, often described as "tingling", in the area of their pain. After an initial successful trial, a permanent stimulator can be implanted that the patient controls with a hand-held device.

This article reviews the equipment used in spinal cord stimulation, patient selection, and the possible mechanisms of this therapy. The process of inserting a stimulator will be described with possible complications, and the effectiveness of this therapy will be analyzed. Finally, possible future uses of SCS will be discussed.

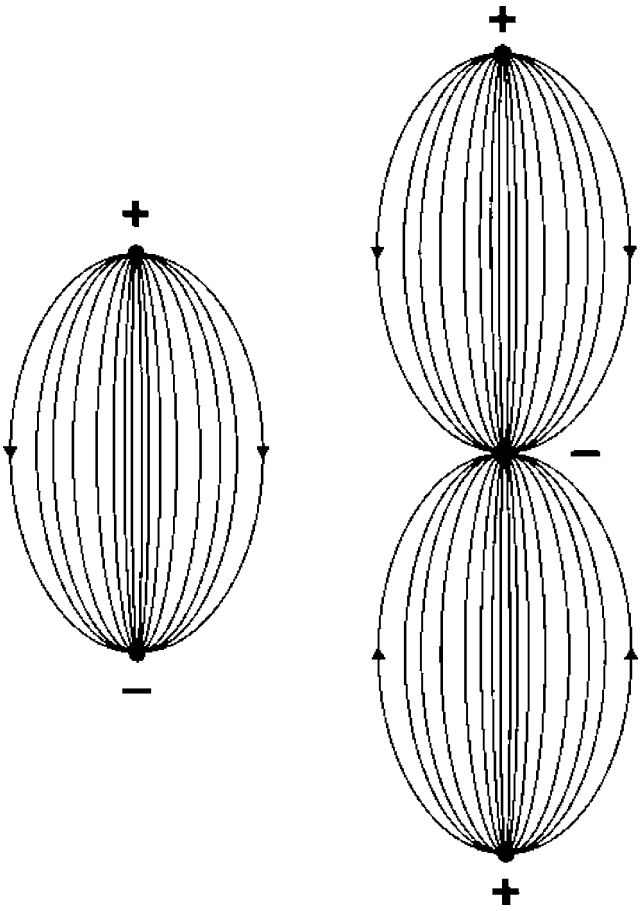


Figure 2. Current distribution between electrodes on SCS leads.

EQUIPMENT

Medtronic, Inc. (Minneapolis, MN), Advanced Neuromodulation Systems, Inc. (Allen, TX) and Advanced Bionics (Valencia, CA) are the primary manufacturers of spinal cord stimulators. There are two types of implantable leads available: the paddle (surgical) lead and the tubular percutaneous lead (see Fig. 1). The percutaneous lead can have four or eight contact points (electrodes), whereas the surgical leads are available with two, four, or eight electrodes. Each electrode can be programmed to function as an anode or cathode for the electrical current used in stimulation (see Fig. 2). The paddle lead is shielded on one side, such that stimulation is produced only on the side with the electrodes (see Fig. 3). This finding has the advantage of directing the entire electrical field toward the spinal cord, as opposed to the percutaneous lead that produces an electrical field circumfrentially around the lead, including away from the spinal cord. Hence, the paddle lead can produce SCS at lower amperage, prolonging battery life. The surgical lead also has the potential advantage of greater stability (less likely to move postimplantation) as it is sutured to surrounding tissue (3). However, the paddle lead requires a minilaminotomy, whereas the percutaneous lead is placed less traumatically via a 15 gauge touhy needle.

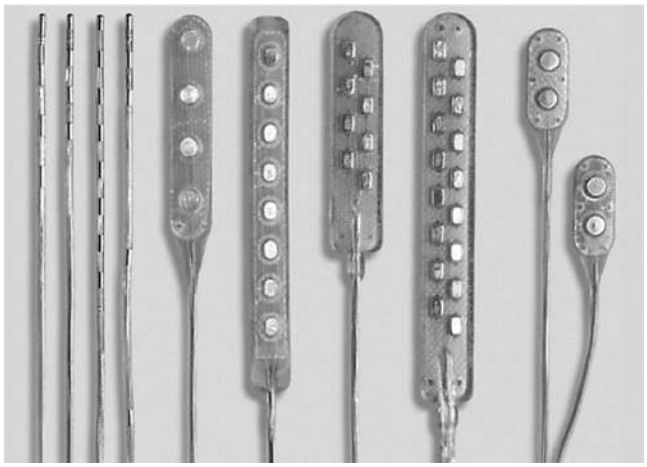


Figure 1. Various percutaneous and paddle leads. (Used courtesy of ANS, Inc.)

The percutaneous lead is made of inert polyurethane with an outside diameter of ~1.3 mm. On its distal end, it has four or eight electrodes made of platinum iridium. These are spaced 4, 6, or 12 mm apart. The electrodes are 3–6 mm long. The plate lead has a two- or four-midline circular or eight parallel rectangular electrodes. There are several options to provide current to the electrodes. An implantable pulse generator (IPG) can be placed subcutaneously (usually in the low abdomen), similar to a

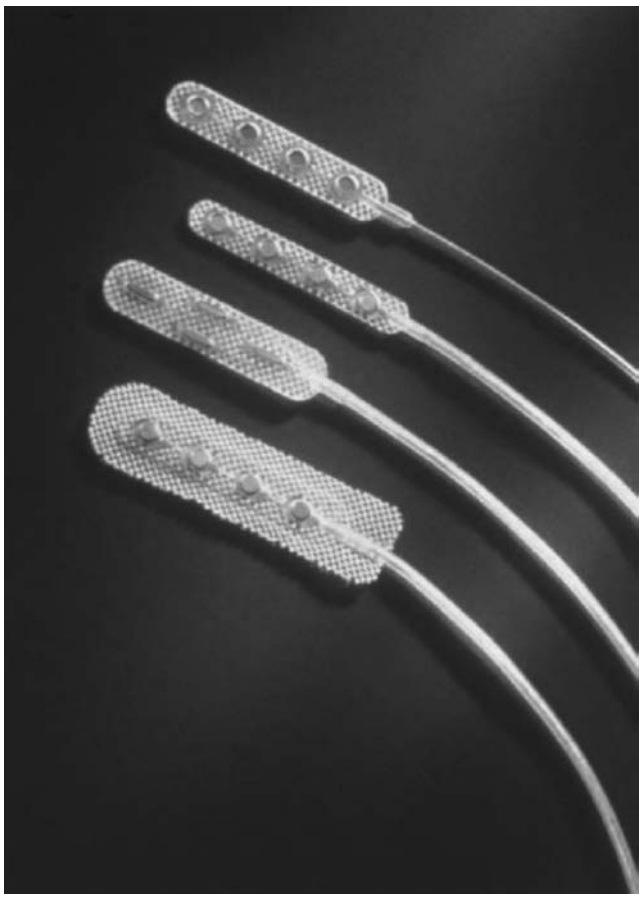


Figure 3. Examples of surgical (paddle) leads. (Used courtesy of Medtronic, Inc.)

pacemaker generator. Recently, a rechargeable IPG has become available. Depending on use, the Pt will percutaneously recharge the IPG every few days to weeks. Finally, a radio frequency (rf) receiver can be placed subcutaneously and powered by an external rf transmitter coil that is held over the device (see Fig. 4). In either case, a cable is tunneled subcutaneously from the power source to the lead in the spine.

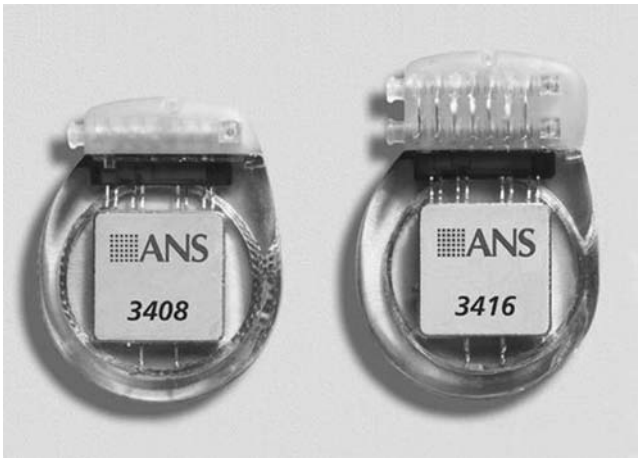


Figure 4. Implantable rf receivers. (Used courtesy of ANS, Inc.)

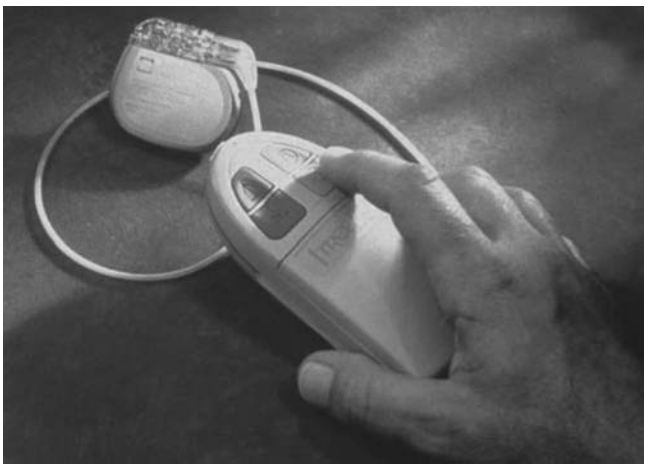


Figure 5. Implantable IPG with hand-held controller. (Used courtesy of Medtronic, Inc.)

The IPG provides pulses of electrical current that are rectangular and biphasic. Programmable features include pulse width, amplitude, and rate. Rates generally used are 30–80 Hz, amplitude is 0–12 V, and pulse width is 200–450 μs. Battery life varies depending on rate, amplitude, and time the device is used. Unless the patient uses the stimulator constantly with high rate and amplitude, the battery should last at least several years. One obvious advantage of an rf receiver system is that there is no need to periodically replace the battery. The energy source of the IPG is a hermetically sealed silver vanadium oxide cell. The power source and electronics are sealed in an oval shaped titanium shield (see Fig. 5).

PATIENT SELECTION

As noted earlier, spinal cord stimulators are used primarily for neuropathic pain. Although there are many types of neuropathic pain, chronic unilateral lower extremity neuropathic pain seems to respond best to this type of therapy. A typical patient may have the failed back surgery syndrome with residual leg pain, or the patient may have some type of neural compressive lesion that is not operable and refractory to medical management. Spinal cord stimulation may also be indicated for chronic arachnoiditis, complex regional pain syndrome, peripheral neuropathy of the lower limb, and phantom limb syndrome. Patients with idiopathic pain, mechanical low back pain, or other forms of nociceptive pain have a lower success rate when compared to those with neuropathic conditions. Clinical experience and studies have shown that SCS is most efficacious when the entire painful area is covered with paresthesia. The more diffuse the patient's pain, the more difficult it will be to cover with SCS.

Selection for SCS often includes psychological screening. Patients with untreated depression, anxiety, or drug abuse issues are not good candidates. Obviously, the patient needs to be able to understand how to use the stimulator. A trial of stimulation is probably the best predictor of long term success (4). Although the value of

psychological testing in predicting success with SCS is controversial, there is no question that patients with chronic pain are best managed with a multidisciplinary approach. This may include physical therapy, psychological and spiritual support, medications, and surgical procedures as indicated.

MECHANISMS

Both animal and human research studies have provided a partial understanding of the mechanisms of SCS (5). Melzack and Wall's gate control theory (1) suggested that stimulation of large cutaneous A-β fibers would inhibit nociceptive input from the smaller A-δ and C fibers. Since SCS has been shown to be more effective for neuropathic pain than nociceptive pain, the mechanism must include more than simple inhibition of nociceptive input. Endorphins or other endogenous opiates do not seem to be involved. In patients with ischemic lower extremity pain or refractory angina, the mechanism of SCS appears to be an increased local blood flow (i.e., microcirculation). This may be due to both inhibition of the sympathetic nervous system and activation of vaso-active chemicals (6).

Animal studies have supported the contention that A-β fiber stimulation is one of the mechanisms of SCS. Animal models of neuropathic pain can be created by lesioning the sciatic nerve, which creates tactile allodynia in the animal, a phenomenon mediated by A-β fibers. Spinal cord stimulation has been seen to suppress this sign. Another effect of SCS is on wide-dynamic range neurons. Wide-dynamic range (WDR) neurons are second order neurons in the dorsal horn of the spinal cord. They receive input from a variety of sensory neurons. In the face of continuous stimulation from injured neurons, the WDR neurons will "wind-up," that is, fire at lower depolarization thresholds. Spinal cord stimulation may decrease this WDR response while simultaneously decreasing the central excitatory neurotransmitters glutamate and aspartate. γ-Aminobutyric acid (GABA), a central inhibitory neurotransmitter, is simultaneously released; therefore, SCS may have beneficial effects on both excitatory and inhibitory pain mechanisms (7).

Recent computer modeling of SCS has led to a greater theoretical and empirical understanding of the interaction of current with spinal structures (8). These models demonstrate how the depth of cerebral spinal fluid and the distance of the electrodes from both the dorsal columns and dorsal roots can affect the patient's paresthesia perception.

IMPLANTATION TECHNIQUE

Before a spinal cord stimulator is implanted, the patient needs to be informed of potential risks. These include infection, bleeding, nerve damage, allergic reaction, and failure of the stimulator to adequately cover or reduce the patient's pain. The patient may experience swelling around the site of the generator and a seroma may develop requiring drainage. If the lead or the generator becomes infected, it may have to be removed. Lead displacement, fracture, or movement can occur such that an initially adequate pattern of stimulation becomes inadequate. Lead and battery revision may become necessary at some point. After placement of a SCS, the patient is instructed not to drive an automobile with the device turned on. Furthermore, they should not undergo a magnetic resonance imaging (MRI) scan or any type of diathermy.

Once consent has been obtained, a trial of SCS is performed. This consists of placing a trial lead in the epidural space, and if adequate coverage of the patient's area of pain can be achieved, allowing the patient to use the stimulator on an outpatient basis for 5–7 days. In 1993, Barolat et al. published a database of 106 patients in whom they had placed spinal cord stimulator leads (9). The electrodes were placed between the C1 and L1 spinal levels for chronic pain management, and the areas the patients felt stimulation were mapped. These maps provide a guideline as to which body areas will be stimulated by implanted electrodes. Barolat also noted that certain body areas were difficult to cover with paresthesia, including the low back, neck, and perineum. Clinical experience has shown that patients with bilateral extremity pain or pain in both the low back and legs may require bilateral lead placement to obtain adequate coverage. The placement of more than one lead in the epidural space allows not only wider paresthesia coverage, but also the use of complex stimulation programs that can be tailored to meet the patient's needs (see Fig. 6). With bilateral eight electrode leads, the possible stimulation combinations (anodes and cathodes) reach the thousands.

To insert the lead, the patient is placed in the prone position and sedated. The operative area is sterilely prepped and draped, and the skin is anesthetized with local anesthetic. When treating lower extremity pain, the puncture site is usually at the L1-2 level. The epidural space is entered with a touhy needle, through which a lead is advanced in a cephalad direction. Using fluoroscopy, the lead is observed to move up the spinal canal until it reaches approximately the T9–T10 level. The lead can be manipulated to direct it slightly to the side corresponding to the patient's pain. For upper extremity pain, the skin is usually punctured at T1–T2, and the tip of the lead is placed at the C3–C4 level. The presence of scar tissue or other anatomic barriers can make lead placement difficult and occasionally impossible.

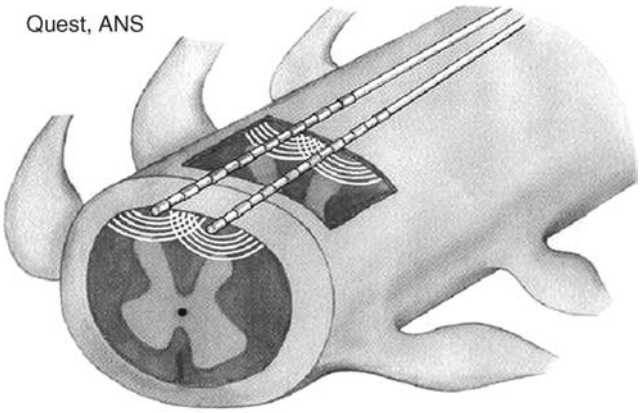


Figure 6. Dual leads allow winder parasthesia coverage. (Used courtesy of ANS, Inc.)

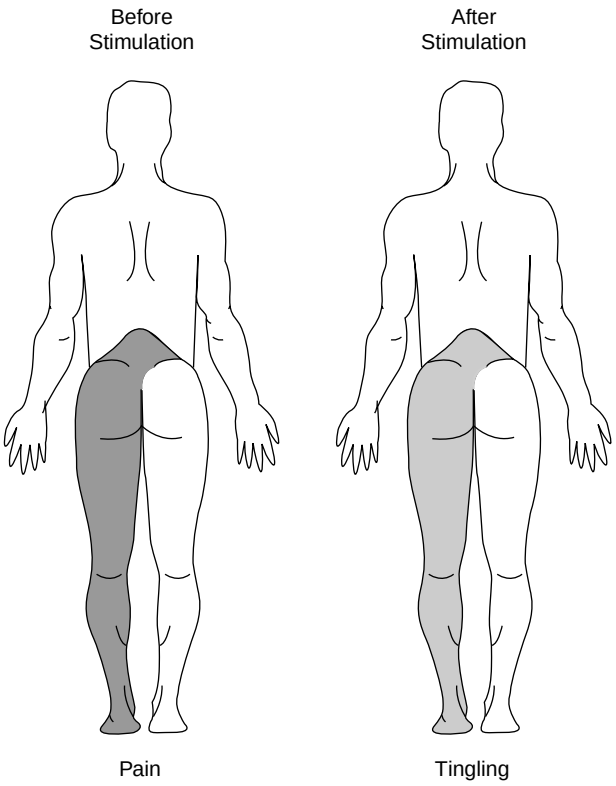


Figure 7. SCS covers painful area with pleasant paresthesias. (Used courtesy of Medtronic, Inc.)

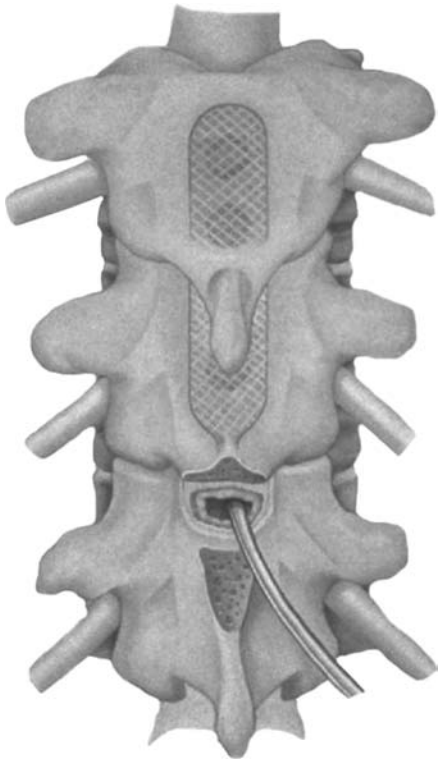


Figure 8. Paddle lead surgically inserted into the epidural space. (Used courtesy of Medtronic, Inc.)

Once the lead is felt to be in proper position, it is attached via a cable to the trial generator. This is an external programmer that allows various combinations of electrodes to be stimulated in an effort to cover the patient's pain with pleasant paresthesia, usually described as "tingling" (see Fig. 7). If the amplitude is set too high, the patient may experience discomfort or muscle stimulation. The patient must be awake enough at this point to answer questions and describe where they feel the stimulation. It is not unusual to need to adjust the position of the lead(s) several times before adequate coverage is obtained.

Once adequate coverage is obtained, the trial lead is secured with tape and/or suture. After recovery from anesthesia, the patient is given instructions as how to operate the stimulator. The patient is allowed to turn the device on or off, and can adjust the amplitude and rate to comfort. Reprogramming the pulse width and lead combinations is generally reserved for the pain specialist. The patient is told not to drive a car with the stimulator turned on, and excess twisting or raising the arms above the head is discouraged. As noted above, the patient will return to the clinic for removal of the trial lead in 5–7 days; however, the patient is encouraged to call sooner should anything change with the function of the device.

During the follow-up visit, several decisions are made. The patient is asked if the stimulator continued to cover their painful area, and if so, did it reduce the discomfort. Ideally, the patient obtained at least a 50% reduction in their pain during the trial. If the patient has received significant pain relief and they want to proceed with permanent implantation, the type of lead (surgical vs. percutaneous) is selected. Placement of the surgical lead requires a minilaminotomy, which includes removal of part of the inferior portion of the lamina and a portion of the *ligamentum flavum*, followed by insertion of the paddle lead into the exposed epidural space (see Fig. 8). Villavicencio et al. followed 27 patients who underwent placement of SCS leads (3). Patients who had electrodes placed via a laminectomy had significantly better long-term effectiveness than the patients with percutaneous leads. Nonetheless, permanent placement of percutaneous leads remains a viable and effective option that does not require a minilaminotomy. The placement of the permanent SCS lead requires that a generator (IPG) be inserted in a subcutaneous pocket, usually in the low abdomen or over the buttock. A cable is tunneled under the skin to the lead inserted in the spine. After permanent implantation, the spinal cord stimulator is controlled with a hand-held device.

OUTCOMES

A number of studies have looked at outcomes after SCS. Van Buyten et al. described 10 years of experience with SCS in 254 patients, 217 of whom had permanent stimulators placed (10). Before the study began, 10% of the patients had died and another 10% had undergone explantation. Reasons for explantation included ineffectivity (4.6%), infection, allergy, and recovery from pain. An independent review of the remaining patients who could be contacted and would participate in the study ($n = 123$) showed that 68% of them graded the treatment as excellent to good (excellent, very good, good, moderate, weak, no

improvement, worse). After excluding retirees and others not pursuing a career, 31% of the patients who had been working before their pain began had returned to work. The authors noted that their success rate is one of the highest reported.

Kay et al. published another retrospective study of SCS covering 13 years (11). Of 70 patients treated with SCS, there were 72 surgical revisions, including electrode (32), connecting cable (6), or generator revision (22). Battery depletion was the single most common indication for generator revision (16/22). Of the 72 revisions, 12 were for explantation. Of 48 patients who responded to a questionnaire, 60% rated their pain relief as substantial (>50%).

Bhdrakant et al. prospectively studied 29 patients over a 2-year period (12). The primary indication for SCS was failed back syndrome. Four of the 29 failed to obtain relief during the SCS trial. Of the 25 patients with permanent implants, SCS was beneficial in 50%. This result is similar to North et al. results for a large series ($n = 320$), where 52% of patients reported at least 50% pain relief (13).

Although these and other studies support the use of SCS for certain chronic pain syndromes, methodological problems preclude drawing final conclusions. Its retrospective nature, lack of controls, and heterogeneous patient populations flaw much of the research on SCS. More prospective studies, perhaps looking at SCS for individual pain syndromes, will be needed before this expensive technology becomes widely accepted.

FUTURE USES OF SCS

Greater understanding of both peripheral and central pain mechanisms combined with evolving technology have expanded the potential uses of SCS. Multiple-electrode configurations have allowed coverage of diffuse pain generators and made reprogramming simple when coverage is lost or new pain symptoms arise. Lumbosacral placement of SCS leads may allow treatment of refractory pelvic neuropathic conditions including sacral neuralgia, vulvodynia, or coccydynia. Urinary incontinence may also be treatable with this technique (14). Peripheral placement of SCS leads has been used for a number of conditions, including occipital neuralgia (i.e., spinally transformed migraine) and trigeminal neuralgia (15,16).

Other current and evolving uses for SCS include chronic regional pain syndromes (RSD and causalgia), postherpetic neuralgia, and postamputation pain. Patients with peripheral vascular disease suffering from rest and night pain seem to benefit from SCS; this indication is used more commonly in Europe than the United States. Spinal cord stimulation has also been shown to be effective in refractory angina (17). Other conditions treated with SCS include severe Raynauds phenomena, Buerger's disease, and diabetic neuropathy.

BIBLIOGRAPHY

1. Melzack R, Wall P. Pain mechanism: A new theory. Science 1965;150:951–979.
2. Shealy C, Mortimer J, Reswick J. Electrical inhibition of pain by stimulation of the dorsal columns: Preliminary clinical report. Anesth Analg 1967;46:489–491.
3. Villavicencio AT, Leveque J, Rubin L, Vulsara K, Gorecki JP. Laminectomy versus percutaneous electrode placement for spinal cord stimulation. Neurosurgery 2000;46:399–406.
4. Barolat G, Ketcic B, He J. Long term outcome of spinal cord stimulation for chronic pain management. Neuromodulation 1998;1:19–29.
5. Oakley J, Prager J. Spinal cord stimulation, mechanisms of action. Spine 2002;27:2574–2583.
6. Kumar K, Toth C, Nath RK, Verma AK, Burgess JJ. Improvement of limb circulation in peripheral vascular disease using epidural spinal cord stimulation: a prospective study. J Neurosurg 1997;86:662–669.
7. Meyerson BA, Linderhoth B. Mechanisms of spinal cord stimulation in neuropathic pain. Neurolog Res 2000; 22:285–292.
8. Alo KM, Holsheimer J. New trends in neuromodulation for the management of neuropathic pain. Neurosurgery 2002; 50:690–704.
9. Barolat G, Massaro F, He J, Zeme S, Ketcik B. Mapping of sensory responses to epidural stimulation of the intraspinal neural structures in man. J Neurosurg 1993;78:233–239.
10. Van Buyten J, Zundert JV, Vueghs P, Vanduffel L. Efficacy of spinal cord stimulation: 10 years of experience in a pain centre in Belgium. Eur J Pain 2001;5:299–307.
11. Kay AD, McIntyre MD, Macrae WA, Varma TRK. Spinal cord stimulation—a long–term evaluation of patients with chronic pain. Bri J Neurosurg 2001;15(4):335–341.
12. Bhadrakant K, Rosenfeld JV, Hutchinson A. The efficacy of spinal cord stimulation for chronic pain. J Clin Neuro Sci 2000;7(5):409–413.
13. North RB, Kidd DH, Zahurak M, James CS, Long DM. Spinal cord stimulation for chronic, intractable pain: Experience over two decades. Neurosurgery 1993;32(3):384–394.
14. Alo KM, Gohel R, Corey CL. Sacral nerve root stimulation for the treatment of urge incontinence and detrusor dysfunction utilizing a cephalocaudal intraspinal method of lead insertion: A case report. Neuromodulation 2001;4(2): 53–58.
15. Lou L. Uncommon areas of electrical stimulation for the relief of pain. Curr Rev Pain 2000;4:407–412.
16. Weiner L, Reed KL. Peripheral neurostimulation for control of intractable occipital neuralgia. Neuromodulation 1999; 2(3):217–221.
17. DeJongste MJL. Spinal cord stimulation for ischemic heart disease. Neurolog Res 2000;22:293–298.

See also BIOELECTRODES; BLADDER DYSFUNCTION, NEUROSTIMULATION OF; ELECTRONEUROGRAPHY; FUNCTIONAL ELECTRICAL STIMULATION; PERIPHERAL VASCULAR NONINVASIVE MEASUREMENTS; TRANSCUTANEOUS ELECTRICAL NERVE STIMULATION (TENS).

SPINAL IMPLANTS

MICHELE MARCOLONGO
Drexel University
Philadelphia, Pennsylvania

ABHIJEET JOSHI
Abbott Spine
Austin, Texas

INTRODUCTION

Spinal implants constitute the fastest growing segment of the orthopedic medical device industry. The area has until

the last 5 to 10 years been vastly under-studied for the proportion of patients who are afflicted with diseases and injuries to the spine. Consequently, new spine medical devices and medical device companies are emerging every day with new and better treatment strategies for prevalent spine disorders. This article will explore the physiological conditions and disease states that require treatment and then demonstrate some treatment strategies that are being used today. The devices in this text are not comprehensive (we would need a much larger space to do that) but do allow an understanding of the state-of-the art in medical treatment of spinal disorders.

Human Spine

The human spine is a mechanical structure as it performs three fundamental biomechanical functions simultaneously (1). First, it transfers the weights (and resultant bending moments) of the head, trunk, and any weights being lifted to the pelvis. Second, it allows the sufficient physiological motion among the head, trunk, and pelvis. Third, and most important, it protects the delicate spinal cord from the potential damaging forces (and moments) resulting from the physiological motions and trauma (1).

Figure 1 show a schematic of the human spine, which is divided into three main regions: the upper region with 7 vertebrae (cervical spine), the middle region with 12 vertebrae (thoracic spine), and the lowermost with 5 vertebrae (lumbar spine). At the distal end of the spine, there is a basin-shaped structure, the pelvis, that supports the spinal column and is made of sacrum and coccyx with fused vertebrae. The human spine is not a straight structure, but it has specific curvature. The spine in the cervical and in the lumbar region is slightly convex anteriorly, whereas in the thoracic and sacral region, it is slightly convex posteriorly. The specific shape allows the increased flexibility while maintaining the overall spinal stability. It also facilitates increased shock-absorbing capacity along with adequate stiffness (1).

Each vertebra is made up of several parts. Figure 2 shows schematic of the vertebrae in a vertebral column. The body of the vertebra is the primary weight-bearing area. Between the vertebrae lie the intervertebral disks, which separate the adjacent vertebrae and act as cushions between them while allowing the movement of one vertebra relative to another. There is a large hole in the center part (spinal canal) that is covered by the lamina. The spinal cord runs through this spinal canal. There is a protruded bone in the central posterior region, called the spinous process. There are pairs of transverse processes that are orthogonal to the spinous process and provide attachment sites for the back muscles. Four facet joints are also associated with each vertebra. Four facet joints in two pairs (superior and inferior) interlock with adjacent vertebrae and provide the stability to the spine (1). An intervertebral disk is situated in between adjacent vertebrae. The disks are labeled with respect to the vertebrae levels, between which they are located. Thus, the T12/L1 disk is located between the twelfth thoracic and first lumbar vertebrae, whereas the L3/L4 disk is located between the third and fourth lumbar vertebrae.

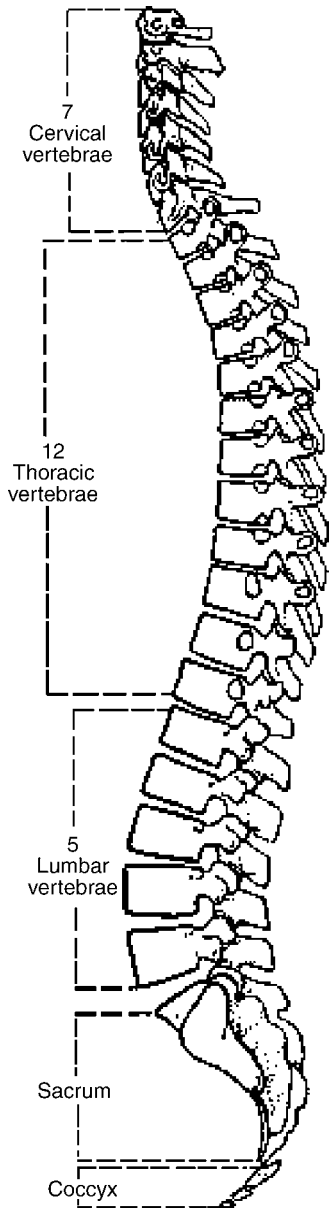


Figure 1. Schematic of human spine (2).

The intervertebral disk is basically a composite structure made up of three different tissues; the central core is called the nucleus pulposus (Fig. 3), which is attached radially to the multilayered fibers of the annulus fibrosus and attached superiorly and inferiorly to cartilaginous end plates (1). The nucleus is predominantly water in a matrix of proteoglycan, collagen, and other matrix proteins. The water content of the nucleus is very high at birth (approximating 90%) and then decreases through the aging cycle down to 70% or less. The annulus surrounds the nucleus with successive layers of tissue with collagen fibers oriented in alternating directions. The annulus is under tension when the nucleus absorbs water and swells. The cartilaginous end plates have multiple perforations that allow exchange of water and nutrients into the disk (4–6).

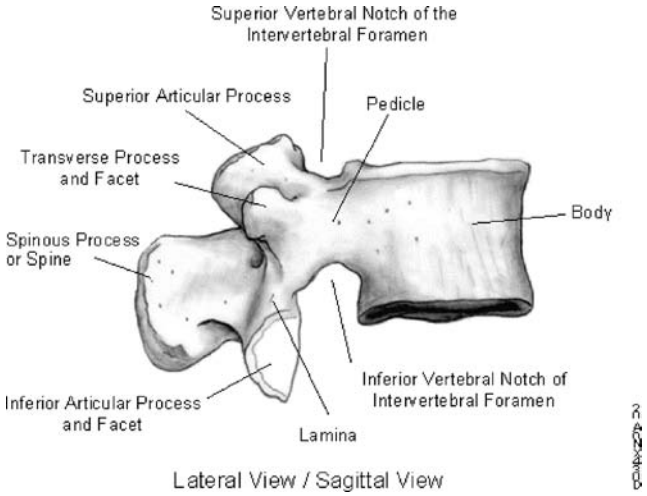


Figure 2. Spinal vertebrae (3).

SPINAL DISORDERS

Various spinal disorders are observed in humans; some manifest in pediatric patients, whereas others affect middle-aged and older patients. The most common spinal disorders can be generally described by three different categories: developmental bone deformities (scoliosis and kyphosis), bone degeneration (vertebral compression fractures), and degenerative disk disease (herniation, rupture and spinal cord stenosis).

Developmental Bone Deformities

The spine is a complex biomechanical structure and performs complex functions. This puts spine under greater strains, and bone deformities may develop during the course of time while performing the demanding functions such as supporting the cranium and trunk, or absorb stresses generated during daily physiological activities. Most common spinal disorders under this category include scoliosis and kyphosis.

Scoliosis. Scoliosis is a lateral curvature of the spine. The symptoms of the scoliosis include uneven waist, different height shoulders, raised/prominent hip, and leaning of body to one side (7). Some causes of scoliosis include congenital deformity, cerebral palsy, atrophy, and neuromuscular problems.

Scoliosis can either be structural or functional. Structural scoliosis is referred in case of adjacent vertebrae rotation upon each other. This is generally followed by deformation of rib cage. In case of functional scoliosis, there is no fixed vertebral rotation or fixed deformity in the thoracic region. The rate of curve progression is not constant; however, the lumbar curves progresses more rapidly than thoracic curves. The scoliosis is generally classified as adolescent idiopathic scoliosis (AIS), adult scoliosis (with or without degenerative changes), and *de novo* scoliosis (which develops secondary to degenerative changes of the lumbar spine, especially in older age) (7).

The most common tools used for diagnosis of scoliosis include plane radiograph, computed tomography (CT) scans, and magnetic resonance imaging (MRI). Treatment options depend on the various factors, including the age, curvature angle, progress rate, location, flexibility, and spinal maturity. Conservative management (no treatment) is commonly incorporated when the curvature is mild (less than 20°).

Orthopedic braces are recommended in case of curvature angle of 25–40°, to prevent further spinal deformity, especially in children. The bracing, however, merely prevents the worsening of the existing curvature and does not restore normal alignment (8). Many types of braces are commercially available in the market. The brace, depending on the type and application, may extend from neck to pelvis (Milwaukee Brace) with plastic pelvic girdle, neck ring, and pressure pads (Fig. 4) (9), or it may just cover below the breast to the initial pelvic region only (Boston Brace). The use of braces has been generally effective in case of children, to prevent the further worsening of the scoliosis, but there is still a lack of consensus about the indications for the brace, type, and wearing time over the body.

Surgical options are used only in case of severe scoliosis (curvature angle greater than 45°) or for the curves that do not respond to nonsurgical treatments. The goals of the

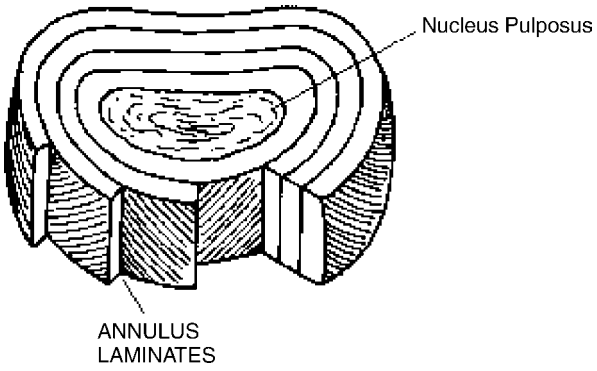


Figure 3. Schematic of an intervertebral disk (1).

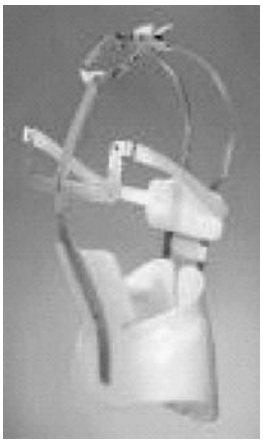


Figure 4. Milwaukee Brace (9).

surgical treatment are to prevent the progression of the curve and correct the deformity using instrumentation (7). The most common method to treat severe scoliosis is spinal fusion (anterior or posterior) and bone grafting/substitute. Bone graft can either be autologous iliac crest, from rib or allograft. In general, the facetctomy is followed by placement of bone graft and fixation. Various instrumentation options are available to surgeons for fixation. Basically, fixation is achieved by use of single/dual rods, posteriorly or anteriorly, along with screws and wires to the fixation point(s). The structures of the vertebral body, such as pedicles, sublaminar region, facets, and processes, may serve as fixation points for fusion (7). Anterior structural support is generally provided by mesh cages or ring allografts. Single thoracic curves are generally treated posteriorly using posterior instrumentation (hooks) and fusion.

Kyphosis. When viewed from the side, the normal spinal column is not completely straight. There are several gentle curves due to the shape and alignment of the vertebrae. Kyphosis is an exaggerated curvature of the spine or a rounded, "hunched" back. Most causes (metabolic, neuromuscular conditions, osteogenesis, spina bifida, among others) of the kyphosis are due to shortening of the anterior column, a weakening or lengthening of the posterior column, or both (7). The symptoms of kyphosis include difference in shoulder heights, forward bend of head compared with the rest of body, and tight hamstring (back thigh) muscles.

As in scoliosis, plane radiographs are also useful in diagnosing kyphosis. These help in defining the nature of sagittal deformities. Cobb (angle) measurements on these radiographs are performed to quantify the deformity in the sagittal and coronal plane. The angle is measured using the adjacent vertebral endplates (plane) as the basis of calculation. CT scans and MRI also find a place as useful diagnostic tools for better assessment of the spinal deformity.

The use of braces is recommended when the curve angle is between 40° and 60° on X ray. Surgical treatment is recommended when the curvature deformity is progressive, and the deformity may lead to neurological compromise. Again, spinal fusion (anterior or posterior) is referred for cases of severe deformity. In the case of young patients, posterior fusion might be considered, which would allow continuous anterior growth to partially correct the deformity (anterior release with posterior instrumentation).

Bone Degeneration

Compression fractures are generated in vertebrae when the bone tissue becomes weak due to degenerative changes. In most cases, the cause of the compression fracture is reduction in bone mineral density leading to weakening of bone (osteoporosis) (1,7). Osteoporosis causes both inorganic and organic phase bone loss. Loss of bone crystal weakens the bone to compressive loading, whereas loss of the organic matrix of bone makes the tissue more brittle, making the bony construct more susceptible to fracture. Other manifestations of osteoporosis include hyperkyphosis with chronic spinal pain and osteoporotic burst fractures. However, the most common manifestation of the bone loss is a vertebral compression fracture (VCF).

To diagnose vertebral compression fractures, plane radiographs are used. To follow the progression of bone density loss throughout the osteoporotic disease process, dual photon absorptiometry (DPA) (which measures axial skeletal bone mass density) and dual energy X-ray absorptiometry (DEXA) (which measures baseline bone density with precision) are used. Quantitative CT can also be used in diagnosis of compression fractures.

Surgical treatment available for reduction of compression fractures is vertebral body augmentation: either kyphoplasty or vertebroplasty (7). Vertebroplasty is a procedure performed to relieve the pain and strengthen the weak vertebrae. During the procedure, an image-guided (X-ray) bone needle may be passed through the patient's back to have precise control over its location. Bone cement (polymethylmethacrylate or PMMA) is pushed through the needle to stabilize the fractured location of the vertebra. After curing, the PMMA biomaterial serves to stabilize the vertebra and to minimize the pain associated with the fracture.

Kyphoplasty is another method of vertebral augmentation, which uses bipedicular approach and balloon tamps to create voids in the bone. The instrumentation (cannula) used in kyphoplasty differs that from used in vertebroplasty. The void created by balloon is filled with PMMA. Other materials, such as calcium phosphate, hydroxyapatite, polymeric hydrogels, and combinations thereof, are being investigated as an alternative solutions to PMMA.

Degenerative Disk Disease

Lower back pain is one of the most prevalent socioeconomic diseases and one of the most important health-care issues today. Over five million Americans suffer from lower back pain, making it the leading cause of lost work days next only to upper respiratory tract illness (10–14). On an average, 50–90% of the adult population suffers from lower back pain (15), and lifetime prevalence of lower back pain is 65–80% (16). It is estimated that 28% experience disabling lower back pain sometime during their lives, 14% experience episodes lasting at least 2 weeks, whereas 8% of the entire working population will be disabled in any given year (16). The total cost of the lower back disabilities is in the range of $50 billion per year in the United States (17) and £12 billion per year in the United Kingdom alone (18). The causes of lower back pain often remain unclear and may vary from patient to patient. It is estimated that 75% of such cases are associated with lumbar degenerative disk disease (DDD).

Many conservative treatment options exist for lower back pain. These generally aim at reducing the pain arising out of nerve root impingement and inflammatory response because of the migrated nucleus. The most commonly used surgical treatments include discectomy and spinal fusion and are sought when conservative treatments fail.

Progression of Degenerative Disk Disease. As the human life progresses, significant changes occur in the tissues of the intervertebral disk. DDD can be simply defined as the

loss of normal disk architecture accompanied by progressive fibrosis. At birth, the water content of the annulus fibrosus is about 80% and that of nucleus pulposus is about 90%. Through the degenerative process, this water content decreases to as low as 70% for the nucleus (19). Microscopic changes such as fragmentation of fibers, mucinous degeneration of fibers leading to cyst formation, and focal aggregation of the collagen to form round aggregates of amorphous material are observed in early stage of degeneration (20). The salient features of the DDD can be denoted as the loss of gelatinous nucleus pulposus, gradual disappearance of the originally well-defined border between the nucleus and the annulus, coarsening of the annulus lamellae, progressive fibrosis, and later fissuring of the annulus fibrosus with the deposition of the aging pigment (21–24).

The load transfer mechanism is clearly altered in the case of a dry nucleus. As a result, the end plates are subjected to reduced pressure at center and more pressure around the periphery. The stress distribution in the annulus is also altered significantly. Essentially, the nucleus does not perform its function of load transfer and the load transfer occurs through end plate—annulus—end plate route (1). The annulus is subjected to abnormal stresses and is more prone to injuries, and cracks/fissures first develop into the annulus.

With continued degeneration, the central nucleus may migrate through the crack developed in the annulus toward the periphery. The migration of the nucleus material is referred to as "disk herniation" (17). Approximately 90% of the disk herniation would occur at the L4-L5 and L5-S1 levels. The migrated material may impinge on the nerve root. The contact of the migrated nucleus with the nerve root irradiates debilitating back pain. Also, the herniated material elicits an inflammatory response because of the avascular nature of the nucleus. It is difficult to distinguish between the effects of aging from that of degeneration on the biomechanical behavior of the lumbar disk. The biomechanical behavior of the disk is dependent on its state of degeneration, which in turn depends on the age.

In case of the normal disk, any load acting on the disk is transferred to the annulus by means of swelling pressure (intradiscal pressure) generated by the nucleus (1). The water binding capability of the nucleus is a function of chemical composition of the nucleus. However, with aging and/or degeneration, changes occur to the proteoglycans as proteases and MMPs attack the molecules. The result is a decrease in the proteoglycan/collagen ratio, which leads to the lower water binding capability of the nucleus (25,26).

The load transfer mechanism in case of such a dehydrated disk is significantly altered (Fig. 5). The nucleus cannot generate sufficient intradiscal pressure to maintain disk height and normal mechanical function (25,27,28). Although it is not well understood, the consequence of the structural and mechanical changes to the disk may be a cause of lower back pain.

Stenosis. The reduction in the disk volume leads to instability, resulting in the growth of bone, end plates, and ligaments to compensate for this volume loss (stenosis). Stenosis is narrowing of the spinal canal (29). It occurs as a result of aging and/or degenerative disk disease. The water content of the nucleus decreases, causing an abnormal load transfer mechanism within the disk. The disk height is reduced, and this dry/hardened disk may bulge into the spinal canal space. Additionally, the facet joints may become thick, thus narrowing the spinal canal further. Spinal stenosis in the lumbar spine may result in cauda equina syndrome and loss of bowel and bladder function. In general, the symptoms are not observed with stenosis. However, when present, the symptoms may include low back stiffness, leg weakness, numbness in the back/legs, and cramping.

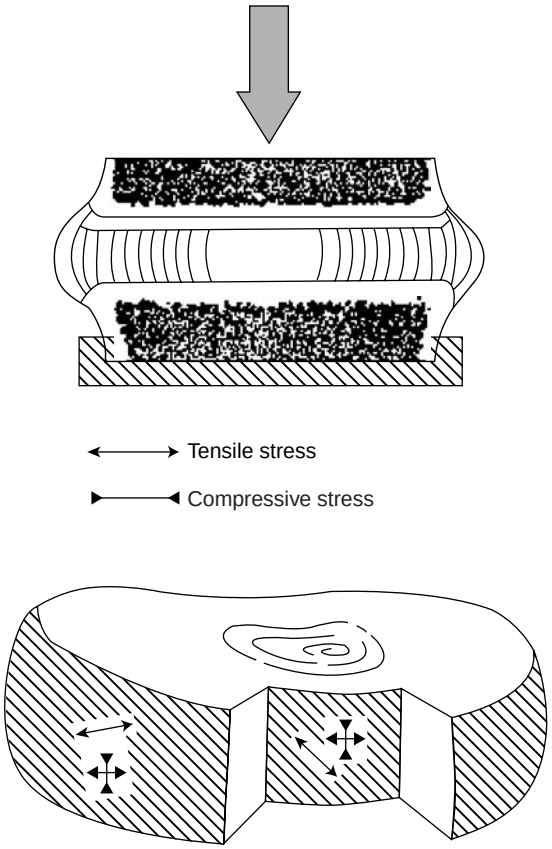


Figure 5. Degenerated disk (1).

Most common methods to diagnose and analyze the stenosis are plane X ray, MRI, and CT scan (7). These treatments, alone or in combination, provide valuable information about the patient's spine structure, location, and the extent of the disease. In particular, the following information can be revealed:

- Disk space narrowing
- Endplate osteophytes and sclerosis
- Facet enlargement and osteophytes formation
- Loss of lumbar lordosis

If conservative treatments such as medication, physical therapy, and spinal injections fail, a surgical approach may be recommended in the cases with persistent back pain

and/or progressive leg weakness. The indications for surgical treatment include radicular pain or neurogenic claudication with MRI or CT. In general, the goals of surgery are pain relief, increased mobility, and improvement in the patient's quality of life. Most common surgical treatments are laminectomy (in case of simple stenosis) and spinal fusion. Fusion is recommended when there is a stenosis in conjunction with

- Degenerative scoliosis or kyphosis
- Degenerative spondylolisthesis

The goal of the laminectomy or, lumbar decompression surgery, is to widen the spinal canal (30) to allow more space for spinal nerves. The treatment would ideally relieve the leg pain and, to a certain extent, back pain. When there is a vertebrae slippage relative to each other (spondylolisthesis), an abnormal motion would occur, which might require spinal fusion along with decompression.

Spondylolisthesis. Spondylolisthesis is defined as displacement or slippage of one vertebra on another (7). Osteoarthritis of the facet joints (degenerative arthritis that breaks the cartilage between the fact joints) can lead to instability of the vertebral segments. The L4-L5 motion segment has most flexion-extension movement and is more prone to such slippage, as a result of weakened facet joints.

The most common symptoms are pain irradiating in lower extremities and cauda equina compression along with incontinence of bowel or bladder. Like most other spinal disorders, surgical treatment is recommended only when the nonsurgical treatments such as activity modification and physical therapy show no significant improvement in the patient's quality of life. The goals of the surgical treatment are pain reduction, prevention of further slip, and stabilization of spine (7). Surgical treatments include spinal fusion (with or without decompression), slip reduction or instrumentation, and interbody fusion.

Recommended operative treatments of degenerative spondylolisthesis are decompressive laminectomy (removal of lamina and medial joints), decompression with postero-lateral fusion (complete laminectomy and partial facetectomy along with fusion of the transverse process), and decompression with instrumental fusion.

TREATMENT OPTIONS AND MEDICAL DEVICES

Most spinal medical devices involve permanently fusing vertebral bone to correct a deformity or to limit a motion segment to stabilize the joint segment and relieve pain. Spine medical devices have their origin in plates, screws, and rods made from stainless steel and titanium, and today, these components comprise the majority of the implantable devices today.

Implants for Developmental Spine Deformities

Implants for developmental bone deformities such as scoliosis and kyphosis, generally use metal rods, screws, plates, and rib cages. For example, CD Horizon from Medtronic (Fig. 10) can be used for treatment of such deformities. Recently, the Food and Drug Administration (FDA) approved the Vertical Expandable Prosthetic Titanium Rib (VEPTR) from Synthes Spine (41,43), which is a surgically implanted device used to treat thoracic insufficiency syndrome (TIS) in pediatric patients. TIS is a congenital condition where severe deformities of the chest, spine, and ribs prevent normal breathing and lung growth and development.

Figure 6. VEPTR for the treatment of pediatric scoliosis from Synthes Spine (41).

The VEPTR device (Fig. 6) is a curved metal rod that is attached to ribs near the spine using hooks located at both ends of the device. The VEPTR device helps straighten the spine and separate ribs so that the lungs can grow and fill with enough air to breathe. The length of the device can be adjusted as the patient grows. It is hoped that the device will accomplish more normal growth pattern without spinal growth limitations, decreased chest, spine, and rib deformity and increased lung volume (43).

Implants for Degenerative Bone Disease

In case of kyphosis caused by osteoporosis (decrease in bone mass density with increased bone brittleness), minimally invasive methods such as vertebroplasty or (balloon kyphoplasty) are being used (7). Vertebroplasty involves the percutaneous injection of (PMMA) into a fractured vertebral body. Balloon kyphoplasty is another surgical approach to treat the kyphosis or deformity of the spine. In this procedure, an inflatable balloon is inserted between the vertebrae space to increase the disk height. Increase in disk height helps to reduce the deformity. The extra space created by balloon is filled with bone cement (PMMA), which when cured, binds the fracture. The hardened cement thus provides the strength to fractured/weak vertebrae and stability to the spine along with reducing the pain.

Implants for Degenerative Disk Disease

The treatment options for the patient would vary based on the age, pain history, and severity. When conservative treatments (such as rest, medications, physical therapy, etc.) fail, the patient is advised to undergo surgery. The goal of the surgery is to alleviate the pain. Most popular surgeries include lumbar microdiscectomy, lumbar laminectomy, microendoscopic surgery, and arthroscopic lumbar discectomy (30,31).

Lumbar microdiscectomy (or lumbar decompression) is a proven technique to reduce the back pain associated with herniated disks. In this treatment, a small portion of the

bone over the nerve root is removed to get relief from pain. A microscope is used to aid in visualizing the pinched nerve and subsequent microsurgical procedure to remove the excess portion of the herniated disk. Similarly, an open decompression (lumbar laminectomy) is another type of surgery that is performed to reduce the pain caused by neural impingement, which is particularly effective as a treatment for spinal stenosis. It is typically done with a posterior approach. The spine is approached by cutting the left and right back muscles off the lamina and removing the lamina itself. The facet joints are then trimmed to allow more space to nerve roots (30).

Even with these surgical treatments, pain may not be relieved for disks that are more severely degenerated. In these cases, fusion is required to restrict the motion of the segment and thus attempt to relieve pain. A discussion of fusion technologies and the associated implants follows. More recent advances in non-fusion technologies are aimed at preserving the motion segment while relieving pain. Such non-fusion technologies include total disk replacement, nucleus replacement, and annulus repair. Numerous new companies and new medical implant strategies are being explored currently and will hopefully prove to be clinically relevant pain relief and function restoring solutions to DDD.

Fusion Solutions. Spinal fusion is recommended when the discectomy approach may not be clinically relevant, and the goal is to relieve pain by stopping the motion of a spine segment. Spinal fusion instrumentation is essentially of three main types: pedicle screws, anterior interbody cages, and posterior lumbar cages. The bone generally fuses more effectively when its motion is minimized; hence, these devices are used to limit the motion of the fused segment. Similarly, the spinal fusion is based on the assumption that if the joint does not move, it will not create pain.

Pedicle screws (Fig. 7) are the means of providing anchor to the spine. They are used in combination with the short rod to grip the spine and are made from biocompatible metals such as medical-grade stainless steel or titanium. After a sufficient time, these screws can be removed by doing a surgery; however, most surgeons recommend keeping the screws unless it causes discomfort to patient.

Figure 7. Pedicle screws and instrumentation from Medtronic (32).

Figure 8. Jaguar Interbody Cage from Depuy Spine (33).

Anterior interbody cages (Fig. 8) have been recently approved by the FDA to use in the disk space. The cages are made from titanium and are porous, which allows the bone graft to grow. Cages are also made of novel composite materials (e.g., Jaguar from Depuy) such as carbon fiber-reinforced polymeric materials. The bone graft grows through the cage from one vertebra to another.

These cages are placed in front (anterior) of the lumbar spine and, hence, the name. The cages can be inserted using either mini-laparotomy or endoscopy, however, the former is preferred. In general, a 3 to 5 in. (7.6 to 12.7 cm) incision on the left side of the abdomen is made to approach the damaged disk.

Anterior lumbar interbody fusion (ALIF) surgery is often combined with posterior lumbar interbody fusion (PLIF) surgery to provide more rigid fixation. When the cages are placed from back of the spine, it is called posterior fusion. Coda (Fig. 9) is a titanium alloy device for PLIF with pedicle screws from Abbott Spine and features intraoperative adjustment for lordosis.

There is another form of the fusion surgery: transforaminal lumbar interbody fusion (TLIF), which is considered as an extended form of PLIF. In TLIF, an entire facet joint is removed to get a better access to disk space as compared with PLIF. This facilitates better visualization and more removal of the disk material and placement of larger bone graft/implant. The success rate of the cages almost entirely depends on the vertebrae condition. The surgery is not recommended in the case of osteoporosis because the vertebral body would not sustain the cage, leading to eventual failure of the end plates. In that sense, the pedicle screws are better than anterior cages as a

Figure 9. Coda PLIF device with pedicle screws from Abbott Spine (34).

Figure 10. CD Horizon Legacy from Medtronic (32).

fixation device. The anterior/posterior fusion is performed in the case of severe spinal instability or in revision surgery. The advantage of the anterior/posterior fusion is that it provides more surface area for the bone fusion to occur.

The gold standard in the case of fusion is considered to be postero-lateral gutter fusion surgery (30). A bone graft is placed in the postero-lateral portion of the spine. The transverse process of the vertebral body serves as an attachment site to the bone graft, which eventually grows to complete the fusion at the site.

The recent trend is to offer the spinal systems that can be used for multiple spinal treatments. For example, the CD Horizon Legacy Spinal System (Fig. 10) can be used as a posterior, noncervical, nonpedicle screw fixation system for treatment of DDD, spinal stenosis, spondylolisthesis, spinal deformities like scoliosis, and kyphosis. When used as a pedicle screw fixation system of the noncervical posterior spine, it may be indicated for degenerative spondylosthesis, kyphosis, and scoliosis.

Novel concepts for spine care (e.g., Dynesys from Zimmer Spine and Wallis from Abbott Spine) are recently introduced in the market. Dynesys (Fig. 11) is a posterior

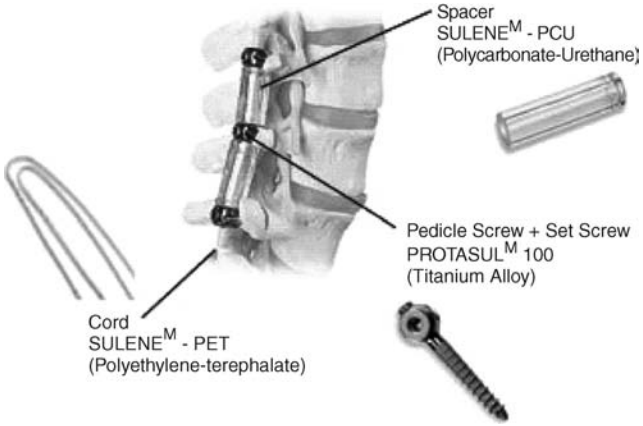


Figure 11. Dynesys from Zimmer Spine (35).

Figure 12. Wallis from Abbott Spine (34).

dynamics stabilization system, which is designed to bring the lumbar vertebrae back into more anatomical position while stabilizing the affected segments. The system used flexible materials threaded through pedicle screws, achieving dynamics stabilization.

The Wallis device (Fig. 12) from Abbott Spine is also another non-fusion spinal stabilization device that is under clinical trials in United States. The system is designed to treat the pain caused by initial stage DDD and aims to stabilize the lumbar spine without fusion, with a minimally invasive procedure.

The bone grafts used for the fusion can be either taken from patient's iliac crest (autograft) or from a human cadaver (allograft), such as Puros (Fig. 13) from Zimmer Spine (34). Autografts have the obvious advantage of compatibility with the patient's body. It helps in osteoconduction (bone growth) by means of providing calcium scaffold along with osteoblasts (bone growing cells) and morphogenic proteins (bone growing proteins). The allografts, in comparison, do not have osteoblasts and morphogenic proteins and merely provide the calcium scaffold for the fusion to occur. However, autografts lead to higher and longer postoperative pain as the bone graft is taken from the patient's own body.

Recently, synthetic bone grafts are introduced (e.g., Infuse), which represents an rhBMP-2 (recombinant human bone morphogenetic protein- 2) formulation combined with a bovine-derived absorbable collagen sponge (ACS) carrier. The INFUSE Bone Graft/LT-CAGE Lumbar Tapered Fusion Device, from Stryker Spine (36) (Fig. 14) is indicated for spinal fusion procedures in skeletally mature patients with DDD at one level from L4-S1, who may also have up to grade I spondylolisthesis at the involved level.

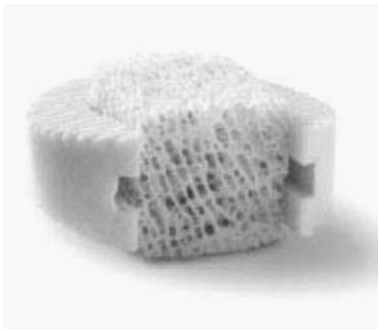


Figure 13. Puros allograft from Zimmer Spine (35).

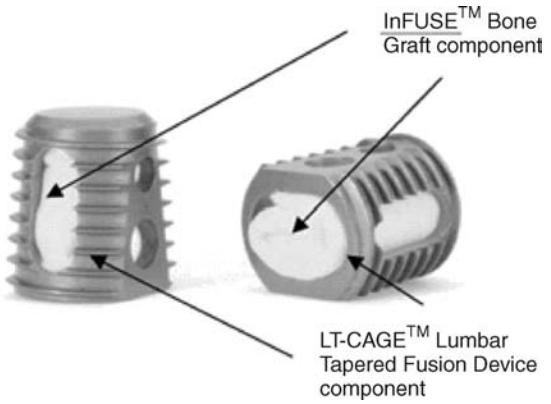


Figure 14. Infuse synthetic bone graft from Stryker Spine (36).

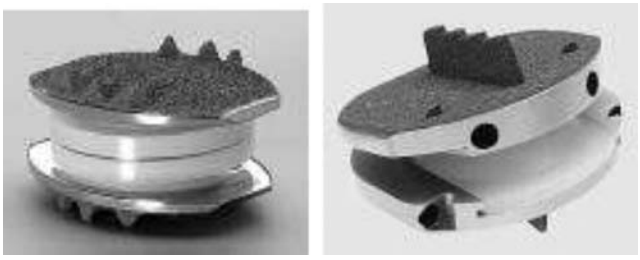


Figure 15. Charite from Depuy Spine (33) and ProDisc from Synthes Spine (41).

Bone stimulators offer another approach that potentially aid in spinal fusion. These externally applied devices emit low electrical current (30). This is aimed at facilitating stimulation of bone growth and increasing the chance of achieving spinal fusion. These are used in the case of patients who have a potentially very slow rate of obtaining solid fusion or in the case of the revision surgery.

The fusion and discectomy relieve pain but do not restore the normal spinal motion (37,38). The motivation behind exploration of the new and better solutions for the treatment of lower back pain is the failure of current treatments (conservative and surgical) in terms of restoring the motion and normal disk biomechanics. This is further aggravated by the complications that may occur after the surgical treatments, such as discectomy and/or spinal fusion.

Non-Fusion Solutions. Total disk replacement, where an entire diseased disk is removed and replaced by a synthetic implant, and nucleus pulposus arthroplasty, where only the nucleus of the disk is replaced either by a synthetic implant or recreated using tissue engineering approach, are the emerging approaches as alternatives to current surgical procedures for the treatment of the lower back pain (39). Annulus repair techniques, where defects in the annulus are either modified or repaired, also finds a place in emerging techniques and can be potentially used either alone or in combination with nucleus pulposus arthroplasty procedures, depending on the degenerative state of the intervertebral disk.

Total Disk Replacement. Total disk replacement targets later stages of disk degeneration (Galante grade IV), where the annulus is severely degenerated and is beyond repair (40). The diseased disk is entirely removed and replaced by a medical device that provides motion to the joint segment. Disk replacement may serve to eliminate the back pain and restore the physiological motion. A similar approach for total knee and hip replacement is highly successful. Total disk prostheses may be better options to spinal fusion and/or diskectomy as it allows the physiological motion between the adjacent vertebrae. Another advantage would be that the effectiveness of the surgery will not be dependent on the integrity of the annulus or degeneration state. To simulate the natural structure and function of the spinal unit, total disk prostheses also provide adequate fixation to the vertebrae.

There are a variety of total disk replacement design strategies, but two of the concepts that are furthest along are the Charite and the ProDisc, which are each based on metallic end plates that are porous coated and allow fixation to the superior and inferior end plates as well as an ultra-high molecular end plate polyethylene core, which provides a low friction articulation of the adjacent vertebrae. The use of artificial disks (Fig. 15) as a replacement to the damaged disk is currently in various phases of development and clinical trials. The Charite received FDA approval in 2004, and ProDisc, Maverick, and Flexicore are under clinical evaluation at the time of this writing (30).

Nucleus Replacement. The nucleus pulposus is a major component of the intervertebral disk and is actively involved in the disk function and load transfer mechanism. It is also involved with the pathologic changes of the disk. Researchers began to consider replacement of the nucleus alone because this tissue seems to degenerate before the annulus fibrosus. If this tissue alone can be replaced, preserving the annulus fibrosus, this may prolong the life of the disk and postpone or prevent the need for a more aggressive procedure such as fusion or total disk replacement. Nucleus replacement, as in case of total disk replacement, aims for restoration of the normal disk mechanics and functions, in contrast with the current surgical procedures of the diskectomy and the spinal fusion.

There are several nucleus implants in the various phases of development and clinical trials. Some are already implanted in humans in Europe (e.g., RayMedica PDN, Disc Dynamic's DASCOR), whereas most other nucleus implants are undergoing bench-top testing and/or investigational device exemption (IDE) submissions. The Raymedica prosthetic nucleus device (Fig. 16) has the longest history of all nucleus implants on the market. The clinical results of the PDN have been promising for pain relief and disk height restoration (41), but they are troubled by expulsion of the device from the annulus. Alternative implant designs and surgical procedures have limited this complication, but it remains a major challenge for these types of devices.

To perform surgical intervention on the intervertebral disk (e.g., in the case of nucleus replacement), outer

Figure 16. Single prosthetic nucleus device by Raymedica (42).

annulus fibrosus needs to be compromised. If the artificial incision in the annulus is not repaired, there is very high risk of nucleus implant expulsion, even under mild physiological loading. The main idea is to seal the annulus incision and/or prevent the expulsion of the nucleus implant from the created window. A barrier can be placed in between the nucleus and the annulus to prevent expulsion. These technologies are currently being explored in early clinical trials and in preclinical evaluations.

APPENDIX 1

CONCLUSIONS

This is perhaps the most exciting time in the development of medical devices for spinal applications. Never before have so many academic researchers, clinicians, and corporations so aggressively pursued solutions to spinal conditions that have a potential to be solved with medical devices. Along with this tremendous interest is the interest in better understanding of the anatomical structure, biochemistry, and function of the spinal structures. As more information becomes available, further refinements in treatments through medical devices will be improved offering more tools to the surgeon and a better chance of relieving pain while preserving the function of the spine.

APPENDIX 2. TERMINOLOGY

Allograft The transplant of an organ or tissue from one person to another.

Autograft The transplant of an organ or tissue from one body site to another body site of the same person.

Spinal Disorder	Treatment Options	Device/Implant
Stenosis	-Laminectomy -Spinal fusion	-Fusion Instrumentation • Cages • Pedicle screws • Metal rod • Autograft/Allograft
Spondylolisthesis	-Laminectomy -Spinal fusion	Fusion instrumentation • Cages • Pedicle screws • Metal rod • Autograft/Allograft
Scoliosis	-Bracing -Spinal fusion	-Braces -Fusion instrumentation • Cages • Pedicle screws • Metal rod • Autograft/Allograft
Kyphosis	-Balloon kyphoplasty	-Bone cement
Vertebral compression Fracture	-Vertebroplasty -Kyphoplasty	-Bone cement-Emerging materials
Disk degeneration and herniation	-Diskectomy -Spinal fusion -Total disk prosthesis -Nucleus replacement	-Fusion instrumentation • Cages • Pedicle screws • Metal rod • Autograft/Allograft -Facet replacement -Nucleus implant device -Total disk arthroplasty

Compression Fracture Collapse of the bone of the vertebral body, mostly due to osteoporosis and trauma.

Disk DegenerationThe loss of normal disk architecture accompanied by progressive fibrosis. This is seen as loss of hydration of the disk material and loss of disk height. This complex process alters the normal biomechanics of the spine and may cause back pain.

Disk Herniation Migration of the central nucleus pulposus of the disk toward disk periphery through cracks or fissures in outer annulus.

Discectomy A procedure in which in an excess portion of the disk impinging on the nerve root is cut off.

Kyphosis An exaggerated curvature of the back bones (vertebrae) in the upper back area or a rounded, "hunched" back.

Kyphoplasty A procedure that combines the ver tebroplasty technique with balloon catheter techno logy to treat the osteoporotic vertebral compression fractures.

Laminectomy A procedure in which the lamina (roof) of the vertebra is trimmed to create more space for the spinal nerves.

Nucleus Implant An artificial material, which can be used as a replacement to the degenerated nucleus of the intervertebral disk to relieve back pain and preserve the normal motion.

Osteoporosis A disease in which bones become fragile and brittle, making it prone to break easily.

Scoliosis A curvature of the spine.

Spinal Fusion A procedure in which an intervertebral disk between the adjacent vertebrae is replaced by bone graft. The procedure is performed to relieve the back pain and stabilize the spinal segment by fusing the vertebrae together, with or without spinal instrumentation.

Spondylolisthesis Slippage of one vertebra on another.

Stenosis Narrowing of the spinal canal.

Total Disk Prosthesis An artificial device, which can be used as a feasible replacement of the degenerated disk to relieve back pain and preserve the motion.

Vertebroplasty A procedure that stabilizes the collapsed vertebra with the injection of the medical-grade bone cement into the spine.

BIBLIOGRAPHY

1. White AA, Panjabi MM. Clinical Biomechanics of the Spine. II ed. Philadelphia: J.B. Lippincott Company; 1990.
2. http://www.ab.ust.hk.
3. http://www.spinalstenosis.org/.
4. Iatridis JC, et al. Is the nucleus pulposus a solid or a fluid? Mechanical behaviors of the nucleus pulposus of the human intervertebral disk. Spine 1996;21:1174–1184.
5. Ayad S, Weiss JB. Biochemistry of the intervertebral disk. In: MIV J, editor. The Lumbar Spine and Back Pain. 3rd ed. New York: Churchill-Livingstone; 1987. 100–137.
6. Buckwalter JA. Aging and degeneration of the human intervertebral disk. Spine 1995;20:1307–1314.
7. Vaccaro A. Core Knowledge in Orthopaedics. Spine. St. Louis, MO: CV Mosby; 2004.
8. http://www.scoliosisrx.com.
9. http://www.bostonbrace.com/ superstructure.htm.
10. MedPro Month. 1998; V.VIII:Number 1.
11. Andersson GBJ. Epidemiologic Aspects of low-back pain in industry. Spine 1981;6(1):53–60.
12. Hedman TP, et al. Design of an intervertebral disk prosthesis. Spine 1991;16(6):S256–S260.
13. Cats-Baril WL, Frymoyer JW. Identifying patients at risk of becoming disabled because of low-back pain—the Vermont Rehabilitation Engineering Center predictive model. Spine 1991;16(6):605–607.
14. Sehgal N, Fortin JD. Internal disk disruption and low back pain. Pain Physician 2000;3(2):143–157.
15. Heliovaara M, et al. Determinants of sciatica and low-back pain. Spine 1991;16(6):608–614.
16. Manchikanti L. Epidemiology of low back pain. Pain Physician 2000;3(2):167–192.
17. Bao QB, Yuan HA. Artificial disk technology. Neurosurgical Focus 2000;9(4):1–9.
18. Bibby S, et al. The pathophysiology of the intervertebral disk. Joint Bone Spine 2001;68:537–542.
19. Vernon-Roberts B. Age-related and degenerative pathology of intervertebral disks and apophyseal joints. In: Jayson MIV, editor. The Lumbar Spine and Back Pain. New York: Churchill Livingstone; 1992. 17–41.
20. Vernon-Roberts B. Disk pathology and disease states. In: Ghosh P, editor. The Biology of the Intervertebral Disk. Boca Raton: CRC Press; 1988. 73–120.
21. Coventry MB, Ghormley RK, Kernohan JW. The intervertebral disk: Its microscopic anatomy and pathology. Part III. Pathologic changes in the intervertebral disk. J Bone Joint Surg 1945;27A:460–474.
22. Friberg S, Hirsch C. Anatomical and clinical studies on lumbar disk degeneration. Acta Orthop Scand 1949;19: 222–242.
23. Harris RI, Macnab I. Structural changes in the lumbar intervertebral disks. Their relationship to low back pain and sciatica. J Bone Joint Surg 1954;36B:304–322.
24. Hoof VD. A: Histological age changes in the annulus fibrosus of the human intervertebral disk. Gerontology 1964;9:136–149.
25. Bao QB, et al. The artificial disk: Theory, design and materials. Biomaterials 1996;17:1157–1166.
26. Akeson WH, et al. Biomechanics and biochemistry of the intervertebral disk. Clin Orthop Rel Res 1977;129:133–139.
27. McNally DS, Adams MA. Internal Intervertebral disk mechanics as revealed by stress profilometry. Spine 1992; 17:66–73.
28. Osti OL, et al. Annular tears and disk degeneration in the lumbar spine. J Bone Joint Surg (Br) 1992;74B:678–682.

29. Snyder DL, Doggett D, Turkelson C. Treatment of degenerative lumbar spinal stenosis. American Family Physician 2004;70(3):517–520.
30. http://www.spine-health.com.
31. http://www.texasspinecenter.com.
32. http://www.medtronic.com.
33. http://www.depuyspine.com.
34. http://www.abbottspine.com.
35. http://www.zimmerspine.com.
36. http://www.stryker.com/spine.
37. Kambin P, Savitz MH. Arthroscopic microdiscectomy: An alternative to open disk surgery. Mount Sinai J Med 2000; 67(4):283–287.
38. Weber H. Lumbar disk herniation: A controlled prospective study with ten years of observation. Spine 1993;8:131–140.
39. Joshi A. Mechanical behavior of the human lumbar intervertebral disk with polymeric nucleus implant: An experimental and finite element study. Ph.D. Thesis. Drexel University, 2004.
40. Galante JO. Tensile properties of the human annulus fibrosus. Acta Orthop Scand 1967; (Suppl. 100):4–91.
41. http://www.synthes.com.
42. http://www.raymedica.com.
43. http://www.fda.gov.

See also HUMAN SPINE, BIOMECHANICS OF; SCOLIOSIS, BIOMECHANICS OF.

SPINE. See HUMAN SPINE, BIOMECHANICS OF.

SPIROMETRY. See PNEUMOTACHOMETERS.

STATISTICAL METHODS

ARNAUD DELORME
University of San Diego,
La Jolla, California

INTRODUCTION

Statistics can be called that body of analytical and computational methods by which characteristics of a population are inferred through observations made in a representative sample from that population. Since scientists rarely observe entire populations, sampling and statistical inference are essential. Although, the objective of statistical methods is to make the process of scientific research as efficient and as productive as possible, many scientists and engineers have inadequate training in experimental design and in the proper selection of statistical analyses for experimentally acquired data. Gill (1) states: "...statistical analysis too often has meant the manipulation of ambiguous data by means of dubious methods to solve a problem that has not been defined." The purpose of this article is to provide readers with definitions and examples of widely used concepts in statistics. This article first discusses some general principles for the planning of experiments and data visualization. Then, since it is expected that most readers are not studying this article to learn statistics, but to find practical methods for analyzing data, a strong emphasis has been put on choice of an appropriate standard statistical model and statistical inference methods (parametric, nonparametric, resampling methods) for different types of data. Then, methods for processing multivariate data are briefly reviewed. The section following it deals with clinical trials. Finally, the last section discusses computer software and guides the reader through a collection of bibliographic references adapted to different levels of expertise and topics.

DATA SAMPLE AND EXPERIMENTAL DESIGN

Any experimental or observational investigation is motivated by a general problem that can be tackled by answering specific questions. Associated with the general problem will be a population. For example, the population can be all human beings. The problem may be to estimate the probability by age bracket for someone to develop lung cancer. Another population may be the full range of responses of a medical device to measure heart pressure and the problem may be to model the noise behavior of this apparatus.

Often, experiments aim at comparing two subpopulations and determining if there is a (significant) difference between them. For example, the frequency occurrence of lung cancer of smokers compared may be compared to nonsmokers or the signal/noise ratio generated by two brands of medical devices may be compared and determined which brand outperforms the other with respect to this measure.

How can representative samples be chosen from such populations? Guided by the list of specific questions, samples will be drawn from specified subpopulations. For example, the study plan might specify that 1000 presently cancer-free persons will be drawn from the greater Los Angeles area. These 1000 persons would be composed of random samples of specified sizes of smokers and nonsmokers of varying ages and occupations. Thus, the description of the sampling plan will imply to some extent the nature of the target subpopulation, in this case smoking individuals.

Choosing a random sample may not be easy and there are two types of errors associated with choosing representative samples: sampling errors and nonsampling errors. Sampling errors are those errors due to chance variations resulting from sampling a population. For example, in a population of 100,000 individuals, suppose that 100 have a certain genetic trait and in a (random) sample of 10,000, 8 have the trait. The experimenter will estimate that 8/10,000 of the population or 80/100,000 individuals have the trait, and in doing so will have underestimated the actual percentage. Imagine conducting this experiment (i.e., drawing a random sample of 10,000 and examining for the trait) repeatedly. The observed number of sampled individuals having the trait will fluctuate. This phenomenon is called the sampling error. Indeed, if sampling is truly random, the observed number having the trait in each repetition will fluctuate randomly $\sim$10. Furthermore, the limits within which most fluctuations will occur are estimable using standard statistical methods.

Consequently, the experimenter not only acknowledges the presence of sampling errors, but he can estimate their effect.

In contrast, variation associated with improper sampling is called nonsampling error. For example, the entire target population may not be accessible to the experimenter for the purpose of choosing a sample. The results of the analysis will be biased if the accessible and nonaccessible portions of the population are different with respect to the characteristic(s) being investigated. Increasing sample size within the accessible portion will not solve the problem. The sample, although random within the accessible portion, will not be representative of the target population. The experimenter is often not aware of the presence of nonsampling errors (e.g., in the above context, the experimenter may not be aware that the trait occurs with higher frequency in a particular ethnic group that is less accessible to sampling than other groups within the population). Furthermore, even when a source of nonsampling error is identified, there may not be a practical way of assessing its effect. The only recourse when a source of nonsampling error is identified is to document its nature as thoroughly as possible. Clinical trials involving survival studies are often associated with specific nonsampling errors (see the section dealing with clinical trials below).

DESCRIPTIVE STATISTICS

Descriptive statistics are tabular, graphical, and numerical methods by which essential features of a sample can be described. Although these same methods can be used to describe entire populations, they are more often applied to samples in order to capture population characteristics by inference.

The two main types of data samples will be differentiated: qualitative data samples and quantitative data samples. Qualitative data arises when the characteristic being observed is not measurable. A typical case is the "success" or "failure" of a particular test. For example, to test the effect of a drug in a clinical trial setting, the experimenter may define two possible outcomes for each patient: either the drug was effective in treating the patient, or the drug was not effective. In the case of two possible outcomes, any sample of size n can be represented as a sequence of n nominal outcome $x_1, x_2, \ldots, x_n$ that can assume either the value success or failure.

By contrast, quantitative data arise when the characteristics being observed can be described by numbers. Discrete quantitative data is countable, whereas continuous data may assume any value, apart from any precision constraint imposed by the measuring instrument. Discrete quantitative data may be obtained by counting the number of each possible outcome from a qualitative data sample. Examples of discrete data may be the number of subjects sensitive to the effect of a drug (number of success and number of failure). Examples of continuous data are weight, height, pressure, and survival time. Thus, any quantitative data sample of size n may be represented as a sequence of n numbers $x_1, x_2, \ldots, x_n$ and sample statistics are functions of these numbers.

Table 1. Result of a Hearing Aid Device Satisfaction Survey in 1000 Patients Showing the Frequency Distribution of Each Response

Satisfaction rank	Number of Responses
0	38
1	144
2	342
3	287
4	164
5	25
Total	1000

Discrete data may be preprocessed using frequency tables and represented using histograms. This is best illustrated by an example. For discrete data, consider a survey in which 1000 patients fill in a questionnaire for assessing the quality of a hearing aid device. Each patient has to rank product satisfaction from 0 to 5, each rank being associated with a detailed description of hearing quality. Table 1 represents the frequency of each response type. A graphical equivalent is the frequency histogram illustrated in Fig. 1. In the histogram, the heights of the bars are the frequencies of each response type. The histogram is a powerful visual aid to obtain a general picture of the data distribution. In Fig. 1, notice a majority of answers corresponding to response type 2 and a 10-fold frequency drop for response types 0 and 5 compared to response type 2.

For continuous data, consider the data sample in Table 2, which represents amounts of infant serum calcium in mg·100 mL^{-1} for a random sample of 75 week old infants whose mothers received vitamin D supplements during pregnancy. Little information is conveyed by the list of numbers. To depict the central tendency and variability of the data, Table 3 groups the data into six classes, each of width 0.03 mg·100 mL^{-1}. The "frequency" column in Table 3 gives the number of sample values occurring in each class. The picture given by the frequency distribution in Table 3 is a clearer representation of central

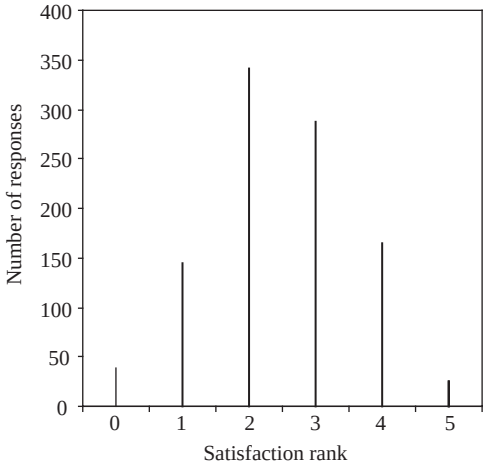


Figure 1. Frequency histogram for the hearing aid device satisfaction survey of Table 1.

Table 2. Serum Calcium (mg·100 mL^{-1}) in a Random Sample of 75 Week Old Infants Whose Mother Received Vitamin D Supplement During Pregnancy

9.37	9.34	9.38	9.32	9.33	9.28	9.34
9.29	9.36	9.30	9.31	9.33	9.34	9.35
9.35	9.36	9.30	9.32	9.33	9.35	9.36
9.32	9.37	9.34	9.38	9.36	9.37	9.36
9.36	9.33	9.34	9.37	9.44	9.32	9.36
9.38	9.39	9.34	9.32	9.30	9.30	9.36
9.29	9.41	9.27	9.36	9.41	9.37	9.31
9.31	9.33	9.35	9.34	9.35	9.34	9.38
9.40	9.35	9.37	9.35	9.32	9.36	9.35
9.35	9.36	9.39	9.31	9.31	9.30	
9.31	9.36	9.34	9.31	9.32	9.34	

tendency and variability of the data than that presented by Table 2. In Table 3, data are grouped in six classes of equal size and it is possible to see the centering of the data about the 9.325–9.355 class and its variability: The measurements vary from 9.27 to 9.44 with ~95% of them between 9.29 and 9.41. The advantage of grouped frequency distributions is that grouping smoothes the data so that essential features are more discernible. Figure 2 represents the corresponding histogram. The sides of the bars of the histogram are drawn at the class boundaries and their heights are the frequencies or the relative frequencies (frequency/sample size). In the histogram, the distribution of the data centered about the point 9.34 is clearly seen. Although grouping smoothes the data, too much grouping (that is choosing too few classes) will tend to mask rather than enhance the sample's essential features.

There are many numerical indicators for summarizing and describing data. The most common ones indicate central tendency, variability, and proportional representation (the sample mean, variance, and percentiles, respectively). We assume that any characteristic of interest in a population, and hence in a sample, can be represented by a number. This is obvious for measurements and counts, but even qualitative characteristics (described by discrete variables) can be numerically represented. For example, if a population is dichotomized into those individuals who are carriers of a particular disease and those who are not, a 1 can be assigned to each carrier and a 0 to each noncarrier. The sample can then be represented by a sequence of zeroes and ones.

The most common measure of central tendency is the sample mean:

$$M = (x_1 + x_2 + \ldots + x_n)/n \quad \text{also noted } \overline{X} \qquad (1)$$

Table 3. Frequency distribution of infant serum calcium data

Serum Calcium mg·100 mL^{-1}	Frequency
9.265–9.295	4
9.295–9.325	18
9.325–9.355	24
9.355–9.385	22
9.385–9.415	6
9.415–9.445	1
Total	75

where $x_1, x_2, \ldots, x_n$ is the collection of numbers from a sample of size n. The sample mean can be roughly visualized as the abscissa of the horizontal center of gravity of the frequency histogram. For the serum calcium data of Table 2, $M = 9.34$, which happens to be the midpoint of the highest bar of the histogram (Fig. 2). This histogram is roughly symmetric about a vertical line drawn through M, but this is not necessarily true of all histograms. Histograms of counts and survival times data are often skewed to the right (long-tailed with concentrated mass at the lower values). Consequently, the idea of M as a center of gravity is important to bear in mind when using it to indicate central tendency. For example, the median (described later in this section) may be a more appropriate index of centrality depending on the type of data and the kind of information one wishes to convey.

The sample variance, defined by

$$s^2 = \frac{1}{n-1}[(x_1 - M)^2 + (x_2 - M)^2 + \cdots + (x_n - M)^2] = \sum_{i=1}^{n} \frac{(x_i - M)^2}{n-1} \qquad (2)$$

is a measure of variability or dispersion of the data. As such, it can be motivated as follows: x_i-M is the deviation of the ith data sample from the sample mean, that is, from the "center" of the data; we are interested in the amount of

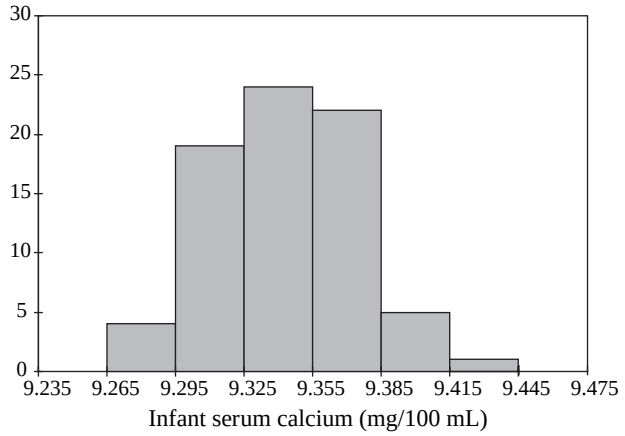


Figure 2. Frequency histogram of infant serum calcium data of Tables 2 and 3. The curve on the top of the histogram is another representation of probability density for continuous data.

deviation, not its direction, so the sign is disregarded by calculating the squared deviation $(x_i\text{-}M)^2$; finally, the squared deviations are averaged by summing them and dividing by the sample size -1. (Division by $n-1$ ensures that the sample variance is an unbiased estimate of the population variance.) Note that an equivalent and often more practical formula for computing the variance may be obtained by developing Eq. 2:

$$s^2 = \frac{\sum x_i^2 - nM^2}{n-1} \tag{3}$$

A measure of variability in the original units is then obtained by taking the square root of the sample variance. Specifically, the sample standard deviation, denoted s, is the square root of the sample variance.

For the serum calcium data of Table 2, $s^2 = 0.0010$ and $s = 0.03$ mg·100 mL^{-1}. The reader might wonder how the number 0.03 gives an indication of variability. Note that for the serum calcium data $M \pm s = 9.34 \pm 0.03$ contains 73% of the data, $M \pm 2s = 9.34 \pm 0.06$ contains 95% and $M \pm 3s = 9.34 \pm 0.09$ contains 99%. It can be shown that the interval $M \pm 3s$ will include at least 89% of any set of data (irrespective of the data distribution).

An alternative measure of central tendency is the median value of a data sample. The median is essentially the sample value at the middle of the list of sorted sample values. We say essentially because a particular sample may have no such value. In an odd-numbered sample, the median is the middle value; in an even-numbered sample, where there is no middle value, it is conventional to take the average of the two middle values. For the serum calcium data of Table 3, the median is equal to 9.34.

By extension to the median, the sample p percentile (say, e.g., 25th percentile) is the sample value at or below which p% (25%) of the sample values lie. If there is no value at a specific percentile, the average between the upper and lower closest existing round percentile is used. Knowledge of a few sample percentiles can provide important information about the population.

For skewed frequency distributions, the median may be more informative for assessing a population center than the mean. Similarly, an alternative to the standard deviation is the interquartile range: it is defined as the seventy-fifth minus the twenty-fifth percentiles and is a variability index not as influenced by outliers as the standard deviation.

There are many other descriptive and numerical methods (see, e.g., Ref. (2)). It should be emphasized that the purpose of these methods is usually not to study the data sample itself, but rather to infer a picture of the population from which the sample is taken. In the next section, standard population distributions and their associated statistics are described.

PROBABILITY, RANDOM VARIABLES, AND PROBABILITY DISTRIBUTIONS

The foundation of all statistical methodology is probability theory, which progresses from elementary to the most advanced mathematics. Much of the misunderstanding and abuse of statistics comes from the lack of understanding of its probabilistic foundation. When assumptions of the underlying probabilistic (mathematical) model are grossly violated, derived inferential methods will lead to misleading and irrational conclusions. Here, only enough probability theory to provide a framework for this article is discussed.

In the rest of this article, experiments that have more than one possible outcome, the actual outcome being determined by some chance mechanism will be studied. The set of possible outcomes of an experiment is called its sample space; subsets of the sample space are called events, and an event is said to occur if the actual outcome of the experiment is a member of that event. A simple example follows.

The experiment will be the toss of a pair of fair coins, arbitrarily labeled coin number 1 and coin number 2. The outcome (1,0) means that coin No. 1 shows a head and coin No. 2 shows a tail. Then, the sample space by the collection of all possible outcomes can be specified:

$$S = \{(0,0)(0,1))(1,0)(1,1)\} \tag{4}$$

There are four ordered pairs so there are four possible outcomes in this coin-tossing experiment. Consider the event A "toss one head and one tail", which can be represented by $A = \{(1,0)\ (0,1)\}$. If the actual outcome is (0,1) then the event A has occurred.

In the example above, the probability for event A to occur is obviously 50%. However, in most experiments it is not possible to intuitively estimate probabilities, so the next step in setting up a probabilistic framework for an experiment is to assign, through some mathematical model, a probability to each event in the sample space.

Definition of Probability

A probability measure is a rule, say P, which associates with each event contained in a sample space S a number such that the following properties are satisfied:

1. For any event, A, $P(A) \geq 0$.
2. $P(S) = 1$ (since S contains all the outcomes, S always occurs).
3. $P(\text{not } A) + P(A) = 1$.
4. If A and B are mutually exclusive events (that cannot occur simultaneously) and independent events (that are not linked in any way), then

$$P(A \text{ or } B) = P(A) + P(B) \quad \text{and} \quad P(A \text{ and } B) = 0$$

Many elementary probability theorems (rules) follow directly from these definitions.

Probability and Relative Frequency

The axiomatic definition above and its derived theorems dictate the properties that probability must satisfy, but they do not indicate how to assign probabilities to events. The major classical and cultural interpretation of probabilities is the relative frequency interpretation. Consider an experiment that is (at least conceptually) infinitely

repeatable. Let A be any event and let n_A be the number of times the event A occurs in n repetitions of the experiment; then the relative frequency of occurrence of A in the n repetitions is n_A/n. For example, if mass production of a medical device reliably yields 7 malfunctioning devices out of 100, the relative frequency of occurrence of a defective device is 7/100.

The probability of A is defined by $P(A) = \lim n_A/n$ as $n \to \infty$, where this limit is assumed to exist. The number $P(A)$ can never be known, but if the experiment can in fact be repeated a large number of times, it can be estimated by the relative frequency of occurrence of A.

The relative frequency interpretation is an objective interpretation because the probability of an event is assumed to be independent of judgment by the observer. In the subjective interpretation of probability, a probability is assigned to an event according to the assigner's strength of belief that the event will occur, on a scale of 0–1. The assigner could be an expert in a specific field, for example, a cardiologist that provides the probability for a sample of electrocardiograms to be pathological.

Probability Distribution Definition and Probability Mass Function

It has been assumed that all data can be numerically represented. Thus, the outcome of an experiment in which one item will be randomly drawn from a population will be a number, but this number cannot be known in advance. Let the potential outcome of the experiment be denoted by X, which is called a random variable in statistics. When the item is drawn, X will be realized or observed. Although the numerical values that X will take cannot be known in advance, the random mechanism that governs the outcome can perhaps be described by a probability model. Using the model, the probability that the random variable X will take a value within a set or range of numbers can be calculated.

One such popular mathematical model is the probability distribution of a discrete random variable X. It can be best described as a mathematical equation or table that gives, for each value x that X can assume, the probability associated with this value $P(X = x)$. For example, if X represents the outcome of the tossing of a coin, there are two possible outcomes, tail and head. If it is a fair coin $P(X = tail) = 0.5$ and $P(X = head) = 0.5$. In statistics, the function $P(X = x)$ is called the probability mass function of X.

It follows from the relative frequency interpretation of probability that, for a discrete random variable or for the frequency distribution of a continuous variable, relative frequency histograms estimate the probability mass functions of this variable. For example, in Table 3, if the random variable X indicates the serum calcium measure, then

$$\hat{P}(X \text{ is in the first bin}) = \hat{P}(9.265 \le X < 9.295) = 4/75$$

the ^ symbol on P indicating estimated probability values, since actual probabilities describe the population itself and cannot be calculated from data samples. Similarly the probability that X is in the second bin, the third bin, ... can be estimated and the collection of these probabilities constitute an estimated probability mass function.

Probability Density Function for Continuous Variables

The probability mass function above best describes discrete events, but what probabilities can be assigned to continuous variables? Since a continuous variable X can assume any value on a continuum, the probability that X assumes a *particular* value is 0 (except in very particular cases that will not be discussed here). Consequently, associated with a continuous random variable X, is a function f_X, called its probability density function that can be used to compute probability. The probability that a continuous random variable X assumes a value between values x_1 and x_2 is the area under the graph of f_X over the interval x_1 and x_2; mathematically

$$P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} f_X(x)dx \tag{5}$$

For example, for the infant serum data of Table 2 (see also Table 3), it can be estimated that the probability that an infant whose mother received a vitamin D supplement during pregnancy has between 9.35 and 9.38 mg·100 mL^{-1} calcium is 22/75 or 0.293, which is the relative frequency of the 9.355–9.385 class in the sample. For continuous data, a smooth curve passing through the midpoint of a histogram bars' upper limit should resemble the probability density function of the underlying population.

There are many mathematical models of probability distribution. Three of the most commonly used probability distribution models described below are the binomial distribution and the Poisson distribution for discrete variables, and the normal distribution for continuous variables.

The Binomial Distribution

The scenario leading to the binomial distribution is an experiment that consists of n independent, repeated trials, each of which can end in only one of two ways arbitrarily labeled success or failure. The probability that any trial ends in a success is p (and hence $q = 1-p$ for a "failure"). Let the random variable X denote the total number of successes in the n trials, and x denote a number in $\{0; \ldots; n\}$. Under these assumptions:

$$P(X = x) = \binom{n}{x} p^x q^{n-x} \quad x = 0, 1, \ldots, n \tag{6}$$

with

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{7}$$

where $n! = 1^*2^*3...^*n$ is n factorial.

For example, suppose the proportion of carriers of an infectious disease in a large population is 10% ($p = 0.1$) and that the number of carriers follows a binomial distribution. If 20 individuals are sampled ($n = 20$) and X is the number of carriers (successes) in the sample, then the probability

that there will be exactly one carrier in the sample is

$$P(X=1)=\binom{20}{1}(0.10)^1(0.90)^{20-1}=0.27$$

More complex probabilities may be calculated with the help of probability rules and definitions. For instance the probability that there will be at least two carriers in the sample is

$$\begin{aligned}P(X\geq 2) &= 1-P(X<2)\\ &\quad \text{(see third probability definition)}\\ &= 1-P(X=0\,\text{or}\,X=1)\\ &= 1(P(X=0)+P(X=1))\\ &\quad \text{(see fourth probability definition)}\\ &= 1-\binom{20}{0}(0.10)^0(0.90)^{20}-\binom{20}{1}(0.10)^1(0.90)^{19}\\ &= 1-0.12-0.27=0.61\end{aligned}$$

Historically, single trials of a binomial distribution are called Bernoulli variates after the Swiss mathematician James Bernoulli who discovered it at the end of the seventeenth century.

The Poisson Distribution

The Poisson distribution is often used to represent the number of successive independent events of a specified type (e.g., cases of flu) with low probability of occurrence (<10%) in some specified interval of time or space. The Poisson distribution is also often used to represent the number of occurrence of events of a specified type where there is no natural upper limit, for example, the number of radioactive particles emitted by a sample over a set time period. Specifically, X is a Poisson random variable if it obeys the following formula:

$$P(X=x)=e^{-\lambda}\lambda^x/x! \quad x=0,1,2,\ldots \tag{8}$$

where $e=2.178\ldots$ is the natural logarithmic base and λ is a given constant. For example, suppose the number of a particular type of bacteria in a standard area (e.g., $1\,\text{cm}^2$) can be described by a Poisson distribution with parameter $\lambda=5$. Then, the probability that there are no more than 3 bacteria in the standard area is given by

$$\begin{aligned}P(X\leq 3) &= P(X=0)+P(X=1)+P(X=2)+P(X=3)\\ &= e^{-5}5^0/0!+e^{-5}5^1/1!+e^{-5}5^2/2!+e^{-5}5^3/3!\\ &= 0.265\end{aligned}$$

Note that the Poisson and the binomial distributions are closely related. In the case of a rare event ($p<10\%$), the binomial distribution (described by probability p and n events) is well approximated by the Poisson distribution with the constant $\lambda=np$. The Poisson distribution was named after the French mathematician Siméon-Denis Poisson, who discovered it in the early part of the nineteenth century.

The Normal Distribution

The binomial and Poisson distributions describe discrete events, but there are also many distributions describing continuous variables. The most important one is the normal distribution (also called Laplace–Gauss distribution as it was discovered by the French astronomer Pierre–Simon Laplace and the German mathematician Karl Friedrich Gauss in the early nineteenth century). Normal distributions arise as a result of many small random fluctuations about some general average (e.g., repeated recordings of a constant body temperature using a noisy electronic thermometer). A random variable X is said to be a normal or Gaussian random variable with mean parameter μ and standard deviation parameter σ if its probability density function is

$$f_X(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{(x-\mu)^2/2\sigma^2} \qquad -\infty<x<\infty \tag{9}$$

The normal probability density function graphed in Fig. 3, is bell shaped with tails rather rapidly receding to zero height. Because f_X represents probability density, the total area bounded by the curve is 1 (see Eq. 9). The area between two values of variable X (x_1 and x_2 where $x_1<x_2$) represents the probability that X lies between x_1 and x_2, (Eq. 5).

As shown in Fig. 3, if X is normal (μ, σ), it can be calculated that $P(\mu-3\sigma\leq X\leq\mu+3\sigma)=0.997$, which, according to the relative frequency interpretation of probability, states that $\sim 99.7\%$ of a large sample from a "normally distributed population" will be contained in the interval mean plus or minus three standard deviations ($\mu\pm 3\sigma$).

Note that there is a relation between the normal and the binomial distribution. Using the same notation as in Eq. 6, if n, the number of samples, is large enough then the variable z defined as

$$z=\frac{x-np}{\sqrt{npq}} \tag{10}$$

is approximately normally distributed with mean 0 and standard deviation 1. In a coin throwing experiment, throwing the coin a large number of times and counting

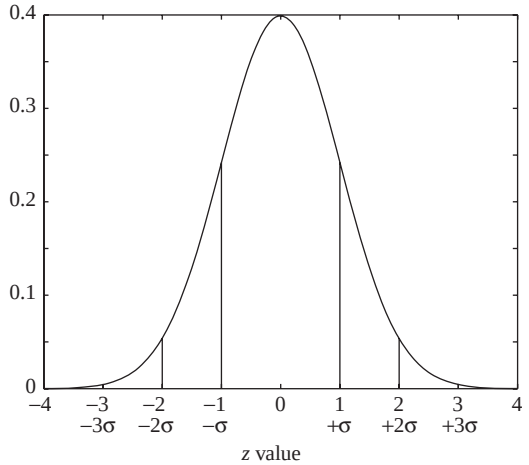


Figure 3. The normal probability density function showing symmetry about a vertical line through μ and the role of σ as a variability parameter. Vertical bars indicates $\pm\sigma$, $\pm 2\sigma$, $\pm 3\sigma$.

the number of heads x, then building a histogram for the value z, the histogram will be close to a normal distribution (as shown in Fig. 3). Similarly, there is a relation between the Poisson and the normal distribution, the variable z defined as $z=(x-\lambda)/\lambda$ is normally distributed for large values of λ.

Many statistical inferential methods described in the next section assume that the data is approximately normally distributed. Much abuse occurs, however, when these methods are applied blindly with no verification of the normality assumption. Incidentally, methods that incorporate assumptions of normality often can be applied to non-normal situations because under certain conditions, the normal distribution can approximate other distributions, such as the binomial and the Poisson distributions. Sometimes, the data can also be preprocessed to fit the normal distribution. For example, a histogram might indicate nonnormality, while a histogram of the logarithms of the data would fit the normal distribution, indicating that normal-based models can be applied to the log-transformed data. These transformations are discussed in most experimental design textbooks.

The importance of the normal distribution in statistics is also due to the central limit theorem in statistics that states that the distribution of any linear mixture of two or more independent random variables is more normal (has a shape closer to the normal distribution) than the distribution of the random variables themselves. This property is used by some algorithms processing multivariate data (as described in a later section).

There are many other continuous probability distributions besides the normal distribution. For example, the most commonly used distribution in survival analysis is the Weibull distribution. The von Mises distribution allows parametric statistical tests for periodic data (i.e., seasonal).

Characteristics of Probability Distributions

Just as there are numerical indexes for sample description, for example, sample means, variances, and percentiles, there are numerical characteristics of probability distributions. The expectation or mean (not sample mean) of an random variable X is

$$\begin{aligned} E(X) &\equiv \sum_x xP(X=x) \quad X \text{ discrete} \\ &\equiv \int_{-\infty}^{\infty} x f_X(x)dx \quad X \text{ continuous} \end{aligned} \tag{11}$$

The expectation E is a measure of central tendency for a population (i.e., the center of gravity of the probability distribution about the y axis). The variance of X is defined in terms of expectation by

$$\text{Var}(X) \equiv E\{[X-E(X)]^2\} \tag{12}$$

In words, Var(X) is the expected squared deviation of X from $E(X)$, and in this sense is a measure of variability or dispersion for a population. The standard deviation of X is the square root of its variance. Table 4 indicates mean and variance for the binomial, the Poisson, and the normal distribution.

Numerical descriptors of populations are often the very things we want to know about populations. They should not be confused with their sample counterparts; the sample numerical descriptors are the basis for drawing inferences regarding their population counterparts, which are of primary interest.

Table 4. Mean and Variance for Standard Distributions (see text for details)

	Binomial	Poisson	Normal
Mean	$\mu=Np$	$\mu=\lambda$	μ
Variance	$\sigma^2=Npq$	$\sigma^2=\lambda$	σ^2

Statistical Inference

A statistical hypothesis is a statement about the probability distribution of populations using one or more data samples. Typical questions are Is this single data sample consistent with this theoretical distribution of values?, Are these two data samples originating from the same population?, Are these n data samples originating from the same population?. Associated with each of these questions, in statistics, two hypotheses are usually formulated.

Hypothesis H_0: All data samples originate from the same population (or the single data sample is consistent with a given theoretical distribution).

Hypothesis H_1: Some data samples do not originate from the same population (or the single data sample is not consistent with the given theoretical distribution).

The test is called significant if hypothesis H_0 is rejected with respect to a user-defined confidence interval (e.g., 5% of chance of wrongly rejecting H_0). It is important to remember that inference tests can never disprove hypothesis H_0. Instead, based on the significance threshold and on the inference test chosen, it can be said that the data support rejecting H_0. The test is called nonsignificant if the hypothesis H_0 and reject hypothesis H_1 is accepted. Accepting H_0 means that we failed to find any significant difference with respect to our user-defined confidence interval. Because the error in accepting H_0 is usually large (see error types below), in general we should avoid drawing any conclusion about the experiment when accepting H_0.

Degree of Freedom. Elementary tests usually depend on the data sample size as well as the number of parameters (e.g., mean or variance) that have to be estimated from the sample, to run the test. Specifically, the number of degrees of freedom of a statistics is defined as the number of independent observations minus the number of population parameters, which must be estimated from sample observations. Details will be provided for each test.

***p*-Values.** Once hypotheses H_0 and H_1 have been defined, that a test has been chosen to address these hypotheses (see below), and that parameters for this test have been calculated, one must choose a level of significance. The term $p<0.05$ is the arbitrary value that is generally accepted to be significant. This means that there must be $<$ a 5% possibility of falsely detecting a significant difference. Now describe how the p value relates to the different types of errors associated with elementary tests.

Type I and Type II Errors. If a hypothesis H_0 is rejected when it should be accepted, it can be said that a type I error has been made. If a hypothesis H_0 is accepted when it should be rejected, it can be said that a type II error has been made. In either case, a wrong decision or judgment has occurred. This is not a simple matter because decreasing one error type usually leads to increasing the other error type. One way of getting around this problem is just to set your significance level at 0.05 (and not at 0.01 or 0.001). In this way, you are balancing between type I and type II errors in your decision making process. One way to decrease both error types is to increase the size of the sample. However, two ways of analyzing the same size dataset (i.e., two types of inference test) might have different efficiency, so that the more efficient might give better performance on both error types. As an example of type I and type II errors, let us imagine that there is a significant difference between the average of blood pressure measured from a population of patients and the general population at $p = 0.05$. Then there will be a 5% chance that our statement is false (type I error). This means that if we repeat the test 100 times, when in fact no real effects are present, we will draw a wrong conclusion ~5% of the time that we observe a significant difference. In contrast, if we state that there is no such difference between population of patients at $p = 0.05$, there is not a 5% chance of being wrong, but usually more (type II error). This is why, in general, when accepting hypothesis H_0, no conclusions should be drawn about the results of an experiment. The exact calculation of type II error usually depends on the size of the actual effect in the population, hence it is usually described by curves as a function of effect magnitude.

Correction for Multiple Comparisons. When multiple tests are performed, the probability that one of them is significant by chance becomes larger. As for type I error, if 100 tests are performed with significance threshold of $p = 0.05$, when in fact no real effects are present, then on average about five of them will indicate significance, but will be false positives. This is the case, for example, when processing biophysical images, such as magnetic, resonance imaging data: a collection of values is acquired for each coordinate on a three-dimensional (3D) grid and a statistical test must be performed on this data. The same problem may arise when processing time series data. The standard conservative approach developed by Bonferroni (3) consists of dividing the p-value threshold by the number of comparisons performed. For example, for 100 tests performed at $p = 0.05$, the corrected p value is $0.05/100 = 0.0005$. This is a conservative approach and a less stringent method has been developed by Holm (4): first choose a significance level $p = \alpha$ (e.g., $p = 0.05$). Then compute the exact p-value for each test, which is usually possible using modern computerized approaches. Rank the collection of p-values from smallest to largest. The smallest p-value is tested against α/N, where N is the number of tests. If the smallest p-value is not $<\alpha/N$, stop the procedure. However, if it is $<\alpha/N$, proceed to test the second smallest p-value against $\alpha/(N-1)$, and so on. A variant of the Holm's procedure consist of testing the first p-value against α/N, the second one against $2\alpha/N$, the third one against $3\alpha/N$, and so on. Technical details and theory about multiple comparisons may be found in (5).

Paired/Unpaired Samples. Table 5 distinguishes between paired versus unpaired data samples. For

Table 5. Which Statistical Inference Test to Use for Which Type of Data

	Dataset		
Goal	Binomial or Discrete	Continuous measurement (from a normal distribution)	Continuous measurement, Rank, or Score (from non-normal distribution)
Example of data sample	List of patients recovering or not after a treatment	Readings of heart pressure from several patients	Ranking of several treatment efficiency by one expert
Describe one data sample	Proportions	Mean, SD	Median
Compare one data sample to a hypothetical distribution	χ^2 or Binomial test	One-sample t test	Sign test or Wilcoxon test
Compare two paired samples	Sign test	Paired t test	Sign test or Wilcoxon test
Compare two unpaired samples	χ^2 Fisher's exact test	Unpaired t test	Mann-Whitney test
Compare three or more unmatched samples	χ^2 test	One-way ANOVA[a]	Kruskal-Wallis test
Compare three or more matched samples	Cochrane Q test	Repeated-measures ANOVA[b]	Friedman test
Quantify association between two paired samples	Contingency coefficients	Pearson correlation	Spearman correlation

[a]All statistical tests in this table are described in the text and often instantiated using a numerical example.
[b]Analysis of Variance = ANOVA

unpaired data samples, there is no direct correspondence between values. This may be the case when a specific measure (e.g., blood pressure) is taken from two distinct populations of patients (e.g., patients suffering from heart failure and control patients). The two data samples corresponding to the two groups of patients are said to be unpaired because there is no relationship between them. In contrast, for paired samples, each value in one sample corresponds to a value in the other sample. In the previous example, it could be the case if each patient tested had a twin volunteering to be a control patient. This would also be the case if two assessments were performed on the same patients (e.g., measure of blood pressure before and after taking a drug). Note that paired groups must necessarily be of the same size. Matched/unmatched data samples are an extension of paired/unpaired data samples when there are more than two samples.

Sampling With or Without Replacement. Sampling with replacement means that each item is put back in the data sample after being sampled (so it may be sampled more than once and appear twice or more in a data sample). Sampling with replacement satisfies the requirement that the trials are independent, but when the sample size is small relative to the size of the population, sampling with or without replacement makes little difference. In elementary statistics, a representative sample is synonymous with the concept of a random sample. When sampling from a population of finite size, a sample of n items is a random sample if it is chosen in such a way that any other sample of size n would be equally likely to be chosen. Sampled items can be chosen with or without replacement. Although impractical in many situations, sampling with replacement leads to easier mathematical analysis. When the population is large relative to the sample size, the analytical methods developed for sampling with replacement yield good approximations. A random sample can be chosen by assigning a number to each member of the population, and then choosing at random n numbers (with or without replacement). This can be done by the so-called Monte Carlo method (consisting of random draws) that uses computer-generated (pseudo)-random numbers.

Table 5 indicates which statistical test should be used depending on data type and question type. Most types of questions have already been described when hypotheses H_0 and H_1 were defined. The last row of Table 5, which is concerned with the relationship between data samples (or more specifically the relationship between variables underlying two paired data samples) was not covered. The corresponding question may be formulated as 'Is there any relationship between the two variables (e.g., two paired measurements)?' The H_0 hypothesis is that there is no relationship between the two variables.

Columns of Table 5 contain elementary tests for different types of variables. Elementary tests cover confidence interval estimation and parametric hypothesis testing for situations involving normally distributed samples, including two-sample situations where the purpose is to compare two populations with respect to their means or variances. Other types of elementary confidence intervals are for proportions and difference of proportions, usually based on the binomial distribution or based on the normal approximation to the binomial distribution. Confidence interval estimation for parameters of nonnormal distributions are much more difficult and closed form formulas often do not exist. In these cases, experimenters must use nonparametric statistical tests that only take into account rank ordering of data samples. They may also use resampling statistical tests, which estimates confidence intervals using many computer-generated random resamplings. For practicality, in Table 5, hypothesis testing was divided into three main categories: hypothesis testing on discrete variables, parametric statistical testing on continuous variables, and nonparametric statistical testing on continuous variables. In a separate section, resampling methods will be dealt with, since it may be applied to any type of data. The list of tests is not exhaustive but instead seeks to provide, within the limited scope of this short article, a range of methods to perform statistical inference on different types of data.

Which type of test to use is often one of the most delicate choices an experimenter has to make. For continuous data, for example, one could use at least three tests: a parametric, a nonparametric, or a resampling inference test. Different tests make different assumptions: parametric test, such as the t-test, make the hypothesis that the data is normally distributed. Nonparametric tests make fewer assumptions about the population distribution but require more data samples. Resampling tests make the assumption that the data samples are an accurate representation of the population. There is no ideal test (although some applied statisticians would argue that resampling methods are indeed superior to other methods), and the test to choose often depends on the type of data being processed or common usage in one specific field of research.

Testing Hypothesis on Discrete Variables

For discrete variables, data is most often represented by proportions of different outcomes. As shown in the first column of Table 5, specific tests have been designed to deal and compare proportions between data samples. Some of these tests (as indicated below) can only deal with binomial data samples (success or failure).

Goodness of Fit to Distribution for One Data Sample. A goodness of fit test may be used to compare one data sample to a hypothetical value or distribution. In a goodness-of-fit test the hypotheses are concerned with the distribution itself. For example, a drug has been repeatedly tested on adults and has shown minor side effects in 2.5% of the cases in which it was administered. To validate this drug for treating children, it is given to a sample of 300 children. The goal of this study is to determine if children showed more or less side effects than adults. The hypothesis H_0 is that the distribution of sample data values for children is generally the same as the hypothetical distribution for adults. The hypothesis H_1 is that the distribution of sample data values for children generally differs from the hypothetical distribution for adults. Table 6 indicates that 13 out of 300 children showed an abnormal reaction to the drug. The

Table 6. Measured and Expected Frequencies of Side Effect for 300 Children Treated With a Test Drug

	Children	Expected value
Side effect	13	7.5
No side effects	287	292.5
Total	300	300

second column in Table 6 indicates the expected values from the theoretical distribution (2.5% of cases for 300 subjects is 7.5 individuals; it is not so important that the expected value is not a whole number since this distribution is only theoretical).

The χ^2 value is then simply calculated by comparing the expected frequencies e_1 (7.5 individuals showing side effects) and e_2 (292.5 individuals showing no side effects) to the observed frequencies O_1 (13) and O_2 (287) using the formula:

$$\chi^2 = \frac{(O_1 - e_1)^2}{e_1} + \frac{(O_2 - e_2)^2}{e_2}$$

or more generally

$$\chi^2 = \sum_i \frac{(O_i - e_i)^2}{e_i} \tag{13}$$

where O_i is the frequency observation in row i and e_i is the corresponding expected frequency. The degrees of freedom is equal to $(n-1)$, where n is the number of rows in the table. Once the χ^2 value and the degrees of freedom have been calculated, the critical value for χ^2_{crit} can be looked up in Table 7 for a given level of significance. If $\chi^2 > \chi^2_{\text{crit}}$, we reject hypothesis H_0 in favor of hypothesis H_1 and conclude that the data support the hypothesis that there is a difference between the sample data and the theoretical distribution at the 5% level of significance.

In the example shown in Table 6,

$$\chi^2 = \frac{(13 - 7.5)^2}{7.5} + \frac{(287 - 292.5)^2}{292.5} = 4.13$$

with 1 degree of freedom (2 rows minus 1). For a test at the 5% level of significance ($p = 0.05$) with 1 degree of freedom, χ^2_{crit} in the χ^2 table (Table 7) is equal to 3.84. Since $4.13 > 3.84$, the hypothesis that the proportion of children having side effects in the same as that of adults is rejected. Comparing actual and expected frequencies in Table 6, we conclude that children have higher occurrences of side effects than adults for this drug.

Note that the construction of the χ^2 table is relatively simple. One can simply assume that a known population (whose expected distribution is known) is sampled several times and that the χ^2 value is computed for each of these samples. The histogram of these observed χ^2 values, when in fact no real effects are present, is an approximation to the χ^2 distribution for the null hypothesis (Fig. 4). The tails of this distribution may be used to set thresholds for significance testing (if an observed χ^2 value ends up in the tail of the distribution, then it is likely that it does not originate from the known population). For example, the χ^2 value for a data sample is significantly different from the χ^2 standard distribution at $p = 0.05$ if it lies in the lower or upper tails each containing only 2.5% of the values of the standard χ^2 distribution.

Table 7. χ^2 Distribution of Critical Values[a]

	$p = 0.05$	$p = 0.01$	$p = 0.001$
1df	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46
7	14.07	18.48	24.32
8	15.51	20.09	26.13
9	16.92	21.67	27.88
10	18.31	23.21	29.59
11	19.68	24.73	31.26
12	21.03	26.22	32.91
13	22.36	27.69	34.53
14	23.69	29.14	36.12
15	25.00	30.58	37.70
16	26.30	32.00	39.25
17	27.59	33.41	40.79
18	28.87	34.81	42.31
19	30.14	36.19	43.82
20	31.41	37.57	45.32
21	32.67	38.93	46.80
22	33.92	40.29	48.27
23	35.17	41.64	49.73
24	36.42	42.98	51.18
25	37.65	44.31	52.62
26	38.89	45.64	54.05
27	40.11	46.96	55.48
28	41.34	48.28	56.89
29	42.56	49.59	58.30
30	43.77	50.89	59.70
35	49.80	57.34	66.62
40	55.76	63.69	73.41
50	67.51	76.15	86.66
60	79.08	88.38	99.62
70	90.53	100.42	112.31
80	101.88	112.33	124.84
90	113.15	124.12	137.19
100	124.34	135.81	149.48

[a]To use this table, choose a p value (column) and read the value for your calculated degrees of freedom (*df*). If your calculated χ^2 value is larger than the one you read in the table, the test you performed is significant (see text for details).

Binomial Test for Binomial Variables. For data samples that are assumed to be obeying the binomial distribution, it is possible to compute exact p values as explained in a previous section. For example, a coin is tossed 10 times to determine if it returns fair results or not. It returns 9 heads. The hypothesis H_0 is that the coin is fair and that the probability of obtaining a head is 0.5. The H_1 hypothesis is that the coin is biased toward head. Using the binomial distribution, the probability of obtaining an equal or more extreme number of heads than the one measured needs to be computed. The probability of obtaining 9 heads

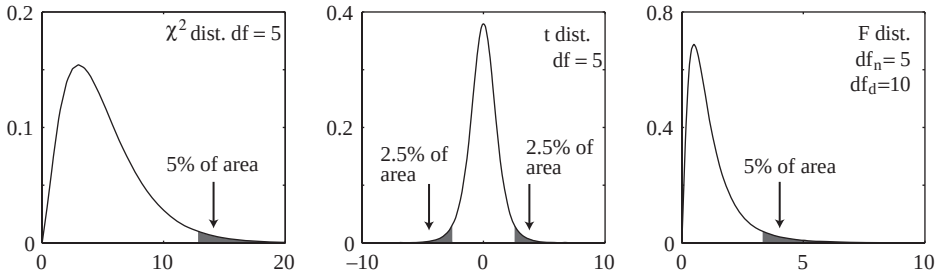


Figure 4. Standard distributions (χ^2, t, and F). Tails of these distributions are used to determine significance thresholds (see text).

or more is

$$P(X \geq 9) = P(X = 9) + P(X = 10) = \binom{10}{9} 0.5^9 (1 - 0.5)^{10-9} + \binom{10}{10} 0.5^{10} = 0.011$$

It appears that this result would appear by chance in only 1.1% of coin tossing trials if the coin is returning fair results. If $p < 0.05$ is considered to be the standard threshold for significance, it can be concluded that the coin does return more heads than a fair coin at the 5% level of significance. Note that this was a one-sided test, assuming that there was prior knowledge that the coin will be biased toward heads (based, e.g., on the aspect of the coin): for a two-sided test that would assess if the coin is fair in returning both faces and heads, it would be necessary to add the probability of obtaining both 9 to 10 heads and 9 to 10 faces.

Sign Test to Compare Paired Samples. This test is best illustrated by an example. To determine if drug A is more effective than a drug B for pain control, 10 patients are tested with these two drugs (with an interval of 7 days to prevent carry over effects) and asked if the drug was effective in controlling their pain. Hypothetical results are shown in Table 8, with + signs indicating a positive effect of the drug and − signs indicating no effect of the drug. The last row indicates the sign of the difference between the first two rows: a + sign indicates that drug A is performing better than drug B and a − sign indicates that drug B is performing better than drug A. When the outcome is the same, the cell is left empty. If the two drugs are equally effective, and if the sample is large enough, then there should be approximately equal numbers of + and − signs in the last row. The expected number of + signs (6 out of 7 nonempty cells) using binomial probability (note that for a large number of values, the approximation of the binomial distribution by the normal distribution may be used). We need to compute the probability of obtaining an equally or more extreme number of + or − than the one obtained, hence to compute $P(0, 1, 2, 5, 6, 7)$:

$$P(+ = 0, 1, 2, 5, 6, 7) = \binom{7}{0} 0.5^0 (1 - 0.5)^{7-0} + \binom{7}{1} 0.5^1 (1 - 0.5)^{7-1} + \cdots = 0.45$$

Although it seems that drug A is a better pain killer than drug B, the p value did not reach significance ($p > 0.05$). In other words, H_0, the hypothesis that the two drugs are performing equally well cannot be rejected. This type of test applied to binomial variables is also sometimes called the Mc Nemar's test.

χ^2 Test to Compare Two or More Unpaired Samples. The χ^2 test allows the comparison of proportions observed in several groups under two or more conditions. Suppose that we wish to determine which of four prosthetic devices perform better for improving muscular response. Each of the four devices is implanted in four random samples of 100 patients each. For each patient, a clinician then estimates if there has been no improvement or partial to full restoration. Data is cross-classified as shown in Table 9. The test described here is usually called the χ^2 test of independence because it aims at finding if results from different groups can or cannot originate from the same population.

Here, the objective is to determine whether improvement is independent of the type of device. If it is the case, then the proportion of responses with no improvement and partial to full restoration should be similar for all four types of devices. The χ^2 test allows the comparison of the actual proportion of responses to each type of device to the idealized proportions, where all types of devices perform equally well. These proportions (also called expected frequencies) are calculated by pooling the responses for all types of devices. For instance, in Table 9, irrespective of the device type, there are 120 patients showing no restoration and 280 patients showing some degree of restoration, so the expected frequency for no restoration is 30% and the expected frequency for partial to full restoration is 70%.

Table 8. Success (+) or Failure (−) of Drug A and B for Reducing Pain in 10 Patients

Patient	1	2	3	4	5	6	7	8	9	10
Drug A	+	+	+	−	+	+	+	+	+	+
Drug B	+	−	−	+	+	−	−	+	−	−
Sign		+	+	−		+	+		+	+

Table 9. Results of Improvement in Muscular Response Following Implantation of an Electronic Device Available in Four Types

	Type of Device				
	1	2	3	4	Total
No improvement	35(30)[a]	40(30)	35(30)	10(30)	120
Partial-to-full restoration	65(70)	60(70)	65(70)	90(70)	280
Total	100	100	100	100	400

[a]Numbers are observed frequencies and number in parentheses are expected frequencies.

As for the simpler example earlier in this section comparing a sample data distribution to a theoretical distribution, the χ^2 is simply calculated by comparing the expected frequencies, denoted by $e_{i,j}$ for device i (where i ranges from 1 to 4) and outcome j (where $j=1$ indicates no restoration and $j=2$ indicates partial to full restoration), to the observed frequencies $O_{i,j}$ using the formula:

$$\chi^2 = \sum_{i,j} (O_{ij} - e_{ij})^2 / e_{ij} \tag{14}$$

The degrees of freedom is equal to *(r – 1) (c – 1)*, where r and c are the number of rows and columns in the table. Once the χ^2 value and the degrees of freedom have been calculated, the critical value for χ^2_{crit} can be looked up in Table 7 for a given level of significance. If $\chi^2 > \chi^2_{\text{crit}}$, then there is a significant difference between the groups being compared.

For the example shown in Table 9,

$$\chi^2 = \frac{(35-30)^2}{30} + \frac{(65-70)^2}{70} + \frac{(40-30)^2}{30} + \cdots = 26.2$$

The degrees of freedom is $(4-1)(2-1)=3$. In this example, for a test at the 5% level of significance ($p=0.05$) and three degrees of freedom, Table 7 indicates that $\chi^2_{\text{crit}} = 7.82$. Since $26.2 > 7.82$, the hypothesis that all four devices are equally effective is rejected. It can be seen that device type 4 is most effective. In fact, further analysis supports the conclusion that differences between the other device types can be explained by sampling variation, and that there is a statistically significant difference between the first three device types taken together and the fourth device type. The additional analysis is sensible because the first three types are different vintages of essentially the same design, whereas the type four device is an experimental version of a fundamentally different design.

The χ^2 test may be used on a table of any size and not necessarily on binomial variables. For the example shown in Table 9, three possible outcomes could be imagined: no improvement, partial restoration, and full restoration. This would have added a row to Table 9 but the χ^2 formula (Eq. 14) would still apply.

Quantify Relationship between Variables. Classification in a table often reflects characteristics of individuals or objects, so they are often referred to as attributes. A measure of the degree of relationship, association, or dependence of two attributes (and the associated variables in the population) is called the coefficient of correlation. It is given by

$$r = \sqrt{\frac{\chi^2}{N(\min(\text{No. rows}, \text{No. columns}) - 1)}} \tag{15}$$

where χ^2 represents the value computed from the χ^2 table; N is the total number of observations, and min(No. rows, No. columns) represents the smaller number between the number of rows (No. row) and the number of columns (No. columns). r can only take values between 0 and 1. The closer r is to 1, the greater the association between the two (or more) columns of the table. To determine if a value of r is significant or not, χ^2 tests previously described in this section may be used.

Parametrical Hypothesis Testing on Continuous Variables

A parametric statistical hypothesis assumes that the data sample originates from a population that fits a specific model (most often the normal model). This is usually the case when recording a measure that fluctuates around a fixed mean because of environmental noise. Before running any statistical tests, one must verify that the data distribution is consistent with the normal distribution. First, plot the histogram to check that the distribution's overall shape is similar to that of the normal distribution. You may then perform a goodness of fit test with the normal distribution. In a goodness-of-fit test, the hypotheses are concerned not with the parameters, but with the distribution itself. For example, H_0: X has a normal distribution; H_1: X does not have a normal distribution. This may be done using the χ^2 goodness of fit test (mentioned in the previous section) applied to the data histogram frequencies compared to expected values calculated from the normal distribution (by integrating Eq. 9 using Eq. 5). Other goodness-of-fit tests are the Kolmogorov–Smirnov, Cramer–Von Mises, and Anderson–Darling. There are also tests when H_0 involves some specific distribution, for example, the Shapiro–Wilk test for normality. Most computer packages incorporate such tests.

One-Sample t-Test to Compare One Data Sample to a Hypothetical Distribution. This test is used to determine if a data sample belongs to a population with mean μ and standard deviation σ (the hypothesis H_0 is that it does belong to this population). This test applies to continuous or noncontinuous data that have a distribution that is not significantly different from normal. First, check that the standard deviation of the data sample is similar to the

Table 10. The *t* Distribution of Critical Values[a]

One-Tailed	0.1	0.05	0.025	0.01	0.005	0.0001
Two-Tailed	0.2	0.1	0.05	0.02	0.01	0.0002
df						
1	3.078	6.314	12.71	31.82	63.66	318.3
2	1.886	2.920	4.303	6.965	9.925	22.33
3	1.638	2.353	3.182	4.541	5.841	10.21
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.611
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
50	1.295	1.676	2.009	2.403	2.678	3.261
60	1.296	1.671	2.000	2.390	2.660	3.232
80	1.292	1.664	1.990	2.374	2.639	3.195
100	1.290	1.660	1.984	2.364	2.626	3.174
1000	1.282	1.646	1.962	2.330	2.581	3.098
inf.	1.282	1.64	1.960	2.326	2.576	3.091

[a]To use this table, find your degrees of freedom in the *df* column (or a lower one if yours is not present in the table). Then, look up the probability in the top row ($p = 0.05$ is a test of significance at 5%). If your calculated t value is larger than the one you read in the table, the test you performed is significant (see text for details).

population's standard deviation σ (within a twofold range). For a data sample containing N values that has a mean M and standard deviation SD, the variable t is defined as

$$t = \frac{M - \mu}{SD}\sqrt{N} \tag{16}$$

The degrees of freedom associated with t is equal to $df = N - 1$. After calculating t and df, set up a threshold for significance (i.e., $p < 0.05$) and look up t_{crit} critical value in Table 10. In the t-test table, you may choose either one- or two-tailed t-test critical values. One-tailed t-tests are used when there is some prior knowledge to predict the direction of the difference. Most commonly, two-tailed t-tests are used when there is no such knowledge. If the calculated t value is $> t_{crit}$, there is a statistically significant difference between the data sample and the hypothetical distribution (the null hypothesis H_0 is rejected).

As for the χ^2 table, building the t-test table is straightforward. One may assume that, for a given degree of freedom, a known population (with a normal distribution) is sampled several times and that the t value is computed for each of these samples (M should on average be the same as μ since it is the theoretical population which is being sampled). The histogram of these observed t values, obtained when in fact no real effects are present, is an approximation to the t distribution (Fig. 4, middle panel). The tails of this distribution allow for threshold setting for significance (as for the χ^2, if an observed value ends up in the tail of the distribution, then it is likely that it does not belong to this distribution). Note that for an infinite number of degrees of freedom, the t distribution is equal to the normal distribution.

For example, in the past, a machine has been producing washers having a thickness of 0.06 in. (0.15 cm) on average.

To test if the machine is still working properly, 10 washers of size (0.065; 0.062; 0.060; 0.059; 0.061; 0.064; 0.067; 0.064; 0.061; 0.062) are produced. The sample mean is 0.0625 and the sample standard deviation is 0.0025. The t value is equal to

$$t = \frac{0.0625 - 0.06}{0.0025}\sqrt{10} = 3.16$$

A two-tailed t-test is used since there is no *a priori* knowledge about the sampled distribution. At the 5% significance level, $t_{\text{crit5\%}}$ is equal to 1.83. Since $t > 1.83$, it can be concluded that there is a significant difference between the expected washer thickness and the observed one (reject hypothesis H_0, which assumes that the sample distribution has a mean of 0.06 in., 0.15 cm). However, at 0.5% significance level, $t_{\text{crit1\%}}$ is equal to 3.25. Since $t < 3.25$, it cannot be concluded that such a difference exists at this level of significance (hypothesis H_0 cannot be rejected).

Paired t-Test to Compare Paired Data Samples. This test applies to two paired samples of continuous or noncontinuous data that have a distribution nonsignificantly different from normal and similar standard deviations (with less than twofold difference). First calculate the difference between each pair and average them (D_{av}) (note that differences in values also have to be normally distributed). Then calculate the value of t using

$$t = \frac{D_{av}}{SD}\sqrt{N} \tag{17}$$

where SD is the standard deviation of the difference between each pair. Since the accuracy of a statistic is influenced by the population size, the degrees of freedom (df) or the number of independent parameters used in the calculation of the test statistic must then be calculated. The degrees of freedom is equal to the degrees of freedom used in calculation the sample SD, that is, the number of pairs minus 1: $df = N - 1$.

Finally, as for the one sample t-test, set up a threshold for significance, look up t_{crit} critical value in Table 10, and compare it to the calculated value. If the calculated t value is $> t_{\text{crit}}$, there is a statistically significant difference between the two groups (the null hypothesis H_0 is rejected).

For example, to test if a newly designed electronic blood pressure (BP) device returns similar (hypothesis H_0) or different (hypothesis H_1) readings compared to an old manual blood pressure device, readings on 10 patients are performed and presented in Table 11 (only systolic pressure in Hgmm is reported in the table).

First, ensure that the two standard deviations are similar (14.1 for the electronic BP device and 13.5 for the manual BP device). To calculate the t value, compute the difference between each pair, check that their distribution is normal, and then average them, $D_{\text{av}} = ((121-115)+(130-131)+\cdots)/10 = 2.8$. The standard deviation of the difference is SD = 2.57, and the degrees of freedom is 9 (10 readings minus 1). Thus the t value is equal to

$$t = \frac{2.8}{2.57}\sqrt{10} = 3.44$$

At the 5% level of significance, for 9 degrees of freedom, $t_{\text{crit5\%}}$ is equal to 2.26. Since $t > 2.26$, it can be concluded that the two devices return different averages (reject hypothesis H_0). The newly devised electronic BP device probably has to be recalibrated to better match the readings of the manual one.

Unpaired t-Test to Compare Unpaired Data Samples. An unpaired t-test aims to compare two unpaired data samples and applies to continuous or noncontinuous data that have a distribution not significantly different from normal. Sample sizes should be similar (with less than twofold difference) for the two groups and, if $n < 30$, variances should also be similar (with less than twofold difference). If the t-test is used in other circumstances, the results will have no meaning.

The most common way of calculating the t-statistics for unpaired data samples is to use the pooled variance estimate (it is also possible to use unpooled variance estimates, but this is less common and will not be presented here). First, calculate the unbiased pooled variance estimate:

$$V = \frac{V_{\text{A}}(N_{\text{A}} - 1) + V_{\text{B}}(N_{\text{B}} - 1)}{N_{\text{A}} + N_{\text{B}} - 2} \tag{18}$$

Then estimate the standard error of the difference of the means:

$$\text{SE} = \sqrt{V(1/N_{\text{A}} + 1/N_{\text{B}})} \tag{19}$$

Then the t statistics is the difference of the means divided by its estimated standard error:

$$t = \frac{M_{\text{A}} - M_{\text{B}}}{SE} \tag{20}$$

where M_{A}, and M_{B} are the means of groups A and B, respectively, and where V_{A} and V_{B} are the variances of groups A and B, respectively. For this test, the number of degrees of freedom is equal to the total number of points minus 2, because two means are estimated.

$$df = (N_{\text{A}} + N_{\text{B}}) - 2$$

Finally, set up a threshold for significance ($p < 0.05$, e.g.), and look up the critical value t_{crit} in Table 10 (see the section above on one sample t-test for the difference between one- and two-tailed t-tests). If the calculated t

Table 11. Systolic Blood Pressure in Hg·mm Measured in 10 Patients Using Either a New Electronic Device or an Old Manual Device

Patient	1	2	3	4	5	6	7	8	9	10
Electronic BP device	121	130	129	113	145	132	110	116	125	155
Manual BP device	115	131	127	111	140	131	111	111	121	150

Table 12. Heart Rate in Beats per Second of Control and Test Patients Suffering from Heart Failure

HF patients	78	81	88	76	93	112	83	96		
Control patients	80	71	68	80	95	67	85	69	85	77

value is $> t_{crit}$, there is a statistically significant difference between the two groups (the null hypothesis H_0 is rejected).

For example, to test if patients diagnosed with heart failure have similar (hypothesis H_0) or higher (hypothesis H_1) heart rates than control patients, 15 readings are performed at rest for these two groups of patients A and B of matched age, sex, and ethnicity. Heart rate is reported in beating per minutes in Table 12.

After testing for normality (see for how to test for normality at the beginning of this section), it is ensured that standard deviations for the two data samples are similar ($SD_A = 11.5$ and $SD_B = 9.1$). To calculate the t-value, it is necessary to compute the mean heart rate for each group. For patients suffering from heart failure, $M_A = 88.4$, and for control patients, $M_B = 77.7$(variances are $V_A = 140.3$ and $V_B = 82.9$). Thus the pooled variance estimate is $V = 108$, the standard error of the mean is 4.93 and the t value is equal to

$$t = \frac{88.4 - 77.7}{4.93} = 2.17$$

At 5% significance level for 16 degrees of freedom (10 heart failure patients plus 8 control patients minus 2), $t_{crit1\%}$ is equal to 2.12. Since $t > 2.12$, the data support the fact that patients with heart failure have higher heart rate than controls (hypothesis H_0 cannot be rejected).

One-Way ANOVA for Unmatched Samples. One-way ANOVA is used to test the hypothesis that two or more samples are drawn from the same distribution of values and have the same mean and variance. Unpaired student t-test is a particular case of one-way ANOVA applied to two data samples. As for t-test, ANOVA test applies to continuous or noncontinuous data that have a distribution that is not significantly different from normal. Sample sizes should be similar (with less than twofold difference) for all sample groups and, if $n < 30$, variances should also be similar (less than twofold difference). If the test is used in other circumstances, the test outcome will lead to erroneous conclusions. The basis of ANOVA is the F (Fisher) variable, which combines the unbiased variance between sample groups ($V_{interGroup}$) and the variance within sample groups ($V_{withinGroup}$).

$$F = \frac{V_{interGroup}}{V_{withinGroup}} \tag{21}$$

For several data samples A, B, C, ... of the same size, intergroup variance is defined as

$$V_{interGroup} = \frac{N_A(M_A)^2 + N_B(M_B)^2 + N_C(M_C)^2 + \cdots - N_T(M_G)^2}{N_G - 1} \tag{22}$$

where M_A, M_B, and M_C are the means of sample A, B, C, ... and N_A, N_B, N_C, ... are the number of values in samples A, B, C, ... M_G is the average of all values from all sample groups and N_G is the number of samples. The within sample group variance is defined as

$$V_{withinGroup} = \frac{(N_A - 1)(SD_A)^2 + (N_B - 1)(SD_B)^2 + (N_C - 1)(SD_C)^2 + \cdots}{N_T - N_G} \tag{23}$$

where SD_A, SD_B, SD_C, ... are the standard deviations of group A, B, C, ... and N_T represents the total number of observations (for all data sample pooled together). Degrees of freedom for the numerator of F and the denominator of F are defined as

$$df_{numerator} = N_G - 1$$
$$df_{denominator} = N_T - N_G$$

Note that each variance in Eqs. 22 and 23 is divided by the appropriate degrees of freedom to give unbiased estimate of population variance (assuming the null hypothesis H_0 is true). As for other inference tests, the computed F value is tested against critical F values (Table 13) obtained from the tail of null-hypothesis F distribution (Fig. 4, right panel).

For example, a clinician planning to purchase equipment for electroencephalography compares the signal to noise ratio for three sets of electroencephalographic equipment. For each system that has been made available to him, he records 10 new patients performing standard psychophysical tasks and measures the signal to background noise ratio of the encephalographic equipment (Table 14).

After testing for normality, we must ensure that standard deviations are similar (i.e., no twofold differences). Standard deviation for Brand A is equal to $SD_A = 1.11$; Brand B: $SD_B = 0.75$; Brand C: $SD_C = 0.94$. After calculating $V_{intergroup} = 0.44$ and $V_{winthinGroup} = 0.89$, F may be calculated using Eq. 21

$$F = \frac{0.44}{0.89} = 0.49$$

The degrees of freedom for the numerator is $df_{numerator} = N_G - 1 = 2$. The degrees of freedom for the denominator is $df_{denominator} = 30 - 3 = 27$. Reading $F_{crit} = 2.95$ in Table 13, we may conclude that there is no significant difference (since $F < 2.95$) in terms of signal/noise ratio between the three sets of EEG equipments (accept hypothesis H_0).

One-Way ANOVA for Matched Samples. One-way ANOVA may also be used to compare paired sample groups. In fact, since for matched samples, one may

Table 13. The F Distribution of Critical Values at $p = 0.05$ for ANOVA tests[a]

$df2 \backslash df1$	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	100	∞
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62	8.59	8.57	8.55	8.54
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.46	4.43	4.41	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.77	3.74	3.71	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.83	2.79	2.76	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.70	2.66	2.62	2.59	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.57	2.53	2.49	2.46	2.41
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.47	2.43	2.38	2.35	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.38	2.34	2.30	2.26	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.31	2.27	2.22	2.19	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.25	2.20	2.16	2.12	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.19	2.15	2.11	2.07	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.15	2.10	2.06	2.02	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.11	2.06	2.02	1.98	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.07	2.03	1.98	1.94	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04	1.99	1.95	1.91	1.84
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	1.98	1.94	1.89	1.85	1.78
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.94	1.89	1.84	1.80	1.73
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.90	1.85	1.80	1.76	1.69
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.87	1.82	1.77	1.73	1.66
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84	1.79	1.74	1.70	1.62
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.04	1.96	1.88	1.79	1.74	1.68	1.63	1.56
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.74	1.69	1.64	1.59	1.51
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05	1.97	1.89	1.81	1.71	1.66	1.60	1.55	1.47
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.87	1.78	1.69	1.63	1.58	1.52	1.44
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.65	1.59	1.53	1.48	1.39
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.89	1.81	1.72	1.62	1.57	1.50	1.45	1.35
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.88	1.79	1.70	1.60	1.54	1.48	1.43	1.33
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.77	1.68	1.57	1.52	1.45	1.39	1.28
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.80	1.72	1.62	1.52	1.46	1.39	1.32	1.19
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.77	1.69	1.59	1.48	1.42	1.35	1.28	1.12
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.76	1.68	1.58	1.47	1.41	1.33	1.26	1.08
∞	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.46	1.40	1.32	1.25	1.03

[a]To use this table, read the value at the intersection of the numerator's degrees of freedom ($df1$) and the denominator's degrees of freedom ($df2$). If your calculated F value is larger than the one you read in the table, the test you performed is significant (see text for details).

analyses either the rows or the columns of a table, the formula given here may be used both for rows or columns, and are usually associated with two-way ANOVA. The formula for the F (Fisher) variable is now equal to

$$F = \frac{V_{\text{interGroup}}}{V_{\text{error}}} \tag{24}$$

The variance due to error or chance is defined as

$$V_{\text{error}} = \frac{\sum_{j,k}(x_{jk} - M_{j.} - M_{.k} - M)^2}{(N_R - 1)(N_C - 1)} \tag{25}$$

where x_{jk} are all the elements in the array, $M_{j.}$ are the row means, $M_{.k}$ are the column means, M is the global array mean, N_C is the number of columns, and N_R the number of rows. The degrees of freedom for the numerator and denominator are now defined as

$$df_{\text{numerator}} = N_R - 1 = N_G - 1$$
$$df_{\text{denominator}} = (N_R - 1)(N_C - 1)$$

Using the same example as shown in Table 14, and now assuming that the data samples are paired (EEG systems were tested with the same patients), the intersubject variance $V_{\text{errort}} = 1.04$ can be computed, and

$$F = \frac{0.44}{1.04} = 0.42$$

The degrees of freedom for the numerator is $df_{\text{numerator}} = N_G - 1 = 2$. The degrees of freedom for the denominator is $df_{\text{denominator}} = (N_R - 1)(N_C - 1) = (3 - 1)(10 - 1) = 18$. For a test at 5%; significance, reading

Table 14. Signal/Noise Ratio for 10 Patients and for Three Brands of EEG Systems

Brand A	1.87	3.88	2.68	1.19	0.93	0.38	2.69	1.8	0.39	1.62
Brand B	2.48	1.71	3.05	1.58	1.7	3	0.47	2.11	2.18	2.22
Brand C	2.29	1.49	2.52	1.26	3.71	2.14	2.33	2.79	2.61	0.29

Table 15. Example of Table for a Two-Factor Experiment

	Protocol 1			Protocol 2			Protocol 3		
Brand A	6	8	8	1	0	2	1	4	4
	10	8	2	2	0	1	4	2	2
	10	6	2	1	3	3	3	5	0
Brand B	2	2	10	4	9	8	3	3	6
	6	10	10	5	5	9	5	4	2
	6	2	6	3	7	7	3	5	6
Brand C	6	0	4	5	3	2	6	6	8
	6	4	4	1	1	1	6	0	10
	4	8	8	3	3	2	4	8	6

$F_{\text{crit}} = 3.55$ in Table 13, it can be concluded that there is no significant difference (since $F < 3.55$) in terms of signal to noise ratio between the three sets of EEG equipments.

Note that one could argue that instead of using ANOVA analysis, t-tests could be performed between each pair of samples. Although this is possible, the ANOVA test is more sensitive than a series of paired t-tests because it processes all data samples simultaneously.

Two-Way ANOVA for Two-Factor Experiments. This type of test is being used for experiments with two factors or two attributes. In the example above, to test the reliability of the EEG equipment, the clinician might want to perform three experimental protocols and measure the signal to noise ratio in each of these protocols. The two factors are now the three sets of EEG equipment and the three protocols as shown in Table 15.

In each of the cells of Table 15, the clinician recorded nine values. In the case of only one value per cell, the analysis would be similar to the one-way ANOVA (row and column data may be analyzed separately using one-way ANOVAs for matched samples). However, if several values are recorded for each cell (several subjects, e.g.), one must use the repeated measures two-way ANOVA test. This test is especially interesting because it is possible to test for interaction between variables. Hypothesis H_0 would be that there is no significant relationship between brands and type of protocol and Hypothesis H_1 would be that there is indeed such a relationship. Running a repeated measures two-way ANOVA test under any software will return 3 p-values: the first value is for significant differences between rows; the second value is for significant differences between columns; the last p-value is for the interaction between columns and rows. In the case of Table 15, the p-value for the columns (protocol) is 0.0004 indicating a significant difference between protocols. As observed in Table 15, the values for the first protocol are indeed higher than the values for other protocols. The p-value for the different rows (device brand) is not significant ($p = 0.22$). The p-value for the interaction between brand and protocol is 0.0006. In fact, it appears that the device of brand B returns higher values for protocol 2 than other brands, and that the device of brand C returns higher values for protocol 3 than other brands.

Experimental design and ANOVA in its many variations is perhaps the most important statistical methodology for experimenters, and the literature is immense. Extreme care should be taken when choosing an ANOVA test. For example, there are different ways to treat multifactor ANOVAs analytically when the number of observations is unequal among the treatment combinations (called unbalanced designs). A nontechnical discussion is the classic Planning of Experiments by Cox (6). Other general introductions are Refs. (1,7–12).

Regression and Correlation. Regressions and correlations aim at determining relationship between variables. We may wish to determine if there is a significant correlation between independent and dependent variables, the independent variable being set by the experimenter, and the dependent variable being measured. For example, to test the reliability of a device, an experimenter may change the temperature of the room where the device is being tested (independent variable), and see if this change affects measures returned by the tested device (dependant variable). Regression and correlation can also be used to estimate the relationship between two (or more) dependent variables.

The first step in determining the relation between two variables is to plot values of one variable versus values of the other variable. This is usually called a scatterplot (Fig. 5). From the scatterplot it is often possible to visualize a smooth curve that approximates the data. If it is a straight line, then the least-squares regression method may be used. Otherwise, other curve fitting procedures may be used. It is sometimes useful to plot scatterplots of transformed variables (e.g., log transformation of values the in first variable versus values of the second variable).

The method of least-squares computes the best linear regression between two variables. Specifically, for two variables X and Y, the data consist of n pairs $(x_1, y_1),\ldots,(x_n, y_n)$. For all values of X and Y, we wish to find the parameter a and b such that

$$Y = aX + b \tag{26}$$

Assuming the jittering of points along the straight line is normally distributed, parameters a and b may be obtained using the formula

$$a = \frac{N\sum x_i y_i - (\sum x_i)(\sum y_i)}{N\sum x_i^2 - (\sum x_i)^2} \tag{27}$$

$$b = \frac{1}{N}\left(\sum y_i - a\sum x_i\right) \tag{28}$$

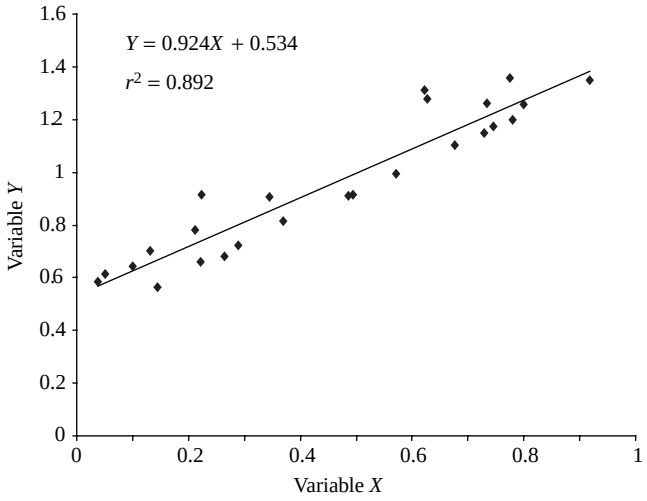


Figure 5. A typical scatterplot with the least-square line drawn through the data points. The r^2 value as well as the best fit equation is indicated on the diagram. The t-value is equal to 9.24 and indicate a sigdnificant relationship between X and Y (at $p = 0.05$, for 22 degrees of freedom, $t_{\text{crit}} = 2.07$).

To draw the linear regression line, y_i^{est} values may be calculated using Eq. 44 for all values of X. A sample-based measure of the strength of the linear association between the X and Y variables is the sample correlation coefficient (also known as the Pearson correlation coefficient) defined by

$$r = \pm\sqrt{\frac{\text{explained-variation}}{\text{total-variation}}} = \pm\sqrt{\frac{\sum(y_i^{est} - M_Y)^2}{\sum(y_i - M_Y)^2}} \quad (29)$$

r may also be expressed using the original variables X and Y.

$$r = \frac{\text{cov}(X,Y)}{SD_X.SD_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - M_X)(y_i - M_Y)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - M_X)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - M_Y)^2}} \quad (30)$$

where M_X and M_Y (SD_X and SD_Y) are the mean (the standard deviation) for X and Y, respectively, and cov(X,Y) is the covariance between X and Y (the numerator on the right of Eq. 30 is equal to cov(X,Y) and the denominator is equal to $SD_X{}^* SD_Y$). Necessarily $-1 \le r \le 1$. Positive (respectively negative) values of r indicate that large values (respectively small values) of X are associated with large values (respectively, small values) of Y. Values of r near 0 indicate little or no linear association. Interpretation must be done with care because there are many reasons for the presence or absence of a correlation. Also, comparing r values may be misleading as a value of $r = 0.6$ does not mean that the linear relationship is twice as strong as $r = 0.3$. On the other hand, r^2, called the sample coefficient of determination, represents the proportion of the total variation in the sample values of Y that can be "explained" by a linear relationship as in Eq. 26. Thus $r^2 = (0.3)^2 = 0.09$ versus $r^2 = (0.6)^2 = 0.36$ indicates a 9% versus 36% accountability for total variability by the proposed linear relationship.

To test if the linear correlation between the two variables is significant, different tests may be used. The null hypothesis H_0 states that there is no relationship between the two variables. A t-test (with degrees of freedom equal to $N - 2$) may be used if the expected population correlation coefficient between variable X and Y is 0 and if we expect the correlation coefficient to be normally distributed when random samples of X and Y are drawn. The variable t is defined as

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \quad (31)$$

More details for determining if correlation coefficients are significant or to compare between correlation coefficients may be found in Spiegel and Stepens (2).

As shown in Fig. 5, most regression computer packages will output scatterplots, and correlation coefficients. Residuals plots (not shown here) indicate if the distribution of distance between estimated and actual values of Y. A histogram of these residuals should be normally distributed (computing the parameter a, b, and the coefficient of correlation r requires that these residuals are normally distributed with mean 0 and a constant standard deviation irrespective of the X values).

A comprehensive presentation of regression methods for linear and nonlinear regression is given in Refs. 2,7, 13.

Nonparametric Testing

Elementary tests mentioned in the previous section require that the distribution of values in the population be normally distributed. In practice, this assumption may not hold so statisticians have devised tests that are less dependent of population distribution. Nonparametric or distribution-free statistical methods generally are not concerned with inferences about parameters of distributions and assume little or no knowledge about the distributions of the underlying populations. Their primary advantage is that they are subjected to less restrictive assumptions than their parametric counterparts. Moreover, the data need not be quantitative (data values may indicate ranks on an ordinal scale). However, a disadvantage of nonparametric methods is that that they may not utilize all the information in a sample, consequently requiring a larger sample than the parametric version to attain the same Type II error (see error types).

The χ^2 goodness-of-fit tests previously mentioned is an example of a nonparametric test. Other nonparametric tests make various hypotheses for medians (or means of a symmetric distribution) and differences in location and/or variability of two populations. There are also tests for randomness, independence, and association among random variables. Relatively elementary texts that give a fairly broad and complete coverage of nonparametric methods are Refs. 14 and 15.

Compare Sample Distribution to a Hypothetical Distribution. As for binomial and discrete data, a χ^2 goodness-of-fit test may be performed. For continuous data, a χ^2

Table 16. Heart Rate Variability for 10 Patients While Their Pacemaker Is Switched On or Off, and Calculation of Signed Rank for Wilcoxon Test

Patient	1	2	3	4	5	6	7	8	9	10	Sum
Pacemaker off	0.15	0.32	0.25	1.1	0.82	0.83	0.94	0.42	0.48	0.21	
Pacemaker on	0.12	0.19	0.28	0.56	0.37	0.52	0.24	0.73	0.81	0.13	
difference	0.03	0.13	−0.03	0.54	0.45	0.31	0.70	−0.51	−0.43	0.08	
abs difference	0.03	0.13	0.03	0.54	0.45	0.31	0.70	0.51	0.43	0.08	
Rank of abs difference	1.5	4	1.5	9	7	5	10	8	6	3	
Signed rank	1.5	4	−1.5	9	7	5	10	−9	−6	3	23

goodness-of-fit test may be used on the frequency distribution (histogram) of the data compared to a hypothetical distribution.

Sign Test and Wilcoxon Test for Paired Samples. As for binomial and discrete data, a sign test allows the comparison of paired samples (see the beginning of the section for a definition of paired and unpaired samples). A sign test simply involves pair-wise comparisons of measures between the two sample data sets (see sign test for binomial and discrete data). A variation of this test is called the Wilcoxon test, which takes into account the signed rank of the difference between each pair (instead of using all the signs). This is best illustrated using an example. To test if a pacemaker device has any effect on heart rate variability (defined as the standard deviation of heart beat intervals in seconds), 10 patients' heart rate variability are measured while the pacemaker was either switched on or off (Table 16).

The Wilcoxon test begins by taking the difference in heart rate variability between the two conditions for each patient (forth row of Table 16). If a difference is equal to 0 it is eliminated from further consideration, since it provides no useful information. The second step consists of taking the absolutes of the differences, which is accomplished simply by removing all the positive and negative signs (fifth row of Table 16), then ranking these absolute differences from lowest to highest, with tied ranks included where appropriate. Tied rank means that if two values are equal they are first-ordered randomly and then assigned their average rank (see the first and third columns of the sixth row in Table 16). Finally, reattach to each rank the positive or negative sign that was removed from the difference in the transition from row four to row five, and sum up these values. In our case, $W = 23$ and the number of values used in this sum is 10 (degrees of freedom).

If two sets of sample values from the same distribution (which verify hypothesis H_0 that the two samples belong to the same distribution) were to drawn repeatedly and W values were calculated, it would be realized that the distribution (histogram) of W values is close to normal. In fact,

$$z = \frac{W}{SD_W} \tag{32}$$

may be defined, where z is normally distributed with mean 0 and variance 1, and SD_W is the standard deviation of W, which can be shown to be equal to

$$\mathrm{SD}_W = \sqrt{\frac{N(N+1)(2N+1)}{6}} \tag{33}$$

For $N = 10$ values, $\mathrm{SD}_W = 19.6$, so $z = 23/19.6 = 1.17$. As mentioned earlier, the t-distribution is equal to the normal distribution for infinite degrees of freedom. Looking in the last row of the t-table (Table 10), for a significant threshold at $p = 0.05$ (two-tailed), $z_{\mathrm{crit}} = 1.64$ is obtained. Since $z < 1.64$, hypothesis H_0 cannot be rejected. Although it seems that heart rate variability is higher when the pacemaker is switched on, the difference did not reach significance.

Mann–Whiney U Test for Unpaired Samples. The Mann–Whitney U test is similar to the Wilcoxon test. Once more, this test will be illustrated using an example. To compare sensitivity of two hearing aids, the minimum sound a patient can hear using each brand is measured (in dB) and reported in Table 17, where 10 different patients tested each prosthetic device (unpaired samples).

To perform a Mann–Whitney test, first combine all values in an array and assign a rank from 1 to 20 to all these values, assigning tied ranks where appropriate (see Wilcoxon test). The rank for each value is indicated in Table 18.

Then, sum up the ranks for each brand, where $R_A = 80$ is the sum for brand A and $R_B = 130$ is the sum for brand B. A significant difference between the two rank sums implies a significant difference between the two samples. Calculate the U statistic to test the difference between the ranks:

$$U = N_A N_B + \frac{N_A(N_A+1)}{2} - R_A \tag{34}$$

Note that the formula above is symmetrical with respect to A and B. In the hearing aid example, $N_A = 10$ and

Table 17. Patient Maximal Sensitivity (in dB) for Two Brands of Hearing Aids

Brand A	0.1	−1	4.1	2.4	−2.3	3.8	0.9	1.4	0.4	1.2
Brand B	2.7	3.1	5.2	2.1	4.7	1.5	−1.2	3.7	2.8	3.1

Table 18. Rank of Measures for Table 17

Brand A	4	3	18	11	1	17	6	8	5	7
Brand B	12	14.5	20	10	19	9	2	16	13	14.5

$N_B = 10$, so

$$U = 10 * 10 + \frac{10(10+1)}{2} - 80 = 75$$

There is no table for U values. Instead, as for the Wilcoxon test, the table for z values is used because of a property of the U distribution. When calculating the U value repeatedly on samples known not to be statistically different (e.g., two data samples drawn from the responses of the same device), then it can be shown that the repeated U values ($U1$, $U2$, $U3$, ...) have a Gaussian distribution with mean M_U and standard deviation SD_U defined as

$$M_U = \frac{N_A N_B}{2} \tag{35}$$

$$\mathrm{SD}_U = \sqrt{\frac{N_A N_B (N_A + N_B + 1)}{12}} \tag{36}$$

This means that the U distribution can be normalized and that

$$z = \frac{U - M_U}{SD_U} \tag{37}$$

is normally distributed with mean 0 and variance 1.

In the example above, $M_U = 10 * 10/2 = 50$ and $\mathrm{SD}_U = 13.2$, so $z = 3.78$. Looking up the last row of the t-table (Table 10) for a significance level of 5%, we read $z_{crit} = 1.64$. Since $z > 1.64$, hypothesis H_0 can be rejected and it can be concluded that one hearing aid performs better than the other one. Looking at the mean or median for each brand, or for this simple example simply at Table 17, brand A clearly allows patients to hear sounds of smaller amplitudes than brand B. Note that the calculations above are usually not necessary since most statistical software will return the value of U along with its significance level.

Kruskal–Wallis Test for Unmatched Samples. The Kruskal–Wallis H test is a generalization of the Mann–Whitney U test to more than two samples (e.g., three brands A, B, and C of sample sizes $N_A, N_B, N_C, \ldots$ with the total number of samples equal to N). As for the Mann–Whitney test, values from all distributions are sorted and once the sum of the rank for each sample is calculated R_A, R_B, R_C, ... the value of H is given by

$$H = \frac{12}{N(N+1)}\left(\frac{R_A}{N_A} + \frac{R_B}{N_B} + \frac{R_C}{N_B} + \cdots\right) - 3(N+1) \tag{38}$$

It can be shown that, after collecting repeated measures of H from several samples from the same population (verifying the hypothesis H_0 that they originate from the same population), the histogram of H values is very close to a χ^2 distribution with degrees of freedom equal to the number of groups minus one (so the χ^2 table may be used for H). Thus, to use the Kruskal–Wallis test, first calculate H, then compute the degrees of freedom (number of groups minus one), and look up the χ^2 critical value in Table 7. If the calculated H value is larger than the critical value, reject hypothesis H_0.

Friedman Test for Matched Samples. Suppose it is wished to determine if three spectroscopy machines A, B, and C returns the same hematocrit density (density of blood cells in a blood sample). We test the three machines using 20 blood samples (the same blood sample is used for all machines). Since preliminary analysis shows that the readings are not normally distributed, nonparametric test will have to used. To do so, for each blood sample, rank the machines (from 1 to 3) and compute the total rank for each machine T_A, T_B, and T_C. The parameter T_{all} being the sum of all the ranks, the squares deviate SS is equal to

$$SS = \frac{(T_A)^2 + (T_B)^2 + (T_C)^2}{N_G} - \frac{(T_{all})^2}{N_G N} \tag{39}$$

where N_G is the number of groups and N is the number of samples in each group. As for the Kruskal–Wallis test, we may use the χ^2 distribution with degrees of freedom equal to $df = N_G - 1$. In the Friedman test, simply refer to this value as χ^2

$$\chi^2 = \frac{SS}{N_G(N_G+1)/12} \tag{40}$$

If the calculated χ^2 value is larger than the critical value for the specified degrees of freedom, reject hypothesis H_0.

The Spearman'S Rank Correlation Test. Rank methods may also be used to determine the correlation between two variables. Instead of using exact variable values, their ranks may be used. For two sample A and B of the same size, corresponding to two variables X and Y (e.g., lifespans and prices of a family of devices), rank each sample value from 1 to N separately for A and B. Then calculate the difference $D_1, D_2, D_3, \ldots$ between the sorted rank for A and B and compute

$$r_S = 1 - \frac{6((D_1)^2 + (D_2)^2 + (D_3)^2 + \cdots)}{N(N^2-1)} \tag{41}$$

If r_S is close to 0, there is no correlation between the two variables, whereas if it is close to 1 or -1, there is a strong correlation between the two variables. To test if r_S is significantly different from 0, the same t-test as for the Pearson correlation coefficient may be used (replacing r by r_s and using the same degrees of freedom $df = N - 2$).

Resampling Methods

Resampling methods help provide confidence intervals for parameters in situations where these are difficult or impossible to derive analytically. Resampling methods also help

perform statistical inference without assuming a known probability distribution for the data. The bootstrap method consists of drawing random subsamples and the randomization method consists of shuffling data samples.

Bootstrap Method. The bootstrap method is the most recently developed method to estimate errors and other statistics. It is not primarily aimed at performing inference although it may be used to do so, since it provides confidence intervals for the measure of interest. The term "bootstrap" derives from the phrase "to pull oneself up by one's bootstrap" (Adventures of Baron Munchausen, by Rudolph Erich Raspe). Suppose we have a data sample and an estimator (e.g., mean). The basic idea involves sampling with replacement to produce random samples of size N from the original data sample (of size larger than N). Each of these samples is known as a bootstrap sample and provides an estimate of the parameter of interest. Repeating the sampling a large number of times provides information on the variability of the estimator and help define confidence limits. There are N to the power of N, N^N, possible samples, called the ideal bootstrap samples. It is important to emphasis that subsamples are drawn with replacement: for example, for an empirical distribution composed of 2 values (5 and 8), the bootstrap samples are (5,8), (5,5), (8,8), and (8,5) (note that there are $2^2 = 4$ of them). Getting all ideal bootstrap samples becomes unrealistic as N becomes larger, so the Monte Carlo approach (which consists of random draws) is used. The sampling is said to be balanced if each sample value is drawn the same number of times. For each bootstrap sample, let us suppose that the mean is calculated. The standard deviation of the bootstrap distribution for the mean correspond to the standard error (Eq. 19) and may be used in parametrical t-test to compute the t value (Eq. 20), and perform inference testing (assuming normality of the distribution of course). However, this mixture of bootstrap and parametric t-test is relatively unconventional, and it is better to estimate the bootstrap distribution of t-values as explained below.

To perform a statistical inference test using bootstrap, first state a null hypothesis H_0. Null hypotheses for resampling tests are usually vague because there may be many reasons (based on the shape of the distribution) why two samples may differ (whereas when performing a parametric t-test, the nonnull hypothesis states clearly that the means are nonequal). Moreover, bootstrap statistics use the implicit assumption that data samples are representative of the underlying population and in fact do as if the data samples were the population itself. Therefore it is not possible to draw direct conclusions about the underlying population either.

In the case of the heart rate study of Table 12, for example, where comparing a measure (i.e., heart rate) for patients suffering from heart failure (sample A) and control subjects (sample B), the null hypothesis would be "patient suffering from heart failure have abnormal heart rate". One way to test this hypothesis is to perform a bootstrap t-test. Two bootstrap samples are first drawn from the pooled distribution of A and B: sample A′ and B′ of the same size as A and B, respectively. The t- value is then

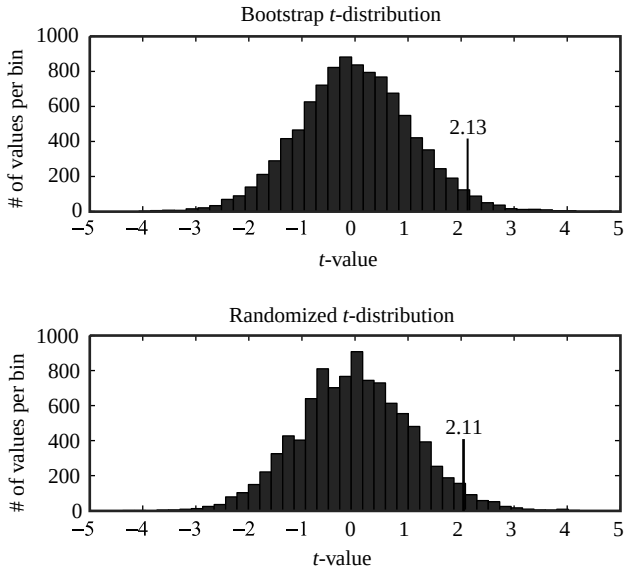


Figure 6. Bootstrap t-distribution for Table 12 (top) and randomized t-distribution for Table 12 (bottom). Since the actual t-value obtained from the original samples in Table 12 ($t = 2.17$) belong to the rightmost 2.5% value in both the bootstrap and the randomized distribution (the 2.5% limit being indicated by a vertical line), it may be considered significant at 5%.

computed using the two bootstrap samples as in Eq. 16. The operation is repeated m times to obtain the distribution of t-values for the null hypothesis. Note that, even if a t- value, is computed it is not assumed normality for the data samples since the distribution of t-values for the null hypothesis is estimated using bootstrap samples. The actual t-value is calculated for the original data samples A and B and tested against the bootstrap t-distribution. If it lies in the lower 2.5% or upper 2.5% tails, then the bootstrap test may be considered to be significant at the 5%; level of significance. In Fig. 6 (top), 10,000 bootstrap t-values were accumulated for the two samples in Table 12. Since the original t-value for Table 12 is equal to 2.17 (see the t-test section) and since it lies in the upper 2.5% of the bootstrap t-value distribution, it may be concluded that the data support the hypothesis that heart rate is affected in patients suffering from heart failure at the 5% significance level.

There are other ways to test for significance using bootstrap, such as the bootstrap-percentile method, or the bootstrap-bca method (see Ref. 16 for a comprehensive reference). In general, it should be remembered that bootstrap methods are designed primarily for estimating characteristics of data samples, not for performing inference tests. Resampling methods specializing in statistical inference are called randomization methods and are describe below.

Randomization Methods. For the purpose of performing paired or unpaired comparisons, randomization methods consist of random permutations of data. Randomization methods are also often called permutation methods or surrogate methods. Specifying the null H_0 hypothesis is

the same as for the bootstrap and involves a vague formulation about the result of the experiment, such as "patient suffering from heart failure have abnormal heart rate" or "the drug treatment does not have an effect on blood pressure".

Randomizing the data is straightforward. Using the same example as for the bootstrap distribution with two unpaired samples A and B of sizes N_A and N_B, a randomization method consists of pooling the data of A and B together (into C), then randomly drawing from C (without replacement) two groups A′ and B′ that have the same size as A and B, respectively (17). Then, compute the estimator (e.g., t-value) for each randomized pair of samples. Repeat this procedure a large number of times to obtain the distribution of the estimator (e.g., t-value) for the null hypothesis. Significance is assessed as for the bootstrap t-test. For example, in Fig. 6 (bottom), 10,000 randomized t-values have been accumulated for the two samples in Table 12. Note that irregularities in the distribution are due to the fact that we are randomizing a relatively small number of values. As for the bootstrap, since the original t-value ($t = 2.17$) lies in the upper 2.5% of the randomized t-values, it may be concluded that the data support the hypothesis that heart rate is affected in patients suffering from heart failure at the 5% significance level. It is reassuring to notice that the upper 5% significance threshold t-value for the bootstrap ($t_{crit} = 2.13$), the randomized ($t_{crit} = 2.11$), and the normal distribution ($t_{crit} = 2.12$) are all similar.

For paired comparisons, the principle is slightly different since we are now randomizing not the sample values but the pairs. For example, for the data of Table 11, one-half of the pairs are selected randomly then shuffled (the value for the first device is now attributed to the second device and vice versa) and the paired t-test value is recalculated (Eq. 17). This procedure is repeated many times. To assess significance, as in the previous paragraph, the original t-value computed using the nonrandomized samples is compared against the distribution of randomized t-value.

This procedure may be generalized to compare an arbitrary number of samples. For example, to compare several unpaired sample, data sample values may be randomized among groups and one-way ANOVA values may be calculated repeatedly. The ANOVA value for the nonrandomized groups is then compared against this ANOVA randomized distribution. Web Ref. 18 provides a clear introduction to resampling methods.

MULTIVARIATE METHODS

Previously the probability distributions involving one variable were discussed, but in many situations there are two or perhaps many interdependent variables, for example, height, weight, daily caloric intake, genetic strain. Data samples involving several variables are called multivariate. Many multivariate analytical methods involve inference for the parameters (means, variances, and correlation coefficients) based on multivariate normal distribution. One such method is known as discriminant analysis and is concerned with the problem of distinguishing between two or more populations on the basis of observations of a multivariate nature. Principal components analysis, cluster, and factor analysis seek to determine relatively few out of possibly many variables that will serve to explain the variability or the interrelationships in the variables. Principal component analysis (PCA) would specifically make each successive component account for as much as possible of the remaining variability uncorrelated with previously determined components. In Fig. 7, data points from two variables are represented. Coordinates of data points on the abscissa axis correspond to values of the first variable and coordinates on the ordinate axis correspond to values of the second variable. The PCA is able to find a first principal axis (labeled one) that accounts for most of the variance of the data. The second principal axis (labeled two) has to be perpendicular to the first principal axis and accounts for the remaining of the variance.

Recent progresses in signal processing and information theory have seen the development of blind source

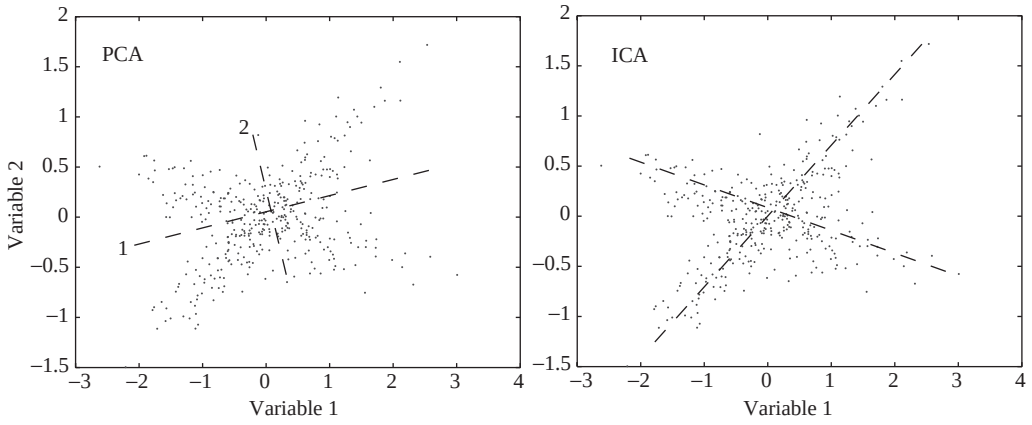


Figure 7. Illustration of PCA and ICA algorithms. The PCA finds axis with maximum variance. By contrast in ICA, the projection of data point on ICA axis is maximally independent.

separation methods, which attempt to find a coordinate frame onto which the data projections have minimal overlap. For example, if two sources of sounds (e.g., a conversation and a CD player) are recorded simultaneously in the same room on two microphones, the sound signal from the two sources are mixed on both microphones. Coordinates of data points in Fig. 7 could represent the signal recorded from the two microphones. Separating the two sound sources from the microphone signal is called blind source separation. Independent component analysis (ICA) is a family of linear blind source separation methods. The core mathematical concept of ICA is to minimize the mutual information among the data projections. PCA components are orthogonal as shown in Fig. 7, which is usually not a realistic assumption for biophysical data. To find biologically plausible sources, PCA must be followed by an axis rotation procedure, and ICA can be viewed as a powerful rotation method. The ICA seeks to find axes for which the projection of data is maximally nonnormal (i.e., contains the maximum amount of information). It uses the property of the central limit theorem in statistics, stateing that any linear mixture of two or more source activities is more normal that the original source activities, so, by finding axes that maximize nonnormality, source separation may be achieved. As can be seen in Fig. 7, ICA is free to adapt to the actual projection patterns of source generators, if their activity time courses are (near) independent of one another. Performing ICA decompositions is most appropriate when sources are linearly mixed in the recorded signal, without differential time delays.

The ICA is being applied to various biomedical signal processing problems that include performing speech and noise separation (19), decomposing functional resonance imaging data (20), and separating brain area activities and artifacts mixed in electroencephalography scalp sensors (21).

Texts that give broad coverage of multivariate analysis are (22–24).

CLINICAL TRIALS

A clinical trial is not a method *per se*, but is a term applied to any form of planned experiments that involves human patients. The purpose of a clinical trial is to evaluate and verify the efficacy and safety of a new treatment or sets of treatments for a given medical condition. Although most of the analytical methods employed for clinical trials are the same as in other contexts, there is a special effort to avoid bias, which leads to some unique designs. Another distinguishing characteristic of clinical trials is the constraint imposed by studying living patients and the often difficult ethical considerations that must be addressed.

Double Blind. The usual method to avoid bias in experimental designs is the radom allocation of experimental subjects to treatments, but this will generally not suffice in clinical trials. A major potential source of bias is when subjects or evaluators in a trial know which treatment (e.g., placebo or active) is being received. In double-blind trials, neither the subject nor the evaluators are aware of which treatment is being received. Sometimes ethical or practical considerations make double-blinding infeasible, and sometimes partial blinding, for example, independent blinded evaluators only, may be sufficient to reduce bias in treatment comparison.

Within Patient Studies versus Across Patient Studies. Most clinical trials are conducted as parallel studies in which two or more treatments are evaluated concurrently in separate groups of patients. As many researchers remain reluctant to assign patients randomly to new or standard treatments, current patients on the new treatment may be compared with data external to the study containing patients who had received standard treatments. Such an approach invites severe bias, since there is no assurance that treatment and control groups do not differ with respect to some factors other than the treatment itself. In crossover studies, each patient receives in succession two or more treatments. When feasible, such within-patient studies require smaller sample sizes than between-patient studies to achieve the same level of significance.

Lifetime Variables. Some clinical studies are conducted as life data analysis and survival studies, and require specific statistical tools. In such studies, a variable represents the time to the occurrence of some event of interest, and is called a lifetime variable. In the engineering context, a life test consists of monitoring the operation of a sample of devices and to observe causes of and times to failure for all or some of the devices. In the clinical context, a survival study may involve observing cause of death (and time from entry to the study until death occurs) for some potentially fatal or, in the case of animal studies, induced disease. Alternatively, the event of interest may be time to relapse or time to remission for some diseases or conditions. The purpose of life tests or survival studies is to estimate or to compare lifetime or survival between different treatment groups.

Statistical Test for Lifetime Variables. Since a lifetime variable must be positive (number of remissions, e.g.), the normal distribution is not usually a suitable model. The normal-based methods of multiple regression and analysis of variance cannot be used in the usual manner and in general requisite mathematical and computational methods are much less tractable than normal-based methods. Consequently, a nonparametric, partially parametric, or nonnormal distributional analytic approach is taken. Data is usually visualized using Kaplan–Meier survival curves where censored patients (patients that have left the study) are explicitly indicated on the curve. Comparing between unpaired groups usually involve a log-rank test or a Mantel–Haenszel test. Conditional proportional hazards regression may be used to compare between two or more paired groups. Finally, Cox proportional hazard regression may be used to compare between more than two unpaired groups and perform regression analysis.

Censoring. As mentioned above, a further complicating factor for survival studies is censoring. Under censoring,

exact lifetimes are known only for a portion of the experimental units, the remainder known only to exceed certain censoring times. Censoring is usually a practical necessity and must be preplanned. For example, a life test on a random sample of 100 devices that has median time-to-failure of 2500 h will likely take over a year to complete if the tests were to continue until all devices fail. Instead, the test might be terminated at some predetermined time (e.g., 1000 h), or immediately upon achieving some predetermined number of failures (e.g., 30). These are called Type I and Type II censoring, respectively, and are the simplest to deal with. A distinguishing characteristic of survival studies involving human patients is that censoring times are often random. For example, suppose patients with a certain cancer are undergoing different chemotherapy treatments. Patients may enter the study in a random manner and patients may survive the termination time of the study or may die due to causes unrelated to the cancer. There are probability models that incorporate these data and lead to appropriate statistical inferential techniques. For example, some techniques assess the effectiveness of different treatments by comparing estimated mean survival times with the effect of unrelated causes of death removed. Other methods used for dealing with censoring will not be discussed. It is sufficient to say that the special problems of statistical inference in the presence of censoring necessitate the use of large sample approximations and computer- aided numerical solutions. Some of these methods incorporate strong assumptions that users should be aware of.

Extensive treatment of methods for censoring and the analysis of survival data is given in Refs. 25–27. Nontechnical discussions of clinical trials and the special statistical treatments they require are given by Pocock (28) and Shapiro (29).

STATISTICAL COMPUTING AND SOFTWARE

Standardized computer programs aiming at performing a variety of statistical analyses were developed through the 1960s at several universities and became widely available in the 1970s. There is now a large number of them and the one to use will depend on the users expertise in statistics and field of research. For infrequent usage on small data samples and testing of simple hypothesis (χ^2, t-test, ANOVA), MS Excel, which is usually already installed on many computer desktops, may be sufficient. Note the availability of extra statistical functions when one selects the Analysis Toolpack add-in (installed but inactive by default). However, MS Excel is not a statistical software *per se*, so to go beyond exploratory analysis stages it is better to rely on professional statistical software.

The best known and most comprehensive of these, now all under privately managed companies, are the Statistical Package for the Social Sciences (www.spss.com), the Statistical Analysis System (www.sas.com), and JMP (www.jmp.com). As its name suggests, SPSS, was developed primarily for use by social scientists and is relatively easy to learn by individuals with limited statistical and computer backgrounds. The SPSS graphical interface is organized as tabular spreadsheets similar to MS Excel. The programs comprising SPSS, their output format, and the examples in the manuals retain a social science flavor. The SAS has evolved into a widely utilized and extremely flexible package that is generally regarded to be more statistically sophisticated and complete than SPSS. JMP, also developed by the SAS institute, is a user-friendly graphical interface that sequentially guides the user through all stages of the experimental design and data analysis.

Apart from the graphical packages mentioned above, most other statistical softwares rely on command line calls, where users call functions from a prompt (note that most of these softwares also include menus). The free R software (www.r-project.org) offers powerful functions contributed by leading statisticians in the world. Because it is an open source project, it is used by many scientists and its extensive libraries are probably the place to look for rare statistical procedures. The Biomedical Programs (BMDP) (www.statsol.ie) contains a large variety of elementary and advanced statistical procedures. The programs are widely applicable, but some are particularly appropriate in biomedical contexts, such as repeated measures ANOVA designs (see ANOVA). The S-plus software is also very popular (www.insightful.com) and very similar to R. It is based on the S language developed at AT-T. Finally, a widely used package in academia, as well as in industry is a package called MINITAB (www.minitab.-com), which is one of the most user-friendly command line software.

There are many smaller, less comprehensive statistical analyses packages available for computers. These range from packages that perform elementary, mostly descriptive analyses, to some that are rather sophisticated. For bootstrap and surrogate statistics, SAS software is preferred among graphical software, although it is possible to program bootstrap and surrogate data routines in SPSS. The R software contains the majority of such user-contributed routines and S-Plus also contains a few of them. Finally, MATLAB (www.mathworks.com), an interpreted language widely used in engineering, also has a large number of user-contributed bootstrap and surrogate statistics routines available.

Caution against the ignorant use of computerized statistical analyses cannot be overemphasized. In planning studies, the methods of analysis and the constraint they impose on experimental designs should be taken into consideration in advance. If not, much work and data collection efforts could be wasted. Worse still, misleading and even meaningless results are often given undeserved weight merely because they represent the voluminous output of computer programs. How often do we hear that "a computer analysis shows. . ..", but such programs can be totally inappropriate. For example, the mathematical methods underlying repeated-measures ANOVA incorporate restrictive assumptions on the normality of the data and the experimental design for appropriate randomization of events. Although these considerations are often ignored, researchers should systematically assess the degree to which test-related assumptions are satisfied. These facts notwithstanding, computer-aided data

management and analysis can be of great benefit if used properly and wisely.

REFERENCES USED

This list is not meant to be comprehensive. For the naïve reader, a basic introduction to statistics with a plethora of exercises is given in the Schaum's outline series on statistics (2). For the nonnaïve reader in statistics, a more technical yet still accessible reference is Ref. 30. Other texts dealing with general statistical methods, particularly regression and analysis of variance are Refs. 31 and 32. Comprehensive web references are Refs. 18, 33, and 34.

Statistical books have also been written for specific research topics. For example, see Ref. 35 for a beginner's reference in designing biology experiments and Refs. 6, 8–10 for more detailed references. As already mentioned, see Refs. 28, 29, 36, and 37 for clinical trials. Finally, a recent development in statistics is statistical process control that deals with optimizing production and quality in the industry (38).

ACKNOWLEDGMENTS

Parts of this article were adapted from a previous version of this encyclopedia paper by P. Sullo and L.E. Ostrander from the Rensselaer Polytechnic Institute. The author also wishes to thank anonymous reviewer for their invaluable comments.

BIBLIOGRAPHY

1. Gill JL, Design and Analysis of Experiments in the Animal and Medical Sciences. Ames (IA): Iowa State University Press; 1978.
2. Spiegel MR, Stepens LJ, Schaum's Outlines in Statistics. 3 ed. New York: McGraw Hill; 1999.
3. Bonferroni CE. Sulle medie multiple di potenze, Bollettino dell'Unione Matematica Italiana, 5 third series, 1950. p 267–270.
4. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6, 65–70.
5. Hoppe F. Multiple Comparisons, Selection, and Applications in Biometry: A Festschrift in Honor of Charles W. Dunnett. New York: Marcel Dekker; 1992.
6. Cox DR. Planning of Experiments. New York: John Wiley & Sons; 1992.
7. Norman R, Draper NR, Smith H. Applied Regression Analysis. 2nd ed. New York: John Wiley & Sons; 1998.
8. Lorenzen TJ, Anderson VL. Design of Experiments: A No-Name Approach. New York: Dekker; 1993.
9. Box GEP, Hunter WG, Hunter JS. Statistics for Experimenters. New York: John Wiley & Sons; 1978.
10. Cochran WG, Cox GM. Experimental Designs. New York: John Wiley & Sons; 1992.
11. Hicks CR, Turner KV. Fundamental Concepts in the Design of Experiments. New York: Oxford University Press; 1999.
12. Winer BJ, Brown DR, Michels KM. Statistical Principles in Experimental Design. 3rd ed. New York: McGraw-Hill; 1991.
13. Neter J, Wasserman W, Applied Linear Statistical Models. 4th ed. New York: McGraw-Hill/Irwin; 1996.
14. Conover WJ. Practical Nonparametric Statistics Methods. 3rd ed. New York: John Wiley & Sons; 1998.
15. Hollander M, Wolfe DA. Nonparametric Statistical Methods. 3 ed. New York: John Wiley & Sons; 1999.
16. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman and Hall; 1994.
17. Blair R, Karniski W. An alternative method for significance testing of waveform difference potentials. Psychophysiology, 1993. p 518–524.
18. Howell DC. Resampling Statistics: Randomization and the Bootstrap. 2005. Availble at http://www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html.
19. Park H -M, Jung H -Y, Lee T -W, Lee S -Y. On subband-based blind signal separation for noisy speech recognition. Elect Lett 1999;35:2011–2012.
20. Duann JR, Jung TP, Makeig S, Sejnowski TJ. fMRLAB: An ICA Toolbox for fMRI Data Analysis. Presented at Human Brain Mapping, Sendai, Japan, 2002.
21. Delorme A, Makeig S. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 2004;134:9–21.
22. Anderson TW. An Introduction to Multivariate Statistical Analysis. New York: John Wiley & Sons; 2003.
23. Gnanadesikan R. Methods for Statistical Data Analysis of Multivariate Observations. New York: John Wiley & Sons; 1997.
24. Stone J. Independent Component Analysis : A Tutorial Introduction. Cambridge (MA): MIT Press; 2004.
25. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. 2nd ed. New York: John Wiley & Sons; 2002.
26. Lawless JF. Statistical Models and Methods for Lifetime Data. New York: John Wiley & Sons; 2002.
27. Miller R. Survival Analysis. New York: John Wiley & Sons; 1998.
28. Pocock SJ. Clinical Trials: A Practical Approach. New York: John Wiley & Sons; 1984.
29. Shapiro SH. Clinical Trials. New York: Dekker; 2004.
30. Hays W, Statistics. 5th ed. New York: Wadsworth Publishing; 1994.
31. Afifi A, Clark VA, May S. Computer-Aided Multivariate Analysis. 4th ed. New York: Chapman & Hall/CRC; 2004.
32. Snedecor GW, Cochran WG, Statistical Methods. 8 ed. Ames (IA): Iowa State University Press; 1989.
33. Lowry R, Concepts and Applications of Inferential Statistics. 1999. Available at http://faculty.vassar.edu/lowry/webtext.html.
34. Wasson J. Statistics in Educational Research—An Internet Based Course. Available at http://www.mnstate.edu/wasson/ed602.htm.
35. Cann A. Maths from Scratch for Biologists. John Wiley & Sons; 2002.
36. Tukey JW. Some thoughts on clinical trials, especially problems of multiplicity. Science 1977;198:679.
37. Chalmers TC, Celano P, Sacks HS, Jr. JS. Bias in treatment assignment in controlled clinical trials. N Engl J Med 1983;309:1358.
38. Amsden R, Butler H, Amsden D. SPC Simplified: Practical Steps to Quality. Quality Resources, 1998.

See also BIOINFORMATICS; COMPUTER-ASSISTED DETECTION AND DIAGNOSIS; QUALITY-OF-LIFE MEASURES, CLINICAL SIGNIFICANCE OF.

STANDARDS FOR MEDICAL DEVICES. See CODES AND REGULATIONS: MEDICAL DEVICES.

STEREOTACTIC RADIOSURGERY. See RADIOSURGERY, STEREOTACTIC.

STEREOTACTIC SURGERY

ANTONIO A.F. DE SALLES
ALESSANDRA GORGULHO
UCLA Medical School
Los Angeles, California

INTRODUCTION

Stereotactic surgery evolved from the need of neuroscientists and neurosurgeons to approach areas deep in the brain with minimal disruption of its structure. In (~1000 AD, the Incas already had some idea of which regions of the brain could be operated on without causing detectable functional deficits (1). They concentrated their trepanations in the right frontal area knowing that lesions in this region of the brain would be safe (2). The scientific literature, however, registers the first minimally invasive attempt to approach the noneloquent areas of the brain in the late 1800s. Zernov, a Russian anatomist, described the first device used to localize the sensory-motor areas of the brain (3,4) This device allowed the surgeon to use noneloquent areas as approaches to targets in the depth of the brain. He designed a frame that was attached to the patient's head and supported a hemisphere with a drawing of the brain gyri. This drawing guided the surgeon where to perform the craniotomy avoiding eloquent areas (4). This device was crude, however, and needed to be replaced by a precision instrument capable of being applied to patients' individual anatomy and not on a generic drawing.

Cartesian coordinates, developed by Renee Descartes (5), were called upon to guide the neurosurgeons in their endeavor. Initially developed for use in laboratory animals, the first stereotactic frame based on Cartesian spatial localization was designed by an electrophysiologist and a neurosurgeon. The Cartesian system provides precise localization of a point in space by the distance that the point is located from each of the three planes. Clarke and Horsley applied a stereotactic frame in animals to guide electrodes into the depth of the cerebellum to study neuronal function (6,7) (Fig. 1). This apparatus inspired the Canadian bioengineer Aubrey Mussen to develop the first true stereotactic device for humans (8,9). This device was never applied to a patient, since Mussen could not convince any neurosurgeon in Canada to use it. It is believed to have been created 1918, as it was found several years later wrapped in a newspaper of that year. It is currently in exhibit at the Montreal Neurological Institute (10).

All early localization of brain structures was based on cranial landmarks. Unfortunately, these bony reference points frequently failed to guide the surgeon to the precise area of the brain to be approached. Only when angiography (11) and ventriculography (12,13) became available, stereotactic surgery based on intracranial reference points became reliable. It took another 17 years for the first human stereotactic surgery based on Cartesian coordinates to be performed. Spiegel and Wycis performed the first stereotactic surgery in 1947, in Philadelphia (14). It was a thalamotomy to treat psychiatric disease (15).

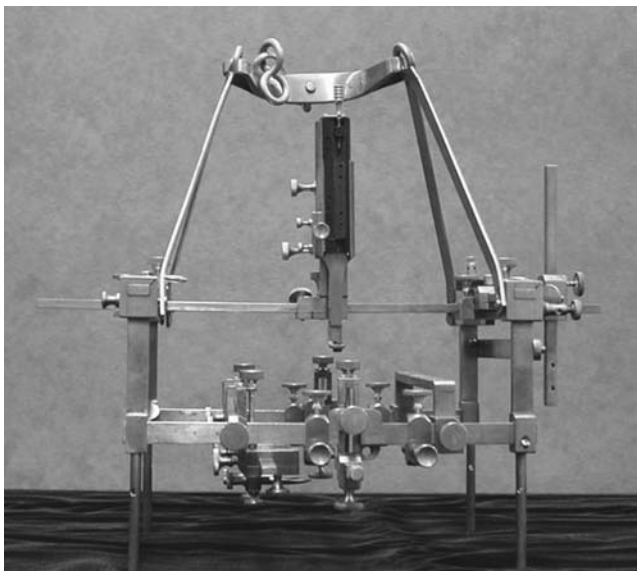


Figure 1. A. Horsley–Clark stereotactic device: Horsley, a neurosurgeon and Clark, a physiologist, developed the first stereotactic apparatus. It was used in experimental studies in animals.

Functional neurosurgery, mainly movement disorders, was the major field for stereotactic surgery at the beginning. During the 1930s and 1940s, multiple central nervous system procedures involving the motor cortex to the cervical spine were used in an attempt to releave parkinsonian symptoms. Despite tremor improvement, none of the procedures resulted in rigidity and akinesia amelioration, and usually resulted in paralysis of the affected limb. In 1940, Meyers published his patients series in whom either tremor and rigidity were improved (16–18). Meyers performed a craniotomy to section the pallidofugal fibers as they emerged the medial Globus Pallidus. In 1947, Spiegel and Wycis developed the stereotactic apparatus, therefore avoiding a craniotomy. It was with Irving Cooper, in 1952, that the surgical treatment of Parkinson's Disease gained popularity by the inadvertent lesion of the anterior choroid artery during a pyramidotomy (19). Cooper et al. provoked a pallido-thalamic infarction due to the lesion and observed an effective relief of the parkinsonian tremor. Thereafter, more precise targets were defined (16–24) for the treatment of Parkinson's disease (PD), essential tremor, dystonia and others. Initial results were encouraging, but recurrence and morbidity, specially in bilateral procedures, were not irrelevant. After the introduction of levodopa therapy in 1968, the indication of surgical procedures declined. Nevertheless, levodopa therapy proved not to be as effective after a medium period between 5 and 10 years (22,25). Over this period of time, tolerance to levodopa occurs and a dose increase is necessary to maintain the prior therapeutic effects. The extra amount of medication leads to a complication known as dyskinesia. The patient

presents involuntary movements that may be mild or severely disabling during the peak period of the dose. Since medication was highly effective only for a period of time, attention was directed once more toward surgical treatment. In the 1980s, stereotactic procedures for movement disorders again became routine (26,27).

The applications of the stereotactic techniques have grown since that time (28–30). New devices were developed to facilitate surgery and increase precision. Frames are still widely used in clinical practice; however, guidance techniques independent of skull fixation are improving and may completely replace the use of the stereotactic frame (31–35). This article discusses the evolution of the stereotactic instrumentation during the last century and describes the directions of the stereotactic technique at the beginning of the millennium.

PRINCIPLES OF STEREOTACTIC SURGERY

The stereotactic frame establishes the stereotactic space. It is described mathematically as a cube with *X*, *Y*, and *Z* coordinates corresponding to lateral, anteroposterior (AP) and vertical measurements, respectively. When the head is placed inside the stereotactic device, precise coordinates can be assigned to any location within the brain. Initially, orthogonal approaches were used, which were adequate because only orthogonal images were available, as represented by plain X rays. The AP projection offered the *X* and *Z* values and the lateral projection offered the *Y*, but also *Z* values (10).

The early stereotactic devices allowed only orthogonal approaches to the brain in relation to the applied stereotactic frame. This is a straightforward method in which the probe is perpendicular to a square base fixed in the skull. Other mathematical approaches were used over the years: the burr hole mounted system (Ward and McKinney), the interlocking arcs system (Brown–Robbers–Wells, BRW), the phantom-based system (Riechert–Mundinger), and the arc centered system (Leksell). In the burr hole mounted system, the depth of the probe is dependent on the angle of the burr hole mounted apparatus. A minor variance of angulation can lead to a major error in the deep as well as anteroposterior and lateral position of the probe. Therefore, this system is not used nowadays. The interlocking arcs system requires the adjustment of individual arcs to define the trajectory making calculations very complex. Despite that, the first computer-based computed tomography (CT) stereotactic was performed with the interlocking arcs (BRW) system. The arc centered system, the most currently used nowadays, was developed by Lars Leksell in 1949 (36). The trajectory of the probe is perpendicular to the arc (vertical axis) and the quadrant (horizontal axis). A spatial spherical shape is defined by the arc-quadrant. When the probe reaches maximum depth, it will always be in the center or at a focal point of that sphere, independent of the entry point. The arc centered became the instrument of choice since the end of the twenty century (37).

Historically, neurosurgeons have been dependent on neuroimaging to perform their craft. Before that, neurosurgery was dependent on symptomatic clues to approach the brain. Only *postmortem* correlation studies served as references for the surgeon to operate in the living brain. Ventriculography was the first great step in the development of stereotactic surgery (12,13). It provided the indispensable brain landmarks for the neurosurgeon to start electrophysiological brain mapping. Functional and anatomic atlases of the brain were developed to guide the stereotactic operations. To date, atlases developed during the 1950s and 1960s are still used in stereotactic surgery (38). Initially, the foramen of Monroe and the pineal gland were the landmarks used for internal guidance of the stereotactic surgeon. These landmarks were well seen on air ventriculography, which was the first contrast material medium used in neurosurgery to delineate the internal structures in a plain X-ray film. Soon after, the positive iodine contrast became available for neurosurgery. The positive contrast injected into the ventricular system provided exquisite delineation of the anatomy of the third ventricle. The anterior commissure (AC) and posterior commissure (PC) could be promptly defined. They became the landmarks of choice for stereotactic surgeons. The main atlases of the brain were developed based on these two landmarks (38,39). (Fig. 2).

The Cartesian planes were largely based on the initial intercomissure plane. This imaginary plane hinged on the AC and PC line being parallel to the skull base. Two other planes perpendicular to this plan and to each other composed the necessary three planes to determine the Cartesian system. The coronal plane passes through the midcommissural point and is perpendicular to the midsagittal plan. The mathematical challenge during stereotactic surgery consists of transforming the numbers generated

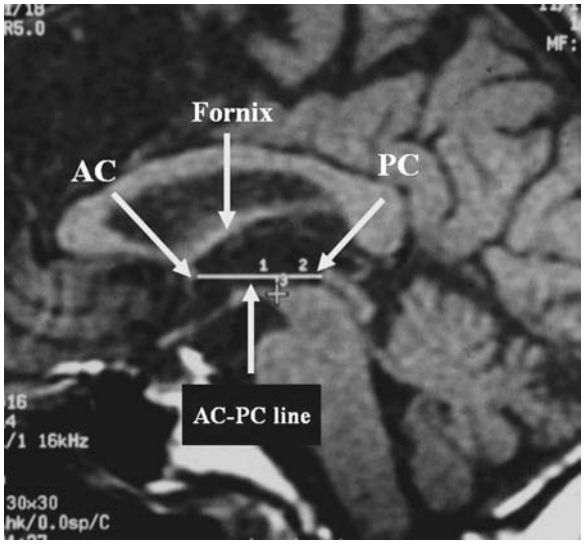


Figure 2. Sagital T1 weighted magnetic resonance imaging (MRI) scan showing the AC, which is defined at the anterior wall of the third ventricle. The PC is placed at the posterior wall of the third ventricle, directly above the beginning of the Aqueduct. The AC–PC line is of the utmost importance in functional neurosurgery. The coordinates established in the major stereotactic atlases were defined using this line as reference.

in the internal Cartesian system of the brain to one of the stereotactic devices. This simple mathematical transformation was done in the operating room by stereotactic surgeons, now this is automatically generated by computer software.

The indirect visualization of the brain by shadows of the ventricular system obtained with plain X rays of the skull was sufficient for functional localization of sites in the brain. These indirect visualizations and calculations allowed stereotactic surgeons to accumulate a wealth of knowledge of the brain's electrophysiology. Standardized atlases allowed a degree of stereotactic accuracy during functional procedures. Pathological anatomy secondary to space occupying lesions did not require the use of standardized atlases. This knowledge was limited to directing surgeons in the treatment of brain tumors and other morphologic diseases of the brain. Visualization of brain lesions while inside of the stereotactic frame was necessary for reaching these lesions for biopsy and possible therapy. This soon became possible with the angiography since indirect targeting based on deformation of the vascular anatomy made the localization reliable. However, due to the vascular nature of the lesions visualized, biopsy would be too risky. Stereotactic biopsy only flourished with the advent of CT and direct soft tissue visualization. During the angiography period, arteriovenous malformations (AVMs) were treated with radiation and tumors were treated with implants of isotopes guided by stereotactic surgery.

STEREOTAXIS BASED ON CT

Targeting based on tomographic images resulted in a remarkable increase of the use of stereotactic techniques in many centers. It also expanded its applicability to different fields. Focused radiotherapy became possible because CT allowed reliable targeting when compared to early angiography and pneumoencephalography–ventriculography generated targets.

The X and Y coordinates were acquired from one axial CT slice. Determining the Z coordinate by CT was a challenge. The initial idea was to correlate the vertical displacement of the CT table from a reference point to the target slice. This parameter was not reliable and a new form of calculating vertical displacement was elaborated. The fiducial markers were made possible by obtaining an accurate three-dimensional (3D) target (37). The fiducials (Fig. 3) are placed perpendicular to the image acquisition plane, either axial, sagittal, and coronal. The fiduciary system is composed of vertical outer bars and a diagonal internal one. According to the slice obtained, the distance between the diagonal line to the vertical one is going to be specific and the precise 3D localization of a point inside the stereotactic space was easily obtained. This same arrangement is applied for MRI image guidance.

STEREOTAXIS BASED ON MRI

Magnetic resonance imaging became a major diagnostic tool for many reasons: multiplanar capability, high spatial resolution, excellent soft tissue contrast, absence of ionizing radiation, and bony artifact on adjacent soft tissue (40). These characteristics made MRI attractive for framed procedures and also enabled the use of MRI in frameless guided procedures (Fig. 4).

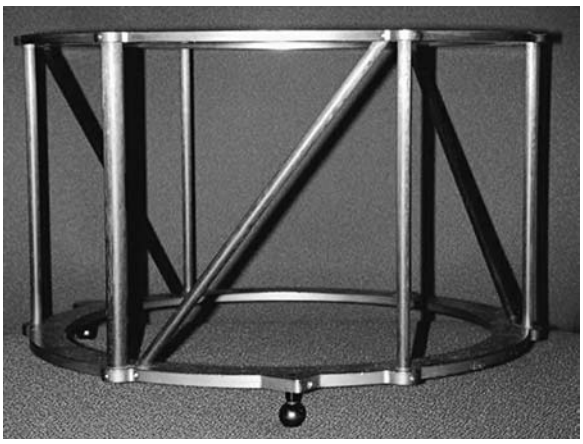


Figure 3. The CRW CT scan localizer (Radionics, Burlington, MA): The fiduciary system is composed of vertical outer bars and a diagonal internal one. Fiducial markers allow a precise 3D targeting.

However, there are some disadvantages. The MRI is more prone to geometric distortion than CT. The maximum localization error reported in the literature ranges from 1 to 5 mm (33,40). The variation in the maximum error may be accounted for partly by the use of different stereotactic systems and different reference imaging modalities to which MRI is compared. Actually, the geometrical accuracy of MRI is not significantly different from the accuracy of conventional stereotactic frames (41,42). The spatial accuracy of MRI depends on the linearity and calibration of magnetic field gradients and on magnetic susceptibility artifacts that manifest as spatial distortions. Meticulous

Figure 4. View of the UCLA interventional MRI (Sonata, Siemens, Erlangen, Germany) operating room.

quality control should decrease errors into magnetic field linearity and calibration. High bandwidth signal acquisition reduces spatial distortion due to susceptibility effects. Several researchers have proposed practical correction algorithms to further improve the spatial accuracy of MRI (40,43,44).

Functional and biopsy cases may be planned based only on the MRI scan since there are other means to confirm final target localization. However, targeting based only on the MRI is not accurate enough when performing radiosurgery (SRS) or stereotactic radiotherapy (SRT) (45). In SRS–SRT cases, only the fusion of the CT and MRI offers the reliability required for the delivery of focused radiation to the target selected.

Several stereotactic systems with frames designed for MRI localization are commercially available: the BRW and CRW (Radionics, Burlington, MA), Leksell (Elekta Instruments Inc, Atlanta, GA), Laitinen (Sandstrom Trade & Technology Inc, Ontario, Canada), and others (10).

STEREOTACTIC FRAME-BASED PROCEDURES

Functional Stereotaxis

Functional procedures are performed under MRI guidance at UCLA. Under local anesthesia and sedation, MRI-compatible stereotactic frame (Leksell, Elekta Instruments Inc, Atlanta, GA) is attached to the patient's head. The stereotactic frame is aligned to the MRI coil. The T1 weighted images 3 mm in thickness are obtained in the axial, coronal, and sagital planes. The anterior and posterior commissures are identified to define the AC–PC line. The coordinates are obtained from the Shaltenbrandt and Wahren atlas. The *X*, *Y*, and *Z* coordinates can be reproduced in the axial, sagital, and coronal MRI views by computer manipulations. The same target can be pinpointed on each of these images. The target determination is therefore better than the thickness of the imaging acquisition. Distortions of the MRI scan are not taken into account since the final position of the electrode is checked with microelectrode recording (MER) and macroelectrophysiology.

The patient is taken to the operating room either for lesion or deep brain electrode implant. The localizing microelectrode is introduced through a burr hole placed in the frontal region. Mapping with MER is obtained and the trajectory is adjusted if necessary. Parameters of electrical stimulation are manipulated aiming to improve symptoms until the point side effects are detectable. The final location of the definitive electrode or lesion site is therefore established (Table 1). The patient undergoes an intraoperative MRI to rule out bleeding and ascertain proper position of the target. Fusion of preoperative with intraoperative imaging is used to compare the accuracy of the surgery (10). This same approach is being used for cell transplantation, growth factors injection, and gene therapy (Table 2).

Table 1. Functional Stereotactic Procedures

Procedure	Indication
Behavior	
Amygdalotomy	Violence, aggressiveness
Anterior capsulotomy	Obsessive compulsive disorder (OCD)
Cingulotomy	Anxiety, depression, OCD
Posteriormedial Hypothalamotomy	Aggressiveness
Subcaudate tractotomy	Anxiety with depression
Pain	
Cingulotomy	Chronic pain emotionally charged
Dorsomedian thalamotomy	Chronic pain emotionally charged
Pulvinotomy	Intractable pain
Mesencephalotomy	Intractable pain
Periaqueductal stimulation	Intractable pain
Periventricular stimulation	Intractable pain
Cortical stimulation	Intractable pain
Movement Disorders	
Ventrolateral thalamotomy	Parkinson's disease, tremor, dystonia
Pallidotomy	Parkinson's disease
Campotomy (Forel's fields)	Parkinson's disease, athetosis, myoclonus
Zona incerta	Parkinson's disease, tremor, torticollis
Dentatotomy	Spasticity
Striatum fetal tissue transplant	Parkinson's disease
Epilepsy	
Amygdalofornicotomy/anterior Commissurotomy	Temporal lobe seizures
Pallidoamygdalotomy or Centromedian lesion	Salaam seizures
Deep electrode	Seizure focus determination

Table 2

Morphologic	Functional
Diagnostic	*Diagnostic*
Stereotactic biopsy	Deep electrode for seizure focus determination
Therapeutic	*Therapeutic*
Radiosurgery	DBS[1] for pain control, movement and behavior disorders
Frameless neuronavigation	Ablative lesion for pain, movement and behavioral disorders
Hyperthermia	Tissue transplantation for movement disorders
Stereotactic craniotomy	Gene therapy delivery
Brachytherapy	Injection of active chemicals (growth factor)

[a] Deep brain stimulator = DBS.

Stereotatic Biopsy

The determination of the biopsy target follows the same targeting procedure described above, however, electrophysiology is not used for confirmation. Instead, the histology obtained with a frozen section confirms the adequacy of the target.

Once the patient has the burr hole placed, the needle biopsy is introduced through the driver attached to the frame arc. The planning is established based on the three plan views: axial, coronal, and sagital. Samples of tissue are collected and frozen pathology is performed. After obtaining histological diagnosis, the operative wound is closed and the patient is submitted to a postoperative MRI. Usually an air bubble is observed at the site of the biopsy, confirming the target (28). Pre- and postoperative images are fused. Minimally invasive approaches to vital structures, such as brainstem, became possible only after the development of high definition imaging methods (Fig. 5).

Radiosurgery

Radiosurgery is likely the most frequent reason for stereotactic frame application nowadays. Under local anesthesia, a CRW (Radionics, Burlington, MA) or SRS frame (BrainLab, Heimstetten, Germany) is attached to the patient's head. The patient is submitted to a CT scan and the CT-framed image is fused to the preoperative MRI. The planning starts by the drawing of the lesion, that is, in reality, the target determination. After conclusion of the planning, the patient is attached to a Novalis (BrainLab, Heimstetten, Germany) couch. Focused radiation is delivered with high precision achieved with frame-based spatial localization. At the end of the treatment, the frame is removed. In SRT cases, since the placement of a frame for 26–30 times in a patient is impracticable, a facial mask is manufactured. The frame and the fiducials are applied to the mask. The patient undergoes CT, which is fused with the MRI previously obtained. The reproducibility of accuracy with the facial mask, on a daily basis, has already been demonstrated to be ~ 2 mm (Solberg, personal communication). New radiosurgery devices, as Novalis and Cyberknife, allow the use of this frameless technique for radiosurgery of extracranial targets.

Stereotactic Craniotomy

This technique was popularized in the 1980s (46). Precise placement of the craniotomy based on computerized presurgical planning, guided retractors, and the use of microscopy were the major advancements in this area. Kelly et al. developed a very elaborate stereotactic system with integration of several imaging techniques. The purpose of this system, named COMPASS, was to guide the craniotomy and removal of deep-seated tumors under image and microscope guidance. Inspired by this idea, other similar equipments were developed. However, all these techniques lack the flexibility that neurosurgery requires. The frame frequently interferes with the craniotomy site and the limits defined for the boundaries of the lesion are not the same once the dura-mater is opened. Nowadays, neuronavigation has replaced stereotactic craniotomy.

Frameless Stereotactic Technique

The advent of image-guided neurosurgery and frameless stereotactic localization, also known as Neuronavigation, has advanced a new concept of stereotaxis (47–49). The development of multiple systems that utilize unique imaging and guidance technologies has established neuronavigation as a commonly used tool in many areas of neurosurgery, such as microsurgery for tumors (19,50–54), vascular lesions (51,52,55,56), biopsies (51,52,58–61) and epilepsy surgery (51,62,63). Moreover, modern neuronavigation techniques have begun to replace traditional image-guided tools, such as fluoroscopy and X rays (64–68). The success of neuronavigation hinges on the practical nature of this technique. It completely replaced the frame-based

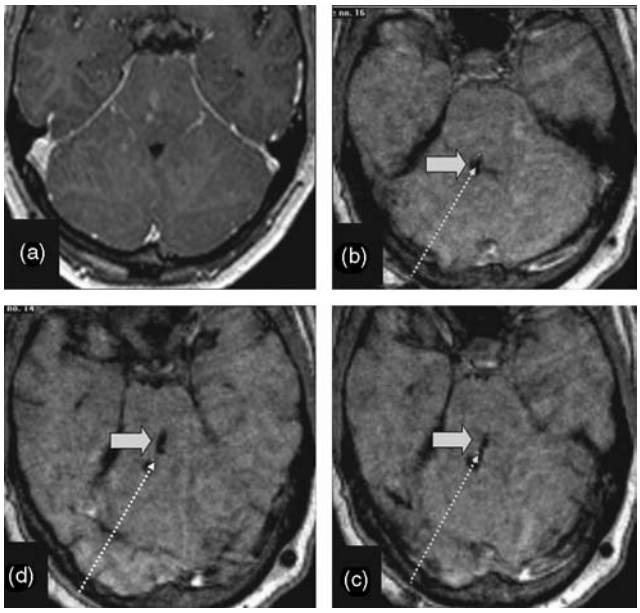


Figure 5. (A) Preoperative MRI scan showing a lesion in the brainstem, anterior to the fourth ventricle (B, C, and D). The pre- and postoperative images were fused. The site of the biopsy (arrow) and the planned trajectory of the needle (dotted arrow) can be noticed.

stereotactic craniotomy approach, which was not well accepted by the general neurosurgeon.

New approaches to lesions situated in critical areas have been developed. For example, frameless guided biopsy of lesions inside the cavernous sinus through the *foramen ovale* is now possible (28). Not long ago, procedures like this would require a craniotomy to be performed. This demonstrates how neurosurgery has become sophisticated to the point of reaching multiple brain areas through natural pathways that could not be accessed before due to lack of precise guidance.

Neuronavigation has set the stage for the application of multiple and complementary imaging techniques to aid the surgical approach. For example, the fusion of magnetoencephalography (MEG), positron emission tomography (PET), and functional MRI (*f*MRI) has been reported in the literature (31,69,70). Additionally, new MRI modalities, such as diffusion tensor imaging (DTI), also fused to other functional images, have made possible the recognition of important white matter structures, further increasing the safety of surgery in the depth of the brain. The orientation of the eloquent tracts in the white matter can be mapped using the fact that only diffusion anisotropy along the principal direction of the magnetic field gradient is visible (31,70).

The frameless neuronavigation system is composed of four elements: (1) registration of the image to the real anatomy of the patient, (2) interactive localization device (ILD), (3) computer interface, (4) integration of virtual and real-time data (71). Fiducial markers are the most reliable method of registration. There are two types of fiducials: mobile and rigid. Soft tissue fixation (mobile fiducials) are less accurate. Rigid fiducial markers require a minor surgical procedure for placement. On the other hand, the improvement in precision is worthy. A physical pointer can be used to identify points or surfaces to be registered with ILD.

Two types of ILD can be used: linked or nonlinked. These terms refer to the presence of an unbroken physical connection between the patient and the device. An example of the linked type is the robotic arm, while one example of unlinked is triangulation, which can be sound or infrared (IR) light based. Although of historical importance, the localizing articulating arm has been replaced by the triangulation techniques, especially by the IR reflecting devices (28,52,72–74). Most recently, radio frequency (rf) emitters and receivers were an option to replace IR triangulation devices to eliminate line of sight problems in the operating room. Localization techniques using a magnetic field created to encompass the surgical area continue in development. Its advantage is localization independent of direct linear relation of the light source and probe. It allows the utilization of curvilinear probes or even catheters.

A new issue with the real-time data acquisition is the consistency of the brain. Current registration techniques assume the brain to be a solid and nondeforming structure. However, the brain is more dynamic and moves according to its environment. Preoperative image guidance, which assumes solid brain consistency, is not necessarily accurate at all times. Brain moves after dura opening due to release of cerebral spinal fluid (CSF). The magnitude of brain shift is variable depending on age, brain atrophy, and positioning for the procedure (50,75). This dynamic brain shifts can be overcome by real-time imaging during surgery.

Interventional MRI

New open designs of MR scanners with in-room image monitors allow MRI-guided interventional procedures. (Fig. 4) The first step toward practical interventional MRI of the brain is the development of instruments that function satisfactorily and safely in the strong magnetic field of clinical MRI scanners. Some concerns derive from this assertive.

Traditionally, surgical instruments are made of stainless steel material. Susceptibility artifacts usually develop at the extremities of the instrument. At the moment, considerable effort has been converged to test different biomaterials in the MRI magnetic field. The goal is to define the most suitable material: minimal artifact and high resistance. Less artifact is produced by ceramics, zirconium seems to produce the smallest artifact among the ceramics (76). On the other hand, ceramics present lower retention force and lower flexibility when compared to metallic materials. Susceptibility artifacts are also dependent on the pulse sequence applied. Spin echo sequences are less sensitive to time-independent local magnetic field variations while gradient spin–echo sequences are more susceptible to time-dependent and time-independent local field changes.

The strength of the magnetic field also interferes with the degree of distortion. It is possible to minimize significantly the amount of distortion using a short echo–time, thinner slices, small field of vision (FOV) and higher readout gradients. Any material used in manufacturing the instruments will produce some amount of artifacts.

The actual location of the instrument, for example, an electrode inside the artifact, is another important issue. This becomes important to evaluate the final placement of the electrode according the planned target comparatively. It is known that artifacts develop parallel to the direction of the main magnetic field applied. The precise position of the electrode inside the distorted image was reported to be in the middle of the artifact (77).

Another concern is related to electromagnetic interference between the MRI scanner and interventional electronics. For example, the rf generator may emit electromagnetic radiation that interferes with image acquisition and generates artifacts. At UCLA, the use of MER in the MRI is routine, either the 0.2 (70) or the 1.5 T (Siemens, Erlangen, Germany), with special attention to work in the fringes of the strong magnetic field. Interferences can be safely avoided in so far as attention is centered in the strategic location of the devices inside the MRI room.

To approximate real-time imaging, either image acquisition or image reconstruction has to be of very high speed. Currently, millimetric resolution in the 0.2 T and submillimetric in the 1.5 T can be achieved with the parameters reported at Table 3.

Interventional MRI scanners offer the possibility of image acquisition in advance for planning the procedure, intraoperatively for compensation of brain shift that occurs

Table 3. Imaging Acquisition Parameters[a]

Magnetic Field Strength	Low Field, 0.2 T		High Field, 1.5 T		
Pulse sequences	T2W FL2D	3 DF	T1W MPRAGE	T2W TSE	FL 2D
Region	WB	WB	WB	BG	WB
TR, ms	1500	56	2050	2800	800
TS, ms	60	25	4.4	84	15
FOV, mm	230	280	280	280	280
Matrix	224 × 256	256 × 256	256 × 256	256 × 256	168 × 256
Voxel Size, mm	1.0 × 0.9 × 6.0	1.1 × 1.1 × 3.0	1.1 × 1.1 × 1.0	1.1 × 1.1 × 2.0	1.5 × 1.1 × 7.0
Scan Time, min:s	11 : 16	7 : 42	17 : 31	2 : 47	4 : 28

[a]T1W = T1 weighed. TSE = turbo spin echo. T2W = T2 weighed. WB = whole brain. FL 2D = flair 2D. BG: basal ganglia. 3 DF = 3D flash. FOV = field of view. MPRAGE = magnetization prepared rapid acquisition gradient echo.

after dura opening (50,78) and postoperatively for early detection of complications and confirmation of target localization. The reliability of intraoperative image acquisition is superior to the method based only on the preoperative data setting for targeting.

APPLICATIONS

Stereotactic surgery is widely applied for morphological and functional procedures in the brain. Functional and morphologic stereotaxis can be subdivided into diagnostic and therapeutic procedures. A brief summary of these applications is presented in Tables 1 and 2.

FUTURE DIRECTIONS

Advances in imaging techniques are characterized by faster acquisition sequences, more precise definition of eloquent structures, and minimization of artifacts. It is feasible to predict a more sophisticated integration among imaging, virtual reality computer, and robotics in a near future. Possibly new surgical tools as thermal ablation, cryoablation, and chemoablation will become routine, not only for the brain, but for the whole body. The MRI would also be able to provide sensitive monitoring of temperature changes and tissue injury with high temporal resolution. This would be an advantage once minimal injury inside or surrounding an eloquent area may lead to a major deficit (79), obviating the need of large surgical access to remove tumors.

The increasing amount of publications (57,80–86) discussing preliminary results using interventional MRI technique, exploring types of biomaterials (76,87–92), and new sequences of image acquisition (31,70) clearly show that interventional MRI is in its early stages of development. Whether frameless procedures under intraoperative MRI guidance will replace frame stereotaxis in the future is still an open question. Studies already claim that the same accuracy obtained with framed technique is achieved (33,40,41,87).

Frameless stereotactic concepts of registration of images to patients' anatomy will continue to be the basis of surgical guidance, even with real-time imaging. Technological advances as interventional MRI carry issues like high costs. Nevertheless, due to its minimally invasive nature compared with other open surgical approaches, image guidance married to interventional MRI may emerge as a means of actually lowering the overall cost of medical care.

BIBLIOGRAPHY

1. Bakay RA et al. Delayed stereotactic transplantation technique in non-human primates. Prog Brain Res 1988;78:463–471.
2. Marino RJ, Gonzales-PortilloM. Preconquest Peruvian neurosurgeons: a study of Inca and pre-Columbian trephination and the art of medicine in ancient Peru. Neurosurgery 2000;47:940–950.
3. DN Z. L'encéphalomètre. Rev Gen Clin Ther 1980;19:302.
4. Lichterman BL. Roots and routes of Russian neurosurgery (from surgical neurology towards neurological surgery). J Hist Neurosci 1998;7:125–135.
5. RD. Discours de la Methode. Paris: Vrin; 1992.
6. Clarke RH, Horsley V. On a method of investigating the deep ganglia and tracts of the central nervous system (cerebellum). Br Med J 1906;1799–1802:1799–1800.
7. Horsley VC R. The structure and functions of the cerebellum examined by a new method. Brain 1908;31:45–124.
8. Olivier A, Bertrand G, Picard C. Discovery of the first human stereotactic instrument. Appl Neurophysiol 1983;46:84–91.
9. Picard C, Olivier A, Bertrand G. The first human stereotaxic apparatus. The contribution of Aubrey Mussen to the field of stereotaxis. J Neurosurg 1983;59:673–676.
10. De Salles AA. Stereotactic applications. In: De Salles AA, Goetsch S. editors. Stereotactic Surgery and Radiosurgery. Madison (WI): Medical Physics Publishing; 1993.
11. Moniz E. L'encéphalographie arterielle, son importance dans la localisation des tumeurs cerebrales. Rev Neurol 1927;2:72–90.
12. Dandy WE. Localization of brain tumors by cerebral pneumography.Am J Roentgenol 1923;10:610–612.
13. Dandy WE. Ventriculography following the injection of air into the cerebral ventricles. Ann Surg 1918;68:5–11.
14. Spiegel EA WH, Marks M, Lee ASJ. Stereotactic apparatus for operations of the human brain. Science 1947;106:349–350.
15. Spiegel EA WH, Baird HW. Studies in stereoencephalotomy.I. Topical relationships of subcortical structures to the posterior comissure. Confin Neurolog 1952;12:9–133.
16. Benabid AL, et al. Chronic electrical stimulation of the ventralis intermedius nucleus of the thalamus as a treatment of movement disorders. J Neurosurg 1996;84:203–214.
17. R. M. Surgical interruption of the pallidofugal fibers: its effect on the syndrome of paralysis agitans and technical considerations in its application. NY State J Med 1942;42:317–325.

18. Vitek JL, et al. Microelectrode-guided pallidotomy: technical approach and its application in medically intractable Parkinson's disease. J Neurosurg 1998;88:1027–1043.
19. Kelly PJ. Computer-assisted stereotaxis: new approaches for the management of intracranial intra-axial tumors. Neurology 1986;36:535–541.
20. Hirai T, Miyazaki M, Nakajima H, Shibazaki T, Ohye C. The correlation between tremor characteristics and the predicted volume of effective lesions in stereotaxic nucleus ventralis intermedius thalamotomy. Brain 1983;106 (Pt. 4):1001–1018.
21. Laitinen LV. Pallidotomy for Parkinson's disease. Neurosurg Clin N Am 1995;6:105–112.
22. Matsumoto K, Shichijo F, Fukami T. Long-term follow-up review of cases of Parkinson's disease after unilateral or bilateral thalamotomy. J Neurosurg 1984;60:1033–1044.
23. Nagaseki Y, et al. Long-term follow-up results of selective VIM-thalamotomy. J Neurosurg 1986;65:296–302.
24. Pollak P, et al. Long-term effects of chronic stimulation of the ventral intermediate thalamic nucleus in different types of tremor. Adv Neurol 1993;60:408–413.
25. Burchiel KJ. Thalamotomy for movement disorders. Neurosurg Clin N Am 1995;6:55–71.
26. Lozano A, et al. Methods for microelectrode-guided posteroventral pallidotomy. J Neurosurg 1996;84:194–202.
27. Tronnier VM, Fogel W, Kronenbuerger M, Steinvorth S. Pallidal stimulation: An alternative to pallidotomy? J Neurosurg 1997;87:700–705.
28. Frighetto L, et al. Image-guided frameless stereotactic biopsy sampling of parasellar lesions. Technical note. J Neurosurg 2003;98:920–925.
29. Grunert P, et al. Frame-based and frameless stereotaxy in the localization of cavernous angiomas. Neurosurg Rev 2003;26:53–61.
30. Walker DG, Ohaegbulam C, Black PM. Frameless stereotaxy as an alternative to fluoroscopy for transsphenoidal surgery: Use of the InstaTrak-3000 and a novel headset. Clin Neurosci 2002;9:294–297.
31. Kamada K, et al. Visualization of the eloquent motor system by integration of MEG, functional, and anisotropic diffusion-weighted MRI in functional neuronavigation. Surg Neurol 2003;59:352–361. discussion 361–352.
32. Roessler K, et al. Frameless stereotactic lesion contour-guided surgery using a computer-navigated microscope. Surg Neurol 1998;49:282–288. discussion 288–289.
33. Samset E, Hirschberg H. Image-guided stereotaxy in the interventional MRI. Minim Invasive Neurosurg 2003;46:5–10.
34. Samset E, Hirschberg H. Neuronavigation in intraoperative MI. Comput Aided Surg 1999;4:200–207.
35. Vitaz TW, Hushek SG, Shields CB, Moriarty TM. Interventional MRI-guided frameless stereotaxy in pediatric patients. Stereotact Funct Neurosurg 2002;79:182–190.
36. L. L. A stereotaxic apparatus for intracerebral surgery. Acta Chir Scand 1949;99:229–233.
37. PL G. Principles of stereotaxis and instruments In:. SJ DSAG editors. Stereotactic Surgery and Radiosurgery. Madison, (WI): Medical Physics Publishing, 1993.
38. Talairach JP. Co-Planar Stereotaxic Atlas of the Human Brain. New York: Thieme; 1988.
39. Shaltenbrand GWW. Atlas of Stereotaxy for the Human Brain. Chicago: Thieme, 1977.
40. De Salles AA, Gronemeyer D, Seibel R, Lufkin R. Instrumentation for Interventional MRI of the Brain. In: De Salles A, Lufkin R., editors. Minimally Invasive Therapy of the Brain. New York: Thieme; 1997.
41. Carter DA, Parsai EI, Ayyangar KM. Accuracy of magnetic resonance imaging stereotactic coordinates with the cosman-roberts-wells frame. Stereotact Funct Neurosurg 1999;72: 35–46.
42. Galloway RL Jr, Maciunas RJ, Latimer JW. The accuracies of four stereotactic frame systems: An independent assessment. Biomed Instrum Technol 1991;25:457–460.
43. Bakker CJ, Moerland MA, Bhagwandien R, BeersmaR. Analysis of machine-dependent and object-induced geometric distortion in 2DFT MR imaging. Magn Reson Imaging 1992;10: 597–608.
44. Fitzpatrick J, et al. A technique for improving accuracy in positron and intensity within images acquired in the presence of field inhomogeneity. New York: SMRM; 1990.
45. Solberg TD MP, DeMarco J, De Salles AAF, Selch MT. Technical aspects of LINAC radiosurgery for functional disorders. J Radiosurgery 1998;1:115–127.
46. Kelly PJ, Alker GJ Jr, Kall BA, Goerss S. Method of computed tomography-based stereotactic biopsy with arteriographic control. Neurosurgery 1984;14:172–177.
47. Kelly PJ, Alker GJ Jr, Goerss S. Computer-assisted stereotactic microsurgery for the treatment of intracranial neoplasms. Neurosurgery 1982;10:324–331.
48. Roberts DW, et al. A frameless stereotaxic integration of computerized tomographic imaging and the operating microscope. J Neurosurg 1986;65:545–549.
49. Watanabe E, et al. Three-dimensional digitizer (neuronavigator): new equipment for computed tomography-guided stereotaxic surgery. Surg Neurol 1987;27:543–547.
50. Barnett GH, Kormos DW, Steiner CP, WeisenbergerJ. Use of a frameless, armless stereotactic wand for brain tumor localization with two-dimensional and three-dimensional neuroimaging. Neurosurgery 1993;33:674–678.
51. Golfinos JG, Fitzpatrick BC, Smith LR, Spetzler RF. Clinical use of a frameless stereotactic arm: Results of 325 cases. J Neurosurg 1995;83:197–205.
52. Gumprecht HK, Widenka DC, Lumenta CB. BrainLab VectorVision Neuronavigation system: Technology and clinical experiences in 131 cases. Neurosurgery 1999;44:97–104. discussion 104–105.
53. Reinhardt HF, Zweifel HJ. Interactive sonar-operated device for stereotactic and open surgery. Stereotact Funct Neurosurg 1990; 54–55. 393–397.
54. Suess O, et al. Intracranial image-guided neurosurgery: Experience with a new electromagnetic navigation system. Acta Neurochir (Wien) 2001;143:927–934.
55. Reinhardt HF, Horstmann GA, Gratzl O. [Microsurgical removal of deep vascular malformations using sonar- stereometry]. Ultraschall Med 1991;12:80–83.
56. Zamorano L, et al. Interactive image-guided surgical resection of intracranial arteriovenous malformations. Comput Aided Surg 1998;3:57–63.
57. Barnett GH, Miller DW, WeisenbergerJ. Frameless stereotaxy with scalp-applied fiducial markers for brain biopsy procedures: experience in 218 cases. J Neurosurg 1999;91:569–576.
58. Germano IM, Queenan JV. Clinical experience with intracranial brain needle biopsy using frameless surgical navigation. Comput Aided Surg 1998;3:33–39.
59. Grunert P, et al. Stereotactic biopsies guided by an optical navigation system: Technique and clinical experience. Minim Invasive Neurosurg 2002;45:11–15.
60. Sawin PD, Hitchon PW, Follett KA, Torner JC. Computed imaging-assisted stereotactic brain biopsy: A risk analysis of 225 consecutive cases. Surg Neurolog 1998;49:640–649.
61. Ulm AJ, Bova FJ, Friedman WA. Stereotactic biopsy aided by a computer graphics workstation: Experience with 200 consecutive cases. Surg Neurol 2001;56:366–371. discussion 371–362.

62. Olivier A, Germano IM, Cukiert A, Peters T. Frameless stereotaxy for surgery of the epilepsies: Preliminary experience. Technical note. J Neurosurg 1994;81:629–633.
63. Wurm G, et al. Advanced surgical approach for selective amygdalohippocampectomy through neuronavigation. Neurosurgery 2000;46:1377–1382. discussion 1382–1373.
64. Choi WW, Green BA, Levi AD. Computer-assisted fluoroscopic targeting system for pedicle screw insertion. Neurosurgery 2000;47:872–878.
65. Dresel SH, et al. Meckel cave lesions: Percutaneous fine-needle-aspiration biopsy cytology. Radiology 1991;179:579–582.
66. Foley KT, Simon DA, Rampersaud YR. Virtual fluoroscopy: computer-assisted fluoroscopic navigation. Spine 2001; 26:347–351.
67. Sheporaitis LA, et al. Intracranial meningioma. AJNR Am J Neuroradiolog 1992;13:29–37.
68. Welch WC, Subach BR, Pollack IF, Jacobs GB. Frameless stereotactic guidance for surgery of the upper cervical spine. Neurosurgery 1997;40:958–963. discussion 963–954.
69. Reithmeier T, et al. Neuronavigation combined with electrophysiological monitoring for surgery of lesions in eloquent brain areas in 42 cases: A retrospective comparison of the neurological outcome and the quality of resection with a control group with similar lesions. Minim Invasive Neurosurg 2003;46:65–71.
70. Tummala RP, Chu RM, Liu H, T T, Hall WA. Application of diffusion tensor imaging to magnetic-resonance-guided brain tumor resection. Pediatr Neurosurg 2003;39:39–43.
71. Maciunas RJ. Approaches to Frame-Based and Frameless Stereotaxis. New York: Thieme; 1997.
72. Bohinski RJ, et al. Glioma resection in a shared-resource magnetic resonance operating room after optimal image-guided frameless stereotactic resection. Neurosurgery 2001; 48:731–742. discussion 742–734.
73. Germano IM, Villalobos H, Silvers A, Post KD. Clinical use of the optical digitizer for intracranial neuronavigation. Neurosurgery 1999;45:261–269. discussion 269–270.
74. Schroeder HW, Wagner W, Tschiltschke W, Gaab MR. Frameless neuronavigation in intracranial endoscopic neurosurgery. J Neurosurg 2001;94:72–79.
75. Nabavi A et al. [Neuronavigation. Computer-assisted surgery in neurosurgery]. Radiologe 1995;35:573–577.
76. Matsuura H, et al. Quantification of susceptibility artifacts produced on high-field magnetic resonance images by various biomaterials used for neurosurgical implants Technical note. J Neurosurg 2002;97:1472–1475.
77. Yelnik J, et al. Localization of stimulating electrodes in patients with Parkinson disease by using a three-dimensional atlas-magnetic resonance imaging coregistration method. J Neurosurg 2003;99:89–99.
78. Nabavi A, et al. Serial intraoperative magnetic resonance imaging of brain shift. Neurosurgery 2001;48:787–797. discussion 797–788.
79. Farahani K, et al. Effect of field strength on susceptibility artifacts in magnetic resonance imaging. Comput Med Imaging Graph 1990;14:409–413.
80. Bernays RL, et al. Histological yield, complications, and technological considerations in 114 consecutive frameless stereotactic biopsy procedures aided by open intraoperative magnetic resonance imaging. J Neurosurg 2002;97:354–362.
81. Black PM, et al. Development and implementation of intraoperative magnetic resonance imaging and its neurosurgical applications. Neurosurgery 1997;41:831–842. discussion 842–835.
82. Hall WA, et al. Brain biopsy using high-field strength interventional magnetic resonance imaging. Neurosurgery 1999;44:807–813. discussion 813–804.
83. Bradford R, Thomas DG, Bydder GM. MRI-directed stereotactic biopsy of cerebral lesions. Acta Neurochir. (Suppl). (Wien) 1987;39:25–27.
84. Dorward NL, Paleologos TS, Alberti O, Thomas DG. The advantages of frameless stereotactic biopsy over frame-based biopsy. Br J Neurosurg 2002;16:110–118.
85. Fahlbusch R, Ganslandt O, Nimsky C. Intraoperative imaging with open magnetic resonance imaging and neuronavigation. Childs Nerv Syst 2000;16:829–831.
86. Moriarty TM, et al. Frameless stereotactic neurosurgery using intraoperative magnetic resonance imaging: stereotactic brain biopsy. Neurosurgery 2000;47:1138–1145. discussion 1145–1136.
87. Koyama T, Handa J. Porous hydroxyapatite ceramics for use in neurosurgical practice. Surg Neurol 1986;25:71–73.
88. Nitatori T, Hanaoka H, Hachiya J, Yokoyama K. MRI artifacts of metallic stents derived from imaging sequencing and the ferromagnetic nature of materials. Radiat Med 1999;17:329–334.
89. Port JD, Pomper MG. Quantification and minimization of magnetic susceptibility artifacts on GRE images. J Comput Assist Tomogr 2000;24:958–964.
90. Shellock FG, Shellock VJ. Ceramic surgical instruments: Ex vivo evaluation of compatibility with MR imaging at 1.5 T. J Magn Reson Imaging 1996;6:954–956.
91. Shellock FG, Shellock VJ. Spetzler titanium aneurysm clips: compatibility at MR imaging. Radiology 1998;206: 838–841.
92. Tominaga T, et al. Magnetic resonance imaging of titanium anterior cervical spine plating systems. Neurosurgery 1995; 36:951–955.
93. Dorward NL, et al. Accuracy of true frameless stereotaxy: In vivo measurement and laboratory phantom studies. Technical note. J Neurosurg 1999;90:160–168.

See also HYPERTHERMIA, INTERSTITIAL; RADIOSURGERY, STEREOTACTIC; TISSUE ABLATION.

STERILIZATION.

See CONTRACEPTIVE DEVICES.

STERILIZATION OF BIOLOGIC SCAFFOLD MATERIALS

DONALD O. FREYTES
STEPHEN F. BADYLAK
University of Pittsburgh
Pittsburgh, Pennsylvenia

INTRODUCTION

Biologic materials such as collagen, tissue grafts, and extracellular matrix (ECM) derived scaffolds are used for tissue and organ repair and are typically regulated as devices by the U.S. Food and Drug Administration (FDA). As such, these materials must be properly processed and sterilized prior to clinical use. Sterilization of biologic materials involves unique considerations (e.g., the

shrink temperature of collagen, changes in the quaternary ultrastructure of the matrix materials, potential inactivation of any bioactive components, and effects upon surface chemistry and architecture). The purpose of this article is to discuss currently used methods of sterilization for biomaterials with emphasis upon biologic materials, the mechanism by which these sterilization methods destroy or neutralize microbes of interest, and the potential effects of the sterilization methods upon the structure and function of the naturally occurring biologic materials and their inherent bioactive constituents.

STERILIZATION METHODS

The FDA recognizes several effective methods of sterilization for medical devices. These methods include dry heat, moist heat (autoclave), ethylene oxide, ionizing radiation, and liquid chemical sterilants. Each of these methods involves a different mechanism of action for killing microbes, and therefore, the physical and chemical effects of these methods upon naturally occurring molecules will differ. A glossary of relevant terms for this article and the topic of sterilization is provided at the end of the article. A list of current international standards for each of the four sterilization methods described herein can be found in Table 1.

Sterilization Versus Disinfection

There is a clear distinction between the terms sterilization and disinfection. The FDA (1997) defines disinfection as the destruction of pathogenic and other types of microorganisms by thermal or chemical means (1). Sterilization is defined as a process intended to remove or destroy all viable forms of microbial life, including bacterial spores, in order to achieve an acceptable level of sterilization (1). Stated differently, sterilization implies the inactivation and removal of all forms of life. The term terminal sterilization refers to the last sterilization step performed prior to use or commercial distribution of a device.

Bioburden and Sterility Assurance Level

Before sterilizing a device, a certain amount of microbial debris remains (including bacterial wall remnants: pyrogens and endotoxins) on and within each device as a result of production and presterilization processing steps. This microbial debris is referred to as the *bioburden*. The bioburden that exists prior to sterilization is directly proportional to the difficulty of sterilization of a medical device. It is almost impossible to remove all organisms from some materials, and therefore acceptable levels of bioburden have been identified for which devices can be considered as sterile. These acceptable levels are referred to as the sterility assurance level (SAL), which represents the number of microorganisms that would be tolerable, or conversely, the probability that the device is nonsterile (e.g., 10^{-3} means there is a 1 in 10^{-3} chance that an organism survived the sterilization process). The specific SAL for each device will depend on the intended clinical application for the device and the standards established by regulatory agencies. The FDA recommends that implantable devices have a SAL of 10^{-6} while devices contacting intact skin may have a SAL of 10^{-3} (1,2).

Table 1. Summary of Sterilization Standards[a]

Sterilization Method	Standard(s)[b]
Heat Sterilization	ISO 20857
Steam Sterilization	ISO 11134, ISO/DIS 17665-1, ISO/DIS 17665-2
Ethylene Oxide	ISO/DIS 11135-1, ISO/DIS 11135-2, ISO 10993-7 (Residuals)
Radiation	ISO/DIS 11137-1, ISO/DIS 11137-2, ISO/DIS 11137-3
Others	ISO 10993 (1-18), ASTM E1766-95, F2347-03, F2103-1, F2064-00

[a]See http://www.iso.org; http://www.astm.org.
[b]ISO – International Organization for Standardization; ASTM – American Society for Testing and Materials; DIS – Draft International Standard.

Validation of Sterilization Methods

Validation studies of the chosen sterilization method must be performed to ensure proper reduction of endotoxins and pyrogens and appropriate SAL levels, while minimizing exposure to the sterilant. During these validation studies, the microorganisms are quantified by conventional techniques (1,3) following the sterilization process. Various parameters of the sterilization procedure are evaluated (e.g., time, heat, gas concentration), and then the devices are tested for sterility. A classic example of this procedure is the determination of the time needed to sterilize a device by dry heat at a fixed temperature. Briefly, devices are placed at the chosen temperature and samples are periodically removed at different time intervals (e.g., every 5 or 10 min) depending on the overall duration of the test. The number of surviving organisms is quantified at each collection time, the logarithm of the surviving organisms computed, and the values plotted against exposure time. From this curve, the exposure time to achieve the target SAL can be extrapolated as shown in Fig. 1. The same concept can be applied to gas and chemical exposure times.

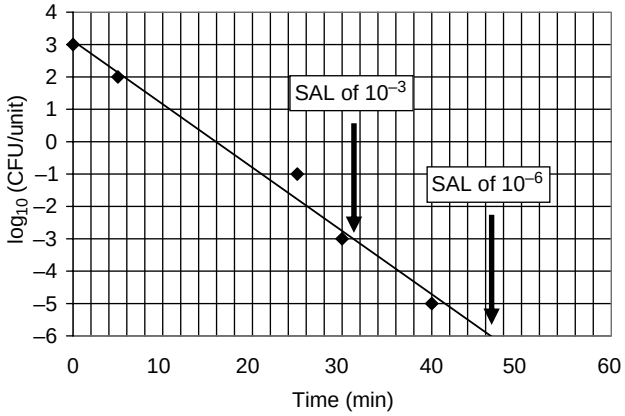


Figure 1. Example of data obtained from a fractional sterilization run to determine the time required to achieve the desired SAL at a fixed temperature. *Note*: The data may not have a linear form and will depend on the microbial load and the type of microbes present.

The determination of the presence of pyrogens and endotoxins levels in a device is another aspect of the validation process. Pyrogens are fever inducing agents most commonly associated with Gram-negative bacteria, but can be produced by most microorganisms. Endotoxins are lipopolysacharides from the cell wall of Gram-negative bacteria capable of inducing certain inflammatory responses. The most common test used to determine the levels of bacterial endotoxins is the limulus amoebocyte lysate (LAL) test (1,3). Pyrogen and endotoxin removal can be achieved via the use of acids, alkaline hydrolysis, hydrogen peroxide, dry heat destruction, and filtration (4).

HEAT STERILIZATION

Heat sterilization can be either applied in a dry or a moist form. Dry heat sterilization is a commonly used method for the sterilization of metallic instruments, powders, and petroleum products (5). Moist heat sterilization is commonly used for surgical instruments and equipment used for cell culture. Moist heat sterilization has also been used for bone grafts to destroy human immunodeficiency (HIV) and hepatitis viruses (6). Except for the sterilization of bone grafts, dry or moist heat is rarely used for biologic materials. Each form of heat sterilization will be discussed separately.

Dry Heat Sterilization

Mechanism of Action. Dry heat sterilization makes use of heated air to inactivate and/or destroy microorganisms. The mechanism of microbial death involves denaturation and coagulation of nucleic acids and proteins. A reduction in the water content of bacterial spores is also considered to be a mechanism by which dry heat sterilization is effective. Although oxidation is possible, it is less widely accepted as a mechanism by which dry heat sterilization exerts its effects (5).

The application of dry heat sterilization utilizes a dry oven-like environment. The air inside the chamber is heated and allowed to equilibrate to a constant temperature (from 120 to 170 °C depending on the duration). The duration of the procedure is a factor of the applied temperature and is determined during the validation process as described earlier. The time and temperature may also vary depending on the nature of the material being sterilized. Table 2 provides guidelines for the sterilization of medical devices using dry heat.

Table 2. Temperatures and Suggested Sterilization Time for a Typical Medical Device[a]

Temperature	Sterilization Time
170 °C (338 °F)	60 min (1 h)
160 °C (320 °F)	120 min (2 h)
150 °C (302 °F)	150 min (2.5 h)
140 °C (284 °F)	180 min (3 h)
121 °C (250 °F)	Overnight

[a]Times suggested by Perkins (5).

Advantages And Disadvantages. The major advantages of dry heat sterilization are low cost and the compatibility with anhydrous oils, powders, and materials not affected by high temperatures. By definition, the moisture content of dry heat sterilization is low. Disadvantages of dry heat sterilization include a relatively long sterilization time and the use of elevated temperatures. Heat diffusion will depend on the ability of the dry heat to diffuse throughout the chamber and the type of material–device being sterilized. The killing rate may be slow due to delayed heat conduction of some materials in which the amount of heat delivered to the microorganisms is limited. Dry heat sterilization cannot be used for most liquids, heat labile substances (i.e., proteins), and heat sensitive materials (1,5).

Physical Effect upon Materials. High temperature can alter bulk properties of polymers and composites and can melt polymers and materials that have a melting temperature below or near the temperature being used for sterilization. The melting temperature of most linear polymers is below the temperatures commonly used in dry heat sterilization (7). Even when the melting temperature is above the temperatures employed, oxidation may still occur in polymers (e.g., nylon) (7). However, polytetrafluoroethylene (PTFE) and silicon rubbers may be effectively sterilized using dry heat (7). High temperatures can adversely affect structural proteins (e.g., collagen and elastin) and modify the mechanical properties of biologic materials (8–10). The high temperature used for dry heat sterilization may also denature bioactive molecules (e.g., growth factors and bioactive peptides present in tissue grafts and naturally occurring biomaterials).

Moist Heat Sterilization (Autoclave)

Mechanism of Action. The mechanism by which moist heat sterilization kills microbes is similar to the mechanism described for dry heat sterilization. Protein and nucleic acid coagulation and denaturation occur quickly once a critical temperature is reached. For the same reasons that protein and nucleic acid destruction inhibit the ability of bacteria to replicate or continue metabolic processes, this method is generally unacceptable for biologic materials that are composed of naturally occurring proteins (5). The high temperatures used with moist heat sterilization are effective for the inactivation of viruses and bacterial spores.

Moist heat (steam) sterilization or autoclaving is conducted under pressurized conditions with saturated steam usually at 121–125 °C. The process typically lasts from 15 to 30 min to ensure that all surfaces are exposed to the moist heat.

Advantages And Disadvantages. The advantages of moist heat sterilization are its efficacy, the relatively low temperature requirements (compared to dry heat sterilization), speed, process simplicity, and the lack of toxic residues when compared to methods such as ethylene oxide (ETO) sterilization (discussed in the next section). Since moist heat sterilization is performed in a pressurized

environment, it can also be used to sterilize liquids. The disadvantages of moist heat sterilization include the presence of water, the use of elevated temperatures, and the possible deposition of impurities present in the steam. Moist heat sterilization is obviously not suitable for heat labile materials or synthetic and natural polymers that are readily degraded by hydrolysis.

Physical Effect upon Materials. Moist heat sterilization can change bulk properties of polymers by hydrolysis and the hydrolytic byproducts may result in the formation of contaminants. For example, exposure of methyl diisocyanate based polyurethanes to prolonged steam sterilization results in the formation of methylene dianiline (due to hydrolysis), which leads to decreased lung function when used in lung perfusion devices (11). Polymers that are prone to hydrolytic degradation include poly(vinyl chlorides) (PVC), polyacetals, polyethylenes, and polyamides (7). For example, the mechanical properties of PVC can be adversely affected by repeated steam sterilization due to rearrangement of macromolecular chains (12). A separate example of the potential harmful effects of moist heat sterilization is the formation of oligomer crystals on the surface of Dacron grafts as a result of steam sterilization that can cause hemolysis when this material is used as a vascular graft (13). With respect to biologic materials, steam sterilization can have adverse effects upon the mechanical properties of tissue grafts as in the case of bone allografts (14). For obvious reasons, the denaturing effects of steam sterilization upon protein structures caused by the elevated temperatures makes this method generally unsuitable for most biologic materials.

In summary, both dry and moist heat forms of sterilization, although effective for nonbiologic materials, are typically unsuitable for the sterilization of biologic materials, due to adverse effects upon the structure and function of protein and non-protein constituents.

ETHYLENE OXIDE STERILIZATION

Ethylene oxide sterilization is a commonly used method for the sterilization of heat-sensitive materials, medical equipment, and biologic materials including those composed of ECM, such as TissueMend (fetal bovine skin from TEI Biosciences), and OaSis (small intestinal submucosa/SIS extracellular matrix from Cook Biotech, Inc.). Sterilization via ETO has also been explored for demineralized bone, tendons, dura mater, and fascia lata (15).

Mechanism of Action. Ethylene oxide is an unstable ring structure capable of reacting via alkylation with functional groups found in nucleic acids and proteins (1,16,17). Examples of reactions between functional groups and ETO are listed in Fig. 2. Sterilization via ETO exposure begins with the placement of the target device into a pressurized sterilization chamber. The humidity within the chamber is controlled by the introduction of moisture (40–90% humidity) and the temperature is maintained between 40 and 50 °C. The ETO is then introduced into the chamber at concentrations ranging from 600 to 1200 $mg \cdot L^{-1}$ for a sufficient time (typically 2–48 h) to achieve the desired SAL. Following sterilization, room air is used to flush the vessel for removal of residual ETO and its toxic byproducts. Longer flushing time minimizes the presence of toxic byproducts but increases the overall sterilization time.

Advantages and Disadvantages. The advantages of using ETO sterilization include the relatively low temperature requirements (compared to heat sterilization), and the high degree of penetration of the ETO gas into the target device. Temperature and moisture sensitive materials are more readily sterilized via ETO than by heat sterilization methods. Perhaps the greatest disadvantage of ETO sterilization is the toxicity and carcinogenicity of the residual byproducts: ETO, ETC, and ETG (see Fig. 3). For samples weighing >100 mg, acceptable levels of ETO, ETC, and

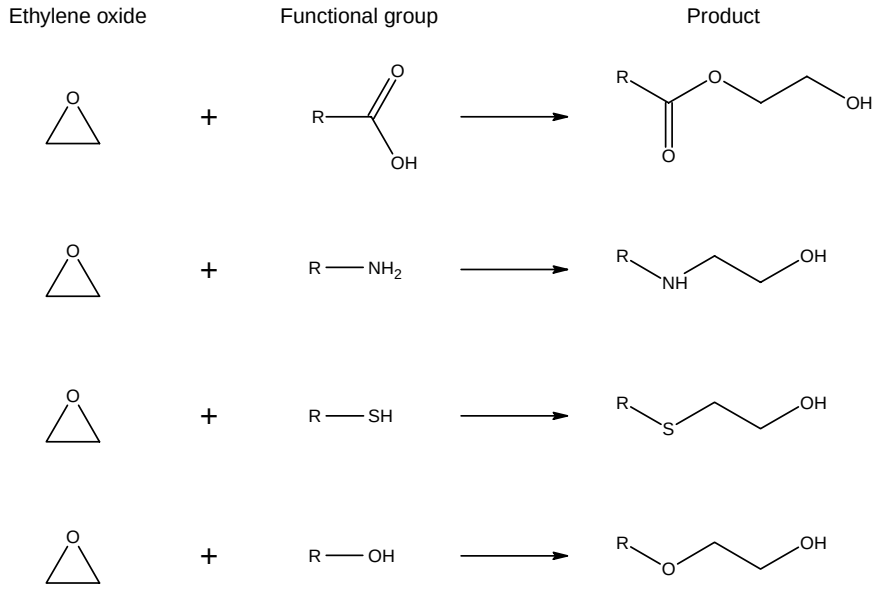


Figure 2. Examples of possible reaction between ETO and common functional groups in biological molecules (16,17). For a more complete explanation of the reactions please refer to (16,19).

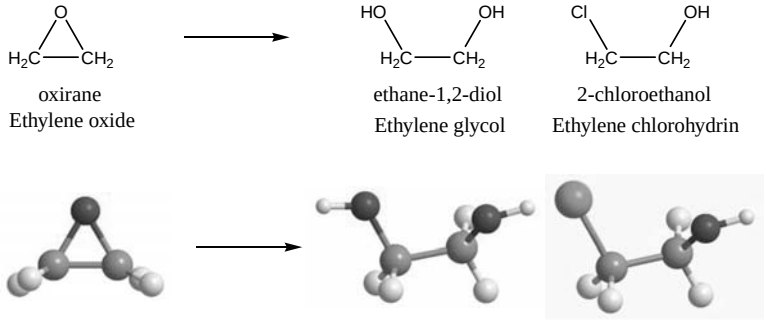


Figure 3. Structure of ETO and its byproducts: Ethylene glycol (ETG) and ethylene chlorohydrin (ETC). The ETO is flammable and highly explosive and has been associated with potential mutagenic, teratogenic, or carcinogenic properties. Both ETG and ETC are toxic as well (18).

ETG are 2500, 2500, and 5000 ppm, respectively (1,17). In addition to the alkylation reaction referred to above, ETO may also react with functional groups on the surface of biomaterials and with constituent proteins causing changes in bioactivity and bulk properties of the device.

Physical Effect upon Materials. Materials coated with proteins (e.g., albumin and heparin) may lose the benefit of these coatings as a result of ETO interactions with these surface molecules (20). Protein cross-links, alkylation reactions, and other chemical changes that occur as a result of ETO interactions may also have adverse effects upon the bioactivity and the mechanical properties of biologic materials (e.g., allografts) (15,21–23).

In summary, ETO sterilization with proper aeration and/or removal of toxic by products may be safely used to sterilize biologic materials. However, changes in bioactivity and structural properties can still occur using this method of sterilization.

IONIZING RADIATION STERILIZATION

Sterilization via radiation is the most commonly used technique to sterilize biologic materials at the present time. Examples of such devices include Restore™ (SIS-ECM used for orthopedic applications, DePuy Orthopedics), CuffPatch™ (carbodiimide crosslinked SIS-ECM, Arthrotech), and Permacol™ (crosslinked porcine dermis, Tissue Science Laboratories). Ionizing radiation is the preferred method of sterilization for other biologic materials such as heart valves, skin, fascia, dura mater, bone, and tendon grafts (24). There are different types of irradiation, including gamma irradiation and electron beam irradiation, that are approved by the FDA and commonly used by medical device manufacturers. These methods will be discussed separately.

Gamma Irradiation

Mechanism of Action. Gamma radiation arises from the decay of cobalt (^{60}Co) or cesium isotopes (^{137}Cs). Gamma irradiation exists in the form of photons generated from the transition of an atomic nucleus from an excited state to a ground state. The resultant high energy particles induce ionization by transforming an uncharged atom to a charged atom with the subsequent release of an electron that in turn collides with other atoms. The resulting discharge of secondary electrons creates oxidizing free radicals that damage proteins and deoxyribonucleic acid (DNA) molecules by dimerization of bases and altering the sugar phosphate backbone. These changes reduce microbe's capacity to replicate or continue necessary metabolic functions (25).

Gamma sterilization is conducted by placing the target device in front of a radiation source, usually cobalt (^{60}Co) or cesium (^{137}Cs), that is usually directed by a window in a lead shielded container (Fig. 4). The dose of gamma irradiation is adjusted by varying the distance between the source and the device or by varying the exposure time. The dose or radiation absorption is commonly expressed in terms of "radiation absorbed dose" (rad) or grays (Gy). The former represents the absorption of 10^{-5} joules (J) per gram (g) ($J \cdot g^{-1}$) while the later represents the absorption of 1 $J \cdot kg^{-1}$. Hence, 1 kGy = 0.1 Mrads. Typical sterilizing doses range between 6 and 25 kGy.

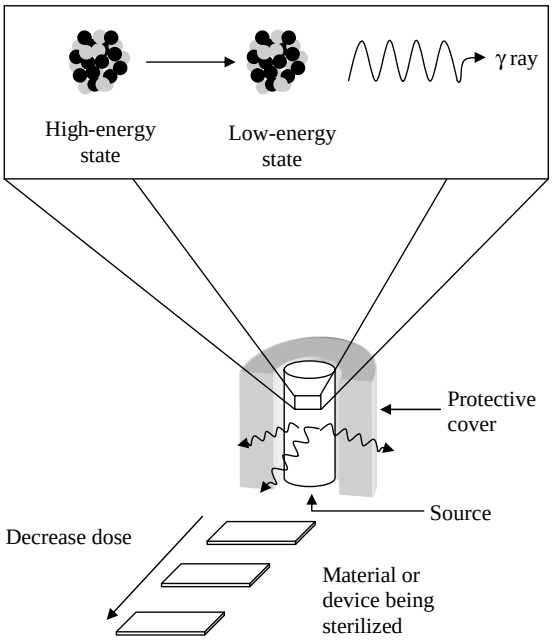


Figure 4. Schematic representation of a gamma source and how devices are typically sterilized. Dosage may be adjusted by calibrating the distance between the source and the device. Other forms may be employed depending on the facilities and the intended application.

Advantages And Disadvantages. The advantages of gamma irradiation include its compatibility with many materials, the minimal amount of toxic residues, and the low temperature requirements. Disadvantages of gamma irradiation include its cost, complexity (e.g., requires highly specialized facilities), and incompatibility with some polymers due to degradation (e.g., polymethylmetacrylamide and cellulose derivatives) or cross-linking (e.g., polyethylene and polystyrene) of the polymer (7). Gamma irradiation also may adversely affect proteins in a dosage dependent manner by the introduction of free radicals and has been associated with reduction in strength of grafts and collagenous biomaterials (26–30).

Physical Effect Upon Materials. Ionizing radiation affects the backbone of synthetic and natural polymers and may cause cross-linking, oxidation, or chain scission (7,24,25). These changes may affect degradation kinetics and material properties (e.g., tensile strength, elastic modulus, elongation, and the color of the material). In general, polymeric materials, including biologic polymers, can be safely sterilized via gamma irradiation. Safe levels of gamma irradiation include a range of 2.5 Mrads (collagen-based materials, ECM scaffolds, polypropylene, polytetrafluoroethylene) to 1000 Mrads (polystyrene and epoxy resins) (7). The lower doses of gamma irradiation (e.g., 2.5 Mrads) effectively sterilize biologic materials with minimal adverse effects to the physical properties of the material.

e-Beam Irradiation

Mechanism of Action. e-Beam irradiation involves the generation of a beam of electrons from a linear accelerator. The beam of electrons can be manipulated to achieve the desired dose by varying power and the exposure time. The sterilizing effect of e-beam irradiation on polymeric materials is similar to that described for gamma irradiation except for less penetrability of the beam into target materials. The main effect of e-beam irradiation comes from the ionizing electrons that result in the destruction of nucleic acids and proteins required for cellular processes. One primary difference between gamma radiation and e-beam is that the electron beam can be focused on the target material and turned off when the sterilization process is completed.

Advantages And Disadvantages. The advantages of e-beam irradiation compared to gamma irradiation include the fast sterilization time, less safety concerns (i.e., the source can be turned off), greater power, and the source does not deplete as occurs with the source of gamma irradiation. The disadvantages of e-beam irradiation include its cost and less penetrability than gamma irradiation. As in the case with gamma irradiation, electron beam irradiation has also been associated with a decrease in strength of collagenous materials (27–30).

Physical Effect Upon Materials. The overall effect of e-beam irradiation is similar to those described for gamma irradiation. Free radicals are created following exposure to the electron beam and if oxygen is present, peroxyl free radicals can be formed that increase the rate of chain scission (25). Polylactic-*co*-glycolic acid (PLGA) exposed to e-beam shows a decrease in the molecular weight (M_w) and thermal properties mainly due to chain scission and crosslinking (31,32). e-Beam has also been shown to reduce the degradation rate of PVC and polypropylene (PP) when compared to gamma sterilization, mainly due to faster free radical termination reactions (25).

The effects of e-beam irradiation upon naturally occurring, biologic scaffold materials is dose dependent. In the lower dose range ($\sim$ 2.5 Mrad), there are minimal changes upon the physical and biologic properties of the irradiated materials (24). However, changes in strength (33,34) and degradability (35) of collagenous materials and the creation of cross-links may still occur as a result of e-beam exposure (33,34).

Table 3 summarizes some of the naturally occurring biomaterials currently available and the method used to terminally sterilize the device. Table 4 summarizes the advantages, disadvantages, and the typical doses for the sterilization methods discussed.

ALTERNATIVE METHODS OF STERILIZATION

Alternative approved methods of sterilization that are recognized by the FDA include liquid chemical sterilants,

Table 3. Example of Currently Marketed Biologic Devices and the Form of Terminal Sterilization Employed

Product	Company	Source	Sterilization
CuffPatch™	Arthrotech, Inc.	Crosslinked porcine small intestine	Gamma
Pelvicol™	Bard, Inc.	Crosslinked porcine dermal collagen	Gamma
Bard® Dermal Gllograft	Bard, Inc.	Cadaveric dermis minus epidermal layer	Gamma
FasLata® Allograft	Bard, Inc.	Cadaveric fascia lata	Gamma
OaSis®	Cook Biotech, Inc.	Porcine small intestine	ETO
Stratasis®	Cook Bioetch, Inc.	Porcine small intestine	ETO
Surgisis®	Cook Biotech, Inc.	Porcine small intestine	ETO
Restore™	Depuy Orthopedics, Inc.	Porcine small intestine	Gamma
TissueMend®	TEI Bioscience, Inc.	Fetal bovine skin	ETO
Permacol™	Tissue Science Laboratories, Inc.	Porcine dermis	Gamma
Tutopatch®	Tutogen Medical, Inc.	Bovine pericardium	Gamma
Tutoplast®	Tutogen Medical, Inc.	Human fascia lata	Gamma

Table 4. Summary of the Most Commonly Used Sterilization Methods and Their Advantages, Disadvantages, and Typical Doses as Discussed in this Article

Sterilization Method	Advantages	Disadvantages	Typical Dose
Dry heat	Inexpensive; Ease of use	Thermal damage to proteins	Dry air at 120–170 °C (250–338°F) for 1–18 h
Moist heat	Faster than dry heat and less heat requirement; Inactivates some viruses	Thermal damage to proteins	Saturated steam at 121–125 °C for 15–30 min (pressure may be adjusted)
Ethylene oxide	Lower temperatures than heat sterilization and less moisture	Toxic by-products; Potential reactions with functional groups	ETO is added at 600–1200 mg·L^{-1} for 2–48 h (40–90% humidity at 40–50 °C)
Ionizing radiation	Minimal thermal damage; Good penetration	Cost; Changes in material properties	6–24 kGy

high intensity light, ultraviolet light, vapor systems (combination of hydrogen peroxide and peracetic acid), exposure to chlorine dioxide, and filtration methods (510 k Sterility Review Guidance K90-1). In addition, low temperature gas plasma and machine-generated X rays can be used for sterilization purposes.

Chemicals are often used to reduce the bioburden, disinfect, and sterilize biologic materials. Table 5 lists a few of the most commonly used chemicals for the disinfection and/or sterilization of grafts and naturally occurring biomaterials. Glutaraldehyde is currently used to sterilize and remove the antigenicity of many porcine-derived products (e.g., heart valves and pericardium). Formaldehyde and a combination of formaldehyde and low temperature steam is another form of sterilization used for medical devices. However, this method is seldom used for biologic materials due to its high toxicity.

Extracellular matrix based bioscaffolds (e.g., Restore, DePuy Orthopaedics) and OaSis® (Cook Biotech, Inc.)™ undergo presterilization disinfection with a peracetic acid and ethanol treatment followed by water and phosphate buffered saline washes to remove any chemical residues. Table 6 summarizes a typical disinfection method used for naturally occurring biomaterials prior to sterilization. Collagen-based products (e.g., Contigen, Bard, Inc.) rely on sterile processing techniques, acidic environments, chemical cross-linking (e.g., glutaraldehyde), and sterile filtration (e.g., via 0.22 μ filters) to achieve the desired sterility.

Table 5. Common Chemical Sterilants and Disinfectants Used for Biologic Materials

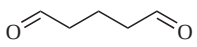
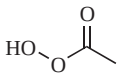
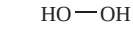
Chemical	Uses	References
 Glutaraldehyde	Used for sterilization and cross-linking of tissues (e.g., heart valves, pericardium, tendons, and collagen based biomaterials).	(1,36–39)
 Peracetic acid	Used to disinfect ECM derived bioscaffolds and disinfect and sterilize medical devices. It is usually combined with hydrogen peroxide and ethanol.	(1,36,40–47)
 Hydrogen peroxide	Used to sterilize grafts and medical devices. It is usually combined with steam sterilization or with other sterilants such as peracetic acid.	(1,23,48–52)
 Ethanol	Used to disinfect ECM derived bioscaffolds and grafts. It is usually combined with peracetic acid and it is also used to disinfect or sterilize instruments.	(1,36,39,45,53–55)

Table 6. Example of the Disinfection Process Used to Reduce the Bioburden of Naturally Occurring Biomaterials[a]

Step	Solution	Time
Peracetic acid disinfection	0.1% (v/v) peracetic acid 4% (v/v) ethanol, and 95.9% (v/v) sterile water	2 h
Phosphate buffer saline wash	Sterile 1X PBS pH 7.4	30 min
Water wash	Sterile water	30 min

[a]See Ref. 56.

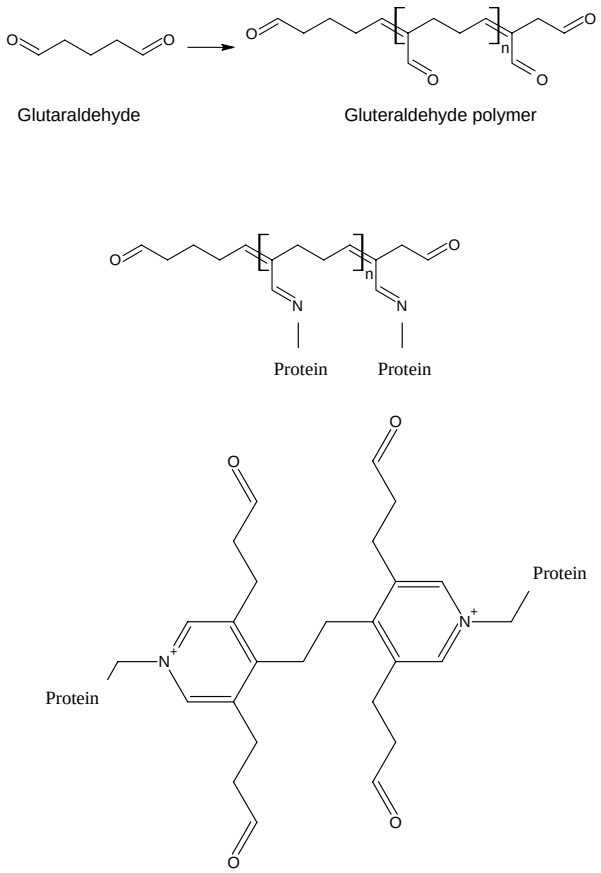


Figure 5. Reaction of glutaraldehyde polymers with amino groups and a pyridinium cross-link resulting from protein exposure to glutaraldehyde. (Redrawn from Ref. 1).

High levels of peracetic acid and combinations of peracetic acid (PAA), ethanol, isopropyl alcohol, hydrochloric acid, antibiotics, hydrogen peroxide, and detergents have been used for sterilization of demineralized bone grafts (48). A combination of steam sterilization and hydrogen peroxide may also be used for sterilizing allogeneic bone, but risks of denaturing collagen molecules still exists. Peracetic acid and hydrogen peroxide have been used for the disinfection and sterilization of acellular dermal matrices (24). The negative effects of these alternative sterilization methods upon biological materials include the molecular cross-linking that is caused by the chemical sterilant (see Fig. 5), collagen and protein denaturation, and the adverse effects of acids (e.g., the disruption of sulfhydryl and sulfur bonds by PAA) upon material properties at the concentrations needed for sterilization.

SUMMARY

There are several options for the sterilization of biologic materials and each option has its advantages and disadvantages. The majority of biologic materials classified as medical devices are sterilized by either gamma or e-beam irradiation or by ethylene oxide. These methods tend to have dose-dependent effects upon the structure and function of biologic materials (e.g., strength of irradiated materials), but doses that effectively sterilize the material can be used with minimal effects upon the structure and function of currently marketed biologic devices.

GLOSSARY

Allograft	Organ or tissue graft obtained from the same species.
Bioburden	The microbiologic load or the number of contaminating organisms in the product before sterilization.
Disinfection	The destruction of pathogenic and nonpathogenic microorganisms by thermal or chemical means. The FDA (1997) definitions as given in Ref. 1.
Endotoxins	Toxic components of the outer membrane of Gram-negative bacteria that cause fever.
Extracellular matrix (ECM)	An insoluble network of polysaccharides and structural and non-structural proteins secreted by host cells and present in all tissues and organs.
Inactivation	Removal or inactivation of the activity of microorganisms by killing or inhibiting reproductive or enzymatic activity. The FDA (1997) definitions as given in Ref. 1.
Medical device	A device used for diagnosis, cure, treatment, or prevention of a disease or condition. A device that affects the structure and function of the body, does not achieve intended use through chemical reaction, and it is not metabolized. The FDA (1997) definitions as given in Ref. 1.
Pyrogens	Fever inducing agents (e.g., signals from inflammatory cells, chemicals, endotoxins, cell remnants from Gram-positive bacteria, and fungi).
Spore	The dormant state of an organism, typically a bacterium or fungus, which exhibits a lack of biosynthetic activity and reduced respiratory activity, and has resistance to heat, radiation, desiccation, and various chemical agents. The FDA (1997) definitions as given in Ref. 1.

Sterility assurance level (SAL)	An indicator that no greater than the predetermined number of microorganisms exists in a product usually represented as the number of colony forming units per device.
Sterilization	A process intended to remove or destroy all viable forms of microbial life, including bacterial spores.
Xenograft	Organ or tissue graft obtained from a different species.

BIBLIOGRPHY

1. Block SS. Disinfection, Sterilization, and Preservation. 5th ed. Philadelphia (PA): Lippincott Williams & Wilkins; 2001. p 1481.
2. Ratner BD. Biomaterials science: An introduction to materials in medicine. 2nd ed. 1996, San Diego: Academic Press; 1996. p 484.
3. Baird RM, Hodges NA, Denyer SP. Handbook of Microbial Quality Control. Pharmaceuticals and Medical Devices. New York: Taylor & Francis; 2000.
4. Halls NA. Achieving Sterility in Medical and Pharmaceutical Products. Vol. 64. New York: Marcel Dekker, Inc.; 1994.
5. Perkins JJ. Principles and Methods of Sterilization in Health Sciences. 2nd ed. Springfield (IL): Thomas; 1969. p 560.
6. Pruss A, et al. Comparison of the efficacy of virus inactivation methods in allogeneic avital bone tissue transplants. Cell Tissue Bank 2001;2(4):201–215.
7. Bronzino JD. The Biomedical Engineering Handbook. Boca Raton (FL): CRC Press: IEEE Press; 1995. p 591.
8. Chen SS, Humphrey JD. Heat-induced changes in the mechanics of a collagenous tissue: Pseudoelastic behavior at 37 degrees C. J Biomech 1998;31(3):211–216.
9. Jun JH, et al. Effect of thermal damage and biaxial loading on the optical properties of a collagenous tissue. J Biomech Eng 2003;125(4):540–548.
10. Harris JL, Humphrey JD. Kinetics of thermal damage to a collagenous membrane under biaxial isotonic loading. IEEE Trans Biomed Eng 2004;51(2):371–379.
11. Mazzu AL, Smith CP. Determination of extractable methylene dianiline in thermoplastic polyurethanes by HPLC. J Biomed Mater Res 1984;18(8):961–968.
12. Habermann V, Waitzova D. On the safety evaluation of extracts from synthetic polymers used in medicine. Arch Toxicol (Suppl) 1985. 8:458–460.
13. Berger K, Sauvage LR. Late fiber deterioration in Dacron arterial grafts. Ann Surg 1981;193(4):477–491.
14. Speirs AD, et al. Biomechanical properties of sterilized human auditory ossicles. J Biomech 1999;32(5):485–491.
15. Prolo DJ, Pedrotti PW, White DH. Ethylene oxide sterilization of bone, dura mater, and fascia lata.for human transplantation. Neurosurgery 1980;6(5):529–539.
16. Fraenkel-Conrat H. The action of 1,2-epoxides on proteins. J Biol Chem 1944;154:227–249.
17. Gad SC, ebrary Inc. Safety Evaluation of Medical Devices. 2nd ed. New York: Marcel Dekker; 2002. p 558.
18. Star EG. [The toxic effect of ethylene chlorohydrin and ethylene glycol on experimental animals and human cell cultures (author's transl)]. Bakteriol Mikrobiol Hyg [B] 1980;171(1): 25–32.
19. Parker RE, Isaacs NS. Mechanisms of eopoxide reactions. Chem Rev 1959;59:737–797.
20. Guidoin R, et al. A compound arterial prosthesis: the importance of the sterilization procedure on the healing and stability of albuminated polyester grafts. Biomaterials 1985; 6(2):122–128.
21. Kearney JN, et al. Evaluation of ethylene oxide sterilization of tissue implants. J Hosp Infect 1989;13(1):71–80.
22. Kudryk VL, et al. Toxic effect of ethylene-oxide-sterilized freeze-dried bone allograft on human gingival fibroblasts. J Biomed Mater Res 1992;26(11):1477–1488.
23. Thoren K, Aspenberg P. Ethylene oxide sterilization impairs allograft incorporation in a conduction chamber. Clin Orthop Relat Res 1995;318:259–264.
24. Gaughran ERL, Goudie AJ, Johnson and.Johnson Inc. Sterilization of medical products by ionizing radiation: International conference. Vienna, Austria, April 25–28, 1977. Sterilization by Ionizing Radiation. Vol. 2. Montreal: Multiscience Publication; 1978. p 408.
25. Woo L, Purohit KS. Advancements and opportunities in sterilisation. Med Device Technol 2002;13(2):12–17.
26. Olde Damink LH, et al. Influence of ethylene oxide gas treatment on the in vitro degradation behavior of dermal sheep collagen. J Biomed Mater Res 1995;29(2): 149–155.
27. De Deyne P, Haut RC. Some effects of gamma irradiation on patellar tendon allografts. Connect Tissue Res 1991;27(1): 51–62.
28. Roe SC, et al. The effect of gamma irradiation on a xenograft tendon bioprosthesis. Clin Mater 1992;9(3–4):149–154.
29. Yahia LH, Drouin G, Zukor D. The irradiation effect on the initial mechanical properties of meniscal grafts. Biomed Mater Eng 1993;3(4):211–221.
30. Godette GA, Kopta JA, Egle DM. Biomechanical effects of gamma irradiation on fresh frozen allografts in vivo. Orthopedics 1996;19(8):649–653.
31. Loo JS, Ooi CP, Boey FY. Degradation of poly(lactide-co-glycolide) (PLGA) and poly(L-lactide) (PLLA) by electron beam radiation. Biomaterials 2005;26(12):1359–1367.
32. Loo SC, Ooi CP, Boey YC. Influence of electron-beam radiation on the hydrolytic degradation behaviour of poly (lactide-co-glycolide) (PLGA). Biomaterials 2005;26(18):3809–3817.
33. Chuck RS, et al. Biomechanical characterization of human amniotic membrane preparations for ocular surface reconstruction. Ophthalmic Res 2004;36(6):341–348.
34. Fujisato T, et al. Cross-linking of amniotic membranes. J Biomater Sci Polym Ed 1999;10(11):1171–1181.
35. Grimes M, Pembroke JT, McGloughlin T. The effect of choice of sterilisation method on the biocompatibility and biodegradability of SIS (small intestinal submucosa). Biomed Mater Eng 2005;15(1–2):65–71.
36. Sprossig M, et al. [Sterilization of biologic heart valve prostheses with glutardialdehyde]. Z Exp Chir 1973;6(4):248–251.
37. Wallace RB. Tissue valves. Am J Cardiol 1975;35(6):866–871.
38. Munting E, et al. Effect of sterilization on osteoinduction. Comparison of five methods in demineralized rat bone. Acta Orthop Scand 1988;59(1):34–38.
39. Sung HW, Hsu HL, Hsu CS. Effects of various chemical sterilization methods on the crosslinking and enzymatic

degradation characteristics of an epoxy-fixed biological tissue. J Biomed Mater Res 1997;37(3):376–383.
40. Mucke H, Wenzel KP. [Preparation of heart valves for grafting after sterilization with peracetic acid]. Z Exp Chir 1973;6(4):252–255.
41. Sprossig M, et al. [Sterilization of heart valve transplants with peracetic acid]. Helv Chir Acta 1973;40(3):357–362.
42. Wutzler P, et al. [Combined cleaning and cold sterilization procedure for cleaning rooms in virological establishments]. Z Med Labortech 1975;16(5):253–259.
43. Wenzel KP. [Final sterilization of formaldehyde-preserved bioprostheses using peracetic acid]. Z Exp Chir 1982; 15(4):261–263.
44. Lomas RJ, et al. Assessment of the biological properties of human split skin allografts disinfected with peracetic acid and preserved in glycerol. Burns 2003;29(6):515–525.
45. Pruss A, et al. Peracetic acid-ethanol treatment of allogeneic avital bone tissue transplants—a reliable sterilization method. Ann Transplant 2003;8(2):34–42.
46. Huang Q, et al. Use of peracetic acid to sterilize human donor skin for production of acellular dermal matrices for clinical use. Wound Repair Regen 2004;12(3):276–287.
47. Lomas RJ, et al. Effects of a peracetic acid disinfection protocol on the biocompatibility and biomechanical properties of human patellar tendon allografts. Cell Tissue Bank 2004; 5(3):149–160.
48. Glowacki J. A review of osteoinductive testing methods and sterilization processes for demineralized bone. Cell Tissue Bank 2005;6(1):3–12.
49. Brown SA, et al. Effects of different disinfection and sterilization methods on tensile strength of materials used for single-use devices. Biomed Instrum Technol 2002;36(1):23–27.
50. Lambert RJ, Johnston MD, Simons EA. A kinetic study of the effect of hydrogen peroxide and peracetic acid against Staphylococcus aureus and Pseudomonas aeruginosa using the bioscreen disinfection method. J Appl Microbiol 1999;87(5):782–786.
51. Rutala WA. Disinfection and sterilization of patient-care items. Infect Control Hosp Epidemiol 1996;17(6):377–384.
52. Rutala WA, Weber DJ. Disinfection of endoscopes: review of new chemical sterilants used for high-level disinfection. Infect Control Hosp Epidemiol 1999;20(1):69–76.
53. Pruss A, et al. Validation of the sterilization procedure of allogeneic avital bone transplants using peracetic acid-ethanol. Biologicals 2001;29(2):59–66.
54. Hodde J, Hiles M. Virus safety of a porcine-derived medical device: evaluation of a viral inactivation method. Biotechnol Bioeng 2002;79(2):211–216.
55. Scheffler SU, et al. Biomechanical comparison of human bone-patellar tendon-bone grafts after sterilization with peracetic acid ethanol. Cell Tissue Bank 2005;6(2):109–115.
56. Freytes DO, et al. Biaxial strength of multilaminated extracellular matrix scaffolds. Biomaterials 2004;25(12):2353–2361.

See also BIOMATERIALS FOR DENTISTRY; BIOMATERIALS: TISSUE ENGINEERING AND SCAFFOLDS; ENGINEERED TISSUE.

STETHOSCOPE. See LUNG SOUNDS.

STOMACH, ELECTRICAL ACTIVITY. See ELECTROGASTROGRAM.

STRAIN GAGES

PAUL C. DECHOW
QIAN WANG
A & M University Health Science Center
Dallas, Texas

INTRODUCTION

Strain gages are devices that allow measurement of the change in the dimensions, displacement, or deformation of an object. These devices have been used extensively in a wide variety of engineering applications over the past century and, in recent decades, their use in biomedical applications, such as in the production of transducers for biomedical monitoring and research, has increased tremendously. These transducers measure a variety of parameters, including strain, displacement, pressure, acceleration, force, and temperature. The chief types of strain gages are mechanical, optical, acoustical, and electrical. There are several types of electrical gages including capacitance, inductance, semiconductor, and resistance strain gages. The electrical resistance strain gages are the most widely used types of gages in engineering and biomedical applications today and receive the most extensive treatment in this article. Before discussing the various types of strain gages in more detail, it is useful to have a working understating of what strain and stress are and their relationship to each other.

STRAIN AND STRESS

Strain is a dimensionless unit, which is defined as the ratio of the change in unit length over the original length,

$$\varepsilon = \Delta L/L \tag{1}$$

where ε is strain, L is the original length of the object along an axis, and ΔL is the change in length along that axis. Measurements of strain are usually given in microstrain ($\mu\varepsilon$) where $1.0\ \mu\varepsilon = 1 \times 10^{-6}\ \varepsilon$. By convention, if an object is shortened in length, the strain is compressive and takes a negative value. If an object is lengthened, the strain is tensile and takes a positive value.

Although strain gages directly measure a change in electrical resistance that is proportional to a change in dimension, they can also be used to measure force or pressure if some of the properties of the test material are known. For many metals and other solids, a constant can be used to define a linear relationship between deformation or strain (ε) and stress (σ), where stress is defined

as force per unit area. This relationship, also known as Hooke's law, can be expressed as

$$\sigma = E\varepsilon \tag{2}$$

Where stress (σ) is directly proportional to strain (ε). The constant E is called the modulus of elasticity, the elastic modulus, or Young's modulus. Since strain is a dimensionless quantity and stress is defined in units of force per area, such as pascals (newtons per meter squared), E is also given in units of force per area.

Strain gages usually measure strain on a small portion of the surface area of a structure. A series of strain gages affixed to a structure allows determination of the way in which the structure is being deformed; the deformation can result from either axial, shearing, twisting, or bending loads. If a structure is loaded so that it is deformed in a predictable way, then strain gages can be affixed to the structure, calibrated, and used to measure the magnitude of the loading. An example of this kind of transducer would be a load cell, which can be used to measure force when load axially.

TYPES OF STRAIN GAGE

Depending on the material and the test situation, the parameters of strain measurement vary widely and require selection of a strain gage appropriate for the particular problem. Considerations in the selection of a gage include factors, such as knowledge of the required accuracy and stability of the gage, the maximum deformation of the material, the duration of the test, patterns and amount of loading of the gage, and the location and situation of gage installation. The following discussion describes some features and limitations of strain gages that are currently in use. Due to their widespread use, greater consideration is given to electrical resistance strain gages.

Mechanical Strain Gages

Mechanical strain gages have been in use longer than other types. However, due to their large size and relative inaccuracy, their use in engineering applications today is limited. These gages are most commonly used to measure strain in industrial applications where it is appropriate to summate strain over a range of several inches. Mechanical strain gages typically consist of a system of two points or knife-edges that can be securely attached to a structure. Then a series of compound levers within the gage magnify the displacement allowing a reading to be taken. The limited range of accuracy of these gages, with measurements restricted to magnifications of up to 2000 (500 $\mu\varepsilon$), contrast with the greater accuracy of electrical resistance strain gages (see below). Yet such gages are most useful where the size of the gage is not an issue and the ability to take a reading from a simple vernier or dial scale, without any associated electric instrumentation, is at a premium.

An additional type of mechanical strain gage, the electromechanical strain gage or extensometer, provides greater accuracy than other types of mechanical strain gages and is used in a variety of current industrial applications and in materials testing.

Optical Strain Gages

Optical techniques have been used in a variety of ways to measure strain. The simplest optical strain gages are similar to mechanical strain gages except that light rays are substituted for mechanical levers in magnifying the displacement. This change serves to decrease the size and inertia of the gage while making it appropriate for use at low frequencies in dynamic applications. The use of lasers as collimated light sources has led to the development of several optical strain gages, including those based on principles of diffraction. The diffraction strain gage, like some of the mechanical gages, has two blade-like edges that are bonded or welded to the test specimen. This gives the gage the advantage of automatic temperature compensation, if the blades are constructed of the same material as the test specimen, making the gage suitable for some use in extreme temperatures.

Another interesting optical gage is the interferometric strain gage. This gage measures strain by examining interference patterns caused by directing a light source, such as a helium–neon laser, at two V-grooves placed on the surface of a specimen. This method is most useful in test situations where it is preferable not to actually attach a strain gage to the test specimen, thus eliminating problems of bonding, inertia of the gage, and temperature compensation. A similar noncontacting device is the infrared (IR) extensometer, which is used in materials testing when deformations are large, such as with elastomeric and highly extensible materials, where strain gages or contracting extensometers cannot be readily attached. Resolutions of up to 5 μm can be obtained with these devices.

Electrical Strain Gages: Capacitance and Inductance Types

These two types of electrical strain gages are not as widely used as the semiconductor and resistance types, yet find uses in specialized applications and in the production of transducers. The capacitance strain gage uses a parallel plate capacitor where the positional relationship of the two plates varies with the capacitance. For example, increasing the distance between the two plates will effect such a change.

There are several types of inductance strain gages. One notable example is the linear variable differential transformer (LVDT). This device is useful for measuring displacements that range from several micrometers to several centimeters, thus making it useful in situations that require measurements of larger displacements than can be measured by electrical resistance strain gages. The device can also be modified with auxiliary mechanisms to measure velocity, acceleration, force, pressure, or flow rate. The LVDT consists of a freely moving iron core surrounded by a primary and two identical secondary coils. The device can be designed such that the voltage output is a linear function of the displacement of the core within the coils over a specific range.

Electrical Strain Gages: Semiconductor Type

Semiconductor strain gages are widely used gages, especially in the production of load cells, miniaturized

transducers, and other applications requiring measurement of very small strains. These gages provide a large signal output relative to strain, as indicated by their large gage factors (GF for a definition see below), which can be as high as 200. By contrast, GFs for electrical resistance strain gage are usually two or less. Unfortunately, semiconductor strain gages also suffer from an extreme sensitivity to temperature. In addition, the piezoresistive effect varies with strain giving these gages ranges of nonlinearity. Semiconductor gages are typically constructed from silicon or germanium. Because of the high receptivity of these materials, the gages consist of a small filament of a single crystal of the semiconductor material mounted on some type of carrier. The receptivity and strain-measuring properties of the crystal can be varied by altering the amount of impurities in the crystal. The fatigue life of semiconductor gages for cyclic strains is less than that of electrical resistance-type gages. These gages are then commonly employed in fields with low strains where frequent loading does not lead to gage failure.

Electrical Strain Gages: Resistance Type

The electrical resistance strain gages are the most widely used gages in engineering and biomedical fields (Fig. 1). The principles behind these gages were first discovered by Lord Kelvin in 1856 when he noted (1) that the resistance of a metal wire increases with increasing strain and decreases with decreasing strain and (2) that different materials have different sensitivities to strain.

Electrical resistance strain gages are also sensitive to temperature changes, but not to the extent of semiconductor gages. Most gages have a several hundred degree range under which they function best. However, within that range, temperatures must be monitored carefully in order to compensate for apparent strain caused by shifts in temperature. However, gages are available that have self-temperature compensation. In these gages, the thermal expansion coefficient of the gage is matched to that of the test material.

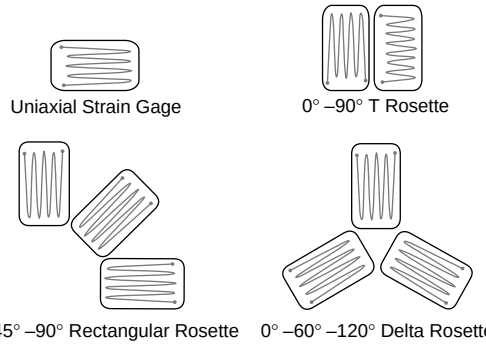


Figure 1. Four basic gage patterns (From eFunda Inc., with permission). Top left: single-element gage. This gage can be used to measure either maximum or minimum strains. Top right: T-style gage with two perpendicular elements. This gage can be used to measure maximum, minimum, and shear strains if the gage is aligned with the principle strains. Bottom left: rectangular rosette strain gage. This gage, as well as the delta rosette strain gage (bottom right), can be used to measure maximum, minimum, and shear strains and need not be aligned with the principal strains.

Electrical resistance strain gages are available on a variety of backing materials ranging from various metal foils to polyamide and epoxy. The backing materials affect the performance of the gage and are varied to achieve the desired temperature range, extensibility, and fatigue characteristics for the gage. Many electrical resistance strain gages fail only when strain levels exceed 20,000 $\mu\varepsilon$ (2%) and one million cycles, making them useful for many static and dynamic applications.

An important characteristic of an electrical resistance strain gage is the gage factor (GF). This factor can be defined as follows:

$$\mathrm{GF} = (\Delta R/R)/\varepsilon \tag{3}$$

$\Delta R/R$ is the change in resistance divided by the resistance, which is then divided by the strain (ε) in order to calculate GF. Thus, GF is a dimensionless constant for any given gage and a measure of sensitivity. Its value is affected by the pattern of the foil, the size of the gage, the geometry of the gage, and the temperature at which measurements are made. Commercially available gages come with calculated GFs throughout their temperature range as well as values for their resistance. A reworking of equation 3 can be used then to calculate ε if $\Delta R/R$ is measured. The parameter $\Delta R/R$ can be easily measured through the use of a Wheatstone bridge circuit or potentiometer, as discussed below. Typical, GFs of electrical resistance strain gages are ~ 2, indicating the low sensitivity of these gages compared to semiconductor strain gages, which have high GFs. As a result, signals from electrical resistance strain gages require extensive conditioning to amplify the signal to a level acceptable to most recording devices.

A rosette strain gage is a combination of three electrical resistance strain gages configured adjacent to each other or stacked on top of one another on a single backing (Fig. 1). Rosettes can be used to calculate the direction and magnitude of the principal strains (minimum and maximum strain), and shear strain on a test material. They are available in two configurations: (1) a delta configuration with the principal axes of the three gages oriented at 60° apart and (2) a rectangular configuration with the principal axes of the three gages oriented at 45° apart. Another gage configuration, the T configuration (Fig. 1), consists of two gages at right angles to each other. This gage can also be used to calculate minimum strain, maximum strain, and shear strain if the directions of the minimum and maximum strains are already known and the elements of the gage are oriented along these axes on the test material.

A wide variety of insulation materials and bonding substances are available for use with electrical resistance strain gages. These include appropriate substances for constructing preparations for various conditions of temperature and moisture for static and dynamic applications.

Electrical Strain Gages: Elastic Resistance Type

Elastic resistance strain gages are specialized types of electrical resistance strain gages that are used in

cardiovascular and respiratory dimensional and plethysmographic (volume-measuring) determinations. The gages are made of a narrow silicone-rubber tubes, usually with an inside diameter of ~ 0.5 mm, and they range in length from 3 to 25 cm. The gages are filled with mercury or with an electrolyte or conductive paste and the ends are sealed with copper, silver, or platinum electrodes. When the gage is stretched, the diameter of the tube decreases and the length increases, increasing the resistance. These gages allow measurement of large-dimensional changes. They are accurate in the 10,000–100,000 $\mu\varepsilon$ range. Strains as high as 300,000 $\mu\varepsilon$ (30%) can be measured with distortion as small as 4%.

Wireless Strain Gage: Telemetry

Strain gages can be wireless, if coupled with telemetry transmitters and receivers. The elimination of the need for trailing wires from experimental objects makes possible data gathering from inconvenient or unsafe monitoring locations and it may reduce some of the noise of electrical interference (4,5).

SIGNAL PREPARATION AND AMPLIFICATION

Bridges

Two electric circuits, the potentiometer and the Wheatstone bridge, are commonly used to convert $\Delta R/R$ to a voltage that can be measured and used to determine ε. The Wheatstone bridge is the more widely used of the two types of circuits. There are some other variations; for further information, see the *Reading List*. A new circuit, the Anderson loop circuit, believe d to outperform the Wheatstone bridge, shows much future promise.

Potentiometer. Figure 2 illustrates a potentiometer circuit (3). This circuit consists of a voltage source (E_i), two resistors (R_1 and R_2), and an output voltage (E_0). A strain gage can take the place of either or both of these resistors and standard circuit equations can be used to compute the changes in E_0 and ΔE_0 and thus $\Delta R/R$. These equations are not given here; for further information, see the *Reading List*.

A major limitation of the potentiometer circuit is that the output voltage E_o is usually quite large compared to

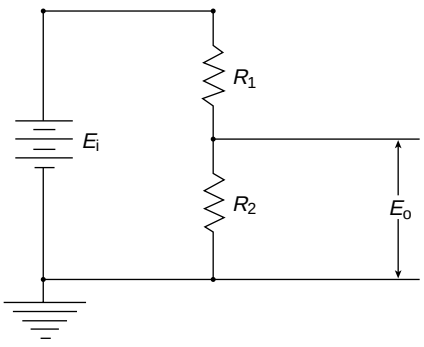


Figure 2. Diagram for a potentiometer circuit (3). Abbreviation: E_i-voltage source, R_1 and R_2-resistors, E_o-output voltage.

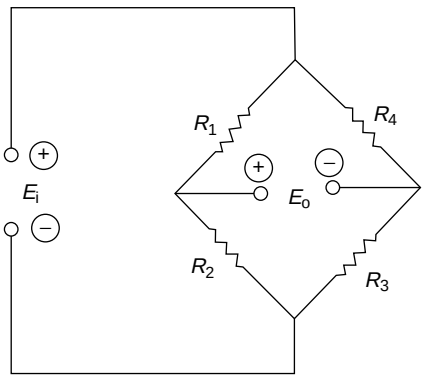


Figure 3. Diagram for a Wheatstone bridge circuit (3). Abbreviation: R_3 and R_4-resistors, others as in Fig. 2.

ΔE_o, making it difficult to obtain an accurate measurement of ΔE_o. This limits the usefulness of the circuit for static loading applications. However, a filter can be employed to block out E_o while allowing voltage pulses of ΔE_o to be measured. Such a filter enables the circuit to be useful in dynamic applications.

The Wheatstone Bridge. Figure 3 illustrates a Wheatstone bridge circuit (1). Note that the circuit contains four resistors. Strain gages can be used in the place of any combination of these resistors depending on the desired measurement. The advantage of this circuit over the potentiometer is that the bridge can be balanced so that the output voltage (E_o) is zero. The ratio of ΔE_o to the input voltage (E_i) can then be used to calculate the change in resistance in the strain gage ($\Delta R/R$) and the strain (ε). This feature makes the Wheatstone bridge circuit useful for the measurement of both static and dynamic strains. The circuit equations used to calculate strain differ depending on the configuration of gages in the Wheatstone bridge. For additional information on these equations, see the *Reading List*.

A limitation of the Wheatstone bridge circuit is that it is nonlinear when ΔR is $> 1\%$ (2). This is generally not a problem within the usual range of resistance changes of electrical resistance strain gages. However, large strains ($> 3000\,\mu\varepsilon$) measured with semiconductor gages may require the use of different circuitry (3). An additional advantage of the Wheatstone bridge is that is allows signals to be added or subtracted in multiple gage installations. In fact, sensitivity of a transducer can be quadrupled by using four active gages instead of one in the application. This can be accomplished, for example, by placing the two strain gages associated with the negative arms of the Wheatstone bridge on the compressive surface of the test specimen, and the two strain gages associated with the positive arms if the Wheatstone bridge on the tensile surface. Temperature compensation can also be achieved with Wheatstone bridges, through the used of a dummy gage, without sacrificing sensitivity.

The Anderson Loop Circuit. Figure 4 illustrates the topology of the Anderson loop circuit. This circuit consists

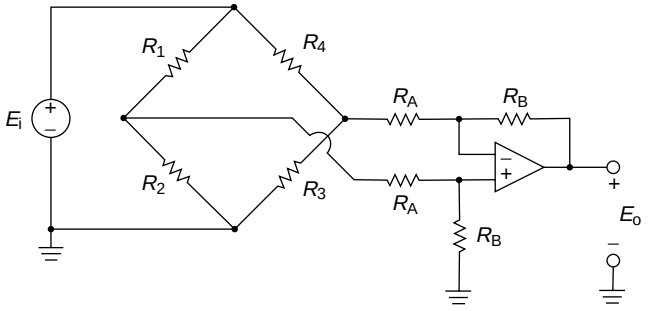


Figure 4. Diagram for a simple differential amplifier circuit for use with a Wheatstone Bridge (Modified from a drawing by O. Poroy ⟨www.engineering.uiowa.edu/∼bme080/lab/S05Lab7.pdf⟩). E_o is amplified by differential gain of sensors R_A and R_B compared to input resistance incurred by the source voltage (E_i).

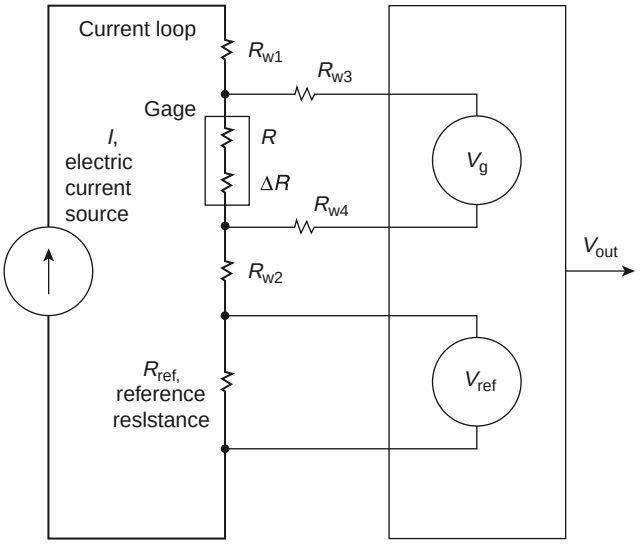


Figure 5. Diagram for an Anderson loop circuit (From Karl F. Anderson/Valid Measurements Inc., with permission)

of one or more sensors (strain gages) and typically one reference element (R_{ref}) connected in a series loop circuit. The same excitation current flows through each circuit element in the loop. The unique feature of this technology is the use of a subtractor function. In order to calculate strain from $\Delta R/R$, this function compares the voltage drop across various circuit elements in the loop, typically to determine the difference between each loop sensor voltage drop and the voltage drop across a reference element. Subtracting the voltage drop across two sensors yields a difference in $\Delta R/R$ between the two sensors.

Compared to the Wheatstone bridge, the Anderson loop circuit does not require four resistive elements. The loop provides a set of linear outputs that are twice that of the typical Wheatstone bridge for the same voltage across (and power dissipation in) each sensor. The loop topology overcomes some limitations normally encountered with the Wheatstone bridge, especially the detrimental effects of varying lead wire and connector resistances. These features open a vista for advanced transducer design. The impact of the Anderson loop circuit on the future of measurement and control is likely to be profound (6). This new technology was invented and developed in projects by NASA, who has placed it in the public domain. For more information, visit the Valid Measurements website, http://www.v-m.us.

Amplifiers

There are several good amplifiers and signal conditioners available commercially for use with strain gages (see *Further Reading* for some sources of information on specific amplifiers). An example for a simple differential amplifier circuit for use with a Wheatstone bridge is illustrated in Fig. 5, where the output voltage (E_o) is amplified by the including of sensors R_A and R_B. In general, the better amplifiers should have a variety of features including some of the following:

1. The ability to complete a variety of Wheatstone bridge circuits ranging up to a full bridge configuration with all resistors in the bridge replaced by strain gages. Dummy gages for a less than a full bridge configuration should be available for several of the standard resistances used in strain gages (including 120 and 350 Ω resistors).
2. A range of excitation voltages should be available for the Wheatstone bridge circuits.
3. The amplifier should be able to balance the Wheatstone bridge over a wide range of output voltages, although some computer-linked devices now automate balancing, making this task transparent to the user.
4. Variable gain to allow accurate readings of both large and small strains with a variety of gages.
5. Wide band operation, high input impedance, low noise level, and low temperature coefficient.
6. The ability to condition not only electrical resistance strain gages installed as part of Wheatstone bridge circuits but also potentiometer circuits, and the ability to condition and amplify output from piezoresistive (semiconductor) gages.
7. Additional features, such as digital readout and filters, depending on specific applications.

APPLICATIONS

Strain gages have enabled significant advances in many biological and biomedical fields. They have been used extensively in a wide variety of biomedical and clinical applications, chiefly to measure strain, stress, pressure, force, or displacement in or produced by living structures, or *in vitro* simulations. It is the aim of the following section to document several of these applications in order to give the reader a practical knowledge of the possibilities for using strain gages in basic biological and biomechanical research as well as in clinical sciences and medical diagnosis.

Biomechanical Research and Bone Health

An area of investigation with speculations that extend back into the nineteenth century and before is the study of the relationship between function and loading in bone. Of great mechanical interest are parameters of bone morphology and physiology, such as bone shape and size and the rate and amount of skeletal remodeling. In 1892, Wolff summarized much previous research in his Law of Bone Transformation that stated that every change in the function of a bone results in changes in trabecular orientation and in the external shape of the bone (7). Use of electrical resistance strain gages has led to much new research in this area.

Strain gages were first used in the 1940s to measure bone strain (8,9). Rosette configurations of electrical resistance strain gages have been most useful in these types of studies since they can provide information about changes in both the direction and magnitude of the strains on the skeleton.

A variety of experimental studies have used strain gages to assess how the craniofacial skeleton is loaded during normal oral and masticatory activities. Early studies used skulls with strain gages bonded to them to assess these loads, strain patterns and their links to bone morphology by simulation of muscular and occlusal forces (10). Later advances in strain gage technology and biological experimental techniques have enabled strains to be measured directly from bone *in vivo* in experimental animals (11–13; Fig. 6). For example, studies have used electrical resistance strain gages to assess how the craniofacial skeleton is deformed during normal oral and masticatory activities in primates (14,15). These studies cannot be repeated on humans because of their invasive nature. However, the animal studies have led to speculations and inferences about the biomechanics of mastication in humans, that can be explored through *in vitro* experiments using human cadaver specimens (9,16). Other interesting cranial studies have used strain gages to understand the behavior of open sutures and their effects on the function, growth, and adaptation of craniofacial skeleton (17,18). Moreover, patterns of craniofacial bone fracture have been related to bone loading and strain distribution determined with strain gage techniques (19–21).

Other studies have used strain gages in the postcranial skeleton to study loading patterns in long bones during locomotion (22). These techniques have also been applied to study skeletal loading during human locomotion (23) and to obtain better understanding of the function of prosthetic devices such as hip replacements (24). Other postcranial studies have used strain gages to measure tension in tendons or in tendons and bone simultaneously (25,26). In general, experimental studies have demonstrated a relationship between magnitudes and rates of loadings in bone and skeletal remodeling (22,27), although the precise nature of this relationship is the subject of considerable debate (27–29).

Coupled with techniques of measuring bone material properties and finite element analysis, the application of strain gages in studies of functional morphology contributes greatly in medical and biological fields (30–34). This research provides tremendous knowledge about bone morphology, adaptation, and modeling and remodeling during aging, normal function, and under abnormal conditions, and provides rationales for bone health maintenance and bone reconstruction.

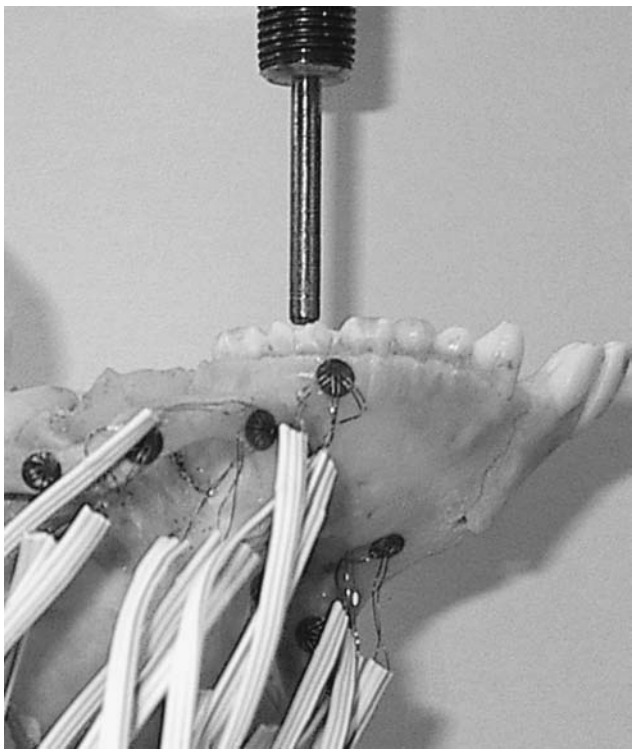


Figure 6. Experimental *In vitro* strain analysis in a skull of *Macaca fascicularis*. A couple of rectangular rosette strain gages were bonded to bone surface of craniofacial skeletons.

Dental and Craniofacial Research

In clinical dentistry and experimental craniofacial biology, measurements of occlusal forces have been used to give an overall assessment of functional capacities of the masticatory system (35–37). Such measurements are made with bite force transducers. These transducers have been used throughout much of the twentieth century and were originally mechanical devices. More recently, they have been constructed with electrical resistance strain gages or semiconductor gages. Figure 7 illustrates one such transducer constructed with electrical resistance strain gages (38). This transducer uses four strain gages in a full bridge configuration for an optimal voltage output. An individual would bite on the distal ends of the two bars to generate a change in resistance in the four gages. This change can then be measured and correlated with a specific biting force. Another feature of this transducer is that the gages are arranged in the circuit such that each upper and lower beam function as differential strain beams. In such beams, only the difference in strain between the two gages on the beam is measured. Thus force can be placed on the beam anywhere distal to the most distal gage and an identical reading will result. For a bite force transducer, this creates an advantage that the bite point dose not have

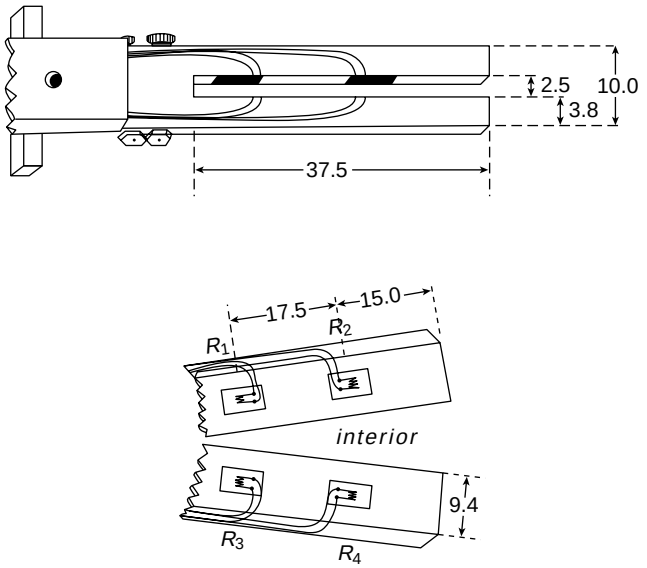


Figure 7. Diagram for a bite force transducer (38). This is an example of a force transducer constructed with four electrical resistance strain gages in a full bridge configuration. Each beam of the device is a differential strain beam. (a) Distal end of the device (missing the handle) illustrating the dimensions. (b) Inner surface of the beams showing the arrangement of the strain gages. Units in millimeters.

to be precisely controlled in order to obtain comparable readings.

Strain gages are used to study teeth and the surrounding alveolar bone during normal masticatory function, and in the presence of orthodontic and functional appliances. For example, strain gages are used to monitor the behavior of a dental prosthetic device (39–41), and its impact on the bone strain environment and eventual bone structure and strength (42). Strain gages are also used in studies of dental materials. A transducer has been constructed with electrical resistance gages to measure dimensional changes in dental amalgam during setting (43). These studies using strain gages provide rationales for orthodontic and prosthodontic treatments.

Physiological Research and Clinical Applications

Strain gages were used to study blood pressure (44), infracranial pressure (45), and cerebrospinal fluid pressure (46). The plethysmograph, made of elastic resistance strain gages, has been used diagnostically for several decades. For example, the venometer uses plethysmography to detect deep vein thrombosis through monitoring changes in certain muscle dimensions while venous outflow is occluded by a cuff placed around body parts (47,48).

A number of transducers constructed with electrical resistance strain gages are available to study the contractile properties of muscle. For example, strain gages were used to design quantitative assessment of neuromuscular deficits (49). Strain gages have also been used for many years to study cardiac muscle physiology *in vitro*. The Walton–Brodie strain gage arch has been used since the 1950s to measure forces exerted by the heart (50). For example, studies, have used the Walton–Brodie arch along with other strain gage transducers to examine the effect of halothane and enflurane on right ventricular performance (51), antiarrhythmic drugs on the heart rate (52), and biventricular mechanical alternans (53).

Elastic resistance or mercury strain gages have been used in cardiovascular and respiratory, and urological research. These studies have examined, for instance, changes in the volume of the larynx and movements of the rib cage and abdomen during respiration (54), the spontaneous contractions of trachea (55), and the change of bladder pressure with the use of catheters (56). Electrical resistance strain gages are also used to make transducers to study gut motility (57,58).

Strain gages were used in pharmacological studies to test physiological reactions to certain drugs (59), in psychophysiological research to examine how body physiology changes with emotion (60) and pain (61), and in sleep studies to evaluate sleep quality by recording eye movement during sleep (62,63) and to diagnose sleep related syndromes such as sleep apnea (64).

A unique application of electrical resistance strain gages was used to construct a transducer to evaluate erectile function in men by using strain gages to measure or monitor penile circumference and rigidity on various conditions (65,66). For example, a transducer called tonometer, measured force at which the penis buckles, giving a quantitative measurements of penile tumescence (67).

Sports Medicine

Strain gages have had been used in sports medicine, such as for performance evaluation, protection, and healing. For example, electric resistive strain gages have been used to construct instruments to measure hand strength (68), and to measure shoulder strength in football players, which is the most vulnerable body part in this sport (69). Strain gage transducers were used to document elongation of the human anterior cruciate ligament during various activities following a sprain. The transducer was temporarily placed in the ligament during arthroscopic surgery (70). Strain gages have also been used to evaluate sports equipments. For example, experiments were conducted to reveal how shoe gear can affect tibial strains in humans during dynamic loading, such as treadmill walking and free running while wearing various sport shoes (71).

Robotic Medicine

Within the concept of robotic surgery and telemedical care, medical practice is significantly enhanced with robotic systems incorporating tactile sensors made of essential strain gage elements. Examples include a tactile sensor system using a piezoelectric transducer to simulate the properties of the human hand for use as a surgical support instrument for conducting micro- or telesurgery (72), or a palpation probe to evaluate patterns of softness and elasticity of human skin (73).

BIBLIOGRAPHY

1. Dechow PC. Strain gage. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation. New York: Wiley; 1988. p 2715–2721.
2. Shull LC. Basic circuits. In: Hannah RL, Reed SE, editors. Strain Gage Users' Handbook. London: Elsevier; 1992. p 79–132.
3. Dorsey J. Semiconductor strain gages. In: Hannah RL, Reed SE, editors. Strain Gage Users' Handbook. London: Elsevier; 1992. p 365–448.
4. Chadwick EK, Nicol AC, Floyd S, Gray TG. A telemetry-based device to determine the force-displacement behaviour of materials in high impact loading situations. J Biomech 2000; 33:361–365.
5. Schatzker J, Sumner-Smith G, Hoare J, McBroom R. A telemetric system for the strain gauge determination of strain in bone *in vivo*. Arch Orthop Trauma Surg 1980;96:309–311.
6. Anderson KF. The Anderson Loop: NASA's successor to the Wheatstone bridge. ISA Transactions 1997;36:351–356.
7. Wolff J. Das Gesetz der transformation der knochen. Berlin; 1892.
8. Gurdjian ES, Lissner HR. Mechanism of head injury as studied by the cathode ray oscilloscope. Preliminary report. J Neurosurg 1944;1:393–399.
9. Evans FG. Methods of studying the biomechanical significance of bone form. Am J Phys Anthropol 1953;11:413–436.
10. Endo B. Experimental studies on the mechanical significance of the form of the human facial skeleton. J Fac Sci Univ Tokyo Sect 5 1966;3:(Pt I).
11. Cochron GVB. A method for direct recording of electromechanical data from skeletal bone in living animals. J Biomech 1974;7:563–565.
12. Wright TM, Hayes WC. Strain gage application on compact bone. J Biomech 1979;12:471–475.
13. Hylander WL, Bays R. An *in vivo* strain-gauge analysis of the squamosal-dentary joint reaction force during mastication and incisal biting in *Macaca mulatta* and Macaca fascicularis. Arch Oral Biol 1979;24:689–697.
14. Hylander WL. *In vivo* bone strain in the mandible of *Galago crassicaudatus*. Am J Phys Anthropol 1977;46:309–326.
15. Ross CF. *In vivo* function of the craniofacial haft: The interorbital "pillar". Am J Phys Anthropol 2001;116:108–139.
16. Daegling DJ, Hylander WL. Biomechanics of torsion in the human mandible. Am J Phys Anthropol 1998;105:73–87.
17. Behrents RG, Carlson DS, Abdelnour T. *In vivo* analysis of bone strain about the sagittal suture in *macaca mulatta* during masticatory movements. J Dent Res 1978;57:904–908.
18. Herring SW, Rafferty KL. Cranial and facial sutures: functional loading in relation to growth and morphology. In: Davidovitch Z, Mah J, editors. Biological Mechanisms of Tooth Eruption, Resorption and Replacement by Implants. Boston: Harvard Society for Advanced Orthodontics; 2000. p 269–276.
19. Ahmad F, et al. Strain gauge biomechanical evaluation of forces in orbital floor fractures. Br J Plast Surg 2003;56:3–9.
20. Ekenman I, et al. Local bone deformation at two predominant sites for stress fractures of the tibia: an *in vivo* study. Foot Ankle Int 1998;19:479–484.
21. Unnewehr M, et al. Fracture properties of the human mandible. Int J Legal Med 2003;117:326–330.
22. Lanyon LE. Analysis of surface bone strain in calcaneus of sheep during normal locomotion—strain analysis of calcaneus. J Biomech 1973;6:41–49.
23. Lanyon LE, Magee PT, Baggott DG. Relationship of functional stress and strain to the processes of bone remodeling—experimental-study on the sheep radius. J Biomech 1979;12:593–600.
24. Lanyon LE, et al. *In vivo* strain-measurements from bone and prosthesis following total hip-replacement—an experimental-study in sheep. J Bone Joint Surg 1981;63:989–1001.
25. Draganich LF, Reider B, Miller PR. An *in vitro* study of the Muller anterolateral femorotibial ligament tenodesis in the anterior cruciate ligament deficient knee. Am J Sports Med 1989;17:357–362.
26. Salmons S. *In vivo* tendon tension and bone strain measurement and correlation. J Biomech 1975;8:87.
27. Hart RT. Bone modeling and remodeling: theories and computations. In: Cowin SC, editor. Bone Mechanics Handbook. Boca Raton (FL): CRC Press; 2001. 31: p 1–42.
28. Martin RB, Burr DB, Sharkey NA. Skeletal Tissue Mechanics. New York: Springer; 1998.
29. Lanyon LE. et al. Osteocytes, strain detection, bone modeling and remodeling. Calcified Tissue International 1993;53: S102–S107.
30. Dechow PC, Hylander WL. Elastic properties and masticatory bone stress in the macaque mandible. Am J Phys Anthropol 2000;112:553–574.
31. Lertchirakarn V, Palamara JE, Messer HH. Finite element analysis and strain-gauge studies of vertical root fracture. J Endod 2003;29:529–534.
32. Rohlmann A, Mossner U, Bergmann G, Kolbel R. Finite-element-analysis and experimental investigation in a femur with hip endoprosthesis. J Biomech 1983;16:727–742.
33. Strait DS, et al. Modeling elastic properties in finite element analysis: How much precision is needed to produce an accurate model? Anat Rec 2005;283A:275–287.
34. Wang Q, et al. *In vitro* strain of monkey facial sutures. Am J Phys Anthropol 2005;126:S40–223.
35. Dechow PC, Carlson DS. Occlusal force after mandibular advancement in adult rhesus monkeys. J Oral Maxillofac Surg 1986;44:887–893.
36. Dechow PC, Carlson DS. Occlusal force and craniofacial biomechanics during growth in rhesus monkeys. Am J Phys Anthropol 1990;83:219–237.
37. Throckmorton GS, Ellis 3rd E, Buschang PH. Morphologic and biomechanical correlates with maximum bite forces in orthognathic surgery patients. J Oral Maxillofac Surg 2000;58:515–524.
38. Dechow PC, Carlson DS. A method of bite force measurement in primates. J Biomech 1983;16:797–802.
39. Throckmorton GS, Ellis 3rd E, Winkler AJ, Dechow PC. Bone strain following application of a rigid bone plate: An *in vitro* study in human mandibles. J Oral Maxillofac Surg 1992;50:1066–1074.
40. Dechow PC, Ellis 3rd E, Throckmorton GS. Structural properties of mandibular bone following application of a bone plate. J Oral Maxillofac Surg 1995;53:1044–1051.
41. Watanabe F, et al. Analysis of stress distribution in a screw-retained implant prosthesis. Int J Oral Maxillofac Implants 2000;15:209–218.
42. Kim YH, Kim JS, Cho SH. Strain distribution in the proximal human femur. An *in vitro* comparison in the intact femur and after insertion of reference and experimental femoral stems. J Bone Joint Surg Br 2001;83:295–301.
43. Lemaitre L, Moors M, Vanpeteghem AP. Method for the measurement of dimensional change of dental amalgam. J Biomed Mat Res 1979;13:887–892.
44. Nielsen PE, Rasmussen SM. Indirect measurement of systolic blood pressure by strain gauge technique at finger, ankle and toe in diabetic patients without symptoms of occlusive arterial disease. Diabetologia 1973;9:25–29.
45. Fryer TB, Silverberg GD, Corbin SD, Schmidt G. Telemetry of intracranial-pressure. Biotelemetry 1978;5:46.

46. Brosnan RJ, et al. Effects of ventilation and isoflurane end-tidal concentration on intracranial and cerebral perfusion pressures in horses. Am J Vet Res 2003;64:21–25.
47. Cooperman M, et al. Detection of deep venous thrombosis by impedance plethysmography. Am J Surg 1979;137:252–254.
48. Maskell NA, et al. The use of automated strain gauge plethysmography in the diagnosis of deep vein thrombosis. Br J Radiol 2002;75:648–665.
49. Andres PL. et al. Quantitative assessment of neuromuscular deficit in ALS. Neurol Clin 1987;5:125–141.
50. de V Cotton M. Circulartory changes affecting measurement of heart force in situ with strain gage arches. Am J Physiol 1953;174:365–370.
51. Mote PS, Pruett JK, Gramling ZW. Effects of halothane and enflurane on right ventricular performance in hearts of dogs anesthetized with pentobarbital sodium. Anesthesiology 1983;58:53–60.
52. Sarel O, Hasin Y, Rogel S. Myocardial conduction time and antiarrhythmic drugs. J Electrocardiol 1981;14:261–266.
53. Hasin Y, Sarel O, Rogel S. Electrical and mechanical response in biventricular mechanical alternans. Arch Int Physiol Biochim 1979;87:19–28.
54. Cavallo SA, Baken RJ. Prephonatory laryngeal and chest wall dynamics. J Speech Hearing Res 1985;28:79–87.
55. Souhrada JF, Dickey DW. Mechanical activities of trachea as measured *in vitro* and *in vivo*. Respir Physiol 1976;26:27–40.
56. Flack FC, James ED. Case study using simultaneous bladder pressure and urine loss measurements. Urol Int 1975; 30:103–108.
57. Johnson CP, et al. Effects of intestinal transplantation on postprandial motility and regulation of intestinal transit. Surgery 2001;129:6–14.
58. Pascaud XB, Genton MJ, Bass P. A miniature transducer for recording intestinal motility in unrestrained chronic rats. Am J Physiol 1978;235:E532–538.
59. Ueda S, Wada A, Umemura S. Methodological validity and feasibility of the nitric oxide clamp technique for nitric oxide research in human resistant vessels. Hypertens Res 2001;27:351–357.
60. Bigelow N, et al. A preliminary report on a study of a correlation between emotional reactions and peripheral blood circulation, using a strain gauge plethysmograph. Psychiatr Q 1955;29:193–202.
61. Forgione AG, Barber TX. A strain gauge pain stimulator. Psychophysiology 1971;8:102–106.
62. Coakley D, Williams R, Morris J. Minute eye movement during sleep. Electroencephalogr Clin Neurophysiol 1979;47:126–131.
63. Mamelak A, Hobson JA. Nightcap: A home-based sleep monitoring system. Sleep 1989;12:157–166.
64. Miyazaki S, et al. Using an air-pad sensor for the diagnosis of sleep apnea: A trial study. Psychiat Clin Neurosci 2002;56: 315–316.
65. Janssen E, Vissenberg M, Visser S, Everaerd W. An *in vivo* comparison of two circumferential penile strain gauges: The introduction of a new calibration method. Psychophysiology 1997;34:717–20.
66. Kiely ME, Thavundayil JX, Lal S. Effect of blood sampling on apomorphine-induced penile tumescence in erectile impotence: A case report. J Psychiat Neurosci 1995;20:233–235.
67. Hahn PM, Leder R. Quantification of penile buckling force. Sleep 1980;3:95–97.
68. An KN, Chao EY, Askew LJ. Hand strength measurement instruments. Arch Phys Med Rehabil 1980;61:366–368.
69. Burnham RS, Bell G, Olenik L, Reid DC. Shoulder abduction strength measurement in football players: Reliability and validity of two field tests. Clin J Sport Med 1995;5:90–94.
70. Henning CE, Lynch MA, Glick Jr KR. An *in vivo* strain gage study of elongation of the anterior cruciate ligament. Am J Sports Med 1985;13:22–26.
71. Milgrom C, et al. The effect of shoe gear on human tibial strains recorded during dynamic loading: A pilot study. Foot Ankle Int 1996;17:667–671.
72. Omata S, Murayama Y, Constantinou CE. Development of a novel surgical support instrument and virtual system incorporating new tactile sensor technology. Stud Health Technol Inform 2004;98:288–290.
73. Iida I, Noro K. An analysis of the reduction of elasticity on the ageing of human skin and the recovering effect of a facial massage. Ergonomics 1995;38:1921–1931.

Further Reading

These books contain much useful information on strain gages, stress and strain analysis, and instrumentation for strain gages. Internet searches using strain gage or strain gauge as key words also yield a wide variety of information about bridges, strain gages, associated products, and instrumentations, manufacturers, technique supports, and training sessions.

Dally JW, Riley WF. Experimental Stress Analysis. 3rd ed. New York: McGraw-Hill; 1991.
Hannah RL, Reed SE, editors. Strain Gage Users' Handbook. London: Elsevier; 1992.
Khan AS, Wang X. Strain Measurements and Stress Analysis. Upper Saddle River (NJ): Prentice Hall; 2000.
Kost GJ, editor. Handbook of Clinical Automation, Robotics, and Optimization. New York: Wiley; 1996.
Perez R. Design of Medical Electronic Devices. London: Academic Press; 2002.
Prutchi D, Norris M. Design and Development of Medical Electronic Instrumentation: A Practical Perspective of the Design, Construction, and Test of Medical Devices. Hoboken, NJ: Wiley-Interscience; 2004.
Webster JG, editor. Medical Instrumentation: Application and Design. 3rd ed. New York: Wiley; 1998.
Window AL. Strain Gauge Technology. 2nd ed. Burlington (MA): Elsevier Science; 1992.
Ballantyne GH, Marescaux J, Giulianotti PC, Primer of Robotic & Telerobotic Surgery. Hagerstown MD: Lippincott Williams & Wilkins; 2004.

See also BIOMATERIALS, TESTING AND STRUCTURAL PROPERTIES OF; BONE AND TEETH, PROPERTIES OF; LIGAMENT AND TENDON, PROPERTIES OF; REHABILITATION AND MUSCLE TESTING; TOOTH AND JAW, BIOMECHANICS OF.

STRESS TESTING, CARDIOVASCULAR. See EXERCISE STRESS TESTING.

SURFACE PROPERTIES OF BIOMATERIALS. See BIOMATERIALS, SURFACE PROPERTIES OF.

SURGERY, FROZEN. See CRYOSURGERY.

SURGERY, INTRAUTERINE. See INTRAUTERINE SURGICAL TECHNIQUES.

SURGERY, MINIMALLY INVASIVE. See MINIMALLY INVASIVE SURGERY.

SURGERY, STEREOTACTIC. See STEREOTACTIC SURGERY.

SWEAT TEST FOR CYSTIC FIBROSIS. See CYSTIC FIBROSIS SWEAT TEST.

SYSTEMIC HYPERTHERMIA. See HYPERTHERMIA, SYSTEMIC

T

TACTILE STIMULATION

CHRISTOPHER J. POLETTO
National Institutes of Health

INTRODUCTION

Definitions

In the most general terms, tactile stimulation is the deliberate elicitation of any of a range of sensations perceived through the sense of touch. The stimulation can be delivered by any means and to any part of the body where touch can be felt. Most commonly, tactile stimulation is applied to the skin, usually to the fingertips, abdomen, or forearms, but tactile stimulation is also used on the tongue as a possible communications tool and at the back of the throat during therapy in people with dysphagia. Clinically relevant tactile stimulation can be as simple as a caregiver stroking a baby to encourage healthy development or involve of thousands of sophisticated actuators in arrays working in concert to present a virtual tactile environment.

Tactile perception should be distinguished from kinesthetic and haptic perception. Tactile perceptions include temperature, skin curvature and stretch, vibration, slip, pressure and local contact force. Kinesthesia is the perception of the relative position and movement of our body parts (proprioception), as well as the sensation of muscular effort exerted while touching or manipulating objects. Although M. Dessoir (1892) originally defined haptic (*haptik* in German, modified from the Greek word *haptikos*) to mean "the study of touch and tactile sensations, especially as a means of communication" (Oxford English Dictionary), the term haptic is applied much more broadly now. Haptic perceptions combine tactile and kinesthetic perceptions to provide a sense of environmental or object properties such as shape (1). Many devices referred to as "tactile displays", especially those intended for virtual reality or telepresence applications, actually combine tactile and kinesthetic feedback, but this article is concerned only with devices that deliver tactile stimulation. The focus will be on the mechanical (vibrotactile and shape displays) and electrical (electrotactile) stimulators that elicit sensations of vibration, pressure, and local contact force, because these types of stimulators are most often used in current tactile stimulation applications.

Applications

Telepresence and telerobotic technologies are emerging as important fields with great potential for use in biomedical applications such as the control of surgical robots. These applications require mechanisms for the feedback of information to the human operator from a set of remote sensors. This feedback has been visual and auditory in nature, but developing technologies are permitting the use of certain forms of haptic feedback such as force reflection to communicate information about a manipulated object's compliance, viscosity, mass, size, and gross shape. Researchers have made less progress toward communicating tactile information such as surface texture, fine shape, roughness, slip, vibration or temperature. These later object properties could potentially be important in helping the user distinguish between morphologically similar objects or, in the case of surgical or medical imaging applications, to assess tissue type or of the degree and extent of tissue damage (2,3). Indeed, in cases where visual information is not available, tactile information may form a user's only basis for decision making.

Other applications for tactile stimulation include sensory substitution or augmentation systems such as those for the blind or deaf. For example, using a computer poses a particular challenge for the visually impaired. Textual information can be read aloud by the computer, but as graphical computer interfaces become more and more ubiquitous, the blind user is placed at a greater disadvantage. One solution is to represent the graphical information through a tactile display. Further applications are discussed below in the section on stimulation techniques.

TACTILE PHYSIOLOGY

Understanding how tactile stimulation techniques function requires a basic understanding of the physiological mechanisms that mediate our sense of touch. When we touch an object, the distribution of contact forces deforms our skin. The temporal and spatial distribution of the deformation determines which mechanoreceptive nerve terminals are excited. Each mechanoreceptive nerve fiber terminates as one or more dendrites. The dendrites have stretch-gated ion channels that allow ions to flow and depolarize the dendrite when their membranes are stretched or deformed. If the depolarization is sufficient, nearby voltage gated channels initiate an action potential that travels up the nerve to the central nervous system. The particular way that the deformation of the skin leads to bending or stretching of the dendrites varies from one type of mechanoreceptor to another and is due to the structure of the sensory receptor at the end of the fiber. The firing frequencies of the mechanoreceptive nerve fibers are then decoded by the central nervous system to ultimately give rise to our sense of touch.

The skin is made up of three distinct layers: the epidermis, dermis, and subcutaneous fat. Made up of primarily keratinocytes, the epidermis is the outermost layer and serves as a protective interface with the outside world. The outermost portion of the epidermis is the *stratum corneum*, which is made up of dead, flat cells that shed and are replenished about every 2 weeks. The epidermis does not contain any blood vessels and is dependant on the deeper skin layers for its oxygen and nutrients. A very thin membrane, the basement membrane, attaches the

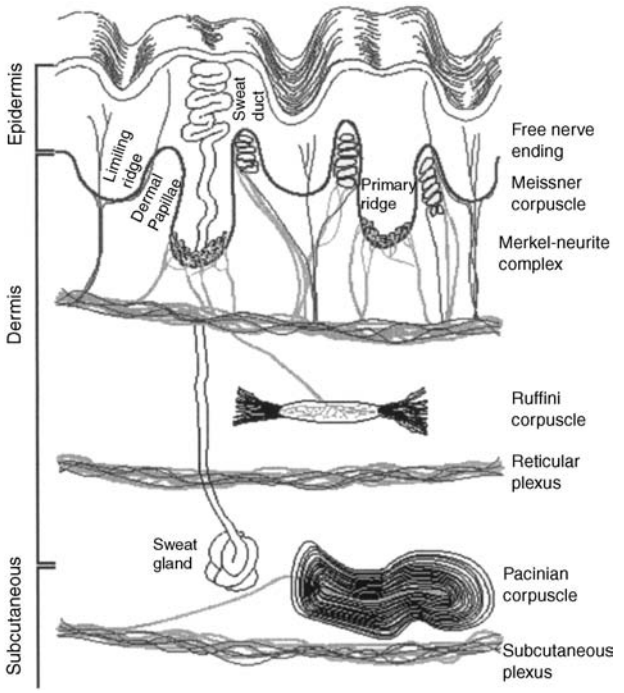


Figure 1. Mechanoreceptors in skin. Mechanoreceptive innervation of human hairless (glabrous) skin. (Diagram provided by Kenneth Johnson.)

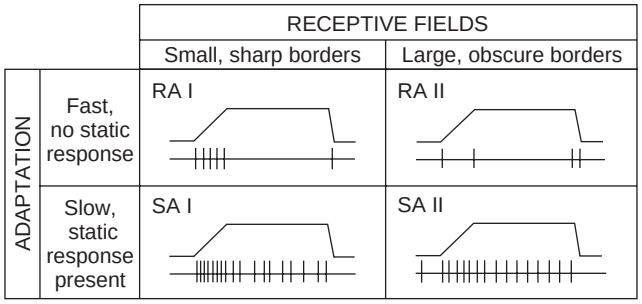


Figure 2. Response of RA and SA fibers. Responses of typical rapidly adapting (RA) and slowly adapting (SA) mechanoreceptors with large or small receptive fields to ramped and sustained indentation. The onset of the indentation stimulates rapid volleys of action potentials in both sets of mechanoreceptors, but the RAs adapt and cease firing after a few milliseconds. The SAs continue firing until indentation offset, which may trigger anther burst from the RAs. (Figure redrawn from Ref. 124.)

epidermis firmly, though not rigidly, to the second layer, the dermis. The dermis contains blood vessels, nerves, hair roots, and sweat glands. Below the dermis lies a layer of fat, the subcutaneous fat, which serves as a cushion between the skin and the underlying muscle and bone. The depth of this layer differs from one person to another.

Four major types of mechanoreceptors are present in human glabrous skin (the hairless skin of the lips, soles, palms, and finger pads) that respond to touch or vibration: Pacinian corpuscles (PC or RAII), Meissner corpuscles (rapidly adapting or RAI), Merkel endings (slowly adapting or SA I), and Ruffini endings (slowly adapting or SA II) (Fig. 1). In hairy skin, the Meissner corpuscles are replaced by hair receptors with similar mechanoreceptive properties. Only a very basic overview of mechanoreceptors is provided here; see (4–6) for reviews of mechanoreceptor morphology. A recent review of the roles that mechanoreceptors play in perception can be found in Ref. 7.

Each type of mechanoreceptor is tuned so that its threshold for excitation is lowest for a particular range of vibration frequencies, although there is considerable overlap between receptor types. The receptors that respond best to higher frequency stimulation are referred to as rapidly adapting, while those that respond best to lower frequencies are called slowly adapting. A sudden but sustained skin indentation initially results in rapid firing of all the affected mechanoreceptors. The rapidly adapting ones soon cease firing (however, they may fire again at stimulus offset), whereas the slowly adapting mechanoreceptors continue firing (albeit at a lower rate) until contact is broken (Fig. 2).

The area in which stimulation leads to a response by a particular mechanoreceptive neuron is known as the receptive field of that neuron. The skin tissue acts as a low pass spatial filter so that the deeper a mechanoreceptor lies in the skin, the larger its receptive field. The receptive field also depends on how many receptors a mechanoreceptive neuron innervates and how widely they are dispersed. Mechanoreceptors with small receptive fields are more suitable for fine spatial discrimination than those with large receptive fields. Mechanoreceptors with small receptive fields are referred to as type I mechanoreceptors and those with large receptive fields are called type II.

The frequency tuning of a mechanoreceptive fiber (i.e., whether it is rapidly or slowly adapting) is due to its mechanical structure. For example, the Pacinian corpuscle receptor is shaped like an onion with multiple layers wrapped around the dendrite. When the skin is indented, the onion-shaped receptor is compressed. With sufficient pressure to deform the corpuscle, parts of the neural membrane are stretched and action potentials are stimulated. The Pacinian rapidly adapts to constant pressure because with maintained pressure, the gel-filled layers of the corpuscle begin to slide across each other in a way that relieves the stress on the dendrite and stops the generation of action potentials. A vibratory stimulus, on the other hand, repeatedly deforms the corpuscle, eliciting continued neuronal firing. The Pacinians are sensitive to a wide range of frequencies (50–1000 Hz), but are exquisitely sensitive to vibrations between 250 and 350 Hz, requiring only submicron surface displacements for excitation. The Pacinians lie deep in the dermis, so pressure anywhere over a fair amount of skin area can compress it enough to affect the dendrite, providing it with a large receptive field. Isolated stimulation of a particular Pacinian nerve fiber leads to a sensation of diffuse vibration. Even though their deep position renders them unsuitable for fine stimulus localization, the Pacinians aid in the perception of surface texture by detecting the vibrations elicited when the skin is brushed over a rough object (8). Pacinian corpuscles are also thought to play a role in tool use, specifically in our ability to attribute actions to the end of the tools we hold, by sensing the vibrations transmitted through the tool.

The ultrastructure of the Meissner corpuscle is more complicated than that of the Pacinian corpuscle, but Meissners are also rapidly adapting. The Meissners respond best to vibratory stimuli with frequencies between 30 and 70 Hz, but also respond to frequencies as low as 3 Hz. The Meissner corpuscles are located just below the dermal–epidermal boundary and have smaller receptive fields than Pacinians. Meissners are considered the primary mediators of the sensations of light touch and flutter while the Pacinians are considered the primary mediators of high frequency vibration and deep pressure.

The primary types of slowly adapting mechanoreceptors are the Merkel complexes and the Ruffini corpuscles. Merkels have a receptive field diameter of 3–4 mm, and a frequency range of 2–32 Hz (9). The Merkel sensory neurons have nonencapsulated nerve endings (neurites) that form a complex arrangement with disk-shaped Merkel cells in the basal layer of the epidermis. There is conflicting data about just how this structural arrangement accounts for the mechanoreceptive properties of the receptors, but there is general agreement that the Merkel–neurite complexes mediate the perception of steady skin indentation and their population response is thought to account for our perceptions of form and texture (7).

Ruffini corpuscles are relatively large spindle-shaped structures that are rooted in the connective tissue of the dermis. They have been studied less than other mechanoreceptors because we derive much of our understanding about mechanoreceptor physiology from studies performed on monkeys and Ruffini corpuscles have never been observed in neurophysiological studies of monkey hands. Additionally, it is not clear which perceptions should be attributed to them, because microstimulation of Ruffini fibers produces no sensation (10), and elicits no cortical somatosensory evoked potentials (11). There is some evidence, however, that Ruffini endings may participate in tactile sensations as SAII receptors (12). Currently, it is believed that Ruffinis mediate perceptions of directional skin stretch. Directional stretch sensitivity would allow them to participate in the perception of hand and finger position as well the perception of the direction of object motion or force (13,14).

In addition to the mechanoreceptors discussed above, the skin contains other sensory fibers that might be important for tactile stimulation that are outside the scope of this article, but deserve mention. The previously mentioned cutaneous and subcutaneous mechanoreceptors have large, fast, myelinated (Aβ) fibers, but there are also mechanically sensitive cutaneous free nerve endings with small, slower myelinated (Aδ) fibers that respond to strong mechanical stimulation, especially by sharp objects. These mechanoreceptive *nociceptors* (receptors that respond to potentially damaging stimuli) mediate primarily pricking pain, but some also respond to thermal stimuli. Additionally, there are even smaller and slower, unmyelinated (C) fibers present that respond to noxious mechanical and/or thermal stimuli and are primarily responsible for sensations of burning pain. Although almost all sensory C fibers are traditionally considered to mediate pain, there is growing evidence that some mediate touch that is specifically pleasant and emotionally salient, such as from stroking by another human (15,16). It is unknown what role C-fibers will play in tactile stimulation devices in the future.

TACTILE STIMULATION TECHNIQUES AND APPLICATIONS

The two main approaches used to artificially stimulate the tactile senses are based on exciting the receptors through mechanical perturbation of the skin or electrically exciting the mechanoreceptive nerve fibers that innervate the receptors. In both approaches, the goal is to manipulate the tactile information through variations in the four tactile primitives that code tactile information: intensity, frequency, spatial pattern, and temporal pattern (17). In addition to these classical primitives, there is evidence for a multidimensional electrotactile primitive that could be termed "color" or "quality", of which frequency may be one dimension (18,19).

Efforts have also been reported to create tactile displays that illicit sensations of warmth or cold using thermal elements such as Peltier cells (20–22). Although interesting and potentially useful, these thermal displays are beyond the scope of this article.

Mechanical

Mechanical tactile stimulators predominately fall into one of two classes: vibrotactile or shape displays (Fig. 3). Vibrotactile displays use contactors that vibrate at a particular, usually fixed, frequency. Shape displays provide small-scale indentation, pressure, or shear profiles to the skin, often using an array of pins that can each be pressed into or withdrawn from the skin.

A useful vibrotactile display can be as simple as the early Tactaid (Audiological Engineering Corporation, Somerville, MA). The Tactaid is a single vibrator that is affixed to the chest, back of the neck, or wrist that transduces acoustic signals into vibration to convey certain speech information (the modern Tactaid VII uses an array of seven vibrators in a line). Single vibrators are also becoming common on force feedback devices such as joysticks and robotic manipulators. They are even becoming popular for home computer devices such as Logitech's iFeel MouseMan computer mouse.

Relatively simple vibrotactile stimulators also offer clinical benefit for individuals with impaired tactile sensation. For example, Collins et al. (23) recently demonstrated that three small vibrating *tactors* (tactile stimulators), imbedded in gel-based shoe insoles, helped to maintain balance and improve postural sway during quiet standing in healthy young and elderly individuals, as well as patients with stroke, and patients with diabetic neuropathy (24,25). This application is particularly interesting because the stimuli were essentially random noise delivered at intensities *below* sensory threshold.

The subthreshold noise enhances the user's ability to detect useful sensory signals from the sole of the foot through the principle of *stochastic resonance.* This principle, found to apply in many nonlinear physical and biological systems, describes systems in which the ability to detect a signal in the presence of noise is actually enhanced by adding a critical amount of noise. Collins contends that

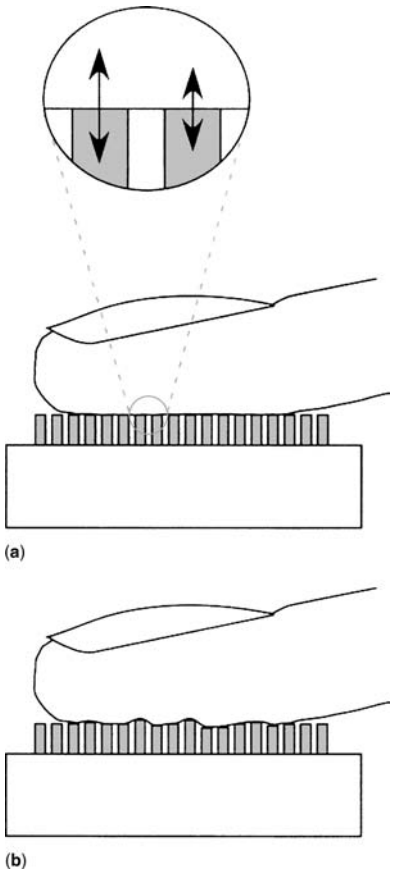


Figure 3. Vibrotactile versus shape displays. Vibrotactile displays (**a**) use pins that vibrate around a common set point. Larger oscillations are used to produce sensations of higher intensity (top inset). Shape, or static, displays (**b**) present more naturalistic surface contours and use larger indentations to encode regions of higher intensity.

such noise-based devices could ameliorate age-related impairments in balance control. Similar devices using electrotactile stimulation are offered by Afferent, a company Collins helped found.

A small number of vibrotactile elements can also be useful in teleoperation applications. For example, a single vibrating element encoding manipulator force was found to improve the ease of use of a deep sea remote manipulator (26,27). Similarly, Debus et al. (28) used a cylindrical handle with four embedded vibrating elements to demonstrate that force feedback vibrations can significantly improve a teleoperated force control task by reducing the mean force errors by 35% compared to visual feedback alone.

Vibrotactile displays that utilize dense arrays of tactors can communicate more information than single oscillators. Tactile stimulators for use outside the laboratory were initially developed as sensory substitution aids for the sight and hearing impaired. One of the first tactile arrays used vibrating contacts on the users back to communicate visual information to blind users (29), an effect called "Tactile Television". Another example of an early vibrotactile array for sensory substitution was the Optacon (OPtical to TActile CONverter), a portable reading device for the blind (30) that was originally developed in 1966 by John Linvill and JC Bliss to assist Linvill's vision impaired daughter. The Optacon allowed the visually impaired to read printed material by feeling a tactile version of a visual image. A small hand-held camera scanned pictures or text in any document. The image was presented to the user via an array of vibrating metal rods that stimulated an index finger. The stimulator consisted of an array of 144 (24 rows $\times$ 6 columns) of vibrating metal pins mounted on top of the controller. Each of the vibrating pins was driven by a piezoelectric bimorph reed actuator. The device was a technological success; many blind individuals reported that it provide increased autonomy. Debbie Stein, the First Vice President of the National Federation of the Blind of Illinois, and a long-time Optacon user stated (31) "The Optacon has given blind people a level of autonomy and flexibility unparalleled in history." As technically successful as the compact, lightweight, and highly portable Optacon was, it was not a financial success so Telesensory, Inc. discontinued manufacturing the Optacon in 1996, citing high cost and low demand.

From the time of the first vibrotactile displays, power consumption has been recognized as an important design factor. Higher power consumption requires larger and more expensive electronics, larger and heavier batteries, increased heat production, and less portability overall. Matching the display parameters (e.g., frequency, amplitude, tactor size, tactor density) to the mechanoreceptor characteristics (e.g., frequency response, sensitivity, receptive field) allows reduction of the power requirements, and many theoretical studies have pursued this goal (32–34). For example, the frequency at which the pins in a vibrotactile display oscillate is often selected to maximize the amplitude of the perceived signal (typically 200–300 Hz) while minimizing the power requirements of the display. The standard psychophysical term used to refer to the perceived amplitude of a vibratory stimulus is *Loudness*. For a given amount of actuator power, the loudness of a sinusoidal stimulus is maximized at $\sim$250 Hz (35–37), the frequency Pacinian corpuscles are most sensitive to. Accordingly, several vibrotactile stimulators use sinusoidal vibrations at or near this frequency (200 Hz in the case of the Optacon). An average young, healthy user can feel 250 Hz vibrations <1 μm in amplitude (38).

The use of rectangular pulses instead of sinusoidal waveforms can result in lower power requirement, especially if short pulses are used so that multiple pins can share the same driver stages. The electronics required to generate rectangular pulses can also be much simpler, potentially as simple as a single transistor operating as a switch (34), further reducing the size and cost of the display. A recent study that measured power consumption experimentally for electromechanical and piezoelectric vibrotactile actuators for sinusoidal and rectangular pulse waveforms (39) found that power consumption for the piezoelectric reed bimorph transducers was two orders of magnitude lower than for the electromechanical transducers. The piezoelectric transducers were more efficient despite their higher voltage requirements, because they required almost no current. Curiously, within the electromechanical transducers, the least power required to reach threshold was using short (0.7 ms low duty cycle)

rectangular pulses delivered at a rate of 25/s, but for the piezoelectric transducers the most efficient waveform was sinusoidal at 250 Hz. Clearly, the most efficient stimulus waveform depends on the type of actuators chosen for the display.

Shape displays tend to have low temporal resolution (unless constructed of piezoelectric actuators), but potentially very high spatial resolution. Static shape displays are perhaps the original example of tactile displays with the primary example being Braille. Blinded at the age of three, Louis Braille (1809–1852), invented the familiar tactile shape display reading system that bears his name. Individual letters, numbers, or common combinations of letters are represented by combinations of six or eight raised dots.

Ideally, a dynamic (changing) tactile shape display could be developed that would present patterns that would be indistinguishable to the user from direct contact with the original (or virtual) surface. The display must therefore deform the skin in the same way [or at least produce the same strain patterns at the receptor level (40–41)] that the represented surface would. The slow development of these dynamic fingertip tactile displays is due to the numerous unknown factors that characterize the sense of touch.

Designing an optimal dynamic tactile display requires precise knowledge of the biomechanics of the skin, which is both tough and deformable at the same time (42). When scanning a texture, we use contact forces ranging between 0.3 and 4.5 N and a scanning speed between 1 and 25 cm/s with an average of 2 cm/s (43). Theoretically, to have full control of the surface stresses, the ideal tactile display system needs an infinitely dense array of actuators, each with three degrees of freedom and infinite stiffness. Nearly ideal function would also require a display with actuators capable of delivering at least 50 N/cm^2 pressure, able to indent the skin at least 4 mm (with 10% resolution), an actuator density of 1/mm^2, and at least 50 Hz refresh bandwidth, requiring a power density of 10 W/cm^2(44). In fact, it may even be advantageous to provide a refresh rate that matches the highest frequency response of the RA receptors, in the kilohertz. The optimal display would also be small enough for use with several fingers at once over a large range of positions and orientations. Due to compromises that have to be made between power, bandwidth, manufacturability, cost, size, actuator stiffness, ease of maintenance, and other factors, no tactile display currently comes close to satisfying these requirements.

One goal of some shape displays, such as those for telesurgery or telerobotics, is to communicate information about the compliance of the manipulated (or virtual) object. Human discriminability of softness (compliance) of objects depends on the object having a deformable surface (45–46). For deformable surfaces, the spatial pressure distribution within the contact region depends on object compliance, so that information from cutaneous mechanoreceptors is sufficient for the user to gauge subtle differences in compliance. When the surface is rigid, however, we require kinesthetic information for discrimination, and our ability to gauge softness is much poorer than that for objects with deformable surfaces. It is likely that the spatiotemporal variation of pressure on the skin (or, equivalently the skin displacement and its derivatives) form the basis for the perception of softness of compressible objects. Consequently, a shape display that aims to communicate compliance information must itself be deformable (compliant) (45). This compliance might be achieved through either passive or active means.

Designers of shape displays have generally taken one of two approaches; (1) control the force the actuators exert on the skin (in the direction perpendicular to the skin), or (2) control the displacement (again, perpendicular) that the actuators impose on the skin. A smaller number have controlled the lateral stresses or strains applied to the skin. Almost all shape displays have been designed for the fingertips. Shape display designs have used solenoids (47), electrostatics (48–49), voice coil actuators (50), shape memory alloys (51–52), pneumatics (44,53), RC servomotors (54), MEMS (55), and even air jets (56). A good summary of the technologies and their relative merits can be found in Ref. (17), with an updated version currently available at http://www.cim.mcgill.ca/~jay/index_files/research_files/actuators.htm.

Electrotactile

The use of mechanical systems to communicate detailed tactile information is limited by the need for dense arrays of end effectors with sufficient mechanical compliance to overcome the stiffness of the skin and deeper tissues. Such stimulators require bulky actuators that are difficult to arrange in the dense arrays required for texture communication. For these reasons, researchers have turned to electrical stimulation of the tactile senses. Electrical stimulation that aims to produce a localized tactile perception is referred to as "electrotactile" or "electrocutaneous" stimulation, whereas "transcutaneous" stimulation describes the more generalized stimulation of nerve bundles [e.g., transcutaneous electrical nerve stimulators (TENS)]. Electrotactile stimulation has the advantages that dense arrays of electrodes can be fabricated and that the stimulator itself can be remotely located so it does not limit the spacing or rigid presentation of the electrodes. Electrotactile stimulation displays also typically require much less power than vibrotactile devices.

In contrast to mechanical stimulators, which deform the receptor at the end of a mechanoreceptive nerve fiber, electrical stimulators depolarize the membrane of the mechanoreceptive nerve fibers directly by passing current through the skin from one electrode (or group of electrodes) to another. The current passing through the skin produces an electric field that extends some distance into the skin. This electric field creates currents within the nerve fibers that depolarize the neural membrane. If this depolarization is sufficient, the neuron fires an action potential that propagates to the central nervous system. In general, the smaller the stimulating electrodes and the closer they are spaced, the more superficial the electric field will be. For a given current, the strength of the electric field beneath an electrode varies inversely with the surface area of the electrode so stimulation will always occur at the smaller of the two electrodes. The usual practice is to use a large

remote electrode or a large surrounding electrode (or group of electrodes) as the current return and a smaller electrode as the stimulating electrode. The negatively charged electrode is referred to as the *cathode* and the positively charged one is the *anode*. Cathodic stimulation is therefore when the negatively charged electrode is the stimulating electrode, whereas anodic stimulation uses the positively charged electrode for stimulation. Cathodic stimulation is the most conventional and is best for exciting nerve fibers that are passing by the electrode, while anodic stimulation is best for exciting fibers that terminate in the vicinity of the electrode. The stimulating phase of the waveform is usually followed by a charge-recovery phase of opposite polarity and much lower amplitude. The reason for the charge-recovery phase is to avoid electrode corrosion and tissue damage by reversing as much as possible the chemical reactions that occur at the electrodes.

Electrocutaneous and transcutaneous electrical stimulation are used in a wide variety of applications including TENS units for pain control and sensory substitution systems such as speech aids (57,58) for the deaf and visual substitution systems (59,60) for the blind. Electrocutaneous stimulation also has been used to improve the utility of upper extremity (61–68) and lower extremity (69–71) prostheses for amputees. Subdermal electrocutaneous stimulation has been used to provide sensory feedback to users of an upper extremity neuroprosthesis (72). Electrocutaneous arrays on the forehead have been used to help people with advanced cases of Hansen's disease (leprosy) to perform detailed manual tasks with reduced chance of injury (73). An early electrotactile device that is still in widespread use is the Tickle Talker speech perception device for the severely hearing impaired (57). The Tickle Talker stimulates eight tactile sensory nerves as they pass between the second and third knuckles of the fingers with speech-encoding patterns and has been shown to provide benefits in supplementing lipreading or residual hearing for hearing-impaired adults and children (54). Visual information has been successfully presented to blind users through electrotactile arrays located on the abdomen (75,76), back, or tongue (81). The tongue is also the target region for a newly commercialized electrotactile device from Wicab, Inc. The 144-point electrotactile array sits on the roof of the mouth against the tongue surface and communicates patterns for a wide range of proposed sensory substitution applications such as providing crude visual information for the blind (78) or restoring postural stability in patients with vestibular deficits (79).

Although the trunk or tongue may be most appropriate for applications requiring independent use of the hands, the fingertips present a much more natural means to explore a virtual surface or environment. Indeed, active haptic exploration leads to better perception of two-dimensional (2D) shapes than does passive static touch (1). Additionally, both mechanical and electrostatic displays require fixed, flat displays, but electrotactile arrays can be mounted in the fingers of a glove, allowing the user to explore arbitrary tactile surfaces in a more natural manner. As with mechanical displays, arrays of tiny electrodes are needed to utilize the high spatial acuity of the fingertip.

The development of a fingertip, electrotactile interface with good spatial resolution will require techniques to selectively activate nerve fibers to produce well localized, precisely placed sensations. Such techniques will necessarily exploit the dynamical and geometrical characteristics of fingertip mechanoreceptive afferents while accounting for similar characteristics of the nociceptive afferents. One drawback of electrotactile stimulation, however, is that producing well-localized, painless sensations remains challenging (75,80–82). Previously successful attempts to use electrocutaneous stimulation as a communications tool have typically been limited to the use of a small number of large electrodes, a paradigm unsuitable for telepresence and most other fingertip applications. Another drawback of electrotactile stimulation for fingertip displays is that we know little about how surface electrical stimulation excites mechanoreceptors. In particular, there is still much to be learned about how to selectively activate one population of mechanoreceptors over another. Mechanical stimulation excites the mechanoreceptive sensors themselves, while electrical stimulation can potentially excite the mechanoreceptive neurons anywhere along their length. It is therefore very difficult to control both the quality and the location of the electrically induced perception.

There has been some progress toward better control of the perceived quality and location of stimuli presented through small electrodes. Poletto and Van Doren (83) performed a study a few years ago to investigate the effects of pulse width, electrode diameter, and stimulus polarity on mechanoreceptive and nociceptive thresholds, as well as their effects on the perceived location and quality of the induced sensations (83). Touch and pain thresholds were measured for every combination of six pulse widths (0.05, 0.126, 0.315, 0.792, 1.991, and 5 ms), four electrode diameters (1, 2, 4, and 8 mm), and two stimulus polarities (anodic and cathodic). After each threshold measurement, the subject reported the perceived location (local or distal to the electrode) of the induced touch and pain sensations, as well as a qualitative descriptor of the pain sensation. The perceived location results indicated that anodic stimulation through small electrodes can be used to reliably excite fibers terminating near the electrode without concurrent excitation of near-by axons of passage. On the other hand, traditional cathodic stimulation, particularly through large electrodes, is much more likely than anodic to excite axons of passage, resulting in a sensation distal to the electrode (Fig. 4). These results (along with other related trends evident in the threshold data) were echoed in the model simulations only when fiber morphologies similar to those observed microscopically were incorporated. A complete discussion of these experiments, as well as neurophysiological models that were created to explain them, can be found in (83). Similar modeling results and explanations for the experimental data have been found by Kajimoto et al. (84), who used a model of a different form.

Another challenge for electrotactile stimulation is that the difference between the pain thresholds and the sensation thresholds (this difference is known as "dynamic range") is smaller for electrical stimulation through small electrodes than for large. This could potentially limit the maximum intensity that can be presented through small

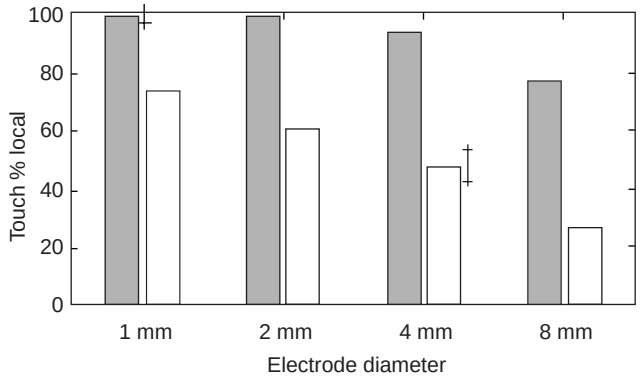


Figure 4. Qualitative results graphs. Electrotactile stimulation can produce sensations that are local or distal to the electrode. The graph shows the percentage of touch sensations that were felt local to (as opposed to distal to) the stimulating electrode, calculated as the percentage of all trials (summed across subject, session, pulse width and repetition, 271–288 total trials per bar) that the subject reported as "local". Bar color indicates stimulus polarity: black, anodic; white, cathodic. Error bars indicate smallest and largest 95% confidence intervals. [Data taken from (83).]

electrodes and could limit the desirability of using electrotactile displays. One way to increase the dynamic range would be to inactivate the pain fibers by applying a long depolarizing prepulse (DPP) prior to the stimulus pulse (the DPP would only be introduced significantly above sensation threshold, to prevent the concurrent elevation of both thresholds). This approach has been shown to significantly increase the pain thresholds, thereby increasing the dynamic range (83).

Despite the difficulties associated with fingertip electrotactile displays using small electrodes, some groups have been successful using electrotactile stimulation to present simple patterns to the fingertips. Kaczmarek et al. (81) developed a 49-point fingertip-scanned electrotactile display that consists of 0.89 mm diameter, flat-topped stainless steel electrode "pins", each surrounded by a 2.36 mm diameter air gap insulator arranged in a square 7×7 array (see Fig. 5 for a 24×24 array from the same group). They found that, using the highest comfortable current levels, subjects were able to correctly identify several simple geometric patterns $\sim 90\%$ of the time compared to near chance levels when stimulation was set to the lowest subthreshold current levels tested (81).

Other groups have also been working on electrotactile displays for selective stimulation of the various types of mechanoreceptors (56,84,86). Because the mechanoreceptors share the same neurophysiology (Aδ nerve fibers), the only way to selectively stimulate them electrically is to take advantage of their differences in geometry and location. This is possible by focusing the current to flow deeply or superficially into the skin by changing the current source distribution at the skin surface. In this way, it is possible to manipulate the electric field applied to each type of mechanoreceptive fiber (83,84). Kajimoto et al. developed a tactile display capable of selectively eliciting sensations of pressure, high frequency vibration and low frequency vibration, which they refer to as "tactile primary colors", analogous to the three primary colors for vision. This display was mounted on a computer mouse for tactile feedback in virtual reality (Fig. 6) and also on a surface with optical sensors underneath for conversion of light–dark patterns to tactile patterns (Fig. 7). The mouse-mounted application is innovative because it uses a force transducer beneath the fingertip array to allow the user to control the intensity of stimuli. This approach avoids the unpleasant sensation of sudden shocks that have discouraged the use of other electrotactile displays (87).

Electrotactile displays do not need to present pictographic information to be useful. In fact, even stimuli that are too weak to be felt can provide functional benefit. Low (subthreshold) levels of current (zero mean, white noise, 1 kHz bandwidth) delivered through large electrodes on the foot have been shown to enhance tactile sensitivity in the region between the electrodes in older adults (88). The effect is thought to result from the same principle of stochastic resonance as the vibrating insoles described above.

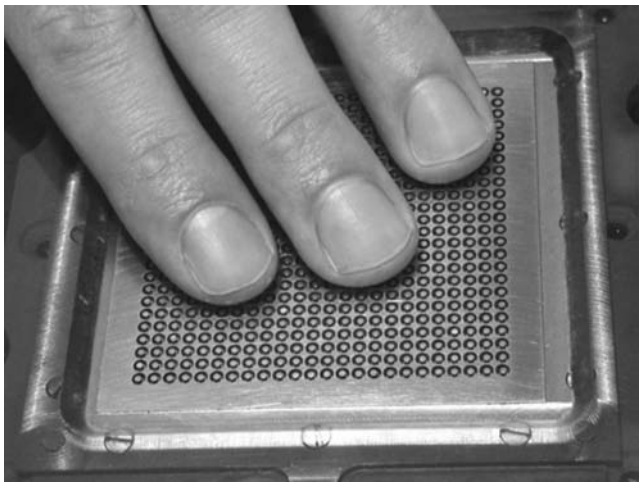


Figure 5. Fingertip electrotactile array. Electrotactile array with 576 electrodes (24×24) that can be scanned with multiple fingers. (Photo provided by Kurt Kaczmarek.)

Figure 6. Mouse-mounted tactile display. Mouse-mounted tactile display with 64 electrodes. The stimulating current is controlled by a force sensor located under the electrodes, allowing the user to control stimulus intensity and reducing frequency of painful shocks. (Photo provided by Hiroyuki Kajimoto.)

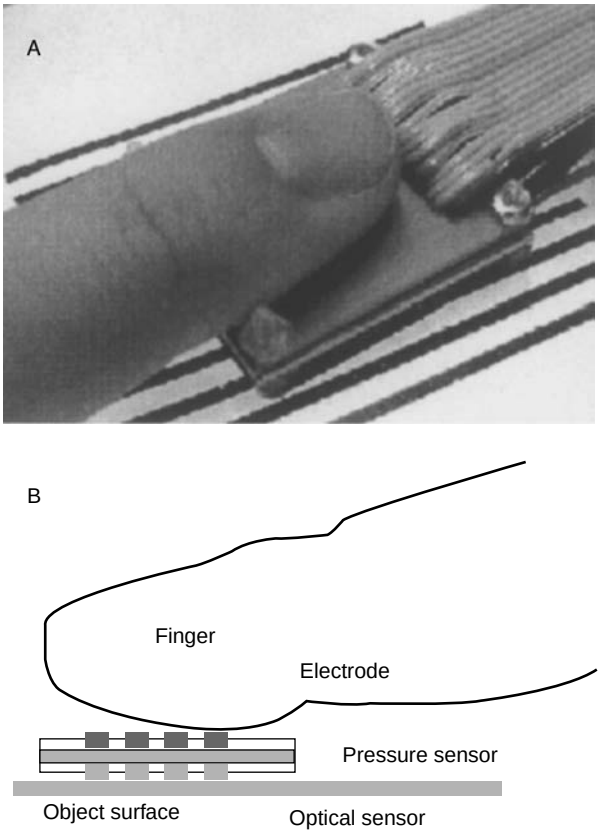


Figure 7. Smart Touch. Prototype of SmartTouch: (**a**) Visual image (black and white stripes) is captured by optical sensors and displayed through electrical stimulation. (**b**) Cross-section. (Photo provided by Hiroyuki Kajimoto.)

FUTURE DIRECTIONS

For most of its history, tactile stimulation has been developed and used primarily for sensory substitution or sensory augmentation systems for persons with sensory impairment. Many of these projects have been generally successful in restoring function and greater autonomy to their users, but target audiences have been small and large scale corporate investment has not been forthcoming. Although applications of tactile stimulation intended to help individuals with impairments are seen as important and will likely continue, it is also likely that applications with mass market appeal will drive the greater part of tactile stimulation technology development. Such applications are currently emerging in immersive virtual reality for entertainment and training purposes, teleoperation and telepresence for industry and medicine, and in personal communications.

There appear to be two separate trends in the technical progression of tactile displays. One trend is toward the development of fingertip tactile interfaces with greater sophistication and complexity for comfortably and portably communicating large amounts of data, such as surface texture. Such development will be important for applications where realistic surface perception is required, such as teleoperation, telesurgery, and immersive virtual reality. The other trend is toward simpler, more intuitive and portable systems to interface with wearable computing or personal communication systems.

Sophisticated array designs for communicating realistic texture sensations will require greater collaboration between material scientists, engineers, physiologists, and psychophysicists. Optimizing the information transfer rates, increasing usability, and improving the naturalism of the displays will require improvements in all areas of display design. Both mechanical and electrical designs could benefit from greater reliability and ease of maintenance, as well as better power supply and control strategies. Mechanical displays require improved actuator technology for lower power consumption, more degrees of freedom, and greater actuator density. Ongoing efforts include adding lateral skin stretch to fingertip mechanical arrays as a means to reproduce novel sensations and provide improved power performance (89). Preliminary studies showed that subjects perceived small tangential displacements as intense as larger normal displacement (90). Novel dimensions such as slip and thermal stimulation could also be combined with mechanical or electrotactile displays (21).

Optimal display geometries beyond the common row and column design will also need to be addressed. Although at least three decades of work have provided us with a reasonable understanding of the physiology and psychophysics of single tactor (or, to a lesser extent, a single electrode) stimulation, mechanical display designs could be improved through better theoretical understanding of the mechanics of skin deformation and mechanoreceptor excitation. Recent modeling efforts provide a good start (91) and further studies in this area need to include the viscoelastic properties for dynamic simulation, the further refinement of relevant stimuli for each type of mechanoreceptor, as well as investigating the role of finger ridges in tactile neural coding.

Much work is also still needed to understand the physical, physiological, psychophysical, and cognitive interactions that exist between tactile stimuli presented at different sites. One example is the so-called rabbit effect, or sensory saltation, in which the perceived stimulus location falls between the locations of two stimulators (92–94). This effect can be elicited through mechanical (95,96) or electrical (97,98) stimulation and may be useful in increasing the effective density of stimulation sites. As well documented as saltation is for stimulators that lie on a common line, its application for arrays is not as straightforward as one might expect. When researches using an array of nine electrodes arranged in a square on a subject's back attempted to use saltation to produce the sensation of a square being traced, the evoked sensation was actually that of a circle being traced (99). Clearly, when hundreds of stimulation sites are being used simultaneously, unanticipated interactions of this sort could be expected.

Substantial surround-masking effects have also been observed with two dimensional electrotactile arrays (100). Further research will be required to elucidate the impact of surround masking on fingertip array function. The precise application will most likely determine whether surround masking is beneficial (e.g., improving contrast and edge

detection, as in the retina) or detrimental (e.g., blurring the image or suppressing valuable tactile information).

Masking also occurs between two consecutive stimuli (101), and must be taken into account when predicting elicited percepts.

Continued improvements are also necessary in our understanding the way that the complex interactions between the various stimulus parameters influence stimulus perception. For example, the information transfer rate for a single stimulator can be improved through coherent modulation of stimulus frequency and amplitude, but when the frequency and amplitude are modulated independently, the information transfer rate of each channel is reduced due to cognitive processing interactions (103). Similar interactions may exist for all of the dimensions along which stimulus parameters can be modulated.

Many challenges still face designers of fingertip electrotactile arrays. Physically, such an array must be held on the finger so that firm contact is maintained even during exploratory finger movements. Embedding the array in a custom-made glove should help solve this problem, but perspiration tends to accumulate between adjacent electrodes, providing a low impedance current pathway that circumvents the skin (103–106). Possibly, the material or chemical properties of the glove itself could be used to wick perspiration away or inhibit its secretion. Note that in an array where the electrodes positions are fixed relative to the fingertip, this problem is much less severe than when the finger is actively scanned over a flat array (108). Kajimoto et al. (108) suggested that for some applications, the electrodes could even be printed directly on the skin using conductive ink.

In addition to the physical problems, significant physiological and psychophysical problems of adaptation (habituation) and interaction must be addressed. It is well known that habituation, that is, a change in the perceived magnitude for constant stimulus amplitude, occurs over time with either mechanical (12,109–111) or electrical (68,112,113) stimulation of mechanoreceptors. Compensation for adaptation may be complicated when stimulating through an electrotactile array since each electrode site will be stimulated at a different rate, and therefore the amount of habituation will vary across the array.

Recent evidence shows that a few hours of fingertip tactile stimulation results in cortical reorganization that improves spatial discrimination but impairs frequency discrimination performance (114). It is not yet known how long the impairment lasts and whether extensive long-term use of vibrotactile or electrotactile stimulators could lead to learned functional sensory deficits.

The emerging field of wearable computing is giving rise to a new set of applications for tactile stimulation. Wearable computers for industrial and military maintenance applications have been reported for at least a decade and applications for the mass consumer market have been proposed, with several in current development. Tactile displays are considered viable alternatives to visual or auditory displays for wearable computers because they can be unobtrusive, socially discrete, and do not interfere with normal vision or sound. The goal of wearable tactile displays can be conceptually different than the sensory substitution, virtual reality and teleoperation displays as discussed above. Most of the displays described above were concerned primarily with directly translating real or virtual visual, tactile, or audio information into tactile stimulation. The tactile information displays envisioned for wearable computers, on the other hand, present information that is not directly based on visual, tactile or audio information. Instead, they aim to communicate data on the state of the worn computer, the environment, or the person wearing it. A very simple example is the vibration of a cell phone in vibrate mode. The vibration is intended to alert the user to an incoming call, not to communicate the voice message itself. MacLean and Enriquez suggested that such tactile alerts will become more common and more complex, taking into account optimization along several perceptual dimensions to optimize information transfer (115). They coined the term "haptic icons" to refer to brief synthetic haptic or tactile signals to convey information such as event notification, identity, content or state.

According to Gemperle et al. (116), wearable tactile displays must be lightweight, silent, small, and physically discreet. They must also have low power requirements, be able to be felt through the clothing they are embedded in or worn on, and they must be held tightly enough to the body to maintain reliable contact. Initial designs would also benefit from being flexible enough to be used in multiple applications with little modification.

One wearable tactile display application currently being pursued by several independent groups is an aid for a personal navigation. Tactile navigation aids have been proposed for pilots (117), drivers (118), scuba divers (119), walkers or hikers (116), astronauts (120,121) and the blind (122). The basic principle is that vibrating tactors embedded in the user's clothing signal either a particular direction (e.g., north, up, the direction to return the vehicle) or the distance to and direction of the next turn or position correction required to reach a destination.

Other applications of wearable tactile displays include providing silent and private alerts, socially subtle transmission of information, providing biofeedback of physiological states such as blood sugar or blood alcohol levels. Future work includes the development of standardized tactile display clothing elements (123), and requires continued research on where to define the spaces on the human body where solid and flexible forms can rest without interfering with fluid human movement. The widespread use of wearable tactile displays depends on consideration not only of function, but of comfort, mobility and social acceptance factors.

BIBLIOGRAPHY

1. Loomis JM, Lederman SJ. Tactual perception. In: Boff KR, Kaufman L, Thomas JP, editors. Handbook of Perception and Human Performance: Vol. II, Cognitive Processes and Performance. New York: John Wiley & Sons; 1986. p 31-1 to 31-41.
2. Perez A, et al. What is the value of telerobotic technology in gastrointestinal surgery? Surg Endosc 2003;17(5):811–813.
3. Rassweiler J, Binder J, Frede T. Robotic and telesurgery: Will they change our future? Curr Opin Urol 2001;11(3):309–320.

4. Halata Z. The mechanoreceptors of the mammalian skin: Ultrastructure and morphological classification. Adv Embryol Cell Biol 1975;50:1–77.
5. Chouchkov C. Cutaneous receptors. Adv Anatomy Embryol Cell Biol 1978;54(5):1–62.
6. Iggo A, Andres KH. Morphoogy of cutaneous receptors. Ann Rev Neurosci 1982;5:1–31.
7. Johnson KO, Hsiao SS, Yoshioka T. Neural coding and the basic law of psychophysics. Neuroscientist 2002;8(2):111–121.
8. Srinivasan MA, Whitehouse JM, LaMotte RH. Tactile detection of slip: Surface microgeometry and peripheral neural codes. J Neurophysiol 1990;63(6):1323–1332.
9. Johansson RS, Landstrom U, Lundstrom R. Responses of mechanoreceptive afferent units in the glabrous skin of the human hand to sinusoidal skin displacements. Brain Res 1982;244(1):17–25.
10. Ochoa J, Torebjörk. Sensations evoked by intraneural microstimulation of single mechanoreceptor units innervating the human hand. J Physiol 1983;342:633–654.
11. Kunesch E, et al. Somatosensory evoked potentials elicited by intraneural microstimulation of afferent nerve fibers. J Clin Neurophysiol 1995;12:476–487.
12. Bolanowski SJ, et al. Four channels mediate the mechanical aspects of touch. JASA 1988;84(5):1680–1694.
13. Johnson KO. The roles and functions of cutaneous mechanoreceptors. Curr Opin Neurobiol 2001;11(4):455–461.
14. Olausson H, Wessberg J, Kakuda N. Tactile directional sensibility: peripheral neural mechanisms in man. Brain Res 2000;866(1–2):178–187.
15. Olausson H, et al. Unmyelinated tactile afferents signal touch and project to insular cortex. Nat Neurosci 2002;5(9):900–904.
16. Wessberg J, et al. Receptive field properties of unmyelinated tactile afferents in the human skin. J Neurophysiol 2003;89(3):1567–1575.
17. Pasquero J. A tactile display using lateral skin stretch. In: Mechanical Engineering. Montreal, Canada: McGill University; 2003.
18. Aiello GL. Multidimensional electrocutaneous stimulation. IEEE Trans Rehabil Eng 1998;6(1):95–101.
19. Kaczmarek KA, Haase SJ. Pattern identification and perceived stimulus quality as a function of stimulation waveform on a fingertip-scanned electrotactile display. IEEE Trans Neural Syst Rehabil Eng 2003;11(1):9–16.
20. Ino S, Shimizu S, Odagawa T, Sato M, Takahashi M, Izumi T, Ifukube T. A Tactile Display for Presenting Quality of Materials by Changing the Temperature of Skin Surface. In: IEEE International Workshop on Robot and Human Communication. Tokyo: 1993.
21. Benali-Khoudja M, Hafez M, Alexandre J-M, Kheddar A, Moreau V. VITAL: A VIbroTActiLe Interface with Thermal Feedback. In: 2004 IEEE International Conference on Robotics & Automation; New Orleans, LA: 2004.
22. Caldwell G, Gosney C. Enhanced tactile feedback (teletaction) using a multi-functional sensory system. In: IEEE International Conference on Robotics and Automation; Atlanta, GA: 1993.
23. Collins JJ, Imhoff TT, Grigg P. Noise-enhanced tactile sensation. Nature 1996;383(6603):770.
24. Liu W, et al. Noise-enhanced vibrotactile sensitivity in older adults, patients with stroke, and patients with diabetic neuropathy. Arch Phys Med Rehabil 2002;83(2):171–176.
25. Khaodhiar L, et al. Enhancing sensation in diabetic neuropathic foot with mechanical noise. Diabetes Care 2003;26 (12):3280–3283.
26. Dennerlein J, Millman P, Howe R. Vibrotactile feedback for industrial telemanipulators. In: Sixth Annual Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, ASME International Mechanical Engineering Congress and Exposition. Dallas: 1997.
27. Dennerlein J, Shahion E, Howe R. Vibrotactile feedback for an underwater teleoperated robot. In: International Symposium on Robotics with Applications (ISORA). Maui Hawaii: 2000.
28. Debus T, et al. Multichannel vibrotaction display for sensory substitution during teleoperation. In: Telemanipulator and telpresence technology VIII. Newton, MA: SPIE; 2001.
29. Collins CC. Tactile telvision—Mechanical and electrical image projection. IEEE Trans Man-Mach Sys 1970;MMS-11:65–71.
30. Linvill JG, Bliss JC. A direct translation reading aid for the blind. Proceedings of the IEEE 1966;54(1):40–51.
31. Stein DK. The Optacon: Past, present, and future. In: The Braille Monitor; 1998.
32. Saunders FA. Information transmission across the skin: High-resolution tactile sensory aids for the deaf and the blind. Int J Neurosci 1983;19(1–4):21–28.
33. Kaczmarek K, et al. A tactile vision-substitution system for the blind: Computer-controlled partial image sequencing. IEEE Trans BME 1985;32(8):602–608.
34. Perez CA, Holzmann CA, Jaeschke HE. Two-point vibrotactile discrimination related to parameters of pulse burst stimulus. Med Biol Eng Comput 2000;38(1):74–79.
35. Verrillo RT, Fraioli AJ, Smith RL. Sensation magnitude of vibrotactile stimuli. P&P 1969;6:366–372.
36. Gescheider GA, et al. Vibrotactile intensity discrimination measured by three methods. JASA 1990;87(1):330–338.
37. Summers IR, et al. Vibrotactile and electrotactile perception of time-varying pulse trains. JASA 1994;95(3):1548–1558.
38. Johansson RS, Landstrom U, Lundstrom R. Sensitivity to edges of mechanoreceptive afferent units innervating the glabrous skin of the human head. Brain Res 1982;244(1): 27–35.
39. Perez CA, et al. Power requirements for vibrotactile piezoelectric and electromechanical transducers. Medical & Biological Engineering & Computing 2003;41(6):718–726.
40. Hayward V, Cruz-Hernandez M. Tactile display device using distributed lateral skin stretch. In: Haptic Interfaces for Virtual Environment and Teleoperator Systems Symposium. Orlando, Florida, USA: ASME IMECE2000; 2000.
41. Phillips JR, Johnson KO. Tactile spatial resolution. III. A continuum mechanics model of skin predicting mechanoreceptor responses to bars, edges, and gratings. J Neurophysiol 1981;46(6):1204–1225.
42. Vincent FVJ. Structural Biomaterials. Princeton University Press; 1991.
43. Caldwell DG, Sagarakis NT, Giesler C. An integrated tactile/shear feedback array for stimulation of finger mechanoreceptors. In: IEEE International Conference on Robotics & Automation; Detroit: Michigan; 1999.
44. Moy G, et al. Human psychophysics for teletaction system design. Haptics-e 2000;1(3).
45. Srinivasan MA, LaMotte RH. Tactual discrimination of softness. J Neurophysiol 1995;73(1):88–101.
46. LaMotte RH, Srinivasan MA. Neural encoding of shape: responses of cutaneous mechanoreceptors to a wavy surface stroked across the monkey fingerpad. J Neurophysiol 1996;76(6):3787–3797.
47. Fischer H, Trapp R. Tactile optical sensor for use in minimal invasive surgery. Stud Health Technol Inform 1996;29:623–629.

48. Strong RM, Troxel DE. An electrotactile display. IEEE Trans Man Machine Systems 1970;MMS-11(1):72–79.
49. Tang H, Beebe DJ. A microfabricated electrostatic haptic display for persons with visual impairments. IEEE Trans Rehabil Eng 1998;6(3):241–248.
50. Pawluk DT, et al. Controla nd pattern specification for a high density tactile array. In: Proceedings of the ASME Dynamic Systems and Control Division; New York, NY: ASME; 1998.
51. Howe R, Kontarinis D, Peine W. Shape memory alloy actuator controller design for tactile displays. In: Proc of the 34th IEEE Conf. on Decision and Control; IEEE 1995.
52. Hasser CJ, Weisenberger JM. Preliminary evaluation of a shape-memory alloy tactile feedback display. DSC: Advances in Robotics, Mechatronics and Haptic Interfaces 1993;49:73–80.
53. Cohn MB, ML, Fearing RS. Tactile feedback for teleoperation. In: SPIE Conf. 1833, Telemanipulator Technology; Boston, MA: SPIE; 1992.
54. Wagner CR, Lederman SJ, Howe RD. A tactile shape display using RC servomotors. In: The Tenth Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. Orlando, FL: 2002.
55. Ghodssi R, et al. Development of a tangential tactor using a LIGA/MEMS linear microactuator technology. In: Micro-Electro-Mechanical systems; International Mechanical Engineering Congress and Exposition. New York, NY: 1998.
56. Asamura N, Yokoyama N, Shinoda H. Selectively stimulating skin receptors for tactile display. IEEE Computer Graphics and Applications 1998;98:32–37.
57. Blamey PJ, Clark GM. Psychophysical studies relevant to the design of a digital electrotactile speech processor. JASA 1987;82(1):116–125.
58. Cowan RSC, et al. Perception of multiple electrode stimulus patterns: Implications for design of an electrotactile speech processor. JASA 1991;89(1):360–369.
59. Bach-y-Rita P. Tactile vision substitution: Past and future. Intern J Neurosci 1983.
60. Collins CC. Electrotactile visual prosthesis. In: Hambrecht TF, Reswick JB, editors. Functional Electrical Stimulation. New York: Marcel Dekker; 1977. p 189.
61. Rohland TA. Sensory feedback in upper-limb prosthetic systems. Inter-Clinic Information Bulletin 1974;13(9):1–8.
62. Schmidl H. The importance of information feedback in prostheses for the upper limbs. Prosthetics and Orthotics International 1977;1:21–24.
63. Shannon GF. Sensory feedback for artificial limbs. Medical Progress Technology 1979;6:73–79.
64. Prior RE, et al. Supplemental sensory feedback for the VA/NU myoelectric hand. Background and preliminary designs. Bull Prosth Res 1976;11:171–191.
65. Scott RN, et al. Sensory-feedback system compatible with myoelectric control. Med Biol Eng Comput 1980;18:65–69.
66. Scott RH. Feedback in myoelectric prostheses. Clin Orth Related Res 1990;256:58–63.
67. Lovely DF, Hudgins BS, Scott RN. Implantable myoelectric control system with sensory feedback. Med Biol Eng & Comput 1985;23:87–89.
68. Szeto AYJ, Lyman J. Comparison of codes for sensory feedback using electrocutaneous tracking. Ann Biomed Eng 1977;5:367–383.
69. Kawamura J, et al. Sensory feedback systems for the lower-limb prosthesis. J Osaka Rosai Hospital 1981;5(2):104–112.
70. Sabolich JA, Ortega GM. Sense of feel for lower-limb amputees: A phase-one study. J Prosthetics Orthotics 1994;6:36–41.
71. Schmid HP, Bekey GA. Tactile information processing by human operators in control systems. IEEE Trans Syst Man Cybern 1978;8(12):860.
72. VanDoren CL, Riso RR, Milchus K. Sensory feedback for enhancing upper extremity neuromuscular prostheses. J Neurol Rehab 1991;5:63–74.
73. Collins CC, Madey JMJ. Tactile sensory replacement. In: Proc San Diego Biomed Symp 1974; 15–26.
74. Cowan R, et al. Improved electrotactile speech processor: Tickle Talker. Ann Otol Rhinol Laryngol Suppl 1995;166:454–456.
75. Saunders FA. Electrocutaneous displays. In: Cutaneous communication systems and devices. Monterey, CA: Psychonomic Society; 1974.
76. Tyler M, et al. A New Electrotactile Prosthesis for the Blind. Unitech Res Inc.; 1993.
77. Kaczmarek KA, Tyler ME. Effect of electrode geometry and intensity control method on comfort of electrotactile stimulation on the tongue. In: ASME Dyn Sys Contr Div. Orlando, FL: ASME; 2000.
78. Sampaio E, Maris S, Bach-y-Rita P. Brain plasticity: 'visual' acuity of blind persons via the tongue. Brain Res 2001;908(2):204–207.
79. Tyler M, Danilov Y, Bach YRP. Closing an open-loop control system: vestibular substitution through the tongue. J Integr Neurosci 2003;2(2):159–164.
80. Saunders FA. Recommeded procedures for electrocutaneous displays. In: Hambrecht TF, Reswick JB, editors. Functional Electrical Stimulation. New York: Marcel Dekker; 1977. p 303.
81. Kaczmarek KA, et al. Electrotactile and vibrotactile displays for sensory substitution systems. IEEE Trans Biomed Eng 1991;BME-38(1):1–16.
82. Poletto CJ, Van Doren CL. Elevating pain thresholds in humans using depolarizing prepulses. IEEE Trans Biomed Eng 2002;49(10):1221–1224.
83. Poletto CJ. Fintertip electrocutaneous stimulation through small electrodes. In: Biomedical Engineering. Cleveland, OH: Case Western Reserve University; 2001.
84. Kajimoto H, et al. Electrocutaneous display as an interface to a virtual tactile world. In: Virtual Reality Conference. Yokohama, Japan: IEEE Computer Society; 2001.
85. Kaczmarek KA, Haase SJ. Pattern identification as a function of stimulation current on a fingertip-scanned electrotactile display. IEEE Trans Neural Syst Rehabil Eng 2003;11(3):269–275.
86. Asamura N, Yokoyama N, Shinoda H. A method of selective stimulation to epidermal skin receptors for realistic touch feedback. In: IEEE Conference on Virtual Reality. 1999.
87. Kajimoto H, et al. Electro-tactile display with force feedback. World Multiconference on Systemics, Cybernetics and Informatics. Orlando, FL: 2001.
88. Dhruv NT, et al. Enhancing tactile sensation in older adults with electrical noise stimulation. Neuroreport 2002;13(5):597–600.
89. Pasquero J, et al. Display of virtual Braille dots by lateral skin deformation: A pilot study. In: Eurohaptics. Munich, Germany: 2004.
90. Biggs J, Srinivasan MA. Tangential versus normal displacements of skin: Relative effectiveness for producing tactile sensations. In: 10th International Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. Orlando FL: IEEE Computer Society; 2002.
91. Dandekar K, Raju BI, Srinivasan MA. 3-D finite-element models of human and monkey fingertips to investigate the mechanics of tactile sense. J Biomech Eng 2003;125(5):682–691.

92. Geldard FA, Sherrick CE. The cutaneous "rabbit": A perceptual illusion. Science 1972;178(57):178–179.
93. Geldard FA, Sherrick CE. The cutaneous saltatory area and its presumed neural basis. Perception and Psychophys 1983;33(4):299–304.
94. Geldard FA. Saltation in somesthesis. Psychological Bulletin 1982;92(1):136–175.
95. Cholewiak RW. Exploring the condition that generate a good vibrotactile line. In: Psychonomic Socitety Meetings. Los Angeles: 1995.
96. Cholewiak RW, Sherrick CE, Collins AA. Studies of saltation. In: Princeton Cutaneous Research Project. Princeton University, Department of Psychology; 1996.
97. Tanie K, et al. Basic study on discriminability of mental location of electrocutaneous phantom sensation. Transactions of the Society of Instrument and Control Engineers 1979;15(4):505–512.
98. Tanie K, et al. Information transmission characteristics of two-dimensional electrocutaneous phantom sensation. Transactions of the Society of Instrument and Control Engineers 1980;16(5):732–739.
99. Tan HZ, Lim A, Traylor R. A psychophysical study of sensory saltation with an open response paradigm. In: DSC-ASME Dynamic Systems and Control Division. ASME; 2000.
100. Szeto AY, Saunders FA. Electrocutaneous stimulation for sensory communication in rehabilitation engineering. IEEE Trans Biomed Eng 1982;29(4):300–308.
101. Craig JC. Identification of scanned and static tactile patterns. Percept Psychophys 2002;64(1):107–120.
102. Poletto CJ, VanDoren CL. Perceptual interactions between electrocutaneous loudness and pitch. IEEE Transactions Rehab Eng 1995;3(4):334–342.
103. Melen RD, Meindl JD. Electrocutaneous stimulation in a reading aid for the blind. IEEE Trans Biomed Eng 1971;18(1):1–3.
104. Tyler M, et al. A dynamic multi-point electrotactile display: The effects of peak voltage, pulse-phase width, number of pulses, and geometric area on the perception of haptic threshold. IEEE Transactions on Rehabilitation Engineering.
105. Kaczmarek K. Tactile displays. In: Barfiled W, TF III, editors. Virtual Environments and Advanced Interface Design. Oxford: Oxford University Press; 1995.
106. Kaczmarek K, Tyler ME, Bach-y-Rita P. Electrotactile haptic display on the fingertips: Preliminary results. In: 16th Annual International Conference IEEE Eng Med Biol Soc; Baltimore, MD: IEEE; 1994.
107. Kaczmarek K. Sensory augmentation and substitution. CRC Handbook of Biomedical Engineering. 1995.
108. Kajimoto H, et al. SmartTouch: Electric skin to touch the untouchable. IEEE Comput Graph Appl 2004;24(1): 36–43.
109. Gescheider GA, et al. Vibrotactile forward masking: Psychophysical evidence for a triplex theory of cutaneous mechanoreception. JASA 1985;78(2):534–543.
110. Hahn JF. Tactile adaptation. in The Skin Senses. 322–326.
111. Verrillo RT, et al. Vibrotactile masking: Effects of one- and two-site stimulation. P&P 1982;33(4):379–387.
112. Szeto AYJ, Farrenkopf GR. Optimization of single electrode tactile codes. Ann Biomed Eng 1992;20:647–665.
113. Kaczmarek KA. Electrotactile adaptation on the abdomen: Preliminary results. IEEE Trans Rehabil Eng 2000;8(4): 499–505.
114. Hodzic A, et al. Improvement and decline in tactile discrimination behavior after cortical plasticity induced by passive tactile coactivation. Neuroscience 2004;24(2):442–446.
115. MacLean K, Enriquez M. Perceptual Design of Haptic Icons. In: EuroHaptics 2003. Dublin, UK: 2003.
116. Gemperle F, Ota N, Siewiorek D. Design of a wearable tactile display. In: Fifth International Symposium on Wearable Computers. Zürich, Switzerland: 2001.
117. van Erp JBF, Veltman JA, van Veen HAHC. A tactile cockpit instrument to support altitude control. In: Human Factors and Ergonomic Society 47th Annual Meeting. 2003.
118. van Erp JB, Padmos P. Image parameters for driving with indirect viewing systems. Ergonomic 2003;46(15):1471–1499.
119. McTrusty T, Walters C. Swimmer Inshore Navigation System (SINS) Tactile Situation Awareness System (TSAS) Test Report. Panama City, FL: Coastal Systems Station; 1997.
120. Rochlis JL, Newman DJ. A tactile display for international space station (ISS) extravehicular activity (EVA). Aviat Space Environ Med 2000;71(6):571–578.
121. van Erp JBF, van Veen HAHC. A multipurpose tactile vest for astronauts in the International Space Station. In: Eurohaptics. Dublin, Ireland: 2003.
122. Ross D, Blasch B. Development of a Wearable Computer Orientation System. Personal Ubiquitous Computing 2002;6(1):49–63.
123. Toney A, et al. A shoulder pad insert vibrotactile display. In: Seventh International Symposium on Wearable Computers. 2003.
124. Johansson RS, Vallbo ÅB. Tactile sensory coding in the glabrous skin of the human hand. Trends in Neurosci 1983;6(1):27–32.

See also FUNCTIONAL ELECTRICAL STIMULATION; HEAT AND COLD, THERAPEUTIC; SKIN, BIOMECHANICS OF; TRANSCUTANEOUS ELECTRICAL NERVE STIMULATION (TENS)

TEETH, PROPERTIES OF. See BONE AND TEETH, PROPERTIES OF.

TELEMETRY. See BIOTELEMETRY.

TELERADIOLOGY

H.K. HUANG
University of Southern California
Los Angeles, California

INTRODUCTION

Telemedicine and teleradiology have become increasingly important as our country's healthcare delivery system gradually changes from fee-for-service to managed, capitated care. During the past several years, we have seen the trend of primary care physicians joining health maintenance organizations (HMOs). These HMOs purchase smaller hospitals and form hospital groups under the umbrella of HMOs. Also, academic institutions form consortia to compete with other local hospitals and HMOs. This consolidation allows the elimination of duplication and the streamlining of healthcare services among hospitals. As a result, costs are reduced, but at the same time because of the downsizing, the number of experts available for service also decreases. Utilization of telemedicine and

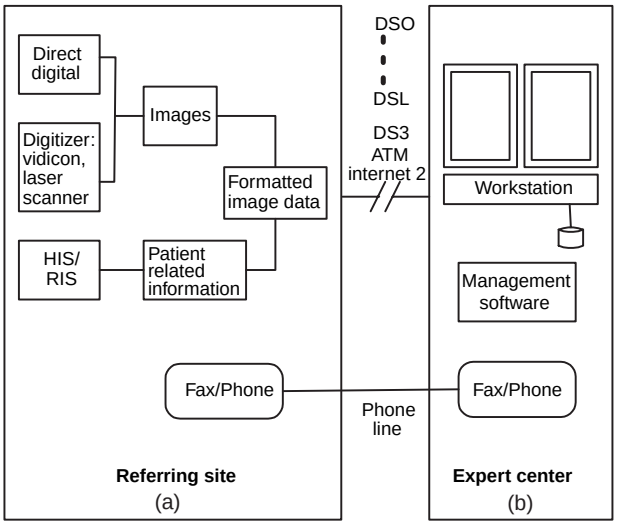


Figure 1. A generic teleradiology setup. (a) Referring site, (b) expert center.

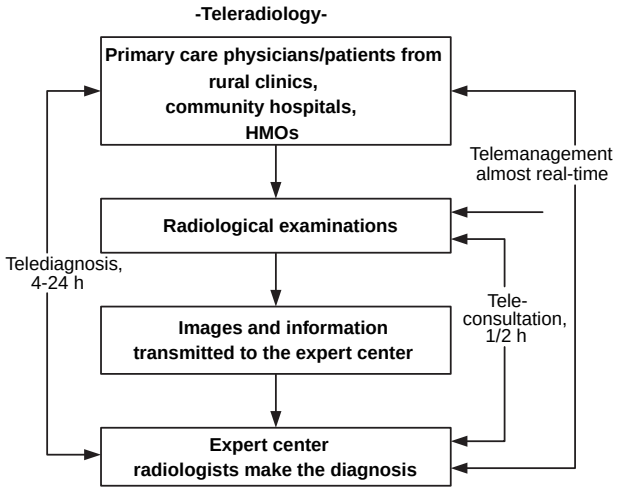


Figure 2. The expert center in teleradiology with three reading modes: telediagnosis, teleconsultation, and telemanagement. These three reading modes dictate the image communication requirements.

teleradiology is a method to alleviate the diminishing of experts, streamline the diagnosis process, and save healthcare costs.

Teleradiology is a subset of telemedicine operation focusing in remote diagnosis of medical images. Teleradiology utilizes computer, display, and telecommunication technologies for radiologists to make remote diagnosis from radiological images generated at distant examination sites. The diagnostic report is sent to the examination site where a primary physician can provide proper treatments to the patient immediately. Figure 1 shows a generic teleradiology set-up illustrating that teleradiology is not a single medical device or an instrument, instead, it is a system integration of various imaging devices using communication technology and system software connecting multiple imaging centers and expert centers together (1–3). Dependent on the required turn around time in obtaining the diagnosis from the examination site, the expert center has three reading modes: telediagnosis, teleconsultation, and telemanagement, which are shown in Fig. 2. These reading modes dictate the communication requirements of transmitting the images between the sites. Teleradiology operation can be very simple or extremely complicated. In the simple case, a radiology resident may send an image set from a CT (computed tomography) scanner using low quality teleradiology equipment and slow speed communication technology in the evening to the radiologist's home for a second opinion. This type of teleradiology operation does not require highly sophisticated equipment. A conventional telephone and simple desktop personal computer with modem connection and display software are sufficient to perform the teleradiology operation. This type of application originated in early 1970.

The complicated teleradiology operation can have different models starting from simple to complicated in ascending order as shown in Table 1. The complications occur when the current examination requires historical images for comparison, and when the radiologist needs information from the radiology information system (RIS) to make a diagnosis. In addition, complications arise when the images and the corresponding diagnosis report are required to be archived to the patient data file. Teleradiology is relatively simple to operate when neither retrieval nor archive of previous information and images is required. However, when both archive and retrieval are required, the operation becomes extremely complicated.

TELERADIOLOGY AND PACS

Picture Arching and Communication System (PACS) is a hospital integrated imaging management system developed in the early 1990s (Fig. 3) (4). The infrastructure of PACS is shown in the upper three rows. Two types of servers in the bottom of the figure are for various PACS applications. When teleradiology service requires patient's historical images as well as related information, technologies used in both teleradiology and PACS become very similar. Table 2 shows technologies used in teleradiology and PACS, the major differences are in image capture, communication, and storage. Some current teleradiology operations still use a film digitizer as the primary method of converting a projection film image-to-digital format, although the trend is moving toward direct digital capture. In PACS, direct digital image capture using Digital Imaging Communication in Medicine (DICOM) standard format is mostly used. In networking, teleradiology uses slower speed wide area networks (WAN) compared with the higher speed local area network (LAN) used in PACS.

Table 1. Four Models of Teleradiology Operation According to Its Complexity

	Historical Images/RIS	Archive
Most simplistic	No	No
Simplistic	Yes	No
Complicated	No	Yes
Most complicated	Yes	Yes

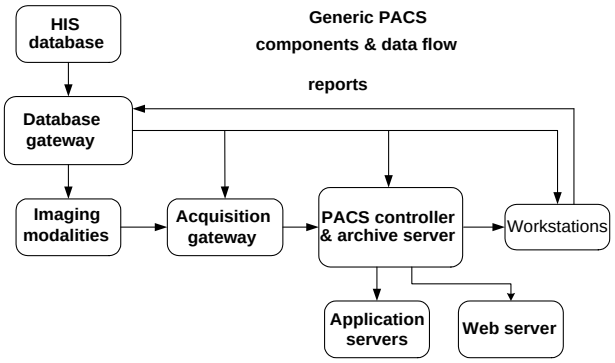


Figure 3. A generic PACS components and data flow. The upper three rows are the infrastructure, the bottom row is PACS applications.

Table 2. Differences in Technology Used between Teleradiology and PACS

Function	Telerad	PACS
Image capture	Digitizer	DICOM
Display technology	Same	Same
Networking	WAN	LAN
Storage	Short	Long
Compression	Yes	May be

In teleradiology, image storage is mostly short term, whereas in PACS it is long term. Teleradiology relies heavily on image compression, whereas PACS may or may not.

PACS and teleradiology use medical images shown in Table 3 for radiologists to make diagnosis. In Table 3, the first and second columns give types and sizes of some common medical images. In clinical applications, one examination is composed of many images of different views and anatomical emphasis, since a single image is generally not sufficient for making the proper diagnosis. In general, a typical examination generates between 10 and 20 MB, although some current CT and MR imaging modalities can generate up to 3000 images per examination. The fourth column shows an average size of one typical examination in each of these image modalities. The high extreme is in digital mammography, which routinely requires 160 MB. To transmit 160 MB of images through WAN for teleradiology requires a very high bandwidth communication technology.

WHY DO WE NEED TELERADIOLOGY?

The managed care trend in healthcare delivery expedites the formation of teleradiology expert centers. However, even without the healthcare reform, teleradiology is still an extremely important component in radiology practice for the following reasons: First, teleradiology secures images for radiologists to read so that no images will be accidentally lost in transit. Second, teleradiology reduces the reporting cycle time after the image is generated. Third, since radiology is subdivided into many subspecialties, a general radiologist requires a specialist's second opinion on occasion. The availability of teleradiology will facilitate seeking a second opinion. Fourth, teleradiology increases radiologists' income since no images would accidentally be lost and subsequently not reported. The healthcare reform adds two more reasons. (1) It saves healthcare costs since an expert center can serve multiple sites reducing the number of radiologists required. (2) It improves the efficiency and effectiveness of healthcare because the report turn-around time would be reduced and there would be no loss of image (5).

TELERADIOLOGY COMPONENTS

A generic schematic of teleradiology depicted in Fig. 1 shows teleradiology components and their connections. Modalities generating images in teleradiology applications include CT, MR, CR, DR, US, NM, DSA–digital fluorography (DF), and film digitizer. Their respective image and examination file sizes Images are shown in Table 3. These acquisition devices are first generated from the examination site and then sent through communication networks to the expert center if they are already in digital format. Or, if these images are stored on films, then they need to be digitized by a film scanner at the examination site.

Table 3. Data Size of Some Common Medical Images and Examinations

Instrument	One Image, bits	No. of Images/Exam	One Examination
Nuclear medicine (NM)	$128 \times 128 \times 12$	30–60	1–2 MB
Magnetic resonance imaging (MRI)	$256 \times 256 \times 12$	60–3000	8 MB up
Ultrasound (US)[a]	$512 \times 512 \times 8(24)$	20–240	5–60 MB
Digital subt. Angiography (DSA)	$512 \times 512 \times 8$	15–40	4–10 MB
Digital microscopy	$512 \times 512 \times 8$	1	0.26 MB
Digital color microscopy	$512 \times 512 \times 24$	1	0.79 MB
Color light images	$512 \times 512 \times 24$	4–20	3–15 MB
Computed tomography (CT)	$512 \times 512 \times 12$	40–3000	20 MB up
Computed/digital radiog (CR/DR)	$2048 \times 2048 \times 12$	2	16 MB
Digitized X rays	$2048 \times 2048 \times 12$	2	16 MB
Digital mammography	$4000 \times 5000 \times 12$	4	160 MB

[a]Doppler US with 24 bit color images.

Image Capture

In image capture, if the original image data are on film, then either a video frame grabber or a laser film digitizer is used to convert them to digital format. A video frame grabber produces low quality digital images, but is faster and cheaper. On the other hand, laser film digitizers produce very high quality digital data, but take longer and cost more compared to the video frame grabber. During the past several years, direct Digital Imaging and Communication in Medicine (DICOM) standard output images from CR, DR, CT, and MR have been used extensively in teleradiology.

Data Reformatting

After images are captured, it is advantageous to convert these images and related data to industry standards because multiple vendors' equipment can be used in the teleradiology chain. The two common standards used in medical imaging industry are the DICOM (6) for images and Health Level 7 (HL7) (7) for textual data. The DICOM standard includes both the image format as well as the communication protocols based on the standard TCP/IP. Health level 7 is the standard for textual data, it uses the TCP/IP communication protocols.

Image Storage

At the expert center, a local storage device is used before images are displayed. The capacity of this device can range from several hundred megabytes to many gigabytes. A long-term archive, such as a small DLT (digital linear tape) library, is used for teleradiology applications that require historical images and diagnostic reports, related patient information, and current images and diagnosis.

Display Workstation

For an inexpensive teleradiology system, a low cost 512-line single monitor can be used for displaying images. However, high resolution multimonitor display workstations are needed for the primary diagnosis.

Table 4 shows the specifications of a 2000- and a 1600-line workstation used for teleradiology primary readings. These state-of-the-art technology diagnostic workstations, use two monitors with over 2 GB of local storage, and can display images and reports from the local storage in 1–2 s. A 2000-line LCD monitor workstation costs from $20,000 to 30,000, and a 1,600 line costs from $15,000 to $20,000. User-friendly image display software is necessary for easy and convenient use by the radiologist at the workstation.

Table 4. Specifications of High-End 2000 and 1600 Line Workstations for Teleradiology

Two LCD Monitors
1–2 week local storage for current + previous exams
1–2 s display of images and reports from local storage
HL7 and DICOM conformance
Simple image processing functions

Table 5. Transmission Rate of Current Wide Area Network Technology

DS-0	56 kbits/s
DS-1	56 to (24 × 56) kbits/s
DSL	144 kbits/s – 8 Mbits/s
DS-1 (T1)	1.5 Mbits/s
ISDN	56 kbits/s to 1.5 Mbits/s
DS-3 (T3)	28 DS-1 = 45 Mbits/s
ATM (OC-3)	155 Mbits/s and up
Internet-2	100 Mbits/s and up

Communication Networks

An important component in teleradiology is communication networks used for the transmission of images and related data from the acquisition site to the expert center for diagnosis. Since most teleradiology applications are not within the same hospital complex, but through inter-healthcare facilities in metropolitan areas or at longer distances, the communication technology involved requires wide area network (WAN) technology. Wide area network can be wireless or with cables. In wireless WAN, some technologies available are microwave transmission and communication satellites. Wireless WAN has not been used extensively in teleradiology due to its higher cost. Table 5 shows cable technology available in WAN from the low communication rate DS-0 with 56 kb/s, to DSL (Digital Subscriber Line, 144 kb/s to 8 Mb/s, depending on data traffic and the subscription), T-1 and T-3 lines starting from 1.5 Mb/s, to very high broadband communication DS-3 with 45 Mb/s (8). These WAN technologies are available through either a long distance or local telephone carrier, or both. The cost of using WAN is a function of transmission speed and the distance between sites. Thus, within a fixed distance, for a DS-0 line with low transmission rate, the cost is fairly low compared to DS-3, which is much faster, but very expensive. Most of the private lines, for example, T-1 and T-3, are point-to-point and the cost depends on the distance between connections. Table 6 gives an example showing the relative cost of the DSL and the T-1 between University of Southern California and St. John's Health Center ~ 15 miles apart in the Greater Los Angeles Metropolitan Area.

Table 6 demonstrates that the initial investment for the DSL is minimal since the WAN carrier pays for the DSL Modem for the network connection. The lowest

Table 6. WAN Cost Using DSL (144 kB/s–8 MB/s) and T-1 (1.5 MB/s) between USC and St. John's Health Center—20 miles

DSL		T-1	
Up front Investment	Minimal	Up Front Investment	$ 5000
Modems (2)	None	T1 DSU/CSU[a] WAN interface (2)	
		Router (2)	$ 4000
Installation (2)	Minimal	T-1 Installation	$ 1000
Monthly charge: (the lowest rate)	$40	T-1 Monthly Charge:	$600

[a]DSU/CSU: Data service unit/ Channel service unit as of June, 2003.

monthly cost is ~$40/month. On the other hand, for T-1 service, the up front investment is $4000 for the two T-1 routers and $1000 for installation. The monthly cost is $600. The up-front investment for the T-1 is much higher than DSL, and for longer distances, its monthly charge is expensive. For example, the charge between Los Angeles and Washington, D.C. for a T-1 line could be as high as $10,000/month. However, T-1 is a point-to-point private line, and it guarantees its 1.5 MB/s specification, and provides communication security. The disadvantages of DSL are (1) it is through shared networks, and hence has no security; (2) its performance depends on the load of the DSL carrier at that moment; and (3) it is not available everywhere. Using T-1 and DSL for teleradiology is very popular. Some larger IT (Information Technology) companies lease several T-1 lines from telephone carriers and sublease portions of them to smaller companies for teleradiology applications.

Another wide area network listed in Table 5, Internet 2 (I2) technology, is ideal for teleradiology application because of its speed of transmission and low cost of operation after the site is connected to the I2 backbone (9). The I2 network is a national infrastructure of high speed communication backbones [> 10 GB/s using gigabit switches and Asynchronous Mode Technology (ATM)] supported by the National Science Foundation (NSF), currently consisting of the vBNS (very high performance Backbone Network Service), the CalREN-2 (California Research and Education Network), and the Abilene. In the global level, vBNS, Abilene and CalREN-2, provide readily available high speed backbones and administrative infrastructure. In the local level, the users have to learn how to connect the hospital and clinic environments to these backbones. Table 7 shows the current performance of the I2 between some sites in the United States. The advantages of using Internet 2 for teleradiology are its high speed and low operational cost once the local site is connected to the backbones. The disadvantages are (1) the local site has to upgrade its conventional Internet infrastructure to be compatible with the high speed I2 performance, which is costly; (2) not enough experts know how to connect from the radiology department to the backbone; and (3) I2 is not yet open for commercial use.

Table 7. Current Internet 2 Performance between Sites in the United States

Test Sites	Response Time (32 Bytes)	Throughput*
CHLA/USC LAN (Childrens Hospital/ U Southern California)	<1 ms	9.8 MBytes/s
CHLA/USC-UCLA	4 ms	2.7 MBytes/s
CHLA/USC-Stanford U	23 ms	900 KBytes/s
CHLA/USC-UCSF	24 ms	700 KBytes/s
CHLA/USC-NLM (National Library of Medicine)	67 ms	320 KBytes/s
CHLA/USC –U Hawaii	76 ms	700 KBytes/s

*units used are MBytes/s and KBytes/s which are different from those used in Table 5.

User Friendliness

User friendliness includes both the connection procedure of the teleradiology equipment at both the examination site and the expert center, and the simplicity of using the display workstation at the expert center.

User friendliness means that the complete teleradiology operation should be as automatic as possible requiring only minimal user intervention. For the image workstation to be user friendly requires three criteria:

1. Automatic image and related data prefetch.
2. Automatic image sequencing and hanging protocol at the monitors.
3. Automatic look-up table, image rotation, and unwanted background removal from the image.

Image and related data prefetch means that all necessary historical images and related data required for comparison by the radiologist should be prefetched from the patient folder at the imaging sites and send to the expert center. When the radiologist is ready to review the case, these prefetched images and related data are already available at the expert center. Automatic image sequencing and hanging protocol at the display workstation means that all these images and related data are sequentially arranged so that at the touch of the mouse, properly arranged images and sequences are immediately displayed on the monitors. Prearranged data minimizes the time required for the searching and organizing of data by the radiologist at the expert center. This translates to an effective and efficient teleradiology operation. The third factor, automatic look-up table, rotation, and background removal, is necessary because images acquired at the distant site might not have the proper look-up table set up for optimal visual display, images might not be generated in the proper orientation, and might have some unwanted white background in the image due to radiographic collimation. All these parameters have an effect on the proper and efficient diagnosis of the images. Figure 4 shows an example of automatic splitting of a CT examination of both the chest and the abdomen into a chest and an abdomen sequence for automatic display using the concept of presentation of grouped procedures (PGP) in IHE (Integrating the Healthcare Enterprise) profile technology (10).

Image Compression

Teleradiology requires image compression because of the slow speed and high cost of using WAN. For lossless image compression, current technology can achieve between 3:1 and 2:1 compression ratios, whereas in lossy compression using cosine transform based MPEG and JPEG hardware or software, 20:1–10:1 compression ratios can be obtained with acceptable image quality. The latest advance in image compression technology is the wavelet transform (11 and JPEG 2000), which has the advantages over cosine transform for higher compression ratio and better image quality, however, hardware wavelet compression is not yet available. Some Web-based teleradiology systems use progressive wavelet image compression techniques. In this

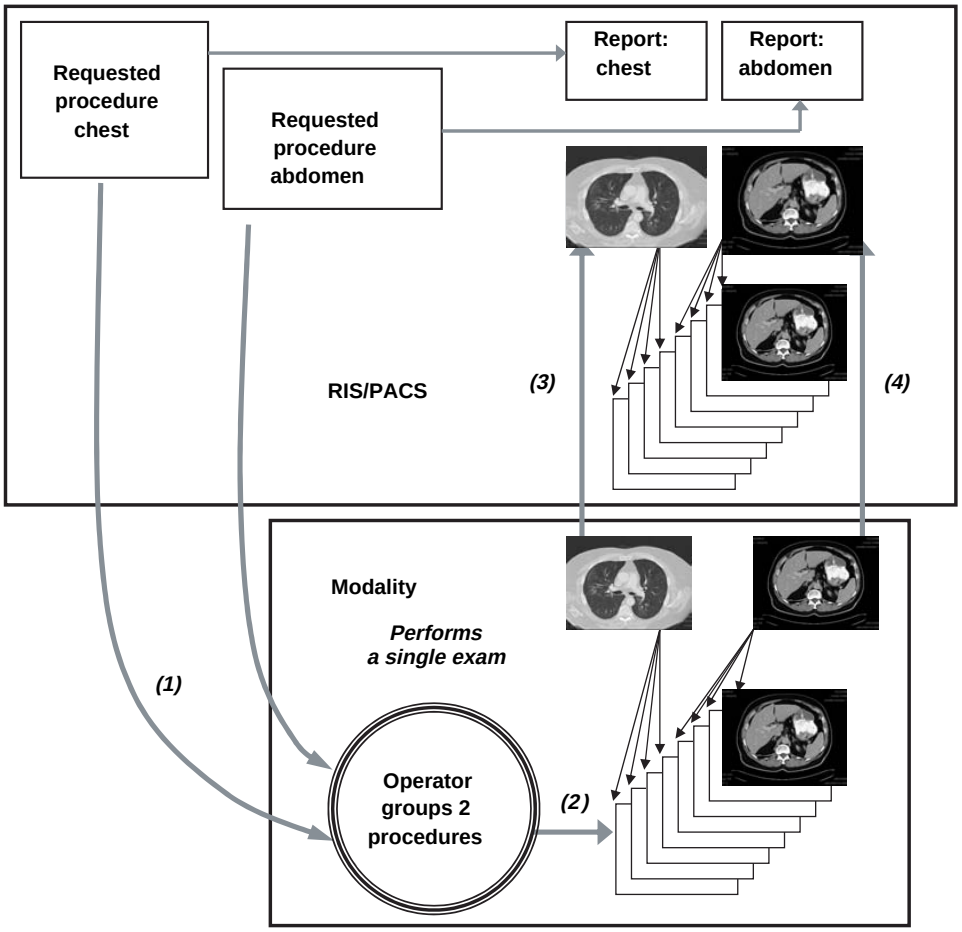


Figure 4. Integrating the Healthcare Enterprise (IHE) Presentation of Grouped Procedures (PGP) Profile. Example shows a single CT acquisition procedure is used to acquire both chest and abdominal scans. The scheduled workflow profile provides information for the PGP to split the acquired images into two subsets, one for the chest and the other for the abdomen (10). Numerals represent the workflow steps in the IHE PGP.

techniques, image reconstruction from the compressed file is in a progressive manner in that a lower resolution image is first reconstructed almost instantaneously and displayed upon request. The user would have the psychological effect that the image is transmitted through the network in real-time. Higher quality images are continuously reconstructed to replace the previous ones until the original image is reconstructed and displayed.

Image Data Privacy, Authenticity, and Integrity

Image transmission in teleradiology is mostly through public networks, for this reason, trust in image data becomes an important issue. Trust in image data is characterized in terms of privacy, authenticity, and integrity of the data. Privacy refers to denial of access to information by unauthorized individuals. Authenticity refers to validating the source of the image. Integrity refers to the assurance that the image has not been modified accidentally or deliberated during the transmission. Privacy and authenticity are the responsibility of the public network provider based on firewall and password technologies, whereas integrity is the responsibility of the end user.

Imaging integrity is mostly done based on the concept of public and private keys digital signature encrypted with mathematical algorithms during the process of image generation. In general, the public and private keys digital signature concept consists of seven steps (12):

1. Private and Public Keys: Set up a method in assigning public and private keys between the examination site and the expert center.
2. Image preprocessing: To segment the object of interest in the image from the background (e.g., the head in a CT image is the object of interest), and extract patient information from the DICOM image header at the examination site while the image is being generated.
3. Image digest: To compute the image digest (digital signature) of the object of interest in the image based on its characteristics using mathematical algorithms.
4. Data encryption: To produce a digital envelope containing the encrypted image digest and the corresponding patient information from the image header.
5. Data embedding: To embed the digital envelope into the background of the image as a further security. The background is used because the embedding would not alter the image quality of the object of interest. In cases where the image has no background, such as a chest radiograph, a more involved lossless embedding technique can been used.
6. The image with the embedded digital envelope is sent to the expert site.

7. The expert center receives Item 6, decrypts the image and the signature. It compares the two digital signatures. One comes with the image, the second is computed from the received image to validate the image integrity.

TELERADIOLOGY OPERATION MODELS

In this section, we discuss four teleradiology operation models that are common in current practice.

Off-Hour Reading

An off-hour reading model is to take care of the off-hour reviewing of the images including evenings, weekends, and holidays when most radiologists are not available at the examination sites. In this set up, image acquisition devices at different examination sites including hospitals and clinics are connected to an off-hour reading center with medium or low grade transmission speed (like the DSL) because the turn around time is not critical except for emergency cases. The connections are mostly direct digital with the DICOM standard. The reading center is equipped with network switches and various types of workstations compatible to the images generated by imaging devices at examination sites. The staffing includes technical personnel taking care of the communication networks and workstations, and radiologists who come in during the evening, weekend, and holiday shifts and perform on-line digital reading. They provide preliminary impression of the exam and transmit it to the examination site instantaneously after reading. The regular radiologists at the examination sites verify the readings and sign off the report the next day. This type of teleradiology set up is low technology but it serves its purpose of solving the shortage of radiologists during off hours.

ASP Model

Application Service Provider (ASP) model is a business venture taking care of the radiological image diagnosis for examination sites where on site radiology interpretations are not available. This model can be for supplying equipment only or for both equipment and radiologists. In the former, an ASP entity sets up a technical center housing network equipment and workstations. It also provides turnkey connectivity for the examination site where images would be transmitted to the center. The examination site can hire its own radiologists to perform reading at the center. In the latter, the center provides both technical support as well as radiologists for reading.

Web-Based Teleradiology

Web-based teleradiology is mostly used by hospital or larger clinics to distribute images to various parts of the hospitals or clinics, or outside of the hospital. A web server is designed where filtered images from PAC systems are either pushed from the PACS server to, or pulled by the Web server. Filtered images mean that the Web server has a predetermined directory to manage the image distribution based on certain criteria like what types of images to where and to whom, and so on. The clients can view these filtered images from the client workstation through the web server. The clients can be referring physicians who just want to take a look at the images or for radiologists to make a remote diagnosis. Web-based teleradiology is very convenient and low cost to set up because most technologies are readily available, especially within the hospital intranet environment. The drawback is that since Web is a general technology, the viewing capability and conditions are not as good as that in a regular PACS workstation where the set up is geared for radiology diagnosis. In order to have full DCIOM image resolution for visualization and manipulation at the clients, modifications have to be made at the Web Server to receive full 12 bits/pixel data from the PACS server, and additional display software at the clients.

PACS and Teleradiology Combined

Teleradiology can function as a pure teleradiology operation shown in Fig. 5. In this operation, the teleradiology management center serves as the operation manager. It receives images from different imaging centers, 1,..., N, keeps a record, but not the images, routes images to different expert centers, 1,..., M according to need for reading. Reports comes back to the management center, it records the reading reports, forwards reports to the appropriate imaging centers. The management center is

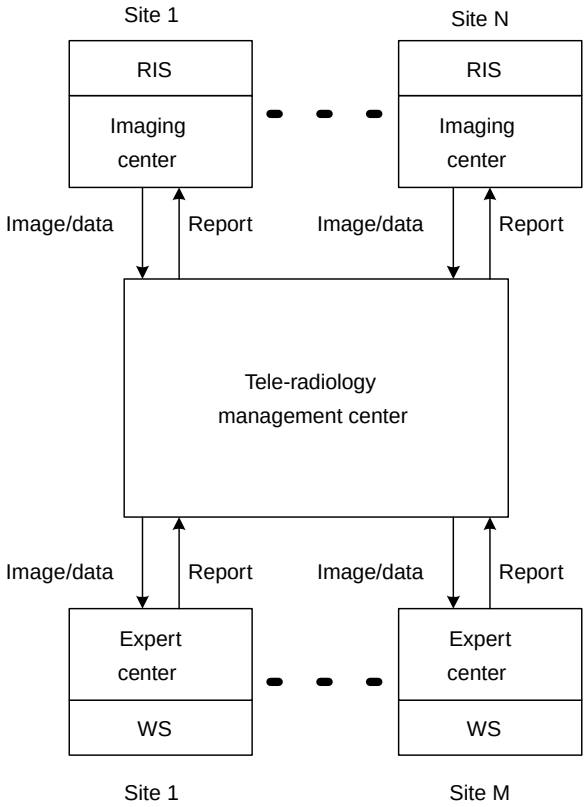


Figure 5. Pure teleradiology model. The management center monitors the operation to direct workflow between imaging centers and expert centers.

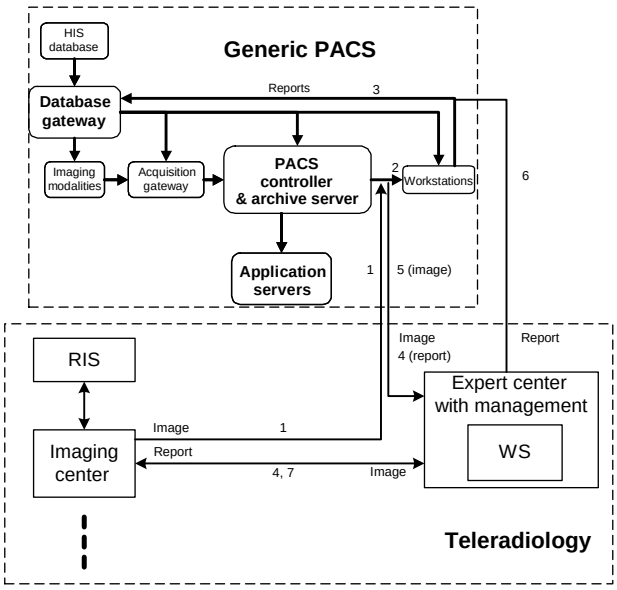


Figure 6. PACS and Radiology Combined Model: The PACS supports either the imaging centers, or PACS and Teleradiology support each other. See text for explanation of workflow steps in numerals.

also responsible for the billing and other administrative functions like image distribution and workload balancing. The networks used for connection between image centers, the management center, and expert centers can be mixed with various performances dependent on the requirements and costs.

Teleradiology can be combined together with PACS as a healthcare enterprise operation, as shown in Fig. 6. The two major components in the combined model are the PACS (see Fig. 3) shown inside the upper dotted rectangle, and the pure teleradiology (see Fig. 5) model shown in the lower dotted rectangle. The workflow of this combined model is as follows:

1. The image center can send images to the expert center for reading as in the pure teleradiology model (7).
2. Radiologists at PACS workstations can read exams from outside imaging centers as well (1).
3. After reading by in-house radiologists from its own workstations (2), reports are sent to the database gateway for its own in-house record (3), or to the expert center from where the report is also sent to the imaging center (4).
4. The PACS can also send exams directly to outside expert center for reading (5). The expert center then returns report to the PACS database gateway (6).

The combined Teleradiology and PACS model is mostly used in a large enterprise level healthcare center with satellite imaging centers, or in back-up radiology coverage between the hospital and imaging centers.

Enterprise Level PACS and Teleradiology Combined with Grid Computing

The PACS and teleradiology combined model described in the last section can be extended to the enterprise level using the grid computing infrastructure. The enterprise level PACS and teleradiology combined model is for very large-scale PAC systems and teleradiology applications. This large-scale model is becoming more and more popular in today's enterprise healthcare delivery system (13).

Grid computing is the integrated use of geographically distributed computers, networks, and storage systems to create a virtual computing system for solving large-scale, data-intensive problems in science, engineering, and commerce (14). A grid is a high performance hardware and software infrastructure providing scalable, dependable, and secure access to the distributed resources. Unlike distributed computing and cluster computing, the individual resources in grid computing maintain administrative autonomy and are allowed system heterogeneity; this aspect of grid computing guarantees scalability and vigor. Therefore, the grid's resources must adhere to agreed-upon standards to remain open and scalable. A popular standard grid computing toolkit is called Globus 3.0 (15). The grid computing provides the user with the following services: computational, data, application, and knowledge. For these reasons, grid computing infrastructure is the ideal technology for large-scale enterprise PACS and teleradiology combined model implementation.

Using the concept of grid infrastructure along with the DICOM standards and IHE workflow profiles, the PACS and teleradiology combined model shown in Fig. 6 can be extended to an enterprise level model conceptually shown in Fig. 7. Grid computing is still in its infancy for medical imaging application. The concept shown in Fig. 7 would require several years before it can materialize.

SOME IMPORTANT ISSUES IN TELERADIOLOGY

Relationship between Teleradiology Technologies and Operation

There are two sets of trade-off parameters in teleradiology. The first set relates to the operation consisting of image quality, reading turn-around time, and cost; and the second set relates to technologies used in the operation including image capture, workstation, compression, communication, and data security requirements. Table 8 shows the relationship between these two sets of parameters.

Image Data Security

In image data security, the patient confidentiality as well as image integrity are important. Since teleradiology uses a public communication method to transmit images that have no security, the question arises as to what type of protection one should provide to assure the patient's confidentiality, and the authentication of the sender. The second issue is the image integrity. After the image is created in digital form, can we assure that the image created has not been altered either intentionally or unintentionally during the transmission? To guarantee patient

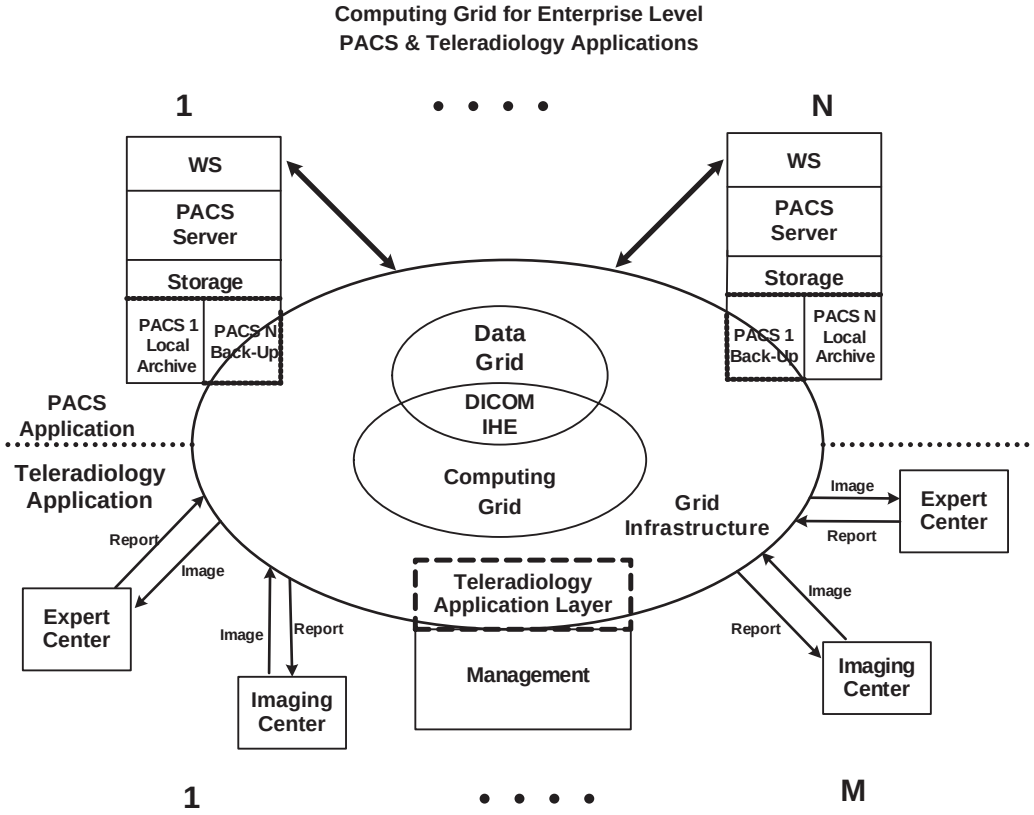


Figure 7. Enterprise level large-scale PACS and teleradiology combined model using the grid computing technology. The center ellipse of the figure is the grid computing infrastructure consisting of the data and computation grid with the DICOM standard and IHE workflow profiles. The data grid handles the image data, and the computational grid takes care of the workflow and management (see Chapter Section 19.2.2.4, Ref (4) for the concept of grid computing and the functions of the data grid.) The top row above the dotted line is the enterprise PACS, which consists of several PAC systems 1,..., *N* rectangles under numerals. The bottom row under the dotted line is the pure teleradiology model described in Fig. 5. The connection between PACS and teleradiology in the enterprise is not a straight line of data communication, instead it goes through the grid computing infrastructure for resource allocation and management, as well as image data acquisition and distribution.

confidentiality and image authenticity, methods such as hardware and software firewalls used routinely in information technology can be set up. To protect image integrity, data encryption and digital signatures can be used. These techniques have been in the domain of defense research for many years, which can be modified for teleradiology application. If high security is imposed on image data, it will increase the cost of decryption and decrease the easy access due to many layers of passwords. The trade-off between cost and performance, confidentiality, and reliability has become a major socioeconomic issue in teleradiology. Since altering a digital image is fairly easy in today's computer technology, developing methods to protect the integrity of image data is essential in teleradiology applications.

Table 8. Relationship between Technologies and Teleradiology Operation

	Image Capture	WS	Compression	Communication	Image Integrity
Image quality	X	X	X		
Turnaround time	X	X	X	X	X
Cost	X	X	X	X	X

Medical–Legal Issues

There are four major medical–legal issues in teleradiology: privacy, licensure, credentialing, and malpractice liability. The ACR (American College of Radiology) Standard for Teleradiology adopted in 1994 defines guidelines for "qualifications of both physician and nonphysician personnel, equipment specifications, quality improvement, licensure, staff credentialing, and liability. Guidelines to these topics, although much is still uncertain, have been discussed extensively by others (16–19). It is important that these issues should be considered thoroughly before a teleradiology operation is set up.

TRENDS IN TELEMEDICINE AND TELERADIOLOGY

The concept of telemedicine and teleradiology originated in the 1970s, however, technology was not ready for real clinical applications for teleradiology until several years ago. As teleradiology is being integrated into daily clinical service, the associated socioeconomic issues discussed also surface. The trends in teleradiology are to balance the cost with the requirements of image quality, and turn-around time for the service. Costs are affected by technology used in image capture, workstation, image compression, communication, and image security. We see that teleradiology will become a necessity in medical practices of the twenty-first century, and will be an integral component of telemedicine as the not so distant method for healthcare delivery.

Teleradiology uses the Web and Internet technologies. Issues that must be resolved immediately are how to lower the communication cost and to bundle textual with image information effectively and efficiently to assure efficient operation, and image security. For the former, Internet 2 appears to be an excellent candidate, and for the latter, ePR (electronic Patient Record) will evolve as a potential winner. [4, Ch 21].

BIBLIOGRAPHY

1. Stahl JN, Zhang J, Zeller C, Pomerantsev EV, Lou SL, Chou TM, Huang HK. Tele-conferencing with dynamic medical images. IEEE Trans Inform Tech Biom 2000;4(1): 88–96.
2. Zhang J, Stahl JN, Huang HK, Zhou X, Lou SL, Song KS. Real-time teleconsultation with high resolution and large volume medical images for collaborative health care. IEEE Trans Inform Tech Biom 2000;4(1):178–185.
3. Stahl JN, Zhang J, Chou TM, Zellner C, Pomerantsev EV, Huang HK. A new approach to tele-conferencing with intravascular ultrasound and cardiac angiography in a low-bandwidth environment. RadioGraphics 2000;20: 1495–1503.
4. Huang HK. March, PACS and Imaging Informatics: Principles and Applications. Hoboken, NJ: John Wiley & Sons; 2004.
5. Huang HK. Teleradiology technologies and some service models. Volume 20, Computerized Medical Imaging and Graphics. 1996. p 59–68.
6. DICOM Standard 2003, http://www.dclunie.com/dicom-status/status.html#BaseStandard2001
7. HL7 Version 3.0: Preview for CIOs, Managers and Programmers, http://www.neotool.com/company/press/199912_v3.htm#V3.0_preview
8. Stahl JN, Tellis W, Huang HK. Network latency and operator performance in teleradiology applications. J Digital Imag 2000;13(3):119–123.
9. Huang HK. 2003, Research trends in medical imaging informatics. In: Hwang NHC, Woo SLY, editors. Frontiers in Biomedical Engineering based on the Proc WCCBME (World Congress for Chinese Biomedical Engineers Proceedings), Chapt. 17, New York: Kluwer Academic Publisher; 2002; p. 269—281.
10. Carr C, Moore SM. IHE: A model for driving adoption of standards. Comp Med Imaging Graphics 2003;Issues 2–3: 137–146.
11. Wang J, Huang HK., Three-dimensional image compression with wavelet transform. In: Bankman IN, editor-in-chief, Rangayyan RM, Woods RP, Robb RA, Huang HK, editors. Hanbook of Medical Imaging. Academic Press; 2000; Chapt. 52, p. 851–862.
12. Zhou X, Huang HK. Authenticity and integrity of digital mammography image. IEEE Trans Medical Imaging 2001; 20(8):784–791.
13. Huang HK. Enterprise PACS and image distribution. Comp Med Imaging Graph 2003;27(2–3):241–253.
14. Bernman F, Fox G, Hey T. Grid Computing. Hoboken, NJ: John Wiley & Sons; 2003.
15. GlobusToolkit 3 Core White Paper, http://www-unix.globus.org/toolkit/documentation.html
16. James AE, Jr, James E, III, Johnson B, James J. Legal considerations of medical of medical imaging. Leg Med 1993; 87–113.
17. Berger SB, Cepelewicz BB. Medical-legal issues in teleradiology. Am J Roentgenol 1996;166:505–510.
18. Berlin L. Malpractice issue in radiology-teleradiology. Am J Roentgenol 1998;170:1417–1422.
19. Kamp GH. Medical-legal issues in teleradiology: A commentary. Am J Roentgenol 1996;166:511–512.

See also COMPUTER-ASSISTED DETECTION AND DIAGNOSIS; RADIOLOGY INFORMATION SYSTEMS.

TEMPERATURE MONITORING

PETER FRASCO
Mayo Clinic Scottsdale
Scottsdale, Arizona

INTRODUCTION

The notion that illness and fever (elevation of body temperature above normal) are linked has been known since the time of Hippocrates and Galen. However, the concept of temperature as a quantifiable vital sign (like pulse rate and blood pressure) that could be measured and recorded is a relatively recent phenomena. Although the thermometer is as ancient as the microscope and older than the stethoscope, its use as an instrument for physical diagnosis in medicine is a relatively recent phenomenon (1).

The thermometer was invented in the fourteenth century, but its use in clinical medicine did not become commonplace until the early twentieth century (2). In 1625, Sanctorio Sanctorius described a device that was used to measure oral temperature. The practice of measuring body temperature with mercury in a glass thermometer began in the early eighteenth century with the work of the Hemann Boerhaave in the Hollan and his students in Vienna (1,2). One of his students, Anton DeHaen, noted changes in temperature with shivering or fever. He also described an increase in heart rate with increased body temperature (1). His contemporaries were unimpressed and the thermometer was largely neglected until the nineteenth century. In 1791, K.A. Wunderlich established the normal range of body temperature from 36.3 to 37.5 °C after recording nearly 1 million readings in 25,000 patients (1,2).

The most plausible explanations for the relative delay from the discovery that body temperature could be measured and monitored to the routine use to the technology relate to the complexity of the early instruments, which were nearly 12 in. (30 cm) long and required ~20 min to record a single measurement. In 1870, T.C. Allbut produced a thermometer that was portable (6 in. (15 cm) long) and reliable (2). It could record a temperature within 5 min. The Allbut thermometer was the forerunner for the mercury in glass thermometer in use presently.

In addition, the late nineteenth century was also an era of intense interest in the specific organ systems and the instruments with which to study them. During this time, the concept of disease shifted away from a holistic model toward a more organ-specific model. Illness was defined by the foci of alterations of function and structure. Overall body heat, as measured by the thermometer, which represented a general phenomena, did not fit easily into this local, organ specific disease concept.

This article will focus upon the technical aspects of temperature monitoring. This emphasis on basic science of the instrumentation of temperature monitoring will be balanced with a discussion of the clinically important topics of heat loss and heat conservation in the surgical patient.

DEFINITIONS

Heat

Heat is the form of energy that is transferred across a boundary of a system at a given temperature to another system at a lower temperature by virtue of the temperature difference between the two systems. This transfer of energy can occur through radiation, conduction, convection, and evaporation. The standard unit of heat (in the International System of Units, SI) is the calorie, which is the amount of energy needed to raise the temperature of 1 g of water by 1 °C. The British thermal unit (Btu), which is not frequently used in medicine, is defined as the amount of heat needed to raise the temperature of 1 lb of water by 1 °F (see discussion of temperature scales below). Heat is transferred from the substance at the higher temperature to the substance at the lower temperature.

Temperature

Temperature is defined as a measurement of the thermal state of a matter, which determines whether it will give heat to another substance, object or energy source, or receive heat from it (3,4). Temperature is a measure of the kinetic energy of the molecules or atoms of a substance. This energy is directly proportional to the velocity of the particles. Temperature will increase as heat energy is added and will decrease as heat energy is lost.

Body Temperature

Body temperature is best defined as the measure of the total kinetic energy within the body. This temperature represents the net thermal effect of total body heat production and heat loss. Body temperature will vary in different parts of the body depending on perfusion, exposure, metabolic activity, local heat gain, or local heat loss. The physiologically important temperature is the "core temperature", which represents the temperature of the body's vital organs (brain, heart, lungs, gut). Any true change in body temperature (ΔT) represents an imbalance in the dynamic between production and loss and can be defined by the following formula (5):

$$\Delta T = \frac{\text{heat production} - \text{heat loss}}{\text{body mass} \times \text{specific heat}}$$

SCALES

Temperature scales are constructed by defining two points based upon a predictable and preferably linear change in the physical property of a given substance at a constant pressure, assigning temperature values to each and then defining the unit of increment between the fixed points. The relationship between temperature and the physical property can be defined as follows:

$$t(x) = ax + b$$

where t is the temperature of the substance. This temperature changes as the property x of the substance changes. The constants a and b depend on the substance used (e.g., mercury, ethanol, copper). The constants are also determined by specifying two points on the scale.

Kelvin

A reading of 1 K [named after William Thompson, a.k.a. Lord Kelvin (1824–1907)] is defined as $1/273.16(3.6609 \times 10^{-3})$ of the thermodynamic temperature of the triple point of pure water (3). The triple point of water is the point at which the solid, liquid, and gaseous phases of water are in equilibrium. Absolute zero, the absence of all heat, when the pressure of the ideal gas is zero, is a temperature of 0 K. The triple point of water in the Kelvin scale is 273 K. Kelvin is the SI unit of temperature.

Centigrade

The centigrade scale was first described by Carolus Linnaeus. The freezing point of water was set at 0 and the boiling point at 100.

Celsius

The celsius temperature scale was originally defined by Anders Celsius. He set the boiling point of water at 0 and the freezing point at 100. As such, the Celsius scale was the reverse of the Centigrade scale. In 1948, the centrigrade scale was replaced with a newly defined Celsius scale that was based upon setting the triple point of water at 0.01 °C and the boiling point of water at 99.975 °C. The single degree increments in the Celsius scale are equal in magnitude to those in the Kelvin scale. To convert from celsius to kelvin the following formula can be used

$$\text{K} = {}^{\circ}\text{C} + 273$$

Table 1. Characteristics of Common Thermometers

Type	Liquid Expansion	Resistance Coil	Thermistor	Thermocouple	Liquid Crystal	Infrared
Accuracy	±0.2 °C	±0.1 °C	±0.1 °C	±0.1 °C	±0.4 °C	±0.3 °C
Sites	Oral, Rectal	Skin	All sites	All sites	Skin	Oral Rectal Otic
Cost	Inexpensive	Expensive	Inexpensive	Moderate	Inexpensive	Very Expensive
Design	Expansion	Electrical	Electrical	Electrical	Chemical	Radiation

Fahrenheit

The fahrenheit [named after Gabriel Fahrenheit (1686–1736)] scale (°F), like the Kelvin scale, is also based upon the triple point and boiling point of water. The scale was originally calibrated with a mercury thermometer using a mixture of salt, ice and water as the zero point. The second point was obtained when the salt was eliminated from the ice and water. This was set at 30 °F. The boiling point of water is 212 °F on this scale. The freezing point of water was adjusted to 32 °F to allow the interval between the triple point and boiling point to be a more rational number (180).

To convert from Fahrenheit to Celsius the following formula can be used

$$°\mathrm{F} = (1.8 \times °\mathrm{C}) + 32$$

TYPES OF THERMOMETERS

A thermometer measures temperature of a system in a quantitative way (Table 1).

NONELECTRICAL METHODS

Changes in Physical Dimensions (See Table 1)

Liquid Expansion Thermometers. The nonelectrical methods of thermometry are loosely based upon the second perfect gas law, Charles' or Guy Lussac's law, which states that at constant pressure the volume of a given mass of gas varies directly with the absolute temperature (4). As heat is added to a substance and temperature increases, the volume of liquids and gases increase. Mercury and ethanol are the most commonly used materials for the expansion based or liquid in glass thermometers.

Ethanol is an alternative to mercury in glass thermometers. Although cheaper than mercury, ethanol thermometers may be unsuitable for high temperatures since ethanol boils at 78.5 °C. Ethanol thermometers have a range from –75 to 120 °C. This is unimportant in the clinical setting of measuring and monitoring body temperature but may be important in other aspects of medicine. In addition, the scale of the ethanol-based thermometer may be less linear than that of the mercury thermometer (3,5).

The design of the liquid expansion thermometer has changed little in last century. The most sensitive location of the liquid in glass thermometer is the bulb, where the largest volume of liquid exists. However, the entire thermometer is temperature sensitive. These thermometers are manufactured with a constriction at the lower portion of the mercury column near the bulb. This prevents the liquid from retracting into the bulb and allows the maximum temperature measured to be displayed until the thermometer is shaken.

Liquid expansion thermometers are impractical for continuous use due to the inherent rigidity of the device, the risk of breakage and the typically slow response (1–3 min) compared to the techniques listed below.

Bimetallic. The sensing element is comprised of two dissimilar metals bonded together in a coil, spiral, or disk. One end of the coil is attached to a lever, the other end is fixed to a point within the device. As temperature changes, the metals expand (with heat) or contract (with cold) by different amounts and the coil tightens or loosens, respectively, and moves a lever over a scale. These thermometers are not very accurate, but are relatively stable over time, require little maintenance, and are inexpensive. Bimetallic thermometer sensitivity to small changes in temperature can be increased by using a long strip wound into a tight coil. The most common clinical application for the bimetallic spring thermometer is use with thermostat devices, temperature alarm devices and operating room temperature monitoring.

The radius of curvature for a bimetallic thermometer is inversely proportional to the difference between the bonding temperature for the strip and the temperature being measured (3,5).

$$R = \frac{t\{3(1+m)^2 + (1+mn)[m^2 + (1/mn)]\}}{6(\alpha_2 - \alpha_1)\,(T - T_0)(1+m)^2}$$

where t = total thickness of the bimetallic strip; m = ratio of thicknesses (low/high expansion materials); n = ratio of Young moduli of elasticity (low/high expansion materials); α_1 = lower coefficient of thermal expansion (1/°C); α_2 = higher coefficient of thermal expansion (1/°C); T = temperature, °C; T_0 = initial bonding temperature, °C.

CHANGES IN ELECTRICAL PROPERTIES (SEE TABLE 1)

Resistance Thermometer

The conductivity of any metal depends on the movement of electrons through its crystal lattice. Resistance to this movement of electrons varies with temperature. Resistance temperature detectors (RTDs) utilize metallic conductors with positive coefficients of resistance or positive temperature coefficients (PTC). As temperature increases, resistance increases almost linearly. The metals that have nearly linear PTC include platinum, tungsten, nickel and nickel alloys. Each metal has a specific resisitivity, ρ, which

varies directly with temperature (3,5–7).

$$\rho_T = \rho_0 [1 + a(T - T_0)]$$

$$\text{Resistance}(\Omega) = \frac{(\rho)L}{A}$$

where L = metal wire length and A = cross-sectional area.

The RTD response relationship is nearly linear and is defined by the following equation (3–7):

$$R = R_0[1 - a(T - T_0) + b(T - T_0)^2]$$

Although the response is nearly linear throughout a wide range of temperature, it is linearity over the smaller temperature interval that is important in clinical medicine. In this scenario, resistance varies according to the following formula:

$$R_t = R_0 (1 + a^* T)$$

Where a and b are calibration constants for resistor material and purity, T is measured temperature, and R_0 reference resistance measured at T_0.

A simple RTD system is comprised of a metal resistor (e.g., platinum wire fashioned into a coil), a source of electrical potential and an ammeter calibrated to indicate temperature. Platinum is the preferred component due to its resistance to corrosion and large positive coefficient of resistance. Adding a Wheatstone bridge with an array of resistors increases the sensitivity of an RTD system.

The RTDs are the most accurate, most stable of the electrical methods for temperature measurement and are nearly linear over a relatively wide range of temperatures. They are, however, slow and expensive compared to thermocouples and thermistors (3,5). In addition, unlike thermocouples, an external current source is required to produce a voltage drop across the sensor. This source of current is a source of self-heating of the RTD if the current is not limited.

Thermistor

A thermistor (Fig. 1)is similar in principle to a resistance thermometer in that the ability to measure temperature depends on the changes in resistance of various metals in response to temperature. There are some important differences. Unlike RTDs, which are made conductors with a positive coefficient to resistance, thermistors are made from semiconductor materials that have a large negative coefficient of resistance. Therefore, as temperature increases, the resistance within the thermistor decreases.

$$\text{Resistance}(\Omega) = A^{(B/\text{absolute temperature})}$$

The relationship between resistance and temperature is nonlinear. In addition, thermistors operate within a relatively high resistance range (1 kΩ–100 kΩ) compared to RTDs.

A digital readout thermistor-based thermometer is illustrated in Fig. 1.

Thermistors are made from heavy metal oxides (cobalt, iron, nickel, manganese, zinc). The metal is shaped into a bead that is sealed into a small measuring tip to which electrodes are attached. The tip can be sealed into a glass tube, cardiac catheter or stainless steel probe. Thermistors are accurate to $\pm$ 0.5 °C over a range of −80 to 150 °C (3). Other advantages of thermistors include relatively small size, low production cost, reproducibility, very high sensitivity, and resolution and relative insensitivity to shock and vibration. Compared to RTDs, however, thermistors are less stable and are more susceptible to internal heating or self-heating. In addition, recalibration may be necessary because the resistance of the metal oxide increases over extended periods of time for unclear reasons. Calibration is defined by the following equation (3,8):

$$\text{Resistance} = R_0 e^{\beta(1/T - 1/T_0)}$$

where R_0 = reference resistance measured at T_0; T = measured temperature; β = material constant.

Thermocouple

When heat is applied to one end of a metallic conductor, electrons at this "hot" junction acquire increased thermal energy relative to those electrons at the unheated or "cold" junction. The electrons diffuse from the hot junction to the cold junction, and in doing so lose this thermal energy. Heat is conducted along the conductor and an

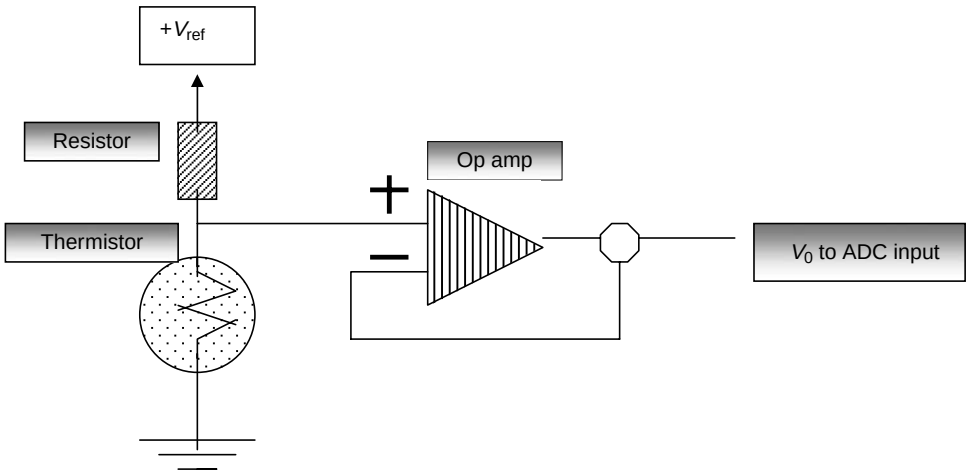


Figure 1. Digital thermistor thermometer ADC.

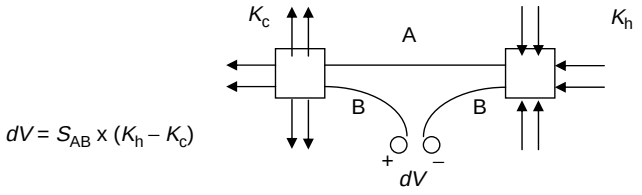


Figure 2. Seebeck principle.

accumulation of electrons at the cold end results in the production of an electric field between the ends of the metal (4). A small voltage is produced at the junction point (in an open circuit) of two dissimilar metals when heat is applied. The electromotive force or voltage across the terminals of an open circuit comprised of two dissimilar metals is proportional to the temperature difference of the two junction temperatures. Platinum, platinum–rhodium, nickel–chromium, and nickel–aluminum can be used as thermocouples. The thermocouple is enclosed in a metal or ceramic shield that protects the device from corrosion and other environmental factors. This voltage does not depend on the temperature along the metals between the junctions.

The thermocouple is based upon the Seebeck principle (Figs. 2 and 3). In 1821, T.J. Seebeck discovered that current is produced in a closed-circuit comprised of two dissimilar metallic conductors when heat is applied to one of the two junctions or if a temperature difference exists between the two junctions. Not every combination pair of metals is acceptable for thermocouple usage, since the direction and magnitude of the current produced is a function of the thermal properties of the metals that comprise the circuit and the temperature difference between the junctions.

In Fig. 2, dV is the voltage difference, K_c is the temperature of the cold junction, K_h is the temperature of the hot junction, S_{AB} is the factor of proportionality.

If the potential difference is to be measured, a second junction is required to produce a complete electrical circuit. This second junction will have its own thermoelectric electromotive force. The electromotive force generated is quantitatively and linearly related to the temperature difference between the two junctions. In order for the thermocouple to be used as a thermometer, this second junction (reference junction, see Fig. 3) must be either maintained at a constant temperature or possess some form of mechanical compensation. The reference junction can be maintained at a constant temperature by immersion in an ice bath or kept at a precisely controlled, constant temperature. Mechanical compensation can be obtained by using a bimetallic spring attached to a coil suspension system. This suspension system functions as the galvanometer. The bimetallic spring is selected according to the range of the instrument. In solid-state instruments, an electrical zero can be used or a thermistor or RTD is used to monitor the reference or cold junction. The signal from the thermistor or the RTD is used to compensate for the cold junction temperature. Figure 3 illustrates a schematic of a thermistor and the relationship between temperature and voltage at the measuring function.

Thermocouples are versatile, inexpensive, small, and durable. They are accurate within a wide temperature range and can be manufactured into a wide variety of physical forms. The low thermal capacity of thermocouples allows for a rapid response time for measurement. An added advantage over RTDs and thermistors is that thermocouples are self-powered. However, compared to the other contact sensors [thermistors, RTDs, and integrated circuit thermometers (ICs)], thermocouples are less sensitive and stable. Other disadvantages include necessity for compensation of the reference junction and, compared to RTDs, a relatively nonlinear voltage to temperature relationship.

Changes in Chemical Phase

Quartz Crystal Thermometry. Liquid crystals (LCs) have the optical properties of a single crystalline solid, but possess the mechanical properties of a liquid. Temperature changes can affect the color of certain liquid crystals. Crystals at specific temperatures, when exposed to light, will scatter some of the light that in turn produces iridescence. This property allows LCs to be used for temperature measurement. Liquid crystals are broadly categorized as either lyotropic or thermotropic. Lyotropic crystals are used in the production of soaps and detergents and react to the type of solvent with which they are mixed (9).

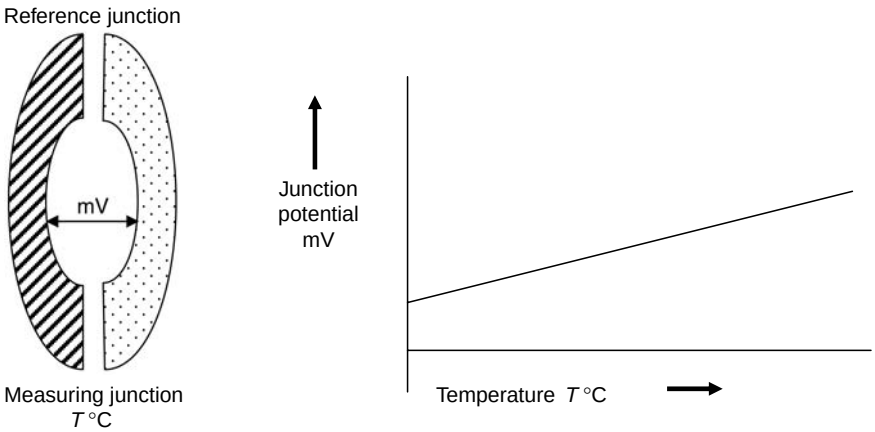


Figure 3. Thermocouple.

Thermotropic liquid crystals react to changes in temperature or pressure (9). Thermotropic crystals are further categorized as either isotropic or nematic (9). Isotropic crystals are random in their arrangement, while nematic crystals have a definite pattern or order. Nematic liquid crystals can be arranged in layers (smectic) or in spirals (cholesteric) (9). Temperature indicators can be constructed by embedding microencapsulated cholesteric liquid crystals into adhesive strips (3,8,9). These crystal devices can be applied to any accessible or visible location. These devices are more useful as trend devices during surgical procedures since they are only accurate to 0.5 °C (3,8). Other disadvantages of LC thermometry include a lack of interfaces to other hemodynamic or respiratory monitors and inherent subjectivity, as colors may be interpreted differently by individual observers.

CHANGES IN EMITTED THERMAL RADIATION (SEE TABLE 1)

Infrared Thermometers

The electromagnetic spectrum is divided into a number of wavelength regions called bands. Light at frequencies less than red, which is at the low frequency edge of the visible portion of the spectrum, are called infrared (IR) and are not visible to the human eye. The IR band of the spectrum covers wavelengths between 0.7 and 1000 μm (4). All bodies with a temperature >0 K radiate energy in the IR band. This heat energy induces electron vibrations that cause electromagnetic emission. This energy travels like light (as an electromagnetic wave or photon) in all directions. The frequency of this emitted radiation is temperature dependent. The amplitude of the emission is dependent on the emissivity of the substance. Emissivity is the ratio of the energy radiated by an object at a given temperature to the energy radiated by a blackbody (perfect radiator) at the same temperature. Emissivity depends on the surface finish, color, aging, and oxidation state of the substance in question. For example, a highly polished metallic object would have a high reflectivity and a low emissivity.

Radiation striking a surface is reflected, absorbed and/or transmitted.

$$\rho_{(\text{reflectivity})} + \varepsilon_{(\text{emissivity})} + \tau_{(\text{transmitivity})} = 1$$

$$\varepsilon_{(\text{emissivity})} = \alpha_{(\text{absorbtivity})}$$

Infrared thermometers measure the amount of IR energy emitted from the object of interest. Contact with the surface or object of interest is not required. Infrared thermometers are composed of a lens (collection of energy emitted from an object), a sensor (thermal, photoelectric or photon detector), and a signal converter (converts thermal energy into an electrical signal). Infrared thermometers have very rapid response times (milliseconds), do not require contact with an object or surface (avoid contamination), are simple to use, and require little if any maintenance.

$$T_{\text{b}} = \frac{[\chi(N_{\text{T}} - N_{\text{T0}}) + T_0^4]}{4}$$

A basic IR thermometer is comprised of four parts: (1) a waveguide that collects and focuses the energy emitted by the target, (2) a pyroelectric sensor that converts the energy to an electrical signal, (3) a microprocessor that adjusts emissivity allowing thermometer calibration to match the emitting characteristics of the object being measured, and (4) a sensor temperature compensation circuit that ensuring that temperature variations within the thermometer are not transferred to the final output. which are fed through the amplifier, multiplexer (MUX), and analogue-to-digital converter (ADC) to the microprocessor for processing and display. The microprocessor that handles emissivity adjustment also performs temperature compensation and calculates the patient temperature.

CLINICAL APPLICATIONS

With the exception of monitoring temperature during anesthesia and surgery or during care in the intensive care unit, temperature is measured and recorded intermittently. During surgery, according to the guidelines established by the American Society of Anesthesiologists (www.asahq.org), the capability to monitor temperature must be readily available. In clinical situations where changes in body temperature are anticipated (long abdominal (10) or thoracic surgical procedures, pediatric) or intended (cardiopulmonary bypass (8,11), neurosurgical procedures), temperature must be monitored. In these situations, temperature is measured continuously with either internal (invasive) or external devices.

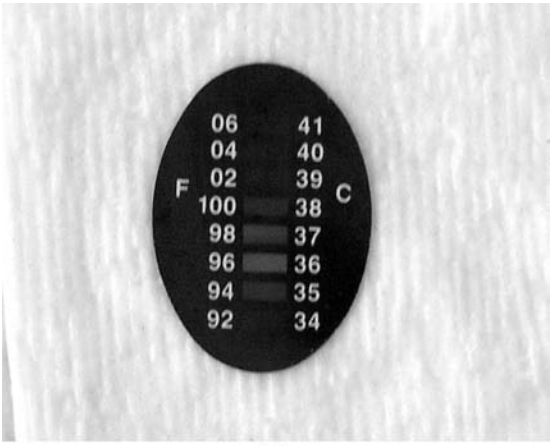
Sites

Accepted measurement sites of core temperature fall into four categories, with distinct advantages, disadvantages, and practical limitations: peripheral (skin); brain (tympanic, nasopharyngeal); visceral (bladder, esophageal, rectal); blood [Pulmonary artery catheter (PAC), esophageal].

There are distinct advantages and disadvantages for each site. These can be broadly categorized into ease of access, risk–benefit analysis, accuracy in measuring core temperature, and precision (5,8,12). For example, the pulmonary artery temperature clearly and accurately reflects core and cardiac temperature (except during the early transition periods during hypothermic cardiopulmonary bypass with extracorporeal circulation) (11), but requires access to the central circulation.

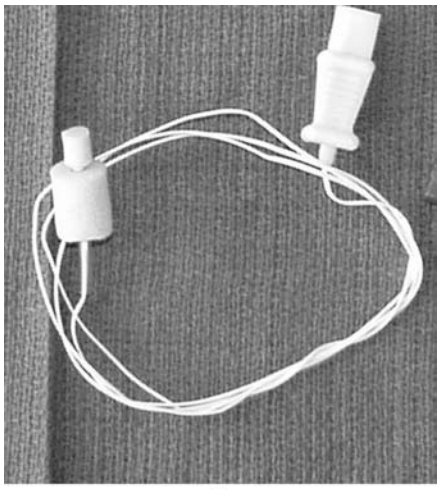
Skin

The skin is easily accessed for temperature measurement, but skin temperature can correlate poorly with core temperature. Disposable LCs or reusable metallic disks attached to thermistors or thermocouples can be used (see Fig. 4a). The LC devices are commonly applied to the forehead where fat distribution is minimal and where regional blood flow is adequate. The reusable probes can be placed in a number of locations including the back, chest or abdominal wall, axilla, or distal extremities. Measurement of skin temperature is also affected by moisture



(a)

(b)



(c)

Figure 4. (a) Liquid-crystal disk and contact-type (skin) thermistor. (b) Infrared otic (tympanic membrane) thermometer. (c) Contact-type thermistor for tympanic membrane temperature.

(sweat) and pressure applied to the probe. An increase in the amount of pressure on the probe will increase temperature measured. The unpredictable relationship between skin and core temperature during surgery and anesthesia makes the skin probe unreliable as an accurate marker for core temperature. Skin probes can be useful as a convenient and essentially risk-free trend monitor (as long as environmental temperature and skin perfusion are constant).

Oral Cavity

The mercury or ethanol in glass thermometers can be used in the oral cavity. As noted previously, the response time is slow relative to other techniques. Thermistors or thermocouples can be manufactured into probes that can be inserted into the oral cavity for rapid, intermittent temperature measurement. In addition, probes can be produced to allow continuous monitoring in the oral cavity. The oral cavity temperature can be up to 1.1 °C lower than simultaneous measurements from the pulmonary artery catheter (5,12).

Tympanic Membrane

The tympanic membrane is a useful and accessible site for IR thermometry. It is a relatively flat, uniformly textured structure that is (1) supplied by branches of the external carotid artery and (2) close to the internal carotid artery, which is the main supply of blood to the brain. As such, it is an acceptable location for core temperature measurement that accurately reflects hypothalamic temperature. The IR probe (see Fig. 4b) (with the aforementioned components) is configured into an otoscope-like device. This probe is covered with a disposable cone-shaped probe cover. The temperature is then displayed on a liquid crystal or light emitting diode display. They are, however, impractical for continuous measurements in medicine due to size, shape, and design of the current devices.

Contact type probes can be produced using thermocouples or thermistors. These probes can be used for continuous measurement of temperature during surgery and anesthesia. The probes are usually manufactured with a foam insulator so the tip of the probe is the only site of contact and measurement. Alternatively, a cotton or felt ball may be inserted to insulate and stabilize a wire probe (see Fig. 4c). There is little lag in the tympanic temperature during core cooling and warming studies, in contrast to most of the other available sites.

Nasopharynx

The nasopharyngeal probes are similar in construction (thermocouple or thermistor) to the tympanic membrane probes. The nasopharyngeal temperature is an adequate marker for brain temperature as long as the probe is placed posterior to the soft palate. The probe is quite uncomfortable in the awake patient. In addition, in the intubated patient, the probe is subject to airflow currents if leakage around the endotracheal tube occurs. These currents can adversely affect the accuracy of the probe in the nasopharynx.

Esophagus

Midesophageal temperature can be measured using a thermistor or thermocouple attached to a flexible probe. The probe can also be integrated with a stethoscope that can be used to monitor heart and breath sounds during anesthesia and surgery.

Position within the esophagus is important in predicting the accuracy of esophageal temperature as a marker of core temperature. Temperature can vary from 1 to 6 °C, depending on the position of the probe in the esophagus. For example, if the probe is inserted in a more distal location (eg., stomach), the temperature measured may be higher than the true core temperature due to warming caused by the normal metabolism in the liver, which is adjacent to the stomach. If the probe is positioned in more proximal position (adjacent to the trachea), accuracy relative to core temperature is affected by cooling secondary to ventilation with room temperature gases via the endotracheal tube. In addition, the probe may be cooled by infusion of cold fluids for irrigation into a chest incision for thoracic or cardiac procedures. The temperature will be affected by the cooling and warming phase of cardiopulmonary bypass with extracorporeal circulation. Optimal location of an esophageal temperature probe in an adult is between 35 and 45 cm from the nostrils.

Urinary Bladder

The bladder temperature correlates well with other measures of core temperature. Bladder temperature is recorded with specially modified urinary catheters. These catheters are commonly used to monitor urine output during surgical procedures or during treatment in the intensive care unit. A thermistor or thermocouple is attached to the patient end of the catheter.

Changes in bladder temperature during hypothermic cardiopulmonary bypass with extracorporeal circulation will occur later than changes in nasopharyngeal, tympanic membrane, and pulmonary artery temperature.

Rectum

The rectal mucosa was previously used as a site for core temperature measurement. It is now clear that the rectum is a relatively peripheral site. As such, it does not accurately reflect core temperature. In fact, at times, rectal temperature may exceed core temperature due to the heat produced by metabolism of the fecal bacterial flora. At other times, the presence of feces in the rectum may insulate the probe from contact with the rectal mucosa. In addition, the rectal veins receive blood from the lower extremities. This blood can have a cooling effect on the rectal temperature.

Rectal probes are seldom used for continuous temperature monitoring due the inaccuracies listed above. In addition, awake patients may find them uncomfortable.

Intermittent rectal temperature measurements are often made in the pediatric population using mercury or ethanol in glass thermometers. In addition, a digital thermistor or thermocouple probe may be used in this location.

Intravascular

The most accurate location for measuring core temperature is the central circulation. The inferior vena cava and the superior vena cava drain into the right atrium, and this mixed-venous blood enters the right ventricle and exits via the pulmonary artery to the pulmonary circulation. The pulmonary artery temperature can be measured with a thermistor that is positioned at the tip of a pulmonary artery (or Swan–Ganz) catheter. This thermistor is actually intended for use in calculation of the cardiac output using the thermodilution technique, display of the temperature is a byproduct of this design.

The pulmonary artery catheter is inserted via an introducer or sheath that is placed into one of the large veins of the body (subclavian, internal jugular, or femoral) typically by percutaneous cannulation. While the temperature of mixed-venous blood accurately reflects true core temperature, there are variables that contribute to the accuracy of any given measure. The pulmonary artery temperature is affected by local heating or cooling during the rewarming phase or the hypothermic phase of cardiopulmonary bypass, the instillation of warm or cold irrigation fluids into the chest, and application of ice or slush to the pericardium. In the absence of these factors, the pulmonary artery temperature is an accurate monitor for core temperature.

Insertion of the pulmonary artery catheter involves significant risks including infection, hematoma, pneumothorax, and arrhythmia and is not appropriate for use solely as a temperature monitor. However, the cases in which pulmonary artery catheters tend to be used, including cardiac surgery, major vascular, and major abdominal surgery, tend also to be associated with the potential for profound changes in core temperature.

Temperature Regulation

The body's core temperature may change without any true change in the total heat content due to redistribution of heat to or from the periphery (7,13–15). This occurs most commonly during surgical procedures due to peripheral dilatation secondary to anesthesia.

The temperature control center is located in the anterior hypothalamus. The hypothalamic regulatory center receives afferent input from peripheral skin receptors and initiates appropriate responses, such as reduced heat loss through skin vasoconstriction, increased heat production through raised muscle tone, or shivering (7,13–15). Impairment of thermoregulatory control can occur due to a number of factors including (7,13–15): anesthetic drugs; hypoxemia; extremes of age; shock (Blood loss or hypovolemia); hypothyroidism; hypoglycemia; malnutrition; extreme exertion; central nervous system (CNS) dysfunction.

Heat production varies by age and gender. At rest, the average male produces $1\,\text{kcal}\cdot\text{kg}^{-1}$ of heat per hour and the average female $0.93\,\text{kcal}\cdot\text{kg}^{-1}\cdot\text{h}^{-1}$ (16,17). Heat production occurs primarily by metabolic activity in the liver and skeletal muscles. Heat production in skeletal muscle increases with voluntary activity (movement and exercise) and shivering (see below). With exercise, heat production can increase to 3–$9\,\text{kcal}\cdot\text{kg}\cdot\text{h}^{-1}$ (16,17). This heat is transferred from the muscles to the blood supply of the muscular bed. This blood then

enters the liver raising core temperature. The specific heat of a human is 0.83 $kcal \cdot kg^{-1} \cdot {}^{\circ}C^{-1}$, with a 70 kg male having to gain 58.1 kcal to raise core temperature by 1 °C (16,17). With basal heat production of only 1 $kcal \cdot kg^{-1} \cdot h^{-1}$, endogenous heat production, if normal, would still raise body temperature 1.2 $^{\circ}C \cdot h^{-1}$; assuming sufficient insulation to prevent any other heat loss from any other mechanisms.

The loss of body heat during anesthesia and surgery is the most significant contributing factor to the development of hypothermia. Heat loss occurs as a consequence of loss through multiple mechanisms, including (10,13,14): evaporative (15–30%); conductive (20–30%); convective (15–30%); radiant (30–50%).

Evaporation is the conversion of a liquid into a vapor at a temperature below its boiling point. Evaporation is always accompanied by a reduction in temperature. Molecules of water at the skin surface and mucosal surfaces in the airways and viscera with enough heat energy to overcome the cohesion of neighboring molecules will vaporize. The average energy of the remaining water molecules will decrease reducing the surface temperature. The intraoperative evaporative loss occurs via the skin, the surgical incision, and, most importantly, the lungs, secondary to controlled ventilation of cold, dry gases. Evaporative losses are dependent on the surface area of the exposed surface, minute ventilation, relative humidity, and airflow velocity. These losses can be reduced by increasing the humidity of respiratory gases. The anesthetic drugs, which produce peripheral and cerebral vasodilatation, certainly exaggerate the normal evaporative skin loss of heat and prevent peripheral vasoconstriction to conserve heat. However, sweating, which increases the amount of moisture available, increases evaporative losses. It is estimated that a general anesthetic reduces the shivering threshold to < 34.5 °C, and muscle relaxants will completely abolish this protective response (13,14).

Conduction is the transfer of heat energy by random atomic or molecular motion between two objects. The transfer always occurs down the temperature gradient from the warmer surface to the colder substance. Conductive loss occurs from (1) placing the patient on the room temperature operating room table and (2) patient contact with the room temperature surgical instruments (18). Conduction is less important than radiation, evaporation, and convection clinically since the layer of air that surrounds a body insulates against heat loss unless air movement is present (see Convection below).

The administration of intravenous fluids, cold blood bank products, as well as surgical wound irrigation provides additional significant conductive heat losses (19). The thermal stress of infusing 1 L of unwarmed crystalloid is ~17 kcal and the stress from a liter of unwarmed bank blood is ~30 kcal.

Calculation of the thermal stress from intravenous fluids or transfusions can be made with the formula:

$$\text{IV "Lost" kcal} = (T_c - T_f) \times V$$

Where T_c = body temperature, T_f = fluid temperature, and V = volume of fluid infused in liters. Hence, the administration of 3 L of room temperature (20 °C) crystalloid to a patient represents a 61 kcal challenge whose core temperature is 37 °C ([37–20 °C] × 3 L).

Convection is the transfer of heat secondary to air currents. Air adjacent to the skin is warmed by conduction. This warmed is less dense and rises. As air flows over the body, the current carries heat away from the body. Convective heat flux is defined by the following equation (4,5,20):

$$Q_{cv} = \gamma(T_s - T_g)$$

where γ is the proportionality constant ($J \cdot g^{-2} \cdot K^{-1}$), T_s is the surface temperature, and T_g is the air temperature.

Forced convective heat loss is caused by gas flow caused by external means (fan or pump). Free convective loss occurs due to the gas flow that occurs secondary to temperature differences (4,5,18,20).

Clinical convective losses occur secondary to the requirement of maintaining the air conditioner and/or the air handling system at the colder settings due to multiple layers of sterile clothing worn by the surgical staff. In addition, the rates of operating room air turnover may be significantly higher than in other ambient settings. This increase in air turnover serves to reduce the likelihood of infection, but increases convective losses. Convective losses can be reduced by trapping the layer of air between the skin and external environment with a barrier, such as a thermal blanket or forced air device (21).

Radiation heat loss occurs due to the transfer of heat in the form of electromagnetic energy. The magnitude of heat transfer is dependent on the surface area of the emitting object (4,13,18). It is not dependent on the presence of a material medium therefore no direct contact is required.

SUMMARY

Iatrogenic hypothermia predisposes the patient to profound physiological consequences, including delayed recovery from anesthesia, increased oxygen consumption, increased vascular resistance, cardiac instability and potential ischemia or infarction, coagulopathy, diminished patient satisfaction, and increased recovery room costs. In considering the need to provide the best patient outcomes, hypothermia induced complications can and should be prevented. Prevention of hypothermia requires monitoring of body temperature. The devices that can be used vary in accuracy, convenience and degree of invasiveness. Appropriate choices will influence patient outcome.

BIBLIOGRAPHY

1. Lyons A, Keiner M. The seventeenth century: anatomical and physiological advances: The thermometer. In: Lyons A, Petrucelli R, editors. Medicine: An Illustrated History. New York: Abradale Press; 1987. p 437–439.
2. Pearce J. A brief history of the clinical thermometer. QJM 2002;95(4):251–252.
3. Liptak B. Temperature Measurement. 3rd ed. Radnor (PA): Chalton Book Company; 1993. p 134.
4. Halliday D, Resnick R, Walker J. Fundamentals of Physics. 7th ed. Hoboken (NJ): Wiley; 2005.
5. Parbrook G, Davis P, Parbrook E, editors. Basic Physics and Measurement in Anaesthesia. 3rd ed. Oxford: Butterworth-Heinemann Ltd.; 1990. p 344.

6. Chang H. Inventing temperature: measurement and scientific progress. Oxford: Oxford University Press; 2004. p 286.
7. Szocik J, Barker S, Tremper K. Fundamental principles of monitoring and instrumentation. In: Miller R, et al., editors. Anesthesia. Philadelphia: Churchill Livingstone; 2000. p 1053–1077.
8. Cork R. Temperature monitoring. In: Blitt C, Hines R, editors. Monitoring in Anesthesia and Critical Care Medicine. New York: Churchill Livingstone; 1995. p 543–556.
9. Collins P. Liquid Crystals: Nature's Delicate Phase of Matter. 2nd ed. Princeton: Princeton University Press; 2002. p 204.
10. Joachimsson P, Hedstrand U, Tabow F, Hansson B. Prevention of intraoperative hypothermia during abdominal surgery. Acta Anaesthesial Scand 1987;31(1):330–337.
11. Phillips P, Skov P. Rewarming and cardiac surgery: a review. Heart Lung 1988;17(5):511–520.
12. Cork R, Vaughn R, Humphery L. Precision and accuracy of intraoperative temperature monitoring. Anesthesia Analg 1983;62:211–217.
13. Sessler D. Mild perioperative hypothermia. N Engl J Med 1997;336:1630–1637.
14. Sessler D, Rubinstein E, Moayeri A. Physiologic responses to mild perianesthetic hypothermia in humans. Anesthesiology 1991;75(4):594–610.
15. Sessler D, Schroeder M. Heat loss in humans covered with cotton hospital blankets. Anesth Analg 1993;77:73–77.
16. Guyton A, Hall J. Textbook of Medical Physiology. 10th ed. Philadelphia: Saunders; 2000. p 1064.
17. Schafer J. Body temperature regulation. In: Johnson L, editor. Essential Medical Physiology. San Diego: Elsevier; 2003. 921–932.
18. Fallacaro M, Fallacaro N, Rachel T. Inadvertent hypothermia. Etiology, effects and prevention. AORN Journal 1986;44(1): 54-7–60-1.
19. Evans J, et al. Cardiovascular performance and core temperature during transurethral prostatectomy. J Urol 1994;152: 2025–2029.
20. Bejan A. Convection Heat Transfer. 3rd ed. New York: Wiley; 2004. p 728.
21. Augustine S. Hypothermia in the post anesthesia care unit. J Post Anesthesia Nursing 1990;5:254–263.

See also BIOHEAT TRANSFER; INTEGRATED CIRCUIT TEMPERATURE SENSOR; MONITORING IN ANESTHESIA.

TEMPOROMANDIBULAR JOINT. See TOOTH AND JAW, BIOMECHANICS OF.

TENDON, PROPERTIES OF. See LIGAMENT AND TENDON, PROPERTIES OF.

TENS DEVICES. See TRANSCUTANEOUS ELECTRICAL NERVE STIMULATION (TENS).

THERMISTORS

PEDRO LOPES DE MELO
State University of Rio de Janeiro
Brazil

INTRODUCTION

The normal body temperature of human beings is considered to be constant ~37 °C. Heat energy is stored in the body and it is essentially constant, as long as we are alive. Most of the heat produced in the organism comes from deep organs, mainly the liver, brain, and heart, as well as the skeletal muscles during exercise (1). A major part of this heat is dissipated at the skin surface by radiation, convection, evaporation, and respiration. However, as the environment temperature varies frequently, the body presents mechanisms to generate and loose heat, controlling and regulating the body temperature. It is done almost exclusively using feedback mechanisms that operate by means of regulatory centers located at the hypothalamus. Thus, if the environment temperature decreases, the body generates more heat and tries to keep it, while the heat generation is decreased and part of the body heat is dissipated when the environment temperature increases. Disease states introduces modifications in this equilibrium that can be indirectly evaluated by body temperature. That is the reason why it is one of the more important physiological parameter, measured virtually in all patients in every hospital bed, being used as a clinical index of disease, as well as an aid for diagnostic and prognostic purposes. Temperature can be measured at the human body in two basic ways: at skin surface and by systemic measurements. Although, in practice, systemic measurements are much more commonly used, both kinds provide valuable diagnostic information. Skin temperature is the result of the relationship between the heat supplied by blood surface circulation, the environmental temperature, and the air circulation around the area at which the measurement is being taken. Skin temperature data are usually obtained from sensors in contact with the skin's surface. However, the use of self-adhering foam patches to affix temperature sensors to the skin appears to procedure artificially higher mean temperatures (2). Systemic temperature is the temperature of the internal regions of the body. This parameter can be measured by temperature sensing devices placed at the mouth, in the rectum, or under the armpits. The oral temperature of a health person is ~ 37 °C. The rectal temperature is typically 0.5 °C higher than the oral temperature, while the underarm temperature is about 1 °C lower than the oral one. Some clinical applications may need more specific measurements, as skull and core temperatures, for example. Skull temperature may be obtained by placing the thermal probe into the nasopharynx near the base of the skull, while core temperature may be obtained from a probe inserted into the esophagus. The basic characteristics of devices used in temperature measurements are ease of operation, cleaning and sterilization, and guarantee of patient safety. It is also important to mention the size and thermal mass of the probe, since these determine the disturbance imposed by the measurement and the speed of response.

The simplest phenomenon for temperature sensing is, perhaps, thermal expansion, which is the basis of the mercury-in-glass thermometer. In clinical routine measurements, however, electronic temperature recording and display are advantageous, since they permit the use of automatic continuous measurement of temperature and remote applications by connecting to computerized systems. Another convenient characteristic of these devices is their small size. There are numerous ways of measuring temperature electronically, most commonly by transducers

based on temperature dependence of electric resistance (RTDs and thermistors) and thermoelectric effects (thermocouples). Thermistors offer a high sensitivity and degree of interchangeability at lower cost than either RTDs or thermocouples. It makes them ideal for healthcare products that incorporate sensor probes that can be discarded after using and replaced with new probes of the same specification, without recalibration. In this section, the attention will be focused on thermal-sensitive resistors, their properties, and the basic instrumentation used in temperature measurements. Examples of commercial devices dedicated to temperature measurements are also discussed, as well as systems for clinical and research applications in several branches of medicine to measure flow, detect the presence of fluids, and evaluate the properties of tissue based on temperature measurements using thermistors.

THEORY

Thermistor (the contracted name of thermally sensitive resistor) is a general term used for both positive and negative temperature coefficient types of semiconducting thermal transducers. These devices are constructed of ceramic materials whose electrical conduction properties are temperature sensitive. At a fixed environmental temperature, a thermistor exhibits a specific ohmic value. However, if the environment temperature varies, this resistance changes. Thermistors are able to sense temperatures from -50 to 300 °C, which is a small range compared with metal wire sensors. A platinum wire, for example, can be used to measure temperature from -160 to 1000 °C. This range, however, is much greater than most of the medical temperature measurement tasks would require.

Thermistors usually have high negative temperature coefficients (NTC), resulting in a decrease of the thermistor resistance with increasing temperature. The negative temperature coefficient in nonmetallic materials (silver sulfide), was first observed by Michael Faraday in 1833, but it was not until 1940 that thermistors where developed that were able to produce reproducible results. The NTC thermistors are typically made of transition-metal oxides. The most usual oxides are those of manganese, nickel, cobalt, and iron (3,4). In the basic fabrication process, a mixture of two or more metal oxide powders are first combined with suitable binders, and are molded to a desired geometry. Then the products are dried and sintered at an elevated temperature. The units are finally coated with an epoxy layer for final protection and stabilization. Varying the types of oxides used, their relative proportions, the sintering atmosphere, and the sintering temperature, a wide range of resistivities and temperature coefficient characteristics can be obtained. More detailed descriptions of NTC manufacturing techniques can be obtained in Refs. 3–5. Since their first use in practical electronic thermometer in the early 1950s, thermistor technology has been enhanced continuously, resulting in improvements that probably situate this device as the most widely used temperature transducer for medical applications nowadays. These devices have been developed to be very sensitive to variations in temperature (~ -3 to -5 °C), present excellent long-term stability ($\pm 0.2\%$ of nominal resistance value per year), and be small in size. Because of this small size, these sensors present a fast response to variations in fluid temperature. Thermistors are also relatively inexpensive.

Beyond the advantages of high sensitivity, interchangeability, and low cost cited previously, another major advantage offered by thermistors is that, unlike RTDs and thermocouples, thermistors are virtually unaffected by lead resistance, since they are high resistance devices. Specifically comparing them with thermocouples, we observe that thermistors can be used with less complex and expensive instruments. It happens because the thermal electromotive force (EMF) values produced by thermocouples are around a few microvolts per degree, requiring high gain and low noise amplification of the signal. Moreover, thermocouple demands additional circuits for compensation of cold junction temperature (4). There are many physical configurations in which thermistors are found, varying from very small bead thermistors, that are spherical and have diameters as small as 0.1 mm, to large flat disks having diameters of several centimeters. Some of these configurations are described in Fig. 1.

Thermistors with positive temperature coefficient (PTC) may also be constructed, by sintering barium and strontium titanate mixtures. These thermistors are often called switching thermistors because of their resistance–temperature characteristics. As temperature increases, the zero-power resistance of this device remains essentially constant until reaching the switching temperature or Currie point, where there is a sharp upward increase in the resistance. The switching temperature can be from -20 to $+125$ °C. The PTC thermistors are frequently used as thermostats to sense and regulate oven temperature (6) and for circuit protection.

THERMISTOR TERMINOLOGY

Dissipation constant (DC or δ) is the amount of power required to raise the temperature of the thermistor 1 °C above the surrounding temperature in steady-state conditions (6). It means that the resistance changes by an equivalent of 1 °C for each dissipation constant rating ($\mathrm{mW}\cdot{}^{\circ}\mathrm{C}^{-1}$) for the selected device. It depends on the heat transfer from the thermistor to its surroundings (by conduction through the leads, free convection in the medium and radiation), the relative motion of the medium in which the thermistor is located, as well as the thermal conductivity. A typical value of the dissipation constant of a thermistor with a 0.24 cm outer diameter, coated with epoxi or phenolic layers, is 2 $\mathrm{mW}\cdot{}^{\circ}\mathrm{C}^{-1}$ in still air (7). Its parameter increases with thermistor mass and in water is $\sim$5–10 times the value measured in air.

Maximum operating temperature is the maximum body temperature at which the thermistor will operate for an extended period of time with acceptable stability of its characteristics. This temperature can be the result of internal or external heating, or both, and should not exceed the maximum value specified (6).

Self-heating is a process observed when a current flowing through a thermistor causes sufficient heating

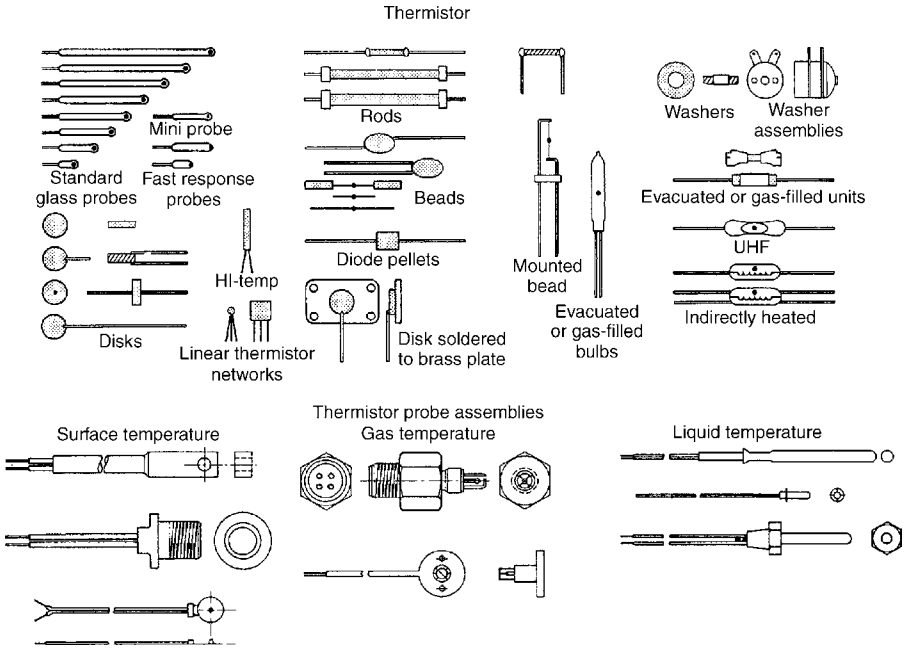


Figure 1. Some thermistor configurations (from Interfacing Sensors to the IBM PC, by W. J. Tompkins and J. G. Webster, 1988 Reprinted by permission of Prentice-Hall).

by power dissipation to raise the thermistor's temperature above the ambient (6). As the effects of self-heating are not always negligible (or may even be intended), a distinction has to be made between the characteristics of an electrically loaded thermistor and those of an unloaded thermistor. The properties of an unloaded thermistorare also termed as zero-power characteristics. This effect depends on the thermal dissipation factor (δ) and the geometry of the thermistor itself. In general, the smaller the device, the smaller is the permissible current before self-heat.

Stability is the ability of a thermistor to retain specified characteristics after being subjected to designated environmental or electrical test conditions ($^{\circ}C{\cdot}year^{-1}$). This parameter is dependent on environmental conditions (e.g., humidity, excessive temperature, and thermal shock), which should be minimized to guarantee stability (7). Physical reasons for this may be thermal stress causing a change in concentration of lattice imperfections, oxygen exchange with the environment (with unprotected, nonglass-encapsulated thermistors), or diffusion in the contact areas of metalizalized surface contacts. To enhance long-term thermistors stability, they are usually subjected to an ageing process directly after manufacture (8). This results in chemically stable devices that are not significantly affected by aging, exhibiting typically $<0.02\,^{\circ}C{\cdot}year^{-1}$ of thermometric drift.

Zero-power resistance (R_0) is the direct current (dc) resistance value of a thermistor at a specified temperature, with negligible electrical power to avoid self-heating (6).

Temperature coefficient of resistance (α) is a useful measurement of the thermistor's sensitivity and is defined as the relative change in resistance referred to change in temperature at a specified temperature ($\%{\cdot}^{\circ}C^{-1}$) under zero-power conditions (6,8).

$$\alpha = \frac{1}{R_T}\frac{dR}{dT} \tag{1}$$

Because the relationship between resistance and temperature is not linear, α is a function of temperature and it is usually specified over the temperature range where the temperature variation is expected. One of the most important characteristics of a thermistor is, without question, its extremely high temperature coefficient of resistance. A typical value is $-4.3\%{\cdot}^{\circ}C^{-1}$ at 37 °C.

Thermal time constant (TC or τ) describes the heat inertness of thermistors (5) and it is the time (s) required for a thermistor to change 63.2% of the total difference between its initial and final body temperature, when subjected to a step function change in temperature under zero-power conditions. This parameter is directly affected by the mass of the thermistor, the thermal properties of the medium surrounding the device, the thermal coupling to the environment, the motion of the medium, the conduction through the leads, and the radiation losses. An epoxy coated thermistor with a 0.24 cm outer diameter will typically have a time constant of 0.75 s in stirred oil and 10 s in still air (7).

Interchangeability tolerance: The rated thermistor parameters values are subject to manufacturing tolerances. They are determined by the composition and structure of the various metal oxides being used in the device production. The result will be a variation from unit to unit within a production lot, as well as from lot to lot. Interchangeability tolerance is the value of how far a specific family of devices may be from the nominal resistance versus temperature curve. For example, if a family of devices has an interchangeability tolerance of ±1.0 °C over the range from 0 to 70 °C, it means that, for this range, all devices of this family are within ±1.0 °C of the resistance versus temperature curve. This feature results in accurate temperature measurements to ±1.0 °C, no matter the substitution of thermistors. Modern thermistor technology results in the production of devices with tight zero-power resistance tolerances. Over the medical temperature range, interchangeable tolerances to ± 0.1 °C are typical (9,10).

Maximum power rating is the maximum power (mW) that a thermistor will dissipate for an extended period of time with acceptable stability of its characteristics (11).

Standard reference temperature is the thermistor body temperature at which nominal zero-power resistance is specified and is usually 25 °C (11).

RESISTANCE–TEMPERATURE CHARACTERISTICS

Thermistors are one member of the family of resistive temperature sensors. Unlike the other members of this family, which presents linear relationship between resistance and temperature (platinum, nickel and cooper, e.g.), thermistors are very nonlinear. This can be seen in Fig. 2, where the characteristics of some typical NTC thermistors and a platinum RTD are compared, showing the higher sensitivity to temperature changes of the first ones (3). In thermistor literature, the most frequently used characteristic relationship between the thermistor resistance, and the ambient temperature is

$$R_T = R_0 \exp\left[\beta\left(\frac{1}{T} - \frac{1}{T_0}\right)\right] \tag{2}$$

where R_T is the thermistor resistance (Ω) at temperature T, R_0 is the thermistor resistance (Ω) at temperature T_0, which is the standard reference temperature (K, K = °C + 273.15), and β is the material constant for thermistor (K). A typical thermistor resistance may vary from 5000 Ω at 0 °C, until 100 Ω at 150 °C, while the reference temperature T_0 is usually taken as 298 K (25 °C). Equation 2 can be rearranged to solve for β:

$$\beta = \frac{T\,T_0}{T - T_0} \ln \frac{R_0}{R_T} \tag{3}$$

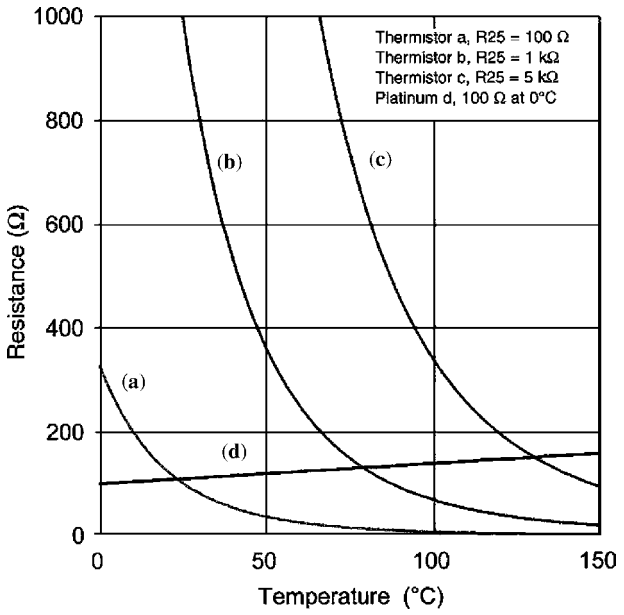


Figure 2. Resistance–temperature relationship of thermistors (a,b,c) and RTD (d). Over the presented range, NTC thermistors offer higher sensitivity to temperature changes compared to RTD. (Reproduced from (3) with permission).

The constant β is usually between 2000 and 5000 K and increases slightly with temperature. Its value is determined by the material properties and it is usually calculated and specified by thermistor manufacturers using two temperatures over a given range. Applying the definition of the temperature coefficient of resistance (eq. 1) in equation 2 results in

$$\alpha = -\frac{\beta}{T^2} \tag{4}$$

It is important to point out that equations 2–4 are valid only over small temperature spans. The Steinhart–Hart equation is, however, more accurate over wider temperature ranges.

$$\frac{1}{T} = a + b(\ln R_T) + c(\ln R_T)^3 \tag{5}$$

where T is the temperature in K, a, b, and c are the coefficients derived from measurements. The Steinhart–Hart equation is an empirically developed polynomial that has been determined to be the best mathematical expression for resistance–temperature relationships of NTC thermistors and probes (12). Solving for resistance, when temperature is known, the form of the equation changes to

$$R_T = e^{\left[\left(-\frac{\chi}{2} + \left\{\frac{\chi^2}{4} + \frac{\psi^3}{27}\right\}^{1/2}\right)^{1/3} + \left(-\frac{\chi}{2} - \left\{\frac{\chi^2}{4} + \frac{\psi^3}{27}\right\}^{1/2}\right)^{1/3}\right]} \tag{6}$$

where, $\chi = (a - 1/T)/c$ and $\psi = b/c$. The a, b, and c coefficients, can be solved measuring the thermistor resistance (R_1,R_2,R_3) at three different temperatures (T_1,T_2,T_3) and using simultaneously equation 5. The data calculated by equations 5 and 6 will be accurate to better than ±0.01 °C, when $T_1 \leq -40$ °C, $T_3 \leq 150$ °C, $|T_1 - T_2| \leq 50$ °C and $|T_2 - T_3| \leq 50$ °C. Parameters T_1, T_2, and T_3 are evenly spaced and at least 10 °C apart.

VOLTAGE–CURRENT CHARACTERISTICS

If a constant electrical power is applied to the thermistor its temperature will first increase considerably, but this change will decline with time. After some time a steady state will be reached, where the power is dissipated by thermal conduction or convection. The voltage drop on the thermistor as a function of the flowing current under conditions of thermal equilibrium is

$$V\,I = \delta(T_T - T_A) \tag{7}$$

If the dissipation constant (δ) variations are negligible for a determined medium and established conditions, and the resistance–temperature characteristic is known, the static current–voltage characteristics can be obtained. Since $V = R_T I$, where R_T is the temperature dependent NTC resistance,

$$I = \sqrt{\frac{\delta(T_T - T_A)}{R_T}} \tag{8}$$

and

$$V = \sqrt{\delta(T_T - T_A) R_T} \tag{9}$$

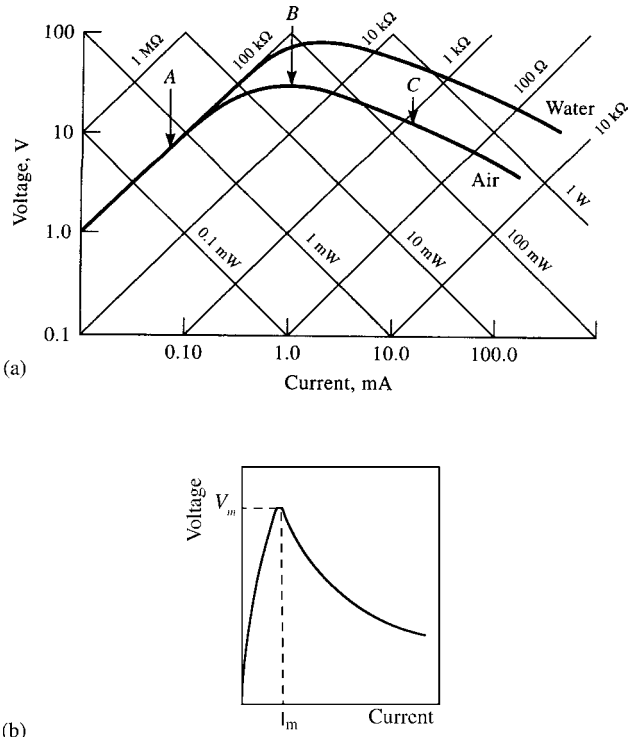


Figure 3. Thermistor voltage versus current characteristics with the device in air and water on log-log coordinates (A), and characteristics on linear coordinates (B). Figure (A) is from Interfacing Sensors to the IBM PC, by W. J. Tompkins and J. G. Webster, 1988 (Reprinted by permission of Prentice-Hall) and Figure (B) is from Ceramic Sensors: technology and applications, T. G. Nenov and S. P. Yordanov, 1996, (courtesy of Technomic Inc, Pennsylvania, USA).

using the above equations the voltage–current curves can be plotted on log–log coordinates, where lines of constant resistance have a slope of $+1$ and lines of constant power have a slope of -1 (Fig. 3a) (6,13). For some applications, it is more convenient to plot the voltage–current characteristic on linear coordinates (Fig. 3b) (5,13).

The voltage–current characteristics of a NTC thermistor has three different sections: first, for small currents, the amount of power dissipated in the thermistor is negligible, and the voltage–current characteristic will describe a constant resistance that is equal to the zero-power resistance of the device at the specified ambient temperature (section A in Fig. 3a). In this condition, the temperature of the thermistor is that of the surroundings. This curve section is usually used when thermistors are employed as temperature sensors. As the current continues to be increased, in the second curve section, there is a rise in the thermistor temperature above that of the surrounding medium since this power cannot be completely removed from the thermistor and the effects of self-heating become evident, resulting in a decrease in its resistance (section B in Fig. 3a). In this condition, more current flows due to decreasing resistance (6). Subsequent incremental increases in current introduce a corresponding decrease in resistance. Hence, the slope of the voltage–current characteristic ($\Delta E/\Delta I$) decreases with increasing current. This continues until a current value is reached (I_m), for which the slope becomes zero and the voltage reaches a maximum value (V_m) (Fig. 3b). As the current is increased above the value of I_m, the third curve section (section C in Fig. 3a) is entered. The slope of the characteristic curve continues to decrease and the thermistor exhibits a negative resistance characteristic (5). This last section of the operating area of NTC thermistors, when they are self-heated, is used in applications such as liquid level detection, air flow detection, and thermal conductivity measurement. However, care should be taken when designing a circuit for self-heated application, because the thermistor is vulnerable to destruction in this region by thermal runaway. This can be avoided by passing constant currents coherent with the thermistor dissipation characteristics. As the power dissipated is proportional to I^2R, when the current is constant, further increases in temperature causes decreased resistances and decreased power, protecting the device.

Anytime a thermistor is applied for temperature measurement, it is very important to avoid its self-heat. Since a thermistor resistance changes with the temperature, this self-generated heat will change this resistance value, producing an erroneous reading. For example, if the dissipation constant of the thermistor selected is $5\,\text{mW}\cdot^\circ\text{C}^{-1}$ and the power dissipated by the device is $20\,\text{mW}\cdot^\circ\text{C}^{-1}$, then a $4\,^\circ\text{C}$ error is induced due to the effect of self-heating. This effect is more pronounced when dealing with a still fluid (i.e., neither flowing nor agitated), because there is less carry-off of the heat generated. This kind of problem does not arise with thermocouples, that are essentially zero-current devices. To maintain a higher degree of accuracy, the temperature error caused by self-heating should be an order of magnitude less than the required sensor accuracy. As an example, if the power dissipation constant of a thermistor is $\sim 2\,\text{mW}\cdot^\circ\text{C}^{-1}$ in still air, in order to keep the self- heat error $<0.1^\circ\text{C}$ the power dissipation must be $<0.2\,\text{mW}$. Very low current levels are required to obtain such a low power dissipation factor. This mode of operation is usually called zero-power sensing.

THERMAL CHARACTERISTICS

The power dissipated by a thermistor (P) is equal to the rate at which thermal energy (H) is supplied to the thermistor. This is the same as the rate at which energy is lost from the thermistor to its surroundings (H_L), plus the rate at which energy is absorbed (H_A) (13,14).

$$P = \frac{dH}{dt} = \frac{dH_L}{dt} + \frac{dH_A}{dt} \tag{10}$$

The rate at which thermal energy is lost from the thermistor is proportional to the temperature raise of the thermistor,

$$\frac{dH_L}{dt} = \delta(T_T - T_A) \tag{11}$$

The dissipation constant (δ) is typically measured under equilibrium conditions. It is not a true constant, since it varies slightly with temperature and with an increase in

temperature. The following relation can express the rate at which thermal energy is absorbed by the thermistor to produce a specific amount of rise in temperature:

$$\frac{dH_A}{dt} = sm\frac{dT_T}{dt} \tag{12}$$

where s is the specific heat and m is the mass of the thermistor. Thus, the heat-transfer equation for an NTC thermistor, at any instant after the application of power to the circuit, can be expressed as

$$P = RI^2 = \frac{dH}{dt} = \delta(T_A - T_A) + sm\frac{dT_T}{dt} \tag{13}$$

The thermal transient conditions at turn-on is given by the solution of equation 13, which is obtained considering P constant:

$$T_T - T_A = \frac{P}{\delta}\left[1 - \exp\left(-\frac{\delta}{sm}t\right)\right] \tag{14}$$

It means that when a significant amount of power is dissipated in a thermistor, its body temperature will rise above the ambient temperature as a function of time. In steady-state condition, when thermal equilibrium is achieved ($dT_T/dt = 0$ in equation 13, or when $t \gg ms/\delta$ in equation 14), the rate of heat loss is equal to the power supplied to the thermistor (remind equation 7, where $VI = \delta(T_T - T_A,)$. When self-heating is negligible ($P \cong 0$), the heat-transfer equation 13 can be rewritten,

$$\frac{dT_T}{dt} = \frac{-\delta}{sm}(T_T - T_A) \tag{15}$$

which can be solved to

$$T_T = T_A + (T_I - T_A)\exp\left(\frac{-t}{\tau}\right) \tag{16}$$

where T_I is the initial body temperature and τ is the thermal time constant ($\tau = ms/\delta$), which depends on the same environmental factors as δ.

All of the preceded discussions of thermal properties of NTC thermistors have been based upon a single device structure with a single time constant. When these devices are encapsulated into sensor housings, the simple exponential response no longer describes adequately the system response. The mass of the housing and the thermal conductivity of the materials used in the sensor housing will normally increase the system dissipation constant, increasing the thermal response time. In this case, the thermal properties are somewhat difficult to predict by mathematical modeling, usually requiring experimental tests of the system to obtain data on the resulting response time and dissipation constant (13). In general, the thermal response time may be reduced keeping the thermal resistance between the actual temperature sensor and the tissue being measured as low as possible.

THERMISTOR LINEARIZATION

The inherent nonlinearity of the resistance versus temperature characteristics of thermistors is rather troublesome in many applications. Even if we are interested in temperature measurements close to the body temperature, with variations of only a few degrees, the thermistor must be linearized. This can be accomplished basically in three ways: modifying the transducer circuitry, using analog circuits, or using digital techniques (15).

In the first option, the output of the transducer can be linearized over a limited temperature range with the addition of series or parallel resistors (6,16,17). It causes the voltage or the resistance of a simple fixed resistor thermistor to have zero error along a linear temperature scale at three equidistant points (Fig 4). The series resistance (R_s) required to make the conductance–temperature characteristic of a thermistor approximately linear may be calculated as follows (16,18):

$$R_s = R_{T_m}\left(\frac{\beta - 2T_M}{\beta + 2T_M}\right) \tag{17}$$

Where R_{T_m} is the resistance of the thermistor at the midscale temperature T_M. As can be observed in Fig. 4, this procedure reduces the sensitivity of the transducer. However, as the sensitivity of a thermistor is relatively high, the reduction is often a satisfactory tradeoff. The temperature coefficient of the series combination (α_S) is given by (6)

$$\alpha_s = \frac{-(\beta/T_M)^2}{(G_{T_M}/G_s) + 1} \tag{18}$$

Where G_{T_M} is the conductance of the thermistor at the midscale temperature and G_S is the conductance of the series resistance. An alternative method is to define the low, mid, and high end of the desired temperature range at three

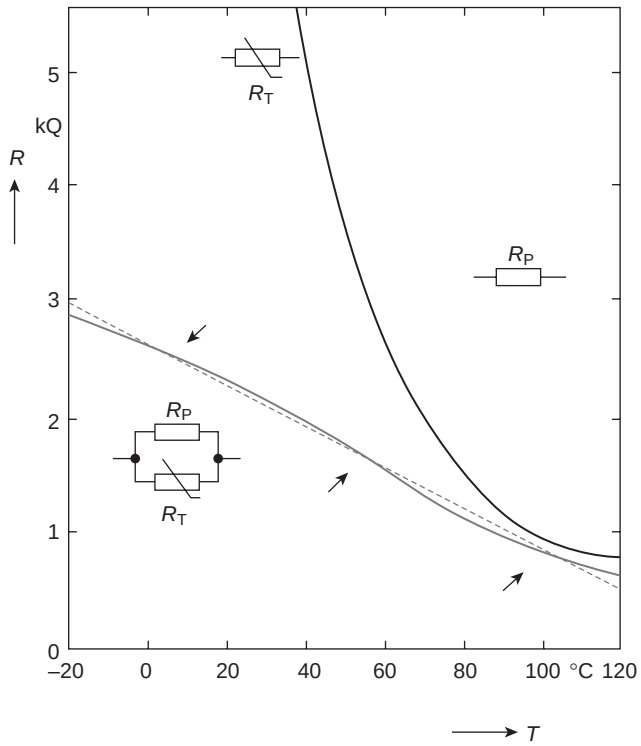


Figure 4. Linearization of a thermistor using a parallel resistor. (Adapted from (17) with permission).

equidistant points, and calculate R_s by (19,20):

$$R_s = \frac{[R_{T_M}(R_{T_{LO}} + R_{T_{HI}}) - (2R_{T_{LO}}R_{T_{HI}})]}{[(R_{T_{LO}} + R_{T_{HI}}) - (2R_{T_M})]} \qquad (19)$$

where R_{LO}, R_{T_M}, and R_{HI} are the thermistor resistance at the low, mid, and high end of the range, respectively. In this procedure, the addition of series resistor R_s forces the equivalent resistance of the fixed resistor thermistor to have zero error along a linear temperature scale in the three points chosen. The temperature range of the application determines the maximum error. For example, if the range is taken to be $-50\,^{\circ}$C to $100\,^{\circ}$C, the errors are 0 at -50, 25, and $100\,^{\circ}$C, and the errors elsewhere are distributed in an S-shaped. If the temperature range is reduced, the errors become smaller; being $2.0\,^{\circ}$C over a $60\,^{\circ}$C range, $0.05\,^{\circ}$C over a $30\,^{\circ}$C range, and $0.01\,^{\circ}$C over a $10\,^{\circ}$C range. Another way to linearize thermistors includes the use of a Wheatstone bridge. In most applications, the bridge consists of a linear thermistor voltage divider and a fixed resistor voltage divider, as described in Fig. 5a (21). This circuit is designed to produce 1 V at $25\,^{\circ}$C and 200 mV at $45\,^{\circ}$C, with an output voltage that is linear within $\pm 0.06\,^{\circ}$C in this temperature range (Fig. 5b). Hoge (22) showed that circuits based on resistors (serial, parallel, or bridge) are equal in their ability to linearize thermistor characteristics. Thus, the choice of a linearizing circuit should be made on other grounds, such as simplicity or convenience.

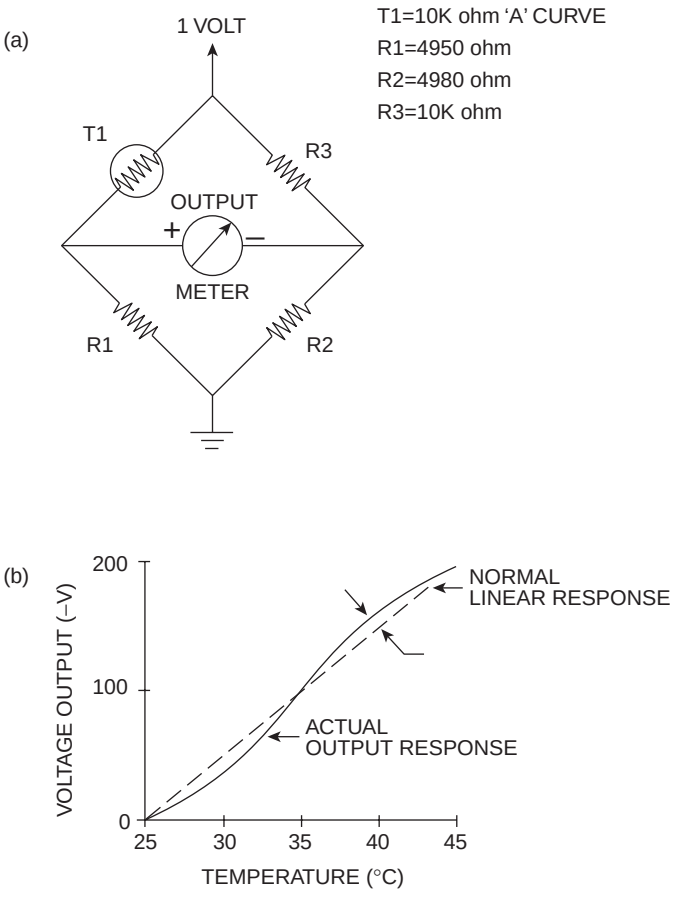
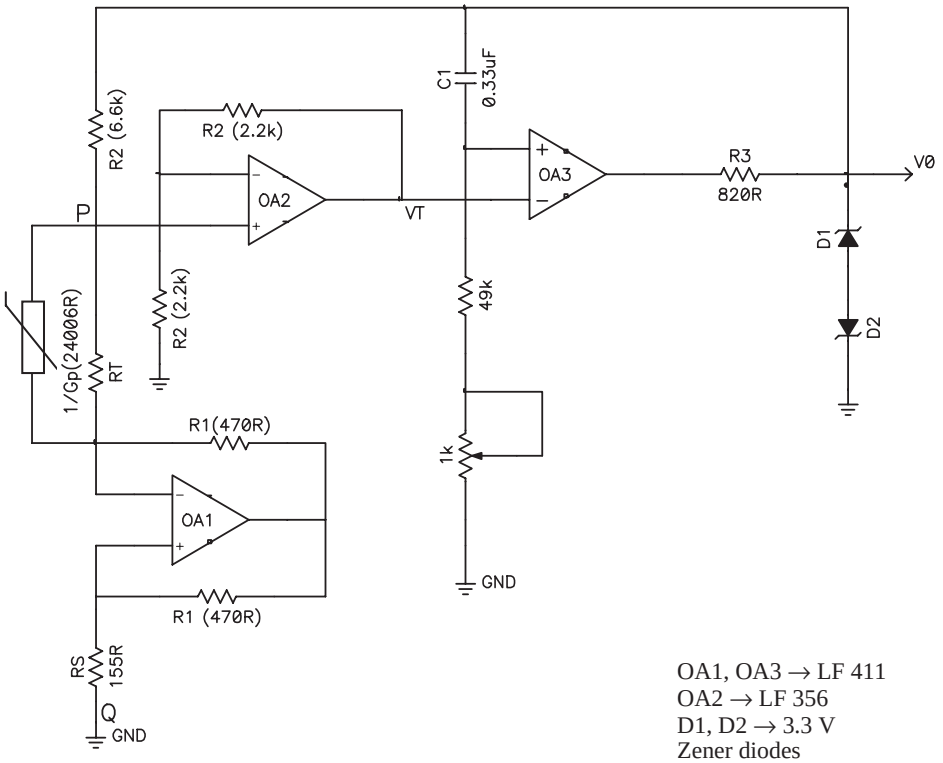


Figure 5. Use of a thermistor in a Wheatstone bridge for temperature measurements (A) and respective output voltage as a function of temperature (B) (courtesy of Alpha Sensors, 2121 Palomar Airport Rd., Carlsbad, CA 92009, USA).

Theoretically, an improved linearization over wider temperatures can be achieved using more complex circuits. For many applications, it can be a better option to linearize the transducer at some point of the analog process, as in cases where no digital processing is used. It is also true when limited processing capability and/or memory are available, and the analog processing can be done simply and at low cost. Analog circuitry using piecewise-linear approximations has been developed to be used with temperature transducers (23). However, these circuits are complex and costly, and are not usually used in practice. Voltage-to-frequency converters (24), logarithmic (25), and temperature to frequency (26) circuits, were proposed for this objective. In 1990 Slomovitz and Joskowicz (27) compared the quality of these active linearizing circuits. They concluded that the errors (2, 1, and 0.7 K, respectively) were in the same order as those obtained in single resistor circuits. The authors attributed the origin of the errors to the fact that all of them were based on the simple exponential approximation. More recently, Kaliyugavaradan et al. (28) proposed a method for linearization of thermistor response using series-parallel resistors based on a new four-constant curve fit method, which resulted in a temperature-to-frequency converter that provides accuracy better than 0.2 K. The same researchers obtained even more accurate results by using the circuit described in Fig. 6, which works essentially as an astable multivibrator. Experimental results obtained using a standard thermistor in the range of 35–95 °C revealed a peak error less than ± 0.1 K (29). Recently, a signal-conditioning circuit by generating a compensating, pseudologarithmic response function was proposed (30).

If the data are to be digitized and processed digitally, as soon as possible, it probably makes sense to perform any needed linearization in the digital domain. The techniques used in this case allow linearization to be done much more efficiently and accurately in software, and eliminate the need of tedious manual calibration using multiple and sometimes interacting trimpots. The principal techniques involve look-up tables and computational algorithms (20,31,32). A look-up table may be constructed, for example, by using a read only memory (ROM) hardwired to the analog-to-digital converter output. Each level from the converter corresponds to an address in ROM, and the word stored in that address is the linearized value. It can be constructed by using the device characteristics, available from the manufacturer. The Steinhart–Hart equation can be used to create the table, which implements a third-order linearization formula that provides high accuracy. If memory is limited, and the measurement is made infrequently, but rapid mathematics is available (as, e.g., in digital signal processing systems), a mathematical function that approximates the inverse of the nonlinear relationship, or the difference between the ideal signal and the actual signal, can be derived and stored in program memory. Then, whenever the measurement is made, the processor computes the correct value, based on its mathematical relationship to the measured input variable (23). In system



OA1, OA3 → LF 411
OA2 → LF 356
D1, D2 → 3.3 V
Zener diodes

Figure 6. Linear temperature to frequency converter (Reproduced from (29) with permission).

including microcontrollers, in many cases the mathematical functions may contain complex polynomial and exponential functions, placing a great burden on the program memory, RAM, and execution speed of most low cost devices. Digital piecewise linear interpolation may be a good choice for sensor linearization in such systems due to its fast execution speed, reduced program memory requirements, and easy of implementation (33).

ELECTRONICS FOR TEMPERATURE MEASUREMENTS

In the circuit project, simulation and analysis environment, SPICE subcircuits for thermistor modeling are useful, allowing for a realistic simulation of thermistor parameters for all standard analysis [transient, alternating current(ac), and dc]. Useful subcircuits have been described by Keskin (14), Hagerman (34) and Wangenheim (35), and are also commercially available (http://www.catena.uk.com; http://www.intusoft.com/products). As discussed earlier, there are a variety of circuits in which thermistors may be used for temperature measurements. Caution must always be taken, however, to insure that the power dissipated in the thermistor is held at a minimum and that the current flow is insufficient to cause self-heating, since temperature measurements require that the thermistor be operated in a zero-power condition. The most common technique implies the use of a constant current source, and the measurement of the voltage developed across the thermistor.

Current Sources

In the section dedicated to the thermal characteristics of thermistors, it was shown that the use of current source would prevent thermal runaway in thermistors. The simplest circuit that can approximate a current source is a high voltage source with a high resistance connected in series (voltage divider), as can be seen in Fig. 7. In this circuit, the output voltage is taken across the fixed resistor, and, the higher this resistor in relation to the thermistor resistance, and the higher the voltage, the closer this circuit will conform with an ideal current source. From the plot of the output voltage (Fig. 7b) it can be observed that there is a range of temperatures where the circuit is reasonably linear with good sensitivity. However, this simple circuit does not act like an ideal current source, since modifications in the thermistor resistance introduces alterations in current. An even better approximation to the desired current source would be obtained using standard circuits based on operational amplifiers (16,36). Integrated current sources are also useful, allowing for the configuration of regulated current sources of varying magnitudes. Fig. 8 shows a typical example, which is based on a device containing two low current regulators (37). One of current regulators supplies 100 μA to the thermistor. The temperature of the thermistor is converted into a voltage that is increased by R3, filtered by R2-C1 and amplified by U1B. The second current source is used to provide the reference voltage in combination with R1 and U1a. This circuit is a useful framework for thermistor temperature measurements using analog-to-digital converters (37).

Wheatstone Bridges

The wheatstone bridge is a widely used means of measuring temperature using thermistors, since the bridge aids the linearization of the NTC (Fig. 5). The condition for the

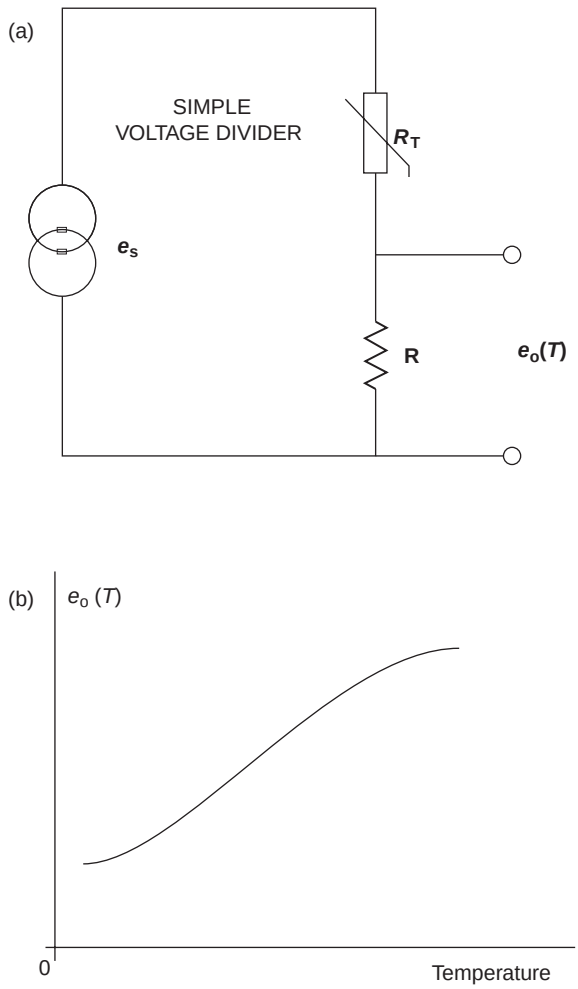


Figure 7. Simple series circuit to approximate a constant current source (A) and corresponding output voltage (B).

balance of the bridge is $T_1R_2 = R_1R_3$. In a normal operation, R_3 is usually a variable resistor that is used to zero the bridge. If the bridge is excited with a constant dc voltage, the bridge output may be displayed in a dc meter, calibrated directly in temperature units. In medical care, it is often necessary to accurately determine very small temperature differences. For example, it may be necessary to measure the difference in temperature between two regions of an organ. In cases involving differential temperature measurements, matched thermistors are placed in the two voltage dividers. In Fig. 5, the second thermistor replaces R_3. With this configuration, temperature differentials as close as 0.001 °C can be readily detected.

Temperature-to-Frequency or Temperature-to-Time Interval Converters

For applications in which digital processors or microcontrollers are used for data acquisition and signal processing, the transducer response must assume a form suitable for conversion to digital format. Temperature-to-frequency or temperature-to-time interval conversions are convenient methods to measure temperature in this case, since they permit an easy and low cost interface, no ADC is needed and only one bit of input is necessary (6). Another advantage is that the optical isolation circuits used in this kind of application presents lower cost than that used in linear systems. Fig. 9 shows two simple circuits in which the frequency of oscillation varies with the temperature. Because a thermistor resistance varies with temperature, the RC time constant will change accordingly. And since the 555 timer determines the frequency corresponding to the RC time constant, the actual resistance is proportional to the number of counts recorded by the 555 timer (Fig. 9a). The frequency, or the number of counts in a given time window, can be easily converted to a temperature value (38,39). In the circuit described in Fig. 9(b), the thermistor is placed in the feedback loop of a histeresis-based oscillator, and the output frequency is related to temperature,

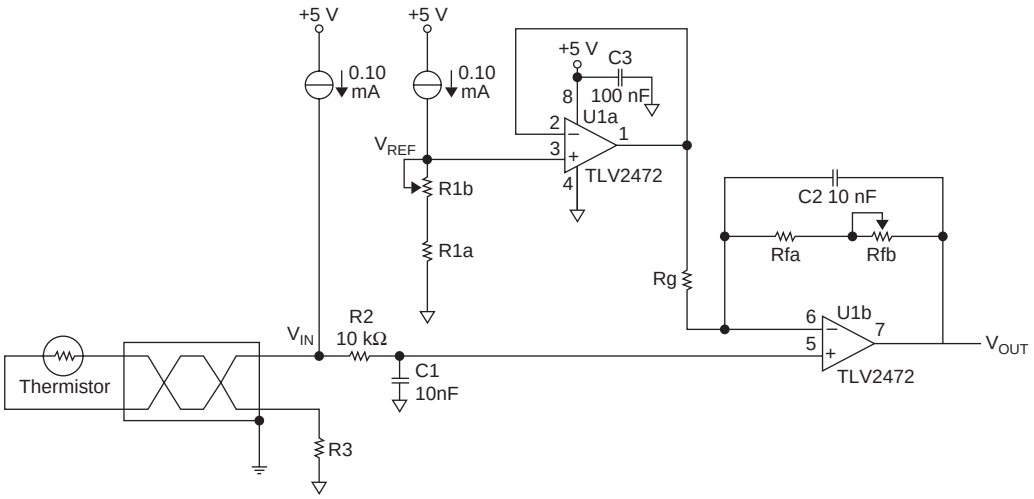


Figure 8. Circuit containing two integrated low current regulators. Courtesy of Texas Instruments Inc. 13532 N. Central Expressway M/S 3807 Dallas, Texas, USA (Adapted from (37) with permission).

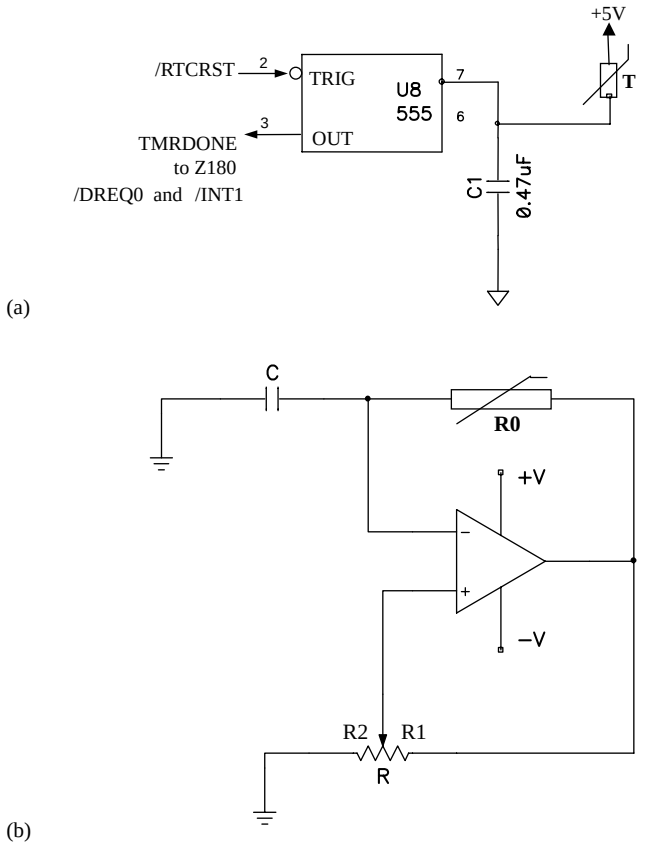


Figure 9. Simple, low-cost nonlinear temperature to frequency converters based on oscillators constructed around 555 (A) and operational amplifier (B). Circuit (A) is a courtesy of ZWorld, 2900 Spafford Street Davis, California 95616, USA, and circuit (B) is from Interfacing Sensors to the IBM PC, by W. J. Tompkins and J. G. Webster, 1988. (Reprinted by permission of Prentice-Hall).

according to the following relation (6):

$$f = \frac{\ln[(1+\gamma)/(1-\gamma)]\exp(\beta/T_0)\exp(-\beta/T)}{2CR_0} \qquad (20)$$

where R_0 and C determine the nominal frequency and $(\gamma = R_2/(R_1+R_2)$ enables small adjustments. Note that in both circuits described in Fig. 9, the frequency of oscillation varies nonlinearly with the temperature. Linear variations of temperature to frequency can be obtained using the circuit described in Fig. 6. Another example is seen in Fig. 10 (15), where the differential output of a bridge circuit employing an YSI 44018 linearized thermistor is amplified by an AD522 instrumentation amplifier, and converted to frequency by a 452L 100 kHz full-cale V/F converter. The circuit operates in a temperature range of 0–100 °C with accuracy to within 0.15 °C. The pulse transformer connected to the collector of the 2N2222A provides galvanic isolation.

Low Cost Systems

In handheld applications, accurate temperature measurement can easily be accomplished by interfacing a Wheatstone bridge including a thermistor and a digital voltmeter integrated circuit, as illustrated in Fig. 11. The IC is comprised of an analog-to-digital converter with built-in $3\frac{1}{2}$ digits LCD driver providing resolution of 0.1 °C. This digital thermometer makes it possible to achieve an overall system accuracy of ±0.4 °C from 0 to 100 °C. Anytime more than one thermistor temperature probe must be measured, a scanning digital panel meter may be a useful adjunct. Fig. 12 shows another simple circuit, using a microammeter in series with a potentiometer and a thermistor connected to a potential source. The meter can be calibrated in terms of temperature, providing a system usually used in low cost applications (6).

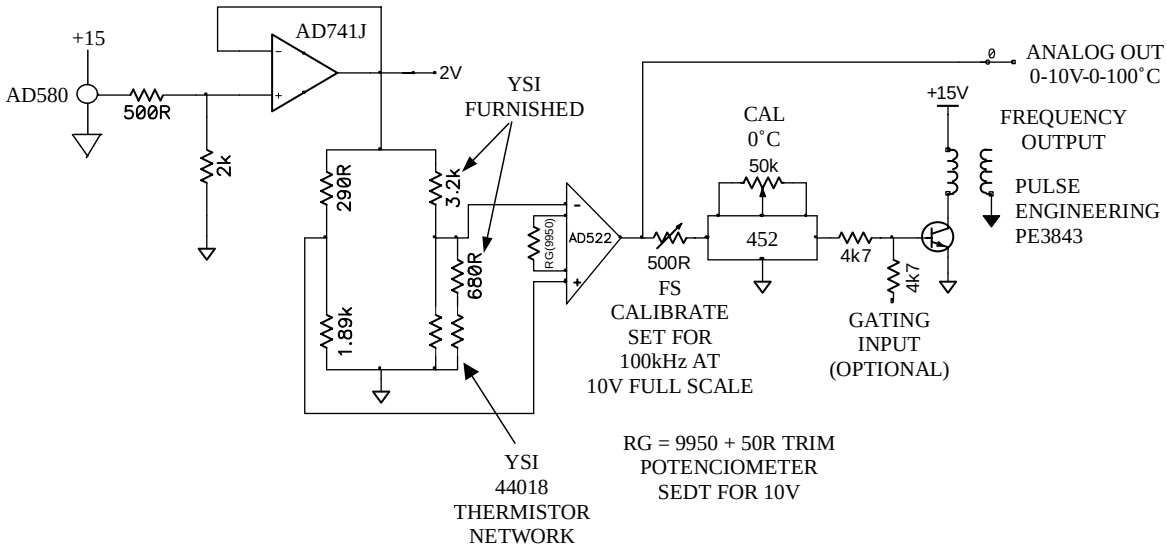


Figure 10. Linear thermistor-to-frequency converter. (From D. H. Sheingold 1981, Transducers Interfacing Handbook, Analog Devices, Norwood, MA. Used by permission).

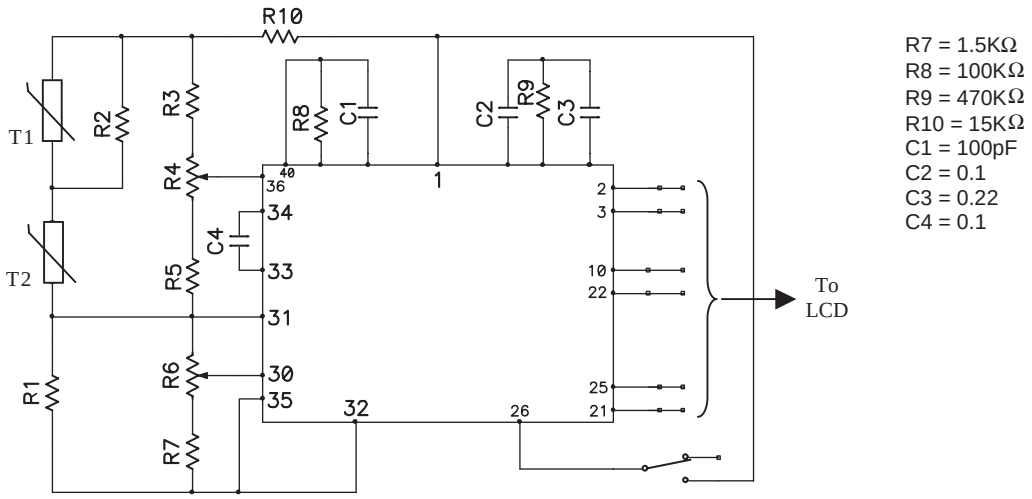


Figure 11. Digital thermometer based on 3-1/2 digits voltmeter (courtesy of Alpha Sensors, 2121 Palomar Airport Rd., Carlsbad, CA 92009, USA).

High Resolution Measurements

Very low modifications in temperature can be sensed by using lock-in amplifiers (40). These amplifiers are used to measure the amplitude and phase of repetitive ac signals buried in noise. It is achieved by their ability of acting as a narrow bandpass filter, which removes unwanted noise while allowing through the signal that is to be measured. The ac frequency of the signal to be measured is used as a reference signal to set the passband region of the filter, and must be supplied to the lock-in amplifier along with the unknown signal. Thus, ac must excite the thermistor. Studying perfusion changes associated with cerebral blood flow, Wei et al. (40) were able to detect temperature changes of 0.001 °C using lock-in amplifiers, Wheatstone bridges and matched thermistors.

Measurements without Physical Contact

Evans and Hajnayebi (41) developed a temperature acquisition system using proximity telemetry. The system consists of two units, a module that can be worn by the patient containing a temperature-to-pulse width converter (astable oscillator), and a handheld interrogator, which incorporates a radio frequency generator and an automatic-gain controlled (AGC) temperature readout (Fig. 13). Usually in measuring a patient's temperature, the clinician must wait until the thermometer probe reaches body temperature and then log the readout. In some patients, however, frequent temperature measurement is essential, as, for example, in severe anemic patients undergoing slow blood flow transfusion where it may be necessary to monitor the body temperature each 15 min, for periods of up to 48 h. In such circumstances, the physical disturbance to the patient becomes extremely unpleasant. The system described in Fig. 13 does not require physical contact, allowing for thermistor to be interrogated by a handheld unit held in close proximity to the patient. Compared with conventional electronic clinical thermometers, such design has an additional advantage, since it does not require a separate display module for each patient.

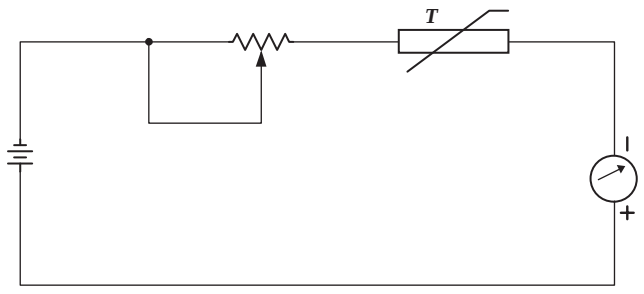


Figure 12. A simple low-cost circuit for temperature measurements based on a microammeter (from Interfacing Sensors to the IBM PC, by W. J. Tompkins and J. G. Webster, 1988, Reprinted by permission of Prentice-Hall).

Systems Based on Microcontrollers

The advent of low cost microcontrollers provides the design engineer new design possibilities for medical temperature measurement. Microcontrollers are comprised of a built-in microprocessor, analog-to-digital converter, RAM, and several digital inputs–outputs. The complete system utilizes the microcontroller, multiplexer, EPROM, digital display,

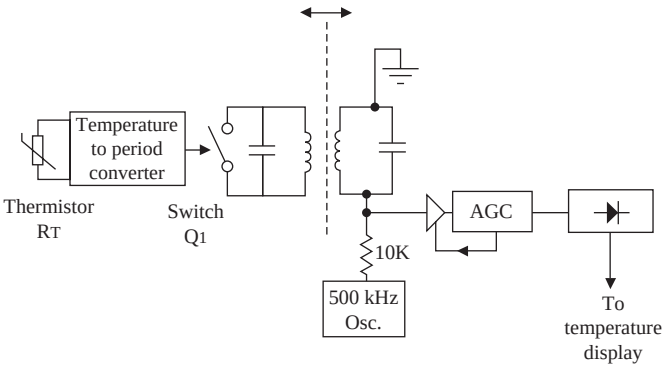
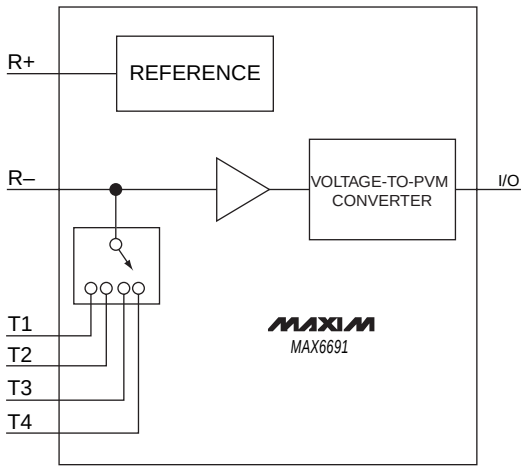
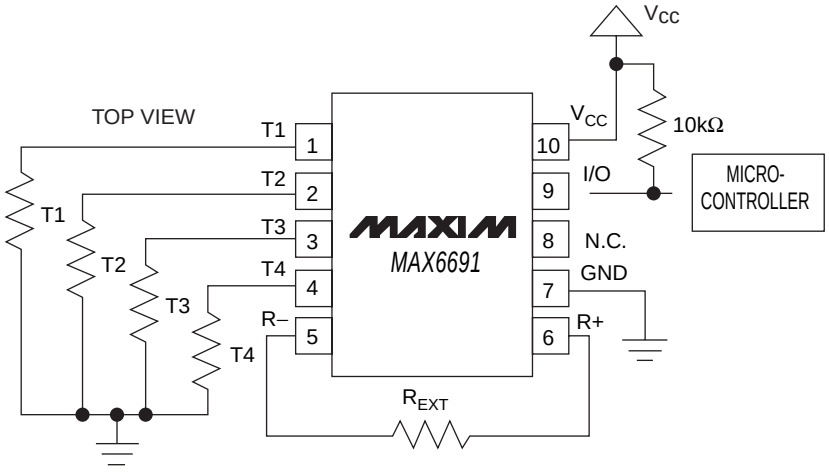


Figure 13. Proximity clinical thermometer hardware. Reprinted from (41), with permission from Elsevier.



(a)



(b)

Figure 14. Functional diagram (A) and typical application circuit (B) of the MAX6691, four-channel temperature-to-pulse-width converter. Copyright Maxim Integrated Products (www.maxim-IC.com). Used by permission.

keypad, and display driver, being programed in assembler language (www.microchip.com; www.motorola.com). These systems are relatively inexpensive to produce, yet offer high temperature accuracy and various software-controlled outputs (39,42). For example, a microcontroller system utilizing thermistor sensors can monitor the temperature in several locations in a patient. After programming, the microcontroller converts the resistance of the thermistor into a temperature reading by using the resistance versus temperature algorithm based on the Steinhart–Hart equation (33) or a look-up table, as discussed before.

Integrated Circuits

Recently, the electronic industry made available integrated circuits specifically dedicated to use with thermistors (43) and well suited for this task (19). The MAX6691 is a four-channel thermistor temperature-to-pulse-width converter that measures the temperatures of up to four thermistors and converts them to a series of pulses whose widths are related to the thermistors temperatures (43). This device can be readily connected to a variety of microcontrollers with a simple single-wire interface. Operating under the supervision of a microcontroller, the MAX6691 powers the thermistors only when a measurement is under course. This minimizes the power dissipation in the thermistors, virtually eliminating self-heating. In the intervals between conversions, the MAX6691 falls into a 10 μA (max) sleep mode, where the voltage reference is disabled and the supply current is at its minimum. In handheld healthcare systems, where power is at a premium, it may be an interesting characteristic. These integrated circuits also have internal voltage reference that isolates thermistor from power-supply noise, as described in the functional diagram presented in Fig. 14a.

The AD7711 is an integrated circuit from Analog Devices with signal conditioning and A/D conversion functions that is well suited for temperature measurement applications using thermistors (19). This integrated circuit includes a programable gain amplifier, current sources and a voltage reference on the chip, allowing thermistors to be excited in either a constant current or voltage mode (Fig. 15a). The reference input is also differential, allowing ratiometric operation on the front end. The component can achieve >16 bits of peak-to-peak resolution and update rates of 100 Hz. Filtering is also provided as part of the ΣΔ process. This on-chip filtering may be useful when transducer excitation frequencies must be removed from the input signal. In addition, the filter profile also provides

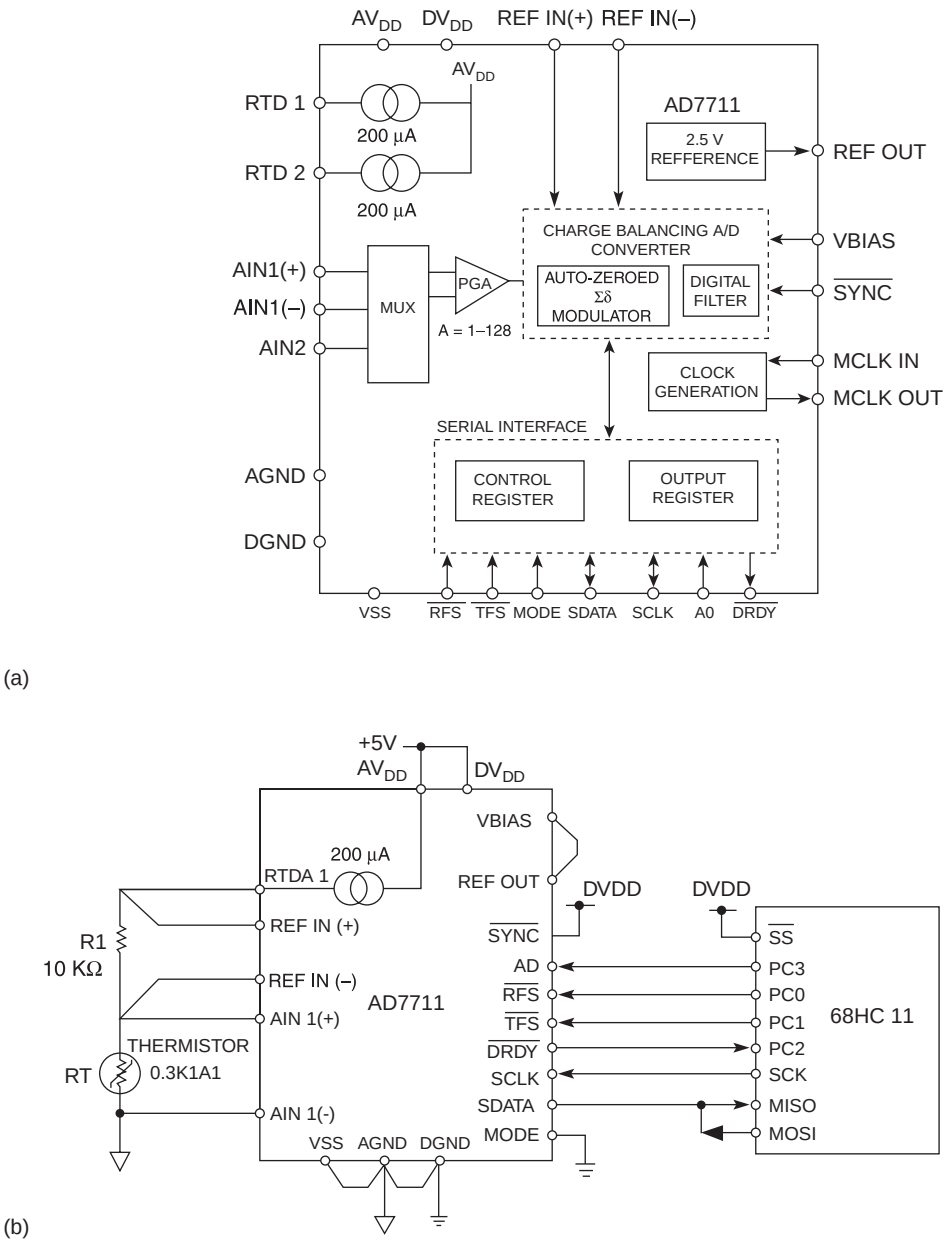


Figure 15. Functional diagram (A) and typical application circuit (B) of the AD7711, 24-Bit Sigma-Delta, Signal Conditioning ADC. Reproduced from Ref. (19) with permission.

notches with 120 dB attenuation that can be placed at 50 or 60 Hz, minimizing line frequency components from the system. An example of the AD7711 use is presented in Fig. 15b. The on-chip 200 μA current source acts as the excitation for the thermistor and generates the reference for the converter. In this case, the circuit is fully ratiometric, ensuring that changes in the excitation current will not affect the performance of the circuit. The diagram also shows a serial interface to a 68HC11 microcontroller, which may control the A/D converter, take data readings, and run a linearization algorithm.

Personal Computer Interfacing

Thermistors can be integrated to data acquisition systems and an IBM PC or compatible computer, in order to implement a flexible system for application is medical thermometry (6,44). These PC-based DAQ systems requires some signal conditioning hardware to interface the thermistor to the measurement device, such as a plug-in DAQ board (Fig. 16). The DAQ board or module performs the analog-to-digital conversion. A system to interface the thermistor and its output signal to this stage should include an excitation current or voltage source, signal amplification, low pass filtering and isolation signal conditioning hardware, which serves to electrically isolate the thermistor sensors from the measurement system, protecting the patient. Concerning the last topic, it is important to point out that electrical-safety codes and standards in healthcare facilities must always be strictly followed (45). Several of useful circuits to perform these tasks were discussed earlier. A typical plug-in DAQ board has eight to 16 analog

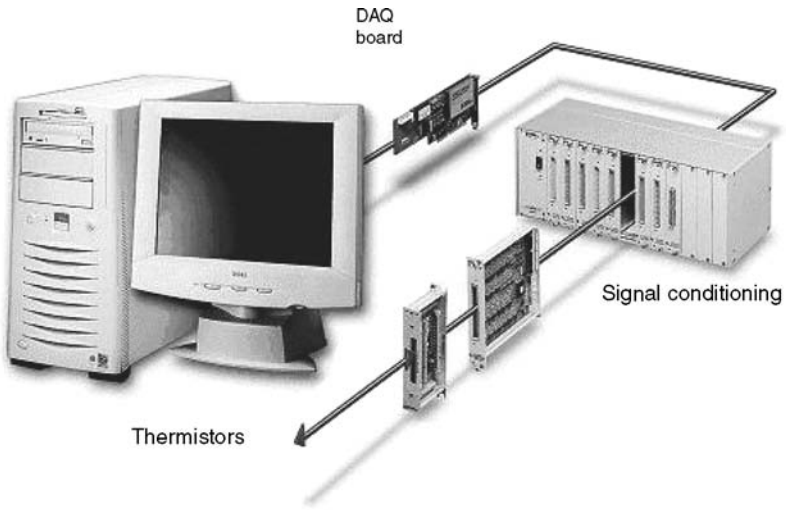


Figure 16. Block diagram of a PC-based DAQ system for temperature measurements with thermistors (courtesy of National Instruments Corporation, Austin Texas, USA).

input channels. External multiplexing systems can be included if it is necessary to expand the number of input channels that can be connected to a DAQ board. The multiplexing system sequentially switches multiple channels to a single-input channel of the DAQ board. Software choices for controlling the data acquisition system include a general purpose programming language (e.g., C, Delphi, or Pascal, under DOS or Windows). Alternatively, the LabVIEW environment can be used to control the system. In this case, special routines for implementing the Steinhart–Hart equation to convert measured voltages from thermistors into temperatures are available.

EXAMPLES OF COMMERCIALLY AVAILABLE THERMISTORS AND THERMISTOR PROBES

In the earliest days, thermistors acquired a reputation for being unpredictable and unstable devices, due to problems in manufacture and in application. Nowadays, however, with increased user familiarity with these devices and modern manufacturing technologies, they can be used with a great deal of confidence (15). A large number of companies produce NTC thermistors. Moreover, there is a great variety in the constructional characteristics and parameters of these components, depending on their use. In this section, a survey will be made concerning NTC thermistors for medical use manufactured by some companies. Detailed information can be found in available catalogs of the manufacturers or at the WWW pages listed at the end of this section.

Matched interchangeable thermistors eliminates the need for individual resistance temperature calibration, as well as permits the standardization of circuit components and simplification of design and replacement problems. Honeywell/Fenwall makes devices with interchangeability to within 0.2 °C, from 0 to 70 °C (Uni-Curve series, Fig. 17a) (10). Highly interchangeable thermistors are also manufactured by Alpha Sensors. These devices present tolerances to ±0.1 °C over the medical temperature range (21).

The Honeywell/Fenwall LTN (Linear Thermistor Networks, Fig. 17b) series are designed to produce a resistance change or voltage output that varies linearly with temperature (typical maximum linearity deviation of 0.256 °C from −30 to 50 °C). They consist of one twin thermistor and two precision resistors, and are also available in probe assemblies, providing a sensitivity that is hundred folds greater than that of thermocouples.

The standalone thermistor element is relatively fragile and cannot be placed in a rugged environment. In order to overcome this problem in biomedical applications, thermistor probes, which are thermistor elements embedded in protective tubes, are widely used. For example, Fig. 17c shows a pediatric probe for oral or rectal use (Ysi 423) with the sensor located at the tip of a semiflexible nylon tube. These probes can be disposable (Ysi 4400) or reusable (Ysi 400 ou 700). When disinfections are critical, autoclavable probes are also available for skin, esophageal, or rectal use in adult or pediatric measurements (Ysi 400AC probes, Fig. 17d).

For research or medical applications requiring small size and rapid response, as, for example, long-term subcutaneous measurement, thermal dilution, and flow measurement studies, Ysi manufactures catheter probes with time constant of 0.2 s, electrically isolated and operating in the temperature range of 0–70 °C. Honeywell/Fenwall manufactures Small Bead Thermistors, that offer ultrafast time response (T.C. 1 s maximum) and are highly sensitive to electric power ($\delta = 0.1$ mW·°C^{-1} minimum). They are suited for usage in low heat capacity applications and their micro size (0.36 mm outer diameter) makes them adequate for using in extremely small assemblies, (e.g., catheter and hypodermic needles). In addition, they are also adequate for used in self-heat applications. If a standard probe does not fulfill the needs, custom design is available by several manufacturers (please, see WWW pages at the end of this article).

SOME FIELDS OF CLINICAL APPLICATIONS OF NTC THERMISTORS

As pointed out before, thermistors are probably the most widely used transducer for medical temperature measurements. Their characteristics have been contributing to

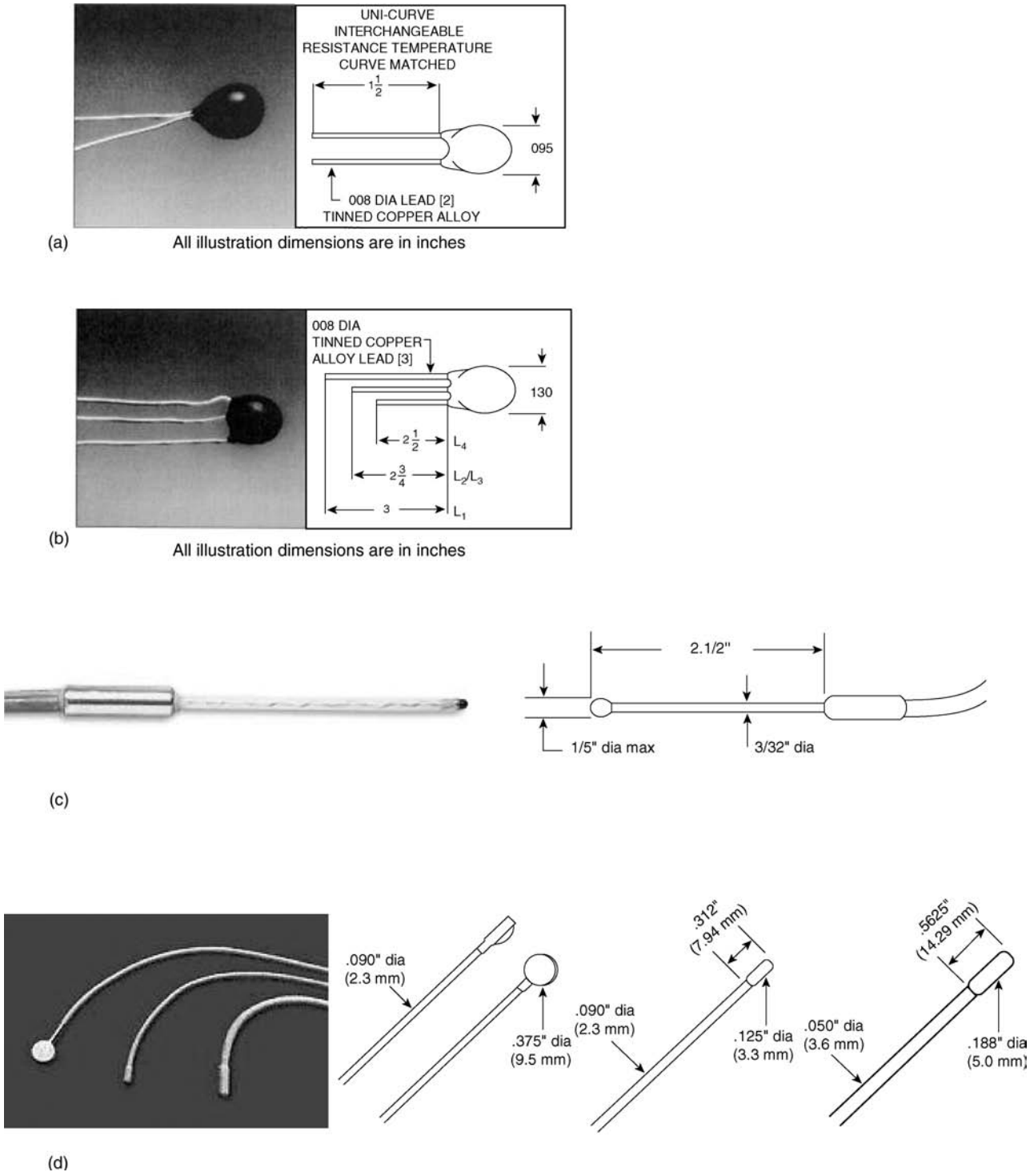


Figure 17. Examples of thermistors commercially available. Highly interchangeability devices (Uni-Curve® series) (A), linear thermistor networks (LTN® series) (B) (courtesy of Honeywell International Inc. 101, Columbia Road Morristown, NJ 07962, USA). Pediatric probe for oral or rectal use (C), and autoclavable probes (D) for skin, esophageal or rectal use in adult or pediatric measurements (courtesy of Ysi, 1700/1725 Brannum Lane, Yellow Springs, Ohio, 45387, USA).

facilitate clinically difficult measurements, that includes testicular temperature measurements in reproductive medicine (46), hypothermia (47), transcutaneous measurements during cardiopulmonary bypass (48), continuous monitoring of preterm infants (49), and personal heat strain monitoring in occupational medicine (50). However, beyond their utility in temperature measurements, these devices have found widespread use in a variety of

systems for clinical and research applications to measure flow, thermal conductivity and diffusivity of biomaterials, and to detect the presence of liquids. Therefore, this section intends to have a brief discussion about some of these applications, showing the branches of medicine where technology based on thermistor is used, and how this technology can assist each of them.

Cardiovascular Monitoring

Cardiac output, the volume of blood ejected by the heart each minute, is a key parameter in cardiovascular medicine, which is used to obtain diagnostic information about the heart and for continuous monitoring of heart function in critically ill patients. The thermodilution method uses thermistor type catheters to estimate this parameter. The tip of a Swan–Ganz catheter is inserted into a large vein, typically one in the right side of the neck, and advances through the right heart into the pulmonary artery (Fig. 18). Other sites can be used for catheter insertion (e.g., the groin or the arm). A cold saline or a dextrose solution, whose volume and temperature are known, is injected into the blood stream through one of the catheter lumens. The solution mixes with the blood in the right atrium and is diluted as it is carried downstream to a thermistor located at the surface of another catheter lumen. At the thermistor location, the temperature of the blood-injected mixture is measured over a period of time, and then the cardiac output (efficiency) is computed from this temperature–time response data.

At the level of the capillary network, the tissue blood flow (perfusion) is a primary factor in the local transport of heat, drugs, oxygen, nutrients, and waste products. This fundamental parameter holds the key to the diagnosis and subsequent treatment of numerous medical problems. The high sensitivity to small changes in temperature shown by

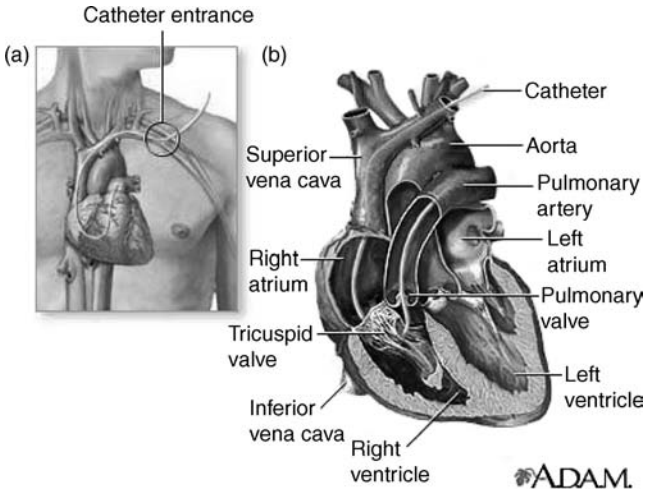


Figure 18. Placement of the thermistor-based catheter in the termodilution method of estimating cardiac output. Overview (A), and a detail of positioning in the heart (B). This method involves the passage of a catheter into the right side of the heart to obtain diagnostic information about the heart and for continuous monitoring of heart function in critically ill patients. From Medical Encyclopedia (Medline Plus), reprinted by permission.

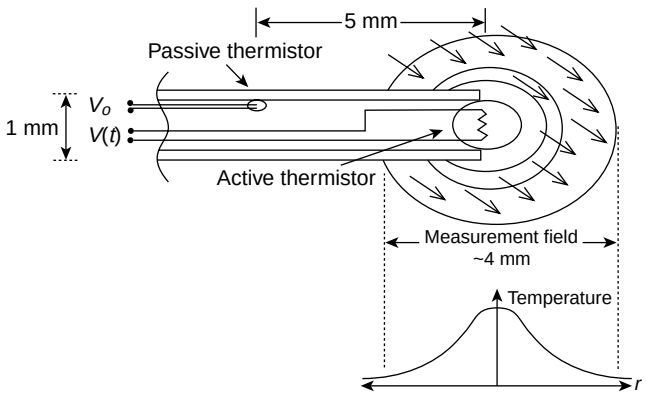


Figure 19. Example of thermal diffusion probe. It can be seen an active, heated thermistor, at the probe tip which produces a thermal measurement field that is dependent of the tissue thermal properties and the perfusion. The passive thermistor, mounted proximal to the probe tip, monitors the baseline temperature variations. Reproduced from (51) with permission of the IFMBE.

a thermistor can be employed to sense the small amounts of heat involved in the thermal techniques developed for the measurement of tissue blood flow (51). Minimally invasive probes constructed around two thermistors are usually applied in these techniques. An example of such probes is presented in Fig. 19. In these probes, one thermistor is used in the self-heat mode, operating as both a heat source and a temperature sensor. Changes in perfusion cause changes in its temperature, which is used as an indirect index of blood flow. As the baseline body temperature fluctuates, a second thermistor with the same size and electrical characteristics is used to measure and compensate for the changes in reference temperature. Systems like these may help clinicians in several application areas, providing early warning for ischemic events, targeting therapy rapidly and accurately, monitoring of patients in organ transplantations, and evaluating tumor and cerebral blood flow (51).

Respiratory Measurements

Obstructive and restrictive respiratory diseases present a huge public health problem. Prevalence rates of asthma and chronic obstructive pulmonary disease are ~4% each. These diseases are characterized by airflow limitation that results from modifications in lung parenchyma and airways. Thermal convection flowmeters measures the local speed of a fluid by measuring the heat loss from a heated element in the flow path (52,53). These instruments are used in respiratory medicine for gas flow analysis, and commonly thermistors are used as the sensing elements. They are also known as thermistor pneumotachometers (16). In these instruments, a small thermistor is placed in the flowing fluid. It operates in the self-heat mode in order to maintain an average temperature above that of the surrounding fluid. This is accomplished using a feedback circuit (16,52,53), as described in Fig. 20. A change in temperature tends to cause a variation in the thermistor resistance, which in turn affects the amplifier output voltage and the current through the sensor. Considering the

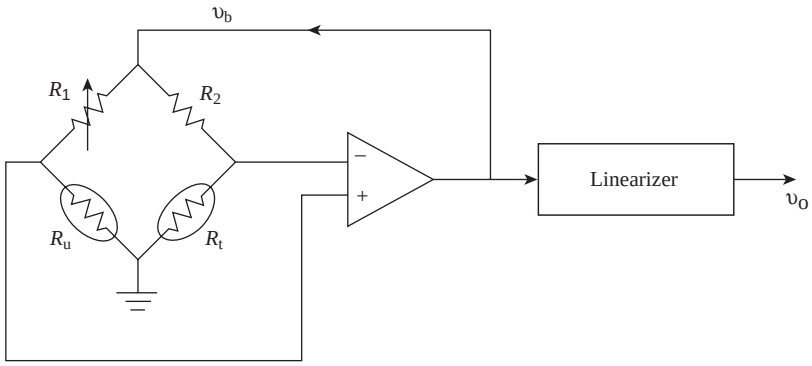


Figure 20. Feedback circuit used in thermal convection flowmeters. From Medical Instrumentation: Application and Design, by J. G. Webster. Reprinted by permission of John Wiley and Sons, Inc.

amplifier as presenting infinite constant gain and bandwidth, the bridge is always statically maintained balanced, and hence the thermistor resistance and temperature are constant. The sensor resistance is maintained constant equal to R_1 under any operating condition. This circuit operates satisfactory if ambient temperature is constant. If changes in ambient temperature are expected, a second unheated thermistor can be included in the circuit to compensate it (R_t in Fig. 20). One of the main advantages of this configuration is that the high negative gain feedback divides the sensor time constant by a factor equal to the loop gain, improving the frequency response (52). Thermistors loose heat at a rate dependent on the local mass flow, temperature, specific heat, kinematic viscosity, and thermal conductivity of the fluid. Thus, when a patient exhales through an breathing device in which the thermistor is mounted, it is cooled and the circuit needs to provide more electrical power to keep the thermistor temperature constant. This power is proportional to the airflow. When gas flow properties are sufficiently constant, the output voltage of these circuits is a nonlinear function of mass-flow rate only. Linear mass–flow relationships may be obtained using a linearization stage (Fig. 20), based on analog or digital implementations of piecewise linear or polynomial approximations (53).

Since the thermistor is cooled equally for both directions of velocity, the system with a single sensor provides an output of the same polarity, independent of the flow direction. This feature limits the use of these sensors to unidirectional flow, which can be satisfactory in some applications, as for example, in forced expiratory testes. Directional sensitivity can be provided by putting multiple sensors at separate points along the flow.

Speech Therapy

Spoken language is a highly developed skill. It involves the use of parts of the brain that deal with hearing, understanding speech, sound production, and conversion of the thoughts into speech. Patients suffering from cleft palate or similar defects exhibit a type of speech that is characterized by excessive nasal escape of air and by an abnormal resonance compared with normal speakers. This may be treated by plastic surgery, palatial prostheses, speech therapy, or even a combination of these treatments (54). Trained listeners and also experienced therapists have been already used to detect the presence of nasal air escape. These evaluations are, however, subjective and hinder objective comparisons with previously made assessments, especially if these were made by another therapist (55). In order to allow a qualitative and quantitative analysis of nasal air escape, anemometers based on thermistors were proposed (54–56). Fig. 21 is a cross-section showing details of the device. The thermistor is positioned in the longitudinal path of the nasal airflow. The system uses a Wheatstone bridge; in one arm of the bridge a thermistor is connected, while in the opposite arm a second thermistor is used, as temperature-compensation element. In the practical use of this kind of system, clear differences were obtained when nonnasal and nasal subjects were asked to talk a nonnasal word (cheese) and a nasal word (missing) (54). The system can provide a numerical figure of merit to indicate the extent of the defect and the effectiveness of a treatment, as, for example, palatal training.

Sleep Medicine

One of the most used methods to sense breathing patterns is to detect airflow using a nasal thermistor sensor. The principle here is that of exhaled air ($\sim$37 °C) being slightly warmer than inhaled air (room temperature), resulting in modifications in the thermistor resistance correlated with the respiration rate. The small size of the thermistor contributes to prolonged, minimally intrusive measurements of breathing pattern, which are particularly important for respiratory surveillance in newborn intensive care, biofedeback studies, and cyrcadian rhythm analysis (57).

These characteristics also made thermistors a traditional device in the diagnostic of sleep-disordered breathing (SDB) (58), which is a widespread disease estimated to be present in $\sim$2–4% of middle-aged adults. In these patients, the pharyngeal airway narrows with sleep onset, initially producing harsh respiratory breathing with some evidence of inspiratory flow limitation. Then, with further narrowing snoring is generated, and finally there is a complete collapse producing a full obstructive sleep apnea (OSA). Even if upper-airway obstruction is incomplete (hypopnea), increased upper airway resistance will still cause clinical symptoms similar to OSA because of respiratory effort-related arousals. Although rare compared to obstructive events, sleep apnea episodes can also be present in the absence of upper airway obstruction. In this case, known as central sleep apnea (CSA), the apnea events results from a decreasing in central controller output to the inspiratory pump muscles. The classic daytime manifestation of SDB is the excessive sleepiness, although other

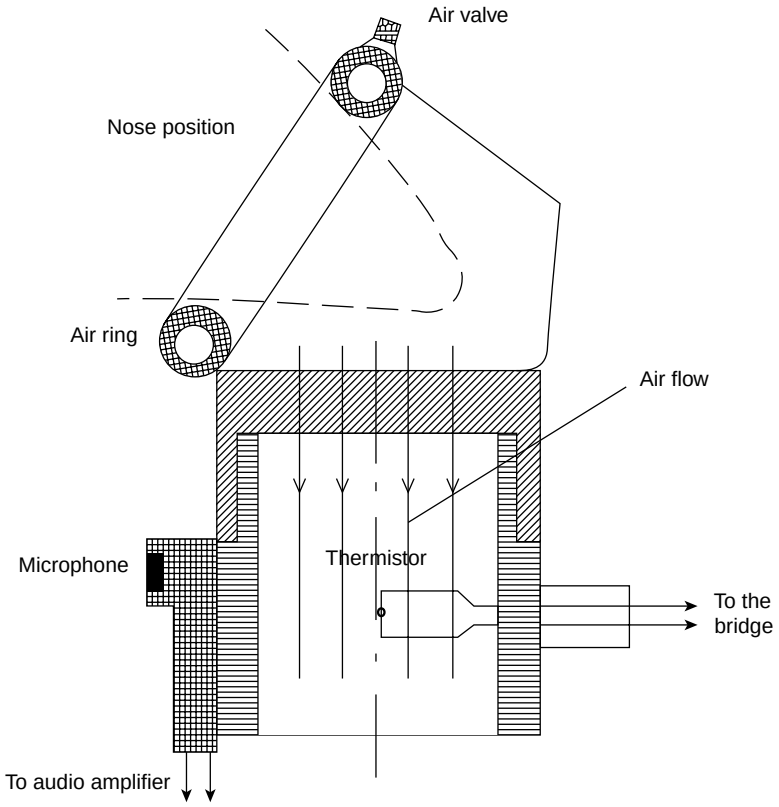


Figure 21. Cross-section showing a thermistor in a mask used by speech therapists to evaluate nasal air escape (Reproduced from (54) with permission of the IFMBE).

symptoms, such as unrefreshing sleep, poor concentration and fatigue are commonly reported. Automobile and industrial accidents have been associated with a poor quality of sleep as well as hypertension and intellectual deficits. The severity of the disease is defined in terms of the apnea/hypopnea index (AHI). This measurement reflects the average number of apneas plus hypopneas observed per hour of sleep, and it is usually derived by identifying and counting each respiratory disturbance, with subsequent division of the sum by the number of hours slept. This way, scientific and diagnostic studies of sleep-disordered breathing are critically dependent of the performance of the measuring device used to detect abnormal respiratory events. Thermistors usually used in SDB diagnosis presents a long time constant (58). This way, they are not able to provide a good characterization of fast events like hypopneas. The effect of the thermistor high measurement time constant can be seen in Fig. 22, in which a simultaneous acquisition of the thermistor and a fast-responding system were done in a awake subject during spontaneous breathing. Fig. 22 shows that ventilatory details were lost by the thermistor system. This loss can be explained by the lowpass filter action of the thermistor, in which the higher order harmonics of the respiration process are discarded. This characteristic limits the accuracy of the clinical diagnosis based in thermistor in sleep studies, because it may introduce underestimation of hypopneas events. It happens because, during hypopneas, the respiratory flow tends to change from a quasisinusoidal (normal breathing) to an almost square-wave pattern. Similar limitation to describe dynamic respiratory events was also observed for thermistors used to monitor air stream temperatures in exercising asthmatic patients (59). Recently, Togawa et al. (60) presented a technique that improved the accuracy of fluctuating temperature measurement by thermistors. This technique seems to present a great potential to improve the performance of these devices in the monitoring of fast events, a field still open to research.

Evaluation of the Thermal Properties of Tissue

The evaluation of thermophysical properties of tissue assume importance with the increasing use of hyperther-

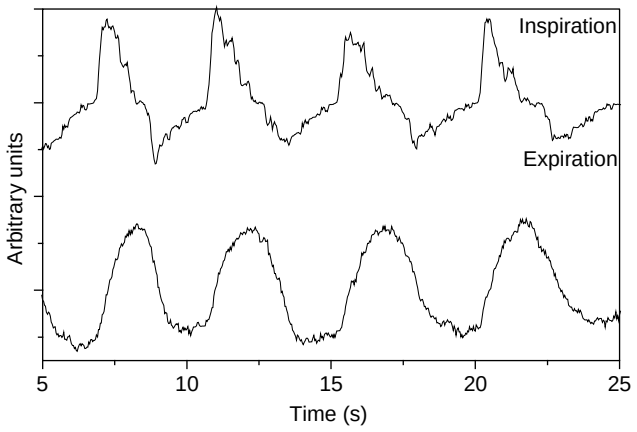


Figure 22. Typical signal morphology obtained in studies with an awake subject during spontaneous breathing by the thermistor (botton trace) and NPR systems (top trace). Note that respiratory details were lost by the thermistor signal (Reproduced from (58) with permission—© 2004, American Institute of Physics).

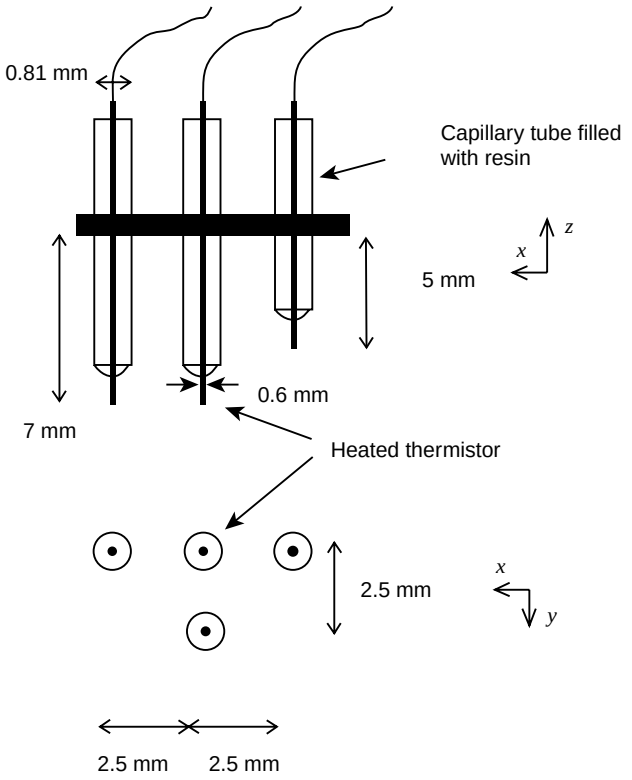


Figure 23. Thermistor thermal probe for the evaluation of thermal conductivity and the thermal diffusivity of biomaterials (Reproduced from (61) with permission—© 2001 IEEE).

mia and therapeutic procedures based on heat delivery, such as radiofrequency, microwave, laser, and ultrasound (61). The thermal conductivity and the thermal diffusivity of biomaterials are measured using thermal probe techniques. Various thermal diffusion probe techniques have been developed from Chato's first practical use of the thermal probe (62). Physically, for all of these techniques, heat is introduced to the tissue at a specific location and is dissipated by conduction through the tissue and by convection with the blood perfusion. The similarity of these techniques is the use of a thermistor bead, either as a heat source or a temperature sensor. The probes are usually constructed by placing a miniature thermistor at the tip of capillary tubes. Fig. 23 shows one example of probe assembly (61). Four probes were used with one heated thermistor and three sensing thermistors in three different locations. The distribution of three sensing thermistors orthogonal to the central heating thermistor permits the determination of the directional nature of thermal properties. The probe is positioned invasively within the tissue of interest. Electrical power is delivered to the heating thermistor and the resulting temperature rise is measured by the sensing thermistors. An empirical relationship between the power delivered by the first thermistor and the temperature raise recorded by the sensing thermistors is used to evaluate the thermal conductivity along the line joining each of the thermistors. The delay between the application of a power pulse in the heated thermistor and the temperature pulse in the sensing thermistors is calculated by cross-correlation and used to determine the diffusivity of the material in each direction. This system allowed the authors to measure the thermal properties of a swine left ventricle *in vivo*, as well as to investigate the effect of ablation. The thermal conductivity and the diffusivity of the tissue dropped after ablation. These data will help to generate three-dimensional (3D) thermal models of the ablation process.

Geriatrics

In the geriatric setting, the urinary incontinence is present in 7–15% of the elderly. The prevalence in elderly living in chronic care hospitals and nursing homes is considerably higher, up to 62% (63). The high sensibility and small size of the thermistor contribute to its application in the management of urinary incontinence, allowing for the development of monitors that can help to reduce elderly reliance on incontinence pads, both improving their quality of life, and reducing the cost of the treatment (64). These monitors are based on the changes in temperature produced by the urinary incontinence event, being developed for long-term ambulatory monitoring (63,64) and for home healthcare (65).

BIBLIOGRAPHY

1. Guyton AC, Hall JE. Textbook of Medical Physiology. 10th ed., Philadelphia: Sounders; 2002.
2. Buono MJ, Ulrich RL. Comparison of mean skin temperature using "covered" and "uncovered" contact thermistors. Phys Meas 1998;19:297–300.
3. Lavenuta G. Negative temperature coefficient thermistors. SENSORS, May 1997, p. 46.
4. Wang CC, Akbar SA, Madou MJ. Ceramic based resistive sensors. J Electroceramics 1998;2(4):273–282.
5. Nenov TG, Yordanov SP. Ceramic Sensors: Technology and Applications. Pennsylvania: Technomic; 1996.
6. Lin CH, Jadvar H. Interfacing temperature sensors. In: Tompkins WJ, Webster JG, editors. Chap. 7. Interfacing Sensors to the IBM PC. Englewood Cliffs, (NJ): Prentice-Hall; 1988. pp. 183–224.
7. Anonymous (1998). Technical Notes, [Online]. Alpha Sensors, Inc. Available at http://www.alphasensors.com/technical.html. Accessed 2005, March 11.
8. Anonymous (2002). General technical information [Online]. Epcos-Electronic Parts and Components. Available at http://www.physics.leidenuniv.nl/edu/courses/Experimentele%20-Natuurkunde/NTC.pdf). Accessed 2005, March 11.
9. Anonymous (2001). YSI Precision Medical Probes and Accessories Catalog, [Online]. Ysi Corporation. Available at http:// http://www.advindsys.com/ysi.htm. Accessed [2005, March 11].
10. Anonymous (No date). Thermistors catalog, [Online]. Honeywell Corporation (Fenwall). Available at http://content.honeywell.com/sensing/hss/thermal/product/thermisters.asp. Accessed 2005, March 17.
11. Anonymous (No date). Thermistor glossary, [Online]. RTI Electronics Inc. Available at http:// http://www.rtie.com/ntc/glossary.htm. Accessed 2005, March 11.
12. Anonymous (2004, December, 14). Application Notes, [Online]. Ametherm Inc. Available at http://www.ametherm.-com/ Aplications/. Accessed 2005, March 11.

13. Anonymous (No date). NTC Thermistors, [Online]. Thermometrics Inc. Available at http:// http://www.thermometrics.-com/assets/images/ntcnotes.pdf. Accessed 2005, March 11.
14. Keskin AU, Yanar TM. Steady-state solution of loaded thermistor problems using an electrical equivalent circuit model. Meas Sci Technol 2004;15:2163–2169.
15. Sheingold DH. Transducer Interfacing Handbook. Noorwood (MA): Analog Devices; 1981.
16. Normann RA, Principles of Biomedical Instrumentation. New York: John Wiley & Sons; 1988.
17. Anonymous (2002). Application Notes [Online]. Epcos—Electronic Parts and Components. Available at http://www.epcos.-com/inf/50/db/ntc02/00290045.pdf. Accessed 2005, March 11.
18. Beakly WR. The design of thermistor thermometers with linear calibration. J Sci Instrum. 1951;28:176.
19. O'Grady A. Building a more perfect union: combining thermistors and high-resolution $\sum\Delta$ A/D converters, Sensors Online Available at www.sensorsmag.com/articles/0100/42/main.shtml, Accessed 2000.
20. Lyons P, Waterworth P. The use of NTC thermistors as sensing devices for TEC controllers and temperature control integrated circuits, technical report, Betatherm Ireland Ltd.
21. Anonymous (1998). Application Notes, [Online]. Alpha Sensors, Inc. Available at http://www.alphasensors.com/tappnotes.html. Accessed 2005, March 11.
22. Hoge HJ. Comparison of circuits for linearizing the temperature indications of thermistors. Rev Sci Instrum 1979;50:(3).
23. Sheingold DH. Nonlinear Circuits Handbook. Noorwood, (MA): Analog Devices; 1976.
24. Sundqvist B. Simple, wide-range linear temperature-to-frequency converters using standard thermistors. J Phys E Sci Instrum 16:261–264.
25. Patranabis D, Ghosh S, Bakshi C. Linearizing transducer characteristics. IEEE Trans Instrum Meas. 1988;37:66–69.
26. Sengupta RN. A widely linear temperature to frequency converter using a thermistor in a pulse generator. IEEE Trans Instrum Meas 1988;37:62–65.
27. Slomovitz D, Joskowicz J. Error evaluation of thermistor linearizing circuits. Meas Sci Technol 1990;1:1280–1284.
28. Kaliyugavaradan S, Sankaran P, Murti VGK. A new compensation scheme for thermistor and its implementation for response linearization over a wide temperature range. IEEE Trans Instrum Meas 1993;42(5):952–956.
29. Kaliyugavaradan S, Sankaran P, Murti VGK. Hardware linearization of thermistor response using series-parallel resistors for temperature-to-time conversion. Meas Sci Technol 1994;5:786–788.
30. Woodwards S. Pseudologarithmic thermistor signal conditioning spans wide temperature range, EDN 2005;50(1):66–67.
31. Ghosh D, Patranabis D. Software linearization of thermistor type nonlinearity. IEE Proc Circuits Devices systems 1992;139(3):339–342.
32. Finnie B. Software linearization techniques for thermistors. Electronic Design Available at www.elecdesign.com/Articles/Index.cfm?AD=1&ArticleID=6338. Accessed 1998.
33. Day J, Bible S. (2004). Piecewise linear interpolation on PIC 12/14/16 series microcontrollers, application note no. 942. [Online]. Microchip Technology Inc. Available at http:// ww1. microchip.com/downloads/en/AppNotes/00942A.pdf. Accessed 2005, March 11.
34. Hagerman J. Model thermistors with spice. Elect Design 1991;39(5):85.
35. Wangenheim L. SPICE subcircuit models thermistors. EDN July 1997.
36. Boylestad RL, Nashelsky L. Electronic Devices and Circuit Theory. Englewood Cliffs (NJ): Prentice-Hall; 2001.
37. Bishop J. Thermistor temperature transducer-to-ADC application. Analog Appl J (Texas Instr Inc). 2000; 44–47.
38. Joyce D. (No date). Practical aspects of thermistors, Technical note no. 124. [Online]. Zwold Inc. Available at http:// www.zworld.com/support/techNoteswhitePapers.shtml. Accessed 2005, March 11.
39. Anonymous (March, 2004). A high resolution/precision thermometer using ST7 and NE555, [Online]. ST Microelectronics, Application note no. AN1755. Available at http:// http:// mcu.st.com/familiesdocs-15.html. Accessed 2005, March 11.
40. Wei D, Saidel GM, Jones SC. Thermal method for continuous measurement of cerebral perfusion. Med Biol Eng Comput 1994;32(5):481–488.
41. Evans NE, Hajnayebi HR. Clinical temperature acquisition using proximity telemetry. J Biom Eng 1991;13:83–86.
42. Baker B. Thermistors in single supply temperature sensing circuits, AN685, Microchip Technology Inc., 1999.
43. Anonymous (2002). Four-channel termistor temperature-to-pulse-width converter, datasheet, [Online]. Maxim Integrated Products. Available at http://www.maxim-ic.com. Accessed 2005, March 11.
44. Potter D. (June, 1995). Measuring temperature with thermistors – a tutorial, [Online]. National Instruments Inc., Application note no. 65. Available at http:// http://www.seas.upenn.edu/courses/belab/ReferenceFiles/Thermisters/an065.pdf. Accessed 2005, March 11.
45. Olson WH. Electrical safety. Webster JG, editor. Medical Instrumentation: Application and Design. 3rd edition. New York: John Wiley & Sons; 1998.
46. Harrison RM, Smith SD, Roberts JA. Testicular temperatures measured by thermistor probe and contact thermography. Fertility Sterility 1990;54(1):173–174.
47. Thoresen M, Whitelaw A. Cardiovascular changes during mild therapeutic hypothermia and rewarming in infants with hypoxic-ischemic encephalopathy. Pediatrics 2000;106:92–99.
48. Sakuragi T, Mukai M, Dan K. Deep body temperature during cardiopulmonary bypass. Br J Anaest 1993;71(4):583–585.
49. Dollberg S, Rimon A, Atherton HD, Hoat SB. Continuous measurement of core body temperature in preterm infants. Am J Perinatol 2000;17(5):257–264.
50. Muir IH, Bishop PA, Lomax RG, Green JM. Prediction of rectal temperature from ear channel temperature. Ergonomics 2001;44(11):962–972.
51. Martin GT, Bowman HF. Validation of real-time continuous perfusion measurements. Med Biol Eng Comput 2000;38:319–325.
52. Webster JG. Measurement of flow and volume of blood. Webster JG, editor. Medical Instrumentation: Application and Design. 3rd edition. New York: John Wiley & Sons; 1998.
53. Primiano Jr. FP. Measurement of the respiratory system. Webster JG, editor. Medical Instrumentation: Application and Design. 3rd edition. New York: John Wiley & Sons; 1998.
54. Besar SS, Kelly SW, Greenhalgh PA. Nasal airflow measurement using a compensated thermistor anemometer, Part 1 system desciption and quantitative analysis.. Med Biol Eng Comput 1989;27:628–631.
55. Besar SS, Kelly SW, Greenhalgh PA. Nasal airflow measurement using a compensated thermistor anemometer, Part 2 Computer signal processing and quantitative analysis. Med Biol Eng Comput 1990;28:127–132.
56. Mirlohi HR, Kelly SW, Manley MCG. New technique for assessment of velopharingeal function. Med Biol Eng Comput 1994;32:562–566.
57. Jovanov E, Raskovic D, Hormigo R. Thermistor-based Sensor for circadian rhythm evaluation. Biomed Sci Instr 2001;37:493–497.

58. Mesquita Jr JA, Melo PL. A respiratory monitoring system based on pressure measurement for the analysis of sleep breathing disorders: Static and dynamic errors reduction and comparisons with thermistors and pneumotachographs. Rev Sci Instrum 2004;75(3):760–767.
59. Clary AL, Fouke JM. Fast-responding automated airway temperature probes. Med Biol Eng Comput 1991;29:501–504.
60. Tagawa M, Kato K, Ohta Y. Response compensaion of thermistors: Frequency response and identification of thermal time constant. Rev Sci Instrum 2003;74(3):1350–1358.
61. Naresh CB, et al. Measurement of directional thermal properties of biomaterials. IEEE Trans Biomed Eng 2001; 48(2):261–267.
62. Chato JC. A method for the measurement of thermal properties of biological materials. Proceedings of the Symposium on Thermal Problems in Biotechnology New York, paper LCN068-58741, 16–25, 1968.
63. Hurk PRB, et al. Long-term ambulatory monitoring of urine leakage in the elderly: An evolution of the validity and clinical application of the thermistor signaling. J Med Eng Technol 1998;22(2):91–93.
64. Cusick G, et al. A system for logging incontinence events using a simple sensor. Proceedings of The Institution of Mechanical Engineers part H. J Eng Med 2003;217(H4): 30–310.
65. Tamura T, et al. A warning detector for urinary incontinence for home health care. Biom Instrum Technol 1995;29:343–349.

Further Reading

Books: The following books contain several useful information concerning thermistors, its use and instrumentation. Tompkins WJ, Webster JG. Eds. Design of Microcomputer-Based Medical Instrumentation. New Jersey: Prentice Hall; 1981.
Enderle J, Blanchard S, Bronzino J. Introduction to Biomedical Engineering. London: Academic Press; 2000.
Doebelin EO. Measurement Systems: Application and Design. New York: McGraw-Hill; 1990.
Cobbold RSC. Transducers for Biomedical Measurements: Principles and Applications. New York: John Wiley & Sons; 1974.
Papers: Thermistor general characteristics and technology: Keskin AU. A simple analog behavioral model for NTC thermistors including self-heating effect. *Sensors and Actuators-A* 2005; 118-2:244–247.
Veijola T. Electrothermal simulation models for NTC and PTC thermistors. *Proceedings of CSC 98 1998*;2, Piraeus: 950–955.
Broughton MB. Analysis and design of almost-linear one thermistor temperature transducers. *IEEE Trans Instrum Meas* 1974. 23(1):1–5.
Jung JS, et al. Reliability evaluation and failure analysis for NTC thermistor. *Int J Mod Phys B* 2003;17(8–9):1254–1260.
Pathan ZB, Shaligram AD. A PC-based characterization set-up for reliable parameter testing of thermistor. *Indian J Pure Appl Phys* 1995;33(4):215–219.

Clinical Applications of NTC Thermistors

Farré R, et al. Accuracy of thermistors and thermocouples as flow-measuring devices for detecting hypopneas. Eur Resp J 1998;11:179–182.
Bhavaraju NC, et al. Measurement of directional thermal properties of biomaterials. IEEE Trans Biom Eng 2001;48(2):261–267.
dos Santos I, et al. Measurement of ejection fraction with standard thermodilution catheters. Med Eng Phys 2002;24(5):325–335.
Cui R, et al. A needle temperature microsensor for in vivo and real-time measurement of the temperature in acupoints. Sensors Actuators A: Phys 2005;119(1):128–132.

Technical Reports

Doug C. Implementing Ohmmeter/Temperature Sensor. AN512, Microchip Technology Inc.
Rodger R. Resistance and Capacitance Meter Using a PIC16C622. AN611, Microchip Technology Inc.
Joyce D. (No date). Practical aspects of thermistors, Technical note no. 124. [Online]. Zwold Inc. Available at http://www.zworld.com/support/techNoteswhitePapers.shtml. Accessed 2005, March 11.

WWW Pages : Manufactures of Thermistors and Associated Products

Honeywell/Fenwall: http://content.honeywell.com/sensing/hss/thermal/.
Advanced Thermal Products: http://www.atpsensor.com/.
Alpha Sensors: http://www.alphatechnicsonline.com/.
Betatherm: http://www.betatherm.com/products/index.php.
Precision Engineering: http://www.pel-ltd.co.uk/english/spec.htm.
Thermometrics: http://www.thermometrics.com/.
Yellow Springs Instruments Inc.: http://www.ysi.com/temperature.htm.
Omega Engineering: http://www.omega.com.

See also HYPERTHERMIA, SYSTEMIC; TEMPERATURE MONITORING.

THERMOCOUPLES

JOAKIM WREN
DAN LOYD
Linköping University
Linköping, Sweden

INTRODUCTION

Accurate temperature measurement is of utmost importance in many medical, biological, and biomedical applications. Technological development has increased both the performance and the range of equipment for temperature measurement, but at the same time the apatite for better measurements has increased as well. Opposite to what might be expected, this in fact increases the demands on the choice and especially the use of equipment for temperature measurement, as aspects associated with the particular application becomes relatively more important.

Despite the large variety of sensors and associated instrumentation, there is a very limited amount of principles or "physics" behind the different sensor types, at least when the most widely used sensors are considered (1); thermocouples, thermistors, and resistance based sensors (e.g., Pt-100). A thermocouple is an active sensor meaning that by itself it can give rise to a detectable signal. Thermistors and resistance-based sensors are passive as they demand an outer-power source. What all these sensors

have in common is that they only measure their own temperature. Thus, they interact thermally with the object to be measured, which can give significant measurement errors.

In order to carry out measurements with sufficient accuracy at a reasonable cost, several demands are put on the equipment and its use. All sensors and measurement systems have their benefits and limitations, and this should be taken into account together with the aspects of the present measurement situation when the measurement equipment is chosen. This process often involves analysis of accuracy, stability, time dependence, and environmental factors, for example, exposure to electromagnetic (EM) and thermal disturbances.

THEORY

Temperature Measurement and the Laws of Thermodynamics

Temperature measurement is based on the zeroth and first laws of thermodynamics. The zeroth law, as the name implies, was stated after the first and second laws, but it was considered to be of such importance that it was denoted the zeroth law. It states that if two bodies are in thermal equilibrium with a third body, they are also in thermal equilibrium with each other. For a further thermodynamic discussion see Ref. 2. From a temperature measurement perspective, it is interesting to replace the third body by a thermometer, for which the zeroth law can be restated as two bodies are in thermal equilibrium if they have the same temperature even if they are not in contact. The law thus tell us what we know from intuition, namely, that systems of different temperatures and in thermal contact with each others strive to equalize their temperatures. This is a very fundamental statement, although in many situations we must consider other aspects as, for example, heat transfer inside the considered systems and between the system(s) and the(ir) surroundings.

The above discussion emphasizes the importance of the first law of thermodynamics, which states that energy is conserved. In reality, all bodies are exposed to a thermal environment at another and often varying temperature that gives a time-dependent heat flux between the bodies and their environment. The difference and/or the variation in temperature and corresponding heat flux may sometimes be small; often, however, they are significant and can be considered as one of the main sources of measurement errors, which will be discussed in the next section. The principle of conservation of energy does not only cause trouble in the form of measurement errors, it can also be beneficial to use in the frequently used "lumped heat-capacity method" (see, e.g., Ref. 3).

Thermocouples

Thermocouple temperature measurement is based on the electrothermal phenomena known as the Seebeck effect after its discoverer T.J. Seebeck in 1821. If two dissimilar metals are connected in a circuit, as in Fig. 1, a temperature dependent electromotive force (EMF) arises in the circuit. The magnitude of the EMF is dependent on the materials and the temperature at the junctions.

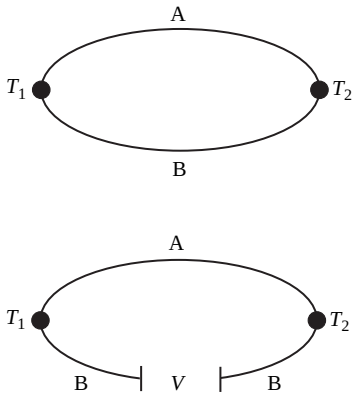


Figure 1. (a) Closed and (b) open thermocouple circuits with conductors A and B and junction temperatures T_1 and T_2.

Since the junctions are electrically connected and subject to different electrical potentials, a current will flow in the circuit; thus, thermal energy is converted into electric energy (see, e.g., Ref. 4). The current flow makes the Seebeck effect related to the Peltier and Thomson effects (see, e.g., Ref. 5). Peltier discovered in 1834 that if a current flows in a circuit made of two dissimilar conductors as in Fig. 1, one of the junctions becomes cold and the other junction hot. Thomson later found that for a circuit subject to a temperature gradient (heat flux), heat is rejected in any point where current and heat flows in the same direction, and is absorbed where the flows are countercurrent.

If a voltmenter is inserted in the circuit, as in Fig. 2b, it is possible to measure the EMF and relate it to the junction temperatures, as the voltage is dependent on the difference between temperatures T_1 and T_2. Temperature measurement using a thermocouple is in fact measurement of a temperature difference between the two junctions.

It is important that the voltmeter has a sufficiently high resistance to keep the current small in order to eliminate disturbances related to the Peltier and Thomson effects (see, e.g., Ref. 5). This is normally not a problem, as modern voltmeters have sufficiently high impedance for these effects to be completely negligible.

Before modern measurement systems became available, the thermocouple was normally used by keeping one of the junctions (the reference junction) at a known temperature; typically an insulated ice bath that is very close to 0 °C, whereas the other junction (the sensor) was used for temperature measurement. The measured voltage is proportional to the temperature difference between the junctions,

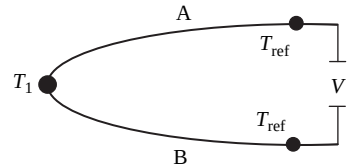


Figure 2. The simplest form of a thermocouple measurement circuit. The sensor measures temperature T_1, whereas the reference junction temperature T_{ref} is integrated in the measurement system.

and since the ice bath temperature is known, the temperature of the other junction can be determined if the thermoelectric properties of the wires are known.

Today, the reference junction is more or less always integrated in the measurement system. The determination of the reference temperature is normally optimized for the instrument at room temperature (20–25 °C), although temperatures close to this interval can often be used accurately for most instruments. The circuit corresponding to a standard connection is given by Fig. 2, where the reference temperature is measured at the transition between the thermocouple leads (or extension leads) and the instrument connection. For two conductors A and B, the relation between the sensor temperature T_1, the reference temperature T_{ref} and the measured EMF (voltage, V) follows the equation

$$V = \int_{T_{\text{ref}}}^{T_1} S_{\text{A}}(T)dT - \int_{T_{\text{ref}}}^{T_1} S_{\text{B}}(T)dT = \int_{T_{\text{ref}}}^{T_1} S_{\text{AB}}(T)dT \qquad (1)$$

where $S_{\text{A}}(T)$, $S_{\text{B}}(T)$ and $S_{\text{AB}}(T)$ are the Seebeck coefficients (also called the thermoelectric power–sensitivity) for the respective conductors A and B, and the thermocouple AB. The EMF corresponding to more complex thermoelectric circuits is discussed in depth in Ref. 6.

If a commercial measurement system is not available or a designed circuit for inclusion in a larger system or stand alone is preferred, today there are special-purpose integrated circuits with built-in cold junction compensation (see, e.g., the example in Fig. 3).

Thermocouple Types

All metallic conductor pairs give rise to an electric potential, and thus have the potential to be used as a thermocouple. It is, however, convenient to use one of the standardized thermocouple types that has emerged, which offers obvious advantages, such as compatibility with commercial instruments, predetermined Seebeck coefficients, and a potential–temperature relationship and International Electrotechnical Commission (IEC)-specified tolerance classes.

There are about a dozen types of thermocouples that complement each other in terms of the measuring signal, temperature range, and tolerance to different environments. Type K is the most commonly used and it is a good compromise of price and performance for many situations, although other types might be the choice for very high/low temperatures, hostile environements, and so on. Perhaps the most important difference between the various types is the output signal determined by the relative Seebeck coefficient of each material. Types S, B, and R contain platinum and are therefore more expensive than the others, but they are also more stable in, for example, e.g., oxidative environments.

There are organizations/national organizations that provide standards for thermocouples, but the standards of the IEC are internationally recognized and should be followed whenever possible. The IEC 584-1, last revised in 1995, contains reference tables and calculation polynomials for the output signals of the standardized thermocouple types as a function of temperature.

THERMOCOUPLE DESIGN

Wire Sensors

In its simplest form, a thermocouple element consists of a pair of wires that are connected together at a measuring junction. The junction must be electrically conducting and can be formed by soldering, crimping, or twisting the wires together, depending on what the situation requires. Wire sensors are mainly used to measure low temperatures in favorable environments. One limitation is imposed by the insulation material [e.g., poly(vinyl chloride) (PVC), nylon, Kapton or poly(tetra fluoroethylene) (PTFE). Another limitation is that the measuring junction is exposed to the atmosphere. The PVC insulation can withstand temperatures up to $\sim$ 100 °C, whereas certain ceramic materials are tolerant of temperatures up to 1000 °C or more.

Sheathed Thermocouples

In sheathed thermocouples, the wires are normally insulated by densely packed magnesium oxide enclosed in a metal alloy suitable for the given thermocouple element.

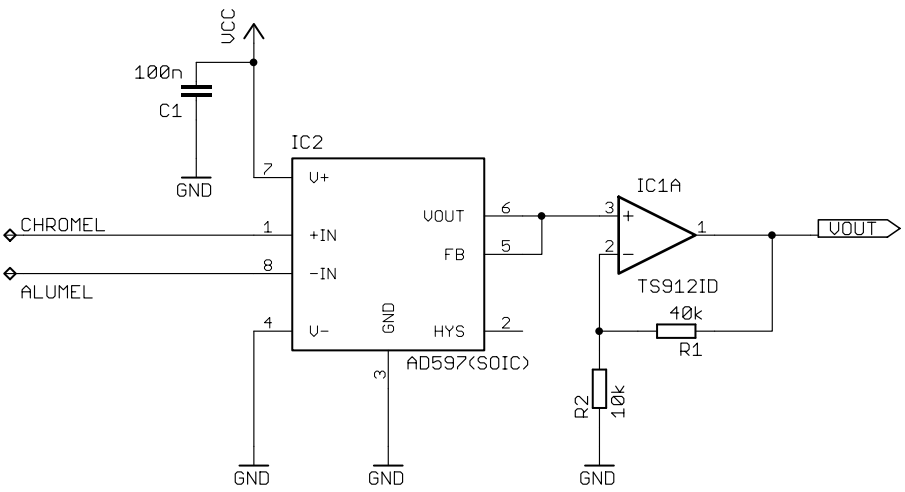


Figure 3. An example of a circuit for temperature measurement using a thermocouple type K. The thermocouple controller AD597 (Analog Devices, Norwood, MA) contains a cold junction compensation circuit, and the OP-amplifier TS912ID (ST Microelectronics, Geneva, Switzerland) amplifies the output "V OUT" so that 0 V equals 0 °C and 5 V equals 100 °C with the present impedances. The thermocouple is connected to the "+IN" and "−IN" connections, and the low pass filtered voltage supply is connected at "V+". There are several manufacturers of similar circuits; for details see the companies product-specific information.

Disadvantages associated with sheathed thermocouples are their relatively high prices and the absence of very thin sensor. Sheathed thermocouples have several advantages, such as (*1*) The metal sheathing is hermetically sealed, which makes the sensor useful in hostile environments. (*2*) It is tolerant to higher temperatures than wire sensor of corresponding size. (*3*) Sheathing can be easily bent to, and withhold, various shapes. (*4*) Vibration resistant.

Sheathing is available in different diameters and with exposed, grounded, and insulated measurement junctions (7,8). The exposed junction protrudes from the sheath and the tube is sealed, for example, by glass. The advantage is the shortest possible response time for sheathed thermocouples. However, several advantages of the fully enclosed sheathed thermocouple are lost; tolerance to high temperatures is one example. For the grounded junction, the wires are welded to the bottom of the sheathing tube. This gives a fairly fast response and good environmental protection, but sensitivity to ground currents. The insulated junction is not electrically connected with the sheath, which gives good protection against EM interference, but a slow temperature response.

MEASUREMENT ERRORS

Reliable temperature measurement is dependent on many factors that, more or less, inevitably leads to measurement errors. Consider, for example, a thermal treatment, such as radio frequency (RF) current ablation. During such a treatment, the temperature should ultimately be known in the entire heated area at every instant. However, this is obviously not possible. In practice, the temperature measurement is limited to a few single locations. In some situations, it might not even be possible to measure the temperature in any location inside the treatment area. In such cases, temperature information must be extracted from measurement outside of the treatment area, which gives space, time, and temperature uncertainties.

Insufficient Contact between Sensor and Measurement Object

Surface-mounted sensors are not only particularly sensitive to thermal shunting, but also to errors associated with an insufficient contact between the sensor and the measurement object. Suppose, for example, that a spherical sensor is placed in contact with a flat surface surrounded by air. Only a fraction of the circumference of the sensor will be in contact with the object to be measured, while the rest of the sensor is exposed to the temperature of the air. Thus, it is essential that the maximum possible contact surface be achieved between the sensor and the object to be measured, and/or that the sensor is thermally insulated from the surrounding.

Effects from adjacent hot or cold surfaces are especially treacherous as thermal radiation can effect the sensor temperature over relatively vast distances. An example is measurement of (hot) gas temperature in a chamber with cold walls; the walls will exchange energy with the sensor resulting in sensor temperature that is too low.

Thermal Shunting

A temperature sensor conducts heat, which gives what is called thermal shunting. The problem is particularly marked when the temperature is measured on pipes and other surfaces. The greater the temperature difference, the greater quantity of heat is "drawn out" of the body by the sensor. There is a sharp drop in temperature at the point where the sensor protrudes from the object to be measured. This is caused by a load being imposed on the object to be measured. The temperature drop is greatest when the object to be measured is a poor conductor of heat (low thermal conductivity). To overcome this measurement problem, the point at which the sensor leaves the body to be measured must be at a suitable distance from the measuring junction.

Temperature Gradients and Thermal Conduction

Heat losses through the probe occur when a sensor connects a warm zone with a colder one, especially if the sensor is the easiest path for the transfer of heat. Heat is conducted away from the object being measured to the probe. The heat flux gives a temperature drop and the sensor measures a lower temperature than the true value. Such losses can be effectively countered if we ensure that the thermal conductivity in the object and across the probe tip is much higher than it is along the sensor. In other words, more heat can be transferred to the sensor than conducted away by the sensor sheath.

Response Time

Since the sensor only measures its own temperature, there will almost always be a temperature deviation between the temperature of the sensor and its surrounding. In theory, the sensor temperature asymptotically approaches, but never reaches, the surrounding temperature. For practical purposes, the response time can be regarded as finite, but it can be both negligibly small and cause severe measurement errors.

The response time is defined as the time needed for the sensor to reach some fraction of a stepwise temperature change of the surrounding, often 90% or $1-(1/e)=67\%$. However, the response time of a sensor must be accompanied by information about the conditions under which the response time was obtained. The reason is that heat must be transferred from the surroundings to the sensor. This transfer takes some time, and the time needed differs greatly between different conditions, such as still air and stirred water. Too often, information of the conditions associated with a particular response time are left out, making the information useless. The response time is connected to the sensor and its surrounding and not the sensor itself.

Some critical factors affecting the response time of a sensor in contact with a body or medium include the following:

The heat capacity of the sensor. The greater heat capacity and mass, the longer response time.

The heat transfer in the materials. Air gaps and insulation, for example, reduce the heat transfer.

The contact surface between the sensor and the body or medium to be measured.

The response time can change over time; it can become shorter if for example, the junction of a sheathed probe is moved closer the sheath, or longer if, for example, a glue become brittle and porous and thereby gets different heat-transfer properties.

Self-Heating

Errors due to self-heating can be quite important in medical and biomedical applications depending on the small sensors often required (9). Resistance-based sensors and especially thermistors may be substantially affected by this type of error, but thermocouples are not affected as they are active sensors. All these types can, however, be affected by EM interference (see the sections Electromagnetic Interference and Electromagnetic Aspects).

Material Defects and Ageing

All types of thermocouple are subjected to varying degrees of wear and ageing depending on the environments in which they are used. It is therefore essential that all types of sensors are regularly inspected and calibrated. The type K thermocouple is the most widely used, and therefore the best documented (see, e.g., Ref. 7).

Use of Alternative Sensor Materials. Some thermocouples are expensive and it can therefore be tempting to use other materials as extension leads. As explained earlier, however, the Seebeck effect is created throughout the measuring circuit, so the use of other materials having a different Seebeck coefficient can give rise to a faulty output signal (7). There are materials, known as compensating leads, that have the same electrical properties as the thermocouple within limited temperature ranges, but it is always better to use thermoelectric material throughout the circuit. If this is not possible, compensating leads for the type of thermocouple used must be employed.

Alternative Probe/Junction Materials. Other materials are sometimes used for the measuring junction. A typical case is when the thermocouple wires are soldered to a tab, which is secured by screws at the required measuring point. This works if the tab has good electrical conductance and the temperature is the same at both ends of the wire. The wires must be as close together as possible so that they get the same temperature that constitutes the measured value.

Ageing. The ravages of time make their mark on all types of thermocouples. In general the process is accelerated by high temperatures ($\gtrsim 200\,^{\circ}\mathrm{C}$), vacuum, and various atmospheres. It is therefore of interest for only a limited amount of medical applications.

At high temperatures, short-range order (SRO) is a hysteretic phenomenon that is based on atomic migration inside the thermocouple, especially type K thermocouples. Also type S thermocouples, which has wires of platinum–rhodium and pure platinum, are vulnerable to atomic migration. At high temperatures, the rhodium vaporizes and drifts across to the pure platinum wires, which results in a gradual fall in the output signal (see, e.g., Ref. 7.)

Connection Errors

Circuit Break (Open Circuit). A sensor wire has fractured, come adrift, or is making poor contact with the instrument. Modern instruments often trigger an alarm, for example, by Open appearing on the display.

Short Circuit. If the insulation has chafed and a short circuit occurs, a new measuring junction is created. The instrument will then display the temperature at the short circuit point, instead of at the tip of the probe. Sometimes, it can be very difficult to detect this type of error.

Reversed Polarity of Entire Measuring Circuit. If the polarity has been reversed, the instrument will also operate in reverse, that is, a temperature increase will be recorded as a temperature decrease.

Reversed Polarity within the Measuring Circuit. The extension lead must have the same polarity as the thermocouple wires. If the polarity of the thermocouple element is reversed, opposing voltages occur. The reading obtained will then be twice the temperature in the terminal head minus the temperature at the measuring junction.

Double Reversed Polarity. If the polarity of the extension lead has been reversed at both ends, the temperature at the ends will affect the output signal. The reading will be the temperature at the measuring junction less twice the temperature difference between the terminal head and the reference junction.

Electromagnetic Interference

Electromagnetic fields interact with materials inside the field. Thermocouples and other measurement sensors–probes are no exception, and the frequent occurrence of EM fields, especially in the medical–biomecical engineering sector, has contributed to many measurement errors.

The properties of an EM field change greatly with frequency. Interaction with the field is highly dependent on the geometry and the material of objects inside the field, which together with the great frequency spectrum of interest, makes evaluation of measurement interference multifaceted and intricate. Thus it is very difficult to predict and evaluate.

Measurement errors can arise due to direct effects, for example, induced current in the thermocouple wires (10). Remember that $1\,^{\circ}\mathrm{C}$ corresponds to $\sim 50\,\mu\mathrm{V}$ for some typical thermocouples. Also, small currents can give significant measurement errors. Induced currents can also heat the wires and the measurement junction, leading to a too high temperature reading. Indirect disturbances can arise, for example, if the inserted probe disturbs the EM field. Such indirect effects are not only associated with metallic probes–material, but plastics as well (11).

THERMOCOUPLES IN MEDICINE AND BIOMEDICAL ENGINEERING

Thermal aspects, and thereby temperature measurement, are important in numerous applications in medicine and biomedical engineering. Thermocouple thermometry is still the dominant technique. Recent publications cover such varying fields as heating of pacemaker leads during magnetic resonance imaging (MRI) (12), estimation of brain protection during cardiopulmonary surgery (13), cryogen spray cooling (14), and thermal treatment using cryotherapy of breast fibroadenomas (15). Although the applications vary over a wide range, many of the fundamental questions are the same.

The following review discusses some applications in medicine and medical devices with a focus on temperature measurement during thermal treatment (treatment of disease and/or symptoms using heat or cold as one of the therapeutic agents). This is one of the major areas of interest for temperature measurement in medicine, and the application involves most of the interesting aspects, problems, and sources of errors associated with thermocouple thermometry. A more thorough review of the use and sources of errors of thermocouples in medicine can be found in Ref. 16, and some practical limitations are discussed in depth in Ref. 17.

Thermal Aspects

Methods, techniques, and equipments of thermal treatment cover a broad spectrum of applications, and the underlying medical as well as physical aspects are often substantially different. Many thermal and temperature measurement considerations are, however, equivalent, which is an important conclusion, as thermal aspects are often the dominant source of measurement errors (see, e.g., Ref. 18.

During thermal treatment, as in many other applications in medicine and elsewhere, it is often difficult to measure the temperature at the most important–interesting location(s). Furthermore, it is rarely a single or even a few temperature values that are solely of interest. More likely it is an entire temperature distribution:the temperature field. The temperature distribution is affected by many factors, and therefore is difficult to predict. During thermal treatment, time, temperature, and individually dependent blood perfusion are the most important and difficult aspects to consider. This is also the situation for other applications associated with human–animal temperature measurement at cites with a heterogenous temperature distribution. Regardless of the application used, the relation between the measured temperature and the corresponding temperature distribution must always be taken under consideration.

An obvious example is functional neurosurgery using RF thermoablation. During such a treatment, the temperature is increased to 70–90 °C during 60 s in a small volume ($\sim 100\,\mathrm{mm}^3$) surrounding the treatment electrode in order to destroy a malfunctioning tissue area. It is important to monitor the temperature in order to control the therapy, but practical difficulties have so far made intratissue temperature measurement impossible. The thermocouple temperature reading is instead carried out inside the electrode that inevitably leads to both spatial and temporal temperature errors as both the maximum temperature and the interesting temperature distribution is located outside the electrode (19). The only possibility to carry out the treatment safely and efficiently is to map the temperature reading with the temperature distribution surrounding the electrode; experiments *in vivo*, *in vitro*, as well as modeling and computer simulations, can be useful tools (see, e.g., the review in Ref. 18).

Another major problem from a temperature measurement perspective is large temperature gradients that can give rise to measurement errors associated with uncertainty in probe positioning (20). Temperature gradients of $10^4\,^\circ\mathrm{C}\cdot\mathrm{m}^{-1}$ and even more are common during thermal treatment. The obvious effect is that an erroneous positioning of only a few tenths of a millimeter in such a steep gradient gives a measurement uncertainty of several degrees. This effect is particularly problematic since it by definition is an uncertainty and thus difficult to predict and compensate for.

Large temperature gradients also give multifacetted thermal conduction effects (21,22) as the probe in general has a different thermal conductivity compared with the measurement object (e.g., tissue). If the thermal conduction is larger in the probe compared with the tissue (as is the normal case), the probe will locally drain heat from the (heated) measurement location affecting the temperature distribution, while the probe itself experiences a temperature different from the adjacent tissue. Both these effects are highly time dependent; rapid temperature disturbances can initially be obtained during, for example, probe insertion and heating onset, although the probe-induced change of the temperature distribution can take much more time. Fortunately, it is possible to reduce the errors associated with these effects. The thermal conduction in the probe relative to the measurement object (e.g., tissue) can be estimated and compensated for, for example, by setting up a computer model that simulates the situation with and without the conduction effects (see the discussion in Ref. 18). Some theory of thermal conduction effects together with analytical, numerical, and experimental results, can be found in Ref. 23.

Electromagnetic Aspects

The use of EM energy for tissue heating purposes makes verification of the EM power deposition important. Since it is difficult to measure and model the EM distribution in general heterogeneous tissue and to obtain from this information an estimate of the corresponding heating effect, a more straightforward approach is by direct estimation of the heating deposition outgoing from the temperature measurement. This has been carried out using thermocouples in order to obtain information of the specific absorption rate (SAR), for example, in deep body regional hyperthermia (24).

When a probe is inserted in an EM field, a field perturbation followed by a changed power absorption in the vicinity of the probe might occur. This yields for all EM

fields, but the significance increases with the field strength that can cause troubles in, for example, thermal treatment and MRI. The volume occupied by a metallic material (probes) has substantially different electromagnetic properties compared with tissue. This can substantially change the field locally, and thereby also change the induced heating followed by an increased temperature.

Although thermocouples are not immune to either direct or indirect EM interference, the possibility to use very small sensors (conductor diameter $\sim$0.1 mm) may for some applications keep this effect within reasonable or even negligible limits. In addition to the size, the shielding (11), the direction of the wires relative to the EM field (25), and the strength and frequency of the EM field, are all of importance to reduce EM interferance. Perhaps the most serious source of direct EM interference is joule heating of the sensor–wires due to induced currents, which can cause heating of the sensor itself, as well as the surrounding tissue (21). Interference can also be reduced by an appropriate design of the measurement electronics and/or suppressed by filters (see, e.g., Refs. 10,16).

BIBLIOGRAPHY

1. Doebelin EO. Measurement Systems—Application and design. 4th ed. New York: McGrawHill; 1990.
2. Cengel Y, Turner R. Fundamentals of Thermal-Fluid Sciences. 2nd ed. New York: McGraw-Hill; 2005.
3. Holman JP. Heat Transfer. 9th ed. New York: McGraw-Hill; 2002.
4. Reed R. Manual on the Use of Thermocouples in Temperature Measurement. American Society for Testing and Materials; 1993.
5. Dike PH. Thermoelectric Thermometry. Philadelphia: Leeds and Northrup; 1954.
6. Moffat RJ. The gradient approach to thermocouple circuitry. Temp Meas Control Sci 1962;3(2).
7. Pentronic AB (No date). Thermocouples. [Online]. Available at http://www.pentronic.se/eng. Accessed 2005.
8. Omega Engineering Inc. (No date). Introduction to thermocouples. [Online]. Available at http://www.omega.com/techref/themointro.html. Accessed 2005.
9. Valvano JW, Nho S, Anderson GT. Analysis of the Weinbaum-Jiji model of blood flow in the canine kidney cortex for self-heated thermistors. ASME—J Biomech Eng 1994;116: 201–207.
10. Chakraborty DP, Brezovich IA. Error sources affecting thermocouple thermometry in RF electromagnetic fields. J Microwave Power 1982;17(1):17–28.
11. Chan KW, Chou CK, McDougall JA, Luk KH. Changes in heating pattern due to perturbations by thermometer probes at 915 and 434 mhz. Inter J Hyperther 1988;4(4): 447–456.
12. Luechinger R, et al. *In vivo* heating of pacemaker leads during magnetic resonance imaging. Eur Heart J 2005;26(4):376–383.
13. Kaukuntla H, et al. Temperature monitoring during cardiopulmonary bypass-do we undercool or overheat the brain? Eur J Cardiothor Surg 2004;26(3):580–585.
14. Wangcun J, Aguilar G, Wang G, Nelson S. Heat-transfer dynamics during cryogen spray cooling of substrate at different initial temperatures. Phys Med Biol 2004;7 (49).
15. Littrup P, et al. Cryotherapy for breast fibroadenomas. Radiology 2005;234(1):63–72.
16. Carnochan P, Dickinson RJ, Joiner MC. The practical use of thermocouples for temperature measurement in clinical hyperthermia. Inter J Hyperthermia 1986;1:1–19.
17. van der Zee J, et al. Practical limitations of interstitial thermometry during deep hyperthermia. Inter J Rad Oncol Biol Phys 1998;40(5):1205–1212.
18. Wren J. On Medical Thermal Treatment—Modelling, Simulation and Experiments. Ph.D. thesis dissertations. Linköpings universitet No. 763, 2002.
19. Wren J, Eriksson O, Wårdell K, Loyd D. Analysis of temperature measurement for monitoring radio-frequency brain lesioning. Med Biol Eng Comput 2001;39:255–262.
20. Wren J. Evaluation of three temperature measurement methods used during microwave thermotherapy of prostatic enlargement. Inter J Hyperther 2004;20(3).
21. Constable RT, Dunscombe P, Tsoukatos A. Perturbation of the temperature distribution in microwave irradiated tissue due to the presence of metallic thermometers. Med Phys 1987; 14(3):385–388.
22. Ryan TP, et al. Thermal conduction effects associated with temperature measurements in proximity to radio frequency electrodes and microwave antennas. Inter J Rad Oncol Biol Phys 1989;16:1557–1564.
23. Samulski TV, Lyons BE, Britt RH. Temperature measurements in high thermal gradients: Analyses of conduction effects. Inter J Rad Oncol Biol Phys 1985;11:963–971.
24. de Leeuw AAC, Crezee J, Lagendijk JJW. Temperature and sar measurements in deep-body hyperthermia with thermocouple thermometry. Inter J Hyperther 1993;9(5):685–697.
25. Dunscombe PB, McLellan J. Heat production in microwave-irradiated thermocouples. Med Phys 1986;13(4):457–461.

See also TEMPERATURE MONITORING; THERMOMETRY.

THERMODILUTION.

See CARDIAC OUTPUT, THERMODILUTION MEASUREMENT OF.

THERMOGRAPHY

HAIRONG QI
University of Tennessee
Knoxville, Tennessee
NICHOLAS A. DIAKIDES
Advanced Concepts Analysis, Inc.
Falls Church, Virginia

INTRODUCTION

The first documented application of infrared (IR) imaging in medicine was in 1956 (1), when breast cancer patients were examined for asymmetric hot spots and vascularity in IR images of the breasts. Since then, numerous research findings have been published (2–4) and the 1960s witnessed the first surge of medical application of the IR technology (5,6), with breast cancer detection as the primary practice. However, IR imaging has not been widely recognized in medicine nowadays, largely due to the premature use of the technology, the superficial understand-

ing of IR images, and its poorly controlled introduction into breast cancer detection in the 1970s (7).

Recently, advances in a couple of related areas have pushed forward series of activities to reappraise the role of IR imaging in medicine (7–12). These advances, including the development of the new-generation IR technology, smart image processing algorithms, and the pathophysiological-based understanding of IR images, will provide a cost-effective, noninvasive, nondestructive, and patient-friendly approach to health monitoring and examination, as well as to assisting diagnosis. These new developments are discussed in detail in this article.

Temperature is a long established indicator of health. The Greek physician, Hippocrates, wrote in 400 bc. "In whatever part of the body excess of heat or cold is felt, the disease is there to be discovered (13)." The ancient Greeks immersed the body in wet mud and the area that dried more quickly, indicating a warmer region, was considered the diseased tissue. The use of hands and thermometers to measure heat emanating from the body remained well into the sixteenth through the eighteenth centuries. Nowadays, we still rely on thermometers a lot when performing health examination.

All the above-mentioned methods are contact based. Since the British astronomer, Sir William Herschel, discovered the existence of IR radiation in 1800, major advances have taken place with IR imaging that do not need direct contact with the patient.

Infrared radiation occupies the region between visible and microwaves of the spectrum. All objects in the universe emit radiations in the IR region as a function of their temperature. As an object gets hotter, it gives off more intense IR radiation, and it radiates at a shorter wavelength (11). The human eye cannot detect IR rays, but they can be detected by using the IR cameras and detectors. Figure 1 illustrates the IR spectral band in finer scale. The boundaries between different IR spectral regions are not agreed upon and can vary. The boundaries that we adopt here are based on Refs. (14–18).

In general, IR radiation covers wavelengths that range from 0.75 to 1000 μm, among which the human body emissions that are traditionally measured for diagnostic purposes only occupy a narrow band at wavelengths of 8–12 μm (19). This region is also referred to as the long-wave *IR* (LWIR) or body infrared rays. Another terminology that is widely used in medical IR imaging is thermal infrared (TIR), which, as shown in Fig. 1, covers wavelengths beyond ~1.4 μm. Within this region, the IR emission is primarily heat or thermal radiation, and hence the term thermography. The image generated by TIR imaging is referred to as the thermogram. The near infrared (NIR) region occupies wavelengths between 0.75 and 1.4 μm. The IR emission that we observe in this region is not thermal (17). Although the NIR and mid-wave IR (MWIR) regions are not traditionally used in human body screening, the new generation detectors have enabled the use of multispectral imaging in medicine, in which both NIR (20) and MWIR (21) are observed in different diagnostic cases.

In this article, we discuss IR imaging in medicine across the full IR spectral region with a focus on the thermal IR region, including the pathophysiological understanding of IR imaging, the development of new generation of IR imagers, and the advanced image processing algorithms of IR images.

PATHOPHYSIOLOGICAL-BASED UNDERSTANDING OF INFRARED IMAGING

Infrared imaging is a physiological test that measures the subtle physiological changes that might be caused by many conditions, for example, contusions, fractures, burns, carcinomas, lymphomas, melanomas, prostate cancer, dermatological diseases, rheumatoid arthritis, diabetes mellitus and associated pathology, deep venous thrombosis (DVT), liver disease, bacterial infections. These conditions are commonly associated with regional vasodilation, hyperthermia, hyperperfusion, hypermetabolism, and hypervascularization (19,22–27), which generate higher temperature heat source. Unlike imaging techniques, such as X-ray radiology and Computed Tomography (CT) that primarily provide information on the anatomical structures, IR imaging provides functional information not easily measured by other methods. Thus correct use of IR images requires in-depth physiological knowledge for its effective interpretation.

Human Thermal Models

The heat emanating on to the surface from the heat source and the surrounding blood flow can be quantified using the

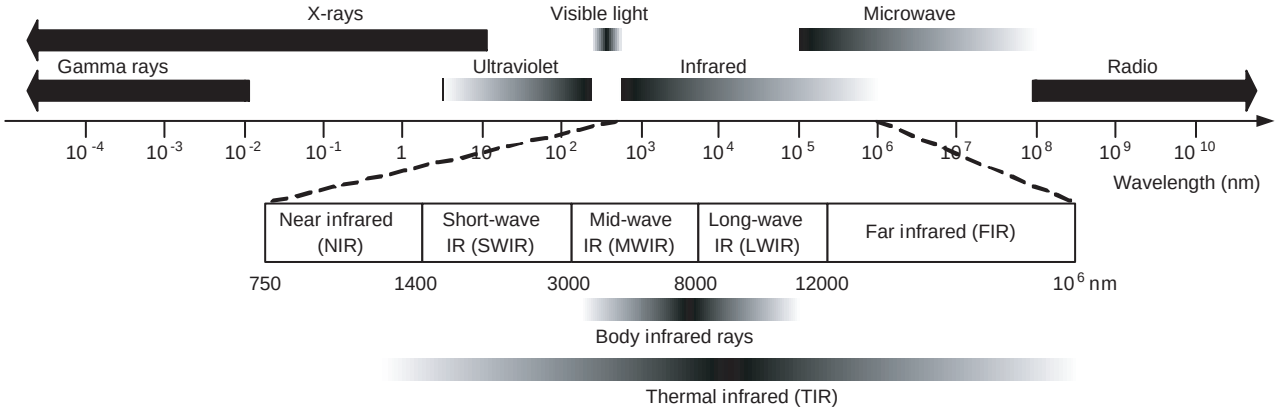


Figure 1. The electromagnetic spectrum and the IR region.

Pennes' bio-heat equation (28). This equation includes the heat transfer due to conduction through the tissue, the volumetric metabolic heat generation of the tissue, and the volumetric blood perfusion rate whose strength is considered to be the arteriovenous temperature difference (29). The equation is given as

$$k\Delta^2 T - c_b w_b (T - T_a) + q_m = 0 \quad (1)$$

where k is conductivity, q_m is volumetric metabolic rate of the tissue, $c_b w_b$ is the product of the specific heat capacity and the mass flow rate of blood per unit volume of tissue, T is the unknown tissue temperature, and T_a is the arterial temperature. In theory, given the heat emanating from the surface of the body measured by TIR imaging, by solving the inverse heat transfer problem, we can obtain the heat pattern of various internal elements of the body. Different methods of solving the bio-heat transfer equation have been presented in literature (30,31). Although it is possible to calculate the thermal radiation from a thermal body by thermodynamics, the complexity of the boundary conditions associated with the biological body makes this approach impractical.

One of the biggest hurdles in the diagnosis using to thermograms is the various thermal environmental conditions that could affect detection and evaluation to a great extent. A computer model was presented in (32) that aims to simulate the heat-transfer phenomenon within the human body and predict the internal temperature as well as the skin surface temperature, providing a reference model that thermograms taken under different thermal conditions can be converted between each other. Figure 2 illustrates the 16-cylinder-segment model this work is based on as well as the simulated body temperature profile.

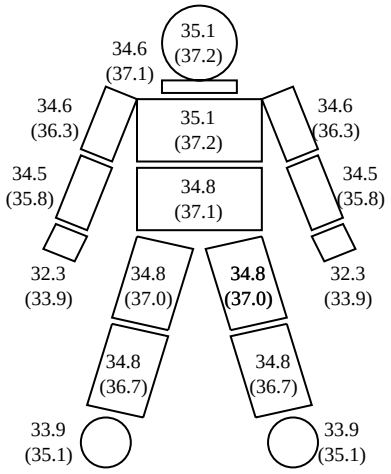


Figure 2. Simulated human body temperature profile based on the 16-cyclinder-segment model, after the first 60 min. The air temperature and mean radian temperature were maintained at 30 °C for 60 min, then changed to 24 °C and maintained at that temperature for another 60 min. The value in the parentheses is the temperature at the center of the segment. (Redrawn from Ref. 32.)

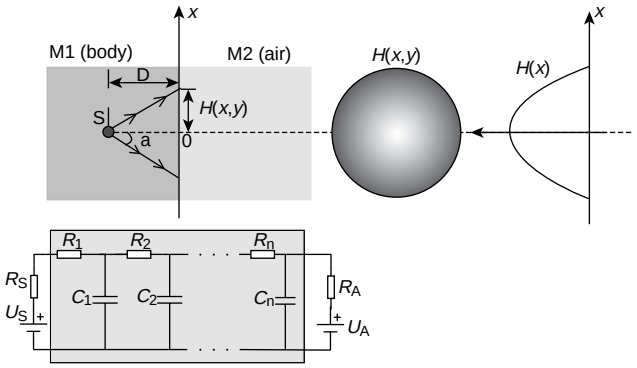


Figure 3. The thermal-electric analog.

The Thermal-Electric Analogue

Liu et al. (33,34) presented a new method for analyzing a thermal system based on an analogy to electric circuit theory; referred to as thermal-electric analog. This method does not require a direct solution to the inverse heat-transfer problem. Figure 3 illustrates the analogy between thermodynamics systems and the electrical circuit, where a battery with voltage U_S is used to simulate the heat source S inside the human body and the heat loss inside the heat source can be simulated as the heat loss on a resistor R_S. Hence, we can establish the correspondence between the temperature of the heat source and the voltage of the battery, as well as between the heat current and the circuit current. The set of R_i and C_i values correspond to the unit heat resistance and heat capacity along each radiation line. The circuit in Fig. 3 only shows the analogy for one radiation line. If the medium between the heat source (S) and the surface is homogeneous, then the radiation pattern sensed by the IR camera at the surface should have a distribution like Gaussian as shown in Fig. 3. If the medium is not homogeneous, then the surface radiation pattern can be represented as a linear combination of different Gaussain distributions.

This analogy can be used to estimate the depth of the heat source (34), and furthermore, help understand the metabolic activities undergoing within the human body, through a so-called slicing technique. The method has been used in early breast cancer detection and has achieved high sensitivity. It has also recently been used for the diagnosis of severe acute respiratory syndrome (SARS) patients, as reported in (35).

The Abnormal Thermogram Patterns

As mentioned before, one of the biggest hurdles in IR image interpretation is the lack of standardized image handling procedures. The human thermal model used in the section Human Thermal Models is one attempt in solving this problem. Fujimas also did some pioneer work (36) in 1998 by proposing eight thermophysiological expressions to identify abnormal thermogram patterns, referred to as the *thermatomes*.

Angiological thermatomes: Abnormal temperature regions caused by organic vascular abnormalities.

Functional angiological thermatomes: Abnormal temperature regions caused by vascular disfunctions.

Neurodermatomal thermatomes: Abnormal temperature bands caused by somatosensory neuronal disorders.

Myotomal thermatomes: Abnormal temperature regions suspected by abnormal muscular blood flow rate.

Metabolic thermatomes: Abnormal hot and/or cold spots caused by excessive and/or lower heat production and blood flow.

Dynamic thermatomes at environmental temperature stress: Regions with abnormal reactions when a patient received an applied thermal load.

Dynamic thermatomes at medication: Regions with abnormal reactions when a patient is given a medication.

Dynamic thermatomes at various kinds of stress: Regions with abnormal reactions when a patient receives a load (various in type).

IR Imaging in Early Breast Cancer Detection

Because IR imaging has been mainly used in breast cancer detection since its introduction to the medical field, in the following, we focus on the potential of IR imaging, especially TIR imaging, in *early* breast cancer detection.

Cancer cells are resulted from permanent genetic change in a normal cell triggered by some external physical agents such as chemical agents, X rays, ultraviolet (UV) rays, and so on. All types of cancer cells have an imbalanced metabolic activity that leads to the utilization of a large amount of blood glucose and the release of large amounts of lactate into blood. In addition, the high metabolic rate of cancer cells causes an increase in local temperature as compared to normal cells. These factors have enabled IR imaging as a viable technique to visualize the abnormality. The IR image provides more dynamic information of the tumor since the tumor can be small in size but can be fast growing making it appear as a high temperature spot in the IR image (37,38).

Many imaging modalities can be used for breast screening, including mammography using X-ray, IR, magnetic resonance imaging (MRI), CT, ultrasound, and positron emission tomography (PET) scans. Although mammography has been the baseline approach, it depends primarily on structural distinction and anatomical variation of the tumor from the surrounding breast tissue (7). Unless the tumor is beyond certain size, it cannot be imaged as X rays essentially pass through it unaffected. Other modalities like MRI and PET scan could provide valuable information to diagnosis, but they are not popularly adopted for various reasons including high cost, complexity and accessibility issues (12). Compared to mammography, MRI, CT, ultrasound, and PET scans that are also called the after-the-fact (a cancerous tumor is already there) detection technologies, IR imaging is able to detect breast cancers 8–10 years earlier than mammography (39,40). Keyserlingk reported in (7) that the average tumor size undetected by IR imaging is 1.28 cm versus 1.66 cm by mammography.

Samples of Other Advanced Interpretations

We would also like to mention two interesting work conducted recently although their influence on diagnosis is yet to be investigated. Alexjander and Deamer (41) propose to study the sound (rhythms and frequencies) made within the human body through the access of the IR frequencies of DNA bases. Imagine if we can "hear" the body, would a pleasing pattern to the ear indicate a healthy subject? Or would different patterns present a sign of a certain disease? Through nonlinear heat transfer modeling, Pavlidis and Levine (42) show that the periorbital blood flow in anxious states can be used to extract subtle facial temperature fluctuation patterns and thus assist in traditional polygraph examination. Perhaps if we go beyond imagination, more exciting applications of IR imaging can come into the light.

NEW GENERATION INFRARED TECHNOLOGIES

Infrared technology owes its origin to military research. Since IR imaging was first introduced to medical diagnosis in the 1960s, most of the IR equipment used has not been specifically designed for the medical market (23). Some of the problems associated with IR cameras at that time, for example, narrow field of view ($<20^{\circ}$) and low spatial resolution ($\sim$200 optical lines), although are not issues in military applications, they have affected the effectiveness and accuracy of diagnosis to a great extent. Some recent advances in IR sensor design expect to solve these problems and make IR sensors adequate for medical applications. In the following, we focus our discussion on the advances in the detector technologies, especially the uncooled camera development.

Cooled versus Uncooled Thermal Detectors

To some extent, the main factor that determines which wavelengths are included in which IR region is the type of detector technology used to capture IR radiation (17). The NIR radiations are observed in very similar way as the visible light, except that special IR detectors need to be used. On the other hand, TIR imaging generally requires the use of a cooling system in the form of a nitrogen or compressed air cooling bottle, which contains crystals like germanium whose electrical resistance is very sensitive to heat. Figure 4 shows the two main assemblies of the EYE-Z640 cooled FLIR (forward looking IR) camera from OPGAL Optronic Industries Ltd. (43). The detector is InSb cryogenically cooled and needs extra gadgets like the cooler and the dewar to support the cooling system. Compared to uncooled IR cameras, although cooled systems generally present better sensitivity, they consume more power, need a relatively longer cooling down time (e.g., a few minutes), and are more expensive. In addition, the average time that cooled IR cameras will function before failing is very limited (around a few thousand hours). Due to the size, weight, and complexity, these systems were limited to fixed deployment like tripod mounting.

The advance in solid state models has made a new class of sensors possible, the uncooled detector design. In the

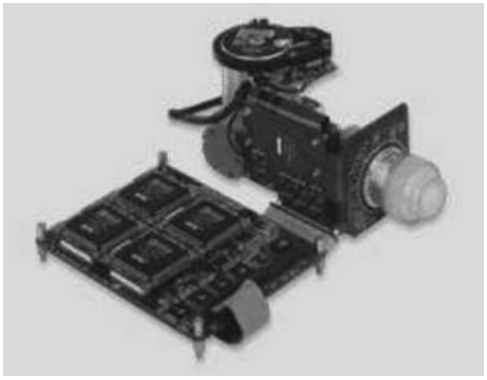


Figure 4. EYE-Z640 InSb cryogenically cooled FLIR camera from OPGAL (44) with the electronic card (lower left corner) and the DDCE (upper right corner: Detector, Dewar, Cooler, and Electronics).

1980s, the Department of Defense (DoD) sponsored companies like Honeywell and Texas Instruments (TI) with large classified contracts to develop uncooled IR detector technology (45). Honeywell's microbolometer and TIs pyroelectric sensors are both successful deliverables from these programs. In 1999, the Defense Advanced Research Projects Agency (DARPA) issued a Broad Agency Announcement (BAA) (46) that solicits proposals for increasing the performance of the uncooled IR sensors to their theoretical limit. The objective for the thermal sensitivity is set at $<$10 mk with the pixel size less than or equal to 25 μm. As far as the array size, high performance arrays for long-range systems can be as large as 960×1280 elements, while arrays for microsensors may be as small as 240×320 elements (46).

Compared to other uncooled IR detector technologies, like ferroelectric and pyroelectric, microbolometer sensors are less expensive, providing higher dynamic range, broader spectral range, and lower cross-talk. Therefore, this type of uncooled sensors are more popularly used, especially after DoD declassified the microbolometer technology in 1992.

The microbolometer technology, which is thermalelectric in nature, converts IR energy to a change in resistance. Each microbolometer detector consists of a silicon nitride microbridge that lies above a silicon substrate and is supported by silicon nitride legs, as shown in Fig. 5. A bolometer is a thermal detector that is deposited on the bridge. When heated by incoming radiation, the bolometer detector can result in a temperature rise that is sensed as a change in the element resistance.

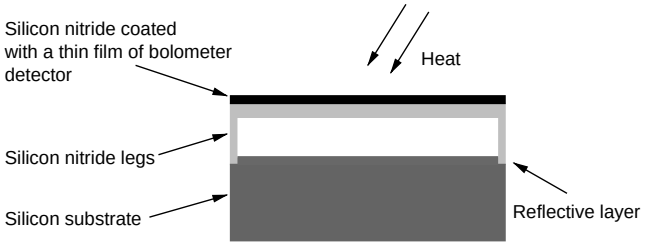


Figure 5. The microbolometer detector layout.

Figure 6. Bioyear's PRISM 2000 thermal metabolic imaging system (Left: Gantry that hosts the FLIR camera. Middle: Console. Right: Bed). (Courtesy of Bioyear, Inc.)

Because the uncooled cameras do not require a cooling system, they are much lighter, smaller, more reliable and less expensive compared to the cooled cameras. Currently, the uncooled cameras are approaching to the thermal sensitivity of the cooled ones (0.05° or 0.02 °C of uncool vs. 0.01 °C of cool) and are very popular in breast imaging. Figure 6 shows the PRISM 2000 Thermal Metabolic Imaging system manufactured by Bioyear, Inc. (47), in which a FLIR microbolometer detector is installed on the gantry. The detector consists of 320×240 pixels with a sensitivity of 0.05 °C.

Although we have listed quite a few advantages of cooled IR cameras, the competition between uncooled and cooled detectors continues, especially with the recent development of the deeply cooled QWIP (quantum well photodetector) detectors. Wiecek conducted a brief comparison (48) between uncooled thermal detectors and QWIP and discussed the limits in both technologies.

Three Generations of Development

Since its first appearance, the development of thermal imagers has gone through three generations of evolution.

The first generation thermal imagers were fielded in the 1970s. They use a single detector or small-size linear array detectors. In order to generate the picture, two scanning mirrors are used. This generation imagers generally have the white out problem (or over saturation over high intensity sources). Although mechanical brightness controls are used later to address the problem, the images still lack clarity.

Second generation imagers appeared in the 1980s. They use a relatively larger linear array ($\sim$120 elements) or small two-dimensional (2D) focal plane array (FPA) ($\sim 64 \times 64$ elements) and the scanning mirrors are still used to generate the picture.

Third generation imagers upgrade the size of the 2D FPA a great deal, some of which contain as many elements as 1024×1024. In addition, the image processing capabilities are integrated on the FPA, hence the so-called on-chip image processing. According to Xenics' definition (49), FPA is a matrix of detector cells that attached to a semiconductor chip. Each cell is responsive in IR wavelengths, in which it absorbs IR radiation, converts it into electrons, and sends a voltage signal in response to form an image. The FPA can capture multicolor images and brings great advantages to image capturing, including emissivity correction, lower

atmosphere influence on the temperature measurement, and so on. The third generation does not use mirrors that largely improves the image quality as the less moving part in the camera, the more reliable the system, and the less mechanical noise. Currently, the third generation FPA detectors can capture wavelength from 3 to 5 or 8–12 μm.

One of the most important features that distinguishes the third generation design is the employment of the time-delay integration (TDI) technique for image integration and enhancement. The TDI is a specialized detector readout mode. Instead of reading out the entire chip as a single large image, the image is read out continuously, line by line from the bottom of the detector chip. If the readout rate of the detector and the velocity of the object being imaged matches each other, then there will not be motion blurs.

Smart Image Processing Approaches to IR Images

Computer-aided diagnosis (CAD) has been playing an important role in the analysis of IR images, as human examination of images is often influenced by various factors like fatigue, being careless, and so on. The detection accuracy is also confined by the limitations of human visual system. On top of all these factors, a shortage of qualified radiologists also put an urgent demand on the development of CAD technologies. Currently, research on smart image processing algorithms on IR images tends to improve the detection accuracy from three perspectives: smart image enhancement and restoration algorithms, asymmetry analysis of the thermogram including automatic segmentation approaches, and feature extraction and classification.

Smart Image Enhancement and Restoration Algorithms

One of the problems with thermograms that has put IR imaging in a somewhat disadvantage situation is its lack of resolution due to blur compounded by rather high levels of noise. Snyder et al. (50) developed an algorithm to increase the effective resolution of thermograms by a 2:1 ratio while at the same time removing the noise and preserving edges in the image. This algorithm is based on a minimization strategy known as mean-field annealing, which takes into account processes of blur, noise, and image correlations, to make an optimal estimate of the missing pixels.

MITs researchers attempt to enhance the resolution of IR images through another route. The Minimally Invasive Optical Biopsy System developed at MIT (51) uses IR light in conjunction with an intravenously injected dye and special computer software to create a clear, high contrast image that could easily allow physicians to detect breast masses and determine if they are benign or malignant.

Kaczmarek and Nowakowski (52) proposed the use of active dynamic thermography (ADT), commonly adopted in nondestructive testing of materials, to further enhance the image quality. ADT analyzes thermal transients after the application of external thermal excitation. Some preliminary results have shown the promise of this approach.

Asymmetry Analysis

Making comparisons between contralateral images are routinely done by radiologists. When the images are relatively symmetrical, small asymmetries may indicate a suspicious region. From the human thermal model depicted in Fig. 2, the symmetry in temperature distribution between the left and the right part of the human body is very obvious. However, these small asymmetries might not be easy to detect and it is important to design an automatic approach to eliminate human factors. Figure 7 shows four case studies on how asymmetric thermal signatures have indicated different types of disease. These studies are originally conducted by Meditherm (53). Figure 7a shows a patient with complex regional pain syndrome (CRPS) in the right foot, which is developed after a fractured calcaneum 18 months previously. The thermogram indicates the right foot is 3.7 °C colder than the left foot. It is reported from the same study that some cases of CRPS are misdiagnosed as psychological or hysterical pain states but thermography is able to show characteristic changes. Figure 7b shows a patient with right knee surgery followed with a painful effusion in the early post operative period. In this case, thermography is able to confirm a significant inflammatory reaction and 30 ml of blood-stained fluid was aspirated. Figure 7c shows a patient with the left wrist injured three years ago. Thermogram indicates obvious temperature change in the left wrist and hand.

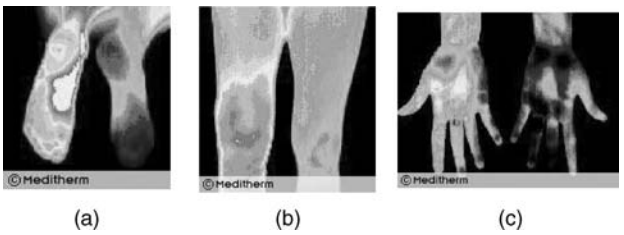


Figure 7. Case studies from Meditherm (53).

Head et al. (54,55) recently analyzed the asymmetric abnormalities in IR images. In their approach, the image is segmented first by operator. Then breast quadrants are derived automatically based on unique point of reference, that is, the chin, the lowest, rightmost and leftmost points of the breast.

Qi and Head (56) developed an automatic approach to asymmetry analysis in IR images. It includes automatic segmentation and pattern classification. Hough transform is used to extract the four feature curves that can uniquely segment the left and right breasts. The feature curves include the left and the right body boundary curves, and the two parabolic curves indicating the lower boundaries of the breasts. Figure 8 shows the segmentation results of two patient images obtained using the Inframetrics 600M camera, with a thermal sensitivity of 0.05 K. The images are collected at Elliott Mastology Center. The results include the intermediate images from edge detection, feature curve extraction, to segmentation. From the figure, we can see that Hough transform can derive the parabola at the accurate location.

Mabuchi et al. (57) designed a computerized thermographic system, which would produce images of the distribution of temperature differences between the affected side and the contralateral healthy side. Because there is no standard skin surface temperature existed, the system

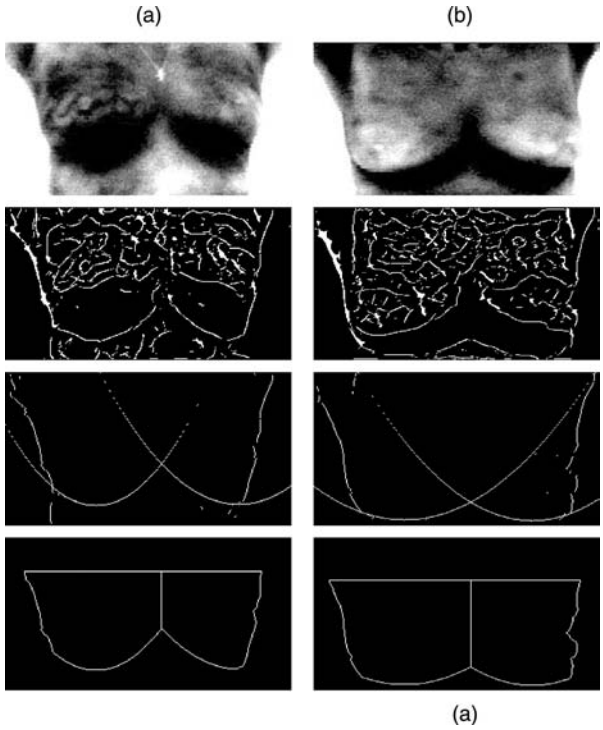


Figure 8. Segmentation results of two images. (a) results from cancerous patient. (b) Results from normal patient. From top to bottom: original image, edge image, four feature curves, segments.

measures the body-surface temperature of each pixel in the affected area and subtract from it the body-surface temperature of the corresponding pixel in the symmetrically located contralateral healthy area to generate the difference image.

Feature Extraction and Classification

Upon segmentation, different features can be extracted from the segments. Asymmetric abnormalities can then be identified based on mature pattern classification techniques. In this process, feature extraction is crucial to the success of computer-aided diagnosis (58) shows that the high order statistics (e.g., variance, skewness, and kurtosis) and joint entropy are the most effective features to measure the asymmetry, while low order statistics (e.g., mean) and entropy do not assist asymmetry detection. Jakubowska et al. (59) also addressed the importance of using statistical parameters (first and second order) in extracting thermal signatures for asymmetry analysis. From the figure, we observe that the high order statistics are the most effective features to measure the asymmetry, while low order statistics (mean) and entropy do not assist asymmetry detection. Figure 9 compares the effectiveness of the features used to analyze the symmetry of the segments derived in Fig. 8. The first data point along the x-axis indicates entropy, the second to the fifth points indicate the four statistical moments (means, variance, skewness, and kurtosis). The y axis shows the closeness metric we defined as the absolute difference between 1 and the ratio between the feature value from the left segment and that from the right segment. Hence, the smaller the difference, the more

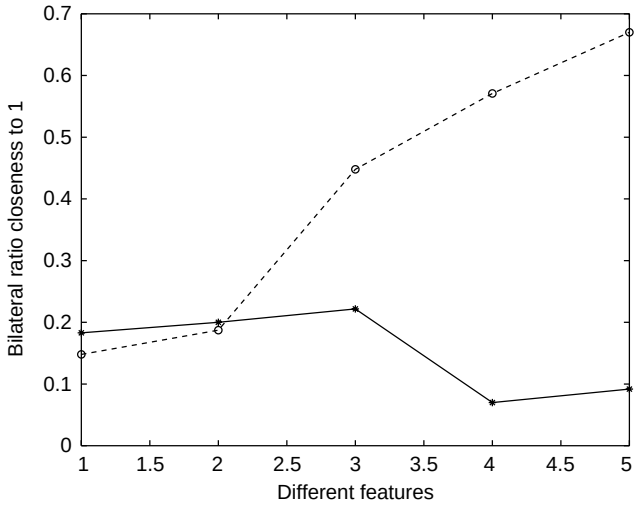


Figure 9. Performance evaluation of different feature elements. Solid line: noncancerous image; Dash line: cancerous image. The five data points along the x axis indicate (from left to right): entropy, mean, variance, skewness, kurtosis.

symmetric the two segments. We observe from the figure that the high order statistics are more effective features to measure the asymmetry than low order statistics (mean) or entropy.

Szu et al. (60) proposed a new paradigm shift that uses at least two dual-band (mid and long) IR imaging cameras operating simultaneously on the patient. This system enables a smart brain-like neural network algorithm, the Lagrange Constraint Neural Network (LCNN), to achieve submilimeter scaling of the close-up breast imaging for the vascular and angiogenesis effects as well as stage-zero detection of ductal carcinoma *in situ*.

The Dynamic Area Telethermometry Technique

To be able to yield objective clinical diagnosis, Anbar et al. (61) proposed the dynamic area telethermometry (DAT) technique. It has been demonstrated to be applicable to any quantitative patho-physiological assessment. The authors demonstrated that using classical fast Fourier transform (FFT) and elementary statistics, the large amount of sequential observations can be reduced to a single quantitative diagnostic parameter without the participation of human experts. Other related work also reported in (62,63).

The above mentioned techniques are just samples of activities reported in recent conferences, workshops, and symposia. Another trend of effort that is worth mentioning is the transition of automatic target recognition (ATR) algorithms developed for military application to medicine."From tanks to tumors" (64,65) has been the theme of this transition and the rich collection of ATR algorithms that the military has sponsored will greatly improve the state-of-the-art of CAD development.

Concept Validation

Concept validation is an important procedure in the promotion of IR-based breast screening where blind diagnosis

and clinical evidence are necessary. Although there have been a lot of clinical trials conducted so far (66), there has not been a well-designed, standard database created for the purpose of concept validation.

The Advanced Concept Analysis, Inc. located at Falls Church, VA was awarded in 2000 to manage the creation of such a database. The project is sponsored by the Deputy Assistant Secretary of the Army for Installations and Environment (Environmental Safety and Occupational Health), the Office of the Deputy Undersecretary of Defense for Science and Technology (ODUSD/S&T), the Air Force Research Laboratory (AFRL), and the Office of Naval Research (ONR). An Internet-based Virtual Distributed Laboratory (VDL) at AFRL will house >8000 images from >2000 patients provided by E.H.H. Breast Cancer Research and Treatment Center, Baton Rouge, LA and Ville Marie Medical Center & Women's Health Center, Montreal, Canada. Each center will use the collaboration tools and evaluation procedures on the VDL to conduct blind diagnoses of the images provided by the other. Blind test results will be compared with actual clinical evidence stored with the imagery. VDL access may be applied for at (67).

INTERNATIONAL IR IMAGING ACTIVITIES IN MEDICINE

Before this article is summarized lets take a quick scan of activities reported worldwide on the usage of IR imaging in medicine.

United States of America and Canada: Infrared imaging is beginning to be reconsidered in the Unites States, largely due to the factors discussed early, that is new IR technology, advanced image processing, in-depth pathophysiological understanding of IR images. Currently, there are several academic institutions with research initiatives in IR imaging. Some of the most prominent include National Institute of Health (NIH), Johns Hopkins University (JHU), University of Houston, and University of Texas. The NIH has several ongoing programs, such as vascular disorders (diabetes, DVT), monitoring angiogenesis activity (Kaposi Sarcoma, pain reflex sympathetic dystrophy (68), monitoring the efficacy of radiation therapy, organ transplant) perfusion, multispectral imaging, and so on (69–71). The JHU does research in microcirculation, monitoring angiogenic activity in Kaposi Sarcoma and breast screening, laparoscopic IR images for renal disease. University of Houston just created an IR imaging laboratory to investigate with IR the facial thermal characteristics for such applications as lie detection and other behavioral issues (fatigue, anxiety, fear, etc.) (42). There are two medical centers specializing in breast cancer research and treatment which use infrared routinely as part of their first line detection system, which also includes mammography and clinical exam. These are: EHH Breast Cancer and Treatment Center, Baton Rouge, LA. and Ville Marie Oncology Research Center, Montreal, Canada

China: China has a long-standing interest in IR imaging. More recently, the novel method Thermal Texture Mapping (TTM) (34,35) has added increased specificity to static imaging. It is known that this method is widely used in this country. Introduction of TTM has been made to NIH. They have been using this method successfully in detection and treatment in Kaposi Sarcoma (associated with AIDS patients). There are further possibilities for high-level research for this method in the United States and abroad. The TTM technology developed by Bioyear (47) has also been adopted in the EHH Center and the Ville Marie Center mentioned above.

Japan: Infrared imaging is widely accepted in Japan by the government and the medical community. More than 1500 hospitals and clinics use IR imaging routinely (70). The government sets the standards and reimburses clinical tests. Their focus is in the following areas: blood perfusion, breast cancer, dermatology, pain, neurology, surgery (open-heart, orthopedic, dental, cosmetic), sport medicine, oriental medicine. The main research is performed at the following universities: University of Tokyo—organ transplant; Tokyo Medical and Dental University (skin temperature characterization and thermal properties); Toho University—neurological operation); Cancer Institute Hospital (breast cancer). In addition, around forty other medical institutions are using infrared for breast cancer screening.

Korea: Began involvement in IR imaging during the early 1990s. More than 450 systems have been used in hospitals and medical centers. Primary clinical applications are neurology, back pain/treatment, surgery, oriental medicine. Yonsei College of Medicine is one of the leading institutions in medical IR imaging research along with three others.

United Kingdom: The University of Glamorgan is the center of IR imaging; the School of Computing has a thermal physiology laboratory which focuses in the following areas: medical IR research, standardization, training (university diploma),"SPORTI" Project funded by the European Union Organization. The objective of this effort is to develop a reference database of normal, thermal signatures from healthy subjects. The Royal National Hospital of Rheumatic Diseases specializes in rheumatic disorders, occupational health (Raynaud's Disease, Carpal Tunnel Syndrome and Sports Medicine). The Royal Free University College Medical School Hospital specializes in vascular disorders (diabetes, DVT, etc.), optimization of IR imaging techniques and Raynaud's Phenomenon).

Germany: University of Leipzig uses IR for open-heart surgery, perfusion, and microcirculation. There are several private clinics and other hospitals that use infrared imaging in various applications. EvoBus-Daimler Chrysler uses IR imaging for screening all their employees for wellness/health assessment (occupational health). InfraMedic, AG, conducts breast cancer screening of women from 20–85 years old for the government under a 2 year grant and IR is the sole modality used.

Austria: Ludwig Boltzmann Research Institute for Physical Diagnostics has done research in IR for many years and it publishes the Thermology International (a quarterly journal of IR clinical research and instrumentation). The General Hospital, University of Vienna, does research mainly in angiology (study of blood and lymph vessels) diabetic foot (pedobarography).

Poland: There has been a more recent rapid increase in the use of IR imaging for medicine in Poland since the

Polish market for IR cameras was opened up. There are >50 cameras being used in the following medical centers: Warsaw University, Technical University of Gdansk, Poznan University, Lodz University, Katowice University and the Military Clinical Hospital. The research activities are focused on the following areas: Active Infrared Imaging, open-heart surgery, quantitative assessment of skin burns, ophthalmology, dentistry, allergic diseases, neurological disorders, plastic surgery, thermal image database for healthy and pathological cases and multispectral imaging (IR, visual, X-ray, ultrasound) (52).

Italy: Much of the clinical use of IR imaging is done under the public health system, besides private clinics. The ongoing clinical work is in the following areas: dermatology (melanoma), neurology, rheumatology, anaesthesiology, reproductive medicine, and sports medicine (72). The University of G. d'Annunzio, Chieti, has an imaging laboratory purely for research on infrared applications.

SUMMARY

This article discussed recent research achievements in medical theomography. The objective is to show that due to the advances in IR technology, image processing techniques, and the pathophysiological-based understanding of theromograms, IR imaging is mature to be used as a first line supplement to both health monitoring and clinical diagnosis. We have established a website (73) to facilitate researchers working in the field of medical thermography to exchange research findings. We welcome contributions to enrich this list of collections.

BIBLIOGRAPHY

1. Lawson RN. Implications of surface temperature in the diagnosis of breast cancer. Can Med Assoc J 1956;75:309–310.
2. Handley RS. The temperature of breast tumors as a possible guide to prognosis. Acta Unio Int Contra Cancrum 1962; 18:822.
3. Lawson RN, Chughtai MS. Breast cancer and body temperatures. Can Med Assoc J. 1963;88:68–70.
4. Lloyd-Williams K, Handley RS. Infrared thermometry in the diagnosis of breast disease. Lancet 1961; (2):1378–1381.
5. Gershen-Cohen J, Haberman J, Brueschke EE. Medical thermography: a summary of current status. Radiol Clin N Am 1965;3:403–431.
6. Haberman J. The present status of mammary thermography. Ca - A Cancer J Clin 1968;18:314–321.
7. Keyserlingk JR, et al. Functional infrared imaging of the breast. IEEE Eng Med Bio Mag May–June 2000; pp. 30–41.
8. Anbar M. Quantitative and dynamic telethermometry—a fresh look at clinical thermology. IEEE Eng Med Bio Mag Jan.–Feb. 1995;14(1):15–16.
9. Head JF, Elliott RL. Infrared imaging: making progress in fulfilling its medical promise. IEEE Eng Med Bio Mag Nov.–Dec. 2002;21(6):80–85.
10. Head JF, Wang F, Lipari CA, Elliott RL. The important role of infrared imaging in breast cancer. IEEE Eng Med Bio May/June 2000; 52–57.
11. Jones BF. A reappraisal of the use of infrared thermal image analysis in medicine. IEEE Trans Med Imaging December 1998;17(6):1019–1027.
12. Keyserlingk J. Time to reassess the value of infrared breast imaging? Oncology News Int 1997;6(9).
13. Thermology. Available at http://www.thermology.com/history.htm.
14. Electromagnetic spectrum. Available at http://www.lbl.gov/MicroWorlds/ALSTool/EMSpec/EMSpec2.htm. Last updated August 31, 2001.
15. HyperPhysics. Available at http://hyperphysics.phy-astr.gsu.edu/ hbase/ems1.htmlc1.
16. Infrared. Available at http://en.wikipedia.org/wiki/Infrared.
17. Near, mid and far-infrared. Available at http://www.ipac.caltech.edu/Outreach/Edu/Regions/irregions.html.
18. Thermal imaging. Available at http://www.ibd.nrc-cnrc.gc.ca/english/specethermal.htm.
19. Whale J. Radiometric thermal diagnostics and dielectric resonance management procedures. http://www.positive-health.com.
20. Mansfield JR. Tissue viability by multispectral near infrared imaging: a fuzzy C-means clustering analysis. IEEE Trans Med Imaging December 1998;17(6):1011–1018.
21. Office of Naval Research Press Release. Detecting breast cancer with a new algorithm and a multi-spectral infrared imaging system. Available at http://www.onr.navy.mil/media/article.asp?ID=14. September 2002.
22. Anbar M. Clinical thermal imaging today. IEEE Eng Med Bio Mag July–Aug. 1998;17(4):25–33.
23. Bale M. High-resolution infrared technology for soft-tissue injury detection. IEEE Eng Med Bio Mag July–Aug. 1998; 17(4):56–59.
24. Harding JR. Investigating deep venous thrombosis with infrared imaging. IEEE Eng Med Bio Mag July–Aug. 1998;17(4): 43–46.
25. Jones BF, Plassmann P. Digital infrared thermal imaging of human skin. IEEE Eng Med Bio Mag Nov.–Dec. 2002;21(6): 41–48.
26. Ring EFJ. Progress in the measurement of human body temperature. IEEE Eng Med Bio Mag July–Aug. 1998;17(4): 19–24.
27. Szabo T, et al. Cardiothermographic assessment of arterial and venous revascularization. IEEE Eng Med Bio Mag May–June 2000;19(3):77–82.
28. Pennes HH. Analysis of tissue and arterial blood temperature in resting human forearm. J Appl Physiol 1948;2:93–122.
29. Ng EYK, Sudarshan NM. Numerical computation as a tool to aid thermographic interpretation. J Med Eng Technol March/April 2001;25(2):53–60.
30. Chan CL. Boundary element method analysis for the bioheat transfer equation. ASME J Heat Transfer, 1992;114:358–365.
31. Hsu TR, Sun NS, Chen GG. Finite element formulation for two dimensional inverse heat conduction analysis. ASME J Heat Transfer 1992;114:553–557.
32. Kakuta N, Yokoyama S, Mabuchi K. Human thermal models for evaluating infrared images. IEEE Eng Med Bio Mag Nov.–Dec. 2002;21(6):65–72.
33. Liu Z, Wang C. Method and apparatus for thermal radiation imaging. U.S. Pat. 6,023,637 (2000).
34. Qi H, Kuruganti PT, Liu Z. Early detection of breast cancer using thermal texture maps. IEEE International Symposium on Biomedical Imaging: Macro to Nano, Washington (DC): 2002. p 309–312.
35. Liu Z, et al. Thermal texture maps (ttm): Concept, theory, and applications. In: Di-akides NA, editor. Biomedical Engineering Handbook. Vol. Infrared Imaging Section. New York: CRC Press; 2005.
36. Fujimasa I. Pathophysiological expression and analysis of far infrared thermal images. IEEE Eng Med Bio Mag July–Aug. 1998;17(4):34–42.

37. Hay GA. Medical Image: Formation, Perception and Measurement. New York: America, The American Institute of Physics and Wiley; 1976.
38. Watmough DJ. The role of thermographic imaging in breast screening. discussion by C R Hill. Medical Images: Formation, perception and measurement 7th L H Gray Conference: Medical Images. 1976; pp. 142–158.
39. Gautherie M. Atlas of breast thermography with specific guidelines for examination and interpretation. Milan. Italy: PAPUSA; 1989.
40. Ng EYK, et al. Statistical analysis of healthy and malignant breast thermography. J Med Eng Technol November/December 2001;25(6):253–263.
41. Alexjander S, Deamer D. The infrared frequencies of DNA bases: Science and art. IEEE Eng Med Bio Mag March–April 1999;18(2):74–79.
42. Pavlidis I, Levine J. Thermal image analysis for polygraph testing. IEEE Eng Med Bio Mag Nov.–Dec. 2002;21(6):56–64.
43. Opgal Optronic Industries Ltd. FLIR cameras. Available at http://www.opgal.com/teflir.htm. 2001.
44. Opgal Optronic Industries Ltd. EYE Z640 cooled FLIR. Available at http://www.opgal.com/z640.htm, 2001.
45. Infrared Solutions. Historical facts of the microbolometer techonology. Available at http://www.infraredsolutions.com/html/technology/historicalFactsF.shtm. 2005.
46. DARPA MTO Program. Baa99-30: Uncooled thermal imaging sensors. Commerce Bus Daily; July 12 1999.
47. Bioyear, Inc. Prism 2000 thermal metabolic imaging system. Available at http://www.bioyear.net/English/prism.htm. 2000.
48. Wiecek B. Advances in infrared technology—quantum well versus thermal detectors. Proc Second Joint EMBS/BMES Conf. Vol. 2. 2002; p 1135–1136.
49. Infrared glossary of terms—xenics. Available at http://www.xenics.com/Products/Glossary.php. 2004.
50. Snyder WE, et al. Increasing the effective resolution of thermal infrared images. IEEE Eng Med Bio Mag May–June 2000;19(3):63–70.
51. Braunstein M, Chan RW, Levine RY. Simulation of dye-enhanced near-ir transillumination imaging of tumors. Proc 19th EMBS Annu Int Conf, Vol. 2, Chicago: 1997; p 735–739.
52. Kaczmarek M, Nowakowski A. Analysis of transient thermal processes for improved visualization of breast cancer using ir imaging. Proc 25th Annu Int Conf IEEE EMBS. Vol. 2. 2003; p 1113–1116.
53. Meditherm—digital infrared thermal imaging. Available at http://www.meditherm.com/thermpage9.htm, Last Updated: September 3, 2003.
54. Head JF, Lipari CA, Elliott RL. Computerized image analysis of digitized infrared images of the breasts from a scanning infrared imaging system. Proc 1998 Conf Infrared Tech Appl XXIV, Part I, Vol. 3436. San Diego SPIE; 1998. p 290–294.
55. Lipari CA, Head JF. Advanced infrared image processing for breast cancer risk assessment. Proc 19th Int Conf IEEE/EMBS, Chicago: IEEE: Oct. 30–Nov. 2 1997. p 673–676.
56. Qi H, Head J. Asymmetry analysis using automatic segmentation and classification for breast cancer detection in thermograms. Proc 23rd Annu Int Conf IEEE EMBS. Vol. 3, Turkey: IEEE: Octover 2001. p 2866–2869.
57. Mabuchi K, et al. Evaluating asymmetrical thermal distributions through image processing. IEEE Eng Med Bio Mag March–April 1998;17(2):47–55.
58. Kuruganti PT, Qi H. Asymmetry analysis in breast cancer detection using thermal infrared images. In Proc 2nd Joint EMBS-BMES Conf. Vol. 2. October 2002; p 1129–1130.
59. Jakubowska T, Wiecek B, Wysocki M, Drews-Peszynski C. Thermal signatures for breast cancer screening comparative study. Proc 25th Annu Int Conf IEEE EMBS. Vol. 2. 2003; 1117–1120.
60. Szu H, et al. Lagrange constraint neural net de-mixing enabled multispectral breast imaging. IEEE EMBS 2002.
61. Anbar M, et al. Detection of cancerous breasts by dynamic area telethermometry. IEEE Eng Med Bio Mag Sept.–Oct. 2001;20(5):80–91.
62. Fujimasa I, Chinzei T, and Saito I. Converting far infrared image information to other physiological data. IEEE Eng Med Bio Mag May–June 2000;19(3):71–76.
63. Gulyaev V Yu, Markov GA, Koreneva GL, Zakharov PV. Dynamical infrared thermography in humans. IEEE Eng Med Bio Mag Nov.–Dec. 1995;14(6):766–771.
64. Irvine JM. Targeting breast cancer detection with military technology. IEEE Eng Med Bio Mag Nov.–Dec. 2002; 21(6):36–40.
65. Paul JL, Lupo JC. From tanks to tumors. IEEE Eng Med Bio Mag Nov.–Dec. 2002;21(6):34–35.
66. Amalu WC. A review of breast thermography. Available at http://www.iact-org.org/articles/ articles-review-btherm.html.
67. AFRL. Medatr database using vdl. Available at http://www.vdl. afrl.af.mil/access/. 2003.
68. Hassan M, et al. Infrared thermographic imaging for the assessment of temperature asymmetries in reflex sympathetic dystrophy. Proc 25th Annu Int Conf IEEE Eng Med Biol Soc, Vol. 2, 2003; p 1102–1105.
69. Hassan M, et al. Non-invasive multi-modality technique to study angiogenesis associated with Kaposi's sarcoma. Proc Second Joint EMBS/BMES Conf. Vol. 2. 2002. p 1139–1140.
70. Harrison B, Mabuchi K. Biomedical infrared imaging in Japan. IEEE Eng Med Bio Mag July–Aug 1998;17(4):66–70.
71. Hassan M, et al. Noninvasive infrared imaging for quantitative assessment of tumor vasculature and response to therapy. Proc 26th Annu Int Conf IEEE Eng Med Biol Soc, Vol 2. 2004; p 1200–1202.
72. Merla A, et al. Infrared functional imaging applied to Raynaud's phenomenon. IEEE Eng Med Bio Mag Nov.–Dec. 2002;21(6):73–79.
73. Qi H. Thermal infrared imaging in early detection of breast cancer—a survey of recent research. Available at http://aicip.ece.utk.edu/research/irsurvey.htm., 2003.

See also MAMMOGRAPHY; THERMOMETRY; X-RAYS, INTERACTION WITH MATTER.

THERMOMETRY

THEODOROS SAMARAS
Aristotle University of Thessaloniki Department of Physics
Thessaloniki, Greece

INTRODUCTION

High body temperature has been an indication of illness since the time of Hippocrates, when only the hand could be used to detect heat or cold. The development of thermometers as we know them was a slow process; it was not before the late 19th century that thermometry became an important tool in clinical practice (1). Medical thermometry is routinely used nowadays across the spectrum of medical specialties and in all environments, ranging from the home to the critical care unit. New techniques have

appeared, which allow the measurement not only of the core body temperature, but also of local temperatures or regional temperature distributions, which can be used as health indicators.

Catheterization for thermometry measurements is used in the thermodilution technique for the assessment of the cardiac output and in intracoronary thermography for the prognosis of atherosclerotic plaques. Non-contact infrared thermometry has been used in the screening of travelers for the severe acute respiratory syndrome (SARS), but also in rheumatic diseases, vascular disorders, and the detection of breast cancer. The most extensive use of thermometry, however, is in thermal therapies (hyperthermia and tissue ablation). The physical method used (RF and microwave radiation, lasers or ultrasound), the treatment site, the target volume size, and the anticipated temperature rise determine the temperature measurement technique. In fact, several practical requirements in this area have resulted in the recent technological advancement of medical thermometry.

SPECIFICATIONS OF A TEMPERATURE MEASURING SYSTEM

To achieve a fair comparison among the various temperature measuring systems and to choose the appropriate one for each application, a set of parameters has to be considered. In the following article, these characteristic parameters are defined as closely as possible to the international vocabulary of basic and general terms in metrology (2).

Accuracy is the ability of a measuring system to provide a value of temperature close to its true value. The *resolution* of a temperature measuring system is the smallest change in the value of temperature that causes a perceptible change in the corresponding indication, whereas the *resolution of a displaying device* is the smallest difference between indications of a displaying device that can be meaningfully distinguished. *Stability* is the ability of a temperature probe to maintain its metrological characteristics constant with time.

The concept of *time constant* has been replaced by the *step-change response time of a measuring system*. As a consequence, the response time of a temperature probe is the duration between the instant when the temperature at the input of the system is subjected to a step change between two values and the instant when the corresponding indication settles within 63% of the step change. The factors affecting the response rate of a temperature probe are

1. The mass of the probe surrounding the active temperature sensitive point.
2. The thermal conductivity of materials used in manufacturing the probe (e.g., sheathing, protective, or insulating coating).
3. The mass and conductivity of the measured material.

TYPES OF TEMPERATURE MEASURING SYSTEMS

Liquid-in-Glass Thermometers

Thermal expansion of fluids is a physical phenomenon that can be used for accurate temperature measurements, as is manifested by the adoption of gas thermometers in the lower ranges of the current international temperature scale, ITS-90 (3). The expansion of solids finds an application not only in thermostats but also in bimetallic thermometers, which are used in industry (furnaces, hot water pipes, vapor chambers). In medicine the liquid-in-glass thermometer has dominated medical thermometry in the last few centuries. However, the use of liquid-inmetal thermometers is not unknown to the chemical industry, as they can present robust and accurate low-cost solutions in a hostile environment.

A liquid-in-glass thermometer comprises a capillary tube sealed at both ends supported in a stem with a suitable scale. At the basis of the tube, a tiny reservoir (bulb) contains the liquid, which expands inside the tube and raises its height, when it gets hot, because glass expands considerably less compared with it. For the liquid to expand without difficulty, the remaining space of the tube is either empty (vacuum) or filled with a compressible gas. The scale on the stem of the thermometer is calibrated in such a way that temperature is proportional to the liquid height. This requirement dictates the use of liquids, for which the volume increases linearly with temperature. Moreover, it implies that the capillary bore has to be of the same diameter. If, due to manufacturing uncertainties, the inner diameter of the capillary tube changes, inaccuracies can occur in temperature measurement.

The choice of the liquid depends on the use of the thermometer and the temperature range of interest. Mercury, alcohol, and some synthetic oils are most commonly employed. Intended use and safety against breakage (e.g., mercury is a toxic liquid) also play a role. The main advantages of liquid-in-glass thermometers are their low cost, user-friendliness, and credibility (their accuracy can be as low as $\pm 0.01\,^{\circ}\mathrm{C}$, or even half of this value for laboratory thermometers). Their drawbacks can be summarized in their low resolution, which strongly depends on the operator, their slow response, and their fragility, which calls for an environment free of vibrations. Another disadvantage is that they have to be read locally, because they have to be in good contact with the measured medium (sometimes even fully immersed in it) and, in some situations, their size. Medical applications fall within the temperature range of both mercury (from about -30 to $500\,^{\circ}\mathrm{C}$) and alcohol thermometers (from about -80 to $70\,^{\circ}\mathrm{C}$).

Further details on the use of liquid-in-glass thermometers can be found in Reference 4.

Electrical Resistance Thermometers

The change of resistance with temperature in conductors is related to changes in free electrons' motion and atomic lattice vibrations. In fact, any conductor could be used in principle to build a resistance temperature detector (RTD). However, manufacturing limitations have led to the choice of specific metals, like copper, gold, nickel, silver, and platinum. The increase of resistance with temperature in metals is often expressed in the form

$$R_T = R_0(1 + \alpha T) \tag{1}$$

where R_0 is the conductor resistance at temperature of 0 °C, T_R is the conductor resistance at temperature of T (in degrees Centigrade) and $\alpha(^\circ C^{-1})$ is the *temperature coefficient of resistance*, which is characteristic for the metal.

The accuracy of RTDs is very high, especially when constructed with metals like platinum, for which a high degree of purity can be achieved. This is why platinum resistance thermometers are used in defining ITS-90 in the range between the triple point of hydrogen (13.8033 K) and the freezing point of silver (1234.93 K), which comprises the biological temperature range. However, the characteristic curve of resistance against temperature is usually modeled as a higher order polynomial for standardization purposes. Assuming a linear relationship, like above, results in an error smaller than 0.4 °C at 50 °C in the temperature range 0 to 100 °C (5).

The temperature sensor of an RTD consists of a wire wound on a ceramic core or a thick film coated on a ceramic surface. The sensor is encapsulated within a ceramic casing to form the temperature probe. The two ends of the wire are connected to a Wheatstone bridge (Fig. 1). The value of resistance R_V is varied until the indication on the digital voltmeter is zeroed, i.e., until it matches the unknown resistance of the temperature sensor. Therefore, the bridge imbalance, which indicates that resistance changes can be readily calibrated to reflect temperature changes by assuming a linear relationship. One problem with the circuit of Fig. 1 is that lead resistance also changes with temperature. Therefore, other arrangements of the lead wires should be implemented, which directly measure and subtract this latter resistance from the sensor resistance. Inaccuracies in RTDs can occur due to self-heating, because the current, which must be passed through the sensor to measure its resistance, causes ohmic heating. This kind of error can be minimized by reducing the flowing current and ensuring a good thermal contact between the sensor and the surrounding medium. As mentioned, RTDs can be very accurate. However, they have a large time constant and are relatively large in size for biological implantation. Moreover, they are fragile and have a high cost.

Thermistors

The principle of resistance changes with temperature is also used in thermistors. These were introduced to overcome the very small temperature coefficients of the cheaper metals used in RTDs. Thermistors are based on semiconductors, fabricated by mixing metal oxides (usually of manganese, nickel, chromium, and cobalt). The alteration of semiconductor resistance with temperature is founded on a different physical principle as compared with conductors. The exchange of electrons (and holes) from the bound valence band to the mobile conduction band depends exponentially on temperature, giving resistance changes of typically 1 Ω/0.01 °C, whereas RTDs operate with changes on the order of 5 mΩ/°C.

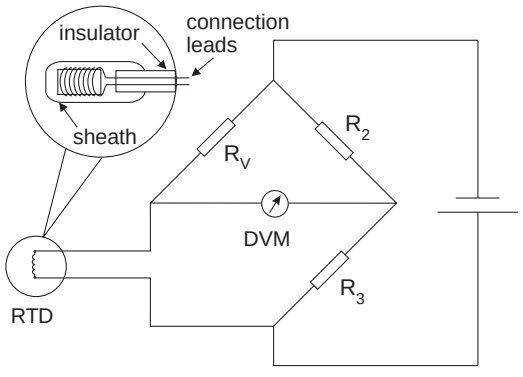


Figure 1. Schematic diagram of an RTD measuring circuit.

Thermistors can present both reduction and increase of resistance with temperature. The change of resistance can be approximated by the relationship

$$R_T = R_0 e^{B\left(\frac{1}{T}-\frac{1}{T_0}\right)} \tag{2}$$

where R_0 is the resistance at temperature T_0 and B is a constant, which characterizes the thermistor material. The values of B are in the range of 3000 to 5000 K. Thermistors are specified by their resistance at 25 °C, which ranges from several ohms to some kiloohms.

They are manufactured in the shape of beads, disks, or rods with conducting leads attached to them and encapsulated in epoxy resin or glass sheaths. Beads are used most of the time; they are available in very small sizes with a diameter less than 0.1 mm. They have response times in the range of some seconds and an accuracy of ±0.1 to ±0.5 °C.

A noteworthy development in thermistor fabrication took place in the mid-1970s, when R. Bowman introduced an elegant design of a temperature probe, which was highly insensitive to radio-frequency electromagnetic radiation (6). The probe had an outer diameter of 1 mm, with a thermistor of 0.5 mm in size, and a response time of 0.2 s. It was based on high-resistance, plastic readout-leads with resistances of about 160 kΩ/cm, which secured a heating error in the lines of less than 0.005 °C for a heating rate of 1 °C/min.

The use of standard photolithography has made possible the creation of miniaturized thermistor probes. For example, the evaporation of amorphous germanium (a-Ge) on a Pyrex glass substrate has resulted in a temperature sensing area of 50 μm × 100 μm × 0.25 μm with large resistance (4 MΩ) that needs only a 20 nA measuring current (7). The accuracy of this probe is close to 0.05 °C in the temperature range 0 to 60 °C and its response time only 14 ms.

Self-heating is a source on inaccuracy for thermistors, as it is for RTDs. However, the high resistivity of thermistor materials can allow for reductions in the measuring current, down to values that ensure the desired resolution. The high resistivity offers another advantage, namely that of eliminating the need for complicated circuitry configurations, because variation of lead resistance with temperature becomes less important. Thermistors need frequent calibration and may show a drift in their characteristics due to changes of the semiconductor materials. Their main drawback is their nonlinearity.

Thermocouples

The operation of thermocouples is based on the *Seebeck effect* or *thermoelectricity*, or the ability of heat energy to

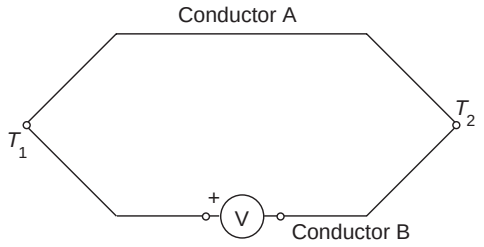


Figure 2. The Seebeck effect (thermoelectricity) is the physical principle for the operation of a thermocouple.

free electrons from a metal's surface into the free state. When a pair of dissimilar metals is connected into the form of a loop and the junctions are kept at different temperatures, an electromotive force (emf) develops (Fig. 2). If the two junctions have the same temperature, then no net emf will appear. Therefore, if one junction is kept at a constant temperature and the temperature of the other is varied, the measured emf will be proportional to the temperature difference of the two junctions.

Many thermocouple material combinations are listed by Kinzie (8). The combination determines the magnitude and the polarity of the resulting emf. The criteria for selecting a thermocouple can include cost, temperature range, chemical stability, physical properties of the measured medium, and duration of measurement. The most commonly used metals for the construction of thermocouples are rhodium, copper, iron, nickel, chromium, and aluminum, as well as some of their alloys. Some metal combinations are standardized and designated by a letter (*T, J, E, K, N, B, S, R*).

Thermocouples of the *T, J, E*, and *K* types operate in temperature ranges, which include the temperatures of biological and medical interest. The first three types are assembled with constantan (an alloy of nickel and copper) as one of the metals. Although they have lower stability than other thermocouple types, they are inexpensive and show good linearity and moderate and sensitivity (30 to 50 μV/$^\circ$C) in the biological temperature range. Their dimensions can be very small. Two metal wires of some microns in diameter (70 to 140 μm in practice) are insulated (in PVC, Teflon, or glass fiber) and connected at their distant end (Fig. 3) to form the measuring (primary) junction, which is in the size of the two wires combined. The small

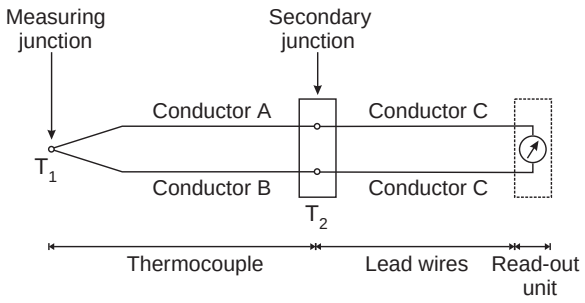


Figure 3. Schematic diagram of a thermocouple measuring circuit.

size of the wires and junctions allows for fast response times, especially when encapsulated in a thin sheath. Furthermore, it facilitates the introduction of thermocouples inside a hypodermic needle for interstitial temperature measurements more easily compared with thermistors and RTDs.

To operate a thermocouple, it is necessary to connect it to a voltage-measuring instrument. However, the connection to the readout module is usually achieved through the use of leads, which are made of a metal, different from the ones that make up the thermocouple. In the case of a type *T* thermocouple and copper leads, there is going to be one junction, namely that of thermocouple constantan to copper lead, which will introduce one more thermocouple junction. The advantage of the type *T* thermocouple is clear; for other types of thermocouples, for which none of the metals is copper, two new thermocouple junctions will be created at the connection of terminals to the readout module, whose emfs will be in series with the emf of the measuring junction. Therefore, the temperature of these secondary (Fig. 3) junctions (also called *reference* or *cold junctions*) must be stable and their emfs known to maintain the calibration of the primary junction. Historically, this used to be accomplished by immersing the reference junctions in an ice bath of 0 $^\circ$C, so that the total measured voltage was adjusted to zero when the probe temperature was also 0 $^\circ$C. A more convenient solution, however, is to provide an electronic bridge circuit between the secondary junction and the voltage measuring equipment. This circuit incorporates a resistance temperature device, whose voltage changes with temperature by the same amount as the reference junction, canceling out any variations of the latter (the temperature-sensitive device is connected in such a way that its voltage is subtracted from the emf appearing at the secondary junction).

Sources of uncertainty in temperature measurements with thermocouples include the spurious emfs from external electric and magnetic fields, temperature measurement of reference junctions, cable specifications, and drift and uncertainty of the voltage measuring instrument. Like in the case of RTDs and thermistors, thermal conduction along the metal readout wires gives another source of inaccuracy.

Interesting and practical information on thermocouple thermometry can be found in References 8 and 9.

Fiber-Optic Probes

The use of fiber-optic thermometers is necessary in situations, where electrical insulation for safety or electromagnetic immunity of the sensor is of concern. The most common medical applications for which these requirements are of paramount importance are cancer treatment with microwave or RF hyperthermia, temperature monitoring during Magnetic resonance imaging (MRI) and cardiac output measurement with the thermodilution technique. There are mainly three reasons for which conventional thermoelectric devices (RTDs, thermistors or thermocouples) should not be used inside electromagnetic fields:

1. The incident electromagnetic fields will be perturbed and scattered by the metal parts of the thermometer devices.
2. Currents will flow in the metal parts of the devices, resulting in their ohmic heating.
3. The currents in the devices can lead to spurious readings.

Fiber-optic sensors work on the principles of light absorption, reflection, scattering or interference, as well as with the effect of induced fluorescence. With respect to the implementation of the physical mechanisms, they either operate in the time domain or they involve intensity or wavelength modulation.

The simplest solution to temperature measurement with a fiber-optic probe is the use of a gallium arsenide (GaAs) crystal as the sensor. One implementation uses two optical fibers, a transmitting and a receiving one. The light transmitted by a light-emitting diode, after having been partially absorbed in the GaAs sensor at the tip of the probe, returns to the detecting module in the readout equipment. It is known that some of the light energy that gets absorbed in the crystal is used to raise electrons from the valence band to the conduction band. As the energy gap between the two bands is a known function of the crystal's temperature, the amount of absorbed power can be related to the temperature of the GaAs sensor. A second implementation with semiconductor sensors uses the same crystal and a dielectric mirror (Fig. 4) at the end of a single optical fiber and takes advantage of wavelength, instead of intensity, modulation due to temperature variations. The transmission spectrum of the crystal moves to larger wavelengths as its temperature rises (Fig. 5). This is known as the absorption/transmission shift and occurs because the energy of a photon is inversely proportional to its wavelength. When the temperature of the GaAs crystal increases, reducing the energy gap between the semiconductor's electron state bands, photons with less energy (longer wavelengths) are absorbed, making the transmission spectrum shift toward higher wavelengths. The advantages of this implementation include that the readout is independent of light intensity and, consequently, factors, which usually contribute to the attenuation in optical fibers, such as length, splices, and bending.

One of the first techniques that have been commercialized for fiber-optic probes is that of induced fluorescence. This technique makes use of the change in fluorescence decay time (lifetime) with temperature. At the tip of the probe, a thermosensitive phosphor sensor is located. This

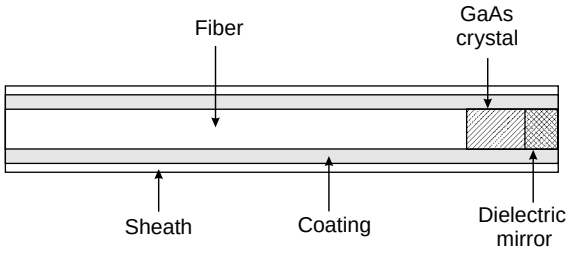


Figure 4. Diagram of a fiber-optic probe.

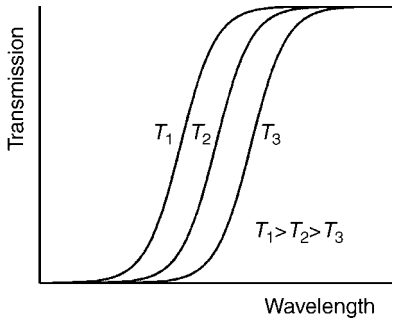


Figure 5. The absorption/transmission shift with changing temperature in semiconductors.

sensor is stimulated by the red light of a pulsed light-emitting diode and emits light over a broad spectrum in the near infrared region. The time needed for this fluorescence effect to decay depends on the temperature of the phosphor sensor; the higher the temperature, the shorter the decay time. The fluorescent signal continues to be transmitted for some milliseconds through the same fiber back to a detector in the readout equipment, even after the stimulating light from the led is off. The fluorescent decay time is then measured by a multipoint digital integration of the decay curve. The use of the decay-time technique eliminates the need to reference the output light intensity to that incident on the sensor, a significant problem experienced in intensity modulation schemes.

Another approach to temperature measurement with fiber-optic probes is Fabry–Perot interferometry, in which two parallel, partially reflecting surfaces are brought very close to each other to form an optical reflecting cavity, also known as etalon. If the distance between these surfaces (due to different coefficients of thermal expansion of the materials they are made of) or the refractive index of the material between them changes, the reflectance spectra and the interference fringes change accordingly.

Apart from electromagnetic immunity, other advantages of fiber-optic probes include minimal thermal conduction along the probe and small size, down to 0.5 mm in diameter. Their temperature range is very wide, their accuracy can reach 0.1 °C, and their response time is in the order of several hundred milliseconds.

MRI Thermometry

The use of minimally invasive surgery is very attractive due to low costs and high effectiveness. It is usually conducted in the guidance of ultrasound imaging or MRI. Thermal therapies combined with MRI have gained recognition in the recent years and include laser-induced thermotherapy (LITT), RF ablation, hyperthermia, and focused ultrasound. In these techniques, it is not recommended to use temperature sensors with metal parts, due to the reasons described above. Instead, the signals collected to reconstruct the image in the MRI devices can be used to create temperature maps inside the patients in three dimensions.

The effect of temperature on physical parameters measured by MRI devices has been known for a long time; the first study on temperature measurement with such a

device appeared in 1983 (10). This first report was based on changes of the longitudinal relaxation time (T_1). In a space free of magnetic fields, the magnetic orientations of atomic nuclei in a biological sample are directed randomly. Once placed into a magnetic field, these nuclei take two different preferred orientations, aligned with or against the external magnetic field, and the sample becomes magnetized. The transition from the random distribution of the unmagnetized sample to a magnetized state requires some exchange of quantized energy. The process of magnetization is exponential, and the rate at which a sample becomes magnetized is characterized by a quantity known as the *longitudinal relaxation* time (T_1):

$$\vec{M}_z = \vec{M}_0(1 - e^{-t/T_1}) \tag{3}$$

If the availability of the exact energy required to flip the nuclei between their two energy states is low, then T_1 will be long. The energy for changing the magnetic state of nuclei is obtained from molecular motion, and thus, it depends on temperature. At absolute zero, where there is no molecular motion, T_1 approaches an infinite value. In fact, for the hydrogen protons, which are in abundance in a biological sample, the spontaneous state change would take place once every about 10^{25} years. Smaller molecules, like water, exhibit a great deal more motion than larger ones, although one should keep in mind the difference between the freely moving bulk water in tissues and that bound at surfaces of proteins and membranes, which is less mobile. The change in T_1 with temperature can be described by

$$T_1 = T_1(\infty)e^{-E_a/kT} \tag{4}$$

where E_a is the activation energy of the longitudinal relaxation process, k is the Boltzmann constant, and T is the temperature (K).

Temperature measurement with T_1, using conventional sequences in MRI, suffers from long acquisition times, which would render the technique insufficient for most thermal procedures. Consequently, other sequences are used, like RF-spoiled gradient echo imaging, where the changes in T_1 can be assessed faster, but with limited signal-tonoise ratio (SNR). Another major problem of T_1 temperature measurements is the errors that occur from movement, which call for high-quality image registration.

Meanwhile, new measurement techniques have emerged based on two other parameters known to be affected by temperature, namely the diffusion coefficient and the proton resonance frequency (PRF) of tissue water. It is known that the diffusion coefficient D depends exponentially on temperature, because it relates the random Brownian motion of molecules with the diffusion process:

$$D \approx e^{-E_a/kT} \tag{5}$$

where E_a is the activation energy for water diffusion in this case. The problem with diffusion temperature mapping is that the motion of water in tissues is strongly dependent on the existence of biological structures, e.g., membranes. The permeability of these structures is dependent on temperature, making the diffusion process nonlinear. Moreover, biological structures of larger dimensions, e.g., muscle fibers or myelin sheaths, lead to the anisotropic diffusion of water, which leads in a tensor for the diffusion coefficient, requiring the determination of its nine elements and long acquisition times.

The most popular technique for MRI temperature measurements is based on the PRF of water. The local magnetic field B_l, perceived by the hydrogen protons in a biological sample, relates to the main magnetic field of the MRI device B_0, with the following equation:

$$B_l = (1 - \sigma_t)B_0 + \delta B_0 \tag{6}$$

where σ_t is the total screening constant of the proton and δB_0 includes all local deviations from B_0, which are not temperature dependent. As the total screening constant increases linearly with temperature and the phase distribution in an image volume (slice) depends on the local field distribution, it is possible to obtain temperature information by directly subtracting phase images. According to the Larmor equation, the phase within a volume at temperature T is given by

$$\varphi(T) = \gamma T_E[(1 - \sigma_t(T))B_0 + \delta B_0] \tag{7}$$

where γ is the gyromangetic ratio and T_E is the echo time of the gradient echo pulse sequence, which is used for the acquisition of the phase images. Therefore, the phase difference between two images at two different temperatures is

$$\begin{aligned}\Delta\varphi &= \varphi(T') - \varphi(T) = \gamma T_E[\sigma_t(T') - \sigma_t(T)]B_0 \\ &= \gamma T_E \alpha \Delta T B_0\end{aligned} \tag{8}$$

where α is the proportionality constant of linear temperature dependence of σ_t, the value of which has been measured at approximately 0.01 ppm/°C (11). The main advantage of the PRF method is its independence of tissue composition. A potential artifact may arise from the presence of lipids, because the PRF of lipid hydrogen protons is independent of temperature. Nevertheless, lipids can be suppressed in gradient echo imaging by frequency-selective slice excitation. It is clear from equations 7 and 8 that the method is sensitive to changes in δB_0 between image acquisitions, due to drift of the external field, movement of the measured object, or change of magnetic susceptibility.

Comparisons among the three methods of MRI thermometry have shown that the PRF method is the one with higher precision (12). An accuracy of 0.5 °C with an estimated resolution of 0.3 °C could be reached in a heterogeneous human phantom (13).

Radiation Thermometry

The random agitation of electrical loads or dipoles in matter at a temperature above absolute zero is associated with the generation of an electromagnetic wideband noise signal with a spectrum extending from RF waves to gamma rays. *Infrared thermography*, i.e., the recording of the temperature distribution of the body using the IR radiation emitted from its surface, and some millimeters beneath it, exploits the wavelengths between 0.8 and

20 μm. *Microwave radiometry* uses the microwave energy emitted at larger wavelengths ranging from 1 mm to 1 m; this means that temperature is collected from a depth of some centimeters inside the body, requiring a huge amount of data to solve the inverse problem formulated (14).

By definition, a *black body* absorbs all radiation incident on it. Yet, if it is at an absolute temperature T, it emits electromagnetic radiation, which is described by the Planck's radiation law. According to the latter, the power P per unit area emitted into solid angle $d\Omega$ within the frequency bandwidth Δf is given by

$$P = \frac{2h f^3 \Delta f \, d\Omega}{c^2[\exp(\frac{hf}{kT}) - 1]} \tag{9}$$

where h is Planck's constant and c is the speed of light. The emissivity ϵ of any surface at a given direction and frequency is then defined as the ratio of the power emitted through the surface to the emissive power through the black-body surface at the same frequency.

The major problem, when measuring temperature with radiation thermometers, is the knowledge of the surface emissivity, which depends on temperature. The techniques for measuring this value are complex and expensive. Apart from the uncertainty in the emissivity value, other sources of inaccuracy include the attenuation of radiation between the target and the thermometer (taking into account humidity and distance), the background radiation present (some of it will be reflected by the measured object) and errors in the radiation detectors. The latter can be either thermal detectors, using the absorbed electromagnetic energy to increase their temperature and sense a change in a physical property, such as resistance or dielectric constant, or photon detectors, which measure the direct effect of incident photons in matter with the excitation of electrons. For example, in a quantum well IR photodetector (QUIP), the incident photons produce electron-hole pairs, which are carried away by an applied voltage, giving rise to a pulse of charge.

Commercial IR cameras have a resolution of about 0.1 °C at the biological temperature range and an accuracy of 2% of the temperature reading (in degrees Centigrade). With the use of continuous calibration techniques, the accuracy can reach 0.04 °C (15). The most popular and controversial class of radiation thermometers are *infrared ear thermometers* or *infrared tympanic thermometers* (ITTs), which have been in the middle of a debate on both their accuracy (16) and their calibration (17).

More details on radiation thermometry can be found in Reference 18.

Electrical Impedance Tomography

The electrical impedivity (impedance of a unit cube) of tissue decreases by about 1.7% $°C^{-1}$ with increasing temperature, due to changes in ionic mobility (19) inside the intra- and extracellular fluids. Electrical impedance tomography (EIT) gives pictures of the conductivity distribution inside the body. Therefore, it can be used, in principal, to determine any variations induced to conductivity through temperature changes. The results in phantoms and *in vitro* were encouraging (20), despite poor resolution in central regions, nonuniform sensitivity, and equipment interference.

However, the conductivity change with temperature in the living tissue is more complicated, because it involves temperature-induced changes in the interstitial fluid volume and the cell membrane. As a consequence of thermoregulation, vasodiltation can lead to changes in tissue conductance in the same order of magnitude as the changes due to ionic mobility of electrolytes (21). Nevertheless, *in vivo*, the method correlated with direct temperature measurements within 1.5 °C, although large errors (>5 °C) did exist (22). The low cost and fast response of EIT are two reasons for which it seems reasonable to continue the effort of improving its application in temperature measurement, in particular for thermotherapy.

Ultrasound

Three methods can be used to estimate with ultrasound temperature changes inside tissues. The first one exploits the time shift of received echoes due to changes in tissue thermal expansion and speed of sound, which result in actual and apparent displacements of scattering regions, respectively. These displacements can be related to changes in temperature $\Delta T(z)$ along the direction of propagation z according to (23)

$$\Delta T(z) = \frac{c_0}{2(\alpha - \beta)} \frac{\delta t(z)}{\delta z} \tag{10}$$

where $t(z)$ is the estimated time-shift at depth z, c_0 is the speed of sound before heating, α is the linear coefficient of thermal expansion, and the coefficient $\beta = (1/c_0)(\delta c/\delta T)$ describes the change in the speed of sound with temperature. It is assumed that the speed of sound varies linearly with temperature up to about 45 °C, whereas the term $(\alpha - \beta)$ depends on tissue type. It is clear that the limitation of the above method is the requirement for prior knowledge of the speed of sound and thermal expansion coefficients.

The second method is based on the changes of ultrasound attenuation with temperature, which, however, are more pronounced at temperatures larger than 50 °C. Therefore, this technique is a good candidate for temperature monitoring in thermal ablation. The third method makes use of the changes on backscattered energy, which could be as much as 5 dB over the temperature range from 37 to 50 °C for individual scatterers (24).

Although temperature measurements by ultrasound are a convenient and inexpensive alternative to MRI, the performance of all of the above methods needs to be evaluated *in vivo*, where sophisticated motion tracking techniques have to be employed to correct for tissue movement.

BIBLIOGRAPHY

1. Pearce JMS. A brief history of the clinical thermometer. QJM Mon J Assoc Phys 2002;95:251–252.
2. ISO/DGuide 99999, International vocabulary of basic and general terms in metrology (VIM). 3rd ed. 2004
3. Preston-Thomas H. The International Temperature Scale of 1990 (ITS-90). Metrologia 1990;27:3–10.

4. Nicholas JV, Liquid-in-glass thermometers. In: Webster JG editor. The Measurement, Instrumentation and Sensors Handbook. Boca Raton, FL: CRC Press; 1999
5. Childs PRN, Greenwood JR, Long CA. Review of temperature measurement. Rev Sci Instrum 2000;71:2959–2978.
6. Bowman RR. A probe for measuring temperature in radio-frequency-heated material. IEEE Trans Microwave Theory Techn 1976;24:43–45.
7. Schuderer J, Schmid T, Urban G, Samaras T, Kuster N. Novel high-resolution temperature probe for radiofrequency dosimetry. Phys Med Biol 2004;49:N83–N92.
8. Kinzie PA. Thermocouple Temperature Measurement. New York: Wiley; 1973.
9. Kerlin TW. Practical Thermocouple Thermometry. Research Triangle Park, NC: ISA International; 1998.
10. Parker DL, Smith V, Sheldon P, Crooks LE, Fussell L. Temperature distribution measurements in two-dimensional NMR imaging. Med Phys 1983;10:321–325.
11. Peters RD, Hinks RS, Henkelman RM. Ex vivo tissue-type independence in proton-resonance frequency shift MR thermometry. Magn Reson Med 1998;40:454–459.
12. Wlodarczyk W, Hentschel M, Wust P, Noeske R, Hosten N, Rinneberg H, Felix R. Comparison of four magnetic resonance methods for mapping small temperature changes. Phys Med Biol 1999;44:607–624.
13. Gellermann J, Wlodarczyk W, Ganter H, Nadobny J, Fõhling H, Seebass M, Felix R, Wust P. A practical approach to thermography in a hyperthermia/magnetic resonance hybrid system: Validation in a heterogeneous phantom. Int J Radiat Oncol Biol Phys 2005;61:267–277.
14. Leroy Y, Bocquet B, Mamouni A. Non-invasive microwave radiometry thermometry. Physiol Meas 1998;19:127–148.
15. Baker JM, Noman JM, Kano A. A new approach to infrared thermometry. Agr Forest Meteorol 2001;108:281–292.
16. Craig JV, Lancaster GA, Taylor S, Williamson PR, Smyth RL. Infrared ear thermometry compared with rectal thermometry in children: A systematic review. Lancet 2002;360:603–609.
17. Pušnik I, van der Ham E, Drnovšek J. IR ear thermometers: What do they measure and how do they comply with the EU technical regulation? Physiol Meas 2004;25:699–708. Erratum in Physiol Meas 2004;25:1337.
18. DeWitt DP, Nutter GD, editors. Theory and Practice of Radiation Thermometry. New York: Wiley; 1988.
19. Brown BH. Electrical impedance tomography (EIT): A review. J Med Eng Technol 2003;27:97–108.
20. Conway J, Hawley M, Mangnall Y, Amasha H, van Rhoon GC. Experimental assessment of electrical impedance imaging for hyperthermia monitoring. Clin Phys Physiol Meas 1992; 13(Suppl A):185–189.
21. Gersing E. Monitoring temperature-induced changes in tissue during hyperthermia by impedance methods. Ann N Y Acad Sci 1999;873:13–20.
22. Paulsen KD, Moskowitz MJ, Ryan TP, Mitchell SE, Hoopes PJ. Initial in vivo experience with EIT as a thermal estimator during hyperthermia. Int J Hyperthermia 1996;12:573–591. discussion 593–594.
23. Arthur RM, Straube WL, Trobauch JW, Moros EG. Non-invasive estimation of hyperthermia temperatures with ultrasound. Int J Hyperthermia 2005;21:589–600.
24. Straube WL, Arthur RM, Theoretical estimation of the temperature dependence of backscattered ultrasonic power for noninvasive thermometry. Ultrasound Med Biol 1994;20: 915–922.

See also HYPERTHERMIA, INTERSTITIAL; HYPERTHERMIA, SYSTEMIC; HYPERTHERMIA, ULTRASONIC; NONIONIZING RADIATION, BIOLOGICAL EFFECTS OF; THERMOCOUPLE; THERMISTOR.

THERMOREGULATION. See BIOHEAT TRANSFER.

THERMOTHERAPY. See HEAT AND COLD, THERAPEUTIC.

TISSUE ABLATION

DIETER HAEMMERICH
Medical University of
South Carolina,
Charleston, South Carolina

INTRODUCTION

Tissue ablation (literal translation "removal") is the destruction of diseased (*pathologic*) body tissue, with the aim to cure a disease. Tissue destruction is achieved by thermal methods, or by application of chemical substances (e.g., ethanol). Thermal methods cause either local heating or cooling of the tissue to lethal temperatures (typically below -40 °C, or above 50 °C). A number of different physical principles are employed for heating and cooling tissue, such as radio frequency (rf) electric current, microwaves, laser, ultrasound, and cryogenic cooling.

Current clinical applications include treatment of heart arrhythmia, cancer (liver, lung, brain, kidney, bone, prostate), uterine bleeding, varicose veins, and enlarged prostate (benign prostate hyperplasia), with other emerging applications (see Table 1 for an overview). Typically, an applicator is introduced under imaging guidance [ultrasound imaging, fluoroscopy, computerized tomography (CT), magnet resonance imaging (MRI)] into the body, to the treatment site. Then the tissue region around the applicator is ablated, destroying the diseased tissue area.

This article describes physical principles, clinical devices, and applications of the different ablation modalities.

PHYSICAL PRINCIPLES OF TISSUE ABLATION

Thermal Ablation Methods

All thermal ablation methods rely on thermal conduction to some extent to heat or cool a region of tissue near the applicator. The problem of thermal ablation can mathematically be described by following heat-transfer equation:

$$\rho c \frac{\partial T}{\partial t} = \nabla \cdot k \nabla T + Q_{\mathrm{A}} - Q_{\mathrm{p}} \qquad (1)$$

Energy Q_{A} ($\mathrm{W \cdot m^{-3}}$) is applied to the tissue by the applicator, resulting in heating (or cooling) of the tissue. Some energy Q_{p} is carried away by blood perfusion; depending on tissue type, perfusion may be negligible (e.g., heart), or may have major impact (e.g., liver). The left-hand side of equation 1 describes how local tissue temperature T changes, depending on tissue density ρ and tissue specific heat c. The first term on the right-hand side describes how thermal energy is conducted through the tissue, depending on thermal conductivity k. Different models are available in the literature that approximate tissue perfusion, the

Table 1. Frequency of Use of Different Ablation Modalities in Clinical Applications[a]

Ablation Modalities	rf	Cryo	Microwave	Ultrasound	Laser	Chemical
Cardiac ablation	[b]	[c]	[d]	[d]	N/A[e]	N/A[e]
Tumor ablation (liver, lung, kidney)	[b]	[c]	[d]	N/A[e]	[d]	[d]
Endometrial	[c]	[c]	[d]	N/A[e]	N/A[e]	N/A[e]
Prostate (cancer, enlarged prostate)	[d]	[c]	[c]	[d]	N/A[e]	N/A[e]
Intervertebral disk	[d]	N/A[e]	N/A[e]	N/A[e]	[d]	N/A[e]
Endovascular	[c]	N/A[e]	N/A[e]	N/A[e]	[c]	N/A[e]
Cornea	[c]	N/A[e]	N/A[e]	N/A[e]	[b]	N/A[e]

[a]Adapted from Jie Zhang, University of Wisconsin-Madison.
[b]Used frequently.
[c]Used sometimes.
[d]Used rarely/under investigation.
[e]not applicable, not used clinically.

most widely used being Pennes' Bioheat equation (1). In the Bioheat equation, blood perfusion is modeled as a distributed heat sink term, according to

$$Q_p = \rho_{bl} c_{bl} w_{bl} (T - T_{bl}) \tag{2}$$

where ρ_{bl} (kg·m^{-3}), c_{bl} [J (kg·K)$^{-1}$] and T_{bl} are density, specific heat, and temperature of blood, respectively. The parameter T is the tissue temperature, and w_{bl} is the blood perfusion (1·s^{-1}).

Figure 1 shows the heat transfer problem on the example of cardiac radio frequency (rf) ablation.

For thermal ablation methods the resulting zone of tissue death is usually called *thermal lesion*, or *coagulation zone* (for heat-based methods). In cardiac ablation literature, *lesion* is the accepted term. The term lesion should be avoided in tumor ablation applications (though it is used, especially in earlier literature), since lesion is a general medical term referring to a pathological part of tissue (e.g., often tumors are called lesions).

Following each of the different ablation principles will be described. A comparison between thermal ablation methods is given in Table 2.

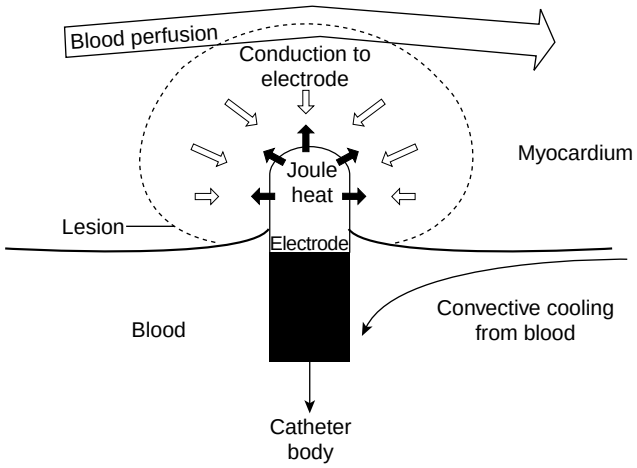


Figure 1. Heat transfer during cardiac rf ablation. Tissue close to the electrode is heated by resistive heating from rf current. Heat is then conducted thermally into the tissue. Heat is lost due to blood perfusion in the myocardium, due to thermal conduction through the electrode, and from convective cooling due to blood flow in the heart chamber. (From Ref. 2.)

Radio Frequency Ablation. Radio frequency ablation is the currently most widely used ablation method. The rf current in the frequency range of typically 450–500 kHz is applied to the tissue via the inserted electrode (catheter); a dispersive electrode (grounding pad) is required to serve as current return path. Inside the tissue electric current is carried by ions (mainly sodium, potassium, and chloride), that is, the ions oscillate due to the alternating electric field. The ion movement results in resistive heating (joule heating) of the tissue due to ionic friction, with most heat generated at the location of maximum rf current density (typically closest to the catheter). One parameter that is often used to characterize electrode performance is the specific absorption rate (SAR), expressed in units of W·kg^{-1}. The SAR determines how much mass related power is deposited at a certain location in the tissue. To determine the SAR resulting from a specific rf electrode, we first have to solve the electric field problem to determine the electric field strength E in the tissue. The SAR can then be calculated from the local version of Ohm's law according to

$$\mathrm{SAR} = \frac{\sigma}{\rho}|E|^2 = \frac{1}{\sigma \cdot \rho}|J|^2 \tag{3}$$

where σ is the electrical tissue conductivity, ρ is the tissue density, and J is the electric current density.

Figure 2 shows the tissue temperature distribution around a cardiac rf catheter after 45 s.

Maximum tissue temperature during rf ablation is limited to ~110 °C; above that temperature tissue water vaporizes, forming an electrically insulating barrier preventing further energy deposition. If too much rf power is applied, tissue around the electrode can char; this area of carbonization is electrically insulating and irreversible. Applied rf power therefore has to be controlled to prevent tissue charring (carbonization). In cardiac applications, maximum temperature is further limited to prevent tissue cavitation from rapidly expanding vapor.

There are three control methods currently used in clinical rf devices:

Power Control. Applied rf power is kept constant throughout the ablation procedure.

Temperature Control. The ablation catheter has one or more thermal sensors (thermocouples or thermistors) embedded in the electrode tip, or at a specified distance

Table 2. Comparison of Heat-Based Ablation Modalities

Modality	Advantages	Disadvantages
Radio frequency	Simple applicator design	Limited by tissue charring Dispersive electrode (ground pad) required Not usable under MRI (rf interference)
Microwave	High tissue temperatures Short application times Constructive interference of microwaves from multiple applicators	
Ultrasound	Directional applicators possible Can be used noninvasively	
Laser	Simple applicator design	Limited by tissue charring
Cryo	Iceball visible under ultrasound imaging Reversible tissue damage (for short application times)	

from the catheter. Applied rf power is controlled to keep the measured temperature constant.

Impedance Control. Applied rf power is controlled depending on tissue impedance, as measured between the active electrode and the grounding pad. Initially, impedance drops since electrical tissue conductivity increases with temperature due to higher ion mobility. As tissue vaporizes, an increase in impedance results. When impedance exceeds a certain threshold, rf power is temporarily shut down to allow vapor to settle, and then reapplied at a lower level. Figure 3 shows a typical time course of impedance during an impedance-controlled ablation procedure.

Microwave Ablation. Microwaves (MW) are electromagnetic (EM) waves in the frequency spectrum from 300 MHz to 300 GHz. During MW ablation, a MW antenna is inserted into the tissue, radiating microwaves into the tissue. These EM waves cause polar water molecules in the tissue to align with the applied alternating electric field. The resulting rotating water molecules cause frictional heating of the tissue. For MW ablation, microwaves at frequencies of 915 MHz or 2.45 GHz are used due to Federal Communications Commission (FCC) restrictions, with wavelengths in the centimeter range inside the tissue. To determine the SAR during MW ablation, first the Maxwell equations have to be solved to determine the electric field distribution in the tissue. Then the SAR can be calculated using equation 3; note that the tissue conductivity σ is strongly dependent on frequency. Since the propagation of microwaves is not hindered by vapor or charred tissue, much higher tissue temperatures compared to rf (up to 150 °C) can be obtained (3).

Different antenna types have been proposed. The dipole antenna has been commonly used (see Fig. 4); other common antenna types are slot antennas and monopoles (4,5). Note that the SAR of MW antennas is dependent on insertion depth of the antenna, and it is significantly different at smaller insertion depths.

Contrary to other ablation methods, there has not been use of any advanced control methods to adjust applied power during MW ablation. So far, constant power (typically 40–100 W, depending on application) has been used. For MW ablation, impedance match between antenna and tissue is important. If there is mismatch in impedance, significant amounts of power are reflected, resulting in

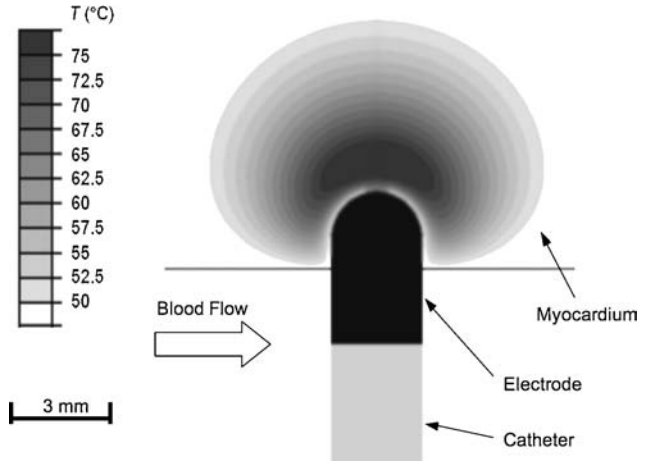


Figure 2. Tissue temperature (computer simulation) after 45 s of cardiac rf ablation. Catheter is 2.3 mm in diameter, with 5 mm electrode length, and inserted 2 mm into tissue. Note that location of maximum temperature is ~1 mm distant from the electrode due to electrode cooling from blood flow in the heart chamber. Outermost boundary (50°C) marks the thermal lesion boundary.

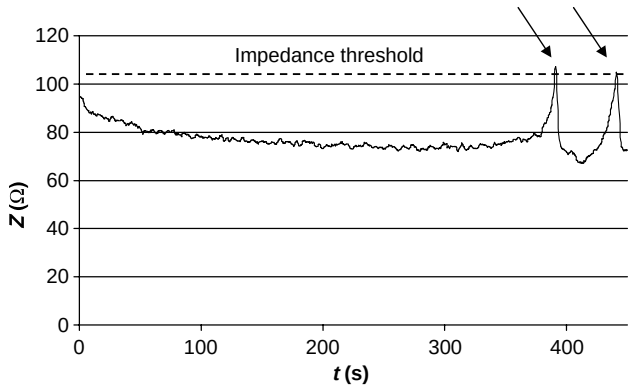


Figure 3. Typical time course of impedance Z (measured between electrode and grounding pads) during rf ablation. Initially, impedance decreases as electrical tissue conductivity increases from heating. When tissue starts to vaporize around the electrode (*arrows*), impedance rises. When impedance exceeds threshold (*dotted line*), power is shut down for 15 s, and then reapplied.

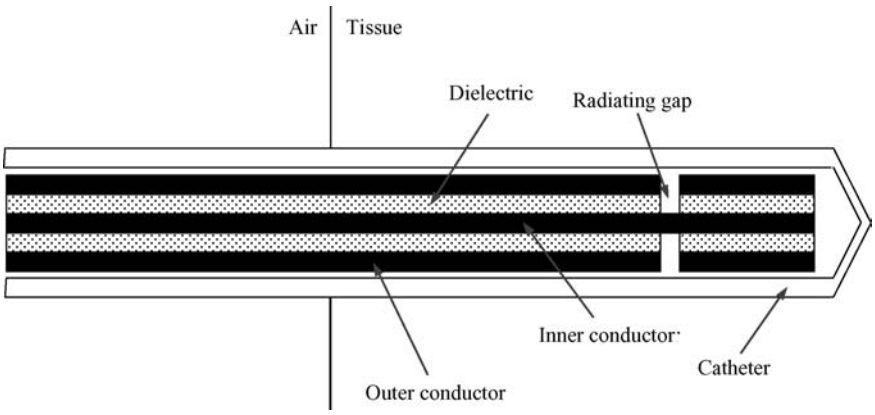


Figure 4. Cross-section of a dipole antenna for microwave ablation. Microwaves are radiated from the gap into the tissue.

undesirable heating of the antenna shaft and cable. Proper matching is complicated by the fact that tissue properties change significantly during heating.

Antenna Arrays. For all thermal ablation methods, the SAR is only significant very close to the applicator; most of the tissue heating is caused by thermal conduction. This can be a disadvantage when large tissue volumes need to be heated, or tissue close to large blood vessels is heated. Microwaves have an advantage in that regard; microwaves from multiple sources can produce constructive interference when the sources (i.e., antennas) are placed accordingly (6). Thereby, large SAR at far distances from the antennas can be achieved. Figure 5 shows the SAR of a square array of four microwave antennas, both for 915 MHz and 2.45 GHz microwaves. Depending on distance of the antennas and wavelength (i.e., frequency), different interference patterns result. An additional way to adjust the SAR pattern is to modify the phase angle at which the microwaves are supplied to the antennas in the array (7).

Ultrasound Ablation. Ultrasound as used in medical applications is typically in the frequency range of 0.5–20 MHz. At high enough intensities, the absorbed mechanical energy results in tissue heating. There are two fundamentally different ways of delivering ultrasound energy to the tissue. Ultrasound can be applied noninvasively from the outside of the body, and focused at the treatment site (High Intensity Focused Ultrasound, HIFU). In addition catheters with integrated ultrasound transducers can be inserted into the tissue, similar to other ablative techniques (Interstitial Ultrasound Thermal Therapy, or Direct Ultrasound Ablation).

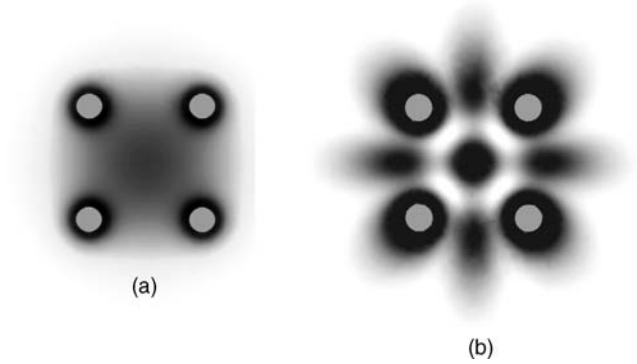


Figure 5. Four dipole antennas are placed in an array 2 cm apart, and driven in-phase at frequencies of 915 MHz (a) and 2.45 GHz (b).. The SAR is shown in linear scale (black = maximum). Note constructive interference at array center (*left*), and complex interference pattern with constructive and destructive interference (*right*) due to shorter wavelength at 2.45 GHz. (Images provided courtesy of Deshan Yang, University of Wisconsin-Madison.)

High Intensity Focused Ultrasound. Ultrasound has a high penetration depth, and can therefore be applied from outside the body and still reach deep tissue sites. Typically, an array of ultrasound transducers is placed outside the body, and the waves are coupled into the tissue by a gel. The ultrasound waves are focused at the desired location, resulting in high intensity and heating near the focal spot (Fig. 6). The location and depth of the focal spot can be adjusted by modifying the phase difference between the transducer elements. A large area can be ablated by moving the focal spot and "painting" the desired area. The transducer may be attached to a computer controlled mechanical positioning system (8).

Interstitial Ultrasound Thermal Therapy (Direct Ultrasound Ablation). One or more transducer elements are placed inside a catheter. Figure 7 shows the schematics of such a catheter, and Fig. 8 shows achieved coagulation zones at different power levels. The catheter is inserted invasively into the application site, and tissue close to the catheter is heated by ultrasound. Sector transducers that emit ultrasound at angles between 30 and 270° (10) can be used, allowing for directional ablation which is important in certain applications like prostate treatment; on the other hand rf, MW, and laser applicators provide typically uniform heating (axial symmetric) around the applicator. Another potential advantage of direct ultrasound ablation is the use of the transducer elements also for imaging; an ultrasound image from inside the treatment zone can be obtained allowing the monitoring of tissue heating.

Laser Ablation (Laser Interstitial Thermal Therapy, LITT). When high intensity laser light is applied to tissue, the light is absorbed, resulting in tissue heating. The penetra-

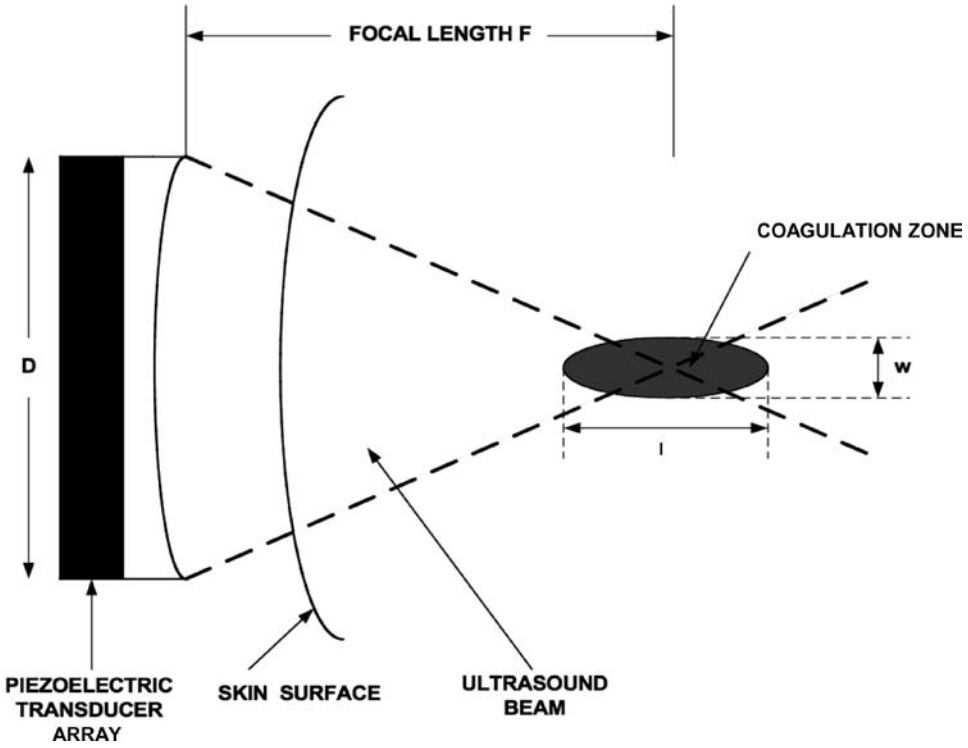


Figure 6. Schematic Diagram showing the principle of HIFU. An array of ultrasound transducers emits ultrasound waves into the tissue, which are focused and result in tissue coagulation around the focal point. The location of the focal point can be shifted by varying the phase between the array elements. (Image provided courtesy of Abhijit Sathaye, University of Wisconsin-Madison.)

tion depth is dependent on wavelength, where penetration increases with wavelength. The most widely used laser type, due to large wavelength and penetration is the Nd:YAG laser with 1064 nm wavelength (near-infrared, IR). This laser has a penetration depth of 3–4 mm. A quartz fiber with diffuser element at the tip is used to introduce the laser light to the treatment site. The transmission of the light into tissue changes during tissue heating. After tissue coagulation occurs, transmission is reduced to 69% of normal and further decreases to 15% of normal with onset of carbonization. Control of applied power to avoid carbonization is therefore essential. Typically, either constant power, or temperature feedback control where applied power is controlled so that the fiber tip is kept at constant temperature is used. Unlike most other ablation modalities, laser ablation has the advantage that the fibers are MRI compatible, allowing use of MR imaging (11).

Applicator Cooling. In the first heat-based ablation devices, the size of the achieved coagulation zone was insufficient for many applications like tumor ablation. Even today, there is a trend toward larger coagulation zones to enable treatment of larger volumes. One successful method that has been applied to virtually all heat-based methods is cooling the applicator (e.g., catheter). In all catheter-based ablation methods, the highest SAR and subsequently highest temperature is obtained close to the catheter. In extreme cases, this can lead to tissue carbonization, which should be avoided as discussed earlier. Applicator cooling is achieved by circulating cooled water inside the catheter. Thereby, tissue close to the catheter is cooled and carbonization is prevented. Furthermore, the location of highest tissue temperature is more distant from the applicator (see Fig. 9). Since maximum tissue temperature is limited (e.g., to $\sim$110 °C for rf ablation), this shift of maximum temperature results in an increase of the size of coagulation zone. The radial dimension of the coagulation zone can be increased by up to a factor of two when applicator cooling is used. Several commercial devices employ applicator cooling.

Cryoablation (Cryotherapy, Cryosurgery). Cryoablation is historically older than other ablative methods, and was introduced in the early 1960s (12). However, it

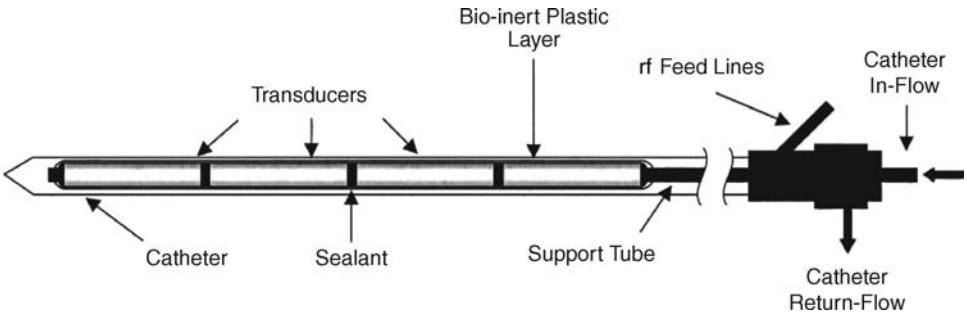


Figure 7. Catheter for Interstitial Ultrasound Thermal Therapy. An array of four transducers is used, and the catheter is cooled by circulating water. (From Ref. 9.)

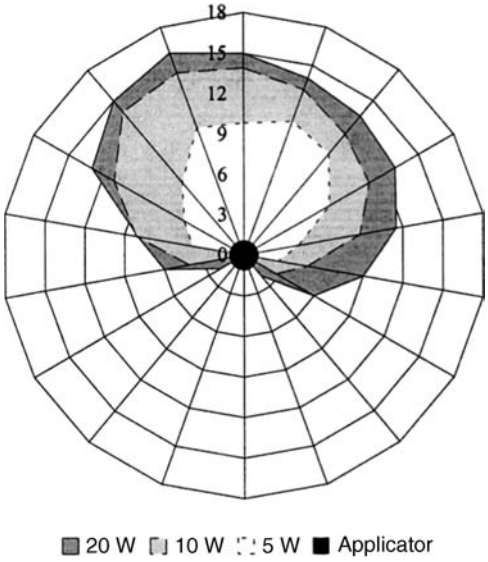


Figure 8. Thermal lesion sizes obtained with cooled ultrasound applicator at different power levels. Directional ultrasound arrays are used, which allows for better control of the coagulation zone. (From Ref. 9.)

did not receive considerable interest until the 1990s, when the development of interoperative ultrasound imaging allowed guidance of probe placement and the monitoring of the procedure.

Cryoablation relies on cooling (freezing) tissue to cause injury. Contrary to most heat-based methods, cryoablation relies solely on thermal conduction. A cryoprobe (see Fig. 10) is introduced into the tissue, and cooled by circulating a cryogen inside. Cryogens used for cryoablation include liquified gases, such as nitrogen, argon, helium, and nitrous oxide. Liquid nitrogen has a boiling temperature of –196 °C. Temperatures reached at the cryoprobe are down to –160 °C, resulting in formation of an iceball around the cryo probe. Nitroous oxide has a higher boiling temperature of –88 °C, but has the advantage of being safer in case of leakage into the body compared to other gases.

Cooling with argon and helium is based on the Joule–Thompson effect, which involves expansion of a gas

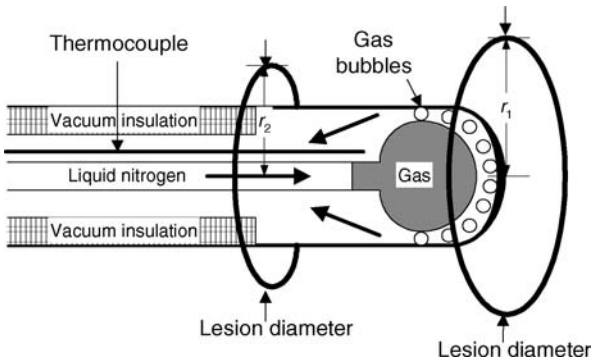


Figure 10. Internal structure of a typical cryoprobe based on liquid nitrogen (LN_2). The probe has vacuum insulation to prevent freezing of the probe shaft and subsequent destruction of normal tissue. The LN_2 changes phase when it hits the warm metal surface of the probe tip. Thus, a thin film of gas bubbles is formed on the metal surface resulting in lowest temperatures and largest ice ball near the tip. (From Ref. 13.)

through an orifice. The Joule–Thompson effect can produce either heating or cooling, depending on the type of gas, temperature, and pressure before expansion. For cryoablation, argon is decompressed from ~3000 psi (21,000 kPa) to ~150 psi (1000 kPa), resulting in cooling down to –186 °C (the boiling temperature of argon). Figure 11 shows a typical temperature distribution in tissue surrounding a cryoprobe.

One advantage of cryoablation over heat based methods is the visibility of the ice ball using ultrasound imaging (14). The interface between the ice ball and surrounding tissue is evident in ultrasound imaging (see Figure 12), and allows real-time monitoring of the ablation procedure.

Chemical Ablation. During chemical ablation a cytotoxic agent is injected into the tissue site to be treated, and diffuses into the tissue from the injection site. The most commonly used agents are ethanol and acetic acid. When ethanol is injected (ethanol ablation or percutaneous ethanol injection) into tissue, it causes cell death by cell dehydration, protein denaturation, and thrombosis of small vessels. Acetic acid has the ability to dissolve lipids

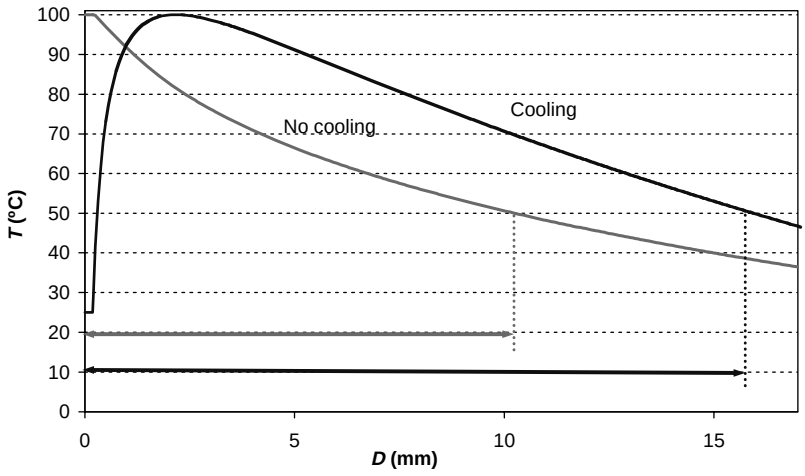


Figure 9. Tissue temperature (T) as a function of distance (D) from applicator. With catheter cooling, location of maximum tissue temperature is moved farther away from the applicator, resulting in an increase in diameter of the coagulation zone. Ablation zone dimensions are indicated by arrows as the regions with temperatures >50 °C. This image shows tissue temperature during rf ablation (from computer model), but the principle of cooling is applicable to all heat-based ablation methods.

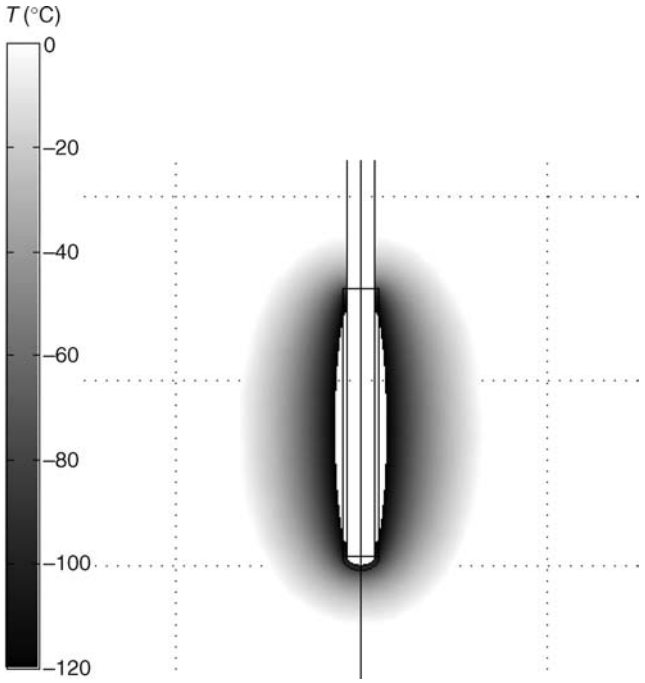


Figure 11. Tissue temperature next to a cryoprobe after 12 min. Outer gray boundary represents the border of the ice ball. Dotted lines are 20 mm apart. (Image provided courtesy of Cheolkyun Kim, University of Wisconsin-Madison.)

(e.g., cell membrane) and possesses a higher toxicity than ethanol.

PRINCIPLES OF THERMAL TISSUE INJURY

Tissue Injury from Heating

The normal range of human body temperature is between 36 and 38 °C; with fever, body temperature can rise up to

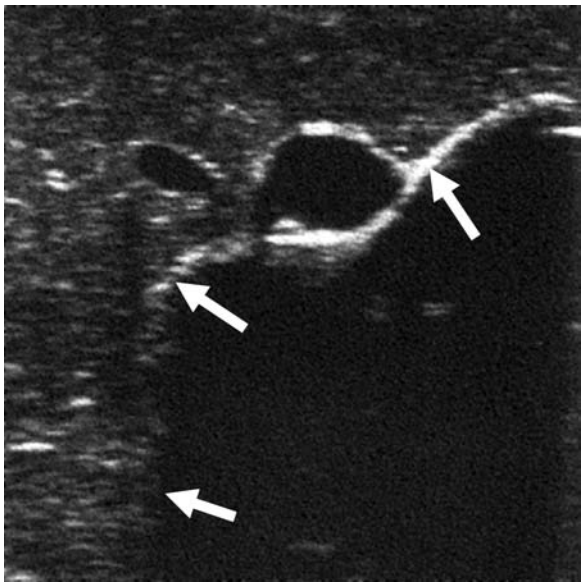


Figure 12. Ice ball forming during cryoablation is visible under ultrasound imaging (*arrows*). (From Ref. 14.)

Table 3. At Elevated Tissue Temperatures Different Effects Take Place Depending on Temperature and Time

Temperature, °C	Tissue Effects
36–38	Normal physiological range
38–42	Fever
>42	Elevated rates of enzyme activity, cell death possible
45	Protein coagulation (after 1–2 h)
50	Protein coagulation (after 2–3 min)
60–100	Protein coagulation (instantaneous)
>100	Tissue vaporization, carbonization

42 °C, which is the maximum temperature a person can typically sustain. Significant cell damage occurs above 42 °C (hyperthermia) with possible cell death, depending on time duration for which temperature is increased (15).

There are several different cell responses to elevated temperature. At low temperatures (42–45 °C), a number of subtle changes in metabolic activity, pH, blood flow, and vascular permeability occur. From ~45 °C protein coagulation occurs after 1–2 h of elevated temperature. Above ~60 °C instantaneous protein coagulation occurs, and above 100 °C tissue vaporization takes place, with possible carbonization at even higher temperatures. Table 3 lists the different temperature ranges and effects on tissue.

Since very high temperatures in excess of 60 °C are obtained during ablative therapies, cell death (necrosis) due to coagulation (i.e., coagulative necrosis) is the most important mechanism of tissue damage; the region of cell death is practically the region where tissue coagulation occurs. From Table 3, we see that we can roughly define the region of cell death (the coagulation zone) as the region where tissue reaches temperatures > 50 °C, since ablative therapies have application times of typically between 1 and 30 min. Figure 13 shows a typical coagulation zone in liver tissue, cut right after rf ablation.

To determine exactly whether cell death results at a certain location, the time history of temperature has to be taken into account. It has been shown in many types of tissue that there is an exponential relationship between cell death, temperature, and time. Above 43 °C, the time required to cause cell death is cut in half with each degree centigrade of temperature increase (18). Even though the time may vary between different tissue types, this exponential time–temperature relationship is the same for all tissues. Figure 14 shows the time required for cell death for different types of tissue, plotted on a double-logarithmic scale. The exponential time–temperature relationship can be modeled mathematically by the Arrhenius model. The time–temperature relationship can then be expressed by an isoeffect equation:

$$t_1 = t_2 \times R^{(T_1 - T_2)} \tag{4}$$

where t_1 and t_2 are the treatment durations at treatment temperatures T_1 and T_2, respectively. The parameter R can be assumed a constant with a value of 0.5 above 43 °C, and 0.25 below 43 °C. Because the onset of appreciable tissue damage occurs at ~43 °C, it has been suggested by Sapareto and Dewey (18) to quantify tissue damage by a

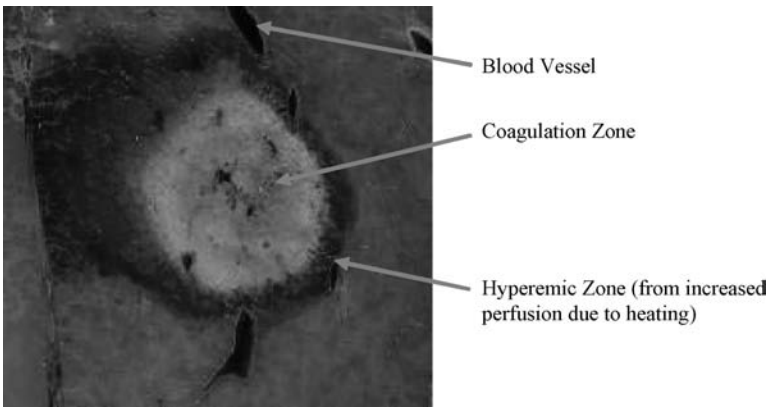


Figure 13. Gross pathology of a coagulation zone created by rf ablation *In vivo* in porcine liver. Liver was sliced right after the ablation procedure. The hyperemic zone contains viable cells.

thermal dose at 43 °C. This thermal dose is expressed as cumulative equivalent minutes at 43 °C (CEM_{43}); that is, a certain thermal treatment has the same effect as keeping the tissue at 43 °C for CEM_{43} minutes. If we set T_1 to 43 °C in equation 4 and allow temperature to be changing with time (as it does during ablative treatments), we obtain

$$CEM_{43} = \int R^{[43-T(t)]} dt \tag{5}$$

Once CEM_{43} exceeds a certain critical value, the tissue can be considered to be destroyed (i.e., ablated). The critical value of CEM_{43} has been measured for many tissues, and is ~340 min for liver; that is, tissue with $t_{43} > 340$ min can be considered to be destroyed. Even though the relationship stated in equation 5 may show inaccuracies in certain cases (e.g., long application times), it is an accurate approximation for the time durations and temperatures that occur during thermal ablation procedures.

Tissue Injury from Freezing

There are two basic mechanisms of tissue injury due to very low temperatures, dehydration and intracellular ice formation (19). In general, extracellular water will freeze before intracellular water. The salinity of the remaining extracellular water increases, resulting in water transport from intra- to extracellular space due to difference in solute concentration. This water transport causes *dehydration* of the cells, and may result in cell death. While mainly dehydration happens at lower cooling rates (<50 °C·min^{-1}), *intracellular ice formation* occurs at higher cooling rates. Formation of ice crystals inside the cells results in damage of cell membranes, organelles, and ultimately cell death. Generally, it is assumed that a minimum temperature of −40 °C is required to ensure cell death.

Other mechanisms assumed to contribute to cell death are loss of blood supply and bursting of cells due to water pouring back into the cells during thawing. Multiple freeze–thaw cycles are typically used during cryoablation to accentuate these effects. An animal study has shown that the iceball boundary visible under ultrasound correlates with the boundary of cell death within 1 mm (20).

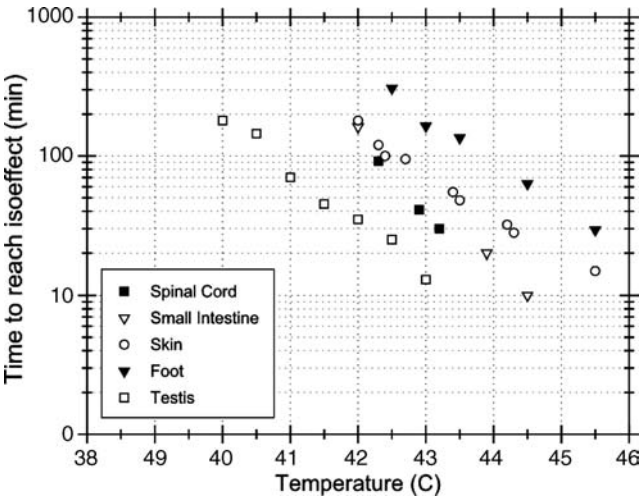


Figure 14. Time–temperature relationship of cell death for different cell types. Note that all cell types exhibit the same exponential relationship (i.e., parallel lines), even though they have different sensitivity (i.e., different times to cell death). (From Ref. 17.)

CLINICAL APPLICATIONS AND DEVICES

In the following section, different applications of tissue ablation are discussed. The areas where ablative therapies are used most widely clinically are treatment of cardiac arrhythmia and cancer. For each application, we will briefly discuss the clinical background, describe devices, and review imaging modalities used for guidance and monitoring.

Cardiac Catheter Ablation

Cardiac ablation is now a standard treatment method for different types of *cardiac arrhythmia* (i.e., abnormal heart beats) (21). Even though other ablative methods have been investigated, rf ablation is the most widely clinically used technique. Recently, cryoablation systems have become commercially available and are clinically used. Microwave systems have also become available. Other modalities like laser, ultrasound, and chemical ablation are mainly found in the research literature with no commercial devices currently available.

One major difference between cardiac ablation and ablation at other sites is that the catheter electrode is in direct contact with the blood pool inside the heart chamber. High temperatures can result in blood clot formation,

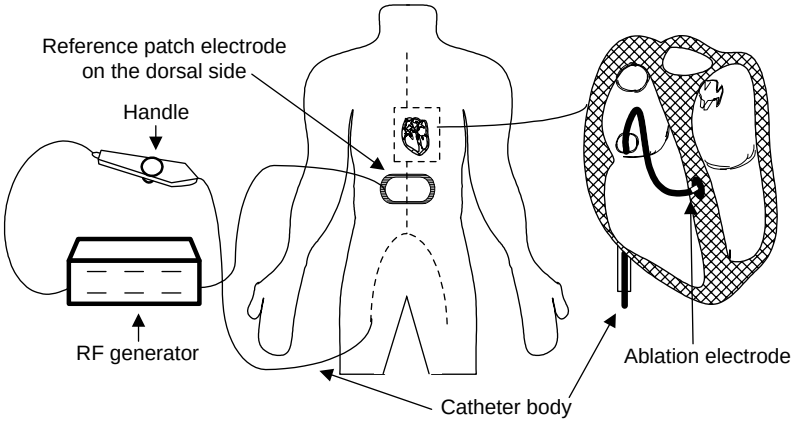


Figure 15. Overview of rf ablation system for treatment of cardiac arrhytmia. An ablation catheter is inserted through a leg vein, and steered into the heart to the treatment site. The procedure is guided by X-ray fluoroscopy. (From Ref. 22.)

which has to be avoided. Blood flow inside the heart results in cooling of the electrode (see Fig. 1), resulting in varying sizes of thermal lesion depending on blood velocity at the specific location inside the heart.

Clinical Background. Cardiac arrhythmias result from abnormalities in the conduction pathways in the cardiac tissue. The types of cardiac arrhythmias treated by ablative methods can be broadly classified into two categories.

Regular Tachycardias with a Discrete Mechanism. This is the condition of the heart (atrium or ventricle) beating too rapidly, typically at a rate >150 beats·min^{-1} at rest. In a normal heart, the excitation (and associated contraction) starts at the sinoatrial (SA) node, the heart's pacemaker. Tachycardia results from excitation originating from locations other than the SA node, or from circular conduction involving abnormal pathways. Tachycardia can further be divided depending on where it originates, into supraventricular [above the ventricle, meaning the atria, the atrioventricular (AV) node and the bundle of His], and ventricular tachycardia (within the ventricle or purkinje system). These tachycardias can often be treated successfully with ablation at a single point or small area of the heart.

Atrial Fibrillation. Atrial fibrillation (AF) is the result of disorganized excitation of the atria, resulting in irregular contraction of the ventricles. The blood flow through the atria is hampered, possibly forming blood pools in the atria. Eventually, a blood clot may form that could lead to a stroke or heart attack (myocardial infarct). Atrial fibrillation is the most common type of arrhytmia, with $\sim$2.2 million people affected in the United States, and 160,000 new cases diagnosed each year. Atrial fibrillation appears to often originate with abnormal excitations from the entry points of the heart, most commonly the pulmonary (lung) veins and less often the systemic veins from the body. Continuation of AF may depend on large areas of the atria. Because of the complexity of both the initiation and maintenance of AF, ablation typically requires a much broader treatment area than that for regular tachycardias. Since its introduction in the 1980s, cardiac catheter ablation has become a standard treatment for many of these types of arrhythmias.

Procedure. The ablation and associated electrophysiology study are performed in a specially equipped laboratory. The patient is either anesthetized or awake with sedation to reduce discomfort and facilitate relaxation. Electrodes and sensors attached to the patient report blood pressure, blood oxygen saturation, and electrocardiogram (ECG). Next, the locations where multiple diagnostic catheters and the ablation catheter will be inserted (typically groin and/or neck) are locally anesthetized. The catheters are inserted into a blood vessel, and guided into the heart (see Fig. 15). The physician performs mapping as described below to determine the mechanism of the arrhythmia, and site of ablation. The primary imaging modality used for guidance is fluoroscopy, but other techniques are now used to assist in both guidance and mapping, as described below. The patient is typically discharged a few hours after the procedure, but may be monitored in the hospital overnight.

Devices

Radio Frequency Ablation. Figure 16 shows a typical rf ablation catheter tip with ablation electrode, and mapping electrodes for measurement of biopotentials from the heart. Most rf devices use temperature control, where power is controlled to keep the temperature measured by a thermocouple or thermistor embedded in the catheter tip constant (typically 60–80 °C). Typical application times are 45–60 s.

For treatment of atrial fibrillation, linear (i.e., elongated) thermal lesions are required. Linear lesions can be created by dragging or sequentially moving a standard catheter (Fig. 16), but some special multielectrode catheters can be

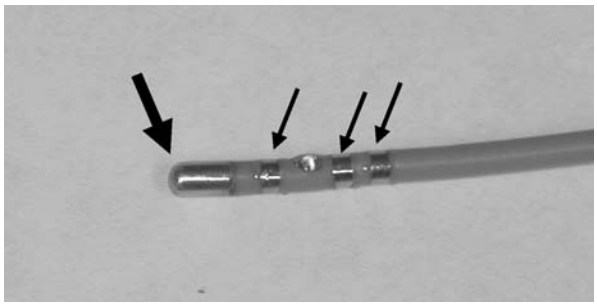


Figure 16. Cardiac rf ablation catheter (7F = 2.3 mm diameter) with ablation electrode (large arrow), and mapping electrodes (small arrows).

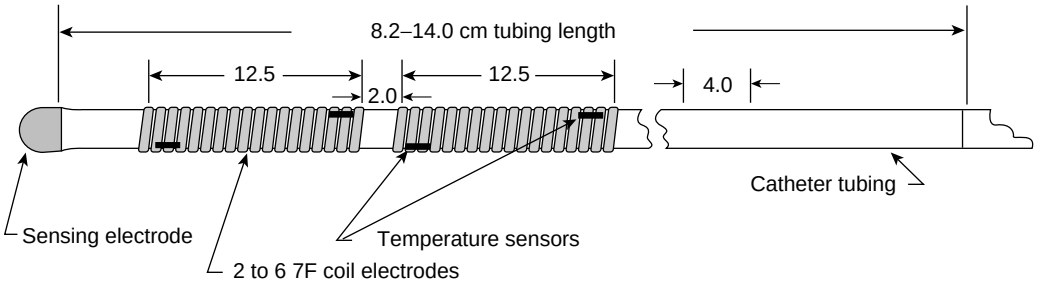


Figure 17. Cardiac multielectrode catheter for creation of elongated (linear) thermal lesions. Each electrode has two temperature sensors located near the edges for independent control. (From Ref. 23.)

more effective (Fig. 17). To keep all electrodes in the array at ideal temperatures, each electrode is controlled separately with multiple thermocouples placed at the edges of the electrodes.

Cryoablation. Cryoablation has the advantage of potential reversibility over heat based methods like rf ablation. For moderate hypothermic temperatures, tissue function can be restored after thawing. Thus, the correct location of a catheter can be confirmed by freezing at moderate temperatures (cryomapping). If the correct site is found (i.e., the arrhythmia stops or conduction is blocked), a longer freezing cycle with lower temperature is performed to destroy the tissue at that location. Once the tissue starts to freeze, the catheter sticks to the tissue (cryoadhesion), whereas with heat-based methods there is a risk of catheter movement. Other advantages include reduced complication rates compared to rf ablation. There should be minimal risk of tissue perforation, neighboring vessels like the coronary arteries seem to be preserved, and there is minimal risk of thrombus formation. Cryothermal lesion sizes are typically smaller than rf, and application times are longer ($\sim$4 min).

Microwave Ablation. Recently, microwave ablation catheters have become available for creating linear lesions (Microwave Ablation System Flex, Guidant, Indianapolis, IN), and are currently in clinical trials for treatment of atrial fibrillation. Microwave ablation has the potential benefits of directional heating, and deeper penetration compared to rf heating.

Cardiac Mapping and Imaging. Before ablation can be attempted, the treatment site in the heart has to be located, which is done by a procedure termed cardiac mapping. During the mapping procedure, local biopotentials known as electrograms are recorded typically from multiple catheters placed at different sites within the heart (Fig. 18). The ablation catheter itself has multiple mapping electrodes to record the electrograms (Fig. 16). This mapping procedure is performed under fluoroscopy to visualize the location of each catheter in the heart. These catheters can also be used for pacing the atria and/or the ventricles. From the temporal relationship between the electrograms, and from the relationship between these potentials and the ECG, the physician determines the mechanism of the arrhythmia. During the mapping procedure, the mapping–ablation catheter is moved to locate the exact site that causes the arrhythmia, where ablation is then performed.

Multiple Electrode Mapping. Mapping using the above procedure can be long and cumbersome. Mapping with multielectrode catheters can speed up the mapping procedure by acquiring simultaneous electrograms, and enabling pacing from multiple locations. Fig. 19 shows the Constellation catheter (Boston Scientific), a basket-type catheter that is expanded within the heart, and can pace or record signals from 64 electrodes.

Another multielectrode catheter, the Ensite Array (St. Jude Medical), also employs 64 electrodes, but by using an inverse solution it estimates $\sim$3000 signals on the interior (endocardial) surface of the heart chamber. It is deployed in without direct contact with the heart tissue (Figure 20). This has the advantage of not disturbing the beating heart and providing a highly detailed map, but is prone to errors since signals are extrapolated without direct tissue contact. Since fluoroscopy is not required during mapping, the radiation dose for the patient is highly reduced.

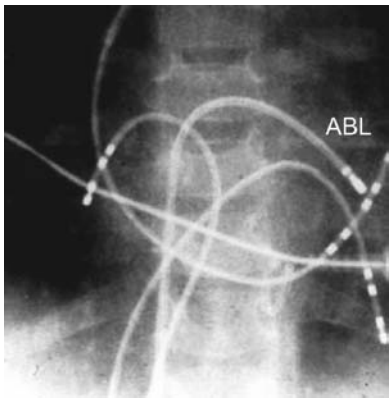


Figure 18. Mapping procedure using multiple catheters. This is an X-ray image of the heart showing the ablation cathater (ABL), and additional catheters for recording electrograms inside the heart. From the relationship between the electrograms at the different sites, the physician can determine the mechanism of the arrhythmia, and site of ablation. (Image provided courtesy of J. Phil Saul, Medical University of South Carolina.)

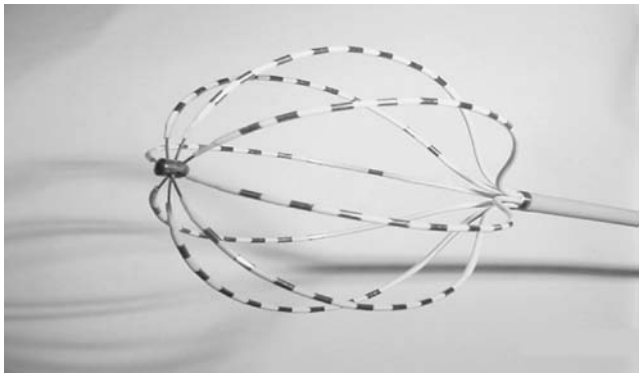


Figure 19. Constellation catheter has a total of 64 electrodes for cardiac mapping. The catheter is expanded inside the heart chamber. (From Ref. 24.)

Electroanatomic Mapping System. The CARTO system (Biosense-Webster) uses electromagnetic methods to determine the exact location of a specially designed catheter (25). A magnetic emitter employing three orthogonal coils is placed under the operating table, below the patient's chest. The reference catheter has a set of antennas located at the tip, which picks up the three magnetic fields. From the strength of the magnetic fields, the distance of the catheter from the magnetic emitter, the exact orientation of the tip (and subsequently the location in the heart) is determined. At each location, the reference catheter measures electrograms within the heart. By moving the catheter within the heart, a three-dimensional (3D) surface depicting the spread of activation through the heart (activation map) is created. From the activation map, the physician can determine the mechanism of the cardiac arrhythmia, and design a treatment plan. As with the Ensite Array, fluoroscopy dose can be markedly reduced with the CARTO system.

Figure 20. Ensite array catheter has 64 electrodes located on the surface of an expandable balloon catheter. The catheter records signals while floating inside the heart without contact. (Image provided courtesy of Endocardial Solutions.)

Impedance Mapping. The impedance signal between each catheter electrode and a set of reference electrodes can also be used to localize catheter positions. However, the accuracy of the location will depend somewhat on the variations in the impedance of the tissues through which the signal traverses. Two commercial systems, LocaLisa (Medtronic) and NAVEX (St. Jude Medical) use this technique to track catheter locations in real time, reducing mapping complexity and fluoroscopy time.

Tumor Ablation

After cardiac catheter ablation became clinically accepted, ablative methods were investigated for cancer treatment starting in the early 1990s. An estimated 70,000 clinical tumor ablation procedures had been carried out in 2004, with numbers still rising.

Clinical Background. The cancer type where tumor ablation was applied first is liver cancer. There was a need for a new treatment modality for liver cancer because surgery, the standard treatment, is only possible in ~15–20% of the cases, and chemotherapy does not work well for liver cancer.

As for many patients there was no viable treatment option, tumor ablation quickly became a standard treatment for unresectable (i.e., not treatable by surgery) liver cancer. More recently, tumor ablation techniques have been applied to other sites such as kidney, lung, and bone.

Tumor ablation can be performed during open surgery, using laparoscopy, or through a small incision in the skin (percutaneously). The treating physician is either a surgeon (open surgery, or laparoscopically) (26), or a radiologist (percutaneous approach) (27). Figure 21 shows a typical patient setup. The patient is either under conscious sedation, or light general anesthesia. The applicator is inserted under imaging guidance [typically ultrasound or computed tomography (CT)] into the tumor (in this case liver). Progress of ablation is monitored usually by ultrasound. Successful ablation is typically confirmed by CT. If

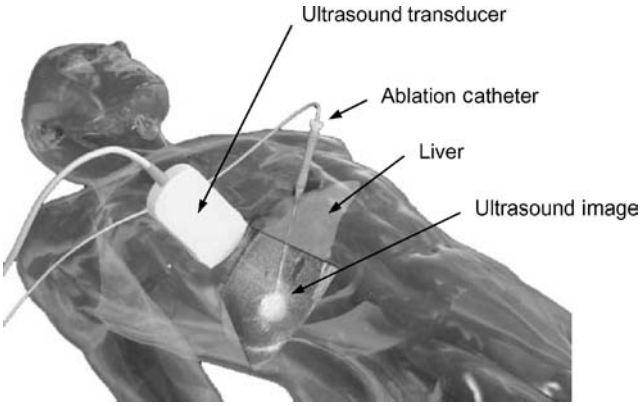


Figure 21. Patient setup for liver tumor ablation. The ablation catheter is inserted into the liver tumor through a small incision in the skin. Ultrasound imaging is used for guiding the electrode into the tumor, and monitoring the ablation. The white region in the overlaid ultrasound image represents areas of gas bubbles due to high temperatures. (From Ref. 28.)

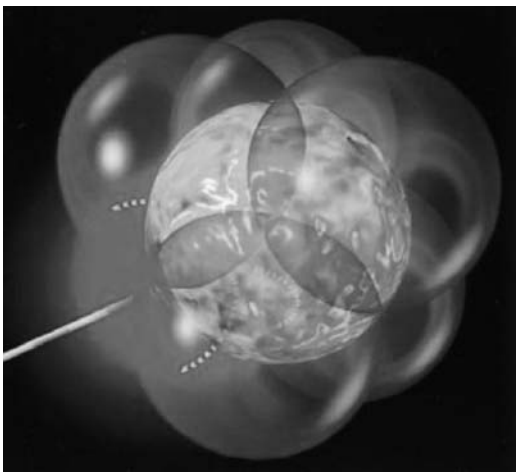


Figure 22. Multiple overlapping ablation zones are required for large tumors to ablate tumor and a 1 cm rim of surrounding tissue. (From Ref. 28.)

the procedure is performed minimally invasively, the patient can leave the hospital the same or the next day.

To successfully treat a tumor, the whole tumor including a 1 cm surrounding zone of normal tissue has to be ablated. It has been shown that the chance of recurrence (i.e., regrowth of the tumor due to incomplete treatment) increases when <1 cm of surrounding tissue is ablated. A single ablation can only treat a limited volume of tissue. Therefore, for larger tumors, multiple overlapping ablations have to be performed (see Fig. 22). This can be done with multiple applicators, but for devices that only support a single applicator (like most rf devices), multiple sequential ablations have to be performed.

Currently, long-term results are only available for liver cancer. For small (<3 cm) primary liver tumors (hepatocellular carcinoma), complete tumor necrosis is achieved in typically 80–90% of the cases, with 3 year survival rates of ~75%. Results are less favorable for metastatic cancer, where complete necrosis in 52–67% of small tumors is achieved, with 3 year survival of 40% (28). Treatment results are significantly worse with tumors >3 cm, where multiple sequential ablations are required.

Devices. Cryoablation was historically the first ablation method applied to cancer treatment. However, cryoablation has been so far mainly limited to use during open surgery. Bleeding from the insertion site could result in internal bleeding if cryoablation were done minimally invasively. The fist thermal ablation method that was applied minimally invasively through a small incision in the skin (percutaneously) was rf ablation (Chemical ablation was used even earlier). Today, rf ablation is the most widely used tumor ablation method, with competing device technologies, like MW ablation, emerging on the market.

Radio Frequency Ablation. There are currently three rf devices for tumor ablation commercially available on the U.S. market (29). They use applied power between 200 and 250 W, application times of 12–45 min, and create coagulation zones of 3–6 cm in diameter. All manufacturers employ different electrodes and power control algorithms. Figure 23 shows the different electrode types. Grounding pads (2–4) for rf current dissipation are typically located on the patient's thighs, equidistant from the rf electrode. One disadvantage of most current rf devices compared to other methods (MW, cryo, laser) is that only one electrode can be used at a time. This prolongs treatment of larger tumors (>3 cm diameter), with up to several hours procedural time. Below we compare the three most widely used systems.

RITA Medical. This system employs a tree shaped multiprong electrode (Fig. 23a). The prongs are stepwise expanded during the ablation procedure. Five of the prongs have thermocouples located at the prong tips that monitor tissue temperature. The rf power is regulated to keep tip temperatures ~100 °C. Different electrodes are available depending on tumor size. In one electrode design, saline is infused into the tissue during the ablation procedure. Local cooling and increase in thermal and electrical conductivity from the saline results in increased coagulation zones, up to 6 cm in diameter with 45 mins application time.

Boston Scientific. This system employs an umbrella shaped multiprong electrode (Fig. 23b). The prongs are completely expanded once the catheter is placed in the tumor. Power is controlled depending on impedance, which is measured between the electrode and the ground pad. If impedance exceeds a certain threshold resulting from tissue vaporization near the electrode, power is turned off for 15 s, and then applied at 70% of previous power level.

Figure 23. The rf electrodes of three manufacturers currently commercially available in the United States, and in clinical use. Multiprong electrodes by RITA medical (a), and Boston Scientific (b). Cooled needle electrode (single and triple cluster) by ValleyLab (c).

Valleylab. This system uses needle electrodes, which are cooled internally by circulating chilled water. Two types of electrodes are available (Fig. 23c): a single needle electrode and a triple-cluster electrode that achieve 3 and 4 cm coagulation zone diameters, respectively. This system employs impedance control, similarly to the Boston Scientific system. Treatment time is 12 min. Another recently introduced system originally developed at the University of Wisconsin, allows the use of up to three electrodes simultaneously, allowing rapid treatment of large tumors (30).

Cryoablation. Different cryoablation systems that employ either argon or liquid nitrogen cooling are clinically available. Cryoprobes in different sizes are available, though the general shape is often similar (Fig. 10). Cryoablation is mainly carried out during open surgery, though recently, small cryoprobes (17 gauge = 1.2 mm diameter) for minimally invasive treatment have become commercially available. Cryoablation has the advantage that the iceball is visible under ultrasound imaging, allowing real-time monitoring of the procedure (20).

Microwave Ablation. Microwave (MW) ablation has been used for small tumors with generic antennas in the Asian region for several years. Only recently has a MW ablation device for ablation of larger tumors become commercially available in the US (Vivant Medical, Mountain View, CA). Potential advantages over rf ablation are shorter treatment times and higher tissue temperatures in addition the possibility of use of multiple antennas takes advantage of constructive interference. Preliminary studies have indicated that MW ablation may show superior performance close to large vessels compared to rf ablation.

Imaging. Imaging serves different purposes during ablative treatment. Initially, the presence of a tumor is identified, typically by CT or MRI (though frequently additional tumors are found during interoperative ultrasound imaging, requiring change of the treatment plan). Figure 24 shows a contrast-enhance CT image of a patient with a liver tumor. Placement of the ablation applicator is typically guided by ultrasound imaging, and often confirmed by CT. Figure 25a shows an ultrasound image of a tumor with inserted rf electrode. Real-time monitoring of the ablation procedure is done by ultrasound imaging. Figure 25b shows the bright (hyperechoic) region resulting from microbubbles due to tissue heating during rf ablation; however, this bright area does not correspond to the zone of tissue destruction. Figure 12 shows the dark region (hypoechoic) depicting the boundary between frozen and unfrozen tissue during cryoablation. This boundary corresponds well with the boundary of cell death (20), which is one of the advantages over heat-based methods. After the ablation procedure, destruction of tumor and sufficient margin of normal tissue is confirmed by another CT scan or MRI.

Figure 24. Tumor (*arrow*) can be identified from this contrast-enhanced CT image. Tumors are visible because they are typically hypervascular (i.e., have higher density of blood vessels than normal), and take up more of the contrast agent. (Image provided courtesy of Bradford J. Wood, NIH.)

Contrast agents for ultrasound imaging have recently become available in Europe (31), and should be available in the United States within the next years. These contrast agents employ microbubbles that are visible under ultrasound to visualize vasculature. Tumors are typically hypervascular, and show up using these contrast agents. As the tissue coagulates during ablation and blood perfusion stops, the contrast agent cannot penetrate the

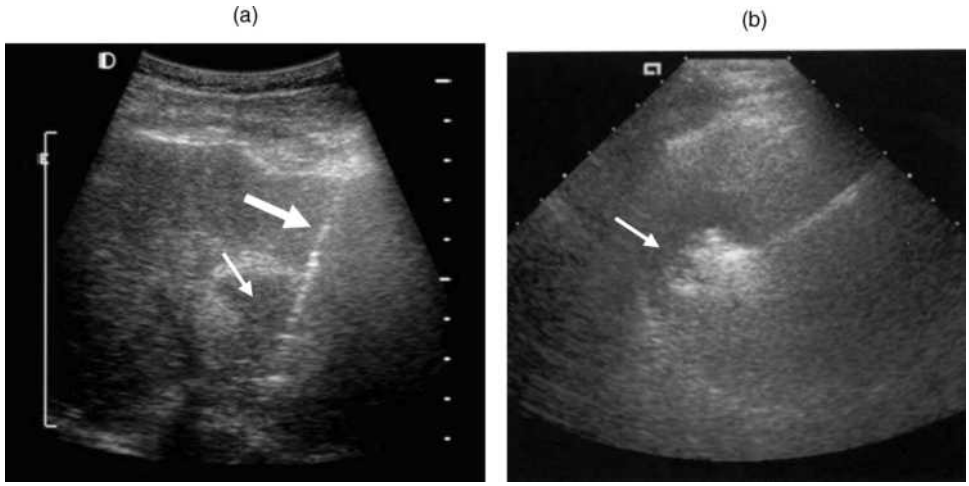


Figure 25. Ultrasound imaging is typically used to guide and monitor the ablation procedure. (a) Shows the rf needle electrode (large arrow) inserted into the tumor (small arrow). (b) Shows bright area (arrow) due to formation of microbubbles as a result of tissue heating. (Images provided courtesy of Bradford J. Wood, NIH.)

coagulated areas anymore. Ultrasound imaging using contrast agents allows real-time monitoring of the ablation procedure contrary to CT and MRI, which can typically only be performed after the procedure.

Future Directions. A continuing trend in tumor ablation is toward larger coagulation zones and shorter treatment times. High tumor recurrence rates close to large vessels are still problematic, with improved applicators attempting to address this issue. Adjuvant therapies (e.g., radiation, chemotherapy agents) can kill cells at lower elevated temperatures (above $\sim$45 °C) than heat alone (above $\sim$50 °C). Thereby, the zone of cell death may be extended (cf. Fig. 13), which may be beneficial close to large vessels where it is difficult to achieve sufficient heating.

Unavailability of adequate monitoring of the extent of the coagulation zone is a major problem for all heat-based methods, especially when damage to nearby structures is to be avoided. This problem is partially addressed by ultrasound contrast agents as discussed above. Other potential imaging solutions include thermometry, using MRI or ultrasound imaging, and ultrasound elastography, which allows imaging of the coagulation zone due to change in elasticity after coagulation. Some companies are working on computer-assisted treatment planning systems that guide the physician in optimal applicator placement, and estimate coagulation zones.

Endometrial Ablation

Clinical Background. Ablation has become a standard procedure for women with dysfunctional uterine bleeding, with $\sim$30,000 annual procedures. Endometrial Ablation (endometrium = lining of the uterus) is indicated for patients that do not respond to standard treatments like drugs and curettage (scraping tissue from the endometrium with a spoon-shaped instrument). Initially either laser or rf energy applied by a rollerball electrode was used; the physician had to manually direct the catheter to ablate the whole *endometrium*, which was time consuming. A number of newer U.S. Food and Drug Administration (FDA) approved devices that are more effective with shorter treatment times are now available (32). These second generation devices treat the whole lining at once, and provide success rates of 67–80%.

Devices. There are currently five second generation devices available on the U.S. market. Two devices use heated fluid to ablate the endometrium. In one device, heated saline (80–90 °C) is circulated within the uterine cavity for 8 min (Hydro Thermablator, Boston Scientific). Another device employs a balloon catheter, inside which heated Dextrose solution (5%) at 87 °C is circulated for 10 min (Thermachoice UBT, Gynecare). One device uses cryoablation with a freeze–thaw cycle (10 min freezing) at two locations in the uterine cavity (Her Choice, AMS). The shortest treatment time is achieved by a device that uses bipolar rf ablation employing mesh electrodes (Novasure, Cytyc Corp.; see Fig. 26). The rf energy is applied for 90–120 s, and controlled by tissue impedance. The last device uses microwave ablation at 9.2 GHz, and 42 W power for 3 min (MEA, Microsulis).

Figure 26. NovaSure catheter used for endometrial ablation. This catheter has an electrode consisting of multiple electrically isolated meshes, with rf energy applied bipolar between different meshes. Image provided courtesy of Cytyc Corporation and affiliates.

Prostate

Clinical Background. Two types of prostate diseases are prevalent in men. Prostate enlargement (Benign prostate hyperplasia) is common in men >50 years of age. The enlarged prostate imparts pressure on the urethra resulting in restriction, at which point it has to be treated to reduce prostate size. Cancer is the second common disease affecting the prostate, and is the most common form of cancer affecting men.

Current standard treatment options for enlarged prostate include medication, surgery where part of the prostate is removed, and minimally invasive ablative methods. For prostate cancer, treatment options are surgical removal (radical prostatectomy) and external radiotheraphy. Ablative methods are currently only used in cases when conventional treatment fails, or in cases with advanced disease to reduce tumor volume. Ablative treatment of prostate cancer requires more precise control of the ablation zone than for enlarged prostate, since the cancer is usually located in the periphery of the prostate (away from the urethra) making treatment from inside the urethra more difficult (33).

Commercial devices are available that employ cryo, rf, microwave, and ultrasound. Depending on the device, the prostate is treated by inserting the ablation catheter into the urethra, rectum, or through the skin in the region between the scrotum and anus (perineum). Urethra and rectum have to be protected from damage by cooling when heat-based devices are used, and by heating when cryoablation is used.

Devices for Enlarged Prostate Treatment

Microwave Ablation. Several devices are available that use microwave antennas inserted through the urethra.

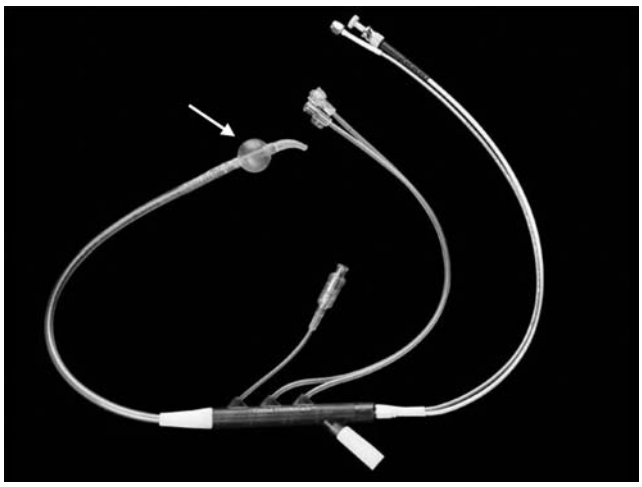


Figure 27. Targis catheter for treatment of enlarged prostate. The different ports are used for inflating the balloon (arrow) inside the bladder, and for perfusing the catheter with cooling water. (Image provided courtesy of Urologix.)

Two devices (Targis and Prostatron, Urologix) use cooled antennas to avoid damage to the urethra. Figure 27 shows the Targis device. The balloon is inflated inside the bladder to facilitate proper positioning. The Thermatrx device (AMS) uses low energy and does not required cooling. The CoreTherm device (ACMI) employs temperature sensors to monitor tissue damage and control applied power. Treatment time is typically 30–60 min, and tissue temperatures are in the range of 45–50 °C.

Radio Frequency Ablation. The transurethral needle ablation device (TUNA, Medtronic) uses a catheter, from which two needle electrodes project from the urethra into the prostate. Maximum power of 30 W is applied for 4 min, so that tissue temperature as measured by sensors located within the needles tips reaches 100 °C. Since each needle ablates only a small tissue volume, multiple insertions are required.

Devices for Prostate Cancer Treatment

Cryoablation. While the rf and MW devices described above are used for treatment of enlarged prostate, cryoablation is used for treatment of prostate cancer. The probes are typically inserted into the prostate through the skin (percutaneously) in the region between the scrotum and anus (perineum). The diameters of current probes are between 1.22 (SeedNet, Galil Medical) and 3 mm (Cryocare CS, Endocare) for the two systems currently available in the United States. A warming catheter perfused with 40–42 °C saline is placed in the urethra to protect it from damage.

Ultrasound Ablation. Two devices that employ high intensity focused ultrasound are commercially available in Europe and Asia, though not yet FDA approved in the United States. Both systems (Ablatherm, EDAP; Sonablate, Focus Surgery) use catheters inserted into the rectum, and catheters have both imaging and treatment transducers. When the treatment transducer is activated, it creates a single cylindrical thermal lesion ∼20 mm long and 1–2 mm in diameter. The focal point is gradually moved according to a treatment plan to ablate the treatment region.

Radio Frequency Ablation. The same devices described under "Devices section for Tumor Ablation" are also under investigation for treatment of prostate cancer.

Endovascular Ablation

Clinical Background. Varicose veins are visibly dilated and twisted veins near the skin surface, most often affecting legs and thighs. Insufficiencies of the venous valves result in blood pooling and enlargement of the veins. Varicose veins affect ∼10% of the population, mostly between ages 30 and 60.

Several treatment options are available, all aiming to close the affected veins. Veins can be treated using surgical stripping, sclerotherapy (injection of an agent that causes vein swelling, and closure), and ablation. During treatment, a catheter is introduced into the vein and the vessel wall is heated. The heat results in shrinkage of the collagen in the wall, eventually closing off the vessel.

Devices

Radio Frequency Ablation. A bipolar, multielectrode device (ClosurePlus, VNUS Medical) is commercially available (see Fig. 28) in two sizes (2 and 2.7 mm diameter). The catheter is inserted into a vein and advanced to the treatment location. The rf energy is applied between two sets of electrodes (2.7 mm catheter), or between outer electrodes and inner ball electrode (2 mm catheter). Power is controlled so that vein wall temperature, as measured by a thermocouple (located in the outer electrodes tips), reaches 85 °C. Heparinized saline is infused through the central lumen of the catheter to prevent blood coagulation. Once the target temperature is reached, the operator moves the catheter at a rate of 2–3 cm per minute while keeping

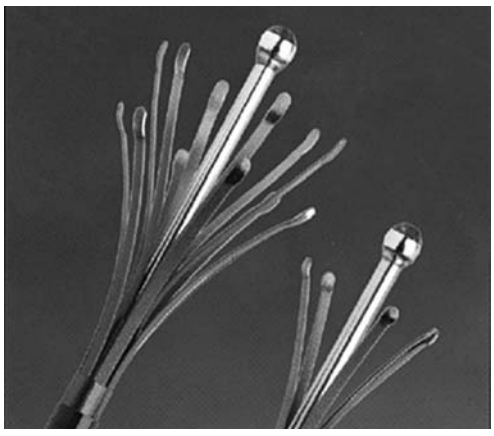


Figure 28. ClosurePLUS endovascular rf catheters for treatment of varicose veins in two sizes. The catheter is introduced with retracted prongs, and expanded at the treatment site. The rf current is passed between the center electrode, and surrounding prongs (left, smaller 2 mm catheter), or between two sets of prongs (right, larger 2.7 mm catheter), while catheter is pulled along the vein to shrink and close of a section of the vein. (Image provided courtesy of VNUS Medical Technologies, Inc.)

Table 4. List of Manufacturers of Ablation Devices

Application	Company	Device Name (type)
Cardiac	Biosense-Webster (Diamond Bar, CA) www.biosensewebster.com Tel. 800-729-9010	Stockert 70 (rf ablation) Carto XP (cardiac mapping system)
	Boston Scientific (Natick, MA) www.bsci.com Tel. 888-272-1001	Cobra (rf ablation) Chilli (rf ablation) EPT-1000XP (rf ablation) RPM (cardiac mapping system)
	Cardima (Fremont, CA) www.cardima.com Tel. 800-354-0102	Revelation (rf ablation)
	Cryocath (Montreal, Canada) www.cryocath.com Tel. 877-694-1212	Freezor (Cryoablation) SurgiFrost (Cryoablation)-FrostByte (Cryoablation)
	Cryocor (San Diego, CA) www.cryocor.com Tel. 858-909-2200	CryoBlator (Cryoablation)
	Medtronic (Minneapolis, MN) www.medtronic.com Tel. 763-514-4000	Atakr (rf ablation) LocaLisa (cardiac mapping system)
	St. Jude Medical (St. Paul, MN) www.sjm.com Tel. 800-328-9634	Livewire (rf ablation) Ensite Array (cardiac mapping system) NAVEX (cardiac mapping system)
Tumor	Boston Scientific (Natick, MA) www.bsci.com Tel. 888-272-1001	RF 3000 (rf ablation)
	Celon (Teltow, Germany) www.celon.com Tel. + 49-3328-3519-0	CelonSurgical (Bipolar rf ablation)
	Endocare (Irvine, CA) www.endocare.com Tel. 800-418-4677	Accuprobe (Cryoablation)
	Galil Medical (Westbury, NY) www.galilmedical.com Tel. 516-794-7020	Seednet (Cryoablation) CryoHit (Cryoablation)
	Rita Medical (Mountain View, CA) www.ritamedical.com Tel. 650-314-3400	Model 1500X (rf ablation)
	Valleylab (Boulder, CO) www.valleylab.com Tel. 800-255-8522	Cool-Tip (rf ablation)
	Vivant (Mountain View, CA) www.vivantmedical.com Tel. 650-694-2900	VivaWave (Microwave ablation)
Endometrial	AMS (Minnetonka, MN) www.americanmedicalsystems.com Tel: 800-328-3881	Her Choice (Cryoablation)
	Boston Scientific (Natick, MA) www.bsci.com Tel. 888-272-1001	Hydro Thermablator (heated saline)
	Cytyc (Marlborough, MA) www.cytyc.com Tel. 800-442-9892	Novasure (Bipolar rf ablation)
	Gynecare (Somerville, NJ) www.gynecare.com Tel. 888-496-2273	ThermaChoice UBT (Balloon ablation, heated fluid)
	Microsulis Medical (Hants, UK) www.microsulis.co.uk Tel. +44-23-9224-0011	MEA (Microwave ablation)
Prostate	ACMI (Southborough, MA) www.acmicorp.com Tel. 888-524-7266	CoreTherm (Microwave ablation)
	AMS (Minnetonka, MN) www.americanmedicalsystems.com Tel: 800-328-3881	Thermatrx (Microwave ablation)
	EDAP TMS (Lyon, France) www.edap-tms.com Tel. +33-472-1531-50	Ablatherm (Focused ultrasound)
	Endocare (Irvine, CA) www.endocare.com Tel. 800-418-4677	Cryocare CS (Cryoablation)
	Ethicon Endo-Surgery (Cincinnati, OH) www.ethiconendo.com 800-873-3636	Indigo Laser System (Laser ablation)
	Focus Surgery (Indianapolis, IN) www.focus-surgery.com Tel. 317-541-1580	Sonablate (Focused ultrasound)
	Galil Medical (Westbury, NY) www.galilmedical.com Tel. 516-794-7020	SeedNet (Cryoablation)
	Medtronic (Minneapolis, MN) www.medtronic.com Tel. 800-328-2518	TUNA (rf ablation)
	Urologix (Minneapolis, MN) www.urologix.com Tel. 800-475-1403	Targis and Prostatron Systems (Microwave ablation)
Endovascular	Diomed (Andover, MA) www.diomedinc.com Tel. 987-475-7771	EVLT (Laser ablation)
	VNUS Medical Technologies (Sam dose, CA) www.vnus.com Tel. 888-797-8346	ClosurePlus (Bipolar rf ablation)

Table 4. (*Continued*)

Application	Company	Device Name (type)
Cornea	Alcon Laboratories (Fort Worth, TX) www.alconlabs.com Tel. 800-757-9195	LADARVision (Laser ablation)
	Bausch & Lomb (Rochester, NY) www.bausch.com Tel. 800-553-5340	Technolas 217A (Laser ablation)
	Nidek (Fremont, CA) www.nidek.com Tel. 510-226-5700	NAVEX (Laser ablation)
	Refractec (Irvine, CA) www.refractec.com Tel. 800-752-9544	Viewpoint CK (rf ablation) NearVision CK (rf ablation)
	VISX (Santa Clara, CA) www.visx.com Tel. 408-773-7321	STAR S4 (Laser ablation)
Intervertebral	Smith & Nephew (Andover, MA) endo.smith-nephew.com Tel. 800-343-5717	SpineCath (rf ablation)
	Valleylab (Boulder, CO) www.valleylab.com Tel. 800-255-8522	discTRODE (rf ablation)

[a]Sorted by application type and alphabetically.

the temperature within 3 °C of the 85 °C target. Treatment success is confirmed via ultrasound imaging, where no flow should be present.

Laser Ablation. The second commercially available device (EVLT, Diomed) uses laser to produce heating, shrinkage of the vein. A laser fiber is inserted into a vein and advanced to the treatment location. Laser light of 810 nm wavelength at 14 W power is applied, while pressure on the vein ensures contact between fiber and vein. The fiber is advanced at a rate of 2–3 mm per second.

A study comparing the ClosurePlus and EVLT devices showed much higher temperatures when the laser device was used, resulting in vein perforations and reduced performance compared to the rf device (34).

Cornea Ablation

Clinical Background. Many vision disorders are a result of imperfections in the shape of the cornea. Conditions treated include astigmatism (cornea has oblong shape), myopia (nearsightedness), hyperopia (farsightedness), and presbyotia (blurred vision at close range, due to age-related loss of elasticity).

Treatment options are surgical treatment (part of the cornea is surgically removed), and ablative methods.

Devices

Laser Ablation (Laser Refractive Surgery). Laser is the most widely used ablative treatment method for corneal ablation. For most applications the excimer laser is used, which is an Argon laser operating in the ultraviolet (UV) range (193 nm wavelength); the laser is used in pulsed mode with 10–60 pulses (each a few ns) per second. The excimer laser does not cause tissue damage as described in the "Tissue injury from Heating" section, but causes the corneal tissue to vaporize. Part of the cornea is thereby removed, resulting in the desired change of shape.

In some applications, the Holmium:YAG laser is used, which operates in the IR region (2100 nm wavelength). This laser causes heating, and shrinkage of the collagen in the cornea, resulting in change of shape.

Radio Frequency Ablation (Conductive Keratoplasty). More recently, rf devices have become available for the treatment of hyperopia and presbyotia. RF energy is applied by fine electrode (90 μm diameter, 450 μm long) in pulsed fashion (exponentially damped rf pulses, pulse rate ∼8 kHz), with power levels of ∼1 W, and application times in second range. The rf heating results in collagen shrinkage, and change of cornea shape.

Intervertebral Disk Ablation

Degenerative diseases of the intervertebral disks are a major cause of lower back pain. Ablative techniques have been introduced in the late 1990s, and are now used in certain patient populations (35). Even though exact mechanisms of pain alleviation are not known, it is assumed that two mechanisms are responsible: shrinkage of the disk reduces pressure on nerve fibers, and destruction of sensitive nerve fibers. Currently, two devices are commercially available, both of which use rf ablation to heat tissue.

Other Applications

Ablative techniques are investigated for a number of other applications. Treatment of different types of cancer other than the ones discussed above such as in the breast and esophagus, are being investigated. In the brain, treatment of deep-seated tumors and other disorders, like Parkinson's disease and Epilepsy, is examined. Ablation is investigated for treatment of chronic pain by ablating responsible nerve fibers. In dentistry, laser ablation is investigated as a potential replacement for mechanical drills.

DEVICE MANUFACTURERS

See Table 4.

BIBLIOGRAPHY

1. Pennes HH. Analysis of tissue and arterial blood temperatures in the resting human forearm. J Appl Physiol 1948;1:93–122.

2. Lai YC, et al. Lesion size estimator of cardiac radiofrequency ablation at different common locations with different tip temperatures. IEEE Trans Biomed Eng 2004;51:1859–1864.
3. Wright AS, Lee FT Jr, Mahvi DM. Hepatic microwave ablation with multiple antennae results in synergistically larger zones of coagulation necrosis. Ann Surg Oncol 2003;10:275–283.
4. Labonte S, et al. Monopole antennas for microwave catheter ablation. IEEE Trans Microw Theory 1996;44:1832–1840.
5. Saito K, et al. Heating characteristics of array applicator composed of two coaxial-slot antennas for microwave coagulation therapy. IEEE Trans Microw Theory 2000;48:1800–1806.
6. Camart JC, et al. 915 MHz microwave interstitial hyperthermia. Part ii: Array of phase-monitored antennas. Int J Hyperthermia 1993;9:445–454.
7. Trembly BS, et al. Effect of phase modulation on the temperature distribution of a microwave hyperthermia antenna array in vivo. Int J Hyperthermia 1995;10:691–705.
8. Hynynen K, McDannold N. Mri guided and monitored focused ultrasound thermal ablation methods: A review of progress. Int J Hyperthermia 2004;20:725–737.
9. Nau WH, Diederich CJ, Burdette EC. Evaluation of multielement catheter-cooled interstitial ultrasound applicators for high-temperature thermal therapy. Med Phys 2001;28:1525–1534.
10. Diederich CJ, et al. Catheter-based ultrasound applicators for selective thermal ablation: Progress towards mri-guided applications in prostate. Int J Hyperthermia 2004;20:739–756.
11. Vogl TJ, et al. Mr-guided laser-induced thermotherapy (litt) of liver tumours: Experimental and clinical data. Int J Hyperthermia 2004;20:713–724.
12. Gage AA. History of cryosurgery. Semin Surg Oncol 1998;14: 99–109.
13. Webster JG, editor. Minimally Invasive Medical Technology. Bristol, UK: IOP Publishing; 2001.
14. Lee FT Jr, Mahvi DM, Chosy SG, Onik GM, Wong WS, Littrup PJ, Scanlan KA. Hepatic cryosurgery with intraoperative us guidance. Radiology 1997;202:624–632.
15. Miller MW, Ziskin MC. Biological consequences of hyperthermia. Ultrasound Med Biol 1989;15:707–722.
16. Dewhirst MW, et al. Basic principles of thermal dosimetry and thermal thresholds for tissue damage from hyperthermia. Int J Hyperthermia 2003;19:267–294.
17. Sapareto SA, Dewey WC. Thermal dose determination in cancer therapy. Int J Radiat Oncol Biol Phys 1984;10:787–800.
18. Bischof JC, et al. Cryosurgery of dunning at-1 rat prostate tumor: Thermal, biophysical, and viability response at the cellular and tissue level. Cryobiology 1997;34:42–69.
19. Weber SM, et al. Hepatic cryoablation: US monitoring of extent of necrosis in normal pig liver. Radiology 1998;207:73–77.
20. Zipes DP, Haissaguerre M. Catheter ablation of arrhythmias. 2nd ed. Armonk, NY: Futura Publishing; 2001.
21. Panescu D, Whayne JG, Fleischman SD, Mirotznik MS, Swanson DK, Webster JG. Three-dimensional finite element analysis of current density and temperature distributions during radio-frequency ablation. IEEE Trans Biomed Eng 1995;42:879–890.
22. Panescu D, Fleischman SD, Whayne JG, Swanson DK, Mirotznik MS, McRury I, Haines DE. Radiofrequency multielectrode catheter ablation in the atrium. Phys Med Biol 1999;44:899–915.
23. Panescu D. Intraventricular electrogram mapping and radiofrequency cardiac ablation for ventricular tachycardia. Physiol Meas 1997;18:1–38.
24. Shpun S, et al. Guidance of radiofrequency endocardial ablation with real-time three-dimensional magnetic navigation system. Circulation 1997;96:2016–2021.
25. Poon RT, et al. Locoregional therapies for hepatocellular carcinoma: A critical review from the surgeon's perspective. Ann Surg 2002;235:466–486.
26. McGahan JP, Dodd GD. Radiofrequency ablation of the liver: Current status. Am J Roentgenol 2001;176:3–16.
27. Dodd GD, et al. Minimally invasive treatment of malignant hepatic tumors: At the threshold of a major breakthrough. Radiographics 2000;20:9–27.
28. Pereira PL, et al. Radiofrequency ablation: In vivo comparison of four commercially available devices in pig livers. Radiology 2004;232:482–490.
29. Haemmerich D, et al. Large-volume radiofrequency ablation of ex vivo bovine liver with multiple cooled cluster electrodes. Radiology 2005;234:563–568.
30. Solbiati L, et al. Guidance and monitoring of radiofrequency liver tumor ablation with contrast-enhanced ultrasound. Eur J Radiol 2004;51 (Suppl): S19–23.
31. Cooper J, Gimpelson RJ. Summary of safety and effectiveness data from fda: A valuable source of information on the performance of global endometrial ablation devices. J Reprod Med 2004;49:267–273.
32. Shinohara K. Thermal ablation of prostate diseases: Advantages and limitations. Int J Hyperthermia 2004;20:679–697.
33. Weiss RA. Comparison of endovenous radiofrequency versus 810 nm diode laser occlusion of large veins in an animal model. Dermatol Surg 2002;28:56–61.
34. Bass EC, et al. Heat-induced changes in porcine annulus fibrosus biomechanics. J Biomech 2004;37:233–240.

See also CRYOSURGERY; ELECTROSURGICAL UNIT (ESU); STEREOTACTIC SURGERY.

TISSUE ENGINEERING

KRISTYN S. MASTERS
WILLIAM L. MURPHY
University of Wisconsin
Madison, Wisconsin

INTRODUCTION (1–4)

Definition

Tissue engineering represents a unique convergence of work from the worlds of clinical medicine, engineering, and basic science. The most commonly cited definition of tissue engineering originates from an influential 1993 paper by Langer and Vacanti (1): "Tissue engineering is an interdisciplinary field that applies the principles of engineering and the life sciences toward the development of biological substitutes that restore, maintain, or improve tissue function."

Although the term tissue engineering had existed for several years prior to this 1993 publication, and the concept of tissue engineering had existed for several decades, it is Langer and Vacanti's paper that is ultimately credited with stimulating broad awareness and acceptance of this description. This definition also served to unify seemingly diverse lines of research, as it encompasses three general strategies for the creation of new tissue, namely, the use of isolated cells or cell substitutes; tissue-inducing substances; or cells placed on or within matrices.

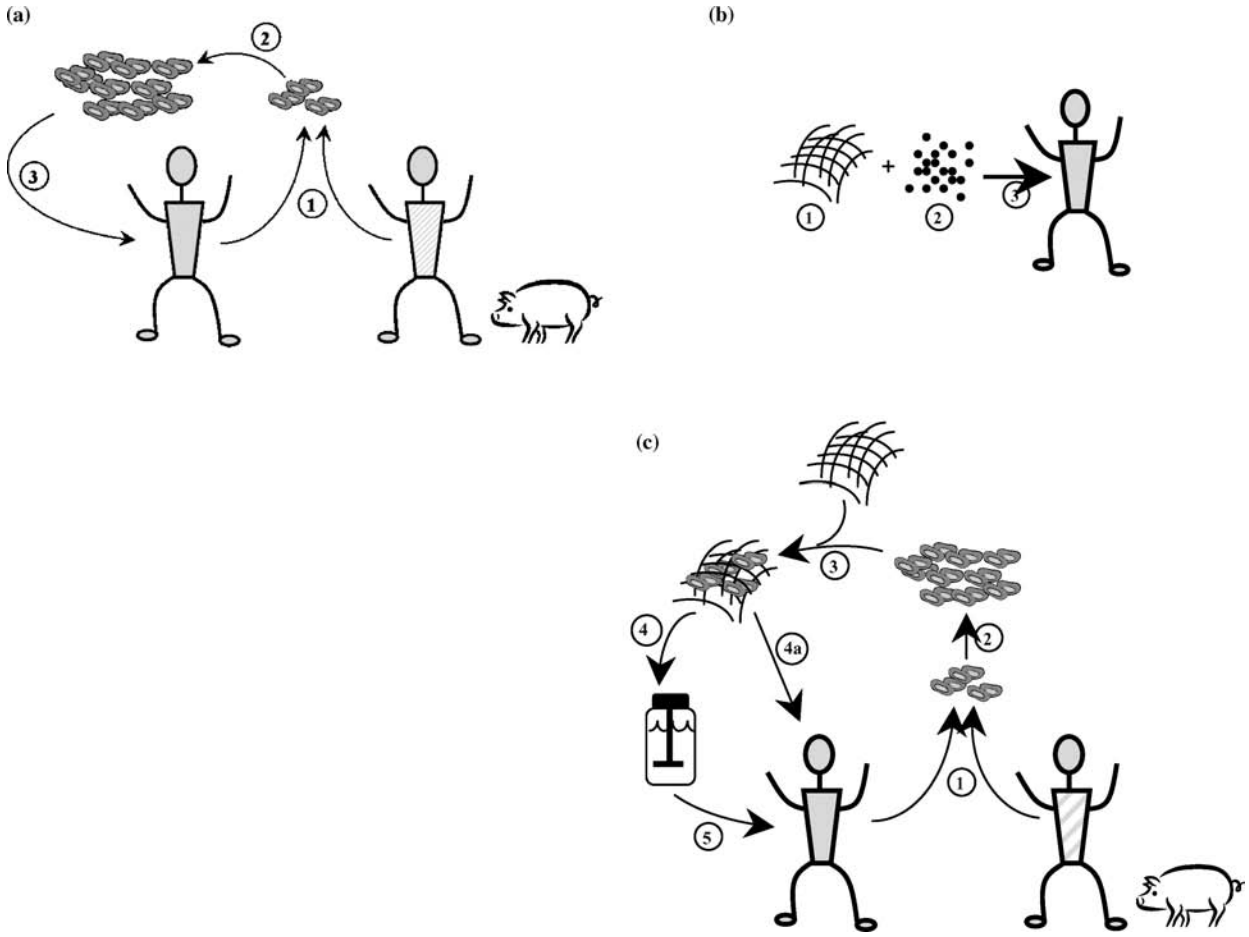


Figure 1. Depiction of three main tissue engineering strategies. In **(a)** cell transplantation, cells are obtained from a donor (*1*), who is either the patient themselves, another person, or an animal, then expanded *in vitro* (*2*), and reimplanted into the diseased/injured site (*3*), Regeneration via scaffold alone is represented by **(b)**, where a degradable, and possibly bioactive scaffold (*1*) may be combined with inductive agents (*2*), and then implanted (*3*). Lastly, **(c)** depicts the implantation of cell-seeded scaffolds, as cells are isolated from a donor (*1*), expanded *in vitro* (*2*), and combined with a biomaterial scaffold (*3*), followed by either immediate implantation (*4a*) or further culture in an *in vitro* bioreactor (*4b*) prior to implantation (*5*).

Thus, tissue engineering includes a wide variety of therapies, including cell transplantation, the implantation of biomaterials alone as scaffolds for *in situ* cell growth, and the *in vitro* or *in vivo* development of cell-seeded materials; these strategies are depicted in Fig. 1. The choice of tissue engineering approach is dependent on the specific type of tissue to be repaired; factors, such as cell availability, tissue regeneration potential, and mechanical stresses in the tissue environment, are all important considerations in the selection of a tissue engineering strategy. Ultimately, the goal of tissue engineering is the formation of neotissues that are mechanically and biologically integrated in the patient's body.

Motivation

Therapeutic options for patients with diseased or dysfunctional tissues or organs are currently limited. While organ transplantation has become a successful means of replacing diseased organs, this option continues to be seriously limited by problems with immune rejection and organ availability. Historically, the number of organs available for transplantation has been exceeded by the number of patients requiring an organ transplant, and this gap continues to widen. The number of needed organs continues to grow at an annual rate that is over twice the rate of increase of donated organs. Over 4000 people die each year while on the UNOS (United Network for Organ Sharing) transplantation waitlist, while 100,000 others die from organ failure without even qualifying for the waitlist.

Certain structural elements of the body (e.g., bone, blood vessels, and skin) may also be replaced via autologous tissue grafts. Such grafting procedures transfer the patient's own healthy tissue to a location that requires assistance with repair, thus circumventing immunogenicity problems associated with receiving tissues from other sources. However, tissue grafting is also accompanied by significant drawbacks and challenges, including limited availability of healthy donor tissue, donor site morbidity, and the need for multiple surgical sites. Grafted or reconstructed tissues are often

functionally inferior and less durable than the natural organs that they replace. Moreover, tissue grafting is appropriate for only select types of tissue.

Lastly, artificial prostheses and permanent implants have been developed to take the place of diseased or defective tissues, particularly in the areas of orthopedics and cardiovascular medicine. While these implants are capable of improving both the patient's lifespan and quality of life, tissue replacement with a permanent, artificial structure results in the loss of that tissue's natural biological functions. Most notably, artificial prostheses are unable to grow or remodel, meaning that they are often unsuitable for pediatric patients and are incapable of responding to changes in the body's needs or environment. Furthermore, significant issues arise from the permanent implantation of synthetic structures, and the types of tissues that can be mimicked by current prosthesis technology are very limited.

The field of tissue engineering holds the potential to overcome the aforementioned challenges associated with organ and tissue availability, immunogenicity, and retention of tissue function. Tissue engineering promises to provide a means of regenerating or replacing diseased or dysfunctional tissues and organs while leaving no permanent implant. While organ regeneration may be the most explicable motivation for the creation of engineered tissues, the objectives of tissue engineering as a field can be quite broad, as tissue engineering can be used to offer an alternative to drug therapy, gene therapy, and whole-organ transplantation; inspire or control the normal processes of tissue repair and healing; replace cells that are missing within an otherwise functional tissue or organ; use cellular control mechanisms to enhance drug delivery; lead to new models of human physiology.

The principal variables in the process of tissue engineering are (*1*) cell source; (*2*) scaffold type and properties; (*3*) method of tissue development (bioreactor design); (*4*) inclusion of inductive factors (i.e., cytokine delivery, gene therapy). This article will introduce and provide background on these key components of tissue engineering, and then discuss how tissue engineers have manipulated these variables in the development of two different types of engineered tissue: bone and cardiovascular tissue.

History

The National Science Foundation (NSF) (2) defines 1987 as the year that tissue engineering became formalized as a field; it is in this year that the NSF became consciously involved in tissue engineering. Although a defined label did not exist to describe their work, several pioneering researchers did perform tissue engineering research in the decades preceding the field's "official" emergence. In the vascular area, the concept of a resorbable vascular graft was introduced in the 1960s, with the first fully resorbable graft reported in 1979. In 1978, researchers observed improved healing after seeding synthetic vascular grafts with endothelial cells, and the first attempt to tissue engineer a vascular structure *in vitro* using collagen and cultured vascular cells was described by Burkel et al. in 1982 (see Ref. 3 for details). The origins of dermal tissue engineering reach into the 1950s, when Billingham and Reynolds demonstrated that cultured epidermal cells could be applied to a graft bed to reconstitute an epidermis. By the 1970s, the ability to culture cells *in vitro* had significantly advanced, enabling the formation of multilayered epidermal sheets that could be transferred intact to a wound bed. The first living skin equivalent (LSE) was created using fibroblasts seeded on a collagen matrix and described in 1981 in a notable paper by Bell et al. (see Ref. 2 for details). In the area of orthopedics, the development of nondegradable biomaterial implants and the discovery of osteo- and chondrogenic cytokines and growth factors constituted the majority of pre-1987 tissue engineering work. Lastly, early (and current) tissue engineering of organs (e.g., the kidney, pancreas, and liver) was (and continues to be) hindered by the functional complexity of these structures and the difficulty of cellular expansion *in vitro*. Several types of cell-seeded, nonimplantable bioreactors were first developed during the 1970s to replace critical metabolic functions provided by these organs; however, there remains debate regarding whether such bioartificial devices fall under the definition of tissue engineering, as they are never incorporated into the body's reparative and homeostatic mechanisms.

In the years since these early forays into tissue engineering, the field has expanded substantially and has experienced some clinical and commercial successes. In 1988, Langer and Vacanti described a method of seeding cells on a resorbable polymer matrix for cell transplantation, which ultimately became the most important enabling technology for advancement, expansion, and recognition of the field of tissue engineering. This specific technique of seeding cells on a three-dimensional (3D), porous, biodegradable scaffold catalyzed an explosion of tissue engineering research in the late-1980s–mid-1990s. While the exploration of fundamental concepts underlying tissue engineering's viability generally took a backseat to the practice of tinkering with various combinations of cells plus materials, these investigations did significantly advance tissue engineering by enabling identification and a better understanding of the obstacles facing tissue engineers. By the year 2000, mainstream media outlets were touting the promise of this new field, with Time magazine proclaiming Tissue Engineer as the number one projected career in the twenty-first century.

Today's tissue engineering research sees the collaboration of engineers with clinicians, biologists, chemists, and many other scientists to effect the creation of numerous types of engineered tissues. Present tissue engineering ventures have addressed the regeneration or replacement of components found in every system of the body (cardiovascular, musculoskeletal, neural, endocrine, digestive, reproductive, and respiratory). While significant progress has been made toward the recreation of many complex tissues, the clinical and commercial success stories of tissue engineering represent less complicated structures (e.g., skin, cartilage, and the bladder). Yet, through the incorporation of emerging technologies, (e.g., the use of embryonic stem cells) the field of tissue engineering continues to evolve and progress in order to meet its promise of revolutionizing regenerative medicine.

COMPONENTS OF TISSUE ENGINEERING

Cell Sources (1,2,4–7)

As discussed in earlier sections, tissue engineering strategies may include the transplantation of cells alone or the seeding of cells upon implantable scaffolds. Cell sourcing is a formidable challenge in current tissue engineering techniques, as it is essential to select cells of appropriate origin and maturity for each tissue engineering application. In this section, cell sources for tissue engineering are arranged and discussed with respect to cell maturity, as the regeneration potential of a cell source is a crucial consideration in designing a tissue engineering strategy.

Mature Cell Sources. Mature cells are differentiated cells that are committed to performing given cell type-specific functions. These cells are generally obtained via primary cultures originating from small pieces of donor tissue. For example, a small skin biopsy may be performed in order to isolate keratinocytes and dermal fibroblasts, the two main cell types in skin. Once isolated, these cells may be cultured and expanded *in vitro* to yield the number of cells required for creation of tissue-engineered skin. However, the expansion potential of these cells is not without limit; over time, mature primary cells drift from their original phenotype, and will either die or become genetically unstable after a certain number of population doublings. One advantage of using mature cells in tissue engineering strategies is that it enables tissue–organ regeneration using the patient's own cells, thus eliminating concerns of immunogenicity. Additionally, decades of research have concentrated on characterizing many mature cell types and their responses to various biological factors, thus facilitating the tissue engineer's ability to predict and control cell behavior and tissue formation. Thus, in situations where a tissue sample of proliferative cells can be obtained using minimally invasive means, the use of primary cell lines as a component of the tissue engineering platform represents a viable strategy. The regeneration of cartilage via Carticel (Genzyme, Inc.) is an excellent example of the successful use of mature autologous cells in tissue engineering. In the Carticel process, healthy, mature chondrocytes are isolated from a patient with large articular cartilage lesions, propagated *in vitro*, and then reimplanted into the patient. Carticel is approved by the U.S. Food and Drug Administration (FDA) and has been clinically used on $>10{,}000$ patients. Note, however, that isolation via primary culture and subsequent subculturing–expansion is not a feasible option for all cell types. For example, while cells from the liver (hepatocytes) readily regenerate *in vivo*, their *in vitro* growth is difficult and results in rapid loss of hepatocyte-specific functions. Furthermore, not all mature cell types are capable of regeneration. The cells of the adult myocardium (cardiomyocytes) are terminally differentiated and incapable of proliferation, even in the *in vivo* environment.

Immature Cell Sources. Historically, successful tissue engineering schemes have used mature, adult cells isolated from a specified tissue type as tools to produce their tissue of origin. A common example, described above, involves use of dermal fibroblasts and keratinocytes to engineer skin tissue. This general strategy has gained a measure of success in cases where cells are accessible and expandable with minimal patient trauma, as in the case of skin and cartilage regeneration. However, many tissues do not contain readily accessible mature cells that can grow outside the body and remain capable of generating their corresponding tissue type. This limitation has led investigators to search for other cell types that are capable of generating functional tissues.

Various "stem cell" types have emerged as a potentially important cell source in tissue engineering (5). These cell types possess several properties that are ideal for regenerative applications, including: (*1*) an ability to make copies of themselves, or self-renew, which may allow for creation of an endless cell source; (*2*) the ability to differentiate into multiple mature cell types; and (*3*) the ability to differentiate in response to environmental cues. These capabilities, in principle, could allow for production of highly complex tissues and organs from a renewable stem cell source. However, the capabilities of stem cells are largely based on observation of natural stem cell activities *in vivo*, and recreation of these activities in engineered systems has been a significant challenge. The following paragraphs delineate the stem cell types in current use in tissue engineering and highlight important challenges in stem cell-based tissue engineering.

Adult Stem Cells. Several investigators are working toward generation of functional tissues using stem cells isolated from a multitude of adult tissues, including skin, skeletal muscle, retina, adipose tissue, dental pulp, blood vessels, and bone marrow. These cell types are often termed tissue-specific stem cells, to distinguish them from the more primitive and pluripotent embryonic stem cells (described in the next section). The bone marrow has been a particularly fruitful cell source, yielding two multipotent cell types: hematopoietic stem cells (HSCs), and bone marrow stromal cells. HSCs were initially defined by their ability to generate all of the mature blood cell types. However, recent studies indicate that these cells can also be coaxed to differentiate into nerve cells or liver cells when delivered to the central nervous system or the liver, respectively. Therefore, HSCs appear to have applications in engineering of several key tissue types. The marrow stromal tissue has also been an intriguing source of stem cell types. Cells from this source have been termed mesenchymal stem cells, bone marrow stromal cells, or marrow-derived mesenchymal stem cells, depending on their isolation and selection procedures. Marrow-derived cells have been used to generate bone, skeletal muscle, cartilage, adipose, and vascular tissues. It is important to note that similar mesenchymal stem cells have also recently been isolated from synovial joints, adipose tissue, and umbilical cord tissue, and have been driven to differentiate into mature cell types, including bone and cartilage cells. A subpopulation of the bone marrow-derived mesenchymal stem cells, which are termed multipotent adult progenitor cells (MAPCs), have recently demonstrated an ability to differentiate into an even more extended range of adult cell

types, including liver, nerve, blood, and lung cells. The identification of these cell types suggests that tissue-specific stem cells, which have historically been considered limited in their differentiation potential, may be capable of transforming into a wider range of mature cell types than anticipated. However, no adult stem cell type has been shown to be pluripotent, and therefore able to give rise to all types of specialized adult cells, and adult cells are more limited in their ability to self-renew in culture (< 60 population doublings) when compared with the more primitive embryonic stem cells (hundreds of population doublings), described in the next section.

Embryonic Stem Cells. Another stem cell type, the human embryonic stem cell (ESC), has also generated a great deal of excitement due to its potential use in tissue engineering applications. Human ESCs are pluripotent cells derived from the inner-cell mass of blastocysts generated via *in vitro* fertilization. These ESCs have also been isolated from embryos created *in vitro* via somatic cell nuclear transfer, which is often termed therapeutic cloning. In addition, similar cell types called embryonic germ cells have been isolated from the fetal gonadal ridge. Each of these cell sources produces cells that are capable of self-renewing for extended periods in culture without differentiating, and they are considered capable of generating all of the mature cell types in the body. It is therefore possible, in principle, to use ESCs as a renewable cell source to engineer any human tissue. Although this potential is intriguing, the use of ESCs in tissue engineering applications has not been extensive. These cells are primitive and often require complex signaling environments to direct them to become a specified mature cell type. Furthermore, the complex signals required to generate specified cell types are incompletely understood. These challenges, along with the social and ethical questions associated with ESC isolation and use, present significant roadblocks to ESC-based tissue engineering. However, recent studies have successfully utilized human ESCs combined with material carriers to generate a variety of adult tissue types. In addition, direct injection of human ESC-derived cell types into pathological sites is an active area of study. These efforts and others suggest that ESCs may be an important component of emerging tissue regeneration approaches.

Challenges. Each of the stem cell types described are being explored for stem cell based tissue engineering approaches, which have included (*1*) direct transplantation of the cells into a pathological location; (*2*) transplantation of cells upon or within a biomaterial carrier; or (*3*) differentiation of the cells in a cell culture bioreactor prior to implantation with or without a biomaterial carrier (Fig. 2) (6). A particular challenge in these stem cell based approaches is delivery of signals to stem cells to direct their differentiation and, in turn, new tissue formation. Stem cells are primitive by nature, and they therefore require instructions, which are provided in the natural stem cell microenvironment (or niche) (7). Many of the emerging approaches to stem cell based tissue regeneration attempt to adopt aspects of the natural stem cell niche to imitate natural tissue development. Design of new biomaterials and new inductive approaches, coupled with new insights about the signals that direct stem cell based tissue formation, may allow for a higher level of control over tissue regeneration in the future.

Biomaterials (8–12)

Tissue engineering strategies include the implantation of material scaffolds alone or in combination with cells. A material that can be used for tissue engineering applications must meet a number of requirements; namely, materials should be biocompatible; biodegradable to nontoxic products within appropriate time frame for application; easily processed to form complex shapes with appropriate porosity; able to support cell growth and proliferation; mechanically suited to an application. There exist numerous classes of biocompatible materials, and design and synthesis of new materials remain active areas of investigation. Several considerations and concerns are common to the design of all biomaterials, regardless of whether the material is derived from natural or synthetic sources.

Naturally Derived Biomaterials (8,10,11). Natural extracellular matrices (ECMs), which are protein-based matrices that surround most cell types in the body, can be considered the quintessential biomaterials. During natural tissue development and repair processes these natural ECMs serve many of the functions that are important for successful tissue engineering, including (*1*) provision of a space-filling scaffold for infiltration of cells and synthesis of

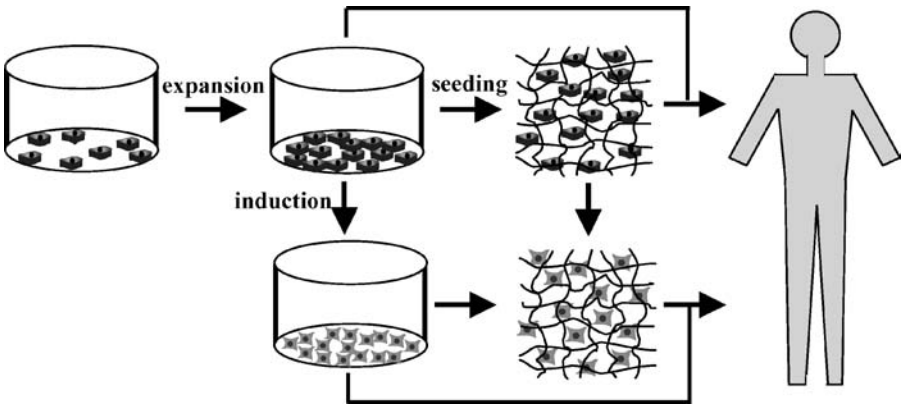


Figure 2. Schematic representation of current approaches to stem cell based tissue engineering. Stem cells are isolated from a variety of tissues, expanded in culture, and implanted into a tissue defect with or without a biomaterial carrier. In some cases, the cells are induced in culture to differentiate into a particular mature cell type prior to implantation.

a new tissue; (*2*) delivery of signals that influence cell activity during new tissue formation; and (*3*) the ability to degrade and remodel while a new tissue forms. It is therefore not surprising that a large fraction of biomaterials used in tissue engineering are derived directly from the natural ECM (8). Most of the natural biomaterials used currently are derived from connective tissues, including ligaments, tendons, blood vessels, skin, and bone. More specifically, the most common natural materials in general clinical and scientific practice include collagen (both native and chemically modified), fibrin, hyaluronic acid, and alginate. The following paragraphs provide a brief description of each of these biomaterials.

Collagen. Collagens are a class of vertebrate proteins (with >25 variations) that fold into a characteristic triple helical structure. In skin, tendon, and bone tissues the predominant type of collagen is Type I, and it is this type that is used extensively in tissue engineering applications. Type I collagen can be formed into a cross-linked hydrogel network, a fiber mesh, or a porous sponge, depending on the processing method. Therefore, it is a tremendously flexible base material, and applications have included skin, bone, cartilage, nerve, and liver tissue engineering. Collagen-based materials can also be readily mixed with cells, proteins, or plasmid DNA during processing, which has led to the use of collagen scaffolds in cell-based and inductive tissue engineering strategies.

Fibrin. Fibrin is an integral protein component of clots that form during blood coagulation. It is in this context that fibrin provides a temporary matrix during natural wound healing. Therefore, the natural function of fibrin demonstrates its potential use in wound healing and tissue engineering applications. Fibrin is capable of forming a hydrogel upon protein cross-linking, and these hydrogels adhere strongly to connective tissues, promote cell attachment and wound healing, and degrade in response to protease activity. Therefore, fibrin-based hydrogels serve as excellent matrices for blood vessel ingrowth and healing of connective tissues (e.g., skin, bone). In addition, synthetic biomaterials have been designed to mimic the biological properties of fibrin, including cross-linking, cell adhesion, and protease degradation. These biomimetic approaches highlight the importance of natural biomaterials as templates for design of synthetic biomaterials.

Hyaluronic Acid. A critical component of the extracellular matrix in many tissues, hyaluronic acid (HA) (10) is a relatively simple, yet unusual, high molecular weight polysaccharide. Hyaluronic acid is present in all mammals, playing a vital role in embryonic development, extracellular matrix homeostasis, wound healing, and tissue regeneration. This acid possesses unique biological and mechanical properties due to its hydrophilic, polyanionic composition, and the influence of HA on cell function is highly dependent on its molecular weight. Degradation of HA may occur via several cell-secreted enzymes, or nonenzymatically via free radicals or other various treatments. The intrinsic physicochemical and biological properties of HA have generated widespread use of this molecule in clinical therapies, and many HA-based products have been approved by the FDA for clinical use in osteoarthritis, ophthalmology, wound healing, gastrourology, and prevention of postsurgical adhesions. Due to the integral involvement of HA in tissue regeneration and the ease of chemically modifying HA to form numerous derivatives, the use of HA has also been extensively investigated in a wide array of tissue engineering applications. Moreover, as HA plays a crucial role during the morphogenesis of many organs, it may be capable of providing specific signals to cells to initiate tissue or organ regeneration. Derivatization or chemical cross-linking of HA yields HA-based materials with a wide range of mechanical, chemical, and biological properties. These scaffolds can take numerous physical forms including fibrous meshes, sponges, microspheres, and hydrogels.

Alginate. Alginate is a natural polysaccharide isolated from seaweed. In the presence of divalent cations (e.g., calcium) or multivalent polymers (e.g., polylysine) alginate chains become cross-linked into a network hydrogel with intriguing properties. The utility of alginate hydrogels in biomaterials and tissue engineering applications is largely a result of their ability to repel proteins. Proteins do not intrinsically interact with alginate hydrogels and, in turn, cells do not bind to these gels. Therefore, alginate hydrogels provide useful platforms for presentation of specific molecules to cells, allowing these gels to be used as controllable synthetic matrices. In addition, alginate hydrogels do not bring about a substantial immune response, so they are useful in applications designed to shield cells or tissues from the immune system (e.g., pancreatic islet delivery). Another attractive property of these hydrogels is their ability to gel *in situ* without the use of harsh cross-linking agents. This allows for minimally invasive implantation of a material, and may enable minimally invasive cell, protein, and gene delivery. Alginate hydrogels have been used in applications ranging from cell encapsulation to bone tissue engineering.

Other Naturally Derived Materials. Numerous other naturally derived materials have demonstrated promise for tissue engineering applications, including chondroitin sulfate, gelatin, agarose, chitosan, and dextran. A more detailed discussion of the role of naturally derived materials in tissue engineering is available in several reviews (8,10–12).

Limitations. Clearly, natural polymers can be used in a wide variety of tissue engineering applications, as they are generally biocompatible, biodegradable, and easy to process. However, there are limitations to the use of these materials. There is often substantial batch-to-batch variation in the properties of these polymers, which can be dependent on the species and tissue of origin or the harvesting procedure. In addition, because these materials are protein or polysaccharide based they are typically not amenable to standard polymer processing schemes that involve high temperatures or harsh organic solvents. In addition, the nature of protein-based natural materials makes them vulnerable to immune responses *in vivo*,

particularly in cases where the proteins are implanted into a host species that differs from their source. These concerns and others provide an impetus to develop synthetic biomaterials that exploit the advantages of natural materials and limit their disadvantages.

Synthetic Materials (9,11,12). Although synthetic materials lack the inherent bioactivity and biologic recognition possessed by natural substances, their advantages are numerous. In a broad sense, the primary advantage associated with synthetic materials is the researcher's ability to accurately control material characteristics. Controlled alterations in material chemistry, polymerization, or method of scaffold formation can produce changes in the type and time of material degradation, porosity, stiffness–elasticity, texture, shape, hydrophilicity, and bioactive ligand presentation. The ability to construct synthetic materials from scratch enables researchers to tailor material properties to the desired specifications required for a specific application. While many such manipulations are also possible with natural materials, it is widely accepted that synthetic materials are more amenable to modification. This section will discuss several of the most prevalent synthetic materials used in tissue engineering applications.

Poly(α-esters). Poly(glycolic acid) (PGA) and poly(lactic acid) (PLA) remain the most widely used synthetic materials for tissue engineering. These polymers are produced via a ring-opening polymerization of glycolide (R = H) or lactide ($R = CH_3$), respectively (Fig. 3).

Poly(glycolic acid) is a highly crystalline, hydrophilic poly(α-ester) that was used as early as the 1960s as the first biodegradable suture material. Simple, random hydrolysis of the ester bonds leads to bulk degradation of PGA, with the degradation rate dependent on the degree of PGA crystallinity. The degradation product of PGA is glycolic acid, which is processed through normal metabolic pathways in the body and ultimately eliminated via the respiratory system as carbon dioxide. Because the degradation products of PGA act as components of the body's natural metabolic pathways, this polymer may be classified as a bioresorbable material. For tissue engineering applications, PGA is most often processed via extrusion to form a 3D, porous mesh structure composed of uniform PGA strands or fibers. Cells seeded on these scaffolds readily adhere to the PGA fibers, then spread, proliferate, and produce extracellular matrix proteins. The PGA has proven to be an excellent carrier for numerous cell types and other biological agents. Due to its rapid degradation and poor mechanical properties, however, use of PGA alone is not suitable for many applications.

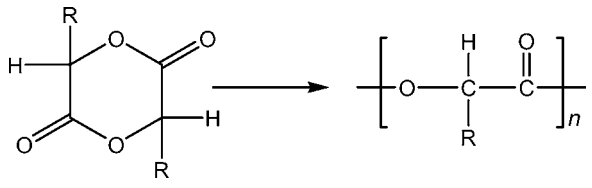


Figure 3. Ring-opening reaction of glycolide (R = H) or lactide ($R = CH_3$) to form PGA or PLA, respectively.

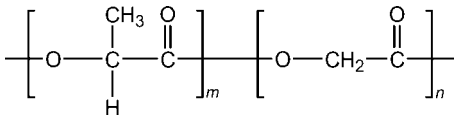


Figure 4. Copolymers of PGA with PLA can be synthesized to form poly(lactic-co-glycolicacid) (PLGA), whose characteristics represent a combination of the two original polymers.

Similar to PGA, PLA is a resorbable poly(α-ester) that degrades via hydrolysis to generate a product (lactic acid) that is readily metabolized by the body. However, PLA is more crystalline and hydrophobic than PGA, resulting in significantly slower degradation characteristics. The PLA is exceptionally strong for a biodegradable material and its superior strength makes it particularly well suited for regeneration of load-bearing tissues (e.g., bone and cartilage).

The synthesis of PLA–PGA copolymers (PLGA, Fig. 4) has yielded tissue engineering scaffolds with a wide range of mechanical and degradation characteristics. Because of their excellent biocompatibility and versatility, PLGA scaffolds have been widely used in tissue engineering applications with much success. The polymers PGA, PLA, and PLGA are all highly processable; scaffolds may be fabricated via extrusion, injection molding, compression molding, or solvent casting. They can be formed into complex structures, and there exist numerous methods for introducing pores. Nanofibrous scaffolds have also been fabricated from PLGA and PLA. These scaffolds mark an interesting advancement in the design of matrices for tissue engineering, as their physical structure (an array of fibers 50–500 nm in diameter) mimics the structure of fibrillar collagen, a component of the native extracellular matrix. A further reason that PGA and PLA remain the materials of choice for tissue engineers is that they are among the few synthetic scaffold materials approved by the FDA. The primary drawbacks of the poly(α-ester) family of polymers are their brittleness and tendency to crumble, lack of chemical functionalities other than end-groups, and production of acidic degradation products that often cause inflammation.

Poly(Anhydrides). Polyanhydrides are notable as a scaffold material for tissue engineering in that they undergo surface degradation via hydrolysis of the anhydride group, as opposed to the bulk degradation experienced by poly(α-esters). Polyanhydrides are also the first new synthetic degradable material approved by the FDA in >20 years. Poly(sebacic acid) (SA) and poly(*p*-carboxyphenoxyhexane) (CPH) are two commonly synthesized polyanhydrides for tissue engineering and drug delivery applications (Fig. 5).

Because homopolymers of SA degrade over a period of days, while polymers based upon CPH degrade over periods of months to years, copolymerization of SA with CPH offers a means of creating polyanhydride materials with a wide range of controlled surface-erosion characteristics. Surface-degrading materials have better retention of mechanical strength than those that degrade via bulk mechanisms; scaffold structural integrity is compromised early in the bulk degradation process, and the majority of mass loss occurs toward the end of the degradation process,

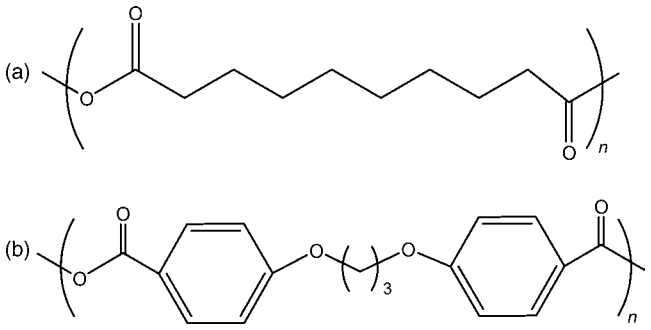


Figure 5. Polyanhydrides, such as SA (a) and CPH (b) are commonly used in tissue engineering and drug delivery applications.

often resulting in inflammation following the sudden burst of acidic degradation products. Such drawbacks make surface-eroding polyanhydrides advantageous for several tissue engineering related applications. Polyanhydrides have proven to be excellent vehicles for drug delivery, and are therefore particularly appropriate for the scaffold-only approach to tissue engineering, where these materials provide a structural platform that elicits *in situ* tissue ingrowth and regeneration through the release of bioactive and tissue-inductive agents. The tissue engineering potential of this class of materials was recently enhanced through the modification of SA and CPH to contain photopolymerizable moieties, meaning that the starting monomers are cross-linked to form scaffolds via exposure to a certain wavelength of light. As discussed in following sections, formation of scaffolds via photopolymerization affords several advantages over many other polymer processing techniques.

Poly(Ethylene Glycol) and Poly(Vinyl Alcohol). Hydrogels synthesized from poly(ethylene glycol) (PEG) or poly(vinyl alcohol) (PVA) represent yet another major class of biomaterials used in tissue engineering applications. An interesting feature of both PEG and PVA is that they are relatively biologically inert, meaning that they are resistant to protein and cell adhesion. While it may appear counterintuitive that nonadhesivity makes these materials attractive for tissue engineering purposes, the reason that tissue engineers value this property of PEG and PVA comes back to the need to control the biomaterial environment. Both PEG and PVA ultimately provide tissue engineers with a "blank" template that can be systematically modified in a controlled manner to possess a wide range of defined chemical, mechanical, and biological properties.

Unlike the polyesters and polyanhydrides described earlier in this section, PEG and PVA are not inherently degradable. However, PEG and PVA can be easily modified to contain either enzymatic or hydrolytic degradation sequences, or both. Thus, not only can the tissue engineer tailor the hydrolytic degradation time for a PEG scaffold, but they may also incorporate moieties that are sensitive to cell-secreted enzymes in order to facilitate creation of a system where scaffold degradation coincides with tissue growth.

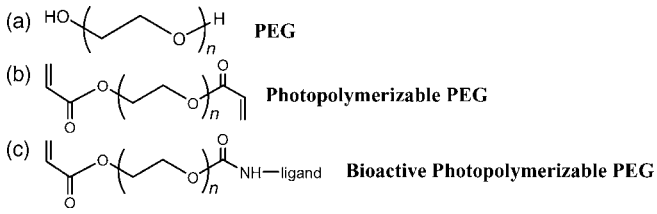


Figure 6. (a)Poly(ethylene glycol) may be chemically modified with vinyl groups (b) to form a material that is polymerized into hydrogel structures via exposure to light (photopolymerized PEG). (c) Bioactive molecules may also be covalently incorporated into these structures by modifying one end group of photoactive PEG with a biological ligand.

Fabrication of PEG and PVA scaffolds is commonly performed via photopolymerization. The end groups of these polymers are modified to contain vinyl groups (Fig. 6) that enable light-induced polymer crosslinking and hydrogel formation under specified conditions. Unlike other processing methods, the conditions of photopolymerization are very mild and can be executed in the presence of cells and sensitive bioactive agents. Photopolymerization may be performed *in situ*, thus offering a minimally invasive method of implanting biomaterials into the site of desired tissue repair. Moreover, the photopolymerized material can assume complex shapes, as it can fill irregularly sized defects *in situ* without complicated molding or shaping techniques. The hydrogel structures that result from photopolymerization are transparent in appearance and possess varying mechanical properties that are dependent on polymer molecular weight and concentration. These materials may also be covalently modified to present various bioactive agents that stimulate cell adhesion to the scaffolds, as discussed in later sections. The PEG and PVA hydrogels possess excellent biocompatibility and are very successful in supporting cell viability, growth, and function. Disadvantages of these materials include their poor tensile strength and relatively small pore size.

Other Synthetic Materials. Numerous other synthetic materials have demonstrated promise for tissue engineering applications, including poly(acrylamides), poly(urethanes), poly(orthoesters), poly(lactones), poly(propylene fumarates), and poly(phosphazenes). A more detailed discussion of the role of synthetic materials in tissue engineering is available in several reviews (9,11,12).

Limitations. Synthetic materials alone cannot provide biological cues to influence cell behavior and tissue formation; these materials serve as only structural supports for cell growth. Thus, it is difficult to recreate a cell's natural biological environment using only synthetic materials. Moreover, issues with material biocompatibility persist. Biocompatibility of scaffolds and their degradation products remains a significant obstacle in the development of new synthetic materials. Furthermore, biocompatibility does not necessarily imply hemocompatibility (blood compatibility), creating further challenges in the creation of materials for blood-contacting applications. Finally, because most synthetic materials degrade via hydrolysis,

tissue engineers continue to struggle with the problem of tailoring material degradation to tissue regeneration such that the material remains for a sufficient period of time to support tissue growth, but not so long as to inhibit tissue formation.

Bioreactors

The previous sections highlighted the importance of isolating cells that are capable of generating new tissues and creating materials that can support new tissue growth. Once the appropriate cell types are procured and material scaffolds are constructed, what is the optimal manner in which to combine these components to generate new tissues? Furthermore, what is the appropriate environment for cultivation of a new tissue as it develops? These are critical questions in tissue engineering approaches, and a variety of bioreactor (13,14) systems have been built to control and optimize cell engraftment cell survival, and proper cell function within biomaterial scaffolds.

Cell Seeding. An initial challenge posed in cell-based tissue engineering schemes involves simply placing cells within a biomaterial construct, a process termed cell seeding. In most strategies, it is important to encourage as many cells as possible to engraft within a biomaterial scaffold to promote generation of new tissues and to avoid squandering valuable cellular components. More efficient cell seeding limits the size of a biopsy needed for cell sourcing and may also reduce the extent of cell expansion necessary to produce adequate cellular components. In addition, studies aiming to generate cartilage, bone, and cardiac tissues have demonstrated that cell seeding density and homogeneity directly influence the growth of new tissues. Higher cell densities lead to enhanced tissue formation, while more homogeneous cell seeding distributions result in more uniform tissue formation. In view of these previous results, investigators have developed a variety of strategies to encourage efficient and homogeneous cell seeding.

The simplest method of cell seeding, termed static seeding, involves simply adding a cell suspension onto a scaffold construct and passively allowing the suspension to permeate the material. This method often leads to inefficient cell seeding and heterogeneous cell distribution. In fact, static seeding approaches often result in growth of only thin $\sim$100 μm layers of tissue due to limitations in cell seeding density and homogeneity as well as limited nutrient diffusion. The limitations of this static approach have led to development of more active cell seeding approaches, which include stirred bioreactor systems, direct perfusion bioreactors, and rotating wall bioreactors. In a stirred bioreactor, a dilute suspension of cells is continuously stirred around a stationary, porous scaffold construct, allowing for convective flow. Similarly, direct perfusion bioreactors encourage infitration of cells by directly flowing a cell suspension through a stationary porous scaffold. Rotating wall bioreactors encourage mixing of cells with scaffolds by rotating the entire bioreactor casing around a central axis. Each of these strategies has enhanced the efficiency and homogeneity of cell seeding, resulting in more copious and uniform new tissue development.

Mass Transport. Encouragement of mass transport to and from a developing tissue is among the most formidable challenges in tissue engineering. This challenge is particularly daunting *in vitro*, where a developing tissue cannot be exposed to a functional vascular network. Mass transport is vital in tissue engineering approaches, as it provides a means for delivery of oxygen and nutrients to cells within a developing tissue. Tissues that are grown in the absence of facilitated mass transport typically contain a shell of viable tissue surrounding an inner core of necrotic tissue. The thickness of healthy tissue is commonly $<$1 mm, which is not an appropriate scale for a majority of intended tissue engineering applications. To address this limitation several investigators have developed bioreactor systems that encourage mass transport throughout a developing tissue. These systems use mechanisms that are similar to the aforementioned cell seeding bioreactors, and in this case convective flow is used to transport oxygen, nutrients, and wastes during new tissue development. The result is a significant increase in the total amount of tissue grown, both *in vitro* and *in vivo*. Bioreactor systems that enhance mass transport have become important in bone, liver, and skeletal muscle tissue engineering approaches, and the clear benefit of mass transport during tissue development *in vitro* has motivated strategies aimed at increasing mass transport *in vivo* (e.g., by promoting vascular tissue in growth).

Inductive Signaling. Efficient and homogeneous cell seeding and optimal cell survival have a substantial impact on the success of a tissue engineering approach. However, it is also important to expose growing tissues to signals that direct their development into functional tissues. This is particularly crucial in emerging stem cell-based tissue engineering schemes, in which cells must be exposed to signals that direct differentiation and induce formation of the tissue of interest by cells capable of generating multiple mature tissue types. Therefore, there is a need to develop bioreactor systems that provide a range of signals to cells to encourage functional tissue growth *ex vivo*. To that end, investigators have developed bioreactor systems that are capable of delivering both biochemical signals (e.g., protein growth factors) and mechanical forces (e.g., fluid shear forces, compressive forces). As a result, new bioreactor systems allow for growth of tissues that more effectively mimic natural tissue structure and function. Illustrative examples include cartilage, bone, skeletal muscle, ligament, and cardiovascular tissues.

Of course, in many cases the most effective bioreactor system for new tissue development is actually the *in vivo* implantation site. Therefore, an important consideration in tissue engineering is how long, if at all, a cell-scaffold construct should be cultured *ex vivo* prior to implantation. For example, several recent approaches have focused on encouraging vascular tissue infiltration *in vivo* to address the challenge of mass transport to and from a developing tissue. For tissues that are particularly dependent on high oxygen tension (e.g., liver tissue) inducing vascular tissue

ingrowth *in vivo* may be a more effective strategy when compared with convective fluid transport in a bioreactor *ex vivo*. However, for poorly vascularized tissues (e.g., cartilage) that rely heavily on the local mechanical environment during their development, the mechanically and biochemically controlled environments of a bioreactor system may allow for optimization of tissue development *ex vivo*. Other tissue types may benefit from an initial culture period *in vitro* for cell expansion and early tissue development, followed by implantation into a supportive *in vivo* site. The amount of *in vitro* culture time that is appropriate prior to implantation of a tissue engineered construct is an important open question in several approaches.

Mechanical Signaling. The mechanical environment is a key factor during development of several tissue types, and mechanical forces become important as early as the eight-cell stage of embryonic development. Similarly, mechanical forces are important parameters during development of several engineered tissue types. Numerous physical factors have been shown to influence growth of engineered cartilage tissue, including hydrodynamic forces and cyclic mechanical compression and tension. In addition, systems for application of tensile stresses to growing skeletal muscle influence the ultimate contractile properties of the engineered muscle tissue. Mechanical forces also have a clear effect on bone development, repair, and regeneration, and investigators have taken advantage of these effects to build bioreactors that provide mechanical stimulation (see the section Case Studies in Tissue Engineering). To address the importance of mechanics, several investigators have generated instrumentation for application of biomechanical forces to developing tissues. Bioreactors can be designed to include screws to apply static strain or motorized load cells for dynamic actuation, and the biomechanical components are engineered to interface with other bioreactor components, including stirring tools for mass transport and media containing inductive agents. Biomechanical forces have played a particularly prominent role in emerging functional tissue engineering approaches, which often aim to provide factors that directly mimic the characteristics of the corresponding *in vivo* environment and improve the properties of the engineered tissue.

Emerging Designs. Although they are beyond the scope of this article, it is important to note that new concepts are emerging in bioreactor design at the cellular and subcellular scale. Several investigators are actively developing microfluidic systems that are capable of culturing multiple cell types simultaneously in highly controlled environments. These systems are being customized for coculture of multiple cell and tissue types (often termed organ-on-a-chip approaches), as well as highly controlled drug delivery. Microfluidic systems can also be interfaced with biomaterials to provide a powerful platform for highly controlled tissue development *ex vivo*. These approaches and others are enhancing the level of complexity that is possible in bioreactor design, and may allow for more direct mimicry of natural microenvironments.

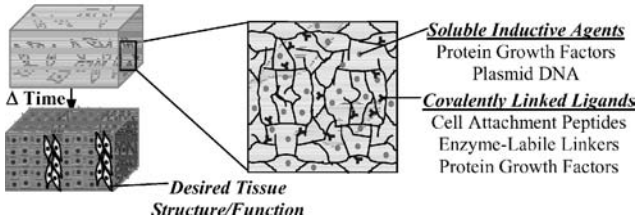


Figure 7. Schematic representation of inductive tissue engineering approaches. Cells are exposed to 3D environments that are engineered to contain covalently linked or soluble agents. The agents are chosen to influence cell activity, ultimately leading to induced growth of a desired tissue.

INDUCTIVE APPROACHES

A variety of investigators have designed 3D matrices to interface directly with mammalian cells, and these materials were described in the biomaterials section above. Both natural (e.g., collagen gels) and synthetic [e.g., poly(L-lactic acid)] materials have been used for decades as scaffolding to support tissue formation from cells *in vitro* or *in vivo*. These materials have historically been passive, in that they have not exerted a high level of control over the signals presented to resident cells. In the past decade, matrices have been designed to interact with cells via specific cell attachment ligands, or designed to deliver soluble signals, so that characteristics of the cell's microenvironment can be tightly controlled (Fig. 7). Bioactive matrices have been used to delineate the effects of specific cell attachment ligands, material properties (e.g., mechanics), and soluble signals on attachment, proliferation, migration, and differentiation of a wide variety of mature cell types. These strategies are often termed inductive approaches, as they are designed to induce a particular cell activity (e.g., tissue formation). The following paragraphs describe inductive schemes that have been used to add a bioactive component to biomaterial matrices.

Covalent Modification of Biomaterials

One widespread approach used to confer bioactivity involves covalently linking biologically active molecules to a biomaterial (15). In this general approach, a biological molecule is engineered to include a chemically reactive group, which is then reacted with a functional group within a biomaterial. The result is a material that has physical properties defined by the original material, and with biological properties defined by the linked biological molecule. This approach has been primarily used to link peptide sequences that promote cell attachment, including RGDS, RGDY, YIGSR, and IKVAV. For example, synthetic PEG hydrogels and natural alginate hydrogels, which do not intrinsically interact with cell surface receptors, can now be readily modified with peptides that promote cell attachment and desired function. In addition to these methods that deliver signals to cells, investigators have also developed materials that respond to cell activity. One example of this general approach is materials that are held together, or cross-linked, by biochemically labile chemistries. These

materials are designed so that their labile linkages are degraded in response to enzymes that are produced by resident cells. Thus, by creatively designing these linkages investigators have been able to generate materials that respond to cell migration and differentiation, resulting in materials that can effectively "listen" to cells. Taken together, the inductive biomaterials created by covalent modification can be considered truly cell interactive, as they are able to deliver signals (e.g., cell adhesion ligands) to cells, and respond to cell activity (e.g., by degrading).

Delivery of Soluble Factors

Another important focus of recent tissue engineering approaches has been delivery of proteins, called growth factors, which can directly modulate cell activity (16). Traditional growth factor delivery approaches have focused on embedding proteins in plastic micron-scale spheres or suspending proteins in highly hydrated gels. In each case, the protein is released via diffusion as the carrier materials degrade, resulting in a high local concentration of the protein. The advent of these technologies has had a revolutionary effect on medicine, and the worldwide market for drug delivery technologies is expected to exceed $100 billion by 2005. Although these technologies have been useful in a wide variety of biomedical applications, their application to tissue engineering is pragmatically limited. Plastic microspheres do not provide a structural matrix for tissue ingrowth and are difficult to process into structural matrices while maintaining the biological activity of the embedded protein.

Hydrated gels are also nonideal carriers, as growth factors typically diffuse out of the gel matrix rapidly, resulting in limited signaling. Investigators have recently developed innovative approaches to address the limitations of these systems and allow for longer term protein release within structural matrices that support tissue development. For example, plastic microspheres have been engineered to form aggregates with cells or to form a matrix with interconnected pores, and these new strategies allow for inclusion of biologically active proteins. The proteins are then released into the cell population in a sustained manner as the material degrades. These methods are flexible in the type of growth factors that can be included, and have recently been extended to delivery of multiple active growth factors simultaneously. Another recent approach involves covalent immobilization of active growth factors within a hydrogel matrix to locally contain growth factors and limit diffusion out of the hydrogel. This approach has been applied to delivery of vascular and neural growth factors, and could potentially be applied to other proteins, provided they maintain biological activity when covalently immobilized. Taken together, these approaches have been quite successful in actively influencing cell activity within structural matrices during new tissue development. The next generation of protein delivery approaches is likely to exert a higher level of control over where and when cells are exposed to inductive signals. Spatial patterning and tightly regulated timing of protein delivery are routine in natural tissue development, and may be similarly important in complex tissue engineering applications.

Gene Delivery

Another set of inductive approaches does not rely on delivery of peptides or proteins to cells from a synthetic system, but instead focuses on genetic manipulation of cell activity (17). Investigators have developed a variety of strategies to deliver genes to cells, and the genes are typically designed to induce cells to secrete inductive proteins. The gene delivery approaches often utilize the same core technologies that have been used for protein delivery, including sustained release from polymer scaffolds and release from hydrogel matrices, but they substitute plasmid DNA for proteins. Plasmids are circular strands of DNA, which typically contain a single gene under the control of a promoter, which is chosen to enable production of an encoded protein in the target cell type. The encoded protein, often a protein growth factor, is chosen to influence cell activity. Therefore, this approach generates cells that act as bioreactors to produce an inductive protein, and the protein, in turn, encourages new tissue development. This gene delivery approach has been successfully applied to engineering of bone, skeletal muscle and vascular tissues, and the general concept could find broad applicability in engineering of essentially any tissue type.

CASE STUDIES IN TISSUE ENGINEERING

Bone Tissue Engineering Strategies

Vertebrate bone tissue is a dynamic organ with a range of vital functions. Bone serves as a storage depot for mineral ions, a protective barrier for internal organs, and a solid support for muscle actuation. Perhaps the most unique and intriguing property of bone is its ability to continuously regenerate its structure, thereby maintaining consistent structural and mechanical properties. A coordinated set of cell and molecular activities mediates this dynamic remodeling process, and these activities are orchestrated by biochemical, mechanical, cellular influences. The dynamic nature of natural bone tissue makes it perhaps the ultimate "smart material", and the characteristics of natural bone have served as an inspiration to chemists and materials scientists who aim to develop stimulus-responsive and self-healing materials. Furthermore, the constant generation of new bone tissue in natural systems serves as an excellent model for tissue engineering, a field that similarly strives to regenerate natural tissue structure and function. Therefore, bone can be considered an exemplary system for a variety of tissue engineering approaches, including conductive, inductive, and cell-based strategies.

It is perhaps surprising that there is any need for engineered bone tissue at all in view of the efficiency of bone regeneration in vertebrates. However, there are several pathological conditions that result in permanent bone loss or damage. Costs of musculoskeletal conditions represent an average of 3% of the gross domestic product of developed countries (an estimated $254 billion annually in the United States) and >600,000 inpatient fracture reduction procedures are performed in the United States annually. Furthermore, bone and joint diseases account for one-half of all chronic conditions in people over the age

of 50. The predicted doubling of this age group's population by 2020 suggests that the tremendous need for novel bone repair and replacement therapies will continue to grow rapidly. Regeneration of natural tissue represents a promising approach to replace bone, and could both supplant many of the current metallic hardware-based bone replacement methods and expand the range of bone loss conditions that can be effectively treated. Filling of bone voids in non-union fractures or maxillofacial deformities, bridging of gaps in spine fusion surgeries, and stabilization of vertebral compression fractures provide illustrative potential targets for new bone tissue engineering approaches. Current clinical strategies aimed at repairing or replacing natural bone tissue are typically passive, relying on mechanically sound hardware (e.g., titanium, polyethylene) to replace the structural properties of natural bone. These approaches do not exert a high level of control over the process of new bone formation in a defect site. Limitations in these approaches invoke new therapies for bone replacement, including bone tissue engineering (18–20). The following paragraphs give a brief description of representative tissue engineering approaches that have been developed to repair or replace bone tissue. A more focused and detailed treatment of this topic is given in several outstanding review papers (18–20).

Cell Sources. Osteoblasts are the chief bone-forming cells in natural bone tissue. Based on the regenerative capacity of natural bone tissue it seems obvious that mature osteoblasts would be an excellent candidate as a cell source for bone tissue engineering. These cells can be isolated from the patient (autologous transplantation) or from a donor (allogeneic transplantation), and the typical donor sites are the calvaria or the iliac crest. Investigators have used osteoblasts seeded within various biomaterial scaffolds to engineer bone tissue. However, there are significant limitations to the use of these differentiated cell types. Isolation procedures result in a limited number of autologous or allogeneic osteoblasts, and these cells are difficult to expand in culture. In addition, allogeneic cells harbor the potential to bring about unwanted immune responses. These limitations have led rapidly to the identification and use of alternative cell sources in bone tissue engineering.

The best characterized cell sources in common use in current bone tissue engineering approaches are bone marrow derived fibroblasts. These cells include the aforementioned mesenchymal stem cells, as well as other more committed osteoblast precursors, called preosteoblasts, isolated from bone marrow. These cell types, which are typically distinguished from other cells that reside in the marrow based on their ability to attach and form colonies on standard cell culture substrates, have a long history of use in bone biology and orthopedic regeneration. Between 1968 and 1974, Friedenstein and co-workers (see Ref. 5 for details) undertook a series of studies on the bone-forming capacity of bone marrow derived cells. The studies resulted in isolation of a specific cell type, termed the colony-forming unit fibroblast, which was characterized by its ability to attach to standard cell culture substrates and form isolated colonies in culture. It was not until two decades later that these cells were further characterized by Caplan and coworkers (see Ref. 5 for details) and called mesenchymal stem cells (MSCs). The pioneering work of Caplan's laboratory in skeletal tissue engineering using MSC sources, coupled with a high profile publication by Pittenger et al. (see Ref. 5 for details) describing the tremendous multipotential of these stem cells, has facilitated use of these cells in bone tissue engineering applications. Recent studies have also identified other adult precursor cells that are capable of forming bone tissue, including skin-, adipose-, and dental pulp-derived mesenchymal stem cells. Investigators are now using each of these cell types as integral components in bone tissue engineering schemes.

Biomaterials. Natural long bones (e.g., the femur) develop upon a cartilage template matrix during a process termed endochondral ossification. In many ways, this natural cartilage template serves as an ideal support for development of new bone tissue. The cartilage matrix: (*1*) is porous and therefore allows for infiltration of bone-forming cells and vascular tissue; (*2*) is capable of withstanding the mechanical environment in an orthotopic location; (*3*) provides a substrate that is conducive to new bone formation; (*4*) is biodegradable and can be remodeled by infiltrating bone cells; and (*5*) contains signals that induce new bone formation. These natural characteristics mirror the parameters that are important for design of natural and synthetic scaffolds for bone tissue engineering, including pore structure, mechanical properties, degradability, osteoconductivity, and osteoinductivity. Osteoconductivity is generally defined as the ability of a material to support formation of bone by bone-forming cells (e.g., in a bone defect), while osteoinductivity is defined as the ability of a material to induce formation of bone tissue in conditions that are not otherwise conducive to bone formation (e.g., in a nonorthotopic location). Investigators have developed several classes of biomaterials to address these parameters and to successfully engineer functional bone tissue.

A broad range of natural and synthetic materials have been explored in bone tissue engineering, including poly (anhydrides), poly(fumarates), self-assembling peptide amphiphiles, and alginate hydrogels. Although these materials are well characterized in the context of bone engineering, the most commonly used bone tissue engineering scaffolds have been poly (α-hydroxy esters), type I collagen-based materials, and calcium phosphate based minerals. Each of these base materials can be processed into porous scaffolds that degrade into nontoxic byproducts, allowing for formation of new bone tissue in concert with material degradation. In addition, the mechanical properties of these matrices are dictated by their structure and composition, allowing for significant mechanical tailoring. A particular advantage of calcium phosphate based materials is a property known as bioactivity, which is the ability of these materials to serve as an excellent template for synthesis of mineralized tissue by bone cells and bone precursor cells *in vitro* and *in vivo*. For example, mesenchymal stem cells differentiate and form bone tissue more readily on calcium phosphate materials when compared with other types of scaffold, including polyesters and

protein-based hydrogels, and this phenomenon is attributed to calcium phosphate bioactivity. Many bone tissue engineering approaches exploit this bioactivity by processing or synthesizing new types of calcium phosphate based materials, including both natural (e.g., coralline hydroxyapatite) and synthetic (e.g., sintered hydroxyapatite) scaffolds. Recent approaches have also combined natural and synthetic polymers with calcium phosphate minerals to create hybrid materials for bone tissue engineering. These materials directly mimic the organic–inorganic composite structure of the natural bone extracellular matrix, and they have shown substantial promise as supportive scaffolds for new bone formation.

Bioreactors. *In vitro* engineering of bone tissue has been an active area of study for over a decade. However, early attempts at bone tissue engineering demonstrated limitations that highlight the importance of bioreactor design. Previous studies indicate that growth of bone tissue within 3D scaffolds in static culture conditions is limited to the outer 200–800 μm of a scaffold construct. The poor tissue growth and cell death observed at locations deeper within these scaffolds was likely caused by poor nutrient transport into the developing tissue. These studies and others have provided an impetus to create new bioreactor systems that enhance mass transport and allow for development of more homogeneous bone tissue *in vitro*. Spinner flasks and rotating wall bioreactors have been shown to substantially enhance bone synthesis by mesenchymal stem cells cultured within 3D polyester scaffolds. In addition, perfusion culture systems successfully enhanced generation of bone tissue by osteoblasts and bone marrow-derived precursor cells *in vitro*. In each case, the success of these approaches is attributed to improved nutrient transport within the developing tissue, which is generated by fluid flow.

Other *in vitro* bioreactors have been designed to understand and manipulate the effect of mechanics on engineered bone tissue growth. Mechanical forces have a well-known influence on remodeling of adult bone tissue and, in turn, regulation of bone strength over time. Mechanical stimulation is also vital to proper early development of cartilage and bone, and the mechanical properties of developed bone tissue can be correlated to the load applied during development. In view of these mechanical effects on bone maintenance and development, it is not surprising that mechanical forces have a pronounced effect on bone formation by osteoblasts and osteoblast precursor cells in cell and tissue culture. Culture systems that apply compressive or shear forces to bone-forming cells have shown that the cells respond to mechanical stress by enhancing synthesis of bone matrix.

Inductive Approaches. It is clear that protein growth factors are a vital component of natural bone development and repair processes, due to their effects on bone forming cells and blood vessel ingrowth. Based on these observations, investigators are developing several novel delivery systems that allow for controlled delivery of protein growth factors to bone defect sites to improve or accelerate bone healing. The most potent known growth factors related to bone regeneration are the class of molecules known as bone morphogenetic proteins (BMPs). Discovery of the unique ability of demineralized bone matrix to induce bone tissue formation in extraskeletal sites led to the isolation and discovery of BMPs as inductive factors. Since their discovery, BMPs have been delivered from several materials to induce formation of new bone tissue in a variety of skeletal and extraskeletal sites. The BMP delivery vehicles examined thus far include demineralized bone matrix, polyester scaffolds, β-tricalcium phosphate, and hydroxyapatite. In each case, BMPs in conjunction with a natural or synthetic carrier material *in vivo* induced dramatic increases in the quantity and functionality of regenerated bone tissue when compared with the carrier materials alone.

Recent inductive approaches have also addressed the importance of angiogenesis in natural bone growth, development, and repair. Bone does not grow, develop, or heal properly when angiogenesis is artificially blocked. However, until recently there had been little interest in specifically inducing ingrowth of a functional vascular supply to support developing bone tissue. Studies have recently demonstrated that delivery of growth factors that induce blood vessel ingrowth (e.g., vascular endothelial growth factor) can substantially increase bone repair and regeneration. These studies provide a new mechanism for actively inducing bone regeneration.

Inductive gene delivery strategies have also been explored in bone tissue engineering. Investigators have delivered plasmid deoxyribonucleic acid (DNA) encoding bone morphogenetic protein-4 (BMP-4) and parathyroid hormone from collagen-based scaffolds. Localized delivery of these plasmids led to the formation of local bone foci similar to natural bone. In addition, multiple studies have achieved transfection of fibroblasts with a gene encoding for bone morphogenetic proteins, resulting in more extensive bone formation. These genetic approaches will likely gain more significance and acceptance with the emergence of more efficient methods for gene transfer, and they could be used in conjunction with novel stem cell based approaches.

Combination Approaches. In this section, bone tissue engineering components are separated into distinct categories to facilitate the readers understanding of the field. However, it is important to note that the vast majority of strategies for bone tissue engineering involve a combination of multiple components, including materials, cells, and biological molecules. In fact, emerging tissue engineering schemes almost exclusively utilize combinations of two or more of the categorized components in this section, and it has become more common for investigators to unite materials, molecules, cells, and highly controlled bioreactor environments to direct bone tissue development. Development of combined approaches requires a more complete understanding of the interdependence of distinct components. For example, the combination of BMPs with a carrier material is critical, and the identity of the carrier material can significantly influence BMP activity. In addition, surface characteristics and geometry of the scaffold material significantly influence induction by BMPs. The complex interplay between substrates and growth factors in synthetic tissue engineering scaffolds emulates the

cross-talk between extracellular matrix signals and soluble signals in natural tissue development and regeneration. Indeed, as the complexity of bone tissue engineering systems increases, researchers may approach the intricacy and control demonstrated during natural bone development and regeneration.

Cardiovascular Tissue Engineering Strategies (21–23)

Cardiovascular disease is a significant cause of morbidity and mortality in the United States and developed countries. Successful treatment has often been limited by the poor performance of synthetic materials utilized for tissue replacement, as hemocompatibility remains a significant challenge in biomaterial design. A lofty goal for cardiovascular tissue engineering is the development of a completely tissue engineered heart. Progress toward this goal will likely be made through the parallel development of effective tissue engineered components of the cardiovascular system. These individual components include blood vessels, heart valves, and cardiac muscle. Each of these structures will be briefly discussed with respect to the cell source, biomaterial, and bioreactor considerations.

Heart Valves. The human heart contains four valves, each with slightly different characteristics and each experiencing a different hemodynamic environment. While heart valves are relatively small, thin structures (on the order of a few hundred microns thick), their composition is surprisingly complex. Valve dysfunction may occur via a variety of mechanisms, and the current treatment for diseased valves is predominantly valve replacement. Over 290,000 people received heart valve replacements in 2003; this number has been steadily rising over the last decade and is expected to reach 850,000 by 2050. Currently available valve substitutes have enabled these patients to experience an enhanced quality of life and have extended patient survival. Yet, in 50–60% of patients with substitute valves, complications associated with these valve replacements necessitate reoperation or cause death within 10 years postoperatively. Present heart valve substitutes consist of either mechanical prosthetic valves or tissue valves that are derived from either human or animal tissue. Perhaps the greatest shortcoming of current valve substitutes is their inability to grow or remodel in response to the physiological environment. Valve replacement is particularly challenging for pediatric patients, who not only outgrow mechanical valves quickly, but also experience rapid calcification (pathological hardening) of transplanted tissue valves. Valve replacement in the elderly has also become more complex; as the average life span increases, but the age at which heart valve disease occurs does not, patients > 65 now need valves that will last 20 years, not just 10. Fabrication of a tissue engineered valve using a biodegradable scaffold may enable the creation of functional valve tissue capable of growth and remodeling in response to changes in its physiological environment. Significant advances have recently been made toward the creation of a functional tissue engineered heart valve via the immobilization of cells within a variety of natural and synthetic matrices, as will be briefly discussed here. A durable, nonobstructive, nonthrombotic, self-repairing tissue valve that would grow with the patient and remodel in response to *in vivo* stimuli is the current goal for tissue engineered heart valves.

Cell Sources. The source of cells to use in valve regeneration remains a significant concern. Valvular interstitial cells (VICs) comprise the majority of the cell population in heart valves, with a thin layer of endothelial cells (ECs) providing the valve with a nonthrombogenic surface. The adaptive, complex, and dynamic structure of heart valves can be primarily attributed to the VICs, which are responsible for valve extracellular matrix production as they constantly remodel and repair the valve. The organization and relative proportions of the valve matrix are paramount to valve function, thus emphasizing the importance of these interstitial cells. Although these cells readily proliferate and function in *in vitro* culture, they are very difficult to obtain from a patient or living donor. Other mature cell types, such as smooth muscle cells, dermal fibroblasts, and myofibroblasts, can perform some of the functions of VICs, but ultimately do not exactly mimic all of the properties of VICs. Recent studies suggest that marrow stromal cells may be a viable immature cell source for valve tissue engineering. Embryonic stem cells also hold promise as a cell source for VICs, although the method of inducing their differentiation to VICs has not yet been identified.

Construction of fully functional heart valves will also require endothelialization of the valve surface. The ECs play a critical role in both the maintenance of valve homeostasis and the pathogenesis of valvular disease. Mature cell sources for ECs include dermal microvascular endothelial cells (HDMECs), which are isolated from human foreskin, and endothelial cells isolated from human umbilical veins (HUVECs). However, neither of these sources provide autologous cells for a patient in need of valve replacement. Mesenchymal stem cells have recently been shown to differentiate into ECs, indicating that MSCs are a promising source for autologous ECs. Researchers have also identified techniques to induce the differentiation of ECs from embryonic stem cells.

Biomaterials. Construction of a tissue engineered heart valve requires a material that is readily processed into a complex shape, can withstand harsh hemodynamic stresses, and yet is flexible enough to allow for valve opening and closure. To this end, synthetic scaffolds have been constructed using PGA, PLGA, polyhydroxyalkanoate (PHA), or poly-4-hydroxybutyrate (P4HB). Studies using PGA/P4HB in particular have been very successful in large animal models. Much work has also focused upon using decellularized native valves as a scaffold material; in this approach, immunogenicity of a donor valve is lessened by removing the cells, leaving only the extracellular matrix structure. The primary advantage of this strategy is that the appropriate matrix composition of the valve is already present and does not have to be reconstructed. These valve scaffolds can then be recellularized with autologous (nonimmogenic) cells, with the goal of restoring the valve's regenerative capacity. Finally, hydrogels made from

several naturally derived materials, such as collagen or hyaluronic acid (HA), have also been used for heart valve tissue engineering. Although HA-based valve scaffolds have not yet progressed to *in vivo* studies, the use of HA is particularly exciting, as this polysaccharide is required for cardiac morphogenesis and native heart valve formation.

Bioreactors. Because it is difficult to survive with only a partially formed or semifunctional heart valve, tissue engineered matrices must first be cultured in a bioreactor environment prior to implantation. Culture of tissue engineered valves in bioreactors aims to create mechanically stable and reliable valve tissue with significant (or complete) degradation of the material scaffold prior to implantation. In addition to providing appropriate nutrients to the cell-seeded scaffolds, such bioreactors are intended to provide physiologically relevant flow and shear stresses in order to condition the tissue for *in vivo* implantation and function. Furthermore, tissue engineered heart valves exposed to pulsatile fluid shear stresses *in vitro* display improved function over constructs cultured in bioreactors without a mechanical signal. Investigators working on tissue engineered heart valves have made significant advances in bioreactor design, and these bioreactors may have utility for the culture of other tissues that require pulsatile stresses.

Cardiac Muscle. Because adult cardiac muscle cells (cardiomyocytes) cannot regenerate, myocardial damage induced by a myocardial infarction (heart attack) or other injury results in loss of cardiac function and progressive deterioration leading to congestive heart failure. Necrotic cardiomyocytes in infarcted ventricular tissue are replaced by fibroblasts, leading to the formation of scar tissue and creating regional contractile dysfunction. Current treatments consist of mechanical support using left ventricular assist devices and, ultimately, heart transplantation. The incidence of heart failure at 1 year postheart attack is >20%, and the 1 year postheart attack mortality rate (~30%) did not change from 1975–1995, highlighting the lack of effective treatments to combat this problem. Multiple tissue engineering strategies, including cell-only transplantation, inductive approaches, and implantation of cell-seeded scaffolds have been investigated in order to generate a viable method of repairing damaged heart tissue.

Cell Sources. Cardiomyocytes (CMs) are highly differentiated cells that comprise 70–80% of the heart's mass, yet only 20–30% of the total cardiac cell number. The remaining cell population consists of cardiac fibroblasts and endothelial cells, although CMs are primarily responsible for the contractile activity of the heart. Electromechanical coupling between CMs enables the transduction of electrical signals into muscular contractions. Use of mature autologous cells is feasible for some tissue repair applications, but is likely not possible for myocardial repair, primarily because adult CMs do not proliferate, and removal of any cardiac tissue to obtain CM progenitor cells could impair the function of the already-injured heart.

Cellular cardiomyoplasty (CCM) is a promising therapy that has recently emerged for the repair of damaged cardiac muscle and involves the injection of a suspension of cardiac-related cells into the injured heart (cell transplantation). The therapeutic goal of CCM is the replacement of dead heart muscle with functionally competent and contractile myocardium. Sources for these injected cells have included skeletal myoblasts, fetal cardiomyocytes, hematopoietic stem cells, mesenchymal stem cells, embryonic stem cells, and endothelial progenitor cells, yet the functional fate of these cells following transplantation remains unresolved, and normal electromechanical coupling between implanted cells and host myocardium has been absent. While CCM has been shown to augment myocardial function, there exists debate about whether the transplanted cells are actually functioning as cardiomyocytes. Strategies that do not involve regeneration of cardiomyocytes are limited to rescuing injured tissue and are unable to contribute directly to the restoration of contractile function or increase systolic function. Hence, if the transplanted cells are not differentiating into cardiomyocytes, then contractile, electrically coupled, fully integrated cardiac tissue is not being regenerated, leading to suboptimal improvement of cardiac function. Furthermore, because only 20% of the heart's cells are CMs, stem cells injected into the heart simultaneously receive developmental cues from many different cell types, making it difficult to predict the differentiation pathway that the transplanted cells will follow. Because researchers currently cannot regulate the differentiation factors to which cells are exposed *in vivo*, transplantation of stem cells that are committed to the cardiomyocyte lineage may be a more effective method of regenerating CMs *in vivo*. These cell sourcing issues observed with cellular cardiomyoplasty also extend to the tissue engineering strategy of implanting cell-seeded biomaterial constructs.

Biomaterials. While cellular cardiomyoplasty has been successful in augmenting cardiac function, it has still not been proven to reliably regenerate cardiac muscle. Thus, many researchers are combining cardiac-related cells with biomaterials in order to develop an approach that enables greater control over tissue formation. Numerous synthetic, natural, and biomimetic materials have demonstrated promise for cardiac tissue engineering applications. Seeding of cardiomyocytes on PGA- or PLGA-based materials has produced cardiac grafts whose *in vitro* structural and electrophysiological properties approach those of native heart muscle. However, the acidic degradation products and brittleness of these materials have hindered their *in vivo* success. Several collagen-based matrices have been used to create small myocardial structures that morphologically resemble and possess many of the properties of native myocardium, while alginate matrices also show good promise as scaffold materials for cardiac repair. Bioactive or inductive materials may also be used in a scaffold-only approach for cardiac tissue regeneration; recent studies have demonstrated the existence of a small number of cardiac progenitor cells within heart muscle, and implantation of a bioactive material containing appropriate agents may recruit these myocardial progenitors to the site of injury. While the number of polymers that have

been used to develop cardiac muscle is small at this point compared to the number used for other tissues, (e.g., bone, skin, and blood vessel) this number is likely to increase over the next decade as the push to develop a completely tissue engineered heart increases.

Bioreactors. *In vitro* development of cardiac muscle requires the use of appropriate nutrient delivery systems. Cell-seeded constructs cultured in perfusion bioreactors contain more uniformly distributed cells with enhanced differentiation. While direct perfusion of cell-scaffold constructs has been shown to be beneficial, culturing the constructs under laminar flow rather than turbulent flow has also been found to enhance engineered cardiac muscle. Laminar flow conditions can be provided by using a rotating bioreactor rather than a spinner flask. Another approach for *in vitro* culture of cardiac muscle is to subject cell-scaffold constructs to mechanical stimuli, such as pulsatile flow and cyclic stretch, as would be experienced *in vivo*. Bioreactors that combine perfusion mechanisms with mechanical signals, such as compression or stretching, significantly improve cardiac tissue formation and function over perfusion-only systems. The nutrient requirements for cardiac tissue are quite high, and this tissue is particularly sensitive to ischemic conditions. The thickness of a tissue engineered myocardial construct that can be developed using current techniques is limited to 100 μm, contrasting the 1 cm thickness of native heart muscle. Recent developments in the design of bioreactors for heart valves, which employ adjustable pulsatile flow and varying levels of pressure, may be applicable to the culture of engineered myocardium.

Large Diameter Blood Vessels (>6 mm). Vascular disease is a prominent problem in the United States, with approximately 500,000 coronary artery bypass surgeries performed each year. Natural tissues, primarily saphenous veins or internal mammary arteries, are generally used for coronary artery replacement. In general, the results have been quite favorable for these procedures. Unfortunately, as discussed earlier, the availability of donor vessels and issues with donor site morbidity often preclude the use of natural tissue grafts. Moreover, grafted vessels are often unable to adjust to the increased pressure and wall shear stress in the grafted position, resulting in inadequate vessel performance. Blood vessel replacements composed of entirely synthetic materials were first developed in the 1950s using polymers, such as polyesters, polyethylene terephthalate (PET, Dacron), and expanded polytetrafluoroethylene (ePTFE, Gore-Tex). However, implantation of these materials is permanent, with no regeneration of the biological blood vessel. Furthermore, several problems such as platelet adhesion and activation, and decreased mechanical compliance compared with adjacent arterial tissue, are also frequently associated with synthetic grafts. These problems have motivated investigations of tissue engineered alternatives.

Cell Sources. Blood vessels consist of three cell types: fibroblasts, smooth muscle cells, and endothelial cells. Most tissue engineering endeavors focus upon the smooth muscle and endothelial cells, which are generally obtained from harvested autologous blood vessels. However, issues, such as donor site morbidity, vessel availability, and performance of an invasive surgery complicate the use of these mature, autologous cells. A solution to these problems may lie in the use of mesenchymal stem cells, which are capable of differentiating into vascular smooth muscle cells and endothelial cells, and may therefore represent an excellent cell source for vascular tissue engineering.

Biomaterials. Attempts to tissue engineer blood vessels date back to the 1970s, and the number of different materials that have since been used in these endeavors is numerous. Construction of a tissue engineered blood vessel requires a material that is resistant to platelet adhesion (nonthrombogenic), can withstand transmural pressure acting normal to the vessel wall and tangential shear stresses, and has similar mechanical properties to that of the adjoining native vessel. Much research has focused upon the use of collagen gels for vascular tissue engineering. Although collagen gels have poor mechanical properties, proper alignment and function of vascular cells in these gels has been observed. The PGA Polymer has been used alone and in combination with other synthetic materials (polyhydroxyalkanoate) to form tissue engineered vessels that remained functional for several months following implantation in an animal model. Recent research has focused upon the use of PEG hydrogels as vascular tissue engineering scaffolds, as the physical and biological properties of these materials are well suited for vascular applications. Finally, a unique vascular tissue engineering approach used no material, but instead rolled sheets of vascular cells into tubular shapes and observed excellent mechanical and functional properties of the tissue.

Bioreactors. With respect to shear stresses, flow, and application of appropriate mechanical signals, the bioreactor culture of tissue engineered vascular grafts possesses many similarities to the other cardiovascular applications discussed above. Application of fluid shear stresses is necessary in order to generate the cellular alignment found in native vessels and to stimulate proper function of vascular cell types. Because *ex vivo* studies of native intact vessels often have similar culture requirements as tissue engineered vascular grafts, a diverse range of researchers has been involved in the development of such bioreactors, yielding numerous configurations and approaches. Variables in the design of vascular bioreactors include constant or pulsatile flow through the vessel lumen, application of longitudinal strain, and perfusion of exterior surface of the vessel.

BIBLIOGRAPHY

1. Langer R, Vacanti JP. Tissue engineering. Science 1993;260: 920–926.
2. National Science Foundation. Report: The emergence of tissue engineering as a research field; 2003.
3. Saltzman WM, Tissue engineering: Engineering principles for the design of replacement organs and tissues. New York: Oxford University Press; 2004.
4. Lavik E, Langer R. Tissue engineering: Current state and perspectives. Appl Microbiol Biotechnol 2004;65:1–8.

5. Lanza R, et al., editors. Handbook of Stem Cells. Volumes 1 and 2. New York: Elsevier Academic Press; 2004.
6. Zandstra PW, Nagy A. Stem cell bioengineering. Ann Rev Biomed Eng 2001;3:275–305.
7. Li L, Xie T. Stem cell niche: structure and function. Ann Rev Cell Dev Biol 2005;21:605–631.
8. Yannas IV. Natural materials. In: Ratner BD, Hoffman AS, Schoen FJ, Lemons JE, editors. Biomaterials Science. New York: Elsevier Academic Press; 1996.
9. Lee KY, Mooney DJ. Hydrogels for tissue engineering. Chem Rev 2001;101:1869–1879.
10. Baier Leach J, et al., editors. Encyclopedia of Biomaterials and Biomedical Engineering. New York: Marcel Dekker; 2004.
11. Cannizzaro S, Langer R. Biomaterials: Synthetic and engineering strategies. In: Dillow AK, Lowman AM, editors. Biomimetic Materials and Design: Biointerfacial Strategies, Tissue Engineering and Targeted Drug Delivery. New York: Marcel Dekker; 2002.
12. Thomson RC, Ishaug SL, Mikos AG, Langer R. Polymers for biological systems In: Meyers RA, editor. Encyclopedia of Molecular Biology: Fundamentals and Applications. New York: VCH Publishers; 1996.
13. Ratcliffe A, Niklason LE, Bioreactors and bioprocessing for tissue engineering. Ann NY Acad Sci 2002;961:210–215.
14. Martin I, Wendt D, Heberer M. The role of bioreactors in tissue engineering. Trends Biotechnol 2004;22:80–86.
15. Lutolf MP, Hubbell JA. Synthetic biomaterials as instructive extracellular microenvironments for morphogenesis in tissue engineering. Nat Biotechnol 2005;23:47–55.
16. Saltzman WM, Olbricht WL. Building drug delivery into tissue engineering. Nat Rev Drug Discov 2002;1:177–186.
17. Bonadio J, Goldstein SA, Levy RJ. Gene therapy for tissue repair and regeneration. Adv Drug Del Rev 1998;33:53–69.
18. Salgado AJ, Coutinho OP, Reis RL. Bone tissue engineering: State of the art and future trends. Macromol Biosci 2004;4: 743–765.
19. Jadlowiec JA, Celil AB, Hollinger JO. Bone tissue engineering: Recent advances and promising therapeutic agents. Exp Opin Biol Ther 2003;3:409–423.
20. Green D, Walsh D, Mann S, Oreffo ROC. The potential of biomimesis in bone tissue engineering: Lessons from the design and synthesis of invertebrate skeletons. Bone 2002;30:810– 815.
21. Leor J, Amsalem Y, Cohen S. Cells, scaffolds, and molecules for myocardial tissue engineering. Pharmacol Ther 2005;105:151–163.
22. Masters KS, Mann BK. Tissue engineered heart. In: Nedovic V, Willaert R, editors. Applications of Cell Immobilisation Biotechnology. New York: Kluwer Academic; 2005.
23. Mann BK, West JL. Tissue engineering in the cardiovascular system: Progress toward a tissue engineered heart. Anat Rec 2001;263:367–371.

See also BIOMATERIALS: TISSUE ENGINEERING AND SCAFFOLDS; BIOSURFACE ENGINEERING; ENGINEERED TISSUE; SKIN TISSUE ENGINEERING FOR REGENERATION; SKIN SUBSTITUTE FOR BURNS, BIOACTIVE.

TISSUE ENGINEERING. See BIOMATERIALS: TISSUE ENGINEERING AND SCAFFOLDS.

TISSUE EQUIVALENTS. See PHANTOM MATERIALS IN RADIOLOGY.

TMJ. See TOOTH AND JAW, BIOMECHANICS OF.

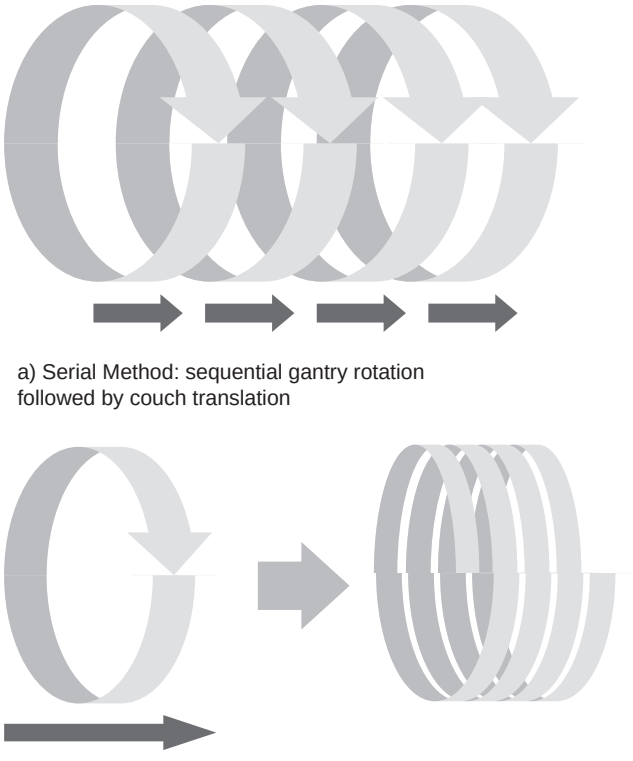
TOMOTHERAPY

TIMOTHY HOLMES
St. Agnes Cancer Center
Baltimore, Maryland

THOMAS R. MACKIE
University of Wisconsin
Madison, Wisconsin

INTRODUCTION

Tomotherapy is intensity-modulated rotational radiotherapy utilizing a photon fan beam (1). The term tomotherapy derives from *tomographic radiotherapy*, literally meaning "slice" radiotherapy. Tomotherapy treatment delivery is conceptually similar to computerized tomographic (CT) imaging where a three-dimensional (3D) image volume is acquired as a stack of two-dimensional (2D) cross-sectional images. By analogy, a tumor volume can be subdivided into a stack of slices that are independently irradiated to achieve a 3D conformal dose distribution. Dose conformation is achieved by modulating the intensity pattern of the incident X-ray beam during rotation of the X-ray source using a fan-beam multileaf (MLC) collimator whose leaves are pneumatically driven to achieve near instantaneous leaf transitions between open and closed states. As shown in Fig. 1, the table motion can be incrementally stepped the width of the fan beam after each completed arc (serial tomotherapy) (2,3) or it can be



a) Serial Method: sequential gantry rotation followed by couch translation

b) Helical Method: simultaneous gantry rotation and couch translation

Figure 1. Tomotherapy methods: (a) serial tomotherapy, (b) helical tomotherapy.

(a) (b)

Figure 2. Components of the NOMOS Peacock System: (a) gantry mounted MIMiC binary MLC and controller, and (b) table mounted precision positioner for accurately stepping the table between gantry rotations. (Photos courtesy Jim Hevzi, PhD, Cancer Treatment and Research Center, San Antonio, TX.)

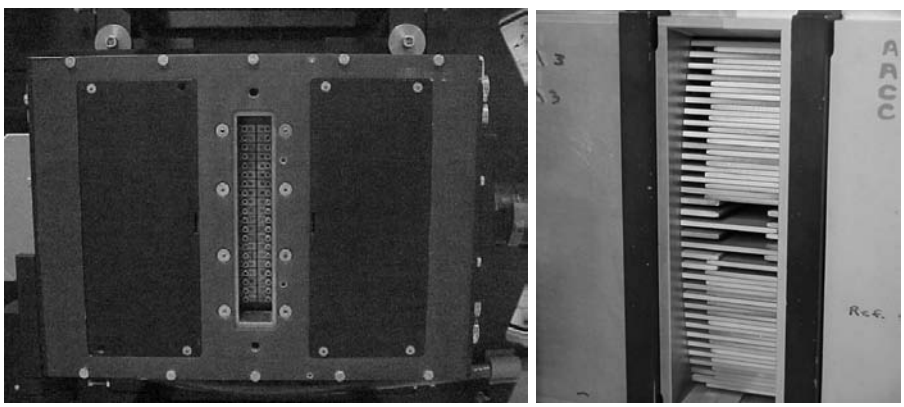
translated simultaneously with continuous source rotation (helical tomotherapy) (4).

Two types of tomotherapy, serial and helical, have been developed and implemented clinically. Serial tomotherapy, embodied in the NOMOS Peacock System (North American Scientific - NOMOS Radiation Oncology Division, Cranberry Township, PA) is an add-on component for conventional C-arm medical linac gantries (Fig. 2a). It is feasible to deliver non-coplanar serial tomotherapy treatments using this implementation. Alternatively, the Hi-ART II helical tomotherapy system (TomoTherapy, Inc., Madison, WI) is a dedicated radiotherapy treatment unit built upon a helical CT ring gantry. The constraint of a ring gantry is minimal since few patients are treated with noncoplanar radiation fields and IMRT diminishes the need for these types of field arrangements. Most importantly, a ring gantry is a very stable platform for CT scanning and is used in all diagnostic CT scanners.

Both tomotherapy implementations have a high level of system integration with common components being (1) an inverse treatment planning system, and (2) computer controlled beam modulation using a pneumatically driven multileaf collimator (Fig. 3). One major difference between the two implementations is the integration of MVCT image guidance into the treatment delivery process of the Hi-ART II (5). The MVCT imaging is primarily used for daily target location to improve precision and accuracy of IMRT treatments. In addition, the detector can be used for machine quality assurance, and for reconstructing daily estimates of delivered dose using measured profiles of X-ray fluence transmitted through the patient during treatment delivery (6). This latter capability, called *dose reconstruction* is useful to the process of *adaptive radiotherapy* wherein a daily discrepancy between planned and delivered dose due to errors in patient setup or changes in gross tumor volume can be corrected in future treatments.

HISTORICAL BACKGROUND ON THE DEVELOPMENT OF TOMOTHERAPY

Rotational treatment techniques were very popular in the era of orthovoltage (250 kVp) X rays prior to the development of megavoltage photon sources. Rotational delivery allowed improved dose at the tumor relative to the surface compared to treatments using a few fixed-beams portals. This was a consequence of the peak dose for orthovoltage X rays occurring at the skin surface, which for nonrotational deliveries was the dose-limiting feature of the treatment. The development of high energy photon beams from Co-60 and medical linear accelerators reduced the need for rotational treatments since the peak dose was shifted below the skin surface resulting in improved skin sparing when few field techniques were used. As a consequence, treatment techniques shifted almost exclusively to simpler few field beam arrangements, typically of opposed beams, that were easy to verify using planar radiography. This in turn eventually led to the development of dual energy medical linear accelerators, or linacs, with two X-ray energies: a low X-ray energy (4, 6 MV) for treatments in the



(a) (b)

Figure 3. Tomotherapy binary MLCs: (a) the NOMOS MIMiC design has 20 pairs of opposed leaves that allow two slices to be treated per rotation. (Photo courtesy Jim Hevzi, PhD, Cancer Treatment and Research Center, San Antonio, TX). (b) The TomoTherapy binary MLC design consists of 64 interdigitating leaves that are used to treat a single slice per rotation. (Photo courtesy TomoTherapy, Inc, Madison, WI.)

head–neck region, and a high X-ray energy (10 MV or greater) for treating deep seated lesions in the trunk and pelvis. In situations where a single low energy machine was the only treatment unit available, rotational techniques were often used for deep-seated lesions, owing to the fact that the relative shape of the rotational dose distribution is insensitive to the X-ray energy used.

In 1988, Anders Brahme of the Karolinska Institute published an article that described the advantages of using nonuniform beam profiles to achieve conformal dose distributions of arbitrary shape (7). A key feature of this work was the application of a mathematical optimization method to determine the nonuniform profiles for few-field delivery. Subsequent efforts by Bortfeld et al. (8), Webb (9), and Holmes et al. (10) applied iterative tomographic image reconstruction methods to inverse planning of intensity modulated beams with the latter two investigations providing methods applicable to rotational IMRT. Webb's application of the simulated annealing algorithm for planning rotational IMRT (9) was implemented in the first commercial inverse treatment planning system, PeacockPlan (NOMOS Corp., Sewickley, PA), to support serial tomotherapy. The work of Holmes et al. (10,11) and Swerdloff et al. (12) under the direction of Dr. T. R. Mackie at the University of Wisconsin Department of Medical Physics began the formulation of a dedicated helical tomotherapy unit (1).

SERIAL TOMOTHERAPY: NOMOS PEACOCK SYSTEM

The Peacock System was developed during the early 1990s by the NOMOS Corporation under the leadership of Mark Carol, a neurosurgeon who became interested in the radiation therapy problem during his medical training (2,3). This system was designed as an add-on device to existing medical linear accelerators, many of which in the early 1990s lacked the features of a computer controlled MLC and a record and verify system needed to perform IMRT delivery. The first serial tomotherapy treatments were performed at the Methodist Hospital, Baylor College of Medicine, in March of 1994 under an Investigational Device Exemption with clearance from the Food and Drug Administration (FDA) coming in 1996. Since then, >100 systems have been installed clinically with over 10 thousand IMRT treatments completed successfully.

The Peacock System consists of four components: (1) CORVUS: an inverse treatment planning system; (2) MIMiC: a computer controlled multileaf collimator; (3) the controller: a dual computer system for control and continuous checking of the MIMiC; and (4) the AutoCrane: a computer controlled table positioner. The latter three components are shown in Fig. 2.

CORVUS was the first commercial implementation of inverse treatment planning for IMRT. The planning system consists of (1) tools for import and fusion of CT and MRI images for subsequent contouring of treatment and avoidance regions including organs at risk; (2) a finite-size pencil beam 3D dose computation model; (3) an optimization engine based on simulated annealing and gradient-based optimization algorithms; (4) algorithms to convert optimized X-ray intensity maps to multileaf collimator field shapes and their associated treatment times; and (5) tools to map patient treatments onto dosimetry phantoms for treatment plan delivery verification.

The MIMiC is a binary multileaf collimator. It consists of 20 pairs of opposed tungsten leaves that are 8 cm tall and project to nominally 1 cm wide at 100 cm from the radiation source. The leaves move quickly in and out of a fan beam to provide the intensity modulation. The motive power for the leaves was compressed air pushing and pulling on pistons. The binary collimator was licensed from the University of Wisconsin patent invented by Swerdloff. et al. (12). The opposed leaf design allows two adjacent treatment slices to be treated simultaneously during a single arc producing a more efficient delivery. The treatment slice width can be set to 1 or 2 cm using mechanical stops, or to 0.4 cm using a tertiary collimator called the BEAK. The maximum size of the "modulated" volume is a 20 cm diameter cylinder whose length is determined by the longitudinal motion of the treatment table.

Tomotherapy treatments typically require an order of magnitude greater number of monitor units than conventional 3D conformal radiotherapy, consequently the leakage characteristics of the collimation system and the treatment unit are of concern as they influence the whole body dose equivalent received by the patient and the risk of a second malignancy that might occur as a consequence of treatment (13,14). Leakage transmission through a MIMiC leaf and through the gap between leaves is <1%; the leakage transmission outside of the defined slice is determined by the leakage characteristics of the treatment unit jaws and head shielding, which is typically 0.1%. These leakage characteristics were improved upon in the development of the dedicated helical tomotherapy unit described in the next section.

Attached to the MIMiC housing is a pc-controller. This device is a dedicated record and verification system used to monitor the (1) the angular position of the treatment gantry using onboard inclinometers, and (2) the open and closed state and transition times (<140 ms maximum) of the MIMiC leaves during treatment delivery. During treatment delivery, the gantry angle defines the open–close state of each leaf as defined by the treatment plan, which is created assuming constant dose rate and gantry rotation speed. Unacceptable changes in gantry speed will be detected by the controller causing treatment interruption by tripping the interlock circuit of the treatment room door thereby terminating the radiation beam. The controller also verifies the integrity of the plan data, which is stored on a 3.5 in. password protected floppy disk that is originally created by CORVUS as a means of transferring the plan to the controller. Data integrity is ensured using CRC check sums of the data files on the floppy disk. Modification of a plan requires the user to decommission the floppy disk using a password while inserted in the CORVUS planning station.

A continuous gantry rotation is modeled as a set of fixed beams spaced at 10° intervals from $\sim\pm160°$ from the vertical gantry position. A 10–15° interval is required at startup to allow the gantry speed and beam output to stabilize before leaves are opened to modulate the beam. Intensity modulation is achieved by opening a leaf for a

fraction of the interval with opening and closing occurring at angles symmetrical to the center of the interval. For example, a 50% modulation is achieved by opening the leaf at -2.5° and closing it at $+2.5^\circ$ about the center of a 10° interval. Typically, 10 intensity levels are used to approximate an intensity profile so leaf transitions occur at 1° boundaries within the 10° interval.

Serial tomotherapy requires precise positioning of adjacent treatment slices at the submillimeter level to avoid unacceptable dosimetry errors in the abutment region between slices (15). In recognition of this issue, the initial implementation of the Peacock System used a manual table positioner called the Crane that was a professional grade photography studio stand (Cambo BV, Kampen, The Netherlands) modified to attach to the treatment table side rails with table position determined by high precision digital linear scales. This device was subsequently replaced by a computer-controlled positioning device called the Auto-Crane, that was originally developed as the Xlator by the University of Texas Health Sciences Center, San Antonio, TX (16).

HELICAL TOMOTHERAPY: TOMOTHERAPY, INC. HI-ART II

Figure 4 shows a TomoTherapy Hi-ART II tomotherapy unit (TomoTherapy, Inc., Madison, WI) with its covers off and its various subsystems labeled. The linac is a 30 cm long, 6 MV, S-band (nominal 3 GHz) magnetron-powered device with a gridded-gun and a solid-state modulator. The linac is used for both CT imaging and treatment delivery, with imaging carried out using a lower beam current to reduce patient dose, and lower X-ray energy to improve image contrast. Given its compact size, the linac is aligned parallel to the beam axis with the flattening filter removed to increase output to 8–10 Gy/min at the gantry rotation axis located 85 cm from the fixed target. Elimination of the field-flattening filter improves the energy spectrum of the beam and reduces scatter outside the field boundary, thereby simplifying treatment beam modeling and CT image reconstruction.

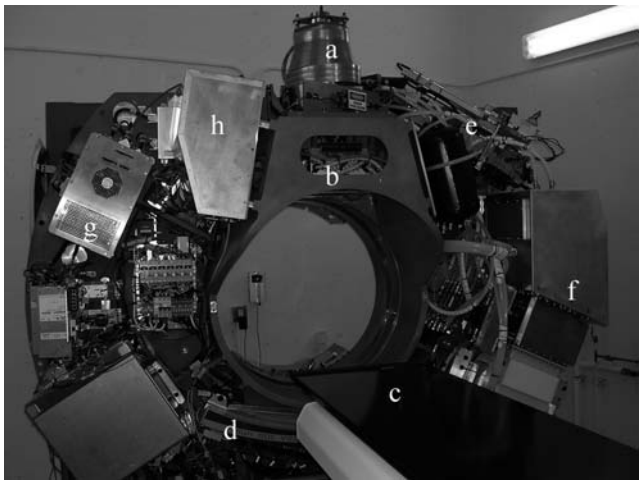


Figure 4. Helical tomotherapy unit in a factory test cell. The major components are: (a) linac, (b) binary MLC, (c) couch, (d) detector, (e) RF system, (f) modulator, (g) gantry control, and (h) RF power.

In recognition of the increased number of monitor units required for tomotherapy delivery, the treatment head shielding was designed to limit the primary leakage to 0.01% of the primary beam, or one-tenth the limit (0.1%) used for conventional radiotherapy treatments. The shielding includes integrated primary collimators that are used to define field widths from 5 to 50 mm (nominal). Below the primary collimators resides the binary multileaf collimator module, consisting of 64 leaves made from 10 cm high tungsten with leakage $<0.5\%$. A 13 cm thick lead counter weight is attached to the ring gantry opposite the treatment head that acts as a rotating primary barrier or beam stop.

Each binary MLC leaf completely blocks a portion of the fanbeam with a projected shadow of 6.25 mm (nominal) at the gantry rotation axis. The 64 leaves define a 40 cm diameter treatment field of view, which combined with up to 160 cm (nominal) couch travel enables very large treatment volumes to receive IMRT. Intensity modulation is achieved using pneumatic control of the binary MLC leaves by rapidly ($\sim$20 ms) switching the open–closed state of leaves during gantry rotation. As in the MIMiC collimator, the intensity level is proportional to the time a leaf is open and effectively there are $\sim$50 intensity levels that that can be delivered.

The linac and gantry systems of the tomotherapy system are highly favorable for CT imaging where mechanical stability of the source-detector positions during rotation and a small source size are desirable. The gantry sag of the tomotherapy system is $\sim$0.1 mm so no sag corrections are required in the CT reconstruction algorithm. The size of the electron beam on the target is $\sim$1 mm so that the resolution is $\sim$1.2–1.6 mm, which is comparable to a conventional CT scanner for high contrast objects. Operating at an average dose to the patient of 1–3 cGy, the images produced have soft tissue contrast of 2–3%, which is poorer than a modern CT scanner, yet are of sufficient quality for adaptive radiotherapy processes. The tomotherapy unit's xenon gas detector elements have tungsten septa separating ionization cavities. In addition to the ionization collectors, the tungsten plates are embedded photon converters intercepting the megavoltage photons and yet are thin enough to let an appreciable fraction of the electrons set in motion to deposit energy in the xenon gas. The interception of the beam by the tungsten means that the quantum efficiency of the system is $\sim$25%, which is much more than the few percent collection efficiency of modern portal imaging systems.

Modeling the treatment delivery process requires discretization of the continuous motions of the gantry and table as well as the continuous intensities of the modulated beams. Proper sampling reduces the chance for computational aliasing that can produce "streak" or "thread" artifacts in the dose distribution (17). Consequently, each 360° gantry rotation is modeled as 51 beams spaced at 7.06° apart: a number chosen to allow a 40 cm diameter target volume to be homogeneously treated with a 2.5 cm completely blocked central avoidance structure (18). Following

optimization, the intensity levels are discretized for treatment delivery, with 50 levels chosen to reduce the uncertainty in the target dose due to intensity discretization to <0.1%. Discretization of couch travel is determined by the pitch ratio: the ratio of the couch travel distance per rotation to the field width defined at the axis. In helical tomotherapy delivery, the pitch is usually set to be less than one-half to avoid thread-like dose artifacts developing near the edge of the field becoming clinically significant (17). Given a typical pitch of 0.3 for a 25 mm field width, the table motion is modeled by offsetting adjacent beams by 0.147 mm (e.g., 0.3×25 mm/51) increments parallel to the direction of table motion.

A Hi-ART II treatment delivery typically takes <5 min for small target volumes like a prostate and <10 min for larger volumes such as head and neck. Mackie et al. (19) provided an expression for estimating the irradiation time given the target length L, the beam width W, the prescribed dose D, the average dose rate at the target R, and the modulation factor M:

$$T = MD(L + W)/WR \quad (1)$$

The modulation factor is a user selectable planning parameter along with the beam width and pitch. It is defined as the ratio of the maximum leaf open time of any leaf to the average leaf opening time of all the nonzero values, and is typically selected in the range of 1.5–3.0. As an example, a 7 cm prostate volume irradiated at an average dose rate at the target volume of 4 Gy/min by a 2.5 cm wide fan beam with a modulation factor of 2.5 would yield a beam-on time of ~4.75 min.

THE HI-ART II CT IMAGING SYSTEM

The detector resolution of the Hi-ART II unit projected to the axis of rotation is ~0.6 mm in the transverse direction and equal to the slice width in the longitudinal direction. The rotation period of the Hi-ART II helical tomotherapy unit is 10 s (6 rpm) and the typical slice thickness used is 4 mm, however, a small slice width (e.g., 2 mm) could be used for the fine resolution needed for small target volumes. The unit takes ~800 projections (or views) per rotation. For each rotation, two CT slices can be obtained. Pitches of 1, 1.5, and 2 are available, which means that up to 0.8 cm length can be scanned in 10 s. A typical tumor of 8–10 cm length would take as little as 2 min to acquire 25 CT slices. Longer lengths and smaller pitches take more time proportionately. Acquisition occurs on the fly so there is little delay following acquisition for the images to be analyzed. Pixel resolution at the center of the image is dominated by the detector resolution and at the edge of the circle of reconstruction by the number of projections. The reconstructed images shown in Fig. 5 indicate that the Hi-ART II is capable of resolving 1.2–1.6 mm objects near the edge of a 30 cm diameter phantom.

The verification CT uses an ~3.5 MV beam, which means that the photons interact almost exclusively by Compton interactions so that the linear attenuation coefficient is linear with the electron density of the medium (20). Metal artifacts arise in conventional CT scanners because the attenuation of the metal is greatly enhanced due to the photoelectric effect. In helical CT, the beam is penetrating enough to eliminate artifacts arising from metal objects like hip prostheses and dental filings. This means that the

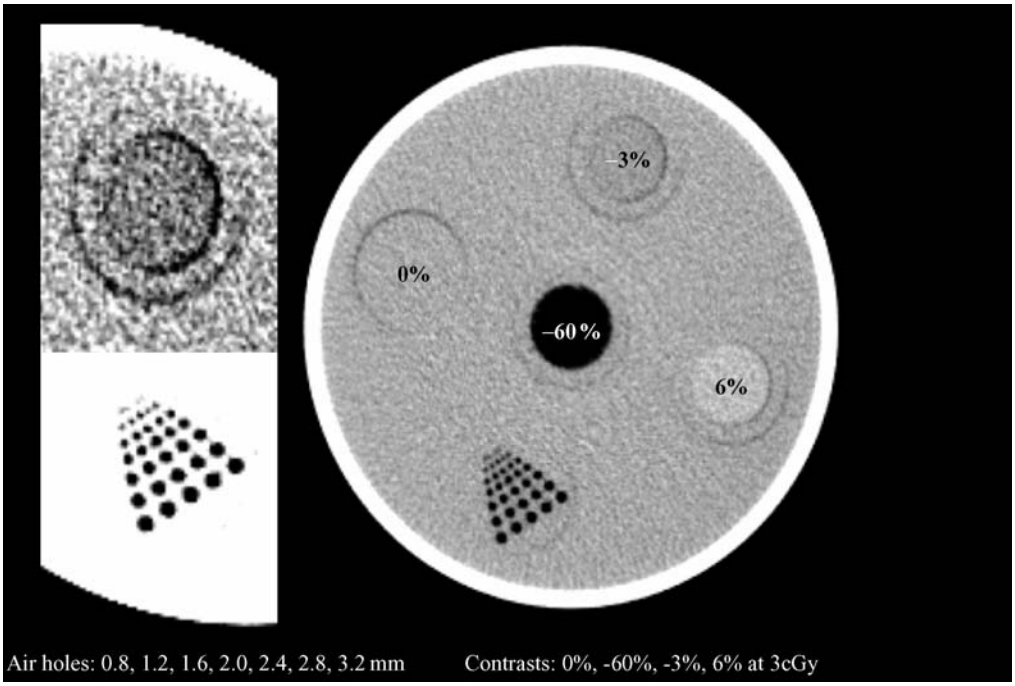


Figure 5. Verification CT at megavoltage (MV) energies of a RMI Solid Water CT phantom. The 3% contrast plug is clearly seen as are the 1.2 mm and 1.6 mm air holes in the solid water phantom. The dose was estimated at 3 cGy at the center of the phantom.

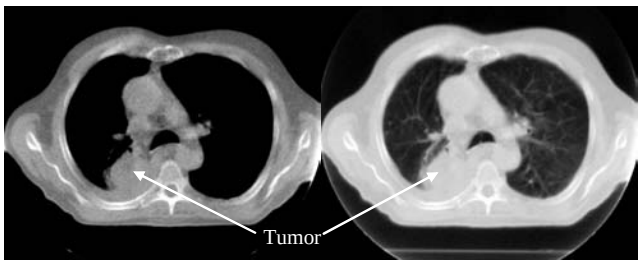


Figure 6. Verification CT of a lung patient. The panel on the left is a soft-tissue window and the panel on the right is a lung window. The difference between muscle and fat and bone and soft tissue is clearly distinguishable. Some of the vascular structures are visible in the lung. The tumor boundary in the lung is discernable, but its extension into the mediastinum is not visible. The dose to the patient was ~3 cGy.

representation supplied by a verification CT is a more reliable CT system for patients with metal implanted appliances. Figure 6 illustrates that bone has less contrast than a conventional CT scan, but it is still clearly discernable on the Hi-ART II unit. The boundary of lung with internal major airways and vascular structures are evident. The boundary between fat and muscle is clearly distinguished, which means that the breast and prostate are discernable. Other organs such as the kidney and bladder are visible. Unlike the highest quality conventional CT scanners, the contrast between white and gray matter in brain is not visible.

The CT imaging is performed prior to each treatment to reduce the possibility of a geometrical miss of the target and sensitive structures. An automated comparison of verification and planning image sets is carried out immediately following image acquisition to guide the adjustment of the patient setup. The patient is assumed to be a rigid object requiring translations and rotations to bring the target anatomy and important sensitive structures into alignment with the treatment plan. The patient is positioned by aligning the patient's skin marks with lasers located outside of the bore of the unit. A sagittal representation of the patient's planning CT is shown on the operator console to aid in selecting the slices to be scanned. A verification scan is taken and reconstructed during the acquisition. The patient is then transported to the same position outside of the gantry bore while the verification image set is fused onto the planning image set and the translation and rotation offsets are reported. Typically the image fusion is first done automatically using a mutual information algorithm (21). Following automated registration, the patient registration can be fine-tuned manually. This allows the operator to take into account, as best as possible, the nonrigid nature of the transformation. Once

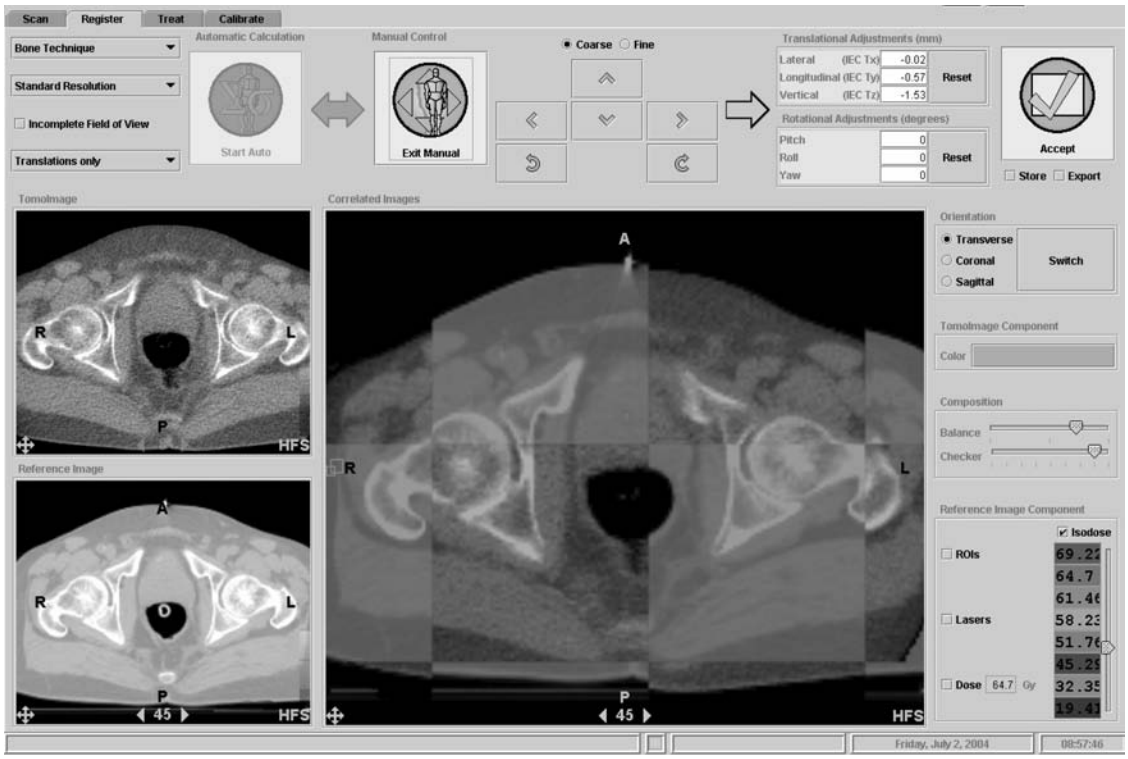


Figure 7. The registration window for a prostate patient. The gray squares are from the planning CT. The verification CT is shown in the upper left and the planning CT is in the lower left. The yellow squares in the large panel are from the tomotherapy verification CT. The rectal boundary and the fat pad surrounding the prostate are clearly aligned on these transverse images. Note that the skin boundary and the leg bones are not as well aligned as the prostate. Regions of interest and the dose distribution obtained from the planning system can be superimposed on the images, but these capabilities have been turned off in this presentation. The translational alignments suggested were 0.0 mm lateral, −0.6 mm longitudinal, and −1.5 mm vertical.

the image registration is completed, the offsets also describe how the patient must be adjusted. Figure 7 shows an example of the graphical user interface for the registration utility being used to register a prostate patient.

If the patient requires adjustment the patient can be translated accordingly. The Hi-ART II CT couch has automated vertical (elevation) and longitudinal translations, and automated gantry start angle adjustments to account for patient roll. The couch top can be manually adjusted in the lateral direction (x direction). Yaw and pitch rotations can be accommodated using angularly calibrated immobilization and positioning aids, which are especially useful for the head and neck. The Hi-ART II includes a set of moveable CT-simulator lasers so that the modified position of the patient can be confirmed.

IMAGE GUIDANCE AND ADAPTIVE RADIOTHERAPY

Adaptation of future treatments from information gleaned from past treatments is a further refinement based on daily CT verification. The daily CT image set forms a model of the patient that can be used to compute an estimate of the dose delivered for that treatment. *Dose reconstruction* is a determination of the dose delivered superimposed on the CT at the time of treatment (6). The CT detector runs at the time of treatment recording the treatment beam exiting through the patient and couch. Using the CT image set acquired just before treatment, the energy fluence incident on the patient can be computed and used to estimate the dose distribution in the patient using the same convolution–superposition method used for helical tomotherapy treatment planning. The total dose delivered up to the current treatment is analyzed for regions of under- and overdosage and these regions reoptimized to bring them inline with the original plan with the corrections applied in one or more future treatments: a process called *adaptive radiotherapy*.

The CT imaging capability of the Hi-ART II has proven extremely useful for reducing daily geometrical misses, which are the most important errors that can compromise treatment efficacy. The increased confidence provided by daily CT setup verification has allowed clinicians to reduce geometrical margins that account for setup uncertainty, and to explore accelerated treatment protocols using a larger dose per fraction than conventional treatments (22). It is expected that CT image guidance technology will facilitate further changes in radiotherapy practice in the coming decade as more facilities adopt this capability.

BIBLIOGRAPHY

1. Mackie TR, Holmes T, Swerdloff S, Reckwerdt P, Deasy JO, Yang J, Paliwal B, Kinsella T. Tomotherapy: A new concept for the delivery of dynamic conformal radiotherapy. Med Phys 1993;20:1709–1719.
2. Carol MP. Peacock: A system for planning and rotational delivery of intensity-modulated fields. Int J Imag Syst Technol 1995;6:56–61.
3. Curran B. Where goest the Peacock? Med Dos 2001;26(1):3–9.
4. Mackie TR, Holmes TW, Reckwerdt PJ, Yang J. Tomotherapy: Optimized planning and delivery or radiation therapy. Int J Imaging Sys Tech 1995;6:43–55.
5. Mackie TR, Kapatoes J, Ruchala K, Lu W, Wu C, Olivera G, Forrest L, Tome W, Welsh J, Jeraj R, Harari P, Reckwerdt P, Paliwal B, Ritter M, Keller H, Fowler J, Mehta M. Image guidance for precise conformal radiotherapy. Int J Radiat Oncol Biol Phys 2003;56(1):89–105.
6. Kapatoes JM, Olivera GH, Balog JP, Keller H, Reckwerdt PJ, Mackie TR. On the accuracy and effectiveness of dose reconstruction for tomotherapy. Phys Med Biol 2001;46:943–966.
7. Brahme Optimization of stationary and moving beam radiation therapy techniques. Rad Oncol 1988;12:129–140.
8. Bortfeld T, Burkelbach J, Boesecke R, Schlegel W. Methods of image reconstruction from projections applied to conformation radiotherapy. Phys Med Biol 1990;35(10):1423–1434.
9. Webb S. Optimization by simulated annealing of three-dimensional conformal treatment planning for radiation fields defined by a multileaf collimator. Phys Med Biol 1991;36(9):1201–1226.
10. Holmes T, Mackie TR, Simpkin D, Reckwerdt P. A unified approach to the optimization of brachytherapy and external beam dosimetry. Int J Radiat Oncol Biol Phys 1991;20 (4):859–873.
11. Holmes TW. A Model for the Physical Optimization of External Beam Radiotherapy, Ph.D. dissertation. Madison (WI): Department of Medical Physics, University of Wisconsin; 1993.
12. Swerdloff S, Mackie TR, Holmes TW. Method and apparatus for radiation therapy. US patent 5,317,616. 1994.
13. Followill D, Geis P, Boyer A. Estimates of whole-body dose equivalent produced by beam intensity modulated conformal therapy. Int J Radiat Oncol Biol Phys 1997;1;38(3):667–672. Erratum in: Int J Radiat Oncol Biol Phys 1997;1;39(3):783.
14. Mutic S, Low DA. Whole-body dose from tomotherapy delivery. Int J Radiat Oncol Biol Phys 1998;1;42(1):229–232.
15. Low DA, Mutic S, Dempsey JF, Markman J, Goddu SM, Purdy JA. Abutment region dosimetry for serial tomotherapy. Int J Radiat Oncol Biol Phys 1999;1;45(1):193–203.
16. Salter BJ, Hevezi JM, Sadeghi A, Fuss M, Herman TS. An oblique capable patient positioning system for sequential tomotherapy. Med Phys 2001;28(12):2475–2488.
17. Kissick MW, Jeraj R, Kapatoes JM, Keller H, Mackie TR, Olivera GH. The thread effect of helical tomotherapy. The XIVth International Conference on the Use of Computers in Radiation Therapy; 2004. p 185–186.
18. Mackie TR, Olivera GH, Kapatoes JM, Ruchala KJ, Balog JP, Tome WA, Hui S, Kissick M, Wu C, Jeraj R, Reckwerdt PJ, Harari P, Ritter M, Forrest L, Welsh J, Mehta MP. Helical tomotherapy. In: Palta J, Mackie TR, editors. Intensity-Modulated Radiation Therapy: The State of the Art. College Park (MD): American Association of Physicists in Medicine; 2003. p 247–284.
19. Mackie TR, Hughes J, Olivera GH, Kapatoes J, Ruchala K, Ramsey C, Kissick M, Jeraj R. The delivery time for helical tomotherapy. The XIVth International Conference on the Use of Computers in Radiation Therapy; 2004. p 750–752.
20. Jeraj R, Mackie TR, Balog J, Olivera G, Pearson D, Kapatoes J, Ruchala K, Reckwerdt P. Radiation characteristics of helical tomotherapy. Med Phys 2004;31(2):396–404.
21. Ruchala KJ, Olivera GH, Kapatoes JM. Limited-data image registration for radiotherapy positioning and verification. Int J Rad Onc Biol Phys 2002;54:592–605.
22. Fowler JF, Tome WA, Fenwick JD, Mehta MP. A challenge to traditional radiation oncology. Int J Radiat Oncol Biol Phys 2004;60(4):1241–1256.

See also COMPUTED TOMOGRAPHY; RADIATION THERAPY, INTENSITY MODULATED; RADIOTHERAPY, THREE-DIMENSIONAL CONFORMAL.

TONOMETRY, ARTERIAL

JOSEPH S. ECKERLE
SRI International
Menlo Park, California

INTRODUCTION

The arterial tonometer is an instrument for measuring arterial blood pressure. It differs from the familiar sphygmomanometer in that, rather than measuring the pressure only at greatest contraction and greatest heart dilation (systolic and diastolic), it provides continuous measurement throughout the heart's pumping cycle. Typically, the instrument sensor is placed over a superficial artery; the radial artery pulse point at the wrist is one convenient site for tonometer measurements. Figure 1 shows how a tonometer sensor would be placed for measurements at this site.

A catheter can be used for accurate, continuous measurement of blood pressure, but the instrument is invasive and numerous risks are associated with its use. In contrast, the noninvasive tonometer can provide an accurate, continuous blood pressure measurement with negligible risk.

Sphygmomanometric instruments of several types are available for noninvasive blood pressure measurements. However, these instruments are generally not capable of continuous blood pressure measurement, nor is their long-term use feasible. The familiar blood pressure cuff hinders venous return and results in peripheral edema. In contrast to sphygmomanometric instruments, the tonometer can be used for beat-by-beat blood pressure measurement over long periods of time with minimal edema. Important disadvantages of the tonometer include its sensitivity to sensor placement and movement artifacts, effects of anatomical variations, and the greater complexity and cost relative to a conventional sphygmomanometer.

In what follows, the physical principles that form the theoretical basis for tonometric blood pressure measurement are first presented; these include techniques for identifying the location of an artery beneath a tonometer sensor and for adjusting the force with which the sensor is pressed against the skin. Next, the technical evolution of tonometer transducers is discussed, leading to a description of modern, multiple-element transducers. This is followed by discussion of several considerations that are unique to tonometric measurements, and may influence the usefulness of the technique for certain applications. Various applications of tonometry, particularly surgical monitoring and cardiovascular evaluation, are discussed next. Finally, the measurement accuracy of tonometry is addressed.

PRINCIPLES OF OPERATION

General Principles

The fundamental principles underlying arterial tonometry are similar to those for ocular tonometry (1,2). Figure 2 shows an idealized model that helps illustrate these principles. In Figure 2, P represents the blood pressure in a superficial artery and F is the force measured by a tonometer transducer. The membrane is the artery wall. Figure 2b is a "free body diagram" showing all the forces and moments acting on the frictionless piston of Fig. 2a. As shown in Fig. 2b, an ideal membrane transmits only a tensile force, T, and does not transmit any bending moment. The tension vector shown, T, is perpendicular to the pressure vector, so the force, F, is independent of T and depends only on the blood pressure and the area of the frictionless piston, A. Following common practice, the integrated effect of arterial pressure acting on the segment of arterial wall is represented by a vector of magnitude PA, oriented perpendicular to the wall. Thus, measurement of the force, F, permits one to directly infer the intraarterial pressure.

Figure 3 shows a superficial artery and a tonometer sensor in cross-section. The tonometer sensor is represented schematically, and is modeled as an assemblage of springs with spring constants, K, as shown. By careful design of the tonometer sensor and selection of an appropriate superficial artery, it is possible to satisfy several conditions:

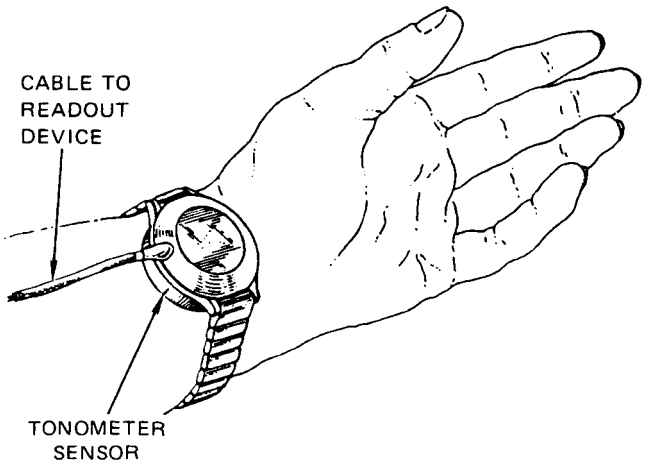


Figure 1. Arterial tonometer sensor on radial artery.

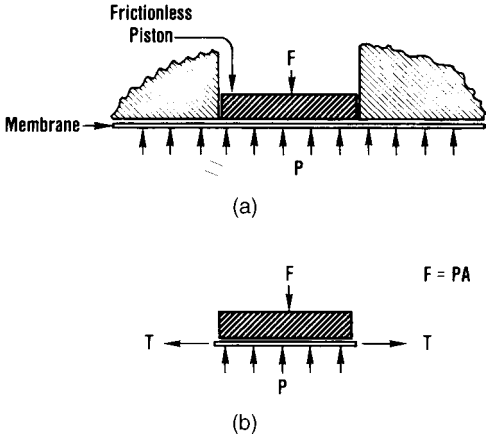


Figure 2. Idealized model for a tonometer.

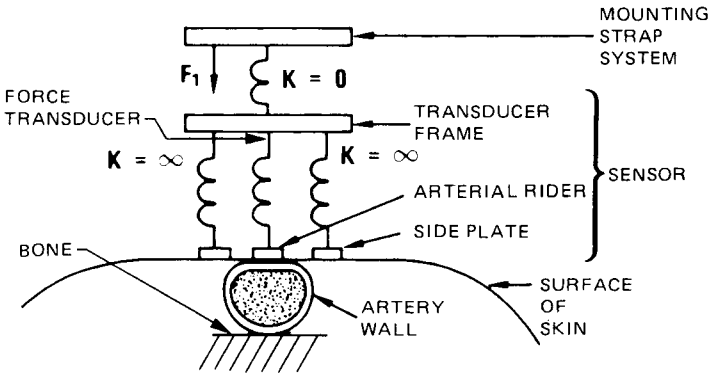


Figure 3. Schematic diagram of a tonometer sensor and a superficial artery.

1. The artery is supported from below by bone (e.g., the radius).
2. The hold-down force, F_1, flattens a portion of the artery wall, but does not occlude the artery.
3. The thickness of the skin over the artery is insignificant, compared to the artery diameter.
4. The artery wall behaves essentially like an ideal membrane.
5. The arterial rider is smaller than the flattened area of the artery, and is centered over the flattened area.
6. The spring constant of the force transducer, K_T, is large compared to the effective spring constant of the artery.

When all these conditions are satisfied, Pressman and Newgard (3) showed theoretically that the conditions of Fig. 2 apply where the arterial rider of Fig. 3 corresponds to the frictionless piston of Fig. 2. Thus, the electrical output signal of the force transducer is directly proportional to the intraarterial blood pressure.

The arterial tonometric measurement depends on the membranelike behavior of the artery wall. Drzewiecki et al. (4) have shown analytically and by experiments with an excised canine femoral artery (5) that the desired behavior can be obtained, provided that the artery is flattened sufficiently. This work provides an important theoretical foundation for the arterial tonometer and helps to explain the observations of several prior *in vivo* studies.

Multiple-Element Sensors

A major practical problem with the simple arterial tonometer of Fig. 3 is the requirement that the arterial rider be precisely placed over the superficial artery. Reliable measurements can be obtained only after painstaking adjustment of the sensor location by a trained operator (6). Apparently, there are differences of opinion concerning the severity of this positioning problem and these are discussed in greater detail below.

To ameliorate this problem, multiple-element tonometer sensors, shown schematically in Fig. 4, have been developed (7–9). The sensor worn by the patient actually consists of a multiplicity of individual sensors. Typically, the sensors are arranged to form a linear array of force transducers and arterial riders. The array need only be positioned with enough precision so that some element of the array is centered over the artery. A computer then automatically selects the sensor element that is correctly positioned over the artery.

One algorithm for selection of the correct element from the multiple-element sensor array exploits two characteristics of the pressure distribution in the vicinity of the artery (8). The first is that the pulse amplitude (i.e., the pressure difference between the systolic and diastolic points on the transducer output waveforms) exhibits a broad maximum over the artery. The algorithm searches for the largest pulse amplitude; the corresponding sensor element will then be within about one artery diameter of the center of the artery. However, this element will, in general, not be precisely centered over the artery.

To identify the centered sensor element more precisely, the algorithm then exploits a second phenomenon, illustrated in Fig. 5. This figure shows a multiple-element tonometer sensor and the underlying, partly compressed artery. For purposes of illustration, assume that the diastolic pressure in the artery is 80 mmHg (10.7 kPa). At the instant of diastole, the pressure measured by each element of the sensor is shown plotted at the top of the figure. Elements 4–6, which all lie over the flattened part of the artery wall, measure the intraarterial pressure [80 mmHg (10.7 kPa)] with good accuracy. However, the pressures measured by elements 2, 3, 7, and 8 are all significantly greater than the intraarterial pressure. This higher pressure can be explained by noting that the artery wall is bent to a very small radius in the regions below the latter elements. As a result, large bending moments are transmitted by the artery wall and are manifested as increased

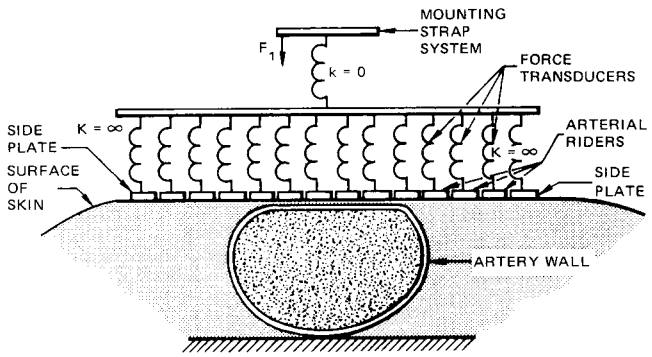


Figure 4. Multiple-element arterial tonometer.

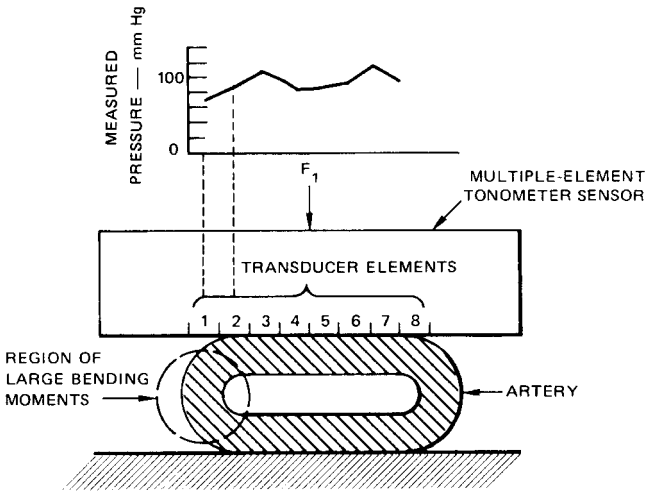


Figure 5. Illustration of pressure distribution near a superficial artery.

pressure on the adjacent sensor elements. A more detailed treatment of this phenomenon can be found in Drzewiecki et al. (4). The element-selection algorithm exploits this phenomenon by searching for a (spatial) local minimum in diastolic pressure (e.g., element 5 of Fig. 5) in a region near the maximum pulse amplitude. The term "local minimum" has a precise meaning in mathematics: When moving in either direction from a local minimum, the value of the function increases. The sensor element corresponding to the local minimum is then assumed to be centered over the artery and blood pressure is measured with this element (10).

Recently, there have been several efforts (11–13) to improve the basic algorithm described above. One algorithm or another might be most effective depending on numerous factors, such as the application, the patient population, and the precision of the sensor and associated amplifiers.

To further ameliorate the positioning problem, Shinoda and others (14,15) have developed motor-driven mechanisms to move a multiple-element sensor laterally within a larger housing strapped to the wrist.

Hold-Down Force

Adjusting the tonometer sensor's transverse location with respect to the artery is not enough; the degree of arterial flattening is also important for accurate tonometric pressure measurement. Arterial flattening depends on the interaction of anatomical factors with the value of the hold-down force, F_1, in Fig. 3. The appropriate value of the hold-down force must be determined for each subject before accurate tonometric measurements can be made. The procedure commonly employed involves increasing (or decreasing) the hold-down force gradually while recording the signal from the tonometer sensor. Figure 6 is an example of such a recording (16). In Fig. 6, hold-down force decreases with time through regions A, B, and C. Region B, where the pulse amplitude is greatest, is considered (3,16,17) to be the region where the most accurate blood pressure measurements are made. This region corresponds to flattening of the artery (as shown in Fig. 5) that is insufficient to cause its occlusion.

Drzewiecki et al. (5) discuss the effects of hold-down force from a theoretical perspective. Eckerle (18) developed algorithms for automatic identification of the center of region B (Fig. 6) and for recognition of difficult subjects (see below) based on various parameters that define this region. Briefly, the algorithm fits a third-order polynomial to the data of Fig. 6. The locations of the regions of Fig. 6 can then be directly computed from the polynomial coefficients. Recently, alternative algorithms (19,20) to control hold-down force have been developed. Again, the optimum algorithm may depend on factors such as application, population, and equipment precision.

TONOMETER SENSOR DESIGN

The size and precision requirements for tonometric blood pressure sensors are severe. Eckerle et al. (21) conclude that, ideally, the arterial rider should be less than ~0.2 mm wide and the associated transducer should be accurate to better than ±2 mmHg (270 Pa). For multiple-element sensors, at least 25 elements with interelement spacings of ~0.2 mm are desirable. As of 1984, these design goals had been approached but not met (21). Then, in 1990, an integrated circuit (IC) based tonometer sensor array was

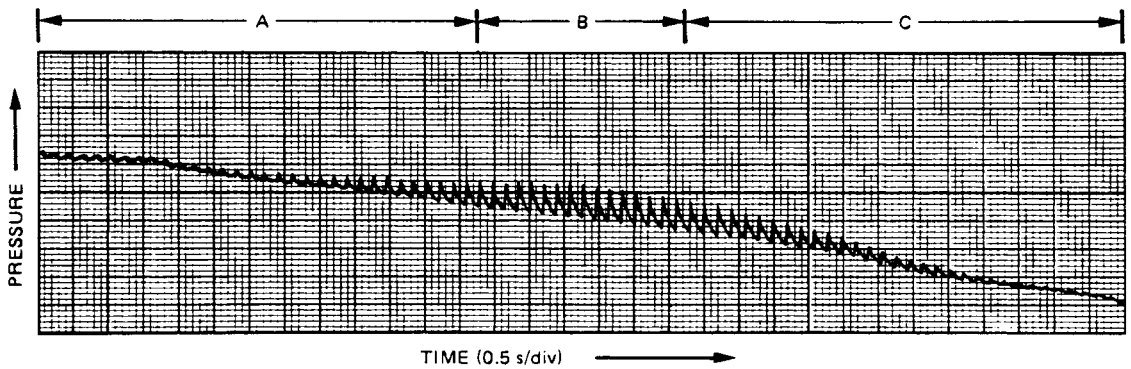


Figure 6. Effect of hold-down force on a tonometer output signal.

reported (9) that substantially achieved these goals. The silicon-and-glass sensor die included 31 tonometer sensor elements in a linear array ~6 mm long.

The first single-element tonometer sensors were constructed of aluminum. Strain gages were attached to miniature beams supporting the arterial rider (3). This approach made effective sensors, but they suffered from the positioning problems previously discussed. Subsequently, various workers devised alternative single-element tonometric sensors. Bigliano et al. (22) reported a pressure sensor with a thin membrane that controlled air flow through a narrow passage. Stein and Blick (23) used a modified myographic force transducer to record blood pressure waveforms from the radial artery. Bahr and Petzke (17,24) used a semiconductor pressure transducer placed over the radial artery. Kelly et al. (6) used a pressure sensor intended for insertion via a catheter. This sensor was mounted in the tip of a pencil-shaped probe, which was then applied to the skin. Borkat et al. (25) placed a pressure capsule, consisting of a rubber bladder attached to a pressure transducer, over the radial artery. Borkat's device is probably the most inexpensive and simple of these alternatives, but it fails to meet the size and stiffness requirements that apply for accurate tonometric measurements.

Multiple-element tonometer sensors have been fabricated from a monolithic silicon substrate using anisotropic etching to define pressure-sensing diaphragms about 10 μm thick in the silicon (26). The IC processing techniques are then used to create piezoresistive strain gages in the diaphragms. External circuitry measures the resistance of the strain gages to determine the pressure exerted on each sensor element. Figure 7 is a photomicrograph of an eight-element sensor fabricated in this way. The diaphragms are square and are arranged in two staggered rows of four each. Note the scale in the figure. The arterial riders in this device are 0.75 × 0.75 mm. Figure 8 is a further magnified view of one element of the Figure 7 sensor. Two radial and two tangential piezoresistive strain gages can be seen together with aluminum metallization used for connection to external circuitry. The performance of these sensors is representative of the best that had been achieved (for tonometers) as of 1984 (21) and is summarized in Table 1.

Achieving the size and accuracy requirements for multiple-element tonometer sensors noted above (21) is difficult, even using the latest advances in IC sensor fabrication. One fundamental problem involves the difficulty of placing independent pressure sensors side by side while minimizing interaction between them. In 1990, Terry et al. (9) reported a clever configuration to address this problem. Briefly, a multiplicity of independent pressure transducers shared a single, long, narrow silicon diaphragm. Some performance parameters of this sensor are also shown in Table 1. This sensor was the first to substantially achieve the size and precision requirements proposed by Eckerle et al. (21).

While the single-diaphragm sensor of Terry et al. (9) has superior performance, relative to the Fig. 7 device, it requires a more complex manufacturing process, and is therefore more expensive. For cost-sensitive applications, a sensor such as Fig. 7 may be preferred. Other groups are investigating fiber-optic transducers (27) and capacitive transducers (28,29) for use in multiple-element tonometer sensors.

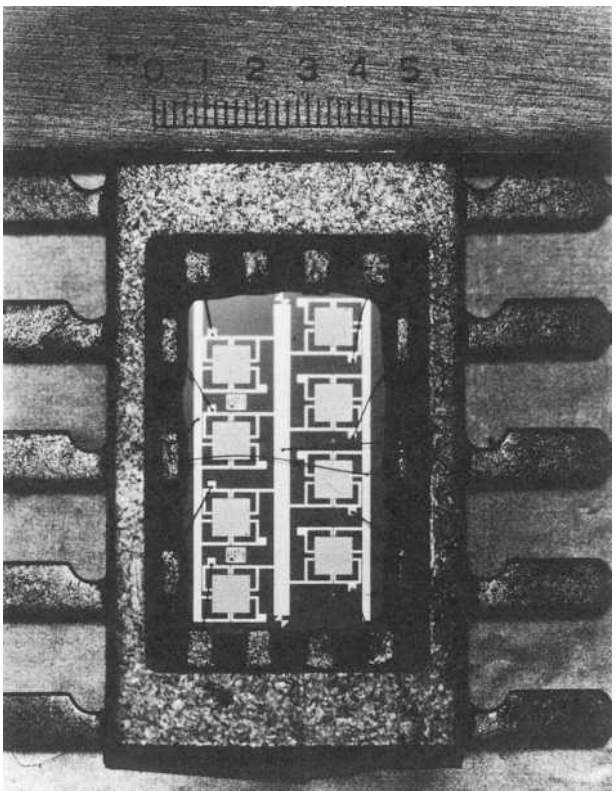


Figure 7. Eight-element tonometer sensor.

Proper mounting of an IC tonometer sensor is very important. The mounting arrangement must protect the fragile sensor while faithfully transmitting the pressure of the patient's skin to the sensor. Consideration should be given to measurement drift caused by thermal effects or material creep. Finally, the shape of the mounted sensor can be chosen to conform to the local anatomy (e.g., the nearby radius in the case of a radial artery tonometer). Fujikawa and Harada (30) and others (31,32) developed

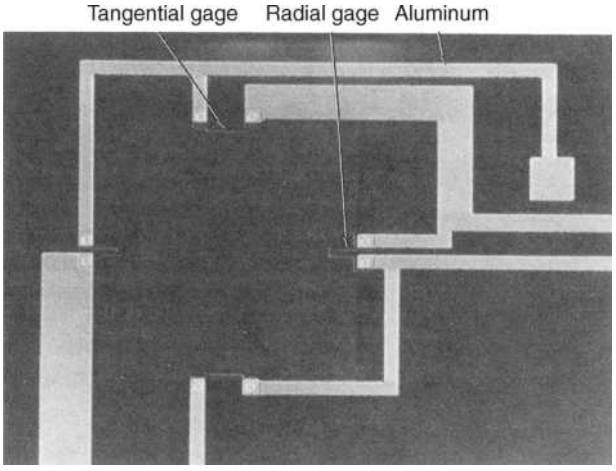


Figure 8. Pressure-sensitive diaphragm of tonometer sensor.

Table 1. Typical Performance of Silicon Tonometer Sensors

Parameter	Typical Value[a] Ref. 26	Typical Value[a] Ref. 9	Units
Number of elements	8	31	Each
Element spacing	0.75	0.2	mm
Sensitivity	25	25–50	μV/V mmHg
Nonlinearity	3	n.r.	mmHg
Temperature coefficient of sensitivity	−0.25	n.r.	%·°C^{-1}
Temperature coefficient of offset (uncompensated)	1	n.r.	mmHg·°C^{-1}
Offset error due to temperature (with temperature compensation)	±1 (estimate)	n.r.	mmHg
Frequency response (flat to <1.0 dB)	>50	n.r.	Hz
Noise	<0.5	n.r.	mmHg

[a]n.r. = not reported.

methods for mounting IC sensors for tonometry in clinical environments.

A clinical tonometer instrument, incorporating many of the features described above, is shown in Fig. 9. Comparable instruments may be obtained from suppliers such as Colin Medical Technology Corp., Komaki, Japan; Hypertension Diagnostics, Inc., Eagan, Minnesota; and AtCor Medical, Sydney, Australia.

MISCELLANEOUS CONSIDERATIONS

One significant advantage of the arterial tonometer is its ability to make noninvasive, nonpainful, continuous measurements for long periods of time. There are several reasons for this superiority over the sphygmomanometer:

1. The sensor is adjusted to partly flatten, but not occlude the radial artery. In contrast, sphygmomanometric systems typically occlude the artery in order to determine systolic pressure.
2. The part of the tonometer sensor representing the arterial riders and side plates of Fig. 3 is ∼20 mm in diameter. There may be occlusion of some veins beneath this part of the sensor, but numerous parallel venous paths in the remainder of the wrist allow return venous flow. [The reader may wish to try the following experiment: Place a dime (a U.S. coin ∼18 mm in diameter) over your radial artery pulse. Holding your wrist with your free hand, press down on the dime with the thumb to occlude the radial artery. You are now experiencing greater venous occlusion than a tonometer sensor would cause.]
3. There are several venous paths through the wrist. Some of these pass between bones that apparently help protect them from pressure that may be applied externally to the wrist by the mounting strap and other parts of the tonometer sensor.

These factors all help to minimize development of edema distal to a tonometer sensor. Discomfort and development of edema will depend on factors, such as subject-to-subject

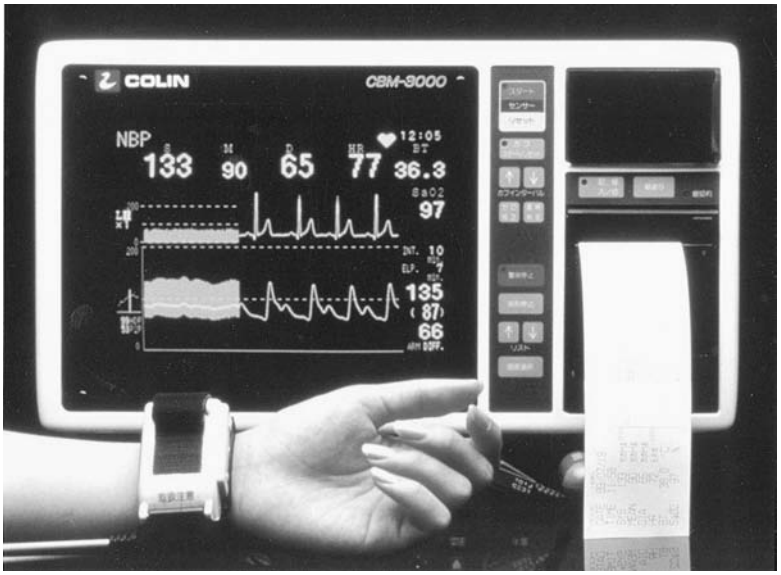


Figure 9. Clinical tonometer instrument.

variations in wrist anatomy, measurement duration, and the hold-down force used. In a study involving 20 conscious, healthy subjects (33) the blood pressure waveform was monitored continuously by tonometry for more than 30 min. The subjects reported no significant discomfort due to the tonometer sensor.

The tonometer's ability to measure blood pressure continuously with minimal distraction to the subject makes it attractive for ambulatory blood pressure monitoring. This application should be feasible in the future, but present instruments do not have sufficient artifact-rejection capability for reliable measurements on ambulatory subjects (21). This is not to say that the arterial tonometer is unusually sensitive to movement artifacts. In a study involving 20 subjects (34), the tonometer was found to be less sensitive to subject movement than photoplethysmographic, quadrapolar impedance plethysmographic, and sphygmomanometric sensors.

Eckerle et al. (35) have taken the first step toward an ambulatory tonometer for blood pressure by developing an ambulatory pulse rate sensor using tonometry. By use of a curved spring, the sensor can be mounted unobtrusively in a watchband <8 mm thick.

Taking an analog signal-processing viewpoint, the blood pressure can be described as the sum of an alternating current (ac) component and a direct current (dc) component. Briefly, the ac component is more easily measured by tonometry than the dc component. More specifically, under clinical conditions, the six conditions listed above may not always be satisfied. When this happens, the tonometer sensor often continues to measure the ac component with good fidelity, while the measurement accuracy of the dc component (on which the familiar systolic and diastolic depend) becomes degraded. Shinoda and others (36–38) have used sphygmomanometric measurements of blood pressure to correct errors in the dc component and thereby produce an output with good waveform fidelity (ac component) and accurate measurement of systolic and diastolic (dc component).

The radial artery is not the only site at which a tonometer may be applied, but most research to date has used this site. Other sites suitable for tonometric measurements include the brachial artery at the inner elbow (the antecubital fossa), the temporal artery in front of the ear, the carotid, and the dorsalis pedis artery on the upper foot.

Anatomical variations of the wrist can make the location and support of the radial artery unsuitable for tonometric pressure measurement. Unsuitability occurs in only a small fraction of the population and such difficult subjects can be recognized by the computer used with multiple-element sensors. In one study involving 6 subjects (7,8) there was one difficult subject, while in another involving 20 subjects (33) there were none.

As described above, a single-element tonometer sensor, such as Fig. 3, can be used for tonometric blood pressure measurement if six conditions are met. The multiple-element tonometer was developed to help simplify tonometric measurements for the clinician (relative to a single-element instrument). Many clinicians (39–45) have reported using multiple-element instruments. Many other clinicians (6,46–54) have used single-element instruments. Chen et al. (44) have used both types. They observe that "Although [the hand-held, single-element tonometer]. . .was probably adequate for brief steady-state data, manual recording was too unstable for accurate pressure tracking during hemodynamic transients, and it introduced an element of user dependence and thus potential bias to the data. The automated [multiple-element tonometer] system circumvented these limitations." We may also observe that the two groups are generally pursuing different applications, suggesting that the choice of optimum sensor type is application dependent.

APPLICATIONS

Perhaps the first clinical application of tonometry was for blood pressure monitoring in surgery and other procedures. Stein and Blick (23) used radial artery tonometry during a catheterization procedure. Kemmotsu et al. (39) used a multiple-element tonometer for surgical monitoring. Several others (40–43,55) have also evaluated the instrument under various surgical and postsurgical situations.

Figure 10 shows a typical blood pressure waveform, obtained with an arterial tonometer from the radial artery of an adult male. Note that systolic and diastolic pressures, as well as pulse rate and dicrotic notch information, can be obtained from the waveform. The subject performed a Valsalva maneuver at the time indicated. The attendant changes in blood pressure and pulse rate are quite clearly indicated. Of course the tonometer measures only the pressure in the underlying artery, not central arterial

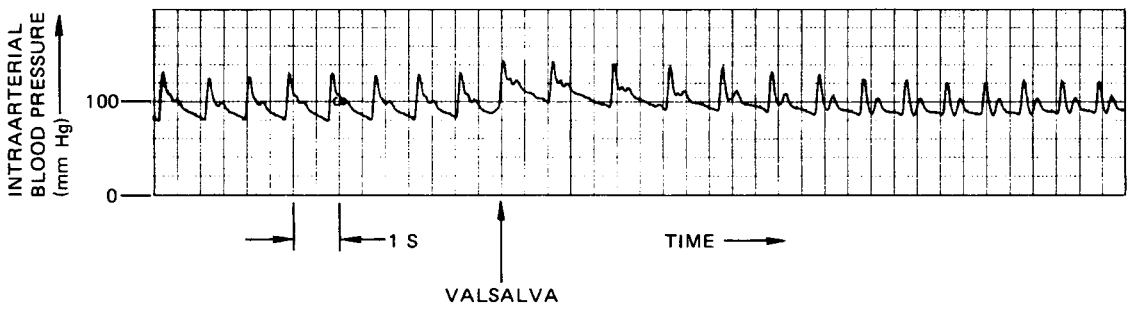


Figure 10. Intraarterial blood pressure waveform.

(aortic) pressure. Central aortic pressure is useful for clinical applications, so there has arisen a desire to calculate central aortic pressure based on a radial pressure obtained with a tonometer.

A brief discussion of circulation dynamics may now help reveal the motivation for some other tonometer applications. Physiologists have long struggled to devise a useful physical model for the circulatory system. Briefly, two complementary models have arisen, a windkessel model and a pulse wave velocity (PWV) model (56,57). In the windkessel model, the circulatory system is represented (often using an electronic analogy) as a small number of lumped elements such as capacitors, inductors, and resistors. In the PWV model, the circulatory system is represented as one or more pipes in which a pulse wave, generated by the heart, travels at a certain velocity. Physiologists and clinicians seek to determine the values of the lumped elements and the magnitude of the PWV by means of noninvasive measurements. Various aspects of the ac component of the pressure waveform of Fig. 10 can be used to estimate the windkessel parameters and PWV.

Focusing on the windkessel model, one can devise a transfer function that relates the pressure at a peripheral site (e.g., the radial) to an input pressure waveform in the aorta (58,44). It is then straightforward (in theory) to compute the inverse of this transfer function. Using this inverse transfer function, one may then calculate the pressure waveform in the aorta based on the peripheral waveform. This technique is potentially very powerful because it allows making a previously invasive measurement (the aortic pressure waveform) by using a noninvasive instrument: a radial artery tonometer.

Several investigators (44,46,58) have evaluated the above technique to determine aortic pressure by mathematical manipulation of a radial artery pressure determined by tonometry. Others (47,59) have evaluated a similar technique to determine aortic pressure from a tonometric carotid pressure measurement.

Numerous investigators (48–51) have used tonometry to estimate windkessel parameters, particularly arterial elasticity. Other investigators (45,52–54) have used it to estimate PWV.

For certain applications, it is sufficient to measure relative, rather than absolute, blood pressure. These include polygraph (lie detector) tests and studies of the physiological effects of various cognitive and physical stressors (e.g., weightlessness and high -*g* aircraft maneuvers). A multiple-element tonometer designed specifically for relative blood pressure measurement has been developed by Eckerle et al. (33) and tested extensively on 20 subjects of both sexes. Subjects were subjected to stressors such as mental arithmetic and cold pressor, and their blood pressures were recorded continuously for periods in excess of 30 min.

MEASUREMENT ACCURACY

Because the accuracy of tonometric blood pressure measurements depends on the size and positioning of the sensor, the value of hold-down force, and the accuracy of the sensor itself, some variability in reported tonometer accuracy can be expected when these factors are not well controlled. In nearly all of the measurements reported prior to 1987 hold-down force was adjusted manually, and in many the tonometer was manually positioned as well. Some improvement in tonometer accuracy and repeatability can be expected from automatic sensor positioning and hold-down force adjustment.

An early validation of the principles of tonometry (60) involved comparison of pressures measured in an exposed canine femoral artery by both a tonometer and an intraarterial catheter. One series of tests involved 15 animals with blood pressure changes being induced by drug injection and vagal stimulation. For this entire series of tests, the difference between pressures measured by the two instruments was never $> 5\%$.

Stein and Blick (23) compared tonometric measurements with direct arterial blood pressures on 20 patients undergoing cardiac catheterization. They found that the tonometer reproduced both the intraarterial waveform and any abrupt changes in pressure caused by various interventions with remarkable fidelity. Specifically, during interventions that increased or decreased systolic pressures by as much as 30%, the tonometric pressure measurement followed the direct arterial pressure within 10% in 92% of the observations.

Weaver et al. (8) compared sphygmomanometric blood pressures with those obtained with a multiple-element tonometer sensor on five subjects of both sexes ranging in weight from 135 to 225 lb (61–102 kg). The standard deviation between the two measurements of systolic and diastolic pressures was 6.5 mmHg (870 Pa), indicating that the tonometer was at least as accurate as the sphygmomanometer.

In 1989, Kemmotsu et al. (39) reported the accuracy of an automated, multiple-element tonometer instrument used for monitoring during surgery. The tonometric measurements showed good correlation ($r = 0.94$, $P < 0.001$ for systolic) with invasive measurements on the contralateral radial artery. Kemmotsu subsequently performed additional studies (40,41) involving surgical monitoring. In the later study (41), involving 60 patients, the mean absolute value of error ranged from 3.6 to 6.6 mmHg (480 to 880 Pa). Standard deviations ranged from 4.5 to 6.2 mmHg (600 to 830 Pa).

Other investigators (42,43,55) have evaluated multiple-element tonometers for surgical and postsurgical monitoring. They found standard deviations ranging from 1.7 (for a single patient) to 14.2 mmHg (230 to 1960 Pa).

Sato et al. (61) evaluated the accuracy of a multiple-element tonometer instrument on conscious subjects subjected to a Valsalva maneuver and a tilting test. This study is of particular interest because they directly addressed the question of frequency response of the tonometer/tissue/artery system. Good frequency response is important for computation of windkessel parameters or PWV as described above. Sato et al. found the frequency response of the tonometer/patient system to be essentially flat from 0 to 5 Hz.

BIBLIOGRAPHY

1. Mackay RS, Marg E. Fast, automatic electronic tonometers based on an exact theory. Acto Ophthalml 1959;37:495–507.
2. Mackay RS. Fast automatic ocular pressure measurement based on an exact theory. IRE Trans Med Electron 1960; ME-7:61–67.
3. Pressman G, Newgard P. A transducer for the continuous external measurement of arterial blood pressure. IEEE Trans Bio-Med Electron 1963;BME-10:73–81.
4. Drzewiecki GM, Melbin J, Noordergraaf A. Arterial tonometry: Review and analysis. J Biomech 1983;16:141–152.
5. Drzewiecki GM, Melbin J, Noordergraaf A. Deformational forces in arterial tonometry. Proc Annu Conf IEEE/Eng Med Biol Soc, 6th 1984;26:642–645.
6. Kelly R, Hayward C, Avolio A, O'Rourke M. Noninvasive determination of age-related changes in the human arterial pulse. Circulation 1989;80:1652–1659.
7. Eckerle JS, Newgard PM, A non-invasive transducer for the continuous measurement of arterial blood pressure. Proc Annu Conf Eng Med Biol 1976;29:98.
8. Weaver CS, et al. A study of non-invasive blood pressure measurement techniques. Noninvasive Cardiovasc Meas 1978;167:89–105. Prepared for a conference sponsored by NASA Office of Technology Utilization and Cardiology Division, Stanford University School of Medicine. Society of Photo–Optical Instrumentation Engineers, Bellingham (WA).
9. Terry S, Eckerle JS, Kornbluh RD, Ablow CM. Silicon pressure transducer arrays for blood-pressure measurement. Sensors Actuators 1990;A 21–A 23:1070–1079.
10. Eckerle JS. Noninvasive blood pressure monitoring transducer. US Patent 4,269,193. 1981.
11. Niwa M. Pulse wave detecting apparatus. US patent 5,103,831. 1992.
12. Kobayashi I. Pulse wave detecting apparatus. US patent 5,170,796. 1992.
13. Takaya M. Continuous blood-pressure monitor apparatus. US patent 6,394,959 B1. 2002.
14. Shinoda M, Ogletree WA. Pulse wave detecting apparatus. US patent 4,784,152. 1988.
15. Kaida N, et al. Pulse wave detecting apparatus. US patent 4,901,733. 1990.
16. Eckerle JS. SRI Tonometer Blood Pressure Measurement System, Operator's Instruction Manual. Final Rep., SRI Proj. No. 4908, Washington (DC): Lab. Div., Federal Bureau of Investigation; 1984.
17. Petzke JC, Bahr DE. Blood pressure measuring apparatus. US patent 3,926,179. 1975 Dec 16.
18. Eckerle JS. Blood pressure monitoring method and apparatus. US patent 4,799,491. 1989.
19. Niwa M. Pulse wave detecting apparatus. US patent 5,119,822. 1992.
20. Ohmori K, Narimatsu K, Kobayashi I. Blood pressure monitoring apparatus. US patent 5,590,661. 1997.
21. Eckerle JS, Fredrick J, Jeuck P. Toward a practical tonometric blood pressure instrument. In: Semmlow JL, Welkowitz W, editors. Proceedings of the Annual Confrence LEEE/English Medical Biology Society 6th; 1984. p. 635–641.
22. Bigliano RP, Molner SF, Sweeney LM. A new physiological pressure sensor. Proc Ann Conf Eng Med Biol 1964;6:82.
23. Stein PD, Blick EF. Arterial tonometry for the atraumatic measurement of arterial blood pressure. J Appl-Physiol 1971;30:593–596.
24. Bahr DE, Petzke JC. The automatic arterial tonometer. Proc Annu Conf Eng Med Biol 1973;15:259.
25. Borkat FR, Kataoka RW, Silva J. An approach to the continuous non-invasive measurement of blood pressure. Proceeding of the San Diego Biomedical Symposium; 1976.
26. Weaver CS, et al. 1976: Wearable Blood Pressure and ECG Recording System. (Grant HL 17604-01A1), Interim Report to the National Heart and Lung Institute, Bethesda (MD); 1976.
27. Drzewiecki GM, Butterfield RD, and Ciaccio EJ, Pressure waveform monitor. US patent 5,363,855. 1994.
28. Guckel H, Burns DW. Planar processed polysilicon sealed cavities for pressure transducer arrays. Proc IEEE Int Electron Devices Meet 1984.
29. Corcuera M, Aravamudhan S, Bhansali S. (No Date). A non-invasive microsystem for blood pressure measurement, Research Experience for Undergraduates, Spring 2004 Symposium, Poster No. EE.3. University of South Florida. [Online]. Available at www.eng.usf.edu/∼schlaf/REU/Symposium/Spring2004/Corcuera2004.pdf Accessed 2005April 8.
30. Fujikawa K, Harada C. Semiconductor pressure pulse wave system. US patent 5,101,829. 1992.
31. Harada C, et al. Contact pressure sensor. US patent 5,179,956. 1993.
32. Narimatsu K, Kawamura N. Pressure wave sensor. US patent 5,467,771. 1995.
33. Eckerle JS, et al. Cardiovascular Activity Monitoring for Polygraph Examination, SRI Final Rep., Contract J-FBI-82-108, Washington (DC): Federal Bureau of Investigation; 1984.
34. Eckerle JS, The arterial tonometer: A non-invasive blood pressure instrument for trauma victims. Non-Invasive Neurologic Evaluation of the Combat Casualty Victim, Bethesda (MD) Naval Medical Research and Development Command; 1985.
35. Eckerle JS, et al. Pulse rate sensor system. US patent 5,243,992. 1993.
36. Shinoda M, Lippincott HW. Continuous blood pressure monitoring system having a digital cuff calibration system and method. US patent 5,165,416. 1992.
37. Aung Y, Takaya M, Nishibayashi H. Blood pressure monitor system. US patent 5,261,414. 1993.
38. Kawamura N, Nakagawa T, Aung Y. Blood pressure monitor system. US patent 5,279,303. 1994.
39. Kemmotsu O, et al. A non-invasive blood pressure monitor based on arterial tonometry. Anesthes Analges (Suppl.) 1989;68:S145.
40. Kemmotsu O, et al. Blood pressure measurement by arterial tonometry in controlled hypotension. Anesthes Analges 1991; 73:54–58.
41. Kemmotsu O, et al. Arterial tonometry for non-invasive, continuous blood pressure monitoring during anesthesia. Anesthesiology 1991;75(2):333–340.
42. Searle NR, Perrault J, Ste-Marie H, Dupont C. Assessment of the arterial tonometer (N-CAT) for the continuous blood pressure monitoring in atrial fibrillation. Can J Anaesthesiol 1993;40(4):388–393.
43. Siegel LC, Brock-Utne JG, Brodsky JB. Comparison of arterial tonometry with radial artery catheter measurements of blood pressure in anesthetized patients. Anesthesiology 1994;81(3): 578–584.
44. Chen C, et al. Estimation of central aortic pressure waveform by mathematical transformation of radial tonometry pressure validation of generalized transfer function. Circulation 1997; 95:1827–1836.
45. Narimatsu K, Takatani S, Ohmori K. A multi-element carotid tonometry sensor for non-invasive measurement of pulse wave velocity. Frontiers Med Biol Eng 2001;11(1):45–58.

46. Adji A, O'Rourke MF. Determination of central aortic systolic and pulse pressure from the radial artery pressure waveform. Blood Pressure Monitoring 2004;9:115–121.
47. Chen C, et al. Validation of carotid artery tonometry as a means of estimating augmentation index of ascending aortic pressure. Hypertension 1996;27:168–175.
48. Collins VR, Finkelstein SM, Cohn JN. Evaluation of pulse contour technique for measuring arterial elasticity. Circulation 1980;62(Suppl. II):1111–1120.
49. Tanaka H, et al. Aging, habitual exercise, and dynamic arterial compliance. Circulation 2000;102:1270–1275.
50. Zimmerman A, et al. Loss of oscillatory arterial compliance is detectable in young patients by radial artery pulse contour analysis. Am J Hypertension April 2000;13: (4, Part 2) Abstract No. B026.
51. Zimlichman R, et al. The seven European sites study of arterial elasticity—using the blood pressure waveform analysis—reliability, repeatability and establishment of normal values for healthy European population with comparison to healthy US population. Am J Hypertension May 2003;16: (No. 5, Part 2):P–315.
52. Blacher J, et al. Aortic pulse wave velocity as a marker of cardiovascular risk in hypertensive patients. Hypertension 1999;33:1111–1117.
53. Lacy PS, et al. Increased pulse wave velocity is not associated with elevated augmentation index in patients with diabetes. J Hypertension 2004;22:1937–1944.
54. Salvi P, et al. Validation of a new non-invasive portable tonometer for determining arterial blood pressure wave and pulse wave velocity: The PulsePen device. J Hypertension 2004;22:2285–2293.
55. Weiss BM, et al. Radial artery tonometry: Moderately accurate but unpredictable technique of continuous non-invasive arterial pressure measurement. Br J Anaesthesiol 1996; 76(3):405–411.
56. Toy S, Melbin J, Noordergraaf A. Reduced models of arterial systems. IEEE Tran. Biomed Eng 1985;BME-32(2).
57. Quick CM, Berger DS, Noordergraaf A. Apparent arterial compliance. Am J Physiol—Heart 1998;274:1393–1403.
58. Karamanoglu M, O'Rourke MF, Avolio AP, Kelley RP. An analysis of the relationship between central aortic and peripheral upper limb pressure waves in man. Eur Heart J 1993; 14:160–167.
59. Stergiopulos N, Westerhof BF, Westerhof N. Physical basis of pressure transfer from periphery to aorta: A model-based study. Am J Physiol 1998;274(Heart Circ. Physiol. 43): H1386–H1392.
60. Pressman G, Newgard P. Development of a Blood-Pressure Transducer for the Temporal Artery NASA CR-293, Contract NAS 2-1332, Menlo Park (CA): Stanford Research Institute; 1965.
61. Sato T, Nishinaga M, Ozawa T, Takatsuji H. Accuracy of continuous blood pressure monitor based on arterial tonometry. Hypertension 1993;21:866–874.

Further Reading

Geddes LA. The Direct and Indirect Measurement of Blood Pressure. Chicago: Year Book Medical Publishers, Inc.; 1970.
O'Rourke MF, Kelly RP, Avolio AP. The Arterial Pulse. London: Lea & Febiger; 1992.
Nichols WF, O'Rourke MF, Hartley C. McDonald's Blood Flow in Arteries. 4th ed. New York: Oxford University Press; 1998.

See also BLOOD PRESSURE MEASUREMENT; IMPEDANCE PLETHYSMOGRAPHY; PERIPHERAL VASCULAR NONINVASIVE MEASUREMENTS.

TOOTH AND JAW, BIOMECHANICS OF

SINAN MÜFTÜ
Northeastern University
Boston, Massachusetts

ALI MÜFTÜ
Tufts University
Boston, Massachusetts

INTRODUCTION

The relations among forces, motion, and deformation are studied in the field of mechanics. Biomechanics seeks to understand the mechanics of living systems (1). The teeth and jaw perform the critical function of initiating the digestion by breaking the food into smaller sizes. This increases the surface area of the food and improves the effectiveness of the enzymes involved in digestion. *Mastication* is defined as the action of chewing foods. The masticatory system is composed of

1. The dention
2. The bones (the maxilla the mandible, and the temporal bone),
3. The ligaments
4. The muscles
5. The temporomandibular joint (TMJ):

The periodontal ligament (PDL), which attaches the teeth to the bones, the articular disk, and the cartilage, which are located on the articulating surfaces of the TMJ, and the other soft tissues such as the blood vessels and the nerves are also parts of this complex. The three main functions of the masticatory system are chewing, swallowing, and speech. This system also helps in expression of emotions and respiration. The temporomandibular system is controlled by the nervous system, and its successful function requires a harmonious relation of these components. Biomechanical investigations of the mastication system aim to understand the fundamental relations between anatomy and function and thus either aid the available treatment modes or help design new ones.

During function (mastication) or rest, various components of the temporomandibular complex work together. Explanation of the complex interrelations between the components of the masticatory system, requires knowledge of the functions of each component. A comprehensive review of the biomechanics of tooth and jaw was given in Reference 2 up to 1988. The aim of the current article is to review the recent developments in this field. In the last few years, the cost and the speed of performing analysis using computational techniques has improved dramatically (3). Thus, many new investigations were enabled particularly by, but not limited to, using the finite element method. This article reflects some of these new analyses. In general, this article is divided into three sections related to the biomechanics aspects of the anatomy, function, and treatments of the masticatory system.

The article starts with an overview of the functional anatomy of the mastication system. Brief anatomical

descriptions of the dentition, the skeletal components, the musculature, the temporomandibular joint/disk, and the connective tissues, such as the ligaments, are given. The monographs given in References 4 and 5 can be consulted for more detailed information.

Next, the biomechanical models of mastication are introduced. This involves investigating the relations between the forces applied by the muscles, and the reaction forces that develop on the teeth, the TMJ, and the ligaments, as well as the deformations of the mandible. In-depth understanding of these relations help treatment of TMJ disorders (such as bruxism), design of endosseous implants, reconstructive prosthetics, and various tooth replacement modalities. Definitions of key engineering concepts such as external and internal forces, internal stresses and strains, and elastic material (stress-strain) behavior are provided. The material properties of all components of the masticatory system are briefly reviewed. The biomechanical fundamentals of mastication are introduced by using force balance relationships, where the mandible is treated as a rigid body. In addition, investigating the deformations of the mandible is briefly covered. The monographs given in References 3,4 and (6–11) are only some of the many references that review these topics in more detail.

Many treatment modalities involving the masticatory system benefit significantly from the studies performed in the fields of biomechanics and biomaterials. These include, but are not limited to, prosthodontic and orthodontic treatments, reconstructive surgery, and reconstructive prosthetics. In this article, the biomechanical considerations in prosthetic dentistry, including dental implant treatments, intracoronal and extracoronal restorations, and fixed and partial dentures, are reviewed.

DENTITION AND SUPPORTIVE STRUCTURES

Human dentition is composed of a total of 32 teeth (Figs. 1 and 2). Each tooth has a part that is visible, called the *crown*, and a part located inside the bone called the *root* as shown in Fig. 3. The teeth are attached to the bone by a soft tissue called the PDL, which serves to connect the tooth to the surrounding bone and distributes the occlusal loads to the bony tissue. On the superior aspect (top portion) of the masticatory system, 16 teeth are attached to the maxillary arch. Likewise, on the inferior aspect (bottom portion), 16 teeth are attached to the mandibular arch. As the sizes of the maxillary teeth are larger, the teeth in the maxilla are situated on a larger arch.

The teeth on each jaw can be classified in four groups, according to their function as shown in Fig. 2. The *incisors* are used to cut the food into smaller pieces. They are located on the anterior (frontal) section of the mouth. The incisors have a relatively small cross-section; there-

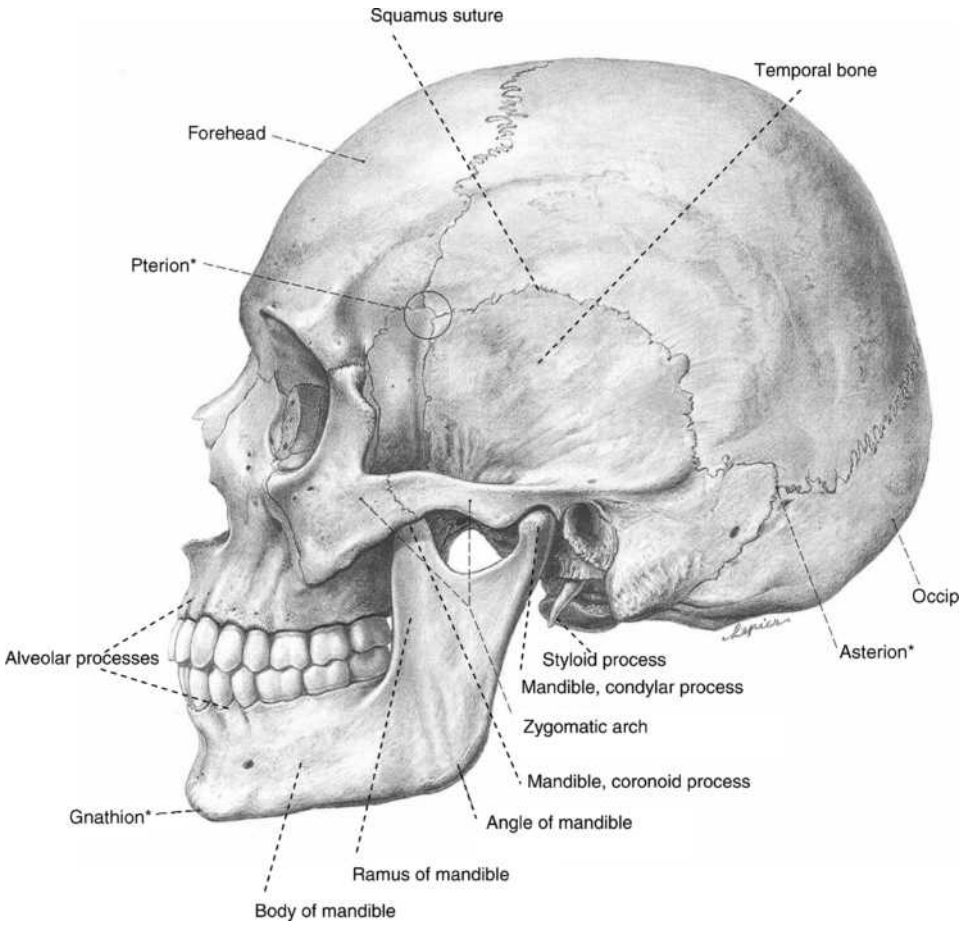


Figure 1. Lateral view of the skull showing the skeletal components of the masticatory system. (From Sobotta Atlas of Human Anatomy, Vol. 1, 13th ed.).

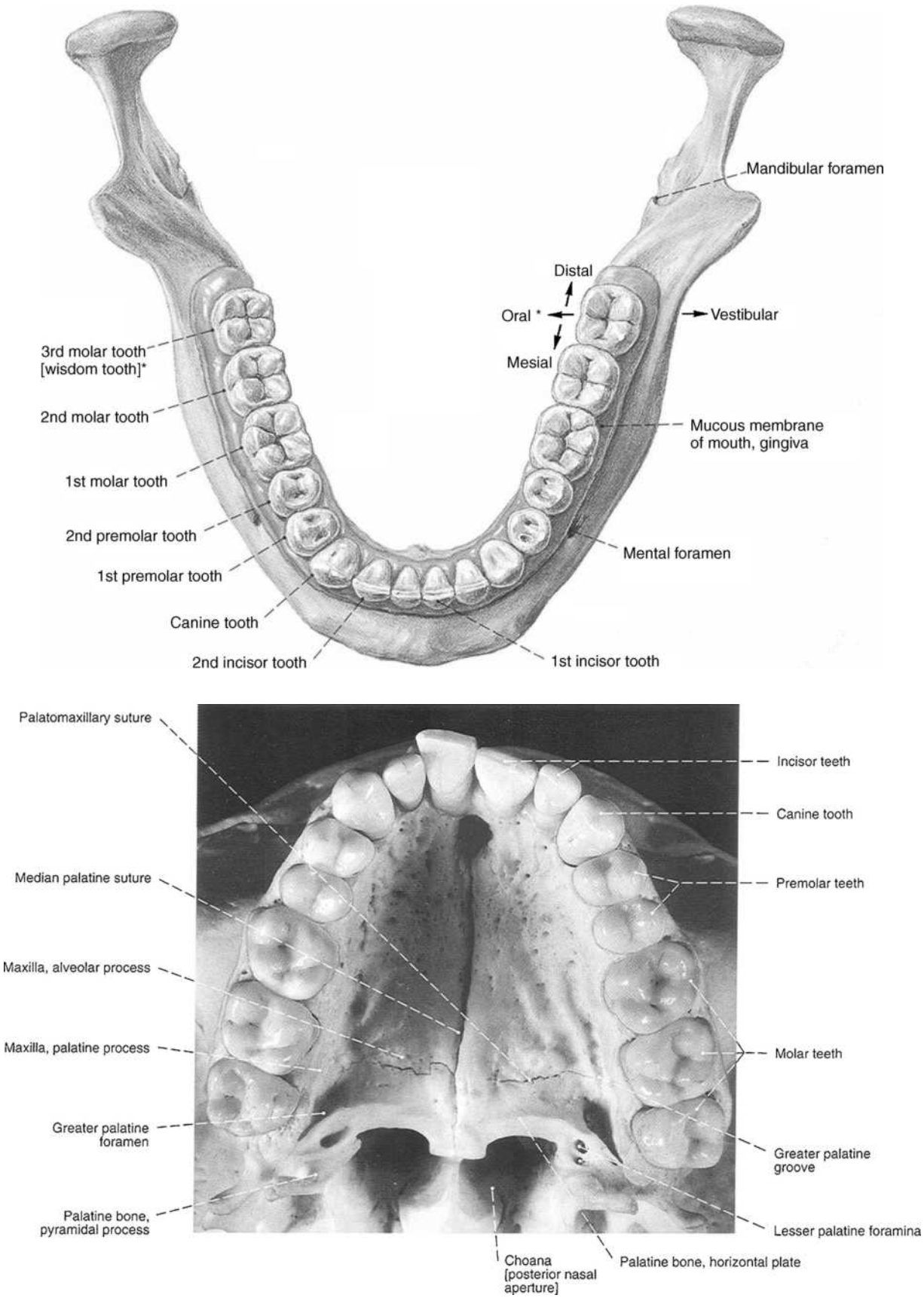


Figure 2. (a) The mandibular dental arch and (b) the maxillary dental arch. (From Sobotta Atlas of Human Anatomy Vol. 1, 13th ed.).

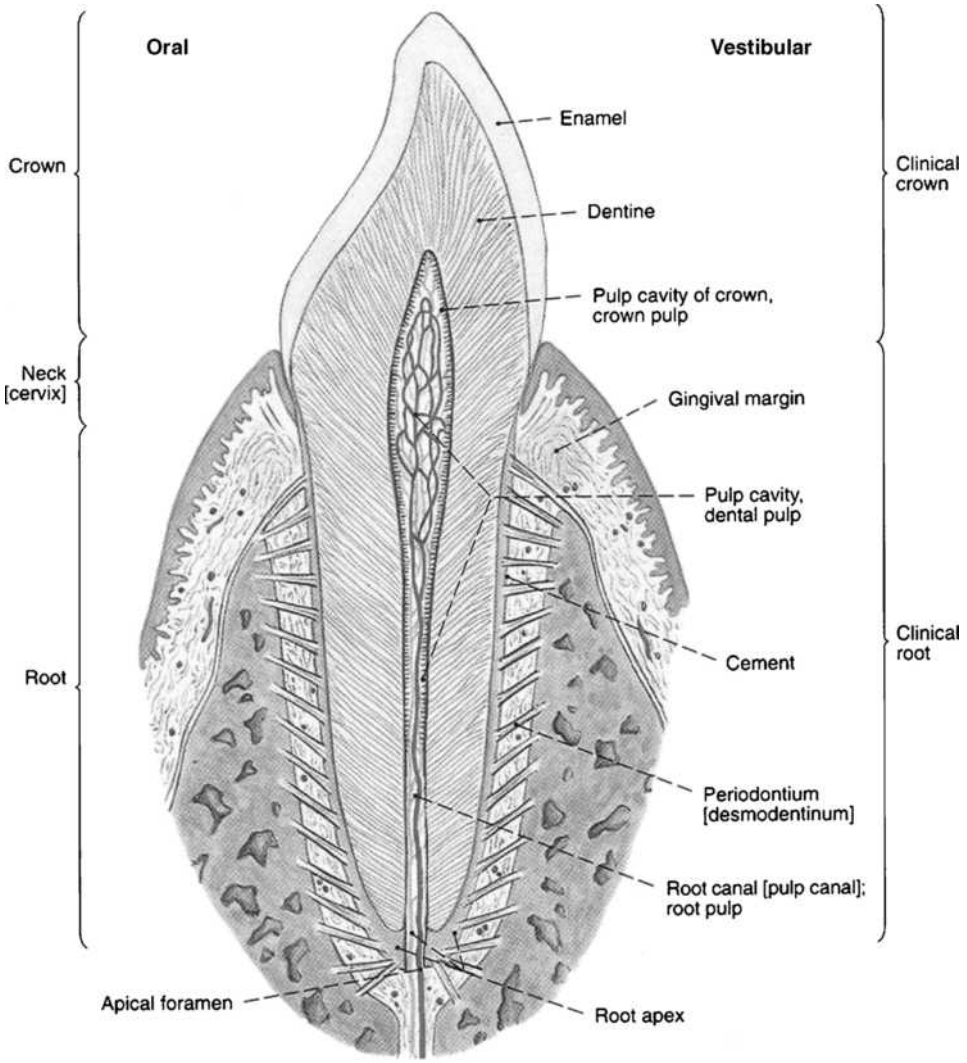


Figure 3. Schematic representation of an incisor tooth showing tooth socket and periodentum. (From Sobotta Atlas of Human Anatomy, Vol. 1, 13th ed.).

fore, they can apply large pressures on the food to provide effective cutting. Four incisors are located on the maxilla and the mandible, each. The *canines* are the long teeth whose function is to tear the food. In humans they mostly function as the incisors and cut the food. There are two incisors on the mandible and two on the maxilla. The *premolars* crush the food to smaller sizes during rhythmic, repetitive phases of mastication. Six maxillary and six mandibular *molars* are located at the ends of lower and upper dental arches. Molar and premolar teeth have multiple cusps and groves, which provide a relatively large surface area, enabling more effective crushing of the food.

The Teeth

Each tooth has two main sections, a *crown* and a *root* or roots as shown in Fig. 3. Each section is further subdivided to crown, cervix (neck), root, and apex. The tooth is solid except for the *pulp cavity* centered within it. The major portion of the tooth is made of *dentin*. A layer of *enamel* covers the crown portion of the tooth mostly above the gum line. A thin layer of *cementum* covers the root set in the bony tissue. The *cementum* is true bone. The pulp cavity may be divided into two portions; the *pulp chamber*, which is mostly in the crown, and the *pulp canal* traversing the interior of the root, ending in a constricted opening at the root apex. The pulp cavity contains the *dental pulp*, a soft tissue containing connective tissue, blood vessels, and nerves (12).

Both the dentin and the enamel contain collagen, hydroxyapatite (HAP) (an inorganic molecule), and water. However, the distributions of these constituents are different in these structures: The dentin contains 47% HAP, 30% collagen, and 23% water; and; the enamel contains; 92% HAP, 2% organic material, and 6% water, by volume. The microstructure of the enamel is characterized by parallel rods of HAP. In the dentin the tubules are connected by organic material. Because of the difference of HAP content and the microstructure, the modulus of the enamel can be as much as four to six times as high as that of dentin (4). See Table 1 for a summary of the elastic properties of the tooth and the periodentum.

The PDL

The general name given to the attachment mechanism of the teeth is the *periodontum*, which consists of the

Table 1. Elastic Properties of the Enamel, Dentin Layers of a Tooth and the PDL

	Elastic Modulus E [GPa]	Poisson's ratio v	Tensile Strength [MPa]	Compressive Strength [MPa]	Shear Strength [MPa]
Enamel	80[b,c]	0.3[b]	10[a]	288–400[a]	8[a]
Dentin	14[d], 15[b], 18[a], 20[b]	0.15[d], 0.31[b,c]	48[a]	232–297[a]	20[a]
PDL	0.002[c], 0.003[a] 0.05[b], 10[d]	0.45[c], 0.49[b,d]			

[a]From Toparli et al.(13)
[b]Rees and Jacobsen (14)
[c]Arola et al.(15)
[d]Imanishi et al.(16)

cementum, the PDL, the alveolar bone, and a portion of the gingiva (Fig. 3) (4). The PDL is a connective tissue present between the root of the tooth and the alveolar bone, whose function is (1) to provide support for the teeth and (2) to control the distribution of the occlusal loads on the bony tissue (5). Its thickness varies between 0.15 and 0.38 mm, in humans. Perhaps a good indication of its function can be observed when it is realized that the magnitude of occlusal loads affects the thickness of the PDL, where increased load results in thickening of the PDL.

Among other cellular components of the PDL are the osteoblast and osteoclast cells associated with the alveolar bone; cementoblast and cemontclast cells that associated with the cementum; and fibroblast cells that are responsible for collagen generation. The extracellular components of the of the PDL are the collagen fibers, oxytalan fibers, nerves, vessels, and the ground substance, which is composed of hyaluranic acid, glycoproteins, proteoglycans, and water (4).

Approximately 65% of the PDL's volume is occupied by dentoalveolar fiber bundles. The collagen molecules (type-I and III) of PDL are wrapped into collagen *fibrils* (55 nm diameter), which are wrapped into *fibers*. The fiber bundles are arranged into networks having a complex three-dimensional overlapping arrangement (4). These fiber bundles provide the load-bearing capacity to the PDL.

It has been mentioned that one of the primary functions of the PDL is to distribute the forces acting on the teeth to the bony tissue. The force–displacement relationship of the PDL is nonlinear. When a tooth is subjected to an external force, its initial displacements are caused by relatively small forces until the force magnitude reaches 1 N. Thereafter, increasingly higher forces are required to displace the PDL. The PDL is stiffer under axial loads acting on the tooth as compared with tangential loads. The same level of load applied tangentially causes a larger displacement of the PDL, as compared with the axial load. Like similar connective tissues, the PDL also shows viscoelastic material behavior.

THE SKELATAL COMPONENTS

The major bones of the human skull that are involved in mastication are the *mandible*, the *maxilla*, and the *temporal* bone. The *maxilla* is composed of two parts that are connected at the midpalatal suture, as shown in Fig. 2b. A large portion of the facial bone is composed of the maxillary bones. On the frontal plane, the maxilla is connected to the *zygomatic bone* and the *nasal bone*, and on the sides of the skull, it is connected to the temporal bone as shown in Fig. 1. The teeth on the maxillary arch are connected to the maxilla on the alveolar ridges. These teeth are considered to be the fixed part of the masticatory system, as the maxilla is fixed to the skull.

The *mandible* is the arch-shaped bone that forms the lower part of the facial skeleton and the masticatory system as shown in Fig. 2a. On the posterior (back) sides, the mandible extends vertically. This vertical (ascending) extension of the mandible is the *ramus*. The *mandibular angle* located at the posterior (back) part of the ramus is shown in Fig. 1. The superior (top) extension of ramus forms two processes: The anterior (frontal) one is the *coronoid process*, and the posterior (back) one is the *condyle*. These two processes make critical attachments to the rest of the masticatory system. The mandible is connected to the temporal bone at the condyle through the TMJ. The TMJ allows pivoting and sliding of the mandible with respect to the fixed part of the masticatory system, and it consists of various ligaments, the articular disk, the synovial capsules, and the synovial fluid. The articulating surfaces of the bones in the TMJ, namely, the condyle and the fossa, are covered with a fibrous tissue called the *articular cartilage*. The coronoid process serves as one of the endpoints of the temporal muscle. Various muscles and ligaments are attached to the mandible, enabling the mastication function. On the front part of the mandible, the teeth are connected through the *alveolar ridges*.

Two *temporal* bones are located on each lateral side of the human skull as shown in Fig. 1. The temporal muscles are connected to the wide area called the squamus part of the temporal bone. The mandible articulates in the concave part of the temporal bone called the *mandibular fossa*, located below the squamus part and near the zygomatic process as shown in Fig. 1. The mandibular fossa is also called the *articular fossa*; or the *glenoid fossa*. The *articular eminence* is located immediately in front of the mandibular fossa. The condyle of the mandible articulates on the maxilla through the TMJ. In the initial phases, of the opening of the jaw, the condyle rotates in the mandibular fossa; however, in the later phases; it moves forward and slides along the articular eminence. The thickness of the bone in the posterior part of the mandibullar fossa is relatively thin, whereas thicker bone is found in the anterior part and in the articular eminence. This is an indication of the load-bearing nature of these surfaces, where bone thickness is larger in sections subjected to larger stresses.

THE MUSCULATURE OF THE MASTICATION SYSTEM

The masticatory muscles are divided into *depressor* and *elevator* groups (6). The depressors are the *digastric*, *suprahyoid*, and *infrahyoid* muscles, and the elevators are the *temporal*, the *masseter*, the *medial pterygoid*, and the *lateral pterygoid* muscles (5).

The depressor muscles are located in the floor of the mouth. The supra- and infrahyoid muscles connect the hyoid bone to the mandible. The digastric muscle connects the mastoid process of the skull with the mandible, and it is attached to the hyoid bone through a tendon. These muscles are primarily involved during jaw opening and swallowing.

The three muscles, which form the fan-shaped muscle, shown in Fig. 4a, are collectively called the *temporal muscle*. The *anterior* (*front*) *temporal muscle* is composed of vertically oriented muscle fibers. The fibers orientation turns gradually toward the horizontal direction in the *middle temporal muscle* and the *posterior* (*back*) *temporal muscle*. The muscle fibers are attached to the temporal bone on the open part called the squamus of the temporal bone. These fibers come together as they descend downward through the zygomatic arch and form a tendon. This tendon attaches to the coronoid process and anterior border of the ascending ramus of the mandible.

The temporal muscle is both an elevator and a positioner. It can function unilaterally, bilaterally, or in sections to position and elevate the mandible. In bilateral closure, this muscle moves the condyle into the mandibular fossa. In unilateral action, the posterior temporal muscle moves the mandible toward the active side.

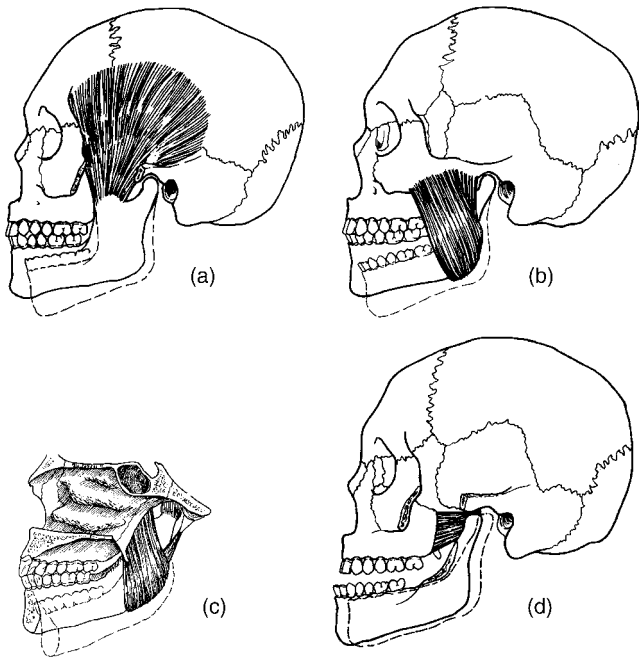


Figure 4. The elevator muscles of mastication: (a) the temporalis, (b) the masseter, (c) the medial pterygoid, and (d) the lateral pterygoid. (Modified from Temporomandibular Disorders and Occlusion by Okeson JP, 5th ed., Mosby, 2003.)

The *masseter* is located on the distal (outer) lateral sides of the ramus of the mandible, as shown in Fig. 4b. This muscle extends from the angle of the mandible upward and attaches to the zygomatic arch. The masseter is the most powerful elevator muscle in the masticatory system, and it is responsible for the high loads in the molar area. The crushing forces on the molar area could become high when the masseter acts together with the internal pterygoid and the anterior temporal muscles.

The *medial* (*internal*) *pterygoid* muscle is the internal counterpart of the masseter as shown in Fig. 4c. The fibers of this muscle originate from the *ptyregoid fossa* of the *sphenoid bone*, located internally in the mouth, extend downward, internally, and connect to the internal face of the mandibular angle, as shown in Fig. 4c. The medial pterygoid and the masseter function in a coordinated manner, and together they can generate high loads. Although the primary function of the medial pterygoid is to elevate the mandible, it can also move the mandible medially (toward the center).

The *lateral pterygoid* muscle: is the collective name of the two muscles: the *upper* (superior) and *lower* (*inferior*) *lateral ptyrogoid*, as shown in Fig. 4d. The upper lateral pterygoid muscle originates from the sphenoid bone located internally in the mouth, extends horizontally, and attaches to the condylar neck, the capsular ligament, and the articular disk. This muscle is active along with the elevator muscles during closing the teeth together for chewing or clenching. The upper lateral pterygoid muscle is a weaker muscle as compared with the lower.

The lower lateral pterygoid muscle originates at the pterygoid plate (Fig. 4d). It extends backward and upward connecting to the neck of the condyle. Bilateral contraction of the right and left lower pterygoids pulls the condyles down, out of the articular eminences, and the mandible moves forward, as shown in Fig. 4d. Unilateral contraction causes movement of the mandible to the opposite side.

THE TMJ

The relative motion of the mandible with respect to the temporal bone takes place at the TMJ, as shown in Fig. 5a. This is where the condyle of the mandible sits against the temporal articulating surfaces of mandibular fossa and the articular eminence. The TMJ is composed of the condyle of the mandible, the articulating surfaces of the temporal bone, the articular cartilage and disk, the ligaments, and the muscles. The articullar disk is considered to act as a non-ossified bone, between the mandibular fossa and the condyle; hence, the TMJ is considered a compound joint of the human body.

The articular disk deserves special attention, as it facilitates the relative motion to take place without direct bone-to-bone contact. The disk is located between the condyle and the temporal fossa. This is a fibrous tissue, which does not have blood vessels and nerves for the most part. The articular disk is mainly a mesh of *collagen fibers* whose interstices are filled with *proteoglycans*. In the TMJ disk, the collagen fibers help maintain its shape during loading, whereas the elastin fibers function to recover the form after

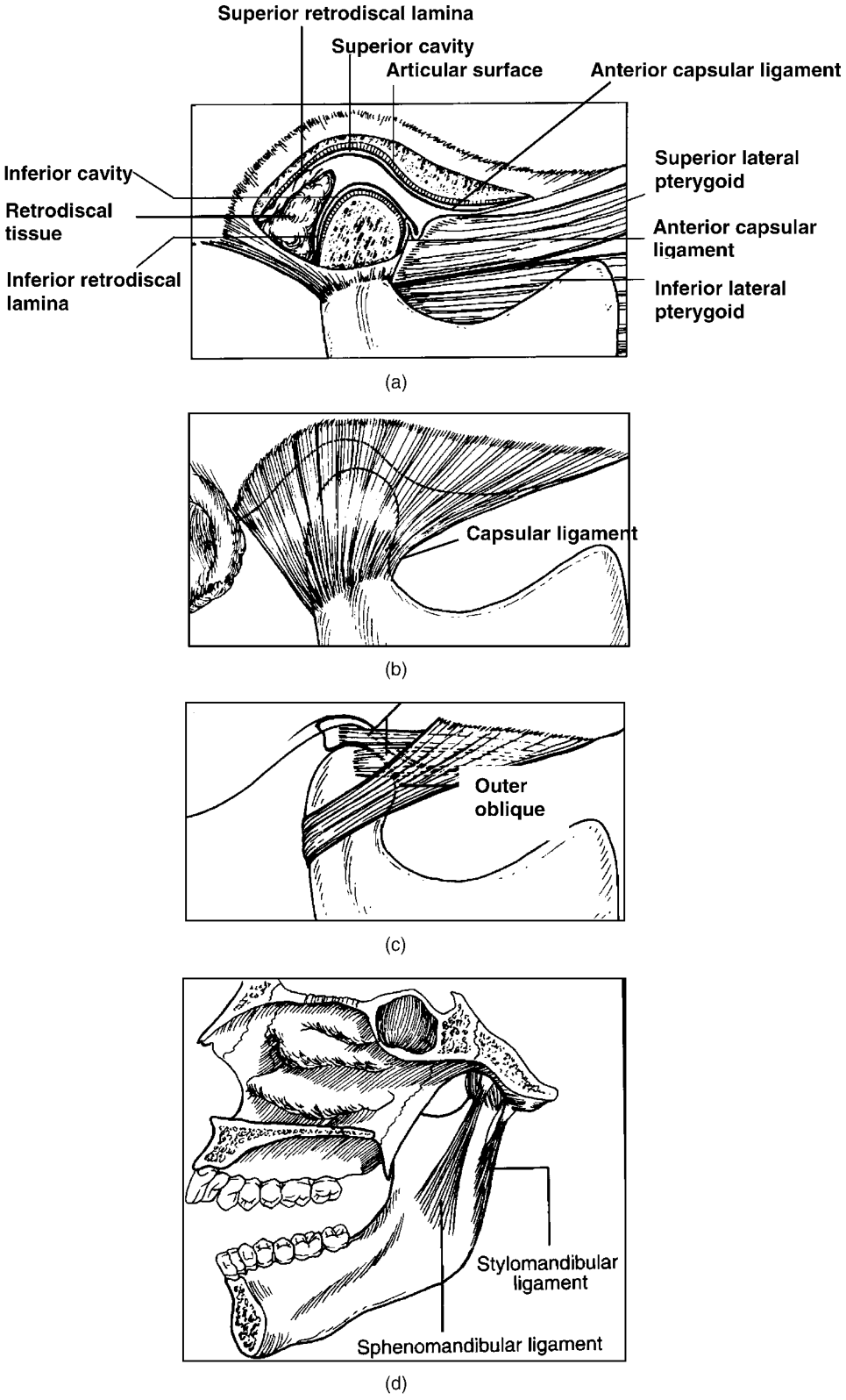


Figure 5. Schematic representation of the articular disk and the ligaments of the masticatory system: (a) the articular disk and its connecting tissues, including the collateral ligaments; (b) the capsular ligament; (c) the tempo-romandibular ligament; and (d) stylomandibular and spheno-man-dibular ligaments. (Modified from Temporomandibular Disorders and Occlusion by Okeson JP, 5th ed., Mosby, 2003.)

unloading. The mechanical properties of the articular disk strongly depend on the collagen fiber and the proteoglycan composition and organization (17). Experimental evidence suggests that the elastic modulus of the articular disk increases with age (18). The load-displacement (stress-strain) behavior of the articular disk is nonlinear and viscoelastic. A poroelastic material behavior model where the solid matrix behavior was hyperelastic showed several similarties with dynamic indentation tests of articular disks (19).

The articular disk is mainly supported by ligaments, as shown in Fig. 5a. In the posterior (back) region of the TMJ,

the articular disk is attached to a loose connective tissue named the *retrodiscal tissue* (RT). The superior (top) portion of the RT is attached to the tympanic plate by the *superior retodiscal lamina*, and its inferior (bottom) part is attached to the condyle by the *inferior retrodiscal lamina*. The medial, lateral, anterior, and posterior surfaces of the articular disk are attached to the *capsular ligament*, which surrounds most of the surfaces of the condyle and the articular surface of the temporal bone. In the front, the articular disk is also attached to the superior (upper) lateral pterygoid muscle (5).

The capsular ligaments divide the joint into the *upper* and *lower cavities*. The internal surfaces of the cavities are covered with endothelial cells. The cavities are filled with *synovial fluid* produced by these cells. The synovial fluid serves as a lubricant in the TMJ and reduces the frictional forces between articulating surfaces. Depending on the joint speed and load, different lubrication regimes are thought to be responsible for this effect (7). These include *boundary lubricatio*n, *elastohydrodynamic lubrication*, and *hydrodynamic* lubrication modes (20). Another effect, which possibly contributes to joint lubrication, is the *weeping lubrication*, which takes place as the synovial fluid, retained in the articular cartilage, is forced out of the cartilage into the synovial cavity, under sufficient normal pressure. The weeping lubrication is thought to be more prevalent during elevation and clenching (5).

Two- (21) and three- (22,23) dimensional finite element models of the quasi-static opening of the jaw that included the contact relations among the articular disk, the condyle, and the articullar eminence showed that the disk moves together with the condyle. These models predicted that the superior lateral pterygoid muscle and the ligaments attached to the disk do not play a significant role in the disk movement during jaw opening (21,23). The biconcave shape of the disk is sufficient to move the disk with the condyle. These studies also showed that the articular disk is primarily loaded in its intermediary (central) region.

THE LIGAMENTS

Ligaments are the connective tissues between the bones. The main function of the ligaments of the masticatory system is to prevent the mandible from undergoing extreme relative motion. Ligaments also protect the nerves and the vessels that connect to the mandible. In the masticatory complex, there are five main ligaments as shown in Fig. 5: the *collateral* (*discal*) *ligament*, *capsular ligament*, the *temporomandibular ligament*, the *sphenomandibular ligament*, and the *stylomandibular ligament*.

The *collateral ligaments* attach the medial (inner) and distal (outer) surfaces of the articular disk to the condyle of the mandible. These attachments along with the anterior (front) and posterior (back) attachments of the articular disk to the capsular ligament create the *synovial cavities*. The attachment of these ligaments permit the motion of the articular disk front-to-back, or in the anterior–posterior direction. Therefore, this ligament allows the disk to travel with the condyle.

The *capsular ligament* encloses the entire TMJ, as shown in Fig. 5b, and thus it provides a sealing function for the synovial fluid. The entire circumference of the articular disk is also attached to this ligament. The capsular ligament is attached superiorly (top) to the temporal bone and inferiorly (bottom) to the neck of the condyle. This ligament resists lateral or inferior (downward) forces.

The *temporomandibular* (TM) *ligament* consists of an inner *horizontal* and an *outer oblique part*, as shown in Fig. 5c. The TM ligament plays an important role in the pivoting action of the TMJ. Both parts of the TM ligament originate from the zygomatic arch. The inner horizontal part extends from the zygomatic arch horizontally and attaches to the anterior (frontal) neck of the condyle. The outer oblique part extends from the zygomatic arch and attaches to the posterior (back) part of the neck of the condyle, as shown in Fig. 5c. The outer oblique part of this ligament resists excessive dropping of the mandible. During the initial phase of the mouth opening, the condyle can pivot around a fixed point, while this ligament is becoming stretched. When the stretching of this ligament reaches its limit, then the condyle moves downward and forward across the articular eminence to continue opening. It is believed that the *inner horizontal part* of this ligament limits the backward (posterior) movement of the condyle and the articular disk.

The full effect of this ligament in limiting the motion of the articular disk has been debated (9); biomechanical models of the disk movement during mandibular opening and closing have shown that the disk could move in the anterior and posterior directions due to the favorable contact conditions provided by its bicuspal shape (21,24).

The *sphenomandibular ligament* extends from the sphenoid bone and attaches to the *lingula* on the inner (medial) surface of the ramus as shown in Fig. 5d. The function of this ligament is not well understood. However, its main function could be protection of nerves and blood vessels from dislocation and trauma, and it can prevent extreme anterior and lateral dislocations (4).

The *stylomandibular ligament* is another accessory ligament. It is located between the styloid process and the back of the ramus as shown in Fig. 5d. It limits excessive protrusive movements of the mandible (4).

MASTICATION AND ITS BIOMECHANICAL MODELS

Mastication is the action of chewing foods. Mastication involves rhythmic and repetitive motion of the mandible with respect to maxilla. The mastication cycle starts with the *opening phase*, followed by the *crushing phase* and *grinding phase*, which occur during closure, as shown in Fig. 6a. The motion of the mandible during mastication is a three-dimensional, complex motion, which has been described as having the shape of a teardrop or a pear. When the motion of the incisors in an idealized mastication cycle is viewed from the frontal plane (Fig. 6a), this analogy becomes clear. During the opening phase, the mandible drops vertically for about 15–18 mm; thereafter, it moves

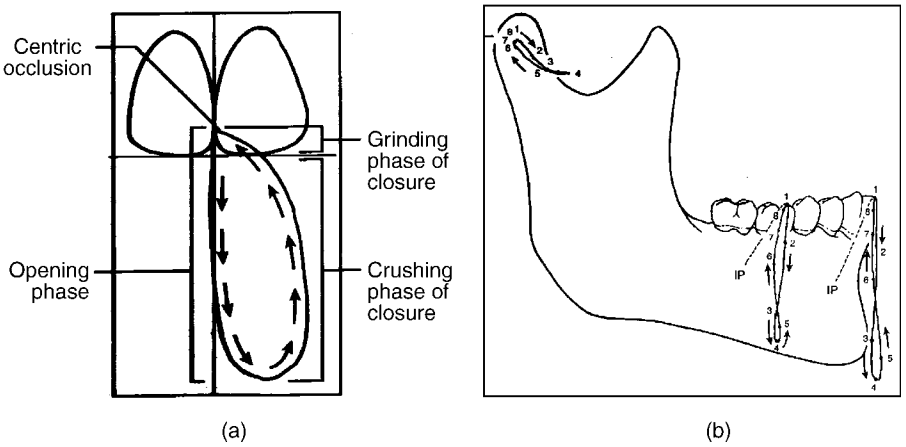


Figure 6. The mastication cycle viewed from (a) the frontal plane; and (b) the sagittal plane. (Modified from Temporomandibular Disorders and Occlusion by Okeson JP, 5th ed., Mosby, 2003.)

laterally and the closure phase begins. During the initial phase of closure, the incisors move 4–5 mm laterally while crushing the food. This phase lasts until incisors are located 3–4 mm laterally and 3 mm vertically with respect to their initial (intercuspal) condition. During the grinding phase, which follows, the food is sheared between the cusps of the incisors. The view of the motion of the incisors from the side (sagittal) plane on the working side, shown in Fig. 6b, shows that during the opening phase, the incisors move slightly in the frontal (anterior) direction, followed by a posterior motion during closing. The TMJ and the premolars also follow a path similar to that of the incisors in the sagittal (side) plane on the working side.

The amplitude of the anterior and the lateral movement of the mandible depend on the stage of mastication. During the initial phases, the incisors are used to cut the bolus (food) and the anterior and lateral movement are relatively large. During the later phases, the posterior teeth are used more, and the lateral movement of the teeth is reduced. The consistency of the food also affects the lateral movement, where harder foods require larger lateral movements and more chewing cycles, as shown in Fig. 7.

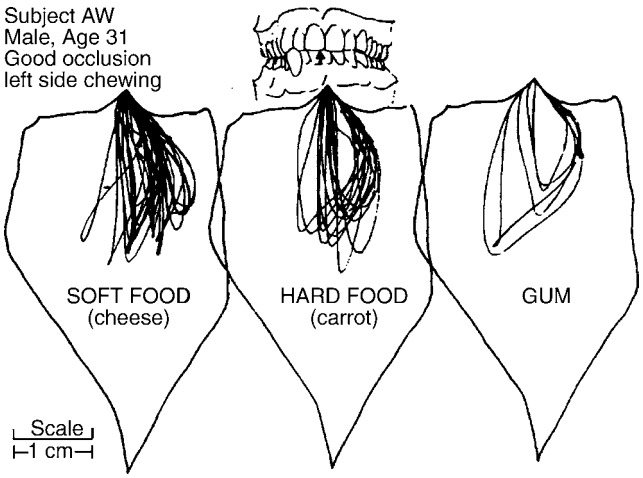


Figure 7. Frontal view of the chewing stroke from a 31-year-old man with good occlusion, for soft food, hard food, and chewing gum. (From Advances in Occlusion, by Lundeen HC, Gibbs CH, Boston, 1982).

In biomechanical analysis of mastication, static analysis are carried out to determine maximum clenching forces that can be applied by the muscles; dynamic analysis, on the other hand, provides information about the muscle TMJ interactions during the open and close cycles, as well as laterodeviations. In these analyses, the mandible is modeled as a rigid body, and the muscle forces are modeled as concentrated forces, as shown in Fig. 8. The attachment points and the three-dimensional vectorial orientations of the muscles are typically, carefully, measured from cadaver specimens. The biomechanical analysis of mastication also considers the deformations of the mandible due to external forces.

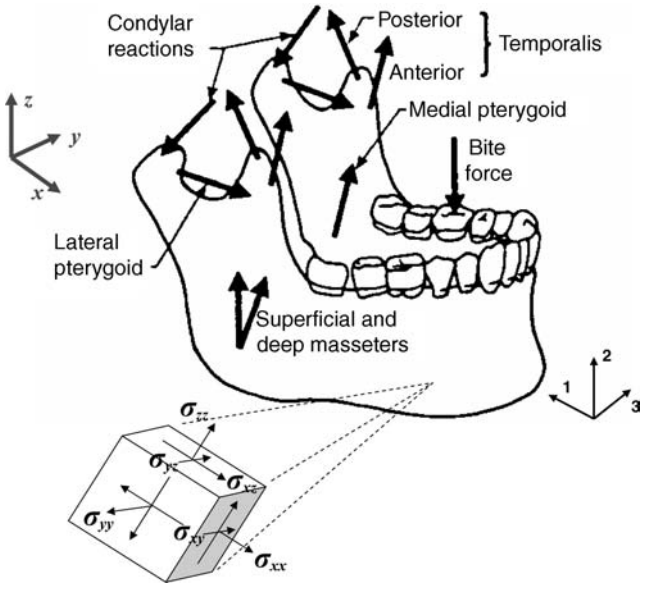


Figure 8. The locations and directions of the muscle force vectors, the bite force vector, and the condoylar reaction force vectors are shown schematically. The *x*-, *y*-, and *z*-coordinate system indicates the coordinate system in which the forces are measured. The 1-, 2-, and 3- coordinate system indicates the transversely isotropy directions for material properties of the mandible. The small stress cube indicates the internal stresses generated at a point inside the mandible in response to external loading. (Modified from Faulkner et al., "A three dimensional investigation of temporomandibular joint loading," Journal of Biomechanics 1987;20:997–1002.)

In the next section, a classification of the internal and external forces acting on a deformable body will be described. The material property definitions and deformation behavior of the tissues involved in the masticatory system is described next. Finally, some static, dynamic, and elastic models of mastication developed in the last 20 years will be described.

In general, the forces acting on a structure can be classified as *external forces* and *internal forces*. The external forces can be further classified as *surface tractions* (external pressure), *concentrated forces*, and *body forces*.

External Forces

Surface tractions are external forces distributed over a finite area, on the surface of the structure. For example, the contact pressure between the condyle of the mandible and the articular disk takes place over a relatively large area. This contact pressure is a normal traction acting on the mandible (and the disk). Note that distribution of this traction can vary from point-to-point. In the metric system, the unit of traction is Pascal (Pa), where 1 Pa = $1\,\mathrm{N/m^2}$.

Concentrated force is a term used for an idealized traction acting over an infinitesimally small area. In the metric system, the unit of the concentrated force is Newton (N), where $1\,\mathrm{N} = 1\,\mathrm{kg{\cdot}m/s^2}$. Use of concentrated forces is common in mechanics as they simplify the analysis in many instances. For example, in many mechanical models of the mandible, the forces exerted by the muscles, the forces experienced by the teeth during clenching or mastication, and the reactions on the condyle are idealized as concentrated forces as shown in Fig. 8. While idealizing the muscle forces, investigators spend a great deal of effort to ensure that the idealized forces represent the (actual) tractions in a mechanically equivalent manner.

Both surface tractions and concentrated forces are transferred from one body onto another through physical contact. Forces exerted by muscles on the bones are in this category. On the other hand, a *body force*, such as that due to gravity, is a force that acts over a distance, without requiring direct physical contact. For example, if it was not for the slight state of contraction of the clenching muscles, the mandible would stay open due to the effect of gravity.

Internal Forces and Internal Stresses

The internal forces develop inside a structure in response to external forces and enable the structure to stay together. The internal forces can only be "visualized" by taking a virtual cross-section of the structure. Depending on the location and orientation of the cross-section, the internal forces will vary in magnitude and direction. The resultant of the internal forces acting on an infinitesimally small cross-sectional area (dA) of the structure can be decomposed into a *normal force* (dF^n) acting perpendicular to the cross-section and a *shear force* (dF_s) acting in the plane of the cross-section. *Stress* at a point inside the structure can then be defined as the limit of the internal forces acting on the area dA as it becomes infinitesimally small. In general, the internal stress state of a structure can be expressed by six independent stress components σ_{xx}, σ_{yy}, σ_{zz}, σ_{xy}, σ_{xz}, σ_{yz} as shown in Fig. 8. (Note that six independent stress components assume that no internal twisting moments exits in the structure. In case these exist, then nine stresses are necessary for this description) Note that in this representation, the first subscript refers to the direction of the normal of the plane on which the stress is acting, and the second subscript refers to the direction of action of the internal force. Thus, σ_{xx}, σ_{yy}, and σ_{zz} are the *normal stress* components, and σ_{xy}, σ_{xz}, and σ_{yz} are the *shear stress* components. It is important to remember that internal stresses are defined at a point inside the structure, and in general, they can vary from point-to-point. Whether they do vary depends on many factors such as the shape of the structure, external force distribution, and material properties. In the metric system, like the tractions, the unit of stress is Pascal.

Deformation and Strain

In general, a structure that is properly fixed on its boundary will deform in response to external forces. The deformation of a point on the structure can be characterized by its displacements. In general, a point $P^{\rightarrow}$ before deformation will be displaced to a new location $P^{\rightarrow\prime}$ after deformation. For small displacements, the displacement vector $d^{\rightarrow}$ can be expressed in a Cartesian coordinate system. The components of $d^{\rightarrow}$ along the x-, y-, and z-axes are u, v, and w, respectively.

Relations governing the mechanics of deformable bodies are expressed in terms of strain, rather than the displacements. In general, *strain* is a nondimensional measure of deformation at a point. The *normal strains* ε_x, ε_y, and ε_z represent the change in length per unit of initial length, along the x-, y-, and z- axes, respectively. The *shear strains* γ_{xy}, γ_{xz}, and γ_{yz} represent the decrease in the right angle initially formed by the sides parallel to the x-y, x-z, and y-z axes, respectively (26).

Material Properties

The load-displacement behavior of a structure depends on the type of material involved. In general, the material properties of a structure are determined by uniaxial tension test, shear test, and hardness test. A uniaxial tension test and a pure-torsion test establish the relations between, for example, σ_{xx} and ε_x and σ_{xy} and γ_{xy} by Hooke's law:

$$\sigma_{xx} = E\varepsilon_x \quad \text{and} \quad \sigma_{xy} = G\gamma_{xy} \tag{1}$$

where the proportionality constants are the elastic (or Young's) modulus E and the shear modulus G. In general, the three-dimensional stress-strain relations are expressed with the generalized Hooke's law, which can be given in matrix form as

$$\{\sigma\} = [E]\{\varepsilon\} \quad \text{or} \quad \{\varepsilon\} = [C]\{\sigma\} \tag{2}$$

where the *stress vector* is $\{\sigma\} = \{\sigma_{xx}\ \sigma_{yy}\ \sigma_{zz}\ \sigma_{xy}\ \sigma_{yz}\ \sigma_{zx}\}^{\mathrm{T}}$, the *strain vector* is $\{\varepsilon\} = \{\varepsilon_x\ \varepsilon_y\ \varepsilon_z\ \gamma_{xy}\ \gamma_{yz}\ \gamma_{zx}\}^{\mathrm{T}}$, the *elasticity matrix* is $[E]$, and the *compliance matrix* is $[C]$ $(=[E]^{-1})$. The elasticity matrix of an *anisotropic* material has 21 independent components. In case the material has three orthogonal planes of symmetry, the material is said to be *orthotropic*. The compliance matrix for an orthotropic

material is defined by nine independent material properties (27):

$$[C] = \begin{bmatrix} 1/E_1 & -v_{21}/E_2 & -v_{31}/E_3 & 0 & 0 & 0 \\ -v_{12}/E_1 & 1/E_2 & -v_{32}/E_3 & 0 & 0 & 0 \\ -v_{13}/E_1 & -v_{23}/E_2 & 1/E_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/G_{23} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/G_{31} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/G_{12} \end{bmatrix} \quad (3)$$

where E_i are the Young's modulii, G_{ij} are the shear modulii, and v_{ij} are the Poisson's ratios in the respective directions, with the restriction that $v_{12}/E_1 = v_{21}/E_2$, $v_{13}/E_1 = v_{31}/E_3$, and $v_{32}/E_3 = v_{23}/E_2$. The number of independent elastic constants is reduced to five for a *transversely isotropic* material, such as the bone, where each plane through a longitudinal axis is a plane of elastic symmetry. For fully *isotropic* materials, the material properties are the same in all directions, and two independent properties (E, v) are sufficient to describe the stress-strain relationship.

Mechanical Properties of Bone

Bone is composed of collagen, water, hydroxyapatite mineral, and small amounts of proteoglycans and noncollagenous proteins. Collagen is the structural protein that gives the flexibility and tensile strength, also found in ligaments and the articular disk. Hydroxyapatite, $Ca_{10}(PO_4)(OH)_2$, is a crystal with hexagonal symmetry located within and between collagen fibers, found in the form of needles, plates, and rods. The internal structure of bone is porous. Two distinct porosity ranges give rise to the *cortical* (compact) bone with low porosity (5–10%) and the *trabecular bone* with 75–95% porosity. The interstices are filled with marrow, a tissue composed of blood vessels, nerves and various types of cells (28).

Among these cells are the *osteoclasts*, which are responsible for formation of new bone, and *osteoblasts*, which are responsible for resorption of bone. *Remodeling* of bone involves repair of internal cracks, and so on, due to coupled activity of osteoclasts and osteoblasts. In addition to self-healing, the bone also has the ability to *adapt* to variations in imposed stresses (Wolff's law). It has been hypothesized that the strain in the bone tissue stimulates the biological response resulting in adaptation (28). The physiologic loading zone has been reported to be in the 1000–3000 microstrain range (28). Bone adaptation has significant biomechanical consequences, in design of endosseous implants, and prosthetics. The loading conditions created in the bone by these should be carefully considered to prevent excessive or insufficient loading.

The cortical bone is a transversely isotropic material, as two of its principal material directions have similar mechanical constants. The stiffest direction of the mandibular cortical bone is along a tangent to the parabolic curve of the mandible, i.e., direction -3 in Fig. 8 (29). The mechanical properties of the *cortical bone* depend on several factors, including the porosity of the bone, mineralization level, bone density, collagen fiber organization, and rate of deformation (28). The ultimate stress of the cortical bone has been reported to be higher in compression (170 MPa) than in tension (100 MPa) (28).

Table 2. Elastic Properties of Cortical and Trabecular Bones Measured for a Transversely Isotropic Material Model

	Transversely Isotropic	
	Cortical Bone[a] [GPa] (30)	Trabecular Bone[b] [Gpa] (30)
$E_1 = E_2$	13.0	0.27
E_3	19.0	0.82
G_{12}	5.3	0.12
$G_{23} = G_{31}$	5.9	0.12
$v_{12} = v_{21}$	0.22	0.19
$v_{31} = v_{32}$	0.42	0.34
$v_{13} = v_{23}$	0.29	0.11

[a]Measured for mandible using ultrasonic techniques.
[b]Measured for human tibia.

The mechanical properties of the *trabecular (cancellous) bone* depend on the porosity, the orientation of the trabecular architecture, and the material properties of the tissue in the individual trabeculae (28). The strength of the trabecular bone has been reported to be the same in tension and compression, and it is approximately 2–5 MPa (28).

Experimental data for the transversely isotropic material properties of the cortical bone of the mandible were obtained by Carter as reported by Hart et al. (30). These values are reported in Table 2. This table also contains the trabecular bone properties measured for human tibia. In other studies, many investigators assumed that the jaw bone deforms in an isotropic manner. The elastic modulus values used in these finite element analysis models ranged between 11 and 20 GPa for the cortical bone and between 0.2 and 7.9 GPa for the trabecular bone. The Poisson ratio was typically taken in the range of 0.3–0.33 (31–35).

Mechanical Properties of the Cartilaginous Tissue of the Masticatory System

The cartilagenous tissue of the masticatory system is the *articular cartilage* on the articular surfaces of the TMJ and the *articular disk*. In general, this type of tissue is composed of cells (chondrocytes) imbedded in intercellular matrix, permeated by a system of fibers (1). Depending on the function of the cartilage tissue in the body, the appearance and the nature of the intercellular matrix can be different. For example, the *yellow elastic fibrocartilage* is found in external ears and larynx; the *hyaline cartilage* is found in nasal, tracheo-bronchial, and articular surfaces; and the *white fibrocartilage* is found in intervertebral and articular disks (1).

In indentation experiments, the cartilage tissue shows an instantaneous recovery, followed by a time-dependent creep deformation, when it is released after compression. Full recovery to the initial state takes place in a finite amount of time. Such material behavior is termed *viscoelastic*. The cartilage tissue is porous, and the synovial

fluid fills the interstices of the tissue. Under compressive stress, the fluid can move out of the tissue, and upon relaxation, it will return to interstitial openings. This fluid flow provides lubrication to the joints as well as nutrients to the cartilage cells (1). The viscoelastic behavior is due to (1) the inherent viscoelastic nature of the solid matrix and (2) the velocity differences between the solid matrix and the liquid (11).

The mechanical response, biomechanical function, and biological integrity of the cartilaginous tissue are governed by the movement of the interstitial fluid (1) *within* cartilage and (2) *across* the articular surface. The synovial fluid is transported through the porous membrane due to a pressure gradient across the tissue. Flow also occurs due to the deformation of the cartilage matrix; the pressure of the fluid internal to the matrix rises in response to deformation. The resulting pressure gradient with respect to the articular surface causes the fluid flow. The biphasic model for the mechanical behavior of the articular cartilage was formulated by Mow et al. (36). In this model, the tissues are modeled as consisting of a soft, permeable, elastic solid, mixed with water. The fluid and the solid phases are coupled through the pressure. Experimentally observed viscoelastic behavior of the articular cartilage can be explained by this biphasic model (36).

Although the articular cartilage behaves as a viscoelastic material due to its poroelastic nature, it can undergo large deformations. Experimental evidence further suggests that the material properties are nonhomogeneous and anisotropic. The experimental measurements of the load-displacement behavior of the articular cartilage of the rabbit TMJ have indicated that the elastic modulus (E) depends on the location with respect to the joint, varying between 0.95 and 2.34 MPa (37). These values are on the same order of magnitude as that reported by Woo et al. obtained for bovine humeral articular cartilage (11). The elastic modulus and Poisson's ratio of the articular cartilage of the temporal joint from these references are given in Table 3.

The articular cartilage is a connective tissue that provides very small frictional resistance to the sliding motions of the joints, in general. Experimental studies to determine the coefficient of friction (COF) show that the presence of the synovial fluid plays a significant role in its low level. Although it has been found that the COF can be as low as 0.005, it has also been shown that it depends on the level of the normal stress and the motion of the joint. In general, under large normal stress values, the COF is found to increase. Moreover, the static friction coefficient is larger than the dynamic friction coefficient (1,18,38). This behavior is in contrast to the behavior of the COF between most other materials (20).

Table 3. Elastic Properties of the Articular Cartilage

Reference	Location	Elastic Modulus E[MPa]	Poisson's Ratio v
Hu (37)	Anterior	2.34[a]	0.46
	Central	1.48[a]	0.39
	Posterior	1.51[a]	0.41
	Medial	1.11[a]	0.38
	Lateral	0.95[a]	0.31
Woo (11)	–	0.79	0.4

[a]Measured from rabbit mandibular condyle.

Tension and compression tests of human *articular disk* show nonlinear stress-strain behavior (14,39). Chin et al. measured the viscoelastic properties of human TMJ disks (40). Chen et al. (39) assumed that the solid matrix of the articular disk can be modeled as a hyperelastic material of the Mooney–Rivlin type (41), where the strain energy of an incompressible rubber-like material is expressed as

$$U = c_1(I_1 - 3) + c_2(I_2 - 3) \tag{4}$$

where I_1 and I_2 are the first and second invariants of the Cauchy–Green deformation tensor and c_1 and c_2 are empirically determined constants (41). Chen et al. used $c_1 = 27.91$ MPa and $c_2 = -20.81$ MPa, based on experimental data obtained from dog articular disks (39).

More recently, Tanaka et al. obtained experimental results for human articular disks (18). Although their data show the same trend predicted by the Mooney–Rivlin type of hyperelastic behavior, they approximated the curve by a single elastic modulus calculated at 2% strain level, and used a linear FEA. Chen and Xu (42) and Hu et al. (37) used a linear piecewise continuous material model. Other investigators assumed that the articular disk behaves strictly in a linear-elastic manner (19,21,43). The data for elastic modulus (E) used in these work are reported in Table 4.

Mechanical Properties of the Ligaments

The physiological function of a joint ligament is to provide stability to joints and to limit their range of motion (45). Ligaments can resist only tensile forces. The structural component responsible for the tensile strength of ligaments is the protein, collagen. Other components of this tissue are elastin, proteoglycans, and glycoproteins. The basic molecular collagen unit is a left-handed molecule. Collagen molecules are wound into super-helices, microfibrils, and fibrils in a hierarchical manner. At each level, the orientation of the helix is reversed. This alternating helix direction is useful in converting the axial tension to circumferential compression, and it is a contributing factor to the strength of the ligaments (45).

The load-displacement curve of the ligament under tensile loading initially displays nonlinear behavior with tangent modulus increasing with increasing strain. Increasing the strain even further, eventually a linear

Table 4. Elastic Properties Used for the Articular Disk Assuming That It Behaves in a Linear Elastic Manner, Used in the FEA of the TM

Reference	Elastic modulus E[MPa]	Poisson's Ratio v
DeVocht (21)	1.8	0.4
Beek (1)	6.0	0.4
Beek (1)	6.8	0.4
Tanaka (18)	47.1[a]	
Hu (44), Chen(42)	44.1[b] and 92.4[b]	0.4

[a]This value is measured at 2% strain, and it is the average of seven human specimens.

[b]E = 44.1 MPa if $\sigma < 1.5$ MPa and E = 92.4 MPa if $\sigma > 1.5$ MPa. These data are based on specimens obtained from dogs.

portion is reached. The post-yielding region shows a decreasing tangent modulus with increasing strain. The elastic modulus (E) of the nuchal (neck) ligament is 7.5 MPa, the ultimate strength, and the ultimate strain are 2.4 and 1.25 MPa, respectively (45). Chen and Xu used nonlinear springs, which can only carry tension, to model the upper, and lower posterior ligaments and the anterior ligament (42). They used data from the talofibular ligament of human ankle and assumed that the stiffness of the ligament would be proportional to its cross-sectional area. Thus, they calculated spring stiffness values that are on the order of 10.9–16.35 kN/m.

Static Models of Mastication

Static models of mastication aim to develop an understanding of the maximum forces applied by muscles of mastication. These models typically treat the mandible as a rigid body. The motion of a rigid body is fully described by the displacements and rotations of its center of mass. During occlusion, the external loads acting on the masticatory system are the forces on the TMJ, the ligaments, the teeth, and the muscles, as shown in Fig. 8.

The static equilibrium, for example, in clenching, requires that the sum of the external force and moment vectors be in balance. Considering the subdivisions of the muscles of mastication (i.e., anterior, medial, posterior temporalis; superficial and deep masseter; superior and inferior lateral pterygoid; and, the medial pterygoid), it can be observed that there will be $N_m = 16$ muscle force vectors ($\vec{F}_i^{(m)}$) acting on the mandible, where N_m is the total number of muscles. Note that the subscript i ($1 \le i \le N_m$) is an integer counter used to identify each muscle. Similarly, the reaction force vectors due to the ligaments, due to the contacting teeth, and at the joints of the TMJ are indicated by $\vec{F}_i^{(l)}$, $\vec{F}_i^{(t)}$, and $\vec{F}_i^{(j)}$, respectively. Note that these forces can only attain positive values; muscles only apply tensile forces; the ligaments provide no appreciable resistance when they are compressed; and when the TMJ and the teeth are in contact, the reaction forces are considered to be positive. The static equilibrium of external forces is ensured by the following vector equation:

$$\sum_{i=1}^{N_m} \vec{F}_i^{(m)} = \sum_{i=1}^{N_l} \vec{F}_i^{(l)} + \sum_{i=1}^{N_t} \vec{F}_i^{(t)} + \sum_{i=1}^{N_j} \vec{F}_i^{(j)} = 0 \qquad (5)$$

where N_l, N_t, and N_j are the total number of force vectors related to the ligaments, teeth, and joint reaction forces. The location of each force vector is represented by a location vector $\vec{r}_i^{()}$. Each force vector causes a moment with respect to the origin of the location vector represented by the vector cross-product $\vec{r} \times \vec{F}$. The moment balance with respect to the origin is expressed by the moment equilibrium equation:

$$\sum_{i=1}^{N_m} \vec{r}_i^{(m)} \times \vec{F}_i^{(m)} + \sum_{i=1}^{N_l} \vec{r}_i^{(l)} \times \vec{F}_i^{(l)} + \sum_{i=1}^{N_t} \vec{r}_i^{(t)} \times \vec{F}_i^{(t)} + \sum_{i=1}^{N_j} \vec{r}_i^{(j)} \times \vec{F}_i^{(j)} = 0 \qquad (6)$$

Equations 5 and 6 are a compact representation of the force and moment vectors. In a Cartesian coordinate system, each vector is represented as the vector sum of its components acting along the x-, y-, and z-axes. Thus, equations 5 and 6 each represent three equations of equilibrium along these axes, resulting in a total of six equations.

As mentioned, there are four elevator muscles on each side of the skull. Some investigators consider only these eight muscles during function; therefore, they use $N_m = 8$ (46). In other studies, depending on the level of detail, the total number of muscle forces used can vary up to $N_m = 26$ (47). Similar comments apply for the number of forces on the ligaments, teeth, and joints. Thus, it can be easily recognized that the six equations of equilibrium, represented by equations 5 and 6, are not sufficient in solving for the magnitudes of the numerous unknown forces. Such systems are statically indeterminate.

To find a solution for a statically indeterminate system, additional equations and assumptions are necessary. Osborn and Baragar (47) assumed that the *strain sensors* in the muscle tissue and pressure sensors in the TMJ are activated at a rate proportional to the magnitude of the muscle tension and joint reaction forces. The total output f of these sensors then becomes

$$f = \sum_{i=1}^{N_m} c_i^{(m)} |\vec{F}_i^{(m)}| + \sum_{i=1}^{N_j} c_i^{(j)} |\vec{F}_i^{(j)}| \qquad (7)$$

where and $c_i^{(m)}$ and $c_i^{(j)}$ are the sensor output rates specific to the muscles and joints, respectively. They further postulated that the central nervous system minimizes the output of the (cost) function f. Thus, the problem of finding the reaction forces in mastication is transformed into a problem where the cost function in equation 7 is minimized subject to satisfaction of equations 5 and 6 and conditions $|\vec{F}_i^{(m)}| \ge 0$ and $|\vec{F}_i^{(j)}| \ge 0$. This defines a *linear programming* problem, typically encountered in economics, and it can be handled with various available approaches (48). Osborn and Baragar investigated the effects of minimizing only the joint forces ($c_i^{(j)} = 1$, $c_i^{(m)} = 0$) or only the muscle forces ($c_i^{(m)} = 1$, $c_i^{(j)} = 0$).

The following muscle groups were considered in their study: the masseter (superficial, deep, anterior, and posterior), the medial pterygoid (anterior and posterior), the temporalis (anterior, medial, and posterior), the lateral pterygoid (superior, upper, inferior), and the digastric. Thus, they had $N_m = 23$. They neglected the effect of ligaments ($N_l = 0$), and assumed a single reaction force ($N_j = 1$) acts on each TMJ. The bite force on different teeth were treated as known external forces. They showed that the model where the cost function involves only the minimization of muscle forces results in more realistic mastication force scenarios. Their model predicted that muscle elements with longer moment arms relative to the joint are activated first. As the bite force increases, a ripple activity spreads into muscles with shorter moment arms.

Hatcher et al. investigated the muscle forces involved in unilateral clenching by using a model that is symmetrical around the mid-saggital plane, including six muscles: the posterior and anterior temporalis, the deep and superficial masseter, the medial pterygoid, and the lateral pterygoid (49).

Due to the assumption of symmetry, $N_m = 6$. They assumed that the left and right TMJ forces were identical ($N_j = 1$). They compared the mathematical results with the results obtained from an *in vitro* model of a skull. The muscle action in the *in vitro* model was simulated by applying external forces, and occlusal and TMJ forces were measured. The study concluded that more realistic results are obtained in the case where the applied force magnitudes are based on the cross-sectional area of the muscles and the electromyogram (EMG) data, in contrast to when they are solely based on the cross-sectional area of the individual muscles. In a different study, the same authors predicted that the balancing side condylar reaction force is larger than the working side, when occlusion occurs on either one of the molars (25). They also showed that the ipsilateral (same side) condylar reaction force varies considerably as the occlusion direction varies in the parasagittal plane.

Smith et al. used a total of six muscles: the right and left lateral pterygoid, the right and left temporalis, and the right and left masseter/medial pteryoid muscle complex (46). They considered the TMJ load and one bite force. They used a minimization technique in which their cost function was the root-mean-square of the TMJ reaction force. This analysis predicted that the TMJ is loaded over the normal functional range of bite force positions and angles. The magnitudes of the reaction forces on the TMJ were found to vary between 5% and 60% of the occlusal force depending on which teeth the occlusion takes place.

Koolstra et al. (50) considered the effects of the eight clenching muscles: the deep and superficial masseter; the medial pterygoid; the anterior, posterior, and deep temporalis; and the superior and inferior lateral pterygoid. They treated the left and the right sides as being independent; hence, a total of 16 muscles were included. The magnitudes of the TMJ reaction forces and the muscle forces were unknown, but their directions of application were measured from the skull. The bite force, on the other hand, was treated as a known quantity. The total number of unknowns were 18, and the linear programming technique was used. The cost function f was based on the postulation that the relative activity of the most active muscle among all muscles should be as small as possible. The relative activity $|\vec{F}_i^{(m)}|$ of a given muscle i was defined as

$$|\vec{F}_i^{(m)}| \leq \mu |\vec{F}_i^{(m)}|_{\text{Max}} \tag{8}$$

where $|\vec{F}_i^{(m)}|_{\text{Max}}$ is the maximum possible for of the muscle and μ is a coefficient defined below. The maximum muscle force of a muscle element is assumed to be proportional to its physiological cross-section, and it is a known quantity. The objective function was given as

$$f(|\vec{F}_i^{(m)}|) = \mu \tag{9}$$

subject to N constraints defined by equation 8. Using this model, the maximum possible bite forces, for which the masticatory system can remain stable, were predicted. Thus, biteforce and joint-force envelopes for mastication at different teeth were evaluated. *In vivo* validation of these results were carried out by Koolstra and van Eijden (51). An overall good agreement between measured and predicted values was observed.

Dynamic Modeling of Jaw Opening and Closing

Biomechanical analysis of the jaw opening and closing was modeled by Koolstra and van Eijden (24,52–54). The dynamic modeling of the jaw opening and closing involves equations of dynamic equilibrium, where the jaw is considered rigid. Thus, the motion of the mandible can be determined by six degrees of freedom of its center of mass; the three rigid-body displacements $\{u\} = \{u_x\, u_y\, u_z\}^{\mathrm{T}}$ in the x-, y-, and z- directions, and the three rotations $\{\theta\} = \{\theta_x\, \theta_y\, \theta_z\}^{\mathrm{T}}$ about the x-, y-, and z-axes. The six equations of motion are given in the general matrix form as follows:

$$[M]\{\ddot{u}\} + [C_u]\{\dot{u}\} = \sum \vec{F} \tag{10}$$

$$[\bar{I}]\{\ddot{\theta}\} + [C_\theta]\{\dot{\theta}\} = \sum \vec{M}_G \tag{11}$$

where $[M]$ is the mass-matrix, $[C_u]$ is the damping matrix for the displacements, $\{\dot{u}\}$ is the velocity vector, $\{\ddot{u}\}$ is the acceleration vector, $[\bar{I}]$ is the rotational inertia matrix calculated with respect to the center of gravity of the mandible, $[C_\theta$ is the damping matrix for the rotations, $\{\dot{\theta}\}$ is the angular speed vector, and $\{\ddot{\theta}\}$ is the angular acceleration vector. The sum of the external forces ($\sum \vec{F}$) and the sum of the moments ($\sum \vec{F}_G$) with respect to the center of mass are calculated using the same relations given in equations 5 and 6 (55).

In Reference 52 the muscle forces are treated as known external forces. This restriction is removed in later studies. The forces on the TMJs and the teeth were calculated based on contact constraints. This aspect of the model makes it nonlinear. The condyle was approximated as a sphere, and the shape of the articular fossa and articular eminence were approximated by a third-order polynomial, based on the dimensions measured from a skull. The cartilage was modeled as a nonlinear spring. Damping of the jaw motion due to friction, which originates from surrounding soft tissues, was included in the model. The forces on the ligaments were neglected.

Using this model, Koolstra and van Eijens performed simulations of *jaw-closing* as a result of isotonic forces (10 N) generated by various pairs of masticatory muscles (52). This model demonstrated that the normally observed swing-slide condylar movement along the articular eminence can be generated by various masticatory muscles. However, different parts of the masseter and the medial pterygoid muscle seemed to be most suitable for this motion (52).

The mastication forces are controlled by the central nervous system, and they are modulated by the physical limitations of the sarcomeres, the force generating units of the muscles (53). The force generated by the sarcomeres depends on the length and contraction velocity of the sarcomeres. In addition, a passive elastic force is generated depending on the amount of stretching. These are determined empirically and expressed as the force-length (F_L), force-velocity (F_v), and stretching (F_p) factors (53). The magnitude of the instantaneous muscle force $F(t)$ can then be formulated as (54):

$$F(t) = F_{\max}[A(t)F_L(t)F_v(t) + F_p(p)] \tag{12}$$

where $A(t)$ is the acitivation level of the muscle.

The jaw-opening and closing simulations were performed by Koolstra and van Eijens using equation 12 in their dynamic model (54). In this work, they included the jawopening muscles, in addition to the closing muscles. The activation level $A(t)$ was kept constant, and all muscles were activated simultaneously. It was found that the level of activation of the temporalis muscle parts was critical in jaw-closing movements. The amount of jaw opening was limited by the passive forces of the jaw-closing muscles. However, the amount of jaw closing was not significantly influenced by the passive forces of the jaw opening (54). The TMJ remained loaded throughout the jaw movements. They also concluded that the average moments generated by the jaw-closing muscles, with respect to the center of mass of the mandible, are responsible for stable operation of the TMJ (52,54).

The lateral deviation of the jaw from the closed position was also modeled by Koolstra and van Eijens (24). They found that laterodeviations that conform to the naturally observed ones could be generated by unilateral muscle contractions. Their analysis concluded that movements of the jaw predominantly depend on the orientation of the contributing muscles with respect to the center of mass, and not on the TMJ ligaments or passive elastic muscle properties (24).

Deformations of the Mandible

Forces applied to the mandible cause internal stress response within the mandible. Highest masticatory loads are experienced when the maxillary and mandibular teeth are in contact. During typical chewing, such high load contacts last on the order of 0.1 s for chewing strokes and longer for swallowing. At the end of the day, 15–30 min of high loads are experienced by the masticatory system. persons with parafunctional habits, such as bruxism, may experience longer durations of high levels of loading (4).

The stress and strain distribution in the mandible change in response to the loss of teeth (30), mandibular reconstruction (56,57) TMJ reconstruction, presence of dental implants (58), or prosthesis (59); consequently, the mandible changes its geometry and material property distribution (30). Analysis of the internal stresses in the mandible has, therefore, a potential to aid treatment. The deformations and internal stresses of the masticatory system could be evaluated by modeling the mandible as a curved beam (60), using the photoelastic method (4) or by strain measurements. The nonuniformity of the cross-sectional area and that of the trabeculae of the mandibular bone, and the presence of the teeth, prevent the beam theory from being an accurate tool of analysis. On the other hand, the photoelastic method provides a good visualization of the internal stress distribution (4). But the trabecular architecture is generally not included in the photelastic models. The strain measurements *in vivo* or *in vitro* can only be performed on the surface of the bone.

The finite element method offers a suitable alternative to model the geometric and material variations of the mandible. Detailed analysis using FEA comes at a high computational expense. However, combined with experimental strain measurements and the photoelastic method, for verification purposes, the FEA can be a useful tool; and, it has been used to analyze the internal stress distribution in the mandible (30,61–63). The FEA of a partially edentulous mandible, by Hart et al., showed that the strain distribution in the mandible is extremely complex (30). This work showed that an asymmetrically edentulous mandible could experience different condylar reactions even in bilateral loading. The anterior portion of the ramus, the coronoid process, and the attachment locations of the muscles experience relatively high tensile stresses under various loading conditions. The condyle experiences relatively high compressive stresses. In general, these results were found to be in agreement with photoelastic stress analysis (4,30).

BIOMECHANICAL PROPERTIES OF TOOTH REPLACEMENT MODALITIES

The treatment objectives of prosthodontic services include maintaining and enhancing quality of life by providing the function, comfort, and aesthetics that is compromised or lost due to oral disease. Amount of damage to oral structures and properties of materials are two important factors that dictate treatment techniques. *Restoration* is a broad term applied to any material or prosthesis that restores or replaces lost tooth structure, teeth, or oral tissues (17). Patient desires and expectations aside, depending on the extent of damage or loss, treatment options may be classified as (1) intracoronal restorations, (2) extracoronal restorations, (3) fixed partial dentures, and (4) removable partial dentures. Each of these treatment modalities has several different choices that can be used by the clinician. In addition to these treatment options, dental implants offer an anchoring function for single tooth replacement or support for various denture scenarios.

To meet functional, biological, and esthetic requirements, a "fixed" restoration (filling, crown, bridge) must remain firmly attached to the tooth. In general, *retention form* indicates the feature of a tooth preparation that resists dislodgement of a crown in a vertical direction or along the path of placement, whereas *resistance form* is the feature to resist dislodgement in all other directions. Tooth preparation also involves removal of sufficient tooth material (reduction) so that the "replacement material" can have enough bulk for structural durability.

Intracoronal Restorations

Absence of tooth structure due to caries (decayed tissue), trauma, or developmental defects is the main indication for *intracoronal* restorations. Before placement of such a restoration, a damaged tooth is "prepared" by the dentist to a certain geometric form to remove all caries, to protect the remaining tooth structure, and to minimize the chances of dislodgement of the restorative material during function (i.e., eating and swallowing) and parafunction (i.e., tooth grinding and clenching).

Metal direct filling materials are gold foil and silver amalgam, which are supplied in many different compositions. *Esthetic direct filling* materials, composite resins, are BISGMA acrylic resins filled with inorganic materials.

The preparation of the teeth for metal direct filling restorations is different than for composite resins. Mechanical locks and undercuts provide retention and resistance to gold foil and amalgam restorations. Preparations are usually extended into the dentin layer to have at least 1–1.5 mm thickness of the material and to take advantage of the elasticity difference between dentin and enamel. Composite restorations, on the other hand, require less invasive preparations. Analyses of preparation designs for various clinical scenarios via photoelastic or finite element studies, as well as bench tests on extracted teeth, have been conducted to provide insight to the optimum preparation form for a specific material to protect the tooth from further damage. Strength (15), wear characteristics (64,65), and reaction to temperature changes are examples of important factors that have been investigated.

Inlays are alternatives to direct filling materials. Preparations for this type of restoration require occlusally diverging walls because these restorations are made indirect and require a path of insertion. Gold alloys, composite resins, or porcelain can be used as inlay materials. This type of restoration is indicated in non-stress-bearing, two-surface preparations (mesio-occlusal, disto-occlusal). Their use in stress-bearing, three-surface (mesio-occluso-distal) situations has been associated with a "wedging" effect, risking cusp fractures.

Increased tooth loss or extensive caries necessitates modification of preparations from ideal forms. At this point, the clinician has to either employ auxilliary retention methods for direct fillings via retentive pins or choose cuspal coverage with indirect restorations (onlay, crown).

Extracoronal Restorations

As the damage to a tooth gets more extensive, an extracoronal type restoration (crown and dowel) is indicated. Major factors that affect retention of a crown include, but are not limited to, the degree of taper of the walls of the prepared tooth, total surface area of the luting cement, area of the cement under shear forces, and roughness of the tooth surface. The resistance form of a crown includes consideration of length of the preparation, tooth width, and taper. Crowns can be made of gold alloys, porcelain, or a combination of both(porcelain fused to metal).

Very often an endodontically treated tooth has a very limited amount of coronal dentin. This is in part due to the required access opening to perform the root canal, and in part to previous tooth loss, i.e., existing fillings. In such situations, foundation restorations are indicated. Frequently, these restorations require utilization of *dowels* (posts). A dowel is fitted into a prepared root canal of a natural tooth. When combined with an artificial crown or core, it provides retention and resistance for the restoration. Many different geometric configurations (e.g., custom cast, prefabricated tapered, parallel sided, threaded), materials of different strength and fatigue characteristics, and corrosion resistance properties (gold alloy, stainless steel, titanium alloy, carbon, ceramic, fiber) have been manufactured for this purpose. The effects of post length and diameter on retention have been widely studied.

Fixed Partial Dentures (Bridges)

Fixed partial dentures (FPDs) can be used to replace missing teeth. Whether these FPDs are supported by natural teeth or dental implants presents different biomechanical issues. On natural teeth, variables such as pericemental support of the abutments (teeth that support the FPD), crown-to-root ratio, configuration of the roots (i.e., single vs. multiple and converging vs. diverging) are among the biomechanical factors to be considered. Available space could become an issue, as the span of the restoration, which is dictated by dimensions of the missing tooth/teeth, is inversely proportional to the strength of the restoration. Similarly, height of the abutment teeth influence height of the connector (part of the bridge where missing tooth replacement, pontic(s), connect(s) to the retainers) and its strength. Hence, material selection is very important. Difference in mobility of abutment teeth may affect seating of the restoration. Another clinical scenario involves the presence of an intermediate abutment when a natural tooth located between terminal abutments serves to support a fixed prosthesis. Under occlusal loads, this "pier" abutment acts as an undesired fulcrum, and the abutment with the least retention fails. To overcome this issue, segmental bridges have been used.

Horizontal components of bite forces are transmitted in part from posterior to anterior directions, because of proximal contact of teeth within the same arch. These forces are well tolerated by the bridge system. However, forces that create movement toward buccal or lingual side (cheek or tongue side) may be detrimental.

When implants are used to support and retain a fixed prosthesis, the difference in mobility of implants is less of a concern, because successful osseointegration is assumed and mobility is more likely related to deformation of bone. However, passive fit of the restoration is a difficult task to achieve due to limitations of materials and techniques. Misfit of restorations is reported to cause more maintenance problems, rather than jeopardizing osteointegration. Unlike natural teeth-supported FPDs, the dentist has the control over the number, size, and location of implant placement when an implant-supported FPD is planned (Fig. 9).

Removable Dentures

Removable dentures are used in complete and partial edentulism. Indications for removable partial dentures include, but are not limited to, situations where the edentulous span is very long, periodontal support is reduced, posterior abutments are absent, and tissue loss is excessive. Typically, a removable partial denture has a cast metal framework, which extends from one side of the dental arch to the other, regardless of the location of the missing teeth. This "cross arch stabilization" by the rigid framework helps reduce detrimental horizontal forces on the abutment teeth.

Components of a removable partial denture (RPD) framework include major connectors, rests, direct and indirect retainers, and minor connectors. One of the most important aspects of removable partial denture fabrication involves design of the framework to achieve retention,

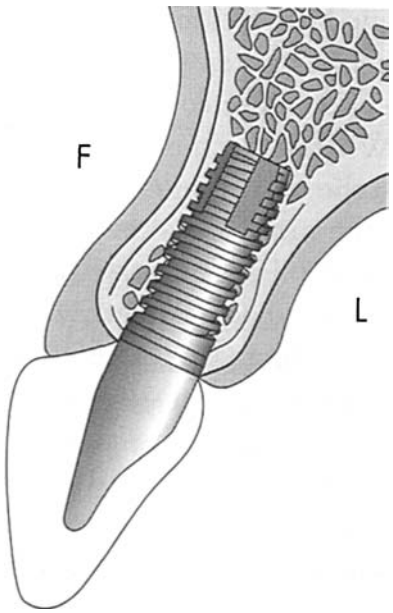


Figure 9. Schematic representation of an osseointegrated, endosseous implant with the abutment and the restorative crown. (From "Single tooth implants," by Misch CE, in Implant Dentistry, 2nd ed., Mosby, 1999).

support, stability, comfort, function, and aesthetics. Removable partial dentures are subject to a combination of forces arising from three different fulcrums on horizontal, sagittal, and vertical planes. As described for other treatment modalities, planning is aimed at directing forces along the long axes of the teeth as much as possible. Factors that influence the amount of stress transmitted to abutment teeth include length of the edentulous span, quality of the ridge support, clasp design flexibility and material, and occlusal harmony.

Tooth loss in an arch varies significantly in number and location; therefore, removable partial dentures are often designed based on available support (tooth supported, or tooth–tissue supported). When the tooth loss pattern allows a tooth-supported RPD, design issues are less complex. However, posterior "free end" scenarios present specific problems because the denture will be supported both by teeth and tissues. Under function, healthy teeth may be displaced as much as 0.2 mm, whereas tissue may be displaced 1.0 mm or more. Thus, prevention of moments on the terminal abutment teeth (teeth that are next to the edentulous space) has been a major concern for tooth–tissue-supported prosthesis. To minimize moments, various clasp assembly designs, resilient precision attachments, and "stress breaking" mechanisms have been proposed.

Recently dental implants have been used to assist RPDs in retention and support.

Complete dentures are indicated in total edentulism. For patients that cannot adapt to complete dentures, use of dental implants is indicated. Retentive mechanisms for implant-assisted dentures vary from individual attachments to bar systems, joining multiple implants. There are many variations to individual attachments as well as to bar systems. These result in different levels of retention, support and freedom, and lack of movement in a specific plane.

Dental Implants

A dental implant is a prosthetic device of alloplastic material implanted into the oral tissues beneath the mucosa and periosteal tissues and into the jaw bone to support a fixed or removable prosthesis. A dental implant system serves as the anchor for the prosthetic reconstruction of missing teeth by supporting a fixed or removable prosthesis. The system mainly consists of an implant and an abutment. *A prosthetic attachment* is typically fixed on the *abutment* by one of the following methods: cementation, use of an occlusal screw, or a socket arrangement that allows retention of a removable prosthesis.

The *implant* is the component implanted into the bone tissue and serves the function of the root. Upon surgical placement of the implant, a healing period of 2–6 months is allowed during which osseointegration takes place. *From the patient's point of view*, "a fixture is considered osseointegrated if it provides a stable and apparently immobile support of a prosthesis under functional loads, without pain, inflammation or loosening" (66). During osseointegration, new bone forms in contact with the implant and a direct structural and functional connection is established, without initiating an immune response (rejection). Mineralized tissue is found to be in contact with the implant surface over most of the surface within nanometers (66). Relative movement (micromotion) between the implant and the bone at the time of placement is more likely to favor development of an fibroosseous interface (28). The presence of the implant modifies the mechanical environment in its surrounding, by altering the normal physiology, distribution of the fluids, and force transmission (28). Nevertheless, dental implants have high long-term success rates (67,68) due to careful bioengineering of the choice of materials (69), surface topography and coatings, overall size and shape of the implant body, and thread shape.

The *abutment* is the component that supports and/or retains the prosthesis (70). The abutment is secured to the implant with a mechanical attachment method, and ideally, it should stay fixed with respect to the implant throughout the life of the implant. Currently, three methods are used to attach an abutment onto an implant. In the most common mechanical attachment method, an abutment retaining-screw is used to fix the abutment with respect to the implant (70). The mechanics of this type of screwattachment are analyzed by classic methods (71) and the FEA (72). Another approach is to use a screw with a relatively large tapered end (73). Finally, in some implant systems, a tapered interference fit between the abutment and the implant is also used to provide the connection (74,75).

From a bioengineering perspective, an important issue is to design the implant with a geometry that will minimize the peak bone stress caused by standard loading (76). The complex geometry of the implants prevents the use of closed-form solutions in stress analysis, where simple formulas relate the effect of external loads to internal stresses

and deformation. The FE method has been applied to the dental implant field to predict stress distribution patterns in the implant–bone interface not only by comparison of various root-form implant designs (76–80), but also by modeling various clinical scenarios (1,58,81,82) and prosthesis designs (32,83–85). This method offers the advantage of solving complex structural problems by dividing them into smaller and simpler interrelated sections by using mathematical techniques (41).

FEAs, which investigate the relation between implant design and stress distribution, have addressed the overall shape and size of the implant body, implant neck geometry, and thread geometry for threaded implants. Rieger et al. showed that a tapered design made of a material with high elastic modulus would be most suitable to serve as a free-standing implant (78). Holmgren et al. suggested considering application of oblique load to FEAs, indicating that these were more realistic occlusion directions capable of causing the highest localized stress in the cortical bone (77). These authors found the stepped implant design to exhibit a more even stress pattern than a straight cylindrical design.

Threaded implants exhibit geometric variations in terms of thread pitch, shape, and depth. Threads are used to increase the surface area of the implant (86). Use of different thread configurations for different bone qualities has been proposed as thread geometry may play an important role in the type of force transmitted (70,87–89). Chun et al. showed that the maximum stress in compact bone is higher for the plateau design compared with the triangular or square designs and their variations. According to these authors, screw pitch had a significant impact on the stress distribution (90).

The transosteal region of the implant body has been defined as the "crest module" (86). For most systems, this neck portion of the implant is smooth. Different designs include parallel, converging, and diverging sides. One particular implant investigated by Hansson using the FE method included both taper and retention elements up to the crest of the implant and was found to have much lower interfacial shear stresses compared with a smooth neck design (76).

BIBLIOGRAPHY

1. Fung YC. Biomechanics, Mechanical Properties of Living Tissues. New York: Springer-Verlag; 1981.
2. Brunski JB. Biomechanics of tooth and jaw. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation. New York: Wiley; 1988. 2776–2788.
3. Koolstra JH, Number crunching with the human masticatory system. J Dent Res 2003;82:672–676.
4. Caputo AA, Standlee JP. Biomechanics in Clinical Dentistry. Chicago: Quintessence Books; 1987.
5. Okeson JP. Management of Temporomandibular Disorders and Occlusion. St. Louis: Mosby; 2003.
6. Koolstra JH. Dynamics of the human masticatory system. Crit Rev Oral Biol Med 2002;13:366–376.
7. Mow VC, Ateshian GA, Spiker RL. Biomechanics of diarthroidal joints: A review of twenty years of progress. J Biomechan Eng 1993;115:460–467.
8. Mow VC, Mak AF. Lubrication of diarthroidal joints. In: Skalak R, Chien S, editors. Handbook of Bioengineering. New York: McGraw-Hill; 1988. 1–34.
9. Osborn JW. The disk of the human temporomandibular joint: Design, function and failure. J Oral Rehabil 1985;12:279–293.
10. van Eijden TMGJ. Biomechanics of the mandible. Crit Rev Oral Biol Med 2000;11:123–136.
11. Woo SL-Y, Mow VC, Lai WM. Biomechanical properties of articular cartilage. In: Skalak R, Chien S, editors. Handbook of Bioengineering. New York: McGraw-Hill; 1988;4:1–44.
12. Anonymous. Dorland's Illustrated Medical Dictionary. 25th ed. Philadelphia: Saunders; 1974.
13. Toparli M, Gökay N, Aksoy T. Analysis of a restored maxillary second premolar tooth by using three-dimensional finite element method. J Oral Rehabil 26:157–164.
14. Rees JS, Jacobsen PH. The effect of cuspal flexure on a buccal Class V restoration: A finite element study. J Dentistry 1998; 26:361–367.
15. Arola D, Galles LA, Sarubin MF. A comparison of the mechanical behavior of posterior teeth with amalgam and composite MOD restorations. J Dentistry 2001;29:63–73.
16. Imanishi A, Nakamura T, Ohyama T, Nakamura T. 3-D finite element analysis of all-ceramic posterior crowns. J Oral Rehabil 2003;30:818–822.
17. Anonymous. The glossary of prosthodontic terms. J Prosthet Dent. 1999;81:39–110.
18. Tanaka E, Sasaki A, Tahmina K, Yamaguchi K, Mori Y, Tanne K. Mechanical properties of human articular disk and its influence on temporomandibular joint loading studied with finite element method. J Oral Rehabil 2001;28:273–279.
19. Beek M, Koolstra JH, van Eijden TMGJ. Human temporomandibular joint disk cartilage as a poroelastic material. Clin Biomech 2002;18:69–76.
20. Williams JA. Engineering Tribology. Oxford: Oxford University Press; 2000.
21. DeVocht JW, Goel VK, Zeitler DH, Lew D. A study of the control of disk movement within the temporomandibular joint using the finite element technique. J Oral Maxillofacial Surg 1996;54:1431–1437.
22. Akca K. H. Iplikcioglu H. Finite element stress analysis of the influence of staggered versus straight placement of dental implants. Int J Oral Maxillofacial Implants 2001;16:722–730.
23. Tanaka E, Rodrigo dP, Tanaka M, Kawaguchi A, Shibazaki T, Tanne K. Stress analysis in the TMJ during jaw opening by use of a three-dimensional finite element model based on magnetic resonance images. Int J Oral Maxillofacial Surg 2001;30:421–430.
24. Koolstra JH, van Eijden TMGJ. Three dimensional dynamical capabilities of the human masticatory muscles. J Biomech 1999;32:145–152.
25. Faulkner MG, Hatcher DC, Hay H. A three-dimensional investigation of temporomandibular joint loading. J Biomech 1987;20:997–1002.
26. Malvern LE. Introduction to the Mechanics of a Continuous Medium. Englewood Cliggs, NJ: Prentice-Hall; 1969.
27. van Rietbergen B, Huiskes R. Elastic constants of cancellous bone. In: Cowin SC, editor. Bone Mechanics Handbook. 2001.
28. Martin RB, Burr DB, Sharkey NA. Skelatal Tissue Mechanics. New York: Springer; 1998.
29. Rigsby DF, Bidez MW, Misch CE. Bone response to mechanical loads. In: Misch CE, editor. Contemporary Implant Dentistry. 2nd ed. St. Louis: Mosby; 1999. 317–328.
30. Hart RT, Henebbel V, Thongpreda N, van Buskirk WC, Anderson RC. Modeling the biomechanics of the mandible: A three dimensional finite element study. J Biomech 1992;25:261–286.
31. Baimonte T, Abbate MF, Pizzarello F, Lozada J, James R. The experimental verification of the efficacy of finite element modeling to dental implant systems. J Oral Implantol 1996;12:104–110.

32. Papavasiliou G, Kamposiora P, Bayne SC, Felton DA. Three-dimensional finite element analysis of stress-distribution around single tooth implants as a function of bony support, prosthesis type, and loading during function. J Prosthet Dent 1996;76:633–640.
33. Sakaguchi RL, Borgersen SE. Nonlinear finite element contact analysis of dental implant components. Int J Oral Maxillo Facial Implant 1993;8:655–661.
34. Teixeria ER, Sato Y, Akagawa Y, Shindo N. A comparative evaluation of mandibular finite element models with different lengths and elements for implant biomechanics. J Oral Rehab 1998;25:299–303.
35. Van Oosterwyck H, Duyck J, Vander Sloten J, Van Der Perre G, De Cooman M, Lievens S, Puers R, Naert I. The influence of bone mechanical properties and implant fixation upon bone loading around oral implants. Clin Oral Impl Res 1998;9:407–418.
36. Mow V, Kuei SC, Lai WM, Armstrong CG. Biphasic creep and stress relaxation of articular cartilage in compression: Theory and experiments. J Biomech Eng 1980;102:73–84.
37. Hu K, Radhakrishnan P, Patel RV, Mao JJ. Nanomechanical and topographic properties of the articular fibrocartillage of the rabbit mandibular condyle. J Struct Biol 2001;136:281–288.
38. Linn FC. Lubrication of animal joints. I, The arthrotripsometer. J Bone Joint Surg 1967;49A:1079.
39. Chen J, Akyuz U, Xu L, Pidaparti RMV. Stress analysis of the human temporomandibular joint. Med Eng Phys 1998;20: 565–572.
40. Chin LPY, Aker FD, Zarrinnia K. The viscoelastic properties of the human temporomandibular joint disk. J Oral Maxillofacial Surg 1996;54:315–318.
41. Belytschko T, Liu WK, Moran B. Nonlinear Finite Elements for Continua and Structures. Chichester: Wiley; 2004.
42. Chen J, Xu L. A finite element analysis of the human temporomandibular joint. J Biomech Eng 1994;116:401–407.
43. Tanaka E, Kazuo T, Sakuda M. A three-dimensional finite element model of the mandible including the TMJ and its application to stress analysis in the TMJ during clenching. Med Eng Phys 1994;16:316–322.
44. Hu K, Qiguo R, Fang J, Mao JJ. Effects of condoylar fibrocartillage on the biomechanical loading of the human temporomandibular joint in a three-dimensional, non-linear finite element model. Med Eng Phys 2003;25:107–113.
45. Viidik A. Properties of tendons and ligaments. In: Skalak R, Chien S, editors. Handbook of Bioengineering. New York: McGraw-Hill; 1988. 1–19.
46. Smith DM, McLachlan KR, McCall WD. A numerical model of temporomandibular joint loading. J Dent Res 1986;65:1046–1052.
47. Osborn JW, Baragar FA. Predicted pattern of human muscle activity during clenching derived from a computer assisted model: Symmetric vertical bite forces. J Biomech 1985;18:599–612.
48. Strang G. Introduction to Linear Algebra. Cambridge, MA: Wellesley-Cambridge Press; 1998.
49. Hatcher DC, Faulkner MG, Hay H. Development of mechanical and mathematic models to study temporomandibular joint loading. J Prosthetic Dentistry 1986;55:377–384.
50. Koolstra JH, van Eijden TMGJ, Weijs WA, Naeije M. A three-dimensional mathematical model of the human masticatory system predicting maximum possible bite forces. J Biomech 1988;21:563–576.
51. Koolstra JH, van Eijden TMGJ. Application and validation of a three-dimensional mathematical model of the human masticatory system in vivo. J Biomech 1992;25:175–187.
52. Koolstra JH, van Eijden TMGJ. Biomechanical analysis of jaw-closing movements. J Dental Res 1995;74:1564–1570.
53. Koolstra JH, van Eijden. Dynamics of the human masticatory muscles during a jaw open-close movement. J Biomech 1997; 30:883–889.
54. Koolstra JH, van Eijden TMGJ. The jaw open-close movements predicted by biomechanical modeling. J Biomech 1997; 30:943–950.
55. Tongue BH, Shepard SD. Dynamics. New York: Wiley;
56. Cox T, Kohn MW, Impelluso T. Computerized analysis of resorbable polymer plates and screws for the rigid fixation of mandibular angle fractures. J Oral Maxillofacial Surg 2003; 61:481–487.
57. Fernandez JR, Gallas M, Burguera M, Viano JM. A three-dimensional numerical simulation of mandible fracture reduction with screwed miniplates. J Biomech 2003;36:329–337.
58. Ishigaki S, Nakano T, Yamada S, Nakamura T, Takashima F. Biomechanical stress in bone surrounding an implant under simulated chewing. Clin Oral Impl Res 2003;14:97–102.
59. Zarone F, Apicella A, Nicolais L, Aversa R, Sorrentino R. Mandibular flexure and stress build-up in mandibular full-arch fixed prostheses supported by osseointegrated implants. Clin Oral Impl Res 2003;14:103–114.
60. Hylander WL. Stress and strain in the mandibular symphasis of primates: A test of competeing hypothesis. Am J Phys Anthro 1984;64:1–46.
61. T.Korioth WP, Hannam AG. Mandibular forces during simulated tooth clenching. J Orofacial Pain 1994;8:178–189.
62. Hirayabashi M, Motoyoshi M, Ishimaru T, Kasai K, Namura S. Stress in mandibular cortical bone during mastication: Biomechanical considerations using a three dimensional finite element method. J Oral Sci 2002;44 1–6.
63. Vollmer D, Meyer U, Joos U, Vegh A, Piffko J. Experimental and finite element study of a human mandible. J Cranio-Maxillofacial Surg 2000;28:91–96.
64. Wassell RW, McCabe JF, Walls AW. A two-body frictional wear test. J Dent Res 1994;73:1546–1553.
65. Wassell RW, McCabe JF, Walls AW. Wear characteristics in a two-body wear test. Dent Mater 1994;10 269–274.
66. Skalak R, Branemark P-I. Definition of osseointegration. In: Branemark P-I, Rydevik BL, Skalak R, editors. Osseointegration in Skelatal Reconstruction and Joint Replacement. Chicago, IL: Quintessenece Books; 1994.
67. Behneke A, Behneke N, d'Hoedt B. The longitudinal clinical effectiveness of ITI solid-screw implants in partially edentulous patients: A 5-year follow-up report. Int J Oral Maxillofacial Implants 2002;15:633–645.
68. Haas R, Polak C, Furhauser R, Mailath-Pokorny G, Dortbudak O, Watzek G. A long-term follow-up of 76 Branemark single-tooth implants. Clin Oral Implants Res 2002;13 38–43.
69. Lemons JE, Dietsh-Misch F. Biomaterials for dental implants. In: Misch CE, editor. Contemporary Implant Dentistry. 2nd ed., St. Louis: Mosby; 1999. 271–302.
70. Misch CE, Hoar J, Beck G, Hazen R, Misch CM. A bone quality-based implant system: A preliminary report of stage I & stage II. Implant Dent 1998;7:35–42.
71. Shigley JE, Mischke CR. Mechanical Engineering Design. 5th ed. Cambridge, MA: McGraw Hill; 1989. 450–457.
72. Lang LA, Kang B, Wang RF, Lang BR. Finite element analysis to determine implant preload. J Prosthet Dent 90:539–546. 2003.
73. Bozkaya D, Müftü S. Mechanics of the taper integrated screwed-In (TIS) abutments used in dental implants. J Biomech 2005;38:87–97.
74. Bozkaya D, Müftü S. Mechanics of the tapered interference fit in dental implants. J Biomech 2003;36:1649–1658.

75. Bozkaya D, Müftü S. Efficiency considerations for the purely tapered interference fit (TIF) abutments used in dental implants. J Biomech Eng 2004;126:393–401.
76. Hansson S. The implant neck: smooth or provided with retention elements: A biomechanical approach. Clin Oral Implants Res 1999;10:394–405.
77. Holmgren EP, Seckinger RJ, Kilgren LM, Mante F. Evaluating parameters of osseointegrated dental implants using finite element analysis—a two-dimensional comparative study examining the effects of implant diameter, implant shape, and load direction. J Oral Implantol 1998;24:80–88.
78. Rieger MR. Finite element stress analysis of root-form implants. J Oral Implantol 1988.; 14:472–484.
79. Rieger MR, Adams WK, Kinzel GL. A finite element survey of eleven endosseous implants. J Prosthet Dent 1990;63:457–465.
80. Siegele D, Soltesz U. Numerical investigations of the influence of implant shape on stress distribution in the jaw bone. Int J Oral Maxillofacial Implants 1989;4:333–340.
81. Pierrisnard L, Hure G, Barquins M, Chappard D. Two dental implants designed for immediate loading: A finite element analysis. Int J Oral Maxillofacial Implants 2002;17:353–362.
82. Van Oosterwyck H, Duyck J, Vander Sloten J, Van Der Perre G, Naert I. Peri-implant bone tissue strains in cases of dehiscence: A finite element study. Clin Oral Implants Res 2002; 13:327–333.
83. Papavasiliou G, Tripodakis AP, Kamposiora P, Strub JR, Bayne S. Finite element analysis of ceramic abutment-restoration combinations for osseointegrated implants. Int J Prosthodont 1996;9:254–260.
84. Stegaroiu R. Kusakari H, Nishiyama S, Miyakawa O. Influence of prosthesis material on stress distribution in bone and implant: A 3-dimensional finite element analysis. Int J Oral Maxillofacial Implants 1998;13:781–790.
85. Stegaroiu R, Sato T, Kusakari H, Miyakawa O. Influence of restoration type on stress distribution in bone around implants: A three-dimensional finite element analysis. Int J Oral Maxillofacial Implants 1998;13:82–90.
86. Misch CE, Waren Bidez M. A scientific rationale for dental implant design. In: Misch CE, editor. Contemporary Implant Dentistry. St. Louis: Mosby; 1999. 329–343.
87. Misch CE, Waren Bidez M, Sharawy M. A bioengineered implant for a predetermined bone cellular response to loading forces. A literature review and case report. J Periodontol 2001; 72:1276–1286.
88. Misch CE, Dietsh-Misch F, Hoar J, Beck G, Hazen R. A bone quality-based implant system: first year of prosthetic loading. J Oral Implantol 1999;25:185–197.
89. Misch CE. Implant design considerations for the posterior regions of the mouth. Implant Dent 1999;8:376–386.
90. Chun HJ, Cheong SY, Han JH, Heo SJ, Chung JP. Evaluation of design parameters of osseointegrated dental implants using finite element analysis. J Oral Rehabil 2002;29:565–574.

See also BONE AND TEETH, PROPERTIES OF; BIOMATERIALS FOR DENTISTRY; JOINTS, BIOMECHANICS OF.

TOTAL ARTIFICIAL HEART. See HEART, ARTIFICIAL.

TOTAL JOINT PROSTHESES. See MATERIALS AND DESIGN FOR ORTHOPEDIC DEVICES.

TOTAL PARENTERAL NUTRITION. See NUTRITION, PARENTERAL.

TRACER KINETICS

EDWARD V. R. DI BELLA
University of Utah
Salt Lake City, Utah

INTRODUCTION

Tracer kinetics refers to the changing distribution of a tracer that provides information regarding a biological system. The tracer may be introduced by injection into the bloodstream, or inhaled, or in some cases be native. By definition, the tracer is present in relatively small (trace) amounts so that it has little or no effect on the system being studied.

A large variety of tracers, along with methods for their measurement, are employed daily in numerous medical applications. Imaging function with tracers can provide complementary or more relevant information than anatomical imaging. For example, positron emission tomography (PET) and single photon emission computed tomography (SPECT) imaging can track radiolabeled tracers to detect and stage cancer, to assess neural function, to follow response to therapy, and to detect heart disease. Magnetic resonance imaging (MRI) with paramagnetic contrast acting as a tracer is a rapidly growing area, as is ultrasound with tracer-like microbubbles. Over 1 million patients are imaged annually in the United States for oncology applications. Over 7 million/year have a cardiac SPECT scan, mostly for the diagnosis of coronary artery disease.

How do these tracers work? Two brief examples will help to illustrate the process. The first example is of tracer applications that are suitable for analysis at equilibrium. That is, the kinetics portion of Tracer Kinetics is not used directly, but rather serves only to bring the tracer to some equilibrium state. The second example is of a radiolabeled glucose analogue, which is most often analyzed as if it were at equilibrium, but offers more information when its changing distribution over time is considered as well. This will provide a feel for the field of tracer kinetics and for the controversy sometimes surrounding the use of tracer kinetics with quantitative modeling methods, compared to visualization of static images of tracer distribution.

"Static" Tracers

By far, the bulk of clinical radionuclide studies use static images acquired while the tracer distribution is changing slowly or is fixed. Microsphere-type tracers are the simplest case of this type of tracer. These tracers provide a static distribution after their first pass and may be, for example, ^{99m}Tc-labeled albumin, or in studies with animals that will be sacrificed, 10–15 μm diameter carbonized microspheres tagged with radioactive or fluorescent labels. Such tracers lodge in capillaries in proportion to flow. The microspheres must be infused into the left atrium or left ventricle or will all be trapped in the lungs. Absolute flow values in milliliters per minute per gram (mL/min/g) can be computed using these tracers. However, the amount of tracer in the region of interest must be known. For animal studies with microspheres, this is done by excising the

regions of interest after sacrifice. The small tissue regions are counted in a type of well counter if radiolabeled microspheres were used, or digested and read in a fluorimeter if fluorescent microspheres were used.

Absolute flow calculation with microspheres also requires knowledge of the flow to a "virtual" organ. This virtual organ can be at the aortic valve at the output of the left ventricle, in which case it is known that the full amount of tracer enters the virtual organ at a flow rate equal to the cardiac output. Thus if cardiac output is known and the number of microspheres is known, then flow to any region of interest can be computed as cardiac output x (number of microspheres in region)/(number of microspheres injected). The units will be milliliters per minute (mL/min), which can be made into mL/min/g after dividing by the weight of the region of interest.

The virtual organ could instead, for example, be in the femoral artery if a catheter was inserted into the artery and a blood sample withdrawn at a known flow rate during the first pass of the microsphere injection. This knowledge of number of microspheres in the virtual organ given a known flow rate allows computation of the flow rates in other regions.

Other tracers that lodge in position in proportion to flow can provide relative flow measures though it is difficult to provide absolute flow measures. One example of a widely used study is ^{99m}Tc-sestamibi SPECT. Sestamibi is injected after exercise or pharmacological vasodilation, and rapidly equilibrates with deposition in proportion to flow. The distribution then changes very slowly, as it is "stuck" in the myocardium. The timing of the SPECT scan is then not critical. The half-life of radioactive decay of ^{99m}Tc is ~6 h, so decay is not a central issue in the imaging. These myocardial perfusion SPECT scans are done at exercise or with a vasodilator in order to be sensitive to stenosed coronary vessels. An excellent reference for the use of tracers in Nuclear Cardiology, including a chapter on tracer kinetics, is found in Ref. 1.

Dynamic Tracers

In general, a thorough understanding of the kinetics of the particular tracer of interest leads to the best method for processing the measurements of the tracer. Consider fluoro-deoxyglucose (^{18}FDG); ^{18}FDG is a positron emitting tracer of glucose. However, ^{18}FDG is trapped at a certain stage in the glycolysis cycle (Fig. 1). As with glucose, hexokinase catalyzes its conversion from ^{18}FDG to phosphorylated ^{18}FDG, ^{18}FDG-6-PO$_4$. However, ^{18}FDG-6-PO$_4$ does not continue along the glycolysis pathway as does glucose-6-PO$_4$, and very little of it leaks back into the interstitial space and vasculature. The changes in ^{18}FDG concentration over time in an area can be represented with a physiological three compartment model as seen in Fig. 1. Compartment models are defined in detail in the next section. The use of dynamic (time series) images of ^{18}FDG acquired with PET requires the application of tracer kinetic principles to provide semiquantitative or quantitative absolute measurements of regional metabolic rate of glucose usage. A mathematical model that can predict or fit the PET data and provides estimates of meaningful physiological parameters is needed. Such models require the arterial input function (AIF), which is the time–activity function the tissue of interest sees as supplying itself.

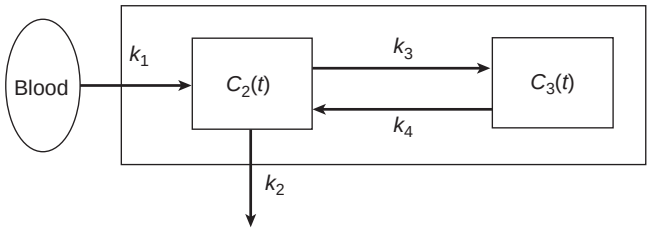


Figure 1. Three-compartment model. For FDG, k_1 represents uptake into a free compartment outside the vasculature. The parameter k_2 is the washout of free FDG back into the vasculature; k_3 represents the rate constant of phosphorylation of FDG to FDG-6-PO4; and k_4 represents the rate constant of dephosphorylation and is relatively small. Note the concentration measured in the tissue, $C_t(t)$ in the text, is $C_t(t) = C_2(t) + C_3(t)$ (illustrated by the largest box).

Yet, in typical clinical ^{18}FDG PET scans, one tomographic volume image is acquired over ~5–30 min, at a time point when the distribution of FDG is changing only slowly. Thus the dynamic data is not used. The images are simply analyzed visually by an experienced reader. Thus at least to date, the complexities of acquiring the data dynamically and measuring an accurate input function and performing the required processing seemingly outweigh the gain of having a quantitative measure of the metabolic rate of glucose usage. This is not the case with some other dynamic tracers such as radiolabeled water, where visual analysis is not useful and kinetic analysis is nearly essential.

Nomenclature

The study of tracer kinetics began prior to 1950, and, due to its myriad applications in a breadth of fields, has had a number of different terms used to describe the same things. Some unified nomenclature was proposed for use with tracer kinetics for biomedical research in 1990 (2). The paper includes a 170 term glossary, but has not prevented the field from having a rich array of discrepant notation and terminology.

Tracer kinetics have been rapidly growing in use with MRI applications. Sufficient research has been published using dynamic contrast enhanced (DCE) MRI with confusing terminology to motivate a paper by MRI researchers regarding terminology for the popular two compartment Kety–Schmidt model. The paper proposes using K^{trans} and k_{ep} for the washin and washout rate constants in the two compartment model (see Fig. 2 and the section on two compartment models below).

Pharmacokinetics is a large related field of study that uses tracer kinetic techniques, sometimes with different nomenclature. The pharmacokinetics field generally seeks to understand drug and biochemical delivery patterns. Drugs of interest may be radiolabeled or tagged with fluorescent or other substances so that they can be tracked and their spatiotemporal behavior identified.

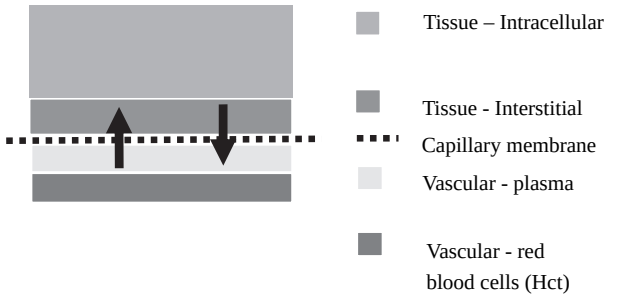


Figure 2. A compartment model type view of a section of capillary (lower two segments) and a section of tissue. The arrows represent the exchange that would occur with an extracellular tracer such as gadolinium. Water would distribute throughout all four regions shown.

COMPARTMENT MODELING

By kinetic modeling we refer to using any of a number of quantitative approaches to extracting information from a time series of data. All of these approaches invoke some sort of mathematical model in order to parameterize the data. The idea is that the parameters can provide more information than processing without models or simple visual examination of the data can provide.

Depending on the type of tracer or contrast agent used, and on the fidelity and resolution of the signal, appropriate models can be selected. That is, the models should not include detail that cannot be discriminated by the measured data, even if it is known to exist physiologically. In practice, almost all models will be required to be simplified versions. The models also need to have useful parameters for the task at hand. Compartment models are widely used for a number of applications due to their relative simplicity and physiological relevance.

A compartment refers to a well-mixed volume in a system of interest that can be considered to have a homogeneous concentration of the tracer. A widely used compartment model of interest is the two compartment model (Fig. 2). This case is often an extremely useful approximation to the system of interest. For example, one compartment is considered to be the vasculature, and one compartment is the tissue of interest. The washin and washout rate constants k_1 and k_2 (or K^{trans} and k_{ep}) govern the concentration in the tissue compartment. The driving compartment, the arterial input function, is usually considered to be measured or otherwise known. For this reason, some authors call this a one compartment model.

Note that a compartment model is often used even in the analysis of multiple heterogeneous regions. Typically, it is assumed that all are driven by the same input function. That is, the blood "compartment" is the same for all of the regions, but the tissue kinetic parameters of uptake and washout of each region may vary. The model remains the same, but the parameters are different for each region.

Kety Model

The roots of compartment modeling approaches are found in Kety's seminal work in the 1940s (3,4). Kety and Schmidt were the first to successfully determine the cerebral blood flow (CBF) and oxygen consumption in a relatively noninvasive way. They had normal volunteers breathe air with a ~5–10% concentration of nitrous oxide. The nitrous oxide that was used as a tracer is an inert gas that is soluble in the brain tissue. They then applied the Fick principle to determine CBF. The Fick principle is simply mass balance for convective systems. Use of the Fick principle gives

$$Q_{\text{b}}(t) = \text{CBF} \int_0^t C_{\text{a}}(\tau) - C_{\text{v}}(\tau) d\tau \tag{1}$$

where Q_{b} is the mass quantity of tracer per 100 g in the brain, CBF is the flow of the tracer (and of the blood) in mL/(min·100 g), C_{a} is the concentration of the tracer in g/mL, and C_{v} is the concentration of the tracer in the venous blood. Alternatively, for experiments with radiolabeled tracers, Q could instead be in terms of activity in Bequerels and concentration in terms of activity/milliliters.

Rearranging and evaluating at a time $t \sim 15$ min when the system is at equilibrium gives the equation Kety used for global CBF:

$$\text{CBF} = \frac{Q_{\text{b}}}{\int_0^t C_{\text{a}}(\tau) - C_{\text{v}}(\tau) d\tau} = \frac{\lambda C_{\text{v}}(t)}{\int_0^t C_{\text{a}}(\tau) - C_{\text{v}}(\tau) d\tau} \tag{2}$$

Here the venous concentration times λ, which is the solubility constant or partition coefficient for nitrous oxide, is used as a measure of the amount of tracer in the brain at equilibrium. The partition coefficient refers to the concentration of the tracer in the tissue divided by the concentration in the blood at equilibrium (at equilibrium arterial and venous blood ideally have the same concentration). Another way to consider the partition coefficient is as a virtual volume, relative to the volume the tracer distributes in a unit of blood. Thus λ is also called the distribution volume. For water, $\lambda = 1$. For nitrous oxide $\lambda = 0.48$ mL/mL. The parameter λ is also sometimes given in units of g/g or mL/g when the density of blood is ~1.06 g/mL or the specific gravity of gray or white matter ~1.05 are used.

For an extracellular tracer such as gadolinium-diethylenetriamine penta-acetic acid (DTPA), $\lambda \sim 0.3$. Figure 2 shows a graphic interpretation of the distribution volume for gadolinium.

In practice, venous measurements of nitrous oxide concentration are taken from an internal jugular vein, and arterial concentrations from a peripheral arterial line. This gives an accurate measure of whole brain CBF, ~50 mL/(min·100 g) in normal humans.

PET $H_2{}^{15}O$ Compartment Model

Interestingly, at first glance Kety's use of the Fick principle does not appear to link to most of today's compartment models. The approach today is often to write a system of differential equations to describe the behavior of the tracer, and then solve the equations to get a form such as:

$$C_t(t) = C_a(t) \otimes k_1 e^{-k_2 t} \tag{3}$$

where ⊗ denotes convolution, and the arterial input function $C_{\text{a}}(t)$ convolved with one or more exponential terms is

fit to the measured image data $C_t(t)$. Though the Kety form has no derivatives or convolutions, Kety's Eq. 2 can be rearranged in a similar form. This rearrangement is illustrated by describing the use of a two compartment model for PET $H_2\,{}^{15}O$. Water rapidly diffuses across capillary and cellular membranes, so one can postulate based on Fig. 2 that the rate of change of concentration in the tissue compartment is equal to the input to the compartment (the rate constant k_1 times the arterial concentration) minus the output from the compartment (C_t times the rate constant k_2).

$$\frac{dC_t}{dt} = k_1 C_\mathrm{a} - k_2 C_t \tag{4}$$

For the tracer water, $k_2 = k_1$ since the capillary membrane appears symmetrical to this tracer and the distribution volume is 1. Solving this first-order differential equation gives Eq. 3. Note that integrating both sides of Eq. 4 gives the form seen in Kety's Eqs. 1 and 2.

The model can be made more general for tracers that are not freely diffusible everywhere (as water is) by including the partition coefficient. That is, for the more general case of diffusible tracers, $k_1 = F$ and $k_2 = F/\lambda$. This formulation allows estimation of both flow to the region of interest (perfusion) and the volume of distribution of the tracer. Both of these measures can be useful clinically.

Tracer Kinetics of Gadolinium Measured with MRI

Less straightforward tracers, such as the paramagnetic moiety Gd-DTPA used in dynamic contrast MRI, can also be approximated with the two compartment model. Gadolinium diffuses out of the vasculature into the interstitial space, but does not enter cells. Thus the distribution volume reflects the size of the interstitial space. This lack of distribution into cells also creates complications when water exchange is considered. That is, Gd-DTPA is not detected directly with MRI as is the case with radionuclide detection. It is detected by how it changes the relaxation rates of hydrogen in its local microenvironment by interaction of its electrons and the hydrogen protons in blood or tissue (5). So, if water (hydrogen) is moving between where the Gd-DTPA is in the interstitial space and the intracellular space free of gadolinium, it experiences different microenvironments and thus different relaxation rates. Assuming the rate of exchange of water molecules across the cell membrane is rapid, one can neglect this effect on the image signal intensity. Otherwise, more complicated modeling including the water exchange is needed.

Another interesting example of tracer kinetics is the use of gadolinium contrast enhanced MRI for determining if an area of cardiac muscle is dead or alive. A static scan ~15 min after contrast injection can show that gadolinium is present at a higher concentration in infarcted tissue compared to normal tissue. This delayed enhancement is most easily understood through a tracer kinetics approach. Figure 3 shows the time curves from a normal and abnormal region and one can see the abnormal region washes in more slowly, but the washout is a slower process. Another way of interpreting this is the distribution volume is larger. This corresponds intuitively to the agent having more

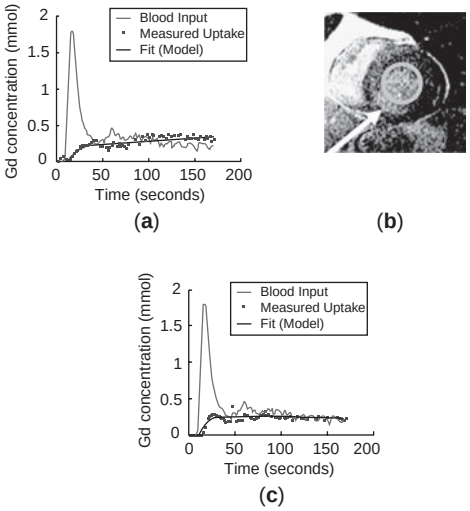


Figure 3. (**a**) Time curves from measured uptake in region corresponding to arrow in the middle image, distribution volume $v_\mathrm{e} = 0.58$ mL/g. The blood input is from a region of interest drawn on the left ventricle cavity. The fit is from a two compartment model for gadolinium-DTPA. (**b**) "Delayed" image of a short axis slice of the heart(right ventricle on left, left ventricle the large doughnut structure with its endocardial wall traced in green) from MRI acquired ~15 min after contrast injection, showing subendocardial inferior infarct (arrow). (**c**) Curve from a remote normal region, $v_e = 0.37$. The magnitude and spatial distribution of v_e may help characterize the infarct better than delayed enhancement alone.

volume to distribute in within regions where cell membranes have been destroyed, since gadolinium is an extracellular tracer.

The volume/g of tissue that is occupied by cells in the heart muscle is estimated as ~70%. The interstitial space, or distribution volume for Gd-DTPA, has been reported as $v_\mathrm{e} \sim 0.3$ mL/g in healthy tissue, where v_e stands for volume of extracellular extravascular space and is used in place of the symbol λ. This distribution volume increases if cell membranes are disrupted as in acute myocardial infarction (heart attack). Scar tissue, composed of a collagen matrix, will still accumulate gadolinium if it is still perfused. This is likely because the collagen matrix has more interstitial space, but is still not completely understood (6).

Another issue for most tracers is that tissue regions of interest contain on the order of 10% of their volume as vasculature. Thus the time curve $C_t(t)$ is biased by a contribution from something similar to $C_\mathrm{a}(t)$. [Due to hematocrit changes in smaller vessels and delay and dispersion, it may not be exactly $C_\mathrm{a}(t)$.] One method of accounting for this is to include another parameter in the fit to estimate the blood volume in the region of interest:

$$C_t(t) = C_\mathrm{a}(t) \otimes k_1 e^{-k_2 t} + v_\mathrm{b} C_\mathrm{a}(t) \tag{5}$$

The v_b parameter can also be used to model blood signal in tissue that arises from partial volume effects. Figure 4 illustrates the use of v_b to account for blood spillover into the tissue in a dynamic cardiac MRI study.

Intravascular tracers that do not leave the vasculature may be used to obtain measures of blood volume and to

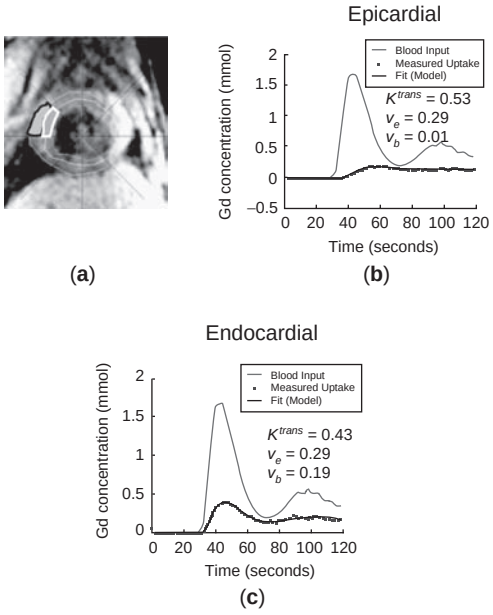


Figure 4. Short axis dynamic cardiac MRI time series study. (**a**) MRI image precontrast. Epicardial (light blue) and endocardial (orange) regions that give the tissue curves in (**b**) and (**c**) are marked. Parts (**b**) and (**c**) use the same blood input function and should have similar flows as implied by the K^{trans} ($K^{\text{trans}} = k_1$) values given. The point here is that due to partial volume effects, typical upslope or percent enhancement semiquantitative measures would incorrectly give very different flow indices for the two regions. The model used Eq. 5 to obtain blood volume v_b values to compensate for partial volume effects.

measure flow and can be analyzed as

$$C_t(t) = C_{\text{a}}(t) \otimes h(t) \tag{6}$$

where in this case $h(t)$ is not parameterized and the peak value of $h(t)$ can be shown to reflect flow (7).

Impulse Response

The response of a system to the input of an impulse, or a delta function $\delta(\,)$, completely characterizes the system if it is a linear time-invariant system. Equation 3 shows that the impulse response for a two compartment model with rate constants k_1 and k_2 is $k_1 * \exp(-k_2 t)$.

While many physiological systems are not linear time invariant, linear time-invariant systems often serve as a good approximation. Examples of nonlinear systems include enzyme kinetics, saturable receptors, or any process that saturates (as the input continues to double the output does not). Too much of any tracer may force a system to nonlinearity, or at least to some abnormal physiological state. For example, at some point there could be so much FDG in the blood stream that doubling the amount only changes the uptake in the tissue by a small amount. More complicated nonlinear and time-varying models are needed for processes such as insulin kinetics (8).

Three or More Compartments

More complex systems, for example, when a tracer diffuses out of the vasculature and may either diffuse back, or remain freely available in the interstitial space, or undergoes a chemical trapping such as phosphorylation, requires a more complex model. The two compartment model may not be sufficient since the tissue now has two different pools of nonuniform concentrations. Note that these pools generally are not measured separately (Fig. 1). The two compartment model can still be used as an approximation, but a three-compartment model (Fig. 1) is likely more appropriate. A more complex model with four compartments has also been proposed for the case of FDG uptake in skeletal muscle (9).

Order of Model

How does one determine if the lower order model or the higher order model is best to use? This is certainly very task dependent, and is an open question. One approach is the Akaike Information Criteria, or AIC (10). This measure is based on the intuitive notion that as the number of parameters increases, the fit to the data should improve in proportion (AIC = residual sum of squares $+2^*P$, where P is the number of parameters being fit).

If the tracer kinetics are not well understood and the order of model to use is unclear, more exploratory methods such as spectral analysis are appropriate. Spectral analysis models the impulse response as a summation of a number of exponentials (11). That is, spectral analysis is a type of generalized compartment model where the impulse response is a linear combination of N decaying exponentials: $h(t) = \sum_{i=1}^{N} \alpha_i \exp(-\beta_i t)$. For example, if the data truly is from a two compartment model, then only one of the exponential terms will have a nonzero coefficient. The drawback, or advantage, depending on one's viewpoint for a particular problem, is that the coefficients are fixed so that the problem becomes linear as only 's need be estimated. The selection of the spectrum of exponentional decay rates can thus be quite important. A logarithmic distribution of has been used in PET (11).

Even if the model is well-known, spectral analysis may be a desirable approach since the linearized fitting can be more rapid than standard nonlinear compartment modeling. Several published studies in PET and SPECT have found high correlations between compartment modeling and spectral analysis methods. Such studies are reviewed in Ref. 12.

A higher level of realism can be obtained by using a distributed model of multiple parallel pathways of capillaries feeding the tissue region of interest. XSIM is freely available software useful for modeling such distributed systems where blood–tissue exchange units form basic building blocks that can have differing delays or flows or permeability (13). Most often the measurement system, particularly medical imaging, does not have the spatial and temporal resolution and is too noisy to robustly use such a complex model. Such models are, however, of great importance in understanding the underlying physiology (8,13).

Model-Free Methods

It is not required to use a specific model. Some "model-free" methods are easier to use or have the advantage of making fewer assumptions regarding the tracer distributions. The

downside is that the parameters may not offer as much insight into physiological mechanisms underlying the imaging results.

The mean transit time is a measure of how long the tracer requires to pass through the capillary bed from the arterial to the venous side. One can consider the impulse response as a histogram of transit times through the region. The expected value of the impulse response provides the mean transit time if there is no recirculation [$\tau = \int th(t)dt$]. Or, for a short bolus, the area divided by the height at time zero of the tissue uptake time curve gives τ.

$$\tau = \int_0^\infty C_t(t)dt / C_t(0)$$

The mean transit time is on the order of 5 s for nondiffusible intravascular tracers in the brain, and on the order of minutes for diffusible tracers (14). Or one can define the mean transit time as the inverse of the washin rate constant, $1/k_1$(14).

Other easily computable parameters such as maximum upslope, percent signal enhancement, area under the curve, or the time until peak signal have been suggested and used as model-free approaches.

ESTIMATING KINETIC PARAMETERS

Weighted Least Squares

The estimation of kinetic parameters from noisy data is a well-studied problem. The errors in the parameters can be estimated based on the variance in the data. The model can be written as:

$$C_t(t) = h(t) \otimes C_a(t) + \varepsilon \tag{7}$$

where ε is an additive noise process. The impulse response can be cast in a toeplitz matrix H and the equations expressed

$$\boldsymbol{y} = H\boldsymbol{b} + \varepsilon \tag{8}$$

where $\boldsymbol{y} = C_t(t)$, possibly for multiple regions, and $\boldsymbol{b} = C_a(t)$, and the unknowns are the parameters $\boldsymbol{p}$ that form H. The weighted least-squares fitting process minimizes $\boldsymbol{y} - H\boldsymbol{b}$, weighted by the data covariance matrix $\Phi = \text{cov}(\mathbf{y})$:

$$\min_p (\boldsymbol{y} - H\boldsymbol{b})^T \Phi^{-1} (\boldsymbol{y} - H\boldsymbol{b}) \tag{9}$$

If the measured data $\boldsymbol{y}$ have uniform Gaussian noise then Φ^{-1}will be a diagonal matrix with on the diagonal and will not be any different from an unweighted fit. Technically, Φ should be the covariance matrix of the residuals $\boldsymbol{y} - H\boldsymbol{b}$, since $\boldsymbol{b}$ is typically a measured quantity with noise. Huesman and Mazoyer compared the results with the use of cov($\boldsymbol{y}$) and cov($\boldsymbol{y} - H\boldsymbol{b}$) as the weighting function in the fitting of dynamic PET data with a noisy arterial input function (15).

One can consider the system of equations (Eq. 8) in the more "standard" form of $\boldsymbol{y} = A\boldsymbol{p}$, where $\boldsymbol{p}$ are the unknown parameters. If we assume the model is linear in the parameters (i.e., $d\boldsymbol{y}/d\boldsymbol{p} = A$) then the best linear unbiased estimate, the estimate with minimum variance (16), is given by

$$\boldsymbol{p} = (A^T \Phi^{-1} A)^{-1} A^T \Phi^{-1} \boldsymbol{y} \tag{10}$$

This gives the parameters that minimize the cov($\boldsymbol{p}$) matrix $E[(p - \hat{p})(p - \hat{p})^T]$. The point is that given this form, one can show that the resulting cov matrix of the parameter estimates is the Fisher information matrix

$$\text{cov}(\boldsymbol{p}) = (A^T \Phi^{-1} A) \tag{11}$$

where A are the partial derivatives with respect to the parameters of the model, also called the senstivity matrix.

More commonly, and for the case of a compartment model, the matrix A will be a function of the parameters. So the model will be nonlinear in the parameters. Equation 11 still holds if A is evaluated at the true parameters $\boldsymbol{p}$, and the measured data are Gaussian (17).

The cov($\boldsymbol{p}$) is the combination of the sensitivity matrix A and the data covariance matrix Φ. In addition to providing a way to estimate the error in the fitted parameters, one can also compare data acquisition strategies in this manner. Mazoyer and Huesman compared input function shapes and temporal sampling strategies by comparing the determinants of their information matrices (17).

Note that Eq. 8 need not only represent a single measured region. The measured curves $\boldsymbol{y}$ and impulse responses can be stacked to form a larger system of equations using the same AIF $\boldsymbol{b}$ and different parameters for each region. (see, e.g., Ref. 18).

Bayesian Approaches

Two quite different Bayesian approaches have been published. One class of methods combines tomographic reconstruction with compartment models (19,20). The method in (19) assumes the AIF is known a priori and then encourages the reconstructed time courses of each voxel to fit a two compartment model. The reconstruction is performed simultaneously to satisfy the tomographic reconstruction criteria that the projections of the estimated image match the measured projections. The method was applied to dynamic teboroxime SPECT data and resulted in improved washin estimates.

Another type of Bayesian approach is to use multiple (normal) regions to estimate the kinetic parameters using a two step process (21). The first iteration calculates the parameters for the regions independently. These are then averaged and their covariance matrix computed and used in the second step in a penalized least squares, or Bayesian, formulation:

$$\min_p (y - Hb)^T \Phi^{-1} (y - Hb) + (p - \mu)^T \Omega^{-1} (p - \mu) \tag{12}$$

where $\boldsymbol{p}$ are the parameters to be estimated, is the average of the parameters estimated in the standard fashion, and

$$\Omega = 1/M \sum_{m=1}^{M} (p^m - \mu)(p^m - \mu) + 1/M \sum_{m=1}^{M} (A^m \Phi^{m^{-1}} A^{m^T})^{-1} \tag{13}$$

That is, Ω is the standard deviation of the parameters estimated in the first step and the estimated uncertainty

of the parameter estimates. The parameter M is the number of regions (assumed homogeneous) and m is used to index the M regions. This approach is equivalent to a specific case of ridge regression (22). These Bayesian methods represent a new class of analysis techniques for tracer kinetics.

IMPROVING AND AUTOMATING ANALYSIS OF THE KINETIC TRACER CURVES

While some applications have sufficient signal-to-noise ratio (SNR) to perform pixelwise analysis of the kinetic tracer curves, other applications can benefit greatly from spatial averages of the data. Methods such as factor analysis and clustering can be useful for automatically identifying blood and tissue curves from the image series data (23). This approach may give better kinetic parameter estimates and be more automated than manual delineation of regions.

Factor Analysis

Factor analysis, also termed FADS, which stands for Factor Analysis of Dynamic Structures, has been somewhat widely used in nuclear medicine applications (24–27). The FADS is essentially principal components analysis followed by a "rotation" to satisfy nonnegativity conditions. While this does not provide a unique solution, several different FADS methods have been validated as clinically useful for certain tasks. As well, a number of modifications to the basic method to incorporate additional information have been proposed, for example, (27).

Clustering

Clustering is becoming a popular method for a data-driven approach to choosing the regions. K-means clustering, also termed c-means clustering, groups alike time–activity curves (or time-concentration curves) by computing a distance between each pair of curves and grouping curves into the cluster that they are closest to. The cluster centers are updated (continuously or at discrete intervals). Most often a Euclidean distance is used to determine to which cluster a voxel belongs.

It is still an open issue to make factor analysis, clustering, and automated approaches robust and clinically practical. Quantitative imaging finds particular importance for longitudinal studies. These studies also need robust registration and processing methods to perform well.

Since a number of the analysis methods are nonlinear, it is typically left to empirical task-specific studies to determine if specific image processing methods can significantly impact the accuracy or robustness of the results.

ISSUES WITH THE ARTERIAL INPUT FUNCTION, AIF

Importance of an Accurate AIF

In order to obtain accurate absolute measures of kinetic parameters, it is important to use an accurate arterial input function. This is critical for any of the models or model-free approaches that use the AIF. Accurate input functions, or at least some measure of what the tissue "sees" is essential for obtaining quantitative parameters and is intrinsic to model-based methods. Many "model-free" methods also rely on measures of input and output to obtain the mean transit time from the images. As other issues key to acquiring good quality dynamic data (scanners, computer speed, and memory) are surmounted, more effort is being focused on estimating the AIF more accurately and automatically.

In many situations, it is difficult to obtain an accurate AIF. If the tracer of interest binds to red blood cells or plasma proteins, then the AIF obtained from the images does not reflect the concentration of the tracer that is available for uptake by the tissues. Or the contrast agent in MRI may be saturated and not provide a true measure of gadolinium concentration when the concentration is high, or there may be flow effects that change the MRI signal. Arterial blood samples and subsequent processing are required to obtain accurate AIFs in these settings.

Methods for Improving the AIF Measurement

A number of techniques for improving the AIF have been proposed. For example, in contrast MRI studies of the brain, either a cluster or a single voxel that meets certain criteria can be automatically obtained and used as the AIF. In MRI cardiac studies, either a small bolus given before the main bolus to give the shape of the AIF curve without saturation, or a pulse sequence modification to obtain a nonsaturated AIF are current methods to improve AIF estimation.

Blind Deconvolution—Methods for Estimating an Unmeasured AIF

There have recently been efforts to estimate the arterial input function jointly with the parameters of interest. That is, given the measured data $\boldsymbol{y}$, estimate both H and $\boldsymbol{b}$ in the equation:

$$\boldsymbol{y} = H\boldsymbol{b} + \varepsilon \tag{14}$$

In the field of telecommunications, this is termed blind estimation or blind deconvolution, since the input is not known (28). While this is not possible given a single region of interest, multiple regions with differing kinetic parameters, driven by the same input, can provide a system of equations with sufficient measurements to estimate the AIF within a global scale factor. Then, Eq. 14 is composed of a number of stacked matrices (18). Another way to see this is to consider two regions without using the matrix formulation:

$$\begin{aligned} y_1(t) &= b(t) \otimes h_1(t) \\ y_2(t) &= b(t) \otimes h_2(t) \end{aligned} \tag{15}$$

Then convolving both sides of Eq. 16 with $h_2(t)$ and substituting from Eq. 12 gives the "cross-relation" expression (29):

$$\begin{aligned} h_2(t) \otimes y_1(t) &= h_2 \underbrace{(t) \otimes b}(t) \otimes h_1(t) \\ y_1(t) \otimes h_2(t) &= y_2(t) \otimes h_1(t) \end{aligned}$$

This last expression can be solved for $\boldsymbol{p}$ (recall h is a function of $\boldsymbol{p}$) with no knowledge of the AIF! The solution is unique with a two compartment model and the use of two different regions, to within a global scale factor (18). A three-compartment model requires more than two regions (or the use of other constraints) to obtain a unique solution (30).

The blind deconvolution method is sensitive to noise (18). Future hybrid approaches incorporating some measurements and other a priori information such as population expectations with the blind deconvolution technique will likely result in less sensitivity to noise and artifact.

FUTURE ISSUES

The last few topics—the Bayesian fitting approaches, factor analysis, clustering, and blind deconvolution represent cutting-edge directions for analysis of tracer kinetics data. These techniques need more research and validation before being widely used in clinical applications.

More and more dynamic studies of tracer kinetics will be performed as the equipment (in particular computer speed and disk space) allows dynamic acquisition and processing with smaller penalties. While the analysis of the studies will be task dependent, insight into the fundamentals of acquiring and analyzing tracer kinetics will be applicable to all fields.

Tracer development is perhaps in its highest gear ever since the largest advances, particularly in molecular imaging, are likely to result from new, more specific, and more physiologically relevant tracers. A hot area is the design of tracers to follow genes of interest, reporter genes, and to track stem cells and disease processes (31).

While the ideal tracer may go in proportion to the site of interest and be stuck there so that static imaging provides full information, it is much more likely that tracers will continue to exhibit complex properties. Thus research into optimal acquisition and analysis of the temporal and spatial distributions of the tracer will remain an important field and will likely even grow in significance. Analysis of dynamic data can offer quantitative and possibly even absolute measures of many clinically important parameters. The use of tracer kinetic principles and image time series from modalities such as optical, MRI, ultrasound, PET, and SPECT will continue to have very important application to cardiac, brain, renal, liver, and other dynamic systems.

BIBLIOGRAPHY

1. Zaret BL, Beller GA. Clinical Nuclear Cardiology: State of the Art and Future Directions. 3rd ed. New York: Mosby-Year Book; 2005.
2. Rescigno A, Thakur AK, Brill AB, Mariani G. Tracer kinetics: A proposal for unified symbols and nomenclature. Phys Med Biol 1990;35:449–465.
3. Kety SS, Schmidt CF. The nitrous oxide method for the quantitative determination of cerebral blood flow in man: Theory, procedure, and normal values. J Clin Invest 1948;27:476–483.
4. Kety SS. The theory and applications of the exchange of inert gas at the lungs and tissues. Pharmacol Rev 1951;3:1–41.
5. Donahue KM, Burstein D, Manning WJ, Gray ML. Studies of Gd-DTPA relaxivity and proton exchange rates in tissue. Mag Res Med 1994;32:66–76.
6. Thomson LE, Kim RJ, Judd RM. Magnetic resonance imaging for the assessment of myocardial viability. J Mag Res Imaging 2004;19:771–788.
7. Ostergaard L, Weisskoff RM, Chesler DA, Gyldensted C, Rosen BR. High resolution measurement of cerebral blood flow using intravascular tracer bolus passages. Part I: Mathematical approach and statistical analysis. Mag Res Med 1996;36:715–725.
8. Carson E, Cobelli C. Modelling Methodology for Physiology and Medicine. New York: Academic Press; 2001.
9. Bertoldo A, Peltoniemi P, Oikonen V, Knuuti J, Nuutila P, Cobelli C. Kinetic modeling of [18F] FDG in skeletal muscle by PET: A four-compartment five-rate-constant model. Am J Physiol Endocrinol Metab 2001;281:E524–536.
10. Akaike H. A new look at the statistical model identification. IEEE Trans Auto Control 1974;19:716–723.
11. Cunningham V, Jones T. Spectral analysis of dynamic PET studies. J Cereb Blood Flow Metab 1993;13:15–23.
12. Murase K. Spectral analysis: Principle and clinical applications. Ann Nuclear Med 2003;17:427–434.
13. Bassingthwaighte JB, Goresky CA. Modeling in the analysis of solute and water exchange in the microvasculature. In: Renkin EM, Michel CC, editors. Handbook of Physiology—The Cardiovascular System. Bethesda: American Physiology Society; 1984.
14. Cherry SR, Sorenson JA, Phelps ME. Physics in Nuclear Medicine. 3rd ed. Saunders; 2003.
15. Huesman RH, Mazoyer BM. Kinetic data analysis with a noisy input function. Phys Med Biol 1987;32:1569–79.
16. Strang G. Introduction to Applied Mathematics. Wellesley: Wellesley-Cambridge Press; 1986.
17. Mazoyer BM, Huesman RH, Budinger TF, Knittel BL. Dynamic PET data analysis. J Comput Assist Tomogr 1986;10:645–653.
18. Riabkov DY, DiBella EVR. Estimation of kinetic parameters without input functions: Analysis of three methods for multichannel blind identification. IEEE Trans Biomed Eng 2002;49:1318–1327.
19. Kadrmas DJ, Gullberg GT. 4D maximum a posteriori reconstruction in dynamic SPECT using a compartmental model-based prior. Phys Med Biol 2001;46:1553–1574.
20. Meikle SR, Mattherw JC, Cunningham VJ, Bailey DL, Livieratos L, Jones T, Price P. Parametric image reconstruction using spectral analysis of PET projection data. Phys Med Biol 1998;43:651–666.
21. Bertoldo A, Sparacino G, Cobelli C. Population approach improves parameter estimation of kinetic models from dynamic PET data. IEEE Trans Med Imag 2004;23:297–306.
22. O'Sullivan F, Saha A. Use of ridge regression for improved estimation of kinetic constants from PET data. IEEE Trans Med Imaging 1999;18:115–125.
23. DiBella EVR, Sitek A. Time curve analysis techniques for dynamic contrast MRI studies. Information Processing in Medical Imaging 2001;LNCS 2082:211–217.
24. Barber DC. The use of principal components in the quantitative analysis of gamma camera dynamic studies. Phys Med Biol 1980;25:283–292.
25. Samal M, Karny M, Surova H, Marikova E, Dienstbier Z. Rotation to simple structure in factor analysis of dynamic radionuclide studies. Phys Med Biol 1987;32:371–82.
26. Buvat I, Benali H, DiPaola R. Statistical distribution of factors and factor images in factor analysis of medical image sequences. Phys Med Biol 1998;43:1695–1711.

27. Sitek A, DiBella EVR, Gullberg GT. Factor analysis with *a priori* knowledge—application in dynamic cardiac SPECT. Phys Med Biol 2000;45:2619–2638.
28. Tong L, Perreau S. Multichannel blind identification: from subspace to maximum likelihood methods. Proc IEEE 1998;86:1951–1968.
29. DiBella EVR, Clackdoyle R, Gullberg GT. Blind estimation of compartmental model parameters. Phys Med Biol 1999;44:765–780.
30. Riabkov DY, DiBella EVR. Blind identification of the kinetic parameters in three-compartment models. Phys Med Biol 2004;49:639–664.
31. Massoud TF, Gambhir SS. Molecular imaging in living subjects: seeing fundamental biological processes in a new light. Genes Dev 2003;17:545–580.

See also FLOWMETERS, ELECTROMAGNETIC; PHARMACOKINETICS AND PHARMACODYNAMICS; PHYSIOLOGICAL SYSTEMS MODELING; RADIONUCLIDE PRODUCTION AND RADIOACTIVE DECAY; STATISTICAL METHODS.

TRANSCUTANEOUS ELECTRICAL NERVE STIMULATION (TENS)

ANJU MADNANI
SANJAY MADNANI
RANDALL CARK
JASON WELLS
LSU Medical Centre
Shreveport, Louisiana

INTRODUCTION

The field of pain management is a rapidly expanding one, and new treatment modalities are being discovered or rediscovered. Electroanalgesia has a long and sometimes dubious history, dating back to the ancient Egyptians. However, the publishing of "the gate control theory" of pain transmission in 1965 by Melzack and Wall transformed our understanding of pain, its transmission, and how it is modulated. With this discovery, electroanalgesia underwent a revolution, and transcutaneous electrical nerve stimulation (TENS) was developed and is continually being refined. Today, our understanding of the mechanism by which TENS produces analgesia continues to expand does its potential applications.

This article provides a review of pain, its definition, types, and physiology. It provides background information and theories surrounding the mechanism of analgesic action of TENS and the development of electroanalgesia. It discusses the usage, design, applications, and warnings surrounding TENS.

WHAT IS PAIN?

Pain is an unpleasant sensory and emotional experience associated with actual or potential damage, or described in terms of such damage. Pain serves as an essential defense mechanism to protect one's body from potential damage. Indeed, the disastrous consequences of diminished or absent pain signaling become readily apparent in diseases and conditions that result in partial or complete damage of the nerves that innervate the extremities (e.g., diabetic neuropathy, tables dorsalis, tuberculoid leprosy, and many others). While serving an essential function, pain can often present for physiologically inappropriate reasons, continue far past the removal of noxious stimuli, remain long after wound healing, or even present for purely psychological reasons. This maladaptive and uncontrolled pain cycle afflicts an estimated 40 million Americans (1), and research into the causes and cures of pain is a rapidly expanding branch of medical science and forms the basis for a multibillion dollar a year, multidisciplinary industry.

Pain can be categorized either temporally as in acute or chronic pain or by the mechanism. Nocioceptive–inflammatory pain is produced after an appropriately perceived tissue injury. Neuropathic pain, however, is produced by nerve injury that is inappropriately perceived due to neuroplasticity. Often described as a burning or electric sensation, neuropathic pain can persist long after an injury or for completely idiopathic reasons. Even simple light touch or changes in temperature are enough to trigger severe bouts of extreme pain, lasting seconds to hours or longer (i.e., trigeminal neuralgia).

Phantom limb pain is another incompletely understood neuropathic phenomenon and occurs in 50–67% of postsurgical amputation patients (2). It is often described as a minor-to-severe cramping or, less commonly, as a burning sensation (3). While this commonly subsides with time, in $\sim$10% of patients, this pain persists and is often refractory to NSAID or opiate therapy, traditional first and second line agents in the treatment of pain.

THE PHYSIOLOGY OF PAIN

The process of nocioception is complicated, but can be divided into four distinct physiological processes transduction, transmission, modulation, and perception. Transduction, the translation of noxious stimuli into electrical activity at the sensory endings of nerves, occurs at unspecialized mechano-, thermo-, or polymodal (thermal and chemical) nociceptors, as well as at unspecialized nerve endings.

Polymodal nociceptors respond to a variety (i.e., chemical, mechanical, and temperature extremes) of intense noxious stimuli. Thermonociceptors are distinct from thermoreceptors that transmit non-noxious temperature information. This class of nocioreceptors functions from temperature ranges of roughly <5 to >45 $^{\circ}$C. Mechanonociceptors are activated when intense pressure stimulates them; as with thermonociceptors, the mechanonociceptors are distinct from the receptors that transmit non-noxious light and strong touch, vibratory information, and so on. Additionally, visceral "silent" nociceptors exist in a default dormant state and are usually activated only in the presence of inflammatory mediators. These silent nociceptors likely are involved in hyperalgesia as discussed below.

In the peripheral nervous system, small unmyelinated C polymodal nocioceptive fibers, as well as the larger,

lightly myelinated Aδ mechano- and thermonocioceptive fibers transmit noxious stimuli to the dorsal horn of the spinal column. The small C fibers are responsible for what is termed slow pain and transmit data at under $2.5\,m{\cdot}s^{-1}$. These small fibers outnumber the larger, lightly myelinated Aδ fibers, responsible for fast pain, which conduct at a rate of $4–30\,m{\cdot}s^{-1}$, by a ratio of $\sim 7:1$ in the epithelium. The concept of slow and fast pain is easily conceptualized by a hypothetical injury of one stepping on a nail. The initial sharp sensation, or fast pain, is transmitted by the larger Aδ fibers, while the nagging dull ache, or slow pain, is transmitted by the smaller, unmyelinated C fibers (Fig. 1).

The Aδ fibers synapse with projecting neurons in lamina 1 of the dorsal horn of the spinal cord. In addition to this direct, afferent input, these projecting neurons receive indirect input from the stalk cell neurons in lamina II. These stalk cell interneurons of lamina II receive their afferent input from the C fibers that synapse with them. The projecting neurons of lamina V receive afferent input from the large myelinated Aβ, non-noxious, sensory fibers via dendtritic synapse in lamina IV, from synapse with Aδ fibers in lamina V, and project both to lamina III as well as higher cortical centers (5,6) (Fig. 2).

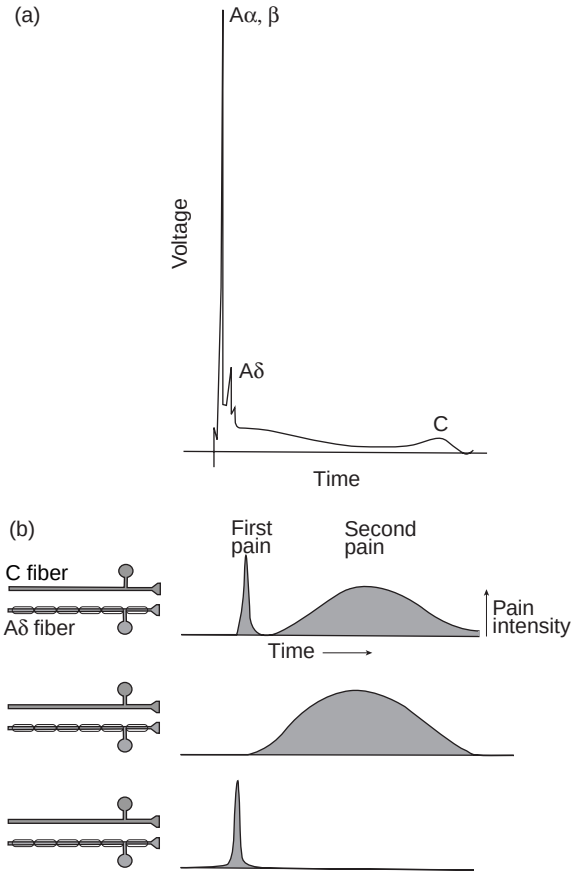


Figure 1. Propagation of action potentials in sensory fibers results in the perception of pain. (Modified from Ref. 4). (a) This electricle recording from a whole nerve shows a compound action potential representing the summated action potentials of all the component axons in the nerve. Even though the nerve contains mostly nonmyelinated axons, the major voltage deflections are produced by the relatively small number of myelinated axons. This is because action potentials in the population of more slowly conducting axons are dispersed in time, and the extracellular current generated by an action potential in a nonmyelinated axon is smaller than the current generated in myelinated axons. (b) First and second pain are carried by two different primary afferent axons. First pain is abolished by selective blockade of Aδ myelinated axons (middle) and second pain by blocking C fibers (bottom).

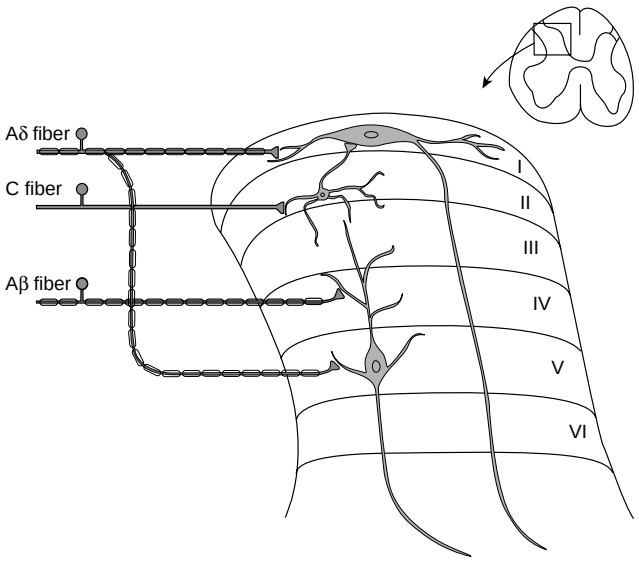


Figure 2. Nociceptive afferent fibers terminate on projection neurons in the dorsal horn of the spinal cord. Projection neurons in lamina I receive direct input from myelinated (Aδ) nociceptive afferent fibers and indirect input from unmyelinated (C) nociceptive afferent fibers via stalk cell interneurons in lamina II. Lamina V neurons are predominately of the wide dynamic-range type. They receive low threshold input from the large-diameter myelinated fibers (Aβ) of mechanoreceptors, as well as both direct and indirect input from nociceptive afferent fibers (Aδ and C). In this figure the lamina V neuron sends a dendrite up through lamina IV, where it is contacted by the terminal of an Aβ primary afferent. A dendrite in lamina III arising from a cell in lamina V is contacted by the axon terminal of a lamina II interneuron. (Adapted from Ref. 4.)

In the dorsal horn of the spinal cord at the synapse level, the afferent pain signal can be modulated to either lessen or amplify the body's response to the pain signal. Serotonin as well as norepinephrine act either directly presynapticly to inhibit the propagation of the pain signal or via activating inhibitory interneurons. The enkephalins, endogenous δ and μ opiate receptor agonists, function at this level to serve a similar inhibitory function. The neuromodulator peptide, substance P is released, along with glutamate from the C fibers, and both work alostericly to amplify the pain signal transmission to higher levels.

Once in the dorsal horn of the spinal cord and after synapse, the afferent pain signal is transmitted to higher cortical centers via either the spinothalamic, spinoreticular, spinomesencephalic, cervicothalamic, or spinohypothalamic pathways. Perception is the final process where all above processes as well as prior physical and psychological experiences interact and create the final subjective and emotional experience of pain. The opioids, both endo-

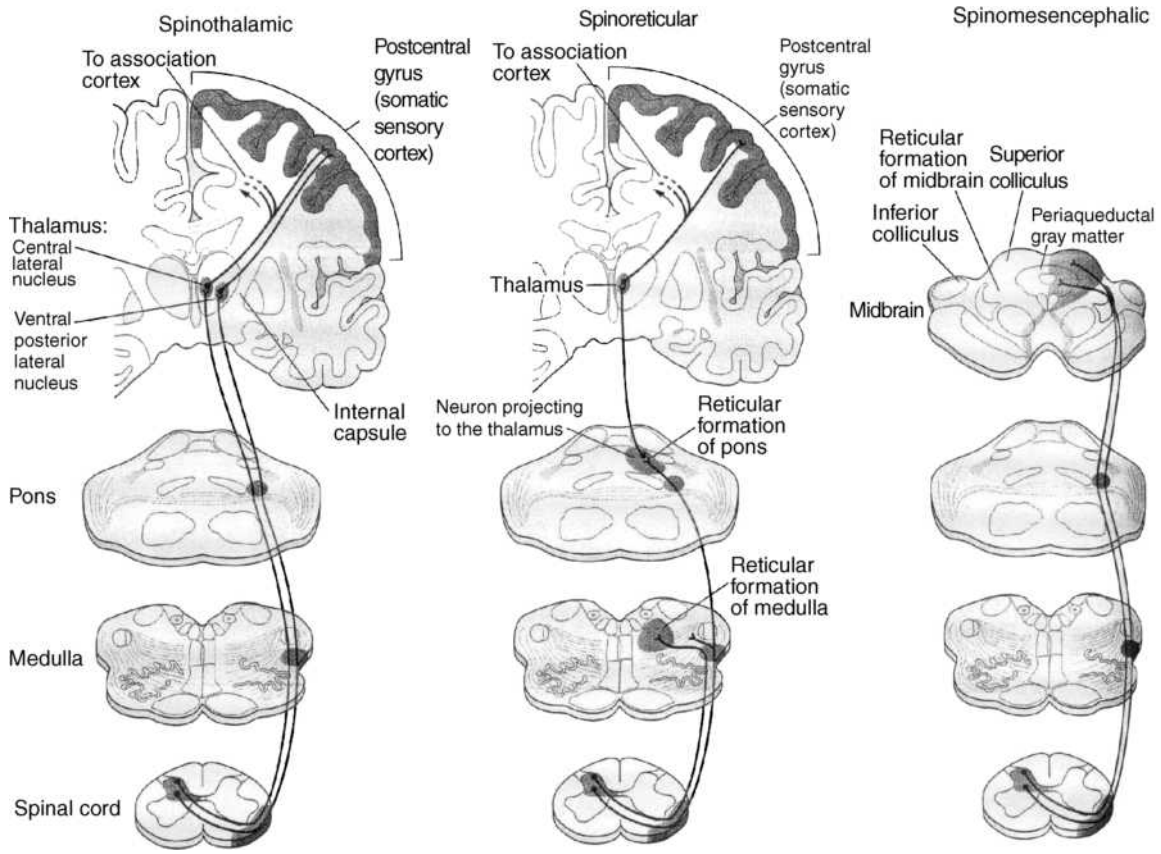


Figure 3. Three of the major ascending pathways that transmit nociceptive information from the spinal cord to higher centers. The spinothalamic tract is the most prominent ascending nociceptive pathway in the spinal cord. (Adapted from Ref. 7.)

genous and exogenous, function to alter perception of pain at the cortical level, as well as to activate inhibitory interneurons in the periaqueductal gray area (Fig. 3).

HYPERALGESIA AND SENSITIZATION

In certain situations, nociceptors can become exquisitely sensitive to stimulation or activated in greater numbers than usual. This results in hyperalgesia and is termed sensitization; this process occurs via distinct mechanisms both peripherally as well as centrally. While peripheral sensitization occurs in both acute and chronic phases of injury, central sensitization generally occurs in the chronic phase of insult, after repetitive noxious events.

Upon peripheral injury, for example, an epithelial incision, inflammation is affected via a large number of chemical mediators, such as prostaglandins, leukotrienes, bradykinin, serotonin, substance P, histamine, potassium, and others, released from both damaged, as well as surrounding tissues (5). These inflammatory mediators serve not only to result in inflammation, but also serve to decrease the threshold to stimulate surrounding nociceptors. This can be done by directly acting to affect sensitization or by working in tandem to sensitize nociceptors via another chemical mediator. For example, bradykinin is an important and extremely potent mediator of hyperalgesia. It works not only to directly sensitize the nociceptive fibers (i.e., C and Aδ fibers), but also serves to stimulate local tissue to produce prostaglandins, which themselves result in sensitization. In addition to bringing about sensitization of nociceptors, some chemical mediators directly activate nociceptors, for example, histamine activating polymodal nociceptors (Table 1).

With continued C fiber pain signal transmission due to persistent noxious insult, increased glutamate is released from their end plates in the dorsal horn. With this increased glutamate release, continued opening of postsynaptic calcium ion channels results. This is mediated by postsynaptic *N*-methyl-d-aspartate (NMDA)-type glutamate receptors. This process, termed "wind-up", results in a continual increase in dorsal horn neuron response to the pain signal. This is an example of pain signal modulation. In addition to this progressively increasing response to the pain signal, dorsal horn neurons can become more easily excitable to a lesser peripheral signal. This process, termed central sensitization, is also mediated by NMDA-type glutamate receptors. Additionally, there is an upregulation in production of a variety of neurotransmitters, neurohormones, and their receptors. Effectively, these changes of excitability and magnitude of C fiber response constitute a pain "memory" and also result in progressively larger areas of peripheral tissue coverage of the dorsal horn neuron. Central sensitization with resultant

Table 1. Naturally Occurring Agents that Activate or Sensitize Nociceptors[a]

Substance	Source	Enzyme Involved in Synthesis	Effect on Primary Afferent
Potassium	Damaged cells		Activation
Serotonin	Platelets	Tryptophan hydroxylase	Activation
Bradykinin	Plasma kininogen	Kallikrein	Activation
Histamine	Mast cells		Activation
Prostaglandins	Arachidonic acid–damaged cells	Cyclooxygenase	Sensitization
Leukotrienes	Arachidonic acid–damaged cells	5-Lipoxygenase	Sensitization
Substance P	Primary afferents		Sensitization

[a]Modified from Ref. 4.

hyperexcitability helps explain allodynia, the perception of a non-noxious stimulus, such as light touch, as a painful stimulus. In light of these changes, it is obvious not only why chronic pain can be so difficult to treat, but also why it is important to break pain "cycles" before chronic changes begin to occur.

Allodynia is classically seen in several different chronic neuropathic pain syndromes for reasons that are not always completely understood, but likely stem from the chronic changes outlined above. Herpes zoster is perhaps the best known of these conditions with many sufferers reporting a severe dermatomal burning pain long after the peripheral nerve damage has healed. Allodynia is common following an attack, and severe bouts of pain can be precipitated from the friction between ones shirt and ones skin. Trigeminal neuralgia is another such chronic condition where allodynia is common. In this condition, lightly stroking one's cheek or the process of eating can precipitate attacks of severe, stabbing transient pain, followed by longer periods of a moderate to severe burning sensation.

PSYCHOLOGICAL ASPECTS OF PAIN

As mentioned earlier, perception of pain is an individualized phenomenon. It is affected by culture, mood, and individual experiences (8,9). Chronic pain can be termed as pain that persists for a certain period, usually ~6 months, after an injury has healed or when the noxious source is idiopathic and central sensitization has occurred. The field of pain management employs a diverse, polymodal disciplinary strategy toward treating pain that extends far beyond simple pharmacotherapy. It includes interventional therapy, physical therapy, meditation, biofeedback, acupuncture, psychiatric therapy, electroanalgesia, and many other treatment modalities. There is a definite psychological component to chronic pain that can cause it to be perceived as much more severe than acute pain and make it refractory to traditional therapy, and chronic pain is frequently associated with depression.

THEORIES OF PAIN

Gate Control Theory

The gate control theory, published in 1965 by Ron Melzack and Pat Wall (10), was the theory from which modern electrotherapy has evolved and that has helped revolutionize treatment and our understanding of pain. The theory states that pain perception depends on the balance of large, afferent sensory Aβ and small diameter afferent nociceptive Aδ and C fiber activity. According to the theory, nonnociceptive sensory fibers can activate neurons in the substantia gelatinosa. These neurons can decrease or inhibit the pain signals of nociceptive neurons prior to higher level transmission. This theory explains the common practice of rubbing an acute wound to lessen pain. It is worth noting that the inhibitory effect of nonnocioceptive neurons is a local one. No analgesic effect exists when rubbing one's toes after an injury to one's fingers (Fig. 4).

The theory's emphasis on the modulation of inputs in the spinal dorsal horn and the dynamic role of the brain in pain processes had a clinical, as well as a scientific impact, and after this theory several methods were developed to modulate the input. Physical therapists and other healthcare professionals began developing and refining different modulation techniques, such as implantable dorsal spinal electrostimulation, and later transcutaneous electrical nerve stimulation devices as well as rediscovering old ones, such as acupuncture. After this discovery, TENS became an important part in treating the acute and chronic pain. See below for a much more thorough discussion of the history of electroanalgesia.

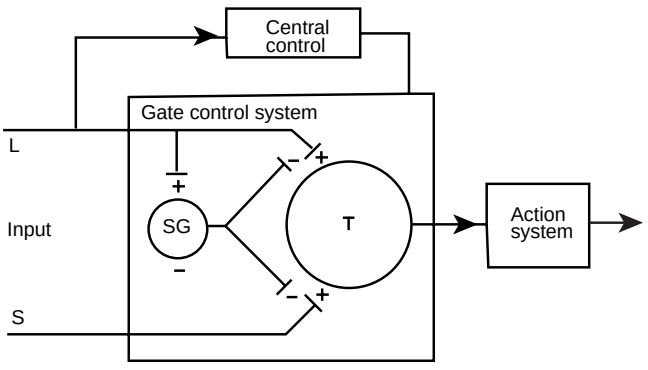


Figure 4. Schematic diagram of the Melzack–Wall gate control theory of pain mechanisms. Large- (L) and small-diameter (S) afferent fibers project to the substantia gelatinosa (SG) and first central trasmission (T) cells. The inhibitory effect (–) of SG on the afferent terminals is increased (+) by activity in L fibers and decreased by activity in S fibers. A specialized system of L fibers (the central control trigger) activates certain cognitive processes that influence the modulating properties of the apinal gating mechanism via descending fibers. (From Ref. 10 R. Melzack and P. D. Wall, *Science*, 150:971–979, 1965, © 1965, AAAS.)

Other Theories Regarding TENS' Analgesic Effect

Other theories have been developed to explain the effectiveness of TENS, namely, the enkephalin and endorphin theories, and likely all three contribute to the analgesic effect. Multiple studies have demonstrated an increase in dorsal horn enkephalin production (11,12) as well as have demonstrated that blockade of opiate receptors lessens or even prevents analgesia from TENS (13–15). As briefly described earlier, enkephalins are μ and δ opiate receptor agonists and function as inhibitory neurotransmitters. Release of enkephalins from inhibitory interneurons decrease Ca^{2+} influx into the nocioceptive neuron, thereby preventing, or lessening depolarization time, prevents or lessens excitatory neurotransmitters, such as glutamate, release, thereby negatively modulating the pain signal. Additionally, enkephalins function postsynapticly to activate K^{+} conductance, thereby hyperpolarizing the dorsal horn projecting neuron and inhibiting pain transmission to further higher cortical centers (Fig. 5).

While enkephalins have a short half-life and function locally, recent studies (16–18) have demonstrated increased levels of circulating endorphins. In contrast to enkephalins, endorphins are circulating μ agonist neurohormones. As such, they act not only on the μ receptors in the dorsal horn of the spinal cord, but also function on central μ receptors to alter the perception of pain and negatively modulate the signal. The discovery that TENS increases endorphin levels is significant. The effect of increasing enkephalins produces a transient effect that lasts only as long as the electrical signal is applied, as is with direct nonnocioceptive stimulation as described in the gate theory. However, use of TENS produces an increase in circulating endorphins that is proportional to usage. The net effect is an analgesic effect that persists after the TENS unit is removed and increases in potency and duration with repeated usage.

The Evolution of Electroanalgesia

The use of electroanalgesia is an ancient practice, and to thoroughly understand the theory and application of TENS, it is important to understand the evolution of electroanalgesia. The powers of certain fish, namely, the Nile Catfish *(Malopterurus electricus)*, Torpedo Fish *(Torpedo mamorata)*, and the Electric Eel *(Gymnotus electricus)* to deliver electrical shocks resulting in paralysis and temporary sensory loss in affected limbs has long been known. The Nile Catfish appeared on walls of various Egyptian tombs, dating from ~2750 bc, and represents the earliest known documentation of this phenomenon of electrical discharge. While it is not known exactly when ancient man discovered the analgesic or anesthetic properties of such fish, it is quite likely that since their discovery by primitive man, these properties were readily apparent.

Exactly when the electrical properties were used for medicinal benefit is unclear, but the earliest known writings describing this benefit were made by made in ad 46 by Scribonus Largus, a Roman physician who described the usage of the torpedo fish as a treatment for intractable gout and headache pain (19). Quoting from his treatise *Compositiones Medicae*, Largus describes these remedies.

> *For any type of gout a live black torpedo should, when the pain begins, be placed under the feet. The patient must stand on a moist shore washed by the sea and he should stay like this until his whole foot and leg up to the knee is numb. This takes away present pain and prevents pain from coming on if it has not already arisen. In this way Anteros, a freedman of Tiberius, was cured* (20).
>
> "*Headache even if it is chronic and unbearable is taken away and remedied forever by a live torpedo placed on the spot which is in pain, until the pain ceases. As soon as the numbness has been felt the remedy should be removed lest the ability to feel be taken from the part. Moreover several torpedoes of the same kind should be prepared because the cure, that is, the torpor which is a sign of betterment, is sometimes effective only after two or three*" (21).

As time progressed, the usage of electroanalgesia spread as a treatment for varying medical conditions. Pedanius Discorides around 80 ad describes the usage of the torpedo fish for rectal prolapse. This represents likely the first description of electrical stimulation for intentional muscular contraction (19). Likewise, these treatments were used and espoused by Galen in the second century.

The knowledge of the usage of the electrical properties of such fish was not limited to Europe. Ibn-Sidah, an Islamic physician described placing an electric catfish on someone suffering a seizure sometime in the eleventh century (21). The use of the electric fish continued to grow and by the sixteenth century the number of remedies had expanded and included treatments for various arthralgias, myalgias, headaches, epilepsy, vertigo, and for inducing sleep both by European, Indian, and Middle Eastern physicians. By the seventeenth century the application of artificially generated electricity was made possible by the development of the electrostatic generator by Otto Von Guericke.

Major revolutions in electroanalgesia came in the mid-nineteenth century from Guillaume Benjamin Amand Duchenne. He introduced the usage of moistened electrodes, as opposed to the more painful dry electrode, as a means of delivering electroanalgesia for treatment of neuropathic pain. His focus on muscle contractions from electrotherapy and making strides toward to a somewhat standardized system of electrode placement were important advances as well.

Throughout the world, electrotherapy became increasingly popular toward the end of the nineteenth and beginning of the twentieth centuries. However, with this rise in popularity came a rise in dubious to downright fraudulent applications and practitioners treating all manners of maladies from skin ailments to weight loss. With an ever increasingly savvy public, the rise of fraudulent applications, the rise of modern pharmacotherapy, X rays, and the like, electrotherapy begin to fall out of favor, or at least popularity, in the early twentieth century (19).

However, technological advances in electrical storage and delivery along with new understandings of pain have produced a resurgence in application and research in electroanalgesia. Shortly after the publishing of the gate control theory, a flurry of exciting discoveries in this

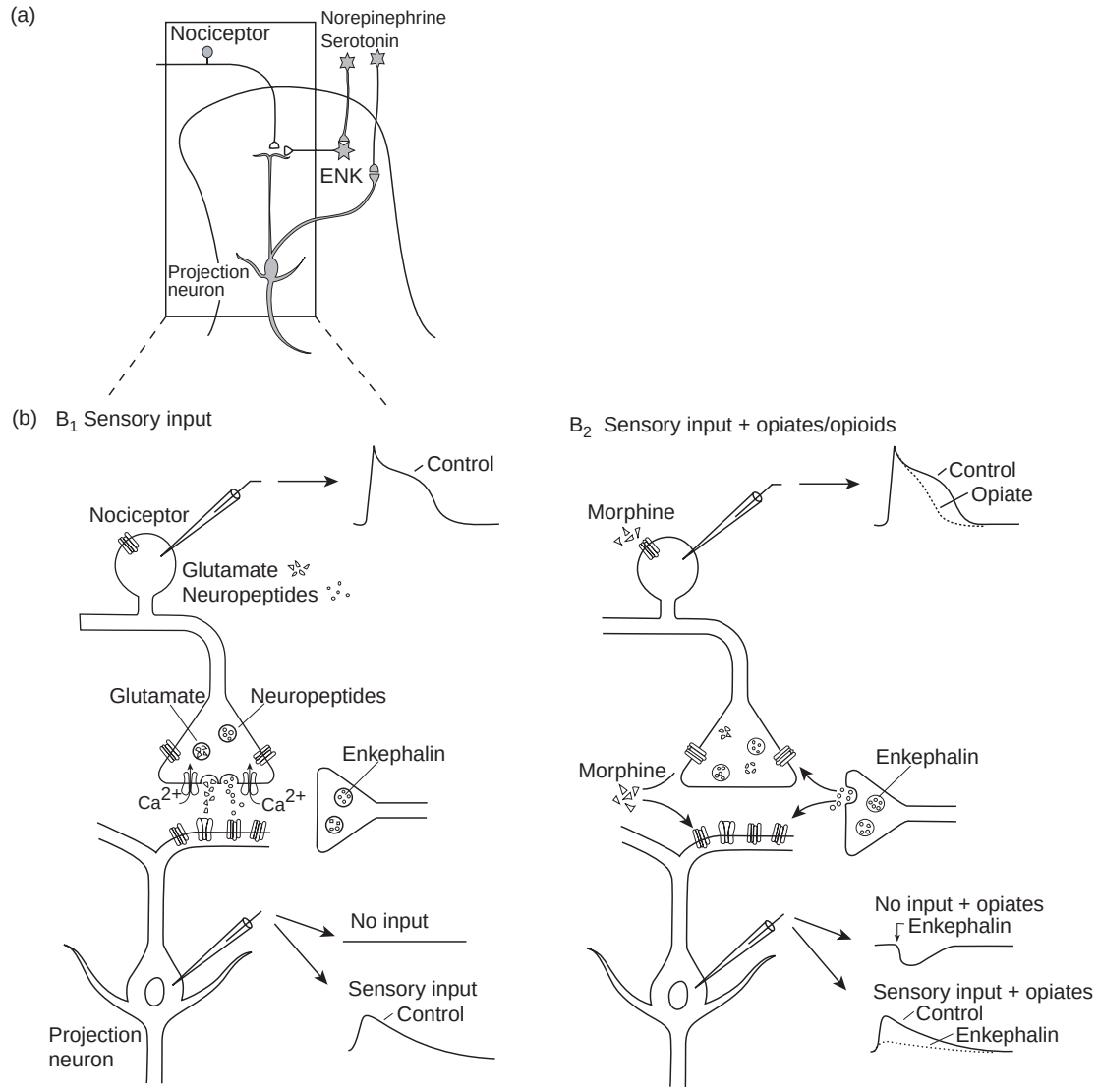


Figure 5. Local-circuit interneurons in the superficial dorsal horn of the spinal cord integrate descending and afferent pathways. (a) Possible interactions between nociceptor afferent fibers, local interneurons and descending fibers in the dorsal horn of the spinal cord. Nocicaptive fibers terminate on second-order spinothalamic projection neurons. Local enkaphalin-containing interneurons (ENK) exert both presynaptic and postsynaptic inhibitory actions at these synapses. Serotonergic and noradrenergic neurons in the brain stem activate the local opioid interneurons and also suppress the activity of spinothaiamic projection neurons. (b) 1. Activation of nociceptors leads to the release of glutamate and neuropeptides from sensory terminals in the superficial dorsal horn, thus depolarizing and activating projection neurons. 2. Opiates decrease the duration of the nociceptor's action potential. probably by decreased Ca^{2+} influx and thus decrease the release of transmitter from primary afferent terminals. In addition, opiates hyperpolarize the mambrane of the dorsal horn neurons by activating a K^{+} conductance. Stimulation of the nociceptor normally produces a fast excitatory postsynaptic potential in the dorsal horn neuron opiates decrease the amplitude of the postsynaptic potential.

realm took place starting with the 1967 demonstration by Sweet and Wall that *In vivo* electrostimulation of peripheral nerves produces analgesia as well as Shealy and Long's work in the area of dorsal and anterior spinal cord surgically implanted stimulators. Shealy and Long discovered that peripheral nerve stimulation done in surgical candidates prior to an electrospinal implant placement produced nearly comparable analgesia to the actual dorsal horn implant (17)! This discovery laid the foundation for TENS development and widespread utilization today (22,23).

While somewhat beyond the scope of this article, it is worth noting that electroaccupunture experienced a widespread rediscovery in China in the 1950s. Though based on a different system of understanding of human physiology than traditional western medicine, this modality of

Figure 6. This represents one of the earliest families of stimulators, with the original model seen on the *left*. The first personal patient treatment model is depicted on the *right*, and a prototype for a miniaturized design is shown in the *center*. The original sponge electrodes are depicted in the *foreground*.

electroanalgesia is beginning to garner increasing interest interest in western medicine (19).

Transcutaneous Electrical Nerve Stimulation

Today electrotherapeutic treatment is one of the most important parts of multidisciplinary approach to treat acute and chronic pain. The TENS units themselves have undergone an evolution from large bulky units to the much smaller units of today. While there are numerous units available, each generally consists of one or more channels for electrodes, a display (either analog or digital), and various options to adjust the various parameters of the electrical current.

One of the First TENS Units Created (below)

The unit (*left*) is one of the first TENS unit available and is large and bulky with an all analog interface. Subsequent units (*center and right*) still remain analog but were more compact, though nowhere near the level of today's units (Fig. 6).

Several Modern TENS Units (below)

Pictured are just several of the numerous commercially available TENS units. Note the compact size of the models compared to older units as well as the digital TENS unit (*bottom*) (Fig. 7 and Table 2).

Electrode

The discovery that transcutaneous peripheral nerve stimulation provided nearly identical analgesic levels as dorsal root stimulation revolutionized electroanalgesia, and it almost goes without saying that the noninvasive, easy to employ nature of TENS is one of the modality's biggest assets. The electrode, the interfacing agent between the skin and machine, has undergone almost as much evolution as the TENS unit itself. The very nature of peripheral transcutaneous nerve stimulation is such that electrical currents must be applied for longer periods of

Figure 7. Several commercially available TENS units.

time in greater amounts. While the process of transferring an electrical current from machine to peripheral nerve may seem relatively simple on the outside, several notable problems present ranging from the actual transfer of the current to skin irritation to cost. Several distinct solutions currently are in use, and all present with a variety of tradeoffs (Table 3). Generally speaking, properties of a good electrode for TENS use include low cost/use ratio, good adhesion, comfortable, nonirritating to skin, good electrical conductivity, and easy to use.

Standard EKG or EEG electrodes were initially used for TENS with limited success, as these were designed for much lower amperage and much shorter usage than is needed for effective TENS. It quickly became apparent based on excessive skin irritation and poor adhesion and subsequent nonuniform current distribution that new electrode solutions were needed. One of the most popular current solutions involves silicone impregnated with carbon (Table 3). These carbon silicone electrodes provide the best cost/use ratio of the commercially available electrodes and can often last for months if properly cared for. However, a tradeoff exists in terms of convenience with these carbon silicone pads, as electroconductive gels, rich in suspended ions to facilitate the transfer of electric current from the TENS unit across the epidermis, must be applied

Table 2. Comparisons of Selected Modern TENS UNITS

Manufacturer/unit	Size, cm	Weight, g	Power source	Digital/ Analog	Pulse Width, μs	Stimulus Modes	Burst	Channels	Waveform	Output Current, mA	Output Rate, Hz
Rehabilicare/ ProMax	7.1 × 10.1 × 3.4	122.2	3 AAA Batteries	Digital	50–400	SD, Modulation, Normal, Burst	8 pulses/burst;	2	Asymmetric rectangular biphasic with zero net dc	0–100	2–160
Rehabilicare/ Maxima3	8.4 × 2.5 × 6.3	121	9 V Battery	Analog	50–400	SD, Normal, Burst	8 pulses/burst; Burst at 2 Hz	2	Biphasic, asymmetrical with zero net dc	0–100	2–160
Rehabilicare/ SMP-Plus	9.5 × 6.4 × 2.5	136.2	9 V Battery	Digital	40–300	SMP, Constant, Burst, Modulated Rate, Modulated Width, Multi-modulated	8 pulses/burst; Burst at 2 Hz	2	Biphasic, asymmetrical with zero net dc	0–60	2–125
Empi/Epix VT				Digital	0–400	ELF, Dual Pulse, High Frequency, Ramped Burst, Alternating Ramped Burst, Modulated Amp., Random Modulated, Cycle Burst, Rate Modulation, Multi modulation.	Varies	2	Balanced asymmetrical biphasic	0–60	2–150
Empi/Epix XL				Analog	0–400	Continuous, Burst, Modulated Rate, Multi-modulated	Varies	2	Symmetrical biphasic square zero net charge,	0–60	2–150
BioMedical Life Systems/ BioMed 2000	9.9 × 6.98 × 2.54	132	9 V Battery	Analog	50–250	Constant, Burst, Width modulation,	8 pulses/burst; Burst at 2 Hz	2	Asymmetric rectangular biphasic	0–80	2–150
BioMedical Life Systems/ BioStim A6	9.9 × 6.98 × 2.54	132	9 V Battery	Analog	10–250	Constant, Burst, Width modulation, Rate Modulation, rate/width modulation, cycled burst	Cycled	2	Asymmetric rectangular biphasic	0–100	2–200
BioMedical Life Systems/ BioStim LX	9.5 × 6.3 × 3.2	226	4 AA Batteries	Digital	10–250	Constant, Burst, Burst Programmable, Width modulation, Cycled Burst	Cycled	2	Asymmetric rectangular biphasic	1–98	1–150
BioMedical Life Systems/ BioStim M7	8.2 × 7.0 × 4.5	266	4 AA Batteries	Digital	10–250	Constant, Burst, Burst Programmable, Width modulation, Rate Modulation, rate/ width modulation, cycled burst	Cycled	2	Asymmetric rectangular biphasic	0–98	1–200

Table 2. (*Continued*)

Manufacturer/unit	Size, cm	Weight, g	Power source	Digital/ Analog	Pulse Width, μs	Stimulus Modes	Burst	Channels	Waveform	Output Current, mA	Output Rate, Hz
BioMedical Life Systems/ BioStim Plus	9.9 × 6.98 × 2.54	132	2 AA Batteries	Digital	10–250	Constant, burst, cycled burst, Pulse rate modulation, pulse width modulation	Cycled/ 2 presets	2	Asymmetric biphasic, square wave	0–98	1–150
Vital/TENS Deluxe	9.1 × 6.4 × 2.3	130	9 V Battery	Analog	50–250	Burst, modulation, constant	7 pulses/burst; Burst at 2 Hz	2		variable	2–120
Kingly Star/ KS-168	40.6 × 18.5 × 6.8	1060.5	UM-1 × 6 Battery or 9 V dc adapter	Analog		Constant, Burst, Modulation		4		0–100	2–200
Pain Management Technologies/ Medscope Pro	10.1 × 5.9 × 2.37	140	9 V Battery	Digital	50–260	Constant, Burst, pulse rate modulation, pulse width modulation	7 pulses/burst; Burst at 2 Hz	2	Biphasic square wave with zero net dc	0–80	2–150
Pain Management Technologies/ Bioscope	9.85 × 6.05 × 2.45	134	9 V Battery	Analog	60–250	Constant, Burst, Pulse rate and Pulse width modulation	~7 pulses/burst; Burst at 2 Hz	2	Asymmetric biphasic square	0–15	2–150
ProMed Specialties/ ProM 100	9.1 × 6.4 × 2.4	130	9 V Battery	Analog	40–250	Constant		2	Asymmetric biphasic square with zero net current dc	0–80	2–150
ProMed Specialties/ ProM 200	9.1 × 6.4 × 2.3	130	9 V Battery	Analog	40–260	Burst, Constant, Modulation	9 pulses/burst; Burst at 2 Hz	2	Asymmetric biphasic square with zero net current dc	0–80	2–120
ProMed Specialties/ ProM 300	9.1 × 6.4 × 2.3	130	9 V Battery	Analog	40–260	Burst, Constant, Modulation	9 pulses/burst; Burst at 2 Hz	2	Asymmetric biphasic square with zero net current dc	0–80	2–50
Shining World Health Care Co./ SW 325	12.5 × 6.6 × 2.8	138	9 V Battery	Digital	200	Constant	n/a	2	Biphasic Spiked Wave	0–50	
Body Clock Healthcare/ Profile	10.5 × 6.5 × 2.75	100	2 AA Batteries	Digital	25–250	Constant, Burst, Width modulation, Rate modulation	Variable	2	Symmetric biphasic rectangular	0–100	1–200
Med-Dyne/TA3	9.1 × 6.4 × 2.3	130	9 V Battery	Analog	40–260	Constant, Burst, Modulation	9 pulses/burst; burst at 2 Hz	2	Asymmetric biphasic square with zero net current dc	0–80	2–160

Table 3. Basic Types of Electrodes

General electrode Type	Number of Uses	Typical Retail Cost Electrode (pair)	Typical features			
			Adhesion and Conduction	Composition Materials	Advantages	Disadvantage
Disposable	1 use	Under $3.00	Pressure-sensitive tape surrounding conductive area (wet get in sponge)	Nonwoven Foam	Easiest to use Very good adhesion Comfort	Cost/use Skin irritation Poor electrical performance
Semireusable	3–10 uses	$5.00–10.00	Conductive adhesive over entire surface	Foam Plastic film Carbon silicone	Ease of use Low skin irritation Comfort	Weak adhesion Care and storage Medium cost per use Good electrical performance
Reusable	>100 uses	$4.00[a]	Requires addition of conductive gel and tape adhesion	Carbon silicone	Very good electrical performance (if applied properly) Lowest cost per use	Most preparation to use Skin irritation Messy Requires gel and tape Skill required for optimal performance Poor adhesion Not as flexible in use

prior to each usage. The electrodes must then be affixed with tape to the skin. This process can be laborious if not impossible for the end user to do, depending on electrode location as well as physical disability. While numerous medical tapes exist, care must be taken in their selection. Some users display mild-to-severe allergic reactions to adhesives in various tape products. Likewise, certain adhesive tapes adhere too firmly to the skin and can result in injury with repeated application and removal.

In applications where cost is no object, sterility is needed, or convenience must be maximized, single use adhesive electrodes are used that consist of thin, porous material impregnated with adhesive electroconductive gel covered with cloth or plastic on one side to prevent adhesion to clothing or electrical current escape. These electrodes are used extensively in hospital or rehab facilities where numerous patients are seen and reusing electrodes is neither feasible nor sanitary as well as in individuals who desire or require maximum convenience.

A third option in electrode selection includes so-called "dry gel" electrodes manufactured from a conductive polysaccharide gum, called Karaya or Sterculia gum when taken from the Sterculia Urens tree native to India or a manufactured comparable material. These dry gel electrodes represent a good compromise between the disposable and reusable carbon silicone electrodes, as they are self adhesive, do not require electroconductive gel application, may be reused several times, and represent a significantly lower cost/use ratio than disposable pads.

Patients should be informed of the various pros and cons of the various electrodes as well as counseled regarding proper usage. According to Szeto, several factors should be considered when selecting the proper electrode:

1. How long will each application of the electrodes last, and therefore what is the level of adhesion needed?
2. Is the stimulation site readily accessible or will there be someone to assist? How simple must the application of the electrodes be?
3. What is the patient's skin type (hairiness, oiliness, hyperallergic)? Will special pregelled electrodes be required?
4. Where is the painful are? This factor will help to determine the best electrode size, shape, cosmesis, and number.
5. Does the TENS user lead an active life? If so, a high performance electrode in terms of adhesiveness, pliability, and nonirritability would be needed.
6. Can the user take good care of the electrodes, and what are the financial arrangements? These issues will affect the cost-effectiveness of disposable or semireusable type of electrode (24).

Electrode size is another factor to consider in selection and depends on the location of pain, area required for stimulation, and personal preference. Numerous sizes and shapes of electrode pads are available and suitable for virtually any application.

Sample Electrodes

Electrode Placement

Electrode placement is crucial in maximizing positive outcome with TENS units. Most units employ two or more channels of current, which splits to two electrodes, and it is often advisable for multiple channels to be used to cover maximal areas, as many pain syndromes often do not exhibit pain localized to a specific point source. Numerous books on TENS or manufacturer information as well as various anatomical charts provide users with possible electrode placements. While it is impossible to describe

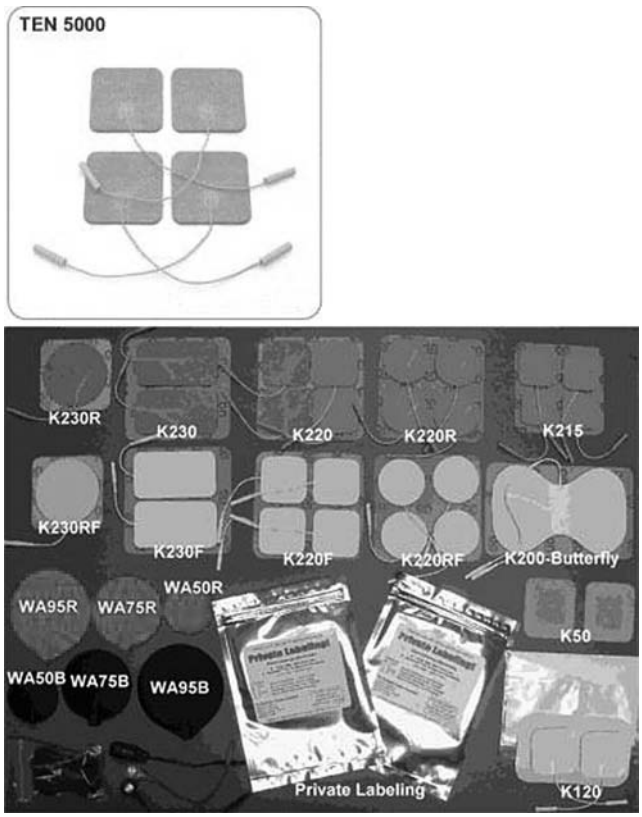


Figure 8. TENS electrodes.

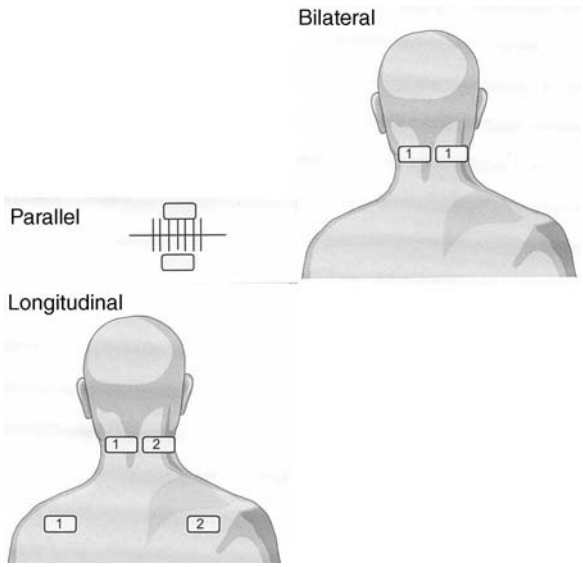


Figure 9. Bilateral pain electrode placement.

proper electrode placement for every pain syndrome, certain electrode arrangements are frequently used.

For the purposes of this discussion, the channels will be referred to as I and II and the negative electrode as "a" and the positive as "b". Parallel placement with one channel is utilized for relatively localized areas of pain, such as point pain or pain from a surgical incision. Electrode Ia is placed on one side of the incision, while Ib is on the other, producing a current that flows between the two with the area of pain in between the electrodes. Bilateral placement is similar, but generally defined as meaning Ia and Ib electrodes are placed on either side of the spine, symmetrically and close together, useful for localized, nonradiating neck or back pain. For radiating neck or back pain, a longitudinal arrangement is often utilized in which electrodes of one channel are on the same side of the spine and placed along the pain pathway. If the pain is bilateral, electrodes of another channel on the can be placed on the opposite side of the spine along the pain pathway (Fig. 9).

A crossed, or interferential, placement is useful for pain localized to large joints, such as shoulder, elbow, or knee. In this pattern, Ia and IIa are placed side by side with IIb below Ia and Ib below IIa, creating a square pattern with electrodes of the same channels diagonally opposite each other with the area of pain in the center of the square. Bracketed placement is generally reserved for treating the dermatomal neuralgia that frequently is associated with shingles, varicella zoster, out breaks. In this arrangement, electrodes Ia and Ib are placed along the dermatome cranial to the neuralgia and electrodes IIa and IIb placed along the dermatome caudal to the neuralgia (Fig. 10).

Occasionally, localized pain is so severe that the user cannot tolerate electrode placement over the affected site, and in this case a contralateral placement in the nonaffected hemisphere over the same anatomical area as on the affected side is utilized. This arrangement will sometimes permit sufficient pain relief for the user to eventually tolerate direct stimulation. While the exact mechanism of analgesia is not known, it is hypothesized the analgesic effect is the result of central inhibitory pathways (24). Certain syndromes, such as Reflex Sympathetic Dystrophy, lend themselves to this placement, and reflex vasodilatory effects may explain contralateral analgesia in these syndromes (25) (Fig. 11).

Certain pain syndromes, such as phantom limb pain, glove–stocking distribution peripheral neuropathy, or acute burns, fractures, lacerations, or other injuries of the hands or feet lend themselves to a placement of the electrodes proximal to the actual source of the pain. In this placement, the electrodes of one channel are simply placed along the dermatome of the pain source, but proximal to the pain (Fig. 12).

The final placement method to be covered is a linear, unilateral, overlapping pattern useful for pain along some, as in myofacial pain, or all, as in radicular pain, of an extremity, and follows a placement procedure outlined by Wolfe (25). After the dermatomal distribution of the pain is elucidated, electrode Ia is placed at the most proximal location where the user experiences pain. Distal to this electrode, the user identifies the site of maximal pain and places IIa here. At the most distal site of pain, electrode IIb is placed, and between IIa and IIb, electrode Ib is placed, being careful to keep all electrodes in the affected dermatome (Fig. 13).

It is important to note with the above placement, the electrical current covers the entire length of the pain the

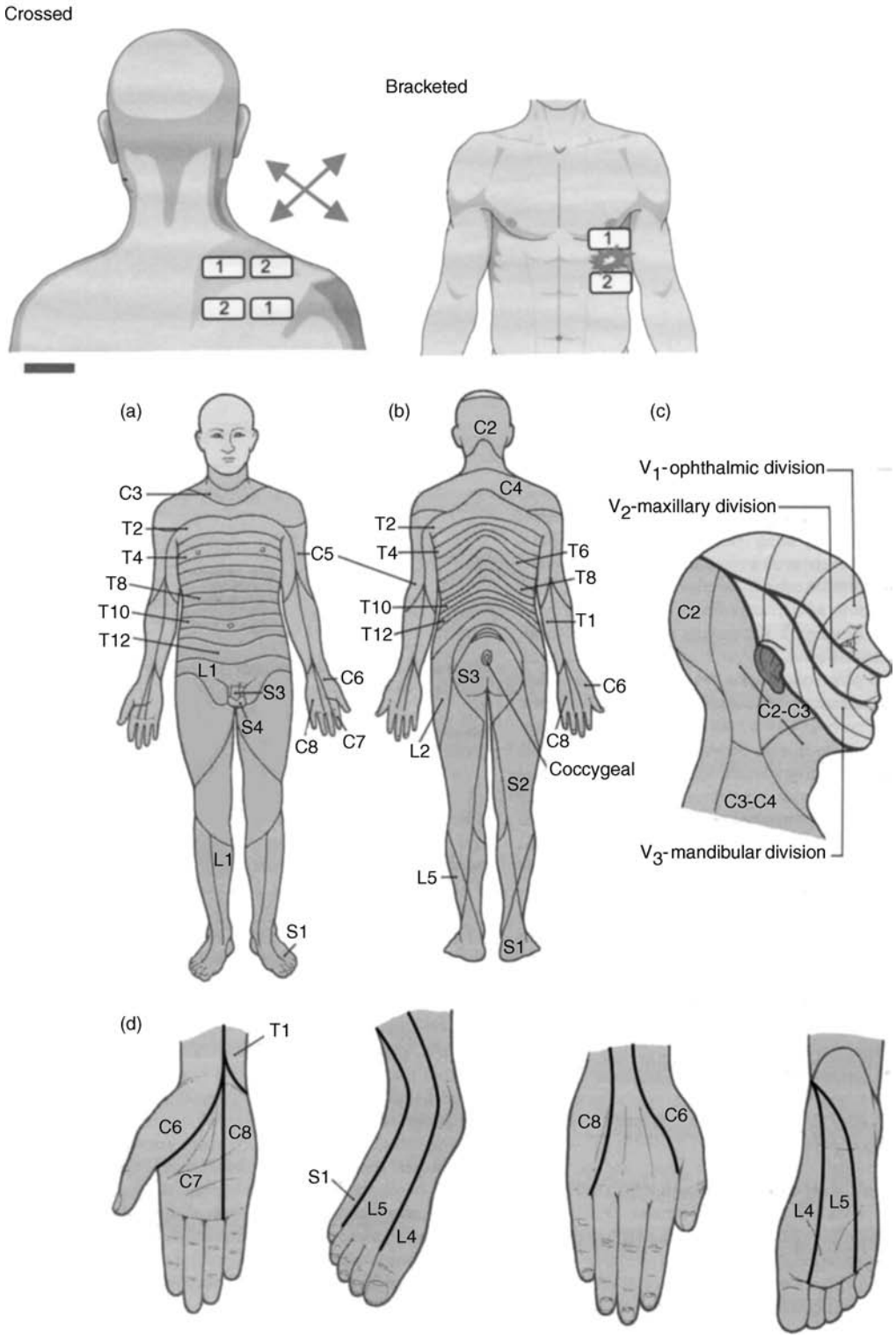


Figure 10. Dermatomal maps of the peripheral distribution of spinal nerves (a and b) and trigeminal nerve (c) d. Details of termatomal maps on anterior and posterior surfaces of the hand and foot.

user experiences. If the user inadvertently places electrodes in a nonoverlapping pattern (i.e., Ia and Ib both proximal to IIa and IIb), the current will not cover the entire pain pathway; instead it will only travel between electrodes of the same channel, leaving the area between Ib and IIa "uncovered". The following should be generally avoided due to poor current coverage area: unilateral, linear (Fig. 14).

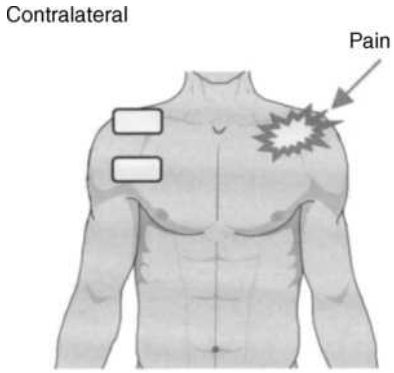


Figure 11. Contralateral electrode placement.

Electric Amplitude–Frequency Selection

Once the proper electrode type is chosen, as well as optimal electrode placement ascertained, the optimal electrical signal must be selected. Generally speaking, the most used currents include "classical" TENS with high intensity–low frequency currents for up to 12 h at a time, low intensity–high frequency currents for up to 45 min several times a day, and intermittent low frequency bursts are used. The varying current intensity–frequency produces analgesia via the different mechanisms as discussed above.

Classical TENS employs high frequency (60–200 pulses per second)/low intensity (15–60 mA) stimulation and produces a distinct "electrical tingling" sensation in the area of electrode pad placement that most users find pleasant. This current is not of significant intensity to produce significant muscular contraction. Pain relief from this form of stimulation is transient, occurring quickly once stimulus is applied and disappearing once current is removed, and the gate control theory likely explains the mechanism of analgesia. The high frequency pulses activate *A*β sensory afferent fibers and inhibit pain transmission in the dorsal column of the spinal cord.

Low frequency (1–5 pulses per s)/high intensity stimulation, producing sustained muscle contractions, results in slower onset pain relief that persists after the stimulus is removed. Numerous studies have demonstrated partial to near total inhibition of analgesic effect via administration of naloxone (13–15). The endorphin and enkephalin theories described previously likely largely account for the mechanism of analgesic activity, especially the endorphin theory and long-term analgesia. While effective at inducing long-term analgesic effects, the low frequency–high intensity method of stimulation is often perceived by many patients as less pleasant than high frequency stimulation.

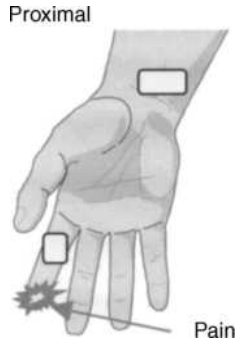


Figure 12. Proximal electrode placement.

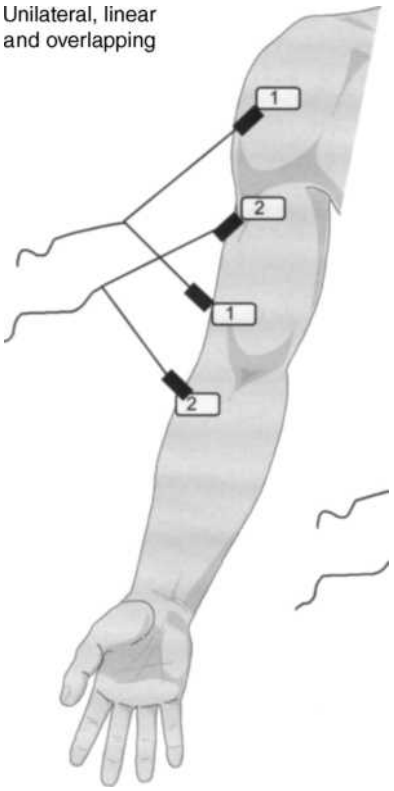


Figure 13. Unilateral, linear, and overlapping electrode placement.

As the long-term effects of low frequency intense stimulation are desirable, manufactures have devised means of producing a more pleasurable sensation while at the same time stimulating muscle contraction enough for long-term analgesia via modulation of the current. To understand the modulation of current in TENS, a brief review of the current waveforms it employs is needed. Briefly, biphasic waveforms, consisting of both a positive and negative phase are used, and these may be either symmetric or

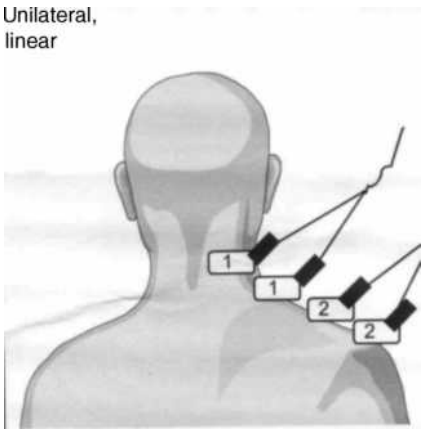


Figure 14. Unilateral, linear electrode placement.

asymmetric. If the current amplitude is equally positive and negative, the current is termed "electrically balanced," also referred to as zero net charge (znc) or no net dc current. While both balanced and unbalanced electrical currents are employed, unbalanced current transmission can result in pH changes in the skin with long-term usage, do to ion exchange, which can result in skin irritation. Additionally, the current employed in TENS is a pulsatile current with interspaced periods of electrical activity and electrical silence. The periods of electrical silence may be either uniform or varying. The frequency of electrical pulses may be given in units of hertz (Hz), cycles per second (cps), or pulses per second (pps) (25). It is important to note that while frequency may be a constant 100 cps, the period between the pulses may be variable.

All of these variables in the current waveform may be adjusted to achieve a net effect that is both pleasant to the patient while sufficient to achieve muscular contraction. For example, the amplitude of the current may be varied over a constant time interval, the duration of the pulse may be varied, the time between pulses may be varied, or a combination of some or all of the previous may be used. As pain severity and character can frequently change, models that allow modulation of electrical current via one or more characteristic offer distinct advantages in patients' individualizing their therapy as well as help prevent physiologic adaptation to the electrical stimulation. While numerous studies have delved into optimizing the electrical waveform (26–28), their conclusions have been varied, and it is likely, there is no optimal waveform. As such, TENS therapy is an individualized one, and patients should have frequent follow ups with their physician to ensure the patient is receiving optimal care for their specific complaint (Table 4).

Table 4. INDICATIONS for Use of TENS

Systemic Pain	
Bursitis	Phantom limb syndrome
Cancer	Raynaud's syndrome
Causalgia	Rheumatoid arthritis
Multiple sclerosis	Synovitis
Neuralgia	Diabetic peripheral
Osteoarthitis	Neuropathy
Fibromyalgia	
Head and Neck Pain	
Cluster headaces	Suboccipital headaches
Dental disorders	TMJ Syndrome
Migraine headaches	Torticollis
Spondylosis	Trigeminal neuralgia
Sprains/strains	Whiplash
Abdominal Pain	
Diverticulosis	Labor
Dysmenorrhea	Postoperative pain
Back Pain	
Facet syndrome	
Intercostals neuralgia	Radiculitis
IVD Syndrome	Sprains/strains
Lumbago	Thoracodynia
Lumbosacral pain	Whole back pain
Lower Extremity Pain	
Ankle pain	Passive stretch pain
Foot pain	Sciatica
Fractures	sprains/strains
Ischialgia	tendonitis
Knee pain	Thrombophlebitis
Upper Extremity Pain	
Epicondylitis	
Frozen shoulder	Sprains/strains
Hand pain	Subdeltoid bursitis
Peripheral nerve	Wrist pain
Injury	

TREATMENT PLANS

Treatment with TENS is an extremely variable and personalized process, and this cannot be underscored enough. It is vital for close healthcare supervision for a user to obtain the maximum therapeutic benefit from tens. TENS may only be used in an acute phase for a short period of time (e.g., incisional pain postsurgery) or for months or years (e.g., those suffering from chronic back pain). For chronic pain sufferers classical TENS may be used for several hours continuously daily. Modulated or low frequency/high intensity may be used for ~30 min three times a day for an indefinite period of time. When using TENS it is important to use as strong or nearly asstrong a current as the user can tolerate to achieve best results.

Fibromyalgia is a poorly understood chronic pain condition that presents unique management challenges not only because it is poorly understood, but also because it is often refractory to traditional treatment modalities. A recent double-blinded study by Cork et al. (29) explored cranial electrotherapy stimulation (CES) as a possible treatment for fibromyalgia. In this study, using electrodes clipped to the participants' ear lobes, the Alpha-Stim CES device, delivered either modified square-wave stimulation at 100 μA and a 50% duty cycle at 0.5 Hz for 1 h daily for 3 weeks or sham therapy (see Fig. 15). While there were no

Figure 15. The Alpha-Stim CES device.

Figure 16. Electrode placement for CES.

differences in baseline pain scores of the participants in either group prior to beginning the study, after 3 weeks of CES therapy those in the treatment group displayed significantly lower Pain Intensity Scores, Tenderpoint Scores, and POMS Scores compared to the sham group. After 3 weeks the study was unblinded, and 23 of the 35 participants in the Sham group elected to switch over to active treatment for 3 weeks. Those switching from sham therapy to active CES therapy displayed significant reductions in the aforementioned pain scores as well (29) (Fig. 16).

WARNINGS AND CONTRAINDICATIONS

TENS is contraindicated in individuals with pacemakers, especially those with demand-type pacemakers as the electrical stimulation could cause misfiring or other malfunction of the pacemaker. Electrode placement in areas of sensory or circulatory deficits should be avoided due to the potential for burns or excessive muscular contraction. Electrodes should not be placed over the carotid sinuses to prevent a vasovagal reflex reaction with resultant hypotension. Electrodes should not be placed over the anterior neck due to potential to induce laryngospasm and subsequent asphyxiation. Electrodes should not be placed over the eyes. TENS should be avoided in pregnant women due to the potential to induce contractions and premature labor. Caution should be used in patients with implanted spinal stimulators as well as intrathecal opiate pumps. The unit should not be used with other electrical medical equipment, such as ECGs, EEGs, pulse oximeters, and electrocautery devices.

PRECAUTIONS

Tens has not been proven to have curative value and should be used only under the supervision of a physician. Patient selection is crucial, and not all patients will respond to TENS. The TENS has not been shown to exhibit curative value and should not be used for pain of unknown origin. The unit itself as well as wires and electrodes should be kept out of reach of children and water.

BIBLIOGRAPHY

1. Medical Data International, Market and Technology Reports, U.S. Markets For Pain Management Products, June 1999 Report RP-821922.
2. Bone M, Critchley P, Buggy DJ. Gabapentin in postamputation phantom limb pain: A randomized, double-blind, placebo-controlled, cross-over study. Reg Anesth Pain Med 2002 Sep–Oct; 27(5):481–486.
3. Zuurmond WW, van der Zande AH, de Lange JJ. Phantom pain following leg amputation: Retrospective study of incidence, therapy and the effect of preoperative analgesia. Ned Tijdschr Geneeskd 1996 May 18; 140(20):1080–1083.
4. Fields, HL. Pain. New York: McGraw-Hill, 1987.
5. Jessell T, Kandel E, Schwartz J, editors. Principles of Neural Science. New York: McGraw-Hill; 2000 p 472–492.
6. Haines D, editor. Fundamental Neuroscience. Philidelphia: Churchill Livingstone; 2002. p 273–292.
7. Willis, WD Jr. The Pain System: The Neural Basis of Nociceptive Transmission in the Mammalian Nervous System. New York: Karger, 1985;
8. Melzack R. The Puzzle of Pain. New York: Basic Books; 1973. p 55–56.
9. Zborowski M. Cultural components in responses to pain. J Soc Issues 1952;8(4):16–30.
10. Melzack R, Wall PD. Pain mechanisms: A New theory Science 1965;150:971–979.
11. Luo F. A study on the cumulative effect of repeated electroacupuncture on chronic pain. Sheng Li Ke Xue Jin Zhan 1996 Jul; 27(3):241–244.
12. Han JS, et al. Effect of low-and high-frequency TENS on Met-enkephalin-Arg-Phe and dynorphin A immunoreactivity in human lumbar CSF. Pain 1991 Dec; 47(3):295–298.
13. Kalra A, Urban MO, Sluka KA. Blockade of opioid receptors in rostral ventral medulla prevents antihyperalgesia produced by transcutaneous electrical nerve stimulation (TENS). J Pharmacol Exp Ther 2001 Jul; 298(1):257–263.
14. Sluka KA, et al. Spinal blockade of opioid receptors prevents the analgesia produced by TENS in arthritic rats. J Pharmacol Exp Ther 1999 May; 289(2):840–846.
15. Han JS, Chen XH, Yuan Y, Yan SC. Transcutaneous electrical nerve stimulation for treatment of spinal spasticity. Chin Med J (Engl) 1994 Jan; 107(1):6–11.
16. Rodriguez E, et al. Effects of transcutaneous nerve stimulation on the plasma and CSF concentrations of beta-endorphin and the plasma concentrations of ACTH, cortisol and prolactin in hysterectomized women with postoperative pain. Rev Esp Anestesiol Reanim 1992 Jan–Feb; 39(1):6–9.
17. Liss S, Liss B. Physiological and therapeutic effects of high frequency electrical pulses. Integr Physiol Behav Sci 1996 Apr–June; 31(2):88–95.
18. Kho HG, Kloppenborg PW, van Egmond J. Effects of acupuncture and transcutaneous stimulation analgesia on plasma hormone levels during and after major abdominal surgery. Eur J Anaesthesiol 1993 May; 10(3):197–208.
19. Gordon G. Electroanalgesia: Historical and Contemporary Developments. 1998 Available at URL:http://freespace.virgin.net/joseph.gadsby/index.htm.
20. Schechter DS. Origins of electrotherapy. N Y State J Med 1971 May 1; 71(9):997–1008.
21. Kellaway P. The William Osler Medal Essay: The part played by electric fish in the early history of bioelectricity and electrotherapy. Bull Hist Med 1946;20:112–137.
22. Long DM. Fifteen Years of transcutaneous electrical stimulation for pain control. Stereotact Funct Neurosurg 1991;56(1):2–19.
23. Hymes A. Introduction: A review of the historical uses of electricity. In: Mannheimer JS, Lampe G, editors. Clinical

Transcutaneous Electrical Nerve Stimulation. Philadelphia: Davis Company; 1984 p 1–5.
24. Szeto A. Pain relief using transcutaneous nerve stimulation. In: Webster JG, editor, Encyclopedia of Medical Devices and Instrumentation. Vol 4 New York: John Wiley & Sons; 1988. p 2203–2220.
25. Wolfe P. A Practical Approach to Transcutaneous Electrical Nerve Stimulation (TENS). New Brighton: Rehabilicare, Inc.
26. Barr JO, Nielsen DH, Soderberg GL. Transcutaneous electrical nerve stimulation characteristics for altering pain perception. Phys Ther 1986 Oct; 66(10):1515–1521.
27. Katims JJ, Long DM, Ng LK. Transcutaneous nerve stimulation. Frequency and waveform specificity in humans. Appl Neurophysiol 1986;49(1–2):86–91.
28. Repperger DW, et al. Microprocessor based spatial TENS (transcutaneous electric nerve stimulator) designed with waveform optimality for clinical evaluation in a pain study. Comput Biol Med 1997 Nov; 27(6):493–505.
29. Cork R, et al. The Effect of Cranial Electrotherapy Stimulation (CES) on Pain Associated with Fibromyalgia. Int J Anesthesiol 2004;8(2). Available at http://www.ispub.com/ostia/index.php? xmlFilePath=journals/ija/vol8n2/ces.xml.

See also BLADDER DYSFUNCTION, NEUROSTIMULATION OF; ELECTROANALGESIA, SYSTEMIC; ELECTROPHYSIOLOGY; FUNCTIONAL ELECTRICAL STIMULATION.

TRANSPLANTATION, LIVER. See LIVER TRANSPLANTATION.

TRAUMA MANAGEMENT. See CARDIOPULMONARY RESUSCITATION.

TWEEZERS, OPTICAL. See OPTICAL TWEEZERS.

U

ULTRASONIC HYPERTHERMIA.

See HYPERTHERMIA, ULTRASONIC.

ULTRASONIC IMAGING

OLIVER KRIPFGANS
University of Michigan
Ann Arbor, Michigan

INTRODUCTION

Medical imaging has many modalities and most of them provide clinicians with unique features of a volume of interest (VOI) resulting from a chosen modality. Ultrasonic imaging is one technique for collecting anatomical and physiological information from within the human body. It can be used for diagnosis (imaging) and for image-guided therapy, where therapeutic intervention can be applied with direct image-based feedback. Other modalities include X ray (roentgen radiation), CT (computed tomography), MRI (magnetic resonance imaging), PET or PET/CT (positron emission tomography), and SPECT (single photon emission computed tomography). In contrast to most other imaging techniques, ultrasonic imaging is very attractive to professionals because it is cheap, real time (with >100 full frame images per second, >100 Hz), and it uses nonionizing radiation. Moreover, current clinical ultrasound machines can be integrated into laptop computers with very little external hardware for maximum portability and versatility. These combined features allow the use of ultrasonic imaging in a wide variety of settings, from private physician practices, to ambulances with on-site paramedics, to battle field situations, where very robust and lightweight equipment is required. Many other uses of ultrasonic imaging are found in science and industry these include, for example, ultrasonic microscopy, nondestructive testing and touch sensitive screens.

PHYSICAL PRINCIPLES

Ultrasonic imaging is based on ultrasound, which is sound produced at frequencies beyond those detectable in human hearing, that is, >20 kHz. In the same way that ultraviolet (UV) light is invisible to the human eye, ultrasound is inaudible to the human ear. Often objects that serve as carriers for ultrasound waves need to be treated as waveguides. Nonlinear effects become apparent for ultrasound propagation when leaving the range of elastic deformation during the propagation of waves through a medium. Physical material constants form ultrasound parameters, for example the speed of sound or the attenuation of sound. Very high frequency sound waves are treated by quantum acoustic laws. Historically, ultrasound was produced by oscillating platelets (1830), or pipes (1876). Magnetostriction (1847) and the piezoelectric effect followed (1880 by Curie), and are still very much relevant mechanisms for medical and industrial ultrasound. In 1918, it was found that the use of oscillating crystals could be used to stabilize frequencies. The upper frequency for sound in a given solid material is determined by the separation of neighboring atoms in the host medium. This upper frequency limit is met when neighboring atoms, assuming the linear chain model, oscillate with a 180° phase shift, the so-called optical branch of oscillations in a solid (1–15).

Sound Waves

Sound waves are mechanical waves by nature and a medium is needed to carry them. These spatial–temporal oscillations occur nonsynchronously throughout a medium and cause density fluctuations, which in turn cause temperature oscillations if the rate of such fluctuations is larger than the time constant for thermal equalization. Typical properties to describe sound waves are

Displacement $\boldsymbol{s} = \boldsymbol{s}(\mathrm{t})$ of a particle due to a traversing wave
Sound particle velocity $\frac{d}{dt}\boldsymbol{s} = \frac{d}{dt}\boldsymbol{s}(t) = v(t)$
Instantaneous mass density $\boldsymbol{\rho} = \boldsymbol{\rho}(t)$
Instantaneous pressure $\boldsymbol{p} = \boldsymbol{p}(t)$

where the later most is the deviation from the ambient static pressure. Mechanical properties, such as strain tensor σ_{ij} and stress tensor s_{ij} can be used to develop relationships between the above mentioned properties of sound waves (Fig. 1).

For mostly lossless media, such as water, one can use the laws of conservation of momentum and conservation of mass to derive the wave equation for sound waves,

$$\nabla p(r) = \frac{1}{c^2}\frac{\partial^2}{\partial t^2}p(r) \qquad \frac{B}{A} = \mathrm{const}\left[\frac{\partial c}{\partial p}\right]_T + \mathrm{const}\left[\frac{\partial c}{\partial T}\right]_p \quad (1)$$

where spatial variations in pressure ∇p are coupled to temporal variations $(\partial p)^2/(\partial t^2)$ via the speed of sound c. This simple relationship represents only a linear approximation, and is therefore valid only for waves of small displacements. The isentropic nonlinearity parameter B/A measures the amount of density change (ρ, $c = (\rho K)^{-1/2}$ (see Eq. 10) for a given pressure (p) and temperature (T) (6). Note that only second-order temporal derivatives in the pressure result in a spatial pressure change, that is, only accelerations can result in sound.

Planar and Spherical Waves. In general, mechanical waves can either travel as longitudinal waves by compressing and expanding the host medium itself or as transversal waves by exerting shear force on the host medium. Water is a very good host medium because it bears minimal

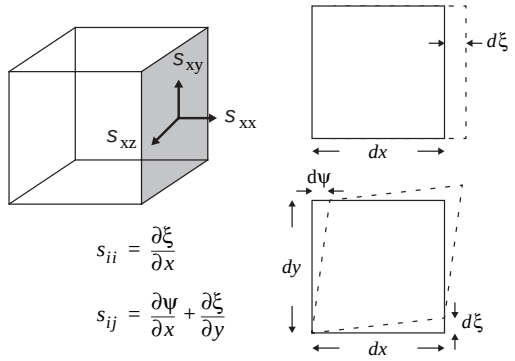


Figure 1. Ultrasonic waves travel through media by elastic deformation of matter. This deformation can be written in terms of strain σ_{ij} and stress s_{ij} tensors.

energy loss for traveling sound waves. The human body is 55–60% (8) composed of water, which ensures good acoustic accessibility. Exceptions are areas blocked by either bone or air, since bone and air provide poor ultrasound transmission characteristics. However, water mostly supports longitudinal waves. Transverse waves are attenuated at a high rate in water and can therefore be neglected. The three-dimensional (3D) wave equation in Eq. 1 reduces then to a one-dimensional (1D) equation with the general solution of an inward and outward propagating wave:

$$\left.\begin{aligned} p(x,t) &= F(x-ct) + G(x+ct) \\ p(r,t) &= \frac{1}{r}(F(r-ct) + G(r+ct)) \end{aligned}\right\} \quad \begin{matrix}\text{for Planar and} \\ \text{spherical waves}\end{matrix} \tag{2}$$

where x and r are the direction of propagation, t is time, and F and G are general, but continuously differentiable functions. The rationale for these arbitrary functions is their argument $x \pm ct$ and $r \pm ct$. This expression ensures the character of the wave as a traveling entity. Whatever the function F represents at time $t = 0$ at position $x = 0$, it will travel to position x/c after time t. In other words, by keeping the argument of the function $F = 0$, one can compute where and how fast the wave travels. Vice versa for the function G, except that it travels in the opposite direction (Fig. 2). For harmonic waves, these two functions are represented by harmonic functions, that is, sin, cos, or more general e^{ix}. Planar and spherical waves follow as:

$$\begin{aligned} p(x,t) &= pe^{ik(x \pm ct)} && \text{planar waves} \\ p(r,t) &= \frac{p}{4\pi r} e^{ik(r \pm ct)} && \text{spherical waves} \end{aligned} \tag{3}$$

Here, k is the wave number, defined as $2\pi/\lambda$, which is the conversion between spatial coordinates of wavelength to

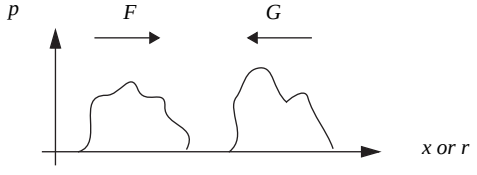


Figure 2. Illustration for the general solution of equation 1; two traveling waves F and G in opposing directions, $+r$ and $-r$, respectively.

radians in the complex exponential. Acoustic attenuation or absorption can be mathematically written as a imaginary valued wave number k_i, leading to the complex valued total wave number $k = k_r + k_i$.

Quantifying Sound

Sound intensity $I(\mathrm{W \cdot cm^{-2}})$ and acoustical power P [W]

$$\boldsymbol{I} = \overline{p\boldsymbol{v}}, \quad P = \oint_S \overline{p\boldsymbol{v}}_n dS \tag{4}$$

are measures of the strength of the acoustic waves. In both equations, the temporal average of the product of pressure and particle velocity is computed. In the equation for intensity in equation 4, the temporal average value is a vector quantity, while in the equation for power it is a scalar because the velocity is taken as the normal component to the encapsulating surface element dS. In addition to acoustic intensity and power, very often one refers to a measure for the acoustic pressure. In SI units, pressure is given by Newton per square meter $(\mathrm{N \cdot m^{-2}})$. However, sound pressures extend over a large range, and therefore a logarithmic scale, the dB scale, is commonly used to measure pressure.

$$\text{dB value} = \begin{cases} 20 \cdot \log p_{\mathrm{rms}}/p_{\mathrm{ref}} \\ 10 \cdot \log I_{\mathrm{rms}}/I_{\mathrm{ref}} \end{cases} \tag{5}$$

As can be seen from Eq. 5, a reference value must be used to compute the pressure level on a dB scale. Typically, this reference is chosen to be the peak output level of the system under test or it can be a fixed value such as when 1 mW into 50 Ω is used on some oscilloscopes. Sound pressure level (SPL) and sound intensity level (SIL) reference values in underwater ultrasonics are 1 μPa and $10^{-12}\ \mathrm{W \cdot m^{-2}}$, respectively. In contrast, SPL for air bourne sound is 20 μPa, whereas SIL remains at $10^{-12}\ \mathrm{W \cdot m^{-2}}$.

Acoustic Impedance

Impedance is a term commonly known from electromagnetism. However, it also applies to sound waves and relates sound pressure and particle velocity in a manner analogous to Ohm's law. One distinguishes at least four types of acoustic impedance: specific acoustic impedance (z) is used to compute the transmission of an acoustic wave from one medium into another; acoustic impedance (Z) is used to estimate the radiation of sound from vibrating surfaces; mechanical impedance (Z_m) is the ratio of a complex driving force and the resulting complex speed of the medium; and radiation impedance (Z_r), which is used to couple acoustic waves to a driving source or a load driven by a force.

$\boldsymbol{z} = \boldsymbol{p}/\boldsymbol{v}$	Specific acoustic impedance
$\boldsymbol{Z} = \boldsymbol{p}/\boldsymbol{U} = \boldsymbol{z}/\boldsymbol{S}$	Acoustic impedance
$\boldsymbol{Z}_m = \boldsymbol{f}/\boldsymbol{U}$	Mechanical impedance
$\boldsymbol{Z}_r = Z/S = \int d\boldsymbol{f}\boldsymbol{s}/\boldsymbol{v}$	Radiation impedance

Here, $\boldsymbol{p}$ is the acoustic pressure as a function of space and time, $\boldsymbol{v}$ is particle displacement velocity as a function of

space and time, $\boldsymbol{U}$ is a volume velocity, $\boldsymbol{S}$ is the surface that emits the sound, $\boldsymbol{f}$ is a complex driving force of the sound source, such as the force of a coil that makes the membrane of a loudspeaker move, and subsequently $\boldsymbol{u}$ is the complex speed at which the forced medium is moving.

Moreover, a quantity called characteristic impedance is analogous to the wave impedance $\sqrt{\mu/\varepsilon}$ of a dielectric medium. Its analytical form depends on the type of wave. Equation 7 shows the closed form expression for planar and spherical waves. It should be noted that the characteristic impedance for spherical waves can be complex valued and therefore pressure (p) and velocity (v) are not required to be in phase.

characteristic impedence:

$$\begin{aligned} z &= \frac{p}{v} = \rho_0 \cdot c && \text{For planar waves} \\ z &= \frac{p}{v} = \rho_0 c \cdot \frac{kr}{(1+(kr)^2)^{1/2}} e^{j\theta} && \text{For spherical waves} \end{aligned} \tag{7}$$

For these two special cases one can see that planar waves have pressure (p) and particle velocity (v) in phase. The ratio of pressure to velocity is a real number and is constant ($\rho_0\, c$). Spherical waves, however, behave differently. Pressure and particle velocity are out of phase when measured close to the sound source. The ratio of pressure to velocity is a complex value (due to the $e^{j\theta}$ term and $\cot\theta = kr$) and it changes with distance (r). For large r, the spherical wave solution approaches the planar wave solution, that is, when $kr \gg 1$.

The parameter Z, the acoustic impedance, is often referred to as a frequency independent constant. The importance of this property lies in the nature of ultrasound. When imaging the human body, the sound waves travel through many layers of varying impedances, such as skin, fat, connective tissues, and organs. The crossing of each tissue boundary changes the sound waves in several ways. Typically, sound both transmits and reflects from tissue layers. While sound is mostly transmitted, small reflections are recorded and displayed as gray levels in a so-called B-mode image, where larger amplitudes of reflected waves is displayed as brighter pixels. More complicated scenarios include mode conversion between longitudinal and transverse waves. Reflection and transmission coefficients for pressure are directly related to the *change* in acoustic impedance as given in the following equation:

$$R = \frac{Z_2 - Z_1}{Z_2 + Z_1} \quad T = \frac{2Z_2}{Z_2 + Z_1} \quad 1 + R = T \tag{8}$$

Here Z_1 and Z_2 are the impedances of the proximal and distal side of the interfacial surface. Therefore, no reflection will be seen from sound entering a layer of equal impedance, but varying density and speed of sound compared to the current medium. Reflection and transmission coefficients for intensities are derived by squaring R and T in Eq. 8 (Fig. 3).

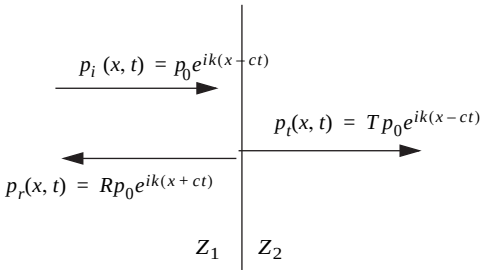


Figure 3. Acoustic propagation is altered when sound waves encounter an impedance change (Z_1 to Z_2), that is, a change in the product of local speed of sound and mass density.

Attenuation

Acoustic attenuation manifests itself in many ways. Sound can be attenuated by mechanisms of reflection, absorption, or scattering. Reflection is caused by impedance changes, whereas absorption and scattering occur due to the internal structure of the medium. Viscous forces cause sound absorption following Lambert–Beer–Bouguer's law

$$\begin{aligned} dI &= -\beta I dx \\ I(x,\beta) &= I_o e^{-\beta x} \\ p(x,t) &= p(x=0,t) \cdot e^{ikx \pm ct + ik_i x} \end{aligned} \tag{9}$$

where k_i is imaginary, and therefore $p(x, t)$ decays exponentially for $x > 0$. In general, acoustic waves in medical imaging are attenuated by traveling through layers of different tissues due to reflection and also due to attenuation inside tissue. Typical acoustic attenuation in biological tissues is of the order of 0.1–1.0 dB $\text{MHz}^{-1} \cdot \text{cm}^{-1}$, that is, a acoustic wave of 1 MHz center frequency, which travels 0.5 cm deep into tissue (1 cm round trip), is diminished by 0.1–1.0 dB, or its amplitude is reduced by ~1–11%. However, a 2.25 MHz wave penetrating the abdomen to a depth of 15 cm at 0.7 dB $\text{MHz}^{-1} \cdot \text{cm}^{-1}$, will be amplitude attenuated by 47.25 dB or 99.6%. Good ultrasound systems have signal to noise and amplification capabilities up to 120 dB, and therefore allow penetration to a reasonable depth at megahertz frequencies. Typical frequency selections are 7.5–15 MHz for 1–3 cm depths and 2.25–3.5 MHz for 12–15 cm depths.

Pulse–Echo

Most medical imaging via ultrasound uses a pulse–echo method to obtain images. That is, sound waves are transmitted into the body and echoes from within the body are registered, and their origin is computed using complex algorithms. Pulse–echo is somewhat unique to ultrasonic imaging. Other modalities use transmission (X ray and CT) or register preexisting radiation from within the body (PET, SPECT).

Multiple transmissions at the same physical location can reveal the motion of targets. A fundamental assumption is the speed of sound in the investigated volume. Typically, water is assumed to be 1485 $\text{m} \cdot \text{s}^{-1}$, and human tissue varies between 1450 and 4080 $\text{m} \cdot \text{s}^{-1}$ (see Table 1), with an average soft tissue value of 1540 $\text{m} \cdot \text{s}^{-1}$ (6). In general, the speed of sound is inversely related to the compressibility K ($\text{m}^2 \cdot \text{N}^{-1}$) and mass density ρ ($\text{kg} \cdot \text{m}^3$)

Table 1. Speed of Sound in Various Human Tissues[a]

Tissue	Mean Velocity, $m \cdot s^{-1}$	Tissue	Mean Velocity, $m \cdot s^{-1}$
Air	330	Brain	1541
Fat	1450	Blood	1570
Human tissue (mean)	1540	Skull bone	4080
		Water	1485

[a]See Ref. 6.

of the host medium:

$$c = \frac{1}{\sqrt{\rho K}} \tag{10}$$

Figure 4 shows the radio-frequency (rf) data for three firings at a set of moving targets. In this depiction, one can assume that the firings shown in a–c occur at a 1 ms interval. As time progresses, the scatterers move farther away from the transducer. At (a) the particles are $1/2 \cdot 26\,\mu s \cdot 1485\,m \cdot s^{-1}$, that is, 19.3 mm, away from the transducer; in (b) the particles shifted to $1/2 \cdot 32.5\,\mu s \cdot 1485\,m \cdot s^{-1}$, that is, 24.1 mm, away from the transducer. This shift of 6.5 μs or 4.8 mm is related to the interfiring time referred to by either pulse repetition interval or frequency (PRI or PRF). A PRI of 1 ms leads to the conclusion that the set of particles is moving at a speed of $4.8\,m \cdot s^{-1}$.

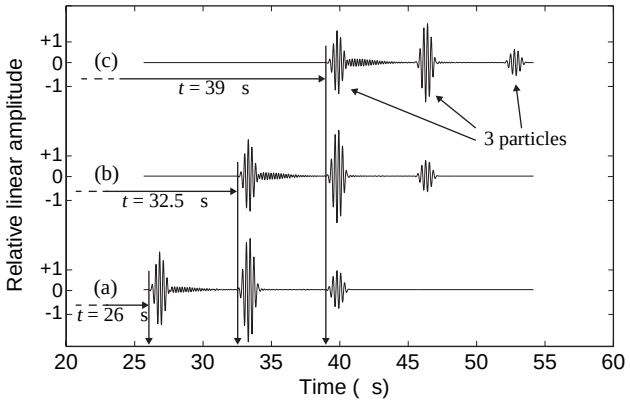


Figure 4. A set of three particles in water moves away from the ultrasound transducer during a set of three transmissions and receptions. (a) At time $t = 0$ ms the first backscatter signal of the set is received after 26 μs. (b) For the second acquisition the first backscatter is received at 32.5 and 39 μs at (c). Signal travel time is directly related to travel distance by means of the speed of sound.

ULTRASOUND GENERATION

Physics

Sound is produced by anything that moves in an accelerated fashion. Nowadays, most practical materials are piezoelectric, such as quartz (SiO_2), polyvinylidene fluoride (PVDF), and lead zirconate titanate (PZT). Piezoelectricity is an effect that is associated with the crystalline structure of the materials. A piezoelectric crystal yields a voltage across its surface when under strain, and the reverse effect facilitates mechanical oscillations of the crystal in response to an alternating current (ac) electric field applied across its surface.

Transducer Construction

Ultrasound transducers are made from piezoelectric materials, as described above. Typically, a layer of material is used to create a surface area for creation and transmission or reception of ultrasonic waves. The thickness of this layer is a function of the material properties and the desired acoustic frequency. As seen in Fig. 3, acoustic waves are reflected by impedance changes. An oscillating layer of piezoelectric material produces mechanical waves that propagate in the oscillation direction. These waves can be either compressional or shear waves. Here, the focus will be on compressional waves. Constructive interference of waves launched or reflected from the front surface and from the back surface of the crystal yield maximum pressure generation. High frequency transducers are made from very thin crystals due to their short wavelength and low frequency transducers are made from larger thickness crystals. For example, a 4 MHz transducer can be made from a 0.55 mm thick crystal. Table 2 lists the speed of sound in PZT_{5A} as $4400\,m \cdot s^{-1}$. The wavelength in PZT_{5A} at 4 MHz is 1.1 mm. Transducer crystals are typically machined to a thickness of λ/2, that is, 0.55 mm for 4 MHz. The rationale for this thickness is in the constructive interference of acoustic waves inside of the crystal. Figure 5 shows the bottom of the crystal moving up and down. Mechanical waves will launch from this surface and travel to either side of it. Assume that the

Table 2. Piezoelectric Properties of Typical Materials Used for Fabrication of Ultrasound Transducers

Property	Units	PVDF	PZT_{5A}	Quartz (x-cut)
Density	$g \cdot m^{-3}$	1.78	7.6	2.6
Relative permitivity	$\varepsilon/\varepsilon_0$	12	1700	4.52
Elastic modulus	$10^{10} N \cdot m^{-1}$	0.3	4.9	
Piezoelectric constant	$10^{-13} C \cdot N^{-1}$	$d_{31} = 20$	$d_{31} = 180$	
		$d_{33} = 30$	$d_{33} = 360$	
Coupling constant		0.11	$k_{31} = 0.35$	
			$k_{33} = 0.69$	
Speed of sound	$m \cdot s^{-1}$		4400	5740
Characteristic impedance	MRayl			15.2

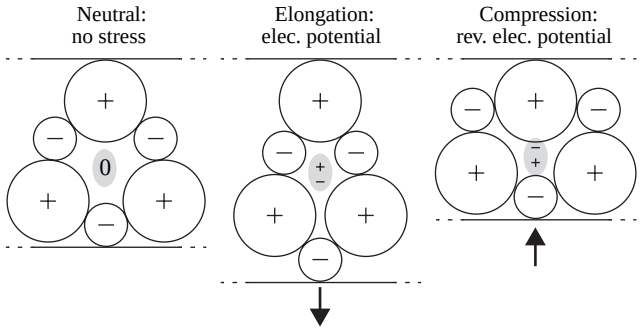


Figure 5. A piezoelectric crystal exhibits an electric charge on its surface when under mechanical stress (shown as the elementary cell response). The reverse effect is used to produce ultrasound; applying an alternating electrical potential across a piezoelectric crystal causes the crystal to vibrate along a given direction.

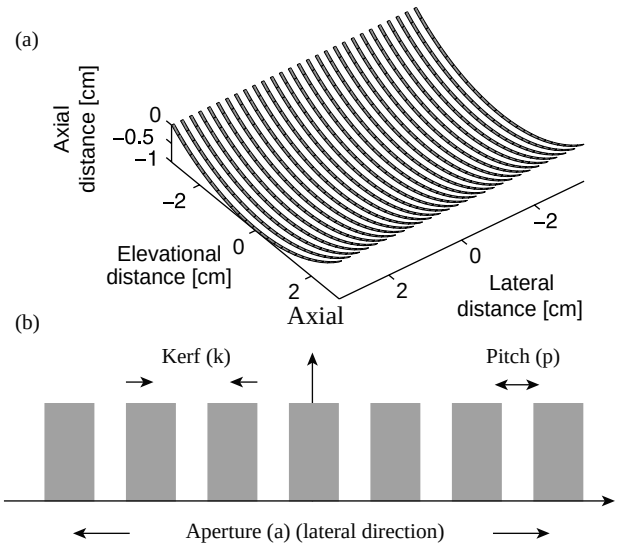


Figure 6. (a) Schematic of the piezoelectric elements of an array ultrasound transducer. Elevational curvature as well as spacing (white area, called kerf) between shaded elements is exaggerated. (b) Definition of geometric parameters: kerf, pitch, and aperture.

bottom side of the surface is facing air and that there is no sound transmitted into it. The sound wave traveling toward the top surface will be transmitted beyond that surface into the desired medium (e.g., tissue).

Reflected waves will travel downward and interfere with upward traveling waves. Moreover, reflected waves experience a phase shift of 180°. This is the reason that $\lambda/2$ is the required thickness and not λ.

Array Design. Modern clinical ultrasound imaging arrays are composed of hundreds of individual piezoelectric elements. Mostly these elements are arranged in a linear, 1D fashion, hence their name: linear arrays. The 3D in front of an imaging array are denoted axial, lateral, and elevational. Axial and lateral axes lie in the imaging plane, with the axial distance extending away from the transducer. The lateral axis is parallel to the transducer surface, inside the imaging plane, whereas the elevational axis extends perpendicular to the imaging plane. By convention, the origin is located at the transducer surface in the middle of the active aperture (see the section Acoustic Imaging). Each element may be rectangular with a fixed curvature for focusing in the elevational direction. Curvature as well as the elevational size of each array element determine the thickness of an image plane, which can be 1 mm. Typical element widths range from $\lambda_m/2$ to $3/2\ \lambda_m$. The wavelength λ_m is the wavelength within the medium where the wave is launched. Arrays with element sizes of $\lambda/2$ or smaller are referred to as fully sampled. Element sizes are typically 0.5–1 cm in the elevational direction and hundreds of micrometers in the lateral direction. The size of the space between individual elements is called the kerf (Fig. 6) and it is meant to diminish acoustic crosstalk between adjacent elements. A major design criterion for arrays is the distance between the centers of elements, called pitch. This distance determines the location and amplitude of acoustic grating lobes. Variations in the pitch, either due to changes in kerf or element width cause the grating lobes to shift. Moreover, the total extent of the aperture is directly related to the full width at half the maximum (fwhm) amplitude of the main lobe and the location of the side lobes. Equation 11 gives the fwhm of the main lobe and the angular position of the side lobes as well as grating lobes. Side lobes result from a transmit and receive aperture being small relative to the wavelength of the acoustic wave (λ/a), whereas grating lobes result from a steered and undersampled aperture, that is, an aperture with too few elements per wavelength.

$$\begin{aligned} &\text{fwhm main lobe} && \Lambda = \frac{aw}{p} \\ &\text{side lobes} && \theta_s = \arcsin\frac{\lambda n}{a} \quad n \in N_0 \\ &\text{grating lobes} && \theta_g = \arcsin\frac{\lambda n}{p} \end{aligned} \tag{11}$$

The consequence of improper selection of pitch and kerf for a given frequency is illustrated in Fig. 7a. At an imaging depth of 10 cm one can see a main lobe of almost 3 mm fwhm and strong side lobes and grating lobes. In this case the ratio of λ to element pitch is 0.4. For a ratio of λ to element pitch of 1.6, the width of the main lobe is 1.2 mm and the side lobes are mostly suppressed (Fig. 7b). Moreover, the grating lobes have disappeared when the ratio of λ to element pitch is increased.

High spatial resolution imaging is achieved by a proper selection of these geometry factors. Of additional importance to the acoustic field pattern are acoustic output, field of view, and practicality. The smaller the actual radiating area, the lower the acoustic pressure in the generated field. Moreover, a large pitch and/or element width will, for a given number of elements, cause the active aperture to increase in size, which limits the possible shift of the active aperture across the physical aperture of the array (see the section Acoustic Imaging). A

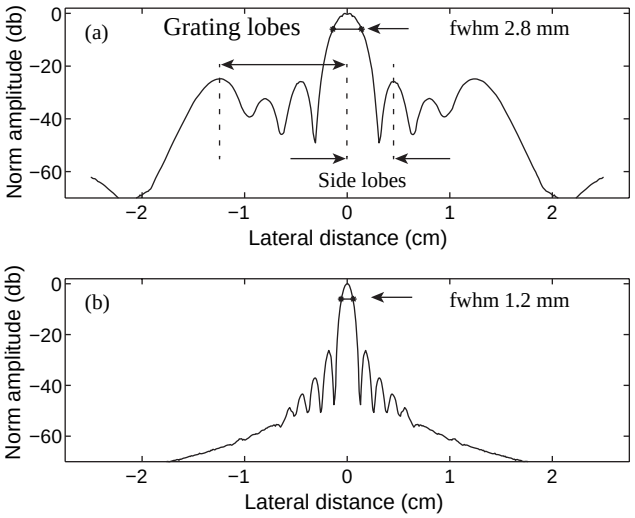


Figure 7. (a) Four elements (200 μm kerf, 1000 μm pitch) using 4 mm aperture, yield a main sound lobe width of 2.8 mm fwhm. (b) Forty elements (50 μm kerf, 250 μm pitch) using 1 cm aperture, cause the main lobe to narrow to 1.2 mm fwhm and the sidelobes and grating lobes to diminish. Half maximum corresponds to −6 dB.

large number of elements is beneficial for focusing and field geometries, but it is physically difficult to electrically wire a substantially larger number of elements. Furthermore, it directly increases costs, since each element requires cabling and multiplexing electronics. Channels with transmit and receive electronics are in general multiplexed to the physical elements. The complications involved with cabling hundreds or thousands of elements as well as the challenge of real-time data acquisition of a large number of channels are directly related to the still limited usage of two-dimensional (2D) imaging arrays. A good introduction and very detailed overview can be found in Shung and Zipparo (16), as well as Angelsen et al. (1,17).

Acoustic Fields. Acoustic fields of transducers can be analytically derived from basic principles. Logically, this derivation originates at the sound source, a moving object or surface. Its motion and surface shape/orientation define the so-called source strength Q (Eq. 12). Using this source strength one can compute the actual pressure field at a distance r from the source. In order to do so, it is necessary to derive the fact that for all simple geometries, the ratio of source pressure P_1 to its source strength Q_1 and the ratio of a second source pressure P_2 to its source strength Q_2 are equal at the same distance (assuming the same frequency). However, the derivation of this equality is beyond the scope of this text. The pressure of a circular piston can now be written as a function of its source strength, as well as the pressure and source strength of a known source, typically a small sphere. Equation 13 gives the analytical expression for the pressure field of the piston transducer, where r is the distance of the observation point from the center of the aperture, θ is the angle between the axis extending perpendicularly from the center of the transducer and the line from the center to the observation point, t is time, ρ is the mass density of the surrounding water, c is the speed of sound in water, U_0 is the sound particle velocity on the aperture, and ω and k are angular frequency and wave number, respectively. The integral is simplified for circular geometry and taken over the entire surface of the transducer's aperture. In *real* world simulations, one could take into account that the circumference of the aperture might be clamped or for other reasons not be able to oscillate with the same amplitude as the center of the aperture. To do so, one would keep U_0 inside the integral as a function of radius or even radius and in-plane angle of the aperture.

$$Q = \int_S (\boldsymbol{u} \cdot \boldsymbol{n}) dS \tag{12}$$

$$\begin{aligned} &\frac{P_1}{Q_1} = \frac{P_2}{Q_2} \Rightarrow P_p = \frac{P_s}{Q_s} Q_p \\ &P_p = \frac{\rho c}{-i2\lambda r} \int_S (\boldsymbol{u} \cdot \boldsymbol{n}) dS \\ &P_p(r, \theta, t) = \frac{i\rho c U_0}{r} e^{i\omega t} \int_S \frac{\mathrm{e}^{-ik\sqrt{r^2+s^2}}}{\sqrt{r^2+s^2}} 2\pi s ds \end{aligned} \tag{13}$$

After deriving the general pressure field, one can compute special cases that are of particular interest, such as the central axis, as well as the far-field angular distribution of the radiation pattern. Further simplification of Eq. 13 yields approximate expressions for both cases and plots are shown in Fig. 8.

$$\begin{aligned} &p(r, 0, t) = \rho c U_0 \left(e^{-ikx} - e^{-ik\sqrt{x^2+a^2}} \right) e^{i\omega t} \qquad \text{(central axis)} \\ &p(r \gg a, \theta, t) \\ &\quad = \frac{i\rho_0 c U_0}{2r} a(ka) e^{i(\omega t - kr)} \frac{2J_1(ka \sin\theta)}{ka \sin\theta} \qquad \text{(far field)} \end{aligned} \tag{14}$$

From the axial dependence one can see strong interference for locations close to the transducer surface. This region is called near the field or Fresnel region and it extends approximately $r = 4$ aperture diameters a away from the transducer. Beyond that point, the pressure falls off following a $1/r$ dependence, and this region is called the far field or Fraunhofer region. As a rule of thumb, the far field starts at $a^2(2\,\lambda)$. Imaging is impractical or impossible in the near field. However, one should keep in mind that this result is true only for a *single* element transducer, and subdividing the aperture into an array of small elements shifts the near-field–far-field transition toward zero based on the actual dimension of the array. Moreover, this transition point is also a function of the emitted frequency as represented by angular frequency and wave number in Eq. 13 and 14. For illustration purposes, a ka value of 8π was chosen for Fig. 8a. When plotting the angular field pattern in Fig. 8b. However, $ka = 4\pi$ was chosen to reduce the number of sidelobes and make the plot more readable. Similarly to Fig. 7, a strong main lobe and additional side lobes are evident in Fig. 8b, that interfere with the main lobe in the sense that appreciable sound will be detectable in nearly all directions. In fact, for this particular example echo amplitudes

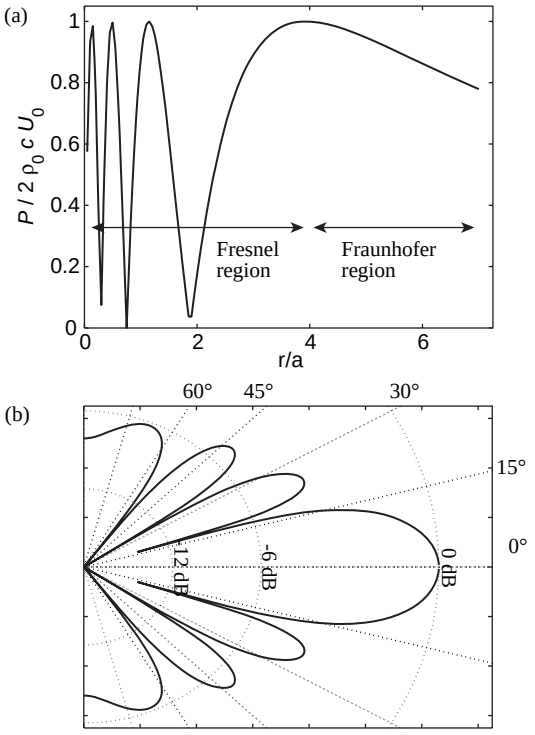


Figure 8. Acoustic field of a circular single element ultrasound source (transducer). (a) The axial transducer response can be divided into near field and far field, with an asymptotic $1/r$ dependence for large r. Phase interference is dominating in the near field. (b) Angular radiation patterns are described by the Fourier transform of the aperture function. A circular piston transducer yields a function that is defined as a Bessel function of the first type divided by its argument $[J_1(x)/x]$.

60–70° off the center axis will be only 50% lower than the main lobe. Note that this example is for educational purposes and is not at all a suitable design for imaging. Aperture sizes and frequencies as shown in Fig. 7, where a narrow and dominant main lobe can yield lateral and elevational spatial selectivity and time range gating, can yield depth information. In general, radiation patterns can be derived as Fourier transforms of the emitting aperture. Circular apertures are described by Bessel functions of the first type divided by their argument, that is, $J_1(x)/x$. For rectangular apertures the sine function $[sin(x)/x]$ directly describes the field.

ACOUSTIC IMAGING

Ultrasonic imaging yields 2D images. Pixel columns represent the lateral extent and rows of pixels display reflections of the transmitted acoustic waves from progressively deeper tissues within the body.

The most rudimentary way of obtaining lateral image data is done by using a single element transducer and wobbling it back and forth over a chosen sector. The most elegant way is to use an electronically controlled array of very small transducers and scanning and or steering an imaging beam across the region of interest. The former method is rarely used anymore. The latter method is the modern standard for ultrasonic imaging of single image planes and a few clinical scanners are even already available for 3D image acquisition of steered elevational image planes.

Depth information is encoded in the time that an acoustic wave takes to travel to a tissue site and back to the transducer. Transmitted sound can be received after a theoretically predictable wait time (see top section of Fig. 9). Such prediction requires the knowledge of the speed of sound along the traveled path. Unfortunately, an assumption of $1540\,\mathrm{m}\cdot\mathrm{s}^{-1}$ for soft tissue is not always precise. Various tissues in the human body differ from each other in terms of their specific speed of sound. When performing abdominal scans, aberration distortions can become significant due to the change in the speed of sound between connective tissues, fat layers, muscles, and abdominal organs. Other anatomical sites that provide difficulties for ultrasonic imaging include the human brain

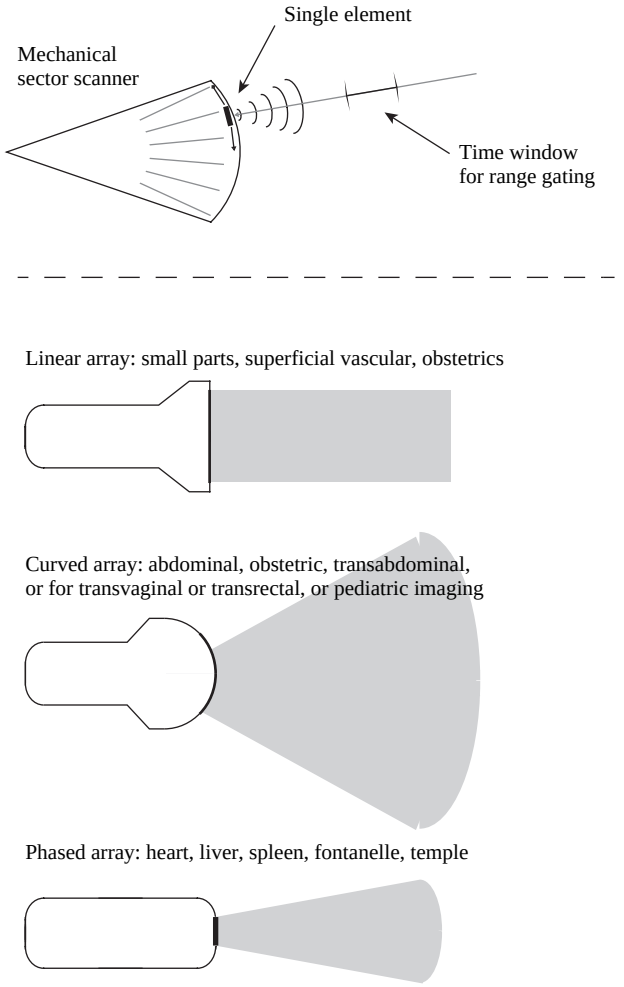


Figure 9. Transducer geometries include (curved) linear arrays and phased arrays. Linear arrays transmit and receive acoustic beams perpendicular to their surface area, whereas phased arrays steer the acoustic beam by a phasing scheme (Fig. 10).

and the heart. Fortunately, for newborns and infants ultrasound can be used to image the brain. Using a phased array (see bottom section of Fig. 9), pediatricians can image the brain directly through the fontanelle, which provides enough acoustical access for imaging. Premature newborns tend to have bleeding in the brain and develop larger ventricles, both of which can be imaged very easily through the fontanelle. However, when this acoustic access window closes, it is very difficult if not impossible to use ultrasound to image the brain. In adults, the temple fissure can be used to image (using a phased array Doppler system at 2.25 MHz) the germinal matrix near the foramen of Monroe. However, transcranial Doppler requires some guesswork on the orientation of vessels. Because of the limited access via the temple fissure, only small aperture and low frequency arrays with sub-optimal imaging capabilities can be employed.

The simplest imaging device is a linear array (range of transducer frequencies: 3–12 MHz). As its name suggests, this type of transducer has a linear arrangement of individual transducer elements. Images from a linear array are generally rectangular, and the image width corresponds to the width of the array. A set of adjacent elements (a subaperture) is used to fire a single image line or a portion thereof. Figure 10 shows how a subaperture can be used to steer and focus a beam. On the left side of each drawing, single sinusoids symbolize the electrical signals being used to excite the individual transducer elements (thick-lined vertical bars) of the array. After excitation an elementary spherically spreading wave launches from each element. Appropriate delay times applied to each electrical signal allow the ultrasound scanner to steer and focus the emitted wave front.

Additional wave conditioning includes amplitude shading of the subaperture. Typically, Gaussian-type functions or approximations thereof are used to weaken the transmit power for the outer most elements of a subaperture. The process is called apodization and yields lower side lobes since the side lobes are related to the Fourier transform of the aperture function. A gaussian apodization will result in a gaussian envelope for the side lobes, whereas no apodization (i.e., a constant amplitude excitation) would yield a sinc function ($\sin(x)/x$) side lobe envelope.

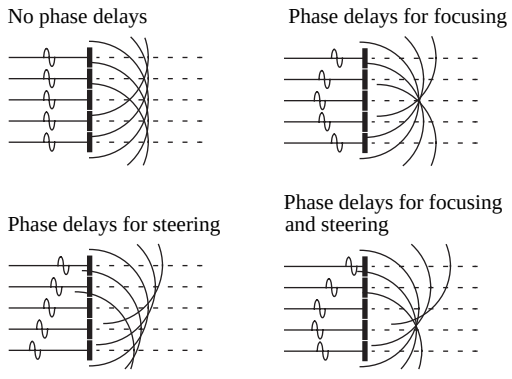


Figure 10. Delaying or advancing the phase of waves emitted from neighboring elements relative to the center element can achieve focusing and steering of the acoustic wave front.

To render an image, a subaperture is formed and linearly moved through the whole physical aperture. A major drawback of this imaging scheme is the dead time of the scanner, which occurs while waiting to receive the backscatter from the maximum depth. Fifteen centimeters of penetration require a wait time of at least (two fold for the round trip time):

$$t = \left[\frac{0.15\,\text{m}}{1540\,\text{m}\cdot\text{s}^{-1}}\right] \times 2 = 200\,\mu\text{s}$$

In addition to this delay, some additional wait time may be required to suppress echoes from even larger depths. Image lines might be separated by 250 μm. For a physical aperture of 5 cm, for example, one has to transmit and receive 201 scan lines, which is equal to a time of 200 μs * 201 lines, that is, 40.2 ms per one full frame or a frame rate of 25 Hz. This frame rate seems reasonable, but one has to keep in mind that no extra wait time was added nor any other imaging overhead such as occurs during blood flow imaging using Doppler.

Nonetheless, if one takes into account the finite lateral width of a transmitted wave, one can subdivide the total image width into independent image segments in which beams can be fired simultaneously, or at least with minimal delay (see Fig. 11). Therefore an aperture of 5 cm could eventually be imaged with five simultaneous beams or a five+fold higher frame rate. Such high frame rates allow real-time ultrasonic imaging of the body and additional overhead, such as is mentioned above for Doppler or multizone focusing. This type of focusing is used when a large depth image would cause the acoustic beam to diverge too much before and after the focal point. By firing the same image line two, three, or four times, the same number of foci can be formed for tighter acoustic beams at larger depths or for shallow regions. This scheme will work to the limit that the beams are not overlapping, that is, there is no signal coming from adjacent image lines.

Other imaging modalities, such as X ray, do not suffer from slow acquisition time due to low wave speeds. Another method to overcome the speed of sound limitation is the use

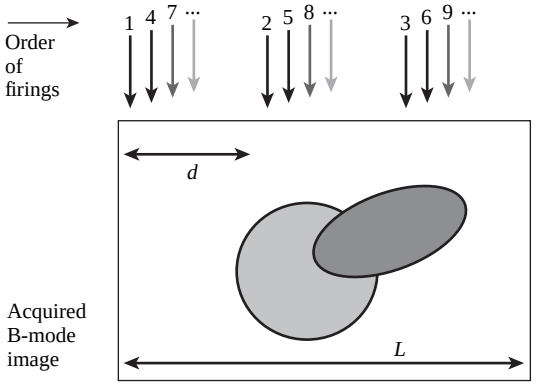


Figure 11. Interleaving scheme for image data acquisition. By minimizing the segment separation d, a maximum of L/d semisimultaneous transmit and receive zones can be achieved. The beam shape determines the smallest value of d, such that no overlap between adjacent firings will occur.

of *synthetic aperture* imaging. This method is already used in astronomy. Instead of 201 firings for 201 image lines, only one transmit is fired, and only one receive is recorded. Image reconstruction and especially spatial resolution will be computed/extracted from received data by extensive postprocessing. This scheme does not yield as much detailed backscatter data; however, it does yield very high frame rates. If only the speed of sound is limiting the data acquisition, then for the above example, the frame rate will increase by a factor of 201. Only the vast processing power of current computers makes synthetic aperture imaging possible. Typical applications for this type of image formation include full screen flow imaging, 3D, or 4D imaging (see respective subsections).

Other types of arrays include curved arrays and phased arrays which are popular for scans that require a larger image width then can be achieved by simply extending the physical aperture. Curved arrays, as seen in Fig. 9, form sector images. Because of the shape of the aperture, a relatively wide image can be achieved using a smaller footprint aperture. Scan lines are no longer parallel to each other but form a fan beam arrangement with field of view angles of up to 85° (150° for some endocavitary arrays). Typical bandwidths range from 2–8 MHz, which is a lower range than for linear arrays since this type of probe is intended for large penetration depths where frequency dependent attenuation prohibits very high frequencies. The use of nonionizing radiation to achieve real-time imaging with large fields of view makes theses probes ideal for abdominal interoperative guidance of, for example, biopsy needles or radiofrequency (RF) ablation tumor treatments. These features are also ideally suited for obstetrics (see Fig. 12).

Phased arrays are also designed to form sector images. Contrary to curved arrays, where the natural shape of the physical aperture provides the basis for the sector shape, phased arrays steer the beam to form the image. As illustrated in Fig. 10, specific timing delays for the subaperture can not only focus to a specified depth, but also steer the beam in the lateral direction. Large fields of view can be achieved this way, but the development of increased side and grating lobes are a trade-off. Anatomical locations with small diameter access to larger distal regions can be imaged with this type of ultrasound array. For example, temple access can be used to image the frontal brain, and extension of the carotid artery above the jaw line is possible. Cardiac imaging typically relies on phased arrays due to acoustic shadowing from the rib cage, where one needs to image between narrowly spaced ribs in order to interrogate the much larger sized heart chambers.

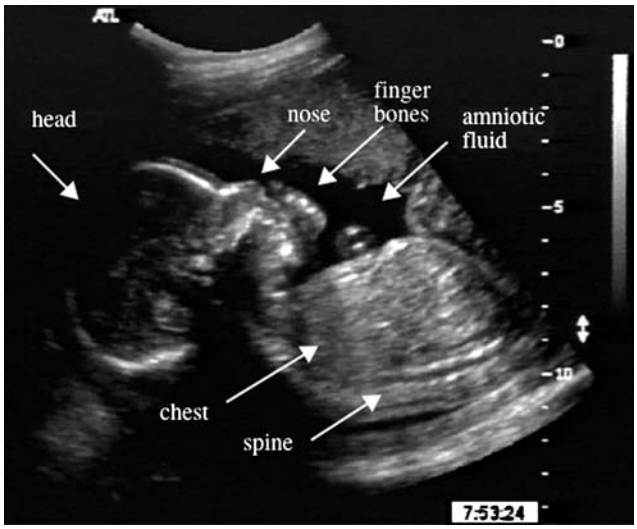


Figure 12. B-mode image of a fetus at the end of the second trimester (cross-sectional sagittal view). Low backscatter amniotic fluid is surrounding her head and upper body. The video reveals that the embryo is sucking her thumb. A curved array was used in this obstetrics exam.

B-Mode

B-mode is one of the most commonly used operation modes of a clinical ultrasound scanner. As explained earlier, ultrasound is a reflection or scattering based imaging modality, and the sophisticated generation of a sound wave allows the focusing of the sound to a specific location. Each transmission yields one scan line around the targeted focal point. If only one focal point is selected, one scan line extends over the total depth range, which is user defined in the current imaging settings. In order to record a full image frame using a linear array, the imaging software of the scanner electronically moves the active aperture of the array across the physical aperture to transmit and receive at a given line density. Typically, hundreds of scan lines are generated this way and displayed on a monitor. Figure 12 shows the cross-sectional sagittal (front to back, vertical slice) image of a fetus *in utero*. She is sucking on her thumb, as real-time video reveals. On the left of Fig. 12 one can see the head and the strong reflection of the skull bone. The very left side of the image is black, an artifact that could be due to maternal bowel gases that scatter the sound away from the transducer. The remainder of the skull is clearly visible from the forehead to the chin and from the back of the head to the neck area. Bones reflect sound waves well and result in a bright signal in the image. The black surroundings of the fetus are regions of amniotic fluid, which does not scatter sound due to the homogeneous nature of the fluid.

In front of the mouth, one can see the hand of the fetus. Once again the bones of the hand, namely, the knuckles, are pronounced since they scatter more ultrasound than the soft tissue of the hand. In the same fashion one can see the reflections of the spine.

M-Mode

M-mode (also called motion-mode imaging) does not yield full frame images per se, but rather one selected image line is rendered as a function of time. This is used for displaying motion of, for example, the periodic movement of heart valves. Any abnormalities or temporal variations can be directly seen as an image on the screen. The B-mode cross-section of a carotid artery is shown in Fig. 13a. Proximal and distal vessel wall delineates the dark vessel interior, as indicated by the arrows to the right.

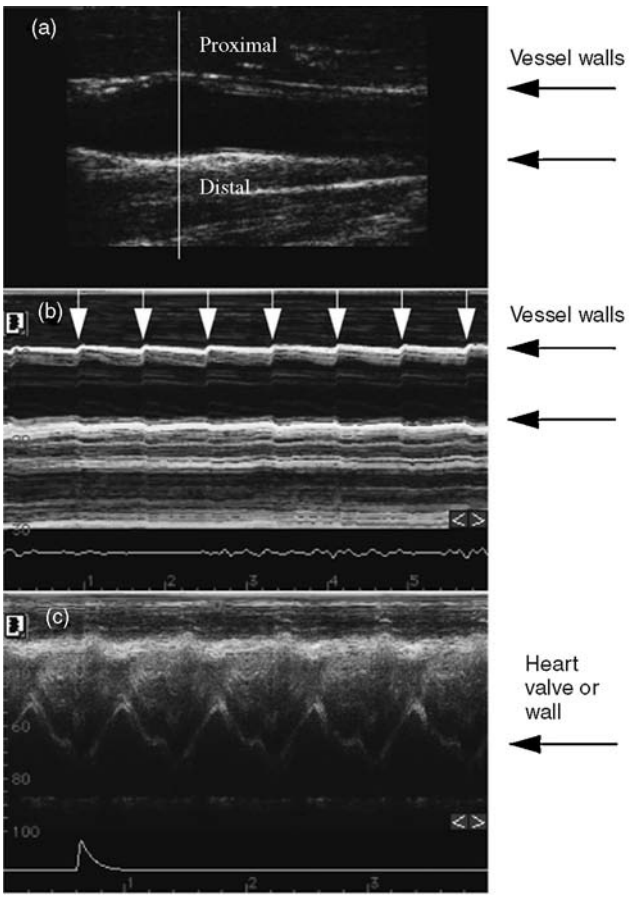


Figure 13. a. Longitudinal cross-section of carotid artery. The pixels along the vertical white line in (a) are plotted in (b) as a function of time. Black arrows indicate the proximal and distal vessel wall. Such walls are in motion as blood is pumped through the vessel in a pulsatile fashion. The repetitive pulsatile wall motion can be seen in the M-mode image in (b).

In a M-mode representation in Fig. 13b, pixels along the white vertical line in (a) are repeated parallel to each other over time. Figure 13b shows 5 s of repeated scans. For each heart beat a pulsatile wave travels through the arterial blood pool locally expanding the blood vessels. This expansion can be seen in B-mode as well as in M-mode representation. However, in B-mode it is an event in time occurring over several image frames, whereas in M-mode this event is plotted as the horizontal axis and therefore easy to detect. White arrows in Fig. 13b indicate the temporal expansion of the blood vessel. Figure 13c shows a much more pronounced motion. The transducer was pointed toward the heart and is therefore either imaging the heart wall or one of the heart valves, showing the typical cardiac pattern.

Doppler Imaging

Acoustic transmission of multiple beams along the same line can reveal temporal changes in the human body. As seen in Fig. 4, pulse echo firings in a rapid fashion can track flow, as well as flow changes in time. A more formal derivation of the mathematical framework shall follow here.

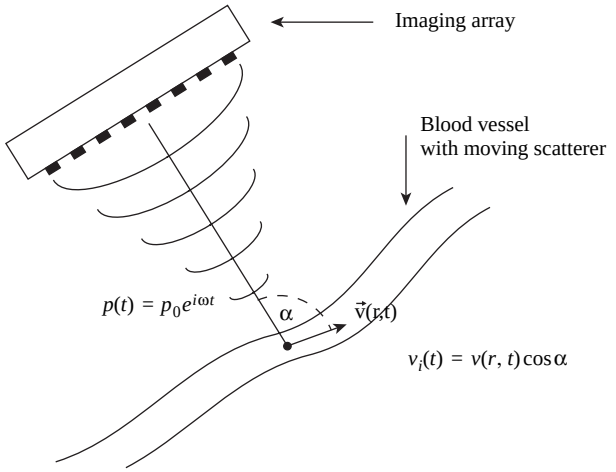


Figure 14. Illustration of Doppler imaging of a blood vessel.

Assume a vessel that is imaged at some angle a, which has acoustic scatterers such as red blood cells flowing at speed $v(r, t)$, as is depicted in Fig. 14. A recorded acoustic echo has an amplitude, frequency, and phase $(\alpha \cdot e^{i(\omega t+\phi)})$. The measured phase is the sum of temporal and spatial phase components. The first term $\omega\Delta t$ in Eq. 15 changes in time with ω and the second term changes in time with the velocity $v(t)$ of flowing red blood cells. Flow speed $v(t)$ and direction α $(\cos\alpha)$ determine the magnitude of the measurable phase shift. Due to the cos α term, any displacement that occurs parallel to the aperture will not be detected.

$$\Delta\phi(\Delta t) = \omega\Delta t - 2\pi\frac{v(t)\cos\alpha\Delta t}{\lambda} = \omega\Delta t\left(1 + 2\frac{v(t)}{c}\right) \quad (15)$$

It is assumed that the time between firings Δt is small enough that the scatterer does not move out of the main lobe of the beam pattern (Fig. 7). Moreover, it is assumed that $v(t)$ is constant during Δt. The absolute and relative received Doppler shift frequency can be directly derived from the change in phase, as the temporal derivative of the phase angle (Eq. 16).

$$\begin{aligned} f_{\text{receive}} &= \frac{1}{2\pi}\frac{\partial}{\partial t}\Delta\phi(t) = \frac{\omega}{2\pi}\left(1 + 2\frac{v_i(t)}{c}\right) \\ &= f_{\text{transmit}}\left(1 + 2\frac{v_i(t)}{c}\right) \\ \Rightarrow \frac{f_{\text{receive}}}{f_{\text{transmit}}} &= \left(1 + 2\frac{v_i(t)}{c}\right) = \left(1 + 2\cos\alpha\frac{v(t)}{c}\right) \end{aligned} \quad (16)$$

In the following example, a simulated scatterer is imaged and its traveling speed is measured. In the computation, the scatterer is travelling away from the imaging array from an axial distance of 2.50–2.52 cm, that is, 200 μm. On that path, the scatterer is imaged 32 times, once every 0.36 ms. Its speed is $2\,\text{cm}\cdot\text{s}^{-1}$. Figure 15a displays the backscatter of a scatterer as shown in Fig. 4, except that only one scatterer is imaged. Thirty-two

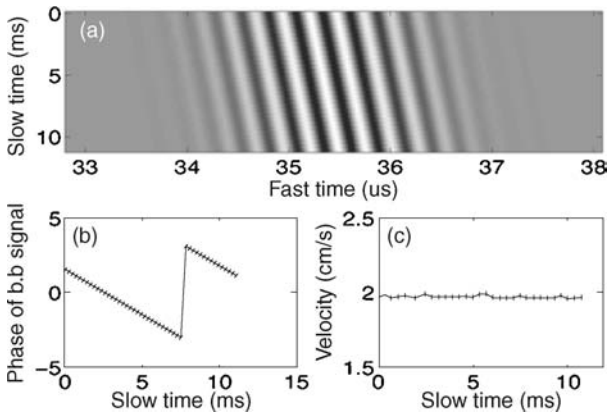


Figure 15. Doppler processing for velocity estimation. (a) Shows stacked (slow time) backscatter signals (fast time) of a moving scatterer. Complex analysis of such signals reveals the change in phase of the backscatter signal (b), and subsequently the scatterer velocity can be computed (c).

backscatter signals are stacked vertically, and the ordinate is labeled with the time at which the signal was measured ("slow" time in ms), whereas the abscissa shows the time frame of the measured radiofrequency signal ("fast" time in μs) same as in Fig. 4. This arrangement of backscatter data is very similar to that used in M-mode imaging. Signal amplitude is displayed as gray scale with gray for zero, black for negative amplitudes, and white for positive amplitudes. Two major steps are performed in order to estimate the velocity of the scatterer from the backscatter signal. At first the signal $f(t)$ is transformed into a complex valued signal $f^*(t)$ by performing a Hilbert transform (Eq. 17). Measured signals are always real valued quantities. However, for computational purposes it is desirable to have complex valued data $f^*(t)$. This step allows us to directly measure the phase of the backscatter signal and yield the velocity of the scatterer after basebanding, which is the second step.

$$\begin{aligned} f^*(t) &= \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{f(\tau)}{t-\tau} d\tau \\ &= \frac{1}{\pi} \lim_{\varepsilon \to 0^+} \left(\int_{-\infty}^{t-\varepsilon} \frac{f(\tau)}{t-\tau} d\tau + \int_{t+\varepsilon}^{\infty} \frac{f(\tau)}{t-\tau} d\tau \right) \end{aligned} \tag{17}$$

Basebanding is a mathematical procedure used to remove the carrier frequency, the main transmit frequency, from a rf-signal. For a complex valued signal, $f^*(t)$, this is done by multiplying the signal with a complex harmonic of the same, but negative frequency $(-\omega_0)$ as the carrier to obtain the complex valued envelope or amplitude modulation $a(t)$ and the phase ϕ (Eq. 18).

$$f^*(t) e^{-i(-\omega_0)t} = a(t) e^{i\phi(t)} \tag{18}$$

The phase is constant for each rf line in tissue, but varies between firings as targets, such as red blood cells, move. Figure 15b shows the phase of the basebanded signal and therefore the position of the scatterer. At ~7.5 ms, the phase wraps from $-\pi$ to $+\pi$ and continues to decrease. This phase wrap was detected and unwrapped before computing the velocity as proportional to the derivative of the phase. This phase unwrapping is not performed in clinical ultrasound scanners. Rather one will see flow of the opposite direction being displayed on the screen as the Doppler processing unit concludes that the sudden increase in phase from $-\pi$ to $+\pi$ must be due to flow in the opposite direction. This artifact is called aliasing and is typically avoided by increasing the pulse repetition frequency (PRF), that is, the rate at which Doppler firings are repeated along the same scan line in order to track backscatter from blood cells. Inverting Eq. 16 for $v(t)$ and using data processed via Eq. 17 yields the flow velocity as given in Eq. 19. A comprehensive description of medical Doppler and Doppler physics can be found in and McDicken and Evans (18).

$$v(\tau) = \frac{c}{2\omega_0} \frac{d}{d\tau} \phi(\tau) \tag{19}$$

Pulse Wave Doppler. Pulse wave Doppler (PW Doppler) is used for measuring blood flow. The user can position a Doppler scan line and Doppler window to any location within the B-mode image, as seen in Fig. 16. The two short horizontal lines at a depth of 6.9 cm in the top B-mode image in Fig. 16 represent the sample volume. This is where the Doppler data is acquired. Typically, the beamformer of the scanner is set to the same sample volume location for transmit and receive. The axial size of this window can be adjusted and is displayed on the screen (here 2 mm, see Size in the right side data column of Fig. 16). Changes in the window's size will affect the duration of the transmitted tone burst cycles. Commonly scanner software adjusts the duration of the transmit pulse to be twice as long as the chosen sample volume. Additionally, an angle (α) can be selected along which a blood vessel is oriented (here 0°, for example, along the Doppler scan line). As shown in Fig. 14 and Eq. 16, the measured flow

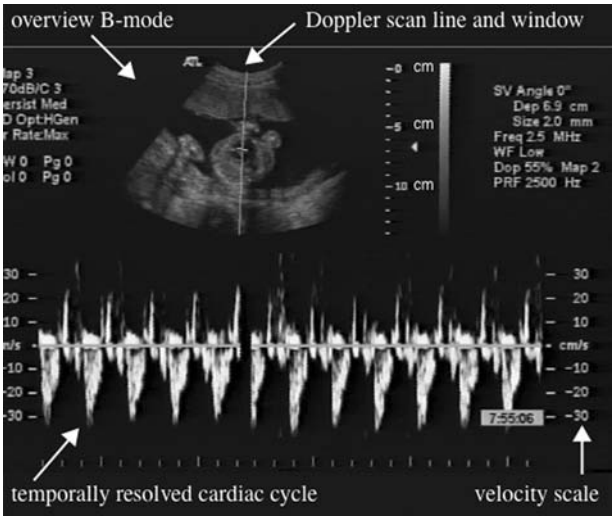


Figure 16. Pulse wave Doppler example. See text for more description.

velocity is only the projection of the actual flow vector onto the acoustic beam, which results in a cos α term. By manually choosing the correct α, the actual flow velocity can be computed and displayed from the measured flow.

The bottom part of the screen image in Fig. 16 shows the temporally resolved blood velocity, where the abscissa represents time (here, a total of 5 s), and the ordinate represents velocity (here, from -30 to $+30\,\mathrm{cm}\cdot\mathrm{s}^{-1}$).

Traditional processing computes the power at each frequency and subsequently the associated velocity, as outlined in Eq. 20.

Here the phase ϕ of the basebanded backscatter signal $a(t)e^{i\phi(t)}$ is digitized along the slow time scale (ms). Fourier analysis is used to determine the frequency or rate of change (ω_n) of the phase $\phi_n = \phi(t_n)$ More precisely, the so-called spectral power $P(\omega_n)$ at each frequency ω_n is computed by Fourier analysis. This quantity yields how much contribution to the power there is for a given rate of change or speed v_n. These two quantities $P(\omega_n)$ and v_n are plotted in the velocity graph in Fig. 16 (lower plot). The gray level for a given point in the graph is determined by $P(\omega_n)$, whereas v_n or ω_n and the time t determine the location of the pixel. The indicated cardiac cycle in Fig. 16 shows contributions from high velocities that yield high Doppler frequencies. At the end of this cycle the blood flow slows down and contributions to high frequencies diminish and formerly white pixels are now plotted in black. Other operations such as windowing of the phase signal ϕ_n are neglected here for simplicity.

$$\begin{aligned} f^*(t)e^{-i(-\omega_0)t} &= a(t)e^{i\phi(t)} \\ P(\omega_n) &= |\sum_n \phi_n T_n e^{i\omega_n n}|^2 \\ v_n(t) &= \left(\frac{\omega_n(t)}{\omega_0 T_{rep}}\right)\left(\frac{c}{2}\right)\cos\alpha \end{aligned} \tag{20}$$

Color Flow Doppler. Color flow Doppler allows the user to see a 2D map of flow in the current B-mode image. Instead of measuring flow only along a single scan line as in Pulse wave Doppler (PW Doppler), all lines inside a chosen region of interest (ROI) are fired repetitively (4–16 times) and analyzed for flow. Velocity resolution and frame rate are limited due to the large number of acoustic transmissions and the computational burden of analyzing the resulting received waveform data. In the same fashion as in B-mode, interleaved imaging can be used to counterbalance the reduction in frame rate caused by the necessary increase in (Doppler) scan lines (see Fig. 11).

In Fig. 17, one can see a regular B-mode image of a carotid artery, extending from the left to the right side of the image, parallel to the transducer face. The gray scale B-mode image is overlaid with Doppler information inside a user-defined color flow box, which is either a rectangle or a parallelogram slanted either to the left or right. For this specific example, a rhomboid with 20° rightward steering was used in order to measure the velocity of blood flow. One should remember that vessels parallel to the transducer do not yield a Doppler or phase shift and therefore can not be identified with Doppler methods. However, a 20° rightward steering provides enough angle deviation to measure flow

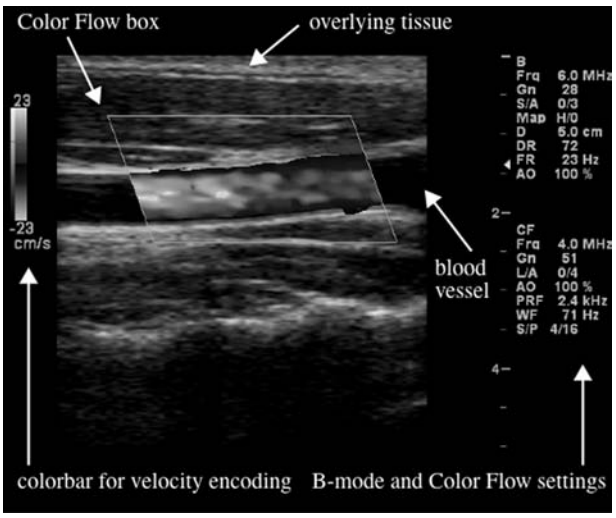


Figure 17. Color flow example showing blood flow in the carotid artery. The color data (here in gray scale) encodes magnitude and direction of flow. A color bar allows the quantitative conversion to actual speed, here a maximum of $\pm 23\,\mathrm{cm}\cdot\mathrm{s}^{-1}$.

and to display velocities everywhere in the chosen color flowbox. Ultrasonic waves are transmitted parallel to the slanted rhomboid.

Typically, blood flow is encoded from dark red to red to yellow when it is approaching the transducer, indicating low, moderate, or high velocities, respectively. Alternatively, it is colored in shades of blue to cyan when it is moving away at low to high speed. The maximum detectable velocity magnitude is directly related to the PRF used. A PRF of 2.4 kHz was chosen in the example shown, resulting in a maximum detectable velocity of $\pm 23\,\mathrm{cm}\cdot\mathrm{s}^{-1}$. Note that the PW Doppler example in Fig. 16 used a 4% greater PRF but allows a 30% greater velocity range. This is related to the greater burst length used in PW Doppler (15–20 cycles) compared to CF Doppler (2–6 cycles) and the subsequent data processing. The PW Doppler is designed for high velocity resolution and simultaneous temporal resolution, whereas CF Doppler is designed for great spatial resolution. Moreover, one should notice the very low backscatter level of blood relative to the surrounding tissue, which can be as much as 40 dB below that of soft tissue. The blood vessel in Fig. 17 appears black relative to the tissue on the proximal and distal side of it.

As mentioned above, CF Doppler demands not only more acoustic transmissions, but also more computing power to estimate the actual flow velocities from the measured acoustic backscatter. This imaging modality became practical when Kasai et al. (12) succeeded in designing a method (see Eq. 21) by which the mean Doppler frequency, that is, the mean velocity in every pixel, could be computed in real time by cross-correlations of the quadrature components, I and Q, of the analytic (complex) backscatter signal. Quadrature components I and Q are obtained by mixing the rf-signal at the hardware level with $\sin(\omega_0 t)$ and $\cos(\omega_0 t)$. This corresponds to the complex base banding given in Eq. 18, since $e^{-i\omega t} = \cos\omega t\text{-}i\,\sin\omega t$. It follows that I and Q are the real- and complex- valued parts of the

basebanded signal, $a(t)e^{i\phi(t)} = I + iQ$.

$$\phi = \frac{\Sigma_{n=1}^{N-1} Q[n]I[N+n] - I[n]Q[N+n]}{i\Sigma_{n=1}^{N-1} Q^2[n] + I^2[n]}$$
$$\bar{v} = \frac{c}{2\omega_0 T_{rep}} \arctan\phi \tag{21}$$

Power Doppler

The previous two methods of flow quantification suffer from a lack of good flow detection. In perfusion studies, it is often necessary to detect very small amounts of flow volume travelling through capillaries at velocities of the order of $1\,\text{mm}\cdot\text{s}^{-1}$, which is $\sim$0.1% of the speed observed in the carotid artery. In order to overcome the poor sensitivity of PW and CF Doppler, power Doppler displays the integral of the power spectrum $P(w)$, shown in Eq. 20. Typically, the integration value is also averaged over a very long period of time (several heart beats). Averaging at least one cardiac cycle results in a nearly constant value for flow. Physically, this value represents the amount of blood flowing, but not the velocity, since it is the integral of all detected velocities.

First, imaging capillary flow yet remains difficult, even in power Doppler mode, partially because blood cells do not scatter much of the transmitted acoustic signal. Small blood vessels, such as the capillaries, provide small fractional blood volume, which further decreases the total backscattered signal. Second, at $1\,\text{mm}\cdot\text{s}^{-1}$, flow velocities in capillary beds are difficult to differentiate from static soft tissue background at zero frequency shift, without being suppressed by the wall filter. This filter is used to prevent tissue motion from being incorrectly ascribed as real flow. In Fig. 17 flow that causes <71 Hz frequency shift per Doppler firing is filtered out of the Doppler data to eliminate flow speeds of $<x\,\text{cm}\cdot\text{s}^{-1}$.

The ultrasound machine transmits Doppler pulses every 0.42 ms (recipocal of 2.4 kHz). Backscatter will contain phase shifts between $-\pi$ and $+\pi$, which corresponds to $+23$ and $-23\,\text{cm}\cdot\text{s}^{-1}$ flow speed.

The acoustic wavelength of the color Doppler (CF panel on the right side in Fig. 17) transmits equals 0.385 mm in tissue ($c = 1540\ \text{m}\cdot\text{s}^{-1}$, $Frq = 4\,\text{MHz}$, $\lambda = c/Frq$) and each firing is separated 0.42 ms (recipocal of PRF = 2.4 kHz). The Doppler electronics measures phase shift, therefore it can not measure more than a shift of 2π or $\pm\pi$. Two-pi corresponds to λ or $\pm\pi$ to $\pm\lambda/2$, hence the maximum detectable speed of $v = s/t = 45.8\,\text{cm}\cdot\text{s}^{-1}$, with $s = \lambda/2$ and $t = 1/\text{PRF}$. However this value does not match the displayed $23\,\text{cm}\cdot\text{s}^{-1}$. Doppler pulses work in pulse-echo mode, in which any displacement dx results in a time shift of 2-times dx/c. Finally the maximum detectable flow speed is given by equation 22. For a given wallfilter (WF) the minimal detectable flow is given by the ratio of the PRF to wall filter times the maximum flow, that is, $(2.4\,\text{kHz}/71\,\text{Hz})^{*}23\,\text{cm}\cdot\text{s}^{-1} = 0.68\,\text{cm}\cdot\text{s}^{-1}$.

$$v = \frac{\delta s}{\delta t} = \frac{0.5\cdot c/(2f)}{1/\text{PRF}} = \frac{0.5\cdot 1540/(2\times 3.75\times 10^6)}{1/2.4\times 10^3}$$
$$= 23\,\text{cms}^{-1} \tag{22}$$

3D Imaging

Current ultrasound images are naturally 2D because ultrasound imaging arrays are only 1D. One-dimensional arrays are still predominant in the market. Even so, great efforts in the ultrasound community are pushing ultrasonic imaging toward 2D arrays. The transition from 1D to 2D is especially apparent in the naming scheme of current arrays:

- 1D: Is the classic linear or focused array, which has one row of elements that allows focusing and steering in the lateral imaging plane. The elevational focus is constant due to a fixed elevational curvature of each element.
- 1.25D: Extra rows of elements on either side of the main row allow changes in the elevational aperture, but there is no electronic elevational focusing, nor steering.
- 1.5D: This class of arrays has a 2D set of elements, where the elevational elements are connected symmetrically to the center row. This array can focus in the elevational direction but not steer.
- 1.75D: A 2D set of individually driven elements is available for this type of array, but the number of elements in the elevational direction is much less than in the lateral direction. Elevational focusing is possible, but only limited elevational steering is available.
- 2D: Elevational and lateral directions should be equivalent and indistinguishable for a true 2D array. Full apodization, steering, and focusing is possible in 3D. Currently there are some commercial systems that use 2D arrays particularly in cardiac imaging.

Hardware and software implementations allow the 3D reconstruction of a scanned volume even when using 1D arrays. Sophisticated 3D hardware position sensors allow ultrasound scanners to register the position and orientation of a 1D array in 3D space. Therefore, any acquired image in a set of many can be aligned with others in the set to render a 3D volume (see Fig. 18). However, these hardware additions are costly and can be inconvenient. Moreover, they might show limitations due to interference with electromagnetic fields or nearby metallic objects. Software solutions use correlations between adjacent image frames to determine the transducer translation or rotation. Figure 18 rudimentarily illustrates how individual frames taken in freehand fashion are "stitched" together to form a 3D volume, which can be rendered in various ways. Figure 19 shows an anatomical example of the bifurcation of the ascending carotid aorta rendered as a 3D volume. Some implementations on clinical scanners, however, already use the 4D nomenclature by adding time as the fourth dimension.

CONTRAST IMAGING

As in every other clinical imaging modality, agent based imaging enhancements are available for ultrasonic

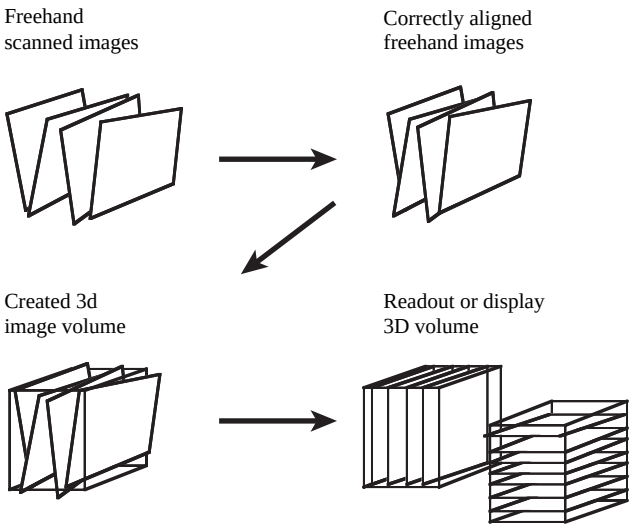


Figure 18. Illustration of 3D image reconstruction. Sets of spatial misaligned images are stacked according to their spatial position and orientation and used to fill a 3D image volume. Afterward this volume can be read out in any slice plane direction or even rendered as a 3D volume, such as shown in Fig. 19.

imaging. However, a limited number of clinical applications is approved by the Food and Drug Administration (FDA). As of 2004, the only FDA approved application for ultrasound contrast agents is for cardiac procedures, and more precisely, for outlining the border of the heart chamber. Other countries or regions such as Europe and Japan have a variety of agents approved. However, considerable research has been performed on ultrasound contrast agents, and it is likely that more FDA approvals will follow in the future. Contrast agents are used to enhance ultrasound image quality; therefore, imaging techniques implemented in current ultrasound scanners will be described.

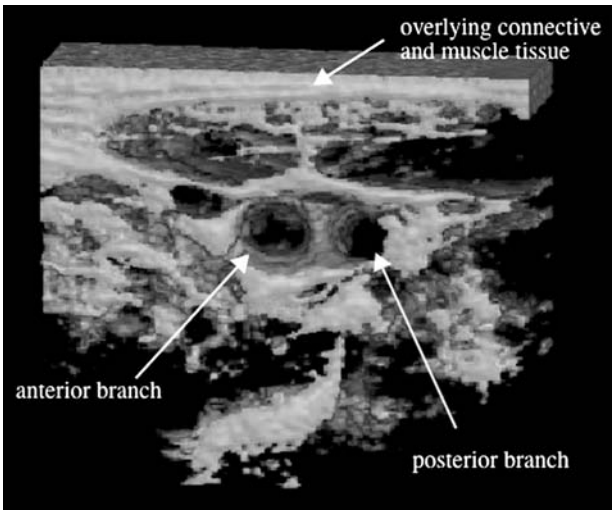


Figure 19. Three-dimensional reconstruction of the ascending carotid artery. This imaging mode uses 3D correlations and compounding to align individual images as the imaging array is swept across the vessel.

Moreover, the physics of contrast agents as well as their optimal clinical use will be discussed.

Clinical Background. Every year 135 million (Source: Amersham Health Inc. owned by GE Healthcare.) ultrasound scans are preformed in American hospitals. Only 0.5% of these procedures actually use contrast-enhancing products. For example, better diagnosis of myocardial perfusion and better visualization of fine capillary-level vasculature will be possible when ultrasound contrast agents are certified by the FDA. Ultrasound is a relatively inexpensive imaging technique, and better diagnostic information can obtained.

New contrast-agent-based ultrasound imaging modes include: Harmonic imaging, Pulse inversion (with harmonic or power mode), Microvascular imaging, Flash contrast imaging, as well as Agent detection imaging.

Enhancing Contrast. A major duty of ultrasound contrast agents is the improvement of ultrasound based image acquisition. The definition of contrast is given in Eq. 23, where I_1 and I_2 are the echo intensities before and after contrast administration, respectively. Even though this is a very simple formula, the mechanism for contrast improvement can be rather complicated.

$$\Lambda = \frac{I_2 - I_1}{I_1} \tag{23}$$

The key for contrast improvement for ultrasonic contrast agents lies in the physical principles of sound transmission, reception, and the nonlinear characteristics of bubbles themselves. For example, an increase in backscatter *amplitude* in the presence of contrast agents relative to the average human tissue backscatter level would improve the overall image. Furthermore, the creation of acoustic *frequencies* that occur in the backscatter signal of bubbles but not in the transmit signal nor in the backscatter of tissue, would provide a mechanism by which bubbles can improve the overall image. An important fact to keep in mind is that ultrasound contrast agents do not enhance the visibility of human tissue nor of blood, but the bubbles themselves can be visualized better than tissue or blood. Nevertheless, imaging perfusion of tissue or measurement of the amount of blood flowing through a vessel can be greatly improved by the usage of ultrasound contrast agents.

Modern Agents. Modern agents are not just gas bubbles. A sophisticated shell coating is used to prevent coalescence and reduce diffusion of the interior gas into the surrounding medium. This shell can be made of serum albumin. Lipids are also used as stabilizing agents. The gases filling the interior of the shell are chosen for low diffusion rates from the bubble into the blood stream as well as because of their low solubility in blood. Table 3 lists commercially available contrast agents. Currently FDA approved contrast agents include Imavist by Alliance Pharma/ Photogen, Definity by Bristol-Myers Squibb Medical Imaging Inc., Albunex by Molecular Biosystems, and Optison by GE/Amersham.

Table 3. Modern Ultrasound Contrast Agents[a]

Manufacturer	Agent Name	Interior Gas	Shell Material
Acusphere	Al-700	Perfluorocarbon	Copolymers
Alliance Pharma. / Photogen	Imavist	Perfluorohexane, air	Surfactant
Bracco	SonoVue	Sulfur hexafluoride	Phospholipid
Bristol-Myers Squibb Medical Imaging, Inc.	Definity	Perfluoropropane	Lipid bilayer
	MRX-815-stroke	Perfluoropropane	Lipid bilayer
Molecular Biosys.	Albunex	Air	Albumin
	Oralex	Air	Dextrose
GE/ Amersham	Optison	Perflutren	Albumin
Nycomed Imaging AS	Sonazoid	Perfluorobutane	Lipid
Schering AG	Echovist	Air	Galactose
	Levovist	Air	Lipid layer
Sonus Pharma.	EchoGen	Dodecafluoropentane	Albumin
	SonoGen		Charged surfactant

[a]See Refs. 3,7, and 9.

Acoustic Bubble Response. Ultrasound contrast agents can be viewed as systems known as harmonic oscillators, with a given amplitude, phase, and frequency. Min-naert has derived Eq. 24, which gives the resonance frequency of a gas bubble as a function of its size. For example, a 3 μm radius (R_0) air bubble (adiabatic coefficient κ) in water (mass density ρ_L) under atmospheric pressure P_0 has a resonance frequency f of 1.1 MHz (Table 3). This is a very fortunate relationship since capillaries of the human circulatory system are as small as 8 μm in diameter and typical clinical frequencies used are 1–10 MHz.

$$f(R_0) = \frac{1}{2\pi}\frac{1}{R_0}\sqrt{\frac{3\kappa P_0}{\rho_L}} \tag{24}$$

The amount of acoustic scattering of the bubble surface (scattering cross-section σ_S) is described by the Rayleigh equation. This equation is used for scatterers that are small (μm) relative to the acoustical wavelength used (mm). For gas bubbles in water, the Rayleigh equation can be written with a series of mathematical terms for corresponding physical oscillation modes.

$$\sigma_S = 4\pi a^2 (ka)^4 \left[\left(\frac{\kappa - \kappa_0}{3\kappa}\right)^2 + \frac{1}{3}\left(\frac{\rho - \rho_0}{2\rho + \rho_0}\right)^2\right] \tag{25}$$

The first term represents a monopole type bubble oscillation, whereas the second term describes a dipole term. One can see that the monopole term dominates the scattering due to the large compressibility (κ) difference between water and air–gas. Density differences (ρ) between water and air/gas are large too, however, the monopole term dominates the acoustic scattering (see Table 4).

Mathematical and Physical Modeling. The equation of motion of a bubble can be readily derived from an energy balance of kinetic (T) and potential energies (U) using the Lagrange formalism ($L = T - U$). It should be mentioned that the momentary inertial mass of the bubble as an oscillator does change with time. This is a major reason why ultrasound contrast agents are nonlinear systems, as will be shown shortly. One of the first equations describing the motion of a gas bubble excited by ultrasound was derived by Rayleigh–Plesset and is given in Eq. 26. This formula is derived under the assumption that the interior gas follows the ideal gas law, and other forces acting on the bubbles are comprised of the internal vapor pressure p_d, the external Laplace pressure caused by the surface tension σ, the viscosity η_L of the surrounding host medium (water), the mass density ρ_L of the water, as well as the static p_0 and acoustic $p_\infty(t)$ pressures.

$$R(t) = \frac{1}{R(t)}\left(-\frac{3}{2}\left(\frac{\partial}{\partial t}R(t)\right)^2 + \frac{1}{\rho_L}\left(\left(p_0 + \frac{2\sigma}{R_0} - p_d\right)\left(\frac{R_0}{R(t)}\right)^{3\kappa} + \cdots + p_d - \frac{2\sigma}{R(t)} - \frac{4\eta_L \frac{\partial}{\partial t}R(t)}{R(t)} - p_0 - p_\infty(t)\right)\right) \tag{26}$$

Figure 20 shows a simulation of the bubble response to a short tone burst excitation. Transmitted acoustic pressures of 1 and 50 kPa were simulated. Graph (a) shows a 1.1 MHz and 50 kPa pressure waveform as transmitted by a simulated ultrasound transducer. Graph (c) shows the subsequent radial oscillations of the simulated bubble (resting radius of 3 μm). Graph (b) shows the spectral response of these oscillations for a sound pressure of 1 kPa. The bubble oscillates mostly at the driving frequency of 1.1 MHz. An increase in sound pressure amplitude (i.e., 50 kPa) reveals the nonlinear nature of gas bubbles. In panel (d), in addition to 1.1 MHz one can also see higher harmonics of 2.2 MHz, 3.3 MHz, and so on.

Table 4. Scattering Coefficients for Monopole and Dipole Terms in Eq. 25 of a Water or Air Filled Sphere Under Water

Material	Bulk Modulus κ, MPa	Density ρ, kg · m^{-3}	Monopole Magnitude	Dipole Magnitude
Water	2250	1000	0	0
Air	0.14	1.14	2.9×10^7	0.33

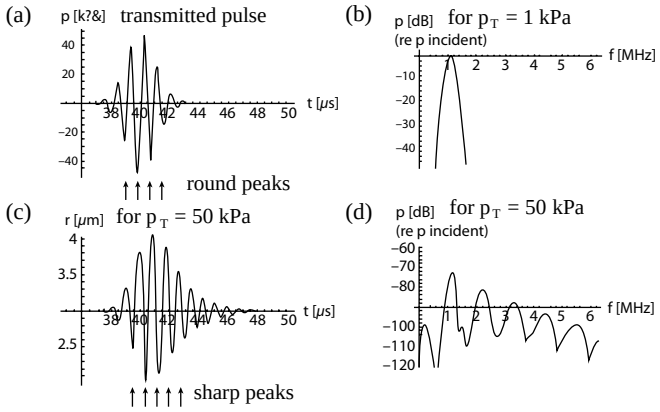


Figure 20. For higher transmit sound pressures (p_T) gas bubbles exhibit nonlinearities at similar magnitudes as the fundamental frequency [(b) vs. (d)]. Higher harmonics increase in their relative amplitude with sound pressure. Radials oscillations of the gas bubble contrast agent shows the nonlinear response during the compressional phase (arrows at small radii). Harmonic imaging takes advantage of the strong nonlinear character of contrast agent, which can be stronger than tissue depending on the number density of the bubbles.

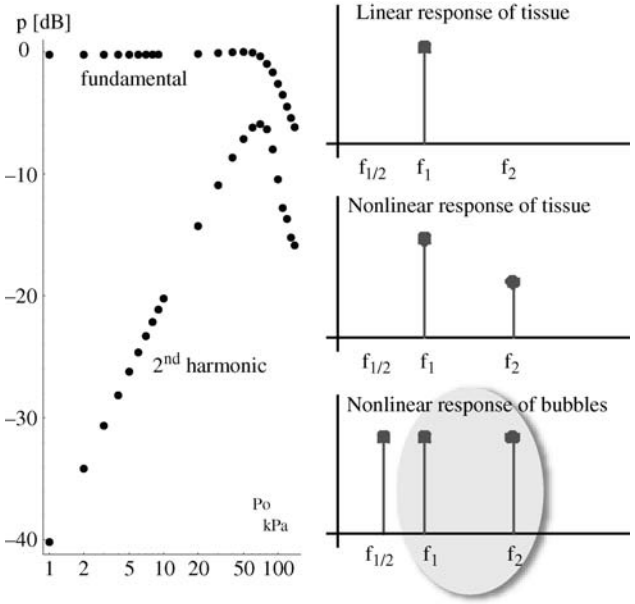


Figure 21. Higher harmonics increase in their relative amplitude with sound pressure (left), P_{harmonic}/P_0. Harmonic imaging takes advantage of the strong nonlinear character of contrast agent, stronger than tissue, depending on the number density of the bubbles. Subharmonic emissions $f_{1/2}$ are also characteristic for bubbles.

Modern ultrasound contrast agents cannot be modeled using the free gas bubble Rayleigh–Plesset model. Elastic layer-based models presented by de Jong (5), Church (4), and Hoff (10) contain additional parameters, such as the mass density of shell material, a second surface tension term, a second viscosity term, or the elastic modulus of the shell material.

Imaging Modes

The following sections will cover imaging modes that rely on nonlinear backscatter, either originating from body tissue or due to nonlinear reflections from contrast agents. Each mode will be theoretically described and illustrated with examples.

Harmonic Imaging

Nonlinear tissue and contrast agent backscatter is the basis for harmonic imaging. Human tissue, as well as ultrasound contrast agent, can be driven in a nonlinear fashion such that the backscattered signal contains not only the original frequency f_0, but also $2x\,f_0$, $3x\,f_0$, and so on (see Fig. 20d). Higher harmonics increase in their relative amplitude with sound pressure. Harmonic contrast imaging takes advantage of the strong nonlinear character of contrast agent, which is stronger than tissue. As a result, nonlinear backscatter from tissue will be smaller in amplitude than that of contrast agent and the vascular system will be visible over tissue. Subharmonic emissions are also characteristic for bubbles ($1/2x\ f_0$, $1/3x\ f_0$, etc.) and can be used to distinguish bubbles from tissue.

Figure 21 is comprised of a direct gas bubble simulation and an illustration of harmonic bubble behavior. As can be seen from the simulation results, harmonic scattering of gas bubbles increases with increasing incident sound pressure. Specifically, the normalized amplitudes of the fundamental and the second harmonic frequency are shown here ($p_{\text{fundamental}}/p_0$ and p_{harmonic}/p_0). Increasing the incident sound pressure causes a proportional increase in the fundamental response, while the contribution of the harmonics increases more strongly. At 50–100 kPa, the backscatter amplitude of the second harmonics peaks and for higher pressures more and more energy is distributed over a wide range of frequencies, from sub-harmonics f_0/n to higher harmonics n^*f_0. Because the model as described in Eq. 26 does not take into account any losses, such as those induced by radiation or viscous damping, backscatter predictions for large excitation pressures will not be accurate. However, at modest amplitudes tissue scatters in a linear fashion, gas bubbles contribute harmonic signals, and the amplitude ration between fundamental and harmonic components can approach 1 for bubbles driven at sufficiently large amplitudes (illustration in lower right panel of Fig. 21).

Figure 20 shows the oscillation of a gas bubble in a large amplitude acoustic field. Panel (a) shows the excitation sound pressure waveform of 50 kPa amplitude. Radial excursions as well as the sound pressure spectrum at a distance of 5 cm from the bubble are plotted on the bottom panels (c) and (d), respectively. Most notably are the sharp peaks for bubble radii smaller than the initial bubble radius (see arrows in Fig. 20c). At positive incident sound pressures, the bubble is compressed and exhibits a large internal pressure. At the same time, the bubble's resonance frequency changes. This change in resonance is the reason

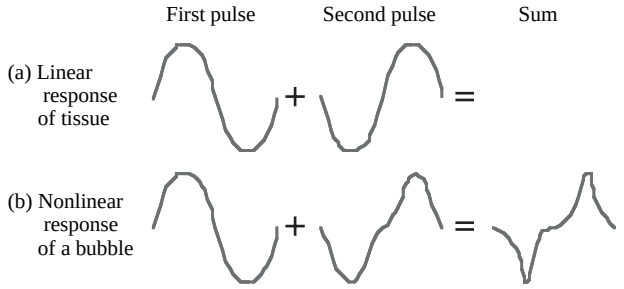


Figure 22. Tissue and contrast agent can yield different backscatter when exposed to an acoustic field of moderate intensity (typically of an mechanical index (MI) of less then 0.2). (a) Tissue responds in a linear fashion, that is, it yields a 0 and a 180° phase-shifted signal. The sum of both waves is zero. (b) For ultrasound contrast agent, the backscatter will include higher harmonics. A 180° phase shift in the transmit frequency f_0 will result in a 360° phase shift in the second harmonic $2\,f_0$. However, a 360° phase shift will result in the original signal. Therefore, the sum of a regular pulse and a inverted pulse will cancel for f_0 frequency components but not for $2\,f_0$ components.

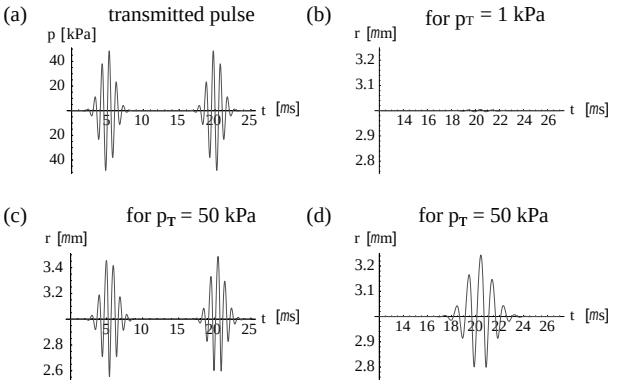


Figure 23. Two transmitted pressure waveform pulses with 180° phase shift with respect to each other, are displayed at time equal 5 and 20 μs in panel (a). This concept of inverted pulses is illustrated in Fig. 22(a). For 1 kPa sound pressure amplitude the sum of the oscillations of these two pulses is near zero (b). However, for a sound pressure of 50 kPa, the radial excursions of the contrast agent bubbles is nonlinear (as shown in Fig. 20) and the sum of the 0 and 180° signal is almost as great (3.2 μm in d) as the oscillation itself (3.4 μm in c). This technique is used in pulse inversion contrast agent detection scheme.

for their highly nonlinear character. Panel (b) shows the scattered sound pressure spectrum for a low incident pressure of 1 kPa. No harmonic contributions can be seen within 40 dB of the pressure amplitude at the fundamental frequency.

Pulse Inversion Mode

As its name suggests, pulse inversion uses inverted pulses to gain contrast in the image. Figure 22 illustrates how this imaging mode works. As opposed to regular B-mode imaging, this mode requires *two* transmissions per image line. The first transmission does not differ from regular B-mode. However, the second transmission is 180° phase shifted, that is, inverted with respect to the first pulse transmitted. In case of tissue the backscatter for both pulses will mostly remain the same in magnitude. However, contrast agents respond with harmonic contributions which differ for 0 and 180° pulses. For very low acoustic pressures, this difference might not be distinguishable from the linear backscatter, but a mechanical index of 0.2 or less is sufficient to perform pulse inversion imaging. The mechanical index is a measure for the sound pressure amplitude and will be defined in the bioeffects section.

High sound pressure amplitude pulses can have two effects. First, they cause nonlinear tissue backscatter, thus reducing the contrast between tissue and contrast agent in the vascularity. Second, they destroy the contrast agent. One has to keep in mind that contrast agents are comprised of encapsulated gas bubbles. Therefore the bubble can shatter and the contained gas can subsequently dissolve into the blood pool. This shattering is possibly a source for bioeffects of ultrasonic imaging and will be discussed in a later section.

Theoretical modelling of the oscillatory behavior of gas bubbles using the above-introduced model directly shows the nonlinear response of gas bubbles. Figure 23a shows the two-pulse sequence required for pulse inversion imaging. For transmit pressures of 1 kPa, the difference in bubble response to 0 and 180° is very small. However, on the same scale, the response to a 50 kPa excitation yields two signals whose fundamentals f_0 cancel, but whose second harmonic contributions $2\,f_0$ add in phase (see Figs. 22 and 23 d).

A major downside of pulse inversion imaging is its sensitivity to motion. Spatial shifts cannot be distinguished from changes in back-scatter due to changes in transmit phase. A clever work-around to this problem is illustrated in Fig. 24, where a three-pulse transmit of varying phase is shown. Situations in which motion is anticipated or intestinal peristalsis or chest wall excursions during breathing. The latter is on the order of $2\,\mathrm{cm}\cdot\mathrm{s}^{-1}$, or for a 1 kHz PRF, 20 μm per firing. A signal of 1 MHz center frequency will experience a phase shift of 4.8° when spatially translated 20 μm. It is assumed that the time duration in which the three pulses are fired is small compared to that of the body motion. If so, then the motion can be approximated as being linear on the time scale of the firings. The first firing is transmitted with zero phase, the second one with $5 + 180°$, the third one with 10° phase shift. Linear response of tissue will result in backscatter signals of 0°, 5°+180°, and 10° phase shifted signals. Averaging the first and third backscatter signal will yield an average of 5° phase shift. Adding this average signal to the second backscatter signal will yield a value of zero for linear tissue, even while it is in motion. Bubbles, however, will yield a similar, nonzero, signal whether in motion or not.

Coded Excitation and Coded Harmonic Excitation. *Coded excitation* is a way to overcome poor signal to noise ratios (SNR) in ultrasonic imaging. It was already used in radar imaging for the same purpose, before its introduction to medical ultrasound. The most simple solution for poor

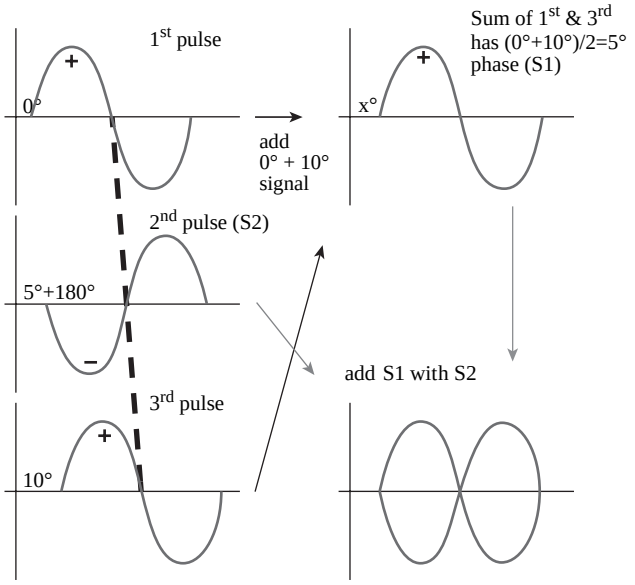


Figure 24. A three pulse sequence with a linear shift in transmit phase (0°, 5°+180°, and 10°) is used to suppress tissue signal, even when the tissue is in motion. The 0 and 10° backscatter signals are averaged to yield a 5° signal from tissue (SI). The 5° phase-shifted and inverted transmit signal (S2) is then added to the averaged signal SI. For moving tissue the result will be zero. If the tissue motion is linear during the time span of the three transmit pulses (∼1 ms), then pulse will be separated by an offset of x°. This offset is in addition to the initial 5° offset between first & second and second and third pulse. Therefore tissue will yield zero for S1+S2. Ultrasound contrast agents, however will yield S1≠S2 as shown in Figs. 22 and 23.

signal/noise ratios is to increase the amplitude of transmitted sound. However, the FDA regulates sound pressure amplitudes because of the likelihood of acoustic cavitation in the presence of large sound pressure amplitudes. Moreover, the number of acoustic transmissions per unit time (pulse repetition interval, PRI) as well as the length of individual pulses (burst length) are regulated by means of the overall deposited energy that will eventually result in tissue heating. Both effects will be discussed in the Bioeffects section below. In addition to regulatory and safety concerns, increasing the number of cycles in a traditional transmit pulse also reduces the axial resolution.

However, improvements in signal to noise ratios without sacrificing axial resolution can be achieved despite regulations on pressure amplitudes and burst lengths. Transmission of specially designed and unique signals can significantly improve their detectability. This concept is called Coded harmonic excitation.

Various code types shall be discussed here to illustrate how Coded excitation works. Figure 25 illustrates two codes. On the left side is the most simple code, a so-called pulse train (or uncoded tone burst), that is, a series of pulses or sinusoids. When transmitting four cycles at a certain frequency one can use a frequency filter that is sensitive at the transmit frequency, but only for signals that are four cycles long. A more complicated type of coded excitation is the code shown on the right side of Fig. 25.

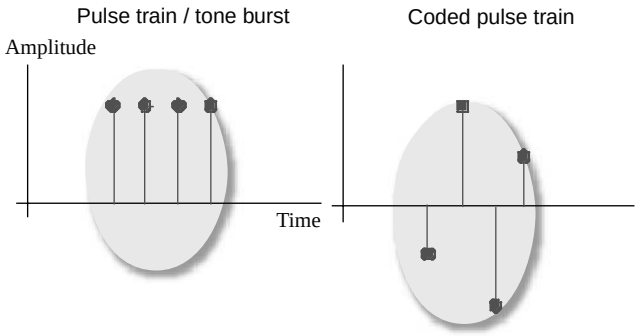


Figure 25. Some signal to noise problems can be overcome by the use of coded excitation. Instead of a single pulse, a pulse train (commonly also known as a tone burst) or coded pulse train is transmitted for better backscatter detection. See Fig. 26 for temporal /spatial resolution.

This signal also transmits four cycles, but now the four cycles all differ in sign as well as in amplitude. Both features make this code more unique and therefore more detectable when ambient noise lowers SNR.

Above mentioned frequency filters are explained in Figs. 26 and 27. The first column in Fig. 26 illustrates the transmit pressure waveforms and the second column shows the receive filter used. A mathematical technique termed convolution is used to match the transmitted waveform with the anticipated receive signal by means of an appropriately designed filter. The average reader might not have adequate signal processing background to be familiar with this concept. Therefore, Fig. 27 will be used to explain Coded Excitation for the example of Golay code

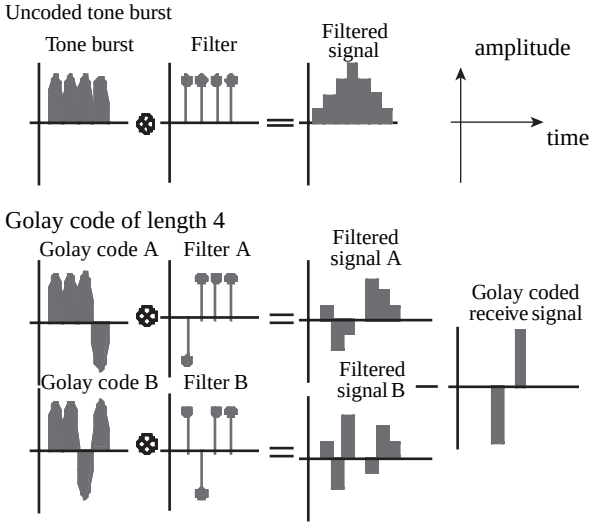


Figure 26. Coded excitation is implemented on modern ultrasound scanners to overcome signal loss and low SNR. A simple solution is the increase in burst length. However, this simultaneously results in a decrease in axial/temporal resolution. Specially coded waveforms are designed for an increase in SNR by reasonably maintaining spatial resolution. Golay codes, for example, can be designed for a 10-fold increase in SNR by loosing only a factor of two in temporal resolution.

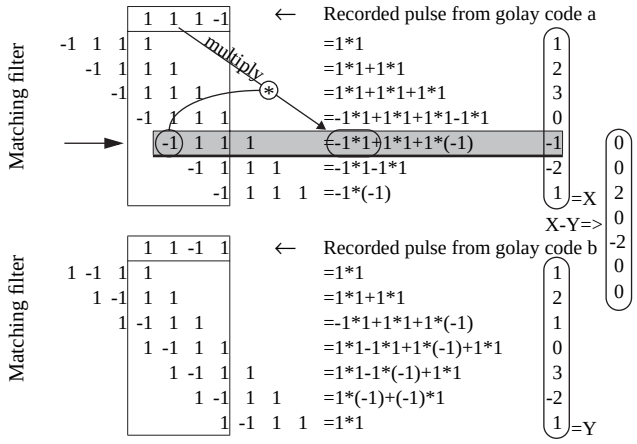


Figure 27. The mathematical process of convolution is described in this figure. For this purpose the reader should refer to the Golay code of length 4 given in Fig. 26 and to the text section on Coded excitation for a detailed description.

of length 4 (see Fig. 26). This code consists of two transmit pulse packages. The first and second packages are four pulse sequences of [+1, +1, +1, −1] and [+1, +1, −1, +1], respectively. Such pulses are shown on the top row of each box in Fig. 27. The pulses are transmitted into the body and the resulting echoes, which should show the same pattern, are recorded. Matching filters are shown below the pulse packages and have been replicated at a total number of 7 positions. Each filter replication has an overlap with the transmit pulse ranging from 1 to 4 digits. For a given position, numbers in the same column are multiplied and then the resulting values of the overlapping columns are added.

For clarification the highlighted case in the top system for position 5 (arrow) is explained in detail. The filter [−1, 1, 1, 1] and the recorded pulse [1, 1, 1, −1] overlap for three digits and the computational method of convolution is equivalent to computing the inner product of two vectors: $[1, 1, -1]^*[-1, 1, 1] = 1^*(-1)+1^*1+(-1)^*1$, which sum to −1. This value is displayed in the column on the right. This convolution procedure is done for both transmits. Subtracting the two resulting 7 digit vectors $X = [1, 2, 3, 0, -1, -2, 1]$ and $Y = [1, 2, 1, 0, 3, -2, 1]$ will yield [0, 0, 2, 0, −2, 0, 0], a receive vector (called Golay-coded receive signal in Fig. 26) with good temporal/spatial resolution. In the given case it is essentially three time steps long: [−2, 0, 2]. The uncoded tone burst is 7 time steps long (see top of Fig. 26).

Back to Fig. 26, the result of the convolution is shown in the third column. For the uncoded tone burst one can see a triangular shaped signal that peaks where the receive filter perfectly matches the incoming pressure wave. The temporal duration of the convolution increases as the duration of the tone burst increases. This unwanted effect of loss of spatial resolution is overcome by the use of more sophisticated codes. Golay transmit codes and receive filters, for example, can mostly maintain spatial resolution while improving SNR. Two differing transmit codes are used and the resulting convolutions are subtracted from each other, similar to pulse inversion imaging.

Coded excitation in the form of a Chirp-type pulse train yields a signal to noise gain of $\sqrt{\Delta F\, T}$; for example, a signal with a 1 MHz frequency sweep and $T = 16\,\mu$s pulse duration will yield a fourfold increase in SNR, which corresponds to 12 dB. Chirps are tone burst of continuously increasing transmit frequencies within the pulse train. Continuously in frequency decreasing pulse trains are called pirchs, which almost spells chirp backward. General Barker and Golay codes perform well too. For a standard Barker sequence the side lobe amplitudes of the transmit/receive signal are −22 dB down relative to the amplitude main lobe. A Golay code with 16 pulses per burst yields a ten-fold increase in SNR. Here, only a length-4 sequence is shown for clarity (see Fig. 26).

More detailed mathematical background, as well as clinical feasibility can be found in publications by Nowicki (19) and Misaridis (20), for example.

BIOEFFECTS

Ultrasound related bioeffects are centered on two main categories: cavitation and heating. Acoustic output (and therefore to some degree acoustic cavitation) is regulated by the FDA through the implementation of the mechanical index (MI), as well as the temperature index (TI) in the so-called Output Display Standard (Standard for Real Time Display of Thermal and Mechanical Acoustic Output Indices on Diagnostic Ultrasound Equipment (1992), published by the National Electrical Manufacturers Association as UD-3). The output display standard is used to obtain approval by the FDA for medical use of ultrasound scanners.

Cavitation is a well-studied effect and manifests itself, for example, by the erosion of ship propellers and for therapeutic purposes in lithotripsy. Hydrodynamically generated microscopic gas bubbles imploding in close proximity to the metallic surface of the propellers punch small holes into the metal. This occurs due to the existence of water jets generated by asymmetrically imploding gas bubbles.

To date, there are only very limited clinical reports on the occurrence of bioeffects in diagnostic ultrasound. Nevertheless, *in vitro* cell cultures and animal models are used to study bioeffects at acoustic parameters beyond the limits for diagnostic imaging. No limits for single parameters (e.g., pressure amplitude, wave frequency, or pulse length) were identified *per se due* to the complicated biological end points (21). However, two fundamental quantities, MI and TI, were introduced as parameters by which to judge the probability of bioeffects (11,21). Both quantities gain in presence, especially with the MI being used by the FDA to limit the acoustic output of scanners.

$$\mathrm{MI} = \frac{p[\mathrm{MPa}]}{\sqrt{f[\mathrm{MHz}]}} \quad \mathrm{TI} = \frac{W\, p[\mathrm{Nm \cdot s^{-1}}]}{W_1^o[\mathrm{Nm \cdot s^{-1}}]} \tag{27}$$

Pressure p and frequency f are quantities that can be directly measured experimentally. Power W_p, however, is a derived quantity and is related to the pressure waveform as

shown in Eq. 28 (22).

$$PII = \frac{\int_{t_1}^{t_2} v_h(V(t))^2 dt}{10^4 \rho c M_L(f_c)^2} \left[\mathrm{J \cdot cm^{-2}}\right]$$
$$I_{\mathrm{SPTA}} = \mathrm{PII}(j, v, z_{mj\mathrm{PII}}) \times \mathrm{PRF}\,[\mathrm{W \cdot cm^{-2}}] \quad (28)$$

The pulse intensity integral *PII* is computed over a truncated duration of the acoustic wave, namely, from t_1 to t_2. These times correspond to the 10 and 90% indexes of the accumulated total energy in the tone burst and can in most circumstances be regarded as the burst length. When calibrating an acoustic transmitter, one often measures an electric waveform produced by a underwater microphone, called a hydrophone. A conversion function v_h, is then used to determine the actual pressure of the measured electric waveform $V(t)$. M_L is the hydrophone's frequency dependence. Together, v_h, and M_L allow the complete conversion of the electrical hydrophone signal to the pressure of the measured wave. Mass density and speed of sound of the medium in which the wave travels are labeled as ρ and c. Even when the burst length is taken into account, one does not know the total transmitted power without the PRF. Spatial peak and temporal average intensities (I_{SPTA}) are computed with knowledge of the PRF, yielding units of watts per centimeter squared ($W \cdot \mathrm{cm^{-2}}$). For computing the TI, however, one needs to multiply that value with the geometric cross-section of the acoustic beam. In reality, the computation of the TI is more complicated and takes into account more parameters such as the chosen exposure conditions, as well as nominal perfusion parameters. In clinical use, perfusion parameters are estimated from user selected target tissue types.

An additional step toward reducing the possibility of bioeffects is the usage of the principle of ALARA (as low as reasonably achievable), that is, the reduction of acoustic output and reduction of exposure time to the lowest reasonable minimum. Typical diagnostic acoustic procedures operate at or below the FDA limit for $I_{\mathrm{SPTA.3}}$ of $720\,\mathrm{mW \cdot cm^{-2}}$ (see Table 5 and (23)). Currently, the FDA regulates output limits via Tracks 1 and 3 of 510(k). Track 1 uses the I_{SPTA} levels shown in Table 5, whereas Track 3 allows device manufacturers to increase their output to a general $720\,\mathrm{mW \cdot cm^{-2}}$ if their device provides user feedback via MI and TI display standards (24).

Biological effects of ultrasound can be investigated after the acoustic output of a sound source is quantified and qualified as described in the previous paragraphs. *In vitro* cell studies are typically the first method for research on the biological effects of ultrasound. Such studies can be performed in a highly controlled environment and are therefore repeatable. This is not necessarily the case for patient studies where many factors cause unavoidable variabilities.

It was found in a study on anesthetized rats that the interaction of the incident sound field with ultrasound contrast agent can cause pete-chial hemorrhages (punctuate sites of bleeding from blood vessels) in heart tissue. In a study by Li et al. (13,14) ultrasound contrast agents were injected in a similar fashion as is used in clinical contrast echocardiography. Heart tissue was observed in real time using a phased array ultrasound scan head (1.7 MHz transmit frequency) operating in a harmonic mode native to the clinical scanner used. Postmortem heart tissue analysis showed that bleedings scaled monotonically, proportionally to the square of the peak rarefactional pressure ($\mathrm{P_{rare}}$) amplitude of the sound field. Pressure amplitudes of 0.6–1.8 MPa P_{rare} were used in that study. This pressure range corresponds to MIs between 0.5 and 1.5. Real-time ultrasonic imaging showed that premature ventricular contractions (PVC) were triggered by ultrasound in the presence of contrast agents. However, no significant PVC were observed for MIs of 0.5 and 1.0, but up to 40 PVCs were observed for a 3 min exposure at a MI of 1.9, the maximum allowed by the FDA.

Currently, there are very limited clinical reports on unanticipated bioeffects of ultrasound. Intended effects exist since ultrasound can be used to treat kidney stones and during such procedures it is likely that tissue bleeding occurs. There have been reports of effects associated with diagnostics, but the review of these reports has yielded no establishment of a causal relationship with ultrasound exposure. However, a study by van der Wouw in healthy male volunteers concludes:"Imaging of contrast agents with high acoustic pressures can cause PVCs if end-systolic triggering is used. This effect is related to both the dose of contrast agent and acoustic pressure. It does not occur during end-diastolic triggered imaging. Precautionary measures would include using lower MIs or end-diastolic triggering" (25).

It has been seen that the level of bioeffects varied with the ultrasound contrast agent used. This is probably due to different shell materials and internal gases, which cause the bubbles to oscillate at different amplitudes. If agent rupture occurs, cavitation type damage in produced in the surrounding tissue (13,14). Thermal and nonthermal effects exist and they are under investigation. According to Natori (26) temperature increases of no more than 1.5°C above normal are considered clinically acceptable and nonthermal, that is, cavitation based effects can only be found where gas bodies are present, such as postnatal lung and intestines, or via ultrasound contrast agents. Therefore regular B-mode imaging is considered by many unaffected (26).

Finally, it should be mentioned that there are no bioeffects in the absence of contrast agent and also no bioeffects in the presence of contrast agents but absence of an ultrasonic field. Moreover, low doses ($10–50\,\mu\mathrm{L \cdot kg^{-1}}$)

Table 5. Acoustical Output Limits for Clinical Ultrasound Scanners per FDA Regulation

Tissue	$I_{\mathrm{SPPA.3}}$, $\mathrm{mW \cdot cm^{-2}}$	$I_{\mathrm{SPTA.3}}$, $\mathrm{W \cdot cm^{-2}}$	MI
Peripheral vasculature	720	190	1.9
Cardiac	430	190	1.9
Fetal and other	94	190	1.9
Ophthalmic	17	28	0.23

[a]Intensity values are derated for tissue attenuation with an acoustic attenuation of $0.3\,\mathrm{dB \cdot cm^{-1} \cdot MHz^{-1}}$.
[b]I_{SPTA} and MI are defined in the text, I_{SPPA} is defined as the intensity of the spatial peak pulse averaged waveform.

of ultrasound contrast agent yielded little if any bioeffects. In general, radiological contrast agents, or the imaging procedure itself, may bear the risk of bioeffects. That risk has to be balanced with the medical need for the procedure.

ACKNOWLEDGMENTS

I wish to thank Katy, Kim, Mario, John, and Jessi for their invaluable efforts for proof reading and their technical comments. Also, I would like to thank the reviewers for their contributions and very helpful commentary. Most of all, I wish to thank my wife Naki for her patience during many lost hours of family time.

BIBLIOGRAPHY

1. Angelsen BAJ. Ultrasound Imaging Waves, Signals, and Signal Processing; 2000.
2. Averkiou M, et al. Ultrasound contrast imaging research. Ultrasound Q 2003;19:(1):27–37.
3. Becher H, Burns PN. Handbook of Contrast Echocardiography—LV Function and Mycardial Perfusion. Berlin: Springer-Verlag; 2000.
4. Church CC. The effects of an elastic solid surface layer on the radial pulsations of gas bubbles. J Acoust Soc Am 1995; 97(3):1510–1521.
5. deJong N, et al. Absorption and scatter of encapsulated gas-filled microspheres: Theoretical considerations and some measurements. Ultrasonics 1992;30:95–105.
6. Duck FA. Physical Properties of Tissue. San Diego: Academic Press; 1990.
7. Feinstein SB. The powerful microbubble: From bench to bedside, from intravascular indicator to therapeutic delivery system, and beyond. Am J Physiol Heart Circ Physiol 2004;287: H450–H457.
8. Gray H, Williams PL, Bannister LH. Gray's anatomy: The anatomical basis of medicine and surgery. New York: Churchill Livingstone; 1995.
9. Grayburn PA. Current and future contrast agents. Echocardiography 2002;19(3):259–265.
10. Hoff L. Acoustic characterization of contrast agents for medical ultrasound imaging. Ph.D. dissertation at the Norwegian University of Science and Technology in Trondheim, Norway; 2000.
11. Holland CK, Apfel RE. Thresholds for transient caviation produced by pulsed ultrasound in a controlled nuclei environment. J Acoust Soc Am 1990;88(5):2059–2069.
12. Kasai C, Namekawa K, Koyano A, Omoto R. Real-time two-dimensional blood flow imaging using an autocorrelation technique. IEEE Trans, Ultrasonics, Ferroelectrics, Frequency Control 1985;SU-32(3):458–464.
13. Li P, Cao et al. Impact of myocardial contrast echocardiography on vascular permeability: An in vivo dose response study of delivery mode, pressure amplitude and contrast dose. Ultrasound Med Biol 2003;29(9):1341–1349.
14. Li P, Armstrong WF, Miller DL. Impact of myocardial contrast echocardiography on vascular permeability: Comparison of three different contrast agents. Ultrasound Med Biol 2004; 30(1):83–91.
15. Lindsay RB. The story of acoustics. J Acoust Soc Am 1965; 39(4):629–644.
16. Shung KK, Zipparo M. Ultrasonic transducers and arrays. IEEE Eng Med Biol 1996; 20–30.
17. Angelsen BAJ, et al. Which transducer array is best. Eur J Ultrasound 1995;2:151–164.
18. McDicken WN, Evans DH. Doppler Ultrasound: Physics, Instrumentation and Signal Processing. New York: John Wiley & Sons, Inc.; 2000.
19. Nowicki A, Litniewski J, Secomski W, Lewin PA. Estimation of ultrasonic attenuation in a bone using coded excitation. Ultrasonics 2003;41:615–621.
20. Misaridis TX, et al. Potential of coded excitation in medical ultrasound imaging. Ultrasonics 2000;38:183–189.
21. Abbott JG. Rationale and derivation of Mi and Ti—A review. Ultrasound Med Biol 1999;25(3):431–441.
22. American Institute of Ultrasound in Medicine (AIUM). Standard Specification of Echoscope Sensitivity and Noise Level Including Recommmended Practice for Such Measurements; 1978.
23. O'Brien WD, et al. Acoustic output upper limits proposition. J Ultrasound Med 2002;21:1335–1341.
24. Nyborg WL. History of the American Institute of Ultrasound in Medicine's efforts to keep ultrasound safe. J Ultrasound Med 2003;22:1293–1300.
25. van derWouw PA, et al. Premature ventricular contractions during triggered imaging with ultrasound contrast. J Am Soc Echocardiogr 2000;13(4):288–294.
26. Natori M. Ultrasound safety: Overview and what we do need in daily clinics for a safe use of diagnostic ultrasound. Inter Congress Ser 2004;1274:125–128.

See also COMPUTED TOMOGRAPHY; ECHOCARDIOGRAPHY AND DOPPLER ECHOCARDIOGRAPHY; IMAGING DEVICES; MAGNETIC RESONANCE IMAGING.

ULTRAVIOLET RADIATION IN MEDICINE

J. J. LLOYD
Regional Medical Physics Department
Royal Victoria Infurmary
Newcastle-upon-Tyne,
United Kingdom

INTRODUCTION

It is to the philosophers and physicians of the ancient civilizations that we should attribute the earliest history of ultraviolet radiation (UVR) in medicine (1). For example, the Greek sun god, Apollo, was also the spiritual god of healing, providing the first documentation of an association between sunlight and health. In 525 BC, Herodotus observed that the strength of a human's skull was related to sunlight exposure, >2000 years before the formal discovery of the role of sunlight in vitamin D metabolism. At about the same time the Egyptians were using psoralens from plant extracts and sun exposure in the treatment of vitiligo. However, it was not until Jonathan Ritter in 1801 discovered the UV region in the solar spectrum that the science of photobiology could really begin.

The Danish physician Niels Finsen (1860–1904) is regarded by many as the father of modern UV therapy. In a series of articles published between 1893 and 1896,

Figure 1. Patients receiving sun heliotherapy for the treatment of tuberculosis. (Picture courtesy of the University of Denver.)

Finsen stressed that it was the UVR in the solar spectrum that was responsible for sunburn and not the radiant heat, as the name implies. In parallel with his scientific investigations, Finsen was also an active clinician. He is best remembered for his successful treatment of *lupus vulgaris* (tuberculosis of the skin, mainly on the face) and in 1903 was awarded the Nobel Prize for medicine in recognition of this work. The photograph shown in Fig. 1 shows patients with *lupus vulgaris* being treated with sunlight (heliotherapy). Treatment of disease using artificial sources of UVR is known as actinotherapy or phototherapy.

Following the pioneering work of Finsen, the early part of the twentieth century saw the rapid expansion of heliotherapy and actinotherapy throughout Europe and the United States. The practice of actinotherapy continued to expand through the middle part of the twentieth century and was accompanied by an enormous literature on the subject during the 1920s and 1930s. This rapid growth is reflected by the many revisions of the handbook *Actinotherapy Technique* first published by the Sollux Publishing Company in 1933 and reprinted for the ninth time (7th ed.) in 1949 (2). (A copy of this can be seen on-line at http://www.meridianinstitute.com/eaem/hanovia/hanocont.html) Most of the irradiation protocols for the countless number of diseases, such as diabetes and angina, described in this book are now of historical interest only. The advent of effective antibiotics and the realization that the successes claimed in many of these diseases were little more than anecdotal have resulted in a more limited role of UVR in clinical medicine. However, today there is considerable interest in photobiology, both in treatment of diseases such as psoriasis, and in research into the photobiological basis of skin aging, carcinogenesis, and photodermatoses. Greater leisure time and ease of travel has led to potential for greater exposure to solar UV and greater potential for short- and long-term deleterious effects. This article presents a review of photobiology with particular emphasis on sources and measurement of UV relevant to medical diagnosis and treatment, but also encompassing biological effects of UV, natural UV exposure, medical applications, and hazard assessment.

THE ULTRAVIOLET SPECTRUM

Ultraviolet radiation is part of the electromagnetic spectrum and lies between the visible and the X-ray regions. Different wavelengths in the UV spectrum show enormous variations in ability to cause biological damage, and for this reason the UV spectrum is divided into three spectral regions: UVA, UVB, and UVC. The notion to divide the UV spectrum into different spectral regions was first put forward at the Copenhagen meeting of the Second International Congress on Light held during August 1932. It was recommended that three spectral regions be defined as follows:

UVA 400–315 nm
UVB 315–280 nm
UVC 280–100 nm

The subdivisions are arbitrary and differ somewhat depending on the disciplines involved (3). The boundary between UVA and UVB is sometimes set at 320 nm and that between UVB and UVC is sometimes regarded as 290 nm. The short wavelength limit for UV is sometimes quoted as 100 nm and sometimes as 200 nm. The UVA region has recently been divided into UVAI (340–400 nm) and UVAII (320–340 nm).

Due to potential confusion with this terminology and because of rapidly changing biological effects as a function of wavelength, it is recommended that, in publications in photobiology, bandwidths are quoted explicitly and ideally the full spectrum of UV sources is described.

BIOLOGICAL EFFECTS OF ULTRAVIOLET RADIATION

An understanding of biological effects of UV is vital in appreciating the requirements for sources and detectors in medical applications. As UV does not penetrate tissue readily, it is the eyes and skin that are organs of particular interest. Comprehensive reviews of the health effects of UV are given in a recent book from the National Radiation Protection Board in the United Kingdom (4). Briefer reviews can be found elsewhere (5,6). In this section, the effects of UV in normal subjects is briefly described; effects in disease are discussed in later sections.

In order to cause a biological effect, UVR must be absorbed and initiate a photochemical process. The biological molecule that absorbs the radiation is known as a chromophore. A plot of the effectiveness of a chromophore at absorbing radiation as a function of wavelength is the absorption spectrum. A plot of the effectiveness of the UVR of different wavelengths in causing a given biological effect is called an action spectrum. The shape of the action spectrum will depend both on the absorption spectrum for the chromophores initiating the effect and also the optical properties of the skin that influences the radiation reaching the chromophores. The transmission of UVR

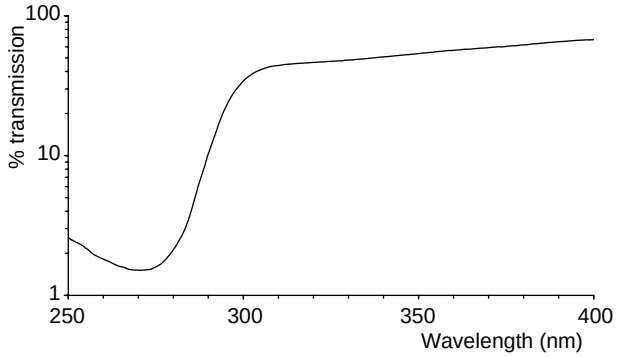


Figure 2. Fractional UVR penetration through the epidermis as a function of wavelength (B.6 L. Diffey, personal communication).

through the epidermis is shown in Fig. 2; UVA has much higher transmission than UVB or UVC, and therefore penetrates deeper into the skin.

Effect of UV Exposure on the Eyes

Both UVC and UVB radiation is predominately absorbed by the conjunctiva and cornea. The lens absorbs radiation in the range 300–370 nm, whereas radiation in the range 400–1400 [visible blue light to near-infrared (IR) radiation] is absorbed on the retina. Acute effects of UVR overexposure are photokeratitis (corneal injury) and photoconjunctivitis. Photokeratitis may occur in mountaineers as a result of high solar UV at altitude and high UV reflection from snow; this is referred to as "snow-blindness". Many epidemiological studies suggest a link between chronic ocular UV exposure and ocular disorders such as cataracts and the appearance of yellow/brown deposits in the cornea. Absorption in the UVC range is of little interest in epidemiological studies since there is no natural exposure at these wavelengths. However, corneal absorption is very high between 190 and 220 nm and radiation in this range is used therapeutically in laser photorefractive surgery. Also certain artificial sources, such as germicidal lamps, emit significant UVC, and therefore care is required in their use.

Effect of UV Exposure on the Skin

The acute effects of UVR on the skin include erythema (sunburn), skin thickening, tanning, and vitamin D production. Long-term effects include induction of skin cancer and premature skin aging. Individual response to UVR varies greatly. Fair skinned people burn more easily in the sun, find it difficult to obtain a tan, and have a higher risk of skin cancer compared to darker skinned individuals. In photobiology it is usual to categorize skin into six types, as shown in Table 1.

Table 1. Characteristics of Different Skin Phototypes[a]

Skin Type	Skin Color	Sensitivity to Sunburn	Ability to Tan	Skin Cancer Risk
I	White	Very high	Virtually nil	High
II	White	High	Poor	High
III	White	Medium	Good	Medium
IV	Olive	Low	Very good	Low
V	Brown	Very low	Very good	Very low
VI	Black	Very low	Very good	Very low

[a]From Ref. 65.

Acute Effects. Erythema is skin redness caused by inflammation and dilation of small blood vessels. Despite being extensively studied the underlying cause of UV induced erythema is poorly understood (7). The erythemal response of the skin in an individual can be defined by the minimum dose required to produce a just perceptible redness. This is referred to as the Minimal Erythemal Dose (MED). A standard erythemal action spectrum has been defined taking into account many published studies (8) (see Fig. 3). Skin sensitivity is maximum to radiation in the UVC and UVB range up to 298 nm. At longer wavelengths, the sensitivity drops rapidly reducing to be about 10,000 times less at 400 nm. The MED is a threshold measurement made by visual assessment. For quantitative studies a device measuring reflectance (11) can be employed to quantify erythema. This device works by measuring the decreased green reflectance, relative to red, from haemoglobin in the dermal blood vessels.

Within an individual the sensitivity to erythema can vary considerable from site to site (12,13). Minimal erythemal dose measured on the forearm may be twice that on the back, and the buttock skin is more sensitive than the back.

Chronic Effects. There are three types of skin cancer associated with UVR: squamous cell carcinoma (SCC), basal cell carcinoma (BCC), and malignant melanoma (MM). The SCCs appear as persistent red crusted lesions on exposed sites and have an incidence about a quarter of BCC. The BCCs appear as raised translucent nodules, normally on the face. Basal cell carcinoma appears to be related to cumulative UVR exposure, whereas SCC and MM may be related to intermittent exposure to UVR. In support of this idea is that fact that BCC and SCC tend to occur on habitually sun-exposed sites whereas MM occurs more commonly on intermittently exposed sites. Although MM has a much lower incidence than non-melanoma skin cancer, it accounts for ~80% of all skin cancer deaths.

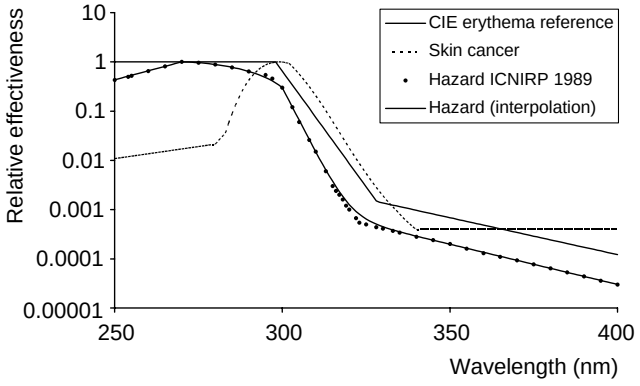


Figure 3. Standard action spectra for erythema (9), non-melanoma skin cancer (10), and UV hazard (6).

Melanoma incidence is increasing rapidly, probably due to increased recreational sun exposure, although other non-UV factors also play a part. Since the 1970s melanoma has seen the largest increase in incidence compared to all other major cancers in the United Kingdom and now stands at 8/100,000 person-years. It is predicted that the incidence will continue to rise, perhaps doubling over the next 30 years in the United Kingdom (14).

Photoageing is the term used to describe features of chronically sun exposed skin. These features include: wrinkles, "age spots", and thick leathery skin.

Effects on the Immune System

There is some evidence that exposure to UVR may have a systemic immunosuppressive effect. It is possible that UV exposure may increase the incidence of infectious diseases and play a role in the promotion as well as initiation of skin cancers.

ULTRAVIOLET SOURCES

Ultraviolet radiation is emitted during transition of a molecular electron from an excited, high energy, to a less energetic state. As the possible energy levels are fixed for a given molecule, radiation is emitted with distinct photon energies. When excitation is by heating, the release of radiation is termed incandescence. Alternatively, excitation can be generated by passage of an electrical discharge through a gas. This is the basis for most artificial UV sources used in medicine, for example, mercury arc lamps and fluorescent UV lamps. The UV emitting LEDs (light emitting diodes) are available, although these tend to have low output and be restricted to long wavelength UV. A further recent development has been the production of UV emitting lasers.

Solar Ultraviolet Radiation

Although direct use of solar radiation as a medical therapy (heliotherapy) is no longer employed, an understanding of solar UV can be helpful clinically, particularly in the investigation of abnormal reactions to sunlight. In clinical phototesting solar simulating sources may be used. A brief summary is therefore given here, for more information the reader is directed to various recent reviews (15–18).

The spectrum of extraterrestrial solar radiation approximates that of a black body at ~5800 K. At the earth's surface, this spectrum is modified by atmospheric attenuation. The stratospheric ozone layer prevents almost all radiation with wavelengths <290 nm and a substantial proportion of UVB (70–90%) reaching the earth. When the sun is lower in the sky, the path length through the atmosphere is greater. The UV intensity is therefore reduced at all wavelengths, but more so for UVB than UVA. The relative intensity of UVA compared to UVB is therefore greater in the winter than summer in the United Kingdom. At mid-day in the summer in the United Kingdom the ambient total UV is ~4 mW/cm^2, of which UVB contributes ~5%. However, due to its greater deleterious effect, UVB contributes to ~80% of the harmful effects of solar UV.

Incandescent Sources

Incandescent sources emit a smooth broad spectrum of radiation with the peak wavelength inversely related to the absolute temperature. Conventional tungsten bulbs used for domestic lighting have peak emission in the infrared (IR) region and emit very little UVR. Tungsten halogen bulbs operate at higher temperatures and may produce rather more UVR. These sources are not used for medical applications due to the low UV output, although, they may be used as reference sources for UV meter calibration.

Mercury and Metal Halide Arc Lamps

The radiation emitted from a mercury-vapor arc lamp arises from two mechanisms. Line or characteristic radiation is produced as a result of excitation of the constituent atoms, together with a spectral continuum that is chiefly due to ion and electron recombination. Lamps can be produced that operate at different pressures. A low pressure mercury arc lamp consists of a fused silica tube filled with argon at ~1 Pa and containing a drop of mercury. A discharge occurs between the electrodes sealed into the ends of the tube. More than 90% of the radiant energy produced by the discharge is at 253.7 nm. Various other characteristic mercury spectral lines occur, for example, at 313 and 365 nm, and these sources are useful for wavelength calibration of spectroradiometers.

If the pressure in the lamp tube is increased, then the spectral lines broaden and also more radiation occurs in the continuum (see Fig. 4). The "alpine sunlamp" is a type of medium pressure arc lamp that was widely used for phototherapy, but has since been superseded by fluorescent lamps. The addition of metal halides to a high pressure mercury discharge lamp greatly enhances the UV output. These lamps were common in phototherapy departments at one time, but have now been mostly replaced by fluorescent UVB lamps. These lamps are still occasionally used for cosmetic tanning. By incorporating optical filters between the lamps and the irradiated subject, absorption of the UVC or UVC and UVB components can be achieved (see Fig. 5).

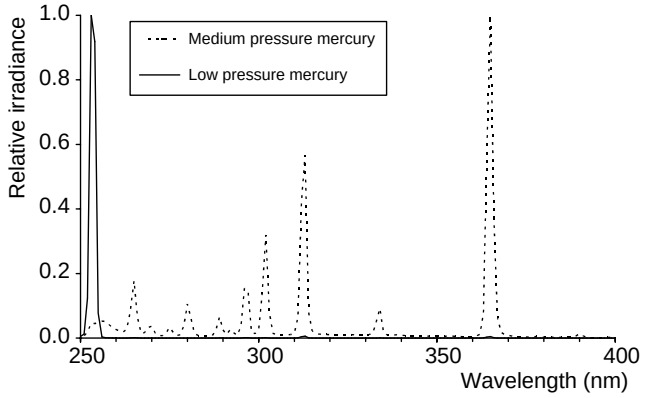


Figure 4. Ultraviolet emission spectra from low and medium pressure mercury lamps.

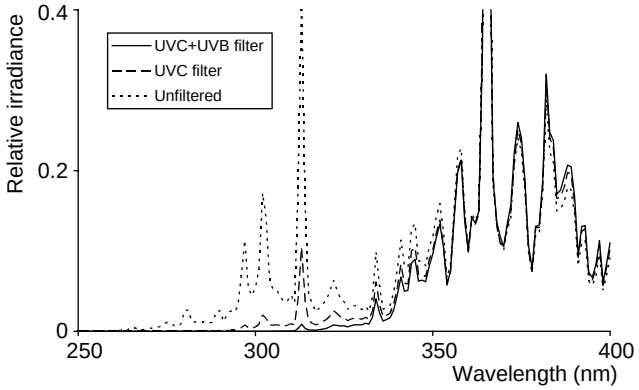


Figure 5. Ultraviolet emission spectra from metal halide lamp with and without filters.

The medium pressure mercury lamp is also the source in the "Wood's lamp", although in this case optical filtration is incorporated into the lamp housing to limit the short-wavelength UV and visible light emissions. This results in an approximate monochromatic source of 365 nm. The Wood's lamp is used in diagnostic fluorescence techniques in dermatology.

Fluorescent Lamps

A fluorescent lamp is a low pressure mercury discharge lamp that has a phosphor coating applied to the inside of the envelope. Fluorescence radiation is produced by the excitation of the phosphor by the 253.7 nm radiation. The spectral power distribution of the fluorescence radiation is a property of the chemical nature of the phosphor material. In addition to a continuum due to the phosphor, the mercury characteristic lines are superimposed. These lines are present in all mercury fluorescent lamps irrespective of the phosphor material. However, it is possible to suppress the emission of certain lines either by using a lamp envelope material that absorbs unwanted short-wavelength radiation or by incorporating a suitable filter in the lamp housing. Because fluorescent lamps are relatively cheap, stable, and efficient and can produce a large high intensity irradiation field, they are now very widely used in phototherapy and cosmetic tanning.

The spectral power distribution of UVA lamps used in cosmetic tanning and PUVA phototherapy are shown in Fig. 6. Broad-band UVB fluorescent lamps (e.g., TL12, Phillips, Eindhoven, the Netherlands) incorporate a phosphor that results in a continuous spectrum from 270 to 380 nm, with a peak at 313 nm. This lamp is used in phototherapy, however, it is being replaced by the narrow-band lamp (Phillips TL01), which has ~80% of its output within 2 nm of the peak at 311 nm (Fig. 7).

The black light fluorescent lamp emits a similar spectral power distribution to the UVA fluorescent lamp in the region 315–400 nm, but with suppression of the mercury lines in the visible spectrum. This results from using a visible-absorbing, UVA transmitting, glass envelope. When switched off, the lamp envelope appears almost black. A purplish light is perceived when the lamp is operating.

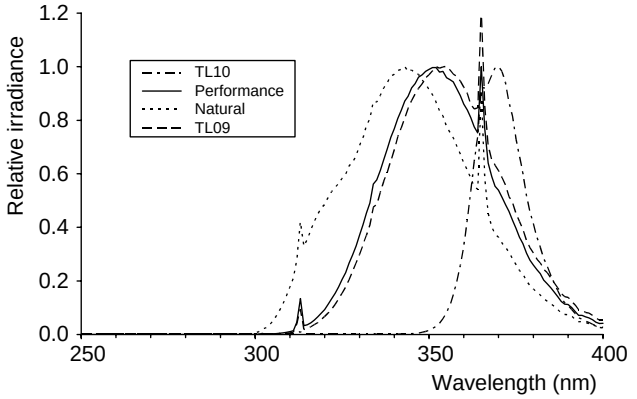


Figure 6. Ultraviolet emission spectra from fluorescent UVA lamps used in PUVA therapy and commercial tanning studios (Cleo Performance and TL09). Also shown are the Phillips TL10 fluorescent lamps emitting longer wave UVA, that were popular for tanning at one time and the Phillips natural fluorescent lamp that have been promoted by the manufacturers as "a more natural way to tan".

Xenon Arc Lamps

In the xenon arc the radiation is emitted primarily as a continuum, unlike the mercury arc, which essentially emits a line spectrum. The production of the continuum is optimum under conditions of high specific power, high current density, and high internal pressure, leading to compact, bright sources. Because of high operating temperatures the lamp envelope is normally constructed from fused silica. Unlike arc lamps containing mercury that has to vaporize, xenon lamps contain a permanent gas filling and the full radiation output is available immediately after switching on so that no run-up period is necessary.

Solar Simulators

Because of the similarity of the spectrum to that of the solar spectrum, the xenon lamp has been employed as a laboratory source of sunlight, the so-called solar simulator. In order to improve the match to sunlight, a WG320 filter

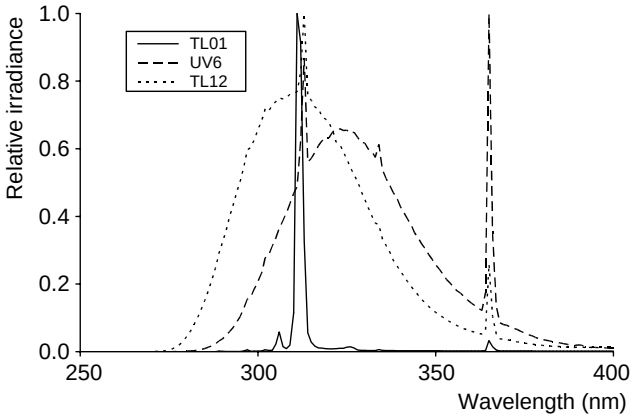


Figure 7. Ultraviolet emission spectra from fluorescent UVB lamps.

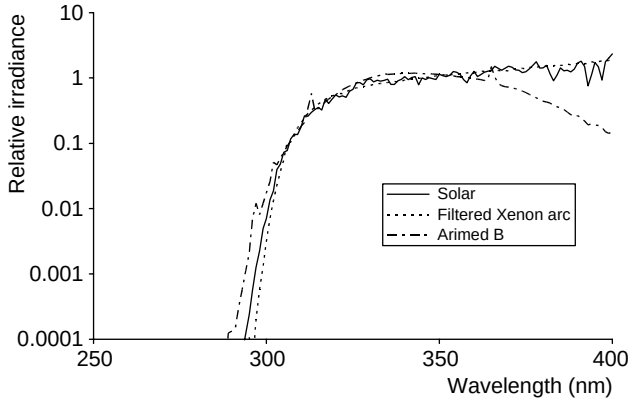


Figure 8. Ultraviolet emission spectra from solar simulator sources together with the solar spectrum obtained at mid-day on a sunny summer's day in Newcastle-upon-Tyne (UK). The relative irradiance has been displayed on a logarithmic scale to allow comparison of small differences in the UVB range.

(Schott AG, Mainz, Germany) should be used and often a visible light absorbing filter (e.g., Schott UG5 or UG11) is also employed. A drawback of the arc lamp solar simulator is that it is difficult to irradiate a large area. Fluorescent lamps can be used for this purpose and the best match to sunlight is probably given by the Arimed B lamp. The suitability of a lamp as a solar simulator can be assessed by comparing the percentage relative cumulative erythemal effectiveness (%RCEE) for a number of wavebands and compare this with the %RCEE values for a "standard sun" (17). Various solar simulator sources are shown in Fig. 8.

Monochromatic Radiation

A radiation monochromator consists of a xenon arc lamp together with a double diffraction grating monochromator. By adjusting the angle of the diffraction grating, and by adjusting the input and output slit widths a narrow range of UV wavelengths can be obtained. These instruments are widely used in photobiology studies, both *in vivo* and *in vitro*. They are also routinely used for testing the erythemal sensitivity of patients in phototesting clinics. Although the radiation is referred to as monochromatic, in fact bandwidths of between 30 and 5 nm are typically employed. Narrowing the bandwidth <5 nm may lead to unacceptably low irradiance.

Lasers

Lasers produce truly monochromatic radiation. There are a number of lasers that produce UV radiation with emissions possible in the UVA, UVB, and UVC range. Argon Fluoride lasers produce 193-nm radiation (UVC) and are used for corneal refractive surgery (19). Dye lasers have been used to investigate the action spectrum for erythema in human skin (20).

THE MEASUREMENT OF ULTRAVIOLET RADIATION

Techniques for the measurement of UVR may be divided into three classes; biological, chemical, and physical. In general, physical devices measure power, while chemical and biological systems measure energy.

Table 2. Radiometric Terms

Term	Unit
Wavelength	Nanometer, nm
Radiant energy	Joule, J
Radiant flux	Watt, W
Radiant intensity	Watt per steradian, W/sr
Radiance	Watt per square meter per steradian, $W/m^2/sr$
Irradiance	Watt per square meter, W/m^2
Radiant exposure (often referred to as "dose")	Joule per square meter, J/m^2

Quantities and Units

In clinical and photobiological UVR dosimetry, it is customary to use the terminology of radiometry rather than that of photometry. Photometry is based on visible light measurements weighted by the human eye's response curve, and therefore not relevant as the eye does not respond to radiation at wavelengths <380 nm. The common radiometric terminology is listed in Table 2. These radiometric quantities can also be expressed in terms of wavelength by adding the prefix "spectral". In clinical photobiology, the derived unit of milliwatt per square centimeter is commonly used and radiant exposure tends to be referred to as dose. Note that dose in this context differs from the term used in radiobiology where dose indicates energy absorbed per unit mass of tissue.

The most frequent radiometric calculation is to determine the time for which a patient who is prescribed a certain dose (in J/cm^2) should be exposed when a radiometer indicates the irradiance in mW/cm^2. The relationship between these quantities is

$$\text{exposure time(min)} = \frac{1000 \times \text{dose}(J/cm^2)}{60 \times \text{irradiance}(mW/cm^2)}$$

Weighted Irradiance

It is useful to derive quantities by weighting the spectral irradiance by an appropriate action spectrum. The erythemally effective irradiance E_{eff}, is defined as

$$E_{\text{eff}} = \sum s(\lambda)E(\lambda)\Delta\lambda$$

Where $s(\lambda)$ is the CIE erythemal action spectrum (8), $E(\lambda)$ is the measured spectral irradiance, and $\Delta\lambda$ is the bandwidth of measurement. The "Global UV index" (21) can be used as a standard way of expressing the erythemally weighted solar irradiance. A measure of the integrated effective irradiance received is given by the Standard Erythemal Dose (SED) (9). One SED is equal to an effective dose of $100\ J/m^2$. In the past, some have used the MED as a standard measure of erythemal dose. However, this is not correct, MED should be reserved to describe an individual's erythemal response and not a standard measure of dose received. The number of SED required to cause just

perceptible erythema ranges from ~1.5 to 6 for skin types I to IV. It can be demonstrated that erythemal responses to UV sources with different spectra are similar if the doses expressed are SED (7).

Physical UV Detectors

Physical UV detectors utilize either thermal or photon mechanisms (22).

In thermal detectors, the absorption of radiation increases the temperature in the detector element and this rise in temperature is measured by some means. Thermopile UV detectors are the simplest and commonest thermal device used to measure UV irradiance. A multijunction thermopile is formed from a number of thermocouples in series, which generate a voltage that is proportional to incident energy in the form of heat. Thermopiles intended for use with UV radiation are fitted with quartz windows, which transmit well in the UV range. The advantage of thermal detectors is that they have a relatively flat spectral response over a wide wavelength range.

Photon detectors operate by absorbing discrete quanta of photon energy, and therefore have a threshold wavelength above which no radiation is detected. The lower wavelength limit is related to optical properties of the detector or associated filters. The response of photon devices are therefore inherently wavelength dependent and thus have to be calibrated for each source of interest.

Photoemissive detectors have a photocathode from which electrons are ejected when photons are absorbed. These electrons are then collected by an anode and a current is produced. The simplest of this type of detector is the vacuum phototube consisting of an evacuated tube with a potential difference applied between the cathode and the anode. This device has a gain of unity and a low responsivity (~0.05 A/W). A gas-filled phototube has a gain of ~10 due to secondary ionization of the gas in the tube that has the effect of producing a greater anode current. Photomultiplier tubes have a series of electrodes (called dynodes) having successively greater potential differences applied between them. As electrons hit each dynode they release further electrons that in turn release more electrons at the next dynode, and so on, leading to a high overall gain (typically 10^6). The responsivity of photomultiplier detectors tends to be high ($\sim 5 \times 10^4$ A/W).

Photodiodes, or junction photodetectors, have a depletion region formed by the junction of n and p doped semiconductor material. On absorption of a photon in this region, electron hole pairs are formed that are then swept out of the region and cause a current to flow in an external circuit. These devices can either be operated in a zero bias or reverse bias mode. For UV detection, GaAsP, GaP, or Si photodiodes are used. These photodiodes are small, cheap, and rugged with good responsivity (~0.1 A/W), and are therefore ideal UV detectors.

UV Radiometers

A radiometer is a complete UV measurement device, consisting of a detector and a meter to amplify and display the detector output. Narrow-band UV radiometers are used to measure the irradiance of a source in the different UV bands. The detector consists of a diffuser to collect the UVR, a filter, and the sensor (e.g., a photodiode) itself. Lambert's law states that the irradiance falling on a surface varies with the cosine of the incident angle. A good diffuser should possess an angular response close to the ideal cosine response—of particular importance in phototherapy, where arrays of tubes are used leading to a large source area irradiation.

A good example of a hand-held radiometer for phototherapy is the IL1400A UV radiometer system (International Light, Newburyport, MA) used with the following probes (see Fig. 9):

- SEL 240 solar blind vacuum phototube, fitted with a SCS280 filter.
- SEL 033 Silicon photodiode detector, fitted with a UVA filter.

Both of these detectors are fitted with domed Teflon diffusers that have a good angular response. The SEL 240 responds over the 185–320 nm range, and the SCS280 filter is a sharp-cut filter, removing all wavelengths <280 nm. The combination of the two gives a system that responds to wavelengths in the 265–332 nm range, which is well matched to the UVB band. The SEL 033 detector responds

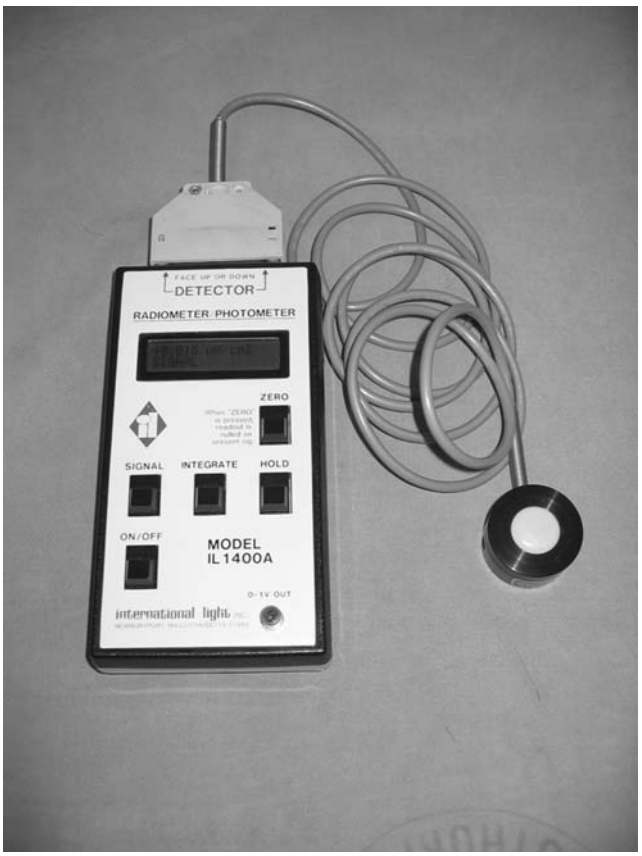


Figure 9. A hand-held UV radiometer suitable for routine monitoring of phototherapy equipment. Model shown in an International Light IL1400A radiometer with UVA detector (see text for further details).

over the 200–1100 nm range, therefore the UVA filter used is a wide-band filter that restricts this response to the UVA range (325–388 nm). It is important to note that narrow-band radiometers have a response that varies with wavelength, and therefore they must be calibrated for each type of source they will be used to measure.

A broad-band radiometer possesses no filter, and therefore measurement of the total irradiance from a source is obtained. Typically these incorporate a thermopile detector connected to an amplifier and display electronics.

Spectroradiometry

Spectroradiometry is concerned with the measurement of the spectrum of a source of optical radiation. In many cases, spectral measurements are not required as ends in themselves, but for application to the calculation of biologically weighted radiometric quantities.

The three basic requirements of a spectroradiometer system are (1) the input optics, designed to conduct the radiation from the source into (2), the monochromator, which usually incorporates one or two diffraction gratings as the wavelength dispersion elements, and (3) an optical radiation detector, either a photomultiplier tube or a solid-state photodiode. The UVR is collected by a diffuser, which should have good angular response, and transmitted to the monochromator by a light guide. A quartz optical fiber is required as quartz has a very high UV transmission, unlike ordinary glass. The monochromator allows the separation of polychromatic radiation into very narrow bandwidths (typically 1 nm) thereby allowing the spectral irradiance of a source to be measured. This is achieved by using a diffraction grating, which disperses the incident radiation. Scanning spectroradiometer systems operate by measuring at a given wavelength, then changing the angle of the diffraction grating, making another measurement, and so on. For photobiology applications it is recommended that a double diffraction grating system is used. The second diffraction grating achieves a reduction in the radiation transmitted outside the waveband of measurement (so-called "stray radiation"). Very small amounts of stray radiation may have large implications in the measurement of erythemally weighted irradiance if it occurs at biologically highly effective wavelengths. The spectroradiometer system requires a high responsivity detector as narrow bandwidths are selected. All of the spectral examples in this article were obtained with this type of spectroradiometer system (the majority using a model DMc150 monochromator from Bentham Instruments, Reading, U.K.). Spectroradiometers need to be calibrated with reference to a standard lamp, which in turn has a output calibrated by a national standards laboratory. The disadvantage of double-diffraction grating scanning spectroradiometers systems are the high cost, large size, and relatively long time required to obtain a complete spectrum.

Recently small, rugged, and relatively cheap spectroradiometers systems have become available (e.g., SolaScope 2000, 4D controls Ltd, Redruth, U.K.). These devices consist of an input optic, single diffraction grating, and solid-state multidetector array. In this device, there are no moving parts and a complete spectrum is rapidly obtained. A major disadvantage for photobiology application is that the stray light level is high (>10%), although this can be subtracted from the measured spectrum to give a reasonable level of compensation (23).

Calibration of Meters for Phototherapy. It is recommended that radiometers are routinely used in phototherapy practice and regularly calibrated in a manner traceable to a national standards laboratory (24). This can be done in one of two ways: either by reference to a calibrated spectroradiometer (25) or by reference to a calibrated meter (26). Surveys of radiometer accuracy have shown very disappointing results, particularly for the measurement of narrow-band UVB radiation (26,27). This is of particular concern since narrow-band UVB phototherapy is becoming increasingly popular, surpassing traditional PUVA therapy. Also, the increase in erythema with dose is greater with narrow-band UVB than PUVA, and therefore the need for accurate dosimetry is greater.

Chemical and Biological Methods

Detailed description of chemical and biological methods is beyond the scope of this article, but they are mentioned briefly as they may be used in a few medical applications. Polysulphone film (28) changes its optical properties with the absorption of UVR. It is possible to measure this change and relate it to the exposure dose received by the film. These devices are useful for measuring human UV exposure as they are unobtrusive for individuals to wear. Biological dosemeters (29) make use of the inactivation of bacteria or viruses as a function of UVR dose. There are certain applications, such as measuring UV doses in water flowing in a disinfection plant, where biological dosemeters are the only reliable way to measure UV dose.

DIAGNOSTIC USES OF ULTRAVIOLET RADIATION

Fluorescence Techniques in Diagnosis

The principle of fluorescence diagnosis is based on the absorption of radiation by a fluorophore in tissue and the subsequent emission of photons of light at a longer wavelength. Typically, the exciting radiation is short wavelength visible light or UVR and the fluorescence is longer wavelength visible light. Certain tissue types may be distinguished (e.g., neoplastic tissue) from surrounding tissue due to differences in endogenous fluorophore. Other techniques use exogenous photosensitisers that may be selectively accumulated in neoplastic tissues. The most common of these photosensitisers is 5-aminolevulinic acid (5- ALA). Originally, applications were limited to visual observation of fluorescence at easily accessible sites, such as the skin and mouth. However, technological advances in fiber optics, light sources, digital video cameras, and computer enhanced imaging has led to considerable interest in endoscopic fluorescence imaging (30). There is a wide range of excitation and emission wavelengths that can be used in these applications, often these are both in the visible range, but techniques using a Wood's lamp have been used for many years by dermatologists.

Tinea capitis (ringworm) is a fungal infection of the scalp. Hairs that are infected with the fungi *Microsporum audouini* or *Microsporum canis* will fluoresce with a bright blue-green color under Wood's light. The diagnosis can be confirmed by removing the fluorescent hairs for direct microscopic examination and culture. Erythrasma is a superficial bacterial infection of the skin, usually between two surfaces of skin that rub together, such as the groin or toe webs. Erythrasma produces scaling and cracking and may be accompanied by itching. The responsible organism (*Corynebacterium minutissimum*) produces a porphyrin that fluoresces a bright coral-red color on irradiation with a Wood's lamp. Irradiation of the oral cavity with a Wood's lamp will produce fluorescence that may prove useful in the diagnosis of various dental disorders such as early dental caries (tooth decay). Normal teeth fluoresce with a light blue color. The presence of calculus on the teeth will result in a yellow-orange fluorescence under UVA illumination.

Photosensitivity Investigations

Sunlight is capable of inducing or aggravating skin diseases. This group of diseases is collectively known as the photodermatoses (31,32). The classical appearance of a photodermatosis is a rash confined to regions exposed to sunlight; the face, neck, arms, and hands. The appearance may be an exaggerated sunburn, or there may be eczematous changes with polymorphic lesions that can include papules, vesicles, and bullous eruptions. The principal photodermatoses are shown in Table 3.

The accurate diagnosis of a suspected photosensitive patient demands, above all, a clear and detailed history. For example, the season of the year in which the symptoms occur may give some guide to the wavelengths in the sun's spectrum that are responsible. The time between exposure and the appearance of the lesion may also be a helpful guide. However, in equivocal cases, phototesting investigations are desirable (33). The object is twofold: to reproduce the disease so as to confirm the diagnosis, and to ascertain those wavelengths of sunlight that are responsible for the photosensitivity, so that suitable preventative measures can be taken.

Provocation Testing. The aim of provocation testing is to expose a relatively large area of skin (say 10–20 cm^2) to UVR and attempt to reproduce the rash of which the patient complains. The methodology used and the success of the technique is variable. The provocation source used in the past has been a medium pressure mercury arc lamp in combination with optical filters to isolate broad spectral regions (e.g., UVB or UVA). Recently, the use of a filtered xenon arc solar simulator source has been recommended as optimal (34). However, others have had good success with narrow-band UVB and UVA fluorescent lamps (35).

Minimal Erythemal Dose Determination. The MED is usually determined using a radiation monochromator. This enables the patient to be irradiated in narrow wavelength intervals. The patient's back is normally chosen for irradiation since it has a large surface area with reasonably uniform sensitivity to UV radiation. A geometric series of doses at each of several different wavelengths is used (33,36) and the results of irradiation are observed 24 h later. In most centers visual inspection is used, although for quantitative research studies an erythema meter may be used (11). Comparing the response in the UVA and UVB range may be helpful in diagnosis (37). For example, in drug induced photosensitivity the UVA MED may be low, and the UVB MED may be normal, whereas in chronic actinic dermatitis, both UVA and UVB MEDs are low (see Fig. 10).

Photopatch Testing. Certain chemicals when activated by UVR in contact with the skin may cause a photocontact allergic reaction in sensitive individuals. This is not to be confused with a phototoxic reaction that occurs in all subjects when exposed to a photosensitizer such as psoralen. The most common photocontact allergens are sunscreens. When UVR is absorbed by a chemical sunscreen it becomes altered and the patient may have an allergic

Table 3. The Principal Photodermatoses, Main Clinical Features, and Typical Phototest Responses

Condition	Typical Clinical Features	Typical Erythemal Response	Provocation Test Response	Photopatch Test Response
Polymorphic light eruption	Itchy, papular rash occurring after a few days sun exposure. Persists a few weeks	Usually normal MEDs, occasionally low MEDs	Positive response (50–80% sensitivity) at suberythemal doses	n/a
Chronic actinic dermatitis	Eczematous rash on sun exposed site in summer months	Low or very low MEDs within UVA and UVB waveband, sometimes sensitivity to blue light also	n/a	n/a
Solar urticaria	Urticarial reaction a few minutes after sun exposure. Fades rapidly	Rapid reaction (within minutes) that may include UVB, UVA and blue/green light sensitivity	n/a	n/a
Drug induced photosensitivity	Rash on sun exposed sites: features vary with responsible drug	Low MEDs in UVA, normal MEDs in UVB range	n/a	n/a
Photocontact allergy	Confluent rash on exposed sites: may be severe	Normal	n/a	Positive when chemical is exposed to UV

Figure 10. Abnormal erythemal sensitivity test obtained at 24 h after exposure in a patient with chronic actinic dermatitis. Dose increments of 40% were used at the 300, 320, and 350 nm. The ⋙ marks indicate the direction of decreasing dose. In a normal subject, only a few exposed areas would become red.

reaction to this altered chemical. Photopatch testing (33,38) consists of placing several small patches of chemicals on the back in two duplicate sets. After 24 h, one set is exposed to UVR (usually 5 J/cm^2 UVA from a fluorescent lamp or radiation monochromator). Reactions are examined after a further 48 h. If a chemical causes a reaction at exposed and unexposed sites this is a contact allergic reaction. If the reaction occurs at the UVR exposed site, only then this is a photocontact reaction.

THERAPY OF SKIN DISEASE WITH ULTRAVIOLET RADIATION

The principal application of UVR for therapy is to treat a variety of skin diseases. The UVR may be administered on its own (phototherapy) or in conjunction with adjunctive agents applied topically or photoactive drugs taken systemically (photochemotherapy).

Ultraviolet Phototherapy

In 1994, a survey of UVB phototherapy in the United Kingdom (39) found that an appreciable number of centers were using equipment that was old and was suboptimal in producing a therapeutic effect. A similar survey in 2002 (24) showed that improvements in equipment were being made. Mercury discharge lamps had mostly been replaced by fluorescent tube irradiators and broad-band UVB lamps had mostly been replaced by narrow-band UVB lamps. The use of narrow-band UVB therapy has increased recently and now surpasses the use of psoralen plus UVA photochemotherapy (PUVA) (40). A recent review and guidance on the use of narrow-band UVB therapy by the British Photodermatology Group (40) concluded that there was good evidence to support its use in the treatment of psoriasis, chronic atopic eczema, vitiligo, and polymorphic light eruption. A range of other diseases have been treated with varying success.

Action Spectrum for Clearance of Psoriasis Using UVR Alone. Psoriasis is a common genetically determined skin disease with a 2% incidence worldwide (41). It is characterized by disfiguring, often distressing, red scaly plaques that may become confluent. The severity varies in any individual at any given time, but it often persists indefinitely. Parrish and Jaenicke (42) studied psoriasis response at different UV wavelengths to derive an action spectrum for psoriasis clearance. Figure 11 shows the clearance action spectrum together with the action spectrum for erythema for one subject in that study; similar results were found for other subjects. It can be seen that radiation with wavelengths <290 nm is ineffective at clearing psoriasis, but effective at producing erythema. Broad-band UVB sources (such as the Phillips TL12 lamp) that have traditionally been used for phototherapy tend to have relatively high emissions at wavelengths <290 nm. The narrow-band UVB lamp (Phillips TL01) has only 2.3% erythemally effective radiation at wavelength <290 nm. It is therefore not surprising that clinical studies comparing traditional broad- and narrow-band (TL01) UVB lamps for psoriasis treatment (40) have usually shown that narrow band is more effective. There are fluorescent broad-band UVB lamps available (e.g., UV6; Sylvania, Brussels, Belgium) with little or no emission <290 nm and the efficacy of these for psoriasis clearance has not yet been established. Spectra of various UVB fluorescent sources are shown in Fig. 7.

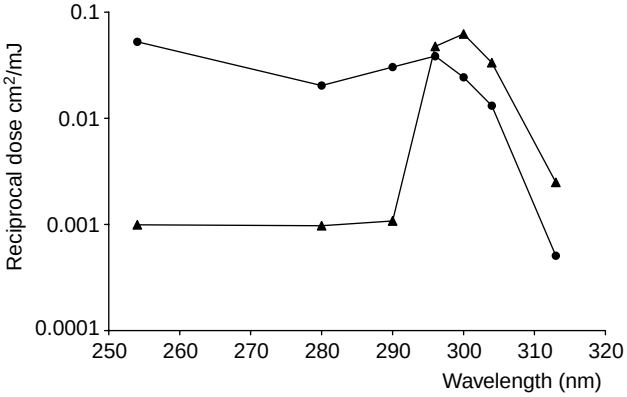


Figure 11. Psoriasis clearance action spectrum and erythema action spectrum in an individual subject. (From Ref. 42). The triangles indicate the reciprocal of the lowest daily dose to clear psoriasis and the circles are the reciprocal of the minimal erythemal dose. Similar responses were seen for the other three subjects in this study.

Psoralen Photochemotherapy

This form of treatment, known as PUVA, involves the combination of the photoactive drug psoralen and UVA irradiation to produce a beneficial effect. Psoralen photochemotherapy has been used to treat many skin diseases (43), although its principal success has been in the management of psoriasis. The psoralen may be applied to the skin either topically (44) or systemically (45); the latter route is generally preferred. The psoralen is usually administered as 8-methoxypsoralen (8-MOP). The patient ingests the 8-MOP tablets and is then exposed to UVA radiation 1 or 2 h later when the photosensitivity of the skin is at a maximum. The mechanism of the treatment is thought to be that psoralen binds to DNA in the presence of UVA, resulting in a subsequent transient inhibition of DNA synthesis and cell division. The UVA lamps used for PUVA therapy (such as the Phillips TL09 or Phillips Cleo Performance lamps) (Fig. 6), typically have a broad spectrum from ~315–400 nm peaking at ~352 nm. Within the UVA range, the action spectrum for psoriasis clearance is thought to be similar to the erythema action spectrum in psoralen sensitised skin and these peak near 320 nm (46). A study using narrow-band UVB and psoralen found no significant difference in response compared to conventional PUVA treatment (47). It is possible that UVA lamps conventionally used for PUVA treatment do not have the optimum spectral characteristics. However, narrow-band UVB phototherapy is now more widely used than PUVA and there is less interest in pursuing optimisation of PUVA treatment.

Treatment Regimes for UV Therapy

It is the custom in phototherapy and photochemotherapy to use the patient as their own biological monitor. The exposure of UVR at a given wavelength required to produce a given degree of erythema in normal Caucasian skin can vary by a factor of 5 or more (48), depending on an individual's susceptibility to sunburn. For photochemotherapy the degree of variability in erythema in sensitized skin may be greater due to variations in psoralen metabolism. Therefore, before embarking upon a course of irradiation, the minimum dose to cause just perceptible erythema should be determined. For phototherapy, this is termed the MED, whereas for photochemotherapy it is termed the minimal phototoxic dose (MPD). Either MED or MPD measurement is made by exposing the skin to increasing doses of UV using a lamp with the same spectral output as the treatment lamp. Devices using a series of filters have been designed to facilitate this (49) (see Figs. 12 and 13). For phototherapy, the MED is usually determined at 24 h. However, erythema in psoralen sensitized skin does not reach a maximum until ~72–96 h, so the MPD is determined at this time. Starting doses for treatment are usually set at ~50–70% of the MED or MPD. Treatment is typically given two or three times weekly until the psoriasis clears. At each treatment, the dose is increased to allow for acclimatisation of the skin. If erythema occurs then the dose may be reduced. The total time until clearance varies considerably from one patient to another, and in some cases complete clearing of the lesions is never achieved. However, typically ~25 treatments are required for clearing of the psoriatic lesion in many patients over a period of ~10 weeks. Various different treatment regimes are used with varying starting doses and number of treatments per week. Examples are given in British Photodermatology Group guidelines (40,44,45) and by Diffey and Hart (50).

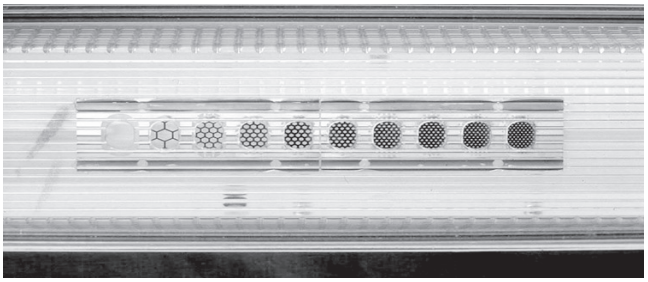


Figure 12. An erythemal skin testing device for use prior to phototherapy, consisting of a TL01 fluorescent lamp and a number of apertures covered by metal grid filters [based on design by Diffey et al. (49)].

UVA1 Therapy

There have been several studies into the use on UVA1 (340–400 nm) therapy (51). As yet this treatment is only available in a few centers and insufficient work has been done to reach a firm conclusion on its efficacy.

Risks of Phototherapy

The major acute risk in phototherapy is that of erythema. In phototherapy, it is the erythemal response in the unaffected skin that limits the treatment dose and frequency. Therefore, it is important to have an understanding of how erythema increases over time and with increasing dose for different therapy sources. Erythema peaks at ~8–12 h after exposure for solar simulated radiation, UVC and UVB (7,52). The degree of erythema varies rapidly with the dose of UVR. When the log of dose is plotted against erythema index, a sigmoid shaped curve is seen with a relatively linear portion above the MED. The steepness of this curve will show variation from individual to individual.

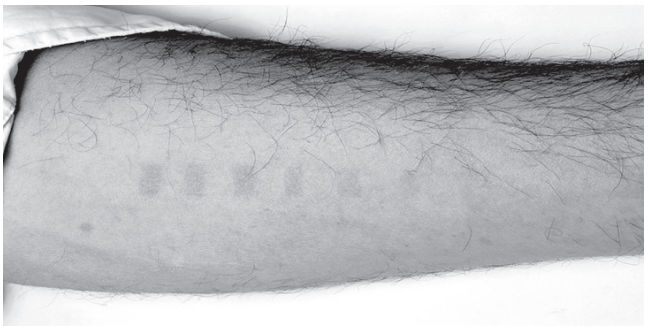


Figure 13. Normal erythemal response at 24 h obtained using the device shown in Fig. 12 in a subject with skin type 2. The MED is seen at site 6 or possibly 7.

Figure 14. A common type of whole body phototherapy therapy supplied by Waldmann (Herbert Waldmann GmbH & Co., Villingen-Schwenningen, Germany.) This illustrates the direct method of measuring UV irradiance; after the door is closed lamps are lit and measurements are made in a number of positions.

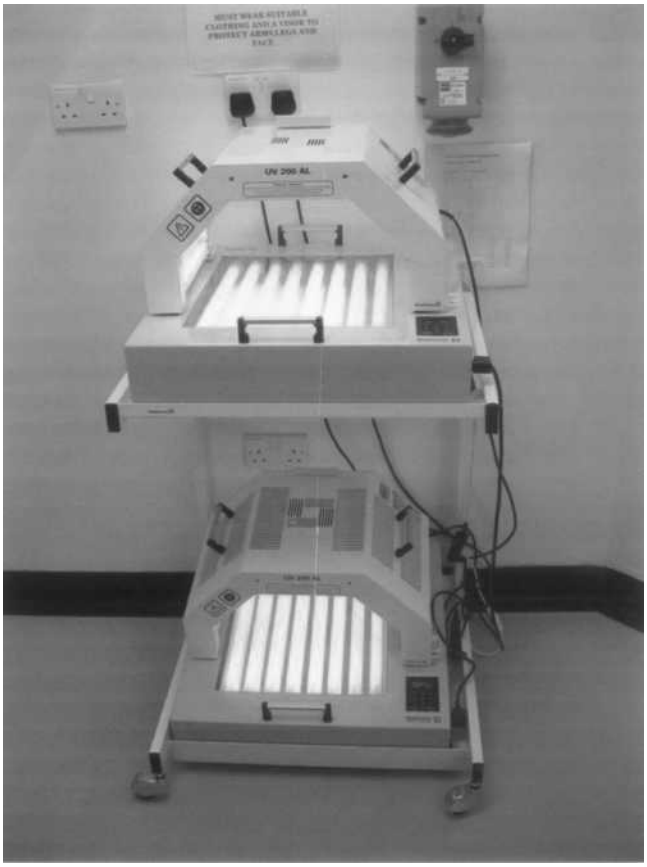


Figure 15. A small PUVA therapy unit suitable for treatment of the hands and feet.

However, on average, for UVB sources doubling the dose will cause a change from mild to moderate erythema (53), whereas for psoralen sensitized skin a quadrupling of the dose is required to cause a similar change (54). It is now firmly established that long-term PUVA treatment leads to a rise in SCC incidence (50). A cumulative dose of <500 J/cm^2 is unlikely to result in significant risk, whereas for a cumulative dose of 2000 J/cm^2 the risk of SCC is $\sim$50%. Phototherapy with UVB alone is thought to carry a lower risk of skin cancer than PUVA (40).

Equipment for Ultraviolet Therapy of Skin Diseases

Ultraviolet therapy is usually delivered using large cabins housing $\sim$24–48, 6 ft vertical fluorescent lamps Fig. 14. Compact units for treating just the hands and feet or smaller areas are also available (Fig. 15). All treatment devices usually incorporate timers to terminate treatments and most now have in-built UV meters. This enables treatment to be entered in dose units; the machine then terminates the treatment when the dose reaches the preset level. The benefits of using in-built meters are that changes in lamp output over time are compensated for and there is less opportunity for human error in calculating the required treatment time for a given dose. The disadvantage is that if the meter sensitivity changes (maybe due to dirt covering the detector) or the sensor becomes covered or shadowed by a large patient, then the patient will receive an overexposure. This problem has been partly overcome in some cabin designs that have 2 m and treatment is terminated if the two readings differ by a set amount. Most systems have safety features to limit the maximum time or dose that can be delivered. An external backup alarm (55) can also help to prevent overexposure.

Equipment Quality Control in Phototherapy

It is recommended that a quality control regime is implemented for phototherapy equipment (24). The approach used is different for cabins with in-built meters to those without. If in-built meters are not used then measurement of irradiance is required for calculation of exposure time for a given prescribed dose. The frequency of measurement required will depend on usage, but typically monthly reading are desirable due to lamp ageing. Daily checks are required to identify any tube failures. If in-built meters are used, compensation is automatically made for changes in cabin irradiance. However, the accuracy and reproducibility of the meter needs to be assessed. It is important to note that the irradiance at the patient surface during treatment is less than that

measured in the same position with the cabin unoccupied, because some of the energy received at the patient surface has been reflected from other parts of the cabin by mirrors behind the fluorescent tubes (56). Without the patient present this reflected fraction is greater. The ratio between "direct" (i.e., with patient in cabin) and "indirect" measurements ranges from ~0.65 to 0.90 (57,58), varying with cabin geometry, lamp spectrum, and patient size. The irradiance within a cabin will vary with position, so it is usual to determine the mean irradiance over several positions for an average sized patient; this is termed the Designated Patient Irradiance (DPI). In order to avoid undue staff hazard, the DPI may be measured using the direct method initially (Fig. 14) and subsequently using the indirect method with a correction for the direct–indirect ratio. Automated systems may be employed to move the detector to a number of predetermined positions within the cabin (59,60). The in-built meter reading should agree with the DPI measured using an independent meter to within 20%. Measurement of irradiance from all equipment should be regularly recorded in a standard manner so that trends over time can be assessed. It is important that there is consistency between calibration of equipment used for MED or MPD testing and that used for treatment. The meter used should be correctly calibrated for the lamp being measured and be recalibrated with reference to national standards annually.

SUNBEDS AND TANNING

One could argue that a sunbed is not a medical device, since its primary purpose is cosmetic tanning. However, information on sunbeds is included here for a number of reasons:

- Certain sunbeds could be classed as medical devices according to British Standards.
- There are concerns and considerable media interest regarding the safety of long-term sunbed use that may impact on health services.
- Many people with skin diseases try using sun beds to treat themselves.

Sunbed Lamps

Prior to the mid-1970s, the only artificial way to achieve a tan was using a "sunlamp" at home. These lamps were unfiltered medium or high pressure mercury arc lamps that emitted a wide spectrum including significant UVC and UVB components. Because of the significant short wavelength contribution, exposure times were short (a few minutes) and the incidence of burning was high. In the early 1980s tanning using UVA fluorescent lamps became popular, largely because the tanning industry promoted the idea that UVA tanning was safe (61). Typical UVA lamps used were similar to those used for PUVA therapy (e.g., Philips TL09 lamp) with ~0.7% of the total UVR in the UVB range. In the mid-1980s, a new lamp was introduced for tanning (Philips TL10) that had a spectrum extending from 340 to 400 nm with a UVB component of only 0.05%, but these lamps are not often used today (62). It has been proposed that the "least bad" way to acquire an artificial tan may be to use a lamp that emits UVB in a similar proportion to natural sunlight (63). Philips have recently introduced the Cleo Natural lamp with this in mind (64). A survey of commercial tanning establishments in Scotland (62) found 13 different models from 7 manufacturers. However, the lamps fell into three categories; 81% were type 1 UVA lamps (e.g., Philips Cleo Performance, emitting ~0.7% UVB); 8% were type 2 UVA lamps (e.g., Philips Cleo Professional, emitting ~1.4% UVB); and 11% were filtered high pressure metal halide lamps. Other new higher UVB emitting lamps are now available such as the Philips Cleo Advantage and Swift (64), but surveys to assess the popularity of these lamps have not been reported. Spectra of UVA lamps used for tanning are shown in Fig. 6.

McGinley et al. (62) found that the UVB irradiance from an average sunbed was similar to that of Glasgow, UK, on a sunny summer day, whereas mean UVA irradiances were three or four times higher. They estimated that the total UVR received during 20 sunbed sessions was similar to that received during a week sunbathing in the Mediterranean.

Tanning

In the tanning process, melanocytes are stimulated to produce melanin that is transferred to keratinocytes. It is widely believed that obtaining a tan protects against sunburn. However, the degree of protection may only be equivalent to a sunscreen with a sun protection factor (SPF) of 3 (65) and there is evidence to suggest that a UVA tan provides even less protection than that due to natural sunlight. The action spectra for tanning and erythema are very similar for people who do not tan easily (42). For those who tan easily it is possible to obtain a UVA tan without burning.

Risks of Sunbed Use

Short-term risks of sunbeds include sunburn, itching, skin rashes, nausea conjunctivitis, and photokeratitis (66). Sunbeds can also provoke the common photodermatoses, polymorphic light eruption. A serious concern over long term sunbed use is that of developing skin cancer (65). The erythema, tanning and squamous cell carcinoma action spectra (see Fig. 3) are all similar, and therefore tanning by any part of the UV spectrum carries an approximately equal risk for SCC. There is also now increasing evidence that sunbed use is associated with increased melanoma risk (65).

Legislation

Sunbeds are used by a significant minority of the population and recently they have become popular with young people. British and European standards (67) classify sunbeds according to the erythemally weighted irradiance in the UVA and UVB range. According to this classification, the typical UVA sunbed is a type 3 device and a sunbed

with Philips Natural lamps is a type 4 device (68). Type 4 devices should be marked with the following warning: "to be used following medical advice only". There is, however, no statutory regulation of sunbed use in the United Kingdom. Some countries or U.S.A. states have regulated access to sunbeds to a certain extent (65). Autier (61) proposes strict controls to minimize and regulate sunbed use. However, Diffey (69) believes that outright prohibition is unnecessary since the excess deaths from sunbed use in the United Kingdom is relatively small compared to that from natural UV exposure and other voluntary risks (e.g., smoking and alcohol).

Sunbeds and Home Phototherapy

A survey has shown that many patients use sunbeds at home to treat psoriasis and believe it to be effective (70). However, a randomized placebo controlled trail of UVA sunbed therapy for psoriasis (71) found only a small degree of improvement. A consensus report of the British Photodermatology Group (72) concluded that home phototherapy represented suboptimal treatment with greater risks compared to hospital based treatment. Das et al.(68) carried out a randomized comparison between conventional UVA sunbed lamps and lamps with higher UVB component (Philips Natural). They found no difference in response for equal erythemal doses. However, when equal exposure times were used (similar to that used in a conventional tanning session) significant improvement in psoriasis were seen. The authors conclude that this home sunbed treatment is likely to be less effective compared to conventional hospital treatment, but may be appropriate for some patients who cannot access hospital therapy. Several groups have investigated installing conventional phototherapy systems in patient's homes. A recent project in Scotland (73) installed narrow-band UVB in patients home and provided them with treatment schedules and training. They concluded that the treatment was effective, safe and cost effective. Overall, home phototherapy probably is worth considering for certain patients and centers with a very rural catchment area.

OTHER USES OF UV IN MEDICINE

Laser Eye Correction

Energy from an ArFl (argon fluoride) excimer laser at 193 nm is very well absorbed by proteins within the cornea. Molecular bonds are broken as the photon energy is greater than the bond strength within these proteins. The process of tissue removal is not like that of a scalpel, or other lasers. Rather, a broad beam is used that removes tissue by a process called "ablative photodecomposition". Since the laser energy is absorbed very near the surface, little tissue damage beyond this region occurs. The degree of tissue removal is computer controlled so that the cornea is reshaped to give the appropriate degree of refractive correction for the patient. Several different effective techniques are used (74,75), Photorefractive Keratectomy (PRK) uses the laser alone to reshape the cornea, whereas Laser-Assisted *In Situ* Keratomileusis (LASIK) is a combined surgical and laser technique. During LASIK surgery, a thin flap on the cornea is cut and folded back and the laser is then used to remove a precise amount of corneal tissue. The LASIK method is currently the dominant procedure and has the advantage of maintaining the central corneal epithelium (76). Current evidence on LASIK suggests that it is efficacious in selected patients with mild or moderate myopia. Evidence is weaker for its efficacy in patients with severe myopia or hyperopia, and there are concerns about the safety of its use in the long term (77). The U.K. National Institute of Clinical Excellence is due to issue guidance on the use of PRK during 2005.

Photopheresis

Photopheresis (also known as extra-corporal photochemotherapy or ECP) was first described by Edelson et al. in 1987 (78) and its use has recently been reviewed by McKenna et al. (79) The technique involves the following process: venous blood is removed from the patient, the white cells are separated and mixed with 8-mythoxypsoralen, these cells are then irradiated using UVA lamps and then returned to the patient. A treatment session takes $\sim$4 h and this is repeated every few weeks for several months. Currently, the technique is expensive and only used on a relatively small number of patients in a limited number of centers. It has been investigated in a wide range of diseases with varying degrees of success. The most promising results have been in the treatment of cutaneous T-cell lymphoma and graft versus host disease. There is some debate about the mechanism of action; it is not likely to be a direct effect on circulating lymphocytes since only a few percent are treated during each session. It is thought that it may work by stimulating the patient's immune response to the malignant lymphocytes.

Blue Light Phototherapy

Blue light phototherapy (50) is used in the treatment of neonatal jaundice. The condition is caused by the inability to clear bilirubin. Bilirubin is formed as the body scavenges iron from heme molecules released during the destruction of red blood cells. Irradiation of bilirubin by blue light causes its conversion to forms that are less toxic and more easily removed from the body. As the liver matures it becomes more able to excrete bilirubin so that phototherapy is only required for a short time. There is uncertainty about the optimum spectral output for lamps used in phototherapy, although it is thought that the most effective wavelength is $\sim$450 nm. Either fluorescent lamps or metal halide lamps are used for blue light phototherapy. The hazards of blue light are almost entirely to the retina. Note, however, that many blue light phototherapy lamps emit radiation in the UV range. Normally, this is removed using an acrylic sheet. However, it has been shown that lamps may emit $\sim$0.2 mW/cm^2 UVR with the shield in place and $\sim$10 times this with the shield absent.

Disinfection and Hygiene

Ultraviolet radiation of wounds and ulcers has been used in the past (80). The success of the treatment relies on the

bactericidal properties of UVR and for this reason it is important that lamps that emit UVC are used. Advances in antibiotic therapy meant that the technique became less common. There is, however, renewed interest in this treatment for methicillin resistant *Staphylococcus aureus* (MRSA) infected wounds (81,82).

UVC irradiation is also used to kill bacteria in the air in operating theaters and there has recently been increased interest in this technique as a method of dealing with deliberate terrorist release of pathogens in buildings (83).

Many flying insects are attracted for UVA radiation (~350 nm). Certain insect killing devices attract insects with UVA, and then electrocute them as they fly toward the source. This method of dealing with flying insects is the technique of choice in many hospital kitchens (84).

Stimulation of Vitamin D Production

Production of vitamin D is the only definite beneficial effect of human UVR exposure. Low level UVR has been used to stimulate vitamin D production in residents of an old people's home (85).

HAZARD ASSESSMENT AND PROTECTION

Exposure Limits and Hazard Assessment of UV Sources

The International Commission on Non-ionizing Radiation Protection has published guidelines on MPE limits for UV sources. The latest revision of the guidelines (6) contains updated information on hazards, but does not change the value of the limits published in 1989 (86). For most individuals, if exposure is below the limits, then they will not experience acute effects and the risks from long-term effects will be at an acceptable level. The limits are based on an envelope action spectra for all acute and chronic deleterious effects on the eye and skin. Certain individuals with disease or sensitized by chemicals or drugs may experience effects at exposure levels below the limits. To be below the MPE there are two criteria that must be satisfied:

1. Total effective spectrally weighted UV radiation (180–400-nm range) incident on the unprotected skin or eye should not exceed 30 J/m^2 over an 8 h period
2. Total unweighted UV radiant exposure in the spectral region 315–400 nm on the unprotected eye should not exceed 10^4 J/m^2 over an 8 h period

The effective spectrally weighted exposure is calculated using the following expression:

$$E_{\text{eff}} = \sum E_\lambda \cdot S_\lambda \cdot \Delta_\lambda$$

Where E_λ is the measured irradiance of the source (W/m^2), S_λ is the relative spectral effectiveness (unitless), and Δ_λ is the wavelength measurement interval. The weighting factors used form a "hazard action spectrum" (6), similar to that used when calculating erythemal effective irradiance. Figure 3 shows a comparison of the two spectra.

The hazard weighted effective irradiance can be measured in two ways: A spectroradiometer can be used to measure the absolute spectral irradiance of the source in question, and the relevant weighting factors can be applied accordingly. A less accurate (but simpler) method is to use a radiometer fitted with an appropriately weighted filter. However, attempts to develop filters that have a good match to the hazard action spectrum have not been particularly successful. Measurements of various UV sources in the author's laboratory with a spectroradiometer and narrow-band meter with hazard weighted filter (International Light, IL1730A UV Actinic radiometer, fitted with an ACT270 filter) are shown in Table 4. While the broad-band meter provides a reasonable approximation for some sources, there are very large errors for others due to the poor matching of the filter characteristics and the hazard action spectrum. It is important to consider both the weighted exposure to the skin and eye and the separate limit for UVA exposure to the eye. As can be seen from the examples in Table 4, the former limit usually dominates, but for the source with high UVA output it is the latter that is more restrictive.

Table 4. UV Hazards Measured in the Author's Laboratory (all at 10 cm)[a]

UV Source	Weighted Irradiance, W/m^2 Measured with Spectroradiometer	Weighted Irradiance, W/m^2 Measured with Radiometer	UVA Unweighted Irradiance, W/m^2 Measured with Spectroradiometer	Time min to Exceed 30 J/m^2 Weighted Irradiance, Skin Hazard	Time min to exceed 10^4 J/m^2 Unweighted UVA Irradiance, Eye Hazard
Bank of six 4 ft TL01 tubes	0.43	0.40	4	1	45
Bank of eight 2 ft TL09 tubes	0.02	0.14	67	30	2
Single 2 ft TL12 tube	0.73	1.1	2	1	80
Single 6 ft UV6 tube	0.22	0.26	4	2	40
Xe Arc lamp	0.20	0.24	2	3	70

[a]The weighted irradiance was obtained either by using a radiometer with a hazard weighted filter (International Light IL1730 with ACTS filter) or by spectroradiometer measurements combined with INIRPC hazard spectrum (6). The times given in the final two columns are those required to exceed to the maximum permissible exposure for eye and skin calculated from the spectroradiometer measurements.

Protection Measures

The National Radiological Protection Board (NRPB) in the United Kingdom has issued advice on protection against UVR (5). For workers exposed to artificial sources, administrative, engineering, and protective measures should be in place to ensure limits are not exceeded. For other situations, such as members of the public exposed to the sun, strict observance of limits is not practicable. If possible, exposure to artificial sources should be limited by engineering controls. For example, phototherapy cabins are usually fitted with interlocks so that the lamps go out when the door is opened. It is essential that staff working with UV sources are well trained and understand the nature of UV hazards. Access to suitable trained Medical Physics staff for advice on UV hazards is also important. Warning signs and access restrictions should be used where exposure above the MPE level is possible. Personal protective equipment (such as face shield and gloves) should be used if exposure to UVR above the MPE cannot be avoided.

The British Standards Institute provide information on the performance specifications of UV filters for personal eye protection (87). Percentage UV transmission should be $<0.0003\%$ at 313 nm and $<7\%$ at 365 nm and luminous transmission should be between 43 and 58%.

Patient Safety

Good staff training is vital to minimize risks of patient UV overexposure during phototherapy. There must be adequate protection against electrical hazards. Patients should not come into contact with bare lamps. There is a risk of injury if the patient falls against the lamps and it is now common practice to place a UV transmitting sheet between the lamps and the patient.

Because psoralen is deposited in the lens of the eye, there is a risk of cataract induction if the eye is exposed to UVA in the 12 h following psoralen ingestion. Consequently, the patient should avoid unnecessary sunlight exposure for the remainder of the day following PUVA treatment and they should wear appropriate eye protection. These may either be sunglasses, prescription lenses, or clear safety spectacles. Contact lenses even if marked as UV protective may not afford sufficient protection (88). The suitability of UV protective eyewear may be assessed using a spectrophotometer to fully characterise the transmission and comparison with limits suggested by Moseley et al. (89). Alternatively, a simpler approach (50) is to measure the lens transmission of radiation from PUVA therapy lamps using a hand-held narrow-band meter.

Hazards from Ozone

Ozone is a colorless, toxic irritant gas formed by a photochemical reaction between short wavelength UVR and oxygen in the air. It is possible to find ozone near UV lamps, especially where radiation <250 nm is transmitted through the envelope of the lamp. If ozone production is suspected, then adequate ventilation should be provided to remove the hazard.

ACKNOWLEDGMENTS

This article is based on that in the first edition by Prof. Brian Diffey and certain sections (e.g., the historical introduction) are taken directly from that source. Brian Diffey provided spectra for some lamps and the skin cancer action spectrum. Brian Diffey and Peter Farr were both very helpful in discussing certain parts of the text and providing useful references. Some of the practical measurements were made by John Frame. Steve Burnet and Muzz Hanliffa provided some of the pictures.

BIBLIOGRAPHY

1. McGregor JM. The history of human photobiology. In: Hawk JLM, editor. Photodermatology. London: Arnold; 1999.
2. Sollux publishing. Actinotherapy techniques. Slough: Sollux Publishing; 1933.
3. Diffey BL. What is light? Photodermatol Photoimmunol Photomed 2002;18(2):68–74.
4. National Radiological Protection Board, Health effects from ultraviolet radiation. Didcot: NRPB; 2002.
5. NRPB. Advice on protection against Ultraviolet radiation. Didcot: NRPB; 2002.
6. International Commission on Non-Ionizing Radiation Protection. Guidelines on limits of exposure to ultraviolet radiation of wavelengths between 180 nm and 400 nm (incoherent optical radiation). Health Phys 2004;87(2):171–186.
7. Harrison GI, Young AR. Ultraviolet radiation-induced erythema in human skin. Methods 2002;28(1):14–19.
8. McKinley A, Diffey B. A reference action spectrum for Ultraviolet induced erythema in human skin. CIE J 1987;6:17–22.
9. CIE. Erythemal reference action spectrum and standard erythemal dose (CIE S 007/E-1998); 1998.
10. CIE. Report 138/2: Action spectrum for photocarcinogenesis (non-melanoma skin cancers). Wien: CIE; 2000.
11. Diffey BL, Oliver RJ, Farr PM. A portable instrument for quantifying erythema induced by ultraviolet radiation. Br J Dermatol 1984;111(6):663–672.
12. Gordon PM, Saunders PJ, Diffey BL, Farr PM. Phototesting prior to narrowband (TL-01) ultraviolet B phototherapy. Br J Dermatol 1998;139(5):811–814.
13. Waterston K, Naysmith L, Rees JL. Physiological variation in the erythemal response to ultraviolet radiation and photoadaptation. J Invest Dermatol 2004: 123(5):958–964.
14. Diffey BL. The future incidence of cutaneous melanoma within the UK. Br J Dermatol 2004;151(4):868–872.
15. Diffey B. Human exposure to solar ultraviolet radiation. J Cos Dermatol 2002;1:124–130.
16. Diffey BL. Human exposure to ultraviolet radiation. In: Hawk JLM, editor. Photodermatology. London: Arnold; 1999.
17. Diffey BL. Sources and measurement of ultraviolet radiation. Methods 2002;28(1):4–13.
18. World Health Organisation. Intersun, the global UV project; 2003.
19. Hersh P, Carr J. Excimer laser photorefractive keratectomy. Ophthalmic Practice 1995;13:126–133.
20. Anders A, Altheide HJ, Knalmann M, Tronnier H. Action spectrum for erythema in humans investigated with dye lasers. Photochem Photobiol 1995;61(2):200–205.
21. World Health Organization. Global solar UV index: a practical guide. 2002.
22. Wilson AD. Optical radiation detectors. In: Diffey B, editor. Radiation measurement on photobiology. London: Academic Press; 1989.

23. Oliver H, Moseley H. The use of diode array spectroradiometers for dosimetry in phototherapy. Phys Med Biol 2002;47:4411–4421.
24. Taylor DK, Anstey AV, Coleman AJ, Diffey BL, Farr PM, Ferguson J, Ibbotson S, Langmack K, Lloyd JJ, McCann P, Martin CJ, Menage Hdu P, Moseley H, Murphy G, Pye SD, Rhodes LE, Rogers S. Guidelines for dosimetry and calibration in ultraviolet radiation therapy: a report of a British Photodermatology Group workshop. Br J Dermatol 2002; 146(5):755–763.
25. Coleman AJ, Collins M, Saunders JE. Traceable calibration of ultraviolet meters used with broadband, extended sources. Phys Med Biol 2000;45(1):185–196.
26. Lloyd JJ. Variation in calibration of hand-held ultraviolet (UV) meters for psoralen plus UVA and narrow-band UVB phototherapy. Br J Dermatol 2004;150(6):1162–1166.
27. Diffey BL, Roelandts R. Status of ultraviolet A dosimetry in methoxsalen plus ultraviolet A therapy. J Am Acad Dermatol 1986;15(6):1209–1213.
28. Diffey B. Ultraviolet radiation dosimetery with polysulphone film. In: Diffey B, editor. Radiation Measurement in Photobiology. London: Academic Press; 1989.
29. Cabaj A, Sommer R. Measurement of Ultraviolet radiation with biological dosemeters. Rad Protection Dosimetry 2000;91:139–142.
30. Ell C. Improving endoscopic resolution and sampling: fluorescence techniques. Gut 2003;52(Suppl IV): iv30–iv33.
31. Ferguson J. Diagnosis and treatment of the common idiopathic photodermatoses. Aust J Dermatol 2003;44(2):90–96.
32. Ferguson J, Ibbotson S. The idiopathic photodermatoses. Semin Cutan Med Surg 1999;18(4):257–273.
33. Bilsland D, Diffey BL, Farr PM, Ferguson J, Gibbs NK, Hawk JL, Johnson BE, Magnus IA, Moseley H, Murphy GM. Diagnostic phototesting in the United Kingdom. British Photodermatology Group. Br J Dermatol 1992;127(3):297–299.
34. van de Pas CB, Hawk JL, Young AR, Walker SL. An optimal method for experimental provocation of polymorphic light eruption. Arch Dermatol 2004;140(3):286–292.
35. Das S, Lloyd JJ, Walshaw D, Farr PM. Provocation testing in polymorphic light eruption using fluorescent ultraviolet (UV) A and UVB lamps. Br J Dermatol 2004;151(5):1066–1070.
36. Farr PM. Irradiation testing of the skin. In: Hawk JLM, editor. Photodermatology. London: Arnold; 1999.
37. Diffey BL, Farr PM. The normal range in diagnostic phototesting. Br J Dermatol 1989;120(4):517–524.
38. Bruynzeel DP, Ferguson J, Andersen K, Goncalo M, English J, Goossens A, Holzle E, Ibbotson SH, Lecha M, Lehmann P, Leonard F, Moseley H, Pigatto P, Tanew A. Photopatch testing: a consensus methodology for Europe. J Eur Acad Dermatol Venereol 2004;18(6):679–682.
39. Dootson G, Norris PG, Gibson CJ, Diffey BL., The practice of ultraviolet phototherapy in the United Kingdom. Br J Dermatol 1994;131(6):873–877.
40. Ibbotson SH, Bilsland D, Cox NH, Dawe RS, Diffey B, Edwards C, Farr PM, Ferguson J, Hart G, Hawk J, Lloyd J, Martin C, Moseley H, McKenna K, Rhodes LE, Taylor DK. An update and guidance on narrowband ultraviolet B phototherapy: a British Photodermatology Group Workshop Report. Br J Dermatol 2004;151(2):283–297.
41. Green C, Diffey BL, Hawk JL. Ultraviolet radiation in the treatment of skin disease. Phys Med Biol 1992;37(1):1–20.
42. Parrish JA, Jaenicke KF. Action spectrum for phototherapy of psoriasis. J Invest Dermatol 1981;76(5):359–362.
43. Ortel B, Honigsmann H. Phototherapy and Photochemotherapy. In: Hawk JLM, editor. Photodermatology. London: Arnold; 1999.
44. Halpern SM, Anstey AV, Dawe RS, Diffey BL, Farr PM, Ferguson J, Hawk JL, Ibbotson S, McGregor JM, Murphy GM, Thomas SE, Rhodes LE. Guidelines for topical PUVA: a report of a workshop of the British photodermatology group. Br J Dermatol 2000;142(1):22–31.
45. British Photodermatology Group, British Photodermatology Group guidelines for PUVA. Br J Dermatol 1994;130(2):246–255.
46. Farr PM, Diffey BL, Higgins EM, Matthews JN. The action spectrum between 320 and 400 nm for clearance of psoriasis by psoralen photochemotherapy. Br J Dermatol 1991;124(5):443–448.
47. de Berker DA, Sakuntabhai A, Diffey BL, Matthews JN, Farr PM. Comparison of psoralen-UVB and psoralen-UVA photochemotherapy in the treatment of psoriasis. J Am Acad Dermatol 1997;36(4):577–581.
48. Mackenzie LA. The analysis of the ultraviolet radiation doses required to produce erythemal responses in normal skin. Br J Dermatol 1983;108(1):1–9.
49. Diffey BL, De Berker DA, Saunders PJ, Farr PM. A device for phototesting patients before PUVA therapy. Br J Dermatol 1993;129(6):700–703.
50. Diffey BL, Hart G. Ultraviolet and blue-light phototherapy—principles, sources, dosimetry and safety. New York: 1997.
51. Dawe RS. Ultraviolet A1 phototherapy. Br J Dermatol 2003;148(4):626–637.
52. Farr PM, Besag JE, Diffey BL. The time course of UVB and UVC erythema. J Invest Dermatol 1988;91(5):454–457.
53. Das S, Lloyd JJ, Farr PM. Similar dose-response and persistence of erythema with broad-band and narrow-band ultraviolet B lamps. J Invest Dermatol 2001;117(5):1318–1321.
54. Ibbotson SH, Farr PM. The Time-Course of Psoralen Ultraviolet A (PUVA) Erythema. J Invest Dermatol 1999;113(3): 346–350.
55. Allan W, Diffey BL. A device for minimizing the risk of overexposure of patients undergoing phototherapy. Photodermatol Photoimmunol Photomed 2002;18(4):199–200.
56. Langmack KA. An insight into the contributions of self-shielding and lamp reflectors to patient exposure in phototherapy units. Phys Med Biol 1998;43(1):207–214.
57. Moseley H. Scottish UV dosimetry guidelines, "ScUViDo". Photodermatol Photoimmunol Photomed 2001;17(5):230–233.
58. Martin CJ, Clouting H, Aitken A. A study of the correction factor for ultraviolet phototherapy dose measurements made by the indirect method. Br J Dermatol 2003;149(6):1227–1231.
59. Currie GD, Evans AL, Smith D, Martin CJ, McCalman S, Bilsland D. An automated dosimetry system for testing whole-body ultraviolet phototherapy cabinets. Phys Med Biol 2001;46(2):333–346.
60. Evans AL, Martin CJ, Smith DC, Currie GD, McCalman S, Bilsland D, Dunn S. Instrument for scanning the angular variation of irradiance in ultraviolet phototherapy cabinets. J Med Eng Technol 2002;26(3):126–131.
61. Autier P. Perspectives in melanoma prevention: the case of sunbeds. Eur J Cancer 2004;40(16):2367–2376.
62. McGinley J, Martin CJ, MacKie RM. Sunbeds in current use in Scotland: a survey of their output and patterns of use. Br J Dermatol 1998;139(3):428–438.
63. Diffey BL, Farr PM. Tanning with UVB or UVA: an appraisal of risks. J Photochem Photobiol B 1991;8(2):219–223.
64. Phillips. Welcome to Phillips Tanning; http//www.lighting.philips.com 2004.
65. Young AR. Tanning devices—fast track to skin cancer? Pigment Cell Res 2004;17(1):2–9.
66. Diffey B. Sunbeds: What are they, who uses them and what are the health effect? London: Health Education Authority; 1997.

67. British Standards Institute. Safety of household and similar electrical appliances. Part 2 Section 2.7 Skin Exposure to ultraviolet and infrared radiation (BS EN 60335-2-27: 1997); 1997.
68. Das S, Lloyd JJ, Walshaw D, Diffey BL, Farr PM. Response of psoriasis to sunbed treatment: comparison of conventional ultraviolet A lamps with new higher ultraviolet B-emitting lamps. Br J Dermatol 2002;147(5):966–972.
69. Diffey BL. A quantitative estimate of melanoma mortality from ultraviolet A sunbed use in the U.K. Br J Dermatol 2003;149(3):578–581.
70. Turner RJ, Farr PM, Walshaw D. Many patients with psoriasis use sunbeds (letter). Br Med J 1998;317:412.
71. Turner RJ, Walshaw D, Diffey BL, Farr PM. A controlled study of ultraviolet A sunbed treatment of psoriasis. Br J Dermatol 2000;143(5):957–963.
72. Sarkany RP, Anstey A, Diffey BL, Jobling R, Langmack K, McGregor JM, Moseley H, Murphy GM, Rhodes LE, Norris PG. Home phototherapy: report on a workshop of the British Photodermatology Group. December 1996. Br J Dermatol 1999;140(2):195–199.
73. Cameron H, Yule S, Moseley H, Dawe RS, Ferguson J. Taking treatment to the patient: development of a home TL-01 ultraviolet B phototherapy service. Br J Dermatol 2002;147(5):957–965.
74. Wilson SE. Clinical practice. Use of lasers for vision correction of nearsightedness and farsightedness. N Engl J Med 2004;351(5):470–475.
75. Bower KS, Weichel ED, Kim TJ. Overview of refractive surgery. Am Fam Phys 2001;64(7):1183–1190.
76. Ambrosio R, Jr., Wilson S. LASIK vs LASEK vs PRK: advantages and indications. Semin Ophthalmol 2003;18(1):2–10.
77. National Institute of Clinical Excellence Interventional Procedure Guidance 102: Laser in situ keratomileusis for the treatment of refractive errors. 2004.
78. Edelson R, Berger C, Gasparro F, Jegasothy B, Heald P, Wintroub B, Vonderheid E, Knobler R, Wolff K, Plewig G. Treatment of cutaneous T-cell lymphoma by extracorporeal photochemotherapy. Preliminary results. N Engl J Med 1987;316(6):297–303.
79. McKenna KE, Whittaker S, Taylor P, Lloyd J, Ibbotson S, Rhodes LT, Russell-Jones R. Evidence based practice of photopheresis: a report of a workshop of the British Photodermatology Group and UK Skin Lymphoma Group. Br J Dermatol 2005; in press.
80. Roelandts R. The history of phototherapy: something new under the sun?. J Am Acad Dermatol 2002;46(6):926–930.
81. Thai TP, Houghton PE, Campbell KE, Woodbury MG. Ultraviolet light C in the treatment of chronic wounds with MRSA: a case study. Ostomy Wound Manage 2002;48(11):52–60.
82. Conner-Kerr TA, Sullivan PK, Gaillard J, Franklin ME, Jones RM. The effects of ultraviolet radiation on antibiotic-resistant bacteria in vitro. Ostomy Wound Manage 1998;44(10):50–56.
83. Brickner PW, Vincent RL, First M, Nardell E, Murray M, Kaufman W. The application of ultraviolet germicidal irradiation to control transmission of airborne disease: bioterrorism countermeasure. Public Health Rep 2003;118(2): 99–114.
84. Diffey B, Langley FC. IPEM report 49: Evaluation of Ultraviolet Radiation Hazards in Hospitals. London: Institute of Physical Sciences in Medicine; 1986.
85. Chuck A, Todd J, Diffey B. Subliminal ultraviolet-B irradiation for the prevention of vitamin D deficiency in the elderly: a feasibility study. Photoderm Photoimm Photomed 2001;17(4): 168–171.
86. International Non-ionizing Radiation Committee of the International Radiation Protection Association. Proposed change to the IRPA 1985 guidelines on limits of exposure to ultraviolet radiation. Health Phys 1989;56(6):971–972.
87. British Standards institute. BS EN 270, Personal Eye protection—Ultraviolet filters—transmittance requirements and recommended use. 2002.
88. Anstey A, Taylor D, Chalmers I, Ansari E. Ultraviolet radiation-blocking characteristics of contact lenses: relevance to eye protection for psoralen-sensitised patients. Photodermatol Photoimmunol Photomed 1999;15(5):193–197.
89. Moseley H, Cox NH, Mackie RM. The suitability of sunglasses used by patients following ingestion of psoralen. Br J Dermatol 1988;118(2):247–253.

See also FLUORESCENCE MEASUREMENTS; SKIN, BIOMECHANICS OF.

UMBILICAL ARTERY AND VEIN MONITORING. See MONITORING, UMBILICAL ARTERY AND VEIN.

V

VAGINA, EXAMINATION OF. See COLPOSCOPY.

VASCULAR GRAFT PROSTHESIS

KANDICE KOTTKE-MARCHANT
The Cleveland Clinic Foundation
Cleveland, Ohio

COBY LARSEN
Case Western Reserve University
Cleveland, Ohio

INTRODUCTION

The system of blood vessels in the body is crucial to transport cells, oxygen, and nutrients to the vital organs. When injured by diseases, such as atherosclerosis, the blood vessels may become occluded, leading to decreased blood flow and organ damage, or may be weakened and rupture due to aneurysmal dilatation. One method of treatment is to use surgery to bypass diseased blood vessels by using an artificial blood vessel substitute, or vascular graft prosthesis. Materials currently used in vascular graft prostheses include polyethylene terephthalate (Dacron) and expanded polytetrafluoroethyelene (ePTFE). These devices successfully function as vascular grafts in larger diameter applications, but suitable materials for small diameter prostheses (<5 mm) are still being developed. Current trends in the development of small diameter prostheses include surface modification to prevent thrombosis, incorporation of endothelial cells, and a tissue engineering approach with the development of a biological blood vessel based upon a biodegradable material scaffold.

THE CLINICAL NEED FOR VASCULAR GRAFT PROSTHESES

Normal Blood Vessel Anatomy

The largest artery leaving the left ventricle of the heart is the aorta, which is ~20 mm in diameter. The aorta branches into progressively smaller arteries (<5 mm internal diameter) that feed blood and nutrients to organs, such as the brain (carotid arteries), liver (hepatic artery) and kidneys (renal arteries), and to the extremities of the arms (cephalic arteries) and legs (iliac and femoral arteries). The smaller arteries branch further to become smaller arterioles (<1 mm), which branch even further to become the smallest capillaries (<100 μm), through which only single cells can pass. The capillaries are the site of most oxygen and nutrient exchange into the tissues. Capillaries then join together to form venules, small veins, and then larger veins that carry deoxygenated blood back to the lungs and heart. The design of replacement blood vessels should take into account both the structure and function of native vessels for optimal function.

Normal arteries have three different layers: the internal intima, the middle media, and the external adventitia (1). (see Fig. 1). The intima is composed of a layer of endothelial cells, with underlying extracellular matrix proteins delimited by the internal elastic lamina. Endothelial cells are responsible for preventing blood vessel thrombosis under normal physiologic conditions by preventing platelet adhesion, coagulation, and thrombus formation (2,3). (see Fig. 2). This is accomplished by the expression and secretion of antiplatelet agents, such as ecto-ADPase, prostacyclin (PGI_2) and nitric oxide (NO) (2). Endothelial cells express thrombomodulin (TM) and release tissue factor pathway inhibitor (TFPI) to prevent activation of the coagulation cascade. Endothelial cells also express heparan sulfate to bind circulating antithrombin and rapidly inhibit local thrombin formation (2). If a fibrin clot is formed, endothelial cells can release factors that stimulate the fibrinolytic system to degrade the thrombus, such as tissue plasminogen activator (tPA) (2). Upon cytokine or thrombin stimulation, endothelial cells can develop a procoagulant phenotype with expression of tissue factor (TF) and release of vWf, PAI-1, and coagulation factors (6). The intima is present in all vessels, with the smallest capillaries composed of only a single layer of endothelial cells resting upon a thin basement membrane of extracellular matrix proteins.

The media varies in composition from large elastic arteries, such as the aorta, where it is composed of layers of elastic tissue and glycosaminoglycans to the smaller muscular arteries, such as the coronary arteries in the heart, where it is composed predominantly of vascular smooth muscle cells (1). The elastic and mechanical properties of the media allow vessels to contract or relax to maintain the lumenal pressure and volumetric flow rate. Arterioles may have only a single layer of smooth muscle cells and capillaries have no media at all.

The external blood vessel layer, or adventitia, serves to supply nutrients to the cells in the blood vessel itself. It consists of a loose layer of connective tissue, with small arterioles and capillaries (vasa vasorum or "vessels of the vessel") feeding the blood vessel. It also contains a network of nerves that can stimulate medial smooth muscle cell contraction or relaxation. The adventitia is prominent in larger vessels, becoming progressively smaller as the vessel diameter decreases. Arterioles and capillaries lack an adventitia completely.

Coronary Atherosclerosis

Atherosclerotic cardiovascular disease, in its many guises, is the leading cause of mortality in the United States, resulting in coronary artery disease, myocardial infarction, stroke, aneurysms, and peripheral vascular disease. Atherosclerosis is a multifactorial disorder of the arterial vascular tree where lipid dysregulation and vascular

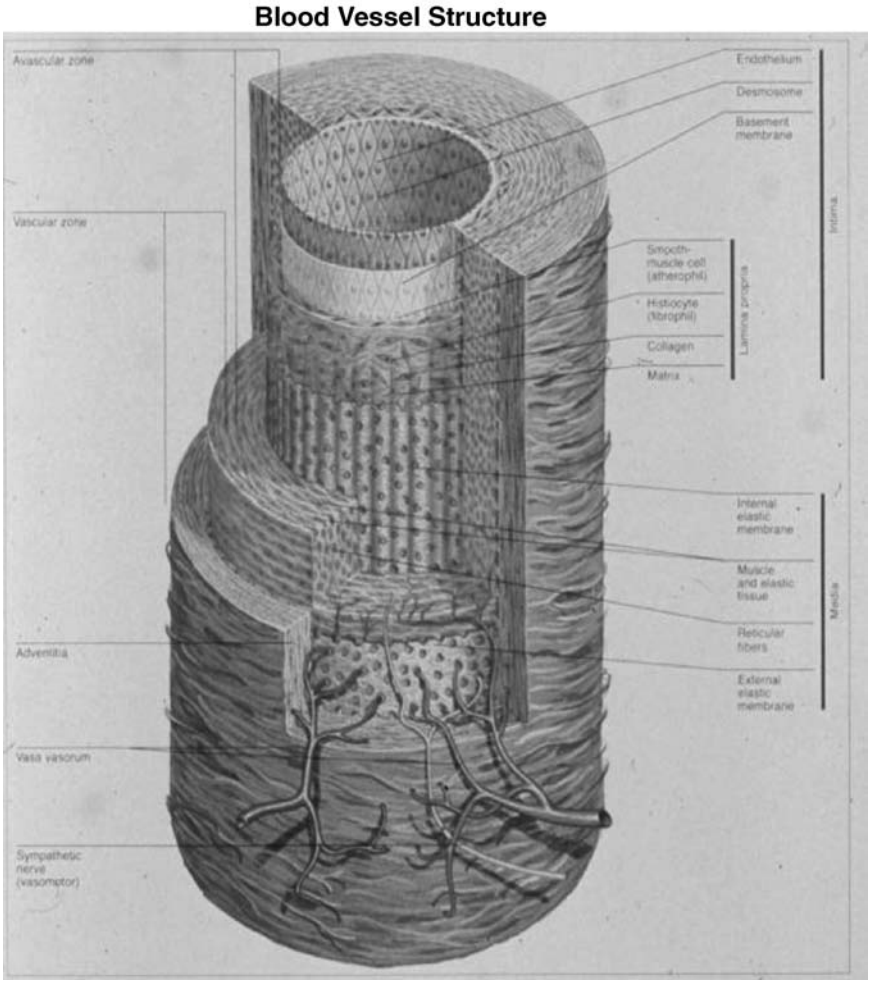


Figure 1. Arteries are composed of three layers: the intima, media, and adventitia. The intima is lined by endothelial cells and bounded by an internal elastic lamina. The media maintains vascular tone and is composed of smooth muscle cells or layers of elastic tissue. The adventitia supports the vessel and provides nutritional and neural stimulation.

inflammation leads to endothelial injury, vascular lipid accumulation, and calcification with the development of lipid-laden deposits, or "plaques" in the vessel wall that eventually lead to stenosis of the vascular lumen (7). The resulting stenosis or narrowing of the blood vessels compromises blood flow through the artery. Rupture of the atherosclerotic plaque with subsequent thrombosis may lead to complete occlusion of the blood vessels. Depending on the location of the artery, this stenosis or occlusion leads to decreased distal blood flow and end-organ dysfunction. Examples include infarction of cardiac muscle with complete coronary artery occlusion or stroke with occlusion of carotid or cerebral arteries.

Despite advances in prevention and early diagnosis, cardiovascular disease claims more lives than the next five leading causes of death, combined (8,9). Coronary atherosclerosis alone, leading to stenosis and occlusion of the small coronary arteries that nourish the cardiac muscle, is the single leading cause of death in American today, responsible for 494,000 deaths in 2002 (8,9). The current modalities for treating coronary artery disease include intravascular angioplasty, stenting, and bypass of the stenotic lesions using saphenous vein or internal mammary artery grafts. Coronary bypass surgery was performed in over one-half million individuals in the United States alone in 2002 (10,11), with the majority of

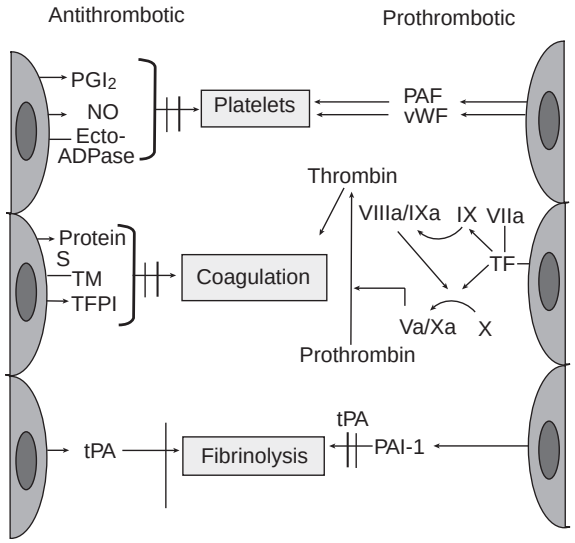


Figure 2. Endothelial cell phenotypes. Endothelial cells can have an antithrombotic (left) or prothrombotic (right) phenotype. (Adapted from Refs. 4 and 5). PGI_2-prostacyclin, NO-nitric oxide, Ecto-ADPase ecto-adenosine diphophatase, TM-thrombomodulin, TFPI-tissue factor pathway inhibitor, tPA-tissue plasminogen activator, PAF-platelet activating factor, vWf-von Willebrand factor, TF-tissue factor, PAI-1-plasminogen activator inhibitor-1.

individuals requiring bypass of three or more vessels. Unfortunately, in a significant number of individuals, autologous veins and arteries are not available, and there is a need for readily accessible, functional, alternative small diameter vascular graft prostheses.

Cerebrovascular Disease

Atherosclerotic vascular disease also affects the cerebral vessels and peripheral arterial tree, leading to stroke, renal failure, and arterial claudication. Stroke is the third leading cause of death, with ~275,000 Americans dying of stroke in 2002 (9). One of the leading causes of stroke is atherosclerotic stenosis of the carotid arteries in the neck, leading to thromboembolic stroke (11). Due to the danger of any further thrombosis or embolism, vascular prostheses have not been used to treat carotid artery disease, and therapy is currently limited largely to carotid endarterectomy, with removal of vascular plaque, and limited use of carotid stents (12). Use of vascular prostheses for cerebrovascular disease will not be feasible until prostheses are developed that pose no risk for distal thromboembolism.

Peripheral Vascular Disease

Peripheral arterial disease, with atherosclerotic occlusion of peripheral arteries predominantly in the lower extremities, leads to claudication and leg pain in ~8–12 million Americans (13,14); while it is not a major cause of mortality, it is associated with significant morbidity (9,13). Current therapies for treating peripheral arterial disease include angioplasty and vascular bypass. Prosthetic grafts have been successfully employed to treat atherosclerosis of the aorto-iliac bifurcation, but saphenous vein bypass is still the most widely used therapy for femoral-popliteal and below-the-knee bypass surgery (15–17). Suitable saphenous veins may not be available in ~15% of patients (17), often due to prior use for coronary artery bypass surgery. There is a clinical need to develop small diameter vascular prostheses that remain patent for long periods to more effectively treat peripheral arterial disease, and provide a viable alternative to saphenous veins.

Aneurysms

An aneurysm is a dilatation of a blood vessel wall that may be circumferential (fusiform aneurysm) or may involve only a portion of the vessel circumference (saccular aneurysm). Aneurysms most frequently affect the abdominal aortic wall and are a significant cause of mortality due to aneurysm rupture, resulting in 1.3% of all deaths (18). Aneurysms can affect other arteries, such as the cerebral arteries, renal arteries, and splenic arteries. Aneurysms usually do not occlude blood flow, but the dilated vessel wall is weakened mechanically and can rupture, leading to extravasation of blood into the surrounding tissues. Infrarenal aortic aneurysms 5.5 and 6.5 cm have annual rupture rates of 11 and 26%, respectively (19). Abdominal aortic aneurysms are usually observed in older patients in association with atherosclerosis, but other factors, such as proteolytic degradation of elastin and collagen by matrix metalloproteinases and plasmin is thought to be contributory (20–22). Therapy has generally been surgical, with the placement of an intralumenal vascular graft (23,24) or an endolumenal stent-graft (25).

Trauma

Vascular trauma from gun shot wounds, knife wounds, motor vehicle accidents or accidental dismemberment may result in laceration or transection of blood vessels. This injury may necessitate replacement with a vascular graft, if the vessel wall cannot be surgically repaired or an autologous saphenous vein is not available (26). Traumatic injury of large diameter vessels, such as the pulmonary artery or aorta usually requires a synthetic vascular graft or patch, as autologous tissue of large diameter is not available. One unique consideration with vascular trauma is bacterial contamination due to open wounds or colon injury, and remote bypass grafts through clean subcutaneous tissues may be desirable. In cases of dismemberment, the blood vessels usually retract into the severed tissue, and microvascular repair methods are usually required to reestablish blood flow (27). This is also a scenario where an off-the-shelf small diameter vascular prosthesis would be desirable.

THE HISTORY OF VASCULAR REPLACEMENTS

Arterial aneurysms have been noted back to antiquity, when the Roman historian Antyllus first reported their tendency to rupture in the second century AD (28). Treatment dates back to 1684, when Moore attempted to induce thrombosis of an aneurysm by introducing large masses of intralumenal wires (23,29). The first reported surgical therapy was ligation of the abdominal aorta proximal to a leaking iliac artery aneurysm by Sir Astley Cooper in 1817 (28). Aneurysmectomy, with excision of the aneurysm and vascular repair was reported by Cooley and DeBakey in 1952 (4). Early vascular grafts employed aortic wrapping or the use of preserved or freeze-dried homograft replacement (30,31). Use of an intralumenal graft replacement was instituted by Creech in 1966, and this remains the standard treatment today (23). In 1991, Parodi introduced a new technique of treating abdominal aortic aneurysms with a stent-graft that could be inserted endolumenally through an incision in the femoral artery (25).

The use of synthetic materials as vascular substitutes for aneurysms and other vascular bypass grafts was initiated in 1952 with the use of Vinyon-N cloth as an arterial conduit (5). This was followed by the use of polyethylene terephthalate (Dacron) as a vascular conduit in 1957 (32,33). In 1969, Gore patented a microexpanded ePTFE, which was first used as an arterial conduit in 1973 (34,35). Umbilical veins were further developed with use of glutaraldehyde cross-linking and an external wrapping of polyester mesh (36). These synthetic prostheses showed some success in large diameter vascular applications, but were plagued by thrombosis and occlusion due to intimal hyperplasia in small diameter (<5 mm) applications. Lining of synthetic vascular prostheses with endothelial cells to improve thromboresistance was first described by Herring in 1978 (37). Development of a blood

vessel by in vitro culture of biologic tissues (i.e. tissue engineering) was first described in 1986, with a multilayer structure of collagen and vascular cells (38).

CURRENT VASCULAR GRAFT PROSTHESES

Vein Grafts

The saphenous vein is a long vein arising in the calf and thigh with extensive collateral circulation. It has been utilized extensively as a vascular graft due to its long size and diameter that is similar to coronary arteries and smaller arteries in the leg, such as the femoral or popliteal arteries. The use of veins as bypass conduits to surgically treat coronary atherosclerosis was first described by Kunlin in 1949 (39), and this procedure remains the standard method of coronary artery bypass surgery to this day. For coronary bypass, saphenous veins are harvested surgically or endoscopically and the quality of the veins assessed by inflation, with valves stripped prior to use, if necessary (40). Current coronary artery bypass techniques using saphenous veins have an occlusion rate of 10–15% at 1 year after surgery, increasing to 30–40% at 10 years after surgery (41). Improved patency has been observed using arterial conduits, such as the internal mammary artery and radial artery grafts (42).

When used as peripheral vascular bypass grafts for the femoral or popliteal arteries in the leg, saphenous veins can be dissected and the direction reversed to allow flow through the valves, but they can also be utilized *In situ* and the valves stripped to allow retrograde flow (43). It is thought that the *In situ* graft should have lesser disruption to the vein endothelium and adventitia compared to the reversed vein conduit (44), but 2 year clinical patency rates for femoropopliteal bypass are similar at ~82% for both procedures (45). Not all patients have an appropriate saphenous vein for either cardiac or peripheral bypass surgery. This is most frequently due to prior surgical harvest, but also may be from poor vein quality due to insufficiently small diameter or increased branching. In these patients, suitable synthetic vascular conduits are desirable.

Dacron

Dacron is a polyester fabric, polyethylene terephthalate. Clinically utilized Dacron vascular prostheses are manufactured as textiles, either knitted or woven (34) (Fig. 3). The original design of a Dacron vascular graft as a textile versus a solid tube was an attempt to allow tissue healing of the prosthesis through the pores of the material. The knitted prostheses have higher porosity than the more densely fabricated woven prostheses and require preclotting or use of a surface coating, such as gelatin, collagen, fibrin glue, or albumin, to prevent translumenal leaking after implantation (34,46,47) (Fig. 4). The woven grafts are most commonly used in the thoracic aorta and for ruptured abdominal aortic aneurysms to minimize blood loss (43). Most Dacron graft prostheses are crimped like a soda straw to improve graft flexibility and decrease lateral compressibility (Fig. 4). An alternative to crimping that results in a thinner prosthesis wall is the use of an external spiral winding of a stiff polymer, such as poly(propylene). In a

Figure 3. Scanning electron micrograph (SEM) of knitted Dacron vascular graft prostheses. The knitted prostheses are of lower porosity than the woven Dacron prostheses and must be sealed by preclotting prior to implantation. SEM, Original Magnification 1000×.

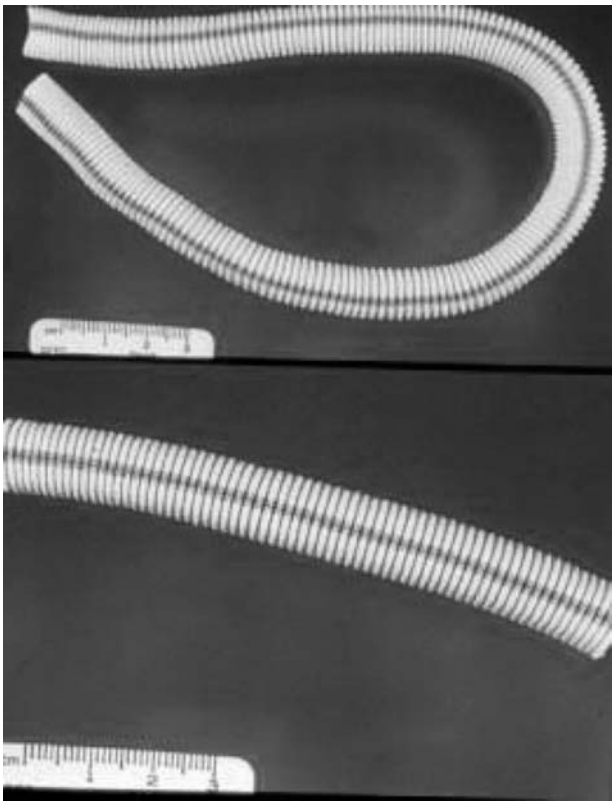


Figure 4. (top) A typical Dacron vascular graft prosthesis. Notice the crimped appearance of the prosthesis, which is utilized for flexibility and lateral compression-resistance. The blue line is to facilitate implantation without twisting, that would lead to kinking (bottom). An albumin-coated Dacron prosthesis. The albumin coating is used to seal the graft interstices and eliminate the need to preclot the graft prior to implantation. This type of coating often fails due to flaking and removal from the surface.

further attempt to foster tissue incorporation and vascular graft healing, some prostheses employ a velour surface, where thousands of individual fiber loops are pulled almost perpendicularly from the material, resulting in a surface that has a markedly increased surface area (48,49). Lumenal velours are employed to trap fibrin and platelets, while external velours are used to increase adventitial healing and cellular incorporation (50).

Dacron vascular grafts have most successfully been employed in the surgical treatment of thoraco-abdominal and abdominal aortic aneurysms, where they are manufactured in a both a tubular and bifurcated aorto-bifemoral configuration. They have also been utilized for longer bypass applications, such as axillary–femoral or femoral–femoral surgery. In these longer length applications, the reinforced, eternally wound prostheses are typically utilized to prevent kinking when transversing joints. The successful treatment of abdominal aortic aneurysms with Dacron vascular grafts is proportional to the size of the aneurysm; for smaller aneurysms, the survival rates for surgical and nonsurgical populations is similar (51). The 5 year patency rates are 93% for aortic bifurcation grafts (52), with similar patency for knitted and woven grafts (53).

Dacron has not been widely used for lower extremity bypass, especially when the bypass entails vessels below the knee, due to poor patency rates. However, a recent study utilizing heparin-bonded Dacron showed favorable patency rates for femoro–popliteal bypass grafting for modified Dacron versus ePTFE (55% vs. 42%; $p < 0.044$) at 3 years, with similar patency rates at 5 years (45% vs. 35%; $p < 0.055$) (54).

Expanded Polytetrafluoroethylene

Polytetrafluorethylene (PTFE–Teflon) is a chemically inert, hydrophobic polymer that resists long-term *In vivo* degradation. As mentioned above, PTFE has been manufactured in a microporous fabric for vascular grafts (ePTFE). The ePTFE is manufactured by a heating, stretching, and extruding process to form solid nodes of PTFE separated by many thin fibrils, with an internodal distance of $\sim 30\,\mu m$ (34) (see Fig. 5). This same material, known as Goretex, is also popular for waterproof sporting applications, such as jackets and shoes, as the material hydrophobicity prevents gross water flux while allowing air transport. The same material properties that make ePTFE attractive for sporting materials make it attractive as a vascular graft. The hydrophobicity is designed to prevent surface thrombosis, while the microporous texture prevents translumenal graft leaking, while encouraging cell incorporation and healing. In applications where increased structural stability is required, an external winding or wrap of polypropylene is often employed.

Patency rates of ePTFE for aorto-bifemoral bifurcation grafts are 95% at 5 years, similar to Dacron (52,55). In extraanatomic bypass, ePTFE has shown a cumulative patency of 83% for femoro–femoral bypass and 75% for axillo–femoral bypass at 5 years (56). In lower extremity revascularization, saphenous vein bypass remains the graft of choice for below-knee applications. However, there is still controversy regarding the optimal material for above-knee femoral–popliteal bypass. Many studies report five year patency rates <50% for ePTFE in femoral–popliteal bypass (54,57), with 5 and 10 year saphenous vein patency rates at 77 and 50% (58,59). However, in a retrospective study, the primary patency rate at 4 years was not significantly better for vein bypass (82.2%) versus ePTFE (80.6%) (60).

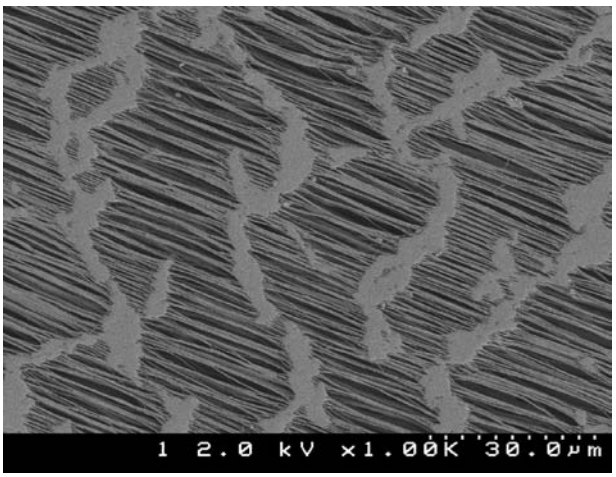


Figure 5. An SEM of expanded ePTFE vascular graft. Note the solid nodes of ePTFE separated by thin fibrils. The typical internodal distance of most clinical prostheses is $30\,\mu m$. Due to the hydrophobic nature of ePTFE, this structure makes the material microporous, allowing limited cell migration, but limiting bulk water leakage. Original magnification 1000×.

A somewhat unique use for ePTFE grafts is the arteriovenous graft used to provide vascular access to patients with renal failure undergoing hemodialysis. These grafts are interposed between the radial artery and cephalic vein in the wrist (61). Other options are the creation of a direct arterio-venous fistula without the use of a graft, and a central venous catheter. Primary patency of arteriovenous grafts is inferior to that of fistulas, irrespective of graft material or postoperative treatment with anticoagulants (62–64).

Endovascular Stent-Grafts

Historically, surgical repair of abdominal aortic aneurysms required major abdominal surgery with its attendant morbidity. For this reason, a procedure to repair aneurysms endolumenally from within the vasculature was attractive. In 1991, Parodi (25) reported the first successful use of a combined stent-graft that could be placed into the aortic aneurysm endolumenally through a femoral artery catheter without open surgery. During the past decade, many different stentgrafts have been developed for this application. They typically are composed of a fabric graft (Dacron or ePTFE) and an expandable metallic skeleton (23,65). They are usually inserted in a closed configuration and then opened via a balloon catheter. The stent is typically held in place with proximal and distal hooks. The metallic

skeletons employed are stainless steel, Elgiloy (nickel, cobalt, and chromium alloy) and Nitinol (nickel titanium alloy).

The stent-grafts may have a lower operative mortality and less frequent complications than open repair (18). However, two large European registries indicate a 3% yearly failure rate for endovascular repair versus 0.3% for open repair (66,67). Failure mechanisms unique to endovascular stent-grafts include endoleak, with blood flow between the exterior of the graft and the lumen of the aneurysm. There are also difficulties with stent-graft migration due to movement of the anchoring hooks (65). Some devices have problems with durability, due to repeated rubbing of the graft material against stent metal, with either fabric degradation or corrosion or stress cracking of the metal (68). In other patients, the aneurysm may continue to expand despite the presence of the stent-graft, termed endotension, due to transmission of lateral wall pressure through the graft.

Vascular Graft Healing

Implantation of synthetic vascular grafts that are not preclotted is associated with rapid protein adsorption, followed by platelet adhesion, inflammatory cell adhesion and variable degrees of fibrin formation (69,70). The fibrin formation is usually limited to a thin layer on the lumenal surface, but this is variable depending on the diameter of the graft. The fibrin layer also includes entrapped macrophages and neutrophils, and reaches a maximal thickness ~2 weeks postimplantation (43). In preclotted Dacron grafts, there is already a fibrin–platelet layer in the graft interstices created by the preclotting process. This may be associated with increased platelet adhesion upon implantation (71). Later cellular healing may come from blood-borne cells, migration of cells from the anastomosis, or in-growth of cells from the adventitia. This may or may not be accompanied by endothelial growth and development of a neo-intima, either due to migration from the anastomoses or implantation of circulating endothelial progenitor cells from the blood stream (72). In the absence of an endothelial layer, the fibrin matrix and inflammatory cells forms a pseudointima. There also may be an in-growth of smooth muscle cells and fibroblasts from the adventitia of the graft, depending on the porosity of the graft material. The adventitia is often associated with fibrosis and a foreign body giant cell response (see Fig. 6).

Most ePTFE grafts implanted in humans do not develop an endothelialized lumen spontaneously (73), and endothelialized pannus in-growth from the anastomosis is usually limited to only a few centimeters. Similarly, Dacron grafts usually do not re-endothelialize, but typically maintain a compact fibrin layer at the lumen. In lower porosity grafts, the graft interstices usually remain acellular. In ePTFE grafts and higher porosity Dacron prostheses, the interstices become populated with fibroblasts, macrophages, and less frequently, smooth muscle cells. Neovascularization, with new capillary in-growth, is rare.

Failure Mechanisms

The healing of implanted vascular grafts is not optimal due to lack of endothelialization or neovascularization. This may lead to clinical complications, such as thrombosis, intimal hyperplasia, and infection. Failure may be acute, midterm, or late (74). Early or acute failure is usually due to technical surgical problems or acute thrombosis. Early infection may be tied to contamination during surgery. Midterm failure (3 months to 2 years) may be due to cellular proliferation and intimal hyperplasia or infection. Late failure may be due to development of atherosclerotic lesions in the graft (75).

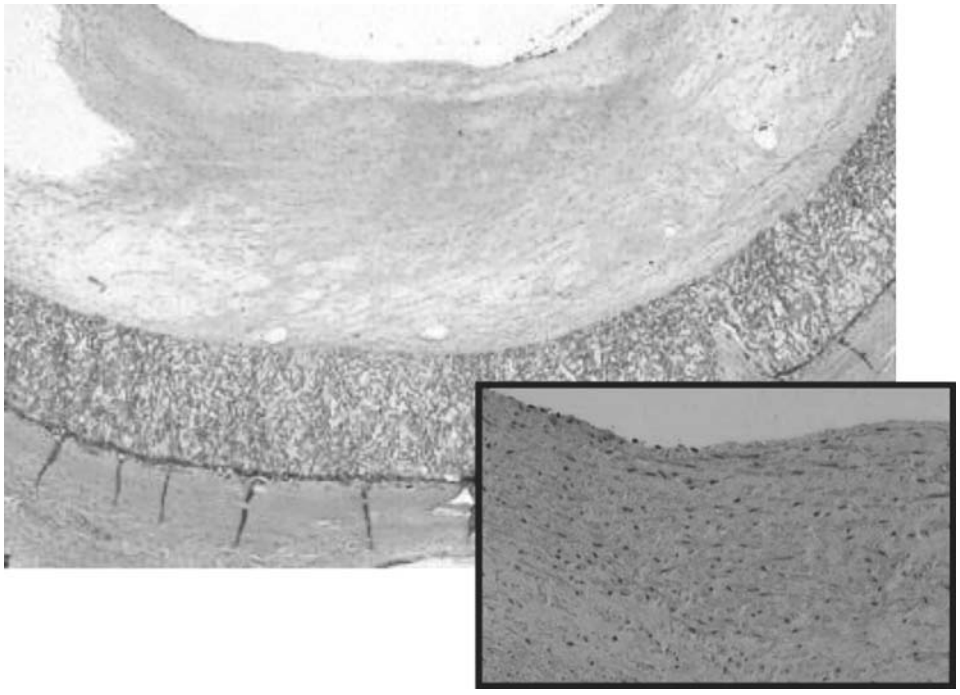


Figure 6. Healing of an ePTFE vascular graft. Note the cellular pseudointima without an obvious endothelial lining. The graft itself is sparsely populated with cells. The adventitial layer is dense compact fibrous tissue with a lining of multinucleated giant cells. Hematoxylin and eosin. Original magnification, 10×. Inset, 20×.

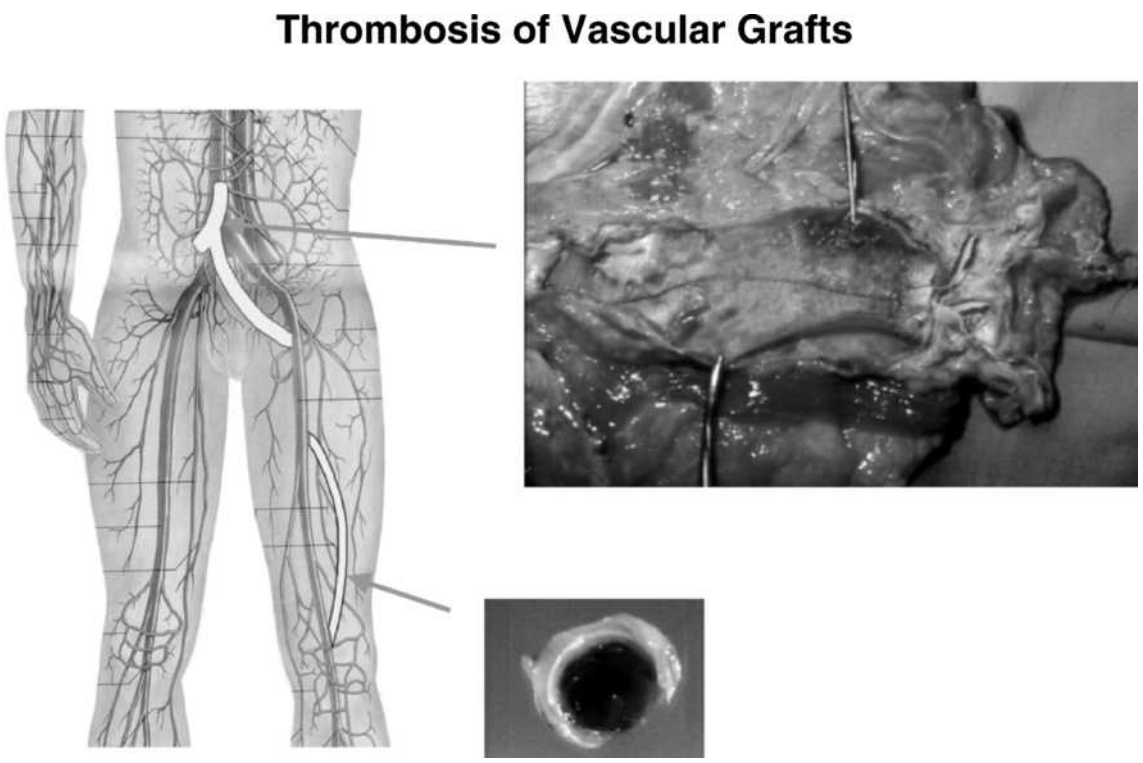


Figure 7. Thrombosis of a Dacron vascular graft in the aorta. In such a large diameter prosthesis, this degree of lumenal thrombosis was of little clinical consequence. A similar degree of thrombosis in a small diameter vessel could result in complete occlusion.

The rapid adhesion of platelets and development of a lumenal fibrin layer can lead to thrombosis (Fig. 7). Thrombosis is a leading cause of vascular graft failure, especially in smaller diameter prostheses, where it leads to decreased flow or occlusion. It is usually an acute or subacute failure mechanism, but may be a cause of late failure due to thrombosis super-imposed upon stenosis due to other causes of vessel narrowing, such as atherosclerosis or intimal hyperplasia. Antithrombotic therapy, particularly antiplatelet therapy, has been shown to be beneficial in decreasing graft occlusion up to 2 years after surgery (76–78).

The majority of stenotic vascular graft failures are due to hyperplasia of the tissues near the anastomosis, particularly the distal anastomosis (79,80), termed anastomotic intimal hyperplasia (81,82). This is particularly a problem with arteriovenous grafts used for dialysis access where the distal intimal hyperplasia may lead to significant stenosis of the lumen (83). The development of intimal hyperplasia is multifactorial, and its etiology is not completely known. It evolves from the tissue healing response at the anastomosis, but inciting factors include chronic inflammation, platelet adhesion with release of platelet derived growth factor (PDGF), vessel wall injury and mechanical factors, such as disturbed hemodynamic flow and compliance mismatch between the native vessel and the more rigid prosthesis (43,83–85). Alteration in cell phenotype due to interaction with the prosthesis or due to disturbed flow may lead to smooth muscle cell proliferation from altered production of basic fibroblast growth factor (bFGF) or PDGF by endothelial cells (86,87) or tumor necrosis factor-alpha (TNFa) by inflammatory cells (88). Indeed, antibodies to bFGF have been shown to decrease smooth muscle cell proliferation in ePTFE grafts experimentally (89). Recent gene-transfer experiments have suggested that increased expression of tPA and increased fibrinolysis may be associated with increased intimal hyperplasia, while increased expression of nitrous oxide synthase (eNOS) and increased expression of the platelet inhibiting molecule nitrous oxide may decrease the hyperplastic response (90).

Infection occurs in 1–6% of arterial vascular graft prostheses (43,91). The sources of infections may be contamination of the prosthesis during implantation or hematogenous seeding from bacteremia. The earlier infections are often due to *Staphylococcus aureus*, while the later infections are often due to a lower virulence organism, such as *Staphylococcus epidermidis* (92). Clinical symptoms of graft infections include fever, leukocytosis, and bacteremia; newer imaging techniques that are able to detect acute inflammation may be useful in diagnosis of graft infections (93). These infections may be resistant to antibiotics due to the lack of vascularity in the graft interstices. Complete or partial graft excision is often necessary for treatment (94). Graft modifications to decrease the risk of infection have included bonding antibiotics to graft surfaces or including antibiotics in the blood used to preclot the graft.

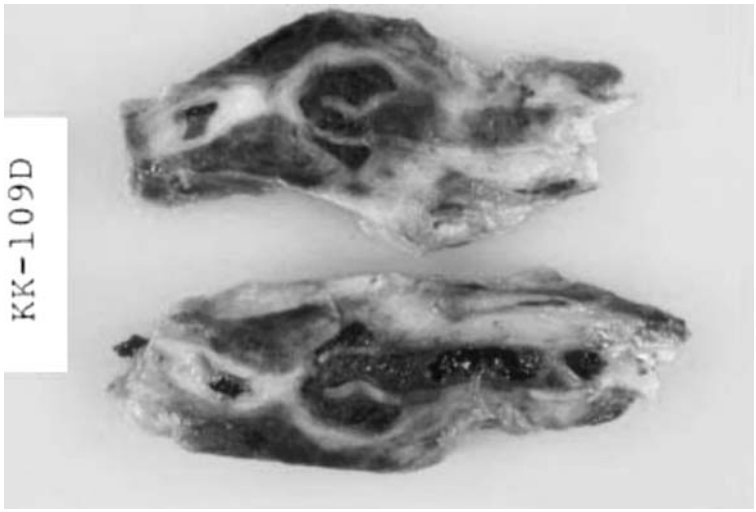


Figure 8. Gross photograph of a pseudoaneurysm at a graft anastamosis. This is not a true aneurysm as there is no failure of the graft material, but a defect between the graft and the native vessel, leading to blood leaking in the interstitial tissues. This type of pseudoaneurysm may cause graft occlusion due to external pressure from the leaking blood.

Material failure of Dacron or ePTFE grafts has been quite uncommon. Neither are substantially susceptible to *In vivo* degradation, although Dacron grafts may expand slightly *In vivo* due to flattening of the crimps or relaxation of the knitted or woven structures. Suture failure and blood leakage at the anastomosis with pseudoaneurysm formation has also been described (Fig. 8).

NEW DEVELOPMENTS IN VASCULAR GRAFT PROSTHESES

Surface Modification

Attachment of an anticoagulant, such as heparin or dipyridamole, to the graft lumen has been investigated as a method to improve small-diameter graft patency (76,95). As an example, Lin et al. (96) modified the lumenal surface of a standard ePTFE with an aldehyde-modified heparin bound to layers of crosslinked polyethyleneimine (PEI), Carmeda bioactive surface (CBAS), and dextransulfate. This graft (4 mm ID), as an exteriorized baboon femoral arteriovenous shunt, showed an 80% reduction in 1–4 h platelet deposition. When compared to contralateral uncoated grafts in a 4 week baboon aortoiliac bypass model, there was significant reduction in neointimal cell proliferation and area at both proximal (0.26 vs. 0.56 mm^2) and distal (0.29 vs. 0.63 mm^2) anastomoses for the heparincoated grafts (96). One major concern with this strategy is the duration of heparin function. Delamination of portions of the layered coating or the presence of physical barriers deposited on the graft surface in the form of blood components would lead to failure for this type of approach. A heparin-bonded Dacron graft is currently available on the European market, with the heparin attached using a tridodecil-methyl-ammonium chloride (TMAC) and the outer portion of the graft coated with collagen to decrease porosity (97). This heparin-bonded Dacron had similar 5 year patency rates to ePTFE for femoral-popliteal bypass grafting. Other surface modifications for ePTFE grafts have included carbon coating to decrease thrombogenicity (98,99), or incorporation of growth factors to facilitate the healing process (100,101) and coating with polyurethane to enhance endothelialization (102).

Newer Biomaterials

There has been some interest in polyurethane elastomers as vascular graft materials, in an effort to develop radially compliant vascular grafts (34,85). Initial vascular grafts with polyester and polyether-based polyurethanes resulted in failure due to material degradation (103). Vectra, a polyetherurethaneurea vascular access graft, showed 12 month patency similar to ePTFE (104), and received U.S. Food and Drug Administration (FDA) clearance in 2000. More recently developed polyurethanes for vascular grafts utilize polycarbonate-based polyurethanes to impart hydrolytic and oxidative stability (85,105). One of these prostheses, Myolink, is available commercially for hemodialysis access. Some experimental hybrid polyurethane grafts have been developed in which the polyurethane is coupled with a gelatin–heparin matrix and reinforced with Dacron mesh (106). Other hybrid polyurethane grafts include a polyurethane with a biodegradable poly(ethylene glycol)–poly(lactic acid) copolymer coating (107). Polyurethanes have recently been modified to incorporate a diazeniumdiolatemodified nitric oxide (NO)-producing peptide, which can release NO, a potent platelet-inhibiting compound (108). A similar NO-producing polyurethane has recently been shown to decrease thrombus formation in a sheep model (109).

Endothelial Incorporation

Current clinically used large-diameter vascular graft materials, such as polyethylene terephthalate (PET) and ePTFE, suffer from early occlusion and thrombosis or late intimal hyperplasia. When employed in small-diameter applications, these failure modes are more prominent and have largely prevented the use of these materials in applications where the diameter is <5 mm. Since these prostheses characteristically do not develop a lumenal

endothelial layer spontaneously, there has been considerable effort expended to develop surface modifications of ePTFE in particular, in order to encourage growth of a confluent endothelial layer, subsequent to the initial report by Herring in 1978 (37).

Two different procedures have been described, endothelial seeding, and sodding. Endothelial seeding involves treating the graft with a low concentration of endothelial cells prior to implantation with *In vivo* expansion of the cells. A two-stage endothelial seeding technique involves treating the graft with a low concentration of cells, followed by extended *In vitro* culture time to expand the cells to a confluent monolayer (110,111). Conversely, endothelial sodding entails coating the graft with a high concentration of endothelial cells to quickly form a confluent endothelial layer. The ePTFE is a very hydrophobic material and few, if any, endothelial cells attach directly to the ePTFE with either direct seeding or sodding techniques. Without modification, hydrophobic ePTFE grafts show attachment of only $10 \pm 7\%$ of applied cells, with EC retention only $4 \pm 3\%$ (112,113). Surface modification of the ePTFE is necessary to encourage endothelial attachment and growth. The source of autologous human endothelial cells can range from harvested veins, adipose tissue capillaries, and endothelial progenitors (75,114,115).

Many different surface modifications of ePTFE to promote endothelialization have been studied, such as attaching cell-binding peptides (116) or fibrin glue (117). A few of these have reached human clinical trials. One such modification involved the application of fibrin glue to the lumenal surface of a small-diameter ePTFE graft to which autologous endothelial cells can attach (110,118). These coated grafts were seeded with autologous endothelial cells, harvested 4–6 weeks prior to graft implantation and expanded *In vitro*. The seeded graft constructs were allowed to mature for 8–10 days before implantation as either one of 21 coronary artery bypass grafts (110) or 153 infrainguinal grafts (118). After a mean postoperative follow-up of 27.7 months, the 4 mm diameter coronary artery graft patency was 90.5%. Angiograms of all 19 patent aortocoronary bypass grafts showed smooth lumenal borders without stenotic regions. Percutaneous translumenal angioscopic evaluation showed a glossy white and smooth endolumenal graft surface without any fibrin, platelet, or erythrocyte deposits. The EC seeded grafts showed an improvement in patency versus unseeded ePTFE aortocoronary bypass grafts (110). For the 6–7 mm diameter infrainguinal reconstructions, Kaplan–Meier analyses revealed a primary patency rate of 84% after 4 years and 63% after 7 years, comparable to primary patency rates for vein grafts (118).

These clinically applied ePTFE modifications were not without incident. In one trial, *In vitro* flow experiments revealed a washout of ECs between 10–15% in the first minute after the application of a physiologic pulsatile flow. Additionally, one patient had a perioperative MI caused by the immediate occlusion of the EC-seeded ePTFE graft because of poor runoff and possible culture contamination with fibroblasts (110). In the other trial, 5% of patients had EC that failed to grow in culture despite rescue serum treatment. At implantation, 5% of seeded grafts were subconfluent and 5% had patches devoid of EC (118). These clinical trials highlight common difficulties faced in many ePTFE endothelialization attempts. Subconfluent EC coatings allow for possible thrombus formation. Extended and technically demanding *In vitro* EC culture is required to generate enough cells for confluent seeding. An additional surgery is required for vessel harvest to isolate primary EC.

To circumvent these difficulties, a potential way to encourage autologous cellular healing *In vivo* is to alter ePTFE pore size. The ePTFE is a hydrophobic material that is microporous due to thin fibrils of material stretched between nodes of solid PTFE; the internodal distance controls the porosity of the ePTFE material. In standard grafts with an internodal distance of 30 μm, endothelialization typically proceeds slowly, if at all, due to EC migration from the anastomoses. Animal studies employing larger pore sizes have shown improved cell in-growth and tissue healing for prostheses with larger internodal distances (119–121). Some (122,123) have suggested that an internodal distance of 60 μm allows optimal cell in-growth from the adventitia while not being as macroporous as prostheses with internodal distances greater than 90 μm (119).

Neovascularization has been demonstrated in animal models with vessels penetrating the larger graft pores (124), however, similar neovascularization and endothelialization is not observed in humans (125). Pore size alone, due to the hydrophobic nature of ePTFE, may be insufficient to influence cell in-growth into ePTFE prostheses.

A significant improvement in ePTFE graft design would be the development of *In vivo* endothelialization without the need for EC preseeding, due to capture of circulating endothelial progenitor cells *In vivo* or facilitating rapid neovascularization from the adventitia, also termed transmural endothelialization. A confluent lumenal monolayer of endothelial cells would serve to eliminate the primary thrombotic mechanism of vascular graft failure. In this regard, the important consideration with ePTFE is how to facilitate endothelial cell attachment. Methods to improve endothelial cell retention include fibronectin preadsorption, covalent fibronectin attachment or glow discharge modification to facilitate protein adsorption. In a study of amide–amine glow-discharge modification, endothelial cell surface adhesion was fourfold higher in treated versus untreated ePTFE grafts (113). Additionally, the endothelial population on treated grafts remained shear stable, whereas untreated grafts showed a >90% decrease in endothelial cell population after application of shear stress (113). An animal implantation model showed improved graft healing after covalent binding of fibronectin to the graft materials (126). In a different approach, Kidd et al. attempted to promote *In vivo* endothelialization in a 1 mm ID ePTFE rat abdominal aortic implant model by facilitating transmural endothelialization (127). Squamous epithelial cell lines were allowed to attach and elaborate extracellular matrix (ECM) on the ablumenal surface of the ePTFE graft for 8 days prior to implantation. The cells were removed, leaving an ePTFE graft with incorporated ECM for

implantation. Upon removal at 5 week, all samples were patent with ECM-modified grafts exhibiting extensive ablumenal vascularization and tissue incorporation compared to nonmodified samples. Additionally, ECM modified grafts possessed a lumenal cellular lining, while nonmodified grafts were void of a cellular lining except for limited pannus in-growth. These results indicate that lumenal EC coverage can be derived from in-growth of capillaries through porous grafts (127).

Tissue Engineering

The recent thrust in developing a successful small-diameter vascular graft has been the use of a tissue engineering approach, instead of starting with an established vascular graft material (128,129). Strategies have included developing degradable polymeric biomaterials (130), utilizing decellularized biological conduits (131), and fabrication of a totally tissue-engineered vessel (38). Typically, a tissue engineering approach utilizes some type of degradable scaffold in which layers of endothelial and smooth muscle cells are grown to recapitulate an arterial architecture. These approaches have unique advantages and limitations. Ideally, tissue engineered vascular constructs should be nonthrombogenic, vasoreactive, and biostable, but should also be biocompatible and not prone to excessive inflammation, immune response, or infection. The cells should also be able to regenerate and angiogenesis should be supported. One challenge for tissue engineered vascular grafts is to combine degradability of the scaffold with sufficient mechanical strength to withstand pulsatile arterial pressures without developing aneurysmal dilatation. From an operative standpoint, tissue engineered grafts should be stored in a readily implantable form, be individualized to particular implant requirements of size and length and not prone to leaking.

Scaffolds. Synthetic polymer scaffolds for tissue engineering applications have utilized bioresorbable polymers, such as polyglycolic acid, polylactic acid, and polydioxanone (34). Polylactic acid is semicrystalline with high mechanical strength, with a naturally occurring degradation product, L-lactic acid (34). In one example, Niklason et al. (132) utilized polyglycolic acid (PGA) scaffolds chemically modified with sodium hydroxide. These were seeded with bovine smooth muscle cells and allowed to incubate under conditions of pulsatile radial stress for 8 weeks. Bovine aortic endothelial cells were seeded onto the lumen of the constructs and continuous perfusion was applied for the final 3 days of culture. These vessels cultured under pulsed conditions with appropriate supplements had average rupture strengths of 2150 ± 708 mmHg (286.6 ± 94.3 kpa) after 8 weeks with 50% dry weight collagen content. These values were very similar to native vessel rupture strength [1680 ± 307 mmHg (223.9 ± 40.9 kPa)] and collagen content (45% dry weight). The engineered vessels also displayed measurable contractions in response to vasoactive substances, such as serotonin, endothelin-1, and prostaglandin F_{2a}.

Another approach is to use two or more bioresorbable polymers with different degradation rates to optimize mechanical properties. Matsumura et al. (133) fabricated two types of bioresorbable polymer for use in a tissue engineering vascular autograft for pediatric patients. One polymer system was comprised of a 50:50 copolymer of lactide and ε-caprolactone [P(CL/LA)] that was reinforced by a nonwoven fabric made with PGA. Endothelial cells were harvested from saphenous vein 1–2 months prior to surgery, expanded *In vitro* and seeded onto the scaffold 10 days prior to surgery. This graft construct was used in a limited clinical study in the pulmonary artery, a low pressure application, with some success. Another hybrid scaffold was similar, but the reinforcing material was changed to a woven fabric made from poly-L-lactic acid (PLLA) to increase durability and the polymer scaffolds were seeded with bone marrow cells aspirated at the time of cardiac surgery and allowed to incubate with the scaffold in culture medium for 2–4 h before implantation. This BMC seeded polymer scaffold was used in 22 patients without thrombogenic complications, stenosis, or obstruction of tissue-engineered autografts. The benefits of this type of tissue engineered vascular autograft is that it has growth potential, reduced incidence of calcifications, and no risk of rejection due to use of autologous cells, but long term *In vitro* cell culture may increase the risk of infection and changes in cell phenotype could lead to thromboembolic complications.

The degradable polymer scaffolds have also been modified to include cell-binding peptides to further encourage cell growth and incorporation, such as arginine-glycine-aspartic acid (RGD), a well-characterized amino acid peptide that is a ligand to several cell surface integrin receptors (134–136). In one study, a biotinylated PLA–polyethylene glycol copolymer was reacted with streptavidin–RGD to facilitate endothelial cell attachment (137). Controlling the density of RGD groups on the polymer surface has been shown to modulate endothelial cell migration rate and shear stability (138,139).

A different approach to synthetic materials is the use of deoxyribonucleic acid (DNA) technology to engineer synthetic proteins. These have been used with some success to incorporate both RGD functionality and elastin-based peptides to facilitate endothelial cell growth and function, while allowing cross-linking and control of mechanical properties (140). Degradation can be controlled by incorporating sequences susceptible to degradation by enzymes, such as collagenase. Incorporation of cell binding amino acid sequences into these matrices has been shown to control cell functionality (141).

Biological materials have also been used as scaffolds, including fibrinbased or collagenous matrices and decellularized materials, such as extracellular matrix, arteries, small intestinal submucosa, and peritoneum–fascia (75,142–145). Cross-linked collagen is an attractive scaffold material, as it can have significant mechanical strength, depending on cross-link density, and contains several cell-binding peptides to encourage cell incorporation. A collagen scaffold was used in the first tissue engineered prosthesis (38), but this failed due to suboptimal mechanical properties. A fibrin gel-based graft has recently

shown some success, albeit in a low pressure jugular vein model (142). Methods to improve the scaffold mechanical properties have included mechanical preconditioning and use of fabricating techniques, such as electrospinning (146,147). Matsuda and co-workers used a rodlike mandrel around which a solution of collagen and smooth muscle cells was formed (148). The mandrel was then removed and the lumen seeded with endothelial cells. This was unsuccessful due to low burst pressures, and external reinforcement was needed to impart suitable mechanical properties (149,150). A collagen microsponge has been combined with a biodegradable polymeric scaffold to produce an implantable graft material for *In vivo* cell repopulation (151). Recently, a hyaluronan-based material (Hyaff-11) has been tested as a scaffold for endothelial cells (152).

Use of biological tissues with a preexisting extracellular matrix, such as arteries (153) or ureters (154), would be attractive, but antigenic differences between xenogenic species and humans preclude their use due to immunoreactivity and rejection. For this reason, there have been attempts to remove the cells from these scaffolds by enzymatic and detergent treatment and repopulate the decellularized scaffolds with autologous human cells (155). The xenogenic matrix still is potentially immunogenic (156), which may be alleviated by cross-linking (157) or population with autologous cells. Decellularized porcine carotid arteries treated with heparin and repopulated with canine cells have been shown to function for up to 18 weeks in a canine aortic graft model (153). Decellularized porcine small intestinal submucosa has also been studied as a tissue engineering scaffold. These constructs have been tested in animal models with some success, but intimal hyperplasia and incomplete endothelial coverage have been observed (158).

Source of Cells. In most tissue engineering constructs, the endothelial and smooth muscle cells are isolated from autologous human tissue, such as abdominal fat aspirates or veins, then cultured *In vitro* together with the construct for up to several weeks prior to implantation. There has been some recent interest in the use of bone marrow-derived endothelial progenitor cells as a novel source of endothelial cells (154). Cho et al. (159) examined tissue engineering of small diameter vascular grafts using bone marrow cells (BMCs) and decellularized arteries. Bone marrow mononuclear cells (BMMNCs) were isolated and two distinct fractions, one containing smooth muscle (SM) a-actin/SM myosin heavy-chain (SMMHC) positive cells and one containing vWF/CD31 positive cells, were cultured for 3 weeks using different culture media and supplements. Other studies have shown that when using BMCs, sufficient cells for seeding can be obtained on the day of surgery, thus obviating the need for extra vein harvesting surgery and prolonged cell culture.

SUMMARY

Vascular prostheses are vital in the treatment of many vascular diseases, from coronary artery disease and aortic aneurysms to peripheral arterial disease. Several synthetic vascular graft materials, particularly Dacron and ePTFE, have been successful as arterial substitutes in large diameter applications. However, these devices can fail due to thrombotic occlusion, intimal hyperplasia, and infection. There is a need to develop novel vascular prostheses that function well in small diameter applications. Progress has been made on methods to facilitate the incorporation of endothelial cells into vascular grafts. The tissue engineering approach, with advances in new materials that can control graft mechanical properties plus cell attachment and proliferation, holds promise for the next generation of these crucial medical devices.

BIBLIOGRAPHY

1. Stehbens WE. General features, structure, topography and adaptation of the circulatory systems. In: Stehbens WE, Lie JT, editors. Vascular Pathology. New York: Chapman and Hall; 1995. p 1–18.
2. Wu KK, Thiagarajan P. Role of endothelium in thrombosis and hemostasis. Ann Rev Med 1996;47:315–331.
3. Gertler JP, Abbott WM. Prothrombotic and fibrinolytic function of normal and perturbed endothelium. J Surg Res 1992;52(1):89–95.
4. Cooley DA, DeBakey ME. Surgical consideration of intrathoracic aneurysms of the aorta and great vessels. Ann Surg 1952;135:660–680.
5. Voorhees AB Jr, Jaretzke AL III, Blakemore AH. The use of tubes constructed from Vinyon-'N' cloth in bridging arterial defects. Ann Surg 1952;135:332–336.
6. Sporn L, Huber P. Endothelial cell biology, in Hemostasis and Thrombosis: Basic principles and Clinical Practice. In: Colman RW, editor. Philadelphia: Lippincott, Williams, and Wilkins; 2001. p 615–623.
7. Hansson GK. Inflammation, atherosclerosis, and coronary artery disease. N Engl J Med 2005;352(16):1685–1695.
8. Cardiovascular Disease Statistics. Available at http://www.americanheart.org/presenter.jhtml?identifier=4478: Accessed 2005 April 30.
9. American Heart Association. Heart Disease and Stroke Statistics—2005 Update. Dallas: American Heart Association; 2005.
10. Open-Heart Surgery Statistics. Available at http://www.americanheart.org/present.jhtml?identifier=4674 Accessed 2005 April 30.
11. Benavente O, Hart RG, Sherman DG. Primary Prevention of Transient Ischemic Attach and Thromboembolic Stroke. In: Verstraete M, Fuster V, Topol EJ, editors. Cardiovascular Thrombosis: Thrombocardiology and Thromboneurology, 2nd ed. Philadelphia: Lippincott-Raven Publishers; 1998. p 585–595.
12. Ferro JM. Cardioembolic stroke: an update. Lancet Neurol 2003;2:177–188.
13. Hirsch AT. Peripheral Arterial Disease Detection, Awareness, and Treatment in Primary Care. JAMA 2001;286:1317–1324.
14. Dunbard RL, Mohler ER 3rd. The unsung perils of peripheral arterial disease: a malady in search of a patient. Prev Cardiol 2005;8(2):108–113.
15. Londrey GL, et al. Infrapopliteal bypass for severe ischemia: comparison of autogenous vein, composite and prosthetic grafts. J Vasc Surg 1991;13:631–636.
16. Veith FJ, et al. Six-year prospective multicenter randomized comparison of autologous saphenous vein and expanded

polytetrafluoroethylene grafts in infrainguinal arterial reconstructions. J Vasc Surg 1986;3:104–114.
17. Sala F, et al. Long-term outcome of femoral above-knee popliteal artery bypass using autologous saphenous vein versus expanded polytetrafluoroethylene grafts. Ann Vasc Surg 2003;17:401–407.
18. Sakalihasan N, Limet R, Defawe OD. Abdominal aortic aneurysm. Lancet 2005;365:1577–1589.
19. The UK Small Aneurysm Trial Participants. Mortality results for randomized controlled trial of early elective surgery or ultrasonographic surveillance for small abdominal aortic aneurysms. The UK Small Aneurysm Trial Participants. Lancet 1998;352:1649–1655.
20. Sakalihasan N, et al. Modification of the extracellular matrix of aneurysmal abdominal aortas as a function of their size. Eur J Vasc Surg 1993;7:633–637.
21. Dobrin PB, Mrkvicka R. Failure of elastin or collagen as possible critical connective tissue alterations underlying aneurysmal dilatation. Cardiovasc Surg 1994;2:484–488.
22. Shah PK. Inflammation, metalloproteinases, and increased proteolysis: an emerging pathophysiological paradigm in aortic aneurysm. Circulation 1997;96:2115–2117.
23. Kamineni R, Heuser RR. Abdominal aortic aneurysm: A review of endolumenal treatment. J Interven Cardiol 2004;17:437–445.
24. Moore WS, Kashyap VS, Vescera CL, Quinones-Baldrich WJ. Abdominal aortic aneurysm: A 6-year comparison of endovascular versus transabdominal repair. Ann Surg 1999; 230:298–306.
25. Parodi JC, Palmaz JC, Barone HD. Transfemoral intralumenal graft implantation for abdominal aortic aneurysms. Ann Vasc Surg 1991;5:491–499.
26. Perry MO. Arterial Injuries. In: Bell PRF, Jamieson CW, Ruckley CV, editors. Surgical Management of Vascular Disease. London: W. B. Saunders; 1992. p 905–925.
27. Lumley JSP. Microvascular Surgery. In: Bell RF, Jamieson CW, Ruckley CV, editors. Surgical Management of Vascular Disease. London: W.B. Saunders; 1992. p 941–954.
28. Millis JM, Brown SL, Busuttil RW. Thoracic and Abdominal Aneurysms. In: Bell PRF, Jamieson CW, Ruckley CV, editors. Surgical Management of Vascular Disease. London: W. B. Saunders; 1992. p 797–828.
29. Keen KW. Surgery: Its Principles and Practice. Philadelphia: W.B. Saunders; 1921.
30. Rea CE. Surgical treatment of aneurysm of the abdominal aorta. Minn Med 1948;31:153.
31. Dubost C, Allary M, Oeconomos N. Resection of an aneurysm of the abdominal aorta: Reestablishment of the continuity by a preserved human arterial graft, with result after five months. AMA Arch Surg 1952;64:405–408.
32. Cooley DA. Early development of Surgical treatment for aortic aneurysms: Personal recollections. Texas Heart Institute J 2001;28:197–199.
33. Hess F. History of (micro)vascular surgery and the development of smallcaliber blood vessel prostheses. Microsurgery 1985;6:59–69.
34. Xue L, Greisler HP. Biomaterials in the development and future of vascular grafts. J Vasc Surg 2003;37:472–480.
35. Matsumoto H, Hasegawa T, Fuse K. A new vascular prosthesis for a small caliber artery. Surgery 1973;74:518–523.
36. Dardik H, et al. Glutaraldehyde-tanned human umbilical vein grafts. In: Stanley JC, et al. editors. Biologic and Synthetic Vascular Prostheses. New York: Grune & Stratton; 1982. p 433.
37. Herring M, Gardner A, Glover J. A single-staged technique for seeding vascular grafts with autogenous endothelium. Surgery 1978;84:498–504.
38. Weinberg CB, Bell E. A blood vessel model constructed from collagen and cultured vascular cells. Science 1986;231:397–400.
39. Kunlin JL. Le traitement de l'arterite obliterante par la greffe veinuse. Arch Mal Coeur 1949;42:371.
40. Yun KL, et al. Randomized trial of endoscopic versus open vein harvest for coronary artery bypass grafting: sixmonth patency rates. J Thorac Cardiovasc Surg 2005;129:496–503.
41. Lytle BW. Prolonging patency—choosing coronary bypass grafts. N Eng J Med 2004;351:2262–2264.
42. Desai ND, Cohen EA, Naylor CD, Fremes SE, the Radial Artery Patency Study Investigators. A randomized comparison of radial-artery and saphenous-vein coronary bypass grafts. N Engl J Med 2004;351:2302–2309.
43. Chervu A, Morre WS. Vascular Grafts and Sutures. In: Bell RF, Jamieson CW, Ruckley CV, editors. Surgical Management of Vascular Disease. London: WB. Saunders; 1992. p 367–389.
44. Boyd JH, Stevens R, Havey A, Silver D. Intimal integrity and fibrinolytic potential of reversed and in situ vein grafts. J Vasc Surg 1987;5:614–621.
45. Lawson JA, Tangelder MJ, Algra A, Eikelboom BC. The myth of the in situ graft: superiority in infrainguinal bypass surgery? Eur J Vasc Endovasc Surg 1999;18:149–157.
46. Snyder RW, Botzko KM. Woven knitted and externally supported Dacron vascular prostheses. Stanley JC, et al. editors. Biologic and Synthetic Vascular Prostheses. New York: Grune & Stratton; 1982. p 485.
47. Cziperle DJ, et al. Albumin impregnated vascular grafts: albumin resorption and tissue reactions. J Cardiovasc Surg 1992;33:407–414.
48. Hall CW, et al. Velour fabrics applied to medicine. J Biomed Mater Res 1967;1:179–196.
49. Lindenauer SM, Lavanway JM, Fry WJ. Development of a velour vascular prosthesis. Curr Top Surg Res 1970;2: 491.
50. Sauvage LR, et al. An external velour surface for porous arterial prostheses. Surgery 1971;70:940–953.
51. Norman PE, Semmens JB, Lawrence-Brown MM. Long-term relative survival following surgery for abdominal aortic aneurysm: a review. Cardiovasc Surg 2001;9:219–224.
52. Friedman SG, et al. A prospective randomized comparison of Dacron and polytetrafluoroethylene aortic bifurcation grafts. Surgery 1995;117:7–10.
53. Quarmby JW, et al. Prospective randomized trial of woven versus collagen-impregnated knitted prosthetic Dacron grafts in aortoiliac surgery. Br J Surg 1998;85: 775–777.
54. Devine C, McCollum C. for the North West Femoro-Popliteal Trial Participants. Heparin-bonded Dacron or polytetrafluoroethylene for femoropopliteal bypass: Five-year results of a prospective randomized multicenter clinical trial. J Vasc Surg 2004;40:924–931.
55. Davidovic L, et al. Aortobifemoral grafting: factors influencing long-term results. Vascular 2004;12:171–178.
56. Gupta SK, Ascer E, Veith FJ. Expanded polytetrafluoroethylene arterial grafts: An eight-year experience. In: Sawyer PN, editor. Modern Vascular Grafts. New York: McGraw-Hill; 1987. p 181.
57. Berglund J, Bjorck M, Elfstrom J. for the SWEDVASC Femoro-popliteal study group. Long-term results of above knee femoro-popliteal bypass depend on indication for surgery and graft-material. Eur J Vasc Endovasc Surg 2005;29:412–418.
58. Taylor LM, Edwards JM, Porter JM. Present status of reversed vein bypass grafting: five-year results of a modern series. J Vasc Surg 1990;10:220–225.

59. Donaldson MC, Mannick JA, Whittemore AD. Femoro-distal bypass with in situ greater saphenous vein. AnnSurg 1991; 213:457–465.
60. Sala F, et al. Long-term outcome of femoral above-knee popliteal artery bypass using autologous saphenous vein versus expanded polytetrafluorethylene grafts. Ann Vasc Surg 2003;17:401–407.
61. Wood RFM. Vascular access in dialysis. In: Bell RF, Jamieson CW, Ruckley CV, editors. Surgical Management of Vascular Disease. London: W.B. Saunders; 1992. p 1049–1067.
62. Francis DMA. More vein, less plastic. Nephrology 2005;10: 10–14.
63. Bosman PJ, et al. A comparison between PTFE and denatured homoogous vein grafts for haemodialysis access: A prospective randomized multi-center trial. Eur J Vasc Endovasc Surg 1998;16:126–132.
64. Crowther MA, et al. Low-intensity Warfarin is ineffective for the prevention of PTFE graft failure in patients on hemodialysis: A randomized controlled trial. J Am Soc Nephrol 2002;13:2331–2337.
65. Towne JB. Endovascular treatment of abdominal aortic aneurysms. Am J Surg 2005;189:140–149.
66. Laheij RJ, et al. Need for secondary interventions after endovascular repair of abdominal aortic aneurysms: intermediate-term follow-up results of a European collaborative registry (EUROSTAR). Br J Surg 2000;87:1666–1673.
67. Hallett JW Jr, et al. Graft-related complications after abdominal aortic aneurysm repair: reassurance from a 36-year population-based experience. J Vasc Surg 1997;25:277–284.
68. Jacobs TS, et al. Mechanical failure of prosthetic human implants: A 10 year experience with aortic stent graft devices. J Vasc Surg 2003;37:16–26.
69. Anderson JM, Kottke-Marchant K. Platelet interaction with biomaterials and artificial devices. CRC Crit Rev Biocompat 1985;1:111–204.
70. Greisler HP, et al. Biointeractive polymers and tissue engineered blood vessels. Biomaterials 1996;17:329–336.
71. Kottke-Marchant K, Anderson JM, Rabinovitch A. The platelet reactivity of vascular graft prostheses: An In vitro model to test the effect of preclotting. Biomaterials 1986;7:441–448.
72. Lin Y, Weisdorf DJ, Solovey A, Hebbel RP. Origins of circulating endothelial cells and endothelial outgrowth from blood. J Clin Invest 2000;105:71–77.
73. Davids L, Dower T, Zilla P. The lack of healing in conventional vascular grafts. In: Zilla P, Greisler HP, editors. Tissue engineering of vascular prosthetic grafts. Austin, TX: R.G. Landes Co.; 1999. p 3–44.
74. Conte MS, et al. Genetic interventions for vein bypass graft disease: a review. J Vasc Surg 2002;36:1040–1052.
75. Rashid ST, et al. Engineering of bypass conduits to improve patency. Cell Prolif 2004;37:351–366.
76. Kidane AG, et al. Anticoagulant and antiplatelet agents. Their clinical and device application(s) together with usages to engineer surfaces. Biomacromolecules 2004;5:798– 813.
77. Collins TC, Souchek J, Beyth RJ. Benefits of antithrombotic therapy after infrainguinal bypass grafting: a meta-analysis. Am J Med 2004;117:93–99.
78. Dorffler-Melly J, et al. Antiplatelet agents for preventing thrombosis after peripheral arterial bypass surgery. Cochrane Database Syst Rev 2003;3:CD000535.
79. LoGerfo FW, et al. Anastomotic hyperplasia: A mechanism of failure in Dacron arterial grafts. Ann Surg 1983;197:479–483.
80. Rotmans JI, et al. Rapid, arteriovenous graft failure due to intimal hyperplasia: A porcine, bilateral, carotid arteriovenous graft model. J Surg Res 2003;113:161–171.
81. Clowes AW, Gown AM, Hanson SR, Reidy MA. Mechanisms of arterial graft failure. Role of cellular proliferation in early healing of PTFE prostheses. Am J Pathol 1985;118:43–54.
82. Clowes AW, Kirkman TR, Clowes MM. Mechanisms of arterial graft failure. II. Chronic endothelial and smooth muscle cell proliferation in healing polytetrafluoroethylene prostheses. J Vasc Surg 1986;3:87–884.
83. Haruguchi H, Teraoka S. Intimal hyperplasia and hemodynamic factors in arterial bypass and artierovenous grafts: a review. J Artif Organs 2003;6:227–235.
84. Purcell C, Tennant M, McGeachie J. Neo-intimal hyperplasia in vascular grafts and its implications for autologous arterial grafting. Ann R Coll Surg Engl 1997;79:164–168.
85. Tiwari A, Salacinski H, Seifalian AM, Hamilton G. New prostheses for use in bypass grafts with special emphasis on polyurethanes. Cardiovasc Surg 2002;10:191–197.
86. Zubilewicz T, et al. Injury in vascular surgery — the intimal hyperplastic response. Med Sci Monit 2001;7:316–324.
87. Sapienza P, et al. Release of PDGF-BB and bFGF by human endothelial cells seeded on expanded polytetrafluoroethylene vascular grafts. J Surg Res 1998;74:24–29.
88. Mattana J, Effiong C, Kapasi A, Singhal PC. Leukocytepolytetrafluoroethylene interaction enhances proliferation of vascular smooth muscle cells via tumor necrosis factor-alpha secretion. Kidney Int 1997;52:1478–1485.
89. Randone B, et al. Suppression of smooth muscle cell proliferation after experimental PTFE arterial grafting: a role for polyclonal anti-basic fibroblast growth factor (bFGF) antibody. Eur J Vasc Endovasc Surg 1998;16:401–407.
90. Yu H, et al. Neointimal hyperplasia on a cell-seeded polytetrafluoroethylene graft is promoted by transfer of tissue plasminogen activator gene and inhibited by transfer of nitric oxide synthase gene. J Vasc Surg 2005;41:122–129.
91. Goldstone J, Moore WS. Infection in vascular prosthesis: Clinical manifestations and Surg management. Am J Surg 1974;128:225–233.
92. Shell DH, et al. Comparison of small-intestinal submucosa and expanded polytetrafluoroethylene as a vascular conduit in the presence of gram-positive contamination. Ann Surg 2005;241:995–1001.
93. Stadler P, Bilohlavek O, Spacek M, Michalek P. Diagnosis of vascular prothesis infection with FDG-PET/CT. J Vasc Surg 2004;40:1246–1247.
94. Hart JP, et al. Operative strategies in aortic graft infections: is complete graft excision always necessary?. Ann Vasc Surg 2005;19:154–160.
95. Aldenhoff YBJ, et al. Performance of a polyurethane vascular prosthesis carrying a dipyridamole (Persantin) coating on its lumenal surface. J Biomed Mater Res 2001;54:224–233.
96. Lin P, et al. Small-caliber heparin-coated ePTFE grafts reduce platelet deposition and neointimal hyperplasia in a baboon model. J Vasc Surg 2004;39(6):1322–1328.
97. Lambert AW, et al. Experience with heparin-bonded collagen-coated grafts for infrainguinal bypass. Cardiovasc Surg 1999;7:491–494.
98. Walpoth BH, et al. Improvement of patency rate in heparin-coated small synthetic vascular grafts. Circulation 1998;98:II 319–II 323.
99. Akers DL, Du YH, Kempczinski RF. The effect of carbon coating and porosity on early patency of expanded polytetrafluoroethlene grafts: An experimental study. J Vasc Surg 1993;18:10–15.
100. Greisler HP, et al. Endothelialization of expanded PTFE grafts by heparin binding growth factor-type1 pretreatment. Surgery 1992;112:244–255.

101. Gray JL, et al. FGF-1 affixation stimulates ePTFE endothelialization without intimal hyperplasia. J Surg Res 1994; 57:596–612.
102. Wang C, Zhang Q, Uchida S, Kodama M. A new vascular prosthesis coated with polyamino-acid urethane copolymer (PAU) to enhance endothelialization. J Biomed Mater Res 2002;62:315–322.
103. Zhang Z, et al. Vascugraft polyurethane arterial prosthesis as femoro-popliteal and femoro-peroneal bypasses in humans: pathological, structural and chemical anlysis of four excised grafts. Biomaterials 1997;18:113–124.
104. Glickman MH, et al. Multicenter evaluation of a polyurethane vascular access graft as compared with the expanded polytetrafluoroethylene vascular access graft in hemodialysis applications. J Vasc Surg 2001;34:465–472.
105. Jeschke MG, Hermanutz V, Wolf SE, Koveker GB. Polyurethane vascular prostheses decreases neointimal formation compared with expanded polytetrafluoroethylene. J Vasc Surg 1999;29:168–176.
106. Wilson GJ, et al. The composite Corethane/Dacron vascular prosthesis. Canine In vivo evaluation of 4 mm diameter grafts with 1 year follow-up. ASAIO Trans 1991;37:M475–M476.
107. Izhar U, et al. Novel synthetic selectively degradable vascular prostheses: A preliminary implantation study. J Surg Res 2001;95:152–162.
108. Jun H-W, Taite LJ, West JL. Nitric oxide-producing polyurethanes. Biomacromolecules, 2005;6:838–844.
109. Fleser PS, et al. Nitric oxide-releasing biopolymers inhibit thrombus formation in a sheep model of arteriovenous bridge grafts. J Vasc Surg 2004;40:803–833.
110. Laube H, Duwe J, Rutsch W, Konertz W. Clinical experience with autologous endothelial cell-seeded polytetrafluoroethylene coronary artery bypass grafts. J Thorac Cardiovasc Surg 2000;120(1):134–141.
111. Deutsch M, et al. Clinical autologous *In vitro* endothelialization of infrainguinal ePTFE grafts in 100 patients: a 9 year experience. Surgery 1999;126:847–855.
112. Kent K, Oshima A, Whittemore A. Optimal seeding conditions for human endothelial cells. Ann Vasc Surg 1992;6(3): 258–264.
113. Tseng DY, Edelman ER. Effects of amide and amine plasma-treated ePTFE vascular grafts on endothelial cell lining in an artificial circulatory system. J Biomed Mater Res 1998;42: 188–198.
114. Sharp WV, Schmidt SP, Meerbaum SO, Pippert TR. Derivation of human microvascular endothelial cells for prosthetic vascular graft seeding. Ann Vasc Surg 1989;3:104–107.
115. Boyer M, et al. Isolation of endothelial cells and their progenitor cells from human peripheral blood. J Vasc Surg 2000;31:181–189.
116. Chan BP, et al. *In vivo* performance of dual ligand augmented endothelialized expanded polytetrafluoroethylene vascular grafts. J Biomed Mater Res Part B: Appl Biomater 2005; 72B:52–63.
117. Kumar TRS, Krishnan LK. A stable matrix for generation of tissueengineered nothrombogenic vascular grafts. Tissue Eng 2002;8:763–770.
118. Meinhart J, et al. Clinical autologous *In vitro* endothelialization of 153 infrainguinal ePTFE grafts. Ann Thorac Surg 2001;71(5 Suppl):S327–331.
119. Hazama K, et al. Relationship between fibril length and tissue ingrowth in the healing of expanded polytetrafluoroethylene grafts. Surg Today 2004;34:685–689.
120. Golden MA, et al. Healing of polytetrafluoroethylene arterial grafts is influenced by graft porosity. J Vasc Surg 1990; 11:838–845.
121. Hirabayashi K, et al. Influence of fibril length upon ePTFE graft healing and host modification of the implant. J Biomed Mater Res 1992;26:1433–1447.
122. Kuzuya A, et al. Healing of implanted expanded polytetrafluoroethylene vascular access grafts with different internodal distances: A histologic study in dogs. Eur J Vasc Endovasc Surg 2004;28:404–409.
123. Contreras MA, Quist WC, LoGerfo FW. Effect of porosity on smalldiameter vascular graft healing. Microsurgery 2000; 20:15–21.
124. Clowes AW, Kirkman TR, Reiday MA. Mechanisms of graft healing. Rapid transmural capillary ingrowth provides a source of intimal endothelium and smooth muscle in porous PTFE prosthesis. Am J Pathol 1986;123:221–230.
125. Sauvage LR, et al. Interspecies healing of porous arterial prosthesis. Observations, 1960 to 1974. Arch Surg 1974;109: 6989–6705.
126. Shimada T, et al. Improved healing of small-caliber, long-fibril expanded polytetrafluoroethylene vascular grafts by covalent bonding of fibronectin. Surg Today 2004;34:1025–1030.
127. Kidd KR, Patula VB, Williams SK. Accelerated endothelialization of interpositional 1-mm vascular grafts. J Surg Research 2003;113:234–242.
128. Matsuda T. Recent progress of vascular graft engineering in Japan. Artif Organs 2004;28:64–71.
129. Nerem RM, Seliktar D. Vascular tissue engineering. Ann Rev Biomed Eng 2001;3:225–243.
130. Pachence JM, Kohn J. Biodegradable polymers. In: Lanza RP, Langer R, Vacanti J, editors. Principles of tissue engineering. 2nd ed. San Diego: Academic Press; 2000. p 263–277.
131. Wilson GJ, et al. Acellular matrix: a biomaterial approach for coronary artery bypass and heart valve replacement. Ann Thorac Surg 1995;60(2 Suppl):S353–S358.
132. Niklason LE, et al. Functional arteries grown *In vitro*. Science 1999;284:489–493.
133. Matsumura G, et al. Successful application of tissue engineered vascular autografts; clinical experience. Biomaterials 2003;24:2303–2308.
134. Drumheller PD, Hubbell JA. Polymer networks with grafted celladhesion peptides for highly biospecific cell adhesive substrates. Anal Biochem 1994;222:380–388.
135. Hersel U, Dahmen C, Kessler H. RGD modified polymers: biomaterials for stimulated cell adhesion and beyond. Biomaterials 2003;24:4385–4415.
136. Shin H, Jo S, Mikos AG. Biomimetic materials for tissue engineering. Biomaterials 2003;24:4353–4364.
137. Patel N, et al. Spatially controlled cell engineering on biodegradable polymer surfaces. FASEB J 1998;12:1447–1454.
138. Sagnella S, et al. Human Microvascular Endothelial Cell Growth and Migration on Biomimetic Surfactant Polymers. Biomaterials 2004;25:1249–1259.
139. Sagnella S, Kligman F, Marchant RE, Kottke-Marchant K. Biomimetic Surfactant Polymers Designed for Shear Stable Endothelialization on Biomaterials. J Biomed Mater Res 2003;67A(3):689–701.
140. Urry DW, Pattanaik A. Elastic protein-based material in tissue reconstruction. Ann NY Acad Sci 1997;831:32–46.
141. Richman GP, Tirrell DA, Asthagiri AR. Quantitatively distinct requirements for signaling-competent cell spreading on engineered versus natural adhesion ligands. J Controlled Rel 2005;101:3–12.

142. Swartz DD, Russell JA, Andreadis ST. Engineering of fibrin-based functional and implantable small-diameter blood vessels. Am J Physiol Heart Circ Physiol 2005;288:H1451–H1460.
143. Bader A, et al. Engineering of human vascular aortic tissue based on a xenogeneic starter matrix. Transplantation 2000; 70:7–14.
144. Clarke DR, et al. Transformation of nonvascular acellular tissue matrices into durable vascular conduits. Ann Thorac Surg 2001;71:S433–S436.
145. Sarac TP. *In vivo* and mechanical properties of peritoneum/fascia as a novel arterial substitute. J Vasc Surg 2005; 41:490–497.
146. Seliktar D, Black RA, Vitro RP, Nerem RM. Dynamic mechanical conditioning of collagen-gel blood vessel constructs induces remodeling in vitro. Ann Biomed Eng 2000;28:351–362.
147. Matthews JA, et al. Smooth muscle cell migration in electrospun poly(lactic acid) and collagen/elastin. Cardiovasc Pathol 2002;11:13–18.
148. Hirai J, Matsuda T. Venous reconstruction using hybrid vascular tissue composed of vascular cells and collagen: tissue regeneration process. Cell Transplant 1996;5:93–105.
149. He H, Matsuda T. Newly designed compliant hierarchical hybrid vascular graft wrapped with microprocessed elastomeric film — II: Morphogenesis and compliance change upon implantation. Cell Transplant 2002;11:75–87.
150. He H, Shirota T, Yasui H, Matsuda T. Canine endothelial progenitor cell-lined hybrid vascular graft with non-thrombogenic potential. J Thorac Cardiovasc Surg 2003;126:455–464.
151. Iwai S, et al. Biodegradable polymer with collagen microsponge serves as a new bioengineered cardiovascular prosthesis. J Thorac Cardiovasc Surg 2004;128:472–479.
152. Turner NJ, Kielty CM, Walker MG, Canfield AE. A novel hyaluronan-based biomaterial (Hyaff-11) as a scaffold for endothelial cells in tissue engineered vascular grafts. Biomaterials 2004;25:5955–5964.
153. Tamura N. New acellular vascular prosthesis as a scaffold for host tissue regeneration. Artif Organs 2003;26:783–792.
154. Shirota T, He H, Yasui H, Matsuda T. Human endothelial progenitor cell-seeded hybrid graft: Proliferative and antithrombogenic potentials *In vitro* and fabrication processing. Tissue Eng 2003;9:127–136.
155. Wilson GJ, et al. Acellular matrix: a biomaterials approach for coronary artery bypass and heart vlave replacement. Ann Thorac Surg 1995;60(2 Suppl):S353–S358.
156. Allaire E, et al. The immunogenicity of the ECM in arterial xenografts. Surgery 1997;122:73–81.
157. Courtman DW, Errett BF, Wilson GJ. The role of crosslinking in modification of the immune response elicited against xenogenic vascular acellular matrices. J Biomed Mater Res 2001;55:576–586.
158. Nemcova S, et al. Evaluation of a xenogeneic acellular collagen matrix as a smalldiameter vascular graft in dogs — preliminary observations. J Invest Surg 2001;14:321–330.
159. Cho S-W, et al. Small-diameter blood vessels engineered with bone marrow-derived cells. Ann Surg 2005;241:506–515.

See also BIOCOMPATIBILITY OF MATERIALS; BIOMATERIALS, SURFACE PROPERTIES OF; POROUS MATERIALS FOR BIOLOGICAL APPLICATIONS.

VASCULAR MEASUREMENTS. See PERIPHERAL VASCULAR NONINVASIVE MEASUREMENTS.

VENOUS SHUNT. See VASCULAR GRAFT PROSTHESIS.

VENTILATION, HIGH-FREQUENCY. See HIGH FREQUENCY VENTILATION.

VENTILATORS, ACUTE MEDICAL CARE

PAUL C. TAMUL
KENNETH SCOPE
LEONARD CRAIG
Feinberg School of Medicine of Northwestern University
Chicago, Illinois

VENTILATOR USE

Introduction

A ventilator, like most critical care devices, is simply a machine. However, it is not a "simple machine." Over the past five decades, the means of delivering gases to patients have become more complex. This has led to an evolution of more than 30 critical care ventilators being used in the United States. The numerous modes of ventilation and our unfortunate attempt to reach a consensus on the terminology used when referring to mechanical ventilation presents a challenge to even the most skilled clinician. Mushin's description of ventilators in his archetypal text (1) is appreciated by many and addresses the classification scheme that can be adopted to the less-sophisticated mechanical ventilator. An innumerable count of authors has used this classification method when writing about mechanical ventilator support (2–4). Today's mechanical ventilators are designed to provide life support using an array of computerized logic systems, software utilities, and even artificial intelligence models to deliver gas to patients. This evolution of ventilators has led to a modernistic classification system (5) to describe today's microprocessor-controlled mechanical ventilator. A classification system that is specific and well defined is intended to enhance communications between clinicians (6). However, some respected organizations (7) and clinicians (8) have been reluctant to accept this new system for understanding mechanical ventilation despite its acceptance by leading members in the pulmonary and critical care medicine community (9–11).

Basic Concepts

A mechanical ventilator is a system designed to alter, transmit, and direct applied energy in a predetermined manner to perform useful work (12). This work was generally accomplished with some pneumatic component in earlier (first-generation) ventilators. Current systems employ electronic circuitry that is microprocessor controlled.

The energy we put into ventilators is in the form of electricity or compressed gas. Mechanically speaking, a

ventilator will take the power input and convert or transmit that energy using the control scheme circuit to arrive at a desired output. Put more plainly, that energy is transmitted or transfigured by the ventilator in a predetermined manner to provide partial or full support to the patient's respiratory muscles.

Input Power

The input power or power source is the energy needed to perform the work required to ventilate the patient.

Generally, two forms of input energy are used to operate mechanical ventilators:

Electric

AC (alternating current)
DC (direct current)

Pneumatic. Many critical care ventilators in use today incorporate electric and pneumatic power sources to function properly because they employ advanced mechanical systems in combination with microprocessor control.

Electric

Critical care ventilators for use in the United States use common electrical voltage (110–115 V) to power them. Some manufacturers have adapted their machines with rechargeable batteries as an alternative in case of transport from one location to another or in case a power failure erupts. Some disadvantages to the type of batteries (lead–acid) used is that they supply limited power, lasting approximately 1 h and generally require a 12–24 h recharging time.

Pneumatic

The availability of compressed gases of oxygen and air in many hospital intensive care units (ICUs) makes it an ideal energy source (energy = pressure × volume) to power today's mechanical ventilator. Although the standard regulated source gas pressure is 50 pounds per square inch (psi), periodic fluctuations in gas pressure may occur (13). Ordinarily, ventilators have internal regulators that work at a lesser pressure (30–45 psi) than the source pressure. This will inherently maintain normal function despite inconsistent source gas pressure. Aside from unpredictable gas pressures, a pneumatically powered ventilator is useful in environments using magnetic resonance imaging (MRI) or when the need for transport arises. To date, only a few machines are capable of functioning on pneumatics only. Most ICU ventilators that are pneumatically powered still require electricity to support their control functions.

Power Conversion and Transmission

Power conversion and transmission refer to the mechanism(s) used to provide gas delivery. Sometimes this is referred to as the drive mechanism. This is only *partially* true. The drive mechanism generates the actual force (force = pressure/area) to deliver gas under pressure. This is usually done with a compressor (internal or external) and/or motor linkage. Power transmission and conversion systems also consist of an output control mechanism. This is usually in the form of one or a series of valves that is used to regulate gas flow to the patient. In more complex machinery, a computerized, closed–loop feedback system is established or provides consistent gas delivery in the event of system disturbances. More recent developments of power transmission and conversion mechanisms are described here. A more detailed list can be found elsewhere (14,15).

Drive Mechanisms. This section will investigate the various drive mechanisms required to accomplish lung inflation. The post-polio epidemic introduced an upsurge in the development of many types of drive mechanisms. Although a pressure gradient (required for lung inflation) can be accomplished by one method, many forms of drive mechanisms were developed and used. Whereas each drive mechanism had a unique way of generating a particular pressure and flow pattern during a positive pressure breath, the ideal mechanism should mimic the physiologic breath and restrain adverse effects. Unfortunately, that is still not accomplished today.

There are two categories of drive mechanisms: (1) direct application of compressed gas via a pressure reducing valve and (2) indirect application via an electric motor or compressor.

Compressed Gas and Pressure Reducing Valve. When compressed gas is used for the drive mechanism, its force is adjusted via a pressure-reducing valve. A pressure-reducing valve (PRV) reduces high input pressure to a lower constant output pressure. This pressure may be set as high as 50 psi or be adjusted to a few cmH_2O. The functioning principle of an adjustable PRV can be described by

$$P_L = \frac{(F_{S1} - F_{S2}) - (P_H \times a)}{A}$$

where
P_L = Low or reduced pressure (generated pressure)
F_{S1} = Force of the large spring (adjusting spring)
F_{S2} = Force of the small spring (sealing spring)
P_H = High input pressure (source pressure)
a = Area of the small seat
A = Area of the large diaphragm

After gas is reduced from the PRV, it is delivered to the patient by one of the following approaches:

1. Directly inflate the lung, as the Puritan-Bennett 7200.
2. Prime a spring weighted bellows as in the Servo 900C.

Electric Motor and Compressor. The drive system in electrically powered ventilators consists of an electrical motor that drives a compressor. The compressor may use either a rotating crank and piston, rotary vane, or linear drive motor. Of particular interest is the type of motor and linkage system employed, because this determines the waveform the ventilator will produce.

Electric Motor/Rotating Vane

Referred to as nonlinear pistons (16), eccentric wheel (17), and rotary pistons (18), these units have a constant speed-rotating wheel with a piston rod connecting a piston to the outer edge of the wheel. As the wheel turns, the piston's actual movement is not constant. Consequently, the pressure and flow developed by this mechanism varies with the motion of the piston.

Electric Motor/Rotating Vane

A rotary vane compressor consists of an electric motor with vanes (blades) attached to the motors shaft, which rapidly spin inside a cylinder. This action generates a constant level of pressure to serve as a drive mechanism. A series of electrically controlled switches (solenoids) serves to assure that the patient is connected to the breathing circuit only during inspiration. During exhalation, the pressure generated by the blower is vented to the atmosphere.

Electric Motor/Linear Piston

Linear piston drive systems may be a rack-and-pinion gear system, a high-tension spring driven system, or a high-pressure pneumatic drive system. This system may be constructed as a single-circuit or double-circuit scheme. In a single circuit, gas is sent directly to the patient from the device's power circuit. The piston's movement is constant throughout its stroke. Therefore, the flow of air being delivered to the patient is also constant, resulting in a square flow wave and linearly increasing pressure curve. A double–circuit design is constructed so that gases from a power circuit are used only to compress a bag–type apparatus (bellows) containing separate gases, which are sent to the patient through a patient circuit. This results in a progressively increasing flow pattern and initially an accelerating pressure pattern. Once the pressures start to equilibrate, a drop in circuit pressure occurs producing a notch in the pressure curve.

Output Control Systems. An output control system is designed to control the flow of gas to a patient. Most ventilators used in intensive care units employ a series of valves to accomplish this. In earlier models (Puritan Bennett MA-1), it was a simple ON/OFF exhalation valve. Current technology enables a valved component to be controlled with microcomputer intelligence. This allows for precise control over gas flow delivery that was not feasible with the compressor–bag delivery system used in earlier generations of mechanical ventilators. For example, the Puritan Bennett 7200 ventilators can precisely manage the movement of its output control system allowing for 3095 movements within 0.003 of an inch (19). The most common valves are discussed below.

Pneumatic Poppet Valve. Sometimes referred to as electromagnetic valves, Plunger valves, and Poppet valves (20,21). This type of device uses a magnetic force to control gas pressure in an ON/OFF fashion. A given valve regulates flow by opening or closing a particular opening. By digitally controlling the open/close sequence of the valve, flow output can be varied.

Proportional Solenoid Valves. Proportional solenoid valves can control flow incrementally instead of turning flow "ON and OFF." This allows for shaping of the inspiratory gas flow being delivered to the patient. Most current generation ICU ventilators use some type of proportional solenoid technology.

Control Scheme

Mechanical ventilators coordinate the delivery of gases based on four variables: flow, volume, time, and pressure. Therefore, the control variable becomes the variable that the ventilator manipulates to create inspiration. A single breath does not reflect significant changes in compliances and resistance. However, flow, volume, time, and pressure can shift and are regulated by the ventilator. In general, if flow, volume, time, or pressure are individually preprogrammed, the others must respond. The control variable will remain constant in the presence of changes in impedance (resistance and compliance). For example, a ventilator is a volume controller if the volume is measured and used to control the volume signal. Many of today's volume ventilators are truly *flow controllers*. They derive a volume measurement from a flow signal. The signal is measured and calculated and displayed as a volume unit of measure on the ventilator display window.

Now that we have discussed *what* is controlled, let us examine *how* it is controlled. In the open–loop system, the user selects the variable (volume, pressure, and flow), for input and observes the output variable (volume, pressure, and flow), which mirrors the actual valve. The operator handles any needed adjustments. The closed–loop system analyzes the data using a feedback signal and automatically modifies the input in an attempt to match the reference value.

Phase Variables. Control of the breath requires the ventilator to perform four phases during a breathing cycle (22):

1. End of expiratory phase (beginning of inspiration)
2. Inspiration
3. End of inspiration (beginning of expiratory phase)
4. Expiration

During these phases, the mechanical ventilator assumes all or part of the work of breathing. During each stage, a certain variable is used to start (trigger), sustain (limit), and end (cycle) the breath.

Trigger. Inspiration is initiated when one of these variables reaches a preset value. This describes

the "trigger" phase of the ventilator. With patient triggering, the ventilator senses the patient effort to initiate gas delivery. The breath is flow, volume, or pressure triggered if the patient initiates the breath. Most modern adult ventilators have a frequency, breathing rate, or respiratory rate control knob that serves as the time trigger. Manual trigger is available on most ventilator control panels as well.

Limit Variable. A limit variable is a parameter that can reach a preset level before inspiration ends, but it does not terminate the inspiration phase (23). This is often confused with a cycle variable. A limit variable does not terminate inspiration; it sets the upper boundaries for pressure, volume, or flow. An example present in many critical care ventilators is the ability for an operator to perform an "inflation hold" of plateau maneuver. The ventilator is allowed to deliver a preset volume, but inspiration does not end until a preset time interval has elapsed. The volume is limited (held and maintained) until the preset pause time is reached.

Cycle. Inspiration will end when either time, pressure, flow, or volume reaches its preselected value. Unlike the "limit" variable, the inspiration phase will end (cycle off).

Baseline Variable. The variable controlled during the expiratory phase is the baseline variable (24). Although flow or volume can serve as the baseline variable, pressure is the most commonly used. Expiratory pressure can be set to atmospheric—Zero. This is called ZEEP (zero end expiratory pressure). ZEEP is obtained when the ventilator's exhalation valve opens to the atmosphere, exposes the patient's airway to the atmosphere, and exposes the patient's airway to a relative pressure of zero. Pressures exceeding atmospheric are referred to as PEEP (positive end expiratory pressure) or CPAP (continuous positive airway pressure). Two well-documented phenomena applicable to PEEP/CPAP therapy are reduced functional residual capacity (FRC) and extravascular lung water (25–27).

Conditional Variables. Specific patterns of breath delivery are examined by the microprocessors control logic system. The ventilator may be set to keep the breath constant or perform more complex maneuvers. The ventilator algorithm for a particular set of breath delivery patterns makes this determination.

Modes of Ventilation. A mode of ventilation represents a combination of control, phase, and conditional variables that establish a set pattern of mandatory and/or spontaneous breaths (28).

Output. Most ventilator waveforms can be classified as one of the following:

1. Exponential
2. Ramp
3. Rectangular
4. Sinusoidal

No ventilator is an ideal controller, and ventilators are designed to only approximate a particular waveform (29,30).

Alarm Systems. Ventilator alarm systems have increased in number and complexity (31) and have recently begun to draw national attention on its proper use in and out of the critical care area (32).

VENTILATOR OPERATION

Introduction

The scope of medical practice expanding to accommodate the needs of those with respiratory diseases is immense. The field of respiratory therapy alone encompasses over 35,000 persons in the United States (33). As one may glean from the previous section of this article, not unlike other areas of medicine, when the need for new technology and therapeutic interventions is required, along with it will come a vast array of specialists, technical support, ancillary staff, and commercial entrepreneurs. The use of ventilators is expensive and represents approximately 10% incidence and contributes significantly to hospital costs (34). Although these costs are impressive, so are the diseases from which the patients suffer, and the true cost is patient mortality, which is invariably high (35). To that end, we have experienced increased understanding of the pathophysiology of their diseases as well as have enjoyed a significant increase in the technologic advances to support them.

The initial aim of ventilatory support has been justified to aid those persons suffering from neuromuscular disorders such as poliomyelitis or tetanus. The practice of medicine precludes one from solely defining a disease and indicating an intervention. But, rather, it identifies a derangement from normal and determines whether it causes harm. Should this be the case, then one should implement a therapeutic change. In the case of mechanical ventilatory support, we refer to the concept of work of breathing (WOB) (36).

$$\text{Work} = \text{Force} \times \text{Distance}$$
$$\text{Force} = \text{Pressure} \times \text{Area}$$
$$\text{Distance} = \text{Volume}/\text{Area}$$

$$\text{Work} = (\text{Pressure} \times \text{area})(\text{Volume}/\text{Area})$$
$$\text{Work} = \text{Pressure} \times \text{Volume}$$

$$\text{WOB} = \text{Tidal Volume} \times \text{Transpulmonary Pressure}$$

The tidal volume that is normally generated results from a relatively small pressure gradient from the intrapleural space to the atmosphere drawing gas inward. Exhalation is then passive when the pressure gradient is reversed, and the pressure then exceeds atmospheric. The relationship between pressure and volume is known as compliance. Pulmonary parenchymal (Cp) compliance is one component of total lung compliance (C_T) that would

also include the chest wall and rib cage (36).

$$C_T(\text{L/cm H}_2\text{O}) = \Delta V(\text{L})/\Delta P(\text{cm H}_2\text{O})$$
$$1/C_T = 1/C_L + 1/C_{CW}$$

Suffice it to say that most disease processes that impart ictus on patients and require ventilatory support alter pulmonary compliance, usually decreasing it. For example, a patient with pneumonia will generally have areas of alveoli, which are normally air-filled, collapsed, and filled with a pus-like edema matter. This fluid decreases lung compliance in the patient and increases the work (WOB) one must do to generate the same tidal volume to breathe. When this tidal volume is sacrificed, minute ventilation is diminished unless respiratory rate (RR) is increased.

$$\text{Minute ventilation} = \text{RR} \times \text{tidal volume}$$

Minute ventilation is a universal measure of the amount of gas that moves in and out of the lungs and sometimes indicates their global ability to remove carbon dioxide. Blood is delivered to the alveoli, and over a tremendous surface area, CO_2 diffuses freely and expires. As CO_2 diffusion is efficient, this process is uncommonly compromised. However, ventilatory failure is common. When a disease process, either of pulmonary etiology or otherwise, overcomes the normal compensatory mechanisms of the body, such that the patient can no longer maintain a normal pH *and* a normal $P\text{aCO}_2$, acute ventilatory failure is present. In this circumstance, the blood pH falls below a normal level and the $P\text{aCO}_2$ of arterial blood rises. The imposed, detrimental, WOB requires that the lungs and muscles of ventilation (i.e., chest wall, intercostals, sternocleidomastoid, and diaphragm) consume the oxygen that is delivered to them, hence, complicating matters because this oxygen is intended for the rest of the patient.

The remaining physiologic systems will begin to falter and malfunction as the acidemia worsens due to the fact that many proteins in the body operate within a narrow pH range (heart, brain, kidneys). This process typically will reveal a patient who is breathing rapidly and shallowly. The associated signs often include diaphoresis and the use of accessory muscles. However, of most concern, patients will describe a feeling of breathlessness, even that they are about to die. These are the patients with a clinical diagnosis of impending ventilatory failure.

Respiration Versus Ventilation

The terms "respiration" and "ventilation" are often incorrectly used interchangeably. It is imperative to offer a brief discussion depicting their differences. Respiration refers to the to and fro movement of a gas across a membrane. In the specific instance of breathing, we will consider the gases oxygen and carbon dioxide. Whereas oxygen is inhaled and taken up by the capillary endothelium and removed by the pulmonary venous system, carbon dioxide is delivered to the lungs by the pulmonary artery where it is eliminated by freely diffusing across the alveolar membrane. Typically, diffusion of carbon dioxide is very rapid, in contrast to oxygen, which requires approximately one third of the course of the capillary (37). The cardiopulmonary system maintains oxygen and carbon dioxide tensions within the normal range (and a normal pH) despite changing metabolism. Any derangement from the normal compensatory homeostatic mechanisms (i.e., disease) may be life-threatening (38).

Ventilation simply defines the movement of a gas in and out of the pulmonary system (38). Clinical medicine makes a distinction and refers to carbon dioxide as a marker for ventilation. This is essentially because the tension of CO_2 inspired is usually zero, and is therefore a better indicator for the cardiopulmonary system's ability to excrete it (38). Modern ventilators, although complex, are designed simply to improve oxygenation, ventilation, or both. Germane to the following, they are not always mutually exclusive.

Mechanical Ventilatory Support: Indications

Many have classified respiratory distress and the need for mechanical ventilatory support into categories that are useful. In addition, a thorough understanding of the physiologic derangement allows selection of the most optimal mode of support. However, simple definitions that are clear allow straightforward clinical judgment so that an intervention can be made and later fine tuned, (Table 1).

If one was to obtain an arterial blood gas and find a pH of less than 7.30 *and* a $P\text{aCO}_2$ greater than 60 mm Hg, they would meet criteria for acute ventilatory failure. This is because, at both this pH and $P\text{aCO}_2$, the CSF hydrogen ion concentration creates a maximal stimulus to breathe and represents an *acute* process. Impending ventilatory failure refers to the patient who is clearly in respiratory distress and frequently is found to be diaphoretic, tachypneic, using accessory muscles, and claims to be short of breath as described above. Apnea is intuitive, and the clearest example is a patient suffering from cardiac arrest. Although intermittently controversial, yet clinically important, hyperventilation for those patients with intractable intracranial hypertension will benefit from lowering carbon dioxide tensions with mechanical ventilatory support. Lastly, the concept of total ventilatory support is reserved for those persons with acute lung injury or adult respiratory distress syndrome at risk for ventilator-associated injury and severely decreased pulmonary compliance.

Mechanical Ventilatory Support: Goals

It seems appropriate to ascertain salient endpoints once a therapeutic intervention has been implemented such as mechanical ventilatory support. In this circumstance, if the ventilator can assume some or all of the patient's WOB and alleviate the detrimental component, it would be helpful to classify the level of support provided to the patient

Table 1. Indications for Mechanical Ventilatory Support

Indication	Comments
Acute ventilatory failure	ABG diagnosis
Impending ventilatory failure	Physical diagnosis
Apnea	Cardiac arrest
Therapeutic hyperventilation	Neurosurgery
Total ventilatory support	Lung protective strategy

Table 2. Levels of Ventilatory Support

Support Level	Work of Breathing	Eucapnia and Normal pH
Partial ventilatory support	Significant	Yes
Full ventilatory support	Not required	Yes
Total ventilatory support[a]	Not required	No

[a]For use in patients with abnormal lung compliance.
Reprinted with permission from Reference 39.

based on the amount of work the *patient* is *required* to provide to achieve a normal pH and a normal carbon dioxide tension in arterial blood (Table 2).

Full ventilatory support is likely the most common mode of ventilatory support employed in the postoperative period or immediately after intubation for one of the above indications for institution of mechanical ventilatory support. The aim of this intervention is to alleviate all of the patient's WOB until whatever medical or surgical disease has begun to resolve. Once a patient's respiratory status has stabilized, it seems prudent to allow one to resume spontaneous ventilatory effort and a significant amount of WOB.

It seems intuitive that once a breathing tube has been placed and ventilatory support has been instituted, it would be wise to work toward retuning the patient to a state free from those devices. Interestingly, there is evidence to support the concept that patients may not demonstrate the hypotensive effects of positive pressure ventilation in the presence of hypovolemia or decreased pulmonary compliance (40). In most circumstances, it is commonplace to transition to partial ventilatory support using various volume-or pressure-limiting modes (see below). As one is "weaned" from support, it is imperative that several items are taken into consideration. It has been suggested that some protocol be followed to standardize the routine and enhance safety. One such protocol, involves a daily screen and a spontaneous breathing trial (41).

Advantages that have been suggested include shorter days on mechanical ventilation, shorter intensive care unit stays and days in the hospital, and lower cost. To that end, safety still seems to be an appropriate concern. Patients must be monitored appropriately for any sudden deterioration in clinical status. Aside from standard intensive care unit monitoring, recommended additional parameters would include evidence of adequate ventilation by arterial blood gas analysis. Should this not be available, continuous end-tidal carbon dioxide capnometry is an alternative. However, intermittent arterial pH and $PaCO_2$ should be obtained to correlate with the $EtCO_2$ values. The best monitor for the adequacy of oxygenation and ventilation remains the patient. Even while ventilated, it is imperative to engage in frequent assessments.

Modes of Support: Volume Versus Pressure

Despite all best efforts, at times there is a need for postoperative mechanical ventilation for a period of time. This may be due to residual anesthetic effects or the effects of surgery, especially abdominal or thoracic, on an already compromised respiratory status. When providing this support, several different modes of mechanical ventilation can be used, depending on the clinical situation at hand. Generally, during a period of acute stabilization, it is best to use a volume preset–pressure variable mode of ventilation (i.e., synchronized intermittent mechanical ventilation) and then to change to a volume variable–pressure preset mode (i.e., pressure support) when the patient is ready for withdrawal of mechanical ventilation (42). However, in postoperative patients, even in those with severe lung disease, the mode of ventilation is not as important as the endpoints of ventilation that are chosen (42).

The flexibility of modern mechanical ventilators has allowed for other considerations such as the degree of patient triggering, whether the breath is limited by pressure or flow, and the degree to which these pressures and volumes support the patient (Table 3).

An understanding of the limitations of the various modes is essential to success. For example, if one was to place a patient on assist/control mode and subsequently decrease the respiratory rate in an effort to wean the patient from support, the patient would continue to receive the preset tidal volume at his or her intrinsic respiratory rate and would not gain any intrinsic WOB. To the contrary, this method of weaning is perfectly acceptable in the synchronized intermittent mandatory ventilation mode (SIMV). In such a case, decreasing the SIMV rate would leave the patient to generate whatever tidal volume he or she could as determined by their transpulmonary pressure and pulmonary compliance (Fig. 1).

An alternative approach would be to change a patient from a volume-targeted mode of ventilation (i.e., SIMV or A/C) and convert to pressure support ventilation (PSV) for weaning purposes, perhaps. Having some discrete

Table 3. Characteristics of Volume-Targeted Ventilation (VTV) and Pressure-Targeted Ventilation (PTV)

Variable	Volume TV	Pressure TV
Trigger	Patient or time	Patient or time
Limit	Flow	Pressure
Cycle	Volume	Time or flow
Tidal volume	Constant	Variable
Peak pressure	Variable	Constant
Modes	Assist/control (synchronized) intermittent Mandatory ventilation	Assist/control (synchronized) intermittent Mandatory ventilation Pressure support

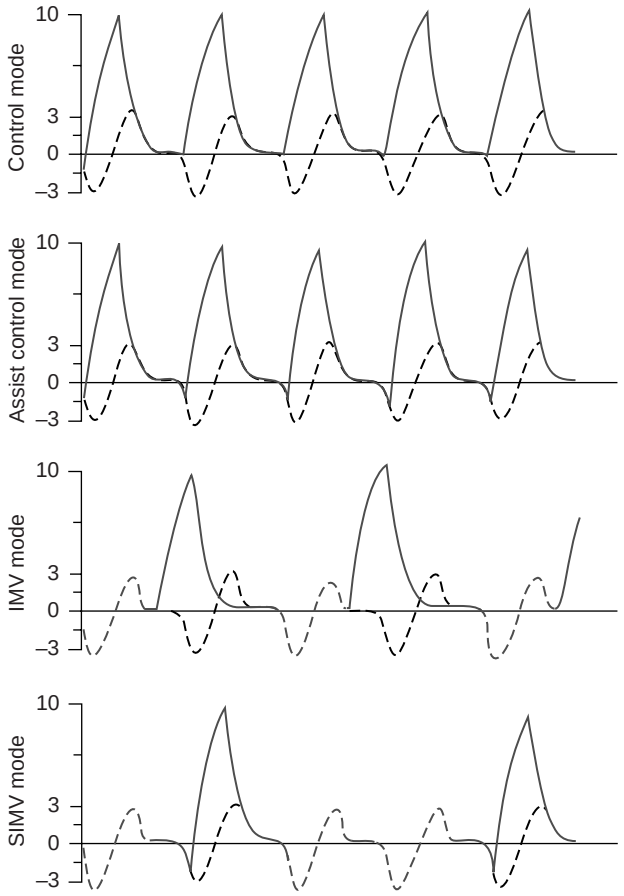


Figure 1. Airway pressure tracings of the four standard volume-preset modes. *Thick solid lines* represent ventilator breaths; *thick dotted lines* represent spontaneous breaths; and *thin dotted lines* refer to what the spontaneous pattern would have been without the ventilator breaths. IMV, intermittent mandatory ventilation; SIMV, synchronized IMV. Reprinted with Permission from Reference (36).

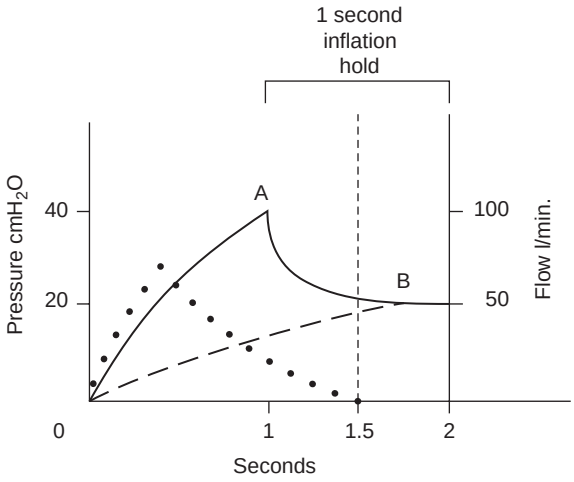


Figure 2. Schematic illustration of inspiratory cycle for volume-preset (1000 mL) tidal volume delivered by square-wave generator in 1 s, followed by 1-s inflation-hold maneuver. Note that the inspiratory cycle is 2 s, although tidal volume is delivered by the ventilator in 1 s. Normal compliance and resistance are assumed. Proximal airway pressure is shown as a *solid line*, alveolar pressure as a *dashed line*, and gas flow in airways as a *dotted line*. Peak airway pressure (*A*) is achieved near the end of tidal volume delivery, at which point there is considerable gradient between peak airway pressure and alveolar pressure that results in continued flow to the lung. Airway pressure rapidly diminishes as flow continues toward alveoli. Note that measurable flow in the system has essentially ceased 0.5 s after the ventilator has delivered tidal volume (*B*); however, a gradient remains between airway and alveolar pressures. An elevated peakpressure (*A*) in the presence of normal plateau pressures indicates increased airway resistance, which may include bronchospasm. To attain true plateau pressure, in which the airway pressure reflects the average peak alveolar pressure, an additional 0.5 s of inflation hold is required after the absence of measurable flow in the system. Reprinted with permission from Reference 42.

knowledge of the patient's arterial blood gas and a minute ventilation set on the ventilator commensurate with adequate oxygenation and ventilation, alternating between either, mode can mode be done easily. For example, if one has a ventilator set to deliver an SIMV respiratory rate of 10 by 700 mL, this then equates to about a 7 L minute volume. If the plateau pressure (point B in Fig. 2) generated is roughly 20 cm H_2O, a reasonable pressure support setting would then be 20 cm H_2O. Provided the patient respires about 10 times a minute, the same degree of oxygenation and ventilation should be anticipated. The minute volume would roughly be similar, and the volume-pressure relationship remains in tact provided there is no abrupt change in the patient's clinical condition.

As a patient's pulmonary disease regresses, lung compliance should improve. Using the ventilator monitor, the tidal volume that is generated from a pressure support breath should increase or the plateau pressure generated from a volume cycled breath should decrease (point B in Fig. 2). This can be a valuable piece of information often overlooked in the daily management of critically ill patients. Any difference between the peak and plateau pressures (points A and B in Fig. 2) would indicate an increase in airway resistance from bronchospasm, inspissated secretions, kinking of the tracheal tube, or some other form of incomplete obstruction.

PEEP/CPAP

Many patients suffer from a deficit in arterial oxygenation. The use of CPAP or PEEP is an effective means to remedy the malady. To clarify, the physiologic effect of CPAP/PEEP is identical; however, the nomenclature will vary depending on whether the patient is receiving positive pressure ventilation. One should use the term CPAP if the patient is not receiving positive pressure ventilation and is spontaneously ventilating and PEEP if the patient is receiving positive pressure.

Increasing the fraction of inspired oxygen, although simple and seemingly effective, to treat arterial hypoxemia may not be entirely benign. There is evidence that exposure of alveolar membranes to prolonged periods of high concentrations of oxygen will induce an oxygen-free radical species and subsequently lung injury (38). In addition,

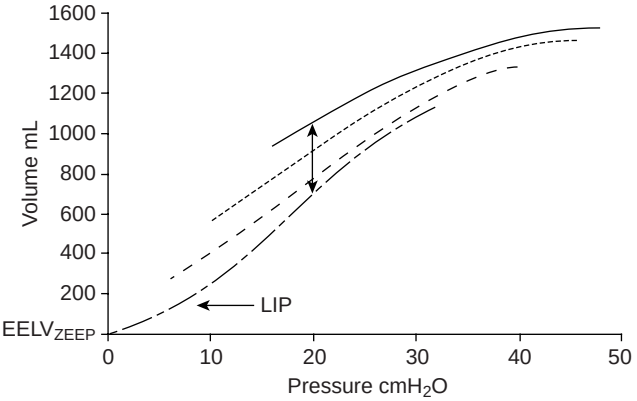


Figure 3. Multiple pressure/volume (P/V) curves of the respiratory system recorded from different levels of PEEP (– – – : PEEP 5; ═: PEEP10; —: PEEP15) and from zero end-expiratory pressure (ZEEP; -·-·-), and related to the elastic equilibrium volume of the respiratory system at ZEEP (EELV_{ZEEP}). The curves are aligned on the same P/V diagram after correcting the starting points of each PEEP P/V curve for the volumes measured during passive expired spirograms performed at each PEEP level. The upward shift of PEEP P/V curves indicates alveolar recruitment. The recruited volume with 15 cm H_2O PEEP (double-headed arrow), compared with ZEEP, is quantified by the volume difference between the curves, for the same level of airway pressure (20 cm H_2O in this example). In other terms, when PEEP is applied, for the same level of airway pressure, lung volume is greater than without PEEP, suggesting the reopening of some alveolar units previously collapsed at ZEEP. Note that recruitment continues far above the value of pressure at the lower inflection point (LIP) and above to the upper inflection point. The different P/V curves tend to join at higher lung volumes, suggesting that the maximal lung volume is approached. Reprinted with Permission from Reference 43.

even shorter periods of high levels of inspired oxygen fractions limit the amount of nitrogen that may "stent" open; an alveolar unit can contribute to alveolar collapse, a phenomenon termed *denitrogenation absorption atelectasis* (38). The Law of LaPlace states that the pressure that tends to keep an alveolus open is twice the wall tension (T) divided by the radius and is represented by

$$P = 2T/r$$

To that end, many employ CPAP/PEEP therapy to improve arterial oxygenation and decrease WOB. WOB may be decreased by shifting a patient to a more favorable portion of the pulmonary compliance curve (Fig. 3).

On the flatter part of the curve, less tidal volume is generated with a substantial change in pressure. According to Fig. 3, on the steeper portion of the curve, less pressure is needed to generate a larger volume, which results in less work. By definition, the alveolar units would remain open at end-expiration, continually participating in gas exchange. It is this very concept of "open lung" ventilation that is thought to be the mainstay of avoiding ventilator-induced injury and ARDS (44,45).

CPAP/PEEP in most instances will have a significant impact on arterial oxygenation. Bringing alveoli into closer contact with the pulmonary capillaries will provide a greater surface area for gas exchange, and as such, the pulmonary venous blood leaving the lungs will have a higher oxygen content. If one could make existing alveoli larger, opening previously closed alveoli, arterial oxygen tension would improve. At CPAP/PEEP levels less than 10 cm H_2O, existing open alveoli are made larger. When CPAP/PEEP levels exceed 15–50 cm H_2O, the collapsed alveoli are then opened and can participate in gas exchange, and thus they are recruited.

This fluid typically does not move out of the lung to some other area of the body. Rather, it shifts within the small spaces of the alveolar unit to the lose connective tissue-filled peri-bronchiolar areas that do not affect gas exchange. Perfusion will better match ventilation, thus increasing the oxygen content of arterial blood.

Mechanical Ventilatory Support: Adverse Reactions

Just as with any drug or medication that is prescribed, ventilators are not without iatrogenic potential. Typical side effects may include inadvertent hyperventilation or patient dysynchrony. These issues can be avoided with careful monitoring of the patient. Meticulous attention to artificial airway maintenance is mandatory. Relative to mechanical ventilatory support, the usual culprit of iatrogenic injury is excessive positive pressure termed *barotrauma* and large tidal volume in specific patients at risk for acute lung injury or ARDS referred to as *volutrauma*.

Barotrauma is simply the disruption of the large conduction airways when exposed to very high pressures. When this occurs, gas leaks into the tissues of the neck

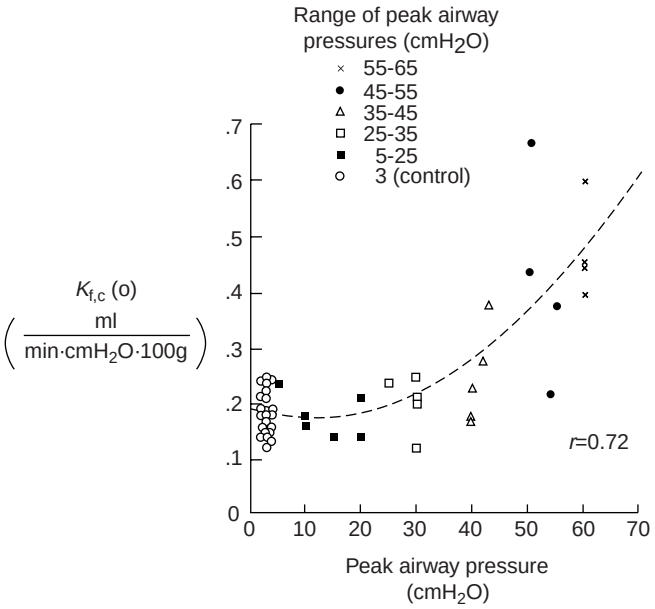


Figure 4. Changes in the capillary filtration coefficient ($K_{f,c}$) of isolated blood-perfused lobes of dog lungs given 20 min of intermittent positive pressure ventilation. Moderate (up to 30 cm H_2O) increases in peak airway pressure did not affect $K_{f,c}$, whereas higher peak pressure produced a steep increase in $K_{f,c}$. Reprinted with Permission from Reference 47.

and chest such as the mediastinum (pneumomediastinum) or the pleural space (pneumothorax). The most life-threatening of these is the tension pneumothorax, which allows gas to fill and accumulate in the pleural space. If there is no mechanism to remove it, the accumulated gas compressed the mediastinum and cardiac chambers compromising cardiac output to the point where venous return ceases and left ventricular output approaches zero.

ARDS, also referred to as acute lung injury, has been studied at great length. It has been noted that more patients survive in certain circumstances if lower tidal volumes are used to ventilate their lungs. In addition, it has been proven in several instances that chemical mediators of inflammation as released and adversely affecte the body of large tidal volumes that are used (46) (Fig. 4). Recently, much has been learned about ARDS, and patient populations at risk for ventilator-induced injury have been identified.

BIBLIOGRAPHY

1. Mushin M, Rendell-Baker W, Thompson PW, Mapleson WW. Automatic Ventilation of the Lungs. Oxford: Blackwell Scientific Publications; 1980. 62–166.
2. Dupuis YG. Ventilators—Theory and Clinical Application. St. Louis, MO: CV Mosby Company; 1986.
3. McPherson SP. Respiratory Therapy Equipment, 3rd ed. St. Louis, MO: CV Mosby Company; 1985.
4. Cairo JM, Pilbeam SP. Respiratory Care Equipment, 6th ed. St. Louis, MO: CV Mosby Company; 1999.
5. Chatburn RL. Classification of mechanical ventilator. Respiratory Care 1992;37:1009–1025.
6. Kacmarek RM. Critical Care Ventilators, Foundations of Respiratory Care.
7. ECRI. http://www.ecri.org.
8. Blanch PB, DeSaultels DA. Chatburn's ventilator classification scheme—poor substitute for the classic approach. Respiratory Care 1994;39:762.
9. Consensus Statement on the essentials of mechanical ventilators—1992. Respiratory Care 1992;37:1000–1008.
10. Chatburn RL. Classification of mechanical ventilators. Respiratory Care 1992;37:1009–1025.
11. Branson RD, Chatburn RL. Technical description and classification of modes of ventilator operation. Respiratory Care 1992;37:1026–1044.
12. Morris W. The American Heritage Dictionary of the English Language. Boston, MA: American Heritage and Houghton Mifflin; 1975.
13. McPherson SP. Respiratory Therapy Equipment, 3rd ed. St. Louis, MO: CV Mosby Company; 1985.
14. Dupuis YG. Ventilators—Theory and Clinical Application. St. Louis, MO: CV Mosby Company; 1986.
15. McPherson SP. Respiratory Therapy Equipment. 3rd ed. St. Louis, MO: CV Mosby Company; 1985.
16. Dupuis YG. Ventilators—Theory and Clinical Application. St. Louis, MO: CV Mosby Company; 1986.
17. Chatburn RL. Classification of mechanical ventilators. In: Branson RD, Hess DR, Chatburn RL, editors. Respiratory Care Equipment. Philadelphia, PA: JB Lippincott Company; 1995.
18. Spearman CB, Sheldon RL, editors. Egans Fundamentals of Respiratory Therapy. St. Louis, MO: CV Mosby Company; 1982.
19. http://www.puritanbennell.com.
20. Chatburn RL. Classification of mechanical ventilators. In: Branson RD, Hess DR, Chatburn RL, editors. Respiratory Care Equipment. Philadelphia, PA: JB Lippincott Company; 1995.
21. Chatburn RL, Scanlan CL. Ventilator modes and functions. In: Scanlan CL, Wilkins RL, Stoller JK, editors. Egan's Fundamentals of Respiratory Care. 7th ed. St. Louis: Yearbook Mosby; 1999.
22. Mushin M, Rendell-Baker W, Thompson PW, Mapleson WW. Automatic Ventilation of the Lungs. Oxford: Blackwell Scientific Publications; 1980.
23. Chatburn RL. Classification of mechanical ventilators. Respiratory Care 1992;37:1009–1025.
24. Kacmarek RM. Critical Care Ventilators, Foundations of Respiratory Care.
25. Daly BDT, Edmonds CH, Norman JC. In vivo alveolar morphometrics with positive end expiatory pressure. Surg Forum 1973;24:217.
26. McIntyre RW, Laws AK, Ramachandran PR. Positive expiratory pressure plateau: Improved gas exchange during positive pressure ventilation. Can Anesth Soc J 1969;16:477.
27. Pare PD, Wariner B, Baile M, et al. Redistribution of pulmonary extravascular water with positive end expiratory pressure in canine pulmonary edema. Am Rev Respir Dis 1983;127:590.
28. Chatburn RL. Classification of mechanical ventilators. Respiratory Care 1992;37:1009–1025.
29. Kacmarek RM. Critical Care Ventilators, Foundations of Respiratory Care.
30. Chatburn RL. Classification of mechanical ventilators. In: Branson RD, Hess DR, Chatburn RL, editors. Respiratory Care Equipment. Philadelphia, PA: JB Lippincott Company; 1995.
31. Chatburn RL. Mechanical ventilators: Classification and principles of operation. In: Hess DR et al., editors. Respiratory Care, Principles and Practice. Philadelphia, PA: WB Saunders Company; 2002.
32. JCAHO Alarm event. http:/www.jcaho.org.
33. http://www.aarc.org.
34. Neddham DM, Bronskill SE, Sibbald. Mechanical ventilation in Ontario, 1992–2000: Incidence, survival, and hospital bed utilization of non-cardiac surgery adult patients. Criti Care Med July 2004.
35. Depuydt PO, Benoit DD, Vandewoude KH. Outcome in noninvasively and invasively ventilated hematologic patients with respiratory failure. Chest October 2004.
36. Miller. Anesthesia. 5th ed. Philadelphia, PA: Churchill Livingstone.
37. Guyton. Textbook of Medical Physiology. 10th ed.
38. Shapiro BA, Peruzzi WT, Templin R. Clinical Application of Blood Gasses, 5th ed.
39. Peruzzi WT, Shapiro BA. Changing practices in ventilator management: A review of the literature and suggested clinical correlations. Surgery 1995;117:121–133.
40. Stock C, Perel A. Handbook of Mechanical Ventilatory Support. 2nd ed. Philadelphia, PA: Williams and Wilkins; 1997.
41. Murray J, Coursin DB, Pearl RG. Critical Care Medicine. Philadelphia, PA: Lippincott Williams, Wilkins.
42. Tamul PC, Peruzzi WT. Assessment and management of patients with pulmonary disease. Crit Care Med April 2004.
43. Maggiore SM, Richard JC, Brochard L, What has been learned from P/V curves in patients with acute lung injury/acute respiratory distress syndrome. Eur Respir J 2003.
44. Grasso S, Mascia L, Del Turco M. Effects of recruiting maneuvers in patients with acute respiratory distress syndrome ventilated with protective ventilatory strategy. Anesthesiology April 2002.

45. Schreiter D, Reske A, Stichert B. Alveolar recruitment in combination with sufficient positive-end expiratory pressure increases oxygenation and lung aeration in patients with severe chest trauma. Crit Care Med April 2004.
46. Amato MB, Barbas CS, Medeiros DM. Effect of a protective-ventilation strategy on mortality in the acute respiratory distress syndrome. NEJM Feb 1998.
47. Dreyfuss D, Sauman G. Ventilator-induced lung injury: Lessons from experimental studies. Am J Respir Crit Care Med 1998; 157.

See also CARDIOPULMONARY RESUSCITATION; CONTINUOUS POSITIVE AIRWAY PRESSURE; VENTILATORY MONITORING.

VENTILATORY MONITORING

STEVEN R. KNOPER
STUART F. QUAN
University of Arizona College of Medicine

INTRODUCTION

One of the most important aspects of the clinical care of patients with known or suspected cardiopulmonary disease is monitoring the adequacy or status of their ventilatory function. However, in order to understand why various parameters of lung function are monitored, a few aspects of basic respiratory physiology require review. Although the lungs perform important metabolic and endocrine functions, their primary function is gas exchange. They provide the vehicle by which oxygen (O_2) in inspired air moves into venous blood and carbon dioxide (CO_2) in venous blood is released into expired gas. The mechanisms by which these processes occur are complex, but can be subdivided into four phases. The first phase involves the movement of gas into and out of the lungs themselves. This aspect of respiratory physiology is termed lung mechanics. During spontaneous breathing, contraction of the diaphragm and intercostal muscles increases negative pressure within the intrapleural space, leading to lung expansion. This results in a pressure gradient along the airways causing gas to flow into the lungs and to eventually reach the alveoli. With relaxation of the inspiratory muscles, the elastic recoil of the lungs reverses the pressure gradient resulting in gas being expired through the mouth into the atmosphere. In the second phase, gas exchange occurs in the alveoli. By the process of diffusion across the alveolar capillary membrane, O_2 enters and CO_2 exits the vascular system. The third phase involves the transport of O_2 and CO_2 in the blood, and is intimately related to adequate cardiovascular and hematologic functioning. Oxygen is carried in the blood primarily in a complex with hemoglobin in the red blood cells. A small amount of O_2 is also dissolved in plasma. The total amount of O_2 delivered throughout the body is a function of the volume of O_2 in the blood, and the rate of delivery or cardiac output. Transport of CO_2 in the body is more complex, but also is related to the properties of hemoglobin and cardiac function. The fourth phase is cellular respiration (defined as the metabolic consumption of O_2 and production of CO_2 and H_2O), and involves the mechanisms by which cells in the body utilize O_2 and excrete CO_2. All of these phases are related to each other, and are essential for normal respiration to occur. Ventilatory monitoring encompasses methods by which the functional status of the aforementioned aspects of lung function can be determined. This article reviews the indications, principles, and techniques used in clinical ventilatory monitoring. A detailed discussion of pulmonary function testing is not included, and is reviewed elsewhere (1).

INDICATIONS FOR VENTILATORY MONITORING

Ventilatory monitoring usually is performed for four general clinical indications: (1) diagnostic, (2) therapeutic, (3) prophylactic, and (4) safety considerations. With respect to diagnostic use, the goal of monitoring is to obtain information pertinent to making a specific etiologic diagnosis of a disease entity or process. For example, in myasthenia gravis patients who are having an acute deterioration in ventilatory function, vital capacity (VC) measurements (the maximum volume excursion of which the lungs are capable by voluntary effort) are obtained after injection of a short-acting cholinesterase inhibitor (edrophonium, Tensilon). Documentation of an increase in VC suggests the need for an increase in anticholinesterase medication, whereas a deterioration is evidence for excessive effect from medication. Another frequently encountered situation where ventilatory monitoring is used diagnostically is the continuous monitoring of airflow and ventilatory effort in sleep disorders laboratories to detect evidence of sleep disordered breathing (2).

A number of ventilatory parameters are useful in monitoring the therapeutic effect of specific treatment modalities. The most frequently encountered situation is where arterial O_2 and CO_2 gas tensions are sampled following institution of supplemental O_2 or mechanical ventilation. Measurement of respiratory system compliance following application of positive end-expiratory pressure (PEEP) in mechanically ventilated patients with acute respiratory failure is another example of where ventilatory monitoring is used to determine the effects of a treatment modality (3). In both the aforementioned examples, therapy frequently is adjusted using results obtained from the ventilatory parameters monitored.

Ventilatory monitoring commonly is performed prophylactically with the goal of detecting a perturbation in normal or baseline respiratory function so that treatment can be instituted early. For example, apnea monitors are employed in infants suspected to be at risk for sudden infant death syndrome. In this clinical situation, alarms are activated if life-threatening apnea or bradycardia occurs so that the infant can be resuscitated before a full cardiorespiratory arrest ensues.

Finally, safety considerations require ventilatory monitoring when certain categories of biomedical life support devices are used. In patients with pulmonary disease, a

Table 1. Levels of Ventilatory Monitoring

Visual observation
Intermittent objective measurements
Continuous or automated measurements
Computerized analysis with treatment algorithms

commonly encountered example is the use of mechanical ventilators. With mechanical ventilation, a number of ventilatory parameters with appropriate alarms are monitored so as to avoid inadvertent disconnection and dangerously high airway pressures. Another example is patients undergoing general anesthesia. Inspired gas concentrations delivered by anesthesia machines frequently may be monitored to insure that the concentrations of anesthetic gases and O_2 are appropriate (4).

LEVELS OF VENTILATORY MONITORING

As emphasized elsewhere (5), there are several levels of intensity with respect to ventilatory monitoring (Table 1). On the most basic level is visual observation and examination by experienced personnel. Although inherently simple, examination of patients is an attribute of clinical medicine that is irreplaceable. The clinical "gestalt" obtained by an experienced healthcare practitioner from observing the work of breathing and the breathing frequency of a patient in respiratory distress has not yet been duplicated by any automated monitoring device. Furthermore, physical examination of the respiratory system is an "art" that should not be lost on future physicians, physician's assistants, nurses, and respiratory care practitioners. In many patients with pulmonary disease, little monitoring other than periodic observation and examination is required.

In some patients with pulmonary disease, more objective data than can be obtained with visual observation and examination are needed. Therefore, the next level of monitoring intensity is the use of intermittent objective measurements of ventilatory function. Probably the most common ventilatory parameter obtained at this level of monitoring are arterial blood gases. Periodic sampling of arterial blood to determine pO_2, pCO_2, and pH is used in almost all patients with the presence or suspicion of significant ventilatory dysfunction. Arterial blood gases are invaluable in following the clinical course and the effect of ventilatory therapy rendered in these patients. Arterial blood gas tensions currently are considered the "gold standard" against which any other measurement of gas exchange is compared. Abnormalities in pO_2 and pCO_2 now are used as definitions of the degree of respiratory impairment. Other frequently used intermittent monitors of ventilatory function are measurements of the VC, the volume of gas exhaled in the first second during a forced expiration (FEV_1), the peak expiratory flow rate during a forced expiration (PF), and the maximum negative airway pressure generated during inspiration against an occluded airway (maximum inspiratory force, MIF). All of these parameters may be useful in following the clinical progress of patients with certain types of pulmonary disease.

Patients who are critically ill may require the third level of monitoring intensity, which is continuous measurement of various ventilatory parameters over time. In most intensive care units, devices are available to continuously track breathing frequency. Patients receiving mechanical ventilation have continuous monitoring of their airway pressures and exhaled tidal volumes for safety considerations. Recent technological advances now allow for continuous monitoring of arterial and venous oxygen saturations. In the most sophisticated intensive care units, on-line monitoring of exhaled gases from ventilated patients and complex measurements of lung mechanics, such as resistance and compliance, can be performed.

Finally, the highest level of monitoring is the use of computer algorithms to not only monitor various ventilatory parameters, but to also adjust therapy according to the values obtained without human intervention. The technology required for such monitoring is undergoing continuous development, and there is increasing use in patient care applications. This is particularly evident on newer microprocessor ventilators where some ventilator modes allow for automatic adjustments to the level of ventilator support according changes in patient lung mechanics.

PARAMETERS USED IN MONITORING LUNG MECHANICS

Breathing Frequency

The most common measurement of ventilatory function is an assessment of the patient's breathing frequency. Usually, this is performed by visually counting the breathing frequency over a 30–60 s period of time, although biomedical devices also can be used for this purpose. Normal individuals breathe at a frequency of ~12 min. However, this rate is markedly altered in patients with lung disease. Reductions in spontaneous respiratory rate are diagnostic of ventilatory dysfunction and suggest central depression of ventilatory control centers in the brain stem. Examples of clinical situations in which this might occur are narcotic overdose, incomplete recovery from a general anesthetic, and primary central alveolar hypoventilation. Whereas reductions in breathing frequency indicate ventilatory dysfunction, an increase in spontaneous breathing frequency (tachypnea) is more nonspecific since individuals can voluntarily increase their breathing frequency in response to physical or psychological stress, or an increase in metabolic demands. However, if a nonpulmonary basis for an increase in breathing frequency can be excluded, tachypnea generally indicates the presence of lung disease. Common examples are pneumonia, congestive heart failure with pulmonary edema, and asthma.

Lung Volume and Airflow. An assessment of the volume of air inspired and expired with each breath (tidal volume, V_T) is an important ventilatory parameter since adequacy of ventilation is determined not only by the breathing frequency, but also by the V_T. The spontaneous V_T in normal individuals varies according to their size, but in an average adult is ~500 mL. Ventilatory dysfunction, whether from processes such as parenchymal pulmonary disease or a reduction in central respiratory drive, generally results in a

decrease in V_T. In mechanically ventilated patients, exhaled V_T is often measured to monitor for inadvertent disconnection or leaks in the ventilator circuit. Sudden declines in exhaled V_T herald disconnection or large ventilator circuit leaks, and on most ventilators trigger audible alarms.

The VC and FEV_1 are frequently monitored intermittently in ambulatory patients with diseases characterized by airflow limitation such as asthma, asthmatic bronchitis, and emphysema. In these diseases, a reduction in comparison to predicted norms (6) is observed in both these parameters. In addition, the ratio of the FEV_1 to the forced VC (FVC), which is normally >0.7, is decreased. The VC by itself is a useful parameter to follow in patients with acute respiratory failure. In spontaneously breathing patients, one criterion for intubation and mechanical ventilation is a VC <10 mL/kg body weight. Conversely, in mechanically ventilated patients, a VC <10 mL/kg body weight is used as a criterion to wean mechanical ventilatory support. Obviously, many other factors are involved in a decision to remove or supply mechanical ventilation, but VC definitely is one key determinant.

The rate of inspiratory air flow is an important ventilatory parameter during mechanical ventilation. It determines the duration of inspiration, and influences airway pressures and the distribution of ventilation in the lungs. Higher peak inspiratory flow rates result in a shortening of the duration of inspiration, higher peak airway pressures, and less uniform intrapulmonary distribution of inspired gas. However, if inspiratory flow rates are too low, inspiration becomes longer than expiration. This may lead to incomplete exhalation, gas trapping in the lung, and an increase in the functional residual capacity (volume of the lung at end-expiration, FRC). These physiologic effects may have deleterious consequences with respect to a decline in cardiac output and an increase in the incidence of pulmonary barotrauma (i.e., pneumothorax, pneumomediastinum, and subcutaneous emphysema).

Monitoring of expiratory flow rates is useful in following the clinical course of patients with airflow limitation. Measurement of flow rates during the middle of a FVC maneuver is a frequently performed pulmonary function test and will not be discussed further. However, periodic determination of PF is a useful monitoring technique to follow the progress of patients with acute and chronic asthma, and also is used to monitor pulmonary function in outpatient studies of patients with suspected bronchospasm resulting from occupational or environmental exposures. Normal values are <500 L/min, but measurements <120 L/min frequently are observed during acute exacerbations of asthma.

Airway and Esophageal Pressures. In patients receiving mechanical ventilation, perhaps one of the most important ventilatory parameters that can be monitored is airway pressures. With volume constant ventilators, a feedback loop is present so that the peak airway pressure cannot exceed a manually preset level. When such a situation occurs (e.g., a patient coughing while inspiratory flow is being delivered from the ventilator), inspiration is terminated when the preset pressure is reached, and an alarm is activated. In this case, peak airway pressure monitoring acts as a safety feature to prevent overpressurization of the airway and the possibility of barotrauma. In contrast, when pressure limited ventilators are used, peak airway pressure is set as a ventilator control parameter and is one of the factors that determines the V_T. Airway pressures are also important in monitoring for inadvertent disconnection or unrecognized leaks in the ventilator circuit during mechanical ventilation. A sudden marked reduction in airway pressure is indicative of a leak in the ventilator circuit, whereas a reduction in the airway pressure to atmospheric pressure would signal a disconnection. On most ventilators, auditory alarms are activated when such events occur. Inspection of the airway pressure curve during mechanical ventilation also can yield important information. As examples, a marked downward concavity at the onset of inspiration is a sign of an insufficient inspiratory flow rate to meet patient demand. During expiration, the failure of the airway pressure curve to return to baseline with an end-expiratory occlusion is indicative of auto-PEEP or air trapping. Finally, airway pressure measurements are an important component of derived ventilatory parameters such as airway resistance and lung compliance.

Monitoring of the maximum inspiratory force (MIF) in patients who are receiving mechanical ventilation is an important ventilatory parameter in determining whether they can be weaned from mechanical ventilation. The MIF can be considered an indicator of the underlying strength of the inspiratory muscles. Values more negative than -25 cm H_2O usually indicate that a patient will be able to maintain adequate ventilation without mechanical ventilatory support, although other factors obviously must also be taken into account.

Measurement of esophageal pressure is used as a reflection of intrapleural pressure (7). It is therefore a component of calculated ventilatory parameters such as lung compliance and the work of breathing. Esophageal pressure catheters also are used to quantify ventilatory effort during sleep studies (polysomnography). However, although there are commercially available devices that incorporate esophageal pressure monitoring to measure lung mechanics, they have not found common acceptance, because placement of an esophageal catheter is invasive and technical factors make accurate determinations difficult (7).

Resistance, Compliance, Volume–Pressure Curves. The total pulmonary resistance to the flow of gas into the lungs is comprised of two factors: (1) friction between the molecules of flowing gas and the airways (airways resistance, R_{aw}), and (2) resistance of the tissues (lungs, rib cage, diaphragm, abdominal contents) as a result of their own displacement (tissue resistance) (8). Elevations of airways resistance are observed in patients with diseases characterized by airflow limitation such as asthma and emphysema. In addition, superimposition of external devices such as narrow endotracheal or tracheostomy tubes can increase airways resistance. Tissue resistance can be increased in patients with such disorders as kyphoscoliosis or pulmonary fibrosis. Total pulmonary resistance is not a frequently used clinical measurement since it requires the simultaneous measurement of lung volume, airway, and esophageal (as a reflection of pleural) pressures (8). Similarly,

airways resistance is also not performed on most patients since it is necessary to use a body plethysmograph (8). No direct measurement of tissue resistance is available, but it is generally calculated as the difference between total pulmonary resistance and airways resistance. Normally, the pulmonary tissue resistance represents ~20% of the total pulmonary resistance. In spite of the difficulty in obtaining a precise measurement of pulmonary resistance in mechanically ventilated patients, an estimate of the pressure needed to overcome airways resistance can be obtained by measuring the difference between the peak airway pressure (p_{max}) and the airway pressure observed after a 1 s inspiratory hold (p_{st}). If airflow measurements (V) at the point of peak airway pressure are available, the equation:

$$R_{aw} = (p_{max} - p_{st})/(V)$$

is an estimate of airways resistance. In general, however, estimates of resistance have not been shown to provide important information relating to the care of patients.

Measurement of compliance or elasticity of the lung is of more importance to the care of patients than resistance. The compliance of a biological system can be defined as the change in volume occurring in response to a change in distending pressure. In addition, compliance measurements can be obtained during static conditions (no flow), or during dynamic conditions (continuous flow). With respect to the respiratory system, static respiratory system compliance (C_{st}) is defined as the change in lung volume occurring with a change in airway pressure over the tidal volume range. Similarly lung compliance (C_L) is defined as the change in lung volume with respect to the change in pleural pressure over the tidal volume range. The relationship between C_{st} and C_L is defined by the following equation:

$$\frac{1}{C_{st}} = \frac{1}{C_L} + \frac{1}{C_t}$$

where C_t is the compliance of the thorax.

Dynamic lung compliance (C_{dyn}) is the change in lung volume occurring with a change in pleural pressure between two points of instantaneous zero flow during breathing (9).

In the respiratory system, changes in compliance are a function of two factors: the elasticity of the tissues and lung volumes. In the upright position, the normal value for C_{st} in adults is ~100 mL/cm H_2O, and the normal value for C_L is 2.0 L/kPa. However, many types of lung pathology such as pulmonary edema and pneumonia result in a reduction of both C_{st} and C_L. In addition, normal values decline with decreasing lung volume. Thus, the normal C_{st} of a small animal or infant is less than an adult human. Dynamic lung compliance and C_L are virtually identical in the absence of significant parenchymal lung disease. However, with airways disease, all alveoli do not ventilate evenly during dynamic conditions, with some gas exchanging areas ventilating slower than others. This phenomenon also has been described as time constant disparity. Because of this disparity in time constants, these slow ventilating areas do not participate fully in gas exchange, and in effect a smaller lung volume is ventilated with each breath. In contrast, during static conditions, a disparity in time constants does not influence the distribution of ventilation. Therefore, C_{dyn} is $<C_L$ in the presence of significant airways disease. Another measure of lung elasticity called the "dynamic characteristic" (C_{dch}) (10) also has been described. It is calculated in patients receiving mechanical ventilation as V_T divided by the difference between P_{max} and PEEP. However, C_{dch} is not just a reflection of lung compliance, but also is influenced by changes in airways resistance.

Measurements of both C_{dyn} and C_L require an estimate of intrapleural pressure that, in most cases, is obtained using an esophageal balloon. Because accurate esophageal pressures are difficult to obtain, especially in critically ill patients, clinical measurement of these ventilatory parameters are uncommon. However, C_{st} only requires measurement of V_T and airway pressure, and thus is easily obtained in most patients receiving mechanical ventilation. Since, in most patients, changes in C_{st} are primarily a result of changes in C_L, serial measurements of C_{st} frequently are used to follow changes in lung elasticity. For example, maximizing C_{st} during application of PEEP has been shown to correspond to the best levels of oxygen transport (product of the cardiac output and the arterial oxygen content) (3). Serial measurements of both C_{st} and C_{dch} are useful in many patients with acute respiratory failure who are receiving mechanical ventilation. A change in both these parameters usually signifies an alteration in the properties of the lung parenchyma such as increasing pulmonary edema or pneumonia. However, a change in C_{dch} without a change in C_{st} is indicative of an increase in airways resistance such as might occur with the onset of bronchospasm or mucous plugging of the airways (10).

Measurements of compliance and resistance summarize pressure–volume relationships in the lung. These can be depicted graphically with two-dimensional (2D) plots of volume versus pressure. A decrease in compliance will be represented by a decrease in slope of this relationship, whereas bowing of the volume–pressure relationship is observed when airways resistance is increased. Analysis of the inspiratory volume–pressure curve may aid in determining the optimum level of PEEP for patients with the acute respiratory distress syndrome who require mechanical ventilation. The optimum level of PEEP is suggested by the inspiratory airway pressure at which the slope of the volume–pressure curve appears to increase (11). In some intensive care units, on-line plots of volume versus pressure can be obtained, but their value in the care of patients has not been proved.

Work of Breathing. In physics, linear work is defined as the product of force times distance. The analogous situation when applied to the movement of air into the lungs is described by the product of pressure times volume. Therefore, the work of breathing is equivalent to the cumulative product of the transpulmonary pressure and the volume of air moved at each instant in time (12). Work of breathing measurements have been proposed as a method of determining whether mechanically ventilated patients are able to breathe spontaneously (13). Unfortunately, accurate determinations require simultaneous measurements of

pleural pressure (usually estimated from esophageal pressure), and airflow. Therefore, work of breathing measurements have not found great utility in the clinical management of patients.

Other Parameters. The rapid shallow breathing index (RSBI) is frequently used to assess the ability of a patient to breath without ventilator support (14). It is calculated by measuring the breathing frequency and dividing it by the V_T in liters. High values (>105) indicative of low V_T and a high breathing frequency suggest that a patient cannot be removed from the mechanical ventilator. However, there has been some controversy concerning the predictive value of this parameter in all clinical situations.

Ventilatory drive is another parameter that can be monitored. Although measurement of the minute ventilation in response to hypoxia or hypercapnia are indexes of central drive in the absence of impairment of neuromuscular or mechanical pulmonary disease, measurement of the airway pressure generated within the first 100 ms after airway occlusion ($p_{0.1}$) is generally recognized as the best indicator of the central drive (15). Unfortunately, this is difficult to accomplish without specialized equipment.

PARAMETERS USED IN THE ASSESSMENT OF GAS EXCHANGE

Oxygenation

The amount of oxygen contained in arterial blood (oxygen content, CaO_2) is determined by the arterial oxygen tension (PaO_2), the percent saturation of hemoglobin (Hgb) with oxygen (S_aO_2), and the Hgb concentration in arterial blood:

$$CaO_2 = 1.34 \times (\text{Hgb}) \times (SaO_2) + 0.003 \times (PaO_2)$$

where CaO_2 = mL O_2/100 mL blood, 1.34 = mL O_2/g Hgb, Hgb = g/100 mL blood, SaO_2 = % saturation, 0.003 = mL O_2 dissolved in blood/mmHg, PaO_2 = mmHg.

In most circumstances, adequacy of oxygenation is monitored by measurement of the PaO_2 or the SaO_2. Although the PaO_2 and SaO_2 are intimately related to each other, their relationship is not linear (Fig. 1), but rather is described by a sigmoid curve. At levels of oxygen tension greater than ~60 mmHg, there is little increase in SaO_2 with increasing values of PaO_2. In contrast, when the PaO_2 falls <60 mmHg, SaO_2 declines rapidly with decreases in PaO_2. Therefore, monitoring of SaO_2 as an estimate of PaO_2 is of little value when the PaO_2 is >70–80 mmHg, since significant alterations in PaO_2 will not be reflected by corresponding changes in SaO_2. At sea level, the normal PaO_2 is ~100 mmHg and the SaO_2 is 97%. However, normal values may vary according to alterations in barometric pressure and increasing age (16). Occasionally, arterial oxygen content (normal value = 9.5 mL O_2/100 mL blood) is monitored as an indicator of oxygenation. Because most of the oxygen in blood is carried in combination with Hgb, this may be of value when Hgb levels are severely abnormal or fluctuating rapidly.

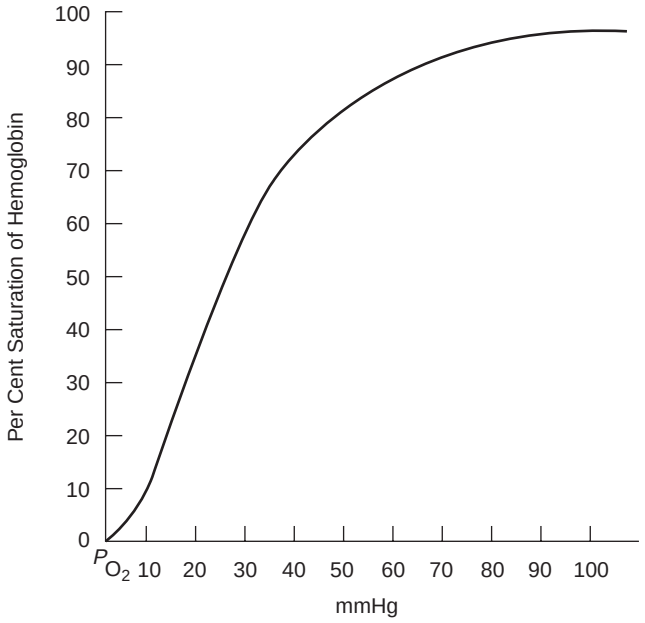


Figure 1. Oxyhemoglobin dissociation curve. The graph shows the relationship between hemoglobin saturation and oxygen tension. Above 60 mmHg, the curve becomes progressively flatter so that little additional saturation occurs when the PO_2 is <60–70 mmHg.

Mixed venous oxygen tension ($P_{\bar{v}}O_2$) or saturation ($S_{\bar{v}}O_2$) generally is regarded as an indicator of cardiac output or metabolic consumption of oxygen in the tissues. However, during continuous monitoring of $S_{\bar{v}}O_2$ in critically ill patients, an abrupt fall in $S_{\bar{v}}O_2$ may herald a sudden deterioration in arterial oxygenation.

There are several parameters derived from measurements of P_aO_2, S_aO_2, $P_{\bar{v}}O_2$, and $S_{\bar{v}}O_2$that are occasionally used in monitoring the status of oxygen gas exchange. Higher inspired oxygen concentrations (F_IO_2) result in higher P_aO_2 values. In order to compare PaO_2 values obtained at differing F_IO_2 values, the P_aO_2/F_IO_2 ratio or its reciprocal, the F_IO_2/P_aO_2 ratio, can be computed. It is assumed, however, that these ratios remain constant with changing F_IO_2 in the presence of lung disease, and that the P_aCO_2 is constant. Since changes in P_aCO_2 result in approximately reciprocal changes in P_aO_2, the status of oxygen gas exchange is sometimes determined by calculating the alveolar–arterial oxygen tension difference ($AaDO_2$; A-a gradient), which controls for changes in oxygenation resulting from changes in the $P_\alpha CO_2$. This is performed by calculating the alveolar oxygen tension (p_AO_2) using the alveolar air equation:

$$P_AO_2 = (P_B - PH_2O) \times F_IO_2 - P_ACO_2/R + [P_ACO_2 \times F_IO_2 \times (1-R)/R]$$

where P_B is barometric pressure, PH_2O is water vapor pressure, P_ACO_2 is alveolar PCO_2 tension, and R is the respiratory gas exchange ratio. It is generally assumed that the $P_ACO_2 = P_aCO_2$ and that R is ~0.8. After the P_AO_2 is computed using the alveolar air equation, the $AaDO_2$ is

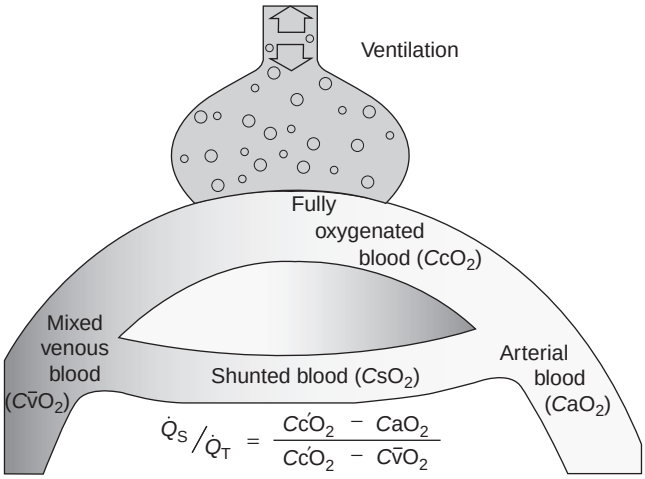


Figure 2. Two-compartment lung model and shunt equation. Conceptually, mixed venous blood either perfuses normal lung and is completely oxygenated or is shunted past the lungs and receives no oxygen. Both oxygenated blood and shunt blood combine to form arterial blood.

the difference between the P_AO_2 and the PaO_2. AaDO_2 changes with age, FIO_2, and position. In a young, healthy person breathing air at sea level, this value is usually 10 mmHg, and 100 mmHg on 100% FIO_2.

The principal disadvantage to the aforementioned indicators of oxygen gas exchange is that they all fail to account for changes in cardiac output, confounding the interpretation of changes in these indicators in the presence of lung disease. In general, impairments in oxygen gas exchange can be quantified using a two-compartment model. As shown in Fig. 2, mixed venous blood can be thought of as either becoming fully oxygenated by the lungs or "shunted" through the lungs without receiving any oxygen. The percentage of the total cardiac output ($\dot{Q}_t$) that is "shunted" ($\dot{Q}_s$) is termed the shunt fraction ($\dot{Q}_s/\dot{Q}_t$). Since changes in $\dot{Q}_t$ may result in changes in the oxygen content of mixed venous blood ($C_{\bar{v}}O_2$), arterial oxygenation may be altered without any change in $\dot{Q}_s/\dot{Q}_t$ (Fig. 3). Therefore, a change in arterial oxygenation may be misconstrued as resulting from an alteration in gas exchange in the lungs when a change in cardiac output is actually responsible. To circumvent this problem, $\dot{Q}_s/\dot{Q}_t$ occasionally is calculated:

$$\dot{Q}_s/\dot{Q}_t = C'_cO_2 - C_aO_2/C'_cO_2 - C_vO_2$$

where C'_cO_2 is the pulmonary capillary oxygen content. Usually C'_cO_2 is calculated by assuming that the hemoglobin in pulmonary capillary blood is fully saturated with oxygen and that the $p_AO_2$reflects the amount of dissolved oxygen in pulmonary capillary blood. Unfortunately, computation of $\dot{Q}_s/\dot{Q}_t$ requires the presence of a pulmonary artery catheter for sampling of mixed venous blood, which makes it unfeasible as a ventilatory parameter in many patients. In the vast majority of patients, however, measurement of the p_aO_2 and, in some situations, the S_aO_2, will provide satisfactory indicators of oxygen gas exchange. Calculation of the aforementioned derived parameters should be reserved for selected critically ill patients in whom such information is necessary for their care.

Ventilation. Adequacy of ventilation is defined by the levels of arterial pCO_2 (p_aCO_2) and pH. These parameters are usually obtained from arterial blood gas samples, and are normally 40 and 7.40 mmHg, respectively. Although hypoventilation is defined as a p_aCO_2 significantly >40 mmHg and hyperventilation is present when the p_aCO_2 is <40 mmHg, the appropriateness of the lung's ventilatory response is determined by the arterial pH. For example, in the presence of a metabolic acidosis, hyperventilation is appropriate if the pH approaches 7.40. However, the hyperventilatory response would be considered inadequate if the pH remained significantly <7.40.

Precise determination of the p_aCO_2 and arterial pH can only be obtained by sampling of arterial blood. Venous pCO_2 and pH occasionally can be used as indirect indicators of arterial values, but this is not recommended since these parameters also reflect cellular metabolism. Of more use as a reflection of p_aCO_2 are measurements of end tidal pCO_2 ($p_{et}CO_2$). In normal individuals, the composition of the gas at the end of a normal exhalation primarily represents gas originating from alveolar areas of the lung. Therefore, the pCO_2 of the gas at the end of a normal tidal breath should approximate the alveolar pCO_2 (p_ACO_2). Since the alveolar–arterial pCO_2 gradient across the alveolar capillary membrane is only 1–2 mmHg, $p_{et}CO_2$ is a good estimate of the p_aCO_2 in normal individuals. In the presence of lung disease, however, the distribution of

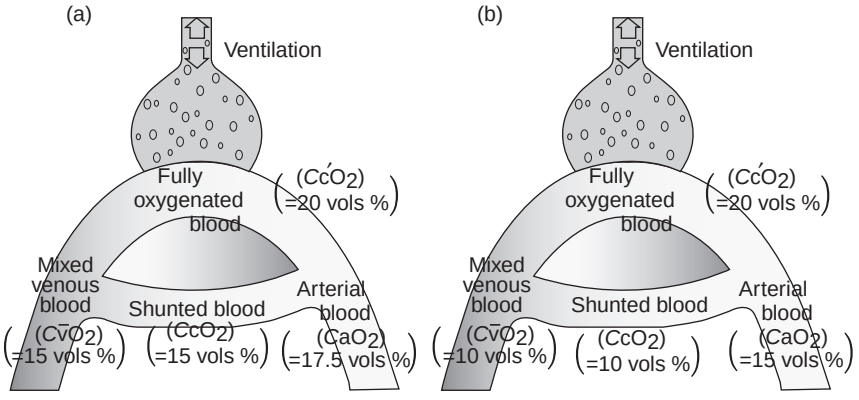


Figure 3. Effect of changes in mixed venous oxygen content on arterial oxygen content in a two-compartment lung model. (1) Assuming that shunted blood flow equals 50% of the cardiac output, mixture of shunted blood with an oxygen content equal to 15 vol% with oxygenated pulmonary capillary blood equal to 20 vol% yields arterial blood with an oxygen content of 17.5 vol%. (b) If the shunt fraction remains 50% and the mixed venous oxygen content decreases 10 vol%, the arterial oxygen content decreases to 15 vol%.

ventilation becomes more uneven. In such situations, the gas composition at the end of an exhalation is composed not only of alveolar gas, but also of gas from poorly perfused or nonperfused areas of the lung (dead space). Therefore, in this situation the $p_{et}CO_2$ will be lower than the p_aCO_2. However, even in situations where lung disease has caused the $p_{et}CO_2$ to be lower than the p_aCO_2, observation of trends in $p_{et}CO_2$ may be useful for suggesting similar changes in p_aCO_2. For example, a slow increase in $p_{et}CO_2$ in a patient with an acute pulmonary illness who is breathing spontaneously would suggest that the patient was beginning to hypoventilate. In addition, an acute increase in the $paCO_2$–$p_{et}CO_2$ difference is indicative of an increase in dead space ventilation such as what would be observed with a pulmonary embolism. However, the most common use of measuring the $p_{et}CO_2$ is to assist in the recognition of an esophageal intubation. In contrast to intubation of the tracheal, inadvertent esophageal intubations are associated with negligible amounts of CO_2.

Estimates of the efficiency of CO_2 gas exchange can be obtained by measurements of expired minute ventilation (V_E) and dead space ventilation to tidal volume ratio (V_D/V_T). In individuals without lung disease, a normal p_aCO_2 can be maintained with a minute ventilation of ~5–6 L/min. Situations where a higher minute ventilation is required to maintain a normal p_aCO_2, suggests either that there is increased ventilation to nonperfused areas of the lung (dead space) or that the production of CO_2 (VCO_2) by the body has increased. The latter can often be excluded on clinical grounds (i.e., absence of fever, sepsis, hyperthyroidism), but both the V_D/V_T and the VCO_2 can be measured. The V_D/V_T can be measured using the Enghoff modification of the Bohr equation:

$$V_D/V_T = p_aCO_2 - p_ECO_2/p_aCO_2$$

where p_ECO_2 represents the pCO_2 in a sample of mixed expired gas. The VCO_2 is determined by measuring the fraction of CO_2 contained in a timed collection of expired gas. Monitoring of V_E, V_D/V_T, and VCO_2 occasionally are helpful in determining the efficiency of CO_2 gas exchange in patients where the etiology of their high minute ventilation requirements is unclear.

METHODS AND DEVICES USED IN VENTILATORY MONITORING

Measurement of Breathing Frequency, Airflow, and Lung Volumes

Aside from visual observation, there are several automated methods of continuously monitoring breathing frequency and effort (Table 2). Since patients in the intensive care unit are universally monitored for cardiac rhythm, one of the most conveniently employed techniques is impedance pneumography. This method is based on the principle that electrical impedance across the chest wall varies with inspiration and expiration (17). Chest wall impedance is measured by passing a constant, sinusoidal, low intensity current between two electrodes affixed to the chest wall. Usually, impedance devices are incorporated into electrocardiogram electrodes, thus permitting the simultaneous monitoring of cardiac rhythm and respiratory effort. Since the relationship between changes in impedance and lung volume are linear, it is possible to obtain measurements of lung volume. However, because the relationship between impedance and volume varies between patients (2.8–4.6 Ω/L), a different calibration must be obtained for each patient. While accurate measurement of lung volumes has been obtained by some investigators (18), others have observed poor reproducibility (17), which may be a result of variability induced by the changing contribution of intercostal and diaphragmatic contraction with alterations in body position. Therefore, most impedance devices are used only qualitatively to assess the frequency and depth of respiratory effort. In addition to their use in an intensive care setting, impedance pneumograms are used as monitors for sleep apnea studies (polysomnograms), and sudden infant death syndrome diagnosis and monitoring (pneumocardiograms and apnea monitors). Since impedance pneumography only measures respiratory effort, however, a major drawback is its inability to distinguish between normal respiratory effort and respiratory effort associated with airway obstruction.

Table 2. Measurement of Breathing Frequency, Airflow and Lung Volumes

Breathing Frequency or Effort
Impedance pneumography
Pleural pressure measurements
Strain gages
Magnetometers
Intercostal and diaphragmatic electromyography
Inductance pneumography
Airflow
Thermistor flowmeters
Nasal pressure measurements
Fleish pneumotachograph
Variable orifice pneumotachograph
Ultrasonic flowmeters
Gas ionization
Oscillating vane spirometer
Mechanical spirometers
Infrared CO_2 absorption
Mass spectrometry

A quantitative indicator of respiratory effort can be obtained from measurements of pleural pressure. Usually, this is performed indirectly by measurement of esophageal pressure with an esophageal balloon and a pressure transducer. However, esophageal pressure measurements also have been obtained using fluid filled catheters (19). With the use of a properly sized and inflated esophageal balloon or fluid catheter placed in the lower third of the esophagus and with the patient in the sitting or lateral decubitus position, esophageal pressures are accurate reflections of pleural pressure. Measurements recorded in the supine position are inaccurate because of artifacts produced by the weight of the mediastinum. Although a commercial esophageal balloon incorporated into a standard nasogastric tube is available (20), recording of accurate pressures is technically difficult in a research environment (7), and even more so in a clinical setting. Nevertheless, esophageal pressure measurements are continuously recorded as an

index of respiratory effort in some sleep disorders laboratories. Otherwise, however, they are not frequently used clinically for continuous monitoring. Direct measurement of pleural pressure using a small catheter placed into the pleural space has also been reported (21), but is not commonly performed.

Respiratory effort also can be monitored using strain gauges (2) and magnetometers (22) placed around the chest and abdomen. Magnetometers measure the change in magnetic fields, produced by electrical coils placed on opposite sides of the body. If expansion of the lungs can be considered as a volume change with only 2° of freedom, the rib cage being 1° of freedom, and the abdomen the other, then lung volume changes are nearly linearly related to changes in the anteroposterior diameters of the chest wall and abdomen. Although the accuracy of magnetometers is limited for the same reasons mentioned for impedance devices, their reliability may be increased if additional magnetometers are used to measure lateral chest wall movement. However, for clinical purposes, both strain gauges and magnetometers are not usually used for quantification of tidal volume. Rather, these methods are used primarily in sleep disorders laboratories for a qualitative estimate of respiratory effort.

Intercostal EMG measurements can be obtained by placement of surface electrodes in the intercostal spaces in the anterior axillary line, and can be employed as an indicator of respiratory effort (2). However, when used in diagnostic studies for sleep disordered breathing, intercostal EMG activity may not be representative of respiratory effort during rapid eye movement sleep because of the inhibition of skeletal muscle tone during this stage of sleep. Frequency analysis of the intercostals and diaphragm EMG has been used by some investigators to study fatigue of the respiratory muscles (23). Although EMG evidence of respiratory muscle fatigue can be detected before clinical findings of acute respiratory failure, routine application of processed respiratory muscle EMG data to patient care has not occurred.

A more reliable method of quantifying breathing frequency and airflow is inductance pneumography. Similar to magnetometers, this technique considers lung volume to be a result of alterations in two compartments: the rib cage and the abdomen. Therefore, the volume of each tidal breath is theoretically attributable to the independent movement of both these compartments. Hence, an inspired volume will be equal to the sum of the volume changes from both compartments. To measure the volume changes in each compartment, bands of insulated wire coils are placed around the chest and abdomen and held in place with a mesh vest. With appropriate calibration, changes in inductance of the coils are accurate reflections of both alterations in lung volume, and the relative contributions of the chest and abdomen to the volume changes (24). Clinical usage of inductance pneumography has been limited primarily to ventilatory monitoring for polysomnography.

There are several categories of devices available that primarily measure airflow (Table 2). With many of these devices, flow rate can be integrated over time to derive a measurement of V_t or VC. Thermistor flow meters are commonly used in sleep disorders laboratories, and in portable ventilation monitors. These devices are based on the principle that the resistance of a thermistor varies with temperature (25). Using a Wheatstone bridge, changes in resistance can be easily measured. The determinants of heat transfer between the thermistor and a gas are gas density and flow rate, and the temperature difference between the thermistor and the gas. In general, an increase in gas flow cools the thermistor and increases thermistor resistance, whereas a decrease in gas flow has the opposite effect. By calibrating the resistance changes to flow rate, quantitative measurements of flow and volume can be obtained. Unfortunately, the changes in resistance with airflow are nonlinear, and correlation with directly measured airflow is relatively poor. Nevertheless, less complex devices are frequently used in sleep disorders laboratories to qualitatively assess airflow during polysomnography.

Flow rate can be calculated by measuring the pressure difference between two sequential points. The most commonly used device employing this principle is the Fleish pneumotachograph (25). With this device, airflow is directed through a mesh screen having little resistance. The pressure drop across the screen is measured using a sensitive differential pressure transducer. Over the flow range for which the device has been designed, flow rate varies with the pressure changes. Inaccuracies result from condensation of water vapor in the device, but can be prevented by heating the pneumotachograph. The Fleish pneumotachograph is primarily used in research applications although some mobile pulmonary function units measure airflow using this device. The pressure difference across a variable orifice also can be used to measure airflow (25). With this device, flow is directed past an elastic flap that acts as a variable orifice. As flow rate increases, the flap opens larger. Similar to the Fleish pneumotachograph, the pressure drop across the orifice is proportional to airflow.

Ultrasonic principles can be employed to monitor ventilation in several ways (26). One method utilizes the principle of vortex shedding. Devices using this principle direct airflow past precisely sized struts creating vortices within the gas stream. An ultrasonic transducer directs sound waves through the air stream to a receiver. These sound waves are interrupted by the formation of vortices within the gas stream. The rate of vortex formation is proportional to airflow. Another method is based on the transit time of an ultrasonic burst between two crystals placed at an angle to the direction of gas flow. Frequencies between 3 and 10 MHz are transmitted, and the transit time can be related to airflow. Advantages of ultrasonic flowmeters are that no moving parts are present, and that there are no problems with condensation in the system.

Ionization of gases by a radioisotope (americium 241) has been proposed as a method of measuring airflow (27). With this technique, gases are ionized by alpha particles emitted by the radioisotope. The alpha particles displace electrons from some of the gas molecules, leaving them positively charged. A downstream electrode measures the amount of ionized particles reaching it. Since the ionized gas particles quickly neutralize themselves by recombination, the number of ionized gas particles reaching the

downstream collector is proportional to the gas flow rate. Similar to ultrasonic devices, potential advantages of this technique are the absence of moving parts, and the lack of interference from condensation in the system.

Interruption of a light beam by a small flapping vane in the gas stream has been used to measure airflow. With this device gas flow oscillates a flapping vane in the gas stream (25). Oscillation of the vane interrupts a light beam from a photoelectric cell. As air flow increases, the frequency of oscillation of the vane and, consequently, the frequency of interruption of the light beam increase. A disadvantage of this type of device is that high flow rates and condensation can cause malfunction.

Airflow can be measured by a variety of mechanical type devices (25,26). The most common example of this category used at the bedside is the Wright Respirometer. This device, and others that are quite similar, direct gas flow through rotating vanes. The vanes, which spin with the flow of gas, are connected through a series of gears to a calibrated dial. The dials on these devices usually are calibrated to measure V_t, VC, and V, but not flow rate. An exception is the Wright Peak Flow Meter, which also uses the principle of gas hitting a rotating vane. However, this device is calibrated to measure flow rate and has a brake that keeps the needle indicator on the highest flow rate obtained until the device is reset. The main disadvantage of these devices is that undermeasurement of airflow occurs at flow rates <3–4 L/min and that the units can be damaged if flow rates are >300 L/min. Nevertheless, their compact size makes them attractive to use.

An indirect estimate of airflow through the nose can be obtained by measuring changes in nasal presssure using a simple oxygen cannula and a pressure transducer. Changes in nasal pressure correlate well with airflow. In addition, flattening of the peak of the nasal pressure tracing during inspiration is indicative of elevations in upper airway resistance. However, it is susceptible to artifact due to mouth opening that results in signal loss. This technique is increasingly used during polysomnography.

A qualitative estimate of breathing frequency and airflow can be obtained by monitoring expired CO_2. Since inspiratory gas has negligible amounts of CO_2, the expiratory phase of breathing is marked by the appearance of measurable amounts of CO_2 in the exhalate. Periods during the ventilatory cycle where no CO_2 is detected indicate the presence of apnea. Carbon dioxide in respiratory gases is usually monitored with an infrared (IR) absorption system or a mass spectrometer (see below). The most common application of the qualitative monitoring of expired CO_2 is in sleep disorders laboratories where the absence of expired CO_2 is a marker for apneic episodes.

Measurement of Oxygen in Blood

Development of the polarographic pO_2 electrode by Dr. Leland Clark in 1953 revolutionized the diagnosis and treatment of patients with impairments in oxygen gas exchange. Today, measurement of arterial oxygen tension is nearly as common place as determination of the hematocrit and can be performed using a modern blood gas analyzer within a minute. The standard pO_2 electrode consists of a platinum cathode sealed in a glass tip and a silver or silver–silver chloride anode. Electrical contact between the anode and the cathode is made by placing them both in a potassium chloride and hydrogen phosphate buffer solution. The blood sample to be analyzed is separated from the electrodes by a gas permeable membrane (polypropylene, Teflon, or Mylar). A constant polarizing voltage is applied between the anode and cathode. Oxygen in the blood sample diffuses across the membrane and combines at the cathode with water and electrons in the electrolyte to form hydroxide ions. The hydroxide ions then are attracted to the anode where electrons are transferred to form silver oxide and water. The current represented by the transfer of electrons at the anode is proportional to the oxygen tension in blood (28).

Blood samples for measurement of oxygen tension usually are obtained from indwelling intravascular catheters or percutaneous needle puncture of a vessel. However, an estimate of p_aO_2 sometimes can be obtained from capillary samples. Warming the skin to induce vasodilation in skin capillaries will cause the pO_2 in these capillaries to approach arterial levels, so-called arterialization. A small amount of this arterialized blood can be obtained by performing a skin prick in the area that was heated, and collecting the blood with a capillary tube. The sample then can be analyzed in a blood gas analyzer designed to accommodate small samples. In infants, such samples frequently are obtained from the heel ('heel stick blood gases'). However, blood gas tensions obtained in this manner are less accurate than direct arterial sampling, and may be inaccurate in situations where skin perfusion is poor (29).

Although the intermittent sampling of arterial and mixed venous blood for oxygen tension is now considered an essential element in the care of critically ill patients in intensive care units, continuous monitoring may be more desirable in very unstable patients. In the late 1970s, a device incorporating a sampling catheter attached to a portable gas chromatograph was introduced into clinical use (30). With this device, a 0.7 mm external diameter probe was inserted through a standard 18-gauge arterial pressure monitoring catheter. The probe consisted of a chamber made from heparin-bonded gas-permeable rubber that was connected through tubing to the gas chromatograph. Oxygen and carbon dioxide were allowed to diffuse from the blood into the probe, and then were carried by helium in the tubing to the chromatograph for analysis. Both pO_2 and pCO_2, which represented the average values of the previous 3.5 min, were displayed every 4 min. Although reasonably accurate, the device was not a commercial success in part because of the requirement that the probe be inserted into an 18-gauge catheter in comparison to the smaller and more commonly used 20-gauge catheter. Continuous pO_2 monitoring also can be performed using miniature Clark-type electrodes inserted intravascularly. One such device has an external diameter of 0.65 mm, making it suitable for insertion into a peripheral artery in adults (31). In addition, these devices have been used to continuously monitor pO_2 circulating externally in heart–lung oxygenators during cardiac bypass surgery. Although potentially attractive for monitoring patients with a very unstable respiratory status, continuous intravascular pO_2

monitoring has yet to find any substantial clinical acceptance.

A continuous approximation of oxygen tension in arterial blood also can be obtained by measuring transcutaneous oxygen tensions (T_cpO_2). Using a small polarographic oxygen electrode operating on a principle similar to that employed in a standard blood gas analyzer, the tension of oxygen diffusing through the skin can be measured. In infants, when the skin is heated to 40 °C, the T_cpO_2 approximates the p_aO_2, and can be used as an indicator of arterial oxygenation. However, the response time of the electrode is relatively slow (95% response time 50–100 s). In adults, skin thickness is greater, and the amount of O_2 able to diffuse to the skin surface is less. Therefore, the T_cpO_2 in adults is significantly less than the p_aO_2 (32). In addition, T_cpO_2 measurements are affected by skin perfusion. When perfusion of the skin falls, as would occur with any condition producing a decrease in cardiac output, T_cpO_2 becomes perfusion dependent. In this situation, a decrease in cardiac output results in a decrease in T_cpO_2 that may not be related to a change in p_aO_2.

Oxygen tension measured from the conjunctiva using a small polarographic electrode has been shown to reflect the p_aO_2 in the absence of hemodynamic impairment. Conjunctival pO_2 values generally range from 50 to 75% of simultaneously measured p_aO_2 values. However, when blood flow to the eyelid is diminished, conjunctival pO_2 measurements are more accurate indicators of changes in peripheral perfusion than approximations of oxygen gas exchange (33). Although measurements of pO_2 in the conjunctiva and the skin appear to monitor similar physiological changes, there are several advantages of measuring the pO_2 of the conjunctiva instead of the skin. First, since the conjunctiva lacks a keratinized surface and has its capillary bed lying just below its surface, it is not necessary to supply extrinsic heating. Second, the capillaries supplying the conjunctiva are branches of the ophthalmic artery, which in turn is a branch of the internal carotid artery. Therefore, measurements of the conjunctival pO_2 may represent a method of indirectly monitoring cerebral perfusion and oxygenation. Last, variability in values resulting from using different skin sites is not a factor in conjunctival pO_2 measurements.

The oxygen saturation of hemoglobin in blood can be calculated from the pO_2, pH, and pCO_2 of the blood sample and the body temperature using a standard nomogram (34), or directly measured using an oximeter. However, use of calculated values is strongly discouraged since they do not account for such factors as the level of carboxyhemoglobin or shifts in the oxyhemoglobin dissociation curve resulting from changes in 2,3-diphosphoglycerate. The latter is a phosphate compound that can shift the oxyhemoglobin dissociation curve independently of pH, pCO_2, and temperature. Oximeters measure oxygen saturation by applying observations made by both Lavoisier and Priestly in 1774 that the color of blood and atmospheric oxygen were related. Subsequently, techniques were developed that allowed determination of the concentrations of different species of hemoglobin by measuring differences in their absorption or reflection of light to different wavelengths. Current blood oximeters pass at least two wavelengths of light through the blood sample to a photodetector whose output is used to calculate absorbances. The oxygen saturation can then be determined by the relative absorbances of each wavelength. Usually, when measuring the relative amounts of saturated or oxyhemoglobin and unsaturated or deoxyhemoglobin, wavelengths of light are chosen so that, with at least one wavelength, the absorbance difference between the two is maximum, and with the other wavelength, the difference approaches zero.

In addition to performing oximetry directly on blood samples, indwelling fiberoptic probes are now available that permit continuous monitoring of the oxygen saturation in mixed venous and umbilical artery blood. In the case of continuous monitoring of mixed venous oxygen saturation, the fiberoptic probe is incorporated into a flow-directed pulmonary artery catheter. Therefore, pulmonary artery pressures and cardiac outputs can be obtained simultaneously with mixed venous oxygen saturation. When compared to oxygen saturation measurements obtained with intermittent blood sampling, these devices have been shown to measure oxygen saturation quite accurately (35). Although continuous monitoring of arterial saturation with such devices is possible, there has not been a large demand for this clinical application since noninvasive methods of measuring arterial oxygen saturation are available (see below), and because there does not appear to be notable clinical utility for measuring mixed venous oxygen saturations continuously. The principle involved with continuous monitoring of oxygen saturation using fiberoptic probes is reflection oximetry. A fiberoptic bundle transmits light of several different wavelengths down the bundle to the blood where light reflected by hemoglobin is transmitted back to a photodetector through a separate fiberoptic bundle. The amount of light reflected by the different wavelengths then can be converted into hemoglobin saturation.

Oxygen saturation can be continuously monitored noninvasively *in vivo* using principles similar to that employed with blood oximeters. The earliest of these devices (Hewlett Packard Ear Oximeter) measures the oxygen saturation of blood flowing through the ear lobe. The technique involves heating the ear to arterialize the blood flow and measuring the optical transmittance of light passed through the ear at 8 wavelengths in the 650–1050 nm range. Oxygen saturation can be computed using a model based on the Beer–Lambert law (36), which states that the absorbance for any wavelength of light is a function of the layers, concentration, and thickness of absorbers. Using the model and converting absorbance to transmittance allows for calculation of the oxygen saturation. The major clinical disadvantage of this device was its bulkiness, and the requirement that it be placed on an ear lobe. More recently, a variety of pulse oximeters have been introduced into clinical use (37). With pulse oximeters, any pulsating arterial vascular bed such as a fingertip, is placed between a light source transmitting different wavelengths and a detector. Expansion and relaxation of the vascular bed by the arterial pulsations alter the length of the light path. The amount of the light detected therefore varies, and this results in the

production of a plethysmographic waveform. The amplitude of the waveform is a function of the wavelengths of light used, the hemoglobin saturation, and the size of the pulse changes. Since the wavelengths of light used are known, beat-to-beat arterial oxygen saturation can be calculated using the magnitude of the pulse changes and a mathematical model based on the Beer–Lambert law. Current pulse oximeters are quite compact, and can be used for ambulatory and home monitoring. In comparison to ear oximeters, they also provide continuous monitoring of the arterial pulse rate. Their use in clinical medicine has proliferated exponentially. The major disadvantage to both ear and pulse oximeters is the inability to measure arterial saturation in low cardiac output states when tissue perfusion is impaired.

The oxygen content of blood can be calculated from the hemoglobin concentration, the saturation of hemoglobin with oxygen, and the pO_2 in the blood sample, as previously described for the arterial oxygen content. In most situations, calculated oxygen contents are used when determination of oxygen contents is required. However, devices and approaches have been developed to directly measure oxygen content. In one approach that is used commercially, the hemoglobin is lysed to release all of the oxygen in the sample into solution. The amount of oxygen is then analyzed using one of a variety of techniques (see below) More cumbersome and difficult is to measure oxygen content by carbon monoxide displacement, or by volume extraction and manometric measurement. The first technique involves displacement of oxygen in the sample to be analyzed by addition of carbon monoxide, and subsequently measuring the increase in pO_2 in the resultant mixture (38). The second technique, described by Van Slyke and Neill (39), involves extraction of the gas from solution with various reagents over mercury, creating a Torricellian vacuum, and measuring pressure changes with a manometer. Both of these techniques are seldomly used today and are primarily of historical interest.

Measurement of Oxygen in Gas

Several techniques are available to measure the concentration of oxygen in gas:

- Paramagnetic analyzers
- Electrical analyzers
- Galvanic fuel cell
- Polarographic gas analyzers
- Ionized oxygen electrode
- Scholander volumetric technique
- Gas chromatography
- Mass spectrometry

Paramagnetic analyzers use the principle that oxygen alters a magnetic field when introduced into it (40). With one such device, oxygen introduced into the magnetic field causes a glass dumbbell, filled with nitrogen and suspended in the field, to rotate. The amount of rotation of the dumbbell is proportional to the percentage of oxygen concentration. Since the alteration in the magnetic field is a response to the amount of oxygen present in the sample gas, a paramagnetic analyzer more accurately reflects the partial pressure of oxygen rather than the oxygen percentage.

Electrical oxygen analyzers compare the resistance of a wire in a reference gas to that of a wire in contact with the sample gas using an electrical current that is proportional to the number of oxygen molecules in the Wheatstone bridge apparatus (40). As the oxygen concentration in the sample gas increases, the resistance of the wire decreases, allowing more current to flow through the wire in contact with the sample gas in the Wheatstone bridge. This phenomenon occurs because oxygen, having a greater mass than nitrogen, cools the wire and decreases its resistance. These devices are sensitive to the presence of gases, other than oxygen and nitrogen, that have a high mass. For example, significant concentrations of carbon dioxide will result in a falsely high oxygen percentage. Since it is the resistances between a sample gas and a reference gas that are being compared, electrical analyzers more accurately reflect oxygen percentage rather than oxygen partial pressure.

Another commonly used type of oxygen analyzer is the galvanic fuel cell (26,40). With this type of device, a semipermeable membrane separates the gas sample from an hydroxide bath. The bath consists of a lead anode and a gold cathode. Oxygen diffuses across the membrane where electrons from the gold electrode and water in the bath combine with the oxygen to form hydroxyl ions. The hydroxyl ions are then attracted to the anode where another chemical reaction occurs that yields lead oxide, water, and electrons. The released electrons produce an electric current that is proportional to the concentration of oxygen in the sample gas. Fuel cell devices sense the amount of oxygen molecules in a sample gas, and therefore most accurately sense oxygen partial pressure. Their readings are adversely affected by excessive moisture, gas under high pressure such as with PEEP, and altitude.

The polarographic principle used in the pO_2 electrode of a blood gas analyzer also can be applied to measurement of oxygen in a gas (40). In fact, many blood gas analyzers will allow processing of gas samples, thus providing an alternative method of measuring the pO_2 in a small gas sample. Factors adversely affecting polarographic units are the same as those for fuel cell devices. In addition, polarographic units most accurately reflect partial pressure. Although both fuel cell and polarographic-type devices utilize positive and negative electrodes and measure the current generated by an electrochemical reaction with oxygen, a faster response time is observed with polarographic units because their electrodes are polarized with a battery. However, the electrodes with galvanic units may have greater longevity.

Another method of measuring the concentration of oxygen in a gas is the use of an ionized oxygen electrode. With this technique, a thin coating of platinum is layered on both sides of a zirconium oxide ceramic tube that is permeable to oxygen. When the tube is heated to a high temperature, O_2 is ionized to O_2^-. The zirconium oxide tube becomes selectively permeable to these ions, and an electrical potential is

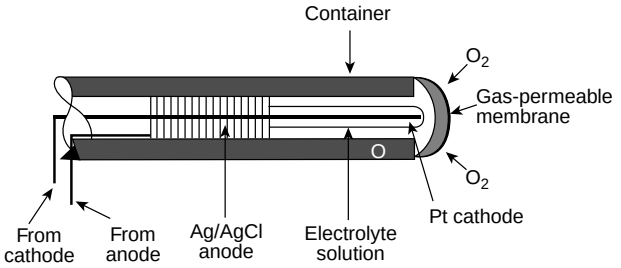


Figure 4. The pO_2 electrode. Sealed platinum cathode and silver–silver chloride anode are bathed in a potassium chloride and hydrogen phosphate electrolyte buffer solution.

generated between the two platinum surfaces that is proportional to the oxygen concentration outside the tube (4).

The most cumbersome method of measuring oxygen concentration in a gas is the Scholander volumetric technique (41). With this technique, volume changes in the sample gas are measured as the various gaseous components are absorbed from the sample. This method is of little clinical utility for continuous monitoring of oxygen concentrations, but is an important reference technique against which more portable units can be compared.

Another, uncommonly used, gas analytical technique in clinical medicine is gas chromatography. With this method, the gas sample to be analyzed is added to the flow of a carrier gas. The carrier gas (usually helium) transports the sample through a column that contains material having a differential affinity for the components of the sample. The column impedes the passage of components with higher affinities, and therefore separation of the components occurs. After passage through the column, the components of the gas sample can be measured using a nonspecific sensor. Usually, this is done by detecting changes in thermal conductivity induced by the components of the gas sample, but other types of sensors also can be employed (42). Gas chromatography is an accurate method for analyzing O_2, CO_2, and anesthetic gas concentrations. However, the slow response time (minutes) is a major drawback. Nevertheless, it has been used for *in vivo* measurements of arterial pO_2 (see the section Measurement of Oxygen in Blood).

The most expensive, but also most versatile technique to analyze oxygen and other gases is mass spectrometry (MS) (4,26). With MS, the sample gas is passed through an electron gun that ionizes the gas particles. The ionized gas particles are then drawn into a magnetic field, where the particles are deflected toward a collection plate. The larger the mass of the gas particle, the farther it will travel down the collection plate. The collection plates count the number of particles of each molecular mass impacting on the plate over time. From this information, the relative composition of the sample gas can be computed. Mass spectrometers are quite accurate, but very expensive. However, one unit can be used to sample gas from multiple sources in a similar way to the method a mainframe computer uses to perform multiple tasks. Such arrangements are used in some operating rooms and intensive care units.

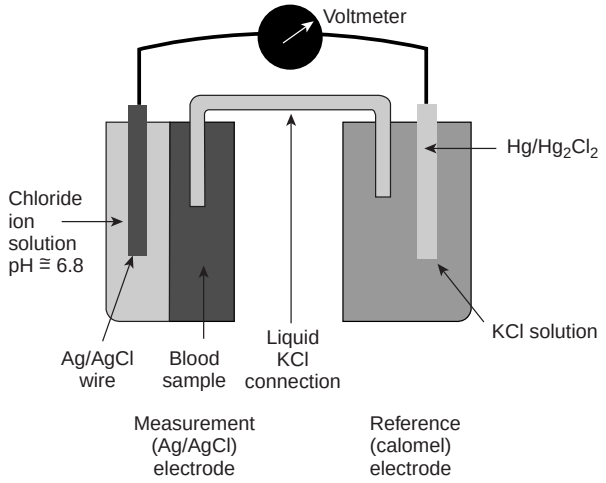


Figure 5. Schematic of the pH electrode. The pH electrode consists of a measurement (silver–silver chloride) electrode and a reference (calomel) electrode.

Measurement of pH and Carbon Dioxide in Blood

Measurement of the pH of blood or another body fluid such as pleural fluid is performed using a pH electrode (Fig. 5). With all pH electrodes, a sample of unknown pH, such as one containing arterial blood, is placed on one side of a pH sensitive glass membrane and a reference pH solution is in contact with the other side. A potential difference develops that is proportional to the pH difference of the two solutions. Since the pH of the reference solution is known, the pH of the sample can be calculated. The actual pH electrode consists of two separate electrodes—a pH-sensitive glass electrode and a reference electrode. Components of the pH sensitive electrode are a silver–silver chloride wire connected to a voltmeter surrounded by a chloride ion buffer with a pH of ~6.8, and a pH-sensitive glass membrane. The reference electrode is constructed so that a platinum wire attached to a voltmeter is in contact with a mercury chloride paste. A 20% potassium chloride solution is used as a wetting agent to insure that the platinum wire maintains contact with the paste. The pH and the reference electrodes are placed in proximity so that the sample is introduced on one side of the pH-sensitive membrane with the buffer of known pH on the other side.

The pCO_2 electrode is actually a modification of the pH electrode. With the pCO_2 electrode, a pH-sensitive glass electrode and a reference electrode are in contact with an electrolyte solution behind a gas-permeable membrane. When a sample is placed on the opposite side of this membrane, CO_2 diffuses across the membrane in both directions and equilibrates with the electrolyte solution. Hydration of the CO_2 results in the production of carbonic acid and a consequent increase in hydrogen ion activity. The pH electrode detects the change in hydrogen ion concentration and develops a voltage change that is proportional to the pCO_2 in the sample.

Whereas reliable methods are available to continuously monitor oxygenation in blood, there are few techniques that are currently used to measure pCO_2.

Intravascular pCO_2 has been measured using a portable gas chromatograph as previously described (30). However, an intravascular sensor using optode technology (see below) to measure pCO_2 and pH, and a miniaturized intravascular Clark pO_2 has been successfully used in humans (43). Devices to measure transcutaneous pCO_2 (T_cpCO_2) are commercially available, and employ a small pCO_2 electrode similar to that used in a standard blood gas analyzer (44). In general, T_cpCO_2 values are higher than corresponding $p_aCO_2$values. Similar to T_cpO_2 measurements, the T_cpCO_2 is adversely affected by changes in skin perfusion, and the response time of the electrode is slow (63% response time: 2 min) (31).

Measurement of Carbon Dioxide in Gas

For monitoring purposes, determination of the concentration of CO_2 in a gas is necessary for calculations of V_D/V_T, VCO_2, and measurement of $p_{et}CO_2$. In addition, it is required in pulmonary function laboratories for performance of CO_2 and O_2 response testing. Mass spectrometry (see the section Measurement of Oxygen in Gas) is a rapid and accurate method of measuring the concentration of CO_2 in a gas, and also is adaptable to measuring gas samples from multiple sources at a central location. However, mass spectrometers are very expensive and not generally portable. The major alternative method of measuring the concentration of CO_2 in a gas is by IR absorption. This technique is based on the principle that all gases whose molecules contain either more than two atoms (e.g., CO_2 or SO_2) or two atoms of different elements (e.g., CO) will absorb IR radiation. With an IR CO_2 analyzer, IR radiation is directed through the gas sample to be analyzed where the sample absorbs a small amount of energy. After passing through the sample, the remainder of the IR radiation is directed into another chamber containing pure CO_2 where the rest of the energy is absorbed. Absorption of IR energy generates heat that can be easily quantified. As the concentration of CO_2 in the gas sample increases, less IR radiation will reach the second chamber that contains pure CO_2. Therefore, the heating effect measured in this chamber will be inversely proportional to the concentration of CO_2 in the gas sample (45). Unfortunately, if a gas sample contains two different gases that can absorb IR energy, the analyzer will be unable to distinguish between the concentrations of the two gases. This problem occasionally arises during monitoring of $p_{et}CO_2$ during nitrous oxide anesthesia. However, placing a chamber filled with nitrous oxide between the IR radiation source and the gas sample to be analyzed will filter out any IR absorption secondary to nitrous oxide.

Practical and Technical Considerations in Blood Gas Analysis

Currently manufactured blood gas analyzers incorporate pO_2, pCO_2, and pH electrodes into a single blood gas analyzer unit so that a blood sample can be analyzed simultaneously for all three parameters. There are several practical considerations in the collection of blood samples for blood gas analysis. First, care should be exercised in the collection of the sample to eliminate air bubbles and not use excessive amounts of anticoagulant. Heparin, a commonly used anticoagulant, is acidic. Excessive amounts therefore will decrease the pH and increase the pCO_2 artifactually. Large air bubbles will produce a decrease in pCO_2 as CO_2 diffuses into the bubble. Oxygen tension will either increase or decrease depending on whether the pO_2 of the sample is higher or lower than the pO_2 of the air bubble. Second, plastic syringes have been observed to perform equally as well as glass syringes provided that the analysis is undertaken within an hour of obtaining the sample. Third, samples should be placed in ice for transport immediately after being obtained to minimize metabolic consumption of oxygen by white blood cells. This may be a cause of spurious hypoxemia in patients with leukemia, who may have pathologically elevated white blood cell counts. Fourth, the patient's temperature should be noted so that correction for the effect of temperature on the relative solubility of oxygen in blood can be made. Several technical factors may introduce error into blood gas analysis. Since oxygen electrodes generally are calibrated using a gas with a pO_2 between 130 and 150 mmHg, measurement of a sample with a higher pO_2 may be in error. In such cases, the electrode should be recalibrated with tonometered blood containing a high pO_2 prior to analysis. In addition, the sample chamber should be flushed with 100% oxygen to leach out any residual wash fluid containing a low oxygen tension. Most blood gas analyzers perform their measurements at a temperature of 37 °C. Variation of the temperature in the sample chamber will alter pH by 0.015 units, pO_2 by 7%, and pCO_2 by 4.5% for each degree of temperature change.

Alarms. Alarms are used during ventilatory monitoring to either signal the occurrence of life-threatening conditions, or indicate the presence of conditions that may not be immediately life-threatening, but are abnormal. Apnea and ventilator disconnection are the most common clinical situations where life-threatening alarms are used. Apnea is detected by monitoring parameters that indicate either absence of ventilatory effort or airflow. However, devices that only detect ventilatory effort such as impedance pneumographs may not be able to distinguish normal respiratory effort from respiratory effort during an obstructive apneic event. Therefore, detection of apnea by monitoring airflow is more sensitive. Ventilator disconnection can be monitored by either a low inspiratory-airway-pressure alarm or a low exhaled-volume alarm. On many current model ventilators, both parameters are monitored. In situations where alarms are used to signal life-threatening events, alarms should be designed such that it is not possible to silence them except for very short periods of time. Otherwise, alarms may be inadvertently left off, and the patient placed at risk.

Inspired oxygen concentration and high airway pressure are two other important parameters that are frequently monitored on a continuous basis. With respect to inspired oxygen concentration, it is important to differentiate between alarms that monitor the gas pressure delivered to a device and alarms that actually sense oxygen concentration. In the first situation, true oxygen concentration is not being measured and in the unlikely situation where an oxygen source becomes contaminated with

another gas, such an alarm would not sound. High pressure alarms on mechanical ventilators not only detect when airway pressure exceeds a certain level, but they also provide a feedback signal to terminate inspiration so that the alarm pressure is not exceeded. Alarms can be used on almost any device used for continuous ventilatory monitoring. However, too many alarms can be ill-advised. If alarms are constantly sounding, all alarms, including life-threatening ones, tend to be ignored.

Evolving Technology in Ventilatory Monitoring. In vivo continuous measurement of O_2, CO_2, pH, and other substances in the future may be clinically useful using optode technology (46,47). An optode is a fiberoptic light bundle coupled to a microcuvette where a chemical reaction can occur. For example, the substance to be analyzed diffuses through a semipermeable membrane into the cuvette whereupon a chemical reaction would be initiated. The chemical reaction would produce a change in the optical properties of the reagent in the cuvette that would be detected by light directed along the fiberoptic light bundle. Wavelength specific absorbance can be used to quantify p_aCO_2 and pH, and fluorescence is useful for p_aO_2, as well as p_aCO_2 and pH. Imprecision of the p_aO_2 sensor has led to a hybrid probe, with a miniaturized Clark pO_2 electrode combined with absorbance p_aCO_2 and pH sensors, as well as a thermocouple for temperature measurement (43). Clinical challenges for continuous on line measurement, utilizing an *in vivo* placement, primarily deal with inaccuracies of p_aO_2 measurement secondary to arterial wall contamination or limited blood flow because of thrombus formation and vasospasm. In addition, probe longevity and cost, as well as pulse pressure dampening over time, have been issues. Because of these problems, an on demand *ex vivo* placement in the arterial line has been tested and has sufficient accuracy, but does not have the advantage of continuous measurement (48).

The ion selective electrode (ISE) or the field effect transistor-based chemical sensor (ISFET) are other new methods of continuously monitoring pH and pCO_2 that may become feasible in the future (47,49). With both these devices, the ion to be analyzed is isolated from the blood using an ion selective membrane. An ISE consists of an electrode surrounded by a buffer solution that is separated from a biological fluid or tissue by an ion-selective membrane. The electrical potential at the measurement electrode is compared to that of a reference electrode, and the potential difference is proportional to the concentration of the ion being measured. The ISFET is an insulated gate field-effect transistor without a metal electrode at the gate. Measurement of the concentration of a specific ion is a function of modulation of current flow within the semiconductor induced by the ion in the surrounding biological tissue or fluid. However, similar to optodes, both the ISE and the ISFET are still in the early developmental stages. Current problems include long-term stability, ambient light and temperature sensitivity, and interference from extraneous ions. With both the ISE and ISFET, control thrombus formation on the probe remains a problem. In addition, an adequate O_2 sensor has not been developed.

Although novel methods of ventilatory monitoring are being developed or are currently under investigation, perhaps the greater challenge will not be to simply prove efficacy. Rather, given the economics of healthcare delivery, the cost-effectiveness of any new device or technique on patient care will need to be unequivocally demonstrated for it to be clinically adopted.

BIBLIOGRAPHY

1. Burrows B, Knudson RJ, Quan SF, Kettel LJ. Respiratory Disorders–A Pathophysiologic Approach. Chicago, IL: Year Book Medical Publication; 1983. p 69–93.
2. Kryger MH. Monitoring Respiratory and Cardiac Function. In: Kryger MH, Roth T, Dement WC, editors. Principles and Practice of Sleep Medicine. Philadelphia: W.B. Saunders; 1989. p 984–993.
3. Suter PM, Fairley HB, Isenberg MD. Optimum end-expiratory airway pressure in patients with acute pulmonary failure. New Engl J Med 1975;242:284–289.
4. Severinghaus JW. Monitoring anesthetic and respiratory gases. In: Blitt CD, editor. Monitoring in Anesthesia and Critical Care Medicine. New York: Churchill-Livingstone; 1985. p 265–290.
5. Fallat RJ. Respiratory monitoring. Clin Chest Med 1982;3:181–194.
6. Hyatt RE, Scanlon PD, Nakamura M. Interpretation of Pulmonary Function Tests. Philadelphia: Lippincott Williams and Wilkins; 1997. p 103–109.
7. Agostoni E. Mechanics of the pleural space. Physiol Rev 1972;52:57–125.
8. Comroe JH, Forster RE, Dubois AB, Briscoe WA, Carlsen E. The Lung. Chicago, IL: Year-Book Medical Publishers; 1962. p 178–190.
9. Murray JF. The Normal Lung. Philadelphia, PA: Saunders; 1976. p 104–105.
10. Bone RC. Thoracic pressure–volume curves in respiratory failure. Crit Care Med 1976;4:148–150.
11. Matamis D, Lemaire F, Rieuf P. Augmentation de la capacité residuelle fonctionnelle induite par la ventilation en pression positive expiratore. Prediction par la courbe pression-volume thoracopulmonaire. Ann Fr Anesth Reanim 1984;3:199–204.
12. Comroe JH, Forster RE, Dubois AB, Briscoe WA, Carlsen E. The Lung. Chicago, IL: Year Book Medical Publications; 1962. p 191–195.
13. Henning RJ, Shubin H, Weil MH. The measurement of the work of breathing for clinical assessment of ventilator dependence. Crit Care Med 1977;5:264–268.
14. Yang KL, Tobin MJ. A prospective study of indexes predicting the outcome of trials of weaning from mechanical ventilation. New Engl J Med 1991;324:1445–1450.
15. Johnson DC, Kazemi H. Central control of ventilation in neuromuscular disease. Clin Chest Med 1994;15:607–617.
16. Burrows B, Knudson RJ, Quan SF, Kettel LJ. Respiratory Disorders–A pathophysiologic Approach. Chicago, IL: Year Book Medical Publishers; 1983. p 61–65.
17. Grenvik A, Ballou S, McGinley E, Millen JE, Cooley WL, Safar P. Impedance pneumography. Chest 1972;62:439–443.
18. Baker LE, Hill DW. The use of electrical impedance techniques for the monitoring of respiratory pattern during anaesthesia. Br J Anaesth 1969;41:2–17.
19. Karason S, Karlsen KL, Lundin S, Stenqvist O.. A simplified method for separate measurements of lung and chest wall

mechanics in ventilator-treated patients. Acta Anaesthesiol Scand 1999;43:308–315.
20. Leatherman NE. An improved balloon system for monitoring intraesophageal pressure in acutely ill patients. Crit Care Med 1978;6:189–192.
21. Downs JB. A technique for direct measurement of intrapleural pressure. Crit Care Med 1976;4:207–210.
22. Mead J, Peterson N, Brinby G, Mead J. Pulmonary ventilation measured from body surface movements. Science 1967;156: 1383–1384.
23. Cohen CA, Zaghelbaum G, Gross D, Roussos CH, Macklem PT. Clinical manifestations of inspiratory muscle fatigue. Am J Med 1982;73:308–316.
24. Cohn M. Respiratory monitoring during sleep: Respiratory inductive plethysmography. In: Guilleminault C, editor. Sleeping and Waking Disorders–Indications and Techniques. Menlo Park, CA: Addison-Wesley; 1982. p 213–224.
25. McPherson SP. Respiratory Therapy Equipment. St. Louis, MO: Mosby; 1981. p 206–214.
26. Sergejev IP. Monitoring of respiratory function during anesthesia. Int Anesthesiol Clin 1981;19:31–59.
27. Jeretin S, Martinez LR, Tang IP, Ito Y. pneumotachography by the ionization principle–a new approach. Anesthesiology 1971;35:218–223.
28. Laver MB, Seifen A. Measurement of blood oxygen tension in anesthesia. Anesthesiology 1965;26:73–101.
29. Siggard-Andersen O. Acid-base and blood gas parameters: Arterial or capillary blood? Scand J Lab Invest 1968;21:289–292.
30. Richman KA, Jobes DR, Schwalb AJ. Continuous in-vivo blood gas determination in man. Anesthesiology 1980;52:313–317.
31. Ledingham IM, MacDonald AM, Douglas IHS. Monitoring of ventilation. In: Shoemaker WC, Thompson WL, editors. Critical Care Medicine–State of the Art. 1981. p E1–E52.
32. Tremper KK, Shoemaker WC. Transcutaneous oxygen monitoring of critically ill adults, with and without low flow shock. Crit Care Med 1981;9:706–709.
33. Isenberg S, Fink S, Shoemaker W. The eye as a peripheral sensor. Ann Ophthalmol 1984;16:1105–1108.
34. Severinghaus JW. Blood gas calculator. J Appl Physiol 1966;21:1108–1116.
35. Baele PL, McMichan JC, Marsh HM, Sill JC, Southorn PA. Continuous monitoring of mixed venous oxygen saturation in critically ill patients. Anesth Analg (Cleveland) 1982;61:513–517.
36. Hill DW. Physics Applied to Anaesthesia. London: Butterworth; 1976. p 339–340.
37. Yelderman M, New W. Evaluation of pulse oximetry. Anesthesiology 1983;59:349–352.
38. Duke GS, Newhouse YMC. Micromethod for measuring blood oxygen content by determining oxygen tension after saturation with carbon monoxide. Am Rev Respir Dis 1974;110:814–816.
39. Van Slyke DD, Neill JM. The determination of bases in blood and other solutions by vacuum extraction and manometric measurement. I. J Biol Chem 1924;61:521–573.
40. McPherson SP. Respiratory Therapy Equipment. St. Louis, MO: Mosby; 1981. p 153–154.
41. Scholander PF. Analyzer for accurate estimation of respiratory gases in one-half cubit centimeter samples. J Biol Chem 1947;167:235–250.
42. Banner N, Olsen RJ. Basic knowledge of gas chromatography and gas chromatographic instruments. In: Ettre LS, Zlatkis A, editors. The Practice of Gas Chromatography. New York: Wiley-Interscience; 1967. p 6–13.
43. Menzel M, Henze D, Soukup J, Engelbrecht K, Senderreck M, Clausen T, Radke J. Experiences with continuous intra-arterial blood gas monitoring. Minerva Anestesiol 2001;67:-325–331.
44. Tremper KK, Shoemaker WC, Shippy CR, Nolan LS. Transcutaneous PCO_2 monitoring on adult patients in the ICU and the operating room. Crit Care Med 1981;9:752–755.
45. Mapleson WW. Physical methods of gas analysis. Br J Anaesth 1962;34:631–636.
46. Peterson JI, Vurek GG. Fiber-optic sensors for biomedical applications. Science 1984;224:123–127.
47. Eberhart RC. Indwelling blood compatible chemical sensors. Surg Clin North Am 1985;65:1025–1040.
48. Mahutte CK. On-line arterial blood gas analysis with optodes: Current status. Clin Biochem 1998;31:119–130.
49. Cheung PW. Chemical sensors. In: Gravenstein JS, Newbower RS, Ream AI, Smith NT, editors. Essential Noninvasive Monitoring in Anesthesia. New York: Grune & Stratton; 1980. p 183–216.

See also MONITORING IN ANESTHESIA; PULMONARY PHYSIOLOGY; RESPIRATORY MECHANICS AND GAS EXCHANGE.

VESSELS, PROPERTIES OF. See ARTERIES, ELASTIC PROPERTIES OF.

VISION CORRECTION DEVICES. See CONTACT LENSES.

VISUAL FIELD TESTING

AMOL D. KULKARNI
University of Wisconsin Madison
Madison, Wisconsin

INTRODUCTION

The visual field is the total area where objects can be seen in the peripheral vision while the eye is focused on a central point. Visual field testing is a critical part of the eye examination and is mandatory for the detailed evaluation of unexplained visual loss. Of the various reasons for conducting a visual field examination, the most common disorder is glaucoma. In glaucoma, field testing is essential not only to establish the diagnosis, but also to follow-up and determine the effectiveness of treatment. There are various modalities available for visual field testing customized to particular eye disorders. This article provides a concise review of the various methods of visual field testing and there applications for clinical and research purposes.

METHODS OF VISUAL FIELDS TESTING

There are various ways of testing the visual field (1). It can be done in a simple manner in the clinic or may involve sophisticated equipments. The commonly used modalities are described below.

1. **Confrontation visual field exam:** It is a quick evaluation of the visual field done by a physician sitting directly in front of you. With one eye covered,

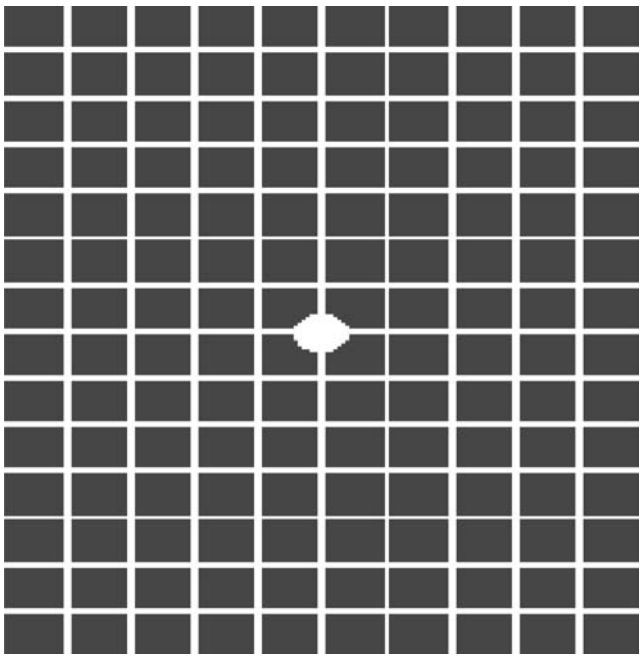


Figure 1. Amsler grid.

the patient is asked to look at the examiner's eye. The patient has to alert the examiner when they can see the examiners hand in the various zones on the nasal, temporal, superior, and inferior aspect while still looking at the examiner's eye. This is a simple method of testing in the clinic and can detect gross visual field loss, such as hemifield loss.

2. **Amsler grid:** It is a quick method of self-assessment of the central 10° of the visual field. It consists of small squares of 1 × 1 mm with a central dot as shown in Fig. 1. After wearing the necessary correction, the Amsler grid is held at 14 in. (35.56 cm) from the eye and the central dot is focused. If any changes in the patterns of adjacent lines or squares are disrupted, a visual defect is suspected. It can detect small central or para-central visual defects called as scotoma especially in disorders of macula.
3. **Tangent screen exam:** The patient is asked to sit 3 ft. (91.44 cm) from a screen with a target in the center. While staring at the target, the patient alerts the examiner on visualizing an object brought into their peripheral vision. This helps in mapping the extent of peripheral vision.
4. **Perimetry:** It is a systematic measurement of the total area where objects can be seen in the peripheral vision while the eye is focused on a central point. The two most commonly used modalities of perimetry are Goldmann kinetic perimetry and threshold static automated perimetry.

Perimetry

There are two types of perimetry based on the stimulus characteristics (1). In kinetic perimetry, a stimulus is moved from a nonseeing area of the visual field to a seeing area along a set meridian. The procedure is repeated with the use of the same stimulus along other meridians spaced at every 15°. In static perimetry, the size and location of the stimulus remain constant. The retinal sensitivity at a specific location is determined by varying the brightness of the stimulus. Based on the measurements from perimetry the normal visual field is defined as the area perceived by a fixating eye extending ~ 60° into the superior field, 75° into the inferior field, 110° temporally, and 60° nasally. This creates an island of vision with a steep central peak that corresponds to the fovea, the area of greatest retinal sensitivity. The shape of the island is defined by conducting measurement at various locations in the field of vision (Fig. 2).

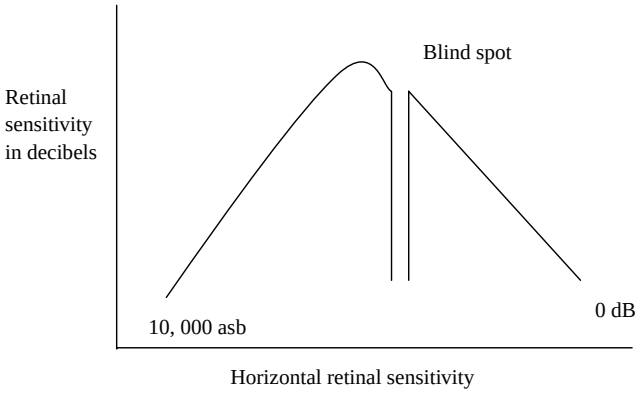


Figure 2. Normal hill of vision.

Goldmann Perimetry

The Goldmann perimeter (1) is the most commonly used instrument for manual perimetry. It is a calibrated bowl projection instrument with a background intensity of 3 1.5 apostilbs (asb). The size and intensity of targets can be varied to plot the visual field kinetically and to determine local static thresholds. The stimuli used to plot a visual field are identified by a Roman numeral, a number, and a letter. The Roman numeral represents the size of the object, from Goldmann size 0 (1/16 mm^2) to Goldmann size V (64 mm^2). Each size increment equals a twofold increase in diameter and a fourfold increase in area. The visual field corresponding to each stimulus size is called as isopter. The Goldmann visual field testing is primarily used for determining field defects caused by disorders of the brain (neuroophthalmology). However, the reliability of field mapping depends on the expertise of the operator (perimetrist).

Automated Perimetry

The introduction of computers and automation lead to the development of a new generation of perimeters (1). These enabled static testing in an objective and standardized fashion with minimal perimetrist bias. A quantitative representation of the visual field can be obtained more easily than with manual testing. The computer allows stimuli to be presented in a pseudorandom, unpredictable fashion.

Patients do not know where the next stimulus will appear, so fixation is improved, thereby increasing the

reliability of the test. Random presentations also increase the speed with which perimetry can be performed, and thereby bypass the problem of local retinal adaptation. Static computerized perimetry measures retinal sensitivity at predetermined locations in the visual field. These perimeters measure the ability of the eye to detect a difference in contrast between a test target and the background luminance. The differential light threshold is designated as the dimmest target seen 50% of the time. Suprathreshold stimuli are brighter than threshold stimuli, and they will be seen $> 50\%$ of the time. Infrathreshold stimuli are dimmer than threshold stimuli, and they will be seen $< 50\%$ of the time. The various perimeters used in clinical practice include Humphrey's visual field analyser, and Octopus perimeter.

Comparison of static and kinetic perimetry (2,3) reveals that kinetic evaluation can clearly outline the normal visual field. However, kinetic perimetry may miss shallow scotomas and poorly define the flat slope seen nasally. Static perimetry readily detects shallow scotomas and can define the slope of both shallow and steep scotomas.

GLAUCOMATOUS FIELD DEFECTS

Perimetry is of paramount significance in the management of glaucoma. In glaucoma, there is damage of the nerve fiber layer that causes loss of visual field. Automated perimetry is widely used and the commonly used programs for glaucoma are the Octopus program 32 and the Humphrey program 30-2. These programs are tests of the central 30° with 6° of separation between locations. The various field defects in glaucoma include diffuse depression, localized nerve fibre bundle defects, paracentral defects, arcuate scotomas, nasal step defects, temporal wedge-shaped defects, early visual field defects, and blind spot changes.

NEWER MODALITIES OF FIELD TESTING

Blue–yellow perimetry, also known as short wavelength automated perimetry (SWAP), represents a recent advance in the early identification of glaucomatous visual field loss. It differs from standard automated perimetry only in that blue light is used as the stimulus, and yellow light is used for the background illumination. Moreover new algorithm, such as SITA, which stands for Swedish Interactive Thresholding Algorithm (4) have been introduced to reduce the length of a visual field test while enhancing sensitivity and specificity.

Frequency doubling technology (FDT) is used to isolate a particular pathway of visual stimulus. It involves use of alternate bars of white and black and helps in early detection of glaucoma.

FUTURE DIRECTIONS

There have been considerable developments in visual field testing in the past decade, and it has opened up areas for the development of new testing and analysis programs. A larger database for patients with glaucoma will improve the accuracy and detection of glaucomatous visual field defects. Additional clinical studies are needed to determine the role of these and other modalities in the future.

BIBLIOGRAPHY

Cited References

1. Johnson CA, Keltner JL. Principles and techniques of the examination of the visual sensory system. In: Newman NJ, Miller NR, editors. Walsh and Hoyt's Clinical neuro-ophthalmology. 5th ed. Baltimore (MD): Williams and Wilkins; 1997; p 194.
2. Trope GE, Britton R. A comparison of Goldmann and Humphrey automated perimetry in patients with glaucoma. Br J Ophthalmol 1987;71:489–493.
3. Tschopp C, et al. Automated static perimetry in the child: methodologic and practical problems. Klin Monatsbl Augenheilkd 1995;206:416–419.
4. Sekhar GC, et al. Sensitivity of Swedish interactive threshold algorithm compared with standard full threshold algorithm in Humphrey visual field testing. Ophthalmology 2000;107: 1303–1308.

VISUAL PROSTHESES

Jean Delbeke
Claude Veraart
Catholic University of Louvain
Brussels, Belgium

INTRODUCTION

Minute electrical stimuli delivered to the retina, the optic nerve, or the occipital cortex can induce light perceptions called phosphenes. The visual prosthesis aims at exploiting these phosphenes to restore a form of vision in some cases of blindness. Very schematically, a camera or a picture capturing device transforms images into electrical signals that are then adapted and passed on to some still functional part of the visual pathways, thus bridging the defective structures. The system has at least some parts implanted, including electrodes and their stimulator circuits. A photosensitive array in the eye could provide the necessary image input, but most approaches use an external miniature camera. Typically, the visual data handling requires an external processor and the power supply as well as the data are provided to the implant by a transcutaneous transmission system.

Despite a first pioneering attempt by Brindley and Lewin as early as 1968 (1) only very few experimental visual prostheses have been implanted in humans so far. The limited accessibility of the involved anatomical structures, the poorly understood neural encoding, and the huge amount of information handled by the visual nervous system have clearly hampered a development that can not yet be compared with the far more advanced evolution of cochlear implants (see article on Cochlear Implants in this encyclopedia). The visual prosthesis is still at a very

early stage, exploring different methodological directions, and seeking minimal performances that would justify clinical applications in the most severe cases of blindness. The first fully fledged clinical study has yet to begin and the experimental character of existing systems limits all trials to a restricted number of well-informed adult volunteers.

Vision Basics and Retinotopy

A basic knowledge of the anatomy and physiology of the visual system is necessary for a proper understanding of the visual prosthesis in its various forms.

The eye can be compared to a camera with an adjustable optics including the cornea and the lens, focusing inverted images of the external world onto the retina. The retina is in fact a thin slice of brain that has expanded into the eye during the embryo development. It is made up of several cell layers. Among these, the photosensitive cells called cones and rods form the external layer with, as a result that light entering the eye has to cross the complete retina before to reach them (see Fig. 1c).

The 120 million or so photosensitive cells are not uniformly distributed. Their density is highest at the fovea, a region that corresponds to the fixation point, near the center of the visual field. Cones outnumber the rods in the central retina and are the only kind of photosensitive cells at the fovea. They discriminate colors. Rods predominate at the periphery. These very sensitive sensors play a major role in night vision, but do not contribute to high resolution nor color perception. The cellular hyperpolarization generated by light impinging on the photoreceptors is carried over to other the neuronal cell layers ultimately connected to output layer represented by the ganglion cells. With the exception of the fovea (most central part) and especially in the periphery of the retina (i.e., of the visual field), there is a great deal of convergence and encoding of the visual signal that must be squeezed from the analog modulation of $\sim$ 130 million photoreceptor potentials into the discharge bursts of little more than 1 million ganglion cells. Roughly, the bipolar cells represent the main converging vertical link between the photoreceptors and the ganglion cells. The horizontal cells at the junction between photoreceptors and bipolar cells and the amacrine cells at the junction between bipolar cells and the ganglion cells provide sideways connections. Functionally, the retinal network can be subdivided in a number of parallel channels each focusing on the transmission of one aspect of images. These include a color coding system and movement detecting circuits. Bipolar cells and ganglion cells also subdivide in ON OFF types. The ON cells increase their activity on exposure of the center of their receptive field to light while their ongoing firing slows down when light strikes in the periphery of that region. The OFF cells react in the opposite way. Ganglion cells also respond more or less

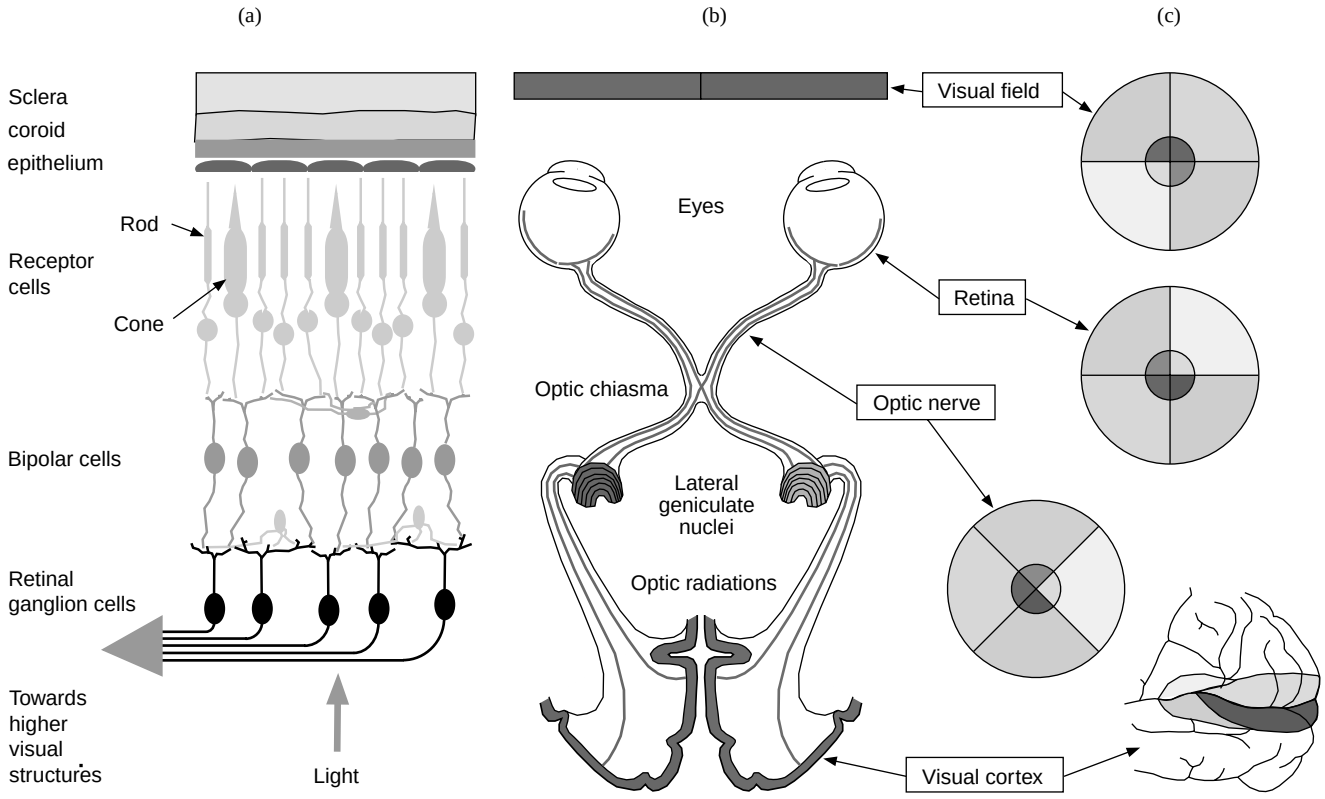


Figure 1. (a) Schematic representation of the visual pathways from the eye to the visual cortex with an indication of the hemifield segregation at the level of the chiasma. (b) Retinotopy mapping all along the visual pathways. Note also the cortical amplification, that is, disproportional cortical representation of the central visual field. (c) Enlarged view of a cut through the retina showing the various layers and their position in relation to light entering the eye.

transiently. The result is a spatiotemporal retinal filter providing the optic nerve with a complex and still not well understood neural code Ganglion cell axons first runs over the inner surface of the retina to joint the optic nerve head located slightly nasally from the center of the retina. There, all ganglion cell axons join before to leave the eye and form the optic nerve.

In the orbit, the optic nerve is relatively slack, laying between extraocular muscles, fat tissue and various blood vessels and nerves. It is protected by a strong sleeve of dura mater as well as a very thin pia mater with cerebrospinal fluid in between. When the optic nerve enters the skull, it looses its dura mater sleeve, which now lines the inner surface of the skull. After a little >1 cm, the two intracranial optic nerves meet and exchange fibers at the level of the chiasma (see Fig. 1a). Ganglion cell axons of the nasal hemiretina or temporal visual field cross the midline and join the temporal axons of the other eye to form the optic tract. Foveal axons split into one branch toward each side. Some fibers (not represented) corresponding to accessory functions leave the mainstream visual pathway to reach the hypothalamus, pretectum and superior colliculus. The axons directly involved in vision end in the lateral geniculate nucleus. This structure performs further signal processing and receives control signals from various parts of the brain. From the lateral geniculate nucleus, the visual information reaches the occipital lobe of the brain through the optic radiation. Interestingly, corresponding inputs from both eyes are arranged in close proximity, but remain segregated up to the level of the primary visual cortex. From there, signals corresponding to various aspects of the visual stimulus are dispatched to different brain locations for further processing.

Consequently, of the fiber exchanges at the level of the chiasma, except for the representation of the fovea, one visual cortex receives only information about the contralateral visual field. In addition, albeit with much distortion, cells of the visual cortex tend to retain the topological relationship of the retinal location from which they receive their input. The resulting point-to-point correspondence with the retinal locations, and hence the visual field is called retinotopy. Some form of retinotopy is found at most levels of the visual pathways up to the cortex (see Fig. 1b). Despite an obvious need for corrective remapping, retinotopy of the structure to be stimulated will be an essential consideration for the development of a visual prosthesis. Right from the level of the retinal ganglion cells on, retinotopy nevertheless remains an approximation that does not take into account other significant aspects of the neural signal encoding.

Blindness

Under ideal conditions, the human minimum angle of resolution (MAR) is $\sim$0.5 arc min (20/10 vision). However, the standard definition of normal visual acuity (20/20 vision) is the ability to resolve a spatial pattern separated by a visual angle of one minute of arc ($\sim$4 μm on the retina). The Snellen visual acuity measures the pattern recognition acuity as the ratio d/D, where D is the distance at which a sign subtends 5 min of arc, and d is the distance at which they can be recognized. Reference patterns subtend 5 min of arc at a distance of 60 m and have features of 1 min of arc, corresponding to the standard normal acuity. A visual acuity of 3/60 means that such a sign can only be recognized at a distance of 3 m. For practical purposes, the Snellen chart is made up of different sized letters such that the examination can be performed at a single distance. Because the normal acuity corresponds to 1 min of arc, the MAR value in arc minutes has the same numeric value as the reciprocal of the Snellen fraction.

However, sight is a multidimensional ability that cannot be measured by acuity alone. The effect of a visual field defect or a reduced sensitivity to light cannot be compared directly to acuity.

The International Statistical Classification of Diseases and Related Health Problems of the World Health Organization (ICD-10) uses codes 1–5 to describe moderate, severe, profound, near total, and total visual impairments, respectively. Within that range, the label low vision (categories 1 and 2) designates a visual acuity $>$6/18 (0.3), but $<$3/60 (0.05) in the better eye with optimal correction or a visual field between 10 and 30°. The label blindness encompasses categories 3–5. Code 3 corresponds to a visual acuity $<$0.05 on the Snellen scale for the best eye using appropriate correction, or a central visual field diameter of $<10^\circ$ in its largest diameter. An acuity of $<$0.02 or a visual field $<5^\circ$ is coded 4 (near total) while total visual impairment means deprived of light perception.

To measure a visual acuity in near total visual impairment, alternative methods are used including close range on the Snellen chart reading, finger counting, the detection of hand motion, or the perception of light.

Blindness can result from any cause potentially affecting the visual pathways: genetic abnormalities, infections, metabolic diseases, trauma, vascular deficiency, or cancer for example. In one subgroup, mainly represented by retinitis pigmentisa (RP), age related macular degeneration, and stargardt's disease, it has been shown that blindness can result from a total destruction of the photosensitive cell layer while a proportion of other cells of the retina and the remaining of the visual pathways survive. The hereditary disease retinitis pigmentosa in particular, is a relatively frequent cause of severe and incurable blindness in developed countries with a prevalence of 1 in 5000 (2).

An essential distinction must also be made between early and late blindness. Indeed, several years of normal behavioral experience are required after birth for the human visual system to fully organize and suitably weight its synaptic connections (3). Those years are referred to as the critical period. The visual brain of people who are born blind is functionally different from that of those who lost sight later in life (4). Early blindness indicates a loss of vision before the end of the critical period. There is no sharp separation between early and late blindness and different visual functions each having their own critical period. Motion detection would probably mature before spatial localization, followed by object recognition, color vision, and stereopsis. According to ophthalmologic observations in diplopia, the critical development period for binocular vision ends around the age of 6. In addition, the visual impairment might result from a progressive

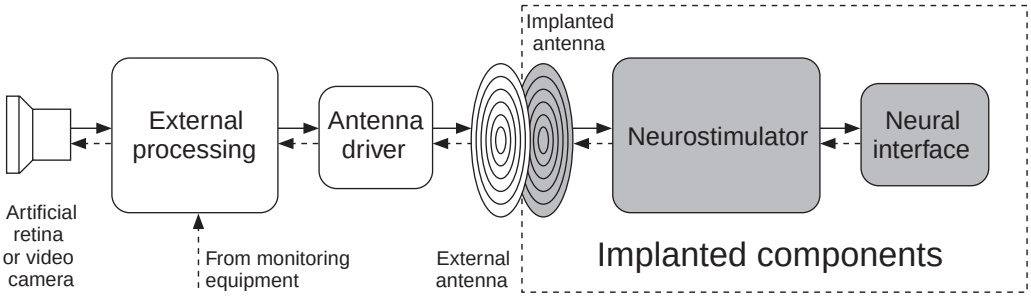


Figure 2. General layout of a visual prosthesis, with the external components on the left and the implanted components (shaded) on the right.

disease. Early blindness type increased metabolic activity has been reported in cases where progressive total blindness occurred around the age of 10 (4). By contrast, late blindness features are observed when the first visual defect appears before the age of 8–12, but only evolve toward total blindness thereafter, or when vision is accidentally lost at the age of 12 or later.

A prosthesis implanted after the critical period is doomed to failure. Age of blindness onset, however, is not the only factor resulting in a deviant visual system. In diseases such as retinitis pigmentosa, the loss of photoreceptors itself results in important remodeling of the remaining retinal network (5), resulting in aberrant neural connections and shielding by scar tissue, all potentially limiting more or less severely the efficiency of a visual prosthesis.

History

The idea of an electrical treatment for blindness is perhaps as old as the discovery of electricity itself. The first real attempt to implant a visual prosthesis, however, dates back from 1968 (1). A set of 80 electrodes were implanted over the occipital pole of the cortex of a blind person. Each electrode could be activated transcutaneously by an equal number of implanted radio receivers. Small precisely located phosphenes were obtained, suggesting that a useful prosthesis could indeed become possible, but the subdural cortex surface electrodes had very high thresholds and did not provide an adequate resolution. While intracortical electrode arrays were being developed using technologies derived from the semiconductor industry (6), it became clear that in terminal retinitis pigmentosa, a significant fraction of the ganglion cell population remains functional despite total blindness (7,8). In such cases, a visual prosthesis interfacing with the surviving layers of the retina or with the optic nerve, all referred to as the anterior type could be as useful as the more complex cortical implant. As a result, starting around 1990, perhaps partially dragged by the success of the cochlear prosthesis, a renewed and still growing interest in visual prosthesis rapidly expanded worldwide.

Basic Theory of Operation of the Visual Prosthesis

In all published visual prosthesis approaches, the visual system is very schematically seen as a transmission chain in which retinal image pixels are encoded into series of electrical pulses ultimately activating the visual cortex. The visual prosthesis is meant to replace or by-pass the defective link in the neural chain. At the cost of not using still functional body parts (eye optics, e.g.), most visual prosthesis designs replace the whole front end of the sensory chain, requiring only the prosthesis output to be connected to the nervous system (see Fig. 2). A picture of the external and implanted components of a prototype of the optic nerve visual prosthesis is given in Fig. 3 as an example.

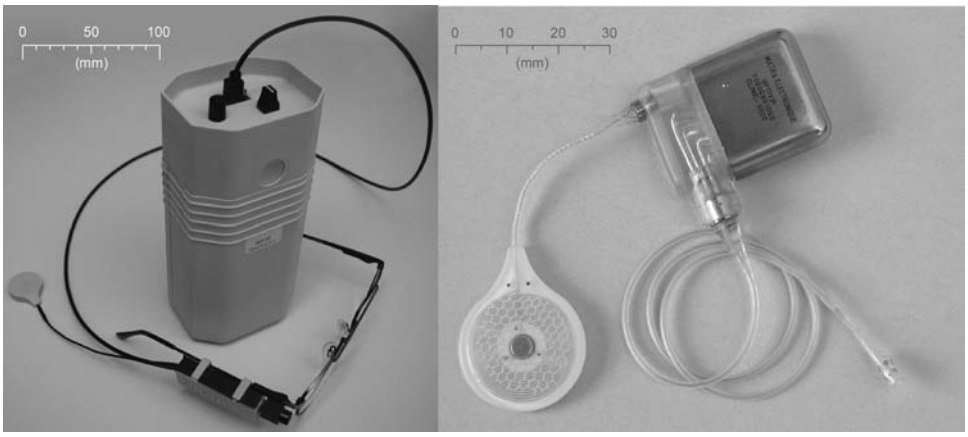


Figure 3. Implementation example: picture of the optic nerve visual prosthesis prototype. On the left, the external processor linked to glasses on which a camera and the external antenna are attached. To the right, the implanted components with the internal antenna, the neurostimulator and the spiral cuff electrode with their leads and connectors.

Depending on the approach, the prosthesis sensors can be any imaging device from an implanted photoarrays (9) to a miniature cameras worn on a pair of glasses. In this last case, a small-size camera is also an asset for esthetical reasons.

A processor is necessary to transform the video images in properly encoded and modulated commands to the implanted stimulator (10). The prosthesis must allow real time object localization and identification. Visual information quickly represents a huge amount of data imposing high demands on the computational performances while the processor including its power supply, must remain easy to carry around. This portability often represents a technological challenge by itself. Adaptability to new and rapidly emerging developments is also an essential requirement in this quite young technology. Transcutaneous electromagnetic transmission of data and power using magnet fixated antennas is now standard for cochlear implants. This principle was already applied in a more primitive way by Brindley back in 1968 (1). He then solved the problems of parallel transmission through the implantation of an array of transmitters. Recent projects use combined power and data electromagnetic telemetry systems similar to those of the cochlear implant systems but with improved power and data performances.

Different Approaches

The neural interface typically represents the most critical and challenging component involved. In keeping with this, the different approaches to develop a visual prosthesis are often classified according to their connection point along the visual pathways: subretinal, epiretinal, transretinal, hybrid, optic nerve, cortex surface, or intracortical. The interface can be chemical or electrical. Until now, the chemical approach has been limited to some work on retinal devices. Many more projects use electrical stimulation through specifically designed electrodes.

Chemical Stimulation. The basic idea of this approach is to use neurotransmitters (L-glutamate, e.g.) or other specific chemicals that are known to activate neural or retinal cells. Directly over the retina, photochemical reactions (typically requiring ultraviolet, UV light) liberate active components from an inactive parent molecule or from 'caging' molecules (C_{60}, fullerene) (11). The result is a direct translation of retinal images into a corresponding neural activation pattern. Alternatively, an electronically activated multichannel microfluidic device could deliver the needed chemicals locally and be used in a visual prosthesis as a substitute to electrodes (12). At present, most research efforts are still devoted to basic problems such as to reduce the required light energy level, biocompatibility, transport of the chemicals and reservoir refill (13). In the future, however, chemical stimulation could have several advantages. There is no electrode corrosion. Stimulation selectivity can bear on the subgroups of retinal cells, and so mimic the physiological activation achieved by synaptic transmission. In addition, the proposed stimulating electrodes can be made on soft materials supposed to be less damaging for the retina than electrode arrays (12).

Subretinal Implants. The most appealing aspect of the subretinal implants is that it exploits the supposedly healthy eye optics and interfaces to the visual pathways before any neural processing has blurred the neural code. The aim is indeed to replace the damaged photosensors by an array of passive miniature photosensitive devices transforming the retinal light image into local electrical stimulating currents. These currents would in turn activate the surviving neuronal circuits of the retina in keeping with the light intensity they receive, much in the same way as photosensitive cells do. The idea is straightforward and logical. Natural accommodation and physiological eye movements would remain functional. Very thin (100 μm) flexible electrode construction and perforations allowing nutrient and other metabolic exchanges between the retina and the choroid could insure biocompatibility (14). The use of glycoprotein coating has been suggested to improve the biocompatibility of the components. Small implants can be quickly and securely trapped between the neural elements and the pigmentary epithelium, which seems to pump out this space (15). The light wavelength sensitivity of the microphotodiodes is in the 500–1100 nm range, which corresponds reasonably well with the visible spectrum of 400–700 nm.

Chow's group developed a 25 μm thick subretinal implant of $\sim$ 5000 subunits on a 2 mm diameter chip, enough to provide a tunnel vision of a little $> 8^\circ$. Such devices have been implanted in a number of retinitis pigmentosa patients. After 6–18 months, no significant side effect has been noticed and some patients reported a transient improvement not related to the implant position in the visual field. All the implants were electrically functional, but no visual response of the implant themselves was demonstrated (9). These devices are only powered by incident light but, as shown by Zrenner's team, currents generated by microphotodiodes are by far too low to activate bipolar cells. Available devices, would require 12 klx (16) to do so while normal ambient light typically reaches $\sim$ 8 lx. An active amplification is therefore necessary, finally sharing with other approaches the need for an external power supply raising again the problems of bulkiness, heat dissipation, energy and data transmission.

Epiretinal Implants. In the epiretinal approach, microelectrode arrays are attached on the vitreal side of the retina, just over the inner layer that contains the ganglion cells and their axons. The stimulation contacts are supposed to activate local ganglion and/or the bipolar cells. Acute stimulation of the retina in human volunteers has gone very far in demonstrating the feasibility of a prosthesis with a resolution of $\sim 1.75^\circ$ (17). Concerns have been raised about the possibility to activate passing-by optic nerve fibers as well, which would result in aberrant perceptions (18). Also, because of the retinal preprocessing, activity in the ganglion cells can no longer be seen as a point-to-point representation of the retinal image and the encoding at that level already exploits the time dimension as well. Humayun's team developed a prosthetic device with an array of 16 electrodes connected to an implanted stimulator located outside the eye. An external system for

image acquisition and processing sends data and power to the implanted electronics by telemetry. This device was implanted in several blind RP patient. Initial results seem encouraging (19).

The basic principle in most visual prosthesis approaches is that stimulation through a small electrode will result in the perception of a point-like phosphene of corresponding retinotopic localization. An array of such electrode contacts would produce a number of phosphenes that can be distinguished by their location in the visual field. After correction for any nonconformal localization, an image perception could then be obtained by activation of the corresponding pixel electrodes. It will be seen that this pixel phosphene method to selectively activate a subset of fibers or ganglion cells does not hold in the case of the optic nerve stimulation.

The Transretinal Approach. A transretinal approach (20) has also been suggested whereby a needle placed in the vitreous is used as a single cathode facing a miniature array of anodes slit under (or in the sclera), thus yielding a transretinal stimulation. Evoked potentials have been obtained in animals using eight contact on a $2 \times 4 \times 0.18$ mm electrode with polyimide substrate. There is no indication yet as to which cells form the primary target.

The Hybrid Approach. The team of Yagi and Tano has started research to grow transplanted neural cells from a subretinal implant to the central nervous system using axon-guiding substrates. This approach could also be applied when ganglion cells are destroyed. This work remains very much preliminary and no results are available yet.

The Optic Nerve Approach. A direct stimulation of the ganglion cell axons with an optic nerve electrode can be seen as an alternative to the epiretinal stimulation. The basic idea here is that the simultaneous activation of a number of contacts can focus the stimulation on a chosen subset of the axon bundle by controlling the applied electric field. As a result, electrode contacts around the nerve can yield a selective activation (21). The number of different fiber subsets that can be stimulated independently and thus the number of phosphene perceptions that can be obtained is much larger than the number of electrode contacts available. This principle could be applied to all electrodes carrying contacts at some distance from the target cells. Focal stimuli are thus generated serially by each multicontact activation instead of in parallel through individual contacts.

This concept has been validated in a human implantation (22). The results confirm a retinotopic organization of the intracranial optic nerve and phosphenes are obtained with safe electric charges. Interestingly, due to the signal encoding in the optic nerve, phosphenes do not reproduce the distribution of the fiber activation and are much smaller than expected. Their position in the visual field can be controlled, which is of course essential in the prospect of the visual prosthesis development, making it possible to convey image information, even without resorting to more complex selective stimulation schemes (e.g., superficial fiber blocking).

It has been suggested (23) that a penetrating electrode could increase the number of available independent responses, but the damage inflicted to the nerve has not yet allowed to validate such an alternative.

The first optic nerve electrode was implanted behind the orbit just in front of the chiasma. A new surgical technique has been developed to implant an eight contact electrode in the orbit. Avoiding intracranial surgery is certainly reassuring for the patient, but the intraorbical approach is technically difficult and has several drawbacks. At that location, the optic nerve is indeed covered by the dura mater, which shields off the fibers from the stimulation and therefore results in higher thresholds. Also, somatic sensory nerve fibers as well as blood vessels are present in the dura mater, and eye movements could limit the stability of the electrode contacts. Nevertheless, the feasibility of this approach has been demonstrated recently.

The Cortical Approach. The very first implanted human visual prosthesis prototype (1) included an array of 80 electrode contacts placed over the occipital cortex of a blind person and linked to an equal number of miniature transmitters placed under the scalp. High thresholds and poor selectivity have led to the conclusion that intracortical rather than cortex surface electrodes were necessary (24). Subsequently, a two-dimensional (2D) device known as the Huntington electrode (25) and a three-dimensionally (3D) structured, single plane, Utah electrode (26) were proposed for intracortical implantation. Resolutions of ~ 400 μm can be obtained (27) where the surface electrodes of Brindley could only resolve minimal distances of 2–3 mm. In an acute experiment, 38 intracortical microelectrodes have been implanted , for a period of 4 months in the right visual cortex of a human volunteer (6). Two-point resolution was about five times better than had typically been achieved with surface stimulation. All phosphenes were located in the left hemi-field with the majority above the horizontal meridian. There was a clustering of most of the phosphenes within a relatively small area of the visual space (6). As opposed to subdural electrodes, intracortical devices have the potential advantage to reach the hidden parts of the cortex in the depth of the calcarine fold corresponding to the big gap in the region of the horizontal meridian of the visual field as observed by Brindley.

Brindley himself stressed the important variability of the cortical maps among individuals. Individual mapping of each cortical implant is thus expected to be necessary. Biocompatibility is still a major point of concern, owing to the large number of electrode contacts and stimulator connections, especially for chronic human implantation. However, compared with the prechiasmatic approaches, the intracortical alternative would in the long run have the advantage of being applicable to many more conditions, not just diseases of the retinal photosensors.

A consortium led by Troyk has undertaken the development of an intracortical prosthesis based on a 256-channel stimulator module and a 1024-contact array. Recently, 152 intracortical microelectrodes have been chronically implanted in area V1 of a male macaque.

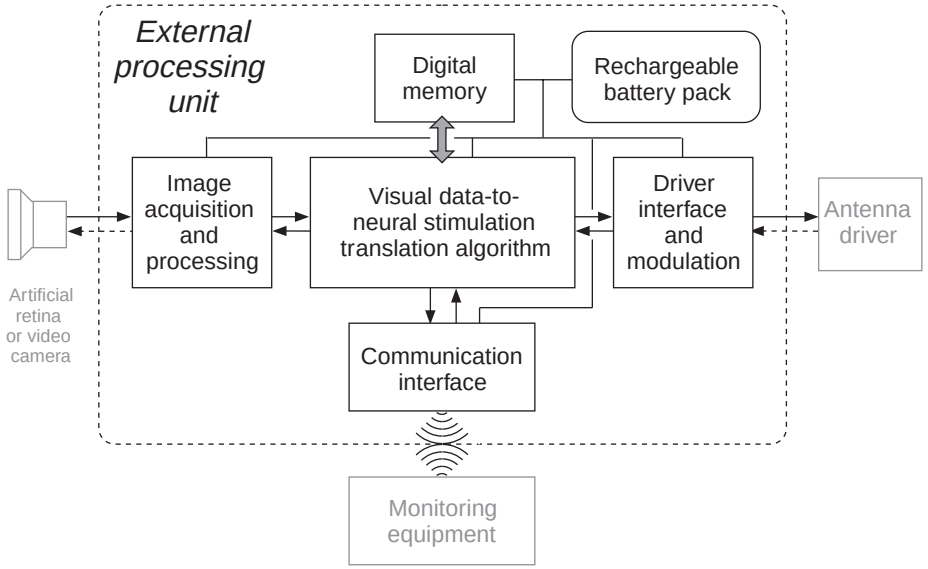


Figure 4. Detailed representation of the external components of Figs. 2 and 3.

Receptive field mapping was done using a memory saccade task (28). It is expected that such new animal psychophysical tests will compensate for the lack of a linguistic report and allow further developments in animals before finally turning to human trials.

ENGINEERING ASPECTS

The Hardware

Typically, the hardware of a visual prosthesis is composed of an external system and an implanted part. The external system (see Fig. 4) includes some image capturing device, an external processor and the external part of a transmission unit. Implanted (see Fig. 5) are the other one-half of the transmission system, stimulators, and an electrode. Depending on the approach, there are important variations on this basic scheme. In subretinal projects, a photosensor array implanted directly in the eye could replace the external camera. Some specifications of the example of the optic nerve prosthesis are given as an example in Table 1.

External Components

The Image Grabber. A miniature video camera typically mounted on glasses provides the image capturing device of the visual prosthesis. A low weight camera with good esthetical appearance is of importance to the blind person and can be improved by miniaturization. It could further be stated that the most trivial imaging devices largely outperforms the needs of all present day visual prostheses (29). Nevertheless, a good image quality including some

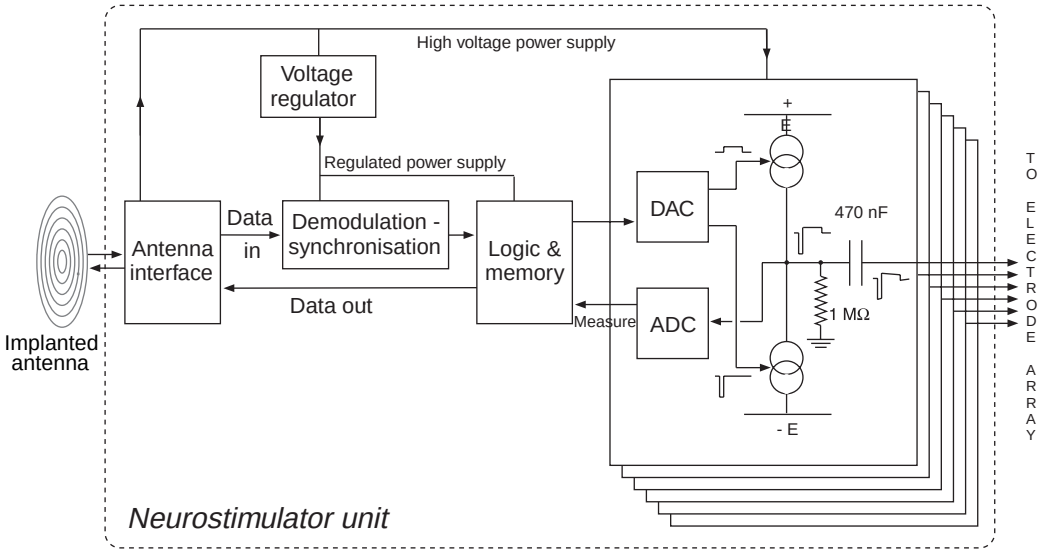


Figure 5. Detailed representation of the implanted components of Figs. 2 and 3.

Table 1. Example: Main Specifications of the Optic Nerve Prosthesis

Image processing	
Pixel size	$1 \times 1^\circ$
Processed visual field	$32 \times 64^\circ$
Telemetry	
Forward carrier frequency	12 MHz
Forward data rate (Frequency Shift Keying modulation)	$3\,\text{Mbit} \cdot \text{s}^{-1}$
Data frames length (including 6 bit CRC)	64 bit
Return carrier frequency	24 MHz
Return data rate	$1.5\,\text{Mbit} \cdot \text{s}^{-1}$
Maximal power transmission (class E)	120 mW
Maximal antennae separation distance	5.5 mm
Implant	
Power consumption	80 mW
No. of independent current sources	8
Current sources	
Time resolution	21 μs
Output voltage range	−9 to +9 V
Maximum output current	3.2 mA
Current resolution (exponential)	8 bit + sign
Electrodes	
Number of contacts	8
Contact recess depth	80 μm
Contact area (platinum)	$0.2\,\text{mm}^2$

standard correction features for luminosity compensation, autofocus, and avoidance of glare (e.g., could greatly simplify the later required image analysis). Later, when the usefulness of some image analysis procedures, such as edge enhancement or nonuniform resolution will have been demonstrated, it might become worthwhile to consider implementing such features in the front end hardware. The video camera would then be replaced by a specific imaging device that could evolve into a real artificial retina.

In conditions, such as blindness, due to retinitis pigmentosa, the main target group for all prechiasmatic types of visual prosthesis, the optics of the eye can still be functional. In the subretinal approach, a photosensor array attached to the retina would transform the eye in an artificial video camera, preserving an essential functionality of the natural eye, namely, gaze orientation . However, in addition to biocompatibility requirements, mounting such a device as well as the necessary control electronics will not be easy (15). Provision must also be made to power the device and send its output signal to an external processor, all with an acceptable level of power dissipation (30).

The external processor. The bottleneck of all visual prosthetic systems is the rather limited quantity of information that can be handled by the neural interface. The amount of data to be transmitted must therefore be reduced by all possible means including image analysis. This topic is likely to become very important in the near future. Limiting images to black and white, thresholding, and edge detection are just preliminary steps. More sophisticated image processing techniques will have to be implemented. Therefore, powerful belt wearable processors will be necessary to translate visual data to neural stimulation (see Fig. 4). Quite unlike the situation with cochlear implants, however, little is known about the precise encoding of visual information in the visual pathways and the first human implants will contribute to such knowledge (31) allowing more efficient algorithms to be developed.

Image processing can be subdivided in several steps including: analysis, selection, mapping, and encoding. The purpose of the image analysis is essentially to reduce the amount of visual data to be transmitted through the prosthesis. After an image data reduction step, a selection procedure should allow only the most important features to be send through. Mapping refers to the method used to establish a correspondence between selected image pixels or features and the phosphenes that can be generated. Finally, a control signals must be generated such that the implanted current sources will issue the intended stimulus. This last encoding step is entirely defined by the characteristics of the implanted device. Mapping on the other hand is linked to the neural code in the neural interface. Much of it is unknown and still requires experimental work with implanted volunteers. Later on, because of the anatomical variability, at least some individual remapping will be required before revalidation itself can be started with a visual prosthesis. Finally, except for the most basic and empirical aspects, image analysis and item selection will have to resort to further psychological studies on perception. Some of these studies could be done with healthy volunteers using virtual reality simulations.

A detailed description of one example of stimulation algorithm is given below. It should be stressed that a communication interface with the processor (see Fig. 4) is an important tool in the development and adaptation of the translation algorithm. Using monitoring equipment, the perceptions of volunteers can be explored and the system can be customized or adapted according to the findings.

Typically attached to the external processor, and probably the major weight to carry along, are the power supply batteries. Note that all published visual prosthesis designs use externally powered implants. Therefore, current is also drawn from this main battery to provide power to the implant. The main specifications to be taken into account are the user friendliness of wearing and reloading the rechargeable batteries. Their size will be defined by a trade-off between weight and autonomy.

The Transmission Unit. Electromagnetic telemetry is based on the classical cochlear implant design. It typically uses two similar antennas holding a small biocompatible (stainless steel encapsulated) magnet in their center. One of these is implanted under the scalp just above the mastoid, behind the ear. The magnets maintain the external antenna over its internal counterpart. This turns out to be the most popular transcutaneous transmission system. The normal skin thickness separating the antenna coils is from 3.5 to 5 mm. However, just after surgery, swelling up to $\sim$7 mm can be observed that can take $>$3 months to recede. During this period, the increased distance might cause malfunction of the antennas.

A telemetry return channel from the implant to the external system is an important feature, providing acknowledgment signals as well as technical diagnostic and monitoring information. For example the output voltage of the implanted controlled current sources gives an indication about the proper operation of the system as well as an estimation of the electrode contact impedance. Also, a measurement of the supply voltage indirectly proves that the power transmission is working adequately.

Alternative antenna arrangements have been considered. For example, as proposed by the Boston group (see Fig. 6), an external primary antenna or coil could be mounted on spectacles and the secondary coil could be implanted on the eye surface or in the anterior segment of the eye. Other transmission techniques exist. A transcutaneous socket on the head (32) has been used but it exposes the patient to severe infectious complications (33) and it is not really acceptable on esthetical grounds.

An all-in-the-eye alternative has been suggested whereby an infrared (IR) (820 nm) laser would transmit power and signal to an intraocular prosthesis through the transparent media of the eye. Data can be transmitted efficiently but heat dissipation in the implanted components is still not compatible with the power requirements of practical devices. Eye movements would also represent a tremendous challenge.

Implanted Components

The Stimulator Case. The following description pertains to the intra-orbital optic nerve visual implant. The purpose of choosing one example is to provide a set of realistic numbers, but the principles apply to most visual prosthesis approaches. The antenna is connected to an 8 mm thick titanium case half engraved in the parietal skull. This neurostimulator hybrid circuit (see Fig. 5) contains the transmission electronics as well as control logic circuits and the

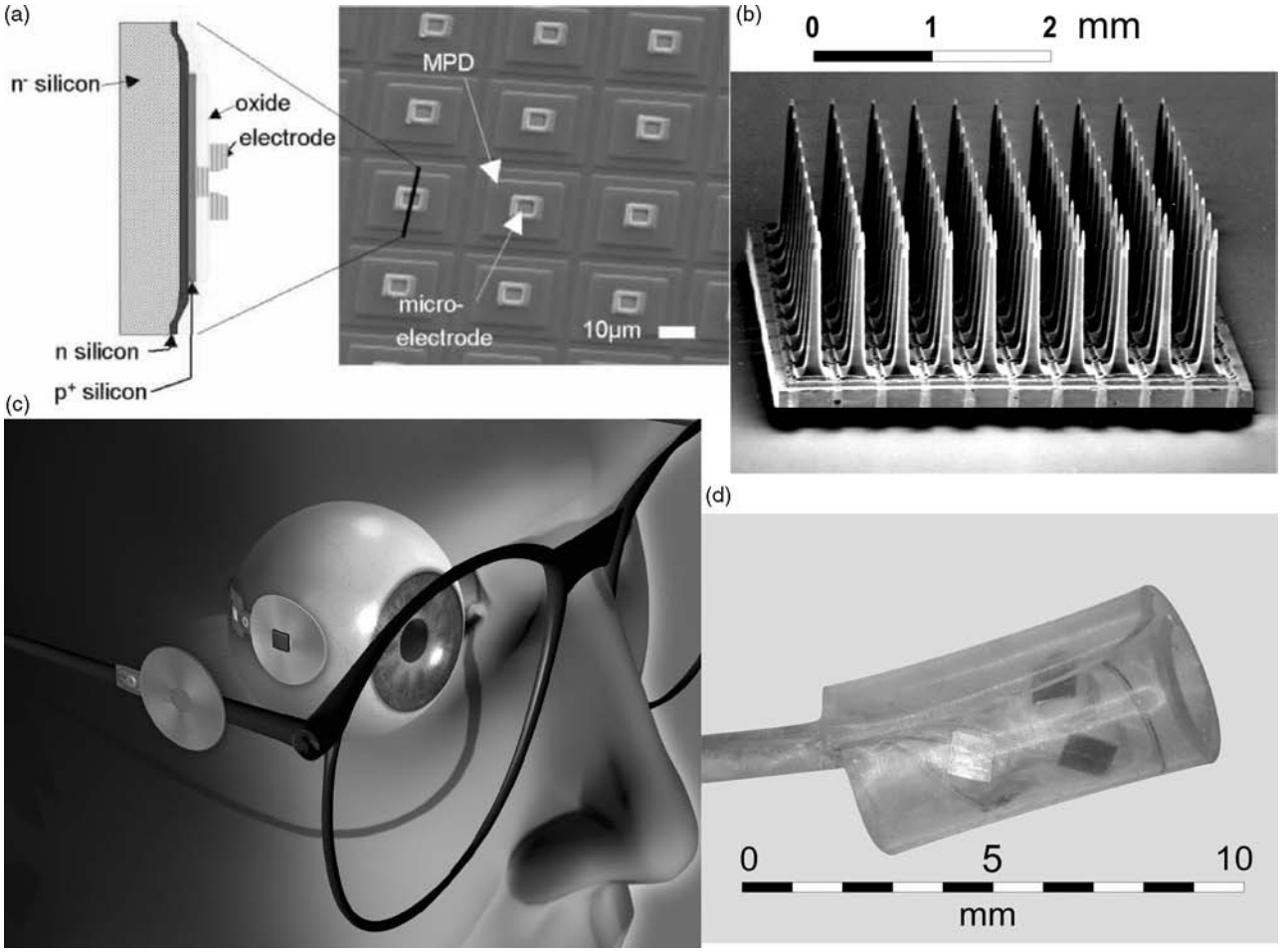


Figure 6. Picture of a number of different electrodes characterising the various approaches to the visual prosthesis. (a) Details of a MicroPhotoDiode (MPD) chip intended to be implanted subretinally in the space normally filled up by cones and rods. (Courtesy of Eberhart Zrenner.) (b) The Utah Cortical Electrode Array. (Courtesy of Richard A. Normann.) (c) An overview of the epiretinal prosthesis with wireless radio frequency transmission being built by researchers at the Boston VA hospital, Harvard Medical School and Massachusetts Institute of Technology. Only a small square end on the left will enter the eye. (Courtesy of Joseph Rizzo.) (d) The optic nerve spiral cuff electrode.

current sources. This circuit occupies the major part of a titanium encasing is closed by a laser welded plate carrying 10 feedthrough connections. This number of feedthroughs is a main factor in determining the minimal size of the stimulator case. The metal encasing is internally connected as the common output reference electrode. A biocompatible polymer cap carries two connectors (a two-contact one for the antenna and an eight-contact connector to the electrode) and protects their junction to the feedthrough wires.

Each stimulator connection receive has two current sources, one for each polarity, as required for the application of biphasic charge balanced pulses. As a rule, implanted current sources are coupled to the stimulation electrode through large output capacitors for safety reasons. These capacitors (470 nF), significantly contribute to the size of the implant.

In an attempt to reduce the chip size while providing a large number of channels, alternative designs (34) do multiplex the output of a single stimulator to a number of electrode contacts instead of having one dedicated stimulator circuit for each contact. Also, the telemetry data rate could be reduced if less degrees of freedom were given in the stimulus definition. The possibility to modify the shape of each pulse individually does indeed require a high rate data transfer.

The Neural Interface or Electrode Assembly. Electrodes (see Fig. 6) are the most characterizing component of the different visual prosthesis approaches. They represent a rather critical and important component in electrophysiology applications (see the chapter on Bioelectrodes). In a nutshell, electrodes form the interface between the electric current carried by electrons along metallic conductors and the ionic conduction found in aqueous solutions, such as the body tissues. This potentially harmful transformation is chemical in nature, called a reduction reaction at the cathode and an oxidation process at the anode. A local pH change is induced, becoming more acid at the anode and more alkaline at the cathode. Of course, the exact nature of the chemical reactions taking place will be influenced by the electrode metal as well as the chemicals present in the solution, their concentrations and the applied potential among other factors. Irreversible chemical reactions cause corrosion and might severely limit the electrode lifetime. However, electric charge limits can be defined within which the chemical changes remain mostly reversible. Maximal values of $0.4\,\mathrm{mC \cdot cm^{-2}}$ (35) are given in the literature for platinum and $4\,\mathrm{mC \cdot cm^{-2}}$ for iridium oxide (36). These metals are most often used in implantable electrode contacts. The reversible charge limits will preserve the electrode contacts, but they have no direct bearing on the potential damage to the living tissues being stimulated.

As a rule, biphasic stimulation pulses are used in implants in order to reduce the risk of deleterious effect. The cathodic phase is meant to activate the neural tissue while the anodic phase, often of smaller amplitude, but with a correspondingly longer duration, is supposed to compensate for the injected electric charge and thereby revert the local chemical changes. Because the currents are not uniformly distributed over a contact area, the electrode geometry should also be taken into account in estimating the allowable charge limit.

Table 2. End Pulse Voltage–Current Ratio[a]

Pulse Duration	42 μs	342 μs
	Cathodic Current	
0.4 mA	2.4 kΩ	5.0 kΩ
1.6 mA	2.3 kΩ	3.2 kΩ

[a]Platinum contacts of 0.2 mm^2 referred to titanium case as anode

The contact impedance or voltage–currents is an important characteristic of the electrodes. Dependent on a large number of factors, it is nonlinear and decreases with the current strength while increasing with the pulse duration. Table 2 gives average values as measured in the example of the optic nerve prosthesis. With epiretinal stimulation (37), using 1 kHz 10 μA sine waves, impedance values are found to be distributed from $<1\,\mathrm{k\Omega}$ to $40\,\mathrm{k\Omega}$ for the same contact area 0.2 mm^2 and larger values for 0.05 mm^2 contacts. The higher values correspond to lower thresholds and are found when the contacts are closer to the retina.

The Prosthesis at Work

Phosphene Generation

Phosphene Description. A first step in the rehabilitation process with a visual prosthesis is to explore the available tools, that is, the identifiable phosphenes that can be generated by the system. A large number of stimulus variable (selected contact, pulse amplitude, pulse duration, train duration and frequency) vectors are systematically tested. The resulting perceptions must be carefully recorded for later use. A more straightforward relationship between the stimulation parameters and especially the contact being activated and the perceived phosphene localization can be expected with the retinal approaches. Still, however, the thresholds will have to be determined and data will be required to correct for the expected (18) mapping distortions.

With the optic nerve prosthesis, the phosphene diameters are variable from a point-like perception to $>10^\circ$. The distribution of these values has modal peaks $\sim$2, 4, and 7° depending on the stimulus parameters. The luminosity is most often reported as weak. Some phosphenes, especially at the periphery of the visual field have the appearance of solid patches while others have variable textures, such as orderly arranged rows or columns of points or small lines. Phosphenes are colored sometimes with contrasting colors between texture elements and the local background. The global background of the visual field is usually described as black, but can sometimes be perceived as gray or slightly colored. Fluctuations of the diffuse background or spontaneous phosphenes occasionally hinder the perception of induced phosphenes.

In the optic nerve example, phosphenes remain located in a quadrant of the visual field that reflects the retinotopical position of the contact used. The center of the phosphenes is located peripherally if near threshold, high frequency and prolonged train stimuli are used and their position in the visual field takes on a more central position for stronger

stimuli, the strength being expressed as a stimulation to threshold current ratio. The most eccentric position that can be accessed by a given train stimulus is located nearer the center for short and/or low frequency trains (31). Phosphenes generated by single pulses are systematically located in the center of the visual field. It is thus possible to model and control the average phosphene center position in the visual field. Unfortunately, in the first implanted volunteer, phosphenes could only be induced in a small region of the visual field (from 8.5° on the left, to 11.5° on the right, and from 6° up to 47° down). This limitation is likely to be linked to the retinitis pigmentosa itself as it is known that only a fraction of the ganglion cells survive this disease (8) and the periphery is more affected than the central retina (38).

With retinal stimulation as well, it has been found that the form of the percept did not always match the stimulation pattern (18) despite the fact that small phosphenes could be obtained at visual field positions corresponding to the point of the retina being stimulated (18).

For all approaches, the phosphene position is clearly referred to retinal coordinates and therefore to the eye orientation at the time of stimulation. Any change in gaze direction or head movement preceding the stimulation will accordingly shift the perceived phosphene location. Thus, for every visual prosthesis with an image capturing device that is not attached to the eye, the users must learn to keep their eyes fixed in the orbit.

Phosphene perceptions are described as short-lived, flash-like. With repeated stimuli, flicker fusion has been found to occur $\sim$ 8–10 Hz and is associated with a perception threshold decrease. At very low repetition frequencies (1–2 Hz) flashing phosphenes remain of stable brightness as long as the stimulation is maintained. By contrast, at higher frequencies, successive phosphenes are perceived a with decreasing brightness and disappear after 1–3 s.

Brindley (cortex, subdural stimulation) did not obtain a flicker fusion of cortical stimuli. Flicker fusion frequency was found to be similar for normal vision and for electrically generated perceptions using epiretinal electrodes (39). In this last approach, the perceived stimulus brightness increased with increasing stimulus rate as well as with increasing the stimulus current, which is not the case with optic nerve stimulation whereby the perceived phosphenes become larger with stronger stimuli. This difference might be related to the fact that the epiretinal stimulus can activate bipolar cells in addition to ganglion cells.

Simulation on Mathematical Models. Computational models play an important role in this kind of work. Numerical evaluation of the electric potential distribution in inhomogeneous and anisotropic media (volume conduction model), and of the resulting neural activation (neural membrane model), can provide a better understanding of the observed responses and provide a predictive design tool. As later confirmed experimentally (40), modeling the epiretinal stimulation has shown that cells or the initial segments have lower threshold than the passing axons (41). For pulses longer than 500 μs, however, the bipolar cells become activated first. Confronting modeling results with experimental data has been useful in the interpretation of the results of optic nerve stimulation studies as well, showing that the retinotopic extension of the perceived phosphenes is quite unlike the corresponding topologic distribution of the activated axons in the optic nerve.

Other techniques including an hybrid association between an adaptive neural network model and analytical expressions of physiological laws can describe the expected localization of phosphenes generated by optic nerve stimulation. These tools will in the future efficiently help to develop individual lookup tables derived from preliminary open loop testing sessions so that these can be kept as short as possible. Such tables or mapping systems are of primary importance for the development of a visual prosthesis, no matter the approach.

Stimulation Thresholds. Despite the functional importance of the activation threshold, values applicable to a given situation are hard to find in the literature. A first reason for this is the lack of standardization. Current-controlled stimulators being used in most approaches, threshold current intensity could seem to be the logical variable to describe. However, as shown by the strength–duration relationship, the electric charge required for activation approaches a minimum asymptotically for short pulses, but increases very fast for pulses longer than the above the chronaxy value (see the article on Functional Electrical Stimulation). A threshold electric charge for realistically (in terms of the electronic circuits and tissue or stray capacitance) short stimuli can thus be seen as an appropriate expression, with the advantage that it directly relates to potential electrochemical damage to the electrode and surrounding tissue. The charge per unit area of electrode contact would be an even better characteristic value including the size factor as a minimal geometric parameter.

In an homogeneous medium, and at a distance much larger than the electrode contact diameter, the thresholds current is proportional to the square of the distance to target ($I_{\text{th}} = k\,r^{a}$), where $a = 2$. This power law can be seriously distorted in real tissues. For example in the retina (ganglion cell stimulation) exponent a can vary between 0.84 and 3.19 depending on electrode geometry and anatomical factors (29). Threshold values can further be expected to depend on factors such as the pulse shape, the nature of the primary target and the animal species.

The nerve membrane activation is not necessarily linked to the production of a cortical evoked potential neither the perception of a phosphene. For example, with optic nerve stimulation, long high frequency train stimuli have been found to yield much lower perception threshold than identical single pulses while in both cases expected to activate the axons at the same level (30).

Another aspect to be considered is that a disease, such as retinitis pigmentosa itself, can be responsible for an important increase in the stimulation threshold (42). The design of a prosthesis must therefore take this fact into account as well.

With a subretinal array, 1 nC or 10 $\mu C \cdot cm^{-2}$ is reported to be sufficient to excite surviving retinal neurons (43). Phosphenes are generated by epiretinal stimuli of 350 $\mu C \cdot cm^{-2}$ through an electrode contact with a diameter

of 520 μm (19). However, this threshold can vary between 24 and 702 μA depending on the proximity to the retina (37). The same authors indicate that there is no simple relationship between the threshold and the electrode contact area. For the transretinal stimulation, a threshold value of 56 $\mu C \cdot cm^{-2}$ is given, corresponding to a total charge of 28 nC (20). In the case of the optic nerve prosthesis, an estimation for trains of 25 pulse (200 μs duration) at 200 Hz would yield a threshold of 60 μA with the intracranial cuff (without dura mater) and 700 μA with the intracortical cuff. The respective threshold charges are 6 and 70 $\mu C \cdot cm^{-2}$. The corresponding chronaxy values are 130 and 192 μs. Brindley (1) gives a value of 13 V on an impedance of 3 kΩ, yielding 4.3 mA to reach threshold with a good electrode driven at 30 Hz with 200 μs pulses. This represents a charge of 860 nC. Thresholds down to 0.4 nC were obtained with 200 μm^2 intracortical electrodes, which corresponds to 1.9 μA for 200 μs pulses at 200 Hz (24).

Stimulation Upper Limits. The upper limit of the stimulation range (44) is even more important to the designer than the threshold. The maximal current will indeed define the required minimal supply voltage as calculated from the expected load impedance. Any stimulus strength above a full activation of the target would represent a waste of the stimulator range and lead to unwanted physiologic response. However, despite the fact that they require higher current values, short stimulation pulses in fact activate the structures with lower charges, and therefore with a lower risk for local tissue and electrode damage. Unfortunately, the maximal current available is directly related to the size of the stimulator ASIC, which should be minimized, especially in a multichannel implant.

Increasing the electrode area to reduce the current density might seem to be an alternative. However, much of the gain in safe stimulus strength could easily be lost in the higher threshold characteristics of larger contacts that cannot be placed close to the target. Every approach will thus lead to a trade off between the stimulus strength required for proper activation, the safety limits and the size of the implant. In the example of the intracranial stimulation of the optic nerve, stable threshold levels have been observed for >6 years of compliance with an upper limit of 150 $\mu C \cdot (cm^2 \cdot phase)^{-1}$ for charge compensated biphasic pulses up to 50 Hz and <50 $\mu C \cdot (cm^2 \cdot phase)^{-1}$ at higher frequencies.

From Phosphene to Visual Perception (Stimulation Algorithm). Once phosphene perceptions can be elicited in a controlled way, one is left with the question as how to use them in order to convey visual information to the visual system of the blind person. In the optic nerve visual prosthesis example, a look-up table is first established using a phosphene position model as a mean to average and interpolate limited experimental data. From the collection of theoretically elicitable phosphenes, only those obtained with a charge density <300 nC/phase (0.2 mm^2 contact area) are considered. Train stimuli longer than 40 ms total duration are excluded as well. The phosphene center has to be at least 1° of visual angle apart and, in cases of choice, the phosphene produced with the lowest stimulus strength is selected. These criteria resulted in a set of 109 individually addressable phosphene with defined position. The portion of the visual field covered is limited to 14° horizontally and 41° vertically. Even within that region, there are holes where no phosphene can be obtained. In addition, the phosphene area is usually >1° and there is thus a clear overlap in the patches of visual field covered by neighbors. The usable look-up table is thus far from representing a complete set of nicely tiled point sized light perceptions covering the entire visual field.

Therefore, black and white images from a 108° head-mounted camera are cropped and digitized to an array of 32° × 64° field of view with one square degree pixels and 8 bit resolution. Next, thresholding is applied in order to further reduce the amount of visual information. In some tasks, image processing also includes edge detection. In real time, the processed image is then superimposed on the position coordinates of the phosphenes of the look-up table. If there is a coincidence between a phosphene position and any part of the image, the corresponding stimulation variable values are selected and send to the optic nerve stimulator. A list of the last 10 occurrences is continuously updated in order to avoid repeatedly inducing the same visual sensation when there are several coincident phosphenes. The least frequently used coincident phosphene is always chosen as the next stimulus. When a single phosphene is generated for each frame captured by the camera, 25 phosphenes can be induced per second.

The processor software and data can be accessed through a communication port (see Fig. 4) allowing to modify parameters, data tables, or the applied algorithms. For example, the random selection above could be replaced by a nearest neighbor selection or reference tables and parameters could be adapted to the perceptions reported by the volunteer. Some authors (45) have proposed a fitting optimization algorithm comparing the input images with the generated perception. This could work as an automatic processor training method. A major problem, however, is to make the subjectively perceived image available for comparison with the input counterpart. It is likely that much *a priori* knowledge will always have to be included in any system. A large share of that knowledge is still not available and will come, among others, from the contribution of first blind volunteers to the preliminary experiments.

From Image to Phosphene Production (Image Processing). No matter what encoding algorithm is used, it is obvious that the amount of information that can be transferred on by present day visual prosthesis prototypes remains extremely poor compared to real world images. Although some improvement can be expected from further interface developments, severe image reductions will be unavoidable for quite some time.

Using some form of virtual reality in healthy volunteers, a few teams have explored the minimum requirements to obtain a useful pixelized vision according to the behavioral task involved. An array of ~600 dots is found to be a minimum for useful reading performance (46). Face recognition could be obtained with 10 × 10–32 × 32 grids (47). By contrast, some subjects can recognize simple objects and

symbols using a 4×4 pixel simulated prosthetic vision (48). Subject's performances clearly increase with training.

It should be stressed, however, that quite unlike the image pixels used in simulations phosphenes are not identical point-like ordered spots that neatly tile a surface. Also, these laboratory experiments only deal with experimental objects presented in an otherwise empty environment. In the real world, subjects will first have to localize and segregate target objects by some preprocessing. However, little has been achieved along these lines hitherto.

Vision Rehabilitation. The ultimate goal of a visual prosthesis is to rehabilitate a visually handicapped person. Results must therefore be evaluated from the blind person's performances point of view and not in terms of device features. Issues, such as the number of available phosphenes or their density, although contributing to the overall performance, cannot be considered as representative of the value of a prosthetic system. Furthermore, isolated analytic characteristics, such as the visual acuity, often mean very little because they can be adjusted by accessories, such as a straightforward optical compensation.

A measurement of the performances will most reliably be obtained in laboratory conditions, but it is the usefulness in real-life that will decide of the success of a prosthesis. For mobility, the prosthesis will be judged against available alternative obstacle detectors and rehabilitation means, such as the long cane or the guide dog in mobility tasks. Stationary visual tasks, such as object localization, identification, and grasping as well as character reading, face recognition, and scene identification are less likely to be available through alternative means. In such tasks, the error rate and the time to task completion will probably be major criteria for acceptance.

Some evaluations of visual prosthesis implant prototypes have been published (19), showing that light and movement can be detected and simple shapes recognized. With the optic nerve implant, basic patterns formed by bars of 22×320 mm and backprojected on a screen as black shapes against a white background can be recognized with the optic nerve prosthesis. The volunteer, sitting at a distance of $\sim$0.5 m from the screen, uses scanning head (and hence camera) movements to explore the pattern, then draws the perceived figure using aluminum rods. A learning effect can be demonstrated as well as an improvement of the results with the number of elicitable phosphenes used in this test (49). A score of 87% of correct recognition is obtained with 109 phosphenes after training. Simultaneously, the task time decreased from $>$2 min to $\sim$53 s (49). With the same system and after substantial training, the volunteer was able to localize, discriminate, and grasp objects on a table in front of them. Three among six familiar objects lay each in one of the nine subdivisions defined on a table surface. Grasping a specified object among the three was systematically successful in $\sim$60 s. In both experiments, the scanning strategy probably explains the prolonged task completion times.

The emergence of multiple alternative designs and improvements for the prosthesis now call for evaluation standards. As suggested by these early results, these should include stationary tests (pattern identification, figure orientation, object localization, object discrimination, and grasping) as well as mobility trials (obstacle localization and avoidance, landmark localization, and identification). Evaluation of the usefulness of the visual prosthesis in a natural environment will be essential. Further down the road to improvement, face recognition, scene identification, and finally reading tests will perhaps also be considered but much better resolutions are still required (29).

HUMAN AND MEDICAL ASPECTS

Candidates for a Visual Prosthesis

No matter what approach is chosen, all visual prosthesis system require a functional visual brain to ultimately interpret their output. This means that, with today's limited performances, only people losing sight after the critical period of development can be considered as suitable candidates.

Another selection criterion is the severity of blindness. As long as the performances of the systems are questionable, only totally blind persons should be considered as candidates. The risk of interfering with residual vision can indeed only be taken if the expected results are significantly better than the remaining visual abilities. In addition, the evaluation of the rather limited performances of an implant could be obscured by any surviving visual functionality.

Other selection criteria are dependent on the chosen approach. All the prechiasmatic approaches (subretinal, epiretinal, optic nerve) require the survival of retinal ganglion cells and their axons in the optic nerve. Terminal retinitis pigmentosa emerges as the condition most typically fulfilling all the selection criteria. Except in cases of brain lesions, the cortical approach would be more generally applicable, including in many cases of acute blindness where the psychological distress is usually more important. Finally, the individual's general health should also be considered because satisfactory candidates are usually terminal cases of RP and therefore rather old while the implantation surgery requires a prolonged anesthesia.

A complete assessment procedure must precede implantation.

A standard ophthalmologic examination is a good starting point. Because of the chronic nature of the condition, the diagnosis should be checked according to up-to-date knowledge. Some candidates have not had an ophthalmologic investigation for many years and only know that they have an incurable eye disease. The blind patient could be unaware of some additional eye problem or other interfering condition. A proper evaluation of the total degree of blindness is necessary and objective tests, such a absent VEP and ERG, are very useful for comparison with the postimplantation evaluation.

Taking into account the heterogeneous nature of retinitis pigmentosa, this diagnostic label is not enough to warrant the survival of ganglion cells in a given patient. Eyelid surface stimulation is the technique of choice here. A cathode is placed on the closed eyelid while an anode is stuck on the heterolateral mastoid (50). Small current

pulses allow to generate a phosphene perception. In healthy subjects, for pulses >2 ms duration, phosphene perception occurs well before the stimulus can be felt. A threshold strength–duration curve (rheobase of 0.28 mA, chronaxy of 3.07 ms in sighted subjects) can show the perceptions to be genuine. Electrically evoked potentials (51) is an alternative technique that would not have to rely on the patient's subjective perceptions. However, much of this added confidence is lost in important stimulation artifacts and the possible confusion with somatosensory potentials, especially in RP patients in whom thresholds are much higher. Some patients describe relatively abundant spontaneous phosphenes and these can be enhanced by the surface electrical stimulation. The same phenomenon could completely jeopardize the working of a visual prosthesis and perhaps induce permanent unpleasant symptoms.

A psychological evaluation is essential as long as the procedure remains experimental. People do accept the idea of pioneering research, but quite rightly want to make sure they will not be misused as guinea pigs for the sake of science alone. The motivation put forward is to help in the development of treatments or simply for the satisfaction of an active contribution or to give their grandchildren a better chance in the frame of their hereditary disease. Esthetic aspects are questioned right from the first contacts. Visibility of a camera is a point of concern.

When the visual prosthesis will have become a clinical treatment, expectations will still have to be confronted with the systems limitations. Also, much attention should be paid to human factors, such as the impact of an implant on a person's social integration.

Magnetic resonance imaging (MRI) is necessary for some visual prosthesis approaches. For example, the optic nerve size and diameter must be estimated on MRI images for an appropriate nerve cuff electrode to be selected for the optic nerve stimulation. The cortical approach might also benefit from a detailed anatomical image before surgery. An MRI examination could be dangerous and will yield very distorted images after implantation of a prosthesis. If for any reason, it is felt that such images will later be useful, then they should be acquired before implantation.

Classical presurgical investigations including thorax X rays and an electrocardiogram (ECG) are standard presurgical procedures.

Surgical Methods

The surgical method is very specific to each of the approaches. The cortical approach can obviously start with a standard craniotomy. Implantation of intracortical electrodes, however, can require a more specific method including specifically designed instruments such as the pneumatic insertion device (52).

The optic nerve approach has resorted to two kinds of surgical methods, one to place the electrode intracranially just in front of the chiasma and the other to implant the electrode in the orbit. Basically, the intracranial method uses a standard pterional transsylvian approach. That is, after right temporo-fronto-parietal cutaneous incision and preparation of the temporal muscle, a craniotomy is performed at the meeting line between the great wing of the sphenoid with the frontal, parietal and temporal bone (pterion). Opening the dura gives access to the sylvian fissure. From there, surgery further proceeds under a microscope to carefully dissect the arachnoid, opening the sylvian fissure until, in the depth, the optic nerve can be separated from it's surroundings. Only minimal retraction of the basal posterior aspect of the frontal lobe is required. The electrode is then wrapped around the optic nerve and the lead suture to the dura.

The second surgical method involves the implantation of the spiral cuff electrode around the intracortical optic nerve. A medial orbital approach is used. After detaching the internal rectus to allow careful retraction of the eye, a thread can be inserted behind the optic nerve and be used to pull the cuff in place.

The electrode leads exit temporally from the skull (intracranial implant) or from the orbit (intraorbital implant) and run backward under the scalp toward the neurostimulator half buried in a recessed well made in the parietal bone. The neurostimulator is also connected to the antenna inserted under the scalp above the mastoid.

The epiretinal system involves the implantation of a similar neurostimulator and wireless link unit as described above. The placement of the electrode is of course completely different and works (19) as follows (19): The periocular space is reached through a lateral canthotomy. The cable and electrode are passed subconjunctivally all around the eye behind each of the four recti muscle insertions and then introduced into the eye through a 5 mm circumferential scleral incision placed 3 mm posterior to the limbus. Prior to the introduction of the implant, the majority of the vitreous gel is removed. The electrode array is the positioned just temporal to the fovea and a single retinal tack is inserted through the electrode array and into the sclera. The attachment of the electrode to the retina is a main issue here. Several solution have been proposed including biocompatible glues, but recently developed types of miniature nail-like devices called retinal tacks appear to work well.

At least two different subretinal electrode implantation methods have been developed (15). The ab interno technique, follows established vitroretinal and submacular surgery procedures. Surgical instruments are inserted into the eye. The vitreous body is then removed while the intraocular pressure is maintained with infusion. Finally, the retina is locally incised and the electrode array is inserted in the subretinal space using a special forceps. The ab externo implantation is designed to avoid damage to the internal structures of the eye. In this procedure, a flexible foil is inserted into the subretinal space through an incision in the sclera and choroid. The implant is then slit into a macular position along the guiding support foil. The path opened by this implantation can also be used for any required external energy supply leads.

Risks of Active Implants

A complete risk evaluation can only be performed on the basis of a review of a significant number of cases after a long postoperative period. Unfortunately, only a limited number

of human trials have been described so far and most of them using methods too different to allow any globalization. A risk analysis will thus have to consider the various aspects of the implantation in the light of similar procedures. The required anesthesia is in itself a well known, low but finite life threatening factor linked to the duration of the procedure as well as the age and general condition of the volunteer. The surgery as such is often mentioned as the most feared aspect by patients. For most of the prechiasmatic approaches, only structures of a nonfunctional organ exposed and the worst failure would thus result in the loss of the possibility to implant a new prosthesis. Risks linked to the visual prosthesis itself can probably be considered to be similar to those for cochlear implants where they are reported to be negligible (53). In some approaches (intracranial optic nerve and cortical), a craniotomy is performed and electrodes are placed in the direct vicinity or in the brain. In such cases, the possibility of an infection or of an abnormal inflammatory reaction or even direct damage to the brain are potential hazards of major consequences.

The possibility of an infectious metastasis around the foreign material must be borne in mind and is well known from passive implant applications. Similarly, experience with other active implants can help to evaluate the burden represented by possible electromagnetic interference, including airport or other safety systems, mobile telephone, interference between multiple implants and the fact that magnetic resonance imaging is no longer available to these patients.

The possibility of a total or partial failure of the implanted system must be considered. Again, however, useful figures, such as the typical lifetime of an implanted system and the percentage of failure in the initial period, cannot be estimated from the present heterogeneous and limited trials.

Heat production by implanted components and electrochemical changes at the level of the electrode contacts are potential hazards insofar that they are not easy to predict because many factors are involved. Safety limits for functional electrical stimulation are difficult to establish. Electrode failure or tissue damage are real risk factors. Also, unwanted stimulation of neighboring structures can lead, for example, to pain or abnormal muscular contractions.

Living tissues can suffer from the activation itself. Axonal potentials are propagated at the cost of cellular metabolic energy. Too strong a functional demand on some structures could create a state of imbalance between the energy demand and supply, leading to cell dead. This type of limit will vary very much in different tissues. For example, peripheral motor axons typically discharge at frequencies $\sim$ 20 Hz and could be damaged by chronic stimulation at 50 Hz while 100 Hz is a typical frequency for afferent activities in the optic and cochlear nerves.

The electrical stimulation could also induce more subtle changes. For example, with cortical stimulation, there is a possibility to trigger repetitive firing and even epileptic fits. For high stimulation currents, Brindley has indeed described phosphenes lasting minutes after the stimulation has ended (1). Such after-discharges could be minimized by reducing the stimulus charge and avoiding prolonged regular stimuli (24).

Psychological complications represent another possible issue that should be monitored. Blindness is indeed a severe disability to which most of the implantation candidates as well as their surroundings have adapted over time. An effective prosthesis will shake this equilibrium as well as the person's social insertion with consequences that could look paradoxical if only the technical success of the prosthesis was to be considered.

Ethical Aspects

With the project to fight one of the most basic human fears using high tech methods supposed to carry out miracles, the visual prosthesis is likely to enjoy a high profile to the layman and to trigger suspiciousness to scientists. This is thus a very emotional and sensitive subject that could be driven by many political and psychological forces alien to the interest of patients. That is why ethical questions should be dealt with most cautiously, especially in the early stages of development.

Basic ethical rules still derive from the Nuremberg code of 1947. As a consequence of the Second World War, it appeared that compliance with national laws could sometimes lead to unacceptable human behaviors. Some more fundamental ethical principles were raised above the law. This gave ethics a very special status. It is not a set of rules dictated by any form of power but pertains to every human being alike, above national or cultural differences (54). Ethics has neither organized a hierarchical structure nor an absolute reference. Progressively, from conferences to declarations, sets of principles gain universal acceptance.

The World Medical Association Declaration of Helsinki, now at its fifth revision since 1964 (55), is most often considered as the main reference. Laws in democratic countries as well as many organization including scientific publishers enforce these basic principles. The European governments have extended these rules in the Convention on Human Rights and Biomedicine (56).

It is generally admitted that the implementation of ethical principles is very dependent on cultural factors. As a result, compliance of research projects with ethical principles is considered to be ideally insured by submission to an independent local ethical committee. National laws and institutional rules tend to organize the working of such committees. Typical questions investigated by these committees are the quality of the information to the volunteer, signature of an informed consent, the risk/benefit ratio to the volunteer, the evaluation of motivations and free decision as well as a specific insurance coverage including for removal of the implanted material if requested. An absolute preservation of the volunteers private life is a must usually requiring anonymity.

The initial development of a visual prosthesis requires a prolonged collaboration with volunteers. In that frame, it was found that the organization in collaboration with the Ethics Committee in charge, of within project meetings between members of the experimental team and representatives of the volunteers could often be a very useful place to solve communication issues and take some consensus decisions with a volunteer.

A more general ethical question in human research is at what stage to move experiments from animals to humans

(57). Because human applications are the final goal, there is always a point where one will have to decide on the involvement of human volunteers. The prosthesis implantation is not trivial. Although blindness is a severe burden, the expected benefits are not life saving and for the time being, at best limited. The balance between risks and benefits is therefore not established. As a result, the informed consent and the volunteer's motivation become central issues. The adequate information to be provided to a lay candidate pertains to a very technical field and the many scientific uncertainties are even more difficult to explain. As long as the proposed visual implants are rather experimental devices, candidates for an implantation must be explicitly informed and accept that preliminary status. They should also be aware of the fact that the postimplantation fine tuning and rehabilitation might take much more time than what would be expected from a well-established treatment.

The volunteer's motivation on the other hand can be biased by difficulties to cope with a heavy handicap and the possible hereditary character of the disease involved. Time spent with a candidate volunteer can certainly solve many of the information and decision criteria issues, avoiding above all to misuse volunteers as mere study objects.

MAIN ISSUES

Evaluation

The systems presently implanted in human volunteers have the capacity of producing either 16 differently located phosphenes simultaneously (19) either >100 phosphenes send to the optic nerve at a pace of 25 phosphenes per second (49). At first sight, it might seem that the serial phosphene generation used in the optic nerve approach will end up in less information per unit of time than in the parallel scheme. The difference might perhaps not be as important as initially thought considering the power required to drive each stimulator output and the size of each stimulation circuit. In a stimulator with a large number of outputs, some multiplexing scheme is required that makes the system serial as well. The tradeoffs between the distance to target, the electrode contact area, the required stimulation charge, the power dissipation, the size of the current sources and the biologic tolerance might represent the real bottleneck. In 2005, despite measurable results, the performances of the visual prosthesis prototypes remain limited and achieved through an unpractical systematic scanning movement. No global image perception has been obtained so far. Learning effects have been demonstrated however and much can still be done on the side of image analysis to obtain an appropriately reduced amount of visual data. A more optimistic view can perhaps be derived from the past experiences with the now well established and accepted cochlear implants. Based on the number of papers published, the visual prosthesis seems to lag ~20 years behind its auditory counterpart.

Hurdles and Limitations

The number of pixels is not the only issue in the evaluation of the quantity of visual data that can be transferred. In an image, each pixel also carries luminance and color information. Although the phosphene perceptions generated can be colored, no approach has hitherto developed means to control it and this dimension can therefore not be exploited. The brightness of phosphene perceptions is modulated by the stimulus intensity, but intensity also modulates the spatial recruitment and therefore, localization and brightness are not independent. In the optic nerve approach, strong stimuli will yield only large phosphenes centrally located in the visual field. With the retinal approaches a stronger stimulus increases the perceived luminosity but is likely to stimulate different structures and interfere with the neural code. In the case of stronger stimuli on the visual cortex surface, Brindley reported the appearance of additional phosphenes and their persistence up to 2 min after the stimulus ceased. Until now, epiretinal or optic nerve approaches do not explicitly take into account matters, such as the existence of ON as well as OFF ganglion cells. Stimulating more than a single cell or axon perhaps leads to some cancellation of their perceptive effect. In addition, the electrical stimulation may interfere with ongoing spontaneous firing of the ganglion cells. The link between stimulus and light perception can thus be expected to be rather complex.

The number of phosphenes that can be generated is still by far not large enough to come to a realistic pixelization. On the other hand, the central visual network does not build an internal projection of the outside world images as implied by such principles. The photosensors in the eye capture images as pixels, but that is also where the pixel structure stops. All other layers of the retina analyze and encode the image according to many characteristics other than a mere pixel position. Beyond a rough retinotopy, edge enhancement, movement detection, and a few other known features, the spatiotemporal encoding appears to be very complex. Except for the subretinal approach, recreating a natural image perception will require to interface a prosthesis with this complex and largely unknown code. This hurdle is likely to be even more challenging for cortical implants than for epiretinal or optic nerve approaches.

Another difficulty is that as long as a near perfect perception will not be achievable, bilateral implantation will not be able to convey distance information. Other distance clues are also linked to a relatively high quality vision. This means that before visual prosthesis reach a very high degree of performance, indirect systems will be required to provide the important distance information.

Another direction for future developments is set by the need for image stabilization. The subretinal approach uses the eye optics with the result that eye movements are included in the normal physiological network of vision. In all other approaches, an external camera is used. This camera is attached to spectacles and follows head movements instead of eye movements. For a correct localization of the surroundings, the prosthesis user must learn to inhibit eye movements as far as possible and only use slow head movement to scan the environment. This is of course a severe limitation. Eye and head movement tracking methods will at some point be necessary to stabilize images. The most advanced solution would be to implant or integrate a miniature camera in the eye as well. A feedback system taking into account eye movements in the image analysis

might be an intermediary step in the near future, before the intraocular camera becomes a reality.

Further progress in electrode design, stimulation algorithms, and electronic miniaturization will perhaps create the prospect of a system capable to generate a high density of phosphenes issued at cinematic rate. However, only that part of the visual field that is still matched by surviving ganglion cells will be available for rehabilitation, at least in all prechiasmatic approaches. As a result, although it is supposed to access the entire visual field, the optic nerve approach appears to reactivate only a restricted field of view. On the other hand, retinal approaches will have to match the electrode location with the remaining retinal cells in order to avoid an useless stimulation of death tissue. Clearly, terminal cases of retinitis pigmentosa will never regain a complete field using these methods. Better results can of course be expected in less affected patients where the prosthesis could be combined to some method to prevent further degradation.

Future Perspectives, Advanced Applications

The visual prosthesis in no way represents a treatment of the cause of blindness. The degenerative processes of retinitis pigmentosa can thus evolve further. The effect of chronic stimulation on sick ganglion cells is not known. The stimulation could either speed up the degeneration, have no influence on the natural course of the disease or, to the contrary, prevent further cell losses (58). The cause of retinitis pigmentosa lays in the biochemistry of the photosensor cells and the degeneration of the other cell layers of the retina could be considered as the result of disuse as well as the consequence of abnormal chemical intercellular exchanges. Keeping ganglion cells active might help them to survive. If this is true, as has been demonstrated in animals equipped with a cochlear implant (59), then the implantation should be done as early as possible, when a minimal number of bipolar and ganglion cells have degenerated. Recovery of a broader visual field can surely be expected in incomplete blindness stages but interference with existing remnants of vision might create new problems to be solved.

Again in keeping with the evolution of the clinical use of cochlear implants, there might be a good reason to foresee a visual prosthesis implantation in blind children. An efficient prosthesis implanted before the end of the critical period might perhaps induce the development of a functional visual cortex and preserve the individual's chances to later benefit from similar devices in adulthood. The limited efficacy of present day visual prosthetic devices does, however, clearly not yet justify such an undertaking.

BIBLIOGRAPHY

Cited References

1. Brindley GS, Lewin WS. The sensations produced by electrical stimulation of the visual cortex. J Physiol 1968;196(2): 479–493.
2. Easty LE, Sparrow JM. Oxford Textbook of Ophthalmology. Oxford (UK): Oxford University Press; 1999.
3. Hubel DH, Wiesel TN. The period of susceptibility to the physiological effects of unilateral eye closure in kittens. J Physiol 1970;206(2):419–436.
4. Wanet-Defalque MC, et al. High metabolic activity in the visual cortex of early blind human subjects. Brain Res 1988;446(2): 369–373.
5. Marc RE, Jones BW, Watt CB, Strettoi E. Neural remodeling in retinal degeneration. Prog Retin Eye Res 2003;22(5):607–655.
6. Schmidt EM et al. Feasibility of a visual prosthesis for the blind based on intracortical microstimulation of the visual cortex. Brain 1996;119(Pt 2):507–522.
7. Santos A et al. Preservation of the inner retina in retinitis pigmentosa. A morphometric analysis. Arch Ophthalmol 1997;115(4):511–515.
8. Stone JL et al. Morphometric analysis of macular photoreceptors and ganglion cells in retinas with retinitis pigmentosa. Arch Ophthalmol 1992;110(11):1634–1639.
9. Chow AY et al. The artificial silicon retina microchip for the treatment of vision loss from retinitis pigmentosa. Arch Ophthalmol 2004;122(4):460–469.
10. Merabet LB et al. What blindness can tell us about seeing again: merging neuroplasticity and neuroprostheses. Nat Rev Neurosci 2005;6(1):71–77.
11. Iezzi R et al. Biocompatibility of Caging Chromophores for Use in Retinal and Cortical Visual Prostheses, ARVO Annual Meeting, Fort Lauderdale, FL, 5-5-2002. E-Abstract. Available at 4478; http://www.iovs.org/search.dtl.
12. Peterman MC et al. The Artificial Synapse Chip: a flexible retinal interface based on directed retinal cell growth and neurotransmitter stimulation. Artif Organs 2003; 27(11): 975–985.
13. Peterman MC et al. Localized neurotransmitter release for use in a prototype retinal interface. Invest Ophthalmol Vis Sci 2003;44(7):3144–3149.
14. Kohler K, Hartmann JA, Werts D, Zrenner E. Histological studies of retinal degeneration and biocompatibility of subretinal implants. Ophthalmologe 2001;98(4):364–368.
15. Zrenner E. Will retinal implants restore vision?. Science 2002;295(5557):1022–1025.
16. Peyman G, et al. Subretinal semiconductor microphotodiode array. Ophthalmic Surg Lasers 1998;29(3):234–241.
17. Humayun MS, de Juan Jr. E. Artificial vision. Eye 1998; 12(Pt 3b):605–607.
18. Rizzo JF et al. Perceptual efficacy of electrical stimulation of human retina with a microelectrode array during short-term surgical trials. Invest Ophthalmol Vis Sci 2003;44(12):5362–5369.
19. Humayun MS et al. Visual perception in a blind subject with a chronic microelectronic retinal prosthesis. Vision Res 2003;43(24):2573–2581.
20. Nakauchi K et al. Transretinal electrical stimulation by an intrascleral multichannel electrode array in rabbit eyes. Graefes Arch Clin Exp Ophthalmol 2005;243(2):eFIRST-6 Dec 2004.
21. Veraart C, Grill WM, Mortimer JT. Selective control of muscle activation with a multipolar nerve cuff electrode. IEEE Trans Biomed Eng 1993;40(7):640–653.
22. Veraart C et al. Visual sensations produced by optic nerve stimulation using an implanted self-sizing spiral cuff electrode. Brain Res 1998;813:181–186.
23. Branner A, Normann RA. A multielectrode array for intrafascicular recording and stimulation in sciatic nerve of cats. Brain Res Bull 2000;51(4):293–306.
24. Hambrecht FT. Visual prostheses based on direct interfaces with the visual system. Baillieres Clin Neurol 1995;4(1):147–165.

25. McCreery DB, Agnew WF, Bullara LA. The effects of prolonged intracortical microstimulation on the excitability of pyramidal tract neurons in the cat. Ann Biomed Eng 2002;30(1):107–119.
26. Jones KE, Campbell PK, Normann RA. A glass/silicon composite intracortical electrode array. Ann Biomed Eng 1992;20(4):423–437.
27. Normann RA et al. High-resolution spatio-temporal mapping of visual pathways using multi-electrode arrays. Vision Res 2001;41(10–11):1261–1275.
28. Bradley DC et al. Visuotopic mapping through a multichannel stimulating implant in primate V1. J Neurophysiol 2005;93(3):1659–1670.
29. Weiland JD, Liu W, Humayun MS. Retinal prosthesis. Annu Rev Biomed Eng 2005;7:361–401.
30. Gosalia K, Weiland J, Humayun M, Lazzi G. Thermal elevation in the human eye and head due to the operation of a retinal prosthesis. IEEE Trans Biomed Eng 2004;51(8):1469–1477.
31. Delbeke J, Oozeer M, Veraart C. Position, size and luminosity of phosphenes generated by direct optic nerve stimulation. Vision Res 2003;43(9):1091–1102.
32. Dobelle WH. Artificial vision for the blind. The summit may be closer than you think. ASAIO J 1994;40(4):919–922.
33. Normann RA, Maynard EM, Rousche PJ, Warren DJ. A neural interface for a cortical vision prosthesis. Vision Res 1999; 39(15):2577–2587.
34. Jones KE, Normann RA. An advanced demultiplexing system for physiological stimulation. IEEE Trans Biomed Eng 1997; 44(12):1210–1220.
35. Brummer SB, Robblee LS, Hambrecht FT. Criteria for selecting electrodes for electrical stimulation: theoretical and practical considerations. Ann N Y Acad Sci 1983;405:159–171.
36. Weiland JD, Anderson DJ, Humayun MS. *In vitro* electrical properties for iridium oxide versus titanium nitride stimulating electrodes. IEEE Trans Biomed Eng 2002;49(12 Pt 2): 1574–1579.
37. Mahadevappa M, et al. Perceptual thresholds and electrode impedance in three retinal prosthesis subjects IEEE Trans Neural Syst Rehabil Eng 2005;13(2):201–206.
38. Humayun MS, Prince M, de Juan Jr. E, Barron Y, Moskowitz M, Klock IB, Milam AH. Morphometric analysis of the extramacular retina from postmortem eyes with retinitis pigmentosa. Invest Ophthalmol Vis Sci 1999;40(1):143–148.
39. Humayun MS, et al. Visual perception elicited by electrical stimulation of retina in blind humans. Arch Ophthalmol 1996;114(1):40–46.
40. Jensen RJ et al. Thresholds for activation of rabbit retinal ganglion cells with an ultrafine, extracellular microelectrode. Invest Ophthalmol Vis Sci 2003;44(8):3533–3543.
41. Greenberg RJ et al. A computational model of electrical stimulation of the retinal ganglion cell. IEEE Trans Biomed Eng 1999;46(5):505–514.
42. Rizzo III JF et al. Methods and perceptual thresholds for short-term electrical stimulation of human retina with microelectrode arrays. Invest Ophthalmol Vis Sci 2003;44(12): 5355–5361.
43. Gekeler F et al. Subretinal electrical stimulation of the rabbit retina with acutely implanted electrode arrays. Graefes Arch Clin Exp Ophthalmol 2004;242(7):587–596.
44. Delbeke J et al. The microsystems based visual prosthesis for optic nerve stimulation. Artif Organs 2002;26(3):232–234.
45. Eckmiller R. Learning retina implants with epiretinal contacts. Ophthalmic Res 1997;29(5):281–289.
46. Sommerhalder J et al. Simulation of artificial vision: II. Eccentric reading of full-page text and the learning of this task. Vision Res 2004;44(14):1693–1706.
47. Thompson Jr. RW, Barnett GD, Humayun MS, Dagnelie G. Facial recognition using simulated prosthetic pixelized vision. Invest Ophthalmol Vis Sci 2003;44(11):5035–5042.
48. Hayes JS et al. Visually guided performance of simple tasks using simulated prosthetic vision. Artif Organs 2003;27(11): 1016–1028.
49. Veraart C et al. Pattern recognition with the optic nerve visual prosthesis. Artif Organs 2003;27:996–1004.
50. Delbeke J et al. Electrical stimulation of anterior visual pathways in retinitis pigmentosa. Invest Ophthalmol Vis Sci 2001;42(1):291–297.
51. Potts AM, Inoue J, Buffum D. The electrically evoked response of the visual system (EER). Invest Ophthalmol 1968;7(3):269–278.
52. Rousche PJ, Normann RA. A method for pneumatically inserting an array of penetrating electrodes into cortical tissue. Ann Biomed Eng 1992;20(4):413–422.
53. Arnoldner C, Baumgartner WD, Gstoettner W, Hamzavi J. Surgical considerations in cochlear implantation in children and adults: a review of 342 cases in Vienna. Acta Otolaryngol 2005;125(3):228–234.
54. Pellegrino ED. Intersections of Western biomedical ethics and world culture: problematic and possibility. Camb Q Health Ethics 1992;1(3):191–196.
55. T.W.M.A.I WMA. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2000;284(23):3043–3045.
56. S.C.o.B CDBI. (1996). The convention for the protection of human rights and dignity of the human being with regard to the application of biology and medicine: Convention on human rights and biomedicine. Secretary General of the Council of Europe, Strasbourg. [Online.] Available at http://conventions.coe.int/Treaty/EN/cadreprincipal.htm.
57. Troyk P et al. A model for intracortical visual prosthesis research. Artif Organs 2003;27(11):1005–1015.
58. Leake PA, Hradek GT, Snyder RL. Chronic electrical stimulation by a cochlear implant promotes survival of spiral ganglion neurons after neonatal deafness. J Comp Neurol 1999;412(4):543–562.
59. Snyder RL et al. Chronic intracochlear electrical stimulation in the neonatally deafened cat. i: expansion of central representation. Hear Res 1990;50(1–2):7–33.

Reading List

Alteheld N, Roessler G, Vobig M, Walter P. The retina implant–new approach to a visual prosthesis. Biomed Tech (Berlin) 2004;49(4):99–103.

Archambeau C, Delbeke J, Veraart C, Verleysen M. Prediction of Visual Perceptions with Artificial Neural Networks in a Visual Prosthesis for the Blind. Artif Intell Med 2004;32(3):183–194.

Bagnoud M, Sommerhalder J, Pelizzone M, Safran AB. Necessary visual information for restoring reading with a retinal implant in a blind patients with massive retinal degeneration of photoreceptors. Klin Monatsbl Augenheilkd 2001;218(5):360–362.

Bak M et al. Visual sensations produced by intracortical microstimulation of the human occipital cortex. Med Biol Eng Comput 1990;28(3):257–259.

Beebe X, Rose TL. Charge injection limits of activated iridium oxide electrodes with 0.2 ms pulses in bicarbonate buffered saline. IEEE Trans Biomed Eng 1988. 35(6):494–495.

Berardi N, Pizzorusso T, Maffei L. Critical periods during sensory development. Curr Opin Neurobiol 2000;10(1):138–145.

Besch D, Zrenner E. Prevention and therapy in hereditary retinal degenerations. Doc Ophthalmol 2003;106(1):31–35.

Blakemore C, Van Sluyters RC. Reversal of the physiological effects of monocular deprivation in kittens: further evidence for a sensitive period. J Physiol 1974;237(1):195–216.

Chow AY, Peachey NS. The subretinal microphotodiode array retinal prosthesis. Ophthalmic Res 1998;30:195–196.

Chow AY, Peachey N. The subretinal microphotodiode array retinal prosthesis II. Ophthalmic Res 1999;31(3):246.

Chow AY et al. Subretinal implantation of semiconductor-based photodiodes: durability of novel implant designs. J Rehabil Res Dev 2002;39(3):313–321.

Chowdhury V, Morley JW, Coroneo MT. An *in-vivo* paradigm for the evaluation of stimulating electrodes for use with a visual prosthesis. ANZ J Surg 2004;74(5):372–378.

Chowdhury V, Morley JW, Coroneo MT. Surface stimulation of the brain with a prototype array for a visual cortex prosthesis. J Clin Neurosci 2004;11(7):750–755.

DeMarco SC et al. Computed SAR and thermal elevation in a 0.25 mm 2D model of the human eye and head in response to an implanted retinal stimulator. Part I: models and method. IEEE Trans Antennas Propagat 2002;(May 30):1–10.

Elfar SD, Cottaris NP, Iezzi R, Abrams GW. Rapid Mapping of Cortical Multi-Electrode Arrays and Its Application for the Evaluation of Retinal Prostheses, Annual meeting of the Association for Research in Vision and Ophthalmology, Ft. Lauderdale, Florida, E-Abstract 3403. Available at http://www.iovs.org/search.dtl 2004.

Fang X et al. Direct stimulation of optic nerve by electrodes implanted in optic disc of rabbit eyes. Graefes Arch Clin Exp Ophthalmol 2005;243(1):49–56.

Fernandez E, Ferrandez J, Ammermuller J, Normann RA. Population coding in spike trains of simultaneously recorded retinal ganglion cells. Brain Res 2000;887(1):222–229.

Finkelstein D, et al. Visual prostheses and visual rehabilitation in low vision research, assessment, and management. Curr Opin Ophthalmol 1991;2(6):729–732.

Greenberg RJ et al. Electrical stimulation of the human retina: an update. Invest Ophthalmol Vis Sci 1995;36(4):S234.

Grumet AE. Electric stimulation parameters for an epi-retinal prosthesis, Ph.D. dissestation. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1999.

Grumet AE, Wyatt Jr. JL, Rizzo III JF. Multi-electrode stimulation and recording in the isolated retina. J Neurosci Methods 2000;101(1):31–42.

Guven D et al. Long-term stimulation by active epiretinal implants in normal and RCD1 dogs. J Neural Eng 2005;2(1): S65–S73.

Hallum LE, Suaning GJ, Lovell NH. Contribution to the theory of prosthetic vision. ASAIO J 2004;50(4):392–396.

Hubel DH, Wiesel TN. Ferrier lecture. Functional architecture of macaque monkey visual cortex. Proc R Soc London B Biol Sci 1977;198(1130):1–59.

Humayun MS et al. Pattern electrical stimulation of the human retina. Vision Res 1999;39(15):2569–2576.

Humayun MS. Intraocular retinal prosthesis. Trans Am Ophthalmol Soc 2001;99:271–300.

Jensen RJ, Ziv OR, Rizzo III JF. Thresholds for activation of rabbit retinal ganglion cells with relatively large, extracellular microelectrodes. Invest Ophthalmol Vis Sci 2005; 46(4):1486–1496.

Johnson L et al. Electrical stimulation of isolated retina with microwire glass electrodes. J Neurosci Methods 2004;137(2): 265–273.

Kanda H et al. Electrophysiological studies of the feasibility of suprachoroidal-transretinal stimulation for artificial vision in normal and RCS rats. Invest Ophthalmol Vis Sci 2004;45(2):560–566.

Kato S, Saito M, Tanino T. Response of the visual system evoked by an alternating current. Med Biol Eng Comput 1983;21(1):47–50.

Kerdraon YA et al. Development and surgical implantation of a vision prosthesis model into the ovine eye. Clin Experiment Ophthalmol 2002;30(1):36–40.

Lakhanpal RR et al. Advances in the development of visual prostheses. Curr Opin Ophthalmol 2003;14(3):122–127.

Laube T et al. Chronically implanted epidural electrodes in Gottinger minipigs allow function tests of epiretinal implants. Graefes Arch Clin Exp Ophthalmol 2003;241(12): 1013–1019.

Laube T et al. Optical energy transfer for intraocular microsystems studied in rabbits. Graefes Arch Clin Exp Ophthalmol 2004;242(8):661–667.

Lazzi G et al. Computed SAR and thermal elevation in a 0.25 mm 2D model of the human eye and head in response to an implanted retinal stimulator. Part II: results. IEEE Trans Antennas Propagat 2002;(May 30):1–8.

Liu W et al. Electronic visual prosthesis. Artif Organs 2003;27(11):986–995.

Loewenstein JI, Montezuma SR, Rizzo III JF. Outer retinal degeneration: an electronic retinal prosthesis as a treatment strategy. Arch Ophthalmol 2004;122(4):587–596.

Margalit E et al. Retinal prosthesis for the blind. Surv Ophthalmol 2002;47(4):335–356.

Marr D, Poggio T. Cooperative computation of stereo disparity. Science 1976;194(4262):283–287.

Maynard EM, Nordhausen CT, Normann RA. The Utah intracortical Electrode Array: a recording structure for potential brain-computer interfaces. Electroencephalogr Clin Neurophysiol 1997;102(3):228–239.

Maynard EM. Visual prostheses. Annu Rev Biomed Eng 2001;3:145–168.

Montezuma SR, Rizzo III JF, Ziv OR. Differential recovery of the electroretinogram, visually evoked cortical potential, and electrically evoked cortical potential following vitrectomy: Implications for acute testing of an implanted retinal prosthesis. J Rehabil Res Dev 2004;41(2):113–120.

Normann RA et al. The Utah 100 microelectrode array: an experimental platform for cortically based vision prosthesis, ARVO abstract # 192-B103. Invest Ophthalmol Vis Sci 1997;38(4):S41.

Palanker D et al. Migration of retinal cells through a perforated membrane: implications for a high-resolution prosthesis. Invest Ophthalmol Vis Sci 2004;45(9):3266–3270.

Pardue MT et al. Neuroprotective effect of subretinal implants in the RCS rat. Invest Ophthalmol Vis Sci 2005; 46(2):674–682.

Pardue MT et al. Possible sources of neuroprotection following subretinal silicon chip implantation in RCS rats. J Neural Eng 2005;2(1):S39–S47.

Parrini S, Delbeke J, Legat V, Veraart C. Modelling analysis of human optic nerve fibre excitation based on experimental data. Med Biol Eng Comput 2000;38(4):454–464.

Peterman MC et al. Fluid flow past an aperture in a microfluidic channel. Anal Chem 2004;76(7):1850–1856.

Piyathaisere DV et al. Heat effects on the retina. Ophthalmic Surg Lasers Imaging 2003;34(2):114–120.

Potts AM, Inoue J, Buffum D. The electrically evoked response of the visual system (EER). Invest Ophthalmol 1968;7(3):269–278.

Radner W et al. Increased spontaneous retinal ganglion cell activity in rd mice after neural retinal transplantation. Invest Ophthalmol Vis Sci 2002;43(9):3053–3058.

Rizzo III JF et al. Retinal prosthesis: an encouraging first decade with major challenges ahead. Ophthalmology 2001;108(1):13–14.

Rizzo III JF et al. *In vivo* electrical stimulation of rabbit retina with a microfabricated array: strategies to maximize responses for prospective assessment of stimulus efficacy and biocompatibility. Restor Neurol Neurosci 2004;22(6):429–443.

Sachs HG, Gabel VP. Retinal replacement–the development of microelectronic retinal prostheses–experience with subretinal implants and new aspects. Graefes Arch Clin Exp Ophthalmol 2004;242(8):717–723.

Sakaguchi H et al. Electrical Stimulation with a Needle-type Electrode Inserted into the Optic Nerve in Rabbit Eyes. Jpn J Ophthalmol 2004;48(6):552–557.

Sakaguchi H et al. Transretinal electrical stimulation with a suprachoroidal multichannel electrode in rabbit eyes. Jpn J Ophthalmol 2004;48(3):256–261.

Schwahn HN et al. Studies on the feasibility of a subretinal visual prosthesis: data from Yucatan micropig and rabbit. Graefes Arch Clin Exp Ophthalmol 2001;239(12): 961–967.

Sommerhalder J et al. Pelizzone, Simulation of artificial vision: I. Eccentric reading of isolated words, and perceptual learning. Vision Res 2003;43(3):269–283.

Suaning GJ, Lovell NH, Schindhelm K, Coroneo MT. The bionic eye (electronic visual prosthesis): a review. Aust N Z J Ophthalmol 1998;26(3):195–202.

Sugiyama T et al. Optic cup enlargement followed by reduced optic nerve head circulation after optic nerve stimulation. Invest Ophthalmol Vis Sci 2001;42(12):2843–2848.

Suzuki S et al. Comparison of electrical stimulation thresholds in normal and retinal degenerated mouse retina. Jpn J Ophthalmol 2004;48(4):345–349.

Thylefors B, Negrel AD, Pararajasegaram R, Dadzie KY. Global data on blindness. Bull World Health Organ 1995;73(1):115–121.

Troyk PR, Schwan MA. Closed-loop class E transcutaneous power and data link for microimplants. IEEE Trans Biomed Eng 1992;39(6):589–599.

Troyk PR. (2003). Multichannel transcutaneous cortical stimulation system. [Online.] Available at http://npp.ninds.nih.gov/npp/sow/stm256.htm?format=printable.

Uhlig CE, Taneri S, Benner FP, Gerding H. Electrical stimulation of the visual system. From empirical approach to visual prostheses. Ophthalmologe 2001;98(11):1089–1096.

Veraart C et al. Vision rehabilitation in the case of blindness. Expert Rev Med Devices 2004;1(1):139–153.

Walter P et al. Successful long-term implantation of electrically inactive epiretinal microelectrode arrays in rabbits. Retina 1999;19(6):546–552.

Warren DJ, Fernandez E, Normann RA. High-resolution two-dimensional spatial mapping of cat striate cortex using a 100-microelectrode array. Neuroscience 2001;105(1):19–31.

Weiland JD et al. Understanding the origin of visual percepts elicited by electrical stimulation of the human retina. Graefes Arch Clin Exp Ophthalmol 1999;237(12):1007–1013.

Weiland JD, Humayun MS. Past, present, and future of artificial vision. Artif Organs 2003;27(11):961–962.

Wilms M, Eger M, Schanze T, Eckhorn R. Visual resolution with epi-retinal electrical stimulation estimated from activation profiles in cat visual cortex. Vis Neurosci 2003;20(5): 543–555.

Yamauchi Y et al. Comparison of electrically evoked cortical potential thresholds generated with subretinal or suprachoroidal placement of a microelectrode array in the rabbit. J Neural Eng 2005;2(1):S48–S56.

Yanai D et al. The value of preoperative tests in the selection of blind patients for a permanent microelectronic implant. Trans Am Ophthalmol Soc 2003;101:223–228.

Zhou DD, Greenberg RJ. Microsensors and microbiosensors for retinal implants. Front Biosci 2004;10:166–179.

Ziv OR, Rizzo JF, Jensen RJ. *In vitro* activation of retinal cells: estimating location of stimulated cell by using a mathematical model. J Neural Eng 2005;2(1):S5–S15.

Zrenner E et al. The development of subretinal microphotodiodes for replacement of degenerated photoreceptors. Ophthalmic Res 1997;29(5):269–280.

Zrenner E et al. Subretinal implants. Ophthalmic Res 1998; 30:197–198.

Zrenner E et al. Reply to the letter of drs. Chow and peachey: the subretinal microphotodiode array retinal prosthesis II. Ophthalmic Res 1999;31(3):247.

Zrenner E et al. Can subretinal microphotodiodes successfully replace degenerated photoreceptors?. Vision Res 1999; 39(15):2555–2567.

VOCAL REHABILITATION. See LARYNGEAL PROSTHETIC DEVICES.

X

X-RAY EQUIPMENT DESIGN

MARTIN TORNAI
Duke University
Durham, North Carolina

INTRODUCTION

Radiography is a form of physical diagnosis in which X rays are used to obtain medically useful information about a patient. X rays are generated in a controlled way, and a beam of X rays is passed through the patient. The differential absorption of ionizing X-ray radiation by different tissues modulates this beam. The transmitted beam is detected, and its information content is recorded. In addition, radiological imaging methods are used to control, guide, and monitor both diagnostic and therapeutic manipulations such as needle biopsy, percutaneous transluminal angioplasty (see CORONARY ANGIOPLASTY), and radiation therapy.

X-ray equipment is composed of many subsystems and components, each of which presents unique choices and tradeoffs for the equipment designer. These components include means for X-ray production, spatial and spectral shaping of the X-ray beam, patient handling, as well as means for image detection and capture, intermediate image handling, image processing, image display, and data storage. Auxiliary equipment and logic are customarily included in the design and construction of X-ray apparatus for the control, synchronization, and automation of the examination process. The X-ray equipment designer must balance different and often conflicting requirements to obtain an optimum solution for each examination category or application. For example, fine spatial detail requires minimization of both geometric blur (implying a fine focal spot along with low power and low magnification) and motion blur (implying short exposure time with high power and a large focal spot). This leads to specialization of the equipment for different medical applications. In addition to clinical requirements, the apparatus must also be designed to meet a wide variety of industrial and professional standards as well as comply with governmental regulations.

The functional organization of an X-ray imaging system is shown in Fig. 1. In one form or another, these elements are present in all X-ray imaging systems. This article will follow the logical flow shown in Fig. 1, with emphasis given to the material contained in the double-sided boxes. Other articles in this Encyclopedia discuss the remaining topics.

Power Components

The primary purpose of the power components in the generator is to control the X-ray tube voltage, current, and exposure time. The contrast and signal-to-noise ratio (SNR) of a radiographic image, as well as dose delivered to the patient, are partially controlled by adjustments of the X-ray tube voltage. Other influencing factors include the X-ray beam filtration, the X-ray attenuation characteristics of the object (e.g., patient), and the nature of the image receptor. The voltage applied to the tube has a major influence on the shape of the X-ray spectrum emitted from the tube (Fig. 2). This change in spectral shape as voltage is varied produces changes in image quality, including the common image metrics of contrast and SNR (see X RAYS, PRODUCTION OF). An increase in voltage leads to decreased image contrast (see X-RAY ATTENUATION IN MATTER) and increased SNR. In addition, the total quantity of radiation emitted from the tube is a power function of the voltage applied across the tube (approximately a cubic function, depending on the voltage range and beam filtration), resulting in increasing dose to the patient. In the United States, it is common to use voltage as the primary control element for regulating the amount of X rays exiting from the patient (termed penetration).

The amount of radiation emitted from a tube is also controlled by the product of the current flowing through the tube and the time during which the current flows (expressed in units of milli-ampere-seconds, or mAs). Use of exposure time is somewhat limited in an attempt to minimize patient motion. Applied current can greatly increase tube heating and reduce tube life, causing tube cooling techniques to be a major design consideration, especially in applications such as tomography, where high rates of tube exposures are experienced due to the requirement for multiple sequential images to be produced and detected. Tube current is controlled by adjustment of the X-ray tube's filament temperature. In most generators, time is controlled by switches in the primary circuits. Some generators use switching tubes in the high-tension circuit of the generator, or a control element in the X-ray tube. Figure 3 is a simplified schematic diagram of conventional single-phase, line-operated generator. Such a system draws its power directly from the main electrical distribution system. Conventional generators may be designed to operate using either single- or three-phase power. When one considers power-line impedance and generator self-loading, single-phase equipment is usually limited to lower peak power devices, whereas three-phase equipment may be found at any power level. The consequences of power-line loading are discussed briefly at the end of this section.

The need for high-power mobile X-ray generators in hospitals has led to the construction of equipment incorporating energy storage devices, either capacitors or batteries. These generators accumulate energy from a low-capacity power line at appropriately low-power levels, and then they discharge the stored energy through the X-ray tube at a much higher power level. The currently preferred design is the battery-operated mobile generator with a high-frequency inverter.

Line-operated X-ray generators place unusual loads on the hospital's electrical distribution system. Most generators draw X-ray tube power directly from the lines during

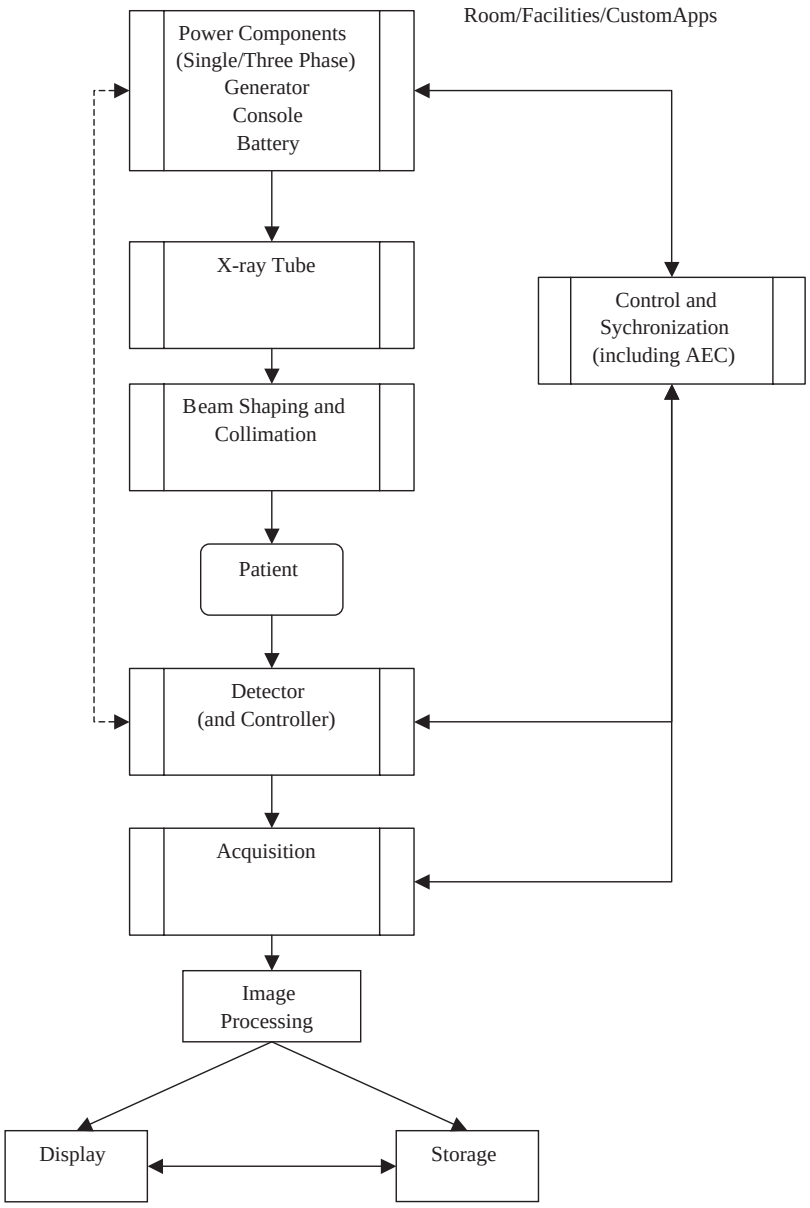


Figure 1. Functional diagram of an X-ray imaging system. Any system used for medical imaging must provide the minimum facilities for delivering a controlled X-ray beam to the patient, positioning the patient relative to the beam, detecting the modulated X-ray beam passing through the patient, and delivering the image to the radiologist.

the time of X-ray exposure. Classic X-ray circuits (Fig. 3) set the X-ray tube voltage by means of measurement of the primary voltage before exposure. The combination of these two effects requires that the power lines be well conditioned and protected from extraneous loads. In addition, other equipment in the hospital must be protected from the effects of the brief high-power demands placed on the line by the X-ray equipment. Also, in this era of ubiquitous digital control, two-way isolation is required to protect both the X-ray equipment and the remainder of the hospital from the broadcast of digital noise, or real digital control signals along the power lines. Power for X-ray generators should be obtained from separate low-impedance supplies, which include dedicated distribution transformers and appropriate size conductors. To minimize conductor size, high-power equipment is usually connected using a higher line voltage. Line-operated X-ray equipment, with the exception of low-power mobile units, is seldom operated from 110 V supplies. Single-phase and low-power three-phase equipment typically operates using a 220 V line. Higher power equipment may require 440/480 V.

An undesirable characteristic of some X-ray generators is ripple, defined as the percent variation of voltage during an X-ray exposure. A high ripple factor can result in undesirable X-ray flux and spectral variations leading to unpredictability in image quality and patient dose. Single-phase, full-wave rectified generators suffered from very large ripple factors, and therefore they have been virtually replaced by the constant potential generator for the most demanding imaging applications, medical imaging being one of them. Although these types of generators are generally larger, more complex, and expensive, they provide extremely low ($<5\%$) ripple factors, and they are essentially constant in their voltage output. These generators typically use single- or three-phase 50/60 Hz line voltages of 220 V or higher. One common type of constant potential

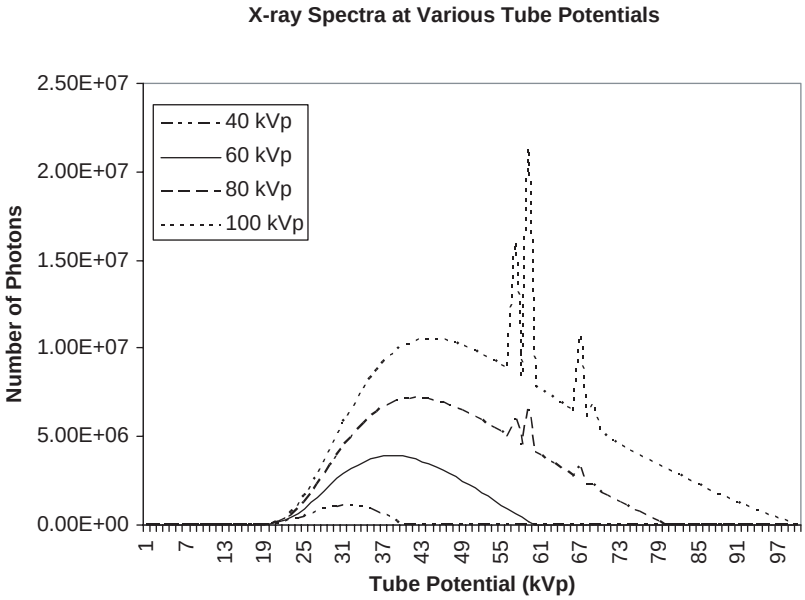


Figure 2. Sample spectra, generated via validated computer simulation, using various tube potentials, 1 mAs exposure, and with 0.5 cm thick Al extrinsic filtration, illustrate the effect on beam shape.

generator is the three-phase generator, using a three-phase transformer with three sets of primary and secondary windings. The newest type of constant potential generator is the medium or high-frequency generator (also called the inverter generator), where the constant potential voltage is controlled by the frequency of the current provided by the chopper/inverter. This generator type is popular both in battery-powered and in mobile units, as well as in the higher power applications of modern equipment. Starter speed is also a consideration available to designers. The starter controls the boost and operation of the rotating anode. Typical low-speed anodes boost in approximately 3 seconds and reach a rotation speed of 3000 rotations per minute (rpm). High-speed starters are often available as add-ons that allow faster boost as well as rotation speeds in excess of 10,000 rpm. Most generators also include a thermal switch that will automatically stop the X-ray exposure when an overheat signal is received from the tube. Additional tube details are provided in the following section.

X-Ray Tubes

The normal X-ray tube is a vacuum diode. Figure 4 is a photograph of a typical X-ray tube housing as well as a cutaway diagram of the typical interior tube components. Electrons are generated by thermionic emission from the filament of the tube. The electron stream is electrostatically focused onto a small target on the anode by means of a carefully designed, shaped, and sometimes charged, highly polished nickel cathode focusing cup.

X rays are produced by interactions between the electrons and the target (see X RAYS, PRODUCTION OF). Most electrons emitted by the hot filament become current carriers across the tube. One can, therefore, set tube current

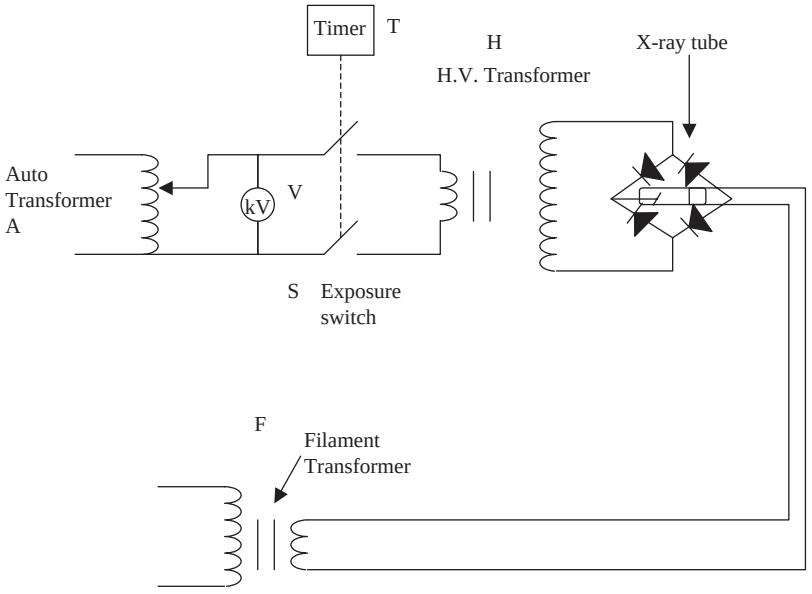


Figure 3. Schematic diagram of a single-phase, full-wave rectified generator. Feedback and stabilization elements are not shown. X-ray tube voltage is adjusted by setting the autotransformer before exposure. Tube current is adjusted by setting the filament voltage and hence its temperature. A, Autotransformer; H, high-voltage transformer; F, filament transformer; S, exposure switch; T, radiographic timer; V, prereading voltmeter used for kV indication.

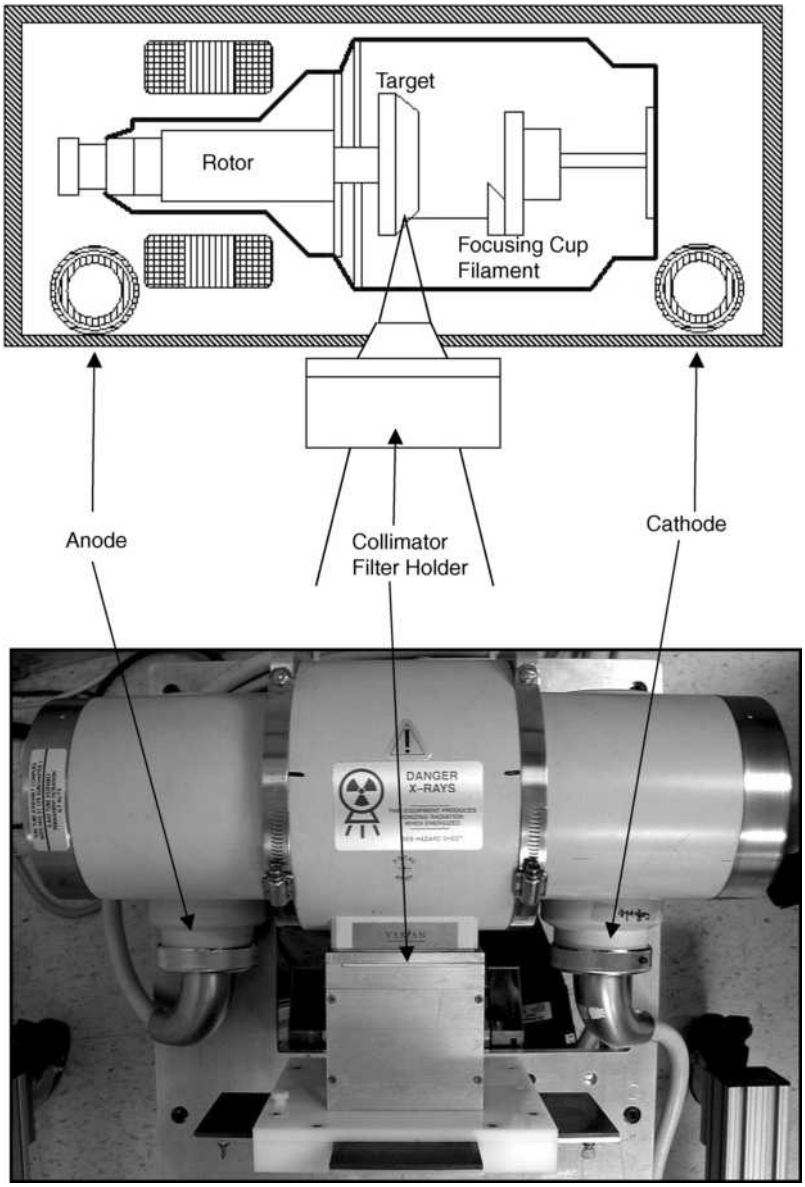


Figure 4. Diagram of a generic X-ray tube and photograph of tube enclosed in its protective housing. The useful X-ray beam emerges from the beam port at the bottom center. The anode is rotated by coupling magnetic fields through the glass envelope of the tube. The stator coils that are used for this purpose surround the tube on the left.

(by adjusting the filament temperature) and tube voltage (by adjusting primary voltage) as independent parameters. Some X-ray tubes function as a triode with a bias voltage applied between the filament and the cathode cup. At low voltage, this bias can be used to modify the size and shape of the focal spot by focusing the electron beam in the tube. A further increase of the bias voltage can serve as a switching device by gating the electron beam ON or OFF.

The X-ray tube converts a very small fraction of the electrical energy delivered to it into useful X rays. In the voltage region corresponding to diagnostic radiology (~20–150 kVp), the physical processes involved in X-ray production (primarily collisional deceleration of electrons in the target) result in an X-ray production efficiency of <1%. The remaining electrical energy is converted into heat. Tungsten is used as the target material for most general-purpose diagnostic X-ray tubes. Its high atomic number ($Z = 74$) maximizes X-ray production efficiency. Tungsten's high melting point (3400 °C) and reasonable specific heat capacity ($C_p = 130$ J/kg-K) help with the thermal problems associated with the waste heat. As special cases, molybdenum (Mo), rhodium (Rh), or Mo/Rh alloy targets are used to produce the specific X-ray spectra needed for mammography. The use of these lower atomic number metals as targets and filters isolates the K-characteristic X rays produced by the target, resulting in a narrow spectral beam at relatively low energy, useful for improving subject contrast in screen-film imaging of the compressed breast. The lower efficiency of X-ray production and lower melting point makes these metals unacceptable for general radiography. Other materials, including silver, cerium, and other exotic materials, continue to be proposed for target materials for specialized applications ranging from small animal imaging to computed tomography and additional industrial applications.

A major task in X-ray tube design is the provision of means for dissipating waste heat before tube structures are damaged. Different tube structures place different thermal

limitations on radiographic techniques and ultimately limit the speed of performing examinations. Tube rating charts that account for these different thermal restrictions are available from all X-ray tube manufacturers. The first consideration is the melting of the tungsten target at the focal spot, the target area bombarded by the electron beam. For sharp projection imaging, the geometry of the X-ray "optics" requires that the focal spot be as small as possible. Design elements such as high-speed rotating anodes (3,000–10,000 rpm) and shallow target angles (5°–15°) are used to limit the power density incident on the physical target location while creating a small effective focal spot, yielding improved resolution. Effective short-term loading in excess of 50 kW/mm^2 of effective focal spot is realized using these techniques. Some tubes even provide user-selectable dual-anode angles or dual-position focal spots for increased flexibility.

Once heat has entered the target region of the anode, it is conducted away from the focal track into the bulk material of the anode. This bulk material usually consists of molybdenum or prolytic graphite. Massive anodes (several kilograms) are used to temporarily store the thermal energy before radiating it to the tube housing and eventually into the environment. The large anode mass, needed for significant thermal energy storage (e.g., for applications in computed tomography and cinefluorography), places limitations on the starting time of the tube, due to the moment of inertia of the anode, and on the length of life of the anode's bearings. The technical requirement of high anode heat storage capacity is optimally met with a heavy smaller diameter disk. The clinical requirement for short radiographic exposure time demands high instantaneous power levels. This requirement is optimally met with a light large-diameter rotating anode disk. These conflicts result in the design and use of many types of X-ray tubes mechanically specialized to meet specific examination requirements.

The tube housing serves several technical purposes. Figure 4 (top) is a diagram of such a tube in its protective housing. The housing is part of the electrical isolation between the high-voltage circuits and the environment. It also provides radiation protection for both patient and operator. Tube housings are lead lined to keep the amount of leakage radiation below legal limits (this requirement assures that the major source of irradiation outside of the beam comes from scatter and the useful beam in the patient). The tube housing illustrated is a variety commonly used, whereby the low atomic number (usually Beryllium) exit window for the X rays is located in the center of the tube between the two poles (cathode and anode) of the power supply from the generator. It is also possible to obtain unipolar X-ray tubes for special applications. With only one single pole for the power supply, it becomes possible to locate the exit window nearer the edge of the tube to allow closer proximity imaging in special-purpose applications, such as in computed mammotomography. Also, these tubes are smaller and lighter than standard dual pole tubes. Drawbacks of such tubes include much lower maximum power (i.e., on the order of ~100 W to 6 kW, compared with the 100+ kW available on other tubes), static anode and, consequently, the need for a larger focal spot to reduce concentration of heat on the anode. These tubes are usually liquid cooled.

Finally, the tube housing is a key portion in the waste-heat handling system. Housings for tubes used at low mean power levels (<100 W or so) can be adequately air cooled, with or without a fan. As the mean load increases (i.e., for applications such as angiography and computed tomography), air cooling becomes inadequate. Additional cooling may be obtained by circulating liquid through a heat exchanger contained in the tube housing or by circulating insulating oil through an external radiator.

Collimation and X-Ray Beam Definition

The geometry of projection X-ray imaging requires that the produced X-ray photons be directly projected from the focal spot onto a detector system. Radiation produced elsewhere in the X-ray tube (off-focal radiation) and scattered radiation from the vicinity of the tube contribute to geometric blur in images. Scattered radiation emanating from the vicinity of the patient, patient supports, and image receptor may not cause observable blur but will still degrade image SNR and contrast. Patient integral radiation dose is minimized, and image quality is maximized by confining the beam of radiation to the smallest possible area. Patient dose and radiological contrast are also influenced by the X-ray spectrum. This section will discuss means for controlling the spatial extent and spectral shape of the X-ray beam.

The production of scatter is reduced by limiting the spatial extent of the X-ray beam by means of a device called a collimator on the source side. The conventional radiograph is produced using a two-dimensional, four-sided cone of radiation. Minimizing the projected beam size by restricting it to the area of interest reduces both patient dose and image degradation due to scatter. Most systems are equipped with adjustable source collimators. Specialized equipment may be found that uses fixed collimation. Modern collimators, both fixed and adjustable, use several sets of apertures to minimize off-focal radiation. X-ray photons emerging from the tube whose real or virtual source is not the focal spot are termed off-focal radiation. Off-focal radiation is produced by electrons scattering from the focal spot with sufficient energy to produce X rays when they collide with some other portion of the anode structure. The use of metallic or partially metallic tubes, in place of all-glass tubes, reduces the production of off-focal radiation by providing alternative return conduction paths for scattered electrons. X-ray photons produced in the focal spot or elsewhere in the tube may be internally scattered and seem to be produced anywhere in the tube. Simple, single-plane diaphragms limit the size of the X-ray beam emerging from the beam port, but they are not very efficient in minimizing the emission of off-focal radiation and internal scatter. Some tubes also include a hood around the anode to absorb off-focal radiation and prevent it from exiting the tube.

In the United States, the law requires that the collimator and the image receptor must be linked so that the maximum settable beam size does not exceed the dimensions of the receptor. The influence of scatter can be further reduced by collimating to and using a one-dimensional fan

of radiation or a pencil beam of radiation, provided that the scatter geometrically misses the image receptor or is attenuated before it reaches the receptor. The former method is accomplished by having a small scanning detector aligned with the useful beam (e.g., in slot scanning). The latter technique involves the use of a secondary collimator between the patient and image receptor that scans synchronously with the X-ray beam. Slot scanning systems have recently become commercially available. These systems use an X-ray fan beam synchronized with a moving collimated slot of detectors to perform a planar patient scan. The effect is to provide a digital radiograph with significantly reduced scatter fractions (relative to grid systems) compared with conventional commercial cone beam systems. One drawback is a slightly longer scan with potentially increased patient motion artifacts, as well as other possible electrical or mechanical artifacts due to the use of moving parts.

The energy spectrum of the X-ray beam emerging from the target is modified (filtered) by its passage through extrinsic filtration before it enters the patient. The X-ray spectrum is further modified as it passes through the patient. These spectral modulations are due to the inherent energy dependence of the X-ray attenuation coefficients, which are intrinsic characteristics of an elemental or compound material.

An improper selection of the target material or filter, relative to the examination being performed, can result either in diminished image quality or excessive patient dose. A certain amount of filtration is beneficial to the patient. Low-energy photons that enter and are absorbed by the patient's tissues contribute to dose without contributing to the image. These photons may be removed from the X-ray beam by inserting additional filters between the tube and the patient. High-purity aluminum is the filter material that is commonly used for general radiography. Because the attenuation coefficient of aluminum decreases across the radiographic energy range, aluminum preferentially absorbs the low-energy photons emitted from the target and is therefore said to "harden" the X-ray beam.

Beam quality is usually described in terms of its half-value layer (HVL) of aluminum. This is the amount of aluminum needed to reduce the beam intensity by 50%. The U.S. Public Health Service has placed minimum HVL requirements on diagnostic X-ray beams. Too high an HVL is also undesirable as it leads to poor SNR in screen-film imaging systems. Alternatively, the molybdenum filter used for mammography functions in a different manner. As pure metals are relatively transparent to their own characteristic radiations, a molybdenum filter passes most of the K-characteristic radiation from a molybdenum target. At an appropriate tube voltage, the filter preferentially removes both low- and high-energy photons. Such a matched filter-target combination results in the transmission of a narrow energy band of X rays.

Several other methods are available for generation of narrow X-ray spectra with an inherent improved dose efficiency when the mean energy is matched to the object and imaging task. Use of heavy (i.e., 7–10 HVLs) K-edge filtration (Fig. 5) can provide a practical, low-complexity, and relatively inexpensive means of using heavy attenuation for beam hardening and near-complete elimination of lower energy photons and using the K-edge to nearly completely eliminate higher energy photons, thereby producing a narrow, quasi-monochromatic beam. Higher tube exposures are necessary, accompanied by increased tube heating and reduced tube life, although the actual object exposure may be similar to that with regular filtration. Bragg diffraction can also be used to produce quasi-monochromatic beams. Although Bragg diffraction can produce a very narrow beam, devices using Bragg diffraction suffer from longer scan times because only narrow fan beams can be obtained; indeed, the object needs to be moved relative to the beam because sweeping the beam would compromise beam quality. Synchrotron light sources have also been proposed for monochromatic radiation but so far have proved far too expensive and impractical for routine medical imaging applications.

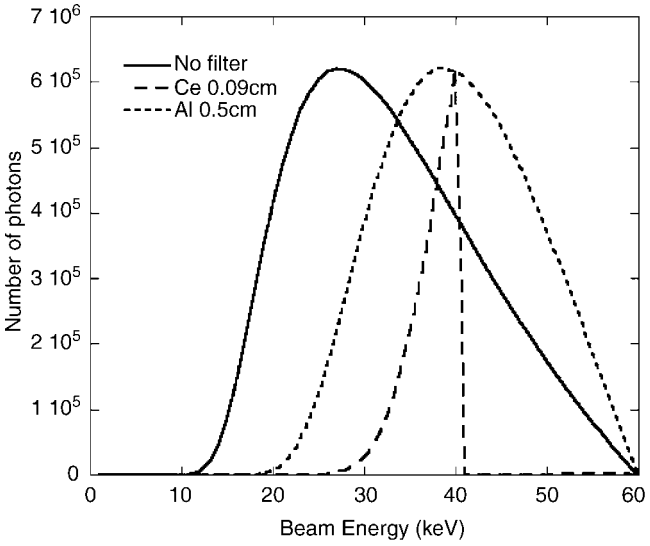


Figure 5. Sample spectra generated via simulation using 60 kVp tube potential, 1 mAs exposure with various extrinsic filters to illustrate effect on beam shape. Illustrated are spectra for no filtration, 0.5 cm thick aluminum, and 0.09 cm thick cerium extrinsic filtration. These represent unfiltered, minimally filtered (similar to standard filtration in commercial systems), and heavily K-edge filtered beams. Note that the cerium filtered beam is quasi-monochromatic (i.e., 7% full width at half the maximum height), which can, by design, provide maximal dose efficiency in a uniformly attenuating, homogeneous object.

Bowtie filters can be employed within the collimator attached to X-ray tubes (Fig. 6, top). These filters conform to the general contour of the patient in that they are thinnest (i.e., allow maximum flux) at the center where the patient is thickest and become thicker near the edge (i.e., allow minimum flux) where the patient contour gets thinner. This allows reduction of dose to the patient while providing a more uniform optical density (in the case of screen-film) or SNR (in the case of digital detection) across the image.

The interactions of X rays with the patient's tissues and other structures in the beam path result in the production, by the Compton process, of copious quantities of scattered radiation. A portion of this scattered radiation is directed

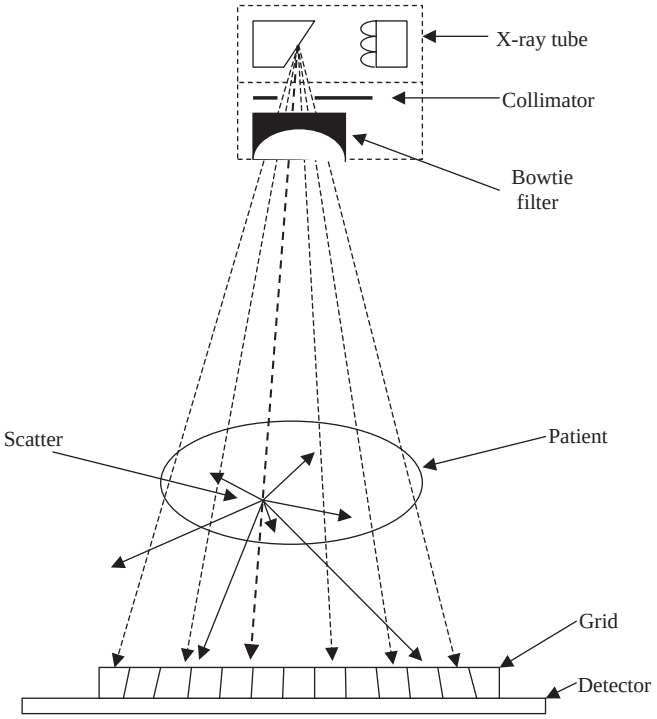


Figure 6. Schematic of the radiation trajectories and various components of interaction in the X-ray generation and detection system. Near the source is the source collimation apparatus. Distal to the source, the radiographic grid consists of thin lead strips, usually aligned with the focal length of the X-ray tube. Useful imaging photons "see" narrow grid strips and wide interspaces. Scattered photons see wider lead strips; hence, they are more likely to be absorbed by the grid.

toward the image receptor. The scattering process destroys the imaging geometry, which consists of straight lines from the focal spot to the image receptor. Scatter does not contribute to the useful imaging process, but because of its diffuse nature, scatter produces a generally uniform background intensity, reducing radiologic contrast, and decreasing the SNR of the image. In chest, for example, surprisingly most (80–90%) X rays on film are scattered radiation, although the image quality of a chest radiograph is still good and full of fine detail.

The use of a grid (Fig. 6) exploits the lack of geometrical coherence of scatter and helps to remove it. The lead strips of the grid are geometrically aligned toward the focal spot of the X-ray tube. Unscattered photons "see" only narrow grid strips and relatively wide grid interstitial spaces. Only a small fraction of these photons (the photons useful in the imaging process) are absorbed by the grid. Scattered photons produced in the patient have a wider angular distribution. On average these scattered photons see much wider grid strips and much narrower interstitial spaces. Thus, a grid selectively removes scattered radiation.

When a grid is used, however, a comparatively higher skin dose is required for the same SNR. The improvement in image quality achieved with a grid usually justifies the increased patient dose. Grids are constructed with thin lead or tungsten strips, and interstitial spaces are composed of a variety of uniformly thick materials including aluminum, organic material (solid or foam), or air. Grids are available in many different aspect ratios (the height of the strips divided by the space between adjacent strips). A higher ratio grid will reject more scattered radiation (necessitating higher incident flux to maintain SNR) but requires more precise positioning than a lower ratio device. Grids, especially coarser ones, can also be moved during the exposure so that the grid lines are blurred, and essentially eliminated by the motion. The combination of a grid and its movement mechanism is referred to as a Potter-Bucky diaphragm. Grids are now available with varying focal lengths and material compositions and can be constructed by extrusion, layering, or chemical etching.

One functional drawback to the grid is that the solid geometry (of the lines of sight) is fixed, so the source to image (SID) cannot be changed without changing the grid. Several alternatives to the anti-scatter grid are available to the designer for dealing with scatter, some or all of which can be used in combination. These include the use of air gaps between the object and imaging device, the slot scanning technique, and algorithmic correction (including the beam stop and deconvolution methods). Employment of algorithmic correction is especially attractive in digital systems.

Image Receptors

Although digital detector usage is increasing greatly, conventional film and film-screen systems are still the most widely used of all radiologic image receptors (see SCREEN-FILM SYSTEMS). The radiographic cassette contains a relatively high density and high atomic number scintillating screen for converting X-ray energy to visible light. Visible light provides better film exposure than the direct interaction of X rays in the silver halides embedded in film. When irradiated, the screen fluoresces and exposes the film to light. Most film blackening results from this fluorescent light. A screen is held in close contact with the film in a light tight cassette so that there is an optimal transfer of the light to the film, with minimal resolution blur caused by the bloom of light generated in the scintillator. Modern film-screen systems have a detective quantum efficiency (DQE) around 50% (DQE is a measure of the ability to convert SNR presented to the film into recorded SNR and includes metrics of resolution, noise, and efficiency). This means that about half of the ultimate information in the modulated X-ray beam is detected by the system. The sensitivity (speed) of film-screen systems may be chosen by selection of screen and film types. Systems that provide proper film blackening at low dose either exhibit a great deal of quantum noise or mask the noise by blurring both the noise and the image detail. This is determined by selection of screen type. Scintillators in the screen can be amorphous (the majority of cases) or highly structured in finely grown pillars. The shape of the gray scale transfer function of the film-screen system is determined by the choice of film. Ideally, the optimum visible absorption efficiency of the film corresponds to the optimum light wavelength output of the screen, although each of these absorption and emission functions vary with wavelength. Different film characteristics are often needed for the

optimal performance of different radiographic examinations. For example, a slow film speed film with finer spatial resolution may be more useful for extremity imaging of fine bone fractures, whereas a faster speed film with coarser resolution is more useful for chest imaging where stopping or minimizing the influence of cardiac motion is desired.

There are both advantages and disadvantages to the development and use of digital X-ray detectors, but a primary advantage is that there is a decoupling of the image capture and image display process. In addition to separating image capture from image display, the digital representation of the radiographic image is key to the successful implementation of a picture archiving and communications system (PACS) system, which has already shown to have a beneficial impact on the delivery of health care. With advances in large-scale miniaturization and packaging, and production of large-area electronics structures with high yield, there are continual advances in production and manufacture of digital detectors with increasingly finer spatial resolution. Indeed, there is a projected 600% growth in demand of digital radiography and digital mammography systems by 2008, along with declining demand for screen-film and computed radiography (CR) systems sales. Digital systems are expected to surpass other film-screen based X-ray imaging systems in the near future. Nevertheless, a large number of these systems currently exist and are part of contemporary radiography.

Photostimulable phosphor systems were introduced several years ago and remain popular in radiographic practice. In a CR system, an image storage plate "traps" an image of the attenuated X-ray distribution with excited state electrons. When such a plate is scanned with a pencil-tip-sized laser beam, the electrons are de-excited and return to their ground state with the emission of visible, phosphorescent radiation. This phosphorescence can be measured, digitized, and subsequently used to produce an image (fluorescent scintillation systems and phosphorescent systems differ in the decay time τ of the scintillation event, where fluorescence occurs in $\tau \sim 10^{-9}$ s, and phosphorescence can be considerably longer, lasting $\tau \sim$hours or days). Such CR systems are capable of reproducing >1; 5 line per millimeter with an acceptable DQE. The phosphors used in these systems have a linear dose-to-photoluminescence response extending over several orders of magnitude of absorbed dose. With proper reading technique, this intrinsic linearity permits a decoupling of dose from film blackening. Intentional or inadvertent exposure errors can be compensated for by the reading electronics, thus reducing retakes. Dose may be selected with regard to the image quality requirements of the individual examination.

The X-ray-sensitive image intensifier, using charge coupled devices (CCDs), is currently the universal image receptor for fluoroscopy and fluorography (see IMAGE INTENSIFIERS). This device converts the modulated X-ray beam into a minified visible light image projected onto a small, fine resolution (cooled) CCD. The conversion factor (image blackening per unit of X-ray input) of modern image intensifiers is high enough that the quantum sink in most imaging chains is the X-ray intensity detected and photoelectrons created at the tube's input screen. The DQE of these detectors is also in the neighborhood of 50%. Most image intensifier-based image receptor systems use optical diaphragms between the output of the image intensifier tube and the input of the CCD. This diaphragm is used to restrict the light intensity reaching the CCD, thereby forcing enough X-ray intensity through the patient so as to produce a statistically meaningful image. In some systems, this diaphragm is externally adjustable, permitting the operator to set the dose to a level that is consistent with the allowable X-ray quantum noise for the examination.

Image intensifier systems commonly use video as the primary fluoroscopic display and recording medium. If the video is used only for fluoroscopic monitoring or for simple playback applications, normal broadcast video standards are employed. For systems that use the video signal for fluorographic applications, such as digital subtraction angiography (DSA) (see DIGITAL ANGIOGRAPHY), better quality nonstandard video formats may be used. A variety of photofluorographic cameras can be used for image capture and the production of hardcopy images (see CINE AND SPOT FILM CAMERAS).

Flat panel detectors are becoming increasingly popular for digital radiography (DR). Although they are more expensive than CR systems and may not be able to be retrofitted into existing gantry's, they provide much faster (i.e., instant) digital data acquisition and higher resolution (now on the order 50 μm or 20 lp/mm) than CR systems. DR detectors are generally divided into indirect and direct detector categories. With indirect detection, a scintillator (e.g., CsI, BGO, Gd_2O_2S, $CdWO_4$) is used to convert X rays into photons in the visible spectrum that are then detected by a thin-film transistor (TFT)-based photodiode array. With direct detection, X rays are detected and converted directly into electric charge by the detection layer (e.g., amorphous-Se, CdZnTe, amorphous-HgI_2) without the need for an intermediate scintillation event. There is an ongoing debate pertaining to the superiority of one approach versus the other. Suffice to say here that direct detectors can have finer spatial resolution leading to a better MTF, and until very recently, indirect detectors with higher overall stopping power had better detection efficiency and hence DQE.

With digital detectors of any kind, one must be concerned with image lag or ghosting characteristics that are specific to the type of detector chosen. Acquisition and/or correction techniques may have to be employed to minimize these residual image effects and generally do not affect image quality for planar projection images in the clinical setting. Tomographic systems that require multiple rapid image acquisitions are more likely to suffer from these degrading effects.

Control and Supervisory Logic

Radiological equipment produces X rays with tube voltages of 25 to 150 kV at power levels extending from less than 100 W to more than 100 kW. Exposure times range from a millisecond to several minutes. A wide voltage range is required to produce the required signal characteristics in the final image. Wide power and exposure time ranges are needed to deliver the appropriate dose rate and total dose

to the image receptor. In addition, digital detection systems require careful synchronization such that frames are not exposed during the frame readout period. High-speed frame-grabbing hardware is also required in specialized applications, such as tomography, where frame rates can be very high. Control, synchronization, regulation, and automation of such elements necessitate the introduction of control and feedback elements into X-ray systems (Fig. 1). In many cases, installed radiographic systems have their control primarily associated with the generator, especially in screen-film systems. For newer digital systems, including special applications such as tomography, the generator becomes a component under the logical control of the overall system host computer, which is responsible for controlling and managing all components of the system.

One example of such synchronization and control is illustrated by a timing diagram (Fig. 7). The diagram shows multiple exposure cycles as would be typical in a tomography application, using a digital detector, rotating anode X-ray tube, and movable gantry, although other implementations may exist. A signal must first be sent to the generator starter to initiate rotation of the anode. Once the anode has reached maximum speed, the detector is checked for status availability. Exposures can be initiated only if the detector is not in the process of reading out the panel. The period between readouts is known as the vertical blanking period. X rays can be exposed during this time (the length of this period depends on the frame rate of the detector). When the detector ready signal is verified, a signal is sent to the X-ray generator to expose X rays. After the X-ray pulse is complete, a signal is sent to the gantry to move the assembly into position for the next image, and simultaneously, a signal is sent to the frame grabbing circuit to acquire the image from the detector controller. Frame grabbers are specialized circuits responsible for quickly obtaining or "grabbing" the exposed frame from the digital detector.

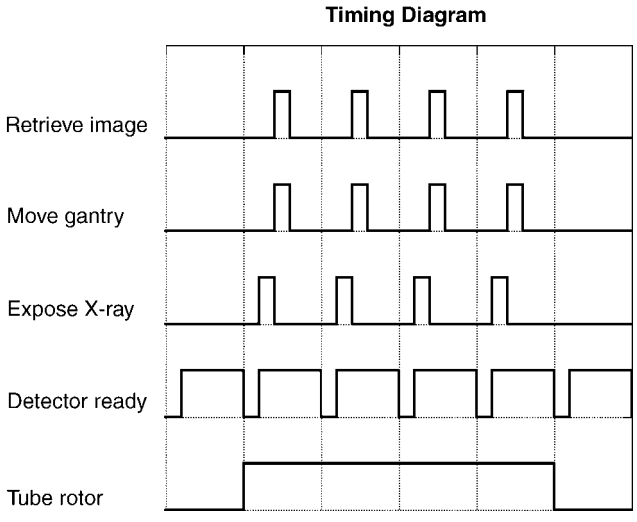


Figure 7. Timing diagram illustrating the necessity for synchronization when implementing an automated system that involves many different system components, including X-ray generator, digital detector, motion control devices, and frame grabbing devices.

The previous scenario gives maximal flexibility to the system designer. Many other options are also available. For example, the detector ready signal could be wired directly to the generator so that exposures are initiated automatically whenever the detector goes into a ready state. In addition, some tomography applications have the X-ray beam constantly on during the multiple acquisitions for tomography and, instead, use a timed, physical shutter for initiation of exposures.

The above scenario also assumes that exposure times are predetermined and will take place at the preset values when an expose signal is initiated. Another control option available on many systems is that of automatic exposure control (AEC). This involves the measurement of X-ray transmission and is used as a control element for generator switching. The commonly used feedback loop, which switches the X-ray beam at a predetermined dose, remains essentially the same as Morgan's original phototimer. Most AEC systems offer the choice of several different measuring fields, with the selection of active fields being examination and projection dependent. With proper positioning, this ensures proper exposure of critical anatomical detail. Modern systems adjust sensitivities within the AEC system to accommodate different film screen systems, compensate for reciprocity loss in radiographic film, select different measuring fields for different examinations and for other control functions. This provides a high degree of uniformity and image consistency via the automatic termination of an exposure for a predetermined desired effect. So the overall hope is that with AEC, the number of reshoots and/or call backs for additional exposures is reduced. The most common transmission measurement device used is an ionization chamber interfaced to the phototiming circuit.

The radiographic technique is the selection of appropriate X-ray exposure factors, taking into consideration patient size, shape, and physical condition, the examination and projection to be performed, and the available choice of radiological materials and supplies. This may, of course, all be done properly by a technologist using his or her own judgment with a manually controlled generator. However, this process has become less a matter of technologist preference and more a part of a departmental standard protocol, both in terms of required radiographic projections and technique factors. Many modern systems also provide anatomical programmed radiography (APR). APR typically provides the technologist with dozens of examination views for each of several different anatomical regions in a menu at the controller. By selecting a particular mode, APR automatically sets the required technique factors, including kVp, mA, time, focal spot, film speed, SID, and imaging receptor. In addition, the operator may input the actual patient thickness providing the optimal technique parameters for that examination.

Patient Support Structures

Patient safety, handling, and comfort are critical elements in the design of medical imaging equipment. The needs of the physician or technologist must also be taken into account. These are not trivial design problems. The engineer must consider, among other things, the patient's

physical condition, attached life support systems, required accessory devices (i.e., traction slings or intravenous bottles), and the patient's mental attitude. In addition, the ergonomics of the equipment should promote the expeditious performance of the examination. For example, the examination table should permit easy patient transfer with a minimum amount of staff assistance. The key component of each radiographic examination is the image receptor. To this end, the technologist must have direct and unimpeded access to the patient from a variety of directions. Nearly all of the technologist's tasks, with the exception of patient handling and positioning, can be automated. The technologist's ability to interact with a sick patient while obtaining the desired imaging results is a fine art. Equipment design and selection must be conducive to this process.

As many rooms often contain Potter–Bucky diaphragms (they are referred to as table and wall Buckys, respectively), the room must be designed to carry such devices. A variant on the radiographic room is the use of a tilting table in place of the flat (nontilting) table. The X-ray tube is usually mounted on a ceiling suspension. A special-purpose apparatus may be used to improve the efficacy of positioning for complex examinations such as skull radiography and trauma-related procedures. C-arms may also be used and must be accommodated.

Another major requirement in room design is that of safety. Rooms must be shielded to prevent radiation from reaching unintended areas while allowing visibility between the operator, who is generally outside the permanent room or behind some radiation shielding, and the patient and scanner. Mobile leaded shields with various sized lead-doped glass are available, and some dedicated systems (e.g., in mammography) have attached operator shielding to minimize exposure from the multiple studies coordinated by the technologist. Thus, rooms must be designed to accommodate both the general requirements common to radiographic systems as well as special requirements for custom applications.

Custom Applications and Equipment Selection

The design of a radiology department in terms of the number and type of procedure rooms and the selection of equipment is an important consideration. The amount to which equipment can be differentiated depends on the size of the institution and the nature of its workload. It is a rare department that can maintain an even patient flow throughout the day. Most departments have a sharp peak early in the morning and a secondary peak in the early afternoon. Sufficient staff and equipment must be provided to handle these peaks.

Radiologic equipment is classified by its technical composition and its clinical function. The simplest apparatus is the dental, mobile, or portable generator. This consists only of the means for X-ray generation and beam control. Patient support, grids, image receptors, and so on, are not included.

In fluoroscopy, the rooms are almost universally equipped with radiographic components, a tilting table, an under-table X-ray tube, and an image intensifier. The under-table tube and image intensifier are linked together and mounted on a spot film tower. The spot film device, located between the patient and the image intensifier, is a specialized cassette holder. Radiographic spot films, obtained during fluoroscopic examinations, are captured using the spot film device and conventional cassettes. The radiologist usually views the fluoroscopic image by means of closed-circuit television or on a computer screen. Fluorographic images are recorded using camera systems directly focused on the output of the image intensifier. The increasing use of fluoroscopic cameras makes the conventional spot film device obsolete.

Angiographic and other interventional or special-procedure laboratories are highly specialized rooms with film changers and cameras adapted to the needs of rapid imaging. Special procedures require especially easy access to the patient by the staff as well as the capability for multiple projection angles of the X-ray beam. The patient's medical condition and any attached life support devices dictate that these examinations be performed with as little patient handling as possible. In addition, the equipment should provide adequate radiation protection for the staff. The patient table and other mechanical equipment in these rooms are therefore highly specialized to meet the needs of different examinations. Under many circumstances, the apparatus permits simultaneous radiography from two directions (i.e., biplane imaging).

Tomographic imaging is another mechanically complex procedure found in radiographic rooms. In a historical tomographic apparatus, the film and X-ray tube were synchronously moved so that overlapping structures in the patient are removed. Tomographic apparatus ranged from attachments to simple Bucky tables to dedicated equipment capable of producing a wide variety of motions. Contemporary computed tomography is still mechanically complex, although manufacturers have gone to great lengths to enable easy patient and technologist access in a clutter-free system (see COMPUTED TOMOGRAPHY).

In radiotherapy, very high (MeV) tube potentials are used for radiation of tumors. These systems may include a separate but integrated keV imaging system for scouting the area for treatment planning before irradiation. MeV imaging systems are also being investigated for their ability to produce their own image without the need for a separate keV system. In the past, direct film detection was used because the incident energy and flux was high enough to sufficiently expose film, but newer flat panel digital devices are rapidly being adapted for this portal imaging, MeV application. Designers must take these into consideration.

Several other emerging areas including tomosynthesis, small animal imaging, computed mammotomography, and other dedicated or application-specific imaging technologies are developing rapidly. Thus, it behooves the designer to keep in mind potential future applications when considering equipment selection options. The primary changes have been rapidly occurring in imaging (i.e., detector) technologies, in moving from analog to digital systems, and the Internet technology associated with their image processing, sharing, and transfer. There are several initiatives to modify the X-ray sources for both improved focal spot characteristics (e.g., smaller, faster, switching capabilities) and target and filtration materials (i.e., for more optimal spectra used

in application specific imaging paradigms). To most appropriately synchronize and use these emerging, digitally based technologies, computer-based, centralized data control and acquisition technologies need to keep pace with these other emerging technologies. Only through properly using and synchronizing these various necessary X-ray system components can the lowest ionizing radiation doses be applied to obtain images that provide the most information content available to the clinician, whose proper interpretation will affect the patient under observation.

FURTHER READING

Curry III TS, Dowdey JE, Murry Jr RC. Christensen's Introduction to the Physics of Diagnostic Radiology. 3rd ed. Philadelphia, PA: Lea & Febiger; 1984. A general overview of the physics and technology of diagnostic radiology. Many U.S. radiology residents use this book as their primary textbook and technical reference.

Hendee WR, Chaney EL, Rossi RP. Radiologic Physics, Equipment and Quality Control. Chicago, IL: Year Book Med. Publ.; 1977. A textbook for advanced technologists.

Johns HE, Cunningham JR. The Physics of Radiology. 4th ed. Springfield, IL: Thomas Publ.; 1983.

Beutel J, Knudel HL, Van Metter RL. Handbook of Medical Imaging, Volume 1: Physics and Psychophysics. Bellingham, WA: SPIE Press; 2000.

See also X-RAYS, PRODUCTION OF; X-RAY THERAPY EQUIPMENT, LOW AND MEDIUM ENERGY.

X-RAY MAMMOGRAPHY. See MAMMOGRAPHY.

X-RAY QUALITY CONTROL PROGRAM

EDWARD L. NICKOLOFF
Columbia University
New York City, New York

INTRODUCTION

In order to improve or maintain an optimal achievable level of healthcare, most medical facilities usually have some type of quality assurance program. The quality assurance programs encompass a variety of functions, such as development of comprehensive safety rules, establishment of infection control guidelines, provisions for the efficient management of critically ill patients, establishment of accurate record-keeping functions, and the development of adequate educational training for the staff (1). In radiology, a portion of the overall quality assurance program is devoted to a quality control program for the X-ray equipment. The American College of Radiology (ACR) defines quality control (QC) programs as a periodic monitoring of aspects of the precision or accuracy of the X-ray equipment pertaining to the successful performance of the equipment, techniques, or tests rather than to the clinical decision making (2). In general, QC involves equipment rather than patient care.

Most X-ray QC programs have several very specific goals toward which they are directed. The list merely enumerates these items. (1) Image quality, (2) patient radiation doses, (3) safety, (4) consistency, (5) economic factors, (6) Regulatory requirements.

The type of QC testing performed is often dependent on the particular goals that are being evaluated. Usually, state or local regulatory agencies require that a QC program for X-ray equipment be conducted on a routine basis. Some regulatory agencies have very specific items that must be measured by a qualified expert or licensed medical physicist at least annually. These items may include typical patient doses (doserate) and generator calibrations. Moreover, the Federal government has published many documents in which suggested survey procedures for X-ray equipment are discussed (3–5). In 1992, the Federal government enacted the Mammography Quality Standards Act (MQSA), which requires: the accreditation of all mammography units, U.S. Food and Drug Administration (FDA) certification of mammography facilities, and at least annually physics QC testing of mammography equipment. In addition, the Joint Commission on the Accreditation of Healthcare Organizations (JCAHO) mandates an adequate QC program that has qualified staff and maintains patient safety. Hence, QC programs for X-ray equipment are mandated by various regulatory agencies and accreditation organizations like the American College of Radiology (ACR) and JCAHO.

Some of the regulatory requirements, however, do not have a major impact upon image quality or economic factors. For example, one regulatory restriction requires that X-ray and light field be congruent with each other to within 2% of the source to image receptor distance (between the X-ray tube focal spot and the image receptor) (SID). Another regulatory requirement is directed toward checking the various interlocks in the X-ray system. Although these regulatory requirements affect radiation protection, they have little impact upon the operation of the equipment or upon image quality. Hence, some of the various QC tests mandated by regulatory agencies, and the results must comply with legal standards of performance.

Another reason for performing QC tests is to assess mechanical, electrical, and radiation safety of the X-ray equipment. These evaluations include inspections directed toward eliminating any electrical shock hazards or the failure of cables, locks, mechanical supports, or counterbalances that could result in the injury to patients or staff The mechanical inspections require that the system components be both visually inspected and the components be manipulated to test their integrity. Radiation levels to the patients and staff must be measured to determine that there is adequate radiation safety and that radiation levels are not excessive.

Another feature of QC programs is to determine the functional performance of an X-ray system. Various controls, switches, and function buttons must be tested to make certain that they are operating properly. These tests are directed toward making certain that the equipment functions as it was designed. A common malfunction of X-ray equipment is a selector switch that indicates some change has been made when in actuality nothing happens.

The optimization of the image quality in radiology procedures is one of the most important reasons for QC programs. The QC programs establish a test procedure by which the image quality can be measured and establish a standard level of image quality one attempts to maintain (6). Thereafter, the image quality can be measured on a regular basis and corrective actions are taken as necessary. Degraded image quality can often hamper the proper clinical diagnosis and thereby have a severe negative impact on patient care. One would not like to deliver radiation doses to the patients that result in limited or no diagnostic information. Hence, a major goal of QC programs in diagnostic radiology departments would be the assessment of image quality.

Quality control programs must also attempt to maintain a consistent level of equipment performance. X-ray outputs that are not reproducible from one exposure to the next will result in over- and under-exposed images. Unusable radiographs result in unnecessary patient radiation dosages, delays in diagnosis and treatment, and economic losses in terms of technologist's and physicians time and film. Similarly, the technologists should be able to expect that adjacent rooms with similar X-ray equipment will have similar X-ray outputs and similar calibrations within reasonable tolerances attributed to differences in the models and manufacturer's designs of the equipment. Automated film processors should be adjusted to produce similar radiographs regardless of where the films are developed. Similarly, digital X-ray imaging like Computed Radiography (CR) and Digital Radiography (DR) must have consistent X-ray exposures, image processing, and display of images. Consistency of a given X-ray room and consistency in comparing the various X-ray equipment is essential to maintaining a standard level of radiographic image quality for the entire facility.

Finally, economic considerations are a major impetus for having an effective QC program. It has been estimated that each radiographic film that must be repeated costs the facility ~$35 (1986 prices) (7). The cost of repeated films include items such as film prices, X-ray tube usage, X-ray equipment depreciation, chemistry usage, processor depreciation, technologist's and physician's time, and facilities cost (rent, heat, and electricity). A study performed by the federal government indicated that repeat rates can often be decreased by a factor of two by effective QC programs (8,9). Without QC programs, repeat rates can be expected to be 10–16%; with QC programs, repeat rates can be expected to be 5–8%. These repeat rates are due to patient motion, positioning errors, or other technologists' errors. Repeat rates can often be drastically reduced by the usage of charts in which the technical factors to be selected are listed for a variety of patient procedures and patient sizes. The development of these technique charts should be considered to be part of a good QC program. Furthermore, a good QC program often identifies and corrects equipment malfunctions before they become serious and result in excessive downtime of the equipment. Thus, a good QC program usually provides cost savings to a X-ray facility that can offset the cost of the program (10,11).

In summary, there are a number of good reasons for the establishment of a QC program for a medical X-ray facility. First, a number of regulatory agencies mandate the need for a QC program. Second, a QC program is necessary to ensure adequate mechanical, electrical, and radiation safety. The QC program may decrease the legal problems that can be encountered from various malpractice suits. In addition, the QC program should result in a good and consistent quality of the diagnostic information obtained from the radiological procedures. Finally, an economic benefit should be realized in terms of reduced repeat rates and diminished equipment downtime.

Even though the justification for, and the goals of, the QC program can be easily identified, the description of the measurement procedures for the program is difficult for a number of reasons. First, there are often several different ways by which a given measurement can be performed with varying amounts of accuracy. The methodology chosen will often depend on the kind of test equipment available and the training of the QC personnel. A rudimentary QC program that would be restrained to have limited expenditures might utilize very basic tools that could be readily operated by inexperienced QC personnel. This type of program would utilize items, such as pocket ionization chambers, step wedges, digital kVp meters, and pennies for field alignment locators (12). A more sophisticated QC program might utilize items, such as oscilloscopes, digital exposure–exposure rate detectors with waveform outputs, high voltage bleeder units for peak kilovoltage measurement, and special test stands. The level of the QC program is usually determined by the size of the X-ray facility, the complexity of the X-ray equipment, the goals of the QC program, the funding for the QC program, the training of the QC personnel, and the caliber of the support provided by the X-ray service organizations. In the material that follows, the parameters that should be measured and acceptable levels of tolerance will be indicated. Although a method for performing the measurements may be indicated, the reader should be aware that there may be several alternative means for performing the measurements. The author has not attempted to list all QC measurement modalities (13–17).

Furthermore, there is a large diversity of X-ray equipment. Each type of X-ray installation requires certain specialized tests unique to that type of installation. In general, diagnostic radiology equipment can be categorized into the following groups: (1) simple radiographic (rad units), (2) chest radiography units, (3) mammographic units, (4) body section tomographic units, (5) dental units, (6) portable radiographic units, (7) portable C-arm fluoroscopic units, (8) ultrasound units, (9) combined radiographic and fluoroscopic (R & F) units, (10) special procedure rooms (with digital and/or cine), (11) computerized axial tomographic (CT scanners), (12) magnetic resonance imaging (MRI), (13) automated film processors, (14) CR and DR units, (15) video display terminals (monitors or VDTs), (16) picture arching and communications systems (PACS).

Even though there is a diversity in the types of diagnostic X-ray installations there are some common components of all units; for example, all X-ray installations (regardless of their type) have an X-ray tube, collimator system, control panel, and an X-ray generator. If the QC program were described in terms of the types of X-ray units, the tests for the various components would be

repeated many times. Instead, it is more efficient to describe the QC testing in terms of the components that would be common to many different types of installations (e.g., the X-ray tube). Items that are unique to a given type of installation are listed separately.

Finally, the level of what is deemed as acceptable performance for the equipment can vary. Older equipment cannot be expected to perform as well as newer equipment. Equipment manufactured after August 1974 is required by Federal regulation to meet certain performance levels and have certain special features. This equipment is designated as compliant equipment. Automatic collimation to the cassette size (PBL, positive beam limitation) was previously required; although this requirement is no longer mandatory, many modern X-ray units still employ PBL (18). Equipment manufactured prior to this date is not as strictly regulated. Moreover, equipment that is well maintained and routinely calibrated should be expected to perform better than equipment not receiving this attention. Hence, performance standards must have some leeway for interpretation. The performance standards that will be quoted here are for newer compliant equipment that is properly maintained.

An evaluation of the performance of X-ray equipment should involve an understanding of the accuracy of the measurements, the condition and age of the X-ray equipment, and its usage. Having stated the various caveats and limitations, the material that follows provides a basic QC outline for various components in the X-ray system.

X-RAY TUBE

The X-ray tubes are the source of the X rays produced in the X-ray equipment. Figure 1 shows a sketch of the X-ray tube. Without exposure initiation, there is usually a standby current passing through the filament. This current is sufficient to keep that filament warm, but it should be less than that necessary for thermionic emission. There are two or more different filaments. The larger filament is used when more thermionic electrons are needed for increased X-ray production. The use of larger filaments, however, results in a larger area on the anode where the electrons collide with the target; this area, as seen from the position of the image receptor, is called the effective focal spot. Large focal spot sizes can result in the degraded image quality due to geometric blur. An important part of a QC program, therefore, is the measurement of effective focal spot size.

Just prior to making an X-ray exposure, the rotor preparation switch is depressed. This switch increases the current through the filament so that sufficient electrons are boiled off in order to yield the correct amount of electron flow (tube current) between the cathode (filament) and the anode (target). Simultaneously, the anode begins to rotate, increasing to its correct speed. The rotation is needed to distribute the heat created by electron bombardment of the target. Upon depression of the exposure switch, the selected high voltage in kilovolts (kVp) is applied between the cathode and the anode. This voltage accelerates the thermionics electrons from the filament and causes them to hit the anode with high speed. The collisions result in X-ray production. The lead lining attenuates X rays in all directions except in the region of the tube port where the lead is missing. Some of the X rays do penetrate through the lead shielding and they are called leakage radiation. This radiation must be measured to determine if it is within legal limits. At the end of the exposure, the high voltage is removed. The filament current then returns to the standby level, and electromagnetic braking is provided to slow the anode to a stop (19–21).

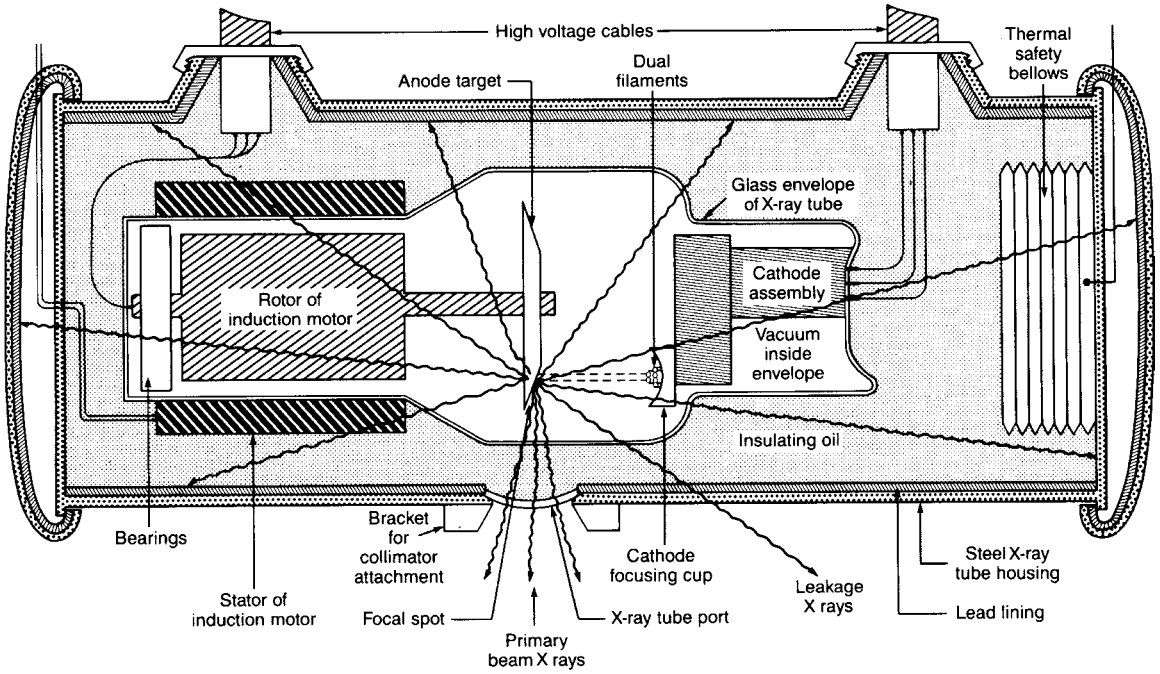


Figure 1. Diagram of an X-ray tube assembly that depicts the major components.

Several types of X-ray tube malfunctions can occur that should be assessed through QC testing. An X-ray tube that has a vacuum leak is known as a"gassy" tube. Gassy tubes have a decreased X-ray output, show spikes in the milliampere and radiation waveforms, and show decreased penetration. A tube that has been overheated repeatedly may cause metal pitting in the anode; the rough anode surface decreases X-ray output (22). The overheating of the anode can also cause metal to evaporate and desposit inside the glass envelop of the X-ray tube; this phenomenon is known as metalization. Metalization can result in electrical arcs which produce spikes in kilovolts, milliamperes, and radiation waveforms. Overheating the filament results in a thinner filament due to evaporation of the metal. This increases the resistance of the filament. Without a tube current feedback circuit, the X-ray production can be decreased. If the filament becomes too thin, the filament will break and X-ray production will cease. Another X-ray tube malfunction is destruction of the bearings upon which the anode rotates due to excessive heat and mechanical wear. As the bearings fail, the anode does not rotate at the proper speed and the heat destroys the target. Bearing problems create excessive noise during anode rotation. Improper warm-up procedures can crack the anode and cause noise during anode rotation. The QC procedures should be able to detect these malfunctions.

MECHANICAL INSPECTION

The X-ray tube housing should be inspected visually for physical damage and/or oil leakage. The alignment of the X-ray tube in the support rings (trunions) should be checked.

ELECTRICAL INSPECTION

All cables into the X-ray tube should be visually inspected for damage to the insulation. The high voltage cables should be checked to make certain that connectors are properly tightened in the X-ray tube housing.

BEARING ROTATION

There are both optical and vibration sensors that can determine the anode speeds. The typical anode speeds are 3600 rpm on low speed and 10,000 rpm on high speed. The best method for checking the bearings is to listen to the sounds made during preexposure boost and slow down. The pitch of the sound should increase for 1–2 s and stabilize as a steady, even hum. During braking, the sound should be rapid 5–10 s deceleration. A metallic clinking sound and other irregular metallic sounds are indicative of bearing failures.

CENTRAL BEAM ALIGNMENT

An X-ray tube can be misaligned in its mountings or the insert can move inside the housing. To check the central beam alignment, a Lucite block with holes drilled in it (or a special hollow plastic cylinder with a central hole on both the top and bottom surface) can be used. The device is placed on a image receptor cassette. The center of the block is positioned at the center of the central beam as shown by the collimator cross-hair. The hole that appears most circular in the film image is at the central axis of the X-ray beam. Off-axis positions have oblong images on the film. This is illustrated in the X-ray image shown in Fig. 2.

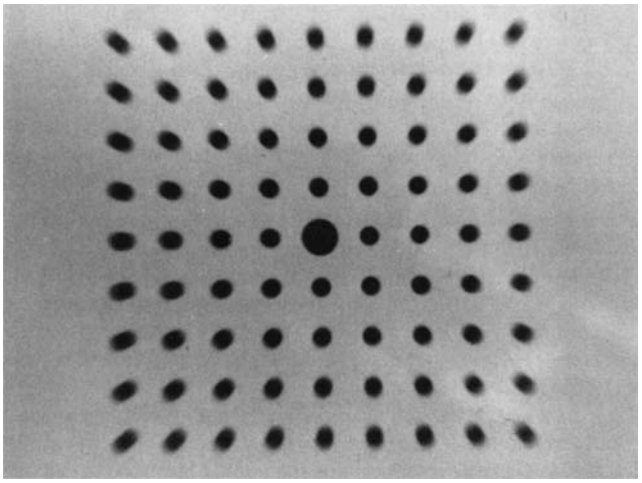


Figure 2. A radiograph of a device utilized to locate the central axis of the X-ray beam.

X-RAY OUTPUT

The X-ray output could decrease significantly for a number of different reasons. Some sources of a decreased output include excessive X-ray beam filtration, low kilovoltage calibration, low milliampere calibration, damage to the anode from overheating, a pitted anode, or gassy X-ray tube. To test the X-ray output, a radiation detector is placed at 1 m from the X-ray tube focal spot and suspended several centimeters above the table top to decrease backscatter. The output should be measured at 100 kVp with a large-focal-spot milliampere setting and exposure times $>1/10$ s. The measured air radiation dose in milligray (mGy) is divided by the mAs used to make the exposure; the measurement can then be converted to a standardized output by expressing the value as mGy per 100 mAs. For a three-phased, 12 pulse X-ray unit, or a high frequency X-ray generator, the X-ray output should be between 7–10 mGy per 100 mAs at 100 kVp and a distance of 1 m. A single-phase, full-wave rectified unit would measure $\sim$50% less. X-ray outputs that are outside these limits by more than $\pm$15% should be considered to be abnormal unless the X-ray tube has significant filtration ($>$3 mm Al equiv or $>$0.1 mm Cu).

RADIATION WAVEFORM

Irregularities in the radiation waveform can be indicative of severe problems with the X-ray tube, cables, and/or X-ray generator. Some sources of irregularities in the radiation waveform include gassy X-ray tubes, pitting of

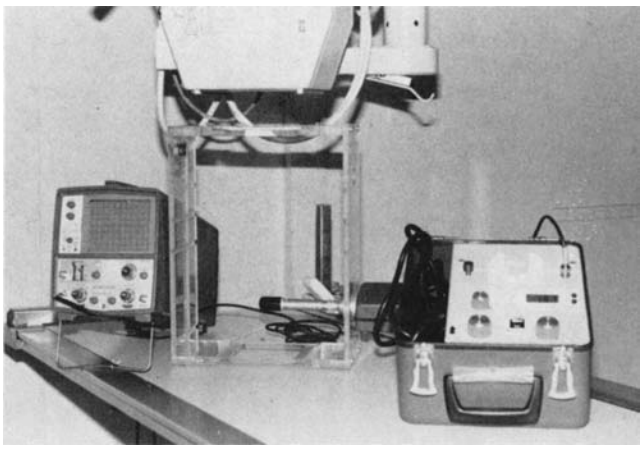


Figure 3. Equipment used to display radiation waveforms. The radiation detector has an output terminal that is connected with a coaxial cable to a storage oscilloscope in order to display the waveform.

the X-ray tube anode, arcing cable connections, malfunctioning milliampere stabilizer, phase imbalance in the power line, rectifier failure, and other generator problems. Modern radiation detectors usually have a BNC or computer output by which the radiation waveform can be monitored. A coaxial cable can be connected to a storage oscilloscope to record the radiation waveform (see Fig. 3). A three-phase, 12-pulse X-ray tube unit usually exhibits a 5–10% ripple in the wave form; high frequency units have a variable ripple (that depends on kVp and mA used) of 3–20%; a three-phase, 6 pulse unit usually exhibits 20–25% ripple; single-phase units exhibit 100% ripple. The peak values of each of the pulses in the waveform, however, should be relatively constant; the peak values should not change by >7%. The display of spikes or dropouts in the waveform are indicative of problems.

FOCAL SPOT SIZES

Focal spot measurements are usually only performed for new X-ray tubes; focal spot measurements are not routinely performed during QC: unless some spatial resolution problems are being investigated. The size of the X-ray tube focal spots can have a significant influence upon the radiographic image quality. Objectional bur can be introduced by excessively large focal spots. One way to measure focal spots is to use a leaded star pattern that is taped to the center of the collimator face plate. For focal spots >0.6 mm, a 2° star pattern is adequate; <0.6 mm focal spot sizes, a 1° pattern should be utilized. Typically 75–80 kVp should be used along with a milliampere value that is midrange for the focal spot size being measured. Focal spot sizes decrease as peak kilovoltage increases and milliamperage decreases. Imaging of the star pattern should use plain film in a paper cassette (without intensifying screens that is called direct film exposures). The magnification factor should be small enough that the outermost interference pattern should appear in the film image (see Fig. 4). In general, a magnification factor of ~1.5× should be utilized

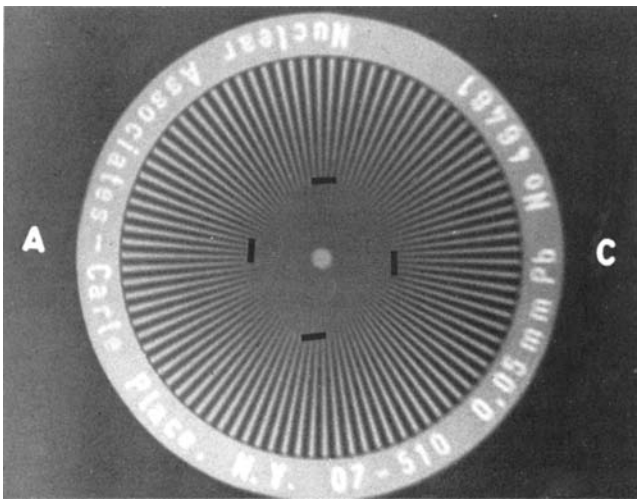


Figure 4. The radiograph of a 2° star pattern used to measure focal spot sizes. The radiograph was taken at 120 k Vp and 100 mA with a magnificaion factor of 2.9. The manufacturer's specified focal spot size was 0.60 mm. The measured focal spot was 0.83 × 0.90 mm.

with a 1.2 mm focal spot, and a magnification factor of ~2.5× should be utilized with a 0.6 mm focal spot. The distance, d, between the outermost interference patterns is measured on the film. The effective focal spot size, α, can be calculated from the following equation (23,24)

$$\alpha = \frac{\pi\theta}{180}\,\frac{d}{(m-1)}$$

where θ is the number of degrees in the star pattern and m is the magnification factor. The measured focal spot sizes should be within −0 to +50% of the manufacturer's stated specifications in the small dimension (width) and within −0 to +100% in the long (length) directions. Small focal spots (<0.4 mm) often have a −0 to + 50% tolerance in both orthogonal directions.

LEAKAGE RADIATION

Since the X rays produced in the anode of the X-ray tube are nearly isotropic in direction, the lead lining within the X-ray tube housing is intended to reduce leakage radiation to acceptable levels. In order to measure the leakage radiation, the highest peak kilovoltages utilized clinically are selected. To accommodate the relatively slow response time of many exposure rate detectors, a relatively long exposure time is required (2–6 s). Federal regulations specify that the leakage radiation be measured at several locations 1 m from the focal spot and averaged; leakage techniques are specified as the highest peak kilovoltage with the highest continuous tube current. The X-ray tube can only be operated continuously at very low tube currents. These currents are typical of those used during fluoroscopy. Unfortunately, radiographic units usually cannot be operated at low currents in a continuous manner; radiographic units are designed for pulsed operations at high currents with short exposure times. To measure leakage radiation with radiographic units, a high peak

kilovoltage and low milliamperage are selected with a several-second exposure time. The adjustable collimators are closed completely and an additional lead sheet (at least 3 mm thick) is placed to block the primary X-ray beam; in this way, only leakage radiation is measured. Readings are taken with an exposure rate meter at several locations 1 m from the other focal spot and averaged. The average value is scaled down to a value corresponding to a low milliampere continuous tube current (typically 5–10 mA). Federal regulations required that the average leakage exposure rate measured at a distance 1 m from focal spot at leakage techniques be <100 mR·h^{-1} (0.87 mGy·h^{-1}). Higher readings are indicative of problems with the lead lining inside the X-ray tube housing. Excessive leakage radiation levels are rare occurrences. Moreover, these measurements need only be performed after X-ray tube replacement; they should not be a part of the routine QC program.

COLLIMATOR ASSEMBLY

The collimator assembly has variable lead shutters that define the size of the X-ray field. The assembly also contains a light bulb and mirror that are used to project a light field identical in size and location to the X-ray field. Collimator assemblies usually contain metallic filters (usually composed of aluminum or maybe also a thin layer of copper) that remove low energy X rays that are not very penetrating. These low energy X rays do not have sufficient energy to penetrate through the patient anatomy that is being radiographed in sufficient quantity to influence the imager quality; these low energy X rays, however, can contribute significantly to the patient radiation dose. Therefore, the X-ray tube plus collimator assembly should contain sufficient metallic filtration to remove the very low energy X rays and harden the X-ray spectra. Figure 5 shows a typical collimator assembly. The QC procedures should evaluate the X-ray field–light field congruency and the penetration of the X-ray beam (19–21).

X-RAY FIELD–LIGHT FIELD ALIGNMENT

If the localizer light is misaligned with the X-ray field, the radiographic image can be too small, too large, or miscentered on the X-ray field. The alignment can be checked by placing a image receptor cassette at a given distance from the focal spot. The typical SID utilized in diagnostic radiology is 1 m. With the light localizer turned on and pointed at the cassette, metal markers are used to delineate the periphery of the light field; pennies are often used for the markers. A radiographic exposure of the film is made. The misalignment of the metal markers with the edge of the X-ray field is determined from the developed image. Figure 6 is an example of a misaligned X-ray–light field. Federal regulations specify that the misalignment in any one orthogonal direction should not exceed 2% of the SID. For automatic collimation (PBL), the error in any one direction must be $<3\%$ and the sum of the errors in two orthogonal directions must be $<4\%$.

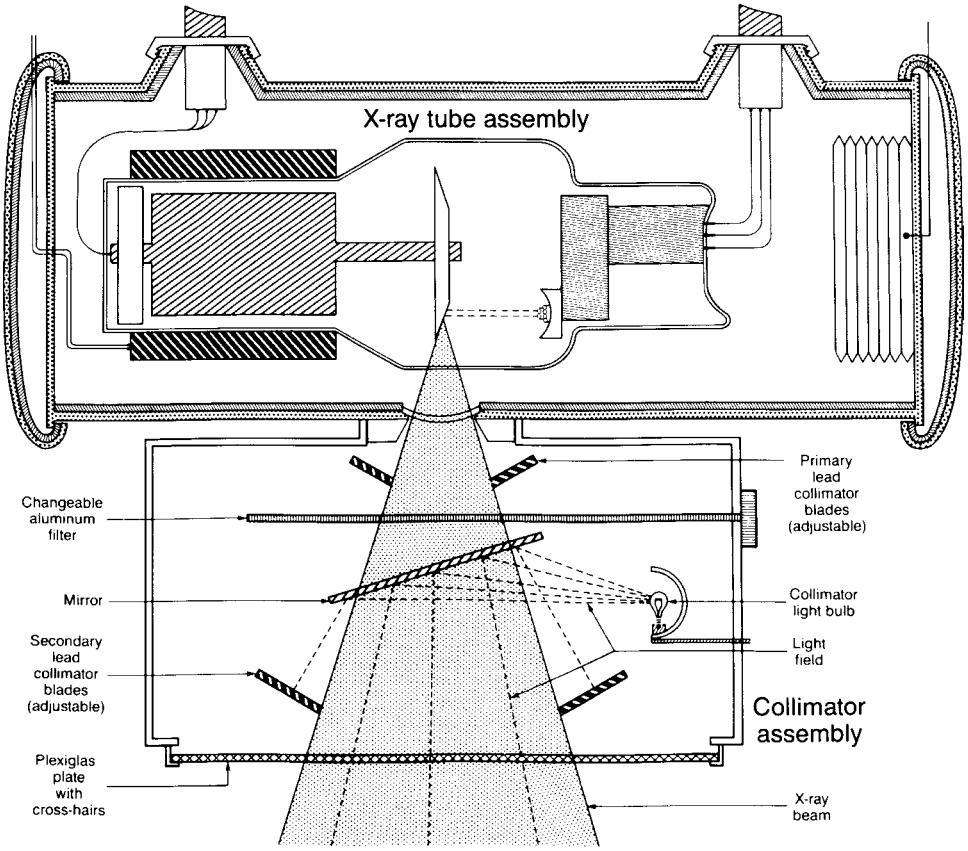


Figure 5. Diagram of an X-ray collimator assembly that depicts major components.

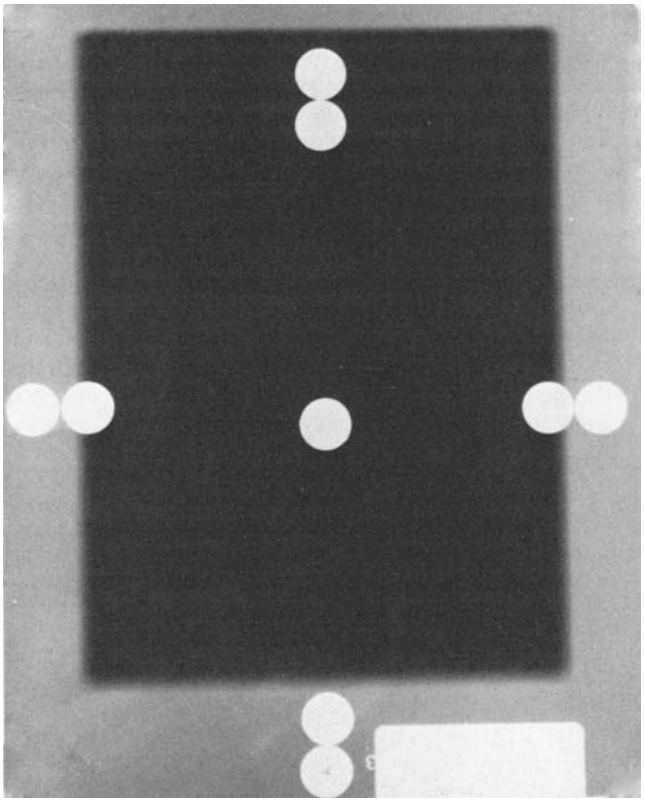


Figure 6. Radiograph of the markers used to measure X-ray–light field congruence. The fields were severely misaligned in this image.

HALF-VALUE LAYER DETERMINATION

The half-value layer (HVL) for each X-ray unit is a measurement of the penetration of the X-ray beam. The HVL is the thickness of aluminum necessary to reduce the X-ray beam intensity by 50%. The HVL is measured in order to determine whether there is sufficient filtration in the X-ray beam. Insufficient filtration will result in increased patient dose. Excessive filtration results in extra stress on the X-ray tube, a loss in radiographic contrast, and increased exposure time. Increased exposure times can result in blur due to patient motion.

The HVL is measured by placing a radiation detector in the X-ray field. Although any selection of peak kilovoltage and mAs can be chosen, usually 80 kVp at 200–400 mA and 0.10 s is selected on the X-ray control. Two exposures are taken with no added filtration in front of the detector; the average value of these two readings represents 100% transmission through the X-ray tube and collimator assembly. Sheets of 1100 aluminum (mammography uses extra high purity aluminum filters) are then placed (one at a time) between the X-ray tube and the detector. The aluminum should be placed closer to the X-ray tube, allowing at least a 30 cm air gap between the aluminum and the detector to minimize scatter effects.

After each new piece of aluminum is added, an exposure is taken and the radiation measured is recorded. When the exposure measured is <50% of the baseline value, the data are plotted on semilog paper and the HVL determined. For 80 kVp, the Federal regulations require the HVL to be >2.3 mm of aluminum equivalent (3). The following table provides the minimum acceptable HVL values for different modalities as a function of the kVp used:

Modality	Equation for Minimum HVL
Mammography	>(kVp/100) + 0.03 mm Al <(kVp/100) + C mm Al Where C = 0.12 Mo/Mo, 0.19 Mo/Rh, and 0.22 Rh/Rh
Radiograph and routine fluoro	>[2.26 (kVp/100)] + 0.48 mm Al
Angiography and cardiac fluoro	>[3.5 (kVp/100)] + 0.08 mm Al

The maximum HVL should not be >6 mm Al equiv at 80 kVp for routine fluoroscopic, and radiographic units nor >9 mm Al equivalent at 80 kVp for angiography and cardiac fluoroscopy. For CT scanners at 120 kVp, typical values of the HVL are between 6 and 10 mm Al equiv.

ILLUMINATION OF THE LIGHT LOCALIZER

It is important that the collimator light be bright enough to visualize the borders of the light field during normal light conditions in an X-ray room. Otherwise, the X-ray field would not be properly determined and collimated. A visual inspection of the collimator localizing light can be conducted under normal room lighting. If a light meter is available, the intensity of the collimator light should be measured with the room lighting turned off. Federal regulations require that the light intensity of the collimator localizer >160 lux (15 ft·C) at a distance of 100 cm.

POSITIVE BEAM LIMITATION

Some X-ray equipment has an automatic collimation system for the X-ray beam. This automatic collimation system is designated as PBL. The PBL restricts the X-ray field size to the size of the cassette placed into the bucky tray. If this sytem malfunctions, the X-ray field size could be larger than the cassette and result in unnecessary X-ray dose to the patient. A smaller field size due to malfunctions could result in a loss of diagnostic information toward the periphery of the film. If the PBL system is present, some state regulatory agencies have performance criteria for the collimation accuracy.

To test the PBL system, two identical cassettes can be used. The collimator field size is opened to its maximum extent, and one cassette is placed into the bucky tray. The localizer light is turned on in the collimator, and the tray is inserted into position. One should visually observe the light field decrease to the approximate size of the cassette. The second cassette is placed upon the table and centered in the light field. The light field should be slightly inside all boarders of the cassette on the tabletop. In fact, for a 1 m SID and an 8 cm distance between the cassette in the bucky tray,and the tabletop, the light field should be 7.5% smaller

than the cassette in each of the two orthogonal directions. A typical criteria for the PBL collimation is an accuracy of $\pm 3\%$ of the SID error in either orthonal direction and $\pm 4\%$ of the SID sum error without regard to sign for both directions.

MINIMUM SOURCE-TO-SKIN DISTANCE

In order to prevent placement of the X-ray tube too close to the patient, the federal regulations specify the minimum distance that the bottom of the collimator must be from the focal spot. This regulation usually does not apply to overhead X-ray tubes where the patient would never be placed next to the collimator under routine clinical conditions. This distance can be determined from geometric magnification effects. If an object of known size (OS) is placed at the bottom of the collimator and radiographed, the image size (IS) will be magnified. The distance from the object to the film should be measured (OFD). The source-to-skin distance (SSD) can be determined from the measurements.

$$\mathrm{SSD} = [\mathrm{OFD}(\mathrm{OS}/\mathrm{IS})]/[1 - (\mathrm{OS}/\mathrm{IS})]$$

For radiographic units, the federal regulations specify that the SSD be >30 cm (12 in.) for mobile systems. For fluoroscopic units, the Federal regulations specify that the minimum SSD should be >38 cm (15 in.) for stationary certified fluoroscopic units, 30 cm (12 in.) for mobile fluoroscopic units, 19 cm (7.5 in.) for mobile fluoroscopy c-arm with an SID $= 45$ cm that are used for extremity imaging and 20 cm (8 in.) for image intensifier fluoroscopic units used for specific surgical applications. In general practice, the minimum SSD for stationary certified fluoroscopic units is usually 45 cm (17.7 in.).

X-RAY GENERATOR

The X-ray generator controls X-ray production. The timer circuit determines the duration of the X-ray exposure. The milliampere selector determines the number of electrons per second that bombard the anode. Hence, the product of milliamperage and time determines the number of X rays produced; this product is known as mAs. kVp selector provides the high voltage between the cathode and anode that accelerates the electrons in an X-ray tube. The peak kilovoltage determines both the energy spectrum and the number of X-ray produced. As peak kilovoltage increases, the penetration of the X-ray beam increases (19–21).

In order to be able to consistently take acceptable radiographs, it is important the X-ray generators be properly calibrated. Moreover, the exposure should be reproducible and the various X-ray equipment should be calibrated to the same standard. In order to meet these criteria, the QC procedures should carefully evaluate the generator performance. Most extremity imaging and some other projections are performed using manual X-ray generator settings.

X-RAY REPRODUCIBILITY

If an X-ray generator is operating properly, the amount of radiation produced for a given selection of technique factors should be constant. If there is a problem with milliampere stabilization, peak kilovoltage waveform, arcing in the X-ray tube, and/or timer viability, the measured X-ray output at any fixed location will vary. A radiation detector should be placed in the central ray $\sim$30–70 cm from the bottom of the collimator. Five-to-ten separate X-ray exposures are recorded for some fixed selection of technique factors (e.g., 80 kVp, 300 mA, and 0.10 s). The average value and the standard deviation for these exposures is calculated. The ratio of one standard deviation divided by the average exposure is also computed. The ratio is called the coefficient of variation (CV). Federal regulations require that the CV be <0.050 for compliant X-ray units and it is desirable that this value really be <0.02 for properly functioning X-ray generators.

TIME ACCURACY

Since the exposure duration timer circuitry partly determines the mAs used, inaccuracy in the timer could result in variations in the radiographic density for film images and mottle for digital images. The duration of X-ray exposures can be easily measured by either special electronic X-ray timers or some digital exposure detectors. One merely places these devices in the X-ray beam and reads the measured exposure time. Other methods utilize the display of a radiation waveform on an oscilloscope and determining the exposure time from the width of the displayed waveform full width at half-maximum (fwhm). It is recommended that the timers be accurate to within $\pm 5\%$ or ± 1 ms, whichever is larger at all timer settings on the control panel; Single-phase equipment should only be accurate to 1 pulse or ± 8.3 ms. Instead of testing all the timer settings, however, 5–10 timer settings from the shortest to the longest should be checked. The X-ray tube chart should be consulted to avoid exposures that would overheat the tube.

TIMER REPRODUCIBILITY

One reason for the lack of exposure reproducibility is an inconsistently functioning X-ray timer circuit. The time for five identical exposures should be measured. Federal regulations require that average exposure time (T) for any fixed selection of technique factors be greater than five times the worse excursion in the exposure duration

$$[5(T_{\max} - T_{\min})]$$

MILLIAMPERAGE ACCURACY

Another factor that influences the selected mAs is the milliampere calibration. Again, inaccuracies among the milliampere stations can result in improper radiographic densities on the film. Moreover, it is desirable that various rooms in a large X-ray department have a similar calibration. Therefore, technique charts for procedures throughout the department can be standardized.

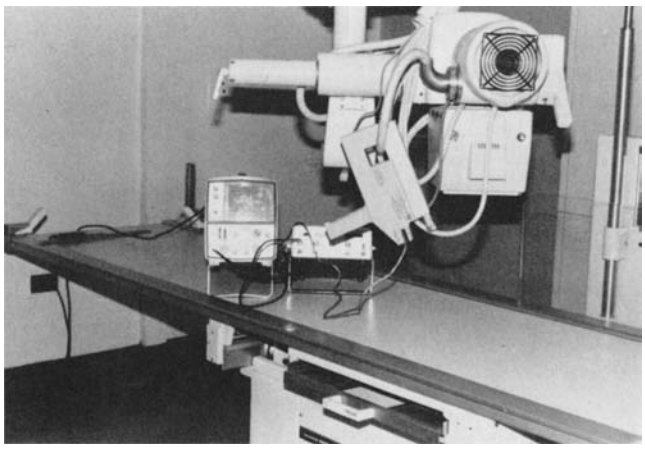


Figure 7. Equipment used to measure and display milliamperage of the generator. The noninvasive probe shown on the X-ray cable measures the milliamperage indirectly by sensing the induced magnetic fields. The storage oscilloscope is used to display the milliampere waveform.

Several different ways to measure the milliamperage accuracy are available. A simple method is the use of noninvasive current probes (25). These probes clamp around the high voltage cables to the X-ray tube and measure milliamperes indirectly by sensing the induced magnetic field around the cable. Figure 7 shows this equipment. This method is usually better for measurement of the larger milliampere settings. It is recommended that all milliampere stations be calibrated to within $\pm 5\%$ or ± 1 mA (whichever is larger) of the indicated value on the control panel. Furthermore, the shape of the milliampere waveform should be observed on an oscilloscope to detect problems, such as arcing, loading, lack of stabilization, rectification problems, and excessive ripple. The waveform should be regular and the peaks in the milliampere waveform should not vary by $>7\%$.

MILLIAMPERE LINEARITY

The assessment of the relationship between the various milliampere settings and their corresponding measured X-ray output is called milliampere linearity. If the milliampere is increased by a factor of 2 with all other settings unchanged, the X-ray output should also increase by a factor of 2. If the milliampere settings on the control are miscalibrated, the actual milliamperage will not increase linearly with the selected values.

In order to make milliampere linearity measurements, one merely places a radiation detector in the X-ray field along the central ray. The detector is kept at a fixed location. One peak kilovoltage and time setting is selected (e.g., 80 kVp and 0.1 s). A series of fixed exposures are measured for each of two (or more) different milliampere settings (e.g., 200 and 400 mA). The average exposure value for each of the 2 mA stations is calculated (E_1 and E_2). The average exposures are divided by the mAs values used to produce the exposures mGy per 100 mAs (or mR mAs^{-1}). The term ($X_1 = E_1 \cdot mAs^{-1}$) should be the same for all selections of milliamperage employed. A linearity factor (LF) is then calculated as follows:

$$LF = |X_1 - X_2|/|X_1 + X_2|$$

Federal regulations require that the LF to be <0.10. Actually, most X-ray generators are capable of LF values <0.05 for the 2 mA settings with the largest discrepancies (if properly calibrated and functioning correctly).

PEAK KILOVOLTAGE CALIBRATION

Improper peak kilovoltage calibration of the X-ray generator has major influences upon image quality and patient dose. High peak kilovoltages reduce patient contrast, increase scatter, and increase subject latitude; they also result in reduction of patient dose. Low peak kilovoltages have the opposite effect. Moreover, miscalibration results in non-uniform X-ray quality throughout a large X-ray department. Thus, a standard for the determination of peak kilovoltage accuracy is important.

There are a number of methods by which peak kilovoltage calibration can be measured. These methods include high voltage bleeder resistor networks and noninvasive kVp meters (26–29). A simple method often utilized is noninvasive kVp meters, such as one shown in Fig. 8. These meters are placed in the X-ray beam and display a digital number for the measured peak kilovoltage. The meters contain several solid-state radiation detectors, each with different amounts of copper filtration. By measuring the amount of penetration of X-ray beam through these filters the peak kilovoltage can be computed. These noninvasive peak kilovoltage meters are usually accurate to within ± 2 kVp. Even though more accurate peak kilovoltage instrumentation exists, the relative ease with which is noninvasive peak kilovoltage measurements can be made is their principal advantage. It is recommended that peak kilovoltage calibration be accurate to within $\pm 5\%$ or ± 2 kVp, whichever is larger at all milliampere

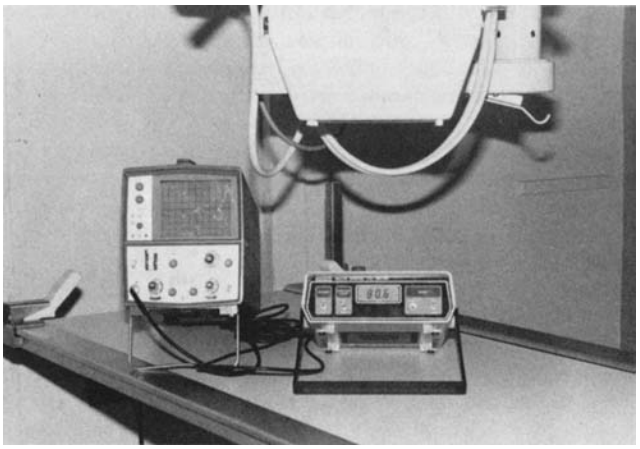


Figure 8. Equipment used to measure and display to peak kilovoltage of the generator. The device shown below the X-ray tube is a noninvasive peak kilovoltage meter. The storage oscilloscope is used to display the peak kilovoltage waveform.

stations for a range of peak kilovoltage settings. For radiographic units, the checks are usually made at 60, 90, and 120 kVp, using a variety of milliampere settings. The shape of the kVp waveform can also be observed by a BNC output connection from the noninvasive kVp meter to an oscilloscope. The peak kilovoltage waveform is monitored to detect problems with arcing, excessive or irregular ripple, stabilization problems, rectification failures, phase imbalances, and noise. The peaks in the kVp waveform should not vary by >7%.

OPERATOR'S CONTROL PANEL

The control panel contains the selection switches that determine the technique factors for the X-ray exposure. The QC program should determine that selection switches function properly and that regulatory requirements are met.

FUNCTION CHECKS

The various selection switches should be tried; it needs to be determined that the switches actually function. Furthermore, all lights, meters, and digital displays should be checked to make certain they are operational.

EXPOSURE CONTROL

Recommendations suggest that the exposure control should be secured at least 40 in. (1 m) from the edge of the secondary radiation protection barrier. It should not be possible to carry the exposure switch into the room and/or make an exposure while inside the X-ray room.

REGULATORY CHECKS

The operator's control panel is required to have certain features. The QC procedures should check for these regulatory items. There must be a warning label with a caution about the hazards of X-ray irradiation. The technique factors must be indicated on the control panel prior to an exposure. During an exposure, there should be a visible beam-on indicator and an audible indicator of exposure termination. Following the exposure, there should be an automatic resetting of the exposure termination timer. For fluoroscopic units, there should be a 5 min timer; after 5 min of fluoroscopy, there must an audible indicator at the and of each 5 min of fluoroscopy. The 5 min timer alarm must continue to sound until the timer is reset (18). However, the reset of the alarm must not affect the cumulative timer that accumulates the total fluoroscopy time during each study.

TABLE AND TUBE STAND

Foremost, the mechanical functions of the table and tube stand motion should be checked. All typical positions utilized clinically should be tried. The accuracy of distance indicator for the SID should be determined; a tape measure can be used to check the SID indicator. The mechanical safety of the table and tube stand should be visually inspected. Anticollison devices should be checked.

AUTOMATIC EXPOSURE TERMINATION

Automatic Exposure Control (AEC) or phototiming utilizes a radiation detector to measure the amount of X-ray radiation incident upon the film–screen or CR cassette. When the detector measures sufficient radiation to produce adequate density on the image or a level of mottle on digital image receptors that is deemed appropriate, the detector sends an electronic signal to the X-ray generator, which terminates the exposure. In this manner, with properly functioning phototimers, radiographs should have satisfactory density (film) or minimally acceptable noise (CR/DR) to prevent degradation of image quality. In order to check the AEC system, the phototimer tracking should be measured.

PHOTOTIMER TRACKING

An accurate AEC circuit can result in a decrease in the repeat rate and a consistent image quality for a variety of patient sizes. However, phototimer circuits can be susceptible to variations with changes in the peak kilovoltage, non-linear responses of film–screen systems, a limited range of patient sizes, or a lack of reproducibility. The best method to check the phototimer is to use acrylic plastic sheets to stimulate the patient and to take radiographs of the plastic. The acrylic sheets are placed in front of the phototimer to simulate various patient sizes. For various peak kilovoltage settings or changes in the plastic thickness, the density of the exposed and developed film (as determined with a densitometer) or Exposure Index Number for CR or DR cassettes should remain constant. If the phototimer circuitry is performing properly, it is recommended that the measured film density be within ± 0.15 du at one standard deviation (or $\pm 10\%$ for digital system index number) of the average value for all variations in peak kilovoltage or simulated patient thickness. Furthermore, for conditions of fixed peak kilovoltage and thickness, reproducibility of the measured film density should be within ± 0.05 du (one standard deviation) or $\pm 5\%$ for the digital system index number.

BUCKY TRAY

The tray into which the film-screen or CR cassette is inserted is called the bucky tray. Above the tray, there is usually a moving grid that is used to remove some of the scattered X rays. It is important that the grid moves properly during the exposure and that it does not contain any defects. A common problem is that the wrong grid is inserted into the table. The detente mechanism that positions the X-ray tube over the cassette may also not be properly centered. The QC tests should attempt to investigate for possible problems.

VISUAL INSPECTION

If possible, the grid should be removed from the bucky assembly on an annual basis. It should be visually inspected for damage and for spills of contrast material. The focal distance and grid ratio marked on the grid should be checked to determine if they are appropriate for the intended radiographic procedures. It should also be determined that the correct side of the grid faces the X-ray tube; inverted grids result in cutoff of image density toward the edges.

GRID UNIFORMITY

In order to ascertain that there are no defects in the grid and that there are no irregular cutoffs, a cassette should be placed in the bucky tray and a rapid exposure made to result in a film density of ~1.20 du (or appropriate digital image index number). Exposure times should be <20 ms in order to stop grid motion. The developed film should have uniform density throughout and the grid lines should be visible. Irregularities in the grid lines or nonuniform density may be indicative of a grid problem.

BUCKY MOTION

Sometimes, the device that causes the grid to oscillate during the exposure malfunctions or is not properly synchronized with the X-ray exposure. To test for bucky motion, a cassette is placed in the bucky tray and an exposure is taken with a longer exposure time selected. Exposure times should be >100 ms. The developed image should have a uniform density and grid lines should not be visible. The bucky motion should blur all the grid lines on the image. If the grid lines still appear, there is a problem with the bucky motion.

CENTERS ALIGNMENT

The miscentering of the X-ray field and the center of the image receptor cassette could result in the loss of diagnostic information. For this test, the detente and collimator light systems are used to position the X-ray tube assembly over the film–screen cassette that has been placed in the bucky tray. The X-ray field size must be smaller than the cassette. An X-ray exposure of the cassette is made and the centers of both the X-ray field and the image are determined. The misalignment of the center of the cassette with the center of the X-ray field must be <2% of the distance between the X-ray tube focal spot and the cassette (SID).

FLUOROSCOPIC X-RAY EQUIPMENT

Some of the components in fluoroscopic X-ray equipment are similar to those in radiographic equipment. The same X-ray generators are often used to operate several X-ray tubes in a single radiology room; often a fluoroscopic and radiographic X-ray tube are connected to the same X-ray generator. However, in addition to the standard radiographic components, fluoroscopic X-ray equipment usually includes additional components, such as image intensifier, cassette spot film devices, television systems, 100 mm (or digital) fluoroscopic spot film cameras, and a tilting table. The fluoroscopic system is designed to operate as a real-time dynamic imager. The X-ray tube operates with low milliampere values for contiuous fluoroscopy so that X-ray production can continue for minutes without overheating the X-ray tube. Pulsed fluoroscopy utilizes large milliampere values (>100 mA), but the duty cycle (actual X-ray production) is ~5–30%. The image intensifier stops the X rays transmitted through the patient and converts them to a light image. The light from the image intensifier can then be recorded by the television system, cassette spot film, or the spot film camera. The QC tests should also include those components that are specific to fluoroscopic systems. The QC procedures for the X-ray generator and tube have been described previously (30).

TRACKING TEST

The image-intensifier assembly (IIA) on fluoroscopic X-ray systems can be positioned at various heights above the table. It is important that the collimator on the X-ray tube adjusts to track the image receptor properly, otherwise, the X-ray field size could possibly exceed the image receptor size. This function is observed by placing an object in the fluoro field and moving the IIA up and down. The collimator should adjust automatically. Federal regulations require compliant fluoro systems to have tracking with SID adjustment.

INTERLOCK TEST

When the IIA is placed in the park position, fluoroscopic X-ray production should be inhibited. If fluoroscopic X-ray production were not inhibited, the patient could receive a radiation dose without any image being recorded (i.e., unnecessary dose). The test is performed by placing the IIA in the park position and then placing a radiation detector on the table surface. The fluoroscopic "on" switch is depressed and the detector indicates whether X rays were produced. IIA interlocks are required for compliant units.

PRIMARY BARRIER TRANSMISSION

The IIA acts as a primary radiation barrier. Any X ray impinging upon it should be drastically attenuated. Transmission of the X rays should be minimal. The test is performed by placing a 1.5 in. (3.8 cm) aluminum penetrameter on the table surface in the primary X-ray beam and depressing the fluoro switch. The entrance radiation exposure rate into the penetrameter is measured as well as the exposure rate of the radiation transmitted through the IIA. Federal regulations require that the exposure rate transmitted through the IIA should not exceed 2 $mR \cdot h^{-1}$ at 10 cm above the IIA surface for each roentgen per minute of entrance exposure.

MAXIMUM ENTRANCE EXPOSURE RATE

In order to limit the patient dose, the federal government has established regulations to limit the maximum fluoroscopic entrance exposure rate. The measurement can be performed with no backscatter material according to regulatory specifications; however, more realistic conditions can be simulated by placing a 1.5 in. (3.8 cm) aluminum penetrameter on the table with a radiation detector next to the penetrameter on the side toward the X-ray tube. Once the detector is centered in the X-ray beam, a lead sheet (3 mm thick) is placed on the penetrameter side toward the IIA. Both the Automatic Brightness Control (ABC) and the manual adjustments are used serially to obtain the highest X-ray radiation levels. The radiation detector is used to measure the exposure rate. Federal regulations require that the entrance rate at the point where the center of the useful beam enters the patient shall not exceed 87 $mGy \cdot min^{-1}$ in air (10 $R \cdot min^{-1}$) in the ABC mode, except during the recording of the fluoroscopic image (cassette or spot exposures) or when provided with an optional high level control. When provided with the optional high level control, the equipment shall not be operable at any combination of tube potential and current that will result in any exposure rate in excess of 87 $mGy \cdot min^{-1}$ in air (10 $R \cdot min^{-1}$) at the point where the center of the useful beam enters the patient unless the high level control is activated. In the higher level fluoroscopy mode, a special buzzer or chime must sound and the maximum patient entrance exposure rate must be <174 $mGy \cdot min^{-1}$ in air (20 $R \cdot min^{-1}$). Furthermore, equipment that does not employ ABC shall not be operable at any combination of peak kilovoltage or milliampereage that will result in an exposure rate in excess of 43 $mGy \cdot min^{-1}$ in air (5 $R \cdot min^{-1}$) at the point where the center of the useful beam enters the patient, except during recording of the fluoroscopic images or when provided with an optional high level control.

TYPICAL PATIENT ENTRANCE EXPOSURE RATES

In order to assess the normal operation of fluoroscopic or cine equipment, typical patient entrance exposure rates should be measured. These rates can be influenced by the improper functioning of the equipment and by adjustments made by the service personnel. A 3.8 cm (1.5 in.) aluminum penetrameter to which 0.5 mm copper has been added can be utilized to simulate an average male patient of $\sim$75 kgm weight; the measurement procedure is the same as that for maximum dose rate, except that the lead is not placed in the X-ray beam. It is recommended that the simulated patient entrance exposure at the point the X-ray beam enters the phantom should not exceed 8.7–26.1 $mGy \cdot min^{-1}$ in air (1–3 $R \cdot min^{-1}$) in the 23 cm (9 in.) Field-of-View (FoV) with ABC. During these measurements, the indicated fluoroscopic peak kilovoltage should >55 kVp and <90 kVp. During digital cine recording in the 23 cm (9 in.) mode, the measured entrance exposure should be less than 200 μGy per frame (23 mR per frame) with the same phantom of 3.8 cm (1.5 in.) of aluminum plus 0.5 mm of copper. Some local regulatory agencies require that typical patient radiation levels be measured annually for all X-ray equipment.

LIMITATION OF THE X-RAY FIELD TO THE IMAGE RECEPTOR SURFACE

The intent of this test is to limit the X-ray field size to the IIA surface such that the patient is only irradiated over a surface for which a TV image can be seen. If the X-ray field size were larger, the patient would be unnecessarily receiving a radiation dose over an area that would not be imaged. In order to perform the measurement of X-ray field size, a leaded ruler is placed in two orthogonal directions upon a film in a cardboard cassette. The film with the ruler can be placed in the X-ray beam anywhere between the tabletop and the IIA surface. The film and ruler are then fluoroscoped. The TV image is viewed to determine the amount of the scale that is visible on the TV monitor. The exposed film is then developed and the size of the fluoro X-ray field is indicated by the darkened portion on the film. Federal regulations specify that the misalignment between the X-ray field size and the image receptor in the fluoroscopic mode must be $<3\%$ of the SID in either orthogonal direction and less than a sum of 4% of the SID for the two orthogonal directions combined.

FLUOROSCOPIC X-RAY OUTPUT

The fluoroscopic X-ray output is valuable information utilized in several ways. These data can then be used to yield patient dose estimates. Furthermore, miscalibrations in peak kilovoltage and/or milliamperage and filtration problems would become apparent in the fluoroscopic X-ray output measurements. Additionally, malfunctions in the X-ray tube, such as anode deterioration, should cause drastic changes in the fluoroscopic output. The measurement is performed by placing an X-ray detector on the table surface (without backscatter) in the X-ray beam to measure the air dose rate in $mGy \cdot (mA \cdot min)^{-1}$ [or exposure rate in $R \cdot (mA \cdot min)^{-1}$] during fluoroscopy for various peak kilovoltage settings. It is recommended that the fluoroscopic X-ray output measured at the table surface ($\sim$46 cm or 18 in.) from the focal spot should be between 13 and 22 $mGy \ (mA \cdot min)^{-1}$ [or 1.5 and 2.5 $R \cdot (mA \cdot min)^{-1}$] at 100 kVp, except for heavily filtered X-ray beams.

RADIATION INTO INPUT RECEPTOR(S)

The operation of fluoroscopic X-ray equipment is closely aligned to levels determined by feedback control loops that measure parameters associated with the input radiation exposure levels. Thus, proper equipment performance necessitates that the levels be adjusted appropriately. Within the guidelines listed below, the levels should be high enough to limit fluoroscopy image noise to tolerable levels, but low enough to limit patient dose and contrast losses associated with the ABC driving to high peak kilovoltage settings. These tests are performed with a 3.8 cm (1.5 in.) aluminum plus 0.5 mm copper in the

X-ray beam and the brightness being controlled by the ABC mode. The IIA should be placed high above the aluminum and copper attenuator in order to limit scattered X-rays from being measured. The radiation detector should be placed at the IIA surface with the grid removed (on some units the grid cannot be easily removed). It is recommended that the following radiation levels for input exposure rates the image receptor be established:

1. Fluoroscopy for 23 cm (9 in.) mode (no grid) $\leq$ 6 mR·min^{-1} (100 R·s^{-1} or 0.9 Gy·s^{-1})
2. Fluoroscopy for 15 cm (6 in.) mode (no grid) $\leq$ 12 mR·min^{-1} (200 R·s^{-1} or 1.8 Gy·s^{-1})
3. Digital cine recording for 23 cm (9 in.) mode (no grid) $\leq$ 20 R/frame or 0.17 Gy·f^{-1}
4. Digital subtraction angiography radiographic exposures (no grid) $\leq$ 1.0 mR per image or 8.7 Gy per image for 23 cm (9 in.) mode

Excessive input radiation levels are indicative of problems, such as the selection of too small an aperture (a diaphragm placed behind the objective lens of the IIA) or deterioration of the image intensifier.

SPATIAL RESOLUTION MEASUREMENTS

The image quality analysis of fluoroscopic X-ray systems requires the measurement of the spatial resolution capabilities of the various imaging modalities. Unfortunately, the TV chain usually degrades the spatial resolution available from the output phosphor of the image intensifier (II). Hence, measurements of II spatial resolution require the removal of the TV camera; direct observation of the II output phosphor with a special telescope is necessary. This measurement is rarely done. Furthermore, the spatial resolution is dependent upon: the magnification mode for the II [6 or 9 in. (15 or 23 cm) field size], the use of a grid, amount of scatter, the type of test pattern, the location of the test pattern on the II surface, the peak kilovoltage employed, and the focal spot size. Focal spot influences are minimized by the placement of the bar test pattern as close as possible to the II input surface; however, clinical usage has the object of interest displaced away from the II input surface. Therefore, a second test should be performed with the test pattern located at a position corresponding to the patient's location. In this case, the combined effect of focal spot and II distortion is included in the spatial resolution measurement.

A variety of measurement procedures for assessing the spatial resolution of the II are utilized; the following methodology, however, is suggested. A 3.8 cm (1.5 in.) aluminum penetrameter should be placed on the tabletop at the center of the X-ray beam. A 0.10 mm lead bar pattern (0.5–5.0 line pairs per millimeter) should be taped at the II surface with the grid removed. The equipment should be operated in the ABC mode with 60–80 kVp. The output phosphor of the II should be observed through a special telescope. An alternative procedure would be to record the image of the bar pattern with the fluoroscopic spot film camera. The measured spatial resolution for the II should he $>$ 3.8–4.0 line pairs per millimeter in the 15 cm (6 in.) mode and 2.5–2.7 line pairs per millimeter in the 23 cm (9 in.) mode. The TV camera should then be reattached and the spatial resolution on the TV monitor should be examined. The lead bars should he oriented at 45 degree to the raster lines on the TV system. For a 525 line TV system, the spatial resolution should be at least 1.8–2.0 line pairs per millimeter in the 15 cm (6 in.) mode and at least 1.2 line pairs per millimeter in the 23 cm (9 in.) mode. For 1023 line TV units, the measured spatial resolution is typically 50–80% better. Flat panel image receptors typically have resolutions of 2.2–2.8 line pair millimeters in all FoV.

CONTRAST RATIO

Due to light scattering from adjacent areas (veiling glare) in the fluoroscopy image, contrast can be degraded. It is therefore important to measure the maximum contrast achievable. To perform this measurement, a lead disk [at least 0.32 cm (1/8 in.) thick] is placed at the center of the input surface. The diameter of the disk should be 10% of the II input area selected. It is suggested that a thin penetrameter be placed in the X-ray field. For modern digital systems (without film recording) a light meter can measure the intensity levels behind the lead disk and the surrounding background on the display monitor using standardized ontrast and brightness control settings. The contrast ratio is a ratio of the light intensity in the area surrounding the disk to the area directly under the disk. For satisfactory fluoroscopy image receptors, the contrast ratio measured with a light meter should be $>$40 : 1–60 : 1. For film type measurements, the ratio should be $>$20 : 1.

CONVERSION GAIN

The conversion gain of an II defines the efficiency by which the device can convert incident X-ray radiation at the II entrance into light output at the output phosphor of the II. Conversion gain (G_x) is measured in units of nit (cd·m^{-2}) of light output per mR·s^{-1} of X-ray radiation input. The conversion gain degrades with age and usage; typical amounts for the decrease in conversion gain are 5–10% year^{-1}.

As the II conversion gain decreases the patient radiation dose will increase and the X-ray tube heat loading will increase. Although II conversion gain measurements can be performed by medical physicists, it is usually better to request the manufacturer's service personnel to do the measurement (31,32). Conversion gain should be measured immediately after installation when the II is new. By the time that the conversion gain drops to 50–60% of its initial value, replacement of the II would be recommended.

SCATTERED RADIATION LEVELS

Information about the scatter radiation levels is relevant to the assessment of radiation protection criteria for radiology

personnel. In order to limit the radiation to the radiologists to 870 Gy (100 mR)per week, the measured levels should be $<$170–220 Gy·h^{-1} (20–25 mR·h^{-1}) for an estimated maximum of 4–5 h of actual fluoro time per week.

The scatter radiation measurements are performed by placing 25 cm of acrylic in the fluoro beam with the collimator wide open. The ABC is used to control the radiation levels. A portable survey meter is used to measure the scatter radiation levels. It is assumed that protective curtains and flaps will be utilized where they are available. Radiation levels are measured for eye, chest, and gonadal levels at standard positions of staff in the fluoroscopic rooms.

MAMMOGRAPHIC X-RAY EQUIPMENT

Mammographic equipment is a special category of X-ray units. These units can either be used with special film–screen cassettes or as digital units. Mammographic units intended for both types of units typically have a molybdenum X-ray tube target with a beryllium window and either a molybdenum or rhodium X-ray beam filter. (Some mammography units are available with either rhodium or tungsten X-ray tube targets.) Digital units typically contain an image receptor with a scintillator phosphor coupled to photodiode array or a direct conversion amorphous selenium radiation detector array. Both types of mammography units operate at low peak kilovoltages usually with X-ray tube potentials of 25–35 kVp. The measured HVLs will depend on the type of unit being tested. Federal regulations specify that the HVL at 30 kVp must be at least 0.33 mm of aluminum equivalent using ultra pure aluminum filters (not 1100 Al) (see table in Half Value Layer Determination).

At these low peak kilovoltages, small irregularities in the compression plate, grid, or cassette can produce artifacts on the radiographs. Even a small piece of tape (or dirt) inadvertently placed on the compression plate can create objectional artifacts. Another consideration is that the female breast is radiation sensitive. Therefore, the radiation dosage during mammography should be kept as low as reasonably possible (ALARA). Finally, mammography strives to achieve better spatial resolution than other types of film–screen radiographs. To achieve these goals, an X-ray tube with a small focal spot size is used to minimize geometric blur. The film–screen cassette is also designed to provide excellent film–screen contact. Beyond the aforementioned factors, the mammographic equipment has many features in common with the standard radiographic units. The QC procedures for mammographic equipment must be expanded to encompass their special features. Specialized tests required for mammographic units are listed below (33–35)

TYPICAL PATIENT DOSE

There are a number of methods to specify the patient dose. One can either measure a skin entrance dose, a mid-breast dose, or an average glandular dose; the average glandular dose is being recommended due to its biological significance (34). Because of the ease of measurement, however, the skin entrance exposure value is most often determined. The measurement is dependent on breast size, breast composition, peak kilovoltage used, anode and filter combination used, and the type of image being recorded (digital or film-screen). To simplify the measurement, 4–5 cm of BR-12 plastic or acrylic plastic can be used to represent a typical breast. The exposure should be taken to obtain normal densities or digital exposure settings for the image. The skin entrance exposure is then measured with a radiation detector that has a good low energy response. The measured radiation with grid should be 5.2–13 mGy (600–1500 mR)dependent on: the type of mammography unit, film–screen or digital selection, image processing conditions, film density, anode type, filtration type, and kVp used. The average glandular radiation dose for a 4.2 cm compressed breast must be $<$3.0 mGy (300 mrad). Typical values for both film–screen and digital units with a 4.2 cm ACR phantom are $\sim$1.5–2.0 mGy (150–200 mrads). Although the use of grids increases the measured exposure values by 2.0–2.5 times, grids are essential to obtain good image contrast.

SPATIAL RESOLUTION

To assess the spatial resolution of the imaging system, a bar pattern (same one described previously in radiography and fluoroscopy section) should be placed on top of a 4–5 cm plastic block (BR-12 or acrylic) that simulates a typical breast. An image should be recorded and the spatial resolution seen on the image should be determined. This measurement combines the effects of the focal spot and image system blur. In general, the spatial resolution of most mammography systems is excellent. It should be anticipated that the measured spatial resolution should be $>$13 line pairs per millimeter with bars along the anode–cathode direction and 11 line pairs per millimeter with bars in the perpendicular direction for film–screen systems. The spatial resolution for digital systems is limited by the detector pixel size; and it is usually in the 5–10 line pairs per millimeter range for current technology.

ARTIFACTS

To identify artifacts, an image of the 3–5 cm acrylic sheets should be imaged. The density should be appropriate and uniform across the surface. Any changes in density are indicative of potential artifacts.

IMAGE QUALITY

Good mammographic image quality involves an assessment of numerous factors. The QC procedures should require that a radiographic image of a phantom that includes several different types of test objects be taken. Figure 9 is an image of the ACR phantom that contains fibrils, calcium specks, simulated masses, and a plastic disk for contrast assessment. For film–screen units, at least 4 fibrils, 3 speck groups, and 3 masses must be clearly

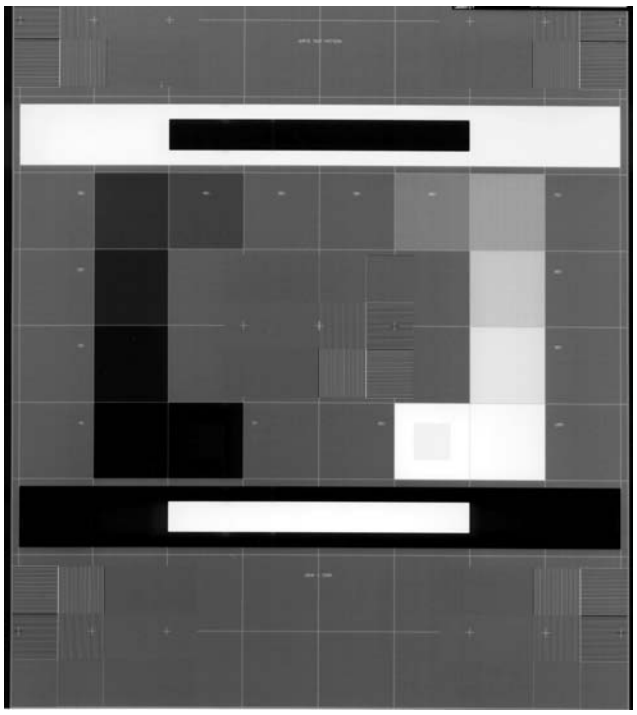


Figure 9. CT image of ACR CT phantom showing plastic rods for CT number linearity and wire for slice thickness evaluation.

visible for a total score of 10 objects (according to ACR and MQSA standards). Digital mammography systems should be able to visualize even more objects in the phantom. The ACR Mammography Accreditation Manual should be consulted for details about many of the mammography equipment tests.

TRAY TRANSMISSION

Since the patient's lap is often situated beneath the cassette tray, it is important that the back surface have sufficient lead lining to significantly attenuate or block the X-ray beam. A radiation detector should be used to measure the amount of radiation transmitted during typical mammographic exposures. It is recommended that the transmission be $<0.01\%$ of the primary beam.

BODY-SECTION TOMOGRAPHIC EQUIPMENT

The purpose of the equipment is to image thin planar sections through the patient's anatomy using film–screen or CR cassettes. The X-ray tube and cassette move in opposite directions about a pivot (or focal) plane to blur objects outside this focal plane. The motion is done such that the anatomy at a preselected depth in the patient will remain in focus on the image receptor cassette. All other anatomy outside the focal plane will be blurred by the motion.

The QC procedures of these tomographic units should include additional tests directed toward analyzing the tube motion and the planar radiographs (36).

PINHOLE IMAGES

The angle through which the X-ray exposure is made influences the thickness of the imaged tomography planar sections. Moreover, it is important that the motion be uniform and symmetric about a central point. All these factors can be assessed through a pinhole image.

To do this test, a lead sheet with a small hole [($\sim\frac{1}{8}$ in. (0.32 cm)] drilled in it should be suspended several inches above the tabletop at the height of the focal plane. A film–screen cassette should be placed on the table top beneath the hole in the lead sheet. With the X-ray tube at zero degree tomo angle, the pinhole should be placed along the central ray and a short exposure taken with no tomo motion to mark this position. Then, a routine tomographic exposure should be made and the cassette film developed. The track on the film should be of uniform density and symmetrical about the zero tomo angle point. The tomographic angle can be computed from the height of the pinhole above the cassette (h) and the width of the track on the film (W). For linear tomography, the tomographic angle is given by the following equation:

$$\theta = 2\tan^{-1}\frac{W}{2h}$$

The measured exposure angle should correspond to the selected angle within $\pm\ 1^{\circ}$.

SLICE THICKNESS

A standard tomographic phantom has been designed to measure the slice thickness and image quality for the body section units (37). The phantom is merely placed on the table and a preselected depth is chosen. A routine tomographic scan is taken and the image developed. Figure 10 shows a tomographic image through this phantom. Examination of the portion of the image that is not blurred yields the depth accuracy, slice thickness, and image quality. It is recommended that both the location of focal plane depth and the slice thickness be accurate to within ± 1.0 mm.

COMPUTED TOMOGRAPHY

Computed tomography scanners are composed of many major subsystems. The hardware contains a variety of complex equipment, such as the X-ray source, the rotating mechanical assemblies in the gantry, radiation detectors, signal-processing electronics, computer systems, display systems, and input–output devices (38). The X-ray source consists of a standard X-ray tube and generator. This portion of the equipment can be tested in the routine manner to check the peak kilovoltage, milliamperage, and time calibration. Because it is difficult to evaluate each of the other CT subsystems individually, the QC tests assess the overall image quality through the use of phantoms. Other tests are directed toward measurement of the patient dosimetry, slice geometry, and table motion. There are several different types of CT phantoms that can be used for QC testing (39–42). The procedures that will be described here are based upon use of the ACR CT Accreditation

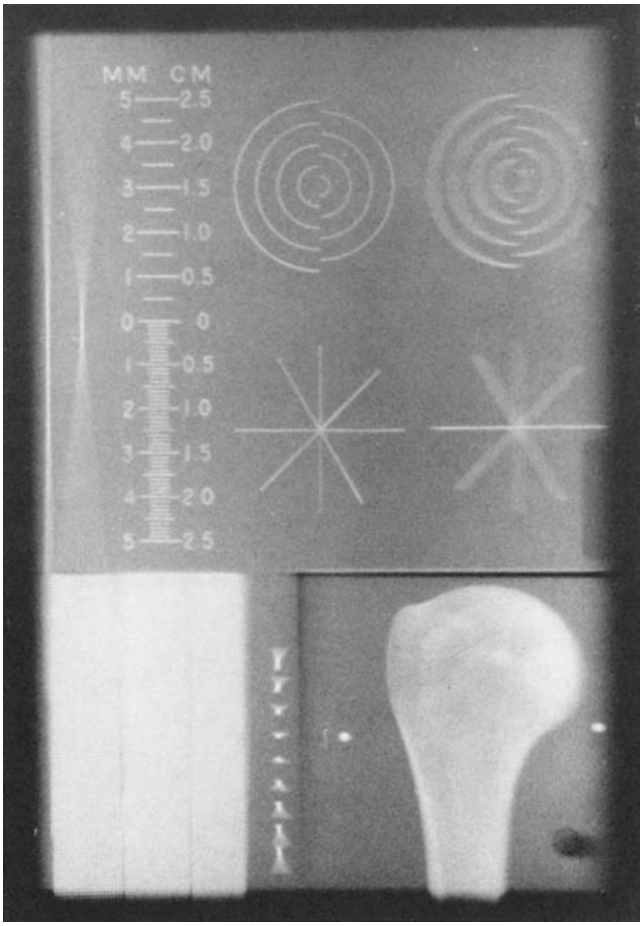


Figure 10. Radiograph of a tomographic phantom used to assess performance of linear tomographic units.

phantom. Similarly, there are many methods by which the CT dosimetry can be performed; however, most often a specialized CT ionization chamber with 100 mm length is used. The chamber is placed in holes drilled in cylindrical acrylic phantoms with a 16 cm diameter cylindrical head and 32 cm diameter body phantom are used for this purpose.

CT NUMBER LINEARITY

The CT numbers of a computerized axial tomographic unit represent the effective linear attenuation coefficients of the various materials being analyzed. In order to be able to differentiate various kinds of body tissues, the CT scanner must be able to measure the effective linear attenuation coefficients correctly for a wide range of materials from air to compact bone, that is, the CT scanner must be able to detect small differences in the way different substances stop X-rays. Usually, the CT numbers (in Hounsfield units) represent the difference in the linear attenuation coefficients between air and water as 1000 units. The CT number linearity is determined by plotting the CT numbers for known materials (usually plastics) versus their effective linear attenuation coefficients, as published in the literature. Linear attenuation coefficients are a function of the photon energy, as well as the X-ray beam filtration. A typical average energy for a well filtered 120 kVp X-ray beam is ~70 keV; this effective energy is often used for this assessment.

One section of the CT phantom contains various rods of plastic embedded in a water equivalent plastic cylinder that has a diameter of 20 cm. The rods consist of air, polyethylene, acrylic, and simulated bone. The anticipated CT numbers for these materials at 120 kVp are approximately −1000, −95, 120, and 955 Hounsfield Units (HU), respectively. Water (or water equivalent plastic) has a reference value of zero. The CT numbers for the various substances are plotted against their attenuation coefficients. The linear regression fit to the data should have a correlation coefficient of >0.99. In order to compare different types of tissue by their CT numbers, it is important to eliminate both CT number offsets and/or slope differences in the linearity scale.

CONTRAST SCALE

The contrast scale for any given CT scanner can be arbitrary. If calibrated properly, however, the CT number of air should be −1000 and the CT number of water should be zero. Compact bone should have a large positive value near +1000 or more. In this manner, an acceptable gray scale will be established for the various anatomical substances shown in scans of patients.

The contrast scale is the change in the effective attenuation coefficient per CT number unit. Usually, measurements are performed with reference to plexiglass and water. The contrast scale (CS) is represented by the following equation:

$$\mathrm{CS} = \frac{\mu_{\mathrm{plex}} - \mu_{\mathrm{water}}}{(\mathrm{CT\,No.})_{\mathrm{plex}} - (\mathrm{CT\,No.})_{\mathrm{water}}}$$

For most CT energies, the numerator of the equation is equal to ~0.024 cm^{-1}. The CS value should be ~0.0002.

In order to determine whether the contrast scale is suitable, one must compute the anticipated CT numbers for air and compact bone based upon the contrast scale measurements.

$$(\mathrm{CT\,No.})x = \frac{\mu_x - \mu_{\mathrm{water}}}{\mathrm{CS}}$$

A number of references list the linear attenuation coefficients for air and compact bone. For 120 kVp, $\mu_{\mathrm{air}} = 0.000208\ \mathrm{cm}^{-1}$ and $\mu_{\mathrm{bone}} = 0.414\ \mathrm{cm}^{-1}$. For water, μ_{water} is usually taken to be $0.190\ \mathrm{cm}^{-1}$. Hence, based upon the measurement of CS, the CT numbers of air, water, and bone can be estimated. Usually the CT units set air at −1000 and compact bone at +1000 HU. (Bone actually has a wide range of CT number values due to the various densities it may have in the human body.) A comparison of the measured values with the ideal values indicates whether the CS is satisfactory.

COMPUTER TOMOGRAPHY NOISE LEVELS

When one images a uniform material in a CT scanner, the CT numbers for a localized region should all be the same. In

practice, the CT numbers vary around some average value. The reason for these variations are changes in the X-ray output, small differences and drifts in the detectors, extraneous electronic noise, and small errors in the computer processing algorithm. All of these parameters introduce variations in the CT numbers. The variations are lumped collectively into a term called the noise of the CT system.

The noise is determined by scanning a uniform water-bath section of the CT phantom and evaluating all the different CT numbers in a small region. The mean value (X) and the standard deviation (σ) of the CT numbers are computed. The CT noise is then expressed as a percentage variation of the linear attenuation coefficient of water.

$$\%\ \sigma\ \text{water} = \frac{(100\,\sigma\,\text{CS})}{\mu_{\text{water}}}$$

The noise value should be measured at various locations in the phantom; both the average noise value for the various locations and the worst case value should be quoted (see Fig. 11).

The noise values for CT scanners are very important performance parameters. The difference in the attenuation coefficients between normal and pathological tissue is small. High CT noise values tend to obscure small tissue differences. For example, differences in the CT values between white and gray matter in the brain are ~0.5–0.6% relative to water (1% on most CT scanners are ~10 CT numbers). In fact, most body tissues have CT numbers of nearly the same value, ±8% (except for bone). Furthermore, the use of lower peak kilovoltages also tends to differentiate tissue better. CT noise levels for a 10 mm slice thickness with 0.50 mm pixel sizes should be below 0.5% of the linear attenuation value of water.

Figure 11. Society of Motion Picture and Television Engineers (SMPTE) test pattern for the evaluation of VDTs used with digital imaging systems.

The measured CT noise levels, however, are a function of many parameters. The noise level decreases as the radiation dose is increased. As the slice thickness is increased, the noise level decreases. The noise levels are also higher for large patients. The relationship to the various factors to the noise level is usually given as the following equation:

$$\%\,\sigma\,\text{water} \propto [B/(HW^3D)]^{0.5}$$

where B is percentage attenuation, H is slice thickness, W is pixel width size, and D is patient dose per scan. In many modern CT units, the noise changes little with pixel size. CT nosie is also closely related to the type of reconstruction kernel chosen. The CT noise is greater for edge enhancement kernels, and CT noise is less for smoothing kernels. To maintain quality control of the CT scanner, noise levels should be monitored on a regular basis. An increase in the noise level is indicative of a developing problem in the CT scanner.

SPATIAL UNIFORMITY

It is extremely important that the response of a CT scanner be uniform across the field being scanned. A CT scan of a uniform material should only exhibit random fluctuations due to system noise. No areas of nonuniform response nor artifacts should be appearing in the scan. This feature is tested by scanning uniform water equivalent plastic section in the CT phantom and examining the CT numbers in different portions of the image. The mean CT numbers should be the same in all portions of the scan. The difference between the average CT number in the various portions of the image should not differ by >2–3 x values the standard deviation of the CT numbers (or ~10 CT numbers).

HIGH CONTRAST SPATIAL RESOLUTION (HOLE PATTERN)

One of the performance criteria of CT scanners is their ability to image small objects. The ability of the scanner to image small objects depends on the subject contrast. The CT scanners have the best spatial resolution with high subject contrast. The ability of the CT scanner to image small, high contrast objects can be measured by a section of the ACR CT phantom; this section has many plastic bars and spaces of different sizes. These bars are of vastly different CT number from the plastic in the spaces to create high contrast, spatial resolution test objects. For each set of bars, a spatial frequency in line pair per cm can be assigned. The frequency ranges from 4 to 12 line pairs per cm. The CT display window and level settings are adjusted to best display bars and spaces. The just barely discernable bar pattern is identified and recorded. It has been suggested that a scanner be capable of resolution approximately twice its smallest pixel size (for pixel-limited resolution). With a 512×512 matrix and a 25 cm FoV, the spatial resolution would be limited by pixel size to ~10 line pairs per cm. For small FoV, the spatial resolution is limited by factors, such as the focal spot size, detector size, and reconstruction algorithms. The modern CT units are capable of spatial resolution of ~8–10 line

pairs per cm for high contrast objects using image zoom display features.

HIGH CONTRAST SPATIAL RESOLUTION (IMPULSE RESPONSE)

Another way of determining scanner resolution is the scanning of a metal pin in a water bath. This scan will produce a point spread response function (PSF). One usually plots the numbers and draws a smooth curve. The point spread for a CT unit can usually be fit with a Gaussian function given below:

$$F(x) = A \exp[-\alpha^2(x - \varepsilon)^2]$$

This equation can be rearranged in order to determine the value of α and ε. The modulation transfer function (MTF) can be computed directly using the value α (43).

LOW CONTRAST RESOLUTION

The CT scanner must not only be able to detect small objects with high contrast, but it must also be able to image small objects with low subject contrast. The difference between normal tissue and pathology is usually small differences in subject contrast (1–2% differences). A good CT scanner should be able to detect small low contrast lesions, that is, low contrast is more important than high contrast resolution.

In order to test low contrast resolution, CT scans are taken through a section in the phantom composed of a plastic with CT numbers only slightly different from water. This section contains various size holes of a slightly different plastic than the surrounding material. The contrast between the plastic and the water is usually 0.4–0.6%. The delectability is dependent upon the radiation dose used, but it would be desirable to have low contrast discrimination of 4–6 mm holes at 0.5–0.6% contrast.

BEAM HARDENING

X-ray beams are polychromatic. Therefore, filtration changes the X-ray spectrum and its effective energy. When CT scanner X-ray beams pass through a lot of bone or long paths through tissue, the X-ray spectrum is altered. The effective linear attenuation coefficients for these hardened X-ray beams change because attenuation is a function of X-ray energy. Therefore, the CT numbers of tissue behind extensive sections of bone may decrease appreciably due to beam hardening. These effects are highly undesirable as some tissues appear abnormal in CT number. This beam hardening artifacts in CT numbers are undesirable.

In order to evaluate beam-hardening effects, 1 in. (2.5 cm) diameter plastic rods with high attenuation coefficients are taped on the opposite sides of the surface of the water section of a CT phantom. This section is scanned, and the CT numbers along the line between the rods are measured. CT scanners should be designed in a manner that the water values are not depressed more than three times the noise standard deviation.

IMAGED CT SLICE THICKNESS

The CT image slide thickness (i.e., sensitivity profile) is measured by scanning a section of the phantom that contains small wires that are displaced along the longitudinal axis of the phatom by a 0.5 mm distance and angled. By counting the number of wires visible in the image, the image slice thickness can be assessed. The measured sensitivity slice thickness should be within ± 0.50 mm of the selected slice values.

PATIENT TABLE TOP INDEXING

It is extremely important that the patient table top moves by the selected distance, otherwise either clinical information or the patient radiation exposure is compromised. This motion is tested by taping a radiation therapy verification film to the outside surface of a cylindrical phantom. Afterward, a series of scans are taken. Each slice thickness setting should be tried; the table motion should be equal to the slice thickness. Three or more consecutive scans should be taken for each collimator setting; the table should index at least 20 mm between different slice thickness settings. Following these scans, the film is developed and examined. There should neither be gaps nor overlaps for consecutive scans. Moreover, the table motion should be accurate to within 1.0 mm for 100 mm of motion in 10 or more steps.

ARTIFACTS

Artifacts can be misinterpreted as anatomical abnormalities in patient scans. Various software and scan techniques should be used to scan the water section of the CT phantom. Since water equivalent plastic should be homogeneous, any artifacts should be clearly visible in the CT image. The various window level and width settings should be utilized to view the image.

PATIENT DOSIMETRY

A special CT ionization chamber (100 mm long) is inserted into cylindrical acrylic phantoms designed to simulate a typical head and body of a patient. The head CT phantom has a 16 cm diameter and the body phantom has a 32 cm diameter. Each phantom has holes drilled at the center and at other locations with different distances from the surface of the phantom. At least one hole is 1 cm inside the outer surface. The ionization chamber is placed in both the center hole and later at the periphery hole. Empty holes are filled with acrylic rods. A single, axial CT slice is taken in the middle of the cylinder, and the exposure measured by the ionization chamber (Em) is recorded. The radiation dose is then computed by the following equation:

$$\text{CTDI} = (C \times f \times \text{Em} \times 100\,\text{mm}/\text{XW})$$

where C = calibration factor for detector, $f =$ exposure to absorbed dose conversion factor, Em= measure exposure, XW = X-ray beam width specified.

The *f*-factor for acrylic is 0.78 $cGy \cdot R^{-1}$, and the *f*-factor for tissue is about 0.92 $cGy \cdot R^{-1}$. The *f*-factor for air is $\sim$0.87 $cGy \cdot R^{-1}$. A weighted CTDI ($CTDI_W$) is calculated using (1/3) of the center value and (2/3) of the peripheral value.

For these data, the actual radiation dose values can be scaled using clinical setting of kVp, milliamperesecond, and pitch. Pitch is the table movement between rotation of the X-ray tube divided by the width of the X-ray beam. The $CTDI_{volume}$ is equal to the $CTDI_W$ divided by the pitch. It is recommended that the $CTDI_{volume}$ for clinical head CT scans of patients be $<$6.0 cGy (rad) and that the body value is $<$3.5 cGy (rad).

FILM PROCESSORS

Because the film processing affects the end product of film–screen radiographic images, this is one of the most important aspects of the QC program for diagnostic radiology. Film processing conditions can affect film contrast, the speed of the film-screen system, and base plus fog levels. Indirectly, inadequate processing usually causes the technologist to alter kilovoltage or mAs that could result in an increased dose to the patient.

In general, a 1 °F (0.6 °C) temperature change in the developer chemistry produces a 0.03–0.07 density change in the developed radiograph. The film density decreases for a decrease in the developer temperature. For an 8–18 °F (4.4–10 °C) decrease in the developer temperature, a 50% increase in the patient exposure is necessary. Similarly, increasing the developer temperature by 5 °F (2.8 °C) can cause a latitude film to appear to have increased contrast. Contamination of the processing chemistry also causes changes in the film density and contrast. The magnitude of the effects are dependent on the type of film, the kind of chemistry, and the processor design.

In order to maintain consistency in the film development, a routine maintenance program for the processor should be established. The replenishment rates should be adjusted properly for the workload. The rollers should be cleaned and inspected on a regular basis and the chemistry should be changed as required. Moreover, steps should be taken to minimize inconsistencies in the chemistry. Within reason, all processors should operate from the same batch of fixer and developer; this can be accomplished by mixing large batches for delivery or using electronic mixers with modular chemistry canisters.

The QC program for processors should utilize daily sensitometry strips that are run through each processor (44). In conjunction with the sensitometry, the processor replenishment rates and temperatures should be checked. Light sensitometers are recommended for making the sensitometry strips (45,46) (see Fig. 12). The film used for sensitometry should be the fastest high contrast film routinely used in the facility. Since variations can be found in different film emissions batches of the same kind of film, strips from both old and new emulsion batches should be run simultaneously during transition periods. The processed sensitometry strips should have the densities measured with a densitometer. A sensitometer step

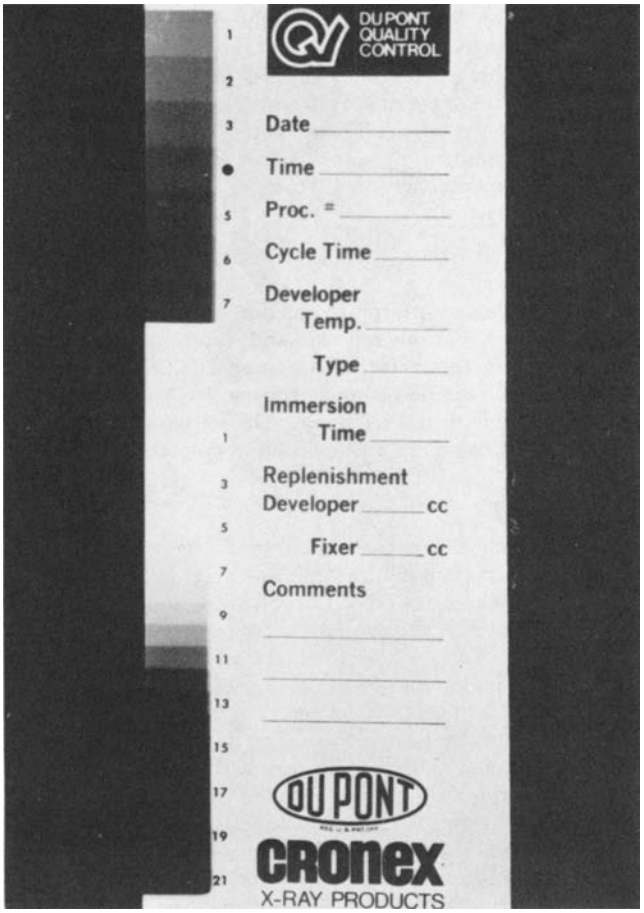


Figure 12. A reproduction of a developed sesitometry strip that is used to evaluate processor performance. The steps used for speed point and contrast index are shown on the strip.

with a density $\sim$1.0 du above the base plus fog level should be selected as the speed point. The difference in densities of two points in the linear portion of the characteristic curve should be used to determine a contrast index. The base plus fog level on the sensitometry strips should also be measured. These three parameters should be compared to optimal values supplied by the film manufacturer and average values should be established for the facility. The speed point should not be allowed to change by more than $\pm$0.15 du about the average value. The contrast index should not change more than $\pm$0.15 units. Since the base plus fog is not very sensitive to variations, a small change is indicative of significant abnormalities in the processing. With good maintenance and QC procedures, the film processors should be very stable and consistent.

DIGITAL IMAGE RECEPTORS

Digital image receptors could be either CR Cassettes with their electronic readers or DR image receptors (47). With both systems, a QC program should be routinely implemented to ensure consistency, identify artifacts and evaluate physicians display monitor (48). Most CR and DR

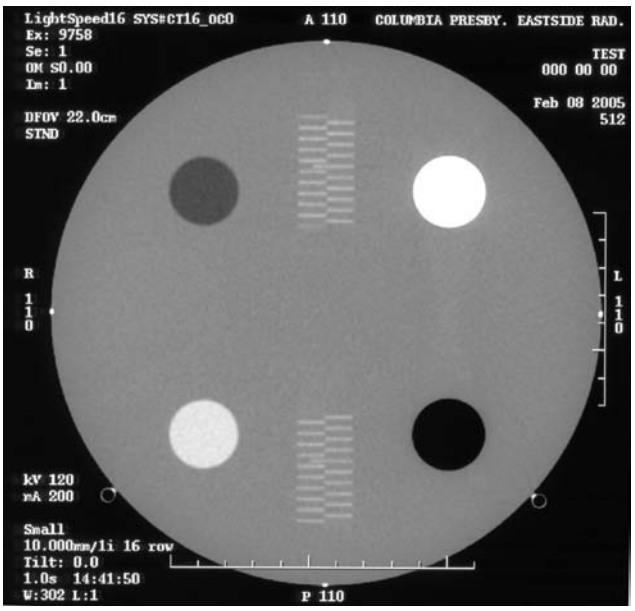


Figure 13. Radiographic image of an ACR Mammography phantom showing fibrils, calcium, speck group, and simulated masses.

systems have an index number to ensure sufficient radiation is utilized to avoid excessive quantum mottle. The QC program should assess that the AEC is properly adjusted to maintain proper radiation exposure. For CR systems, cassette should be exposed to measure spatial linearity, spatial resolution, uniformity of density and proper size and collimation. Both CR and DR units should be evaluated for artifacts and dirt. The physicians' review video monitor should be checked with a video pattern such as the SMPTE (Society of Motion Picture & Television Engineers) for low contrast visibility, spatial linearity, spatial resolution and a reasonable contrast scale. A light meter can then be employed to ensure that maximum luminance and contrast scale are adjusted similarly on all video monitors (Fig. 13).

SUMMARY

There are numerous reasons for providing a QC program in diagnostic radiology. Quality control is mandated by several regulatory agencies; hence a radiology facility must demonstrate that some type of QC is being performed. Another reason for QC programs is to ensure electrical, mechanical and radiation safety for patients and staff. Moreover, the QC program should provide better image quality and a consistency in the radiographs. Finally, the QC program should result in financial savings and improved equipment performance.

The effectiveness of the QC program will depend on the extent of testing performed. A cursory QC program is usually conducted to just meet minimum regulatory requirements. A more extensive program will attempt to establish and maintain a high standard for equipment performance.

Because of the diversity in the diagnostic equipment, the QC program should be directed toward evaluating the special features of each type of equipment. Hence, there are many specialized aspects to QC testing. Although other testing could be used to supplement and strengthen the QC procedures described, the information given in this article provides a basis for establishing an effective QC program for diagnostic radiology equipment.

BIBLIOGRAPHY

1. Joint Commission on Accreditation of Hospitals. Accreditation Manual for Hospitals. Chicago, IL. Available at http://www.JCAHO.org. 2005.
2. American College of Radiology (ACR). Standards. Reston (VA) ACR, 2001–2002 or http://www.acr.org.
3. U.S. Department of Health, FDA, Center for Devices & Radiological Health (CDRH). Routine Compliance Testing Procedures for Diagnostic X-ray Systems. Available at http://www.fda.gov/cdrh/radhlth/xraytestproc.html 2003.
4. U.S. Department of Health, FDA, Center for (CDRN) Devices and Radiological Health (CDRN). Mammograph facility survey, equipment evaluation and medical physical qualification requirements under MQSA. Available at http://www.fda.gov/cdrh/dmqrp/6409.pdf 2000.
5. U.S. Department of Health, FDA, Center for Devices and Radiological Health (CDRH) Resource manual for compliance test parameter of diagnostic X-ray systems. Available at http://www.fda.gov/cdrh/comp/rad_medical.html 2005.
6. Ostinelli A, et al. Quality assurance in diagnostic radiology: Practical outcomes. Rad. Medica Oct 2003;106(4):413–419.
7. Noyes RS. The economics of quality assurance. Radiol/Nucl Med Mag Dec 1980.
8. U.S. Department of Health, Education and Welfare. Quality assurance programs for diagnostic radiology facilities (FDA) 80-8110, Washington (DC): U.S. Govt. Printing Office; 1980.
9. Peer S, et al. Comparative reject analysis in conventional film-screen and digital storage phosphor radiography. Rad Protection Dosim 2001;94(1-2):68–71.
10. Nelson RE, Barnes GT, Wittten DM. Economic analysis of a comprehensive quality assurance program. Radiol Technol 1997;49:129–134.
11. Fields T, Griffith CR, Hubbard LB. What price quality? A quality assurance program for diagnostic radiology. Appl Radiol Jan–Feb 1980;57–64
12. U.S. National Council on Radiation Protection and Measurements (NCRP). NCRP Report No. 99 quality assurance for diagnostic radiology, Bethesda (MD); 1980.
13. Waggener RG, Wilson CR, editors. Quality Assurance in Diagnostic Radiology. New York: American Institute of Physics; 1980.
14. Gray JE, Winkler NT, Stears J, Frank ED. Quality Control in Diagnostic Imaging. Baltimore: University Park Press; 1983.
15. American College of Radiology (ACR). Accreditation Programs (Mammography, MRI, CT, Fluoroscopy and Ultrasound). Available at http://www.acr.org. 2005.
16. American Association of Physicists in Medicine (AAPM). Quality Control in Diagnostic Radiology. Madison (WI): Medical Physics Publishing; 2002.
17. Hospital Physicist's Association. Quality Assurance Measurements in Diagnostic Radiology. Rep. No. 29, London: HPA; 1979.
18. U.S. Department of Health, Education and Welfare. Regulation for the administration and enforcement of radiation control for health and safety act of 1968. (FDA) 75–8003, Washington (DC): U.S. Govt Printing Office; 1976.
19. Seibert JA, Barnes GT, Gould RG. Specifications, Acceptance Testing and Quality Control of Diagnostic X-ray Equipment. New York: Springer-Verlag; 1997.

20. Hendee WR, Ritenour ER. Medical Imaging Physics. 4th ed New York: Wiley; 2002.
21. Bushberg JT, Siebert JA, Leidholdt EM. Essential Physics of Medical Imaging. 2nd ed. Baltimore: Williams & Wilkins; 2001.
22. Stears JG, Felmlee JP, Gray JE. Half-value layer increase owing to tungsten build-up in the x-ray tube: Fact or fiction. Radiology 1986;160:837–838.
23. Spiegler P, Breckinridge WC. Imaging of focal spots by means of the star test pattern. Radiology 1972;102(3):679–684.
24. Hendee WR, Chaney EL. X-ray focal spots: Practical consideration. Appl Radiol 1974;3(3):25–29.
25. Atherton JV, Nickoloff EL. Non-invasive X-ray tube current measurement. Med Phys 1987;14(2):258–261.
26. Chaney EL, Hendee WR. An Instrument with digital readout for indirect determination of kVp. Med Phys 1978;5(2):141–145.
27. Ramirez-Jamenez FJ, et al. Consideration of the measurement of practical peak voltages in diagnostic radiology. BJR Sept., 2004;77(921):745–750.
28. Healey T, Dickson DG, Greenwood MWB. A calibration system for x-ray generators and tube factor. Br J Radiol 1979;52:44–50.
29. Giarratanno JC, Waggener RG, Hevezi JM, Shalek RJ. Comparison of voltage-divide, modified Ardran-Crooks cassette, and Ge (Li) spectrometer methods to determine the peak kilovoltage (kVp) of diagnostics X-ray units. Med Phys 1976;3:142–147.
30. U.S. Department of Health, FDA. Quality Assurance for Fluoroscopy X-Ray Units and Associated Equipment, FDA 80-8095, Washington (DC): U.S. Government Printing Office; 1979.
31. Holm T, Moseley RD. The conversion factor for image intensifier. Radiology 1964;82:898–904.
32. Hay GA, Clarke OF, Coleman NJ, Cowen AR. A set of X-ray test objects for quality control in television fluoroscopy. 1985;April BJR 58(688):335–344.
33. U.S. National Council on Radiation Protection and Measurements (NCRP). NCRP Report No. 85. A user's guide to mammography. Washington (DC); 1986.
34. Hendrick RE, Botsco M, Plott CM. Quality control in mammography. Radiol Clin N Am Nov 1998;33(6):1041–1057.
35. Haus AG. Screen-film mammography update: X-ray units, breast compression, grids, screen-film characteristics and radiation dose. Proc SPIE 1984;486.
36. U.S. Department of Health Education and Welfare. Quality assurance for conventional tomographic X-ray units. (FDA) 80-8096. Washington (DC): U.S. Govt. Printing Office; 1979.
37. Littleton JT. A phantom method to evaluate the clinical effectiveness of a tomographic device. Am J Roentgenol Rad Ther 1970;58:139–145.
38. Nickoloff EL. What to look for when buying CT equipment? Appl Radiol 1982;11(3):69–74.
39. American Association of Physicists in Medicine (AAPM), AAPM Report No. 39 Specification and acceptance testing of computed tomography scanner. College Park (MD): American Institute of Physics (AIP); 1993.
40. McCollough CH, et al. The phantom portion of the American College of Radiology (ACR) CT accreditation program: Practical tips, artifacts, examples and pitfalls to avoid. Med Phys Sept 2004;31(9):2423–2442.
41. Morin RL, Gerber TC, Mc Collough CH. Physics and dosimetry in computed tomography. Cardiol Clinics Nov 2003;21(4): 515–520.
42. Johns HE, et al. Physics of CT scanners: Principles and problems. Int J Rad Oncol Biol Phy 1977;3:45–51.
43. Nickoloff EL, Riley R. A. simplified approach for modulation function determinations in computed tomography. Med Phys 1985;12:437–442.
44. Gray JE. Photographic quality assurance in diagnostic radiology, nuclear medicine and radiation therapy. Vols. 1 and 2 (FDA) 76-8043 and (FDA) 77-8018, Washington (DC): U.S. Govt Printing Office; 1976.
45. Blendl C, Buhr E. Comparison of light and x-ray sensitometric response of double emulsion films for different processing conditions. Med Phys Dec 2001;28(12):2420–2426.
46. Nickoloff EL, Leo F, Reese M. A comparison of five methods of monitoring the precision of automated X-ray film processors. Radiology 1978;129:509–514.
47. Rowlands JA. The physics of computed radiography. Phys Med Biol Dec 2002;47(23):R123–R166.
48. Samei E, et al. Performance evaluation of computed radiography systems. Med Phys March 2001;28(3):361–37l.

See also CODES AND REGULATIONS: RADIATION; PHANTOM MATERIALS IN RADIOLOGY; SAFETY PROGRAM, HOSPITAL.

X-RAY SCREEN-FILM SYSTEMS.

See SCREEN-FILM SYSTEMS.

X-RAY SIMULATOR.

See RADIATION THERAPY SIMULATOR.

X-RAY THERAPY EQUIPMENT, LOW AND MEDIUM ENERGY

C-M CHARLIE MA
Fox Chase Cancer Center
Philadelphia, Pennsylvania

INTRODUCTION

X-rays were used to treat cancer patients within a year of their discovery by Wilhelm Roentgen in November 1895 (1). However, it was not until the invention of the hot cathode tube by Coolidge in 1913 that X-ray beams could be delivered in a stable manner and, hence, the output controlled and measured with any precision. Further improvements in the accuracy of X-ray beam dosimetry occurred with the publication of the first central-axis depth dose tables in 1922 (2) and with the definition of the Roentgen and its use as a standard unit for X-ray dosage in 1928 as recommended by the Second International Congress of Radiology in Stockholm, Sweden.

In those early years, the energy of X-ray beams was limited to 140 kV, and therefore only relatively superficial lesions could be successfully treated. The situation was improved in the early 1920s as higher energy (up to 400 kV) X-ray units became available to permit the treatment of deep-seated lesions (3). Since cobalt-60 treatment units (average γ-ray energy = 1.25 MeV) came into use in the late 1950s and electron linear accelerators capable of producing high energy ($\geq$4 MeV), X rays in the early 1960s, the use of kilovoltage X-rays in radiation therapy has drastically declined for deep-seated tumors.

The use of low energy X rays for the treatment of superficial lesions is still popular, even though the use of electrons has replaced them for treating such lesions in

many cancer clinics. Electron beams have finite ranges that are ideal for treating shallow tumors to spare distal normal tissues. However, clinical accelerators that produce electron beams for radiation therapy are more expensive then X ray units. There is an increased use of low energy X rays for intraoperative radiation therapy, such as for stereotactic brain irradiation or treatment of the dura, for endocavitary irradiation of rectal cancers, and treatment of other malignant skin lesions. The depth-dose properties of X rays are considered by some oncologists to be more favorable than electron beams for adequate surface coverage and for treating macroscopic diseases at depths. Medium energy X rays still play a role in many developing countries, where the technology is easier to support and maintain than electron linear accelerators and safer than cobalt-60 teletherapy units. Medium energy X rays are also widely used in radiation biology research for *in vitro* cell irradiation or *in vivo* animal experiments.

X-RAY DEVICES FOR RADIATION THERAPY

Medium energy X rays (also called *orthovoltage* X rays) are generated by X-ray devices at an accelerating potential between 150 and 500 kV. Low energy X rays (also called *superficial* X rays) are generated by X-ray devices at an accelerating potential between 50 and 150 kV. Very low energy X rays that are generated at an accelerating potential <50 kV have mainly been used in *contact* therapy.

The manner in which X rays are generated is similar between therapeutic and diagnostic X rays. Electrons emitted from a heated filament are accelerated in an evacuated tube onto a tungsten anode (also called target), producing bremsstrahlung photons with peak energy up to the accelerating voltage. Because of the intended applications, however, therapeutic and diagnostic X-ray tubes have some very different design features as discussed below.

One difference is the target design for heat dissipation. Diagnostic tubes are generally run for shorter times (from a fraction of a second to several seconds) at high current (up to 500 mA) leading to high instantaneous rates of heat production. Tungsten is used as the target material since it has a high melting point (3370 °C). In addition, rotating targets are used in diagnostic tubes to spread the heat deposition over a large area, thereby increasing the maximum permissible loading of the tube. Therapeutic tubes are often run for longer times (several minutes being common) at relatively low currents (5–30 mA). The instantaneous rate of heat production is low for therapeutic tubes, but the total amount of heat generated during a therapy session can be quite high. Therefore, therapeutic tubes use stationary targets that can be cooled by oil or water. The target is often made of tungsten and is mounted on a massive copper stem that serves as a good heat conductor.

The second difference is the target angle, which effectively defines the X-ray source size. Small target angles in the range of 6–20° are used in diagnostic tubes in order to produce small focal spots that will result in a small geometric penumbra of the image. In therapeutic tubes, however, geometric penumbra is not of primary concern. Therefore, therapeutic tubes can use larger target angles in order to produce radiation beams with large field sizes (up to 20 × 20 cm) at target-to-surface distances (SSD) that are usually much shorter than the SSDs employed in diagnostic tubes. The resultant X-ray source size of a therapeutic tube may be as large as 1.0 cm in diameter.

Another difference is the target shielding for secondary electrons. For therapeutic tubes running at an accelerating potential >200 kV, damages to the tube may result if secondary electrons from the target are allowed to reach the glass envelope. The target in a therapeutic tube is often surrounded by a double-layered shield as shown in Fig. 1 (4). The inner layer is made of copper that stops most electrons while its low atomic number ($Z = 29$) minimizes bremsstrahlung production. The outer layer is made of tungsten ($Z = 74$), which absorbs most stray bremsstrahlung photons produced in the copper layer. A thin beryllium ($Z = 4$) window in the shield below the target absorbs most secondary electrons, but allows the useful X-ray beam

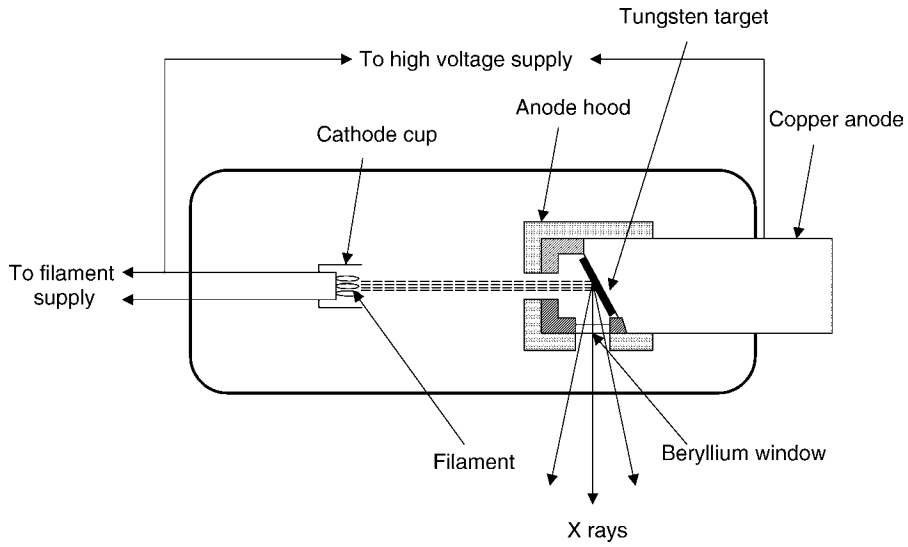


Figure 1. An X-ray tube for radiation therapy constructed with a hooded anode design.

to pass through with little attenuation. This type of shielding design for therapeutic tubes is called a hooded anode. The X-ray tube is enclosed in an oil shield that serves two purposes: (a) to cool the tube and (b) to attenuate stray X rays.

Many major manufacturers have ceased manufacturing kilovoltage X-ray therapy units (5). The three currently available orthovoltage treatment units are Therapax DXT300 manufactured by Pantak, Inc. (East Haven, CN) and Gulmay D3300 and Gulmay D3225 manufactured by Gulmay Medical Ltd. (Mississauga, ON). The same two vendors also produce two superficial therapy units, Therapax HF 150 and Gulmay D2000. Some kilovoltage units are no longer available, but are still in wide use, such as the Philips RT-250 and RT-305 units and the Siemens Stabilipan. The most popular contact unit still in clinical use is the Philips RT-50, formerly manufactured by Philips Medical Systems (Shelton, CN). A new, and novel, contact therapy device available is the Photon Radiosurgery System, Model PRS400, manufactured by Photoelectron Corp. (Lexington, MA).

Orthovoltage Units

Most orthovoltage units operate at 200–300 kV with a tube current of 10–20 mA and have various filters to provide beams with half value layers (HVL) between 1 and 4 mm Cu. The X-ray generators used in these units may be single phase, that is, self-rectified or half-wave rectified, or may employ a Villard-type voltage-doubling circuit or a constant potential circuit (6). The constant potential circuit maintains the accelerating potential at a nearly constant value so that X rays can be produced continuously rather than in pulses. This type of circuit allows for treatments at a shorter treatment time with a higher average X-ray energy.

The energy absorbed by the anode of an X-ray tube is proportional to the product of tube current, accelerating potential, and operating time. For orthovoltage units, which may be operated continuously at high tube voltages, the anode must be cooled efficiently. The copper anodes of orthovoltage units are usually massive to serve as heat conductors. Oil circulated cooling is used instead of water circulated cooling, as shown in Fig. 2 (7), where a hollowed-out anode

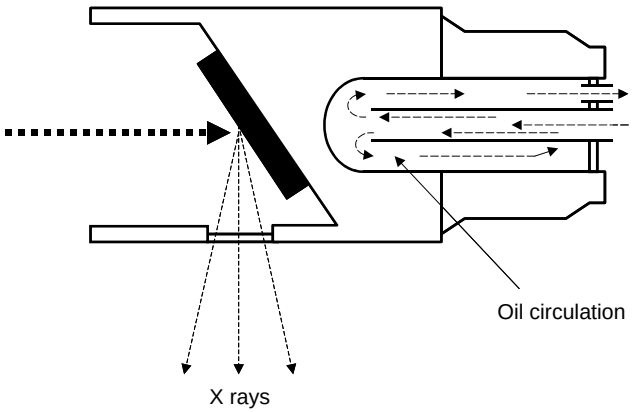


Figure 2. The anode of an orthovoltage X-ray tube. The back of the anode is hollowed out to permit oil circulation.

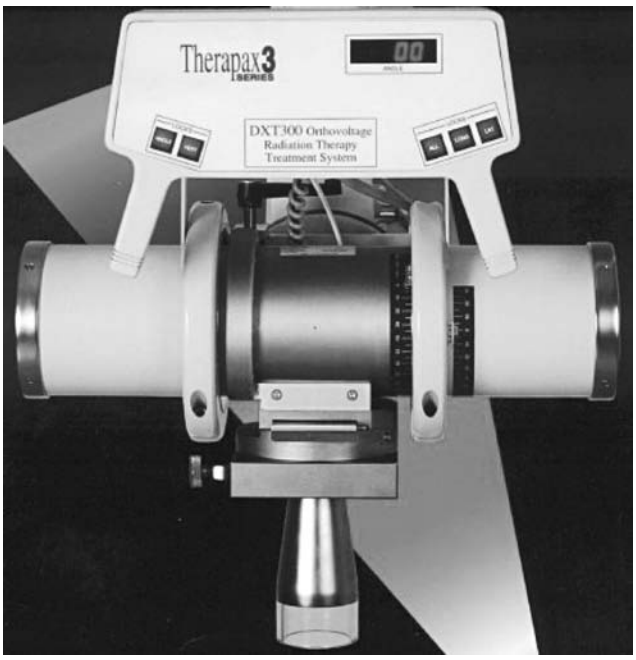


Figure 3. The Therapax DXT300 orthovoltage unit manufactured by Pantak.

provides efficient oil circulation for heat dissipation. The oil then flows through a reservoir where heat is removed by either circulating water or forced-air cooling. Orthovoltage units usually require a warm-up time of several minutes, which allows the tube potential to increase gradually in order to minimize the strain of high voltage on the tube.

Figure 3 shows a currently available microprocessor-controlled orthovoltage treatment unit, Therapax DXT300 (5). This unit features a dual channel dosimetry system in which the primary means of termination of treatment dose can be either a timer or the integrated signal from a transmission ionization chamber. It features a metal ceramic tube, a control panel that incorporates microprocessor technology, a Diamentor dose monitor, and a Cockcroft–Walton multiplier generator that gives a very stable output. The unit can be ceiling or stand mounted. The Therapax DXT300 unit offers energies and filtrations from 300 kV (3 mm Cu) down to 30 kV (0.1 mm Al). The treatment field shape and size can be defined by a variable collimator (rectangular or square fields up to 20 cm $\times$ 20 cm defined at 50 cm SSD) or by fixed square or circular cone-type applicators: open-ended cones (30 cm SSD) or close-ended cones (50 cm SSD).

Superficial Units

Superficial units operate at an accelerating potential 50–150 kV with a tube current 5–15 mA. The X-ray generators in these units are generally half-wave rectified. The X-ray tube has a hooded anode and is enclosed in an oil shield that is constructed with expansion bellows to allow oil to expand as temperature increases. Superficial units may also operate continuously for long treatment times, and therefore require additional cooling in parallel to heat conduction by the copper anode stem and

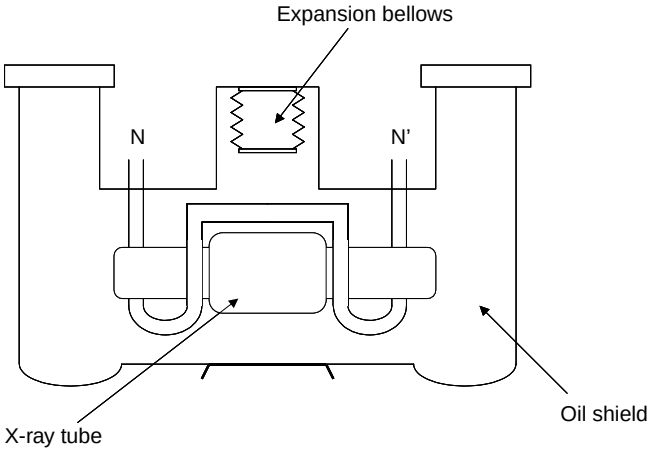


Figure 4. The cooling system for superficial X-ray tubes. Water circulates through the pipe N–N′ to remove the heat from the oil.

convection and absorption by the oil shield. Figure 4 shows a cooling system that employs circulating water to remove heat from the oil continuously (7).

Orthovoltage tubes can also produce X-ray beams in the superficial energy range with matching HVLs. The principal difference between a superficial unit and an orthovoltage unit is that a superficial unit typically has shorter SSD cone applications and, consequently, higher dose rates (5). For example, the Therapax HF150 (Fig. 5) is the companion device to Therapax DXT300 and produces X rays in the range 20–150 kV with HVLs from 0.1 to 10.78 mm Al. The DXT300 can also produce X rays down to 30 kV with an HVL of 0.1 mm Al. The HF150 has 15 cm SSD cone applicators while the DXT300 has 30 and 50 cm SSD cone applicators.

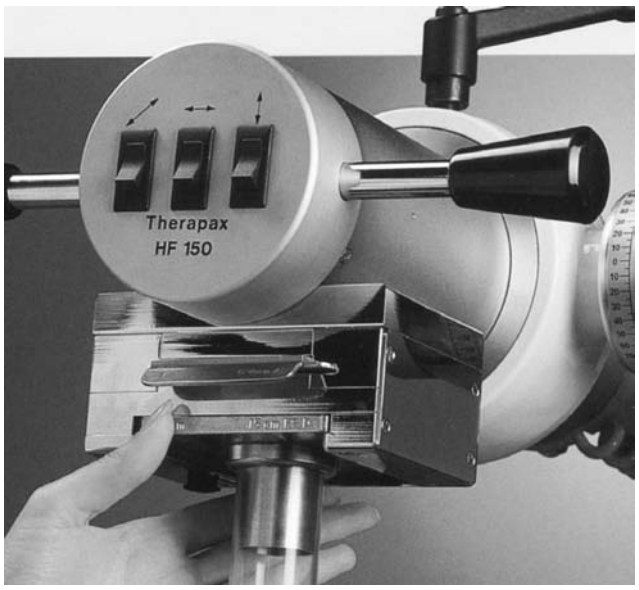


Figure 5. The Therapax HF150 superficial unit manufactured by Pantak.

Contact Units

Both orthovoltage and superficial units can produce X-ray beams in the contact X ray range with complete overlap of available energies and HVLs. The Philips RT-50 unit is strictly a contact therapy unit, operating at energies between 10 and 50 kV. This unit has a slim, long body measuring 43 cm in length and 6 cm in diameter, tapering down to 3 cm at the tip. The SSD is only 4 cm and its dose rate in air is up to 20 $Gy \cdot min^{-1}$. These features make it especially suitable for endocavitary irradiation (e.g., for treating rectal cancers).

The Photon Radiosurgery System (PRS) is a special treatment unit (Fig. 6) consisting of an X-ray source, a control console, and associated calibration devices (8). The X-ray source is small, measuring $17.5 \times 11 \times 7$ cm and weighing 1.6 kg, with X rays being produced at the tip of a 10 cm long, 3.2 mm diameter probe. A plastic, biocompatible sheath covers the probe tip to avoid direct tissue contact when used interstitially. The X-ray energy can be 30, 40, or 50 kV with a tube current of 5, 10, 20, or 40 μA. The whole unit is small enough to be hand-held, though it is designed to work in conjunction with a conventional stereotactic head frame. The system has a built-in internal radiation monitor to measure the output of the beam and to determine the length of the treatment. Additional monitors include a timer and an external radiation monitor that has to be calibrated at the start of each treatment since its output is geometry dependent. Other dosimetry devices are provided to measure the isotropy of the beam, the straightness of the probe, and the relative in-air output of the beam during routine calibration and in the operating room immediately prior to the procedure.

X-RAY BEAM QUALITY

X-ray beams with different energy spectra will differ in their ability to penetrate in a medium such as tissue. The quality of an X-ray beam refers to its penetrating power, with high beam quality indicating deeper penetration. The exact knowledge of beam quality plays an important role in clinical radiotherapy to ensure proper treatment target coverage and adequate skin sparing. It is desirable to use more than one beam quality parameter to specify an X-ray beam (9). The usual quantities used are the accelerating potential and the HVL. However, most dosimetric data for kilovoltage X rays have been given using one beam quality parameter, usually HVL.

The peak voltage across the X-ray tube determines the maximum photon energy in an X-ray spectrum. The accelerating potential is expressed in kilovolts for a kilovoltage X-ray beam to indicate that the X-ray beam actually has an energy spectrum with its maximum energy up to its accelerating potential. Higher accelerating potentials produce more penetrating X-rays and, therefore have a direct effect on the beam quality.

The HVL is defined as the thickness of an absorber that reduces the air kerma rate of a narrow unidirectional X-ray beam at a point distant from the absorber to 50% of that of the nonattenuated beam (9). For orthovoltage X rays, both aluminum (Al) and copper (Cu) can be used as the

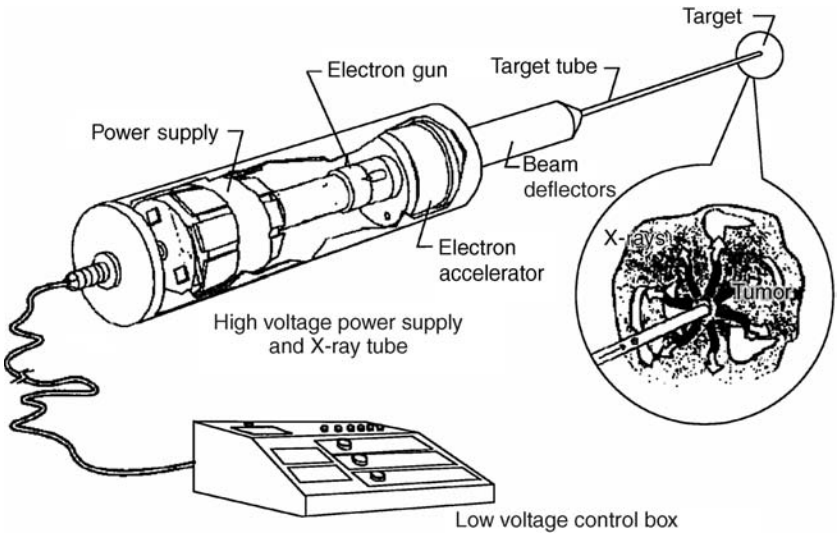


Figure 6. A schematic diagram showing the Photon Radiosurgery System for interstitial treatment.

absorbing material for the determination of HVL although Cu is more frequently used. The beam quality is expressed in mm Al or mm Cu. For superficial and contact X-ray beams, only mm Al is used to express HVL.

The HVL is measured using an ionization chamber with a narrow beam setup, in which a collimating aperture is placed half-way between the X-ray target and the measurement chamber. The field size is reduced just enough to encompass the chamber. The absorbing materials are placed at least 50 cm from the measurement chamber and close to the collimating aperture. The absorber should be made of high purity material and the thickness of the absorber should be measured accurately. The thickness that reduces the air kerma value to one-half is obtained by interpolation. A monitor detector is used to correct fluctuations of air kerma rate. The measurement chamber must have limited beam quality dependence (within 5% between 40–300 kV) for accurate HVL measurements. Thin-walled chambers are used for lightly filtered beams to measure the low energy X-ray components accurately (HVL errors of up to 10% were observed for a lightly filtered 100 kV beam using a Farmer-type cylindrical chamber).

DOSIMETRY CALIBRATION

Methodology

For kilovoltage X-ray beams, the absorbed dose to water is usually determined with an ionization chamber calibrated in air in terms of either exposure or air kerma. The commonly used ionization chambers are generally considered to be "photon detectors" as the well-known Bragg–Gray cavity theory no longer applies to this energy range (10). To determine the dose to water, the measured air kerma is converted to water kerma using the ratio of mass energy absorption coefficients for water to air, evaluated using the energy spectrum at the position of interest. The dose to water is considered to be the same as water kerma for this energy range due to the negligible difference between kerma and collision kerma (electrons have very small ranges) assuming that quasicharged particle equilibrium exists. This is generally true for measurement points at depths equal to or greater than the depth of maximum dose for kilovoltage X rays.

Two different methods have been recommended for kilovoltage X-ray dosimetry calibration by national and international dosimetry protocols (11–14). The in-air method measures the air kerma free in air and then converts it to dose to water with the ratio of mass energy absorption coefficients for water to air and a backscatter factor. The in-phantom method measures the air kerma at the reference depth in water under reference conditions and then converts it to the absorbed dose at the depth of the center of the chamber in undisturbed water using the ratio of mass energy absorption coefficients for water to air and other beam quality- and chamber-dependent correction factors.

The ICRU Report 23 was the first to recommend the in-phantom method for orthovoltage X-ray beams since it was difficult to make accurate measurements in regions at or close to the surface of a phantom, and the dose distribution there, unlike at greater depths, was considerably affected by the details of the beam collimation system (11). For this reason, the *British Journal of Radiology Supplement 11* gave two distinct sets of depth-dose tables for close-ended applicators and open diaphragms, respectively (15). It was also clear that by normalizing the depth-dose curves at a depth rather than at the surface the differences in the recommended depth-dose curves would be virtually eliminated. The dose values at greater depths were more clinically relevant since orthovoltage X-ray beams were primarily used for treating deep-seated tumors in the 1970s.

Several more recent dosimetry protocols still recommended the in-phantom method for reference dosimetry for orthovoltage X-ray beams, but used a reference depth of 2 rather than 5 cm (12–14). This was aimed at reducing the measurement uncertainty at lower X-ray energies. However, the in-air method has been the commonly used method for the whole kilovoltage energy range in clinical radiotherapy, especially in North America (16). This may be explained by the fact that nowadays orthovoltage beams

are mainly used for treating tumors close to the skin. The primary point of interest is the dose near the surface rather than at greater depths. Another reason is that it is more convenient to perform routine beam calibration free in air than in a water phantom.

The American Association of Physicists in Medicine (AAPM) recommended that both methods be used for absorbed dose determination for orthovoltage X-ray beams depending on the measurement point of interest (9). To improve measurement accuracy, the in-air method should be used if the primary point of interest is at the phantom surface. On the other hand, if one is more interested in the dose at large depths than at the surface, the in-phantom method should be used. Better agreement in measured PDD curves at depth = 1 cm and greater can be achieved when they are normalized to the values at the 2 cm reference depth than normalized to the surface values. The AAPM recommended that only the in-air method be used for superficial X-ray beam dosimetry (9).

Measurement Equipment

Phantoms. Water is the recommended phantom material when the in-phantom method is used in orthovoltage X-ray beam dosimetry (9). Conversion factors are used to convert from the air kerma in the sensitive chamber cavity to the dose to water. Plastic phantoms are not recommended for in-phantom reference dosimetry for orthovoltage X rays as the chamber correction factors and conversion factors to derive dose at a depth in water for these phantoms are poorly known. Some water equivalent plastics are commercially available for orthovoltage X-ray measurements. However, their properties must be investigated before they can be used as water substitutes.

Dosimeters. Air-filled ionization chambers are recommended for kilovoltage X-ray beam reference dosimetry (9). Measurements for orthovoltage X rays are performed either free in air (in cases where the surface dose is the primary concern) or at a 2 cm depth in water (dose at greater depths is the primary concern). Cylindrical chambers that have a calibration factor varying with the beam quality by <3% in this energy range are recommended for the beam output measurement. For superficial X rays, only the in-air method is used. Cylindrical chambers that have a calibration factor varying with the beam quality by <5% between 50 and 150 kV are recommended. Extensive studies have been carried out on the correction factors for the commonly used Farmer-type chambers for the in-phantom measurement (16–21). Other cylindrical chambers may also be used. However, correction factors must then be determined by comparing these chambers with a chamber with known correction factors. If measurements are performed in water, a thin waterproofing sleeve should be used and appropriate correction factors should be applied depending on the material and thickness used. If measurements are made in air, the thimbles of cylindrical chambers are thick enough, so no build-up cap is required. For X-ray energies <70 kV, calibrated parallel-plate chambers with a thin entrance window are recommended. Thin plastic build-up foils should be added to the entrance window, if necessary, to provide full electron buildup and to eliminate electron contamination (see Table 1). All measurements should be corrected for temperature, pressure, ion recombination, and polarity effect.

Table 1. Thickness of Build-Up Material for Thin Window Parallel-Plate Chambers Used for In-Air Calibrations with X-Ray Energy <100 kV

Generating Potential, kV	Thickness of Foil, mg cm^{-2}
50	1.5
60	3.0
70	4.7
80	6.6
90	8.7
100	10.9

Electrometers. Electrometers for kilovoltage X-ray therapy measurements should be capable of reading currents on the order of 0.1 nA, with an accumulated charge of 50–100 nC. They are calibrated by a standards laboratory with proper correction factors applied in the measurement. These correction factors are generally close to 1.000; but can occasionally be 5% different from 1.000. Electrometers and ionization chambers can be calibrated either together or sometimes separately and if so their calibration factors must be combined.

Detector Calibration

Modern dosimetry protocols for kilovoltage X rays are based on the air kerma, K_{air}, which is defined as the kinetic energy, ΔE, transferred from X rays to charged particles per unit air mass, Δm (i.e., $K_{\text{air}} = \Delta E/\Delta m$). Assuming that K_{air} is the air kerma at the reference point in air for a given beam quality and M the reading (corrected for temperature, pressure, recombination, and polarity effect) of an ionization chamber to be calibrated with its center at the same point, the air kerma calibration factor, N_K, for this chamber at the specified beam quality is defined as

$$N_K = \frac{K_{\text{air}}}{M} \tag{1}$$

Previous dosimetry protocols for kilovoltage X rays were based on the exposure, X, which is defined as the total electric charge, ΔQ, of all ions of one sign (e.g., electrons), produced in air by X rays in a unit mass of air, Δm (i.e., $X = \Delta Q/\Delta m$). A similar equation can be used to derive the exposure calibration factor, N_X:

$$N_X = \frac{X}{M} \tag{2}$$

The relation between the air kerma calibration factor and the previous exposure calibration factor is given by

$$N_K = N_X \left(\frac{W}{e}\right)_{\text{air}} (1-g)^{-1} \tag{3}$$

where $(W/e)_{\text{air}}$ has the value 33.97 J/C (or 0.876 cGy·R^{-1}) for dry air, $(1\,g)$ corrects for the effect of radiative losses by charged particles to bremsstrahlung photons (g is the

fractional energy lost to bremsstrahlung photons in air, which is practically zero for X-ray beams below 300 kV).

The N_K or N_X factors can be obtained from a standards laboratory for a number of available X-ray beam qualities to match the beam energies in a clinic. In the calculation of dose to a patient for a kilovoltage X-ray beam, the accuracy of the chamber calibration determines the final accuracy. Each chamber involved in reference dosimetry of the clinical kilovoltage setup should be calibrated, since there can be up to 8% differences between the calibration factor, N_K, for different chambers of the same type. In addition, each chamber should be calibrated at a number of beam qualities to determine the energy dependence of its response. One generally expects this energy dependence to be at most 6% between 40 and 300 kV for well-designed chambers.

Determination of Absorbed Dose

Formalism for the In-Air Method. The in-air method can be used for both superficial and orthovoltage X-ray dose determination. If the point of interest is at the phantom surface the in-air method is more accurate than the in-phantom method. In this method, the measurement is performed with the chamber free in air and the dose to water at the phantom surface (theoretically, this only applies for depths beyond the range of the contaminant electrons and where quasicharged particle equilibrium has been established) can be calculated by

$$D_{\mathrm{w}} = MN_{\mathrm{K}}\left(\frac{\overline{\mu}_{\mathrm{en}}}{\rho}\right)^{\mathrm{w}}_{\mathrm{air}} B_{\mathrm{w}} P_{\mathrm{stem,air}} \tag{4}$$

where M is the corrected ionization chamber reading, N_{K} the chamber air kerma calibration factor at the user's beam quality, $(\overline{\mu}_{\mathrm{en}}/\rho)^{\mathrm{w}}_{\mathrm{air}}$ the ratio of mass energy absorption coefficients for water to air, evaluated over the X-ray energy spectrum free-in-air, in the absence of a phantom, B_{w} the backscatter factor that accounts for the effect of the phantom (water) scatter, and $P_{\mathrm{stem,air}}$ the chamber stem correction factor that accounts for the change in photon scatter from the chamber stem between the calibration and measurement (mainly due to the change in field size). Here, $(\overline{\mu}_{\mathrm{en}}/\rho)^{\mathrm{w}}_{\mathrm{air}}$ is independent of the field size used since it is evaluated over the primary beam only (22). The AAPM recommended values of $(\overline{\mu}_{\mathrm{en}}/\rho)^{\mathrm{w}}_{\mathrm{air}}$ and B_{w} for some typical beam qualities are given in Tables 2 and 3 (taken from Ref. (9)). The parameter $P_{\mathrm{stem,air}}$ is taken as unity for a calibrated chamber if, for a given beam quality, the same field size is used in the calibration and the measurement. Otherwise, the $P_{\mathrm{stem,air}}$ factor can be measured by intercomparing the chamber with unknown $P_{\mathrm{stem,air}}$ with a reference chamber for which $P_{\mathrm{stem,air}}$ is known (9). A Farmer-type cylindrical chamber with flat response can be used as a reference chamber for the measurement of $P_{\mathrm{stem,\ air}}$ of another chamber since its stem effect varies little with field size (<1%).

Formalism for the In-Phantom Method. For orthovoltage X rays, the in-phantom method is recommended (9) if the point of interest is at a depth in the patient. For the 2 cm reference depth in water the dose to water for orthovoltage X rays can be calculated using

$$D_{\mathrm{w}} = MN_{\mathrm{K}}P_{\mathrm{Q,cham}}P_{\mathrm{sheath}}\left(\frac{\overline{\mu}_{\mathrm{en}}}{\rho}\right)^{\mathrm{w}}_{\mathrm{air}} \tag{5}$$

where M is the corrected ionization chamber reading, N_{K} the chamber air kerma calibration factor at the user's beam quality, and $(\overline{\mu}_{\mathrm{en}}/\rho)^{\mathrm{w}}_{\mathrm{air}}$ the ratio of mass energy absorption coefficients for water to air, evaluated over the X-ray energy fluence at 2 cm depth in water, in the absence of the chamber. The overall chamber correction factor $P_{\mathrm{Q,\ cham}}$accounts for the effect of the change in chamber response due to photon energy and angular variation between chamber calibration in air and measurement in water, the effect of chamber stem between calibration in air and measurement in water and the effect of displacement of water by the chamber in the measurement in water. The sheath correction factor P_{sheath} is needed when a waterproofing sheath is used (23), which is not directly related to the individual chamber type. The values of $P_{\mathrm{Q,\ cham}}$, P_{sheath} and $(\overline{\mu}_{\mathrm{en}}/\rho)^{\mathrm{w}}_{\mathrm{air}}$ for some typical beam qualities are given in Tables 4–6. The values were taken from Ref. 9.

Table 2. Ratios of Average Mass Energy Absorption Coefficients for Water to Air to Convert Air Kerma to Water Kerma, Free-in-air for Both Superficial and Orthovoltage X Rays

HVL		
mm Al	mm Cu	$(\overline{\mu}_{\mathrm{en}}/\rho)^{\mathrm{w}}_{\mathrm{air}}$ (free-in-air)
0.03		1.047
0.05		1.046
0.08		1.044
0.10		1.044
0.3		1.035
0.5		1.028
0.8		1.022
1.0		1.020
3.0		1.021
5.0		1.029
8.0		1.045
	0.1	1.020
	0.3	1.035
	0.5	1.050
	0.8	1.068
	1.0	1.076
	3.0	1.100
	5.0	1.109

CLINICAL DOSIMETRY

Percentage Depth Dose

Percentage depth dose (PDD) and lateral dose profiles are difficult to measure for kilovoltage X rays because of the significant variation of X-ray energy (especially for lightly filtered beams) and angular distribution with depth and field size. Detectors for the PDD and dose profile measurement should have high spatial resolution and constant energy and angular response. Solid detectors, attractive because of the small size of their sensitive volume (diode detector, TLD, film), usually show significant beam quality

Table 3a. Water Kerma-Based Backscatter Factors, B_w, for Different Field Diameters (d) and SSD for Superficial X Rays

		HVL, mm Al										
SSD (cm)	d (cm)	0.03	0.05	0.08	0.1	0.3	0.5	0.8	1.0	3.0	5.0	8.0
20	1	1.005	1.007	1.011	1.014	1.028	1.036	1.043	1.046	1.061	1.058	1.053
	3	1.005	1.008	1.014	1.019	1.049	1.069	1.092	1.105	1.158	1.165	1.158
	5	1.005	1.008	1.014	1.019	1.054	1.080	1.112	1.131	1.215	1.234	1.236
	10	1.005	1.008	1.014	1.019	1.057	1.088	1.129	1.155	1.291	1.334	1.354
	15	1.006	1.008	1.014	1.019	1.058	1.090	1.133	1.162	1.321	1.380	1.414
	20	1.006	1.008	1.014	1.019	1.058	1.091	1.136	1.165	1.334	1.402	1.444
30	1	1.005	1.007	1.011	1.015	1.027	1.035	1.043	1.047	1.063	1.059	1.053
	3	1.005	1.008	1.014	1.019	1.048	1.069	1.093	1.107	1.164	1.168	1.158
	5	1.005	1.008	1.014	1.019	1.053	1.079	1.111	1.130	1.221	1.242	1.237
	10	1.006	1.008	1.014	1.019	1.057	1.088	1.130	1.157	1.298	1.350	1.367
	15	1.006	1.008	1.014	1.019	1.058	1.091	1.136	1.165	1.332	1.403	1.434
	20	1.006	1.008	1.014	1.019	1.058	1.091	1.138	1.169	1.350	1.428	1.472
50	1	1.005	1.007	1.011	1.014	1.027	1.035	1.042	1.045	1.065	1.059	1.052
	3	1.005	1.007	1.013	1.018	1.049	1.070	1.093	1.106	1.163	1.169	1.160
	5	1.005	1.007	1.013	1.018	1.054	1.081	1.113	1.132	1.226	1.241	1.242
	10	1.006	1.007	1.013	1.018	1.057	1.091	1.134	1.159	1.309	1.352	1.375
	15	1.006	1.007	1.013	1.018	1.058	1.093	1.140	1.169	1.346	1.411	1.448
	20	1.006	1.007	1.013	1.018	1.058	1.094	1.142	1.173	1.363	1.443	1.493
100	1	1.005	1.007	1.011	1.014	1.028	1.036	1.042	1.044	1.062	1.059	1.053
	3	1.005	1.008	1.014	1.019	1.050	1.070	1.092	1.104	1.163	1.169	1.162
	5	1.006	1.008	1.014	1.019	1.055	1.082	1.113	1.131	1.225	1.240	1.243
	10	1.006	1.008	1.014	1.019	1.058	1.091	1.134	1.158	1.311	1.351	1.381
	15	1.006	1.008	1.014	1.019	1.059	1.094	1.140	1.169	1.354	1.417	1.460
	20	1.006	1.008	1.014	1.019	1.059	1.095	1.143	1.172	1.375	1.451	1.508

dependence or large experimental uncertainties. Well-designed cylindrical chambers have nearly constant energy response in this beam quality range and are suitable for in-phantom measurements. However, the measurement depth is limited to no less than the outer radius of the chamber. Parallel-plate chambers have been used for measurements at smaller depths. Those chambers designed for electron beams usually have a calibration

Table 3b. Water Kerma-Based Backscatter Factors, B_w, for Different Field Diameters (d) and SSD for Orthovoltage X rays

		HVL, mmCu						
SSD (cm)	d (cm)	0.1	0.3	0.5	0.8	1.0	3.0	5.0
20	1	1.061	1.055	1.053	1.048	1.045	1.024	1.018
	3	1.158	1.168	1.155	1.147	1.140	1.082	1.057
	5	1.214	1.242	1.233	1.219	1.209	1.127	1.088
	10	1.290	1.352	1.353	1.339	1.326	1.204	1.141
	15	1.320	1.407	1.415	1.403	1.389	1.251	1.174
	20	1.333	1.434	1.447	1.436	1.421	1.278	1.194
30	1	1.063	1.056	1.052	1.047	1.044	1.024	1.018
	3	1.164	1.169	1.155	1.146	1.139	1.084	1.055
	5	1.220	1.242	1.235	1.221	1.211	1.130	1.087
	10	1.297	1.363	1.367	1.347	1.332	1.214	1.147
	15	1.330	1.417	1.438	1.422	1.405	1.270	1.189
	20	1.348	1.446	1.478	1.464	1.446	1.302	1.213
50	1	1.065	1.054	1.052	1.047	1.045	1.025	1.018
	3	1.163	1.170	1.157	1.148	1.140	1.084	1.057
	5	1.225	1.247	1.240	1.226	1.214	1.131	1.089
	10	1.308	1.367	1.376	1.360	1.344	1.222	1.152
	15	1.345	1.433	1.452	1.446	1.428	1.285	1.195
	20	1.361	1.471	1.499	1.495	1.478	1.325	1.226
100	1	1.062	1.055	1.052	1.047	1.045	1.025	1.018
	3	1.163	1.170	1.160	1.150	1.142	1.085	1.057
	5	1.224	1.245	1.241	1.227	1.217	1.132	1.090
	10	1.310	1.370	1.383	1.369	1.353	1.226	1.155
	15	1.353	1.447	1.463	1.458	1.441	1.291	1.204
	20	1.373	1.490	1.513	1.516	1.499	1.334	1.237

Table 4. Ratio of Average Mass Energy Absorption Coefficients for Water to Air at 2 cm Depth in Water for a 100 cm^2 Field Defined at 50 cm SSD for Orthovoltage X rays

HVL		
mm Cu	mm Al	$(\bar{\mu}_{en}/\rho)^{w}_{air}$
0.1	2.9	1.026
0.3	6.3	1.037
0.5	8.5	1.046
0.8	10.8	1.055
1.0	12.0	1.060
2.0	15.8	1.081
3.0	17.9	1.094
4.0	19.3	1.101
5.0	20.3	1.105

factor varying with beam quality by 20–40% for kilovoltage X rays. Significant corrections with depth may be required for the PDD measurement with these chambers. Specifically designed parallel-plate chambers for low energy X rays usually have a flat energy response in air but not at a depth in a phantom. For example, >10% variations in chamber response have been observed for the Capintec PS-033 chambers. Thus, a depth-related correction factor may be required for these chambers to be used in accurate PDD measurements.

The suitability of X-ray detectors for relative dosimetry measurement has been evaluated extensively (16,24,25). As a general requirement for the evaluation of a specific detector, the relative in-air chamber response and the relative in-phantom response should be compared with a well-behaved cylindrical chamber at depths, where reasonable measurements with the cylindrical chamber can be performed. Diamond detectors and the NACP parallel plate chamber (type) have been found to require relatively small depth-dependent correction factors for orthovoltage X-ray beams (16).

Care must be exercised in the measurement of PDD and dose profiles for kilovoltage X-ray beams because of their significant depth and field size dependence. A water tank with a small-volume scanning ionization chamber is ideal for the PDD and profile measurement. A monitor chamber is often needed to eliminate the effect of erratic fluctuations in dose rate. If a thin window parallel plate chamber is used, the chamber must have sufficient buildup material placed over its entrance window (see Table 1). Because of the finite size of the ionization chamber in the beam direction, it is necessary to extrapolate the dose to the surface. Since the dose distribution may be nonlinear near the phantom surface, caution should be exercised in extrapolating over the last few millimeters (24,25).

If desired, the depth-dose distribution can be measured by placing thin sheets of water-equivalent material over

Table 5. Overall Chamber Correction Factors $P_{Q,cham}$ for Commonly used Cylindrical Chambers in Orthovoltage X-Ray Beams[a]

Chamber Type HVL, mm Cu	NE2571	Capintec PR06C	PTW N300 01	NE2611 or NE2561
0.10	1.008	0.992	1.004	0.995
0.30	1.023	1.008	1.021	1.017
0.50	1.025	1.010	1.023	1.019
0.80	1.024	1.010	1.022	1.018
1.0	1.023	1.010	1.021	1.017
2.0	1.016	1.007	1.015	1.011
3.0	1.009	1.005	1.010	1.006
4.0	1.004	1.003	1.006	1.003
5.0	1.002	1.001	1.002	1.001

[a]The data applies to the in-phantom method for a chamber at 2 cm depth in water and a 100 cm^2 field.

Table 6. Correction Factors for PMMA Sheaths When Using Cylindrical Chambers for In-Water Measurements in Orthovoltage X-Ray Beams[a]

HVL		PMMA (Lucite)			
(mm Cu)	(mm Al)	t = 0.5 mm	t = 1 mm	t = 2 mm	t = 3 mm
0.1	3.0	0.998	0.995	0.991	0.986
0.3	6.1	0.998	0.997	0.994	0.991
0.5	8.5	0.999	0.998	0.996	0.994
0.8	11.0	0.999	0.998	0.997	0.996
1.0	12.1	1.000	0.999	0.998	0.997
2.0	15.2	1.000	1.000	1.000	0.999
3.0	17.6	1.000	1.000	1.000	1.000
4.0	19.4	1.000	1.000	1.000	1.000
5.0	20.9	1.000	1.000	1.000	1.000

[a]The data applies to 2 cm depth in water and a 100 cm^2 field.

the chamber in a phantom and moving the chamber back by the same amount to maintain a constant SSD. The water-equivalence of the material in the energy range of interest must be verified. The Poly (methyl methacrylate) (PMMA) is not suitable for this purpose. Strictly speaking, an ion chamber measures the depth-ionization distribution rather than the depth-dose distribution. However, the difference between them is small (16). If a suitable detector for relative dosimetry cannot be identified in the clinic it is recommended to use the data from the *British Journal of Radiology Supplement 25* (26) or published data that match the kV and HVL values of the user's beams (27–31).

Output Factors

Output factors are required for all combinations of SSD and field size used for kilovoltage X-ray treatment. The output factor is defined as the surface dose value for a given SSD and field size relative to that under the reference conditions. Since the scatter contribution from the inside of a cone applicator may be significant, it is not sufficiently accurate to estimate output factors for different applicators using the ratio of the backscatter factors corresponding to the respective field sizes. The output factor for each individual applicator must be measured at each beam quality. If the in-phantom calibration method has been used for orthovoltage X rays it is necessary to obtain the PDD in order to determine the dose at the surface. Note that this may result in large uncertainties in the output factors since the uncertainties are generally high in the PDD values near the surface.

Isodose Curves

Isodose curves are obtained by joining together the points, which have the same dose values. This can be done using the depth dose and lateral profile data if they already exist. A water tank with a small volume ionization chamber is recommended for the dose distribution measurement. A monitor chamber is needed to eliminate the effect of dose rate fluctuations. Dose values at various depths across the beam are measured and the points of equal dose are connected to give the isodose curves. Published dose distributions may be used as reference if a suitable detector for relative dosimetry cannot be identified in the clinic. However, the published data must have the same kV and HVL values as those of the clinical beams.

Quality Assurance

A well-established quality assurance (QA) program will ensure the safe and accurate delivery of radiation therapy treatments using kilovoltage X rays. The safe delivery of radiation therapy is ensured by machine interlocks and strategically placed emergency-off buttons. Accurate delivery of radiation treatment is ensured by maintaining the mechanical accuracy within the specification of treatment unit, maintaining the beam quality in its original condition, and maintaining the accuracy of dosimetry calculation and measurements for treatment planning and plan verification. All dosimetry data and mechanical limits are established during the initial machine commissioning period before the system is first used for treating a patient and thereafter annually, or after any change which may significantly alter the dosimetry data and the mechanical limits. Once the baseline is established, maintaining the baseline data becomes the mission of the dosimetry QA program. Documentation for each full yearly calibration is maintained for 5 years after the completion of calibration. Documentation of weekly and monthly spot check measurements is maintained for 2 years.

The test frequency of each mechanical component and dosimetry data for a treatment machine is determined by the significance of the tests that indicate beam dosimetry changes and mechanical tolerance changes within the limited amount of time. Unlike clinical linear accelerators, the number of combinations of energy and filter for a kilovoltage X-ray machine are high, but only a few combinations are used on a routine basis. The number of patients that are treated using kilovoltage X-ray beams are much less than for megavoltage radiotherapy. Therefore, the check of the beam dosimetry should depend on the frequency of the beam usage. The dosimetry QA items and their frequencies are summarized below:

Daily Checks.

- Beam output constancy for energy and filter combinations in use.
- Functionality of the audiovisual monitor.
- Door and energy interlock circuits and emergency stops.

Monthly Checks.

- Items included in the daily checks.
- Beam flatness and symmetry.
- Timer operation.
- Light-radiation congruence.

Annual Checks.

- Items included in the monthly checks.
- Dose rate for all energy and filter combinations.
- Output factors for each of the applicator (cone).
- Timer accuracy (verification of the timer error).
- Accuracy of the light localizer system.
- Accuracy of distance measuring devices.
- Beam quality.
- Accuracy of depth dose data and isodose charts.
- Accuracy of field size dependence data.
- Agreement of dose rate with distance from target.
- Attenuation in lead for patient block thickness.

BIBLIOGRAPHY

1. Glasser O, editor. The Science of Radiology. London: Bailliere; 1933.

2. Voltz F. Dosage Tables for Deep-Therapy. London: Heinemann; 1922.
3. Paterson R. The Treatment of Malignant Disease by Radium and X-Rays. Baltimore: Williams and Wilkins; 1949.
4. Philips Laboratories, X ray Research. Holland: PL; 1937–1950.
5. Van Dyk J, editor. The Modern Technology of Radiation Therapy. Madison, WI: Medical Physics Publishing; 1999.
6. Jaundrell-Thompson F, Ashworth WJ. X ray Physics and Equipment. Philadelphia: Davis; 1965.
7. Moran EF. Roentgen rays: Generators. In: Glasser O, editor. Medical Physics. Vol. 1, Chicago, IL: Yearbook Publishers; 1944.
8. Dinsmore M, et al. A new miniature X-ray source for interstitial radiosurgery: Device description. Med Phys 1996;23:45–52.
9. Ma C-M, et al. AAPM protocol for 40-300kV x-ray beam dosimetry for radiotherapy and radiobiology. Med Phys 2001;28:868–893.
10. Ma C-M, Nahum AE. Bragg-Gray theory and ion chamber dosimetry in photon beams. Phys Med Biol 1991;36:413–428.
11. ICRU (International Commission on Radiation Units and Measurements). Radiation Dosimetry: Measurement of Absorbed Dose in a Phantom Irradiated by a Single Beam of X- or Gamma Rays. ICRU Report 23. Washington (D.C.): ICRU; 1973.
12. IAEA (International Atomic Energy Agency). Absorbed Dose Determination in Photon and Electron Beams; An International Code of Practice, Vol. 277 of Technical Report Series, Vienna, Austria: IAEA, 1987.
13. Klevenhagen SC, et al. The IPEMB code of practice for the determination of absorbed dose for x-rays below 300 kV generating potential (0.035 mm Al-4mm Cu HVL; 10-300 kV generating potential). Phys Med Biol 1996;41:2605–2625.
14. NCS (Netherlands Commission on Radiation Dosimetry). Dosimetry of Low and Medium Energy X-rays, a Code of Practice for Use in Radiotherapy and Radiobiology. NCS Report 10. Delft, The Netherlands: NCS; 1997.
15. Central axis depth dose data for use in radiotherapy. Br J Radiol (Suppl) 1972; 11.
16. Ma C-M, Seuntjens JP, editors. Kilovoltage X ray Beam Dosimetry for Radiotherapy. Madison (WI): MPP; 1999.
17. Seuntjens JP, Thierens H, Schneider U. Correction factors for a cylindrical chamber used in medium energy x-ray beams. Phys Med Biol 1993;38:805–832.
18. Ma C-M, Nahum AE. Monte Carlo calculated stem effect corrections for NE2561 and NE2571 chambers in medium-energy x-ray beams. Phys Med Biol 1995;40:63–72.
19. Ma C-M, Nahum AE. Calculations of ion chamber displacement effect corrections for medium-energy x-ray dosimetry. Phys Med Biol 1995;40:45–62.
20. Seuntjens JP, Verhaegen F. Dependence of overall correction factor of a cylindrical ionization chamber on field size and depth in medium energy X ray beams. Med Phys 1996;23:1789–1796.
21. Seuntjens J, et al. Determination of absorbed dose to water with ionisation chambers calibrated in free air for medium energy X rays. Phys Med Biol 1988;33:1171–1185.
22. Nahum AE, Knight RT. Consistent formalism for kV x-ray dosimetry. Proceedings of the IAEA International Symposium on Measurement Assurance in Dosimetry. Vienna, Austria: IAEA; 1994. p 451–459.
23. Ma C-M, Seuntjens JP. Correction factors for water-proofing sleeves in kilovoltage x-ray beams. Med Phys 1997;24:1507–1513.
24. Li XA, Ma C-M, Salhani D. Measurement of percentage depth dose and lateral beam profile for kilovoltage x-ary therapy beams. Phys Med Biol 1997;42:2561–68.
25. Ma C-M, Li XA, Seuntjens J. Consistency study on kilovoltage x-ray beam dosimetry for radiotherapy. Med Phys 1998;25: 2376–2384.
26. Central axis depth dose data for use in radiotherapy. Br J Radiol (Suppl) 1996; 25.
27. Podgorsak EB, Gosselin M, Evans MDC. Superficial and orthovotage x-ray beam dosimetry. Med Phys 1998;25:1206–1211.
28. Gerig L, Soubra M, Salhani D. Beam Characteristics of the Therapax DXT-300 orthovoltage therapy unit. Phys Med Biol 1994;39:1377–1392.
29. Butson MJ, Mathur J, Metcalfe P. Dose characteristics of a new 300 kVp orthovoltage machine. Aust Phys Eng Sci Med 1995;18:133–138.
30. Aukett RJ, Thomas DW, Seaby AW, Gittins JT. Performance characteristics of the Pantak DXT-300 kilovoltage X ray treatment machine. Br J Radiol 1996;69:726–734.
31. Kurup RG, Glasgow GP. Dosimetry of a kilovoltage radiotherapy x-ray machine. Med Dosimetry 1993;18:179–86.

See also COBALT 60 UNITS FOR RADIOTHERAPY; RADIATION DOSIMETRY FOR ONCOLOGY; RADIOTHERAPY, INTRAOPERATIVE.

X-RAYS: INTERACTION WITH MATTER

ANIL SETHI
Department of Radiation Oncology Loyola University Medical Center
Maywood, Illinois

INTRODUCTION

X-ray interaction with matter forms the backbone of medical physics. The basic interaction and its applications can be found in many subfields of medical physics. Some of the more common examples are applications in diagnosis and treatment of cancer, shielding design of accelerators, radiation detectors, biological response to radiation, and radiation protection. A thorough understanding of the interaction process is therefore mandatory to appreciate many exciting phenomena encountered in medical physics. Although the title of this article refers to X-ray interaction with a medium, the physics of the process applies equally well to gamma (γ) rays. Furthermore, in this article, all electromagnetic radiation is referred to as photons regardless of whether their origin is atomic (X rays) or nuclear (γ rays). How a given photon will interact with matter depends on the energy of the photon and not on its birthplace.

When a beam of X rays passes through the medium, it undergoes attenuation or loss of intensity. In other words, some photons are removed from the beam. This attenuation may be due to either absorption or scattering of photons by the medium. In absorption, the energy of the photon is completely transferred to the atoms, whereas in scattering, the X-ray beam undergoes a change in direction that may be accompanied by a change in its energy. In both absorption and scattering, the net result is the transfer of energy to the medium. The amount of energy deposited per unit mass of the medium is called dose (measured in the SI

unit, gray, 1 Gy = 100 cGy), which is an important quantity in radiotherapy applications. How much energy will be transferred to the medium depends on the medium composition and the energy of X rays.

X-ray energy transfer to the medium is a two-step process. First, incident photons interact with the medium and release electrons. Next, in traveling through the medium, these electrons lose their energy via excitation or ionization of atoms. Therefore, photons are labeled as "indirectly ionizing" radiation, in contrast to charged particles, such as, electrons that are "directly ionizing." It is the latter that act as a vehicle for photon energy transfer to the medium. If the electrons have sufficient energy, they may also eject secondary electrons from atoms that form their own tracks known as δ rays.

MODES OF PHOTON INTERACTION

A photon can interact with the medium in any one of several competing but independent processes. The main modes of interaction are as follows:

1. Coherent (or Rayleigh) scattering
2. Photoelectric effect
3. Compton scattering
4. Pair production or triplet production
5. Photodisintegration

Each of the above processes is characterized by the photon interacting with a different subatomic particle in the medium. For example, the photon interaction may be with the entire atom (as in photoelectric effect, coherent scattering), atomic electrons (Compton scattering), atomic nucleus (photodisintegration), electric field of electrons (triplet production), or nuclear field (pair production). As a result, the photon may undergo elastic (coherent) scattering, inelastic (incoherent) scattering, or complete absorption. For a beam of photons emanating from a linear accelerator and having a mixture of energies (polyenergetic beam), all of these interactions may be taking place simultaneously. Which interaction process will dominate and govern the fate of most photons in the beam depends on the energy of photons and the type (atomic number, Z) of attenuating material.

The interaction processes listed above are in an increasing order of importance as the photon energy increases. For example, coherent scattering is the most important interaction at very low photon energies, whereas photodisintegration occurs only at very high photon energies. In the energy domain most common in medical physics (few Kiloelectronvolts to several Megaelectronvolts), the interactions of greatest interest are photoelectric effect, Compton scattering, and pair production. It is these interactions that will be the primary focus in this article.

SOME DEFINITIONS OF INTEREST

Before describing each interaction in further detail, it is useful to define terms that are commonly used in describing a photon beam.

Photon Fluence (Φ)

The photon fluence of an X-ray beam is the number of photons crossing a unit area. If we find that N photons pass through an area A perpendicular to the beam, then the photon fluence through the medium is

$$\Phi = \frac{N}{A}$$

Fluence Rate (Flux Density, ϕ)

The rate at which photons pass through an area is called fluence rate. If N photons pass through an area A in time t, then the fluence rate is

$$\phi = \frac{N}{At}$$

Energy Fluence (Ψ)

If the energy of each photon is $h\nu$ (where ν is the frequency and h is Planck's constant, $h = 4.135 \times 10^{-31}$ MeV/s), then the energy fluence is

$$\Psi = \frac{EN}{A} = \frac{dE}{dA} = h\nu\frac{dN}{dA}$$

Energy Fluence Rate (ψ)

Also called the intensity of the beam, and it is defined as ψ (or I)

$$I = \frac{EN}{At} = \frac{d\Psi}{dt}$$

Photon Attenuation

Linear Attenuation Coefficient (μ). Suppose that a monoenergetic beam consisting of N_0 photons is incident on an attenuator of thickness x (Fig. 1). The number of photons N that will pass through the attenuator and get registered by the detector may be written as

$$N = N_0 e^{-\mu x} \tag{1}$$

where μ is known as the *linear attenuation coefficient* of the attenuator and represents the fraction of incident photons that interact in a unit thickness of the medium. Therefore, μ represents the probability that a photon will

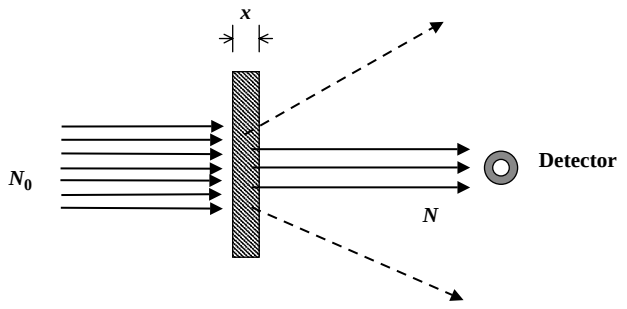


Figure 1. Experimental setup to determine photon beam attenuation in an absorbing medium.

interact in the absorber medium. The attenuation coefficient is a function of both the absorber material and the incident photon energy. As the incident beam is assumed to be monoenergetic, μ is a constant. Because μx must be a dimensionless number, when x is expressed in centimeters, the linear attenuation coefficient μ has units of 1/cm.

Equation 1 may also be written in terms of the beam intensity:

$$I = I_0 e^{-\mu x} \tag{2}$$

where I_0 is the intensity of incident photons and I is the transmitted intensity. The plot of intensity (I) versus thickness of absorber (x) is an exponential curve on a linear graph or a straight line on a semilogarithmic graph (Fig. 2). The plot shows that the percentage of photons removed from the photon beam is the same with each unit increase in absorber thickness.

Half-Value Layer. The thickness of the absorber that will attenuate a photon beam to 50% intensity is known as the half-value layer (HVL). It can be shown from equation 2 that

$$\text{HVL} = \frac{\ln 2}{\mu} = \frac{0.693}{\mu} \tag{3}$$

HVL is a measure of the penetrability of the beam: a high-energy X-ray beam (with low μ) will have a large HVL.

Tenth Value Layer. The tenth value layer (the thickness of the absorber required to reduce the transmitted intensity to one tenth of the original intensity) or TVL can be written as

$$\text{TVL} = \frac{\ln 10}{\mu} = \frac{2.302}{\mu} = 3.323(\text{HVL}) \tag{4}$$

The beam generated from an X-ray tube or a linear accelerator is not mono-energetic, but it is composed of a spectrum of energies. The average energy of such a beam of photons is approximately one third of the maximum photon energy. When this poly-energetic beam passes through the absorber, it is the low-energy X rays that are absorbed or filtered out first. After the low-energy photons have been removed, the filtered beam becomes more energetic or "harder." As the filter thickness increases, the average energy of the beam increases. Therefore, for an X-ray beam, the plot of beam intensity *versus* absorber thickness is not quite a straight line (Fig. 3). Instead, we see that the first HVL in the absorber is smaller than the second HVL; the second HVL is smaller than the third HVL, and so on.

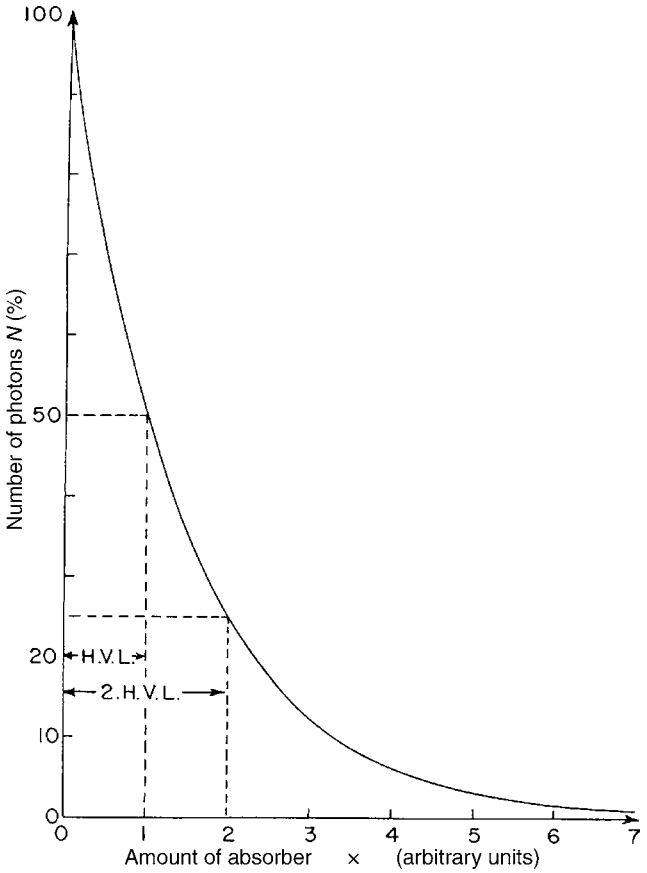


Figure 2. A graph of number (%) of photons transmitted through an absorber versus absorber thickness. Each HVL of absorber reduces the incident number of photons by half.

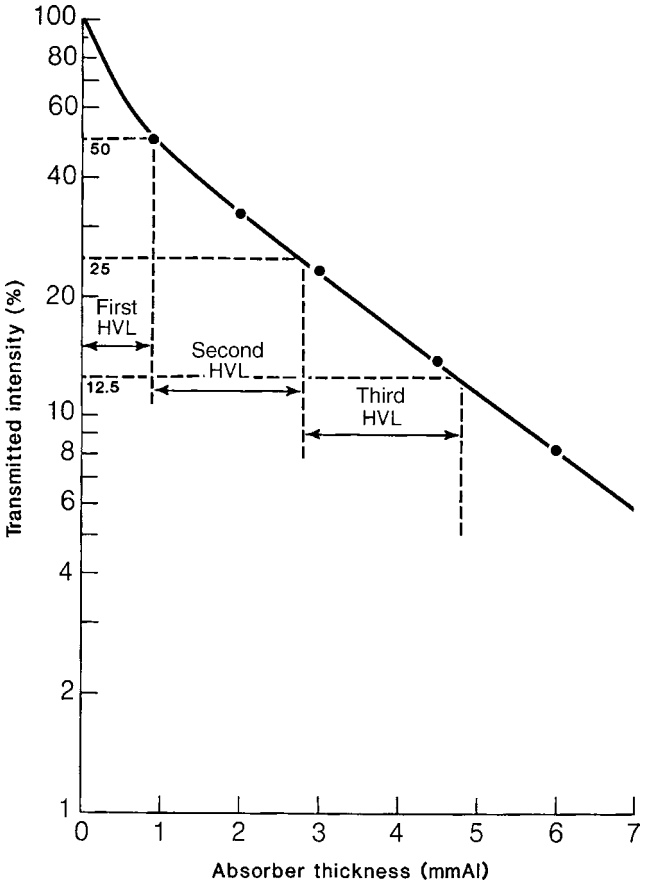


Figure 3. Transmitted intensity for a polyenergetic X-ray beam incident on an absorber. As the beam gets "harder" in passing through the absorber, the corresponding HVL gets larger.

Mean Free Path. As stated above, μ is the probability that a photon will undergo an interaction in a unit thickness of the absorber; therefore, $1/\mu$ can be thought of as the average distance a photon will travel between two successive interactions. Hence, the mean free path (MFP) is

$$\text{MFP} = \frac{1}{\mu} \tag{5}$$

It can be observed from equation 1 that approximately 37% of photons are left in the photon beam after traveling one MFP in the absorber.

Mass Attenuation Coefficient (μ/ρ). The linear attenuation coefficient (μ) varies with the density of the medium. As an example, water, water vapor, and ice have the same atomic composition, but their linear attenuation coefficients are all different due to the difference in physical density. However, if we divide the linear attenuation coefficient μ by the density ρ of the medium, we get a more fundamental quantity, known as the mass attenuation coefficient (measured in squared centimeter/gram):

$$\mu_m = \mu/\rho \tag{6}$$

The absorber thickness, in this case, is ρx and has the units of gram/squared centimeter (i.e., mass of the absorber per unit area). Here the density dependence of the attenuation coefficient has been removed and μ_m is a function of the material composition only. In the aforementioned example, water, water vapor, and ice all have the same mass attenuation coefficient μ_m.

The attenuation coefficient of a mixture of elements (such as water) can be calculated from the attenuation coefficients of individual elements:

$$\left(\frac{\mu}{\rho}\right) = \sum_i w_i \left(\frac{\mu}{\rho}\right)_i \tag{7}$$

where w_i is the fractional weight of the ith element in the mixture.

As photon beam attenuation depends on the number of electrons and atoms in the path of the beam, the corresponding attenuation coefficients can be defined as follows.

Electronic Attenuation Coefficient. When the absorber thickness is expressed as electrons/squared centimeter, the corresponding attenuation coefficient is the electronic attenuation coefficient. If N_e is the number of electrons/gram, then

$$\mu_e = \frac{\mu}{\rho}\left(\frac{1}{N_e}\right) = \frac{\mu}{\rho}\left(\frac{A}{N_A Z}\right) \tag{8}$$

where A is the atomic mass number, Z is the atomic number, and N_A is the Avogadro's number $= 6.02 \times 10^{23}$ atoms per atomic weight in grams. The electronic attenuation coefficient, μ_e (expressed in squared centimeter/electron) represents the probability that an incident photon will have an interaction with an electron in the absorber. As $Z/A \cong 0.5$ for all materials (except hydrogen), the number of electrons/gram is a constant for all materials: $N_e = (N_A Z/A) = 3 \times 10^{23}$ (Table 1). As we will see later, the near constancy of N_e across the periodic table has important implications for the Compton effect.

Table 1. Number of Electrons Per Gram

Material	Density (g/cm^3)	Atomic Number (Z)	Number of Electrons/Gram
Hydrogen	0.0000899	1	6.00×10^{23}
Carbon	2.25	6	3.01×10^{23}
Oxygen	0.001429	8	3.01×10^{23}
Aluminum	2.7	13	2.90×10^{23}
Copper	8.9	29	2.75×10^{23}
Lead	11.3	82	2.38×10^{23}
		Effective Z	
Fat	0.916	5.92	3.48×10^{23}
Muscle	1	7.42	3.36×10^{23}
Water	1	7.42	3.34×10^{23}
Air	0.001293	7.64	3.01×10^{23}
Bone	1.85	13.8	3.00×10^{23}

From Johns and Cunningham, *The Physics of Radiology*.

Atomic Attenuation Coefficient. Similarly, the atomic attenuation coefficient may be defined as the probability that the photon will have an interaction with an atom in the absorber. The coefficient is related to linear attenuation coefficient via

$$\mu_a = \frac{\mu}{\rho}\left(\frac{Z}{N_e}\right) = \frac{\mu}{\rho}\left(\frac{A}{N_A}\right) \tag{9}$$

The atomic attenuation coefficient μ_a is expressed in squared centimeter/atom, when the absorber thickness is in the units of atoms/squared centimeter.

We now look at each interaction in more detail.

COHERENT OR RAYLEIGH SCATTERING

Coherent scattering refers to the elastic scattering of a photon beam with atomic electrons, in which no energy transfer to medium takes place. The incident photon interacts with bound electrons as a whole (hence called coherent scattering). The electrons are temporarily set in motion (or oscillation) by the photon's electromagnetic field, and they return back to their original state by emitting a photon with the same energy as the incident photon (Fig. 4). The incident photon energy is too small to break free any electron from its shell. The scattered photon is emitted at a small angle relative to the incident photon. When a single electron from an atom participates in the process, the reaction is called Thomson scattering.

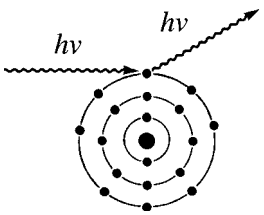


Figure 4. Illustration of coherent scattering: A photon with energy hv scatters off an atom without transferring any energy. The scattered photon has the same energy as the incident energy.

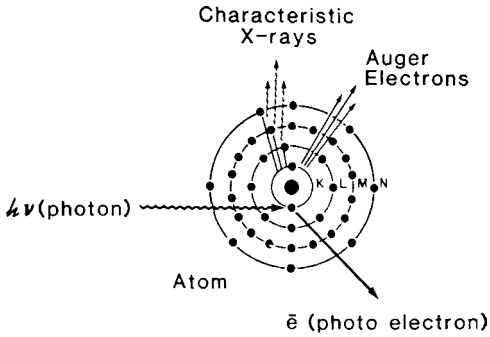


Figure 5. Photoelectric effect: A photon incident on an atom is absorbed with the emission of photoelectron. In the process, characteristic X rays and Auger electrons may also emitted.

The probability of coherent scattering (σ_{coh}) is highest when photon energy is very small and the medium atomic number Z is very large. Likewise, the probability of coherent scattering is lowest when Z of the medium is small and the photon energy is large. Consequently, coherent scattering is not very likely to occur in tissue and therefore is not of much interest in medical physics.

PHOTOELECTRIC EFFECT

In the photoelectric effect, the incident photon interacts with an atom and ejects one of the bound electrons (from K, L, M, or N shells) (Fig. 5). The photon disappears by transferring all of its energy to the atom. Some of the photon's energy is used to overcome the binding energy of the electron, and the rest changes into the kinetic energy of the electron:

$$K.E. = h\nu - E_b \tag{10}$$

where $h\nu$ is the incident photon energy and E_b is the binding energy of the electron. For the photoelectric effect to take place, the energy of the incident photon *must* be larger than the binding energy of the bound electron.

For example, suppose that the incident photon ejects an electron from the K-shell. After the photoelectron is ejected, a vacancy is created in the electron shell and the atom goes into an excited state. The vacancy left by the electron is filled by an outer shell electron (say, from an L-shell) with a lower binding energy. This leads to emission of characteristic X rays with energy $= E_K - E_L$. If the characteristic X rays have sufficient energy, they may also knock out electrons (called Auger electrons) from the surrounding shells. These Auger electrons leave behind more vacancies that in turn leads to generation of more X rays. The cycle is repeated until all of the photon energy is absorbed by the medium. In the tissue, the binding energies of electron shells are very small. Therefore, the characteristic X rays released in tissue are of very low energy and hence are locally absorbed. In the photoelectric energy range, all incident radiation on tissue is locally absorbed with no scatter radiation. In contrast, the characteristic radiation produced in high Z material (e.g., lead) is more energetic and is absorbed some distance away from the site of photoelectric effect.

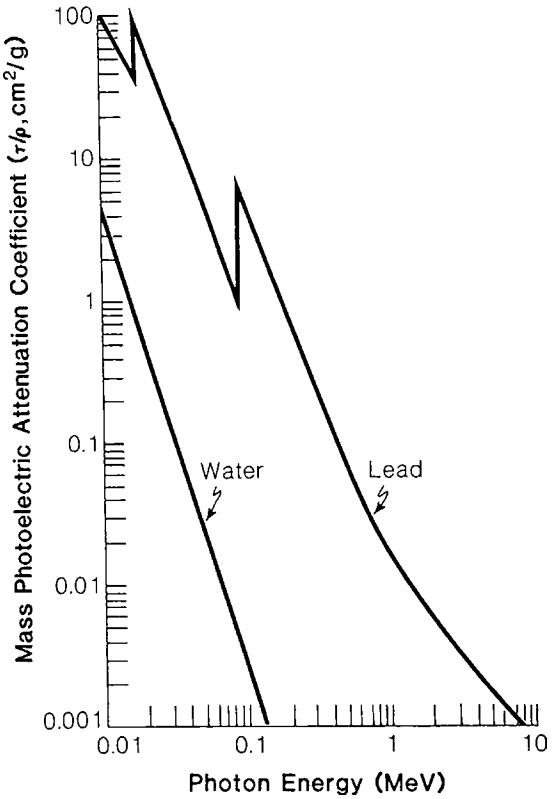


Figure 6. Variation of photoelectric mass attenuation coefficient with incident photon energy. Curves shown are for water ($Z = 7.82$) and lead ($Z = 82$).

The probability that a photon will undergo photoelectric effect depends strongly on its incident energy $E = h\nu$ and the atomic number Z of the absorbing material. In fact, the photoelectric mass attenuation coefficient τ/ρ varies directly as the cube of atomic number and inversely as the cube of photon energy (Z^3/E^3). Figure 6 shows a plot of τ/ρ versus energy for water ($Z = 7.5$) and lead ($Z = 82$). Due to the Z^3 dependence, a photon is 1000 times more likely to undergo photoelectric interaction in lead than in a water-like medium (e.g., tissue). Also, if it has sufficient energy to knock out an electron, a low-energy photon is more likely than a high-energy photon to participate in the photoelectric effect. It can be observed from the figure that as the photon energy increases, the probability of photoelectric interaction decreases rapidly (as $1/E^3$). The curve for lead also shows sharp discontinuities for incident photon energies of 15 keV and 88 keV. These peaks correspond to the binding energies of L- and K-shells, respectively, for lead and are called absorption edges. A photon having energy less than the electron's binding energy cannot undergo a photoelectric effect with electrons in that shell. However, as soon as the photon energy exceeds the binding energy of the electron, the probability of the photoelectric effect increases dramatically (like a resonance) and a sudden jump (or discontinuity) is observed in the plot of photoelectric attenuation coefficient versus energy. The discontinuity is greatest for the K-shell and becomes weaker for higher shells (L, M, N, etc.). Therefore, it is noticed that the most tightly bound

electrons have the greatest chance of undergoing photoelectric interaction. For water, the plot of τ/ρ again shows the same steady decline with energy, although the curve is much smoother. As the K-shell binding energy for a low Z medium is only about 0.5 keV, no discontinuities are observed.

The above behavior of the photoelectric effect has many important applications in medical physics:

1. Due to the Z^3 dependence, differential absorption of photons by bone, muscle, and fat is exaggerated and provides excellent X-ray film contrast in mammography and other diagnostic applications.
2. High Z materials, such as, $BaSO_4$ and Hypaque, are ideally suited for contrast enhancement in CT scanning.
3. As lead is a good absorber of low-energy photons, it is commonly used for shielding in diagnostic radiology procedures (for example, lead aprons).

The angular dependence of emitted photoelectron is a function of the photon's energy. For a low-energy photon, the electron has small kinetic energy and is emitted at 90° relative to the direction of the incident photon. As the photon energy increases, the photoelectrons also have higher energy and are emitted in a more forward direction.

COMPTON SCATTERING

As the incident photon energy increases (beyond 30–40 keV in tissue-like medium), the probability that it will undergo photoelectric effect decreases and the Compton effect becomes the dominant mode of interaction. This is the most important mode of interaction for tissue like materials.

In Compton scattering, the incident photon with energy hv_0 interacts with a loosely bound (or "free") electron from an outer shell and transfers some of its energy to it. The photon is scattered at a lower energy (hv') and scattering angle θ, and the recoil (or Compton) electron is ejected with an energy E at an angle ϕ relative to the incident photon's direction (Fig. 7). From conservation of energy and momentum, the following relations can be obtained:

$$\begin{aligned} E &= hv_0 \frac{\alpha(1-\cos\theta)}{1+\alpha(1-\cos\theta)} \\ hv' &= \frac{1}{1+\alpha(1-\cos\theta)} \\ \cot\phi &= (1+\alpha)\tan\theta/2 \end{aligned} \tag{11}$$

where $\alpha = hv_0/m_0c^2$ and $m_0c^2 = 0.511$ MeV is the rest mass energy of the electron.

In terms of a photon's wavelength, after undergoing a Compton interaction, the scattered photon has a longer wavelength λ' than that of the incident photon λ. The change in wavelength or "Compton shift" is independent of the incident photon energy and depends only on the scattering angle θ:

$$\Delta\lambda = \lambda' - \lambda = h/m_0c(1-\cos\theta) = \lambda_C(1-\cos\theta) \tag{12}$$

where $\lambda_C = h/m_0c = 0.02426$ Å $= 2.426 \times 10^{-10}$ m is the Compton wavelength, or the wavelength of a photon whose energy is just equal to the rest mass energy of the electron.

The scattered photon's energy, and therefore how much energy is imparted to Compton electron, depends on the scattering angle and incident photon energy. Let us first consider the dependence of transferred energy on the photon's scattering angle θ (Fig. 8). If the photon undergoes a "head-on" collision with the electron, the photon is scattered backward ($\theta = 180$) and the electron moves forward ($\phi = 0$). In this case, the photon transfers most of its energy to the electron ($hv' = hv'_{\min}$ and $E = E_{\max}$). If the photon makes a "glancing" hit with the electron ($\theta = 0$), the electron receives very little energy and the scattered photon continues forward with the same energy as the incident photon ($hv' = hv$ and $E = 0$). In summary, when a photon is scattered at small angles ($\theta \to 0$), very little of its energy is transferred to the electron. However, as the photon scattering angle increases ($\theta = 0 \to 180$), a greater fraction of the incident energy is imparted to the electron.

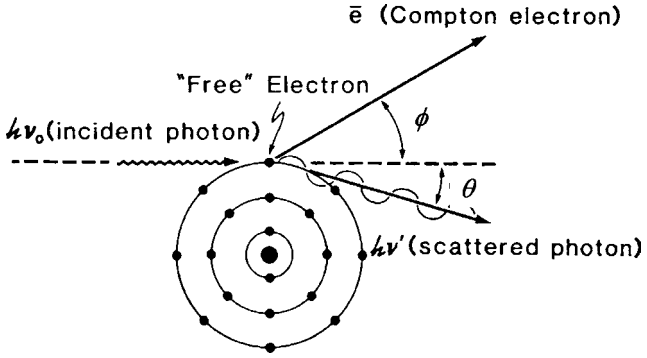


Figure 7. Compton effect: Incident photon scattering off a free electron.

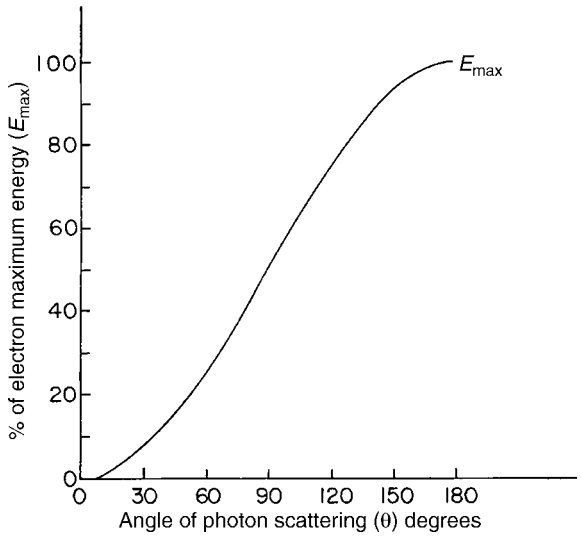


Figure 8. Compton scattering: Relationship between energy transferred to Compton electron and the scattering angle.

Unlike the photoelectric effect, however, the photon does not give all of its energy to the electron. The directional dependence of scattered electron and photon is governed by equation 11.

We next study the dependence of Compton effect on incident photon energy. For a low-energy photon, only a small portion of its energy is imparted to the electron. The scattered photon's energy is almost the same as the incident photon. Note that in the limiting case ($h\nu_0 \rightarrow 0$), this reduces to Rayleigh scattering where the incident photon suffers no loss of energy. As the photon energy increases, the fraction of energy transferred to the electron increases. For a very high-energy photon, the photon loses almost all of its energy to the electron and is emitted with a low energy.

It can be shown that a photon scattered at 90° can have energy of no more than 0.511 MeV regardless of the incident photon energy. If the photon is scattered backward, its maximum energy is only 0.255 MeV. These results are independent of the incident photon energy and have important consequences in the shielding design for treatment rooms in radiotherapy. For example, for side-scattered radiation (90° scatter), one needs to only shield for 0.511 MeV photons, and for backscattered radiation (180° scatter), the maximum required shielding is for 0.255 MeV photons. However, a photon scattered in the forward direction can have any energy up to that of the incident photon. As the incident photon energy increases, the recoil electron angle ϕ becomes smaller; i.e., the electron is more likely to be ejected in the forward direction.

As the Compton interaction involves a free electron, it is independent of the atomic number Z of the medium and depends only on the number of electrons per gram, which is constant for almost all materials (Table 1). Thus, the Compton mass attenuation coefficient σ/ρ is the same for all materials. In other words, gram for gram all materials will undergo the same Compton interaction. However, the linear attenuation coefficient σ will be larger for denser materials: In the Compton energy range, 1 cm of bone will attenuate more than 1 cm of tissue. The inherent contrast in the Compton range is due to density difference and not due to Z dependence as observed in the photoelectric effect. This fact, coupled with the presence of scattered photons, causes the mega-voltage X ray film quality to be inferior to that of kilovoltage films. In the soft tissue, the Compton effect is most important for photons with incident energy 0.1–10 MeV. As the photon energy increases, the probability of Compton interaction decreases.

PAIR PRODUCTION

A photon with energy more than 1.02 MeV may interact with the medium through pair production. In this reaction, the photon interacts with the field of the nucleus and disappears with the creation of a positive and negative electron (e^+/e^-) pair (Fig. 9). This reaction is an example of energy converting into mass. As the rest mass energy for electron or positron is $m_0c^2 = 0.511$ MeV, the threshold for pair production is 1.022 MeV. The total kinetic energy of the electron–positron pair is

$$h\nu - 1.022 = E^+ + E^- \tag{13}$$

The excess energy is generally shared equally between e^+ and e^-; however, any ratio is possible. As the products have equal and opposite charge, the net charge is conserved in the reaction.

The above reaction can also take place in the presence of an electron: a process known as triplet production (e^+/e^- pair and the interacting electron). The threshold energy for triplet production is 2.04 MeV. However, the likelihood of triplet production is small compared with pair production.

The electron and positron formed in the pair- (or triplet-) production lose their energy in passing through matter via ionization and excitation of atoms, until they come to rest. Near the end of its track, a positron combines with an available electron to produce two annihilation photons, each with an energy of 0.511 MeV (Fig. 9). The photons are emitted in opposite directions (180 degrees apart) to conserve momentum. This reaction is opposite to the pair production in that matter is now converted into energy.

The likelihood of pair production is very small when the photon energy is about 1–2 MeV. However, it increases rapidly with energy and becomes the dominant interaction for photons with an incident energy above 10 MeV. This behavior is in contrast to the other photon interaction processes considered so far that decrease in likelihood with increasing photon energy. The pair production interaction also increases with Z of the atomic nucleus.

It varies as Z^2/atom or Z/g. This implies that high-energy X rays will be readily absorbed in high Z materials, leading to "beam-softening." For this reason, lead is not recommended as a flattening filter in high-energy linear accelerators. For the same reason, HVL is not a useful concept in specifying beam quality of high-energy X rays. This can be explained by recalling that HVL is related to the attenuation coefficient as HVL = $0.693/\mu$. In the high-energy range, μ increases (or HVL decreases) as the photon

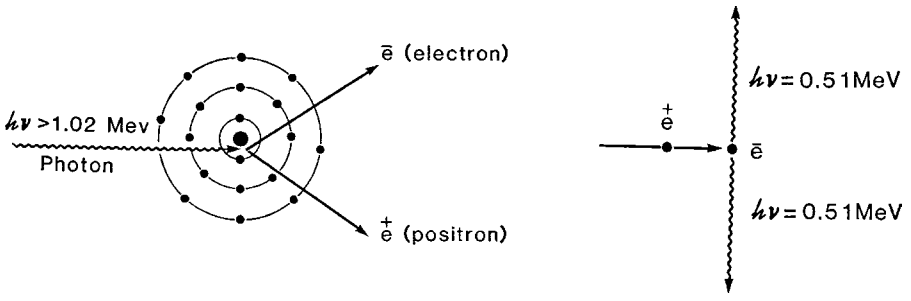


Figure 9. Illustration of pair production process and subsequent production of annihilation radiation.

energy increases. Therefore, HVL is not a meaningful quantity in describing a beam of high-energy photons.

PHOTODISINTEGRATION

A photon with still higher energy may penetrate the nucleus and knock out one of its constituents: a proton, neutron, or alpha particle along with more gamma rays. The incident photon is absorbed, and the nucleus is transformed into an unstable reaction product. The latter next returns back to a stable state via radioactive decay of a nuclear particle, for example, as in (γ, p) and (γ, n) reactions.

The threshold for the photodisintegration reaction is essentially the binding energy of nucleons in the nucleus. For low Z nuclei, this is above 10 MeV (except for Be: 2 MeV and ^{2}H: 1.5 MeV). For heavy nuclei, the nuclear binding energy is about 7 MeV. Due to Coulomb repulsion, the threshold for (γ, p) is lower than that for the (γ, n) reaction. Beyond the threshold energy, the probability for photodisintegration increases rapidly with increasing incident photon energy, reaches a maximum value, and then drops with further increase in energy. This peak is referred to as the *nuclear giant resonance* and is due to the electric dipole absorption of the incident photon.

Compared with the other reactions described above, the probability for photodisintegration is small and hence does not contribute significantly to photon attenuation. However, the interaction has great importance in shielding considerations for radiotherapy room design.

RELATIVE IMPORTANCE OF VARIOUS TYPES OF INTERACTIONS

Up to now we have considered how a monoenergetic beam of photons interacts with a medium having an atomic number Z. However, an X-ray beam from a radiotherapy machine is not monoenergetic, but it consists of a mixture of photons with various energies. Such a beam, in passing through an absorber, will undergo all of the above interactions to various degrees. The *total linear attenuation coefficient* may be written as the sum of the photoelectric, coherent, Compton, and pair production coefficients:

$$\mu_{\text{total}} = \tau + \sigma_R + \sigma_C + \kappa \tag{14}$$

Figure 10 shows a plot of μ_{total} versus photon energy along with contributions from individual component interactions.

The *total mass attenuation coefficient* may be written as

$$\frac{\mu_{\text{total}}}{\rho} = \frac{\tau}{\rho} + \frac{\sigma_R}{\rho} + \frac{\sigma_C}{\rho} + \frac{\kappa}{\rho} \tag{15}$$

The relative probability of various interactions depends on the incident photon energy and the Z of the material. In general, the photoelectric interaction is most common at low photon energies; the Compton effect dominates at intermediate energies and the pair production at high energies. As explained above, as coherent scattering is significant for low-energy photons (<10 keV) incident on high Z materials, it may be ignored.

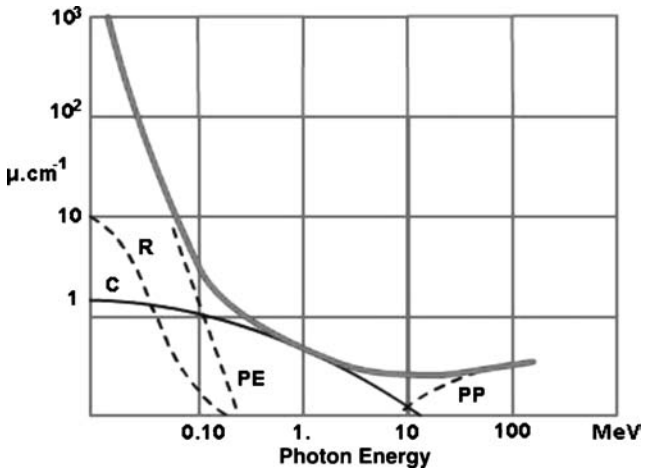


Figure 10. A plot of total linear attenuation coefficient that is composed of Rayleigh, photoelectric, Compton, and pair production processes.

Let us consider the plot of total mass attenuation coefficient as a function of photon energy for water and lead, which, respectively, represent a low Z and high Z material (Fig. 11). At low energies ($\geq$10 keV), the photoelectric effect is dominant. The attenuation coefficient displays Z^3/E^3 behavior in this region. Therefore, μ is much higher for lead (10^3 times) compared with water and decreases rapidly with increasing photon energy.

As the photon energy exceeds electron binding energy ($\sim$100 keV and higher), the Compton effect takes over. As the Compton interaction is independent of Z of medium, lead and water have practically the same attenuation. As the photon energy is further increased, the attenuation coefficient decreases until the pair production becomes the dominant mode of interaction (for photon energies higher than 1 MeV).

In the pair production range, the attenuation coefficient increases with energy as log E. Also, the likelihood of interaction is higher in lead than in water because of the Z dependence of pair production interaction. The above behavior of various modes of interactions is summarized in Table 2 for tissue-like material.

Another way to discuss the relative importance of various interactions is to look for them as regions of dominance in the plot of $h\nu$ versus Z (Fig. 12). The curves display points in the $h\nu$ and Z space for which the Compton effect equals the photoelectric effect and pair production. The Compton effect is most dominant between 1 and 5 MeV

Table 2. Relative Importance of Photoelectric, Compton, and Pair Production Processes in Water

Energy	Photoelectric	Compton	Pair Production
10 keV	95	5	0
25 keV	50	50	0
150 keV	0	100	0
4 MeV	0	95	5
24 MeV	0	50	50
100 MeV	0	16	84

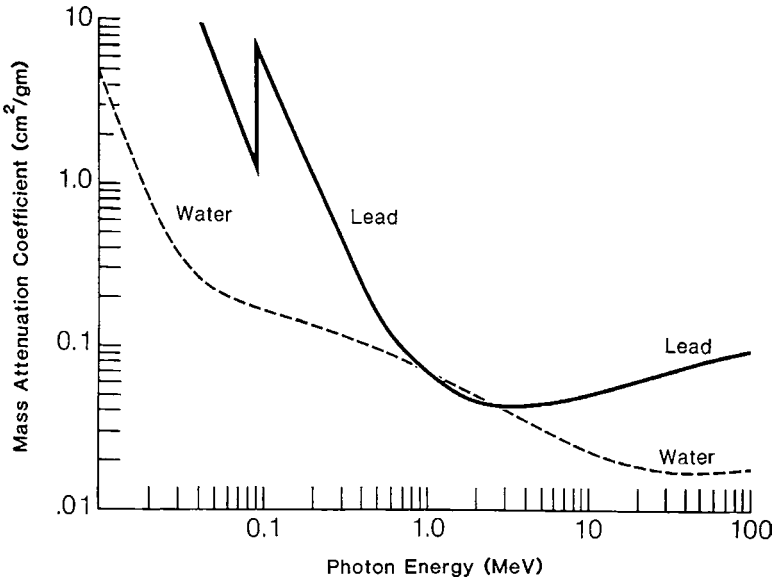


Figure 11. A plot of total mass attenuation coefficient as a function of incident photon energy for lead and water.

regardless of the type of material. According to this figure, for low Z (tissue-like materials), the Compton effect is the main interaction over a much wider energy range. For high Z materials, however, the photoelectric effect is dominant at low photon energies and pair production is the main interaction at high energies.

ENERGY TRANSFER AND ENERGY ABSORPTION COEFFICIENTS

When a beam of photons passes through an attenuator, either part or all of its energy is transferred to the medium (Table 3). The exact fraction of energy transferred depends on the incident photon energy and Z of the medium. If part of the incident energy is transferred, then the remaining photon energy is emitted as scattered photons. The scattered photons may further interact with the medium and lose some or all of their energy. Thus, a photon may undergo *multiple* interactions with electrons in the medium before it is absorbed or escapes out. Suppose hv is the incident photon energy, out of which, $\bar{E}_{tr}$ is the average energy transferred to the medium. Let E_{scat} be the energy of scattered photons; then

$$h_v = \bar{E}_{tr} + \bar{E}_{scat} \tag{16}$$

The fraction of photon energy transferred to electrons kinetic energy per unit absorber thickness can be written as

$$\mu_{tr} = \frac{\bar{E}_{tr}}{hv}\mu \tag{17}$$

where μ_{tr} is the linear energy transfer coefficient. The corresponding mass energy transfer coefficient is μ_{tr}/ρ.

Most of the energy transferred to the electrons is deposited in the medium at the site of interaction via ionization and excitation of atoms and is referred to as the absorbed energy E_{ab} (related to radiation dose delivered). However, a portion of the transferred energy is radiated away in the form of bremstrahlung radiation (E_{rad}):

$$\bar{E}_{tr} = \bar{E}_{ab} + \bar{E}_{rad} \tag{18}$$

Therefore, the energy absorption coefficient may be written as

$$\mu_{ab} = \frac{\bar{E}_{ab}}{hv}\mu \tag{19}$$

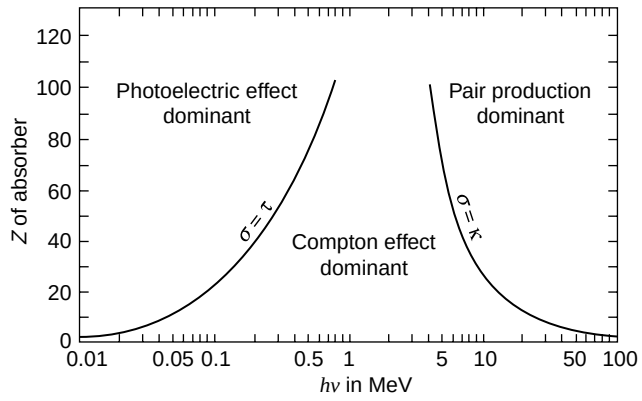


Figure 12. Relative importance of the three major types of X-ray interactions. The curves show the values of Z and hv for which two neighboring effects are equal.

Table 3. Depth of Maximum Dose and Percent Depth Dose at 10 cm Depth for Megavoltage Photon Beams

Photon Beam Energy	Depth of Maximum Dose (cm)	Percent Depth Dose at 10 cm depth (%)
1.25 MV (^{60}Co)	0.4	58.7
4 MV	1	63
6 MV	1.6	66.7
10 MV	2.5	73.2
18 MV	3.5	79.2
25 MV	5	84.5

X-RAY BEAM ENERGY PARAMETERS

As we have noted, the type of interaction an X-ray beam will undergo with a medium depends on its energy. This section presents some ways of defining energy of incident radiation.

The most comprehensive description of an X-ray beam is obtained via spectral energy distribution, i.e., a graph showing the relative population of different energy photons in the beam. The spectral distribution can be measured by several methods, including magnetic spectrometry, photoactivation, Cerenkov detection, and total-scintillation spectrometry, or it can be approximated by computational techniques based on bremstrahlung interaction in a target. Although useful, the exact determination of the spectra is labor intensive. In addition, the spectral distribution consists of a vast amount of information that makes it difficult to compare two X-ray beams using this information. Instead, alternative simpler methods may be used to describe the quality of the X-ray beam and its spectra.

Beam Energy, Maximum Energy, or Peak Energy

Typical X-ray spectrum consists of photons with energy ranging from 0 to a maximum energy (E_{max}). X-ray energy as nominally specified by the manufacturer is generally the energy of the peak intensity. This energy is the energy of the most probable electrons incident on the patient. The peak energy, however, is not a good indication of the X-ray spectrum, as two beams with same peak energy may have different spectrum.

Effective Energy

The X-ray beam is usually composed of photons of a mixture of energies. These will attenuate differently in matter. The effective energy of such a heterogeneous beam is defined as the energy of a monoenergetic beam that has the same HVL as the beam in question.

Half-Value Layer

HVL is the thickness of the absorber required to attenuate the beam intensity to half its original value.

Weighted Mean Energy

The weighted mean energy of a heterogeneous X ray beam is found from the mean mass-attenuation coefficient weighted by the energy fluence of the photons. The mean energy of the heterogeneous beam is approximately one third of the peak energy.

FURTHER READING

Khan FM. The Physics of Radiation Therapy. 3rd ed. Philadelphia, PA: Lippincott Williams and Wilkins; 2003.

Johns HE, Cunningham JR. The Physics of Radiology. 4th ed. Springfield, IL: Thomas; 1984.

Evans RD. The Atomic Nucleus. Malabar, FL: Krueger; 1955.

Cember H. Introduction to Health Physics. 3rd ed. New York: McGraw-Hill; 1996.

Ter-Pogossian MM. The Physics Aspects of Diagnostic Radiology. New York: Harper & Row; 1967.

Hendee WR, Ritenour ER. Medical Imaging Physics. 4th ed. New York: Wiley Science; 2002.

See also RADIATION THERAPY, QUALITY ASSURANCE; SCREEN-FILM SYSTEMS; THERMOGRAPHY.

X-RAYS, PRODUCTION OF

BRUCE HORN
Kaiser Permanente
Los Angeles, California

INTRODUCTION

Soon after Wilhelm Roentgen discovered X rays in 1895, scientists recognized their usefulness to visualize the internal anatomy of humans. For a number of years, X-ray tubes that radiologists used for this purpose were unreliable and produced low X-ray output. In 1913, William Coolidge invented the forerunner of the modern X-ray tube. Unlike earlier designs, this tube featured a heated tungsten filament as the source of electrons and utilized a high quality vacuum instead of the low gas pressure previously employed. The result was a reliable X-ray tube whose X-ray output could be reproducibly controlled over a wide range. Today's wide range of X-ray imaging techniques, from a simple chest radiograph to a multislice computed tomogram, still depend on an X-ray tube to produce X rays. Recent developments in X-ray tube design include improved heat handling capability, increased X-ray output, and reduced focal spot size.

PRODUCTION OF X RAYS

X rays, like radio waves, light, γ rays, and other types of electromagnetic radiation, are defined by their energy and source. X rays are produced by two methods involving the interaction of electrons with matter. In the first method, a stream of electrons directed at a target is decelerated by forces between the incident electrons and the atomic nuclei of the target material. Since a deceleration implies that kinetic energy is lost, one of the ways in which electrons lose energy is by creating photons of electromagnetic energy equal in energy to the kinetic loss. This process is call bremsstrahlung, or "braking radiation". The photons produced by this interaction are commonly called X rays. If a large number of electrons, all having the same initial kinetic energy, interact with a target material, the resulting bremsstrahlung consists of photons, or X rays, with a continuum of energies from zero to a maximum equal to the initial electron kinetic energy. This range of energies is called the X-ray spectrum. The shape of the spectrum remains the same regardless of the kinetic energy of the electrons. Although X rays are produced by this mechanism even if the target consists of a gas, the efficiency of X-ray production by bremsstrahlung is directly proportional to the product of the atomic number (Z) of the target

material and the kinetic energy of the electrons. Since the portion of the kinetic energy of the incident electrons that is not converted to X rays results in heating the target material, it is desirable to utilize high energy electrons and target materials of high atomic number to increase the efficiency of bremsstrahlung production as much as possible. However, in either therapeutic or diagnostic medical applications, X-ray beam penetration is controlled by selecting the initial kinetic energy of the electrons, which is controlled by the electrical potential difference between the electron source (cathode) and the target (anode). Although other considerations may dictate the choice of target material when designing X-ray tubes for some diagnostic applications, such as mammography (molybdenum or rhodium), the target material remains as the variable that can be optimized to provide the most efficient X-ray production. Due to its high atomic number, relative high melting point, good heat-transfer characteristics, and low cost, the best target material for use in diagnostic imaging applications is tungsten. At energies commonly employed in medical applications, approximately ~1% of the kinetic energy of the electrons is converted to bremsstrahlung by interaction with a tungsten target. The remaining 99% of the incident energy results in heating of the target. As will be discussed later, the primary challenge of X-ray tube design is how to most effectively handle heat generated by the X-ray production process.

In the second method, electrons impinging on the target also interact with the orbital electrons of the atoms in the target. If the incident electron has a kinetic energy greater than or equal to the binding energy of an orbital electron, the bound electron may be ejected from the atom as a result of the interaction. The resulting vacancy is then filled by one of the other orbital electrons falling into the potential well of the vacancy. This transition results in the release of a photon with an energy equal to the difference between the electronic states of the vacancy and the orbital electron that fills the vacancy. The released photon is called a characteristic X ray. Due to the quantized energy levels of orbital electrons, characteristic X rays have specific energies independent of the original kinetic energy of the accelerated electrons. For tungsten, the *K*-shell binding energy is 69 keV, and the L-shell binding energy is 11 keV. Characteristic X rays resulting from interactions with *L*-shell electrons are not important for medical applications since the resulting low energy photons are not useful and are severely attenuated by the encapsulating materials used in a practical X-ray tube. As a result, for practical purposes, the X-ray spectrum from a tungsten target produced by electrons with kinetic energies <69 keV will not exhibit characteristic X-ray production. X-ray spectra produced by electrons with kinetic energies >69 keV will have a series of narrow peaks of increased X ray production at photon energies corresponding to the energy differences for the possible K-shell orbital transitions superimposed on the bremsstrahlung spectrum. Figure 1 illustrates the typical shape of bremsstrahlung spectra, as well as the difference in characteristic X-ray production between two accelerating potentials. The existence of characteristic X rays in the spectra is relatively unimportant for most medical diagnostic imaging procedures. On the other hand, the characteristic X rays produced by a molybdenum target X-ray tube are important to the process of image formation for film-screen mammography.

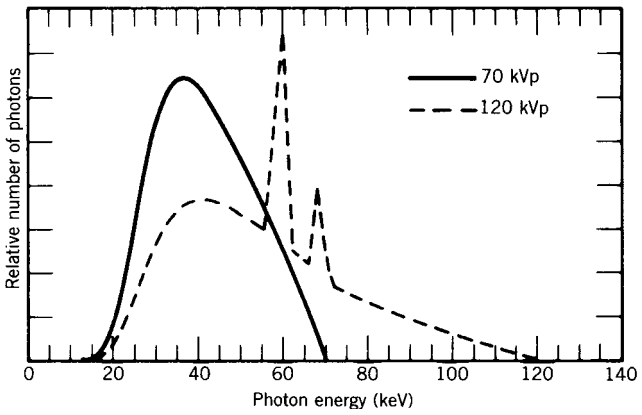


Figure 1. Diagnostic X-ray spectra.

CONSTRUCTION OF X-RAY TUBES

An X-ray tube consists of a cylindrical envelope (tube insert) that contains the primary components necessary to produce X rays in a controlled manner. This envelope is contained within a tube housing that provides physical protection of the tube insert, stray radiation protection and a means of transferring heat generated during the X-ray production process from the insert to the surrounding environment. The insert contains an electron source (filament), a means of directing the electrons in the intended direction (focusing cup), and a target (anode). A typical X-ray tube insert is shown in Fig. 2. Unlike radiation emitted by radioactive material, X-ray production can be electrically switched on and off by controlling the voltage applied to the X-ray tube.

Envelope

The envelope housing the X-ray generating components is usually constructed of blown glass. However, some specialized X-ray tubes have metal or ceramic envelopes to improve the tube's heat handling capabilities. Air is evacuated from the envelope until an acceptable vacuum is achieved so that the electrons used in the X-ray generation process will not interact with gas molecules.

Filament

A small helical coil of tungsten wire, similar to the filament in an ordinary light bulb, is used to produce the electrons needed for the X-ray generation process. Passing an electric current through this tungsten filament, heats it to the point where free electrons are sufficiently energized to escape from the filament. This process is called thermionic emission. The freed electrons form a cloud, or space charge, around the filament. A separate filament is provided for each focal spot size, usually two per tube. During the X-ray generation process, a large negative electrical potential difference is applied to the filament relative to the anode

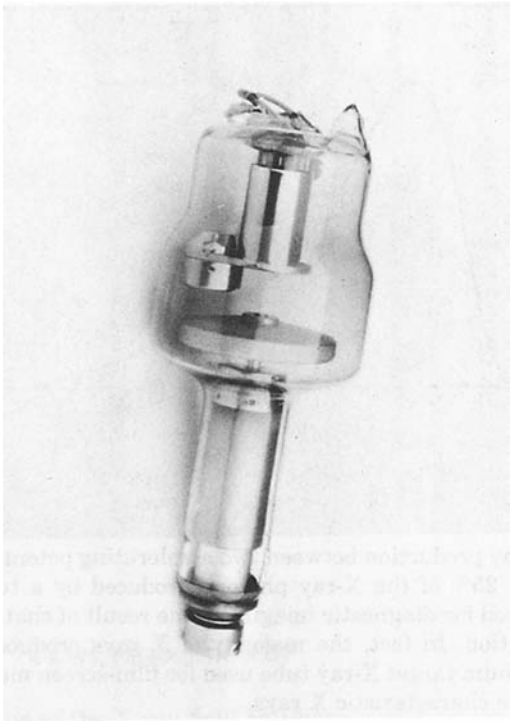


Figure 2. Typical medical X-ray tube insert.

so that electrons supplied by the cloud are repelled by the filament and attracted by the anode. If large quantities of electrons (high X-ray output) are needed at low accelerating potentials, the repulsive force of the space charge may inhibit thermionic emission and reduce the rate at which electrons are released.

Focusing Cup

A metal focusing cup (cathode) surrounds the filament on all sides, except that facing the anode. Figure 3 shows a typical focusing cup and filaments. During the X-ray generation process, the same large negative potential difference that is applied to the filament is also applied to the focusing cup. The surface of the focusing cup adjacent to the filament is shaped in such a way as to focus the beam of accelerated electrons onto the anode. Therefore, the size of the filament and shape of the focusing cup control the size of the focal spot on the surface of the anode.

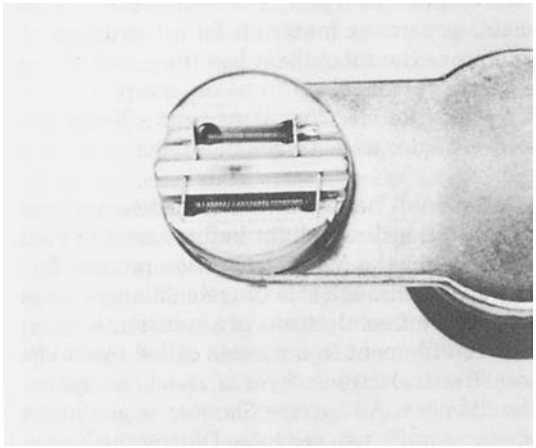


Figure 3. Focusing cup and filaments.

Anode

The anode (target) is the medium with which the electrons interact to produce X rays. In a typical X-ray tube used for medical applications, the anode is constructed of tungsten or an alloy of rhenium and tungsten. In the case of film-screen mammography, molybdenum or rhodium is often used for X-ray tube anodes because its characteristic X rays are better suited for breast imaging than the spectrum of X rays from tungsten. It is not necessary that the entire anode be constructed of the target material. Since the interactions of the electrons with the target take place near the surface, the target can be relatively thin and backed by a material that exhibits better heat-transfer characteristics, such as copper or graphite. During the X-ray generation process, a large positive electrical potential difference is applied to the anode to attract electrons emitted by the filament. As a result, the electrons are accelerated during their passage from the filament to the anode by the potential difference between the filament (cathode) and the anode, typically 20,000–150,000 V. Because the voltage supplied to the X-ray tube from a single-phase, full-wave-rectified source varies from zero to maximum and back to zero every 8.3 ms, this accelerating voltage is specified by its maximum value in units of peak kilovolts (kVp). Although three-phase or high frequency voltage sources exhibit considerably less voltage variation, their voltage waveforms are also described in terms of peak voltage. An X-ray tube with an anode-to-cathode potential difference of 100 kVp produces a spectrum of X rays with a maximum energy of 100 keV.

The surface of the anode is angled slightly from perpendicular with the electron beam. This angle is referred to as the target angle. The area on the anode where the electrons interact to produce X rays is called the focal spot or focus. The geometry of the X-ray tube components is designed so that the X-ray beam exits the tube at a 90° angle with respect to the electron beam inside the tube. The effective size of the focal spot is the size of the spot on the anode projected along the central axis of the X-ray beam. Since the width of the focal spot is the dimension along the axis perpendicular to both the electron beam and the X-ray beam central axis, only the length of the focal spot is affected by the target angle. The effective focal spot size is important because it determines an X-ray tube's ability to image small objects. The focal spot area in the plane of the surface of the anode is important because it limits the number of electrons per unit time that can impinge on the target without overheating. The relationship between the effective focal spot size and the actual size on the anode is given by the following formula:

$$f = F \sin \alpha$$

where f is the length of the focus effective size, F is the length of the actual focus on the anode, and α is the target angle. The concept of having a large focal area, over which

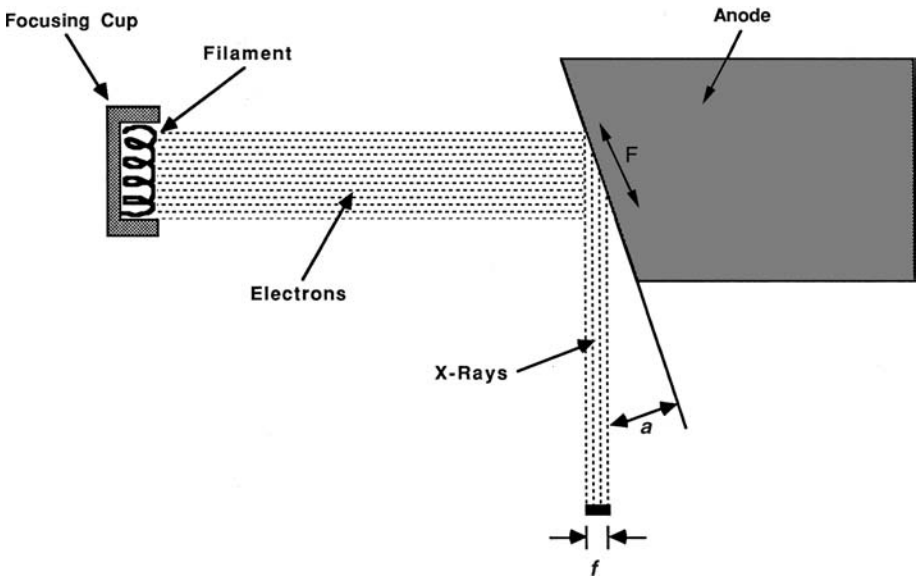


Figure 4. Line-focus principle.

to distribute heat while maintaining a small focal area for imaging by use of an angled target, is called the line-focus principle and is demonstrated in Fig. 4. For example, a 12° target angle with an effective focal spot of 1.0 × 1.0 mm has an actual focus size of 4.8 × 1.0 mm.

As illustrated in Fig. 5, the maximum X-ray field size is the largest field size that can be produced by the tube at a specified distance from the focal spot. Since X rays emitted by the focal spot at an angle greater than the target angle will be absorbed by the anode, the maximum field size (centered on the X-ray beam) along the dimension parallel to the anode–cathode axis of the X-ray tube is dependent on the target angle as follows:

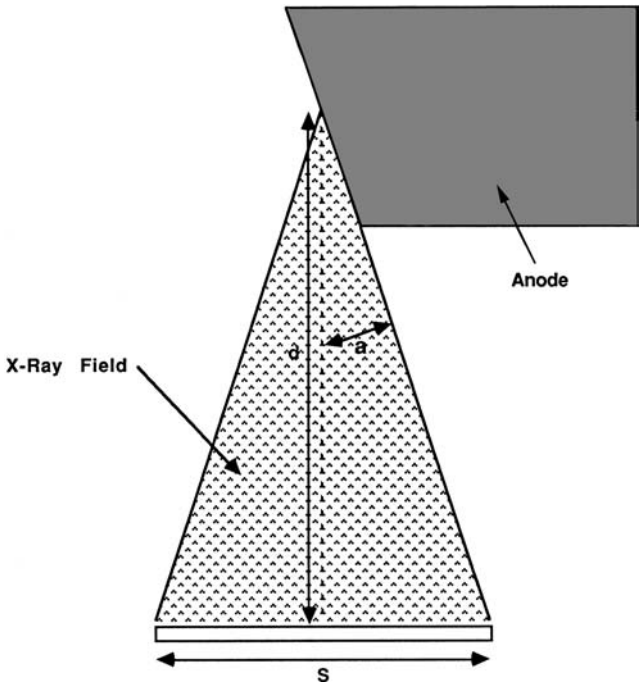


Figure 5. Maximum X-ray field size.

where S is the maximum size of the X-ray beam, d is the distance from the focal spot along the central axis of the X-ray beam, and α is the target angle. Therefore, although a larger target angle will result in a larger maximum field size, it will also result in lower heat loading capability. Conversely, a smaller target angle will result in a higher heat loading capability, but a smaller maximum field size. Simple algebra using the formula for maximum field size results in the conclusion that a target angle of at least 12° is required to cover the typical diagnostic imaging maximum field size of 43 cm at the commonly used distance of 100 cm. Diagnostic X-ray tube target angles typically are in the range of 12–15°, with a few special-purpose X-ray tubes having target angles as small as 9°. The intensity distribution of X rays along the direction parallel with the anode–cathode axis of the X-ray tube is not uniform. As shown in Fig. 6, comparison of the intensity at points equidistant from the central axis of the X-ray beam and equidistant from the focus indicates that the intensity of points toward the anode is less than that of corresponding points toward the cathode. This effect is called the heel effect. Since the X rays are generated from interactions of

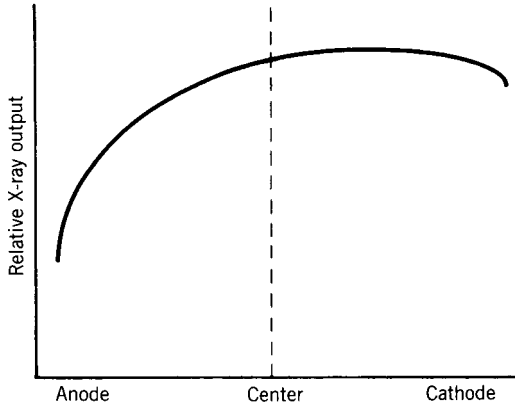


Figure 6. Heel effect.

the electrons slightly below the surface of the anode, X rays at the anode end of the field will have had to travel through additional anode material when compared with the cathode end. This attenuation of the X rays by the anode material produces a reduced intensity at the anode end. It also selectively reduces the number of low energy photons, which causes the spectrum of X-ray energies at the anode end of the X-ray field to have relatively more high energy photons and fewer low energy photons than the cathode end. As a result, the X rays at the anode end have a higher average penetrating power.

Due to the limiting effect of target angle on field size and the desire for higher heat loading capabilities, the rotating anode was developed. Rather than continuously directing the electron beam toward the same area on the anode throughout an X-ray exposure, directing the electron beam to different areas on the anode during the exposure will produce higher heat loading capability. By using a disk-shaped anode and rotating the disk about its central axis, the focus traces a circular path, or focus track, on the anode during an exposure. Since heat is distributed over a larger area, the X-ray tube's heat loading capability is greatly increased. For example, a rotating anode with a 10 cm diameter focus track and a 1.0 mm focal spot is capable of a heat loading >300 times that of the same anode held stationary during the exposure. Due to their exceptional heat loading advantage, essentially all X-ray tubes used for medical imaging have rotating anodes. On the other hand, fixed-anode tubes are used for dental radiography since the techniques employed do not produce enough heat to require the use of a rotating anode. Typical rotating-anode diameters are 7.5–10 cm, with some special purpose X-ray tubes employing anodes that have a diameter of 12.5 cm.

The tungsten disk in a rotating-anode X-ray tube is supported on its central axis by a molybdenum stem that is concentrically connected to a copper cylinder, or rotor. The rotor is supported internally on ball bearings to allow the entire anode/stem/rotor assembly to rotate. Figure 7 shows a typical tungsten rotating-anode assembly. Unlike a fixed-anode tube in which heat is transferred from the anode by conduction, the mounting of a rotating anode is designed to insulate the bearing from the heat contained in the anode and, as a result, to prevent damage to the bearings. Heat transferred to the bearings decreases their lubrication, resulting in premature bearing failure. A rotating anode transfers heat to the surrounding tube housing by radiation, rather than by conduction.

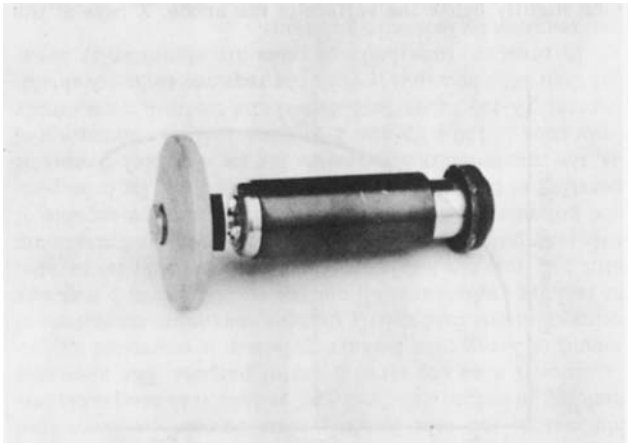


Figure 7. Rotating anode assembly.

Although the electrons are focused to a specific area on the anode, they may rebound from the focus with sufficient energy to interact at other positions on the anode to generate X rays. This off-focus radiation produces imaging artifacts. Generally, rotating-anode tubes generate more off-focus X rays than fixed-anode tubes because they have larger tungsten surfaces with which to interact. Although a significant amount of the output of an X-ray tube may be due to off-focus radiation, a well-designed collimator attached to the tube will reduce the affect of this radiation on clinical imaging.

Focus

As stated previously, a typical diagnostic X-ray tube has two focal spots, one approximately half the size of the other, with nominal sizes ranging from 0.3 to 2.0 mm. The larger focus is used for clinical applications that demand the shortest possible exposure time (consequently maximum heat generation). Conversely, the smaller focus is used for applications that require the capability to resolve the smallest anatomical features. X-ray tubes incorporating two focal spot sizes are usually built so that their foci are superimposed on the surface of the anode, even though two separate filaments are used to produce the two focal spots. This design alleviates alignment problems that would occur if the foci were not superimposed. However, X-ray tubes have been manufactured that employ a biangular anode so that the focus tracks of the two focal spots are on separate areas of the anode with different target angles. This design allows optimization of the target angle for each of the foci, but also causes some alignment problems.

When selecting an X-ray tube, one of the important factors to consider is the size of the focal spot. In order to have a standardized method of defining focal spot size, the National Electrical Manufacturers Association (NEMA) in the United States and the International Electrotechnical Commission (IEC) in Europe have each developed standards that address focal spot size specification and measurement. Since the X-ray tube manufacturing process is somewhat imprecise, the labeled focus size is a useful nominal value when comparing the relative sizes among different tubes. However, the actual focal spot size may vary considerably from the nominal size. In fact, measured focus length as much as 100% greater than the nominal size are considered to be within the tolerance specified by the standards. For example, a nominal 1.0 mm focus may have a measured length as large as 2.0 mm. Several methods of characterizing the focal spot have been developed: (*1*) slit method, (*2*) resolution or star pattern method, and (*3*) pinhole method.

The slit method employs a rectangular slit 10 μm wide and at least 5000 μm long used at enlargement factors of 1.0–3.0. The long axis of the slit is placed perpendicular to the axis of the focal spot whose dimension is desired. The resulting image on the film is a one-dimensional (1D)

distribution of the focal spot intensity along the chosen axis. Two images are therefore necessary to measure both the width and length of the focal spot. Both NEMA and the IEC specify the slit method as the method to be used for measuring the dimensions of focal spots. However, while the NEMA standard specifies measurement by eye using a 5–10X magnifying glass that has a built-in calibrated graticule, the IEC standard specifies the use of a scanning microdensitometer to obtain the line-spread function (LSF) of the width or length of the focal spot. The dimension of the focal spot is specified as the full width at 15% of maximum of the LSF.

The resolution method measures the ability of the focal spot to spatially resolve small objects and specifies the results in terms of spatial frequency (e.g., line pairs per mm), but does not provide a measurement of the dimensions of the focal spot. In this method, a line-pair resolution phantom is placed on the central axis of the X-ray beam between the focal spot and a single-emulsion film. The line-pair phantom consists of alternating lines of absorbing material (usually lead) and spaces, where adjacent lines and spaces have the same width. The width of the line pairs varies systematically over the face of the phantom to provide a range of line-pair sizes. The largest line-pair width, for which the line pair is completely blurred, is proportional to the focal spot size along the X-ray tube axis perpendicular to the line pairs. In order to facilitate the measurement of focal spot sizes by this method, a specialized phantom, commonly called a star pattern, which consists of alternating wedges of absorbing material and spaces in a circular configuration, was developed. Star pattern line-pair widths vary continuously from the outside edge to the center of the phantom. Using this phantom, both dimensions of the focal spot can be characterized from the same exposure. Since the phantom is circularly symmetric, the orientation of the phantom relative to the axes of the X-ray tube is unimportant. The image of the phantom on the film is analyzed by finding the points closest to the outside edge of the phantom for which the line pairs are completely blurred. The focal spot resolution can then be calculated using the following formula:

$$R = 180M/\pi\theta D$$

where R is the focal spot resolution in line pairs per millimeter, θ is the angle in degrees of a single wedge of absorbing material, D is the diameter of the blur on the film in millimeters, and M is the magnification factor of the phantom image. For the typical star pattern phantom with 2° wedges, this formula reduces to

$$R = 28.65M/D$$

The diameter corresponding to the resolution of the focal spot parallel with the anode–cathode axis of the X-ray tube (length) is the blur diameter that is perpendicular to this axis. Likewise, the diameter corresponding to the focal spot resolution perpendicular to the anode–cathode axis (width) is the blur diameter that is parallel with the anode–cathode axis. The principal advantage of the star pattern method of characterizing focal spots is its ease of use. It does not require long exposure times and evaluates both the width and length of the focal spot from one exposure. The disadvantage of this method is that it does not provide an indication of the shape or intensity distribution of the focal spot.

The pinhole method utilizes a small circular pinhole 30–100 μm in diameter to image the focal spot onto a single-emulsion X-ray film for the purpose of determining its overall shape, orientation, intensity distribution and relative location. Although this method was used in the past to measure the dimensions of the focal spot, it is no longer considered suitable for doing so. The pinhole must be carefully aligned with the central axis of the X-ray beam. The distances between the focal spot and pinhole and between the pinhole and the film are adjusted so as to provide an enlargement factor of 1.0 or 2.0, depending on the nominal focus size. The film is exposed for a sufficient amount of time to produce an image of the focal spot with a density of 0.8–1.2 O.D. above base plus fog. Figure 8 illustrates the measurement geometry. The pinhole image is examined by eye using a magnifying glass, or by use of a scanning microdensitometer to quantitatively characterize the focal spot.

The focal spot size depends on the technique factor used during an exposure. Generally, the size of the focus increases as the tube current (mA) increases, and slightly decreases as the tube potential (kVp) increases. An increase in X-ray tube current means that more electrons travel from the filament to interact with the anode. The self-repulsion of the electrons causes the electron beam to slightly defocus, thereby interacting with a larger area on the anode. Since neighboring electrons in the beam are closer and more plentiful for larger tube currents, the effect of self-repulsion increases the focus size. An increase in

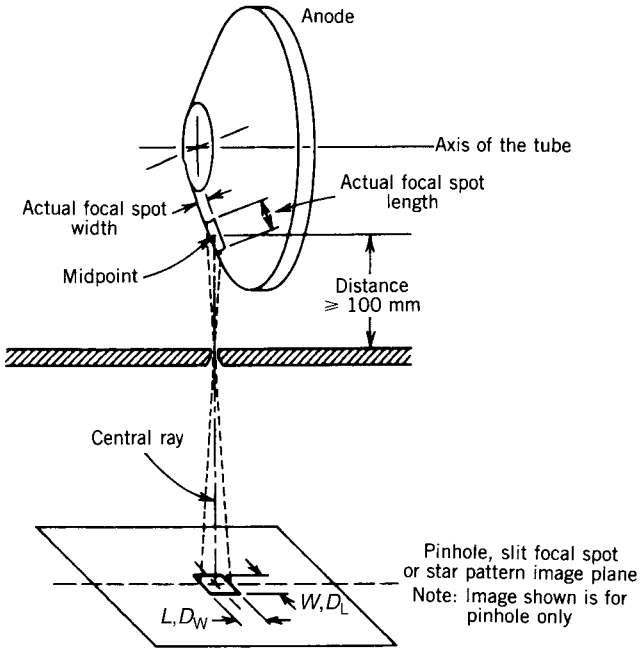


Figure 8. Focal spot size measurement geometry.

X-ray tube potential means that the electrons that travel from the filament to the anode are more energetic, and therefore have higher velocities. In addition, since neighboring electrons experience self-repulsion for a shorter period of time, the effect of self-repulsion is diminished and the focus size decreases. It is because of the effect of tube potential and current on focus size that both NEMA and the IEC specify the exposure factors for which focal spot sizes are measured. For a typical diagnostic X-ray tube, factors of 75 kVp, 50% of the maximum rated tube current at this tube potential, and 0.1 s exposure duration are employed.

Due to the line-focus principle, the projected focal spot size is dependent on the position in the X-ray field from which the actual focus on the anode is viewed. The observed target angle from a point at the cathode end of the X-ray field is larger, thereby producing a larger focal spot. Conversely, the observed focal spot size at the anode end is smaller. For example, the projected focus size at the cathode edge of a 40-cm field is 1.9 mm for an X-ray tube with a 12° target and a nominal focus size of 1.0 mm. All focal spot sizes are therefore specified relative to the central axis of the X-ray beam.

Stator

By surrounding the neck of the X-ray tube with a coil of wire in combination with a copper rotor attached to the anode inside of the tube forms an electric motor that rotates the anode. Exciting the coil with 60 Hz alternating current (ac) results in the anode being rotated at 3600 rpm, or one revolution every 16.7 ms, since it acts as a synchronous motor. However, this rotation speed is insufficient to avoid overheating the focal spot while meeting the clinical need for short, intense X-ray exposures of only a few milliseconds duration. For example, a 3 ms exposure would use only 18% of the available focal track. By applying 180 Hz ac to the stator windings so that the anode is accelerated to ~10,000 rpm within 1–2 s, heat will be spread over a greater percentage of the focal track (e.g., 50% of the track for a 3 ms exposure). The delay while the anode accelerates to the desired speed is called the "prep" time. During the prep time, the filament circuit is boosted so that the filament will have already reached its operating temperature by the time the X-ray exposure is initiated (when the anode has reached its desired speed). As soon as the exposure is completed, the accelerating voltage is removed from the stator. Due to mechanical resonances in the anode bearings, it is necessary to apply a braking voltage to the stator windings to decelerate the anode from 10,000 to 3,600 rpm within ~20 s. This action prevents destruction of the bearings by mechanical resonances. The anode is allowed to coast to a standstill once its speed is <3600 rpm. Maintaining the anode speed at 10,000 rpm for long periods of time causes the bearings to wear out quickly.

Window. In most diagnostic X-ray tubes, the glass envelope of the tube insert serves as a window through which the X-ray photons pass. However, for X-ray tubes designed for applications requiring low energy photons, a beryllium window is used because beryllium, unlike glass, does not significantly attenuate low energy photons.

Tube Housing

The X-ray tube insert must be contained within a metal tube housing, shown in Fig. 9, which provides a number of functions: mechanical protection against breakage of the glass envelope, an electrically grounded enclosure for safety, lead shielding against stray radiation emitted by the X-ray tube, a means to transfer heat from the anode to the outside environment, connections for the electrical cables from the high voltage power supply or generator, as well as a mechanical mounting for attachments. A glass or plastic window in the tube housing that is aligned with the central axis of the X-ray beam serves as a means for the useful X-ray beam to exit the tube. It also offers a means for directly viewing the anode to determine its physical condition. The volume between the insert and the tube housing is filled with oil to aid the transfer of heat deposited in the anode from the insert to the housing.

Although the X-ray photons emitted toward the window are desired, X-rays are also produced in many other directions. For this reason, the tube housing must contain sufficient lead shielding to reduce this "leakage" radiation to acceptable levels, since it does not serve a useful purpose. The allowable amount of X-ray tube leakage radiation is controlled by government regulations that specify that the maximum magnitude of the leakage radiation shall not exceed 100 mR in 1 h at any point 1 m from the focal spot when the X-ray tube is operated at its maximum continuous rated tube current and maximum

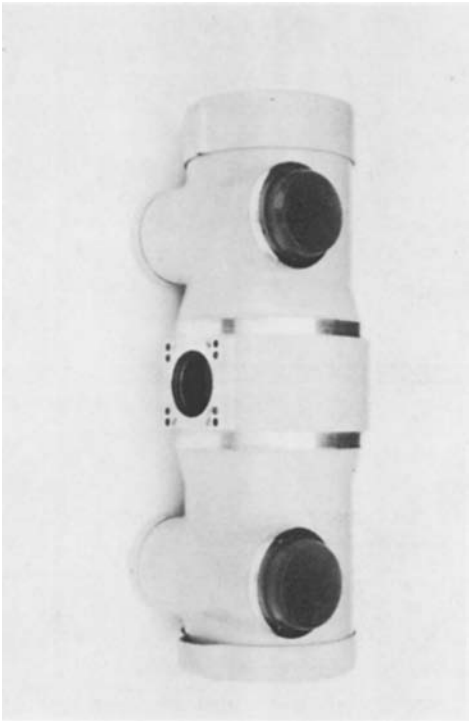


Figure 9. Typical medical X-ray tube housing.

tube potential. Although an X-ray tube may be capable of operating at tube currents as high as 1000 mA for brief periods, the maximum continuous rating for a typical diagnostic X-ray tube is 150 kVp and 3–6 mA. All manufacturers of diagnostic X-ray tubes, both dental and medical, must certify that their tube housings provide the required degree of shielding. In practice, tube leakage rates are significantly lower than required by regulations.

Most X-ray systems balance the tube potential difference between the anode and cathode so that the voltage difference between anode and ground is equal to the voltage difference between ground and the cathode. In other words, a 100 kVp tube potential means that the anode is 50 kVp positive with respect to ground, and the cathode is 50 kVp negative with respect to ground. For this reason, special high voltage connectors are required for both the anode and cathode. The standard high voltage connector contains three receptables, each insulated from ground. In the anode connector, all three receptacles are electrically connected together, as well as to the anode. In the cathode connector, one receptacle is connected to the cathode, one receptacle is connected to the small focus filament, and one receptacle to the large focus filament. The receptacle connected to the cathode also acts as the common connection for the small and large filaments. In this way, the filaments are maintained at the same potential difference from ground as the cathode or focusing cup. The exceptions are X-ray tubes that utilize a biased focusing cup to shape the focal spot distribution, necessitating a cathode high voltage connector with four receptacles instead of three. The fourth receptacle is used to apply a bias voltage between the focusing cup and the filaments. Since the connectors are standardized, X-ray tubes from different manufacturers can be easily used with an X-ray system.

As discussed previously, X-ray photons emitted by the tube have a spectrum of energies. Since lower energy photons have little penetrating power, they tend to be absorbed near the skin surface of a patient undergoing a diagnostic exam and do not contribute to the X-ray image. For patient protection purposes, it is therefore desirable to eliminate or filter these low energy photons before they reach the patient. By reducing the low energy content of the beam, the materials used in the construction of the tube insert and tube housing that are in the path of the useful X-ray beam serve this purpose. Aluminum permanently mounted in the tube housing window provides additional filtration. All of the filtration that is permanently in the useful beam is referred to as the inherent filtration. Government regulations specify the minimum amount of inherent filtration that must be present in the X-ray beam as a function of the maximum tube potential at which the tube is operable.

HEAT DISSIPATION IN X RAY TUBES

Much of the design of an X-ray tube is predicated on the heat generated by the inefficiencies of the X-ray production process. Although the use of a rotating anode increases the ability of the focal spot to handle the instantaneous heat load during the short period of X-ray generation, there are limitations as to the total amount of heat that the anode and tube housing can each store. Basically, there are five parameters that describe the heat dissipation characteristics of an X-ray tube: (*1*) focal spot loading, (*2*) anode heat capacity, (*3*) anode cooling rate, (*4*) housing heat capacity, and (*5*) housing cooling rate.

These heat characteristics are described in terms of one of two units, heat units or watt·seconds. In the United States, heat parameters are specified in heat units (HU). The number of heat units generated by an X-ray exposure is given by the following formula:

$$\mathrm{HU} = CVAt$$

where HU is the number of heat units, C is a waveform factor, V is the tube potential in peak kilovolts, A is the tube current in milliamperes, and t is the exposure duration in seconds. The waveform factor corrects for the differences in heating effect resulting from differences in average power of different high voltage waveforms. For a full-wave-rectified, single-phase waveform, the waveform factor is 1.0. For a 12-pulse, three-phase waveform, the waveform factor is 1.35. For example, an X-ray exposure of 100 kVp, 500 mA, 0.2 s using a single-phase generator would result in 10,000 HU. The same technique using a three-phase generator would result in 13,500 HU.

In Europe, heat parameters are expressed in watt·seconds (W·s) or joules (J). The heat generated by an X-ray exposure is given by the following formula:

$$E = KVAt$$

where E is the heat in watt·seconds or joules, K is a waveform factor, V is the tube potential in peak kilovolts, A is the tube current in milliamperes, and t is the exposure duration in seconds. In this case, the waveform factor for a full-wave-rectified, single-phase waveform is 0.74 and for a 12-pulse, three-phase waveform is 1.0. Therefore, the quantity of heat represented by one heat unit is 26% less than the quantity represented by 1 W·s or joule. Using the technique factors of the previous example (100 kVp, 500 mA, 0.2 s) results in 10,000 J of heat generated.

Focal Spot Loading

The heat handling characteristic of the X-ray tube focal spot is expressed in terms of kilowatts for an exposure duration of 0.1 s. The kilowatt loading of the focal spot is equal to the product of the tube potential in peak kilovolts and the tube current in milliamperes divided by 1000. It is the amount of power, or energy per second, deposited as heat in the anode. Typical X-ray tube inserts used in diagnostic radiology for high power procedures have focal spot loading capabilities of 30–40 kW for a small focus of 0.6 mm, and 100 kW for a large focus of 1.2 mm. The small focus of such a tube could not be used for the example technique of 100 kVp, 500 mA, because the focal spot loading would be exceeded. This demonstrates the trade-off between focal spot size and heat loading capability. A focal spot of 0.3 mm may have a maximum heat loading of as little as 9 kW. A tube insert with this rating could not be

used with a tube current >90 mA at 100 kVp, or >125 mA at 72 kVp, for an exposure duration of 0.1 s.

Anode Heat Storage Capacity

The maximum amount of heat that can be stored in the anode at any one time is its heat capacity. Depending on the manufacturer, the heat capacity is specified in terms of heat units or joules. The size and construction of the anode determine its heat capacity. Rotating anodes 3 in. (7.6 cm) in diameter have less heat storage capacity than anodes 4 in. (10.2 cm) in diameter. The typical "high power" diagnostic X-ray tube has a heat capacity of 300,000–400,000 HU. This is equivalent to the heat generated by 22–29 three-phase exposures of 100 kVp, 500 mA, 0.2 s each. Since the heat storage capacity is defined by the amount of energy deposited in the anode that raises its temperature from some reference temperature to the maximum operating temperature of the anode, the specified heat storage capacity will depend on the choice of reference temperature. For example, the storage capacity of an insert may be 400,000 HU when referenced to a "warm" anode, but is 600,000 heat units when referenced to ambient room temperature, that is, a "cold" anode. For this reason, it is important to be aware of the corresponding reference temperatures when comparing anode heat storage capacities of tube inserts of different manufacturers.

Anode Cooling Rate

Heat deposited in the anode is transferred to the tube housing. The rate at which the anode cools is dependent on the temperature difference between the anode and the tube housing. When the maximum rated amount of heat is stored in the anode, the cooling rate will be maximum; when the anode is cold, the cooling rate will be minimum. For this reason, the cooling rate of the anode is given as a cooling curve of heat stored in the anode as a function of time, with no additional heat deposited, as illustrated in Fig. 10. For an X ray tube with a maximum anode cooling rate of 72,000 HU·min^{-1}, at least 11 s would be required to transfer the heat generated by a 100 kVp, 500 mA, 0.2 s three-phase exposure from the anode to the tube housing. Since the maximum continuous heating rate is equal to the maximum cooling rate, the anode of this tube could be subjected to 1200 HU·s^{-1} continuously (e.g., 100 kVp, 8.9 mA).

Tube Housing Heat Capacity

Just as the anode is limited as to the amount of heat that can be stored in it at any one time, the heat capacity of the tube housing is also limited. Since the housing has a much larger mass than the anode, its heat capacity is significantly greater. The typical diagnostic X-ray tube housing has a heat storage capacity of 1,500,000 HU. This is equivalent to the heat generated by 111 three-phase exposures of 100 kVp, 500 mA, 0.2 s each.

Tube Housing Cooling Rate

Heat deposited in the tube housing is transferred to the surrounding ambient environment. The maximum housing

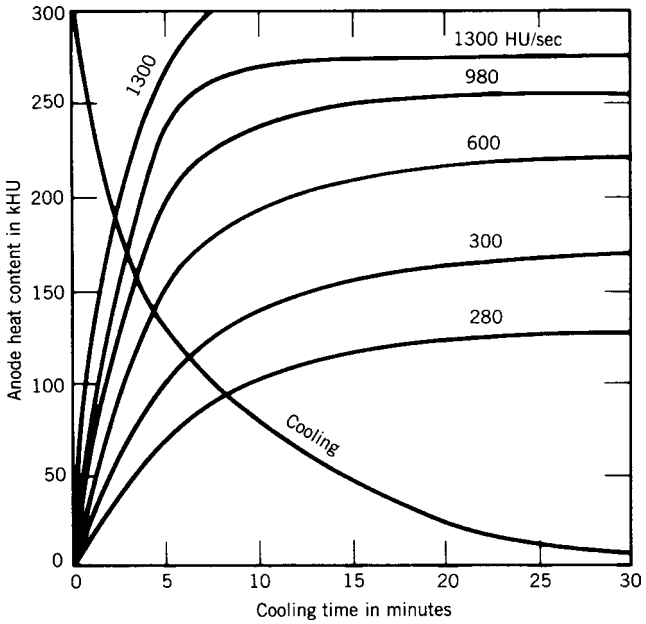


Figure 10. Anode cooling/heating curves.

cooling rate occurs when the housing is heated to its maximum capacity; the minimum cooling rate occurs when the housing contains little heat. The maximum cooling rate of a typical tube housing is ~15,000 HU·min^{-1}, unless additional measures have been taken to increase the cooling rate. A typical tube housing that has a fan to increase air flow over the housing has a maximum cooling rate of ~24,000 HU·min^{-1}. Another method of increasing the cooling rate is to constantly circulate the oil contained in the housing through an external heat exchanger. This method results in a maximum cooling rate of ~50,000 HU·min^{-1}. For high work load applications where X-ray exposures occur frequently enough that the heat generated by the exposures will not be totally dissipated from the X-ray tube before the next exposure, the cooling rate of the housing determines the rate at which heat can be generated in the tube. It is for this reason that manufacturers use the additional cooling measures discussed. For example, at least 54 s would be required to transfer the heat generated by a 100 kVp, 500 mA, 0.2 s three-phase exposure from a housing with a cooling rate of 15,000 HU·min^{-1}. The addition of a fan would reduce this cooling time to 34 s.

Tube Rating Chart

To allow X-ray equipment users to easily determine whether a desired exposure technique exceeds the heat characteristics of an X-ray tube, all tube heat parameters are combined into one chart called a tube rating chart. A chart is provided by the manufacturer for each focal spot size and anode rotation speed. Each chart consists of a set of curves of tube potential versus exposure duration for different tube currents that defines the maximum single exposure capability of the X-ray tube when the tube is cold. An exposure technique that falls "below" the corresponding

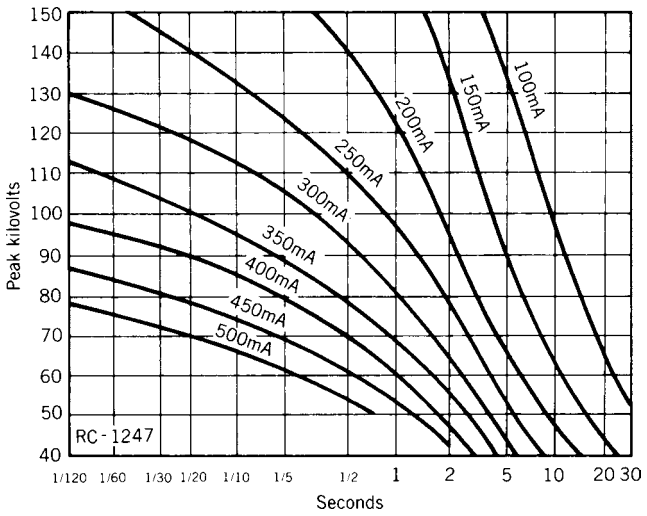


Figure 11. Tube rating chart.

curve on the chart is allowed; a technique that is "above" the corresponding curve is not allowed.

Using the tube rating chart shown in Fig. 11, the point corresponding to a technique of 100 kVp, 500 mA, 0.2 s is above the curve for 500 mA. Therefore, this technique exceeds the heat capabilities of this particular X-ray tube and is not allowed. Examination of this 500 mA rating curve reveals that a tube potential of 100 kVp is not allowed, regardless of the duration of the X-ray exposure. The point corresponding to 100 kVp and 0.2 s indicates that a maximum tube current of ~280 mA is obtainable for 100 kVp and 0.2 s exposure duration. Similarly, the point where the 500 mA rating curve intersects 0.2 s indicates that a maximum tube potential of 54 kVp is obtainable for a 500 mA and 0.2 s exposure technique. Since the amount of film blackening is proportional to the product of the tube current and exposure duration, the technique of 100 kVp, 500 mA, 0.2 s could be changed to 100 kVp, 250 mA, 0.4 s without affecting the resulting diagnostic image. The point on the tube rating chart for this latter technique is below the 250 mA rating curve and the technique is therefore allowed. Some manufacturers provide tube rating charts that provide rating curves of tube current versus exposure duration for different tube potentials, but these can be utilized in a similar manner.

Single-exposure tube rating charts are not useful for determining the feasibility of multiple radiographic exposures (or radiographic exposures combined with fluoroscopic exposures) that may come close to exceeding a tube's heat capabilities. The appropriate cooling and heating curves, combined with direct calculation of heat units produced, must be used in these cases. The data sheets for X-ray tubes designed for serial exposure applications such as cineangiography often include tables of maximum allowable techniques for serial exposures.

X-RAY TUBE FAILURE

Like electric light bulbs, X-ray tubes have a limited length of life. Due to the electrical, mechanical, and heat stresses that an X-ray tube experiences during its normal use, there are several types of X-ray tube failures.

Bearing Wear

Since the X-ray tube insert is a sealed envelope, the bearings of a rotating anode can only be lubricated at the time of manufacture. Although the insert is designed to prevent heat deposited in the anode from reaching the bearings, sufficient heating of the bearings takes place in the long term to gradually reduce their lubrication. Prolonged operation at high rotor speeds also rapidly decreases the lifetime of the bearings. Eventually, the bearings become noisy and, if the tube insert is not replaced, will seize, thus preventing the anode from rotating.

Filament Evaporation

By design, an electric current is used to heat the filament to produce electrons. If this current is excessive, the filament will melt, resulting in an open filament. In addition, evaporation of the filament material is a natural consequence of filament heating. This may result in a loss of filament material sufficient to cause failure.

Cracked Anode

If an X-ray exposure is made with the anode stationary, the anode may crack due to intense heat build-up at the focal spot. This may be caused by either bearing or stator failure. Even if the problem of the stationary anode is resolved (i.e., the stator repaired) the cracked anode presents a safety problem because it may shatter, implode the glass envelope, and pierce the tube housing. It is also possible to crack a rotating anode if the anode is subjected to a very high heat-generating technique when it is cold. For this reason, some manufacturers recommend that X-ray tubes be warmed up by making several low tube-current, long-duration exposures (70 kVp, 75 mA, 6.0 s) whenever the tube has not been used for 1 h or more.

Tungsten Deposition

The heating of the tungsten filament and tungsten anode during normal operation of the X-ray tube causes some of the tungsten to evaporate and to build up gradually as a deposit on the glass envelope of the tube insert. This deposit causes the glass envelope to be less able to repel electrons reflected from the anode because it reduces the electrical insulating properties of the glass envelope, thereby reducing the magnitude of the induced negative charge on the envelope. The electron bombardment of the glass releases trapped gases that cause the electrical breakdown (arcing) of the tube when the amount of gas released becomes great enough.

BIBLIOGRAPHY

Further Reading

Brecher R, Brecher E. The Rays: A History of Radiology in the United States and Canada. Baltimore (MD): Williams & Wilkins; 1969.

Coulam CM, Erickson JJ, Rollo FD, James AE Jr. The Physical Basis of Medical Imaging. New York: Appleton-Century-Crofts; 1981.

Hendee W. Medical Radiation Physics—Roentgenology, Nuclear Medicine and Ultrasound, 2nd ed. Chicago (IL): Year Book Medical Publishers; 1979.

International Electrotechnical Commission. X-Ray Tube Assemblies for Medical Diagnosis—Characteristics of Focal Spots. Publ. No. 60336, Geneva: IEC, 2005.

National Electrical Manufacturers Association. Measurement of Dimensions and Properties of Focal Spots of Diagnostic X-Ray Tubes. Stand. Publ. No. XR 5-1992 (R1999), Washington (DC): NEMA; 1999.

Ter-Pogossian MM. The Physical Aspects of Diagnostic Radiology. New York: Harper & Row; 1969.

See also CODES AND REGULATIONS: RADIATION; SCREEN-FILM SYSTEMS; X-RAY EQUIPMENT DESIGN.

References for Radiotherapy, heavy ion:

BIBLIOGRAPHY

1. Mould RF. A Century of X-rays and Radioactivity in Medicine: Institute of Physics Publishing; 1993.
2. Attix FH. Introduction to Radiological Physics and Radiation Dosimetry: John Wiley & Sons; 1986.
3. Khan FM. The Physics of Radiation Therapy, Second Edition: Williams & Wilkins; 1994.
4. Hendee WR, Ibbott GS, Hendee EG, Radiation Therapy Physics: Wiley-Liss; 2005.
5. Petti PL, Lennox AJ, Hadronic radiotherapy, Ann Rev Nuclear Part Science 1994;44:154–197.
6. Scharf WH. Biomedical Particle Accelerators: AIP Press; 1994.
7. Bewley DK. The Physics and Radiobiology of Fast Neutron Beams: Institute of Physics Publishing; 1992.
8. Podgorsak EB, Metcalfe P, Van Dyk J. Medical Accelerators: Chapter 11 in Van Dyk J ed. The Modern Technology of Radiation Oncology, Medical Physics Publishing; 1999.
9. Chu, WT, Ludewig BA, Renner TR. Instrumentation for treatment of cancer using proton and light-ion beams. Rev Sci Instrum 1993;64:2055–2122.
10. Chen GTY. Radiotherapy, heavy ion; Verhey LJ, Proton beam radiotherapy. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation. New York: John Wiley & Sons; 1988.
11. Castro JR. Results of heavy ion radiotherapy, Radiation and Environmental Biophysics (Historical Archive) 1995;34:45–48.
12. Heavy charged particles in reseach and medicine. Proceedings of a Symposium. Berkeley, California, May 1–3, 1985, Radiat Res Suppl. 1985;8:SI-334.
13. Kraft G. RBE and its interpretation, Strahlentherapie und Onkologie 175 Suppl. 1999;2:44–47.
14. Matsufujil N, Fukumura A, Komori M, Kanai T, Kohno T. Influence of fragment reaction of relativistic heavy charged particles on heavy-ion radiotherapy, Phys Med Biol 2003;48: 1605–1623.
15. Pollock BE ed., Contemporary Stereotactic Radiosurgery: Technique and Evaluation: Futura Publishing Company; 2003.
16. Bomford CK, Kunklery IH, Walter and Miller's Textbook of Radiotherapy Radiation Physics, Therapy, and Oncology 6th Ed. Churchill Livingstone; 2003.
17. Purdy JA, Grant III W, Palta JR, Butler EB, Perez CA eds. 3-D Conformal and Intensity Modulated Radiation Therapy: Physics & Clinical Applications: Advanced Medical Publishing, Inc., 2001.
18. Webb S, The Physics of Three-Dimensional Radiation Therapy: Conformal Radiotherapy, Radiosurgery and Treatment Planning: Institute of Physics Publishing; 1993.
19. Hongstrom Kr, Paciotti MA, Smith AR, Coller M, Comparison of static and dynamic treatment modes for the pion therapy beam at LAMPF. Int J Radiat Oncol Biol Phys 1980;6(12): 1693–700.
20. Sisterson J ed. Particle Therapy Co-operative Group (PTCOG) Newsletter, Mass. General Hospital, Harvard. MA. Jan. 2005.
21. Rossi S, Amaldi U, The TERA programme: status and prospects. In: Larssson B, Crawford J, Wienreich R, eds. Adv. in Neutron Capture Therapy, Vol. I: Elsevier Science; 1997.
22. Podgorsak EB, Podgorsak MB, Special Techniques in Radiotherapy: Chapter 17. In: Van Dyk J, ed., The Modern Technology of Radiation Oncology, Medical Physics Publishing; 1999.
23. Podgorsak FB, Podgorsak MB, Stereotactic Irradiation: Chapter 16. In: Van Dyk J, ed., The Modern Technology of Radiation Oncology, Medical Physics Publishing; 1999.
24. Washington CM, Leaver D, Principles and Practice of Radiation Therapy, Second Edition: Mosby; 2004.
25. Wang CC, ed. Clinical Radiation Oncology: Indications, Techniques, and Results Second Edition: Wiley-Liss; 2000.
26. Hall EJ, Radiobiology for the Radiologist, 4th Edition: J.B. Lippincott; 1994.
27. Rossi HH, Zaider M, Microdosimetry and Its Applications: Springer; 1996.
28. Greene D, Williams PC, Linear Acceleraators for Radiation Therapy, 2nd Edition: Institute of Physics Publishing; 1997.
29. Nias AHW, An Introduction to Radiobiology: John Wiley & Sons, 1990.
30. Prasad KN, Handbook of Radiobiology, 2nd ed.: CRC Press; 1995.
31. Tsipenyuk YM, The Microtron: Development and Applications: Taylor & Francis; 2002.
32. Dobelbower RR, Abe M, Intraoperative Radiation Therapy: CRC Press; 1989.
33. Chiu C, Fomytskyi M, Grigsby F, Raischel F, Downer MC, Tajima T, Laser electron accelerators for radiation medicine: A feasibility study. Medical Physics 2004;31:2041–2042.
34. Bostick WH, Possible techniques in direct-electron-beam tumor therapy, Phys Rev 1950;77:564–565.
35. Shih CC, High energy electron radiotherapy in a magnetic field. Medical Physics 1975;2:9–13.
36. Whitmire DP, D.L. Bernard DL, Peterson MD, Magnetic modification of the electron-dose distribution in tissue and lung phantoms. Medical Physics 1978;5:409–417.
37. Bielajew AF, The effect of strong longitudinal magnetic fields on dose deposition from electron and photon beams. Medical Physics 1993;20:1171–1179.
38. Litzenberg DW, Fraass BA, McShan DL, O'Donnell TW, Roberts DA, Becchetti FD, Bielajew AF, Moran JM, An apparatus for applying strong longitudinal magnetic fields to clinical photon and electron beams. Phys Med Biol 2001;46: N105–N115.
39. Becchetti FD, Sisterson JM, Hendee WR, Point/counterpoint: high energy electron beams shaped with applied magnetic fields could provide a competitive and cost-effective alternative to proton and heavy-ion radiotherapy. Medical Physics 2002;29:2435–2437.
40. Chen Y, Bielajew AF, Litzenberg DW, Moran JM, Becchetti FD, Magnetic confinement of electron and photon radio-

therapy dose — a Monte Carlo simulation with a non-uniform longitudinal magnetic field. Med Phys 2005;32:3810–3818.
41. Maughan RL, Yudelev M, Neutron Therapy: Chapter 21. In: Van Dyk J ed., The Modern Technology of Radiation Oncology. Medical Physics Publishing, 1999.
42. Forman JD, Yudelev M, Bolton S, Tekyi-Mensah S, Maughan R, Fast neutron irradiation for prostate cancer, Cancer and Metastasis Reviews 2002;21:131–135.
43. Zamenhof RG, Busse PM, Harling OK, Goorley JT, Boron Neutron Capture Therapy: Chapter 24. In: Van Dyk J ed., The Modern Technology of Radiation Oncology, Medical Physics Publishing; 1999.
44. Larsson B, Crawford J, Wienreich R, eds. Adv. in Neutron Capture Therapy, Elsevier Science; 1997.
45. Moyers MF, Proton Therapy: Chapter 20. In: Van Dyk J ed., The Modern Technology of Radiation Oncology, Medical Physics Publishing; 1999.
46. Goitein M, Lomax AJ, Pedroni ES, Treating cancer with protons, Physics Today, Sept. 2002;45–50.
47. Coutrakon G, Bauman M, Lesyna D, Miller D, Nausbaum J, Slater J, Johanning J, DeLuca PM, Siebers J, Ludewigt B, A prototype beam delivery system for the proton medical accelerator at Loma Linda. Medical Physics 1991;18:1093–1099.
48. Litzenberg DW, Roberts DA, Lee MY, Pham K, Vander Molen AM, Ronningen R, Becchetti FD, On-line monitoring of radiotherapy beams: experimental results with proton beams. Medical Physics 1999;26:992–1006.
49. Li Q, Kanai T, Kitagawa A, A model to evaluate the biological effect induced by the emitted particles from a beta-delayed particle decay beam. Physics Med Biol 2003;48:2971–2986.
50. Jaekel O, Kraemer M, Karger CP and Debus J. Treatment planning for heavy-ion radiotherapy: clinical implementation and application. Phys Med Biol 2001;46:1101–1116.
51. Li Q, Kitagawa A, Kanazawa M, Urakabe E, Kanai T, Tomitani T, Sato S, Wei Z. The production of 9C beam in the secondary beam line of the HIMAC facility and its potential application in cancer therapy, Nuclear Physics A 2004;746: 288–2292.
52. Miyamoto T, Yamamoto N, Nishimura H, Koto M, Tsujii H, Mizoe J, Kamada T, Kato H, Yamada S, Morita S, Yoshikawa K, Kandatsu S, Fujisawa T, The Working Group for Lung Cancer, Carbon ion radiotherapy for stage I non-small cell lung cancer. Radiotherapy and Oncology 2003;66:127–140.
53. Hirao Y, Ogawa H, Yamada S, Sato Y, Yamada T, Sato K, Itano A, Kanazawa M, Noda K, Kawachi K, M. Endo M, Kanai T, Kohno T, Sudou M, Minohara S, Kitagawa A, Soga G, Takada E, Watanabe S, Endo K, Kumada M, Matsumoto S, Heavy ion synchrotron for medical use – HIMAC project at NIRS – JAPAN, Nuclear Physics A 1992;538:541–550.
54. Krämer M, Jäkel O, Haberer T, Kraft G, Schardt D, Weber U, Treatment planning for heavy-ion radiotherapy: physical beam model and dose optimization. Phys Med Biol 2000; 45/11:3299–3317.
55. Kim J, Marti F, Blosser H, Study of a superconducting cyclotron for heavy-ion therapy, Cyclotrons and their applications, 2001 F. Marti ed., AIP Conference Proceedings Vol 2002;600-1:324–326.

INDEX

Italic numbers preceding page references indicate the volume numbers. Page numbers followed by f denote figures; page numbers followed by t denote tables.